# Application of data science methodologies to explore, model and predict population-level subjective wellbeing outcomes using the New Zealand Integrated Data Infrastructure (IDI).

Anantha Narayanan T L

A thesis submitted to Auckland University of Technology in fulfilment of the requirements for the degree of Doctor of Philosophy

2024

School of Sport & Recreation, Human Potential Centre

Supervisors:

Dr Tom Stewart
Prof Scott Duncan
Prof Gail Pacheco

**Thesis Abstract**

The increasing recognition of population wellbeing as a key indicator of societal prosperity has spurred governments worldwide to formulate policies aimed at enhancing their citizens' wellbeing. In New Zealand, the General Social Survey (GSS) provides subjective wellbeing measures for a subset of the population (~10,000 individuals). Although the GSS sample is representative of the New Zealand population across various sociodemographic groups, including factors such as ethnicity and socio-economic status, the available wellbeing data only covers ~0.2% of the population. This limitation affects the granularity of insights it can offer, hindering in-depth exploration into (1) the determinants of population wellbeing and (2) the effects of government policies on wellbeing, especially within smaller and marginalised subgroups of the population who may be of high policy interest (e.g., people living in deprived regions). To address this challenge, detailed population-level wellbeing data that is sensitive enough to reflect the effects of policy change is essential. The microdata available within New Zealand's Integrated Data Infrastructure (IDI) presents an opportunity to bridge this gap. By leveraging the IDI's extensive dataset, this thesis aims to apply advanced data science methodologies to systematically explore, model, and predict GSS-based subjective wellbeing outcomes for the broader New Zealand population.

To begin, a systematic scoping review was conducted to provide a comprehensive overview of the current literature around modelling health and wellbeing outcomes using machine learning. This foundational review identified prevailing trends, methodologies, and notably, the scarcity of studies predicting population wellbeing outcomes. Next, the thesis delved into understanding New Zealand's current wellbeing through detailed cross-sectional and trend analysis of GSS data. Key associations with subjective wellbeing were observed across diverse demographic categories, including age, gender, ethnicity, and socio-economic status. These

insights set the stage for the next part of the thesis focused on modelling subjective wellbeing outcomes.

The core of the thesis focussed on the development and validation of statistical models for predicting subjective wellbeing within the GSS population. Census-level administrative variables were utilised as predictors, and the Random Forest emerged as an effective model for showcasing how data science techniques can predict wellbeing outcomes. Despite its strengths, capturing the variability of subjective wellbeing proved challenging, prompting a critical discussion on the need for methodological refinement. Lastly, the research extended to applying and then validating these predictive models in the broader New Zealand census population. While predictions generally aligned with GSS estimates across different demographic groups, several disparities underscored the complexities of accurately modelling wellbeing at the population level, which were discussed in the final chapter of this thesis.

The thesis demonstrated the feasibility of predicting subjective wellbeing outcomes in a population with existing routinely collected data and advanced analytical techniques. It acknowledged the challenges of modelling subjective outcomes, and suggested avenues to enhance the precision and applicability of this research methodology. This work contributes to the broader understanding and enhancement of population wellbeing, underscoring the importance of comprehensive, representative data for informing policy and societal progress.

# Table of Contents

## List of Figures



## List of Tables

**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 (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

……

Anantha Narayanan T L, April 2024

**Co-authored works**

***Papers under review or in submission:***

Narayanan A, Stewart T, Duncan S, Pacheco G. (*In Review*). A cross-validation study to investigate the efficacy of census-level socio-demographic factors for predicting subjective-wellbeing outcomes in New Zealand. *Submitted to Scientific Reports.* https://doi.org/10.21203/rs.3.rs-4266983/v1

Narayanan A, Stewart T, Duncan S, Pacheco G. (*In Review*). Application of machine learning for predicting health and wellbeing outcomes from population datasets: A systematic scoping review. *Submitted to Scientific Reports.*

Narayanan A, Stewart T, Duncan S, Pacheco G. (2024). Predicting subjective wellbeing outcomes for the New Zealand population using census-level data. *Preparing for submission to the Australian and New Zealand Journal of Public Health.*

Narayanan A, Stewart T, Duncan S, Pacheco G. (2024). Subjective wellbeing outcomes across different demographic groups and regions in New Zealand – A cross-sectional and trend analysis. *Preparing for submission to the Australian and New Zealand Journal of Public Health.*

***Peer-reviewed conference presentations:***

Narayanan A, Stewart T. (2024). Using machine learning to explore the efficacy of administrative variables in prediction of subjective wellbeing outcomes in New Zealand. *Accepted for presentation at the International Conference on Data Management, Analytics, and Innovation (ICDMAI), Singapore, July 2024*.

Narayanan A, Stewart T, Duncan S (2021). Application of data science methodologies to explore, predict, and model wellbeing outcomes using the New Zealand Integrated Data Infrastructure (IDI). *Presented at the Innovations in Applied Data Symposium.* https://terourou.org/symposium/applied-computing/

Narayanan A. (2022). Modelling and prediction of New Zealand's population wellbeing using machine learning techniques. *Rangahau Aranga: AUT Graduate Review, 1*(1). https://doi.org/10.24135/rangahau-aranga.v1i1.44

**Research chapter contributions**

Chapters 3 to 6 of this thesis are either under review or in preparation for submission to peer reviewed journals. Ananth Narayanan (the principal author) was responsible for designing all studies with assistance from Tom Stewart, Scott Duncan, and Gail Pacheco. Data cleaning and processing for all the studies was performed by Ananth Narayanan. Ananth Narayanan also conducted the analysis and drafted the manuscripts, with critical feedback provided by all other authors. The percentage contribution of each author is presented below.

| | |
|---|---|
| Anantha Narayanan……..... | 85% |
| Tom Stewart……... | 10% |
| Scott Duncan……. | 2.5% |
| Gail Pacheco……... | 2.5% |

**Co-author agreement.**

Dr Tom Stewart Professor Scott Duncan Professor Gail Pacheco

**Acknowledgements**

This research would not have been possible without the financial support from the AUT Vice-Chancellor Scholarship. I am also thankful to the *Te Hotonga Hapori* project team and Prof. Scott Duncan for their support through a stipend.

Ethics approval for the study presented in this thesis was obtained from the Auckland University of Technology Ethics Committee (21/115) in April 2021, with further amendments approved in May 2021 (Appendix B).

Thanks to my primary supervisor, Dr Tom Stewart, for his invaluable guidance and unwavering support throughout this journey. Without Tom's mentorship, this research and thesis would not have been possible. I am also grateful to Prof. Scott Duncan and Prof. Gail Pacheco for their guidance, insightful feedback, and thought-provoking questions that have greatly contributed to the development of this thesis. Lastly, heartfelt thanks to my family and friends whose support and encouragement have been a constant source of strength, pushing me forward to achieve this important milestone.

# Chapter 1 – Introduction

## Background

The concept of wellbeing lies at the core of human existence, influencing the choices we make and the goals we pursue [1]. Our pursuits, whether on an individual or communal level, are inherently connected to the overarching goal of either improving or maintaining wellbeing. Understanding and enhancing wellbeing is not just a personal quest but also a collective endeavour that shapes the foundation of societies [2]. Wellbeing stands as a fundamental driver of human behaviour [3], reflecting the quality of life experienced by individuals. A nation's population thrives when individual wellbeing is high, leading to numerous societal benefits. Elevated levels of population wellbeing correlate with increased productivity, enhanced social cohesion, and a higher quality of life [4, 5]. Individuals in societies with high wellbeing tend to be healthier, more engaged in their communities, and resilient in the face of challenges [6, 7]. Governments worldwide are recognising the pivotal role of population wellbeing in shaping policy and decision-making processes [8]. Integrating wellbeing metrics into policy design becomes crucial as it directly influences the lived experiences of citizens. When policies are crafted with wellbeing as an end outcome, they contribute to creating environments that foster positive emotions, social connections, and a sense of fulfilment [9].

Precise measurement of wellbeing is a crucial step in formulating policies that positively impact population wellbeing. In New Zealand, detailed measures of wellbeing are collected as part of the General Social Survey (GSS), which is based on the New Zealand Treasury's Living Standards Framework (LSF) [10]. The LSF provides a comprehensive structure for understanding and assessing both individual and collective wellbeing across twelve different domains. This thesis will specifically focus on one of these domains known as "subjective wellbeing".

Subjective wellbeing refers to an individual's overall assessment of their life, encompassing both emotional experiences and cognitive evaluations of satisfaction within various domains such as work, relationships, and personal accomplishments [11]. It goes beyond traditional measures of success or material wealth and delves into the subjective aspects of happiness, life satisfaction, and positive emotions [12]. Understanding subjective wellbeing aids policymakers, psychologists, and researchers when developing interventions and policies that promote holistic wellbeing [13]. In the GSS, subjective wellbeing is assessed primarily through four indicators – life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. These will be the main outcomes of interest in this thesis.

Despite being nationally representative, the GSS data have a major drawback – these data are only available for a smaller subset (i.e., ~10,000) of New Zealand's 5.1 million population. This limitation highlights the need for more comprehensive data sources. The government's push for cross-agency data integration over the last several decades has led to the development of the Integrated Data Infrastructure (IDI). The IDI (managed by Stats NZ) is a population-level database containing individual response data (microdata) relating to people and households that can be linked longitudinally over time. It contains anonymised data about education and training, income and work, benefits and social services, health, justice, and housing.

The IDI hosts the GSS and administrative datasets like the census, and these datasets can be linked using a unique identifier ID. Extrapolating the GSS-based wellbeing outcomes to the full IDI population may be possible via advanced modelling techniques, such as machine learning models like random forest [14]. These models could be trained to predict subjective wellbeing outcomes using administrative and environmental variables within in the IDI. If successful, the outcome of this study (a population-level measure of wellbeing) would enable researchers to easily incorporate wellbeing measures into IDI-based policy analysis.

Ultimately, it would improve our understanding of how the political, social, and economic environment influences the wellbeing and functioning of New Zealanders.

**Thesis Rationale**

The importance of population wellbeing is gaining traction and acknowledgment across the globe [15]. Trying to improve economic growth (the market value of all goods and services produced by the New Zealand economy) is the traditional way a government runs a country and tracks its progress. In 2019, the New Zealand government administered the first ever 'wellbeing budget'. This came about after recognising that although New Zealand had sustained economic growth, there were several societal issues that needed attention. For instance, poor mental health is a significant problem in New Zealand, with evidence suggesting that in any given year, one in five New Zealanders will have a diagnosed mental illness [16]. Child poverty is another serious issue in New Zealand, with ~150,000 children in New Zealand living in households that experience material hardship [17]. The Wellbeing Budget was the government's way of refocusing on what matters for the success of the New Zealand people. Despite being representative of broad population groups (i.e., gender, ethnicity), the GSS is not representative of certain groups in the population that may be of high policy interest, such as individuals facing employment challenges. This means it is impractical to explore the determinants of wellbeing and how government policies are associated with wellbeing, in these smaller subsets of the population. To overcome this challenge, population-level wellbeing data, sensitive enough to capture the effects of policy change, are required. Therefore, the overarching aim of this research was to apply data science methodologies (such as machine learning algorithms like random forest) and explore the feasibility of predicting population-level subjective wellbeing outcomes using administrative data from the IDI. The specific objectives to achieve this overall aim were:

1. To systematically review existing literature focused on predicting population-level health wellbeing outcomes using machine learning techniques;
2. To explore the current wellbeing landscape in New Zealand across diverse demographic groups and regions;
3. To develop and validate models to predict GSS-based subjective wellbeing outcomes from administrative variables in the New Zealand census;
4. To extrapolate subjective wellbeing scores to the entire New Zealand population and evaluate these predictions using GSS trends as a reference.

**Thesis Structure**

This thesis includes four distinct publications adapted in chapter format, as seen in Figure 1-1. Chapter 2 sets the thesis context with a brief overview of wellbeing concepts, New Zealand's IDI, and current applications of data science in health, particularly within the New Zealand context using the IDI. Next, Chapter 3 is a systematic scoping review that summarises the use of machine learning to predict health and wellbeing outcomes, focusing on the utilisation of population-level administrative variables as predictors. Chapter 4 examines the recent wellbeing climate in New Zealand, utilising data sourced from the GSS. It explores the distribution of subjective wellbeing outcomes across sociodemographic groups and how these change over time. Chapter 5 describes the development and validation of machine learning models that can predict GSS-based subjective wellbeing outcomes using primarily census-based variables as predictors. These models are then used in Chapter 6 to extrapolate these predictions to the entire census population. Chapters 3 to 6 are either under review in a peer-reviewed journal or under submission. As these chapters were written as separate articles, repetition of some information (e.g., introduction and methodology) was unavoidable. Finally, Chapter 7 concludes by offering a summary of key findings from each study, outlining the limitations of the PhD research project and discussing the implications for future research.

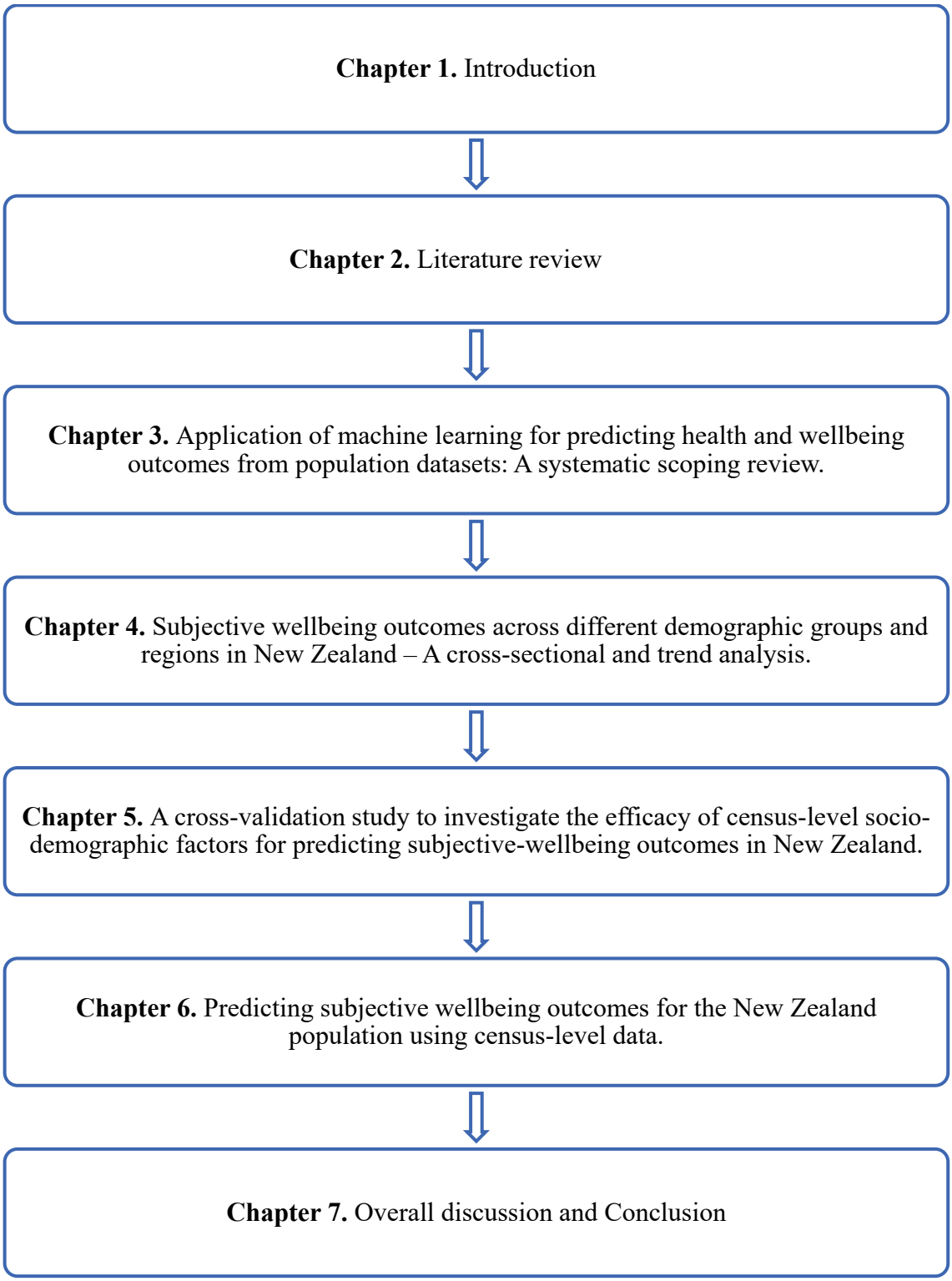



**Figure 1-1.** Structure of the thesis

# Chapter 2 – Literature review

## Preface

The conceptualisation of wellbeing has evolved over time. Concurrently, the advent of big data and the application of advanced data science methodologies are also gaining prominence. The aim of this chapter was to establish the thesis context in preparation for the subsequent research chapters, by conducting a brief review on (1) the concepts of wellbeing, (2) the availability of big data in New Zealand, and (3) the application of data science in public health. The first section reviews popular concepts of wellbeing and highlights the importance of population wellbeing. This section also discusses the factors that influence wellbeing and reviews New Zealand's wellbeing initiatives. The next section provides an overview of the Integrated Data Infrastructure (IDI) and highlights the opportunities that it presents for understanding population wellbeing. The final section briefly summarises the existing application of data science in health. This section concludes by exploring the current applications of data science methodologies in New Zealand using the IDI.

## Background

New Zealand has faced persistent challenges with poor economic and productivity growth over the last two decades [18]. Although, economic growth is a crucial driver of a country's standard of living, it does not guarantee improvements in its citizens' wellbeing [6]. Evidence suggests that among the 38 OECD countries, New Zealand is relatively "healthy" (i.e., high life expectancy, high self-reported health), but faces long-term challenges such as mental health crisis, child poverty, and high levels of suicide [19, 20]. Research has shown that many New Zealanders have low levels of wellbeing, with just 25% reporting high wellbeing [21]. A similar trend of low population wellbeing is also seen in other countries. For example, in the USA, only 20% reported high wellbeing [22], and 20 out of 22 European countries reported that 9–28% of their population had high wellbeing [23]. The World Health Organisation defines health as a "state of complete physical, mental and social wellbeing and not merely the absence of disease or infirmity" [24]. Wellbeing is a meaningful positive outcome which is a combination of how happy and satisfied one is with their life. It indicates "how people perceive their lives", and if they are able to lead it with purpose, balance and meaning [25]. Early researchers in the field studied wellbeing as a single-dimensional objective construct explained by a simple question — "How happy do you feel?" using an individual's objective life conditions such as income, education, age, and relationship status [26, 27]. However, this conceptualisation can be challenged as different individuals tend to react and feel differently (in similar life situations) due to the influence of individual factors such as life expectation and experience [28]. Given the subjectivity of an individual's outlook towards life, researchers began to consider wellbeing as a multi-dimensional subjective concept with various aspects and interactions [11].

**What is wellbeing?**

Early researchers in the field mainly recognised wellbeing as the essence of positive human functioning (referred to as "eudaimonia") [23]. For instance, Jahoda (in 1958) attributed positive functioning to six key elements – 'attitudes of an individual toward his own self', 'self-actualisation', 'integration', 'autonomy', 'perception of reality' and 'environmental mastery' [29]. Over the years, various theoretical frameworks have been developed that describe the eudaemonic perspective of wellbeing [30, 31]. Alternatively, some researchers in the field conceptualised wellbeing as a combination of both Eudaemonic (one's ability to function) and Hedonic (one's emotions/ feelings such as happiness, sense of enjoyment) elements [32]. In the past few decades, some key contributions have been made to model wellbeing and its dimensions. Firstly, the tripartite model of subjective wellbeing (Figure 2-1) proposed by Diener et al. [28] illustrated wellbeing as a combination of three main components – (1) cognitive component, (2) positive affect, and (3) negative affect. Secondly, Seligman and colleagues [33] presented their PERMA model of wellbeing that comprises five key dimensions – Positive emotion, Engagement, Relationships, Meaning, and Accomplishment.

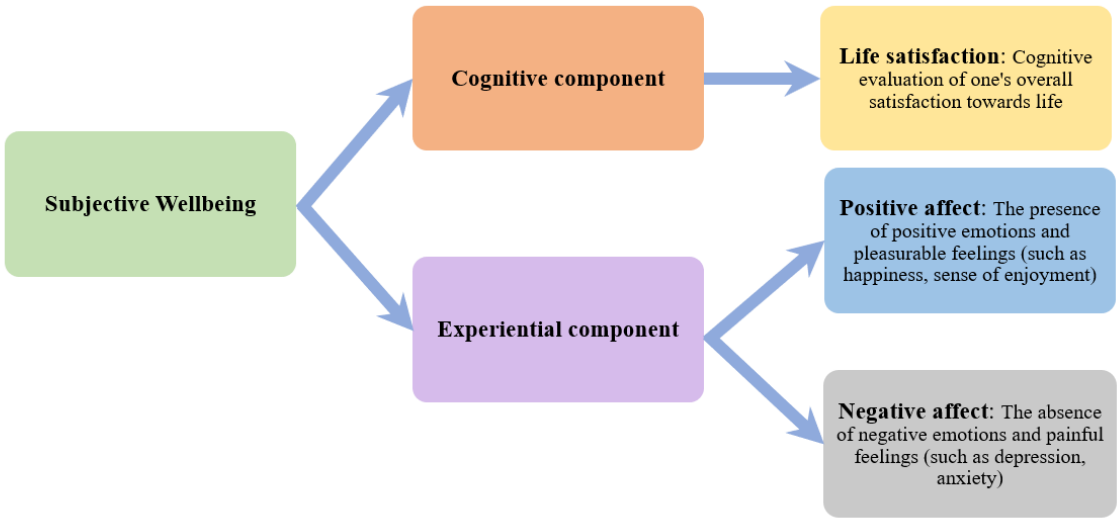



**Figure 2-1**. Tripartite model of subjective wellbeing proposed by Ed Diener (1999)

A high level of wellbeing (termed 'flourishing') is linked with positive feelings and emotions (such as cheerfulness, calmness, and satisfaction), and positive functioning (such as autonomy, competence, and engagement) [34]. Thompson et al. [35] developed a dynamic model of flourishing (see Figure 2-2). In this model, 'Flourishing' is illustrated as a construct which is a combination of one's positive feelings (block 1) and positive functioning (block 2). The model also depicts how one's external conditions (block 3) and their personal resources (block 4) enable them to function effectively (block 2) in their lives and thus experience positive emotions (block 1). When an individual (or a society) experiences positive emotion (both daily and overall) and functions well, it is presumed that they are 'flourishing'.

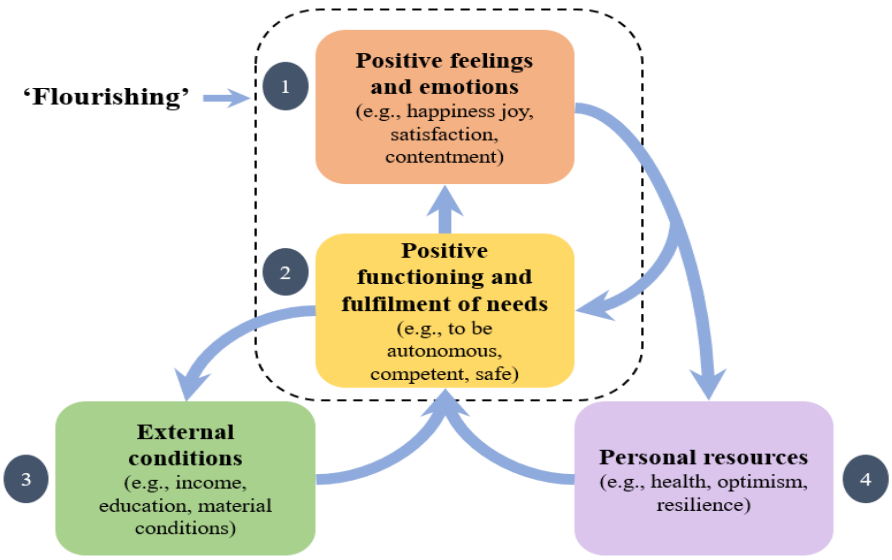



**Figure 2-2.** Dynamic model of Flourishing; proposed by Thompson et al. 2004.

While Diener's tripartite model and Seligman's PERMA model outlines wellbeing as a construct that includes both hedonic and eudaemonic aspects, Thompson's dynamic model also highlights the importance of various factors (external conditions and personal resources) that indicate and potentially influence one's wellbeing. In the context of this thesis, our focus is exclusively on subjective wellbeing indicators, such as life satisfaction.

**Why is wellbeing important?**

Population wellbeing can have a direct impact on a country's overall economic output, social, and health care costs [6]. Individuals with high levels of wellbeing are likely to have positive health outcomes [36], be more productive and have stronger social relationships [13, 37] (both at an individual and societal level). While the link between physical and mental health is well established, researchers have generally tended to study the negative effects of ill-being states (such as depression and anxiety) on physical health, as opposed to the contrary (i.e., impact of high wellbeing on physical health) [36, 38]. However, recent studies have shown that positive mental states (e.g., happiness) can positively influence physiological health [39], and individuals with high levels of wellbeing are less likely to develop cardiovascular diseases [36]

and hypertension [40]. A study conducted by Cohen et al. [38] also revealed that individuals with higher levels of positive emotion had increased resistance against infections (such as colds), lower levels of stress hormones and less reported physiological symptoms (such as aches and pain) than those experiencing a higher level of negative emotion. Given these benefits, there is growing international interest in the measurement and promotion of population wellbeing [15, 41], and understanding what factors are associated with high or low wellbeing. While measuring population wellbeing is not an easy task, a deeper understanding of the association between mental and physical states could be a step towards a healthier and a more flourishing nation.

Over the years, the measurement of individual wellbeing has mainly focussed on mental ill-being or a single score for life satisfaction. There is a substantial body of work exploring the patterns, causes, and effects of mental ill-being [42]. Despite the absence of mental illness, there are several factors (as discussed in the previous section) that can influence the degree to which one is experiencing wellbeing or flourishing. Considering the multi-dimensional construct of wellbeing, traditional wellbeing assessment approaches are inadequate to represent the full range of a population's wellbeing [43]. Recent advancements in the scientific measurement of wellbeing, including improved methodologies and metrics, led to the development of the *Better Life Initiative* in 2011 by the Organisation for Economic Co-operation and Development (OECD) [44]. This initiative was focussed on creating 'best practice' guidelines in the measurement of wellbeing. Figure 2-3 depicts the wellbeing framework developed by OECD as part of the initiative. OECD's wellbeing framework is based on the 'capabilities approach' proposed by Amartya Sen, a pioneer economist [74]. The approach underlines that a country's economic growth should be aimed at expanding the capabilities of its citizens to lead the types of lives they value and have reason to value. This can be achieved by growing various capital stocks available in a country such as human capital

(education, skills), social capital (networks, relationships), natural capital (environmental resources), and economic capital (financial resources, investments). For example, improving education and healthcare systems can enhance human capital, while investing in sustainable infrastructure and environmental conservation can enhance physical and natural capital. By enhancing these capital stocks, a country can provide its population with better access to essential services like healthcare, education, and transportation. This, in turn, expands the capabilities of different groups within the population to utilize these resources to improve their quality of life and achieve their personal and collective goals [74]. Considering this aspect, the OECD's wellbeing framework [44] highlights the importance of measuring one's current and future wellbeing. Current wellbeing is measured across 11 dimensions which can be grouped into two categories – material living conditions (e.g., income and wealth, jobs and earnings) and quality of life (e.g., work-life balance, life satisfaction). Future wellbeing is assessed across four resources – natural, economic, human, and social capital. These resources are constantly affected by today's actions and are essential to sustain future wellbeing [45].

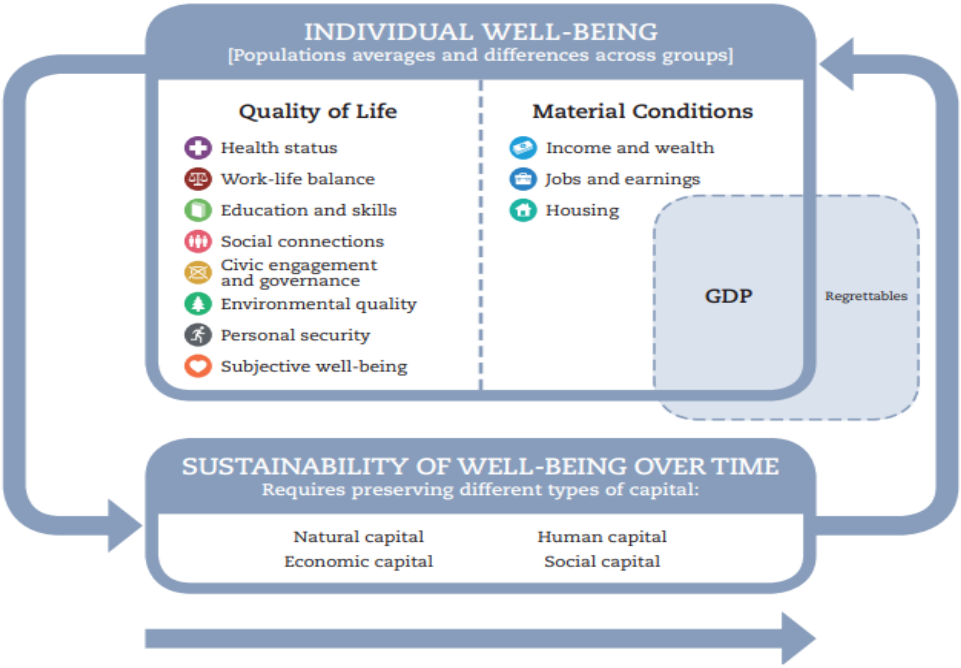



**Figure 2-3**. OECD's wellbeing framework

The OECD guidelines aim to bring uniformity and hence comparability to wellbeing measurement across different countries [43]. These guidelines also serve the purpose of guiding public policy and monitoring progress across nations. Considering these guidelines, the New Zealand government followed a 'wellbeing approach' to administer its first ever Wellbeing Budget in 2019 [46]. Traditionally, decisions about budget allocations and initiatives have been based on the country's fiscal and economic position. Contrastingly, in the Wellbeing Budget, the government made decisions and developed initiatives targeted at improving the wellbeing of New Zealanders. For example, the five key areas identified for improvement in the 2019 budget were: (1) Supporting mental wellbeing for all New Zealanders, (2) Improving child wellbeing, (3) Improving Māori and Pasifika incomes, skills and opportunities, (4) Building a productive nation through innovation, social and economic opportunities, and (5) Transforming the Economy. Evidence from the New Zealand Treasury's Living Standards Framework [10] is primarily used to identify and prioritise the areas of wellbeing where New Zealand could improve.

**New Zealand's Living Standards Framework & General Social Survey (GSS)**

The New Zealand Living Standards Framework (developed by New Zealand Treasury since 2011) was specifically designed based on the OECD wellbeing framework. It is used to monitor and evaluate the impacts of public policy on the lives of New Zealanders by assessing current wellbeing across 12 domains: health, housing, income and consumption, jobs and earnings, time use (leisure and free time), knowledge and skills, safety and security, social connections, cultural identity, civic engagement and governance, environmental quality, and subjective wellbeing [10]. Each wellbeing domain is assessed using indicators that reflect our understanding of how New Zealanders experience wellbeing. These indicators are acquired from government data sources and surveys. For example, the New Zealand General Social

Survey (GSS) [47] holds information about the majority of these indicators. Table 2-1 shows the wellbeing domains and some of the key indicators available in the GSS.

**Table 2-1**. GSS wellbeing domains.

| Wellbeing domain | Key indicators |
|---|---|
| Health | Self-reported health status |
| Housing | Housing suitability, Coldness in house |
| Income and consumption | Household income |
| Jobs and earnings | Job satisfaction |
| Time use | Availability of leisure-time |
| Knowledge and skills | Highest qualification |
| Subjective Wellbeing | Life satisfaction, Sense of purpose, Family-wellbeing |
| Safety and security | Safety feeling |
| Social connections | Loneliness |
| Cultural identity | Ability to be yourself |
| Civic engagement and governance | Trust in people and government organisations |
| Environmental quality | Problems in the natural environment (e.g., air pollution, water pollution) |

The GSS is one of the three household surveys conducted by Stats NZ (a government agency) once every two years. First conduced in 2008, the GSS provides information on the wellbeing of New Zealanders aged 15 years and over. It encompasses a range of social and economic indicators and can be used to show how wellbeing outcomes vary across different groups within the population. Specifically, it includes objective information about circumstance, such as labour force status and income, as well as a personal assessment of different life aspects, such as life satisfaction, health, housing, human rights, and relationships. The GSS data are collected at both an individual and household level through questionnaires administered using computer-assisted personal interviews, employing a three-stage sample selection method [48]. Firstly, primary sample units (i.e., households) that represent the population are randomly chosen from the Household Survey Frame. Next, eligible households are selected to complete the household

questionnaire. Lastly, one individual is randomly selected from each eligible household to complete the individual/personal questionnaire. The total number of individuals generally recruited to take part in the GSS is between 8,000 and 10,000. The data acquired from each survey respondent are linked to New Zealand's IDI (described in detail below).

**Factors that influence wellbeing**

There are various factors that can influence wellbeing. These factors can be broadly classified into four categories: (1) personality traits (including genetics), (2) socio-demographic factors, (3) behavioural factors, and (4) environmental factors. The study of personality traits is a highly researched area in human psychology. The 'five-traits' model [49] attributes five key traits to an individual's personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Studies have shown that these personality traits are influenced (equally) by one's genetic inheritance and their social environment (e.g., family relationships). In the context of wellbeing, research studies have explored the construct of subjective wellbeing and its association with various personality traits. For instance, a study conducted by Steel et al. [50] found that personality traits (neuroticism, extraversion, agreeableness, and conscientiousness) had significant correlations with all aspects of subjective wellbeing (e.g., positive and negative affect, happiness, life satisfaction, and quality of life).

Socio-demographic factors are also major determinants of one's health and wellbeing [51-53]. These factors include age, gender, ethnicity, socio-economic status, education (e.g., highest qualification), employment (e.g., employment status, number of jobs, work hours, personal income), and household factors (e.g., number of people in the household) [54, 55]. Next, behavioural factors such as physical activity, dietary choices, sleep patterns, and substance use also play an important role in shaping an individual's wellbeing. Regular physical activity, such as walking or engaging in sports, contributes significantly to overall mental and physical health, promoting positive mood and reducing the risk of chronic diseases [56]. In contrast, sedentary

behaviour can worsen health issues and negatively impact wellbeing [57]. Similarly, a balanced diet rich in nutrients supports cognitive function and emotional stability [58]. Additionally, consistent sleep and the avoidance of harmful substances like tobacco and excessive alcohol are essential for maintaining optimal health and preventing mental distress [59]. These behavioural factors are closely linked to health-promoting environments, which can further enhance an individual's wellbeing across the life span [60]. For example, an individual's physical activity level (e.g., time spent walking) and food choices (e.g., what food is consumed, and where it is obtained) are some of the key health behaviours that are affected by environmental conditions. These environmental factors include accessibility to recreational facilities, transportation systems, and access to healthy food. On the other hand, environmental factors such as air and noise pollution can also have a direct impact on wellbeing [61, 62]. A recent review [63] revealed that neighbourhood quality, green space (such as parks), land-use mix, industry activity, and traffic volume were linked with psychological distress. Similar studies have shown that people living in areas with green or blue spaces (e.g., parks or ocean) can have increased mental wellbeing and are less likely to experience psychological disorders [64-66]. More recently, Hobbs et al. [67] developed a "Healthy Location Index" (HLI; scored between 1–10) for each meshblock (smallest geographical unit that consists of ~30–60 households [68]) in New Zealand based on the accessibility to various features of the environment (such as access to green space, blue space, physical activity facilities, and fast-food outlets). This study revealed that unhealthy environments were associated with adverse mental health and psychological distress in those aged 15+ years.

**Integrated data infrastructure**

The potential measurement of wellbeing across multiple domains, at the population level, has only become possible due to the existence of a rich research database in New Zealand – the Integrated Data Infrastructure (IDI). Established in 2011, the IDI is a complex data resource

managed by Stats NZ that holds microdata about people and households. All data in the IDI are stored as tables in a Structured query language (SQL) database. The IDI data captures all individuals' (who have ever been NZ residents) interaction with government agencies across various domains such as education and training, income and work, benefits and social services, health, justice, and housing [69]. Data in the IDI are also linked (through tax data) to the Longitudinal Business Database (LBD) – another large database which holds deidentified microdata about businesses. This allows researchers to explore the links between New Zealand people and business. Figure 2-4 shows the basic data links in the IDI.

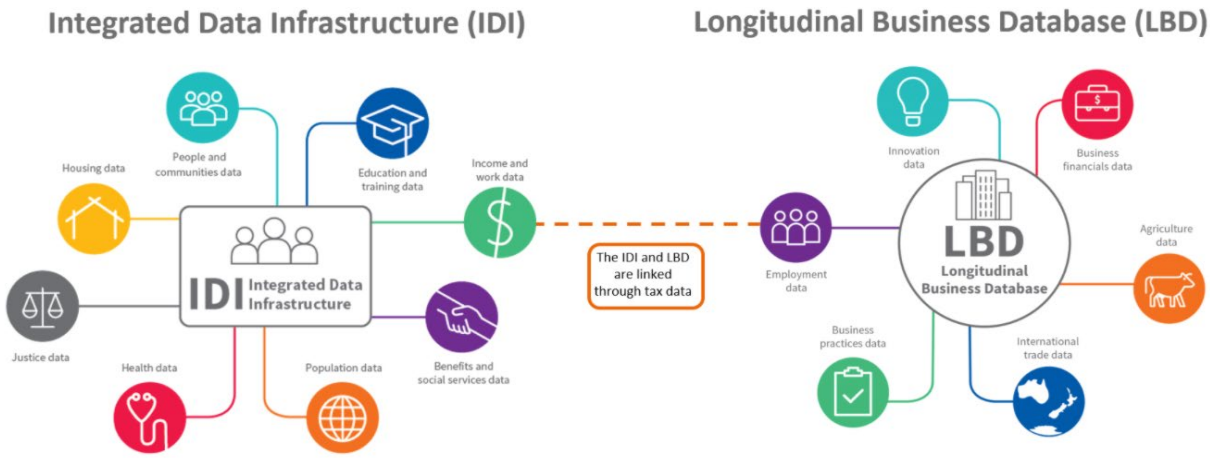



**Figure 2-4.** Data links in the IDI

The IDI comprises of two main components – (1) spine and (2) nodes. The 'spine' is the central dataset in the IDI that holds information for all people ever resident in New Zealand, including citizens, permanent residents, people with temporary visas that allow them to reside, work, or study in New Zealand, and those who live and work in New Zealand without requiring a formal visa [70]. The spine dataset is constructed by compiling data from three major sources – (1) Inland Revenue (tax) data from 1999 onwards, (2) births from 1920 onwards, and (3) New Zealand Immigration visa data from 1997 onwards. It is estimated that the IDI spine contains information for approximately 10 million individuals. The IDI 'nodes' on the other hand, are

individual datasets which are a collection of New Zealand's whole-population administrative data collected by various government agencies. This includes the census and several other questionnaire-based social and socioeconomic surveys from samples of the population. The IDI is generally refreshed up to four times every year. Upon every refresh, several new datasets are added, and existing datasets are updated. All individual datasets (aka nodes) are ultimately linked to the IDI spine using record linkage techniques (such as unique identifier linking and probabilistic linking) [71]. A sample schematic of the IDI structure is shown in Figure 2-5.

Data within the individual 'nodal' datasets may be at the individual, household, or meshblock level. To ensure confidentiality and privacy protection, all individuals and households in the IDI are identified using unique anonymous identifiers, instead of personal information (e.g., all names, addresses, and government agency IDs are removed). In September 2018, there were a total of 166 billion pieces of information in the IDI [72].

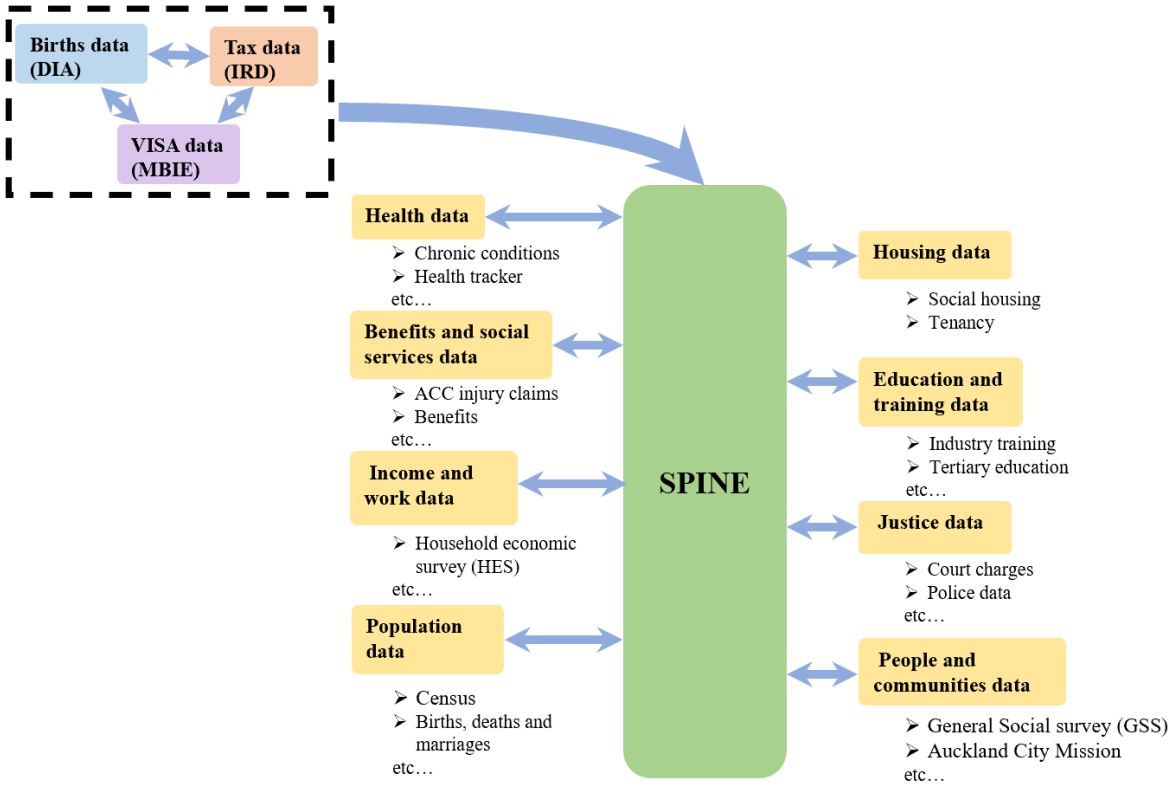



**Figure 2-5**. Schematic representation of the IDI spine and nodes

Detailed measures of subjective wellbeing are only available for a smaller subset of participants within the IDI (via the GSS). Consequently, our understanding of the various factors influencing wellbeing is limited, primarily due to the lack of detailed population-level subjective wellbeing data. Exploring innovative data science and predictive modelling methodologies presents a potential solution to bridge this gap.

