# A Predictive Model for Identifying Health Trends Among Māori and Pacific People – An Analysis from 10-years of New Zealand Public Hospital Discharges

## Abstract

Introduction and Methods: This research focused on the quality of healthcare services for Māori and Pacific Islanders. New Zealand (NZ) public hospital discharges data from 2005 to 2015 has been used, the process of extracting meaningful data was challenging as well as crucial task. The extracted data was imported to the Power BI analysis platform to create corresponding categories and reports. Based on the extracted data, a prediction model has been developed to predict the trends for patients with a specific chronic disease, external injuries and operative procedures based on the previous/historic data.

Analysis: Initial exploration suggests that the service demand increased from 138,656 in 2005 to 163,386 in 2015. People with external injuries increased from 32,726 in 2005 to 48,326 in 2015. The number of operative procedures reached the peak point in 2012 (298,231), then declined to 244,799 in 2015.

Results and Conclusion: The results show the 'disease' rate of Māori and Pacific Islanders is about 17.84% (n=138,656). 'Factors influencing health status and contact with health services' are the leading cause of health services utilization. We successfully analyzed the diseases with highest incidence rate and key characteristics of this group of patients. This research concluded with a series of key findings on the disease types including injuries, procedures, and services.

## Keywords

Predictive model, hospital discharges, machine learning model, data analysis, machine learning, predictive analysis, healthcare delivery, disease prediction, operative procedures, Māori Population and Pacific Islanders.

# 1. Introduction and Background

New Zealand, in common with most developed countries, is expected to experience significant population ageing in the coming decades [1]. The absolute number of people suffering from disease or injuries is increasing rapidly, which results in increased healthcare cost, service demands and management.

The public hospitals in New Zealand face enormous pressure to reduce costs, manage high-service demands and provide quality services for all New Zealanders. In New Zealand, the adverse events are estimated to cost the medical system approximately 870 million NZD, of which 590 million NZD are spent towards treating preventable adverse events. The results suggest that up to 30% of public hospital expenditure goes toward treating an adverse event [2-5]

In this research, 54,360 records were analyzed from NZ public hospital discharge database to present a variety of trends. From the literature review, it is evident that there is a substantial research currently going on in the prediction of re-admissions, hospital length of stay and the cost of a particular disease. This research focused on the analysis and prediction of disease types, external causes, and operative procedures with respect to ethnicity, age and gender, especially with the Māori and Pacific Islanders demographic. The aim is to investigate the common disease found among Māori and Pacific Islanders and its related age group.

# 2. Methods

## 2.1. Design, Material and Data Resource

Firstly, the analysis aims to present interesting trends and patterns based on the age, sex and ethnicity, which is beneficial to decision-making for public hospitals in New Zealand. Despite some limitations, the comorbidity score derived from hospital discharge data provides an important enhancement to population-based disease research [4]. Secondly, build a data mining model to predict the numbers of patient suffer from particular disease according to the historical data, which could help the hospital management for future planning.

The discharge dataset consisted of 10 excel files from 2005 to 2015, which consists of three categories. The first category presents the 'internal' (illness/diseases) causes led to being in a hospital, it contains twenty disease types. The second category is related to external causes such as transportation accidents, falls, mechanical forces, etc. The last category is regarding the operative procedures including the nervous system, eye and adnexa and endocrine system, etc.

The data for each year contains approximately ten excel sheets. In order to make the data ready to be analyzed, it is necessary to convert data into a desirable format. Finally, all of the extracted data were processed and stored in three tables of database. The first table is made of nine columns (ID,Yyear, Diseasetype, Ethnicity, Agegroup, Gender, Meanstay, Daycases, Numbers); the second table consists of nine columns (ID, Yyear, Injurytype, Ethnicity, Agegroup, Gender, Meanstay, Daycases, Numbers) and the third table is formed from nine columns (ID, Yyear, Procedurestype, Ethnicity, Agegroup, Gender, Meanstay, Daycases, Numbers). To visualize this data, Power BI tool was used to develop relevant reports, which is a popular data analysis platform. The data stored in a database could be linked directly to Power BI desktop via the system interface.

### 2.2. Data Description

The total dataset was of 1,566,501 patients in the research data samples. The sample percentage of Māori is found to be 70%, much higher than that of Pacific Islanders and the proportion of male and female is about 6:4. In terms of age group, the rate of fall in diseases for '0-5' years and '20-25'years is 23% and 11% respectively, which is much higher than other age groups. According to the 2013 census data, there were 4.8% of people aged 15-29 years, 6.1% of Māori and 4.8% for Pacific people respectively [3].

