Faculty of Graduate Studies

Master Program in Applied Statistics and Data Science

The Role of Machine Learning in Health Insurance Industry- a case study

Prepared by
Ruba Rishmawi
Student number:1195135

Supervisor
Dr. Hassan Abu Hassan

Submitted in partial fulfillment of the requirements for the "Master Degree in Applied Statistics and Data Science" from the faculty of Graduate Studies at Birzeit University-Palestine

Feb 2022

The Role of machine learning in health insurance industry- a case study

Prepared by:
Ruba Rishmawi
Student number:1195135

Committee:

Name	Signature
Dr. Hassan Abu Hassan	
Dr. Radi Jarrar	
Dr. Weeam Hammoudeh	

Submitted in partial fulfillment of the requirements for the "Master Degree in Applied Statistics and Data Science" from the faculty of Graduate Studies at Birzeit University-Palestine

Feb 2022

Acknowledgments

I would like to express my sincere gratitude to my supervisor Dr. Hassan Abu Hassan for his continuous support, consultancy, and guidance that have helped me extremely during my research and writing of the thesis. His massive knowledge and experience have enabled me to complete this research.

I would also like to thank my family and friends who gave me unconditional support during my research and helped me in finalizing this research within the time frame.

Table of Contents

List of Figures

List of Tables

Abstract

Insurance companies worldwide are exploring how machine learning (ML) can improve customer satisfaction, reduce operational costs and increase profitability.
The greatest opportunity lies in risk prediction and fraud detection, whereas these two topics are the most challenging in the Insurance domain.

The first goal of the study is to predict the risk level in Health Insurance subscribers using machine learning Classification algorithms. The study included around 10 thousand subscribers who were classified on 3 risk levels (High-risk, Mid-risk, and Low-risk). The main objective is to predict accurately the risk level and accordingly assist the company in providing an accurate premium rate for new customers.
Some supervised classification algorithms were analyzed and compared using R statistical programming Language; the Random Forest algorithm had the highest accuracy which was around 94%, sensitivity for the High-risk level was around 72%, and 81% for the Mid-risk level.

The second goal of the study is to detect outliers in medical claims to help predict which claims are suspects of fraudulent activities and assist insurance specialists to investigate and discover cases of fraud.
Unsupervised Outlier Detection algorithms assist in discovering abnormal behavior in medical pattern claims and discover cases of fraud. Three outlier detection techniques were compared, the first using Clustering algorithm (PAM), the second using Density-based local outliers (LOF), and the third using Isolation Forests (IF).

Outliers with abnormal behavior were detected, some had very high costs compared to the clusters/neighbors they belong to, and others had a very small-time interval between clinical visits which may be suspicious behavior.
Claims audits consume lots of time and are very costly for insurance companies, so having machine learning algorithms detect the suspicious claims that require review will save the company huge amounts of operational costs and will increase work efficiency.

الملخص

تقوم شركات التأمين في جميع أنحاء العالم بتطوير منتجات باستخدام كميات كبيرة من البيانات لتقييم المخاطر وتحديدها والتنبؤ بها، وتستكشف كيف يمكن تطبيق أنظمة قائمة على الذكاء الصناعي وخوارزميات التعلم الآلي بهدف تحسين رضا العملاء وتقليل التكاليف التشغيلية وزيادة الربحية، ويعد التنبؤ بالمخاطر والكشف عن الاحتيال من أكثر المواضيع تحديا في مجال التأمين.

الهدف الأول من الدراسة هو استخدام التحليلات التنبؤية للتنبؤ بمستوى المخاطر لدى مشتركي التأمين الصحي باستخدام خوارزميات التعلم الآلي. اشتملت الدراسة على حوالي 10 آلاف مشترك تم تصنيفهم على 3 مستويات مخاطر (عالية المخاطر، متوسطة المخاطر، منخفضة المخاطر). الهدف الرئيسي هو التنبؤ بدقة بمستوى المخاطر وبالتالي مساعدة شركة التأمين في توفير معدل أقساط تأمين متناسبة مع المخاطر للعملاء الجدد.
تم تحليل ومقارنة بعض خوارزميات التصنيف Classification algorithms باستخدام لغة البرمجة الإحصائية R . و حصلت خوارزمية Random Forest على أعلى دقة حيث بلغت حوالي 94٪ ، وكانت الحساسية sensitivity لمستوى المخاطر العالية حوالي 72٪ ، و81٪ لمستوى المخاطر المتوسطة.

الهدف الثاني من الدراسة هو اكتشاف القيم المتطرفة والشاذة في المطالبات الطبية للمساعدة في الكشف عن الاحتيال باستخدام خوارزميات التعلم الآلي ومساعدة المتخصصين في التأمين على التحقيق واكتشاف المطالبات الطبية المشتبه في ارتكابها أنشطة احتيالية.
تساعد خوارزميات التعلم الآلي الخاصة بالكشف عن القيم الشاذة Outlier Detection في اكتشاف السلوك والنمط غير الطبيعي في المطالبات الطبية بهدف اكتشاف حالات الاحتيال. تمت مقارنة ثلاث تقنيات (PAM) Clustering، (LOF)، و(IF). تم الكشف عن مجموعة من القيم المتطرفة ذات السلوك أو النمط غير الطبيعي ، وكان لبعضها تكاليف عالية جدًا مقارنة بالمجموعات التي ينتمون إليهم ، وكان لدى البعض الآخر فترة زمنية قصيرة جدًا بين الزيارات الطبية والتي قد تكون سلوكًا مشبوهًا.
تستهلك عمليات تدقيق المطالبات الكثير من الوقت والجهد وهي مكلفة للغاية لشركات التأمين ، لذا فإن استخدام خوارزميات التعلم الآلي في اكتشاف المطالبات المشتبه بارتكابها أنشطة احتيالية ليتم مراجعتها من قبل الخبراء سيوفر للشركة مبالغ ضخمة من التكاليف التشغيلية حيث سيتم التركيز على المطالبات الأكثر شبهة للاحتيال عوضاً عن مراجعة وتدقيق كميات كبيرة من المطالبات الطبية.

Chapter one

Introduction

1.1 Background

The insurance industry is considered one of the most challenging business domains, as it is highly related to risk, so it has always been dependent on statistical analysis. Nowadays, insurance companies have large amounts of data, and therefore, they are in need of using machine learning techniques. These techniques help in discovering patterns and enhancing business decisions.
Predictive modeling in health insurance has been gaining more interest and it has been proven that it provided higher quality products, gave a competitive advantage, and led to gain sustainable growth and optimize services. (Positive Impact of Machine Learning in the Insurance Industry, 2021)

Machine learning algorithms identify patterns and predict how likely an event is to occur based on historical data; thus, they can be applied in various areas of insurance such as risk assessment, marketing analytics, claims analysis, fraud detection, and others. The main objectives of applying machine learning algorithms are to reduce time & cost, eliminate fraud, adjust policies, and accordingly improve customer experience. (Yadav, 2021)

The most common use cases of machine learning in the health insurance industry fall into these two main categories:

- **Risk assessment:** Companies use machine learning algorithms to detect potential risks and then make adjustments to the premium rates. Setting the right premium rate will help the insurance company obtain new customers and decrease the loss from risky customers (Morse, 2021)
  Risk classification is very common in insurance companies, where customers are grouped according to their estimated level of risk, this is usually done by an underwriter whose job is to evaluate risk and set the premium rates accordingly. However, this process is time & cost-consuming. Hence, to improve the underwriting process machine learning algorithms can classify the risk level based on the available data and accordingly recommend the most adequate premium rate (Boodhun & Jayabalan, 2018)

- **Fraud Detection:** Insurance fraud is a very critical problem for insurance companies, and detecting fraud is very important to reduce costs and increase profits.
  Unfortunately, predicting fraud customers is not an easy task because it is very hard to have a training sample of frauds and use it in the classification methods such as logistic regression.
  Unsupervised learning will assist in discovering abnormal behavior in medical claims' patterns and will help the insurance specialists to investigate and discover cases of fraud.

  Medical fraud claims are classified but not limited to 5 categories: (Pai, Agnihotri, Rajath, & Kumar Jha, 2016)
    - Bills are costlier than they are supposed to be for the medical service.
    - Unnecessary services: claims for services that are not necessary for the medical condition, for example, unnecessary laboratory tests or medicines.
    - Duplicate claims: Have duplication in medical claims within certain time intervals.
    - Using other clients' coverage
    - Filing claims that were not actually received

Claims audits consume lots of time and are very costly for insurance companies, so having machine learning algorithms detect the suspicious claims that require review and that are suspect of abuse or fraud will save the company huge amounts of operational costs and will increase efficiency and customer satisfaction (Mckinsey&Company, 2017)

This thesis intends to apply machine learning algorithms to a dataset of medical insurance subscribers and their medical claims from an insurance company located in Palestine. The main objective is to provide output that assists the company in profiling their subscribers and improving the underwriting policies in addition to detecting outliers in medical claims that may be suspect for customers and medical providers' fraud.

1.2 Problem statement and importance of the study

The health insurance policies in Palestine depend on the principle of risk pooling, which means that costs of risk are shared between individuals in the same group/company, this leads to the reduction of burdens to which the insured person may be exposed to (Health Insurance, n.d.). On the other hand, insurance companies should have high accuracy in predicting the risks in each account to avoid losses caused by health insurance policies.

In this case, a study that includes a sample of health insured subscribers in a Palestinian Insurance company, 23% of subscribers who are considered risky and non-profitable

contribute 67% of medical claims' cost, so predicting the risky subscribers based on their characteristics will assist the policy-makers in the underwriting processes and provide the right premium rate, which will eventually lead to gaining more profit, and this illustrates the importance of using machine learning predictive models to offer accurate premium rates, and to reduce the risk of losing revenues.

As for detecting fraud and abusers, it is so important for insurance companies to find new methods for detecting fraudulent claims, because relying on business rules and manual investigation is very costly and time-consuming, so machine learning studies in this field are giving a new opportunity for companies to reinvent their fraud detection methods. A study shows that a Dutch household pays on average 100 euros extra to compensate for the fraud (Blanken, 2017). The Department of Justice in the United States reports that fraud costs the health insurance industry over 100 billion dollars per year (Sennaar, 2019)

This illustrates the importance of such studies, where recommendations and output will encourage Palestinian Insurance companies to apply machine learning models to enhance their processes and increase their profits.

1.3 Research objectives

The objective of the study is to utilize machine learning algorithms in the health insurance domain to provide accurate predictions and help the company in understanding the patterns and behavior of its subscribers.

In this way, human resources and costs consumed by the operational process and manual inspection can be reduced, which will alternately assist the company in gaining a competitive business advantage and increase its profits.

The thesis will cover 2 areas of study:

- Predictive models for risk assessment. The models will assist the policy-makers in the underwriting processes and provide the right premium rate, which will eventually lead to gaining more profit.
- Fraud detection: The model will help detect suspicious claims that require review and audit, and accordingly will save cost and time and increase fraud detection accuracy.

In addition to the predictive modeling, some exploratory analysis and clustering techniques were applied to help the company understand the characteristics of each segment and help in setting new policies and attracting new customers, also this will

assist the company in marketing campaigns, cross-selling, and up-selling of new products.

In conclusion, the aim of applying machine learning models in insurance is the same as all other industries that are moving towards applying machine learning models to improve processes, make real-time decisions, optimize marketing strategies and of course, increase revenues.

1.4 Definition and explanation of key terminology

- Health insurance: The health insurance policies provide medical and health care for individuals in groups, companies, and foundations operating in Palestine which include medical examination costs, diagnosis, treatment, and physical and psychological support (Health Insurance, n.d.)

- Insurance underwriting: is the process of assessing the risk when setting the prices of insuring a home, car, driver, or an individual's health or life. It determines whether it would be profitable for an insurance company to take a chance on providing insurance coverage to an individual or business and accordingly define prices and coverages.

- The insurance premium is the cost of the insurance.

- Inpatient care requires overnight hospitalization. Patients must stay at the medical facility where their procedure was done (which is usually a hospital) for at least one night, while Outpatient care doesn’t require overnight hospitalization

1.5 Limitations of the study

Some limitations were faced in applying the machine learning models such as:

- Not all desired variables were available in the insurance company, since some variables that are mentioned in the literature review were not stored in the company’s databases and could not be retrieved.
- Most of the used files in preparing the dataset were disorganized which consumed lots of effort in aggregating and grouping the data.
- This problem occurred because the files were provided in CSV formats and the researcher didn’t have direct access to the company’s databases, so data processing was done manually using MS Excel, MS Access, and R programming language, whereas having direct access to the databases could have saved time and efforts by writing simple SQL statements.

- Lack of labeled data to be used for machine learning algorithms.
- The provided datasets didn’t have labeled classes for the risk-level of each customer to be used as the dependent variable, so the researcher had to create a new variable depending on the associated costs and grouped the customers on 3 risk levels < High-risk, Mid-risk & Low-risk>
- Faced some computational power limitations; since some machine learning algorithms require high computational power and due to using a personal laptop, the performance of some algorithms was very slow or even couldn’t converge.

1.6 Research ethics

Prior to the initiation of the study, official written approval was obtained from the CEO of the insurance company to use the data provided for research purposes. All data regarding the privacy of the subscribers were removed before beginning the analysis such as name and ID. The researcher is committed not to misusing or sharing the data.
The company’s official letter is stated in Appendix1

Chapter two

Literature review

Insurance companies have been trying to improve the efficiency of their products, so data mining has become more popular in this domain nowadays. Due to the increase of electronic health records, data is easier to collect and analyze, where healthcare professionals are concentrating on maximizing the efficiency of using data mining in their organizations (Pooja & Jagadeesh , 2019)

Profiling clients and analyzing their behavior can reveal valuable information, it wouldn't only help in segmenting current customers and knowing their characteristics, it would also help in acquisition plans and in creating new products that suit low-risk customers, and will assist in cross-selling and up-selling products. Applying clustering techniques based on customers' behavior will lead to targeted marketing. (Abdul-Rahman, Arifin, Hanafiah, & Mutalib, 2021)

A cluster sampling study was performed to identify the characteristics of high-cost patients and the determinants of the annual medical expenditures of Chinese rural residents. The analysis objectives were to reveal abusers and assign customers to high-cost, moderate-cost, and low-cost (others) groups based on their annual medical expenditures. Age, disease category, inpatient status, healthcare utilization, and utilization level were identified as the determinants of annual medical expenditures, and the study concluded that the medical expenditures of rural residents are clustered at a remarkably high level. Policy-makers shall guide these high-cost segments and manage their utilization of the unnecessary healthcare actions (Zhang, Lu, Niu, & Zhang, 2018)

Customer segmentation assists in offering customized offers and unique customer experiences, in another study on insurance claims data the researcher used different clustering techniques to segment the customer such as Partitioning Algorithms, Hierarchical Algorithms, Density-Based Clustering Methods, and K-means. The dataset was composed of over 800 attributes, a variable selection process was implemented to simplify the model. 11 clusters were the best output of the clustering techniques, the clusters were then profiled based on company and basic demographic KPIs such as Claim Ratio, Age Distribution, Gender Distribution, Relationship with the policyholder, and profitability. The use of the clusters was to fine-tune the marketing approach for customer relationship management and use the knowledge for retention and acquisition campaigns (Zaqueu, 2019)

In another study done by Bücker(2016) the researcher used demographic attributes, tenure with the company, and preferred channel of communication. The objective was to

understand the clusters in order to offer them the most appropriate promotion through the best channel. 6 clusters were the best model. Another segmentation was applied which aimed to understand the behavior of the clients, the variables in the model were relevant to the behavior of clients the previous period such as number of months since last purchase, number of active policies, number of contracts as an insured person, and yearly increase in active product families. The objective of this clustering model is to identify the loyal customers from others who are about to churn or aren't active. Another segmentation was value segmentation where customers were clustered based on policy costs and profit and the objective was to understand the characteristics of profitable and non-profitable segments. (Bücker, 2016)

Cluster Analysis can be done to serve certain business objectives, it can be based on value, the behavior of purchases, satisfaction levels, or communication behavior. It is very useful to set business strategies, especially for sales and marketing. Different clustering techniques can be used; each has its pros and cons.

2.1 Risk assessment

Underwriting nowadays depends a lot on machine learning since it saves time and cost, this helps in giving the firm a competitive advantage. Thus, improving the underwriting process is crucial to enhance customer acquisition and retention. Insurance firms with skilled underwriting teams have powerful impacts on the insurance business. Opposing selection can be avoided by correctly classifying the risk levels of individual applications through predictive analytics (Bhalla, 2012)

A study was done to predict the Heath insurance amounts, the study included the following attributes as independent variables: 'age', 'gender', 'BMI', 'children', 'smoker' and 'charges.' Three prediction models were applied: multiple regression, decision tree, and gradient boosting regression which has the highest goodness of fit (Bhardwaj & Anand, 2020)

Other researchers used massive health insurance claims to predict very high-cost claimants (Maisog, et al., 2019). They showed that such studies can be done using machine learning and not necessarily using actuarial science. The study used machine learning to identify the claimants who exceed 250,000$ per year. They split the model into formulations: predicting cost and binary classification whether a member will exceed a certain cost amount. Many variables were included in the model such as the list of diagnoses, medical procedures, drugs, medical history, age, gender, family size, unemployment, poverty, insurance coverage, education, minority, marital status in addition to more than 100 variables for claims data and their cost. Many models were

applied and the best performing model was the Light Gradient Boosted Tree classifier achieving a sensitivity of 91.5%.

Another study aimed to predict if the health care expenditure will increase next year or won't. It was applied to claims data provided by Helsana Group which is one of the largest health insurance companies in Switzerland (Jödicke, et al., 2019). The study was done on 373,264 patients where the data was for medical claims for two subsequent years and included demographic parameters such as age, gender, area of residence, marital status in addition to medical history such as the number of outpatient visits, health status, prescribed drugs which resulted to 449 groups of drugs. The goal of the research was to evaluate the risk factors for the cost increase.
The researchers used 3 models: Linear regression, Feedforward neural network, and BDT, which is a variant of decision tree methods with a gradient boosting algorithm governing the learning process. The BDT model performed the best leading to 67.6% accuracy. Finally, the researchers made a subgroup analysis on the subscribers who have the highest probability of cost increase to identify the most important features that led to the increase in cost such as type of drugs, diseases, pregnancy, etc. (Jödicke, et al., 2019)

In another risk prediction research and using supervised learning algorithms, the risk was divided into 8 ordered levels, and predictors used were age, height, weight, employment, insurance policy, insurance history, and medical history distributed on 48 dummy variables. Data was processed, missing values were imputed and the researcher used dimension reduction to reduce the number of variables and get efficient modeling, then 4 models were compared, REP tree which is a type of decision tree classifier technique had the lowest MAE, RMSE and was considered the best model. Researchers emphasized the importance of such studies and on the variability in models than can be implemented using such datasets such as in segmentation, marketing, sales, and premium rates predictions (Boodhun & Jayabalan, 2018)

Analyzing healthcare databases is very useful for extracting lots of information that are drivers to the success of health policies either for insurance firms or public health. Same as previous studies' methods a study was done in Ontario, Cañada to predict the high-cost patients who were considered 5% of the community (Fitzpatrick, et al., 2015). Variables including age, gender, income level, education, marital status, ethnicity, food security, residence setting, home-ownership, and others were applied to a logistic regression model and concluded that some factors such as food security, homeownership, income level education were all significant and greatly increased the odds of being a high-cost, these social determinants were important components and need interventions to improve public health, the objective of this study was to understand

components that affect health from a broader perspective and to look beyond cost. Another study targeting the same group of patients in Ontario, Cañada was done (Rosella, et al., 2014). But in this study, the patients were distributed on multiple groups and ranked individuals according to gradients of cost within each CCHS cohort (1, 2-5, 6-50, and lower 50th percentiles); high-cost were defined as the top 5% of users. Multinomial logistic regression was used to predict the risk group and some additional variables regarding health care were added to the model such as physical activity, smoking status, alcohol consumption, and life stress. Variables capturing health status (both self-reported and measured by health care utilization) were very influential. The study is a little different from similar studies since the researchers investigated the effects of health behaviors and health status in addition to multiple socio-demographic measures, in addition, they split the patients into multiple groups of risk not only binary. (Rosella, et al., 2014)

Another study that was applied on Western Denmark patients aimed to compare the standard method which only included 6 variables versus advanced machine learning algorithms in predicting high-cost patients, especially those who move from a lower to an upper level of expenditures within 1 year—that is, 'cost bloomers. The researchers used over 1000 features and applied elastic-net penalized logistic regression, the model achieved a 21% and 30% improvement in cost capture over a standard model for predicting high-cost patients and cost bloomers, respectively. (Tamang, et al., 2017)

The studies emphasize the importance of machine learning algorithms when working with big data and show how risk assessment and predicting high-cost clients can help insurance firms and governments in setting the right premium rates and in adjusting policies. The variables that were common in most studies were demographics such as age, gender, region (rural, urban), marital status, income level & family size in addition to medical history like chronic diseases, mental diseases, injuries, medicines taken, etc. Some studies added additional features that were influential and were relevant to the health condition like BMI, smoking status, physical activity, alcohol consumption… Other studies included socio-economic features such as house ownership, education, poverty, and food security.

Many machine learning algorithms were applied and, in most studies, the researchers tried different models and compared between them.

2.2 Fraud detection

Fraud and abuse detection are the most critical issues in health insurance since it brings massive financial loss to companies every year. Data science platforms and software are always updated to detect fraudulent activity. To make these detection policies on the

platforms, machine algorithm models are built to assist in understanding fraudulent actions. Usually, insurance companies use machine learning models for more efficiency. These models rely on previous fraud data or on unsupervised learning to recognize fraud schemes that were not noticed before. (top-10-data-science-use-cases-in-insurance, 2021)

Unsupervised classification seeks to detect the cases which are most dissimilar from the norm such as outliers. Outliers should then investigate for fraudulent activity, one can think that the objective of the analysis is to return a suspicious score for those unusual observations. (Bolton & Hand, 2002)

A study was made using data from the U.S health care system to discover fraudulent behavior, data included demographic of patients, claims data (including diagnoses information, claim amount, count, date), medical providers' data (ID, type of provider). The diagnosis was considered the control variable in the research where claims were classified according to the principal diagnosis, afterwards, the researcher chose the most common diagnosis to study them. The payment amount and the count are the most common in the existence of fraud in literature so they were entered in the clustering model. As a next step, the researcher studied the suspicious clusters that had the highest distance from the population and had the largest amount of payment. (Liu & Vasarhelyi, 2013)

Fraud can be either caused by medical providers or by clients, a study was applied on a major insurance health organization to detect fraud and abuse. The involved data was physicians' prescription claims. Cluster analysis was applied to identify suspect physicians, and discriminant analysis to assess the validity of the clustering approach. The results identified 2% of physicians as suspects of fraud. Discriminant analysis suggested that the indicators established adequate performance in the detection of physicians who were suspects of fraud. (Joudaki, Rashidian, Minaei-Bidgoli, Mahmoodi, & Geraili, 2015)

Another study used a new methodology in fraud detection where the researcher proposed splitting the model into stages (Johnson, 2016). The first three stages are aimed at detecting outliers among providers, services, and claim amounts. Stage four integrates the first three stages and obtained a risk measure. Stage five aimed to compute risk threshold values. The final step is done by comparing the risk value with the risk threshold to define which claim is considered fraud. As every diagnosis needs some tests and medication thus it can be considered and each diagnosis has an average cost, in this way it was easy to identify the abnormal claims, and all cases above 3 standard deviations are considered abnormal behavior. The second stage of fraudulent behavior computes likelihood values that diagnoses of claims do not belong to groups of population

parameters like age or gender. Stage three calculates the likelihood that claim amounts were overstated. Stage four then computes the risk values of claims using distance and likelihood values. Stage five determines a risk threshold value for each claim and the final stage defines which cases are considered fraud. The used methodology was unsupervised neural network methods.

Clustering is the common method of detecting fraudulent claims, another researcher used the same methodology, first applied cluster analysis based on diagnoses type using the attributes average cost and number of bills, medicines cost, etc. The study first does a Cluster Analysis to identify the number of clusters whose members share a common billing pattern, then used multivariate outlier detection methods – the Maha. distance, the robust distance, and the robust distance. (Macedo, Araia, & Zafari, 2016)

The studies clarify that the best methodology to detect fraud is to use unsupervised learning and detect outliers in the claim's datasets. The used attributes that are common in the studies are average cost per diagnosis, number of bills, the time interval between bills, and providers data. These attributes will be used in this study to detect the outliers that might be suspects of fraudulent behavior.

Chapter three

Methodology

3.1 Introduction

In this study, machine learning algorithms were used to illustrate the importance of ML and data mining in the health insurance sector. Two areas of study were covered:

- Risk Assessment
- Claims fraud detection

Detailed Algorithms and used attributes are explained in Section 3.7

3.2 Data description

3.2.1 Population

The population is health insured subscribers in a local Insurance company located in Palestine called Tamkeen Insurance Co. The subscribers are insured at the corporate level which means the corporate account purchases an insurance policy for its employees and their families including spouses, children, and newborns who are added to the insurance on birth date.

