



Application of Data Mining to Classify Medical Insurance Customers Based on  
Claim Experience: The Case of Awash Insurance Company S.C

**A Thesis Presented**

**by**

**Mebeki Haile Kabeta**

**to**

**The Faculty of Informatics**

**of**

**St. Mary's University**

**In Partial Fulfillment of the Requirements  
for the Degree of Master of Science**

**in**

**Computer Science**

**January, 2021**

# **ACCEPTANCE**

Application of Data Mining to Classify Medical Insurance Customers Based on  
Claim Experience: The Case of Awash Insurance Company S.C

**By**

**Mebeki Haile Kabeta**

**Accepted by the Faculty of Informatics, St. Mary's University, in partial  
fulfillment of the requirements for the degree of Master of Science in  
Computer Science**

**Thesis Examination Committee:**

---

**Internal Examiner**

---

**External Examiner**

---

**Dean, Faculty of Informatics**

**January 2021**

## DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Mebeki Haile Kabeta

---

Full Name of Student

---

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Dr. Getahun Semeon

---

Full Name of Advisor

---

Signature

Addis Ababa

Ethiopia

January 2021

## Acknowledgments

Foremost, my deepest gratitude is to the almighty God for giving me the ability to face challenges and complete this research work.

Next to this, I would like to express my sincerest gratitude to my advisor Dr. Getahun Semeon for his critical reading, guidance and valuable comments.

I would like to thank friends, underwriters and life and Health manager of Awash Insurance Company for their support and encouragement, which gave me strength to successfully complete this work.

Finally I would like to thank my family for supporting spiritually throughout my life.

# Table of Contents

DECLARATION .....	iii
Acknowledgments.....	iv
List of Acronyms .....	ix
List Figures .....	x
List of Tables .....	xi
Abstract.....	xii
CHAPTER ONE.....	1
INTRODUCTION .....	1
1.1 Background.....	1
1.2 Statement of the problem.....	4
1.3 Objective.....	10
1.3.1. General Objective.....	10
1.3.2. Specific Objective .....	10
1.4. Methodology.....	10
1.5. Scope/Limitations .....	11
1.6. Significance of the Study.....	12
1.7. Organization of the Study .....	12
CHAPTER TWO .....	14
LITERATURE REVIEW .....	14
2.1. Concept of Insurance .....	14
2.2. History of Insurance.....	15
2.3. Principles of Insurance.....	16
2.4. Medical Insurance .....	17
2.5. Risk and Insurance.....	17
2.6. Overview of Data Mining .....	18

2.7. Data Mining Methods, Techniques and Algorithms.....	20
2.7.1. Classification.....	21
2.7.2. Clustering .....	21
2.7.3. Prediction .....	22
2.7.4. Association rule.....	23
2.7.5. Neural networks .....	23
2.7.6. Data mining Algorithms.....	24
2.7.6.1. Naïve Bayes algorithm .....	24
2.7.6.2. Logistic Regression .....	25
2.7.6.3. Support Vector Machine (SVM) .....	26
2.8. Data Mining Models .....	27
2.8.1. The KDD Process Model .....	27
2.8.2. The CRISP-DM process model.....	30
2.8.3. The SEMMA process model.....	32
2.9. Application of Data Mining In Insurance Sector.....	33
2.9.1. Acquiring new customers.....	33
2.9.2. Customer level analysis.....	34
2.9.3. Customer Segmentation .....	34
2.9.4. Policy designing and policy selection .....	34
2.9.5. Prediction .....	34
2.9.6. Claims management .....	35
2.9.7. Developing new product lines.....	35
2.9.8. Underwriting and Policy management.....	35
2.9.9. Risk management .....	35
2.9.10. Reinsurance .....	35
2.9.11. Fraud detection.....	36
2.9.12. Trend analysis .....	36

CHAPTER THREE .....	37
RESEARCH METHODOLOGY .....	37
3.1. Research Purpose .....	38
3.2. Research Design.....	38
3.3. Cross-Industry Standard Process for Data Mining processing model.....	39
3.3.1. Business Understanding .....	39
3.3.2. Understanding of the data .....	41
3.2.2.1. Data Collection .....	41
3.2.2.2. Description of the Collected Data .....	44
3.2.3. Data Selection, Preparation and Preprocessing.....	45
3.2.3.1. Data Selection.....	45
3.2.3.2. Data splitting into Training and Testing Data .....	45
3.2.3.3. Data Cleaning .....	46
3.2.3.4. Data integration .....	47
3.2.3.5. Attribute selection.....	48
3.2.3.6. Data formatting.....	49
3.2.4. Data Transformation .....	49
3.2.5. Data modeling .....	51
3.2.6. Evaluation of the discovered knowledge .....	52
3.2.7. Use of the discovered knowledge .....	55
3.2.8. Deployment .....	55
3.2.8.1 Prototype .....	55
CHAPTER FOUR.....	57
EXPERIMENT AND DISCUSSIONS.....	57
4.1. Introduction.....	57
4.2. Experiment.....	58
4.2.1 Support Vector Machine (SVM) modeling.....	58

4.3. Naïve Bayes Model.....	64
4.4. Logistic Regression Modeling.....	67
4.6. Summary of the findings.....	70
4.7. Deployment.....	71
4.7.1. Prototype.....	71
CHAPTER FIVE.....	77
CONCLUSIONS AND RECOMMENDATIONS.....	77
5.1. Introduction.....	77
5.2. Conclusions.....	77
5.3. Recommendations and future works.....	80
References.....	82
Annexes.....	87
Annex 1.....	87
Annex 2.....	89



## List of Acronyms

PN	Policy Number
LALT	Life Assured Limit Type
TP	Total Premium
CA	Claim Amount
DM	Data Mining
AIC	Awash Insurance Company
ML	Machine Learning
KDD	Knowledge Data Discovery
CRISP-DM	Cross-Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, and Assess
LR	Logistic Regression
DM	Data Mining
RN	Risk Name
RTA	Rate to Be Applied
CSV	Comma Separated Values
ANN	Artificial Neural Network
MI	Medical Insurance
BP	Back propagation
SVM	Support Vector Machine
LR	Logistic Regression

## List Figures

Figure 3. 1: underwriting raw data.....	43
Figure 3. 2: Claim raw data.....	43
Figure 4. 1: Snapshot for installing necessary packages and loading the corresponding libraries	59
Figure 4. 2: Dataset snap shot for SVM model.....	59
Figure 4. 3: partitioning snapshot to training and test data for SVM model .....	60
Figure 4. 4: Building a Model snapshot for SVM .....	60
Figure 4. 5: Prediction and model accuracy snapshot for SVM model .....	61
Figure 4. 6: Confusion matrix result snapshot for SVM model.....	62
Figure 4. 7: ROC Curve result snap shot .....	63
Figure 4. 8: Dataset snap shot for Naïve Bayes model .....	64
Figure 4. 9: partitioning snapshot to training and test data for Naïve Bayes model.....	65
Figure 4. 10: Building a Model snapshot for Naïve Bayes model.....	65
Figure 4. 11: Prediction snapshot for Naïve Bayes model.....	65
Figure 4. 12: Confusion matrix result snapshot for Naïve Bayes model .....	66
Figure 4. 13: Dataset snap shot for Logistic Regression model .....	67
Figure 4. 14: partitioning snapshot to training and test data for LR model .....	68
Figure 4. 15: Building a Model and prediction snapshot for LR model .....	68
Figure 4. 16: model accuracy snapshot for Logistic Regression model .....	69
Figure 4. 17: Confusion matrix result snapshot for Logistic Regression model .....	69
Figure 4. 18: Saving a mode to specific directory .....	73
Figure 4. 19: Loading the saved model to R studio .....	73
Figure 4. 20:Prototype 1 .....	74
Figure 4. 21:Prototype 2 .....	75

## List of Tables

Table 3. 1: Description of collected data .....	44
Table 3. 2: Description of underwriting and claim data after the data integration .....	48
Table 3.3: Attribute Transformation .....	50
Table 3. 4: Premium to be calculated based on claim ratio .....	50
Table 3. 5: Life Assured Limit Type transformation .....	51
Table 3. 6: Risk Name transformation .....	51
Table 3. 7: confusion matrix for a two class classifier .....	53
Table 4. 1: Comparison of prediction models.....	71
Table 4. 2: Required parameters for a medical insurance customer .....	72

## Abstract

The main objective of study was to classify medical insurance customers with high claim ratio in order to take an appropriate measures during underwriting process to save profit making customers under medical insurance class of business. Globally insurance companies are spending high amount of claim costs due to medical insurance. It is a concern for companies to have a system that could differentiate whether the customers are profit making or loss incurring from upcoming claims. In the insurance industry the claim costs are needed to be minimized as much as possible. The main cause which result in high claim costs knowing profit making and loss incurring customers without the knowledge of claim experience in the company. To tackle the problem of high claim cost in medical insurance class of business, predictive data mining techniques has been employed using Support Vector Machine, Naïve Bayes and Logistic Regression predictive models. The dataset used for the experiment in this study was collected from Awash Insurance Company specifically from underwriting and claim data tables of medical insurance class of business. After cleaning irregularities and incomplete data in the dataset, a total of 41,151 records have been used to train the models in the ratio of 80:20.

To meet the aforementioned objective of the study, the CRISP-DM methodology, which involves six steps was adopted to undertake data mining process and to address the business problem systematically and iteratively. A six steps process model is used to guide the entire knowledge discovery process. Support Vector Machine, Logical Regression and Naïve Bayes classification algorithms are used to build predictive model.

Experiments are conducted and the resulting models show that the Support Vector Machine (SVM) is found to work well in classifying medical insurance customers with 99.39% classification accuracy. A prototype is developed based on the predictive model. Finally recommendations and future research directions are forwarded based on the results achieved.

*Key words: Predictive data mining, CRISP-DM, medical insurance class of business, SVM*

# CHAPTER ONE

## INTRODUCTION

### 1.1. Background

In every business, from the small corner store to the large manufacturer, there are common challenges with insurance, claims, and risk in general. Buildings can be damaged by fire, people may be sick, someone could slip and fall, vehicle accidents often occur, or losses can occur as a result of defective products. Now, more than ever, it is vital to the success of an organization to understand the extent of risk and control liability. ([www.clearrisk.com](http://www.clearrisk.com))

Currently, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Valuable bits of information are embedded in these data repositories. The only problem is that this storehouse of data has to be mined for useful information. Normally, these terabytes of transaction data are collected, generated, printed, stored, only to be filled and discarded after they have served their short-lived purposes as audit trails and paper trails (Rene, 2010 ).

Insurance is the pooling of money by a company from a group of people or organizations, to pay for the fortuitous losses that any of them may suffer. The money that the people pay to the insurance company is called the premium, and for this premium, the company promises to indemnify any of its customers for covered losses. The company may also provide risk management services.

Risk management has long been a topic worth pursuing, and indeed several industries are based on its successful applications, insurance companies and banks being the most notable. What gives this discipline enhanced attention and renewed prominence is the belief that nowadays we can do a better job of it. This perception is based on phenomenal developments in the area of data processing and data analysis. The challenge is to turn ‘data’ into information, knowledge and deep understanding [1]. Traditionally, actuaries develop risk models by segmenting large population of

policies into risk groups, each with its own distinct risk characteristics. Premiums are then determined for each policy in a risk group based on the risk characteristics of the group, such as mean claim rate and mean claim severity, as well as on the cost structure of the company, its marketing strategy, competitive factors, etc. [2].

However, the basic tenet in the industry is that no rating system can be perfect and competition therefore compels insurance companies to continually refine both the delineations they make among the risk groups and the premium they charge. The analytical methods employed by actuaries are based as much on statistical analysis as they are on experience, expert knowledge, and human insight [3].

Nowadays, there are two general categories of insurance coverage throughout the world; life and Non-Life insurance service (Olayungbo, 2015). Under Life insurance, medical insurance is one of the main insurance services provided by insurers where several insurance companies were formed to specifically underwrite medical insurance business long back 1890's (Rani and Gobena 2017). After acquisition of license from National Bank of Ethiopia as per the insurance business Proclamation No. 86/ 1994, insurance companies in Ethiopia transact insurance business under "general insurance/Non-Life" and/or "long-term/Life" insurance (NBE, 2012).

**Medical Insurance:** Medical insurance covers illness, disease or accidental bodily injury necessitating expenses in respect of treatment/service on the recommendation of a registered medical practitioner up to the annual maximum limit. In Awash Insurance Corporation medical insurance cover is provided in both life and non-life categories. In non-life, it is provided with personal accident and workmen's compensation insurance and in life, it is provided with medical insurance (individual medical and group medical). ([www.awashinsurance.com](http://www.awashinsurance.com))

**Personal and Group Personal Accident:** insures individuals and groups of individuals against the event of death, bodily injuries and medical expense resulting from an accident caused by violent, accidental, external and visible means. It is provided based on an agreed sum insured between the insurer and the insured. Under this insurance type standard, extended illness, worldwide and sport activity covers are provided.

Workmen's Compensation (Employers Liability): covers employees of an insured company against death or disability and medical expense due to work-related accident or disease. It can be extended to cover beneficiaries for workers accident to and from work and in residence. The cover is provided based on the monthly salary of the employees. This one is employer's legal liability insurance.

Individual Medical Insurance: provides payment of medical costs and charges incurred during illness and accidental injury. This policy can only be issued for individuals who have life insurance policy registered in their name. Legal dependents can be added as second insured (spouse) and child. The primary insured has to be self-supporting.

Group Medical Insurance: provides payment of medical costs and charges incurred during illness and accidental injury. This is a type of policy issued for groups with group members greater than 20 and it can be issued without the life insurance. Like that of the individual, each of the group member's legal dependents can be added and the primary insured should be self-supporting.

Unlike other classes of business, medical insurance doesn't require proportional re-insurance arrangement (Rani and Gobena 2017). As a result, the premium collected from medical insurance class of business remain in primary insurance company which motivate insurance companies for the wide spread of this class of business. Among life insurance class of businesses, medical insurance class of business takes the lion share in terms of premium contribution and claim cost which is a common feature in most developing countries (Alhassan and Biekpe, 2015). For Nile insurance reported the premium contribution of medical class of business in 2018/19 as 60% of the total portfolio and 85.7 claim ratio while Awash Insurance reported under the same reporting year premium contribution of medical insurance class of business 65.5% and claim ratio 87%. Thus, a close analysis in this class of business makes insurance companies more profitable and positive contribution to the industry as a whole. Accordingly, there is a need to predict and classify individual customers risk items as loss incurring and /or profit makers in the life class of business using data mining techniques.

## 1.2 Statement of the problem

According to (Berhe and Kaur, 2015) any profit making industry, insurance companies require earning profit to be sustainable in the current competitive environment. However, medical insurance class of business is not an attractive line of business in Ethiopian insurance industry and almost all insurance companies describe in their annual reports that medical insurance is consistently registered a negative results.

Moreover, an investigation of major factors which determine insurance company's profitability has been conducted by many researchers on insurance companies' profitability and came up with different findings. Among the many research works in the insurance claim, some of them have been given more attention in this study.

Yihenewu (2016) conducted to apply data mining clustering and classification algorithms. K-means clustering algorithm is implemented to come up with the natural group of the claim records. The researcher implemented two classification algorithms, J48 decision tree classification algorithms and multiperceptron ANN. Using j48 decision tree classification algorithms different experimentations are conducted. The first experimentation with default parameter values and 10 fold cross validation test options has registered 94.63% accuracy. ANN experimentations were conducted. Accordingly the experimentation with default parameter values with 10-fold cross-validation test option registered 99.58% accuracy. The study also registered an accuracy of 93.74% with percentage split test option by splitting the dataset into 80% to training set and 20% to test set. The result of this study indicates that applying data mining to classify insurance individual customers to predict the risk item is very promising. The above prediction accuracy also indicates that data mining is a powerful tool to measure the uncertainty of losses in EIC. Hence, the research identified future research direction in order to implement applicable system in risk assessment process.

Yunos, Ali, Shamsyuddin, and Ismail (2016), made their research on prediction of motor claim frequency and severity using Artificial Neural Network (ANN) in their paper titled "Predictive Modeling for Motor Insurance Claims Using Artificial Neural Networks". The researchers investigated the capability of ANN as a potential technique to be applied in modeling the motor



insurance claims problem. According to the authors, accurate predictive models in motor insurance claims are used for risk classification which is used as the formulation of different premiums scheme for the same coverage based on their category of the customers. The authors used Back Propagation Neural Network (BPNN) algorithm in ANN with a three-layer network structure of a back propagation (BP) learning algorithm to model the motor insurance claims. For the model evaluation, the researchers used 4 statistical methods, these are mean squared of error (MSE), root mean square of error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The result shows among the statistical methods MAPE error represent a smallest value of error for claim frequency and claim severity. Moreover, the authors described that BPNN model is successful in predictive modeling the Malaysian motor insurance claims by using several of network structures.

Burri, Bojja and Buruga (2019), conducted their research paper titled “Insurance Claim Analysis Using Machine Learning Algorithms” and explained the Machine language as a data facilitator which to be converted to knowledge for decision making. According to Burri, Bojja and Buruga (2019) insurance companies are extremely interested in the prediction of the future. This is because accurate prediction gives them a chance to decrease financial loss for the company in connection with claim cost. Moreover, forecasting the upcoming claims helps to charge competitive premiums that are not too high and not too low. It also contributes to the improvement of the pricing models. This helps the insurance company to be one step ahead of its competitor. The authors used Naïve Bayes Updatable, Naïve Bayes, Multi-Layer Perceptron, J48, Random Tree, Logistic Model Tree (LMT), and Random Forest for this research paper and the result shows Logistic Model Tree (LMT), Random Forest algorithms have given better claim prediction when compared with the rest of classification algorithms.

Dandi (2015), In this study an attempt is made to reveal the high potential of data mining applications for customer segmentation, referring to the optimal usage of data mining methods and techniques to thoroughly analyze the collected historical data and to segments life insurance customers of Ethiopian Insurance Corporation based on their value. The combination of clustering and classification data mining techniques were conducted to build customer segmentation model based on customers value computation. To achieve the objective of the study, K- means clustering

algorithm is applied to identify the characteristics of life insurance policy holders who contribute high or low value to the corporation. Different user-defined parameters were applied to achieve good clustering model. Based on the results of clustering model, J48 algorithm of decision tree was implemented to predict the potential value of customers and to identify the most significant variables that could help to segment life insurance customers based on their values. During the study, decision tree models accuracy rate of the decision tree models shows slight differences when run information were set with different parameter and tests option. High performance (99.98 %,) is registered when cross validation test option and default parameter is set in the run information. This study revealed that a good customer segmentation model can be built by combining K- means clustering and J48 classification algorithms. Besides, J48 decision tree algorithm showed good quality to visualize the clusters and to understand the clusters that lead to profile customers.

Lijia Guo et al., [4] focused on property/casual insurance and described the data mining techniques which used for the process of property insurance. T. L. Oshini Goonetilleke et al., [5] summarized the analysis of customer analysis when mining a life insurance data. He focused on the customer retention and implemented the techniques for customer attrition. That recommended that problem of attrition analysis of life insurance domain was successfully removed with the implementation of data mining techniques.

