National College of Ireland

Identification and Prediction of Factors Impact America Health Insurance Premium

MSc Research Project
Data Analytics

Jun Jun Sun
Student ID: X17162238

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jun Jun Sun ……
Student ID:	X17162238 ……
Programme:	Data Analytics …… **Year:** 2020 ……
Module:	MSc Research Project……
Supervisor:	Dr. Catherine Mulwa……
Submission Due Date:	17/08/2020……
Project Title:	Identification and Prediction of Factors Impact America Health Insurance Premium
Word Count:	9526 …… **Page Count** 25……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:  ……

Date: 17/08/2020 ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identification and Prediction of Factors Impact America Health Insurance Premium

Jun Jun Sun
x17162238
MSc Research Project in Data Analytics
17th August 2020

Abstract

For insurance companies understand the factors that impact user's health insurance premium would be very essential to make the accurate charge, premium always be a user's priority consideration to make appropriate decisions. This project used predictive analytics and insurer attributes to identify the factors that influence health insurance cost, according to the output which demonstrated the majority factors that contribute to health insurance premiums cost are BMI, smoke status, age and children, these four factors have significant correlation impact to health insurance premiums. Through discovery the correlation between individual's attributes, utilized 3 regression models and 1 statistical model to solve the research question and provide meaningful insights for insurance companies, and used another seven classification models to resolve the sub research question. The comparison and evaluation of several model outputs that determine the most effective and best performance model implemented to achieve the research question is Random Forest model with 80% R-square value, Support Vector Machine is a second performance model with 67% accuracy. Also, SVM and Random Forest used to solve research question and sub research question.

Key Words: Machine Learning models, Predictive Analytics, Insurance Premiums, R-square value, Accuracy.

1 Introduction

United State health insurance was extremely high, which covered 91.5% in 2018, but is 0.4% decreased from year 2017 to 2018 (Berchhick, et al., 2019). And 8.5% of the population (27.5 million people) still do not have health insurance (Berchhick, et al., 2019). Much of the recent debate over health care in the United States has focused on the cost of health insurance, which is not affordable for many Americans. The insurance premium is likely to continue to increase, this project was assured everybody understand the factors that would impact cost.

1.1 Motivation and Background

Prediction models focus on future predictions and provide some other useful information, such as the contribution of variables for predictive analysis, such as inferring the dominant factor in a customer's purchasing decision. Prediction model is significant for different health insurers premium and other purposes (Lahirih & Agarwal, 2014). For health insurance industry growing, the healthcare delivery systems precise forecast of cost would assist the business planning, also for insurer's, by advance, knowing their insurance cost would help them choose the different insurance plan type with the appropriate premium.

How much health insurance premiums cost that you need to pay depends several factors, based on the relevant literature research, author (Goleiji L, 2015) analyses the health insurance premiums cost from seven factors which including individuals' age, sex, BMI, children amount,

smoker, region and charger level, use prediction analysis has enhance health insurer's their premium pricing accuracy, establish individualized health insurance plan type, and set up good customer relationship.

Based on a comprehensive understanding of the health insurance characteristics demand, guarantee strong operation of the health insurance market and promote its healthy development. At the same time, studying the factors affecting the demand and pricing of health insurance, it would have positive effect on the product development and innovation of insurance companies, also assist insurance companies define consumer's requirements and preferences, further meet their health protection needs and improve the overall competitiveness of enterprises.

1.2 Project Requirement Specifications

With the rapid development of economies, the health insurance industry has entered the era of fierce competition. The face of a large amount number of generating policy and business, many companies do not have a large amount of data for in-depth analysis and mining, so that large amounts of data play a huge value-added role. The way the data are organized in its business warehouse is to satisfy the paradigm theory, which reflects the way the business or organized, but if is fundamentally different from the business system. First, extraction of information is designed for future analytical applications; second, the data in the business data warehouse are cleaned and all business history information is preserved to maximize the protection of data from distortion, in order to use all the data to generate comprehensive data mining analysis.

1.2.1 Research Questions

RQ: *" To what extent can aspect from data mining analysis of American health insurance industry used supervised machine learning techniques(Multiple Linear Regression, Support Vector Machine, Random Forest)and statistic model (ANOVA) to deliver meaningful insights for health insurance premium base from an individual's behaviour and predicted contributed factors?"*

Sub-RQ: *"Can we predict new insurer (smoke or non-smoke) behaviour based on individual's attribute using machine learning algorithms (Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbour, Support Vector Machine and Random Forest)?"*

1.3 Research Project Objectives

Objective A: Gather online dataset from different sources through several tools and programming techniques and clean the dataset to remove missing value.
Objective B: Compute the score of different factors would impact health insurance cost.
Objective C: Identify the parametric and non-parametric method, used SPSS statistical predictive analysis and determine the factors that impact America health insurance premiums.
Objective D: Implementation, evaluation and review finding of predictive model used data mining techniques, allocated data into a suitable chosen algorithm. Objective D is divided into 7 sub-objectives as in table1.

Table 1 Sub - Objectives

Object D I	Implementation, evaluation and result of Multiple Linear Regression model
Object DII	Implementation, evaluation and result of Random Forest model
Object DIII	Implementation, evaluation and result of Support Vector Machine model
Object DIV	Implementation, evaluation and result of Naïve Bayes model
Object DV	Implementation, evaluation and result of Decision Tree model
Object DVI	Implementation, evaluation and result of Logistic Regression model
Object VII	Implementation, evaluation and result of K - Nearest Neighbour model

1.4 Research Project Contributions

The essential contribution of this project was completely developed several models (Table 2), used data mining technologies to analyse America health insurance premium, and used R configuration functions and libraries in RStudio to developed. Predictive analysis was done used correlation coefficient defined the significant factors that impact health insurance premiums. The regression models represent R-Square results, and classification models represent accuracy result from confusion matrix.

Table 2 Research Contributions

Developed Multiple Linear Regression model and R-square result to solved the RQ
Developed Random Forest model and result to solved the RQ and the Sub-RQ (regression model generated R-square for RQ & classification model generated accuracy for Sub-RQ)
Developed Support Vector Machine model and result to solved the RQ and the Sub-RQ (regression model generated R-square for RQ & classification model generated accuracy for Sub-RQ)
Developed Naïve Bayes model and accuracy result to solved the Sub-RQ
Developed Decision Tree model and accuracy result to solved the Sub-RQ
Developed Logistic Regression model and accuracy result to solved the Sub-RQ
Developed K-Nearest Neighbour model and accuracy result to solved the Sub-RQ

The remaining parts of technical report are as follows: Chapter 2 is indicating the associated related work of predictive analysis within American health insurance industry from year 2019. Addition chapter 3 present the modified CRISP-DM methodology approach. Chapter 4 introduced the implementation, evaluation and finding results of seven different supervised machine learning algorithms. Eventually chapter 5 describe the conclusion of implement results and recommendation of future work.

2 Literature Review of Predictive Analytics in U.S. Health Insurance Industry

2.1 Introduction

Predictive models can alert insurance companies to potentially high risk. Using allocated resources and information effectively analysis and forecasts will assist insurers improve their ability to innovate and respond quickly to market change. From our research there was a lack of prediction model algorithms review of the literature (Burns, et al., 2019). Thus, this project conduct prediction model literature review of American health insurance premiums. Comparing different approaches on the health insurance dataset, applied and evaluated machine learning technologies, justify the result to provide meaningful insight.

Health insurance companies are restructuring the entire value chain to increase transparency, to adapt the changing business environment. In order to meet the changing preferences of users, they are developed value-added services to create more engaging experiences. At the same time, they strategically leverage new technologies such as predictive analytics to address emerging risk management and improve operational efficiency (Watson, 2019).

Health insurance companies have extensive experience in underwriting and price health related risks. However, traditionally the measurement of individual risk index of insurance decision-making is still insufficient. Among big data and internet technology growth, insurance companies can access to large amounts of customer data and generate actionable insights. Rising health care costs are forcing health insurance companies to identify high-risk customers, Insurance companies are using predictive analytics techniques for risk analysis and provide

early intervention. Through insightful predictive analytics, assess potentially high-risk customers this reduces losses and increase profitability (Berchhick, et al., 2019).

Prediction model analysis ensures the completeness an accuracy of the data in the modelling process. 67% world's insurance companies use predictive analytics in their business (Frees, et al., 2016). Society of Actuaries (SOA) generated research from year 2009 to 2015, through calculated insurer's renew insurance policy status, from year 2010 to 2014 that insurance companies lapse rate increased from 65.7% to 69.9% (Michael Ewald, 2015). When developing exist or new products use traditional methods, it normally follows the previous lapse rate plus safe boundary, but according different insurance factors impact, when insurance premiums increased the lapse rate also increased.

From schema and associated that defined substantial of insurers, the predictive model information obtains insights in the health insurance industry, also defined the appropriate factors from a statistical perspective (Fang. R, 2016). SOA tried GLM modelling in 2010, the predictive result accuracy has increased to reduce lapse rate. Insurance companies from data collected forecasted lapse rate (Michael Ewald, 2015).

Machine learning algorithms and predictive model analytics are growing rapidly in the health insurance industry. In many industries of successful machine learning algorithms was well studied academically (Siegel, 2016). Therefore, the aim of this research is to define suitable predictive model by using machine learning algorithms in the health insurance industry, assist business and organization expand more feasible policies while conduct insurance rate plan.

2.2 Concept of Predictive Analytics and its Working

The data science and insurance industry are naturally compatible, and the insurance industry is based on data science. Because the insurance products are based on predictions of the future, and the future must be based on a summary of historical trends. The core of insurance products is developed a unique pricing with demand orientation for different groups of people, identify its potential markets and influencing factors, and establish a functional relationship between demand and the main influence, to predict the future level of market demand and its changes (Sivagnanam & Srinivasan, 2010).

The methods of prediction include qualitative prediction and quantitative prediction (Goldstone, 2008). Qualitative prediction through investigation the people who familiar with the situation, make predictions based on experience and awareness of phenomena, e.g. customer opinion prediction, marketing staff opinion, and expert discussion opinion method (McDonald, 1993). Quantitative prediction is based on the quantitative relationship between variables, such as time relationship and correlation to establish modelling, quantitative prediction which divided into time series and regression prediction (Bao, et al., 2013).

The time series method, also known as the trend prediction method, is a method of scheduling historical data in chronological order, predicting according to the continuity of the development of phenomena to value the insurance claims (Frees, May, 2013). Regression prediction is based on the matching data, put independent variables into the model to make predictions of the dependent variable, the closer the model is to the high data, the better the prediction result, the common regression model is a linear and logistic regression model. Insurance companies using a regression model to forecast the customer's insurance and medical costs (Kim & Basu, 2016).