3.2.2 Sample data

The data set consists of 10,843 subscribers with about 100,000 medical claims, and with 24 attributes, which describe the characteristics of insurance applicants in addition to the claims' data description. The data set comprises nominal, continuous, as well as discrete variables. The sample dataset is based on subscribers' medical claims in 2020 for active subscribers only and excluding the New/ Churned corporate accounts during 2020

The attributes were reviewed with an experienced expert who has good knowledge in insurance fraud prevention and underwriting policies.

The following table shows the attributes present in the data set.

Table (3.1) Attributes used in the study

Attributes used for risk assessment
• Gender • Age • Marital Status: Married, Single • Subscriber type: corporate employee, Spouse, Children, New-born (added within 2020) • Number of children • Age of oldest child • Has chronic Diseases: yes, no • Number of Chronic Diseases • Had Old surgery (during last year): yes, no • Wear glasses: yes, no • Corporate account Type: NGO, educational, medical institution, a private company • New/Old account: started with the insurance company before 2020/or in 2020 • Account Policy coverage: Mid/ High based on the agreement with the Insurance company • Policy cost • Number of claims per subscriber/the year 2020 • The outpatient sum of claims per subscriber/the year 2020 • Inpatient sum of claims per subscriber/the year 2020
Attributes used for Fraud Detection
• Claim ID • Provider ID • Claim type: Physical therapy, X-ray, Dental, Glasses, Clinical visit, Pharmacy, Laboratory, Procedures, Surgery, Inpatient Hospitalization • Claim Date • Diagnosis Type: general medical test, cardiac, Skin test, optical, toxicological examination, genetic, blood analysis, vitamin' deficiency, Back pain, Asthma, …

To apply a risk assessment model, a new attribute was created from the dataset splitting the subscribers into 3 risk levels:

- ✓ High-risk: subscribers who have both inpatient <hospital admissions> and outpatient claims and are considered non-profitable and have the highest cost.
- ✓ Mid-risk: subscribers who have only had outpatient claims with no hospital admissions, but have high outpatient costs that exceed the policy cost. They are considered non-profitable but with lower total costs than High-risk.
- ✓ Low-risk subscribers who only have outpatient claims and they don't exceed the yearly policy cost. They are considered the profitable segment.

This new attribute will be used as the dependent variable in the model to predict the risk level which will assist in predicting the insurance cost.

The following shows how 23% of subscribers contribute 67% of total health insurance cost, and this clarifies how it is important for the insurance company to specify the number of risky subscribers before setting the premium rates in the underwriting process.

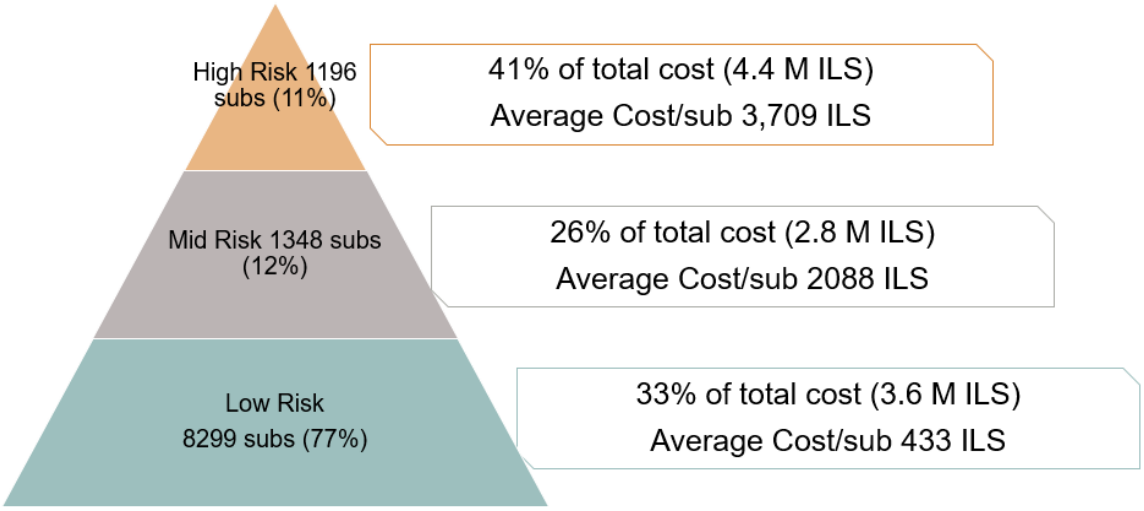


Figure (3.1) Risky vs Non-Risky by (Researcher)

3.3 Data processing

Raw data was converted to a useful and efficient format so it can be used in the machine learning models. Data files were merged and aggregated to create new attributes, such as the number of children, age of the oldest child, number of chronic diseases, number of claims, and total claims amount. Other attributes were grouped such as company type and diagnosis.

The data files used to create the final datasets for the risk assessment and fraud detection models are:

- File1: Subscribers' Demographics
- File2: Subscribers' Medical History including chronic diseases, wear glasses and if the subscriber had previous surgery
- File3: Accounts' Policy cost and coverage
- File4: Medical Claims details in the year 2020 including claim ID, diagnosis, cost, date, claim type, medical provider, and if the claim was inserted manually or through the online system

- File5: Medical Claims details in the year 2019

Some aggregation functions and mathematical equations were applied to the data files to extract useful variables and the files were joined to prepare the final two datasets:

1. Dataset for Risk assessment, where the key variable in the join relationships was the subscriber ID and consisted of 10,843 rows.
2. Dataset for Fraud Detection, where the key variable in the join relationships was the Claim ID for the clinical visit and consisted of 9,612 rows. For each clinical visit join relationships with the laboratory expenses, pharmacy expenses, X-ray expenses, and procedures expenses were done.

   Not all customers were used in the fraud detection dataset due to the following reason: The insurance company in 2020 migrated from a manual system for claims management to an online management system. The claims records that were inserted in the manual system didn't include sufficient information regarding diagnosis type and expenses associated with the clinical visit, accordingly, the sample was reduced to include only customers who had all their claims inserted in the online system and had full information.

No missing values were available in the dataset, so there was no need for imputation of missing values.
The used tool in the study is R software

After collecting and arranging the datasets to be used in the machine learning algorithms, the dataset was split into two subsets:

- ❑ Training set (60%): the subset is used to train the models
- ❑ Testing set (40%): The subset is used to test and evaluate the models that were produced using the training set

3.4 Data exploration

After arranging the raw data, data was explored to check the significance of the variables and to check if there is a high correlation between the variables to avoid multicollinearity.

3.4.1 Data exploration/ risk assessment

The following table describes the distribution of subscribers based on the following factors: gender, marital status, subscriber type, chronic diseases, and having previous surgery

Table (3.2) Data Description/ Categorical variables

	Count	frequency
Subscriber type		
children	4716	43%
employee	3702	34%
newborn	105	1%
spouse	2320	21%
Total	10843	100%
company type		
companies	4999	46%
educational	3207	30%
medical	2000	18%
NGOs	637	6%
Has Chronic Diseases		
No	9802	90%
Yes	1041	10%
Had old surgery		
No	10305	95%
Yes	538	5%
Marital Status		
married	4748	44%
single	6095	56%

The following table shows the mean, minimum, and maximum of the numerical variables:

Table (3.3) mean values of numerical variables

Attribute	Min.	Mean	Max.
Number of chronic diseases	0	0.20	4
Age	1	25	69
Number of kids	0	1	11
policy cost	650	1144	2541
Out-patient Sum Value	1	741	5990
Inpatient Sum Value	613	2345	32,976

Variables significant based on risk factor

As a first step, visualizing the attributes compared to the dependent variable assists the researcher in checking the significance of these attributes and helps in visualizing which attributes might influence the models.

Numeric variables:

Since the dependent variable is multiclass, the ANOVA test was checked and all differences in means were significant for all variables since p-values were less than 0.05. ANOVA tests details are in Appendix(B)

The following chart shows the difference in means based on the factor: High Risk, Mid Risk, Low Risk.

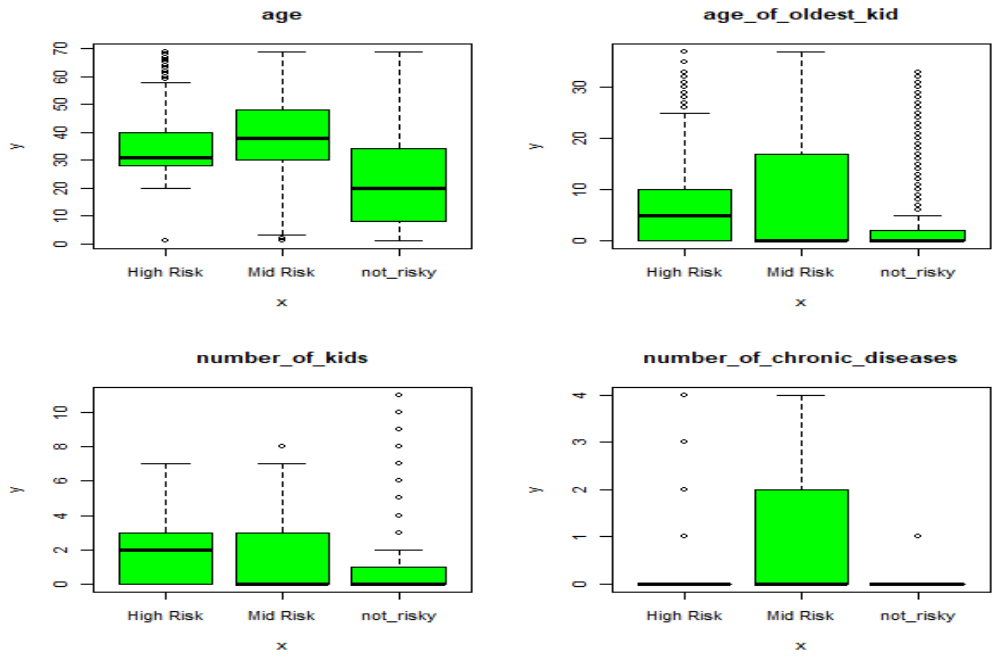


Figure (3.2) Means Differences by Risky factor by (Researcher)

There are significant differences in means between the groups which indicates that those attributes are good to be considered in the machine learning models. (Tzinie, 2020)

The Density plot also shows the difference in distribution based on the risk level, for example, the age distribution in low risk is skewed to the left which indicates that younger people are generally less risky than older people.

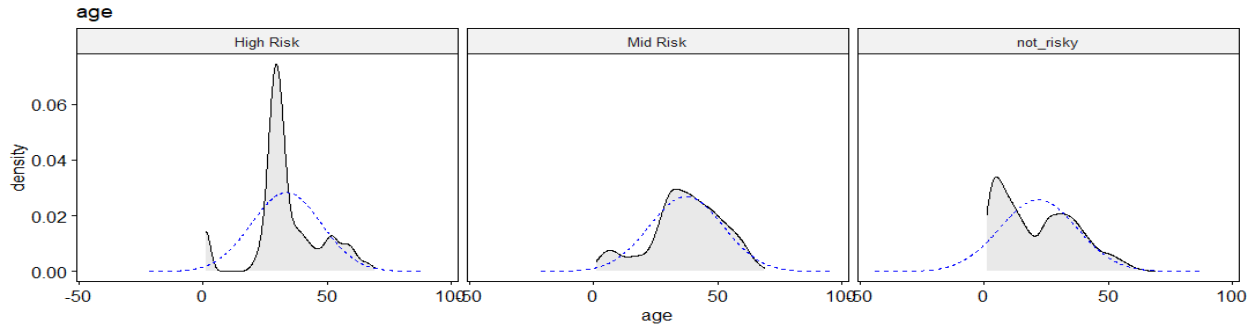


Figure (3.3) Density Plots by Risky factor by (Researcher)

The following graph also shows the effect of the two variables age, the number of chronic diseases on the risk level, and as shown it is clear that when age increases in addition to having chronic diseases the customers tend to have a higher risk.

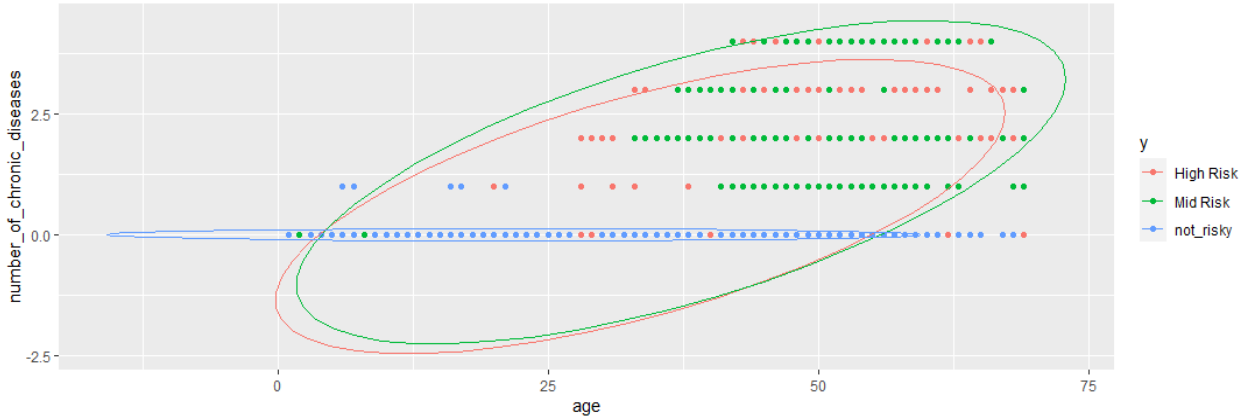
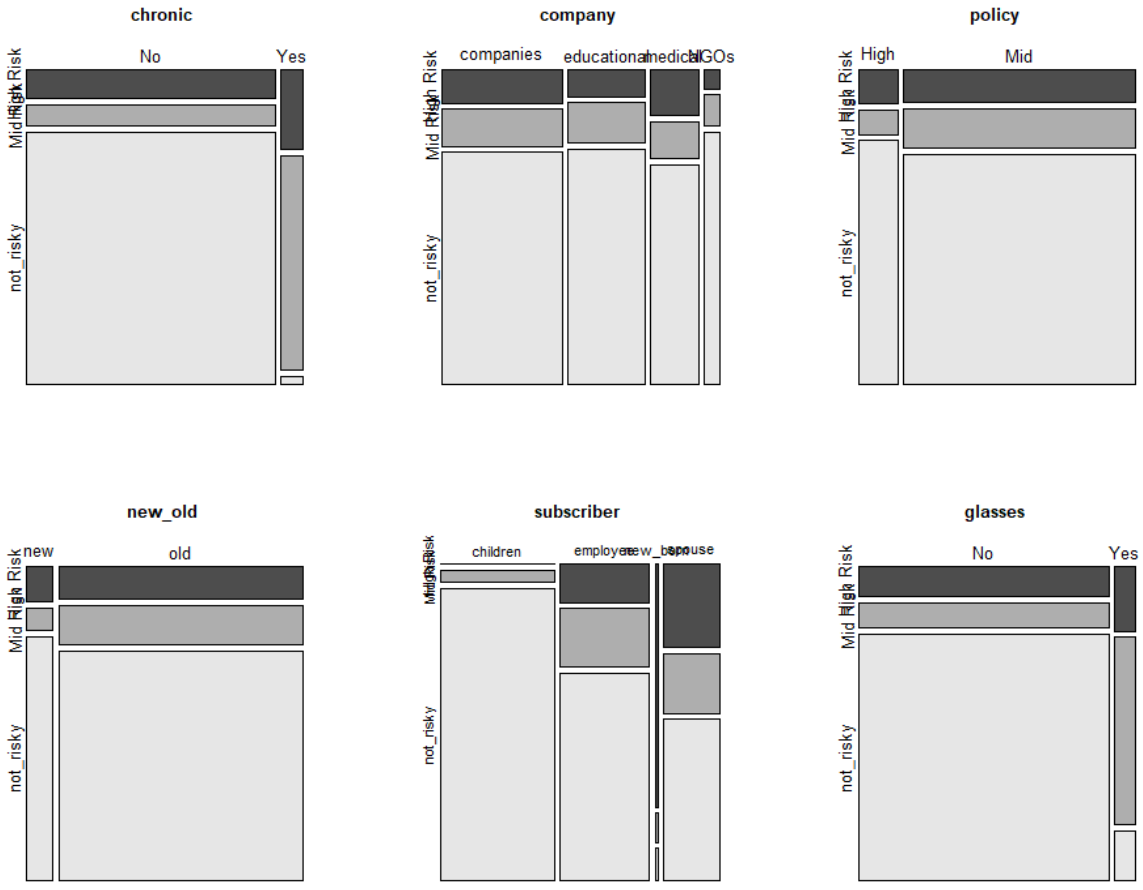


Figure (3.4) Two predictors plot by Risky factor by (Researcher)

The following figure shows the Mosaic plots which show a two-way frequency distribution between risk factor and other categorical variables



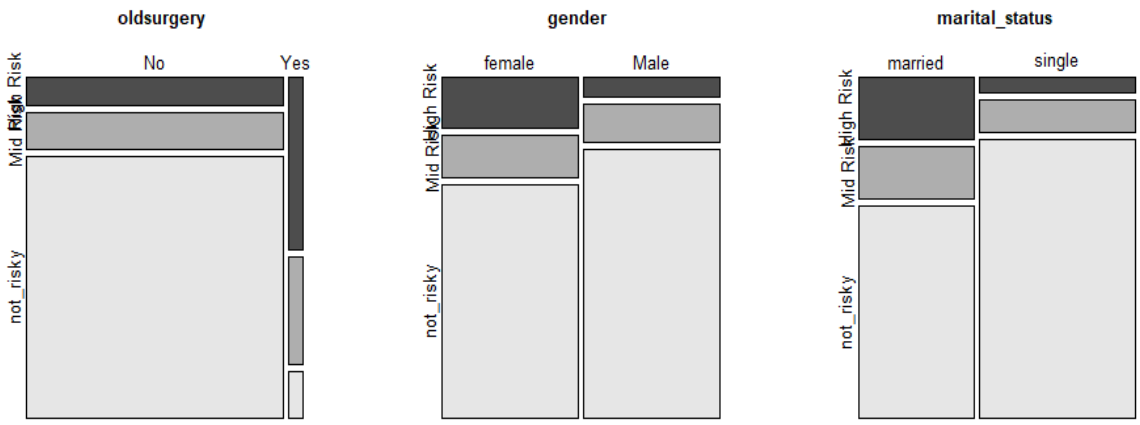


Figure (3.5) Mosaic plots between Risky factor and other attributes by (Researcher)

The percentage differences between the risk levels are clear for many attributes such as chronic diseases where most of the subscribers who have chronic diseases are High & Mid Risk. Same for old surgery where most of the subscribers who had old surgery are risky and that's most probably because they need medical follow-ups or special medicines. Also, married people have a much higher percentage of High-risk than single people.

To confirm that the frequency distribution differences are significant Chi-square tests were conducted and all categorical variables had p-values less than 0.05, which means there were significant differences in the two-way frequency distribution between each attribute and the dependent variable. (Gajawada, 2019)

Table (3.4) Chi-square test for Categorical Independent variables

Variable	p-value
company type	< 0.001
new_old_account	< 0.001
policy coverage	< 0.001
subscriber	< 0.001
wear glasses	< 0.001
old surgery	< 0.001
has chronic	< 0.001
gender	< 0.001
marital status	< 0.001

Multicollinearity check

To check if there is collinearity between the numerical variables a correlation matrix was applied using Pearson's Correlation.

The following chart shows the correlation matrix

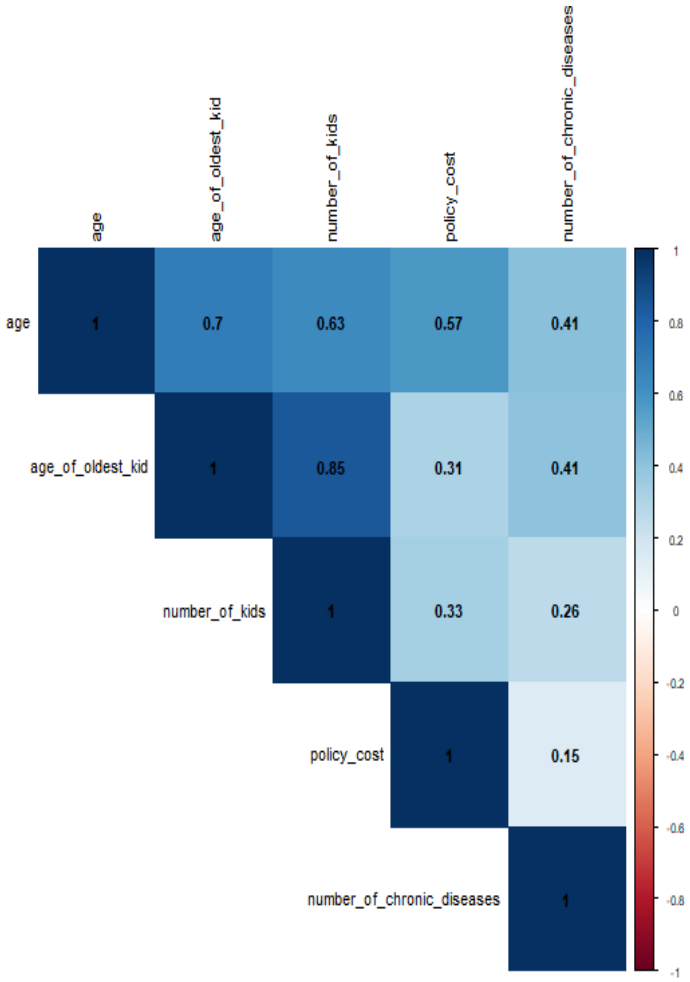


Figure (3.6) Correlation plot between numerical variables by (Researcher)

Collinearity was also checked by the variance inflation factor (VIF) where If values of VIF exceed 10 this often means there is high multicollinearity, also values exceeding 5 may be a cause of concern.

In this model, the VIFs were below 5, so there is minimal multicollinearity and there is no need to drop any variable.

Table (3.5) VIF test for Collinearity check

Variable	VIF
Policy cost	1.54
age	2.89
Age of oldest kid	4.34
Number of kids	3.57
Number of chronic diseases	1.22

Correlation Between Categorical variables

Categorical variables were encoded into a matrix to compute the correlation

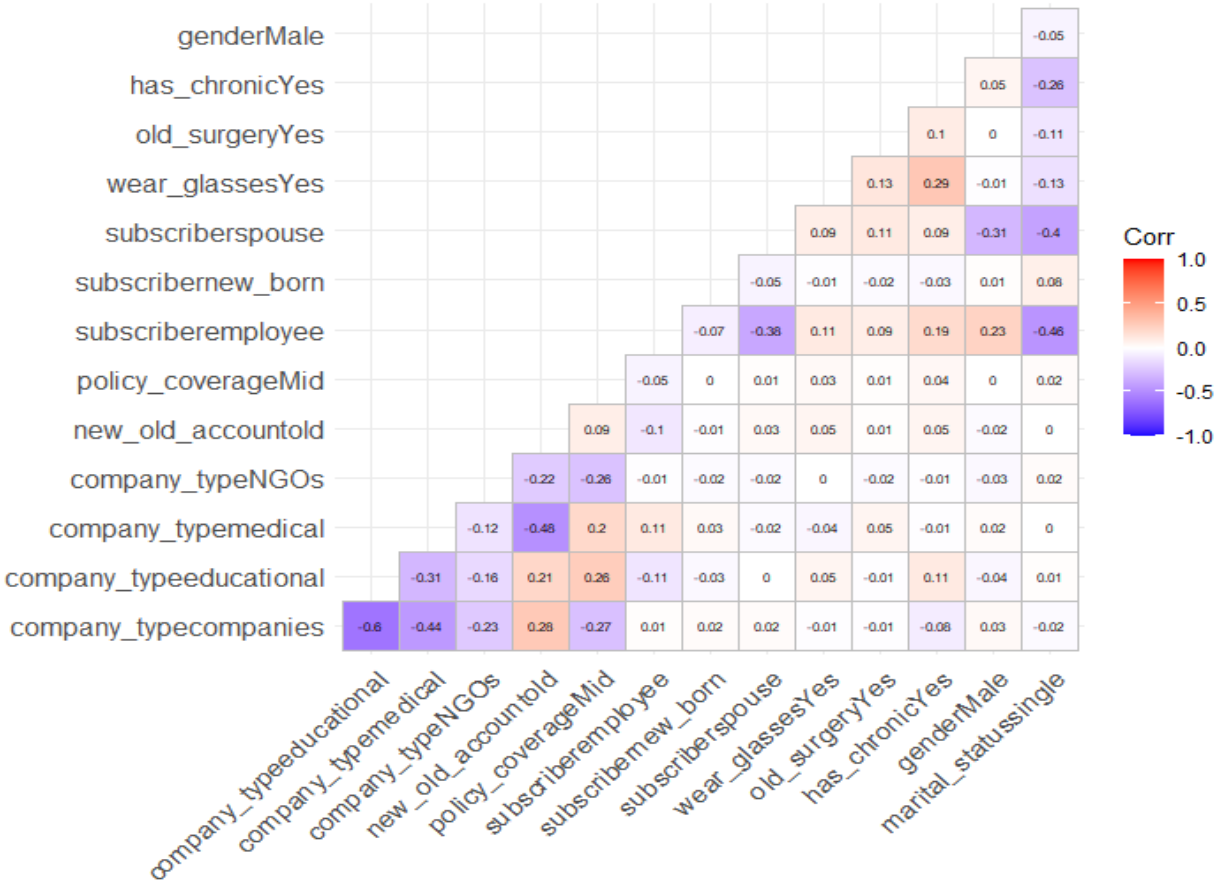


Figure (3.7) Correlation plot for categorical variables by (Researcher)

There is no high correlation between the categorical variables so none of them will be dropped from the prediction models.