### 2.3. Data Preprocessing and Preparation

Data pre-processing is one of the essential steps to prepare data so that data mining technique applied to it produces high-quality and accurate output patterns [9]. The raw data could not be analyzed by Power BI desktop directly, so it was necessary to import it to SQL server 2008. Total of 60 excel sheets were imported

to a database including disease types, external causes and operative procedures with different genders, ethnicities and periods.

Stored procedures were programmed to process raw data, such as adding columns, splitting columns, updating values, creating views, and reducing data range, this was done to extract the desirable data. We faced challenges in pre-processing due to the determination of fine distinctions among 60 excel sheets, it required adjustments to the parameters of stored procedures and also involved filtering redundant. Finally, three tables accommodated the desirable data. There are total nine columns or features, 'ID' is auto-increment column from 1, 'Yyear' stands for the year of disease happened, 'Ethnicity' including Māori and Pacific Islanders, age group '0-' means age range from 0 to 5. 'meanstay' means average long of stay in hospital, 'Daycases' stands for the number of cases happened in daytime. In table 2, 'injurytype' stands for injury types and 'Procedurestype' means operative procedures type over the ten years from 2005 to 2015.

Table 1: Disease Type for Māori

| ID | Yyear | Diseasetype | Ethnicity | Agegroup | Gender | Meanstay | Daycases | Numbers |
|---|---|---|---|---|---|---|---|---|
| 1 | 2005/1/1 0:00 | G00-G99 | Māori | 0- | M | 7 | 403 | 110 |
| 2 | 2005/1/1 0:00 | H00-H59 | Māori | 50- | M | 2.4 | 469 | 29 |
| 3 | 2005/1/1 0:00 | D50-D89 | Māori | 35- | F | 3.7 | 306 | 38 |
| 4 | 2005/1/1 0:00 | A00-B99 | Māori | 15- | F | 3.5 | 457 | 117 |

Table 2: Injury Types for Māori

| ID | Yyear | Injurytype | Ethnicity | Agegroup | Gender | Meanstay | Daycases | Numbers |
|---|---|---|---|---|---|---|---|---|
| 1369 | 2005/1/1 0:00 | V01–V09 | Māori | 0- | F | 8.5 | 15 | 8 |
| 1370 | 2005/1/1 0:00 | V01–V09 | Māori | 10- | F | 8.5 | 15 | 10 |
| 1371 | 2005/1/1 0:00 | V01–V09 | Māori | 15- | F | 8.5 | 15 | 13 |
| 1372 | 2005/1/1 0:00 | V01–V09 | Māori | 20- | F | 8.5 | 15 | 3 |

Table 3: Procedure Types for Māaori

| ID | Yyear | Procedurestype | Ethnicity | Agegroup | Gender | Meanstay | Daycases | Numbers |
|---|---|---|---|---|---|---|---|---|
| 1 | 2005/1/1 0:00 | 1-86 | Māori | 0- | F | 7.6 | 411 | 192 |
| 2 | 2005/1/1 0:00 | 1-86 | Māori | 10- | F | 7.6 | 411 | 55 |
| 3 | 2005/1/1 0:00 | 1-86 | Māori | 15- | F | 7.6 | 411 | 76 |
| 4 | 2005/1/1 0:00 | 1-86 | Māori | 20- | F | 7.6 | 411 | 72 |

Power BI is a cloud-based business analytics service that gives a single view of data. Power BI can read data from SQL server database directly, any data changes would update the reports of Power BI in real time.

## 3. Prediction Model

WEKA was used to build the predictive model, the WEKA program is written in Java, is available freely on the web and comprises a variety of data-mining algorithms [10]. There are five algorithms could be applied to predict numbers such as linear regression, M5P regression and model trees, K nearest neighbor (KNN) and baseline predictor. KNN method is a popular classification method in data mining and statistics because of its simple implementation and significant classification performance [11]. KNN was used to predict classification problems and also used to have a good performance in numeric prediction. In this research, we used the KNN model with adjusted parameters to get reasonable results.

Bagging is a supervised learning approach that allows several models to have an equal vote in classification, which helps if the weak learning algorithm is unstable due to small changes in the input data. Hence, bagging can hardly improve the k-NN prediction on account of its stability [12]. Bagging has the ability to improve the performance of prediction models, especially for avoiding over-fitting [13]. In this research, bagging-based ensemble KNN was used to improve the ability to numeric prediction.

### 3.1. Training and Testing Dataset

Machine learning methods can achieve satisfactory performance with sufficiently well-labeled training dataset [14]. In total there are 14,400 records in the first table 'Table_disease', which consists of two parts, training dataset, and testing dataset. We selected 70% of data for training and 30% for testing. Models were developed on the training dataset and evaluated on the testing datasets [15]. Testing dataset is used to assess the performance of model according to the running results such as accuracy and mean absolute error.

