ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

PREDICTIVE MODEL FOR MEDICAL INSURANCE FRAUD DETECTION: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

By

Mirchaye Mulugeta

June, 2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

PREDICTIVE MODEL FOR MEDICAL INSURANCE FRAUD DETECTION: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science

By

Mirchaye Mulugeta

June, 2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

PREDICTIVE MODEL FOR MEDICAL INSURANCE FRAUD DETECTION: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

By

Mirchaye Mulugeta

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Dereje Teferi (PhD)	**Advisor**	----------	----------
Tibebe Beshah (PhD)	**Examiner**	----------	----------
Ermias	**Examiner**	----------	----------

Table of Contents

LIST OF FIGURES V

LIST OF TABLES V

DEDICATION VII

ACKNOWLEDGEMENT VIII

ABSTRACT IX

CHAPTER ONE ..... - 1 -

INTRODUCTION ..... - 1 -

1.1. BACKGROUND ..... - 1 -

1.1.1. History of Insurance in Ethiopia ..... - 3 -

1.1.2. Medical Insurance ..... - 5 -

1.2. Statements of the problem ..... - 6 -

1.3. Objective of the Study ..... - 7 -

1.3.1. General objective ..... - 7 -

1.3.2. Specific objective ..... - 7 -

1.4. Scope and Limitations of the study ..... - 8 -

1.5. Significance of the Study ..... - 8 -

1.6. Organization of the Thesis ..... - 9 -

CHAPTER TWO ..... - 10 -

2. DATA MINING AND KNOWLEDGE DISCOVERY PROCESS ..... - 10 -

2.1. Data Mining ..... - 10 -

2.2. The Data mining process ..... - 11 -

2.2.1. Business understanding ..... - 11 -

2.2.2. Data understanding ..... - 11 -

2.2.3. Data preparation ..... - 11 -
2.2.4. Modeling ..... - 12 -
2.2.5. Evaluation ..... - 12 -
2.2.6. Deployment ..... - 12 -
2.3. Data mining and statistics ..... - 13 -
2.4. Challenges of Data mining ..... - 13 -
2.5. Data mining functions ..... - 13 -
2.5.1. Predictive Modeling ..... - 13 -
2.5.1.1. Classification ..... - 14 -
2.5.1.2. Regression ..... - 15 -
2.5.2. Descriptive Modeling ..... - 16 -
2.5.2.1. Clustering ..... - 16 -
2.5.2.2. Association Rule Discovery ..... - 17 -
2.6. Types of Data Mining Systems ..... - 18 -
2.7. The Data Mining Process Models ..... - 19 -
2.7.1. The KDD process Model ..... - 21 -
2.7.2. CRISP-DM process Model ..... - 22 -
2.7.3. The six-step KDP Model ..... - 24 -
2.8. Data Mining Tools ..... - 26 -
2.9. Application areas of Data Mining ..... - 27 -
2.9.1. Application of data mining in Insurance ..... - 27 -
2.9.2. Application of Data mining for fraud detection ..... - 28 -
2.9.3. Local Related works ..... - 31 -
CHAPTER THREE ..... - 33 -

3.1. Research Design ..... - 33 -
3.1.1. Understanding of the problem ..... - 33 -
3.1.2. Understanding of the data ..... - 34 -
3.1.3. Preparation of the data ..... - 34 -
3.1.4. Data mining ..... - 34 -
3.1.5. Evaluation of the discovered knowledge ..... - 35 -
3.1.6. Use of the discovered knowledge ..... - 35 -
3.2. Data mining methods for fraud detection ..... - 35 -
3.2.1. Naïve Bayes Classification Technique ..... - 36 -
Naïve Bayes algorithm ..... - 37 -
3.2.2. Decision tree classification technique ..... - 39 -
The j48 decision tree algorithm ..... - 41 -
3.3. Evaluation methods ..... - 42 -
CHAPTER FOUR ..... - 45 -
4. BUSINESS UNDERSTANDING, DATA UNDERSTANDING AND DATA PREPAPRATION ..... - 45 -
4.1. EIC Business Practice ..... - 45 -
4.1.1. Classification of Medical Insurance policies in EIC ..... - 45 -
4.1.2. Underwriting and Claims handling process of Medical Insurance in EIC ..... - 46 -
4.1.2.1. Policy Underwriting ..... - 46 -
4.1.2.2. Claims Processing ..... - 47 -
4.2. Understanding the data ..... - 48 -
4.2.1. Data collection ..... - 48 -
4.2.2. Description of the data ..... - 49 -
4.3. Preparation of Data ..... - 52 -

4.3.1. Data Cleaning..... - 52 -

4.3.2. Data Integration..... - 54 -

4.3.3. Data transformation..... - 55 -

4.3.4. Data Reduction..... - 55 -

CHAPTER FIVE..... - 57 -

5. EXPERIMENTAL RESULTS..... - 57 -

5.1.1. Naïve Bayes Classification Model Building..... - 58 -

5.1.2. J48 Decision Tree Model Building..... - 62 -

5.1.3 Comparison of Naïve Bayes and J48 Decision Tree Models..... - 68 -

5.2. Evaluation of Discovered Knowledge..... - 70 -

5.3. Medical Insurance Fraud Detection System Prototype..... - 74 -

5.4. Maintenance of the Model..... - 76 -

CHAPTER SIX..... - 77 -

6. CONCLUSIONS AND RECOMMENDATIONS..... - 77 -

6.1. Conclusion..... - 77 -

6.2. Recommendation..... - 79 -

References..... - 81 -

APPENDIX -1- Screen shot taken from weka J48 classification of the selected model..... - 88 -

APPENDIX -2- Sample taken from weka classification result of Naïve Bayes..... - 89 -

APPENDIX -3- Sample taken from weka classification result of J48 Decision Tree..... - 90 -

APPENDIX -4- Sample code for implementing rules..... - 91 -

APPENDIX -5- Interview Questions..... - 94 -

DECLARATION..... - 95 -

LIST OF FIGURES

Figure 1 steps of KDD process (Fayyad et al. 1996) ..... - 21 -

Figure 2 phases of the CRISP-DM reference model ..... - 23 -

Figure 3 the six-step KDP model ..... - 25 -

Fig 4: A confusion matrix for positive and negative entries ..... - 43 -

Fig 5.1 Initial user interface ..... - 43 -

Fig 5.2 Screen shot user interface after the validation is done ..... -76 -

LIST OF TABLES

Table 4.1.:- Distribution of initially collected data ..... - 49 -

Table 4.2.:- description of attributes from CLAIM_OBJECTS table ..... - 50 -

Table 4.3.:- description of attributes from CLAIM table ..... - 51 -

Table 4.4.:- description of attributes from INSUREDS/O_ACCINSURED table ..... - 51 -

Table 4.5.:- description of attributes from INSURED_OBJECT table ..... - 51 -

Table 4.6.:- description of attributes from LIFE_RISK_COVERED/GEN_RISK_COVERED table ..... - 51 -

Table 4.7.:- description of attributes from CLAIM_AVAIL_DOCS table ..... - 51 -

Table 5.1: confusion matrix for Experiment 1, Naïve Bayes Classifier with different test options ..... - 58 -

Table 5.2: Classification results for Experiment 1, Naïve Bayes Classifier with different test options ..... - 59 -

Table 5.3: confusion matrix for Experiment 2, Naïve Bayes Classifier with different test options ..... - 60 -

Table 5.4: classification results of Experiment 2, Naïve Bayes Classifier with different test options ..... - 60 -

Table 5.5: confusion matrix for Experiment 3, Naïve Bayes Classifier with different test options ..... - 61 -

Table 5.6: Classification results of Experiment 3, Naïve Bayes Classifier with different test options .... - 61 -

Table 5.7: summary of confusion matrix for experiment 4, j48 model with full dataset..... - 63 -

Table 5.8: summary of classification results for experiment 4, j48 model with full dataset..... - 63 -

Table 5.9: summary of confusion matrix for experiment 4, j48 model with full dataset and applying SMOTE..... - 64 -

Table 5.10: Classification Results for experiment 4, j48 model with full dataset and applying SMOTE. - 65 -

Table 5.11: summary of confusion matrix for experiment 5, j48 model with non-life dataset ..... - 65 -

Table 5.12: Classification results for experiment 5, j48 model with non-life dataset..... - 66 -

Table 5.13: summary of confusion matrix for experiment 5, j48 model with life dataset..... - 66 -

Table 5.14: Classification results for experiment 5, j48 model with life dataset ..... - 67 -

Table 5.15: comparison of J48 and Naïve Bayes models..... - 69 -

DEDICATION

This research work is dedicated to my father, Mulugeta Woldetensae, who has advised me to succeed in my study and carrier despite all the difficulties I faced.

ACKNOWLEDGEMENT

First of all I would like to thank the almighty God who gave me his blessings and put me in this position passing through a long journey of life.

My sincere gratitude goes to my advisor, Dr. Dereje Teferi, for his constructive comments in all of the research phases and critical readings of the full paper. I also like to thank my instructor, Dr. Million Meshesha, who supports me in constructing the research problem.

My heartfelt thanks extended to my husband, Tadiyos Tesfaye, for the support he made in all aspects throughout the course life. I do not have words to express my gratitude to my mother, Fisseha Woldeamanuel, and all my sisters and brothers who had taken care of my lovely boys. For sure without such support it is unthinkable to entertain my education and become at this level. I am also very grateful to my brother-in-laws who were always there with me when it matters most.

Furthermore, I would like to thank EIC staff who have dedicated their time and contribute much knowledge and experience for this research to be finalized within the time allotted.

Last but not least, I want to thank all my friends who gave me their support in all aspects.

ABSTRACT

Insurance fraud is an act that can be seen in different insurance types including medical insurance. Fraud in the case of medical insurance is done by misrepresenting facts to get unauthorized benefit from the expenses covered under medical insurance. Globally companies are spending high amount of claim costs due to insurance fraud. It is a concern for companies to have a system that could differentiate frauds from incoming claims. Data mining tools and techniques can be applied in different fields one of which is fraud detection.

This research is conducted for the purpose of testing the applicability of data mining techniques in detecting fraud suspected medical insurance claims in the case of Ethiopian Insurance Corporation. A six_step hybrid process model is used to guide the entire knowledge discovery process. J48 decision tree and Naïve Bayes classification algorithms are used to build predictive model.

Several experiments are conducted and the resulting models show that the J48 decision tree is found to work well in detecting fraud with 84.01% classification accuracy. A prototype is developed based on the rules extracted from the J48 decision tree model. Finally recommendations and future research directions are forwarded based on the results achieved.

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND

Today, it is difficult to find a field or industry that Information Technology (IT) has not affected greatly. IT is highly responsible for our civilization. Today IT has managed to cover many aspects of computer technologies which can be split to different professional fields. It is a very demanding field since it is always developing as a result of new technologies coming our way every day; IT professionals should re-educate themselves every time. It is very important for anyone in the field to stay up to date with the newly developing technologies related to their industry and if possible to the technology as a whole.

In today's modern technology era almost every business runs its day to day operations using technology which has completely transformed business practices. It is hard to survive these days without having advanced tools. Information Technology has become a complete backbone to almost any business and its ability to be competitive and efficient. The advancements in communication together with the evolution of the IT industry have made it easy to do business throughout the world in real time. Our lifestyles and business opportunities change with the improvement in Information Technology which reduce complications and brings possibilities.

As a result of the advancement in technology there is huge data stored in different business organizations. The size of this data is growing exponentially, but most of the data has once been stored and have never been further processed and used [53]. This data collected from different sources if processed properly can provide immense hidden knowledge. If this knowledge is accessed it can be a key to gaining a competitive advantage for an industry. This creates a need for tools to analyze and model the stored data. The value of data mining is to proactively seek out the trends within an industry and to provide understanding to organizations that maintain substantial amounts of information [53].

While data becomes large in size it will be very difficult to be traced and compiled by experts, here comes the need for data mining techniques to help managers in different levels in support of sound decision making [34]. Data mining (Knowledge discovery in databases (KDD)) aims at the discovery of useful information from large collection of data. As it is explained in [12] KDD refers to overall process of discovering useful knowledge from data, while data mining refers to application of algorithms for extracting patterns from data [53]. The core functionalities of data mining includes applying various methods and algorithms in order to preprocess, classify and cluster the data to discover useful patterns from stored data [53].

Data mining techniques can be applied to a wide variety of data repositories which include databases, data warehouses, spatial data, multimedia data, internet and web based data [4]. Recently, different data mining systems are available to different business data. There are various application areas of data mining such as Customer Relationship Management (CRM), frauds detection, product development, Risk Assessment etc. Specifically in fraud detection credit card fraud, telecommunication fraud, cheque forgery, e-commerce fraud, insurance fraud and tax fraud are major areas of research [35].

There are a lot of cases where managers need to make decisions without having evidence of the dependent variable due to lack of technology support. One of the issues is fraud in insurance claims. The main problem here is classifying claims in two categories fraudulent and valid [9]. The previous statistical approaches to classify frauds such as discriminant analysis, probit, logit, feed forward-back propagation neural networks etc. used the previous records of claims which investigators go through them and classify as fraudulent or non-fraudulent. Based on these information, a model is built a model to be used to classify future claims is done [9]. Since previous fraud data is not registered carefully in practical cases it is very difficult to find this type of information.

Insurance fraud is a usual phenomenon which occurs in different ways. It has probably existed ever since the insurance industry itself [1]. Compared to other types of frauds, it is difficult for the insurer to know whether the claim is fraud or valid unless the fraudulent behavior is recognized [54]. Common medical insurance frauds include billing for services which are not rendered, billing for a non-covered service as a covered service, misrepresenting date and location of service, false/ unnecessary issuance of prescription etc. [1].

1.1.1. History of Insurance in Ethiopia

Insurance is defined in different ways. The legal definition operating in Ethiopia is: Insurance policy is a special type of contract where a person called the insurer undertakes against payment of one or more premiums to pay to a person called a beneficiary, a sum of money where a specified risk materializes [49]. Hence it is appropriate when we want to be protected against a significant monetary loss. There are many types of insurance products that are needed by customers to stay safe and secured.

The first Insurance business was transacted by the bank of Abyssinia, as an agent to a foreign company which to cover Marine and Fire risks. Another Austrian agent was representing Baliose Fire Insurance Company. During the Italian occupation (1936-1941) Italian insurance companies began operating. In 1941 seven companies were opened. The first Motor Insurance policy was issued in 1947 by the South British Insurance company. In 1950 all classes of Insurance cover except life insurance were available. Until 1950 there was no Local Insurance company in Ethiopia, the first Ethiopian insurance company named Imperial Insurance Company was established in 1951 [41].

In 1960 the first insurance law was enacted: before that there was no legal insurance frame of reference. In 1970 insurance proclamation was issued to regulate the insurance market in Ethiopia. As a result of this proclamation foreign companies were not allowed to operate in the country. Minimum requirements were set in this proclamation like Capital, ownership, return

so that some companies were closed due to not meeting the minimum requirement set in this proclamation [41].

In 1975 there were 13 insurance companies operating in Ethiopia and they were nationalized as one and owned by Government which becomes the Ethiopian Insurance Corporation. Recently there are 17 insurance companies operating in Ethiopia in which 16 of them are privately owned share companies and only the EIC is government owned insurance company.

Ethiopian Insurance Corporation (EIC) was established in 1975 by taking the assets and liabilities of the nationalized insurance companies to engage in all classes of insurance businesses in order to provide insurance service to reach the broad masses of the people and to promote efficient utilization of both material and financial resources [41]. Moreover, it is empowered to manage, administer, supervise and direct all insurance business transactions at a national level and also negotiate, arrange, underwrite and correct reinsurance treaties and policies with foreign companies.

Currently the corporation is providing insurance cover for more than 75 types of insurance products which are categorized under life and non life insurance. The major processes of the corporation are currently automated and all the branches are connected through network to use the systems online. As EIC is having the biggest market share in the insurance industry it has many clients buying the different types of insurance covers. Hence the database which resides in ORACLE is very huge which can be further processed and used for different decision making purposes. However, there is no implementation of data mining activity conducted on the corporation's data in regards to medical claims so far.

1.1.2. Medical Insurance

In Ethiopian Insurance Corporation medical insurance cover is provided in both Life and Non-Life categories. It is provided with Personal accident and workmen's compensation insurance and in Life it is provided with the basic life cover [45][46].

Workmen's Compensation: 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 [46].

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 [46].

Life Medical Product (Individual): 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 [45].

Life Medical Product (Group): 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 [45].

1.2. Statements of the problem

In the Insurance Industry claim costs are needed to be minimized as much as possible. There are different causes which result in high claim costs. One of the causes is insurance fraud which makes most insurance companies to incur huge claim cost. According to the Global Health Care Anti-fraud Association public and private organizations around the world are losing $415 billion a year to health care fraud which is on average 7.29% of the annual healthcare expenditure [17].

A discussion was made with the medical insurance experts in EIC on the difficulties they are facing in differentiating valid and fraud suspicious claims. They have mentioned that currently in EIC, most of the claims lodged in medical insurance are simply paid without investigating whether they are real claims or frauds. Only some exaggerated claims are assessed by human investigators for fraud. This is mainly due to shortage of health professionals working at EIC. Furthermore there is no measure kept to check a claim for validity except the fulfillment of the necessary claim documents. But the availability of claim document cannot assure that a claim is valid or fraud. They have also faced difficulty in that a claim is assessed differently by different professionals. The assessments are different and the parameters that the different experts use to evaluate the case are different. Sometimes the fraud case investigation may be done in the middle of processing the claim while the insured is sure that he will be indemnified. This will create a conflict with the customer while wasting unnecessary time on handling a claim which at the end will be rejected.