2.3 Impact of Predictive Analytics Practices

Predictive analytics is also a statistical or data mining solution, that includes algorithms and techniques can be used in structured and unstructured data to determine the future result.

Deployed for many other uses, such as forecasting, optimization and simulation. As well as providing information for the planning process and critical insight into the future of the enterprise. Predictive and hypothesis analysis helps users review and weight the impact of potential decision from historical patterns and probabilities (Dey, et al., 2017).

Machine learning is particularly numerical computation, the validity of final model usually depends directly on whether the amount of training data is large enough and whether the number of algorithm iterations is sufficient. but more data does not necessarily help to predict the future, for example, the most recent year's data trend give more accurate forecast compare to previous three years data of what will happen over the next three months, because some implicit trends have changed over past three years, the recent data trend is more representative (O'Boyle, 2019).

2.4 Comparison of Techniques

Data mining have been well used in various industries which including the insurance industry, and it received obvious benefits, data mining used patterns of customer behaviour analyse and determine suitable plan type for different individual. Machine learning is technology field that pursues relationships between computer science and data information. The type of learning used by computers is conveniently subclasses into categories such as supervised learning and unsupervised learning (Lip GY, 2010). Supervised learning is classification and regression model, it is used when data labelled and are trying to predict future output based on known features (input data variables) often used the applications where historical data predict future value. Unsupervised learning algorithms used clustering and reducing dimensional model, when we have unlabelled data and trying to find hidden patterns and group with similar features in the data, it normally identifies clusters, the mainly used explore the data and find intrinsic structure within the input data (Weng, 2018). Health insurance premium cost prediction model is mainly using supervised learning algorithms.

2.4.1 Critiques of Supervised Machine Learning Models

Supervised learning aims is training sample with attribute labels, which can be understood as a sample with input and output. All regression algorithms and classification algorithms are supervised learning. The algorithm difference between regression and classification is the type of output variable, the quantitative output is called regression, or continuous variable prediction, qualitative output is called classification, or discrete variable prediction, supervised learning methods were conducted for this project.

Multiple Linear Regression: Multiple Linear Regression is one of the most used algorithms for regression tasks, is a simple form of an algorithm that expects to use a hyperplane-fit dataset (a straight line when only two variables are available). Regression minimizes the distance between the observation point and model line. If the curve is a straight line, it is called Linear Regression, if the curve is duplicated, it is called multiple regression. In fact, a phenomenon is often associated with multiple factors, predict or estimate dependent variables by an optimal combination of multiple independent variables, is more efficient and more realistic than using only one independent variable. Therefore, Multiple Linear Regression is more practical than one dimensional Linear Regression. The method of prediction is regression coefficient is multiplied by the input value and then added up to get the predicted value. GLS also called Generalized Linear model is a more flexible process than the Linear Regression model, it enables response variables that have deviation distribution instead of have a normal distribution (Ohlsson, 2010). Author (Freyder, 2015)applied Linear Regression model to estimate the

average cost of health plan after inpatient event and defined females tend is cost less than male. Author (Kuo, et al., 2011) generated 40% R^2 value for healthcare cost prediction analyses.

Random Forest: Random Forest as a new emerging and high flexible machine learning algorithm has wide applications for healthcare and insurance industry, which can be used for regression prediction and classification modelling customer source of marketing. There are two key words in the name of Random Forest, one is "Random" and other is "Forest". Hundreds of trees can be called a forest, Random selects sampling treating sets, each tree's training set is different sample. Random Forest is an algorithm that integrates multiple trees through the idea of integrated learning approach, this also the main idea of Random Forest, the embodiment of ensemble ideas. Random Forest is an algorithm that integrates multiple trees through the idea of the integrated learning approach. Random Forest has demonstrated better performance than other classifier for identifies potential high cost health insurance users (Kim & Park, 2019), it generated 40% R^2 value for large amount healthcare dataset. (S, et al., 2015)

Support Vector Machine: Support Vector Machine (SVM) is widely used for regression and classification, it turns classification problems in finding classification plane, the classification is realized by maximizing the distance of the classification boundary point from the classification plane. SUV method is based on the principle of minimizing the structural risk of computational learning theory, its main idea is to address the problem of two classifications. Find a super plane in high-dimensional space as a split of two types to ensure minimal classification error, the SVM approach is ideal for solving high-dimensional, non-line two classification problem, and because of its principle advantages, its classification accuracy is better than traditional Decision Tree and other classification methods. Generally, the SVM performance better than other algorithms, partially for nonlinearly data (Tian, 2012). For analysis large healthcare data, it conducted 41% R^2 value from article resource (S, et al., 2015).

The discussion and research of SVM are gradually widely applied well in pattern recognition, function approximation, data mining and nonlinear system control. This method has also been applied to the financial field, mainly the prediction and classification of time series, and some scholars have carried out credit rating research using the SVM method (Xu Jianhua, 2004). But it has been rarely literature on the application of SVM to customer churn analysis, the main reason is an unbalanced dataset with differing number of positive and negative samples, these data contain a relatively small number of positive samples but it is necessary to detect them. SVM classification surfaces are affected by the uneven sample distribution, bias to positive samples, resulting a significant reduction in the effectiveness of SVM algorithms (Wu G, 2003).

Naïve Bayes: Naïve Bayes is a statistical classification based on Bayes' theorem, which is classified by predicting the probability that tuple belongs to a particular class. The Naïve Bayes assumes that the effect property value of a given class is independent of other property class conditions. Naïve Bayes is one of the most widely used classification algorithms, it gives the category to be classified, in order to get the probability that each category will appear under the condition severity of this occurrence, which is the largest, which category is considered to fall into the category. Naïve Bayes use features selection methods for boosting procedures (Ch. Ratanamahatana, 2002), performance as well as C4.5 Decision Tree (H. Zhang, 2005). Naïve Bayes classifier accuracy was improved by combining with other methods in several researches (H. Zhang, 2005), experiments have been done with up to 4 attributes and accuracy result between 67% to 70% in health sector (Hickey, 2013).

Decision tree: Decision Tree approach originated in the concept learning system (CLS), then developed to ID3 approach, eventually evolved into C4.5 that could handle continuous properties (Gepp, 2012). Decision tree is a simple but widely used common classification method, its thoughts are very similar to the process of human progressive analysis, make a comparison and conclusion. Decision tree includes binary tree and non-binary tree, every non-leaf node represents a feature property, every branch means the output of this feature property on a valuable domain, every leaf node holds a category. Decision making process using the decision tree which starting at the root node, test the corresponding feature attributes on the category to be classified and select the output branch according to its value, until reach the leaf node, the category in which leaf node is stored as decision. The author conducted experimental work by used CART and C4.5 Decision Tree model to classify the healthcare data, after automatically search for best subset of features selected and applied, the result determination that CART accuracy rate at 80.86% and C4.5 accuracy rate at 77.8%, CART was performing better o C4.5 Decision Tree model (Hota & Dewangan, Sep 2016).

Logistic Regression: Logistic regression is used to deal with the regression of dependent variables as categorical variables, the most common problem is two category or two item distribution, multi category issues can also handle, it is a classification method. The principle of Logistic Regression is like continuous regression algorithms, but often deals with the formal classification algorithms, it has a close relationship with Support Vector Machine classifier. The Logistic Regression model was proposed by statistician David Cox (Hand & A.M.Herzberg, 2005). It is a classification model, distributed by conditional probability. And Logistic Regression is a nonlinear regression model compared to Linear Regression. From the research that define data sources is critical factors impacting Logistic Regression model establish, the limitation of data source would significant effect result of coefficients and lack accuracy (Duan, et al., 2018). Author (Kismiantini, 2018) conducted Logistic Regression model forecasted probability of having health insurance of an age, the method was binary classification, the classification accuracy result is 56.5%, which is moderate model.

K-Nearest Neighbours (KNN): KNN is a simple and practical algorithm in Machine Learning, there are two main differences from the algorithms discussed earlier: It is a non-participation method. There is no need to have a fixed-format model like Linear Regression or Logistic Regression, nor need to fit the parameters. It can be used for both classification and regression. In the classification area, for an unknown point, select the point closest to the K distance (which can be Euclidean distance or other similarity metric), and then count the K-point, the class with the largest frequency number in this K-point as the classification result (Harrison, 2018). With research from (Attewell, et al., 2015)that KNN classifier performance correctly at 75% of insurance data.

2.5 Identified Gaps in Predictive Analytics in U.S. Insurance Industry

Supervised learning method is widely used machine learning method, but it needs to give the expected output in different environment states (Ahmed, et al., 2018). Combination the memory and knowledge without interaction environment would limit the application of this method (Ahmed, et al., 2018). Bases from literature review research, that quality of data accuracy and data access issue have not been significant indicated by another author in the health insurance premium analysis. The big data accessibility, evaluation with the academic research approach is a gap to identify the health insurance premium. but from other research of experimental results in a variety of health domains, define that combine two models have advantages to promotional ability (Wang, et al., 2006).

2.6 Conclusion

Data mining models can assist insurance companies provide appropriate services based on predictive analytics to increase customer satisfaction and brand loyalty. According to above relevant researches, shows that the best knowledge of the candidate has been done with predictive analytics in the health insurance industry, there are many researchers has been done construct insurance used machine learning analysis (Dey, et al., 2017). Most of them focus on health premiums establish to obtain more accurate insight, but not enough research has been done features analysis in the health insurance industry, the big dataset is required from customers for feature analysis. Many Insurance companies believe their predictive analysis met their expectations, but data acquisition remains a huge challenge, insurance companies rely on traditional insurance data in modelling but obtain historical data can be a daunting task. The models used in this paper still used the traditional model without innovative proposed a novel model, this is the direction of follow up research need to continue pay attention.

3 Scientific Methodology used

3.1 Introduction

This project was introduced health insurance relevant fundamental knowledge and technologies. Compared different technologies attribution. This chapter represents a scientific methodology used CRISP-DM method for U.S. health insurance industry, the main approach of CRISP-DM has been successfully used different domain, the following research purpose is introducing scientific methodology process and use in the health insurance field.

3.2 America Health Insurance Methodology Approach

A project plan is fundamental and essential procedures of every research. According to consumer behaviours feature to establish analysed, set the most appropriate segmentation point for the future consumer effect factor of the pricing point to help decision making. The CRISP-DM model provides complete process description for Knowledge Discovery in Databases (KDD) project. Standard according to CRISP-DM, the model of data mining in the insurance industry can divide a KDD project into six different stages (Azevedo & Santos, 2008), but the sequence is not completely unchanged, shows in the figure 1 chart.

