2021-09

# Mining Community Based Health Insurance for Fraud Detection Using Data Mining Techniques

ZELALEM, MENGISTU

# BAHIR DAR UNIVERSITY

# BAHIR DAR INSTITUTE OF TECHNOLOGY

# SCHOOL OF RESEARCH AND POST GRADUATE STUDIES

# FACULTY OF COMPUTING

**MSC THESIS ON**

## Mining Community Based Health Insurance for Fraud Detection Using Data Mining Techniques

**By:**

**ZELALEM MENGISTU**

A thesis submitted In Partial fulfillment of the requirement for the degree of Master of Science in Information Technology

Advisor: Abrham Debasu /Assistant.Professor/

***September, 2021***

***Bahir Dar, Ethiopia***

# BAHIR DAR UNIVERSITY

# BAHIR DAR INSTITUTE OF TECHNOLOGY

# SCHOOL OF RESEARCH AND POST GRADUATE STUDIES

# FACULTY OF COMPUTING

## MSC THESIS ON

## Mining Community Based Health Insurance for Fraud Detection Using Data Mining Techniques

**By:**

**ZELALEM MENGISTU**

A thesis submitted In Partial fulfillment of the requirement for the degree of Master of Science in Information Technology

Advisor: Abrham Debasu /Assistant. Professor/

**September, 2021**

**Bahir Dar, Ethiopia**

## DECLARATION

This is to certify that the thesis entitled –**Mining Community Based Health Insurance for Fraud Detection Using Data Mining Techniques** submitted in partial fulfillment of the requirements for the degree of Master of Science in Information Technology under Faculty of Computing, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during the course of this investigation have been duly acknowledged.

| | | |
|---|---|---|
| **Zelalem Mengistu** | | 17/2/2014 E.C |
| Name of the candidate | signature | Date |

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

**Approval of thesis for defense result**

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of student ***Zelalem Mengistu wondem*** Signature ______ Date 17/2/2014 As members of the board of examiners, we examined this thesis entitled "**Mining Community Based Health Insurance for Fraud Detection Using Data Mining Techniques** " by *zelalem mengistu.* We here by certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of science in "Information Technology".

**Board of Examiners**

| | | |
|---|---|---|
| Name of Advisor | Signature | Date |
| Abrham Debasu | | 10/2/2014 E.C |
| Name of External examiner | Signature | Date |
| Dr. Dereje Teferi | | Oct 22, 2021 |
| Name of Internal Examiner | Signature | Date |
| Mekuanint A. (PhD) | | Oct 22, 2021 |
| Name of Chairperson | Signature | Date |
| Belete B. | | 26/10/2021 |
| Name of Chair Holder | Signature | Date |
| Derejaw L | | 26/10/2021 |
| Name of Faculty Dean | Signature | Date |
| Asegahegn E. | | oct 27, 2021 |

**Faculty Stamp**

ባሕር ዳር ዩኒቨርሲቲ ባሕር ዳር ቴክኖሎጂ ኢንስቲትዩት የኮምፒውቲንግ ፋካልቲ
Bahir Dar University
Bahir Dar Institute Of Technology Faculty Of Computing

# ACKNOWLEDGMENT

First and foremost, I would like to thank God. St. Mary, St. Michael thank you for your blessing and giving me the courage and wisdom to accomplish this thesis. I would also like to thank my Advisor, Abrham/Assistant Professor /for his excellent support, encouragement, and patience. I greatly appreciate your support from title selection to the proposal preparation and the delivery of the course research. Thank you so much for giving me what I required.

# ABSTRACT

*Community-based health insurance schemes (CBHIs) apply the principles of insurance to the social context of communities, guided by their preferences and based on their structures and arrangements. CBHIs can help communities manage health care costs and provide access to basic health care for the poor and other vulnerable groups.*

*Typically, CBHIs are organized and managed by a local community organization. The CBHI plan establishes agreements with various health providers, thereby forming a network of facilities. Most schemes cover basic health care services (e.g., antenatal care, deliveries, and child health care) and family planning services, while some schemes may also cover costs of hospital treatment.*

*Insurance fraud is an act that can be seen in different insurance types including CBHI insurance. Fraud in the case of CBHI is done by misrepresenting facts to get unauthorized benefit from the expenses covered under CBHI in the society. 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 CBHI claims in the case of west gojjam zone. A six step hybrid process model is used to guide the entire knowledge discovery process. J48 decision tree and Naive 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 93.06% 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.*

Table of Contents page

........................................................................ III

ACKNOWLEDGMENT.......................................................... IV

ABSTRACT.................................................................. V

LIST OF ABBREVIATIONS ................................................. IX

List of figures............................................................. X

List of table............................................................... xi

CHAPTER ONE INTRODUCTION ............................................ 1

1.1. Background of the Study ................................................ 1

1.1.1. Community Based Health Insurance (CBHI)........................ 1

1.1.2Community Based Health Insurance in Ethiopia...................... 2

1.1.3. Health Insurance Fraud ................................................ 3

1.2. Statement of the problem .............................................. 4

1.3. Research Question ........................................................ 5

1.4. Objectives of the Study .................................................. 5

1.4.1General Objective: ......................................................... 5

1.4.2. Specific Objectives: ....................................................... 5

1.5. Scope and Limitation of the Study .................................... 6

1.6. Significance of the study ................................................. 6

1.7. Organization of the Thesis ............................................... 6

CHAPTER TWO ................................................................ 8

2.1. Community-based health insurance schemes (CBHIs) ............ 8

2.1.1. Data mining concepts .................................................. 10

2.2. What is data mining? ...................................................... 11

2.3. Knowledge Discovery in Database (KDD) and Data Mining ...... 12

2.4. Data Mining Models ........................................................ 12

2.4.1. KDD Process Model ..................................................... 13

2.4.2. The CRISP-DM process model ........................................ 15

2.4.3. Hybrid model.............................................................. 18

2.5. Data Mining Tasks ........ 19
2.5.1 Supervised/predictive data mining methods ........ 20
2.5.2. Unsupervised/Descriptive data mining methodologies ........ 21
2.6. Predictive Modeling and Classification ........ 24
2.7. Data Mining in Health care ........ 27
2.8. Role of data mining Community-Based Health Insurance ........ 30
2.9. Fraud in Health Insurance ........ 31
2.10. Related work ........ 33
2.11. Data mining methods for fraud detection ........ 37
**CHAPTER THREE** ........ 39
**RESEARCH DESIGN AND METHODOLOGY** ........ 39
3.1 Research design ........ 41
3.2. Source of Data Set and Collection Method ........ 42
3.3. Architecture of proposed model ........ 42
3.4 CBHI Data preparation ........ 47
3.5. Medical service claim process ........ 48
3.6. Data Analysis tools ........ 51
3.6.1. Naïve Bayes Classification Technique ........ 52
**3.6.2. Decision tree classification technique** ........ 54
3.7. Evaluation methods ........ 57
**Chapter four** ........ 60
**4. Results and discussion** ........ 60
4.1. Development and testing environment ........ 60
4.2. Experimental Design ........ 60
4.2. Data preparation, filtering, and selection ........ 60
4.3. Classifiers performance ........ 62
4.2.1. Naïve Bayes model building ........ 63
4.4. Graphical User Interface Development ........ 69
**CHAPTER FIVE** ........ 71

**5. Conclusion and Recommendation** .......................................................................................... 71

**5.1. Conclusion** .......................................................................................................... 71

5.2. Recommendation......................................................................................................... 72

References ........................................................................................................................ 73

Appendix ........................................................................................................................... 78

Appendix B ........................................................................................................................ 81

## LIST OF ABBREVIATIONS

CBHI: Community Based Health Insurance

CHW: Community Health Workers initiative

CRISP-DM: Cross Industry Standard Process for Data Mining

DM: Data Mining

FMOH: Federal Ministry of Health

HEP: Health Extension Program

ID3: Iterative Dichotomiser3

KDD: Knowledge Discovery in Database

SEMMA: Sample Explore Modify Model Assess

WEKA: Waika to Environment for Knowledge Analysis

# List of figures

Figure 1:-Knowledge discovery steps (Source: Fayyad,et al [11].) .................................................. 14

Figure 2:-Cross-Industry Standard Process for Data Mining (CRISP-DM) .......................................... 16

Figure 3:-Original and duplicate claims made by the same patient ................................................... 22

Figure 4:-Classifying the claims ........................................................................................................ 23

Figure 5:-classification succeeds over clustering (outlier)................................................................. 23

Figure 6:-Process of data mining in health care ................................................................................. 29

Figure 7:-1Architecture of proposed model ........................................................................................ 44

Figure 8:-Weka explorer with CBHI training dataset ......................................................................... 62

Figure 9:-Snapshot of decision tree from experiments done in weka ................................................ 69

Figure 10:-Figure 10:-Prototype implementation diagram................................................................. 70

# List of table

Table 1:list of selected attributes and their converted form ........ 49

Table 2: Discretization of the age manually ........ 50

Table 3:-A confusion matrix for positive and negative entries ........ 57

Table 4:-Confusion matrix for experiment 1, naïve Bayes classifier with test options ........ 64

Table 5:-Classification results for experiment 1, naïve Bayes classifier with different test options .. 64

Table 6:- Default variables for J48 Tree algorithm ........ 66

Table 7:- Summary of Classification results for experiment 2, J48 classifier with different test options ........ 67

Table 8:- Comparison of naïve Bayes classifier with j48 classifier ........ 68

# CHAPTER ONE INTRODUCTION

## 1.1. Background of the Study

The Insurance Industry has historically been a growing organization .It plays an important role in insuring the financial benefit of one country. Insurance is a contract that provides safety to an individual or entity against a loss. This safety guarantees reimbursement from an insurance company. There are different types of Insurance systems found in the world including our country. Insurance is a contract between you and an insurance provider called a policy that allows you to be compensated for certain losses. According to the Insurance Information Institute of America, fraud accounts for around 10% of the property/target insurance industry's incurred losses and loss adjustment expenses each year, while this amount varies depending on line of business, economic conditions, and other factors. Also, the US Federal Bureau of Investigation said that health care fraud, both private and public, is estimated to be 3 to 10 percent of total health care expenditures (Kirlidog & Asuk, 2012). The term health insurance is a form of insurance that pays for medical expenses. Most health insurance policies offer particular benefits, and health insurance fraud schemes like over billing for the services received deprive consumers of these advantages. Health insurance is one of the most common and widely used types of insurance in the United States. A health insurance system is an institute that provides health care services to meet the health needs of the target population. It is a type of insurance that covers the price of medical services. If you have health insurance, you pay a yearly fee to a health insurance company, and if you have an accident or need surgery or a medical diagnostic, the health insurance company will cover your medical bills. There are four types of Health Insurance, these are National Health Insurance, Social Health Insurance, Private Voluntary Insurance-commercial and Community based Health Insurance(Jeffrey, 2007) . The study is a primary focus on Community based health insurance system as a case study.

### 1.1.1. Community Based Health Insurance (CBHI)

CBHI (Community-Based Health Insurance) is a concept for ensuring financial security against the costs of disease while also boosting access to high- quality health services for low-income rural households who are not covered by formal insurance (Yilma et al, 2015).CBHI is currently being offered in various developing country rural regions, and research into its influence on the well-Being of the poor in these locations is ongoing. The effectiveness of Community Based Health Insurance, on the other hand, is dependent on the presence of socialcapital in the community. As a result, academics have been looking into the influence of CBHI on the well-being of the poor in rural areas, particularly in terms of social capital. Concerning the impact of social capital on the demand forCBHI, previous research recommends that social capital is a vital asset that contributes not only to success but also to the increased demand and community's readiness to pay for community-based health insurance. This is because social capital facilitates the effective functioning and sustainability of CBHI. Different studies have also reported that people are more willing to pay for CBHI in communities that have sufficient stocks of social capital(V Rawte & Anuradha, 2015).

Health Insurance provided by the Community Is a non-profit health insurance program founded on a mutual-aid ethic among persons in the informal sector and in rural areas. Members of community-based health insurance combine their premium payments into a self-managed communal fund. CBHI is limited in most developing countries and out-of- pocket health care expenditure still makes poor households especially in rural and people in the informal sector.

### 1.1.2Community Based Health Insurance in Ethiopia

The Community Based Health care Insurance (CBHI) strategy in Ethiopia could be considered to originate from an ancient Ethiopian traditional practice known as the Idir (Tariku, 2011). The Idir was a long-standing traditional financial institution established by community members to support local finances and raise funds for crises such as group and family deaths. While the Idir method responds to unplanned events like deaths, CBHI in Ethiopia tries to prevent these unplanned events by changing these funds into a life saver rather than a celebration service fund.

Accordingly, the initiative was launched as the first of its kind in 2011 across 13 woredas (districts) within the four main regions of the country (V Rawte & Anuradha, 2015). Before this, the Ethiopian Federal Ministry of Health (FMOH) tried to pursue a rather vigorous policy on health since 1993, which included the construction of primary health facilities, as well as the development of the human and technical resources required to run them; these human resources involved providing adequate training for personnel(Melih& Cuneyt, 2012). Community-based health insurance aims that one can access the best health care without fearing the financial strain; it helps people to have in mind rather than to have fear.

The CBHI initiative was established in Ethiopia as a community-based health project that collects fees from members into a fund that covers basic health care expenditures, allowing members to visit local health care clinics whenever they are ill. CBHI is the result of the Ethiopian Federal Ministry of Health's (FMOH) efforts to provide total financial protection for health care in order to achieve universal health coverage. The Ethiopian health care system was largely dependent on out-of-pocket spending, baring many households to financial hardship due to very expensive health expenditures or causing them to give up seeking health care, especially in rural Ethiopia.

The amount of data held in insurance databases is continually expanding due to significant advancements in information technology. These massive databases have a plethora of information and could be a lucrative source of useful business data.

### 1.1.3. Health Insurance Fraud

False or misleading information is submitted to a health insurance company in order to have them pay unauthorized benefits to the policyholder, another party, or the entity providing services in this sort of fraud (Melih & Cuneyt, 2012).The insured person or the provider of health care can both be at fault. Fraud and abuse can be found across the health care system. Scams involving doctors, hospitals, nursing homes, diagnostic facilities, medical equipment providers, and advocates have been reported.

A single subscriber can defraud a health insurance company by

- allowing someone else to acquire health care services using his or her name and insurance information
- Using benefits to pay for prescriptions that his or her doctor did not prescribe

Health care providers can commit fraudulent acts by:

- Services, procedures, and/or goods that were never performed are billed.
- pricing for services that are more expensive than those already offered
- providing unneeded services for monetary gain
- treating non–covered treatments as though they were medically necessary
- misrepresenting a patient's diagnosis in order to justify tests, surgeries, or other treatments

- Each phase of a single procedure should be billed as if it were a distinct procedure
- within the terms of the insurance policy, charging a patient more than the agreed-upon co-pay
- paying "kickbacks" for referring motor vehicle accident victims to medical care
- Asking payments for one patient many times/for several districts.

## 1.2. Statement of the problem

Purposefully cheating the health insurance company that results in health care profits being funded criminally to an individual or group is known as health insurance fraud (Kirlidog & Asuk, 2012).Over the years Health Insurance Industries receives millions of claims from health care providers. Out of the millions of claims forwarded to health Insurance industries, a small percentage of the claims are frauds and large percentages cover non-fraud. Health insurance is a cash-strapped industry with a high claims ratio(Osmar , 2008) .As a result, if the health insurance sector is to be free of fraud, it must focus on the eradication or reduction of false claims received through health insurance. The health insurance sector, which includes community-based health insurance, amassed a massive amount of data about their clients or members. To extract information from the whole amount of raw data manually, Insurance companies lost time and effort.

The data extracting process resulted in developing new products and services to meet customers‘needs. In this study, West gojjam Zone Community Based Health Insurance is considered as a case for prediction/detection of frauds. In the West gojjam zone, there are 16 districts and 6 City/Town Administration CBHI Schemes, and a total of 22 CBHI schemes. Today different medical treatment is encountered in the health care and some medical treatments are free or available through government and third party treatments. However, claims for medical services for community-based health insurance are part of the cost, including free or government-funded treatments, and one health care/hospital is getting more paychecks and defrauding the health insurance and the government.

In the 2018/2019 G.C west gojjam zone the numbers of patients who got medical care or visited patients are 289,642 and the amount of split payment is 11,720,449 birr paid by different communities based health insurance schemes. It is difficult to know whether the claim cost is fraud or non-fraud/valid because the auditing system is poor. Therefore, the purpose of this study is to analyze such large documents using data mining documents.

## 1.3. Research Question

1. How to discover the hidden knowledge and patterns in the insurance claims data?
2. What are the determinant factors used to classify a given claim as fraud or valid?
3. Which data mining algorithm is appropriate to develop/construct a predictive model for categorizing medical insurance claims as valid and Frauds?

## 1.4. Objectives of the Study

### 1.4.1General Objective:

The general objective of this research is to develop predictive model for community- based health insurance fraud detection.

### 1.4.2. Specific Objectives:

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

- To identify the medical service included and not included under community-based health insurance and applying the data preparation, data cleaning, and prepossessing.
- Identifying the suitable algorithms to develop a predictive model to classify the claims.
- To review literature related to health insurance fraud
- Developing the predictive model for identifying the medical claim as –fraud and Non-fraud/Valid.
- Preparing a data set and experimenting with the developed predictive model.
- Evaluate the model
- Develop prototype to classify the medical claim as fraud and non-fraud.

## 1.5. Scope and Limitation of the Study

The main aim of this research is to develop and explore the prediction model for the prediction and detection of fraud in the health insurance sector. This research focuses not only on the medical claim but also includes cheating by health care and hospital financial departments. The focus of the research is included only the Community based health insurance scheme. The limitation of the study may include the scarcity of time, internet, etc.

## 1.6. Significance of the study

By providing insights and recommendations on how to design, build, and implement prediction/detection systems employing data mining technologies, the research brings value to the Community Based Health Insurance fraud domain. The research aims to reduce costs in health care thus, in turn, enables better health care that is accessible. When fully applied the prototype enables insurers to identify the performer of fraud and subject them to legal process. A system to detect frauds in insurance helps the insurance company to minimize unreasonably paid insurance claim costs as well as reduce investigation costs. Moreover, it improves the claim handling efficiency of the claim officers as they can differentiate between fraudulent and valid claims. Hence officers can handle the valid ones at the appropriate time.

## 1.7. Organization of the Thesis

The study is organized into five chapters. The first chapter deals with the background of the study, which introduces data mining applications in health care, a statement of the problem, objective, scope, and contribution of the study. The second chapter discusses works of literature to be reviewed, which briefly discusses data mining applications in the health care domain, data mining techniques to be used in the study domain, and mining algorithms applied in the selected data set samples. 6 The third chapter mainly focuses on the research methodology; how the research was conducted, including what procedures are followed to understand the problem, collect, and analyze the data, tools to be used, and algorithms applied in the study. The fourth chapter discusses the detailed description of the data, model developments, and performance measures of the developed models; such attempted to show the task to be done to generate a better quality data set ready to apply data mining tools and techniques. Therefore, prepossessing tasks including data cleaning, transformation, and attribute selection are discussed. And also presents the experimentation was done, performance evaluation, and the analysis of the result using selected hybrid techniques in data mining with selected algorithms. The last chapter focuses on making conclusions and recommendations to show further research directions in the future, which follows references that are cited by the researcher.