It is known that data mining is playing a vital role in different business organizations to manage large datasets. This is due to speed, use of different models to do retrospective analysis which could help to come up with predictive patterns as a result of which organizations become efficient in their day to day activities. Moreover, data mining can identify interesting patterns which human experts are not aware of before [34].

There is an attempt to apply data mining techniques to predict fraudulent claims for vehicle insurance in Ethiopia [47]. As to the researcher knowledge there is no local research done in applying data mining to detect fraud in medical insurance. Hence it is the aim of this research to construct a predictive model that enables to identify fraudulent transactions in medical insurance claims.

To this end, this research attempts to explore and answer the following research questions:

- *Which data mining algorithms are suitable to generate predictive patterns for classifying medical insurance claims as frauds and valid?*
- *What is the pattern to classify a given claim as fraud or valid?*
- *How does the knowledge obtained be implemented to validate incoming claims?*

1.3. Objective of the Study

1.3.1. General objective

The general objective of this research is to construct a predictive model for the purpose of medical insurance fraud detection to help EIC minimize claim costs.

1.3.2. Specific objective

To achieve the general objective the following specific objectives are identified.

- To review literature related to data mining and analyze insurance business nature for medical insurance so as to understand insurance fraud detection;
- To collect and prepare relevant insurance data for the research;
- To explore , compare and select proper data mining algorithm for the specific problem;
- To build a model for detecting fraud and test its performance;
- To develop prototype to make the obtained knowledge useful ;
- To evaluate the performance of the model using test dataset.

1.4. Scope and Limitations of the study

This research attempts to apply data mining techniques in detecting insurance fraud in the case of medical insurance claims based on claim data from Ethiopian Insurance Corporation. Towards this end, a predictive model will be created using classification algorithms. After the model is obtained, a prototype will be developed to be used before processing every incoming claim request and check whether it is valid or fraudulent.

Medical Insurance data of fiscal years 2011 and 2012 is collected from 18 selected branches of the Ethiopian Insurance Corporation. The selection is done based on branch grades that are the main districts with high level (district A's), second level districts (district B's) and third levels (Branch I's).

Medical insurance claim frauds do not only arise from the insured side but also the healthcare provider. A limitation the researcher faced in conducting this research is unavailability of data regarding health care providers which makes the research scope limited to detecting frauds of medical claims arising from customer side only.

1.5. Significance of the Study

Having a system to detect frauds in insurance will help the insurance company to minimize unnecessarily paid insurance claim costs as well as minimize investigation cost. Moreover it improves claim handling efficiency of the claim officers as they are able to differentiate between fraudulent and valid claims. Hence officers can handle the valid ones in appropriate time. This will result in customer satisfaction.

The resulting models can also work with integration of the existing insurance system which the corporation is using. Other insurance companies can also adopt the models since most of them have the same business nature and functioning in similar environments.

Moreover this research can be taken as an input for future research works in general in data mining or specific to insurance fraud detection.

1.6. Organization of the Thesis

This paper is organized in six chapters. The first chapter discusses the background of the research, explanation about the problem area and formulation of research questions, objective of the research and scope of the research.

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 and specific to medical insurance fraud detection 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 Naive Bayes' and the J48 decision tree methods. Evaluation methods of the resulting models are also included.

In chapter four the business understanding, data understanding and data preprocessing phases are explained in detail.

Chapter five is where the main data mining task is explained. In this chapter experimentation and evaluation of the results is discussed in detail. Finally a prototype in which how the discovered knowledge would be implemented in claim handling process is demonstrated. In the last chapter, conclusion and recommendation is forwarded based on the achievements of the study.

CHAPTER TWO

2. DATA MINING AND KNOWLEDGE DISCOVERY PROCESS

2.1. Data Mining

Data mining is extracting knowledge from large amounts of data [26]. When we have large data set, it is no longer enough to get simple and straight forward statistics out of them. The data being big data and the changing business needs together changes the focus from simple retrieval and statistics into complex data mining [26]. Data mining is described as the union of historical and recent developments in statistics, Artificial Intelligence and Machine Learning [53].

As stated in [40] most organizations are rich in data but the ability to make effective fact based decisions is not dependent on the amount of data they have. Success is based on our ability to discover more meaningful and predictive insights from the data we capture. Data mining looks for hidden patterns in the stored data which can be used to predict future behavior. Data mining is essential for improving performance and creating competitive advantage for all types of organizations [40]. In general data mining can help us to rapidly discover new, useful and relevant insights from data, make better decision and act faster, monitor analyses and results to verify their continued relevance and accuracy and effectively manage a growing portfolio of predictive modeling assets [40].

In [27] data mining (KDD) is explained as the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple statistics. It uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. It is defined by different scholars in different terms but the key properties of data mining are automatic discovery of patterns, prediction of likely outcomes, creation of actionable information and focus on large data sets and databases.

Han and Kamber (2006) Describes data mining as one step in knowledge discovery process [20]. The existence of different types of data: spatial data, engineering design data, hypertext and multimedia data time-related stream data and the World Wide Web become the bases for the increase in the focus of data mining researches. Data mining can be applied to all those types of data repositories even though the challenges and techniques of mining differ in each type of databases.

2.2. The Data mining process

The data mining life cycle has six phases, but the sequence of the phases is not rigid. That means moving back and forth between the phases is required [10]. The six phases are Business understanding, Data understanding, Data preparation, Modeling, Evaluation and deployment.

2.2.1. Business understanding

The initial task in data mining process is understanding the business. Here the requirements and objectives of the data mining project will be identified. After the project objectives are identified from the business perspective the formulation of the data mining problem will be carried out in this step.

2.2.2. Data understanding

The data understanding phase involves data gathering and exploration. While we take a closer look at the data, we can determine how well it addresses the business problem and we can see the quality and usefulness of the data for the purpose of the data mining task. Here we may decide to get more data or remove some from the existing.

2.2.3. Data preparation

In this step, we have to go through the data preparation phase since the real world data is unclean and is impossible to be mined directly. Data preparation if done properly will bring a significant improvement in the information that can be obtained at the end of the data mining task. This phase covers all the tasks involved in creating the final dataset to be used in building the model. Data preparation tasks are likely to be performed multiple times, and not in any

prescribed order. Here attribute selection, data reduction, adding new attribute, data cleansing and transformation tasks are accomplished.

2.2.4. Modeling

Modeling is where the data mining algorithms and techniques are selected. In this step we may need to transform the data to a format that is acceptable by the selected data mining tool. Also it may be needed to reduce the data and divide the dataset to training and test dataset. The training set is used to train the data mining algorithms and the test set to test the accuracy of the patterns found. Then finally implement the selected algorithms and build the model.

In modeling it is important to remember that no one model or algorithm is categorized as best model. The choice of model is based on the problem and the nature of the data. That is why there are different models, tools and technologies available [50].

2.2.5. Evaluation

In the evaluation phase the resulting model is evaluated on how well it satisfies the originally stated business goal. All the patterns found by data mining algorithms may not be valid, there is a concept called overfitting: the algorithms might find patterns in the training set that are not present in the general dataset [40]. The testing should be done efficiently since once the validity of the data mining results are tested and applied to a real data it will not be tested every time. It is by the end of this phase that a decision on the use of the data mining model will be reached.

2.2.6. Deployment

Deployment is the use of data mining within the target environment [27]. This step uses visualization techniques to help users understand and interpret the results. The application of models to new data, extraction of model details or the integration of the data mining models within applications , data warehouses or reporting tools are included in this phase of data mining.

2.3. Data mining and statistics

Data mining and Statistics have overlapping concepts since most of the data mining techniques are also available in statistical frameworks. The traditional statistical methods require user interaction for validating the model. Due to this reason statistical methods are difficult to automate. In addition to this fact these methods do not scale well to very large data sets. The statistical methods are mostly used in testing hypothesis or finding correlations based on smaller or representative samples of a large population [27]. The data mining techniques on the other hand can be easily automated. Correctness of the resulting model depends on the amount of data used.

2.4. Challenges of Data mining

With all the advantages that we can get from data mining it has also got some challenges. The main challenge is data mining technique by its nature is dependent on the data accuracy: which is difficult to find in the real world. Another challenge in data mining is data sharing: organizations may not be willing to avail their data to conduct data mining. Building a data warehouse for mining is expensive so start up cost is another issue [22].

2.5. Data mining functions

Data mining functions represent a class of mining problems that can be solved using data mining algorithms [27]. For a specific problem the data mining algorithm and techniques to be used differ. So the type of mining function should be known to go through a data mining process. The functions of data mining can be categorized in to two based on the types of goals that the knowledge discovered is intended to be used [15]. These are verification: the system is used to verify hypothesis and Discovery: to find new patterns. The discovery function is further sub divided in to prediction and description. Since the concern of this study is the discovery part we shall briefly discuss discovery (prediction and description) data mining.

2.5.1. Predictive Modeling

Prediction is finding patterns for predicting future behavior of entities based on the existing data [15]. Predictive modeling falls in the category of supervised learning where a target

variable let Y be explained as a function of other variables X [15]. Classification, Regression, Time series analysis and prediction are some examples of predictive modeling [10]. Classification is a data mining function that assigns items in a collection to target categories or classes [10]. Regression and prediction differs on the type of prediction output where the regression predicts specific value and classification predicts class membership [53].

In time series analysis the values of an attribute is examined as it varies over time and the distance measures are used to determine the similarity between different time series. It is also defined as the analysis of a sequence of measurements made at specified time intervals. Time is the dominating dimension of the data [53].

Classification and regression are the two most common problems of data mining today [53].

2.5.1.1. Classification

Han and Kamber (2006) define classification as a process of finding a model that describes and distinguishes data classes or concepts, for the purpose of using the model to predict the classes of objects whose class label is known [20]. Due to this, classification is categorized as a supervised learning method. Classification algorithms work by first processing a training set which contains a set of attributes and their respective outcome which is included in the prediction attribute. The algorithm tries to discover relationships between the attributes that can make it possible to predict the outcome. Then it will be given a new data set or the testing set which the model has not seen before. The test set contains the same attributes with the training set but the prediction attribute is missing. The algorithm then analyses the inputs and predicts the values for the unknown attribute. The widely used classification algorithms include Decision Tree, Neural Network, Naïve Bayes, k-nearest neighbor, and support vector machine [27].

Decision Tree: is a classifier expressed as a recursive partition of the instance space by forming a rooted tree. This means it has root node which do not have any incoming edges but only

outgoing. All the other nodes have only one incoming edge. Internal nodes are nodes with outgoing edges and the leaves or terminal nodes are those without outgoing edges. The instance space is split according to a function of the input attribute values [24].

Artificial Neural Network: is an information processing paradigm that is works in the way biological nervous system, such as brain does. The structure of information processing system is composed of a large number of highly interconnected elements working together to solve specific problems. Neural Networks derive meaning from complicated data this can be used to extract patterns and detect trends that are complex to be noticed by humans. A trained neural network can be taken as expert in the area of information it has been given to analyze. Therefore it can project new situations based on the training data or initial experience [3].

Naïve Bayes: Naïve Bayes classification methods are supervised learning algorithms based on applying Bayes theorem with the assumption of independence variables [20]. This algorithm will be further explained in chapter three.

k-Nearest neighbor: is a simple classification algorithm which stores all variable cases and classifies new cases based on a similarity measure such as distance functions. In KNN a case is classified by a majority of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function [2].

Support vector machine: performs classification by constructing hyper-planes using support vectors. Hyper-planes are decision planes that separate between a set of objects having different class memberships in multidimensional space. It is used for both linear and non-linear data [20].

2.5.1.2. Regression

Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, temperature could be predicted using regression techniques [27]. Target values are known in the data set for regression task. Regression models are tested by computing various statics that

measures the difference between the predicted values and expected values [27]. It is of four different types: Linear regression, multivariate linear regression, non-linear regression and multivariate non-linear regression.

2.5.2. Descriptive Modeling

Description is finding patterns for presenting the data in an understandable form. Descriptive models are categorized as unsupervised learning techniques [15]. Examples of descriptive modeling include clustering, summarization and Association rule discovery [10].

2.5.2.1. Clustering

Big and high-dimensional data comes with different problems. Data points tend not to be where we think they are scattered very far apart, and can be quite far from the mean. To process the data we have to put together data points that are close and form blobs out of them. This process is known as clustering. Clustering is part of building a model.

Clustering is the process of grouping physical or abstract objects into classes of similar objects. A cluster is a subset of objects which are similar [30]. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it.

An important thing to notice about clustering is that there is no right answer. There are a variety of clusters of the same data. Each of these representations may be useful. There are different methods which are developed in order to come up with an optimal cluster for the intended purpose. In general clustering methods are classified as Partitioning and Hierarchical clustering.

Partitional clustering: Given a database of n objects, constructs k partitions of the data where each cluster optimizes a clustering criterion. The criteria could be minimization of sum of squared distance from the mean within each cluster [30]. These types of clustering algorithms

are complex since they exhaustively enumerate all possible groupings and try to find the global optimum. Even for small number of objects the number of partitions is huge. In general partitional clustering algorithms first compute the values of similarity or distance them they order the results and pick the one that optimizes the measures. Partitioning algorithms further classified as K-means, k-medoids, Bisecting k-means, Expectation Maximization and DBSCAN.

Hierarchical Clustering: these algorithms create a hierarchical decomposition of the objects. They are further classified as Agglomerative (bottom-up) and Divisive (top-down) clustering Techniques.

Agglomerative clustering: these algorithms start with each object being a separate cluster by itself and merge those which have minimal distance between them in to one cluster. The merging stops when all objects are in a single group or at any other point the user wants [30].

Divisive clustering: is a top down clustering strategy, which the entire data set is regarded as one cluster, and then clusters are recursively split until the desired structure of clusters is obtained. Distance between data points is used to check their dissimilarity in order to divide in to clusters [30].

The process of dividing or merging is shown using a dendrogram: a tree like structure used to depict the hierarchy of the clusters. Crossing along the tree with a horizontal line we can come up with different number of clusters. The main difficulty in using hierarchical model is: knowing where to split and how to split.

2.5.2.2. Association Rule Discovery

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rule [27]. In association rule pattern is discovered based on a relationship of a particular item

on other items in the same transaction [57]. Association rules are created by analyzing data for frequent if-then patterns and the most important relationships are identified based on support and confidence criteria. Support is how frequently items appear in a database and confidence is the number of times the if-then statements have been found to be true [27].

Association rule discovery is mostly used in sales transaction analysis. It is also called market basket analysis. It is applicable in areas including direct marketing, sales promotions, for business trends, store lay out, catalogue design, website personalization.

FP-Tree and Apriori are well known data mining algorithms which are used for association rule discovery.

2.6. Types of Data Mining Systems

Data mining researches generate different varieties of data mining systems due to the reason that data mining is the contribution of diversified disciplines. Therefore it is necessary to provide a clear classification of the data mining systems to help users identify those that best matches their problems easily. Data mining systems can be categorized based on the following classification criteria [10].

- Based on the kind of database mined

  Here the types of databases are further classified by data model and type of data so that each types use different data mining techniques. When data model is considered as criteria for classification: we can have relational, transactional, object-oriented, or data warehouse mining systems. And when the type of data is taken as criteria for classification the different data mining systems are spatial data, multimedia data, time-series data, text data and World Wide Web data mining systems.

- Based on the kind of knowledge mined

  Data mining systems can be categorized according to the kinds of knowledge they mine such as characterization, discrimination, association, classification, clustering, etc. An

advanced data mining system should facilitate discovery of knowledge at multiple levels of abstraction. That means generalized knowledge (high level), primitive-level knowledge (raw data level) or multiple level knowledge abstraction.

- Based on the kind of techniques utilized

  Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved example autonomous systems, interactive exploratory systems, query-driven systems, or based on the methods of data analysis employed example database oriented, data warehouse oriented, machine learning, statistics, visualization, pattern recognition, neural networks and so on.

2.7. The Data Mining Process Models

There are different data mining process models which are designed to work for research purpose, industrial projects purpose and hybrid ones. Process model as defined in [31] is a set of framework activities and tasks to get a job done, including inputs and outputs in every task. The final objective of a process model is to make it manageable, repeatable and measureable. The characteristics of a good process model include: *Effectiveness* in producing the right output, *Maintainable* so that we can easily find and remedy faults, *Predictable* in helping put the steps of development so that we can plan and allocate resource for the process, *Repeatable* once it is discovered it should be replicated to other future projects, *Improvable* since development environments and requested products are changing quickly our process will also change to catch up, and *Traceable* which can allow to follow up the project [25].

The efforts to establish KDP model were initiated in the mid 1990s, when data mining is being shaped, researchers started defining multistep procedures to guide users of data mining tools in the complex knowledge discovery world [8]. The main emphasis was to provide a sequence

of steps that would help to execute a KDP is an arbitrary domain. In 1996 the first nine-step model is developed by Fayyad et al [50].