3.4.2 Data exploration/ fraud detection

A sample of clients' claims was used in the fraud detection models. Not all customers were used since some of the claims had insufficient data and were not inserted in the online Insurance system with full information as explained in section 3.3. Accordingly, the sample was reduced to include only customers who had all their claims with full information. The remaining sample included 4278 customers with an average yearly number of clinical visits of 2.3

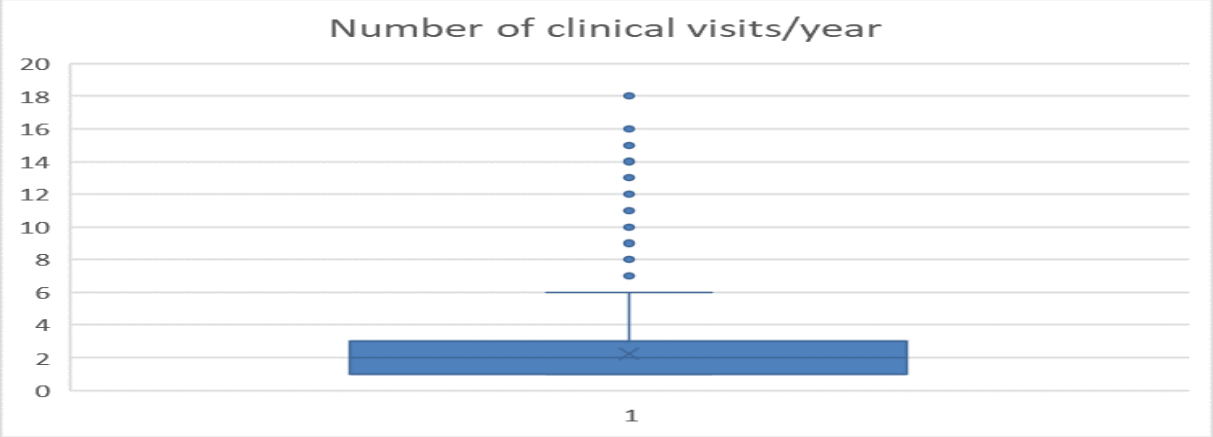


Figure (3.8) Boxplot for average clinical visits/year (Researcher)

Some customers have more visits per year than the norm as shown in the boxplot above.

An ANOVA test was applied to check if there is a significant difference between the total cost of each main diagnosis category, the test indicated that the differences in means were significant since the p-value was less than 0.001 and the following chart shows the means differences

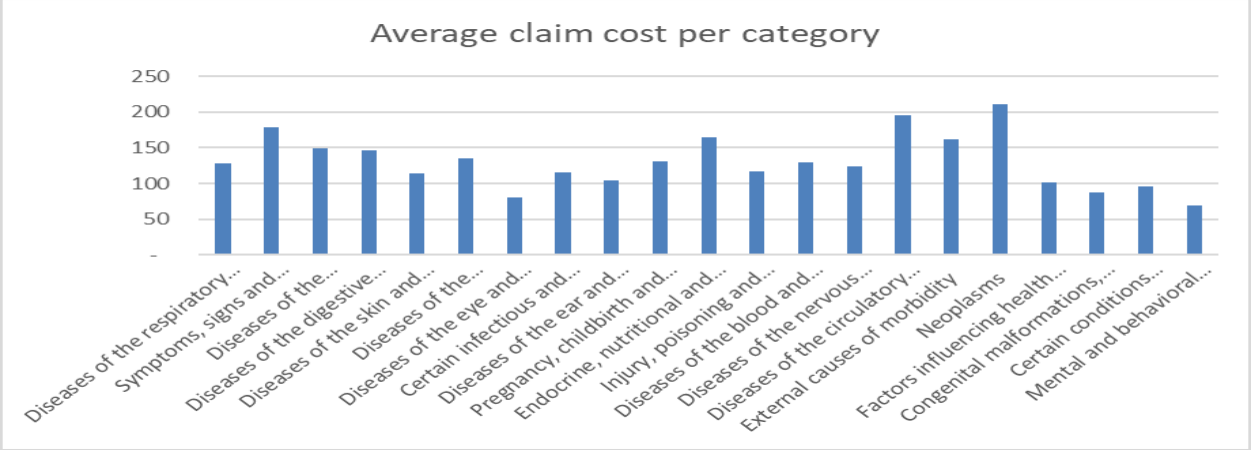


Figure (3.9) Bar-chart for Average Claims Cost based on diagnoses types (Researcher)

The highest diagnosis cost is for the diagnosis category Neoplasms followed by diseases of the circulatory system.

Since the diagnosis type is a significant variable in the total cost so it would be appropriate to add it to the fraud detection model, where the objective is to find claims that are outliers from their similar diagnosis group in lab costs, pharmacy costs, x-rays costs, and procedures costs.

Another significant variable that will assist in detecting fraudulent claims is the time interval between claims for the same customer and whether the previous visit was for the same doctor.

On average, the time interval between the claims is 128 days as shown in the following Boxplot

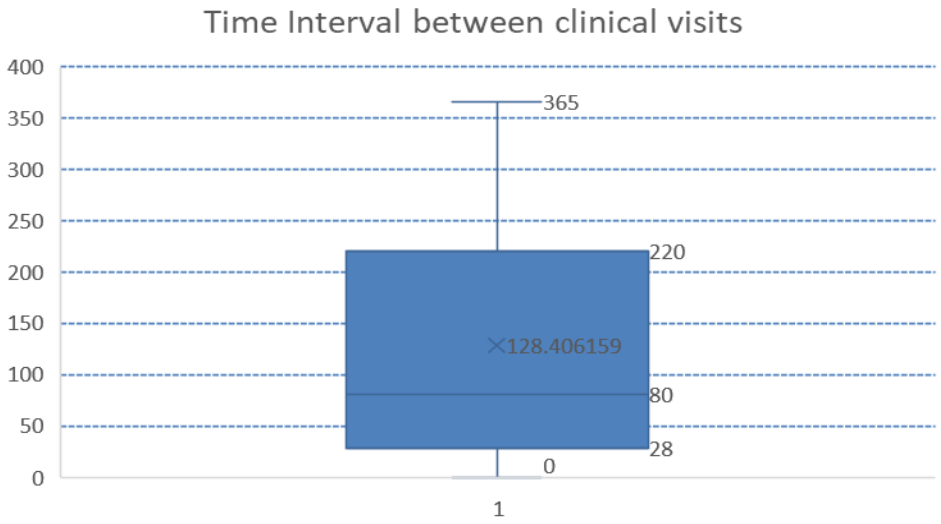


Figure (3.10) Boxplot for the time interval between clinical visits for the same customer (Researcher)

Based on the data exploration, it might be appropriate to add these attributes in the outlier detection models to detect the outliers that might be suspects of fraudulent behavior.

3.5 Machine learning algorithms

3.5.1 Introduction

Machine learning algorithms are divided into supervised and unsupervised algorithms

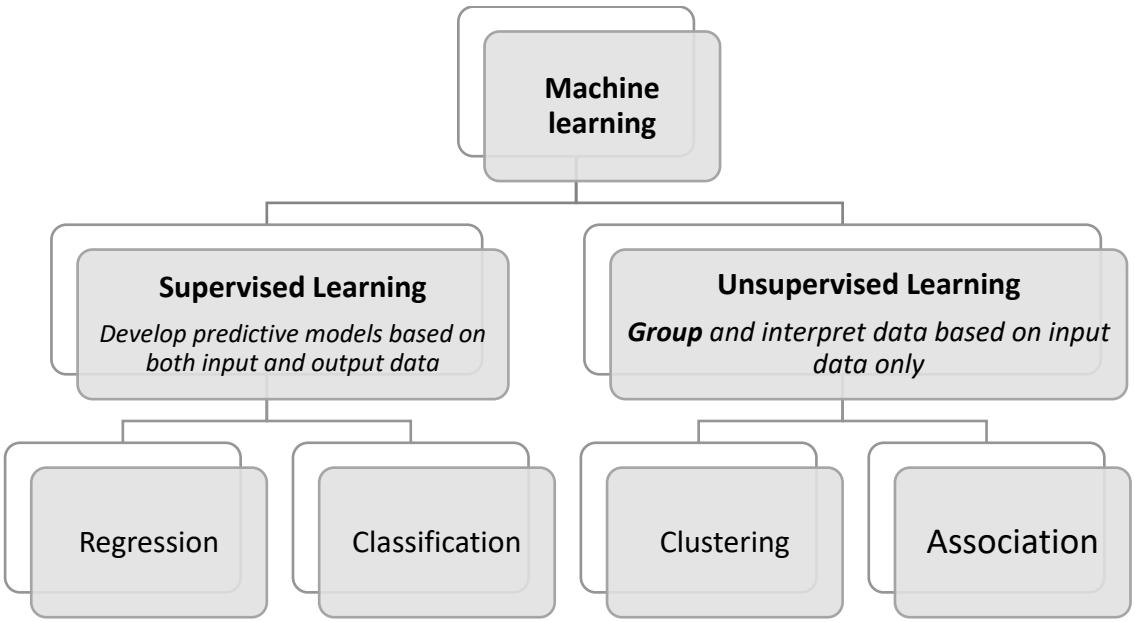


Figure (3.11) ML algorithms by (Researcher)

The main difference between supervised and unsupervised learning is that supervised learning uses labeled data to help predict outcomes, while unsupervised learning doesn't. In supervised learning, the machine learning algorithms use datasets to train the model and predict the outcome, and model accuracy can be measured. As for unsupervised learning, the models analyze the input data, discover hidden patterns, and group/cluster the data accordingly without the need of labeling output data. (Mohri, Rostamizadeh, & Talwalkar, 2018)

Types of Supervised Learning:

I. **Classification:** It is a common technique of data mining and is used to categorize each item into classes or groups within a data set, the model learns from the labeled dataset to predict the future. For example, document classification consists of assigning a class such as sports, politics, or weather to each document. Some common algorithms are Logistic regression, Naïve Bayesian classifier, support vector machines (SVM), decision trees, k-nearest neighbor, random forest, neural network classifiers. (Mohri, Rostamizadeh, & Talwalkar, 2018)

II. **Regression:** It is used to understand the relationship between dependent and independent variables and predict numerical values of the dependent variable.

Some common algorithms are Linear regression, Polynomial regression, Neural network, and Poisson regression. (Abdulhafedh, 2022)

Unsupervised learning models are used for two main tasks:

I. **Clustering** is used for grouping unlabeled data based on their similarities or differences to identify patterns or groups of similar objects within a data set. Useful information can be extracted from unsupervised learning and can be very helpful in market segmentation for example. Some common algorithms are K-means clustering and hierarchical clustering. (Kassambara , 2017)
II. **Association:** It is used to discover rules and find relationships between variables in a given dataset, such as people that buy X also tend to buy Y. It is mostly used for market basket analysis and recommendation engines. (Delua, 2021)

Choosing the right algorithm or approach is based on the goals of the study.

3.5.2 Some common classification machine learning algorithms:

The following are brief descriptions of some common classification algorithms that will be used in this case study.

3.5.2.1 Multinomial logistic regression

Multinomial logistic regression is an extension of binary logistic regression that allows having more than 2 categories for the dependent variable.
It is a classification algorithm and is used for the prediction of the outcome of a categorical variable in which the log odds of the outcomes are modeled as a linear combination of the predictor variables which can be nominal, ordinal, interval, or ratio-level.

The multinomial logistic regression having r levels in the dependent variable estimates (r-1) separate binary logistic regression models having a reference category used in the (r-1) models.
Each regression model explains the effect of the predictors on the probability of success in that category in comparison to the reference category. ach model has different coefficients where the predictors can affect each category differently. (Aggarwal, 2015)

In the Logistic regression model, we are mainly interested in the odds of the logistic curve which is the ratio of something happening (Y=1) to something not happening (Y=0)

$$\textbf{Odds}= \frac{P(Y=1)}{1-P(Y=1)} = e^{B0+B1x1+\cdots}$$

By taking the logarithm of the odds which is called the logit(P) this estimates a multiple linear regression function

Ln(Odds) = $B0 + B1x1 + \cdots$

The coefficients of the Logit function are coefficients of the log-odds of the default class. (What is Logistic Regression?, 2021)

The odds can be retrieved back by taking the exponential of the coefficients, the exponential coefficients measure how relevant an independent variable is and tell us about the direction of the relationship (positive or negative). But their impact is multiplicative where an odds of 1 means no change, so to calculate the magnitude of change in the dependent variable we use *(Exponential of the coefficients -1)*100%*

Multinomial logistic regression is less affected by basic assumptions like normality or equal variances. But we need to check that there is no high multicollinearity among the independent variables and that the sample size is sufficient, in addition, it is preferable not to have outliers in the data. (What is Logistic Regression?, 2021)

The regression coefficients are usually estimated using **maximum likelihood estimation (MLE)** which is a method of estimating the coefficients of a probability distribution by maximizing the likelihood function.

To measure the goodness of fit of the model the **deviance: -2 log-likelihood (-2LL)** is used, it measures how much-unexplained variation there is in the logistic regression model, where it compares the difference in probability between the predicted outcome and the actual outcome for each case and sums these differences together to provide a measure of the total error in the model. But to compare the value of the deviance, it should be compared against a **baseline model**, where it helps to test if the model is significantly more accurate than just guessing the outcome which will be the category with the largest number of cases. (Aggarwal, 2015)

The significance of the model can be tested by the Chi-square test of -2LL difference as follows, where the p-value should be significant.

X^2= [-2LL (baseline)] - [-2LL (new)]

with *degrees of freedom= $k_{baseline}$- k_{new}, where k is the number of parameters in each model.*

The two most common tests to measure the goodness of fit and effectiveness of the model are Hosmer & Lemeshow's R2 and Nagelkerke's R2. Both describe the proportion of variance that the model successfully explains. A value near 1 means a better fit of the model. (Using Statistical Regression Methods in Education Research, 2011)

3.5.2.3 Decision tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to parameters set.
These splits are represented as nodes. The algorithm adds a node to the model every time that an input column is found to be significantly correlated with the predictable column. (Shwartz & David, 2014)

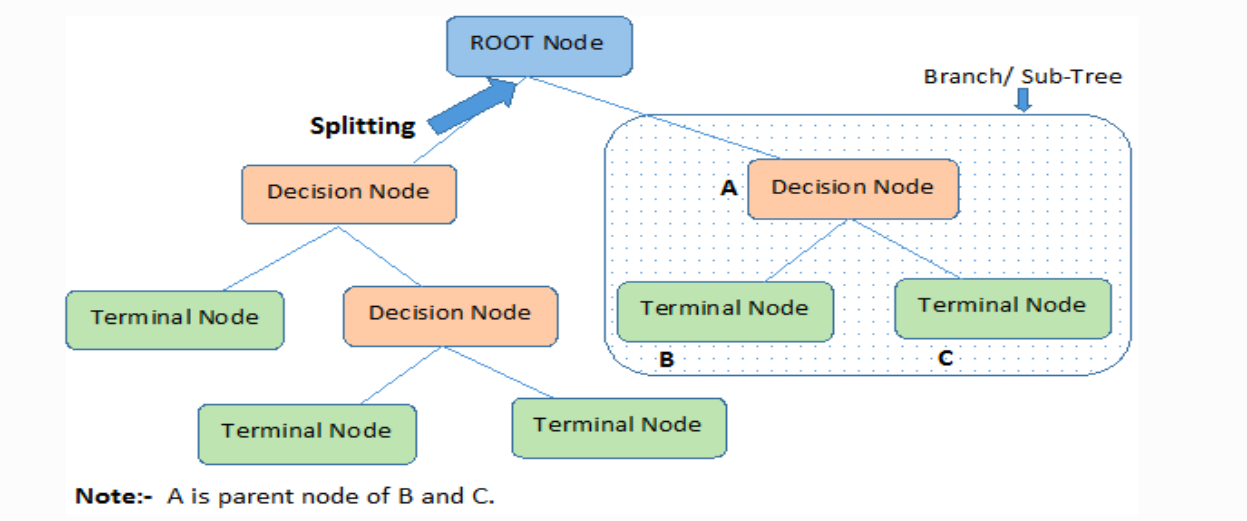


Figure (3.12) Decision tree by (Chauhan, 2020)

Decision Tree's main function is to identify the attribute for the node in each level. This process is called attribute selection.
The most popular attribute selection measure is Information Gain.

Information Gain depends on Entropy
Entropy measures the purity of the split. The higher the entropy is the more, the harder it is to draw any conclusions from that information.

$$E(S) = \sum_{i=1}^{c} -p_i \, log_2 p_i$$

Information Gain is a measure of the change in entropy when the decision tree partitions the training instances into smaller subsets which is the expected reduction in entropy.

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{k} Entropy(j, after)$$

Steps in decision tree algorithm using Information Gain measure:

- ❑ First, it considers the original set S as the root node.

- ❑ On each iteration and for each attribute in the dataset it calculates Entropy(H) and Information gain (IG) of this attribute.
- ❑ Then it selects the attribute which has the Largest Information gain.
- ❑ The set S is then split by the selected attribute and produce the first subset
- ❑ The algorithm continues the same process but with excluding attributes that were selected in previous nodes. (Chauhan, 2020)

Some packages designed for decision trees algorithm use other measures for attribute selection such as Gain Ratio. Gain Ratio is used to normalize the information gain of an attribute against how much entropy that attribute has

$$Gain\ Ratio = IG/Entropy$$

At first, determine the information gain of all the attribute and pick the attribute of higher gain ratio to split. (Tyagi, 2021)

3.5.2.4 Random Forest

Same as Decision Trees, a Random Forest is a type of Supervised Machine Learning that can be used for both classification and regression.
A random forest fits several decision trees on various sub-samples from the dataset and then averages to improve the accuracy and control over-fitting which means it predicts the final output based on the majority votes for each data point.
Random forests depend on the strength of individual decision trees and the correlation among the trees. (Hartshorn, 2016)

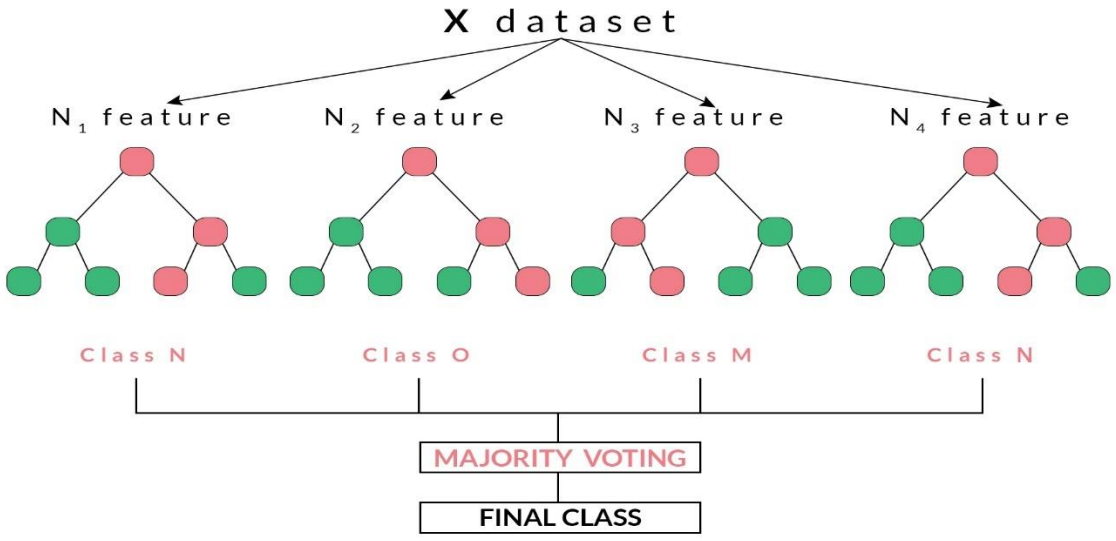


Figure (3.13) Random Forest by (Tahsildar, 2019)

The greater number of trees provides higher accuracy, also the sub-sample size can be controlled in the algorithm with the max samples' parameter.

3.5.2.5 Neural network

Neural Network ANN is an information processing model inspired by the human nervous system; a neural network is a complex adaptive system that can change its internal structure by adjusting weights of input (can learn by examples).
Neural networks provide the best solutions to many problems in image recognition, speech recognition, and natural language processing.
The neural network was designed to solve problems that are easy for humans and difficult for machines such as identifying pictures, identifying numbered pictures. These problems are often referred to as pattern recognition. (Nielsen, 2015)

There are two main types of artificial neural networks: Feedforward and feedback artificial neural networks.

- ❑ In Feedforward neural network Neurons in this layer were only connected to neurons in the next layer (travel in only one direction towards the output layer)
- ❑ Feedback neural networks contain cycles. Signals travel in both directions by introducing loops in the network.

<u>Feedforward neural network</u>

Neural Network is comprised of node layers, the first layer of the neural network receives the raw input, processes it, and passes it to the hidden layers. The hidden layer passes the information to the last layer, which produces the output

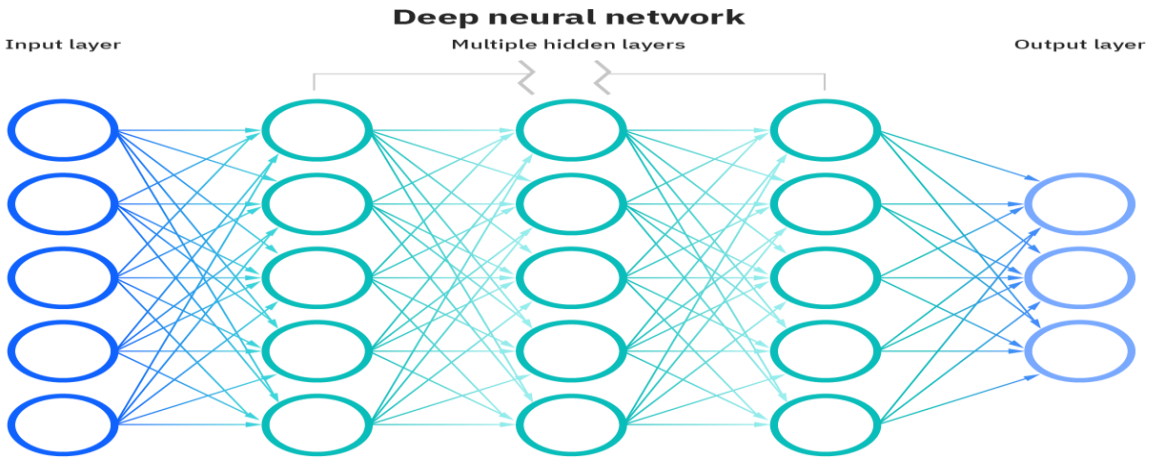


Figure (3.14) An example of a feedforward neural network (IBM Cloud Education, 2020)

Each node, or artificial neuron, connects to another and has an associated weight and threshold. It is like that each node as its own linear model, composed of input data, weights, a bias (or threshold), and an output

$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias$$

If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. This results in the output of one node becoming in the input of the next node. (Patel, 2019)

Activation function

The activation function defines the output of a neuron in terms of a local induced field. Activation functions are a single line of code that gives the neural nets non-linearity such as sigmoid function

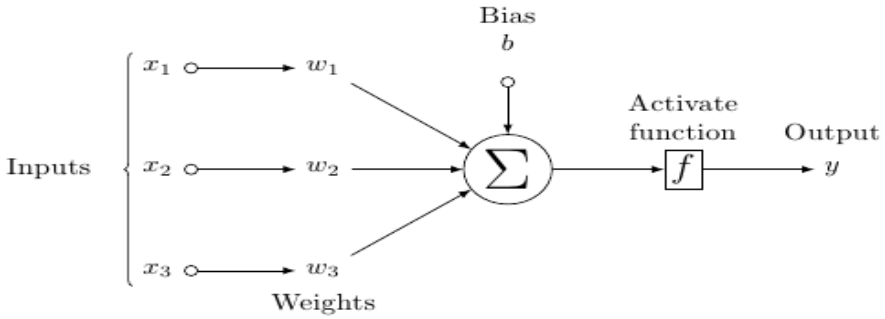


Figure (3.15) activation function in ANN by (Patel, 2019)

The performance of the neural network algorithm used for classification can be measured by calculating the classification accuracy on the hold-out test set. (Nielsen, 2015)

3.5.2.6 Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression problems.