# CHAPTER TWO

## 2. Literature Review

All types of insurance, including health insurance, are vulnerable to fraud. Intentional deception or misrepresentation for the purpose of obtaining a substandard health benefit is health insurance fraud. In massive amounts of insurance claim data, data mining technologies and procedures can be utilized to detect fraud. Anomaly detection is one of the most used data mining approaches for locating fake records. This method is used to identify outlines or anomalies that depart from the norm. An abnormality in a computer network's data traffic pattern, for example, could indicate that a hacked machine is sending out sensitive data, and an aberration in an MR picture could indicate the presence of a malignant tumor (Kumar, 2005). (Spence et al., 2001). Anomaly analyses are used for fraud detection in a variety of domains outside of the insurance industry, including credit cards, mobile phones, and insider training monitoring (Chandola et al., 2009).Data mining systems, unlike most other computer software, do not predict the occurrence of an event with mathematical precision. The anomaly detection technique assesses the possibility or probability of each record being fraudulent by examining historical insurance claims based on a few examples that are known or suspected to be false.

Analysts can then look into the cases that have been flagged by data mining technologies in further depth. Patterns in a vast volume of data are discovered using data mining techniques. Data mining is a subset of knowledge discovery that entails using statistical, mathematical, artificial intelligence, and machine learning approaches to extract and find usable data and knowledge from big databases (Turban et al., 2005). Specific algorithms are used in a variety of data mining functions (Bigus, 1996).

### 2.1. Community-based health insurance schemes (CBHIs)

Community-based health insurance plans (CBHIs) apply insurance ideas to the social context of communities, driven by their decisions and based on their structures and arrangements. Community-based health information systems (CBHIs) can help communities manage health care costs and ensure that the poor and other vulnerable groups have access to basic health care. The plans are particularly beneficial in reaching rural inhabitants and the informal sector, which includes self-employed people and is not easily insured (e.g., farmers, petty traders, and laborers).These people are frequently unable to pay for basic health care at the point of treatment, which, if unabated, could push them into poverty.

Typically, CBHIs are organized and managed by a local community organization. The CBHI plan creates a network of facilities by developing agreements with various health providers. Most plans cover basic health care (such as prenatal care, births, and infant health care) as well as family planning, with some plans also covering hospitalization expenditures. CBHIs are valuable because they include community members as enrolls and volunteers. By pooling resources and augmenting them with external funds, ensure that health services fulfill community needs and that primary health care is accessible and affordable to members. The fact that CBHIs are frequently reliant on government and donor funding is one of their major flaws. Because such plans typically cover a small, low-income group of people, they don't have a large enough risk pool to support their operational costs.

Because most registrants are poor and cannot afford high premiums, premium payments and municipal subsidies are frequently insufficient to cover the costs of health care. Furthermore, while community involvement is good to CBHIs, it can be unsuccessful at times due to poor management and technical abilities of community service providers inside the CBHI structure. CBHIs should be part of a larger package of finance mechanisms that extend health care to under served populations, such as fee exemption schemes, equity funds and vouchers for beneficiaries, and results-based financing. Although CBHIs may not be appropriate in all instances, they can make a significant contribution to health care initiatives.

Valuable Insights on Implementing CBHIs

- ✓ Planning and budgeting. An overall design strategy is insufficient, as it needs to be tailored to each area. Prior to execution, an operational strategy and a realistic budget for the relevant schemes should be prepared.

- ✓ Balancing income and costs. Given the necessity to subsidize health care for low-income individuals, policymakers, planners, and implementer must pay particular attention to finance factors such as premium rates, fixed service delivery costs, and measures to diversify income sources.

- ✓ Engaging local communities. CBHIs rely on community members not only to maintain enrolment rates, but also to help run programs, monitor implementation, and educate people about health issues.

- ✓ Building a service network. Large CBHIs provide a variety of health services through a network of public and private providers. When private providers are involved and regulated to assure the provision of high-quality care, they can provide cost-effective services. Payment arrangements must be carefully considered, and transparency and accountability must be guaranteed.
- ✓ Assessing progress. Managers and funder of the CBHI must conduct a baseline survey, examine service delivery and cost data on a regular basis, and conduct periodic assessments of the program's strengths and weaknesses. Impact assessments are also recommended.

### 2.1.1. Data mining concepts

In today's world, there is a large amount of data saved in real-world databases, and that number is rapidly increasing. As a result, semi-automatic approaches for discovering hidden knowledge in such databases are required. Data mining is the process of mechanically sifting through massive volumes of data in order to uncover previously unknown patterns, generate new insights, and make predictions. The expansion of large databases has been fueled by advancements in digital data capture and storage technology. This can be seen in different sectors; for instance, supermarket transnational data, telephone call details, different governmental statistics, credit card records, different medical records. These days interest has grown in the possibility of extracting information from the databases that might be of value to the owner of the database. The discipline concerned with the task has become known as Data Mining (Mirchaye, 2015). Also, the methods and tools for data collecting, storing, and transferring for different purposes have increased. Massive volume of stored data is necessary to be collect and extracted and suitable for gaining information and knowledge i.e. they have value when extracting efficiently whenever the user needs (Kirlidog & Asuk, 2012).This raises the demand for new tools and techniques which help analyze the massive data for information and knowledge.

This leads to the idea of data mining; Data mining is often set in the broader context of knowledge discovery in databases, or KDD. According to (Yihenew, 2015) the KDD process contains several stages; selecting the target data, prepossessing the data, transforming the data, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures. It is estimated that the amount of data stored in the world's database raises every twenty months at a rate of 100 %( Fayyad & P, 1996). As the size of data grows, the proportion of information in which people could understand decreases considerably. This tells that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results in an increasing information gap. Consequently, people begin to realize this bottleneck and to look into possible remedies.

## 2.2. What is data mining?

Data Mining is a concept which is given different meaning by different researchers. The practice of examining data from many perspectives and summarizing the results into valuable information is known as data mining. According to Fayyad and Piatetsky (Fayyad & P, 1996) data mining is a process of non-trivial extraction of implicit, previously unknown, potentially useful, and actionable information (such as knowledge rules, constraints, regularities) from huge amounts of data in databases. This information enables us to make a serious business decision. Data mining is the computing process of discovering patterns, hidden information, and unknown data, relationships, and knowledge in large data sets. It is a confluence to machine learning, statistics, Artificial Intelligence, databases, and others. It requires analyses of large-scale data, which cannot be handled by traditional statistical methods. It is a crucial phase in the whole process of information discovery in databases, in which intelligent approaches are used to extract patterns (Anagaw et al., 2013).

## 2.2. Data Mining and Machine Learning

Data mining and Machine learning are hot research areas in a field of computer science whose quick development is due to the advances in data analysis research, growth in the database industry, and the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores. Machine learning is an interdisciplinary field of data mining that studies computer algorithms that improve themselves automatically over time (Cios et al., 2007).

Accordingly, Data mining is the process of identifying relationships in a vast and complicated set of data using machine learning algorithms. Data mining is used to extract information from raw data in databases that is expressed in an intelligible manner and may be used for a number of applications, while machine learning provides the technical foundation (Lakshmi, 2013). For instance, the main goal of both data mining and machine learning technologies is learning from data.

Data mining techniques tend to learn models from data. There are two approaches to learning the data mining models. Those are supervised learning, unsupervised learning as mentioned in the next section.

## 2.3. Knowledge Discovery in Database (KDD) and Data Mining

Conventionally the idea of searching relevant patterns in data has been referred to using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms, KDD and data mining are used widely [11].

Knowledge extraction (also known as data/pattern analysis, data archeology, data dredging, information harvesting, and business intelligence) is the automatic or semi- automatic extraction of interesting patterns or knowledge from a large amount of data stored in multiple data sources such as file systems, databases, and data warehouses. KDD is also defined by (Wirth & Hipp, 2000) as the whole process of discovering meaningful knowledge from data, with data mining referring to a specific component in this process.

Hence, Data Mining refers to a set of algorithms for extracting usable meaning from stored data, whereas Knowledge Discovery in a Database refers to the total process of discovering knowledge and incorporates Data Mining as one phase among many.

## 2.4. Data Mining Models

Knowledge Discovery and Data Mining (KDDM) process models are a set of processing processes that practitioners should follow when completing KDDM projects (Joudaki et al., 2015). The methods that are conducted in each of the steps are described in this model, which is primarily used to plan, work through, and lower the cost of any given project. Fayyad and colleagues proposed the model's basic framework. Since then, a number of other KDDM models have been developed in academia and industry.

All process models have several phases that are performed in a sequential order, which frequently involves loops and iterations. Each succeeding step is started when the preceding one has been completed successfully, and it requires a result created by the previous step as one of its inputs.

Another common feature of the proposed models is the span of covered activities; it ranges from the task of understanding the project domain and data through data preparation and analysis, to evaluation, understanding, and application of the generated results.

Consequently, rules, patterns, categorization models, relationships, trends, statistical analysis, and other terminology are commonly used to characterize the newly created information. The following models are common industrial, hybrid, and academic techniques used in DM research.

### 2.4.1. KDD Process Model

Researchers began defining muti step procedures to guide users of DM tools in the complex knowledge discovery world when the DM field was forming in the mid-1990s; researchers began defining muti step procedures to guide users of DM tools in the complex knowledge discovery world when the DM field was forming in the mid-1990s (Ozgul, 2013).

According to Fayyad (Fayyad & P, 1996), the KDD process is the process of employing DM approaches to extract what is deemed knowledge based on the specification of measures and thresholds; using a database and any necessary prepossessing, sub sampling, and database transformation. Accordingly, the KDD process model was the leading academic research model applied in different data mining projects. For instance, KDD as a process consist of the following steps including DM is one them.

**Data cleaning:** is the process of removing noise and useless data from a collection**.** The process of data cleansing is normally expensive, hence it was not possible to do with old technologies. Nowadays, faster computers allow data cleansing to be performed in an acceptable amount of time on a large amount of data.

**Data integration:** Because the data may come from a variety of sources, several heterogeneous data sources are integrated into a single source at this stage**.** Issues of data integration include identifying similar entities, removing redundancy, detecting and removing conflicts and errors.

**Data selection:** The data relevant to the analysis is chosen and extracted from the data collection at this stage**.**

**Data transformation:** It's also known as consolidation of data, and it's the act of converting selected data into a usable format for the mining process.

**Data mining:** It's a crucial stage in the data KDD process. When the data has been cleaned and p reprocessed, it is delivered to intelligent algorithms for categorization, grouping, and similarity searches within the data, among other things. We chose methods that are good for finding patterns in data in this section. Some algorithms are more accurate than others when it comes to knowledge discovery. Thus selecting the right algorithms can be crucial at this point.

**Pattern evaluation:** s a process in which only the most intriguing patterns expressing knowledge are assessed using predetermined criteria.

**Knowledge representation:** The found knowledge is visually represented to the user in this final phase. This crucial stage employs visualization tools to assist users in comprehending and interpreting the data mining findings in such a way as shown in fig

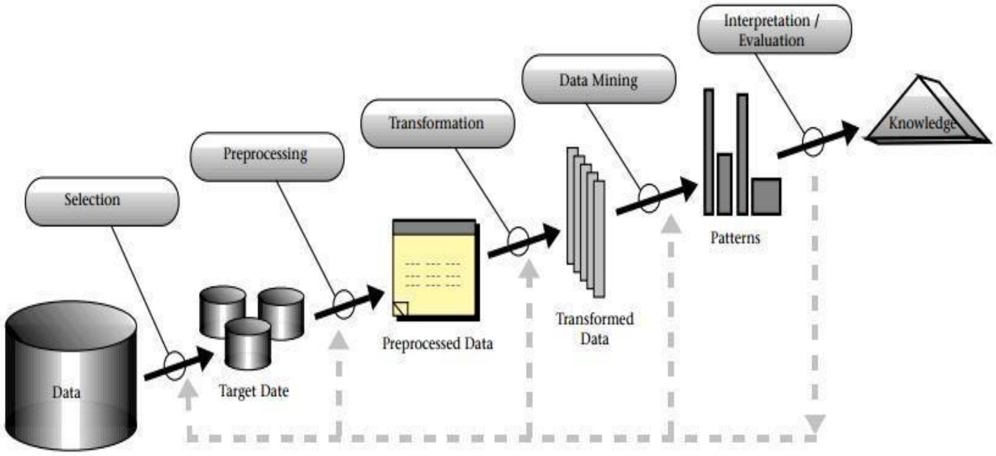


Figure 1:-Knowledge discovery steps (Source: Fayyad,et al [11].)

### 2.4.2 The SEMMA Process

The acronym SEMMA stands for (Sample, Explore, Modify, Model, Assess) .It was developed by SAS Institute which focuses on model development aspects of data mining and involves the following steps;

**Sample:** This stage entails taking a sample of data from a huge data set large enough to contain important information but small enough to alter fast.

**Explore:** This stage entails looking through the data for unexpected trends and abnormalities in order to gain insight and suggestions.

**Modify:** To focus on the model selection process, this stage involves modifying the data by creating, choosing, and altering the variables.

**Model:** This stage entails modeling the data by letting the software to look for a combination of data that reliably predicts a desired outcome on its own**.**

**Assess:** This stage entails reviewing the data by determining the relevance and trustworthiness of the DM process' conclusions, as well as estimating how well it functions.

### 2.4.2. The CRISP-DM process model

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard process model which has been developed by two vendors; Integral Solutions Limited (ISL) (now part of SPSS) and NCR Corporation which is the world‘s leading supplier of data warehouse solutions (Ozgul, 2013). CRISP-DM is a standard process that is a non- proprietary model. It is an application or industry-neutral data mining methodology that mainly focuses on business issues as well as technical analysis.

Because CRISP-DM is a vendor-independent paradigm, it may be used with any DM tool to solve any DM issue (Fayyad & P, 1996). Hence the model defines the phases to be carried out in a DM project with related tasks and the deliverable for each phase as shown in fig.2.

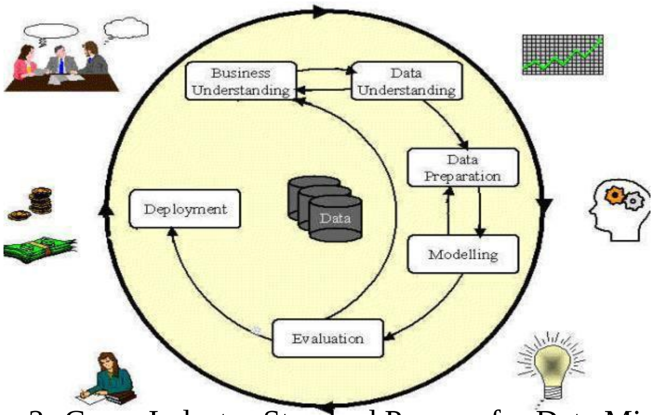


Figure 2:-Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM model is divided into six phases and related tasks as described in Peter [19],

1. **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective. Then, converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives [20].

The Main tasks of this phase are: first, determining business objectives by understanding the client's needs from the business perspective; secondly, assessing situations through investigation of facts about the factors influencing the project; thirdly, determining data mining goals based on the project objectives in technical terms and lastly, producing project plan by preparing a detailed plan to reach the project objectives.

1. **Data understanding:** The data understanding phase begins with data collection and continues with actions to familiarize yourself with the data, find data quality issues, get first insights into the data, or detect intriguing subsets in order to create hypotheses for hidden information**.** Major activities are collecting initial data (initial data collection report), describing data (data description report), exploring data (data exploration report), and verifying data quality (data quality report).

2. **Data preparation:** The data preparation phase includes all of the steps necessary to create the final data set from the raw data. Tasks for data preparation are likely to be repeated and not in any particular order**.** Data preparation and cleansing are a sometimes overlooked yet crucial phase in the data mining process. Frequently, the method used to collect data was not strictly regulated. Thus, the data may contain outline values and impossible data

combinations. Analyzing data that has not been carefully screened for such problems can produce highly misleading results, particularly in predictive data mining [20].

The main tasks in data preparation are; data selection which is a rationale for industry, data cleaning and documenting, constructing data, integrating data, data reformatting, and have data set description.

3. **Modeling:** Various modeling techniques are chosen and implemented in this phase. Their parameters are set to the best possible values. For the same DM problem type, there are usually many approaches. Some techniques have specific requirements in the form of data. Here, modeling technique selection, test design generation to validate the model and test its quality, building model and model assessment through interpretation, evaluation, comparison, and ranking of models according to the evaluation criteria from a data mining perspective are the main tasks of this phase.

4. **Evaluation:** Prior to final model deployment, it is critical to properly examine the model and review the actions taken to construct it to ensure that it meets business objectives. At the end of this phase, a decision should be reached on how to use the DM results through assessment of data mining results to business success criteria to approve models, review of the process, and determine next steps such as possible actions and decisions making based on patterns and relationships.

5. **Deployment:** Model construction isn't always the last step in a project. Even if the goal of the model is to improve data understanding, the information gathered must be arranged and presented in a way that the customer can understand. The main activities in this phase are planning deployment, planning for monitoring and maintenance, producing the final plan and presenting, reviewing a project, and producing documentation.

The sequence of the six stages is not rigid, as is schematize in figure 2.2 CRISP-DM is extremely complete and documented. All his stages are properly organized, structured, and defined, allowing that a project could be easily understood or revised [21]. Besides, CRISP-DM encourages best practices and offers organizations the structure needed to realize better, faster results from data mining. Accordingly due to this reason the researcher selected this model to achieve the objective of the study.

### 2.4.3. Hybrid model

The development of both academics particularly the KDD and industrial oriented (CRISP-DM and other) data mining models has led to the growth of hybrid models, i.e., models that combine the features and job of both. A hybrid model is a six-step KDP model developed by cios [22], the descriptions of the six steps are explained below.

**Step 1: understanding the problem domain:** This initial step involves working closely with domain experts to define the problems and determine the project goals, identifying key people, and learning about current solutions to the problems. The major tasks undertaken in this phase are: understand and learn domain-specific terminology. Explain the problem area in detail and make a boundary for it lastly, convert the project goals into DM goals and select appropriate data mining tools to be used in performing the whole research.

**Step 2: understanding the data:** Collecting sample data and agreeing on data kinds, formats, and sizes was a major responsibility during this phase. These attempts can be guided by background knowledge. Furthermore, data should be examined for completeness, redundancy, missing values, and attribute value plausibility, among other

things. Verification of the data applicability to the DM aims was another key activity in this step.