Later industrial models have followed the academic ones and different approaches were proposed such as the five-step model, IBM model and Industrial model six-step CRISP-DM [8]. The CRISP-DM becomes a leading model in the industry.

The different data mining process models exist are categorized in three categories as KDD related, CRISP-DM related and others [25]. This summary paper reviewed 14 data mining process models and classified in the above categories noting that the KDD is the initial approach and CRISP-DM as a central approach and the others are based on those two categories. According to this paper the models that fall under KDD related category include Human-Centered, SEMMA, Cabena et al., Two Crows and Anand & Buchner. Those that are under CRISP-DM related are Cios et al.(six step KDP), RAMSYS, DMIE and Marban et al. In others category we find 5A's and 6o.

The common features of these steps include the process includes multiple steps executed in sequence. Each subsequent step is initiated upon successful completion of previous step and requires the result generated by the previous step as its input. Another common feature is the range of activities starting from understanding the project domain and data through data preparation and analysis, to evaluation, understanding and application of the generated results. All of them also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process [8]. The main differences between them lie in the number and scope of their specific steps.

We shall see the basic ones the KDD process model and the CRISP-DM model and in addition the six step KDP model (derived from CRISP-DM for academic research purposes) which is selected as a methodology for this research purpose in detail.

2.7.1. The KDD process Model

The KDD refers to the overall process of discovering useful knowledge from data [15]. The basic problem addressed by the KDD process is mapping low-level data in to other forms that might be more compact, more abstract or more useful. It is defined in another way as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [15]. The KDD as a process is defined as interactive and iterative, involving nine-steps with many decisions made by the user. These steps are shown in fig 1 below.

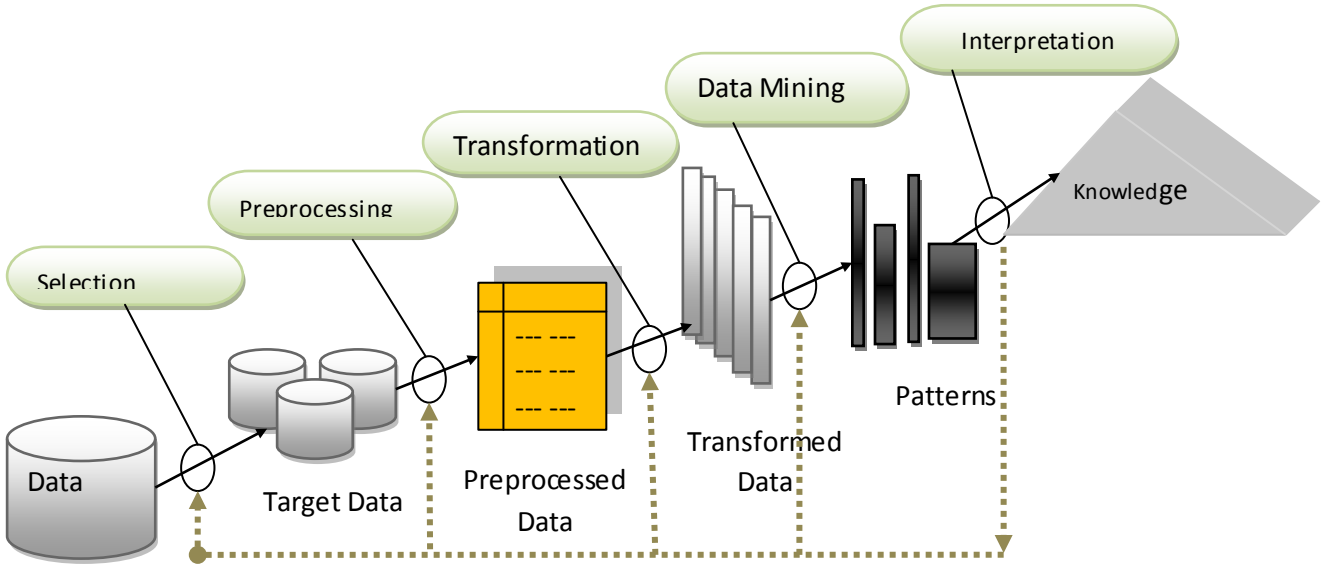


Figure 1 steps of KDD process (Fayyad et al. 1996)

Developing and understanding the application domain: learning the relevant prior knowledge and goals of the application

Creating a target dataset: includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed

Data cleaning and preprocessing: this step includes operations such as removing noise and outliers, decide on strategies of handling missing field and deciding on database management issues.

Data reduction and projection: finding useful features to represent the data depending on the goal of the task and applying dimensionality reduction and transformation methods to reduce the data.

Choosing the data mining task: deciding on the function of the model derived (classification, clustering,…).

Choosing the data mining algorithm: selecting methods to be used for searching the patterns

Data mining: searching for patterns by running the selected data mining algorithms on the prepared data.

Interpreting mined patterns: possible visualization of the mined patterns, removing redundant or irrelevant ones and translating the useful ones into terms understandable by the user

Consolidating discovered knowledge: includes incorporating the discovered knowledge in to the system, taking actions based on it and resolving potential conflicts with the previous and new knowledge

2.7.2. CRISP-DM process Model

Starting from the 1990 s different organizations were having different models for their data mining tasks. The CRISP-DM (Cross Industry Standard Process for Data Mining) was created for the purpose of having a standard for the processes involving data mining projects. It was created by a group of three companies Daimler-Benz, SPSS, and NCR for industry purpose in 1996 [7].

The CRISP-DM process model provides an overview of lifecycles of a data mining project. It contains phases of a project, their respective tasks, and relationships between these tasks. According to this model the life cycle of data mining project consists of six phases shown in figure 2. The sequence of the phases is not rigid moving back and forth between the different phases is possible. The outcome of each phase determines which phase has to be performed

next. The arrows indicate the most important and frequent dependencies between the phases. The outer circle in the figure shows the cyclical nature of data mining. It does not end once the solution is deployed. The lessons learned during the process and from the deployed solution can trigger new [7].

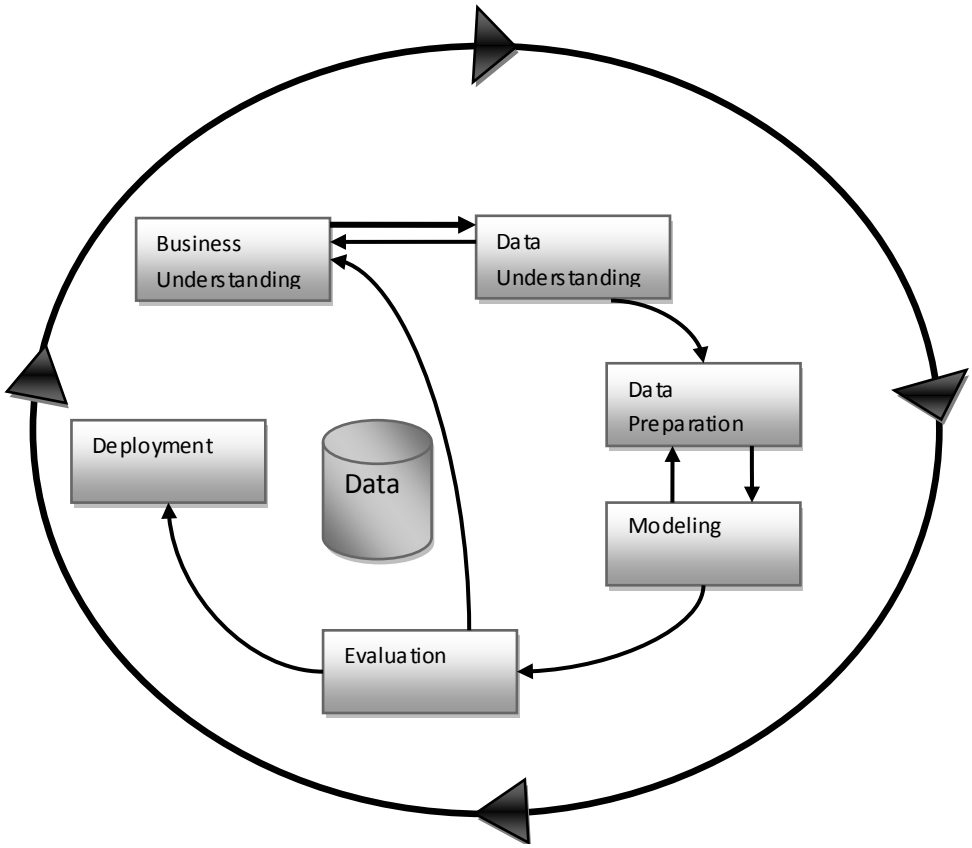


Figure 2 phases of the CRISP-DM reference model

The six phases of CRISP-DM model are explained as follows:

Business understanding: this is the initial phase which focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding: this phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify problems, to discover first insights into the data, or to detect interesting subsets to form hypothesis for hidden information.

Data preparation: this phase covers all activities to construct the final data set from the initial raw data. Data preparation tasks are likely to be performed

Modeling: various modeling techniques are selected and applied on the prepared data, and the optimal value of variables is obtained.

Evaluation: at this stage the model built before processing the final deployment is thoroughly evaluated and the steps executed to construct the model are reviewed to be sure it properly achieves the business objectives. At the end of this phase a decision on the use of the data mining results should be reached.

Deployment: this phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. The deployment step is carried out by the customer itself in many cases.

CRISP-DM is revised again in 2007 based on changes that occurred in the business application of data mining like availability of new types of data, integration of results with operational systems, deployment into real time environments and mining large-scale databases [25].

2.7.3. The six-step KDP Model

The existence of academic and industrial models has led to the development of hybrid models which combine aspects of both [8]. This is the six-step KDP model shown in Figure 3 which is developed by Cios et al. (2006). It is developed by adopting CRISP-DM in to academic researchers in a way that provide more general, research-oriented description of steps, introduces mining step instead of modeling step, has more detailed feedback mechanisms and the last step is modified to use the discovered knowledge to be applied in other domains [8].

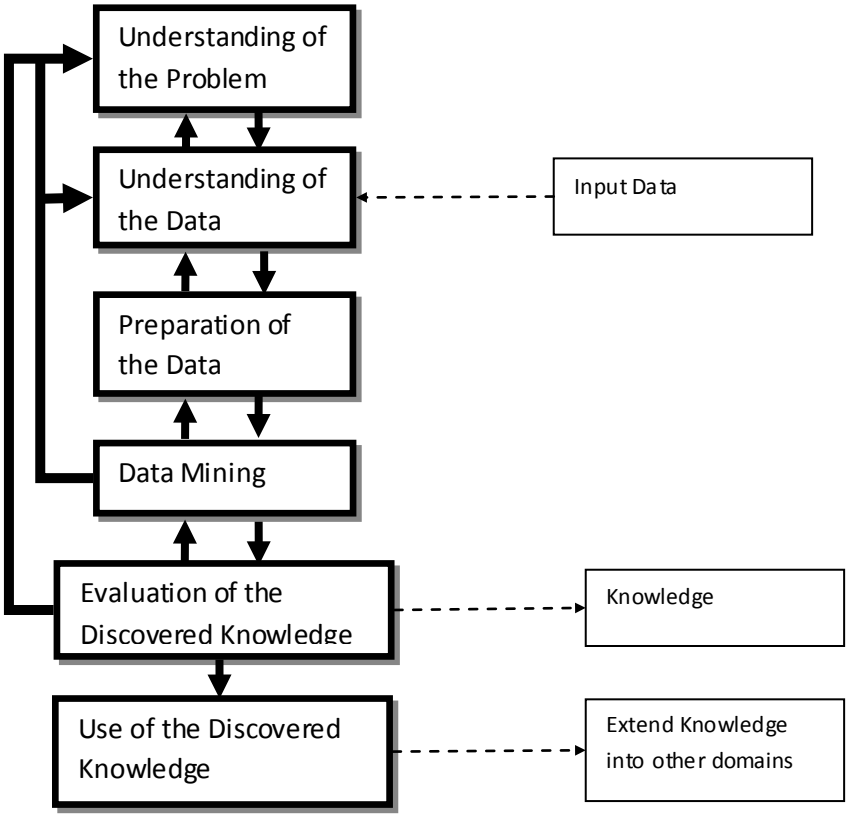


Figure 3 the six-step KDP model

Understanding of the problem domain: this step involves working closely with the domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also includes learning domain specific terminologies. Problem description is prepared and the project goals are translated in to data mining goals. And the initial selection of data mining tools to be used is performed.

Understanding of the data: this step is guided by back ground knowledge and includes collecting and deciding which data to use including the format and size. The data is then checked for completeness, redundancy, missing values. Then the data will be checked for usefulness to the data mining goal.

Preparation of the data: it involves sampling, running correlation and significance tests, and data cleaning tasks are accomplished. The end result is data that meet the specific inputs required for the data mining tools selected.

Data mining: using different data mining methods and algorithms to derive knowledge from the preprocessed data.

Evaluation of the discovered knowledge: includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts. Only approved models are retained.

Use of the discovered knowledge: this is the final step which consists of where and how to use the discovered knowledge. The application area of the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and entire project is documented. Finally, the discovered knowledge is deployed.

2.8. Data Mining Tools

There are different open source data mining tools available to support the tasks of data mining. The tools available in the beginning were invoked from command prompt which requires programming knowledge. Then with contributions of research communities now a days we can get many tools having interactive graphical user interface [60].

The known general-purpose tools for data mining, machine learning and statistics include Weka, R, Tanagra, KNIME and Orange. From which Weka and R are the most popular one's [60]. The choice differs with the implementation language, interactive graphical user interface, documentation, stability, performance, visualization of the models etc…

R is a language and environment for statistical computing and graphics [60]. This tool needs scripting knowledge to manipulate. While Weka is the best known open-source machine learning and data mining tool. It can be accessed through java programming or command line interface or using a Weka explorer graphical user interface.

2.9. Application areas of Data Mining

Data mining tools and techniques can be successfully applied in different fields [5]. Now a day's different organizations are using it as a tool to deal with the competitive environment. It is widely used in health industry, Auditing, telecommunication industry, financial organizations and retail. In the medical area it has applications like genetics to understand the mapping relationship between the variation in human DNA sequences and disease susceptibility and is also very important in diagnoses, prevention and treatment of diseases [5]. In retail markets it is used to understand customer buying behavior, shopping pattern and others based on the large transaction data. The knowledge discovered from the retail transaction data is used to fin change in customer consumption, adjustment of price and variety to promote sales and attract customers. In telecommunication data mining helps to identify and compare system loads, data traffic and profit and telecommunication fraud.

The other area where data mining is widely used is in financial organizations data. Like bank insurance and the like which is the focus of this research. The data in the financial sector is relatively quality data [5]. For banks it is used in specific areas like loan payments and customer credit policies by analyzing income level and payment to income ratio the bank can make decisions of loan granting based on the risk. Another area in financial sectors is to have groups of customers having similar characteristics and facilitate some common service accordingly. Data mining is also used to detect fraudulent customer behaviors [5].

2.9.1. Application of data mining in Insurance

In insurance industries data mining can be used in different areas for customer relation management CRM, product development since there is a customer data which is registered for one time transaction. It can be applied to study customer behavior and living style so that a new product can be designed for different groups of customers according to their standards of leaving. Application areas identified in [12] in Insurance are: Discovery of medical procedures that are claimed together through claims analysis, Identification of customers that are potential

buyers for new policies, Detection of behavior patterns capable of identifying risky customers, Detection of fraudulent behavior.

In addition to these [11] include existing customer's retention as data mining application area in insurance by analyzing their data and provide package discounts for their stay according to their policy behaviors. This is an issue of most Insurance companies because of new customer acquisition cost increases over time greater emphasis is given for customer retention programs [11].

Furthermore identifying risk factors that predict profits, claims and loses, reinsurance, financial analysis, Estimating outstanding claims provision and detecting fraud are listed as application areas of data mining in insurance [51]. The modern data mining models such as decision trees and Neural Networks are found to predict risk than current actuarial models [18]. So that the companies can use the risk assessment results for pricing. For the case of reinsurance which is a field of risk management in which an insurance company transfers a portion of its risks to another insurance or reinsurance company in consideration of premium payment [18]. Data mining can develop predictive models based on previously paid claims data which will be used to identify policies which need reinsurance cover. In case of outstanding claim provision: an estimate amount of claim is taken from the beginning based on severity of claim, likely delay time before settlement and inflation and interest. Data mining techniques can be applied to predict the estimate amount based on analyzing insurance claims data. Insurance also uses data mining for claims management, which is claims data analysis for fraud analysis.

2.9.2. Application of Data mining for fraud detection

There are plenty of specialized fraud detection solutions which protect businesses such as credit card, e-commerce, insurance, retail, telecommunication industries [32]. As stated in [32] the main problem in data-mining based fraud detection systems is scarcity of real data. Also in e-commerce the data mining task is challenging because of the boundary is not clearly known

between network intrusion detection systems and fraud detection systems. As it is a survey paper it has gone through different fraud detection systems available and suggests hybrid algorithms that is using many supervised algorithms or applying both supervised and unsupervised algorithms to process the data and detect the suspicion scores and rules.