**Data science in health**

Data science is a distinct field of study aimed at extracting actionable insights and information (in the form of predictions, automated decisions, and visualisations) from data [73]. These data may be simple or complex, large or small, static or dynamic, structured or unstructured [74]. One of the major advantages of data science (when compared to traditional inferential statistical methods) is the ability to not just discover interesting patterns in the data (that helps to explain the past), but instead, utilise these patterns to make predictions about the future [75]. While inferential statistics are also capable of making predictions, data science methods often excel in handling complex data structures and capturing nonlinear relationships, leading to more robust and accurate predictions in many cases [76, 77]. Data science is an interdisciplinary field that comprises several key domains such as artificial intelligence (using computers to perform tasks that generally requires human intelligence) [78], database management systems, visual analytics (interpreting and analysing data through visualisations) [79], data mining (detection of patterns in big and complex datasets) [80], process mining (using event-data to discover, monitor, and improve processes) [81], and behavioural science (using data to analyse and examine human behaviours) [82]. In the context of this thesis, artificial intelligence (specifically, machine learning) is further explored.

Machine learning is an amalgamation of the technical fields of computer science and statistics, and is one of the integral areas of data science and artificial intelligence [74, 83]. It is currently one of the most rapidly evolving fields around the world. The term “machine learning” refers

to a machine's (i.e., a computer's) ability to make data-driven predictions or decisions by understanding patterns in the input data with no explicit programming [74]. Machine learning algorithms are broadly classified into three types: (1) supervised, (2) unsupervised, and (3) semi-supervised learning. Supervised machine learning is focused on creating prediction models that are developed (or trained) using algorithms that learn to map a set of input data to an outcome measure. The model building process is automated through recursive learning of the input data [84]. Some commonly used supervised machine learning techniques are the support vector machine, random forest, and neural network [14]. On the other hand, unsupervised learning employs mathematical techniques to create clusters or groupings in unlabelled input data (e.g., identifying topics of discussion in social media) [85]. A commonly used unsupervised machine learning technique is k-means clustering [86]. Finally, semi-supervised learning techniques are used to develop models based on a combination of both labelled and unlabelled data [87]. These methods are generally helpful when labelled datasets are limited or expensive. Given there are several types of machine learning techniques, there is no one technique that is best suited for every problem. Therefore, it has been recommended to utilise a range of techniques based on study-specific variables such as the type of input data and the outcome of interest [75].

Recent advancements in machine learning have become possible due to developments in innovative learning algorithms and a global upsurge in the availability and accessibility of data [88, 89]. The application of machine learning is widespread across multiple areas such as computer vision, speech recognition, and natural language processing [83]. In health research, machine learning has been used in biomedicine, bioinformatics, clinical informatics, imaging, and health care [90]. However, the application of machine learning to understanding population health and wellbeing is likely limited – with only one review published on this topic [91]. Monitoring population health is complicated due to the scarcity of regular, timely, and precise

population data [51]. Furthermore, the complexity of population health datasets makes it challenging for researchers to answer key questions related to health and wellbeing (e.g., what are its drivers? where does it deteriorate? and how does it change over time?) [92]. These datasets are complex due to their large size, the variety of data sources, and the diverse types of data they contain, such as medical records, demographic information, and behavioural data. It is posited that machine learning can help to overcome these challenges by predicting precise health outcomes from regularly updated government administrative data. In 2015, Luo and colleagues [51] successfully examined the feasibility of a machine-learning model to predict the prevalence of six non-communicable disease (NCD) outcomes (four NCDs and two major clinical risk factors), based on population socio-demographic characteristics that are widely available and regularly updated through the national census and community surveys. Likewise, another study in the US used machine learning to successfully predict mortality risk among 2,066 residents using variables such as age, gender, self-reported health, physical activity score, smoking history, and chronic health conditions [93].

The availability of linked population datasets in New Zealand (i.e., the IDI) presents significant opportunity to model and understand population health. A recent review revealed that there is considerable scope for the application of machine leaning in the area of mental health, given the enormous volume of data collected over the last decade [94]. However, out of 568 projects that have used IDI data nationally (documented by Stats NZ as of April 2021) only three have applied machine learning to predict health outcomes (cardiovascular disease events and diabetes complications risk) using variables from the administrative datasets [95]. None of these studies have explored population wellbeing. Clearly, there is tremendous scope for the application of machine learning to better understand population health and wellbeing from the IDI. The body of work proposed in this PhD project focuses on exploring, predicting, and modelling wellbeing outcomes in New Zealand, from the IDI. It is anticipated that the findings

will enable government and researchers to make better use of the IDI when evaluating population wellbeing.

# Chapter 3 – Application of machine learning for predicting health and wellbeing outcomes from population datasets: A systematic scoping review.

## Preface

In Chapter 2, the literature review underscored the importance of population wellbeing, highlighted opportunities within New Zealand's Integrated Data Infrastructure (IDI) and showcased the growing potential of data science within the field of public health. Building on this groundwork, this chapter – a systematic scoping review – aimed to summarise studies that have used machine learning to predict health and wellbeing outcomes, specifically utilising population-level administrative variables as predictors. The findings are expected to guide methodological decisions used throughout this doctoral project.

Note: To improve the flow and clarity of the chapter, results are integrated with the discussion. This approach facilitates a more cohesive interpretation of the findings by aligning data with their contextual analysis.

## Introduction

In the field of public health, data-driven approaches play a crucial role in shaping health policy by guiding interventions and providing insights into population health and wellbeing [96]. The wealth of data generated from diverse sources like administrative records, electronic health records, medical imaging, and wearable devices (e.g., activity trackers) is essential for understanding health trends and outcomes within and across populations [97-100].

The prevalence of "big data" has opened a new era of opportunities in public health [101]. As technology advances and global connectivity increases, every field of study has an ocean of information waiting to be explored. Big data's role in deciphering health trends, forecasting disease outbreaks, and shaping public health strategies is becoming increasingly vital. Governments globally are harnessing the potential of big data for research, policy formulation, and decision-making [102]. For example, New Zealand's Integrated Data Infrastructure (IDI) is one such initiative moving in this direction [103].

Traditionally, public health researchers have relied on statistical models to understand associations and predict outcomes. While these models yield valuable insights, they have certain limitations, especially when dealing with a multitude of variables and complex data structures [76]. The more factors, or covariates involved, the reliability and generalisability of traditional statistical models can be questionable. The interpretation of results also becomes challenging, as these models may struggle to accommodate the intricate interplay of variables [104]. Although traditional models can also capture variable interactions by incorporating interaction terms and other techniques, they are most effective on smaller datasets with linear relationships [76]. Their limitations become more apparent when applied to larger and more complex datasets.

To overcome this challenge, researchers have explored advanced statistical tools such as machine learning [104]. Machine learning is an amalgamation of the technical fields of computer science and statistics and is one of the integral areas of data science and artificial intelligence [14]. The evolution of machine learning has brought about sophisticated analytical tools capable of processing complex, multidimensional data. For instance, machine learning algorithms like neural networks and support vector machines can analyse medical imaging data to detect early signs of diseases such as cancer with high accuracy [105]. Additionally, natural language processing (NLP) techniques can be used to extract valuable insights from unstructured clinical notes, enabling better patient management [106]. Predictive models, such as those using random forests and gradient boosting machines, can forecast patient outcomes [107] based on a combination of genetic, environmental, and lifestyle factors, allowing for tailored interventions and preventive measures. These advancements not only promise to enhance predictive accuracy in healthcare but also pave the way for more informed decision-making and personalised treatment strategies [108].

In the past decade, applications of machine learning on administrative data have opened new areas of research [109, 110]. Administrative data, collected primarily for administrative purposes by government agencies, healthcare providers, and other organisations, includes records such as demographics, birth and death registrations, and healthcare service use records [111]. These data, often covering large segments of the population, provides a comprehensive, longitudinal view of various health-related aspects and social determinants. It is a cost-effective, scalable option for large-scale studies, making it an asset for public health research [98].

The availability of extensive administrative datasets worldwide has opened interesting opportunities in public health research, as the results are easily reproducible and comparable across countries [112]. This emerging trend has seen a surge in studies applying data science

techniques on vast administrative datasets to predict population-level health outcomes, accompanied by an expansive array of different methodologies, reflecting the diversity and innovation within the field of public health research [91]. Despite this rapid progress, a comprehensive synthesis of the existing evidence (focused on administrative data) remains a noticeable gap in the literature.

Therefore, the aim of this scoping review was to deliver a comprehensive summary of methodologies in studies that use machine learning techniques, to predict health and wellbeing outcomes using population-level administrative variables. Explicitly, this review: (1) identified and categorised the machine learning techniques applied when predicting health outcomes from population datasets; (2) evaluated the diversity of health outcomes predicted through these techniques; and (3) assessed the methodologies, dataset characteristics, and predictors utilised in these models.

The results of this systematic scoping review are expected to guide methodological decisions in studies aiming to implement predictive modelling within the field of health and wellbeing. It is anticipated that the findings will contribute to best practices in the field, underscoring the transformative potential of machine learning in public health.

**Methods**

This systematic scoping review was conducted by searching the Scopus, Web of Science, and various other databases (such as MathSciNet, MEDLINE, SocINDEX and SPORTDiscus) using the EBSCO search engine. We used a combination of terms related to “administrative data”, “health outcomes”, and “machine learning” to capture relevant literature. The following terms were searched for in abstracts, titles, and keywords:

*[("administrative data" OR "health data" OR "census" OR "population level*" OR "population-level*" OR "nationally representative" OR "nationally-representative" OR "socio-demographic data" OR "socio-economic data")*

*AND ("health outcome*" OR "health status" OR "wellbeing" OR "wellbeing" OR "disease prediction" OR "disease burden" OR "*score*" OR "risk factor*" OR "risk prediction")*

*AND ("machine learning" OR "artificial intelligence" OR "data mining" OR "feature extraction" OR "feature selection" OR "supervised learning" OR "neural network*" OR "random forest" OR "deep learning" OR "unsupervised learning" OR "K-means" OR clustering OR "cluster analysis")].*

The search was limited to peer-reviewed, English-language journal articles published in the last 10 years as of November 2, 2021; conference presentations and grey literature were excluded. The reference lists of all included studies were also screened for any additional papers meeting the inclusion criteria. The initial screening process involved reviewing titles and abstracts for relevance, followed by a full-text assessment of potentially eligible studies by two independent reviewers. Any disagreements in study selection were resolved through discussion to ensure a comprehensive and unbiased collection of relevant studies. The review was written in alignment with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) protocol [113], ensuring adherence to pre-specified objectives and methodologies. The review protocol was registered under the International Prospective Register of Systematic Reviews (PROSPERO) [114] (Registration ID: CRD42021270690).

The eligibility criteria for included studies were focused on the utilisation of machine learning techniques to predict health and wellbeing outcomes from nationally representative administrative datasets. Studies that predicted physiological health outcomes (e.g., risk of a

disease) were also included. Studies that used clinical variables (e.g., laboratory values, blood cell count, cholesterol level) to train machine learning models were excluded. Next, studies that were focused on specific sub-populations. (e.g., people already with a disease) were excluded. The exclusion criteria extended to studies predicting non-physiological health outcomes, such as risk of inpatient falling, risk of hospitalisations or hospital re-admission, patient experience, or need for health care services. These exclusion criteria were applied due to the scope of the review; a focus on exploring broader health outcomes within a national context, rather than clinical settings. This approach also ensured that findings of this review were relevant and generalisable to the broader population.

Data extraction focused on key aspects such as the publication year, geographic location, data source, study period, population characteristics, outcome variables, number and type of predictor variables, and prediction scale (i.e., individual or area level). These data provided a foundation for synthesising information across studies. Additionally, we evaluated the data pre-processing techniques, types of predictors (i.e., independent variables) used, machine learning models and software employed, and methods for evaluating model performance.

A critical component of our review was the appraisal of individual studies. Each study was critically assessed for methodological quality using an adapted version of the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist [115], presented in Table 3-1. This appraisal focused on evaluating the study design, data quality, machine learning model development and validation, outcome measurement, and statistical analysis methods. This critical appraisal was vital in assessing the quality and rigor of the included studies, providing insights into the strengths and limitations of the current literature. It is important to note that the criteria used in this assessment are tailored exclusively to evaluate methodological quality; therefore, caution must be exercised in interpreting the results, considering the specific adaptations made for this quality appraisal. Throughout the

review process, we maintained an awareness of potential biases or perspectives influencing our approach. This reflexivity ensured a balanced and objective methodology, from study selection to data analysis.

**Table 3-1**. Methodological Quality Evaluation Checklist for Reviewed Studies (Adapted from TRIPOD)

| Criteria | Description |
|---|---|
| ***Data*** | Did the study distinctly outline the design or data sources for both the development and validation datasets, and did it specify the dates when participant data was collected? |
| ***Participants*** | Did the study specify key elements of the study setting, including the number and location of centres, and describe eligibility criteria for participants along with details of any treatments received? |
| ***Data preparation*** | Did the study describe any data pre-processing steps, encompassing cleaning, feature engineering, sampling, and measures taken for data quality? |
| ***Outcome*** | Did the study clearly define the predicted outcome, detailing how and when it was assessed? Additionally, were qualifications of outcome assessors and measures for inter- or intra-rater variability described? |
| ***Predictors*** | Did the study describe the choice and definition of initial predictors and mention any methods to address discrepancies in predictor measurement? |
| ***Class imbalance*** | Was there an explicit statement on how class imbalance was addressed? |
| ***Sample size*** | Was there an explanation of how the study size was determined, including any sample size calculations? |
| ***Missing data*** | Did the study describe how missing data were handled and provide reasons for omitting any data? |
| ***Analytical methods*** | • Did the study describe how the data were used in the analysis, including any division into training and testing sets?<br>• Did the study clearly outline the chosen model, key model-building steps (e.g., feature selection, engineering, hyperparameter tuning), validation or updating methods, measures for assessing performance, and the process for selecting the final model?<br>• Was there a report on any assessment of model transportability or generalisability, and were any sensitivity analyses or alternative modelling strategies performed? |
| ***Model Output and Risk Groups*** | Did the study clearly define and report the model output, providing comprehensive details on any risk groups or classifications generated by the prediction model? |
| ***Validation*** | Did the study transparently and systematically validate models, specifying the validation type, employing appropriate dataset splitting techniques, and utilising well-defined performance metrics? |

## Results

A total of 3,005 articles (Scopus, Web of Science, EBSCO) were identified through the database search, and 20 additional articles were included from the reference list of the identified articles. Of the 3,025 articles, 1,188 duplicates were removed, leading to 1,837 potentially

eligible articles whose titles and abstracts were screened. Of these, 1,682 were excluded as they were deemed irrelevant based on the established exclusion criteria, such as the use of clinical variables as predictors, a focus on specific sub-populations, or non-physiological health outcomes. The full-texts of the remaining 155 articles were then assessed for inclusion. A total of 126 studies were excluded with reasons: Did not employ machine learning (e.g., used only regression modelling; $n$ = 55), used clinical variables as predictors ($n$ = 30), clinical study population ($n$ =11), outcome not related to health or wellbeing ($n$ = 9). Additionally, 21 articles were deemed irrelevant for other reasons, such as being reviews or focusing on broader health outcomes like hospital re-admission, pregnancy-related issues, or surgical site infections. A final list of 25 studies were included in the review (see Figure 3-1). A detailed description of all the included studies is presented in Table 3-2, which includes information such as reference, year, location, dataset/survey name, duration of data usage, sample description, outcome variables, number of predictors, scale of prediction (individual or area), machine learning algorithms used, and validation methods. Next, Table 3-3 details the number and types of predictor variables employed across these studies.

Our quality assessment of the reviewed studies revealed varying degrees of adherence to the TRIPOD checklist. The results are summarised in the Appendix Table A-1, providing a comprehensive overview of the presence (Yes), partial presence (Partially), or absence (No) of key methodological criteria across each study. Notably, certain studies were specifically designed to predict area-level outcomes, leading to some categories such as missing data, imbalance, or participants being deemed "NA - Not Applicable". Most studies exhibited robust compliance in fundamental categories such as 'Data,' 'Participants,' 'Outcome,' 'Predictors,' and 'Analytical Methods.' However, 'Class Imbalance' and 'Missing Data' categories were less consistently addressed, with several papers either partially discussing or omitting these aspects. Among the 25 studies, only eight tackled the class imbalance issue, with three partially

addressing it and another six not addressing it. Next, 70% of the studies ($n$ = 17) addressed missing data, while 20% ($n$ = 5) did not.

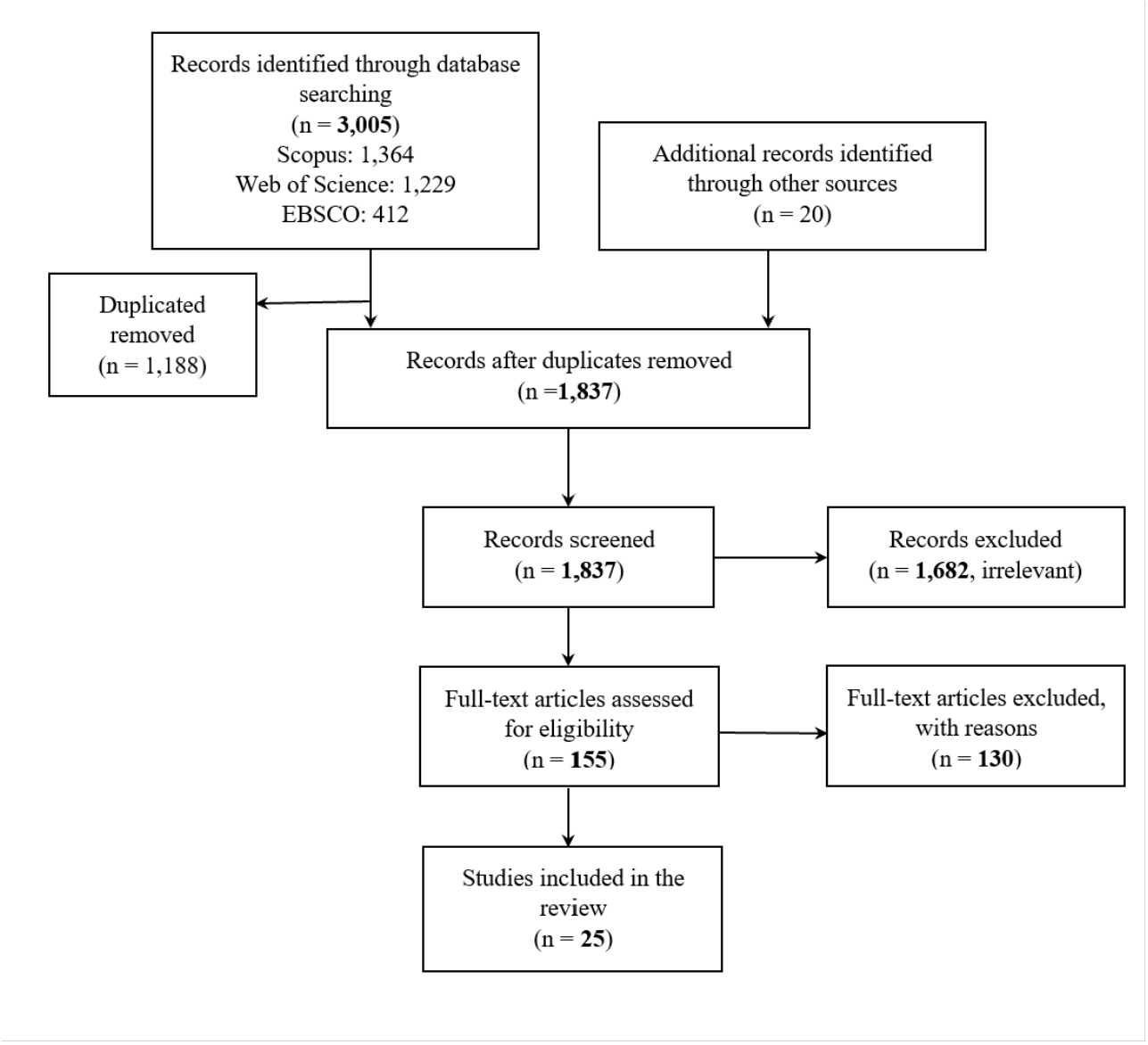



**Figure 3-1.** PRISMA flow diagram.

**Table 3-2.** Description of the studies.

| Ref | Year | Place | Name of the population dataset/survey | How many year(s) data was used? | Sample description | Population type | Outcome variable(s) | Type of outcome variable | Number of predictors | Scale of prediction (Individual or Area) | Machine learning algorithm used | Validation methods |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [116] | 2015 | USA | American Community Survey,<br><br>Behavioural Risk Factor Surveillance System (BRFSS) | 5 years,<br><br>2007 – 2012 | N/A | Adults | 6<br><br>Risk of high blood pressure, Risk of Obesity, Prevalence of: CVD – Angina, Heart attack, Stroke & Diabetes | Physiological health outcome | 7 | Area level, State level | Stepwise regression,<br><br>Lasso regression,<br><br>Random Forest,<br><br>Gaussian regression | Out of sample (Split dataset) validation<br><br>Training data: 2007–2010 dataset,<br><br>Validation data: 2011–2012 dataset |
| [117] | 2018 | Korea | Korea National Health & Nutrition Examination Survey | 5 years,<br><br>2007 – 2012 | 11,628<br><br>Aged 19 and above | Adults | 1<br><br>Suicide Ideation (Yes/no) | Mental health outcome | 15 | Individual level | Random forest | 10-fold CV<br><br>Data split (Training dataset = 90%, Test dataset = 10%) |
| [118] | 2019 | England | English Longitudinal Study of Aging | 10 years (6 waves),<br><br>2002 – 2012 | 6,209<br><br>age <65 (n=1,585),<br><br>age 65–79 (n=3,267),<br><br>and age >80 (n=1,357) | Older adults | 1<br><br>Health status (0-100) | General health outcome | 13 | Individual level | Random forest,<br><br>Linear regression model,<br><br>Deep learning | 10-fold CV |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [119] | 2019 | South Korea | Korean Young Risk Behaviour Web-based Survey | 1 year,<br><br>2017 | 15,294<br><br>Aged 12 – 18 years. | Young adults | 2<br><br>Suicide ideation (Yes/No), Suicide attempt (Yes/No) | Mental health outcome | 25 | Individual level | Logistic regression,<br><br>Random Forest, Support vector machine,<br><br>Artificial neural network,<br><br>Extreme gradient boosting | 5-fold CV<br><br>Data split (Training dataset = 80%, Test dataset = 20%) |
| [120] | 2019 | USA | National Health Interview Survey | 20 years,<br><br>1997 – 2016 | 583,770<br><br>Aged 18 – 85 years | Adults and older adults | 1<br><br>Risk score of colorectal cancer (Score between 0 and 1) | Physiological health outcome | 18 | Individual | Artificial Neural Network | 10-fold CV,<br><br>Out of sample testing on 2017 NHIS dataset |
| [121] | 2019 | USA | Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) data set | 8 years,<br><br>1993 – 2001 | 64,739<br><br>Aged 50 – 78 years | Older adults (women) | 1<br><br>Risk of breast cancer | Physiological Health outcome | 13 | Individual | Logistic regression,<br><br>Gaussian naive Bayes,<br><br>Decision tree,<br><br>Linear discriminant analysis,<br><br>Support vector machine, and | 10 x 10-fold CV |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Feed-forward artificial neural network | |
| [122] | 2020 | USA | US Health and Retirement Study dataset | 16 years,<br><br>1992 – 2008 | 13,611<br><br>Aged 52 – 104 years | Older adults | 1<br><br>Risk of mortality | General health outcome | 57 | Individual | Lasso regression,<br><br>Random Forest | Cross-validation |
| [123] | 2020 | Senegal | Demographic and Health Surveys | 1 year,<br><br>2015 – 2016 | 73 (Malaria dataset),<br><br>2,597 (Anaemia dataset)<br><br>Aged 15 – 59 years | Adults | 2<br><br>Risk of Malaria,<br><br>Risk of anaemia | Physiological health outcome | 46 (Malaria)<br><br>53 (Anaemia) | Individual | Random forest,<br><br>Support vector machine,<br><br>Artificial neural network,<br><br>K-Nearest Neighbours,<br><br>Naive Bayes | Data split (Training data = 80%, Test data = 20%) |
| [124] | 2020 | USA | National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), Community-level data from government agencies. | 2 years,<br><br>2001 – 2002,<br><br>2004 – 2005 | 34,563<br><br>Aged 18 and above | Adults | 1<br><br>Risk of PTSD | Physiological health outcome | 134 | Individual | Bayesian Additive Regression Trees (BART),<br><br>Penalised logistic regression,<br><br>Classification trees | 10-fold CV (for BART) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [125] | 2020 | England | English Longitudinal Study of Ageing (ELSA) dataset | 12 years (7 waves), 2002 – 2015 | 69,478<br><br>Aged 50 and above | Older adults | 2<br><br>Two measures of loneliness | Mental health outcome | 75 | Individual | XGBoost,<br><br>LightGBM,<br><br>Logistic Regression | Data split into test, validation and training datasets<br><br>Test = 1 wave (2014/15)<br><br>Validation = 1 wave (2012/13)<br><br>Training = 5 waves (2002 - 2011) |
| [126] | 2020 | Australia | Medicare Australia enrolment database | 12 years,<br><br>2004 – 2016 | 236,584<br><br>Aged 45 and above | General population | 1<br><br>Prescription for type 2 diabetes | Physiological health outcome | 39 | Individual | Logistic regression,<br><br>Gradient boosting,<br><br>Deep learning, Random forest | 5-fold CV,<br><br>Data split (Training dataset = 70%, Test dataset = 30%) |
| [127] | 2021 | USA | US National Health and Nutrition Examination Surveys (NHANES) | 10 years,<br><br>2007 – 2016 | 8,829<br><br>Aged 20 and above | General population | 1<br><br>Risk of diabetes | Physiological health outcome | 2 | Individual | Hybrid machine learning model (with artificial neural networks) Please refer [127] for more details. | Data split into test, validation and training datasets<br><br>(Training = 2,227 samples,<br><br>Validation = 2,226 samples<br><br>Test = 4,581 samples) |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [128] | 2021 | USA | 1. Centers for Disease Control and Prevention 500 Cities dataset<br>2. Smart Location Database (SLD)<br>3. Social Vulnerability Index (SVI) data,<br>4. 311 service request data. | 1 year, 2015 – 2016<br>3 years, 2010 – 2013<br>1 year, 2016<br>6 years, 2014-2020 | N/A,<br>196 census tracts. | General population | 6<br>Prevalence of six different health outcomes – cancer, coronary heart disease, diabetes, poor mental health, stroke, and | Physiological and mental health outcome | 60 | Area level (Census tract-level) | Ridge Regression, Lasso Regression, Elastic Net, Support Vector Machine, Decision Tree, Random Forest, Extra Trees, Gradient Boosting | 10-fold CV<br>Data split into training and test set<br>(Training data = 80%, Test data = 20%) |
| [129] | 2021 | USA | 1. American Community Survey<br>2. Veteran Healthcare systems dataset | 4 years, 2014 – 2018<br>N/M | N/A,<br>122 Health care systems | Veteran population | 2<br>COVID-19 cases, COVID-19 deaths | Physiological health outcome | 8 | Area-level (Healthcare system level) | Naive ordinary least squares, LASSO, Ridge regression | 5-fold CV |
| [130] | 2019 | USA | 1. Centers for Disease Control and | 1 year, 2017 | N/A | General population | 2 | Physiological | 14 | Area level (Census tract-level) | Random forest | N/M |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Prevention 500 Cities dataset<br><br>**2.** American Community Survey | 4 years, 2011–2015 | 27,066 census tracts | | Prevalence of coronary heart disease (CHD) and stroke. | health outcome | | | | Ranked top predictors using GINI index score |
| [131] | 2020 | USA | 1. Centers for Disease Control and Prevention 500 Cities dataset<br><br>2. American Community Survey<br><br>3. Environmental Justice Screening dataset | 1 year, 2017<br><br>4 years, 2011–2015<br><br>1 year, 2015–2016 | N/A<br><br>26,697 census tracts | General population | 5<br><br>Prevalence of hypertension, high cholesterol, diabetes, coronary heart disease (CHD), and stroke. | Physiological health outcomes | 19 | Area level (Census tract-level) | Bayesian Additive Regression Trees | N/M<br><br>Ranked top predictors using GINI index score |
| [132] | 2021 | USA | National Epidemiologic Survey on Alcohol and Related Conditions | 2 years,<br><br>2001 – 2002,<br><br>2004 – 2005 | 34,653<br><br>Aged 18 years and above | General population | 1<br><br>Risk of suicide attempt | Mental health outcome | 2,985 | Individual | Balanced random forest | 10-fold CV,<br><br>Out of sample testing,<br><br>Validation across demographic variables (sex, age, ethnicity, and income) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [133] | 2019 | Korea | Korea National Health & Nutrition Examination Survey | 7 years,<br><br>2007 – 2012 | 2,654<br><br>Aged 19 years and above | General population | 1<br><br>Risk of suicide attempt | Mental health outcome | 41 | Individual | Random forest | 10-fold CV<br><br>Data split (Training data = 70%, Test data = 30%). |
| [134] | 2021 | USA | Gallup polling data<br><br>American Community Survey (ACS) | 2 years, 2014 – 2016<br><br>4 years, 2014 – 2018 | 52,006 (for overall wellbeing)<br><br>51,870 (for physical wellbeing) | Veteran Population | 2<br><br>Overall wellbeing,<br><br>Physical wellbeing | General outcome | 162 (overall wellbeing)<br><br>116 (Physical wellbeing) | Individual | Support vector machine,<br><br>XGBoost,<br><br>Multi-layer Perceptron classifier | 5-fold CV<br><br>Data split (Training data = 80%, Test data = 20%) |
| [135] | 2020 | USA | National Survey of Children's Health | 2 years,<br><br>2016 & 2017 | 6,630<br><br>Aged 3 – 17 years | General population | 1<br><br>Treatment of ADD/ADHD | Mental health outcome | 770 | Individual | Classification and regression tree analysis,<br><br>Random decision forest,<br><br>Deep neural network | Out of sample validation<br><br>(Training data= 2016 dataset, Test data= 2017 dataset) |
| [136] | 2019 | Bangladesh | Bangladesh Demographic and Health Survey | 1 year,<br><br>2011 | 2013<br><br>Aged 0.5 – 5 years | General population | 1<br><br>Anaemia | Physiological health outcome | 24 | Individual | Linear discriminant analysis,<br><br>Classification and regression trees,<br><br>K-Nearest Neighbours,<br><br>Support vector machines, | 10-fold CV,<br><br>Data split (Training data = 80%, Test data = 20%) |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Random forest | |
| [137] | 2021 | USA | Communities That Care Youth Survey,<br><br>American Community Survey (ACS) | 6 years,<br><br>2011 – 2017 | 174,864<br><br>Aged 12 – 18 years | General population | 1<br><br>Having suicidal thoughts and behaviour (Yes/No) | Mental health outcome | 7,900 | Individual | K-Nearest Neighbours,<br><br>Naive Bayes,<br><br>Logistic Regression,<br><br>Decision Tree classifier,<br><br>XGBoost,<br><br>LightGBM | Data split (Training data = 80%, Validation data = 10%, Test data = 10%) |
| [138] | 2021 | USA | 1. Centers for Disease Control and Prevention 500 Cities dataset<br>2. American Community Survey<br>3. Environmental Justice Screening dataset | 1 year, 2017<br><br>4 years, 2011-2015<br><br>1 year, 2015-2016 | 26,697 census tracts | General population | 1<br><br>Risk of stroke | Physiological health outcomes | 24 | Area level (Census tract-level) | Quantile regression forests | N/M<br><br>Ranked top predictors |
| [139] | 2020 | USA | 1. National Health Interview Survey | 20 years, 1997-2016 | N/M | General population | 1<br><br>Risk of colorectal | Physiological health outcome | 18 | Individual | Artificial neural network, | Out of sample testing |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | (NHIS),<br><br>2. Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) data set | 8 years, 1993 – 2001 | | | cancer (CRC) | | | | Logistic regression,<br><br>Naive Bayes,<br><br>Decision tree<br><br>Random forest,<br><br>Support-vector machine,<br><br>Linear discriminant analysis | (Training dataset = NHIS,<br><br>Test Dataset = PLCO, and vice versa) |
| [140] | 2020 | USA | 1. Centers for Disease Control and Prevention 500 Cities dataset<br>2. American Community Survey<br>4. Environmental Justice Screening dataset | 1 year, 2017<br><br>4 years, 2011-2015<br><br>1 year, 2015-2016 | 26,697 census tracts | General population | 1<br><br>Risk of stroke | Physiological health outcomes | 24 | Area level (Census tract-level) | Bayesian additive regression trees,<br><br>Gradient boosting machines,<br><br>Extreme gradient boosting<br><br>Random forests. | 200 x 5-fold CV |

***Note:*** *CV – cross-validation; N/A – Not Applicable; N/M – Not Mentioned*

**Table 3-3.** Description of predictor variables

| Ref | Number of predictors and their types |
|---|---|
| [116] | 6.<br>**Demographics:** age, gender, ethnicity, household income, employment status, marital status |
| [117] | 15.<br>**Demographics**: gender, age, education,<br>**Other predictors**: Depressed mood over two weeks, stress level in daily life, subjective health status, reasons for unemployment, days of feeling sick or discomfort, average work week, and limitation of daily life and social activities, EuroQoL-5D (EQ-5D): anxiety/depression, EuroQoL-Visual Analogue Scale (EQ-VAS), EQ-5D: mobility, EQ-5D: pain/discomfort, EQ-5D: usual activities. |
| [118] | 13.<br>**Demographics**: gender, age group, quintile of household wealth, formal education, marital status, falls, smoking behaviour, alcohol consumption, physical activity, employment, size of social network<br>**Other variables**: personal-fitted variable (created using a linear model, outcome (health status) ~ all demographic variables), Trend variable (created using a linear model, outcome (health status) ~ timepoint) |
| [119] | 5.<br>**Demographics**: gender, grade, city type, academic achievement, family structure, family socioeconomic status, and education level of father and mother<br>**Health-related lifestyle factors**: current smoking, current alcohol consumption, substance use, physical activity, obesity, sexual experience, and internet addiction<br>**Psychological stress factors**: sadness, stress, self-rated health, sleep satisfaction, self-rated weight, distorted weight perception, school injury, and violence.<br>**Presence of Comorbidities**: asthma, allergic rhinitis, and atopic dermatitis. |
| [120] | 18.<br>**Demographics**:<br>age, gender, body-mass index, smoking frequency, hispanic ethnicity, american Indian, african american, other, or multiple races, vigorous exercise frequency<br>**Presence of comorbidities**: hypertension, coronary heart disease pooled heart conditions, myocardial infarction, diabetes (non-gestational), heart condition/disease, angina pectoris, ulcer (stomach, duodenal, peptic), stroke, emphysema |
| [121] | 13.<br>**Demographics**: age, age at menarche, age at menopause, age at first live birth, ethnicity, number of first-degree relatives who have had breast cancer, BMI<br>**Other variables**: an indicator of current hormone usage, number of years of hormone usage, pack years of cigarettes smoked, years of birth control usage, number of live births, and an indicator of personal prior history of cancer |
| [122] | 57 variables grouped from 6 different classes.<br>1. adverse socioeconomic and psychosocial experiences during childhood<br>2. socioeconomic conditions |

| | |
|---|---|
| | 3. health behaviours<br>4. social connections<br>5. psychological characteristics, and<br>6. adverse experiences during adulthood |
| [123] | 46 (for malaria) and 53 (for anaemia) variables across 5 different classes were used as predictors.<br>1. social history<br>2. housing<br>3. family<br>4. health-care services and<br>5. demographic factors |
| [124] | 134 discrete variables were extracted as predictors from the US Census, and other community-level datasets.<br>Example predictors – demographics, income level, crime rate, unemployment level, educational attainment, GDP |
| [125] | 75 variables from 3 different classes used as predictors.<br>Baseline variables (e.g., sociodemographic status, financial situation, general health condition, and personal behaviour and habit)<br>1. disease-related variables<br>2. disability-related variables |
| [126] | 39 variables from 4 different classes were used as predictors.<br>1. demographics<br>2. medical and family history<br>3. lifestyle indicators<br>4. dietary indicators |
| [127] | 2 variables (age and waist circumference) were used as predictors. |
| [128] | 60 variables that characterise the social environment, the physical environment, and the aspects and degrees of neighbourhood disorder were used as predictors. |
| [129] | 8.<br>Socio-demographic variables were used as predictors (e.g., ethnicity, population density, gender distribution, age distribution, marital status distribution, education distribution, income distribution, and the poverty rate. |
| [130] | 14.<br>**Unhealthy behaviours:** binge drinking, smoking, no leisure-time physical activity, insufficient sleep, and obesity<br>**Prevention measures**: lack of health insurance, routine checkup, cholesterol screening<br>**Socio-demographic factors:** age, gender, race/ethnicity, income, and education |
| [131] | 19.<br>**Unhealthy behaviours:** binge drinking, smoking, no leisure-time physical activity, insufficient sleep, and obesity<br>**Prevention measures:** lack of health insurance, routine checkup, cholesterol screening<br>**Socio-demographic factors**: age, gender, race/ethnicity, income, and education<br>**Environmental measures:** ozone level in air, PM2.5 level in air, traffic proximity and volume, and house built prior to 1960 |
| [132] | 2985 variables (e.g., age, family income, financial crisis, marital status, education level, paternal alcohol misuse, and parental separation) were used as predictors. |

| | |
|---|---|
| [133] | 41 variables from 4 different classes were used as predictors.<br>1. demographics<br>2. physical health,<br>3. substance use<br>4. socioeconomic status |
| [134] | 113 demographic variables and 49 zipcode level demographic variables<br>(e.g., percentage of population employed in a specific sector) were used as predictors |
| [135] | 770 variables across various domains (such as physical and oral health, emotional and mental health, health insurance coverage, health care access and quality, community and school activities, family health and activities, neighbourhood safety and support) were used as predictors. |
| [136] | 24 variables across various domains (such as demographic, socio-economic, health and nutritional history) were used as predictors. |
| [137] | 7900.<br>Variables across various domains (such as demographics, family life, past behaviour, community involvement and perception of norms, detailed information on school involvement and behaviour, drug usage, gambling, religion, and antisocial behaviour)<br>Zip-code level data (such as marriage/divorce rates, racial makeup, labour force details, average household information, income percentages, and educational background percentages etc) were used as predictors. |
| [138] | 24.<br>**Unhealthy behaviours:** smoking, no leisure-time physical activity, insufficient sleep, and obesity<br>**Prevention measures**: lack of health insurance, visits to dentist, colonoscopy screening, routine checkup<br>**Socio-demographic factors:** age, gender, race/ethnicity, income, and education<br>**Environmental measures:** ozone level in air, PM2.5 level in air, traffic proximity and volume, and house built prior to 1960. |
| [139] | 18.<br>**Demographics**:<br>age, gender, body-mass index, smoking frequency, ethnicity, vigorous exercise frequency, family history<br>**Presence of comorbidities**:<br>diabetic status, hypertension, ulcers, a stroke, any liver comorbidity, arthritis, bronchitis, coronary heart disease, myocardial infarction, and/or emphysema |
| [140] | 24.<br>**Unhealthy behaviours:** smoking, no leisure-time physical activity, insufficient sleep, and obesity<br>**Prevention measures:** lack of health insurance, visits to dentist, colonoscopy screening, routine checkup<br>**Socio-demographic factors:** age, gender, race/ethnicity, income, and education<br>**Environmental measures:** ozone level in air, PM2.5 level in air, traffic proximity and volume, and house built prior to 1960. |

## Discussion

### *Dataset description*

Of the 25 studies, 17 (68%) were based on data collected in the USA, while three originated from South Korea, and the remaining five drew data from other countries such as Bangladesh, Australia, and the UK. The majority of these (approximately 80%) were published in the last three years. Notably, nearly half of them (11 in total) drew data from datasets compiled within the last decade, i.e., after 2010. This trend highlights the increasing interest in utilising administrative data in recent years, potentially influenced by various government initiatives aimed at collating administrative data for research purposes. The increasing accessibility to administrative data for researchers in the past decade is a promising sign, suggesting an exponential growth in data science studies that explore population-level outcomes in the coming years. Government-supported administrative datasets play a significant role in facilitating such research.

Across these studies, a total of 23 distinct datasets were employed, with the majority (approximately 60%) boasting a sample size exceeding 6,000. While the principle in data science is often "more data is better," challenges such as data imbalance need to be considered. Given that these datasets are typically nationally representative samples, researchers must exercise caution in assessing data efficacy and quality, which includes addressing issues like honesty in survey responses. An example of this was found in one survey that administered a "fake drug" question to evaluate the honesty of young adults [137]. Applying weighting techniques can potentially offer solutions for dealing with class imbalance issues. It is worth noting that all studies, except two, utilised government administrative datasets or surveys conducted by various government agencies. The remaining two studies were based on datasets collected by university researchers.

Furthermore, approximately 50% of the studies (12 in total) included data from adults aged 18 and above, among which five studies focused exclusively on adults aged 50 and above. Additionally, three studies included data from young people aged between 11 and 18 years, and one focused on toddlers. Across these studies, data from both men and women were used, except for one study that exclusively examined women. Moreover, 23 out of the 25 studies were based on data from the general population, while the other two focused on the veteran population. The wide range of age groups and populations covered in these studies indicates their broad scope and applicability.

***Data pre-processing***

Data pre-processing is an integral step in data science and machine learning modelling. Proper data preparation is essential for several reasons. Training a model with incorrect data, such as mislabelled categories in a variable, duplicated entries, or inconsistent formats, can lead to unreliable and non-generalisable predictions [141]. In the context of population-level data, these datasets serve as ground truth information, and model predictions built on flawed data can have serious implications for policy and decision-making. For instance, if health-related data are used to inform public health decisions, inaccurate or incomplete data can result in suboptimal outcomes.

One common issue in large health datasets is class imbalance [142]. Class imbalance refers to an unequal distribution of classes in a categorical variable. This means that one category has a significantly larger number of cases compared to others. Class imbalance can be categorised into slight imbalance (where the class distribution is only slightly skewed) and severe imbalance (where the class distribution is heavily skewed, e.g., one positive case for every thousand negatives). While a slight class imbalance is not an issue in most cases, the problem with severe class imbalance lies in the insufficient training data available for the minority class

[143]. Many machine learning algorithms perform best when presented with balanced datasets, and in cases of class imbalance, the model tends to be biased towards the majority class, often resulting in poor classification of the minority class.