SVM when used in classification models aims to find a hyperplane that divides the groups of the data points (Mohri, Rostamizadeh, & Talwalkar, 2018), where the algorithm maps the data points to a higher dimensional space and then find the hyperplane that maximizes the margin (which is the distance between the hyperplane and the closest data point) between two classes assuming that all data points lie on the correct side of the hyperplane as shown in the following figure

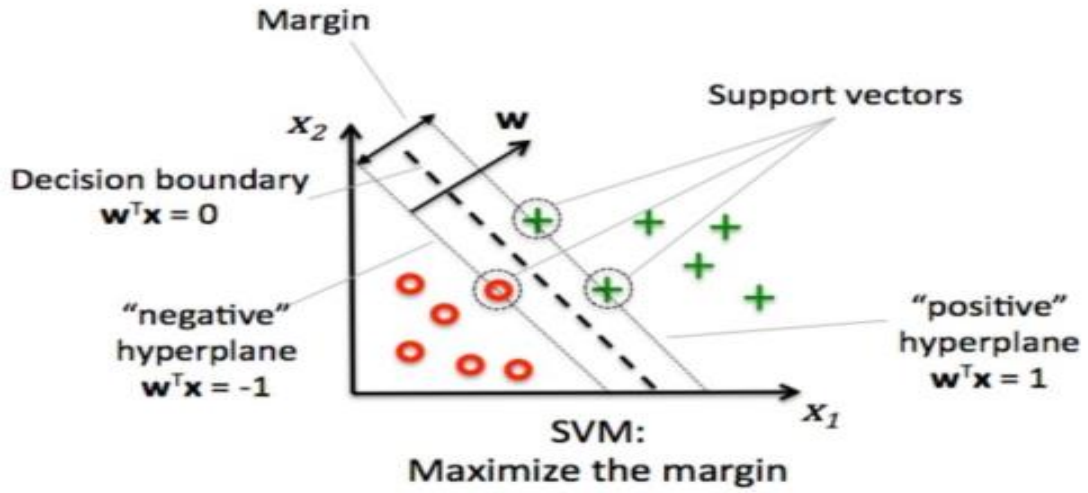


Figure (3.16) Support Vector machine (Demir, 2021)

The Nth dimensional space is called Kernel, the main types of Kernels are:

- Linear Kernels
- Polynomial Kernels
- Radial Basis Function Kernel

Nonlinear SVM maps the data from its original space into a higher dimensional one where it can linearly separate the data points. The learned hyperplane is then expected in its original input space. (Deng, 2013) resulting in having a nonlinear decision boundary as shown in Figure

$$\phi(x_1, x_2) = (z_1, z_2, z_3) = (x_1, x_2, x_1^2 + x_2^2)$$

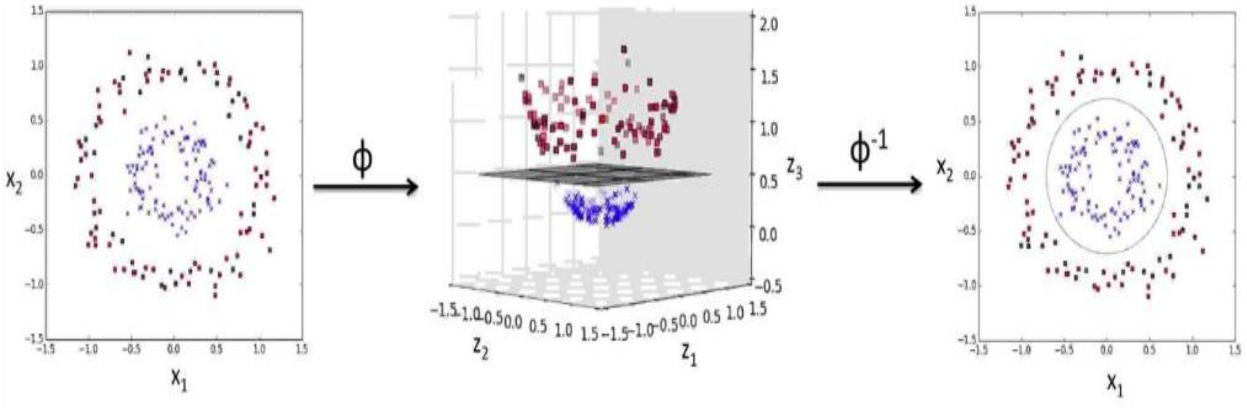


Figure (3.17) Support Vector machine, dimensional space (Demir, 2021)

3.5.2.7 Comparison between classification algorithms

Many algorithms can solve classification models, each has its different statistical method. Logistic regression is very useful for understanding the influence of the predictors and makes no assumptions on the distribution of the data but has its disadvantage since it assumes all predictors are independent of each other and that's not always the case in real-life data.

As for Decision trees, they can be very sensitive to the data they are trained on and causes overfitting which leads to misleading results when applying the data on a new dataset, while random forest performs better since each individual tree uses a random sample from the dataset, so this eliminates the problem of overfitting and assures higher accuracy.

As for the SVM algorithm, the main issue is to define the Kernel correctly, where SVM learns decision boundaries that have the shape in the high-dimensionality space based on the kernel specified, while Neural networks have a different way of operating and don't require kernels.

The main disadvantage of Neural Networks is that the functions don't always guarantee convergence, need computational power, and are generally much slower than SVM.

The conclusion is that the best model depends on the data distribution, data size, and the computational power needed, so the best approach is to test multiple models and compare the results.

3.5.3 Unsupervised learning algorithms

Unsupervised Learning is used to identify patterns in unlabeled data, it can be used for clustering, dimensionality reduction, association, and outlier detection.

There are many used algorithms in unsupervised Learning based on the purpose of the study. In this case study, the used algorithms are K-means & Hierarchal Clustering for clustering the subscribers based on their associated health claims' costs.

And for outlier detection the used algorithms are: Local Outlier Factor (LOF), Automatic PAM clustering algorithm for outlier detection (APCOD) & Isolation Forests (IF)

3.5.3.1 Hierarchical clustering

Hierarchical clustering is a type of unsupervised learning. The objective of cluster analysis is to classify groups that have similar perceptions and to profile these groups.

In Hierarchical clustering, clusters are formed by the composition/decomposition of cases, where larger clusters are formed by merging smaller clusters. It begins by treating every data point as a separate cluster, then identifying the two clusters which are closest together and merging them. The process continues until all clusters are merged. (Kassambara , 2017)

Hierarchical clustering doesn't require the number of clusters k as an input, it computes distances between the clusters using methods like Euclidean and Manhattan distances. There are two algorithms to apply:

- Agglomerative clustering: It works in a "bottom-up" manner. Whereat each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster and keep the iterations until all points are members of just one single big cluster.
- The divisive hierarchical clustering is the reverse of the agglomerative where it works in a "top-down" manner. It begins with all objects in a single cluster and separates the most heterogeneous objects. The process is iterated until all objects are in their own cluster

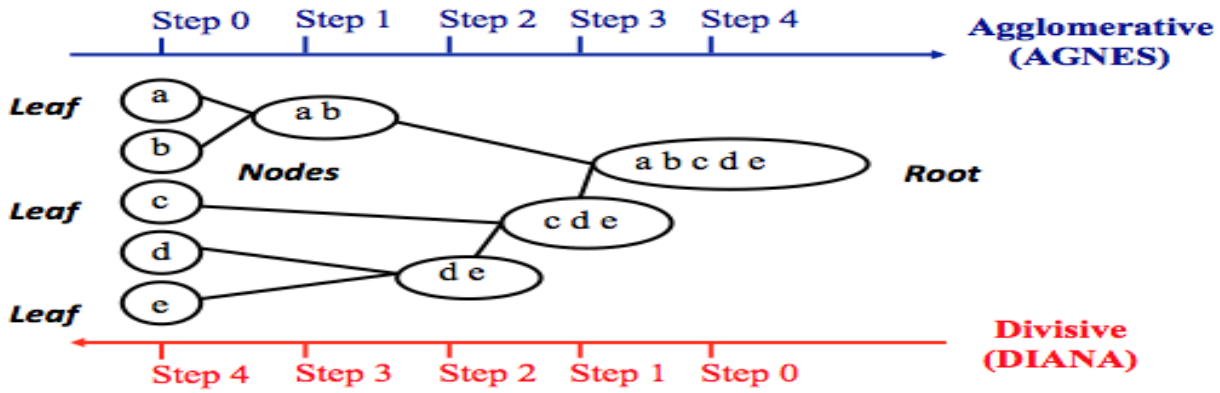


Figure (3.18) agglomerative and divisive clustering. (Kassambara , 2017)

A graphical representation of the clustering tree is Dendrograms. The height provided on the vertical axis in the Dendrogram indicates the (dis)similarity distance between the clusters. The higher the height is the less similar the objects are. This height is called the *cophenetic distance* between the two objects. The dendrogram can be cut at a certain height to define a number of clusters. Another way to cut the tree is to view the percentage increase in the agglomeration coefficient which measures the increase in the

heterogeneity within clusters, so if there is a high increase in the coefficient this means the cluster merging should stop as it will be merging different clusters.

To measure how well the cluster tree generated reflects the data is to compute the correlation between the *cophenetic* distances and the original distance. High correlation reflects the higher accuracy of the model. (Hastie, Tibshirani, & Friedman, 2017)

3.5.3.2 K-means clustering

K-means algorithm technique is to partition the dataset into *K* pre-defined clusters, It tries to make the within-cluster data points as similar as possible while the different clusters as far as possible.
As a first step, Cluster centroid should be set in the model by either inserting predefined values of the clusters' centroids or it can be done by shuffling the dataset and assigning k data points for the centroids. The shuffling process to assign the data points can be repeated to make sure there is no change to the centroids.
After setting the centroids, each data point is assigned to the closest cluster. (Kassambara , 2017)
Note: centroids for the clusters are the average of data points in each cluster.

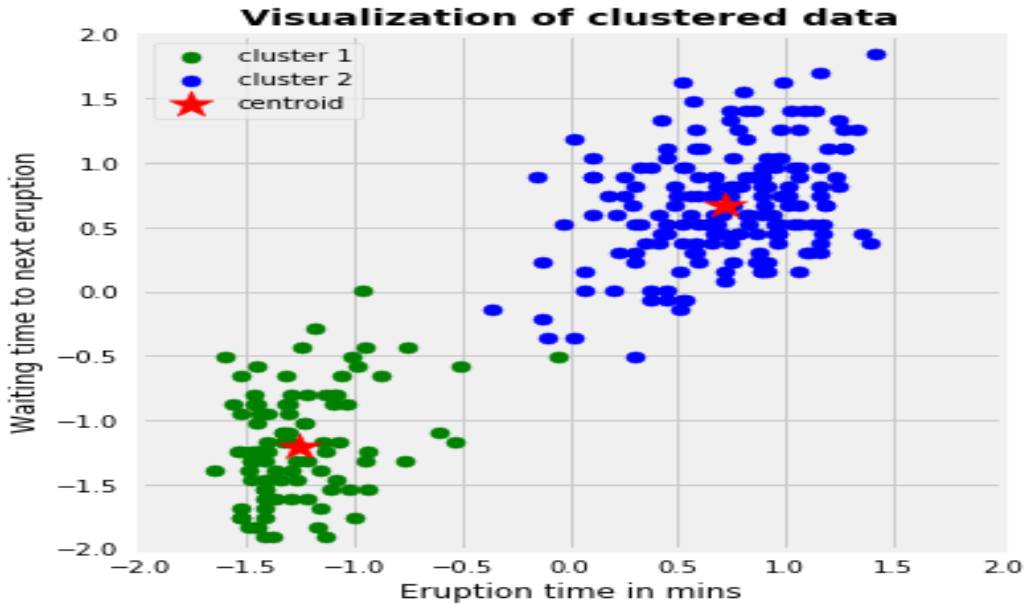


Figure (3.19) Visualization of clustered data (Dabbura, 2018)

To evaluate how well the model performs based on different *K* clusters we can use the elbow method. This method is based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. A suitable number of clusters can be chosen when SSE starts to flatten and makes the shape of an elbow. (Dabbura, 2018)

3.5.3.3 Local outlier factor (LOF)

Local outlier factor (LOF) is an algorithm that identifies the outliers in a dataset, it is derived from the DBSCAN algorithm. In LOF the objective is to find observations that are not alike. LOF allows defining outliers by comparing the local density of an object to the local densities of its neighbors and points that have a lower density than their neighbors are considered to be outliers. It produces an anomaly score by measuring the deviation of a point from its local neighborhood which is called the local density. Local density is determined by calculating distances between data points that are neighbors (k-nearest neighbors). (Aggarwal, 2015)

The primary hyperparameter in LOF is *k*, the number of neighbors. The LOF method scores each data point by computing the ratio of the average densities of the neighbors to the density of the point itself.

By comparing the densities points with similar densities to its neighbors have a LOF approximate to 1, The points with lesser densities than their neighbors are considered outliers and have a high LOF. Usually, if LOF is greater than 1 it can be considered as an outlier, but 1 is not necessarily the threshold value of the LOF, it is dependent on the case and the dataset, but we can say that the higher the LOF is the most probable it is an outlier.

3.5.3.4 Automatic PAM clustering for outlier detection

PAM (Partition Around Medoids) clustering algorithm is a clustering technique used to find Clusters that have minimum average dissimilarity between objects that belong to the same cluster.

PAM is more robust compared to k-means as it handles noise better, but its main disadvantage is that it needs a high computational overhead.

In the k-medoids method (PAM) each cluster is represented by a medoid which is the most centrally located point within the cluster.

To estimate the best number of clusters average silhouette method is used which measures the quality of a clustering. A high average silhouette width indicates a good clustering. The best number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. (Kassambara , 2017)

The methodology of outlier detection comprises two phases:

1. clustering

2. finding outlying score using Silhouette values

Silhouette takes into account both the average distance of a point to other points in its cluster & separation which is the average distance to all points in the nearest cluster. Negative Silhouette values can be assumed as outliers because they don't fit well to the cluster they were assigned to. (Batool & Hennig, 2021)

3.5.3.5 Isolation Forest (IF)

Isolation Forests are a type of algorithm used for outlier detection and are an unsupervised tree-based model, they are similar to Random Forests. It doesn't use distance or density measures. It can handle high-dimensional data, doesn't require high memory, and doesn't need much computational power.
Isolation Forests are a group of binary decision trees. At first, a random sub-sample of the data is selected and assigned to a binary tree, it next branches by selecting a random feature and using a random threshold of the selected feature. If the value is less than the threshold it goes to another branch, and accordingly the node is split. The process continues to construct random binary trees. (Bai, 2021)

During scoring an 'anomaly score' is assigned to each of the data points. A score close to 1 means that the data point is more likely to be an outlier while a score of 0.5 or less means that the observation is more likely a normal observation.

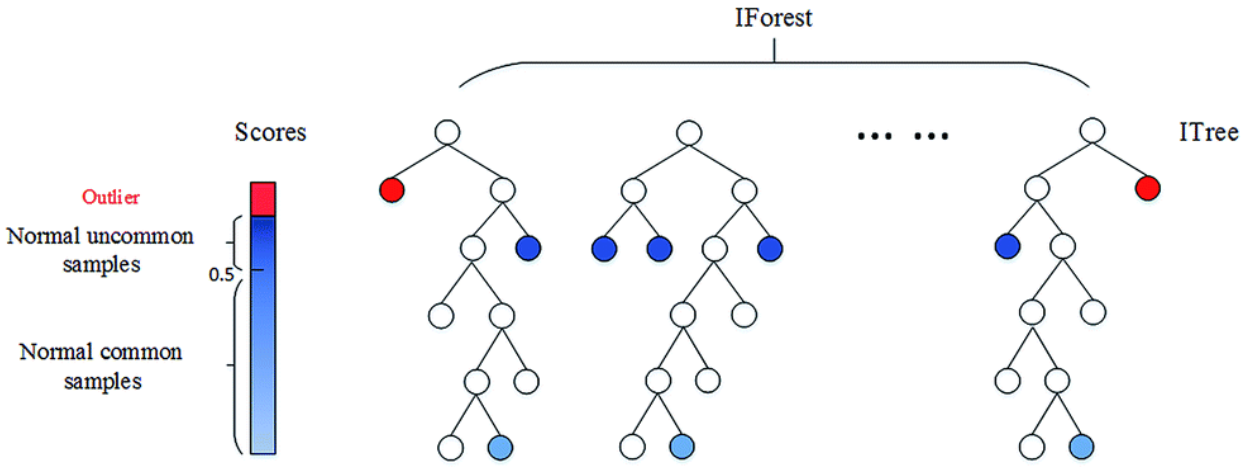


Figure (3.20) Isolation Forest (Bai, 2021)

3.6 Performance measures of Machine Learning algorithms

To measure which model performs better after evaluating their significance and making sure all assumptions are met properly, some measures are used for comparison based on the type of machine learning algorithm.

3.6.1 Classification measures

Some of the most common classification measures that are used to evaluate and compare between classification algorithms are:

- **Accuracy**: it is the ratio of the number of correct predictions to the total number of the input sample

  Accuracy = $\frac{\text{number of correct predictions}}{total\ number\ of\ input\ samples.}$

- **True Positive Rate (Sensitivity= Recall)**: It is the ratio of positive data points that are correctly considered as positive, with respect to all positive data points.

  Sensitivity = $\frac{\text{true positives}}{false\ negatives + true\ positives}$

- **True Negative Rate (Specificity):** It is the ratio of negative data points that are correctly considered as negative, with respect to all negative data points.

  Specificity= $\frac{\text{true negatives}}{false\ positives + true\ negatives}$

- **Precision:** It is the number of true positive results divided by the number of positive results predicted in the model including the false positive.

  Precision= $\frac{\text{true positives}}{false\ positives + true\ positives}$

- **F1 score:** It is the Harmonic Mean between precision and recall, and tries to find the balance between precision and recall. The range for F1 Score is [0, 1], The greater the F1 Score, the better is the performance of the model

  $$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

3 .7 Applied machine learning algorithms in the study

Data mining is extracting knowledge from a large amount of data. After selecting relevant variables, suitable machine learning algorithms should be applied and evaluated.

3.7.1 Cluster analysis

To understand the customers more and define the characteristics of those who cause higher costs than others unsupervised machine learning algorithms were applied and tested. The main objective is to group the customers based on the associated costs and define a suitable number of clusters.

The clusters were built based on the 3 variables:

- Policy cost
- Outpatient costs
- Inpatient costs

To group the customers, cluster analysis was applied and the following two algorithms were used:

- Hierarchical clustering
- K-means clustering

ANOVA and Chi-square tests were applied to profile the clusters and understand the characteristics of each cluster

The output of the cluster analysis assists in understanding how to group customers based on their total costs and helps in defining the most suitable number of risk levels.

3.7.2 Risk assessment

The main objective for predicting the risk level of the new customers is to automate and enhance the underwriting process and achieve higher accuracy in estimating the expected claims' costs by using machine learning algorithms.
The underwriting process is currently done by insurance specialists who gather information about the new clients who request health insurance for their employees and their families and estimate the claims' costs for the requested coverage. Gathered information includes main demographics such as age, gender, marital status, number of children, history of chronic diseases, and number of subscribers who wear glasses. The underwriter analyzes the gathered information and proposes a rate for the premium cost.

Applying machine learning algorithms will assist in the underwriting process, reduce time and effort and achieve higher accuracy. The output of the machine learning algorithms

helps the underwriter in setting the most suitable premium rate, reducing risk, and acquiring new clients by offering competitive prices.

Data was prepared to be used in the predictive models, and subscribers were labeled as per their risk level which was used as the dependent variable in the model.
Since the dependent variable is multiclass with 3 levels <High-Risk, Mid-risk, Low-risk >, then the most suitable machine learning algorithms to be used are supervised classification algorithms.

The independent variables used in the classification models are:

Independent variables
• Gender
• Age
• Marital Status: Married, Single
• Subscriber type: corporate employee, Spouse, Children, New-born (added within 2020)
• Number of children
• Age of oldest child
• Has chronic Diseases: yes, no
• Number of Chronic Diseases
• Had Old surgery (during last year): yes, no
• Wear glasses: yes, no
• Corporate account Type: NGO, educational, medical institution, a private company
• New/Old account: started with the insurance company before 2020/or in 2020
• Account Policy coverage: Mid/ High based on the agreement with the Insurance company

In this study, different algorithms were implemented on the data set that was prepared for the risk prediction models. Each algorithm has its pros and cons as described in chapter 4.
The following algorithms were applied and evaluated:

- Multinomial Logistic regression
- Decision tree
- Random Forest
- Neural classifiers
- SVM

Predicting the number of customers in each risk level will assist in setting a pricing model that will help in the underwriting process and eventually ensure profit. This methodology in insurance underwriting leverages the underwriter to underwrite a greater number of policies in lesser time with higher accuracy.

After choosing the best algorithm for predicting the risk, the predicted cost will be estimated by calculating the weighted average cost for the new insured account based on the predicted risk as follows:

(Number of expected High-Risk subscribers *average cost of High-risk + Number of expected Mid Risk* average cost of Mid-risk + Number of expected Low Risk * average cost of Low-risk) % Number of subscribers

3.7.3 Fraud detection

The data received from the insurance company doesn't have a flag for fraud/abusers then the methodology that was used to detect fraud is unsupervised learning Outlier Detection. The algorithms were applied to help detect suspicious claims that require review and audit and therefore save cost and time and increase fraud detection accuracy. The insurance specialist has to review the outliers instead of having to look at the whole claims or at a random sample.

Three approaches were applied to detect the outliers:

- Local Outlier Factor (LOF)
- Automatic PAM clustering algorithm for outlier detection (APCOD)
- Isolation Forest (IF)

The Outlier Detection algorithms specify outliers in the dataset that are far from the centers of the clusters/ neighborhoods that they belong to. The model will help detect suspicious claims with abnormal behavior that require review and audit.

The dataset was prepared where the control variable is the claim including only claims of clinical visits, where the objective is to detect the outliers in the claims.

The attributes that were included in the outlier detection algorithms are:

- ❑ Claim diagnoses type (21 categories based on ICD10 main categories such as
  - ✓ Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
  - ✓ Diseases of the circulatory system
  - ✓ Diseases of the digestive system
  - ✓ Diseases of the ear and mastoid process
  - ✓ Diseases of the eye and adnexa
  - ✓ Diseases of the genitourinary system
  - ✓ Diseases of the musculoskeletal system and connective tissue
  - ✓ Diseases of the nervous system
  - ✓ Diseases of the respiratory system
  - ✓ ……

- ❑ Medicines cost for the same claim
- ❑ Laboratory cost for the same claim
- ❑ X-ray cost for the same claim
- ❑ Procedures cost for the same claim
- ❑ Time interval from the last claim for the same customer.
- ❑ Same provider for the previous visit: Yes/No

Each Diagnosis has an average cost for medicines, X-rays, procedures & Laboratory costs so outlier detection techniques will be able to detect the outliers that have much higher costs than the norm to be further investigated.

In addition, on average the time interval between claims visits is 128 days, some diagnoses types have lower time intervals, so outlier detection techniques will help in detecting intervals that are far from the norm. for example, within a few days or on the same day.

After detecting the outliers in claims data, an investigation is to be done and some questions to be answered.

- Why do these claims cost higher for the same diagnoses type?
- Is there a significant frequency for some doctor/s in the outliers?
- Are there subscribers who have a high percentage of outliers claim cost/ their total claim cost.

Chapter Four

Results and discussion

4.1 Introduction

There are two main objectives in this study, the first one is to predict the risk and accordingly predict the expected cost per client. This will assist insurance underwriters in recommending the most adequate premium rate for new clients.
And the second main objective is to detect the fraud or abuse risk, this is done by detecting the abnormal claims/ outliers where these outliers can be investigated for fraudulent activity by insurance specialists, assist in discovering new fraud/abuse patterns, and help in setting new policies and track suspicious customers or medical providers.

Before starting the models some explorative analysis was done as explained in chapter 3 which helped to give an overall image of the important attributes. Another way to understand the clients base and define the characteristics of those who cause higher costs is to segment them into clusters, so as a start a Cluster Analysis was done to assist in understanding the current customer base and provide insights that will help in setting marketing strategies.

4.2 Cluster analysis

The variables that were used to cluster the customers are Outpatient sum of claims, Inpatient sum of claims, policy-cost.
Two clustering algorithms were used. First agglomerative Hierarchical clustering was used since it is an explorative algorithm and will give us an indication of the number of clusters. The algorithm will try to group the most similar points and form clusters.