**Step 3: Preparation of the data:** Once the problem is understood and selected complete the appropriate data to the problem under investigation and the next step is selecting representative data. As a result, the researcher is used only the selected data as input for DM methods in the subsequent steps. It entails sampling, correlation and significance testing, and data cleaning, which includes ensuring that data records are complete, eliminating or adjusting for noise and missing values, and so on.Feature selection and extraction procedures (to reduce dimension), derivation of new attributes (say, through discretization), and data summarization can all be applied to the cleaned data (data granularization).The end result is data that meets the DM tools' unique input criteria which are selected in Step 1.This step is the most time-consuming part of all data mining model Cios.

**Step 4: Data mining:** Another key step in the knowledge discovery process is data mining. It involves the use of several DM tools on data prepared in step 3.The application of these tools is to discover new knowledge. The process of discovering new information includes: the data model was constructed using one of the chosen DM tools and training and testing procedures are designed. The then generated data model was verified by using testing procedures. It takes less time than data prepossessing steps. Some of the common data mining techniques implemented using data mining tools are decision trees, neural networks, clustering, support vector machine, rule induction, association rule.

**Step 5: Evaluation of the discovered knowledge:** Understanding the results, determining whether the discovered knowledge was original and interesting, interpreting the results by domain experts, and determining the effect of the discovered knowledge are all part of the evaluation process. Only approved models were kept, and the entire process was re-evaluated to see what more efforts may have been taken to improve the outcomes**.** A list of errors made in the process is also prepared.

**Step 6: Use of the discovered knowledge:** The third phase entails deciding where and how to apply the newly acquired knowledge. The existing domain's application area could be expanded to include other domains. A strategy was developed to track the implementation of the new information, and the entire project would be documented and reported on**.**

**I adopt the** CRISP-DM process model since it includes a vendor-independent model so it can be used with any DM tool which can be applied to solve any DM problems.

## 2.5. Data Mining Tasks

The data mining tasks are different types depending on the use of data mining results. According to (Fang Weiping, 2013) data mining tasks can be classified into two categories: descriptive and predictive. Descriptive data mining activities characterize the data in the database's general qualities, and predictive data mining jobs infer current data to produce predictions.

As presented(Dua S. & Dua P., 2014), brief insights about data mining, information of health care Insurance frauds, and highlights of the advantages of data mining techniques (Bayesian classification) over other methods used in fraud detection. Different data mining classifications exist: with the most accepted and common categorization adopted by machine learning experts dividing the data mining technology into supervised and unsupervised.

### 2.5.1 Supervised/predictive data mining methods

The predictive model has supervised learning because the classes are predefined before the examination of the target data. Classification, prediction, regression, and time series analysis are just a few examples of predictive modeling data mining jobs (VipulaRawte& rinivas, 2015). The technique of assessing the current and previous states of an attribute and projecting its future state is known as prediction. The practice of mapping target data to specified groups or classes is known as classification. The regression involves the learning of a function that maps data items to a real-valued prediction variable. In the time series analysis, the value of an attribute is examined as it varies over time. Distance measures are used in time series analysis to assess how similar distinct time series data are, the structure of the line is investigated to determine its behavior, and the historical time series plot is used to forecast future values of the variable. The most common learning strategy involves training the model with predefined class labels. The class names in the context of health insurance fraud detection could be "legitimate" and "fraudulent" claims. The model can be built using the training data set. Then every new claim may be compared to the model that has previously been trained to determine its classification. If a claim follows a pattern that is comparable to legitimate activity, it will be labeled as non-fraud; otherwise, it will be classified as fraud. The advantages are that all classes have meaning for people and can be used to classify patterns quickly. The disadvantage is the difficulty associated with gathering class labels.

Furthermore, labeling all of the claims when there is bulk input data is expensive, and claims must be identified properly because false positives and genuine negatives might give customers a negative impression of the insurance firm. Skewed distribution – of the class labels in the training data set can result in a model which does not have very good accuracy for prediction. Supervised learning models cannot detect new types of frauds and significant efforts are required from the experts to derive the labeled training samples which will be used to construct the model. A support vector machine, for example, is a supervised learning technique for categorization. It has an initial training phase in which the algorithm is fed data that has already been categorized. SVM can estimate which class fresh incoming data will fall into when the training phase is completed.

The most common learning technique is supervised machine learning, in which the model is trained using predefined class labels. The class names in the context of health insurance fraud detection could be "legitimate" and "fraudulent" claims. The model can be built using the training data set. Then every new claim may be compared to the model that has previously been trained to determine its classification.

If a claim follows a pattern that is comparable to legitimate activity, it will be labeled as non-fraud; otherwise, it will be classified as fraud. The advantages are that all classes have meaning for people and can be used to classify patterns quickly. The difficulty in acquiring class labels is a disadvantage. Furthermore, labeling all of the claims when there is bulk input data is expensive, and claims must be identified properly because false positives and genuine negatives might give customers a negative impression of the insurance firm. Skewed distribution of the class labels in the training data set can result in a model which does not have very good accuracy for prediction. Expert work is necessary to produce the labeled training samples that will be used to develop the model, as supervised learning methods cannot detect novel types of frauds. A support vector

Machine, for example, is a supervised learning technique for categorization. It has an initial training phase in which the algorithm is fed data that has already been categorized. SVM can estimate which class fresh incoming data will fall into when the training phase is completed.

Supervised techniques are normally used for grouping and projection objectives which include traditional statistical methods like support vector machine (SVM), neural networks, Bayesian networks discriminates analysis, and regression analysis. These supervised approaches need certainty in the true or accurate grouping of the records. Decision Tree, genetic algorithms, support Vector machine, and neural network are an example of supervised data mining methodologies that have been applied to expose abuse and fraud in community-based health insurance schemes.

### 2.5.2. Unsupervised/Descriptive data mining methodologies

Descriptive models are unsupervised learning functions (Sharrifa, 2019).These functions do not anticipate a goal value, instead concentrating on the data core structure, relations, interconnection, and so on. Clustering, summation, association rules, and sequence analysis are all methods used in the descriptive task. Descriptive modeling is a mathematical procedure for describing real-world events and the relationships among the forces that cause them. The process is used by consumer-driven organizations to help them target their marketing and advertising efforts. This type of methodologies usually evaluates one's claims aspects and characteristics in connection to other claims and figure out how they are associated or independent from each other. Unsupervised methods are typically used for characterization which includes segmentation techniques like anomaly detection and clustering and association rules extraction like A priority algorithm [1].

Unsupervised Learning: There are no class designations for unsupervised learning. It concentrates on locating instances of aberrant behavior. Unsupervised learning approaches, unlike supervised learning techniques, can detect both old and new types of fraud since they are not limited to fraud patterns with predefined class labels. The advantages are that it seeks to discover anything that deviates from usual behavior, and because of the lack of direction, it can uncover patterns that have previously gone unnoticed. While the disadvantage being because of lack of direction, there may be times when no interesting knowledge has been discovered in the set of features selected for the training.

Advantage of Supervised Technique (Classification) over Unsupervised Technique (Clustering):

Consider there are two claims made by the same patient, out of which one is the original claim and the other one is a duplicate claim as shown in Fig.2.1. A duplicate claim is formed by changing the date but keeping the rest of the patient‘s details the same.

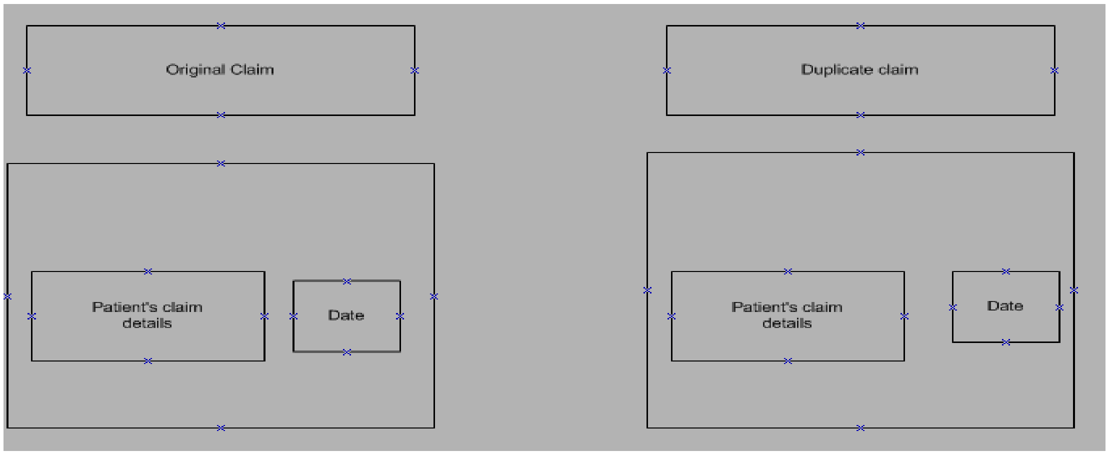


Figure 3:-Original and duplicate claims made by the same patient

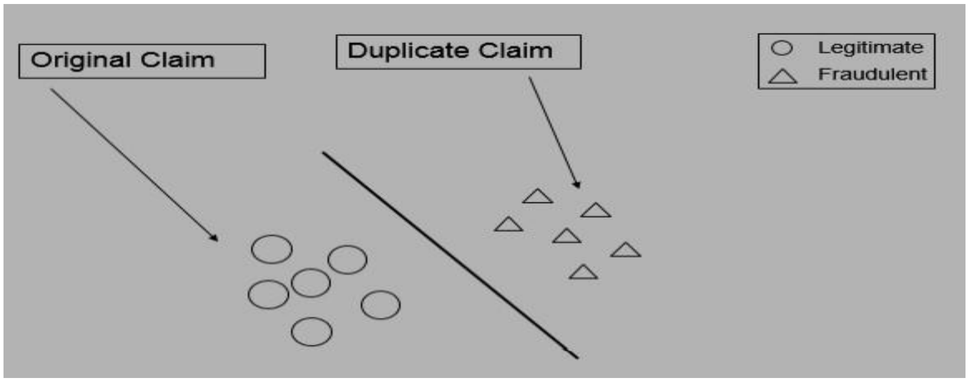


Figure 4:-Classifying the claims

In fig 1.4.Based on the training given to the support vector machines, both original and duplicate claims are divided into their respective classes (SVM). In this case, the duplicate claim is classified as fraudulent and hence identified.

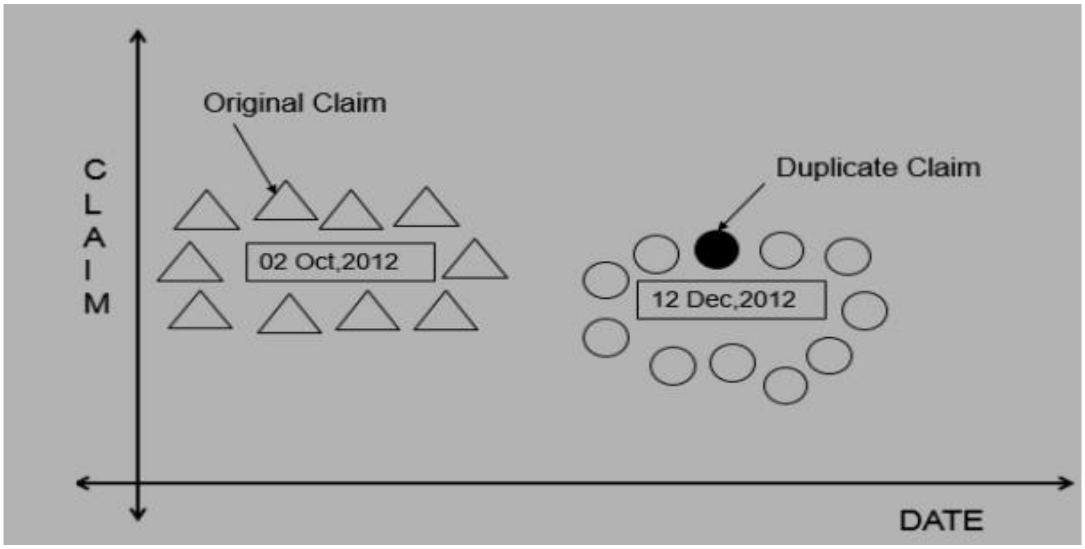


Figure 5:-classification succeeds over clustering (outlier)

But, Fig. 5. This shows that if a clustering-based approach like outline detection is used, then the duplicate claim does not get identified.

## 2.6. Predictive Modeling and Classification

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. As (Adenusi, 2011)further explained Predictive modeling involves using some variables or fields in the data set to predict unknown or future values of other variables of interest and produces the model of the system described by the given data set.

Thus, the main objective of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction, or estimation, among this Classification is commonly used as a data mining technique for prediction [7].

Classification is one of the most common data mining tasks, which is also pervasive in human life. Human beings usually classify or categorize to understand and communicate about the world. For any object or instance, classes are predefined according to the values of a specific field [25]. Han and Kamber [23], also described classification as a process of finding

a set of models or predefined conditions that describes and distinguished data classes or concepts. It is a supervised learning method. In supervised learning, we are provided with a collection of labeled patterns and the problem is to label a newly encountered, yet unlabeled pattern. The given labeled patterns are used to learn the descriptions of classes which in turn are used to label (classify) a new coming pattern. The classification technique maps data into predefined groups. The derived model of classification may be represented in various forms such as the ―If-then‖ rule, decision tree, neural networking, Bayesian networks, etc.

Even though, there are many data mining tasks available and commonly used in various applications, the researcher used decision tree and navies Bayes techniques in this research due to:-

- Relatively faster-learning speed than other classification methods
- Convertible to simple and easy to understand classification if-then-else rules
- Comparable classification accuracy with other methods.
- Does not require any prior knowledge of data distribution, works well on noisy data.

## Decision Tree

Decision trees are a way of representing a series of rules that lead to a class or value. A decision tree is a tree structure that looks like a flowchart, with each node representing a test on an attribute value, each branch representing the test's outcome, and tree leaves representing classes or class distributions (Vipula Rawte & Srinivas, 2015). Decision trees can easily be converted to classification rules depending on the algorithm and each node may have two or more branches. For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed, it is called a multi-way tree. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By exploring the decision tree and determining which branch to take, starting at the root node and going to each succeeding node until a leaf node is reached, a value or class can be assigned to a case. Each node chooses the proper branch based on the data from the cases. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups to a particular target variable (Dua S.& Dua P., 2014).The decision tree model is used to either compute the likelihood that a given record belongs to each of the categories or to classify the record by assigning it to the most likely class when the goal variable is categorical. A decision tree can also be used to estimate the value of continuous variables. A decision tree is a collection of nodes, arranged as a binary tree. The leaves render decisions; in our case, the decision would be ―likes‖ or ―doesn‘t like.‖ Each interior node is a condition on the objects being classified; in our case, the condition would be a predicate involving one or more features of an item [26].

To classify an object, we begin at the root and apply the root predicate to it. Go to the left child if the predicate is true and to the right child if it is false. Then repeat the process at the next node until you reach a leaf. That leaf is categorized as a loved or disliked object. A decision tree's initial state is the root node, which is given all of the examples from the training set. If all of the samples belong to the same class, no additional partitioning decisions are required, and the solution is complete. If the node's examples belong to two or more classes, the node performs a test that results in a split. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution i.e. it may have components that are too specific to noise and outliers that may be presented in the training data. As Apte and Weiss [28] indicated, to relax this over-fitting most decision tree methods go through a second phase called pruning that tries to generalize the tree by eliminating sub trees that seem too specific. Error estimation techniques play a major role in tree pruning. Most modern decision tree modeling algorithms are a combination of a specific type of a splitting criterion for growing a full tree and a specific type of a pruning criterion for pruning tree.

Hence, Decision trees are a simple, but powerful form of multiple variable analyses. They provide unique capabilities to supplement complement and substitute for

- Traditional statistical forms of analysis (such as multiple linear regressions)
- A variety of data mining tools and techniques (such as neural networks)
- Recently developed multidimensional forms of reporting and analysis found in the field of business intelligence.

### Some of Decision tree Algorithms

The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items [29]. Many different algorithms maybe used for building decision trees including J48, CART (Classification and Regression Trees), C4.5 and C5.0, and CHAID (Chi- squared Automatic Interaction Detection) [26].

**J48:-** J48 decision tree algorithms adopt an approach in which decision tree models are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built [23]. According to Bharti [30], the J48 decision tree algorithm is a predictive machine learning model that decides the target values (dependent variable) of a new sample based on various attribute values of the available data. It can be applied to discrete data, continuous or categorical data.

The J48 decision tree can serve as a model for classification as it generates simpler rules and removes irrelevant attributes at a stage before tree induction. In several cases, it was seen that j48 decision trees had a higher accuracy than other algorithms [14]. It offers also a fast and powerful way to express structures in data.

**CART: -**This is a technique that generates a binary decision tree. CART uses binary split based on GINI (recursive partitioning motivated by statistical prediction) techniques and in the CART algorithm there exist exactly two branches from each non-terminal node. In the CART algorithm pruning is performed based on the measure of complexity of the tree.

**C4.5 and C5.0: -** The decision tree algorithm C4.5 is an improvement to ID3 and it produces a tree with multiple branches per node [31]. The numbers of branches are equal to the number of categories of a predictor. C4.5 combines multiple decision trees into a single classifier. In C4.5 pruning is performed based on the error rate at each leaf. C4.5 algorithm has also the ability to handle missing and continuous data. C5.0 is a commercial version of C4.5. Unlike C4.5 the precision algorithm used for C5.0 has not to be disclosed.

**CHAID: -** This is a muti way splitting algorithm using chi-square tests (detection of complex statistical relationship). The number of branches varies from two to the number of predictor categories.

## 2.7. Data Mining in Health care

It is well known that health care is a complex area in which new information is being collected daily at a growing rate. An enormous part of this information is found in the form of paperwork. However, making this information available in electronic form and converting the information into knowledge is not an easy task. All health care institutions need expert analysis of their medical data, which is time-consuming and expensive for human analysts (Durairaj & Ranjani, 2013). The ability to use data in databases to extract useful information for quality health care service is a key for the success of all health care institutions (Divya &Sonali, 2013).

Health care data contains all the information regarding patients as well as the parties involved in health care industries. The size and complexity of such data are increased from time to time. Due to this, the health care sectors need large storage places for their data as well as intelligent technologies to retrieve meaningful information from the complex and dirty data set collections (Parvez, Saqib, & Syed, 2015). Using traditional methods for extracting meaningful information from these complex data sets is impossible. However, due to advancements in the fields of statistics, mathematics, databases, and data mining disciplines, it is possible to extract meaningful patterns from them. Data mining is slowly but increasingly applied to overcome various problems of knowledge discovery in health care (Ruben, 2009).