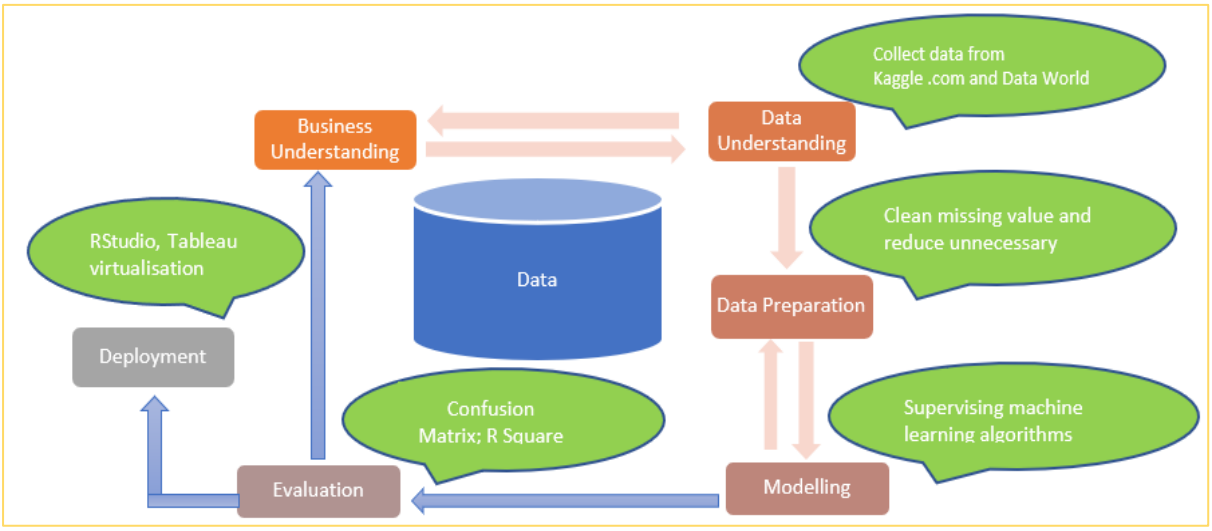


Figure 1: America Health Insurance Methodology Approach

Business understanding – This first stage we must understand the requirements and ultimate objectives of the project from a commercial perspective, the purpose is combining these objectives with the data mining definition and the result. Clear mines the objectives and identify the high-risk claims from customer characteristics that would affect results, guide company in marketing decision.

Data understanding – This stage begins data collection and get familiar with the data, there are five csv file datasets that would contribute to this research project. The first two datasets were extracted from U.S insurance company from Kaggle[1], each of it contains 1338 rows, 8 and 9 variables. Other two csv files include insurance training and testing datasets, it was extracted from Kaggle[2], Training data contain 43,400 rows and 12 variables; Test data contain 18601rows and 11 variables. The last dataset hospital inpatient discharge dataset was extracted from Data World[3], this dataset contains 163,065 rows and 12 variables.

Data preparation – The data preparation stage covers all the work of building the final dataset from the raw dataset that will serve as an analysis object for the modelling tool. Large and complex dataset is pre-processed, the missing data are eliminated, and the target is distinguished from the interpretation of variables. Also, the high-risk customer datasets are filtered from the perspective of description statistic.

Modelling – This stage a variety of modelling methods will be selected and used, model building based on different data mining objectives and data characteristics, the model is establishing using different mining algorithms, evaluation model calibrates its parameters to the most desirable value.

Evaluation – From data analytics perspective, this stage has been established high quality model, but it is important to evaluate the model before making a final model deployment. The models result need be compared to validation, accuracy verification, support verification and other tests to determine the value of the model like confusion matrix evaluation.

Deployment – This last stage is organizing its finding and process into a readable text form, create modelling is not the ultimate purpose of the project. Although modelling is intended to add more information about the data, this information still needs to be organized and presented in a way that customers can use. Only publishing the model to the decision maker can truly reduce probability cost of claims of insurance business.

3.3 Data Extraction and Data Pre-Processing

This project collected five datasets from online resources; first partitioned dataset contains insurance customer individual behaviours factor related to insurance charge. Second partitioned dataset consists the different premiums from different region in America and more variables that not related to project after research.

The following procedures (Figure 2) show the measure of data extraction:

- Research for different website regarding topic defined.
- Download appropriate dataset and import and extract that from online resource into csv file used RStudio programming language.
- After import, the query page used machine learning algorithms to identify what information that will require for this research.

1 https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset
2 https://www.kaggle.com/hiralpandhi/healthcaredataset?select=test_2v.csv
3 https://data.world/healthdatany/gaf8-ac33

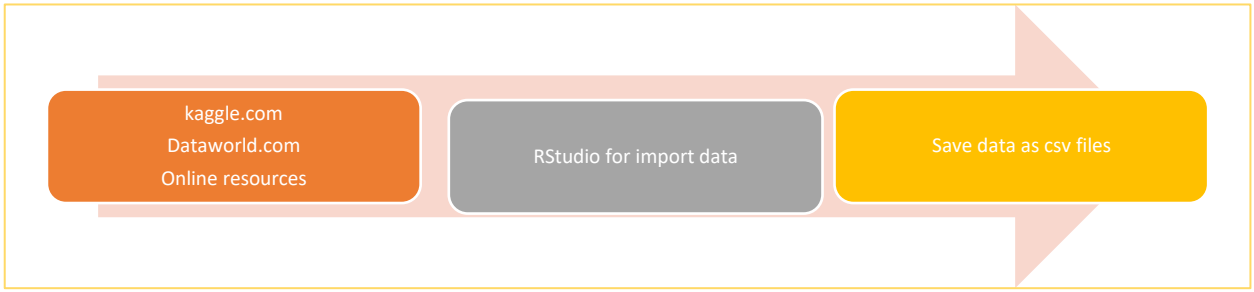


Figure 2: The Process Applied During Data Extraction

Following procedures (Figure 3) show the measures of data pre-processing:

- First converted three variables Sex, Region and Smokers to categorical variables, then converted them to numerical variables to consistent with building modelling.
- Clean dataset and removed missing values for analysis and building models.
- Split data into training and testing data, then evaluation each model output of by R-square (Regression model) and accuracy (Classification model).

Data analytics steps are as follows:

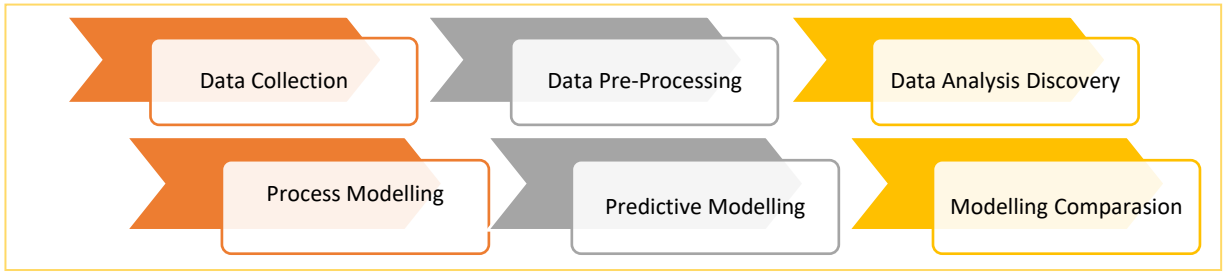


Figure 3: The Process Applied During Data Pre-Processing

3.4 Architectural and Technical Design

This project of predictive analytics research and implementing will utilize three layers skeleton. The first customer layer will contain user interface and visualizing the result from U.S. insurance industry, the second layer is business layers include main predictive analytics using different machine learning algorithms and technologies. The third and final layer is the data persistence layer that used RStudio programming language and SPSS software, the data gathered in RStudio. The below chart (Figure 4) representing each layer process.

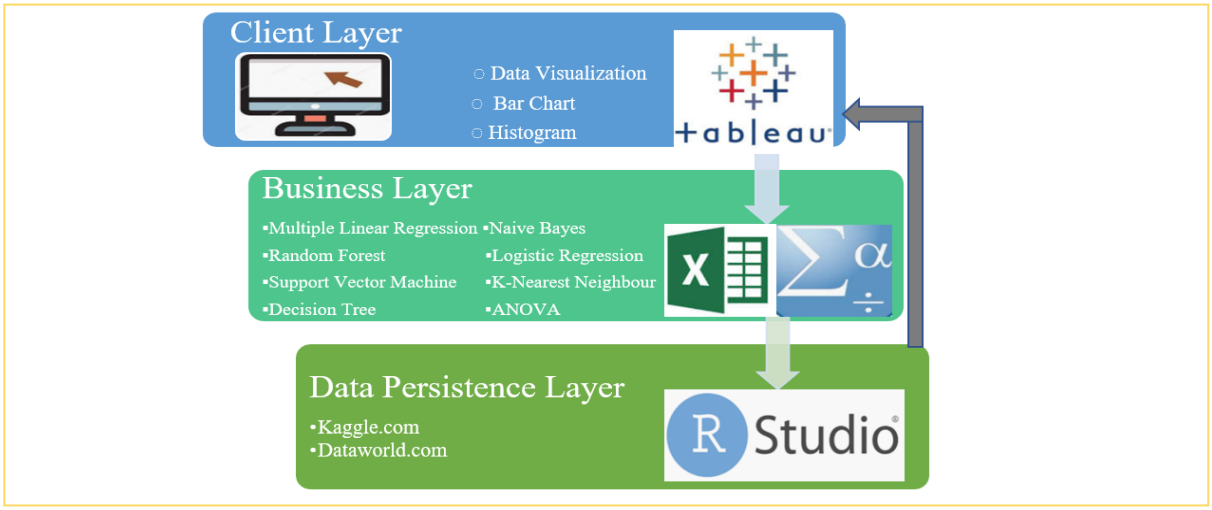


Figure 4: Three Layers Architecture Design

3.5 Conclusion

This chapter represents a scientific methodology used to collect online data sources, this project used CRISP-DM technique approach, CRISP-CM approach has been adopted for this project planning. The process flow determines the workflow of this research from start to finish, there are three tiers framework has been carried for this research project framework, which including client layer, business layer and data persistence layer. The following chapters will focus on how predictive model implements for this research project.

4 Implementation, Evaluation and Result of Models

This chapter presents dataset collection, pre-processing and progress to final utilization. It also contains the implementation and evaluation results of machine learning models and prediction analysis performance. To evaluate the models and determine result by used the confusion matrix method to conduct accuracy result for classification models, and R-square result from compared regression models. This research project is very significant to solve the research question and sub-research question, at the end it will illustrate and compare final overall outputs.