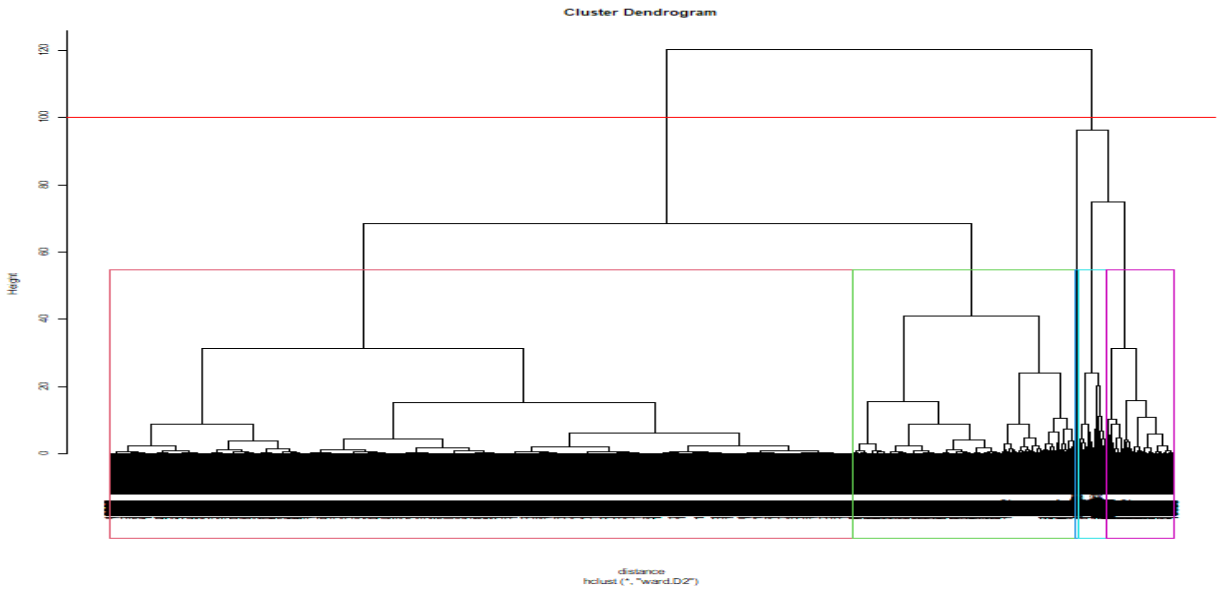


Figure (4.1) Cluster Dendrogram for Hierarchical Clustering by (Researcher)

The above chart is a dendrogram which is a graphical representation of the clusters. The height provided on the vertical axis in the Dendrogram indicates the dissimilarity distance between the clusters, so after setting height 50 where objects are closer together it is noticed that 4 clusters would be appropriate.

The agglomerative coefficient is 0.7023861, which measures the amount of clustering st ructure found, and since the coefficient is closer to 1 this suggests a strong clustering st ructure.

The customers were separated into the following 4 clusters where cluster 1 is considered the lowest cost and cluster 3,4 are the highest.

Table (4.1) Cluster Output for Hierarchical Clustering

cluster	number of customers	mean outpatient cost	mean inpatient cost	mean number claims
1	6354	277	146	4
2	3485	1053	113	13
3	980	2,616	1073	30
4	24	2,059	17,684	19

As a second clustering method, a K-means clustering algorithm was used. The output of the hierarchical clusters' mean was used as centroids for the k-means model to get higher accuracy, and the number of clusters that were used is 4.

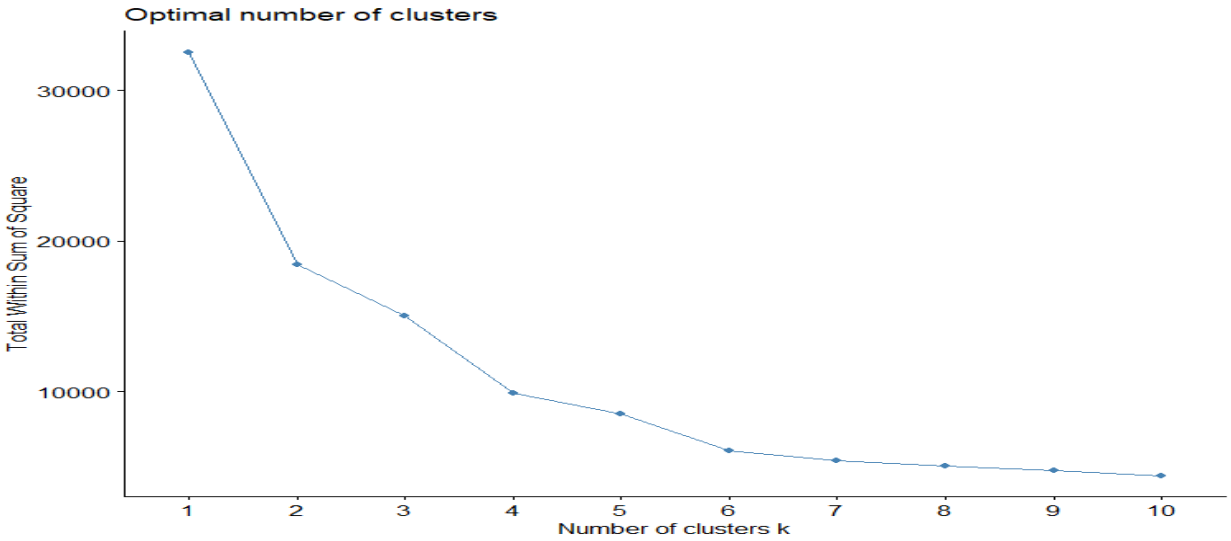


Figure (4.2) Elbow method for K-means Clustering by (Researcher)

The elbow method as shown in the above shape also indicated that 4 Or 5 clusters are a suitable number where the Total sum of squares starts to flatten at cluster 4.

The following is a graphical visualization of the clusters.

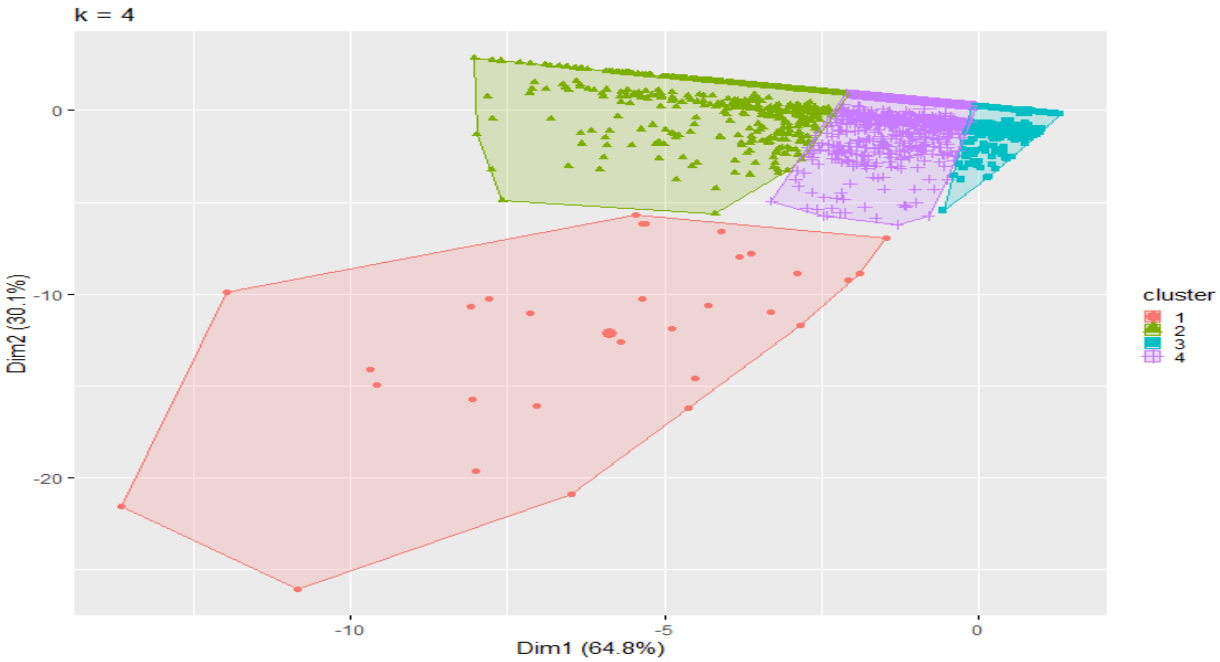


Figure (4.3) Cluster Visualizations for K-means by (Researcher)

As noticed cluster 4 has higher dimensions and is clearly the highest cost segment. The following table shows the number of clusters based on the K-mean algorithm.

Table (4.2) Cluster Output for K-means Clustering

cluster	number of customers	mean outpatient cost	mean inpatient cost	mean number claims
1	7,288	328	61	5
2	2,700	1,202	248	15
3	824	2,848	1,457	32
4	31	2035	15,776	19

Cluster 1 has around 67% of customers and they are considered the low-cost segment, cluster 2 is considered mid-high cost, and clusters 3 & 4 are the very-high-cost segments.

To understand the characteristics and profile of each cluster some descriptive statistics was done, and to check if the differences in means or frequencies were different ANOVA test & Chi-square tests were done and the p-values were less than 0.05 which indicates the differences in means between the clusters for the numerical variables are significant, and the frequency distribution for categorical variables are also different between the clusters and differences are shown in the following table:

Table (4.3) Frequency tables and means of the independent variables for K-means Clusters

				Subscriber				Chronic diseases		Gender	
Cluster	count	Mean Age	Mean # chronic	Child	Employee	newborn	Spouse	No	Yes	F	M
1	7,288	21	0	**57%**	28%	1%	15%	99%	1%	48%	52%
2	2,700	31	0.3	20%	44%	3%	34%	89%	11%	55%	45%
3	824	41	1.7	3%	64%	0%	33%	80%	**20%**	53%	47%
4	31	35	1.4	3%	36%	**29%**	32%	84%	16%	32%	68%

From the above table, it is noticed that cluster1 has a higher percentage of children than other clusters, with only 1% having chronic diseases and 66% being single.
As for cluster 2 children percentage is much lower only 20%, and the percentage of customers with chronic diseases is 11% which is much higher than cluster 1, also females' percentages are higher in this cluster.
Cluster 3 has a much higher age, with a very low percentage of children & most of this cluster are married at 67%.
Cluster 4 can be considered an outlier cluster where it is a very small cluster and customers have very high inpatient costs.

Clusters can be summarized as follows:

Table (4.4) Profiling Cluster Output

Cluster	Description	Count of subscribers	Percentage of subscribers	Conclusion
1	Low-cost	7,288	71%	Younger segment mostly children, with no chronic diseases & mostly single
2	Mid- Cost	2,700	22%	The mid-age segment is mostly married adults with some having chronic diseases
3	High-Cost	824	5%	Older segment most have chronic diseases with average 2 chronic diseases
4	High-Cost (V. High Inpatient Cost)	31	2%	This is a very small cluster with very high cost, which is probably an outlier cluster
Total		**10,843**		

The cluster analysis assisted in predicting the number of clusters based on the cost variables, the output showed that there are 3 main clusters each having different characteristics.
Using the output of the cluster analysis the researcher decided to group the customers into 3 main groups based on the risk level (High, Mid, Low)
Higher risk is mainly customers who need inpatient/ hospitalization care, while Mid Risk has high outpatient costs with much lower Inpatient cost, and the low-risk cluster has the lowest cost in both inpatient & outpatient costs.

4.3 Risk assessment & cost prediction

To set the right premium rates for a new account/company requesting a Health Insurance policy, the insurance underwriters evaluate the risk based on the companies' profile and predict the cost.
The objective of this study is to use machine learning algorithms using the current customers' base in order to predict the cost for a new account.

Based on the Cluster Analysis outputs, the Risk level is distributed into 3 main categories:

- High Risk – Customers having Inpatient cost < who require overnight hospitalization including admissions, operations, and surgeries>
- Mid Risk- Customers have high outpatient costs and are considered non-profitable to the insurance company. Claims costs are higher than the policy cost.
- Low Risk- Outpatient cost includes all other medical expenses such as clinic visits, medicines, clinical procedures, laboratory, X-rays, dental, optical, and physical therapy are less than the policy cost.

The following shows the distribution of customers and costs based on the new dependent variable risk level

Table (4.5) Summary of the Risk-factor

Risk Level	Number of customers	Average Inpatient Cost	Average Outpatient Cost	Average Total Cost
High Risk	1196	2,345	1,364	3,709
Mid Risk	1348	-	2,089	2,089
Low Risk	8299	-	433	433
Total	**10843**	**259**	**742**	**1,000**

To predict the cost two approaches were tried, first regression algorithms were applied to predict the total cost (Inpatient +outpatient costs) together. The results were not adequate and R-square was around 32% only.
The variations in the dependent variable (total cost) were too high and there exist lots of outliers especially for customers having inpatient hospitality costs.

An alternative approach was used to reach higher accuracy is to predict the Risk Level and then use the outcome to calculate the weighted average cost for the new insured account.

(Number of expected High Risk *3,709 ILS + Number of expected Mid Risk*2,089 ILS + Number of expected Low Risk * 433 ILS) % Number of subscribers

4.3.1 Predict risk level using classification models

This section includes output details and accuracy measures for the classification algorithms that were used to predict the risk level.

4.3.1.1 Multinomial logistic regression

Multinomial logistic regression was applied using all independent variables, and as noted in the Analysis of Deviance all variables were significant except the policy coverage.

Analysis of Deviance Table (Type II tests)

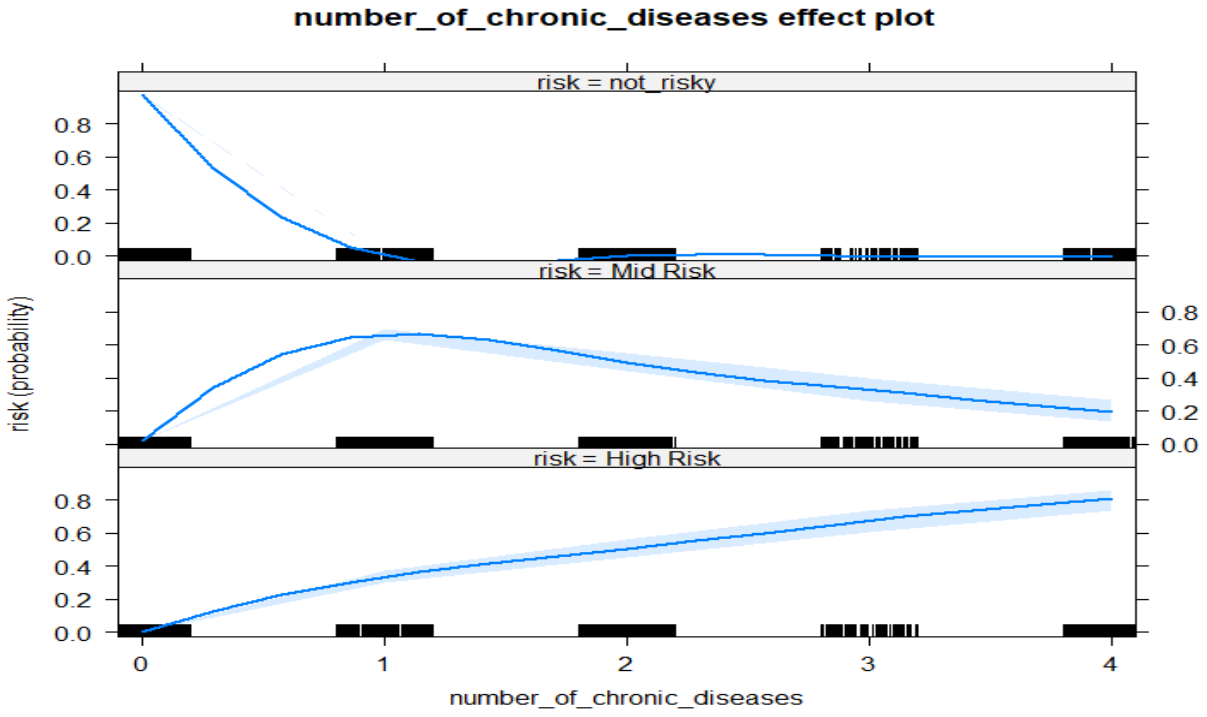
Response: risk

	Chisq	Df	Pr(>Chisq)
new_old_account	6.4	2	0.040766 *
company type	21.19	6	0.001695 **
policy coverage	0.86	2	0.651403
has chronic	153.77	2	< 2.2e-16 ***
subscriber	1194.31	6	< 2.2e-16 ***
wear glasses	830.48	2	< 2.2e-16 ***
old surgery	660.41	2	< 2.2e-16 ***
age	50.18	2	1.268e-11 ***
gender	263.14	2	< 2.2e-16 ***
marital status	238.03	2	< 2.2e-16 ***
number_of_kids	58.76	2	1.740e-13 ***
number_of_chronic_diseases	157.63	2	< 2.2e-16 ***
age_of_oldest_kid	108.34	2	< 2.2e-16 ***

- Naglkerke R square, which measures the goodness of model fit and describes the proportion of variance that the model successfully explains is 67% which is good.

The model uses a Baseline-Category Logit Model, which means it represents the summary of the odds in one category relative to the baseline category which is in this case the “High-Risk Level”.

The following effect plot is a sample of how each independent variable affects the probability of each Risk Level.



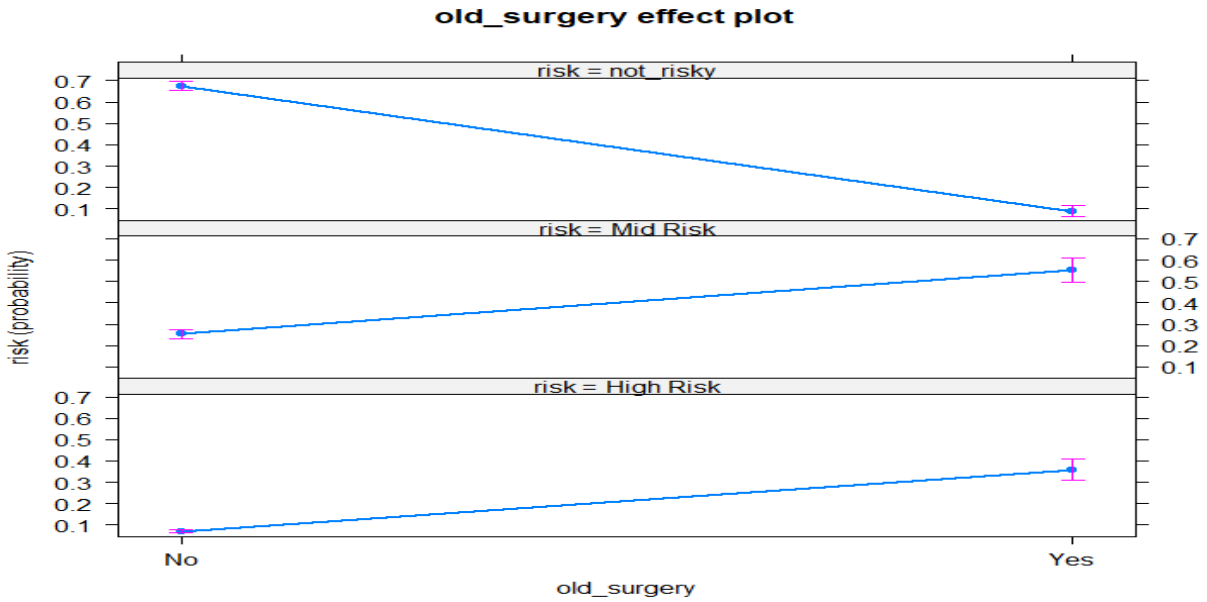


Figure (4.4) Effect plot of the independent variable on the Risk-factor by (Researcher)

The effect plots can show how the change in the independent variable might affect the response probabilities and it is based on the used model.

- Output of the Multinomial Logistic Regression:

Confusion Matrix

Table (4.6) Confusion Matrix for Multinomial Logistic regression

	Reference		
Prediction	**High Risk**	**Mid Risk**	**Low Risk**
High Risk	**404**	156	95
Mid Risk	207	**792**	48
Low Risk	585	400	**8156**

The Main accuracy measures were as follows:

Table (4.7) Accuracy & Sensitivity for Multinomial Logistic regression

	Class		
	High Risk	**Mid Risk**	**Low Risk**
Sensitivity	0.33779	0.58754	0.9828
Overall Accuracy	**0.8625**		

Multinomial Logistic Regression didn't perform well in predicting the High & Mid Risk Classes, and since these classes are the most important in predicting the right price <premium rate>, then this model won't be appropriate.

4.3.1.2 Decision tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split into nodes. The algorithm adds a node to the model every time that an input column is found to be significantly correlated with the predictable column.

Decision trees can create complex trees and cause overfitting, that's why some pruning was done to avoid the overfitting; the minimum number of samples in the leaf node was set to be 30 observations, and Early stopping was set as True.

The decision tree model was fitted using Quinlan's C5.0 algorithm which uses Gain Ratio instead of Information Gain. The algorithm handles continuous numeric data, handles missing data, is capable of pruning, and includes a way of addressing "rare" cases. (Fisher, 2020)

The used attributes in the model are

- ✓ 100.00% old_surgery
- ✓ 100.00% has_chronic
- ✓ 100.00% number_of_chronic_diseases
- ✓ 100.00% age
- ✓ 99.34% marital_status
- ✓ 97.66% subscriber
- ✓ 96.84% wear_glasses
- ✓ 92.82% policy_cost
- ✓ 92.67% gender
- ✓ 55.66% number_of_kids
- ✓ 31.30% company_type
- ✓ 27.35% age_of_oldest_kid
- ✓ 13.18% new_old_account
- ✓ 7.78% policy_coverage

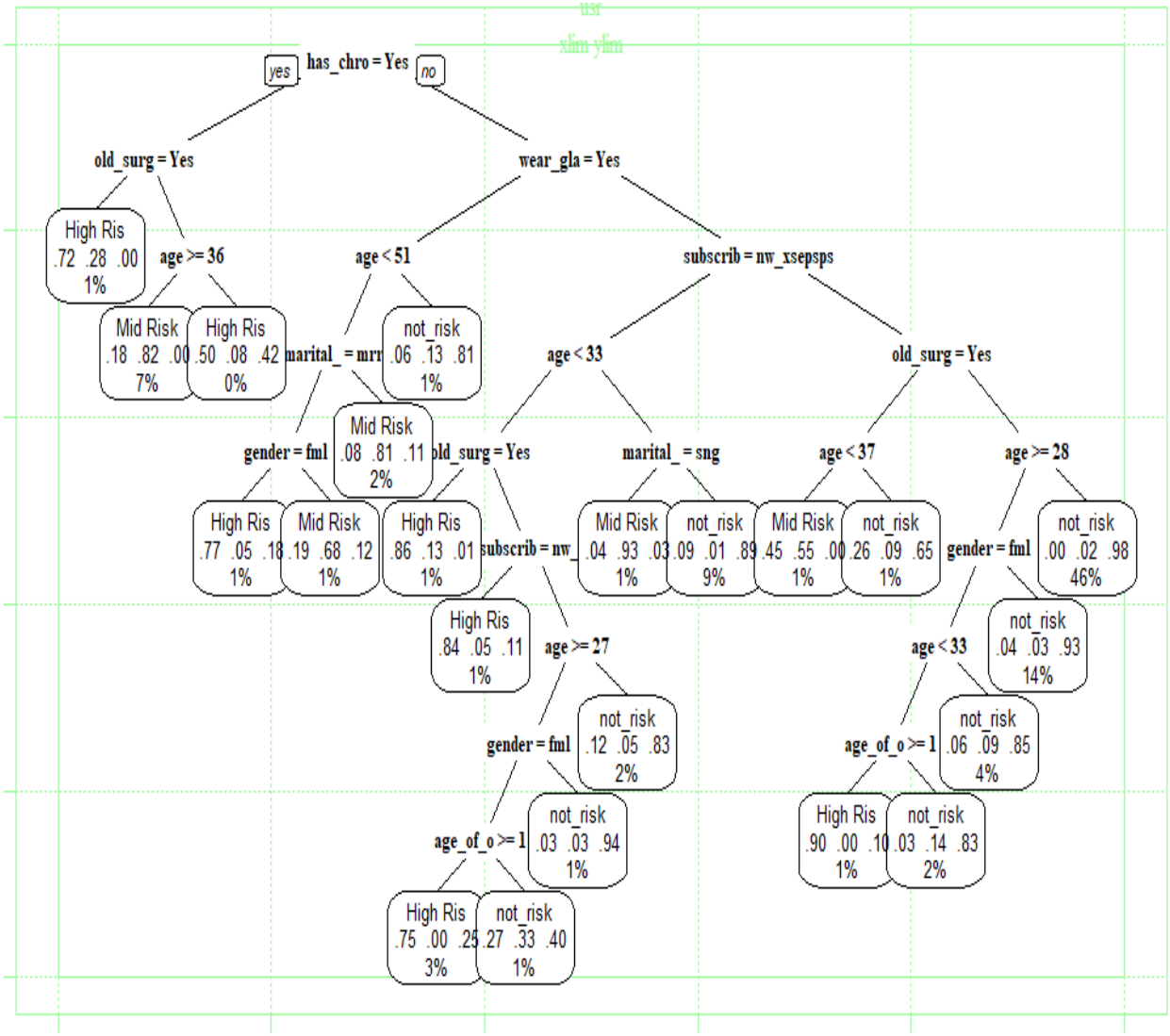


Figure (4.5) Decision Tree Output by (Researcher)

As shown in figure (4.5), the tree starts with the attribute has-chronic, if a customer has chronic diseases and had old surgery then the customer is considered a High-risk.
The data is continuously split into nodes according to the given conditions, and the process continues until reaching the final nodes where nodes can't be classified further.