Data mining is widely applied in the area of genetics and medicine, which is called medical data mining. There is an explosive growth of biomedical data, ranging from those collected in pharmacological studies and cancer therapy, kidney, heart, and brain disease investigations to those identified in genomics and proteomics research. The rapid growth of biotechnology and biological data analysis methods are needs a new field to discover interesting and important patterns from this huge data for better decision making (Pradhan, 2014). Thus, the current signs of progress in data mining research led to the development of many efficient and scalable methods for discovering interesting patterns and knowledge from large databases, ranging from efficient classification methods up to

Clustering, outliner analysis, frequent, sequential, and structured pattern analysis methods, and visualization tools (Tasha et al, 2012). Medical databases are collected and stored in large quantities about patients and their clinical conditions. Relationships and patterns are hidden in this huge data set collections of the sector that could provide new medical knowledge, which helps physicians for disease diagnosis, prognosis, treatment and outbreak prediction, clinics and hospitals to adapt new medical practices, fast and effective medical service, practitioners, and policy makers to identify and develop timely guidelines and medical insurers to provide effective services for customers (Mariammal et al, 2014). Analyzing medical data sets are improved health care by enhancing the performance of patient management tasks like 10 groupings the patients having similar type of diseases or health issues, making better diagnosis and effective treatments, predicting the length of stay of patients in hospital, designing plans

For an effective information management system, analyzing the various factors that are responsible for diseases transmission, helping patients to identify health care institutions that provide services with minimum cost, and accessing the latest information about different types of diseases (Mary K., 2004). Figure 2.2 showed the overall process of data mining in health care (Kamran, 2013).

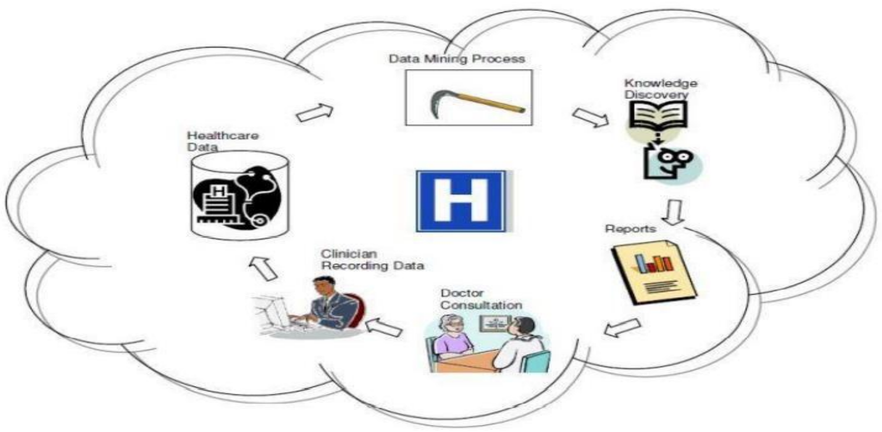


Figure 6:-Process of data mining in health care

The health care data contain details about the hospitals, patients, medical claims, treatment cost, clinical diagnosis reports, pharmaceutical prescriptions, etc( Mary K., 2004). Huge amounts of health care data need to be converted into information and knowledge, which can help to control costs and maintain a high quality of patient care. Without data mining, it is difficult to understand the full potential of data collected from a healthcare organization, which is massive, highly dimensional, distributed, and uncertain. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost savings and decision making by discovering patterns and trends of the data set (Desikan, Hsu, &Srivastava, 2011).

Data mining is the practice of mechanically searching large stores of data to determine patterns and tendencies that go beyond simple analysis. Sometimes data mining is also called knowledge discovery (KDD) [9]. Data mining is about finding new information in a lot of data. The information obtained from data mining is hopefully both new and useful/important. In many cases, data is stored so it can be used later. The data is saved with a specific goal. Saving information makes a lot of data. A large amount of data is usually saved in a database. Extracting new information that can also be useful from data is called data mining. We are in an age often referred to as the information age. In this information age, because People believe that information leads to power and success and thanks to sophisticated technologies and they have been collecting a vast amount of Information‘s from different information source. Initially, with the beginning of computers and means for mass digital storage, people start collecting and storing all sorts of data, counting on the power of computers to help sort through this bulk of information [11]. Unfortunately, these massive collections of data stored on disparate structures very rapidly became. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% [9]. As the volume of data increases, the amount of information in which people could understand decreases and it is difficult to acquire accurate information for making decisions. Data mining is the process of extracting knowledge from huge amounts of data. When we have large data set, it is no longer enough to get simple and straightforward statistics out of them. The data being big data and the changing business needs together change the focus from simple retrieval and statistic into complex data mining.

## 2.8. Role of data mining Community-Based Health Insurance

With the conventional approaches of Community based health insurance fraud and corruption discovery, a small number of auditors manage a lot of healthcare claims forms [17]. The auditors have a minimum time for each form claim, concentrating on particular attributes of a claim form and not focusing their concentration on the exhaustive picture of the provider‘s exploits. Data mining plays a paramount part in preventing and detecting fraud. It enables monitoring and tracking of claims and makes it harder for fraudsters to fake claims. With data mining, one can have a 360-degree view and understanding of the entire process. Insurers have come to learn that no stand-alone technology techniques are adequate. The emerging use of computerized systems and many electronic records has brought about arising prospects for better discovery and uncovering of abuse in health insurance. Developments artificial intelligence and machine learning take interest in programmed ways of detecting fraud. To prevent and detect CBHI fraud, ICT results and different technologies are applied.

As presented [9], brief insights about data mining, information of healthcare Insurance frauds, and highlights of the advantages of data mining techniques (Bayesian classification) over other methods used in fraud detection. Different data mining classifications exist: with the most accepted and common categorization adopted by machine learning experts dividing the data mining technology into supervised and unsupervised.

A 2010 program assessment reported some positive outcomes as well as some challenges. The researchers found that the quality of services was good in terms of contributions to lower maternal and neonatal mortality rates, as well as drug availability and patient satisfaction. The community had an excellent relationship with the MHP. On the other hand, the researchers noted some high turnover among the health providers, who are civil servants, and recommended that the MHP hire private health providers. Financial viability was a key issue. Member attrition was high, and members who used the services infrequently were asking for a discounted premium [44].

## 2.9. Fraud in Health Insurance

Over the years health care- insurance companies receive millions of claims from health care givers or providers. Out of the millions of claims forwarded to health insurance companies, a small percentage of the claims are fraudulent. That small percentage of fraudulent claims costs governments billions of dollars annually. Health insurance fraud in Kenya is still a big headache to the government with industry reports of 2016 on health insurance fraud showing losses and the higher claim rate, a clear indication that fraudsters have invaded the industry. The extensiveness of health insurance fraud inflates costs for all end-users or customers and costs the government and the insurance sector billions of money each year. According to the insurance regulatory Authority (IRA), in a 2016 report, the general insurance business suffered 2.1 billion in losses, with claims rising by 11.8% from 49.05 billion in 2015 to hit 54.86 billion in 2016. Corruption and fraud have been recognized as major barriers to health care [42]. For instance, in the USA, health care insurance fraud is over 30 billion dollars annually and the situation in developing countries like India is not that different[43]. Apima(2018) grouped fraud in health insurance as either opportunistic or professional. Opportunistic fraud is that which is performed by a person who gets a chance to increase the charges of a request or get an overcharged estimation for costs or repairs to her/his insurance institution. Opportunistic fraud is the most common type of fraud though its loss and impact on an insurance company are low. Professional fraud on the other side is mostly performed by organized groups of individuals with many false identities mostly targeting many insurance companies. Globally, insurance companies have become aware of the fraud problem, and different ways of fighting it have been initiated. One most effective and practical approach to handle the fraud problem is effective information support using a way of fraud management systems. Ways of combating fraud can be through prevention, detection, and responding to fraud. These interventions involve the process of discovering past and new and possibly future cases of fraud as quickly as possible. Subelj, Furlan, &Bajec(2011) noted that the main focus so far has been mainly on fraud detection techniques and a lot of literature has been published on that matter.

Concentrated knowledge is however needed to detect deception and misuse or corruption in the health care setting. A number of the health insurance systems in place depend on experienced humans to physically scrutinize insurance claims for the establishment of suspicious ones. This makes the system development and claim to review processes very tedious especially when dealing with big and established national insurance organizations. Lately, there has been an increase in the implementation of electronic claims processing systems. These systems are geared to automatically carry out audits and reviews of health care claims. The systems are modeled for the identification of areas that require proper attention like incorrect or incomplete input data, doubled or identical claims, and creatively not covered services. All those systems are limited as they are applied to find or discover only specific categories of fraud.

To attain efficacious fraud detection, several researchers have endeavored to come up with complicated or advanced ant-fraud techniques which include predictive modeling, data analytic, automated red business rules, geographic data mapping, and link analysis (kizito,2016). Predictive modeling allows insurers to scrutinize past or old fraudulent claims and discover attributes and aspects that can assist in preventing future fraud with the main aim being detecting fraud as early as possible in the claims processing system.

Link analysis explores the connection or association a midst the claims, transactions, and people thus helping tie up distinctive components or actors and then establishes the degree of the connection between the parties there by giving out information that may be used to define aspects that point to any possible fraudulent activity. Automated red business rules are procured that may be incorporated into an insurer's root or main information systems. These procedures are very fundamental in assisting to foresee certain types of doubtful claim activity about past fraud through the recognition of inconsistencies or non-uniformity during the claims processing. Geo-mapping may assist insurers to assess risk and uncover anomalies during underwriting and processing of claims. With go mapping an insurer can ascertain that an incident truly happened where the person alleges it did.(kizito,2016).

In developed countries, laws have been regulated for violation of health insurance. HIPAA conditions that health insurance fraud and abuse is a federal law breaking crime that can have substantial punishment attached to it. Persons that are found guilty of health care insurance fraud are entitled to a sentence of up to ten years in a national prison coupled with a considerable monetary fine.

## 2.10. Related work

Health insurance claims are susceptible to fraud. A survey by(Wafula, orto&Mageto, 2014) found invoicing for services that were not performed, invoicing for additional and costly services than were provided, carrying out medically wasteful or unneeded services to generate insurance compensations, and falsifying a patients examination to validate tests, surgeries or any other medical routines that are not creatively essential as the common fraudulent activities done in most of the Kenyans health insurance claims.

In South Africa, the fraudulent activities in health insurance claims as identified by (Legotlo & Mutezo, 2018) include true/fraud/claims, fluctuating or erratic invoicing of codes, exaggerated invoicing products and facility services, providing of wasteful curative services, replicated claims, exempted benefits and products coasted as insured allowances or payments and claims from illegitimate service providers.

In the global scenario, a survey on health care fraud investigations conducted by the health insurance Association of America private insurers revealed that many of fraud undertaking is affiliated to diagnosis (43%) invoicing services (34%). In Medicare, the very regular forms of fraud include invoicing for services not provided, falsifying the diagnosis to validate or confirm compensation, forging certificates of a medical cause, strategies of prescription and medical records to confirm or validate compensation, asking, giving, and receiving of kickback.(HCFA,2015).

Mirchaye [21] has developed a predictive model for medical insurance fraud detection: the case of the Ethiopian Insurance Corporation. This research is conducted to test 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 were used.

Another research Tariku [1], studied Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company. The research has tried to apply first the clustering algorithm followed by classification techniques for developing the predictive model. K-Means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud. The resulting cluster is then used for developing the classification model. The classification task of this study is carried out using the J48 decision tree and Naïve Bayes algorithms to create the model that best classifies fraud suspicious insurance claims.

Yihenew [3] also research Data Mining Techniques In Support Of Motor Insurance Policy Risk Assessment in the Case of Ethiopian Insurance Corporation (EIC). The research is implemented using the six-step cios et al DM process model. The data collection process in this research was done in two phases. Records about vehicles are collected from INSIS database whereas records about drivers‘are collected manually. The collected data set is p reprocessed using WEKA DM tools and Microsoft Excel to select attributes, derive new attributes, handle missing values and remove outliers. The researcher implemented two classification algorithms, J48 decision tree classification algorithms and multi layer perception (MLP). Using j48 decision tree classification algorithms different experimentation are conducted. The first experimentation with default parameter values and 10 fold cross-validation test options has registered 94.63% accuracy. At last, the researcher tried to recommend future research areas with the specific issue.

Richard &Taghil, (2018) propose a machine learning approach for Medicare fraud detection using publicly available claims data and labels for known fraudulent medical providers. They successfully demonstrated the effectiveness of applying machine learning with random under-sampling to identify Medicare fraud. They employed a C4.5decision tree and logistic regression. The results revealed that C4.5 decision tree and logistic regression learners give the most fraud detection capability, especially for the 80:20 class arrangements giving average AUC scores of 0.883 and 0.882 appropriately with low wrong negative rates.

Hasheminejad & Salimi(2018)propose a novel sliding time and scores window-based method, called FDiBC (Fraud Detection in Bank Club), to detect fraud in bank clubs. In FDiBC, 14 features are produced from each score gained by bank club customers, and five sliding time and scores window-based feature vectors are proposed based on all of the customer members' scores. A positive and a negative label are used to generate training and test data sets from the obtained scores of fraudsters and common customers in a bank's customers' club system, respectively. After generating the training data set, learning is performed through two approaches: 1) For positive data, such as fraudulent customers, clustering and binary classification using the OCSVM method, and 2) SVM- based multi-class classification, C4.5, KNN, and Naïve Bayes methods. The results obtained reveal that FDIBC can detect fraud with 78% accuracy, and thus can be used in practice.

Inês (2017) created an artificial neural network that learns with insurance data and evolves continuously over time, anticipating fraudulent behaviors or actors, and contribute to an institution's risk protection strategies.

Lelenguiya (2015) proposed a model for detecting Non-Technical Loss (NTL) of commercial in electricity utilization utility with the use of data mining technologies like Naïve Bayes, neural network, K-Nearest Neighbor, and Support Vector Machine. He applied the data mining techniques about customer information invoicing or billing system for electricity Utilization in a selection of accounts at Kenya Power Limited. The effectiveness and correctness of the model were verified and assessed to get one accepted technique to be adopted by Kenya Power Limited. From the results of the tested model, the greatest outcome for fraud detection hit rate is attained by support vector machine (SVM) classifier with 86.44% followed by K- Nearest Neighbor with 84.75% and classifier with the least optimal fraud detection rate is the Naïve Bayes at74.58%.

Dharani & Shoba (2015) suggested a data mining technique or approach for discovering fraudulent prescriptions in a big prescription database, the design and built a personalized data mining model for detecting prescription fraud wherein they utilized data mining techniques for allocating a risk score to prescriptions concerning prescribed medicament- diagnosis uniformity, prescribed medicament‘s uniformity within a prescription, prescribed medicament age and sex uniformity and diagnosis- cost uniformity. The suggested model functions substantially well for the prescription fraud detection hitch with a 77.4% true positive rate.

Pal&Pal (2015) suggested the use of varied data mining technologies like ID3, J48, and Naïve Bayes for the discovery of health care fraud. According to the results, J48 has the highest accuracy than naïve **Bayes** having the lowest accuracy of 96.7%.

Joudakietal., (2015) implemented a data mining methodology to a sizable health- insurance institution data set of non-governmental general physicians‘ prescription claims. Thirteen pointers were created in total. More than half (54%) of the general physicians‘ were culprits of performing more or less behavior. The outcomes so determined 2% of physicians as fraud culprits. Discriminates analysis indicated that the indicators showed satisfactory effectiveness in the discovery of physicians who were culprits of committing fraud (98%) and abuse(85%) in a fresh instance of data.

Liu, (2014) has suggested a go-location clustering design that‘s analyses the go- location particulars of Medicare/Medicaid recipients and service gives to identify doubtful claims with the rationale that recipients like to opt for health service providers that are situated in a somewhat shorter distance from where the recipient lives.

Mamo, (2013**)** investigated the potential suitableness of the data mining methodologies in building designs and prototypes that can identify and foretell doubtful nessin levy or tax claims. In his research, he first applied the clustering algorithm to the data set and then finally applied classification techniques to develop the predictive model. The K-Means clustering algorithm is applied to uncover the ordinary classification of the various levy claims as fraudulent or not fraudulent. The subsequent cluster is then applied in the development of the classification model. J48 decision tree and Naïve Bayes classification algorithms were used in this study to build the prototype that can foresee fraud suspicious levy claims best. The model built on the J48decision tree algorithm displayed the highest classification accuracy of 99.98%.The model was assessed with a 2200 testing data set and recorded a prediction accuracy of97.19%

Wei et al, (2013) proposed a novel algorithm, to effectively extract variance patterns and determine unprincipled from real behavior, backed up by a working and usable pattern choice and risk scoring that integrates predictions from independent design models. The results from investigations on a large scale real online banking data prove that the system can attain sizeable higher accuracy and smaller alert volume than the modern bench marking fraud detection systems integrating domain knowledge and conventional fraud discovery techniques.

Ogwueleka, (2011) developed a neural network (NN) design for the credit fraud card identification system with the use of an unsupervised technique that was used to the transactions data to create four groups of low, high, risky, and high-risk groups. The self- grouping map neural network approach was applied for cracking the challenge of bringing out the finest groupings of each record to its related group. The receiver operating curve (ROC) forced it card fraud (CCF) identification watch identified over 95% of fraud incidences without prompting any false panics contrary to other mathematical designs and the two-stage clusters.

Shin et al., (2012) discovered misconduct in internal medicine outpatient clinics' claims by a risk score for showing the extent of the possibility of misconduct by health caregivers; and then grouped the health caregivers with the use of a decision tree. They used a specific interpretation of the outliner score and obtained 38 features for identifying misconduct and corruption.

## 2.11. Data mining methods for fraud detection

One of the most serious issues facing the insurance sector is fraud, which results in significant losses. Fraud in the insurance industry is defined by Gill et al. (1994) as making a fraudulent claim, inflating a claim or adding more items to a claim, or being dishonest in any way with the objective of gaining more than genuine entitlement. Although it is difficult to assess the quantity of losses caused by fraud in Turkey, these losses are reflected in increased premiums charged by insurers. Property and casualty insurance fraud is expected to cost the Canadian insurance business 1.3 billion Canadian dollars a year, or about 10-15 percent of all claims paid out in Canada (Gill, 2009). , It may annoy real clients and cause claims adjudication to be delayed. Investigation fees are also a source of concern. As a result, many insurance companies prefer to pay a claim without conducting an inquiry because it saves them money in the long run (ibid).

According to the Association of British Insurers, fraudulent claims cost the UK insurance industry over 1 billion pounds each year, and fraudsters are constantly devising new schemes (Morley et al., 2006).Fraud can be seen in all insurance types including health insurance in general and CBHI in particular. In the CBHI, fraud is defined as the deliberate deception or manipulation of facts in order to obtain a shoddy health benefit.

In massive amounts of insurance claim data, data mining technologies and procedures can be utilized to detect fraud. Anomaly detection is one of the most used data mining approaches for locating fake records.

This technique aims to detect outliers or anomalies which deviate from the usual patterns. An abnormality in a computer network's data traffic pattern, for example, could indicate that a hacked machine is sending out sensitive data, and an aberration in an MR picture could indicate the presence of a malignant tumor (Kumar, 2005). (Spence et al., 2001).