When we come to insurance fraud detection according to [35] mainly used algorithms in detecting automobile insurance fraud are Bayesian Network to classify the data and Decision Trees to build predictive models. Bayesian network is constructed first for legal cases by training the net with characteristics of legal drivers which is taken from real legal driver's data and then for fraud cases by training the characteristic of fraudsters taken from experts' previous knowledge. From the experiment result the probability of fraud is known then Decision Tree is constructed. At last the performance of the model is tested using accuracy, recall and precision which are derived from the resulting confusion matrix.

In health insurance fraud detection [6] notes that the fraud detection system can be specific to each country, usually based on gaps or weaknesses of legislation. Models are consistently changing because individuals always seek new ways to circumvent the law [6]. This paper lists the following as the main types of fraud that can be identified in Romanian insurance health systems.

- Unusual high number of invoices for a particular insured person in a short time.
- Use of false identities for claiming false hospitalization and false prescription;
- Claiming medical invoices having dates outside the insurance period;
- Claims having payable amounts higher than amounts that the insurance will pay.

This paper briefly describes the analysis methods applied in the field of health insurance fraud detection. Each of the methods in relation with the type of fraud they can be applied is described. And at last recommends hybrid solution in terms of technologies and models of analysis is best solution for fraud detection.

A recent research conducted on medical expense insurance fraud in China has first compared using unsupervised methods and supervised methods for medical insurance fraud detection [44]. As a result the supervised ones are found more accurate but it puts some limitations: one it is difficult and costly to obtain labels for training sample; two, unbalanced data due to the fraudulent cases registered will be few compared to the valid ones and the labeling of dependent variables could be inaccurate [44]. For the research they have used the supervised method called discrete choice model. In the research they have taken medical insurance claim data of one insurer of the year 2009 and 2010 a total of 8073 claim records of which they have considered the non-paid claims as fraudulent and partially and totally paid ones as valid. After the experiment predictive factors of fraudulent medical claims are found. Most of the factors are related to either medical service provider or characteristics of the insurance policy. The result indicates hospital qualification, total cost of health care, policy holders renewal status, claim duration are significant factors of medical insurance fraud in the market. The total percentage of correct classification is found to be 67%.

A research conducted in turkey insurance company for detecting fraud in health insurance performed anomaly detection analysis by running support vector machine (SVM) algorithm [21]. The system is trained to determine a boundary between normal and anomalous records. Then each record is compared with the boundary and classified in to the two categories. For the research they have used 808,348 claim records of nine years. The data mining software calculates the probability of the anomaly if each record. If the probability is greater than 50% then the record is marked as anomalous. 6595 records are found in the range of probabilities with 50% to 67.3%. Then those records found anomalous are analyzed based on criteria's like rejected claims, excessive claims in health center types and excessive claims in health centers. Lastly they concluded that data mining methods such as anomaly detection, clustering and classification can successfully detect outliers in large data sets. This is helpful for insurance

industries to detect fraudulent claims. Once the anomalous claims are detected several analyses will be made. The main task in the analysis is to narrow the target for detecting frauds.

2.9.3. Local Related works

A research was done on tax fraud detection taking the Ethiopian Revenue and Customs Authority as a case [9]. In this case fraud is defined as involvement of one or more persons who intentionally act secretly to deprive the government income and use for their own benefit [9]. In the study k-means (for clustering) and J48 decision tree and Naive Bayes (for classification) algorithms were applied. Hence the model built using the J48 algorithm was tested using test data set with 2200 records and 99.98% classification accuracy and a prediction accuracy of 97.19% is found. The researcher has tried to recommend on future research areas like testing the other data mining algorithms (Neural Network and Decision Tree) to be applied for tax fraud detection.

Tesfaye conducted a research on Nyala Insurance in which he developed a predictive model for insurance risk assessment, which specifically detects behavior patterns to identify risky customers in motor insurance [48]. In the research he used customers and insured vehicle information and applied decision tree and neural network in building the predictive model. The dataset is 1056 records of which 90% is used for training and 10% of it used for testing. Using decision tree 95.69% of the validation set is correctly classified, and the classification accuracy for low, medium and high risk policies are 98.15%, 94.12%, and 92.86% respectively where as neural network model correctly classified 92.24 % of the validation set, high-risk groups are correctly classified, and low and medium-risk groups are classified with accuracy of 98.15% and 76.47% respectively. In the research a new pattern was found between the two models that some policies misclassified by decision tree were correctly classified by neural networks, and vice versa. So he concludes that hybrid of the two models may result in better classification accuracy.

Data mining algorithms are tested to work for vehicle insurance fraud detection [47].The research is done based on data collected from Africa Insurance Sh. Co. In the study clustering algorithm (k-means) and classification algorithms (J48 decision tree and Naive Bayes) algorithm are used to build a predictive model in order to classify motor insurance claims which are frauds and non frauds. Here attributes like age, sex, police report are not included since there is no record found showing those attributes. But those attributes especially the police report are important in fraud detection. With data lacking these values the researcher concludes that the result is promising and with the model developed by J48 decision tree 99.96% classification accuracy and 97.19% prediction accuracy is obtained. At last the researcher tried to recommend future research areas in relation to the specific issue. One of the areas is applying data mining algorithms to detect fraud in other insurance types than motor insurance.

CHAPTER THREE

3. METHODOLOGY OF THE RESEARCH

3.1. Research Design

This research is an experimental research conducted on medical insurance claim data collected from Ethiopian Insurance Corporation. In this chapter the methods, techniques and tools used to conduct the research are discussed in detail based on the steps of hybrid data mining process model selected to guide the entire process of this research.

As it is discussed in chapter two data mining process models are developed for purely academic purposes and also for Industrial purposes. By combining the two a hybrid model is also developed. This Hybrid model is organized by taking the Cross Industry Standard Process for Data Mining (CRISP-DM) model by adopting it to academic researches [8]. The process model adopted to undertake this research is the hybrid one due to the reason that this model describes each of the knowledge discovery process steps in a better way and it is flexible since it has a feedback mechanism in more steps than the CRISP-DM.

The hybrid model has six steps that are: understanding of the problem, understanding of the data, preparation of the data, data mining, Evaluation of the discovered knowledge and use of the discovered knowledge. The research is designed based on steps of this process model.

3.1.1. Understanding of the problem

To understand the problem primary and secondary sources are used. Secondary sources include books, local and international articles written in general on data mining and specific to the area of fraud detection and the different underwriting and claim manuals reviewed; Discussions made with the insurance experts working in the area is taken as primary source. This helps to select the appropriate algorithm and adopt proper methodology for the research. Another important step to be performed here is selecting the tool to be used. In this research WEKA data mining tool is used since it is platform independent and contains a graphical user

interface which is easy to understand. Moreover, it contains different data mining algorithms and the researcher is also familiar with this tool. Furthermore it can be freely downloaded from the internet.

3.1.2. Understanding of the data

After the medical insurance claim data is collected from existing manual as well as electronic files the attributes found in the data are validated with the experts so as to understand the purpose of each of the parameters and whether they are useful for the specific problem area or not. Then attributes are described from different data points.

3.1.3. Preparation of the data

In this step data cleansing tasks are performed. These are removing duplicate records, correcting noisy data, filling missing values by using estimates, removing irrelevant attributes and records i.e. attributes and records which are out of interest of the data mining task, there will also be inconsistencies and redundancies while the data from the different sources is integrated. The cleaned data with the above techniques is further processed for dimensionality and numerocity reduction. For dimensionality reduction purpose the WEKA attribute selection (GainRatioAttributeEval attribute evaluator together with Ranker search method) is used. Finally the input data set for running the classification and clustering algorithms as training and test data is generated.

3.1.4. Data mining

The data is labeled as valid or non-valid (suspicious to fraud) by taking the assumption that all the claims which are rejected while all the claim documentation available are non-valid and the paid ones are valid. This assumption is reached after discussion with the experts and there is no way that they can check further for validity on claims which are already paid as they already settled the cases assuming they are valid. This approach for detecting fraud is also used in the work done for detecting health insurance fraud in China [44].

Classification algorithms Bayesian Network and J48 decision tree are used for developing the predictive model. These classification algorithms are tested for the area of fraud detection in different previous works and they have performed well.

3.1.5. Evaluation of the discovered knowledge

The models which are captured from the data mining step are evaluated. Testing the classification accuracy is done by using 10 fold cross validation method and percentage split test mode by changing the percentage. The algorithm is tested on test data set to see how many of the test set are classified as true positive and false positive and then calculating recall and precision. The performance of the algorithms is also evaluated by percentage of classification and time parameter: the time taken to build the model. At last with the involvement of the domain experts the resulting model is evaluated.

3.1.6. 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. For development of the prototype Microsoft Visual Studio 2012 with SQL server 2008 are used. Those environments are selected because the researcher is familiar with them.

3.2. Data mining methods for fraud detection

The main issue in fraud detection is the fact that the collections of data are impossible to be processed by human brain [6]. In this article the methods and techniques used for fraud detection are categorized in to two as **statistical techniques** and **artificial intelligence**. Statistical data analysis techniques include: Data preprocessing techniques, calculation of different statistical parameters and probability distributions, time-series analysis of time dependent data, clustering and classification and matching algorithms. The artificial intelligence category includes data mining to classify cluster and segment the data and automatically find association and rules in the data that shows interesting patterns related to fraud, expert

systems to encode rules for detecting fraud, Machine learning automatically identifies characteristics of fraud.

The data mining techniques which are tested to work in the area of fraud are Naïve Baye's and Decision tree classification algorithms for tax fraud [9] and vehicle insurance fraud [35][47] detection. They both have used unlabelled data so that before applying the classification algorithms clustering is used. For clustering the data in to different clusters k-means is applied then human experts label in to fraud and valid cases. Then after classification algorithms are used to build a model for predicting the incoming cases.

For the purpose of this research the data will be initially labeled by the assumption that the rejected claims with the appropriate claim documents available are taken as fraud cases and the rest are taken as non-fraud. This means that clustering will not be used and only classification algorithms are used to predict the incoming claims. The already tested algorithms (Naïve Bayes and J48 decision tree) to work for fraud detection are tested for the medical insurance claim fraud detection. We shall see the description of those data mining methods in the following sections.

3.2.1. Naïve Bayes Classification Technique

Bayesian classifiers are statistical classifiers [20] and it is based on Bayes' rule. The formula states that assuming the event of interest A happens under any of hypotheses H_i with a known (conditional) probability $P(A|H_i)$. Assume in addition that the probabilities of hypotheses $H_{1,\ldots,}$ H_n are known(prior probabilities). Then the conditional (posterior) probability of the hypothesis H_i i=1, 2 ,… , n given that event A happened is

$$P(H_i|A)= \frac{P(A|H_i)\,P(H_i)}{P(A)} \text{ , Where } P(A)= P(A|H_1)\,P(H_1)+\ldots+ P(A|H_n)\,P(H_n).$$

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 [56]. 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 [20].

Naïve Bayes is found simple but effective classification algorithm in solving real life problems [55]. This algorithm also performs well even when attribute dependencies exist [13]. 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. [55][13] After referring the different modifications concluded that such modifications lead to complications which deviate from its basic simplicity.

Naïve Bayes algorithm

The Naïve Bayesian classification algorithm works as follows [20]:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X= (x_1,x_2,…,x_n), depicting n measurements made on the tuple from n attributes, respectively ,A_1,A_2,…,A_n.
2. Suppose that m classes, C_1,C_2,…,C_m. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on

X. That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$P(C_i|X)>P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$

Thus we maximize $P(C_i|X)$. in Bayes theorem $P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$ the class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis.

3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need to be maximized . if the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is $P(C_1)=P(C_2)=\ldots=P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i)=|C_i,D|/|D|$, where $|C_i,D|$ is the number of training tuples of class C_i in D.
4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (that is there are no dependence relationships among the attributes). Thus,
   $P(X|C_i)=\prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$

   We can easily estimate the probabilities $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ from the training tuples, Recall that here x_k refers to the value of attribute A_k for tuple X. For each attribute, we look at whether the attribute is categorical or continuous-valued.
5. In order to predict the class label of X, $P(X|C_i)P(C_i)>P(X|C_j)P(C_j)$ for $1 \leq j \leq m$, $j \neq i$ in other words, the predicted class label is the class C_i for which $P(X|C_i)P(C_i)$ is the maximum

3.2.2. Decision tree classification technique

Decision tree builds classification models in the form of a tree structure by breaking down a dataset into smaller and smaller subsets incrementally. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The top most decision node in a tree is called root node [24]. Decision trees can handle both categorical and numerical data.

Decision tree is also defined in [20] as a flowchart-like structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. The goal is to find the optimal decision tree by minimizing the generalization error [24]. Most decision tree classifiers perform classification in two phases: tree-growing or tree building and tree-pruning [52]. The tree-building is done in top-down manner and during this phase the tree is recursively partitioned till all the data items belong to the same class label. In the pruning phase the full grown tree is cut back to prevent over fitting and improve the accuracy of the tree in bottom up fashion.

In the tree-building phase there are different splitting and stopping criteria available. The splitting criteria include Univariate (impurity-based, information gain, gini index, gain ratio, distance measure, and more others) and Multivariate splitting criteria. The stopping criteria include all instances in the training set belong to a single value, the maximum tree depth has been reached, the number of cases in the terminal node is less than the minimum number of cases for parent nodes, the best splitting criteria is not greater than a certain threshold [24].

Applying tight stopping criteria tends to create small and under-fitted decision trees; on the other hand, using loosely ones tends to generate large trees that are over-fitted to the training set [52]. Pruning methods are developed to solve these problems. In pruning first a loosely

stopping criteria is used to make the tree over-fit. Then after, the resulting over-fitted tree is cut back into smaller pieces of trees by removing sub-branches that are not contributing to the generalization accuracy. There are different pruning methods exist like cost-complexity pruning, reduced error pruning, minimum error pruning, error-based pruning, optimal pruning, minimum description length pruning and more others [24].

The first decision tree algorithm is ID3 (Iterative Dichotomizer) developed by J.R. Quinlan [20]. The idea behind decision tree induction is that many correct decision trees can be built for classifying objects in a training set, but it is needed to go beyond the training set and classify those unseen objects correctly to a class where they belong [33]. As it is explained this work to expand to classification of those new objects decision tree should capture some meaningful relationship between an object's class and its values of the attributes. One approach to this task is to generate all the possible decision trees that correctly classify the training set and to select the simplest of them. The ID3 is designed for classification where there are many attribute and the training set contains many objects, but where relatively good decision tree is required without much computation.

ID3 employs a top-down, greedy search through the space of possible branches with no backtracking. Entropy and Information gain are used to construct the decision tree in ID3 [24]. A decision tree built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. In ID3 entropy is used to calculate the homogeneity of a sample. That means if the sample is completely homogeneous then the entropy is 0 and if it is equally divided it has entropy of 1.

$$E(S) = \sum_{i=1}^{n} - p_i \log_2 p_i,$$ where S is the set and pi are examples in set S.

Information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

Gain= Entropy(S) - $\sum_{i=1}^{n} p_i \log_2 (p_i)$, where parent node, S is split onto p partitions

The construction of the tree stops when all instances belong to a single value of target feature or when best information gain is not greater than 0. This algorithm does not apply any pruning procedures and it does not handle numeric attributes or missing values [24].

ID3 is extended to C4.5 by the same person in 1993 and this one uses gain ratio for splitting criteria. The splitting stops when the number of instances to be split is below a certain threshold. After the growing phase it uses error-based pruning. This algorithm can handle numeric attributes and incorporates missing values. The gain ratio is normalization of the information gain.

$$Gain\ ratio = \frac{Information\ gain}{Entropy}$$

Based on the concern that making the algorithms more efficient, cost effective and accurate for different areas in real world different other decision tree algorithms are developed to entertain the difficulties in the existing ones. Some known decision tree algorithms include CART, CHAID, QUEST, BF-TREE.

Decision tree algorithms are most powerful approaches in data mining [23]. They are relatively fast in classifying unknown records, it is able to handle both discrete and continuous attributes but for continuous valued it is not performing well. It is also easy to interpret and performs well in the presence of noisy data. Decision trees provide a clear indication of which fields are most important for prediction and can be implemented in data mining packages over variety of platforms [23].

The J48 decision tree algorithm

The J48 decision tree is an open source java implementation of C4.5 decision tree algorithm in WEKA data mining tool [19]. The basic steps in the algorithm are:

1- In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class

2- The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute
3- Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

The tree building process uses entropy and information gain.

Features of the J48 algorithm

1- Both discrete and continuous attributes are handled by this algorithm. A threshold value is decided by C4.5 for handling continuous attributes. This value divides the data list into those that have their attribute value below the threshold and those greater than or equal to it.
2- It handles missing values in the training data.
3- After the tree is fully constructed, it performs pruning.

3.3. Evaluation methods

The performance of classification algorithms is usually examined by evaluating the accuracy of the classification [20][28]. The other evaluation approaches like time and space can be also measures but they are secondary; which is best is also depends on the interpretation of the problem by users [28]. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified [20]. Confusion matrix is a tool to analyze classification accuracy.

Confusion matrix is given n classes an m x n matrix where $C_{i, j}$ in the first m rows and n columns indicates the number of tuples of class I that were labeled by the classifier as class j. A classifier is said to have good accuracy by drawing an ideal diagonal from $C_{1, 1}$ to C m, n the rest of the entries outside the diagonal are close to zero [20].