Confusion Matrix

Table (4.8) Confusion Matrix for Decision Tree

	Reference		
Prediction	**High Risk**	**Mid Risk**	**Low Risk**
High Risk	**817**	44	52
Mid Risk	141	**1062**	42
Low Risk	238	242	**8205**

The Main accuracy measures were as follows:

Table (4.9) Accuracy & Sensitivity for Decision Tree

	High Risk	Mid Risk	Low Risk
Sensitivity	0.683	0.787	0.9886
Overall Accuracy	**0.93**		

Accuracy and sensitivity are higher than Multinomial Logistic regression.

4.3.1.3 Random forest

A random forest fits several decision trees on various sub-samples from the dataset and then averages to improve the accuracy.
The model was fitted using randomForest which implements Breiman's random forest algorithm. It fits many classification or regression trees (CART) models to random subsets of the input data and uses the combined result for prediction. One of its main features is the ability to estimate the importance of each predictor variable in modeling. (Breiman, 2001)

To have good accuracy two parameters mtry and ntree should be tuned

- mtry: Number of variables randomly sampled as candidates at each split.
- ntree: Number of trees to grow.

There are other parameters, but these are the most likely to have the biggest effect on your final accuracy.
The following chart shows that the error for all dependent variable classes stabilizes after a few trees, so choosing 500 or 1000 trees would be good enough.

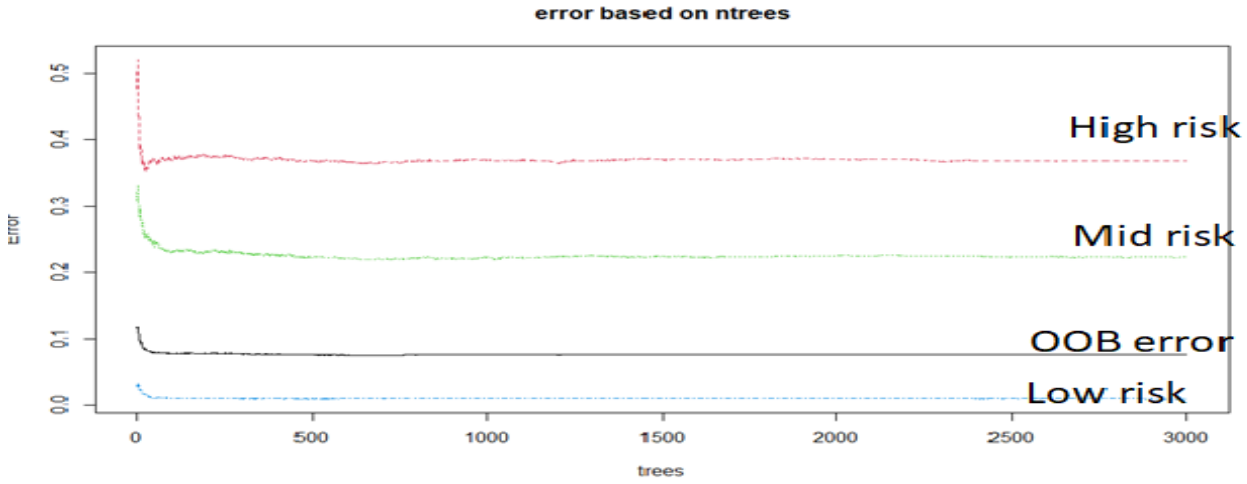


Figure (4.6) Error based on number of trees in Random Forest by (Researcher)

The best mtry to choose is 4 since it has the least OOBError

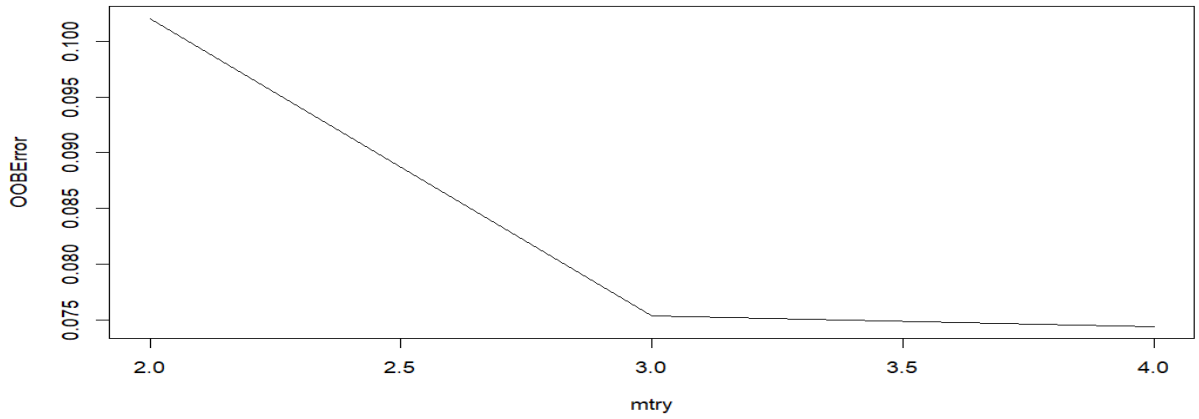


Figure (4.7) Error based on mtry in Random Forest (Researcher)

After choosing the parameters the Random Forest was trained, we got the following Confusion Matrix

Table (4.10) Confusion Matrix for Random Forest

	Reference		
Prediction	**High Risk**	**Mid Risk**	**Low Risk**
High Risk	**930**	12	22
Mid Risk	54	**1112**	30
Low Risk	212	224	**8247**

The Main accuracy measures were as follows:

Table (4.11) Accuracy & Sensitivity for Random Forest

	Class		
	High Risk	**Mid Risk**	**Low Risk**
Sensitivity	0.77759	0.8249	0.9937
Overall Accuracy	**0.9489**		

The following chart shows the mean decrease Gini which is how much the model fit decreases when you drop a variable. The greater the drop the more significant the variable is so it shows the importance of each variable in the forest.

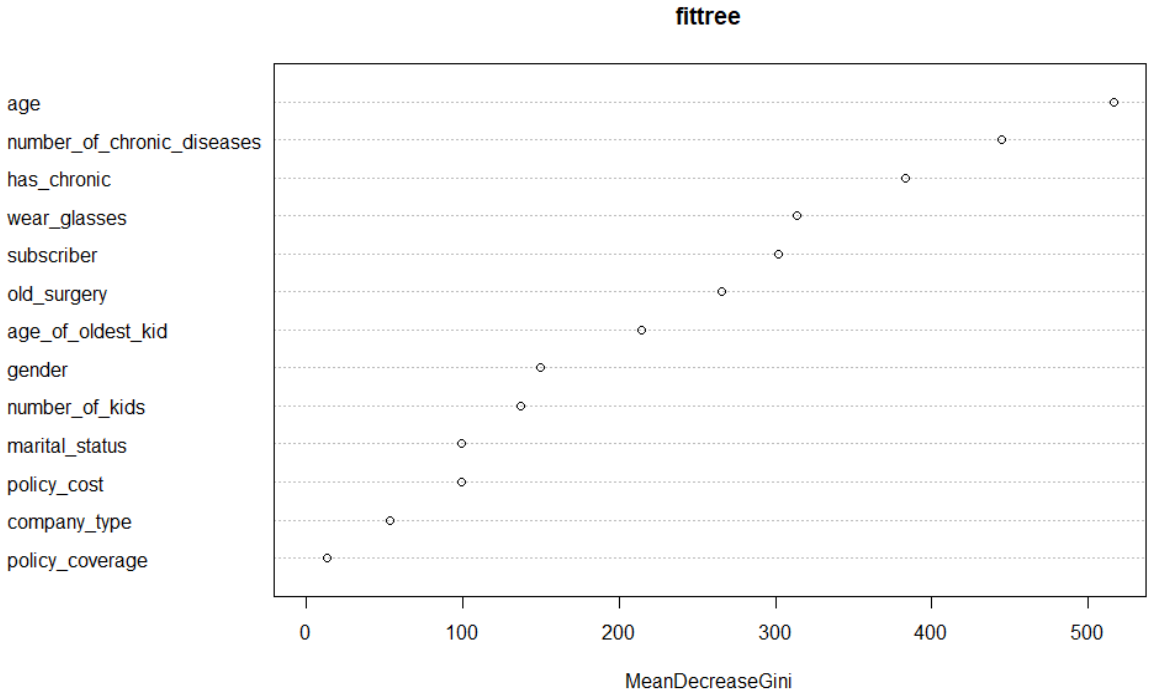


Figure (4.8) Mean Decrease Gini in Random Forest for variable importance by (Researcher)

4.3.1.4 Neural network classifier

Neural Network keeps adjusting the weights to minimize the error and predict the correct class label. In addition, it can predict nonlinear relationships.

A Neural Network classifier was applied to predict the risk level. Multilayer Perceptron was applied and the dataset was normalized before applying the model.
The model was fitted using neuralnet package. The package demands an all-numeric matrix or data frame. It allows flexible settings through custom-choice of error and activation function. (Fritsch, Guenther, & Wright, 2019)

There are 1 hidden layer with 5 nodes, where 5 nodes got the highest accuracy after using systematic experimentation on the number of nodes between (3-7), as for the number of hidden layers the computational power didn't allow to perform more hidden layers, and that might be the reason for not having higher accuracy.

Neural Network got higher overall accuracy than Multinomial and decision trees but Random Forest has got the best accuracy.

Confusion Matrix

Table (4.12) Confusion Matrix for Neural Network

	Reference		
Prediction	**High Risk**	**Mid Risk**	**Low Risk**
High Risk	**805**	75	203
Mid Risk	177	**1017**	65
Low Risk	214	256	**8031**

The Main accuracy measures were as follows:

Table (4.13) Accuracy & Sensitivity for Neural Network

	Class		
	High Risk	**Mid Risk**	**Low Risk**
Sensitivity	0.673	0.754	0.967
Overall Accuracy	0.908		

The following chart shows the importance of the input variables for the risk-level: High Risk

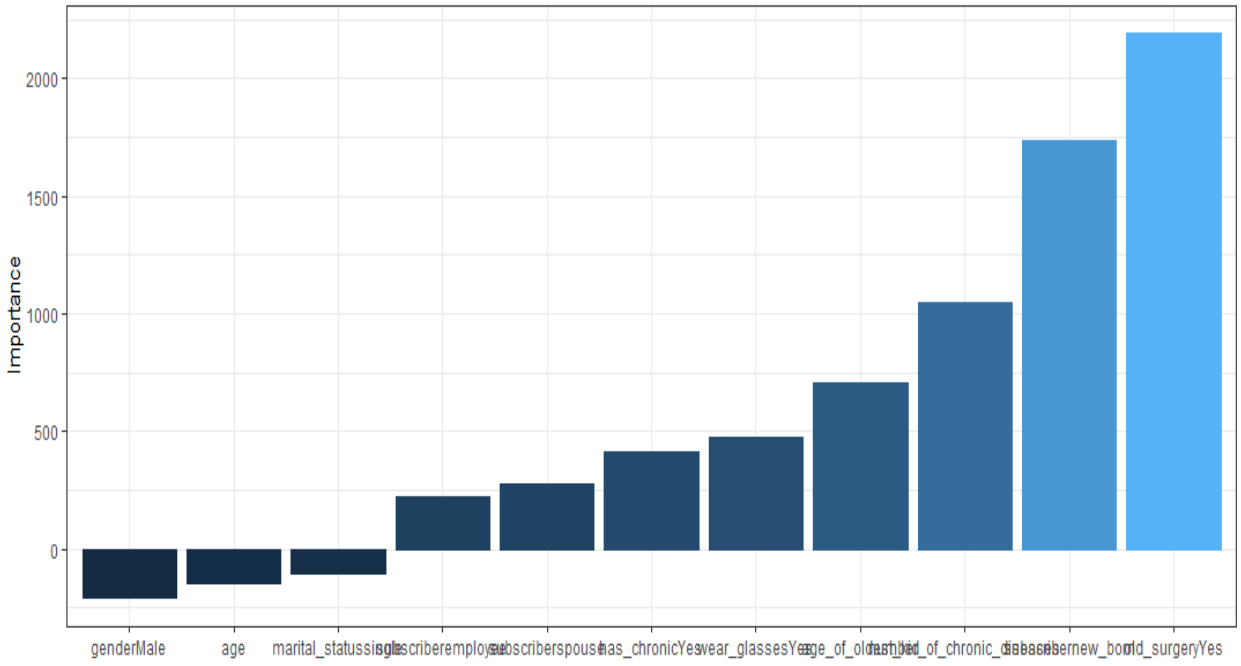


Figure (4.9) Variable importance for High-Risk response in Neural Network by (Researcher)

The highest important variables are Old_surgeryYes, New_born, number of chronic diseases.

4.3.1.5 Support vector machines

SVM algorithm maps the data points to a higher dimensional space and then finds the hyperplane that maximizes the margin which is called a Kernel.

The model was fitted using caret package with using the predictive model SVM-Radial. caret uses an analytical formula to get reasonable estimates of sigma and fix it to that value. In addition, caret cross-validates over a set of cost parameters C. (RPubs, n.d)

The effectiveness of the SVM model depends upon choosing the Kernel, its parameters and the soft Margin Parameter C.

In this dataset, the most suitable Kernel was Radial with the highest accuracy. Also, some parameters were set in the Control which are:
Resampling Method=repeated cross-validation
Number of resampling iterations=10
The “repeats” parameter= 3, this parameter contains the complete sets of folds to compute for our repeated cross-validation.

As for the cost parameter C, it decides how much an SVM should be allowed to bend with the data, so it is the tradeoff between misclassification and simplicity of the model.

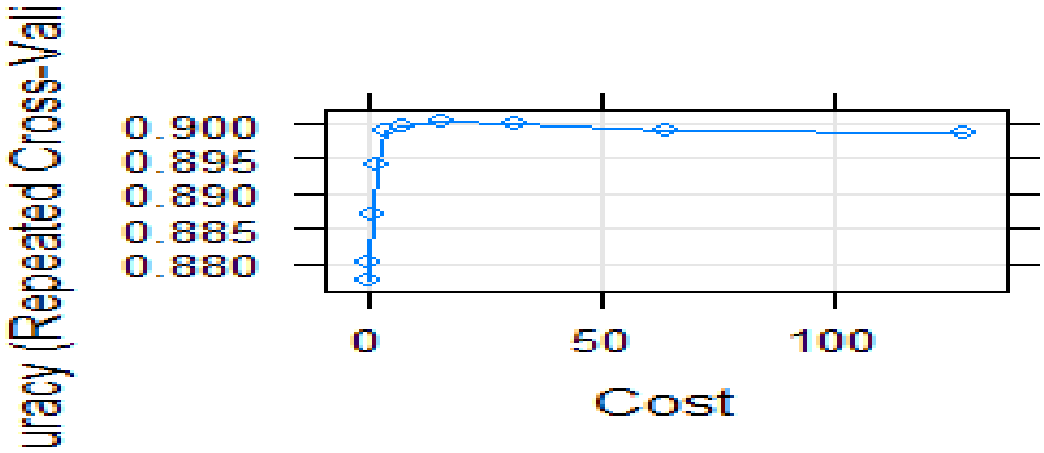


Figure (4.10) Accuracy based on Cost parameter C in SVM-Radial by (Researcher)

Accuracy was used to select the optimal model and the final values used for the model C =18

Confusion Matrix

Table (4.14) Confusion Matrix for SVM

	Reference		
Prediction	**High Risk**	**Mid Risk**	**Low Risk**
High Risk	**869**	48	164
Mid Risk	112	**1067**	36
Low Risk	215	233	**8099**

The Main accuracy measures were as follows:

Table (4.15) Accuracy & Sensitivity for SVM

	Class		
	High Risk	**Mid Risk**	**Low Risk**
Sensitivity	0.72659	0.7915	0.9759
Overall Accuracy	0.9255		

4.3.1.6 Comparison of classification models

The following table shows a comparison between the models that were used to predict the Risk Level of customers, the data was split into training and test data and the following accuracy and sensitivity measures are the output of the model that was trained on the training dataset and predicted on the test dataset:

Table (4.16) Comparison of Models

Algorithm	**Overall Accuracy**	**Sensitivity**		
		High Risk	**Mid Risk**	**Low Risk**
Random Forest	94%	72%	81%	99%
SVM	92%	70%	78%	98%
Decision Tree	92%	67%	78%	99%
Neural Network	90%	64%	76%	97%
Multinomial Logistic regression	86%	36%	61%	98%

The best performing model is Random Forest with Total Accuracy of 94% followed by SVM-Radial.
Overall accuracy is good and will assist the insurance company in predicting an approximate cost when they need to price the insurance premium rate.

Based on the characteristics of the subscribers in a new account, the insurance company will be able to predict the risk level and the approximate cost of a certain account based on the weighted average cost as follows:

= (Number of expected High Risk *average cost of High-risk + Number of expected Mid Risk*average cost of Mid-risk + Number of expected Low Risk * average cost of Low-risk) % Total number of subscribers within an account

= (Number of expected High Risk *3,709 ILS + Number of expected Mid Risk*2,089 ILS + Number of expected Low Risk * 433 ILS) % Number of subscribers

The following table shows the difference between the predicted weighted cost and actual cost for some of the accounts in the dataset:

Table (4.17) Comparison between the actual price and predicted price for a sample of accounts

Company ID	Predicted # of High Risk	Predicted # of Mid Risk	Predicted # of not-risky	Total # of customers	Actual av. cost /account	expected cost /account	difference
7378	137	267	1343	1747	933	943	1%
1835	113	238	1199	1550	980	926	-6%
17592	126	98	906	1130	858	942	10%
28672	95	41	664	800	1,067	907	-15%
8159	67	71	353	491	1,210	1,119	-7%
14612	38	6	168	212	1,192	1,067	-10%
6170	14	11	171	196	892	760	-15%
950	7	27	150	184	828	801	-3%
258	8	17	129	154	703	786	12%
28229	5	11	124	140	717	680	-5%
697	8	7	29	44	1,462	1,292	-12%

The differences between the actual and predicted costs are acceptable for most accounts, noting that some accounts have much higher costs than others which indicates that they have higher risk subscribers.
This proves that machine learning models can assist the Insurance company in setting the right premium rate for new accounts and avoiding high loss and ensuring profitability.

4.4 Fraud detection

This section covers the second objective of the study which is to detect claims' outliers. Unsupervised outlier detection algorithms are used to detect outliers in the claims' datasets aiming to help detect suspicious claims that require review and audit by insurance specialists and therefore will save cost and time and increase fraud detection accuracy in the insurance company.

Three approaches were applied to detect the outliers:

- Automatic PAM clustering algorithm for outlier detection (APCOD)
- Local Outlier Factor (LOF)
- Isolation Forest (IF)

4.4.1 Automatic PAM clustering algorithm for outlier detection (APCOD)

Since the dataset contains both numeric and categorical variables, K-means and Hierarchical clustering can't be used directly, so the approach is to use Gower distance and partitioning around medoids instead of centroids.

The function daisy() in [cluster package] provides a solution (Gower's metric) for computing the distance matrix when the data contain non-numeric columns. Gower distance is a dissimilarity matrix, it works by calculating the distance metric for each variable and then scaled to fall between 0 and 1. Then a linear combination is calculated to create the final distance matrix.

After calculating the dissimilarity matrix, a clustering algorithm is to be used.
Partitioning around medoids (PAM) will be used, it is very similar to K-means clustering but uses clustering around the medoids instead of the centroids which are used in K-means.
Medoids are defined by the observations themselves whereas every observation that yields to the lowest average distance will be assigned as the medoid and the process continues till choosing the best medoids. (Clustering Mixed Data Types in R, 2016)

To select the suitable number of clusters Silhouette width will be used, it measures how similar observation is to its own cluster, where higher values of Silhouette width mean that objects are well matched to their own clusters and poorly matched to the neighbors' clusters.
The following graph shows that 10 clusters is the best number of clusters

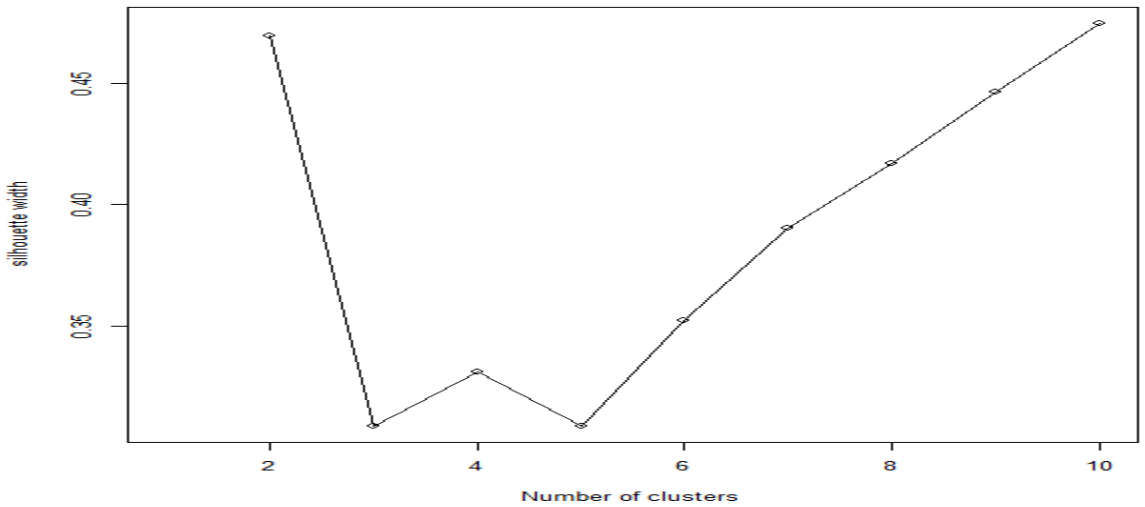


Figure (4.11) Number of clusters based on Silhouette width by (Researcher)

The following graph shows the clusters and it is noticed how there are observations that some observations have negative Silhouette width which identifies that they don't fit the clusters.

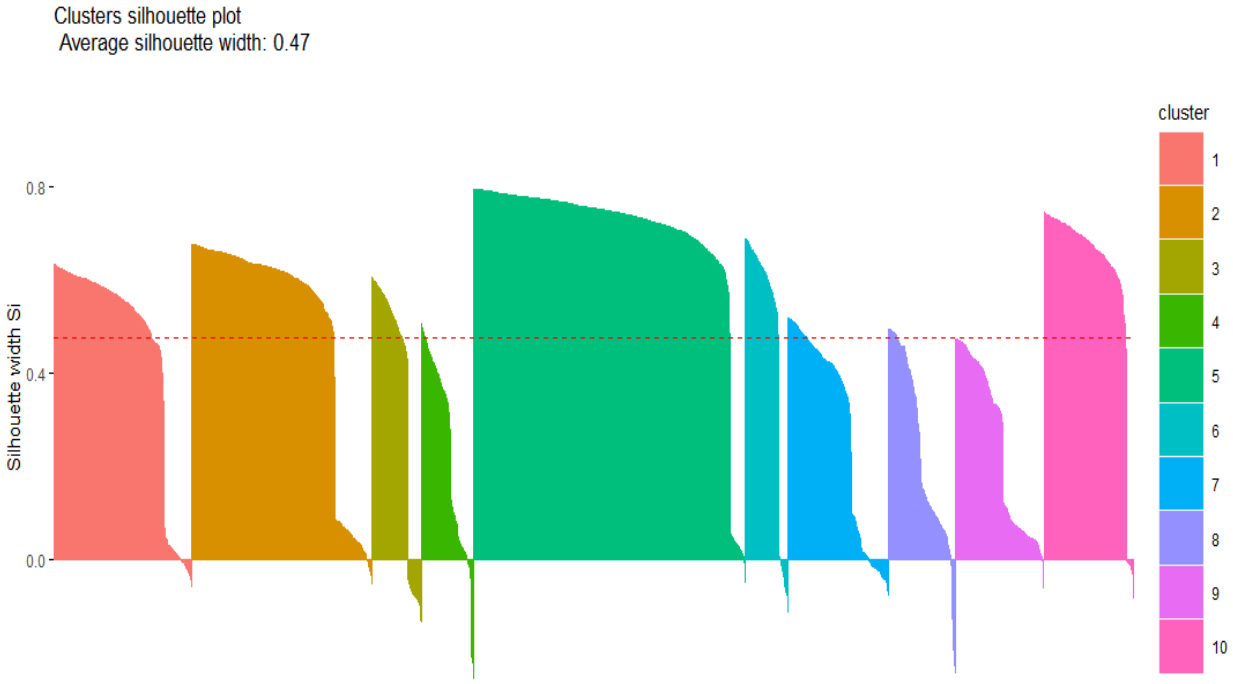


Figure (4.12) Cluster silhouette plot for PAM clustering method by (Researcher)

The following table shows a summary of each cluster:

TABLE (4.18) CLUSTERS SUMMARY BASED ON CLAIMS

	Number of claims in the cluster	Main Disease category in each cluster	number of days between claims	Average claim cost
CLUSTER 1	1223	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified:974	172	155
CLUSTER 2	1603	Diseases of the respiratory system:1273	51	128
CLUSTER 3	442	Diseases of the digestive system:373	153	154
CLUSTER 4	470	Diseases of the genitourinary system:283	166	125
CLUSTER 5	2414	Diseases of the respiratory system:2278	178	127
CLUSTER 6	385	Diseases of the skin and subcutaneous tissue:324	165	110
CLUSTER 7	887	Symptoms, signs, and abnormal clinical and lab oratory findings, not elsewhere classified:557	42	184
CLUSTER 8	601	Diseases of the eye and adnexa:318	167	111
CLUSTER 9	787	Diseases of the musculoskeletal system and connective tissue:422	39	160
CLUSTER 10	800	Diseases of the musculoskeletal system and co nnective tissue:733	169	149

To identify outliers in the cluster analysis, we will use the Silhouette value, whereas mentioned it measures how an object is similar to its cluster and far from neighbors' clusters, Values of Silhouette range from -1 to 1 and values that are close to 1 means observations are fitted to the cluster while values that are less than zero mean they don't fit well and can be considered outliers

Values with Silhouette width less than zero are considered outliers. There are 697 outliers in the dataset, these outliers need further investigation by insurance specialists.