Anomaly analyses are used for fraud detection in a variety of domains outside of the insurance industry, including credit cards, mobile phones, and insider training monitoring (Chandola et al., 2009).In Community based health insurance fraud detection [30], Nave Baye's and decision tree classification algorithms are data mining techniques that have been tested to work in the area of fraud.

They both will use unlabeled data so that before using the classification algorithms, clustering is used. For clustering, the data into different clusters k-means is applied then human experts label into fraud and valid cases. Then after classification algorithms are used to build a model for predicting the incoming cases. For the study the data initially labeled by the assumption that the rejected claims with the right claim documents accessible are taken as fraud cases and the rest is taken as non-fraud/valid. This means that clustering is 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.

# CHAPTER THREE

# RESEARCH DESIGN AND METHODOLOGY

## 3. Methodology of the Research

This chapter covers the research methods, techniques, and tools of data collection used to conduct the study.

The data mining approach is intended to ensure that the data mining effort results in a stable model that solves the problem it was created to tackle. Various data mining approaches have been presented as blueprints for organizing the process of acquiring data, evaluating data, publishing results, implementing results, and tracking changes. The research design is aimed to build a predictive model that predicts the frauds of community-based health insurance. For this particular study, a hybrid of experimental and design-based research techniques is applied. To build the model that analyses and predicts the CBHI claims (in terms of claim status, user address, payment for drugs, labs, and consultants) using predictive data mining techniques, the Hybrid Knowledge Discovery Process for Data Mining is used. This methodology was developed, by adopting the CRISP-DM (Cross-Industry Standard Processes for Data Mining) model to the needs of the academic research community [10, 28].

Community-based health insurance schemes (CBHIs) apply insurance ideas to the social context of communities, based on their structures and arrangements and driven by their choices. CBHIs can assist communities in managing health care expenses and ensuring that the poor and other vulnerable populations have access to basic health care. The plans are particularly beneficial in reaching rural inhabitants and the informal sector, which includes self-employed people and is not easily insured (e.g., farmers, petty traders, and laborers). These people are frequently unable to pay for basic health care at the point of treatment, which, if unabated, could push them into poverty. Typically, CBHIs are organized and managed by a local community organization. The CBHI plan creates a network of facilities by developing agreements with various health providers. Most plans cover basic health care services (such as prenatal care, births, and child health care) as well as family planning, with some plans also covering hospitalization costs. CBHIs are valuable because they include community members as enrollee and volunteers.

The researcher works closely with domain experts to describe the problem and develop the research goals, identify key persons, and learn about current solutions to the problem in the understanding the problem domain step. A description of the problem, including its constraints, is created based on the insights gathered during this phase. The primary (observation and interview) and secondary (data analysis) data gathering methods are used to identify, comprehend, and analyze the business challenges. An interview with domain experts is used to specify feature selection, and observation is used to comprehend some difficult business processes.

The research goals then need to be translated into the data mining goals and include an initial selection of the data mining tools. The Understanding the Data process entails gathering sample data and determining which data will be required, as well as its format and size. If background knowledge is available, some traits may be prioritized. Next, we must ensure that the data is suitable for the data mining objectives.

Data needs to be checked for completeness, redundancy, missing values, the plausibility of attribute values, etc. Preparation of the data is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data is used as input for data mining tools; is decided. It could include data sampling, data cleaning (such as confirming the completeness of data records, eliminating or adjusting for noise, and so on), and so on. The cleaned data can then be processed further using feature selection and extraction techniques (to minimize dimensionality) as well as the derivation of additional attributes (for example, through discretization and/or normalization).

As a result, new data records would be created, which would fulfill particular input requirements for the data mining technologies that were to be employed.

Another important phase in the knowledge discovery process is data mining. Although data mining technologies are used to find new information, their implementation takes less time than data preparation.

This step involves the usage of the planned data mining tools and the selection of new ones. Data mining tools include many types of algorithms; this step involves the use of several data mining tools on data prepared so far. First, the training and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing procedures. An Evaluation of the Discovered Knowledge step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire data mining process may be revisited to identify which alternative actions could have been taken to improve the results. The final step is Use of the Discovered Knowledge. This step is entirely in the hands of the owner of the database/data warehouse. It consists of planning where and how the discovered knowledge will be used. Moreover, the application area in the current domain will be extended to other domains.

## 3.1 Research design

In this chapter, the methods, techniques, and tools used to conduct the research are discussed in detail based on the steps of the hybrid data mining process model selected to guide the entire process of this research. As is discussed in chapter two data mining process models are developed for purely academic purposes and also for Industrial purposes. 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 the steps of this process model.

This research is experimental research conducted on health/medical insurance claim data collected from West gojjam zone community-based health insurance scheme and West gojjamfunselam referral hospital. In the west gojjam zone, CBHI scheme a large amount of medical service claims collected from monthly reports of all districts under the zone.

Besides, the researcher also collects data from West gojjam funselam hospital because different health care refers their patients to the hospital. The total amount of data collected from the above organization is 8033. And hence, the researcher uses this organizational dataset as a primary source for experiments. This data is also passed through data selection, data transformation, normalization and finally extracting knowledge. In this hospital, there are two patient recording systems. The data mining methods that are experimented within this research are the unsupervised learning method (naïve Bayes) and the supervised learning method (decision tree and neural network).

The naive Bayes classifier method is specifically adapted if the dimension of entered data is higher i.e 8033. Naïve Bayesian classifier supposes that the elements or attributes characters are hypothetically independent and there consist no dependence associations among st the attributes. This makes naïve Bayes the most accurate classifier when the presumption holds [30].

Based on these clustering and classification algorithms a predictive model is developed. A decision tree is a versatile and widely used classification and prediction tool. A decision tree is made up of nodes that create a rooted tree, which is a directed tree with no outgoing edges.

## 3.2. Source of Data Set and Collection Method

Data collecting methods are classified into two categories: secondary data collection methods and primary data collection methods.

For corporations, organizations, and even personal use, data collection is critical. Data is one of the most valuable resources at your disposal in the digital age. In every aspect of our lives, we go through the process of data collection. For example, if you want to move to a new city, you collect as much data as you can. When evaluating a new job offer, you gather information about the company's growth, wage scale, and other factors.

The data gathering process and procedures are more formal in a commercial setting, and the results tend to be better as a result. This is due in part to a clear distinction between the many sorts of data that might be gathered. When used correctly, the right data can move your company forward by assisting you in making the best decisions in areas like selecting a market segment, determining the best marketing mix, making financial decisions, and more. When used incorrectly, it can seem like the choices being made by you or your team is always falling short.

How can you make sure you have the right information to make important decisions? By adopting sound data collection methods and analysis.

## 3.3. Architecture of proposed model

To achieve the general objective of this study, a hybrid data mining process model is applied. As indicated by (Sisay, 2019) hybrid data mining process model is selected because of

- providing more general concepts,
- research-oriented description of the steps,
- Introducing a data mining step instead of the modeling step and also its flexibility since it has a feedback mechanism in more steps than CRISP-DM. A hybrid model has six steps that are:

A hybrid model has the following six steps that are:

1. Understanding of the problem domain
2. Understanding of the data
3. Preparation of the data
4. Data mining
5. Evaluation of the discovered knowledge and
6. Use of the discovered knowledge. The research is designed according to the steps of this process model.

Proposed model Architecture

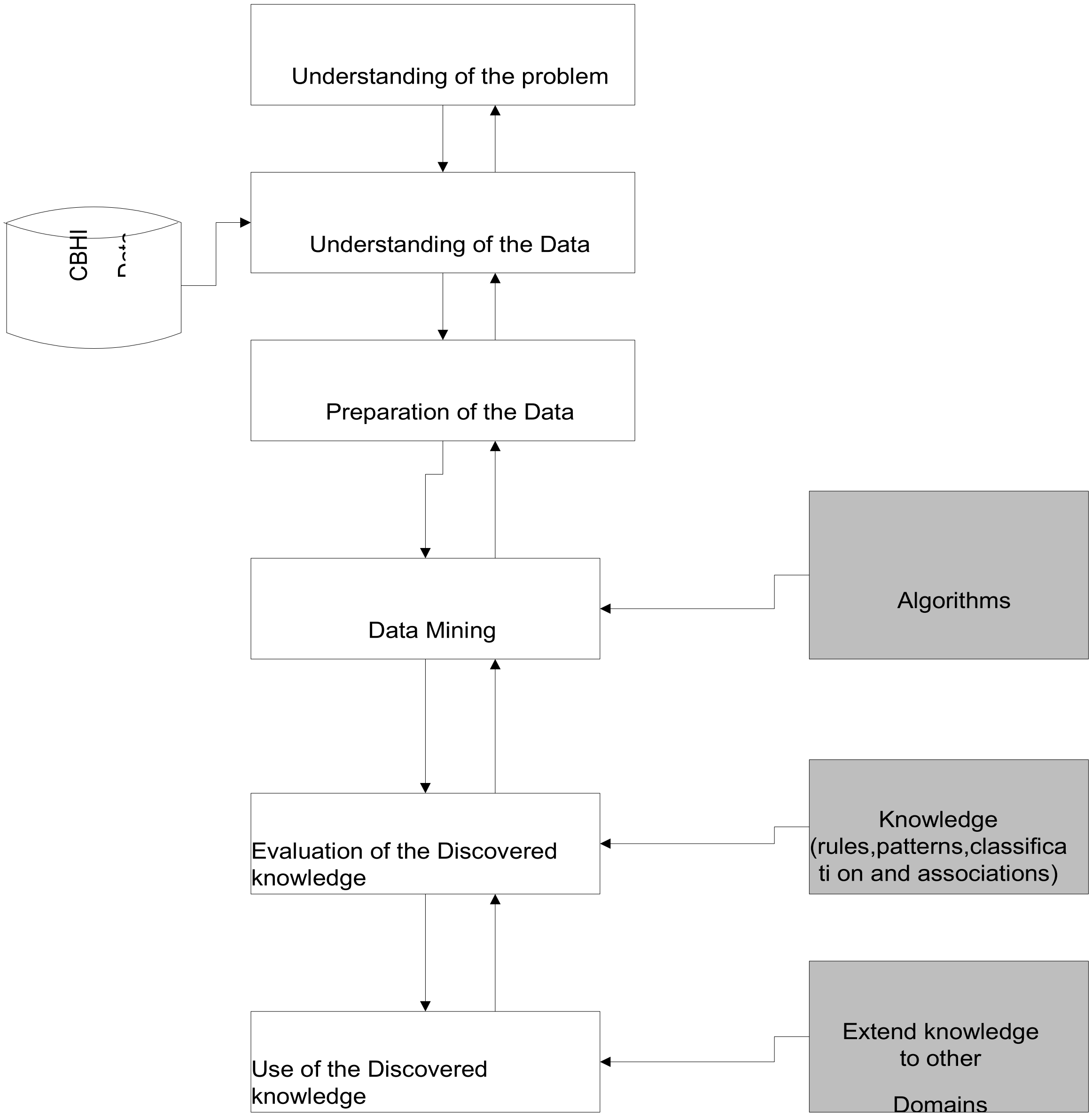

Understanding of the problem
CBHI Data
Understanding of the Data
Preparation of the Data
Data Mining
Algorithms
Evaluation of the Discovered knowledge
Knowledge (rules,patterns,classifica ti on and associations)
Use of the Discovered knowledge
Extend knowledge to other Domains

Figure 7:-1Architecture of proposed model

### 3.3.1. Understanding of the problem domain

To understand the problem primary and secondary sources are used in the study. Secondary data is information that has been collected, analyzed, and formatted by someone or a group other than you.

Primary data, often known as raw data, is information that you acquire and analyses yourself. It's information obtained directly from the source. This could include in-person interviews, audience polls, or even courses.

Primary data is typically obtained with a specific aim in mind, although it can be more difficult to understand for the researcher. This is due to the fact that the data is unstructured and must be organized in a way that allows you to make informed decisions. Primary data collection methods can be divided into two groups: quantitative and qualitative. The research is conducted from west gojjam Zone Health Bureau, funselam hospital, and several CBHI Schemes found in the zone.

Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals, etc…, So that for the study the researcher will use published data from different articles including Ethiopia health insurance` published reports in 2015 and their websites. The research data set is collected from the West gojjam Zone CBHI Scheme in Excel and database format. The data set availed contains both outpatient and inpatient claims records from 2017 to 2019/2020. Despite all these challenges, the data with their selected attributes were collected with maximum care that can be taken. The collected data were entered in Microsoft Excel format to take advantage of easier data manipulation and further ease of accessing the data from the weka interface which was used as a platform for model building purposes. Accordingly, a total of 8033 data were collected from the two hospitals (the data was collected from West gojjam Zone CBHI) and filled in Excel format.

This step helps to select the appropriate algorithm and adopt the proper methodology for the study. Another important step to be performed here is selecting the tool to be used. In this research WEKA, a data mining tool is used since it is platform-independent and contains a graphical user interface that is easy to understand. Moreover, it contains different data mining algorithms and the researcher is also familiar with this tool.

### 3.4.2. Understanding of the data

After the CBHI data is collected from the existing manual as well as electronic files, the attributes found in the data should be discussed with the experts to understand the purpose of each of the parameters and whether they are useful for the specific problem

Area or not. Then attributes have been described from different data points. The CBHI data has been organized in excel format collected using primary and secondary sources.

### 3.4.3. Preparation of the data

In this step, the appropriate data sets for the research were selected. These data sets were cleaned, integrated, and formatted to be adequate for use. Data cleaning activities have been performed including removing duplicated records by WEKA and excel, 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 problems. The cleaned data with the above techniques would be added processed for dimension reduction. For dimension purposes, the WEKA attribute selection is used. Finally, the input data set for running the classification algorithms as training and test data produced.

For the study that has been conducted, the researcher applies the prediction data mining approach, so due to the nature of the data we have determined the suitable data mining task, algorithm and finally, we have employed the data mining algorithms to satisfy the intended objective of the study. In our case, we use the classification technique/ task to develop the model. There are several algorithms present today for data mining but appropriate algorithms have been selected based on the nature of our data nominal, interval like age, and numerical. The training and testing procedures are designed and the data model is constructed using the chosen data mining tools. The researcher adopted decision tree (J48) classification algorithms and Naive Bayes Classifier algorithm. Finally, the researcher evaluates the accuracy of those algorithms.

### 3.4.4. Data mining

Data mining, also known as knowledge discovery (KDD), is the process of extracting patterns and other useful information from big data sets. Because of the advancements in data warehousing technologies and the rise of big data, the use of data mining techniques has exploded in recent decades, supporting businesses in turning raw data into valuable knowledge.

Despite the fact that technology is always evolving to handle massive amounts of data, executives still confront capability and automation issues (Osmar, 2008).

Through smart data analytic, data mining has improved corporate decision-making. The data mining techniques used in these investigations can be classified into two categories: they can either describe the target data set or predict outcomes using machine learning algorithms. From fraud detection to user habits, bottlenecks, and even security breaches, these strategies are used to organize and filter data, revealing the most valuable information.

### 3.4.5 Evaluation of the discovered knowledge

The models which are developed from the data mining step are evaluated. The algorithm is tested on test data set to see how many of the test set is classified as good performance and then calculating recall and accuracy. The performance of the algorithms is also evaluated by the percentage of classification and time parameter: the duration of time to build the model. Finally with the involvement of the domain experts the resulting model is evaluated.

### 3.5.6 Use of the discovered knowledge

This is the last and final step of the KDD process to use the discovered knowledge for different purposes. The discovered knowledge can also be used by interested parties or can be integrated with another system for further action.

Finally how to use the discovered knowledge is perceived by developing a prototype to test each incoming claim before processing.

## 3.4 CBHI Data preparation

The process of cleansing erroneous data is known as data cleaning. The majority of the data is unclean. It indicates that most data can be wrong for a variety of reasons, including device failure, network failure, or human mistake. As a result, data cleaning is required prior to mining.

The data cleaning step includes activities such as removing unnecessary data values 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 are the beneficiary of CBHI.

What are the importance and benefits of data cleansing?

- ✓ Data Cleaning removes major errors.
- ✓ Data Cleaning ensures customers to get more accurate decisions.
- ✓ Data cleaning eliminates discrepancies that are common when many sources of data are combined into a single data set.
- ✓ Data cleaning makes the data-set more efficient, more reliable, and more accurate.

**How to Handle incomplete/Missing Data?**

- Ignore/delete the tuple row containing the erroneous data
- Fill in the missing value manually
- Fill the values automatically by
  - Getting the attribute mean
  - Getting the constant value if any constant value is there.
  - Getting the most probable value by Bayesian formula or decision tree

## 3.5. Medical service claim process

The CBHI schemes and health service provider has an agreement on the payment of medical service claims. According to their agreement, the health care or hospitals can present their medical service payment quarterly. After they provide service to beneficiaries (CBHI members) they must record all patient files such as beneficiary ID, Name, sex, age, CBHI ID, date, types of diseases, cost for drugs, imaging, laboratory, surgery, counseling‘s, etc. The CBHI claim cost is delivered to the insurers and the CBHI scheme pays the claim. In this study, all process is analyzed and evaluated based on the developed model. Table 3.2 describes the attributes of the dataset having the original name and the modified attribute name with their corresponding description.

Fields from the different tables are selectively taken by consulting with business experts. CBHI 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. Besides, 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 to the problem to make further analysis on the claim ratio of fraudulent cases. Some parameters have null values and others are not related to the problem at hand. By merging all the attributes from the different tables a total of 19 attributes are taken and described as follows in table 3.2.

| Attribute name | Modified attribute name | Description | Selected |
|---|---|---|---|
| BeneficiaryCode | Bcode | Unique clients number | □ |
| Beneficiary Name | Bname | Clients name | X |
| Claims start data | Startdate | Registered data | □ |
| Claims end date | EndDate | Finishing date | □ |
| PatientNumber | PatientcardNumber | Unique patient number | X |
| Age | Age | Age of client when he/she registered in CBHI | □ |
| Gender | Sex | | □ |
| Disease type | Diseasetype | Type of diseases | □ |
| Residence | Residence | Residence of the client | □ |
| Employment | Employed | Employment status of the client | □ |
| Family Members | Haschildren | Whether the client has a child or not | □ |
| Fee for consultant | Consult fee | Payment for consulting | □ |
| Fee for labs | Lab fee | Payment for laboratory | □ |
| Fee for images | Image fee | Payment for x-ray | □ |
| Fee for surgery | Surgery fee | Payment for surgery | □ |
| Fee for drugs | Drug fee | Payment for drug | □ |
| Fee for admission | Admission fee | Payment for admission | □ |
| Others fee | Others fee | Payment for others | □ |
| Totalfee | Total fee | Total payments | □ |

Table 1:list of selected attributes and their converted form

For some others, new attributes were derived from the existing ones. They are discussed as follows. –Patient Address‖ in the Patient Registration Form refers to the residence of the patient. Health Facilities capture this information (Woreda/ Kifle-Ketema/ Peasant Association, Kebele, House Number, Telephone number) for effective follow-up. But these details are irrelevant for the Bayesian Network learning process. Possibly, what is important is the person‘s residence category- Rural or Urban. Therefore, the clients were categorized simply as rural or urban. Another attribute, ―Employment‖, contains four values: Working Full time, working Part-time, Not Working due to ill health and Unemployed. Again this attribute was modified as Employed, and its values were assigned simply to be ―Yes‖ if a person is working full-time or part-time, or ―No‖ otherwise. Yet another attribute, family Members-Children‖ details the number of children alive/dead. For most cases, however, only the presence or absence of children was reported and no other details. Here, from the researcher‘s point of view, the presence/absence of children was more important (and reliable) as compared to the partially-filled ―number of children‖ for Bayesian learning. Therefore, this field was renamed as have children‖ and the values for the attribute became ―Yes‖ if the client has at least one child, or ―No‖ if not.