4.1 Introduction

Insurance company calculates the individual premium based on model generated from historical data, this chapter introduce the realization and evaluation several models of prediction analysis, use the solution process of several models as an experiment to representing the best performance model by R-square value and accuracy, the purpose of this analysis is to determine and predict the impact of several factors to health insurance premium.

4.2 Prediction Data Analysis Process Chart

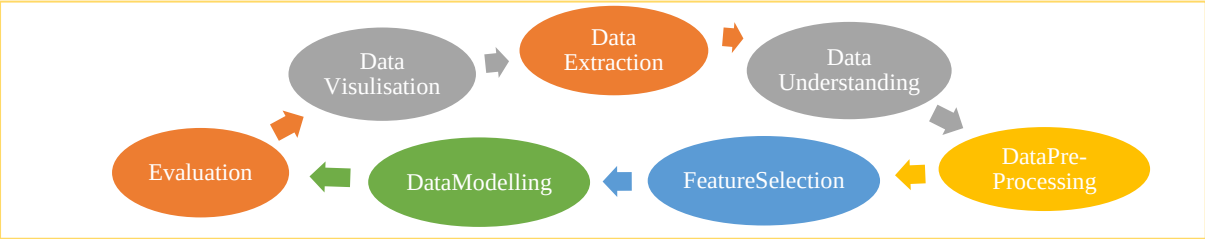


Figure 5 Data Flow Chart

The mathematical model is analysis the potential relationships and predict desire output. The process cycle of flow chart base of prediction analysis shown in figure 5, the flow begins with data extraction from different online resource. After data extraction and data understanding are data pre-processing, data pre-processing includes data cleaning to remove NA values, encode and transform, the data preparation used RStudio analysis the variables and conduct feature selection work. Then 7 machine learning models were used to conduct analysis and evaluate the R-square and accuracy output. Finally presents model visualisation.

4.3 Data Extraction and Pre-Processing

The essential part of this project is the data extraction and pre-processing. There are 5 csv datasets extracted and integrated to conduct this prediction. Four datasets from Kaggle and one from Data World, they are all open source public datasets. All data extraction, data cleaning, transforms and encode work was through RStudio, this section was completed objective A.

4.3.1 Prediction Data Analysis

Predictive analysis uses statistical techniques and 7 different algorithms to forecast the future situation used RStudio and SPSS. Objective B and objective C were generated from this chapter.

4.3.2 Data Extraction

The data extraction workflow covered searched and check data from raw data sources. The first step was analysed process to identify a research question, then selected appropriate method for data extracts and stored them in the provisionally area of relational database for future use.

4.3.3 Data Pre-Processing

Used RStudio and SPSS to conduct data distribution overview.

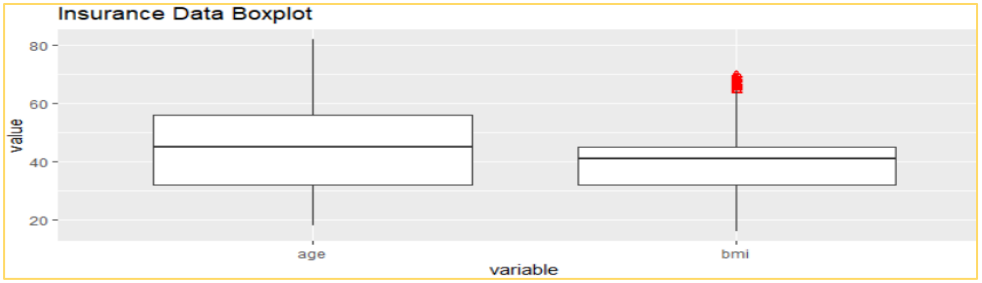


Figure 6 Boxplot of Age and BMI

The above figure 6 applied boxplot used age and BMI to represents this insurance dataset, the median is shown by a middle line of both boxes, it shows the median age for this dataset is approximately around age 45, and median BMI is about 40, this is a quite high BMI value.

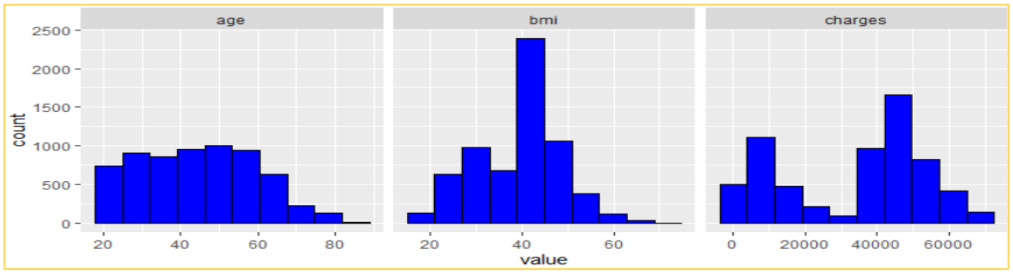


Figure 7 Histogram of Age, BMI and Charges

There are three histograms in above figure 7 which shows the distribution of age, BMI and charges. The age group from 20 – 60 counts are quite comparable. The median of BMI at approximately 40 - 45, and charge value is split into two parts, this cause the dataset extract was from 5 different sources. And the price range was spread distributed.

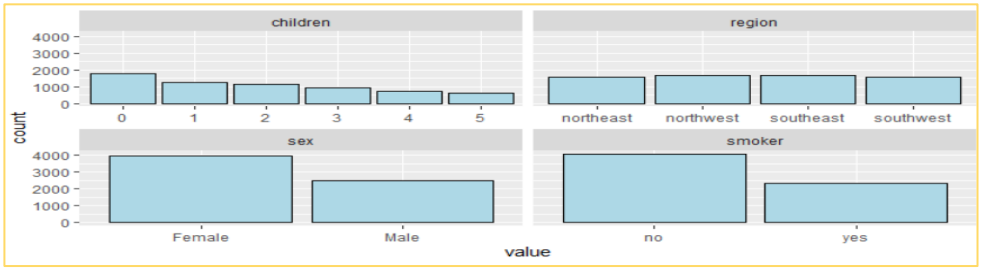


Figure 8 Children, Sex, Region, Smoker Distribution

Figure 8 shows distribution of four variables with value count. Most users have non children, four regions evenly matched. Sex variable presents Females at higher count than Males. Smoker (yes) value is nearly half the amount compared to the clients who do not smoke.

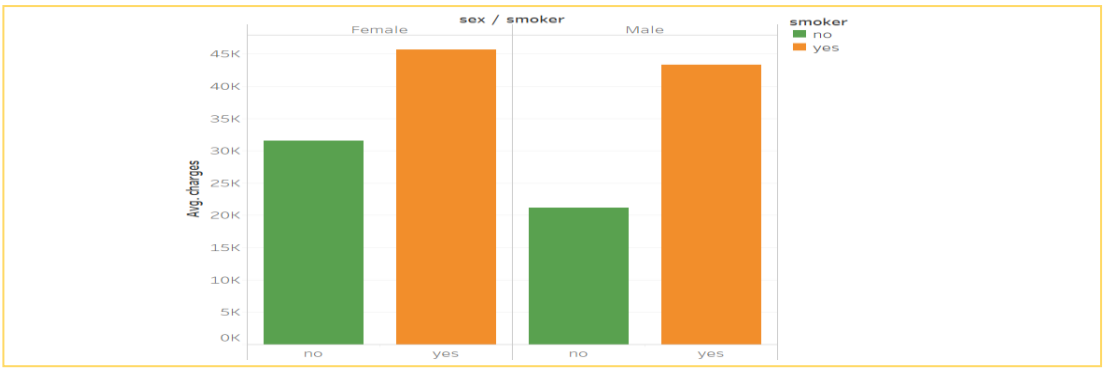


Figure 9 Insurance Premium Distribution for Smoke and Sex

Figure 9 histogram demonstrate the health insurance premium with sex and smoker status, both male and female non-smokers paid less than smokers, which matches our expectation.

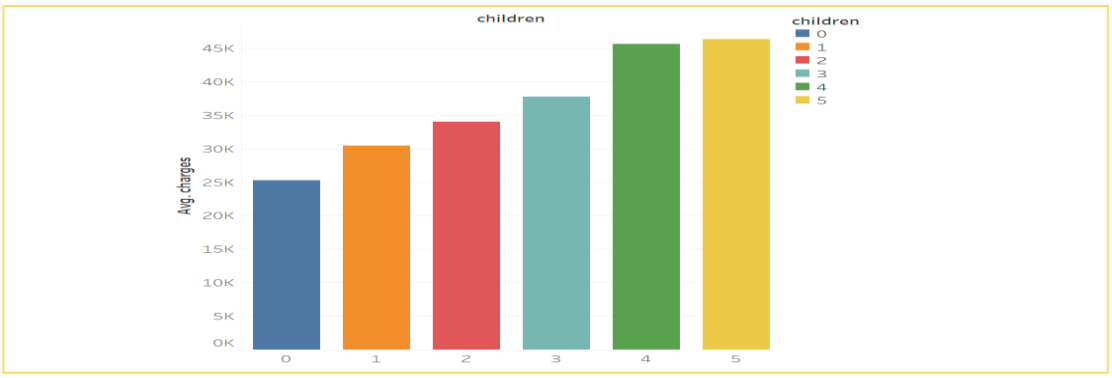


Figure 10 Premium Charges with Children Amount

Figure 10 shows children amount with charge relationship, for users who do not have children pay less health insurance premium on average of 25,000 per annum, compared to users with a highest price at 45,000 per annum with 4-5 children. This also matching our expectation.