### 3.2. Statistical Analysis

Figure 1 shows there are 2,446,318 people who suffer from a disease from 2005 to 2015, the children range from 0 to 5 years old account for most of the total population, with 14% and 7% for Māori and Pacific Islanders respectively. The age ranges for Māori was from 85 to 90 years with n=17,204, the percentage of Māori people who suffer from disease is a little bit more than that of Pacific Islanders. On the whole, the numbers of Māori people suffer from disease is about 2.2 times higher than that of Pacific Islanders, based on the percentage of 68.95% and 31.05% for Māori and Pacific Islanders respectively [16].

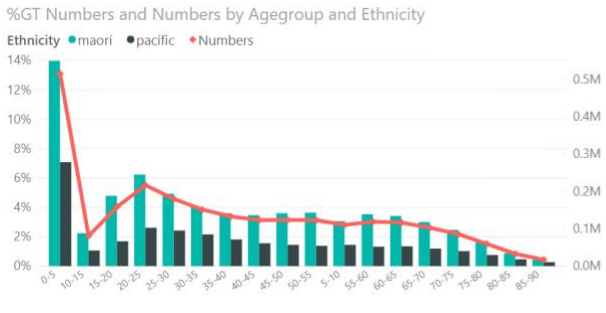


Figure 1 Rates of different age-group suffer from disease

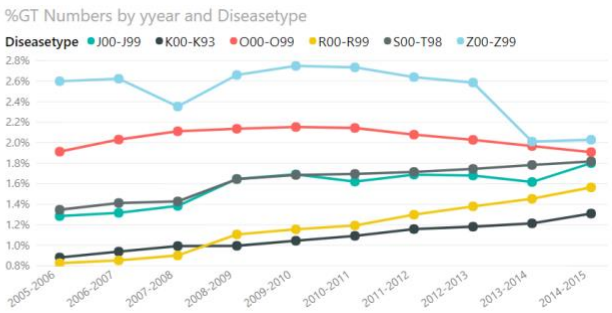


Figure 2 Top six diseases group by years

Figure 2 presents the rates of top six diseases between Māori and Pacific Islanders over the years. The disease types 'Z00-Z99' (Factors influencing health status and contact with health services) is responsible for the nearly 2.6% (40,732) of the total disease types before 2013. A stable increase was found for 'S00-T98' (Injury, poisoning and certain other consequences of external causes) and 'K00-K93' (Diseases of the digestive system). The percentage of 'R00-R99' (Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified) showed most rapid increment over ten years. The trend of disease type

‘O00-O99’ (Pregnancy, childbirth, and the puerperium) did not fluctuate, the proportion was nearly 2% throughout the analysis period.

Figure 3 shows the trend for the total number of people (Māori and Pacific Islanders) suffering from disease year by year. In the 2005, the number was about 210,113 and in 2015 it was 265,798 emphasizing a significant increase. We found the maximum increase among Māori who suffer from a disease with an increase rate is 9.39% was in 2008 (153,523) to 2009 (167,941). In contrast, the increasing trend could be observed in the year of 2007 and 2013. However, the number of Pacific Islanders decreases, which only happened in these two years. In the 2012, there were increasing number for both Māori and Pacific Islanders people, the medical treatment data could be analyzed to decrease the number of people with the particular disease in the future. Thailand has taken on the leadership role and has been able to dominate the normative processes of sub-regional disease control and in doing so has strengthened its own economic and national security [17].