To overcome class imbalance, researchers have utilised various techniques, such as the cost-sensitive classification method [144]. This technique assigns different costs to misclassifying different classes, making the classifier more sensitive to minority classes [144]. Resampling the imbalanced class is another common approach. Researchers either under-sample the majority class or up-sample the minority class, or sometimes use a combination of both strategies [145]. This rebalancing helps ensure that the machine learning model does not favour one class over the other and provides more reliable predictions. Out of the 25 studies reviewed, four employed resampling to address class-imbalance problems in their datasets [117, 119, 133, 137]. Notably, one study [133] applied the Synthetic Minority Over-sampling Technique (SMOTE) [146] to achieve a balanced distribution in their dataset. One study [132] employed the Balanced Random Forest, an ensemble method extending the traditional Random Forest algorithm with a balancing mechanism. This approach has shown promise in handling imbalanced data, providing an alternative for class-imbalance challenges in machine learning [147].

Another challenge in data pre-processing is dealing with missing values, as many machine learning models cannot handle datasets with missing information. Researchers often opt to exclude missing values, but this approach might not always be suitable, especially when training data are limited, leading to insufficient data for model building [148]. Missing data are generally of four types: Structurally Missing Data, Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) [149]. Structural missing data occur mainly from the dataset's design, such as when a respondent skips a question due to the survey's structure (e.g., a student may not answer a question related to employment or

income). MCAR means missingness unrelated to observed responses, MAR implies missingness depending on observed responses, and MNAR indicates non-random missingness related to unobserved values [148].

Of the reviewed studies, seventeen addressed missing data, with more than half (ten studies) excluded observations with missing values in their datasets. Alternatively, five studies implemented diverse imputation techniques to address missing data. These approaches encompassed ensemble methods like Bayesian Additive Regression Trees (BART) [124], a simplistic imputation assigning -1 to missing values [125], and the Multiple Imputation by Chained Equations (MICE) method [117, 133]. Detailed information about the MICE method can be found elsewhere [150]. One study [139] explored multiple imputation techniques, creating distinct datasets for evaluation. Additionally, another study [132] employed variable transformation and the missing-indicator method to preserve information from incomplete datasets. This involved adding two missing categories (structural/true) and incorporating binary variables to signify the presence of missing data. In this method, the variables were first transformed by adding two missing categories (structural/true) and employed the missing-indicator method, which involves adding binary variables to signify missing data presence. Ultimately, the choice of imputation method should be made carefully, as it can significantly impact model performance. It is advisable to understand the type of missing data before addressing it [151].

Finally, three studies [122, 128, 134] standardised predictor variables, while one study [121] normalised predictor variables. These preprocessing techniques aim to align variables to a common scale, typically by rescaling to a mean of 0 and a standard deviation of 1, or normalising between 0 and 1 [152]. These practices can be particularly useful when variables are measured in different units or have diverse ranges. Standardising them ensures that they contribute equally to the model's predictions, preventing one variable from dominating the

others. However, it is also worth noting that the importance of standardisation can vary depending on the type of model employed. For instance, it plays a crucial role in models sensitive to the scale of input predictors, such as k-nearest neighbours. In contrast, its impact might be less pronounced in tree-based algorithms, such as the Random Forest. Therefore, the decision to standardise should be carefully considered within the context of the specific modelling technique employed.

In summary, the reviewed studies highlighted key findings in data pre-processing and machine learning modelling for population-level health data. These include the necessity of addressing class imbalance through techniques such as resampling and cost-sensitive classification methods. Additionally, strategies for managing missing data, such as diverse imputation techniques and variable transformation methods, were explored. Standardising and normalising predictor variables were also recognised as crucial steps in data preprocessing to ensure model reliability. These findings collectively emphasise the importance of robust data pre-processing practices to ensure that machine learning models yield reliable and interpretable results for informed decision-making.

***Outcome variables***

The number of outcome variables varied from one to six per study, although most (21 out of 25) focused on predicting one or two outcomes. In total, these studies collectively predicted 21 distinct outcome variables, as depicted in Figure 3-2.

These diverse outcomes were classified into three main categories: (1) Physiological, (2) Mental, and (3) General outcomes. Among them, 14 studies (approximately 60%) were exclusively dedicated to predicting physiological outcomes. In contrast, six studies solely focused on mental health outcomes, and one study ventured into predicting both physiological and mental health outcomes. Additionally, three studies [118, 122, 134] broadened their scope by predicting general outcomes, such as overall wellbeing.

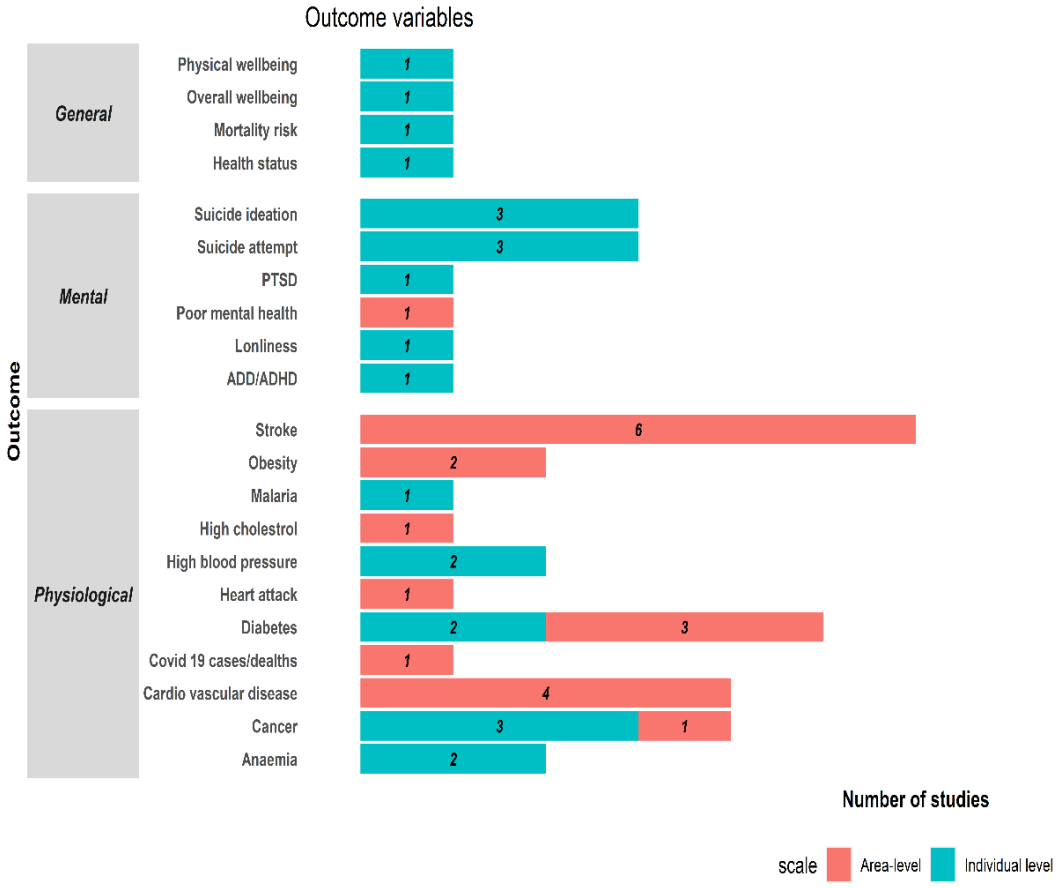



**Figure 3-2.** Type and scale of outcome variables in the reviewed studies

Further categorisation was based on the scale of prediction: (1) Individual level, which pertains to predictions about specific individuals, such as estimating the risk of disease for an individual, and (2) Area level, where predictions relate to groups of individuals residing in a specific region. Notably, a substantial majority of studies (18 out of 25, approximately 72%)

concentrated on predicting outcomes at the individual level. Only seven studies targeted area-level predictions, with census tracts being the most frequently chosen scale (six out of seven).

Among the 15 studies that predicted physiological health outcomes, eight were focused on individual-level predictions. Within this group, the most predicted physiological outcomes were risk of cancer. For area-level predictions, stroke (six studies) and cardiovascular disease (four studies) were the prominent physiological outcomes addressed. In the case of studies predicting mental or general health outcomes (10 studies in total), only one ventured into area-level prediction. The most frequently predicted mental health outcomes were suicide ideation and suicide attempts, with each outcome being explored in three separate studies. Notably, of all 25 studies, only one study tackled the prediction of wellbeing as an outcome, and this was limited to the veteran population.

***Features or Predictor variables***

Each of the studies encompassed a diverse set of predictors (also known as 'features'), ranging from two to thousands. For ease of visualisation and analysis, the features used in these studies were categorised into four groups: (1) under 25 features, (2) between 25 and 50 features, (3) between 50 and 100 features, and (4) above 100 features.

A significant proportion of studies (20 out of 25, or 80%) utilised fewer than 100 features, with 13 of these using fewer than 25 features. This observation suggests that researchers often prefer a more concise set of features when constructing their machine learning models. Conversely, a smaller number of studies (5 out of 25) opted for a more extensive feature set, two of which had more than 2,000. The number of features used in a study can affect model interpretability, with increased features leading to greater model complexity. Additionally, a large feature set also poses challenges for the future usability of these models, as these data must be continually collected and updated in the future.

Nearly half of the 25 studies (12 out of 25, or 48%) involved a feature selection process as a precursor to training the machine learning models. Feature selection is a vital step that reduce the number of features needed in a model, reducing model complexity [153]. One approach is to train models with different sets of features and assess their performance. Out of the 25 studies, three [120, 125, 128] took this approach, constructing various feature sets by excluding different variables in each set and then comparing their performance.

An alternative feature selection method involves ranking the importance of features using various algorithms, eliminating the least important features in a recursive manner. Five studies adopted this approach, employing machine learning models such as Random Forest, Quantile Random Forest, Gradient Boosting, and Bayesian Additive Regression Trees (BART) to identify and select the best features before model training. It is worth noting that tree-based models are frequently employed in ranking features, often using metrics such as the GINI index [130]. This is because tree-based models inherently provide feature importance scores, making them a popular choice for feature selection [154].

Lastly, two studies [123, 138] used correlation testing to select features and minimise the issue of multicollinearity, where several variables together may not be necessary to train the models, as they convey similar information [155]. Another study [116] adopted the Hilbert-Schmidt Independence Criterion Lasso method to select the most relevant features, showcasing the diversity of feature selection techniques across the studies.

***Types of features***

Next, a wide range of features were utilised across the studies, and these can be broadly categorised into seven key groups based on their type and relevance for predicting health outcomes.

1. ***Socio-demographic factors*:** These factors were common in all studies and are recognised as valuable predictors for various health outcomes. Socio-demographic factors include gender, age, ethnicity, education, and socio-economic status. These variables provide crucial insights into individuals' backgrounds and social contexts, making them essential for predictive modelling in public health research.
2. ***Physiology-related variables*:** In nine studies, non-clinical physiological variables have been used, emphasising their importance in predicting physiological health outcomes. These variables are often related to an individual's physiological condition, including the presence of comorbidities, disease or disability-related factors, and medical history. Such variables play a significant role in understanding and predicting health conditions and their interrelationships.
3. ***Psychological variables*:** Approximately four studies integrated psychological variables into their predictive models to forecast mental health-related outcomes, such as suicide ideation. These variables typically encompass emotional states, including levels of anxiety, depression, stress, and sadness.
4. ***General health-related variables***: Twelve studies leveraged general health-related variables, irrespective of the type of health outcome they predicted. Examples of these variables include health status, body mass index (BMI), and healthcare service utilisation. These variables provide a holistic picture of an individual's overall health and wellbeing, offering valuable insights for health predictions.
5. ***Lifestyle behaviours***: Nearly 70% of the studies (17 out of 25) incorporated lifestyle behaviours as predictors. These behaviours encompass factors such as smoking and drinking status, dietary indicators, and physical activity levels. Decades of public health research have established the profound influence of lifestyle choices on health and

wellbeing outcomes, underscoring the significance of these variables in predicting public health outcomes [156, 157].

6. ***Environmental factors***: A smaller proportion of studies (4 out of 25) integrated variables related to the environment, such as ozone levels in the air and proximity to traffic. These studies recognise the impact of the environment on health and wellbeing, with the potential to explore additional environmental factors as predictors. For instance, research has highlighted the positive effects of green spaces on wellbeing, providing a foundation for further investigation [158].
7. ***Community-level variables***: Ten studies employed variables aggregated at the community level, including factors like socio-demographic distribution, the percentage of the population employed in specific sectors, and crime distribution. While seven of these studies predicted community-level outcomes, three focused on individual-level predictions. Incorporating community-level and individual-level factors may offer valuable insights into the social and economic environments in which people reside, further enhancing the accuracy of health outcome predictions.

### *Predictive modelling*

#### *Machine learning software*

The choice of machine learning software varied across studies. Nearly half (11 out of 25) opted for R programming [159] for building and evaluating their machine learning models. The popularity of R in this context can be attributed to its comprehensive packages and libraries specifically tailored for data analysis, making it an attractive choice for researchers in public health. The open-source nature of R promotes transparency and reproducibility, allowing other researchers to employ the same tools and replicate methodologies.

Six studies chose Python, a versatile and widely adopted programming language in the machine learning community. Its extensive ecosystem of libraries and frameworks, including scikit-learn (which offers tools for data preprocessing, model selection, and evaluation) and TensorFlow (a deep learning framework), provides researchers with a powerful toolkit for creating predictive models [160, 161]. Python's popularity extends beyond academia, making it a valuable skill for data scientists and machine learning practitioners [162]. Lastly, two studies preferred MATLAB, known for its numerical computing capabilities, and one study utilised node.js. Ultimately, the choice of software depends on the complexity of the analysis, the availability of suitable libraries and frameworks, and researchers' familiarity and expertise with the programming language.

***Machine learning models***

Selecting the appropriate predictive model is a critical step in machine learning, as it greatly influences the model's performance and its ability to make accurate predictions. The machine learning models employed in the reviewed studies fall into four main categories, each represented by a distinct colour in Figure 3-3: Linear models (depicted by orange blocks), Tree-based models (green blocks), Neural networks (blue blocks), and other models (highlighted in light yellow).

Linear models, which include traditional statistical techniques like linear regression and logistic regression, are characterised by their simplicity and interpretability. These models are particularly useful when the relationship between predictors and outcomes is expected to be linear. Out of the 25 studies, almost half (12 studies) incorporated these traditional models, although they were primarily used in conjunction with more complex models. This highlights the common practice of using linear models for initial analyses and as benchmarks for more complex models.

Next, tree-based models, such as random forests, are machine learning techniques that make predictions by constructing decision trees. These decision trees work by recursively splitting the dataset into subsets based on the most significant predictors, creating a hierarchical structure resembling an inverted tree [154]. These models are known for their ability to capture complex non-linear interactions within the data, making them valuable for predictive modelling. Nearly all the studies ($n$ = 23) included in the review incorporated at least one tree-based model in their analysis. The random forest model was the most popular algorithm, with almost 60% of studies using it.

Neural networks, another category, are particularly powerful for tasks like image recognition and natural language processing due to their ability to model intricate patterns in large datasets. They are inspired by the structure of the human brain, consisting of layers of interconnected nodes or neurons [163]. Each neuron processes information and passes it to the next layer, mimicking the way human brain cells communicate. Six studies out of the 25 utilised artificial neural networks, while two studies employed more complex models such as deep learning. Lastly, there are other models employed in nine studies, such as support vector machines, k-NN, Naïve Bayes, each with its own advantages and use cases.

In summary, the choice of model type depends on the dataset characteristics, the complexity of relationships, computational resources available, and the research objectives. Linear models suit straightforward relationships, while tree-based models excel in capturing complex multilayered interactions. Neural networks are adept at recognising intricate patterns, such as those in medical imaging, and other models serve specific tasks or data types. Each model type has advantages and disadvantages, and selecting the right one depends on balancing these factors with the research goals.

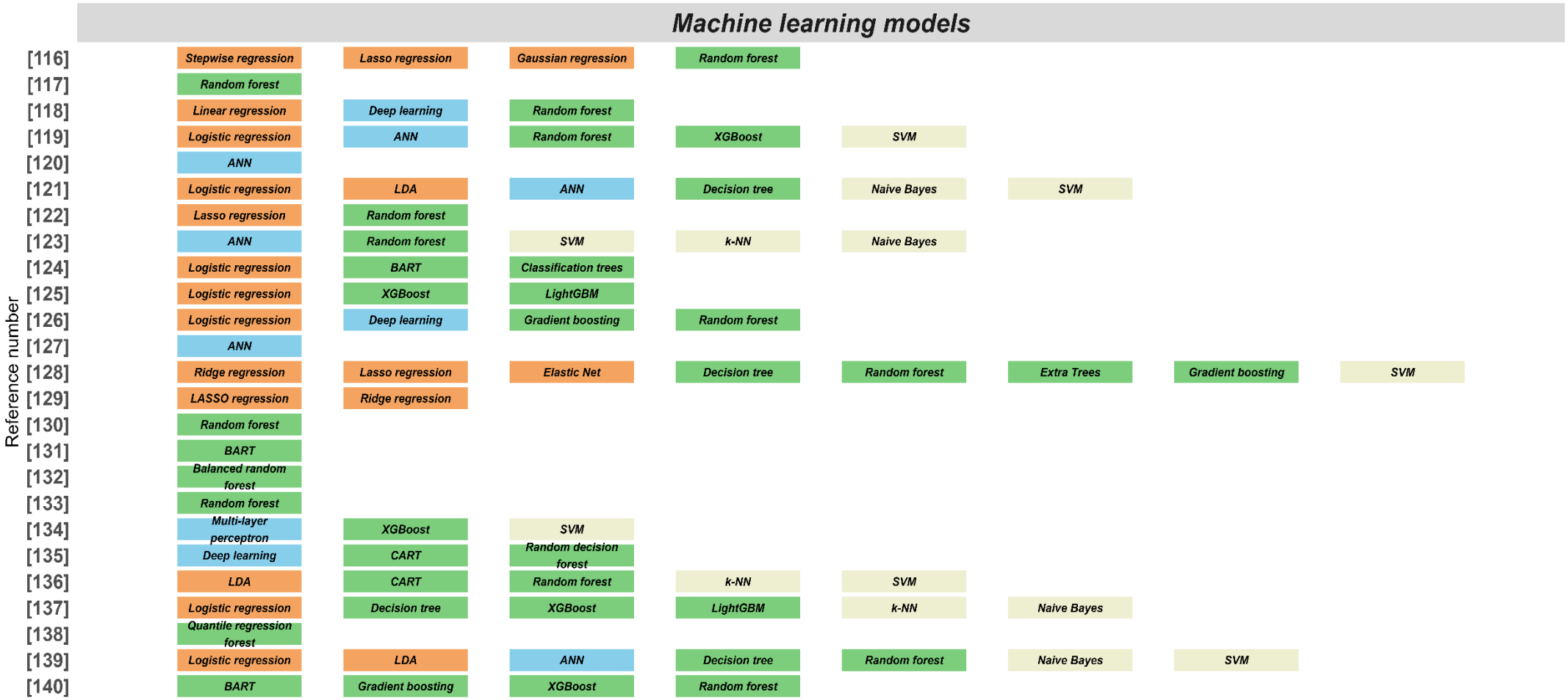



**Figure 3-3.** Types of machine learning models used.

### *Hyperparameter selection*

Hyperparameter selection is another critical step in optimising machine learning models. These parameters are settings that control how a machine learning model learns from data. Unlike the data itself, these settings are chosen by the researcher before the modelling process. For instance, in a neural network, the learning rate is a hyperparameter that determines how quickly the model adjusts its predictions based on new data. Similarly, in a random forest model, the number of trees is a hyperparameter that influences the model's complexity and predictive accuracy [164]. These parameters play a key role in determining the model's performance, robustness, and ability to generalise to new, unseen data. The process generally involves experimenting with different configurations to identify the settings that yield the best results.

In the context of this review, some studies explored various hyperparameter combinations manually ($n = 5$) and validated model performance using techniques like *k-fold* cross-validation (explained in the next section). Alternatively, some studies ($n = 6$) employed the grid search method to systematically adjust hyperparameters. Grid search involves defining a range of values for each hyperparameter and exhaustively testing all possible combinations. This method simplifies hyperparameter tuning by automating the search process and identifying the most suitable configuration. A detailed explanation of all hyperparameters is beyond the scope of this review. Nonetheless, being aware of their significance and making informed choices during model design is crucial.

### *Evaluation of Model Performance*

Cross-validating a machine learning model is crucial for evaluating its true predictive performance on new data and identifying issues such as overfitting. While the ideal scenario involves testing the model on a completely independent dataset, in many cases, a separate dataset is not available. Therefore, an alternative strategy is to split the dataset into a training

set and a testing set. This split can be based on a proportion, such as an 80/20 split (80% training data and 20% test data).

Out of the 25 studies, six employed this method, with the 80/20 split being the most popular. Some studies took a more comprehensive approach by dividing their data into a training set (e.g., 80% of the dataset), a validation set (e.g., 10% of the dataset), and a test set (e.g., 10% of the dataset). This approach allows for fine-tuning the model by experimenting with various hyperparameters on the validation dataset before testing on the test dataset. Another approach is to split the data based on years or data collection waves, mimicking an independent dataset. For example, one study [116] used data from years 2007–2010 as their training data and data from years 2011–2012 from the same dataset as their test data. Another three studies employed this validation method [120, 125, 135].

The most popular method for validating model performance, used by approximately half of the studies, is *k-fold* cross-validation. In this approach, the dataset is divided into k subsets, and the model is trained on k-1 subsets and tested on the remaining subset. This process is repeated *k* times, with each subset acting as the test set, and the results are then averaged. Common choices for *k* include 10-fold or 5-fold cross-validation. This method provides a robust assessment of the model's performance by ensuring that each data point is used for both training and testing, reducing the risk of bias.

Finally, some studies validated model performance by comparing predictions across specific demographic categories. For instance, in one study [132] assessing the risk of suicide attempts, the model's predictions were tested across demographic subgroups based on gender, age, ethnicity, and income. This approach is particularly valuable for understanding the sensitivity of the model's predictions. By analysing how the model performs across different demographic subgroups, researchers gain insights into where the model excels and where it may fall short.

This information is instrumental in guiding further refinements to the model, helping it become more accurate and reliable.

**Limitations**

The broad approach of this review provided a comprehensive summary of methodologies, leading to the inclusion of a diverse array of studies covering a wide spectrum of outcomes. While this approach ensured a general understanding of the field, it precluded the assessment of specific results within each study (i.e., comparing the accuracy of predictions across studies). Furthermore, there was a limited number of studies that specifically focused on subjective outcomes (i.e., life satisfaction or subjective wellbeing). This scarcity reduced the ability to draw specific methodological conclusions about these outcomes. Given subjective outcomes are often linked with complex human perceptions and experiences, it presents a unique challenge in the context of machine learning and is a clear avenue for future research.

**Conclusion**

To summarise, this scoping review underlined the promising potential of machine learning for predicting health and wellbeing outcomes using administrative variables. The findings offered valuable insights into the different methodologies that were used, and the importance of these methodological decisions for model building. Specifically, the review highlighted:

- A growing interest in using administrative data, particularly in recent years.
- The importance of robust data preprocessing, including addressing class imbalance, handling missing data effectively, and employing techniques like resampling and variable transformations.
- The focus on diverse outcomes across physical, mental, and general health domains, with an emphasis on individual-level predictions and physical health outcomes.

- The use of varied predictors, predominantly socio-demographic factors, physiological variables, and lifestyle behaviours, with a trend towards using concise feature sets and employing feature selection methods.
- The use of a mix of machine learning models, primarily tree-based models, along with linear models and neural networks, reflecting a flexible approach based on data characteristics.
- The employment of validation techniques such as train-test splits, k-fold cross-validation, and demographic subgroup analysis for robust model performance assessment, highlighting the importance of thorough validation methods.

Despite these highlights, a notable gap in the literature was observed, with limited studies examining subjective outcomes. Given the importance of subjective wellbeing, there is significant scope for future research in this area, which could contribute to a more comprehensive understanding of how human perception and experience can be captured and modelled from administrative data.

# Chapter 4 – Subjective wellbeing outcomes across different demographic groups and regions in New Zealand – A cross-sectional and trend analysis.

## Preface

The preceding chapter systematically reviewed the literature related to predicting health outcomes from administrative data. This demonstrated the promising potential of machine learning for predicting health and wellbeing outcomes using administrative variables, but also highlighted the lack of focus on psychological (subjective) outcomes such as wellbeing—a central focus of this thesis. This chapter explores the recent wellbeing landscape in New Zealand, utilising data sourced from the General Social Survey. There is a particular focus on examining the distribution of wellbeing outcomes across sociodemographic groups, and if these trends change over time. These insights will be crucial for guiding methodological decisions in subsequent chapters, where subjective wellbeing outcomes will be predicted from administrative variables.

## Introduction

The significance of individual and collective wellbeing is gaining global recognition, prompting governments worldwide to prioritise the understanding and enhancement of population wellbeing [8, 165, 166]. Research indicates that individuals with elevated levels of wellbeing are more likely to exhibit positive health outcomes [167], increased productivity [168], and build stronger social relationships [169]. Furthermore, the influence of population wellbeing extends beyond individual lives, impacting a country's overall economic productivity [170], and shaping the landscape of social progress [171] . In New Zealand, the assessment of population wellbeing is a systematic process conducted through the General Social Survey (GSS). Initiated in 2008, the GSS has seen six consecutive biennial waves, with the 2020 wave delayed until 2021 due to the disruption caused by the COVID-19 pandemic.

The GSS captures a wide array of wellbeing facets across twelve distinct domains, based on the Living Standards Framework (LSF) [48]. These encompass health, housing, income and consumption, jobs and earnings, leisure and free time, knowledge and skills, safety and security, social connections, cultural identity, civic engagement and governance, environmental quality, and subjective wellbeing. This study specifically focuses on the subjective aspect of wellbeing. Often referred to as psychological wellbeing or happiness, subjective wellbeing is a multidimensional construct encompassing an individual's overall evaluation and assessment of their own life [172]. This encompasses various dimensions of life, including emotional experiences, life satisfaction, and a sense of purpose or meaning. It therefore represents a broad perspective, offering a comprehensive view of one's quality of life through both affective and cognitive lenses [12].

Subjective wellbeing holds particular significance in the overall study of wellbeing due to its focus on an individual's personal evaluation and perception of their own life [12, 173]. While objective indicators like income, health, or social connections, are undoubtedly crucial,

subjective wellbeing provides a unique and valuable perspective for several reasons. Firstly, it captures the individual's subjective experience, allowing for a direct understanding of how they feel about their life and circumstances. This personal perspective complements and enriches other objective measures. subjective wellbeing reflects an individual's ability to adapt and cope with life's challenges [3]. Even in adverse situations, individuals can maintain or recover their subjective wellbeing, showcasing resilience that might not be evident in purely objective measures.

The increasing acknowledgment of subjective wellbeing as a crucial policymaking indicator is reflected in global initiatives committed to integrating wellbeing metrics into policy frameworks [9, 171, 174]. The UK's "Measuring National Wellbeing Programme" [175, 176], the French Commission's report on the "Measurement of Economic Performance and Social Progress" [177], and the European Commission's "GDP and Beyond" project [178] stand out as notable examples of these endeavours. These initiatives underscore a shift toward more people-centric governance, where policies oriented toward enhancing subjective wellbeing may contribute to overall societal happiness and satisfaction. Elevated levels of subjective wellbeing have further been associated with improved health outcomes and increased longevity [167, 179] and individuals with a positive sense of wellbeing often adopt healthier lifestyle behaviours and experience lower levels of stress, thereby contributing to overall physical health [179].

Historically, subjective wellbeing studies were limited in scope, often focusing on small groups of individuals [180]. This has recently changed, and researchers are now measuring and understanding subjective wellbeing from large nationally representative samples across the world spanning different cultures [181]. Traditional subjective wellbeing studies were largely dominated by the set-point theory, positing that subjective wellbeing remains stable over time due to the persistence of personality traits and the phenomenon of hedonic adaptation [182,

183]. Nevertheless, the landscape has evolved, recognising the limitations of historical approaches that provided a static snapshot. The past few decades have seen a paradigm shift, emphasising the importance of studying subjective wellbeing over time. This has become possible due to the increasing availability of longitudinal data. For instance, the European Social Survey [184] has been collecting methodologically robust cross-national data for more than 30 European countries every two years since 2002.

Subjective wellbeing is subject to change over time due to various factors, such as the impact of economic shifts, life events like marriage or the loss of a loved one, and changes in the environment, such as living conditions or societal dynamics [185, 186]. Studies show that during economic instability, the resulting financial strain and uncertainty can significantly reduce life satisfaction and happiness [187]. Similarly, research indicates that while significant economic advancements can improve material conditions, these gains may be accompanied by a decline in social capital—trust, social networks, and civic engagement—negatively impacting overall wellbeing over time [188]. This dynamic nature of subjective wellbeing offers researchers and policymakers the opportunity to track and analyse the intricate fluctuations in an individual's and thereby a nation’s wellbeing over both short and long periods [189]. In New Zealand, the assessment of subjective wellbeing in the GSS relies on four primary indicators: life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. However, for the purposes of this study, our focus is specifically on life satisfaction and life worthwhileness. This decision is driven by the availability of data, as these variables span more than two GSS waves, enabling a comprehensive trend analysis. It is important to note that family wellbeing and mental wellbeing variables were introduced more recently, with data being available from 2018 onward, and present opportunities for exploration in future studies.

The overarching goal of this thesis is to model and predict subjective wellbeing outcomes in New Zealand. To achieve this, it is critical to understand the recent wellbeing landscape in New Zealand, and identify factors associated with wellbeing over time. This information can be used to identify key variables that can be utilised in the predictive modelling proposed in subsequent thesis chapters. Importantly, developing predictive models that remain stable over time necessitates an understanding of how wellbeing outcomes and demographic variables evolve over time. Therefore, the primary aims of this study are: (1) to describe the distribution of wellbeing outcomes across different sociodemographic groups in New Zealand, and (2) to explore the association between wellbeing outcomes and various demographic variables, and if these vary over time.

## Methods

### *Participants*

The data used in this study were obtained from the New Zealand GSS. The GSS is a nationally representative survey conducted biennially by Stats NZ that aims to capture New Zealanders' wellbeing across different domains [10]. It consists of two questionnaires: (1) a personal questionnaire, and (2) a household questionnaire. Between 8,000 and 10,000 households (sampled to be representative of the New Zealand population) are shortlisted to complete the household questionnaire. One individual person (aged 15 years or over) is randomly selected from each household to complete the personal questionnaire. Each individual participant's data can be linked to their corresponding household questionnaire using a unique household identifier. This study makes use of data collected across three GSS waves: 2014, 2016 and 2018. It is important to note that the individuals surveyed each year are different, making this a series of repeated cross-sections rather than a longitudinal panel. More information on the GSS can be found elsewhere [190].

### *Measures*

Life satisfaction and life worthwhileness are two different subjective wellbeing outcomes used for this analysis. Both outcomes, detailed in Table 4-1, were chosen for their consistent availability across the three GSS waves. Life satisfaction typically reflects an individual's overall contentment with various aspects of their life, encompassing factors such as relationships, work, and personal achievements [191]. On the other hand, life worthwhileness delves into the perceived meaningfulness or significance an individual attributes to their lives [192]. While life satisfaction may capture an individual's general sense of happiness, life worthwhileness focuses on the existential depth and purpose that individuals find in their daily existence. Past research has predominantly focused on understanding life satisfaction, creating a gap in our understanding of life worthwhileness [192]. However, it's important to recognise that both of these measures are significant components of subjective wellbeing, each capturing distinct dimensions of individuals' experiences and perceptions. They are equally relevant in policy-making contexts, contributing to a comprehensive understanding of subjective wellbeing and its determinants [193, 194]. Next, a concise set of seven sociodemographic variables were selected. The decision to focus on these specific demographic factors was based on their applicability to the whole population and because they are key demographic variables commonly considered as stratifiers in descriptive research and policymaking processes. They are widely used in demographic studies and are essential for understanding population dynamics. Moreover, these variables align well with census data, laying a robust foundation for the development of prediction models in subsequent chapters. These variables have been organised into categories, which are shown in Table 4-2.

**Table 4-1.** Wellbeing outcome summary

| Outcome variable | Survey question in the GSS | Range | Description |
|---|---|---|---|
| 1. Life satisfaction | Where zero is completely dissatisfied, and ten is completely satisfied, how do you feel about your life as a whole? | 0 – 10 | Completely dissatisfied – Completely satisfied |

| | | | | |
|---|---|---|---|---|
| 2. | Life worthwhileness | Where zero is not at all worthwhile, and ten is completely worthwhile, overall, to what extent do you feel the things you do in your life are worthwhile? | 0 – 10 | Not at all worthwhile – Completely worthwhile |

**Table 4-2.** Demographic variable summary

| | Predictor variable | Type of variable | Number of categories | Description |
|---|---|---|---|---|
| 1. | Age range | Categorical | 6 | 15-24 years<br>25-34 years<br>35-44 years<br>45-54 years<br>55-64 years<br>65 years or over |
| 2. | Gender | Categorical | 2 | Male or Female |
| 3. | Ethnicity | Categorical | 5 | New Zealand European<br>New Zealand Māori<br>Pacific<br>Asian<br>Middle Eastern/Latin American/African and Other Ethnic groups |
| 4. | Region | Categorical | 6 | Auckland<br>Wellington<br>Northland group (Northland, Bay of Plenty, Gisborne)<br>Rest of North Island<br>Canterbury<br>Rest of South Island |
| 5. | Household size | Categorical | 6 | One person<br>Two people<br>Three people<br>Four people<br>Five people<br>Six or more people |
| 6. | Personal Income | Categorical | 8 | $0 - $30,000<br>$30,001 - $35,000<br>$35,001 - $40,000<br>$40,001 - $50,000<br>$50,001 - $60,000<br>$60,001 - $70,000<br>$70,001 - $100,000<br>$100,001 or more |
| 7. | Household income | Categorical | 8 | Same as personal income |

***Data Analysis**:*

All analyses were performed in R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria), inside the Stats NZ Integrated Data Infrastructure (IDI) data lab. Various

confidentiality rules (e.g., random rounding) were applied to the results in line with IDI protocols, and were screened by Stats NZ prior to release [195]. Data from all three GSS waves were combined into a single dataset, and the demographic variables were condensed into fewer categories to simplify analysis. Furthermore, any missing values and responses labelled as "Don't know" or "Refused to answer" were excluded from the dataset. The number of valid data points (for each variable) across each wave is presented in the Appendix section – Table A-2. Firstly, descriptive statistics (means and standard deviations) were computed to describe the distribution of each outcome variable (life satisfaction, life worthwhileness) across various demographic subgroups (Aim 1). To determine the association between wellbeing outcomes and each demographic variable (Aim 2), a series of linear regression models were fit. Survey year (i.e., 2014, 2016, 2018) was added as an additional predictor in one set of models, allowing for a trend analysis that aimed to assess how this relationship might change over time. An interaction term was specified to examine the potential varying associations between each wellbeing outcome and sociodemographic variable across different survey years. An analysis of variance (ANOVA) table was generated for each model to summarise the main and interaction effects. The association between each wellbeing outcome and sociodemographic variable was interpreted as varying across time if the interaction term was statistically significant ($p < 0.05$).

**Results**

The descriptive statistics for each wellbeing outcome across each demographic category are presented in Appendix Table A-3. Table 4-3 and Table 4-4 show the linear regression results for the life satisfaction and life worthwhileness outcomes (respectively), encompassing both simple regression models (with a single sociodemographic predictor) and adjusted regression models incorporating survey year as an interaction term. The interaction term reflects how the relationship between each sociodemographic variable and wellbeing outcomes varies across

different survey years. For each sociodemographic variable, the regression coefficient of the first category represents the model intercept, while the subsequent categories are relative to the intercept.

Unadjusted model results highlight that individuals aged 65 and above consistently reported the highest life satisfaction and life worthwhileness scores. Gender differences reveal that females maintain a slightly higher life satisfaction score than males. New Zealand Europeans had higher scores in both life satisfaction and worthwhileness compared to most other ethnic groups. Regional variations are also evident, with residents in the Northland group consistently reporting slightly higher life satisfaction and worthwhileness scores than residents in other regions. Furthermore, a positive relationship between income and life satisfaction and life worthwhileness scores was observed, with higher income associated with high life satisfaction. Lastly, individuals residing in one-person households (i.e., living alone) consistently reported the lowest life satisfaction and worthwhileness scores, highlighting the potential influence of social connections on subjective wellbeing.

The trend analysis revealed a significant negative change in both life satisfaction and life worthwhileness between 2014 and 2018. Despite adjusting for year, most variable coefficients demonstrated minimal changes (when compared to the unadjusted models), suggesting the persistence of observed trends over time. However, the statistically significant interaction terms suggest that the association between life satisfaction and age varied across different years. In contrast, for life worthwhileness, none of the demographic variables show a significant interaction effect with year. This highlights the dynamic and multifaceted nature of the relationships between demographic factors and subjective wellbeing outcomes over the years, emphasising the need for detailed interpretations in understanding their evolving dynamics.