4.4 Prediction Analysis and Feature Selection

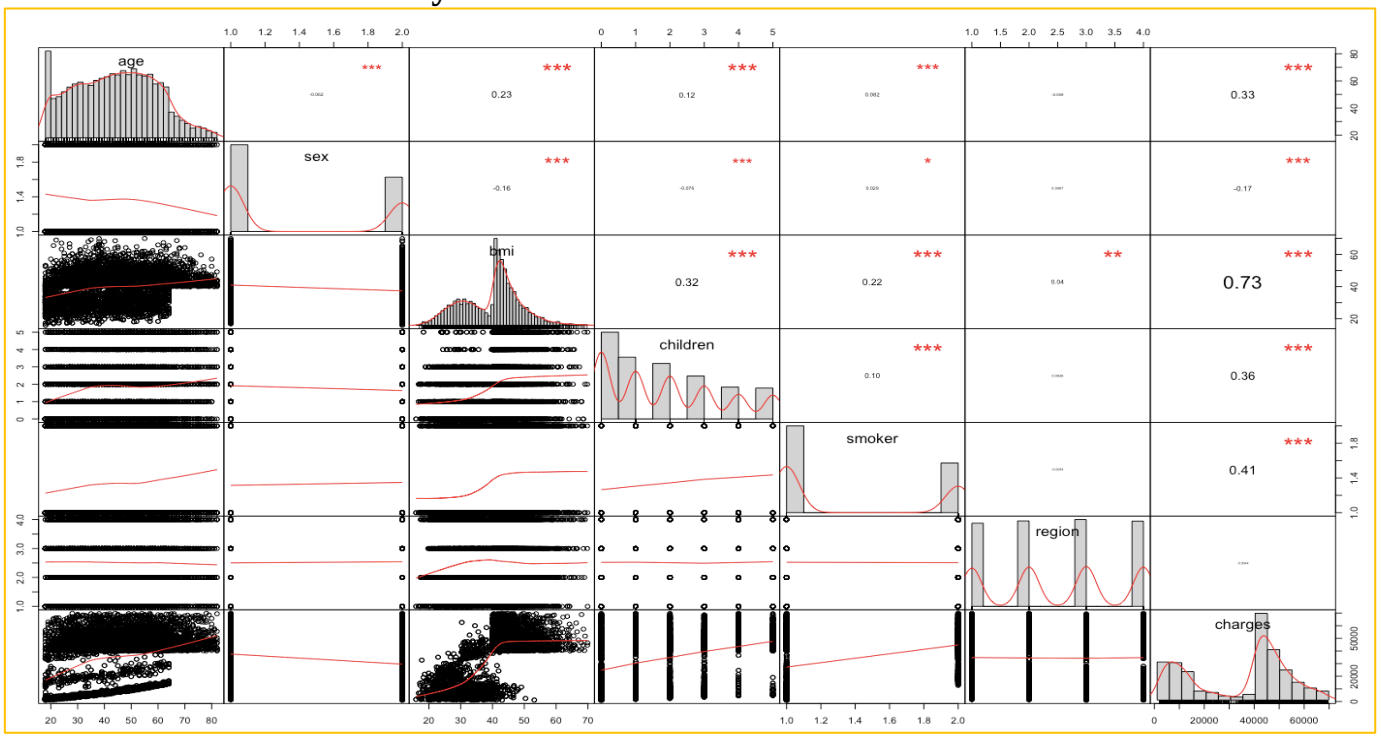


Figure 11 Correlation Coefficient of Different Variables

Feature selection used collected datasets and selected explanatory variables to presented into model, to determine an adequate measurement to introduce the correct variables if they improve the outcome measurement or delete the variable if they do not contribute to the model. Figure 11 above illustrated the correlation coefficient between the dimensions. Objective B was identified in this chart. The closer the correlation coefficient to 1 would represent the more positive correlation, the closer to -1 represents the more negative correlation, the close to 0 represents not related, we can use correlation coefficients to reduce the dimensions and exclude some irrelevant dimensions. BMI level has 73% and smoker has 41%, the obesity index of BMI level is positively correlated to premium, the following by age, children and sex. The region has least correlation with charges. Table 3 below correlation result also can determine the objective B use SPSS analysis.

Table 3 Correlations

Correlations

		age	bmi	children	region	charges	sex	smoker
age	Pearson Correlation	1	.235**	.122**	-.009	.329**	-.062**	.082**
	Sig. (2-tailed)		.000	.000	.470	.000	.000	.000
	N	6406	6406	6406	6406	6406	6406	6406
bmi	Pearson Correlation	.235**	1	.323**	.040**	.727**	-.165**	.219**
	Sig. (2-tailed)	.000		.000	.001	.000	.000	.000
	N	6406	6406	6406	6406	6406	6406	6406
children	Pearson Correlation	.122**	.323**	1	.003	.357**	-.076**	.104**
	Sig. (2-tailed)	.000	.000		.825	.000	.000	.000
	N	6406	6406	6406	6406	6406	6406	6406
region	Pearson Correlation	-.009	.040**	.003	1	-.004	.009	-.008
	Sig. (2-tailed)	.470	.001	.825		.723	.487	.501
	N	6406	6406	6406	6406	6406	6406	6406
charges	Pearson Correlation	.329**	.727**	.357**	-.004	1	-.173**	.410**
	Sig. (2-tailed)	.000	.000	.000	.723		.000	.000
	N	6406	6406	6406	6406	6406	6406	6406
sex	Pearson Correlation	-.062**	-.165**	-.076**	.009	-.173**	1	.029*
	Sig. (2-tailed)	.000	.000	.000	.487	.000		.022
	N	6406	6406	6406	6406	6406	6406	6406
smoker	Pearson Correlation	.082**	.219**	.104**	-.008	.410**	.029*	1
	Sig. (2-tailed)	.000	.000	.000	.501	.000	.022	
	N	6406	6406	6406	6406	6406	6406	6406

**. Correlation is significant at the 0.01 level (2-tailed).

After clean and combined 5 datasets, the final generated datasets have a total number with 6,406 rows and 7 variables, the variable meaning is shown below in table 4.

Table 4 Final Dataset Variable Meaning

Variable Name	Variable Meaning	Data Type
Age	Insurer age between 18-82	Int
Sex	Male=1; Female =2	Factor
BMI	Value of weight Range between 15-70	Numeric
Children	Amount of Child from 0-5	Int
Smoker	Yes=1; No=2	Factor
Region	Four Different Region-North East; North West; South East; South West	Factor
Charges	From $1,121 to $69,833	Numeric

4.5 Implementation, Evaluation and Results of Prediction Modelling – Regression and Classification

This section uses machine learning techniques to implement and evaluate the research question and sub-research question. Research question would apply regression models, and classification models would apply to sub-research question. All the process of evaluation and modelling are conducted in RStudio, the objective C is used SPSS statistical model and objective D is used RStudio to run 7 different machine learning algorithms, different algorithms were generated the

expected results. Regarding the previous literal review of seven different algorithms, output results are compared and analysed with measurement of R-square and accuracy. Each algorithm in different scenario are for different reasons, the regression model is different as a classification model that used confusion matrix to defined indicated accuracy, the regression model is for numerical prediction that normally used P-value or R-square value.

The aim of the research question is prediction analysis the factors impact insurance premium charges, which is numeric outcome, therefore, regression models like Multiple Linear Regression, Random Forest and Support Vector Machine models are utilized for this objective; Sub-research question is classification analysis, therefore this project used Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbour model, also Support Vector Machine and Random Forest classification model for analysis.

All the models used split data training and testing. This was divided up into 70% training data and the remaining 30% for testing data, both data were saved into csv files to be used on each model for reusability to save time, duplication and complexity and guarantee that every run will produce the same output.

Table 5 ANOVA Model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.801[a]	.642	.641	11950.46168	1.724

a. Predictors: (Constant), region, age, sex, smoker, children, bmi
b. Dependent Variable: charges

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.636E+12	6	2.726E+11	1909.060	.000[b]
	Residual	9.139E+11	6399	142813534.4		
	Total	2.550E+12	6405			

a. Dependent Variable: charges
b. Predictors: (Constant), region, age, sex, smoker, children, bmi

Before conducted machine learning models, table 5 used SPSS statistic tool generated ANOVA model, the F-value is very large and P-value is very small below 0.05 alpha value, these two values would determine is correlation between predictive variables (region, children, sex, smoker, age BMI) with charges, also from the R-square value at 0.642, means approximately 64% dependent variable would contribute to the model. It is representing the model is suitable for predictive analysis of health insurance premium charge.

4.5.1 Implementation, Evaluation, and Result of Multiple Linear Regression

Multiple Linear Regression model dependent variable is numerical with two or more independent variables to conduct the prediction analysis. Suppose that dependent variable function interprets other variables and some random noise.

The model generates by following format as $y = \beta o + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk + \varepsilon$
Where dependent variable is y and $\beta o + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk$ are independent variables and ε is random noise(error), βi is the contribution value of independent variables. βo is the y intercept which is the dependent variable value when all the independent variable is zero (Bremer, 2012).

The implementation of Multiple Linear Regression model is used package “library ggplot2”, this is used for display and configure the gg-plot. The application to conduct this implementation is RStudio, creating both plot and result from the lm function model.

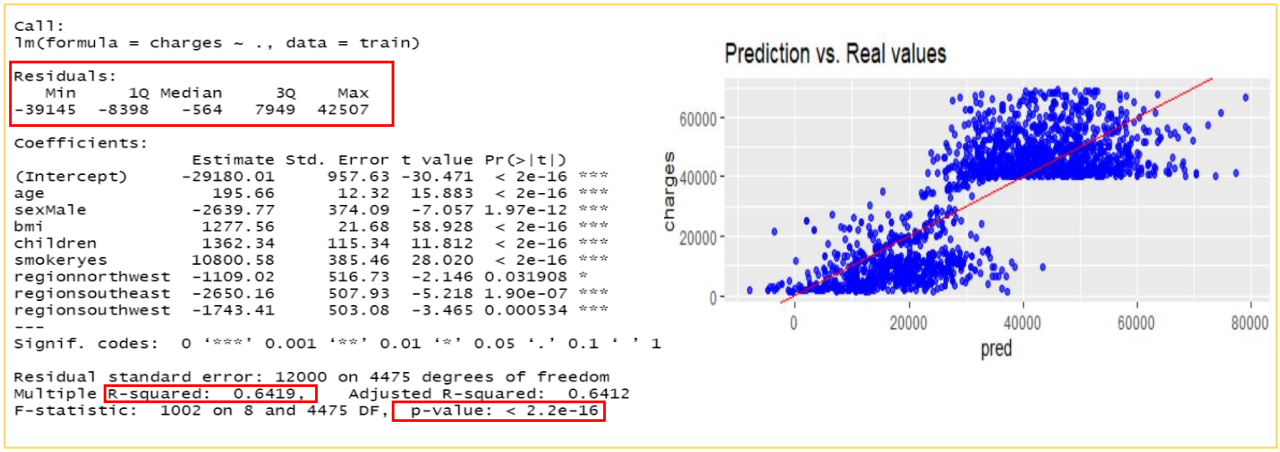


Figure 12 Multiple Linear Regression Model and Chart

Figure 12 above is the output of Multiple Linear Regression model and virtualisation display of predicted value and true value. Residuals mean the difference between the predicted data and the actual data, residual maximum and minimum error values range are quite large from minimum -39,145 to maximum 42,507. It illustrates that Multiple Linear Regression model is not ideal when predicting extreme values. However, it can be seen that the difference between 1Q, Median, 3Q(i.e. in ¼ value, ½ value, ¾ value) is not very large, As can see from the chart to judge the distribution of the entire value similar to the normal distribution, the centralized error is around -564.