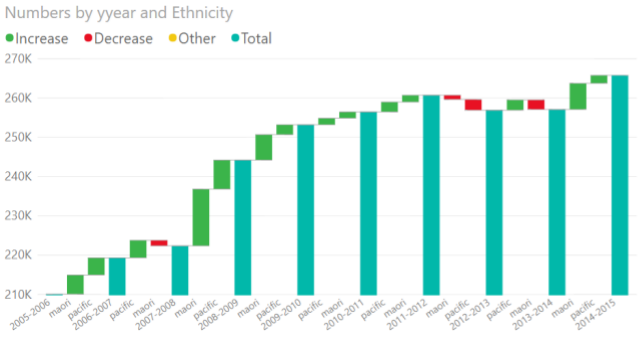


Figure 3 Numbers of two ethnicities suffer from disease order by years

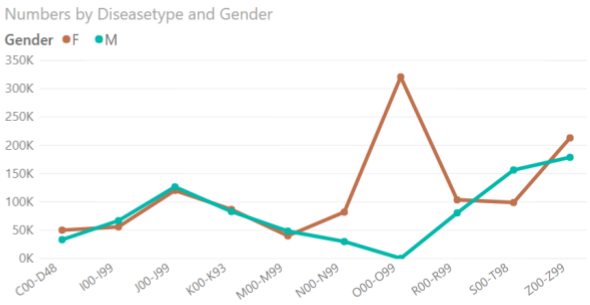


Figure 4 Numbers of two genders suffer from top ten disease

Figure 4 illustrates the number of people with disease from different genders for top ten diseases in the year from 2005 to 2015. Among the top five disease types, the number of males is similar to that of female, however, in other five disease types excluding ‘S00-T98’, the number of female patients was found to be higher than that of male. In the type ‘S00-T98’ (Injury, poisoning and certain other consequences of external causes), the number of the male is around 1.5 times of female over the years from 2005 to 2015.

The bar chart in figure 5 gives information about the proportion of top eight injuries types among Māori and Pacific Islanders people from 2005 to 2015. Figure 5 shows upward trends in the total injury types; the total number went up from 32,726 in 2005 to 48326 in 2015. The red line showed there was a sharp increase from 2008 to 2009. The injuries type 'W20-W49' (Exposure to inanimate mechanical forces) grew slightly each year, in the year of 2015, the amount of 'W20-W49' reached the peak point (14,643). The percentage of 'W00-W19' (Falls) and 'Y83-Y84' (Surgical and other medical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure) also increased gently, but they were still less than that of 'W20-W49'. There were almost no changes in the injury types of 'W50-W64' (Exposure to animate mechanical forces) and 'V40-V49' (Car occupant injured in transport accident), which stayed the number of approximately 1,600 over the ten years. Most injuries types present upward trends except above two injuries types.

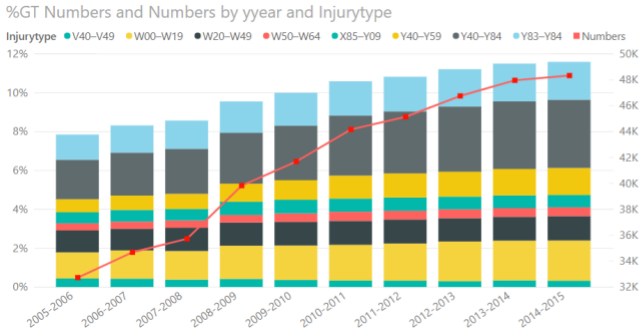


Figure 5 Rates of injuries types group by years

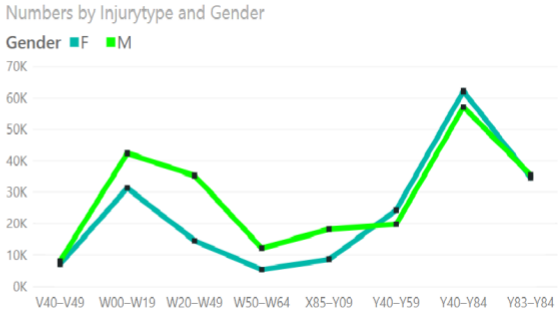


Figure 6 Numbers of two genders group by injuries types

Figure 6 illustrates the number of top eight injury types from both males and females over the 10 years from 2005 to 2015. Overall, there were different tendencies due to different genders and injury types, in the first five injury types, the number of males is more than that of females. However, in the last three injury types, number of females was higher when compared with the male patients. In the injuries type of 'W20-W49', the biggest gap between males and females could be observed, the number was 35,335 and 14,677 respectively. In 'V40-V49' and 'Y83-Y84', we found similar number of males and females, the numbers were about 7,500 and 35,000 respectively. The largest number of people who suffer from injuries is from 'Y40-Y84' (Complications of medical and surgical care), regardless of their gender.

Figure 7 shows the proportion of top ten procedure types for males and females, we found that the procedure type ‘1820-1922’ (Noninvasive, Cognitive and Other Interventions, not elsewhere classified) was the highest, which account for about 16% and 14% for females and males respectively. The percentage of ‘1040-1129’ (Procedures on Urinary System) was relatively low, which was about 2% males and females. Female-only procedure types were ‘1330-1347’ (Obstetric Procedures) and ‘1240-1299’ (Gynecological Procedures). However, for most of procedures types, the proportion of males and females were equal over the ten years. Overall, except ‘Obstetric Procedures’ and ‘Gynecological Procedures’, the number of other operative procedures were not influenced by gender from 2005 to 2015.

