FUZZY ONTOLOGY CASE-BASED REASONING APPROACHES TO PREDICTION OF CARDIOVASCULAR DISEASE

A thesis submitted to Auckland University of Technology (AUT) in fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Computer and Information Sciences

SON MINH HUYNH (STEVE)

2020

School of Engineering, Computer and Mathematical Sciences

ATTESTATION OF AUTHORSHIP

“I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of another university or institution of higher learning, except where due acknowledgements are made.”

Signed

Son Minh Huynh

ACKNOWLEDGEMENTS

There are a number of people that I would like to sincerely express my great appreciation to. Their support during my PhD research journey was necessary and invaluable to help me complete this thesis. First, my two supervisors, Professor Dave Parry and Associate Professor Jacqueline Whalley, gave thoughtful guidance, useful critiques, and enthusiastic encouragement that helped me keep moving forward. Second, I greatly thank the School of Computer and Mathematical Sciences for the PhD fees scholarship, prompt responses to my requests such as the ethics approval request, and empathetic support when I needed more time for my young family. I also send my warm gratitude to other lecturers, staff, and students for your useful and enjoyable conversation and advice. Last but not least are my special thanks and love to my wife Helen, my four-year-old daughter Patricia, and my newborn son Karl. You have always been there for me unconditionally. There were beautiful sunny days, and I did not take you to the beach. There were activities at the childcare and social events, and I could not be there with you. Nevertheless, you never complained a word. I shall spend more time for you after the completion of this PhD degree.

ABSTRACT

Cardiovascular disease (CVD) is a major cause of morbidity and mortality. However, current widely used regression models are known to have a number of drawbacks, including prediction inaccuracy for individuals and for other cohorts, inflexibility of handling intervention, requirement of complete clinical data, deficiency of dealing with inaccurate, vague and uncertain data, and poor explanatory capacity.

Therefore, this research developed a novel prediction model named CRISK—short for CVD Risk—for predicting 10-year risk of CVD. The model was developed based on a combination of fuzzy ontology and case-based reasoning (CBR). Fuzzy ontology can help handle and store vague and uncertain data, which is common in real life. Retrieving the closest cases to the input case, CBR could contribute to the development of a personalised prediction model. The CRISK model retrieves the seven closest cases to the input case and generates prediction outcomes from these seven closest cases. To do this, three algorithms, Retrieve, Reuse, and Revise, were developed. The CRISK model uses 13 risk factors: total cholesterol, low-density lipoprotein (LDL) cholesterol, very-low-density lipoprotein (VLDL) cholesterol, systolic blood pressure (SBP), triglycerides, diastolic blood pressure (DBP), glucose, number of cigarettes smoked a day, high-density lipoprotein (HDL) cholesterol, hematocrit, body mass index (BMI), and lactate dehydrogenase (LDH). Moreover, the model introduced a new way to represent and interpret CVD prediction outcomes when compared with existing models. In CRISK, the prediction outcomes are represented as fuzzy membership values of the "High CVD Risk" and "Low CVD Risk" fuzzy sets. Depending on the fuzzy membership value, a different level of attention is given to the input case. Using this method, not only the predicted risk category but also the prediction of when CVD would happen is provided.

The CRISK model achieved reasonably good predictions. For internal validation, the prediction performance results were True Positive Rate (TPR)=0.8733 (CI=0.0102), True Negative Rate (TNR)=0.8270 (CI=0.0116), Precision=0.2247 (CI=0.0128), F_1-value=0.3574 (CI=0.0147), and Negative Prediction Value (NPV)=0.9913 (CI=0.0029) where CI is the 95% confidence interval. These performance results were obtained from experiments using the Framingham Heart Study (FHS) Offspring Cohort Exam 1 dataset, which was the dataset used to develop the CRISK model. For external validation, experiments on the FHS Original Cohort Exam 11 dataset were performed. This dataset

had two missing risk factors: triglycerides and LDH. The prediction results obtained for this external validation were TPR=0.8167 (CI=0.0434), TNR=0.5041 (CI=0.0560), Precision=0.2866 (CI=0.0507), F_1-value=0.4242 (CI=0.0554), and NPV=0.9185 (CI=0.0307) where CI is the 95% confidence interval. In addition, the CRISK model was analysed to be able to solve or partially solve five out of eight limitations of regression models identified in this research. Moreover, CRISK gave a better prediction performance in comparison with two high-profile existing CVD prediction models.

This research has shown the usefulness of fuzzy ontology CBR approaches in CVD prediction. The achievements from the research are promising. Therefore, it would be worth investing more into fuzzy ontology CBR approaches in building CVD prediction models specifically and in building chronic disease prediction models generally. However, it would not be that a prediction model is built once and used forever. It is rather to continuously perform experimentation and update the model when new datasets arrive, especially datasets from different ethnic groups. These would help keep improving the prediction performance for the model and keep the model up to date.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

AI	Artificial Intelligence
AUC	Area Under Curve
BMI	Body Mass Index
BP	Blood Pressure
CART	Classification and Regression Tree, a popular decision tree machine learning algorithm
CBR	Case-Based Reasoning
CHD	Coronary Heart Disease
CRISK	CVD Risk, the CVD risk prediction model resulting from this research
CSV	Comma-Separated Values
CVD	Cardiovascular Disease
CPM	Clinical Prediction Model
DBP	Diastolic Blood Pressure
DL	Description Logic
DVT	Deep Vein Thrombosis
Eclipse	An Integrated Development Environment (IDE) for the Java programming language
FHS	Framingham Heart Study
FN	False Negatives
FNR	False Negative Rate

FP	False Positives
GUI	Graphic User Interface
HDL	High-Density Lipoprotein
HDL-C	High-Density Lipoprotein Cholesterol
ID3	Interactive Dichotomiser 3, a popular decision tree machine learning algorithm
IDE	Integrated Development Environment
IRI	Internationalised Resource Identifier
IS	Information System
ISRF	Information System Research Framework
IT2FS	Interval Type-2 Fuzzy Set
KNN	K Nearest Neighbours
LDH	Lactate Dehydrogenase, an enzyme that helps produce energy. LDH is found in almost all body tissues
LDL	Low-Density Lipoprotein
LDL-C	Low-Density Lipoprotein Cholesterol
LOOCV	Leave-One-Out-Cross-Validation
MAE	Mean Absolute Error
MI	Myocardial Infarction
ML	Machine Learning
NPV	Negative Predictive Value
OWL	Web Ontology Language
P/N	Positives / Negatives

PAD	Peripheral Arterial Disease
PCE	Pooled Cohort Equation
PE	Pulmonary Embolism
Protégé	A free, open-source ontology editor
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RHD	Rheumatic Heart Disease
ROC	Receiver Operating Characteristics
RQ	Research Question
SBP	Systolic Blood Pressure
SMOTE	Synthetic Minority Over-sampling Technique
RMSE	Root Mean Squared Error
T1FS	Type-1 Fuzzy Set
T2FS	Type-2 Fuzzy Set
TC/HDL-C	Total Cholesterol / High-Density Lipoprotein Cholesterol
TN	True Negatives
TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate

VBA	Visual Basic for Applications. Excel VBA is Microsoft’s programming language for Excel and other Microsoft Office programs.
VLDL	Very-Low-Density Lipoprotein
VLDL-C	Very-Low-Density Lipoprotein Cholesterol
W3C	World Wide Web Consortium
Weka	A free, open-source data mining software
XML	eXtensible Markup Language

Chapter 1

INTRODUCTION

1.1 RATIONALE AND SIGNIFICANCE OF THE STUDY

Cardiovascular disease (CVD), aka heart disease, is the number one cause of death worldwide [1]. In 2016, an estimated 17.9 million people died from CVD, accounting for 31% of all global deaths in the year [1]. This means that one in three people would pass away as a result of heart disease. In New Zealand, the 2018 statistics from the Heart Foundation website[1] report that 33% of deaths annually are caused by heart disease, one in twenty adults are living with heart disease, and every 90 minutes a New Zealander dies from heart disease. According to European Cardiovascular Disease Statistics 2017 [2], CVD accounted for 45% of all deaths in Europe and 37% of all deaths in the European Union (EU).

CVD not only places immense health burdens but also massive economic burdens [1, 3]. The total cost of CVD is estimated to be around 210 billion EUR on the EU economy annually [2]. Of this total cost, 53% (111 billion EUR) is for healthcare costs, 26% (54 billion EUR) is due to productivity losses, and 21% (45 billion EUR) is to informal care of people with CVD. In the US, it is projected that in 2035, more than 130 million adults in the nation’s population (45.1%) will have some form of heart disease, and the total costs for CVD in the year are expected to reach 1.1 trillion USD [3].

Interestingly, the majority of CVD cases can be controlled by addressing behavioural risk factors [1]. Examples of these risk factors are smoking, unhealthy diet, and physical

[1] https://www.heartfoundation.org.nz/

inactivity [1]. As CVD events typically appear suddenly and often result in fatality before medical care can be given [4], early detection of high CVD risk people would be greatly beneficial in preventing CVD events by allowing early intervention for those at high risk. Therefore, with accurate early CVD prediction, a healthy population could be maintained resulting in markedly reduced health and economic burdens [5].

However, the CVD prediction problem remains unsolved despite the numerous prediction models that have been developed. In a systematic review published in 2016 [6], 363 prediction models prior to June 2013 were identified and reviewed. In terms of prediction techniques, most of the developed models use regression methods: Cox proportional hazards modelling (n=160, 44%), accelerated failure time analysis (n=77, 21%), and logistic regression (n=71, 20%). But, according to the authors, the usefulness of most of these models remains unclear. Besides regression, machine learning (ML) has become more and more popular in the medical community generally and in the domain of CVD prediction specifically [7]. Though ML is reported to improve CVD prediction accuracy when compared to traditional established statistical models, there are still certain limitations related to the developed ML models including poor interpretability[2] and overfitting [8]. Therefore, further investigation of the feasibility and acceptability of ML applications is needed before they can be employed in day-to-day clinical workflows [7, 8].

Recently, new approaches using fuzzy ontologies and case-based reasoning (CBR) for prediction of chronic diseases, such as diabetes [9, 10] and chronic kidney disease [11], have been explored. Fuzzy ontologies have been known to be able to represent the vagueness and uncertainty of data [12]. This representation may possibly help resolve the limitations of regression models that are in widespread use but unable to deal with missing or unreliable values [12]. On the other hand, CBR may be an important approach in health care [13, 14]. A CBR system provides a solution for a new case based on the solutions of similar past cases [15, 16]. Therefore, a combination of fuzzy ontology and CBR approaches for CVD prediction would be worth investigating.

[2] Interpretability refers to the ability to understand the causality, i.e. relationships between risk factors and outcomes

1.2 Research Aim, Objectives, and Research Questions

This study aims to develop a fuzzy ontology CBR model called CRISK (Cardiovascular disease RISK) for prediction of CVD within 10 years that would possibly be a candidate to be used in daily clinical practice. Ten years is chosen because this is the common CVD prediction time interval. To be suitable for daily clinical practice the developed model should be able to resolve limitations of the current in-use models and/or perform better in terms of prediction performance.

To achieve this aim, followed objectives need to be achieved. Firstly, current CVD prediction problems need to be identified and thoroughly analysed. Secondly, the developed model must employ techniques that can address the current CVD prediction problems. Thirdly, experiments and validation are used to find which risk factors are important.

The following six research questions (RQs) are proposed for this study. How these research questions were formulated is explained in section 2.2.6.

RQ1. Can a CVD prediction model be developed using a combination of fuzzy ontology and CBR?

RQ2. What risk factors are important in the prediction of CVD using this method?

RQ3. How does the developed model perform in terms of prediction performance?

RQ4. How does the developed model perform in terms of external validation?

RQ5. How does the developed model overcome the limitations of current widely used regression models?

RQ6. How does the developed model compare with current widely used regression models in terms of prediction performance?

Chapter 4, Chapter 5 and Chapter 6 in this body of work help answer the first three research questions (RQ1, RQ2, and RQ3). The content of Chapter 7 is used to answer RQ4. In Chapter 8, section 8.1 helps answer RQ5 and section 8.3 helps answer RQ6. Descriptions of the answers are given in section 9.1.1 in Chapter 9.

1.3 CONTRIBUTIONS OF THIS STUDY

This study provides several contributions to the existing knowledge. These contributions are summarized below.

1. An extensive literature review on the topic of CVD prediction is provided. The literature review gives the current status of the topic as well as highlighting the drawbacks of the existing mainstream regression models. Details of the literature review on CVD prediction are in section 2.2.

2. The CRISK prediction model (Chapter 4) and its associated CRISK system (Chapter 5) are developed and documented. The developed model achieves good prediction performance (section 6.3) and solves or partially solves five out of the eight problems of current regression models (section 8.1). The CRISK system provides modules for creating ontologies, running experiments, and giving CVD prediction for an individual case (Chapter 5).

3. This research shows that fuzzy ontology CBR approaches are useful in CVD prediction. This should encourage future researchers to spend more effort for fuzzy ontology CBR approaches in CVD prediction specifically and in chronic disease prediction generally.

4. This research contributes a new way to represent and interpret CVD prediction outcomes, using fuzzy membership values of “High CVD Risk” and “Low CVD Risk” defined in this thesis. With this new way, not only the predicted risk category but also the prediction of when CVD would occur is provided. Details can be found in section 3.4.2 and section 8.3.

5. This study proposes the idea of continuous experimentation and updates for a CVD prediction model. This would help keep improving the model’s prediction performance. Details can be found in section 9.2.5.

1.4 THESIS ORGANISATION

The rest of the thesis is organised into the following ten chapters:

Chapter 2 first reviews existing CVD prediction models. The models are classified into different categories and reviewed focusing on prediction methods, risk factors, datasets, prediction performance, and limitations. From the review, problems with existing prediction models, and potential and gaps of fuzzy ontology CBR approaches in CVD prediction are identified and these lead to formation of research questions for this body of work. The chapter then covers related information about fuzzy ontology for this body of work. This includes the theories of type-1 and type-2 fuzzy sets, definition of fuzzy ontology, and the advantages of using it. Moreover, state-of-the-art languages and tools for building and managing fuzzy ontologies are provided. Finally, the chapter introduces CBR, popular techniques used for CBR, and tools for building CBR systems. This includes explaining what CBR is, describing the four activities in a CBR cycle, and providing details of three common techniques used for CBR. In addition, well-known tools for building CBR systems are reported.

Chapter 3 explains how the research was approached and carried out. This includes deciding on a research paradigm, employing a research methodology, forming a research framework and research guidelines, and creating strategies and plans to develop the CRISK prediction model. In addition, it covers dataset collection, dataset selection, experimentation design, and data preparation accordingly to the experimentation design. Besides, it describes an evaluation protocol created to assess the developed CRISK prediction model. The protocol consists of evaluation metrics, external validation, and comparison to existing models.

Chapter 4 describes the CRISK prediction model. It first gives an overview of the architecture of the model. It then explains in detail each component of the model. In addition, the main algorithms developed for the model are described as pseudo-code.

Chapter 5 explains the developed CRISK system. Details include how the system was developed, especially focusing on the structure of the CRISK system. The system consists of four modules: Constructor, Experimenter, Batch Experimenter, and Predictor. The purpose and details of each module are also described.

Chapter 6 describes experimentation and the results. The chapter first explains in detail how the experimentation was done. It then reports the results, focusing on finding whether it is worth creating separate prediction models for men and women, and how the prediction performance could be possibly improved in the future. In addition, other findings derived from the experimentation results are also reported.

Chapter 7 gives details of external validation of this research. The chapter first describes how external datasets was prepared. It then explains how prepared datasets were tested. After that, the chapter reports testing results and findings from the testing results.

Chapter 8 is a discussion chapter. It refers back to the list of problems of current regression models and discusses how these problems have been addressed by the developed CRISK model. It then discusses personalised prediction using the CRISK model. After that, CRISK is compared with three high-profile existing CVD prediction models. Finally, the possibility of applying CRISK in daily practice is raised.

Chapter 9 concludes the thesis. Achievements of the study are described. After that, limitations in this research and future directions for it are provided.

Chapter 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter first reviews the current status of CVD prediction (section 2.2). This includes giving an overview of CVD, reviewing existing well-known prediction models, finding problems with current regression models, and reviewing current fuzzy logic, fuzzy ontology, and CBR approaches in CVD prediction. From this review, research questions (stated in section 1.2) are formulated for this study (section 2.2.6).

The chapter then provides an investigation of fuzzy ontology in section 2.3. This section first gives an overview of type-1 and type-2 fuzzy sets. It then provides an overview of fuzzy ontology including the benefits of using it. Finally, the section gives a summary of the languages and tools commonly used to create and maintain fuzzy ontologies. Information from this section is based on in decision making for building the CRISK model, which helps answer RQ1, in section 3.4. Decisions involve whether to use type-1 or to use type-2 fuzzy sets and which languages and tools should be used to develop fuzzy ontologies for CRISK.

After that, the chapter provides an investigation of CBR in section 2.4. First, this section explains what CBR is and gives an update on research activities and application of CBR in various domains. It then provides details of common techniques used in CBR systems. Finally, the section summarises popular CBR tools for developing CBR applications, focusing on programming language used and whether the tool supports ontologies and fuzzy ontologies. From the investigation presented in this section, decisions on whether to use existing tools or developing an own CBR application for CRISK, which helps answer RQ1, and on what techniques to employ are made (section 3.4).

2.2 CARDIOVASCULAR DISEASE PREDICTION

2.2.1 An Overview of CVD

CVD is a group of diseases pertaining to the heart, the vascular system of the brain, or blood vessels [17]. It is estimated that 17.9 million people died from CVD in 2016, representing 31% of deaths globally [1]. CVD includes coronary heart disease (CHD), cerebrovascular disease, peripheral arterial disease (PAD), rheumatic heart disease (RHD), congenital heart disease, and deep vein thrombosis and pulmonary embolism (DVT & PE) [18]. Among these six types of CVD, four are caused by arteriosclerosis of the blood vessels, that is hardening of the arteries. Arteriosclerosis of the blood vessels providing blood to the heart muscle causes CHD [19]. Arteriosclerosis in the blood vessels providing blood to the brain causes cerebrovascular disease [20]. When arteriosclerosis happens in the blood vessels providing blood to the arms and legs, it results in PAD [21]. If arteriosclerosis happens at a vein deep in the body, usually in the lower legs or thighs, and creates a blood clot, the resulting condition is known as DVT [22]. The blood clot restricts or can completely block the blood flow. The blood clot can sometimes dislodge, travel to the heart and then lungs, forming blockages in arteries supplying blood to the lungs. This condition is called a PE [22]. On the other hand, RHD is a disease in which the heart muscle and heart valves are damaged from rheumatic fever, which is caused by streptococcal bacteria [23]. Finally, congenital heart disease refers to malformations in the cardiovascular structure that occur before birth [24]. In these six types of CVD, coronary heart disease (CHD), also known as ischemic heart disease [19], is the most common type of CVD [25]. Globally, CHD accounts for 7.4 million deaths, which is about 43% of deaths caused by CVD, in 2012 [18].

Prediction plays a significant role in reducing disability and premature death caused by CVD. The underlying pathology is atherosclerosis, which develops over years and is usually advanced by the time symptoms occur. Acute coronary and cerebrovascular events typically appear suddenly and can be fatal before medical care can be given [4]. Therefore, CVD prediction techniques have been extensively researched and developed for decades, aiming to provide early intervention for those at risk.

Existing well-known CVD prediction models are identified and classified into the following categories: conventional Framingham models, augmented Framingham models, and alternatives to Framingham models [26]. Sections 2.2.2, 2.2.3, and 2.2.4 review those prediction models.

2.2.2 Conventional Framingham Models

Conventional Framingham models [27-32] were developed as part of the Framingham Heart Study (FHS). This study started in 1948 and had the aim of observing, and as a consequence, understanding risk factors that cause heart disease [33]. Originally, the Framingham Study followed a cohort of 5,209 men and women living in Framingham, Massachusetts, in the United States of America (USA) [27]. The study continuously monitored morbidity and mortality, and medical examinations were carried out every two years to record a variety of characteristics, including blood chemistry, blood pressure, and electrocardiogram [27]. In 1971, the study enrolled a second generation of 5,124 participants, who were children of the first cohort and those children's spouses, called the Framingham Offspring Cohort [28, 34]. As these two cohorts were predominantly white of European descent, 506 ethnic minority residents of Framingham were recruited for the Omni 1 Cohort in 1994 followed by another 410 ethnic minority participants for the Omni 2 Cohort a decade later [33]. In 2002, investigators created the Third Generation Cohort with 4,095 participants, who were children of the Offspring Cohort. One year later, 103 residents, who were spouses of Offspring Cohort participants who were not initially enrolled in the study but had at least two children in the Third Generation Cohort, were signed up for the New Offspring Spouse Cohort [33].

The first Framingham model [27] was developed by Kannel and colleagues in 1976. The study used data of people from the Original Cohort who were initially free of CHD, congestive heart failure, cerebrovascular disease, intermittent claudication, and rheumatic heart disease. It produced risk functions to predict CHD, brain infarction, intermittent claudication, hypertensive heart failure, and total CVD within eight years. Each risk function was a logistic regression model where dependent variables (risk factors) were sex, age, systolic blood pressure (SBP), cigarette smoking, electrocardiographic evidence of left ventricular hypertrophy (ECG-LVH), glucose intolerance, and serum cholesterol. Total CVD prediction includes disease classified to either CHD, brain infarction (cerebrovascular disease), intermittent claudication (PAD), or hypertensive heart failure (congestive heart failure in the absence of coronary or rheumatic heart disease). For each risk function, two different sets of regression coefficients were developed for men and women respectively. Regression coefficients were calculated using the method of Walker-Duncan [35]. Figure 2-1 illustrates the relationships among the prediction model, predicted diseases, risk factors used, and cohorts that the study based on.

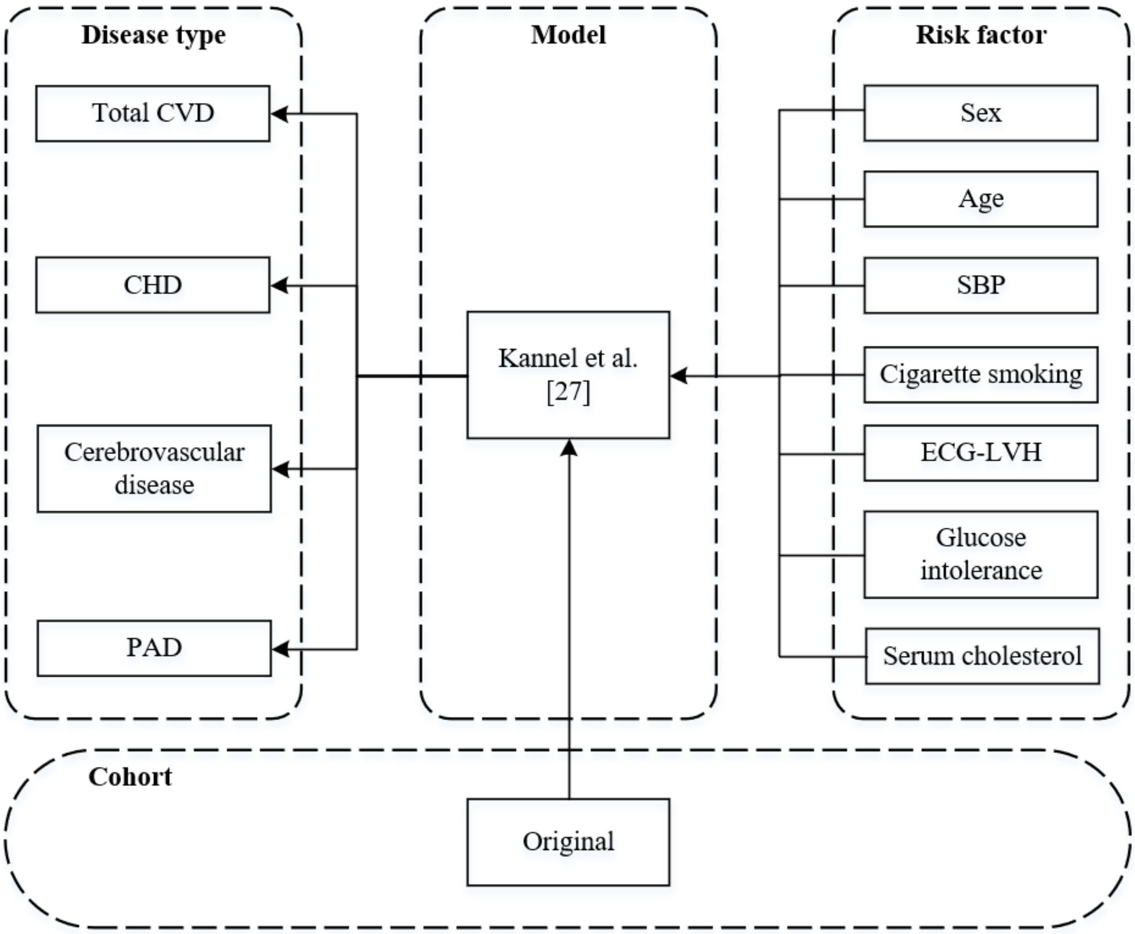


Figure 2-1: Mappings of Kannel et al. [27] model, disease types, risk factors, and cohorts

In 1991, two studies were published by Anderson and colleagues. The first one was an updated CHD risk profile [28] of Kannel's original CHD risk profile [27]. Though only people who were free from CVD were included in the first study [28], the dataset used was larger and more recent as it combined both the Original Cohort and the Offspring Cohort. In addition, another risk factor, high-density lipoprotein (HDL) cholesterol, was added to the regression function for prediction. Risk estimation was done using a parametric regression model, an accelerated failure time regression model [36], where parameters were calculated using a computer software program that implemented the maximum likelihood method and was developed by one of the authors.

A further study [29], based on the updated CHD risk profile [28], aimed to develop equations for predicting additional outcomes. This study also selected members from both the Original and the Offspring cohorts who were free of CVD and cancer, and presented prediction equations for myocardial infarction, CHD, death from CHD, stroke, CVD, and death from CVD. Risk factors used in this study were sex, age, blood pressure, total cholesterol, HDL cholesterol, smoking, glucose intolerance, and left ventricular hypertrophy. Again, the accelerated failure time regression model [36] was used to predict probabilities for each of the outcomes. The parameters were also estimated by using the maximum likelihood method. Figure 2-2 shows mappings among the two models, predicted diseases, risk factors used, and cohorts that the studies are based on.

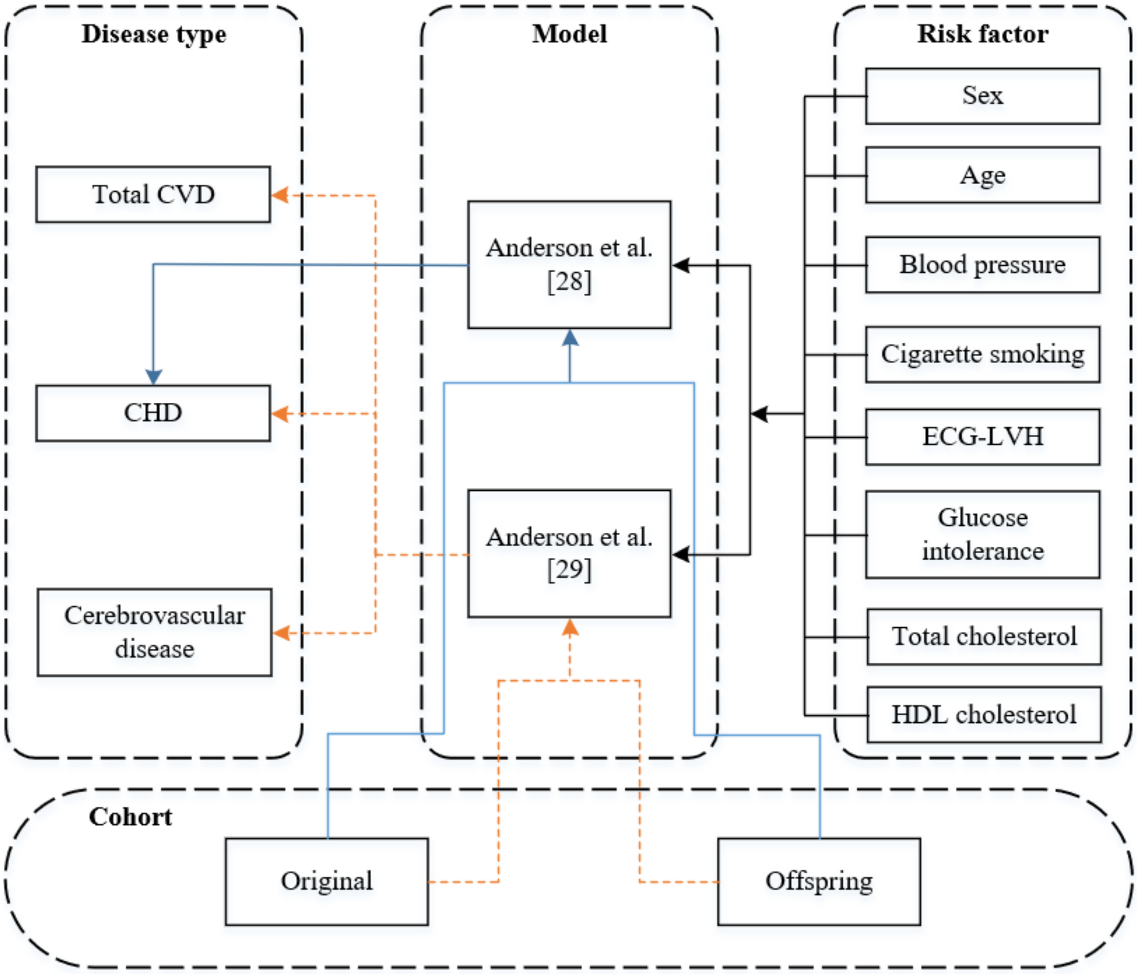


Figure 2-2: Mappings of Anderson et al.'s [28] & [29] models, disease types, risk factors, and cohorts

One advantage of the models developed by Anderson and colleagues [28, 29] over the original model developed by Kannel and colleagues [27] is that the user can specify the number of years (from 4 to 12) ahead that they wish to predict within. This capability to specify prediction interval is a result of using the accelerated failure time regression model [36] rather than a logistic regression model. Unlike logistic regression, this accelerated failure time regression model can provide predictions for different lengths of time [29].

In 1998, Wilson and colleagues developed another Framingham risk profile for CHD [30] using categorical variables that had become part the framework of the Joint National Committee (JNC-V) blood pressure and the National Cholesterol Education Program (NCEP) cholesterol programs in USA. Wilson's study also used data from the Original and Offspring cohorts and produced recommended guidelines to predict CHD risk based on sex, age, blood pressure, diabetes, total cholesterol, and low-density lipoprotein (LDL) cholesterol. The model provides a similar result to Anderson et al.'s first model [28] that used continuous variables. However, Wilson and colleagues' prediction formulation [28] is much simpler than the one used by Anderson et al. [26]. Wilson et al. presented the prediction formulation as score sheets with steps to follow to calculate CHD risk points. Their score sheets were developed from those in the Cox proportional hazards modeling [37]. Wilson et al.'s prediction algorithm was adopted and used by The National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) (NCEP/ATP III) to estimate a person's 10-year risk of developing CHD, in which three levels of risk were defined: less than 10%, 10% to 20%, and greater than 20% [38]. Figure 2-3 displays mappings among Wilson et al.'s model, predicted diseases, risk factors used, and cohorts that the study was based on.

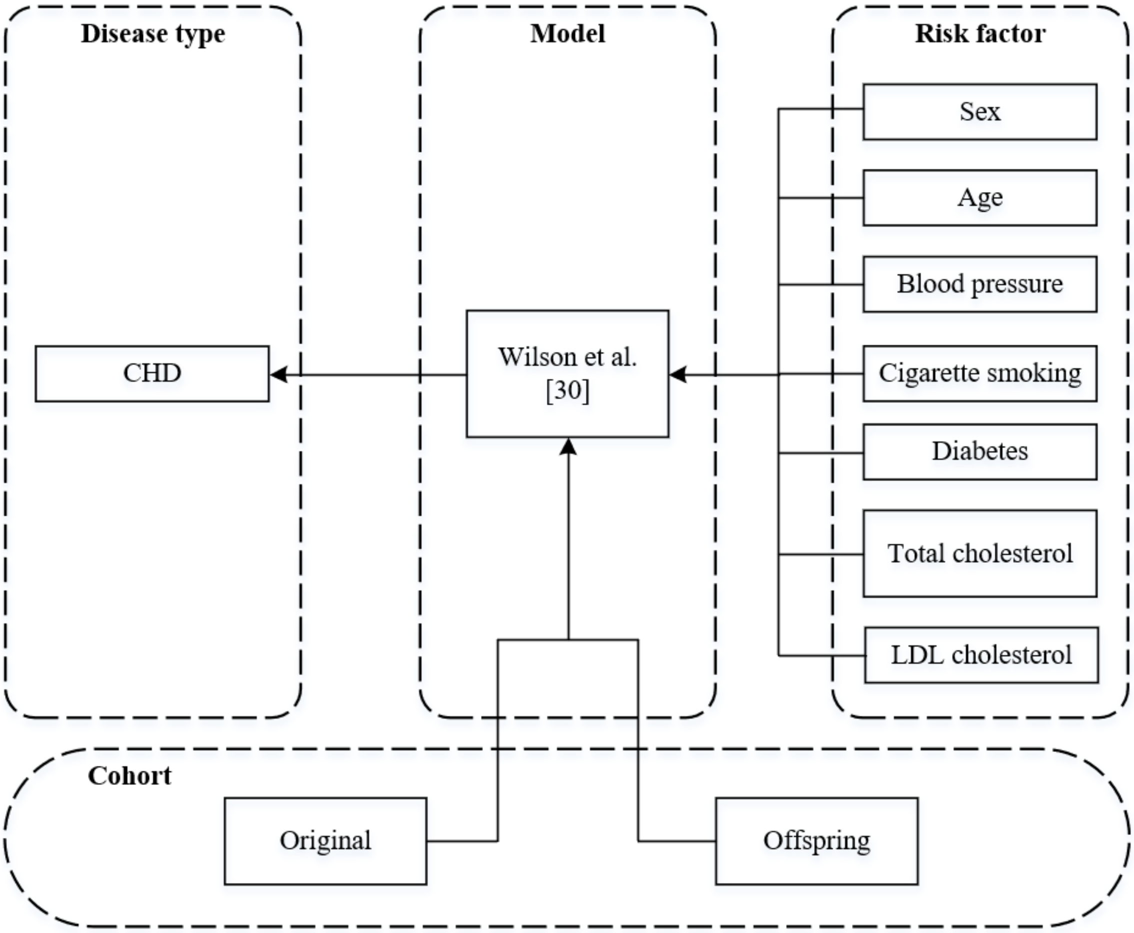


Figure 2-3: Mappings of Wilson et al. [30] model, disease types, risk factors, and cohorts

Later in 2008, D'Agostino and co-authors reported on the development of a gender-specific multivariable risk factor algorithm [31] that can be used to predict total CVD risk and risk of individual CVD events (coronary, cerebrovascular, peripheral arterial disease, and heart failure). The study used more participants (8,491) from the Original and Offspring cohorts than previous studies. The Cox proportional hazard regression method [37] was used to relate risk factors to the incidence of a first CVD event. The authors ended up providing two versions of CVD risk prediction. The first version was based on eight traditional risk factors: sex, age, SBP, treatment for hypertension, cigarette smoking, diabetes, total cholesterol, and HDL cholesterol. In contrast, the second version included non-laboratory-based predictors—body mass index (BMI) was used instead of total cholesterol and HDL cholesterol. Figure 2-4 connects the two model versions from the study with predicted diseases, risk factors, and cohorts.

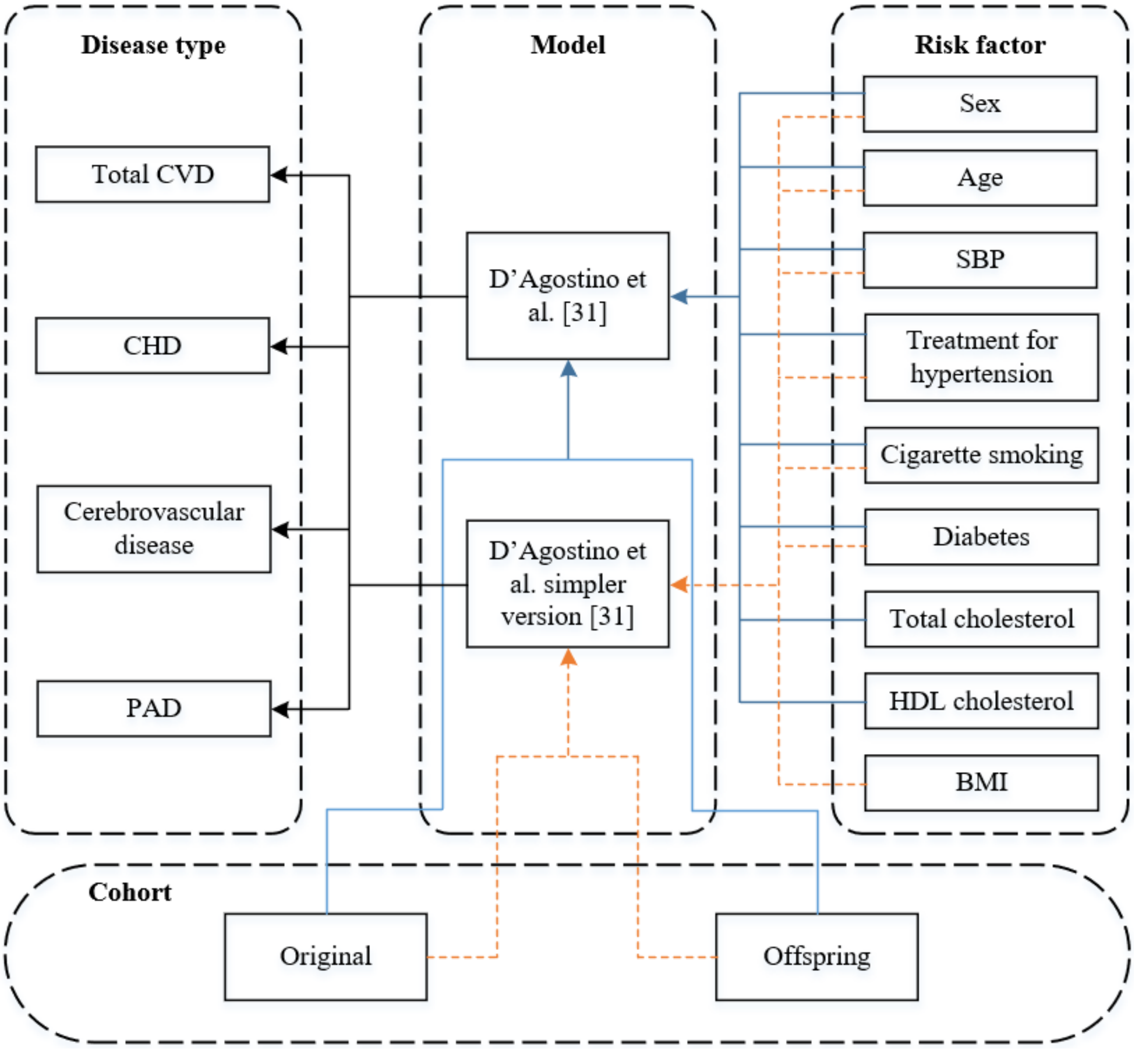


Figure 2-4: Mappings of D'Agostino et al. [31] models, disease types, risk factors, and cohorts

In 2009, Pencina et al. [32] published a model to primarily estimate the 30 year risk of "hard" CVD (coronary death, myocardial infarction, stroke) and secondarily estimate the 30 year risk of "general" ("full" or "total") CVD (coronary death, myocardial infarction, coronary insufficiency, angina, ischemic stroke, hemorrhagic stroke, transient ischemic attack, peripheral artery disease, heart failure). Since the majority of existing models predict risk within the ≤ 10-year risk time frame, this model is a good addition as it provides longer-term estimation and therefore could be more suitable for prediction for people at younger ages and also possibly increase life expectancy. The authors developed the model using a modified Cox regression that allows adjustment for competing risks of non-cardiovascular death. The prediction model has two versions. The first version is based on sex, age, SBP, diabetes, smoker, treated hypertension, total cholesterol, and HDL cholesterol risk factors while the second version replaced total cholesterol and HDL cholesterol with BMI. The simpler version (the second version) performed reasonably well, not far below the performance of the first version [32]. Figure 2-5 connects the two versions of Pencina et al.'s model with predicted diseases, risk factors, and cohorts.

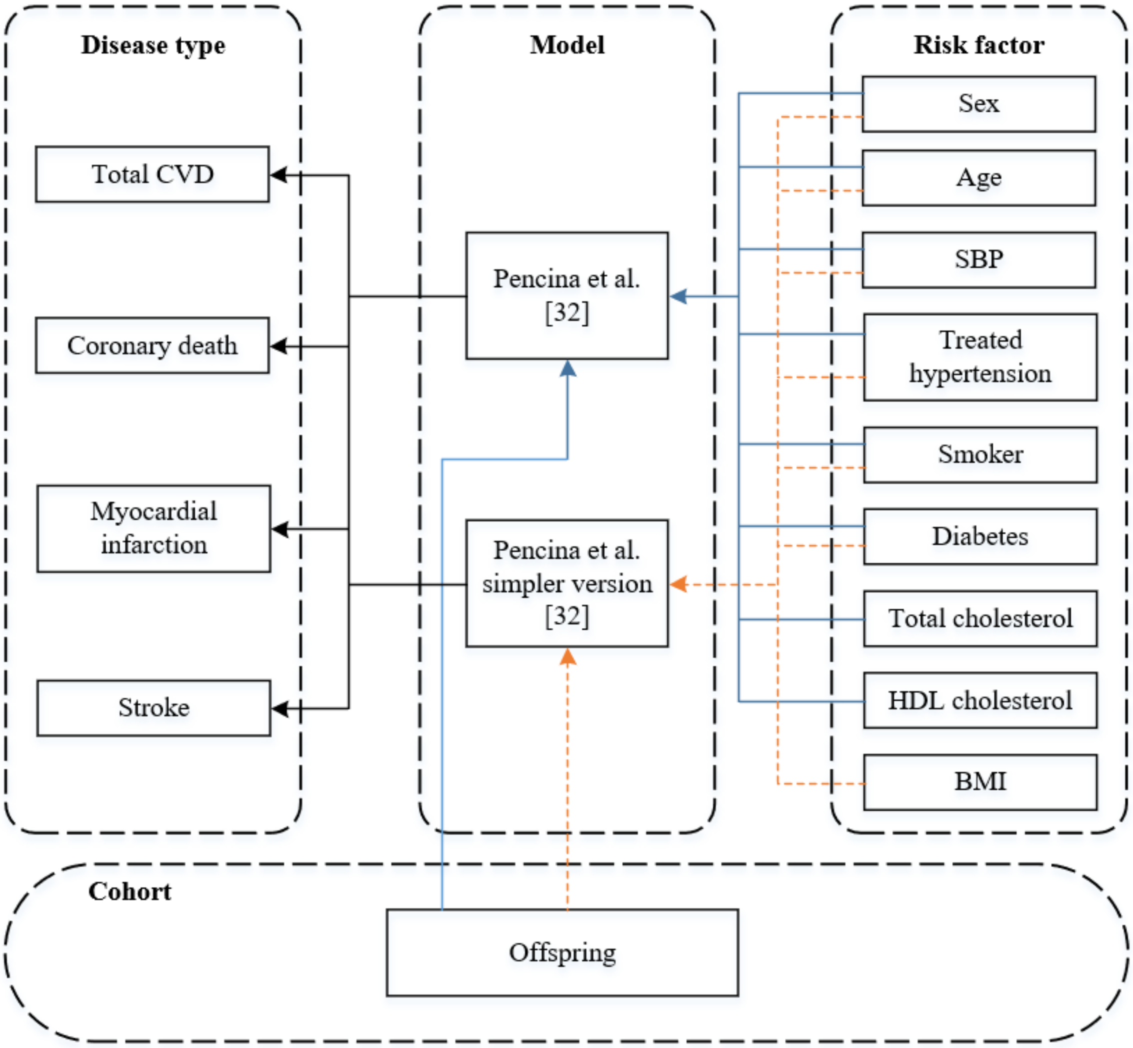


Figure 2-5: Mappings of Pencina et al. [32] models, disease types, risk factors, and cohorts

Table 2-1 summarizes conventional Framingham CVD prediction models in terms of method, number of participants, age, and prediction interval.

Table 2-1: Conventional Framingham CVD prediction models

Model	Method	No. of participants	Age (years)	Prediction Interval
Kannel et al., 1976 [27]	Logistic regression [35]	5,209	35–74	8 years
Anderson et al., 1991 [28]	Accelerated failure time regression [36]	5,573 (2,590 men and 2,983 women)	30–74	4–12 years
Anderson et al., 1991 [29]	Accerlerated failure time regression [36]	5,573 (2,590 men and 2,983 women)	30–74	4–12 years
Wilson et al., 1998 [30]	Categorical variable score sheet, Cox proportional hazards modeling [37]	5,345 (2,489 men and 2,856 women)	30–74	10 years
D'Agostino et al., 2008 [31]	Cox proportional hazards modeling [37]	8,491 (3,969 men and 4,522 women)	30–74	10 years
D'Agostino et al. simpler version, 2008 [31]	Cox proportional hazards modeling [37]	8,491 (3,969 men and 4,522 women)	30–74	10 years
Pencina et al., 2009 [32]	A modified Cox model	4,506 (2,173 men and 2,333 women)	20–59	30 years

The FHS currently recommends models for CVD prediction on their official website.[3] For 10-year CHD risk prediction the Wilson et al. [30] model is recommended. For 10-year general CVD risk prediction, the D'Agostino et al. [31] model and its simpler version should be used while for 30 year general CVD risk prediction, the Pencina et al. [32] model and its simpler version are suggested.

A common limitation of all six Framingham studies [27-32] was that they were solely restricted to white cohorts (Original and/or Offspring) to develop their models. This potentially limits the generalizability to other ethnic groups [32]. Therefore, application of the Framingham models in other populations needed to be verified. Consequently, a

[3] https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/

number of studies, reported in the literature, were carried out to test the Framingham risk functions in different geographical areas and with different ethnic cohorts.

Despite being pioneers in the field of CVD prediction, the most well-known, and the most commonly used both in USA and globally [33, 39], the Framingham models have been shown to overestimate or underestimate risk when applied to populations other than the original cohorts.

The Anderson et al. [28] model was found to overestimate CHD risk (4%) for men in the French PCV-METRA cohort when compared with the risk estimated by the localised French model (2%) [40]. The Anderson et al. [29] model also substantially overestimated CHD risk in German MONICA Augsburg and PROCAM cohorts for both genders [41]. It was reported that the risk predicted by Anderson et al.'s model was double the risk observed in these two cohorts [41]. The Wilson et al. [30] model was found to overestimate CHD risk in an Italian population [42]. Anderson et al.'s [28] and [29] models were also confirmed to significantly overestimate CHD risk for people in the United Kingdom [43]. On the other hand, the Wilson et al. [30] model was reported to underestimate CHD risk in Czech men [44]. These findings indicate that while the Framingham models may be accurate when applied to Framingham cohorts they are probably not as accurate when applied to other cohorts or other populations in the World.

2.2.3 Augmented Framingham Models

To overcome the deficiencies of Framingham equations, researchers have tried to add new variables to them. Usually, one or more biomarkers are included as additional risk factors to the equation [26]. One popular biomarker is the C-reactive protein (hsCRP), a plasma protein synthesised by the liver in response to inflammation [45]. A systematic review by Danesh et al. [46] stated that C-reactive protein is a relatively moderate predictor of CHD; however recommendations for using it in predicting CHD need to be reviewed. Just two years later, Lloyd-Jones et al. [47] found no proof that the including C-reactive protein adds substantial predictive value to CHD prediction over employing the conventional risk factors. Lloyd-Jones's findings are further supported in a later review by McNeill et al. [45] who concluded that the C-reactive protein is unimportant in CHD prediction.

Other biomarkers that have been investigated by researchers include: fibrinogen [48], homocysteine [49], N-terminal fragment brain natriuretic peptide (NT-pro-BNP) [50, 51], small dense lipoproteins [52], apolipoproteins [53, 54], lipoprotein-associated phospholipase A2 [55], lipoprotein (a) [56], cystatin C [57], uric acid [58-60], alanine aminotransferase [61], and gamma-glutamyltransferase [62, 63]. However, in a systematic review by Dent [45], it was suggested that these biomarkers' performances in CVD prediction are inconsistent from study to study and they do not really add value to the prediction.

Another trend in amendment to the Framingham equations was to create different presentations based on the Framingham equations. Examples include the New Zealand risk tables [64], the Joint European Societies' charts [65], and the second Joint British Societies' recommendations [66]. The New Zealand risk tables [64] were based on the Framingham model of Anderson et al. [29] to estimate 5-year CVD risk. A cell in a table is identified for a person based on risk factors. Each cell has a colour that represents the risk level. Similarly, the joint European Societies' charts [65], based on the Framingham model of Anderson et al. [28], also divide the charts into different coloured cells mapping to different CHD risk levels within 10 years. The second Joint British Societies' recommendations [66] were also based on the Framingham model of Anderson et al. [28] but replaced CHD risk with CVD risk to predict the 10-year CVD risk for a person. The CVD risk prediction algorithm was represented as charts, where a chart's area was divided into different coloured contours representing different risk levels. Nevertheless, these table and chart-based approaches gave visual and easier to understand representations of the prediction algorithms but did not improve prediction accuracy nor resolved the problems mentioned in the previous section 2.2.2 of the conventional Framingham models.

2.2.4 Alternatives to Framingham Models

Besides the FHS, researchers around the World have also built up different study cohorts to identify the risk factors associated with CVD and to develop prediction models. Well-known CVD prediction models derived from these studies include the PROCAM model [67], the SCORE model [68], the ASSIGN model [69], the two Reynolds models [70] and [71], the two QRISK models [72] and [73], the 2013 Pooled Cohort Equation (PCE) model [74], the Globorisk model [75], the PREDICT-1° model [76], and the 2018 PCE model [77]. Table 2-2 gives a summary of these models by publication year, prediction disease, dataset, risk factors, method, and prediction interval.

These eleven models can be grouped into three categories: using a pool of cohorts, including additional biomarkers, and including ethnicity/family/social-economic factors. SCORE, 2013 PCE, Globorisk, and 2018 PCE belong to the first category as their datasets are collections of different cohorts. ASSIGN (used family history and social deprivation), QRISK 1 (used family history and area measure of deprivation), QRISK 2 (used ethnicity, family history, and deprivation score), PREDICT-1° (used ethnicity, family history, socioeconomic deprivation), and 2018 PCE (used ethnicity) belong to the third category. PROCAM and the two Reynolds models employed both "additional biomarkers" and "family/social-economic factors" and therefore can be classified as belonging to both the second and the third categories. PROCAM used triglycerides and family history of premature myocardial infarction (MI). The Reynolds model for women used several additional biomarkers ($HbA_{1c,}$ Lp(a), apolipoprotein B-100, hsCRP, and apolipoprotein A-I) and parental history of MI before age 60 years. The Reynolds model for men used hsCRP and parental history of MI before age 60 years.

All eleven models have the common characteristic of using regression prediction methods. Among these models, SCORE was developed using the Weibull proportional hazards model [78] while the rest were developed using the Cox proportional hazards model [37]. Regression prediction models have been found in both existing literature and in this study to have common limitations that are explained in detail in the next section (section 2.2.5).

Table 2-2: Alternatives to Framingham prediction models

Model	Publication Year	Disease	Dataset	Risk Factors	Method	Prediction Interval
PROCAM (Assmann et al. [67])	2002	CHD	• Name: PROCAM • Location: Germany • Size: 5,389 men • Age: 35–65	Age, low-density lipoprotein cholesterol (LDL-C), smoking, high-density lipoprotein cholesterol (HDL-C), SBP, family history of premature myocardial infarction (MI), diabetes, triglycerides	Cox proportional hazards modelling [37]	10 years
SCORE (Conroy et al. [68])	2003	CVD	• Name: a pool from 12 cohorts • Location: Europe • Size: 205,178 people (88,080 women) • Age: 19–80	Sex, age, smoking, SBP, either total cholesterol (TC) or total cholesterol / high-density lipoprotein cholesterol (TC/HDL-C)	Weibull proportional hazards modelling [78]	10 years
ASSIGN (Woodward et al. [69])	2007	CVD	• Name: ASSIGN • Location: Scotland • Size: 13,297 people (6,540 men) • Age: 30–74	Sex, age, social deprivation, family history, diabetes, smoking, SBP, total cholesterol, HDL-C	Cox proportional hazards modelling [37]	10 years

Reynolds risk score for women (Ridker et al. [70])	2007	CVD	• Name: Renolds (women) • Location: US • Size: 24,558 women • Age: ≥ 45	Age, HbA$_{1c}$ (% with diabetes), SBP, smoking, Lp(a), apolipoprotein B-100, hsCRP (C-reactive protein), apolipoprotein A-I, parental history of MI before age 60 years	Cox proportional hazards modelling [37]	10 years
Reynolds risk score for men (Ridker et al. [71])	2008	CVD	• Name: Renolds (men) • Location: US • Size: 10,724 men • Age: 50–79	Age, SBP, smoking, total cholesterol, HDL-C, hsCRP, parental history of MI before age 60 years	Cox proportional hazards modelling [37]	10 years
QRISK 1 (Hippisley-Cox et al. [72])	2007	CVD	• Name: QRISK 1 • Location: UK • Size: 1.28 million people (Of these, 50.4% were women) • Age: 35–74	Sex, age, smoking, SBP, TC/HDL-C, BMI, family history of CHD in first degree relative aged less than 60, area measure of deprivation, existing treatment with antihypertensive agent	Cox proportional hazards modelling [37]	10 years
QRISK 2 (Hippisley-Cox et al. [73])	2008	CVD	• Name: QRISK 2 • Location: England and Wales • Size: 1,535,583 people (773,291 women) • Age: 35–74	Ethnicity, sex, age, smoking, SBP, TC/HDL-C, BMI, family history of CHD in first degree relative aged less than 60, deprivation score, treated hypertension, type-2 diabetes, renal disease, atrial fibrillation, rheumatoid arthritis	Cox proportional hazards modelling [37]	10 years

2013 PCE (Goff et al. [74])	2013	CVD	• Name: pooled cohorts (including ARIC [79], Cardiovascular Health Study [80], CARDIA [81], Framingham Original [82], and Framingham Offspring [83] cohorts) • Location: USA • Size: 24,626 people (11,240 white women, 9,098 white men, 2,641 African-American women, and 1,647 African-American men) • Age: 40–79	Age, sex, total cholesterol, HDL-C, SBP, use of antihypertensive therapy, diabetes, smoking	Cox proportional hazard modelling [37]	10 years and lifetime
Globorisk (Hajifathalian et al. [75])	2015	CVD	• Name: a pool of 8 cohorts • Location: USA • Size: 50,129 people (33,323 men) • Age: ≥ 40	Sex, age, SBP, total cholesterol, diabetes, smoking	Cox proportional hazards modelling [37]	10 years
PREDICT-1° (Pylypchuk et al. [76])	2018	CVD	• Name: PREDICT • Location: New Zealand • Size: 401,752 • Age: 30–74	Age, ethnicity, NZ index of socioeconomic deprivation, family history of premature CVD, smoking, diabetes, history of atrial fibrillation, SBP, TC/HDL-C, blood pressure lowering medication, lipid lowering medication, antithrombotic medication	Cox proportional hazard modelling [37]	5 years

2018 PCE (Yadlowsky et al. [77])	2018	CVD	• Name: pooled cohorts from 6 longitudinal cohort studies, ARIC (Atherosclerosis Risk in Communities Study, 1987 to 2011), CHS (Cardiovascular Health Study, 1989 to 1999), CARDIA (Coronary Artery Risk Development in Young Adults Study, 1983 to 2006), FHS offspring cohort (1971 to 2014), JHS (Jackson Heart Study, 2000 to 2012), and MESA (Multi-Ethnic Study of Atherosclerosis, 2000 to 2012 • Location: USA • Size: 26,689 people • Age: 40–79	Age, sex, race, total cholesterol, HDL-C, SBP, treatment for high blood pressure, diabetes, smoking	Cox proportional hazard modelling [37]	10 years and lifetime

2.2.5 Problems with current regression prediction models

Despite trying to improve the Framingham models by introducing additional biomarkers to the equations or carrying out research on different cohorts, existing regression prediction models suffer from common limitations. These limitations can be attributed to several fundamental issues associated with using traditional regression techniques to build a disease prediction model. Firstly, a statistical regression technique tries to find a mathematical function that best fits the data. Thus, it uses the same function with the same number of fixed independent variables and the same coefficients for all cases. However different cases might need to have different coefficient values to have better predictions of the outcomes. For example, smoking might have a greater impact on the CVD risk in one group but little impact on another group. Secondly, these statistical methods are limited to using a small number of predictors [84] and therefore might miss other factors that are important to the outcomes. Thirdly, the relationships between covariates and risk may be too complex to be presented by a regression function [85]. Another issue is that, over time, new cases arrive, and the original regression function might be no longer suitable as changes in society occur, such as migration, behaviour changes, environmental changes and different models of health care delivery. Table 2-3 lists and explains the common limitations of regression models.

Table 2-3: Limitations with current regression models

Limitation #	**Name**	**Explanation**
Limitation 1	Inaccuracy for individual	A model can be accurate for the population but inaccurate for an individual [86].
Limitation 2	Inaccuracy for other cohorts	A model can perform well for a certain cohort but turns out to overestimate or underestimate for other cohorts, escpecially cohorts of different racial groups [87].
Limitation 3	Inflexibility of handling intervention	How will the prediction result change if the person stops smoking, starts having treatment, etc.? [12]
Limitation 4	Requirement of complete clinical data	To build a regression model, a complete dataset is required while in reality there are often missing data [12].
Limitation 5	Deficiency of handling inaccurate data or result	Clinical recorded data might be inaccurate [12]. With models where prediction results are crisp values, a small error in prediction might completely shift the person to a wrong category, such as from “high risk” to “low risk”
Limitation 6	Deficiency of handling vagueness of data or result	For example, when a person says that they smoke “a lot of cigarettes” a day. It is unknown exactly how many cigarettes they actually smoke a day.
Limitation 7	Deficiency of handling uncertainty of data or result	For example, if the prediction result for a person is 85% chance of belonging to the high risk group, does it mean the chance is exactly 85% or somehere between 80% and 90%?
Limitation 8	Poor explanatory capacity	Regression methods are built with complex equations that are not easy to vizualise or to understand how they are formed.

2.2.6 Current fuzzy logic, fuzzy ontology, and CBR approaches

To overcome the problems of regression prediction models, a few studies have tried to use fuzzy logic and fuzzy ontology approaches to CVD prediction. In 2012, Pal and co-authors [88] described developing an expert system for screening that would help detect CHD at an early stage. The paper focused on rules formulation from doctors and a fuzzy expert system approach was used to cope with uncertainty present in the medical domain. In 2013, Parry and MacRae [12] introduced an approach that used a fuzzified ontology to both improve CVD prediction accuracy and provide personalised predictive capacity. In 2014, Kim et al. [89] proposed a model named Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model that gave content recommendation to CHD patients. The model consists of three parts: a fuzzy membership function, a rule set, and a fuzzy inference. In 2015, Kim et al. [90] used a hybrid approach combining both fuzzy logic and CART decision tree to build their model for prediction of CHD within 10 years.

CBR has been suggested as an important niche for disease prediction [13, 14]. A CBR system solves a new problem from the solutions of existing similar past cases [15, 16]. There has been some success in healthcare using CBR [15, 91-93]. However, according to a 2011 survey by Begum et al. [94], most of these systems were still at the prototype stage and not available in the market as commercial products. Later, in 2016, in another survey by Choudhury and Begum [95], the number of CBR systems in the healthcare domain was found to have increased significantly. Nevertheless, most of the systems do not include an adaptation step and leave the adaptation task to human experts. Details of CBR are explained in section 2.4.

CBR has not been widely used in CVD prediction. In a 2014 review by Sutano et al. [96] only one model, which was developed by Guessoum et al. [97] for the diagnosis of chronic obstructive pulmonary disease, was related to CVD. The survey by Choudhury and Begum [95] in 2016 found two more models, by Koton [98] and Reategui et al. [99], for the diagnosis of heart disease using CBR. More recently, Kalavai [100] proposed a heart disease prediction model that utilised CBR in an image similarity search. However, these models were all for diagnosing heart disease rather than predicting it in a future time interval. To my knowledge, there have not been any models reported that combine fuzzy ontology and CBR as a model for the prediction of CVD.

Fuzzy ontology CBR systems have been used in other domains. In the domain of collision avoidance systems in marine environments, in 2007, Park et al. [101] reported on an ontology-based fuzzy CBR support system for ship collision avoidance. Their system operates in two steps. The first step identifies any dangerous ships and indexes those new cases. The second step retrieves similar cases from the ontology and produces the solution (the new heading) to take to avoid collision. Recently, Ali et al. [102] proposed a type-2 fuzzy ontology to provide accurate information about collision risk and the marine environment during real-time marine operations. The type-2 fuzzy ontology-based approach was proposed as the existing type-1 fuzzy ontology-based approach was not capable of extracting sufficient information to offer solutions due to the intensively blurred image data that results from the hazy marine environment.

In the domain of depression diagnosis, in 2012, Ekong et al. [103] presented a neuro-fuzzy CBR model as a decision support system for the diagnosis of depression based on the overall severity of symptoms. Neuro-fuzzy inference systems provide self-learning intelligent systems that are capable of handling uncertainties in a diagnosis process [104].

In Education, in 2013, Inyang et al. [105] developed a fuzzy clustering technique based on the Fuzzy c-Means (FCM) algorithm to identify at-risk students at an early stage in their academic career. FCM is a method of clustering which allows one piece of data to belong to two or more clusters. Later, in 2015, Vo et al. [106] introduced an algorithmic framework for incomplete educational data clustering using a nearest prototype strategy. Their framework was found to be able to perform data clustering on datasets with large numbers of missing values.

In 2015, in diabetes diagnosis, El-Sappagh et al. [9] proposed a fuzzy-ontology-oriented case-base reasoning framework for semantic diabetes diagnosis. They compared their framework with existing traditional CBR systems and a set of five machine-learning classifiers. The authors claimed that their system outperformed all those systems. However, several limitations are found from their study. First, their dataset consisted of only 60 real diabetes cases. This is quite a small population. To have more confidence in their approach, a larger cohort is needed. Second, the system has not been validated against other diverse population datasets. Third, in this system (relied in the existing jColibri2 CBR framework [107]), risk factor values were recorded as instances while they should actually be represented as literals in the ontology. However, despite the limitations of El Sappagh et al.'s work, their study has shown the potential of fuzzy ontology CBR approaches in the medical diagnosis domain.

As there has not been a fuzzy ontology CBR model for the prediction of CVD yet while fuzzy ontology and CBR have been used and have shown usefulness in other domains, it is interested to know if a CVD prediction model can be developed using a combination of fuzzy ontology and CBR, what risk factors this developed model uses, and the model prediction performance. Hence, the first three research questions (RQ1, RQ2, and RQ3) below are formed.

RQ1. Can a CVD prediction model be developed using a combination of fuzzy ontology and CBR?

RQ2. What risk factors are important in the prediction of CVD using this method?

RQ3. How does the developed model perform in terms of prediction performance?

Besides, the developed model should also be tested using external datasets and this step is called external validation [108]. Otherwise, the model may not be trusted to be used in daily clinical practice [109]. Therefore, the RQ4 below is formulated in this research.

RQ4. How does the developed model perform in terms of external validation?

In addition, it is also interested to know how the developed model solves the limitations of current widely used regression models and how it compares with those existing models on prediction performance. As a result, the further two research questions (RQ5 and RQ6) below are created for this research.

RQ5. How does the developed model overcome the limitations of current widely used regression models?

RQ6. How does the developed model compare with current widely used regression models in terms of prediction performance?

2.3 FUZZY ONTOLOGY

2.3.1 Type-1 Fuzzy Sets

Type-1 fuzzy sets were introduced by Zadeh in 1965 [110]. Unlike crisp sets where an element has a membership of 0 (does not belong to) or 1 (belongs to), each element in a fuzzy set has a degree of membership which is represented by a real number in the interval [0, 1]. Professor Zadeh defined a type-1 fuzzy set as:

> *Let X be a space of points (objects), with a generic element of X denoted by x. Thus, X = {x}.*
>
> *A fuzzy set (class) A in X is characterized by a membership (characteristic) function $f_A(x)$ which associates with each point in X a real number in the interval [0, 1], with the value of $f_A(x)$ at x representing the "grade of membership" of x in A. Thus, the nearer the value of $f_A(x)$ to unity, the higher the grade of membership of x in A. When A is a set in the ordinary sense of the term, its membership function can take on only two values 0 and 1, with $f_A(x)$ = 1 or 0 according as x does or does not belong to A.* [110, p. 339]

The concept of a type-1 fuzzy set, as defined above, can be illustrated in the following example of youngness, which is defined to answer the question "to what degree is a person young?". *X* is the universe of discourse, which is a set of all ages, *A* is the subset of young ages, *x* is the age of the person, and $f_A(x)$ is the degree of youngness of age *x*. If $f_A(x)$ equals to 1, *x* is 100% belongs to *A*. If $f_A(x)$ equals to 0, *x* is 0% belongs to *A*. If $f_A(x)$ equals to 0.3, *x* is 30% belongs to *A*. In this example, the membership function $f_A(x)$ is defined as in Equation (1) and is illustrated by the graph in Figure 2-6.

$$f_A(x) = \begin{cases} 1 & x < 25 \\ (35 - x)/10 & 25 \le x < 35 \\ 0 & x \ge 35 \end{cases} \tag{1}$$

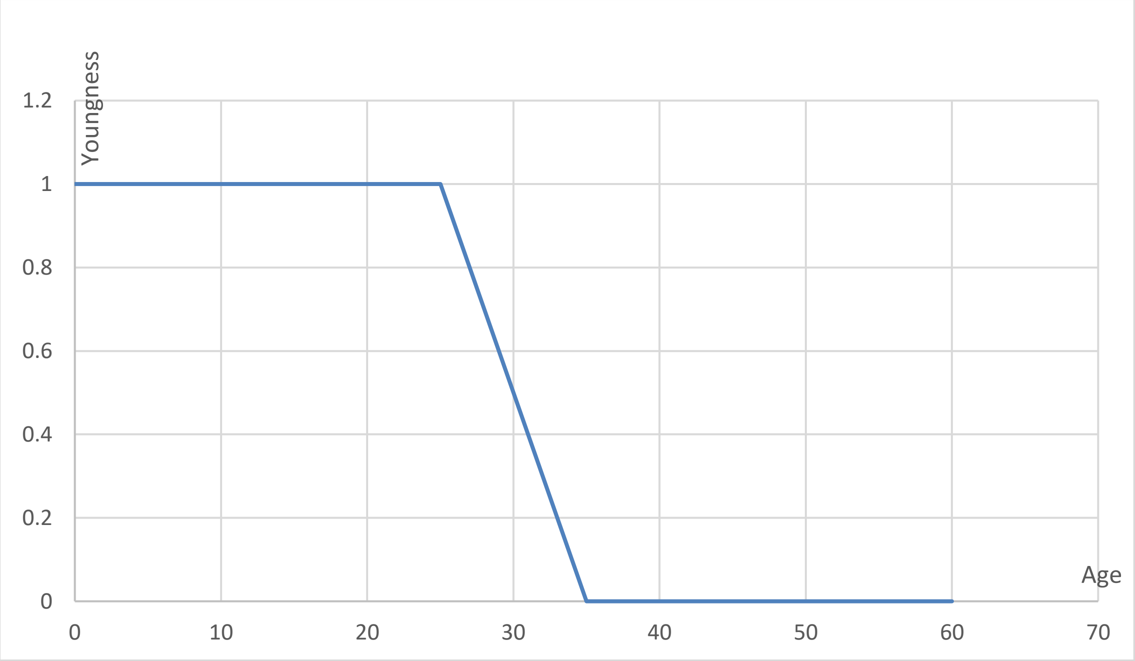


Figure 2-6: Degree of youngness based on age

The membership function is the significant component of a fuzzy set such that operations with fuzzy sets are defined via their membership functions [111]. In practice, the most common types of membership functions are triangular, trapezoidal, bell-shaped, gaussian, and sigmoidal [112, 113].

2.3.2 Type-2 Fuzzy Sets

In 1975, Zadeh introduced type-2 fuzzy sets [114]. Type-2 fuzzy sets allow the incorporation of uncertainty of membership functions into fuzzy set theory. In type-1 fuzzy sets, membership functions are totally crisp [115]. In a type-2 fuzzy set, a membership degree is also fuzzy and can be defined by a type-1 fuzzy set [116].

Therefore, type-2 fuzzy sets can model uncertainty. On the other hand, type-1 fuzzy sets can model vagueness (having a degree of membership), but not uncertainty (the degree of membership is also fuzzy). The membership of the membership (secondary membership) of a type-2 fuzzy set is also fuzzy. Ideally, type ∞ fuzzy sets must be used to completely represent uncertainty. However, this is not practical [117]. All the literature found in this research only deals with type-1 and type-2 fuzzy sets.

By default, a type-2 fuzzy set is called a general type-2 fuzzy set to distinguish it from an interval type-2 fuzzy set. In interval type-2 fuzzy sets, secondary membership functions are interval sets (i.e. the secondary memberships are either zero or one). A reason for having interval type-2 fuzzy sets is that general type-2 fuzzy sets are computationally intensive [118]. It is much simpler to use interval type-2 fuzzy sets than general type-2 fuzzy sets [115]. In fact, interval type-2 fuzzy sets are the most widely used type-2 fuzzy sets in practice [115, 119].

2.3.3 An Overview of Fuzzy Ontology

Although there have been a number of definitions of ontology [120], it can be defined as:

> *Ontology is an explicit specification of conceptualization* [121, p. 199]. *In computer science, ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships among those concepts* [122, p. 43].

There have also been several definitions of fuzzy ontology in the literature [120]. In essence, a fuzzy ontology is an ontology that contains fuzzy concepts (each fuzzy concept is a fuzzy set). In a general sense, *“a fuzzy ontology is a shared model of some domain which is often conceived as a hierarchical data structure containing all concepts, properties, individuals, and their relationships in the domain, where these concepts, properties and so on may be defined imprecisely”* [123, p. 91]. Formally, a fuzzy ontology can be represented as a quintuple $F=<I, C, T, N, X>$ [124, p. 13] where:

- *I is the set of individuals (objects), also called instances of the concepts.*
- *C is the set of fuzzy concepts (classes of individuals, or categories, or types). Each concept is a fuzzy set on the domain of instances.*
- *The set of entities of fuzzy ontology is defined by* $E = C \cup I$.
- *T denotes the fuzzy taxonomy relations among the set of concepts C. It organizes concepts into sub-(super-) concept tree structures. The taxonomic relationship T(i, j) indicates that the child j is a conceptual specification of the parent i with a certain degree.*
- *N denotes the set of non-taxonomy fuzzy associative relationships that relate entities across tree structures, for example:*
    - *Naming relationships, describing the names of concepts*
    - *Locating relationships, describing the relative location of concepts*
    - *Functional relationships, describing the functions (or properties) of concepts*
- *X is the set of axioms expressed in a proper logical language, i.e., predicates that constrain the meaning of concepts, individuals, relationships and functions.*

Ontology brings several advantages over other traditional methods of data management, such as relational database schemas. Using ontologies offers knowledge sharing, reuse of existing knowledge, and information integration [120]. As such, ontology plays a prominent role in the Semantic Web and in other forms of knowledge management [125]. In addition, automated reasoning, which has been the focus from the very start of Artificial Intelligence (AI) [126], is enabled by ontologies. Automated reasoning can be conducted using inference rules [127]. Berners-Lee et al. [127] gave an example of the power of inference rules with real-world data below:

> *If a city code is associated with a state code, and an address uses that city code, then that address has the associated state code. A program could then readily deduce, for instance, that a Cornell University address, being in Ithaca, must be in New York State, which is in the U.S., and therefore should be formatted to U.S.*

> *standards. The computer does not truly "understand" any of this information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user.* [127, p. 10]

Fuzzy ontology, which incorporates fuzzy concepts into an ontology, provides additional advantages to the use of ontology. Fuzzy ontology can represent vague and uncertain information, which is common in real-world scenarios. For example, "*the guest house is a cheap, small and more hospitable hotel*" [120, p. 64]. Therefore, fuzzy ontology can help represent real world knowledge, which is not always crisp but often vague and imprecise.

The following arguments are based on [128]. With a wealth of literature, fuzzy ontology has proved to be very useful in many application domains, including information retrieval, semantics extraction and analysis, knowledge mining, clustering, integration, decision making, and knowledge representation and reasoning. However, research on fuzzy ontology is still in the development stage and important challenges remain such as construction, mapping, integration, query, and storage. These challenges and strategies for overcoming them need to be more deeply investigated.

2.3.4 Fuzzy Ontology Representation Languages

RDFS (Resource Description Framework Schema) and OWL (Web Ontology Language) are the most widely used languages for describing ontologies [128]. RDF (Resource Description Framework) is a model that describes things as triples; each triple is in the form of <*subject*><*predicate*><*object*>, for example "CVD is the number one cause of death". RDFS is considered to be a primitive language providing basic elements for writing ontologies; however, a more powerful language is needed to deal with complex relationships among objects [129]. OWL was built on top of RDF and RDFS adding semantic richness that allows reasoning. For example, if there is an RDF statement "Professor A teaches the Data Mining class", then with OWL, it is also implied that "the Data Mining class is taught by Professor A" [130]. In fact, OWL is the standard language for writing ontologies recommended by W3C (World Wide Web Consortium) [128]. The current version of OWL is OWL 2 [131].[4]

To describe fuzzy ontologies, several approaches have been proposed. These include fuzzy extensions of RDF/RDFS [132-137], fuzzy extensions of OWL 1 [138-140], and frameworks to represent fuzzy ontologies using OWL 2 [141]. The fuzzy extensions of RDF generally allow the addition of a degree of truth to an RDF triple, such as “Auckland is a big city to degree 0.8” [128]. There is not enough richness to fully represent fuzzy ontologies in these approaches. The fuzzy extensions of OWL 1 introduce new syntax, and thus current ontology editors cannot be used [141]. Recently, the approach by Bobillo and Straccia [141] of representing fuzzy ontologies using OWL 2 annotation properties has received significant attention [128]. In addition to the framework to represent fuzzy ontologies using the existing OWL 2 language, Bobillo and Straccia also developed a plugin for Protégé to create and edit fuzzy ontologies, and three parsers to read and transform those created fuzzy ontologies into formats that can be read by fuzzy DL reasoners, such as fuzzyDL [142] and DeLorean [143]. Details of the Protégé plugin and the parsers are mentioned in the next section (section 2.3.5); however, it should be noticed that all the approaches mentioned currently support only type-1 fuzzy logic, not type-2 fuzzy logic.

2.3.5 Fuzzy Ontology Tools

The number of tools for construction and management of fuzzy ontologies seems, based on this investigation, to be low. In addition, all the tools to my knowledge only support type-1 fuzzy logic. Well known tools include:

- *Fuzzy OWL 2 Protégé plugin:* Bobillo and Straccia [141] developed this plugin to make the syntax of the annotations transparent to users when creating fuzzy ontologies. This means that users do not need to know and type the annotations but instead use the plugin’s GUI (graphic user interface) to create fuzzy ontologies. As part of this doctoral research, this plugin was verified as being compatible with Protégé versions 4.1 and 4.3. As mentioned previously, Bobillo

[4] To avoid confusion in this body of work, OWL 1 means the first version of OWL; OWL 2 means the second version of OWL; Using OWL without a version number means the versions can be ignored in the context.

and Straccia also developed three parsers (one general parser and two specific parsers) to translate the fuzzy ontologies created using the plugin into formats suitable for existing fuzzy DL reasoners. The general parser can be customised for any specific fuzzy DL reasoner. Bobillo and Straccia adapted the general parser to create the two specific parsers, one for fuzzyDL and the other one for DeLorean. It should be noted that this plugin and the parsers support type-1 fuzzy sets only, not type-2 fuzzy sets. The parsers are written in Java and use the OWL API 3 [144], a well-known Java API for working with OWL 2 ontologies. The Protégé plugin and the parsers are publicly available for download on their website.[5]

- *Fuzzy Protégé plugin:* Ghorbel et al. [145] created a plugin for Protégé 3.3.1 to build fuzzy concepts and roles, and allow automatic computing of membership degrees. In addition, this plugin allows querying the created fuzzy ontologies based on fuzzy criteria. However, it seems that this plugin is no longer available for download and installation into Protégé.

- *Fuzzy KAON:* Calegari and Ciucci [139, 146] developed a way to define and manage fuzziness directly in the KAON ontology editor. However, KAON is based on RDFS and therefore it is not possible to represent all the constructors and axioms of their developed Fuzzy-OWL, a fuzzy extension of OWL 1 [139]. Therefore, it would be necessary and more useful to define and implement a way to represent fuzzy ontologies in KAON2, a successor to the KAON project (KAON1). KAON2 is based on OWL-DL, a sublanguage of OWL. Both KAON and KAON2 can be downloaded from websites.[6]

[5] http://www.umbertostraccia.it/cs/software/FuzzyOWL/index.html

[6] https://sourceforge.net/projects/kaon/ for KAON, and http://kaon2.semanticweb.org/ for KAON2

2.4 CASE-BASED REASONING

2.4.1 An Overview of CBR

Case-based reasoning (CBR) is a problem solving paradigm that resolves a problem using the specific knowledge of previously experienced cases [16]. This paradigm is associated with a CBR cycle (methodology[7]). The CBR cycle is comprised of four activities [147, p. 303]:

1. ***Retrieve** similar cases to the problem description.*
2. ***Reuse** a solution suggested by a similar case.*
3. ***Revise** or adapt that solution to better fit the new problem if necessary.*
4. ***Retain** the new solution once it has been confirmed or validated.*

As CBR is only a problem-solving paradigm accompanied by the CBR methodology, actual techniques are needed to build CBR systems to solve real-world problems. Common techniques include nearest neighbours, fuzzy logic, and database technology [147]. Details of these techniques are described in section 2.4.2.

Research on CBR has been gaining momentum with more and more practical applications produced in a variety of domains [148]. Some examples of those domains are law [149], education [150], marine [101], software engineering [151], and health care [9]. As the CBR methodology helps provide computational models very close to human reasoning, which mostly uses past experiences to solve daily problems [152], it is reasonable for such high research interest in CBR and the applicability of CBR in various domains.

2.4.2 Common techniques used for CBR

2.4.2.1 Nearest neighbour

Nearest neighbour is perhaps the most widely used technique in CBR to retrieve similar cases [147]. Nearest neighbour algorithms calculate the similarity (distance) between the problem (target) case and an existing case in the case base (case library). The calculation

[7] A system of methods for how things are proceeded

is repeated for every case in the case base to identify *k* nearest neighbours. Outcomes for the target case are decided based on these *k* nearest neighbours using majority voting.

Among the distance functions (such as Euclidean, Cosine Similarity [153], Minkowsky [154], and Chi-square [155]) used in k nearest neighbours (KNN), Euclidean is the most widely used [156]. The Euclidean distance between *A* and *B* is generally represented by Equation (2).

$$dist(A, B) = \sqrt{\sum\nolimits_{i=1}^{m} (x_i - y_i)^2} \qquad (2)$$

where *A* and *B* are feature vectors $A = (x_1, x_2, \ldots, x_m)$ and $B = (y_1, y_2, \ldots, y_m)$, *m* is the number of features of *A* and *B*.

Distances are usually normalised to fall within the [0, 1] range [147]. This helps deal with the issue of sensitivity to a broad range of values in a single feature that may govern the distance. The normalised Euclidean distance is generally represented by Equation (3) [156]. In this case, all features x_i and y_i are unit normals in the [0, 1] range.

$$dist(A, B) = \sqrt{\frac{\sum_{i=1}^{m} (x_i - y_i)^2}{\mathrm{m}}} \qquad (3)$$

2.4.2.2 Fuzzy logic

Fuzzy logic becomes helpful for a CBR system to deal with qualitative terms, lack of certainty in information, or sudden changes in outcome categories due to small changes in features (e.g. risk factors). In a CBR system, numerical features (crisp values) can be converted into qualitative terms (fuzzy values) for indexing and retrieval [157]. In addition, a major task in CBR is to measure similarities, which are inherently fuzzy in nature [157]. For example, colour matching is defined as *excellent*, *good*, *fair*, or *poor* in a CBR system created by General Electric to determine what colourants to use [158]. Moreover, information can be uncertain in real world scenarios and therefore it would be more appropriate to represent it as fuzzy data. For example, smoking status can be represented as *light*, *medium*, or *heavy* smoking as a person may not know exactly how many cigarettes they smoke in a day. On the other hand, many existing prediction models (e.g. CVD risk prediction models) represent outcomes as categories (e.g. low, moderate, high) which are in fact crisp representations. The issue in this case is that small changes in risk factors may move the outcome between categories [12]. As a result, a totally inappropriate treatment plan for the person may be recommended. In this situation, fuzzy logic can smooth the transition in the outcome categories.

2.4.2.3 Database technology

Using database technology to build a CBR system is perhaps the simplest form [147]. Relational Database Management Systems (RDBMS) have proven to be an appropriate means to store and retrieve large volumes of data as they have been widely used in the software industry. The SQUAD system [159] developed by NEC Japan as a software quality control advisory system is an example of using database technology to build CBR systems.

2.4.3 CBR Tools

There have been several tools for building CBR systems [148, 160]. Popular tools are CBR Shell [161], FreeCBR [162], jCOLIBRI [107], myCBR [163], eXiTCBR [164], and IUCBR [165]. Although there may be unknown or unpopular tools that have not been identified in this research, the number of CBR tools is perhaps still low when considering the high research interest in CBR systems and the applicability of CBR to a wide variety of domains.

Table 2-4: CBR tool summary

Tool	**Programming language**	**Support ontology**	**Support fuzzy ontology**
CBR Shell	Java	No	No
FreeCBR	Java	No	No
jCOLIBRI	Java	Yes	No
myCBR	Java	No	No
eXiTCBR	Java	No	No
IUCBR	Java	No	No

Table 2-4 provides a summary of the CBR tools based on the criteria for this research. These criteria are development programming language, whether the tool supports ontology, and whether the tool supports fuzzy ontology. All six tools evaluated were developed in Java. Java is a cross-platform, mainstream programming language and is favoured by open-source and academic communities. Only jCOLBRI has the features needed to work with cases stored in ontologies. None of the tools support fuzzy ontologies.

2.5 CHAPTER SUMMARY

2.5.1 Cardiovascular Disease Prediction

There are a large number of existing CVD prediction models. In this study, well-known ones are identified and classified as conventional Framingham models, augmented Framingham models, and alternatives to Framingham [26]. The conventional Framingham models [27-32] were developed as part of the FHS. The augmented Framingham models tried to add additional risk factors, especially biomarkers such as C-reactive protein, into the Framingham equations, and/or create different presentations for Framingham equations e.g. represented as charts or tables. The alternatives to Framingham [67-77] are models developed from cohorts not from the FHS.

In terms of prediction methods, the majority of the existing models used regression-based techniques, in which Cox proportional hazards modelling [37] dominated. However, regression-based prediction models are known to have limitations. These are inaccuracy for individuals, inaccuracy for other cohorts, inflexibility of handling interventions, requirement of complete clinical data, deficiency of handling inaccurate data or results, deficiency of handling vagueness and uncertainty of data or results, and poor explanatory capacity.

As the problem of CVD prediction has not been solved, this thesis investigates new approaches of using fuzzy ontology and CBR for chronic disease prediction, including CVD prediction. Though there has been no fuzzy ontology CBR system in the CVD prediction domain yet, there are a few studies of such approaches in other domains, for example El-Sappagh et al. [9] in diabetes diagnosis. Therefore, fuzzy ontology CBR approaches appear worthy of investigation. From these, six research questions (stated in section 1.2) were created for this study.

2.5.2 Fuzzy Ontology

A fuzzy ontology is an ontology whose content contains fuzzy concepts (each fuzzy concept is a fuzzy set). Typical types of fuzzy sets are Type-1 and Type-2. A type-1 fuzzy set is different from a crisp set in that the membership value can be any real number in [0, 1]. In a type-2 fuzzy set, the membership value is also fuzzy and can be represented as a type-1 fuzzy set.

A number of approaches have been proposed to describe fuzzy ontologies. These include fuzzy extensions of RDF/RDFS, fuzzy extensions of OWL 1, and methods to represent fuzzy ontologies using OWL 2. Among these approaches, the recent approach by Bobillo and Straccia [141] to describe fuzzy ontologies using OWL 2 annotation properties has been in the spotlight.

However, there is still a lot of room for research and development in terms of languages and tools to create and maintain fuzzy ontologies. The number of tools is still limited. Moreover, none of the existing languages and tools appear to support type-2 fuzzy sets.