Another key point is the multiple R-square, called R-square value, this model R-square value is at 64% which explains that linear prediction is suitable for approximately 64% of case success. Except Northwest region, the rest of the variables all have significant correlation with prediction variable charges at significance at a level lower than 0.05.

4.5.2 Implementation, Evaluation, and Result of Random Forest - Regression

Random Forest is a common and effective algorithm in the machine learning field, it is for regression and classification analysis and very focus on principles and code implementation process. The process of this section implementation main function of regression Random Forest.

The result of below figure 13 shows the R-square value is 80%. Which mean that measurement of 80% variance for dependent variables can be explained and predicted by independent variable, it illustrated that Random Forest model well predicted the factors related and impact America health insurance premium. Also, the plot chart shows the importance of variables impact insurance premium, the top three factors are BMI smoker and age.

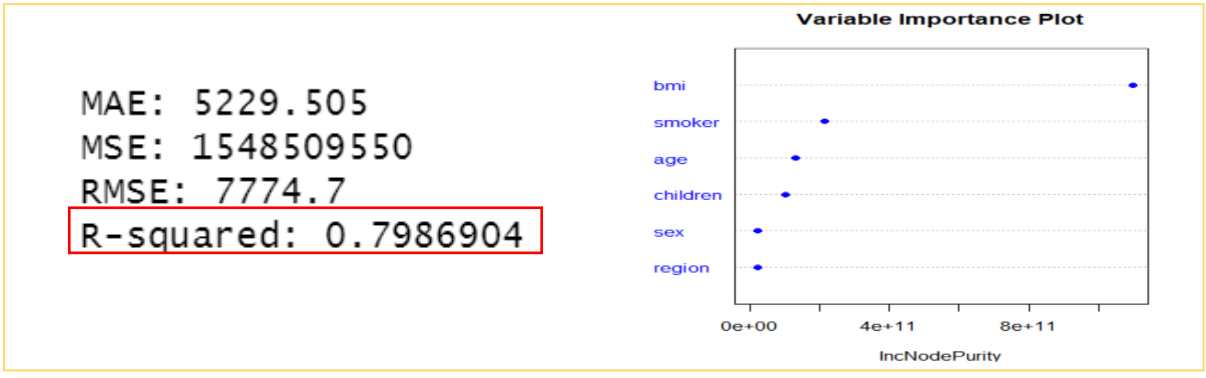


Figure 13 Random Forest Model (Regression) and Chart

4.5.3 Implementation, Evaluation, and Result of Support Vector Machine - Regression

Support Vector Machine (SVM) is supervised classification and regression prediction model, SVM regression main idea is to map data x to high feature space F, and to perform linear regression in this space through a nonlinear mapping. SVM can use to complete the classification by taking prediction research (SUNIL RAY, 2017).

R-square value is the square of the correlation, the SVM model represents 67% R-square value shows below figure 14, which means that 67% variance of dependent that can be explained and contributed for the independent variable, there is a correlation between the factors that can impact health insurance premium. The plot chart shows the prediction charges against real charges, red is the real charge value and blue colour is the predicted value, at the lower range of insurance cost that prediction is more accurate than the higher health insurance cost.

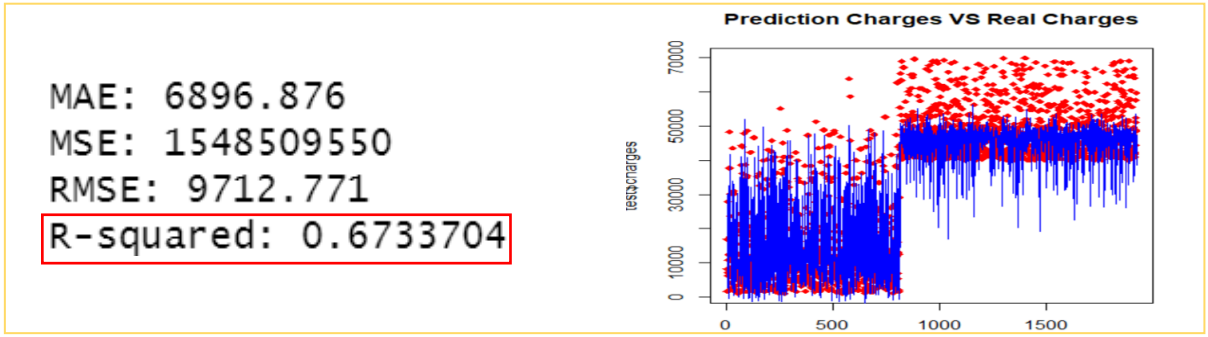


Figure 14 Support Vector Machine Model (Regression) and Chart

4.5.4 Implementation, Evaluation and Result of Naïve Bayes - Classification

Naïve Bayes is the idea of the simple Bayes algorithms, it is forecast the join probability distribution between the feature and the output, when predicting give feature, all possible outputs are extracted by Bayes theorem, the largest of them is taken as the prediction result, the advantage is that the model is simple, efficient and widely use in prediction purpose of insurance domain (Stijn , et al., 2004).

The implementation of the Naïve Bayes model is use package “library e1071”. The software for process this implementation is RStudio, confusion matrix calculates the output and accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  654 164
       yes 543 561

               Accuracy : 0.6322
                 95% CI : (0.6101, 0.6538)
    No Information Rate : 0.6228
    P-Value [Acc > NIR] : 0.2053

                  Kappa : 0.2902

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5464
            Specificity : 0.7738
         Pos Pred Value : 0.7995
         Neg Pred Value : 0.5082
             Prevalence : 0.6228
         Detection Rate : 0.3403
   Detection Prevalence : 0.4256
      Balanced Accuracy : 0.6601

       'Positive' Class : no
```

Figure 15 Naive Bayes Model

Figure 15 above is the output of Naïve Bayes model, the model purpose was to predicted new insurer smoker or non-smoker behaviours, the result of accuracy from this model is 63%, which means the measurement of 63% can explain and identify relationships between variables, P-value is lower the alpha value 0.05 represent the variables chosen were significant relate to the smoke condition, so with the relation independent variables that would assist insurance company identify insurer smoking behaviour in order to control the insurance premium.

4.5.5 Implementation, Evaluation and Result of Decision Tree - Classification

Decision Tree model is according to the tree structure based on the properties of data. Decision Tree model is often used to solve classification and regression problem and commonly use CART algorithms. Decision Tree is a machine learning predictive model that represents a mapping between object properties and object value, it is consist of connected nodes, each node in the tree specifies the judgment of object property values, its branches represent objects that meet the node criteria, each possible value and a line corresponds. The leaf node of the tree represents the prediction to which the object belongs. The high corresponds to the weight of the training set that reaches this node, higher weight is more reliable the distribution of class on each node (Prachi, 2019).

The implementation of the Decision Tree model is use package "library party", this display and configure the Decision tree image. The back-end software in RStudio conducts this implementation, result from Confusion Matrix providing the accuracy of Decision Tree model.

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  574   0
       yes 623 725

               Accuracy : 0.6759
                 95% CI : (0.6544, 0.6968)
    No Information Rate : 0.6228
    P-Value [Acc > NIR] : 7.016e-07

                  Kappa : 0.4101

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4795
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.5378
             Prevalence : 0.6228
         Detection Rate : 0.2986
   Detection Prevalence : 0.2986
      Balanced Accuracy : 0.7398

       'Positive' Class : no
```

Figure 16 Decision Tree Model

The output of the Decision Tree model is shown at figure 16, P-value is lower than alpha value 0.05, which means that the selected variables has a significant correlation with insurer's smoker status. This model accuracy output is 68%, which can explain the variable relationship and model performance.

Figure 17 below represents Decision Tree visualisation. The node begins with the premium charges. From the lower charges that age is a factor relate to smoker status, at middle level of charges that BMI become more significant relate to smoker status, at the highest premium that sex is very different from smoker behaviour, which female smoker more than male (use percentage of smokers multiply by the amount of each sex). This chart can show that the insurer smoke behaviour can be charged differentiated with their health insurance premium.