		Predicted Class	
Actual Class		C1	C2
	C1	True positives	False negatives
	C2	False positives	True negatives

Fig 4: A confusion matrix for positive and negative entries

True positive: if the outcome from a prediction is p and the actual value is also p then it is called a true positive.

False positive: if the outcome from a prediction is p however the actual value is n then it is called false positive.

False negative: positive tuples that are incorrectly labeled as negatives

True negative: if the outcome from a prediction is n and the actual value is also n then it is called a true negative.

Precision and recall are also used as measure of relevance. In this interpretation

Precision = true positive / true positive + false positive

Recall = true positive / true positive + false negative

Accuracy of a classifier is measured on a test set consisting of class labeled entries that were not used to train the model. That is the percentage of test set tuples that are correctly classified by the classifier. On the other hand it can be described by error rates. Error measures calculate how far is the predicted value from the actual known value? The most common error functions include

- Absolute error: $|Y_i - Y_i'|$ where Y_i is the actual value and Y_i' is the predicted value
- Squared error: $(Y_i - Y_i')^2$
- Mean absolute error: $\frac{\sum_{i=1}^{d} |Yi - Yi'|}{d}$, the average absolute error over the dataset

- Mean squared error: $\frac{\sum_{i=1}^{d}(Yi - Yi')2}{d}$, average squared error over the dataset

The average mean squared error exaggerates the presence of outliers while the mean absolute error does not. We can normalize the above errors by the mean value Y.

- Relative absolute error: $\frac{\sum_{i=1}^{d}|Yi - Yi'|}{\sum_{i=0}^{d}|Yi - Y|}$
- Relative squared error: $\frac{\sum_{i=1}^{d}(Yi - Yi')2}{\sum_{i=1}^{d}(Yi - Y)2}$, where y is the mean value of Yi's i.e, $\frac{\sum_{i=1}^{t} Yi}{d}$

Percentage split (holdout), random sub sampling, cross-validation, and bootstrap are common techniques for assessing accuracy based on random sampled partitions of the given data [20]. The cross validation and percentage split methods are applied for the purpose of this work.

10-fold cross validation

In this method the dataset are partitioned in to 10 mutually exclusive subsets of equal size. Then training and testing is performed 10 times. For each iteration i partition D_i is reserved as a test set, and the remaining are used to train the model. Classification accuracy using this method is calculated as the overall number of correct classifications from the 10 iterations, divided by the total number of tuples in the initial dada. And prediction accuracy is equal to the total errors from the 10 iterations divided by the total number of initial tuples [20]. It is recommended to use this method to estimate model accuracy since it is less bias and variance.

Percentage Split

The classifier is evaluated with certain percent of the data which is held out for testing. The percentage amount of data to be held out is specified by user and the accuracy varies based on the data.

CHAPTER FOUR

4. BUSINESS UNDERSTANDING, DATA UNDERSTANDING AND DATA PREPAPRATION

4.1. EIC Business Practice

As it has been explained by the officers interviewed, EIC has implemented Business Process Re-engineering (BPR) in 2010. The processes are structured in one core and six support processes. The core process is insurance Service Process which is also divided in to life insurance and general insurance sub-processes. The six support processes include Business Development and Risk Management, Resource Management, Finance and Investment, Information Technology service Management, Internal Audit and Legal Service processes. EIC is led by the Chief Executive officer (C.E.O.) and administered by the board of management which reports to the public Financial Enterprises Agency which is accountable to the Prime Minister's Office.

The focus of this research is medical insurance claim fraud and it is structured life insurance and general insurance sub-processes. These sub processes are also structured in to corporate and retail business teams. The corporate team is responsible for handling business from organizations and the retail team takes care of businesses from individual clients. Claims of medical insurance are handled in both retail and corporate business teams. The higher officials as well as the performers were responsive and eager to see the findings of the research.

4.1.1. Classification of Medical Insurance policies in EIC

Medical insurance coverage is provided as a cover in the general insurance part that is with workmen's compensation and Personal accident Insurance where as in life insurance part it is provided as product by itself. EIC defines medical expenses as expenses properly incurred by an insured person for medical, surgical, massage, therapeutic, x-ray or nursing treatment including the cost of medical supplies and ambulance hire; but exclude the cost of board and lodging [37].

4.1.2. Underwriting and Claims handling process of Medical Insurance in EIC

Operational jobs in EIC are fully automated. The underwriting and claim processing is accomplished by using the automated system. Implementation of Insurance and accounting systems is done in the year 2010 and it became fully operational by 2011.

4.1.2.1. Policy Underwriting

To underwrite an insurance policy, the insured person should be between ages 14 and 65. Ages for renewal policies could be extended to 70. Information about the insured needed at the time of underwriting includes: name of insured, monthly salary, occupation, age and sex [37].

Medical expense cover provided has maximum aggregate limit of an agreed amount for anyone person in anyone year. The maximum limit set in life is greater than the limit in general insurance category. Expenses incurred for the following are excluded from the insurance coverage:

- In connection with venereal disease, intentional self injury or resulting from drunkenness
- In connection with treatment not undertaken by the direction of a registered medical practitioner
- In connection with any chronic disease an insured person has had in history
- In connection with any claim arising as a result of illness within 21 days from the commencement of the cover.

The group policies could be issued in named or unnamed bases. Most of the time large organizations having large number of employees and frequent entrance and exit of employees will buy policies in unnamed bases but to be adjusted by the end of the policy year when the exact number of employees with their wages is delivered from the insured. In this case employee detail (name, sex, age) is not delivered at the time of underwriting rather the total

number of employees in groups by their occupation and salary. For these product types an agreed rate of premium is applied to underwrite and the details are not taken into consideration for premium calculation [37].

4.1.2.2. Claims Processing

Claim processing is initiated by the insured when s/he notifies an accident or illness happened [37]. Notification could be done through telephone, e-mail or in person but the claim notification should be signed by the insured. When notifying claim for medical insurance in both life and general cases the documents that need to be provided are claim notification slip which is filled by the insured with the necessary details of the claim, Medical certificate, Doctors prescription, Medical board disability certificate /if there is disability, and invoices for paid expenses. Whether the policy is named or unnamed, during claim the insured should deliver the exact salary of the employee at the time of illness and his/her name needs to be mentioned.

After the claim request is presented by the insured, the date of medication and the covers with exclusions of the policy are verified to see if they are all in order. The claim verification is done using the existing Insurance system. After it is registered in the system according to the cover provided in the policy, a hard copy claim file will be opened and numbered according to the system generated claim number.

The assigned claim experts check for validity of the claim simply by tracing the date of the invoices and the prescription as well as the total amount requested is within the policy limit. For the invoices which are in order claim payment will be prepared. Currently since the claim experts are not medical practitioners the medications are not checked if they are related with the said illness. Only when the claim officer is aware of some medications that he will consult a professional to make sure that the claimed medical expense is related with the illness. To date, this is the only way the corporation uses to differentiate between fraudulent and valid claims.

Rejection later will be prepared for invalid claims and for valid claims the prepared payment will be assigned to the respective higher level officer or the team leader for approval. The claim approval authority limit of the different positions differs based on the claim payment amount.

From the different types of insurance categories, the areas mostly perceptible to fraud are vehicle insurance followed by medical insurance [32].

4.2. Understanding the data

To apply data mining, having the data at hand is the prerequisite. This phase includes the initial step of collecting data and understanding the relationship of the data to the data mining problem to be solved by the study. Here the data fields are described and analyzed through discussion with the domain experts.

4.2.1. Data collection

The data used for this study is collected from the centrally managed operational databases of EIC. Life and non-life have two different databases but the tables have almost the same data structures. The number of records used is 20821 from life and 16054 from non-life. The data covers medical insurance claim data taken from different tables related to the problem. Those related to the claim from CLAIM and CLAIM_OBJECTS tables and those related to the insured properties are from P_INSUREDS and GEN/LIFE_RISK_COVERED tables and those related to the state of the claim documents are collected from CLAIM_AVAIL_DOCS table. The data covers the claims that are reported in the year 2011 and 2012 in the different branches of EIC. During the period, life was only operational in the main Life_Addis district and non-life in 18 districts and branches. Rejected claims records are those assumed to be suspects for fraud. In life insurance however, the experts do not register those rejected cases. So for the selected insurance period these records are collected from manual hard copy files and converted to electronic form. The distribution of the data from both categories is shown in table 4.1 below.

Category	General Insurance		Life Insurance		Total
Rejected	2792	**7.6%**	3257	**8.8%**	**6049**
Resolved	13262	**36%**	17564	**47.6%**	**30826**
Total	**16054**		**20821**		**36875**

Table 4.1.:- Distribution of initially collected data

As it is seen from the table the data seems unbalanced that is rejected cases are much less in number from the resolved ones.

4.2.2. Description of the data

Fields from the different tables are selectively taken by consulting with business experts. EIC allows providing data only if it is not confidential information. Therefore, records related to customer identity and addresses are not included even in the crude data from the beginning. In addition, records related to premiums are also hidden according to the company rule. From the data restricted by the company, the premium data could be related with the problem to make further analysis on claim ratio of fraudulent cases. Some parameters have null values and others are not related to the problem in hand. By merging all the attributes from the different tables a total of 20 attributes are taken. The data is directly exported to an EXCEL format. The description of the attributes with the relative tables is shown in tables 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7.

No.	Attribute Name	Data type	Description
1	CLAIM_OBJ_SEQ	NUMBER	A system generated serial number which identifies claim_requests
2	INSURED_ID	NUMBER	Is the system generated id for the specific insured
3	CLAIM_ID	NUMBER	Sequence generated to identify claims
4	INSR_TYPE	NUMBER	The type of insurance provided i.e. 1605 workmen's compensation, 2001 personal accident, 5301 medical individual and 5302 medical group
5	COVER_TYPE	VARCHAR2	The covers under the specific insurance type provided. 'MEDICEXP' medical expense, 'EXTENDIL' extended illness ,'MEDICEXPGR' medical expense group, 'EXTENDILGR' extended illness group, 'HIHOSPSICK' hospitalization and sickness, 'HIMAT' Maternity benefit, 'HIPREGN' Pregnancy checkup, 'HIDENTAL' dental check-up, 'HIEYE' Eye glass benefit
6	RISK_TYPE	VARCHAR2	Specific risks covered under the insurance cover
7	LOSS_TYPE	VARCHAR2	1-property damage and 2- non-property damage
8	CLAIM_STATE	NUMBER	The different claim states i.e. 1-wait for documents, 2-Waiting to be evaluated by experts, 3- evaluated, 4-payment calculated, 5- for confirmation, 6-confirmed, 7- paid ,10-cancelled and 11-annuled
9	INITIAL_RESERV_AMNT	NUMBER	The amount hold initially for the claim
10	LAST_RESERV_AMNT	NUMBER	This is the amount left after claim payments are processed
11	LAST_RESERV_CURRENCY	VARCHAR2	currency
12	LAST_RESERV_DATE	DATE	This is the date which the reserve is adjusted for the last time
13	REGISTRATION_DATE	DATE	The date when the claim is registered
14	EVENT_DATE	DATE	The date when the insured is hospitalized
15	NOTIFICATION_DATE	DATE	The date when insured notifies the claim to the insurance company.

Table 4.2.:- description of attributes from CLAIM_OBJECTS table

No.	Attribute Name	Data type	Description
1	CLAIM_TYPE	VARCHAR2	The type of claim requested
2	OFFICE_ID	NUMBER	The branch offices responsible for the claim

Table 4.3.:- description of attributes from CLAIM table

No.	Attribute Name	Data type	Description
1	SEX	NUMBER	Gender of the insured person, '1' refers to male and '2' refers to male insureds
2	BIRTH_DATE	DATE	Birth date

Table 4.4.:- description of attributes from INSUREDS/O_ACCINSURED table

No.	Attribute Name	Data type	Description
1	OBJECT_TYPE	NUMBER	Is the insured registered as individual '1' or '8' group, this is for the non-life and for life the insured object is differentiated by the insurance type itself.

Table 4.5.:- description of attributes from INSURED_OBJECT table

No.	Attribute Name	Data type	Description
1	INSURED_VALUE	NUMBER	The amount in which the insurer is liable

Table 4.6.:- description of attributes from LIFE_RISK_COVERED/GEN_RISK_COVERED table

No.	Attribute Name	Data type	Description
1	DOCUMENT_STATE	NUMBER	State of the necessary claim documents 1- Not available, 2- Available , 3- Not expected

Table 4.7.:- description of attributes from CLAIM_AVAIL_DOCS table

When the collected raw data is seen from the point of view of data quality, it is believable because the systems are fully operational and it is real transactional data. But it has redundant attributes. As a result of the relationship of data tables most of the columns found in one table are also found in another table. Some attributes have missing values and the tables are a bit complex to understand their relationship. Apart from that records on the health service provider side are not captured in the system. Due to the reason that performers only encode data which are mainly related to premium computation, other fields may not be considered as important even if the field is available. This becomes a constraint to the analysis if fraud is initiated from the health provider.

4.3. Preparation of Data

The data collected originally comes from two different databases and 10 different tables. Quality issues like missing values, redundant attribute, unnecessary data values etc… are also observed in these records, which need to pass through the different stages of data preparation. Data preparation is constructing a final dataset from one or more data sources to be used for modeling [59]. This process is crucial in data mining. If it is not carefully analyzed it will finally produce misleading results. Data preparation tasks performed includes data cleaning, data integration, data transformation, and data reduction which are discussed in the next sections in detail.

4.3.1. Data Cleaning

Data cleaning step includes activities such as removing unnecessary data vales or attributes and filling missing values. From the data under workmen's compensation insurance type the risks are found to be compiled under one COVER_TYPE which is 'WORKMENSCOMP'. In this cover the different types of compensation (death, disability, and medical expense) are included. This makes it difficult to differentiate which of the claims belong to medical insurance. Therefore all 7015 records of this insurance type are excluded from the collected data.

To remove attributes first the columns with null values are all deleted then attributes like CLAIM_OBJ_SEQ and CLAIM_ID which identifies a specific claim are deleted. There are different

dates related to the claim, which are EVENT_DATE, NOTIFICATION_DATE, LAST_RESERV_DATE and REGISTRATION_DATE. Even if they describe different dates two of them are enough for the analysis, therefore EVENT_DATE and NOTIFICATION_DATE are taken and the others are deleted. These dates are taken because the experts suspect fraudulent claims if an insured delays to notify claim to the insurance company.

When insurance type refers to group, insured records like SEX and AGE are not applicable and in the raw data it contains null entries. Hence SEX field for those records is filled with 'NA' value which refers to the value is not applicable for the specific record. Regarding age for group of insured's it is difficult to fill a calculated age value for the attribute. This is because of the product nature, what is provided by the insured at the time of policy underwriting does not include detail for each individual. Therefore, it is not possible to estimate age from the existing data. This case is further elaborated in data reduction (attribute selection) step.

Age of the insured in case of individual insured is missing in 83 records and the method used to handle this case is deleting the records with missing age after calculating the percentage of missing from the total records (0.56%) and assuming it is not that much considerable.

When duration between 'EVENT_DATE' and 'NOTIFICATION_DATE' is calculated in data transformation, 113 records are found to have delay of notification more than a year and even more than four years in some cases. Experts involved assured the researcher that this can never happen in reality and they reason out that since the date lie in the beginning period of implementation of the software those records may be taken for test. Due to this it is better not to consider these cases for the study since they are outliers and may result in variation of prediction.

Another variable containing missing values is LIMIT for personal insurance product type. It is entertained by filling with the minimum limit that the corporation set for this category of

insurance which is 2000. But in the initial agreement with the insured it can be raised with additional premium calculated based on the increased amount of cover.

4.3.2. Data Integration

Data integration is needed since the data comes from different databases and different tables. The structure of the tables in both databases is almost the same but it needs rearrangements of values to bring them to one format. Another issue is redundancy, which occurs when the data from the difference sources is integrated attributes containing similar information are found like COVER_TYPE, RISK_TYPE and CLAIM_TYPE therefore COVER_TYPE is picked and the others are removed.

Non life insurance differentiates group and individual insurance by using the COVER_TYPE and the OBJECT_TYPE while life insurance differentiates by using the INSR_TYPE because field OBJECT_TYPE does not exist for life insurance records. Therefore, the values for this column in life data is filled with the corresponding object type from the non life category. That is for individual medical insurance (5301) OBJECT_TYPE=1 and group medical insurance (5302) OBJECT_TYPE=8.

Currency parameter is found to contain the currency of claim amount for both life and non life in local currency but stored as 'ETB' in non life and as 'BIR' in life category. Therefore it is needed to make the same for both and 'ETB' is replaced by 'BIR'.

Claim documents come from a separate table; one claim request can have different documents to be provided from the insured. One claim id is found to be repeated multiplied by the number of documents. To merge these, the maximum state is selected from claim object states grouping by CLAIM_ID and maps each claim with the state of document. Maximum value is selected because claim document state is 1(not available) from the beginning then when the document is delivered it will be updated to 2(available) and if the document is not needed for

the specific claim it will be updated to 3(not expected). A claim can be transferred to the next stage only if there is no document in state 1.

4.3.3. Data transformation

Attributes like EVENT_DATE and NOTIFICATION_DATE need to be reformatted. Date was previously formatted like '23-NOV-11 10.38.07.000000000 AM' and it is transformed to contain only date, month and year as '23-NOV-11'. Then a new attribute is derived from the two dates by taking the date difference between them and named as 'DELAY_TO_NOTIF'.