According to Hintsay (2016), classifying customers based the risk they involve in the specified insurance company was one the major problems noticed over the year. So the researcher built a classification models using decision tree and neural network. The predefined classes of risks in the study were: low medium and high risk classes. The tools selected in the study were See5 and the Brain Maker Software. Those tools can support decision tree classification and neural network respectively. He pointed see5 software is best for selecting attributes for decision tree classification and the BrainMaker software was selected to train and to build neural network model software .A decision tree classifier, is mainly used in this study for attribute selection that could be used as an input for the neural network. He said that both tools have the facility to partition the dataset randomly into training and testing sets. From the four branches of NISCO, 1332 records were collected. The total variables of the records are 25. Because of the missing values four attributes were totally discarded and other six attributes were extracted from the existing ones. For the study

the data sets selected were 1160. By considering the classes for low risk 629, for medium risk 305 and for high risk 226 were resulted. This record classification where made based on the assessment report made on strong and weak points of the policy's. But the data mining techniques can improve the result by models such as decision tree and neural networks .Then the 1160 dataset were divided into two: 90% (1044) for model building and testing set and the remaining 10% (116) for validation set. He said out of 1044 records which were selected for model building and testing set — 940 (90%) of the 1044 facts were used for training and the remaining 104 (10%) were used for testing purposes. Finally the results for decision tree classification , according to the three predefined classes(low, medium and high risk policies ) the classification accuracy was 98.15%, 94.12%, and 92.86% respectively and the validation test was correctly classified 95.69% of the validation. Next, the neural network model correctly classified 92.24 % of the validation set. He said that high-risk groups are correctly classified and the remaining, low and medium-risk groups, the classification accuracy were of 98.15% and 76.47% respectively.

Another research entitled as “Application of Data Mining for Customer Segmentation: The Case of Buusaa Gonofa Microfinance Institution” was conducted by Reganie (2013). The researcher said due to lack of appropriate tool to segment customers before looking potential customers, the intuition couldn't identify value customers and measure their profits. The goal of the study is to build a model that can help to classify customers for of Buusaa Gonofa microfinance institution. The methodology selected by the researcher was CRISP- DM. The researcher used clustering techniques to segment customers into appropriate number of clusters. K means clustering algorithm was used to cluster customers which exhibit similar characteristics. Then after, he built a predictive model to predict potential customers. For that reason or classification purpose, J48 algorithm was used. Finally, the predictive model achieved 99.5% accuracy. The researcher concluded that, the result of predictive modelling is encouraging. Perhaps, data mining technology can be applied in micro finance industries in order to extract interesting patterns and knowledge. He recommended the institution should consider developing an integrated warehouse to apply data mining techniques.

Taking into account the content of these research papers, different issues related to insurance claim have been addressed. For instance, insurance claim analysis using machine learning algorithms insurance companies are extremely interested in the prediction of the future. This is because

accurate prediction gives them a chance to decrease financial loss for the company in connection with claim cost. Moreover, forecasting the upcoming claims helps to charge competitive premiums that are not too high and not too low. The analysis of policyholders claim information is the other issue that have been addressed. Categorization of policyholders according to the claim cost they brought to the pool was another focal point in the research papers. The purpose of the analysis was to allocate higher premium charge to customers who brought more claim cost to the company. Insurance risk understanding was another very important factor that have been discussed in the research paper. It is also disclosed that understanding the risk that the insurance company assumed during underwriting helps to adjust the premium amount to be collected from the insured.

The aforementioned related works helped the researcher, different issues related to insurance claim have been addressed. From the analysis we can understand that these research papers gave more attention either to find efficient predictive models for insurance claim prediction or finding a way to categorize customers based on the claim amount that they brought to an insurance. Contributing a lot in claim prediction as it is attempted to analyze, most of the studies did not address medical insurance individual customer risk item claim cost analysis in their study. The importance of medical insurance individual customer risk item study of insurance claim prediction is saving other individual customer risk items which are profit makers.

Regarding to profitability of insurance industry in Ethiopian there are some research works conducted so far. Some of the authors who conducted insurance industry include:

*Behailu Kebede (2016), "Factors Affecting Insurance Companies Profitability In Ethiopia."*

*Mehari and Aemiro (2013), "Determinants of insurance companies' profitability Analysis of insurance sector in Ethiopia"*

*Demise and Hailegebreal (2016), "Determinants of insurance companies' profitability Analysis of insurance sector in Ethiopia"*

*Mingizem Birhan (2017), "Determinants of Insurance Company profitability in Ethiopia"*

All these researches conducted on Ethiopian insurance industry mostly focused on assessment and identification of factors affecting the profitability of the industry rather than individual customer

risk item claim cost analysis which is actually the main problem of the industry that should be addressed.

From the discussion made with branch managers of AIC who have given a privilege to settle claims, the critical issue in medical insurance class of business is classifying specific individual customer risk-items which eat the premium of others because of their high claim cost. The life and health manager agreed that medical insurance class of business is not a loss incurring business in general. Some individual customer risk categories from medical insurance of business are very profit making and others are loss incurring. So, the issue is how to classify this individual customer risk category or risk items from the others in the same class of business which is medical class of business. Moreover, senior underwriters and branch manager of the company, emphasized that classification of individual customer risk items in the insurance industry is the most important issue in all Ethiopian insurance companies in general and for their branch offices in particular.

According to branch managers, a study conducted at individual customer risk item helps the industry to prepare their strategic plan accordingly whereas for branches especially in AIC insurance company, most benefits including annual salary increment and bonus for all employees depends on the claim ratio they registered, which is directly related to the profitability of the branch.

Therefore, this study applies data mining technique to classify medical insurance customers risk claim cost at individual customer risk item under medical insurance class of business, more specifically to classify which individual customer risk items brings more claim cost to the pool.

The proposed study was aimed to answer the following research questions:

- How to develop predictive model and employed to predict loss incurring customers that brings high claim cost?

## 1.3. Objective

### 1.3.1. General Objective

The main objective of this study is to classify medical insurance customers based on claim experience using data mining technique by taking Awash Insurance Company S.C. as a case.

### 1.3.2. Specific Objective

To achieve the general objective of the study, the following specific objectives are identified.

- To collect relevant dataset required for the mining, analysis and performance evaluation.
- Preparing the data that will be used for model building by selecting important attributes, and cleaning them.
- To identify attributes that predict customers with high claim cost.
- To explore and identify best classification algorithm that are suitable for data analysis.
- To evaluate the performance of the classification model in characterizing the medical insurance customers of the company.
- To report on the results and make recommendations for further researches.

## 1.4. Methodology

This research is an experimental research conducted on medical insurance claim data collected from Awash Insurance Company. The purpose of this study is to uncover the hidden risk item record patterns in the database of Awash Insurance Company. There are different data mining process models available among which Cross-Industry Standard Process for Data Mining (CRISP) is widely user process model (Olegas Niaksu, 2015). Taking into account the factors mentioned so far, under this study, the researcher also preferred Cross-Industry Standard Process for Data Mining (CRISP) and R programming for the experiment to be conducted.

There are different types of standards and methodologies being used in DM researches. CRISP-DM is the most widely used and highly recommended model to be used for DM researches and projects in several industries including the banking industry. CRISP-DM evolved to become the de facto industry standard (Chapman et al., 2000; Jackson, 2002; Kulikowski, 2011; KURGAN &

MUSILEK, 2006; MARISCAL, MARBAN, & FERNANDEZ, 2010). It is also stated as a neutral process model that can be used with any tool or application by any industry (Gilchrist, Mooers, Skrubbeltrang, & (Corresponding, 2012). It is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem (Ponce & Karahoca, 2009). For these reasons, the researcher adopted the CRISP-DM model to achieve the intended outcomes of the research. The CRISP-DM has six steps that are: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

This study also considered Support Vector Machine, logistic regression models which preferable model for Categorical variable types and Naïve Bayes which is effective and powerful algorithm for predictive modeling (Huang and Lei, 2011 and R programming will be employed as the building tool. In this study the data is collected from Awash Insurance Company database which includes underwriting and claim data.

## **1.5. Scope/Limitations**

The insurance industry deals with risks of various classes of business such as marine, medical, motor private, commercial motor, fire and burglary, and engineering. The scope of this research is to examine the potential feature in Awash Insurance Company medical insurance cover is provided in life provided with Medical Insurance class of business. Medical insurance class of business is one of the most risky class of business. This research attempts to apply data mining techniques to classify medical insurance customers as profit making or losing incurring customer based on their claim cost. Towards this end, a predictive model will be created using classification algorithms.

The limitation of the research is, bounded to Awash Insurance Company's medical insurance class of business provided in life categories; giving more due attention to the claim they incur in the company. All the required data will be collected from Awash Insurance Company database, because of their underwriting and claim records of the customers in the company. Some important attributes like age, location, and medical history of the customer which are very important for this research are not captured in most cases.

## 1.6. Significance of the Study

The aim of this study is to explore the applicability of data mining techniques in the insurance industry and build models that can classify medical insurance customers based on their claim experience. Based on these, the subsequent benefits can be gained from the finding of this study.

- The research is believed to initiate further research in the area for exploiting the potentials of data mining techniques in the company.
- The output of this study provides information to Awash Insurance S.C to determine whether the medical insurance customers are profit making or loss incurring based on claim experience. Those attribute that belong to loss incurring customers take the lion share for the decrease in profitability of medical insurance. This gives the company a chance to filter loss incurring attributes on medical insurance which enables the company to minimize the loss incurred by medical insurance.
- In this research an attempt was made to find out the applicability of data mining technology in the insurance industry. The result of the study shall be used as an input for the development of full-fledged data mining application in supporting insurance claim activity.
- Although the study was aimed at addressing insurance problems in particular, the output of the study may be used as a source of methodological approach for studies dealing with the application of data mining technology on similar problem areas.
- Moreover, the study findings can provide insight for further researches to apply data mining technologies to advance insurance industry.

## 1.7. Organization of the Study

This thesis consists of five chapters. The first chapter deals with the general overview of the study including background of the study, background of the company, statement of the problem, objectives, significance of the study and scope and limitation of the study. The second chapter covers literature review regarding data mining technology. Here the knowledge discovery process tasks with the specific methods and algorithms are discussed. Under this chapter data mining tools and application of data mining in general for insurance sector is also covered through review of related works. The third chapter discusses methodology of the research in which the techniques



tools and methods are explained using the selected process model. In this part specific data mining algorithms which are used. The fourth chapter is deals about data analysis and presentation of findings. The last chapter is devoted to summary of major findings concluding remarks, and recommendations forwarded based on the research findings of the present study.

## CHAPTER TWO

### LITERATURE REVIEW

#### 2.1. Concept of Insurance

Insurance can be defined from individual and economic point of view. From individual point of view insurance is economic device where by the individuals substitute small certain cost (the premium) for a large uncertain financial loss (the contingency insured against) that would exist if it were not for the insurance. From the society point of view, insurance is an economic device for reducing and eliminating risk through the process of combining a sufficient number of homogenous exposures in to a group to make the losses predictable for the group as a whole. (Emmett & Vuauohan, 2008)

According to Osterville (1998), insurance is often defined as a contract of indemnity. That means the insured is not to make any profit out of the insurance but should only be compensated to the extent of the financial loss. Moreover, insurance is a mechanism (or a service) for the transfer of risk of financial loss to someone else called the insurer in exchange of the payment of an agreed fixed amount known as premium. The insured should pay the premium before the contingent claim is serviced by the insurer (Nowadays this is not an issuer in Ethiopia insurance industry since the National Bank of Ethiopia declared “No premium No cover” policy in August 2012). From the insured's point of view, insurance is a transfer of risk whereas from the insurer's point of view, insurance as a "pooling" mechanism of a large number of exposure units or risks (Osterville, 1998).

Insurance can be defined from individual and economic point of view. From individual point of view insurance is economic device where by the individuals substitute small certain cost (the premium) for a large uncertain financial loss (the contingency insured against) that would exist if It were not for the insurance. From the society point of view, insurance is an economic device for reducing and eliminating risk through the process of combining a sufficient number of homogenous exposures in to a group to make the losses predictable for the group as a whole.(Emmett & Vuauohan, 2008)

## 2.2. History of Insurance

The first modern insurance service in Ethiopia was underwritten by Bank of Abyssinia around 1905 .This bank was transacting fire and marine insurance. According to Economic Progress of Ethiopia (1955) as cited in Hailu (2007) the first survey done by Ministry of commerce in 1954 revealed as there were 17 overseas insurance companies operating in the country .Consequently the second survey which was done in 1960, revealed as there were 33 overseas agents .The only local insurance company at that time was Imperial Insurance. The history of life insurance in Ethiopian dates back to1954 where the service was given by the two foreign companies Central Assurance Company Limited and Groupement Francis de Assurances located in Addis Ababa and Dire Dawa.

During this time the insurance service was regulated with the provisions of the Commercial code (1960) and there was no specific proclamation to the insurance industry. The issuance of proclamation number 281/1970 which was aimed to regulate the insurance business in Ethiopia brought remarkable change on the structure of the industry. The first responsible, independent body to regulate the insurance industry, the insurance council and insurance controller office, was created due to this proclamation .In respect to this proclamation the Ministry of commerce. Trade and Tourism issued Regulation (Legal Notice No. 393/ 1971) aiming to implement the proclamation and create conducive insurance market. The regulation resulted in the licensing of fifteen domestic insurance companies, thirty six agents, seven brokers, three actuaries and eleven assessors. In 1972 the number of insurance companies goes down to thirteen. Among these companies, Ethiopian life insurance and American life insurance companies were offering life insurance whereas Lion insurance was offering both life and general insurance .In 1974, just before the Ethiopian revolution, there were seventeen insurance companies, four were giving life assurance, thirteen were giving marine insurance, eleven were giving General Accident Insurance and all of them were giving Fire insurance. The revolution brought a command economy with a socialist credo which resulted in the nationalization of the private insurance companies. The then existed Government decided on January 1, 1975 to transfer the ownership of insurance companies to the Government, as a result those insurance companies were all nationalized. In December 1975 the government issued proclamation No 68/1975 to establish Ethiopian Insurance Corporation; made an autonomous public enterprise effective from January 1, 1976.The insurance council and

office of insurance controller also dissolved. The 1976 proclamation (proclamation 99/19760 empowered national bank to supervise and regulate banks and other financial institutions including insurance. EIC monopolized the insurance market till 1994 (Hailu, 2007).The collapse of the Marxist regime in 1991 resulted in a market economy. The new proclamation (Proclamation Number 86/1994) issued and allowed the re-emergence of private insurance companies and prohibited foreign insurers to operate in Ethiopia .And also mandated the National bank of Ethiopia as the supervisory authority. In addition to restricting the insurance business to domestic investors, it brought structural change to the industry by classifying the business in two, namely the general insurance business and long term insurance business. The long term insurance include life insurance, Medical insurance, pension and disability insurance whereas the general insurance business is property and causality insurance .The proclamation enforced to have a separate account for each class of business (Hailu, 2007).

### **2.3. Principles of Insurance**

Insurance coverage has its own principles which are dedicated and identifiable only for insurance business. The main objective of every insurance contract is to give financial security or compensation and protection to the insured from any future uncertainties and for that to happen, insured must never ever try to misuse this safe financial cover (Akrani, 2011). According to Sibindi (2013), in addition to the common law of contract, the following rules in particular are considered to be the basic principles of insurance and these insurance principles are wel respectful by both parties called the insured and the insurer all over the world (Akrani 2011)

These insurance principles are known as “Principle of Utmost Good Faith” which is the insurance contract must be signed by both parties (insurer and insured) in an absolute good faith or belief or trust, “Principle of Insurable Interest” which is to say a person has an insurable interest when the physical existence of the insured object gives him some gain but its non-existence will give him/her a loss, “Principle of Indemnity” which stands for an insurance Contract signed only for getting protection against unpredicted financial losses arising due to future uncertainties, “Principle of Contribution” means if the insured has purchased more than one policy on the same subject matter. According to this principle, the insured can claim the compensation only to the extent of actual loss either from all insurers or from any one insurer, “Principle of Subrogation”, is applicable when

the insured is compensated for the losses due to damage to his insured property, then the ownership right of such property shifts to the insurer, “Principle of Loss Minimization” which states that the insured must take all possible measures and necessary steps to control and reduce the losses in such a scenario and “Principle of Causa Proximate” this is when a loss is caused by more than one causes, the proximate or the nearest or the closest cause should be taken into consideration to decide the liability of the insurer(Akrani, 2011).

In broad sense, the insurance market is based on two fundamental characteristic that is the transfer of exposure from a single party to a large group and the sharing of all losses by all those in the group. This implies that insurance relies heavily on the Law of Large Numbers which is known as pool management (Sibindi, 2013).

## **2.4. Medical Insurance**

Medical insurance provides for the payment of the costs of medical care that result from bodily injury. Its benefits help to meet the expenses of physicians, hospital, nursing and related services, as well as medications and supplies. Benefits may be in the form of reimbursement of actual expenses may be paid directly to the provider of the services or to the insured. This insurance provides payments in relation to services rendered by a doctor and for many types of expenses associated with hospitalization. This protects the insured from financial loss due to medical care costs. Medical insurance pays all or part of hospitalization, surgery, laboratory tests, medical and other medical care.

## **2.5. Risk and Insurance**

Profitable underwriting depends on accurate risk assessment (Dockrill, et al, 2001).Underwriting management identify the risks associated with a class of business, evaluate the degree of risk involved in terms of severity and frequency, and other external factors. The risk assessment process requires relevant and detailed statistical data, which is available from both internal and external sources. Internally, and most relevant, is the data extracted from the insurer’s own policy and claims records. Externally, insurers have long recognized the value of exchanging claims and related data. It should be noted that merely gathering information, however relevant, serves little purpose. Underwriters must organize the data in such a way that they are enabled to make the

necessary judgments. The data must also be available in a manageable format. As risk can never be certain or predictable, it needs a kind of management. The risk management is nothing but a method to prejudge the risk that may come up sometime in future. It is not prediction but a process of reducing the risk to a minimum level (Sisk, 2018). Risk management involves a number of measures that are used to keep the risk at possible minimum level. In our day to day life also we take many steps to keep the risk at lower level for example most people do not keep valuables at home and rather prefer to keep them in a bank locker by paying certain locker rent to the bank. Risk of life, health or property is reduced by purchasing a proper insurance. All these actions of individual persons are done under fear of uncertainty and unpredictability of future. Likewise in business and commerce also an element of fear of loss always exists if the risk components are not managed properly. Risk is a fear of happening something adverse and in order to prevent such adverse happenings a plan has to be in place to overcome such adverse happenings which is called as risk management (Sisk, 2018).

From insurance company's point of view, everything that the customers or insured brings to the insurance company to be insured are known as risk item. Medical insurance risk is the claim cost that the company is liable to paid to the extent of the sum insured of the customer. For many reasons, the probability of happening of claim for medical class of business is very high compared to other class of other life insurance businesses which is evidenced by annual claim report of almost all insurance companies in Ethiopia. Accordingly throughout this study the risk item refers anything brought to the insurance company to be insured.

## **2.6. Overview of Data Mining**

The definition of data mining or Knowledge Discovery in Databases is the action that extracts some new important information contained in large databases. The target of data mining is to find unexpected characteristics, hidden features or other unclear relationships in the data based on techniques' combination. Today, many applications in a wide and various ranges of business founded and worked in this regulation [6].

According to Fayyad et.al. (1996) note, historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information

discovery, information harvesting, data archaeology, and data pattern processing. Furthermore, the term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities and also gained popularity in the database field.

Berry et.al. (2000), although the rapid pace of change in the past century was felt in nearly every area, it is hard to find examples of anything, anywhere, that has changed as fast as the quantity of stored information. They assert that this information explosion has created new opportunities and new headaches in every field, ranging from marketing to medicine to manufacturing.