4.4.2 Local outlier factor (LOF)

Local density is determined by estimating distances between data points that are neighbors (k-nearest neighbors). The number of chosen neighbors is 100, and for each data point, the outlier score LOF was calculated.

The following chart shows the values of the LOF, and it is shown that most LOF values are around 1, while there are some points with very high LOF

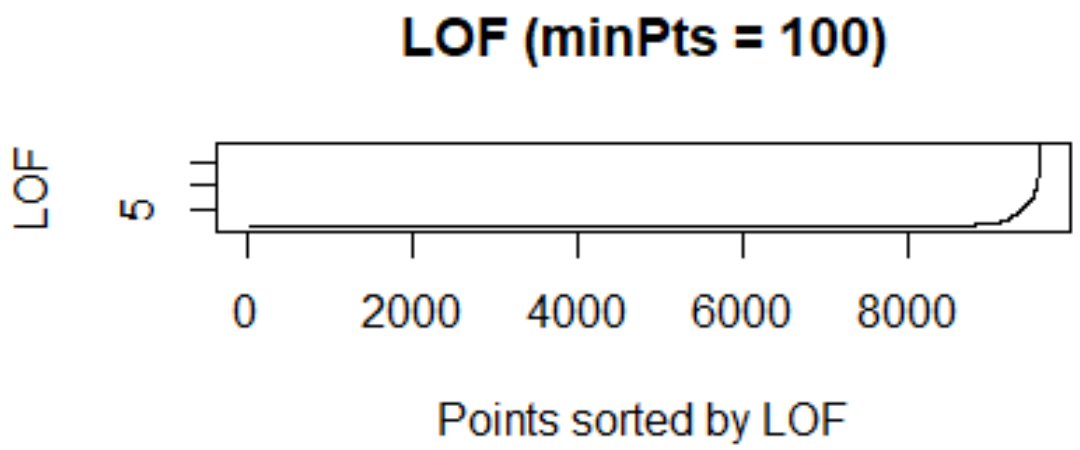


Figure (4.13) Local Outlier Factor (LOF) by (Researcher)

The higher the LOF the more possible the data is an outlier. Since there are very high values of the LOF the threshold was set to 1.5, whereas every point with LOF higher than 1.5 can be considered as an outlier.

- LOF < 1.5 Inlier (similar data point which is inside the density cluster)
- LOF > 1.5 Outlier

761 claims can be considered outliers and need more investigation to check for fraud.

4.4.3 Isolation forest (IF)

The idea of the algorithm is to isolate outliers by creating decision trees over random attributes. During the split an anomaly score is calculated for each point:

- If the value is close to 1 the data point is likely an outlier
- If the value is smaller than 0.5, then the data point is likely to be a regular point (Bai, 2021)

The following chart shows the outliers in blue graphed in a two-dimensional space

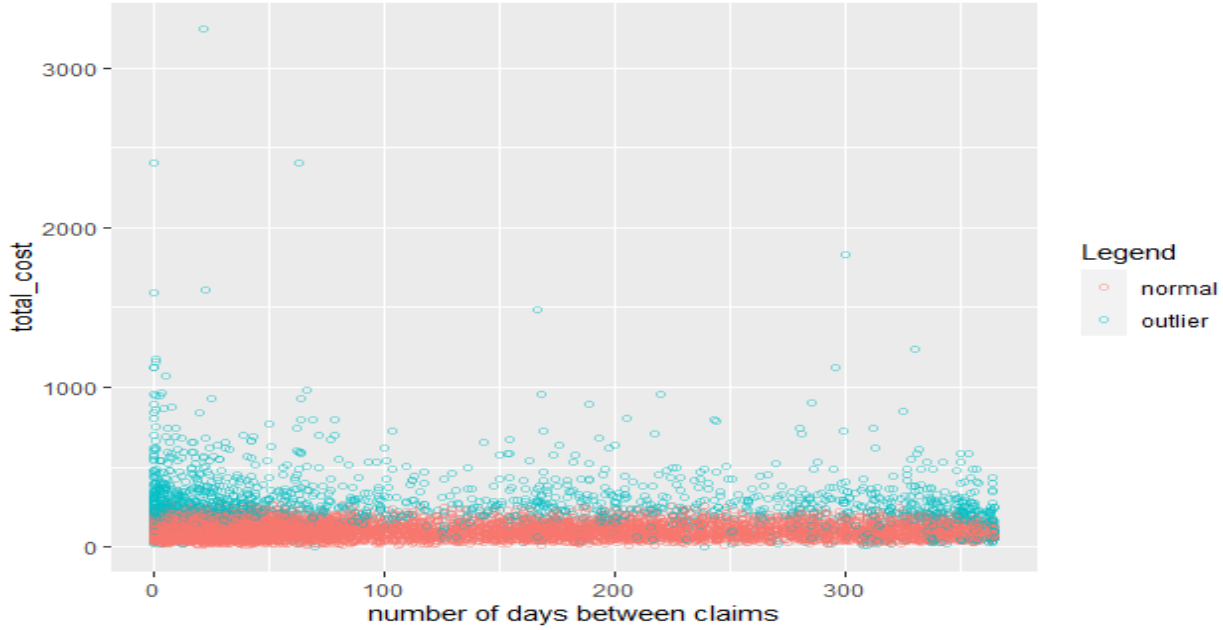


Figure (4.14) Isolation Forest Outliers by (Researcher)

812 outliers were detected using the Isolation Forest algorithm.

Comparing between outlier detection algorithms

The objective of the study is to detect unlabeled outliers, which means there are no accuracy measures to detect which algorithm performed better since knowing for sure which medical claim is a fraud needs further investigation by insurance specialists.

The following table shows a sample of these outliers that are available in the 3 outliers detection methods where 403 claims were joint in the 3 outlier detection methods.
Some outliers were classified outliers because they had very low costs or very high intervals between claims visits. These outliers are surely not considered suspect for fraud.

Table (4.19) Sample of outliers for fraud detection

Clinic DISEASE	same provider for the previous visit	number of days between claims	Main category	total claim cost
Excessive bleeding in the premenopausal period	Yes	0	Diseases of the genitourinary system	75.5
Allergy, unspecified	Yes	0	Injury, poisoning, and certain other consequences of external causes	157.45
Thyroiditis	Yes	0	Endocrine, nutritional and metabolic diseases	174
Benign neoplasm of the thyroid gland	No	0	Neoplasms	1122.9
Delayed puberty	Yes	0	Endocrine, nutritional and metabolic diseases	351.5
Urinary tract infection, site not specified	Yes	0	Diseases of the genitourinary system	86.6
Gonococcal vulvovaginitis, unspecified	Yes	0	Certain infectious and parasitic diseases	171
Conjunctivitis	Yes	0	Diseases of the eye and adnexa	138
Bronchitis, not specified as acute or chronic	No	185	Diseases of the respiratory system	585
Abdominal migraine, intractable	Yes	12	Diseases of the nervous system	560
Rheumatic aortic valve diseasesâ° Syncope and collapse	No	2	Diseases of the eye and adnexa	544.75
Diabetes mellitus due to underlying conditionâ° Other abdominal pain	Yes	84	Endocrine, nutritional and metabolic diseases	517.55
hyperlipidemiaâ° Vitamin B12 deficiency anemiaâ° Vitamin D deficiency	Yes	109	Endocrine, nutritional and metabolic diseases	505.75
Diabetes mellitus due to underlying condition	Yes	0	Endocrine, nutritional and metabolic diseases	458.5
Iron deficiency anemiaâ° Palpitations	No	144	Diseases of the blood and blood-forming organs	439.2

Some outliers may be caused because they have low time intervals between visits, or because the cost of the medicines, lab, x-rays, or medical procedures are much higher than the mean of the same claim type.

Summarizing the outliers by the provider and checking if the outliers claim forms a high percentage from the total claims done by the provider may be an indicator that the provider is abusing the insurance system.
And the same case by the customer, if the customer claims are mostly outliers either having higher cost claims or low time interval between the claims indicate that the customer is suspicious.

The following table shows Providers who have higher outliers' percentage, which indicates that these providers have higher costs per claim for the same diagnosis and further investigation might be required.

Table (4.20) Sample of Providers/ Doctor with a high percentage of outliers from total claims

Provider ID	Count of outliers	Count of total claims by provider	Percentage of outliers from total claims
154	2	14	14%
157	2	14	14%
112	2	17	12%
7	2	21	10%
105	3	23	13%
487	7	33	21%
259	5	43	12%
83	4	49	8%
69	10	150	7%
44	21	186	11%
10	35	447	8%
5	44	622	7%
19	136	1922	7%

The following table shows a sample of customers who have high outliers' percentages, which may be caused by higher claims costs or low time intervals between doctor visits.

Table (4.21) Sample of Subscribers with a high percentage of outliers from total claims

Customer ID	count of outliers	total claims by customer	Percentage
8543xx495	3	16	19%
4106xx591	2	15	13%
8533xx825	2	13	15%
9495xx626	3	11	27%
4010xx206	2	10	20%
4022xx867	4	10	40%
8516xx966	3	10	30%
9117xx737	3	10	30%
4112xx972	5	8	63%
8523xx083	3	8	38%
9841xx484	2	8	25%

Chapter five

Conclusions and recommendations

5.1 Conclusions

This study concluded on the importance of machine learning in the health insurance sector, where predicting the risk level for customers will assist the insurance company to set right premium rates and minimize the risk, as for Fraud detection the machine learning models will assist the company in detecting fraudulent activities done by either medical providers or by customers. This methodology can be automized by Insurance companies to give real-time alerts when outliers occur and accordingly take immediate actions. This will save cost and time and increase fraud and abusers' detection accuracy.

The best-performed model in predicting the risk level was Random Forest with total accuracy of 94%, Sensitivity for High-Risk level 72%, and 81% for Mid-Risk.

Various studies are done in this field aiming to define a model that predicts the risk level accurately, comparing the sensitivity results with other studies the accuracy is good, for example in a similar study the best performing model was the Light Gradient Boosted Tree classifier achieving sensitivity 91.5%, but the researcher used a much higher number of independent variables (Maisog, et al., 2019).

In another study, the researcher distributed the patients on multiple groups and Multinomial logistic regression was used to predict the risk group where overall accuracy was 84% which is a bit lower than the overall accuracy achieved in this case study. (Rosella, et al., 2014)

The studies show that the accuracy rate and used approach were good and emphasize the importance of machine learning algorithms in this field since accurately predicting which patients will incur high costs is very important for healthcare providers.

As for the fraud detection model, the study provides an approach that insurance companies can follow to detect fraud cases. The expected fraud cases need further investigation by insurance specialists to decide on the model accuracy.

A proper mix of machine learning and human investigation can bring fraud detection to a high level of accuracy and objectivity.

5.2 Recommendations

Based on the output of this study few recommendations are given for future work related to the case study:

- Include additional variables related to the case study aiming to increase the accuracy such as:
  - More details regarding the medical history such as type of medicines the person takes/used to take, Family medical history, more details about the chronic diseases, …
  - Variables are relevant to the health condition such as BMI, smoking status, physical activity, alcohol consumption.
  - Socio-economic features such as house-ownership, income level, education, poverty, food security, etc.
- Increase the awareness within insurance companies on the importance of data mining and machine learning in discovering patterns and analyzing the behavior of their customers.
- Examine multiple outlier detection algorithms on different datasets and investigate more on which algorithm performs better in detecting real fraud cases and gives higher accuracy. This should be done by insurance specialists who can classify the outlier as real fraud or normal.
- Encourage Insurance companies to deploy a real-time fraud detection tool that sends notifications to alert about suspicious claims.
- Encourage Insurance companies to use predictive models in their pricing and underwriting processes.
- Perform Machine learning algorithms that require high computational power on separate servers to achieve more efficiency.

References

Abdulhafedh, A. (2022). Comparison between Common Statistical Modeling Techniques Used in Research. *Open Access Library Journal, 9*.

Abdul-Rahman, S., Arifin, N. F., Hanafiah, M., & Mutalib, S. (2021). Customer Segmentation and Profiling for LifeInsurance using K-Modes Clustering and DecisionTree Classifier. *(IJACSA) International Journal of Advanced Computer Science and Applications, 12*(9).

Aggarwal, C. C. (2015). *Data Mining The Textbook.* Switzerland: Springer International Publishing.

Bai, B. (2021, Oct 19). *What is Isolation Forest?* Retrieved from Data Science World: https://dsworld.org/what-is-an-isolation-forest/

Batool, F., & Hennig, C. (2021). Clustering with the Average Silhouette Width. *Computational Statistics and Data Analysis (CSDA), 158*.

Bhalla, A. (2012). Enhancement in Predictive Model for Insurance Underwriting. *International Journal of Computer Science & Engineering Technology (IJCSET), 3*(5), 160-165.

Bhardwaj, N., & Anand, R. (2020). Health Insurance Amount Prediction. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), 9*(5).

Blanken, E. (2017). *The Impact of Big Data and Machine Learning on Insurance, Master Thesis Actuarial Science and Mathematical Finance.* University of Amestrdam.

Bolton, R. J., & Hand, D. (2002). Statistical Fraud Detection: A Review. *Statistical Science17(3)*, 235–249.

Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems,4*, 145–154.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. doi:http://dx.doi.org/10.1023/A:1010933404324

Bücker, T. (2016). *Customer Clustering in the Insurance Sector by Means of Unsupervised Machine Learning, Master Program in Advanced Analytics.* NOVA Information Management School.

Chauhan, N. S. (2020, Jan). *Decision Tree Algorithm, Explained.* Retrieved from kdnuggets: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Clustering Mixed Data Types in R. (2016, Jun 22). Retrieved from Wicked Good Data: http://dpmartin42.github.io/posts/r/cluster-mixed-types

Dabbura, I. (2018, Sep 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.* Retrieved from Towards data science: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Delua, J. (2021, Mar 12). *Supervised vs. Unsupervised Learning: What's the Difference?* Retrieved from IBM: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Demir, A. (2021, Aug 24). *Support Vector Machine (SVM) Classification*. Retrieved from Medium: https://medium.com/geekculture/support-vector-machine-svm-classification-6579184d78e5

Deng, N. T. (2013). *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions (1st ed.).* New York: Chapman and Hall/CRC.

Fisher, J. (2020, Dec 10). *Decision Tree Subtypes*. Retrieved from Data Science Diaries: https://julielinx.github.io/blog/50_trees_subtypes/

Fitzpatrick, T., Rosella, L. C., Calzavara, A., Petch, J., Pinto, A. D., Manson, H., & Goel, V. (2015). Looking Beyond Income and EducationSocioeconomic Status Gradients Among FutureHigh-Cost Users of Health Care. *Am J Prev Med, 49*(2), 161-171.

Fritsch, S., Guenther, F., & Wright, M. (2019, Feb 7). *Package 'neuralnet'*. Retrieved from cran.r-project.org: https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

Gajawada, S. (2019, Oct 4). *Chi-Square Test for Feature Selection in Machine learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees. A Mostly Intuitive Guide, But Also Some Python.* Goodreads, Kindle Edition.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction* (Second Edition ed.). Stanford, California: Springer.

Health Insurance. (n.d.). Retrieved from Palestinian Insurance Federation: https://www.pif.org.ps/articles/view/221

IBM Cloud Education. (2020, Aug 17). *Neural Networks*. Retrieved from IBM: https://www.ibm.com/cloud/learn/neural-networks

Jödicke, A. M., Zellweger, U., Tomka, I., Neuer, T., Curkovic, I., Roos, M., . . . Sargsyan, H. (2019). Prediction of health care expenditure increase: how does pharmacotherapy contribute? *BMC Health Services Research, 19*(1).

Johnson, M. E. (2016, Sep). Multi-Stage Methodology to Detect Health Insurance Claim Fraud. *Health Care Manag Sci, 19*(3), 60-249.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., & Geraili, B. (2015). Improving fraud and abuse detection in general physician claims: a data mining study. *International Journal of Health Policy and Management, 5(3)*, 165–172.

Kassambara , A. (2017). *Practical Guide To Cluster Analysis in R Edition 1.* Montpellier, France: STHDA.

Liu, Q., & Vasarhelyi, M. (2013). Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information. *29th WORLD CONTINUOUS AUDITING AND REPORTING SYMPOSIUM (29WCARS), 1*.

Macedo, P., Araia, S., & Zafari, B. (2016). *Medicare Fraud Analytics Using Cluster, Paper 10761-2016.* George Washington University.

Maisog, J., Li, W.-H., Xu, Y., Hurley, B., Shah, H., Lemberg, R., & Borden, T. (2019). Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. *ArXiv abs/1912.13032*.

Mckinsey&Company. (2017, Sep 1). *Artificial intelligence in health insurance: Smart claims management with self-learning software*. Retrieved from Mckinsey&Company: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/artificial-intelligence-in-health-insurance-smart-claims-management-with-self-learning-software

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations Of Machine Learning second edition.* Cambridge, Massachusetts: The MIT Press.

Morse, S. (2021, May 27). *How health insurance companies use AI to make consumers healthier*. Retrieved from Health Care Finance: https://www.healthcarefinancenews.com/news/how-health-insurance-companies-use-ai-make-consumers-healthier

Nielsen, M. (2015). *Neural Networks And Deep Learning.* Determination Press.

Pai, D. D., Agnihotri, P., Rajath, G., & Kumar Jha, B. (2016). Applications of Data Mining in Detecting Fraudulent Health Insurance Claim. *International Journal of Engineering Research & Technology (IJERT), 4*(22).

Patel, A. (2019, Apr 24). *FeedForward Neural Network and Back Propagation.* Retrieved from medium: https://medium.com/computer-vision-101-with-deep-learning/chapter-2-3-deep-learning-101-feedforward-neural-network-and-back-propagation-d42feff32d0

Pooja, H., & Jagadeesh , P. (2019). A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records. *1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 1*, 51-55.

Positive Impact of Machine Learning in the Insurance Industry. (2021, June 8). Retrieved from Intellectsoft: https://www.intellectsoft.net/blog/machine-learning-in-insurance-automation/

Rosella, L. C., Fitzpatrick, T., Wodchis, W., Calzavara, A., Manson, H., & Goel, V. (2014, Oct 31). High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *BMC Health Services Research, 14*.

RPubs. (n.d). *SVM with CARET*. Retrieved from RPubs by RStudio: https://rpubs.com/uky994/593668

Sennaar, K. (2019, Dec 13). *Artificial intelligence in Health Insurance – Current Applications and Trends.* Retrieved from emerj: https://emerj.com/ai-sector-overviews/artificial-intelligence-in-health-insurance-current-applications-and-trends/

Shwartz, S., & David, S. (2014). *Understanding Machine Learning: From Theory To Algorithms.* Cambridge University Press.

Tahsildar, S. (2019, Mar 3). *Random Forest Algorithm In Trading Using Python*. Retrieved from Quantinsti: https://blog.quantinsti.com/random-forest-algorithm-in-python/

Tamang, S., Milstein, A., Sørensen, H. T., Pedersen, L., Mackey, L., Betterton, J.-R., & Janson, L. (2017). Predicting patient 'cost blooms' in Denmark: a longitudinal population-based study. *BMJ Open, 7*.

top-10-data-science-use-cases-in-insurance. (2021, May 27). Retrieved from activewizards: https://activewizards.com/blog/top-10-data-science-use-cases-in-insurance/

Tyagi, N. (2021, Mar 22). *What is Information Gain and Gini Index in Decision Trees?* Retrieved from analytics steps: https://www.analyticssteps.com/blogs/what-gini-index-and-information-gain-decision-trees

Tzinie, E. (2020, Nov 20). *Feature Selection: Filter Methods*. Retrieved from Medium: https://medium.com/analytics-vidhya/feature-selection-73bc12a9b39e

Using Statistical Regression Methods in Education Research. (2011, Jul 15). Retrieved from ReStore: https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/index.html

What is Logistic Regression? (2021, June 1). Retrieved from statisticssolutions: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

Yadav, P. (2021, Apr 4). *medium.com/mlearning-ai/top-5-applications-of-machine-learning-in-the-insurance-industry*. Retrieved from medium: https://medium.com/mlearning-ai/top-5-applications-of-machine-learning-in-the-insurance-industry-99abe8b840cb

Zaqueu, J. R. (2019). *Customer Clustering in the Health Insurance Industry, Masters Program in Advanced Analytics .* NOVA Information Management School.

Zhang, Y., Lu, S., Niu, Y., & Zhang, L. (2018). Medical expenditure clustering and determinants of the annual medical expenditures of residents: a population based retrospective study from rural China. *BMJOpen, 8:e022721*.

Appendix(A) Permission letter from the insurance company

تمكين للتأمين
TAMKEEN INSURANCE

Date: February 16, 2021

Permission Letter

To: Birzeit University,,

On behalf of Tamkeen Insurance Company, I am writing to grant permission for Ruba Nasri Shihadeh holding ID number "914859814"; a Masters Student at Birzeit University to conduct her research titled *" The Role of Machine Learning in Health Insurance Industry"* using data from Tamkeen Insurance database and is related to the research topic.

The data is to be used for research purposes only and will be shared according to the company's protocols and privacy regulations.