**Table 4-3**. Regression model results for the outcome Life satisfaction

| | | | Unadjusted | | | | Adjusted for Year | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **Variable category** | **N** | **B** | **SE** | **t value** | **p value** | **B** | **SE** | **t value** | **p value** | **Interaction p value** |
| **Year** | 2014 | 8802 | 7.70 | 0.02 | 445.77 | | | | | | |
| | 2016 | 8517 | 0.03 | 0.03 | 1.24 | **0.215** | | | | | |
| | 2018 | 8835 | -0.06 | 0.03 | -2.20 | 0.028 | | | | | |
| **Age range** | 15-24 years | 2727 | 7.73 | 0.03 | 222.47 | | 7.76 | 0.06 | 134.33 | | |
| | 25-34 years | 4107 | -0.11 | 0.04 | -2.40 | 0.016 | -0.24 | 0.08 | -3.12 | 0.002 | |
| | 35-44 years | 4221 | -0.22 | 0.04 | -4.95 | <0.001 | -0.26 | 0.08 | -3.51 | <0.001 | 0.034 |
| | 45-54 years | 4533 | -0.29 | 0.04 | -6.49 | <0.001 | -0.36 | 0.07 | -4.89 | <0.001 | |
| | 55-64 years | 4299 | -0.18 | 0.04 | -4.16 | <0.001 | -0.17 | 0.08 | -2.22 | 0.026 | |
| | 65 years or over | 6276 | 0.42 | 0.04 | 10.15 | <0.001 | 0.45 | 0.07 | 6.43 | <0.001 | |
| **Gender** | Male | 11856 | 7.65 | 0.02 | 454.91 | | 7.70 | 0.03 | 263.35 | | **0.068** |
| | Female | 14310 | 0.07 | 0.02 | 3.25 | 0.001 | 0.01 | 0.04 | 0.13 | **0.898** | |
| **Ethnicity** | New Zealand European | 18000 | 7.73 | 0.01 | 566.10 | | 7.74 | 0.02 | 332.93 | | |
| | New Zealand Māori | 2046 | -0.12 | 0.04 | -2.83 | 0.005 | -0.18 | 0.07 | -2.62 | 0.009 | **0.459** |
| | Pacific | 1173 | -0.07 | 0.06 | -1.26 | **0.207** | -0.09 | 0.10 | -0.91 | **0.362** | |
| | Asian | 2391 | 0.01 | 0.04 | 0.21 | **0.831** | -0.03 | 0.08 | -0.43 | **0.670** | |
| | MELAA / Other | 2541 | -0.20 | 0.04 | -5.24 | <0.001 | -0.18 | 0.07 | -2.56 | 0.010 | |
| **Region** | Auckland | 6780 | 7.66 | 0.02 | 344.45 | | 7.70 | 0.04 | 191/44 | | **0.398** |
| | Wellington | 3114 | 0.02 | 0.04 | 0.44 | **0.662** | -0.03 | 0.07 | -0.40 | **0.688** | |
| | Northland group | 3282 | 0.11 | 0.04 | 2.77 | 0.006 | 0.14 | 0.06 | 2.13 | 0.034 | |
| | Rest of North Island | 5958 | 0.05 | 0.03 | 1.51 | **0.132** | 0.00 | 0.06 | -0.07 | **0.948** | |
| | Canterbury | 3642 | -0.04 | 0.04 | -1.17 | **0.240** | -0.09 | 0.06 | -1.41 | **0.160** | |
| | Rest of South Island | 3393 | 0.09 | 0.04 | 2.21 | 0.027 | 0.04 | 0.07 | 0.56 | **0.573** | |
| **Personal income** | Loss/ $0 - $30,000 | 12069 | 7.59 | 0.02 | 456.21 | | 7.57 | 0.03 | 275.61 | | **0.451** |
| | $30,001 - $35,000 | 1404 | 0.06 | 0.05 | 1.12 | **0.262** | 0.11 | 0.09 | 1.24 | **0.214** | |
| | $35,001 - $40,000 | 1587 | 0.06 | 0.05 | 1.30 | **0.195** | 0.21 | 0.08 | 2.52 | 0.012 | |
| | $40,001 - $50,000 | 2637 | 0.07 | 0.04 | 1.76 | **0.078** | 0.16 | 0.07 | 2.40 | 0.016 | |
| | $50,001 - $60,000 | 2256 | 0.17 | 0.04 | 4.17 | <0.001 | 0.26 | 0.07 | 3.58 | <0.001 | |
| | $60,001 - $70,000 | 1695 | 0.28 | 0.05 | 5.96 | <0.001 | 0.32 | 0.09 | 3.67 | <0.001 | |
| | $70,001 - $100,000 | 2523 | 0.27 | 0.04 | 6.78 | <0.001 | 0.38 | 0.07 | 5.23 | <0.001 | |
| | $100,001 or more | 1992 | 0.42 | 0.04 | 9.52 | <0.001 | 0.52 | 0.08 | 6.18 | <0.001 | |
| **Household income** | Loss/ $0 - $30,000 | 5226 | 7.35 | 0.03 | 291.93 | | 7.36 | 0.04 | 184.81 | | **0.139** |
| | $30,001 - $35,000 | 1188 | 0.34 | 0.06 | 5.77 | <0.001 | 0.38 | 0.09 | 4.31 | <0.001 | |
| | $35,001 - $40,000 | 1080 | 0.23 | 0.06 | 3.70 | <0.001 | 0.16 | 0.11 | 1.40 | **0.162** | |
| | $40,001 - $50,000 | 2001 | 0.26 | 0.05 | 5.49 | <0.001 | 0.35 | 0.08 | 4.35 | <0.001 | |
| | $50,001 - $60,000 | 1932 | 0.24 | 0.05 | 5.00 | <0.001 | 0.21 | 0.08 | 2.58 | 0.010 | |
| | $60,001 - $70,000 | 1758 | 0.29 | 0.05 | 5.80 | <0.001 | 0.28 | 0.08 | 3.34 | 0.001 | |
| | $70,001 - $100,000 | 4701 | 0.42 | 0.04 | 11.36 | <0.001 | 0.44 | 0.06 | 7.21 | <0.001 | |
| | $100,001 or more | 8271 | 0.59 | 0.03 | 18.47 | <0.001 | 0.64 | 0.05 | 11.77 | <0.001 | |
| **Household size** | One person | 6474 | 7.47 | 0.02 | 329.13 | | 7.53 | 0.04 | 194.43 | | |
| | Two people | 9021 | 0.41 | 0.03 | 13.89 | <0.001 | 0.33 | 0.05 | 6.42 | <0.001 | **0.363** |
| | Three people | 4071 | 0.12 | 0.04 | 3.18 | 0.001 | 0.04 | 0.06 | 0.61 | **0.545** | |
| | Four people | 3855 | 0.25 | 0.04 | 6.72 | <0.001 | 0.17 | 0.06 | 2.62 | 0.009 | |
| | Five people | 1680 | 0.33 | 0.05 | 6.68 | <0.001 | 0.24 | 0.09 | 2.73 | 0.006 | |
| | Six or more people | 1056 | 0.26 | 0.06 | 4.22 | <0.001 | 0.26 | 0.11 | 2.45 | 0.014 | |

***Note:*** *All p-values ($p > 0.05$) are highlighted in bold. The interaction p-value indicates the statistical significance of the interaction effect between each demographic variable and the survey year.*

**Table 4-4.** Regression model results for the outcome Life worthwhileness

| | | | Unadjusted | | | | Adjusted for Year | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **Variable category** | **N** | **B** | **SE** | **t value** | **p value** | **B** | **SE** | **t value** | **p value** | **Interaction p value** |
| **Year** | 2014 | 8781 | 8.09 | 0.02 | 503.73 | | | | | | |
| | 2016 | 8511 | 0.02 | 0.03 | 0.86 | **0.390** | | | | | |
| | 2018 | 8826 | -0.05 | 0.03 | -2.13 | 0.033 | | | | | |
| **Age range** | 15-24 years | 2724 | 7.86 | 0.03 | 243.91 | | 8.93 | 0.05 | 147.83 | | **0.588** |
| | 25-34 years | 4104 | 0.15 | 0.04 | 3.55 | <0.001 | 0.05 | 0.07 | 0.70 | **0.484** | |
| | 35-44 years | 4218 | 0.15 | 0.04 | 3.72 | <0.001 | 0.06 | 0.07 | 0.80 | **0.424** | |
| | 45-54 years | 4527 | 0.07 | 0.04 | 1.77 | **0.077** | 0.01 | 0.07 | 0.20 | **0.838** | |
| | 55-64 years | 4302 | 0.20 | 0.04 | 4.86 | <0.001 | 0.17 | 0.07 | 2.36 | 0.018 | |
| | 65 years or over | 6249 | 0.51 | 0.04 | 13.27 | <0.001 | 0.48 | 0.06 | 7.38 | <0.001 | |
| **Gender** | Male | 11841 | 7.96 | 0.02 | 513.03 | | 8.00 | 0.03 | 296.63 | | **0.114** |
| | Female | 14283 | 0.22 | 0.02 | 10.36 | <0.001 | 0.16 | 0.04 | 4.51 | <0.001 | |
| **Ethnicity** | New Zealand European | 17982 | 8.12 | 0.01 | 643.54 | | 8.13 | 0.02 | 378.2 | | **0.277** |
| | New Zealand Māori | 2043 | -0.03 | 0.04 | -0.85 | **0.397** | -0.10 | 0.06 | -1.60 | **0.109** | |
| | Pacific | 1173 | -0.13 | 0.05 | -2.56 | 0.010 | -0.11 | 0.09 | -1.22 | **0.221** | |
| | Asian | 2385 | -0.17 | 0.04 | -4.62 | <0.001 | -0.25 | 0.07 | -3.52 | <0.001 | |
| | MELAA / Other | 2535 | -0.12 | 0.04 | -3.32 | 0.001 | -0.06 | 0.06 | -0.90 | 0.367 | |
| **Region** | Auckland | 6771 | 7.97 | 0.02 | 387.89 | | 7.94 | 0.04 | 213.99 | | **0.616** |
| | Wellington | 3105 | 0.09 | 0.04 | 2.49 | 0.013 | 0.14 | 0.06 | 2.28 | 0.023 | |
| | Northland group | 3282 | 0.22 | 0.04 | 6.15 | <0.001 | 0.27 | 0.06 | 4.45 | <0.001 | |
| | Rest of North Island | 5943 | 0.19 | 0.03 | 6.34 | <0.001 | 0.24 | 0.06 | 4.39 | <0.001 | |
| | Canterbury | 3639 | 0.08 | 0.03 | 2.16 | 0.031 | 0.09 | 0.06 | 1.56 | **0.120** | |
| | Rest of South Island | 3387 | 0.15 | 0.04 | 4.33 | <0.001 | 0.21 | 0.06 | 3.42 | 0.001 | |
| **Personal income** | Loss/ $0 - $30,000 | 12048 | 7.98 | 0.02 | 518.78 | | 7.99 | 0.03 | 314.74 | | **0.970** |
| | $30,001 - $35,000 | 1404 | 0.11 | 0.05 | 2.21 | 0.027 | 0.06 | 0.08 | 0.82 | **0.411** | |
| | $35,001 - $40,000 | 1584 | 0.14 | 0.05 | 3.20 | 0.001 | 0.17 | 0.08 | 2.16 | 0.031 | |
| | $40,001 - $50,000 | 2631 | 0.07 | 0.04 | 2.03 | 0.042 | 0.10 | 0.06 | 1.59 | 0.111 | |
| | $50,001 - $60,000 | 2256 | 0.16 | 0.04 | 4.23 | <0.001 | 0.22 | 0.07 | 3.18 | 0.001 | |
| | $60,001 - $70,000 | 1695 | 0.22 | 0.04 | 5.03 | <0.001 | 0.20 | 0.08 | 2.51 | 0.012 | |
| | $70,001 - $100,000 | 2517 | 0.22 | 0.04 | 5.91 | <0.001 | 0.22 | 0.07 | 3.27 | 0.001 | |
| | $100,001 or more | 1986 | 0.39 | 0.04 | 9.52 | <0.001 | 0.44 | 0.08 | 5.74 | <0.001 | |
| **Household income** | Loss/ $0 - $30,000 | 5211 | 7.83 | 0.02 | 333.51 | | 7.85 | 0.04 | 212.46 | | **0.531** |
| | $30,001 - $35,000 | 1185 | 0.24 | 0.05 | 4.51 | <0.001 | 0.27 | 0.08 | 3.31 | 0.001 | |
| | $35,001 - $40,000 | 1074 | 0.21 | 0.06 | 3.79 | <0.001 | 0.09 | 0.10 | 0.92 | **0.359** | |
| | $40,001 - $50,000 | 1995 | 0.22 | 0.04 | 4.94 | <0.001 | 0.27 | 0.07 | 3.66 | <0.001 | |
| | $50,001 - $60,000 | 1935 | 0.22 | 0.04 | 4.91 | <0.001 | 0.23 | 0.08 | 2.99 | 0.003 | |
| | $60,001 - $70,000 | 1758 | 0.17 | 0.05 | 3.59 | <0.001 | 0.16 | 0.08 | 2.13 | 0.034 | |
| | $70,001 - $100,000 | 4701 | 0.29 | 0.03 | 8.41 | <0.001 | 0.32 | 0.06 | 5.63 | <0.001 | |
| | $100,001 or more | 8262 | 0.41 | 0.03 | 13.88 | <0.001 | 0.42 | 0.05 | 8.30 | <0.001 | |
| **Household size** | One person | 6450 | 7.90 | 0.02 | 375.75 | | 7.95 | 0.04 | 221.26 | | **0.642** |
| | Two people | 9021 | 0.29 | 0.03 | 10.50 | <0.001 | 0.24 | 0.05 | 5.18 | <0.001 | |
| | Three people | 4065 | 0.16 | 0.03 | 4.68 | <0.001 | 0.14 | 0.06 | 2.37 | 0.018 | |
| | Four people | 3858 | 0.22 | 0.03 | 6.53 | <0.001 | 0.12 | 0.06 | 1.94 | **0.053** | |
| | Five people | 1674 | 0.24 | 0.05 | 5.28 | <0.001 | 0.22 | 0.08 | 2.71 | 0.007 | |
| | Six or more people | 1059 | 0.21 | 0.06 | 3.80 | <0.001 | 0.22 | 0.10 | 2.18 | 0.029 | |

***Note:*** *All p-values ($p > 0.05$) are highlighted in bold. The interaction p-value indicates the statistical significance of the interaction effect between the demographic variables and the survey year.*

**Discussion**

This primary aim of this study was to investigate the relationship between subjective wellbeing outcomes (life satisfaction and life worthwhileness) and various sociodemographic variables within the New Zealand context. We utilised repeated cross-sections of GSS data to explore how these relationships vary over time. Our findings reaffirm some established trends such as higher life satisfaction and worthwhileness among older adults, females, and New Zealand Europeans, as well as those with higher incomes, consistent with prior research. Additionally, we observed significant declines in life satisfaction and worthwhileness between 2014 and 2018, highlighting the dynamic nature of individuals' perspectives on wellbeing.

The trend analyses revealed some consistent patterns, such as the elevated life satisfaction scores among older individuals, but also identified some notable variations. For example, 2018 saw a decrease in both life satisfaction and life worthwhileness compared to 2014. Additionally, only the demographic variable ‘age’ demonstrated varying associations with life satisfaction across the survey years. These variations might be indicative of the complex interplay between external factors, life events, and individual adaptability, contributing to the dynamic nature of subjective wellbeing [196]. Collectively, these insights not only deepen our understanding of the current state of subjective wellbeing in New Zealand, but also provide valuable information for identifying variables that are sensitive to capturing variations in wellbeing within the population. Understanding these factors are also essential for developing subsequent studies that aim to predict wellbeing measures based on various administrative variables.

The analysis of life satisfaction revealed a clear age-related pattern, with individuals aged 65 and above consistently reporting the highest scores. Young people, those aged 15 to 24 years, also exhibit higher life satisfaction scores compared to their middle-aged counterparts (aged 25 to 64 years). This finding closely aligns with some established studies, noting a moderate U-

shaped trend in psychological wellbeing and life satisfaction, where scores tend to increase after the age of 50 years [197-199]. The longitudinal examination of our data supports and reinforces the association between age and life satisfaction, highlighting the persistence of this trend over time. A recent study focusing on life satisfaction in New Zealand similarly demonstrated a modest U-shaped relationship with age [199]. However, it is crucial to acknowledge varying perspectives within the research landscape. Some researchers strongly disagree with the simplistic U-shaped pattern between age and life satisfaction, arguing that this relationship is far more complex [200, 201]. A recent meta-analytic review of longitudinal studies revealed a more intricate trajectory, indicating a decline in life satisfaction between the ages of 9 and 16 years, followed by a gradual increase until around the age of 70 years. Surprisingly, a subsequent decrease occurred after the age of 70 years, continuing until the age of 96 years [202]. This decline in life satisfaction after 70 years might be attributed to health issues and a decrease in social connections. In the context of our analysis, it is important to note that our analyses did not include children (below 15 years), and people older than 65 years were grouped together. Given that the primary aim of our study was not to specifically investigate age and life satisfaction, it is beyond our scope to delve deeper into this aspect and comment on the relationship. Future research could provide additional clarity and insights into the complex relationship between age and life satisfaction.

Similarly, the exploration of life worthwhileness scores aligns with the pattern observed in life satisfaction. Older adults, specifically those aged 65 years and above, consistently report higher scores. Notably, unlike life satisfaction, we did not identify a U-shaped pattern in life worthwhileness. Instead, our findings indicate modest variations with age. Over the years, individuals aged 18 to 55 years did not exhibit significant differences in their life worthwhileness scores, while those aged 55 years and above consistently reported significantly higher scores. The observed differences in age-related patterns between life satisfaction and

life worthwhileness scores could be attributed to various reasons such as changing life priorities, evolving perspectives on what makes life worthwhile, or the impact of external factors on these two distinct aspects of subjective wellbeing. Additionally, societal and cultural factors may also play a role in shaping perceptions of life satisfaction and worthwhileness, influencing individuals' responses across different age groups. Future research could delve deeper into these subtle aspects to explore the intricate relationship between age and different dimensions of subjective wellbeing.

Our findings underscore that females consistently demonstrate slightly higher life satisfaction and life worthwhileness scores compared to males, aligning with established patterns observed in prior research spanning diverse cultures and age groups [28, 203]. The topic of gender differences in subjective wellbeing has been a subject of prolonged debate, marked by numerous studies presenting inconsistent results [204-207]. A recent meta-analysis delving into gender differences in subjective wellbeing concluded that there are no discernible differences in life satisfaction between men and women [204]. Some studies propose a connection between societal gender inequality and gender differences in subjective wellbeing [208]. Interestingly, a recent study challenged the notion of inherent gender differences, suggesting that the perceived disparity may stem from the use of different response scales by women and men when reporting their happiness. When these scales are normalised, women appear less happy than men on average [209].

Upon examining trends over the years, our analysis did not identify any significant differences between females and males in life satisfaction scores (although life worthwhileness scores still exhibited significant differences). This observation suggests that while a gender difference in life satisfaction is evident in cross-sectional analyses, this difference diminishes over time, possibly influenced by evolving societal dynamics or other contextual factors. The overall relationship between gender and life satisfaction exhibited variations across the years, as

indicated by the significant interaction effect, suggesting that the gender differences present in the cross-sectional analysis were not consistently observed over time.

The comparison of life satisfaction scores between different ethnicities in New Zealand reveals distinctive patterns. When compared to New Zealand Europeans, individuals of Pacific and Asian ethnicity did not exhibit significantly different life satisfaction scores. However, people of Māori ethnicity and those of Middle Eastern, Latin American, and African (MELAA) ethnicity consistently reported lower life satisfaction scores. Contrastingly, the analysis of life worthwhile scores showed different dynamics. There were no significant differences in the scores between individuals of New Zealand European and Māori ethnicity. However, all other ethnic groups demonstrated lower life worthwhile scores than New Zealand Europeans. The disparity in life satisfaction and life worthwhileness among the Māori could be attributed to a combination of cultural, socio-economic, and historical factors [210, 211]. Research also indicates that Māori have experienced persistent health and wellbeing inequities, often linked to the long-term effects of colonization and socio-economic disadvantages [212]. The distinct constructs of life satisfaction (more externally influenced by factors like income, standard of living), and life worthwhileness (more internally driven) may contribute to these nuanced results. Past research has highlighted the significance of a secure cultural identity derived from cultural and social connections for Māori mental wellbeing [213]. The New Zealand Mental Health Monitor and the Health and Lifestyles Survey indicated that 77% of Māori respondents were satisfied with their lives, while an even higher proportion, 86%, considered their lives to be worthwhile [213]. Cultural values and identity likely play a pivotal role in Māori individuals feeling of life worthwhile despite challenges impacting general life satisfaction. Resilience, coping mechanisms, and community support within minority groups may further contribute to maintaining a strong sense of purpose amidst adversities.

There was no statistically significant difference in life satisfaction scores between individuals residing in Auckland and those in Wellington, the rest of Northland, and Canterbury. Notably, residents of the Northland group, encompassing Northland, Gisborne, and Bay of Plenty, consistently report slightly higher life satisfaction scores than Auckland residents. This pattern does not vary over the years indicating a stable trend in regional life satisfaction. Conversely, for life worthwhileness, individuals residing outside of Auckland consistently report higher scores compared to Auckland residents, with the Northland group residents reporting the highest life worthwhile scores when compared to other regions. This trend remains consistent over the years, suggesting a lasting pattern of higher life worthwhile scores among residents in areas other than Auckland. The only deviation from this overall trend is observed in Canterbury, where residents do not conform to the general pattern seen in other regions. These findings highlight nuanced regional variations in subjective wellbeing, with the Northland group consistently standing out for higher life satisfaction and life worthwhile scores. These variations could be attributed to the influence of distinct regional factors on subjective wellbeing across New Zealand, emphasising the need for further exploration and investigation into the specific dynamics at play in each region.

The observed regional variations in life satisfaction and life worthwhileness in New Zealand could be attributed to the influence of distinct regional factors, each contributing to the unique experiences of residents across different areas. Prior work has underscored the significance of geographic location in influencing one's wellbeing [214]. Findings from the Australian Unity Wellbeing Index indicate that individuals residing in regional or rural areas tend to have better wellbeing compared to their urban counterparts, driven by higher satisfaction with community connectedness and personal safety [215]. Regions with robust economies may exhibit higher life satisfaction due to greater financial stability. However, variations in lifestyle, such as the pace of life and stress levels, particularly in urban areas like Auckland, could contribute to

differing levels of wellbeing. Regions offering a more relaxed lifestyle with greater access to natural environments, green spaces, and reduced pollution could potentially enhance people's life satisfaction and the perception that life is worthwhile.

The impact of economic factors on subjective wellbeing has been a subject of ongoing debate among researchers [37]. While some studies suggest a positive association between high incomes and increased life satisfaction [216-218], others contend that the relationship between income and life satisfaction is relatively weak [219]. Our findings reveal that individuals with higher incomes tend to have higher life satisfaction and life worthwhile scores; however, an interesting trend emerges when exploring deeper into the income brackets. Notably, life satisfaction scores among individuals with personal incomes below $50,000 do not exhibit significant differences. For instance, there is no notable distinction between individuals earning $20,000 and those earning $40,000, despite the income disparity. However, individuals earning over $50,000 demonstrate significantly higher life satisfaction scores than those earning less than $30,000. This trend, however, is not consistent over time, as observed in the longitudinal analysis. Over the years, individuals earning just above $35,000 tend to have better life satisfaction and life worthwhile scores compared to those earning $30,000 or less. Individuals earning $100,000 and above display the highest overall life satisfaction and worthwhile scores consistently. Higher-income households often enjoy a better quality of life, including improved access to housing and healthcare, contributing to improved evaluation of life [220]. However, some studies suggest a limited or even non-existent long-term correlation, labelling the income-happiness connection as a paradox [221-223]. Nevertheless, it is crucial to acknowledge that a higher income does not inherently guarantee higher wellbeing, as individual perceptions and other factors such as social connections, health status and other life circumstances play pivotal roles in this complex relationship [180].

Lastly, numerous studies underscore the significance of social connections and their positive impact on subjective wellbeing [180, 224]. Living alone was associated with the lowest life satisfaction and life worthwhile scores, as revealed by our analyses on household size. However, it is important to note that living alone does not necessarily indicate loneliness; instead, it reflects individuals' choices influenced by various factors like personal preferences, lifestyle, or other circumstances. Next, people in two-person households, often indicative of couples, consistently exhibit the highest scores. This observed trend might be attributed to the potential emotional and social support inherent in close relationships, contributing to a greater sense of life satisfaction and perceived worthwhileness. The shared responsibilities, companionship, and mutual support within two-person households may foster a more favourable environment for overall subjective wellbeing.

While our study provided valuable insights into the wellbeing climate in New Zealand by exploring the associations between various demographic variables and subjective wellbeing outcomes, it is essential to acknowledge certain limitations. Firstly, our study was limited in providing detailed explanations for the observed relationships or delving deeper into the underlying mechanisms related to the wellbeing construct, as such in-depth analysis was beyond its scope. Secondly, while our investigation focused on a concise set of demographic variables, it's important to acknowledge that the exclusion of additional factors, such as employment status or educational background, may limit the comprehensiveness of our findings. These additional variables could potentially provide valuable insights and context to better understand subjective wellbeing among different demographic groups. Therefore, future studies may benefit from including a broader range of variables to capture a more comprehensive picture of the factors influencing subjective wellbeing. Lastly, our study specifically focused on two subjective wellbeing outcomes: life satisfaction and life worthwhileness. We were unable to include the other two subjective wellbeing outcomes

(family wellbeing and mental wellbeing) in our analysis due to their non-availability in more than one wave of the GSS. However, as future waves of the GSS become available in the IDI, researchers could delve deeper into understanding how these additional variables vary across socio-demographic groups and over time.

**Conclusion**

This study offered a comprehensive exploration of subjective wellbeing in New Zealand, highlighting the significant roles of age, gender, ethnicity, region, income, and household size. While reinforcing established trends such as higher life satisfaction among older individuals and slight gender differences, it also revealed small changes in these associations over time. This reflects the dynamic nature of subjective wellbeing, and how it can be impacted by external factors and individual adaptability. Ethnicity emerged as a significant factor associated with subjective wellbeing, with distinct patterns among different ethnic groups, hinting at potential cultural and socio-economic influences. Regional variations further highlighted the potential role of geographic location and lifestyle factors in shaping wellbeing outcomes. Nevertheless, the study also acknowledges its limitations of not contributing to a deeper understanding of the underlying mechanisms between these observed relationships in subjective wellbeing. Overall, this research provides valuable insights into the multifaceted nature of subjective wellbeing, emphasising its intricate and multifaceted nature.

# Chapter 5 – A cross-validation study to investigate the efficacy of census-level socio-demographic factors for predicting subjective wellbeing outcomes in New Zealand

## Preface

The preceding chapter offered valuable insights into the diverse wellbeing experiences of New Zealanders across various demographic groups and regions. In this chapter, we develop and validate statistical models capable of predicting different GSS-based subjective wellbeing outcomes. This primarily involves leveraging population-level administrative variables, such as socio-demographic factors available in the census dataset, alongside meshblock-level environmental variables. Ultimately, the developed model will enable us to extrapolate wellbeing predictions to the broader population, aligning with the overarching goal of this thesis.

## Introduction

The significance of population wellbeing is gaining widespread recognition globally, prompting governments to broaden their evaluative criteria beyond the traditional measure of GDP (Gross Domestic Product) to assess the overall success of their population [23]. While GDP and productivity measures continue to be central for policymaking, there is an emerging shift towards a more comprehensive approach that includes the assessment of wellbeing. Initiatives like the Wellbeing Economy Governments partnership (WEGo) exemplify this shift, where national and regional governments collaboratively advance the concept of Wellbeing Economies [225]. Despite sustained economic growth, New Zealand faces pressing challenges such as high rates of child poverty, homelessness, and suicide. In response, the government introduced its inaugural 'wellbeing budget' in 2019 [46], signifying a renewed commitment to prioritising people's wellbeing alongside economic growth.

Understanding wellbeing presents challenges due to the evolving nature and diverse perspectives around its meaning. Initially, wellbeing was often perceived as positive human functioning, referred to as "eudaimonia," encompassing aspects such as self-actualisation and autonomy [29]. Other researchers have integrated eudaemonic and hedonic components, combining aspects of functioning and emotions [32]. For example, Diener's tripartite model identified cognitive, positive affect, and negative affect components [28], while Seligman's PERMA model introduced positive emotion, engagement, relationships, meaning, and accomplishment as key dimensions [33]. Thompson et al.'s dynamic model of 'flourishing' further highlights the interplay between positive feelings, effective functioning, external conditions, and personal resources [35]. This comprehensive perspective suggests that 'flourishing' or elevated wellbeing emerges from the interplay of positive emotions and effective functioning within an individual's unique circumstances and available resources. Thus, a 'flourishing' nation indicates elevated wellbeing among its citizens.

The increasing significance of incorporating wellbeing indicators into policy decisions is becoming more prominent in New Zealand. Despite this growing importance, there still exists a considerable gap in our understanding of the factors that influence population wellbeing in the country. This knowledge gap is partially attributed to the scarcity of detailed, population-level wellbeing data. The New Zealand General Social Survey (GSS), a biennial survey of around 9,000 individuals [47], offers wellbeing data across twelve domains: health, housing, income and consumption, jobs and earnings, leisure and free time, knowledge and skills, safety and security, social connections, cultural identity, civic engagement and governance, environmental quality, and subjective wellbeing. Designed based on the New Zealand Living Standards Framework [10], which in turn was drawn from the OECD's framework [43], the GSS lays the foundation for wellbeing assessment in New Zealand. In the context of this study, we focus primarily on the subjective wellbeing domain, focussing on indicators such as life satisfaction, sense of purpose, family wellbeing and mental wellbeing.

Although the GSS sample is considered nationally representative, certain subgroups of the population (that may be of significant policy interest) remain underrepresented due to limitations in sample size. For instance, it is impractical to understand the wellbeing experiences of individuals living in government-sponsored social housing. This is because the number of people who participated in the GSS and are also residents of social housing may be very small. Therefore, to assess the impact of government initiatives targeting this specific population sub-group, comprehensive wellbeing measures applicable to the entire population are needed.

To address this challenge, two strategies offer potential solutions. One approach involves collecting regular wellbeing data for the entire population in a census activity; however, this method is resource-intensive and time-consuming. An alternative approach involves leveraging

existing routinely collected data to extrapolate GSS wellbeing measures to the broader population. This may be feasible due to New Zealand's Integrated Data Infrastructure (IDI): a complex database managed by Stats NZ [103]. The IDI contains individual response data (microdata) on people and households, supplemented with anonymised information on education, income, health, justice, and housing. Notably, the IDI facilitates dataset linkage across these areas using a unique identifier variable. Details about this linking process are available elsewhere [226]. Crucially, the IDI houses the GSS data, allowing linkage with the country's Census data which the majority of the nation's population completes (given it is a legal requirement to do so).

The Census is a comprehensive nationwide survey conducted once every five years in New Zealand, with the primary aim of officially counting individuals and households in the country [227]. It also provides a snapshot of various aspects of life, including demographic information, educational qualifications, employment status, and more. Additionally, the Census gathers data on addresses for each household, which are then aggregated at the meshblock level for reporting purposes. A meshblock represents the smallest administrative geographical unit, typically encompassing about 30 to 60 households [68]. Environmental data, such as the extent of green spaces, are also available at the meshblock level and can therefore be linked to the Census data. One notable example is the Healthy Location Index, which captures accessibility of health-promoting environmental features (e.g., green spaces, physical activity facilities) and health-constraining environmental features (e.g., alcohol outlets, fast-food shops) [228]. The ability to link such key environmental information to the Census is crucial, given the established links between the environment and wellbeing [63, 64].

The aim of this study is to predict GSS-derived wellbeing measures from Census-based sociodemographic information and meshblock-level environmental indicators. If successful, such a model could be used to extrapolate these predicted wellbeing scores to the entire IDI

population, thereby creating a population-level estimate of subjective wellbeing. This could yield transformative benefits by facilitating the integration of wellbeing metrics into policy analysis. It also holds the potential to significantly enhance our understanding of how the political, social, and economic landscape impacts the wellbeing and overall functioning of individuals in New Zealand. This could further empower decision-makers to formulate more informed, targeted, and effective policies that address the genuine needs and concerns of New Zealanders.

## Methods

### *Data Sources*

The data used in this study were sourced from three datasets: GSS [47], New Zealand Census of Population and Dwellings [227], and the Healthy Location Index (HLI) [67]. Of these, two are present in the IDI, namely the GSS and the Census. All datasets within the IDI are structured as tables in an SQL database and can be linked with one another using the Stats NZ unique identifier variable [72]. All datasets within the IDI can be accessed only from a Stats NZ data laboratory. A formal application to access the IDI datasets, and the IDI data laboratory was submitted and approved by Stats NZ. The methodology used in this research was approved by the AUT University Ethics Committee (AUTEC #21/115).

The study utilised GSS data during the 2018 year, with a sample size of 8,793. More information regarding the GSS and its data collection methodology can be found elsewhere [48, 190]. The wellbeing outcome variables, unique identifier variable (snz_uid) and the meshblock_code variable was selected from the GSS. The subjective wellbeing outcome variables investigated in this study are listed in Table 5-1.

**Table 5-1.** GSS wellbeing outcome measures

| Outcome variable | Survey question in the GSS | Range | Description |
|---|---|---|---|
| 1. Life satisfaction | Where zero is completely dissatisfied, and ten is completely satisfied, how do you feel about your life as a whole? | 0 – 10 | Completely dissatisfied – Completely satisfied |
| 2. Life worthwhileness | Where zero is not at all worthwhile, and ten is completely worthwhile, overall, to what extent do you feel the things you do in your life are worthwhile? | 0 – 10 | Not at all worthwhile – Completely worthwhile |
| 3. Family wellbeing | Where zero means extremely badly and ten means extremely well, how would you rate how your family is doing these days? | 0 – 10 | Extremely badly – Extremely well |
| 4. Mental wellbeing | Derived variable, this variable is based on WHO-5's wellbeing index score [229, 230] | 0 – 100 | Excellent – Poor |

Next, the Census 2018 dataset was utilised in this study. Further details about the Census and its methodology are available elsewhere [231]. The size of the dataset was approximately 4.9 million observations with over 300 variables, of which 29 demographic variables were selected as predictors. The choice of these variables was guided by their availability for most of the population. To enhance interpretability, some variables were consolidated into fewer categories due to low counts in some specific categories. Table 5-2 shows the full list of demographic variables used in the study.

**Table 5-2.** Predictor variables from the Census 2018 dataset

| Predictor variable | Type of variable | Number of categories | Description |
|---|---|---|---|
| 1. Age (in years) | Continuous | NA | 0 to 120 |
| 2. Gender | Categorical | 2 | Male or Female |
| 3. Ethnicity | Categorical | 5 | European,<br>New Zealand Māori,<br>Pacific,<br>Asian,<br>Middle Eastern/Latin American/African and Other Ethnic groups |
| 4. Region | Categorical | 6 | Auckland,<br>Wellington,<br>Northland group (Northland, Bay of Plenty, Gisborne),<br>Rest of North Island,<br>Canterbury,<br>Rest of South Island |
| 5. Marital Status | Categorical | 5 | Married (not separated),<br>Separated,<br>Divorced or dissolved,<br>Widowed or surviving civil union partner,<br>Never married and never in a civil union |

| | | | | |
|---|---|---|---|---|
| 6. | Birth Country | Categorical | 2 | New Zealand, Other |
| 7. | Highest Qualification | Categorical | 8 | No Qualification,<br>School Qualification,<br>Post-school Qualification,<br>Bachelor's degree and Level 7 Qualification,<br>Post-graduate and Honours Degrees,<br>Master's Degree,<br>Doctorate Degree,<br>Overseas Secondary School Qualification |
| 8. | Personal Income | Categorical | 9 | $0 - $30,000<br>$30,001 - $35,000<br>$35,001 - $40,000<br>$40,001 - $50,000<br>$50,001 - $60,000<br>$60,001 - $70,000<br>$70,001 - $100,000<br>$100,001-$150,000,<br>$150,001 or More |
| 9. | Household Income | Categorical | 9 | Same as Personal Income |
| 10. | Number of income sources | Categorical | 5 | No source of income,<br>One source,<br>Two sources,<br>Three sources,<br>Four sources,<br>Five or more sources, |
| 11. | Workforce Status | Categorical | 4 | Employed Full-time,<br>Employed Part-time,<br>Unemployed,<br>Not in the Labour Force |
| 12. | Study Participation Code | Categorical | 3 | Full-time study,<br>Part-time study,<br>Not studying |
| 13. | Number of Languages spoken | Categorical | 7 | None,<br>One Language,<br>Two Languages,<br>Three Languages,<br>Four Languages,<br>Five Languages,<br>Six Languages |
| 14. | Home Ownership | Categorical | 3 | Hold in a family trust,<br>Own or partly own,<br>Do not own and do not hold in a family trust |
| 15. | Index of Socioeconomic Deprivation Score 2018 [232] | Continuous | Derived variable | 823 – 1552 |
| 16. | Index of Socioeconomic Deprivation 2018 [232] | Categorical | 10 | 1 – Least deprived<br>10 – Most deprived |
| 17. | Dwelling dampness indicator | Categorical | 4 | Always damp,<br>Sometimes damp,<br>Not damp,<br>Don't know |
| 18. | Dwelling mould indicator | Categorical | 4 | Mould over A4 size - always,<br>Mould over A4 size - sometimes,<br>No mould/mould smaller than A4 size,<br>Don't know |
| 19. | Difficulty in Seeing | Categorical | 4 | No difficulty,<br>Some difficulty,<br>A lot of difficulty,<br>Cannot do at all |

| | | | | |
|---|---|---|---|---|
| 20. | Difficulty in Hearing | Categorical | 4 | Same as above |
| 21. | Difficulty in Washing | Categorical | 4 | Same as above |
| 22. | Difficulty in Communication | Categorical | 4 | Same as above |
| 23. | Difficulty in Remembering | Categorical | 4 | Same as above |
| 24. | Difficulty in Walking | Categorical | 4 | Same as above |
| 25. | Disability indicator | Categorical | 2 | Not disabled,<br>Disabled |
| 26. | Crowding Code – Based on Canadian National Occupancy Standard | Categorical | 5 | 2+ beds needed,<br>1 bed needed,<br>no beds needed,<br>1 bed spare,<br>2+ beds spare |
| 27. | Cigarette smoking behaviour | Categorical | 3 | Regular Smoker,<br>Ex-Smoker,<br>Never Smoked Regularly |
| 28. | Have you ever smoked? | Categorical | 2 | Yes or no |
| 29. | Do you smoke regularly? | Categorical | 2 | Yes or no |

Lastly, data related to the environment was acquired from the Healthy Location Index (HLI) dataset [67]. As this dataset is not present in the IDI, it was imported into the IDI data environment by Stats NZ. The HLI data provides a rank (ranging between 1 and 52,593) for every New Zealand meshblock (excluding oceanic meshblocks). This ranking is determined based on the accessibility of each meshblock (i.e., distance proximity) to both health-promoting features of the environment (e.g., physical activity facilities) and health-constraining features of the environment (e.g., fast-food outlets, takeaway outlets). The methodology involves a straightforward assignment of ranks, offering a transparent depiction of how each meshblock compares in terms of accessibility to these environmental factors. More details about this dataset and the methodology involved in developing this measure can be found elsewhere [228]. A total of 13 environmental variables (shown in Table 5-3) were used as predictors in this study. All these variables were measured in deciles, ranging from 1 (indicating the highest decile and closest proximity to the environmental feature) to 10 (representing the lowest decile and the farthest distance from the environmental feature).

**Table 5-3.** Environment related variables from the Healthy Location dataset (HLI)

| Variable | Variable description | Range |
|---|---|---|
| 1. FruitVeg_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest Fruit and Veg shop in meters) | 1 – 10 |
| 2. Supermarket_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest supermarket in meters) | 1 – 10 |
| 3. PhysicalActivity_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest physical activity area in meters) | 1 – 10 |
| 4. Greenspace_rank_dec | Decile of the meshblock rank (ranked based on the median proximity to greenspace in meters) | 1 – 10 |
| 5. Bluespace_rank_dec | Decile of the meshblock rank (ranked based on the median proximity to bluespace area in meters) | 1 – 10 |
| 6. Goods_dec | Decile of the sum of meshblock ranks of all environmetal goods (listed above – 1 to 5) | 1 – 10 |
| 7. FastFood_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest fast food in meters) | 1 – 10 |
| 8. Takeaways_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest takeaways in meters) | 1 – 10 |
| 9. DairyConvenienc_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest dairy/convenience store in meters) | 1 – 10 |
| 10. AlcoholOutlets_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest alcohol outlet in meters) | 1 – 10 |
| 11. GamingVenues_rank_dec | Decile of the meshblock rank (ranked based on the distance to the closest gaming venues in meters) | 1 – 10 |
| 12. Bads_dec | Decile of the sum of meshblock ranks of all environmetal bads (listed above – 7 to 11) | 1 – 10 |
| 13. Env_dec | Decile of the sum of meshblock ranks based on access to access to blue- and greenspace (listed above – 4 and 5) | 1 – 10 |

The GSS dataset was linked with the Census using the unique identifier variable (snz_uid) and to the HLI dataset using the meshblock number. After linking these, the dataset underwent a cleaning process to ensure data quality and consistency. Any observations with missing values

were removed from the dataset ($n$ = 3,135). Unknown or "did not answer" categories in the variables were removed resulting in a data with 5,658 observations and 42 predictor variables (29 Census variables + 13 HLI variables). The demographic distribution of the final dataset (shown in Appendix Table A-4) closely resembles that of the GSS 2018 dataset, indicating a balanced representation of most of the demographic sub-groups without any noticeable over- or under-representation.

***Modelling***

The development of precise predictive models is pivotal in extrapolating GSS data to the broader population. A robust predictive model assists in uncovering patterns within the dataset and establishes a solid foundation for reliable extrapolation. In this study, we employed three distinct predictive models with varying degrees of complexity: (1) Stepwise Linear Regression, (2) Elastic Net Regression, and (3) Random Forest. The modelling process described below was repeated for each of the four wellbeing outcome variables separately (life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing). These models were chosen due to the substantial number of predictor variables ($n$ = 42), and their inherent variable selection properties. Furthermore, assessing their relative performance on the dataset allows us to evaluate whether more complex models have significantly better predictive capability for our outcomes of interest. It also allows us to explore the utility of machine learning techniques in contrast to traditional modelling methods.

To begin, the Stepwise Linear Regression method was utilised. It employed an iterative process to select and remove predictor variables based on their statistical significance, ultimately yielding a subset of relevant variables [233]. While not entirely random, the variable selection process is automated, making it suitable for situations where there are numerous potential predictors. The order of variable selection is determined through statistical criteria rather than

random selection. For more detailed information on this model, please refer to Draper and Smith (1998) [234]. Next, we incorporated the Elastic Net Regression model to evaluate its predictive performance in comparison to the Stepwise method. Elastic Net regression provides a unique set of advantages over other regression methods as it is a combination of both Lasso (L1) and Ridge (L2) regularisation techniques [235]. This combination facilitates automatic variable selection, enhanced model interpretability and reducing overfitting, making it particularly well-suited for regression tasks involving high-dimensional data [235]. Lastly, we introduced a Random Forest model, to compare its performance against the traditional regression models. The Random Forest is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions to enhance accuracy and reduce overfitting [154]. The Random Forest is effective at handling high-dimensional data as it has inbuilt variable selection, and can capture complex non-linear relationships between variables more effectively than traditional linear regression techniques [236].

All models were implemented using the *train* function in the 'caret' package in R, with the appropriate 'method' argument specified as follows: Stepwise regression: ‘glmStepInc', Elastic Net: ‘glmnet', and Random Forest: ‘rf’. Furthermore, to mitigate class imbalances inherent within the dataset, class weights were computed as the inverse of the class frequencies and subsequently integrated into the model training process. These weights, operationalised though the 'weights' parameter in the *train* function, serve to recalibrate the model's focus towards underrepresented classes, thereby improving accuracy in predicting these classes. For instance, a class with substantially fewer instances than others would be assigned a higher weight, incentivising the model to allocate increased computational resources towards accurately predicting instances of this class. This methodological adjustment is crucial in fostering a balanced predictive performance, counteracting the model's inherent propensity to bias predictions in favour of overrepresented classes.

Firstly, the dataset was split into a training set and a testing set in a 70:30 ratio. The training set, consisting of 70% of the data ($n$ = 3,963 observations), was further subjected to a 10-fold cross-validation process to evaluate and select the best model parameters. During this cross-validation process, various combinations of hyperparameters (e.g., *mtry* and *ntree* values for the random forest model) were evaluated, and the optimal values (that yielded the lowest root mean squared error) were used to train the final model. The final models were trained on the entire training dataset using these optimal parameters. The performance of the final models was then evaluated on the testing dataset ($n$ = 1,695 observations) to assess their predictive capabilities and generalisation to unseen data. For the Random Forest, the importance of each variable for improving model performance was estimated using the *varImp* function in the 'caret' R package.

Two variations of the model were fit, one incorporating environment-related variables from the HLI dataset, and another excluding HLI indicators. This was performed to examine how environmental data affected the predictive performance of each model. The performance of all models were assessed using root mean squared error (RMSE), mean absolute error (MAE), and multiple R-squared ($R^2$). As a further check, the Pearson's correlation between the observed and predicted values were also evaluated.

**Results**

We employed three distinct models to predict four wellbeing variables: life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. Table 5-4 provides a summary of both the observed mean and standard deviation, alongside the predicted values for all models. Notably, the Random Forest model exhibited superior performance, with predictions that were closely aligned with the observed values. These results were obtained through the evaluation of model performance on the test dataset, comprising 30% of the original dataset ($n$ = 1,695).

Table 5-5 provides an overview of the performance metrics for all predictive models. Notably, the Random Forest models demonstrated stronger performance with lower RMSE (ranging between 1.5 and 1.6 for life satisfaction, life worthwhileness, and family wellbeing). However, the $R^2$ values were relatively low (~ 0.06), suggesting that these models had limited explanatory capabilities. Traditional models (Stepwise regression and Elastic Net) produced higher RMSE values (~ 2.5) and even lower $R^2$ values (<0.03) for these wellbeing variables. Table 5-5 displays the results with and without the inclusion of environmental variables for the Random Forest model only (given this was the best performing). The incorporation of environmental features had a negligible impact on the model's predictive capacity. Furthermore, we assessed the correlation between the observed and predicted values produced by the Random Forest model (without environmental variables). This correlation ranged from weak to moderate, falling within the range of 0.202 to 0.250 for all wellbeing outcome variables. Appendix Table A-5 (presented in the appendix section) shows the importance of the top 10 predictor variables employed by the random forest model.