The Fuzzy OWL 2 Protégé plugin developed by Bobillo and Straccia [141] was chosen for creating fuzzy ontologies in this study (section 3.4.4).

2.5.3 Case-Based Reasoning

Case-based reasoning (CBR) is a problem-solving paradigm that solves a problem using the solutions of similar past problems. Its cycle involves four activities: Retrieve, Reuse, Revise, and Retain. Popular techniques for building CBR systems include nearest neighbour, fuzzy logic, and database technology.

Though research interest in CBR is high and the applications for CBR are numerous, tools for building a CBR system are still limited. Moreover, to my knowledge, none of the current tools support fuzzy ontologies.

Chapter 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter explains in detail the research methods used and their application in this study. It starts with a description of the choice to adopt a positivist research paradigm, whose beliefs about the world led to the formation of the research questions and guided how the study should be approached. An explanation of how Design Science was chosen as the research methodology is then provided, outlining the systematic way in which this research was carried out. Alongside Design Science, a conceptual research framework, implementing the Design Science methodology in Information Research, customised for this study, and guidelines are also introduced. Next, the chapter describes the strategies and plans to develop the CRISK prediction model with the aim of solving the CVD prediction problem. After that, there is a description of how datasets were collected and selected, how the experiments were designed, and how data were prepared. Finally, an evaluation protocol consisting of evaluation metrics, external validation, and comparison to existing models to assess the developed CRISK model is described.

3.2 RESEARCH PARADIGM

The positivist paradigm was chosen to shape this study. The reason was that it was believed that, in this research that aims to develop a CVD prediction model, knowledge could be discerned using appropriate scientific methods [166]. As such, the results of the research approach can be objectively tested for accuracy and other measures. In terms of data collection and analysis, quantitative methods were used [167].

3.3 METHODOLOGY, FRAMEWORK, AND GUIDELINES

As this research aims to develop a novel CVD prediction model, which can be seen as a new and innovative artifact serving human purposes, Design Science was chosen as a suitable methodology for this research [168]. In Design Science, the research activities are twofold: *build* and *evaluate* [169]. *Build* refers to the activity of constructing the artifact for a specific purpose, showing that such an artifact can be made. *Evaluate* is the activity of developing criteria and assessing the performance of the created artifact against those criteria [169].

To understand, execute, and evaluate Design Science research in Information Systems (IS), a conceptual framework for this body of work was created. This research framework is depicted in Figure 3-1. It was adapted from the Information System Research Framework (ISRF) introduced by Hevner et al. [169] to suit this research. The two activities of the Design Science methodology, forming a spiral model, are seen in the centre of the framework. A spiral model allows ease of management as problems can be identified early and appropriate actions can be taken quickly [170].

As can be seen in Figure 3-1, this research used applicable knowledge from the Knowledge Base to develop an artifact (a CVD prediction model) for business needs from the Environment. The Environment contains goals/tasks/problems/opportunities of CVD prediction that define the business needs. The artifact resulting from the research must be relevant to the business needs of the Environment. The Knowledge Base includes foundations, methodologies, and tools. The research process must be rigorous which can be achieved by utilising appropriate knowledge from the Knowledge Base.

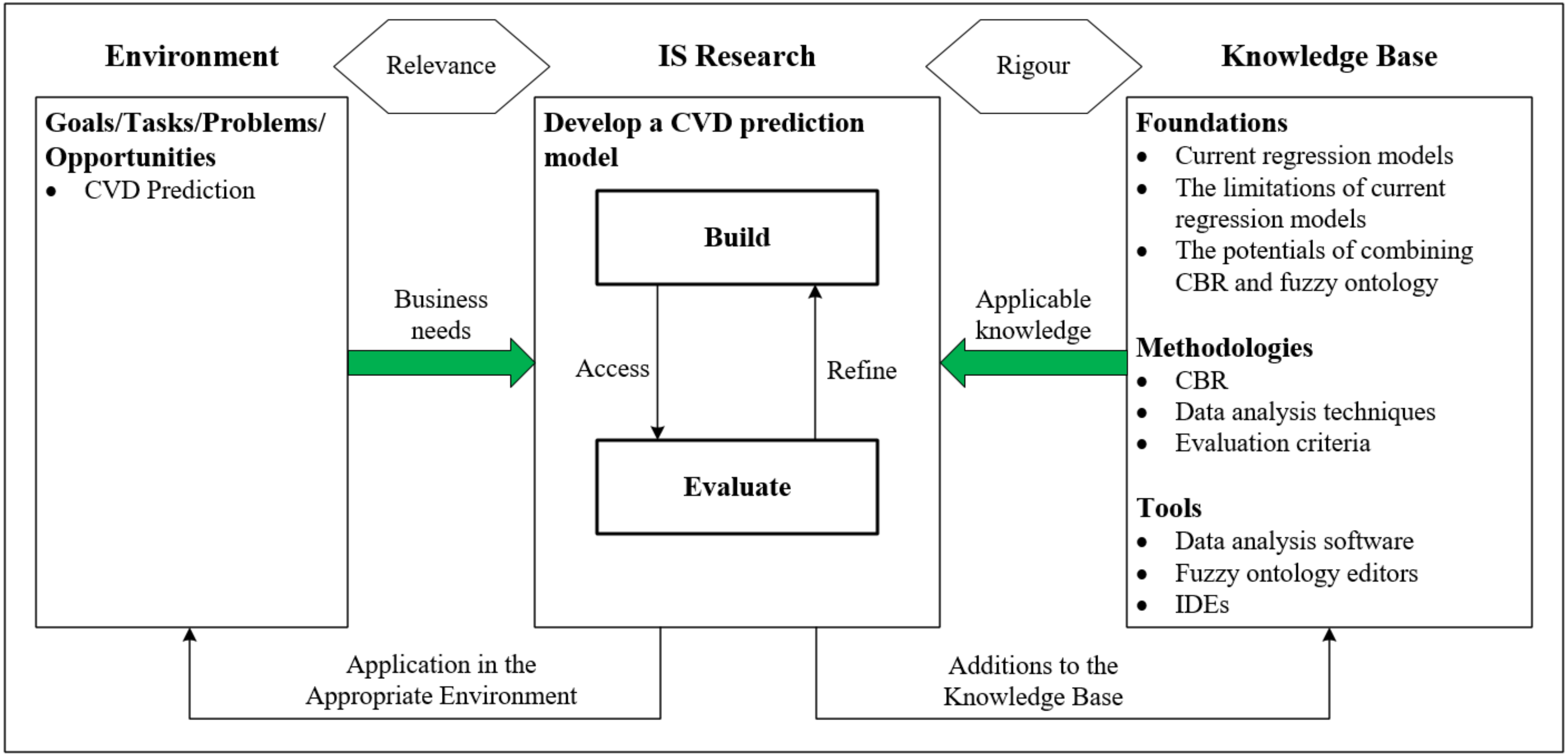

Environment
Relevance
IS Research
Rigour
Knowledge Base
Goals/Tasks/Problems/ Opportunities
• CVD Prediction
Business needs
Develop a CVD prediction model
Build
Access
Refine
Evaluate
Applicable knowledge
Foundations
• Current regression models
• The limitations of current regression models
• The potentials of combining CBR and fuzzy ontology
Methodologies
• CBR
• Data analysis techniques
• Evaluation criteria
Tools
• Data analysis software
• Fuzzy ontology editors
• IDEs
Application in the Appropriate Environment
Additions to the Knowledge Base

Figure 3-1: Research framework (adapted from Hevner et al. [169])

Hevner et al. [169] also suggested seven guidelines (Table 3-1) to assist researchers, reviewers, editors, and readers to understand the requirements of effective Design Science research.

Table 3-1: Design Science research guidelines (adapted from Hevner et al. [169])

Guideline	Description
Guideline 1: Design as an Artifact	Design Science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
Guideline 2: Problem Relevance	The objective of Design Science research is to develop technology based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well executed evaluation methods.
Guideline 4: Research Contributions	Effective Design Science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigour	Design Science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a Search Process	The search for an effective artifact requires utilising available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design Science research must be presented effectively both to technology oriented as well as management-oriented audiences.

Next, section 3.4 describe strategies and plans to develop the CRISK prediction model, which helps answer RQ1, RQ2, RQ3, and RQ5.

3.4 STRATEGIES AND PLANS TO DEVELOP THE CRISK PREDICTION MODEL

3.4.1 CRISK as a CBR system whose case base is a fuzzy ontology

I decided to develop the CRISK prediction model as a CBR system whose case base is a fuzzy ontology. Fuzzy ontologies are able to represent the vagueness and uncertainty of data [12]. On the other hand, CBR has been recommended as being useful in disease prediction [13, 14]. Fuzzy ontology CBR systems have been used in a number of domains and have shown the potential in the medical diagnosis domain (section 2.2.6). A recent study [9] followed this approach of combining fuzzy ontology and CBR and did well in diabetes diagnosis. Therefore, it would be worth trying a combination of fuzzy ontology and CBR in developing the CRISK model for CVD prediction.

Type-1 fuzzy ontology was decided to be used for the case base. There were two reasons for this decision. I would like to start with something simple first (type-1 fuzzy ontology is simpler than type-2 fuzzy ontology, as explained in section 2.3). Another reason for the decision of using type-1 fuzzy ontology was based on the knowledge that none of the existing fuzzy ontology tools support type-2 fuzzy sets (section 2.3.5). Developing a fuzzy ontology tool that supports type-2 fuzzy sets would be not feasible in the three-year timeframe of a PhD.

Table 3-2 restates the problems with current regression models identified in Table 2-3 and adds arguments explaining how a combination of fuzzy ontology and CBR might be able to resolve them.

Table 3-2: Possible solutions for the problems with the current regression models

Limitation #	Name	Might be solved by	Explanation
Limitation 1	Inaccuracy for individual	CBR	CBR targets on individuals, not on populations as generating a solution for a new case based on the solutions of the most similar existing cases.
Limitation 2	Inaccuracy for other cohorts	CBR	A CBR approach will be able to incorporate examples from new cohorts into the case base, or even replace the case base completely when dealing with different populations.
Limitation 3	Inflexibility of handling intervention	N/A	Not sure if CBR or fuzzy ontology would be able to address this limitation at this stage
Limitation 4	Requirement of complete clinical data	CBR	Even there is missing data, it is still possible to retrieve the closest cases to the input case using CBR
Limitation 5	Deficiency of handling inaccurate data or result	Fuzzy ontology	Fuzzy ontology usage can help deal with vagueness and uncertainty of data. In addition, representing prediction results using fuzzy sets can tolerate prediction errors. With models where prediction results are fuzzy values, a small error in prediction might increase or descrease membership values of the prediction outcomes but does not completely shift the person to a wrong category, such as from "high risk" to "low risk".
Limitation 6	Deficiency of handling vagueness of data or result	Fuzzy ontology	
Limitation 7	Deficiency of handling uncertainty of data or result	Fuzzy ontology	

Limitation 8	Poor explanatory capacity	CBR and fuzzy ontology	Fuzzy ontology shows relationships among risk factors, individuals, and CVD outcomes while the usage of CBR helps explain why those outcomes are prediction results based on the closest cases retrieved

3.4.2 A new way to define "high risk" and "low risk" categories

Time before the CVD event (CVD Interval) is important. When a person undergoes an examination that includes calculating their CVD risk, I think they would like to know about when they would have CVD. Thus, it is reasonable that when someone asking what their CVD risk is would be satisfied with a representation that said a CVD event soon means "high risk" and that a CVD event in a long time or no CVD at all can be called "low risk". Adding this CVD Interval and the memberships of the low risk and high risk categories may provide useful information.

This is not the same as most approaches, e.g. a regression model. A typical regression model, such as D'Agostino et al. [31], calculates the probability (risk) of having CVD within e.g. 10 years for a person, then classifies them into a risk category by comparing this risk value to a threshold value, e.g. 20%. This means that if the risk value is greater than or equal to 20%, the person is assigned to the "high risk" category, otherwise the "low risk" category. With this way, the question of "when the person would have CVD" cannot be answered.

I defined CVD Interval as the time interval from the examination to when the first CVD event happens (Figure 3-2).

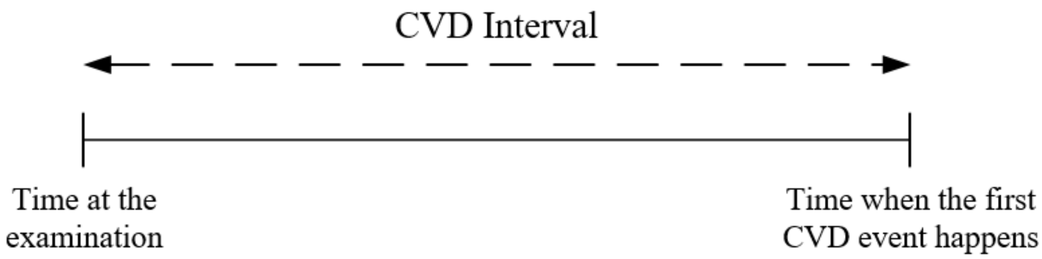


Figure 3-2: Definition of CVD Interval

I defined two fuzzy sets, "High CVD Risk" and "Low CVD Risk", whose membership functions base on CVD Interval, to represent CVD prediction outcomes for a person. The membership functions of these two fuzzy sets are described in Equations (4) and (5) respectively and are illustrated by the graphs in Figure 3-3, where *x* is CVD Interval in the year unit, $\mu_{\text{High CVD Risk}}$ or μ_H is the "High CVD Risk" membership value, and $\mu_{\text{Low CVD Risk}}$ or μ_L is the "Low CVD Risk" membership value.

The reason for using fuzzy sets to represent CVD prediction outcomes was to make transitions between risk categories less abrupt. With risk categories represented by crisp sets, a small change in the prediction result can completely shift an individual between two categories, e.g. from low risk to high risk. Fuzzy sets help mitigate this issue as having overlapping areas between categories (e.g. Figure 3-3).

The reason for defining two risk categories was based on the aim of this research to develop a model for prediction of CVD within 10 years (section 1.2). This simplest approach is to have two risk categories: one category for people who develop CVD within 10 years and another one for those who do not.

The reason for defining "High CVD Risk" and "Low CVD Risk" as trapezoidal membership functions (Figure 3-3) is explained as follows. I assume that when CVD Interval ≤ 5 years, the person completely belongs to the high risk category, as the threshold is 10 years (prediction for within 10 years). Similarly, I assume that when CVD Interval ≥ 15 years, the person completely belongs to the low risk category. Among the five common types of membership functions (triangular, trapezoidal, bell-shaped, gaussian, and sigmoidal), mentioned in section 2.3.1, only trapezoidal is suitable to represent this way of thinking about "High CVD Risk" and "Low CVD Risk".

The reason I chose 5 years as the benchmark to start decreasing the high risk membership value from 1 and increasing the low risk membership value from 0 (Figure 3-3) was that 5 years is in the middle of 0 and 10 years. Different values other than 5 years (e.g. 6 years) could have been chosen and that would likely have resulted in different testing results of prediction performance.

The reason for choosing 15 years as the benchmark where the high risk membership gets down to 0 and the low risk membership reaches 1 (Figure 3-3) is explained as follows. Fifteen years was chosen for symmetry and because this results in the intersection point of the "High CVD Risk" and "Low CVD Risk" graphs having the x coordinate of 10 years, which is needed to conclude if the person will develops CVD within 10 years by comparing the high risk membership with the low risk membership (if $\mu_H \geq \mu_L$, the person will develop CVD within 10 years).

$$\mu_{High\ CVD\ Risk}(x) = \mu_H(x) = \begin{cases} 1 & x < 5 \\ -\dfrac{x}{10} + 1.5 & 5 \leq x < 15 \\ 0 & x \geq 15 \end{cases} \tag{4}$$

$$\mu_{Low\ CVD\ Risk}(x) = \mu_L(x) = \begin{cases} 0 & x < 5 \\ \dfrac{x}{10} - 0.5 & 5 \leq x < 15 \\ 1 & x \geq 15 \end{cases} \tag{5}$$

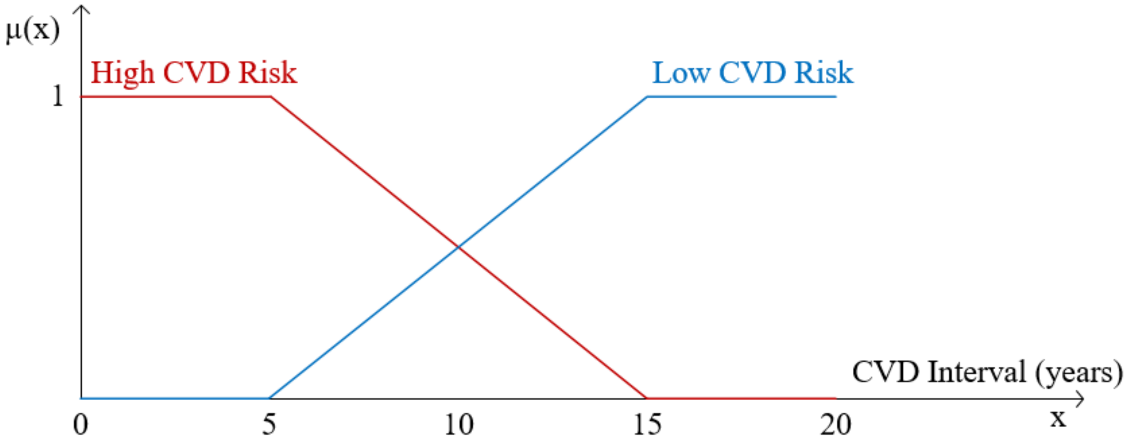


Figure 3-3: High CVD Risk and Low CVD Risk membership functions for this research

When dealing with data whose outcomes are known, a person is assigned to the "High CVD Risk" set (category) or the "Low CVD Risk" set, or to both sets depending on their CVD Interval. Interpretation is below:

If the person has a CVD event within 5 years, they only belong to the "High CVD Risk" set with the membership values $\mu_H = 1$ and $\mu_L = 0$.

If the person has no CVD event within 15 years, they only belong to the "Low CVD Risk" set with the membership value $\mu_L = 1$

If the person has a CVD event between 5 and 15 years' time, they belong to both "High CVD Risk" and "Low CVD Risk" sets. In this case, $0 < \mu_H < 1$ and $0 < \mu_L < 1$.

If the person has a CVD event at exactly 10 years' time, their "High CVD Risk" membership value equals their "Low CVD Risk" membership value, $\mu_H = \mu_L = 0.5$.

3.4.3 Prediction Process Strategies for CRISK

CRISK will calculate the predicted risk class, μ_H, μ_L, and CVD Interval for a new case based on using membership functions and nearest neighbours. Strategies for the prediction process are below:

- A fuzzy KNN algorithm proposed by Keller et al. [171] is based on to retrieve closest cases to the input case
- From these closest cases, μ_H, and μ_L are calculated for the input case using the above fuzzy KNN algorithm. Then, the risk category is decided for the input case based on μ_H and μ_L (if $\mu_H \geq \mu_L$ then "High CVD Risk", otherwise "Low CVD Risk")
- Defuzzification to get the predicted CVD Interval

The prediction process of CRISK is explained in more details in section 4.2.

3.4.4 Plans to develop CRISK

The CRISK prediction system was decided to have four modules as follows:

1. **The Constructor module**: for creating fuzzy ontologies
2. **The Experimenter module**: for experimentation of different datasets, which are fuzzy ontologies created by the Constructor module, to evaluate prediction performance
3. **The Batch Experimenter module**: a command line module designed to run long and repetitive experimentation jobs, for example to find a combination of predictors that creates the most accurate prediction model
4. **The Predictor module**: for predicting CVD risk for each single case (person) Details of each module are further explained in Chapter 5.

Figure 3-4 gives a summary of the plans to develop the CRISK prediction model. These plans were made based on results from the analysis of existing foundations, methodologies, and tools in the Knowledge Base (Figure 3-1), CVD prediction goals/tasks/problems/opportunities in the Environment (Figure 3-1), and with consideration of the time constraint of three years for PhD research.

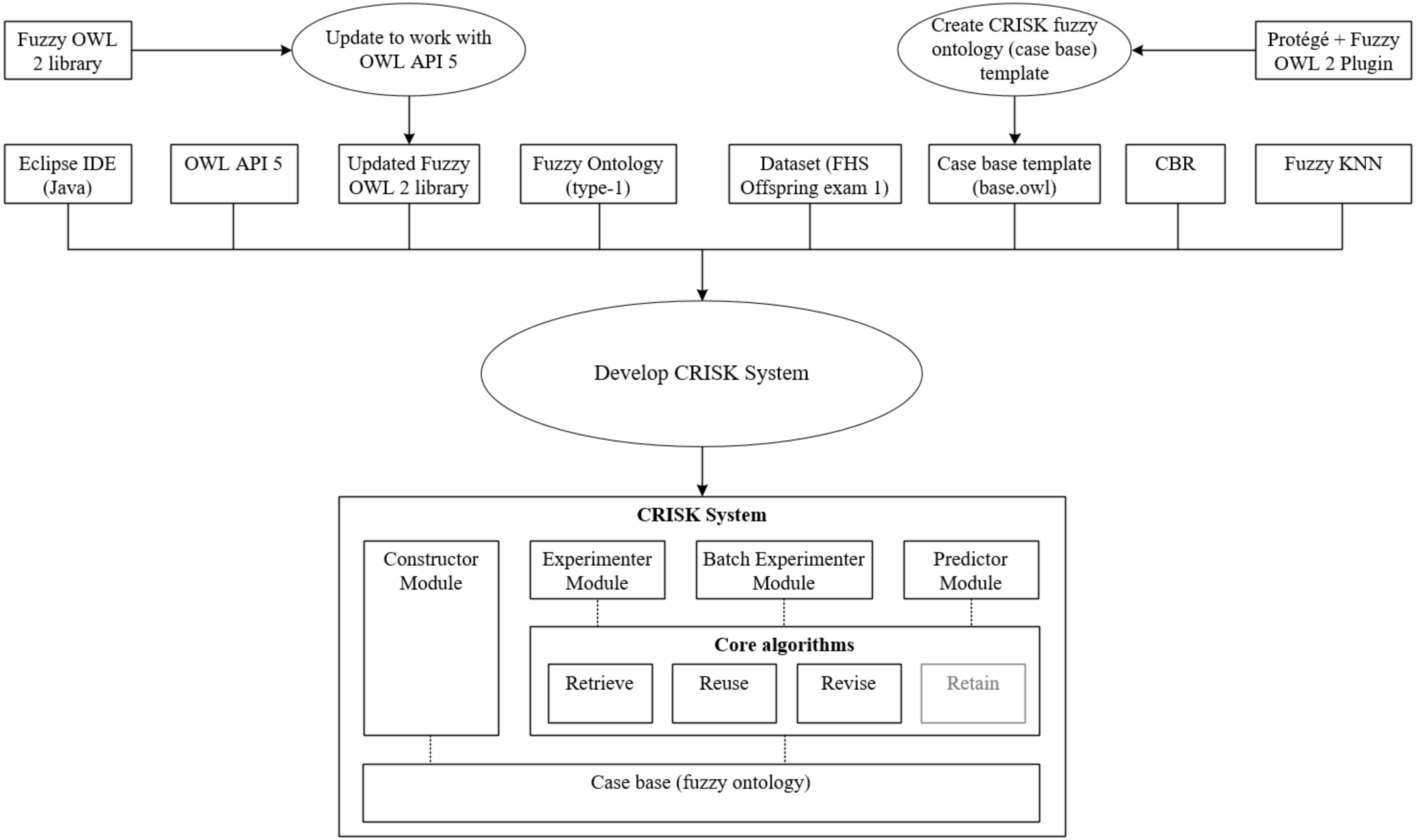

Fuzzy OWL 2 library
Update to work with OWL API 5
Create CRISK fuzzy ontology (case base) template
Protégé + Fuzzy OWL 2 Plugin
Eclipse IDE (Java)
OWL API 5
Updated Fuzzy OWL 2 library
Fuzzy Ontology (type-1)
Dataset (FHS Offspring exam 1)
Case base template (base.owl)
CBR
Fuzzy KNN
Develop CRISK System
CRISK System
Constructor Module
Experimenter Module
Batch Experimenter Module
Predictor Module
Core algorithms
Retrieve
Reuse
Revise
Retain
Case base (fuzzy ontology)

Figure 3-4: Plans to develop the CRISK system

The decision was made to implement the CRISK system from scratch instead of using or extending an existing CBR tool. The main reason for this decision was that there are no existing CBR tools available that support fuzzy ontology (section 2.4.3).

The process for developing the CRISK system can be summarised as follows:

> First, a case base template (base.owl) was created using Protégé 4.3 with the Fuzzy OWL 2 plugin [141]. Details of the base.owl file can be found in section 5.2. It is this template which the CRISK system uses to generate the case base. The template is a type-1 fuzzy ontology containing basic components including the fuzzy concepts of “High CVD Risk” and “Low CVD Risk”. The dataset used for creating the case base was decided to be the FHS Offspring Exam 1 dataset (see section 3.6 for justification of the decision).
>
> Next, the Fuzzy OWL 2 library containing the parsers [141] was updated to work with OWL API 5, which was the latest OWL API at the time that the case base template was designed and created.
>
> After that, the CRISK system was implemented in the Java programming language (Java 8) using Eclipse IDE (version 2018-09). The CRISK application used the OWL API 5 and the updated Fuzzy OWL 2 library to create and manipulate fuzzy ontologies. Core algorithms of the application were designed according to the CBR cycle: Retrieve, Reuse, Revise, and Retain. The Fuzzy KNN algorithm [171] was based on to develop the Retrieve, Reuse, and Revise algorithms. In Figure 3-4, the Retain algorithm is grayed out because it is out of the scope of this research (section 4.8). Details of these algorithms are from section 4.5 to section 4.8.

3.5 DATASET COLLECTION

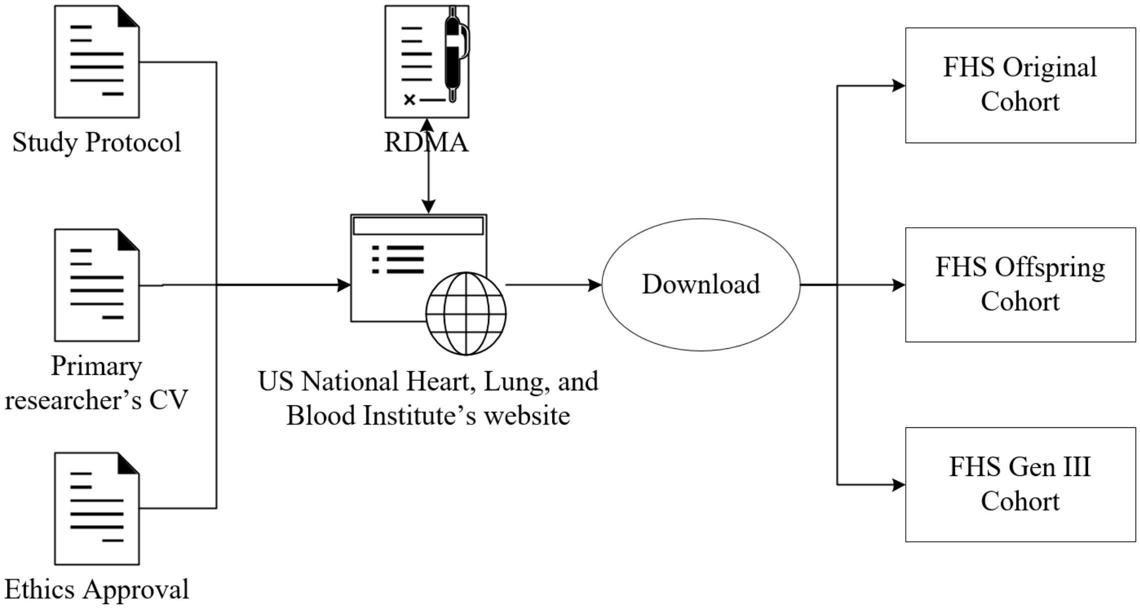


Figure 3-5: The formal process to obtain FHS datasets

Three FHS cohort datasets were obtained from the US National Heart, Lung, and Blood Institute (US-NHLBI) in June 2017. These were Original, Offspring, and Third Generation (Gen III) cohorts. To obtain these cohorts, a formal request process (Figure 3-5) was undertaken through the website of the US-NHLBI.[8] To start this process, a user account had to be created. The request form requires a study protocol, an ethics approval, and the CV of the primary researcher. After that, an RMDA (Research Materials Distribution Agreement) form issued by the US-NHLBI must be signed before the datasets can be accessed and downloaded. The ethics approval and the signed RMDA form can be found in Appendix A and Appendix B respectively.

FHS datasets were used for this research. The FHS has been broadly recognised as a premier longitudinal study whose background and design were reviewed by a large number of studies [172]. The participants in the FHS went through examinations every two years. Justification for fit-for-purpose is listed in [173] as below:

1. *The Framingham town was of adequate size to provide enough participants for the study.*
2. *It was compact enough that the study population could be observed conveniently.*

[8] https://biolincc.nhlbi.nih.gov/studies/framcohort/

3. *It contained a variety of socioeconomic and ethnic subgroups to provide contrasting groups for analysis.*
4. *The population was relatively stable to enable adequate follow-up for a long time. This was partly due to stable economy supported by a diversity of employment opportunities.*
5. *The town was located near a medical center which could provide consultations and the opportunity for educational development of the staff.*
6. *The physicians and other medical professionals in the town were highly supportive of the study and cooperated fully with its objectives.*
7. *Framingham contained two general hospitals at the beginning of the study. However, one closed shortly after the study began, SQ that a major portion of the medical care was provided by a single hospital.*
8. *Framingham, like most towns in Massachusetts, maintained an annual list of its residents.*
9. *The staff of a well-organised health department helped to provide death certificate information and other vital statistics.*
10. *Framingham had been the site of a community study of tuberculosis nearly 30 years before that had had successful participation by the townspeople. It was believed that this spirit of cooperation was still present in 1948.*

3.6 DATASET SELECTION

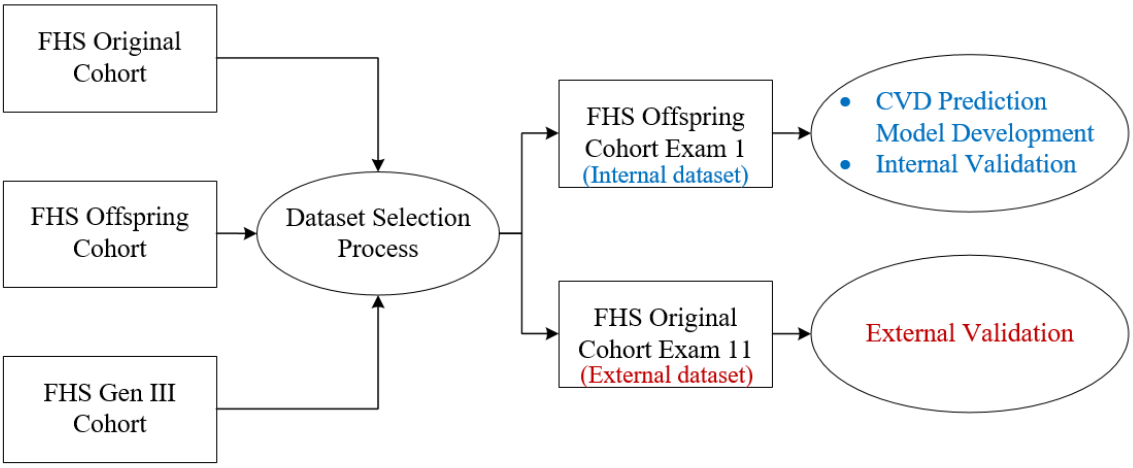


Figure 3-6: Dataset selection process

As the aim of this research was to build a model for prediction of CVD within 10 years, it was necessary to choose cohorts whose follow-ups were not less than 10 years. To know how long the follow-ups were, the column named “cvddate” was used. This date (the number of days since the first exam) was the date that the participant was diagnosed with CVD or, in the case of no CVD occurrence, the date of censoring (the last known date the participant did not have CVD). Figure 3-6 summarises the dataset selection process.

The FHS Original cohort was selected to go to the next step, data preparation. It was initiated in 1948, consisted of 5,079 participants, and went through 32 medical exams. Among these 5,079 participants, 3,189 people developed CVD and 1,890 people were recorded as not having CVD. The CVD dates of those who developed CVD ranged from 0 to 22,301 while ones of the latter ranged from 0 to 22,670. These made the FHS Original cohort eligible for this research as the follow-up had met the 10-year threshold (3,652.4 days).

The FHS Offspring cohort was also selected to go to the next step, data preparation. It was initiated in 1971, consisted of 5,013 participants, and went through 9 exams. Among these 5,013 participants, 1,372 people developed CVD and 3,641 people were recorded as not having CVD. The CVD dates of those who developed CVD ranged from −992 to 14,231 while ones of the latter ranged from 0 to 14,353. These made the FHS Original cohort eligible for this research as the follow-up was more than 10 years after the first medical examination.

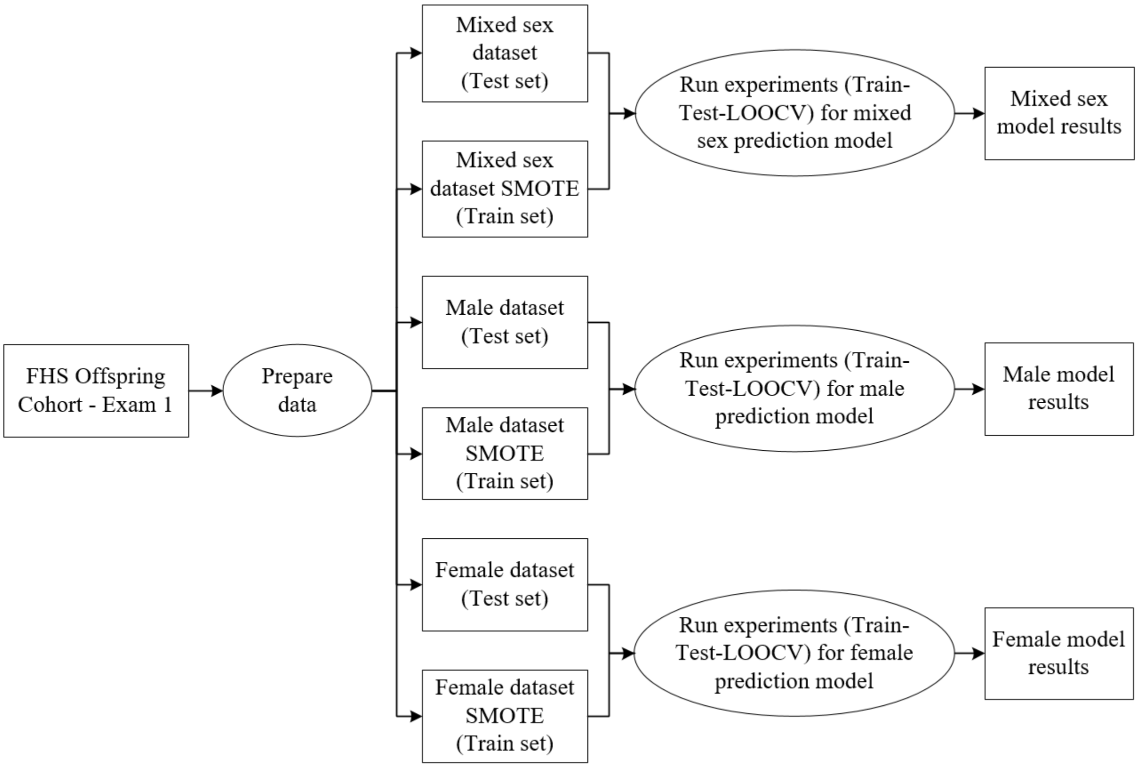
The FHS Gen III cohort was eliminated from the data for this research. The FHS Gen III dataset had 4,078 participants. The CVD date values ranged from −8,824 to 3,196. These made the FHS Gen III cohort ineligible for this research as the last follow-up had occurred within less than 10 years.

Between the two eligible cohorts, the FHS Offspring Cohort dataset collected based on Exam 1 was decided to be used as the main dataset for building the model. The reasons for this decision included the fact that more attributes related to CVD development such as HDL cholesterol [31, 32] were collected early, from Exam 1, for the FHS Offspring Cohort. In addition, in the course of this research, it was found that there was more missing data in the FHS Original Cohort than in the FHS Offspring Cohort.

The FHS Original Cohort dataset based on Exam 11 was chosen as a dataset for external validation. The reason for choosing Exam 11 was that eleven out of the thirteen predictor attributes chosen for the CRISK prediction model (section 6.5) could be found in this exam. The two predictor attributes missing in Exam 11 were triglycerides and lactate dehydrogenase (LDH). Other exams in the FHS Original Cohort did not do as well as

Exam 11 in terms of providing predictor attributes for the external validation purpose for the developed CRISK prediction model. In addition, the time gap between Exam 11 and the latest exam, Exam 32, of the Original Cohort is about 42 years, which is sufficient for a 10-year CVD study.

3.7 EXPERIMENTAL DESIGN



FHS Offspring Cohort - Exam 1
Prepare data
Mixed sex dataset (Test set)
Mixed sex dataset SMOTE (Train set)
Male dataset (Test set)
Male dataset SMOTE (Train set)
Female dataset (Test set)
Female dataset SMOTE (Train set)
Run experiments (Train-Test-LOOCV) for mixed sex prediction model
Run experiments (Train-Test-LOOCV) for male prediction model
Run experiments (Train-Test-LOOCV) for female prediction model
Mixed sex model results
Male model results
Female model results

Figure 3-7: Experimental Design for the development of CRISK

Figure 3-7 summarises this research's experimental design for the development of the CRISK prediction model. Data from FHS Offspring Cohort Exam 1 went through the Data Preparation step (section 3.8) to produce three datasets, mixed sex dataset, male dataset and female dataset, and their "SMOTEd" datasets prepared by applying the Synthetic Minority Over-sampling Technique (SMOTE) [174] to handle dataset imbalance issues. The reason for choosing SMOTE is given in section 3.8.10. After that, the developed CRISK prediction model was run against each dataset to produce corresponding prediction performance results (Chapter 6). These results were then analysed to determine which combination of predictors and number of nearest neighbours yielded the best prediction accuracy for each model. These results were also used to determine whether or not there should be separate prediction models for men and women.

There were a couple of reasons for experimenting with a mixed sex model, a male model, and a female model. In the FHS Offspring Cohort dataset, there are attributes applicable to women only, such as “ovaries removed”, “hysterectomy”, and “periods have stopped 1 year or more”. In addition, from the literature review, there were a number of sex specific models, for example the PROCAM model for men [67], the Reynolds risk score for women [70], and the Reynolds risk score for men [71]. And most importantly, experimentation on mixed sex, male, and female models would help to decide for this research if it is the best to have sex specific prediction models or just a mixed sex model.

Details of the experiments for each prediction model are explained in Chapter 6. In essence, there were two dimensions to the experiment: the number of predictors *n* and the number of nearest neighbours *k*. For the first dimension, a backward elimination technique was used to decrease the number of predictors by 1 each time, i.e. starting with *n*, then $n - 1$, and finally *1* predictor. The predictors were ranked in order of the most important to the least important. For the second dimension, the number of nearest neighbours *k* was trialled with odd numbers from 1 to 17. The main reason for choosing odd numbers was to avoid ties i.e. two class labels having the same number of votes. Later, from the experimentation results in Table 6-3, that k = 7, 7, and 15 respectively were chosen for the mixed sex model, the male model, and the female model indicated that it was not needed to trial with $k > 17$. If *k* had been 17 to yield best prediction performance for any of those three models, it should have been needed to trial with $k > 17$.

The Train-Test-LOOCV method was used for experimental validation in this research. LOOCV stands for leave-one-out-cross-validation. Each dataset was used as the test set while its “SMOTEd” dataset was used as the training set (the case base). Though there is no training step for a nearest neighbour algorithm, the “training set” terminology is used in this body of work to indicate the case base, not the test set. As each test set was a subset of the training set (“SMOTEd”), for each case from the test set, its instance presenting in the training set needed to be removed. Therefore, the validation method was named Train-Test-LOOCV in this research. More details of this method can be found in section 5.3.

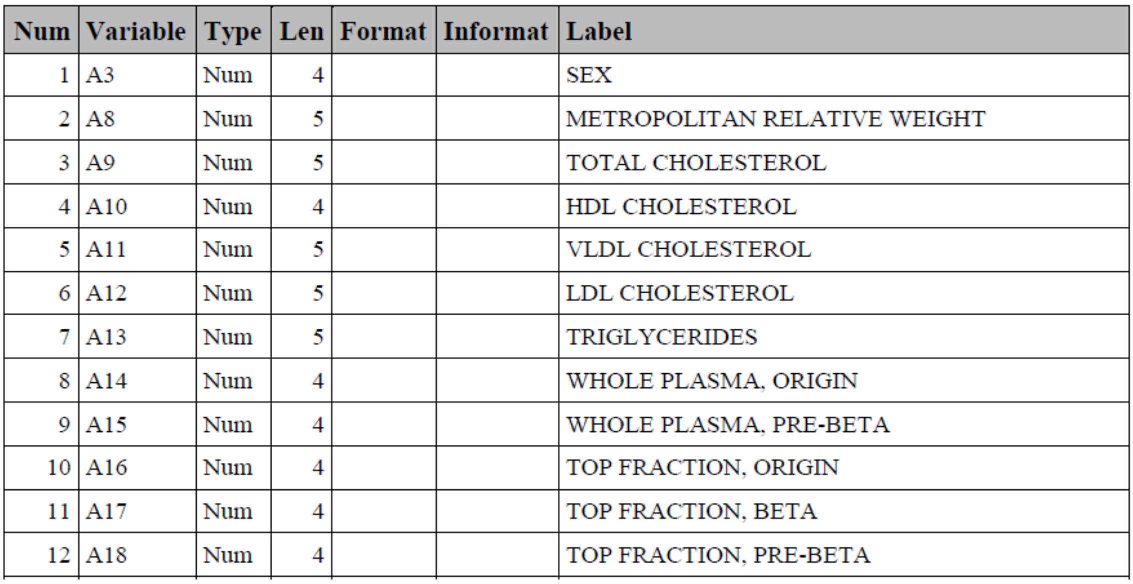
3.8 DATA PREPARATION

As detailed in the experimental design (section 3.7), data from the chosen FHS Offspring Cohort Exam 1 were used to prepare three datasets: a mixed sex dataset, a male dataset and a female dataset. For the mixed sex dataset and the male dataset, attributes that were present only for females were removed. For the male dataset, cases belonging to female participants were eliminated and vice versa. Other than these, the process to prepare data for each dataset was the same. Therefore, the following subsections only describe the steps used to prepare data for the mixed sex dataset. Data preparation for the other two datasets was undertaken in a similar manner and details can be found in Appendix F, Appendix G, Appendix H, Appendix I, Appendix L, Appendix M, Appendix N, and Appendix O.

3.8.1 Attribute Collection

First, from the FHS Offspring cohort raw data downloaded, attributes from multiple CSV files were combined to create an initial dataset containing 139 attributes. This was carried out based on the IDTYPE and PID columns in each individual CSV file. The IDTYPE column identified the cohort (e.g. the value 0 for the Original Cohort and the value 1 for the Offspring cohort). The PID column identified the participant within a cohort. Therefore, IDTYPE and PID together uniquely identified a participant in the FHS.

Next, collected attribute names were transformed into meaningful names. The reason for this was that the original attribute names were coded as e.g. “A3”, “A8”, “A9” etc. The transformation was done by referencing the “Data Dictionary.pdf” file enclosed in the downloaded raw data folder. The attribute name was replaced with the more meaningful corresponding label. Figure 3-8 displays a screenshot of the Data Dictionary file.



Num	Variable	Type	Len	Format	Informat	Label
1	A3	Num	4			SEX
2	A8	Num	5			METROPOLITAN RELATIVE WEIGHT
3	A9	Num	5			TOTAL CHOLESTEROL
4	A10	Num	4			HDL CHOLESTEROL
5	A11	Num	5			VLDL CHOLESTEROL
6	A12	Num	5			LDL CHOLESTEROL
7	A13	Num	5			TRIGLYCERIDES
8	A14	Num	4			WHOLE PLASMA, ORIGIN
9	A15	Num	4			WHOLE PLASMA, PRE-BETA
10	A16	Num	4			TOP FRACTION, ORIGIN
11	A17	Num	4			TOP FRACTION, BETA
12	A18	Num	4			TOP FRACTION, PRE-BETA

Figure 3-8: A screenshot of the Data Dictionary file

3.8.2 Invalid Case Removal

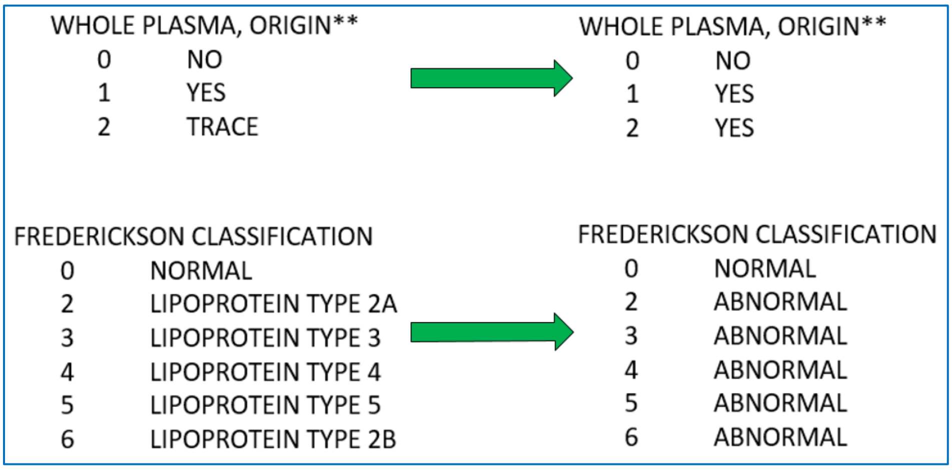
Next, as this research aimed to build a prediction model to predict an CVD event in future for people who are free of CVD, 115 cases having CVDDATE values of less than or equal to 0 (range from −992 to 0) were removed. The CVDDATE attribute recorded the date of CVD status as the number of days since Exam 1. After this removal, the dataset contained 5,013 − 115 = 4,898 cases.

3.8.3 Attribute Reduction

Next, data analysis was done to remove 16 columns, resulting in a dataset containing 123 attributes (ref. Appendix C). Four steps to eliminate these 16 attributes were performed. The first step was to remove the four columns with more than 80% of their data missing. This left 135 attributes with the highest percentage of missing data being 33.85%. Next, SBP and diastolic blood pressure (DBP) measured by nurses were excluded as the duplicate readings taken by physicians had fewer missing data. The third step was to eliminate the eight attributes belonging to females only. Finally, the "QUETELET INDEX", the "METROPOLITAN RELATIVE WEIGHT", and the "H.C.T" columns were removed. "QUETELET INDEX" is a duplicate of the "BMI" column. "METROPOLITAN RELATIVE WEIGHT" is so highly correlated with BMI that these measures of body fatness can be considered to be identical [175]. "H.C.T" is a duplicate of the HEMATOCRIT column.

3.8.4 Raw Data Value Transformation

Next, the collected raw data values were transformed into desired forms. Numeric values representing nominal values (e.g. 1 for Male and 2 for Female) were replaced by the actual nominal values (e.g. "Male" and "Female"). This was done by referencing the coding manuals supplied with the dataset package. For simplification, nominal attributes having more than two values (except "weight compared with 1 month ago" and "weight compared with 1 year ago") were converted into binary nominal values. Examples are shown in Figure 3-9. For "weight compared with 1 month ago" and "weight compared with 1 year ago" attributes, three nominal values of "about same", "5+ lbs lighter" and "5+ lbs heavier" were kept. The reason was that "lighter" and "heavier" go into two opposite directions from “about same” and therefore it would not be appropriate to turn these values into binary nominal values.



WHOLE PLASMA, ORIGIN**		WHOLE PLASMA, ORIGIN**	
0	NO	0	NO
1	YES	1	YES
2	TRACE	2	YES

FREDERICKSON CLASSIFICATION		FREDERICKSON CLASSIFICATION	
0	NORMAL	0	NORMAL
2	LIPOPROTEIN TYPE 2A	2	ABNORMAL
3	LIPOPROTEIN TYPE 3	3	ABNORMAL
4	LIPOPROTEIN TYPE 4	4	ABNORMAL
5	LIPOPROTEIN TYPE 5	5	ABNORMAL
6	LIPOPROTEIN TYPE 2B	6	ABNORMAL

Figure 3-9: Conversion of nominal values into binary nominal values

3.8.5 Prediction Attribute Preparation

Next, two prediction attributes, cvd10 and cvdInterval were created. Their values were derived from the two of the collected attributes namely CVD and CVDDATE. An intermediary attribute called CVDYear was also created. Eventually, the CVD, CVDDATE, and CVDYear columns were removed. Table 3-3 gives a description for each attribute and explains the logic to calculate the values for the two created prediction attributes, cvd10 and cvdInterval. The logic is based on the definitions of “High CVD Risk” and “Low CVD Risk” described in Equations (4) and (5), and their interpretation provided in section 3.4.2.

161 cases having CVD = “No” and follow-up time < 15 years were removed. After the removal, the dataset contained 4,898 − 161 = 4,737 cases.

Table 3-3: Prediction attribute preparation

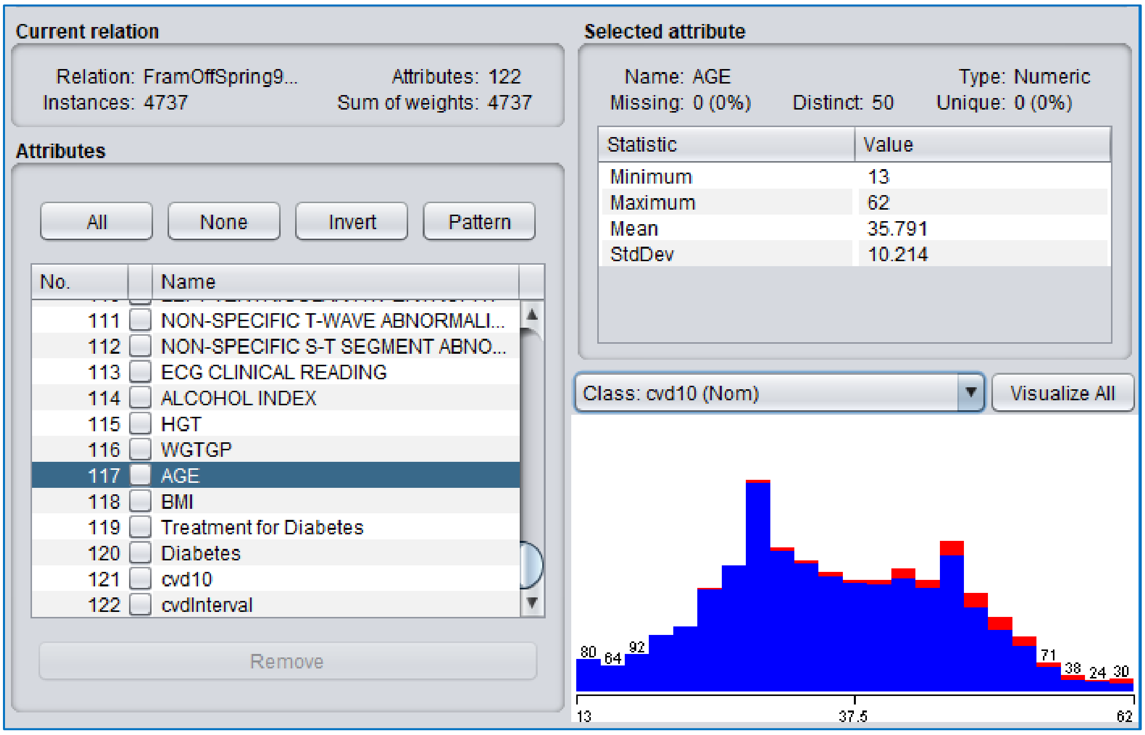
Attribute	Description	Logic for value calculation
CVD	Cardiovascular Disease (CVD) status (0 : No, 1 : Yes)	N/A. The values (0s and 1s) were already in the downloaded dataset, and were already transformed into No and Yes values in section 3.8.4.
CVDDATE	Date of CVD status (Number of days since Exam 1). This date corresponds to the date the participant had CVD or the date of censoring (last known date the participant did not have CVD). This is an integer number data type.	N/A. The values are already in the downloaded dataset.
CVDYear	Year of CVD status (Number of years since Exam 1). The value was calculated from the CVDDATE value and how many days in a year. This is a real number data type.	CVDYear = CVDDATE / 365.242199

cvd10	CVD status within 10 years since Exam 1 (No or Yes)	If CVD = “Yes” Then
		    If CVDYear ≤ 10 Then
		        cvd10 = “Yes”
		        cvdInterval = CVDYear
		    Else
		        cvd10 = “No”
		        cvdInterval = CVDYear
		    End If
		Else */* CVD = “No” */*
		    If CVDYear < 15 Then
		        Remove the case */* Follow-up less than 15 years */*
		    Else
		        cvd10 = “No”
		        cvdInterval = “” */* Don’t have CVD Interval” */*
		    End If
		End If

cvdInterval	Year of CVD status (Number of years since Exam 1). The value was decided from CVD, CVDYear, and cvd10. This is a real number data type. When CVD = “No”, CVDDATE is the last known date the participant did not have CVD. Therefore, in this situation, only keep the case if the follow-up time ≥ 15 years to ensure that the High CVD Risk membership and Low CVD Risk membership are known ($\mu_H = 0$ and $\mu_L = 1$). Otherwise, the memberhip values are unknown.	

3.8.6 Attribute Value Range Analysis

In the next phase of data preparation, attribute value range analysis was undertaken. The result was that all attribute value ranges were within acceptable ranges. This analysis was undertaken using Weka 3.8.3 to visually examine each attribute in the dataset along with their respective descriptive statistics (Figure 3-10). In addition, the value ranges were also checked against the FHS Offspring Cohort Exam 1 Coding Manual (enclosed in the dataset package obtained). This analysis confirmed that the data was ready for the next step, feature selection.



Current relation

Relation: FramOffSpring9... Attributes: 122
Instances: 4737 Sum of weights: 4737

Attributes

No.		Name
111	□	NON-SPECIFIC T-WAVE ABNORMALI...
112	□	NON-SPECIFIC S-T SEGMENT ABNO...
113	□	ECG CLINICAL READING
114	□	ALCOHOL INDEX
115	□	HGT
116	□	WGTGP
117	□	AGE
118	□	BMI
119	□	Treatment for Diabetes
120	□	Diabetes
121	□	cvd10
122	□	cvdInterval

Selected attribute

Name: AGE Type: Numeric
Missing: 0 (0%) Distinct: 50 Unique: 0 (0%)

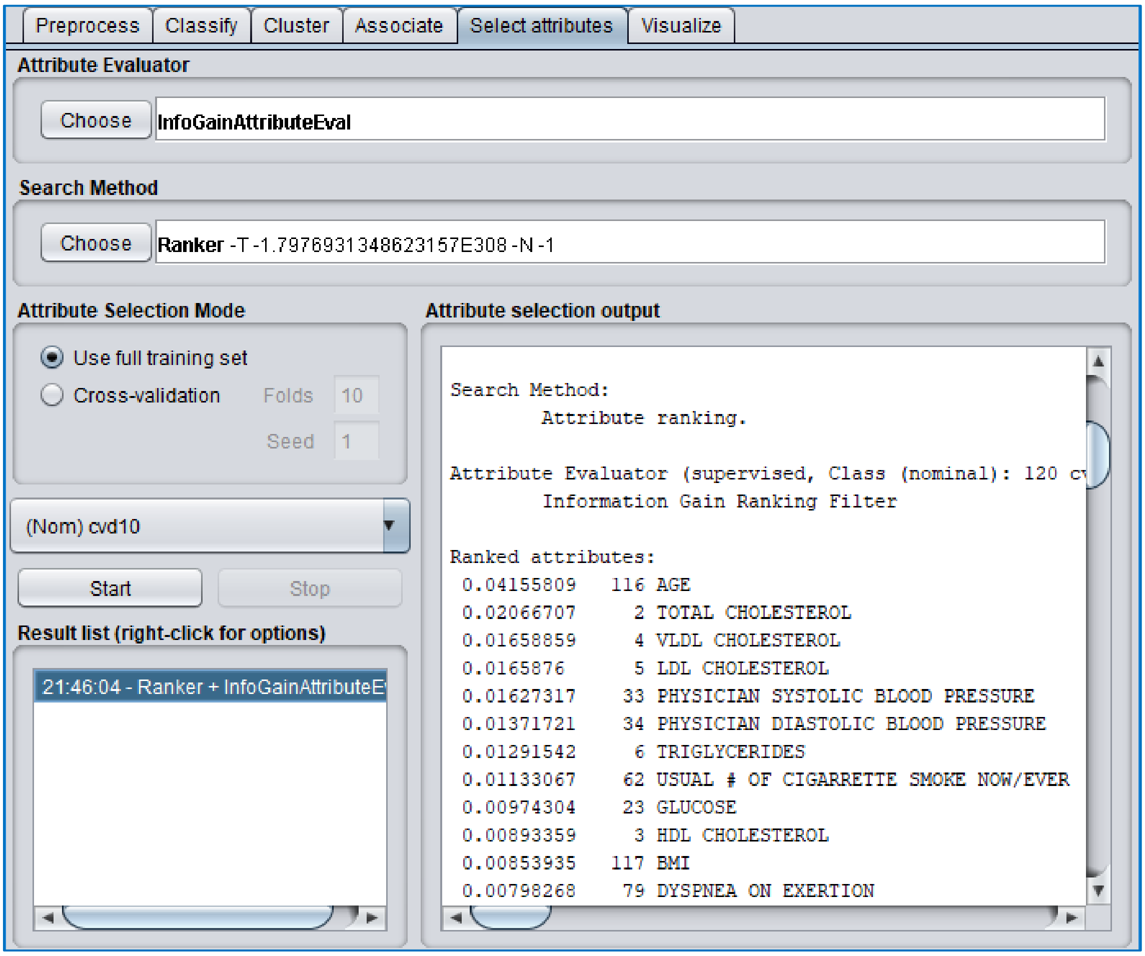
Statistic	Value
Minimum	13
Maximum	62
Mean	35.791
StdDev	10.214

Class: cvd10 (Nom)

Figure 3-10: Attribute value range analysis in Weka

3.8.7 Feature Selection

Initially, 34 predictor attributes (risk factors) were selected. This was done by using Weka's InfoGainAttributeEval attribute evaluator with the Ranker search method (Figure 3-11) to rank the 119 predictor attributes in the prepared dataset. InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class (cvd10). The highest ranked attribute was "age" with an information gain value of 0.041558. The cut-off value was chosen to be 0.004, which was about one tenth of this highest information gain value. This means that attributes having info gain values less than 0.004 were eliminated. Later, the result of using only 13 predictors for the CRISK model from section 6.4 proves that the cut-off value was chosen adequately, i.e. small enough to not missing out predictors. Appendix D contains the attribute selection output resulting from Weka.



Attribute Evaluator

InfoGainAttributeEval

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) cvd10

Result list (right-click for options)

21:46:04 - Ranker + InfoGainAttributeE

Attribute selection output

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 120 cv
        Information Gain Ranking Filter

Ranked attributes:
 0.04155809   116 AGE
 0.02066707     2 TOTAL CHOLESTEROL
 0.01658859     4 VLDL CHOLESTEROL
 0.0165876      5 LDL CHOLESTEROL
 0.01627317    33 PHYSICIAN SYSTOLIC BLOOD PRESSURE
 0.01371721    34 PHYSICIAN DIASTOLIC BLOOD PRESSURE
 0.01291542     6 TRIGLYCERIDES
 0.01133067    62 USUAL # OF CIGARRETTE SMOKE NOW/EVER
 0.00974304    23 GLUCOSE
 0.00893359     3 HDL CHOLESTEROL
 0.00853935   117 BMI
 0.00798268    79 DYSPNEA ON EXERTION
```

Figure 3-11: Feature selection using Weka

Table 3-4 displays 34 selected features and their Weka computed information gain rankings in descending numerical order.

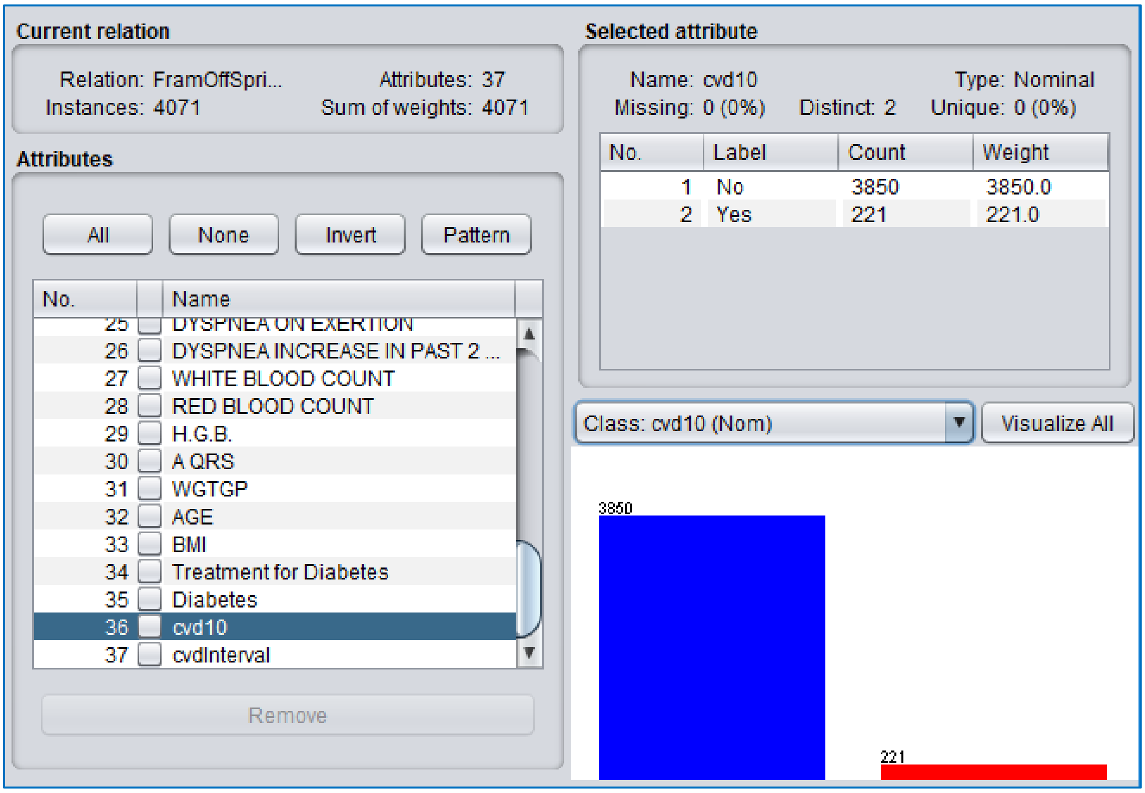
Table 3-4: Selected features and their information gain rakings

Attribute	Information gain
AGE	0.04155809
TOTAL CHOLESTEROL	0.02066707
VLDL CHOLESTEROL	0.01658859
LDL CHOLESTEROL	0.0165876
SYSTOLIC BLOOD PRESSURE	0.01627317
DIASTOLIC BLOOD PRESSURE	0.01371721

TRIGLYCERIDES	0.01291542
USUAL # OF CIGARETTES SMOKE NOW/EVER	0.01133067
GLUCOSE	0.00974304
HDL CHOLESTEROL	0.00893359
BMI	0.00853935
DYSPNEA ON EXERTION	0.00798268
SEX	0.00758577
LDH	0.00752456
ALKALINE PHOSPHOTASE	0.00731425
URIC ACID	0.00697716
WGTGP	0.00626734
HISTORY OF HYPERTENSION	0.00617897
SMOKED AT LEAST 1 YEAR	0.00606969
HEMATOCRIT	0.00593551
WHITE BLOOD COUNT	0.0057986
FREDERICKSON CLASSIFICATION	0.00570935
Diabetes	0.00566588
DYSPNEA INCREASE IN PAST 2 YEARS	0.00564343
H.G.B.	0.00560985
TOP FRACTION PRE-BETA	0.00547475
RED BLOOD COUNT	0.005308
A QRS	0.00528462
SMOKES CIGARETTES	0.00519749
HYPOGLYCEMIC AGENTS	0.00489626
Treatment for Diabetes	0.00489626
PRE-BETA BAND	0.00416658
HYPOTENSIVES	0.0041151
WHOLE PLASMA PRE-BETA	0.00406476

3.8.8 Missing Data Removal

After retaining only attributes selected via Weka's InfoGainAttributeEval method, 666 cases having missing data were removed. This left a dataset of 4,071 cases. Of these, 221 cases have cvd10 = "Yes" and 3,850 cases have cvd10 = "No" (Figure 3-12). Among those 3,850 cases having cvd10 = "No", 2,950 cases do not have CVD Interval.



Current relation

Relation: FramOffSpri... Attributes: 37
Instances: 4071 Sum of weights: 4071

Attributes

No.	Name
25	DYSPNEA ON EXERTION
26	DYSPNEA INCREASE IN PAST 2 ...
27	WHITE BLOOD COUNT
28	RED BLOOD COUNT
29	H.G.B.
30	A QRS
31	WGTGP
32	AGE
33	BMI
34	Treatment for Diabetes
35	Diabetes
36	cvd10
37	cvdInterval

Selected attribute

Name: cvd10 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	No	3850	3850.0
2	Yes	221	221.0

Class: cvd10 (Nom)

Figure 3-12: Data distribution of the cvd10 attribute (the class attribute) visualised in Weka

This posed the issue of dataset imbalance for the cvd10 class where the classification categories are not equally represented [176]. If no action was taken to address this issue, classification performance would be affected. When dataset imbalance happens, classifiers tend to have good accuracy on the majority class but very poor accuracy on the minority class [177].

3.8.9 Feature Ranking

After missing data were removed, the selected attributes were ranked again using Weka (Appendix E). This produced more accurate rankings of the predictors than that of the original rankings (Table 3-4) which was undertaken when there was still missing data in the dataset. Table 3-5 provides these final rankings of the selected attributes, which were used as input for the backward elimination experimentation that is described in detail in Chapter 6.

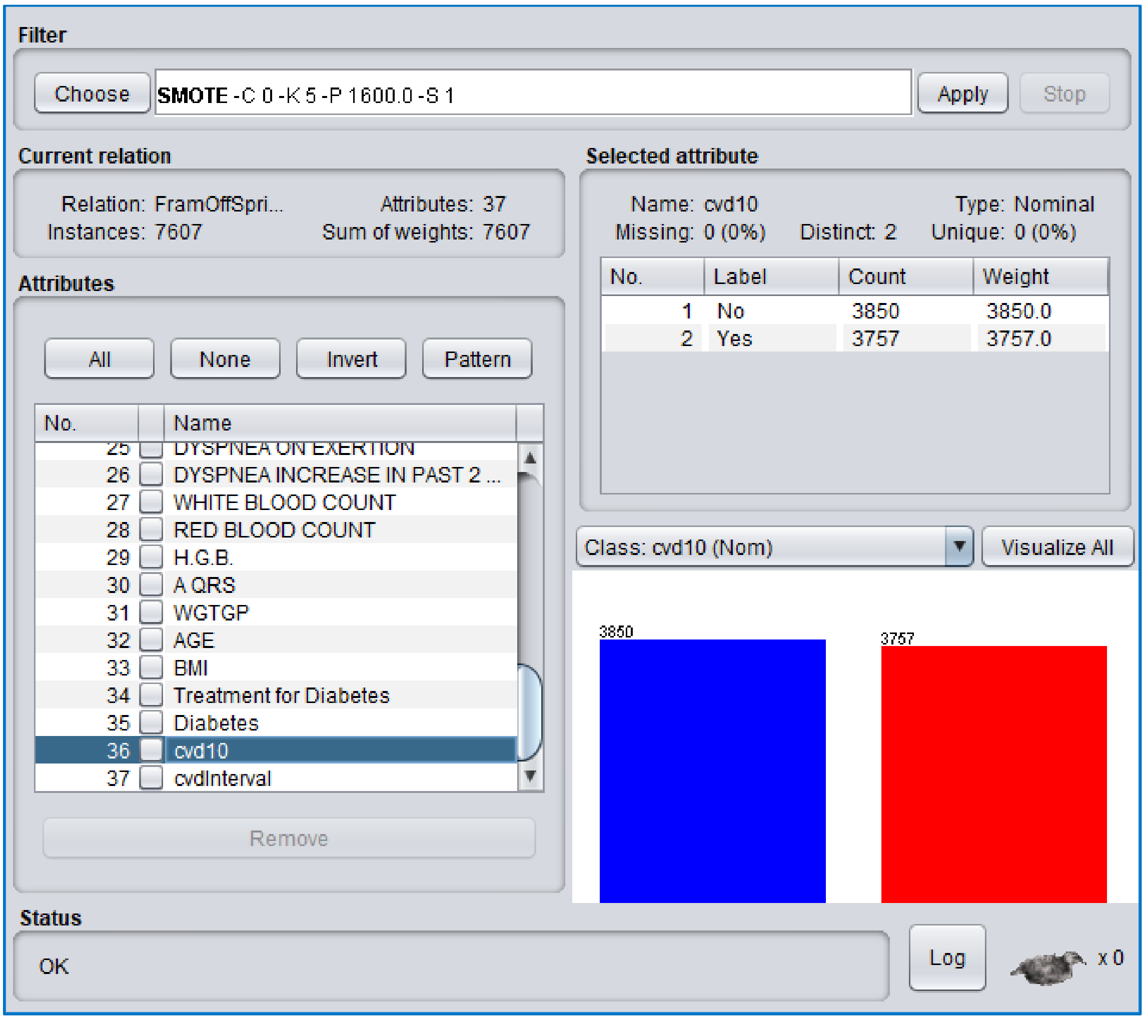
Table 3-5: Selected features and their final information gain rakings (done after missing data removal)

Attribute	Information gain
AGE	0.04384
TOTAL CHOLESTEROL	0.02324
LDL CHOLESTEROL	0.0184
VLDL CHOLESTEROL	0.01733
SYSTOLIC BLOOD PRESSURE	0.01562
TRIGLYCERIDES	0.01392
DIASTOLIC BLOOD PRESSURE	0.01286
GLUCOSE	0.0119
USUAL # OF CIGARETTES SMOKE NOW/EVER	0.01112
HDL CHOLESTEROL	0.01109
HEMATOCRIT	0.00915
BMI	0.00902
LDH	0.00843
SEX	0.00805
WGTGP	0.00741
URIC ACID	0.00736
FREDERICKSON CLASSIFICATION	0.00735
H.G.B.	0.00725
ALKALINE PHOSPHOTASE	0.00719
WHITE BLOOD COUNT	0.00678
DYSPNEA ON EXERTION	0.00673
Diabetes	0.00672
TOP FRACTION PRE-BETA	0.0067
RED BLOOD COUNT	0.00652
SMOKED AT LEAST 1 YEAR	0.00616
Treatment for Diabetes	0.00578
HYPOGLYCEMIC AGENTS	0.00578
A QRS	0.00569
HISTORY OF HYPERTENSION	0.00568
PRE-BETA BAND	0.00543
WHOLE PLASMA PRE-BETA	0.00535
DYSPNEA INCREASE IN PAST 2 YEARS	0.00506
SMOKES CIGARETTES	0.00503

HYPOTENSIVES	0.00445

3.8.10 Imbalanced Dataset Handling using SMOTE

In order to address the dataset imbalance issue (section 3.8.8), the Synthetic Minority Over-sampling Technique (SMOTE) [174], available in Weka, was used. To balance unbalanced datasets, resample techniques (oversampling and/or undersampling) are used. For this research's data distribution (Figure 3-12), oversampling rather than undersampling methods should be chosen because of not having many minority class samples (only 221 positive cases). In addition, undersampling may remove useful samples for building the CRISK model [178]. For oversampling, SMOTE is the most popular oversampling method [179] and has proven successful in variety of applications and domains [178].



SMOTE -C 0 -K 5 -P 1600.0 -S 1

Relation: FramOffSpri... Attributes: 37
Instances: 7607 Sum of weights: 7607

Name: cvd10 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	No	3850	3850.0
2	Yes	3757	3757.0

No.	Name
25	DYSPNEA ON EXERTION
26	DYSPNEA INCREASE IN PAST 2 ...
27	WHITE BLOOD COUNT
28	RED BLOOD COUNT
29	H.G.B.
30	A QRS
31	WGTGP
32	AGE
33	BMI
34	Treatment for Diabetes
35	Diabetes
36	cvd10
37	cvdInterval

Class: cvd10 (Nom)

OK

Figure 3-13: Applying SMOTE in Weka to balance the dataset

Based on the number of positives and negatives (Figure 3-12), the percentage parameter was set to 1,600. This setting helped produce a balanced dataset of 3,850 cases of “No” and 3,757 cases of “Yes” after applying the SMOTE technique (Figure 3-13). Among the 3,850 cases of “No”, 2,950 cases do not have CVD Interval.

An ideal percentage value of 1,642.1 instead of 1,600 could have been calculated and set for SMOTE to result in having the number of positives equals the number of negatives. However, this ideal balance would not be how the model would operate in reality. In reality, it would not practical to always keep the number of positives equal the number of negatives. Therefore, the percentage value of 1,600 was kept and used.

3.8.11 Input file preparation for experimentation

Finally, four input files in CSV format were created for input into CRISK. They were a training dataset file, a test dataset file, a predictors file, and a predictors ranking file. An explanation of these input files is given in Table 3-6.

Table 3-6: Prepared input files for experimentation

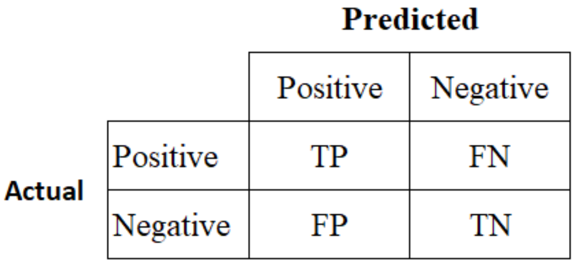
File name	Purpose	Explanation
FramOffspring_SMOTE.csv	To be used as the training dataset	Resulting from applying SMOTE in section 3.8.10
FramOffSpring.csv	To be used as the test dataset	Resulting from missing data removal in section 3.8.8. This was the actual dataset with the real data (imbalanced dataset), before applying SMOTE.
predictors.csv	To describe predictors (risk factors)	The predictors file is shown in Appendix J. Detailed explanation of a predictors file is found in section 5.2.
predictorsRanking.csv	To rank the predictors	Resulting from feature ranking in section 3.8.9. The predictors ranking file is displayed in Appendix K.

3.9 CRISK Model Evaluation Protocol

3.9.1 Evaluation Metrics

Evaluation metrics to measure the performance of the developed CRISK prediction model in this research were chosen with consideration of dataset imbalance. The prepared FHS Offspring Cohort dataset has only about 5% of True Positive (cvd10 = Yes) cases. For a highly imbalanced dataset like this one, the overall accuracy as in Equation (6) cannot be used as an evaluation metric because if a model just predicted all cases to be negative (cvd10 = No), it would achieve about 95% accuracy. Therefore, evaluation metrics other than the overall accuracy should be considered.

The decision made by binary classification can be represented as a 2 × 2 confusion matrix in Figure 3-14 [180]. The matrix has four outcomes. True positives (TP) are positive cases correctly predicted as positive. False negatives (FN) are positive cases incorrectly predicted as negative. False positives (FP) are negative cases incorrectly predicted as positive. And, true negatives (TN) are negative cases correctly predicted as negative.



		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Figure 3-14: Confusion matrix

A number of popular evaluation metrics are derived from the above confusion matrix [180-182]. These include Accuracy, True Positive Rate (TPR), True Negative Rate (TNR), Precision, F-value, and Negative Predictive Value (NPV). These metrics are described by Equations (6) to (11).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

$$TPR = Recall = Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{7}$$

$$TNR = Specificity = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F - value = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{10}$$

$$NPV = \frac{TN}{TN + FN} \tag{11}$$

With the exception of Accuracy, these metrics were used to evaluate the performance of the CRISK prediction model developed in this research. Among them, F-value (the harmonic mean of recall and precision) may not be entirely straight-forward to interpret. However, the F-value statistic is a popular evaluation metric for imbalanced datasets [181]. Usually, the value of β (see Equation (10)) is set to 1 [181]. In this research, β was also set to 1 and thus the F-value became F_1-value whose formula is described in Equation (12).

$$F_1 - value = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{12}$$

When coming to decision making, as to what combination of predictors *n* and number of nearest neighbours *k* yielded the best prediction performance, the F_1-value was favoured over the other four evaluation metrics. The reason was that the F_1-value is a harmonic mean of Recall and Precision. While Recall tells us about the ability of a model to pick how many positive cases out of all actual positive cases, Precision tells us about the ability of a model to pick only relevant cases, meaning how many cases from the ones predicted to be positive are actually positive. In CVD prediction, if a “High Risk” person is predicted to be “Low Risk” because of low Recall, that person may miss out on medical treatment and attention. On the other hand, if a “Low Risk” person is predicted to be “High Risk” because of low Precision, unnecessary medical treatment may occur.

However, when two models produce the same or very similar F_1-values, Recall should be considered with a higher priority than Precision. In CVD prediction, the consequence of missing out a positive case would be a lot worse than wrongly classifying a negative case as positive. A FN case may not be given further examination, treatment, or attention and therefore may lead to the worst scenario—death. On the other hand, unnecessary medical treatment and attention for a FP case could just cost money and time. Moreover, these costs might not be entirely wasted as, for example, a FP patient may gain some health benefits from being advised to consume a healthy diet in order to reduce CVD risk.

This research did not use AUC as an evaluation metric. AUC is abbreviated from Area Under the Receiver Operating Characteristic (ROC) Curve [183]. It is used as a performance metric for a number of existing models such as D'Agostino et al. [31], PROCAM [67], and SCORE [68]. The ROC curve is a two-dimensional graph of False Positive Rate (FPR = 1 − TNR) on the X axis and TPR on the Y axis. Each point on the ROC curve corresponds to one decision threshold set for the algorithm used by the model. For example, in case of a KNN algorithm, the decision threshold (the number of votes) can be varied from 0 to *k* to produce the ROC curve. As a result, AUC does not represent the model operating in the most suitable selected threshold but summarises the performance of the model over regions of ROC space in which one would rarely operate [184]. In this research, the decision threshold was set to be the majority vote. The number of risk factors *n* and the number of nearest neighbours *k* were varied instead. Each combination of *n* and *k* produces one single point of FPR and TPR. TPR, TNR, Precision, F_1-value, and NPV were evaluated rather than producing ROCs.

Though the prediction outcomes include CVD Interval, widely used metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Correlation Coefficient to measure the error of a model in predicting quantitative data were decided not to be used as performance metrics in this research. The reason is twofold. First, there are many cases whose CVD Interval and/or predicted CVD Interval are unknown. Second, the focus of this research, at the current stage, is to correctly predict whether a person will have CVD within 10 years, not on checking the accuracy of the predicted CVD Interval. However, RMSE values are reported in Table 6-5.

3.9.2 External Validation

In this research, external validation was planned and conducted to answer RQ4. External validation means testing a developed model with different cohorts to the one that was used to build the model [108]. Such validation is critical as a model may perform well for a certain cohort but may overestimate or underestimate when applied to other cohorts, especially cohorts of different racial groups [87]. The fact that numerous models have been published without adequate external validation was highly criticised in the review article by Damen et al. [6]. In this study, the FHS Original Cohort Exam 11 dataset was chosen as a test set for the external validation process. The reasons for choosing this dataset were explained in section 3.6. Details of the external validation results are presented in Chapter 7.

3.9.3 Comparison to existing models

For comparison to existing models, three models were chosen: the D'Agostino et al. [31] model, the PREDICT-1° [76] model, and the 2018 PCE [77] model. Though two models were developed in the D'Agostino et al. [31] study, and both of them are currently used on the FHS website for 10-year CVD risk prediction, only the first model was chosen as it performed better than its simpler version (the second model). PREDICT-1° was recently published and was reported to perform better than the 2013 PCE [74] model—a well-known alternative to conventional Framingham models from the American College of Cardiology/American Heart Association. The 2018 PCE model was published after PREDICT-1° and claimed a significant improvement in terms of accuracy in prediction of CVD. Details of comparison of CRISK to these three existing models are provided in section 8.3. These details help answer RQ6.

3.10 CHAPTER SUMMARY

This Research Methodology chapter described how this research was approached, designed, implemented, and evaluated. Design Science was the research methodology for this study. The methodological process was guided by the beliefs of the positivist paradigm. The chosen methodology was equipped with a conceptual framework in Information Research, tailored for this specific study in CVD prediction. In addition, the study was also assisted by a set of guidelines.

The development of the CRISK prediction model, which was the artifact resulting from carrying out this Design Science research to solve the CVD prediction problem, was strategically planned and designed. The prediction model was decided to be a CBR system whose case base is a fuzzy ontology. The CRISK system was decided to be implemented from scratch as none of the known existing CBR tools supports fuzzy ontology. The prediction outcomes from the CRISK system were designed to be represented as fuzzy membership values of “High CVD Risk” and “Low CVD Risk” fuzzy sets.

To have data for the CRISK prediction model development and evaluation, FHS cohorts were obtained and this process was detailed in this chapter. The FHS Offspring Cohort Exam 1 dataset was selected to be used for the development of the prediction model. FHS Original Cohort Exam 11 dataset was determined to be used as the dataset for external validation. SMOTE was used to address the dataset imbalance issue.

For evaluation of the developed CRISK prediction model, a protocol consisting of evaluation metrics, external validation, and comparison to existing models was formed. The evaluation metrics to assess prediction performance included TPR, TNR, Precision, F_1-value, and NPV. The developed model was then compared to the D’Agostino et al. [31] model, the PREDICT-1° [76] model, and the 2018 PCE [77] model (section 8.3).

Chapter 4

CRISK PREDICTION MODEL

4.1 INTRODUCTION

This chapter explains the CRISK prediction model. It first provides an account of the design of the model including input, four CBR activities, the case base, and output. The design shows how information flows between these components of the model. Each component of the model is explained in detail with a focus on the algorithms used in each CBR activity. These algorithms are described using pseudo code. From the information provided in this chapter, this or a similar CVD risk prediction system can be implemented.

This chapter together with Chapter 5 and Chapter 6 help answer the first three research questions (RQ1, RQ2, and RQ3). The answers are described in section 9.1.1.

4.2 CRISK PREDICTION MODEL DESIGN

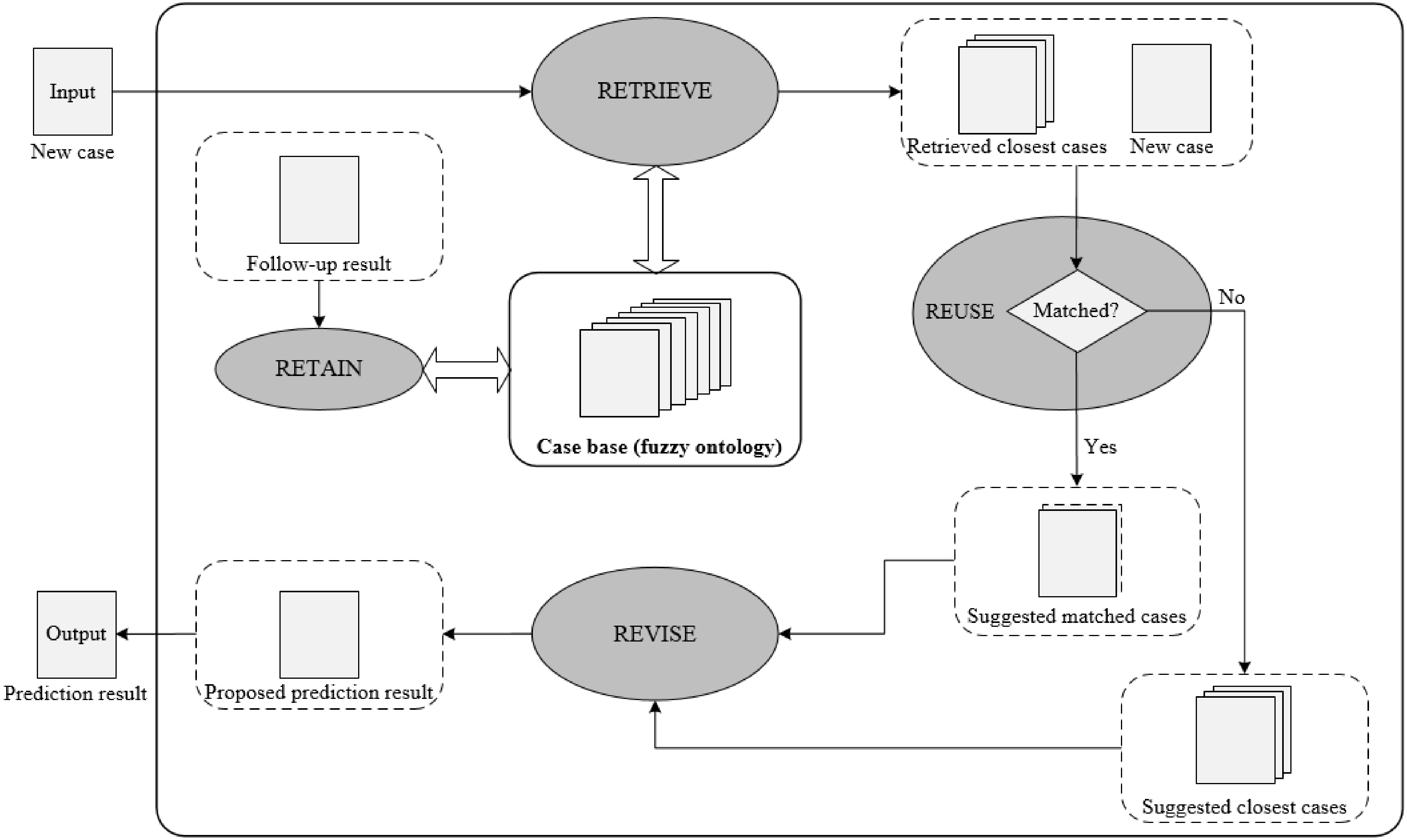


Figure 4-1: CRISK Prediction Model

Figure 4-1 shows an overview of the CRISK model that was designed as a CBR system using fuzzy ontology. Basically, it consists of a case base and four CBR activities, Retrieve, Reuse, Revise, and Retain. Existing cases are stored in the case base, which is a fuzzy ontology. Other than those main components, a prediction result (output) is generated by the model for each new case (input). The prediction process is explained as below:

1. A **new case** is input for CVD prediction.
2. The **Retrieve** algorithm (developed based on the fuzzy KNN algorithm [171]) queries the **Case Base** to retrieve *k* closest cases to the new case (in this research, *k* is decided to be 7 from the experimentation results in Chapter 6).
3. The **Retrieve** algorithm also identifies *h* matched cases to the new case from the *k* closest cases retrieved.
4. *k* closest cases, including *h* matched cases identified, are passed from the **Retrieve** algorithm to the **Reuse** algorithm.
5. The **Reuse** algorithm suggests the *h* matched cases if $h > 0$, otherwise the *k* closest cases, to the **Revise** algorithm.
6. If there are matched cases ($h > 0$), the **Revise** algorithm calculates the prediction outcomes (risk class, “High CVD Risk” membership, “Low CVD Risk” membership, predicted CVD Interval) from these match cases. This includes handling when there is a tie and when CVD Interval cannot be decided as a single value e.g. 25 years.
7. If there is no matched case ($h = 0$), the **Revise** algorithm calculates the prediction outcomes as follows. First, it calculates the “High CVD Risk” membership μ_H and the “Low CVD Risk” membership μ_L for the new case using Equation (14) in section 4.7. Then, it decides the predicted risk class for the new case based on these membership values. After that, it calculates the predicted CVD Interval using EITHER the membership functions declared in Equations (4) and (5) when μ_H and μ_L are less than 1 OR averaging CVD Intervals of the nearest neighbours when μ_H or μ_L equals 1.
8. The **Revise** algorithm proposes the prediction outcomes (prediction result) for the new case.

When there is a need to conclude whether the case develops CVD within 10 years, e.g. for measuring prediction accuracy of the developed model using a confusion matrix, the defuzzification process is as follows:

If the "High CVD Risk" membership value μ_H is greater than or equal to the "Low CVD Risk" membership value μ_L, then the case is considered to belong to "High CVD Risk" and it is predicted that CVD will develop within 10 years.

Otherwise, the case is considered to belong to "Low CVD Risk" and it is predicted that the person will not develop CVD within 10 years.

Details of each component of the model, including the Retrieve, Reuse, and Revise algorithms, are described in the next sections in this chapter. The Retain activity is out of scope of this research (explained in section 4.8).

4.3 CASE BASE

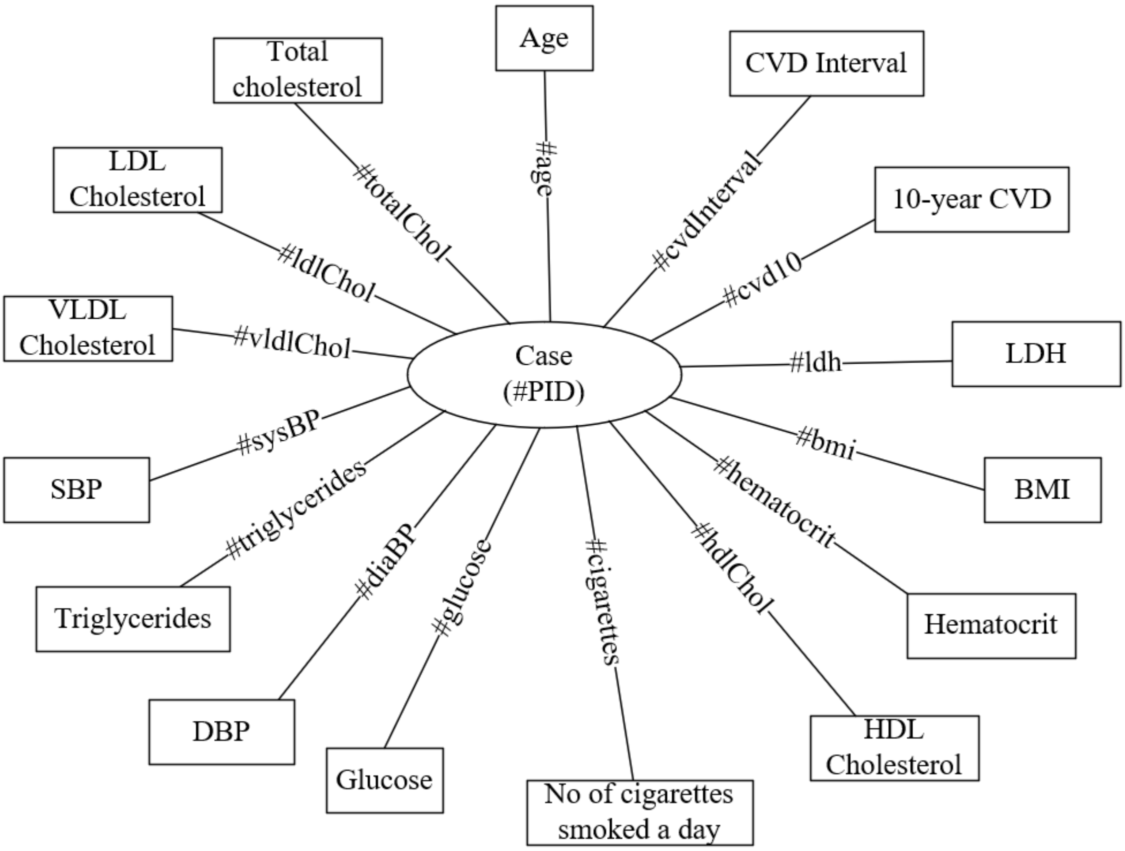


Figure 4-2: Illustration of a case stored in the CRISK ontology

The case base consists of existing CVD cases that already have follow-up results. It is a fuzzy ontology (called the CRISK ontology in this body of work). Figure 4-2 illustrates a case stored in the case base. The case contains 13 risk factors (age, total cholesterol, LDL cholesterol, very-low-density lipoprotein (VLDL) cholesterol, SBP, triglycerides, DBP, glucose, number of cigarettes smoked a day, HDL cholesterol, hematocrit, BMI, and LDH) and two CVD outcomes ("10-year CVD" and "CVD Interval"). The 13 risk factors were selected to be the predictor attributes for CVD prediction based on the results of experiments detailed in Chapter 6.

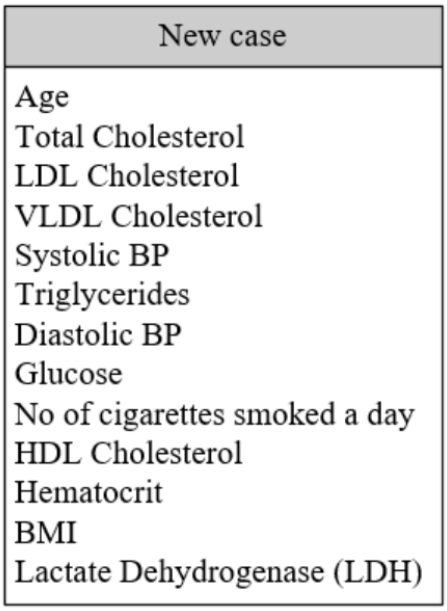
How a CVD case is stored in the case base is further explained. A case is stored in the case base as an individual, uniquely identified by the case ID (PID). A case's attribute (e.g. "age") is represented as a data property (e.g. "#age"). The value of a case's attribute is represented as a literal (represented as a rectangular box in Figure 4-2). A literal can be either a crisp value (e.g. 37) or a membership value of a fuzzy set (e.g. 0.3 of the "young" set). In the case of a membership value of a fuzzy set, the membership function of the fuzzy set must be defined.

Among a case's attributes, "10-year CVD" and "CVD Interval" are prediction attributes (outcomes). "10-year CVD" is a binary value of either "Yes" or "No" indicating whether a CVD event happens within 10 years since the examination. "CVD Interval" is the number of years since the examination that a CVD event happens. Values of "10-year CVD" and "CVD Interval" are calculated using the rules defined in Table 3-3.

The "10-year CVD" attribute that is represented as a binary value of either "Yes" or "No" is not directly used to present prediction outcomes; instead, its fuzzy membership values of "High CVD Risk" and "Low CVD Risk" fuzzy sets are. These fuzzy concepts are declared in the CRISK ontology (the case base) using Protégé with the Fuzzy OWL 2 plugin [141]. Their membership functions are described in Equations (4) and (5), and illustrated by Figure 3-3 in section 3.4.2. Conversions between crisp values of CVD prediction outcomes and fuzzy membership values are explained in the Revise algorithm (section 4.7).

Though a case in the case base is capable of storing fuzzy membership values, at the current stage only crisp values are used in this research. In this research, predictor attributes are not fuzzified. The CVD outcomes ("10-year CVD" and "CVD Interval") are also stored in the case base as crisp values, although they are represented as fuzzy membership values of "High CVD Risk" and "Low CVD Risk" when displaying on the result screen for users or when involving in the CBR activities' algorithms. The algorithms can read the fuzzy concepts declared in the case base and perform the conversion between crisp and fuzzy membership values instead of directly keeping the fuzzy membership values in the case base. However, directly storing fuzzy membership values in the case base is another way and it will achieve the same results.

4.4 INPUT (NEW CASE)



New case
Age Total Cholesterol LDL Cholesterol VLDL Cholesterol Systolic BP Triglycerides Diastolic BP Glucose No of cigarettes smoked a day HDL Cholesterol Hematocrit BMI Lactate Dehydrogenase (LDH)

Figure 4-3: Illustration of an input (new case)

Figure 4-3 illustrates an input example. The new case contains the 13 risk factors: age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, number of cigarettes smoked a day, HDL cholesterol, hematocrit, BMI, and LDH. These risk factors are used to predict CVD for the case. If some of these risk factors are missing from the input case, the CRISK prediction system still provides prediction results. However, the prediction accuracy may be degraded when a case is missing some of the input variables.

4.5 RETRIEVE

The main purpose of the Retrieve activity is to retrieve *k* closest cases to the input case from the case base. The other purpose is to find which cases match with the input case from the list of *k* closest cases retrieved. To do these, a Retrieve algorithm was developed in this research based on the fuzzy KNN algorithm [171].

Let $C = \{c_1, c_2, c_3, \ldots, c_n\}$ be the case base containing *n* cases $c_1, c_2, c_3, \ldots, c_n$. Let *c* be the new case whose CVD risk is to be predicted. Let *L* be the list to contain *k* nearest cases (from *C*) to *c*. Let *M* be the list of *h* ($0 \le h \le k$) cases (from *C*) matched with *c*. The Retrieve algorithm is defined as follows:

```
Retrieve algorithm
/* Get k closest cases from the case base */
For each $c_i$ in C
    Calculate distance $d_i$ from $c_i$ to c
    If L has fewer than k elements
        Add $c_i$ to L
    Else
        Get the last element $l_{k-1}$ of L
        Get/calculate distance $d_l$ from $l_{k-1}$ to c
        If $d_i < d_l$
            Remove $l_{k-1}$ from L
            Add $c_i$ to L
        End If
    End If
    Sort L ascendingly
End of For loop

/* Get h matched cases from the list of k closest cases */
For each element $l_i$ in L
    Get/calculate distance $d_{li}$ from $l_i$ to c
    If $d_{li} = 0$
        Add $l_i$ to M
    End If
End of For loop
```

The Retrieve algorithm has two *for* loops for the two purposes mentioned above. First, it iterates through the case base *C* to find *k* closest cases to the input case and add these *k* closest cases into the list *L*. Second, it iterates through the list *L* of *k* closest cases to find *h* ($0 \leq h \leq k$) matched cases with the input case and add these *h* matched cases into the list *M*.

To support the Retrieve algorithm, the Distance algorithm was developed to calculate the distance between a CVD case in the case base and the input case from their risk factors' values. Let *c* be the case from the case base, c_{input} be the input case, and *d* be the distance between them that is needed to be measured. The Distance algorithm is defined as follows:

Distance algorithm

```
Intitialise d = 0.0
For each risk factor $r_i$ in $c_{input}$
    If c has $r_i$
        Initialise diff = 0.0
        If $r_i$ is nominal data
            If the value of $r_i$ of $c_{input}$ equals the value of $r_i$ of c
                Set diff = 0.0
            Else
                Set diff = 1.0
            End If
        Else /* Numeric data */
            Normalise the value of $r_i$ of $c_{input}$ to be $v_{input}$
            Normalise the value of $r_i$ of c to be v
            Set diff = $v_{input} - v$
        End If
        Set d = d + diff×diff
    End If
End of For loop
d = sqrt(d) /* square root of d */
```

The normalisation follows the min-max rescaling method to rescale a numeric value into the range [0, 1]. The normalisation formula is given in Equation (13) below:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (13)$$

where *x* is the original value and *x'* is the normalised value.

4.6 REUSE

The purpose of the Reuse activity is to suggest either matched cases or closest cases to the input case. As the Retrieve algorithm in section 4.5 already gets *k* closest cases and *h* matched cases, the Reuse algorithm is defined as below:

```
Reuse algorithm
If h > 0      /* There are matched cases */
    Suggest the h matched cases
Else          /* There is no matched case */
    Suggest the k closest cases
End If
```

4.7 REVISE

The Revise activity proposes CVD prediction results for the new input case from the suggested cases output from the Reuse activity. Let *h* be the number of matched cases to the input case from these suggested cases. The Revise algorithm is defined as below:

```
Revise algorithm
If h > 0      /* There are matched cases */
    Use "Revise Matched Cases" algorithm
Else          /* There is no matched case */
    Use "Revise Closest Cases" algorithm
End If
```

Depending on whether there are matched cases or not, the Revise algorithm uses either the "Revise Matched Cases" algorithm or the "Revise Closest Cases" algorithm to generate CVD prediction results for the new input case.

Let *M* be the list of matched cases, *pClass* and *pCVDInterval* respectively be the predicted CVD Class and the predicted CVD Interval for the new input case, μ_H and μ_L respectively be the predicted "High CVD Risk" membership and the predicted "Low CVD Risk" membership for the new input case. The "Revise Matched Cases" algorithm is defined as follows:

Revise Matched Cases algorithm

```
Initialise $count_H$, $count_L = 0$
Initialise $sum_H$, $sum_L = 0.0$
Initialise noCVDIntervalFlag = false, machineDecidableFlag = true,
cvdIntervalPredictableFlag = true
For each case m in M
    If cvd10 of m = Yes  /* High CVD Risk */
        Set $count_H = count_H + 1$
        Set $sum_H = sum_H$ + cvdInterval
    Else   /* Low CVD Risk */
        Set $count_L = count_L + 1$
        If m has CVD Interval
            Set $sum_L = sum_L$ + cvdInterval
        Else
            Set noCVDIntervalFlag = true
        End If
    End If
End of For loop

If $count_H = count_L$ /* There is a tie */
    If noCVDIntervalFlag = true
        Set machineDecidableFlag = false
    Else
        Set pCVDInterval = $(sum_H + sum_L)/h$
        Calculate $\mu_H$ based on pCVDInterval
        Calculate $\mu_L$ based on pCVDInterval
    End If
Else If $count_H > count_L$
    Set pClass = High CVD Risk
    Set pCVDInterval = $sum_H / count_H$
    Calculate $\mu_H$ based on pCVDInterval
    Calculate $\mu_L$ based on pCVDInterval
Else /* $count_H < count_L$ */
    Set pClass = Low CVD Risk
    If noCVDIntervalFlag = true
        Set cvdIntervalPredictableFlag = false
        Set $\mu_H = 0$
        Set $\mu_L = 1$
```

```
37:         Else
38:             Set pCVDInterval = $sum_L$ / $count_L$
39:             Calculate $\mu_H$ based on pCVDInterval
40:             Calculate $\mu_L$ based on pCVDInterval
41:         End If
42:     End If
```

The Revised Matched Cases algorithm iterates through each case in the list of matched cases *M*, counting how many "High CVD Risk" cases *count$_H$*, and how many "Low CVD Risk" cases *count$_L$* are present in *M*. While doing that, it also identifies if there are cases with no CVD Interval. After that, the algorithm bases on the findings to decide CVD prediction outcomes. There are three types of outcomes: CVD Class and CVD Interval predictable, only CVD Class predictable (*cvdIntervalPredictableFlag = false* and *machineDecidableFlag = true*), and neither CVD Class nor CVD Interval predictable (*machineDecidableFlag = false*).

Important notes explaining the Revised Matched Cases algorithm:

- **At line 28:** The algorithm calculates the predicted CVD Interval (*pCVDInterval*) for the new case by only averaging the matched "High CVD Risk" cases. The reason for not averaging all *h* cases' CVD Intervals is because this may result in a *pCVDInterval* conflicts with *pClass*, for example, *pClass* is "High CVD Risk" but *pCVDInterval* is a value greater than 10 years.
- **At line 38:** The algorithm calculates the predicted CVD Interval (*pCVDInterval*) for the new case by only averaging the matched "Low CVD Risk" cases. The reason for not averaging all *h* cases' CVD Intervals is because this may result in a *pCVDInterval* conflicts with *pClass*, for example, *pClass* is "Low CVD Risk" but *pCVDInterval* is a value less than 10 years.
- **At line 35 and 36:** when *pClass* is "Low CVD Risk" and *pCVDInterval* is unknown, set $\mu_H = 0$ and $\mu_L = 1$. Otherwise, if $0 < \mu_L < 1$, *pCVDInterval* can be calculated using the "Low CVD Risk" membership function.

When there is no matched case found, the Revise algorithm uses the following "Revise Closest Cases" algorithm to generate CVD prediction results for the new input case.

Revise Closest Cases algorithm

```
Calculate $\mu_H$ using Equation (14)
Calculate $\mu_L$ using Equation (14)

If $\mu_H \geq \mu_L$
    Set pClass = High CVD Risk
    If $\mu_H = 1$   /* CVD Interval <= 5 years */
        Calculate pCVDInterval = average of CVD Intervals of nearest High Risk cases
    Else         /* 10 years >= CVD Interval > 5 years */
        Calculate pCVDInterval from the High CVD Risk membership function
Else
    Set pClass = Low CVD Risk
    If $\mu_L = 1$   /* CVD Interval >= 15 years */
        Calculate pCVDInterval = average of CVD Intervals of the nearest Low Risk
        cases that have CVD Interval. If all the nearest Low Risk cases don't have CVD
        Interval, set cvdIntervalPredictableFlag = false.
    Else         /* 10 years < CVD Interval < 15 years*/
        Calculate pCVDInterval from the Low CVD Risk membership function
End If
```

The "Revise Closest Cases" algorithm uses Equation (14), which is the core part of the fuzzy KNN algorithm [171], to calculate μ_H and μ_L. From these membership values, the predicted CVD Class can be decided, and the predicted CVD Interval can be calculated. Equation (14) is defined as below:

$$\mu_i(c) = \frac{\sum_{j=1}^{k} \frac{\mu_{ij}}{\left(distance(c_j, c)\right)^2}}{\sum_{j=1}^{k} \frac{1}{\left(distance(c_j, c)\right)^2}} \tag{14}$$

where c is the new input case, c_j is a case in the list of k closest cases, μ_{ij} is the membership value of c_j in the i^{th} class (in this context, there are two classes, "High CVD Risk" and "Low CVD Risk"), the distance between c_j and c is calculated using the Distance algorithm detailed in section 4.5.

Important notes explaining the Revised Closest Cases algorithm:

- **At line 7:** When μ_H = 1, all *k* closest cases must belong to "High CVD Risk". This can be proved by solving Equation (14) to conclude that all μ_{ij} = 1 (*i* is associated with the "High CVD Risk" set, *j* is from 1 to *k*).
- **At line 13:** When μ_L = 1, all *k* closest cases must belong to "Low CVD Risk". This can be proved by solving Equation (14) to conclude that all μ_{ij} = 1 (*i* is associated with the "Low CVD Risk" set, *j* is from 1 to *k*).

4.8 RETAIN

The purpose of the Retain activity is to save the new case when it has follow-up results of CVD statuses. As a result, this case becomes an existing case in the case base contributing to CVD prediction. Though a complete CBR system should have all four activities, Retrieve, Reuse, Revise, and Retain, only the Retrieve, Reuse, and Revise activities were developed in this research. The reasons were due to both the time limitation and the fact that, in conducting this research, existing FHS datasets that already have follow-up results were used. Therefore, the Retain activity is out of scope for this research. Neither design nor implementation was undertaken for the Retain activity.

4.9 OUTPUT (PREDICTION RESULT)

The output provides a CVD prediction result including predicted CVD Class, predicted CVD Interval, predicted High CVD Risk membership, and predicted Low CVD Risk membership. In addition, the CRISK system also displays on the prediction result screen all *k* closest cases. Moreover, any matched cases among these *k* closest cases are also marked. The prediction result screen with all these details is designed to assist medical practitioners (such as doctors) in reviewing and making decision of CVD risk prediction for the new input case. Two screenshots of the prediction result screen are given in Figure 5-32 and Figure 5-33 in the next chapter.

4.10 CHAPTER SUMMARY

The CRISK Prediction model is a CBR system whose case base is a fuzzy ontology. At this stage, only the outcome “10-year CVD” is fuzzified as membership values of “High CVD Risk” and “Low CVD Risk” fuzzy sets, whose membership functions are described in Equations (4) and (5) respectively. Fundamentally based on the Fuzzy KNN algorithm by Keller et al. [171], the algorithms used for the Retrieve, Reuse, and Revise activities of the CRISK Prediction model were developed. Main contributions to the original Fuzzy KNN algorithm include the development of the Distance algorithm, the development of the “Revise matched cases” algorithm, and the enhancement of the “Revise closest cases” algorithm to generate not only CVD risk class but also CVD prediction interval.

Chapter 5

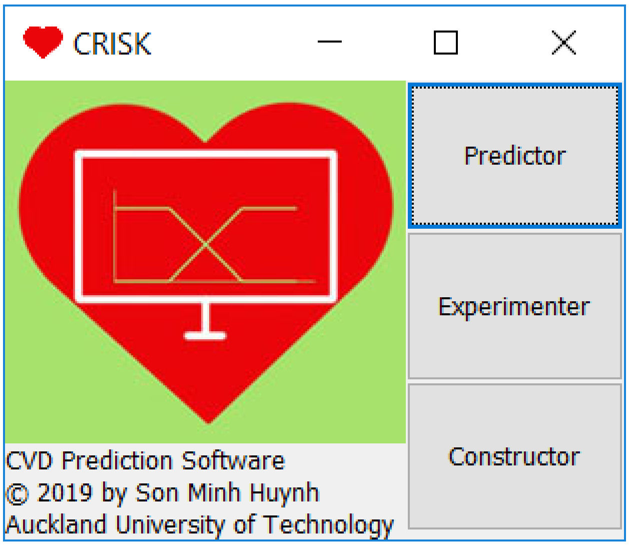
CRISK SYSTEM IMPLEMENTATION

5.1 INTRODUCTION

This chapter describes how the CRISK system was developed. Java (version 8) was the programming language and Java Swing was the GUI widget toolkit for the development. As mentioned previously, the Fuzzy OWL 2 Protégé plugin [141] was used to create the "High CVD Risk" and "Low CVD Risk" fuzzy data types. This chapter together with Chapter 4 and Chapter 6 help answer the first three research questions (RQ1, RQ2, and RQ3). The answers are described in section 9.1.1.

The CRISK system consists of four modules: Constructor, Experimenter, Batch Experimenter, and Predictor. Details of each module are further explained in sections 5.2 to 5.5 respectively.

Figure 5-1 shows the welcome screen of the CRISK application. The Constructor button launches the Constructor module. The Experimenter button opens both the Experimenter module and the Batch Experimenter module. The Predictor button starts the Predictor module.



CVD Prediction Software
© 2019 by Son Minh Huynh
Auckland University of Technology

Figure 5-1: CRISK welcome screen

5.2 CRISK Constructor Module

The Constructor module (Figure 5-2) is used to transform a CVD dataset in CSV format into a fuzzy ontology file in OWL 2 format for use in the Experimenter, Batch Experimenter and Predictor modules. The Constructor module takes a dataset file (Figure 5-5) and a predictors file (Figure 5-6) as inputs from the user. Upon clicking on the “Create Ontology” button, a dialog window is opened asking the user to save the ontology file to be created (Figure 5-3).

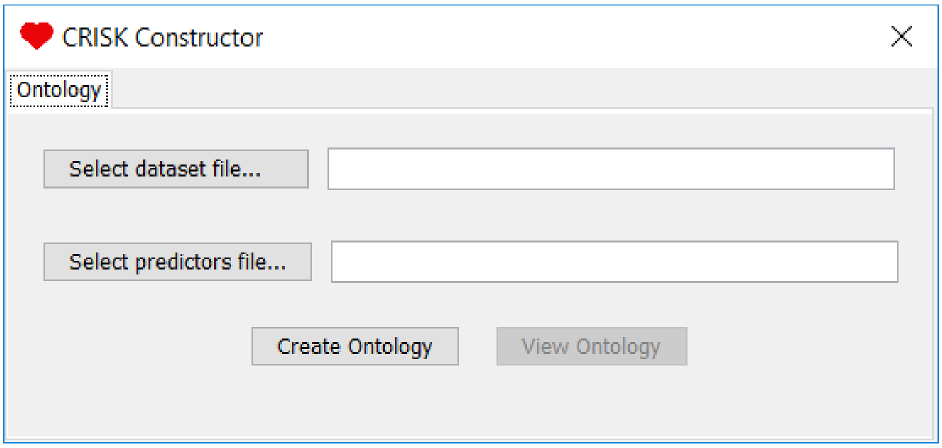
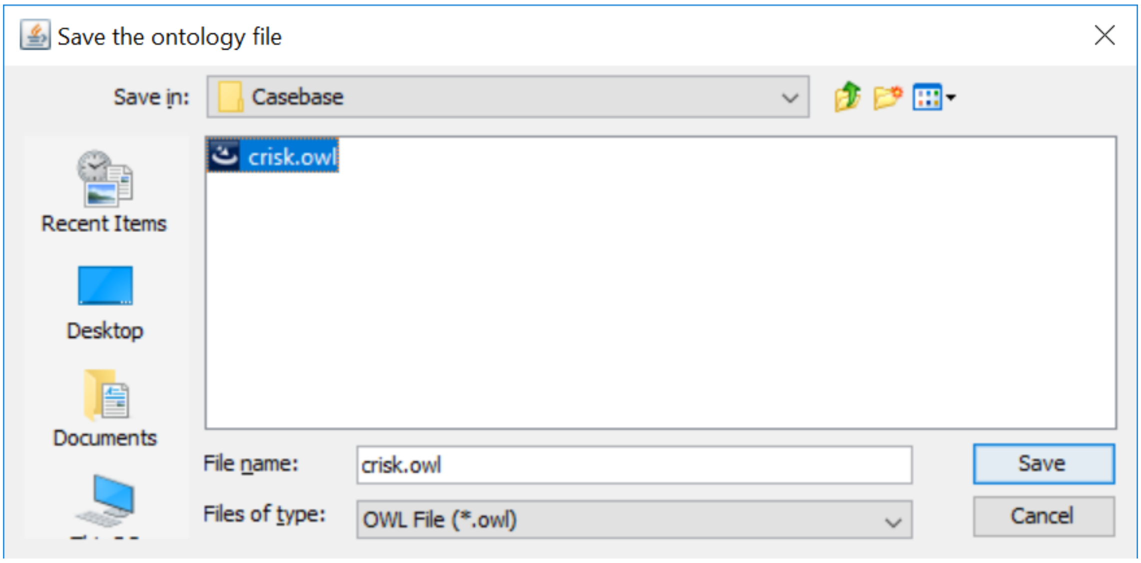


Figure 5-2: CRISK Constructor screen



Save the ontology file

Save in: Casebase

crisk.owl

File name: crisk.owl

Files of type: OWL File (*.owl)

Figure 5-3: CRISK Constructor—Asking the user to save the ontology file

Figure 5-4 shows the ontology construction process. From the dataset file and the predictors file input, the Constructor module uses the CRISK fuzzy ontology template file to create a fuzzy ontology file for the dataset.

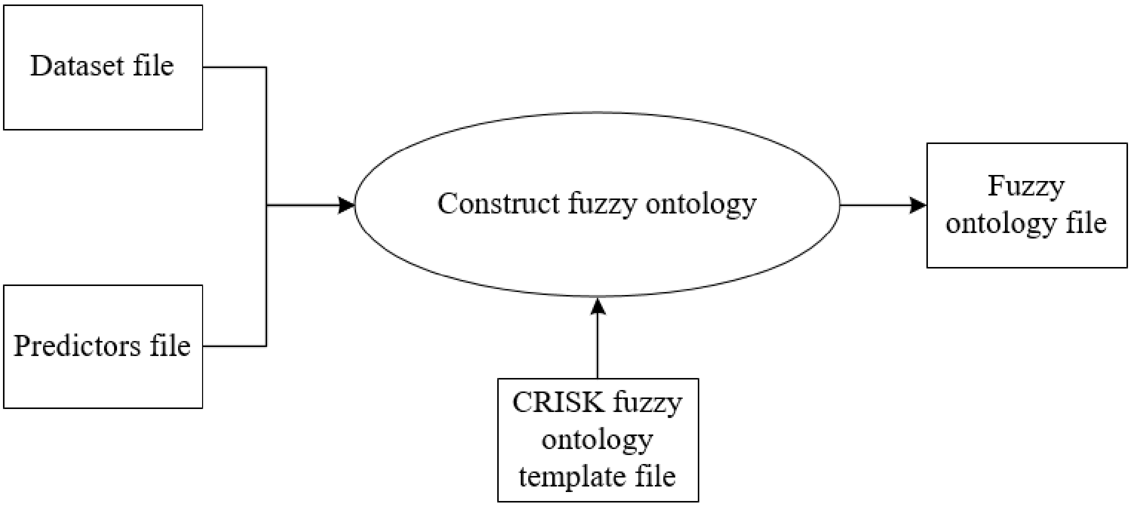
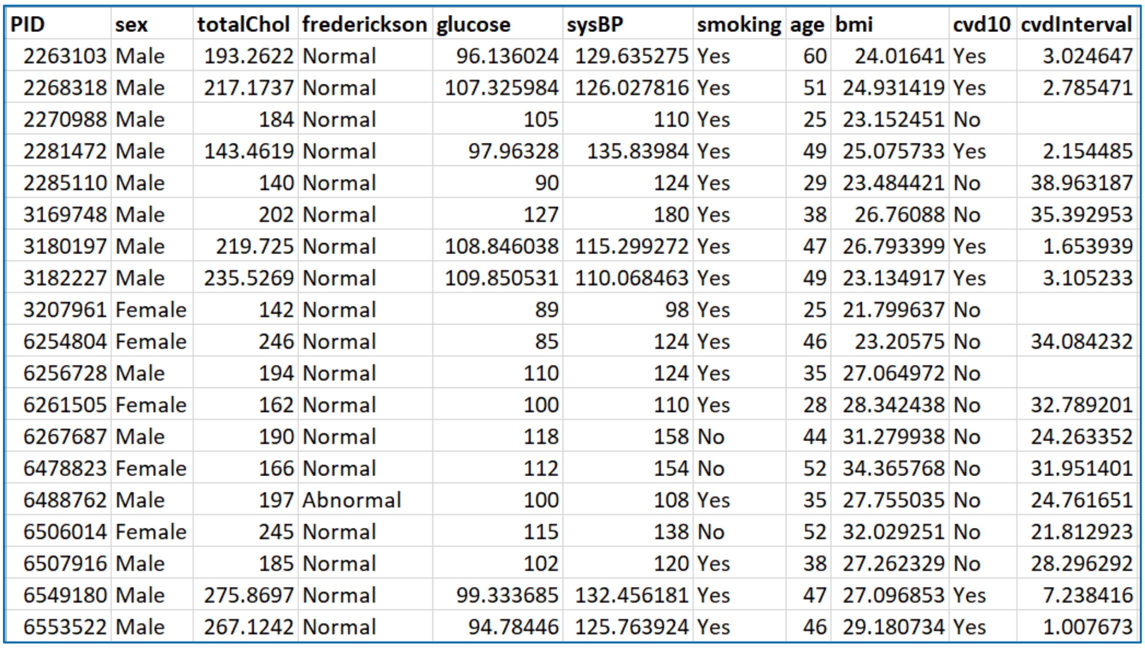


Figure 5-4: CRISK Constructor—Fuzzy ontology construction process

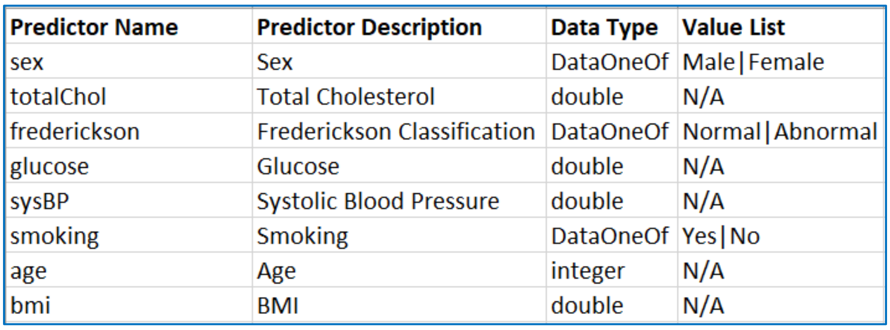
The dataset file is a CSV file containing predictor names as column headings and values for all cases of the dataset. Each row contains the values of a case. The first column is the case unique identification. The last two columns are 10-year CVD and CVD Interval (in years) respectively. The other columns in the middle are predictors. An example of a dataset file is screenshot below (Figure 5-5).



PID	sex	totalChol	frederickson	glucose	sysBP	smoking	age	bmi	cvd10	cvdInterval
2263103	Male	193.2622	Normal	96.136024	129.635275	Yes	60	24.01641	Yes	3.024647
2268318	Male	217.1737	Normal	107.325984	126.027816	Yes	51	24.931419	Yes	2.785471
2270988	Male	184	Normal	105	110	Yes	25	23.152451	No	
2281472	Male	143.4619	Normal	97.96328	135.83984	Yes	49	25.075733	Yes	2.154485
2285110	Male	140	Normal	90	124	Yes	29	23.484421	No	38.963187
3169748	Male	202	Normal	127	180	Yes	38	26.76088	No	35.392953
3180197	Male	219.725	Normal	108.846038	115.299272	Yes	47	26.793399	Yes	1.653939
3182227	Male	235.5269	Normal	109.850531	110.068463	Yes	49	23.134917	Yes	3.105233
3207961	Female	142	Normal	89	98	Yes	25	21.799637	No	
6254804	Female	246	Normal	85	124	Yes	46	23.20575	No	34.084232
6256728	Male	194	Normal	110	124	Yes	35	27.064972	No	
6261505	Female	162	Normal	100	110	Yes	28	28.342438	No	32.789201
6267687	Male	190	Normal	118	158	No	44	31.279938	No	24.263352
6478823	Female	166	Normal	112	154	No	52	34.365768	No	31.951401
6488762	Male	197	Abnormal	100	108	Yes	35	27.755035	No	24.761651
6506014	Female	245	Normal	115	138	No	52	32.029251	No	21.812923
6507916	Male	185	Normal	102	120	Yes	38	27.262329	No	28.296292
6549180	Male	275.8697	Normal	99.333685	132.456181	Yes	47	27.096853	Yes	7.238416
6553522	Male	267.1242	Normal	94.78446	125.763924	Yes	46	29.180734	Yes	1.007673

Figure 5-5: An example of a dataset file

The predictors file is also a CSV file, used to describe the risk factors. The predictors file has four columns, Predictor Name, Predictor Description, Data Type, and Value List. The Predictor Name column lists the predictors from the dataset file in the same order. The Predictor Description column describes the predictor. The Data Type column indicates that whether a predictor is a double, an integer, a DataOneOf (nominal) etc. These are data types in OWL 2. For a predictor whose data type is DataOneOf, a list of values separated by the vertical bar character must be defined in the Value List column. An example of a predictors file is screenshot below (Figure 5-6).

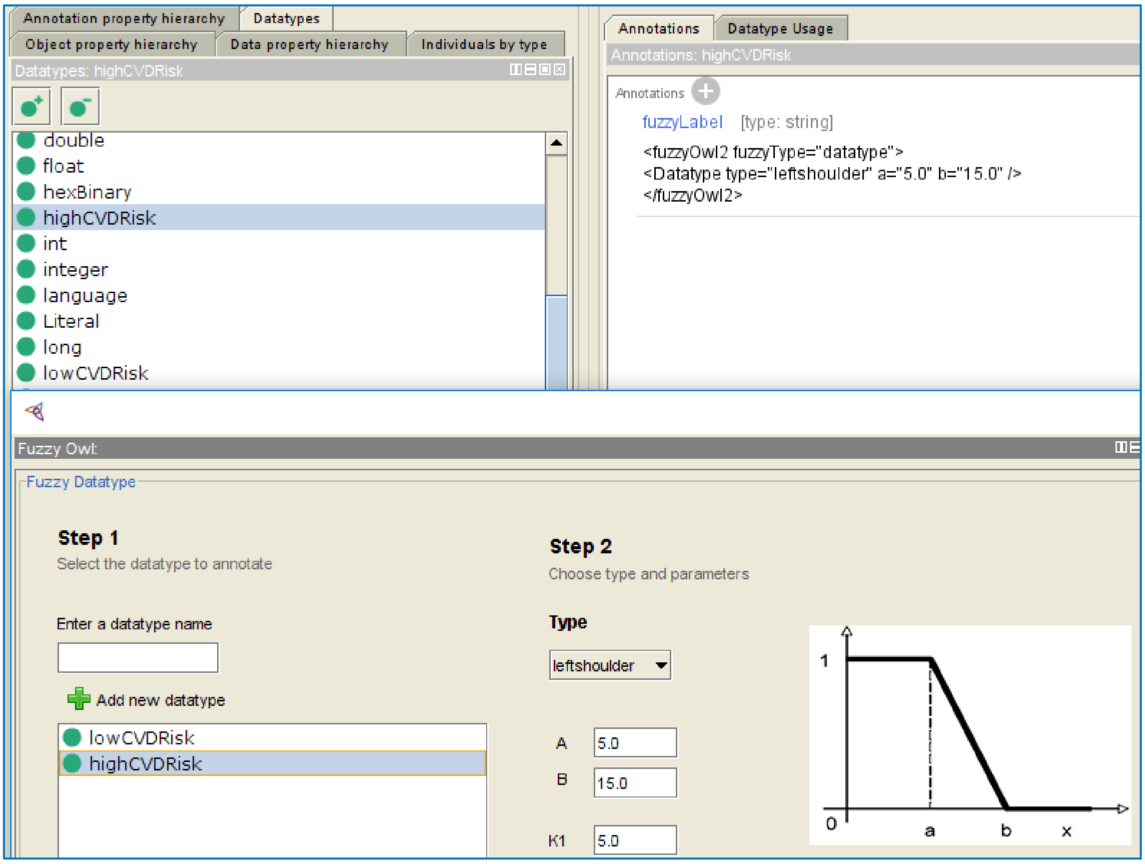


Predictor Name	Predictor Description	Data Type	Value List
sex	Sex	DataOneOf	Male|Female
totalChol	Total Cholesterol	double	N/A
frederickson	Frederickson Classification	DataOneOf	Normal|Abnormal
glucose	Glucose	double	N/A
sysBP	Systolic Blood Pressure	double	N/A
smoking	Smoking	DataOneOf	Yes|No
age	Age	integer	N/A
bmi	BMI	double	N/A

Figure 5-6: An example of a predictors file

The CRISK fuzzy ontology template file is stored in the CRISK system as a resource file named “base.owl”. The content of the file is provided in Appendix P. The file basically contains two fuzzy data types, a class hierarchy, and two data properties for the two prediction attributes (cvd10 and cvdInterval). The two fuzzy data types highCVDRisk and lowCVDRisk were created based on the membership functions defined in Equations

(4) and (5) in section 3.4.2. Figure 5-7 and Figure 5-8 respectively show these two fuzzy data types viewed in Protégé having the Fuzzy OWL 2 plugin installed. The class hierarchy (Figure 5-9) defines two classes, CBR_CASE class and its child class CRISK_CASE. The CRISK_CASE class is where a CVD case belongs. The two data properties for the two prediction attributes are named cvd10 and cvdInterval. Figure 5-10 and Figure 5-11 respectively show them in Protégé.



fuzzyLabel [type: string]

```
<fuzzyOwl2 fuzzyType="datatype">
<Datatype type="leftshoulder" a="5.0" b="15.0" />
</fuzzyOwl2>
```

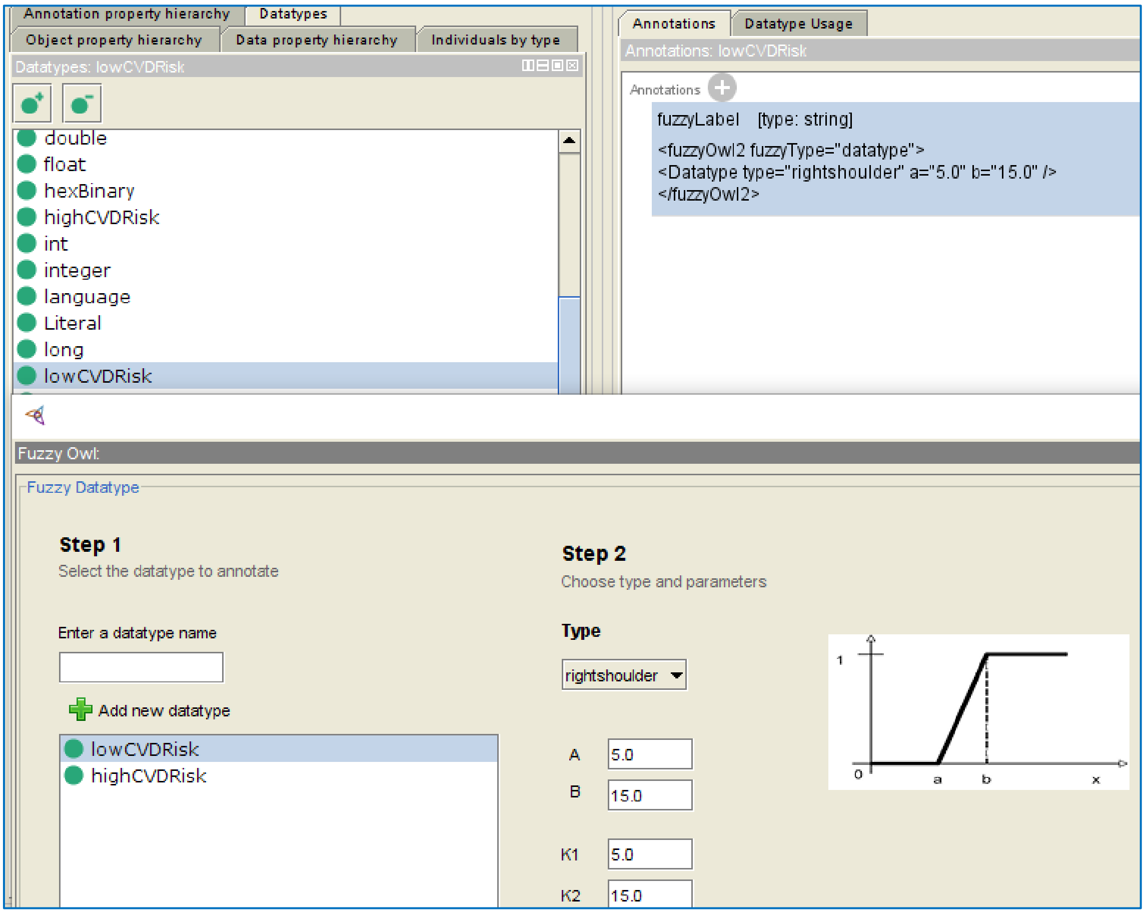
Type: leftshoulder

A: 5.0

B: 15.0

K1: 5.0

Figure 5-7: Viewing highCVDRisk data type in Protégé with the Fuzzy OWL 2 plugin



fuzzyLabel [type: string]