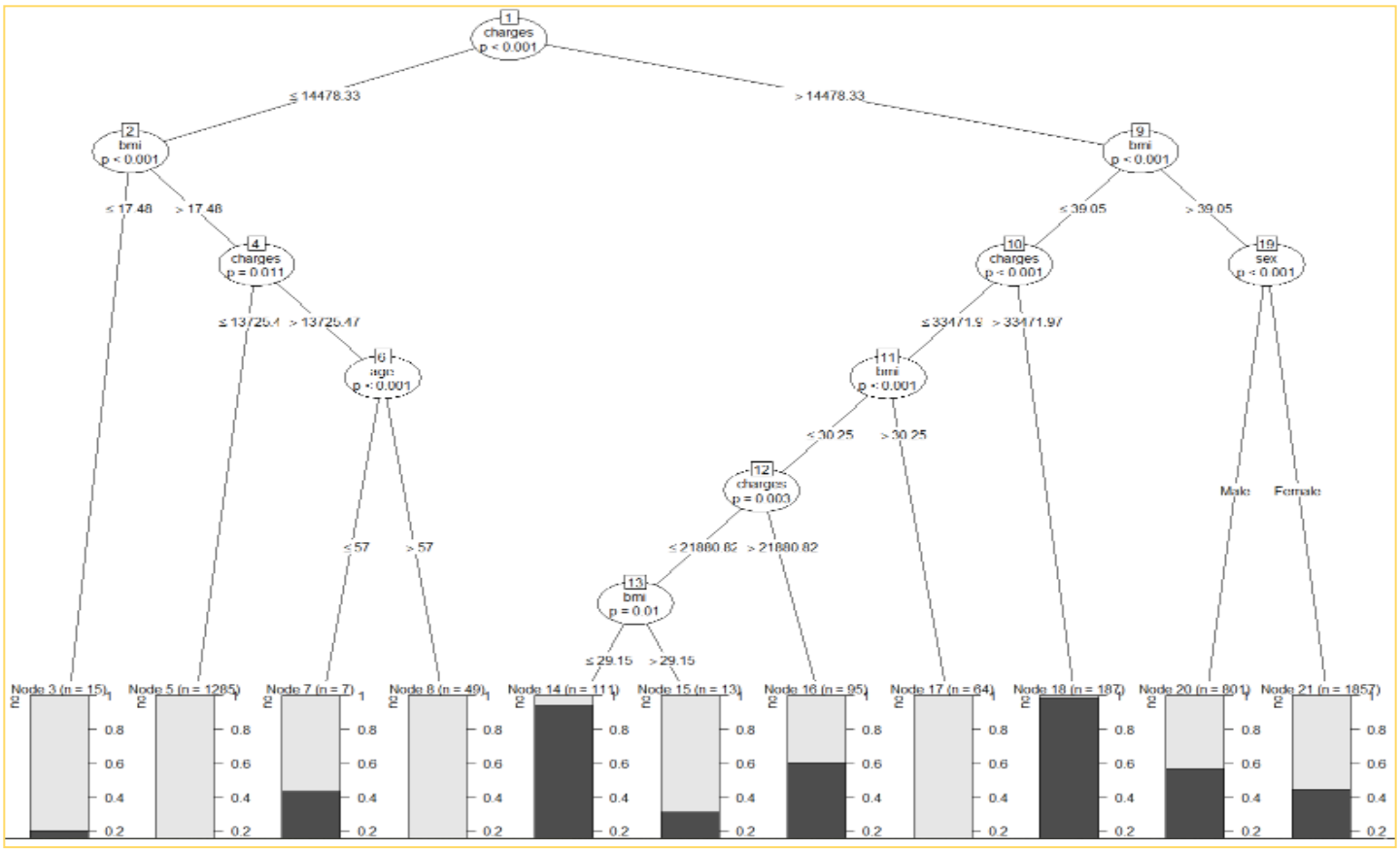


Figure 17 Decision Tree Model Chart

4.5.6 Implementation, Evaluation, and Result of Logistic Regression - Classification

Logistic Regression is a classification model and is often used for two classifications, Logistic Regression is popular with industry because of its simplicity, parallelization and explain ability. The essence of Logistic Regression is to assume that the data obeys this distribution and then use a very plausible estimate to make a parameter prediction. When a binary result variable is predicted by a series of continuous and categorical predictors (Alzen, et al., 2018).

The implementation of the Logistic Regression model did not use any package. The software applied for this implementation in RStudio, the Accuracy output as a result.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1662  -0.9302  -0.4427   1.0914   1.8904

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.322e+00  2.142e-01  -6.171 6.79e-10 ***
age             -8.470e-03  2.405e-03  -3.522 0.000428 ***
sexMale          5.050e-01  7.383e-02   6.840 7.93e-12 ***
bmi             -4.523e-02  5.879e-03  -7.695 1.42e-14 ***
children        -2.284e-02  2.179e-02  -1.048 0.294682
regionnorthwest  9.796e-02  9.956e-02   0.984 0.325178
regionsoutheast  1.706e-01  9.783e-02   1.744 0.081118 .
regionsouthwest -1.233e-02  9.944e-02  -0.124 0.901289
charges          7.396e-05  3.169e-06  23.338  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  959 408
       yes 238 317

               Accuracy : 0.6639
                 95% CI : (0.6423, 0.685)
    No Information Rate : 0.6228
    P-Value [Acc > NIR] : 9.912e-05

                  Kappa : 0.25

 Mcnemar's Test P-Value : 2.947e-11

            Sensitivity : 0.8012
            Specificity : 0.4372
         Pos Pred Value : 0.7015
         Neg Pred Value : 0.5712
             Prevalence : 0.6228
         Detection Rate : 0.4990
   Detection Prevalence : 0.7112
      Balanced Accuracy : 0.6192

       'Positive' Class : no
```

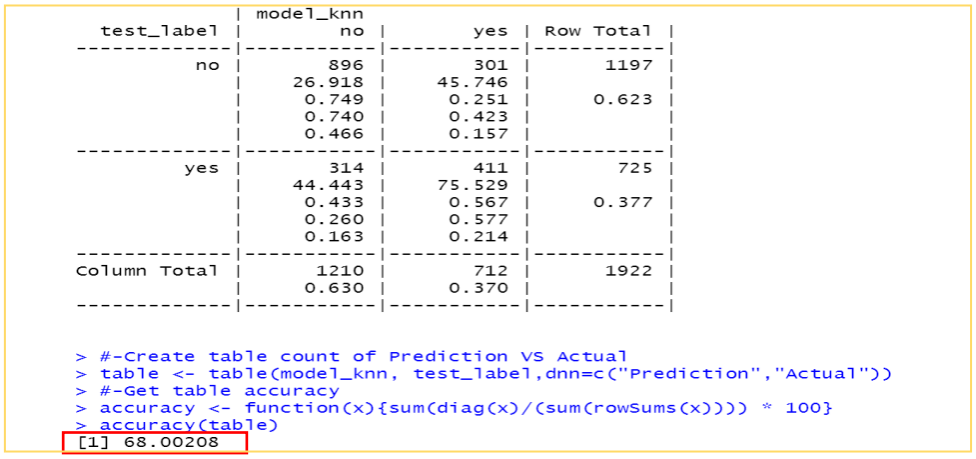
Figure 18 Logistic Regression Model

There are two output from Logistic Regression Model shows in figure 18, the left part shows residual level from minimum and maximum is between -2.1662 to 1.8904, this range is not quite large, also shows the coefficients of age, sex, BMI and charges has significant correlation with smoker status. The right part output from confusion matrix shows the model accuracy is 66%.

4.5.7 Implementation, Evaluation, and Result of K-Nearest Neighbour - Classification

K-Nearest Neighbour is the simplest and most effective classification algorithm for classifying data. Many datasets and infrastructure information that cannot be given, and no way to know what characteristics the average and typical instance sample have. In most of samples in which K is most similar in feature space (i.e. the closest to the feature space) belong to a certain category, the sample also falls into this category. It means that the nearest K sample sits to what kind of class it belong to, and determines that it belongs to the same category (Peterson, 2009).

The implementation of the K-Nearest Neighbour model is use package "library class" for the model. The software conducts this implementation in RStudio, the result for accuracy is from confusion matrix.



test_label	model_knn: no	yes	Row Total
no	896	301	1197
	26.918	45.746	
	0.749	0.251	0.623
	0.740	0.423	
	0.466	0.157	
yes	314	411	725
	44.443	75.529	
	0.433	0.567	0.377
	0.260	0.577	
	0.163	0.214	
Column Total	1210	712	1922
	0.630	0.370	

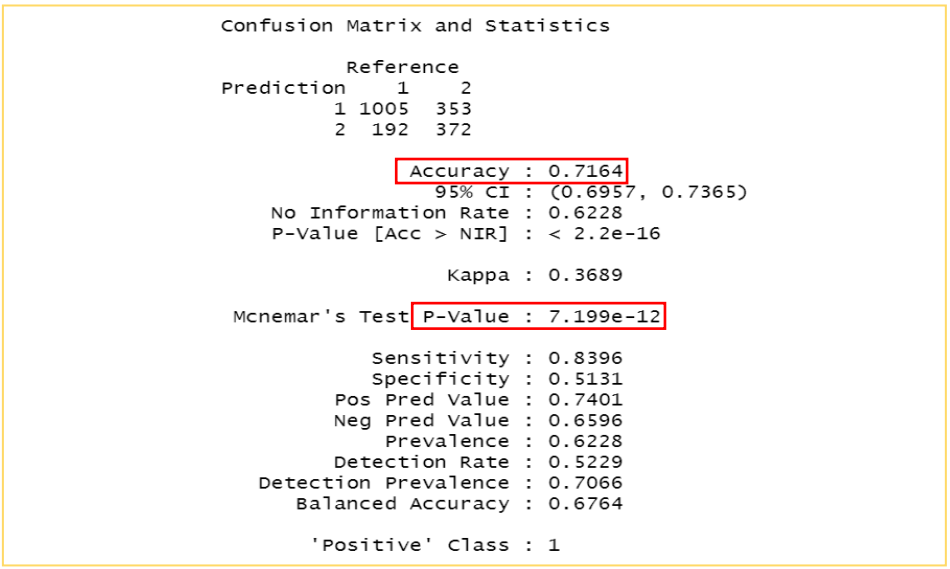
```
> #-Create table count of Prediction VS Actual
> table <- table(model_knn, test_label,dnn=c("Prediction","Actual"))
> #-Get table accuracy
> accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
> accuracy(table)
[1] 68.00208
```

Figure 19 K-Nearest Neighbour Model

Figure 19 used confusion matrix shows the accuracy of this model is 68%, it means the portion of 68% can explain the prediction and real value, which shows the model provide good prediction performance to estimate insurer smoker behaviours relate to other selected variables.

4.5.8 Implementation, Evaluation, and Result of SVM - Classification

This implementation of the Support Vector Machine model is use package "library e1071", the model is for classification purpose and software used for conducting this implementation is done in RStudio, the result accuracy is from Confusion Matrix.



```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 1005  353
         2  192  372

               Accuracy : 0.7164
                 95% CI : (0.6957, 0.7365)
    No Information Rate : 0.6228
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3689

 Mcnemar's Test P-Value : 7.199e-12

            Sensitivity : 0.8396
            Specificity : 0.5131
         Pos Pred Value : 0.7401
         Neg Pred Value : 0.6596
             Prevalence : 0.6228
         Detection Rate : 0.5229
   Detection Prevalence : 0.7066
      Balanced Accuracy : 0.6764

       'Positive' Class : 1
```

Figure 20 Support Vector Machine Model

The SVM model output from figure 20 shows the accuracy is 72%, which means that 72% can explain this model to predict smoker status with other selected variables. P-value is below the alpha value can also interpreted that there is significant correlation between smoker status and other variables.

4.5.9 Implementation, Evaluation and Result of Random Forest - Classification

The implementation of the Random Forest model is used package “library randomForest. Used RStudio generated implementation, the accuracy result is from Confusion Matrix method.

The output of the Random Forest model used confusion matrix to demonstrate at figure 21, the accuracy of model is 71%, which means the portion of 71% accuracy would explain the model for predict analysis smoker status with other selected variables. P-value is very low can also interpret there is significant correlation between smoker status and other variables.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 950 314
         2 247 411

               Accuracy : 0.7081
                 95% CI : (0.6872, 0.7284)
    No Information Rate : 0.6228
    P-Value [Acc > NIR] : 2.389e-15

                  Kappa : 0.3672

 Mcnemar's Test P-Value : 0.005328

            Sensitivity : 0.7937
            Specificity : 0.5669
         Pos Pred Value : 0.7516
         Neg Pred Value : 0.6246
             Prevalence : 0.6228
         Detection Rate : 0.4943
   Detection Prevalence : 0.6576
      Balanced Accuracy : 0.6803

       'Positive' Class : 1
```

Figure 21 Random Forest Model

4.6 Comparison of Developed Models

Table 6 compares regression models output with literature review result used R-square, ANOVA model also generated R-square used statistical analysis. Top 4 out of 6 classification models from Table 7 compares with literature review result used accuracy. Also, generated accuracy output from Support Vector Machine and Random Forest classification model.