**Table 5-4.** Descriptive statistics (obtained from the testing dataset, *n* = 1,695) for observed and predicted wellbeing variables

| Outcome variable | Observed mean ± SD | Predicted mean ± SD | | |
|---|---|---|---|---|
| | | Stepwise Regression | Elastic Net Regression | Random Forest |
| Life satisfaction | 7.75 ± 1.65 | 6.23 ± 1.48 | 6.09 ± 1.19 | **7.76 ± 0.23** |
| Life worthwhileness | 8.15 ± 1.54 | 6.68 ± 1.42 | 6.61 ± 1.35 | **8.14 ± 0.21** |
| Family wellbeing | 7.86 ± 1.61 | 6.69 ± 1.97 | 6.67 ± 1.65 | **7.87 ± 0.21** |
| Mental wellbeing | 63.16 ± 17.75 | 55.41 ± 13.58 | 54.55 ± 10.57 | **63.23 ± 2.32** |

**Table 5-5.** Model performance metrics

| | Outcome variable | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| **Stepwise Regression** | Life satisfaction | 2.534 | 2.081 | 0.028 |
| | Life worthwhileness | 2.463 | 2.012 | 0.013 |
| | Family wellbeing | 2.683 | 2.154 | 0.010 |
| | Mental wellbeing | 21.509 | 17.459 | 0.040 |
| **Elastic net Regression** | Life satisfaction | 2.497 | 2.085 | 0.028 |
| | Life worthwhileness | 2.461 | 2.030 | 0.015 |
| | Family wellbeing | 2.478 | 2.008 | 0.013 |
| | Mental wellbeing | 20.325 | 16.772 | 0.055 |
| **Random Forest** *(with environmental variables)* | Life satisfaction | 1.595 | 1.200 | 0.077 |
| | Life worthwhileness | 1.508 | 1.164 | 0.045 |
| | Family wellbeing | 1.583 | 1.195 | 0.040 |
| | Mental wellbeing | 17.226 | 13.715 | 0.062 |
| **Random Forest** *(without environmental variables)* | Life satisfaction | 1.596 | 1.198 | 0.072 |
| | Life worthwhileness | 1.505 | 1.165 | 0.050 |
| | Family wellbeing | 1.582 | 1.198 | 0.040 |
| | Mental wellbeing | 17.273 | 13.745 | 0.053 |

## Discussion

The primary aim of this study was to evaluate the predictive efficacy of population-level socio-demographic variables in predicting GSS-based subjective wellbeing outcomes, encompassing life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. This analysis was augmented by incorporating environmental data from the HLI. The study employed three distinct predictive models: Stepwise Regression, Elastic Net, and Random Forest. Our results demonstrated the models' ability to predict wellbeing outcomes, as evidenced by their low RMSE values, by utilising a concise set of easily accessible socio-demographic variables from the Census. However, the low $R^2$ values suggest a constrained capacity to account for the extensive variability in the dependent variables. In practical terms, while the models are adept at approximating group-level averages with reasonable precision, they fail to capture the underlying dynamics or variance in wellbeing outcomes. We suspect these limitations may be influenced by various factors, such as dataset characteristics. Nevertheless, our findings highlight the need for further improvements in the predictive modelling of wellbeing outcomes.

This could involve incorporating a wider variety of predictors within the IDI, such as health records, or adopting advanced modelling techniques like deep learning algorithms.

In our investigation, Random Forest models outperformed conventional modelling techniques like Elastic Net and Stepwise regression in terms of predictive capability. This may be because random forest algorithms are capable of capturing complex nonlinear relationships in the data, handling multicollinearity, and reducing overfitting through their ensemble nature [236, 237]. While previous studies have employed similar methodologies to predict clinical outcomes such as the incidence of cardiovascular diseases and other chronic conditions [51], our study stands out by predicting subjective wellbeing outcomes (e.g., life satisfaction) by utilising a straightforward demographic variable set.

The inclusion of environmental variables from the HLI dataset did not result in a significant improvement in model performance when compared to models that solely relied on socio-demographic factors. Yet, these environmental variables ranked among the top 10 important predictors when assessed using the *varImp* function. This suggests that while environmental factors play a role in shaping wellbeing outcomes, their influence is not strong enough to significantly enhance the predictive capabilities of these models. Prior research has indicated a connection between the HLI indicators and deprivation [228], primarily determined using various socio-demographic indicators such as education, income, and housing data from the Census. Given that we have already included a range of these Census-level socio-demographic variables in our analyses, the inclusion of environmental variables may not have offered any additional insights beyond what we had already captured through the Census data.

Additionally, it is worth noting that the environmental variables from the HLI dataset primarily capture proximity to various environmental elements but do not consider the total number, variety, or quality of such facilities. It is known that overall extent and quality of green/blue

space within an area is related to mental health [65, 66], and studies have established the importance of environmental factors in influencing an individual's mental health [238, 239]. Future studies could explore the utility of a more refined selection of environmental variables in the modelling process.

Although our predictions were reasonable, there are limitations in our approach that should be discussed. Firstly, the wellbeing data from the GSS 2018 dataset used to train the models did not have a uniform distribution of responses across the measurement scale. For instance, the outcome variable 'life satisfaction' ranged from 1 to 11, and over 50% of respondents reported a score of either 7 or 8. This imbalance may be inherent to the subjective nature of the question. Despite incorporating weights into the model training process, the majority of our predictions tended to cluster around scores of 7 and 8. Since this range of values closely aligns with that of the observed values, the models achieved a relatively low RMSE (< 1.6). However, a lower correlation between the observed and predicted values (0.20 – 0.25) suggests that the model predictions within this narrow range were not linearly associated with the observed data. This discrepancy can likely be attributed to the clustering of values in the GSS dataset itself, reflecting the nature of subjective wellbeing within populations. It is important to highlight that while our predictions typically fell within a 1 – 2-point range of the true scores, this apparent accuracy could be misleading. This is because the true scores themselves predominantly fell within this same 1–2-point range, and consequently, the proportional error is relatively high.

Understanding subjective wellbeing, especially when collected through surveys, is complex. Unlike quantifying tangible health conditions (e.g., cardiovascular disease, obesity, diabetes), subjective wellbeing relies on self-reported responses, which can vary based on how an individual interprets the question. For example, two people who choose scores of 7 might perceive those scores differently. Moreover, a lower score might not necessarily indicate less satisfaction relative to another person, it could reflect an individual's unique understanding of

the scale. Without a benchmark for validation, it is challenging to confidently interpret model results. Another important consideration is that these outcome scores reflect an individual's overall wellbeing experience over time, not just their feelings on the day of the survey. However, someone generally satisfied with life might choose a lower score if recent unpleasant events influenced their mood. The subjective nature of these outcomes makes their validation difficult.

Another limitation arises from the dataset cleaning process, particularly the exclusion of nearly 3% of the Māori population due to missing values (see Appendix Table 5). This exclusion could have potentially introduced bias into the model's predictions and overall outcomes. Similarly, another limitation pertains to the Census 2018 dataset which had a lower response rate than expected. To address this challenge, Stats NZ employed alternative strategies to impute missing data. These strategies involved leveraging other available microdata within the IDI to fill in the gaps and enhance the completeness of the dataset. Although the data imputation process is beneficial, it introduces a potential source of bias or uncertainty in our results, as the imputed values may not accurately capture the true characteristics of the non-respondent population. Further information regarding this issue can be explored in "2018 Census collection response rates unacceptably low" by Stats NZ (2018) [240].

To enhance the predictive performance for future studies, we recommend exploring additional analyses and engaging in alternate feature engineering strategies. For instance, a broader range of demographic variables could be considered to provide a more comprehensive representation of individual characteristics. One area for future exploration could involve examining how different treatments of the outcome variable impact the model's predictive accuracy. For instance, our model used life satisfaction as a continuous variable, rather than categorising it (e.g., low, medium, high). However, establishing thresholds for these classes is a somewhat subjective decision, and may require guidance from industry experts. Additionally, creating

composite indices that capture multiple dimensions of wellbeing, or integrating other data sources available in the IDI (e.g., health data) as predictors could potentially lead to improved model performance. Considering alternative modelling techniques beyond the ones explored in this study, such as neural networks, may also reveal further insights into predicting wellbeing outcomes.

**Conclusion**

Our findings indicate that a Random Forest model, in conjunction with census-level socio-demographic variables, yields moderate predictive efficacy for a range of GSS-based subjective wellbeing measures. This outcome underscores the potential of this methodological approach. However, it is imperative to acknowledge limitations arising from the subjective nature and distribution characteristics of the outcome variables, as our models could not effectively explain the inherent variability in wellbeing outcomes. While our study offers valuable insights into predicting wellbeing outcomes using predictive modelling techniques, there is significant scope for improvement. By refining the modelling approach, incorporating more diverse data sources (e.g., health records within the IDI), and employing advanced analytical methods (e.g., deep learning), future research could contribute to a more accurate and comprehensive understanding of population wellbeing and offer robust tools for evidence-based policymaking.

# Chapter 6 – Predicting subjective wellbeing outcomes for the New Zealand population using census-level data.

## Preface

In the previous chapter, machine learning models were developed to predict various subjective wellbeing measures from socio-demographic and environmental variables. These models exhibited a low RMSE score, suggesting reasonable predictive accuracy. However, to truly gauge their applicability, it is crucial to evaluate the models' performance at the population level, not just on the GSS dataset. This chapter explores model predictions at this broader scale, by utilising these models to predict wellbeing in the census population. These models' ability to capture variations across different population subgroups will be benchmarked against the subgroup estimates provided by the GSS, which is known for its representativeness. This will provide insights into the models' capacity to reflect real-world conditions, pinpoint specific subgroups where the model shows high accuracy, alongside identifying areas where enhancements are needed. Such an evaluation is vital for understanding the utility for informing policy decisions and addressing wellbeing disparities.

## Introduction

In recent years, the concept of success, both at the individual and societal levels, has evolved significantly. While traditional economic indicators remain important, they are no longer considered the sole measures of prosperity and progress [166]. A paradigm shift towards assessing population wellbeing has gained prominence, prompting governments to reassess their policy priorities and goals in the pursuit of comprehensive societal flourishing (i.e., high levels of wellbeing among citizens) [225]. New Zealand has been at the forefront of this global trend, evidenced by the country's pioneering Wellbeing budget in 2019 [46].

The link between positive mental states, such as happiness, and physiological health is well-established [36, 39]. Individuals with high levels of wellbeing are also less likely to develop cardiovascular disease [36] and hypertension [40]. Cohen et al.'s research [38] revealed that individuals with higher levels of positive emotion exhibited increased resistance against infection, lower levels of stress hormones, and fewer reported physiological symptoms compared to those experiencing higher levels of negative emotion. This growing body of evidence has spurred international interest in the measurement and promotion of population wellbeing [23], along with understanding the factors associated with both high and low levels of wellbeing.

Measuring wellbeing presents unique challenges due to its inherently subjective nature [241]. To comprehensively assess how a country's citizens are faring and how government initiatives impact their wellbeing, a population-wide measure of wellbeing is essential. In New Zealand, wellbeing metrics are primarily available through the General Social Survey (GSS), which is representative of the country's overall population, but may not accurately reflect specific subgroups of high policy interest [242]. Consequently, understanding the determinants of wellbeing within these crucial areas remains a challenge.

Two potential strategies emerge as solutions. One approach involves collecting routine wellbeing data for the entire population through a census activity. However, this method is resource-intensive and time-consuming. An alternative strategy entails utilising existing data collected as part of routine procedures (e.g., socio-demographic data from the census) and using this person-specific information to extrapolate wellbeing measures from the GSS to the broader census population. This approach becomes viable due to the existence of New Zealand's Integrated Data Infrastructure (IDI), a sophisticated database managed by Stats NZ [103]. The IDI comprises individual response data (microdata) concerning people and households, supplemented with anonymised information spanning education, income, health, justice, and housing [72]. Notably, the IDI facilitates data linkage across these domains, additional details of which can be found elsewhere [226]. Critically, the IDI enables the innovative linkage of GSS and the country's Census data, which covers most of the national population, owing to legal requirements.

In a recent study (Chapter 5), this methodology was applied to construct predictive models that utilised socio-demographic data from the census to predict wellbeing measures in the GSS sample, including life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. This study aims to assess the convergent validity of these predictive models by evaluating their predictions in the absence of a definitive wellbeing measure at the population level. Given that the GSS is recognised as the established metric for assessing population wellbeing in New Zealand, we approach this by comparing our predictions against existing GSS measures across different demographic subgroups and regions (see Figure 6-1). Through this comparative analysis, we seek to illuminate the efficacy of our predictions and their congruence with the national wellbeing data currently available.

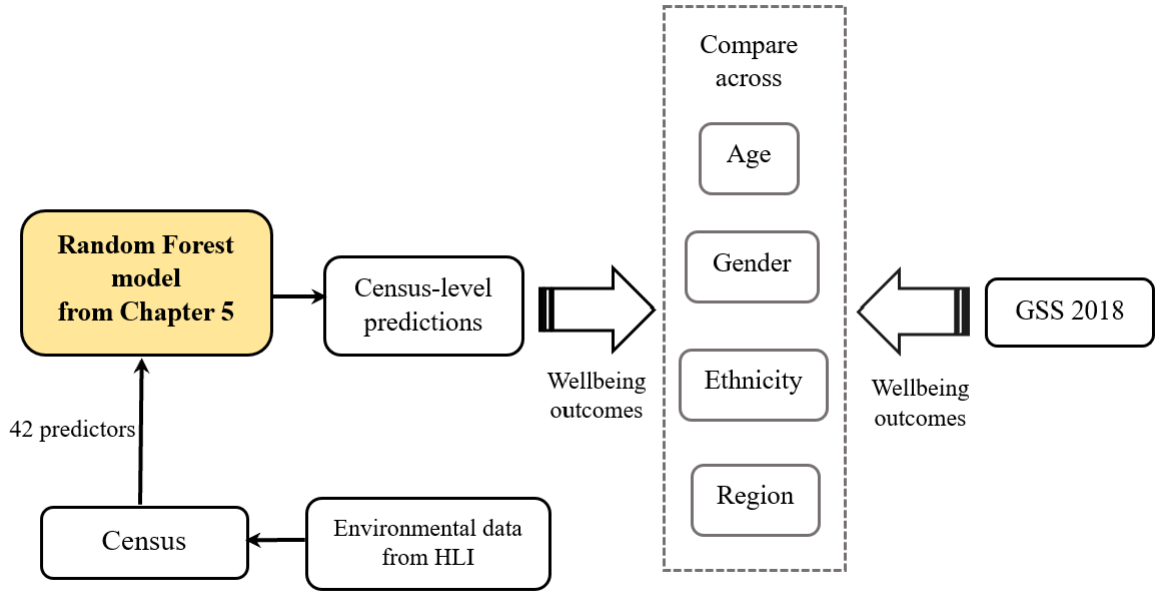



**Figure 6-1.** Validation Approach for Population-level Predictions

## Methods

### *Data Sources*

This study utilises three key datasets: the New Zealand GSS [47], the New Zealand Census of Population and Dwellings [227], and the Healthy Location Index (HLI) [67]. Both the GSS and the Census are present in the IDI, allowing for seamless linkage using a unique identifier variable [72]. The HLI dataset, although external to the IDI, was incorporated into the IDI environment by Stats NZ. All datasets within the IDI can only be accessed from the Stats NZ data laboratory. A formal application to access the IDI datasets and laboratory was submitted and approved by Stats NZ. The methodology in this research was approved by the AUT University Ethics Committee (AUTEC #21/115).

The Census dataset provides a snapshot of the population and dwellings in New Zealand at a specific point in time. The census data are collected by Stats NZ once every five years. Further details about the census and its methodology are available elsewhere [231]. The 2018 Census

dataset, with approximately 4.9 million entries with over 300 variables, was our primary data source. From this, 29 demographic variables were selected.

Next, data related to the environment was acquired from the Healthy Location Index (HLI) dataset [67]. The HLI provides a rank (ranging between 1 and 52,593) for every meshblock in New Zealand based on the accessibility (i.e., distance proximity) to (1) health-promoting features of the environment (e.g., green spaces, physical activity facilities) and (2) health-constraining features of the environment (e.g., fast-food outlets, takeaway outlets). More details about this dataset and the methodology involved in developing this measure can be found elsewhere [67]. A total of 13 environmental variables were selected, each measured in deciles, ranging from 1 (indicating the highest decile and closest proximity to the environmental feature) to 10 (representing the lowest decile and the farthest distance from the environmental feature). A detailed list of all the predictor variables is already presented in Table 5-2 and Table 5-3.

***Data Preparation and Cleaning***

The Census was linked with the HLI dataset using the meshblock code. After linking the datasets, any observations with missing values (across any of the census variables) were removed from the dataset. To ensure data quality and consistency, the dataset underwent a cleaning process. Unknown or "did not answer" categories in the variables were removed resulting in the final dataset with a sample size of 2,451,891 with 42 predictor variables. The demographic distribution of the final census dataset (shown in Table 6-1) closely resembles that of the GSS 2018 dataset, indicating a balanced representation of most of the demographic sub-groups without any noticeable over- or under-representation.

**Table 6-1**. Comparative Distribution of Demographic Data: GSS 2018 vs. Census 2018 (cleaned)

| Variable | Variable category | GSS 2018 dataset distribution (N=8,661) | Census dataset distribution (i.e., after data cleaning, N = 2,451,891) |
|---|---|---|---|
| Age range | 15–24 years | 10% | 15% |
| Age range | 25–34 years | 16% | 16% |
| Age range | 35–44 years | 16% | 16% |
| Age range | 45–54 years | 17% | 17% |
| Age range | 55–64 years | 17% | 16% |
| Age range | 65 years or over | 24% | 20% |
| Gender | Male | 45% | 47% |
| Gender | Female | 55% | 53% |
| Ethnicity | New Zealand European | 67% | 69% |
| Ethnicity | New Zealand Māori | 7% | 4% |
| Ethnicity | Pacific | 4% | 3% |
| Ethnicity | Asian | 10% | 14% |
| Ethnicity | Other | 11% | 10% |
| Region | Auckland | 27% | 33% |
| Region | Wellington | 11% | 12% |
| Region | Northland group | 12% | 10% |
| Region | Rest of North Island | 24% | 20% |
| Region | Canterbury | 13% | 14% |
| Region | Rest of South Island | 13% | 11% |

Lastly, the GSS 2018 dataset was used in the study to compare our predictions with wellbeing measures available in the GSS. Four different wellbeing outcome variables were selected: life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing. Four demographic variables were selected: age, gender, ethnicity, and region. These variables were selected because, they are the key demographic variables commonly considered as stratifiers in descriptive research and policymaking processes. Additionally, Chapter 4's findings underscored the relationships between these demographic variables and subjective wellbeing, reinforcing their significance and further justifying their use.

***Prediction and Analysis***

In this study, we employed the Random Forest model that was trained using GSS data in Chapter 5. The Random Forest model is an ensemble learning technique that constructs

multiple decision trees and combines their predictions to improve accuracy while mitigating overfitting [154]. Given its superior performance over traditional statistical models (as observed in Chapter 5), we exclusively utilised the Random Forest model for our predictions.

Firstly, population-level predictions for the wellbeing outcomes were generated by applying the Random Forest model to the cleaned census dataset, which encompassed over 2.4 million observations with 42 predictors. This was achieved using the *predict* function within the 'caret' package in R. Next, we summarised the predictions across four distinct demographic categories: age, gender, ethnicity, and region. Various descriptive statistics, including mean, standard deviation, and interquartile ranges, were computed for each demographic group. This entire process was repeated four times, employing the appropriate model for each of the four outcomes of interest. These results were then compared with GSS wellbeing measures stratified by each corresponding demographic subgroup. The GSS currently serve as the standard measure of population wellbeing in New Zealand.

To assess potential differences between our predictions and the GSS values across these demographic subgroups, we employed the Mann-Whitney U test. *p*-values derived from this test were used to determine the significance of any disparities observed.

**Results**

Our analysis yields insights into the predictive ability of our model across various population subgroups. Table 6-2 to Table 6-5 detail these findings, comparing predicted values for life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing against the GSS data, respectively. Our model's predictions align with GSS data within specific demographic groups, while most subgroups exhibit statistically significant differences. This suggests that the model's accuracy in predicting wellbeing measures varies considerably across different segments of the population.

The analysis of life satisfaction (presented in Table 6-2) revealed statistically significant differences between predicted scores and their corresponding GSS values across all subgroups, except for individuals of Pacific ethnicity. Table 6-3 indicates the predicted values of life worthwhileness differed from the GSS among all age groups, except for individuals aged 15 to 24. Predicted life worthwhile scores were also different for New Zealand European and Asian ethnic groups, but not Pacific, Māori, or other. In terms of regional comparisons, significant differences were evident in life worthwhile scores across all regions except Northland (includes Northland group and the Rest of North Island). The predicted values closely aligned with GSS values, for the entire North Island, except Auckland and Wellington.

Predicted family wellbeing was different from the GSS across all demographic subgroups (Table 6-4). Finally, Table 6-5 presents the mental wellbeing results. Significant differences were observed across all age groups except for those aged 25 to 34 and 45 to 54. In contrast, no significant differences were found between predicted and original GSS measures for ethnic groups other than New Zealand European, Asian, Māori, or Pacific.

**Table 6-2.** Life satisfaction predictions against GSS values

| Outcome Variable | Demographic group | Demographic Category | N – Census | N – GSS | Predictions Mean ± SD | Predictions Median (25th, 75th percentile) | GSS Mean ± SD | GSS Median (25th, 75th percentile) | *p-value |
|---|---|---|---|---|---|---|---|---|---|
| **Life satisfaction** | Age range | 15–24 years | 372744 | 861 | 7.66 ± 0.22 | 7.66 (7.52, 7.80) | 7.72 ± 1.64 | 8 (7, 9) | p < 0.05 |
| | | 25–34 years | 388650 | 1413 | 7.70 ± 0.20 | 7.70 (7.57, 7.83) | 7.67 ± 1.59 | 8 (7, 9) | p < 0.05 |
| | | 35–44 years | 387615 | 1374 | 7.76 ± 0.21 | 7.77 (7.63, 7.89) | 7.52 ± 1.62 | 8 (7, 8) | p < 0.05 |
| | | 45–54 years | 417729 | 1455 | 7.78 ± 0.21 | 7.79 (7.65, 7.92) | 7.47 ± 1.78 | 8 (7, 9) | p < 0.05 |
| | | 55–64 years | 390849 | 1482 | 7.81 ± 0.22 | 7.82 (7.68, 7.95) | 7.55 ± 1.83 | 8 (7, 9) | p < 0.05 |
| | | 65 years or over | 494304 | 2076 | 7.88 ± 0.21 | 7.90 (7.76, 8.02) | 8.08 ± 1.76 | 8 (7, 10) | p < 0.05 |
| | Gender | Male | 1157151 | 3903 | 7.77 ± 0.22 | 7.77 (7.62, 7.91) | 7.64 ± 1.73 | 8 (7, 9) | p < 0.05 |
| | | Female | 1294740 | 4758 | 7.77 ± 0.23 | 7.78 (7.63, 7.92) | 7.74 ± 1.72 | 8 (7, 9) | p < 0.05 |
| | Ethnicity | New Zealand European | 1697433 | 5829 | 7.80 ± 0.22 | 7.81 (7.66, 7.95) | 7.72 ± 1.73 | 8 (7, 9) | p < 0.05 |
| | | New Zealand Māori | 93312 | 609 | 7.66 ± 0.23 | 7.66 (7.50, 7.82) | 7.72 ± 1.79 | 8 (7, 9) | p < 0.05 |
| | | Pacific | 74397 | 384 | 7.61 ± 0.18 | 7.61 (7.48, 7.73) | 7.50 ± 1.70 | 7 (7, 9) | **0.619** |
| | | Asian | 351687 | 903 | 7.74 ± 0.18 | 7.75 (7.63, 7.86) | 7.73 ± 1.65 | 8 (7, 9) | p < 0.05 |
| | | MELAA and other ethnic groups | 235059 | 936 | 7.70 ± 0.23 | 7.70 (7.54, 7.84) | 7.55 ± 1.79 | 8 (7, 9) | p < 0.05 |
| | Region | Auckland | 809145 | 2343 | 7.74 ± 0.20 | 7.75 (7.62, 7.88) | 7.61 ± 1.67 | 8 (7, 9) | p < 0.05 |
| | | Wellington | 287412 | 993 | 7.72 ± 0.22 | 7.74 (7.59, 7.88) | 7.66 ± 1.62 | 8 (7, 9) | p < 0.05 |
| | | Northland group | 252060 | 999 | 7.80 ± 0.23 | 7.81 (7.65, 7.95) | 7.82 ± 1.72 | 8 (7, 9) | p < 0.05 |
| | | Rest of North Island | 491400 | 2115 | 7.79 ± 0.25 | 7.80 (7.63, 7.96) | 7.74 ± 1.80 | 8 (7, 9) | p < 0.05 |
| | | Canterbury | 337545 | 1104 | 7.79 ± 0.22 | 7.80 (7.64, 7.94) | 7.61 ± 1.80 | 8 (7, 9) | p < 0.05 |
| | | Rest of South Island | 274332 | 1104 | 7.80 ± 0.24 | 7.82 (7.65, 7.96) | 7.79 ± 1.73 | 8 (7, 9) | p < 0.05 |

***Note:*** *MELAA – Middle Eastern/Latin American/African; *p-value from Mann-Whitney U test; All p-values (p > 0.05) are highlighted in bold.*

**Table 6-3.** Life worthwhileness predictions against GSS values

| Outcome Variable | Demographic group | Demographic Category | N – Census | N – GSS | Predictions Mean ± SD | Predictions Median (25th, 75th percentile) | GSS Mean ± SD | GSS Median (25th, 75th percentile) | *p-value |
|---|---|---|---|---|---|---|---|---|---|
| **Life worthwhileness** | Age range | 15–24 years | 372744 | 861 | 7.97 ± 0.15 | 7.96 (7.87, 8.06) | 7.80 ± 1.70 | 8 (7, 9) | **0.116** |
| | | 25–34 years | 388650 | 1413 | 8.10 ± 0.16 | 8.10 (8.00, 8.20) | 8.01 ± 1.54 | 8 (7, 9) | p < 0.05 |
| | | 35–44 years | 387615 | 1374 | 8.16 ± 0.16 | 8.16 (8.06, 8.26) | 8.03 ± 1.55 | 8 (7, 9) | p < 0.05 |
| | | 45–54 years | 417729 | 1455 | 8.19 ± 0.16 | 8.19 (8.09, 8.29) | 7.90 ± 1.60 | 8 (7, 9) | p < 0.05 |
| | | 55–64 years | 390849 | 1482 | 8.23 ± 0.16 | 8.23 (8.13, 8.33) | 8.08 ± 1.65 | 8 (7, 9) | p < 0.05 |
| | | 65 years or over | 494304 | 2076 | 8.26 ± 0.17 | 8.28 (8.17, 8.37) | 8.36 ± 1.63 | 9 (8, 10) | p < 0.05 |
| | Gender | Male | 1157151 | 3903 | 8.14 ± 0.19 | 8.14 (8.01, 8.26) | 7.94 ± 1.62 | 8 (7, 9) | p < 0.05 |
| | | Female | 1294740 | 4758 | 8.18 ± 0.18 | 8.19 (8.05, 8.30) | 8.18 ± 1.61 | 8 (7, 10) | p < 0.05 |
| | Ethnicity | New Zealand European | 1697433 | 5829 | 8.19 ± 0.18 | 8.20 (8.07, 8.32) | 8.12 ± 1.60 | 8 (7, 9) | p < 0.05 |
| | | New Zealand Māori | 93312 | 609 | 8.10 ± 0.18 | 8.10 (7.98, 8.23) | 8.10 ± 1.76 | 8 (7, 10) | **0.991** |
| | | Pacific | 74397 | 384 | 8.00 ± 0.15 | 8.00 (7.90, 8.10) | 7.83 ± 1.61 | 8 (7, 9) | **0.166** |
| | | Asian | 351687 | 903 | 8.09 ± 0.15 | 8.10 (7.99, 8.20) | 7.95 ± 1.56 | 8 (7, 9) | p < 0.05 |
| | | MELAA and other ethnic groups | 235059 | 936 | 8.09 ± 0.19 | 8.09 (7.96, 8.21) | 7.99 ± 1.67 | 8 (7, 9) | **0.053** |
| | Region | Auckland | 809145 | 2343 | 8.12 ± 0.17 | 8.13 (8.01, 8.24) | 7.94 ± 1.58 | 8 (7, 9) | p < 0.05 |
| | | Wellington | 287412 | 993 | 8.13 ± 0.18 | 8.14 (8.00, 8.26) | 7.98 ± 1.58 | 8 (7, 9) | p < 0.05 |
| | | Northland group | 252060 | 999 | 8.19 ± 0.19 | 8.20 (8.07, 8.32) | 8.22 ± 1.69 | 8 (7, 10) | **0.780** |
| | | Rest of North Island | 491400 | 2115 | 8.18 ± 0.20 | 8.19 (8.05, 8.32) | 8.17 ± 1.64 | 8 (7, 10) | **0.079** |
| | | Canterbury | 337545 | 1104 | 8.18 ± 0.18 | 8.19 (8.06, 8.31) | 8.03 ± 1.63 | 8 (7, 9) | p < 0.05 |
| | | Rest of South Island | 274332 | 1104 | 8.19 ± 0.19 | 8.20 (8.06, 8.32) | 8.15 ± 1.59 | 8 (7, 10) | p < 0.05 |

***Note:*** *MELAA – Middle Eastern/Latin American/African; *p-value from Mann-Whitney U test; All p-values (p > 0.05) are highlighted in bold.*

**Table 6-4.** Family wellbeing predictions against GSS values

| Outcome Variable | Demographic group | Demographic Category | N – Census | N – GSS | Predictions Mean ± SD | Predictions Median (25th, 75th percentile) | GSS Mean ± SD | GSS Median (25th, 75th percentile) | *p-value |
|---|---|---|---|---|---|---|---|---|---|
| **Family wellbeing** | Age range | 15–24 years | 372744 | 861 | 7.79 ± 0.19 | 7.80 (7.67, 7.92) | 7.69 ± 1.68 | 8 (7, 9) | p < 0.05 |
| | | 25–34 years | 388650 | 1413 | 7.81 ± 0.18 | 7.81 (7.68, 7.93) | 7.74 ± 1.62 | 8 (7, 9) | p < 0.05 |
| | | 35–44 years | 387615 | 1374 | 7.85 ± 0.18 | 7.86 (7.73, 7.97) | 7.70 ± 1.66 | 8 (7, 9) | p < 0.05 |
| | | 45–54 years | 417729 | 1455 | 7.84 ± 0.18 | 7.85 (7.73, 7.97) | 7.61 ± 1.76 | 8 (7, 9) | p < 0.05 |
| | | 55–64 years | 390849 | 1482 | 7.86 ± 0.17 | 7.87 (7.75, 7.98) | 7.67 ± 1.79 | 8 (7, 9) | p < 0.05 |
| | | 65 years or over | 494304 | 2076 | 8.00 ± 0.17 | 8.01 (7.90, 8.12) | 8.23 ± 1.66 | 8 (7, 10) | p < 0.05 |
| | Gender | Male | 1157151 | 3903 | 7.85 ± 0.19 | 7.86 (7.73, 7.98) | 7.80 ± 1.68 | 8 (7, 9) | p < 0.05 |
| | | Female | 1294740 | 4758 | 7.87 ± 0.19 | 7.88 (7.75, 8.01) | 7.82 ± 1.74 | 8 (7, 9) | p < 0.05 |
| | Ethnicity | New Zealand European | 1697433 | 5829 | 7.87 ± 0.19 | 7.87 (7.74, 8.00) | 7.81 ± 1.71 | 8 (7, 9) | p < 0.05 |
| | | New Zealand Māori | 93312 | 609 | 7.77 ± 0.18 | 7.77 (7.65, 7.89) | 7.61 ± 1.91 | 8 (7, 9) | p < 0.05 |
| | | Pacific | 74397 | 384 | 7.86 ± 0.18 | 7.87 (7.74, 7.98) | 7.94 ± 1.67 | 8 (7, 9) | p < 0.05 |
| | | Asian | 351687 | 903 | 7.94 ± 0.16 | 7.95 (7.84, 8.05) | 8.20 ± 1.45 | 8 (7, 9) | p < 0.05 |
| | | MELAA and other ethnic groups | 235059 | 936 | 7.76 ± 0.18 | 7.77 (7.64, 7.89) | 7.51 ± 1.78 | 8 (7, 9) | p < 0.05 |
| | Region | Auckland | 809145 | 2343 | 7.89 ± 0.18 | 7.90 (7.78, 8.02) | 7.88 ± 1.62 | 8 (7, 9) | p < 0.05 |
| | | Wellington | 287412 | 993 | 7.83 ± 0.20 | 7.84 (7.70, 7.97) | 7.80 ± 1.68 | 8 (7, 9) | p < 0.05 |
| | | Northland group | 252060 | 999 | 7.85 ± 0.19 | 7.85 (7.72, 7.98) | 7.66 ± 1.84 | 8 (7, 9) | p < 0.05 |
| | | Rest of North Island | 491400 | 2115 | 7.85 ± 0.20 | 7.86 (7.72, 7.99) | 7.83 ± 1.74 | 8 (7, 9) | p < 0.05 |
| | | Canterbury | 337545 | 1104 | 7.86 ± 0.19 | 7.87 (7.73, 8.00) | 7.72 ± 1.81 | 8 (7, 9) | p < 0.05 |
| | | Rest of South Island | 274332 | 1104 | 7.85 ± 0.19 | 7.86 (7.73, 7.99) | 7.87 ± 1.64 | 8 (7, 9) | p < 0.05 |

***Note:*** *MELAA – Middle Eastern/Latin American/African; *p-value from Mann-Whitney U test; All p-values (p > 0.05) are highlighted in bold.*

**Table 6-5.** Mental wellbeing predictions against GSS values

| Outcome Variable | Demographic group | Demographic Category | N – Census | N – GSS | Predictions Mean ± SD | Predictions Median (25th, 75th percentile) | GSS Mean ± SD | GSS Median (25th, 75th percentile) | *p-value |
|---|---|---|---|---|---|---|---|---|---|
| **Mental wellbeing** | Age range | 15–24 years | 372744 | 861 | 62.90 ± 1.97 | 62.91 (61.65, 64.14) | 63.66 ± 17.75 | 68 (52, 76) | p < 0.05 |
| | | 25–34 years | 388650 | 1413 | 63.02 ± 1.89 | 63.10 (61.86, 64.25) | 61.99 ± 17.76 | 64 (52, 76) | **0.093** |
| | | 35–44 years | 387615 | 1374 | 63.20 ± 1.90 | 63.28 (62.05, 64.44) | 61.04 ± 18.43 | 64 (48, 76) | p < 0.05 |
| | | 45–54 years | 417729 | 1455 | 63.27 ± 2.01 | 63.40 (62.08, 64.60) | 60.43 ± 18.41 | 64 (48, 76) | **0.929** |
| | | 55–64 years | 390849 | 1482 | 63.66 ± 2.15 | 63.84 (62.40, 65.09) | 62.06 ± 18.47 | 64 (52, 76) | p < 0.05 |
| | | 65 years or over | 494304 | 2076 | 64.02 ± 2.35 | 64.31 (62.64, 65.64) | 66.65 ± 18.44 | 72 (56, 80) | p < 0.05 |
| | Gender | Male | 1157151 | 3903 | 63.52 ± 2.09 | 63.61 (62.24, 64.91) | 64.23 ± 18.12 | 68 (52, 76) | p < 0.05 |
| | | Female | 1294740 | 4758 | 63.24 ± 2.10 | 63.33 (61.94, 64.64) | 61.77 ± 18.55 | 64 (48, 76) | p < 0.05 |
| | Ethnicity | New Zealand European | 1697433 | 5829 | 63.50 ± 2.14 | 63.59 (62.16, 64.95) | 62.62 ± 18.29 | 64 (52, 76) | p < 0.05 |
| | | New Zealand Māori | 93312 | 609 | 62.54 ± 2.21 | 62.65 (61.15, 64.03) | 62.26 ± 18.92 | 64 (48, 76) | p < 0.05 |
| | | Pacific | 74397 | 384 | 62.22 ± 1.87 | 62.31 (61.06, 63.47) | 63.73 ± 18.84 | 64 (52, 80) | p < 0.05 |
| | | Asian | 351687 | 903 | 63.73 ± 1.61 | 63.78 (62.74, 64.78) | 66.04 ± 18.25 | 68 (56, 80) | p < 0.05 |
| | | MELAA and other ethnic groups | 235059 | 936 | 62.63 ± 2.11 | 62.66 (61.31, 63.98) | 61.45 ± 18.42 | 64 (48, 76) | **0.139** |
| | Region | Auckland | 809145 | 2343 | 63.33 ± 1.89 | 63.44 (62.19, 64.59) | 63.30 ± 18.49 | 64 (52, 80) | p < 0.05 |
| | | Wellington | 287412 | 993 | 62.85 ± 2.01 | 62.94 (61.57, 64.22) | 61.40 ± 17.50 | 64 (52, 72) | p < 0.05 |
| | | Northland group | 252060 | 999 | 63.60 ± 2.24 | 63.72 (62.21, 65.16) | 63.37 ± 18.60 | 68 (52, 76) | p < 0.05 |
| | | Rest of North Island | 491400 | 2115 | 63.43 ± 2.26 | 63.51 (62.02, 64.95) | 62.65 ± 18.84 | 68 (52, 76) | p < 0.05 |
| | | Canterbury | 337545 | 1104 | 63.45 ± 2.15 | 63.53 (62.12, 64.85) | 62.85 ± 18.24 | 64 (52, 76) | p < 0.05 |
| | | Rest of South Island | 274332 | 1104 | 63.66 ± 2.21 | 63.77 (62.27, 65.19) | 63.28 ± 18.10 | 68 (52, 76) | p < 0.05 |

***Note:*** *MELAA – Middle Eastern/Latin American/African; *p-value from Mann-Whitney U test; All p-values (p > 0.05) are highlighted in bold.*

**Discussion**

The examination of population-level predictions derived from the Random Forest model reveals important insights into the alignment between predicted values and the corresponding GSS values across various demographic groups. It is evident from our results that these predictions align closely with GSS values within specific demographic groups, even though most subgroups exhibit statistically significant differences. A notable observation is the small standard deviations associated with the predictions compared to the larger standard deviations in the GSS data, which could be attributed to differences in sample sizes.

One plausible interpretation of these results is that, while statistically significant differences exist, the subgroup means of the predicted values are similar to their GSS counterparts. The smaller standard deviations in the predictions may indicate that the model was unable to capture the larger variability present in the GSS reported wellbeing data. This means that the predicted wellbeing scores tended to be more tightly clustered around their respective means. However, in the context of wellbeing measurement, it is crucial to consider not just statistical significance but also practical significance [243]. Even if statistically significant differences exist, they may not be practically meaningful if the absolute differences are small [244, 245].

Additionally, it is also important to consider the ranking nature of the Mann Whitney U test and its implications for interpreting the results. The Mann Whitney U test compares ranks by joining both samples and ranking them from smallest to largest [246]. Given that a significant proportion of individuals were predicted with scores clustered around 7, most of these predictions will fall within the middle ranks. In contrast, the GSS scores are more likely to exhibit rankings at both the low and high ends, potentially contributing to the observed significant difference.

Another important consideration when exploring our results is the effect of sample size on the observed statistical differences. Due to the large sample size in the prediction dataset (derived from the census), even small differences can achieve statistical significance. However, the practical significance of such differences may be limited [247]. This highlights the need to interpret the results not only from a statistical perspective but also from a practical one. As discussed by Osborne (2008), statistical significance does not necessarily imply practical significance. When dealing with large sample sizes, it is crucial to focus on the magnitude of differences [248]. Therefore, while our results showed significant variations in subjective wellbeing scores among certain demographic groups, it's crucial to assess their practical impact on policymaking or intervention strategies, possibly by consulting relevant stakeholders or experts in the field.

The GSS is currently the primary source of wellbeing data in New Zealand. While our model (developed in Chapter 5) and validations are based on GSS data, it is important to recognise the limitations of using self-reported wellbeing data as the benchmark for evaluating model predictions. Other approaches for assessing the validity of these predictions, such as comparing trends within specific population groups, are valuable for understanding the alignment between predictions and GSS data. However, they may not fully demonstrate the sensitivity of our predictions without a well-established gold standard for population wellbeing.

The alignment of predicted values with GSS trends, such as increasing life satisfaction with age, is promising. Our predictions align with previous studies which found that life satisfaction tends to increase with age [198, 249]. Another notable observation is the consistent trend observed in life worthwhile scores among different ethnic groups. Individuals of Pacific ethnicity reported the lowest life worthwhile scores, while New Zealand European ethnicity individuals reported the highest life worthwhile scores, in both the GSS and our predictions.

Collectively, this indicates the model's capability to capture underlying patterns within the wellbeing data, which is crucial for making informed policy decisions.

The subjective nature of wellbeing measurements presents a unique challenge. Wellbeing is a deeply personal experience, and responses can vary greatly within demographic subgroups. Therefore, relying solely on self-reported data to inform policy and decisions should be approached with caution. Establishing a clear 'gold standard' for measuring population wellbeing is challenging due to this subjectivity. Studies have highlighted wellbeing's multidimensional nature, noting that individuals may prioritise different aspects of wellbeing differently [31, 250].

Furthermore, the reliance on self-assessments in the GSS, despite its design to be representative, adds a layer of complexity. The data collected are based on individuals' subjective experiences, reflecting personal perceptions of their wellbeing. This raises a critical question: can these subjective self-reports accurately represent the wellbeing of an entire population?

Studies in wellbeing research have consistently identified challenges with using self-reported measures for policy development [251, 252]. One significant issue is the variation in how different individuals interpret and respond to wellbeing questions. Factors like culture, social context, and personal experiences can significantly influence these perceptions, introducing variability and potential biases into the data. This makes it difficult to ascertain if the scores genuinely reflect a population's overall wellbeing status.

Therefore, it is crucial for policymakers and researchers to use self-reported wellbeing data judiciously in decision-making. While the GSS data offers valuable insights into individuals' perceptions, it should be complemented with objective indicators and contextual information for a more holistic understanding of the wellbeing landscape. Developing a composite index

that integrates both subjective and objective indicators could mitigate the limitations of subjective data, providing a more comprehensive approach to understanding wellbeing by encompassing a wider array of determinants and outcomes.

Utilising future data from the upcoming Census and GSS 2023 datasets would provide an opportunity to revisit and refine our methodology. This iterative approach can further enhance the accuracy of the predictive model and provide insights into whether the observed trends and patterns hold over time. The continuous exploration of methodologies and the integration of new data sources will contribute to our understanding of population wellbeing and its relationship with demographic variables and environmental factors. Future research could also explore alternate validation approaches, other machine learning techniques, or incorporate additional predictor variables to improve accuracy.

**Conclusion**

Our study assessed the predictive accuracy of population-level subjective wellbeing outcomes using a Random Forest model applied to the Census dataset. For each wellbeing outcome, we observed a close alignment between model-predicted means and GSS means for specific demographic groups, indicating the potential utility of our predictive models in estimating group-level subjective wellbeing measures. However, the subjective nature of wellbeing measurements, along with the absence of a gold standard measure, necessitates a cautious interpretation of our findings. Nevertheless, there are promising avenues for refining our predictive modelling approach and improving the accuracy of population wellbeing predictions. Leveraging future data from upcoming Census and GSS datasets, as well as ongoing methodological exploration, offer opportunities to address current limitations and enhance model robustness. In conclusion, while our study marks a significant step toward understanding and predicting population-level subjective wellbeing, ongoing research and

advancements in predictive modelling methodologies are vital for achieving accurate assessments.