Kumar et al (2008) the health care environment is generally perceived as being ‘rich in information’ yet having ‘knowledge poor’. There is a wealth of data available within the health care systems. Baylis (1999) states health care generates large amounts of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce large amounts of clinical data. This data is a strategic resource for health care institutions.

According to Witten and Frank (2005) states that lack of data is no longer a problem at the current stage. However, the inability to generate useful information from data is the problem. As the volume of data increases inexorably, the proportion of people understands decreases, alarmingly. Lying in hidden data, in all these data potentially useful information, i.e. rarely made explicit or taken advantage of.

Data mining is the process of discovering patterns in the large data sets. The purpose of the data mining is to find information from the large data sets and convert it into usable structures so that this information can be used for further processing without any difficulty. It is handled by databases and managed by database management aspects. This is a commonly used word for any kind of large scale data processing. The term data mining was discovered around 1990 in computer science. It is also referred by several other terms like Knowledge Discovery in Databases (KDD) or Predictive Analytics or Data Science ([www.wikipedia.com](http://www.wikipedia.com)). Data mining is generally an iterative and interactive discovery process. The goal of this process is to mine patterns,

associations, changes, anomalies, and statistically significant structures from large amount of data [7]. The mined results should be valid, novel, useful, and understandable.

Data mining can be defined as the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns. In the insurance industry, data mining can help firms gain business advantage. For example, by applying data mining techniques, companies can fully exploit data about customers' buying patterns and behavior – as well as gaining a greater understanding of their business to help reduce fraud, improve underwriting and enhance risk management. This paper discusses how insurance companies can benefit by using modern data mining methodologies and thereby reduce costs, increase profits, acquire new customers, retain current customers and develop new products. Data mining methodology often can improve upon traditional statistical approaches to solving business solutions. For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. Data mining often can improve existing models by finding additional, important variables, identifying interaction terms and detecting nonlinear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. [8]

## **2.7. Data Mining Methods, Techniques and Algorithms**

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. [9]

Machine learning methods are commonly categorized as either supervised or unsupervised learning methods. Supervised approaches require both the input (predictors) variables and the output (response) variable, whereas, unsupervised approaches rely solely upon the input (explanatory) variables [10]. The following are a few of data mining methods with the corresponding tasks.



### 2.7.1. Classification

Classification is the most commonly applied data mining task which employs a set of pre-classified examples with target categorical variable (training set) to develop a model or function that can classify examples whose class labels are unknown (test set). The data classification process involves learning and classification steps. In the learning step, the training data set is analyzed by classification algorithm. In classification step, test data set is used to estimate the accuracy of the model built in the learning step. If the accuracy is acceptable, the model can be applied to new data tuples [11].

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

### 2.7.2. Clustering

Clustering is grouping of a set of objects (whose classes are unknown) into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same

cluster and are dissimilar to the objects in other clusters, that is, the objects are clustered so that the intra-class similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. The objects are assigned to their respective class and their common features in the cluster are summarized to form the class description. Clustering is a type unsupervised learning that can be employed for the purpose of identifying classes of objects in a data set prior to classification task so that cost of classification can be minimized. For instance, cluster analysis can be applied to categorize genes with similar functionality [11].

Examples of clustering methods:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

### **2.7.3. Prediction**

Unlike classification, prediction is used to predict the value of a dependent continuous variable, rather than a categorical label. Hence, regression and numeric prediction are synonymously used. Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable [11].

Regression technique can be adapted for prediction. Regression analysis can be used to model the Relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to

forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

#### **2.7.4. Association rule**

Association is usually used to find frequent item sets among large data sets. It is the process of finding attributes which ‘go together’, mostly in transactional data sets. Association is also known as affinity analysis or market basket analysis. Association rules are of the form ‘If antecedent, then consequent’, together with a measure of the support and confidence associated with the rule. For instance, association rule methods can be applied to determine proportion of cases in which a new drug will exhibit dangerous side effects [10].

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible association rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

#### **2.7.5. Neural networks**

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends

that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example, handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Types of neural networks: Back Propagation

### **2.7.6. Data mining Algorithms**

For this study three predictive models known as SVM, Naïve Bayes and Logistic Regression are used SVM which is generally capable of delivering higher performance for small and medium dataset in terms of classification accuracy (Srivastava and Bhambhu, 2010). According to the author, SVM is generally are capable of delivering higher performance for small and medium dataset (in our case 41,151 records) in terms of classification accuracy. SVMs can learn a larger set of patterns and be able to scale better (R Joseph, Hlomani and Letsholo, 2016). Moreover, SVM has the ability to update the training patterns dynamically whenever there is a new pattern during classification and able to model complex nonlinear decision boundaries and are less prone to over fitting than other methods.

#### **2.7.6.1. Naïve Bayes algorithm**

Naïve Bayes is proved as one of the most efficient and effective algorithms for data mining. An explanation is granted how this algorithm performs and its classification efficiency is high [12]. In this work it is proved that what eventually affects the classification optimality of Naïve Bayes is the distribution of dependencies among all attributes which violates the assumption of class conditional independence i.e. the effect of an attribute value on a given class is independent of the other attributes. Han and Kamber (2006) States that different studies found that the simple Naïve Bayes classification have comparable performance results with decision tree and neural network classifiers and have high accuracy and speed when applied to large databases [13].

Naïve Bayes is found simple but effective classification algorithm in solving real life problems [14]. This algorithm also performs well even when attribute dependencies exist [15]. Different modifications of this algorithm have been introduced by research communities in the area of

statistics, data mining, machine learning and pattern recognition. The researches and explanations given on the algorithm revolve around the assumption of independence. Extensions of the algorithms are made basically by increasing its tolerance of attribute independence or reduce tolerance of dependency but the results do not necessarily lead to significance improvements. [16][17] After referring the different modifications concluded that such modifications lead to complications which deviate from its basic simplicity.

### 2.7.6.2. Logistic Regression

Logistic Regression is intrinsically simple, it has low variance and so is less prone to over-fitting and it is faster and more reliable when the dimension gets large. Moreover, Logistic Regression can easily update the model to take in new data (using an online gradient descent method) (R Joseph, Hlmani and Letsholo, 2016).

(Hlosta, Stríž, Kup, Zendulka, & Hruška, 2013) LR is a machine learning model for binary classification. The method can handle both numeric and categorical variables. Given a learned model, the value of the output variable is computed by applying the logistic function to linear combination of attribute values and weight vector.

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable,  $Y$ , from one or more response variables,  $X$ . However, unlike linear regression the response variables can be categorical *or* continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for non normally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables, which is not always met in orthoscopic datasets. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions. One drawback of LR is that the method cannot produce typicality probabilities (useful for forensic casework), but these values may be substituted with nonparametric methods such as ranked probabilities and ranked interindividual similarity measures (Ousley and Hefner, 2005).

Logistic regression is used for predicting a binary variable. It is a generation of linear regression; the binary dependent variable cannot be modeled directly by linear regression. Logistic regression is a classification tool when used to predict categorical variables such as whether an individual is likely to purchase or not, and a regression tool when used to predict continuous variables such as the probability that an individual will make a purchase. There are several classification and regression techniques including decision trees, neural networks etc. (Dunham, 2006).

### **2.7.6.3. Support Vector Machine (SVM)**

Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities.

Machine learning overlaps with statistics in many ways. Over the period of time many techniques and methodologies were developed for machine learning tasks [18].

SVM which is generally capable of delivering higher performance for small and medium dataset in terms of classification accuracy (Srivastava and Bhambhu, 2010). R. Burbidge *et. al.*, have shown that the support vector machine (SVM) classification algorithm, proves its potential for structure–activity relationship analysis. In a benchmark test, they compared SVM with various machine learning techniques currently used in this field.

The insidious use of Support Vector Machine (SVM) in various data mining applications makes it an obligatory tool in the development of products that have implications for the human society. SVMs, being computationally powerful tools for supervised learning, are widely used in classification, clustering and regression problems. SVMs have been successfully applied to a variety of real-world problems like particle identification, face recognition, text categorization, bioinformatics, civil engineering and electrical engineering *etc.*

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression [19]. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression

prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. Support vector machine was initially popular with the NIPS community and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task [20]. It is also being used for many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik [21] and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization

(SRM) principle, which has been shown to be superior, [22], to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, whereas ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems [23].

## **2.8. Data Mining Models**

Currently there are many data mining process models available for data mining. Some of them are Knowledge Discovery in Database (KDD), Sample, Explore, Modify, Model, and Assess (SEMMA), Cross-industry process for data mining (CRISP-DM) etc. are well known data process models.

### **2.8.1. The KDD Process Model**

KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, Artificial Intelligence, knowledge acquisition for expert systems, data visualizations, and high-performance computing (Cios, Pedrycz, Swiniarski, and Kurgan, 2007). The main goal here is extracting high-level knowledge from low-

level data in the context of large datasets. The data mining, as a component of KDD, uses combined techniques from machine learning, pattern recognition, and statistics to find patterns. KDD focuses on the overall process of knowledge discovery from data including how the data are stored and accessed, how algorithms can be scaled to massive datasets still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported (Sciences Applied, 2010). A driving force behind KDD is the database field. Indeed, the problem of effective data manipulation when data cannot fit in the main memory is of fundamental importance of KDD. Database techniques for gaining efficient data access, grouping and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger datasets. A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps to set the stage for KDD. (Sciences Applied, 2010).

Fayyad (1996) defined data mining as a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumerations of patterns over the data. According to this definition data mining is the step that is concerned with the actual extraction of knowledge from data. To emphasize the necessity that data mining algorithms need to process large amounts of data, the desired patterns has to be found under acceptable computational efficiency limitations. Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (Fayyad, 1996). Based on this definition, data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. This means that the objective of the data mining exercise plays no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as “secondary” data analysis.

The KDD or Knowledge Discovery Databases [24] is the process of extracting the hidden knowledge according from databases. KDD requires relevant prior knowledge and brief



understanding of application domain and goals. KDD process model is iterative and interactive in nature. There are nine different steps or stages of this model and these are given below.

- Developing and Understanding of the application domain: this is the first stage of KDD process in which goals are defined from customer’s view point and used to develop and understanding about application domain and its prior knowledge.
- Creating a target data set: this is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables. It is an important stage because knowledge discovery is performed on all these.
- Data cleaning and pre-processing: this is the third stage of KDD process which focuses on target data cleaning and pre-processing to complete and consistent data without any noise and inconsistencies. In this stage strategies are develop for handling such type of noisy and inconsistent data.
- Data transformation: this is the fourth stage of KDD process which focuses on transformation of data from one form to another so that data mining algorithms can be implemented easily. For this purpose different data reduction and transformation methods are implemented on target data.

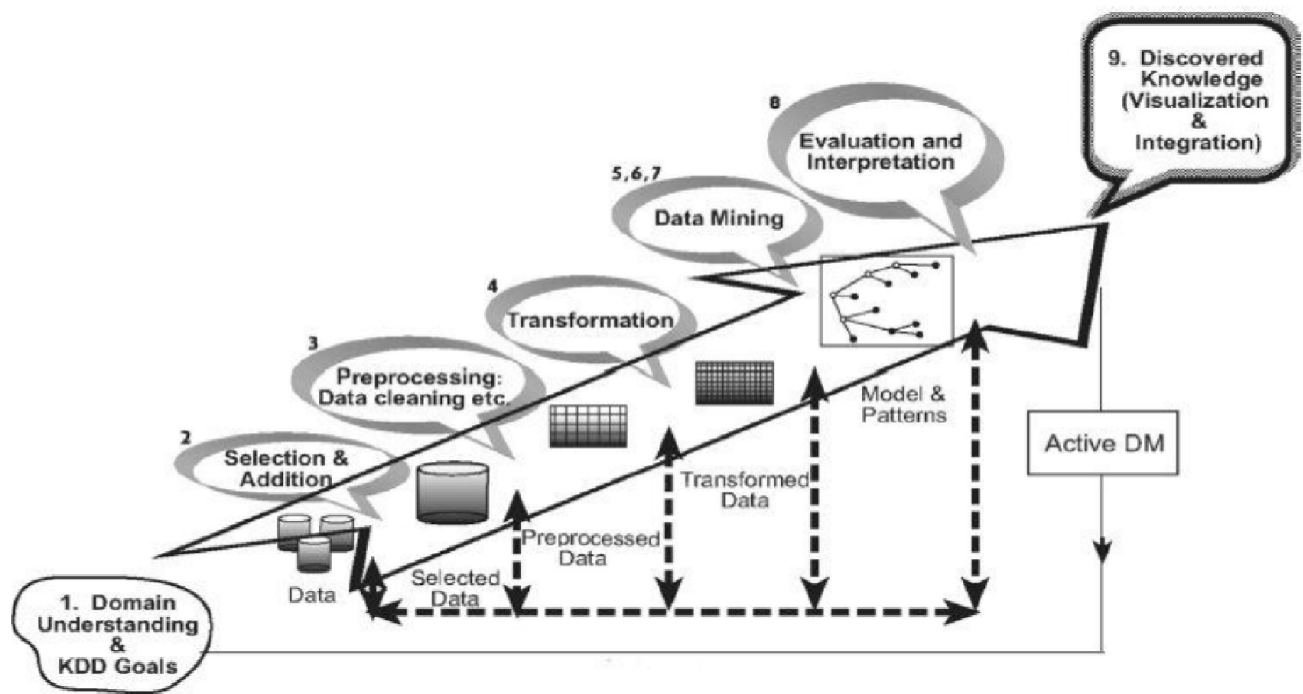


Figure 2. 1: Knowledge discovery databases (KDD) process model

- Choosing the suitable data mining task: this is the fifth stage of KDD process in which appropriate data mining task is chosen based on particular goals that are defined in first stage. The examples of data mining method or tasks are classification, clustering, regression and summarization etc.
- Choosing the suitable data mining algorithm: this is the sixth step of KDD process in which one or more appropriate data mining algorithms are selected for searching different patterns from data. There are number of algorithms present today for data mining but appropriate algorithms are selected based on matching the overall criteria for data mining.
- Employing data mining algorithm: this is the seventh step of KDD process in which selected algorithms are implemented.
- Interpreting mined patterns: this is the eighth step of KDD process that focuses on interpretation and evaluate of mining patterns. This step may involve in extracted patterns visualization.
- Using discovered knowledge: this is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used interested parties or can be integrate with another system for further action.

### **2.8.2. The CRISP-DM process model**

Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR in 1999, CRISP-DM 1.0 version was published and is complete and documented. It provides a uniform framework and guidelines for data miners. It consists of six phases or stages which are well structured and defined [25]. These phases are described below.

- Business understanding: this is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
- Data understanding: this is the second phase of CRISP-DM process which focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information.

- Data preparation: this is the third phase of CRISP-DM process which focuses on selection and preparation of final data set. This phase may include many tasks records, table and attributes selection as well as cleaning and transformation of data.
- Modeling: this is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem.
- Evaluation: this is the fifth stage of CRISP-DM process which focuses on evaluation of obtained models and deciding of how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether achieves the objectives properly or not.
- Deployment: this is the sixth and final phase of CRISP-DM process focuses on determining the use of obtain knowledge and results. This phase also focuses on organizing, reporting and presenting the gained knowledge when needed.

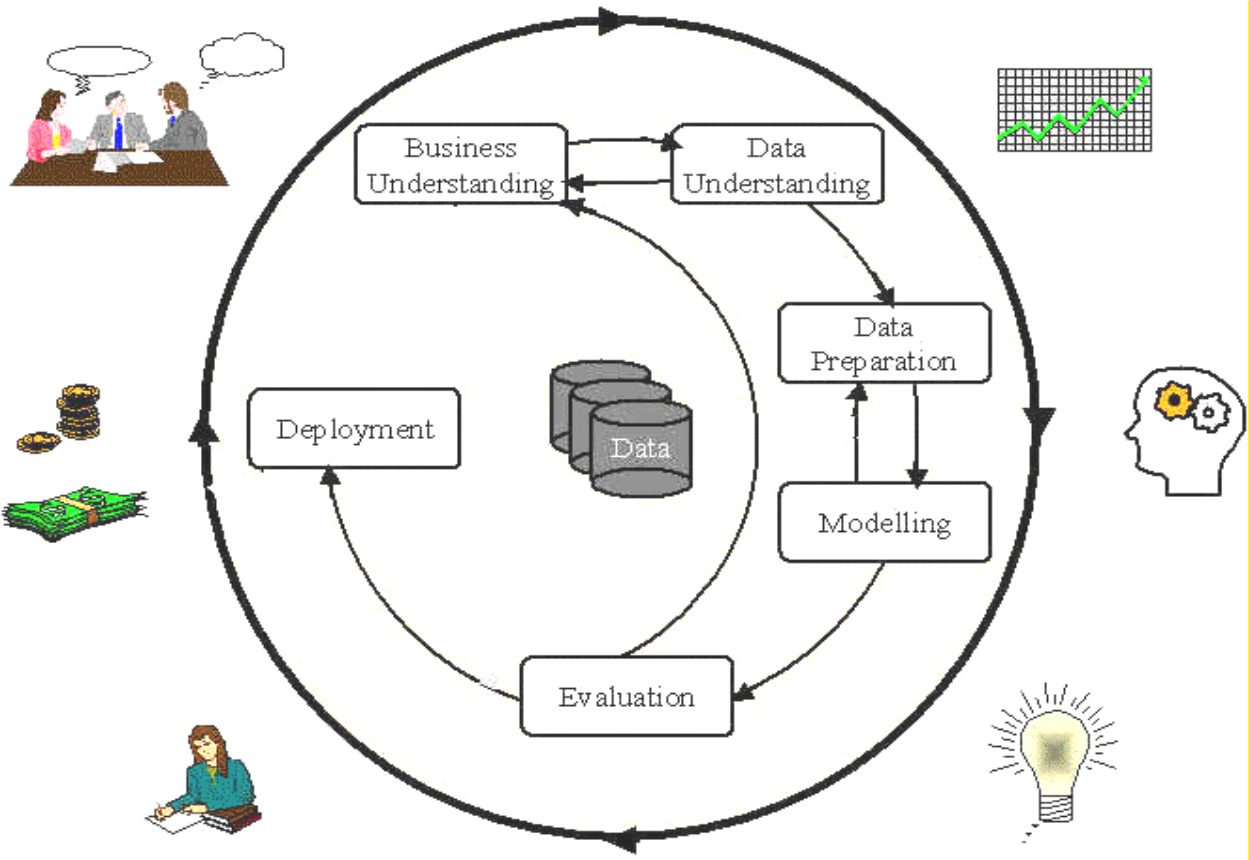
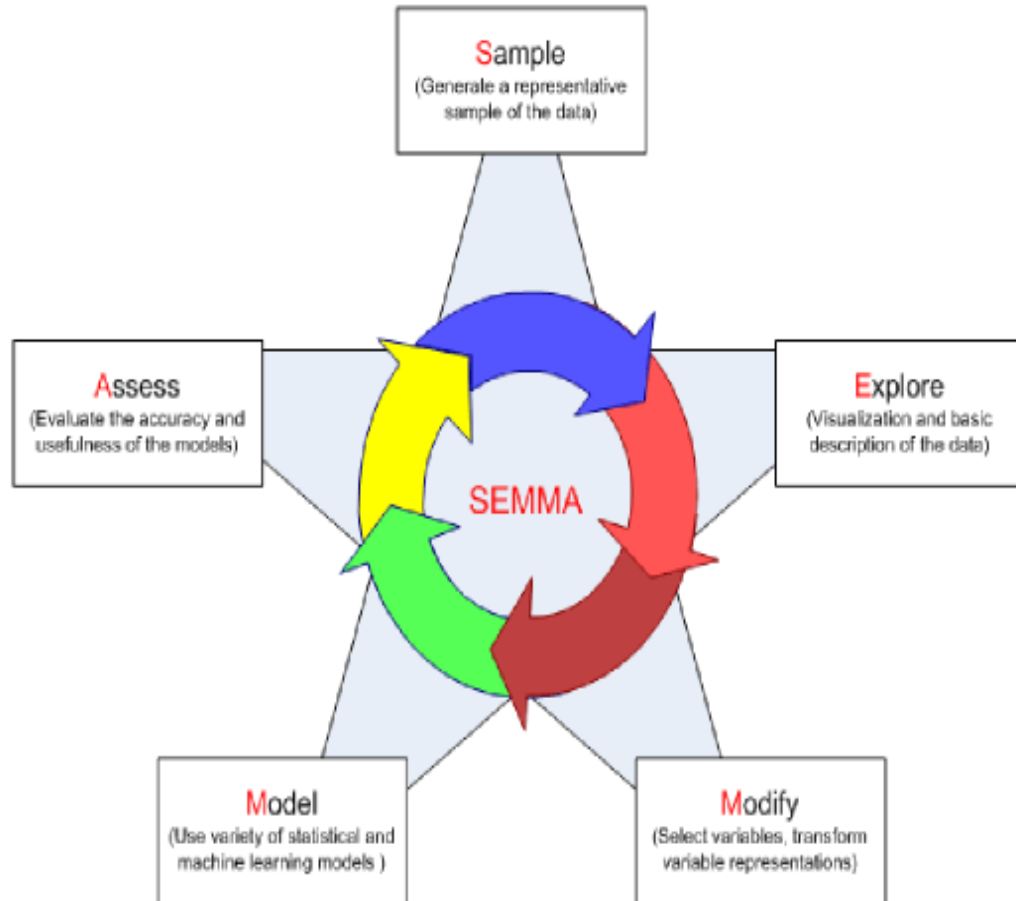


Figure 2. 2: CRISP-DM Process Model, Source: Shafique, &Qaiser, 2017

### 2.8.3. The SEMMA process model

The SEMMA stand for (Sample, Explore, Modify, Model, and Access) is data mining method developed by SAS institute. It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them. It has a cycle of five stages or steps.