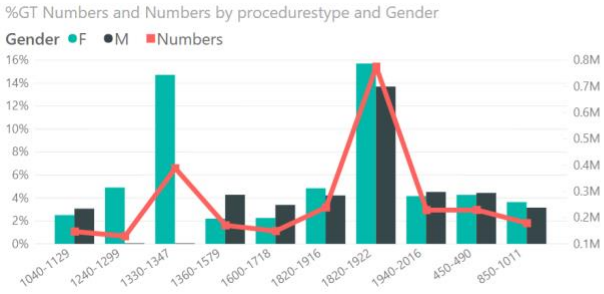


Figure 7 Rates of gender and procedures types

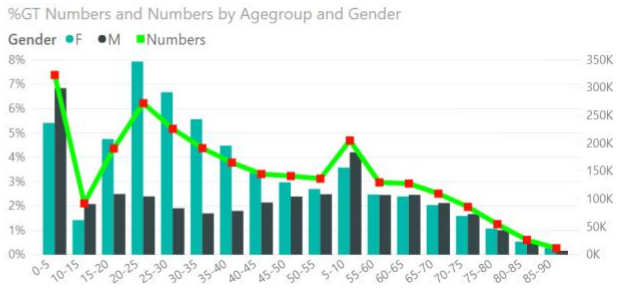


Figure 8 Rates of gender and ages on procedures types

Figure 8 shows the percentage of males and females on procedures from different age groups over the ten years from 2005 to 2015, the age group with the largest proportion is 0-5 years old, with over 5% and 7% for female and male respectively. Whereas the group 85y – 90y had the smallest proportion, the number is about 0.44% (11,560). Apart from age group 0-5, the group ’20-25’ has the largest proportion with 7.93% and 2.39 for female and male respectively. We found that there was a huge gap between male and female for ages ranging from 0 to 40, the total number of female procedures were much more than that of the males. However, the number of females tend to be similar to the number of the males when the age exceeded 40 years old. In the age group ’20-25’, the gap between different genders is largest, which is similar to age group ’30-35’ and ’35-40’.

# 4. Building the Model

## 4.1. Training the Model

The quantity of the dataset was large enough for training and testing the proposed model [18]. For training the model, balanced samples were used for each period [19]. The training dataset consists of 9,838 records from Māori and Pacific Islanders, the age ranges from 0 to 85y, which contains 18 age groups in total. The performance of prediction model mainly depends on the number and the quality of training dataset. The training dataset accounted for about two-thirds of all data on disease types. In terms of algorithms, WEKA was used to build a model and predict the numbers of patients experience from a particular disease. WEKA provided four models which could be used in numeric prediction such as linear regression, M5P regression, kNN and baseline predictor. The first three models were selected and compared for choosing the best model, the mean absolute error, correlation coefficient, root mean squared error and root relative squared error were the four-key metrics in judging the model's overall performance.

Figure 9 shows the results from three different algorithms, the outcomes of training models were output in terms of the mean absolute error, correlation coefficient, root mean squared error and root relative squared error, taking 'mean absolute error' as an example, the results are 18.96, 148.13 and 50.65 respectively. Similarly, other three results can be seen in table 4.

Table 4: Outputs from the three different models – kNN, Linear Regression and M5P Regression

| Algorithm | Dataset | FS | Number of Features | Correlation Coefficient | Mean Absolute Error | Root Mean Squared Error | Root Relative Squared Error |
|---|---|---|---|---|---|---|---|
| kNN | Training | no | 7 | 1.00 | 18.96 | 44.41 | 10.04 |
| Linear Regression | Training | no | 7 | 0.53 | 148.13 | 384.57 | 85.45 |
| M5P Regression | Training | no | 7 | 0.91 | 50.65 | 183.62 | 40.67 |

Based on these four results, kNN was selected as a good performance than other modes. kNN is a non-parametric prediction algorithm. It searches the K most similar feature vectors within the historical database to predict future values [20]. However, the kNN performance showed that our models required further refinement.

### 4.2. Model Evaluation

Data mining consists of a cycle of generating, testing and evaluating various algorithms/models [21]. The evaluating model plays a significant role in the process of numeric prediction, which aims to identify the best model. There are four metrics used to assess the performance of prediction model. Correlation coefficient method was used to evaluate the correlation between two variables in the field of statistics [22]. The value ranges from 0 to 1 where, 1 indicates there is a strong positive relationship between variables.

Mean absolute error was used to measure the difference between the predicted value of the classifier and the actual result. Root mean squared error is to describe the dispersion degree of the sampling distribution of the corresponding sample statistics and the measure of the sampling error size of the corresponding sample statistics.

Root relative squared error sometimes does not reflect the true size of the error, while the proportion of the true value error has a good reflection on true value error. Based on the above running results, the correlation coefficient of (Instance Based Learner) IBK is 1 with a low mean absolute error about 18, so IBK was selected as the best model to predict the number of patients require health services in future. The testing dataset was used to test the model of IBK, but the results were not accurate enough as expected, so a few measures was taken to improve the performance of prediction models, such as adjusting parameters or bagging.