After the removal of the above mentioned attributes, a total of 15 attributes (including the target class) were held for the model building purpose. (See Table 1) Another data reduction technique applied in this research was discretization. The collected and cleaned data attributes here constitute discrete (categorical) values except for the Age attribute whose values were numerical. Since the tool selected to conduct the research was a Bayesian Network tool that accepts only discrete values, the Age attribute was discredited into three discrete classes. The task was performed by BN Power Soft‘s pre- processor, using the equi-width discretization technique. The following table shows the actual values and their corresponding transformations for the Age attribute. The research deals with predicting fraud based on CBHI adherence of the client in different age categories; therefore, the minimum age observed was 16 and the maximum age was 80.

| Age range | Transformed value |
|---|---|
| >=16 and<=29 | Young |
| >30 and <=55 | Middle-age |
| >55 and <=75 | Senior |
| >75 | Old |

Table 2: Discretization of the age manually

## 3.6. Data Analysis tools

There are different open-source data mining tools available to support the tasks of data mining. To complete the study WEKA will be adapted for mining the data. The WEKA software tool is the proposed tool that will be used because it supports various standard data mining operations. WEKA is a set of machine learning algorithms that may be used to solve real-world data mining challenges. The WEKA tool is preferable for the study because of its functionality (support for algorithms), the familiarity of the researcher with the software, and ease of use and runs on almost any platform. WEKA (Waikato Environment for Knowledge Analysis) is an open-source and free machine learning software that is used to extract data. Weka is GNU General Public License-compliant open source data mining software. The University of Waikato in New Zealand developed this approach. The Waikato Environment for Knowledge Analysis (Weka) is an acronym for Waikato Environment for Knowledge Analysis. Weka is a platform that implements cutting-edge data mining and machine learning algorithms. The Weka tool allows users to conduct tasks such as association, filtering, classification, clustering, visualization, regression, and more.

The WEKA platform consists of a package of visualization tools and algorithms for data exploration and projective designing or modeling as well as menu-driven design for easy access. Different algorithms are supported by WEKA i.e. classification, regression, decision trees, and clustering.

This tool enables users to instantly try out and construct independent machine learning techniques on new data sets. Its flexible, extendable framework allows complicated data mining procedures to be modeled from the huge collection of database learning algorithm and tools given. As an additional, the proposed technique that will be applied in this paper is Decision Tree (J48, ID3, and NAÏVE BAYES) because it is powerful classification algorithms. ID3 stands for Iterative Dichotomiser3.It constructs the decision tree one inner node at a time, selecting the attribute that gives the maximum information gain if the instances were divided into subsets based on the values of that attribute at each node.

They stated that the ID3 Kappa static value observed is higher than the J48 and NB algorithms, and that ID3 produces less error than the J48 and NB algorithms [9]. J48, on the other hand, can assist in not just making correct predictions from data but also explaining its patterns. It deals with the problems of the numeric attributes, missing values, pruning, estimating error rates, the complexity of decision tree induction, and generating rules from trees (Witten and Frank, 1999).

### 3.6.1. 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 the hypotheses Hi with a known (conditional) probability P(A|Hi). Assume besides that the probabilities of hypotheses H1,…,Hn are known(prior probabilities). Then the conditional (posterior) probability of the hypothesis $H_i$i=1, 2 …, n given that event A happened is

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

Naïve Bayes is proved as one of the most efficient and effective algorithms for data mining. An explanation is granted of 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) State that different studies found that the simple Naïve Bayes classification has comparable performance results with decision tree and neural network classifiers and has high accuracy and speed when applied to large databases [20].

Naïve Bayes is found simple but effective classification algorithm in solving real-life problems [45]. 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 significant improvements. [13] After referring to the different modifications concluded that such modifications lead to complications that deviate from their 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= $(x_1,x_2,\ldots,x_n)$, depicting n measurements made on the tuple from n attributes, respectively,$A_1,A_2,\ldots,A_n$.

2. Suppose that m classes, $C_1,C_2,\ldots,C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on

   X. That is,If and only if, a naive Bayesian classifier predicts that tuple X belongs to the class Ci.

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

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 posterior hypothesis.

3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ needs 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 $P(X|C_i)$. 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, $P(X|C_i)$ 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.

$\leq \leq$

5. To predict the class label of X, $P(X|C_i)P(C_i)>P(X|C_j)P(C_j)$ for 1 j m, j I, $\neq$

   In other words, the predicted class label is the class $C_i$ for which $P(X|C_i)P(C_i)$ is the maximum

### 3.6.2.Decision tree classification technique

Decision tree builds classification models in the form of a tree structure by breaking down a data set into smaller and smaller subsets incrementally. The final result is a tree with decision nodes and leaf nodes. There are two or more branches in a decision node. A classification or conclusion is represented by a leaf node. The root node is the topmost decision node in a tree [24]. Both category and numerical data can be handled by decision trees.[20] Defines the decision tree as a flowchart-like structure in which each internal node (non-leaf node) symbolizes an attribute test, each branch represents a test outcome, and each leaf node (terminal node) stores a class label.

Decision tree inducers are algorithms that create a decision tree automatically from a dataset. By minimizing the generalization error, the goal is to identify the best decision tree [21]. Most decision tree classifiers perform classification in two phases: tree-growing or tree building and tree-pruning [12]. The tree-building is done in a 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 a bottom-up fashion.

In the tree-building phase, there are different splitting and stopping criteria available. The splitting criteria include unilateral (impurity-based, information gain, gain index, gain ratio, distance measure, and more others) and Multivariate splitting criteria. The best splitting criteria is not greater than a certain threshold, 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, and the number of cases in the terminal node is less than the minimum number of cases for parent nodes [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, loosely stopping criteria are used to make the tree over-fit. Then, 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. 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 Dichotomous) 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 the classification of those new objects decision tree should capture some meaningful relationship between an object‘s class and its values of the attributes.One method for tackling this problem is to build all potential decision trees that correctly classify the training set and then pick the simplest one. The ID3 is intended for classification problems with a large number of attributes and a large number of objects in the training set, but where a relatively effective decision tree is required without a lot of work.

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 is constructed from the top down, beginning with a root node, and involves partitioning the data into subsets containing instances with comparable values. The homogeneity of a sample is calculated using entropy in ID3. That is, if the sample is totally homogenous, the entropy is 0; if the sample is evenly divided, the entropy is 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 the target feature or when the 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 a normalization of the information gain.

Gain ratio =

Based on the concern that making the algorithms more efficient, cost-effective, and accurate for different areas in the real world different 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 the most powerful approaches in data mining [23]. They are relatively fast in classifying unknown records, they can 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 a variety of platforms.

**The J48 decision tree algorithm**

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

1- Because the tree represents a leaf when the instances are of the same class, the leaf is returned by labeling with the same class.

2- A test on the attribute is used to calculate the potential information for each attribute. The gain in knowledge that would arise from a test on the attribute is then determined.

3- Then, based on the current selection criterion, the best characteristic is identified, and that property is chosen 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 number splits the data list into those whose attribute value is less than or equal to the threshold and those whose value is greater than or equal to it.

2- In the training data, it handles missing values.

3- The tree prunes itself after it has completed its construction.

## 3.7. Evaluation methods

The performance of classification algorithms is usually examined by evaluating the accuracy of the classification. 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 [29].The percentage of correctly categorized test set tuples represents the accuracy of a classifier on a given test set. A confusion matrix is a method for determining the accuracy of classification.

A confusion matrix is given n classes an m x n matrix where $C_{i,\ j}$ in the first m rows and n columns indicate 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].

| | C1 | C2 |
|---|---|---|
| C1 | True positives | False negatives |
| C2 | False positives | True negatives |

Table 3:-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 measures of relevance. In this interpretation Precision =

true positive / true positive + false positive

Recall = true positive / true positive + false negative

The 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

3.6.2. Absolute error: $|Y_i - Y_i‘|$ where $Y_i$ is the actual value and $Y_i‘$ is the predicted value □ Squared error: $(Y_i - Y_i‘)^2$

3.6.3. Mean absolute error: $\frac{\sum_{i=1}^{d} |Yi - Yi'|}{d}$ , the average absolute error over the data set

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

The average mean squared error exaggerates the presence of outlines while the mean absolute error does not. We

can normalize the above errors by the mean value Y. $\frac{\sum_{i=1}^{d} |Yi - Yi'|}{\sum_{i=o}^{d} |Yi - Y|}$

- Relative absolute 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,

- Relative squared error: $\frac{\sum_{i=1}^{t} Yi}{d}$

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

The cross-validation and percentage split methods are applied for this work.

**10-fold cross-validation**

In this method, the data set is partitioned into 10 mutually exclusive subsets of equal size. Then training and testing are 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 data. 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 biased and variance.

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

# Chapter four

## 4. Results and discussion

This chapter presents an analysis and results from a discussion of the data mining model and the prototype developed for this study.

### 4.1. Development and testing environment

We have used Windows 10 Home edition as an Operating System with the hardware components Intel® Core™ i7-5500U CPU, 2.40 GHz with 12GB RAM and 1TB Hard disk has been used as a development and testing environment.

### 4.2. Experimental Design

Different experiments were conducted in this research and all experiments with two situations, decision tree with pruning and without pruning, naïve Bayes with all attributes and with selected attribute. The tests were done with the use of a test set and 10-fold cross-validation and on different test split percentages.

To check the quality of the classifier the accuracy of the algorithm is measured. The accuracy of the classifier is measured by true positive rate, true negative rate, F-measure, Recall, Precision. The research also measures the error rate of each classifier mean absolute error, root means squared error, relative absolute error, root relative squared error are measured. WEKA version 3.9 is the tool that is chosen in an experiment to analyze the output of the algorithm. Experimenter and explorer are two mainly used interfaces during this experiment WEKA can perform data preprocessing and modeling technique, and easy to use for the research with its easy to use graphical user interface. WEKA handles different data mining tasks which are, data preprocessing, clustering, classification, regression, visualization, and feature selection.

### 4.2. Data preparation, filtering, and selection

The original data set was classified in different groups i.e approved, denied, and canceled, in litigation. We picked on the approved and denied claims only on the assumption that the approved claims were thought to be non-fraudulent and the denied claims were thought to be fraudulent.

The data was then filtered for selecting claims with only complete and useful data. Weka and excel were applied to remove repeating claims, remove claims with zero (0) amount, and remove claims with missing columns.

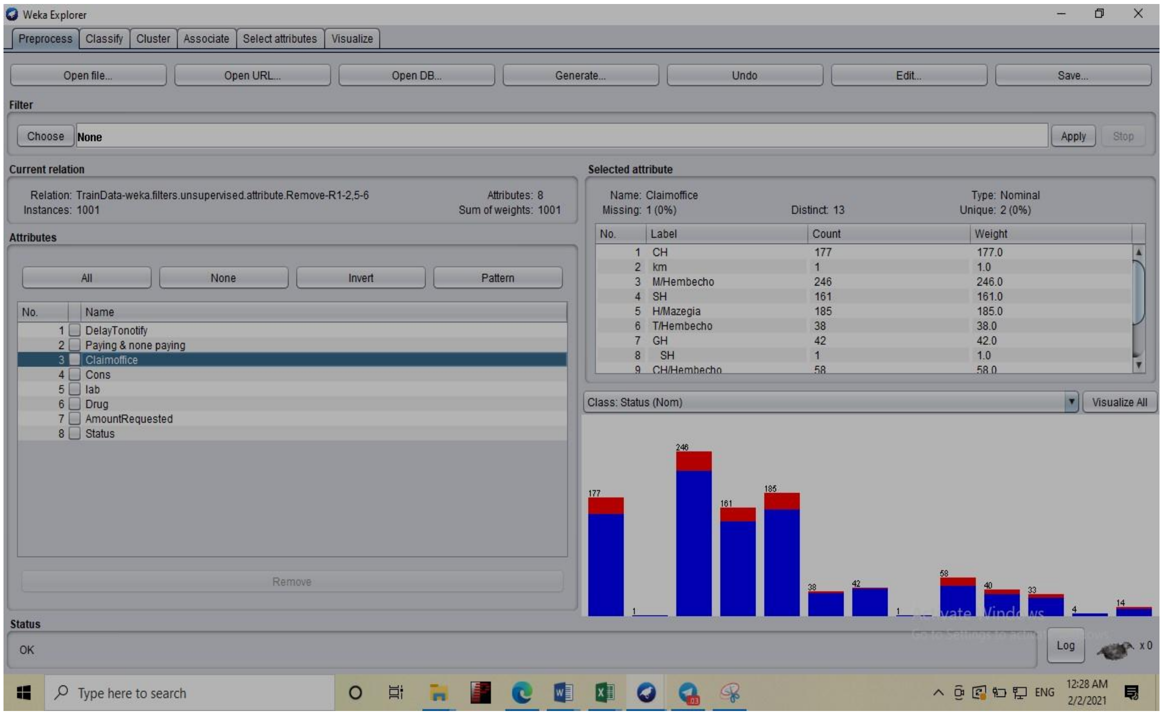
The initial data set provided us with 20 features as follows: Beneficiary code, Beneficiary name, claim start date, claims end date, Age, gender, disease type, residence, employment, family members, fee for consultant fee for labs, fee for images, fee for surgery, fee for drugs, fee for admission, others fee, total fee.

Diagnosis and the total amount charged on a claim based on different providers should give an idea of any fraudulent activity claims. Thus the claims were classified into claim payment amount, diagnosis, and the providers. Given that the accessibility of a facility and payment charges differs with the disease, in addition to status, we used the diagnosis as the control variable in our analysis. So, the features that were considered for prepossessing step were Beneficiary code, Beneficiary name, claim start date, claims end date, paying or nonpaying Age, gender, disease type, residence, employment, family members, total fee.

Later the claims data set was classified according to the principal diagnosis. The five most common diagnoses in the data set were taken. After claims, filtering, and selection, only 8033 claim records remained from the initial 12,187 records collected for the study.

Even though a large number of claims were removed after filtering conditions, the number of claims remaining was more than sufficient for classification on the different classifiers. Therefore, classification methods are used to build the predictive model. Data mining algorithms for classification are naïve Bayes and J48 classification algorithms which are explained in chapter three.

Here is the step in the knowledge discovery process where the real tasks of data mining or model building take place to extract novel patterns hidden in the data set. Each time the model build classification accuracy will be tested using 10 fold cross-validation and a separate test data set. The models built by the two algorithms are compared to each other from the point of view evaluating criteria discussed in chapter three.



Filter

None

Current relation

Relation: TrainData-weka.filters.unsupervised.attribute.Remove-R1-2,5-6 Attributes: 8
Instances: 1001 Sum of weights: 1001

Attributes

| No. | | Name |
|---|---|---|
| 1 | □ | DelayTonotify |
| 2 | □ | Paying & none paying |
| 3 | □ | Claimoffice |
| 4 | □ | Cons |
| 5 | □ | lab |
| 6 | □ | Drug |
| 7 | □ | AmountRequested |
| 8 | □ | Status |

Selected attribute

Name: Claimoffice Type: Nominal
Missing: 1 (0%) Distinct: 13 Unique: 2 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | CH | 177 | 177.0 |
| 2 | km | 1 | 1.0 |
| 3 | M/Hembecho | 246 | 246.0 |
| 4 | SH | 161 | 161.0 |
| 5 | H/Mazegia | 185 | 185.0 |
| 6 | T/Hembecho | 38 | 38.0 |
| 7 | GH | 42 | 42.0 |
| 8 | SH | 1 | 1.0 |
| 9 | CH/Hembecho | 58 | 58.0 |

Class: Status (Nom)

Status

OK

Figure 8:-Weka explorer with CBHI training dataset

## 4.3. Classifiers performance

This is a process in the knowledge finding process in which the real processes of data mining or mode building take place to extract novel patterns hidden in the data set. The training data set was classified into approved and denied claims. Claims with the denied status were taken to be suspicious or fraudulent cases. Therefore classification methods were used to build the fraudulent predictive models. The naïve Bayes and decision tree (J48) algorithms were selected for the classification experiments. These techniques and algorithms were chosen because

- ✓ They have been widely and successfully applied in predictive analytic over time
- ✓ They need reasonably little trouble from users in the preparation of data to get around the scale variations amid st the parameters.

10 fold cross-validation and the percentage-split (66%) classification setups were applied in performing the experiments of training the models. A 10-fold cross-validation setup has been tested to be mathematically good enough in assessing the effectiveness of the

classifier.

To evaluate the accuracy of the classifiers in grouping the claims into defined categories, an analysis of the classification was done. Accuracy denotes the percentage of the correctly made forecasts by the model in comparison to the real or classifications. The classification accuracy of the independent model is recorded and their effectiveness is compared in grouping new instances of records. A different test dataset was applied to test the effectiveness of the models. The models built by the two algorithms were then compared.4-four experiments were done in two different scenarios. The data set was run on each algorithm into two instances to establish the best model. The 10 fold and 66% percentage split were used on each algorithm.

### 4.2.1. Naïve Bayes model building

The Naive Bayes algorithm is an intuitive method that uses the conditional probabilities of each attribute belonging to each class to make a prediction. It employs Bayes' Theorem, a formula for calculating probability from historical data by counting the frequency of values and combinations of values. The maximum likelihood approach is used to estimate parameters in naive Bayes models. Despite oversimplified assumptions, it frequently outperforms in a variety of difficult real-world scenarios. One of the major advantages of the Naïve Bayes theorem is that it requires a small amount of training data to estimate the parameters.

#### Experiment 1

Naïve Bayes was the first data mining technique applied to classify the data. Naïve Bayes classification algorithm or technique functions based on the three conditions

- ✓ The prior probability of a given hypothesis
- ✓ The possibility of the data given that assumption and
- ✓ The possibility of the data itself.

Its classification performance is drawn from the presumption of conditional independence amid st the attributes.