Two new attributes are newly added so as to make the data appropriate for data mining. These are CATEGORY to differentiate the original data source is life or Non-Life and FRAUD_SUSPECTED which takes 'YES' or 'NO' value. The data is labeled as 'YES' for those rejected cases and claim documents are fulfilled and 'NO' for other cases.

Another attribute 'CLAIM_AMOUNT' is derived from 'INITIAL_RESERV_AMNT' and 'LAST_RESERV_AMNT' by taking the absolute value of the difference of these two values. This is due to the reason that the focus is on the amount of claim cost incurred.

4.3.4. Data Reduction

Reduction includes reducing irrelevant and redundant variables which contain no useful information for the data mining task (dimensionality reduction) and reducing data values (numerosity reduction). Data reduction is helpful to obtain a reduced representation of the data. It is applied since making use of complex data for analysis takes long time [36].

The initial data collected after data cleaning steps becomes 29664 and then the number of the rejected cases is not balanced with the settled cases. Therefore data values of the processed cases should be reduced to make the final dataset balanced. For this purpose stratified random sampling is used because the number of records from Life and Non-life are almost equal. The data values are classified in to four strata; Life individual, Life group, Non Life individual and

Non Life Group. Then equal number of data is taken from the different groups by taking different offices into consideration. Finally 15021 records remain for training and testing.

The final data after the previous four steps of data preparation is converted to an arff file format which Weka can recognize by using excel to arff converter. Then the data is fed to Weka tool to perform attribute selection.

Weka attribute selection is used for dimensionality reduction. Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction [36]. Although we collected data with the help of domain experts, the selection of best attributes is critical because it makes difference in meaningfully model the problem [38]. Selection of Attribute evaluator and search methods depends on the given dataset [36]. Therefore, by applying different combination of attribute evaluation and search methods on the final dataset the attributes are ranked based on their usefulness and the result is analyzed with the domain experts. In this regard GainRatioAttributeEval attribute evaluator together with Ranker search method is selected to be used in this study.

Attribute selection is done in two steps. First, AGE attribute value for group medical insurance case is further analyzed for its usefulness. It has values for individual but a null value is assigned for group medical insurance. The data excluding the entire group INSURANCE_TYPE is given to the selected weka attribute selection algorithm and the result shows AGE is ranked as 12 with gain ration of 0 from the total of 14 attributes. That is this attribute is not in the list of selected attributes even in case of individual medical insurance. Therefore this column is removed from the entire dataset. Second, the total data set is passed to weka attribute selection from the result LOSS_TYPE and LAST_RESERV_CURRENCY are found to be useless to the problem. Gain ratio calculated is 0 for both attributes. Final remaining attributes from data preprocessing step are INSURED_ID, INSR_TYPE, COVER_TYPE, AMNT_REQUESTED, DELAY_TO_NOTIF, CLAIM_OFFICE, SEX, CATEGORY, OBJECT_TYPE, LIMIT, DOCUMENT_STATE, and FRAUD_SUSPECTED.

CHAPTER FIVE

5. EXPERIMENTAL RESULTS

In this chapter real data mining experiments conducted on the final dataset are described in detail. The research problem is to test data mining algorithms for detecting fraud suspected claims for the case of medical insurance. In this regard preprocessed data containing a total of 15020 records is prepared.

The data set is initially labeled with the assumption that rejected claims are fraudulent and all claims in process as well as settled are not. Therefore classification methods are used to build a predictive model. Data mining algorithms used for classification are Naïve Bayes and J48 classification algorithms which are explained in chapter three. Here is the step in knowledge discovery process where the real tasks of data mining or model building takes place to extract novel patterns hidden in the dataset. Each time the model build classification accuracy will be tested using 10 fold cross validation and a separate test dataset. The models build by the two algorithms are compared to each other from the point of view of evaluating criteria discussed in chapter three.

The experiment is done in three different scenarios. First with the full dataset and then with life insurance claim data and non-life insurance data separately to check if the prediction of fraudulent claims works better for each of the insurance categories.

The next step to be seen in this chapter would be interpretation and evaluation of the knowledge discovered from the data mining step. This phase includes filtering information to be presented by removing redundant or irrelevant patterns. In interpretation of results with the involvement of domain experts', conflicts of the newly discovered knowledge with the previous one are resolved and the models are translated to understandable forms.

Finally the prototype developed using the rules generated by the data mining algorithms is demonstrated.

5.1.1. Naïve Bayes Classification Model Building

Naïve Bayes classification algorithm works based on the following three conditions: The prior probability of a given hypothesis, the probability of the data given that a hypothesis, and the probability of the data itself. Its classification performance is based on the assumption of conditional independence between the attributes [56].

Weka data mining tool is used to run Bayesian classification with 10-fold cross validation test mode and percentage split test mode respectively. The experiments are done in three different scenarios: first full data set containing 15020 records is supplied and then the data is split in to life and non-life separately to see the classification performance of the algorithms in the two different categories. The results obtained are summarized and presented in tables 5.1 and 5.2.

Experiment -1-

Actual	**Predicted**		
Naïve Bayes classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**3363**	2684	6047
NO	3124	**5849**	8973
Naïve Bayes classifier with percentage split (66%) test mode			
YES	**1112**	926	2038
NO	979	**2090**	3069
Naïve Bayes classifier with percentage split (75%) test mode			
YES	763	725	1488
NO	724	**1543**	2267

Table 5.1: confusion matrix for Experiment 1, Naïve Bayes Classifier with different test options

	Result	Percentage	Time
Naïve Bayes classifier with 10fold cross validation test mode			
Correctly classified	9212	61.33 %	0.03 sec
Incorrectly classified	5808	38.67%	
Naïve Bayes classifier with percentage split (66%) test mode			
Correctly classified	3202	62.70%	0.03 sec
Incorrectly classified	1905	37.30%	
Naïve Bayes classifier with percentage split (75%) test mode			
Correctly classified	2306	61.41%	0.03s
Incorrectly classified	1449	38.59%	

Table 5.2: Classification results for Experiment 1, Naïve Bayes Classifier with different test options

From the results it can be seen that Bayes classifier with 66% percentage split has relatively better performance than the other two cases. 62.69% (3202) of the test data (5107) are classified correctly to the respective class fraud suspected or not.

Experiment-2-

Experiment two is for non-life records independently and the result is as depicted in tables 5.3 and 5.4 below

Actual	**Predicted**		
Naïve Bayes classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**1418**	1373	2791
NO	1228	**3394**	4622
Naïve Bayes classifier with percentage split (66%) test mode			
YES	**517**	463	980
NO	365	**1175**	1540
Naïve Bayes classifier with percentage split (75%) test mode			
YES	**357**	353	710
NO	272	**871**	1143

Table 5.3: confusion matrix for Experiment 2, Naïve Bayes Classifier with different test options

	Result	Percentage	Time
Naïve Bayes classifier with 10fold cross validation test mode			
Correctly classified	4812	64.91 %	0.06 sec
Incorrectly classified	2601	35.09%	
Naïve Bayes classifier with percentage split (66%) test mode			
Correctly classified	1692	67.14%	0.03 sec
Incorrectly classified	828	32.86%	
Naïve Bayes classifier with percentage split (75%) test mode			
Correctly classified	1228	66.27%	0.02s
Incorrectly classified	625	33.73%	

Table 5.4: classification results of Experiment 2, Naïve Bayes Classifier with different test options

It is still true with non-life data set that relatively better classification accuracy is seen when 66% percentage split is applied which is 67.14% of the test data are correctly classified.

Experiment-3-

Life data set is treated separately in this experiment and the results obtained are presented in table 5.5 and 5.6 below

Actual	**Predicted**		
Naïve Bayes classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**2274**	977	3251
NO	1465	**2886**	4351
Naïve Bayes classifier with percentage split (66%) test mode			
YES	**690**	413	1103
NO	421	**1061**	1482
Naïve Bayes classifier with percentage split (75%) test mode			
YES	**532**	275	807
NO	320	**773**	1093

Table 5.5: confusion matrix for Experiment 3, Naïve Bayes Classifier with different test options

	Result	Percentage	Time
Naïve Bayes classifier with 10fold cross validation test mode			
Correctly classified	5160	67.88%	0.02 sec
Incorrectly classified	2442	32.12%	
Naïve Bayes classifier with percentage split (66%) test mode			
Correctly classified	1751	67.73%	0.02 sec
Incorrectly classified	834	32.26%	
Naïve Bayes classifier with percentage split (75%) test mode			
Correctly classified	1305	68.68%	0.02s
Incorrectly classified	552	33.32%	

Table 5.6: Classification results of Experiment 3, Naïve Bayes Classifier with different test options

Here the accuracy of classification is better when applying 75% percentage split test mode which is 68.68%. The life category is seen to have better classified fraudulent claims than the non-life category.

5.1.2. J48 Decision Tree Model Building

As it is discussed in detail in chapter three decision tree classification model can clearly show the variables which are to be used for prediction of incoming data values. J48 is an algorithm used to build decision tree. The dataset is prepared and labeled with fraud suspected or not and this data is fed to the weka tool and j48 data mining algorithm is run in different scenarios.

J48 has different parameters (like confidenceFactor, minNumObj, reducedErrorPruning, Unpruned etc) that have initial default values and depending on data the values can be changed so that the classification accuracy could be increased. In this research also the algorithm is used with the default values and also tested by changing these values to see the change in the classification accuracy. The experiment is carried out accordingly and the result in each case is shown as follows

Experiment -4-

Here the full dataset containing 15020 records is passed to j48 decision tree algorithm of weka data mining tool with 10-fold cross validation test mode. In this case 84.01% of the data are correctly classified and 15.99 % are incorrectly classified. The algorithm generates a tree with size of 1542 and number of leaves is 905. It is difficult to generate a predictive rule from this tree by traversing all the nodes. Therefore in the same experiment it is tested by increasing minNumObj from the default value 2 to 5, 10, 15, 20, 25 and 30 of course the number of leaves and the size of the tree decreases to a better manageable size but the classification accuracy decreases in each case. Case when minNumObj =25 is selected.

Then to check if the classification accuracy is better when the percentage of test set is increased the test mode is switched in to percentage split first with the default value 66% (in

which 5107 of the records are used as test set and the remaining 9913 records are used for building the model) and then percentage is increased to 70% and 75% and the test result is of each of the cases the resulting confusion matrix is summarized and presented in the tables 5.7 and 5.8 below

Actual	**Predicted**		
J48 classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**4784**	1263	6047
NO	1138	**7835**	8973
J48 classifier with percentage split (66%) test mode			
YES	**1552**	486	2038
NO	450	**2619**	3069
J48 classifier with percentage split (75%) test mode			
YES	1145	343	1488
NO	275	**1992**	2267

Table 5.7: summary of confusion matrix for experiment 4, j48 model with full dataset

	Result	Percentage	Time
J48 classifier with 10fold cross validation test mode			
Correctly classified	12619	84.01 %	0.86 sec
Incorrectly classified	2401	15.99%	
J48 classifier with percentage split (66%) test mode			
Correctly classified	4171	81.67%	0.56 sec
Incorrectly classified	936	18.33%	
J48 classifier with percentage split (75%) test mode			
Correctly classified	3137	83.54%	0.58s
Incorrectly classified	618	16.46%	

Table 5.8: summary of classification results for experiment 4, j48 model with full dataset

From the above tables it can be seen that j48 classifier with 10-fold cross validation test mode has better classification accuracy of 84.01%. The data seems unbalanced and it has been tried to increase the classification accuracy by applying weka SMOTE. It is an over-sampling approach to handle skewed datasets [42]. It works by oversampling the minority classes by generating syntactic examples of them and adding to the dataset. This increases the probability of correctly classifying minority classes [42].

Therefore adding filter smote with 5 nearestNeighbours and 100 percent percentage of smote data amount generated the results depicted in tables 5.9 and 5.10.

Actual	**Predicted**		
J48 classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**10896**	1198	12094
NO	1690	**7283**	8973
J48 classifier with percentage split (66%) test mode			
YES	**3635**	478	4113
NO	628	**2422**	3050
J48 classifier with percentage split (75%) test mode			
YES	2735	302	3037
NO	468	**1762**	2230

Table 5.9: summary of confusion matrix for experiment 4, j48 model with full dataset and applying SMOTE

	Result	Percentage	Time
J48 classifier with 10fold cross validation test mode			
Correctly classified	18179	86.29 %	1.03 sec
Incorrectly classified	2888	13.71%	
J48 classifier with percentage split (66%) test mode			
Correctly classified	6057	84.56%	0.98 sec
Incorrectly classified	1106	15.44%	
J48 classifier with percentage split (75%) test mode			
Correctly classified	4497	85.38%	0.97 sec
Incorrectly classified	770	14.62%	

Table 5.10: Classification Results for experiment 4, j48 model with full dataset and applying SMOTE

From the results found in this experiment we can see that j48 classification with 10-fold cross validation has classified the data better than with percentage split. And of course a better accuracy of 86.29% is reached when smote is applied.

Experiment -5-

This experiment is continued by splitting the dataset into life and non-life to see the accuracy of the models separately in the two categories. The case treated in this experiment is the non-life data set. The results are depicted in tables 5.11 and 5.12.

Actual	**Predicted**		
J48 classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**2243**	548	2791
NO	448	**4174**	4622
J48 classifier with percentage split (66%) test mode			
YES	**765**	215	980
NO	159	**1381**	1540
J48 classifier with percentage split (75%) test mode			
YES	556	154	710
NO	108	**1035**	1143

Table 5.11: summary of confusion matrix for experiment 5, j48 model with non-life dataset

	Result	Percentage	Time
J48 classifier with 10fold cross validation test mode			
Correctly classified	6417	86.56 %	0.16sec
Incorrectly classified	996	13.44%	
J48 classifier with percentage split (66%) test mode			
Correctly classified	2146	85.16%	0.16 sec
Incorrectly classified	374	14.84%	
J48 classifier with percentage split (75%) test mode			
Correctly classified	1591	85.86%	0.56 sec
Incorrectly classified	262	14.14%	

Table 5.12: Classification results for experiment 5, j48 model with non-life dataset

The same is true with the non-life dataset. It is also tested after applying smote and a classification accuracy of 86.99% is achieved.

Experiment -6-

In this experiment life insurance dataset is separately treated. The results again are depicted in tables 5.13 and 5.14

Actual	**Predicted**		
J48 classifier with 10fold cross validation test mode	**Yes**	**No**	**Total**
YES	**2561**	690	3251
NO	732	**3619**	4351
J48 classifier with percentage split (66%) test mode			
YES	**834**	269	1103
NO	271	**1211**	1482
J48 classifier with percentage split (75%) test mode			
YES	628	179	807
NO	172	**921**	1093

Table 5.13: summary of confusion matrix for experiment 5, j48 model with life dataset

	Result	Percentage	Time
J48 classifier with 10fold cross validation test mode			
Correctly classified	6180	81.29 %	0.27sec
Incorrectly classified	1422	18.70%	
J48 classifier with percentage split (66%) test mode			
Correctly classified	2045	79.11%	0.19 sec
Incorrectly classified	540	20.89%	
J48 classifier with percentage split (75%) test mode			
Correctly classified	1549	81.53%	0.17 sec
Incorrectly classified	351	18.47%	

Table 5.14: Classification results for experiment 5, j48 model with life dataset

When smoothing is applied, a better classification accuracy of 85.72% is obtained. And it is seen that highest accuracy is found with non-life insurance data. In each of the above experiments normalization of numeric attributes is done to put the values in a range of 0 and 1 to entertain outlier data values but there is no significant difference in the results therefore the results after normalization are not considered.

Finally it is tested with discretization of numeric variables to put in range of values. This makes the tree more manageable to derive rules by traversing through it. But the classification accuracy is significantly decreased. For the full data set 71.8% of the records are correctly classified while it was 84.01% before. Hence the resulting model after discretization is not considered.

In both Naïve Bayes and J48 Decision tree classification models, a further test is conducted by supplying the INSURED_ID in addition to the other variables. Even if it is an id variable and it uniquely identifies an entry when the data is visualized using weka tool it shows that there are repeated INSURED_ID s found in the claim data. This experiment is done to see which types of clients are frequently claiming for medical insurance and which fall in the category of fraud suspected. From the test it has been further traced that most frequently claiming individuals

are classified in fraud suspected category. Insured individuals making claims frequently are also found their claim is processed even over the limit of insurance covered under specific policy.

5.1.3 Comparison of Naïve Bayes and J48 Decision Tree Models

One of the objectives of this research is to explore a proper data mining algorithm for the problem of detecting fraud suspected claims. From what the researcher found in literatures, the algorithms selected to perform well in such cases are Naïve Bayes and J48 decision tree algorithms. Experiments are conducted to verify these classification algorithms are also applicable to medical insurance claim fraud detection by applying them in similar data sets but with different scenarios as discussed in detail in sections 5.1.1 and 5.1.2. . Though they are workable, the classification accuracy is not as good as reported (99.96%) in [44] which is done in classifying motor insurance claim fraud detection. But it is better than that of the SVM classification of medical insurance fraud (67.3%) [21].