### 4.3. Feature Selection

Feature selection was found to be effective to remove the attributes which are not relevant to the prediction results. Once the superfluous attributes were removed from training dataset, the performance of model was improved [23]. In this research, we used wrapper method to select the key features. Wrapper methods depend on a specific learning algorithm in evaluating the selected subset of features, comparing to other families, wrappers are more accurate since they consider the relations between the features themselves [24]. The wrapper selects a subset of features by assessing the performance of learning method. After the feature

selection is implemented, the columns ‘Meanstay’ was removed. Moreover, in order to avoid the influence on the performance of model, the column ‘ID’ was also removed from training dataset.

### 4.4. Refining the Model with Bagging and Adjusting Parameters

Bagging is one of the earliest ensemble methods using the bootstrap sampling technique. The bootstrap technique samples randomly with replacement to generate multiple samples forming a training set [25]. We used bagging with KNN to generate a new model, then compared the results between a new model and other models. Each of the generated subsets is used to construct the decision tree and they are later aggregated into the final model. The output results are in table 5 (K=1, training dataset) with improvement in the mean absolute error, root mean squared error, and root relative squared error except for the correlation coefficient.

A poorly chosen nearest neighborhood parameter leads to an underlying probability density estimate that does not represent the data well [26-27]. A small or large ‘K’ will have a negative influence on the prediction results, so it is important to choose a proper k value for KNN algorithm, which stands the number of neighbors to use. The test results (as shown in table 5 (K=1, testing dataset) were, mean absolute error was about 27, root relative squared error was about 68, a correlation coefficient was 0.9899.

Table 5: Results with feature selection

| Classifier | Dataset | FS | K | Number of Features | Correlation Coefficient | Mean Absolute Error | Root Mean Squared Error |
|---|---|---|---|---|---|---|---|
| Bagging (IBK) | Training | wrapper | 1 | 6 | 0.9955 | 17.7073 | 43.3038 |
| Bagging (IBK) | Training | wrapper | 2 | 6 | 0.9938 | 20.5910 | 50.9241 |
| Bagging (IBK) | Testing | wrapper | 1 | 6 | 0.9899 | 27.1599 | 68.0228 |

# 5. Results

## 5.1. Outcomes

The analysis shows an increase in the number of patients experience multiple diseases such as ‘J00-J99’ (Diseases of the respiratory system), ‘k00-k93’ (Diseases of the digestive system), ‘R00-R99’ (Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified), which ranges from 138,656 in 2005 to 163,386 in 2015, the increase was about 17.84%. Such growth may produce a huge medical burden to an individual and society. Māori people had a larger population of diseases than that of Pacific Islanders, the total female patients were about 1,169,485, which were more than that of the male patients. Moreover, we found that there is not any particular disease which only relates to Māori or Pacific Islanders.

The analysis also contains external injuries information, the number of people suffer from external injuries dramatically rises to 48,326 in 2015 from 32,726 in 2005, the increase rate is about 47.69% although there is no significant increase in terms of the number of injuries types. The injury type ‘Y40-Y84’ (Complications of medical and surgical care) ranks first in all injury types over the ten years. In contrast, the injury type ‘W50-W64’ (Exposure to animate mechanical forces) was found to be stable, and there were no major outbreaks in those years.

In terms of procedures types, the number of people who does the surgery reaches the peak point in the year of 2013, and then declined to 244,799 in the year of 2015, this highlight the variation in number of procedures and a health and economic burden related to diseases and surgeries in New Zealand over the ten years.

# 6. Discussion

This research comprises of data analysis and data prediction techniques and methodologies. Data analysis involves the distribution of disease types, injuries types, and procedure types over the ten years from 2005 to 2015. Utilizing NZ public hospital dataset was challenging to ensures the data accuracy, the collected data consists of different dimensions such as age group, ethnicity, and gender, which provides useful

information for medical research. Power BI was applied to create corresponding reports with plenty of dimensions and the building of model makes it feasible to predict numbers for the next year.

Only Māori and Pacific Islanders people were studied in this research, patients from other ethnicities were not considered, which may fail to make the analysis results very representative in showing the trends on other ethnicities. In terms of prediction, there were five features in the extracted dataset, which could be used to form a training dataset for training model. Prediction model is still needed to improve and acquire a model with better performance, such as parameters optimization, trying other algorithms and enlarge the training dataset.