- **Sample:** this is the first and optional stage of SEMMA process which focuses on sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.
- **Explore:** this is the second stage of SEMMA process which focuses on exploration of data. This can helps in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.
- **Modify:** this is the third stage of SEMMA process which focuses on modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also looks for outliers and reducing the number of variables.
- **Model:** this is the fourth stage of SEMMA process which focuses on modeling of data. The software for this automatically searches for combination of data. There are different modeling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.
- **Access:** this is the fifth and final stage for SEMMA process focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.



*Figure 2. 3: SEMMA model process model*

## 2.9. Application of Data Mining In Insurance Sector

Data mining is becoming common in both the insurance sectors like private and public. Data of the customer are one of the most valuable assets of any firm. The traditional methods, which were used for handling huge amounts of data generated by insurance transactions, are too complex. For transferring huge amount of data for decision making, data mining makes the methodology. Insurance firms use the data mining methodologies to enhance research and increase sales among the customers. The data mining used for various tasks in the insurance sector as follows.

### 2.9.1. Acquiring new customers

Acquisition of new customer is most important scenario of any firm. Traditionally, the insurance companies used the services of brokers to acquire the customers, but today a lot of ways helps to

acquire the new customers [26] Insurance firm focused of both acquiring new customer & retaining existing ones. Cluster Analysis used in the private sector to identify target group of customers. It involves targeting the population who are most likely to become customers or most profitable to the company.

### **2.9.2. Customer level analysis**

Analysis of customer purchase patterns and behavior. Using associated discovery technique, most insurance firms accurately select which policies and services to offer which customers [26]. According [27], it used data mining technology for insurance settlement and analyzed the customer records and also developed function structure model for customer analysis using data mining method.

### **2.9.3. Customer Segmentation**

Segment based products for targeting the customers. Data mining can be used for customer segmentation, for promoting the cross-selling of services, and in increasing customer retention. Customers are assigned to lifestyle segment based on their purchase history. Market segmentation is the key issue for the development of loyal relationships among the customers [28]

### **2.9.4. Policy designing and policy selection**

The insurance firm made the investigation whether people tend to purchase policy for the reason and the policy designed. In that case, to compete successful in the market, insurance companies used data mining technologies.

### **2.9.5. Prediction**

Data mining used for variety of applications such as predicting and classifying customer's and clustering customer characteristics for achievement of profitability. From the customer point of view, predictive analytics provides some benefits such as simplified claim handling process, reduced policy premium for low risk customers, faster and automated claim settlement. [29]. Data mining tools predict behaviors and future trends, allowing businesses to make proactive,

knowledge-driven decisions. It performs inference on the current data (insurance dataset) order to make predictions [30]

### **2.9.6. Claims management**

It is one of the most function is the insurance; data mining handled the claims management function such as claim analysis and fraud analysis.

### **2.9.7. Developing new product lines**

To develop the new product/plan depends on the customer needs. Insurance firms utilized all of their available information to better develop new product and marketing campaigns [26].

### **2.9.8. Underwriting and Policy management**

Data mining can be used in this application to optimize the function of the insurance value chain (Premium Analysis and Loss analysis)

### **2.9.9. Risk management**

One of the main stages in the process of risk management is risk financing is, of course, insurance. Insurance industry is keen in identifying the risks pertaining to their business. Risk management contains the six phases are risk identification, risk analysis, risk prioritization and risk monitoring [31].

### **2.9.10. Reinsurance**

Reinsurance comes under in the fields of risk management. The reinsurer may be either a specialist reinsurance company, which only undertakes reinsurance business, or another insurance company. Data mining tools can develop predictive models to arrive at the reinsurance level for the book of business based on the historical claims data. These predictive models can be identified suitable policies for reinsurance based on the loss experience of similar policies in past.

### **2.9.11. Fraud detection**

Detecting fraud claims is important in the insurance firm. Data mining isolates the factors that lead to fraud waste and abuse. To identify which transactions are most likely to be fraudulent. This is called as Fraud anomaly detection. In medical insurance, various medical insurance agencies suffered due to fraud claim in the health insurance; here he developed a model with three steps for the health insurance fraud detection. And he discussed the characteristics of fraud detection are high claims payment data is incorrect, suspicious data analysis, problem of hospital or physician [32]. The types of fraud and how much of fraud activities in the insurance firm discussed and he developed a claim sorting algorithm for the claim processing systems [33].

### **2.9.12. Trend analysis**

Trend analysis often refers to the science of studying changes in social patterns, including fashion, technology, and consumer behavior. In insurance, to reveal difference between the typical customer this month and last [34]. Data mining used in the different service industry especially in insurance firms the most frequently used applications for customer segmentation, customer retention, risk assessment and fraud detection and Policy approval process.

Insurance firms have a lot of improved changes in Information technology. Insurance industry has historically been a growing industry. It plays an important role in insuring the economic well-being one country. Many insurers use data mining techniques to identify the new customers and other are applying this technique to reduce portfolio risk, and to identify policies that were based on fraudulent information.



## CHAPTER THREE

### RESEARCH METHODOLOGY

The purpose of this study is to classify medical insurance customers based on claim record patterns in the database of Awash Insurance S.C. As it is discussed under data Mining Process Modeling in detail, there are different data mining process models available among which Cross-Industry Standard Process for Data Mining (CRISP) is widely user process model (Olegas Niaksu, 2015). Taking into account the factors mentioned so far, under this study, the researcher also preferred Cross-Industry Standard Process for Data Mining (CRISP) and R programming for the experiment to be conducted.

This research is an experimental research conducted on medical insurance data collected from Awash Insurance Company. In this chapter the methods, techniques and tools used to conduct the research are discussed in detail based on the steps of CRISP-DM data mining process model selected to guide the entire process of this research. Methodology means the steps or procedures that the researcher follows to achieve the objectives stated. It is a road map that shows the direction how the research is going to be done to reach the end.

In this research, the researcher aimed to use CRISP-DM process model to achieve the stated objectives. This is because the model has been widely applied for data mining studies. Besides, it is flexible to account for differences i.e. for different business problems and different data. And also it is open-source and industry standard data mining processes model. Accordingly, this study has followed the following methodologies in order to develop classification model and that predict medical insurance customers based on claim experience on historical data.

The process model adopted to undertake this research is the CRISP-DM. The CRISP-DM model has six steps that are: Business understanding, data understanding, data preparation, modeling, evaluation and deployment .The research is designed based on steps of this process model.

This chapter also describes different evaluation and performance measurement techniques in order to select the best model by applying the selected classification algorithms. Finally the output of

the classification algorithm create predictive model. In order to meet the goal of the study appropriate data mining methodology has to be selected. Therefore, the selection criteria were based on the nature of the problem identified and on the extensive effort required to review related literatures from previous studies in the study area.

### **3.1. Research Purpose**

The goal of the study was to apply data mining technique for classification of medical insurance customers based on claim experience from claim records in the life and health insurance database of Awash Insurance Company. Based on the information gained regarding the business practice and the service given by the Company to its customers and information obtained from the database, the researcher designed methodology to assess the problems and to find solutions for the problems in the aforementioned insurance company. The aim of the study was to build a meaningful model that classify medical insurance customers depending on their claim experience and applying appropriate data mining techniques. The result of this study could help Awash Insurance Company Life and health insurance branch to make better decision and help to assess the profitability of their existing customers.

### **3.2. Research Design**

The technology used in the world today enables the opportunity to collect large amounts of data. The problem is not to find the data, but rather how to analyze the gathered data and extract useful knowledge from it. This process is commonly known as data mining. The need to understand large and complex datasets are typical for all fields of business, science, and engineering [37].

Datasets continue to grow in size and becoming more complex and the need for software tools with automatic and intelligent data analysis has grown [37]. Therefore, the interest in machine learning has increased in the last couple of years. The possibility to find patterns and interpret data without the involvement of humans is a very efficient and powerful technique. Companies can extract useful information from a large dataset. [38].

Data mining and machine learning are complex techniques which involve several steps in the process. For a company to be successful in this field a useful model to follow is required. Today there exists some modeling methods that are more commonly employed than others. CRISP-DM [39] is a framework that consists of six different phases. It is an iterative process that starts with getting a business understanding of the problem. After that, an understanding of the data is established and the data is processed for the modeling step where the actual machine learning algorithm is applied. The result produced by the model is then evaluated. If the result is of good quality the algorithm is being deployed.

### **3.3. Cross-Industry Standard Process for Data Mining processing model**

Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR in 1999, CRISP-DM 1.0 version was published and is complete and documented. It provides a uniform framework and guidelines for data miners (Shafique & Qaiser, 2014). It consists of six phases or stages which are well structured and defined (Shearer, 2000). These phases are described below.

#### **3.3.1. Business Understanding**

It is an initial phase of CRISP-DM. In this phase, an understanding of the goal and the requirements of the project should be formed from a business perspective [40]. This understanding will then be transformed into a definition of data mining problems, to create a project plan for achieving the goals [41]. This initial phase focuses on understanding the study objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives [42].

During the business understanding phase a literature review is performed in order to gain insight with data mining studies which related to customer classification studies which have been solved by the application of data mining techniques and methods in previous researches. The study domain area is insurance organization. The researcher made an effort to study the objective of Awash Insurance Company and the situation of the business it is involved through document analysis and interviews. Awash insurance company is one of the first few pioneer private insurance

companies in Ethiopia launched following the liberalization of the financial sector in 1994. For the last 6 years it is the leading insurance company in the industry from private insurance companies except 2019/20 fiscal year. Currently it has 52 branches and 18 contact offices in the country.

From the interview and discussion made with Life and Health insurance manager as well as underwriting and claim officers, medical insurance class of business takes the lion share of the insurance business in the industry in general and Awash insurance company in particular. According to the life and health manager medical insurance business is not attractive as most of the claim payments go to medical insurance class of business from year to year which is justified by the claim report of the company. For example from July 1, 2019-June 30, 2020 the claim ratio of Awash Insurance Company is 65.2% for medical insurance (AIC CONSOLIDATED CLAIMS ANALYSIS, 2019/21). But no insurance company can avoid it since it is the major line of business through which they can get other class of business like marine, engineering, fire and burglary etc. which are attractive or major profit making class of businesses.

On the other hand from the discussion with branch managers, it is not possible to say medical insurance class of business is a loss making business; because branch managers confirmed that there a lot No Claim Discount(NCD) cases to be provided to their customers every year. That means some customer risk items are less claim cost prone and others are highly claim cost prone from medical class of business. Currently, there is no way to differentiate each customer risk items either as attractive or non-attractive groups. It is too difficult to do this manually as the data is voluminous. Moreover, it is not easy to associate different parameters to a given risk item to say this customer risk item with these parameters is high or low risk ratio. In general, since there are huge amount of medical insurance individual customer risk items in day to day insurance business transaction in underwriting and claim process, it is very difficult to classify which customer medical insurance individual risk items are particularly causing higher amount to claim cost manually. Therefore, the claim report produced by claim department is as traditional as to say from the total claim paid; and the big ratio goes to medical class of business which is very generic.

Therefore, this study applies DM technique to classify loss incurring and profit making customers under MI class of business, more specifically to classify which individual customer brings more claim cost to the pool. Contributing a lot in claim prediction as it is attempted to analyze MI

individual customer risk item based on claim upcoming in their study. The company can maximize its profit by controlling the unacceptable claim cost.

To determine how Machine Learning model could be used to classify MI customers claim cost that they can incur in future during underwriting using predictive models if they are provided an insurance coverage. For this to happen DM application is used to uncover claim record patterns in the database of AIC. Using the best performing model, AIC can provide insurance coverage to its customers based on the amount of claim that they can bring to the company. This encourages those customers with less claim and discourage loss incurring in general. Moreover, the company can maximize its profit by controlling the unacceptable claim cost.

### **3.3.2. Understanding of the data**

After understanding the problem to be addressed, the next step was analyzing and understanding the available data. The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of the available data (Cios et al, 2007). The original attributes and their description was presented.

This phase starts with an initial data collection and will then proceed with the goal of understanding the data. To gain this understanding different activities will be performed, such as identity data quality problems, discover first insights into the data and detect interesting subsets [43].

Next to identifying the problem and building a plan for solving the problem, the researcher proceeded with the central item in data mining process which is data understanding. This includes listing out attributes with their respective values and evaluation of their importance for this research and careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed. Finally, the researcher verified the usefulness of the data with respect to the DM goals.

#### **3.2.2.1. Data Collection**

The source of data for this research has been collected from Awash Insurance Company life and health branch database. Under this study data collection and proceeds with activities in order to

get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

To achieve the objective of this study, the researcher attempted to understand the data reside in the Awash Insurance Company database that can help to classify the customers of AIC based on their claim records. Considering the problem described and the business task, data pertaining to policyholders from 2019 to 2021 the historical data is extracted from AIC data base of medical insurance into MS-excel. To achieve the data mining goal formed from the business problem, relevant of the information were selected.

Most of the crucial parameters are life assured life assured limit, dependent limit, total premium, premium proportion, claim amount and claim ratio. The record contains dependent variable representing either “company rate” for risks whose risk claim ratio is below the industry standard claim ratio which is below 66% or “Loading” for claims whose claim ratio is above industry standard. Claim ratio is a value obtained by dividing claim amount by premium (claim amount/premium collected) customers. Finally, six (7) attributes, five (6) independent and 1 dependent variable and along with 41,151 are selected for data preprocessing task.

This is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms. This research paper also considered logistic regression models which preferable model for Categorical variable types and Naïve Bayes which is effective and powerful algorithm for predictive modeling (Huang and Lei, and R programming will be employed as the building tool. During data collection from Awash Insurance database all the fields are not included. Only selected fields or attributes are selected.

## Underwriting raw data

Proposer Name	Policy Number	Begin Date	End date	Life Assured Limit Ty	Risk Name	Total Premium
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Male	5,050
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Female	20,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Male	10,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	10,100
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	10,000
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	830
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021		Adult Male	2,300
M/S. Agriteam Canada Consulting PLC for TASC Consultancy	AIC /LD/MD/0425/2017	17-02-2020	16-02-2021	Total Limit	Adult Male	2,479

Figure 3. 1: underwriting raw data

## Claim raw data

Policy Number	Claim Number	Life Assured Name(Claimant Name)	Claim Amount
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/000321/20-001	Roman G/Egziabher	1600
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/000321/20-002	Solomon Tilahun	1810
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/000781/20-001	Ekram Wollo	3147.3
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/000781/20-002	Rahwa Aregawi	3314.7
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/001170/20-001	Solomon Tilahun	7263.9
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/001453/20-001	Roman G/Egziabher	1691.1
AIC/LD/MD/ 0562/20	LCLN/LHB/GM/001453/20-002	Tedla Bekele	20000
AIC/LD/MD/0106/02	LCLN/LHB/GM/000604/20-001	Million Regassa	602
AIC/LD/MD/0106/02	LCLN/LHB/GM/000775/20-001	Beletu Gedu	1637.6
AIC/LD/MD/0112/03	LCLN/LHB/GM/000926/20-001	Retta Lema	644
AIC/LD/MD/0115/03	LCLN/LHB/GM/000870/20-001	Abebe Fujiye Mathewos Abebe	817
AIC/LD/MD/0118/05	LCLN/LHB/GM/000566/20-001	Emebet Seyoum	200
AIC/LD/MD/0118/05	LCLN/LHB/GM/000566/20-002	Elsabeth Demissie	2536
AIC/LD/MD/0118/05	LCLN/LHB/GM/000718/20-001	Sirkalem Yimer	525
AIC/LD/MD/0118/05	LCLN/LHB/GM/000718/20-002	Merertu Abera	623.8
AIC/LD/MD/0118/05	LCLN/LHB/GM/000791/20-001	Sintayehu H/Mariam	1435
AIC/LD/MD/0118/05	LCLN/LHB/GM/000792/20-001	Bilise Gameda	2266.13
AIC/LD/MD/0118/05	LCLN/LHB/GM/001074/20-001	Martha Endalew	900
AIC/LD/MD/0118/05	LCLN/LHB/GM/001338/20-001	Belay Lule	5000

Figure 3. 2: Claim raw data

### 3.2.2.2. Description of the Collected Data

Description of the data is very important in data mining process in order to clear understand the data. Without such an understanding, useful application cannot be developed.

*Table 3. 1: Description of collected data*

<b>Attributes</b>	<b>Description</b>	<b>Data type</b>
Policy Number	It represents the policy number since the format of policy number is not convenient for model building it is represented by policy code	Character
Premium Amount	The amount of premium received by Insurance company in exchange of insurance coverage promised	Numeric
Claim amount	The amount of claim paid for specific policy	Numeric
Life assured limit	It represents either an individual or total	Character
Claim Number	It represents the claim number if the claimant	Character
Rate to be applied(RTA)	This is the dependent variable that this experiment is to be conducted to know how to charge the premium for specific risk item (either using “Company rate” or “Loading”). It represents the company or the organization buy the medical insurance product	Numeric
Risk Name	It represents gender of the insured either adult male or adult female	Character
Begin date	It represents the first date the policy holder purchase the policy holder	Numeric
Claim Ratio	Value obtained by dividing claim amount by the total premium	Numeric



End date	Date at which the medical policy holders end	Numeric
Dependent Limit Type	It represents the dependent is individual or total limit type	Character

Note: “Premium Amount” described above is the actual premium collected from the policy holders the dependent variable termed as “Rate to be applied(RTA)” represents a proposed premium to be collected based on the suggested rate type during underwriting for the purpose of the this research.