# Chapter 7 – General Discussion

## Research summary

The primary aim of this thesis was to explore the feasibility of applying data science methodologies such as random forest to predict subjective wellbeing outcomes using administrative data from the Integrated Data Infrastructure (IDI). The overarching goal was to extrapolate these predictions to encompass the broader New Zealand population, ultimately yielding a comprehensive population-level measure of wellbeing. This research represents a novel contribution as the first exploration of predicting population-level wellbeing in New Zealand. By implementing a comprehensive and scalable data-driven methodology, it holds the potential to provide valuable insights into population health and has the potential to inform evidence-based policy decisions, marking a significant advancement in the field of wellbeing research and data science.

Chapter 2 laid the foundation for this work through a narrative literature review, summarising wellbeing theory and emphasising the global and political significance of subjective wellbeing. The review delved into the emergence of big data and the application of data science techniques, particularly machine learning, for modelling public health outcomes. This review provided context for our specific research objectives: (1) systematically review the literature on using machine learning to predict population-level health and wellbeing outcomes, (2) examine the New Zealand wellbeing landscape, (3) develop models for predicting subjective wellbeing from census data, and (4) extrapolate wellbeing predictions to the population-level and evaluating them against existing GSS measures. Each subsequent chapter of the thesis was structured to address these objectives in detail.

Chapter 3 was a systematic scoping review that examined data science methodologies used to predict the health and wellbeing of populations. The findings highlighted the potential of machine learning for this purpose but noted a lack of focus on subjective outcomes to date. This gap reinforced the novelty and significance of our research focus, underscoring the need to explore predictive methodologies tailored to subjective outcomes.

Chapter 4 examined the wellbeing landscape in New Zealand through a cross-sectional and trend analyses focused on understanding the association between subjective wellbeing outcomes and various demographic variables. The results emphasised the multifaceted nature of subjective wellbeing, with notable associations across age, gender, ethnicity, region, and socio-economic status. This study set the stage for the subsequent chapter, where these key variables were utilised in a predictive modelling framework.

Chapter 5 examined the effectiveness of various statistical models for predicting subjective wellbeing outcomes (life satisfaction, life worthwhileness, family wellbeing, and mental wellbeing) using census-level socio-demographic factors. Our results showcased the capability of the random forest for predicting subjective wellbeing, evidenced by low RMSE values. However, the models also exhibited low $R^2$ values, indicating a limited ability to explain the variability in the outcome variables. Despite achieving reasonable predictive accuracy, the study emphasised the need for further refinements to modelling subjective wellbeing outcomes.

The model that was developed in Chapter 5 was then used to extrapolate wellbeing predictions to the broader census population in Chapter 6. These predictions were validated across several population subgroups, by comparing the predicted means with the corresponding means from the GSS. Our results indicated that predictions generally align with GSS values within specific demographic groups, although most subgroups exhibited statistically significant differences. Various factors that could have contributed to these differences were acknowledged and

discussed, noting the importance of refining our methodology and modelling approach to enhance the prediction of population wellbeing in future studies.

## Significance of findings

The following section summarises the significance of our research findings. Firstly, the importance of a population-level measure of wellbeing is revisited, before discussing methodological insights garnered from undertaking this work. Next, the predictive model performance is discussed from a practical lens, exploring the real-world utility of these results in their current form. The subjective nature of wellbeing, particularly when used within an objective modelling framework, is then discussed, before acknowledging the main limitations of this thesis research.

### *A population-level measure of wellbeing*

The assessment of population-level wellbeing has been a long-discussed and intricate topic in fields of sociology, health, and economics. In New Zealand, the GSS serves as a tool to understand individuals' wellbeing experiences. However, as noted throughout this thesis, a limitation arises when examining wellbeing in smaller population subgroups due to the relatively small sample size of the GSS. This limitation emphasises the necessity for a more comprehensive approach to capturing wellbeing at the population-level. These challenges drove the direction of this research, representing the first attempt to address this issue by exploring methodologies capable of capturing population-level wellbeing (as discussed in Chapters 5 and 6).

Governments across the globe have begun to proactively embrace data-driven approaches in public health, demonstrating a commitment to utilising comprehensive health datasets for research, monitoring health trends, and informing evidence-based policymaking [253-255]. These examples showcased the potential of data science techniques in shaping public health

strategies and decision-making processes at a national level. We recognised that within the New Zealand context, the existence of the IDI offered a unique opportunity to explore how data science techniques such as machine learning could be used to achieve a population-level measure of wellbeing.

While our modelling approach has demonstrated considerable strengths, it also has notable limitations, particularly regarding the prediction of lower subjective wellbeing scores (below 7) at the population level. This limitation stems from substantial imbalances within the GSS dataset itself, especially concerning lower wellbeing scores. The current data in the GSS (particularly lower reported scores) may not be adequate for an effective modelling approach, highlighting the need for more comprehensive training data to improve our models' ability to explain outcome variability effectively.

Integrating key wellbeing questions into the Census could provide an ideal solution due to the Census' extensive reach and inclusivity across diverse population groups, enhancing data completeness and accuracy. Additionally, the Census being mandatory and reaching a large proportion of the population reduces potential biases that may arise in voluntary surveys with limited participation, such as the GSS. This comprehensive coverage improves the reliability and robustness of the data, making it a more reliable source for informing public policy decisions and resource allocation related to wellbeing initiatives. However, since the Census occurs once every five years, it may not capture short-term changes in population wellbeing due to policy shifts or external events like the COVID-19 pandemic. For instance, wellbeing data collected in 2018 (pre-COVID) and 2023 (post-COVID) may exhibit similar trends across population groups, masking the true impact of COVID-19 on people's wellbeing during that period. This limitation underscores the importance of our modelling approach, which can offer more frequent updates on population wellbeing by leveraging regularly updated administrative data.

Hypothetically, integrating wellbeing questions into the census for just one wave could potentially generate a significant volume of training data for each wellbeing variable across every score. Leveraging data science methodologies on such a dataset might result in enhanced model performance compared to models relying solely on the GSS. Nonetheless, decisions regarding these implementations depend on governmental priorities and may be influenced by political and social factors. Since the implementation of such a system remains highly uncertain, future work should focus on refining our existing methodology and modelling approach to improve model performance.

***Methodological decisions in predictive modelling***

Prior to modelling wellbeing, the initial parts of this thesis focused on exploring methodological decisions utilised in past work and conducting an initial analysis of wellbeing trends in New Zealand. Both steps were necessary to guide the decisions made in the subsequent chapters. The systematic review revealed the recent increase of big data in this field, particularly the utilisation of population-level datasets developed after 2010. This aligned with the introduction of the IDI in New Zealand in 2011, which facilitated access to multiple data sources collected across many areas of society.

Selecting appropriate predictor variables was identified as a crucial step in model development, as this decision could impact the robustness and generalisability of the models developed. The selection of predictor variables in Chapter 5 was guided by insights from Chapter 4 (i.e., identifying variables significantly associated with wellbeing in New Zealand) and the systematic review (i.e., collating, and synthesising variables used in previous modelling studies). Our results indicated that age and socio-economic deprivation consistently emerged as important predictors; a finding supported by recent research [256] where income and age ranked among the top socio-demographic predictors of subjective wellbeing. When including environmental variables from the Healthy Location Index, our model's performance did not

notably improve. Nonetheless, environmental variables remained among the top predictors, surpassing many socio-demographic variables from the census (see appendix Table A-5). We noted that future studies could explore environmental variables that capture the quality and quantity of environmental attributes in addition to the proximity. It is also likely that these area-level variables (i.e., aggregated by meshblock) may have limited their predictive accuracy at the individual level; a problem known in geography as the modifiable areal unit problem [257].

In this research, our primary focus was on utilising variables available in the census. It is essential to acknowledge that the IDI contains additional data sources, including administrative health records and income tax records. Linking these datasets to the GSS is one of the main advantages of the IDI, which creates opportunities for developing a more comprehensive set of predictor variables. However, linking datasets within the IDI introduces challenges related to sample size, as the selection of additional variables from different datasets increases the likelihood of encountering missing data, particularly when dealing with sensitive information. This missing data can affect both the training of predictive models and when extrapolating predictions to the population. This is both from a practical standpoint (as some model fitting processes cannot handle missing data), and an accuracy standpoint, as missing data can bias predictions, and affect model generalisability.

For example, when linking a health-related question from administrative health records to the GSS, valid responses may be dependent on individuals having a specific health condition. Any participants who have valid GSS data but missing health data might be subsequently excluded from analysis. The more datasets that are linked, exponentially more participants might be excluded from analysis. There are, however, several methods to handle missing data that could help to mitigate these issues. Depending on the extent of missing data, it may be possible to impute missing values using methods such as Multiple Imputation by Chained Equations (MICE) [150] or machine learning methods such as missForest [258]. Alternatively, predictive

modelling algorithms such as K-Nearest Neighbours (KNN) can inherently handle missing values (i.e., participants with missing information are retained during analysis) and may also be valuable [259]. Refining methodological decisions around missing data could enhance access and utilisation of numerous datasets within the IDI.

One notable trend of the studies examined in Chapter 3 was the substantial sample sizes, usually exceeding 6,000 individuals. When modelling the GSS data in Chapter 5, our sample size was approximately 5,000 after the data cleaning process. Instead of combining data from multiple GSS waves (2014, 2016, and 2018) to achieve a larger sample size, the decision to focus solely on the 2018 wave was influenced by the availability of census data for the same year. Recognising the trade-offs between including variables with high predictive capacity, achieving an adequate sample size, and optimising model performance, is crucial. Future studies could delve deeper into exploring these trade-offs.

The selection of an appropriate predictive model was also crucial. Chapter 3 highlighted the prevalent use of tree-based models, especially the Random Forest, observed in about 80% of reviewed studies. Chapter 5 utilised the Random Forest model to predict subjective wellbeing outcomes, with variations of linear regression models (Stepwise, Lasso, and Ridge regression) used as benchmarks. Indeed, a recent study by Vera Cruz et al. (2023) ranked the top 50 predictors of subjective wellbeing in approximately 38,000 older adults from 18 different countries using the Random Forest model, and further employed Generalised Additive Modelling (GAM) to assess predictor significance [256]. Future studies could explore and evaluate more complex models such as deep learning. However, regardless of model complexity, if the underlying variables lack sufficient predictive information or if insufficient training data are available, model performance will still be affected.

***Model performance and validation***

In Chapter 5, we demonstrated that the Random Forest models surpassed traditional statistical approaches, including stepwise regression, lasso, and ridge regression. While the Random Forest model exhibited moderate performance with RMSE values below 1.5 for life satisfaction, life worthwhileness, and family wellbeing (i.e., 1.5 units of error on an 11-point scale; 0–10), and under 20 for mental wellbeing (on a 0–100 scale), the relatively low $R^2$ values (i.e., 0.077 for life satisfaction, ~8% variance explained) implied a restricted ability to explain the variability in subjective wellbeing across individuals. Factors associated within this unexplained variance, including dataset characteristics, are explored in the next section.

Validation is crucial for assessing the practical utility and generalisability of machine learning models. In Chapter 6, we evaluated the validity of our models by comparing their population-level predictions with existing GSS measures of wellbeing across various demographic groups. The results demonstrated that the models' population-level predictions closely aligned with GSS scores within specific demographic groups. While statistically significant differences were observed in most subgroups, the trends observed in the predicted means closely mirrored their GSS counterparts. For instance, the predicted mean life satisfaction scores for individuals aged above 65 were the highest compared to other age groups, which was consistent in the GSS data. This indicates the model's ability to reproduce inherent patterns in population wellbeing, even though statistically significant differences existed between these two measures.

Understanding what constitutes a meaningful change in wellbeing is essential for accurately interpreting the results of our models. Whether a small change in wellbeing is meaningful or not is a difficult question to answer, given the nature of the wellbeing questions (i.e., their subjective interpretation, and the validity of the 0–10 scale). If a small change in wellbeing (say, 0.1 units) is meaningful, then the practical utility of our model may be limited, given the

RMSE of ~1.5 units. However, if a meaningful change is, for instance, a 3-unit change, then our model's practical utility is enhanced. A 2022 study [199] that investigated life satisfaction scores in New Zealand over 11 years showed a change from 7.61 (the highest) in 2007 to 7.23 (the lowest) in 2011; a difference of 0.48 units. Determining if such a change is meaningful is challenging, particularly when factoring in measurement error. A past review has indicated that the test-retest reliability of global 'life satisfaction' measures is only moderate, with correlations of repeat measures averaging $r = 0.59$ [260]. This indicates that the measurement error associated with subjective wellbeing measures may reduce the practical utility of our model predictions, considering the RMSE observed (i.e., any differences in wellbeing could be due to measurement error, rather than error associated with the model).

An alternative validation method to explore could involve conducting a cross-sectional analysis of the predictions across different demographic groups, and statistically comparing them against GSS data, similar to what was conducted in Chapter 4. The results in Chapter 6 alluded to the models' accuracy in this respect, but conducting this analysis formally is the next step. For example, the GSS data suggest people living in Auckland have lower life satisfaction scores compared to those living elsewhere, and if same trend is observed in the predicted scores, then the practical utility of the model is enhanced. The focus of this approach is on inter-group differences and trends rather than the absolute scores, which are arguably more valuable for actionable public policy initiatives, given the subjectivity of wellbeing measures.

***Subjectivity of the outcome variables***

The moderate performance exhibited by the models can be attributed to several factors, with one primary reason being the inherent nature of subjective data. The responses across the measurement scale for wellbeing outcomes did not follow a uniform distribution (i.e., an equal number of responses for each score), exemplified by the life satisfaction variable which ranged

from 0 to 10. Notably, over 50% of respondents reported scores of 7 or 8, showcasing a significant imbalance. This trend persisted in 2021, where more than 80% of individuals rated their overall life satisfaction at 7 or above [261]. While this distribution reflects the subjective nature of the question rather than a dataset limitation, it introduces challenges in developing sensitive models due to the pronounced imbalance, particularly for lower scores.

To address the class imbalance issue, we considered various strategies identified through the review conducted in Chapter 3. Several studies utilised resampling techniques to create a dataset with a uniform distribution, involving either up-sampling (increasing the minority class) or down-sampling (decreasing the majority class). However, due to concerns about potential information loss and model bias associated with resampling techniques [262], we decided against this approach. Instead, we opted for an alternative method by incorporating weights into our training process, aiming to mitigate the challenges posed by the imbalanced data. While applying weights is a common approach, future research could also explore other techniques such as synthetic data generation [263] to address class imbalance and assess model efficacy.

Understanding subjective wellbeing is inherently complex due to its reliance on self-reported responses, introducing variability based on individual interpretations of the question. In our models, approximately 92% of the variance in wellbeing scores remained unexplained. It is difficult to determine how much of this unexplained variance is due to individual interpretations of the question or scale. Unlike easily defined outcomes such as cardiovascular disease or obesity, subjective wellbeing lacks a tangible, universally understood metric. Two individuals selecting the same score may perceive it differently, contributing to the intricate nature of modelling and validating subjective outcomes. Additionally, a lower score does not necessarily indicate less satisfaction; rather, it may reflect an individual's nuanced

understanding of the scale. This inherent subjectivity complicates the modelling and validation of wellbeing outcomes.

While the challenges persist in modelling subjective wellbeing using objective socio-demographic and environmental variables, it remains uncertain whether the same difficulties extend to all subjective variables. Consider a hypothetical scenario: if we can successfully predict another wellbeing indicator, such as "loneliness," with precision, and given the strong correlation between loneliness and one's quality of life and overall wellbeing [264, 265], we could potentially employ this indicator as a predictor of subjective wellbeing [256]. The prospect of predicting one subjective attribute to subsequently infer subjective wellbeing outcomes represents an avenue for future researchers to explore.

Another promising approach involves creating a composite wellbeing score that considers all GSS wellbeing indicators, encompassing both subjective and objective measures across all 12 domains of the Living Standards Framework. Subsequently, researchers can predict this composite score using census-level features. We hypothesise this method as less challenging, as it not only incorporates the subjective components of wellbeing, but also comprises objective conditions.

The assessment of experienced wellbeing through Ecological Momentary Assessment (EMA) has also gained popularity in recent times [266-268]. EMA involves asking subjective wellbeing questions repeatedly within short time intervals, such as multiple times a day over a week. For instance, participants might be prompted with a question like "How happy do you feel?" three times a day at random moments. By leveraging the prevalence of smartphones and recent technological advancements, a mobile app could capture momentary assessments, providing a more nuanced understanding of participants' experienced wellbeing. Modelling these responses using appropriate predictor variables could offer valuable insights into

translating individual experiences into population-level wellbeing outcomes. Integrating EMA-type measures into the GSS and subsequently into the IDI could potentially enable the modelling of a much more precise measure of wellbeing at both individual and population levels.

**Study limitations**

While this thesis makes a novel attempt in predicting subjective wellbeing outcomes using census-level socio-demographic variables as predictors, it is crucial to acknowledge several limitations. Firstly, the study's focus was limited to predicting subjective wellbeing outcomes from the GSS. Given the broad concept of wellbeing, comprising several domains, we chose to examine only subjective wellbeing. This decision was influenced by the novelty of our research focus, as subjective wellbeing data are not readily available from other sources. However, it is important to note that other wellbeing indicators, such as "health status", can be precisely acquired from administrative health data or modelled using the National Health Survey.

Secondly, the review in Chapter 3 initially aimed to investigate the methodological decisions employed by studies predicting only subjective outcomes. However, due to the scarcity of studies in this domain, the review was extended to include all health outcomes, including the incidence of diseases. While these outcomes were objective, insights drawn from these studies informed our choice of appropriate methods for modelling various subjective wellbeing outcomes. In Chapter 4, the cross-sectional study was limited to two wellbeing outcomes (life satisfaction and life worthwhileness) due to the non-availability of the other two outcomes (family wellbeing and mental wellbeing) across more than one data collection wave.

Another limitation pertains to the usage of only Random Forest models for predicting subjective wellbeing outcomes (Chapter 5). While this decision was primarily driven by the

prominence of random forest models identified in the review (Chapter 3), other competent machine learning models, such as deep learning and gradient boosting, were not examined. The selection of an optimal machine learning model is dependent on the specific outcome of interest, and it is recognised that no single model suits all scenarios [269].

**Future Directions**

While many future directions were discussed throughout the thesis, this summary outlines key areas for future research in population-level wellbeing using predictive modelling. Firstly, investigating the potential of including additional predictor variables from administrative datasets, such as health records, while balancing data richness with sample size and availability, is essential. Another crucial area for future research is addressing class imbalance in subjective wellbeing data, which could be achieved by experimenting with resampling techniques or synthetic data generation methods to create a more balanced dataset. Finally, exploring alternative modelling strategies like deep learning and gradient boosting techniques could improve the accuracy and robustness of the predictions. These models excel at handling larger sets of predictors, suggesting their advantage may be evident when extensive feature sets from various sources within the IDI are employed in the modelling process.

Next, exploring the concept of predicting one subjective attribute (e.g., loneliness) to infer other subjective wellbeing outcomes presents a promising avenue for future work. Finally, expanding the scope of research beyond predicting subjective wellbeing outcomes from the GSS is crucial. Future studies should include other wellbeing domains and indicators, such as health status, and incorporate additional waves of GSS data into the modelling process. Additionally, incorporating a temporal element into cross-validation, such as training on one year and testing on another, could add value by demonstrating if models are robust to changes over time. Exploring these avenues of research not only holds promise for enhancing the

accuracy and depth of population-level wellbeing prediction but also contributes to a more comprehensive understanding subjective wellbeing assessment.

**Conclusion**

This body of work explored the feasibility of utilising census-level sociodemographic variables to predict GSS-based subjective wellbeing outcomes using Random Forest models. While these models showed promise by achieving RMSE values below 1.5 for life satisfaction, life worthwhileness, and family wellbeing (scale: 0 – 10), and an RMSE under 20 for mental wellbeing (scale: 0 – 100), there is significant potential for further improvement, particularly in enhancing the model's sensitivity to capture the broader variability among the outcome variables. Exploring alternative modelling and validation strategies holds the key to improving the overall capability of the model. The validation of census-level predictions against existing GSS measures across demographic groups revealed alignment with specific groups, although the predictions replicated the inherent trends seen in the GSS data. In light of the challenges posed by the inherent subjectivity of the outcomes, strategies such as integrating a broader array of administrative variables within the IDI and selecting predictors that closely represent the population group of interest may prove beneficial. Despite the acknowledged limitations, this research contributes to the development of a methodology involving predictive modelling of subjective wellbeing outcomes, thereby encouraging future exploration into diverse predictors and advanced techniques.

## References


1. Estes, R. J., & Sirgy, M. J. (2017). *The pursuit of human well-being: The untold global history*. Springer.
2. Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social Indicators Research*, *40*, 189-216. https://doi.org/10.1023/A:1006859511756
3. Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: does happiness lead to success? *Psychological Bulletin*, *131*(6), 803-855. https://doi.org/10.1037/0033-2909.131.6.803
4. Haddon, J. (2018). The impact of employees' well-being on performance in the workplace. *Strategic HR Review*, *17*(2), 72-75. https://doi.org/10.1108/SHR-01-2018-0009
5. Bian, Y., Zhang, L., & Gao, Y. (2020). Social bonding and subjective wellbeing: findings from the 2017 ISSP Module. *International Journal of Sociology*, *50*(1), 26-47. https://doi.org/10.1080/00207659.2019.1701320
6. Ed, D., & Martin, E. P. S. (2004). Beyond Money: Toward an Economy of Well-Being. *Psychological Science in the Public Interest*, *5*(1), 1. https://doi.org/10.1111/j.0963-7214.2004.00501001.x
7. Diener, E., Seligman, M. E. P., Choi, H., & Oishi, S. (2018). Happiest People Revisited. *Perspectives on Psychological Science*, *13*(2), 176-184. https://doi.org/10.1177/1745691617697077
8. Forgeard, M. J. C., Jayawickreme, E., Kern, M. L., & Seligman, M. E. P. (2011). Doing the right thing: Measuring wellbeing for public policy. *International journal of wellbeing*, *1*(1). https://doi.org/10.5502/ijw.v1i1.15
9. Diener, E. (2009). *Well-being for public policy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195334074.001.0001
10. NZ Treasury. (2018). *The Treasury Approach to the Living Standards Framework*. Wellington, New Zealand
11. Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542-575. https://doi.org/10.1037/0033-2909.95.3.542
12. Das, K. V., Jones-Harrell, C., Fan, Y., Ramaswami, A., Orlove, B., & Botchwey, N. (2020). Understanding subjective well-being: perspectives from psychology and

public health. *Public Health Reviews*, *41*(1), 25. https://doi.org/10.1186/s40985-020-00142-5

13. Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, *55*(1), 34. https://doi.org/10.1037/0003-066X.55.1.34
14. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (Second edition. ed.). Springer.
15. Felicia, A. H., & Timothy, T. C. S. (2013). Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being. *Social Indicators Research*, *110*(3), 837. https://doi.org/10.1007/s11205-011-9966-7
16. Gluckman, P. D. (2017). *Rethinking New Zealand's Approach to Mental Health and Mental Disorder: A Whole-of-government, Whole-of-nation Long-term Commitment*. Office of the Prime Minister's Chief Science Advisor.
17. Duncanson, M., Richardson, G., Oben, G., Wicken, A., van Asten, H., & Adams, J. (2020). *Child poverty monitor 2020: Technical report* (2357-2078). http://hdl.handle.net/10523/10585
18. Strategic Policy Branch. (2023). *What we know (and don't know) about economic growth in New Zealand*. Retrieved from https://www.mbie.govt.nz/dmsdocument/4028-what-we-know-and-dont-know-about-economic-growth-in-new-zealand
19. NZ Treasury. (2023). *Wellbeing in Aotearoa New Zealand 2022*. Retrieved from https://www.treasury.govt.nz/sites/default/files/2022-11/te-tai-waiora-2022.pdf
20. UNICEF New Zealand. *New Report Card shows that New Zealand is failing its Children*. https://www.unicef.org.nz/media-releases/new-report-card-shows-that-new-zealand-is-failing-its-children
21. Mackay, L., Schofield, G., Aaron, J., & Prendergast, K. (2015). *Sovereign wellbeing Index: 2015*.
22. Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist*, *62*(2), 95-108. https://doi.org/10.1037/0003-066X.62.2.95
23. Huppert, F. A., & So, T. T. C. (2013). Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being. *Social Indicators Research*, *110*(3), 837-861. https://doi.org/10.1007/s11205-011-9966-7

24. Sartorius, N. (2006). The meanings of health and its promotion. *Croatian medical journal*, *47*(4), 662-664.
25. Mackay, L., Egli, V., Booker, L.-J., & Prendergast, K. (2019). New Zealand's engagement with the Five Ways to Wellbeing: evidence from a large cross-sectional survey. *Kōtuitui: New Zealand Journal of Social Sciences Online*, *14*(2), 230-244. https://doi.org/10.1080/1177083X.2019.1603165
26. Easterlin, R. A. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In P. A. David & M. W. Reder (Eds.), *Nations and Households in Economic Growth* (pp. 89-125). Academic Press. https://doi.org/10.1016/B978-0-12-205050-3.50008-7
27. Campbell, A., Converse, P. E., & Rodgers, W. L. (1976). *The Quality of American Life: Perceptions, Evaluations, and Satisfactions*. Russell Sage Foundation.
28. Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*, *125*(2), 276-302. https://doi.org/10.1037/0033-2909.125.2.276
29. Jahoda, M. (1958). *Current concepts of positive mental health*. Basic Books. https://doi.org/10.1037/11258-000
30. Ryff, C. D. (1989). Beyond Ponce de Leon and Life Satisfaction: New Directions in Quest of Successful Ageing. *International Journal of Behavioral Development*, *12*(1), 35-55. https://doi.org/10.1177/016502548901200102
31. Ryan, R. M., & Deci, E. L. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, *52*(1), 141-166. https://doi.org/10.1146/annurev.psych.52.1.141
32. Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. Free Press.
33. Seligman, M. E. P. (2012). *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster.
34. Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.-w., Oishi, S., & Biswas-Diener, R. (2010). New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research*, *97*(2), 143-156. https://doi.org/10.1007/s11205-009-9493-y
35. Thompson, S., & Marks, N. (2008). *Measuring well-being in policy: issues and applications*. New Economics Foundation.

36. Boehm, J. K., & Kubzansky, L. D. (2012). The heart's content: the association between positive psychological well-being and cardiovascular health. *Psychological Bulletin*, *138*(4), 655. https://doi.org/10.1037/a0027448
37. Graham, C. (2012). *Happiness around the world: The paradox of happy peasants and miserable millionaires*. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199549054.001.0001
38. Cohen, S., Doyle, W. J., Turner, R. B., Alper, C. M., & Skoner, D. P. (2003). Emotional style and susceptibility to the common cold. *Psychosomatic medicine*, *65*(4), 652-657. https://doi.org/10.1097/01.psy.0000077508.57784.da
39. Cohen, S., & Pressman, S. D. (2006). Positive affect and health. *Current directions in psychological science*, *15*(3), 122-125. https://doi.org/10.1111/j.0963-7214.2006.00420.x
40. Blanchflower, D. G., & Oswald, A. J. (2008). Hypertension and happiness across nations. *Journal of health economics*, *27*(2), 218-233. https://doi.org/10.1016/j.jhealeco.2007.06.002
41. Department of Health and Social Care - UK Government. (2024). *Wellbeing and health policy*. Retrieved from https://www.gov.uk/government/publications/wellbeing-and-health-policy
42. Seligman, M. E. P., & Csikszentmihalyi, M. (2014). Positive psychology: An introduction. In *Flow and the foundations of positive psychology* (pp. 279-298). Springer. https://doi.org/10.1007/978-94-017-9088-8_18
43. OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being*. https://doi.org/doi:https://doi.org/10.1787/9789264191655-en
44. Organisation for Economic Co-operation Development. (2011). *How's life?: measuring well-being*. OECD Paris.
45. Durand, M. (2015). The OECD Better Life Initiative: How's Life? and the Measurement of Well-Being. *Review of Income and Wealth*, *61*(1), 4-17. https://doi.org/10.1111/roiw.12156
46. NZ Treasury. (2019). *The Wellbeing Budget 2019*. https://www.treasury.govt.nz/publications/wellbeing-budget/wellbeing-budget-2019
47. Stats NZ. (2018). *New Zealand General Social Survey 2018*. Wellington, New Zealand Retrieved from https://www.stats.govt.nz/information-releases/wellbeing-statistics-2018

48. Stats NZ. *General Social Survey (GSS)*. https://datainfoplus.stats.govt.nz/Item/nz.govt.stats/2ed50ad6-8ab8-47df-883d-210a51b50043#:~:text=The%20GSS%20uses%20a%20three,characteristics%20of%20the%20whole%20country.
49. Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417-440. https://doi.org/10.1146/annurev.ps.41.020190.002221
50. Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychol Bull*, *134*(1), 138-161. https://doi.org/10.1037/0033-2909.134.1.138
51. Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., Phung, D., Venkatesh, S., & Allender, S. (2015). Is Demography Destiny? Application of Machine Learning Techniques to Accurately Predict Population Health Outcomes from a Minimal Demographic Dataset. *PLOS ONE*, *10*(5), e0125602. https://doi.org/10.1371/journal.pone.0125602
52. Metzler, M. (2007). Social determinants of health: what, how, why, and now. *Preventing chronic disease*, *4*(4).
53. Link, B. G., & Phelan, J. C. (1996). Understanding sociodemographic differences in health--the role of fundamental social causes. *American journal of public health*, *86*(4), 471-473. https://doi.org/10.2105/ajph.86.4.471
54. Khumalo, I. P., Temane, Q. M., & Wissing, M. P. (2012). Socio-Demographic Variables, General Psychological Well-Being and the Mental Health Continuum in an African Context. *Social Indicators Research*, *105*(3), 419-442. https://doi.org/10.1007/s11205-010-9777-2
55. Sirgy, M. J. (2021). Effects of Demographic Factors on Wellbeing. In M. J. Sirgy (Ed.), *The Psychology of Quality of Life: Wellbeing and Positive Mental Health* (pp. 129-154). Springer International Publishing. https://doi.org/10.1007/978-3-030-71888-6_6
56. Mahindru, A., Patil, P., & Agrawal, V. (2023). Role of Physical Activity on Mental Health and Well-Being: A Review. *Cureus*, *15*(1), e33475. https://doi.org/10.7759/cureus.33475
57. Atkin, A. J., Adams, E., Bull, F. C., & Biddle, S. J. H. (2011). Non-Occupational Sitting and Mental Well-Being in Employed Adults. *Annals of Behavioral Medicine*, *43*(2), 181-188. https://doi.org/10.1007/s12160-011-9320-y

58. Gómez-Pinilla, F. (2008). Brain foods: the effects of nutrients on brain function. *Nature Reviews Neuroscience*, *9*(7), 568-578. https://doi.org/10.1038/nrn2421
59. Walker, M. (2017). *Why we sleep: The new science of sleep and dreams*. Penguin UK.
60. Livingston, V., Jackson-Nevels, B., & Reddy, V. V. (2022). Social, Cultural, and Economic Determinants of Well-Being. *Encyclopedia*, *2*(3), 1183-1199. https://doi.org/10.3390/encyclopedia2030079
61. Calderón-Garcidueñas, L., Torres-Jardón, R., Kulesza, R. J., Park, S.-B., & D'Angiulli, A. (2014). Air pollution and detrimental effects on children's brain. The need for a multidisciplinary approach to the issue complexity and challenges. *Frontiers in human neuroscience*, *8*, 613. https://doi.org/10.3389/fnhum.2014.00613
62. Lin, W.-H., Pan, W.-C., & Yi, C.-C. (2019). "Happiness in the air?" the effects of air pollution on adolescent happiness. *BMC Public Health*, *19*(1), 795. https://doi.org/10.1186/s12889-019-7119-0
63. Gong, Y., Palmer, S., Gallacher, J., Marsden, T., & Fone, D. (2016). A systematic review of the relationship between objective measurements of the urban environment and psychological distress. *Environment International*, *96*, 48-57. https://doi.org/10.1016/j.envint.2016.08.019
64. Engemann, K., Pedersen, C. B., Arge, L., Tsirogiannis, C., Mortensen, P. B., & Svenning, J.-C. (2019). Residential green space in childhood is associated with lower risk of psychiatric disorders from adolescence into adulthood. *Proceedings of the national academy of sciences*, *116*(11), 5188-5193. https://doi.org/10.1073/pnas.1807504116
65. Nutsford, D., Pearson, A. L., Kingham, S., & Reitsma, F. (2016). Residential exposure to visible blue space (but not green space) associated with lower psychological distress in a capital city. *Health & place*, *39*, 70-78. https://doi.org/10.1016/j.healthplace.2016.03.002
66. Nutsford, D., Pearson, A. L., & Kingham, S. (2013). An ecological study investigating the association between access to urban green space and mental health. *Public health*, *127*(11), 1005-1011. https://doi.org/10.1016/j.puhe.2013.08.016
67. Hobbs, M., Kingham, S., Wiki, J., Marek, L., & Campbell, M. (2021). Unhealthy environments are associated with adverse mental health and psychological distress: Cross-sectional evidence from nationally representative data in New Zealand. *Preventive Medicine*, *145*, 106416. https://doi.org/10.1016/j.ypmed.2020.106416

68. Stats NZ. (2016). Statistical standard for meshblock. *Wellington: Statistics New Zealand*.
69. June, A., & Tony, B. (2017). New Zealand's Integrated Data Infrastructure (IDI): Value to date and future opportunities. *International Journal of Population Data Science*, *1*(1). https://doi.org/10.23889/ijpds.v1i1.124
70. Black, A. (2016). *The IDI prototype spine's creation and coverage*. (StatisticsNew Zealand Working Paper No 16–03). www.stats.govt.nz
71. Statistics NZ. (2014). *Linking methodology used by statistics New Zealand in the Integrated Data Infrastructure project*. Statistics NZ. www.stats.govt.nz
72. Milne, B. J., Atkinson, J., Blakely, T., Day, H., Douwes, J., Gibb, S., Nicolson, M., Shackleton, N., Sporle, A., & Teng, A. (2019). Data resource profile: the New Zealand integrated data infrastructure (IDI). *International Journal of Epidemiology*, *48*(3), 677e. https://doi.org/10.1093/ije/dyz014
73. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499
74. van der Aalst, W. (2016). Data Science in Action. In W. van der Aalst (Ed.), *Process Mining: Data Science in Action* (pp. 3-23). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4_1
75. Hand, D. J., & Adams, N. M. Data Mining. In *Wiley StatsRef: Statistics Reference Online* (pp. 1-7). https://doi.org/10.1002/9781118445112.stat06466.pub2
76. Michele, B., Ewa, J. K., Karin, H., & Rajesh, M. (2022). Evaluating Similarities and Differences between Machine Learning and Traditional Statistical Modeling in Healthcare Analytics. In F. Marco Antonio Aceves & M. T.-G. Carlos (Eds.), *Artificial Intelligence Annual Volume 2022* (pp. Ch. 2). IntechOpen. https://doi.org/10.5772/intechopen.105116
77. Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, *1*(1), 117-138. https://doi.org/10.1007/s42081-018-0009-3
78. Esposito, A., Faundez-Zanuy, M., Morabito, F. C., & Pasero, E. (2023). *Applications of Artificial Intelligence and Neural Systems to Data Science* (Vol. 360). Springer.
79. Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., & Wrobel, S. (2020). *Visual analytics for data scientists*. Springer.
80. Mining, W. I. D. (2006). Data mining: Concepts and techniques. *Morgan Kaufinann*, *10*(559-569), 4.

81. Wil van der Aalst. (2016). *Process mining: data science in action*. Springer https://doi.org/10.1007/978-3-662-49851-4
82. Buisson, F. (2021). *Behavioral data analysis with R and Python*. O'Reilly Media, Inc.
83. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415
84. El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology* (pp. 3-11). Springer. https://doi.org/10.1007/978-3-319-18305-3_1
85. Teague, S. J., & Shatte, A. B. R. (2018). Exploring the transition to fatherhood: feasibility study using social media and machine learning. *JMIR pediatrics and parenting*, *1*(2), e12371. https://doi.org/10.2196/12371
86. Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2
87. Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130. https://doi.org/10.1007/978-3-031-01548-9
88. Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734
89. Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). Big Data Opportunities for Global Infectious Disease Surveillance. *PLOS Medicine*, *10*(4), e1001413. https://doi.org/10.1371/journal.pmed.1001413
90. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical informatics insights*, *8*, 1-10. https://doi.org/10.4137/BII.S31559
91. Jason Denzil, M., Emmalin, B., Meghan, O. N., Thomas, P., Vivek, G., Daniel, F., Kathy, K., & Laura, C. R. (2020). Predicting population health with machine learning: a scoping review. *BMJ Open*, *10*(10), e037860. https://doi.org/10.1136/bmjopen-2020-037860
92. Flaxman, A. D., & Vos, T. (2018). Machine learning in population health: Opportunities and threats. *PLOS Medicine*, *15*(11), e1002702. https://doi.org/10.1371/journal.pmed.1002702

93. Rose, S. (2013). Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *American Journal of Epidemiology*, *177*(5), 443-452. https://doi.org/10.1093/aje/kws241

94. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426-1448. https://doi.org/10.1017/S0033291719000151

95. Stats NZ. (2021). *Research using Stats NZ microdata*. https://statsnz.contentdm.oclc.org/digital/collection/p20045coll17

96. Chao, K., Sarker, M. N. I., Ali, I., Firdaus, R. B. R., Azman, A., & Shaed, M. M. (2023). Big data-driven public health policy making: Potential for the healthcare industry. *Heliyon*, *9*(9), e19681. https://doi.org/10.1016/j.heliyon.2023.e19681

97. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *24*(1), 198. https://doi.org/10.1093/jamia/ocw042

98. Jutte, D. P., Roos, L. L., & Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual review of public health*, *32*, 91-108. https://doi.org/10.1146/annurev-publhealth-031210-100700

99. Shin, G., Jarrahi, M. H., Fei, Y., Karami, A., Gafinowitz, N., Byun, A., & Lu, X. (2019). Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. *Journal of Biomedical Informatics*, *93*, 103153. https://doi.org/10.1016/j.jbi.2019.103153

100. Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Ginneken, B. V., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, *109*(5), 820-838. https://doi.org/10.1109/JPROC.2021.3054390

101. Mooney, S. J., & Pejaver, V. (2018). Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, *39*, 95-112. https://doi.org/10.1146/annurev-publhealth-040617-014208

102. Giest, S. (2017). Big data for policymaking: fad or fasttrack? *Policy Sciences*, *50*(3), 367-382. https://doi.org/10.1007/s11077-017-9293-1

103. Stats NZ. *Integrated Data Tools*. https://www.digital.govt.nz/showcase/integrated-data-tools/

104. Crown, W. H. (2015). Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions. *Value in Health*, *18*(2), 137-140. https://doi.org/10.1016/j.jval.2014.12.005

105. Radak, M., Lafta, H. Y., & Fallahi, H. (2023). Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies. *Journal of Cancer Research and Clinical Oncology*, *149*(12), 10473-10491. https://doi.org/10.1007/s00432-023-04956-z

106. Technologies, S. (2022). *NLP in healthcare: extracting insights from medical text*. Retrieved 24/11/2022 from https://stagezero.ai/blog/nlp-in-healthcare-extracting-insights-from-medical-text/

107. Huang, Y., Li, J., Li, M., & Aparasu, R. R. (2023). Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Medical Research Methodology*, *23*(1), 268. https://doi.org/10.1186/s12874-023-02078-1

108. Rodrigues, P. M., Madeiro, J. P., & Marques, J. A. L. (2023). Enhancing Health and Public Health through Machine Learning: Decision Support for Smarter Choices. *Bioengineering (Basel)*, *10*(7). https://doi.org/10.3390/bioengineering10070792

109. Mazzali, C., & Duca, P. (2015). Use of administrative data in healthcare research. *Internal and Emergency Medicine*, *10*(4), 517-524. https://doi.org/10.1007/s11739-015-1213-9

110. Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1-12. https://doi.org/10.1016/j.ssresearch.2016.04.015

111. Elias, P. (2014). Administrative data. *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences. Berlin: SCIVERO*, 47-48.

112. United Nations Economic Commission for Europe. (2007). *Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics* (Statistical standards and studies (Conference of European Statisticians), Issue.

113. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., & Weeks, L. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467-473. https://doi.org/10.7326/M18-0850

114. Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev*, *1*, 2. https://doi.org/10.1186/2046-4053-1-2

115. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine*, *162*(1), 55-63. https://doi.org/10.7326/M14-0697

116. Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., Phung, D., Venkatesh, S., & Allender, S. (2015). Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. *PLOS ONE*, *10*(5), Article e0125602. https://doi.org/10.1371/journal.pone.0125602

117. Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. *Psychiatry Investigation*, *15*(11), 1030-1036. https://doi.org/10.30773/pi.2018.08.27

118. Engchuan, W., Dimopoulos, A. C., Tyrovolas, S., Caballero, F. F., Sanchez-Niubo, A., Arndt, H., Ayuso-Mateos, J. L., Haro, J. M., Chatterji, S., & Panagiotakos, D. B. (2019). Sociodemographic indicators of health status using a machine learning approach and data from the english longitudinal study of aging (ELSA). *Medical Science Monitor*, *25*, 1994-2001. https://doi.org/10.12659/MSM.913283

119. Jung, J. S., Park, S. J., Kim, E. Y., Na, K. S., Kim, Y. J., & Kim, K. G. (2019). Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. *PLOS ONE*, *14*(6), Article e0217639. https://doi.org/10.1371/journal.pone.0217639

120. Nartowt, B. J., Hart, G. R., Roffman, D. A., Llor, X., Ali, I., Muhammad, W., Liang, Y., & Deng, J. (2019). Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLOS ONE*, *14*(8), Article e0221421. https://doi.org/10.1371/journal.pone.0221421

121. Stark, G. F., Hart, G. R., Nartowt, B. J., & Deng, J. (2019). Predicting breast cancer risk using personal health data and machine learning models. *PLOS ONE*, *14*(12), Article e0226765. https://doi.org/10.1371/journal.pone.0226765

122. Puterman, E., Weiss, J., Hives, B. A., Gemmill, A., Karasek, D., Mendes, W. B., & Rehkopf, D. H. (2020). Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proceedings of the National Academy of Sciences of the*

*United States of America*, *117*(28), 16273-16282. https://doi.org/10.1073/pnas.1918455117

123. Sow, B., Mukhtar, H., Ahmad, H. F., & Suguri, H. (2020). Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. *Informatics for Health and Social Care*, *45*(3), 229-241. https://doi.org/10.1080/17538157.2019.1582056

124. Worthington, M. A., Mandavia, A., & Richardson-Vejlgaard, R. (2020). Prospective prediction of PTSD diagnosis in a nationally representative sample using machine learning. *BMC Psychiatry*, *20*(1), Article 532. https://doi.org/10.1186/s12888-020-02933-1

125. Yang, H., & Bath, P. A. (2020). Predicting loneliness in older age using two measures of loneliness. *International Journal of Computers and Applications*, *42*(6), 602-615. https://doi.org/10.1080/1206212X.2018.1562408

126. Zhang, L., Shang, X., Sreedharan, S., Yan, X., Liu, J., Keel, S., Wu, J., Peng, W., & He, M. (2020). Predicting the development of type 2 diabetes in a large australian cohort using machine-learning techniques: Longitudinal survey study. *JMIR Medical Informatics*, *8*(7), Article e16850. https://doi.org/10.2196/16850

127. Buccheri, E., Dell'Aquila, D., & Russo, M. (2021). Artificial intelligence in health data analysis: The Darwinian evolution theory suggests an extremely simple and zero-cost large-scale screening tool for prediabetes and type 2 diabetes. *Diabetes Research and Clinical Practice*, *174*, Article 108722. https://doi.org/10.1016/j.diabres.2021.108722

128. Feng, C., & Jiao, J. (2021). Predicting and mapping neighborhood-scale health outcomes: A machine learning approach. *Computers, Environment and Urban Systems*, *85*, Article 101562. https://doi.org/10.1016/j.compenvurbsys.2020.101562

129. Makridis, C. A., Mudide, A., & Alterovitz, G. (2021). How Much Does the (Social) Environment Matter? Using Artificial Intelligence to Predict COVID-19 Outcomes with Socio-demographic Data. *Pacific Symposium on Biocomputing*, *26*, 328-335. https://doi.org/10.2139/ssrn.3706882

130. Li, Y., Liu, S. H., Niu, L., & Liu, B. (2019). Unhealthy Behaviors, Prevention Measures, and Neighborhood Cardiovascular Health: A Machine Learning Approach. *Journal of Public Health Management and Practice*, *25*(1), E25-E28. https://doi.org/10.1097/PHH.0000000000000817

131. Hu, L., Liu, B., & Li, Y. (2020). Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: A Bayesian machine learning approach. *Preventive Medicine*, *141*, Article 106240. https://doi.org/10.1016/j.ypmed.2020.106240

132. De La Garza, A. G., Blanco, C., Olfson, M., & Wall, M. M. (2021). Identification of Suicide Attempt Risk Factors in a National US Survey Using Machine Learning [Review]. *JAMA Psychiatry*, *78*(4), 398-406. https://doi.org/10.1001/jamapsychiatry.2020.4165

133. Ryu, S., Lee, H., Lee, D. K., Kim, S. W., & Kim, C. E. (2019). Detection of suicide attempters among suicide ideators using machine learning. *Psychiatry Investigation*, *16*(8), 588-593. https://doi.org/10.30773/pi.2019.06.19

134. Makridis, C. A., Zhao, D. Y., Bejan, C. A., & Alterovitz, G. (2021). Leveraging machine learning to characterize the role of socio-economic determinants on physical health and well-being among veterans. *COMPUTERS IN BIOLOGY AND MEDICINE*, *133*. https://doi.org/10.1016/j.compbiomed.2021.104354

135. Morrow, A. S., Campos Vega, A. D., Zhao, X., & Liriano, M. M. (2020). Leveraging Machine Learning to Identify Predictors of Receiving Psychosocial Treatment for Attention Deficit/Hyperactivity Disorder. *Administration and Policy in Mental Health and Mental Health Services Research*, *47*(5), 680-692. https://doi.org/10.1007/s10488-020-01045-y

136. Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2019). Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh. *Journal of Data Science*, *17*(1), 195-217. https://doi.org/10.6339/JDS.201901_17(1).0009

137. Weller, O., Sagers, L., Hanson, C., Barnes, M., Snell, Q., & Shannon Tass, E. (2021). Predicting suicidal thoughts and behavior among adolescents using the risk and protective factor framework: A large-scale machine learning approach. *PLOS ONE*, *16*(11). https://doi.org/10.1371/journal.pone.0258535

138. Hu, L. Y., Ji, J. Y., Li, Y., Liu, B., & Zhang, Y. Y. (2021). Quantile Regression Forests to Identify Determinants of Neighborhood Stroke Prevalence in 500 Cities in the USA: Implications for Neighborhoods with High Prevalence. *JOURNAL OF URBAN HEALTH-BULLETIN OF THE NEW YORK ACADEMY OF MEDICINE*, *98*(2), 259-270. https://doi.org/10.1007/s11524-020-00478-y

139. Nartowt, B. J., Hart, G. R., Muhammad, W., Liang, Y., Stark, G. F., & Deng, J. (2020). Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. *Frontiers in Big Data*, *3*. https://doi.org/10.3389/fdata.2020.00006

140. Hu, L., Liu, B., Ji, J., & Li, Y. (2020). Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *Journal of the American Heart Association*, *9*(22). https://doi.org/10.1161/JAHA.120.016745

141. Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022). The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*. https://doi.org/10.48550/arXiv.2207.14529

142. Li, D. C., Liu, C. W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Comput Biol Med*, *40*(5), 509-518. https://doi.org/10.1016/j.compbiomed.2010.03.005

143. Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

144. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10). Springer.

145. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29. https://doi.org/10.1145/1007730.1007735

146. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://doi.org/10.1613/jair.953

147. Agusta, Z. P. (2019). Modified balanced random forest for improving imbalanced data prediction. *International Journal of Advances in Intelligent Informatics*, *5*(1), 58-65. https://doi.org/10.26555/ijain.v5i1.255

148. Kang, H. (2013). The prevention and handling of the missing data. *Korean J Anesthesiol*, *64*(5), 402-406. https://doi.org/10.4097/kjae.2013.64.5.402

149. Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Mining*, *14*(1), 44. https://doi.org/10.1186/s13040-021-00274-7

150. Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*, *20*(1), 40-49. https://doi.org/10.1002/mpr.329

151. Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464-469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

152. Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. *Machine Learning Technical Reports*, *1*(1), 1-6. https://doi.org/10.13140/RG.2.2.28948.04489

153. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182. https://doi.org/10.1162/153244303322753616

154. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

155. Chan, J. Y., Leow, S. M., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8), 1283. https://doi.org/10.3390/math10081283

156. Mokdad, A. H., Marks, J. S., Stroup, D. F., & Gerberding, J. L. (2004). Actual Causes of Death in the United States, 2000. *JAMA*, *291*(10), 1238-1245. https://doi.org/10.1001/jama.291.10.1238

157. Ford, E. S., Bergmann, M. M., Boeing, H., Li, C., & Capewell, S. (2012). Healthy lifestyle behaviors and all-cause mortality among adults in the United States. *Preventive Medicine*, *55*(1), 23-27. https://doi.org/10.1016/j.ypmed.2012.04.016

158. World Health Organization. (2016). *Urban green spaces and health*. https://iris.who.int/handle/10665/345751

159. R Core Team. (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/

160. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., & Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*. https://doi.org/10.48550/arXiv.1603.04467

161. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine

learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://doi.org/10.48550/arXiv.1201.0490

162. Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*, *11*(4), 193. https://doi.org/10.3390/info11040193
163. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539
164. Weerts, H. J. P., Mueller, A. C., & Vanschoren, J. (2020). Importance of tuning hyperparameters of machine learning algorithms. *arXiv preprint arXiv:2007.07588*. https://doi.org/10.48550/arXiv.2007.07588
165. Corbin, J. H., Abdelaziz, F. B., Sørensen, K., Kökény, M., & Krech, R. (2021). Wellbeing as a policy framework for health promotion and sustainable development. *Health Promot Int*, *36*(Supplement_1), i64-i69. https://doi.org/10.1093/heapro/daab066
166. Costanza, R., Hart, M., Talberth, J., & Posner, S. (2009). *Beyond GDP: The need for new measures of progress* (The pardee papers, Issue. https://www.bu.edu/pardee/files/documents/PP-004-GDP.pdf
167. Diener, E., & Chan, M. Y. (2011). Happy People Live Longer: Subjective Well-Being Contributes to Health and Longevity. *Applied Psychology: Health and Well-Being*, *3*(1), 1-43-43. https://doi.org/10.1111/j.1758-0854.2010.01045.x
168. Isham, A., Mair, S., & Jackson, T. (2020). *Wellbeing and productivity: a review of the literature* (CUSP Working Paper Series, Issue. https://cusp.ac.uk/themes/aetw/wp22/
169. Diener, E. (2013). The Remarkable Changes in the Science of Subjective Well-Being. *Perspectives on Psychological Science*, *8*(6), 663-666. https://doi.org/10.1177/1745691613507583
170. Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, *22*(2), 53-72. https://doi.org/10.1257/jep.22.2.53.
171. Vik, M. H., & Carlquist, E. (2017). Measuring subjective well-being for policy purposes: The example of well-being indicators in the WHO "Health 2020" framework. *Scandinavian Journal of Public Health*, *46*(2), 279-286. https://doi.org/10.1177/1403494817724952
172. Diener, E. (2009). *The science of well-being: The collected works of Ed Diener* (Vol. 37). Springer. https://doi.org/10.1007/978-90-481-2350-6

173. Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and Open Questions in the Science of Subjective Well-Being. *Collabra: Psychology*, *4*(1), 15. https://doi.org/10.1525/collabra.115

174. Hicks, S., Tinkler, L., & Allin, P. (2013). Measuring Subjective Well-Being and its Potential Role in Policy: Perspectives from the UK Office for National Statistics. *Social Indicators Research*, *114*(1), 73-86. https://doi.org/10.1007/s11205-013-0384-x

175. Everett, G. (2015). Measuring National Well-Being: A UK Perspective. *Review of Income and Wealth*, *61*(1), 34-42. https://doi.org/10.1111/roiw.12175

176. Spence, A., Powell, M., & Self, A. (2011). Developing a framework for understanding and measuring national well-being. *Office for National Statistics*, 2-8.

177. Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2009). *Report by the commission on the measurement of economic performance and social progress*. The Commission Paris Retrieved from https://www.economie.gouv.fr/files/finances/presse/dossiers_de_presse/090914mesure_perf_eco_progres_social/synthese_ang.pdf

178. Radermacher, W. J. (2015). Recent and Future Developments Related to "Gdp and Beyond". *Review of Income and Wealth*, *61*(1), 18-24. https://doi.org/10.1111/roiw.12135

179. Diener, E., Pressman, S. D., Hunter, J., & Delgadillo-Chase, D. (2017). If, Why, and When Subjective Well-Being Influences Health, and Future Needed Research. *Appl Psychol Health Well Being*, *9*(2), 133-167. https://doi.org/10.1111/aphw.12090

180. Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*, *2*(4), 253-260. https://doi.org/10.1038/s41562-018-0307-6

181. Diener, E., & Diener, M. (1995). Cross-cultural correlates of life satisfaction and self-esteem. *Journal of personality and social psychology*, *68*(4), 653. https://doi.org/10.1037/0022-3514.68.4.653

182. Diener, E., & Lucas, R. E. (1999). Personality and subjective well-being. In *Well-being: The foundations of hedonic psychology.* (pp. 213-229). Russell Sage Foundation. https://doi.org/10.1007/978-90-481-2350-6_4

183. Frederick, S., & Loewenstein, G. (1999). Hedonic adaptation. In *Well-being: The foundations of hedonic psychology.* (pp. 302-329). Russell Sage Foundation.

184. European Social Survey. (2018). *European Social Survey Round 10 Data (Version 10.0)*. http://www.europeansocialsurvey.org/.

185. Moro-Egido, A. l., Navarro, M., & Sánchez, A. (2022). Changes in Subjective Well-Being Over Time: Economic and Social Resources do Matter. *Journal of Happiness Studies*, *23*(5), 2009-2038. https://doi.org/10.1007/s10902-021-00473-3

186. Yap, S. C. Y., Anusic, I., & Lucas, R. E. (2014). Chapter 7 - Does Happiness Change? Evidence from Longitudinal Studies. In K. M. Sheldon & R. E. Lucas (Eds.), *Stability of Happiness* (pp. 127-145). Academic Press. https://doi.org/10.1016/B978-0-12-411478-4.00007-2

187. Alesina, A., Di Tella, R., & MacCulloch, R. (2004). Inequality and happiness: are Europeans and Americans different? *Journal of Public Economics*, *88*(9), 2009-2042. https://doi.org/10.1016/j.jpubeco.2003.07.006

188. Bartolini, S., & Sarracino, F. (2015). The Dark Side of Chinese Growth: Declining Social Capital and Well-Being in Times of Economic Boom. *World Development*, *74*, 333-351. https://doi.org/10.1016/j.worlddev.2015.05.010

189. Maggino, F., & Facioni, C. (2017). Measuring Stability and Change: Methodological Issues in Quality of Life studies. *Social Indicators Research*, *130*(1), 161-187. https://doi.org/10.1007/s11205-015-1129-9

190. Stats NZ. (2018). *About the General Social Survey*. https://www.stats.govt.nz/help-with-surveys/list-of-stats-nz-surveys/about-the-general-social-survey/

191. Pavot, W., & Diener, E. (2008). The Satisfaction With Life Scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology*, *3*(2), 137-152. https://doi.org/10.1080/17439760701756946

192. Hancock, K., Sherar, L., & Downward, P. (2021). *Life satisfaction, worthwhileness of life and leisure activities among older people: assessing the mediating effect of self-reported health limitations*. https://doi.org/10.21203/rs.3.rs-405611/v1

193. Mintrom, M. (2019). New Zealand's Wellbeing Budget Invests in Population Health. *Milbank Q*, *97*(4), 893-896. https://doi.org/10.1111/1468-0009.12409

194. UK's Office for National Statistics. (2018). *Personal well-being in the UK QMI.* https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingintheukqmi

195. Statistics NZ. (2020). *Microdata output guide (fifth edition).* www.stats.govt.nz

196. Lucas, R. E. (2007). Adaptation and the Set-Point Model of Subjective Well-Being:Does Happiness Change After Major Life Events? *Current directions in psychological science*, *16*(2), 75-79. https://doi.org/10.1111/j.1467-8721.2007.00479.x

197. Kaiser, M., Otterbach, S., & Sousa-Poza, A. (2022). Using machine learning to uncover the relation between age and life satisfaction. *Scientific Reports*, *12*(1), 5263. https://doi.org/10.1038/s41598-022-09018-x

198. Stone, A. A., Schwartz, J. E., Broderick, J. E., & Deaton, A. (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the national academy of sciences*, *107*(22), 9985-9990. https://doi.org/10.1073/pnas.100374410

199. Jarden, R. J., Joshanloo, M., Weijers, D., Sandham, M. H., & Jarden, A. J. (2022). Predictors of Life Satisfaction in New Zealand: Analysis of a National Dataset. *Int J Environ Res Public Health*, *19*(9). https://doi.org/10.3390/ijerph19095612

200. Galambos, N. L., Krahn, H. J., Johnson, M. D., & Lachman, M. E. (2020). The U Shape of Happiness Across the Life Course: Expanding the Discussion. *Perspect Psychol Sci*, *15*(4), 898-912. https://doi.org/10.1177/1745691620902428

201. Laaksonen, S. (2018). A research note: Happiness by age is more complex than U-shaped. *Journal of Happiness Studies*, *19*, 471-482. https://doi.org/10.1007/s10902-016-9830-1

202. Buecker, S., Luhmann, M., Haehner, P., Bühler, J. L., Dapp, L. C., Luciano, E. C., & Orth, U. (2023). The development of subjective well-being across the life span: A meta-analytic review of longitudinal studies. *Psychological Bulletin*, *149*(7-8), 418-446. https://doi.org/10.1037/bul0000401

203. Joshanloo, M., & Jovanović, V. (2020). The relationship between gender and life satisfaction: analysis across demographic groups and global regions. *Archives of Women's Mental Health*, *23*(3), 331-338. https://doi.org/10.1007/s00737-019-00998-w

204. Batz-Barbarich, C., Tay, L., Kuykendall, L., & Cheung, H. K. (2018). A meta-analysis of gender differences in subjective well-being: Estimating effect sizes and associations with gender inequality. *Psychological Science*, *29*(9), 1491-1503. https://doi.org/10.1177/0956797618774796

205. Haring, M. J., Stock, W. A., & Okun, M. A. (1984). A research synthesis of gender and social class as correlates of subjective well-being. *Human relations*, *37*(8), 645-657. https://doi.org/10.1177/001872678403700805

206. Wood, W., Rhodes, N., & Whelan, M. (1989). Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological Bulletin*, *106*(2), 249. https://doi.org/10.1037/0033-2909.106.2.249

207. Bartram, D. (2022). The 'Gender Life-Satisfaction/Depression Paradox' Is an Artefact of Inappropriate Control Variables. *Social Indicators Research*, *164*(3), 1061-1072. https://doi.org/10.1007/s11205-022-02986-7

208. Tesch-Römer, C., Motel-Klingebiel, A., & Tomasik, M. J. (2008). Gender Differences in Subjective Well-Being: Comparing Societies with Respect to Gender Equality. *Social Indicators Research*, *85*(2), 329-349. https://doi.org/10.1007/s11205-007-9133-3

209. Montgomery, M. (2022). Reversing the gender gap in happiness. *Journal of Economic Behavior & Organization*, *196*, 65-78. https://doi.org/10.1016/j.jebo.2022.01.006

210. Houkamau, C. A., & Sibley, C. G. (2010). The multi-dimensional model of Māori identity and cultural engagement. *New Zealand Journal of Psychology*, *39*(1), 8-28.

211. Pere, L. M. (2006). *Oho mauri: cultural identity, wellbeing, and tāngata whai ora/motuhake.* Massey University]. Wellington, New Zealand. https://mro.massey.ac.nz/server/api/core/bitstreams/c71bcc9d-bb90-442c-98e6-1764422f8108/content

212. Russell, L., Smiler, K., & Stace, H. (2013). Improving Māori health and reducing inequalities between Māori and non-Māori: has the primary health care strategy worked for Māori. *New Zealand: Health Research Council Of New Zealand and The Ministry Of Health Wellington*.

213. Russell, L. (2018). Te oranga hinengaro: Report on Māori mental wellbeing results from the New Zealand mental health monitor & health and lifestyles survey. *Health Promotion Agency/Te Hiringa Hauora*.

214. Australian Unity Wellbeing Index. (2021). *Location, location, location: Does where you live affect your wellbeing?* Retrieved 30/12/2023 from https://www.australianunity.com.au/wellbeing/what-is-real-wellbeing/does-where-you-live-affect-your-wellbeing

215. Australian Unity Wellbeing Index. (2021). *The location effect: why where you live matters*. https://www.australianunity.com.au/wellbeing/what-is-real-wellbeing/why-where-you-live-matters

216. Biswas-Diener, R., & Diener, E. (2001). Will money increase subjective well-being? A literature review and guide to needed research. *Social Research Indicators*, *57*, 119-169. https://doi.org/10.1023/A:1014411319119

217. Cheung, F., & Lucas, R. E. (2015). When does money matter most? Examining the association between income and life satisfaction over the life course. *Psychology and aging*, *30*(1), 120. https://doi.org/10.1037/a0038682
218. Stevenson, B., & Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin paradox. *National Bureau of Economic Research.* https://doi.org/10.3386/w14282
219. Lucas, R. E., & Schimmack, U. (2009). Income and well-being: How big is the gap between the rich and the poor? *Journal of Research in Personality*, *43*(1), 75-78. https://doi.org/10.1016/j.jrp.2008.09.004
220. Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proc Natl Acad Sci U S A*, *107*(38), 16489-16493. https://doi.org/10.1073/pnas.1011492107
221. Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, *27*(1), 35-47. https://doi.org/10.1016/0167-2681(95)00003-B
222. Khalil, E. L. (2022). Solving the income-happiness paradox. *International Review of Economics*, *69*(3), 433-463. https://doi.org/10.1007/s12232-022-00398-0
223. Clark, A. E., Frijters, P., & Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. *Journal of Economic literature*, *46*(1), 95-144. https://doi.org/10.1257/jel.46.1.95
224. Haller, M., & Hadler, M. (2006). How Social Relations and Structures can Produce Happiness and Unhappiness: An International Comparative Analysis. *Social Indicators Research*, *75*(2), 169-216. https://doi.org/10.1007/s11205-004-6297-y
225. Coscieme, L., Sutton, P., Mortensen, L. F., Kubiszewski, I., Costanza, R., Trebeck, K., Pulselli, F. M., Giannetti, B. F., & Fioramonti, L. (2019). Overcoming the Myths of Mainstream Economics to Enable a New Wellbeing Economy. *Sustainability*, *11*(16). https://doi.org/10.3390/su11164374
226. Kvalsvig, A., Gibb, S., & Teng, A. (2019). Linkage error and linkage bias: A guide for IDI users. *University of Otago*.
227. Stats NZ. *Census*. https://www.stats.govt.nz/topics/census
228. Marek, L., Hobbs, M., Wiki, J., Kingham, S., & Campbell, M. (2021). The good, the bad, and the environment: developing an area-based measure of access to health-promoting and health-constraining environments in New Zealand. *International*

*Journal of Health Geographics*, *20*, 1-20. https://doi.org/10.1186/s12942-021-00269-x

229. Staehr, J. K. (1998). The use of well-being measures in primary health care-the DepCare project. *World Health Organization, Regional Office for Europe: Well-Being Measures in Primary Health Care-the DepCare Project. Geneva: World Health Organization*.
230. Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: a systematic review of the literature. *Psychotherapy and psychosomatics*, *84*(3), 167-176. https://karger.com/pps/article-pdf/84/3/167/3480704/000376585.pdf
231. Stats NZ. (2023). *Census*. https://www.census.govt.nz/
232. Atkinson, J., Salmond, C., & Crampton, P. (2019). *NZDep2018 Index of Deprivation, Interim Research Report*. Wellington: University of Otago.
233. Miller, A. (2002). *Subset selection in regression*. CRC Press.
234. Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.
235. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
236. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.
237. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792. https://doi.org/10.1890/07-0539.1
238. Evans, G. W. (2003). The built environment and mental health. *Journal of urban health*, *80*(4), 536-555. https://doi.org/10.1093/jurban/jtg063
239. Hartig, T., Mitchell, R., de Vries, S., & Frumkin, H. (2014). Nature and Health. *Annual review of public health*, *35*(Volume 35, 2014), 207-228. https://doi.org/10.1146/annurev-publhealth-032013-182443
240. Stats NZ. (2018). *2018 Census collection response rates unacceptably low*. https://www.stats.govt.nz/methods/2018-census-collection-response-rates-unacceptably-low

241. Cooke, P. J., Melchert, T. P., & Connor, K. (2016). Measuring well-being: A review of instruments. *The Counseling Psychologist*, *44*(5), 730-757. https://doi.org/10.1177/0011000016633507

242. Social Investment Agency. (2019). *Measuring the impact of social housing placement on wellbeing.* Wellington, New Zealand

243. Peeters, M. J. (2016). Practical significance: Moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning*, *8*(1), 83-89. https://doi.org/10.1016/j.cptl.2015.09.001

244. Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759. https://doi.org/10.1177/0013164496056005002

245. Mohajeri, K., Mesgari, M., & Lee, A. S. (2020). When Statistical Significance Is Not Enough: Investigating Relevance, Practical Significance, and Statistical Significance. *MIS Quarterly*, *44*(2). https://doi.org/10.25300/MISQ/2020/13932

246. Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution* (pp. 196-202). Springer.

247. Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ*, *4*(3), 279-282. https://doi.org/10.4300/jgme-d-12-00156.1

248. Osborne, J. W. (2008). *Best practices in quantitative methods*. Sage.

249. Diener, E. D., & Suh, M. E. (1997). Subjective well-being and age: An international analysis. *Annual review of gerontology and geriatrics*, *17*(1), 304-324. https://doi.org/10.1891/0198-8794.17.1.304

250. Huppert, F. A., Marks, N., Clark, A., Siegrist, J., Stutzer, A., Vittersø, J., & Wahrendorf, M. (2009). Measuring well-being across Europe: Description of the ESS well-being module and preliminary findings. *Social Indicators Research*, *91*, 301-315. https://doi.org/10.1007/s11205-008-9346-0

251. Dolan, P., & Metcalfe, R. (2012). Measuring Subjective Wellbeing: Recommendations on Measures for use by National Governments. *Journal of Social Policy*, *41*, 409 - 427. https://doi.org/10.1017/S0047279411000833

252. Kahneman, D., & Krueger, A. B. (2006). Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspectives*, *20*(1), 3-24. https://doi.org/10.1257/089533006776526030

253. Health Data Research UK. (2024). *NHS Winter pressures could be addressed by research harnessing big data, machine learning and artificial intelligence*. Retrieved

26/12/2023 from https://www.hdruk.ac.uk/news/nhs-winter-pressures-could-be-addressed-by-research-harnessing-big-data-machine-learning-and-artificial-intelligence/

254. Australian Government. (2023). *National Real Time Prescription Monitoring (RTPM)*. https://www.health.gov.au/our-work/national-real-time-prescription-monitoring-rtpm

255. Singapore Government. (2024). *Smart Health Initiatives*. https://www.smartnation.gov.sg/initiatives/health/

256. Vera Cruz, G., Maurice, T., Moore, P. J., & Rohrbeck, C. A. (2023). Using artificial intelligence to identify the top 50 independent predictors of subjective well-being in a multinational sample of 37,991 older European & Israeli adults. *Scientific Reports*, *13*(1), 11352. https://doi.org/10.1038/s41598-023-38337-w

257. Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and techniques in modern geography*.

258. Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

259. Murti, D. M. P., Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019, 23-24 Oct. 2019). K-Nearest Neighbor (K-NN) based Missing Data Imputation. 2019 5th International Conference on Science in Information Technology (ICSITech),

260. Krueger, A. B., & Schkade, D. A. (2008). The Reliability of Subjective Well-Being Measures. *J Public Econ*, *92*(8-9), 1833-1845. https://doi.org/10.1016/j.jpubeco.2007.12.015

261. Stats NZ. (2022). *Wellbeing statistics: 2021*. Retrieved 15/06/2023 from https://www.stats.govt.nz/information-releases/wellbeing-statistics-2021/

262. Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. *Procedia Computer Science*, *176*, 420-429. https://doi.org/10.1016/j.procs.2020.08.043

263. Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: a narrative review. *PLOS Digital Health*, *2*(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082

264. Richard, A., Rohrmann, S., Vandeleur, C. L., Schmid, M., Barth, J., & Eichholzer, M. (2017). Loneliness is adversely associated with physical and mental health and

lifestyle factors: Results from a Swiss national survey. *PLOS ONE*, *12*(7), e0181442. https://doi.org/10.1371/journal.pone.0181442

265. Beutel, M. E., Klein, E. M., Brähler, E., Reiner, I., Jünger, C., Michal, M., Wiltink, J., Wild, P. S., Münzel, T., Lackner, K. J., & Tibubos, A. N. (2017). Loneliness in the general population: prevalence, determinants and relations to mental health. *BMC Psychiatry*, *17*(1), 97. https://doi.org/10.1186/s12888-017-1262-x

266. Dejonckheere, E., Mestdagh, M., Verdonck, S., Lafit, G., Ceulemans, E., Bastian, B., & Kalokerinos, E. K. (2021). The relation between positive and negative affect becomes more negative in response to personally relevant events. *Emotion*, *21*(2), 326. https://doi.org/10.1037/emo0000697

267. Mengelkoch, S., Moriarity, D. P., Novak, A. M., Snyder, M. P., Slavich, G. M., & Lev-Ari, S. (2024). Using Ecological Momentary Assessments to Study How Daily Fluctuations in Psychological States Impact Stress, Well-Being, and Health. *Journal of Clinical Medicine*, *13*(1), 24. https://doi.org/10.3390/jcm13010024

268. de Vries, L. P., Baselmans, B. M. L., & Bartels, M. (2021). Smartphone-Based Ecological Momentary Assessment of Well-Being: A Systematic Review and Recommendations for Future Studies. *J Happiness Stud*, *22*(5), 2361-2408. https://doi.org/10.1007/s10902-020-00324-7

269. Doupe, P., Faghmous, J., & Basu, S. (2019). Machine Learning for Health Services Researchers. *Value in Health*, *22*(7), 808-815. https://doi.org/10.1016/j.jval.2019.02.012

# Appendices

## Appendix A. Tables

**Table A-1.** Evaluation of methodological quality in the reviewed studies, adapted from TRIPOD.

| Ref | Data | Participants | Data Preparation | Outcome | Predictors | Class Imbalance | Sample Size | Missing Data | Analytical Methods | Model Output | Validation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [116] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *NA* | *Yes* | *Yes* | *Yes* |
| [117] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [118] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *NA* | *Yes* | *No* | *Yes* | *Yes* | *Yes* |
| [119] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *Yes* | *Yes* |
| [120] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [121] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Partially* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [122] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [123] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [124] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [125] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [126] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *No* | *Yes* | *Yes* | *Yes* |
| [127] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [128] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *NA* | *Yes* | *Yes* | *Yes* |
| [129] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *NA* | *Yes* | *Yes* | *Yes* |
| [130] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *Yes* | *Yes* | *Partially* | *No* |
| [131] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *Yes* | *Yes* | *Partially* | *No* |
| [132] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [133] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [134] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *No* | *Yes* | *Yes* | *Yes* |
| [135] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [136] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Partially* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| [137] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *No* | *Yes* | *Yes* | *Yes* |
| [138] | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Partially* | *Yes* | *Yes* | *Yes* | *Partially* | *No* |
| [139] | *Yes* | *Partially* | *Yes* | *Yes* | *Yes* | *No* | *No* | *Yes* | *Yes* | *Yes* | *Yes* |
| [140] | *Yes* | *NA* | *Yes* | *Yes* | *Yes* | *NA* | *NA* | *Yes* | *Yes* | *Yes* | *Yes* |

**Table A-2.** Number of valid datapoints in the GSS across each variable (N).

| | | 2014 | 2016 | 2018 |
|---|---|---|---|---|
| ***Demographic variables*** | Age range | 8817 | 8529 | 8850 |
| | Gender | 8820 | 8529 | 8850 |
| | Ethnicity | 8820 | 8532 | 8850 |
| | Region | 8817 | 8532 | 8850 |
| | Personal income | 8820 | 8532 | 8850 |
| | Household income | 8820 | 8529 | 8850 |
| | Household size | 8820 | 8532 | 8850 |
| ***Outcome variables*** | Life satisfaction | 8799 | 8517 | 8838 |
| | Life worthwhileness | 8787 | 8511 | 8826 |

***Note:*** *Excluded people who answered - Don't know, Refused to Answer, Missing/Not stated/Not Applicable*

**Table A-3.** Descriptive statistics of the outcome life satisfaction and life worthwhileness

| | | Life satisfaction | | | Life worthwhileness | | |
|---|---|---|---|---|---|---|---|
| **Variable** | **Variable category** | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** |
| **Age range** | 15-24 years | 2727 | 7.72 | 1.65 | 2724 | 7.86 | 1.72 |
| | 25-34 years | 4107 | 7.62 | 1.66 | 4104 | 8.01 | 1.57 |
| | 35-44 years | 4221 | 7.5 | 1.73 | 4218 | 8.01 | 1.59 |
| | 45-54 years | 4533 | 7.44 | 1.88 | 4527 | 7.93 | 1.72 |
| | 55-64 years | 4299 | 7.54 | 1.97 | 4302 | 8.07 | 1.77 |
| | 65 years or over | 6276 | 8.15 | 1.85 | 6249 | 8.38 | 1.69 |
| **Gender** | Male | 11856 | 7.66 | 1.83 | 11841 | 7.96 | 1.69 |
| | Female | 14310 | 7.73 | 1.83 | 14283 | 8.18 | 1.69 |
| **Ethnicity** | New Zealand European | 18000 | 7.73 | 1.8 | 17982 | 8.12 | 1.66 |
| | New Zealand Māori | 2046 | 7.61 | 1.96 | 2043 | 8.09 | 1.82 |
| | Pacific | 1173 | 7.67 | 1.89 | 1173 | 7.97 | 1.74 |
| | Asian | 2391 | 7.73 | 1.79 | 2385 | 7.94 | 1.66 |
| | MELAA / Other | 2541 | 7.53 | 1.9 | 2535 | 7.99 | 1.77 |
| **Region** | Auckland | 6780 | 7.66 | 1.77 | 6771 | 7.97 | 1.65 |
| | Wellington | 3114 | 7.69 | 1.75 | 3105 | 8.05 | 1.63 |
| | Northland group | 3282 | 7.76 | 1.87 | 3282 | 8.2 | 1.78 |
| | Rest of North Island | 5958 | 7.71 | 1.88 | 5943 | 8.16 | 1.72 |
| | Canterbury | 3642 | 7.61 | 1.88 | 3639 | 8.04 | 1.72 |
| | Rest of South Island | 3393 | 7.74 | 1.83 | 3387 | 8.13 | 1.66 |
| **Personal income** | Loss/ $0 - $30,000 | 12069 | 7.59 | 2.02 | 12048 | 7.98 | 1.89 |
| | $30,001 - $35,000 | 1404 | 7.64 | 1.92 | 1404 | 8.1 | 1.74 |
| | $35,001 - $40,000 | 1587 | 7.63 | 1.8 | 1584 | 8.11 | 1.6 |
| | $40,001 - $50,000 | 2637 | 7.67 | 1.71 | 2631 | 8.06 | 1.57 |
| | $50,001 - $60,000 | 2256 | 7.76 | 1.65 | 2256 | 8.15 | 1.48 |
| | $60,001 - $70,000 | 1695 | 7.87 | 1.53 | 1695 | 8.21 | 1.39 |
| | $70,001 - $100,000 | 2523 | 7.86 | 1.44 | 2517 | 8.2 | 1.4 |
| | $100,001 or more | 1992 | 8.01 | 1.48 | 1986 | 8.37 | 1.33 |
| **Household income** | Loss/ $0 - $30,000 | 5226 | 7.35 | 2.18 | 5211 | 7.83 | 2.05 |
| | $30,001 - $35,000 | 1188 | 7.66 | 2.07 | 1185 | 8.09 | 1.85 |
| | $35,001 - $40,000 | 1080 | 7.58 | 1.96 | 1074 | 7.99 | 1.74 |
| | $40,001 - $50,000 | 2001 | 7.61 | 1.93 | 1995 | 8.05 | 1.71 |
| | $50,001 - $60,000 | 1932 | 7.6 | 1.88 | 1935 | 8.07 | 1.67 |
| | $60,001 - $70,000 | 1758 | 7.63 | 1.81 | 1758 | 8 | 1.65 |
| | $70,001 - $100,000 | 4701 | 7.76 | 1.67 | 4701 | 8.12 | 1.58 |
| | $100,001 or more | 8271 | 7.95 | 1.52 | 8262 | 8.25 | 1.44 |
| **Household size** | One person | 6474 | 7.47 | 2.04 | 6450 | 7.9 | 1.9 |
| | Two people | 9021 | 7.88 | 1.79 | 9021 | 8.19 | 1.65 |
| | Three people | 4071 | 7.59 | 1.76 | 4065 | 8.06 | 1.61 |
| | Four people | 3855 | 7.71 | 1.65 | 3858 | 8.13 | 1.53 |
| | Five people | 1680 | 7.8 | 1.7 | 1674 | 8.13 | 1.58 |
| | Six or more people | 1056 | 7.73 | 1.78 | 1059 | 8.09 | 1.68 |

**Table A-4.** Demographic data distribution comparison of the GSS 2018 dataset with the Modelling dataset

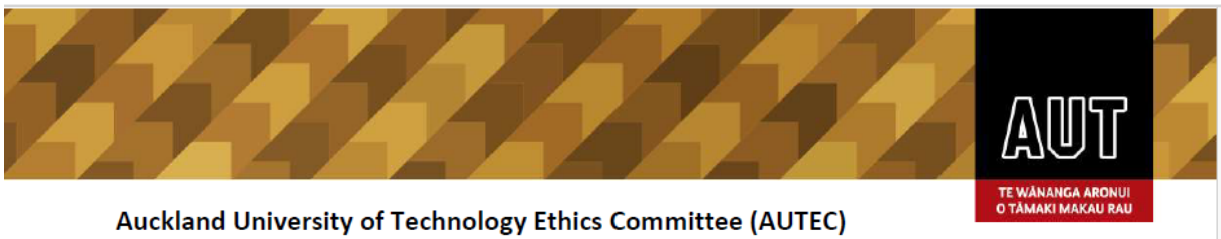
| Variable | Variable choice | GSS 2018 dataset distribution (N = 8,793) | Modelling dataset distribution (i.e., after data cleaning, N = 5,658) |
|---|---|---|---|
| Age_range | 15–24 years | 10% | 9% |
| Age_range | 25–34 years | 16% | 15% |
| Age_range | 35–44 years | 16% | 16% |
| Age_range | 45–54 years | 17% | 17% |
| Age_range | 55–64 years | 17% | 18% |
| Age_range | 65 years or over | 24% | 25% |
| Gender | Male | 45% | 44% |
| Gender | Female | 55% | 56% |
| Ethnicity | New Zealand European | 67% | 74% |
| Ethnicity | New Zealand Māori | 7% | 4% |
| Ethnicity | Pacific | 4% | 3% |
| Ethnicity | Asian | 10% | 11% |
| Ethnicity | Other | 11% | 9% |
| Region | Auckland | 27% | 27% |
| Region | Wellington | 11% | 12% |
| Region | Northland group | 12% | 11% |
| Region | Rest of North Island | 25% | 24% |
| Region | Canterbury | 13% | 14% |
| Region | Rest of South Island | 13% | 13% |

**Table A-5.** Top 10 important predictors in prediction of outcome variables.

| Outcome variable | Top 10 important predictor variables | Importance score |
|---|---|---|
| ***Life satisfaction*** | age_group | 100 |
| | deprivation_score | 84.17 |
| | Bluespace_rank_dec | 54.45 |
| | FruitVeg_rank_dec | 54.37 |
| | Env_dec | 53.58 |
| | AlcoholOutlets_rank_dec | 53.34 |
| | Supermarket_rank_dec | 53.22 |
| | deprivation_code | 53.1 |
| | Greenspace_rank_dec | 52.77 |
| | PhysicalActivity_rank_dec | 52.72 |
| ***Life worthwhileness*** | age_group | 100 |
| | deprivation_score | 85.78 |
| | Bluespace_rank_dec | 55.69 |
| | Greenspace_rank_dec | 54.77 |
| | AlcoholOutlets_rank_dec | 54.68 |
| | Env_dec | 54.65 |
| | deprivation_code | 54.59 |
| | FruitVeg_rank_dec | 54.46 |
| | PhysicalActivity_rank_dec | 54.23 |
| | Goods_dec | 53.87 |
| ***Family wellbeing*** | age_group | 100 |
| | deprivation_score | 83.07 |
| | deprivation_code | 56.84 |
| | Bluespace_rank_dec | 56.81 |
| | Greenspace_rank_dec | 55.95 |
| | Goods_dec | 55.76 |
| | Supermarket_rank_dec | 55.71 |
| | Env_dec | 55.7 |
| | AlcoholOutlets_rank_dec | 55.59 |
| | FruitVeg_rank_dec | 55.4 |
| ***Mental wellbeing*** | age_group | 100 |
| | deprivation_score | 80.05 |
| | FruitVeg_rank_dec | 53.61 |
| | Bluespace_rank_dec | 52.13 |
| | AlcoholOutlets_rank_dec | 51.72 |
| | PhysicalActivity_rank_dec | 51.52 |
| | Env_dec | 51.44 |
| | Goods_dec | 51.12 |
| | Greenspace_rank_dec | 50.92 |
| | Supermarket_rank_dec | 50.64 |

***Note:*** *All scores are scaled gini coefficients, presented relative to the first variable 'age_group', which has a score of 100. A description of these variables are presented in Tables 5-2 and 5-3*

# Appendix B. Ethical approval




## Auckland University of Technology Ethics Committee (AUTEC)

Auckland University of Technology
D-88, Private Bag 92006, Auckland 1142, NZ
T: +64 9 921 9999 ext. 8316
E: ethics@aut.ac.nz
www.aut.ac.nz/researchethics

18 May 2021

Tom Stewart
Faculty of Health and Environmental Sciences

Dear Tom

Re Ethics Application: **21/115 Application of data science methodologies to explore, predict, and model wellbeing outcomes using the New Zealand Integrated Data Infrastructure (IDI).**

Thank you for providing evidence as requested, which satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC).

Your ethics application has been approved for three years until 18 May 2024.

**Standard Conditions of Approval**

1. The research is to be undertaken in accordance with the Auckland University of Technology Code of Conduct for Research and as approved by AUTEC in this application.
2. A progress report is due annually on the anniversary of the approval date, using the EA2 form.
3. A final report is due at the expiration of the approval period, or, upon completion of project, using the EA3 form.
4. Any amendments to the project must be approved by AUTEC prior to being implemented. Amendments can be requested using the EA2 form.
5. Any serious or unexpected adverse events must be reported to AUTEC Secretariat as a matter of priority.
6. Any unforeseen events that might affect continued ethical acceptability of the project should also be reported to the AUTEC Secretariat as a matter of priority.
7. It is your responsibility to ensure that the spelling and grammar of documents being provided to participants or external organisations is of a high standard and that all the dates on the documents are updated.

AUTEC grants ethical approval only. You are responsible for obtaining management approval for access for your research from any institution or organisation at which your research is being conducted and you need to meet all ethical, legal, public health, and locality obligations or requirements for the jurisdictions in which the research is being undertaken.

Please quote the application number and title on all future correspondence related to this project.