Sincerely,

Mohammad Rimawi
General Manager

تمكين للتأمين
TAMKEEN INSURANCE
GM

برؤية إسلامية

Ramallah, Palestine
P.O Box 2222
Tel: +970 2 2944400
Fax: +970 2 2944401

info@tamkeen-ins.ps
www.tamkeen-ins.ps
Tamkeen Insurance
1800 202 202

Appendix(B) ANOVA tests

1. ANOVA test for mean differences in age based on the dependent variable risk level

```
summary(anovatable)
              Df  Sum Sq Mean Sq F value Pr(>F)
risk           2  375134  187567   820.9 <2e-16 ***
Residuals   10840 2476691     228


PostHocTest(anovatable,method="lsd")

  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$risk
                          diff     lwr.ci     upr.ci    pval
Mid Risk-High Risk    3.779423   2.602452   4.956394 3.2e-10 ***
not_risky-High Risk -11.709114 -12.625518 -10.792710 < 2e-16 ***
not_risky-Mid Risk  -15.488537 -16.358611 -14.618463 < 2e-16 ***

Mean Differences`
  data$risk`   `n()`  mean    sd
1 High Risk    1196  33.5  13.8
2 Mid Risk     1348  37.3  14.5
3 not_risky    8299  21.8  15.4
```

2. ANOVA test for mean differences in number of children based on the dependent variable risk level

```
summary(anovatable)
              Df Sum Sq Mean Sq F value Pr(>F)
risk           2   1338   668.9     287 <2e-16 ***
Residuals   10840  25264     2.3


PostHocTest(anovatable,method="lsd")

  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$risk
                          diff     lwr.ci     upr.ci   pval
Mid Risk-High Risk  -0.1730323 -0.2919038 -0.0541607 0.0043 **
not_risky-High Risk -0.9146853 -1.0072401 -0.8221305 <2e-16 ***
not_risky-Mid Risk  -0.7416530 -0.8295286 -0.6537774 <2e-16 ***

Mean Differences`
   data$risk`   n()` mean   sd

1 High Risk    1196 1.70   1.52
2 Mid Risk     1348 1.53   1.89
3  not_risky   8299 0.784  1.46
```

3. ANOVA test for mean differences in number of chronic diseases based on the dependent variable risk level

```
summary(anovatable)
               Df Sum Sq Mean Sq F value Pr(>F)
risk            2   1576   788.0    1945 <2e-16 ***
Residuals   10840   4391     0.4


PostHocTest(anovatable,method="lsd")

  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$risk
                          diff     lwr.ci     upr.ci   pval
Mid Risk-High Risk   0.5009453  0.4513877  0.5505029 <2e-16 ***
not_risky-High Risk -0.5876501 -0.6262362 -0.5490640 <2e-16 ***
not_risky-Mid Risk  -1.0885954 -1.1252307 -1.0519600 <2e-16 ***

Mean Differences
` data$risk` `n()`    mean     sd
1 High Risk    1196 0.590   1.24
2 Mid Risk     1348 1.09    1.37
  3 not_risky    8299 0.00265 0.0514
```

4. ANOVA test for mean differences in policy cost based on the dependent variable risk level

```
summary(anovatable)
               Df    Sum Sq  Mean Sq F value Pr(>F)
risk            2  32364122 16182061     231 <2e-16 ***
Residuals   10840 759320806    70048


PostHocTest(anovatable,method="lsd")

  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$risk
                          diff     lwr.ci     upr.ci    pval
Mid Risk-High Risk   -76.20890  -96.81722  -55.60058 4.5e-13 ***
not_risky-High Risk -161.75348 -177.79936 -145.70759 < 2e-16 ***
not_risky-Mid Risk   -85.54457 -100.77924  -70.30991 < 2e-16 ***


Mean Differences

  data$risk`  `n()` mean    sd
1 High Risk    1196 1278.  230.
2 Mid Risk     1348 1201.  237.
3 not_risky    8299 1116.  273.
```

Appendix(C) R Code for subscribers clustering

```
library(cluster)
library(tidyverse)
numerics2<-
data[c("outpatient_SumOfPAY_VALUE","inpatient_amount","outpatient_CountOfPAY_VALUE")]
numerics3<-scale(numerics2)
# Distance used in clustering
distance<-dist(numerics3,method="euclidean")
hclust_avg<-hclust(distance,method = "ward.D2")
win.graph(10,8,6)
plot(hclust_avg)
rect.hclust(hclust_avg,border = 2:6,h=50)
abline(h=100,col="red")
#compute cophentic distance
res.coph<-cophenetic(hclust_avg)
cor(distance,res.coph)
set.seed(1200)
cut_avg<-cutree(hclust_avg,k=4)
table(cut_avg)
library(dplyr)
table_cluster<-mutate(data,cluster=cut_avg)
summary(table_cluster)
# summary for hierarchal clustering
summary_hie<-table_cluster %>%
  group_by(table_cluster$cluster)%>%
summarise(n(),meanout=mean(outpatient_SumOfPAY_VALUE),meanin=mean(inpatient_amount),meann
umberofout=mean(outpatient_CountOfPAY_VALUE))
#k-means clustering
dataforcluster<-cbind(data[,1:16],numerics3)
dataforcluster<-as.data.frame(dataforcluster)
distance2<-get_dist(numerics3)
# find optimal number of clusters
```

```
set.seed(123)
plotelbow<-fviz_nbclust(numerics3,kmeans, method = "wss")
win.graph(10,8,8)
plotelbow
Kmean<-kmeans( numerics3,4,nstart = 500)
Kmean$size
#plot Kmean clusters
p3 <- fviz_cluster(Kmean, geom = "point",  data = numerics3) + ggtitle("k = 4")
win.graph(10,8,6)
plot(p3)
# summary of Kmeans clusters
dataforcluster1<-cbind(cluster=Kmean$cluster,data)
summarykmeans<-dataforcluster1 %>%
  group_by(cluster)%>%
summarise(n(),meanout=mean(outpatient_SumOfPAY_VALUE),meanin=mean(inpatient_amount),meann
umberofout=mean(outpatient_CountOfPAY_VALUE))
```

Appendix(D) R Code for classification algorithms

```
library(readxl)
data <- read_excel("C:/Users/Simon/Desktop/thesis/data.xlsx")
attach(data)
#plots of distribution
win.graph(8,4,6)
par(mfrow=c(1,5))
for(i in 1:5){hist.default(numerics[,i], main=names(numerics)[i],border="blue",col="green")}
win.graph(8,4,6)
par(mfrow=c(1,5))
for(i in 1:5){boxplot(numerics[,i], main=names(numerics)[i],col="red")$out}

# plot xs versus y /numeric variables
library(ggpubr)
```

```
win.graph(12,4,12)
ggdensity(data,x="age",facet.by="risk",fill="lightgray",title =
"age")+stat_overlay_normal_density(color="blue",linetype="dashed")
ggdensity(data,x="age_of_oldest_kid",facet.by="risk",fill="lightgray",title = "age of oldest
kid")+stat_overlay_normal_density(color="blue",linetype="dashed")
ggdensity(data,x="number_of_kids",facet.by="risk",fill="lightgray",title = "number of
kids")+stat_overlay_normal_density(color="blue",linetype="dashed")
ggdensity(data,x="number_of_chronic_diseases",facet.by="risk",fill="lightgray",title = "number of
chronic")+stat_overlay_normal_density(color="blue",linetype="dashed")
# 1 variable plot / categorical variables
chronic<-table(data$has_chronic,data$risk)
company<-table(data$company_type,data$risk)
policy<-table(data$policy_coverage,data$risk)
new_old<-table(data$new_old_account,data$risk)
subscriber<-table(data$subscriber,data$risk)
glasses<-table(data$wear_glasses,data$risk)
oldsurgery<-table(data$old_surgery,data$risk)
gender<-table(data$gender,data$risk)
marital_status<-table(data$marital_status,data$risk)
win.graph(10,4,12)
par(mfrow=c(1,3))
mosaicplot(chronic,color = TRUE,cex.axis = 1.2)
mosaicplot(company,color = TRUE,cex.axis = 1.2)
mosaicplot(policy,color = TRUE,cex.axis = 1.2)
win.graph(10,4,12)
par(mfrow=c(1,3))
mosaicplot(new_old,color = TRUE,cex.axis = 1.2)
mosaicplot(subscriber,color = TRUE,cex.axis = 1)
mosaicplot(glasses,color = TRUE,cex.axis = 1.2)
win.graph(10,4,12)
par(mfrow=c(1,3))
mosaicplot(oldsurgery,color = TRUE,cex.axis = 1.2)
```

```
mosaicplot(gender,color = TRUE,cex.axis = 1.2)
mosaicplot(marital_status,color = TRUE,cex.axis = 1.2)
#2variables plots
win.graph(12,4,12)
data %>% mutate(y = factor(risk)) %>%
  ggplot(aes(age, number_of_chronic_diseases, fill = y, color=y)) +
  geom_point(show.legend = TRUE) +
  stat_ellipse(type="norm")
# View the significance of variables before building models
library(ggpubr)
library(tidyverse)
library(stats)
library(perturb)
library(DescTools)
# anova age
data %>%
  group_by(data$risk)%>%
  summarise(n(),mean=mean(age),sd=sd(age))
anovatable1<-aov(age~risk,data=data)
summary(anovatable1)
PostHocTest(anovatable1,method="lsd")
# anova age-of oldest kid
data %>%
  group_by(data$risk)%>%
  summarise(n(),mean=mean(age_of_oldest_kid),sd=sd(age_of_oldest_kid))
anovatable2<-aov(age_of_oldest_kid~risk,data=data)
summary(anovatable2)
PostHocTest(anovatable2,method="lsd")
# anova number of kids
data %>%
  group_by(data$risk)%>%
  summarise(n(),mean=mean(number_of_kids),sd=sd(number_of_kids))
```

```
anovatable3<-aov(number_of_kids~risk,data=data)
summary(anovatable3)
PostHocTest(anovatable3,method="lsd")
# anova number of chronic diseases
data %>%
  group_by(data$risk)%>%
  summarise(n(),mean=mean(number_of_chronic_diseases),sd=sd(number_of_chronic_diseases))
anovatable4<-aov(number_of_chronic_diseases~risk,data=data)
summary(anovatable4)
PostHocTest(anovatable4,method="lsd")

# collinearity
model2<-lm(total ~
age+age_of_oldest_kid+number_of_kids+number_of_chronic_diseases+policy_cost,data=data)
summary(model2)
car::vif(model2)
library(mctest)
omcdiag(model2)
# correlation numerical variables
library(corrplot)
cor.matrix<-cor(numerics,method="pearson")
win.graph(12,8,8)
corrplot(cor.matrix,type="upper",method="color",addCoef.col="black",tl.col="black",number.cex=1,mar=c(
0,0,0,0))

library(sjPlot)
categorical[]<-lapply(categorical,as.integer)
win.graph(10,8,6)
sjp.corr(categorical)

library(ggcorrplot)
cat<-data[c(3:9,14:15)]
```

```
win.graph(14,12,12)
model.matrix(~0+. ,data=cat) %>%
          cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag=F,type="lower",lab=TRUE,lab_size=2)
 # Modeling
# Multinomial regression
library(ModelMetrics)
library(nnet)
multinomialmodel<-
multinom(risk~new_old_account+company_type+policy_coverage+has_chronic+subscriber+wear_glasse
s+old_surgery+age+gender+marital_status+number_of_kids+number_of_chronic_diseases+policy_cost+
age_of_oldest_kid,data,maxit = 10000,Hess=TRUE)
summary(multinomialmodel)
multinomialmodel
(round(fitted(multinomialmodel),2))
prediction<-predict(multinomialmodel,data)
tablecomparison<-table(data$risk,prediction)
tablecomparison
confusionMatrix(prediction,data$risk)
library(effects)
win.graph(12,10,10)
plot(Effect("number_of_chronic_diseases",multinomialmodel))
plot(Effect("age",multinomialmodel))
plot(Effect("gender",multinomialmodel))
plot(Effect("old_surgery",multinomialmodel))
multinomialmodel
# multinomial caret pkg
library(caret)
library(base)
library(generics)
fit.control<-trainControl(method = "repeatedcv",number=10,repeats = 10)
set.seed(456)
```

```
fit<-
caret::train(as.factor(risk)~new_old_account+company_type+policy_coverage+has_chronic+subscriber+
wear_glasses+old_surgery+(age_of_oldest_kid)+(age)+gender+marital_status+number_of_kids+number
_of_chronic_diseases+policy_cost,data=data, method="multinom",trControl=fit.control)
pred1<-predict(fit,data)
pred1
confusionMatrix(pred1,data$risk)
library(performance)
library(car)
library(pscl)
Anova(multinomialmodel,type ="II",test="Wald")
round(pR2(multinomialmodel),5)
coef(multinomialmodel)
odds<-round(exp(coef(multinomialmodel)),3)
odds
# decision tree
#decision tree C50
library(C50)
Ctrl<-C5.0Control(subset=TRUE, minCases = 50,noGlobalPruning = TRUE,CF=0.05,earlyStopping =
TRUE)
modeltraining<-C5.0.default(data[c(3:16)],as.factor(data$risk),control = Ctrl,trials=20)
C5predict<-predict(modeltraining,data)
C5predict
summary(modeltraining)
win.graph(26,14,8)
plot(modeltraining,type="s",main="decision tree")
#random forest
library(randomForest)
fittree<-
randomForest(as.factor(risk)~company_type+policy_coverage+has_chronic+subscriber+wear_glasses+ol
d_surgery+(age_of_oldest_kid)+(age)+gender+marital_status+number_of_kids+number_of_chronic_dise
ases+policy_cost,data=data,trControl=fit.control,trace=TRUE,ntree=3000,mTry=4)
```

```
plot(fittree)
predtree<-predict(fittree,data)
library(writexl)
predtree<-cbind(data,predtree)
write_xlsx(predtree,"C:/Users/Simon/Desktop/thesis/checkrisk.xlsx")
win.graph(20,12,12)
plot(fittree,main = "error based on ntrees")
# tuning parameters for random forest
model_tuned<-tuneRF(x=data[,3:16],y=data$risk,ntreeTry = 3000,mtryStart = 3,stepFactor = 1.5,improve
= 0.01,trace = TRUE)
win.graph(20,12,12)
model_tuned

confusionMatrix(predtree,data$risk)
win.graph(20,12,12)
varImpPlot(fittree)
library(caret)
library(mlbench)
# choose mtry, control for random forest
mtry=floor(sqrt(ncol(data)))
control1<-trainControl(method ="repeatedcv",number = 100,repeats = 10 )
metric<-"Accuracy"
tunegrid<-expand.grid(.mtry=mtry)
rf_default<-
train(as.factor(risk)~new_old_account+company_type+policy_coverage+has_chronic+subscriber+wear_g
lasses+old_surgery+log(age)+gender+marital_status+number_of_kids+number_of_chronic_diseases+pol
icy_cost+age_of_oldest_kid,data,method="rf",metric=metric,trControl=control1)
summary(rf_default)
rf_default$results
#svm
# svm using caret
# SVM non linear
```

```
set.seed(3233)
svm_Radial <-
train(as.factor(risk)~marital_status+company_type+policy_coverage+has_chronic+subscriber+wear_glas
ses+old_surgery+age_of_oldest_kid+age+gender+number_of_kids+number_of_chronic_diseases+policy
_cost, data = data, method = "svmRadial",
                    trControl=trctrl,
                    preProcess = c("center", "scale"),
                    tuneLength = 10)
 plot(svm_Radial)
svm_Radial
test_pred_radial <- predict(svm_Radial, newdata = data)
confusionMatrix(as.factor(test_pred_radial), as.factor(data$risk))
grid_radial <- expand.grid(sigma = c(0,0.01, 0.02, 0.03, 0.04,
                                     0.05, 0.06, 0.07,0.08, 0.1, 0.25, 0.5, 0.75,0.9),
                           C = c(0,
                                 1, 1.5, 2,5,8,10,15,20))
set.seed(3233)
svm_Radial_Grid <-
train(as.factor(risk)~marital_status+company_type+policy_coverage+has_chronic+subscriber+wear_glas
ses+old_surgery+age_of_oldest_kid+age+gender+number_of_kids+number_of_chronic_diseases+policy
_cost, data = data, method = "svmRadial",
                        trControl=trctrl,
                        preProcess = c("center", "scale"),
                        tuneGrid = grid_radial,
                        tuneLength = 10)
svm_Radial_Grid
# neural networks for classification
library(neuralnet)
# Normalize the data
maxs <- apply(numerics, 2, max)
maxs<-as.numeric(maxs)
mins <- apply(numerics, 2, min)
```

```
mins<-as.numeric(mins)
scaled <- as.data.frame(scale(numerics, center = mins,scale = maxs - mins))
scaled
dataforneural<-
cbind(scaled,data[c("company_type","new_old_account","subscriber","wear_glasses","old_surgery","has_
chronic","gender","marital_status","risk")])
attach(dataforneural)
View(dataforneural)
m<-
model.matrix(risk~company_type+new_old_account+subscriber+wear_glasses+old_surgery+has_chronic
+gender+marital_status+age+age_of_oldest_kid+number_of_kids+number_of_chronic_diseases+policy_
cost,data=dataforneural)
View(m)
head(m)
set.seed(3233)
nn1 <-
neuralnet(risk~company_typeeducational+company_typemedical+company_typeNGOs+new_old_accoun
told+subscriberemployee+subscribernew_born+subscriberspouse+wear_glassesYes+old_surgeryYes+h
as_chronicYes+genderMale+marital_statussingle+age+age_of_oldest_kid+number_of_kids+number_of_
chronic_diseases+policy_cost,data = m,linear.output = FALSE,hidden = 5,act.fct = "logistic",lifesign =
"minimal",likelihood = TRUE,err.fct = 'ce',threshold = 0.05)
summary(nn1)
library(NeuralNetTools)
win.graph(30,10,14)
olden(nn1)
olden(nn1,out_var='High Risk',bar_plot=FALSE)
yhat<-nn1$net.result
yhat<-data.frame(yhat)

yhat<-ifelse(max.col(yhat[ ,1:3])==1,"High Risk",
      ifelse(max.col(yhat[ ,1:3])==2,"Mid Risk","not_risky"))
summary(yhat)
```

```
table(as.factor(data$risk),as.factor(yhat))
confusionMatrix(as.factor(yhat),data$risk)
plot(nn1)

# Splitting the data into a TRAINING and TESTING data sets
set.seed(3000)
Shuffleddata1<-data[order(runif(10843)),]
traindata1<-Shuffleddata1[1:floor(nrow(Shuffleddata1)*0.6), ]
testdata1<-Shuffleddata1[floor(nrow(Shuffleddata1)*0.6+1):nrow(Shuffleddata1), ]
# Multinomial regression train/test
fit.control<-trainControl(method = "repeatedcv",number=10,repeats = 10)
fit<-
caret::train(as.factor(risk)~new_old_account+company_type+policy_coverage+has_chronic+subscriber+
wear_glasses+old_surgery+(age_of_oldest_kid)+(age)+gender+marital_status+number_of_kids+number
_of_chronic_diseases+policy_cost,data=traindata1, method="multinom",trControl=fit.control)
pred1<-predict(fit,testdata1)
confusionMatrix(pred1,testdata1$risk)

#decision tree train/test
mytree<-
rpart(as.factor(risk)~company_type+policy_coverage+has_chronic+subscriber+wear_glasses+old_surger
y+(age_of_oldest_kid)+(age)+gender+marital_status+number_of_kids+number_of_chronic_diseases+poli
cy_cost,data=traindata1,method = "class",control = control)
mytreepred<-predict(mytree,testdata1)
mytreepred<-apply(mytreepred,1,which.max)
mytreepred<-levels(data$risk)[mytreepred]
confusionMatrix(as.factor(mytreepred),testdata1$risk)
#random forest train/test
fittree<-
randomForest(as.factor(risk)~company_type+policy_coverage+has_chronic+subscriber+wear_glasses+ol
d_surgery+(age_of_oldest_kid)+(age)+gender+marital_status+number_of_kids+number_of_chronic_dise
ases+policy_cost,data=traindata1,trControl=fit.control,trace=TRUE,ntree=3000,mTry=4)
```

```
predtree<-predict(fittree,testdata1)
confusionMatrix(predtree,testdata1$risk)
# SVM train/test
svm_Radial <-
train(as.factor(risk)~marital_status+company_type+policy_coverage+has_chronic+subscriber+wear_glas
ses+old_surgery+age_of_oldest_kid+age+gender+number_of_kids+number_of_chronic_diseases+policy
_cost, data = traindata1, method = "svmRadial",
                trControl=trctrl,
                preProcess = c("center", "scale"),
                tuneLength = 10)
test_pred_radial <- predict(svm_Radial, newdata = testdata1)
confusionMatrix(as.factor(test_pred_radial), as.factor(testdata1$risk))
# neural networks train/test
numerics<-
traindata1[c("age","age_of_oldest_kid","number_of_kids","policy_cost","number_of_chronic_diseases")]
numerics<-as.data.frame(numerics)
# Normalize the data
maxs <- apply(numerics, 2, max)
maxs<-as.numeric(maxs)
mins <- apply(numerics, 2, min)
mins<-as.numeric(mins)
scaled <- as.data.frame(scale(numerics, center = mins,scale = maxs - mins))
dataforneural<-
cbind(scaled,traindata1[c("company_type","new_old_account","subscriber","wear_glasses","old_surgery",
"has_chronic","gender","marital_status","risk")])
attach(dataforneural)
m<-
model.matrix(risk~company_type+new_old_account+subscriber+wear_glasses+old_surgery+has_chronic
+gender+marital_status+age+age_of_oldest_kid+number_of_kids+number_of_chronic_diseases+policy_
cost,data=dataforneural)
nn1 <-
neuralnet(risk~company_typeeducational+company_typemedical+company_typeNGOs+new_old_accoun
```

```
told+subscriberemployee+subscribernew_born+subscriberspouse+wear_glassesYes+old_surgeryYes+h
as_chronicYes+genderMale+marital_statussingle+age+age_of_oldest_kid+number_of_kids+number_of_
chronic_diseases+policy_cost,data = m,linear.output = FALSE,hidden = 5,act.fct = "logistic",lifesign =
"minimal",likelihood = TRUE,err.fct = 'ce',threshold = 0.05)
# create matrix for test data
numerics1<-
testdata1[c("age","age_of_oldest_kid","number_of_kids","policy_cost","number_of_chronic_diseases")]
numerics1<-as.data.frame(numerics1)
maxs1 <- apply(numerics1, 2, max)
maxs1<-as.numeric(maxs1)
mins1 <- apply(numerics1, 2, min)
mins1<-as.numeric(mins1)
scaled1 <- as.data.frame(scale(numerics1, center = mins1,scale = maxs1 - mins1))
dataforneural1<-
cbind(scaled1,testdata1[c("company_type","new_old_account","subscriber","wear_glasses","old_surgery"
,"has_chronic","gender","marital_status","risk")])
attach(dataforneural1)
m1<-
model.matrix(risk~company_type+new_old_account+subscriber+wear_glasses+old_surgery+has_chronic
+gender+marital_status+age+age_of_oldest_kid+number_of_kids+number_of_chronic_diseases+policy_
cost,data=dataforneural1)
# prediction on testdata matrix
predictnn1<-predict(nn1,m1)
yhat<-ifelse(max.col(predictnn1[ ,1:3])==1,"High Risk",
      ifelse(max.col(predictnn1[ ,1:3])==2,"Mid Risk","not_risky"))
confusionMatrix(as.factor(yhat),testdata1$risk)
```

Appendix(E) R Code for Outlier Detection

attach(claim)

library(cluster)

```
library(tidyverse)
library(dplyr)
library(ISLR)
library(Rtsne)
library(ggplot2)
library(Rcpp)
# distance for categorical and numerical
gower_distance<-daisy(claim[c(8,11:13)],metric = "gower",type=list(logratio=3))
summary(gower_distance)
#PAM
sil_width<-c(NA)
for(i in 2:14){
  pam_fit<-pam(gower_distance,diss=TRUE,k=i)
  sil_width[i]<-pam_fit$silinfo$avg.width
}
# plot silhouette width (higher is better)
win.graph(10,10,10)
plot(1:14, sil_width,xlab="Number of clusters",ylab = "silhouette width")
lines(1:14,sil_width)
# chose 10 clusters
pam_fit<-pam(gower_distance,diss=TRUE,k=10)
summary(pam_fit)
pam_fit$clustering
pam_fit$medoids
win.graph(10,8,6)
plot(pam_fit)
# summary
pam_results<-claim %>%
dplyr::select(c(8:16)) %>%
mutate(cluster=pam_fit$clustering)%>%
group_by(cluster)%>%
do(the_summary=summary(.))
```

```
pam_results$the_summary
library(factoextra)
# visualize silhouhette information
sil<-silhouette(pam_fit$cluster,gower_distance)
win.graph(20,10,20)
plot(sil)
fviz_silhouette(sil)
#outliers
#identify observation with negative silhouette
neg_sil_index<-which(sil[,"sil_width"]<0)
sil[neg_sil_index, ,drop=FALSE]
outlierstable<-claim[neg_sil_index,]
library(writexl)
write_xlsx(outlierstable,"C:/Users/Simon/Desktop/thesis/outliers.xlsx")
# using K-means clustering for finding outliers
library(factoextra)
mixedclusters<-kmeans(gower_distance,centers=7)
mixedclusters$cluster
p4 <- fviz_cluster(mixedclusters, geom = "point",  data = claim[c(14:17)],ellipse.type =
"norm",ggtheme=theme_minimal())
summary(p4)
win.graph(10,8,6)
plot(p4)
outliers <- boxplot.stats(gower_distance)$out
# Finding the index positions of the outliers
index_outliers <- which(gower_distance%in% outliers)
# Flag the outliers in the data and create final dataset
kmeans_data_final <- claim %>%
  mutate(index = row_number(),
         cluster = mixedclusters$cluster,
          suspect_behaviour = ifelse(index %in% index_outliers, "F", "NF"))
kmeans_data_final
```

```
# Using LOF Local outlier factor and gower_distance
library(dbscan)
lof <- lof(gower_distance, minPts = 10)
summary(lof)
lof<-as.matrix(lof)
lof_greater2<-which(lof>1.5)
outlierslof<-claim[lof_greater2,]
win.graph(10,10,10)
boxplot(lof)
# distribution of outlier factors
summary(lof)
win.graph(10,10,10)
hist(lof, breaks = 10, main = "LOF (minPts = 10)")
#plot sorted lof. Looks like outliers start around a LOF of 1.5.
plot(sort(lof), type = "l",  main = "LOF (minPts = 10)",
     xlab = "Points sorted by LOF", ylab = "LOF")
# using Isolation forest anomaly detection
library(isotree)
library(Rcpp)
data<-claim[c(8,10,13:15)]
iforest <- isolation.forest(data,ntrees = 5000,ntry=5)
#predict outliers within dataset
data$pred<- predict.isolation_forest(iforest, data, type = "score")
data$outlier <- as.factor(ifelse(data$pred >=0.45, "outlier", "normal"))
data
#plot data again with outliers identified
library(ggplot2)
win.graph(10,10,10)
ggplot(data, aes(y = total_pay, x = `number of days between claims`, color = outlier)) +
  geom_point(shape = 1, alpha = 0.5) +
  labs(y = "total_cost", x = "number of days between claims") +
  labs(alpha = "", colour="Lege
```