```
<fuzzyOwl2 fuzzyType="datatype">
<Datatype type="rightshoulder" a="5.0" b="15.0" />
</fuzzyOwl2>
```

Step 1
Select the datatype to annotate

Enter a datatype name

lowCVDRisk
highCVDRisk

Step 2
Choose type and parameters

Type
rightshoulder

A 5.0
B 15.0
K1 5.0
K2 15.0

Figure 5-8: Viewing lowCVDRisk data type in Protégé with the Fuzzy OWL 2 plugin

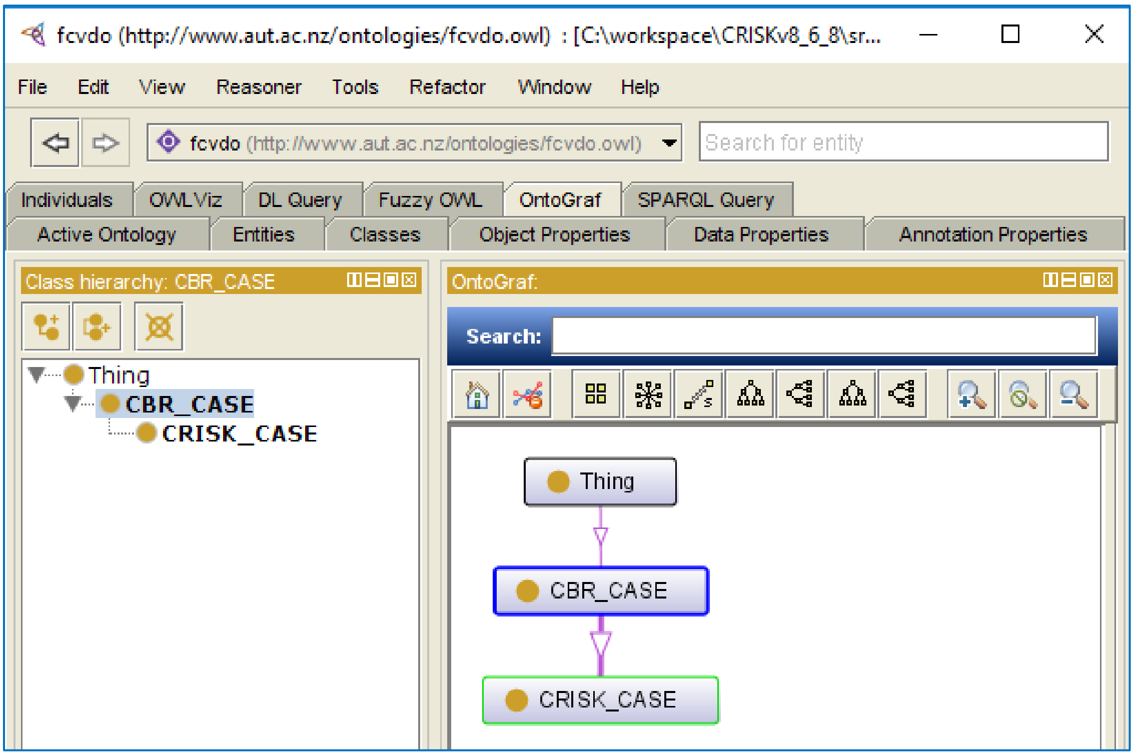
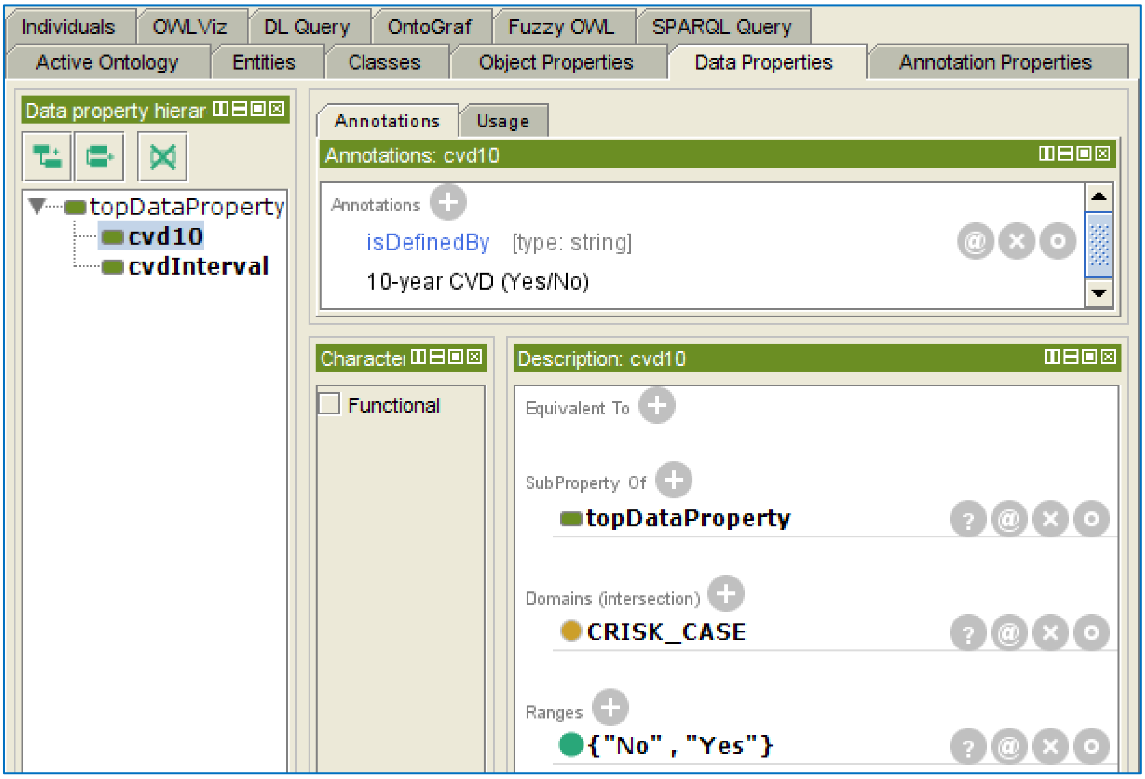


Figure 5-9: Viewing CRISK class hierarchy in Protégé



topDataProperty
- cvd10
- cvdInterval

Annotations: cvd10

isDefinedBy [type: string]
10-year CVD (Yes/No)

□ Functional

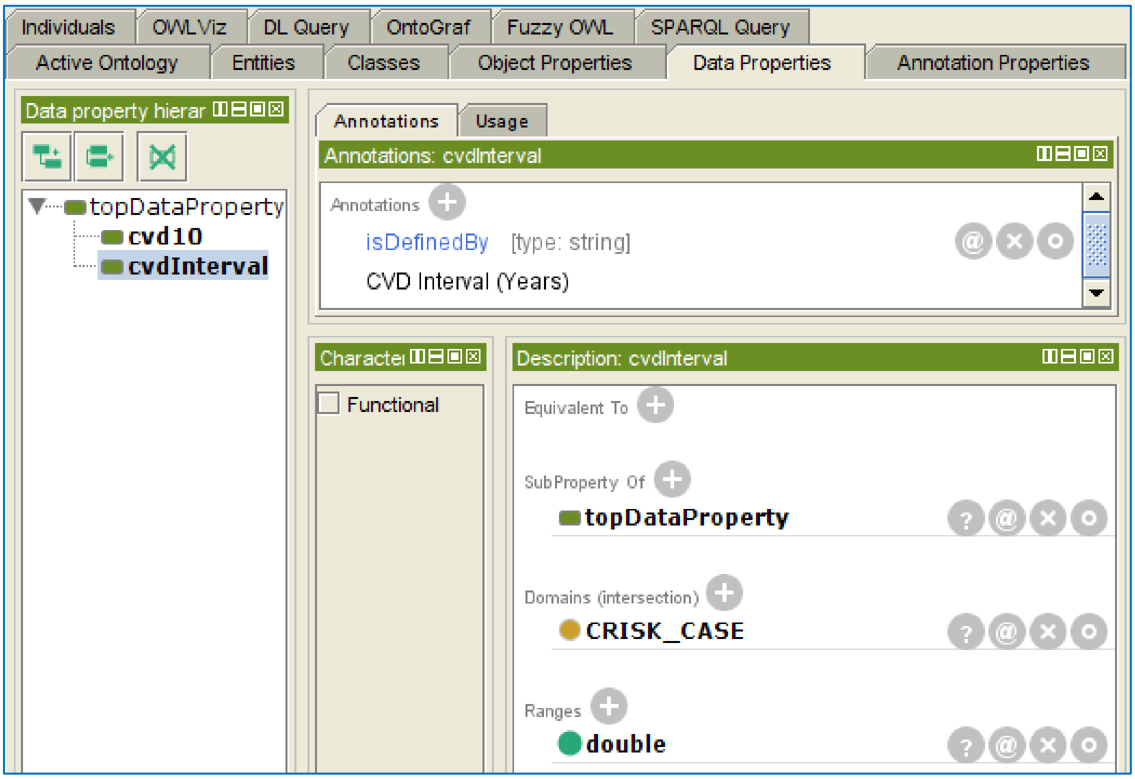
Description: cvd10

SubProperty Of: topDataProperty

Domains (intersection): CRISK_CASE

Ranges: {"No" , "Yes"}

Figure 5-10: Viewing cvd10 data property in Protégé



topDataProperty
- cvd10
- cvdInterval

Annotations: cvdInterval

isDefinedBy [type: string]
CVD Interval (Years)

□ Functional

Description: cvdInterval

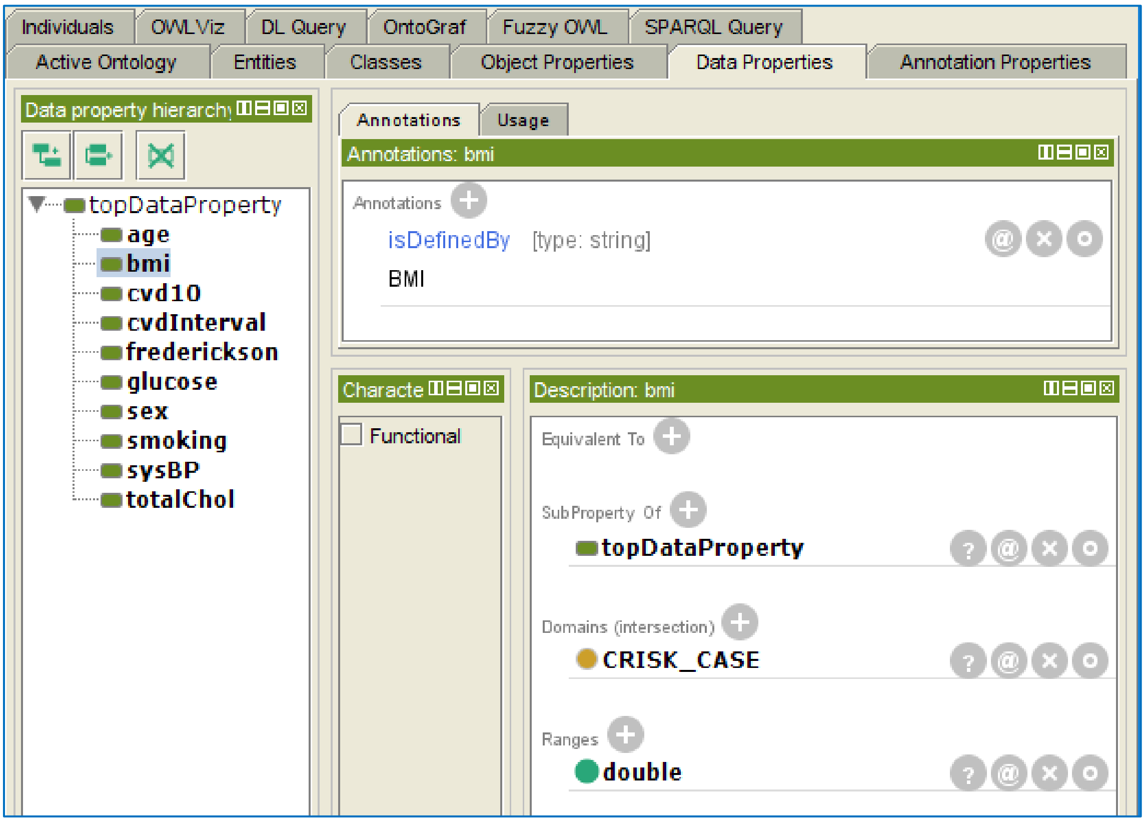
SubProperty Of: topDataProperty

Domains (intersection): CRISK_CASE

Ranges: double

Figure 5-11: Viewing cvdInterval data property in Protégé

The output file generated by the CRISK Constructor module is a fuzzy ontology file. Two fuzzy data types highCVDRisk and lowCVDRisk are created for the ontology from the CRISK fuzzy ontology template file. Each risk factor from the CSV dataset file becomes a data property in the created fuzzy ontology. The two prediction attributes cvd10 and cvdInterval are also created for the ontology as data properties using the CRISK fuzzy ontology template file. Figure 5-12 shows these data properties of a sample fuzzy ontology created for a sample dataset (Figure 5-5) in Protégé. Each case from the CSV dataset file becomes an individual in the created fuzzy ontology. Each individual is uniquely identified by its IRI (Internationalised Resource Identifier) whose value is the PID of the case. Figure 5-13 displays individuals of the sample fuzzy ontology in Protégé. These individuals belong to the CRISK_CASE class, whose parent class is CBR_CASE.



Annotations: bmi

Annotations

isDefinedBy [type: string]
BMI

topDataProperty
- age
- bmi
- cvd10
- cvdInterval
- frederickson
- glucose
- sex
- smoking
- sysBP
- totalChol

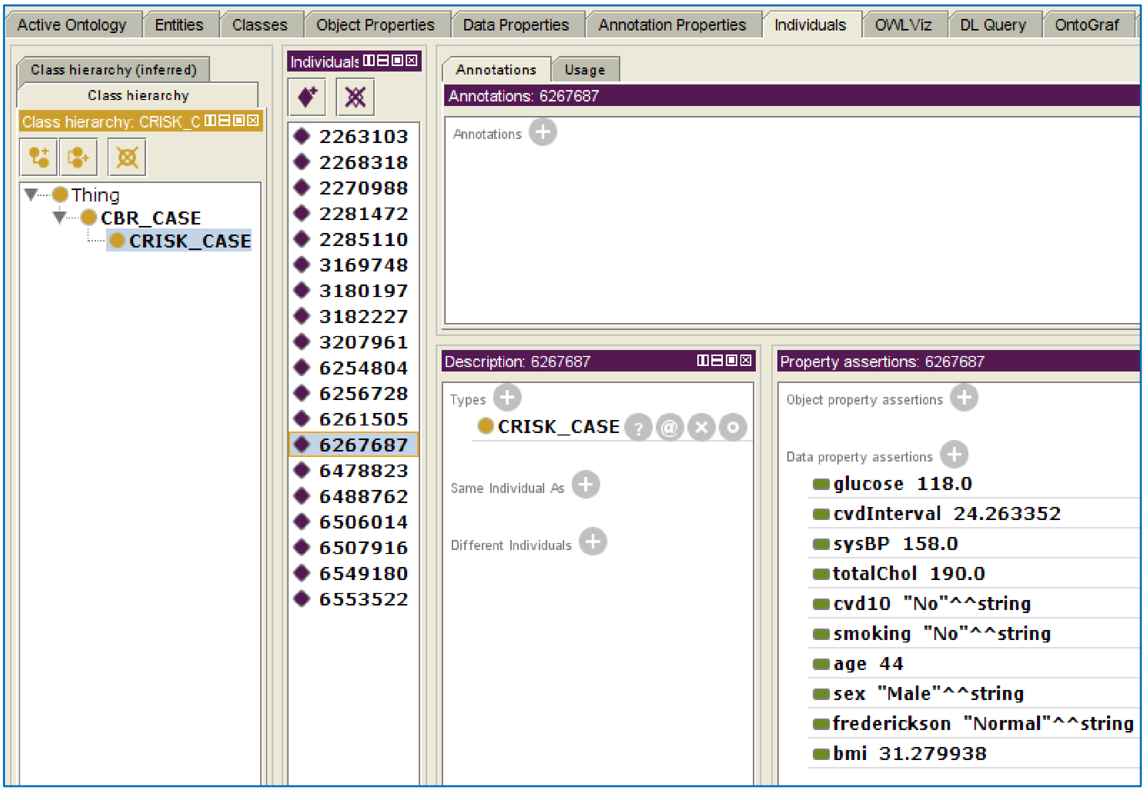
□ Functional

SubProperty Of
topDataProperty

Domains (intersection)
CRISK_CASE

Ranges
double

Figure 5-12: Viewing data properties of the sample fuzzy ontology in Protégé



Thing
CBR_CASE
CRISK_CASE

2263103
2268318
2270988
2281472
2285110
3169748
3180197
3182227
3207961
6254804
6256728
6261505
6267687
6478823
6488762
6506014
6507916
6549180
6553522

Types: CRISK_CASE

Data property assertions:

glucose 118.0
cvdInterval 24.263352
sysBP 158.0
totalChol 190.0
cvd10 "No"^^string
smoking "No"^^string
age 44
sex "Male"^^string
frederickson "Normal"^^string
bmi 31.279938

Figure 5-13: Viewing individuals of the sample fuzzy ontology in Protégé

5.3 CRISK EXPERIMENTER MODULE

The Experimenter module is used to experiment on a CVD dataset stored as a fuzzy ontology for prediction performance based on different numbers of nearest neighbours. The fuzzy ontology input to this model is created using the CRISK Constructor module (section 5.2). The Experimenter module offers three experimentation types: LOOCV, Train-Test-LOOCV, and Train-Test.

For the LOOCV experiment (Figure 5-14), one ontology is selected. The system iterates through all the CVD cases stored in the ontology. For each case, the system removes the case from the case base (the selected ontology), performs prediction for the case, and adds the case back to the case base for the next iteration of the loop.

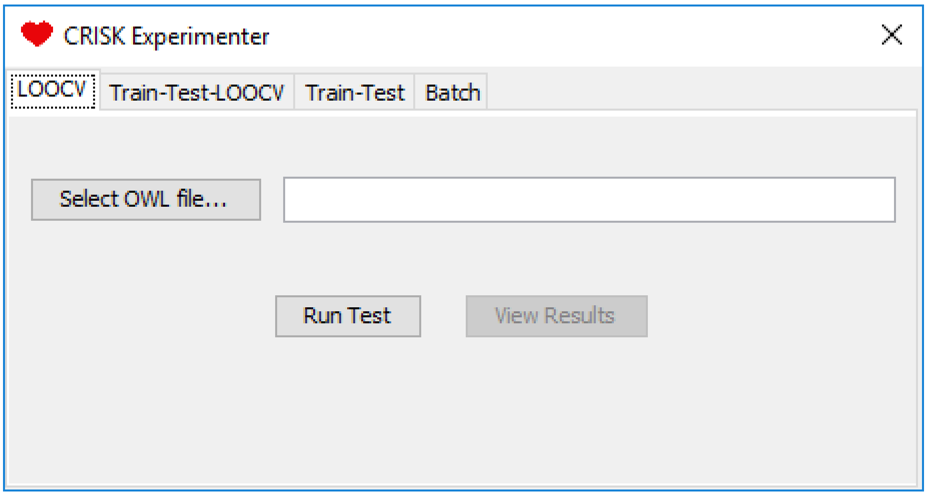


Figure 5-14: CRISK Experimenter—LOOCV screen

For the Train-Test-LOOCV experiment (Figure 5-15), two ontologies files are selected. The testing OWL file is a subset of the training OWL file. This happens when, for example, the training OWL file is generated from the testing OWL file using SMOTE for imbalanced dataset. As a result, for each case in the testing set, the system removes the case from the case base (the training set), performs prediction for the case, and adds the case back to the case base for the next iteration of the loop.

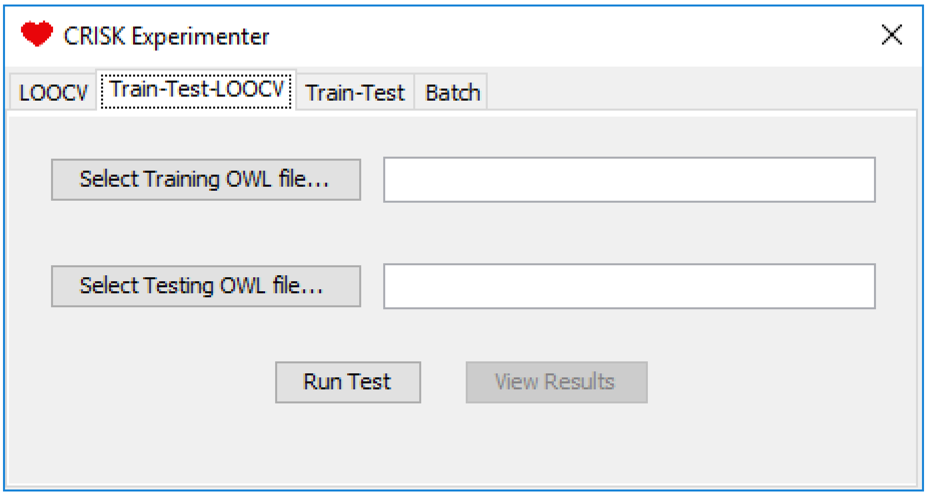


Figure 5-15: CRISK Experimenter—Train-Test-LOOCV screen

For the Train-Test experimentation (Figure 5-16), two ontology files are also selected. However, the testing and training sets are different sets and contain separate cases. Therefore, the system iterates through each case in the testing set and performs prediction for the case without removing any case from the case base (the training set).

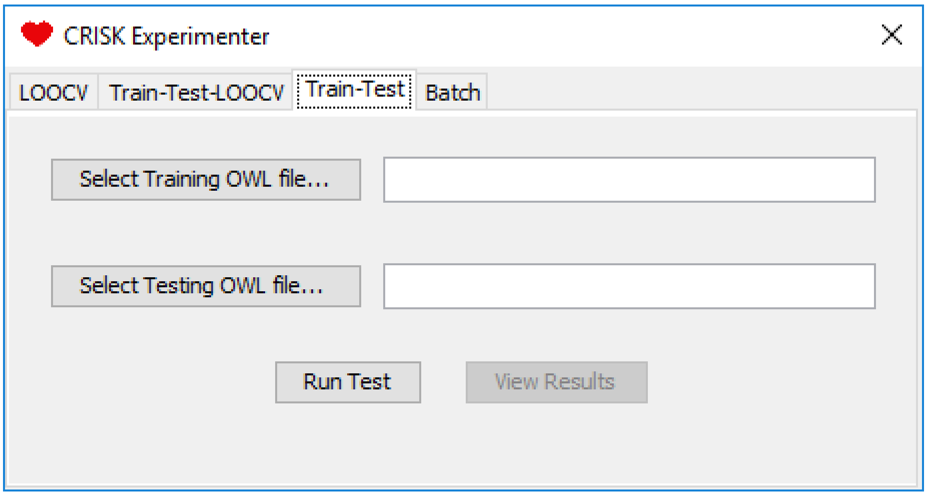
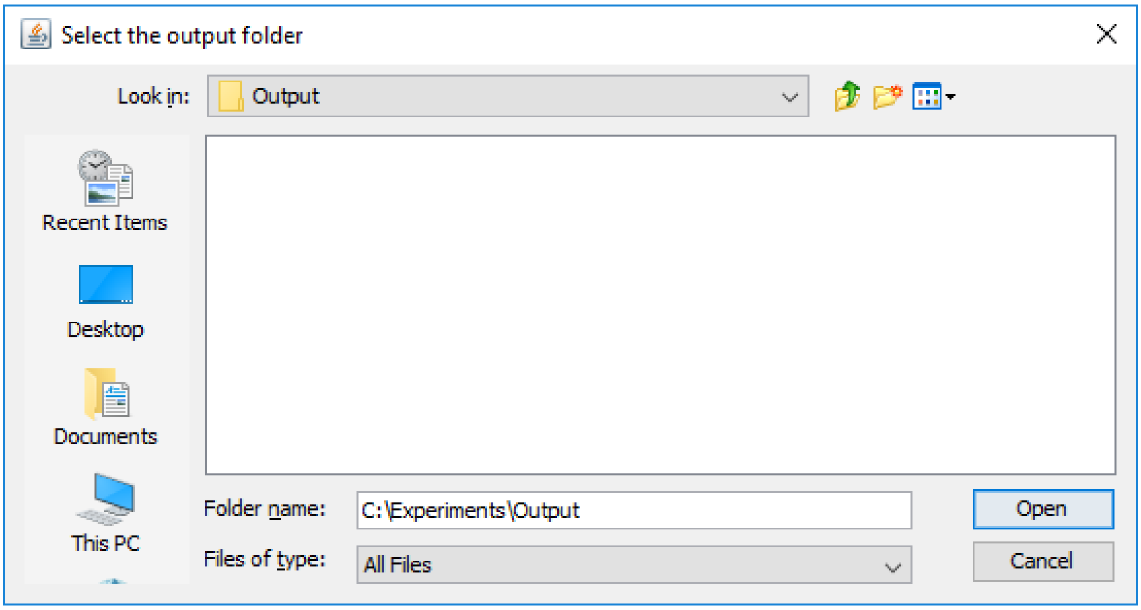


Figure 5-16: CRISK Experimenter—Train-Test screen

For each experimentation type, upon clicking on the “Run Test” button, the system opens a dialog (Figure 5-17) asking the user to select an output folder to store the experimentation results.



Look in: Output

Folder name: C:\Experiments\Output

Files of type: All Files

Figure 5-17: CRISK Experimenter—Output folder selection screen

Figure 5-18, Figure 5-19, and Figure 5-20 respectively provide a high level view of the LOOCV, Train-Test-LOOCV, and Train-Test experimentation processes.

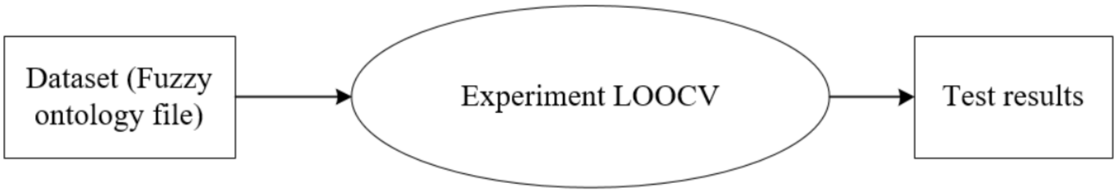


Figure 5-18: CRISK Experimenter—LOOCV experimentation process

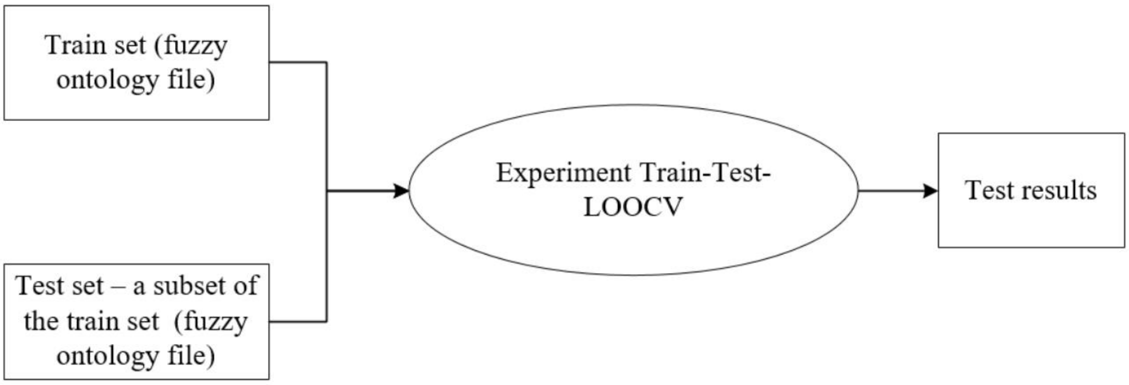

Train set (fuzzy ontology file)
Test set – a subset of the train set (fuzzy ontology file)
Experiment Train-Test-LOOCV
Test results

Figure 5-19: CRISK Experimenter—Train-Test-LOOCV experimentation process

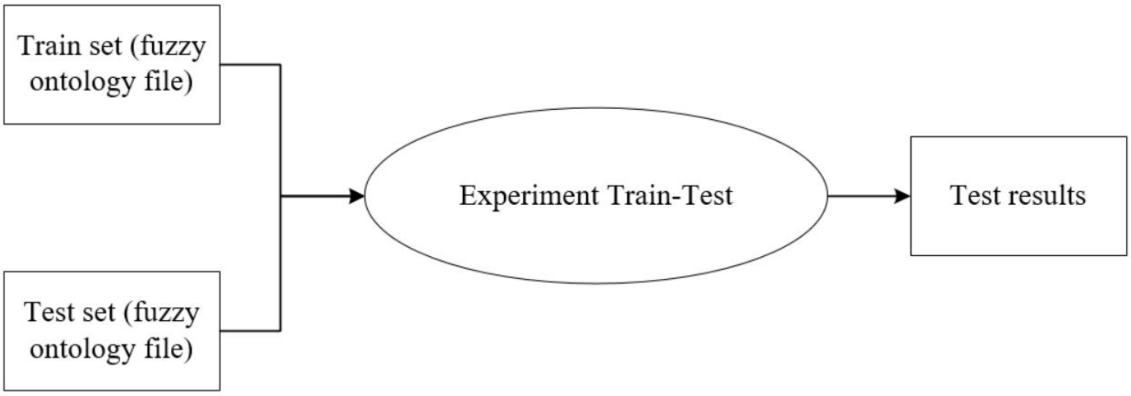
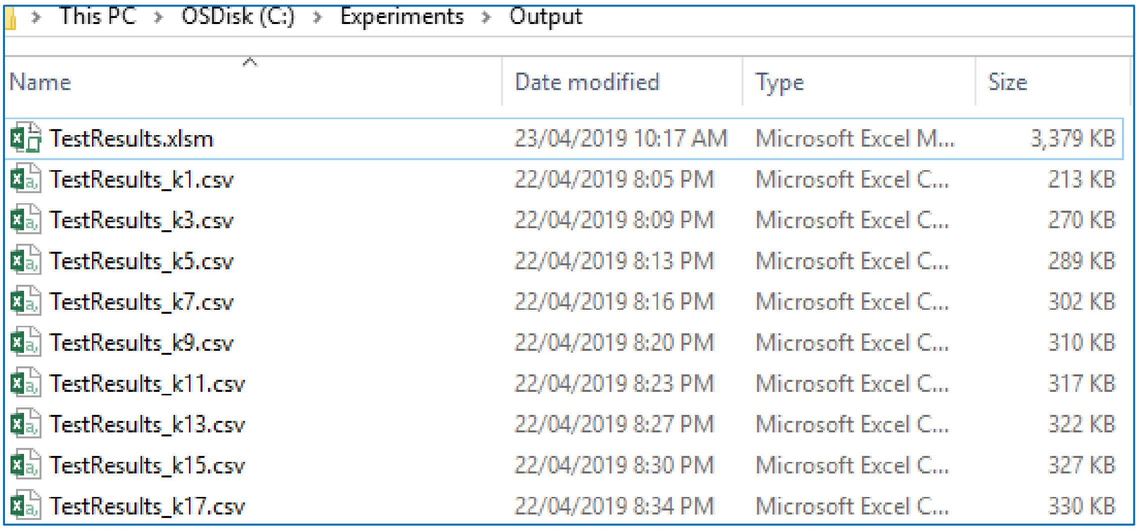

Train set (fuzzy ontology file)
Test set (fuzzy ontology file)
Experiment Train-Test
Test results

Figure 5-20: CRISK Experimenter—Train-Test experimentation process

The results of each experiment, stored inside the user selected output folder, consist of nine CSV files and one macro-enabled Excel file (Figure 5-21). Each CSV file contains the experiment results corresponding to one value of *k* (the number of nearest neighbours). The macro-enabled Excel file collects result data from the result CSV files and produces a summary of prediction performance.

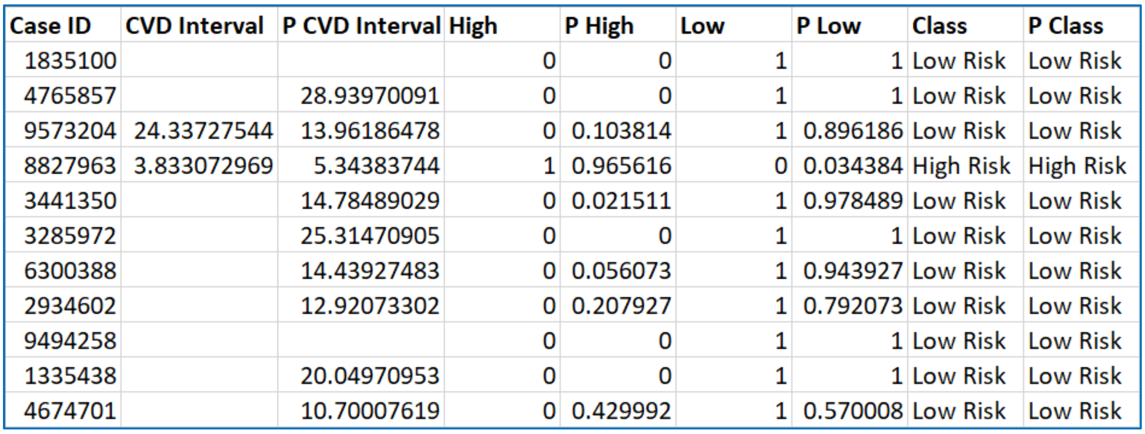


› This PC › OSDisk (C:) › Experiments › Output

Name	Date modified	Type	Size
TestResults.xlsm	23/04/2019 10:17 AM	Microsoft Excel M...	3,379 KB
TestResults_k1.csv	22/04/2019 8:05 PM	Microsoft Excel C...	213 KB
TestResults_k3.csv	22/04/2019 8:09 PM	Microsoft Excel C...	270 KB
TestResults_k5.csv	22/04/2019 8:13 PM	Microsoft Excel C...	289 KB
TestResults_k7.csv	22/04/2019 8:16 PM	Microsoft Excel C...	302 KB
TestResults_k9.csv	22/04/2019 8:20 PM	Microsoft Excel C...	310 KB
TestResults_k11.csv	22/04/2019 8:23 PM	Microsoft Excel C...	317 KB
TestResults_k13.csv	22/04/2019 8:27 PM	Microsoft Excel C...	322 KB
TestResults_k15.csv	22/04/2019 8:30 PM	Microsoft Excel C...	327 KB
TestResults_k17.csv	22/04/2019 8:34 PM	Microsoft Excel C...	330 KB

Figure 5-21: CRISK Experimenter—Experiment result files

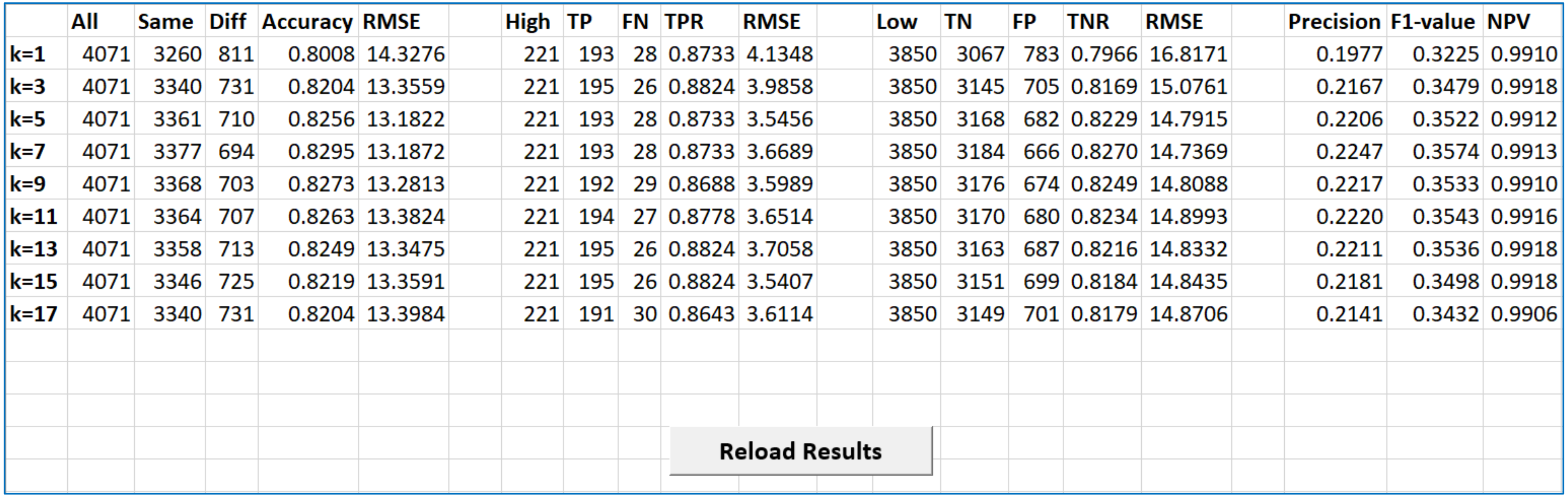
Figure 5-22 displays the content of a result CSV file. The “Case ID” column uniquely identifies a case. The “CVD Interval” column gives the actual CVD Interval in years while the “P CVD Interval” is the system predicted CVD Interval in years. The “High” column is the actual membership value of High CVD Risk while the “P High” is the predicted membership value of High CVD Risk. It is similarly explained for the “Low” and “P Low” columns. Finally, the “Class” and “P Class” columns respectively represent the actual CVD Class, “High CVD Risk” or “Low CVD Risk”, and the predicted CVD Class of a case. A case belongs to “High CVD Risk” when the “High CVD Risk” membership value is greater than or equal to the “Low CVD Risk” membership value. Belonging to “High CVD Risk” also means that the case develops CVD within 10 years (cvd10 = Yes).



Case ID	CVD Interval	P CVD Interval	High	P High	Low	P Low	Class	P Class
1835100			0	0	1	1	Low Risk	Low Risk
4765857		28.93970091	0	0	1	1	Low Risk	Low Risk
9573204	24.33727544	13.96186478	0	0.103814	1	0.896186	Low Risk	Low Risk
8827963	3.833072969	5.34383744	1	0.965616	0	0.034384	High Risk	High Risk
3441350		14.78489029	0	0.021511	1	0.978489	Low Risk	Low Risk
3285972		25.31470905	0	0	1	1	Low Risk	Low Risk
6300388		14.43927483	0	0.056073	1	0.943927	Low Risk	Low Risk
2934602		12.92073302	0	0.207927	1	0.792073	Low Risk	Low Risk
9494258			0	0	1	1	Low Risk	Low Risk
1335438		20.04970953	0	0	1	1	Low Risk	Low Risk
4674701		10.70007619	0	0.429992	1	0.570008	Low Risk	Low Risk

Figure 5-22: CRISK Experimenter—The content inside a result CSV file

Figure 5-23 displays an example of the summary sheet of a result summary macro-enabled Excel file. The sheet summarises the prediction performance of the CRISK system for all cases, “High CVD Risk” cases, and “Low CVD Risk” cases, against each value of the number of nearest neighbours *k*. In the example shown in this figure, the number of cases in total was 4,071. Among those, 221 cases were “High CVD Risk” cases and 3,850 cases were “Low CVD Risk” ones. The “Same” column gives the number of correctly classified cases while the “Diff” column shows the number of incorrectly classified cases. The “Accuracy” column provides the prediction accuracy. An RMSE (Root Mean Squared Error) column gives error measurement for prediction of CVD Interval, for cases having both CVD Interval and predicted CVD Interval. Other columns include “TP”, “FN”, “TPR”, “TN”, “FP”, “TNR”, “Precision”, “F_1-value”, and “NPV”. The meanings of these columns can be found in section 3.9.1. Clicking on the “Reload results” button recollects data from each result CSV file and refreshes the data on the summary sheet.



	All	Same	Diff	Accuracy	RMSE	High	TP	FN	TPR	RMSE	Low	TN	FP	TNR	RMSE	Precision	F1-value	NPV
k=1	4071	3260	811	0.8008	14.3276	221	193	28	0.8733	4.1348	3850	3067	783	0.7966	16.8171	0.1977	0.3225	0.9910
k=3	4071	3340	731	0.8204	13.3559	221	195	26	0.8824	3.9858	3850	3145	705	0.8169	15.0761	0.2167	0.3479	0.9918
k=5	4071	3361	710	0.8256	13.1822	221	193	28	0.8733	3.5456	3850	3168	682	0.8229	14.7915	0.2206	0.3522	0.9912
k=7	4071	3377	694	0.8295	13.1872	221	193	28	0.8733	3.6689	3850	3184	666	0.8270	14.7369	0.2247	0.3574	0.9913
k=9	4071	3368	703	0.8273	13.2813	221	192	29	0.8688	3.5989	3850	3176	674	0.8249	14.8088	0.2217	0.3533	0.9910
k=11	4071	3364	707	0.8263	13.3824	221	194	27	0.8778	3.6514	3850	3170	680	0.8234	14.8993	0.2220	0.3543	0.9916
k=13	4071	3358	713	0.8249	13.3475	221	195	26	0.8824	3.7058	3850	3163	687	0.8216	14.8332	0.2211	0.3536	0.9918
k=15	4071	3346	725	0.8219	13.3591	221	195	26	0.8824	3.5407	3850	3151	699	0.8184	14.8435	0.2181	0.3498	0.9918
k=17	4071	3340	731	0.8204	13.3984	221	191	30	0.8643	3.6114	3850	3149	701	0.8179	14.8706	0.2141	0.3432	0.9906

Figure 5-23: CRISK Experimenter—Inside a result summary macro-enabled Excel file

5.4 CRISK Batch Experimenter Module

The Batch Experimenter module is used to experiment a CVD dataset stored as a fuzzy ontology for prediction performance based on different numbers of nearest neighbours and different combinations of predictors. This also means that the Batch Experimenter adds one more dimension—the combination of predictors—into the Experimenter module detailed in section 5.3. The Batch Experimenter runs as a batch process in a command line window. Usually it takes a long time to run due to the number of experiments related to the different combinations of predictors, which are ranked by a predictors ranking file.

Different combinations of predictors are made using a backward elimination technique. To perform this backward elimination technique, the program begins with all *n* predictors in the first iteration. In the second iteration, it experiments with $n - 1$ predictors by omitting the least important predictor (the last predictor in the ranking). In the third iteration, it experiments with $n - 2$ predictors by omitting the last two predictors. Eventually, in the last iteration, it runs with only 1 predictor, the first predictor in the ranking.

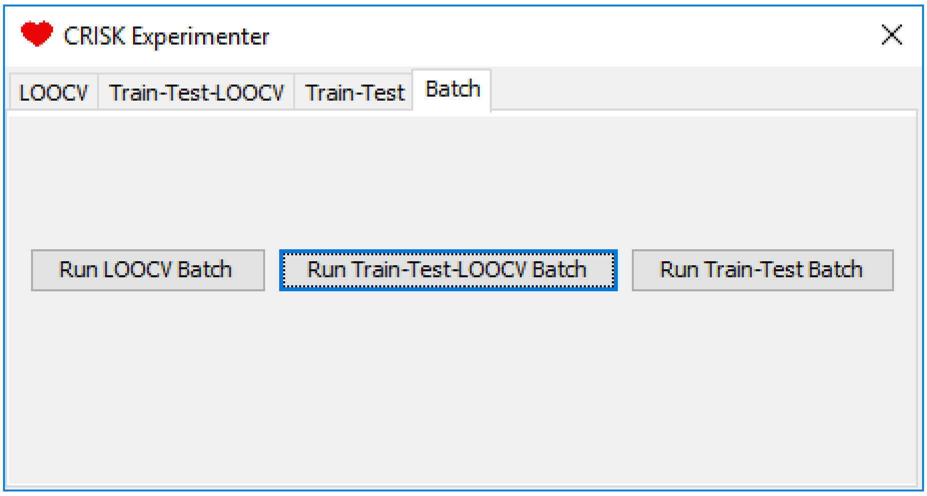


Figure 5-24: CRISK Batch Experimenter—Batch Experimenter screen

Figure 5-24 displays the entry GUI to the Batch Experimenter module. It has three buttons named “Run LOOCV Batch”, “Run Train-Test-LOOCV Batch”, and “Run Train-Test Batch”. Each button opens a command line window for a batch process for that type of experimentation. The “LOOCV Batch” process asks the user to select a dataset file, a predictors file, a predictors ranking file, and an output folder to store testing results. Both “Train-Test-LOOCV Batch” and “Train-Test Batch” processes ask the user to select a training dataset file, a test dataset file, a predictors file, a predictors ranking file, and an output folder. These input files are in CSV format. More information about these files can be found in Table 3-6.

Figure 5-25, Figure 5-26, and Figure 5-27 respectively provide high level views of the LOOCV Batch, Train-Test-LOOCV Batch, and Train-Test Batch experimentation processes. Each dataset file in CSV format is converted into a corresponding fuzzy ontology file before being experimented for different combination of predictors.

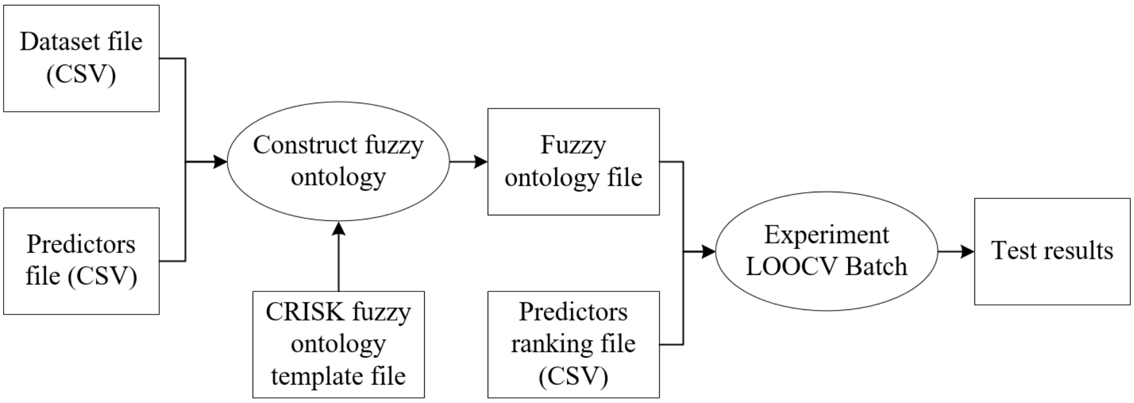


Figure 5-25: CRISK Batch Experimenter—LOOCV batch experimentation process

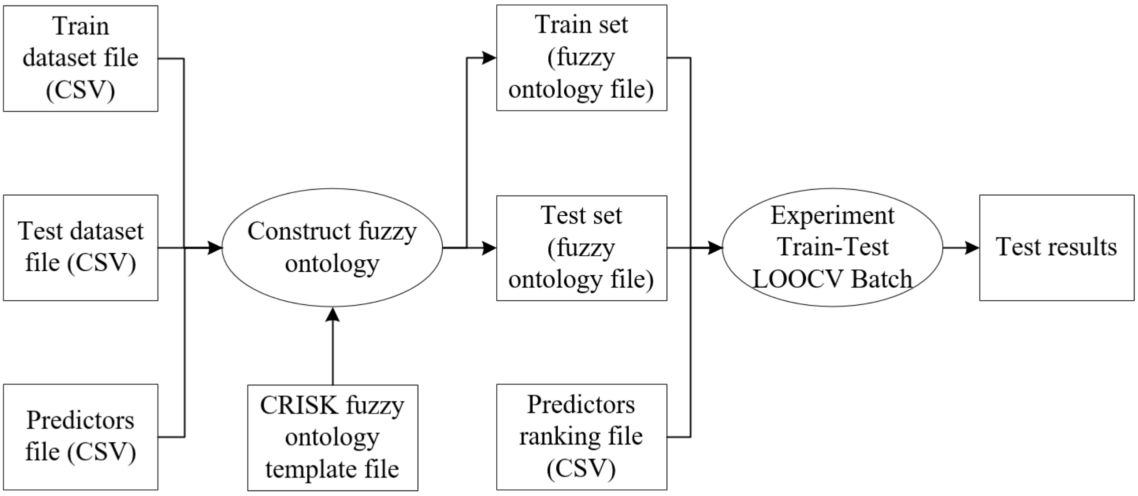

Train dataset file (CSV)
Test dataset file (CSV)
Predictors file (CSV)
Construct fuzzy ontology
CRISK fuzzy ontology template file
Train set (fuzzy ontology file)
Test set (fuzzy ontology file)
Predictors ranking file (CSV)
Experiment Train-Test LOOCV Batch
Test results

Figure 5-26: CRISK Batch Experimenter—Train-Test LOOCV batch experimentation process

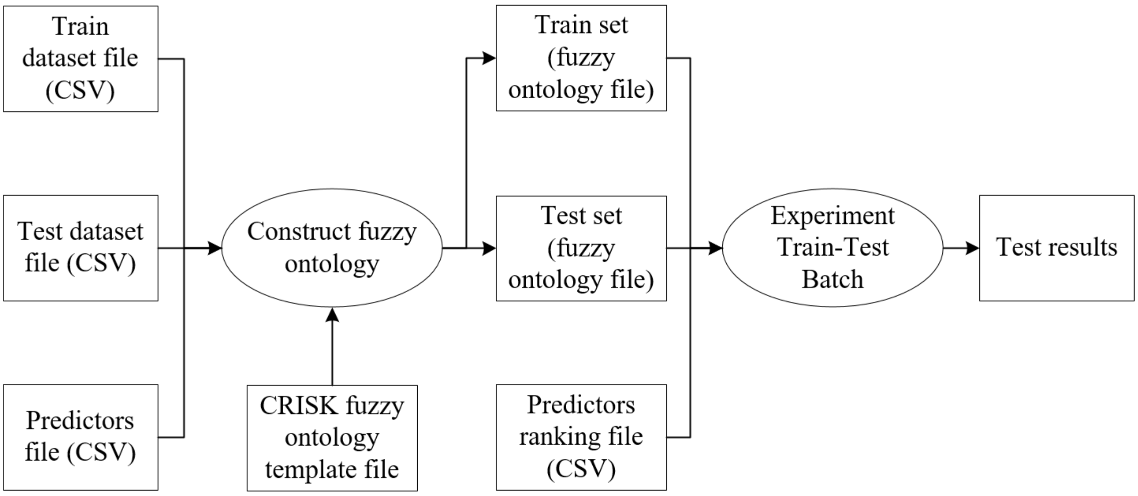

Train dataset file (CSV)
Test dataset file (CSV)
Predictors file (CSV)
Construct fuzzy ontology
CRISK fuzzy ontology template file
Train set (fuzzy ontology file)
Test set (fuzzy ontology file)
Predictors ranking file (CSV)
Experiment Train-Test Batch
Test results

Figure 5-27: CRISK Batch Experimenter—Train-Test batch experimentation process

Figure 5-28 provides a screenshot of an output folder from a batch experimentation. Each subfolder is named according to the number of predictors included in the experiment. Experimentation results for each combination of predictors are stored in that corresponding subfolder the same way as in the Experimenter module (Figure 5-21). The Batch Experimenter module also creates inside the output folder a summary file named "Grand_Summary_TestResults.xlsm". This is a macro-enabled Excel file to gather result data from each "TestResults.xlsm" file inside each subfolder.

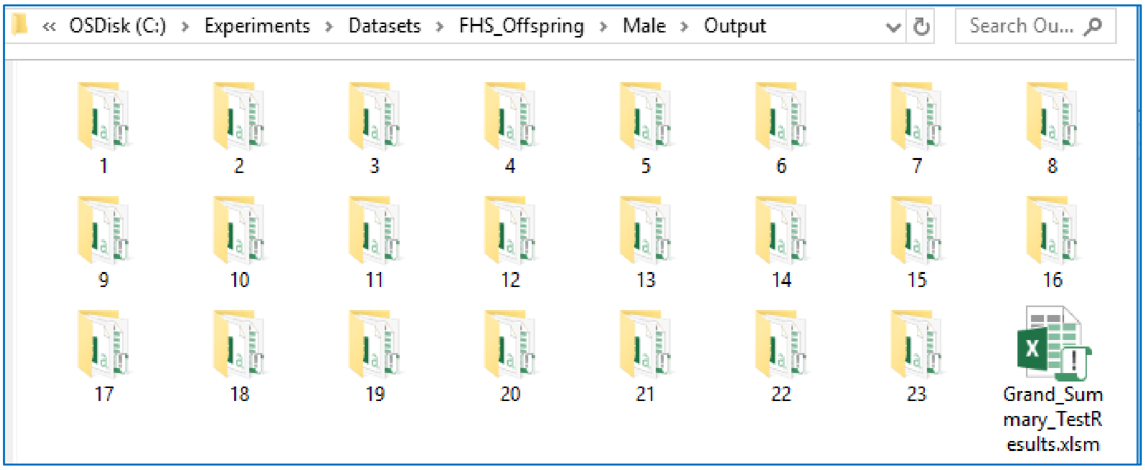
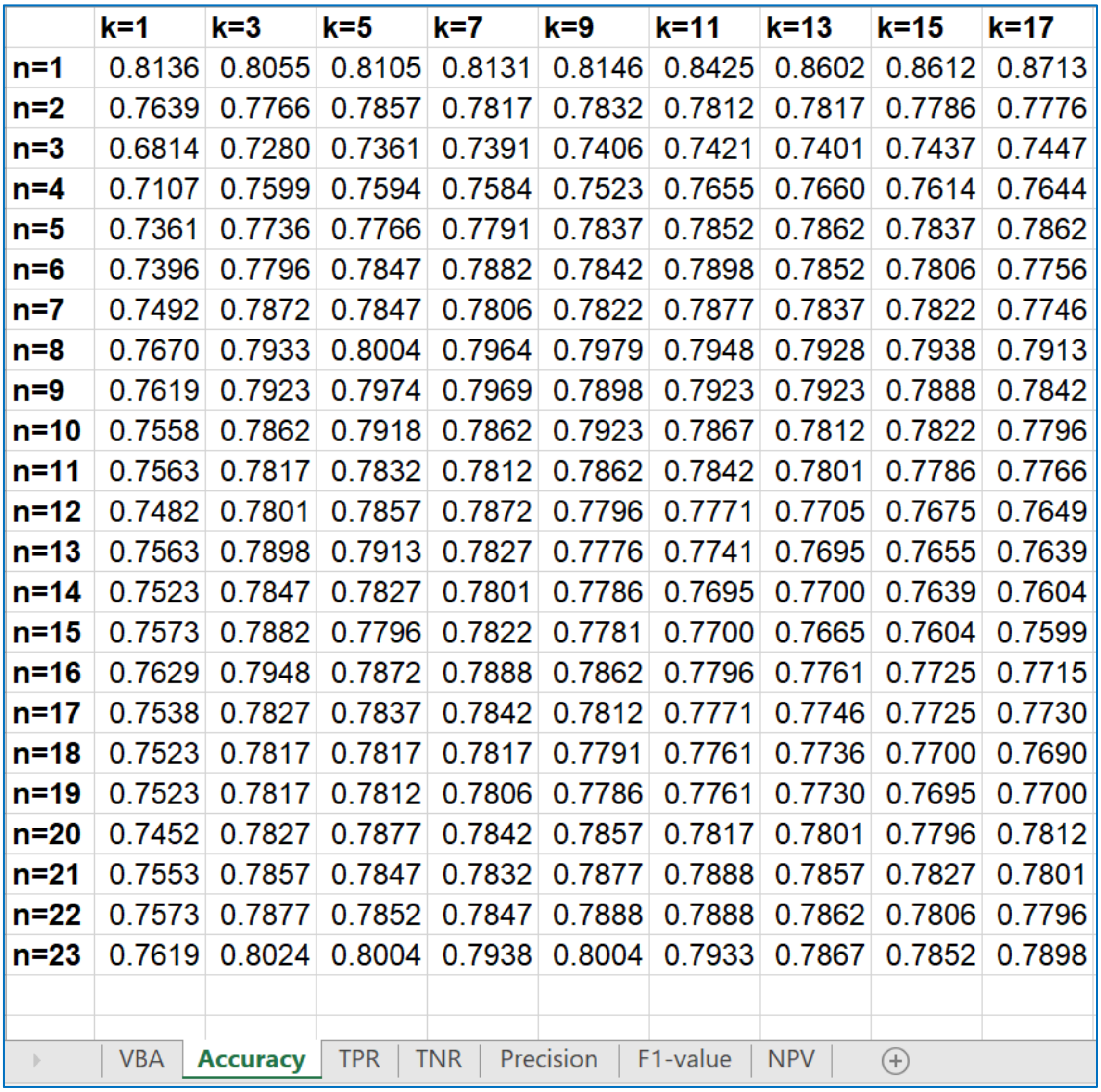

« OSDisk (C:) › Experiments › Datasets › FHS_Offspring › Male › Output

Figure 5-28: CRISK Batch Experimenter—Output folder

Figure 5-29 shows the content of a Grand Summary Test Results file. It collects prediction performance data into six sheets, "Accuracy", "TPR", "TNR", "Precision", "F_1-value", and "NPV". Each of these sheets provides the performance results based on the two dimensions experimented on, the number of predictors *n* and the number of nearest neighbours *k* for that metric.



	k=1	k=3	k=5	k=7	k=9	k=11	k=13	k=15	k=17
n=1	0.8136	0.8055	0.8105	0.8131	0.8146	0.8425	0.8602	0.8612	0.8713
n=2	0.7639	0.7766	0.7857	0.7817	0.7832	0.7812	0.7817	0.7786	0.7776
n=3	0.6814	0.7280	0.7361	0.7391	0.7406	0.7421	0.7401	0.7437	0.7447
n=4	0.7107	0.7599	0.7594	0.7584	0.7523	0.7655	0.7660	0.7614	0.7644
n=5	0.7361	0.7736	0.7766	0.7791	0.7837	0.7852	0.7862	0.7837	0.7862
n=6	0.7396	0.7796	0.7847	0.7882	0.7842	0.7898	0.7852	0.7806	0.7756
n=7	0.7492	0.7872	0.7847	0.7806	0.7822	0.7877	0.7837	0.7822	0.7746
n=8	0.7670	0.7933	0.8004	0.7964	0.7979	0.7948	0.7928	0.7938	0.7913
n=9	0.7619	0.7923	0.7974	0.7969	0.7898	0.7923	0.7923	0.7888	0.7842
n=10	0.7558	0.7862	0.7918	0.7862	0.7923	0.7867	0.7812	0.7822	0.7796
n=11	0.7563	0.7817	0.7832	0.7812	0.7862	0.7842	0.7801	0.7786	0.7766
n=12	0.7482	0.7801	0.7857	0.7872	0.7796	0.7771	0.7705	0.7675	0.7649
n=13	0.7563	0.7898	0.7913	0.7827	0.7776	0.7741	0.7695	0.7655	0.7639
n=14	0.7523	0.7847	0.7827	0.7801	0.7786	0.7695	0.7700	0.7639	0.7604
n=15	0.7573	0.7882	0.7796	0.7822	0.7781	0.7700	0.7665	0.7604	0.7599
n=16	0.7629	0.7948	0.7872	0.7888	0.7862	0.7796	0.7761	0.7725	0.7715
n=17	0.7538	0.7827	0.7837	0.7842	0.7812	0.7771	0.7746	0.7725	0.7730
n=18	0.7523	0.7817	0.7817	0.7817	0.7791	0.7761	0.7736	0.7700	0.7690
n=19	0.7523	0.7817	0.7812	0.7806	0.7786	0.7761	0.7730	0.7695	0.7700
n=20	0.7452	0.7827	0.7877	0.7842	0.7857	0.7817	0.7801	0.7796	0.7812
n=21	0.7553	0.7857	0.7847	0.7832	0.7877	0.7888	0.7857	0.7827	0.7801
n=22	0.7573	0.7877	0.7852	0.7847	0.7888	0.7888	0.7862	0.7806	0.7796
n=23	0.7619	0.8024	0.8004	0.7938	0.8004	0.7933	0.7867	0.7852	0.7898

Figure 5-29: CRISK Batch Experimenter—Grand Summary Test Results file

The "VBA" sheet (Figure 5-30) contains two buttons named "Reload" and "Refresh individual Test Result files and Reload". The first button refreshes data on the other seven sheets by reloading result data from each individual "TestResults.xlsm" file inside each subfolder. The second button does two things. It first refreshes data in each "TestResults.xlsm" file inside each subfolder. It then recollects data from these individual files to refresh data on the "Accuracy", "TPR", "TNR", "Precision", "F_1-value", and "NPV" sheets.

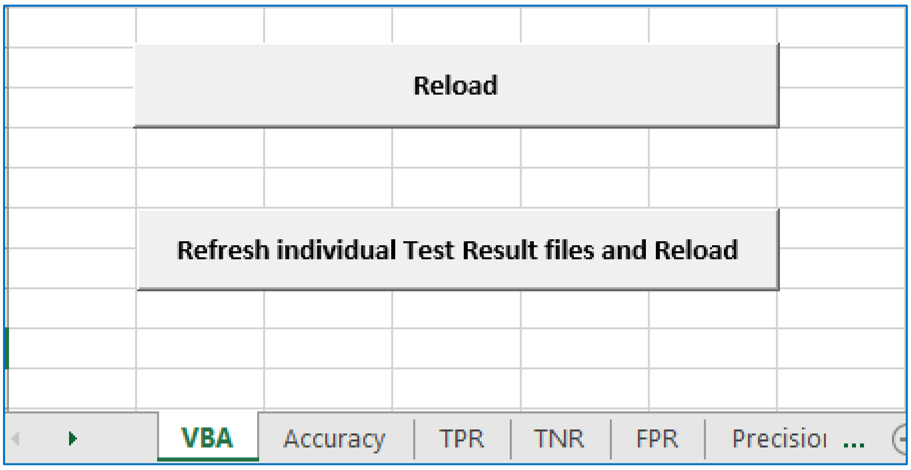
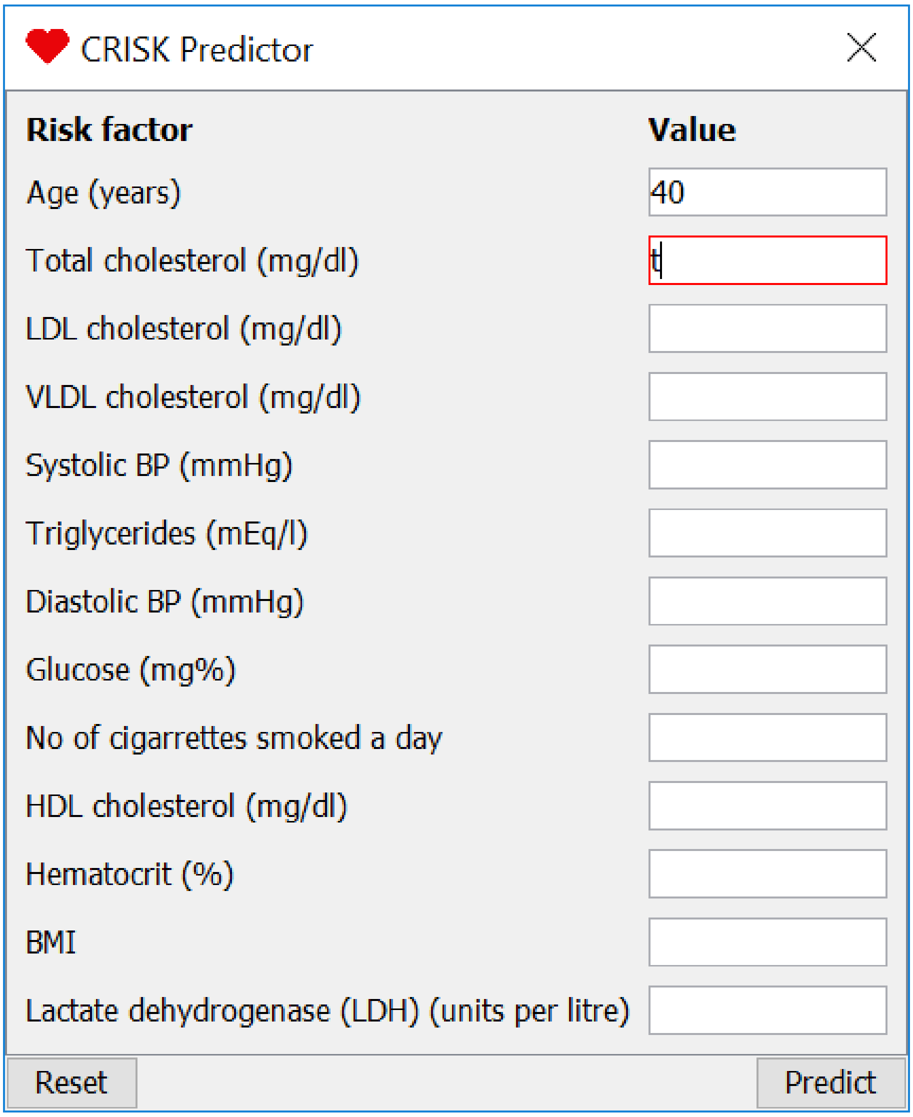


Figure 5-30: CRISK Batch Experimenter—Grand Summary Test Results file’s VBA sheet

5.5 CRISK PREDICTOR MODULE

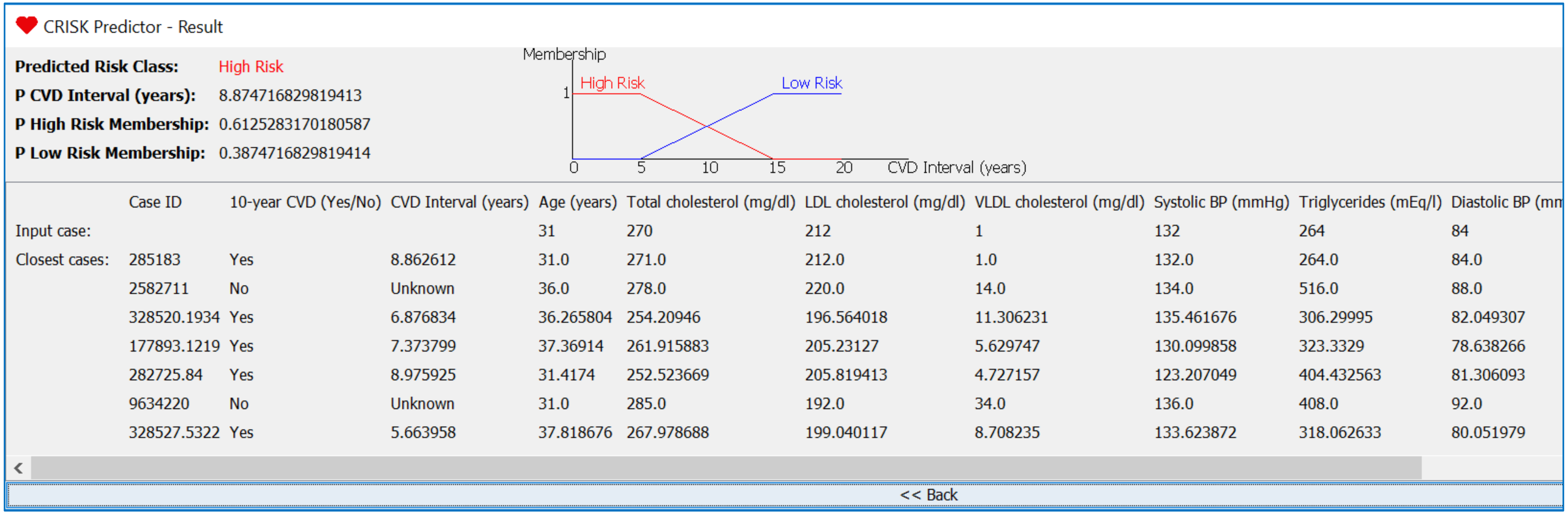
The Predictor module is used to give prediction for an input case (person) as to whether that person belongs to the “High CVD Risk” or the “Low CVD Risk” category.



Risk factor	Value
Age (years)	40
Total cholesterol (mg/dl)	t
LDL cholesterol (mg/dl)	
VLDL cholesterol (mg/dl)	
Systolic BP (mmHg)	
Triglycerides (mEq/l)	
Diastolic BP (mmHg)	
Glucose (mg%)	
No of cigarrettes smoked a day	
HDL cholesterol (mg/dl)	
Hematocrit (%)	
BMI	
Lactate dehydrogenase (LDH) (units per litre)	

Figure 5-31: CRISK Predictor—Input screen

Figure 5-31 shows the input screen of the module. There are 13 risk factors, resulting from experimentation done in Chapter 6 to find out the best combination of predictors and the number of nearest neighbours (n = 13, k = 7) for yielding the best prediction performance. When an invalid value is input, the developed input validation feature notifies and disallows this. Clicking on the “Reset” button clears all input textbox components on the screen. Clicking on the “Predict” button executes the CVD prediction process for the input case.



CRISK Predictor - Result

Predicted Risk Class: High Risk

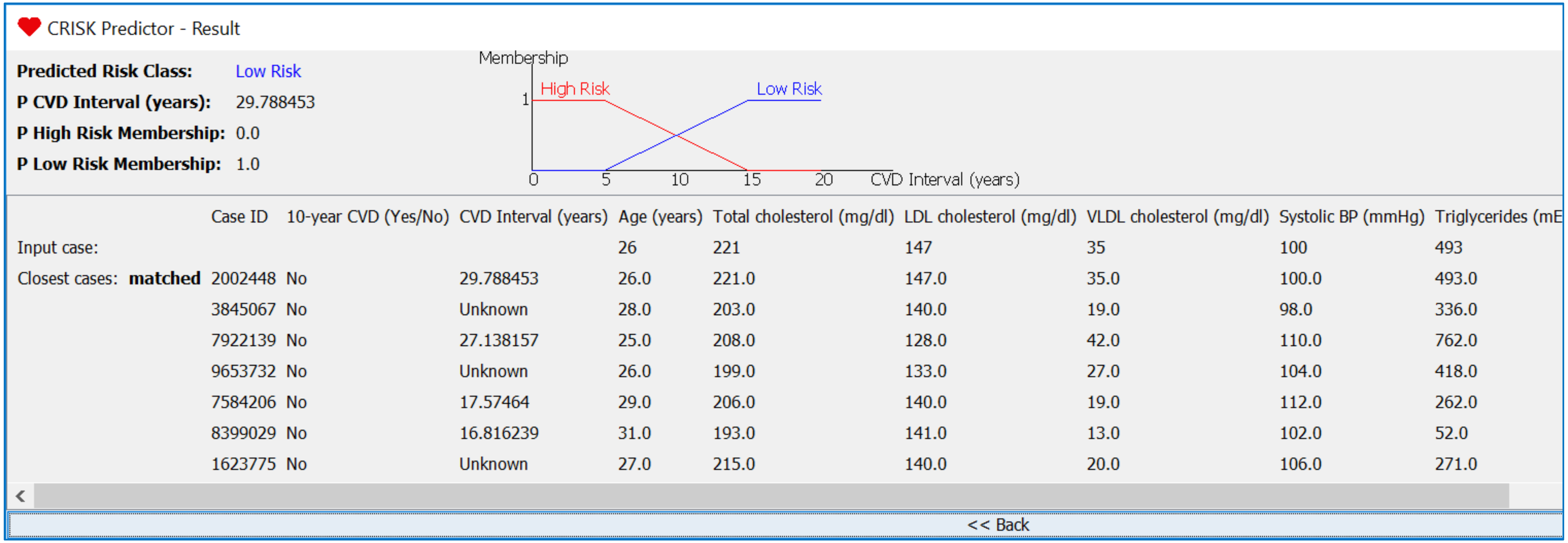
P CVD Interval (years): 8.874716829819413

P High Risk Membership: 0.6125283170180587

P Low Risk Membership: 0.3874716829819414

	Case ID	10-year CVD (Yes/No)	CVD Interval (years)	Age (years)	Total cholesterol (mg/dl)	LDL cholesterol (mg/dl)	VLDL cholesterol (mg/dl)	Systolic BP (mmHg)	Triglycerides (mEq/l)	Diastolic BP (mm
Input case:				31	270	212	1	132	264	84
Closest cases:	285183	Yes	8.862612	31.0	271.0	212.0	1.0	132.0	264.0	84.0
	2582711	No	Unknown	36.0	278.0	220.0	14.0	134.0	516.0	88.0
	328520.1934	Yes	6.876834	36.265804	254.20946	196.564018	11.306231	135.461676	306.29995	82.049307
	177893.1219	Yes	7.373799	37.36914	261.915883	205.23127	5.629747	130.099858	323.3329	78.638266
	282725.84	Yes	8.975925	31.4174	252.523669	205.819413	4.727157	123.207049	404.432563	81.306093
	9634220	No	Unknown	31.0	285.0	192.0	34.0	136.0	408.0	92.0
	328527.5322	Yes	5.663958	37.818676	267.978688	199.040117	8.708235	133.623872	318.062633	80.051979

Figure 5-32: CRISK Predictor—Prediction Result screen with no matched case



Predicted Risk Class: Low Risk

P CVD Interval (years): 29.788453

P High Risk Membership: 0.0

P Low Risk Membership: 1.0

		Case ID	10-year CVD (Yes/No)	CVD Interval (years)	Age (years)	Total cholesterol (mg/dl)	LDL cholesterol (mg/dl)	VLDL cholesterol (mg/dl)	Systolic BP (mmHg)	Triglycerides (mE
Input case:					26	221	147	35	100	493
Closest cases:	**matched**	2002448	No	29.788453	26.0	221.0	147.0	35.0	100.0	493.0
		3845067	No	Unknown	28.0	203.0	140.0	19.0	98.0	336.0
		7922139	No	27.138157	25.0	208.0	128.0	42.0	110.0	762.0
		9653732	No	Unknown	26.0	199.0	133.0	27.0	104.0	418.0
		7584206	No	17.57464	29.0	206.0	140.0	19.0	112.0	262.0
		8399029	No	16.816239	31.0	193.0	141.0	13.0	102.0	52.0
		1623775	No	Unknown	27.0	215.0	140.0	20.0	106.0	271.0

Figure 5-33: CRISK Predictor—Prediction Result screen, showing a matched case

Figure 5-32 shows the prediction result screen. Prediction outcomes include predicted Risk Class, predicted CVD Interval, predicted High Risk Membership, and predicted Low Risk Membership. In addition, a graph depicting the “High CVD Risk” and “Low CVD Risk” functions is also displayed to assist interpretation of the prediction outcomes. The result screen also displays the input case and the seven closest cases retrieved from the case base. Among those seven closest cases, if one matches the input case, there is a “matched” indicator shown next to that case, e.g. in Figure 5-33.

5.6 CHAPTER SUMMARY

The CRISK system, developed in Java, consists of four modules: Constructor, Experimenter, Batch Experimenter, and Predictor. The Constructor module converts a dataset in CSV format into a fuzzy ontology file in OWL 2 format that can be used in the Experimenter, Batch Experimenter and Predictor modules. The Experimenter module offers three types of experimentation, LOOCV, Train-Test-LOOCV, and Train-Test, to experiment a dataset for prediction performance based on different values of *k* (the number of nearest neighbours). The Batch Experimenter module adds another dimension, the combination of predictors, into the experimentation. The Predictor module provides prediction for an input case (a person) for whether that person belongs to “High CVD Risk” or “Low CVD Risk”.

Chapter 6

EXPERIMENTATION, RESULTS AND FINDINGS

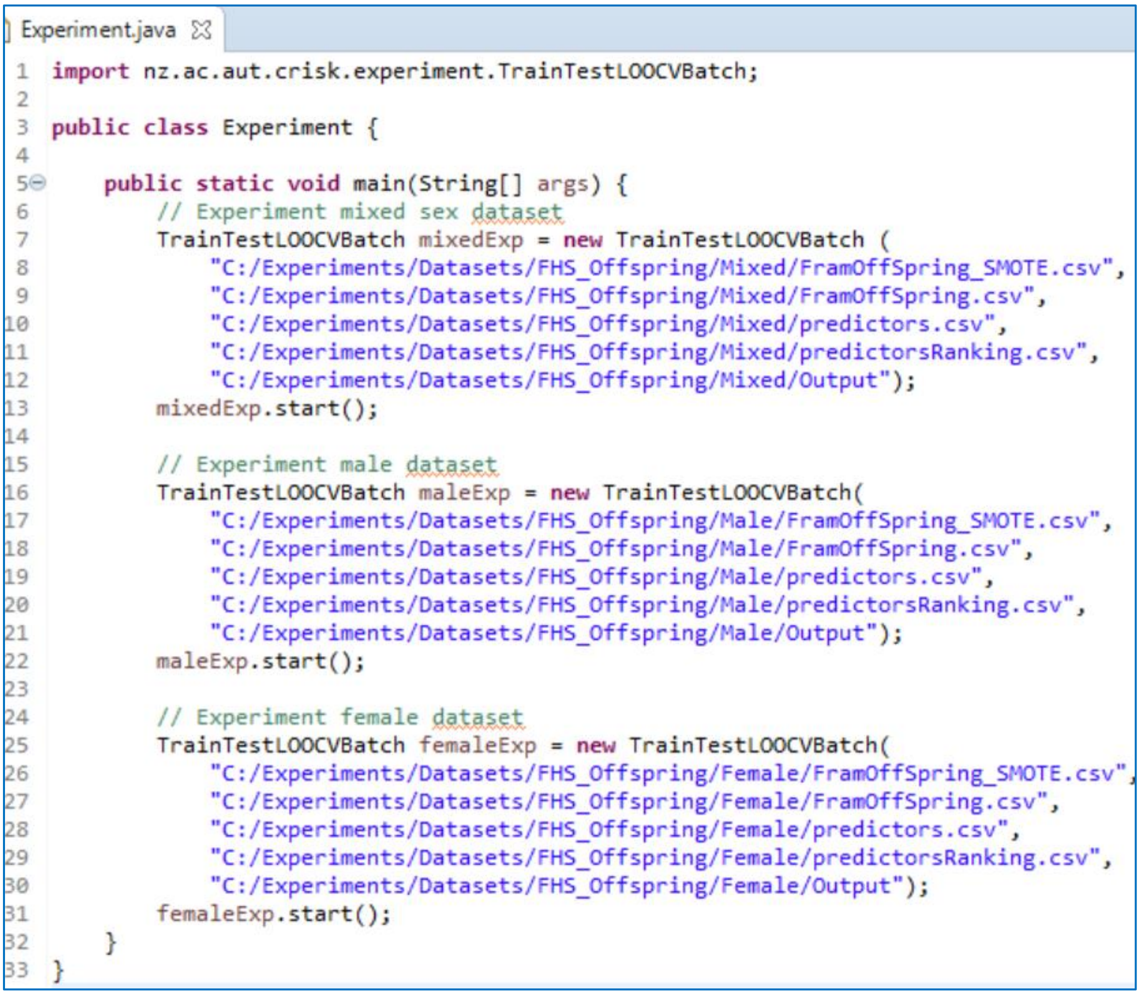
6.1 INTRODUCTION

This chapter reports on the experiments conducted using the CRISK system, the experimental results, and the overall findings of this research. The experimentation was carried out on the mixed sex dataset, the male dataset and the female dataset (described in section 3.8). Two dimensions were explored in these experiments: the combination of predictors and the number of nearest neighbours. The results of these experiments helped uncover whether separate prediction models for males and females should or should not be created. They also helped determine the predictors and the number of nearest neighbours that yielded the best prediction performance. These results informed the construction of the CRISK Predictor Module, described in section 5.5. Moreover, other findings including things that could be done to improve CVD prediction performance, why some popular risk factors mentioned in other prior research did not make it to the list of selected predictors in this research, and data distributions of the chosen risk factors, are also reported in this chapter.

This chapter together with Chapter 4 and Chapter 5 help answer the first three research questions (RQ1, RQ2, and RQ3). The answers are described in section 9.1.1.

6.2 EXPERIMENTATION

The experiments were designed in order to find the optimal combination of parameters (dataset, predictors, and number of nearest neighbours) that yields the best CVD risk prediction performance. Each of the three datasets, mixed sex dataset, male dataset, and female dataset, prepared from section 3.8 were experimented using the CRISK Batch Experimenter Module's Train-Test-LOOCV function (section 5.4). Though the GUI module of the CRISK Batch Experimenter could have been run three times for the three datasets to obtain the same results, a Java class named Experiment was developed to run the Batch Experimenter module as three threads concurrently to save execution time. This Java class was run directly in the Eclipse IDE. Figure 6-1 displays the Java code of the Experiment class. Explanation of the CSV input files for the experimentation can be found in section 3.8.11.



```java
import nz.ac.aut.crisk.experiment.TrainTestLOOCVBatch;

public class Experiment {

    public static void main(String[] args) {
        // Experiment mixed sex dataset
        TrainTestLOOCVBatch mixedExp = new TrainTestLOOCVBatch (
            "C:/Experiments/Datasets/FHS_Offspring/Mixed/FramOffSpring_SMOTE.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Mixed/FramOffSpring.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Mixed/predictors.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Mixed/predictorsRanking.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Mixed/Output");
        mixedExp.start();

        // Experiment male dataset
        TrainTestLOOCVBatch maleExp = new TrainTestLOOCVBatch(
            "C:/Experiments/Datasets/FHS_Offspring/Male/FramOffSpring_SMOTE.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Male/FramOffSpring.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Male/predictors.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Male/predictorsRanking.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Male/Output");
        maleExp.start();

        // Experiment female dataset
        TrainTestLOOCVBatch femaleExp = new TrainTestLOOCVBatch(
            "C:/Experiments/Datasets/FHS_Offspring/Female/FramOffSpring_SMOTE.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Female/FramOffSpring.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Female/predictors.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Female/predictorsRanking.csv",
            "C:/Experiments/Datasets/FHS_Offspring/Female/Output");
        femaleExp.start();
    }
}
```

Figure 6-1: Java class named Experiment developed for experimentation in this research

Table 6-1 below summarises the three datasets. They were all imbalanced with negative cases overwhelming the positive cases, especially in the female dataset. The ratios of negative cases to positive cases for mixed sex, male, and female datasets were 3850/221 = 17.42, 1819/155 = 11.74, and 2035/66 = 30.83, respectively. To balance these datasets for building case bases (training sets), SMOTE was applied to them with the percentages of 1,600%, 1,050%, and 3,000% respectively (refer to section 3.8.10).

Table 6-1: Summary of mixed sex, male, and female datasets

	Mixed sex dataset	**Male dataset**	**Female dataset**
Number of cases	4,071	1,974	2,101
Number of High Risk cases (positive)	221	155	66
Number of Low Risk cases (negative)	3,850	1,819	2,035
SMOTE percentage applied to balance the dataset	1,600%	1,050%	3,000%

6.3 RESULTS

TPR, TNR, Precision, F_1-value, and NPV results for each of the three datasets are reported in the Appendix Q, Appendix R, and Appendix S respectively. For each evaluation metric per each dataset (model), there is one table displaying results for each combination of *n* (predictors) and *k* (number of nearest neighbours). Each table is accompanied by a 3D graph plotting the results. In general, all models achieved higher Recall (TPR) than Precision. All models performed very well for TNR and NPV. Table 6-2 below reports the best performance results by metric for each dataset.

Table 6-2: The best performance results of each model for each metric

	Mixed sex dataset	**Male dataset**	**Female dataset**
TPR max	0.8824 (n=13, k=3, 13, or 15)	0.8903 (n=12, k=1 or 7)	0.6515 (n=8, k=1)
TNR max	0.9065 (n=33, k=5)	0.9247 (n=1, k=17)	0.9238 (n=29,k=17)
Precision max	0.2277 (n=18, k=7)	0.2560 (n=9, k=7)	0.1388 (n=6, k=9)
F_1-value max	0.3574 (n=13, k=7)	0.3966 (n=12, k=7)	0.2252 (n=5, k=15)
NPV max	0.9918 (n=13, k=13)	0.9881 (n=12, k=7)	0.9866 (n=5, k=15)

For the F_1-value, which is the metric that is favoured for decision making in this research (see justification in section 3.9.1), the male model performed the best, followed by the mixed sex model, and then the female model. The male model achieved the highest F_1-value of 0.3966 with the combination (n = 12, k = 7). The mixed sex model achieved the highest F_1-value of 0.3574 with the combination (n = 13, k = 7). The female model achieved the best F_1-value of 0.2252 with the combination (n = 5, k = 15).

Table 6-3: The performance results of each model based on the combinations of *n* and *k* that generated the best F_1-values where CI is the 95% confidence interval.

	Mixed sex dataset with **(*n* = 13, *k* = 7)**	**Male dataset with** **(*n* = 12, *k* = 7)**	**Female dataset with** **(*n* = 5, *k* = 15)**
TPR	0.8733 (CI = 0.0102)	0.8903 (CI = 0.0138)	0.6364 (CI = 0.0206)
TNR	0.8270 (CI = 0.0116)	0.7784 (CI = 0.0183)	0.8698 (CI = 0.0144)
Precision	0.2247 (CI = 0.0128)	0.2551 (CI = 0.0192)	0.1368 (CI = 0.0147)
F_1-value	0.3574 (CI = 0.0147)	0.3966 (CI = 0.0216)	0.2252 (CI = 0.0179)
NPV	0.9913 (CI = 0.0029)	0.9881 (CI = 0.0048)	0.9866 (CI = 0.0049)

Table 6-3 compares the performance results of the three models on the five metrics for the combinations of n and k that generated the best F_1-values. For each metric value, a confidence interval (CI) value of 95% confidence level is also provided. The confidence interval value was calculated according to Equation (15). In this formula, v is the performance metric value and N is the size of the dataset. For the mixed sex dataset, N =

4,071. For the male dataset, N = 1,974. For the female dataset, N = 2,101. The parameter value of 1.96 corresponds to the 95% confidence level.

$$interval\ (CI) = 1.96 \times \sqrt{\frac{v\,(1-v)}{N}} \tag{15}$$

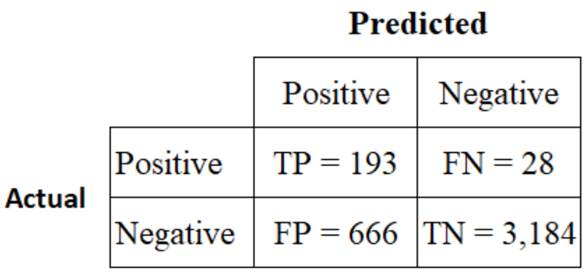
Table 6-4 displays the list of predictors that generated the best prediction performance for each model.

Table 6-4: Lists of predictors that generated the best prediction performance for each model

Model	No. of predictors	List of predictors
Mixed sex	13	Age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarrettes, HDL cholesterol, hematocrit, BMI, LDH
Male	12	Age, total cholesterol, first second volume, total vital capacity, LDL cholesterol, albumin, white blood count, glucose, triglycerides, total bilirubin, LDH, cigarettes
Female	5	Age, total cholesterol, first second volume, total vital capacity, LDL cholesterol

The first paragraph of section 6.4 will explain why the CRISK model was decided to be the mixed sex model with n = 13 and k = 7.

The confusion matrix of the mixed sex model when n = 13 and k = 7 is shown in Figure 6-2.



		Predicted	
		Positive	Negative
Actual	Positive	TP = 193	FN = 28
	Negative	FP = 666	TN = 3,184

Figure 6-2: Confusion matrix of the mixed sex model when n = 13 and k = 7

Table 6-5 displays the RMSE values of the mixed sex model, based on different *k* when the number of attributes is 13, for the CVD Interval prediction of cases that have both CVD Interval and predicted CVD Interval. The first column shows the number of nearest neighbours. The second, third, and last columns display the RMSE values for “All” cases, “High CVD Risk” cases, and “Low CVD Risk” cases that have both CVD Interval and predicted CVD Interval, respectively.

Table 6-5: RMSE values for CVD Interval prediction of cases that have both CVD Interval and predicted CVD Interval of the mixed sex model when n = 13

No. of nearest neighbours	RMSE—All (years)	RMSE—High (years)	RMSE—Low (years)
k=1	14.33	4.13	16.82
k=3	13.36	3.99	15.08
k=5	13.18	3.55	14.79
k=7	13.19	3.67	14.74
k=9	13.28	3.60	14.81
k=11	13.38	3.65	14.90
k=13	13.35	3.71	14.83
k=15	13.36	3.54	14.84
k=17	13.40	3.61	14.87

RMSE for the “High CVD Risk” cases is a lot smaller than RMSE for the “Low CVD Risk” cases. This could be mainly caused by the fact that the variance of the “High CVD Risk” cases is a lot smaller than the variance of the “Low CVD Risk” cases. In the mixed sex dataset, a “High CVD Risk” case has the CVD Interval value in [0.24, 9.96] years, while a “Low CVD Risk” case has the CVD Interval value in [10.01, 38.96] years. An additional cause could be that 2,950 cases out of 3,850 “Low CVD Risk” cases do not have CVD Interval (section 3.8.10), while all “High CVD Risk” cases have a CVD Interval (section 3.8.5). Having many unknown CVD Interval cases in the *k* closest cases retrieved could affect the predicted CVD Interval for a predicted “Low CVD Risk” case with μ_L = 1. In this case, the predicted CVD Interval is the average of CVD Intervals of the nearest “Low CVD Risk” cases that have CVD Interval—see the “Revise Closest Cases” algorithm (section 4.7). When *k* = 7, RMSE for the “High CVD Risk” cases is 3.67.

6.4 FINDINGS

Based on the experimentation results presented in Table 6-3, the CRISK model was decided to be the mixed sex model with the number of predictors *n* = 13 and the number of nearest neighbours *k* = 7, as this model performed better than the two separate gender specific models. For F_1-value, though the male model (F_1*-value* = 0.3966) performed a little better than the mixed sex model (F_1*-value* = 0.3574), the female model (F_1*-value* = 0.2252) performed far worse than the mixed sex model. The male model achieved slightly better than the mixed sex model in terms of F_1-value as a result of performing a little better in both Recall (0.8903 to 0.8733) and Precision (0.2551 to 0.2247). However, the mixed sex model did somewhat better than the male model in all other two metrics, TNR (0.8270 to 0.7784) and NPV (0.9913 to 0.9881). Although it was specified in section 3.9.1 that the F_1-value is favoured among the five performance metrics when comes to decision making, this is only applicable when the compared models are run against the same dataset. Therefore, the CRISK Predictor module (section 5.5) was decided to be developed as a mixed sex model.

In spite of using SMOTE to balance the datasets, the TPR (Recall), Precision, and F_1-value are still proportional to the P/N (positives/negatives) ratio to some degree. This is shown in Table 6-6, accompanied by a chart in Figure 6-3. The male model performed a little better than the mixed sex model in Recall, Precision, and thus F_1-value. The reason might just be that the male dataset had a higher P/N ratio. This examination of the P/N ratio reinforced the decision to construct the CRISK Predictor module as a mixed sex model instead of separate male and female models. The P/N ratio can be increased in the future by, for example, employing more positive cases to the existing case base.

Table 6-6: TPR, Precision, and F_1-value, and P/N ratio of the three datasets

	Mixed sex dataset	**Male dataset**	**Female dataset**
TPR	0.8733	0.8903	0.6364
Precision	0.2247	0.2551	0.1368
F_1-value	0.3574	0.3966	0.2252
P/N	0.0574	0.0852	0.0324

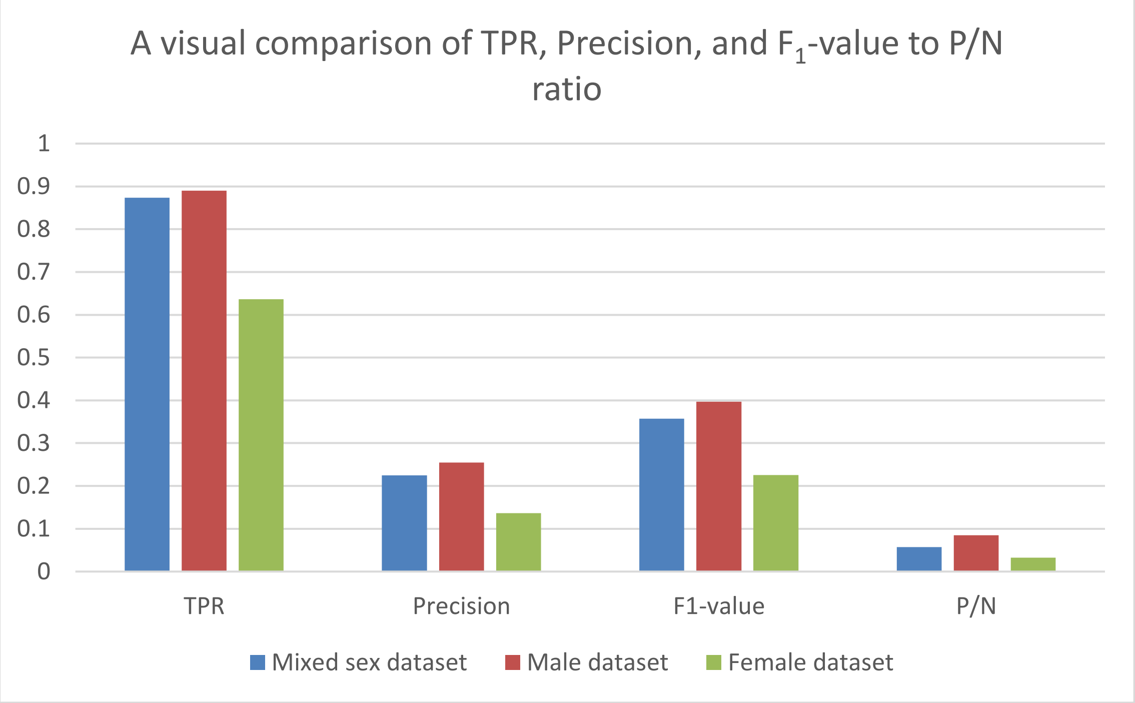


Figure 6-3: A visual comparison of TPR, Precision, and F_1-value to P/N ratio of the three datasets

Another way to increase the P/N ratio is to increase the prediction interval from 10 years to a longer period, such as 20 years. At first, it may look contradictory to the statistics [1] saying that CVD accounts for about 31% of deaths worldwide, while the percentage of positive cases in the mixed sex dataset in this study was only 221/4071=5.4%. The reason for this discrepancy was that the death statistics were recorded for people across their lifetime while this study observed people for only a 10-year period. If the time interval had been set for a longer period e.g. 20 years, the number of positive cases would likely have been more.

The TNR and NPV values are inversely proportional to the P/N ratio to some degree, especially the TNR. Though being inversely proportional to the P/N ratio, there is not much difference in the NPV values of the three datasets. This can be explained from the formula *NPV = TN / (TN + FN)*. For all three datasets, there are very few positive cases when compared with the number of negative cases. Therefore, FN is also very small compared with TN. Thus, NPV is very close to one for all three datasets. The TNR, NPV, and N/P values are provided in Table 6-7. The chart in Figure 6-4 allows a visual comparison of these metrics' values between the three datasets.

Table 6-7: TNR, NPV, and P/N ratio of the three datasets

	Mixed sex dataset	Male dataset	Female dataset
TNR	0.8270	0.7784	0.8698
NPV	0.9913	0.9881	0.9866
P/N	0.0574	0.0852	0.0324

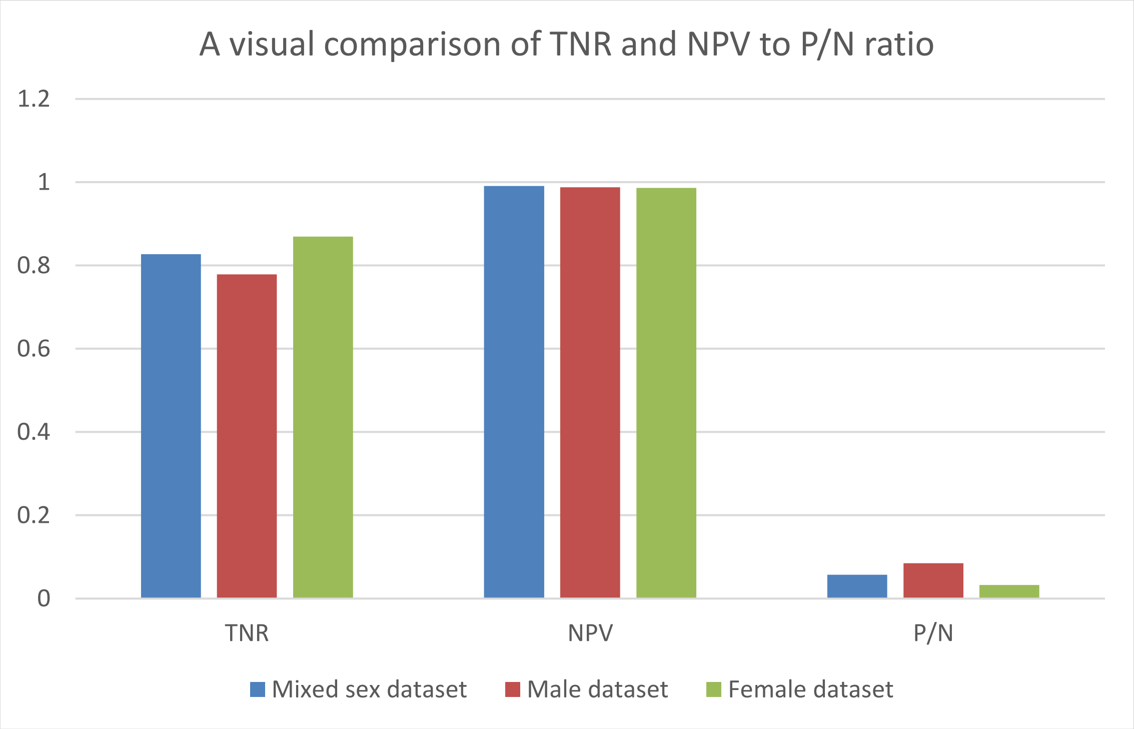


Figure 6-4: A visual comparison of TNR and NPV to P/N ratio of the three datasets

Interestingly, three popular risk factors, sex, diabetes, and smoking, did not make it to the list of predictors for all three models. These three risk factors appear in many existing models, for example Wilson et al. [30], the two models of D'Agostino et al. [31], the two models of Pencina et al. [32], QRISK 2 [73], 2013 PCE [74], Globorisk [75], and 2018 PCE [77]. The sex attribute is a predictor for all existing mixed sex models reviewed in this research, except the PREDICT-1° [76] model. The smoking attribute is a predictor for all of the existing models reviewed in this study. The diabetes attribute is used in almost all the models reviewed in this body of work.

The omission of these common predictors must have occurred for a reason. One possible reason could be that sex, diabetes, and smoking are indirect predictors (indirect causes). For the sex attribute, direct predictors are actually health parameters such as cholesterol, blood pressure, triglycerides, glucose etc. The diabetes attribute is in fact derived from glucose, which already made it to the list of predictors for the mixed sex model and the male model in this research. Similarly, for the smoking attribute, the cigarettes attribute (the number of cigarettes smoked a day) made it to the list of predictors for the mixed sex model and the male model instead. It seems reasonable that CVD outcomes would be more sensitive to how many cigarettes are smoked per day than simply whether that person smokes or not.

Another interesting finding was that not only sex, diabetes, and smoking, but also none of the nominal (categorical) attributes made it to the list of predictors for all three models. This may be because, unlike numerical data, nominal data has a very limited number of values, for example “male” and “female”, “yes” and “no”, or “black”, “blue”, and “brown”. Therefore, when coding for e.g. the Distance algorithm for the Retrieve algorithm (section 4.5), the number of distance values for a nominal attribute is also very limited. For example, for sex, whose value is either “male” or “female”, there are only two possible distance values. The first value is zero, meaning the two cases are of the same sex. The second value is set to e.g. one, meaning the two cases are different in sex. Therefore, the distance calculation for a nominal attribute is not as fine as for a numerical attribute. As a result, nominal attributes may not be as important as numerical attributes for a prediction model whose algorithm is developed based on KNN.

Weka was used to visualise data distributions of the 13 chosen attributes. Univariate attribution distribution graphs of these predictors are recorded in Appendix T. Among them, only HDL cholesterol is inversely proportional to CVD, i.e. the bigger HDL cholesterol (good cholesterol), the smaller the risk of having CVD. Age, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarettes, hematocrit, BMI, and LDH are all proportional to CVD. Total cholesterol is also proportional to CVD in general (Figure T-18). However, it is interesting seeing that at around one eighth area of the distribution graph, the trend goes in the opposite direction (inversely proportional to CVD). This trend may be explained by the fact that total cholesterol is made up of HDL, LDL, and VLDL cholesterols. At this area, the cases having higher levels of total cholesterol also had higher levels of HDL cholesterol, which is the good cholesterol, and this reduces the risk of having CVD. Summaries of the mixed sex dataset with the 13 risk factors chosen for CVD prediction and its “SMOTEd” one are respectively shown in Figure U-42 (Appendix U) and Figure V-43 (Appendix V).

6.5 CHAPTER SUMMARY

The developed CRISK prediction model achieved prediction performance results of TPR=0.8733 (CI=0.0102), TNR=0.8270 (CI=0.0116), Precision=0.2247 (CI=0.0128), F_1-value=0.3574 (CI=0.0147), and NPV=0.9913 (CI=0.0029) where CI is the 95% confidence interval. These results were achieved with the number of predictors $n = 13$ and the number of nearest neighbours $k = 7$ with the mixed sex dataset. Experimentation on different combinations of predictors, different numbers of nearest neighbours, and different datasets (mixed sex, male, and female) helped make the decision to construct the CRISK prediction model as a mixed sex model instead of separating into a male model and a female model, and to set $n = 13$ and $k = 7$. The 13 predictors selected are age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarettes, HDL cholesterol, hematocrit, BMI, and LDH. They are all numerical data.

A couple of findings were derived from the experimentation results. One important finding was that TPR (Recall), Precision, and F_1-value were proportional to the P/N ratio to some degree. This brings the hope to improve the prediction performance for the developed CRISK prediction model in the future by employing more positive cases into the case base. Another finding was that sex did not make it into the list of predictors. This finding was thought to be interesting as sex has been a really popular predictor appearing in all existing mixed sex CVD prediction models, that were reviewed in this study, except the PREDICT-1° [76] model. Among the 13 chosen attributes, only HDL cholesterol decreased CVD risk while the remaining predictors increased CVD risk as the predictor attributes’ values increased.

Chapter 7

EXTERNAL VALIDATION

7.1 INTRODUCTION

This chapter reports details of external validation in this study. The details include external dataset preparation, external dataset testing, results and findings from the results. For external dataset preparation the preparation steps are described. In addition, the quality of external datasets e.g. having the full list of predictor attributes or not, is also reported. The testing results and quality of external datasets helped derive useful findings for the developed CRISK prediction model when performing on external datasets.

This chapter helps answer RQ4. The answer is described in section 9.1.1.

7.2 EXTERNAL DATASET PREPARATION

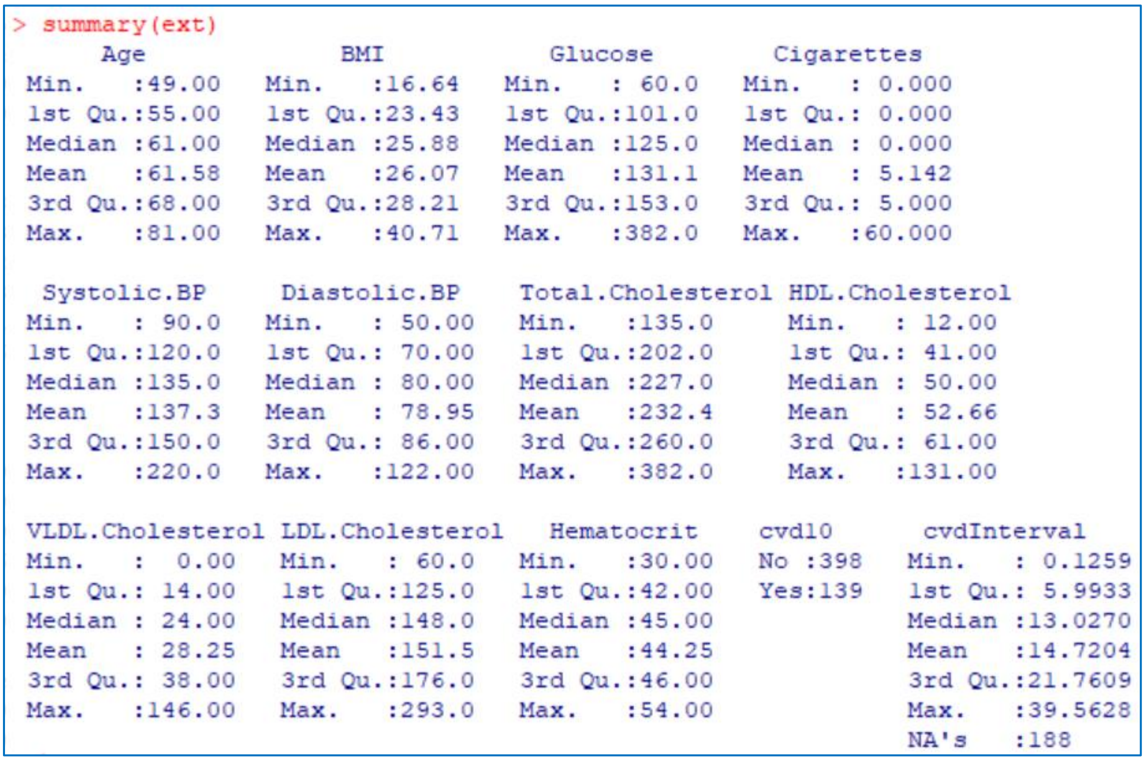
The external dataset was prepared from the FHS Original Cohort Exam 11. The reason for choosing this dataset as an external dataset was explained in section 3.6. Steps for preparing the external dataset were similar to those for preparing the mixed sex dataset from FHS Offspring Cohort Exam 1 as detailed in section 3.8. However, there were a couple of differences. The first one was that the list of predictor attributes was known. It was the 13 risk factors (age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarettes, HDL cholesterol, hematocrit, BMI, and LDH), chosen based on the work reported in Chapter 6, used for the CRISK Predictor module. However, as FHS Original Cohort Exam 11 did not contain triglycerides and LDH, the external dataset prepared contained only eleven predictors. The second one was, when calculating cvdInterval (section 3.8.5), cases having CVD = “No” and CVDYear < 10 were removed instead of the CVDYear < 15 limit applied for data preparation for the experiments detailed in Chapter 8. The reason was that, unlike when creating a case base, there was no need to know μ_H and μ_L. In this set of experiments there was only interest in labelling cvd10 as either “Yes” or “No”. Table 7-1 shows the content of the predictors file describing the predictor attributes of the external dataset.

Table 7-1: Predictors file to describe the predictors of the external dataset

Predictor Name	Predictor Description	Data Type	Value List
age	AGE	double	N/A
bmi	BMI	double	N/A
glucose	GLUCOSE	double	N/A
cigarettes	USUAL # OF CIGARETTES SMOKE NOW/EVER	double	N/A
sysBP	SYSTOLIC BLOOD PRESSURE	double	N/A
diaBP	DIASTOLIC BLOOD PRESSURE	double	N/A
totalChol	TOTAL CHOLESTEROL	double	N/A
hdlChol	HDL CHOLESTEROL	double	N/A
vldlChol	VLDL CHOLESTEROL	double	N/A
ldlChol	LDL CHOLESTEROL	double	N/A
hematocrit	HEMATOCRIT	double	N/A

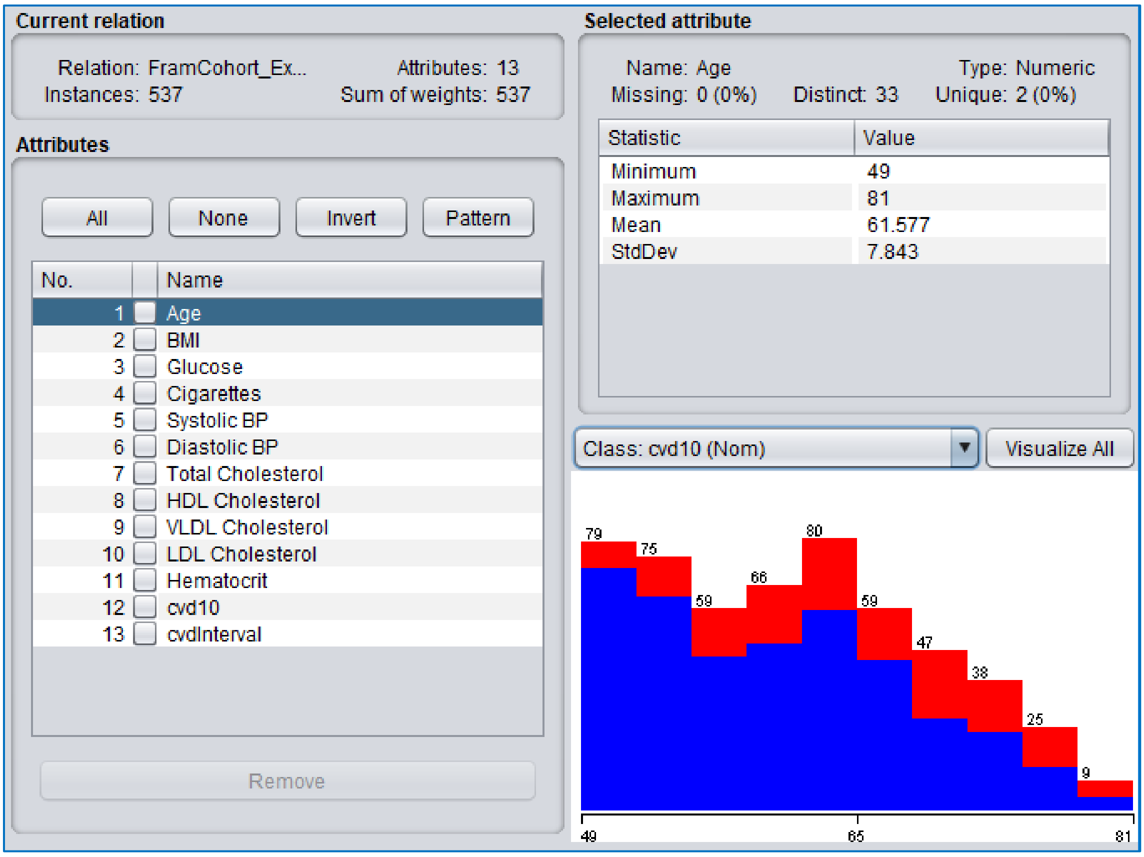
Figure 7-1 summarises the external dataset prepared, which has eleven predictors. The summary descriptive statistics were calculated using R (version 3.6.2). Three attributes in the external dataset (age, glucose, and HDL cholesterol) had values beyond those attributes’ value ranges in the case base (Appendix V). While the external dataset prepared had 537 cases; 231 cases had an age greater than the maximum age (62) in the case base; three cases had glucose greater than the maximum glucose (310) in the case base; one case had HDL cholesterol greater than the maximum HDL cholesterol (123) in the case base. As there were many cases in the external dataset having ages greater than the maximum age value in the case base, the prediction performance based on KNN would be affected as not having cases close enough to those out-of-range cases. Figure

7-2 shows the Weka univariate attribute distribution graph of Age in the prepared external dataset.



```
> summary(ext)
      Age             BMI            Glucose          Cigarettes
 Min.   :49.00   Min.   :16.64   Min.   : 60.0   Min.   : 0.000
 1st Qu.:55.00   1st Qu.:23.43   1st Qu.:101.0   1st Qu.: 0.000
 Median :61.00   Median :25.88   Median :125.0   Median : 0.000
 Mean   :61.58   Mean   :26.07   Mean   :131.1   Mean   : 5.142
 3rd Qu.:68.00   3rd Qu.:28.21   3rd Qu.:153.0   3rd Qu.: 5.000
 Max.   :81.00   Max.   :40.71   Max.   :382.0   Max.   :60.000

  Systolic.BP     Diastolic.BP    Total.Cholesterol HDL.Cholesterol
 Min.   : 90.0   Min.   : 50.00   Min.   :135.0     Min.   : 12.00
 1st Qu.:120.0   1st Qu.: 70.00   1st Qu.:202.0     1st Qu.: 41.00
 Median :135.0   Median : 80.00   Median :227.0     Median : 50.00
 Mean   :137.3   Mean   : 78.95   Mean   :232.4     Mean   : 52.66
 3rd Qu.:150.0   3rd Qu.: 86.00   3rd Qu.:260.0     3rd Qu.: 61.00
 Max.   :220.0   Max.   :122.00   Max.   :382.0     Max.   :131.00

 VLDL.Cholesterol LDL.Cholesterol   Hematocrit     cvd10       cvdInterval
 Min.   :  0.00   Min.   : 60.0   Min.   :30.00   No :398   Min.   : 0.1259
 1st Qu.: 14.00   1st Qu.:125.0   1st Qu.:42.00   Yes:139   1st Qu.: 5.9933
 Median : 24.00   Median :148.0   Median :45.00             Median :13.0270
 Mean   : 28.25   Mean   :151.5   Mean   :44.25             Mean   :14.7204
 3rd Qu.: 38.00   3rd Qu.:176.0   3rd Qu.:46.00             3rd Qu.:21.7609
 Max.   :146.00   Max.   :293.0   Max.   :54.00             Max.   :39.5628
                                                            NA's   :188
```

Figure 7-1: Summary of the external dataset in R



Current relation

Relation: FramCohort_Ex...   Attributes: 13
Instances: 537   Sum of weights: 537

Attributes

No.		Name
1	□	Age
2	□	BMI
3	□	Glucose
4	□	Cigarettes
5	□	Systolic BP
6	□	Diastolic BP
7	□	Total Cholesterol
8	□	HDL Cholesterol
9	□	VLDL Cholesterol
10	□	LDL Cholesterol
11	□	Hematocrit
12	□	cvd10
13	□	cvdInterval

Selected attribute

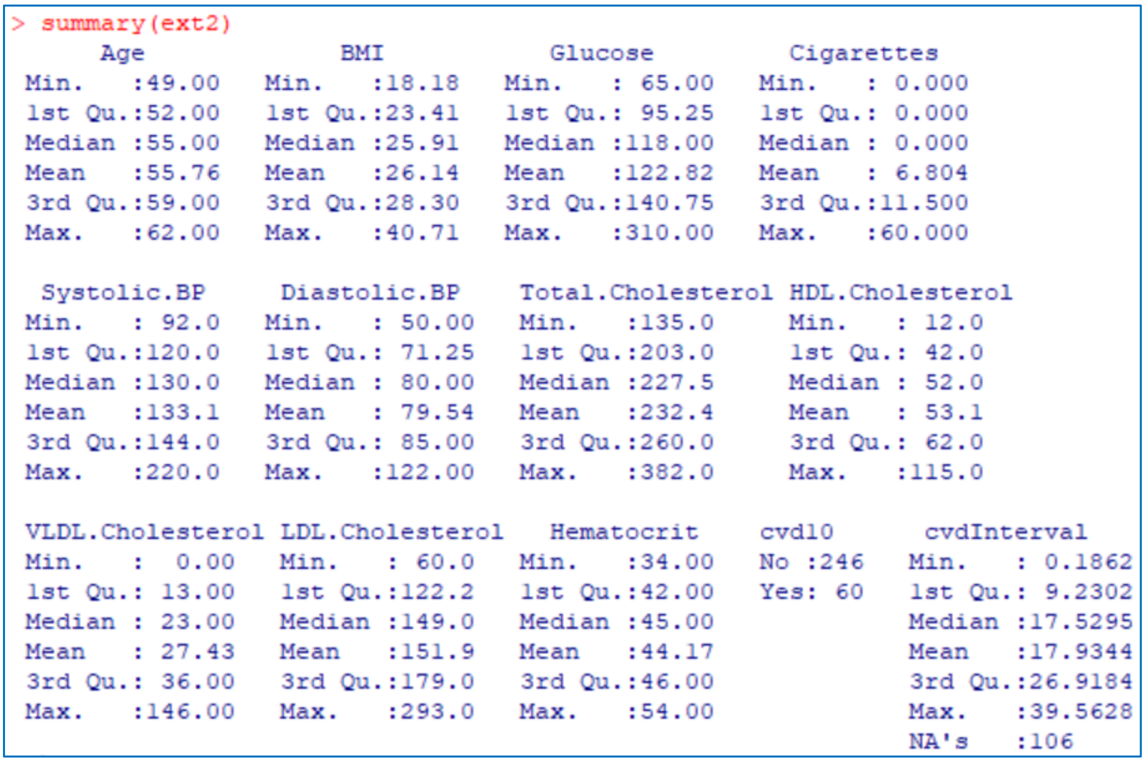
Name: Age   Type: Numeric
Missing: 0 (0%)   Distinct: 33   Unique: 2 (0%)

Statistic	Value
Minimum	49
Maximum	81
Mean	61.577
StdDev	7.843

Class: cvd10 (Nom)

Figure 7-2: Univariate attribute distribution of Age in the external dataset

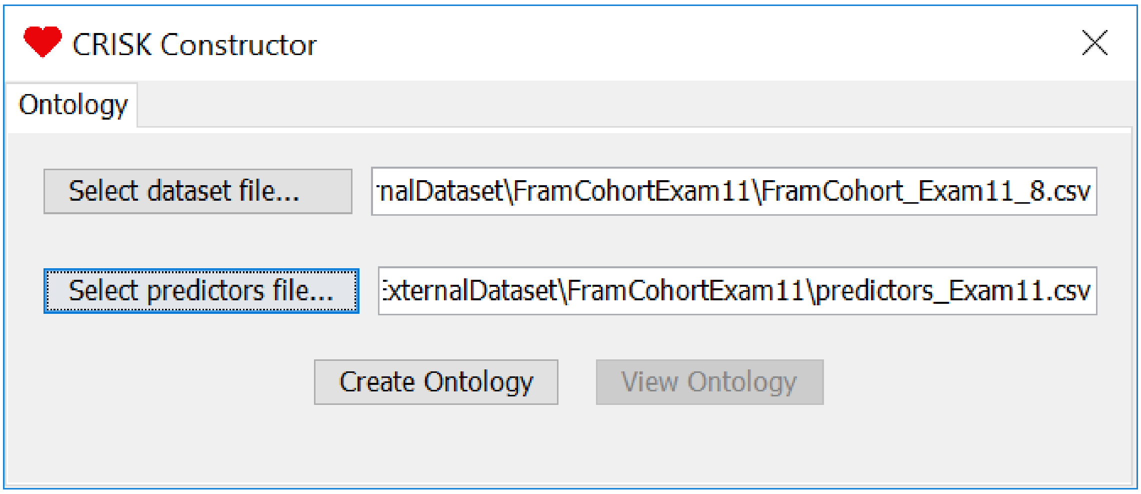
Therefore, another version of the external dataset was created, from the already prepared external dataset, by removing cases having predictor attribute values out of their ranges present in the case base. From here on, the first version of the external dataset will be referred to as External Dataset 1 and the second version as External Dataset 2. Both versions will be used to undertake a comparative evaluation of the prediction performance of the developed CRISK prediction model. Figure 7-3 gives a summary of External Dataset 2.