It is needed to compare the two algorithms applied in the data set of medical insurance claim fraud detection in this research in order to proceed with the development of prototype. The comparison based on classification accuracy and performance is summarized and shown in table 5.15 below.

Except in the case of life insurance data set which treated separately; in all of the experiments, better classification accuracy is achieved when 10-fold cross validation test mode is applied. Therefore the result of this experiment is taken for comparison.

Dataset	Classification model	Correctly classified	Misclassified	Better Classifier
Full data set	Naïve Bayes	9212(61.33 %)	5808(38.67%)	J48 Decision Tree
	J48 Decision Tree	12619(84.01 %)	2401(15.99%)	
Non-Life Dataset	Naïve Bayes	4812(64.91 %)	2601(35.09%)	J48 Decision Tree
	J48 Decision Tree	6417(86.56 %)	996(13.44%)	
Life Dataset	Naïve Bayes	5160(67.88%)	2442(32.12%)	J48 Decision Tree
	J48 Decision Tree	6180(81.29%)	1422(18.71%)	

Table 5.15: comparison of J48 and Naïve Bayes models

As it is shown in table 5.15 J48 decision tree has better classification accuracy in all the scenarios. It is also seen that from the J48 decision tree classification model results of the three different datasets, non-life dataset are classified better although the difference observed between the results is not significant. Therefore, we can use one and same model to both Non-Life and Life insurance categories and this could be the model built by full dataset which has 84.01% prediction accuracy.

In addition to the above metrics the time taken to build model can be used as a measure of performance of classification model. The time taken to build Naïve Bayes model is shorter than J48. Time is recognized when the data becomes larger and larger but still the difference is not significant.

Classification accuracy can be alternatively seen by calculating absolute error, squared error, Mean squared error, Relative absolute error and Relative squared error which can be taken directly from the result of the weka classification process. A screen shot of the results are attached as an appendix. Again the J48 models are also better in this regard.

5.2. Evaluation of Discovered Knowledge

Based on the experiment results, the J48 model built by the full dataset is taken as a working predictive model for the case of detecting fraud suspected medical insurance claims. J48 decision tree uses 7 of the total 11 variables supplied to build the tree. These variables are INSR_TYPE, COVER_TYPE, AMNT_REQUESTED, DELAY_TO_NOTIF, CLAIM_OFFICE, SEX, LIMIT and DOCUMENT_STATE. It takes DELAY_TO_NOTIFF (the number of days between date of examination and date of notification) as the first splitting variable. Due to this, it can give much information for claims to be fraud suspected or not.

Different rules are extracted from the decision tree developed by this model. The rules are presented as follows:

Rule [1] If DELAY_TO_NOTIF <=50 and CLAIM_OFFICE='HAWASSA' then FRAUD_SUSPECTED='YES'

Rule [2] If DELAY_TO_NOTIF <=50 and CLAIM_OFFICE!='HAWASSA' and AMNT_REQUESTED<=10.16 then FRAUD_SUSPECTED='YES'

Rule [3] If DELAY_TO_NOTIF <=50 and CLAIM_OFFICE!='HAWASSA' and AMNT_REQUESTED>10.16 and INSR_TYPE='MED_IND' and LIMIT >=12000 then FRAUD_SUSPECTED='YES'

Rule [4] If DELAY_TO_NOTIF <=2 and CLAIM_OFFICE!='HAWASSA' and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and LIMIT <2750 then FRAUD_SUSPECTED='YES'

Rule [5] If DELAY_TO_NOTIF <=2 and CLAIM_OFFICE NOT IN ('HAWASSA', 'SOUTH_ADDIS') and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and LIMIT >2750 and CLAIM_OFFICE then FRAUD_SUSPECTED='YES'

Rule [6] If DELAY_TO_NOTIF [2, 50] and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and CLAIM_OFFICE ='MEKELLE' then FRAUD_SUSPECTED='YES'

Rule [7] If DELAY_TO_NOTIF [2, 29] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT<=1800 then FRAUD_SUSPECTED='YES'

Rule [8] If DELAY_TO_NOTIF [2, 14] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT [1800,4500] and SEX='MALE' then FRAUD_SUSPECTED='YES'

Rule [9] If DELAY_TO_NOTIF [14,50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED [10.16,2778.60] and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT [2500,4500] then FRAUD_SUSPECTED='YES'

Rule [10] If DELAY_TO_NOTIF [40,50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT [1800,4500] AND then FRAUD_SUSPECTED='YES'

Rule [11] If DELAY_TO_NOTIF [2,50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>10.16 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT[4500 , 9000] then FRAUD_SUSPECTED='YES'

Rule [12] If DELAY_TO_NOTIF [2,50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>9183 and INSR_TYPE!='MED_IND' and COVER_TYPE!='EXTENDILL' and DOC_STATE='AVAILABLE' and LIMIT [9000,12000] then FRAUD_SUSPECTED='YES'

Rule [13] If DELAY_TO_NOTIF [2, 50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>10.16 and INSR_TYPE not in ('MED_IND','PER_ACC') and COVER_TYPE!='EXTENDILL' and DOC_STATE='NOT_AVAILABLE' then FRAUD_SUSPECTED='YES'

Rule [14] If DELAY_TO_NOTIF [2, 50] and CLAIM_OFFICE NOT IN ('HAWASSA', 'MEKELLE') and AMNT_REQUESTED>1802 and COVER_TYPE!= 'EXTENDILL' and DOC_STATE='NOT_AVAILABLE' and INSR_TYPE='PER_ACC' then FRAUD_SUSPECTED='YES'

Rule [15] If DELAY_TO_NOTIF [50,254] and CLAIM_OFFICE !='HAWASSA' and 'AMNT_REQUESTED'<=10.30 then FRAUD_SUSPECTED='YES'

Rule [16] If DELAY_TO_NOTIF [50,254] and AMNT_REQUESTED >4916.5 and CLAIM_OFFICE ='CENTRAL_ADDIS' then FRAUD_SUSPECTED='YES'

Rule [17] Rule -17- If DELAY_TO_NOTIF [50, 66] and CLAIM_OFFICE not in ('HAWASSA', 'CENTRAL_ADDIS') and 'AMNT_REQUESTED'> 4916.5 then FRAUD_SUSPECTED='YES'

Rule [18] If DELAY_TO_NOTIF >254 and CLAIM_OFFICE ='SOUTH_ADDIS' then FRAUD_SUSPECTED='YES'

The model showed new knowledge out of the expectation of the experts consulted for this study. These include:

1- Claims that are reported early are considered as fraud suspected than those with late reporting's.
2- Claims with small amounts are also found fraudulent.
3- It is also seen that in the different offices or regions the reasons to suspect claim as fraud are also different.

A discussion session is conducted with the experts again to clear up with the controversies of their previous expectation and the knowledge discovered after the evaluation of the developed data mining model.

The case why they consider late reported claims are suspicious to fraud is because they assume that clients need time to gather evidences to create a false claim and this assumption is also in line with the model build for motor insurance claim fraud detection by (Tariku, A. 2011). But it is not true for the case of medical insurance and agreed with experts that the claimant is in a difficult condition and may not be able to notify a medical expense claim within a short period and may need some time to recover from his condition. This shows that lately notified claims are real claims in this case.

The second novel knowledge discovered from the chosen model is about small amount of claim requests. As it is discussed with the experts in problem identification stage of this research they suspect claims as fraud when the claim amount is higher, the resulting model shows of course there are claims found fraudulent when the amount requested is greater than 9000 birr but also claims with amount less than 1800 birr and even less than 10.16 birr are found fraudulent. It is discussed with the experts and this case also been supported by a reason that customers claim with small amounts in different times and when the amount is small the experts do not go through further investigation and just pay the requested amount. This makes the customer

create fraudulent claims in different times within the insurance period. Therefore the expert shall consider investigation in small amounts of clams also. This should be supported by systems to count frequency of claims per individual insured.

The researcher also made a telephone discussion with the experts of the different branches how they made suspect claims to be fraudulent and their response was the same. That is late notification and higher claim amounts. But from the final model result the variables to suspect a claim as fraudulent differs in the different regions. Although this reason is discovered by a research conducted by (Bologa, R., et.al.(2013)) it is stated as individuals seek ways to fraud insurance claims using gaps in the legislations and expectations of experts in different countries, it is a new discovered knowledge to the EIC experts.

5.3. Medical Insurance Fraud Detection System Prototype

Prototype development is needed to show that the data mining model developed could be deployed. EIC shall organize data warehousing to make use of data mining applications. But for this research purpose a simple web application is developed based on the rules extracted from the model.

The prototype is done assuming that the claim data already recorded in the operational system can be used as an input to it and there is no need to capture the data again. This system could be integrated with the operational system and after experts encode the claim data in the operational system they will only fetch the claim information to this system through database link functionality. The user interface is only to search the stored data and see the validation result based on the rules set behind. The interface is shown in fig.5.1. And the source code is attached as an appendix.

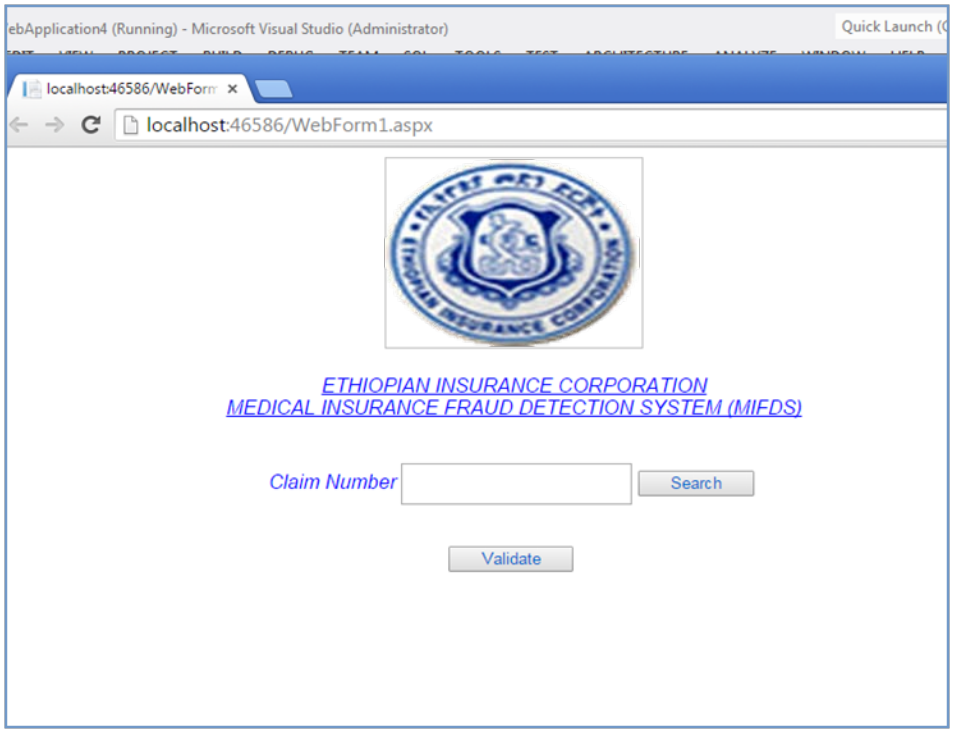


Fig 5.1. Initial user interface

The searching is done by specifying claim number of the claim to be validated and click the search button. Then the claim detail will be populated. When Validate button is pressed the system does the validation and returns the result. This is shown in fig 5.2 below.

Fig. 5.2. Screen shot user interface after the validation is done

5.4. Maintenance of the Model

Maintenance of models is required due to the reason that real time transaction data is changing every second when transactions are inserted, updated or deleted. These changes shall be considered that they can make the predictive model unworkable after some period of time [43]. The data mining research dimension is shifting towards real time data mining instead of updating a data warehouse at every end of day transactions bases.

The change in transactions holds true for case of insurance medical claims fraud predictive model built in this research. The data taken for building the model is of limited number of transactions for the purpose of the research. The resulting model is promising to be workable for the time being. However, when comes to real implementation, it shall be subject to maintenance every six months to one year period since the predictive variables may change through time.

CHAPTER SIX

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusion

Fraud detection is an important area in different sectors and specifically in insurance industries. Inability to discover fraudulent claims and put a solution to prevent them cost companies a lot and it is a critical problem for businesses. Developing a mechanism to entertain frauds shall be considered as one of their business strategies for insurance organizations.

As data becomes larger human experts could not be successful to investigate all the claims for fraud unless there is a mechanism to do so. It is time consuming to trace all the incoming claims and results in delay to entertain the genuine cases. Variables to indicate fraud are rapidly changing so it needs timely analysis and update of those patterns which are useful to differentiate between normal and abnormal operations.

This research is intended to check that data mining can help to differentiate valid medical insurance claims and show how to use the discovered knowledge for the case of Ethiopian Insurance Corporation. For this purpose data is collected from the Oracle database of INSIS system which the corporation is currently using to process the operational transactions. The final preprocessed data used for the experiment contains 15020 records and 11 variables. In order to conduct this research a six step hybrid data mining model developed by Cios et. al [8] is used.

Understanding and Preparation of the data took too much time of the research duration. The data collected initially is from the life insurance category and the non-life category further divided into group insurance and individual insurance covers the researcher faces a difficulty in merging the data in to one data set. There were missing values, values applicable for one cover are not applicable to others and also outliers were found in the data. Another difficulty was data regarding rejected claims of Life insurance category were not registered in the system.

These were collected from manual files and are merged with the date collected from the system.

The research is conducted by taking all the rejected cases with document available as fraud suspected and those in process and settled ones as valid claims by consulting the medical insurance claim experts of EIC. The data was labeled accordingly and Naïve Bayes and J48 decision tree classification algorithms are applied to see the results in different experiments in different scenarios by changing the default parameters and applying some filtering techniques like smote, attribute normalization and discretization to increase the classification accuracy. Classification accuracy is also checked by changing training and testing options from 10-fold cross validation to percentage split while the percentage value changed. WEKA data mining tool is used for this purpose. The result of the two algorithms is also compared and J48 decision tree is found to better classify the case.

It is shown that data mining algorithms can be used to predict medical insurance claim fraud; since resulting model gives 84.01% classification accuracy in which 12619 of the total records are classified correctly in to the respective fraud suspicious and non fraud suspicious categories and the rest 2401 or 15.99% are incorrectly classified. The classification pattern is shown by the rules generated with traversing from the root node along the leaves of the J48 decision tree. Finally, prototype is developed based on the rules to show how to apply the knowledge discovered.

Though it is done for academic research purpose and it is an initial work regarding the medical insurance fraud detection area in the case of EIC, it has positive results to be implemented for the intended purpose. The implementation of these techniques significantly increases quality and timeliness of detecting medical insurance claim fraud. Identifying fraud and valid claims easily will also optimize the time taken to handle the valid ones.

6.2. Recommendation

Classification accuracy mainly depends on the data mining application domain. All the influencing factors that can affect fraud in medical insurance should be considered to come up with high classification accuracy. The input parameters may only represent part of all information that could influence fraud. Other parameters may not be from the insured perspective; it could be from the insurance expert or from the health service provider side which both are not considered in this study.

To take advantage of data mining for further analysis health service provider data is needed to be available in the insurance data base. There is a variable in the data which is filled by a default value 'hospital' but from this study the researcher recommends that the specific hospital data should be recorded in the space provided as it is an one variable for predicting medical insurance fraud.

In group insurance type, two problems are observed in data handling of the insurance company. One: the insured sends name and salary of insured only at the time of claim otherwise for registering insurance policy only the total number of employees grouped by their occupation and their total salary is registered. When claim is notified there is no previously registered information about the individual claimant. This area is open to fraud medical insurance claims. Therefore, the researcher recommends the corporation to register the necessary individual insured data at the time of policy underwriting.

The second problem observed in group policies is even if there is a column to hold the claimant identification number this is filled with system generated serial number. Due to this reason every time a claimant comes he/she is registered as a new one and this is also open to fraud. It is recommended to register ID of the claimant to uniquely identify a specific person. In general the existing practice on group insurance type is open to make fraud either from the insured side

as well as the expert side. Therefore, it is also recommended to the corporation if this type of insurance delivery is stopped at some point.

Workmen's Compensation insurance type is excluded from the dataset because it doesn't have a separate risk type and cover type for medical expense. It is entertained together with other types of claims like death and disability. Therefore it is advisable to the corporation to make use of a different entry for different types of claims.

Furthermore the followings are recommended for further research:

- The classification accuracy achieved (84.01%) shows that some variables which give further information for detecting fraud are not available in the collected data. Causes of medical insurance fraud are not only from the customer side; it could be from the insurance expert itself or from the health care provider. Further research can be conducted including this information.
- From the different types of insurance products fraud detection research is conducted for motor insurance by Tariku (2011) and this research is conducted on medical insurance claim fraud detection. Therefore, other researches could be done focusing on the rest of insurance types.
- A research could also be done to test the prediction performance using other data mining algorithms like Support Vector Machine in the context of our country.
- The researcher has got information from literatures on how much is spent on fraudulent claims and how much of them are for valid claims in case of different countries, but there is no such information in Ethiopia. Therefore, the researcher also recommends an overall analysis could be one research area in this regard.

References

[1] ACL Services Ltd. (2010), *Fraud Detection Using Data Analytics in Insurance Industry*, Global Fraud Study: Report to the Nations on occupational Fraud and Abuse, ACL Services Ltd.