Table 6 Regression Algorithms Comparation Result

Regression Model	R-square (Literature Review)	R-square (This Project)
Multiple Linear Regression	40%	64%
Random Forest	40%	80%
Support Vector Machine	41%	67%
ANOVA	-	64%

Table 7 Classification Algorithms Comparation Result

Classification Model	Accuracy (Literature Review)	Accuracy (This Project)
Naïve Bayes	67% or 70%	63%
Decision Tree	80.86% or 77.8%	68%
Logistic Regression	56.5%	66%
K-Nearest Neighbour	75%	68%
Support Vector Machine	-	72%
Random Forest	-	71%

All models represent good results to identify the research question and sub-research question, the output of Random Forest (Regression) model shows the best result to explain the research question, and Support Vector Machine (classification) model shows the best result to solve the sub-research question.

5 Conclusion and Future Work

Insurance companies usually sale their insurance product based on customer demand, the fundamental is to have a better understanding about customer behaviour, this would support insurance companies' growth. This project focuses on insurer's different behaviours and attributes, use machine learning models and statistic model to assist insurance companies analyse the correlation with insurance premium and insurer's characters.

Through the result of different models that can response the research question of this project " To what extent can aspect from data mining analysis of American health insurance industry used supervised machine learning techniques(Multiple Linear Regression, Support Vector Machine, Random Forest)and statistic model (ANOVA) to deliver meaningful insights for health insurance premium base from an individual's behaviour and predicted contributed factors?" The process was conducted by three regression models to determine the result, from the output result demonstrate that Random Forest model provide the best achievement (80% R-square) and successfully addressed this research question, Support Vector Machine Regression model was 2nd performance solved the research question, but SVM Classification model deliver the best performance to solve the sub-research question, this finding illustrates the same model would have various performances utilizing classification and regression approach.

In order to define the health insurance premium correlation with insurer's behaviours and health habits, based on the discovery and study of this research project, identified that predictive model has great advantage and performance applied on health insurance field after achieved research questions and 4 objectives, health insurance organisation can precisely account the insurance premium according to different individual's attribute.

This will assist insurer get the appropriate amount for their health insurance premium, also it would build trust level relationships between insurance organisations and insurers, which will make insurers continue pay their insurance premium. Through these predictive analyses, health insurance companies would deliver and evaluate the following decision to make an assumption.

- Which insurance plan would suit and desire for different a kind of insurer?
- How much insurance premium should be charged according to personal attributes and behaviours?
- How to reduce risk management regarding to insurer's behaviours?
- How to build good trust relationship between insurance companies and insurers?

Thus, it is very useful information for health insurance companies to collect and analysis insurer's attributes. Insurance companies require a data source from internal and external, these data are relevant to the breakdown of the insurance market. Through depth analysis of customer information and customer behaviour, will predict and understand more customer demanding and recommend suitable products with achievable price, to achieve a distinctive personality precision marketing.

Therefore, there are still some limitations scope for further research, for example, data from insurer's medical and other sickness condition, insurer's previous health insurance claim data

information which is not included in this project, will be good additional part toward future research, assisting insurance companies to have better insights from different individuals behaviours and attribute, also giving more accurately forecast on health insurance premium charges and risk management.

Acknowledge

I would especially thank my Supervisor Dr. Catherine Mulwa supported and guided me through this project. I would also like to acknowledge my husband and my kids who supported me and with me in this pandemic period.

References

Alzen, J. L., Langdon , . L. S. & O, V. K., 2018. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. 28 December.

Ahmed, A. M., Rizaner, A. & Ulusoy, A. H., 2018. A Decision Tree Algorithm Combined with Linear Regression for Data Classification.

Attewell, P., Monaghan, D. B. & Kwong, D., 2015. *Data Mining for the Social Sciences.* s.l.:University of California Press.

Azevedo, A. & Santos, M. F., 2008. Based on data mining algorithms and science analysis. *ResearchGate.*

Bao, Y., Xiong, T. & Hu, Z., 2013. *Multi-step-ahead time series predicton using multiple-output support vector regression,* s.l.: Neurocomputing.

Berchhick, E. R., Barnett, J. C. & Upton, R. D., 2019. *Health Insurance Coverage in the United States:2018,* Washington: United States Census.

Bremer, M., 2012. *Math 261A - Spring 2012,* s.l.: s.n.

Burns, E. et al., 2019. *Considerations for Predictive Modeling in Insurance Applications,* s.l.: Society of Actuaries.

Ch. Ratanamahatana, D. G., 2002. *"Scaling up the Naive Bayesian classifier: using decision trees for features selection," in Proceedings ofWorkshop on Data Cleaning and Preprocessing.* s.l.:at IEEE International Conference on Data Mining (ICDM'02), M.

Couronné,, R., Probst , . P. & Boul, . A.-L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. 17 July.

Dey, L., Meisheri, H. & Verma, I., 2017. Predictive Analytics with Structured and Unstructured data - A Deep Learning based Approach. December.

Duan, Z. et al., 2018. A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform. *International Journal of Financial Studies*, 7 Feb.

Fang. R, P. S. Y. Y. C. C., 2016. Computational health informatics in the big data age:. *a survey ACM Comput Surv,* Volume 49(1), pp. pp. 1-36.

Frees, E. W. (., May, 2013. *Predictive MOdeling of Insurance Company Operations,* s.l.: University of Wisconsin-Madison.

Frees, E. W., Derrig, R. A. & Meyers, G., 2016. *Predictive Modeling Aplications in Actuarial Science Volume ||: Case Studies in Insurance.* s.l.:Cambridge University Press.

Freyder, C. W., 2015. *SING LINEAR REGRESSION AND MIXED MODELS TO PREDICT HEALTH CARE COSTS AFTER AN INPATIENT EVENT,* s.l.: West Virginia University.

Gepp, A. W. J. H. K. K. &. B. S., 2012 . *A comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection.* s.l.: Journal of data science, 10, 537–561. .

Goldstone, J. A., 2008. *Using Quantitative and Qualitative Models to Forecast Instability,* Washington: United States Instittute of Peace.

Goleiji L, T. M., 2015. *Identification of influential features and fraud detection in the Insurance Industry using the data mining techniques (Case study: automobile's body insurance).*. s.l.:Majlesi J Multimed Process 4:1–5.

H. Zhang, L. J. J. S., 2005. Hidden Naiıve Bayes,. In: *in Proceedings of theTwentieth National Conference on Artificial Intelligence.* s.l.:s.n., p. pp. 919–924.

Hand, D. & A.M.Herzberg, 2005. *Selected Statistical Papers of Sir David Cox.* s.l.:s.n.

Harrison, O., 2018. Machine learning basics with the K-Nearest Neighbors algorithm. *Towards data science,* 10 Sep.

Hickey, S. J., 2013. Naive Bayes Classification of Public Health Data. *Communications of IIMA,* 13(2).

Hota, H. & Dewangan, S., Sep 2016. *Classification of Health Care Data Using Machine Learning Technique,* India: Bilaspur University .

Kim, D. D. & Basu, A., 2016. Estimating the medical care costs of obesity iin the United States. In: *systematic review, meta-analysis, and empirical analysis.*. s.l.:s.n., pp. 602-614.

Kim, Y. J. & Park, H., 2019. Improving Prediction of High-Cost Health Care Users with Medical Check-Up Data. *Big Data,* Big DataVol. 7, No. 3(https://doi.org/10.1089/big.2018.0096).

Kismiantini, D. W. A., 2018. *Analysis of Factors Affecting the Health Insurance Ownership with Binary Logistic Regression Model,* Indonesia: Yogyakarta State University.

Kuo, R. N. et al., 2011. Predicting Healthcare Utilization Using a Pharmacy-based Metric With the WHO's Anatomic Therapeutic Chemical Algorithm. *Medical Care,* pp. 1031-1039.

Lahirih, C. & Agarwal, N., 2014. *Predicting healthcare expenditure increase for an individual from medicare data.,* s.l.: Proceedings of the ACMSIGKDD Workshop on Health Informatics..

Lip GY, N. R. P. R. L. D. C. H., 2010. *Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation.* s.l.:US national Library of Medicine National Institutes of Health.

McDonald, J. B., 1993. *Predicting Insurance Insolvency Using Generalized Qualitative Response Models,* s.l.: Worker's Compensation Insurance.

Michael Ewald, Q. W., 2015. *Predictive Modeling: A Modeler's Introspection.* s.l.:Society of Actuaries.

O'Boyle, F., 2019. *Classical Time-Series vs Machine Learning Methods.* [Online]
Available at: https://towardsdatascience.com/classical-time-series-vs-machine-learning-methods-80290850bd5b

Ohlsson, E. &. J. B., 2010. *Non-life insurance pricing with generalized linear models.*. Berlin: Springer.10.1007/978-3-642-10791-7 .

Peterson, L. E., 2009. K-nearest neighbor. *Scholarpedia,* p. 4(2):1883..

Prachi, M., 2019. *Decision Tree Analysis.* [Online]
Available at: https://theinvestorsbook.com/decision-tree-analysis.html

Siegel, E., 2016. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Revised and Updated,,* s.l.: s.n.

Sivagnanam, K. J. & Srinivasan, R., 2010. *Business Economics.* s.l.:Tata McGraw Hill Education Private Limited.

S, S. et al., 2015. *Population cost prediction on public healthcare datasets.* s.l., Proceedings of the 5th international conference on digital health, pp. 1312-1321.

Stijn , V., Derrig , R. A. & Dedene, G., 2004. A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. *IEEE Transactions on Knowledge and Data Engineering,* pp. pp. 612-620, vol. 16.

SUNIL RAY, 2017. Understanding Support Vector Machine(SVM) algorithm from examples (along with code). *Analytics Vidhya,* 13 SEPTEMBER.

Tian, Y. S. Y. &. L. X., 2012. *Recent advances on support vector machines research. ,* s.l.: Technological and Economic Development of Economy.

Wang, L., Li, X., Cao, C. & Yuan, S., 2006. Combining decision tree and Naive Bayes for classification. 13 10.

Watson, K., 2019. Predictive analytics in health care Emerging value and risks. *Deloitte Insights,* 19 July.

Weng, S., 2018. *Introducing machine learning for healthcare research,* s.l.: University of Nottingham.

Wu G, C. E., 2003. *Class-Boundary Alignment for Imbalanced Dataset Learning,* s.l.: Department of Electrical & Computer Engineering, University of California, Santa Barbar.

Xu Jianhua, Z. X. ,. L. Y., 2004. Advances in support vector machines[J]. . In: *Control and Decision.* s.l.:s.n., pp. 19 (5) : 481 - 493..