```
> summary(ext2)
      Age             BMI            Glucose          Cigarettes
 Min.   :49.00   Min.   :18.18   Min.   : 65.00   Min.   : 0.000
 1st Qu.:52.00   1st Qu.:23.41   1st Qu.: 95.25   1st Qu.: 0.000
 Median :55.00   Median :25.91   Median :118.00   Median : 0.000
 Mean   :55.76   Mean   :26.14   Mean   :122.82   Mean   : 6.804
 3rd Qu.:59.00   3rd Qu.:28.30   3rd Qu.:140.75   3rd Qu.:11.500
 Max.   :62.00   Max.   :40.71   Max.   :310.00   Max.   :60.000

  Systolic.BP     Diastolic.BP    Total.Cholesterol HDL.Cholesterol
 Min.   : 92.0   Min.   : 50.00   Min.   :135.0     Min.   : 12.0
 1st Qu.:120.0   1st Qu.: 71.25   1st Qu.:203.0     1st Qu.: 42.0
 Median :130.0   Median : 80.00   Median :227.5     Median : 52.0
 Mean   :133.1   Mean   : 79.54   Mean   :232.4     Mean   : 53.1
 3rd Qu.:144.0   3rd Qu.: 85.00   3rd Qu.:260.0     3rd Qu.: 62.0
 Max.   :220.0   Max.   :122.00   Max.   :382.0     Max.   :115.0

 VLDL.Cholesterol LDL.Cholesterol   Hematocrit      cvd10       cvdInterval
 Min.   :  0.00   Min.   : 60.0   Min.   :34.00   No :246    Min.   : 0.1862
 1st Qu.: 13.00   1st Qu.:122.2   1st Qu.:42.00   Yes: 60    1st Qu.: 9.2302
 Median : 23.00   Median :149.0   Median :45.00              Median :17.5295
 Mean   : 27.43   Mean   :151.9   Mean   :44.17              Mean   :17.9344
 3rd Qu.: 36.00   3rd Qu.:179.0   3rd Qu.:46.00              3rd Qu.:26.9184
 Max.   :146.00   Max.   :293.0   Max.   :54.00              Max.   :39.5628
                                                             NA's   :106
```

Figure 7-3: Summary of External Dataset 2 in R

The CRISK Constructor module was used to create two ontologies from the two external datasets. Figure 7-4 displays the screenshot of the ontology construction step for External Dataset 1. Ontology construction for External Dataset 2 was carried out in the same way.



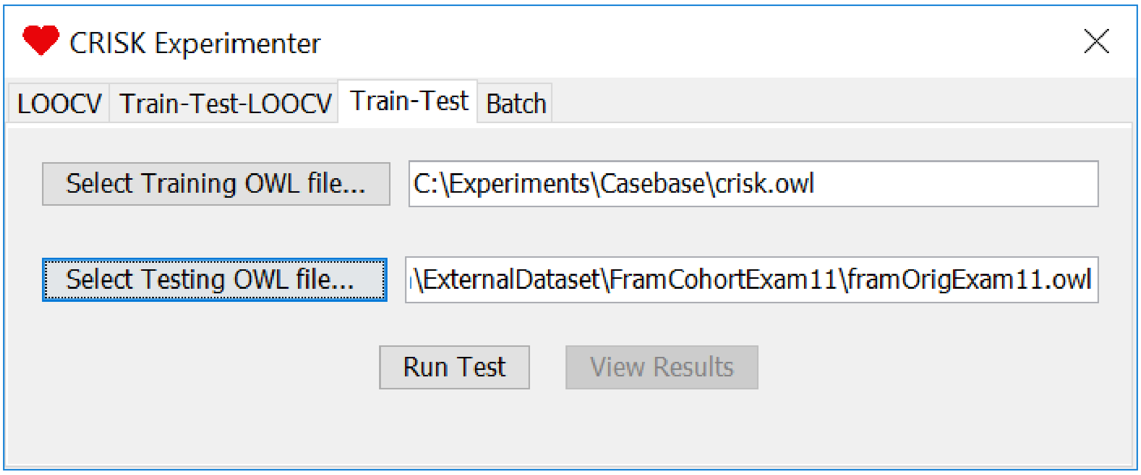
Select dataset file... `nalDataset\FramCohortExam11\FramCohort_Exam11_8.csv`

Select predictors file... `:xternalDataset\FramCohortExam11\predictors_Exam11.csv`

Figure 7-4: Ontology construction for External Dataset 1 using the CRISK Constructor module

7.3 External Dataset Testing

The CRISK Experimenter module (Train-Test experimentation type) was used to test the two external datasets. Figure 7-5 displays a screenshot of this testing step. The testing OWL file was an ontology file created in section 7.2. The training OWL file was the case base used for the CRISK Predictor module.



Select Training OWL file... `C:\Experiments\Casebase\crisk.owl`

Select Testing OWL file... `\ExternalDataset\FramCohortExam11\framOrigExam11.owl`

Figure 7-5: External dataset testing using the CRISK Experimenter module

7.4 Results

The prediction performance results of the developed CRISK prediction model on the two external datasets, External Dataset 1 and External Dataset 2, are displayed in Table 7-2.

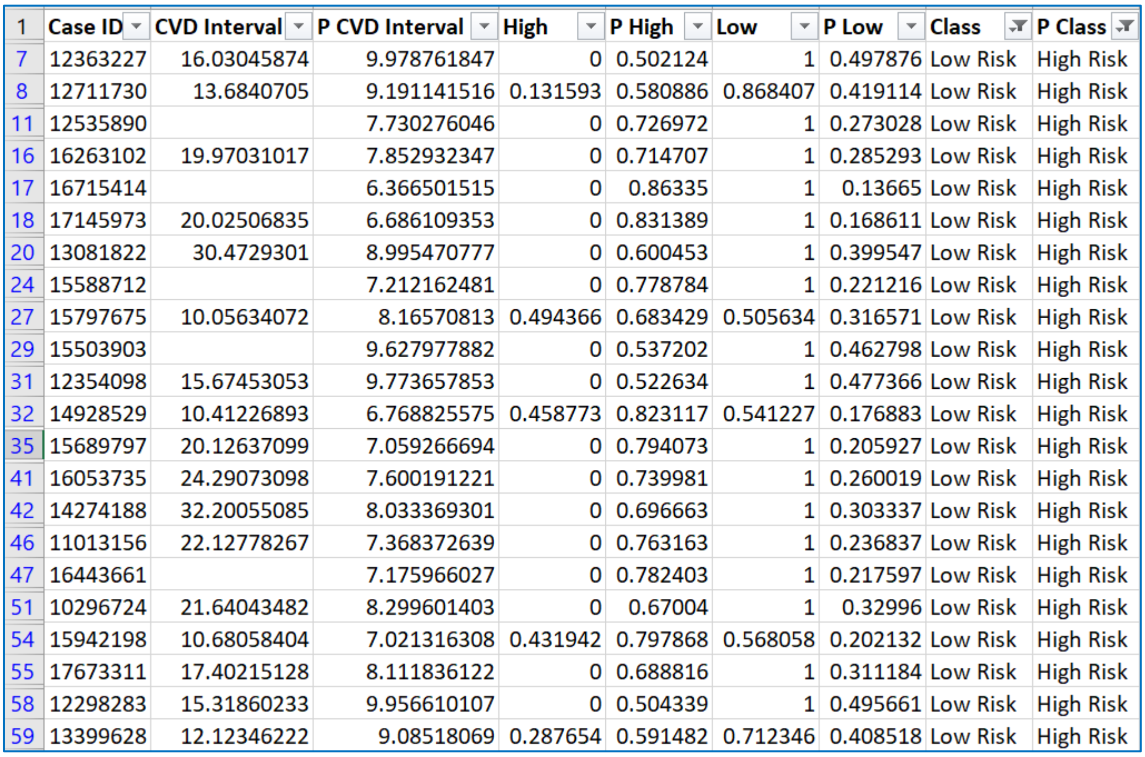
Table 7-2: External validation results where CI is the 95% confidence interval

	External Dataset 1	**External Dataset 2**
TPR	0.7410 (CI = 0.0371)	0.8167 (CI = 0.0434)
TNR	0.4472 (CI = 0.0421)	0.5041 (CI = 0.0560)
Precision	0.3189 (CI = 0.0394)	0.2866 (CI = 0.0507)
F_1-value	0.4459 (CI = 0.0420)	0.4242 (CI = 0.0554)
NPV	0.8318 (CI = 0.0316)	0.9185 (CI = 0.0307)

Prediction performance on External Dataset 2 could be concluded to be better than on External Dataset 1. It was better in terms of TPR, TNR, and NPV. However, it was worse for Precision and slightly worse for F_1-value. This can be explained as resulting from differences in the dataset sizes and P/N ratios. The External Dataset 1 had P = 139 and N = 398 while the External Dataset 2 had P = 60 and N = 246. As Precision is computed as TP / (TP + FP), Precision can achieve a high value when the TP is high. Though the TPR for External Dataset 1 is lower than for External Dataset 2, the TP value is a lot higher for External Dataset 1 than for External Dataset 2 as P is much higher in External Dataset 1. Moreover, P/N = 0.35 for External Dataset 1, which is greater than P/N = 0.24 for External Dataset 2. As noted in section 6.4, the F_1-value is proportional to the P/N ratio. Therefore, prediction performance on External Dataset 2 could be concluded to be better than on External Dataset 1 regardless of having a smaller F_1-value.

7.5 FINDINGS

The developed CRISK prediction model achieved good results on positive cases but terrible results on negative cases. CRISK performed reasonably well based on TPR (0.7410 and 0.8167) and very well on NPV (0.8318 and 0.9185). However, TNR was undesirable (0.4472 and 0.5041).



1	Case ID	CVD Interval	P CVD Interval	High	P High	Low	P Low	Class	P Class
7	12363227	16.03045874	9.978761847	0	0.502124	1	0.497876	Low Risk	High Risk
8	12711730	13.6840705	9.191141516	0.131593	0.580886	0.868407	0.419114	Low Risk	High Risk
11	12535890		7.730276046	0	0.726972	1	0.273028	Low Risk	High Risk
16	16263102	19.97031017	7.852932347	0	0.714707	1	0.285293	Low Risk	High Risk
17	16715414		6.366501515	0	0.86335	1	0.13665	Low Risk	High Risk
18	17145973	20.02506835	6.686109353	0	0.831389	1	0.168611	Low Risk	High Risk
20	13081822	30.4729301	8.995470777	0	0.600453	1	0.399547	Low Risk	High Risk
24	15588712		7.212162481	0	0.778784	1	0.221216	Low Risk	High Risk
27	15797675	10.05634072	8.16570813	0.494366	0.683429	0.505634	0.316571	Low Risk	High Risk
29	15503903		9.627977882	0	0.537202	1	0.462798	Low Risk	High Risk
31	12354098	15.67453053	9.773657853	0	0.522634	1	0.477366	Low Risk	High Risk
32	14928529	10.41226893	6.768825575	0.458773	0.823117	0.541227	0.176883	Low Risk	High Risk
35	15689797	20.12637099	7.059266694	0	0.794073	1	0.205927	Low Risk	High Risk
41	16053735	24.29073098	7.600191221	0	0.739981	1	0.260019	Low Risk	High Risk
42	14274188	32.20055085	8.033369301	0	0.696663	1	0.303337	Low Risk	High Risk
46	11013156	22.12778267	7.368372639	0	0.763163	1	0.236837	Low Risk	High Risk
47	16443661		7.175966027	0	0.782403	1	0.217597	Low Risk	High Risk
51	10296724	21.64043482	8.299601403	0	0.67004	1	0.32996	Low Risk	High Risk
54	15942198	10.68058404	7.021316308	0.431942	0.797868	0.568058	0.202132	Low Risk	High Risk
55	17673311	17.40215128	8.111836122	0	0.688816	1	0.311184	Low Risk	High Risk
58	12298283	15.31860233	9.956610107	0	0.504339	1	0.495661	Low Risk	High Risk
59	13399628	12.12346222	9.08518069	0.287654	0.591482	0.712346	0.408518	Low Risk	High Risk

Figure 7-6: Details of prediction results for External Dataset 2

Details of the test results (the CSV file for k = 7) for External Dataset 2 were examined to find a possible explanation for the poor performance observed on negative cases. Figure 7-6 shows a screenshot of FP cases from the test result file. The “P CVD Interval” column for all FP cases was examined. Its histogram is shown in Figure 7-7.

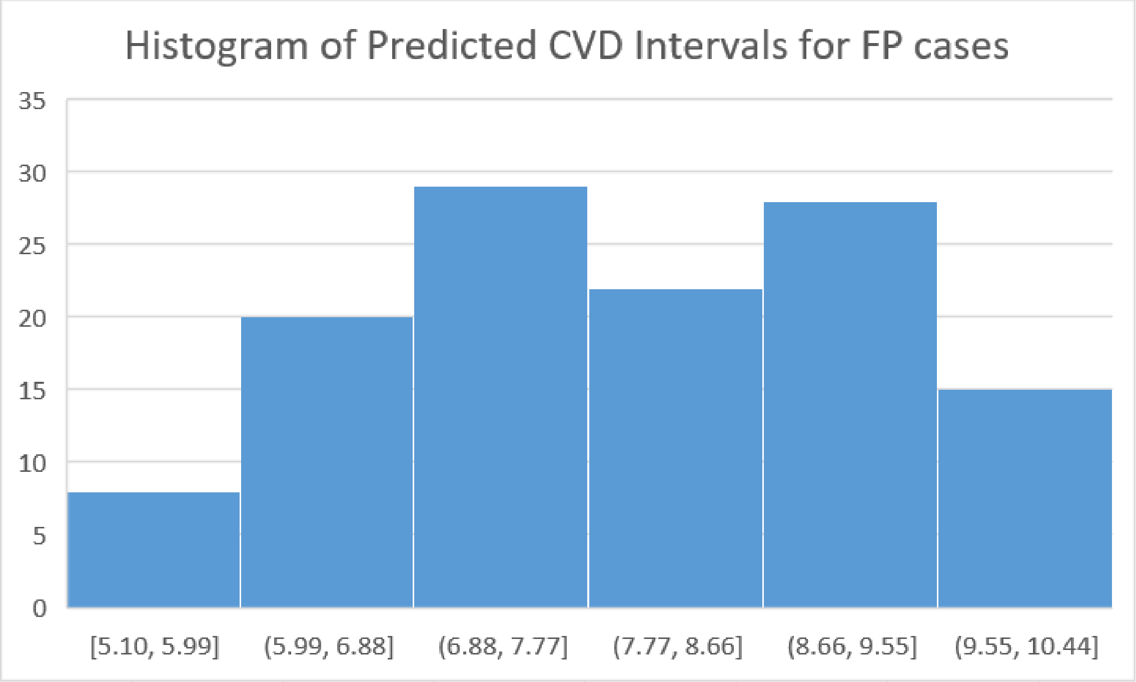


Figure 7-7: Histogram of Predicted CVD Intervals for FP cases in External Dataset 2

The predicted CVD Interval values for 122 FP cases in External Dataset 2 ranged from 5.1 to 9.98 years. It was good to see that none of FP cases had a predicted CVD Interval of less than 5 years. This also means that all FP cases were predicted to belong to both "High CVD Risk" and "Low CVD Risk" classes; however, μ_H is greater than μ_L. Sixty-five (more than a half) of the FP cases had a predicted CVD Interval greater than 7.77 years. Just a small shift to the right of these predicted CVD Interval values means those cases shift from being FP to TN and this shift would result in an increase in the TNR. Taking into account that two predictors (triglycerides and LDH) were missing in the external dataset, it could be possible that if these predictors had not been missing, the TNR may have been higher and therefore prediction performance would have been improved.

7.6 CHAPTER SUMMARY

Though two out of thirteen predictor attributes were missing from the external dataset, external validation achieved the TPR values of 0.7410 and 0.8167 for External Dataset 1 and External Dataset 2 (resulting from removing value-out-of-range cases from External Dataset 1), respectively. Besides, the resulting NPV values were 0.8318 and 0.9185 for External Dataset 1 and External Dataset 2, respectively.

Having out-of-range predictor attribute values (when compared to the ranges in the case base) in the test dataset decreases the prediction performance. This was illustrated by examining the prediction performance results of External Dataset 1 and External Dataset 2. The underlining cause of this decrease in predictive power seems to be related to the fact that the case base did not provide enough close enough neigbours to those value-out-of-range cases rather than directly to the CRISK model itself.

Chapter 8
DISCUSSION

To answer the six research questions stated in section 1.2, a system named CRISK was developed based on fuzzy ontology and CBR. The case base of the system is a fuzzy ontology constructed from the FHS Offspring Cohort Exam 1 dataset. The CRISK model achieved prediction performance results of TPR=0.8733, TNR=0.8270, Precision=0.2247, F_1-value=0.3574, and NPV=0.9913. Important risk factors were found to be age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarettes, HDL cholesterol, hematocrit, BMI, and LDH. External validation on the FHS Original Cohort Exam 11 dataset (External Dataset 2), which had two missing risk factors (triglycerides and LDH), achieved TPR=0.8167, TNR=0.5041, Precision=0.2866, F_1-value=0.4242, and NPV=0.9185.

This chapter introduces and discusses four discussion points related to the CRISK model developed in this research. The first one is whether the model has solved the problems of current widely used regression models. This discussion point also helps answer RQ5. The second point is about that CRISK supports personalised prediction. The third one is to compare CRISK to existing CVD prediction models. This third point also helps answer RQ6 (the answers to RQ5 and RQ6 are described in section 9.1.1). Finally, clinical applicability of the CRISK model is discussed in the fourth discussion point.

8.1 HAS THE CRISK MODEL SOLVED PROBLEMS WITH THE CURRENT REGRESSION PREDICTION MODELS?

The CRISK model designed and implemented as part of this research has possibly solved or at least partially solved five of the eight limitations of the current regression models stated in Table 2-3 in section 2.2.5. Table 8-1 provides justifications for this answer in detail by limitation.

Table 8-1: Which limitations of current regression models have possibly been solved by the CRISK model?

#	Limitation	Solved?	Explanation
1	Inaccuracy for individuals	Yes	The developed CRISK model first retrieves seven closest cases for an input case. It then generates prediction for the input case based on the CVD outcomes of these seven closest cases. Therefore, unlike regression models that are known to predict for populations [86], the CRISK model was designed to predict for individuals.
2	Inaccuracy for other cohorts	N/A	There is not yet an answer to this problem. This could not be addressed in this research because there was insufficient testing undertaken on external datasets and insufficent data. The only external dataset collected for this research was the FHS Original Cohort Exam 1. However, this dataset had two missing risk factors, triglycerides and LDH. In addition, the FHS Original Cohort belongs to the same racial group as the case base constructed from FHS Offspring Cohort, so there was a lack of diversity in the cases available. In future, the CRISK model should also be tested on cohorts from different racial and geographical groups.
3	Inflexibility of handling intervention	No	At this stage, the CRISK model does not handle intervention factors such as the person quits smoking, or the person starts having treatment for CVD or related conditions.
4	Requirement of complete clinical data	Partially Yes	The CRISK model was designed to allow for missing data. Missing data can be both in the case base and in the input case. While the system can make a prediction, missing data may affect the prediction accuracy. This is reasonable. There is currently no automatic mechanism in the system for handling missing data. The system currently just ignores these missing risk factors and retrieves the closest cases based on the risk factors that have values in the input case. Therefore, doctors are the ones responsible for making judgement calls based on the system's CVD prediction outcomes,

			keeping in mind the potential impact of the missing data. More information about missing data handling is given in section 9.2.7.
5	Deficiency of handling inaccurate data or result	Partially Yes	In the CRISK system, the case base is a fuzzy ontology capable of storing both crisp and fuzzy data. When fuzzified, an inaccurate crisp data value could still belong to the correct fuzzy set. For example, a person smokes 45 cigarettes per day but this is inaccurately recorded as 55 cigarettes per day. When fuzzified, smoking 45 cigarettes per day or smoking 55 cigarettes per day could belong to the same fuzzy set e.g. “heavy smoking”. Currently, in the CRISK system, only the CVD prediction outcomes are fuzzy. All values of risk factors are stored in the case base as crisp values. In addition, the current version of the Retrieve algorithm also only works with crisp-valued risk factors. However, with not much effort, the Retrieve algorithm could be updated to work with fuzzy-valued predictors in the future. The CRISK system displays CVD prediction results including membership values of two fuzzy sets “High CVD Risk” and “Low CVD Risk”. Therefore, a case can be predicted to belong to both “High CVD Risk” and “Low CVD Risk” with, most of the time, different membership values. As a result, a FP or FN case may still “be paid attention to” as having a positive membership value belongs to the other fuzzy set (the correct set). Real examples for this were mentioned in section 7.5.
6	Deficiency of handling vagueness of data or result	Partially Yes	Currently the CRISK system works with type-1 fuzzy ontology and therefore can handle vagueness of data or result. However, as noted in Limitation #5, in the current CRISK system, only CVD outcomes are fuzzy. Hence this limitaton is only partially addressed.

7	Deficiency of handling uncertainty of data or result	No	In its current version, the CRISK system is not capable of working with type-2 fuzzy ontology, which is known to handle uncertainty of data.
8	Poor explanatory capacity	Yes	The CRISK system displays the seven closest cases together with prediction outcomes for the input case in order to enhance the system’s explanatory capacity. This extra detail helps the user interpret the CVD prediction outcomes generated by the system.

8.2 PERSONALISED PREDICTION

The CRISK model built based on CBR is for personalised prediction. Unlike many existing models e.g. regression models, a formula is created based on the whole dataset. After that, the same formula is used to give predictions for new people. The formula was created to separate as many people in the dataset as possible into correct classes, focusing on the whole population rather than an individual. Therefore, such models are considered to predict for populations [86]. On the other hand, the CRISK model always tries to retrieve the seven closest cases to the input case first. Then, it generates CVD prediction outcomes for the input case based on CVD outcomes of these seven closest cases. Therefore, it can be considered as a model for personalised prediction (individualised prediction) rather than as population based.

Personalised prediction should be the focus for building a disease prediction model. However, so far, it has not been given enough attention, especially in CVD prediction. None of the CVD prediction models reviewed in this research were designed to be a pesonalised model. Recently, in 2019, there was a PhD dissertation [185] completed that introduced a competing-risk adjusted model called LIFE-CVD to estimate the benefit from lipid-lowering, blood pressure-lowering, and anti-thrombotic therapy and smoking cessation in people without prior CVD. Individual therapy-benefit is expressed as 10-year risk reduction, lifetime-risk reduction, and CVD-free life expectancy. In essence, this is also a regression model adding individualised therapy-benefit estimation. Different individual may have different therapy and therefore may have different CVD risk reduction.

8.3 COMPARISON TO EXISTING MODELS

Table 8-2 gives a general comparison of the developed CRISK model to three existing prediction models, D’Agostino et al. [31], PREDICT-1° [76], and 2018 PCE [77]. Reasons for choosing these existing models were given in section 3.9.3. All three existing models were built using the Cox Proportional Hazards regression method [37]. None of these studies performed external validation when their models were published. However, they can all be externally validated as long as there is an external dataset having the same risk factors, as their CVD risk equations and tools are publicly available for download and use.

Table 8-2: A general comparison of CRISK and three chosen existing CVD prediction models

Model	Risk factors	Method	Prediction Interval	Prediction Performance	External Validation	Comments
CRISK	Age, total cholesterol, LDL-C, VLDL-C, SBP, triglycerides, DBP, glucose, cigarettes, HDL-C, hematocrit, BMI, LDH	Fuzzy ontology CBR	10 years	TPR=0.8733, TNR=0.8270, F_1-value=0.3574, and NPV=0.9913	Yes	External validation on a dataset with two missing risk factors achieved TPR=0.8167, TNR=0.5041, F_1-value=0.4242, and NPV=0.9185.
D'Agostino et al. [31]	Age, sex, SBP, treatment for hypertension, smoking, diabetes, total cholesterol, HDL-C	Cox proportional-hazards modeling	10 years	AUC = 0.763 for men and 0.793 for women	No	The excel based CVD risk calculator tool is available for download.
PREDICT-1° [76],	Age, ethnicity, NZ index of socioeconomic deprivation, family history of premature CVD, smoking, diabetes, history of atrial fibrillation, SBP, TC/HDL-C, blood pressure lowering medication, lipid lowering medication, antithrombotic medication	Cox proportional hazard modelling	5 years	The slopes of regression lines comparing predicted and observed total cardiovascular disease risk in deciles were 0.98 (95% CI 0.93–1.02) for women and 0.98 (0.98–1.01) for men.	No	The model (risk equation) was designed to be externally validated by other studies.
2018 PCE [77]	Age, sex, race, total cholesterol, HDL-C, SBP, treatment for high blood pressure, diabetes, smoking	Cox proportional hazard modelling	10 years and lifetime	No explicit results but stating that the updated 2018 PCE improved accuracy among all race and sex groups comparing to the 2013 PCE.	No	The CVD Risk calculator tool named "ACC/AHA Excel-Based CV Risk Calculator" is available for download.

Nevertheless, for prediction performance comparison against CRISK, only the D'Agostino model and the 2018 PCE model were selected. The reason for choosing these two existing models was twofold. First, all their risk factors were available in the FHS Offspring Cohort Exam 1 dataset, a dataset that was available in this research. Second, their prediction intervals are 10 years, the same as CRISK. On the other hand, PREDICT-1° was not chosen for performance comparison because it not only predicts CVD within 5 years but also uses risk factors that were not available in the datasets of this research.

The test dataset preparation process for performance comparison is summarised in Figure 8-1. First of all, a dataset was prepared to have all the risk factors used by the three models. The way to do this was to collect more risk factors that are used by the existing models into the dataset that was prepared from the FHS Offspring Cohort Exam 1 dataset for CRISK in this research. For "race" (used by 2018 PCE), this risk factor was added, and its value was set to "WH". The 2018 PCE model uses "race" and accepts two values, "AA" (for African Americans) and "WH" (for whites or others). The FHS Offspring Cohort is known as a Caucasian cohort and therefore the value "WH". After all additional columns were added, it was checked to remove cases having missing values; however, there was no such case. Next, as each of the existing models has different acceptable ranges of risk factor values, two comparison scenarios were defined. Test dataset preparation was carried out according to these two scenarios. An example of acceptable ranges of values is that, 2018 PCE only accepts total cholesterol values from 130 to 320 mg/dl.

The first scenario was to have three test datasets for three models. The test dataset for CRISK was the same as the one prepared from the first step above because all risk factor values were within acceptable ranges for CRISK. In fact, the model takes any ranges of values as there was not any acceptable range defined for the model in this research. This dataset had 4,071 cases (P = 221, N = 3,850). The test dataset for the D'Agostino model was formed by removing cases outside acceptable ranges of values for the model. This resulted in a dataset of 2,841 cases (P = 211, N = 2,630). In the same way, the test dataset for the 2018 PCE model was created and had 1,470 cases (P = 166, N = 1,304).

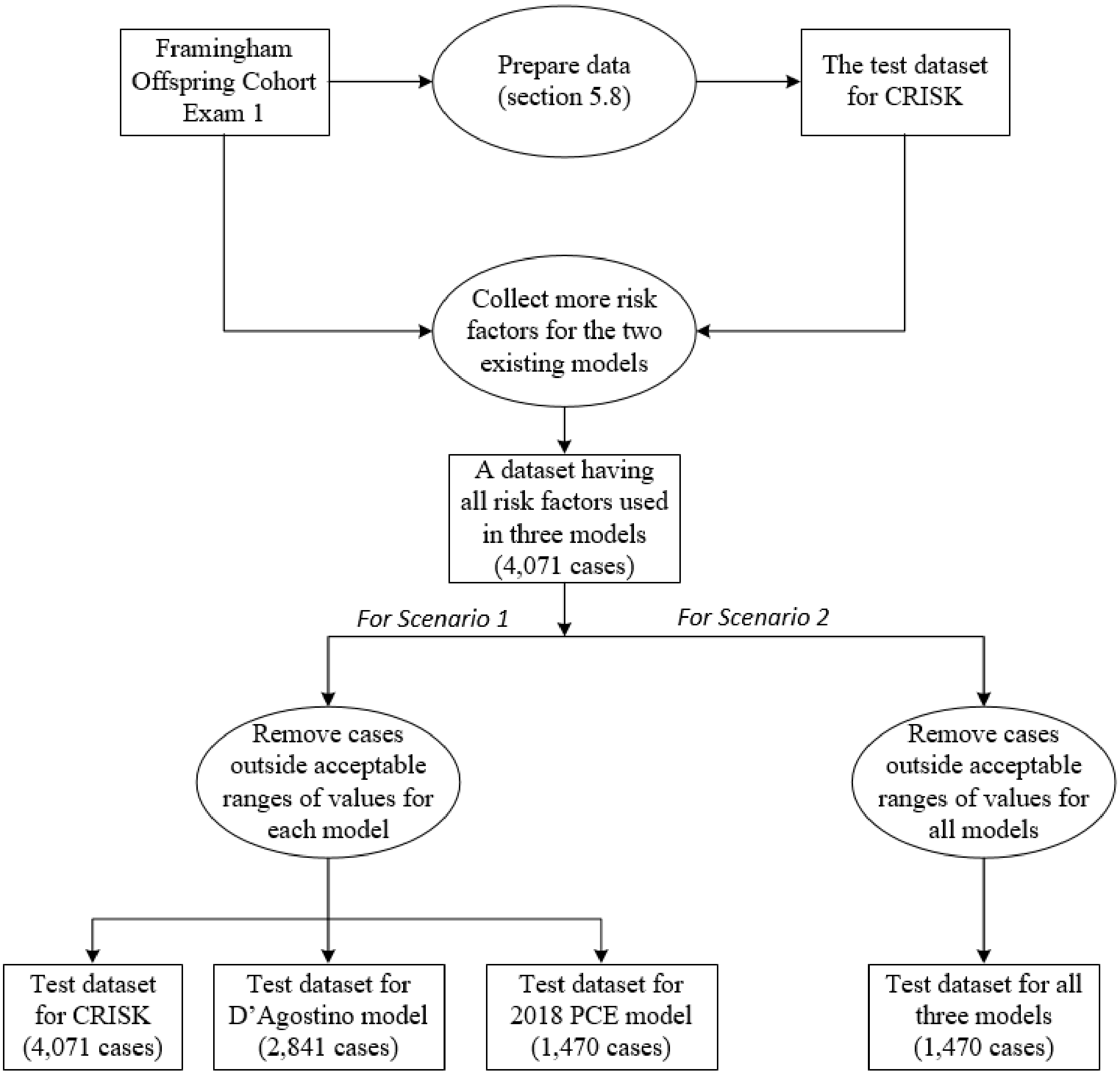

Framingham Offspring Cohort Exam 1
Prepare data (section 5.8)
The test dataset for CRISK
Collect more risk factors for the two existing models
A dataset having all risk factors used in three models (4,071 cases)
For Scenario 1
For Scenario 2
Remove cases outside acceptable ranges of values for each model
Remove cases outside acceptable ranges of values for all models
Test dataset for CRISK (4,071 cases)
Test dataset for D'Agostino model (2,841 cases)
Test dataset for 2018 PCE model (1,470 cases)
Test dataset for all three models (1,470 cases)

Figure 8-1: Test dataset preparation process for performance comparison to existing models

The second scenario was to have one test dataset for use in all three models. This was done by removing cases outside acceptable ranges of values for all three models. The resulting dataset had 1,470 cases (P = 166, N = 1,304).

Table 8-3 and Table 8-4 show prediction performance testing results from scenario 1 and scenario 2, respectively, for the three models. The performance metrics given are TPR, TNR, Precision, F_1-value, and NPV. To have these performance metrics for the D'Agostino model and the 2018 PCE model, a confusion matrix was generated for each model. For D'Agostino, the threshold to classify "High CVD Risk" and "Low CVD Risk" from prediction results was 20%. This means that if the predicted 10-year CVD risk (probability) is 20% or more, the prediction is classified as "High CVD Risk" (or predicted 10-year CVD is "Yes"). For 2018 PCE, the threshold was 10% for "High CVD Risk".

Table 8-3: Performance comparison of CRISK and existing models—Scenario 1

	CRISK	**D'Agostino**	**2018 PCE**
TPR	0.8733	0.3318	0.4096
TNR	0.8270	0.9422	0.8673
Precision	0.2247	0.3153	0.2822
F_1-value	0.3574	0.3233	0.3342
NPV	0.9913	0.9462	0.9203

Table 8-4: Performance comparison of CRISK and existing models—Scenario 2

	CRISK	**D'Agostino**	**2018 PCE**
TPR	0.9217	0.3675	0.4096
TNR	0.5775	0.8896	0.8673
Precision	0.2173	0.2976	0.2822
F_1-value	0.3517	0.3288	0.3342
NPV	0.9830	0.9170	0.9203

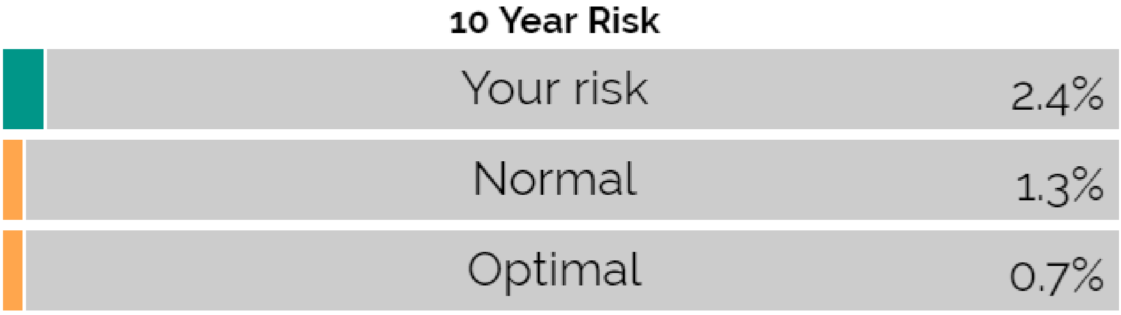
In performance test scenario 1, CRISK can be concluded to perform better than the two existing models. Its TPR was a lot higher than the other two models (0.8733 c.f. 0.3318 and 0.4096). It also performed better in terms of F_1-value and NPV. For TNR and Precision, it did a little worse than the other two models. Although the test dataset for CRISK was a lot less balanced than the test datasets for the other two models (P/N = 221/3850 c.f. 211/2603 and 166/1304), CRISK still achieves a better F_1-value than the two existing D'Agostino and 2018 PCE models (0.3574 c.f. 0.3233 and 0.3342).

In performance test scenario 2, CRISK can also be concluded to perform better than the two existing models. Its TPR was also a lot higher than the other two models (0.9217 c.f. 0.3675 and 0.4096). It also performed better in terms of F_1-value and NPV. For Precision, it did a little worse than the other two models (0.2173 c.f. 0.2976 and 0.2822). Interestingly, for TNR, CRISK performed a lot worse than the other two models (TNR =

0.5775 c.f. 0.8896 and 0.8673). As the three models were run against the same dataset, F_1-value is the decisive metric, followed by Recall (TPR), for decision making (section 3.9.1). Therefore, CRISK can be determined to perform the best among the three models (F_1-value = 0.3517 c.f. 0.3288 and 0.3342, TPR = 0.9217 c.f. 0.3675 and 0.4096).

That CRISK performed badly in TNR in scenario 2 is worth exploring. Its test dataset in scenario 2 was a subset of the test dataset employed in scenario 1. In scenario 1, the test dataset's age range was from 13 to 62 years. In scenario 2, the test dataset's age range was from 40 to 62 years, as cases less than 40 years old were removed to be within acceptable ranges of values for all three models. Looking at the age distribution of the case base in Figure T-17, it can be seen there are a lot more positives than negatives. Therefore, the whole case base was balanced but for the age range from 40 to 62, it was imbalanced and skewed towards positive cases. This could be one of the main reasons for CRISK performing well as determined by TPR but badly in terms of TNR in this age range. This finding is of interest and could lead to future work of dividing the whole dataset into different age ranges and applying SMOTE for each individual age range instead of for the whole dataset.

Besides prediction performance, another focus for comparing these models is how prediction outcomes are presented. Figure 8-2 depicts how CVD prediction outcomes are presented from the D'Agostino model. The other two existing models also provide similar prediction outcome presentations. They all generate prediction outcomes as probabilities for developing CVD within 10 years. For example, in Figure 8-2, the probability of having 10-year CVD is 2.4%. This reaffirms that prediction using regression-based method models is a prediction for populations. In this case, the person does not know whether they belong to the 2.4% of people who would develop CVD or belongs to the 97.6% of people who would not develop CVD within 10 years. Interpretation of the CVD prediction result as conveyed to the person would be that they should try to lower their risk to the normal level and even better to the optimal level.



10 Year Risk

Your risk	2.4%
Normal	1.3%
Optimal	0.7%

Figure 8-2: An example of CVD prediction outcomes presentation of a regression model

The CRISK model generates and presents CVD prediction outcomes differently from the three existing prediction models. This can be seen in Figure 5-32 and Figure 5-33 in section 5.5. The results include predicted Risk Class, predicted CVD Interval, predicted High Risk Membership, predicted Low Risk Membership, a graph depicting "High CVD Risk" and "Low CVD Risk" fuzzy sets, and the seven closest cases to the input case. Information contained in the seven closest cases may help doctors make a decision as to whether to accept or reject the system's prediction result. This information, from the closest seven cases, may also help doctors find useful additional insights into the specific case under consideration.

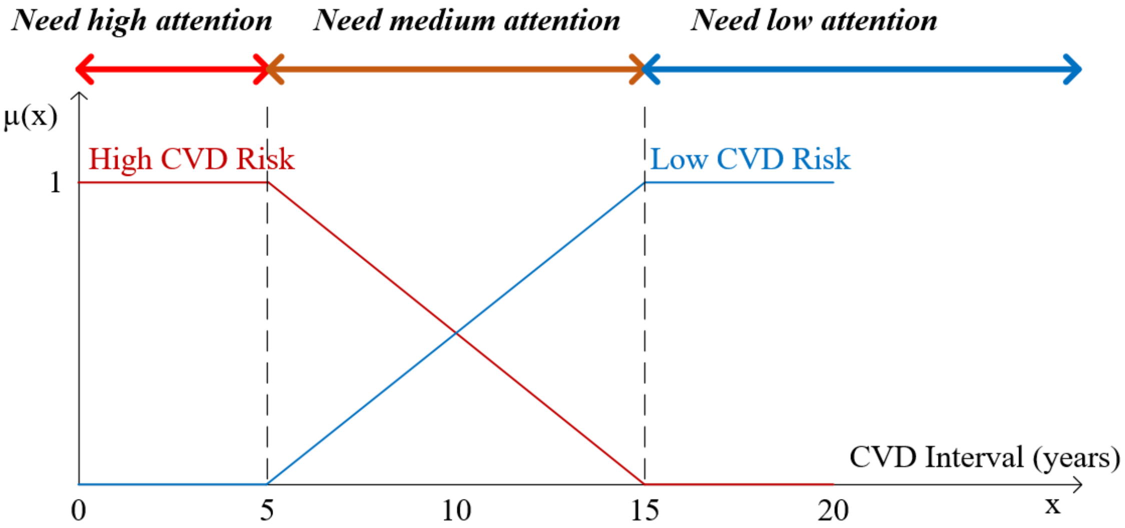


Figure 8-3: Possible interpretation of CVD prediction outcomes from the CRISK model

For interpretation of the CVD prediction result by the CRISK model, a possible way to respond to the prediction result is proposed as illustrated in Figure 8-3. When the result falls into the left (red) area, i.e. $\mu_H = 1$, the person needs high attention. When the result falls into the middle (orange) area, i.e. $0 < \mu_H < 1$ and $0 < \mu_L < 1$, the person needs medium attention. When the result falls into the right (blue) area, i.e. $\mu_L = 1$, the person needs low attention. The person should aim to shift their CVD prediction result towards the right (the blue area). With this new way to interpret the prediction result, a wrongly classified case (FP or FN) may still be paid attention to and therefore not be missed out as may still have a positive membership value belongs to the correct fuzzy set (see real examples in section 7.5).

8.4 CLINICAL APPLICABILITY OF THE CRISK MODEL

The CRISK model could be possibly applied in day-to-day operations in healthcare clinics. There are several reasons for this belief. First, for prediction performance, it achieves TPR=0.8733 and TNR=0.8270 (section 6.3), and performs better than two existing high-profile models (section 8.3). Second, it is designed to predict for an individual, not for a population. Third, besides the prediction result, the system also displays the closest cases to the input case. This information would be useful for e.g. manual checking and manual decision-making by doctors. In addition, it provides a new way to interpret the CVD prediction result using fuzzy set memberships. This new way would be worth further investigation as it may possibly provide more useful and accurate information than the traditional way of using risk probabilities (see justification in section 8.3). Moreover, the CRISK system was designed to be continuously updated. Updates include enriching the case base, updating the list of risk factors, and updating the number of nearest neighbours. Updates of the list of risk factors and the number of nearest neighbours can be achieved using the developed experimentation framework (CRISK Experimenter Module).

Chapter 9
CONCLUSION

This chapter gives a concise and engaging conclusion to this body of work. It first gives a summary of the research achievements. These include providing answers to the research questions, giving reflection on the research, and showing the contribution of the study. Finally, limitations and future directions for this research are provided.

9.1 RESEARCH ACHIEVEMENTS

9.1.1 Answers to Research Questions

As there was no existing fuzzy ontology CBR model for the CVD prediction domain, this research set out to build a fuzzy ontology CBR model for prediction of 10-year CVD (reported in Chapter 4, Chapter 5 and Chapter 6), plus performing external validation (reported in Chapter 7) and having discussions (reported in Chapter 8) to answer the six research questions defined in section 1.2. As a result, this research has given answers to these six research questions, restated below:

RQ1. Can a CVD prediction model be developed using a combination of fuzzy ontology and CBR?

RQ2. What risk factors are important in the prediction of CVD using this method?

RQ3. How does the developed model perform in terms of prediction performance?

RQ4. How does the developed model perform in terms of external validation?

RQ5. How does the developed model overcome the limitations of current widely used regression models?

RQ6. How does the developed model compare with current widely used regression models in terms of prediction performance?

The answer to RQ1 is “Yes”. A way to develop CRISK using a combination of fuzzy ontology and CBR is summarised as follows. Existing cases are stored in a case base which is a fuzzy ontology. The model has four main algorithms Retrieve, Reuse, Revise, and Retain associated with the four CBR activities. The output (prediction result) includes predicted CVD Class, predicted CVD Interval, predicted High CVD Risk membership, and predicted Low CVD Risk membership. Details of the CRISK model and the implemented CRISK system can be found in Chapter 4 and Chapter 5, respectively.

The answer to RQ2 is age, total cholesterol, LDL cholesterol, VLDL cholesterol, SBP, triglycerides, DBP, glucose, cigarettes, HDL cholesterol, hematocrit, BMI, and LDH. However, this answer was based on the case base built from the FHS Offspring Cohort dataset. This list of values might change when data from other datasets are included in the case base. Details of the experiments to find the list of risk factors for the CRISK model can be found in Chapter 6.

For answering RQ3, the CRISK model achieved prediction performance results of TPR=0.8733 (CI=0.0102), TNR=0.8270 (CI=0.0116), Precision=0.2247 (CI=0.0128), F_1-value=0.3574 (CI=0.0147), and NPV=0.9913 (CI=0.0029) where CI is the 95% confidence interval. Details of the experiments and the prediction performance results are in Chapter 6.

For answering RQ4, the CRISK model achieved TPR=0.8167 (CI=0.0434), TNR=0.5041 (CI=0.0560), Precision=0.2866 (CI=0.0507), F_1-value=0.4242 (CI=0.0554), and NPV=0.9185 (CI=0.0307) where CI is the 95% confidence interval for external validation using the FHS Original Cohort Exam 11 dataset. Two risk factors, triglycerides and LDH, were missing in this external dataset. More details of external validation are in Chapter 7.

To answer RQ5, an analysis was done and it concluded that CRISK was able to solve or partially solve five out of eight limitations identified for regression models. A combination of fuzzy ontology and CBR helped build a model that provided several advantages over the current mainstream regression models. Using CBR, the developed model supported personalised prediction by focusing on closest cases to the input case. Moreover, retrieving and showing the closest cases alongside the CVD prediction outcomes helped give a good explanatory capability to the model. Using type-1 fuzzy ontology meant that the CRISK model could handle inaccurate and vague input data and prediction results. In addition, the CRISK model developed was designed to allow for missing data in both input cases and the case base. More details of the analysis can be found in section 8.1.

For answering RQ6, CRISK was compared to two high-profile CVD prediction models, D'Agostino et al. [31] and 2018 PCE [77]. CRISK outperformed these two models in terms of prediction performance when testing on the FHS Offspring Cohort Exam 1 dataset. More details on comparing CRISK to existing models are in section 8.3.

9.1.2 Research Contributions

There are several contributions this research has added to the existing base of knowledge. They are summarised below:

- The thesis contributed an in-depth literature review of CVD prediction models, including conventional Framingham models, augmented Framingham models, and alternatives to Framingham models. In addition, the literature review identified eight problems with current mainstream regression models. It also reviewed current fuzzy logic, fuzzy ontology and CBR approaches in CVD prediction. Moreover, it provided the reasons why a combination of fuzzy ontology and CBR would be able to solve the problems of regression models and would be worth investigating for CVD prediction.
- The main contribution of this research was the design, implementation and evaluation of the CRISK prediction model and its associated CRISK system. The CRISK model achieved prediction performance results of TPR=0.8733

(CI=0.0102), TNR=0.8270 (CI=0.0116), Precision=0.2247 (CI=0.0128), F_1-value=0.3574 (CI=0.0147), and NPV=0.9913 (CI=0.0029) where CI is the 95% confidence interval. These results are reasonably good for a CVD prediction model. In addition, the CRISK model was shown to solve or partially solve five out of the eight problems of current mainstream regression models. Moreover, CRISK performed better when compared to two high-profile existing models, D'Agostino et al. [31] and 2018 PCE [77], by testing all the models against the same dataset—the FHS Offspring Cohort Exam 1. The CRISK system contains modules for creating ontologies, running experiments with different datasets, number of nearest neighbours, and number of risk factors for different scenarios, and providing CVD prediction for an individual case.

- This research showed that fuzzy ontology CBR approaches are useful in CVD prediction. Fuzzy ontology helps deal with vagueness and uncertainty of data. CBR is suitable for personalised prediction. These advantages should encourage future researchers to invest fuzzy ontology CBR approaches in CVD prediction specifically and in chronic disease prediction generally.
- This research contributed a new way to represent and interpret CVD prediction outcomes. The prediction outcomes are represented as fuzzy membership values of "High CVD Risk" and "Low CVD Risk" fuzzy sets. Depending on the fuzzy membership values, different attention is given to the input case. This new way of representing and interpreting CVD prediction outcomes is different from the widely used regression models. A typical regression model displays prediction outcomes as probabilities e.g. 5% probability of developing CVD within 10 years.
- Finally, this research proposed the idea of continuous experimentation and updates for a CVD prediction model and provided a system that enables this process. So far, it has been that a model developed from a cohort turns out to perform poorly on different cohorts. Therefore, it seems reasonable to keep experimenting on new datasets to update the model in order to continuously improve the prediction performance of the model.

9.1.3 Reflection on the Research

One factor that helped arrive at a successful outcome to this programme of research was a properly defined research methodology. The Design Science methodology was accompanied by a research framework, research guidelines, and research strategies and plans. In addition, dataset collection, dataset selection, experimentation design, data preparation, and an evaluation protocol were also defined and documented in detail. This made the research journey go smoothly and resulted in the CRISK model, which is a new and innovative artifact that harnesses the strengths of fuzzy ontology and CBR for CVD prediction. This artifact's success led to prediction performance results that helped successfully answer the six research questions.

This body of work can be easily reproduced. Besides a detailed research methodology, the model design, implementation, experimentation, and external validation were also documented. In particular, the developed Retrieve, Revise, and Reuse algorithms were given in this thesis. Moreover, programming language, plugins, and development tools were also reported in this thesis. Therefore, following this thesis step by step, other researchers can reproduce this research and be able to arrive at the same results.

9.2 LIMITATIONS AND FUTURE DIRECTIONS

In this research, there were several limitations opening up areas for future work. These limitations were attributed to a number of causes, including the time constraint of this three year PhD programme, lack of open and freely available datasets for building and validating the CRISK prediction model developed in this research, and that fuzzy ontology CBR approaches are new in CVD prediction and thus there is a lack of existing resources e.g. algorithms and tools. The following subsections give details of the limitations and future directions identified for this study.

9.2.1 Need for expansion of the current Case Base

The case base of the CRISK system needs to be continuously expanded. There are a couple of reasons for this. Primarily, as it is a CBR system, the richer the case base, the more chance for the system to receive closer cases to the input case. The case base was built from the FHS Offspring Cohort, which was a majority-Caucasian cohort. Racial/ethnic status has been considered as a strong predictor for chronic disease, including CVD [186]. Excluding race from a clinical prediction model (CPM) may lead to inaccurate prognostication and harmful decision making in minority groups [187, 188]. Therefore, the case base needs to be expanded to include cases from other ethnicities. In addition, continuous expansion of the case base will support more experimentation to update the CRISK model regularly to continuously improve CVD risk prediction.

However, to accomplish expansion of the CRISK case base, more datasets need to be made available to researchers in the future. In this study, it was difficult to find open datasets to build and validate the CRISK model. There are reasons, such as confidentiality of healthcare data, for not sharing datasets. However, sharing data freely and openly might help accelerate the progress towards achieving a precise and reliable CVD prediction model.

9.2.2 Experimenting different approaches to balance the case base

As pointed out in section 8.3, the current approach of applying SMOTE to the whole imbalanced dataset for building the case base may not be the best approach. Using SMOTE may have resulted in the poor TNR observed for senior people.

Two alternative approaches may be worth exploring. Firstly, dividing the dataset into different age ranges, for example [20, 29], [30, 39], [40, 49], [50, 59], and so on. SMOTE is then applied to each individual age range separately (refer to section 8.3 for justification). The second option is to gather more real positive cases to the dataset and remove negative cases from the dataset in order to balance the number of positive and negative cases.

9.2.3 Fuzzification of Predictor Variables

In this research, fuzzy logic was applied in representing CVD prediction outcomes, but not predictor variables. There was no guarantee that fuzzification of predictors would yield better prediction for the CRISK model. However, it may be worth trying. A very strong candidate for fuzzification among the 13 predictors used in the model is “cigarettes”. A person may not know exactly how many cigarettes they smoke a day. The number of cigarettes smoked a day could be fuzzified as, for example, “light smoking”, “medium smoking”, and “heavy smoking”. Another candidate for fuzzification is “race”. “White race” set, “black race” set, and “Asian race” set, for example, could be created. A person of mixed race, for example “white” and “black”, could be identified by the degrees of membership values, such as $\mu_{white} = 0.5$, $\mu_{black} = 0.5$, $\mu_{Asian} = 0$.

9.2.4 Usage of type-2 fuzzy ontology

In this study, the capability of handling uncertainty of data of type-2 fuzzy sets (see section 2.3.2) was not explored. The reason was twofold. First, as type-1 fuzzy sets are simpler than type-2 fuzzy sets, type-1 fuzzy sets were chosen to start with in this research to see how things go. Second, there was no fuzzy ontology tool that supported type-2 fuzzy sets. The three-year time constraint of the PhD programme did not allow development of such an ontology tool. However, it would be worth exploiting the capability of handling data uncertainty of type-2 fuzzy sets to improve CVD prediction models in future.

9.2.5 More Experimentation and Continuous Updates of the CRISK model

As CBR is driven by data, continuous experimentation and updates of the CRISK model are recommended. When new cases are added into the case base, the experimentation described in Chapter 6 should be carried out. This will help update the number of nearest neighbours *k* and the list of predictors for the CRISK model. Moreover, when having enough cases from different races, it might be worth experimenting with different models for different races. Results from different experiments should be compared to decide on optimal settings for the CRISK model's parameters.

Not only the CRISK model, but any model (e.g. a regression model) should be continuously updated. In case of a regression model, whose equation was built from a certain dataset, having new datasets from different cohorts, if the model was rebuilt from a combination of the original dataset and the new datasets, would most likely result in a different equation to the original model's equation.

There could be an argument asking: "When will this continuous updating process end?" or "Is the model never completed?" The answer is that the continuous updating process might never end. This process is aimed to periodically improve the CVD prediction performance for a model and/or keep the model up to date. These updates reflect changes in population health and related environmental factors from which contemporary cases are derived. Although a perfect prediction model might never be achieved, continuous experimentation and updates to a developed model on new emerging datasets should result in continuous improvement of the prediction performance of the model.

9.2.6 Introduction of Weights for the Retrieve algorithm

It may be worth trying to add weights to the Distance algorithm that is part of the Retrieve activity (described in section 4.5). Then, line 16 of the Distance algorithm would become $d = d + weight \times weight \times diff \times diff$. As different risk factors have different levels of impact on the CVD outcomes (as ranked by Weka's InfoGainAttributeEval attribute evaluator), appropriate weight values could possibly improve the CVD prediction performance for the CRISK model.

9.2.7 Missing Data Handling Mechanism

It might be good trying to develop an automatic missing data handling mechanism. The CRISK system might then be able to set a boundary for missing data. For example, how many missing risk factor values and which missing risk factors are acceptable for the input case? In addition, the system might also be able to replace missing values from input cases with appropriate values computed using an appropriate imputation method. Currently, the system simply ignores missing risk factors and retrieves closest cases based on risk factors that have values in the input case. As a result, CRISK heavily depends on doctors making judgement calls on CVD prediction outcomes for those cases that are missing data.

9.2.8 CRISK system giving indication for out-of-range input

The developed CRISK system should somehow give indication of when the input case has risk factor values that are outside of value ranges in the case base. This indication would help doctors make judgement calls if the input case is not "so much" out-of-range and accept the prediction outcomes from the system. Out-of-range values proved to affect the prediction performance as shown in section 7.4—prediction performance on External Dataset 2 (no out-of-range values) was better than on External Dataset 1 (having out-of-range values). Implementing this indication feature for the CRISK system would only take a few days.

9.2.9 Clinical Trials

Clinical trials of the developed CRISK model should be one of the next steps following the completion of this PhD research. The CRISK Predictor module of the CRISK system was designed to be used in day-to-day operation in healthcare clinics. However, the usefulness and usability of the developed system cannot be fully realised until clinical trials.

REFERENCES

[1] W. H. Organization, "Cardiovascular Diseases (CVDs)-Fact Sheet 317, May 2017," ed, 2017.

[2] E. Wilkins *et al.*, "European cardiovascular disease statistics 2017," 2017.

[3] E. J. Benjamin *et al.*, "Heart disease and stroke statistics—2018 update: a report from the American Heart Association," *Circulation,* vol. 137, no. 12, pp. e67-e492, 2018.

[4] W. H. Organization, *Prevention of cardiovascular disease: guidelines for assessment and management of cardiovascular risk*. World Health Organization, 2007.

[5] L. J. Shaw *et al.*, "10-Year Resource Utilization and Costs for Cardiovascular Care," *Journal of the American College of Cardiology,* vol. 71, no. 10, pp. 1078-1089, 2018.

[6] J. A. Damen *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," (in eng), *Bmj,* vol. 353, p. i2416, May 16 2016, doi: 10.1136/bmj.i2416.

[7] S. J. Al'Aref *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European heart journal,* vol. 40, no. 24, pp. 1975-1986, 2019.

[8] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PloS one,* vol. 12, no. 4, p. e0174944, 2017.

[9] S. El-Sappagh, M. Elmogy, and A. Riad, "A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis," *Artificial intelligence in medicine,* vol. 65, no. 3, pp. 179-208, 2015.

[10] S. El-Sappagh, M. Elmogy, F. Ali, and K.-S. Kwak, "A case-base fuzzification process: diabetes diagnosis case study," *Soft Computing,* pp. 1-20, 2018.

[11] S. Tahmasebian, M. Langarizadeh, M. Ghazisaeidi, and M. Mahdavi-Mazdeh, "Designing and implementation of fuzzy case-based reasoning system on android platform using electronic discharge summary of patients with chronic kidney diseases," *Acta Informatica Medica,* vol. 24, no. 4, p. 266, 2016.

[12] D. Parry and J. MacRae, "Fuzzy ontologies for cardiovascular risk prediction-A research approach," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013: IEEE, pp. 1-4.

[13] I. Bichindaritz and C. Marling, "Case-based reasoning in the health sciences: What's next?," *Artificial intelligence in medicine,* vol. 36, no. 2, pp. 127-135, 2006.

[14] A. Holt, I. Bichindaritz, R. Schmidt, and P. Perner, "Medical applications in case-based reasoning," *The Knowledge Engineering Review,* vol. 20, no. 03, pp. 289-292, 2005.

[15] S. Chattopadhyay, S. Banerjee, F. A. Rabhi, and U. R. Acharya, "A Case-Based Reasoning system for complex medical diagnosis," *Expert Systems,* vol. 30, no. 1, pp. 12-20, 2013.

[16] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications,* vol. 7, no. 1, pp. 39-59, 1994.

[17] S. Mendis, P. Puska, and B. Norrving, *Global atlas on cardiovascular disease prevention and control*. World Health Organization, 2011.

[18] W. H. Organization, "Cardiovascular diseases (CVDs)," *Fact sheet,* vol. 317, 2015.

[19] K. N. Frayn, S. Stanner, and F. British Nutrition, *Cardiovascular disease : diet, nutrition and emerging risk factors : the report of a British Nutrition Foundation task force*. Oxford, UK; Ames, Iowa, USA: Published by Blackwell Pub. for the British Nutrition Foundation (in English), 2005.

[20] J. Y. Chong, *Cerebrovascular disease* (no. Book, Whole). New York: Oxford University Press, 2013.

[21] C. G. Kevil, S. C. Bir, C. B. Pattillo, and N. I. Akkus, *Peripheral arterial disease: pathophysiology and therapeutics* (no. Book, Whole). San Rafael, California: Morgan & Claypool, 2013.

[22] S. Z. Goldhaber and R. B. Morrison, "Pulmonary embolism and deep vein thrombosis," *Circulation,* vol. 106, no. 12, pp. 1436-1438, 2002.

[23] E. Marijon, M. Mirabel, D. S. Celermajer, and X. Jouven, "Rheumatic heart disease," *The Lancet,* vol. 379, no. 9819, pp. 953-964, 2012.

[24] J. I. Hoffman and S. Kaplan, "The incidence of congenital heart disease," *Journal of the American college of cardiology,* vol. 39, no. 12, pp. 1890-1900, 2002.

[25] G. A. Roth *et al.*, "Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015," *Journal of the American College of Cardiology,* p. 23715, 2017.

[26] T. Dent, "Predicting the risk of coronary heart disease: I. The use of conventional risk markers," *Atherosclerosis,* vol. 213, no. 2, pp. 345-351, 2010.

[27] W. B. Kannel, D. McGee, and T. Gordon, "A general cardiovascular risk profile: the Framingham Study," *The American journal of cardiology,* vol. 38, no. 1, pp. 46-51, 1976.

[28] K. M. Anderson, P. WOLSON, P. M. Odell, and W. B. Kannel, "An updated coronary risk profile: a statement fo rhealth professionals," *Circulation,* vol. 83, no. 1, pp. 356-362, 1991.

[29] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *American heart journal,* vol. 121, no. 1, pp. 293-298, 1991.

[30] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation,* vol. 97, no. 18, pp. 1837-1847, 1998.

[31] R. B. D'Agostino *et al.*, "General cardiovascular risk profile for use in primary care the Framingham Heart Study," *Circulation,* vol. 117, no. 6, pp. 743-753, 2008.

[32] M. J. Pencina, R. B. D'Agostino, M. G. Larson, J. M. Massaro, and R. S. Vasan, "Predicting the 30-year risk of cardiovascular disease The Framingham Heart Study," *Circulation,* vol. 119, no. 24, pp. 3078-3084, 2009.

[33] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective," *The Lancet,* vol. 383, no. 9921, pp. 999-1008, 2014.

[34] S. S. Franklin *et al.*, "Does the relation of blood pressure to coronary heart disease risk change with aging? The Framingham Heart Study," *Circulation,* vol. 103, no. 9, pp. 1245-1249, 2001.

[35] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika,* vol. 54, no. 1-2, pp. 167-179, 1967.

[36] K. M. Anderson, "A nonproportional hazards Weibull accelerated failure time regression model," *Biometrics,* pp. 281-288, 1991.

[37] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 187-220, 1972.

[38] E. S. Ford, W. H. Giles, and A. H. Mokdad, "The distribution of 10-year risk for coronary heart disease among US adults: findings from the National Health and Nutrition Examination Survey III," *Journal of the American College of Cardiology,* vol. 43, no. 10, pp. 1791-1796, 2004.

[39] M. T. Cooney, A. Dudina, R. D'Agostino, and I. M. Graham, "Cardiovascular Risk-Estimation Systems in Primary Prevention Do They Differ? Do They Make a Difference? Can We See the Future?," *Circulation,* vol. 122, no. 3, pp. 300-310, 2010.

[40] D. Laurier, N. P. Chau, B. Cazelles, P. Segond, and P.-M. Group12, "Estimation of CHD risk in a French working population using a modified Framingham model," *Journal of clinical epidemiology,* vol. 47, no. 12, pp. 1353-1364, 1994.

[41] H.-W. Hense, H. Schulte, H. Löwel, G. Assmann, and U. Keil, "Framingham risk function overestimates risk of coronary heart disease in men and women from Germany—results from the MONICA Augsburg and the PROCAM cohorts," *European Heart Journal,* vol. 24, no. 10, pp. 937-945, 2003.

[42] A. Menotti, P. Puddu, and M. Lanti, "Comparison of the Framingham risk function-based coronary chart with risk function from an Italian population study," *European Heart Journal,* vol. 21, no. 5, pp. 365-370, 2000.

[43] P. Brindle *et al.*, "Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study," *Bmj,* vol. 327, no. 7426, p. 1267, 2003.

[44] J. Reissigova and J. Zvarova, "The Framingham risk function underestimated absolute coronary heart disease risk in Czech men," *Methods Inf Med,* vol. 46, no. 1, pp. 43-49, 2007.

[45] T. Dent, "Predicting the risk of coronary heart disease. II: The role of novel molecular biomarkers and genetics in estimating risk, and the future of risk prediction," *Atherosclerosis,* vol. 213, no. 2, pp. 352-362, 2010.

[46] J. Danesh *et al.*, "C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease," *New England Journal of Medicine,* vol. 350, no. 14, pp. 1387-1397, 2004.

[47] D. M. Lloyd-Jones, K. Liu, L. Tian, and P. Greenland, "Narrative review: assessment of C-reactive protein in risk prediction for cardiovascular disease," *Annals of internal medicine,* vol. 145, no. 1, pp. 35-42, 2006.

[48] J. Danesh *et al.*, "Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis," *JAMA: the journal of the American Medical Association,* vol. 294, no. 14, pp. 1799-1809, 2005.

[49] D. S. Wald, M. Law, and J. K. Morris, "Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis," *Bmj,* vol. 325, no. 7374, p. 1202, 2002.

[50] J. De Sutter *et al.*, "Plasma N-terminal pro-brain natriuretic peptide concentration predicts coronary events in men at work: a report from the BELSTRESS study," *European heart journal,* vol. 26, no. 24, pp. 2644-2649, 2005.

[51] C. Kistorp, I. Raymond, F. Pedersen, F. Gustafsson, J. Faber, and P. Hildebrandt, "N-terminal pro-brain natriuretic peptide, C-reactive protein, and urinary albumin levels as predictors of mortality and cardiovascular events in older adults," *Jama,* vol. 293, no. 13, pp. 1609-1616, 2005.

[52] M. Rizzo and K. Berneis, "Low-density lipoprotein size and cardiovascular risk assessment," *Qjm,* vol. 99, no. 1, pp. 1-14, 2006.

[53] A. Thompson and J. Danesh, "Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: a literature-based meta-analysis of prospective studies," *Journal of internal medicine,* vol. 259, no. 5, pp. 481-492, 2006.

[54] A. M. Bennet *et al.*, "Association of apolipoprotein E genotypes with lipid levels and coronary risk," *Jama,* vol. 298, no. 11, pp. 1300-1311, 2007.

[55] C. A. Garza, V. M. Montori, J. P. McConnell, V. K. Somers, I. J. Kullo, and F. Lopez-Jimenez, "Association between lipoprotein-associated phospholipase A 2 and cardiovascular disease: a systematic review," in *Mayo Clinic Proceedings*, 2007, vol. 82, no. 2: Elsevier, pp. 159-165.

[56] E. R. F. Collaboration, "Lipoprotein (a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality," *JAMA: the journal of the American Medical Association,* vol. 302, no. 4, p. 412, 2009.

[57] M. G. Shlipak *et al.*, "Cystatin C and the risk of death and cardiovascular events among elderly persons," *New England Journal of Medicine,* vol. 352, no. 20, pp. 2049-2060, 2005.

[58] A. Strasak *et al.*, "Serum uric acid and risk of cardiovascular mortality: a prospective long-term study of 83 683 Austrian men," *Clinical chemistry,* vol. 54, no. 2, pp. 273-284, 2008.

[59] A. M. Strasak *et al.*, "Serum uric acid is an independent predictor for all major forms of cardiovascular death in 28,613 elderly women: a prospective 21-year follow-up study," *International journal of cardiology,* vol. 125, no. 2, pp. 232-239, 2008.

[60] M. J. Bos, P. J. Koudstaal, A. Hofman, J. C. Witteman, and M. M. Breteler, "Uric acid is a risk factor for myocardial infarction and stroke the Rotterdam study," *Stroke,* vol. 37, no. 6, pp. 1503-1507, 2006.

[61] R. K. Schindhelm *et al.*, "Alanine aminotransferase predicts coronary heart disease events: a 10-year follow-up of the Hoorn Study," *Atherosclerosis,* vol. 191, no. 2, pp. 391-396, 2007.

[62] A. Fraser, R. Harris, N. Sattar, S. Ebrahim, G. D. Smith, and D. Lawlor, "Gamma-glutamyltransferase is associated with incident vascular events independently of alcohol intake analysis of the British women’s heart and health study and meta-analysis," *Arteriosclerosis, thrombosis, and vascular biology,* vol. 27, no. 12, pp. 2729-2735, 2007.

[63] S. Wannamethee, L. Lennon, and A. Shaper, "The value of gamma-glutamyltransferase in cardiovascular risk prediction in men without diagnosed cardiovascular disease or diabetes," *Atherosclerosis,* vol. 201, no. 1, pp. 168-175, 2008.

[64] R. Jackson, "Updated New Zealand cardiovascular disease risk-benefit prediction guide," *Bmj,* vol. 320, no. 7236, pp. 709-710, 2000.

[65] K. Pyörälä, G. De Backer, I. Graham, P. Poole-Wilson, and D. Wood, "Prevention of coronary heart disease in clinical practice: recommendations of the Task Force of the European Society of Cardiology, European Atherosclerosis Society and European Society of Hypertension," *Atherosclerosis,* vol. 110, no. 2, pp. 121-161, 1994.

[66] B. H. S. British Cardiac Society, Diabetes UK, HEART UK, Primary Care Cardiovascular Society, The Stroke Association, "JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice," *Heart,* vol. 91, pp. v1-v52, 2005.

[67] G. Assmann, P. Cullen, and H. Schulte, "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study," *Circulation,* vol. 105, no. 3, pp. 310-315, 2002.

[68] R. Conroy *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project," *European heart journal,* vol. 24, no. 11, pp. 987-1003, 2003.

[69] M. Woodward, P. Brindle, and H. Tunstall-Pedoe, "Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)," *Heart,* vol. 93, no. 2, pp. 172-176, 2007.

[70] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score," *Jama,* vol. 297, no. 6, pp. 611-619, 2007.

[71] P. M. Ridker, N. P. Paynter, N. Rifai, J. M. Gaziano, and N. R. Cook, "C-Reactive Protein and Parental History Improve Global Cardiovascular Risk Prediction The Reynolds Risk Score for Men," *Circulation,* vol. 118, no. 22, pp. 2243-2251, 2008.

[72] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study," *Bmj,* vol. 335, no. 7611, p. 136, 2007.

[73] J. Hippisley-Cox *et al.*, "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2," *Bmj,* vol. 336, no. 7659, pp. 1475-1482, 2008.

[74] D. C. Goff *et al.*, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *Journal of the American College of Cardiology,* vol. 63, no. 25 Part B, pp. 2935-2959, 2014.

[75] K. Hajifathalian *et al.*, "A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys," *The Lancet Diabetes & Endocrinology,* vol. 3, no. 5, pp. 339-355, 2015.

[76] R. Pylypchuk *et al.*, "Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study," *The Lancet,* vol. 391, no. 10133, pp. 1897-1907, 2018.

[77] S. Yadlowsky, R. A. Hayward, J. B. Sussman, R. L. McClelland, Y.-I. Min, and S. Basu, "Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk," *Annals of internal medicine,* 2018.

[78] W. Weibull, "A statistical distribution of wide applicability," *Journal of applied mechanics,* vol. 103, 1951.

[79] T. A. Investigators, "The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives," *American journal of epidemiology,* vol. 129, no. 4, pp. 687-702, 1989.

[80] L. P. Fried *et al.*, "The cardiovascular health study: design and rationale," *Annals of epidemiology,* vol. 1, no. 3, pp. 263-276, 1991.

[81] G. D. Friedman *et al.*, "CARDIA: study design, recruitment, and some characteristics of the examined subjects," *Journal of clinical epidemiology,* vol. 41, no. 11, pp. 1105-1116, 1988.

[82] T. R. Dawber, W. B. Kannel, and L. P. Lyell, "An approach to longitudinal studies in a community: the Framingham Study," *Annals of the New York Academy of sciences,* vol. 107, no. 2, pp. 539-556, 1963.

[83] W. B. Kannel, M. Feinleib, P. M. McNamara, R. J. Garrison, and W. P. Castelli, "An investigation of coronary heart disease in families: the Framingham Offspring Study," *American journal of epidemiology,* vol. 110, no. 3, pp. 281-290, 1979.

[84] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *European Heart Journal,* p. ehw302, 2016.

[85] J. Wolfson *et al.*, "A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data," *Statistics in medicine,* vol. 34, no. 21, pp. 2941-2957, 2015.

[86] D. B. Panagiotakos and V. Stavrinos, "Methodological issues in cardiovascular epidemiology: the risk of determining absolute risk through statistical models," *Vascular health and risk management,* vol. 2, no. 3, p. 309, 2006.

[87] M. Matheny *et al.*, "Systematic review of cardiovascular disease risk assessment tools," 2011.

[88] D. Pal, K. Mandana, S. Pal, D. Sarkar, and C. Chakraborty, "Fuzzy expert system approach for coronary artery disease screening using clinical parameters," *Knowledge-Based Systems,* vol. 36, pp. 162-174, 2012.

[89] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," *Cluster Computing,* vol. 17, no. 3, pp. 881-891, 2014.

[90] J. Kim, J. Lee, and Y. Lee, "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree," *Healthcare informatics research,* vol. 21, no. 3, pp. 167-174, 2015.

[91] M. Goyal, "A Case-Based Reasoning Framework for Prediction of Stroke," in *Information and Communication Technology*: Springer, 2018, pp. 219-227.

[92] H. Malekpoor, "A novel combination of Cased-Based Reasoning and Multi Criteria Decision Making approach to radiotherapy dose planning," University of East Anglia, 2018.

[93] K. L. T. Choy *et al.*, "An intelligent case-based knowledge management system for quality improvement in nursing homes," *VINE Journal of Information and Knowledge Management Systems,* vol. 48, no. 1, pp. 103-121, 2018.

[94] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and M. Folke, "Case-based reasoning systems in the health sciences: a survey of recent trends and developments," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 41, no. 4, pp. 421-434, 2011.

[95] N. Choudhury and S. A. Begum, "A survey on case-based reasoning in medicine," *International Journal of Advanced Computer Science and Applications,* vol. 7, no. 8, pp. 136-144, 2016.

[96] D. H. Sutanto, N. S. Herman, M. Ghani, and K. Abd, "Trend of Case Based Reasoning in Diagnosing Chronic Disease: A Review," *Advanced Science Letters,* vol. 20, no. 10-11, pp. 1740-1744, 2014.

[97] S. Guessoum, M. T. Laskri, and J. Lieber, "RespiDiag: a case-based reasoning system for the diagnosis of chronic obstructive pulmonary disease," *Expert Systems with Applications,* vol. 41, no. 2, pp. 267-273, 2014.

[98] P. Koton, "A medical reasoning program that improves with experience," *Computer methods and programs in biomedicine,* vol. 30, no. 2-3, pp. 177-184, 1989.

[99] E. B. Reategui, J. A. Campbell, and B. F. Leao, "Combining a neural network with case-based reasoning in a diagnostic system," *Artificial Intelligence in Medicine,* vol. 9, no. 1, pp. 5-27, 1997.

[100] V. D. Kalavai, "Heart disease prediction system using ANN, RBF and CBR," *International Journal of Pure and Applied Mathematics,* vol. 117, no. 21, pp. 199-217, 2017.

[101] G.-K. Park, J. L. Benedictos, C.-S. Lee, and M.-H. Wang, "Ontology-based fuzzy-CBR Support System for ship's collision avoidance," in *Machine Learning*

and Cybernetics, 2007 International Conference on, 2007, vol. 4: IEEE, pp. 1845-1850.

[102] F. Ali, E. K. Kim, and Y.-G. Kim, "Type-2 fuzzy ontology-based semantic knowledge for collision avoidance of autonomous underwater vehicles," *Information Sciences,* vol. 295, pp. 441-464, 2015.

[103] V. E. Ekong, U. G. Inyang, and E. A. Onibere, "Intelligent decision support system for depression diagnosis based on neuro-fuzzy-CBR hybrid," *Modern Applied Science,* vol. 6, no. 7, p. 79, 2012.

[104] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence," 1997.

[105] U. G. Inyang and E. E. Joshua, "Fuzzy clustering of students' data repository for at-risks students identification and monitoring," *Computer and Information Science,* vol. 6, no. 4, p. 37, 2013.

[106] V. T. N. Chau, N. H. Phung, and V. T. N. Tran, "A robust and effective algorithmic framework for incomplete educational data clustering," in *Information and Computer Science (NICS), 2015 2nd National Foundation for Science and Technology Development Conference on*, 2015: IEEE, pp. 65-70.

[107] J. A. Recio-García, P. A. González-Calero, and B. Díaz-Agudo, "jcolibri2: A framework for building Case-based reasoning systems," *Science of Computer Programming,* vol. 79, pp. 126-145, 2014.

[108] S. Bleeker *et al.,* "External validation is necessary in prediction research:: A clinical example," *Journal of clinical epidemiology,* vol. 56, no. 9, pp. 826-832, 2003.

[109] K. G. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman, "Prognosis and prognostic research: what, why, and how?," *Bmj,* vol. 338, p. b375, 2009.

[110] L. A. Zadeh, "Fuzzy sets," *Information and control,* vol. 8, no. 3, pp. 338-353, 1965.

[111] H.-J. Zimmermann, *Fuzzy set theory—and its applications*. Springer Science & Business Media, 2011.

[112] J. Zhao and B. K. Bose, "Evaluation of membership functions for fuzzy logic controlled induction motor drive," in *28th Annual Conference of the IEEE Industrial Electronics Society*, Spain, 2002, vol. 1: IEEE, pp. 229-234.

[113] T. J. Ross, *Fuzzy logic with engineering applications*. John Wiley & Sons, 2009.

[114] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Information sciences,* vol. 8, no. 3, pp. 199-249, 1975.

[115] J. M. Mendel and R. B. John, "Type-2 fuzzy sets made simple," *Fuzzy Systems, IEEE Transactions on,* vol. 10, no. 2, pp. 117-127, 2002.

[116] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*. Prentice hall New Jersey, 1995.

[117] J. M. Mendel, "Type-2 fuzzy sets," in *Uncertain Rule-Based Fuzzy Systems*: Springer, 2017, pp. 259-306.

[118] Q. Liang and J. M. Mendel, "Interval type-2 fuzzy logic systems: theory and design," *IEEE Transactions on Fuzzy systems,* vol. 8, no. 5, pp. 535-550, 2000.

[119] J. M. Mendel, "Type-2 fuzzy sets and systems: An overview [corrected reprint]," *IEEE computational intelligence magazine,* vol. 2, no. 2, pp. 20-29, 2007.

[120] F. Bobillo, "Managing vagueness in ontologies," *University of Granada, Granada,* 2008.

[121] T. Guber, "A translational approach to portable ontologies," *Knowledge Acquisition,* vol. 5, no. 2, pp. 199-229, 1993.

[122] D. Man, "Ontologies in Computer Science," *Didactica Mathematica,* vol. 31, no. 1, p. 43, 2013.

[123] Z. Ma, F. Zhang, L. Yan, and J. Cheng, *Fuzzy knowledge management for the semantic web*. Springer, 2014.

[124] R. Fullér, "What is fuzzy logic and fuzzy ontology," in *KnowMobile national workshop, Helsinki*, 2008, vol. 30.

[125] D. Parry, "Fuzzification of a standard ontology to encourage reuse," in *Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on*, 2004: IEEE, pp. 582-587.

[126] B. Smith and C. Welty, "Ontology: Towards a new synthesis," in *Formal Ontology in Information Systems*, 2001, vol. 10, no. 3: ACM Press, USA, pp. iii-x, pp. 3-9.

[127] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34-43, 2001.

[128] F. Zhang, J. Cheng, and Z. Ma, "A survey on fuzzy ontologies for the Semantic Web," *The Knowledge Engineering Review*, vol. 31, no. 3, pp. 278-321, 2016.

[129] G. Antoniou and F. Van Harmelen, *A semantic web primer*. MIT press, 2004.

[130] S. M. S. Huynh, "Towards Semantic Web: current XML resources conversion and RFID employment," Auckland University of Technology, 2014.

[131] W. C. O. W. Group, "{OWL} 2 Web Ontology Language Document Overview," 2009.

[132] M. Mazzieri, "A fuzzy RDF semantics to represent trust metadata," in *1st Workshop on Semantic Web Applications and Perspectives (SWAP2004)*, 2004, pp. 83-89.

[133] M. Mazzieri and A. F. Dragoni, "A fuzzy semantics for the resource description framework," in *Uncertainty Reasoning for the Semantic Web I*: Springer, 2008, pp. 244-261.

[134] Y. Lv, Z. M. Ma, and L. Yan, "Fuzzy RDF: A data model to represent fuzzy metadata," in *Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on*, 2008: IEEE, pp. 1439-1445.

[135] U. Straccia, "A minimal deductive system for general fuzzy RDF," in *International Conference on Web Reasoning and Rule Systems*, 2009: Springer, pp. 166-181.

[136] N. Manolis and Y. Tzitzikas, "Interactive exploration of fuzzy RDF knowledge bases," in *Extended Semantic Web Conference*, 2011: Springer, pp. 1-16.

[137] A. Zimmermann, N. Lopes, A. Polleres, and U. Straccia, "A general framework for representing, reasoning and querying with annotated semantic web data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 11, pp. 72-95, 2012.

[138] M. Gao and C. Liu, "Extending OWL by fuzzy description logic," in *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*, 2005: IEEE, pp. 6 pp.-567.

[139] S. Calegari and D. Ciucci, "Fuzzy ontology, fuzzy description logics and fuzzy-owl," in *International Workshop on Fuzzy Logic and Applications*, 2007: Springer, pp. 118-126.

[140] G. Stoilos, G. Stamou, and J. Z. Pan, "Fuzzy extensions of OWL: Logical properties and reduction to fuzzy description logics," *International Journal of Approximate Reasoning*, vol. 51, no. 6, pp. 656-679, 2010.

[141] F. Bobillo and U. Straccia, "Fuzzy ontology representation using OWL 2," *International Journal of Approximate Reasoning*, vol. 52, no. 7, pp. 1073-1094, 2011.

[142] F. Bobillo and U. Straccia, "fuzzyDL: An expressive fuzzy description logic reasoner," in *Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on*

Computational Intelligence). IEEE International Conference on, 2008: IEEE, pp. 923-930.

[143] F. Bobillo, M. Delgado, and J. Gómez-Romero, "DeLorean: a reasoner for fuzzy OWL 1.1," in *Proceedings of the Fourth International Conference on Uncertainty Reasoning for the Semantic Web-Volume 423*, 2008: CEUR-WS. org, pp. 13-24.

[144] M. Horridge and S. Bechhofer, "The OWL API: a Java API for working with OWL 2 ontologies," in *Proceedings of the 6th International Conference on OWL: Experiences and Directions-Volume 529*, 2009: CEUR-WS. org, pp. 49-58.

[145] H. Ghorbel, A. Bahri, and R. Bouaziz, "Fuzzy protégé for fuzzy ontology models," *Age*, vol. 12, no. 18, p. 30, 2009.

[146] S. Calegari and D. Ciucci, "Fuzzy ontology and fuzzy-OWL in the KAON project," in *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, 2007: IEEE, pp. 1-6.

[147] I. Watson, "Case-based reasoning is a methodology not a technology," *Knowledge-based systems*, vol. 12, no. 5, pp. 303-308, 1999.

[148] A. K. Goel and B. Diaz-Agudo, "What's Hot in Case-Based Reasoning," in *AAAI*, 2017, pp. 5067-5069.

[149] E. L. Rissland, K. D. Ashley, and L. K. Branting, "Case-based reasoning and law," *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 293-298, 2005.

[150] J. L. Kolodner, M. T. Cox, and P. A. González-Calero, "Case-based reasoning-inspired approaches to education," *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 299-303, 2005.

[151] M. J. Khan, "Applications of case-based reasoning in Software Engineering: a systematic mapping study," *IET Software*, vol. 8, no. 6, pp. 258-268, 2014.

[152] M. M. Richter and R. O. Weber, *Case-based reasoning*. Springer, 2016.

[153] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," ed. Cambridge: Cambridge University Press, 2008.

[154] B. G. Batchelor, *Pattern recognition: ideas in practice*. Springer Science & Business Media, 2012.

[155] R. S. Michalski, R. E. Stepp, and E. Diday, "A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts," in *Progress in pattern recognition*: Elsevier, 1981, pp. 33-56.

[156] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, 2016.

[157] S. C. Shiu and S. K. Pal, "Case-based reasoning: concepts, features and soft computing," *Applied Intelligence*, vol. 21, no. 3, pp. 233-238, 2004.

[158] W. Cheetham and J. Graf, "Case-based reasoning in color matching," in *International Conference on Case-Based Reasoning*, 1997: Springer, pp. 1-12.

[159] H. Kitano and H. Shimazu, "The Experience-Sharing Architecture," *Case-Based Reasoning Experiences: Lessons Learned & Future Directions*, 1996.

[160] P. ElKafrawy and R. A. Mohamed, "COMPARATIVE STUDY OF CASE BASED REASONING SOFTWARE," *International Journal of Scientific Research and Management Studies (IJSRMS)*, vol. 1, no. 6, pp. 224-233, 2014.

[161] "AIAI CBR Shell." http://www.aiai.ed.ac.uk/project/cbr/CBRDistrib/ (accessed.

[162] "FreeCBR." http://freecbr.sourceforge.net/ (accessed 2019).

[163] A. Stahl and T. R. Roth-Berghofer, "Rapid prototyping of CBR applications with the open source tool myCBR," in *European conference on case-based reasoning*, 2008: Springer, pp. 615-629.

[164] "eXiTCBR." http://exitcbr.udg.edu/ (accessed.

[165] S. Bogaerts and D. B. Leake, "Increasing AI Project Effectiveness with Reusable Code Frameworks: A Case Study Using IUCBRF," in *FLAIRS Conference*, 2005, pp. 2-7.

[166] G. Makombe Prof, "An Expose of the Relationship between Paradigm, Method and Design in Research," *The Qualitative Report*, vol. 22, no. 12, pp. 3363-3382, 2017.

[167] N. Mackenzie and S. Knipe, "Research dilemmas: Paradigms, methods and methodology," *Issues in educational research*, vol. 16, no. 2, pp. 193-205, 2006.

[168] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision support systems*, vol. 15, no. 4, pp. 251-266, 1995.

[169] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly*, vol. 28, no. 1, pp. 75-105, 2004.

[170] B. W. Boehm, "A spiral model of software development and enhancement," *Computer*, vol. 21, no. 5, pp. 61-72, 1988.

[171] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580-585, 1985.

[172] R. B. D'Agostino and W. B. Kannel, "Epidemiological background and design: the Framingham Study," *Proceedings of the American Statistical Association sesquicentennial invited paper sessions*, pp. 707-718, 1989.

[173] M. Feinleib, "The Framingham Study: sample selection, follow-up, and methods of analyses," *National Cancer Institute Monograph*, vol. 67, pp. 59-64, 1985.

[174] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[175] D. S. Gray and K. Fujioka, "Use of relative weight and body mass index for the determination of adiposity," *Journal of clinical epidemiology*, vol. 44, no. 6, pp. 545-550, 1991.

[176] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*: Springer, 2009, pp. 875-886.

[177] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42-47, 2012.

[178] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863-905, 2018.

[179] F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," *arXiv preprint arXiv:1711.00837*, 2017.

[180] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data--recommendations for the use of performance metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013: IEEE, pp. 245-251.

[181] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25-36, 2006.

[182] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in *2009 21st IEEE international conference on tools with artificial intelligence*, 2009: IEEE, pp. 59-66.

[183] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[184] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography,* vol. 17, no. 2, pp. 145-151, 2008.

[185] N. Jaspers, "Individualized cardiovascular disease prevention: risk factors, risk prediction, and clinical implementation," Utrecht University, 2019.

[186] P. Mody, A. Gupta, B. Bikdeli, J. F. Lampropulos, and K. Dharmarajan, "Most important articles on cardiovascular disease among racial and ethnic minorities," *Circulation: Cardiovascular Quality and Outcomes,* vol. 5, no. 4, pp. e33-e41, 2012.

[187] J. K. Paulus, B. S. Wessler, C. M. Lundquist, and D. M. Kent, "Effects of Race Are Rarely Included in Clinical Prediction Models for Cardiovascular Disease," *Journal of general internal medicine,* vol. 33, no. 9, pp. 1429-1430, 2018.

[188] U. R. Essien and L. R. Jackson, "Race Effects in CVD Prediction Models," *Journal of general internal medicine,* vol. 34, no. 4, pp. 484-484, 2019.

APPENDICES

Appendix A Ethics Approval

AUTEC Secretariat

Auckland University of Technology
D-88, WU406 Level 4 WU Building City Campus
T: +64 9 921 9999 ext. 8316
E: ethics@aut.ac.nz
www.aut.ac.nz/researchethics

11 April 2017

Dave Parry
Faculty of Design and Creative Technologies

Dear Dave

Ethics Application: 17/82 **Fuzzy ontology case-base reasoning approaches to prediction of Cardiovascular disease**

Thank you for submitting your application for ethical review to the Auckland University of Technology Ethics Committee (AUTEC). I am pleased to confirm that your ethics application has been approved for three years until 11 April 2020.

As part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online through http://www.aut.ac.nz/researchethics. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 11 April 2020;
- A brief report on the status of the project using form EA3, which is available online through http://www.aut.ac.nz/researchethics. This report is to be submitted either when the approval expires on 11 April 2020 or on completion of the project;

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to obtain this. If your research is undertaken within a jurisdiction outside New Zealand, you will need to make the arrangements necessary to meet the legal and ethical requirements that apply there.

To enable us to provide you with efficient service, we ask that you use the application number and study title in all correspondence with us. If you have any enquiries about this application, or anything else, please do contact us at ethics@aut.ac.nz.

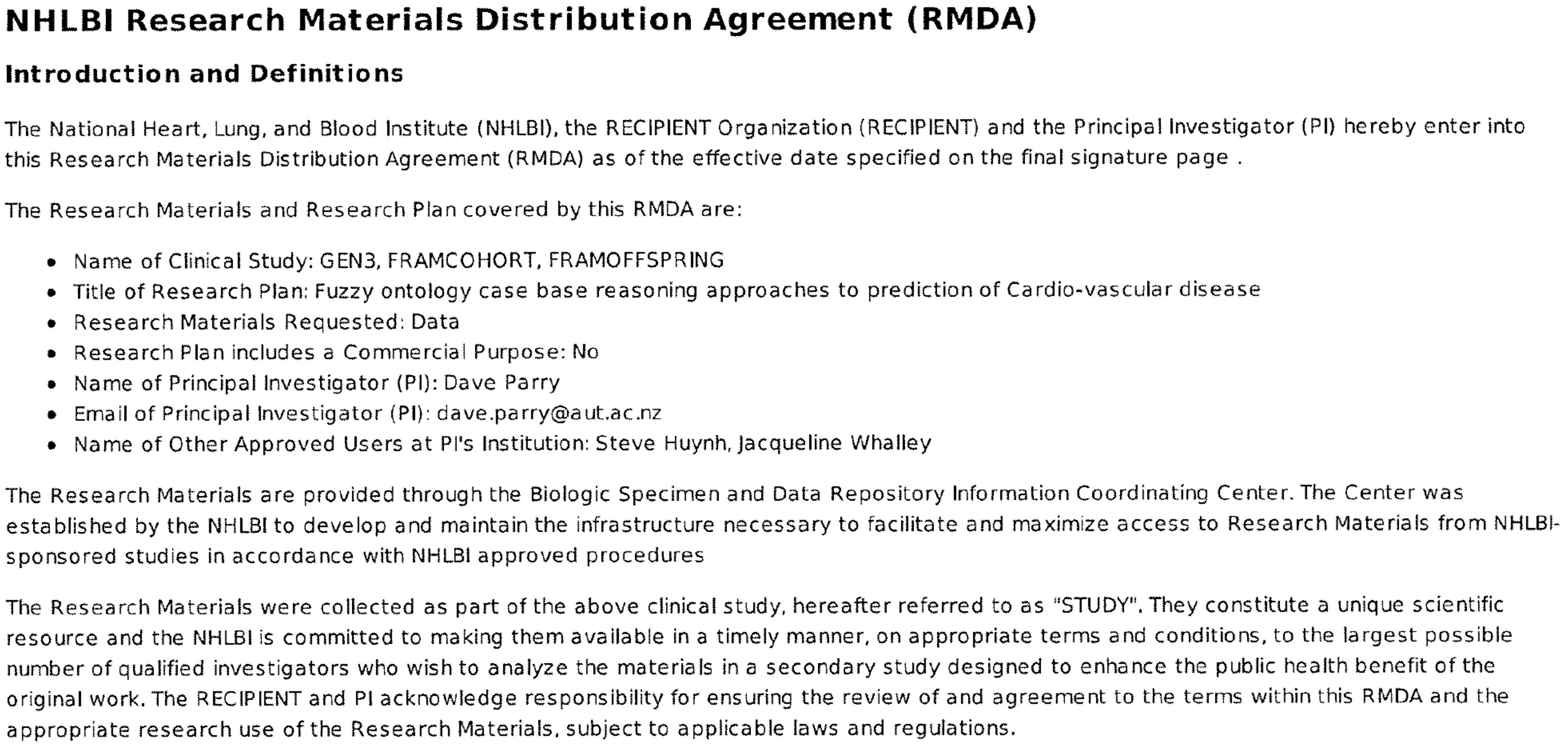
All the very best with your research,

K O'Connor

Kate O'Connor
Executive Secretary
Auckland University of Technology Ethics Committee

Cc:steve.huynh@aut.ac.nz

Appendix B NHLBI RESEARCH MATERIALS DISTRIBUTION AGREEMENT (RMDA)



NHLBI Research Materials Distribution Agreement (RMDA)

Introduction and Definitions

The National Heart, Lung, and Blood Institute (NHLBI), the RECIPIENT Organization (RECIPIENT) and the Principal Investigator (PI) hereby enter into this Research Materials Distribution Agreement (RMDA) as of the effective date specified on the final signature page .

The Research Materials and Research Plan covered by this RMDA are:

- Name of Clinical Study: GEN3, FRAMCOHORT, FRAMOFFSPRING
- Title of Research Plan: Fuzzy ontology case base reasoning approaches to prediction of Cardio-vascular disease
- Research Materials Requested: Data
- Research Plan includes a Commercial Purpose: No
- Name of Principal Investigator (PI): Dave Parry
- Email of Principal Investigator (PI): dave.parry@aut.ac.nz
- Name of Other Approved Users at PI's Institution: Steve Huynh, Jacqueline Whalley

The Research Materials are provided through the Biologic Specimen and Data Repository Information Coordinating Center. The Center was established by the NHLBI to develop and maintain the infrastructure necessary to facilitate and maximize access to Research Materials from NHLBI-sponsored studies in accordance with NHLBI approved procedures

The Research Materials were collected as part of the above clinical study, hereafter referred to as "STUDY". They constitute a unique scientific resource and the NHLBI is committed to making them available in a timely manner, on appropriate terms and conditions, to the largest possible number of qualified investigators who wish to analyze the materials in a secondary study designed to enhance the public health benefit of the original work. The RECIPIENT and PI acknowledge responsibility for ensuring the review of and agreement to the terms within this RMDA and the appropriate research use of the Research Materials, subject to applicable laws and regulations.

The RECIPIENT and PI acknowledge that other researchers are entitled to access to the Research Materials on the same terms as RECIPIENT so that duplication of research may occur. RECIPIENT and PI also recognize that the STUDY Investigators have made a substantial long-term contribution in establishing the Research Materials and the NHLBI encourages appropriate collaborative relationships by outside investigators with the STUDY Investigators and proper acknowledgement of their contributions.

The NHLBI believes that the confidentiality and privacy of the STUDY participants can best be assured by requiring all who are interested in accessing the Research Materials to acknowledge their review of this RMDA and agree to adhere to its provisions. Violation of its confidentiality provisions could lead to legal action on the part of STUDY participants, their families, or the U.S. Government.

Note: RECIPIENT requests access to NHLBI Research Materials for its PI at its sole risk.

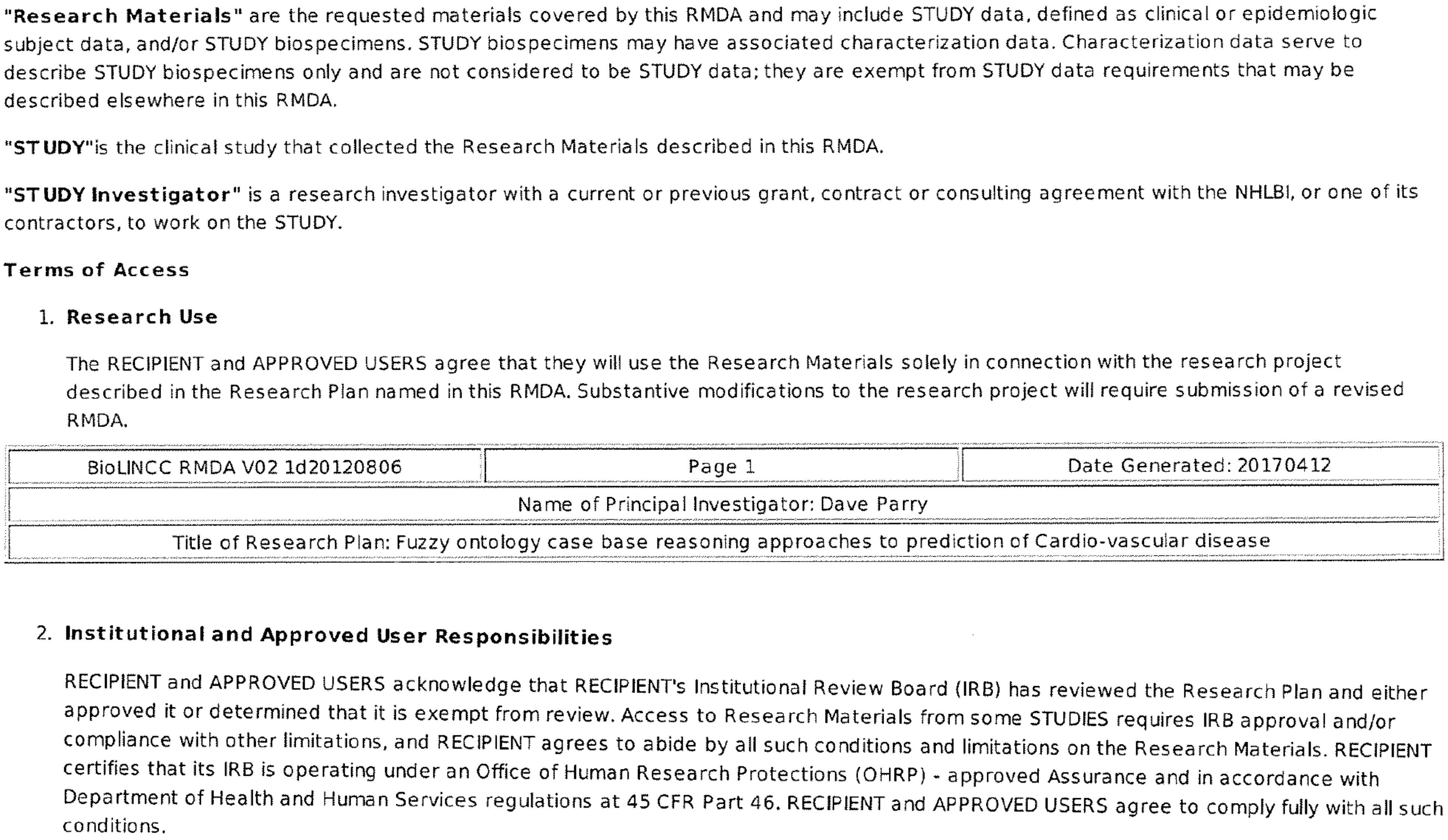
For the purpose of this Agreement

"RECIPIENT" is any organization that is seeking access to STUDY Research Materials, and may be a: Public/State Controlled Institution of Higher Education; Private Institution of Higher Education; Nonprofit organization with 501(c)(3) IRS Status (Other than Institution of Higher Education); Nonprofit Organization without 501(c)(3) IRS Status (Other than Institution of Higher Education); Small Business; For-Profit Organization (Other than Small Business); State Government; Government of a U.S. Territory or Possession; Non-domestic (non-U.S.) Entity (Foreign Organization); or Eligible Agency of the U.S. Government.

"Principal Investigator (PI)" is an individual judged by the RECIPIENT to have the appropriate level of authority and responsibility to lead the scientific investigation proposed in the Research Plan using the requested materials, oversee the supporting staff who are provided access to the Research Materials and contribute to the analytic effort and public disclosure of STUDY results, and assume responsibility for all team members' compliance with the terms and conditions of this RMDA.

"APPROVED USERS" are all individuals specifically identified in the Research Plan, including the PI. Only individuals listed in the Research Plan may have access to the Research Materials.

"Research Plan" is a description of the proposed research that includes the identities of the investigators participating in the research effort. The Research Plan must include the project title, the RECIPIENT's name, the PI's name, the name of other APPROVED USERS, and the proposed research protocol with the research objectives and design. For plans including biospecimens, the biospecimen material type, number, minimum volume, and required characteristics needed to meet the objectives of the protocol must also be included.



"Research Materials" are the requested materials covered by this RMDA and may include STUDY data, defined as clinical or epidemiologic subject data, and/or STUDY biospecimens. STUDY biospecimens may have associated characterization data. Characterization data serve to describe STUDY biospecimens only and are not considered to be STUDY data; they are exempt from STUDY data requirements that may be described elsewhere in this RMDA.

"STUDY"is the clinical study that collected the Research Materials described in this RMDA.

"STUDY Investigator" is a research investigator with a current or previous grant, contract or consulting agreement with the NHLBI, or one of its contractors, to work on the STUDY.

Terms of Access

1. **Research Use**

   The RECIPIENT and APPROVED USERS agree that they will use the Research Materials solely in connection with the research project described in the Research Plan named in this RMDA. Substantive modifications to the research project will require submission of a revised RMDA.

BioLINCC RMDA V02 1d20120806	Page 1	Date Generated: 20170412
Name of Principal Investigator: Dave Parry		
Title of Research Plan: Fuzzy ontology case base reasoning approaches to prediction of Cardio-vascular disease		

2. **Institutional and Approved User Responsibilities**

   RECIPIENT and APPROVED USERS acknowledge that RECIPIENT's Institutional Review Board (IRB) has reviewed the Research Plan and either approved it or determined that it is exempt from review. Access to Research Materials from some STUDIES requires IRB approval and/or compliance with other limitations, and RECIPIENT agrees to abide by all such conditions and limitations on the Research Materials. RECIPIENT certifies that its IRB is operating under an Office of Human Research Protections (OHRP) - approved Assurance and in accordance with Department of Health and Human Services regulations at 45 CFR Part 46. RECIPIENT and APPROVED USERS agree to comply fully with all such conditions.

RECIPIENT and APPROVED USERS agree to report promptly to the NHLBI any proposed change in the Research Plan and any unanticipated problems involving risks to subjects or others. Changes to the Research Plan include changes in the APPROVED USERS list. This RDMA is made in addition to, and does not supersede, any of RECIPIENT's institutional policies or any local, State, and/or Federal laws and regulations that provide additional protections for human subjects.

Evidence of local IRB review and/or approval (where appropriate) from an expedited or convened review to conduct the Research Plan with the requested STUDY data must be included in a supplemental Adobe PDF document that will be uploaded during the application process and attached to the RMDA form.

3. **Public Posting of Approved User's Research Use Statement**

The RECIPIENT and PI agree that information about the proposed research use can be posted on a public web site that describes the project(s) included in the Research Plan. The information will include the PI's name, RECIPIENT institution, project title, and a brief summary of the research. In addition, citations resulting from the use of Research Materials may be posted on the Biologic Specimen and Data Repository Information Coordinating Center Website.

4. **Non-Identification**

The PI agrees not to use the Research Materials, either alone or in concert with any other information, to identify or contact individual STUDY subjects without specific approval to contact STUDY subjects obtained from the IRB(s) responsible for the STUDY.

5. **Non-Transferability of Research Materials**

The RECIPIENT and PI agree to retain control over the Research Materials, and further agree not to release or distribute Research Materials in any form to any entity or individual unless required by NHLBI policies. The RECIPIENT and PI agree to store Research Material data on a computer with adequate security controls (see Section 6), and to maintain appropriate control over the Research Materials at all times. Research Materials data containing individual-level information, in whole or in part, may not be sold to any entity or individual at any point in time for any purpose.

The PI agrees that if his or her relationship with the RECIPIENT terminates and a relationship with a different RECIPIENT is established during the period of the RMDA, a new RMDA from the second RECIPIENT will be submitted and approved before the PI resumes use of the Research Materials. Any versions of Research Material data stored at the first RECIPIENT will be destroyed and their destruction documented. However, if advance written notice and approval by the NHLBI Program Office is obtained to transfer responsibility for the approved Research Plan to a different PI with a relationship with the first RECIPIENT, the Research Material data may not need to be destroyed.

6. **Security of Research Materials**

The RECIPIENT and PI agree to store Research Material data on a computer with security controls adequate to protect sensitive or identifiable information, to ensure that only approved, supervised persons have access to the data, and to maintain appropriate control over the Research Materials at all times. Hard copies of any Research Material must similarly be stored under conditions sufficiently secure to avoid inappropriate access, and shredded prior to discarding.

This RMDA will be in effect for a period of three (3) years from its effective date for the requested STUDY data set. At the end of the three (3) year period, the RECIPIENT and PI agree to destroy all copies of the STUDY data, and all derivatives that contain individual-level information. Characterization data associated with the STUDY biospecimens are exempt from this requirement.

An extension of this RMDA may be permitted by the NHLBI upon submission by the PI and RECIPIENT of evidence of IRB approval for the extended period.

7. **Intellectual Property**

By requesting access to the STUDY Research Materials, the REQUESTER and APPROVED USERS acknowledge the intent of the NHLBI to see that anyone authorized for research access through the attached Research Plan, follow the intellectual property principles within the NIH GWAS Policy for Data Sharing as summarized below:

> Achieving maximum public benefit is the ultimate goal of Research Material distribution through the NHLBI Biological and Data Repository Information Coordinating Center. The NIH believes that Research Materials, such as these covered by this RMDA, should be considered as pre-competitive, and urges APPROVED USERS to avoid making IP claims derived directly from the STUDY Research Materials. However, the NIH also recognizes the importance of the subsequent development of IP on downstream discoveries, especially in therapeutics, which will be necessary to support full investment in products to benefit the public.
>
> It is expected that these NHLBI-provided data, and conclusions derived there from, will remain freely available, without requirement for licensing. The NIH encourages broad use of shared Research Materials coupled with a responsible approach to management of intellectual property derived from downstream discoveries in a manner consistent with the NIH's Best Practices for the Licensing of Genomic Inventions and the NIH Research Tools Policy.

8. **Acknowledgement of BioLINCC Research Resources**

RECIPIENT agrees to acknowledge the contribution of the STUDY in all oral and written presentations, disclosures, or publications resulting from any analyses conducted on the STUDY Research Materials.

If the Research Plan involves collaboration with STUDY Investigators, then the APPROVED USERS will comply with all policies established by the STUDY's publications committee. In addition, the APPROVED USERS will acknowledge the source of the data by including language similar to the following either in the acknowledgment or in the text of the manuscript: "This manuscript was prepared using GEN3, FRAMCOHORT, FRAMOFFSPRING Research Materials obtained from the NHLBI".

If the Research Plan does not involve collaboration with STUDY Investigators or the STUDY has ended, the RECIPIENT will acknowledge the source of the data by including language similar to the following either in the acknowledgment or in the text of the manuscript: "This Manuscript was prepared using GEN3, FRAMCOHORT, FRAMOFFSPRING Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the GEN3, FRAMCOHORT, FRAMOFFSPRING or the NHLBI." Manuscripts and abstracts resulting from the Research Plan should not use the name of the STUDY in the title of the manuscript/abstract unless the title clearly denotes the source of the Research Materials as being from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (e.g., "...An investigation using the <STUDY name and Research Materials>"). The purpose is to delineate manuscripts from the Research PI and APPROVED USERS from manuscripts from the STUDY and STUDY Investigators.

The RECIPIENT and PI agree to ensure that all APPROVED USERS will not include in any manuscripts derived from Research Materials any case studies that describe the characteristics of individual participants, or a small number or groups of participants.

9. **Research Use Reporting**

Prompt publication or other public disclosure of the results of the Research Plan is encouraged.

When requested by the NHLBI, the APPROVED USERS agree to provide general comments regarding topics such as the effectiveness of the NHLBI Biological Specimen and Data Repository Information Coordinating Center Research Material access process (ease of access and use; appropriateness of STUDY data format; challenges in following the policies; suggestions for improving research material access; or the program in general).

10. **Non-Endorsement, Indemnification**

The RECIPIENT and PI acknowledge that although all reasonable efforts have been taken to ensure the accuracy and reliability of Research Materials, the NHLBI, and STUDY Investigators do not and cannot warrant the results that may be obtained by using any Research Materials included therein. The NHLBI and all contributors to these Research Materials disclaim all warranties as to performance or fitness of the Research Materials for any particular purpose.

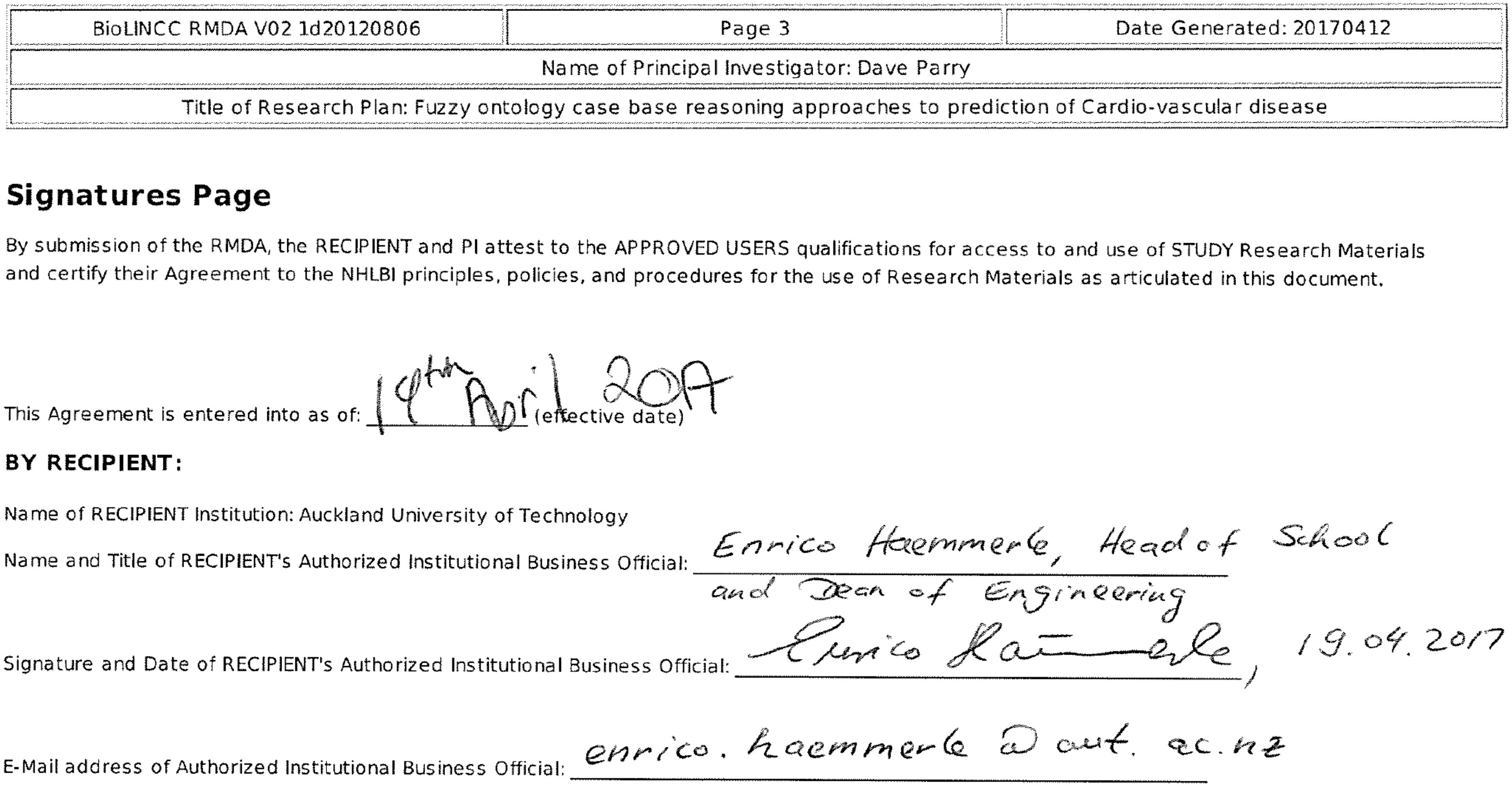
No indemnification for any loss, claim, damage or liability is intended or provided by any party under this Agreement. Each party shall be liable for any loss, claim, damage, or liability that said party incurs as a result of its activities under this Agreement, except that the NIH, as an agency of the United States, assumes liability only to the extent provided under the Federal Tort Claims Act, 28 U.S.C. 2671 et seq.

11. **Termination and Violations**

The NHLBI may terminate this Agreement if RECIPIENT or APPROVED USERS are in default of any of its conditions and such default has not been remedied within 30 days after the date of written notice of such default by an authorized representative of the NHLBI. Past violations will be taken into consideration by the NHLBI for future requests from the RECIPIENT and APPROVED USERS to access NHLBI Research Materials.

12. **Amendments**

Amendments to this Agreement must be made in writing and signed by authorized representatives of all parties.



BioLINCC RMDA V02 1d20120806	Page 3	Date Generated: 20170412
Name of Principal Investigator: Dave Parry		
Title of Research Plan: Fuzzy ontology case base reasoning approaches to prediction of Cardio-vascular disease		

Signatures Page

By submission of the RMDA, the RECIPIENT and PI attest to the APPROVED USERS qualifications for access to and use of STUDY Research Materials and certify their Agreement to the NHLBI principles, policies, and procedures for the use of Research Materials as articulated in this document.

This Agreement is entered into as of: 19th April 2017 (effective date)

BY RECIPIENT:

Name of RECIPIENT Institution: Auckland University of Technology

Name and Title of RECIPIENT's Authorized Institutional Business Official: Enrico Haemmerle, Head of School and Dean of Engineering

Signature and Date of RECIPIENT's Authorized Institutional Business Official: Enrico Haemmerle, 19.04.2017

E-Mail address of Authorized Institutional Business Official: enrico.haemmerle@aut.ac.nz

BY PRINCIPAL INVESTIGATOR:

Name: Dave Parry

Title: Head of Dept. of Computer Science AUT

Surface Mail Address: AUT Tower, Wakefield Street.
Auckland New Zealand.

E-Mail Address: dave.parry@aut.ac.nz

Telephone Number: +64 9921 9999 xtn 8918.

Fax Number: +64 9921 9944

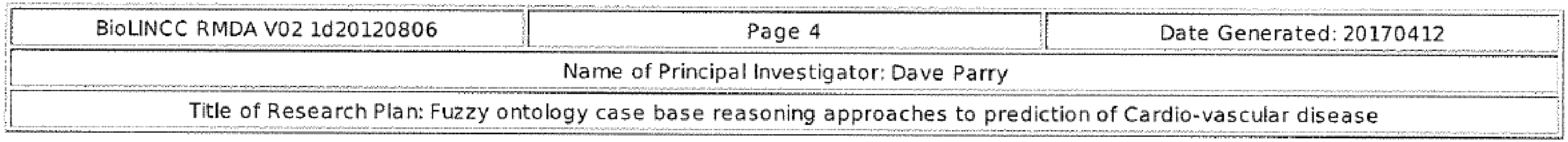
Signature and Date: 19th April 2017.

BY NHLBI Authorized Representative:

Name and Title: Sean A. Coady -S

Signature and Date: Digitally signed by Sean A. Coady -S
DN: c=US, o=U.S. Government, ou=HHS, ou=NIH, ou=People, cn=Sean A. Coady -S, 0.9.2342.19200300.100.1.1=0010154059
Date: 2017.04.21 08:17:47 -04'00'

"**Authorized Institutional Business/Signing Official**" is an individual with the authority to enter into business transactions on behalf of the RECIPIENT.



BioLINCC RMDA V02 1d20120806	Page 4	Date Generated: 20170412
Name of Principal Investigator: Dave Parry		
Title of Research Plan: Fuzzy ontology case base reasoning approaches to prediction of Cardio-vascular disease		

Appendix C ANALYSIS FOR INITIAL ATTRIBUTE REDUCTION

Table C-1 shows analysis to initially remove 16 unsuitable attributes from the initially collected dataset (section 3.8.3).

Table C-1: Analysis to initially remove unsuitable attributes

Attribute	Percentage of missing data (%)	Decision
CONFIRMATION TYPE 3	99.95916701	Remove
T4	84.99387505	Remove
PHYSICIAN SYSTOLIC BLOOD PRESSURE, 2ND	82.48264598	Remove
PHYSICIAN DIASTOLIC BLOOD PRESSURE, 2ND	82.48264598	Remove
IF STOPPED, AGE STOPPED	33.85055125	
AGE START SMOKE CIGARETTE REGULARLY	33.5238873	
WEIGHT AT AGE 25	16.33319722	
COMPLETE BLOOD COUNT	11.47407105	
WHITE BLOOD COUNT	10.73907717	
RED BLOOD COUNT	10.41241323	
H.G.B.	10.41241323	
M.C.V.	10.41241323	
M.C.H.	10.41241323	
M.C.H.C.	10.41241323	
H.C.T.	6.206614945	Remove (Duplicated with the HEMATOCRIT column)
HEMATOCRIT	5.532870559	
CALCIUM	4.838709677	
ALBUMIN	4.797876684	
PHOSPHORUS	4.757043691	

BUN	4.757043691
TOTAL PROTEIN	4.757043691
LDL	4.757043691
Diabetes	4.654961209
SGOT	4.430379747
ALKALINE PHOSPHOTASE	4.389546754
URIC ACID	4.287464271
GLUCOSE	4.205798285
TOTAL BILIRUBIN	4.205798285
GLOBULIN	4.185381788
WEIGHT COMPARED WITH 1 MONTH AGO	2.53164557
TOP FRACTION, ORIGIN	2.164148632
TOP FRACTION, BETA	2.164148632
TOP FRACTION, PRE-BETA	2.164148632
BOTTOM FRACTION, PRE-BETA	2.123315639
SINKING PRE-BETA BAND	2.102899143
FREDERICKSON CLASSIFICATION	2.102899143
FASTING 12 HRS OR MORE	2.082482646
PRE-BETA BAND	2.041649653
HDL CHOLESTEROL	2.021233156
VLDL CHOLESTEROL	2.021233156
LDL CHOLESTEROL	2.021233156
WHOLE PLASMA, ORIGIN	1.93956717
WHOLE PLASMA, PRE-BETA	1.919150674
WEIGHT COMPARE WITH 1 YEAR AGO	1.837484688
TRIGLYCERIDES	1.714985708
TOTAL CHOLESTEROL	1.674152715
WHOLE PLASMA APPEARANCE	1.674152715

INFRANATE AFTER 12 HRS AT 4 DEGREES	1.674152715	
CREAM AFTER 12 HRS OR MORE	1.674152715	
USES FILTER	1.53123724	
INHALES	1.347488771	
ALCOHOL INDEX	1.06165782	
COCKTAIL INTAKE	1.041241323	
FIRST SECOND VOLUME	1.020824826	
TOTAL VITAL CAPACITY	1.020824826	
SMOKES CIGARS	0.979991833	
PAROXYSMAL NOCTURAL DYSPNEA	0.979991833	
BILATERAL ANKLE EDEMA	0.979991833	
NOCTURNAL COUGH OR WHEEZING	0.979991833	
DYSPNEA ON EXERTION	0.959575337	
SMOKES PIPES	0.93915884	
RECENT ORTHOPNEA	0.93915884	
HISTORY OF ENLARGED HEART	0.918742344	
DYSPNEA INCREASE IN PAST 2 YEARS	0.918742344	
HISTORY OF HYPOTHYROID DISEASE	0.898325847	
WINE INTAKE	0.898325847	
BEER INTAKE	0.857492854	
HISTORY OF OTHER KIDNEY AILMENT	0.816659861	
HISTORY OF NEPHROSIS	0.796243365	
BROCHODILATOR OR AEROSOL	0.755410372	
NURSE SYSTOLIC BLOOD PRESSURE	0.734993875	Remove (Use the physicians’ one)
NURSE DIASTOLIC BLOOD PRESSURE	0.734993875	Remove (Use the physicians’ one)
TRANQUILIZERS	0.714577379	
OTHER (C-V DRUGS)	0.694160882	
HISTORY OF HEART MURMUR	0.694160882	

DIURETICS FOR BLOOD PRESSURE	0.653327889	
HYPOTENSIVES (EXCLUDING DIURETICS)	0.592078399	
ANTI-THYROID	0.551245406	
ANTI-COAGULANTS	0.551245406	
THYROID	0.53082891	
HYPOGLYCEMIC AGENTS (SPECIFY)	0.510412413	
LOW CALORIE DIET LAST 2 WEEKS	0.510412413	
Treatment for Diabetes	0.510412413	
ANTI-CHOLESTEROL AGENTS	0.489995917	
DIURETICS FOR FLUID RETENTION	0.46957942	
DIABETIC DIET LAST 2 WEEKS	0.449162924	
LOW CHOLESTEROL DIET LAST 2 WEEKS	0.408329931	
CARDIAC GLYCOSIDES	0.387913434	
SMOKES CIGARETTES	0.387913434	
NITRITES	0.367496938	
QUINIDINE	0.367496938	
PREMARIN	0.367496938	Remove (Female only)
SMOKED AT LEAST 1 YEAR	0.347080441	
USUAL # OF CIGARRETTE SMOKE NOW/EVER	0.347080441	
OTHER (SPECIFY)	0.326663944	Remove (Female only)
AMOUNT OF FOOD LAST 2 DAYS	0.326663944	
AMOUNT OF ALCOHOL	0.326663944	
OVARIES REMOVED	0.306247448	Remove (Female only)
HISTORY OF HYPERTENSION	0.183748469	
PHYSICIAN SYSTOLIC BLOOD PRESSURE, 1ST	0.163331972	
PHYSICIAN DIASTOLIC BLOOD PRESSURE, 1ST	0.163331972	

ORAL CONTRACEPTIVE	0.163331972	Remove (Female only)
AGE AT WHICH PERIODS STOPPED	0.102082483	Remove (Female only)
CAUSE OF CESSATION OF MENSES	0.102082483	Remove (Female only)
WOLFF-PARKINSON-WHITE SYNDROME	0.102082483	
CHEST DISCOMFORT	0.081665986	
P-R INTERVAL	0.06124949	
QT INTERAVAL	0.06124949	
NON-SPECIFIC T-WAVE ABNORMALITY	0.06124949	
NON-SPECIFIC S-T SEGMENT ABNORMALITY	0.06124949	
ECG CLINICAL READING	0.06124949	
ECG FINDING SUMMARY	0.040832993	
HYSTERECTOMY	0.040832993	Remove (Female only)
VENTRICULAR RATE	0.040832993	
QRS INTERVAL	0.040832993	
A QRS	0.040832993	
RIGHT-INTRAVENTRICULAR BLOCK	0.040832993	
LEFT-INTRAVENTRICULAR BLOCK	0.040832993	
HEMIBLOCK	0.040832993	
BIFASCULAR BLOCK	0.040832993	
INCOMPLETE-ATRIOVENTRICULAR BLOCK	0.040832993	
COMPLETE ATRIOVENTRICULAR BLOCK	0.040832993	
PREMATURE BEATS	0.040832993	
OTHER ARRHYTHMIA	0.040832993	
OTHER ECG ABNORMALITY	0.040832993	
TAKING DIGITALIS OR QUINIDINE	0.040832993	
MYOCARDIAL INFARCTION	0.040832993	
LEFT VENTRICULAR HYPERTROPHY	0.040832993	
METROPOLITAN RELATIVE WEIGHT	0.020416497	

QUETELET INDEX (KG/M SQUARED)	0.020416497	Remove (Duplicated with the BMI column)
HGT	0.020416497	
BMI	0.020416497	
PID	0	
SEX	0	
PERIODS HAVE STOPPED 1 YR OR MORE	0	Remove (Female only)
WGTGP	0	
AGE	0	
CVD	0	
CVDDATE	0	

Appendix D MIXED SEX DATASET—FIRST ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka’s InfoGainAttributeEval attribute evaluator to rank 119 risk factors of the mixed sex dataset (section 3.8.7).

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring9_PreparePredictionAttributes2-
weka.filters.unsupervised.attribute.Remove-R122-weka.filters.unsupervised.attribute.Remove-
R1
Instances:    4737
Attributes:   120
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data




=== Attribute Selection on all input data ===


Search Method:
                                                    Attribute ranking.


Attribute Evaluator (supervised, Class (nominal): 120 cvd10):
                                                    Information Gain Ranking Filter


Ranked attributes:
 0.04155809   116 AGE
 0.02066707     2 TOTAL CHOLESTEROL
 0.01658859     4 VLDL CHOLESTEROL
 0.0165876      5 LDL CHOLESTEROL
 0.01627317    33 PHYSICIAN SYSTOLIC BLOOD PRESSURE
 0.01371721    34 PHYSICIAN DIASTOLIC BLOOD PRESSURE
 0.01291542     6 TRIGLYCERIDES
 0.01133067    62 USUAL # OF CIGARRETTE SMOKE NOW/EVER
 0.00974304    23 GLUCOSE
 0.00893359     3 HDL CHOLESTEROL
```

```
0.00853935   117 BMI
0.00798268    79 DYSPNEA ON EXERTION
0.00758577     1 SEX
0.00752456    31 LDL
0.00731425    30 ALKALINE PHOSPHOTASE
0.00697716    25 URIC ACID
0.00626734   115 WGTGP
0.00617897    41 HISTORY OF HYPERTENSION
0.00606969    56 SMOKED AT LEAST 1 YEAR
0.00593551    20 HEMATOCRIT
0.0057986     86 WHITE BLOOD COUNT
0.00570935    19 FREDERICKSON CLASSIFICATION
0.00566588   119 Diabetes
0.00564343    80 DYSPNEA INCREASE IN PAST 2 YEARS
0.00560985    88 H.G.B.
0.00547475    11 TOP FRACTION PRE-BETA
0.005308      87 RED BLOOD COUNT
0.00528462    96 A QRS
0.00519749    59 SMOKES CIGARETTES
0.00489626    52 HYPOGLYCEMIC AGENTS (SPECIFY)
0.00489626   118 Treatment for Diabetes
0.00416658    17 PRE-BETA BAND
0.0041151     47 HYPOTENSIVES (EXCLUDING DIURETICS)
0.00406476     8 WHOLE PLASMA PRE-BETA
0.00381172    27 ALBUMIN
0.00376564    46 DIURETICS FOR BLOOD PRESSURE
0.00329209    64 INHALES
0.00320043    76 CHEST DISCOMFORT
0.00298447    67 WEIGHT AT AGE 25
0.00284326    35 FIRST SECOND VOLUME
0.00282801    73 COCKTAIL INTAKE
0.00277396    37 ECG FINDING SUMMARY
0.00271187   112 ECG CLINICAL READING
0.00259362    95 QT INTERAVAL
0.00256636    36 TOTAL VITAL CAPACITY
0.00230398   110 NON-SPECIFIC T-WAVE ABNORMALITY
0.00227035    42 CARDIAC GLYCOSIDES
0.00195175    81 RECENT ORTHOPNEA
0.00183344    22 PHOSPHORUS
```

```
0.001739     24 BUN
0.00146816  107 TAKING DIGITALIS OR QUINIDINE
0.00131895   14 INFRANATE AFTER 12 HRS AT 4 DEGREES
0.0012996    48 ANTI-CHOLESTEROL AGENTS
0.00128142    7 WHOLE PLASMA ORIGIN
0.00127889    9 TOP FRACTION ORIGIN
0.00125359  109 LEFT VENTRICULAR HYPERTROPHY
0.00120901   57 SMOKES CIGARS
0.00116796  111 NON-SPECIFIC S-T SEGMENT ABNORMALITY
0.00115491   43 NITRITES
0.00115229   53 TRANQUILIZERS
0.00094572   13 WHOLE PLASMA APPEARANCE
0.00093981   44 QUINIDINE
0.00078236   55 OTHER (C-V DRUGS)
0.00062946  104 PREMATURE BEATS
0.00061819  106 OTHER ECG ABNORMALITY
0.00058816   84 NOCTURNAL COUGH OR WHEEZING
0.00049601   16 FASTING 12 HRS OR MORE
0.00048836  102 COMPLETE ATRIOVENTRICULAR BLOCK
0.00046391  101 INCOMPLETE-ATRIOVENTRICULAR BLOCK
0.00046373   75 AMOUNT OF ALCOHOL
0.00044647   15 CREAM AFTER 12 HRS OR MORE
0.00041847   74 AMOUNT OF FOOD LAST 2 DAYS
0.00039265   70 DIABETIC DIET LAST 2 WEEKS
0.00032651   82 PAROXYSMAL NOCTURAL DYSPNEA
0.00030627   40 HISTORY OF HYPOTHYROID DISEASE
0.00027141   78 HISTORY OF ENLARGED HEART
0.00027115   83 BILATERAL ANKLE EDEMA
0.0002689    68 LOW CHOLESTEROL DIET LAST 2 WEEKS
0.00025458   18 SINKING PRE-BETA BAND
0.00025406   12 BOTTOM FRACTION PRE-BETA
0.0002421    99 HEMIBLOCK
0.00021993   45 DIURETICS FOR FLUID RETENTION
0.0001946    63 USES FILTER
0.00018558  103 WOLFF-PARKINSON-WHITE SYNDROME
0.0001834    98 LEFT-INTRAVENTRICULAR BLOCK
0.00017086   77 HISTORY OF HEART MURMUR
0.00015389   51 ANTI-COAGULANTS
0.00013661   54 BROCHODILATOR OR AEROSOL
```

```
 0.0001234    105 OTHER ARRHYTHMIA
 0.00012042    39 HISTORY OF OTHER KIDNEY AILMENT
 0.00011912    50 ANTI-THYROID
 0.00011001    58 SMOKES PIPES
 0.00009091    10 TOP FRACTION BETA
 0.00008331   108 MYOCARDIAL INFARCTION
 0.00007596    66 WEIGHT COMPARED WITH 1 YEAR AGO
 0.00006664   100 BIFASCULAR BLOCK
 0.00003568    69 LOW CALORIE DIET LAST 2 WEEKS
 0.00003507    65 WEIGHT COMPARED WITH 1 MONTH AGO
 0.00001571    38 HISTORY OF NEPHROSIS
 0.00000799    97 RIGHT-INTRAVENTRICULAR BLOCK
 0.00000292    49 THYROID
 0             71 BEER INTAKE
 0            114 HGT
 0             92 VENTRICULAR RATE
 0            113 ALCOHOL INDEX
 0             93 P-R INTERVAL
 0             94 QRS INTERVAL
 0             61 IF STOPPED AGE STOPPED
 0             26 TOTAL PROTEIN
 0             91 M.C.H.C.
 0             28 GLOBULIN
 0             21 CALCIUM
 0             85 COMPLETE BLOOD COUNT
 0             29 TOTAL BILIRUBIN
 0             32 SGOT
 0             89 M.C.V.
 0             90 M.C.H.
 0             72 WINE INTAKE
 0             60 AGE START SMOKE CIGARETTE REGULARLY


Selected attributes:
116,2,4,5,33,34,6,62,23,3,117,79,1,31,30,25,115,41,56,20,86,19,119,80,88,11,87,96,59,52,118,
17,47,8,27,46,64,76,67,35,73,37,112,95,36,110,42,81,22,24,107,14,48,7,9,109,57,111,43,53,13,
44,55,104,106,84,16,102,101,75,15,74,70,82,40,78,83,68,18,12,99,45,63,103,98,77,51,54,105,3
9,50,58,10,108,66,100,69,65,38,97,49,71,114,92,113,93,94,61,26,91,28,21,85,29,32,89,90,72,6
0 : 119
```