### 3.2.3. Data Selection, Preparation and Preprocessing

#### 3.2.3.1. Data Selection

This phase uses to create a target dataset. The whole target dataset may not be taken for the DM task. Irrelevant or unnecessary data are eliminated from the DM database before starting the actual DM function.

Data with outliers or missing values have been removed from the dataset so that the output will give representative information for decision making. As the data in Awash Insurance Company database contain many missing values, inconsistent data, there are many records which have been corrected in consultation with underwriters and branch managers and few of them are removed from the record so as to make it more convenient for machine learning.

Finally 41,151 records have been taken as complete and consistent records for further experimentation. The datasets were checked for completeness and correctness of the required attributes using excel file filtering and sequential arranging programs before analysis and prediction for the quality assurance of the experiment under process.

#### 3.2.3.2. Data splitting into Training and Testing Data

The problem of appropriate data splitting can be handled as a statistical sampling problem. Therefore, various classical statistical sampling techniques can be employed to split the data [May et al., 2010; Lohr, 1999]. It is standard in ML to split data into training and test sets. The reason

for this is very straightforward: if you try and evaluate your system on data you have trained it on, you are doing something unrealistic. Training dataset is used to fit the machine learning model and test dataset is used to evaluate the fit machine learning model.

Supervised machine learning algorithms are tools capable of making predictions and classifications. However, it is important to ask yourself how accurate those predictions are. After all, it's possible that every prediction your classifier makes is actually wrong, we can leverage the fact that supervised machine learning algorithms, have a dataset of pre-labeled data points. In order to test the effectiveness of your algorithm, we'll split this data into training and test data. The training set is the data that the algorithm will learn from and testing set is for evaluating purpose. From the total data sets 32,920 for training set (to build the model) and 8,231 for training data set (for evaluation purpose)

### **3.2.3.3. Data Cleaning**

This phase is used for making sure that the data is free from different errors. Otherwise, different operations like removing or reducing noise by applying smoothing techniques, correcting missing values by replacing with the most commonly occurring value for that attribute (Witten, & Frank, 2005).

Data with incomplete/ missing, duplicated, inconsistent data, noise/outliers and irrelevant data have been removed from the dataset so that the output will give relevant information for decision making. As the data in Awash Insurance Company database contain 3,420 missing values, inconsistent data, there are 1080 records which have been corrected in consultation with senior underwriters and life and health branch manager and few of them are removed from the record so as to make it more convenient for machine learning.

Finally 41,151 records have been taken as complete and consistent data records for further experimentation. The datasets were checked for completeness and correctness of the required attributes using excel file filtering and sequential arranging programs before analysis and prediction for the quality assurance of the experiment under process.

The data set under experiment is to be classified into a category of acceptable claim ratio which is less than 66%, is applied standard claim ration that is fixed by company executive management with consent of board of directors. This company rate is similar throughout the company's underwriting units. This rate is usually known as company rate. For the purpose of this study it simply represented as "company rate". The other category is records with high claim ratio or more than 66% claim ratio. In the company the customers with high claim ratio is subjected to premium loading. For the purpose of this experiment, simply it is termed as "Loading".

#### **3.2.3.4. Data integration**

The data integration process was done before deriving the attributes. As described before the dataset which was discussed above were available in in life and health database. Data integration method for retrieving important fields from different files and tables was done in the effort to prepare the data ready for the DM techniques to be undertaken in this research.

The data collected from the two tables, i.e. from the underwriting and claim data tables are exported to excel format and the column arrangement is done in separate files. Then every claim incurred in the claim data table moved to the corresponding policy number in the underwriting data table. The fact behind is every policy number in the claim data table is available in underwriting data table but the underwriting table data must not available in the claim data table.

Table 3. 2: Description of underwriting and claim data after the data integration

Attributes	Description	Data type
Policy Number	It represents the policy number since the format of policy number is not convenient for model building it is represented by policy code which is basically two values; either motor commercial or motor private.	String
Premium Amount	The amount of premium received by Insurance company in exchange of insurance coverage promised	Numeric
Claim Amount	The amount of Claim paid for medical insurance policy	Numeric
Life assured limit type	It represents life assured limit type either an individual or total limit	Character
Rate To Be applied(RTA)	This is the dependent variable that this experiment is to be conducted to know how to charge the premium for specific risk item (either using “Company rate” or “Loading”.)	String
Dependent Limit Type	It represents the dependent is individual or total limit type	Numeric
Risk Name	It represents gender of the insured either adult male or adult female	Character

### 3.2.3.5. Attribute selection

Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. Most machine learning algorithms are designed to learn in which are the most appropriate attributes to use for making their decisions.

There are many reasons that make the number of attributes to have a significant decrease from the original collection. Some of the reasons are, it uses to speed up the learning process, it makes simple to understand the generated rules. Due to the above reasons selecting attribute for measuring the individual customers risk item is mandatory. The below selected attribute run on R are evaluated using Information Gain. Based on the gain ration value and discussion with help domain experts most important attributes, which helps to classify MI customers of AIC life

insurance based on their claim experience are selected final datasets which were prepared for modeling purpose. During underwriting there are many data to be captured. But all these data are not important from the purpose of this paper perspective. As it is stated earlier, the purpose of this paper is to classify medical insurance customers based on claim experience whether loss incurring or profit making customers. Most of the attributes captured during underwriting do not serve this purpose. For instance the following attributes are not included in the experiment. Proposer Name, Basic sum assured, Begin Date, Gender, Relationship, Basic Rate, End date, etc.

On the other hand the following attributes are considered important for the purpose of this paper. Policy Name, Life Assured limit type, Risk Name, Dependent Limit Type, Total Premium , Claim Amount, and Rate to applied (RTA) are included into the experiment. Taking into account the dataset under experiment, the names and values of the attributes have been changed into some generic symbols for the sake of simplicity to have a more accurate representation of the variables.

#### **3.2.3.6. Data formatting**

After the dataset are categorized based on the objective of the study and the suggestion of domain experts, the dataset is transformed into appropriate extension for both statistical analysis and data mining rule techniques to describe users navigational behaviors .

This study implemented R as a data mining tool. The available dataset should be prepared in a format and data type which is suitable for R programming software. At first the integrated dataset was in an excel file format. To feed the final dataset into the R programming software the file is changed into other file format. The excel file was first changed into a comma delimited (CSV) file format. After changing the dataset into a CSV format the next step was opening the file with the R programming software.

#### **3.2.4. Data Transformation**

The actual data in the dataset are not suitably for prediction techniques as it is. Therefore, it needs some transformation which makes the pattern of the data easy for the machine learning process in

order to make sensible prediction. The dataset in this experiment and analysis contains categorical values that are transformed to binary values or factors.

The values of an attributes are changed as follows.

*Table 3.3: Attribute Transformation*

Attributes	Transformed to
Policy Number	PN
Life Assured Limit Type	LALT
Risk Name	RN
Dependent Limit Type	DLT
Total Premium	TP
Claim Amount	CA
Rate To Be Applied	RTA

*Table 3. 4: Premium to be calculated based on claim ratio*

Rate To Be Applied(RTA)		
Claim Ratio	Changed to	Binary Value
>66%	Loading	1
<66%	Company rate	0

*Table 3. 5: Life Assured Limit Type transformation*

Life Assured Limit Type		
Life Assured Limit Type	Transformed to	Binary Value
Individual Limit	IL	0
Total Limit	TL	1

*Table 3. 6: Risk Name transformation*

Risk Name		
Risk Name	Transformed to	Binary Value
Adult Female	AF	0
Adult Male	AM	1

### 3.2.5. Data modeling

In this phase, a suitable selection of modeling techniques, algorithms, or combinations of them will be done. Then, optimal algorithm parameters' values are chosen (Deshpande and Thakare, 2010). Therefore, this the step at which the process of extracting knowledge from processed data is done.

In this phase, various modeling techniques are selected and applied, and their parameters are set to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary. Using the above data taken from AIC database, this study

use some machine learning algorithm to learn certain pattern from the data provided in training data to predict the likely category of each individual customer risk item along with all the selected testing parameters when the testing data is applied on. As SVM generally are capable of delivering higher performance for small and medium dataset (in our case 41,151 records) in terms of classification accuracy (Srivastava and Bhambhu, 2010), it is chosen for this experiment. Moreover, it is memory efficient. The output we expect from this experiment is two which is “Company rate” and “Loading” and therefore, as Logistic regression model is suitable to predict dichotomous outcomes (Jihye Jeon, 2015); it is also chosen as good algorithm for our dataset. On the other hand as Naives Bayes algorithm is relatively simple to understand, can be trained easily and faster to predict classes (Kaviani and Dhotre, 2017), it is employed in this study too. Moreover, Rstudio is employed as it is widely used among data miners for developing statistical software and data analysis (Tejashree and Sawant, 2016). Some of the fields in the dataset of this paper contains binary data and other are transformed to numeric data type for which Rstudio is very suitable as well. According to Tejashree and Sawant (2016), R’s popularity substantially increased in recent years as a result of its ability to provide quality with ease of statistical and graphical techniques.

### **3.2.6. Evaluation of the discovered knowledge**

Before the model can be deployed the conducted work needs to be evaluated to be sure that the result meets the business requirements. This will be done in this phase. The steps that have been executed to create the result will be reviewed and evaluated thoroughly and at the end of this phase a decision on the data mining result should have been reached.

At this stage, the models obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives. Evaluation is the key to making real progress in data mining. After building a model, we must evaluate its results and interpret their significance (Two Crows Corporation, 2005). This stage consists of the interpretation and evaluation of the mined patterns. This indicates interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing irrelevant patterns, and translating the useful ones into terms understandable by users. The evaluation process is also carried out to identify interesting



patterns representing knowledge based on some lift as an interestingness measures. Both objective and subjective measures have been applied during the association rule analysis. The subjective method requires additional expert knowledge or input, which is not fully available during this study; the researcher used lift as objective measure. Therefore Improvement and further analysis can be made in this area. At the fifth stage of CRISP-DM process model (or models) that appears to have high quality from a data analysis perspective will be built. According to Sastry and Babu (2013), before proceeding to final deployment of the model, it is important to thoroughly evaluate it and be certain that the model properly achieves the business objectives. Therefore, the predicted value will be evaluated against the actual values using Confusion Matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Santra and Christy, 2012).

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

*Table 3. 7: confusion matrix for a two class classifier*

		Predicted	
Actual	Negative	a	b
	Positive	c	d

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect of predictions that an instance negative, and
- d is the number of correct predictions that an instance is positive.

The other pointes related to confusion matrix are:

The accuracy (AC) is the proportion of the total number of predictions that were correct (Santra and Christy, 2012).

$$AC = \frac{a + d}{a + b + c + d} \quad (1)$$

True positive rate (TP) is the proportion of positive cases that were correctly identified (Santra and Christy, 2012).

$$TP = \frac{d}{c + d} \quad (2)$$

False positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive (Santra and Christy, 2012).

$$FP = \frac{b}{a + b} \quad (3)$$

True negative rate (TN) is defined as the proportion of negatives cases that were classified correctly (Santra and Christy, 2012).

$$TN = \frac{a}{a + b} \quad (4)$$

False negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative (Santra and Christy, 2012).

$$FN = \frac{c}{c + d} \quad (5)$$

Precision (P) is the proportion of the predicted positive cases that were correct (Santra and Christy, 2012)

$$FN = \frac{d}{b + d} \dots\dots\dots > \quad (6)$$

The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. (Santra and Christy, 2012).

$$Recall = \frac{a}{a + c} \dots\dots\dots > \quad (7)$$

F-Measure provides a way to combine both precision and recall into a single measure that captures both properties. (Santra and Christy, 2012).

$$F\text{-Measure} = (2 * Precision * Recall) / (Precision + Recall) \dots\dots\dots > \quad (8)$$

### 3.2.7. Use of the discovered knowledge

Finally, how to use the discovered knowledge is seen by developing a prototype to test each incoming claim before processing. The purpose of this study is to uncover claim record patterns in the database of Awash Insurance S.C. As it is discussed under data Mining Process Modeling in detail, there are different data mining process models available among which Cross-Industry Standard Process for Data Mining (CRISP) is widely user process model (Olegas Niaksu, 2015). Taking into account the factors mentioned so far, under this study, the researcher also preferred Cross-Industry Standard Process for Data Mining (CRISP) and R programming for the experiment to be conducted.

### 3.2.8. Deployment

#### 3.2.8.1 Prototype

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the user can use it. Depending on the

requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the user can use it. (Chapman et al, 2000)

## CHAPTER FOUR

### EXPERIMENT AND DISCUSSIONS

#### 4.1. Introduction

As discussed in chapter three, under this chapter different experiments conducted exhaustively. After final data preparation, the total number of record in the dataset is 41,151. Among the total dataset 80% of the total data will use to train the algorithm of the model and 20% of it has been used for testing the performance of the model selected for the experiment.

Hence, the experimentation conducted using machine learning models such as Logistic Regression, Support Vector Machine and Navies Bayes models as discussed in chapter three. Each model evaluated and the best performing model chosen as best machine learning model algorithm for the objective of the experiment under discussion.

R Studio is used for the experiment to be carried out. Under this section different activities have been conducted regarding to running and evaluating model building experiments, selecting the best and appropriate model, and providing explanations on the selected model are vital task of this chapter is a programming language used for data manipulations, statistical analysis, and data visualization (CRAN 2017). Just like any programming language, R includes conditional statements, recursive functions, and input/output commands. But unlike a typical programming language, R includes the following features that make the language especially suited for data science (CRAN 2017). R has characters like open source (One of the main reasons the adoption of R is spreading is its open source nature. R binary code is available for everyone to download, modify, and share back again), Plugin ready (There is a base version of R, containing a group of default packages that are delivered along with the standard version of the software. The functionalities available through the base version are mainly related to file system manipulation, statistical analysis, and data Visualization. And data visualization friendly (R complies with principles and techniques employable to effectively display the information and messages contained within a set of data)

## 4.2. Experiment

### 4.2.1 Support Vector Machine (SVM) modeling

Under this experiment prediction model was built using SVM algorithm. SVM which is generally capable of delivering higher performance for small and medium dataset in terms of classification accuracy (Srivastava and Bhambhu, 2010)

This research intended to assist underwriters during underwriting process to determine how to calculate the premium for customers requesting for medical insurance coverage. The values are binary which is either “Company rate” or “loading “represented by ‘0’ and ‘1’ respectively.

The data set under experiment is to be classified into a category of acceptable claim ratio which is less than 66%, is applied company rate claim ration that is fixed by company executive management with consent of board of directors. This Company rate is similar throughout the company’s branches or underwriting units. This rate is usually known as company rate. For the purpose of this study it simply represented as “company rate”. The other category is records with high claim ratio or more than 66% claim ratio. In the company risks with high claim ratio is subjected to premium loading. For the purpose of this experiment, simply it is termed as “Loading”.

➤ Installing necessary packages and loading the corresponding libraries

As it depicted in figure 4.1, R packages are a collection of R functions, complied code and sample data. They are stored under a directory called "library" in the R environment. By default, R installs a set of packages during installation and more packages are added later.

```

install.packages('naivebayes')
install.packages('dplyr')
install.packages('ggplot2')
install.packages('psych')
install.packages('lattice')
install.packages("caret")
install.packages('e1071')

library(e1071)
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
set.seed(1234)

```

Figure 4. 1: Snapshot for installing necessary packages and loading the corresponding libraries

➤ Importing dataset from csv format to RStudio:

```
>SVMmi<-read.csv (file.choose (), sep = ",")
```

	PN	LALT	RN	DLT	PA	CA	RTA
1	AIC /LD/MD/0425/2017	1	0	1	791	0	0
2	AIC /LD/MD/0425/2017	1	0	1	791	0	0
3	AIC /LD/MD/0425/2017	1	0	1	791	0	0
4	AIC /LD/MD/0425/2017	1	0	1	791	0	0
5	AIC /LD/MD/0425/2017	1	0	1	791	0	0
6	AIC /LD/MD/0425/2017	1	0	1	791	0	0
7	AIC /LD/MD/0425/2017	1	0	1	791	0	0
8	AIC /LD/MD/0425/2017	1	0	1	791	0	0
9	AIC /LD/MD/0425/2017	1	0	1	904	0	0
10	AIC /LD/MD/0425/2017	1	0	1	904	0	0

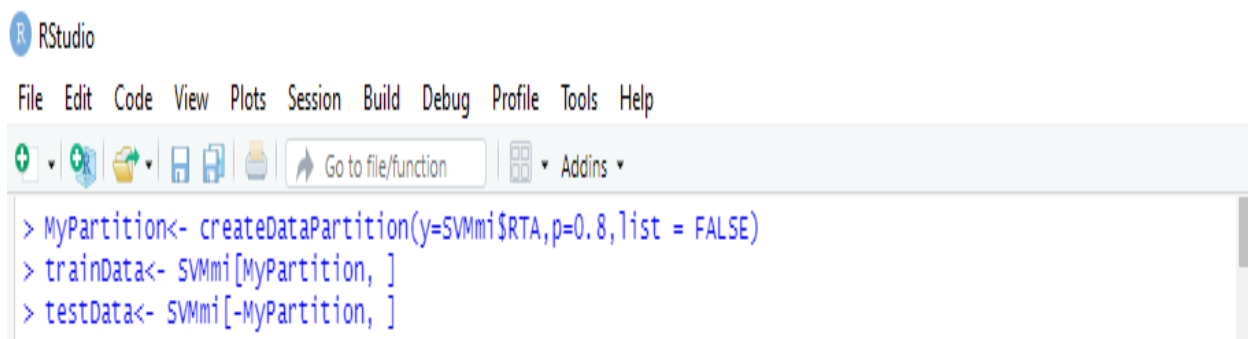
Showing 1 to 11 of 41,151 entries, 7 total columns

Figure 4. 2: Dataset snap shot for SVM model

As it is depicted in figure 4.2 above the dataset consists 41,151 records with 7 parameters. These parameters are Policy name(PN), Life Assured Limit Type(LALT),Risk name( RN),Dependent Limit type( DLT ) and integers for Premium amount( PA ) , Claim Amount(CA) and Rate to be applied(RTA).Therefore, in this experiment 41,151 records and 7 parameters are used.