In line with other similar research, we presented the trends of diseases in age groups range from new born to 85 years, ethnicity including Māori and Pacific Islanders for both male and female. Juyoung et al. [28] conducted a research to determine the prevalence trends of osteoarthritis (OA), rheumatoid arthritis (RA), and other types of arthritis in the United States from 1999 to 2014, which focus on the relationship between age group and OA and RA. A lot of data was analyzed to estimate the total costs of treating head and neck cancers, specifically oropharyngeal, laryngeal and oral cavity cancer, in secondary care facilities in England during the period 2006/2007 to 2010/2011 [29]. Spinal cord injuries nursing data elements are explored as a mechanism used in the prediction of patient length of stay by building prediction model with artificial neural networks [30].

## 7. Conclusion

We successfully analyzed the diseases with the highest incidence, and the characteristics of the population with the highest incidence, such as age, gender, ethnicity, and so on. At the same time, external causes such as ‘Y40-Y84’,’W50-W64’ that lead to hospitalization were also analyzed. Finally, a series of analyses were conducted on the types of procedures in the 10 years (until 2015).

Over the ten years from 2005 to 2015, the number of diseases for Māori and Pacific Islanders increased by 17.84%. ‘Factors influencing health status and contact with health services’ was the leading cause of health

loss, accounting for 25% of total diseases. The number of people with most diseases was on the rise, while the number of minority diseases was on the decline. The disease is concentrated in people aged 0-5 years, and the Māori population was larger than the Pacific Islanders.

Hospitalization caused by various external causes had increased year by year, which increased the burden of public hospitals in New Zealand. In the year of 2015, the corresponding number was about 48,326, which show the trend of rising continually in the future. 'Injury, poisoning and certain other consequences of external causes' had raised to become the third-ranked cause of health loss in females and males. Providing better care for people living with their physical health – is a growing challenge for the health and social sectors.

The procedure types also grew sharply, with the tipping pint reached in 2012 and then declined in 2015, which shows a good trend for the public hospital, although the population was increasing in those years. Summing up the experience leading to reducing operative procedures has a positive influence on allocating the resource of the public hospital.

Several algorithms were selected to build a model, however, IBK algorithm was considered as the best one. In order to reduce prediction error, 'bagging' was combined with IBK algorithm to improve performance. The output of the prediction model might be helpful for the hospital management, policy and decision makers and also towards the future resource planning.

## Compliance with Ethical Standards

Authors declare no conflict of interest.

Ethical approval: Publicly available dataset (de-identified) is used in this study, under the guidelines for research purposes only.

## References


[1] O. A. Aziz, C. Ball, J. Creedy, and J. Eedrah, "The distributional impact of population ageing in New Zealand," New Zealand Economic Papers, vol. 49, no. 3, pp. 207-226, 2014.

[2] P. M. Brown, Colin, "Cost of medical injury in New Zealand: a retrospective cohort study.," Journal of Health Services Research & Policy, Article vol. 7, pp. p29-34, 2002.

[3] J. Zhao, S. Gibb, R. Jackson, S. Mehta, and D. J. Exeter, "Constructing whole of population cohorts for health and social research using the New Zealand Integrated Data Infrastructure," Aust N Z J Public Health, 2018.

[4] D. Y. Lichtensztajn, B. M. Giddings, C. R. Morris, A. Parikh-Patel, and K. W. Kizer, "Comorbidity index in central cancer registries: the value of hospital discharge data," Clin Epidemiol, vol. 9, pp. 601-609, 2017.

[5] J. Park, M. K. Seeley, D. Francom, C. S. Reese, J. T. Hopkins. Functional vs. Traditional Analysis in Biomechanical Gait Data: An Alternative Statistical Approach. Journal of human kinetics. vol. 60, no. 1, pp. Pp 39-49, 2017.

[6] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, "Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC)," Int J Epidemiol, vol. 46, no. 4, pp. 1093-1093i, Aug 1 2017.

[7] N. Cox et al., "Implementing Sustainable Data Collection for a Cardiac Outcomes Registry in an Australian Public Hospital," Heart Lung Circ, vol. 27, no. 4, pp. 464-468, Apr 2018.

[8] K. Gibert, M. Sànchez–Marrè, J. Izquierdo, and K. Gibert, "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining," AI Communications, vol. 29, no. 6, pp. 627-663, 2016.

[9] J. N. A. Svraka, "Importance of Data Pre-processing in Credit Scoring Models Based on Data Mining Approaches," Croatian Society MIPRO, vol. 1046, 2018.

[10] A. K. Sigurdardottir, H. Jonsdottir, and R. Benediktsson, "Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis," Patient Educ Couns, vol. 67, no. 1-2, pp. 21-31, Jul 2007.

[11] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," IEEE Trans Neural Netw Learn Syst, vol. 29, no. 5, pp. 1774-1785, May 2018.