Appendix E MIXED SEX DATASET—SECOND ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka's InfoGainAttributeEval attribute evaluator to rank 34 selected risk factors, after removing missing data, of the mixed sex dataset (section 3.8.9).

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring12_RemoveMissingData-
weka.filters.unsupervised.attribute.Remove-R37-weka.filters.unsupervised.attribute.Remove-R1
Instances:    4071
Attributes:   35
              SEX
              TOTAL CHOLESTEROL
              HDL CHOLESTEROL
              VLDL CHOLESTEROL
              LDL CHOLESTEROL
              TRIGLYCERIDES
              WHOLE PLASMA PRE-BETA
              TOP FRACTION PRE-BETA
              PRE-BETA BAND
              FREDERICKSON CLASSIFICATION
              HEMATOCRIT
              GLUCOSE
              URIC ACID
              ALKALINE PHOSPHOTASE
              LDH
              PHYSICIAN SYSTOLIC BLOOD PRESSURE
              PHYSICIAN DIASTOLIC BLOOD PRESSURE
              HISTORY OF HYPERTENSION
              HYPOTENSIVES (EXCLUDING DIURETICS)
              HYPOGLYCEMIC AGENTS (SPECIFY)
              SMOKED AT LEAST 1 YEAR
              SMOKES CIGARETTES
              USUAL # OF CIGARETTE SMOKE NOW/EVER
              DYSPNEA ON EXERTION
              DYSPNEA INCREASE IN PAST 2 YEARS
              WHITE BLOOD COUNT
              RED BLOOD COUNT
              H.G.B.
              A QRS
              WGTGP
              AGE
              BMI
              Treatment for Diabetes
              Diabetes
              cvd10
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
                                                  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 35 cvd10):
                                                  Information Gain Ranking Filter

Ranked attributes:
 0.04384   31 AGE
 0.02324    2 TOTAL CHOLESTEROL
 0.0184     5 LDL CHOLESTEROL
 0.01733    4 VLDL CHOLESTEROL
 0.01562   16 PHYSICIAN SYSTOLIC BLOOD PRESSURE
 0.01392    6 TRIGLYCERIDES
 0.01286   17 PHYSICIAN DIASTOLIC BLOOD PRESSURE
 0.0119    12 GLUCOSE
 0.01112   23 USUAL # OF CIGARETTE SMOKE NOW/EVER
 0.01109    3 HDL CHOLESTEROL
 0.00915   11 HEMATOCRIT
 0.00902   32 BMI
 0.00843   15 LDH
 0.00805    1 SEX
 0.00741   30 WGTGP
 0.00736   13 URIC ACID
 0.00735   10 FREDERICKSON CLASSIFICATION
 0.00725   28 H.G.B.
 0.00719   14 ALKALINE PHOSPHOTASE
 0.00678   26 WHITE BLOOD COUNT
 0.00673   24 DYSPNEA ON EXERTION
 0.00672   34 Diabetes
 0.0067     8 TOP FRACTION PRE-BETA
 0.00652   27 RED BLOOD COUNT
 0.00616   21 SMOKED AT LEAST 1 YEAR
 0.00578   33 Treatment for Diabetes
 0.00578   20 HYPOGLYCEMIC AGENTS (SPECIFY)
 0.00569   29 A QRS
 0.00568   18 HISTORY OF HYPERTENSION
 0.00543    9 PRE-BETA BAND
 0.00535    7 WHOLE PLASMA PRE-BETA
 0.00506   25 DYSPNEA INCREASE IN PAST 2 YEARS
 0.00503   22 SMOKES CIGARETTES
 0.00445   19 HYPOTENSIVES (EXCLUDING DIURETICS)


Selected attributes:
31,2,5,4,16,6,17,12,23,3,11,32,15,1,30,13,10,28,14,26,24,34,8,27,21,33,20,29,18,9,7,25,22,19 :
34
```

Appendix F MALE DATASET—FIRST ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka's InfoGainAttributeEval attribute evaluator to rank 118 risk factors of the male dataset.

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring10_PreparePredictionAttributes-
weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-
R120
Instances:    2256
Attributes:   119
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data




=== Attribute Selection on all input data ===


Search Method:
                                                  Attribute ranking.


Attribute Evaluator (supervised, Class (nominal): 119 cvd10):
                                                  Information Gain Ranking Filter


Ranked attributes:
 0.0592066922   115 AGE
 0.0275554434    35 TOTAL VITAL CAPACITY
 0.023543638      1 TOTAL CHOLESTEROL
 0.0221832382    34 FIRST SECOND VOLUME
 0.0203427569     4 LDL CHOLESTEROL
 0.0200335985    26 ALBUMIN
 0.0133942659    85 WHITE BLOOD COUNT
 0.0129591726    22 GLUCOSE
 0.0122847245     5 TRIGLYCERIDES
 0.0116015904    33 DIASTOLIC BLOOD PRESSURE
```

0.0102338515	32 SYSTOLIC BLOOD PRESSURE
0.0099748177	55 SMOKED AT LEAST 1 YEAR
0.0098996084	3 VLDL CHOLESTEROL
0.0098467173	61 USUAL # OF CIGARRETTE SMOKE NOW/EVER
0.0090644755	58 SMOKES CIGARETTES
0.008899552	28 TOTAL BILIRUBIN
0.0085976877	30 LDH
0.008378051	78 DYSPNEA ON EXERTION
0.0076409857	51 HYPOGLYCEMIC AGENTS (SPECIFY)
0.0076409857	117 Treatment for Diabetes
0.0074232927	95 A QRS
0.0074042504	118 Diabetes
0.0060329042	2 HDL CHOLESTEROL
0.005622893	75 CHEST DISCOMFORT
0.0052735685	110 NON-SPECIFIC S-T SEGMENT ABNORMALITY
0.0051490572	109 NON-SPECIFIC T-WAVE ABNORMALITY
0.0048315193	63 INHALES
0.0045464337	18 FREDERICKSON CLASSIFICATION
0.004538698	29 ALKALINE PHOSPHOTASE
0.004448376	94 QT INTERAVAL
0.0041216431	25 TOTAL PROTEIN
0.003651327	36 ECG FINDING SUMMARY
0.0034912303	40 HISTORY OF HYPERTENSION
0.003490247	21 PHOSPHORUS
0.003448162	10 TOP FRACTION PRE-BETA
0.00335672	16 PRE-BETA BAND
0.0032350604	7 WHOLE PLASMA PRE-BETA
0.0032248113	111 ECG CLINICAL READING
0.0027752659	79 DYSPNEA INCREASE IN PAST 2 YEARS
0.0027078788	45 DIURETICS FOR BLOOD PRESSURE
0.0025348444	82 BILATERAL ANKLE EDEMA
0.0022848854	54 OTHER (C-V DRUGS)
0.0022079623	46 HYPOTENSIVES (EXCLUDING DIURETICS)
0.0021062347	80 RECENT ORTHOPNEA
0.0020648841	6 WHOLE PLASMA ORIGIN
0.0020600986	8 TOP FRACTION ORIGIN
0.001581056	44 DIURETICS FOR FLUID RETENTION
0.0015237592	76 HISTORY OF HEART MURMUR
0.0013126402	52 TRANQUILIZERS

0.0008738141	74	AMOUNT OF ALCOHOL
0.0008409894	47	ANTI-CHOLESTEROL AGENTS
0.0008044509	101	COMPLETE ATRIOVENTRICULAR BLOCK
0.0007748173	11	BOTTOM FRACTION PRE-BETA
0.0007748173	17	SINKING PRE-BETA BAND
0.0007452462	62	USES FILTER
0.0007381664	14	CREAM AFTER 12 HRS OR MORE
0.0007280777	103	PREMATURE BEATS
0.0006743346	108	LEFT VENTRICULAR HYPERTROPHY
0.0006574618	83	NOCTURNAL COUGH OR WHEEZING
0.0006451987	100	INCOMPLETE-ATRIOVENTRICULAR BLOCK
0.0005806095	38	HISTORY OF OTHER KIDNEY AILMENT
0.0005695691	56	SMOKES CIGARS
0.0005478234	68	LOW CALORIE DIET LAST 2 WEEKS
0.0005239122	41	CARDIAC GLYCOSIDES
0.0004631239	97	LEFT-INTRAVENTRICULAR BLOCK
0.0004060788	105	OTHER ECG ABNORMALITY
0.0003593274	13	INFRANATE AFTER 12 HRS AT 4 DEGREES
0.0003568099	43	QUINIDINE
0.0003568099	42	NITRITES
0.0003512771	98	HEMIBLOCK
0.0002511922	64	WEIGHT COMPARED WITH 1 MONTH AGO
0.0002443124	73	AMOUNT OF FOOD LAST 2 DAYS
0.0002424533	9	TOP FRACTION BETA
0.0002381329	12	WHOLE PLASMA APPEARANCE
0.0002190216	15	FASTING 12 HRS OR MORE
0.0002107705	65	WEIGHT COMPARE WITH 1 YEAR AGO
0.0001897813	39	HISTORY OF HYPOTHYROID DISEASE
0.0001893646	67	LOW CHOLESTEROL DIET LAST 2 WEEKS
0.0001541605	107	MYOCARDIAL INFARCTION
0.0001463647	81	PAROXYSMAL NOCTURAL DYSPNEA
0.0001027499	99	BIFASCULAR BLOCK
0.0000943372	50	ANTI-COAGULANTS
0.0000811353	57	SMOKES PIPES
0.0000729965	77	HISTORY OF ENLARGED HEART
0.0000721089	48	THYROID
0.0000705051	104	OTHER ARRHYTHMIA
0.0000483676	49	ANTI-THYROID
0.0000474103	37	HISTORY OF NEPHROSIS

```
 0.0000022624   102 WOLFF-PARKINSON-WHITE SYNDROME
 0.0000014008    96 RIGHT-INTRAVENTRICULAR BLOCK
 0.0000000551    53 BROCHODILATOR OR AEROSOL
 0.0000000394    69 DIABETIC DIET LAST 2 WEEKS
 0               31 SGOT
 0               89 M.C.H.
 0               88 M.C.V.
 0               91 VENTRICULAR RATE
 0               87 H.G.B.
 0               90 M.C.H.C.
 0               92 P-R INTERVAL
 0               84 COMPLETE BLOOD COUNT
 0              113 HGT
 0               93 QRS INTERVAL
 0              116 BMI
 0              114 WGTGP
 0               86 RED BLOOD COUNT
 0               72 COCKTAIL INTAKE
 0               27 GLOBULIN
 0               71 WINE INTAKE
 0               23 BUN
 0               24 URIC ACID
 0              106 TAKING DIGITALIS OR QUINIDINE
 0              112 ALCOHOL INDEX
 0               60 IF STOPPED AGE STOPPED
 0               20 CALCIUM
 0               70 BEER INTAKE
 0               66 WEIGHT AT AGE 25
 0               19 HEMATOCRIT
 0               59 AGE START SMOKE CIGARETTE REGULARLY

Selected attributes:
115,35,1,34,4,26,85,22,5,33,32,55,3,61,58,28,30,78,51,117,95,118,2,75,110,109,63,18,29,94,25
,36,40,21,10,16,7,111,79,45,82,54,46,80,6,8,44,76,52,74,47,101,11,17,62,14,103,108,83,100,38
,56,68,41,97,105,13,43,42,98,64,73,9,12,15,65,39,67,107,81,99,50,57,77,48,104,49,37,102,96,5
3,69,31,89,88,91,87,90,92,84,113,93,116,114,86,72,27,71,23,24,106,112,60,20,70,66,19,59 :
118
```

Appendix G MALE DATASET—SECOND ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka's InfoGainAttributeEval attribute evaluator to rank 23 selected risk factors, after removing missing data, of the male dataset.

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring14-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R25
Instances:    1974
Attributes:   24
              TOTAL CHOLESTEROL
              HDL CHOLESTEROL
              VLDL CHOLESTEROL
              LDL CHOLESTEROL
              TRIGLYCERIDES
              GLUCOSE
              ALBUMIN
              TOTAL BILIRUBIN
              LDH
              SYSTOLIC BLOOD PRESSURE
              DIASTOLIC BLOOD PRESSURE
              FIRST SECOND VOLUME
              TOTAL VITAL CAPACITY
              HYPOGLYCEMIC AGENTS (SPECIFY)
              SMOKED AT LEAST 1 YEAR
              SMOKES CIGARETTES
              USUAL # OF CIGARRETTE SMOKE NOW/EVER
              DYSPNEA ON EXERTION
              WHITE BLOOD COUNT
              A QRS
              AGE
              Treatment for Diabetes
              Diabetes
              cvd10
Evaluation mode:    evaluate on all training data
```

```
=== Attribute Selection on all input data ===

Search Method:
                                                  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 24 cvd10):
                                                  Information Gain Ranking Filter

Ranked attributes:
 0.05941   21 AGE
 0.02679    1 TOTAL CHOLESTEROL
 0.02532   12 FIRST SECOND VOLUME
 0.02421   13 TOTAL VITAL CAPACITY
 0.0224     4 LDL CHOLESTEROL
 0.02112    7 ALBUMIN
 0.01598   19 WHITE BLOOD COUNT
 0.01556    6 GLUCOSE
 0.01411    5 TRIGLYCERIDES
 0.01013    8 TOTAL BILIRUBIN
 0.00976    9 LDH
 0.00961   17 USUAL # OF CIGARRETTE SMOKE NOW/EVER
 0.00941   15 SMOKED AT LEAST 1 YEAR
 0.00937   11 DIASTOLIC BLOOD PRESSURE
 0.00921    3 VLDL CHOLESTEROL
 0.00881   16 SMOKES CIGARETTES
 0.00874   18 DYSPNEA ON EXERTION
 0.00862   14 HYPOGLYCEMIC AGENTS (SPECIFY)
 0.00862   22 Treatment for Diabetes
 0.00806    2 HDL CHOLESTEROL
 0.00788   10 SYSTOLIC BLOOD PRESSURE
 0.00779   23 Diabetes
 0.00685   20 A QRS

Selected attributes: 21,1,12,13,4,7,19,6,5,8,9,17,15,11,3,16,18,14,22,2,10,23,20 : 23
```

Appendix H FEMALE DATASET—FIRST ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka's InfoGainAttributeEval attribute evaluator to rank 126 risk factors of the female dataset.

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring9_PreparePredictionAttributes-
weka.filters.unsupervised.attribute.Remove-R129-weka.filters.unsupervised.attribute.Remove-
R1
Instances:    2481
Attributes:   127
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data




=== Attribute Selection on all input data ===


Search Method:
                                                  Attribute ranking.


Attribute Evaluator (supervised, Class (nominal): 127 cvd10):
                                                  Information Gain Ranking Filter


Ranked attributes:
 0.02852496   123 AGE
 0.01869903    56 AGE AT WHICH PERIODS STOPPED
 0.01760813    55 PERIODS HAVE STOPPED 1 YR OR MORE
 0.01736669    57 CAUSE OF CESSATION OF MENSES
 0.01333186    34 FIRST SECOND VOLUME
 0.01316599    32 SYSTOLIC BLOOD PRESSURE
 0.01151725    87 DYSPNEA INCREASE IN PAST 2 YEARS
 0.010896       1 TOTAL CHOLESTEROL
 0.01079082    86 DYSPNEA ON EXERTION
 0.00976453     3 VLDL CHOLESTEROL
```

0.00969151	40	HISTORY OF HYPERTENSION
0.0091564	35	TOTAL VITAL CAPACITY
0.00818991	5	TRIGLYCERIDES
0.00808946	4	LDL CHOLESTEROL
0.00794315	58	HYSTERECTOMY
0.00711604	29	ALKALINE PHOSPHOTASE
0.0070442	46	HYPOTENSIVES (EXCLUDING DIURETICS)
0.00694056	33	DIASTOLIC BLOOD PRESSURE
0.00586272	45	DIURETICS FOR BLOOD PRESSURE
0.00544788	79	WINE INTAKE
0.00536259	24	URIC ACID
0.0050902	22	GLUCOSE
0.00493847	41	CARDIAC GLYCOSIDES
0.00461508	2	HDL CHOLESTEROL
0.00424987	122	WGTGP
0.00412152	30	LDH
0.00397126	97	M.C.H.
0.00392025	121	HGT
0.00370363	114	TAKING DIGITALIS OR QUINIDINE
0.00349519	10	TOP FRACTION PRE-BETA
0.00349136	69	USUAL # OF CIGARRETTE SMOKE NOW/EVER
0.0032381	18	FREDERICKSON CLASSIFICATION
0.00323324	59	OVARIES REMOVED
0.00248442	126	Diabetes
0.00245533	42	NITRITES
0.00213577	116	LEFT VENTRICULAR HYPERTROPHY
0.00212337	88	RECENT ORTHOPNEA
0.00197348	60	ORAL CONTRACEPTIVE
0.00193088	52	TRANQUILIZERS
0.00188409	43	QUINIDINE
0.00188145	47	ANTI-CHOLESTEROL AGENTS
0.0018661	16	PRE-BETA BAND
0.00181642	7	WHOLE PLASMA PRE-BETA
0.00168755	77	DIABETIC DIET LAST 2 WEEKS
0.00126579	125	Treatment for Diabetes
0.00126579	51	HYPOGLYCEMIC AGENTS (SPECIFY)
0.00111503	119	ECG CLINICAL READING
0.0010974	73	WEIGHT COMPARE WITH 1 YEAR AGO
0.00106009	13	INFRANATE AFTER 12 HRS AT 4 DEGREES

0.00100238	90	BILATERAL ANKLE EDEMA
0.00099236	113	OTHER ECG ABNORMALITY
0.00092967	15	FASTING 12 HRS OR MORE
0.00086887	66	SMOKES CIGARETTES
0.00086159	63	SMOKED AT LEAST 1 YEAR
0.00086123	36	ECG FINDING SUMMARY
0.00085272	117	NON-SPECIFIC T-WAVE ABNORMALITY
0.00083746	91	NOCTURNAL COUGH OR WHEEZING
0.00076227	110	WOLFF-PARKINSON-WHITE SYNDROME
0.00075842	50	ANTI-COAGULANTS
0.00074358	71	INHALES
0.00071828	83	CHEST DISCOMFORT
0.00066923	89	PAROXYSMAL NOCTURAL DYSPNEA
0.00063807	81	AMOUNT OF FOOD LAST 2 DAYS
0.00062922	85	HISTORY OF ENLARGED HEART
0.00062846	61	PREMARIN
0.00057662	53	BROCHODILATOR OR AEROSOL
0.00056414	62	OTHER (SPECIFY)
0.00055495	12	WHOLE PLASMA APPEARANCE
0.00053141	49	ANTI-THYROID
0.000512	111	PREMATURE BEATS
0.00041051	104	RIGHT-INTRAVENTRICULAR BLOCK
0.00033901	72	WEIGHT COMPARED WITH 1 MONTH AGO
0.00025063	112	OTHER ARRHYTHMIA
0.00021424	75	LOW CHOLESTEROL DIET LAST 2 WEEKS
0.00020537	82	AMOUNT OF ALCOHOL
0.00011656	76	LOW CALORIE DIET LAST 2 WEEKS
0.00011425	48	THYROID
0.00011348	84	HISTORY OF HEART MURMUR
0.00010678	39	HISTORY OF HYPOTHYROID DISEASE
0.00009143	14	CREAM AFTER 12 HRS OR MORE
0.00007469	64	SMOKES CIGARS
0.0000681	44	DIURETICS FOR FLUID RETENTION
0.00005187	54	OTHER (C-V DRUGS)
0.00004595	9	TOP FRACTION BETA
0.00003733	65	SMOKES PIPES
0.00003717	115	MYOCARDIAL INFARCTION
0.00003717	107	BIFASCULAR BLOCK
0.00003717	105	LEFT-INTRAVENTRICULAR BLOCK

```
0.00002137   108 INCOMPLETE-ATRIOVENTRICULAR BLOCK
0.00002021    17 SINKING PRE-BETA BAND
0.00002       11 BOTTOM FRACTION PRE-BETA
0.00001737   118 NON-SPECIFIC S-T SEGMENT ABNORMALITY
0.00001407    70 USES FILTER
0.00000388    38 HISTORY OF OTHER KIDNEY AILMENT
0.00000138   106 HEMIBLOCK
0             21 PHOSPHORUS
0            120 ALCOHOL INDEX
0             20 CALCIUM
0              8 TOP FRACTION ORIGIN
0             19 HEMATOCRIT
0             23 BUN
0            124 BMI
0              6 WHOLE PLASMA ORIGIN
0             92 COMPLETE BLOOD COUNT
0             25 TOTAL PROTEIN
0             96 M.C.V.
0             68 IF STOPPED AGE STOPPED
0             78 BEER INTAKE
0             74 WEIGHT AT AGE 25
0             95 H.G.B.
0             26 ALBUMIN
0             93 WHITE BLOOD COUNT
0             94 RED BLOOD COUNT
0             80 COCKTAIL INTAKE
0             67 AGE START SMOKE CIGARETTE REGULARLY
0             98 M.C.H.C.
0             99 VENTRICULAR RATE
0             28 TOTAL BILIRUBIN
0             27 GLOBULIN
0            100 P-R INTERVAL
0             31 SGOT
0            109 COMPLETE ATRIOVENTRICULAR BLOCK
0             37 HISTORY OF NEPHROSIS
0            102 QT INTERAVAL
0            101 QRS INTERVAL
0            103 A QRS
```

```
Selected attributes:
123,56,55,57,34,32,87,1,86,3,40,35,5,4,58,29,46,33,45,79,24,22,41,2,122,30,97,121,114,10,69,
18,59,126,42,116,88,60,52,43,47,16,7,77,125,51,119,73,13,90,113,15,66,63,36,117,91,110,50,7
1,83,89,81,85,61,53,62,12,49,111,104,72,112,75,82,76,48,84,39,14,64,44,54,9,65,115,107,105,
108,17,11,118,70,38,106,21,120,20,8,19,23,124,6,92,25,96,68,78,74,95,26,93,94,80,67,98,99,2
8,27,100,31,109,37,102,101,103 : 126
```

Appendix I FEMALE DATASET—SECOND ATTRIBUTE EVALUATION BY WEKA

Below was the result output from using Weka's InfoGainAttributeEval attribute evaluator to rank 33 selected risk factors, after removing missing data, of the female dataset. However, after the second attribute evaluation, only 29 risk factors were chosen for the female model as four risk factors had the information gain values of 0.

```
=== Run information ===


Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     FramOffSpring13-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R35
Instances:    2101
Attributes:   34
              TOTAL CHOLESTEROL
              HDL CHOLESTEROL
              VLDL CHOLESTEROL
              LDL CHOLESTEROL
              TRIGLYCERIDES
              TOP FRACTION PRE-BETA
              FREDERICKSON CLASSIFICATION
              GLUCOSE
              URIC ACID
              ALKALINE PHOSPHOTASE
              LDH
              SYSTOLIC BLOOD PRESSURE
              DIASTOLIC BLOOD PRESSURE
              FIRST SECOND VOLUME
              TOTAL VITAL CAPACITY
              HISTORY OF HYPERTENSION
              CARDIAC GLYCOSIDES
              DIURETICS FOR BLOOD PRESSURE
              HYPOTENSIVES (EXCLUDING DIURETICS)
              PERIODS HAVE STOPPED 1 YR OR MORE
              AGE AT WHICH PERIODS STOPPED
              CAUSE OF CESSATION OF MENSES
              HYSTERECTOMY
              OVARIES REMOVED
              USUAL # OF CIGARRETTE SMOKE NOW/EVER
              WINE INTAKE
              DYSPNEA ON EXERTION
              DYSPNEA INCREASE IN PAST 2 YEARS
              M.C.H.
              TAKING DIGITALIS OR QUINIDINE
              HGT
              WGTGP
              AGE
              cvd10
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
                                    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 34 cvd10):
                                    Information Gain Ranking Filter

Ranked attributes:
 0.03042   33 AGE
 0.01967   21 AGE AT WHICH PERIODS STOPPED
 0.01871   20 PERIODS HAVE STOPPED 1 YR OR MORE
 0.01871    4 LDL CHOLESTEROL
 0.01819   22 CAUSE OF CESSATION OF MENSES
 0.01396    1 TOTAL CHOLESTEROL
 0.01379   14 FIRST SECOND VOLUME
 0.01373   12 SYSTOLIC BLOOD PRESSURE
 0.0127    28 DYSPNEA INCREASE IN PAST 2 YEARS
 0.01067    3 VLDL CHOLESTEROL
 0.00976   27 DYSPNEA ON EXERTION
 0.00894   10 ALKALINE PHOSPHOTASE
 0.00877    5 TRIGLYCERIDES
 0.0082    23 HYSTERECTOMY
 0.00807   15 TOTAL VITAL CAPACITY
 0.00766   16 HISTORY OF HYPERTENSION
 0.0076    13 DIASTOLIC BLOOD PRESSURE
 0.00701   17 CARDIAC GLYCOSIDES
 0.0065    19 HYPOTENSIVES (EXCLUDING DIURETICS)
 0.00634   26 WINE INTAKE
 0.00616    8 GLUCOSE
 0.00579   11 LDH
 0.00565   30 TAKING DIGITALIS OR QUINIDINE
 0.00541    2 HDL CHOLESTEROL
 0.00514    9 URIC ACID
 0.00421    6 TOP FRACTION PRE-BETA
 0.00401   18 DIURETICS FOR BLOOD PRESSURE
 0.00395    7 FREDERICKSON CLASSIFICATION
 0.00376   24 OVARIES REMOVED
 0         25 USUAL # OF CIGARRETTE SMOKE NOW/EVER
 0         29 M.C.H.
 0         31 HGT
 0         32 WGTGP


Selected attributes:
33,21,20,4,22,1,14,12,28,3,27,10,5,23,15,16,13,17,19,26,8,11,30,2,9,6,18,7,24,25,29,31,32 : 33
```

Appendix J MIXED SEX DATASET—PREDICTORS FILE

Table J-2 displays the predictors file prepared in section 3.8.11 to describe the predictors (risk factors) of the dataset, which was used for experimenting the mixed sex model.

Table J-2: Predictors file to describe the predictors of the dataset for the mixed sex model

Predictor Name	Predictor Description	Data Type	Value List
sex	SEX	DataOneOf	Male|Female
totalChol	TOTAL CHOLESTEROL	double	N/A
hdlChol	HDL CHOLESTEROL	double	N/A
vldlChol	VLDL CHOLESTEROL	double	N/A
ldlChol	LDL CHOLESTEROL	double	N/A
triglycerides	TRIGLYCERIDES	double	N/A
wholePlasma	WHOLE PLASMA PRE-BETA	DataOneOf	Yes|No
topFraction	TOP FRACTION PRE-BETA	DataOneOf	Yes|No
preBetaBand	PRE-BETA BAND	DataOneOf	Yes|No
frederickson	FREDERICKSON CLASSIFICATION	DataOneOf	Normal|Abnormal
hematocrit	HEMATOCRIT	double	N/A
glucose	GLUCOSE	double	N/A
uricAcid	URIC ACID	double	N/A
alkalinePhos	ALKALINE PHOSPHOTASE	double	N/A
ldh	LDH	double	N/A
sysBP	SYSTOLIC BLOOD PRESSURE	double	N/A
diaBP	DIASTOLIC BLOOD PRESSURE	double	N/A
hypertension	HISTORY OF HYPERTENSION	DataOneOf	Yes|No
hypotensives	HYPOTENSIVES (EXCLUDING DIURETICS)	DataOneOf	Yes|No
hypoglycemic	HYPOGLYCEMIC AGENTS	DataOneOf	Yes|No
smoked1Year	SMOKED AT LEAST 1 YEAR	DataOneOf	Yes|No
smoking	SMOKES CIGARETTES	DataOneOf	Yes|No

cigarettes	USUAL # OF CIGARETTE SMOKE NOW/EVER	double	N/A
dyspnea	DYSPNEA ON EXERTION	DataOneOf	Yes|No
dyspneaIncrease2Yrs	DYSPNEA INCREASE IN PAST 2 YEARS	DataOneOf	Yes|No
whiteBloodCount	WHITE BLOOD COUNT	double	N/A
redBloodCount	RED BLOOD COUNT	double	N/A
hgb	H.G.B.	double	N/A
aqrs	A QRS	double	N/A
wgtgp	WGTGP	double	N/A
age	AGE	double	N/A
bmi	BMI	double	N/A
treatmentForDiabetes	Treatment for Diabetes	DataOneOf	Yes|No
diabetes	Diabetes	DataOneOf	Yes|No

Appendix K MIXED SEX DATASET—PREDICTORS RANKING FILE

Table K-3 displays the predictors ranking file prepared in section 3.8.11 to rank the predictors (risk factors) of the dataset used for experimenting the mixed sex model.

Table K-3: Predictors ranking file to rank the predictors of the dataset for the mixed sex model

No	Info Gain	Predictor
1	0.04384	age
2	0.02324	totalChol
3	0.0184	ldlChol
4	0.01733	vldlChol
5	0.01562	sysBP
6	0.01392	triglycerides
7	0.01286	diaBP
8	0.0119	glucose
9	0.01112	cigarettes
10	0.01109	hdlChol
11	0.00915	hematocrit
12	0.00902	bmi
13	0.00843	ldh
14	0.00805	sex
15	0.00741	wgtgp
16	0.00736	uricAcid
17	0.00735	frederickson
18	0.00725	hgb
19	0.00719	alkalinePhos
20	0.00678	whiteBloodCount
21	0.00673	dyspnea
22	0.00672	diabetes

23	0.0067	topFraction
24	0.00652	redBloodCount
25	0.00616	smoked1Year
26	0.00578	treatmentForDiabetes
27	0.00578	hypoglycemic
28	0.00569	aqrs
29	0.00568	hypertension
30	0.00543	preBetaBand
31	0.00535	wholePlasma
32	0.00506	dyspneaIncrease2Yrs
33	0.00503	smoking
34	0.00445	hypotensives

Appendix L MALE DATASET—PREDICTORS FILE

Table L-4 displays the predictors file prepared to describe the predictors (risk factors) of the male dataset, which was used for experimenting the male model.

Table L-4: Predictors file to describe the predictors of the dataset for the male model

Predictor Name	Predictor Description	Data Type	Value List
totalChol	TOTAL CHOLESTEROL	double	N/A
hdlChol	HDL CHOLESTEROL	double	N/A
vldlChol	VLDL CHOLESTEROL	double	N/A
ldlChol	LDL CHOLESTEROL	double	N/A
triglycerides	TRIGLYCERIDES	double	N/A
glucose	GLUCOSE	double	N/A
albumin	ALBUMIN	double	N/A
totalBilirubin	TOTAL BILIRUBIN	double	N/A
ldh	LDH	double	N/A
sysBP	SYSTOLIC BLOOD PRESSURE	double	N/A
diaBP	DIASTOLIC BLOOD PRESSURE	double	N/A
firstSecondVolume	FIRST SECOND VOLUME	double	N/A
totalVitalCapacity	TOTAL VITAL CAPACITY	double	N/A
hypoglycemic	HYPOGLYCEMIC AGENTS	DataOneOf	Yes|No
smoked1Year	SMOKED AT LEAST 1 YEAR	DataOneOf	Yes|No
smoking	SMOKES CIGARETTES	DataOneOf	Yes|No
cigarettes	USUAL # OF CIGARETTE SMOKE NOW/EVER	double	N/A
dyspnea	DYSPNEA ON EXERTION	DataOneOf	Yes|No
whiteBloodCount	WHITE BLOOD COUNT	double	N/A
aqrs	A QRS	double	N/A
age	AGE	double	N/A
treatmentForDiabetes	Treatment for Diabetes	DataOneOf	Yes|No
diabetes	Diabetes	DataOneOf	Yes|No

Appendix M MALE DATASET—PREDICTORS RANKING FILE

Table M-5 displays the predictors ranking file prepared to rank the predictors (risk factors) of the dataset used for experimenting the male model.

Table M-5: Predictors ranking file to rank the predictors of the dataset for the male model

No	Ranked	Predictor
1	0.05941	age
2	0.02679	totalChol
3	0.02532	firstSecondVolume
4	0.02421	totalVitalCapacity
5	0.0224	ldlChol
6	0.02112	albumin
7	0.01598	whiteBloodCount
8	0.01556	glucose
9	0.01411	triglycerides
10	0.01013	totalBilirubin
11	0.00976	ldh
12	0.00961	cigarettes
13	0.00941	smoked1Year
14	0.00937	diaBP
15	0.00921	vldlChol
16	0.00881	smoking
17	0.00874	dyspnea
18	0.00862	hypoglycemic
19	0.00862	treatmentForDiabetes
20	0.00806	hdlChol
21	0.00788	sysBP
22	0.00779	diabetes
23	0.00685	aqrs

Appendix N FEMALE DATASET—PREDICTORS FILE

Table N-6 displays the predictors file prepared to describe the predictors (risk factors) of the dataset, which was used for experimenting the female model.

Table N-6: Predictors file to describe the predictors of the dataset for the female model

Predictor Name	Predictor Description	Data Type	Value List
totalChol	TOTAL CHOLESTEROL	double	N/A
hdlChol	HDL CHOLESTEROL	double	N/A
vldlChol	VLDL CHOLESTEROL	double	N/A
ldlChol	LDL CHOLESTEROL	double	N/A
triglycerides	TRIGLYCERIDES	double	N/A
topFraction	TOP FRACTION PRE-BETA	DataOneOf	Yes|No
frederickson	FREDERICKSON CLASSIFICATION	DataOneOf	Normal|Abnormal
glucose	GLUCOSE	double	N/A
uricAcid	URIC ACID	double	N/A
alkalinePhos	ALKALINE PHOSPHOTASE	double	N/A
ldh	LDH	double	N/A
sysBP	SYSTOLIC BLOOD PRESSURE	double	N/A
diaBP	DIASTOLIC BLOOD PRESSURE	double	N/A
firstSecondVolume	FIRST SECOND VOLUME	double	N/A
totalVitalCapacity	TOTAL VITAL CAPACITY	double	N/A
hypertension	HISTORY OF HYPERTENSION	DataOneOf	Yes|No
cardiacGlycosides	CARDIAC GLYCOSIDES	DataOneOf	Yes|No
diureticsForBP	DIURETICS FOR BLOOD PRESSURE	DataOneOf	Yes|No

hypotensives	HYPOTENSIVES (EXCLUDING DIURETICS)	DataOneOf	Yes|No
periodsStopped1yrOrMore	PERIODS HAVE STOPPED 1 YR OR MORE	DataOneOf	Yes|No
agePStopped	AGE AT WHICH PERIODS STOPPED	double	N/A
causeOfCessationM	CAUSE OF CESSATION OF MENSES	DataOneOf	Normal|Abnormal|NotStopped
hysterectomy	HYSTERECTOMY	DataOneOf	Yes|No
ovariesRemoved	OVARIES REMOVED	DataOneOf	Yes|No
wineIntake	WINE INTAKE	double	N/A
dyspnea	DYSPNEA ON EXERTION	DataOneOf	Yes|No
dyspneaIncrease2Yrs	DYSPNEA INCREASE IN PAST 2 YEARS	DataOneOf	Yes|No
dOq	TAKING DIGITALIS OR QUINIDINE	DataOneOf	Yes|No
age	AGE	double	N/A

Appendix O FEMALE DATASET—PREDICTORS RANKING FILE

Table O-7 displays the predictors ranking file prepared to rank the predictors (risk factors) of the dataset used for experimenting the female model.

Table O-7: Predictors ranking file to rank the predictors of the dataset for the female model

No	Ranked	Predictor
1	0.03042	age
2	0.01967	agePStopped
3	0.01871	periodsStopped1yrOrMore
4	0.01871	ldlChol
5	0.01819	causeOfCessationM
6	0.01396	totalChol
7	0.01379	firstSecondVolume
8	0.01373	sysBP
9	0.0127	dyspneaIncrease2Yrs
10	0.01067	vldlChol
11	0.00976	dyspnea
12	0.00894	alkalinePhos
13	0.00877	triglycerides
14	0.0082	hysterectomy
15	0.00807	totalVitalCapacity
16	0.00766	hypertension
17	0.0076	diaBP
18	0.00701	cardiacGlycosides
19	0.0065	hypotensives
20	0.00634	wineIntake
21	0.00616	glucose
22	0.00579	ldh
23	0.00565	dOq

24	0.00541	hdlChol
25	0.00514	uricAcid
26	0.00421	topFraction
27	0.00401	diureticsForBP
28	0.00395	frederickson
29	0.00376	ovariesRemoved

Appendix P CRISK FUZZY ONTOLOGY TEMPLATE FILE

Below is the CRISK fuzzy ontology template file (base.owl):

```
<?xml version="1.0"?>

<!DOCTYPE Ontology [
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
   <!ENTITY xml "http://www.w3.org/XML/1998/namespace" >
   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
]>

<Ontology xmlns="http://www.w3.org/2002/07/owl#"
     xml:base="http://www.aut.ac.nz/ontologies/fcvdo.owl"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     ontologyIRI="http://www.aut.ac.nz/ontologies/fcvdo.owl">
   <Prefix name="" IRI="http://www.aut.ac.nz/ontologies/cvdo.owl"/>
   <Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#"/>
   <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
   <Prefix name="xml" IRI="http://www.w3.org/XML/1998/namespace"/>
   <Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSchema#"/>
   <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
   <Annotation>
      <AnnotationProperty IRI="#fuzzyLabel"/>
      <Literal datatypeIRI="&rdf;PlainLiteral">&lt;fuzzyOwl2
fuzzyType=&quot;ontology&quot;&gt;
&lt;FuzzyLogic logic=&quot;zadeh&quot; /&gt;
&lt;/fuzzyOwl2&gt;</Literal>
   </Annotation>
   <Declaration>
      <Class IRI="#CBR_CASE"/>
   </Declaration>
   <Declaration>
      <Class IRI="#CRISK_CASE"/>
   </Declaration>
   <Declaration>
      <DataProperty IRI="#cvd10"/>
   </Declaration>
```

```
<Declaration>
    <DataProperty IRI="#cvdInterval"/>
</Declaration>
<Declaration>
    <AnnotationProperty IRI="#fuzzyLabel"/>
</Declaration>
<Declaration>
    <Datatype IRI="#highCVDRisk"/>
</Declaration>
<Declaration>
    <Datatype IRI="#lowCVDRisk"/>
</Declaration>
<SubClassOf>
    <Class IRI="#CRISK_CASE"/>
    <Class IRI="#CBR_CASE"/>
</SubClassOf>
<SubDataPropertyOf>
    <DataProperty IRI="#cvd10"/>
    <DataProperty abbreviatedIRI="owl:topDataProperty"/>
</SubDataPropertyOf>
<SubDataPropertyOf>
    <DataProperty IRI="#cvdInterval"/>
    <DataProperty abbreviatedIRI="owl:topDataProperty"/>
</SubDataPropertyOf>
<DataPropertyDomain>
    <DataProperty IRI="#cvd10"/>
    <Class IRI="#CRISK_CASE"/>
</DataPropertyDomain>
<DataPropertyDomain>
    <DataProperty IRI="#cvdInterval"/>
    <Class IRI="#CRISK_CASE"/>
</DataPropertyDomain>
<DataPropertyRange>
    <DataProperty IRI="#cvd10"/>
    <DataOneOf>
        <Literal datatypeIRI="&rdf;PlainLiteral">No</Literal>
        <Literal datatypeIRI="&rdf;PlainLiteral">Yes</Literal>
    </DataOneOf>
</DataPropertyRange>
<DataPropertyRange>
    <DataProperty IRI="#cvdInterval"/>
```

```
        <Datatype abbreviatedIRI="xsd:double"/>
    </DataPropertyRange>
    <AnnotationAssertion>
        <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
        <IRI>#cvd10</IRI>
        <Literal datatypeIRI="&xsd;string">10-year CVD (Yes/No)</Literal>
    </AnnotationAssertion>
    <AnnotationAssertion>
        <AnnotationProperty IRI="#fuzzyLabel"/>
        <IRI>#highCVDRisk</IRI>
        <Literal datatypeIRI="&rdf;PlainLiteral">&lt;fuzzyOwl2
fuzzyType=&quot;datatype&quot;&gt;
&lt;Datatype type=&quot;leftshoulder&quot; a=&quot;5.0&quot; b=&quot;15.0&quot; /&gt;
&lt;/fuzzyOwl2&gt;</Literal>
    </AnnotationAssertion>
    <AnnotationAssertion>
        <AnnotationProperty IRI="#fuzzyLabel"/>
        <IRI>#lowCVDRisk</IRI>
        <Literal datatypeIRI="&rdf;PlainLiteral">&lt;fuzzyOwl2
fuzzyType=&quot;datatype&quot;&gt;
&lt;Datatype type=&quot;rightshoulder&quot; a=&quot;5.0&quot; b=&quot;15.0&quot; /&gt;
&lt;/fuzzyOwl2&gt;</Literal>
    </AnnotationAssertion>
    <AnnotationAssertion>
        <AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
        <IRI>#cvdInterval</IRI>
        <Literal datatypeIRI="&xsd;string">CVD Interval (Years)</Literal>
    </AnnotationAssertion>
    <DatatypeDefinition>
        <Datatype IRI="#highCVDRisk"/>
        <DataIntersectionOf>
            <DatatypeRestriction>
                <Datatype abbreviatedIRI="xsd:double"/>
                <FacetRestriction facet="&xsd;minInclusive">
                    <Literal datatypeIRI="&xsd;double">5.0</Literal>
                </FacetRestriction>
            </DatatypeRestriction>
            <DatatypeRestriction>
                <Datatype abbreviatedIRI="xsd:double"/>
                <FacetRestriction facet="&xsd;maxInclusive">
                    <Literal datatypeIRI="&xsd;double">15.0</Literal>
                </FacetRestriction>
```

```
         </DatatypeRestriction>
      </DataIntersectionOf>
   </DatatypeDefinition>
   <DatatypeDefinition>
      <Datatype IRI="#lowCVDRisk"/>
      <DataIntersectionOf>
         <DatatypeRestriction>
            <Datatype abbreviatedIRI="xsd:double"/>
            <FacetRestriction facet="&xsd;minInclusive">
               <Literal datatypeIRI="&xsd;double">5.0</Literal>
            </FacetRestriction>
         </DatatypeRestriction>
         <DatatypeRestriction>
            <Datatype abbreviatedIRI="xsd:double"/>
            <FacetRestriction facet="&xsd;maxInclusive">
               <Literal datatypeIRI="&xsd;double">15.0</Literal>
            </FacetRestriction>
         </DatatypeRestriction>
      </DataIntersectionOf>
   </DatatypeDefinition>
</Ontology>
```

Appendix Q MIXED SEX MODEL EXPERIMENTATION RESULTS

Table Q-8 displays prediction TPR results for experimenting the mixed sex dataset. The top ten TPR values are in dark-red bold text. The highest TPR values have yellow highlighted background. The TPR results are plotted as a 3D graph in Figure Q-1.

Table Q-8: Mixed sex model—TPR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.4751	0.4525	0.4434	0.4344	0.4344	0.4072	0.3846	0.3756	0.3710
n=2	0.4842	0.4389	0.4570	0.4661	0.4615	0.4751	0.4796	0.4887	0.4887
n=3	0.6561	0.6290	0.6244	0.6425	0.6335	0.6471	0.6471	0.6516	0.6425
n=4	0.6878	0.6561	0.6787	0.6968	0.6878	0.6923	0.7104	0.7104	0.7195
n=5	0.7647	0.7466	0.7602	0.7602	0.7738	0.7692	0.7738	0.7647	0.7692
n=6	0.7511	0.7647	0.7602	0.7647	0.7692	0.7647	0.7692	0.7511	0.7692
n=7	0.7466	0.7195	0.7104	0.7466	0.7376	0.7511	0.7376	0.7330	0.7511
n=8	0.7828	0.7557	0.7783	0.7738	0.7919	0.7828	0.7738	0.7828	0.7828
n=9	0.7738	0.7783	0.8009	0.8100	0.7919	0.7783	0.7602	0.7828	0.7873
n=10	0.7919	0.7511	0.7964	0.7738	0.8054	0.7919	0.7964	0.7828	0.7783
n=11	0.8371	0.8145	0.8371	0.8190	0.8145	0.8100	0.8190	0.8100	0.7964
n=12	**0.8778**	**0.8778**	0.8462	0.8597	0.8507	0.8507	0.8462	0.8416	0.8281
n=13	**0.8733**	**0.8824**	**0.8733**	**0.8733**	**0.8688**	**0.8778**	**0.8824**	**0.8824**	0.8643
n=14	0.7195	0.7104	0.7059	0.7149	0.6833	0.6968	0.6968	0.7014	0.6968
n=15	0.7149	0.6878	0.6833	0.6878	0.6968	0.6878	0.6833	0.6787	0.6742
n=16	0.7104	0.7059	0.7014	0.7104	0.7014	0.7014	0.7014	0.6968	0.6923
n=17	0.6606	0.6742	0.6742	0.6697	0.6652	0.6561	0.6561	0.6516	0.6561
n=18	0.6742	0.6652	0.6606	0.6697	0.6606	0.6561	0.6561	0.6471	0.6290
n=19	0.7014	0.6833	0.6606	0.6652	0.6697	0.6561	0.6471	0.6561	0.6516
n=20	0.7014	0.6833	0.6923	0.6742	0.6833	0.6833	0.6923	0.6742	0.6742
n=21	0.6606	0.6380	0.6290	0.6199	0.6199	0.6244	0.6335	0.6290	0.6199
n=22	0.6425	0.5973	0.6018	0.5973	0.5928	0.5928	0.6063	0.6018	0.5973
n=23	0.5566	0.5339	0.5249	0.5294	0.5566	0.5520	0.5475	0.5475	0.5385
n=24	0.5566	0.5430	0.5204	0.5249	0.5475	0.5520	0.5385	0.5385	0.5385
n=25	0.4932	0.4842	0.4842	0.4932	0.5023	0.4977	0.4842	0.4842	0.4842
n=26	0.4932	0.4842	0.4887	0.4887	0.4932	0.4842	0.4751	0.4706	0.4706
n=27	0.4932	0.4842	0.4887	0.4887	0.4977	0.4842	0.4751	0.4661	0.4661
n=28	0.4887	0.4706	0.4796	0.4977	0.4842	0.4661	0.4615	0.4525	0.4570
n=29	0.4299	0.4163	0.4163	0.4163	0.4118	0.4027	0.4072	0.4072	0.3982

n=30	0.4253	0.4118	0.4072	0.4072	0.4072	0.4027	0.4072	0.4072	0.4027
n=31	0.4253	0.4072	0.4072	0.4072	0.4072	0.4027	0.4072	0.4072	0.4027
n=32	0.4344	0.4118	0.3982	0.4027	0.4027	0.3982	0.4027	0.4118	0.4163
n=33	0.4253	0.4027	0.3891	0.3846	0.3801	0.3756	0.3801	0.3891	0.3982
n=34	0.4299	0.3891	0.3891	0.3846	0.3756	0.3756	0.3710	0.3891	0.3937

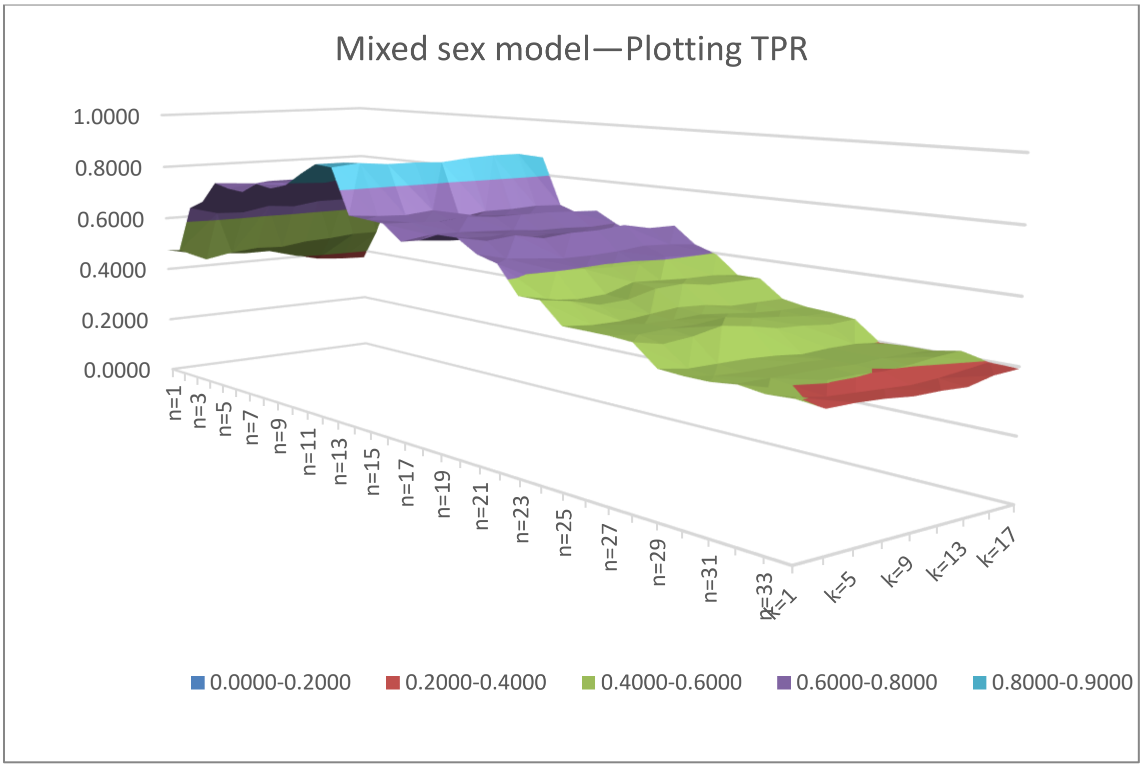


Figure Q-1: Mixed sex model—Plotting TPR

Table Q-9 displays prediction TNR results for experimenting the mixed sex dataset. The top ten TNR values are in dark-red bold text. The highest TNR value has yellow highlighted background. The TNR results are plotted as a 3D graph in Figure Q-2.