Weka software framework was used to model the naïve Bayes model using the naïve Bayes simple algorithms. The 10 fold cross-validation and the percentage split with 66% for training were employed.

| Actual | Predicted | | |
|---|---|---|---|
| Naïve Bayes classifier with 10fold cross validation test mode | Fraud | Non fraud | Total |
| Fraud | 2183 | 540 | 2723 |
| NonFraud | 4668 | 642 | 5310 |
| Naïve Bayes classifier with percentage split(66%) test mode | | | |
| Fraud | 936 | 347 | 1283 |
| NonFraud | 3175 | 552 | 3727 |

Table 4:-Confusion matrix for experiment 1, naïve Bayes classifier with test options

| | Result | Percentage | Time |
|---|---|---|---|
| Naïve Bayes classifer with 10 fold cross validation test mode | | | |
| Correctly classified | 7183 | 89.42% | 0.03sec |
| Incorrectly classified | 850 | 10.58% | |
| Naïve Bayes Classifer with percentage split(66%) test mode | | | |
| Correctly classified | 3808 | 76% | 0.03sec |
| Incorrectly classified | 1202 | 24% | |

Table 5:-Classification results for experiment 1, naïve Bayes classifier with different test options

From the results, it can be seen that the Bayes classifier with 10 fold cross-validation test mode has relatively better performance than the other two cases. 89.42% (7183) of the test data (8033) are classified correctly to the respective class fraud suspected or not.

**Experiment two**

The prediction model is tested with the naïve Bayes model within this experiment test options

### 4.2.2. J48 Decision tree model building

J48 is an open-source Java implementation of a simple C4.5 decision tree algorithm.J48 is an ID3 extension. Accounting for missing values, decision tree pruning, continuous attribute value ranges, rule derivation, and other features are included in J48. J48, as a decision tree classifier, employs a predictive machine-learning model that determines the consequent value of a new sample using various attribute values from the available data. The internal nodes of a decision tree represent the various qualities; the branches between the nodes indicate the possible values for these attributes in the observed samples, while the terminal nodes indicate the dependent variable's ultimate value (classification).

J48 is a classification algorithm that is used to build decision trees. The dataset was prepared and labeled with fraud suspected or not and then fed to the Weka tool and the J48 data mining algorithm is run in different scenarios. J48 has different parameters (like confidence Factor, minNumObj, reduced error pruning, 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, the algorithm was used with default values and also tested by changing these values to see the changes in the classification accuracy.

At first, the classification model was modeled with the default variable values of the J48 classification algorithm. Table 4.1 outlines the default variables with their values for the J48 decision tree classification algorithm.

| Parameter | Description | Default Value |
|---|---|---|
| Confidence-factor | This is the pruning confidence factor(smaller values are pruned more) | 0.25 |
| minNumObj | This is the lowest number of occurrences per leaf | 2 |
| Un-pruned | Shows if pruning is performed | False |

Table 6:- Default variables for J48 Tree algorithm

Experiment 3

The first experiment was done using the default variables. The default 10-fold cross- validation investigation choice was applied to train the classification model. With the default variables, the classification model was built with a J48 decision tree of 17 leaves and a tree size of 25. The decision tree made use of all the attributes in the training set with the status variable being the determining variable. Table 4.2 below shows the confusion matrix of the model.

Table 5. Confusion matrix for experiment two, J48 classifier with different test mode

| Actual | Predicted | | |
|---|---|---|---|
| J48 classifier with 10fold cross validation test mode | Fraud | Non fraud | Total |
| Fraud | 2183 | 540 | 2723 |
| NonFraud | 4668 | 642 | 5310 |
| J48 classifier with percentage split(66%) test mode | | | |
| Fraud | 960 | 371 | 1283 |
| NonFraud | 3175 | 552 | 3727 |

|  | Result | Percentage | Time |
|---|---|---|---|
| J48 classifier with 10 fold cross-validation test mode |  |  |  |
| Correctly classified | 7548 | 93.96% | 0.04sec |
| Incorrectly classified | 485 | 6.04% |  |
| J48 Classifier with percentage split(66%) test mode |  |  |  |
| Correctly classified | 4022 | 80.27% | 0.04sec |
| Incorrectly classified | 988 | 19.73% |  |

Table 7:- Summary of Classification results for experiment 2, J48 classifier with different test options

From the above tables, it can be seen that the j48 classifier with a 10-fold cross-validation test mode has a better classification accuracy of 93.96%. 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 them to the dataset. This increases the probability of correctly classifying minority classes [32].

### 4.1.3 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 works of literature, 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 community-based health insurance claim fraud detection by applying them in similar data sets but with different scenarios, as discussed in detail above. 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 community-based health insurance(CBHI) claim fraud detection in this research to proceed with the development of a prototype. The comparison based on classification accuracy and Performance is summarized and shown below. In all of the experiments, better classification accuracy is achieved when a 10-fold cross-validation test mode is applied. Therefore the result of this experiment is taken for comparison.

| Dataset | Classification model | Correctly | Misclassified | Better classifie |
|---|---|---|---|---|
| Full dataset | J48 classifier | 93.96% | 6.04% | J48 Classifier |
| | Naïve Bayes | 89.42% | 10.58% | |

Table 8:- Comparison of naïve Bayes classifier with j48 classifier

As it is shown in table 4.4 J48 decision tree has better classification accuracy. In addition to the above metrics, the time taken to build the model can be used as a measure of the performance of the classification model. The time taken to build the 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.

Decision trees are a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that a tree in which each node symbolizes a point at which a choice must be taken depending on the input, and we travel from node to node until we reach a leaf that gives you the projected output.

The decision tree generates a tree with branches, nodes, and leaves that allows us to move an unknown data point down the tree, applying the data point's characteristics to the tree until we reach a leaf and can identify the data point's unknown output. We discovered that in order to design a good classification tree model, we must first have a data set with known outputs on which to base our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

The following figure is the snapshot of the decision tree from experiments done in weka.

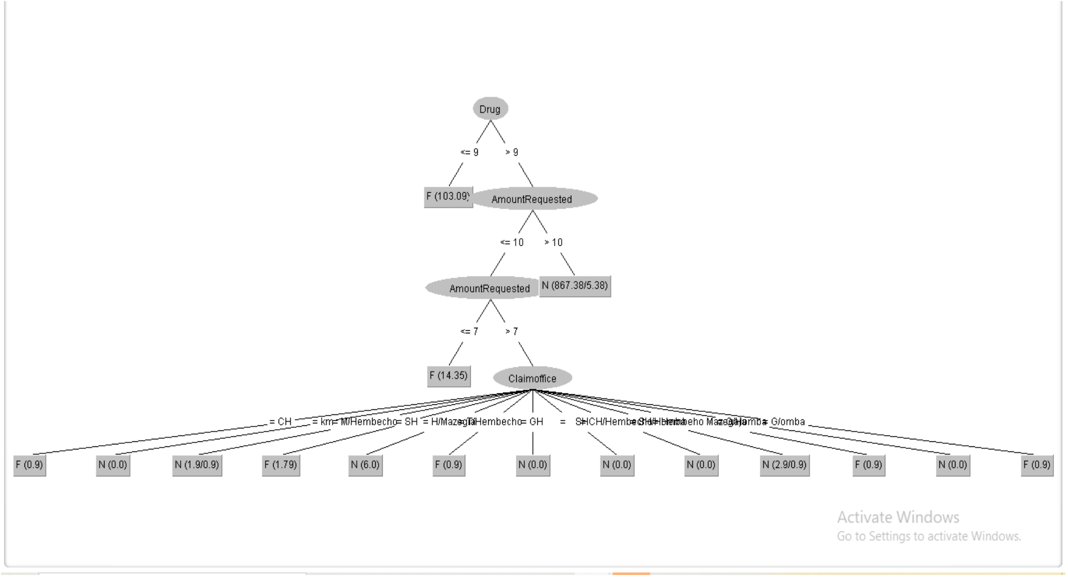
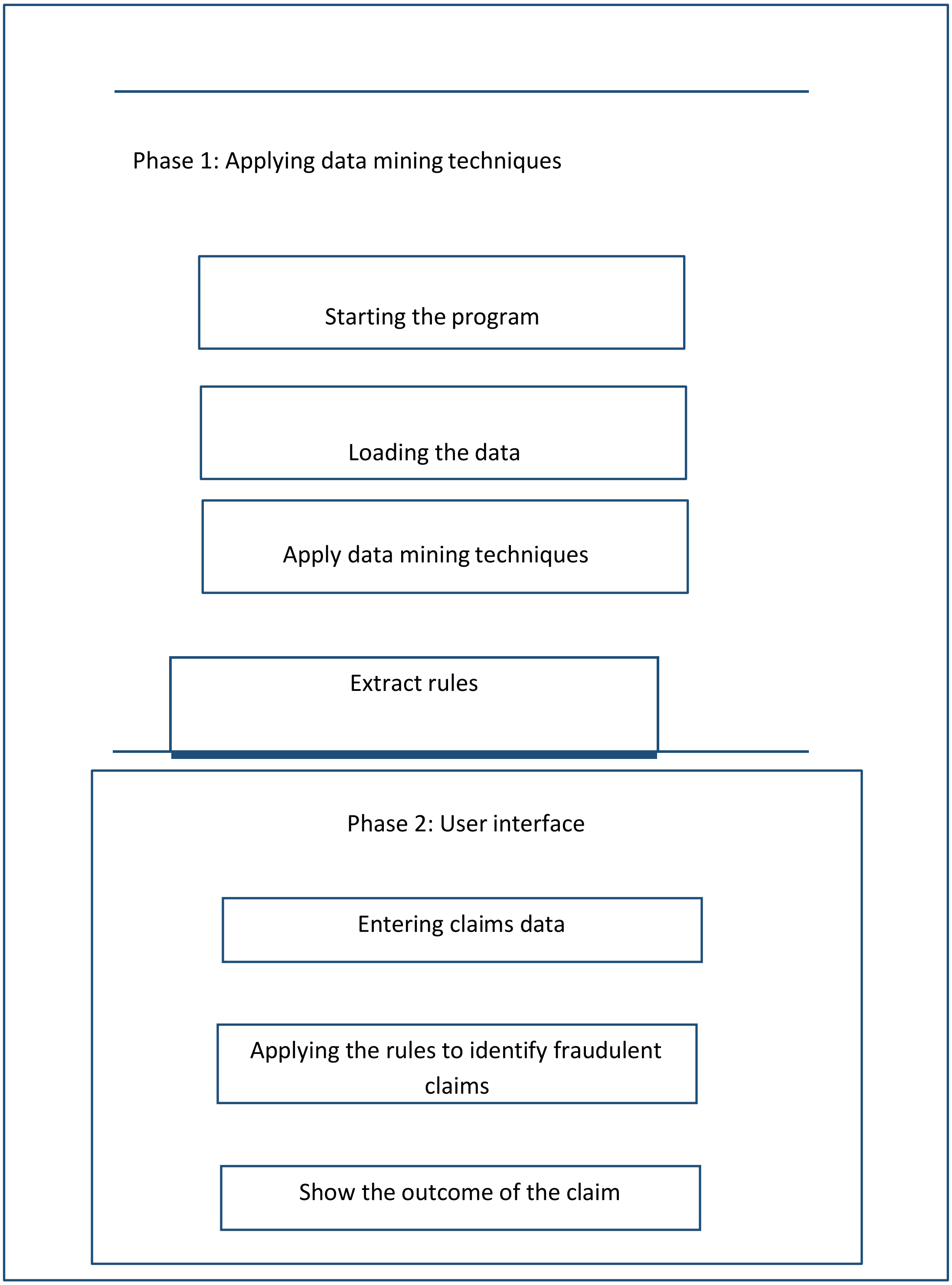


Figure 9:-Snapshot of decision tree from experiments done in weka

## 4.4. Graphical User Interface Development

For the development of a graphical user interface, Microsoft vision 2007 is used. This prototype graphical user interface is developed based on the best models chosen above for the classification of the datasets using selected attributes. Before developing the graphical interface, the researcher is generated rules by using one of the classifiers called the J48 algorithm. The rules used by the researcher are important to design the graphical user interface for predicting occurrences of fraud and non-fraud CBHI documents in a patient‘s record. Thus, the following figure showed the graphical user interface developed for predicting occurrences of fraud and non-fraud CBHI documents inpatient records.

Figure 10:-Figure 10:-Prototype implementation diagram


Phase 1: Applying data mining techniques
Starting the program
Loading the data
Apply data mining techniques
Extract rules
Phase 2: User interface
Entering claims data
Applying the rules to identify fraudulent claims
Show the outcome of the claim

# CHAPTER FIVE

## 5. Conclusion and Recommendation

### 5.1. Conclusion

This research attempted to create a new fraud detection model for community-based health insurance (CBHI) claims processing based on classification algorithms like J48, decision trees.

CBHI 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 West gojjam zone health sectors. For this purpose data is collected from the domain experts and officers of zonal health office in which the corporation is currently using to process the operational transactions. The final preprocessed data used for the experiment contains 8033 records and 11 variables or attributes. In order to conduct this research a six step hybrid data mining model developed by Cios et.al(Cios et al., 2007) is used. Understanding and Preparation of the data took too much time of the research duration. The data collected initially is from the health centers and zonal health. 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 CBHI(community based health insurance) were not registered in the system.

## 5.2. Recommendation

In this research, a number of tasks are completed to find the possible applicability of data mining technology on occurrences of fraud documents in CBHI patient user records. Even though the study, which is conducted has dedicated to the academic exercise, its results are hopeful to be applied in addressing practical problems on prevention and control of infectious disease. The researcher forwarded the following recommendations based on the result of this study:

- The researcher has got information from literature on how much is spent on fraudulent claims and how much of them are for valid claims in case of different region in West Gojjam zone, but there is no such information in Ethiopia. Therefore, the researcher also recommends an overall analysis could be one research area in this regard.
- Due to the shortness period of the CHBI that is started in Ethiopia. The validity of fraud documents available in the zonal health officers was scarce. Thus, for future trying to collect documents by adding this year data it will increase the accuracy of the prediction model.
- The samples do not represented all users of CBHI the country, because it is collected only from West Gojjam Health Bureau, funselam hospital and several CBHI Schemes clinics found in zone.
- However, experiments on data mining require very huge representative sample datasets from different regions of the country and involve private hospitals for collecting samples to get highly effective results which makes the fraudulent and treatment activities are time saver and accurate. Thus, we would like to recommend researchers who wants to work on this area to use the data from different zones and hospitals.

# References

[1] Deogan. (2011). Data Mining: research Trends, Challenges, and Applications [database ontheInternet].http://citeseer.nj.nec.com/deogun97data.html [Accessed on February, 21, 2019].

[2] F.T. (2006). ―Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining," International Conference on Information Acquisition, vol. 1, no. 1.

[3] R. Agrawal. (2015)."Mining Association Rules between Sets of Items in Large Databases." in In Proceedings of SIGMOD, 20716, 1993.According to (Sutch, T. (2015). Using association rules to understand subject choice at AS/A level., 2015.

[4] Witten, I. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann publishers.

[5] T. Sutch. (2015)."Using association rules to understand subject choice at AS/A level.," Cambridge.

[6] Fayyad U, Piatetsky-Shapiro, G., and Smyth, Padharic. From Data Mining to Knowledge Discovery in Databases. (1996). database on the Internet.[Cited April 15,2019].

[7]Mary, K. &Obenshain, M. (2004).Application of Data Mining Techniques to Healthcare Data. Chicago Journals The Society for Healthcare Epidemiology of America, pp. 690-695

[8] Jiawei al et. (2012). Data Mining Concepts and Techniques,3rd Edition. Waltham, USA: Morgan Kaufmann Publisher.

[9] Tariku, A. (2011). *Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company* (Doctoral dissertation, Addis Ababa University).

[10] Mirchaye, M. (2015). *Predictive Model For Medical Insurance Fraud Detection: The Case Of Ethiopian Insurance Corporation* (Doctoral dissertation, Addis Ababa University).

[11]Yihenew, F. (2015).*Constructing predictive model using Data mining techniques in support of Motor insurance policy risk Assessment: the case of Ethiopian Insurance Corporation (EIC)* (Doctoral dissertation, AAU).

[12] Molla,H.Ololo, S., &Megersa, B. (2014). Willingness to join community-based health insurance among rural households of Debub Bench District, Bench Maji Zone, Southwest Ethiopia.*BMC Public Health*, *14*(1), 591.

[13] Anagaw,M. A. D., Sparrow, R., Yilma, Z., Abebaw, D., Alemu, G., &Bedi, A. (2013). Impact of Ethiopian pilot community-based health insurance scheme on health- care utilization: a household panel data analysis. *The Lancet*, *381*, S92.

[14] Rawte, V., &Anuradha, G. (2015, January). Fraud detection in health insurance using data mining techniques.In *2015 International Conference on Communication, Information & Computing Technology (ICCICT)* (pp. 1-5).IEEE.

[15] Lavers, T. (2019). Social protection in an aspiring ‗developmental state': The political drivers of Ethiopia's PSNP. *African Affairs*, *118*(473), 646-671.

[16] Mariam, D. H., &Asnake, M. (2010). 25 Years of Public Health Leadership in Africa: The Ethiopian Public Health Association. *Canadian Journal of Public Health*, *101*(6), 445-446.

[17] Sharrifa,M.R. (2019). *Use of Data Mining To Detect Fraud Health Insurance claims* (Doctoral dissertation, University of Nairobi).

[18] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0:step-by-step data mining guide. Copenhagen: SPSS.

[19] Peter C., Pablo H, Rolf S, Jaap V. and Alessandro Z. (1998). Discovering Data Mining: From Concept to Implementation. Prentice-Hall, Upper Saddle River, NJ.

[20] SPSS, (2004). Improving Tax Administration with Data Mining.Executive

report. Accessed on 20/11/2011 from www.spsslietuva.com/media/collateral/modeli ng/tax.pdf.

[21] Santos, M &Azevedo, C (2005). Data Mining – Descoberta de Conhecimentoem Bases deDados. FCA Publisher.

[22] Cios. 2007. Data Mining: A Knowledge Discovery Approach. Springer, New

[23] Han, J, and Kamber, M 2006, Data Mining: Concepts and Techniques, 2nd ed, Morgan kufman Publishers, San Francisco

[24] Tamrat B., "Fixed Line Telephone Service Failure and Recovery: A case study in Ethiopian Telecommunication Corporation," Master of Business Page | 87 Administration, School of Business and Public Administration Addis Ababa University, Addis Ababa, 2010.

[25] Berry, M. and Linoff, G. (1997). Data mining techniques: For marketing, sales, and customer support. New York. John Wiley and Sons, Inc.

[26] Two Crows Corporation 2005, Introduction to Data Mining and Knowledge Discovery, 3 rdedn, Two Crows Corporation, Potomac: U.S.A.

[27] Hajizadeh, E Ardakani, DandShahrabi, J 2010, Application of Data Mining Techniques in Stock Markets: A Survey‟, Journal of Economics and International Finance, Vol. 2(7), pp. 109- 118.

[28] Apte,C, and Weiss, M 1997, Data Mining with Decision Trees and Decision Rules, Future Generation Computer Systems, New York. www.research.ibm.com/dar/papers/pdf/fgcsapteweiss_with_cover.pdf Access Date: September 20, 2014.