➤ Dividing the dataset into training and test data

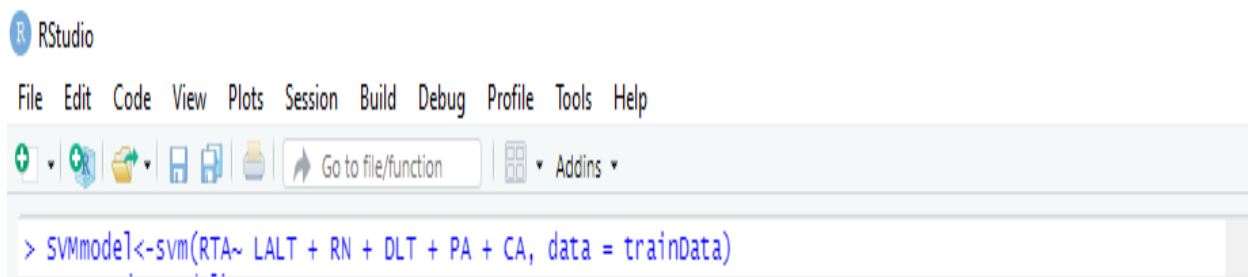
The dataset used in SVM model building was 41,151 medical customers 6 independent and 1 dependent attributes which have been saved as Comma Separated Version (CSV) format. In this experiment, dataset was divided into training data and testing data using 80:20 ratios. The training data was specifically used for the model building and the testing data was used for evaluation of the model. Partitioning dataset in to training data and testing data using “createDataPartition” function in the ration 0.8:0.2 is depicted as follows. Out of 41,151 we use 32,921 for test data and 8,230 records for training data.



```
> MyPartition<- createDataPartition(y=SVMmi$RTA,p=0.8,list = FALSE)
> trainData<- SVMmi[MyPartition, ]
> testData<- SVMmi[-MyPartition, ]
```

Figure 4. 3: partitioning snapshot to training and test data for SVM model

➤ Building model for SVM



```
> SVMmodel<-svm(RTA~ LALT + RN + DLT + PA + CA, data = trainData)
```

Figure 4. 4: Building a Model snapshot for SVM



As depicted above, SVM model building .Here, the dependent variable used is target that means the Rate to be applied (RTA), for the independent variable is Life assured limit (LALT),risk name(RN),Dependent limit type (DLT),Premium amount(PA) and Claim amount(CA).The models built using the training dataset.

➤ Prediction and model accuracy for SVM model

The first screenshot shows the RStudio interface with the following code and output in the console:

```
> table(PredictedValue=pred, AccualValue=testData$RTA)
      AccualValue
PredictedValue 0    1
              0 7289 32
              1   18 890
```

The second screenshot shows the RStudio interface with the following code and output in the console:

```
> (7289 + 890)/(7289 + 890 + 32 + 18)
[1] 0.9939239
```

*Figure 4. 5: Prediction and model accuracy snapshot for SVM model*

Figure 4.5 shows the cross table which is presented as predicted value and actual value. Based on the cross table above, the SVM model performed as follows. The model predicted customer records as customer of “Company rate” which is actually “Company rate” (TP), 305 predicted as “Loading” which is actually “Company rate ” (FN), 32 predicted as “Company rate” which is actually “Loading” (FP), 890 predicted as “Loading” which is actually “Loading”(TN).

Therefore the accuracy is calculated as  $(TP+TN)/ (TP+TN +FP+FN)$  which gives 99.39% of accurate.

➤ Confusion matrix result for SVM

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
> confusionMatrix(table(pred,testData$RTA))
Confusion Matrix and Statistics

pred   0    1
 0 7289   32
 1   18  890

      Accuracy : 0.9939
      95% CI   : (0.992, 0.9955)
  No Information Rate : 0.888
  P-value [Acc > NIR] : < 2e-16

      kappa : 0.9693

Mcnemar's Test P-value : 0.06599

      Sensitivity : 0.9975
      Specificity : 0.9653
  Pos Pred value : 0.9956
  Neg Pred value : 0.9802
    Prevalence   : 0.8880
  Detection Rate : 0.8858
  Detection Prevalence : 0.8897
  Balanced Accuracy : 0.9814

      'Positive' Class : 0
  
```

Figure 4. 6: Confusion matrix result snapshot for SVM model

As the figure 4.6 depicted, the accuracy of SVM model is 99.39%. It is calculated as  $(TP+TN)/(TP+TN+FP+FN) = (7289+890)/(7289+890+32+18)=0.9939=99.39\%$ . This implies error rate will be  $(FN+FP)/ Total value = (18+32)/ (8239) =0.61$ . The precision of the model is when it predicts yes, how often it is correct?  $(TP)/(TP+FP)$  which is  $(7289)/(7289+32) =0.9956=99.6\%$ . The True Positive Rate also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is  $(TP)/Total\ yes$ . This is  $(TP)/(TP+FN) = (7289)/(7289+833) =0.9975=99.75\%$ . False positive rate is when it is actually No, how often the model predicts as Yes. Therefore it is computed as  $(FP)/ (Total\ actual\ No) = (32)/ (32+890) = 0.03$ . True Negative Rate is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is  $(TN)/ (Total\ actual\ No) = (890)/ (32+890) =0.9653=96.53\%$

ROC graphs in machine learning was [Spackman \(1989\)](#), who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community, due in part to the realization that simple classification accuracy is often a poor metric for measuring performance ([Provost and Fawcett, 1997](#); [Provost et al., 1998](#)). In addition to being a generally useful performance graphing method, they have properties that make them especially useful for domains with skewed class distribution

and unequal classification error costs. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes. ROC graphs are conceptually simple, but there are some non-obvious complexities that arise when they are used in research.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
> RCORpre= prediction(pred,y_act)
> RCORperf=performance(RCORpre, "tpr","fpr")
> RCORperf=performance(RCORpre, "tpr","fpr")
> plot(RCORperf, colorize=TRUE,main="ROC curve", ylab="Sensitivity", xlab="1-Specificity")
> library(ROCR)
> auc<-performance(RCORpre,"auc")
> auc<-unlist(slot(auc,"y.values"))
> auc<-round(auc,4)
> legend(0.4,0.6, auc, title = "Area under ROC (AUC)", cex=1)
> |
```

As depicted in figure 4.7 above the ROC is plotted between True Positive Rate (Sensitivity) or (Y axis) and False Positive Rate (1-Specificity) or (X Axis). In the above plot, the area under curve covers the maximum area (ROC curve 99.56%)

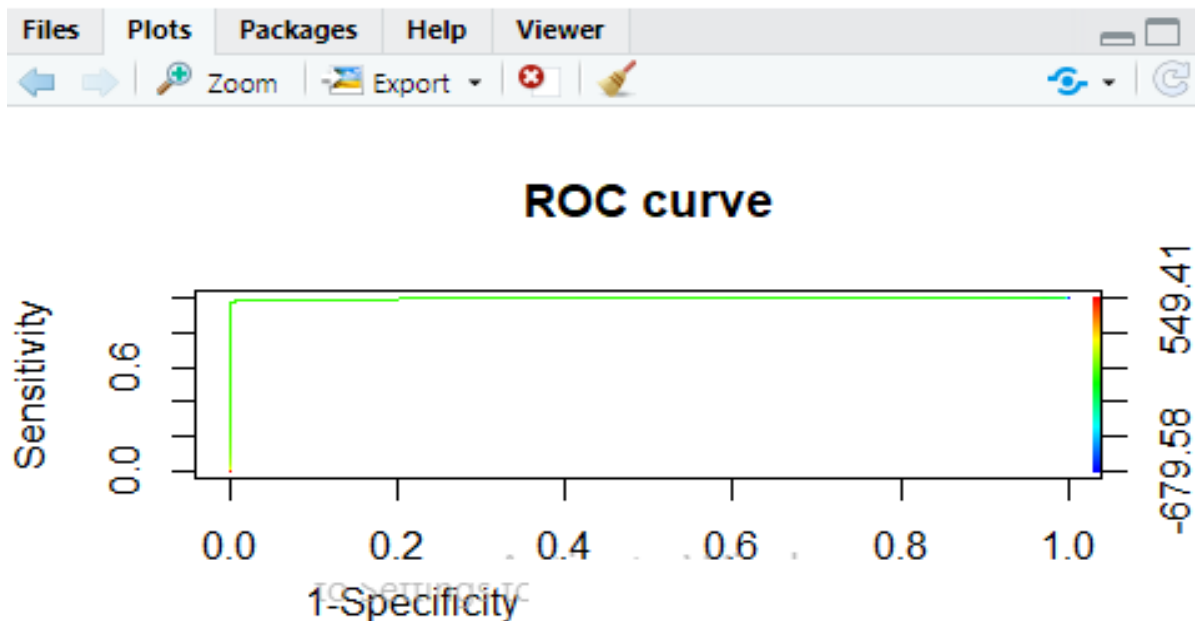


Figure 4. 7: ROC Curve result snap shot

ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area under curve (AUC). The area under the curve (AUC) also referred to as the performance of the ROC curve.

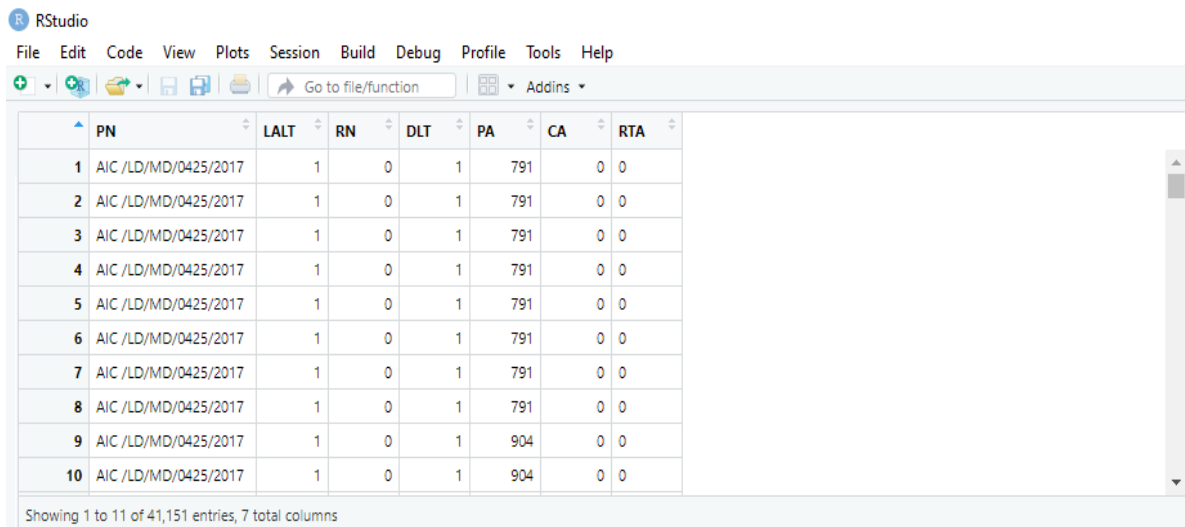
### 4.3. Naïve Bayes Model

Naive Bayes algorithm has an ability of handle noisy data, continuous and discrete data and make probabilistic prediction and the main reason behind its popularity is that it can be written into the code very easily delivering predictions model in very less time. Thus, it can be used in the real-time model predictions (Huang and Lei, 2011)

Under this experiment building prediction model is done using Naive Bayes Classification algorithm and R programming language.

- Importing dataset from csv format to RStudio:

```
>NBmi<-read.csv (file.choose (), sep = ",")
```



	PN	LALT	RN	DLT	PA	CA	RTA
1	AIC /LD/MD/0425/2017	1	0	1	791	0	0
2	AIC /LD/MD/0425/2017	1	0	1	791	0	0
3	AIC /LD/MD/0425/2017	1	0	1	791	0	0
4	AIC /LD/MD/0425/2017	1	0	1	791	0	0
5	AIC /LD/MD/0425/2017	1	0	1	791	0	0
6	AIC /LD/MD/0425/2017	1	0	1	791	0	0
7	AIC /LD/MD/0425/2017	1	0	1	791	0	0
8	AIC /LD/MD/0425/2017	1	0	1	791	0	0
9	AIC /LD/MD/0425/2017	1	0	1	904	0	0
10	AIC /LD/MD/0425/2017	1	0	1	904	0	0

Figure 4. 8: Dataset snap shot for Naïve Bayes model

- Dividing the dataset into training and test data

In this experiment, dataset was divided into training data and testing data using 80:20 ratios. The training data was specifically used for the model building and the testing data was used for evaluation of the model.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ + + + + Go to file/function Addins
> id<-sample(2, nrow(NBmi), replace=T, prob=c(0.8,0.2))
> train<-NBmi[id==1,]
> test<-NBmi[id==2,]
> |

```

Figure 4. 9: partitioning snapshot to training and test data for Naïve Bayes model

➤ Building model for Naïve Bayes model

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ + + + + Go to file/function Addins
> NBmodel<-naive_bayes(RTA~ LALT + RN + DLT + PA + CA , data = train)
> |

```

Figure 4. 10: Building a Model snapshot for Naïve Bayes model

As depicted above, Naïve Bayes model building .Here, the dependent variable used is target that means the Rate to be applied (RTA), for the independent variable is Life assured limit (LALT),risk name(RN),Dependent limit type (DLT),Premium amount(PA) and Claim amount(CA).The models built using the training dataset.

➤ Prediction for Naïve Bayes model

```

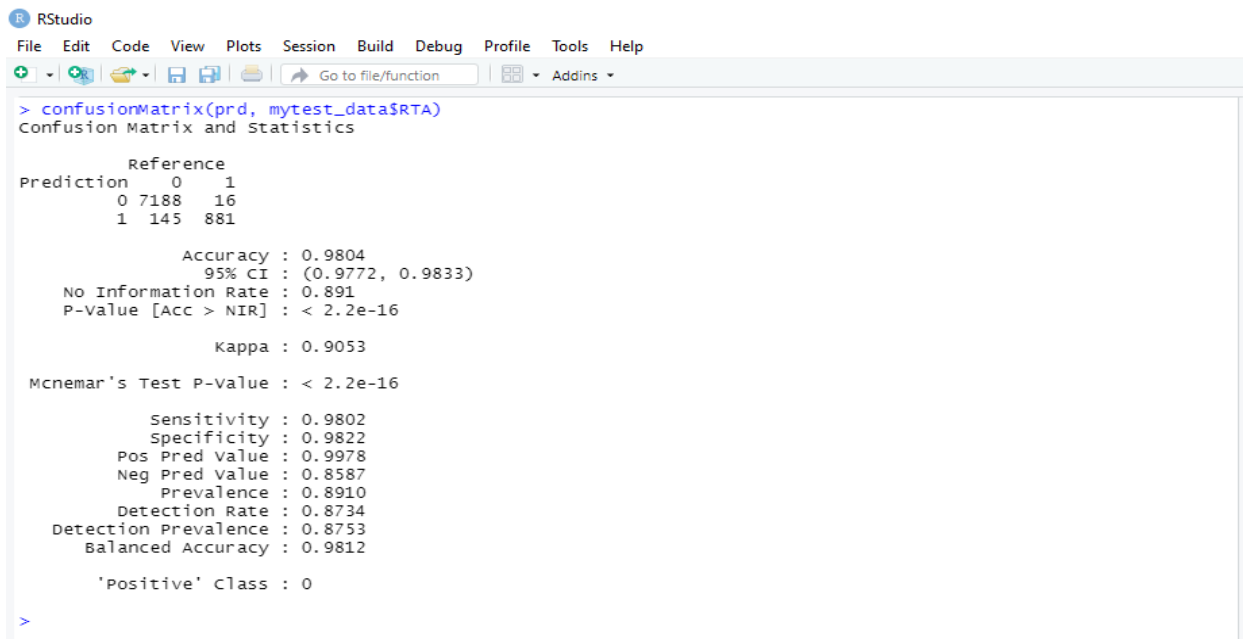
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> prd=predict(NBmodel,newdata = mytest_data)
> table(Predictedvalue=prd, Actualvalue=mytest_data$RTA)
      Actualvalue
Predictedvalue  0    1
              0 7369  22
              1   54 785
> |

```

Figure 4. 11: Prediction snapshot for Naïve Bayes model

Figure 4.11 depicted the cross table which is presented as predicted value and actual value. Based on the cross table above, the Naïve Bayes model performed as follows. The model predicted customer records as customer of “Company rate” which is actually “Company rate” (TP), 29518 predicted as “Loading” which is actually “Company rate ” (FN), 84 predicted as “Company rate” which is actually “Loading” (FP), 841 predicted as “Loading” which is actually “Loading”(TN). Therefore the accuracy is calculated as  $(TP+TN) / (TP+TN +FP+FN)$  which gives 92.01% of accurate.

➤ Confusion matrix result for Naïve Bayes



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
> confusionMatrix(prd, mytest_data$RTA)
Confusion Matrix and Statistics

      Reference
Prediction 0  1
0      7188  16
1      145  881

      Accuracy : 0.9804
      95% CI   : (0.9772, 0.9833)
      No Information Rate : 0.891
      P-value [Acc > NIR] : < 2.2e-16

      kappa   : 0.9053

      McNemar's Test P-value : < 2.2e-16

      Sensitivity : 0.9802
      Specificity : 0.9822
      Pos Pred Value : 0.9978
      Neg Pred Value : 0.8587
      Prevalence : 0.8910
      Detection Rate : 0.8734
      Detection Prevalence : 0.8753
      Balanced Accuracy : 0.9812

      'Positive' class : 0
  
```

Figure 4. 12: Confusion matrix result snapshot for Naïve Bayes model

As the figure 4.12 depicted, the accuracy of SVM model is 99.39%. It is calculated as  $(TP+TN) / (TP+TN+FP+FN) = (7188+881) / (7188+881+16+145) = 0.9804 = 98.04\%$ . This implies error rate will be  $(FN+FP) / \text{Total value} = (145+16) / (8230) = 0.0195$ . The precision of the model is when it predicts yes, how often it is correct?  $(TP) / (TP+FP)$  which is  $(7188) / (7188+16) = 0.9977 = 99.77\%$ .

The True Positive Rate also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is  $(TP) / \text{Total yes}$ . This is  $(TP) / (TP+FN) = (7188) / (7188+18) = 0.9802 = 98.02\%$ . False positive rate is when it is actually No, how often the

model predicts as Yes. Therefore it is computed as  $(FP) / (\text{Total actual No}) = (16) / (16+881) = 0.0178$ . True Negative Rate is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is  $(TN) / (\text{Total actual No}) = (881) / (16+881) = 0.9822 = 98.22\%$

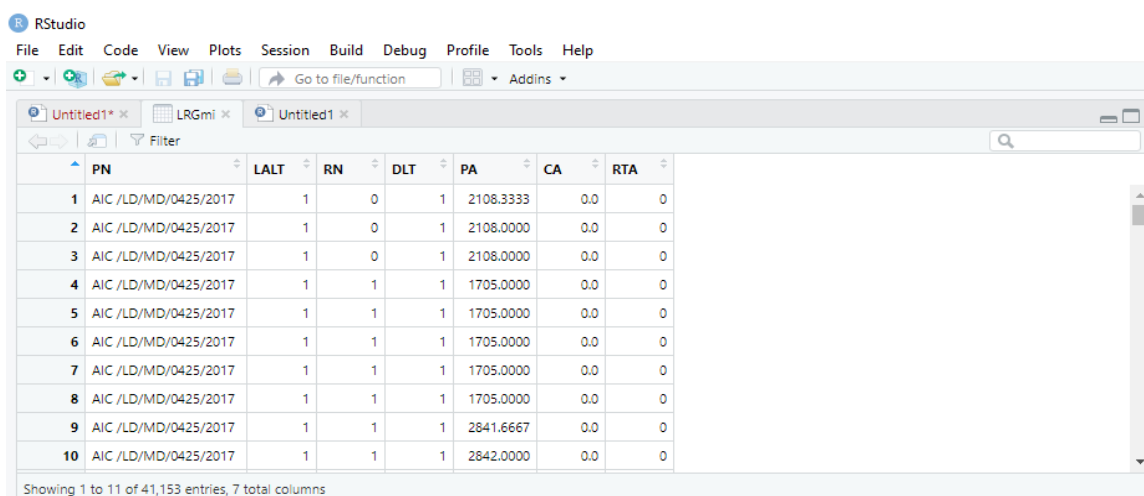
## 4.4. Logistic Regression Modeling

(Hlosta, Stríž, Kup, Zendulka, & Hruška, 2013) LR is a machine learning model for binary classification. The method can handle both numeric and categorical variables. Given a learned model, the value of the output variable is computed by applying the logistic function to linear combination of attribute values and weight vector. The logistic function converts the input value to interval  $[0, 1]$ . The result describes a confidence value for a given case being of the class 1. Typically, threshold  $t = 0.5$  is applied to determine whether an examined example belongs to class 0 or 1. Regression is the analysis, or measure, of the association between a dependent variable and one or more independent variables (B.K. & Srivatsa, 2011).

Under this experiment there are 6 independent variables and 1 dependent variable. The dependent variable Rate to be Applied (RTA) has a categorical values which are either ‘0’ or ‘1’ which represents “Company rate” and “Loading” respectively. And therefore logistic regression is selected for the prediction model under this experiment.

➤ Importing dataset to R Studio:

```
>LRGmi<-read.csv (file.choose (), sep = ",")
```



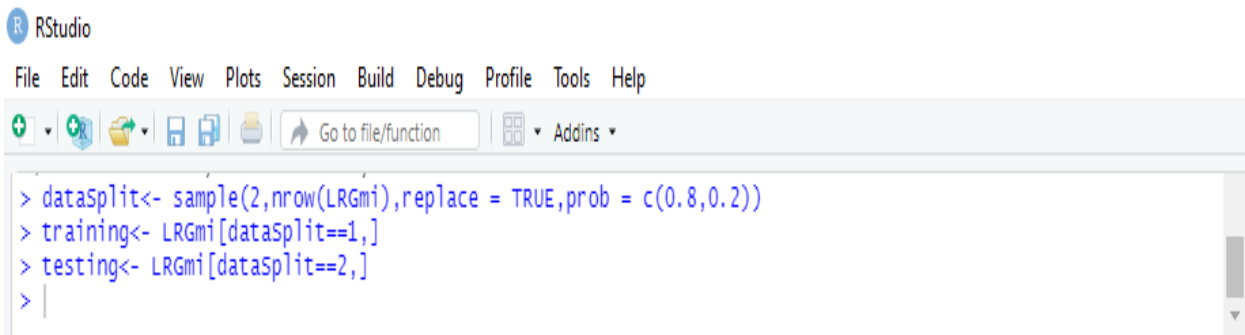
	PN	LALT	RN	DLT	PA	CA	RTA
1	AIC /LD/MD/0425/2017	1	0	1	2108.3333	0.0	0
2	AIC /LD/MD/0425/2017	1	0	1	2108.0000	0.0	0
3	AIC /LD/MD/0425/2017	1	0	1	2108.0000	0.0	0
4	AIC /LD/MD/0425/2017	1	1	1	1705.0000	0.0	0
5	AIC /LD/MD/0425/2017	1	1	1	1705.0000	0.0	0
6	AIC /LD/MD/0425/2017	1	1	1	1705.0000	0.0	0
7	AIC /LD/MD/0425/2017	1	1	1	1705.0000	0.0	0
8	AIC /LD/MD/0425/2017	1	1	1	1705.0000	0.0	0
9	AIC /LD/MD/0425/2017	1	1	1	2841.6667	0.0	0
10	AIC /LD/MD/0425/2017	1	1	1	2842.0000	0.0	0

Showing 1 to 11 of 41,153 entries, 7 total columns

Figure 4. 13: Dataset snap shot for Logistic Regression model

➤ Dividing the dataset into training and test data

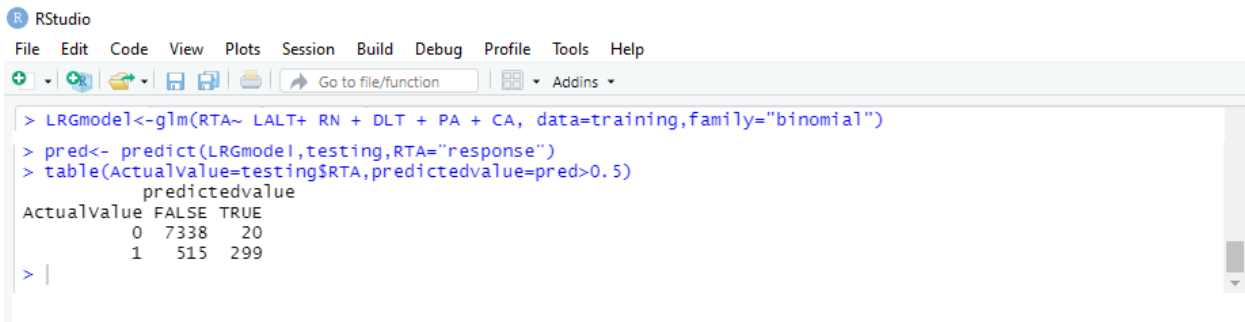
The dataset used in Naïve Bayes model building was 41,151 medical customers 6 independent and 1 dependent attributes which have been saved as Comma Separated Version (CSV) format. In this experiment, dataset was divided into training data and testing data using 80:20 ratios. The training data was specifically used for the model building and the testing data was used for evaluation of the model.



```
> datasplit<- sample(2,nrow(LRGmi),replace = TRUE,prob = c(0.8,0.2))
> training<- LRGmi[datasplit==1,]
> testing<- LRGmi[datasplit==2,]
> |
```

Figure 4. 14: partitioning snapshot to training and test data for LR model

➤ Building model for Logistic Regression model



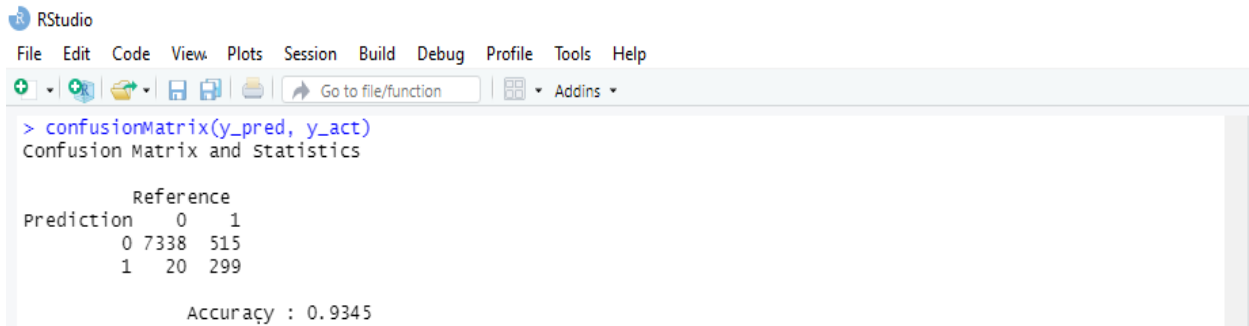
```
> LRGmodel<-glm(RTA~ LALT+ RN + DLT + PA + CA, data=training,family="binomial")
> pred<- predict(LRGmodel,testing,RTA="response")
> table(Actualvalue=testing$RTA,predictedvalue=pred>0.5)
      predictedvalue
Actualvalue FALSE TRUE
0           7338   20
1            515  299
> |
```

Figure 4. 15: Building a Model and prediction snapshot for LR model

As depicted above, Naïve Bayes model building .Here, the dependent variable used is target that means the Rate to be applied (RTA), for the independent variable is Life assured limit (LALT),risk name(RN),Dependent limit type (DLT),Premium amount(PA) and Claim amount(CA).The models built using the training dataset.



➤ Model accuracy for Logistic Regression model



```
> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

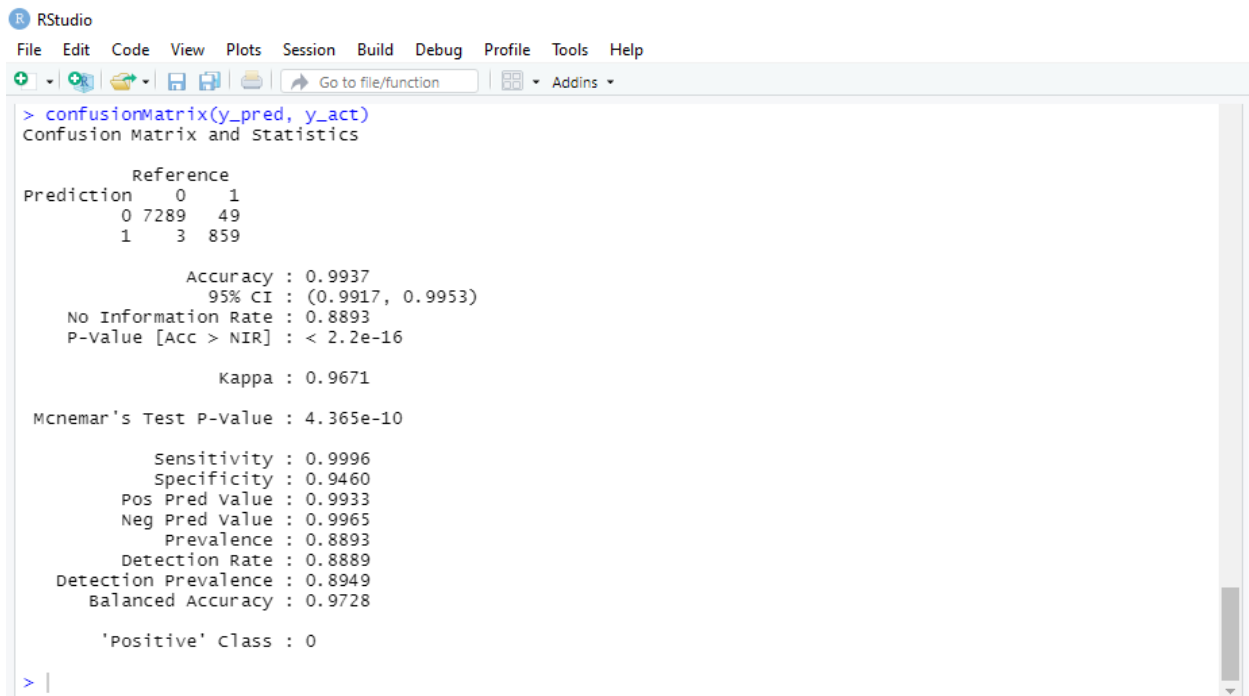
          Reference
Prediction 0      1
 0      7338  515
 1         20  299

      Accuracy : 0.9345
```

Figure 4. 16: model accuracy snapshot for Logistic Regression model

Figure 4.16 shows the accuracy is calculated as  $(TP+TN) / (TP+TN +FP+FN)$  which gives 99.37% of accurate.

➤ Confusion matrix result for Logistic Regression



```
> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

          Reference
Prediction 0      1
 0      7289  49
 1         3  859

      Accuracy : 0.9937
      95% CI   : (0.9917, 0.9953)
  No Information Rate : 0.8893
  P-value [Acc > NIR] : < 2.2e-16

      kappa   : 0.9671

  McNemar's Test P-value : 4.365e-10

      Sensitivity : 0.9996
      Specificity  : 0.9460
      Pos Pred Value : 0.9933
      Neg Pred Value : 0.9965
      Prevalence    : 0.8893
      Detection Rate : 0.8889
      Detection Prevalence : 0.8949
      Balanced Accuracy : 0.9728

      'Positive' Class : 0

> |
```

Figure 4. 17: Confusion matrix result snapshot for Logistic Regression model

As the figure 4.16 depicted, the accuracy of SVM model is 99.37%. It is calculated as  $(TP+TN)/(TP+TN+FP+FN) = (7289+859)/(7289+859+3+49) = 0.9937 = 99.37\%$ . This implies error rate will be  $(FN+FP)/\text{Total value} = (3+49)/(8200) = 0.0063$ . The precision of the model is when it predicts yes, how often it is correct?  $(TP)/(TP+FP)$  which is  $(7289)/(7289+49) = 0.9956 = 99.56\%$ . The True Positive Rate also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is  $(TP)/\text{Total yes}$ . This is  $(TP)/(TP+FN) = (7289)/(7289+833) = 0.9996 = 99.96\%$ . False positive rate is when it is actually No, how often the model predicts as Yes. Therefore it is computed as  $(FP)/(\text{Total actual No}) = (49)/(49+859) = 0.05$ . True Negative Rate is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is  $(TN)/(\text{Total actual No}) = (859)/(49+859) = 0.9460 = 94.60\%$

## 4.6. Summary of the findings

Currently, claim cost trend in Awash Insurance Company is inefficient controlling mechanism. According to the discussion made Life and Health Insurance manager, every branch reports the claim happened monthly and compiled at head quarter and then served to the management table for decision. They have mentioned that currently in Awash Insurance Company, most of the claims lodged in medical insurance are simply paid without having any classification of customers that are profit making or loss incurring. This implies there is no measure kept to check a medical insurance customers based on claim experience to classify whether the customers are profit making or loss incurring. Most of the time the decision will be premium rate revision. Here the problem is, first it difficult to address manually all the cases exhaustively, second the current way of analysis doesn't take into consideration the concept of parameters for every individual customer.

Considering the current problems observed, an experiment was extensively conducted on the dataset using Naïve Bayes and Logistic Regression prediction models. These predictive models are developed using selected attributed under medical insurance class of business after exhaustive discussion with underwriters as well as life and health manager.

The aim of this study is to identify under which premium calculation customer could be categorize under company rate or Loading which include charging additional premium and the experiments performed mainly to identify the best classifier for predicting the customer to the correct category.

Therefore, the predicted value will be evaluated against the actual values using confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Santra and Christy, 2012).

The percentage of correctly classified instances is referred as model accuracy and the accuracy is referred to as model performance (Ponnuraja, Lakshmanan and Srinivasan, 2016). Therefore, as it is observed from table 4.1 below, Support Vector Machine performed better when compared to Naïve Bayes and Logistic Regression both accuracy and precision.

*Table 4. 1: Comparison of prediction models*

<b>Classifier</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
Logistic Regression	99.37 %	99.33%	99.95%
Naïve Bayes	98.04 %	93%	98.06%
Support Vector Machine	99.39%	99.6%	99.97%

The study was answer the research questions: *How best can data mining techniques address existing challenges of insurance companies?*

The best performing prediction model selected depending on the accuracy level of the model and therefore, during medical insurance underwriting process the underwriter can get the claim information from the predicting model about the claim cost of the customer and categorize under appropriate premium calculation group either “company rate” or “loading “and serve the customer. Using the best performing model (in our case Support Vector Machine) the users can capture the values of this variables at underwriting stage and gets better understanding of the potential claim of the customer under process.

## **4.7. Deployment**

### **4.7.1. Prototype**

Finally, a predictive model created using classification algorithms. After the model is obtained, a prototype developed to be used before processing every incoming claim request.

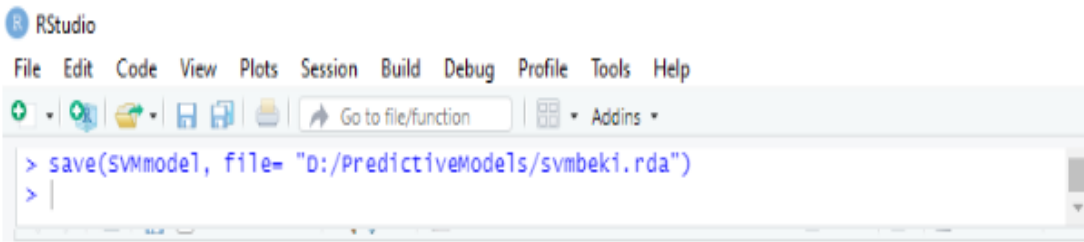
The deployment of this experiment will be at underwriting directorate of the company to be used by each and every branch during underwriting process. As the building of a model takes much time, the model should be built once and saved to some directory to be used anytime in demand.

The discovered knowledge is carried out with Awash Insurance Company user (in our case underwriters), not the customers who will carry out the deployment steps.

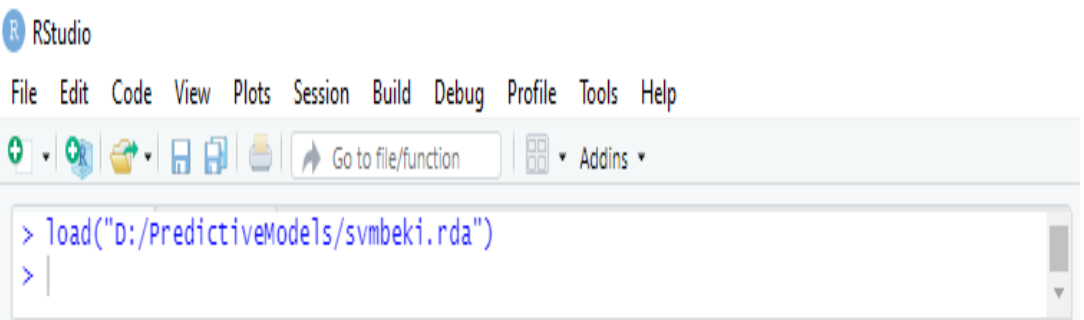
Underwriters are technical peoples and in order to understand knowledge gained, the data analyst will give a training for the underwriters to understand upfront what actions will need to be carried out from discovered knowledge in order to make underwriters to make use of the knowledge gained. When a customer comes to Awash insurance company, to be insured, the underwriters' captures required parameters and predict the claim upcoming.

*Table 4. 2: Required parameters for a medical insurance customer*

Parameters	Represented by	Values	Represented by
Policy Name	PN	AIC/BOL/MTPV/203020	AIC/BOL/MTPV/203020
Life Assured Limit Type	LALT	Total Limit	1
Risk Name	RN	Adult female	0
Dependent Limit Type	DLT	Total Limit	1
Premium Amount	PA	5000	5000
Claim Amount	CA	1254	1254



*Figure 4. 18: Saving a mode to specific directory*



*Figure 4. 19: Loading the saved model to R studio*

Note: the rate of medical insurance customer for premium calculation (RTA) has categorical values, either '0' or '1' representing "Company rate" and "Loading" respectively. The prediction result shows RTA value is 1 for record number 1. It means this medical insurance customer claim cost is high. Therefore, based on the prediction result, the user is advised to load the premium to give insurance cover policy. Moreover, the customer may bring to the insurance company more than one customer. For instance, the number of medical insurance customers are one, two, three and four. Under that situation, the predictive model can be used as follows.