Table Q-9: Mixed sex model—TNR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.7600	0.7577	0.7631	0.7990	0.8016	0.8223	0.8452	0.8512	0.8532
n=2	0.8262	0.8460	0.8483	0.8447	0.8439	0.8426	0.8423	0.8416	0.8400
n=3	0.7587	0.7894	0.7870	0.7881	0.7932	0.7945	0.7951	0.7940	0.7935
n=4	0.7434	0.7849	0.7919	0.7966	0.7943	0.7945	0.7979	0.7956	0.7938
n=5	0.7460	0.7868	0.7896	0.7823	0.7865	0.7849	0.7865	0.7857	0.7852
n=6	0.7564	0.7847	0.7925	0.7842	0.7878	0.7896	0.7888	0.7847	0.7862
n=7	0.7517	0.7935	0.8003	0.8042	0.8060	0.8049	0.8034	0.8036	0.8000
n=8	0.7675	0.7982	0.8091	0.8104	0.8049	0.8057	0.8031	0.8031	0.7990

n=9	0.7974	0.8288	0.8338	0.8325	0.8330	0.8345	0.8353	0.8348	0.8317
n=10	0.8073	0.8421	0.8405	0.8421	0.8410	0.8400	0.8421	0.8379	0.8400
n=11	0.7971	0.8265	0.8325	0.8361	0.8384	0.8358	0.8353	0.8351	0.8330
n=12	0.7995	0.8208	0.8255	0.8257	0.8286	0.8294	0.8255	0.8247	0.8234
n=13	0.7966	0.8169	0.8229	0.8270	0.8249	0.8234	0.8216	0.8184	0.8179
n=14	0.8200	0.8418	0.8509	0.8478	0.8499	0.8486	0.8460	0.8434	0.8439
n=15	0.8291	0.8496	0.8553	0.8553	0.8522	0.8517	0.8506	0.8527	0.8514
n=16	0.8304	0.8481	0.8527	0.8535	0.8517	0.8514	0.8519	0.8512	0.8538
n=17	0.8358	0.8577	0.8590	0.8582	0.8613	0.8623	0.8571	0.8597	0.8608
n=18	0.8400	0.8616	0.8701	0.8696	0.8655	0.8603	0.8621	0.8603	0.8610
n=19	0.8265	0.8499	0.8571	0.8561	0.8540	0.8545	0.8509	0.8491	0.8483
n=20	0.8351	0.8574	0.8587	0.8616	0.8597	0.8579	0.8582	0.8543	0.8545
n=21	0.8358	0.8577	0.8587	0.8610	0.8564	0.8543	0.8543	0.8538	0.8545
n=22	0.8397	0.8597	0.8608	0.8636	0.8584	0.8561	0.8571	0.8556	0.8561
n=23	0.8462	0.8717	0.8748	0.8735	0.8719	0.8727	0.8691	0.8660	0.8691
n=24	0.8483	0.8704	0.8719	0.8701	0.8714	0.8712	0.8686	0.8678	0.8681
n=25	0.8558	0.8784	0.8810	0.8803	0.8787	0.8790	0.8771	0.8784	0.8774
n=26	0.8558	0.8792	0.8805	0.8800	0.8792	0.8790	0.8777	0.8784	0.8769
n=27	0.8561	0.8795	0.8808	0.8800	0.8790	0.8784	0.8769	0.8777	0.8769
n=28	0.8605	0.8826	0.8852	0.8857	0.8805	0.8823	0.8808	0.8792	0.8782
n=29	0.8683	0.8943	0.8979	0.8979	0.8945	0.8948	0.8940	0.8945	0.8922
n=30	0.8686	0.8935	0.8969	0.8969	0.8935	0.8927	0.8927	0.8930	0.8899
n=31	0.8683	0.8935	0.8969	0.8966	0.8938	0.8927	0.8925	0.8927	0.8894
n=32	0.8706	0.8943	0.8984	0.8984	0.8948	0.8930	0.8927	0.8930	0.8894
n=33	0.8766	**0.9013**	**0.9065**	**0.9042**	**0.9013**	**0.9016**	0.8995	0.8997	0.8974
n=34	0.8766	**0.9018**	**0.9062**	**0.9034**	**0.9008**	**0.9010**	0.8995	0.8992	0.8971

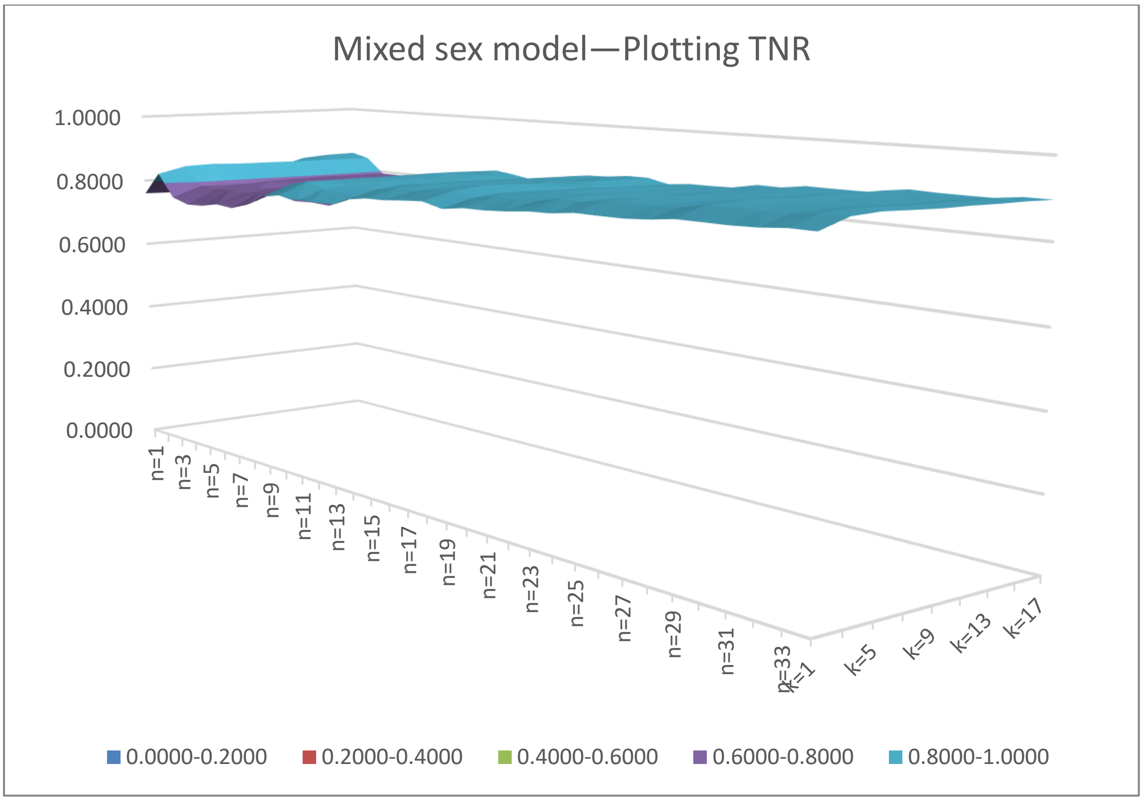


Figure Q-2: Mixed sex model—Plotting TNR

Table Q-10 displays prediction Precision results for experimenting the mixed sex dataset. The top ten Precision values are in dark-red bold text. The highest Precision value has yellow highlighted background. The Precision results are plotted as a 3D graph in Figure Q-3.

Table Q-10: Mixed sex model—Precision

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.1020	0.0968	0.0970	0.1103	0.1116	0.1163	0.1248	0.1265	0.1267
n=2	0.1379	0.1406	0.1474	0.1469	0.1451	0.1477	0.1487	0.1504	0.1492
n=3	0.1350	0.1463	0.1441	0.1482	0.1496	0.1531	0.1534	0.1537	0.1515
n=4	0.1333	0.1490	0.1577	0.1644	0.1610	0.1621	0.1679	0.1663	0.1668
n=5	0.1473	0.1673	0.1718	0.1670	0.1722	0.1703	0.1722	0.1700	0.1705
n=6	0.1504	0.1693	0.1737	0.1690	0.1722	0.1726	0.1729	0.1668	0.1712
n=7	0.1472	0.1667	0.1695	0.1795	0.1791	0.1810	0.1772	0.1765	0.1774
n=8	0.1620	0.1769	0.1896	0.1898	0.1890	0.1878	0.1841	0.1858	0.1827
n=9	0.1798	0.2070	0.2166	0.2172	0.2139	0.2126	0.2095	0.2138	0.2117
n=10	0.1908	0.2145	**0.2228**	0.2195	**0.2253**	0.2212	**0.2245**	0.2171	0.2183
n=11	0.1915	0.2123	**0.2229**	**0.2229**	**0.2244**	0.2207	0.2221	0.2199	0.2149
n=12	0.2008	0.2195	0.2177	0.2207	0.2217	**0.2225**	0.2177	0.2160	0.2121
n=13	0.1977	0.2167	0.2206	**0.2247**	0.2217	0.2220	0.2211	0.2181	0.2141

n=14	0.1866	0.2050	0.2137	0.2124	0.2071	0.2090	0.2062	0.2045	0.2040
n=15	0.1936	0.2079	0.2133	0.2144	0.2130	0.2102	0.2080	0.2092	0.2067
n=16	0.1938	0.2105	0.2147	0.2178	0.2135	0.2132	0.2138	0.2118	0.2137
n=17	0.1877	0.2138	0.2153	0.2133	0.2159	0.2148	0.2086	0.2105	0.2129
n=18	0.1948	0.2162	**0.2260**	**0.2277**	0.2199	0.2123	0.2145	0.2100	0.2062
n=19	0.1883	0.2071	0.2098	0.2097	0.2085	0.2057	0.1994	0.1997	0.1978
n=20	0.1962	0.2157	0.2195	0.2185	0.2185	0.2163	0.2189	0.2099	0.2102
n=21	0.1877	0.2046	0.2035	0.2039	0.1986	0.1974	0.1997	0.1980	0.1966
n=22	0.1871	0.1964	0.1988	0.2009	0.1938	0.1912	0.1959	0.1930	0.1924
n=23	0.1720	0.1928	0.1940	0.1937	0.1997	0.1993	0.1936	0.1900	0.1910
n=24	0.1740	0.1939	0.1891	0.1883	0.1964	0.1974	0.1904	0.1895	0.1898
n=25	0.1642	0.1861	0.1894	0.1912	0.1920	0.1910	0.1845	0.1861	0.1848
n=26	0.1642	0.1871	0.1901	0.1895	0.1899	0.1867	0.1823	0.1818	0.1799
n=27	0.1644	0.1874	0.1905	0.1895	0.1910	0.1861	0.1813	0.1794	0.1785
n=28	0.1674	0.1871	0.1934	0.2000	0.1887	0.1853	0.1818	0.1770	0.1772
n=29	0.1578	0.1844	0.1897	0.1897	0.1831	0.1802	0.1807	0.1815	0.1750
n=30	0.1567	0.1816	0.1848	0.1848	0.1800	0.1773	0.1789	0.1793	0.1735
n=31	0.1564	0.1800	0.1848	0.1844	0.1804	0.1773	0.1786	0.1789	0.1728
n=32	0.1616	0.1827	0.1837	0.1854	0.1802	0.1760	0.1773	0.1809	0.1776
n=33	0.1652	0.1898	0.1928	0.1872	0.1810	0.1797	0.1783	0.1822	0.1822
n=34	0.1667	0.1853	0.1924	0.1860	0.1785	0.1789	0.1748	0.1814	0.1801

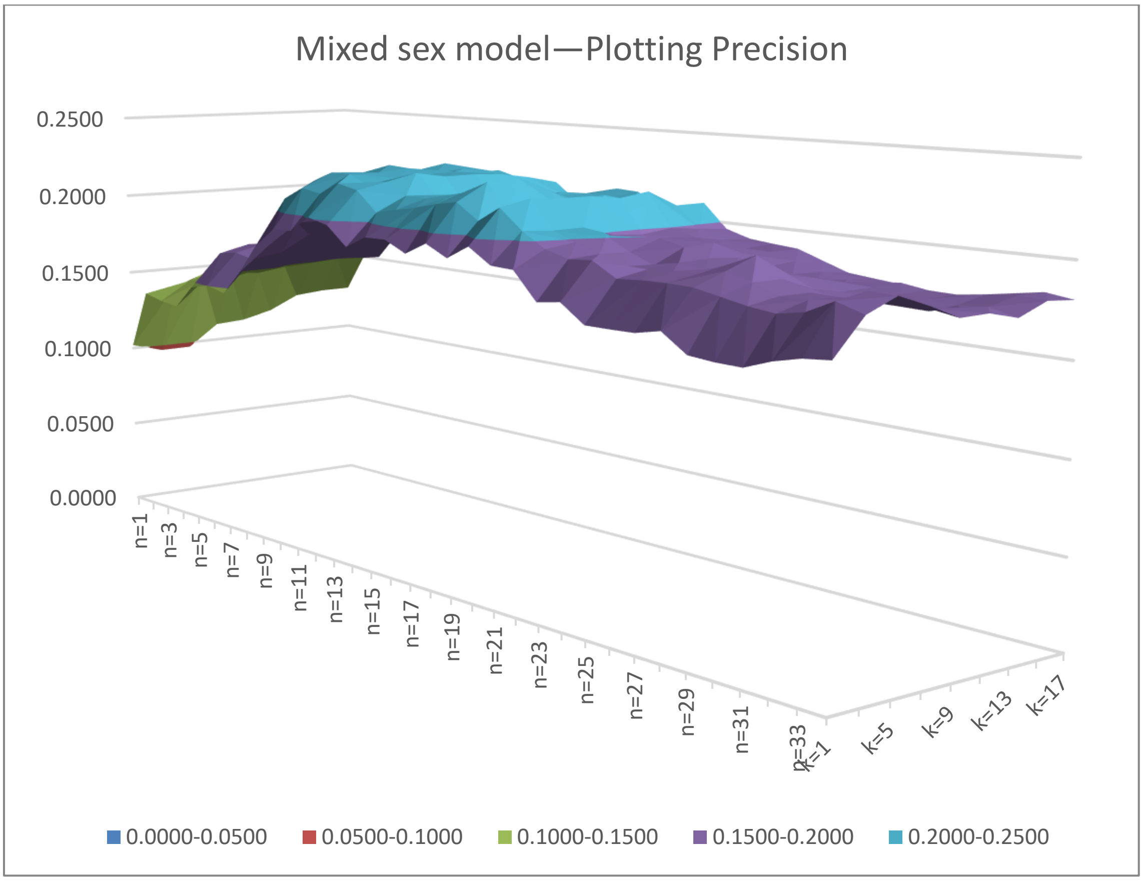


Figure Q-3: Mixed sex model—Plotting Precision

Table Q-11 displays prediction F_1-value results for experimenting the mixed sex dataset. The top ten F_1-values are in dark-red bold text. The highest F_1-value has yellow highlighted background. The F_1-value results are plotted as a 3D graph in Figure Q-4.

Table Q-11: Mixed sex model—F_1-value

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.1680	0.1595	0.1592	0.1760	0.1776	0.1809	0.1885	0.1893	0.1889
n=2	0.2146	0.2130	0.2230	0.2234	0.2208	0.2253	0.2270	0.2300	0.2286
n=3	0.2239	0.2374	0.2341	0.2409	0.2420	0.2476	0.2480	0.2487	0.2453
n=4	0.2234	0.2429	0.2560	0.2660	0.2609	0.2627	0.2716	0.2695	0.2709
n=5	0.2471	0.2734	0.2802	0.2738	0.2817	0.2789	0.2817	0.2782	0.2791
n=6	0.2506	0.2773	0.2828	0.2768	0.2815	0.2817	0.2824	0.2730	0.2801
n=7	0.2459	0.2706	0.2738	0.2895	0.2882	0.2917	0.2857	0.2845	0.2869
n=8	0.2684	0.2867	0.3050	0.3048	0.3051	0.3030	0.2974	0.3003	0.2962
n=9	0.2918	0.3270	0.3410	0.3426	0.3369	0.3340	0.3284	0.3359	0.3337
n=10	0.3076	0.3337	0.3482	0.3420	**0.3521**	0.3458	0.3502	0.3399	0.3409
n=11	0.3117	0.3368	**0.3520**	0.3504	**0.3519**	0.3469	0.3494	0.3459	0.3385
n=12	0.3269	0.3511	0.3463	0.3512	**0.3517**	**0.3527**	0.3463	0.3438	0.3376
n=13	0.3225	0.3479	**0.3522**	**0.3574**	**0.3533**	**0.3543**	**0.3536**	0.3498	0.3432

n=14	0.2964	0.3181	0.3281	0.3275	0.3179	0.3215	0.3182	0.3166	0.3156
n=15	0.3047	0.3193	0.3251	0.3269	0.3263	0.3220	0.3189	0.3198	0.3163
n=16	0.3046	0.3243	0.3287	0.3333	0.3273	0.3270	0.3277	0.3249	0.3266
n=17	0.2923	0.3246	0.3264	0.3235	0.3259	0.3237	0.3166	0.3182	0.3215
n=18	0.3022	0.3263	0.3368	0.3398	0.3299	0.3208	0.3233	0.3171	0.3106
n=19	0.2969	0.3179	0.3184	0.3189	0.3179	0.3132	0.3049	0.3062	0.3035
n=20	0.3066	0.3279	0.3333	0.3300	0.3311	0.3286	0.3326	0.3201	0.3204
n=21	0.2923	0.3099	0.3075	0.3068	0.3008	0.3000	0.3037	0.3012	0.2985
n=22	0.2898	0.2956	0.2989	0.3007	0.2921	0.2892	0.2961	0.2923	0.2911
n=23	0.2628	0.2833	0.2833	0.2836	0.2939	0.2929	0.2861	0.2821	0.2820
n=24	0.2651	0.2857	0.2774	0.2772	0.2891	0.2908	0.2813	0.2803	0.2807
n=25	0.2463	0.2688	0.2723	0.2756	0.2778	0.2760	0.2672	0.2688	0.2675
n=26	0.2463	0.2699	0.2738	0.2731	0.2742	0.2695	0.2635	0.2623	0.2603
n=27	0.2466	0.2702	0.2741	0.2731	0.2760	0.2688	0.2625	0.2591	0.2581
n=28	0.2494	0.2677	0.2757	0.2853	0.2716	0.2651	0.2609	0.2545	0.2554
n=29	0.2309	0.2556	0.2606	0.2606	0.2535	0.2490	0.2503	0.2510	0.2431
n=30	0.2290	0.2521	0.2542	0.2542	0.2497	0.2462	0.2486	0.2490	0.2425
n=31	0.2287	0.2497	0.2542	0.2539	0.2500	0.2462	0.2483	0.2486	0.2418
n=32	0.2356	0.2531	0.2514	0.2539	0.2490	0.2441	0.2462	0.2514	0.2490
n=33	0.2380	0.2580	0.2579	0.2519	0.2453	0.2430	0.2428	0.2482	0.2500
n=34	0.2402	0.2511	0.2575	0.2507	0.2420	0.2423	0.2377	0.2475	0.2472

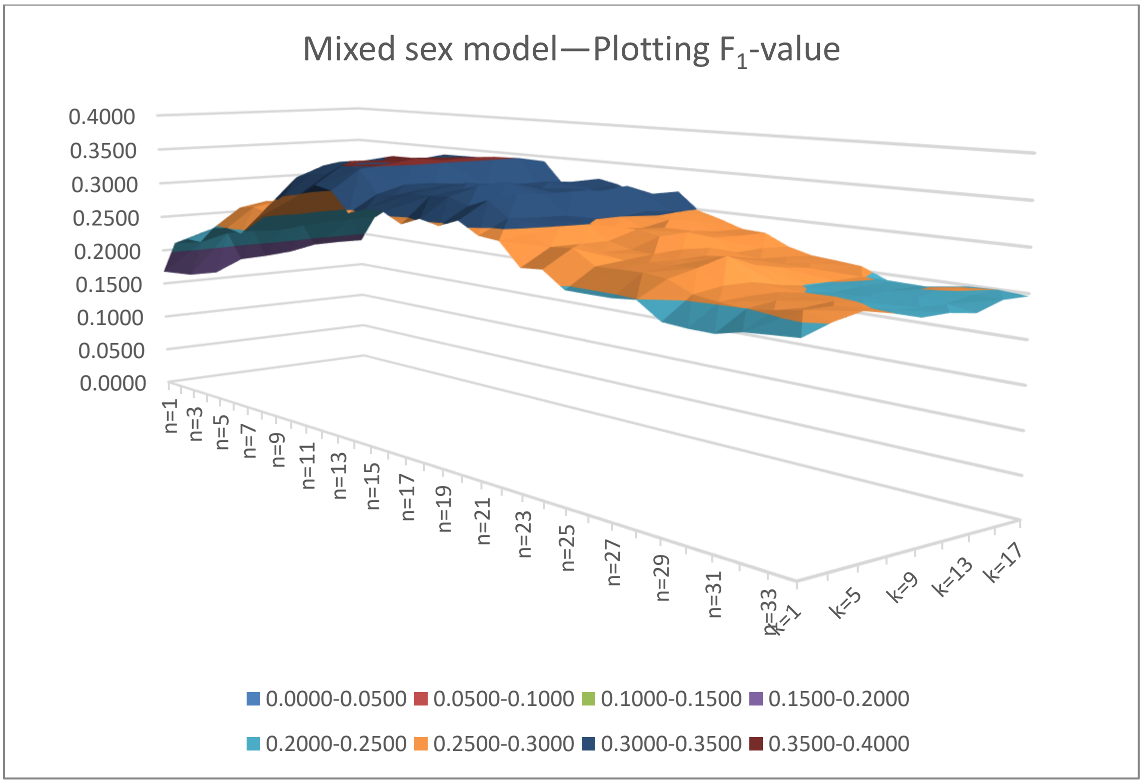


Figure Q-4: Mixed sex model—Plotting F_1-value

Table Q-12 displays prediction NPV results for experimenting the mixed sex dataset. The top ten NPV values are in dark-red bold text. The highest NPV value has yellow highlighted background. The NPV results are plotted as a 3D graph in Figure Q-5.

Table Q-12: Mixed sex model—NPV

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.9619	0.9602	0.9598	0.9609	0.9611	0.9603	0.9599	0.9596	0.9594
n=2	0.9654	0.9633	0.9646	0.9650	0.9647	0.9655	0.9658	0.9663	0.9662
n=3	0.9746	0.9737	0.9733	0.9746	0.9742	0.9751	0.9752	0.9754	0.9748
n=4	0.9765	0.9755	0.9772	0.9786	0.9779	0.9783	0.9796	0.9795	0.9801
n=5	0.9822	0.9818	0.9829	0.9827	0.9838	0.9834	0.9838	0.9831	0.9834
n=6	0.9815	0.9831	0.9829	0.9831	0.9835	0.9832	0.9835	0.9821	0.9834
n=7	0.9810	0.9801	0.9797	0.9822	0.9817	0.9826	0.9816	0.9813	0.9825
n=8	0.9840	0.9827	0.9845	0.9842	0.9854	0.9848	0.9841	0.9847	0.9846
n=9	0.9840	0.9849	0.9865	0.9871	0.9859	0.9850	0.9838	0.9853	0.9855
n=10	0.9854	0.9833	0.9863	0.9848	0.9869	0.9860	0.9863	0.9853	0.9851
n=11	0.9884	0.9873	0.9889	0.9877	0.9875	0.9871	0.9877	0.9871	0.9862
n=12	**0.9913**	**0.9915**	0.9894	0.9903	0.9898	0.9898	0.9894	0.9891	0.9882
n=13	**0.9910**	**0.9918**	**0.9912**	**0.9913**	**0.9910**	**0.9916**	**0.9918**	**0.9918**	0.9906
n=14	0.9807	0.9806	0.9805	0.9811	0.9791	0.9799	0.9798	0.9801	0.9798
n=15	0.9806	0.9793	0.9792	0.9795	0.9800	0.9794	0.9791	0.9788	0.9785

n=16	0.9804	0.9805	0.9803	0.9809	0.9803	0.9803	0.9803	0.9800	0.9797
n=17	0.9772	0.9787	0.9787	0.9784	0.9782	0.9776	0.9775	0.9773	0.9776
n=18	0.9782	0.9782	0.9781	0.9787	0.9780	0.9776	0.9776	0.9770	0.9759
n=19	0.9797	0.9791	0.9778	0.9780	0.9783	0.9774	0.9767	0.9773	0.9770
n=20	0.9799	0.9792	0.9798	0.9788	0.9793	0.9792	0.9798	0.9786	0.9786
n=21	0.9772	0.9763	0.9758	0.9753	0.9752	0.9754	0.9760	0.9757	0.9751
n=22	0.9761	0.9738	0.9741	0.9739	0.9735	0.9734	0.9743	0.9740	0.9737
n=23	0.9708	0.9702	0.9698	0.9700	0.9716	0.9714	0.9710	0.9709	0.9704
n=24	0.9709	0.9707	0.9694	0.9696	0.9711	0.9713	0.9704	0.9704	0.9704
n=25	0.9671	0.9674	0.9675	0.9680	0.9685	0.9682	0.9673	0.9674	0.9674
n=26	0.9671	0.9674	0.9677	0.9677	0.9680	0.9674	0.9668	0.9666	0.9665
n=27	0.9671	0.9674	0.9678	0.9677	0.9682	0.9674	0.9668	0.9663	0.9662
n=28	0.9670	0.9667	0.9674	0.9685	0.9675	0.9664	0.9661	0.9655	0.9657
n=29	0.9637	0.9639	0.9640	0.9640	0.9636	0.9631	0.9633	0.9634	0.9627
n=30	0.9634	0.9636	0.9634	0.9634	0.9633	0.9630	0.9633	0.9633	0.9629
n=31	0.9634	0.9633	0.9634	0.9634	0.9633	0.9630	0.9633	0.9633	0.9629
n=32	0.9640	0.9636	0.9630	0.9632	0.9631	0.9628	0.9630	0.9636	0.9637
n=33	0.9637	0.9634	0.9628	0.9624	0.9620	0.9618	0.9619	0.9625	0.9629
n=34	0.9640	0.9626	0.9627	0.9624	0.9617	0.9617	0.9614	0.9625	0.9627

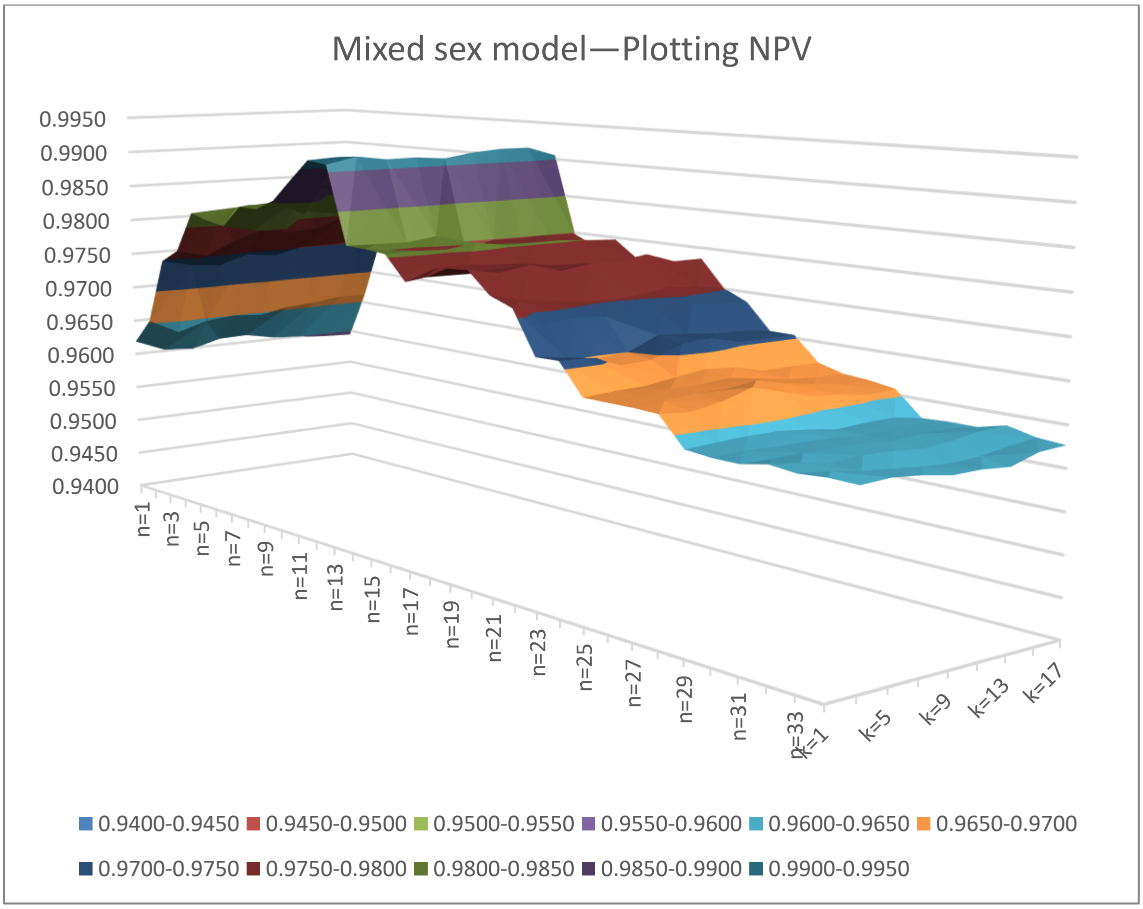


Figure Q-5: Mixed sex model—Plotting NPV

Appendix R MALE MODEL EXPERIMENTATION RESULTS

Table R-13 displays prediction TPR results for experimenting the male dataset. The top ten TPR values are in dark-red bold text. The highest TPR values have yellow highlighted background. The TPR results are plotted as a 3D graph in Figure R-6.

Table R-13: Male model—TPR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.3806	0.4194	0.3935	0.3871	0.3806	0.3226	0.2968	0.2774	0.2452
n=2	0.5355	0.4516	0.4645	0.4710	0.4645	0.4710	0.4774	0.4710	0.4774
n=3	0.7226	0.6839	0.6903	0.6968	0.6968	0.6968	0.6774	0.6968	0.6968
n=4	0.7226	0.7355	0.6903	0.7032	0.7226	0.7290	0.7161	0.7226	0.7226
n=5	0.7290	0.7226	0.6645	0.6839	0.6839	0.7032	0.6903	0.6839	0.6839
n=6	0.7871	0.8000	0.7871	0.8000	0.7742	0.7548	0.7548	0.7613	0.7484
n=7	0.8129	0.8000	0.7484	0.7742	0.7806	0.7742	0.7742	0.7806	0.7742
n=8	0.8387	0.8000	0.7806	0.7742	0.7871	0.7742	0.7677	0.7677	0.7677
n=9	**0.8516**	0.8065	0.8065	0.8323	0.8000	0.7935	0.8065	0.8065	0.7871
n=10	0.8452	0.8194	0.8387	0.8129	**0.8516**	**0.8516**	0.8194	0.8194	0.8258
n=11	**0.8710**	**0.8645**	0.8387	0.8258	0.8452	0.8258	0.8323	0.8194	0.8323
n=12	**0.8903**	**0.8774**	**0.8710**	**0.8903**	**0.8645**	**0.8645**	**0.8581**	**0.8516**	**0.8516**
n=13	0.8194	0.8065	0.8000	0.8065	0.7806	0.7871	0.7742	0.7677	0.7677
n=14	0.8452	0.8129	0.7871	0.8065	0.7935	0.7742	0.7806	0.7806	0.7806
n=15	0.8323	0.8194	0.8129	0.8000	0.8000	0.8000	0.7935	0.7806	0.7806
n=16	0.7871	0.7677	0.7613	0.7484	0.7484	0.7548	0.7484	0.7419	0.7355
n=17	0.6581	0.6258	0.6323	0.6065	0.6065	0.6065	0.6065	0.6065	0.6000
n=18	0.6194	0.6000	0.6000	0.5806	0.5742	0.5677	0.5677	0.5677	0.5613
n=19	0.6194	0.6000	0.6000	0.5742	0.5677	0.5677	0.5613	0.5613	0.5548
n=20	0.6323	0.6000	0.6194	0.6065	0.6065	0.6000	0.5742	0.5806	0.5677
n=21	0.6645	0.6129	0.6258	0.6000	0.6065	0.6129	0.5871	0.5935	0.5871
n=22	0.6581	0.6129	0.6194	0.6000	0.6129	0.6129	0.5935	0.6000	0.5935
n=23	0.6645	0.6258	0.6387	0.6000	0.6065	0.5935	0.5613	0.5484	0.5677

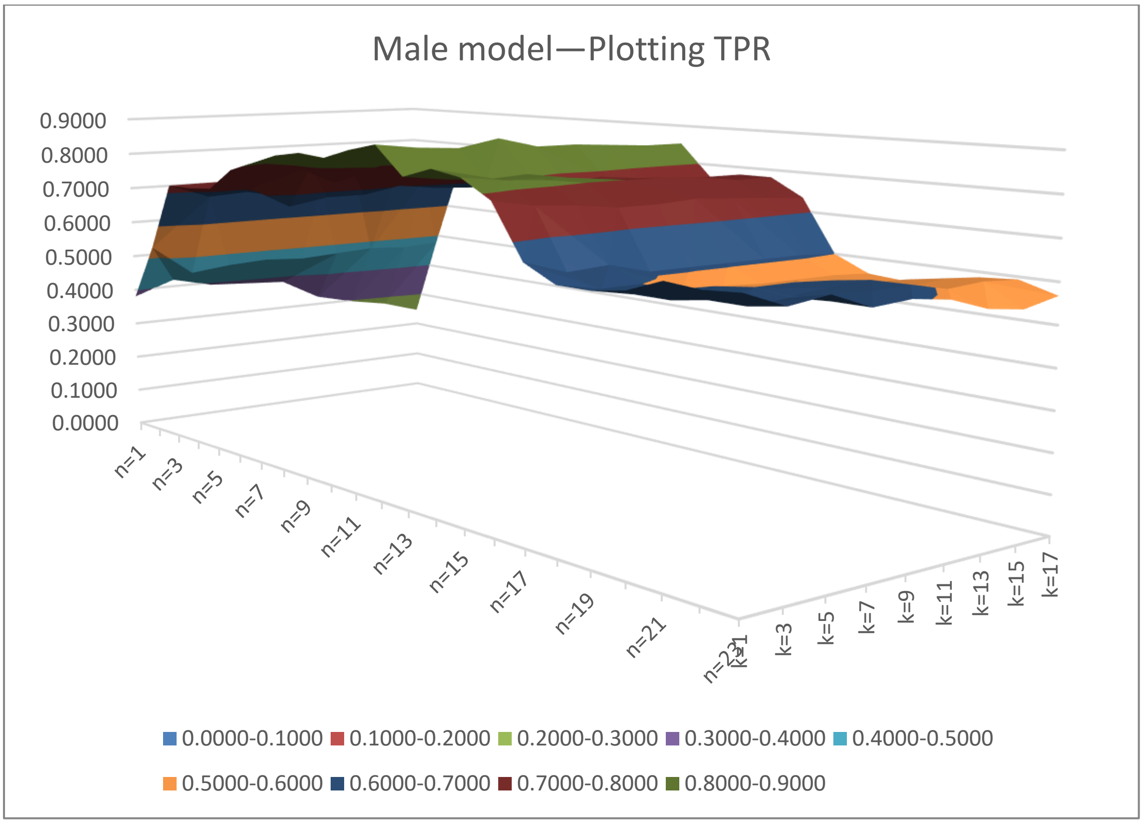


Figure R-6: Male model—Plotting TPR

Table R-14 displays prediction TNR results for experimenting the male dataset. The top ten TNR values are in dark-red bold text. The highest TNR value has yellow highlighted background. The TNR results are plotted as a 3D graph in Figure R-7.

Table R-14: Male model—TNR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	**0.8505**	**0.8384**	**0.8461**	**0.8494**	**0.8516**	**0.8868**	**0.9082**	**0.9109**	**0.9247**
n=2	0.7834	0.8043	0.8131	0.8081	0.8103	0.8076	0.8076	0.8048	0.8032
n=3	0.6778	0.7317	0.7400	0.7427	0.7444	0.7460	0.7455	0.7477	0.7488
n=4	0.7097	0.7620	0.7653	0.7631	0.7548	0.7686	0.7702	0.7647	0.7680
n=5	0.7367	0.7779	0.7861	0.7872	0.7922	0.7922	0.7944	0.7922	0.7949
n=6	0.7356	0.7779	0.7845	0.7872	0.7850	0.7927	0.7878	0.7823	0.7779
n=7	0.7438	0.7861	0.7878	0.7812	0.7823	0.7889	0.7845	0.7823	0.7746
n=8	0.7609	0.7927	0.8021	0.7982	0.7988	0.7966	0.7949	0.7960	0.7933
n=9	0.7543	0.7911	0.7966	0.7938	0.7889	0.7922	0.7911	0.7872	0.7839
n=10	0.7482	0.7834	0.7878	0.7839	0.7872	0.7812	0.7779	0.7790	0.7757
n=11	0.7466	0.7746	0.7784	0.7774	0.7812	0.7806	0.7757	0.7752	0.7719
n=12	0.7361	0.7719	0.7784	0.7784	0.7724	0.7697	0.7631	0.7603	0.7576
n=13	0.7510	0.7883	0.7905	0.7806	0.7774	0.7730	0.7691	0.7653	0.7636
n=14	0.7444	0.7823	0.7823	0.7779	0.7774	0.7691	0.7691	0.7625	0.7587

n=15	0.7510	0.7856	0.7768	0.7806	0.7763	0.7675	0.7642	0.7587	0.7581
n=16	0.7609	0.7971	0.7894	0.7922	0.7894	0.7817	0.7784	0.7752	0.7746
n=17	0.7620	0.7960	0.7966	0.7993	0.7960	0.7916	0.7889	0.7867	0.7878
n=18	0.7636	0.7971	0.7971	0.7988	0.7966	0.7938	0.7911	0.7872	0.7867
n=19	0.7636	0.7971	0.7966	0.7982	0.7966	0.7938	0.7911	0.7872	0.7883
n=20	0.7548	0.7982	0.8021	0.7993	0.8010	0.7971	0.7977	0.7966	0.7993
n=21	0.7631	0.8004	0.7982	0.7988	0.8032	0.8037	0.8026	0.7988	0.7966
n=22	0.7658	0.8026	0.7993	0.8004	0.8037	0.8037	0.8026	0.7960	0.7955
n=23	0.7702	**0.8175**	0.8142	0.8103	0.8169	0.8103	0.8059	0.8054	0.8087

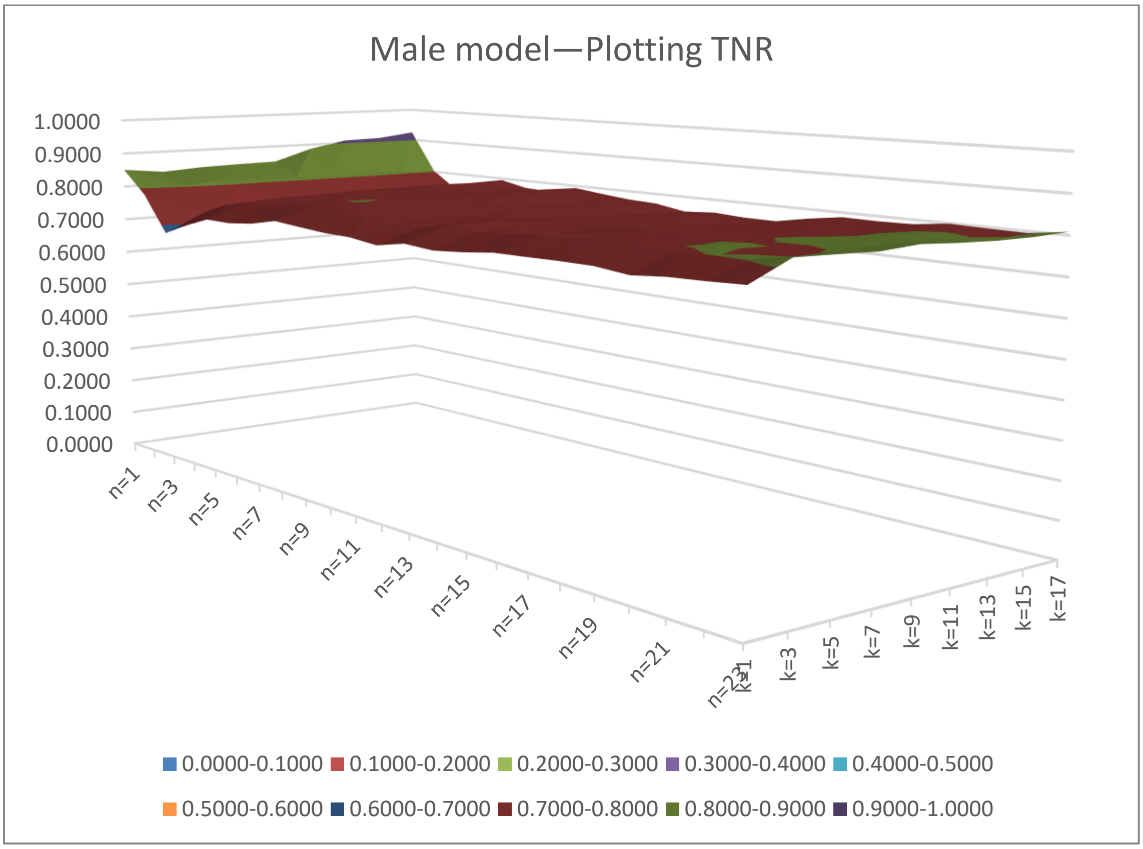


Figure R-7: Male model—Plotting TNR

Table R-15 displays prediction Precision results for experimenting the male dataset. The top ten Precision values are in dark-red bold text. The highest Precision value has yellow highlighted background. The Precision results are plotted as a 3D graph in Figure R-8.

Table R-15: Male model—Precision

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.1782	0.1811	0.1789	0.1796	0.1793	0.1953	0.2160	0.2098	0.2171
n=2	0.1740	0.1643	0.1748	0.1730	0.1727	0.1726	0.1745	0.1706	0.1713
n=3	0.1605	0.1785	0.1845	0.1875	0.1885	0.1895	0.1849	0.1905	0.1912

n=4	0.1750	0.2084	0.2004	0.2019	0.2007	0.2116	0.2098	0.2074	0.2097
n=5	0.1909	0.2171	0.2093	0.2150	0.2190	0.2238	0.2225	0.2190	0.2213
n=6	0.2023	0.2348	0.2374	0.2427	0.2348	0.2368	0.2326	0.2296	0.2231
n=7	0.2128	0.2417	0.2311	0.2317	0.2340	0.2381	0.2344	0.2340	0.2264
n=8	0.2301	0.2475	**0.2516**	0.2464	**0.2500**	0.2449	0.2419	0.2429	0.2404
n=9	0.2280	0.2475	**0.2525**	**0.2560**	0.2441	0.2455	0.2475	0.2441	0.2369
n=10	0.2224	0.2438	**0.2519**	0.2428	**0.2543**	**0.2491**	0.2392	0.2401	0.2388
n=11	0.2265	0.2463	0.2439	0.2402	**0.2476**	0.2429	0.2402	0.2369	0.2371
n=12	0.2233	0.2468	**0.2509**	**0.2551**	0.2445	0.2423	0.2358	0.2324	0.2304
n=13	0.2190	0.2451	0.2455	0.2385	0.2300	0.2280	0.2222	0.2179	0.2168
n=14	0.2198	0.2414	0.2355	0.2363	0.2330	0.2222	0.2237	0.2188	0.2161
n=15	0.2216	0.2456	0.2368	0.2371	0.2335	0.2267	0.2228	0.2161	0.2157
n=16	0.2190	0.2439	0.2355	0.2348	0.2325	0.2276	0.2235	0.2195	0.2176
n=17	0.1907	0.2073	0.2094	0.2048	0.2022	0.1987	0.1967	0.1950	0.1942
n=18	0.1825	0.2013	0.2013	0.1974	0.1939	0.1901	0.1880	0.1853	0.1832
n=19	0.1825	0.2013	0.2009	0.1952	0.1921	0.1901	0.1863	0.1835	0.1826
n=20	0.1801	0.2022	0.2105	0.2048	0.2061	0.2013	0.1947	0.1957	0.1943
n=21	0.1929	0.2074	0.2091	0.2026	0.2080	0.2102	0.2022	0.2009	0.1974
n=22	0.1932	0.2093	0.2082	0.2039	0.2102	0.2102	0.2040	0.2004	0.1983
n=23	0.1977	0.2261	0.2265	0.2123	0.2201	0.2105	0.1977	0.1936	0.2018

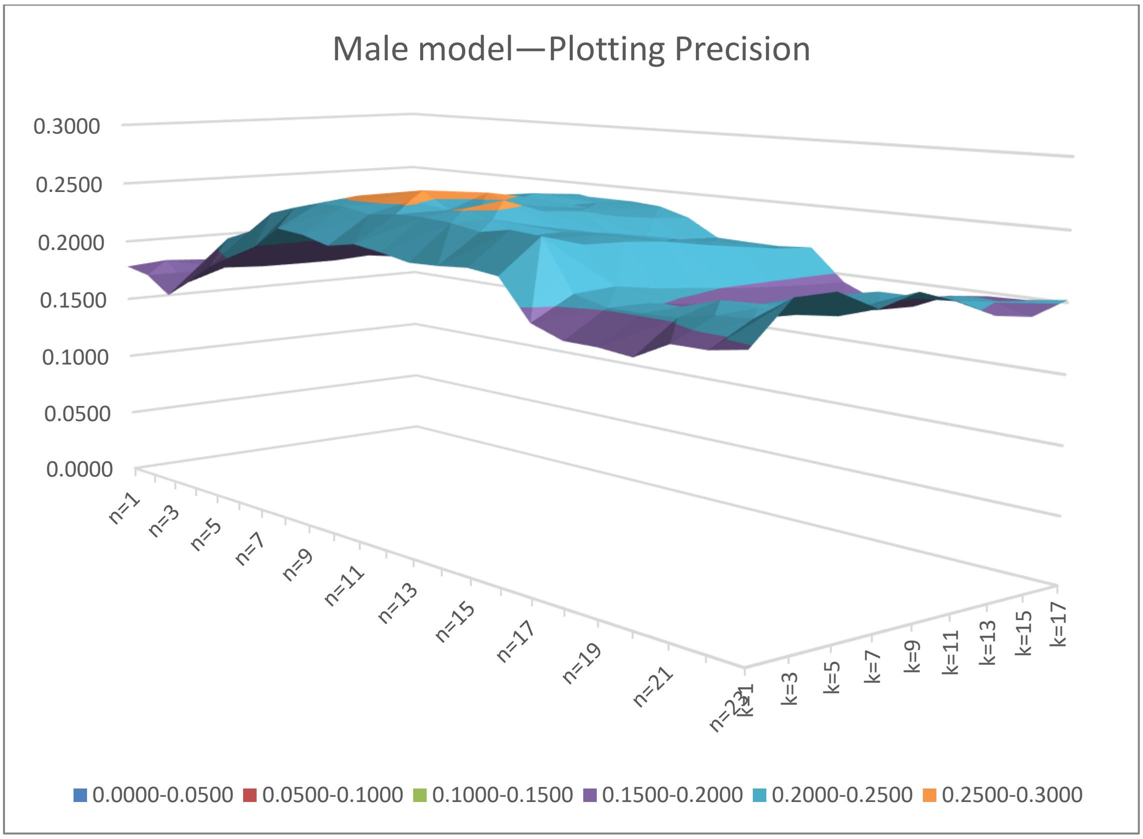


Figure R-8: Male model—Plotting Precision

Table R-16 displays prediction F_1-value results for experimenting the male dataset. The top ten F_1-values are in dark-red bold text. The highest F_1-value has yellow highlighted background. The F_1-value results are plotted as a 3D graph in Figure R-9.

Table R-16: Male model—F_1-value

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.2428	0.2529	0.2460	0.2454	0.2438	0.2433	0.2500	0.2389	0.2303
n=2	0.2627	0.2410	0.2540	0.2530	0.2517	0.2526	0.2556	0.2504	0.2521
n=3	0.2626	0.2830	0.2912	0.2955	0.2967	0.2979	0.2905	0.2992	0.3000
n=4	0.2818	0.3248	0.3106	0.3137	0.3142	0.3280	0.3246	0.3223	0.3251
n=5	0.3025	0.3338	0.3184	0.3272	0.3318	0.3396	0.3365	0.3318	0.3344
n=6	0.3219	0.3631	0.3647	0.3724	0.3604	0.3606	0.3556	0.3528	0.3437
n=7	0.3373	0.3713	0.3531	0.3566	0.3601	0.3642	0.3598	0.3601	0.3504
n=8	0.3611	0.3780	0.3805	0.3738	0.3795	0.3721	0.3679	0.3690	0.3662
n=9	0.3597	0.3788	**0.3846**	**0.3915**	0.3741	0.3750	0.3788	0.3748	0.3642
n=10	0.3522	0.3757	**0.3875**	0.3739	**0.3917**	**0.3854**	0.3703	0.3713	0.3705
n=11	0.3595	**0.3834**	0.3779	0.3721	**0.3830**	0.3754	0.3728	0.3676	0.3691
n=12	0.3571	**0.3853**	**0.3896**	**0.3966**	0.3812	0.3785	0.3700	0.3651	0.3626
n=13	0.3456	0.3759	0.3758	0.3682	0.3554	0.3536	0.3453	0.3395	0.3381
n=14	0.3489	0.3722	0.3626	0.3655	0.3602	0.3453	0.3477	0.3418	0.3385

n=15	0.3501	0.3780	0.3668	0.3658	0.3615	0.3533	0.3479	0.3385	0.3380
n=16	0.3427	0.3701	0.3598	0.3575	0.3547	0.3498	0.3442	0.3387	0.3358
n=17	0.2957	0.3114	0.3146	0.3062	0.3032	0.2994	0.2970	0.2951	0.2934
n=18	0.2819	0.3015	0.3015	0.2946	0.2899	0.2848	0.2825	0.2794	0.2762
n=19	0.2819	0.3015	0.3010	0.2913	0.2871	0.2848	0.2797	0.2766	0.2748
n=20	0.2804	0.3024	0.3142	0.3062	0.3077	0.3015	0.2908	0.2927	0.2895
n=21	0.2990	0.3100	0.3134	0.3029	0.3097	0.3130	0.3008	0.3002	0.2955
n=22	0.2987	0.3120	0.3117	0.3044	0.3130	0.3130	0.3036	0.3005	0.2973
n=23	0.3047	0.3322	0.3345	0.3137	0.3230	0.3108	0.2924	0.2862	0.2978

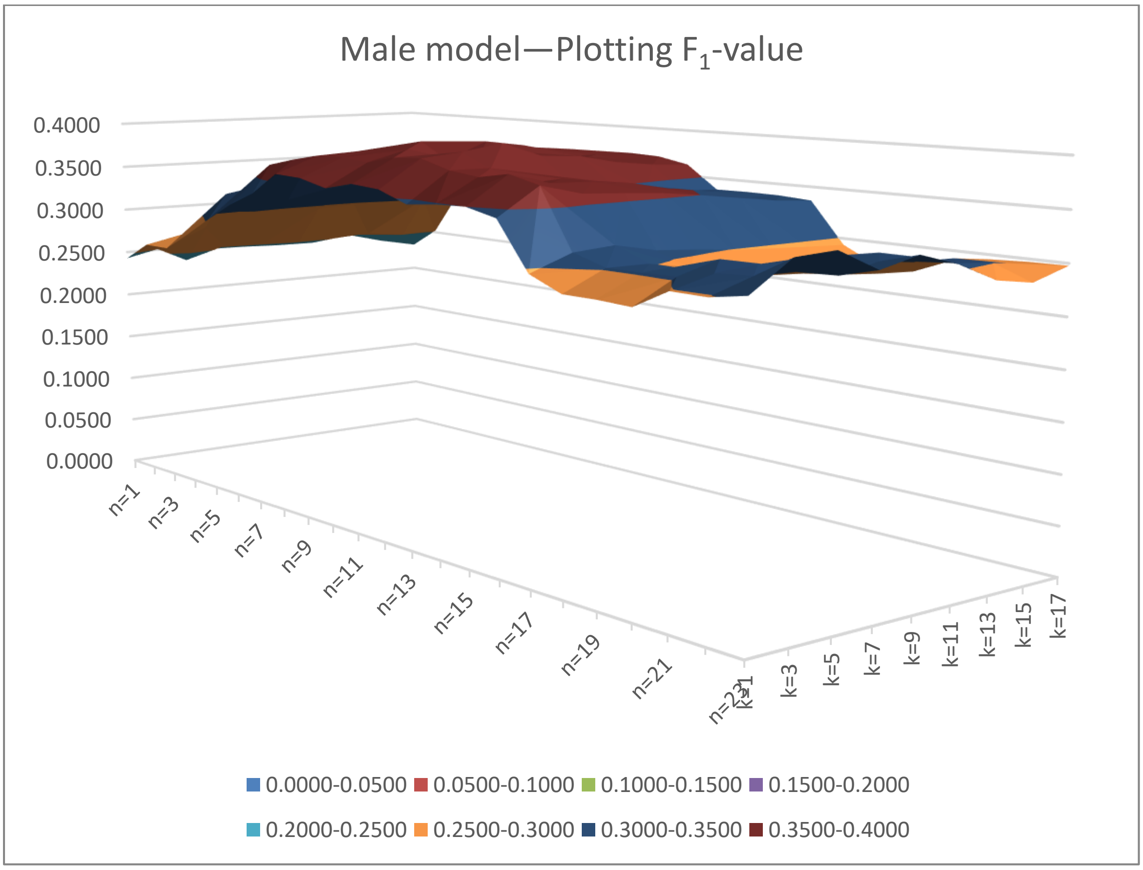


Figure R-9: Male model—Plotting F_1-value

Table R-17 displays prediction NPV results for experimenting the male dataset. The top ten NPV values are in dark-red bold text. The highest NPV value has yellow highlighted background. The NPV results are plotted as a 3D graph in Figure R-10.

Table R-17: Male model—NPV

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.9416	0.9443	0.9424	0.9421	0.9416	0.9389	0.9381	0.9367	0.9350
n=2	0.9519	0.9451	0.9469	0.9472	0.9467	0.9471	0.9477	0.9470	0.9475

n=3	0.9663	0.9645	0.9656	0.9664	0.9665	0.9665	0.9644	0.9666	0.9666
n=4	0.9678	0.9713	0.9667	0.9679	0.9696	0.9708	0.9696	0.9700	0.9701
n=5	0.9696	0.9705	0.9649	0.9669	0.9671	0.9691	0.9678	0.9671	0.9672
n=6	0.9759	0.9786	0.9774	0.9788	0.9761	0.9743	0.9742	0.9747	0.9732
n=7	0.9790	0.9788	0.9735	0.9760	0.9767	0.9762	0.9761	0.9767	0.9758
n=8	0.9823	0.9790	0.9772	0.9765	0.9778	0.9764	0.9757	0.9757	0.9757
n=9	0.9835	0.9796	0.9797	0.9823	0.9789	0.9783	0.9796	0.9795	0.9774
n=10	0.9827	0.9807	0.9829	0.9801	**0.9842**	0.9841	0.9806	0.9806	0.9812
n=11	**0.9855**	**0.9853**	0.9827	0.9813	0.9834	0.9813	0.9819	0.9805	0.9818
n=12	**0.9875**	**0.9866**	**0.9861**	**0.9881**	**0.9853**	**0.9852**	**0.9844**	0.9836	0.9836
n=13	0.9799	0.9795	0.9789	0.9793	0.9765	0.9771	0.9756	0.9748	0.9747
n=14	0.9826	0.9800	0.9773	0.9792	0.9779	0.9756	0.9763	0.9761	0.9760
n=15	0.9813	0.9808	0.9799	0.9786	0.9785	0.9783	0.9775	0.9760	0.9759
n=16	0.9767	0.9758	0.9749	0.9736	0.9736	0.9740	0.9732	0.9724	0.9717
n=17	0.9632	0.9615	0.9622	0.9597	0.9596	0.9594	0.9592	0.9591	0.9585
n=18	0.9593	0.9590	0.9590	0.9572	0.9564	0.9557	0.9555	0.9553	0.9546
n=19	0.9593	0.9590	0.9590	0.9565	0.9558	0.9557	0.9549	0.9547	0.9541
n=20	0.9601	0.9590	0.9611	0.9597	0.9598	0.9590	0.9565	0.9571	0.9560
n=21	0.9639	0.9604	0.9616	0.9591	0.9599	0.9606	0.9580	0.9584	0.9577
n=22	0.9633	0.9605	0.9610	0.9592	0.9606	0.9606	0.9586	0.9589	0.9583
n=23	0.9642	0.9625	0.9636	0.9596	0.9606	0.9590	0.9557	0.9544	0.9564

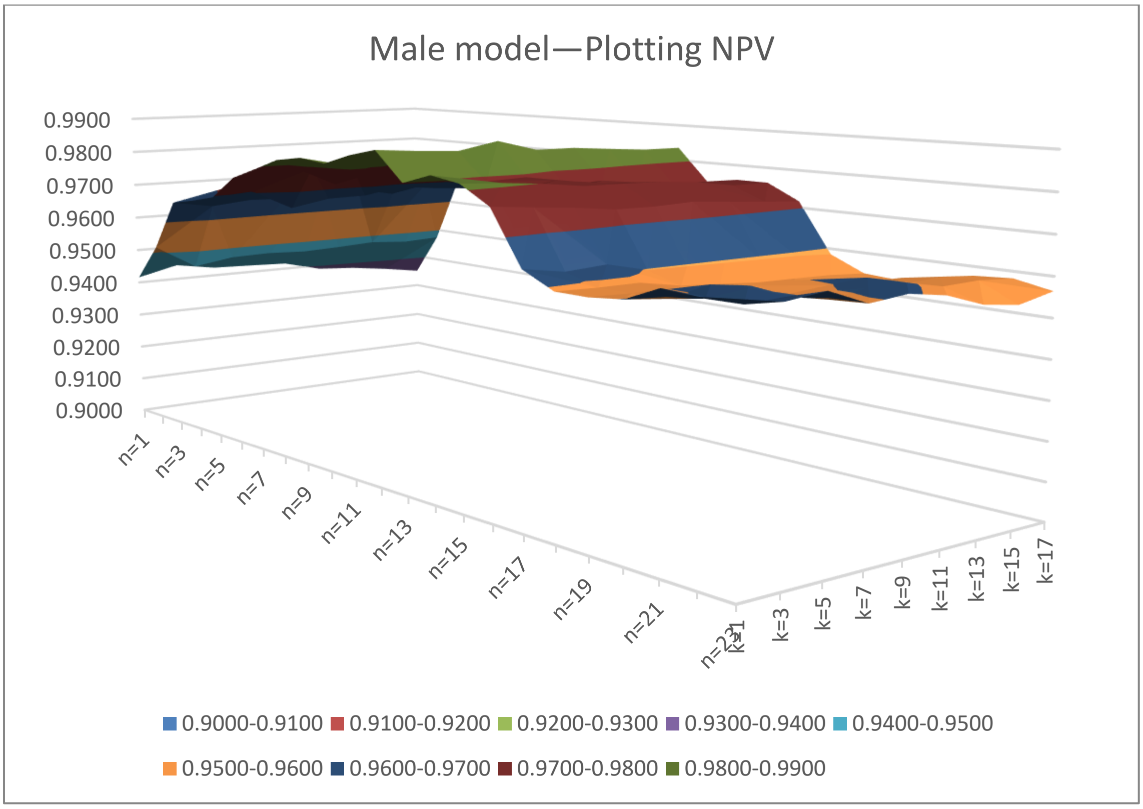


Figure R-10: Male model—Plotting NPV

Appendix S FEMALE MODEL EXPERIMENTATION RESULTS

Table S-18 displays prediction TPR results for experimenting the female dataset. The top ten TPR values are in dark-red bold text. The highest TPR value has yellow highlighted background. The TPR results are plotted as a 3D graph in Figure S-11.

Table S-18: Female model—TPR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.3939	0.4242	0.4242	0.4242	0.3636	0.3333	0.3182	0.3030	0.3030
n=2	0.4091	0.3788	0.3788	0.3788	0.3788	0.3788	0.3788	0.3788	0.3788
n=3	0.3939	0.3636	0.3636	0.3636	0.3636	0.3636	0.3485	0.3485	0.3485
n=4	0.5606	0.5303	0.5303	0.5152	0.5303	0.5303	0.5455	0.5606	0.5758
n=5	0.5303	0.5909	0.6061	0.6061	**0.6212**	**0.6364**	**0.6364**	**0.6364**	**0.6212**
n=6	**0.6364**	0.5909	0.5758	0.5909	0.5909	0.5909	0.5606	0.5758	0.5455
n=7	0.5758	0.5758	0.5758	0.5606	0.6061	0.5758	0.5758	0.5909	0.5909
n=8	**0.6515**	**0.6212**	**0.6212**	**0.6212**	**0.6364**	**0.6212**	0.6061	0.6061	**0.6212**
n=9	0.6061	0.4848	0.4848	0.4848	0.5152	0.5000	0.4848	0.5000	0.5000
n=10	0.6061	0.5152	0.5000	0.5303	0.5152	0.4848	0.5303	0.5000	0.5303
n=11	0.5606	0.5000	0.5000	0.5152	0.5000	0.5000	0.5152	0.5152	0.5000
n=12	0.5303	0.5152	0.5000	0.4545	0.5000	0.4697	0.4697	0.4848	0.5000
n=13	0.5758	0.5152	0.5606	0.5303	0.5152	0.5303	0.5000	0.5152	0.5152
n=14	0.4697	0.4242	0.4545	0.4394	0.4242	0.4242	0.3636	0.3939	0.3788
n=15	0.4545	0.4545	0.4091	0.3939	0.3939	0.3636	0.3939	0.3788	0.3636
n=16	0.3485	0.3939	0.3333	0.3030	0.2576	0.2727	0.3030	0.2879	0.2879
n=17	0.3788	0.3788	0.3333	0.2727	0.2424	0.2879	0.2576	0.2424	0.2576
n=18	0.3788	0.3636	0.3333	0.2576	0.2424	0.2879	0.2576	0.2424	0.2576
n=19	0.3788	0.3485	0.3182	0.2727	0.2424	0.2879	0.2727	0.2576	0.2576
n=20	0.3939	0.3485	0.3182	0.2727	0.2727	0.2879	0.2727	0.2727	0.2576
n=21	0.3939	0.3636	0.3333	0.3030	0.2879	0.2879	0.2879	0.2879	0.2727
n=22	0.4697	0.4091	0.3485	0.3485	0.3182	0.3030	0.3182	0.3030	0.3030
n=23	0.4697	0.4091	0.3485	0.3485	0.3182	0.3182	0.3182	0.3030	0.3030
n=24	0.4697	0.4697	0.4091	0.3788	0.3485	0.3333	0.3333	0.3485	0.3485
n=25	0.4848	0.4545	0.4394	0.3939	0.3636	0.3485	0.3788	0.3939	0.3939
n=26	0.4545	0.3485	0.3939	0.3788	0.3788	0.3485	0.3333	0.3182	0.2879
n=27	0.4242	0.3485	0.3939	0.3788	0.3636	0.3485	0.3182	0.3182	0.2879
n=28	0.3485	0.2727	0.3030	0.3030	0.2879	0.2273	0.2273	0.1970	0.2424
n=29	0.3030	0.2576	0.2727	0.2576	0.2576	0.2273	0.2273	0.2273	0.2273

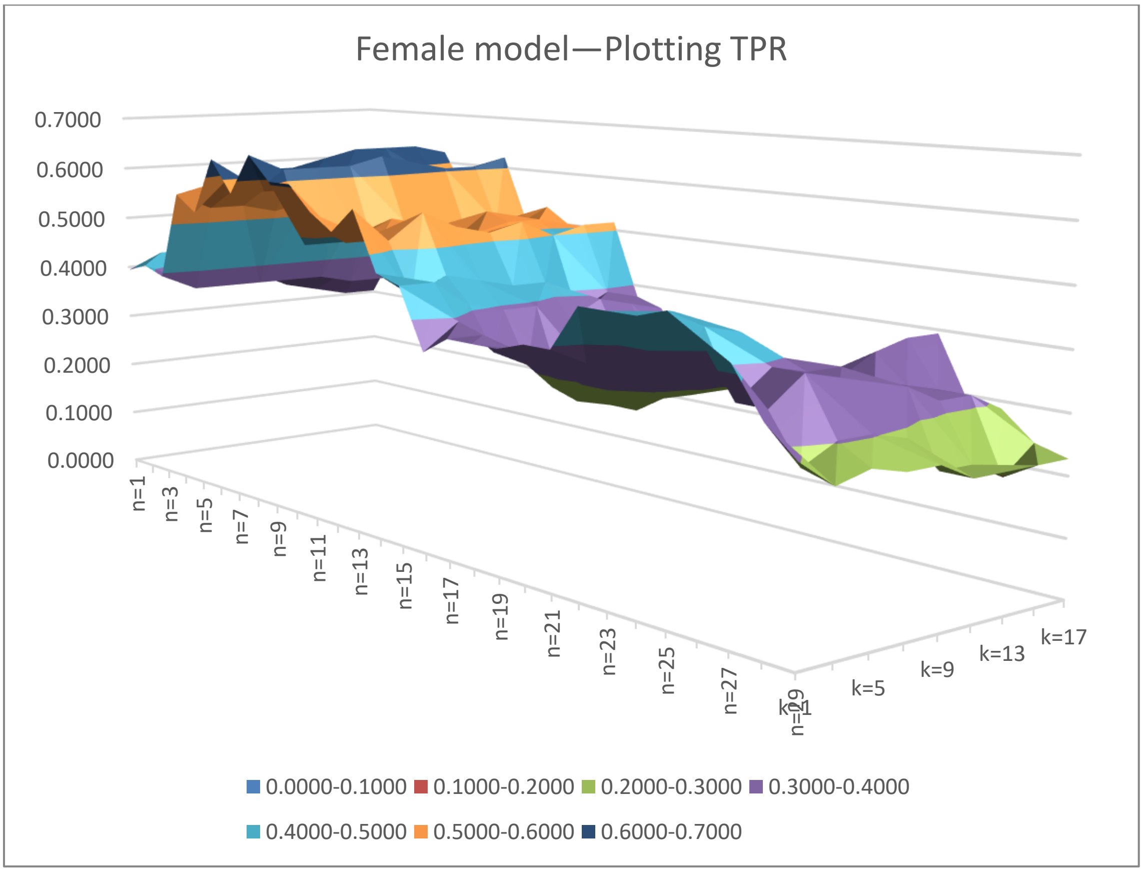


Figure S-11: Female model—Plotting TPR

Table S-19 displays prediction TNR results for experimenting the female dataset. The top ten TNR values are in dark-red bold text. The highest TNR value has yellow highlighted background. The TNR results are plotted as a 3D graph in Figure S-12.

Table S-19: Female model—TNR

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.8732	0.8826	0.8840	0.8840	0.8973	0.9052	0.9091	0.9096	0.9111
n=2	0.8845	0.8963	0.8973	0.8963	0.8958	0.8953	0.8958	0.8963	0.8958
n=3	0.8781	0.8889	0.8914	0.8924	0.8909	0.8929	0.8919	0.8924	0.8924
n=4	0.8467	0.8536	0.8575	0.8550	0.8536	0.8541	0.8526	0.8521	0.8536
n=5	0.8678	0.8722	0.8722	0.8713	0.8688	0.8673	0.8678	0.8698	0.8688
n=6	0.8418	0.8717	0.8781	0.8796	0.8811	0.8806	0.8830	0.8811	0.8840
n=7	0.8314	0.8575	0.8600	0.8644	0.8654	0.8649	0.8698	0.8683	0.8698
n=8	0.8319	0.8614	0.8624	0.8683	0.8678	0.8683	0.8698	0.8708	0.8713
n=9	0.8344	0.8629	0.8639	0.8683	0.8693	0.8693	0.8688	0.8717	0.8698
n=10	0.8369	0.8644	0.8742	0.8727	0.8757	0.8796	0.8767	0.8840	0.8796
n=11	0.8472	0.8703	0.8781	0.8786	0.8860	0.8830	0.8811	0.8835	0.8870
n=12	0.8501	0.8658	0.8771	0.8781	0.8826	0.8885	0.8865	0.8845	0.8855
n=13	0.8452	0.8698	0.8786	0.8855	0.8904	0.8929	0.8939	0.8904	0.8904

n=14	0.8604	0.8899	0.8998	0.9081	0.9106	0.9125	0.9120	0.9101	0.9111
n=15	0.8595	0.8845	0.9017	0.9081	0.9096	0.9061	0.9032	0.9081	0.9091
n=16	0.8644	0.8889	0.9032	0.9086	0.9066	0.9052	0.9012	0.9096	0.9106
n=17	0.8698	0.8870	0.9042	0.9091	0.9091	0.9135	0.9130	0.9174	0.9155
n=18	0.8717	0.8880	0.9052	0.9091	0.9096	0.9145	0.9140	**0.9184**	0.9165
n=19	0.8722	0.8889	0.9052	0.9096	0.9101	0.9150	0.9140	**0.9184**	0.9165
n=20	0.8698	0.8870	0.9032	0.9086	0.9071	0.9115	0.9111	0.9140	0.9135
n=21	0.8688	0.8806	0.9002	0.9022	0.9091	0.9115	0.9101	0.9145	0.9115
n=22	0.8722	0.8889	0.8973	0.8988	0.8983	0.9012	0.8963	0.8968	0.8978
n=23	0.8722	0.8889	0.8973	0.8988	0.8983	0.9012	0.8963	0.8968	0.8978
n=24	0.8717	0.8875	0.8929	0.8963	0.8983	0.8978	0.8998	0.9002	0.9012
n=25	0.8639	0.8811	0.8860	0.8855	0.8894	0.8894	0.8889	0.8914	0.8934
n=26	0.8855	0.9052	0.9027	0.9022	0.9022	0.9047	0.9022	0.9047	0.9052
n=27	0.8865	0.9061	0.9032	0.9017	0.9012	0.9037	0.9002	0.9017	0.9032
n=28	0.8983	0.9179	0.9150	0.9130	0.9145	0.9179	0.9165	0.9170	**0.9199**
n=29	0.9002	**0.9199**	**0.9184**	**0.9184**	**0.9199**	**0.9233**	**0.9204**	**0.9224**	**0.9238**

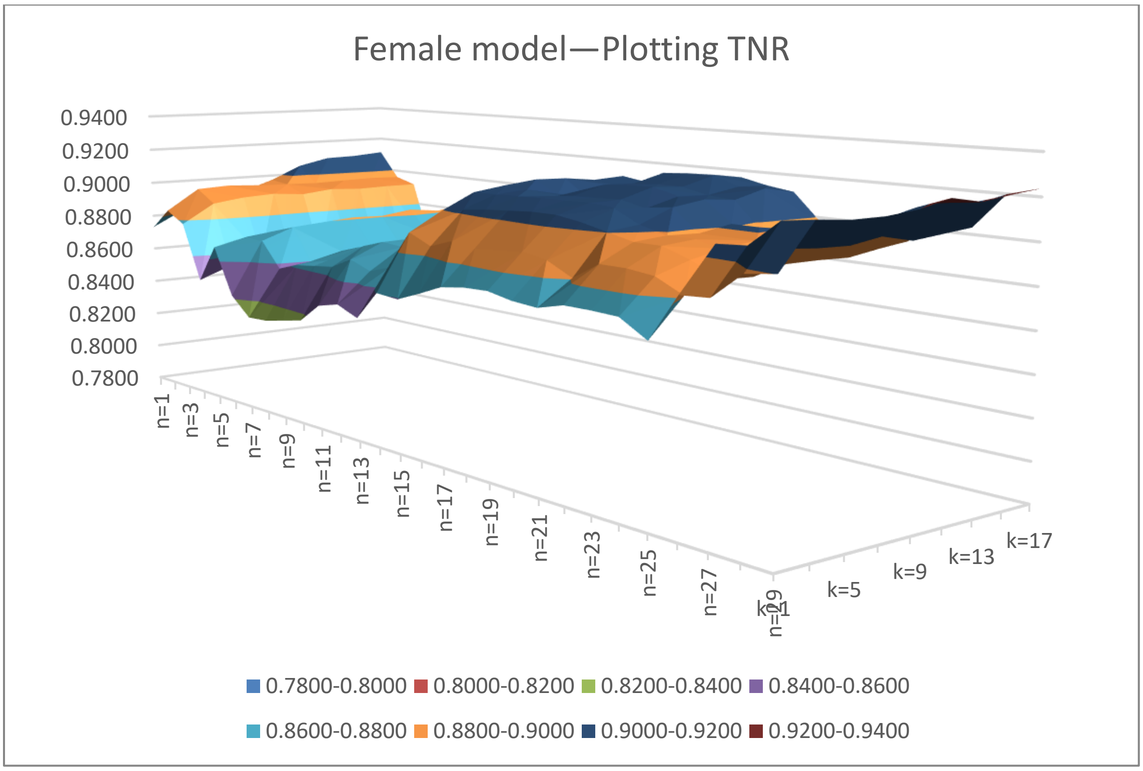


Figure S-12: Female model—Plotting TNR

Table S-20 displays prediction Precision results for experimenting the female dataset. The top ten Precision values are in dark-red bold text. The highest Precision value has yellow highlighted background. The Precision results are plotted as a 3D graph in Figure S-13.

Table S-20: Female model—Precision

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.0915	0.1049	0.1061	0.1061	0.1030	0.1023	0.1019	0.0980	0.0995
n=2	0.1031	0.1059	0.1068	0.1059	0.1055	0.1050	0.1055	0.1059	0.1055
n=3	0.0949	0.0960	0.0980	0.0988	0.0976	0.0992	0.0947	0.0950	0.0950
n=4	0.1060	0.1051	0.1077	0.1033	0.1051	0.1054	0.1071	0.1095	0.1131
n=5	0.1151	0.1304	0.1333	0.1325	0.1331	0.1346	**0.1350**	**0.1368**	0.1331
n=6	0.1154	0.1300	0.1329	**0.1373**	**0.1388**	**0.1383**	0.1345	**0.1357**	0.1324
n=7	0.0997	0.1159	0.1176	0.1182	0.1274	0.1214	0.1254	0.1270	0.1283
n=8	0.1117	0.1269	0.1277	0.1327	**0.1350**	0.1327	0.1311	0.1320	**0.1353**
n=9	0.1061	0.1029	0.1036	0.1067	0.1133	0.1104	0.1070	0.1122	0.1107
n=10	0.1075	0.1097	0.1142	0.1190	0.1185	0.1155	0.1224	0.1227	0.1250
n=11	0.1063	0.1111	0.1174	0.1210	0.1245	0.1218	0.1232	0.1255	0.1255
n=12	0.1029	0.1107	0.1166	0.1079	0.1213	0.1202	0.1183	0.1199	0.1241
n=13	0.1076	0.1137	0.1303	0.1306	0.1323	**0.1383**	0.1325	0.1323	0.1323
n=14	0.0984	0.1111	0.1282	0.1343	0.1333	**0.1359**	0.1182	0.1244	0.1214
n=15	0.0949	0.1132	0.1189	0.1221	0.1238	0.1116	0.1166	0.1179	0.1148
n=16	0.0769	0.1032	0.1005	0.0971	0.0821	0.0853	0.0905	0.0936	0.0945
n=17	0.0862	0.0980	0.1014	0.0887	0.0796	0.0974	0.0876	0.0870	0.0899
n=18	0.0874	0.0952	0.1023	0.0842	0.0800	0.0984	0.0885	0.0879	0.0909
n=19	0.0877	0.0924	0.0981	0.0891	0.0804	0.0990	0.0933	0.0929	0.0909
n=20	0.0893	0.0909	0.0963	0.0882	0.0870	0.0955	0.0905	0.0933	0.0881
n=21	0.0887	0.0899	0.0978	0.0913	0.0931	0.0955	0.0941	0.0984	0.0909
n=22	0.1065	0.1067	0.0991	0.1004	0.0921	0.0905	0.0905	0.0870	0.0877
n=23	0.1065	0.1067	0.0991	0.1004	0.0921	0.0946	0.0905	0.0870	0.0877
n=24	0.1062	0.1192	0.1102	0.1059	0.1000	0.0957	0.0973	0.1018	0.1027
n=25	0.1036	0.1103	0.1111	0.1004	0.0964	0.0927	0.0996	0.1053	0.1070
n=26	0.1141	0.1065	0.1161	0.1116	0.1116	0.1060	0.0995	0.0977	0.0896
n=27	0.1081	0.1075	0.1166	0.1111	0.1067	0.1050	0.0938	0.0950	0.0880
n=28	0.1000	0.0973	0.1036	0.1015	0.0984	0.0824	0.0811	0.0714	0.0894
n=29	0.0897	0.0944	0.0978	0.0929	0.0944	0.0877	0.0847	0.0867	0.0882

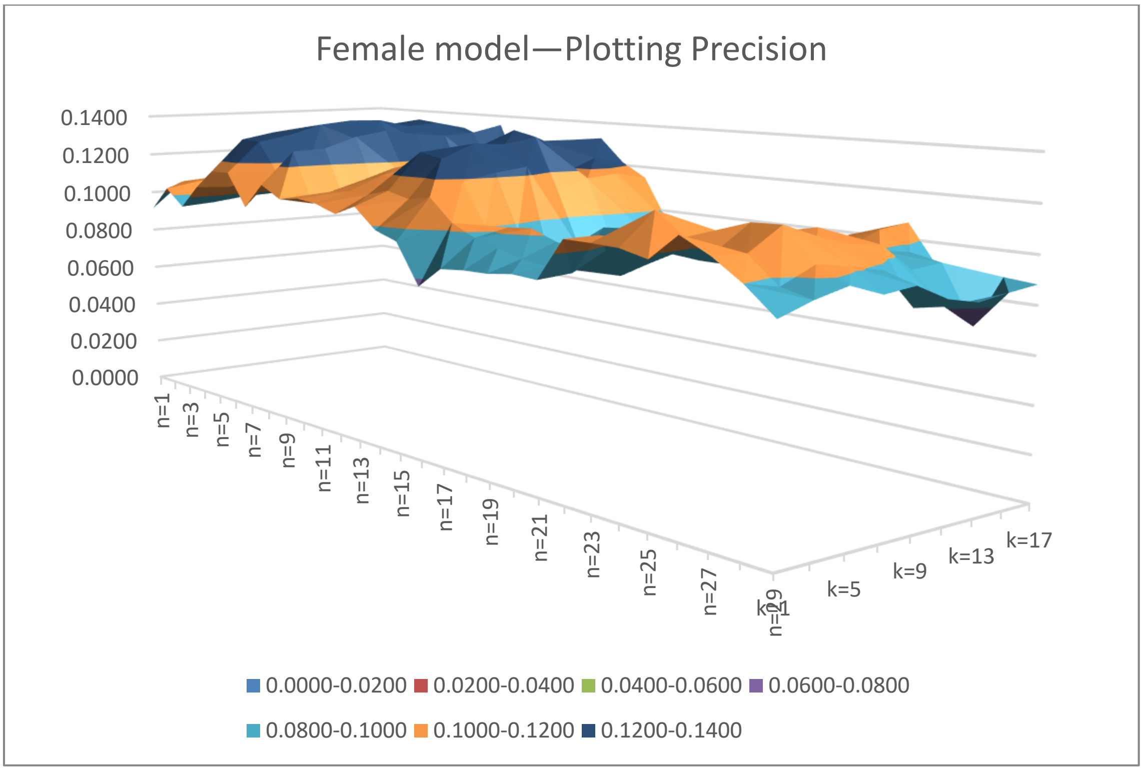


Figure S-13: Female model—Plotting Precision

Table S-21 displays prediction F_1-value results for experimenting the female dataset. The top ten F_1-values are in dark-red bold text. The highest F_1-value has yellow highlighted background. The F_1-value results are plotted as a 3D graph in Figure S-14.