[2] Andoni, A., & Indyk, P. (2006, October). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on* (pp. 459-468). IEEE.

[3] Atkinson, P. M., & Tatnall, A. R. L. (1997). Introduction neural networks in remote sensing. *International Journal of remote sensing*, *18*(4), 699-709.

[4] Ayre, L. B. (2006). Data Mining for Information Professionals. *San Diego, California, USA*.

[5] Bagga, S., & Singh, G. N. (2012). Applications of Data Mining. *International Journal for Science and Emerging Technologies with Latest Trends, Vol. 1, No. 1, 19-23.*

[6] BOLOGA, R., & FLOREA, A. (2013). Big Data and Specific Analysis Methods for Insurance Fraud Detection. *Database Systems Journal, Vol. 4, No.4, 30-39.*

[7] Chapman, P Clinton, J Kerber, R Khabaza, T Reinartz, T Shearer, C and Wirth, R (2000). CRISP-DM 1.0 Step-by-step data mining guide.

[8] Cios, K Witold, P Roman, S and Kurgan A. (2007). *Data Mining A Knowledge Discovery Approach*, Springer

[9] Daniel, M. (2013). *Application of Data Mining Technology to support fraud Protection: The Case of Ethiopian Revenue and Customs Authority,* (Master's Thesis), Addis Ababa, Ethiopia: Department of Information Science, Addis Ababa University.

[10] Deshpande, S. P., & Thakare, D. V. (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS)*, *Vol. 1 No.1, 32-44.*

[11] Devale, A. B., & Kulkarni, R. V.(2012). Applications of Data Mining Techniques In Life Insurance, *International Journal of Data Mining and Knowledge Management Process Vol. 2, No. 4, 31-40*

[12] Dharminder, K. and Deepak, B. (2011). Rise of Data Mining: Current and Future Application Areas. *International Journal of Computer Science Issues* (IJCSI), *Vol. 8, No 1, 256-260*

[13] 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.

[14] Ekina, T., Leva, F., Ruggeri, F., & Soyer, R. (2013). Application of Bayesian Methods in Detection of Healthcare Fraud. *in Chemical Engineering Transaction*, *33*.

[15] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, vol. *17, No.* 3, 37.

[16] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD .Vol. 96, 82-88.*

[17] Gee, J., Button, M., Brooks, G., & VINCKE, P. (2010). The financial cost of healthcare fraud. *University of Portsmouth*

[18] Guo, L. (2003). Applying data mining techniques in property/casualty insurance. *in CAS 2003 Winter Forum, Data Management, Quality, and Technology Call Papers and Ratemaking Discussion Papers, CAS*.

[19] *Hardeep, K. (2014). Classification of data using New Enhanced Decision Tree Algorithm (NEDTA). International Journal of Emerging Technologies in Computational and Applied Sciences ,8(2)147-152*

[20] Han, J. & Kamber, M. (2006), *Data mining: Concepts and techniques,* 2nd edn, Morgan Kufman Publishers, San Francisco

[21] Kirlidog, M., & Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, *62, 989-994.*

[22] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, Vol. *19, No. 2*, 65.

[23] Neeraj, B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *International journal of advance research in computer science and software engineering*, *3*.

[24] Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (Vol. 2). New York: Springer.

[25] Mariscal, G., et al. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, Vol. 25, No.02, 137-166.*

[26] Martin, B. (2012). *Data mining techniques*, IBM Corporation, Chicago.

[27] Oracle Corporation 2008, Oracle Data Mining Concepts, 11*g* Release 1 (11.1).

[28] Patil, T. R., & Sherekar, M. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *Int J Comput Sci Appl*, *6*, 256-261.

[29] Patrick, L. et al. (2002). Fraud Classification Using Principal component Analysis of RIDITs. *The Journal of Risk and Insurance*, *Vol. 69, No. 3, 341-371*

[30] Periklis, A. (2002). Data clustering techniques. *Qualifying oral examination paper, University of Toronto, Toronto, Canada*.

[31] Pressman, R. S. (2005). *Software engineering: a practitioner's approach*. Palgrave Macmillan.

[32] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research.

[33] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106

[34] Rajanish, D. (2002). *Data Mining in Banking and Finance: A Note for Bankers*. Indian Institute of Management, Ahmadabad.

[35] Rekha, B. (2011). Detecting Auto Insurance Fraud by Data Mining Techniques. *Journal of Emerging trends in computing and information sciences, Vol. 2, No.4, 156-162.*

[36] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse, (2014): WEKA Manual for Version 3-6-12

[37] Revised Insurance Manual, Ethiopian Insurance Corporation , 2010

[38] Ritu G., Vijy C. (2013), A Comprehensive Study on Feature Selection Using Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, Vol .3 Issue 9, pp 26-33

[39] SAS Institute Inc. (2010). Combating Health Care Fraud: State-of-the-art methods for detection and prevention of fraud, waste and abuse in the health care industry, *SAS Institute Inc.*

[40] SAS Institute Inc. (2012). Data Mining from A-Z: Better Insights New opportunities, *SAS Institute Inc.*

[41] *Special Issue*, Ethiopian Insurance Corporation, 2001.

[42] Suman, S., Laddhad, K., & Deshmukh, U. (2005). Methods for Handling Highly Skewed Datasets. *Part I-October*, *3*.

[43] Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its applications* (Vol. 29). Springer Science & Business Media.

[44] Sun, Q. et al. (2014). *Detection of Health Insurance Fraud with Discrete Choice Model*: Evidence from Medical Expense Insurance, China. *SSRN 2459343*.

[45] *System Requirement Specification Document For Life Insurance,* Ethiopian Insurance Corporation, 2006.

[46] *System Requirement Specification Document For General Insurance*, Ethiopian Insurance Corporation, 2006.

[47] Tariku, A. (2011). *Mining Insurance Data for Fraud detection: The Case of Africa Insurance Share Company,* (Master's Thesis), Addis Ababa, Ethiopia: Department of Information Science, Addis Ababa University.

[48] Tesfaye, H. (2002). *Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment.* (Master's Thesis), Addis Ababa, Ethiopia: Department of Information Science, Addis Ababa University.

[49] The Commercial code of the Empire of Ethiopia, 1960.

[50] Two Crows Corporation 1999, *Introduction to Data Mining and Knowledge Discovery*, 3rd edn, Two Crows Corporation, Potomac: U.S.A.

[51] Umamaheswari, K., & Janakiraman, S. (2014). Role of Data mining in Insurance, *An international journal of advanced computer technology, Vol. 3, No. 6*

[52] Venkatadri, M., & Lokanatha, C. R. (2010). A Comparative Study on Decision Tree Classification Algorithms in Data Mining. *International Journal of Computer Applications in Engineering, Technology and Sciences*, *2*(2).

[53] Walter A. (2000). *Data Mining Industry: Emerging Trends and new opportunities.* (Master's Thesis), Massachusetts Institute of Technology.

[54] Woodfield, J. (2005), Predicting Workers Compensation Insurance Fraud using SAS Enterprise Miner 5.1 and SAS Text Miner, *Data Mining and Predictive Modeling , SAS Institute Inc. USA.*

[55] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H.,... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *vol. 14, No. 1, 1-37.*

[56] Zhang, H. (2004). The optimality of naive Bayes. *AA, Vol. 1, No.2, 3.*

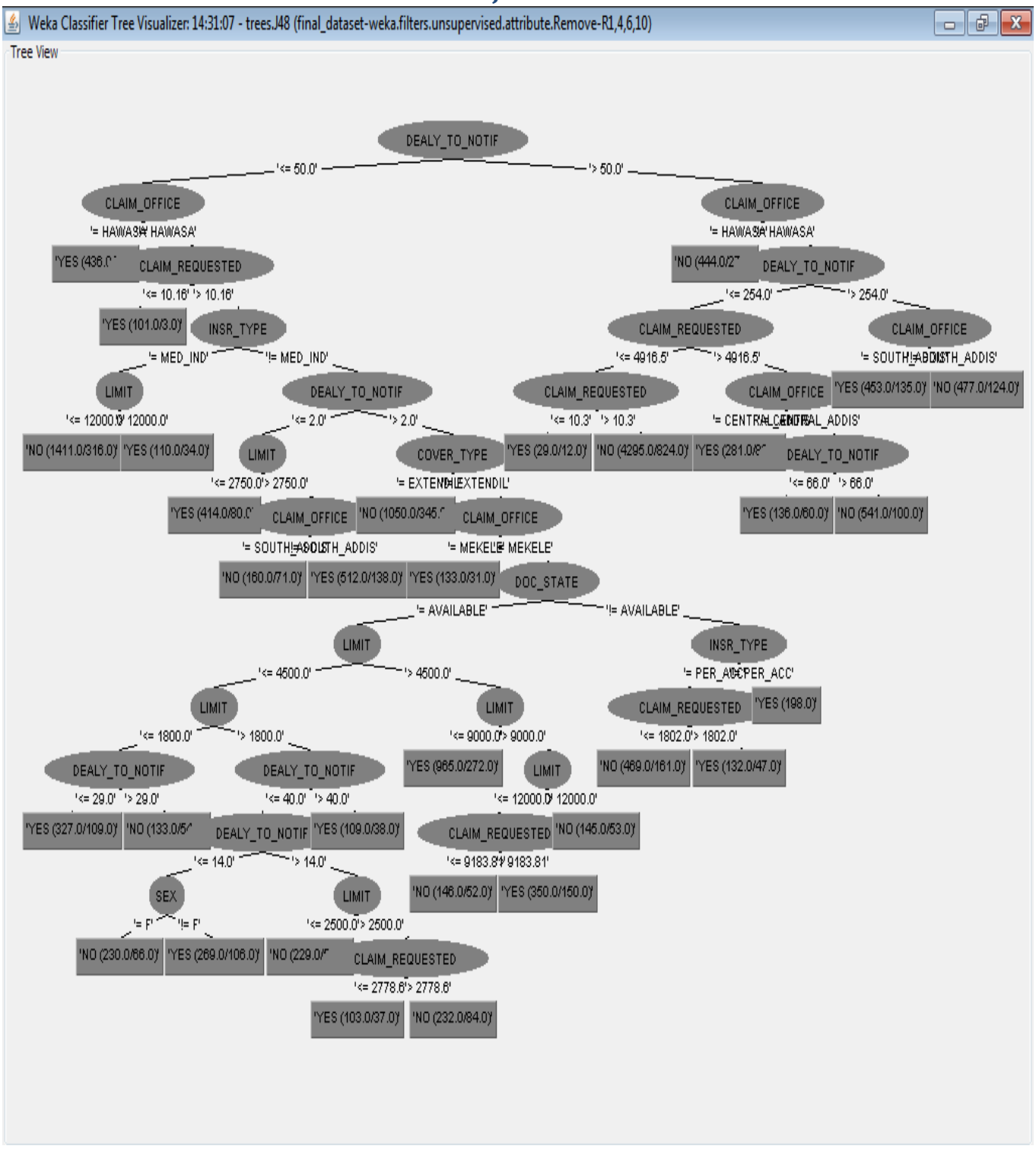
[57] Zhao, Q., & Bhowmick, S. S. (2003). Association rule mining: A survey. *Nan yang Technological University, Singapore*.

[58] Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, *41*(12), 1955-1959.

[59] Zubair K. (2014): A survey of data mining: concepts with applications and its future scope: *International journal of computer science trends and technology vol2 iss3*

[60] Zupan, B., & Demsar, J. (2008). Open-source tools for data mining. *Clinics in laboratory medicine*, *28*(1), 37-54.

APPENDIX -1- Screen shot taken from weka J48 classification of the selected model



APPENDIX -2- Sample taken from weka classification result of Naïve Bayes

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9212               61.3316 %
Incorrectly Classified Instances      5808               38.6684 %
Kappa statistic                          0.2055
Mean absolute error                      0.4591
Root mean squared error                  0.4975
Relative absolute error                 95.4346 %
Root relative squared error            101.448  %
Total Number of Instances            15020

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.556     0.348      0.518     0.556     0.537
0.616    YES
                 0.652     0.444      0.685     0.652     0.668
0.616    NO
Weighted Avg.    0.613     0.405      0.618     0.613     0.615
0.616

=== Confusion Matrix ===

    a    b   <-- classified as
 3363 2684 |    a = YES
 3124 5849 |    b = NO
```

APPENDIX -3- Sample taken from weka classification result of J48 Decision Tree

```
Number of Leaves  :   905

Size of the tree :    1542


Time taken to build model: 0.86 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12619               84.0146 %
Incorrectly Classified Instances      2401               15.9854 %
Kappa statistic                          0.6666
Mean absolute error                      0.2033
Root mean squared error                  0.3467
Relative absolute error                 42.264  %
Root relative squared error             70.699  %
Total Number of Instances            15020

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.791     0.127      0.808     0.791     0.799
0.901    YES
                 0.873     0.209      0.861     0.873     0.867
0.901    NO
Weighted Avg.    0.84      0.176      0.84      0.84      0.84
0.901

=== Confusion Matrix ===

    a    b   <-- classified as
 4784 1263 |    a = YES
 1138 7835 |    b = NO
```

APPENDIX -4- Sample code for implementing rules

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Web;
using System.Web.UI;
using System.Web.UI.WebControls;
using System.Data;
using System.Data.SqlClient;

namespace WebApplication4
{
    public partial class WebForm1 : System.Web.UI.Page
    {
        SqlConnection vid = new SqlConnection("Data Source=(local);Initial
Catalog=MIFD;Integrated Security=True");
        protected void Page_Load(object sender, EventArgs e)
        {

        }

        protected void Button1_Click(object sender, EventArgs e)
        {
            String Str = "Select * from CLAIM where (CLAIM_NO like '%' + @Button + '%')";

            SqlCommand xp = new SqlCommand(Str, vid);

            xp.Parameters.Add("@Button", SqlDbType.NChar).Value = TextBox1.Text;
            vid.Open();
            xp.ExecuteNonQuery();
            SqlDataAdapter da = new SqlDataAdapter();
            da.SelectCommand = xp;
            DataSet ds = new DataSet();
            da.Fill(ds, "CLAIM_NO");
            GridView1.DataSource = ds;
            GridView1.DataBind();
            vid.Close();

         }

        protected void Button2_Click(object sender, EventArgs e)
        {
            string claim_no;
            int insr_type;
            string cover_type;
            DateTime event_date;
            DateTime notification_date;
            int claim_amount;
            string sex;
            string claim_doc_state;
            int delay_to_notify;
            int limit;
```

```
            vid.Open();
            String Claim = "Select * from CLAIM where (CLAIM_NO like '%' + @Search +
'%')";

            SqlCommand comm = new SqlCommand(Claim, vid);
            comm.Parameters.Add("@Search", SqlDbType.NChar).Value = TextBox1.Text;
            SqlDataReader dr = comm.ExecuteReader();

            while (dr.Read())
            {
                claim_no = (dr["Claim_no"].ToString());
                insr_type = Convert.ToInt32(dr["Insr_type"]);
                cover_type = (dr["COVER_TYPE"].ToString());
                event_date = Convert.ToDateTime(dr["event_date"]);
                notification_date = Convert.ToDateTime(dr["notification_date"]);

                delay_to_notify = Convert.ToInt32(event_date.Day -
notification_date.Day);
                claim_amount = Convert.ToInt32(dr["claim_amount"]);
                sex = (dr["sex"]).ToString();
                claim_doc_state = (dr["claim_doc_state"]).ToString();
                limit = Convert.ToInt32(dr["limit"]);


                //rule-1
                if (delay_to_notify <= 50)
                {
                    if (claim_amount <= 10.16)
                    {
                        LabelMessage.Text = "This Claim Request is suspected to fraud! It
needs further investigation";
                    }
                    else
                     LabelMessage.Text = " This Claim is Valid! Please process!";
                }


                // Rule-2
                if (delay_to_notify <= 50)
                {
                    if (claim_amount > 10.16)
                    {
                        if (insr_type == 2001)
                        {
                            if (limit >= 12000)
                            {
                                LabelMessage.Text = "This Claim Request is suspected to
fraud! It needs further investigation";
                            }
                            else
                              LabelMessage.Text = "This Claim is Valid! Please process!";
                        }
                    }
```

```
                }

                //rule 3
                if (delay_to_notify <= 2)
                {
                    if (claim_amount > 10.16)
                    {
                        if (insr_type != 2001)
                        {
                            if (limit < 2750)
                            {
                               LabelMessage.Text = "This Claim Request is suspected to fraud!
It needs further investigation";

                            }
                            else
                              LabelMessage.Text = "This Claim is Valid! Please process!";

                        }
                    }
                }
            }
                vid.Close();

        }

        protected void GridView1_SelectedIndexChanged(object sender, EventArgs e)
        {

        }

            }
        }
```

APPENDIX -5- Interview Questions

1- What does the organization of the company look like?

2- How do you process medical insurance policy underwriting and claim processing?

3- How do you see medical insurance fraud in the case of Ethiopian Insurance Corporation?

4- How do you classify a medical insurance claim as fraud or non-fraud? And what parameters do you use currently to evaluate a claim as valid or not?

5- Please explain the data variables related to medical insurance claims.

DECLARATION

I declare that this thesis is my original work and has not been presented for a degree in any other university.

---------- ----------

Signature Date

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

Dereje Teferi (PhD)