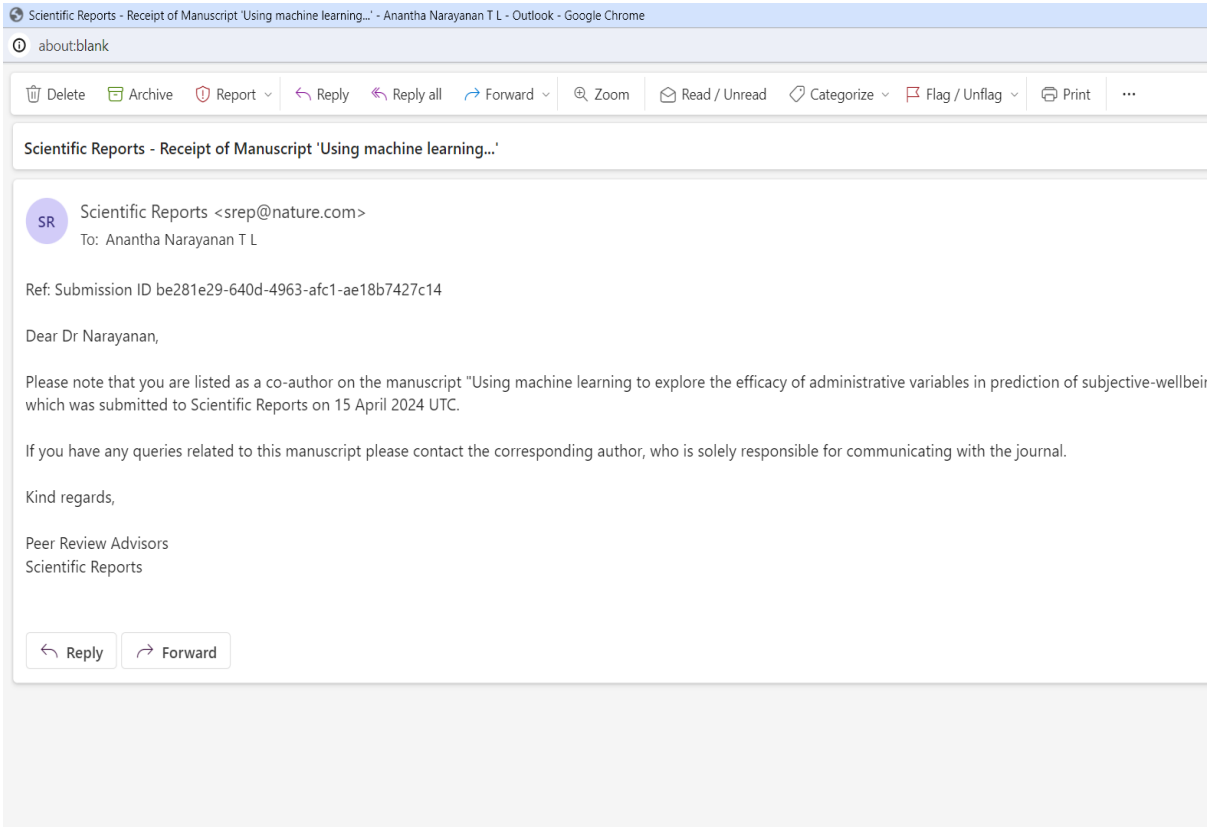
For any enquiries please contact ethics@aut.ac.nz. The forms mentioned above are available online through http://www.aut.ac.nz/research/researchethics

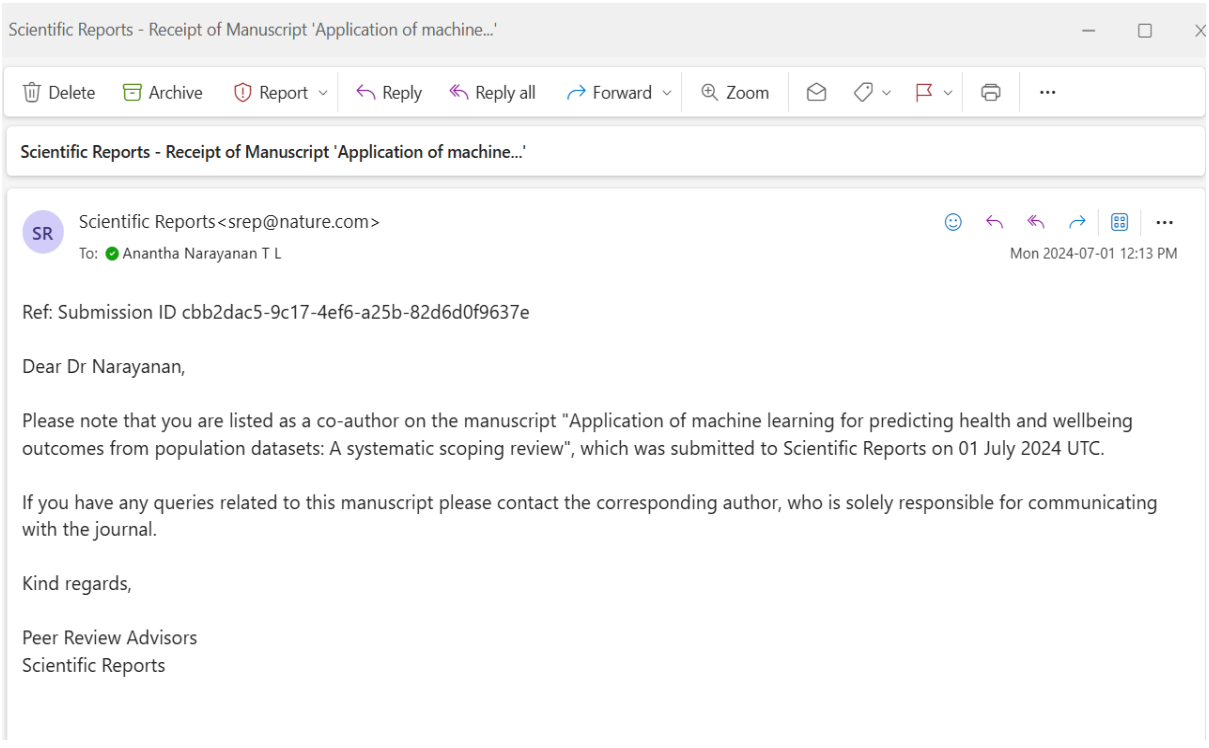
(This is a computer-generated letter for which no signature is required)

The AUTEC Secretariat
**Auckland University of Technology Ethics Committee**

Cc: anantha.narayanan.tl@aut.ac.nz; scott.duncan@aut.ac.nz

# Appendix C. Manuscript submission forms



## Appendix D. R programming code used for analysis.

### Chapter 4 - GSS cross-sectional/longitudinal analysis

#### *Source libraries*

```
library(DBI)

library(odbc)

library(dbplyr)

library(tidyverse)

library(readr)

library(xlsx)

library(writexl)

library(parameters)

library(caret)
```

#### *Check connection to IDI database*

```
dbCanConnect(odbc(), Driver = "ODBC Driver 17 for SQL Server", Server = "PRTPRDSQL3
6.stats.govt.nz,1433", Database = "IDI_Clean_20211020", Trusted_Connection = "yes")
```

#### *Connect to IDI database*

```
con <- dbConnect(odbc(), Driver = "ODBC Driver 17 for SQL Server", Server = "PRTPRD
SQL36.stats.govt.nz,1433", Database = "IDI_Clean_20211020", Trusted_Connection = "y
es")
```

#### *Save SQL tables as RDS files for easy access*

```
sqltoRDS<- function(schema_name, table_name, file_name){

tab <- tbl(con, in_schema(schema_name, table_name)) res_data<- tab %>% collect()

write_rds(res_data, paste0(file_name, ".rds"))}
```

#### *Save GSS files*

```
# Household

sqltoRDS("gss_clean", "gss_household_2008", Datasets/GSS/gss_household_2008")

sqltoRDS("gss_clean", "gss_household_2010", "Datasets/GSS/gss_household_2010")

sqltoRDS("gss_clean", "gss_household_2012", "Datasets/GSS/gss_household_2012")

sqltoRDS("gss_clean", "gss_household", "Datasets/GSS/gss_household_141618")


# Person

sqltoRDS("gss_clean", "gss_person_2008", "Datasets/GSS/gss_person_2008")

sqltoRDS("gss_clean", "gss_person_2010", "Datasets/GSS/gss_person_2010")

sqltoRDS("gss_clean", "gss_person_2012", "Datasets/GSS/gss_person_2012")

sqltoRDS("gss_clean", "gss_person", "Datasets/GSS/gss_person_141618")


# Suuplement
```

```
sqltoRDS("gss_clean", "gss_supp_2014", "Datasets/GSS/gss_supp_2014")
sqltoRDS("gss_clean", "gss_supp_2016", "Datasets/GSS/gss_supp_2016")
sqltoRDS("gss_clean", "gss_supp_2018", "Datasets/GSS/gss_supp_2018")
```

### *Link GSS 2014/16/18 person, household and supplement*

```
gss_p_141618<- read_rds("Datasets/GSS/gss_person_141618.rds")
gss_h_141618<- read_rds("Datasets/GSS/gss_household_141618.rds")
gss_supp_141618<- read_rds("Datasets/GSS/gss_supp_2018.rds")
gss_141618_master<- gss_p_141618%>%
 left_join(., gss_h_141618, by = "snz_uid")%>%
 left_join(., gss_supp_141618, by = "snz_uid")
```

### *Select variables and rename them*

```
gss_141618_ch4 <- gss_141618_master %>%
  select(
    snz_uid, # Unique ID
    gss_pq_collection_code, # Timepoint
    gss_pq_feel_life_code, # Life satisfaction
    gss_pq_life_worthwhile_code, # Life worthwhileness
    gss_pq_dvage_code,  # Age derived variable
    gss_pq_ethnic_grp1_snz_ind,  # European ethnicity
    gss_pq_ethnic_grp2_snz_ind,  # Maori ethnicity
    gss_pq_ethnic_grp3_snz_ind,  # Pacific
    gss_pq_ethnic_grp4_snz_ind,  # Asian
    gss_pq_ethnic_grp5_snz_ind,  # Middle Eastern/Latin American/African
    gss_pq_ethnic_grp6_snz_ind,  # Other
    gss_pq_dvsex_code, # Gender
    gss_pq_Region_Group_code, # Region group
    gss_pq_pers_inc_amt_code, # Personal income
    gss_hq_household_inc1_dev, # Household income
    gss_hq_household_size_dev # Household size
  ) %>%
  mutate_at(vars(-2), as.numeric)


names(gss_141618_ch4) <- c(
  'snz_uid', 'Year', 'Life_satisfaction', 'Life_worthwhile', 'Age_range', 'eth_1',
  'eth_2', 'eth_3', 'eth_4', 'eth_5', 'eth_6', 'Gender', 'Region', 'Personal_income
',
  'Household_income', 'Household_size'
)
```

### *Transform variables into fewer categories*

```
gss_141618_ch4<- gss_141618_ch4%>% mutate(

Age_range = if_else (Age_range <15, 1, if_else(Age_range <25, 2, if_else(Age_range
<35, 3, if_else(Age_range <45, 4, if_else(Age_range <55, 5, if_else(Age_range <65,
6, if_else(Age_range >=65, 7, -1)))))),


Ethnicity = if_else(paste0(eth_1, eth_2, eth_3, eth_4, eth_5, eth_6) == "100000", 1
, if_else(paste0(eth_1, eth_2, eth_3, eth_4, eth_5, eth_6) == "010000", 2, if_else(
paste0(eth_1, eth_2, eth_3, eth_4, eth_5, eth_6) == "001000", 3, if_else(paste0(eth
_1, eth_2, eth_3, eth_4, eth_5, eth_6) == "000100", 4, if_else(paste0(eth_1, eth_2,
eth_3, eth_4, eth_5, eth_6) == "000010", 5, if_else(paste0(eth_1, eth_2, eth_3, eth
_4, eth_5, eth_6) == "000001", 6, if_else(paste0(eth_1, eth_2, eth_3, eth_4, eth_5,
eth_6) == "110000", 7, 8)))))),


Personal_income = if_else(Personal_income <19, 18, if_else(Personal_income >24, 25,
Personal_income)), Household_income = if_else(Household_income <9, 8, if_else(House
hold_income >14, 15, Household_income)),


Household_size = if_else(Household_size > 15 , 16, Household_size), # Compress more
than 5 in a household to one category

)

gss_141618_ch4[is.na(gss_141618_ch4)]<- -1 write_rds(gss_141618_ch4%>% select(-star
ts_with("eth_")), paste0("GSS_141618_cleaned.rds"))
```

### *GSS descriptive analysis*

```
gss_141618<- read_rds("GSS_141618_cleaned.rds")

gss_des<- gss_141618%>%dplyr::select(-snz_uid) %>%

  gather(Outcome_variable, Outcome_value,  c('Life_satisfaction',  'Life_worthwhile
'))%>%

  mutate(Outcome_variable = factor(Outcome_variable, levels = c('Life_satisfaction'
, 'Life_worthwhile'))) %>%

  gather(Demographic_variable, Demographic_category, c('Age_range':'Ethnicity'))%>%

  mutate(Demographic_variable = factor(Demographic_variable,
                                       levels = c('Age_range',
                                                  'Gender',
                                                  'Ethnicity',
                                                  'Region',
                                                  'Personal_income',
                                                  'Household_income',
                                                  'Household_size')))
gss_outcome_list<- gss_des%>%group_split(Outcome_variable, Demographic_variable)

temp<- list()

for(i in 1:length(gss_outcome_list)){
```

```
  # Remove NAs, Refused and 'Don't know' values for each combination of Outcome/ De
mographic variables and summarise mean,median...etc


  temp[[i]]<- gss_outcome_list[[i]]%>%filter((!Outcome_value %in% c(-1, 88, 99, 777
)))%>%
    filter((!Demographic_category %in% c(-1, 88, 99, 777)))%>%
    group_by(Year, Outcome_variable, Demographic_variable, Demographic_category)%>%
    summarise( Number_of_observations = n(),
               outcome_sum = sum(Outcome_value),
               Mean_outcome_value = round(mean(Outcome_value),2),
               Stdev_outcome_valu = round(sd(Outcome_value),2),
               Median_outcome_value = median(Outcome_value),
               Minimum_outcome_value = min(Outcome_value),
               Maximum_outcome_value = max(Outcome_value))}
x<- bind_rows(temp)
write_xlsx(x, "GSS_data_descriptive.xlsx")
```

***Linear model (Cross-sectional), not adjusting for 'year'***

```
  for(i in 1:length(gss_outcome_list)){
    # Remove NAs, Refused and 'Don't know' values for each combination of Outcome/
Demographic variables
    gss_outcome_list[[i]]<- gss_outcome_list[[i]]%>%filter((!Outcome_value  %in% c(
-1, 88, 99, 777))& (!Demographic_category %in% c(-1, 88, 99, 777)))
 }
params <- list()
estimated_means <- list()
model_perf <- list()


for(i in 1: length(gss_outcome_list)){
  x<- gss_outcome_list[[i]]%>% mutate(
  Year = factor(Year),
  Demographic_category = factor(Demographic_category),
  Outcome_value = factor(Outcome_value))


  outcome <-x$Outcome_variable[1] %>% as.character()
  predictor <-x$Demographic_variable[1] %>% as.character()


  x <- x %>%
    dplyr::rename(!!outcome := Outcome_value,
                  !!predictor := Demographic_category
                  )
```

```
  form <- paste(outcome, '~', predictor)
  m <- lm(formula = form, data = x)
  params[[i]] <- parameters(m,exponentiate = TRUE)
  params[[i]]$outcome = outcome
  model_perf[[i]] <- model_performance(m) %>%
   as.data.frame() %>%
   mutate(outcome= outcome,
          predictor = predictor
}
params <- bind_rows(params)
model_perf <- bind_rows(model_perf)
write.xlsx(params, "Model_parameters_unadjusted.xlsx")
write.xlsx(model_perf, "Model_performance_unadjusted.xlsx")
```

***Linear models (longitudinal), adjusted for 'year'***

```
params <- list()
model_perf <- list()
anova_out<- list()
for(i in 1: length(gss_outcome_list)){
  x<- gss_outcome_list[[i]]%>% mutate(
  Year = factor(Year),
  Demographic_category = factor(Demographic_category),
  Outcome_value = factor(Outcome_value))

  outcome <-x$Outcome_variable[1] %>% as.character()
  predictor <-x$Demographic_variable[1] %>% as.character()

  x <- x %>%
    dplyr::rename(!!outcome := Outcome_value,
                  !!predictor := Demographic_category
                  )
   form <- paste(outcome, '~', predictor , '* Year')

  m <- lm(formula = form, data = x)
  params[[i]] <- parameters(m,exponentiate = TRUE)
  params[[i]]$outcome = outcome
  model_perf[[i]] <- model_performance(m) %>%
    as.data.frame() %>%
    mutate(outcome= outcome,
           predictor = predictor)
```

```
anova_out[[i]]<- anova(m) #interaction p-value
}
params <- bind_rows(params)
model_perf <- bind_rows(model_perf)
anova_out<- bind_rows(anova_out)

write.xlsx(params, "Model_parameters_adjusted.xlsx")
write.xlsx(model_perf, "Model_performance_adjusted.xlsx")
write.xlsx(anova_out, "Model_performance_adjusted.xlsx")
```

## Chapter 5 - Modelling subjective wellbeing outcomes.

### *Save Census files*

```
sqltoRDS("cen_clean", "census_individual_2018", "Datasets/Census/census_individual_2018")
sqltoRDS("cen_clean", "census_household_2018", "Datasets/Census/census_household_2018")
sqltoRDS("cen_clean", "census_family_2018", "Datasets/Census/census_family_2018")
sqltoRDS("cen_clean", "census_dwelling_2018", "Datasets/Census/census_dwelling_2018")
```

### *Link Census and GSS 2018*

```
census_2018 <- read_rds("Datasets/Census/census_individual_2018.rds")
census_household_2018 <- read_rds("Datasets/Census/census_household_2018.rds")
census_family_2018<- read_rds("Datasets/Census/census_family_2018.rds")
census_dwelling_2018<- read_rds("Datasets/Census/census_dwelling_2018.rds")

gss_p_2018<- read_rds("Datasets/GSS/gss_person_141618.rds")%>%
  filter(gss_pq_collection_code == "GSS2018")%>%
  select(snz_uid,  # Unique ID
         gss_pq_collection_code,  # Timepoint
         gss_pq_feel_life_code,   # Life satisfaction
         gss_pq_life_worthwhile_code,  # Life worthwhileness
         gss_pq_fam_wellbeing_code, # Family wellbeing
         gss_pq_health_dvwho5_code, # Mental wellbeing
         ) %>%
  mutate_at(vars(-gss_pq_collection_code), as.numeric)

# Merge GSS with census indivudual, household and dwelling
```

```
census_gss <- left_join(gss_p_2018,census_2018, by = "snz_uid")
census_gss <- left_join(census_gss,census_household_2018, by = "snz_cen_hhld_uid")
census_gss <- left_join(census_gss,census_dwelling_2018, by = "snz_cen_dwell_uid")

# Remove NAs from all snz_cen_uid, snz_cen_dwell_uid & snz_cen_hhld_uid

census_gss<- census_gss [-c((which(is.na(census_gss$snz_cen_uid)| is.na(census_gss$snz_cen_dwell_uid)| is.na(census_gss$snz_cen_hhld_uid)))),]
write_rds(census_gss, "census_gss.rds")
```

### *Read data*

```
census_gss<- read_rds("Datasets/census_gss.rds")
HLI<- readxl::read_xlsx("MB2018_exposures_GB.xlsx", sheet = 'MB2018_exposures_GB')%>%
  select(-G_B, -GB_3cat)%>%
  mutate_all(~as.numeric(.))
```

### *Selection of variables*

```
selected_census_gss <- census_gss %>%
                        select(snz_uid,
                               life_satisfaction = gss_pq_feel_life_code,   #Life satisfaction
                               life_worthwhile = gss_pq_life_worthwhile_code,  #Life worthwhileness
                               family_wellbeing = gss_pq_fam_wellbeing_code, #Family wellbeing
                               mental_wellbeing =gss_pq_health_dvwho5_code, #Mental wellbeing
                               age_group = cen_ind_age_code,
                               gender = cen_ind_sex_code,
                               ethnicity = cen_ind_eth_single_comb_grp8,
                               region = cen_hhd_dwell_address_rc,
                               birth_country = cen_ind_birth_country_code,
                               no_of_languages = cen_ind_languages_cnt_code,
                               income_source_count = cen_ind_inc_srce_cnt_code,
                               personal_income = cen_ind_ttl_inc_code,
                               worforce_status = cen_ind_wklfs_code,
                               home_ownership = cen_ind_home_ownsp_code,
                               highest_qualification = cen_ind_standard_hst_qual_code,
                               studying_code= cen_ind_study_prtpcn_code,
```

```
                                smoking_code = cen_ind_smoking_stus_code,
                                disablity_code = cen_ind_dsblty_ind_code,
                                communication_code = cen_ind_dffcl_comt_code,
                                hearing_code = cen_ind_dffcl_hearing_code,
                                remembering_code = cen_ind_dffcl_remembering_code,
                                seeing_code = cen_ind_dffcl_seeing_code,
                                walking_code = cen_ind_dffcl_walking_code,
                                washing_code = cen_ind_dffcl_washing_code,
                                marital_status = cen_ind_legl_mrit_stus_recode,
                                smoke_regularly = cen_ind_smoke_regular_ind_code,
                                smoke_ever = cen_ind_smoke_ever_ind_code,
                                deprivation_code = cen_ind_New ZealandDep2018,
                                deprivation_score = cen_ind_New ZealandDep2018_Score
,
                                household_income = cen_hhd_total_hhld_income_code,
                                crowding_code = cen_hhd_can_crowding_code,
                                dampness_code = cen_dwl_damp_code,
                                mould_code = cen_dwl_mould_code,
                                meshblock_code = cen_hhd_dwell_meshblock_code
                                )
```

***Create freq tables for columns in a df, specifiy start, end and output file***

```
# Example - create_freq_table(selected_census_gss, 2, 39, "Freq_table_variables")
create_freq_table <- function(dataframe, start, finish, output_filename) {
    vector<- colnames(dataframe)[start:finish]
    for (i in 1:length(vector)){
       freq_table<- table(dataframe[vector[i]])

   write.xlsx(freq_table, paste0(output_filename, ".xlsx") ,append = TRUE, sheetNam
e = paste(vector[i]))
     }
  }
```

***Clean data - Remove NAs/invalid categorries, transform variables into categories/fewer categories***

```
clean_selected_census_gss <- selected_census_gss%>%
  mutate(life_satisfaction = ifelse(life_satisfaction %in% c(0,88,99) , NA, life_sa
tisfaction),
         life_worthwhile =  ifelse(life_worthwhile %in% c(0,88,99) , NA, life_worth
while),
         family_wellbeing =  ifelse(family_wellbeing %in% c(0,88,99) , NA, family_w
ellbeing),
```

```
         mental_wellbeing = ifelse(mental_wellbeing %in% c(0,777) , NA, mental_well
being),

         health_status = ifelse(health_status %in% c(88,99), NA, health_status) %>%
factor(.))%>%

  mutate_at(vars(-snz_uid, -life_worthwhile,-family_wellbeing, -mental_wellbeing, -
health_status,-life_satisfaction, -birth_country), ~as.numeric(.))%>%

  mutate(#age_group = ifelse(age_group < 14 , NA, age_group),

         gender = as.factor(gender),

         ethnicity = ifelse(ethnicity == 99  , NA, ethnicity) %>% factor(.),

         region = ifelse(region == 99  , NA, region) %>% factor(.),

         birth_country = case_when(str_starts(birth_country,"1201") ~ "New Zealand"
,
                                   birth_country == '9999' ~ 'Invalid',

                                   TRUE ~ "Other")%>%na_if(.,'Invalid') %>% factor(
.),

         no_of_languages = ifelse(no_of_languages %in% c(0,7,8,9) , NA, no_of_langu
ages),

         income_source_count = ifelse(income_source_count == 99  , NA, income_sourc
e_count),

         personal_income = ifelse(personal_income == 99  , NA, personal_income) %>%
factor(.),

         worforce_status = ifelse(worforce_status == 9  , NA, worforce_status) %>%
factor(.),

         home_ownership = ifelse(home_ownership %in% c(7,9) , NA, home_ownership)%>
% factor(.),

         highest_qualification = ifelse(highest_qualification %in% c(97,99) , NA, h
ighest_qualification)%>% factor(.),

         marital_status = ifelse(marital_status %in% c(777,999) , NA, marital_statu
s)%>% factor(.),

         household_income = ifelse(household_income == 99  , NA, personal_income) %
>% factor(.),

         crowding_code = as.factor(crowding_code))%>%

  mutate_at(vars(home_ownership,studying_code,smoking_code,disablity_code,communica
tion_code, hearing_code,     remembering_code, seeing_code, walking_code, washing_c
ode,

         smoke_regularly, smoke_ever, dampness_code, mould_code ) , ~ifelse(. %in%
c(7,9) , NA, .)%>% factor(.)) %>%

  left_join(., HLI, by = c('meshblock_code' = 'MB2018')) %>% # Remove this line for
modelling without HLI indicators

  na.omit()
```

***Save demographic distribution of the modelling dataset (to check if its comparable with the original GSS dataset)***

```
model_demo_distri <- clean_selected_census_gss %>%

  #select(life_satisfaction: region)%>%
```

```
  mutate(age_group   = if_else (age_group <15, 1,
                          if_else(age_group <25, 2,
                                  if_else(age_group <35, 3,
                                          if_else(age_group <45, 4,
                                                  if_else(age_group <55, 5,
                                                          if_else(age_group <65, 6,
                                                                  if_else(age_group
>=65, 7, -1)))))),
         ethnicity = case_when( ethnicity == 11 ~ 1,
                                ethnicity == 12 ~ 2,
                                ethnicity == 13 ~ 3,
                                ethnicity == 14 ~ 4,
                                ethnicity %in% c(15, 16, 21) ~ 5),

         region =  case_when( region == 2 ~ 11,
                              region == 9 ~ 12,
                              region %in% c(1,4,5) ~ 13,
                              region %in% c(3,6,7,8)  ~ 14,
                              region == 13 ~ 15,
                              TRUE ~ 16))%>%
  gather(variable, choices, life_satisfaction:mould_code, -deprivation_score)%>%
  group_by(variable, choices)%>%
  summarise(count = n(),
            rr3_count = plyr::round_any(count,3))  ## create sepatae column with RR
3 value
write_xlsx(model_demo_distri, "Chapter 5/outputs/Modelling data distribution.xlsx")
```

***Modelling – Random Forest (This will take hours)***

```
outcome_list = c("life_satisfaction", "life_worthwhile" ,"family_wellbeing", "menta
l_wellbeing")


rf_results <- list()
test_result<- list()
RMSE_result<- list()


for (i in 1:4) {


print(paste0("Starting modelling for: ", outcome_list[i]))
t<- proc.time()
```

```
final_data<-clean_selected_census_gss%>%
  select(outcome = outcome_list[i], age_group:ncol(.), -meshblock_code)

set.seed(111)

partition <-createDataPartition(final_data$age_group, p=.70, list = FALSE)

training <- final_data[partition,]
testing<- final_data[-partition,]

weights = data.frame((1/table(training$outcome))*100)%>%mutate(Var1 = as.numeric(Va
r1))

x<- training%>%left_join(.,weights,by = c('outcome' = 'Var1'))

# Set training parameters

fitControl <- trainControl( method  = "repeatedcv",
                            number = 10,
                            repeats = 10,
                            search = 'grid')
# create grid parameters

tunegrid <- expand.grid(.mtry = (2:15))

modelsummary <- list()
modelresults <- list()

## Train model
for (ntree in c((1:10)*200)){

  rfFit <- train(outcome ~ ., data = training,
                 method = "rf",
                 tuneGrid = tunegrid,
                 ntree = ntree,
                 metric = 'RMSE'
                 weights = x$Freq,
                 verbose = FALSE
                   )
```

```
  key <- toString(ntree)
  modelsummary[[key]] <- rfFit
  modelresults[[key]] <- data.frame(ntrees = ntree, rfFit$results)
}
rf_results[[i]] <- bind_rows(modelresults)%>%mutate(outcome = outcome_list[i])
best_model = which.min(rf_results[[i]]$RMSE)
optimal_mtry = rf_results[[i]]$mtry[best_model]
optimal_ntree = rf_results[[i]]$ntree[best_model]

# Train final model
tunegrid <- expand.grid(.mtry = optimal_mtry)
rf_final <- train(outcome ~ ., data = training,
                  method = "rf",
                  tuneGrid = tunegrid,
                  ntree = optimal_ntree,
                  metric = 'RMSE'
                  weights = x$Freq,
                  verbose = FALSE
                  )
saveRDS(rf_final, paste0(outcome_list[i], "_model_with_HLI.rds"))

# Predict on test set
test_result[[i]] <- data.frame (outcome = outcome_list[i],
                            actual = testing$outcome,
                            predicted = predict(rf_final, newdata = testing))

RMSE_result[[i]]<- test_result[[i]]%>%
  group_by(outcome)%>%
  summarise(mtry = optimal_mtry,
            ntree = optimal_ntree,
            rmse = RMSE (predicted, actual),
            mae  = MAE (predicted, actual),
            r2 = R2(predicted, actual))


t_elapsed <- proc.time() -t

print(paste0("Completed modelling for: ", outcome_list[i], ". Time taken = " , t_el
apsed[3], " secs"))
}
```

```
rf_results_final <- bind_rows(rf_results)
final_result<- bind_rows(test_result)
RMSE_final<- bind_rows(RMSE_result)

write_xlsx(rf_results_final, "Model_metrics_no_HLI.xlsx")
write_xlsx(final_result, "model_prediction_results_no_HLI.xlsx")
write_xlsx(RMSE_final, "RMSE_results_no_HLI.xlsx")

# Variable importance for all RF models -----
#Read models

rf_models<- list.files('models', full.names = T)

varimp_overall <- bind_rows(lapply(rf_models,
                      function(x){data.frame(varImp(read_rds(x))$importance) %>%
                          rownames_to_column(., var ="variable")%>%
                          mutate(model_type = substr(x, 8, nchar(x)-4))}))
write_xlsx(varimp_overall, "Variable_importance.xlsx")
```

***Modelling - Elastic Net***

```
test_result <- list()
RMSE_result<- list()

for (i in 1:4) {

  print(paste0("Starting modelling for: ", outcome_list[i]))
  t<- proc.time()

set.seed(111)

final_data<-clean_selected_census_gss%>%
  select(outcome = outcome_list[i], age_group:ncol(.), -meshblock_code)

partition <-createDataPartition(final_data$age_group, p=.70, list = FALSE)

training <- final_data[partition,]
testing<- final_data[-partition,]
weights = data.frame((1/table(training$outcome))*100)%>%mutate(Var1 = as.numeric(Va
r1))
```

```
x<- training%>%left_join(.,weights,by = c('outcome' = 'Var1'))
elastic_net<- train (outcome ~ .,
                     data = training,
                     method = "glmnet",
                     weights= x$Freq,
                     trControl = trainControl (method = "cv", number = 10))

best_model =  which.min(elastic_net$results$RMSE)
optimal_lambda = elastic_net$results$lambda[best_model]
optimal_alpha = elastic_net$results$alpha[best_model]



# Predict on test set
test_result[[i]] <- data.frame (outcome = outcome_list[i],
                                actual = testing$outcome,
                                predicted = predict(elastic_net, newdata = testing)
)
  RMSE_result[[i]]<- test_result[[i]]%>%
    group_by(outcome)%>%
    summarise(alpha = optimal_alpha,
              lambda = optimal_lambda,
              rmse = RMSE (predicted, actual),
              mae  = MAE (predicted, actual),
              r2 = R2(predicted, actual))

t_elapsed <- proc.time() -t

print(paste0("Completed modelling for: ", outcome_list[i], ". Time taken = " , t_el
apsed[3], " secs"))
}
final_result<- bind_rows(test_result)
RMSE_final<- bind_rows(RMSE_result)

write_xlsx(final_result, "model_prediction_results_elasticnet.xlsx")
write_xlsx(RMSE_final, "RMSE_results_elasticnet.xlsx")
```

### *Modelling - Stepwise regression*

```
outcome_list = c("life_satisfaction", "life_worthwhile" ,"family_wellbeing", "menta
l_wellbeing", "health_status")
test_result <- list()
```

```
RMSE_result<- list()
for (i in 1:4) {

  print(paste0("Starting modelling for: ", outcome_list[i]))
  t<- proc.time()
    set.seed(111)
    final_data<-clean_selected_census_gss%>%
    select(outcome = outcome_list[i], age_group:ncol(.), -meshblock_code)

  partition <-createDataPartition(final_data$age_group, p=.70, list = FALSE)



  training <- final_data[partition,]
  testing<- final_data[-partition,]
  weights = data.frame((1/table(training$outcome))*100)%>%mutate(Var1 = as.numeric(
Var1))

  x<- training%>%left_join(.,weights,by = c('outcome' = 'Var1'))

  stepwise<- train (outcome ~ .,
                        data = training,
                        method = "glmStepAIC",
                        weights= x$Freq,
                        trControl = trainControl (method = "cv", number = 10))

  # Predict on test set
  test_result[[i]] <- data.frame (outcome = outcome_list[i],
                                  actual = testing$outcome,
                                  predicted = predict(stepwise, newdata = testing))

  RMSE_result[[i]]<- test_result[[i]]%>%
    group_by(outcome)%>%
    summarise(alpha = optimal_alpha,
              lambda = optimal_lambda,
              rmse = RMSE (predicted, actual),
              mae  = MAE (predicted, actual),
              r2 = R2(predicted, actual))

  t_elapsed <- proc.time() -t
```

```
  print(paste0("Completed modelling for: ", outcome_list[i], ". Time taken = " , t_
elapsed[3], " secs"))


}
final_result<- bind_rows(test_result)
RMSE_final<- bind_rows(RMSE_result)
write_xlsx(final_result, "model_prediction_results_stepwise.xlsx")
write_xlsx(RMSE_final, "RMSE_results_stepwise.xlsx")
```

## Chapter 6 - Predict wellbeing for the population.

### *Create and save population-level data - Don't need to run every time*

```
census_2018 <- read_rds("Datasets/Census/census_individual_2018.rds")
census_household_2018 <- read_rds("Datasets/Census/census_household_2018.rds")
census_family_2018<- read_rds("Datasets/Census/census_family_2018.rds")
census_dwelling_2018<- read_rds("Datasets/Census/census_dwelling_2018.rds")


census_overall <- left_join(census_2018,census_household_2018, by = "snz_cen_hhld_u
id") %>%
  left_join(.,census_dwelling_2018, by = "snz_cen_dwell_uid")


#Remove NAs from all snz_cen_uid, snz_cen_dwell_uid & snz_cen_hhld_uid


census_overall<- census_overall [-c((which(is.na(census_overall$snz_cen_uid)| is.na
(census_overall$snz_cen_dwell_uid)| is.na(census_overall$snz_cen_hhld_uid)))),]


write_rds(census_overall, "census_prediction.rds")
```

### *Transform variables*

```
census_overall<- read_rds("Datasets/Census/census_prediction.rds")
HLI<- readxl::read_xlsx("MB2018_exposures_GB.xlsx", sheet = 'MB2018_exposures_GB')%
>%select(-G_B, -GB_3cat)%>%mutate_all(~as.numeric(.))


census_overall <- census_overall%>%
   select(snz_uid,
          age_group = cen_ind_age_code,
          gender = cen_ind_sex_code,
          ethnicity = cen_ind_eth_single_comb_grp8,
          region = cen_hhd_dwell_address_rc,
          birth_country = cen_ind_birth_country_code,
          no_of_languages = cen_ind_languages_cnt_code,
```

```
          income_source_count = cen_ind_inc_srce_cnt_code,
          personal_income = cen_ind_ttl_inc_code,
          worforce_status = cen_ind_wklfs_code,
          home_ownership = cen_ind_home_ownsp_code,
          highest_qualification = cen_ind_standard_hst_qual_code,
          studying_code= cen_ind_study_prtpcn_code,
          smoking_code = cen_ind_smoking_stus_code,
          disablity_code = cen_ind_dsblty_ind_code,
          communication_code = cen_ind_dffcl_comt_code,
          hearing_code = cen_ind_dffcl_hearing_code,
          remembering_code = cen_ind_dffcl_remembering_code,
          seeing_code = cen_ind_dffcl_seeing_code,
          walking_code = cen_ind_dffcl_walking_code,
          washing_code = cen_ind_dffcl_washing_code,
          marital_status = cen_ind_legl_mrit_stus_recode,
          smoke_regularly = cen_ind_smoke_regular_ind_code,
          smoke_ever = cen_ind_smoke_ever_ind_code,
          deprivation_code = cen_ind_New ZealandDep2018,
          deprivation_score = cen_ind_New ZealandDep2018_Score,
          household_income = cen_hhd_total_hhld_income_code,
          crowding_code = cen_hhd_can_crowding_code,
          dampness_code = cen_dwl_damp_code,
          mould_code = cen_dwl_mould_code,
          meshblock_code = cen_hhd_dwell_meshblock_code)

census_overall <- census_overall%>%
  mutate_at(vars(-snz_uid, -birth_country), ~as.numeric(.))%>%
  mutate(age_group = ifelse(age_group < 14 , NA, age_group),
         gender = as.factor(gender),
         ethnicity = ifelse(ethnicity == 99  , NA, ethnicity) %>% factor(.),
         region = ifelse(region == 99  , NA, region) %>% factor(.),
         birth_country = case_when(str_starts(birth_country,"1201") ~ "New Zealand"
,
                                   birth_country == '9999' ~ 'Invalid',
                                   TRUE ~ "Other")%>%na_if(.,'Invalid') %>% factor(
.),
         no_of_languages = ifelse(no_of_languages %in% c(0,7,8,9) , NA, no_of_langu
ages),
         income_source_count = ifelse(income_source_count == 99  , NA, income_sourc
e_count),
```

```
         personal_income = ifelse(personal_income == 99  , NA, personal_income) %>%
factor(.),

         worforce_status = ifelse(worforce_status == 9  , NA, worforce_status) %>%
factor(.),

         home_ownership = ifelse(home_ownership %in% c(7,9) , NA, home_ownership)%>
% factor(.),

         highest_qualification = ifelse(highest_qualification %in% c(97,99) , NA, h
ighest_qualification)%>% factor(.),

         marital_status = ifelse(marital_status %in% c(777,999) , NA, marital_statu
s)%>% factor(.),

         household_income = ifelse(household_income == 99  , NA, personal_income) %
>% factor(.),

         crowding_code = as.factor(crowding_code))%>%

  mutate_at(vars(home_ownership,studying_code,smoking_code,disablity_code,communica
tion_code, hearing_code, remembering_code, seeing_code, walking_code, washing_code,

                 smoke_regularly, smoke_ever, dampness_code, mould_code ) , ~ifelse
(. %in% c(7,9) , NA, .)%>% factor(.)) %>%

  left_join(., HLI, by = c('meshblock_code' = 'MB2018')) %>%

  na.omit()


write_rds(census_overall, "Datasets/final_prediction_data.rds")
```

***Predict wellbeing for the census population***

```
final_prediction_data<- read_rds("Datasets/final_prediction_data.rds")%>%

  select(-snz_uid, -meshblock_code)

outcome_list = c("life_satisfaction", "life_worthwhile", "family_wellbeing", "menta
l_wellbeing")

final_model<- lapply(outcome_list, function(x)
{read_rds(paste0("models/",x,"_model.rds"))}l) # Use HLI model for prediction


# Takes a while to run

final_prediction_result <- data.frame (

  #snz_uid = final_prediction_data$snz_uid,

                             age = final_prediction_data$age_group,

                             gender = final_prediction_data$gender,

                             ethnicity = final_prediction_data$ethnicity,

                             region = final_prediction_data$region,

                             life_satisfaction_pred = predict(final_model[[1]], newd
ata = final_prediction_data),

                             life_worthwhile_pred = predict(final_model[[2]], newdat
a = final_prediction_data),

                             family_wellbeing_pred = predict(final_model[[3]], newda
ta = final_prediction_data),

                             mental_wellbeing_pred = predict(final_model[[4]], newda
ta = final_prediction_data)

  )
```

```
write_rds(final_prediction_result, "results/final_prediction_result.rds")
```

### *Compare prediction result with GSS*

```
final_prediction_result<- read_rds("results/final_prediction_result.rds")%>%
  mutate_all(~as.numeric(as.character(.)))%>%
  mutate(age   = if_else (age <15, 1,
                                  if_else(age <25, 2,
                                                  if_else(age <35, 3,
                                                          if_else(age <45, 4,
                                                                  if_else(age <55, 5,
                                                                   if_else(age <65, 6,

if_else(age >=65, 7, -1)))))),
         ethnicity = case_when( ethnicity == 11 ~ 1,
                                ethnicity == 12 ~ 2,
                                ethnicity == 13 ~ 3,
                                ethnicity == 14 ~ 4,
                                ethnicity %in% c(15, 16, 21) ~ 5),

         region =  case_when( region == 2 ~ 11,
                              region == 9 ~ 12,
                              region %in% c(1,4,5) ~ 13,
                              region %in% c(3,6,7,8)  ~ 14,
                              region == 13 ~ 15,
                              TRUE ~ 16))%>%
  rename(age_group = age)%>%
  gather(variable, value, 1:4)%>%
  mutate(group_category = paste0(variable, "_", value))%>%
  select(-variable, -value)
gss_2018<- read_rds("GSS_141618_cleaned.rds")%>%
  filter(Year == "GSS2018")%>%
  select(age_group = Age_range,
         gender = Gender,
         ethnicity = Ethnicity,
         region = Region,
         life_satisfaction = Life_satisfaction,
         life_worthwhile = Life_worthwhile,
         family_wellbeing = Family_wellbeing,
```

```
         mental_wellbeing = Mental_wellbeing

        )%>%

  mutate(life_satisfaction = ifelse(life_satisfaction %in% c(0,88,99) , NA, life_sa
tisfaction),


         life_worthwhile =  ifelse(life_worthwhile %in% c(0,88,99) , NA, life_worth
while),

         family_wellbeing =  ifelse(family_wellbeing %in% c(-1,0,88,99) , NA, famil
y_wellbeing),

         mental_wellbeing = ifelse(mental_wellbeing %in% c(0,777) , NA, mental_well
being))%>%

  na.omit()%>%

  gather(variable, value, 1:4)%>%

  mutate(group_category = paste0(variable, "_", value))%>%

  select(-variable, -value)


outcome_list <- c("life_satisfaction", "life_worthwhile", "family_wellbeing", "ment
al_wellbeing" , "health_status")


group_list<- unique(final_prediction_result$group_category)

comparison_output <- list()

for (i in 1:length(outcome_list)){

for (j in 1: length(group_list)){


  predicted_data<- final_prediction_result %>% filter(group_category == group_list[
j]) %>%

    rename(outcome = paste0(outcome_list[i],"_pred"))


  gss_data<- gss_2018 %>% filter(group_category == group_list[j]) %>%

    rename(outcome = outcome_list[i])


comparison_output[[paste0(i,"_",j)]]<-  data.frame(outcome = outcome_list[i],

            group_category = group_list[j],

            N_predicted = plyr::round_any(nrow(predicted_data),3),

            N_gss= plyr::round_any(nrow(gss_data),3),

            mean_predicted = mean(predicted_data$outcome),

            sd_predicted = sd(predicted_data$outcome),

            iqr_25_predicted = quantile(predicted_data$outcome, probs = 0.25),

            iqr_50_predicted = quantile(predicted_data$outcome, probs = 0.50),

            iqr_75_predicted = quantile(predicted_data$outcome, probs = 0.75),

            mean_gss =  mean(gss_data$outcome),

            sd_gss = sd(gss_data$outcome),
```

```
            iqr_25_gss = quantile(gss_data$outcome, probs = 0.25),

            iqr_50_gss = quantile(gss_data$outcome, probs = 0.50),

            iqr_75_gss = quantile(gss_data$outcome, probs = 0.75),


            p_value = wilcox.test(predicted_data$outcome, gss_data$outcome, var.equ
al = FALSE)$p.value


            )}}

output<- bind_rows(comparison_output) %>% mutate(significane = if_else(p_value< 0.0
5, "p < 0.05" , "p > 0.05"))


write_xlsx(output, "Prediction_summary_RR3.xlsx")
```