[12] L. He, Q. Song, and J. Shen, "k-NN Numeric Prediction Using Bagging and Instance-relevant Combination," presented at the 2010 Second International Symposium on Data, Privacy, and E-Commerce, 2010.

[13] Z. Wu et al., "Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings," Energy and Buildings, vol. 173, pp. 117-127, 2018.

[14] S. Li, W. Song, H. Qin, and A. Hao, "Deep variance network: An iterative, improved CNN framework for unbalanced training datasets," Pattern Recognition, vol. 81, pp. 294-308, 2018.

[15] M. Batterham, E. Neale, A. Martin, and L. Tapsell, "Data mining: Potential applications in research on nutrition and health," Nutr Diet, vol. 74, no. 1, pp. 3-10, Feb 2017.

[16] T. M. Lanzieri et al., "A prospective observational cohort study to assess the incidence of acute otitis media among children 0-5 years of age in Southern Brazil," Braz J Infect Dis, vol. 21, no. 4, pp. 468-471, Jul - Aug 2017.

[17] C. WENHAM, "Regionalizing Health Security: Thailand's Leadership Ambitions in Mainland Southeast Asian Disease Control," Contemporary Southeast Asia: A Journal of International & Strategic Affairs, vol. 40, no. Issue 1, 2018.

[18] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100-107, 2018.

[19] A. P. Sinha, J. H. May. Evaluating and tuning predictive data mining models using receiver operating characteristic curves. Journal of Management Information Systems, vol. 21(3), pp. 249-280, 2004.

[20] X. Wang, K. An, L. Tang, and X. Chen, "Short Term Prediction of Freeway Exiting Volume Based on SVM and KNN," International Journal of Transportation Science and Technology, vol. 4, no. 3, pp. 337-352, 2015.

[21] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017 Jan 1;15:104-16.

[22] X. Zha et al., "Research on the Pearson correlation coefficient evaluation method of analog signal in the process of unit peak load regulation," 2017 13th IEEE International Conference on, 522, 2017.

[23] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70-79, 2018.

[24] H. Faris et al., "An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems," Knowledge-Based Systems, vol. 154, pp. 43-67, 2018.

[25] H. Hong et al., "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," Catena, vol. 163, pp. 399-413, 2018.

[26] J. Nordhaug Myhre, K. Øyvind Mikalsen, S. Løkse, and R. Jenssen, "Robust clustering using a kNN mode seeking ensemble," Pattern Recognition, vol. 76, pp. 491-505, 2018.

[27] A. D. Ajak, E. Lilford, and E. Topal, "Application of predictive data mining to create mine plan flexibility in the face of geological uncertainty," Resources Policy, vol. 55, pp. 62-79, 2018.

[28] J. Park, A. Mendy, and E. R. Vieira, "Various Types of Arthritis in the United States: Prevalence and Age-Related Trends From 1999 to 2014," Am J Public Health, vol. 108, no. 2, pp. 256-258, Feb 2018.

[29] S. T. Keeping et al., "The cost of oropharyngeal cancer in England: A retrospective hospital data analysis," Clin Otolaryngol, vol. 43, no. 1, pp. 223-229, Feb 2018.

[30] M. Kraft et al., "Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population," Systems sciences, 2003.

Appendix: Disease code, type and details

| Type | Name |
|---|---|
| A00-B99 | I Certain infectious and parasitic diseases |
| C00-D48 | II Neoplasms |
| D50-D89 | III Diseases of the blood and blood-forming organs and certain disorders involving the im |
| E00-E90 | IV Endocrine, nutritional and metabolic diseases |
| F00-F99 | V Mental and behavioural disorders |
| G00-G99 | VI Diseases of the nervous system |
| H00-H59 | VII Diseases of the eye and adnexa |
| H60-H95 | VIII Diseases of the ear and mastoid process |
| I00-I99 | IX Diseases of the circulatory system |
| J00-J99 | X Diseases of the respiratory system |
| K00-K93 | XI Diseases of the digestive system |
| L00-L99 | XII Diseases of the skin and subcutaneous tissue |
| M00-M99 | XIII Diseases of the musculoskeletal system and connective tissue |
| N00-N99 | XIV Diseases of the genitourinary system |
| O00-O99 | XV Pregnancy, childbirth and the puerperium |
| P00-P96 | XVI Certain conditions originating in the perinatal period |
| Q00-Q99 | XVII Congenital malformations, deformations and chromosomal abnormalities |
| R00-R99 | XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| S00-T98 | XIX Injury, poisoning and certain other consequences of external causes |
| Z00-Z99 | XXI Factors influencing health status and contact with health services |