```

> #when One customer comes to AIC for medical insurance
> PN<-c("AIC/BOL/MTPV/203020/20")
> LALT<-c(1)
> RN<-c(0)
> DLT<-c(1)
> PA<-c(2108)
> CA<-c(0)
> df<-data.frame(PN,LALT,RN,DLT,PA,CA)
> df
      PN LALT RN DLT PA CA
1 AIC/BOL/MTPV/203020/20  1 0  1 2108  0
> pred2<-predict(SVMmodel, newdata = df)
> pred2
1
0
Levels: 0 1
>
> #when Two customer comes to AIC for medical insurance
> PN<-c("AIC/BOL/MTPV/203020/20", "AIC/BOL/MTPV/203021/20")
> LALT<-c(1,0)
> RN<-c(0,1)
> DLT<-c(1,0)
> PA<-c(2108,2450)
> CA<-c(0,2350)
> df<-data.frame(PN,LALT,RN,DLT,PA,CA)
> df
      PN LALT RN DLT PA CA
1 AIC/BOL/MTPV/203020/20  1 0  1 2108  0
2 AIC/BOL/MTPV/203021/20  0 1  0 2450 2350
> pred2<-predict(SVMmodel, newdata = df)
> pred2
1 2
0 1
Levels: 0 1

```

Figure 4. 20: Prototype 1

Figure 4.20 depicted, if the number of medical insurance customer is one or two, the required parameters will be captured. The captured data should be framed in line with the built model format. After capturing and framing the required fields identified during experiment, the saved model should be loaded into R studio. Finally, the new data captured will be used to predict the upcoming claim cost as indicated on figure 4.20.

Note: the user captured identified attributes of the one medical insurance customers based the parameters captured, the model, predicted the premium to be collected from medical insurance premium from medical insurance customer should be calculated using company rate of the company. And the user captured identified attributes of the two medical insurance customers based

the parameters captured, the model, predicted the premium to be collected from medical insurance customer number 2 should be loaded and premium from medical insurance customer number 1 should be calculated using company rate of the company.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
> PN<-c("AIC/BOL/MTPV/203020/20", "AIC/BOL/MTPV/203021/20", "AIC/BOL/MTPV/203022/20")
> LALT<-c(1,0,1)
> RN<-c(0,1,1)
> DLT<-c(1,0,1)
> PA<-c(2108,2450,1500)
> CA<-c(0,2350,1200)
> df<-data.frame(PN,LALT,RN,DLT,PA,CA)
> df
      PN LALT RN DLT  PA  CA
1 AIC/BOL/MTPV/203020/20  1  0  1 2108  0
2 AIC/BOL/MTPV/203021/20  0  1  0 2450 2350
3 AIC/BOL/MTPV/203022/20  1  1  1 1500 1200
> pred2<-predict(SVMmodel, newdata = df)
> pred2
1 2 3
0 1 1
Levels: 0 1
>
> #when Four customer comes to AIC for medical Insurance
> PN<-c("AIC/BOL/MTPV/203020/20", "AIC/BOL/MTPV/203021/20", "AIC/BOL/MTPV/203022/20", "AIC/BOL/MTPV/203022/20")
> LALT<-c(1,0,1,0)
> RN<-c(0,1,1,0)
> DLT<-c(1,0,1,1)
> PA<-c(2108,2450,1500,20000)
> CA<-c(0,2350,1200,200)
> df<-data.frame(PN,LALT,RN,DLT,PA,CA)
> df
      PN LALT RN DLT  PA  CA
1 AIC/BOL/MTPV/203020/20  1  0  1 2108  0
2 AIC/BOL/MTPV/203021/20  0  1  0 2450 2350
3 AIC/BOL/MTPV/203022/20  1  1  1 1500 1200
4 AIC/BOL/MTPV/203022/20  0  0  1 20000 200
> pred2<-predict(SVMmodel, newdata = df)
> pred2
1 2 3 4
0 1 1 0
Levels: 0 1

```

Figure 4. 21: Prototype 2

Figure 4.21 depicted, if the number of medical insurance customers are three or four, the required parameters will be captured. The captured data should be framed in line with the built model format. After capturing and framing the required fields identified during experiment, the saved model should be loaded into R studio. Finally, the new data captured will be used to predict the upcoming claim cost as indicated on figure 4.21.

Note: the user captured identified attributes of the three medical insurance customers based the parameters captured, the model, predicted the premium to be collected from medical insurance customer number 2 and 3 should be loaded and premium from medical insurance customers number 1 should be calculated using company rate of the company. And the user captured identified attributes of the four medical insurance customers based the parameters captured, the model, predicted the premium to be collected from medical insurance customer number 2 and 3 should be loaded and premium from medical insurance customers number 1 and 4 should be calculated using company rate of the company.



## CHAPTER FIVE

### CONCLUSIONS AND RECOMMENDATIONS

#### 5.1. Introduction

The growth of interest in data, information and knowledge management has been helping many organizations to digitize and manage their information resource for effective use in future to prediction about their business processes, product and claim experience of their customers.

Implementation of data mining technologies in this study help to discover pattern for customer's classification in order to enhance service delivery and to maximize profit in the company. Data mining application could be used to classify customers depending on the historical data. Then, the data mining technology could help to predict profit making and loss incurring customers other in taking measures to improve the service delivery and profit generating in the future.

In this experimental research an attempt is made to reveal the high potential of data mining applications for customer classification, referring to the optimal usage of data mining methods and techniques to thoroughly analyze the collected historical data and to classify medical insurance customers of AIC based on their claim experience.

This chapter summarizes and concludes findings of the work done based on the findings of the study. The conclusion and recommendations as well as future work will be discussed briefly as follows.

#### 5.2. Conclusions

The aim of this study was to determine how machine learning model could be used to assess and to classify medical insurance customers upcoming claim cost that they can incur in future during underwriting if they are provided an insurance coverage.

Awash Insurance Company already automated its operation which means all operation and non-operation activities are managed centrally at head quarter. All the transaction performed anywhere in the company are actually performed in the central database at head quarter. The dataset used for the purpose of this study is collected from the company's database or central database server. On underwriting module or proposal page, there are many fields which are not mandatory that a user easily skip without filling them during underwriting which create incomplete data record. As the result there were many records found which are incomplete of required parameter in this experiment. Based on the branch they are originated, extensive interaction has been made with the branch underwriters and manager to complete them. Those records whose policy periods are too long back and enough information couldn't be found are removed from the experiment.

Finally 41,151 complete record of dataset were brought to the experiment. Out of 41,151 dataset records 32,921 records for training dataset and 8,230 for test dataset. From the many parameter of underwriting and claim data record, this experiment employed 7 parameters. Among these 6 are independent parameters and 1 dependent parameter. These are policy name (PN) - the code that identifies policyholders, Life assured limit type (LALT)-identifies limit type of the insured either it is individual or total limit, Risk name (RN)-it represents the gender of the life assured either adult male or adult female, Dependent limit type(DLT)-represents the life assured dependents gender , Premium Amount (PA) - the premium amount collected from customer, Claim Amount (CA) - the claim amount paid for that customer ,Rate to be applied(RTA) – This is the dependent variable that this experiment is to be conducted to know how to charge the premium for specific customer either using “Company rate” or “Loading”. It represents the company or the organization buy the medical insurance product. The importance of these parameters is experimented and the output for all of them is very significant.

The experiment has been conducted on the dataset using Naïve Bayes, Support Vector Machine and Logistic Regression prediction models. All the prediction models were resulted in different accuracy levels. The experiment outcome for each model – SVM, Naïve Bayes and Logistic Regression is 99.39%, 98.04%, 99.37%, respectively. Based on the experiment outcome of each prediction model, Support Vector Machine performed better with prediction accuracy of 99.39% followed by Logistic Regression with prediction accuracy of 99.37% and Naïve Bayes with prediction accuracy of 98.04%. According to the author, SVM is generally are capable of

delivering higher performance for small and medium dataset (in our case 41,151 records) in terms of classification accuracy. SVMs can learn a larger set of patterns and be able to scale better (R Joseph, Hlomani and Letsholo, 2016). Moreover, SVM has the ability to update the training patterns dynamically whenever there is a new pattern during classification and able to model complex nonlinear decision boundaries and are less prone to over fitting than other methods.

Based on the experiments conducted and output found it can be concluded that the Machine learning techniques can be used in an insurance industry to classify medical insurance customers based on claim experience using predictive models by measuring their performance using confusion matrix. Therefore, other insurance companies can also use these predictive models to classify medical insurance customers based on claim cost (automatically during underwriting which increase their profitability as they have a chance to adjust their premium ahead of insuring the customer as a profit making or loss incurring).

The contribution of this study provides information to Awash Insurance Company to determine whether the medical insurance customers are profit making or loss incurring. Those attribute that belong to loss incurring customers take the lion share for the decrease in profitability of medical insurance. This gives the company a chance to filter loss incurring attributes which enables the company to minimize claim cost incurred by medical insurance. So, the company can maximize its profit by controlling the unacceptable claim cost.

From the summary analysis, we can understand that, previous research works gave more attention either to find efficient predictive models for insurance claim prediction or finding a way to classify customers based on claim experience that they brought to an insurance company. Contributing a lot in claim prediction as it is attempted to analyze, most of the studies did not address medical insurance individual customer risk item based on claim upcoming in their study.

The discovered knowledge enables to determine how machine learning model could be used to classify medical insurance customers claim cost that they can incur in future during underwriting if they are provided an insurance coverage. When every individual customers provided the medical insurance coverage the predictive model enables the insurance company whether the customers are profit making or loss incurring depending on their claim cost. This helps the company to

minimize cost before the claim occur and to predict the customer claim cost at early stage. If the customers has a high claim cost either the company increase the premium or reject that customer. And if the claim cost is low the company provided the coverage or discount accordingly. It is, therefore, with this understanding this study is conducted using data mining technology to predict features and report hidden knowledge discovered by data mining techniques in Awash Insurance Company.

Based on the experiments conducted and output found it can be concluded that the Machine learning techniques can be used in an insurance industry to categorize individual customer risk items based on the risk exposure using predictive models by measuring their performance. Therefore, other insurance companies can also use predictive models to classify medical insurance individual customers based on claim exposure of the risk items automatically during underwriting which increase their profitability as they have a chance to adjust their premium ahead of insuring the risk.

### **5.3. Recommendations and future works**

The researcher encouraged the company, to work on the application of data mining techniques for successful achievement of the company goal. Because the finding showed the importance of loss incurring or profit making customers based on their claim records, and how new knowledge can be generated from those data, to improve their service in a new and modern methods than the traditional one. These in turn helped them to identify their market.

The experiment conducted has shown an encouraging result, but it doesn't mean it can be put in place at user's station easily. There might be different factors which need further investigation that requires technically rich professionals. Therefore, other researchers and academicians who are interested in the same area can explore in detail and come up with better deployment method to take this predictive model into practice.

The researcher forwards the following points as a future work:

- The research aimed at computing application of data mining to classify customers based on claim experience of medical insurance policy holders; hence, the researcher considered only medical insurance policy holders for the data-sets construction and classification process. Therefore, further researches should be conducted in the area of other insurance products using appropriate value measurement model.
- The experiment conducted in this study may be used as a way for further investigation so that it can be implemented in insurance companies. In order to deploy this finding practically in insurance companies working environment, researchers interested in the same area needs to integrate the scripts used in this experiment to some different programming language.
- Researchers interested in the same area needs to investigate an optimal user interface (dashboard) so that single click event can trigger the integrated script to manipulate the database for prediction purpose due to end users are not interested into detail of the process to check the probability of claim for each customer, they need simplified user interface.
- The prediction process needs CPU time that might make the production environment busy as the number of users increasing from time to time. Therefore, it requires further investigation to use the database of the company's offline.

## References

- [1] Kahane, Yehuda, Nissan Levin, Ronen Meiri and Jacob Zahavi (2007). "Applying data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance." *Asia-Pacific Journal of Risk and Insurance*, Volume 2, Issue 1:33-50
- [2] Apet, e. a. (1998). Insurance Risk modeling Using Data mining Technology. Volume 6, Issue 4:121-127
- [3] Dockrill, M. et al. (2001). Underwriting Management. Study Course 815. London: CII Publishing Division.
- [4] Lijia Guo. "Applying Data mining Techniques in Property/Casualty Insurance", *Casualty Actuarial Society Forum Casualty Actuarial Society -Arlington, Virginia Winter 2003*.
- [5] T. L. Oshini Goonetille ke and H. A. Ca ldera , "Mining Life Insurance Data for Customer Attrition Analysis ", *Journal of Industrial and Intelligent Information Vol. 1, No. 1, March 2013*
- [6] Eniafe Festus Ayetiran, "A Data Mining-Based Response Model for Target Selection in Direct Marketing", *I.J.Information Technology and Computer Science*, 2012, 1, 9-18.
- [7] K. Umamaheswari and S. Janakiraman —Role of Data mining in Insurance Industry, *international journal of advanced computer technology*, Vol.3 Issue-6, June-2014
- [8] Wikipedia Online. [Http://en.wikipedia.org/wiki](http://en.wikipedia.org/wiki)
- [9] Deshpande S. P. and Thakare V. M. (2010). Data Mining System and Applications: A Review. *International Journal of Distributed and Parallel systems (IJDPS)*, 1(1): 32-44.
- [10]. Milovic B., Milovic M. (2012). Prediction and decision making in health care using data mining. *International journal of public health science*, 1(2): 69-78.

- [11] Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No.4 1-305
- [12] Han, J. & Kamber, M. (2006), *Data mining: Concepts and techniques*, 2nd edn, Morgan Kufman Publishers, San Francisco ISBN: 978-0-470-89045-5
- [13] International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 04, Apr 2019
- [14] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H.,... & Steinberg, D. (2008). Top 1 algorithms in data mining. *Knowledge and Information Systems*, vol. 14, No. 1, 1-37.
- [15] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, Vol. 29, No.2, 103-130.
- [16] Mohammed J. Zaki, "DATA MINING TECHNIQUES", Volume: 06 Issue: 07, August 2003
- [17] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H.,... & Steinberg, D. (2008). Top 1 algorithms in data mining. *Knowledge and Information Systems*, vol. 14, No. 1, 1-37.
- [18] Kanwal garg, Dharminder kumar and m.c.garg, "Data mining techniques for identifying the customer behavior of investment in life insurance sector in india", *International journal of information technology and knowledge management*, Vol.1, pp.51-56.
- [19] Katharina Morik and Hanna Kopcke. "Analyzing Customer Churn in Insurance Data". *Lecture Notes in Computer Science*, vol. 3202, 2004, pp. 325-336.
- [20] Nanthawadee Sucharittham, Thanaruk heeramunkong Choochart Haruechaiyasak, Bao Tu Ho, Dam Hieu Chi, "Data Mining for Life Insurance Knowledge Extraction: A Survey". *International journal of Computer Applications*, vol.51, No.3, pp.22-26. August 2012

- [21] S. Ba laji and Dr. K.Srinivasta, “Naïve Bayes Classification Approach for Mining Life Insurance Databases for Effective Prediction of Customer Preferences over Life Insurance Products”, International journal of Computer Applications, vol.51, No.3, pp.22-26.August 2012
- [22] H.Lookman Sithic, T.Balasubramanian, “Survey of Insurance Fraud Detection Using Data Mining Techniques”, International Journal of Innovative Technology and Exploring Engineering, Vol-2, Issue-3, February 2013.
- [23] Re khaBhowmik, “Detecting Auto Insurance Fraud by Data Mining Techniques”, Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, April 2011
- [24] Ramageri B. M. Data Mining Techniques and Applications. In Indian Journal of Computer Science and Engineering Vol. 1 No. 4 pp 301-305
- [25] Colin Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining”, JOURNAL of Data Warehousing, Volume 5, Number 4, page. 13-22, 2000.
- [26] Kanwal garg, Dharminder kumar and m.c.garg, “,Data mining techniques for identifying the customer behavior of investment in life insurance sector in india”, International journal of information technology and knowledge management, Vol.1, pp.51-56.
- [27] Xu Zhikun<sup>1</sup>, Wang Yanwen<sup>2</sup> and Liu Zhaohui, ”Optional Insurance Compensation Rate Selection and Evaluation in Financial Institutions “, International Journal of u- and e- Service, Science and Technology, Vol.7, No.1, pp.233-242, 2014
- [28] Miguéis V.L., Camanho A.S., João Falcão e Cunha (2012), “Customer data mining for lifestyle segmentation”, Expert Systems with Applications, Vol.39, pp. 9359-9366.
- [29] Ashok Goudar, “Predictive Analytics in insurance Servies”, MPHASIS, Hp company.



- [30] S. Balaji and Dr. K. Srinivasta, "Naïve Bayes Classification Approach for Mining Life Insurance Databases for Effective Prediction of Customer Preferences over Life Insurance Products", *International journal of Computer Applications*, vol.51, No.3, pp.22-26. August 2012
- [31] Dilbag singh, Pradeep Kumar, "Conceptual Mapping of Insurance Risk Management to Data Mining", *International Journal of Computer Applications*, Vol. 39, No.2, pp.13-18 2012.
- [32] Kuo-Chung Lin , Ching-Long Yeh," Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance", *International Journal of Engineering and Technology Innovation*, vol. 2, no. 2, pp. 126-137, 2012
- [33] Derrig R. A., "Insurance fraud," *Journal of Risk and Insurance*, vol. 69.3, pp. 271-287, 2002.
- [36] M. Venkatesh "A Study Of Trend Analysis In Insurance Sector In India ",*International Journal Of Engineering And Science (IJES)* ,Volume 2 ,Issue 6,pp 01-05, June 2013.
- [37] Lijia Guo. "Applying Data mining Techniques in Property/Casualty Insurance", *Casualty Actuarial Society Forum Casualty Actuarial Society -Arlington, Virginia Winter 2003*.
- [38] T. L. Oshini Goonetilleke and H. A. Caldera , "Mining Life Insurance Data for Customer Attrition Analysis ",*Journal of Industrial and Intelligent Information* Vol. 1, No. 1, March 2013
- [39] M. Kantardzic. *Data Mining: Concepts, Models, Methods and Algorithms*. John Wiley Sons, Inc, 2011. ISBN: 978-0-470-89045-5
- [40] P. C. et. al. *Crisp-dm 1.0 - step-by-step data mining guide*.<https://www.the-modeling-agency.com/crisp-dm.pdf>. Accessed: 09.04.2018.
- [41] KDnuggets. *What main methodology are you using for your analytics, data mining, or data science projects?* poll.<https://www.kdnuggets.com/polls/2014/analytics-data-miningdata-science-methodology.html>. Accessed: 02.05.2018.

[42] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinart, C. Shearer, and R. Wirth, CRISP-DM *Step-by-Step Data Mining Guide*, 2000. <http://www.crisp-dm.org>.

[43] Kamber et.al, 2006. “Data mining concepts and techniques”, 2nd edition. ISBN: 978-0-470-89045-5

## Annexes

### Annex 1

```
install.packages('lattice')
install.packages("ggplot2")
install.packages("caret")
install.packages('e1071')
```

```
library(lattice)
library(ggplot2)
install.packages('lattice')
install.packages("ggplot2")
install.packages("caret")
install.packages('e1071')
```

```
library(lattice)
library(ggplot2)
library(caret)
library(e1071)
LRGmi<-read.csv(file.choose(),sep = ",")
str(LRGmi)
#LRGmi$LALT<- as.numeric(LRGmi$LALT)
#LRGmi$RN<- as.numeric(LRGmi$RN)
#LRGmi$DLT<- as.numeric(LRGmi$DLT)
#LRGmi$PA<-as.numeric(LRGmi$PA)
#LRGmi$CA<- as.numeric(LRGmi$CA)
#LRGmi$CR<-as.numeric(LRGmi$CR)
LRGmi$RTA<- as.factor(LRGmi$RTA)
str(LRGmi)
```

```
dataSplit<- sample(2,nrow(LRGmi),replace = TRUE,prob = c(0.8,0.2))
training<- LRGmi[dataSplit==1,]
testing<- LRGmi[dataSplit==2,]

LRGmodel<-glm(RTA~ LALT+ RN + DLT + PA + CA, data=training,family="binomial")
pred<- predict(LRGmodel,testing,RTA="response")
table(ActualValue=testing$RTA,predictedvalue=pred>0.5)
```

## Annex 2

```
install.packages('naivebayes')
install.packages('dplyr')
install.packages('ggplot2')
install.packages('psych')
install.packages('lattice')
install.packages("caret")
install.packages('e1071')

library(e1071)
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
set.seed(1234)
head(empdata)
empdata<-read.csv(file.choose(),sep = ",")
empdata$RTA<-as.factor(empdata$RTA)
id<-sample(2, nrow(empdata), replace=T, prob=c(0.8,0.2))
train<-empdata[id==1,]
test<-empdata[id==2,]
model<-naive_bayes(RTA~ LALT + RN + DLT + PA + CA , data = train)
model
p<-predict(model,train, type = 'prob')
head(cbind(p,train))
pred<-predict(model,train)
(tab<-table(pred, train$RTA))
```