Table S-21: Female model—F_1-value

	k=1	**k=3**	**k=5**	**k=7**	**k=9**	**k=11**	**k=13**	**k=15**	**k=17**
n=1	0.1486	0.1682	0.1697	0.1697	0.1605	0.1566	0.1544	0.1481	0.1498
n=2	0.1646	0.1656	0.1667	0.1656	0.1650	0.1645	0.1650	0.1656	0.1650
n=3	0.1529	0.1519	0.1543	0.1553	0.1538	0.1558	0.1489	0.1494	0.1494
n=4	0.1783	0.1754	0.1790	0.1722	0.1754	0.1759	0.1791	0.1832	0.1891
n=5	0.1892	0.2137	0.2186	0.2174	0.2193	**0.2222**	**0.2228**	**0.2252**	0.2193
n=6	0.1953	0.2131	0.2159	**0.2229**	**0.2248**	**0.2241**	0.2170	**0.2197**	0.2130
n=7	0.1700	0.1929	0.1954	0.1953	0.2105	0.2005	0.2060	0.2091	0.2108
n=8	0.1907	0.2108	0.2119	0.2187	**0.2228**	0.2187	0.2156	0.2168	**0.2222**
n=9	0.1806	0.1698	0.1707	0.1749	0.1858	0.1808	0.1753	0.1833	0.1813
n=10	0.1826	0.1809	0.1859	0.1944	0.1926	0.1866	0.1989	0.1970	0.2023
n=11	0.1787	0.1818	0.1902	0.1960	0.1994	0.1958	0.1988	0.2018	0.2006
n=12	0.1724	0.1823	0.1891	0.1744	0.1953	0.1914	0.1890	0.1922	0.1988
n=13	0.1814	0.1863	0.2114	0.2096	0.2105	**0.2194**	0.2095	0.2105	0.2105
n=14	0.1627	0.1761	0.2000	0.2057	0.2029	0.2059	0.1784	0.1891	0.1838
n=15	0.1571	0.1813	0.1843	0.1864	0.1884	0.1708	0.1799	0.1799	0.1745

n=16	0.1260	0.1635	0.1544	0.1471	0.1245	0.1300	0.1394	0.1413	0.1423
n=17	0.1404	0.1558	0.1555	0.1338	0.1199	0.1456	0.1308	0.1280	0.1333
n=18	0.1420	0.1509	0.1566	0.1269	0.1203	0.1467	0.1318	0.1290	0.1344
n=19	0.1425	0.1460	0.1500	0.1343	0.1208	0.1473	0.1390	0.1365	0.1344
n=20	0.1457	0.1442	0.1479	0.1333	0.1319	0.1434	0.1358	0.1390	0.1313
n=21	0.1448	0.1441	0.1512	0.1404	0.1407	0.1434	0.1418	0.1467	0.1364
n=22	0.1737	0.1693	0.1544	0.1559	0.1429	0.1394	0.1409	0.1351	0.1361
n=23	0.1737	0.1693	0.1544	0.1559	0.1429	0.1458	0.1409	0.1351	0.1361
n=24	0.1732	0.1902	0.1736	0.1656	0.1554	0.1486	0.1507	0.1575	0.1586
n=25	0.1707	0.1775	0.1774	0.1600	0.1524	0.1465	0.1577	0.1661	0.1683
n=26	0.1824	0.1631	0.1793	0.1724	0.1724	0.1625	0.1533	0.1495	0.1367
n=27	0.1723	0.1643	0.1799	0.1718	0.1649	0.1614	0.1448	0.1463	0.1348
n=28	0.1554	0.1434	0.1544	0.1521	0.1467	0.1210	0.1195	0.1048	0.1306
n=29	0.1384	0.1382	0.1440	0.1365	0.1382	0.1266	0.1235	0.1255	0.1271

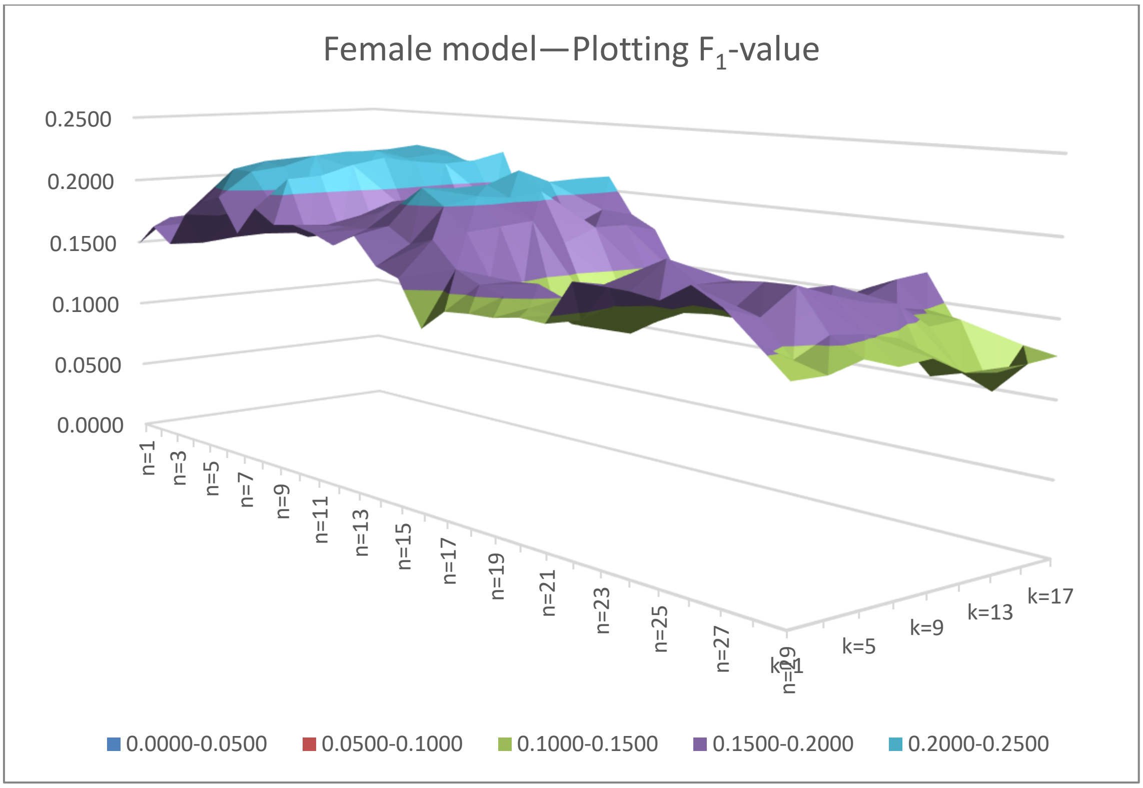


Figure S-14: Female model—Plotting F$_1$-value

Table S-22 displays prediction NPV results for experimenting the female dataset. The top ten NPV values are in dark-red bold text. The highest NPV value has yellow highlighted background. The NPV results are plotted as a 3D graph in Figure S-15.

Table S-22: Female model—NPV

	k=1	k=3	k=5	k=7	k=9	k=11	k=13	k=15	k=17
n=1	0.9780	0.9793	0.9793	0.9793	0.9775	0.9767	0.9763	0.9758	0.9758
n=2	0.9788	0.9780	0.9780	0.9780	0.9780	0.9780	0.9780	0.9780	0.9780
n=3	0.9781	0.9773	0.9774	0.9774	0.9774	0.9774	0.9769	0.9769	0.9769
n=4	0.9834	0.9825	0.9825	0.9819	0.9825	0.9825	0.9830	0.9836	0.9841
n=5	0.9827	0.9850	0.9856	0.9855	**0.9861**	**0.9866**	**0.9866**	**0.9866**	**0.9861**
n=6	**0.9862**	0.9850	0.9846	0.9851	0.9852	0.9852	0.9841	0.9846	0.9836
n=7	0.9837	0.9842	0.9843	0.9838	0.9855	0.9843	0.9844	0.9849	0.9850
n=8	**0.9866**	0.9859	0.9860	**0.9860**	**0.9866**	**0.9860**	0.9855	0.9855	**0.9861**
n=9	0.9849	0.9810	0.9810	0.9811	0.9822	0.9817	0.9811	0.9817	0.9817
n=10	0.9850	0.9821	0.9818	0.9828	0.9824	0.9814	0.9829	0.9820	0.9830
n=11	0.9835	0.9817	0.9819	0.9824	0.9820	0.9820	0.9825	0.9825	0.9820
n=12	0.9824	0.9822	0.9818	0.9803	0.9820	0.9810	0.9810	0.9815	0.9820
n=13	0.9840	0.9822	0.9840	0.9831	0.9826	0.9832	0.9822	0.9826	0.9826
n=14	0.9804	0.9794	0.9807	0.9804	0.9799	0.9799	0.9779	0.9789	0.9784
n=15	0.9798	0.9804	0.9792	0.9788	0.9788	0.9777	0.9787	0.9783	0.9778
n=16	0.9761	0.9784	0.9766	0.9757	0.9741	0.9746	0.9755	0.9752	0.9753
n=17	0.9774	0.9778	0.9766	0.9747	0.9737	0.9753	0.9743	0.9739	0.9744
n=18	0.9774	0.9773	0.9767	0.9742	0.9737	0.9754	0.9743	0.9739	0.9744
n=19	0.9774	0.9768	0.9762	0.9747	0.9737	0.9754	0.9748	0.9745	0.9744
n=20	0.9779	0.9767	0.9761	0.9747	0.9747	0.9753	0.9748	0.9748	0.9743
n=21	0.9779	0.9771	0.9765	0.9756	0.9752	0.9753	0.9753	0.9754	0.9748
n=22	0.9807	0.9789	0.9770	0.9770	0.9760	0.9755	0.9759	0.9754	0.9754
n=23	0.9807	0.9789	0.9770	0.9770	0.9760	0.9761	0.9759	0.9754	0.9754
n=24	0.9807	0.9810	0.9790	0.9780	0.9770	0.9765	0.9765	0.9771	0.9771
n=25	0.9810	0.9803	0.9799	0.9783	0.9773	0.9768	0.9778	0.9784	0.9785
n=26	0.9804	0.9772	0.9787	0.9782	0.9782	0.9772	0.9766	0.9761	0.9751
n=27	0.9794	0.9772	0.9787	0.9781	0.9776	0.9772	0.9760	0.9761	0.9751
n=28	0.9770	0.9749	0.9759	0.9758	0.9754	0.9734	0.9734	0.9724	0.9740
n=29	0.9755	0.9745	0.9750	0.9745	0.9745	0.9736	0.9735	0.9735	0.9736

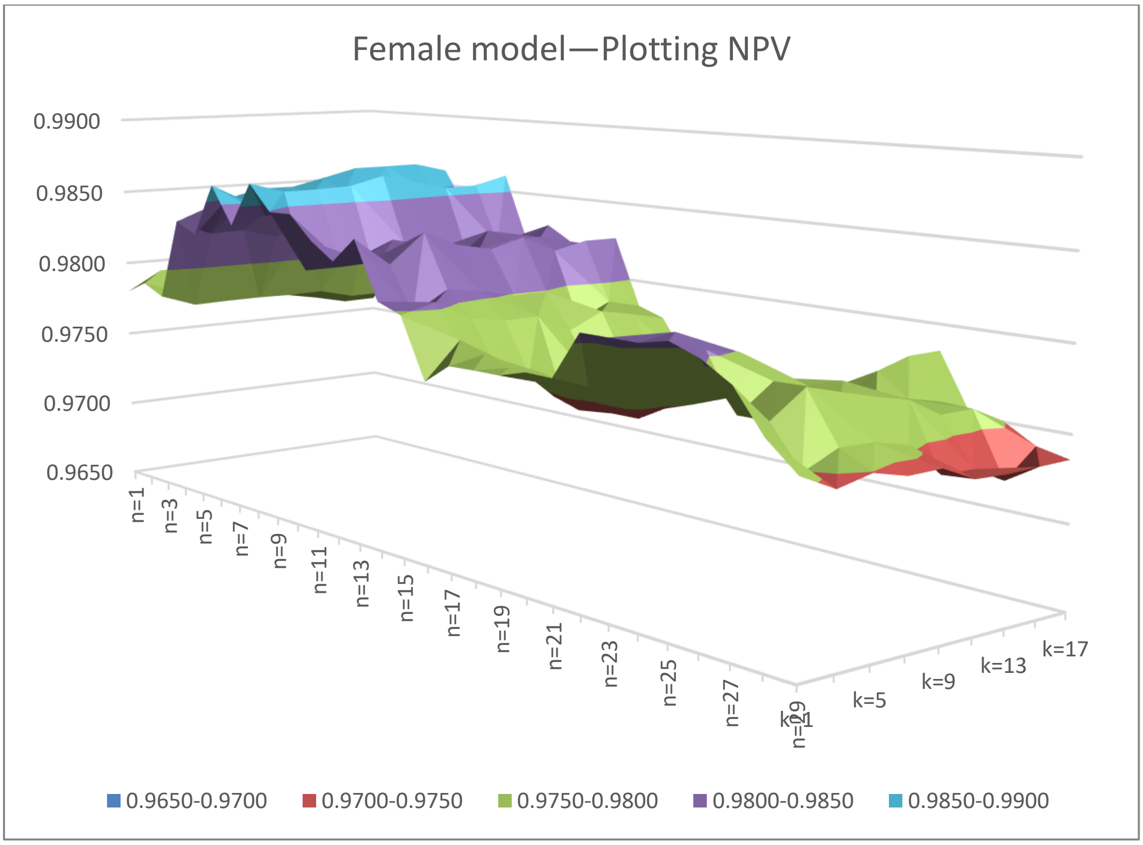


Figure S-15: Female model—Plotting NPV

Appendix T Data Distributions of the Mixed sex dataset and it's "SMOTEd" one

The figures in this appendix show comparison of the mixed sex dataset and its "SMOTEd" one by comparing Weka univariate attribute distributions of the chosen 13 predictors. The red and blue colours refer to positive and negative classes respectively.

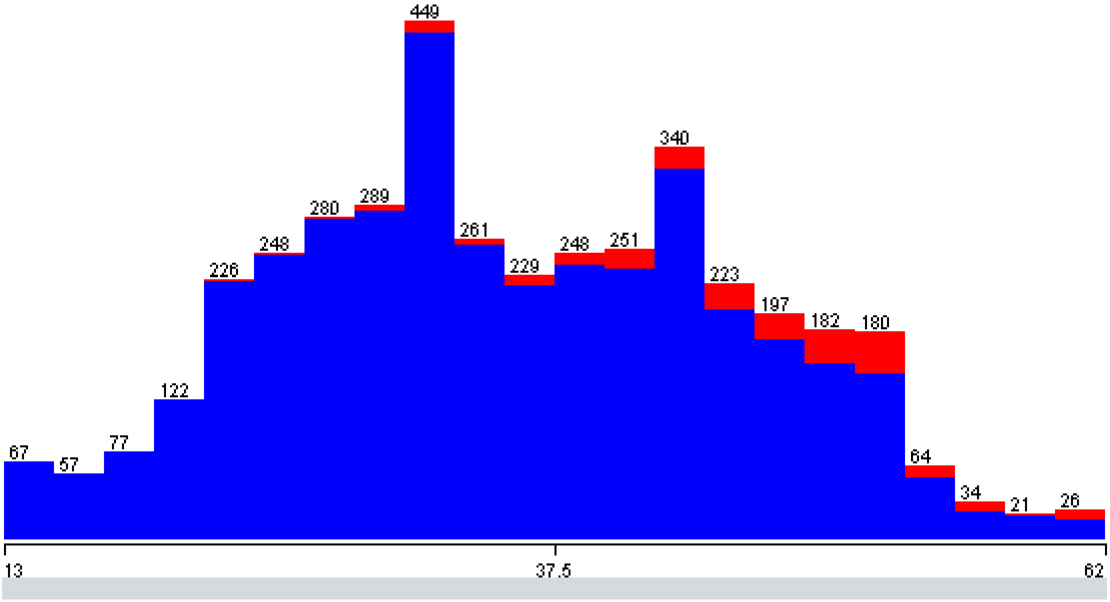


Figure T-16: Univariate attribute distribution of Age in the mixed sex dataset

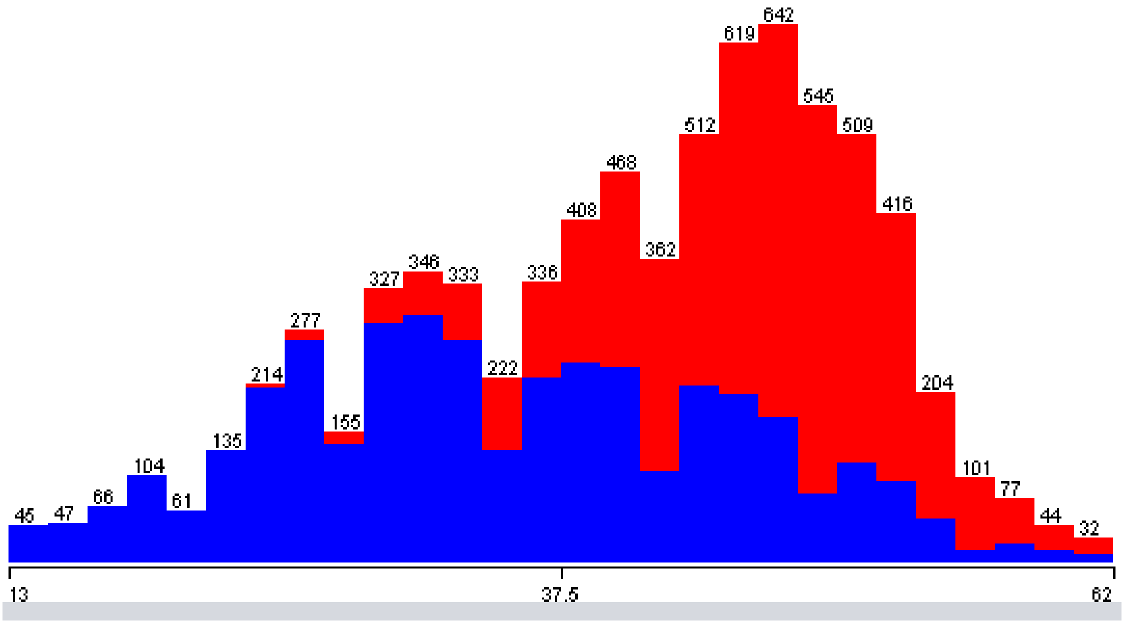


Figure T-17: Univariate attribute distribution of Age in the "SMOTEd" mixed sex dataset

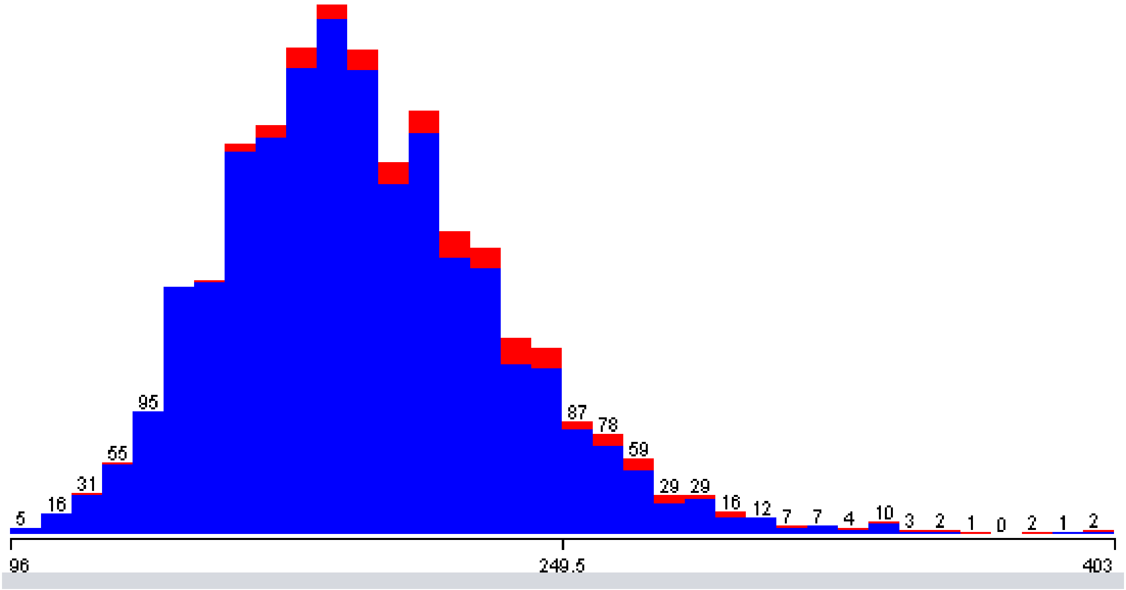


Figure T-18: Univariate attribute distribution of Total Cholesterol in the mixed sex dataset

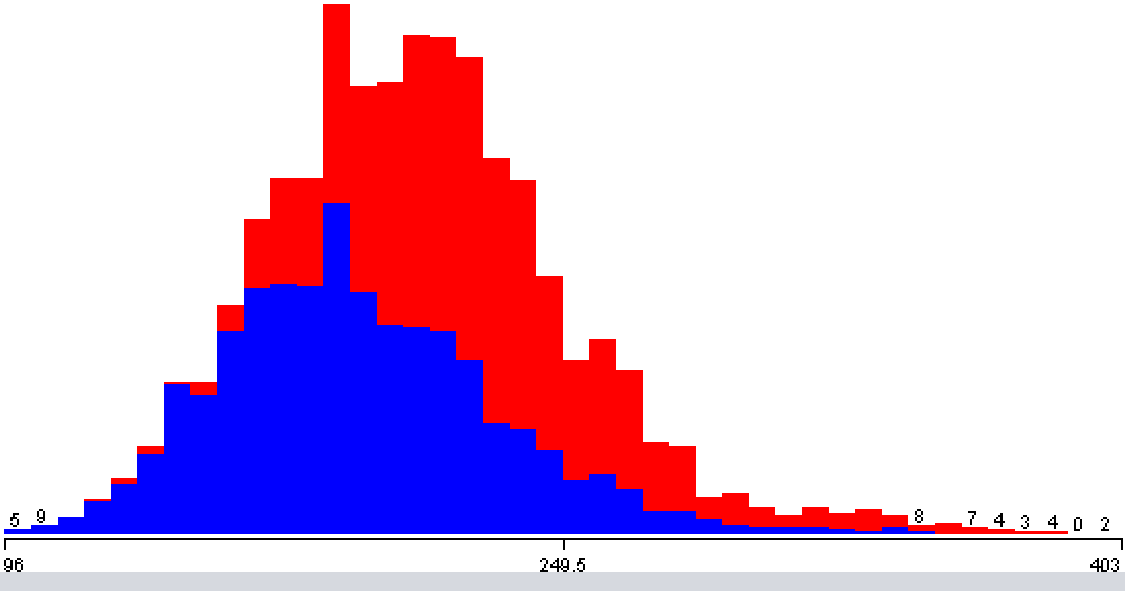


Figure T-19: Univariate attribute distribution of Total Cholesterol in the "SMOTEd" mixed sex dataset

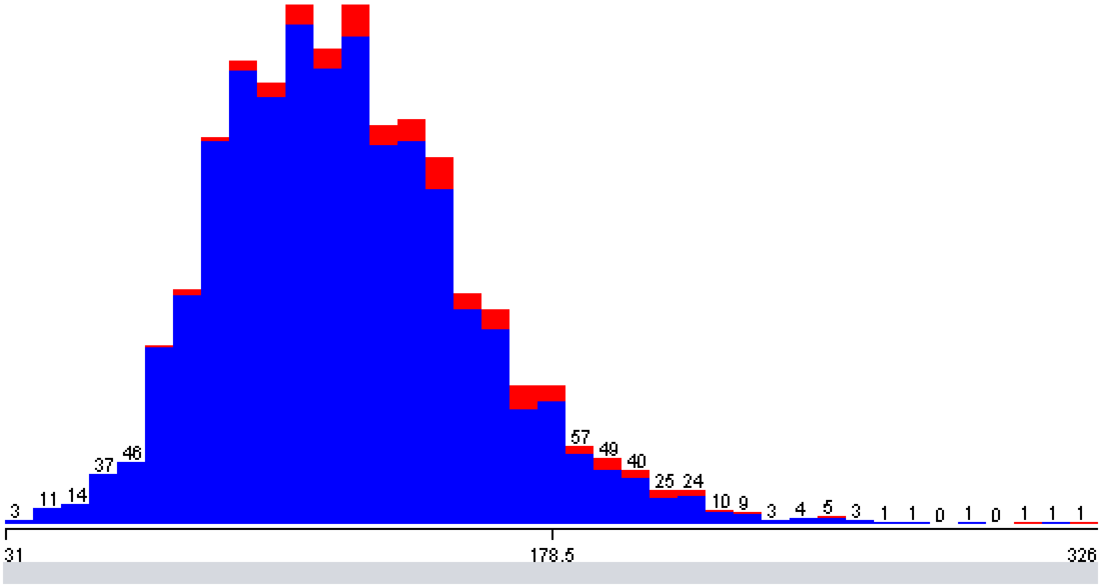


Figure T-20: Univariate attribute distribution of LDL Cholesterol in the mixed sex dataset

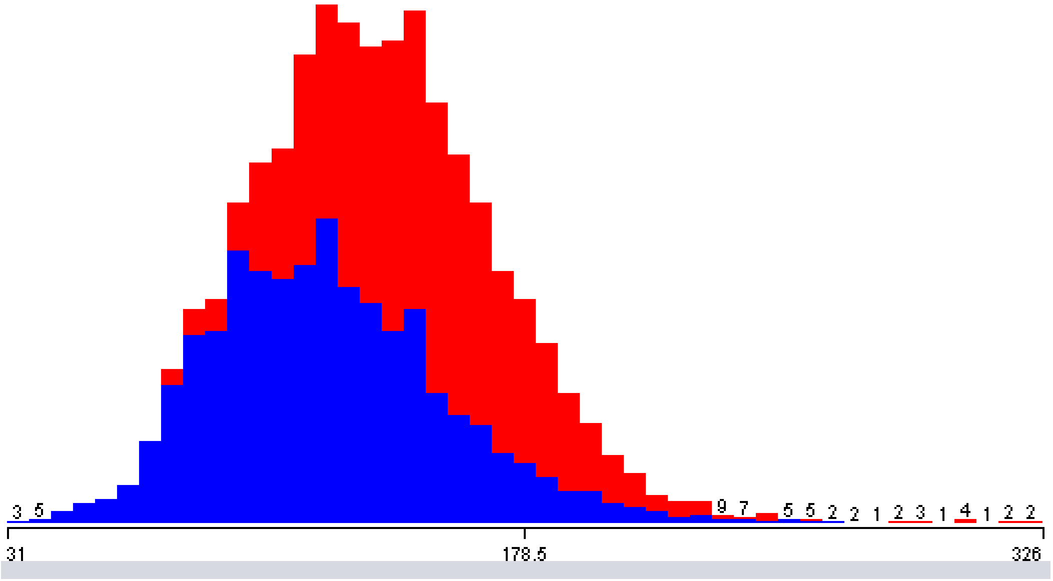


Figure T-21: Univariate attribute distribution of LDL Cholesterol in the "SMOTEd" mixed sex dataset

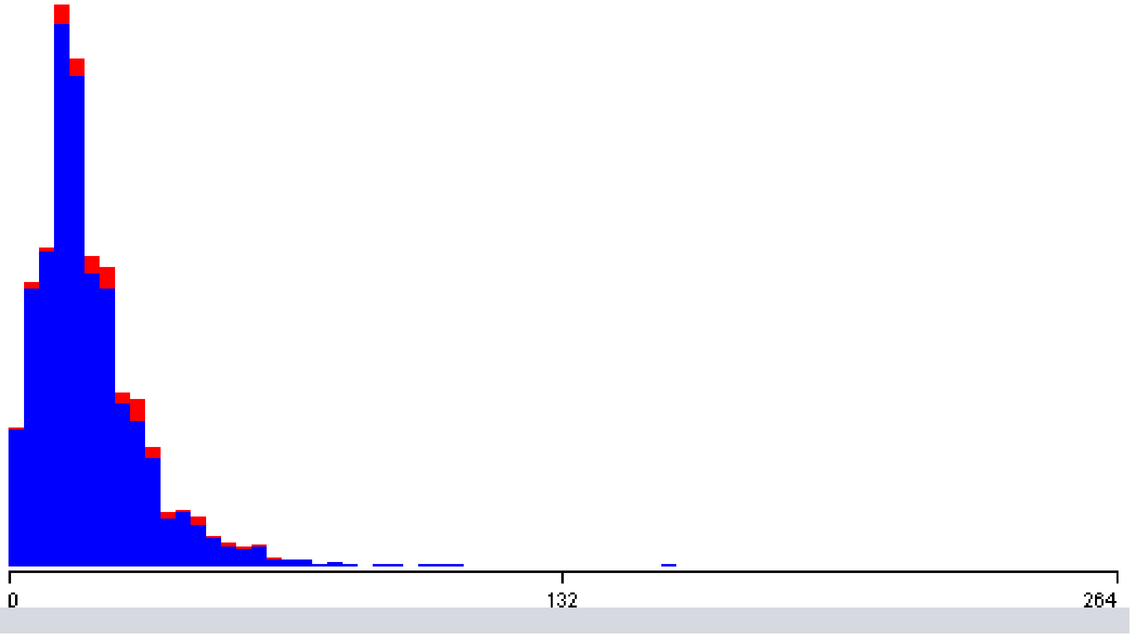


Figure T-22: Univariate attribute distribution of VLDL Cholesterol in the mixed sex dataset

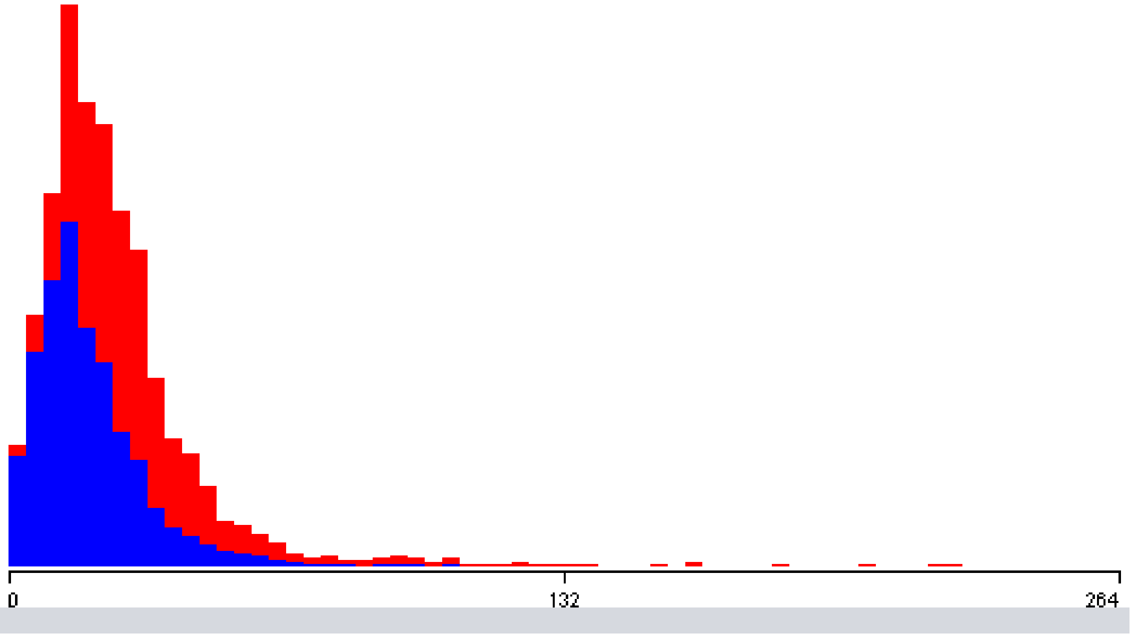


Figure T-23: Univariate attribute distribution of VLDL Cholesterol in the “SMOTEd” mixed sex dataset

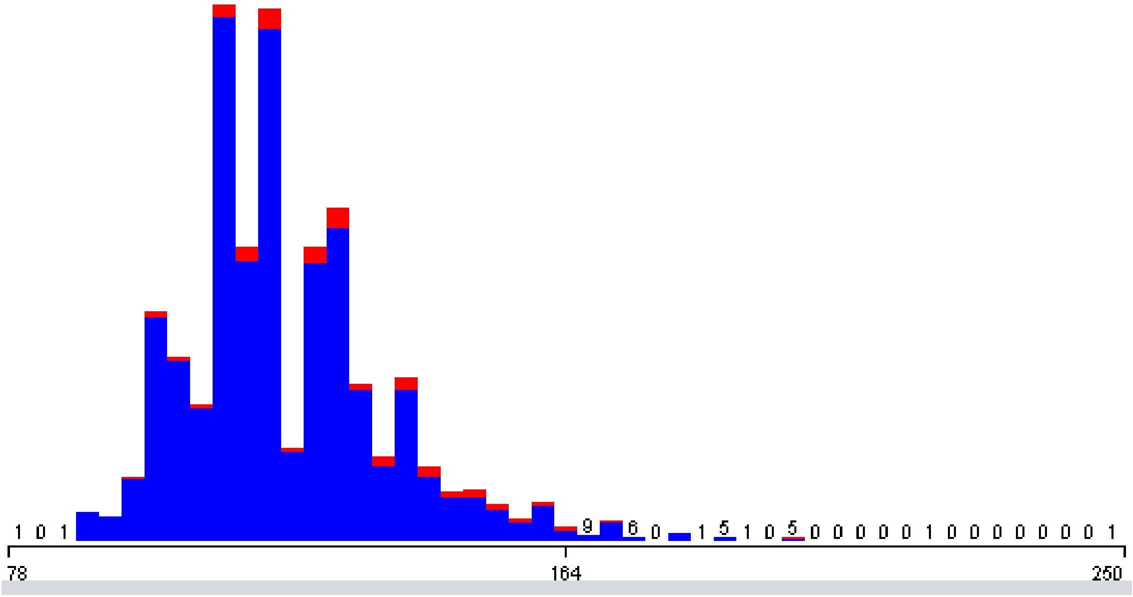


Figure T-24: Univariate attribute distribution of SBP in the mixed sex dataset

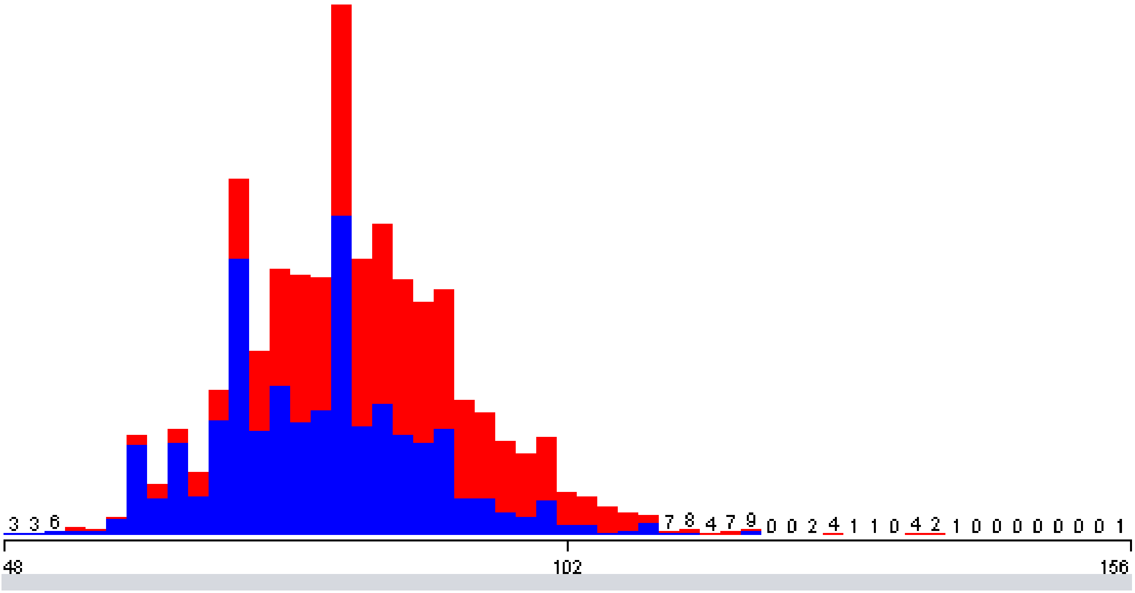


Figure T-25: Univariate attribute distribution of SBP in the "SMOTEd" mixed sex dataset

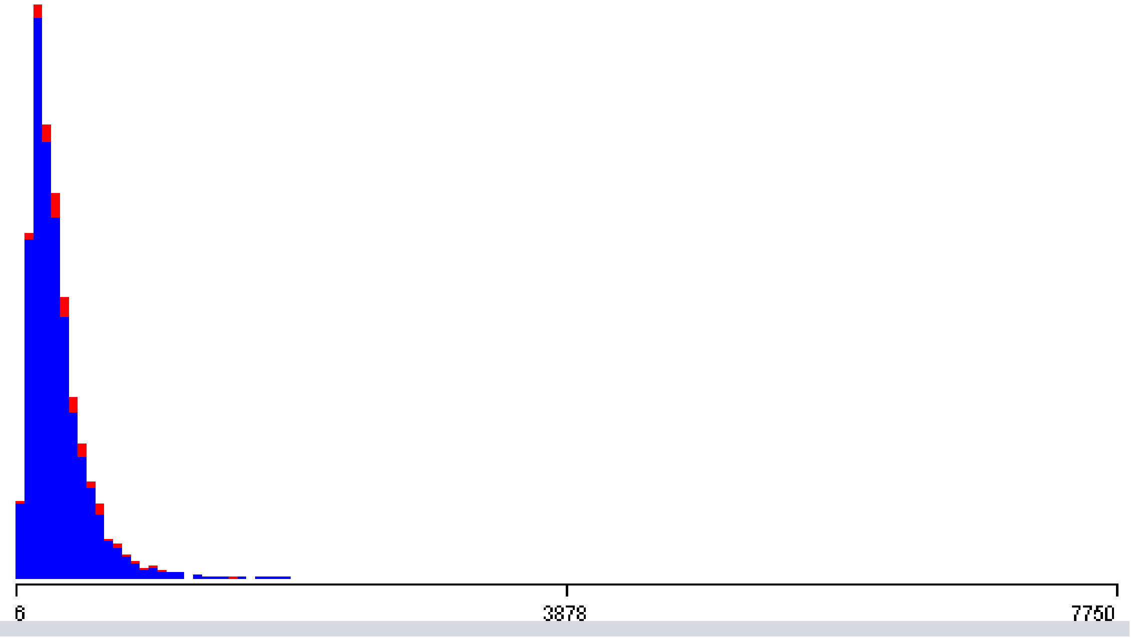


Figure T-26: Univariate attribute distribution of Triglycerides in the mixed sex dataset

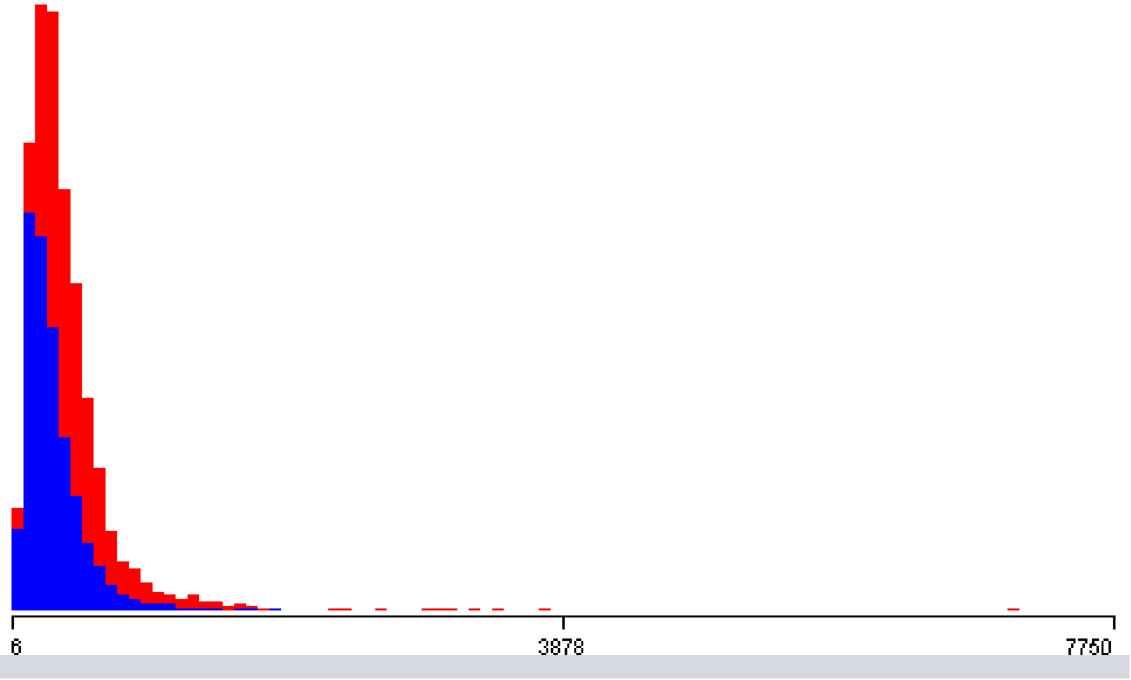


Figure T-27: Univariate attribute distribution of Triglycerides in the “SMOTEd” mixed sex dataset

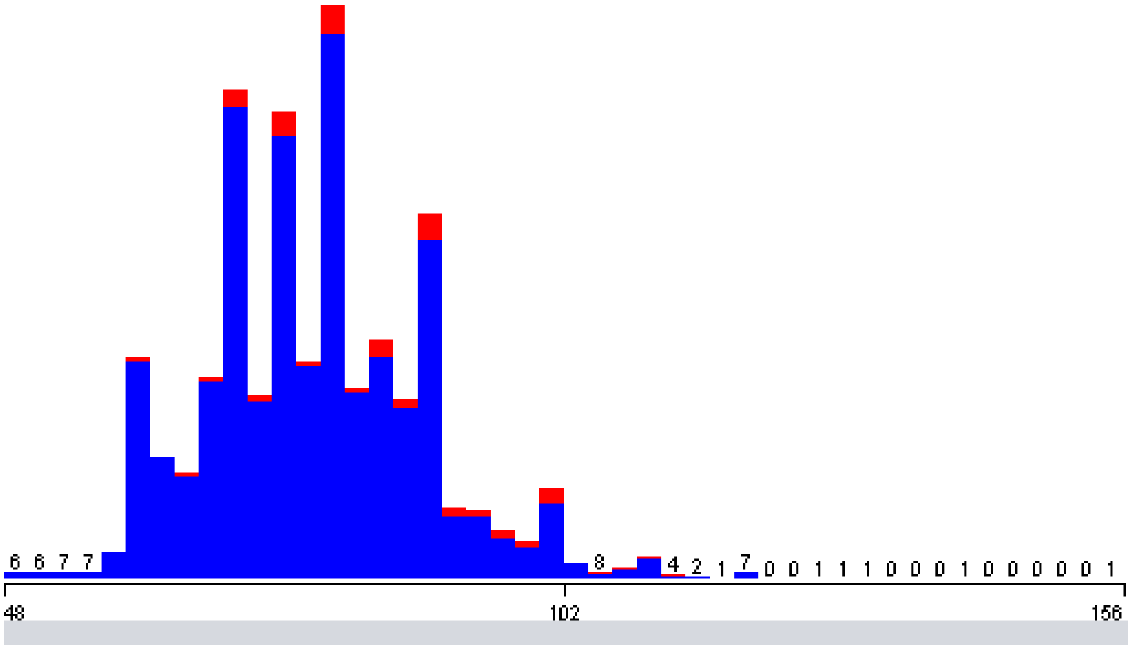


Figure T-28: Univariate attribute distribution of DBP in the mixed sex dataset

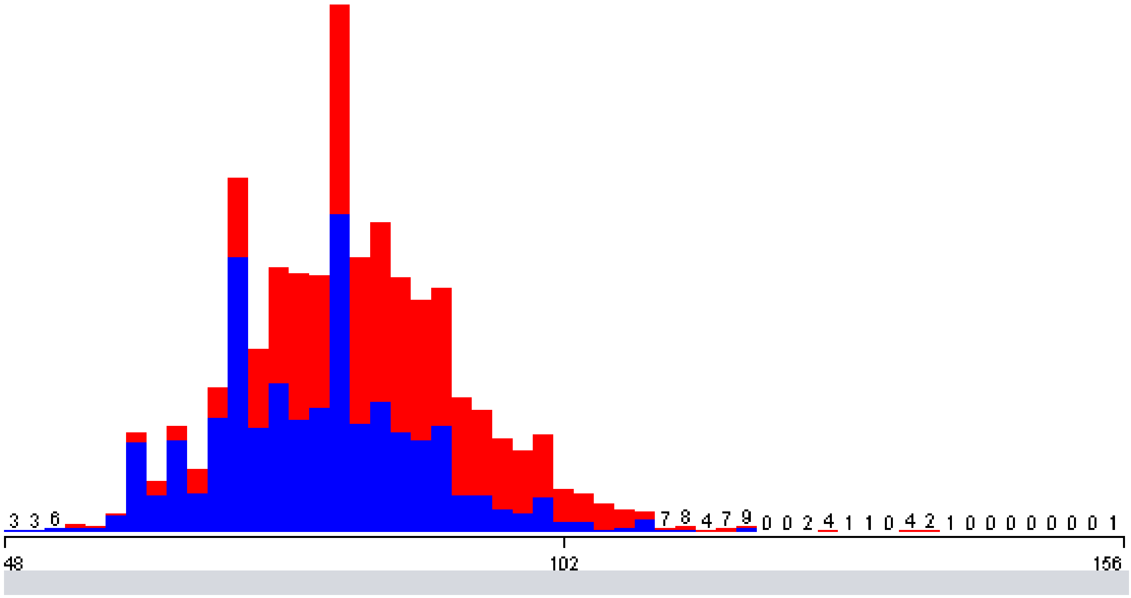


Figure T-29: Univariate attribute distribution of DBP in the "SMOTEd" mixed sex dataset

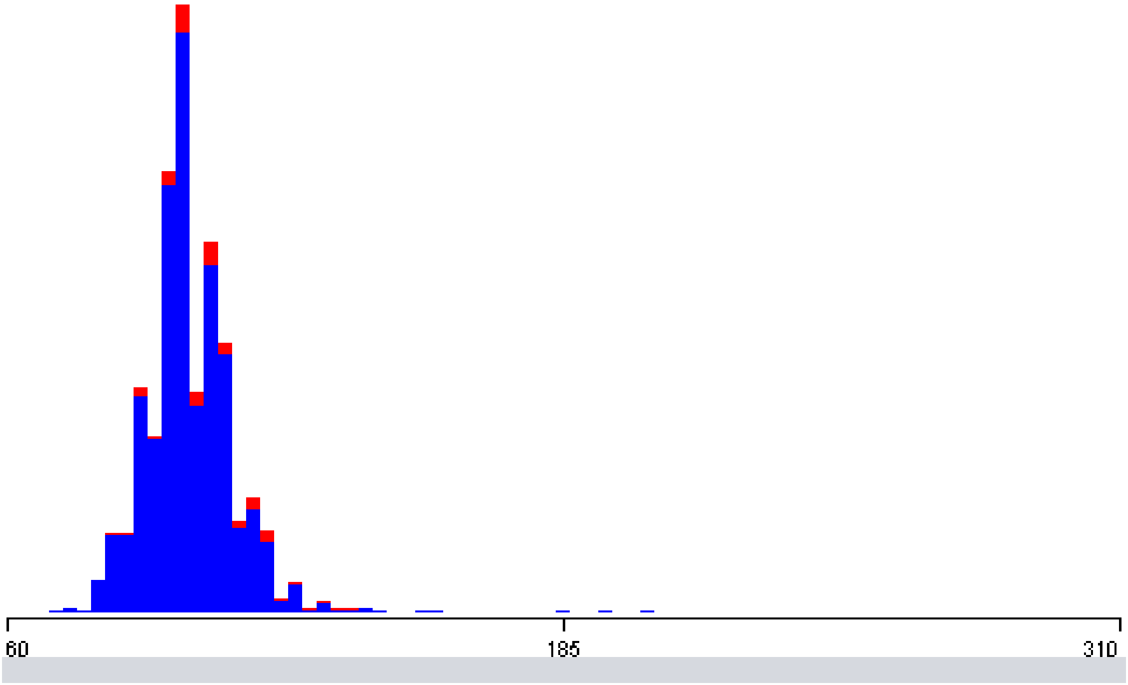


Figure T-30: Univariate attribute distribution of Glucose in the mixed sex dataset

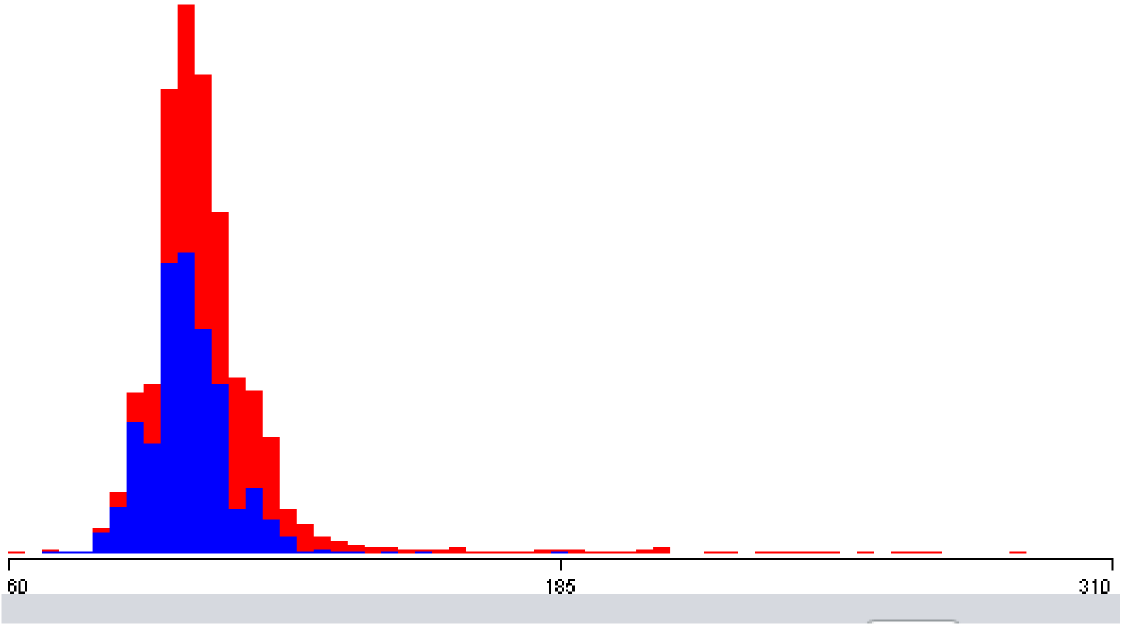


Figure T-31: Univariate attribute distribution of Glucose in the “SMOTEd” mixed sex dataset

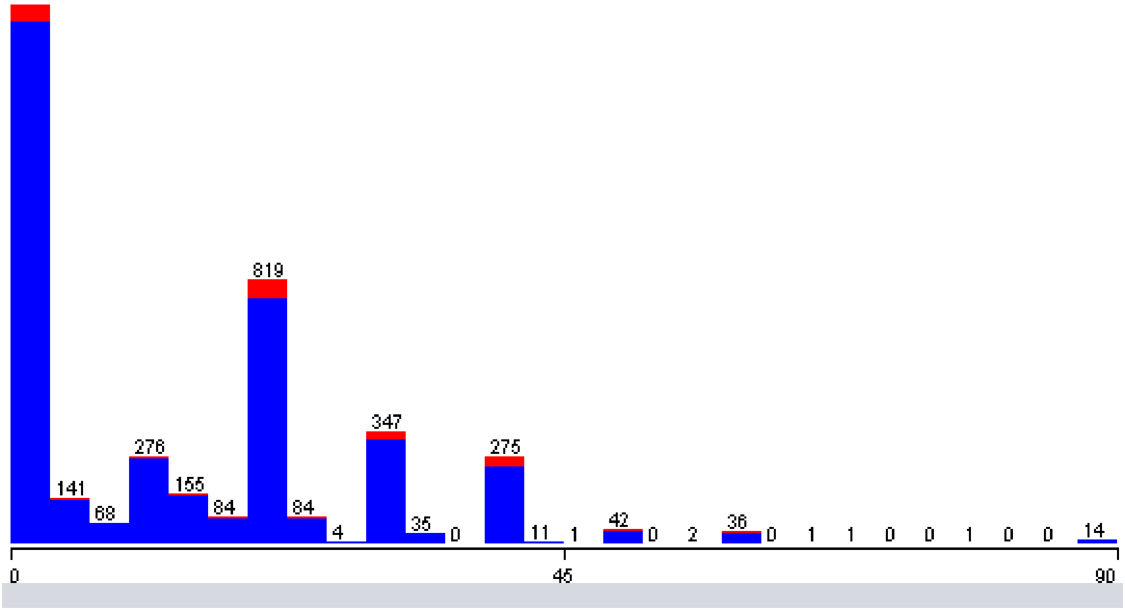


Figure T-32: Univariate attribute distribution of Cigarettes in the mixed sex dataset

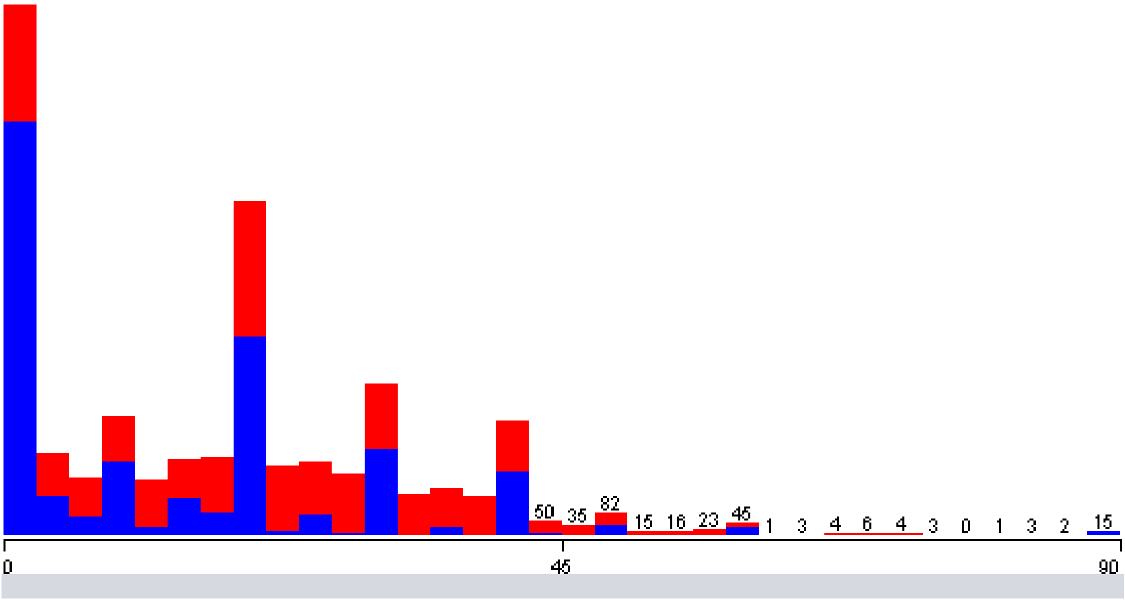


Figure T-33: Univariate attribute distribution of Cigarettes in the “SMOTEd” mixed sex dataset

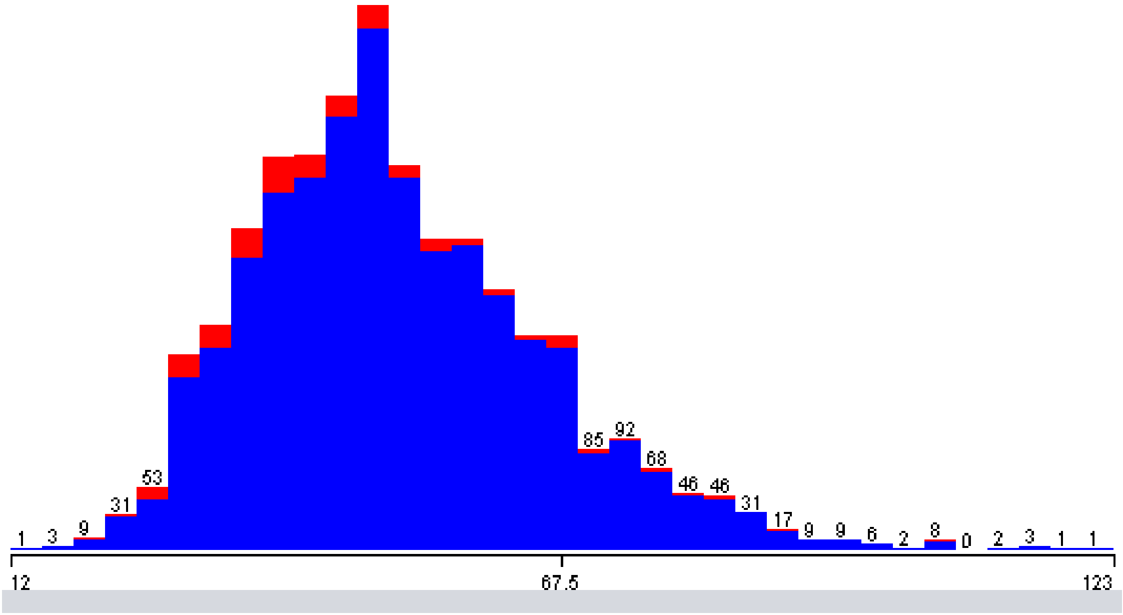


Figure T-34: Univariate attribute distribution of HDL Cholesterol in the mixed sex dataset

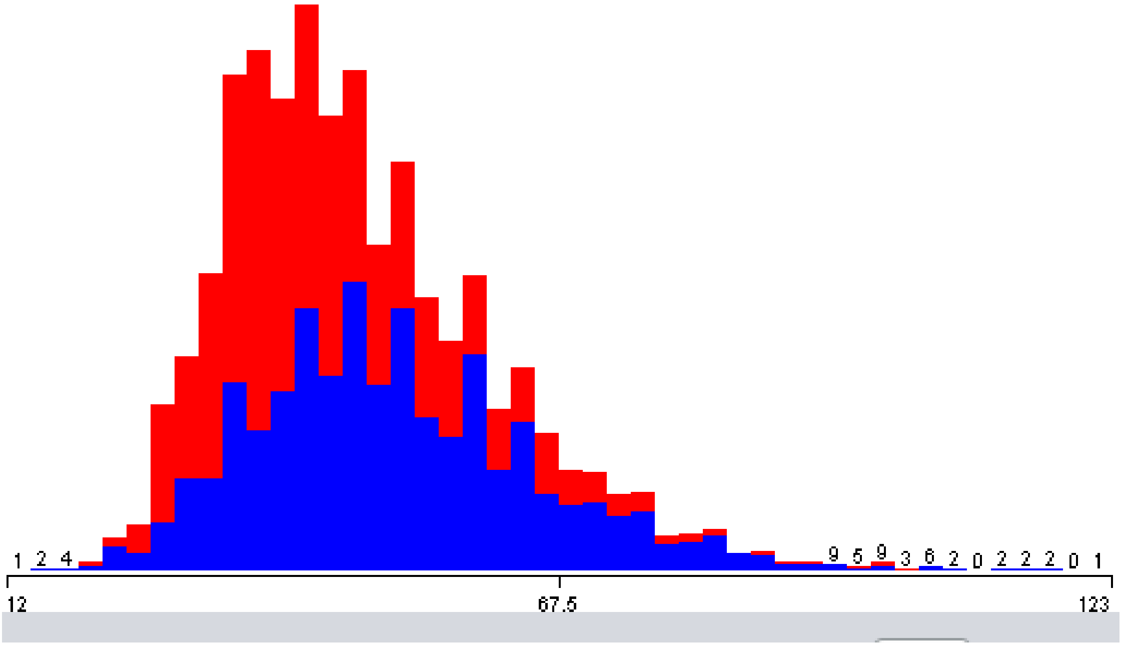


Figure T-35: Univariate attribute distribution of HDL Cholesterol in the "SMOTEd" mixed sex dataset

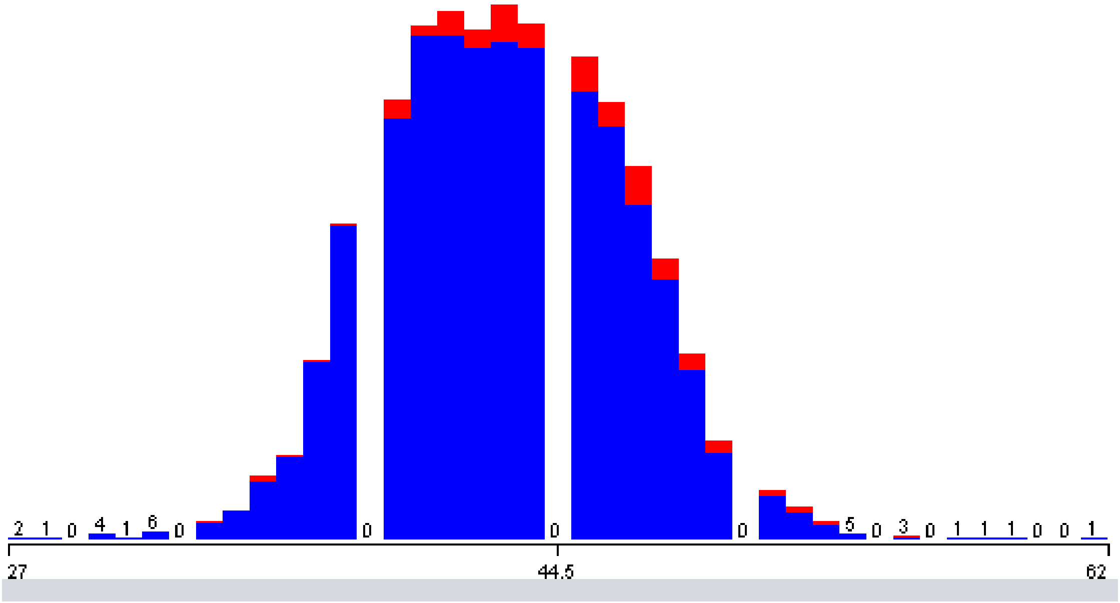


Figure T-36: Univariate attribute distribution of Hematocrit in the mixed sex dataset

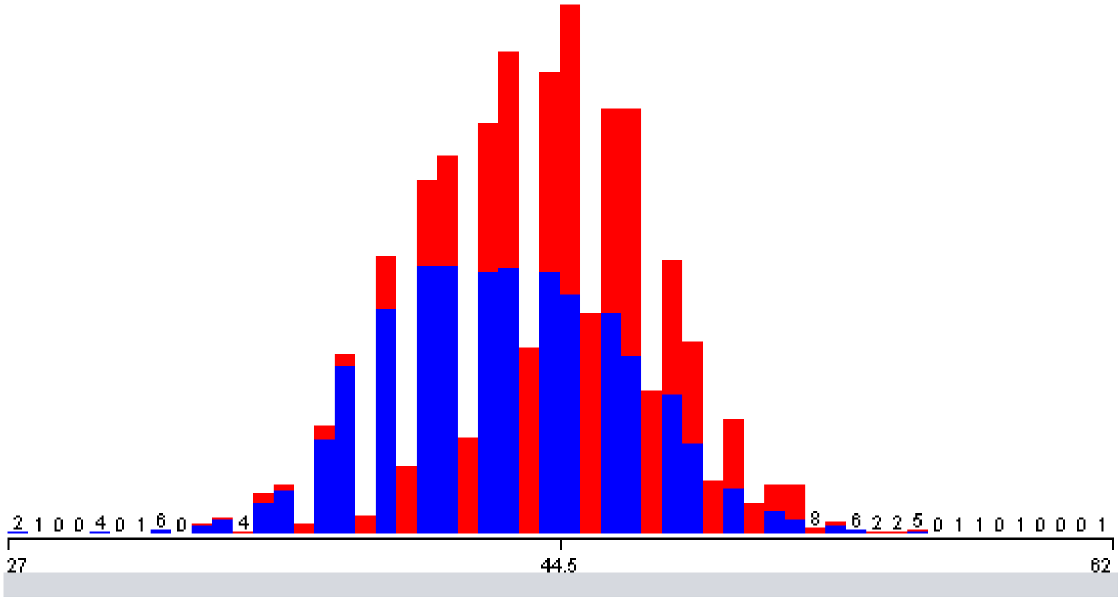


Figure T-37: Univariate attribute distribution of Hematocrit in the “SMOTEd” mixed sex dataset

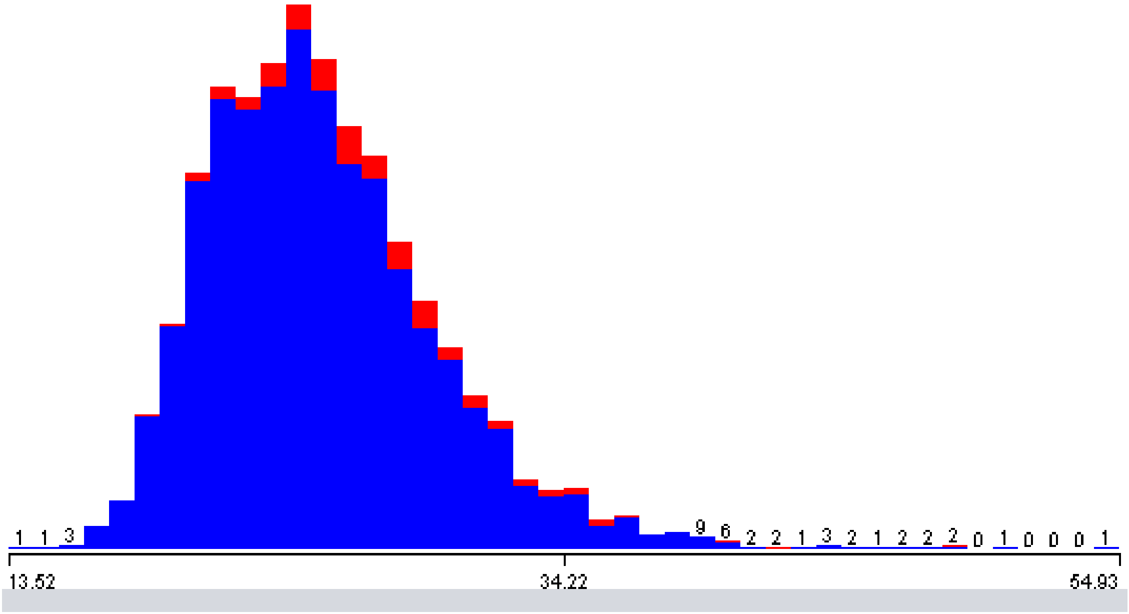


Figure T-38: Univariate attribute distribution of BMI in the mixed sex dataset

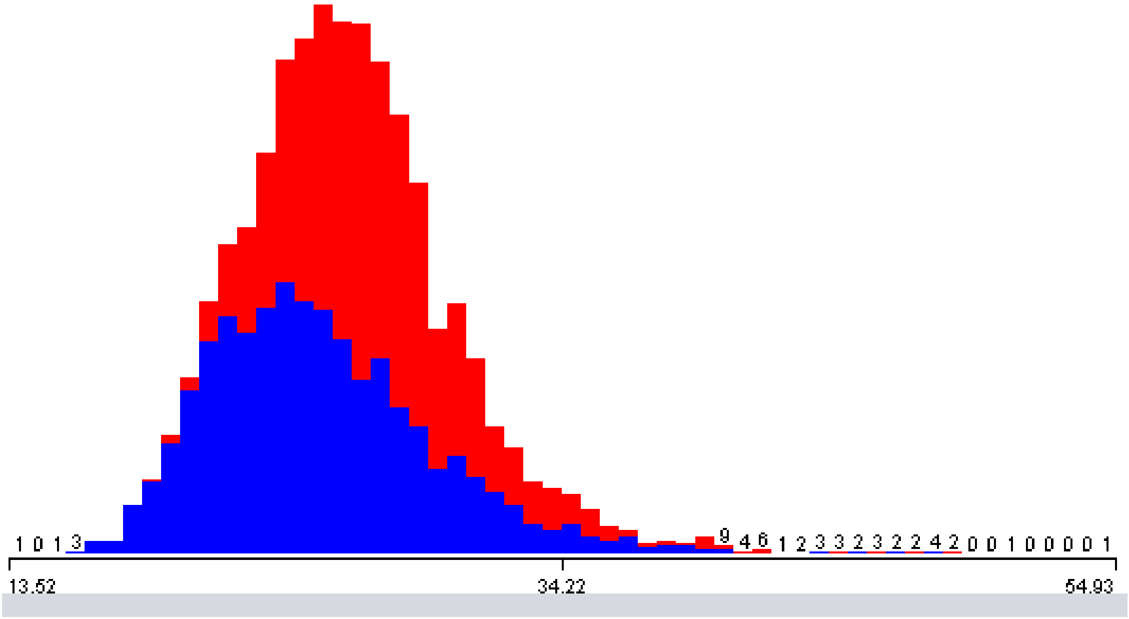


Figure T-39: Univariate attribute distribution of BMI in the “SMOTEd” mixed sex dataset

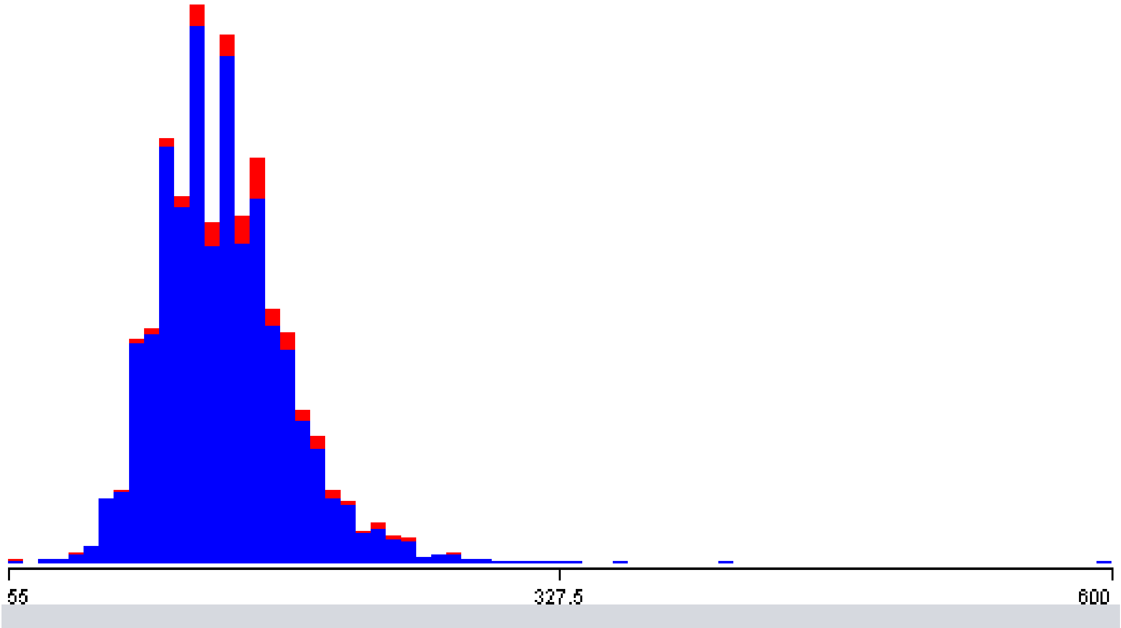


Figure T-40: Univariate attribute distribution of LDH in the mixed sex dataset

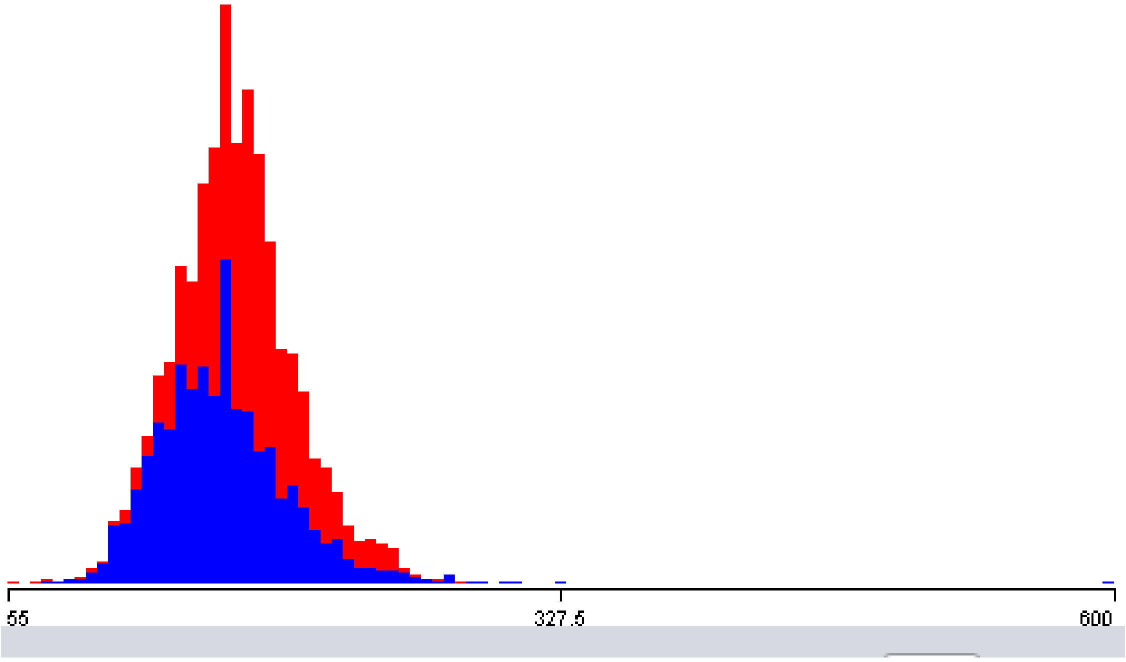
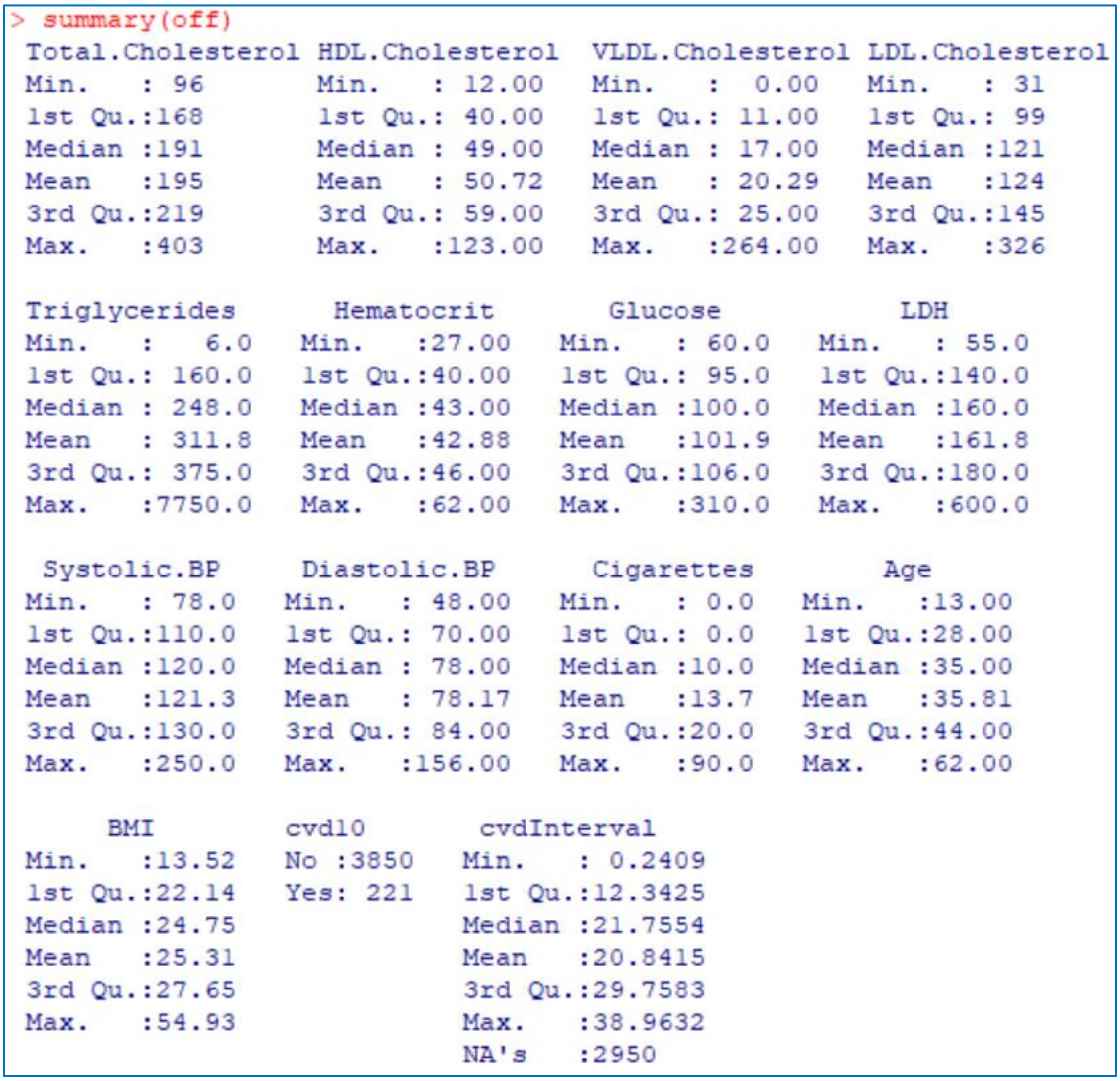


Figure T-41: Univariate attribute distribution of LDH in the “SMOTEd” mixed sex dataset

Appendix U SUMMARY OF THE MIXED SEX DATASET

Figure U-42 summarises the mixed sex dataset with 13 risk factors chosen to be predictors for CVD prediction. The summary was done in R.

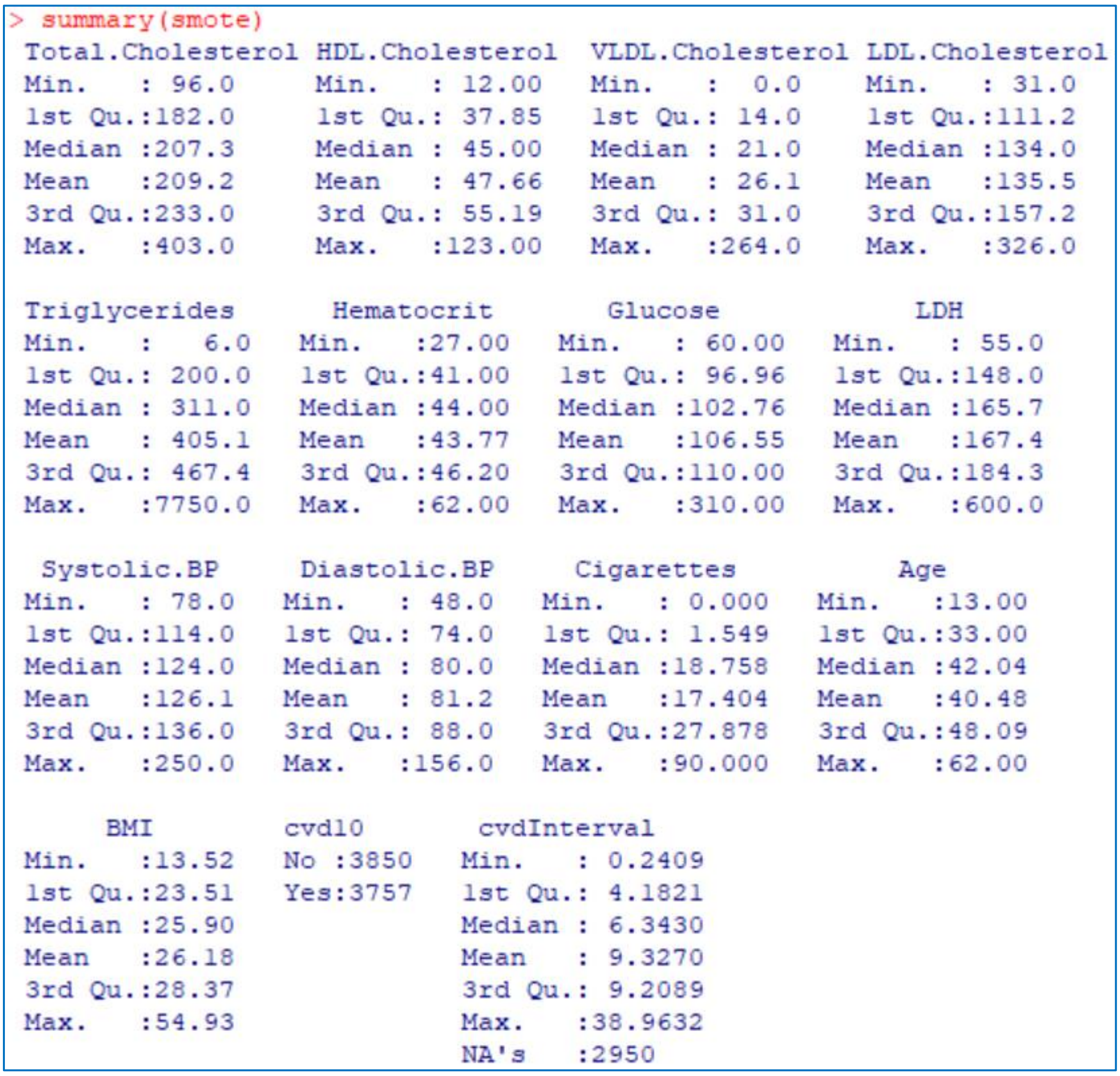


```
> summary(off)
 Total.Cholesterol HDL.Cholesterol  VLDL.Cholesterol LDL.Cholesterol
 Min.   : 96       Min.   : 12.00   Min.   :  0.00   Min.   : 31
 1st Qu.:168       1st Qu.: 40.00   1st Qu.: 11.00   1st Qu.: 99
 Median :191       Median : 49.00   Median : 17.00   Median :121
 Mean   :195       Mean   : 50.72   Mean   : 20.29   Mean   :124
 3rd Qu.:219       3rd Qu.: 59.00   3rd Qu.: 25.00   3rd Qu.:145
 Max.   :403       Max.   :123.00   Max.   :264.00   Max.   :326

 Triglycerides      Hematocrit        Glucose            LDH
 Min.   :   6.0   Min.   :27.00   Min.   : 60.0   Min.   : 55.0
 1st Qu.: 160.0   1st Qu.:40.00   1st Qu.: 95.0   1st Qu.:140.0
 Median : 248.0   Median :43.00   Median :100.0   Median :160.0
 Mean   : 311.8   Mean   :42.88   Mean   :101.9   Mean   :161.8
 3rd Qu.: 375.0   3rd Qu.:46.00   3rd Qu.:106.0   3rd Qu.:180.0
 Max.   :7750.0   Max.   :62.00   Max.   :310.0   Max.   :600.0

  Systolic.BP     Diastolic.BP      Cigarettes         Age
 Min.   : 78.0   Min.   : 48.00   Min.   : 0.0   Min.   :13.00
 1st Qu.:110.0   1st Qu.: 70.00   1st Qu.: 0.0   1st Qu.:28.00
 Median :120.0   Median : 78.00   Median :10.0   Median :35.00
 Mean   :121.3   Mean   : 78.17   Mean   :13.7   Mean   :35.81
 3rd Qu.:130.0   3rd Qu.: 84.00   3rd Qu.:20.0   3rd Qu.:44.00
 Max.   :250.0   Max.   :156.00   Max.   :90.0   Max.   :62.00

      BMI          cvd10       cvdInterval
 Min.   :13.52   No :3850   Min.   : 0.2409
 1st Qu.:22.14   Yes: 221   1st Qu.:12.3425
 Median :24.75              Median :21.7554
 Mean   :25.31              Mean   :20.8415
 3rd Qu.:27.65              3rd Qu.:29.7583
 Max.   :54.93              Max.   :38.9632
                            NA's   :2950
```

Figure U-42: Summary of the mixed sex dataset, with the chosen 13 predictors, in R

Appendix V SUMMARY OF THE "SMOTED" MIXED SEX DATASET

Figure V-43 summarises the "SMOTed" mixed sex dataset with 13 risk factors chosen to be predictors for CVD prediction. The summary was done in R.



```
> summary(smote)
 Total.Cholesterol HDL.Cholesterol  VLDL.Cholesterol LDL.Cholesterol
 Min.   : 96.0     Min.   : 12.00   Min.   :  0.0    Min.   : 31.0
 1st Qu.:182.0     1st Qu.: 37.85   1st Qu.: 14.0    1st Qu.:111.2
 Median :207.3     Median : 45.00   Median : 21.0    Median :134.0
 Mean   :209.2     Mean   : 47.66   Mean   : 26.1    Mean   :135.5
 3rd Qu.:233.0     3rd Qu.: 55.19   3rd Qu.: 31.0    3rd Qu.:157.2
 Max.   :403.0     Max.   :123.00   Max.   :264.0    Max.   :326.0

 Triglycerides      Hematocrit        Glucose             LDH
 Min.   :   6.0   Min.   :27.00   Min.   : 60.00   Min.   : 55.0
 1st Qu.: 200.0   1st Qu.:41.00   1st Qu.: 96.96   1st Qu.:148.0
 Median : 311.0   Median :44.00   Median :102.76   Median :165.7
 Mean   : 405.1   Mean   :43.77   Mean   :106.55   Mean   :167.4
 3rd Qu.: 467.4   3rd Qu.:46.20   3rd Qu.:110.00   3rd Qu.:184.3
 Max.   :7750.0   Max.   :62.00   Max.   :310.00   Max.   :600.0

  Systolic.BP      Diastolic.BP     Cigarettes            Age
 Min.   : 78.0   Min.   : 48.0   Min.   : 0.000   Min.   :13.00
 1st Qu.:114.0   1st Qu.: 74.0   1st Qu.: 1.549   1st Qu.:33.00
 Median :124.0   Median : 80.0   Median :18.758   Median :42.04
 Mean   :126.1   Mean   : 81.2   Mean   :17.404   Mean   :40.48
 3rd Qu.:136.0   3rd Qu.: 88.0   3rd Qu.:27.878   3rd Qu.:48.09
 Max.   :250.0   Max.   :156.0   Max.   :90.000   Max.   :62.00

      BMI          cvd10        cvdInterval
 Min.   :13.52   No :3850   Min.   : 0.2409
 1st Qu.:23.51   Yes:3757   1st Qu.: 4.1821
 Median :25.90              Median : 6.3430
 Mean   :26.18              Mean   : 9.3270
 3rd Qu.:28.37              3rd Qu.: 9.2089
 Max.   :54.93              Max.   :38.9632
                            NA's   :2950
```

Figure V-43: Summary of the "SMOTEd" mixed sex dataset, with the chosen 13 predictors, in R