[29] Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. IEEE Transactions on Systems, Man, and Cybernetics, PartC 35 (4): 476–487.

[30] Bharti, K Jain, S and Shukla, S (2010) Fuzzy K-mean Clustering Via J48 for Intrusion Detection System.International Journal of Computer Science and Information,1(4).

[31] D. VenugopalSetty, T.M.Rangaswamy, and K.N.Subramanya. (2010). A Review on Data Mining Applications to the Performance of Stock Marketing. In: International Journal of Computer Applications (0975 – 8887. Volume 1 – No. 3.

[32] Mrs. Keerti. S. Mahajan& R. V. Kulkarni. (2013). A Review:Application of Data Mining Tools for Stock Market. In: Keerti S Mahajan et al,Int.J. Computer

Technology

&Applications,Vol 4 (1), 19-27. [33] Dr. M.Hanumanthappa1 &Sarakutty.T.K (2011).Predicting the Future of Car Manufacturing Industry using Data Mining Techniques.In: ACEEE Int. J. on Information Technology, Vol. 01, No. 02.

[34] QASEM A. AL-RADAIDEH, ADEL ABU ASSAF & EMAN ALNAGI. Predicting Stock Prices Using Data Mining Techniques (2013). In: The International Arab Conference on Information Technology.

[35] Magdalena Daniela NEMEŞ &Alexandru BUTOI. (2013). Data Mining on Romanian Stock Market Using Neural Networks for Price Prediction. In: Informatics Economică; vol. 17, No. 3/2013

[36] ValtteriAhti.(2009).Forecasting Commodity Prices with Nonlinear Models. In: HECER Discussion Paper No. 268.

[37] Dr. G.ManojSomeswar, B. Satheesh, and G.Vivekanand.(2009).Finance Mining – Analysis of Stock Market Exchange for Foreign Using Classification Techniques. In: International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 4, pp.717-723.

[38] Ramesh A. Medar and Vijay. S. Rajpurohit. (2014). Data mining in Agriculture on Crop Price Prediction Techniques and Applications. From: International Journal of Computer Applications (0975 – 8887) Vol 99, No.12.

[39] Andres M. Ticlavilca, Dillon M. Feuz, and Mac McKee.(2010).Forecasting Agricultural Commodity Prices Using Multivariate Bayesian Machine Learning Regression. In: Paper presented at the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management.

[40] G.M.Nasira and N.Hemageetha. (2012).Forecasting Model for Vegetable Price Using Back

Propagation Neural Network. In: International Journal of Computational Intelligence and Informatics, Vol. 2: No. 2.

[41] Kumar, V. (2005). Parallel and distributed computing for cybersecurity. IEEE Distributed Systems Online, 6(10), 1-9.

[42] Frankfurt,C.&Cuervo,L.G.(2017).E-Government as a tool to advance health. Global Journal of Medicine and Public Health.Vo5

[43] Rawte, V.Anurada, G.(2015,January). Fraud detection in health insurance using data mining techniques.In 2015 International conference on communication, information and computing technology (ICCICT) (pp.1-5).IEEE.

[43]Abubakar S. Magaji and Adeboye K.R. (2014). An Intense Nigerian Stock Exchange Market Prediction: Using Logistic with Back-Propagation ANN Model. In: Science World Journal Vol 9 (No 2).

[43] Wilma Latuny. (2012). Predicting the Selling Price of Dried EUCHEUMACOTTONII in Indonesia with Four Classifiers of Data Mining Techniques. In: ARIKA, Vol. 06, No. 2, ISSN: 1978-1105.

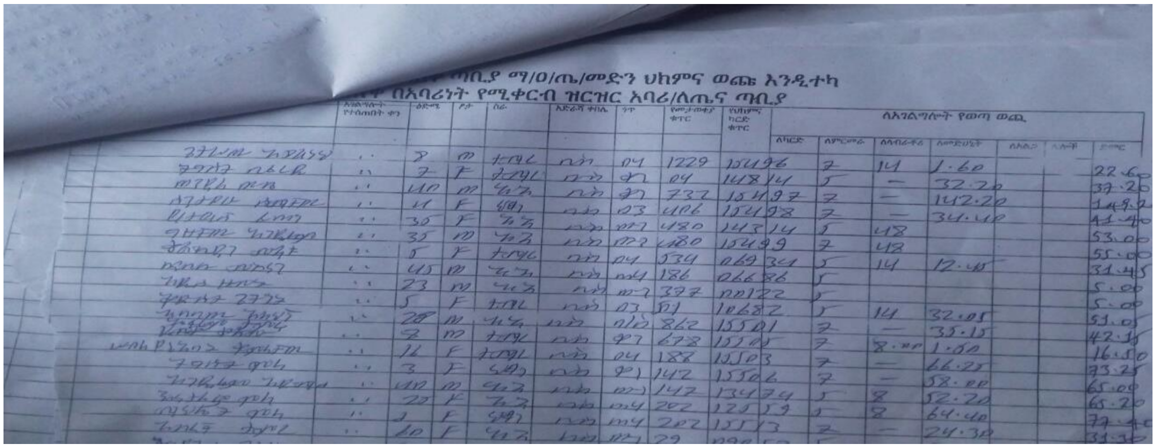
[44] Adenusi, O. 2011. ―Community Health Insurance Schemes—Kwara/Hygeia.‖Conference presentation.

# Appendix

```
=== Run information ===


Scheme:          weka.classifiers.trees.J48 -C 0.25 -M 2


Relation:     TrainData-weka.filters.unsupervised.attribute.Remove-R1-2,5-6

Instances:    1001

Attributes:  8 Delay To notify


              Paying & none paying Claim office


              Cons
lab
              Drug Amount Requested


              Status


Test mode:    evaluate on training data


=== Classifier model (full training set) ===
J48 pruned tree
------------------
Drug <= 9: F (103.09)
Drug > 9
|    Amount Requested<= 10
|    |    Amount Requested<= 7: F (14.35)
|    |    Amount Requested> 7
|    |    |    Claim office = CH: F (0.9)
|    |    |    Claim office = km: N (0.0)
|    |    |    Claim office = M/Hembecho: N (1.9/0.9)
|    |    |    Claim office = SH: F (1.79)
|    |    |    Claim office = H/Mazegia: N (6.0)
|    |    |    Claim office = T/Hembecho: F (0.9)
|    |    |    Claim office = GH: N (0.0)
|  |  |  Claim office =   SH: N (0.0)

|  |  |  Claim office = CH/Hembecho: N (0.0)

|  |  |  Claim office = SH/Homba: N (2.9/0.9)

|  |  |  Claim office = HembehoMazegaja: F (0.9)

|  |  |  Claim office = G/Homba: N (0.0)

|  |  |  Claim office = G/omba: F (0.9)

|    Amount Requested> 10: N (867.38/5.38)


Number of Leaves  :                                          16


Size of the tree:                                            20


Time taken to build model: 0.07 seconds


=== Evaluation on training set ===


Time is taken to test model on training data: 0.06 seconds
```

```
=== Summary ===

Correctly Classified Instances         994               99.3007 %

Incorrectly Classified Instances         7                0.6993 %

Kappa statistic                          0.9683

Mean absolute error                      0.0129

Root mean squared error                  0.0763
```

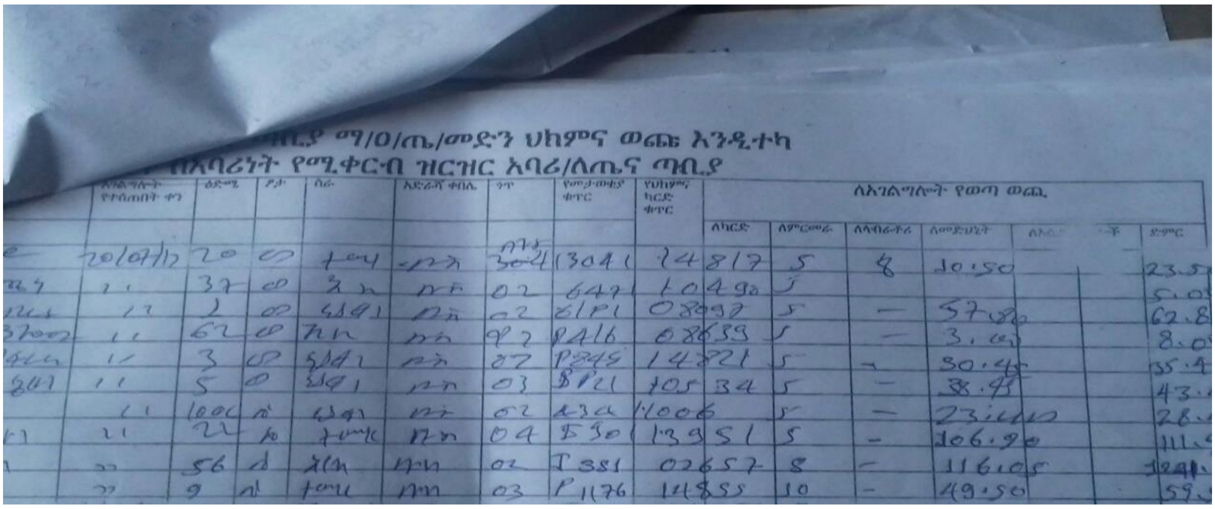
# Appendix B



ጣቢያ ማ/ዐ/ጤ/መድን ህክምና ወጪ እንዲተካ

በአባሪነት የሚቀርብ ዝርዝር አባሪ/ለጤና ጣቢያ

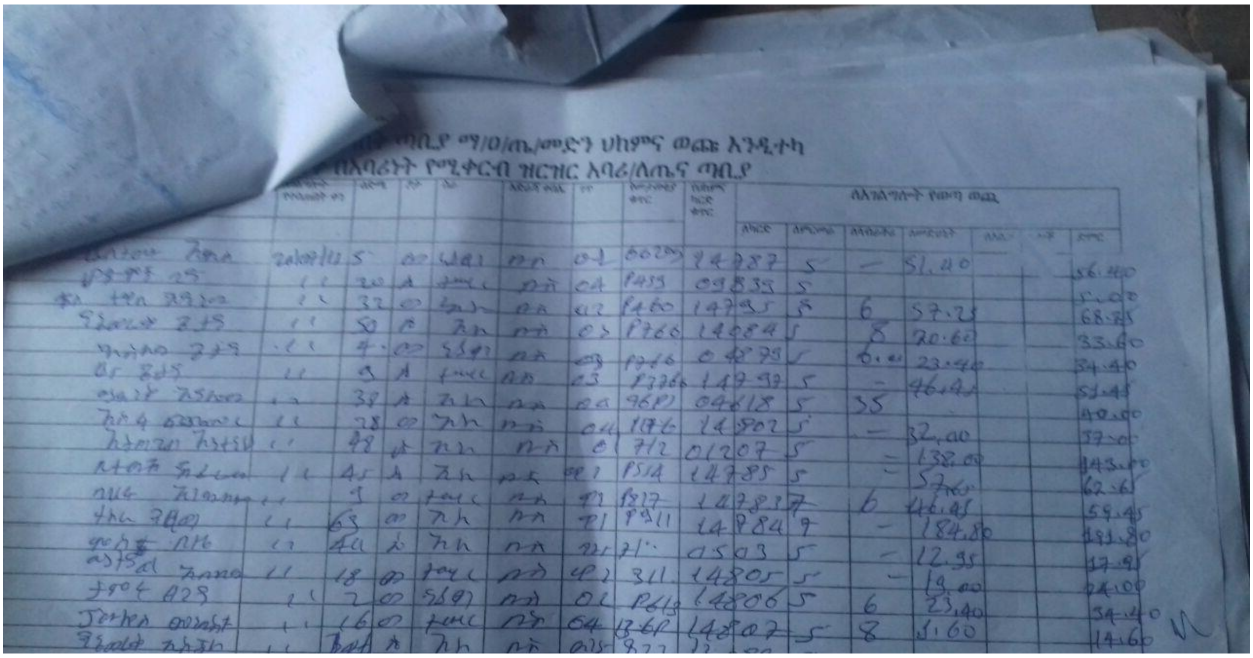
| | አገልግሎት የተሰጠበት ቀን | ዕድሜ | ጾታ | ስራ | አድራሻ ቀበሌ | ጎጥ | የመታወቂያ ቁጥር | የህክምና ካርድ ቁጥር | ለአገልግሎት የወጣ ወጪ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | ለካርድ | ለምርመራ | ለላብራቶሪ | ለመድሀኒት | ለአልጋ | ሌሎች | ድምር |
| [illegible] | 〃 | 7 | M | [illegible] | [illegible] | [illegible] | 1229 | 15496 | | 7 | 14 | 1.60 | | | 22.60 |
| [illegible] | 〃 | 7 | F | [illegible] | [illegible] | [illegible] | 04 | 14814 | | 5 | — | 32.20 | | | 37.20 |
| [illegible] | 〃 | 40 | M | [illegible] | [illegible] | [illegible] | 732 | 15497 | | 7 | — | 142.20 | | | 149.20 |
| [illegible] | 〃 | 11 | F | [illegible] | [illegible] | [illegible] | 406 | 15498 | | 7 | — | 34.40 | | | 41.40 |
| [illegible] | 〃 | 35 | F | [illegible] | [illegible] | [illegible] | 480 | 14314 | | 5 | 48 | | | | 53.00 |
| [illegible] | 〃 | 35 | M | [illegible] | [illegible] | [illegible] | 480 | 15499 | | 7 | 48 | | | | 55.00 |
| [illegible] | 〃 | 5 | F | [illegible] | [illegible] | [illegible] | 534 | 06934 | | 5 | 14 | 12.45 | | | 31.45 |
| [illegible] | 〃 | 45 | M | [illegible] | [illegible] | [illegible] | 186 | 06686 | | 5 | | | | | 5.00 |
| [illegible] | 〃 | 23 | M | [illegible] | [illegible] | [illegible] | 377 | 00122 | | 5 | | | | | 5.00 |
| [illegible] | 〃 | 5 | F | [illegible] | [illegible] | [illegible] | 511 | 10682 | | 5 | 14 | 32.05 | | | 51.05 |
| [illegible] | 〃 | 28 | M | [illegible] | [illegible] | [illegible] | 862 | 15501 | | 7 | — | 35.15 | | | 42.15 |
| [illegible] | 〃 | 2 | M | [illegible] | [illegible] | [illegible] | 678 | 15505 | | 7 | 8.00 | 1.50 | | | 16.50 |
| [illegible] | 〃 | 16 | F | [illegible] | [illegible] | [illegible] | 188 | 15503 | | 7 | — | 66.25 | | | 73.25 |
| [illegible] | 〃 | 3 | F | [illegible] | [illegible] | [illegible] | 142 | 15506 | | 7 | — | 58.00 | | | 65.00 |
| [illegible] | 〃 | 40 | M | [illegible] | [illegible] | [illegible] | 147 | 13474 | | 5 | 8 | 52.20 | | | 65.20 |
| [illegible] | 〃 | 20 | F | [illegible] | [illegible] | [illegible] | 202 | 12559 | | 5 | 8 | 64.40 | | | 77.40 |
| [illegible] | 〃 | 2 | F | [illegible] | [illegible] | [illegible] | 202 | 15513 | | 7 | — | 24.30 | | | 31.30 |
| [illegible] | 〃 | 60 | F | [illegible] | [illegible] | [illegible] | 29 | [illegible] | | [illegible] | | | | | [illegible] |



...ቢያ ማ/ዐ/ጤ/መድን ህክምና ወጪ እንዲተካ

...በአባሪነት የሚቀርብ ዝርዝር አባሪ/ለጤና ጣቢያ

| | አገልግሎት የተሰጠበት ቀን | ዕድሜ | ፆታ | ስራ | አድራሻ ቀበሌ | ጎጥ | የመታወቂያ ቁጥር | የህክምና ካርድ ቁጥር | ለአገልግሎት የወጣ ወጪ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | ለካርድ | ለምርመራ | ለላብራቶሪ | ለመድሀኒት | ለአ... | ...ች | ድምር |
| [illegible] | 20/07/12 | 20 | ወ | [illegible] | [illegible] | [illegible] | 13041 | 14817 | | 5 | 8 | 10.50 | | | 23.5 |
| [illegible] | " | 37 | ወ | [illegible] | [illegible] | 02 | 647 | 10490 | | 5 | | | | | 5.0 |
| [illegible] | " | 2 | ወ | [illegible] | [illegible] | 02 | [illegible] | 08097 | | 5 | — | 57.80 | | | 62.8 |
| [illegible] | " | 62 | ወ | [illegible] | [illegible] | [illegible] | [illegible] | 08639 | | 5 | — | 3.00 | | | 8.0 |
| [illegible] | " | 3 | ወ | [illegible] | [illegible] | 07 | P849 | 14821 | | 5 | — | 30.45 | | | 35.4 |
| [illegible] | " | 5 | ወ | [illegible] | [illegible] | 03 | [illegible] | 10534 | | 5 | — | 38.45 | | | 43. |
| | " | 100 | [illegible] | [illegible] | [illegible] | 02 | [illegible] | 11006 | | 5 | — | 23.00 | | | 28. |
| [illegible] | " | 22 | [illegible] | [illegible] | [illegible] | 04 | [illegible]530 | 13951 | | 5 | — | 106.90 | | | 111. |
| [illegible] | " | 56 | [illegible] | [illegible] | [illegible] | 02 | I 351 | 02657 | | 5 | — | 116.05 | | | 121. |
| | " | 9 | [illegible] | [illegible] | [illegible] | 03 | P 1176 | 14855 | | 10 | — | 49.50 | | | 59. |



...ቢያ ማ/ዐ/ጤ/መድን ህክምና ወጪ እንዲተካ

...በአባሪነት የሚቀርብ ዝርዝር አባሪ/ለጤና ጣቢያ

| | አገልግሎት የተሰጠበት ቀን | ዕድሜ | ፆታ | ስራ | አድራሻ ቀበሌ | ጎጥ | የመታወቂያ ቁጥር | የህክምና ካርድ ቁጥር | ለአገልግሎት የወጣ ወጪ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | ለካርድ | ለምርመራ | ለላብራቶሪ | ለመድሀኒት | ለአ... | ...ች | ድምር |
| [illegible] | 20/07/12 | 5 | ወ | [illegible] | [illegible] | 01 | 6627 | 14787 | | 5 | — | 51.40 | | | 56.40 |
| [illegible] | " | 20 | [illegible] | [illegible] | [illegible] | 04 | P439 | 09839 | | 5 | | | | | 5.00 |
| [illegible] | " | 32 | ወ | [illegible] | [illegible] | 02 | P460 | 14795 | | 5 | 6 | 57.25 | | | 68.25 |
| [illegible] | " | 50 | [illegible] | [illegible] | [illegible] | 03 | P766 | 14684 | | 5 | 8 | 20.60 | | | 33.60 |
| [illegible] | " | 4 | ወ | [illegible] | [illegible] | [illegible] | P916 | 04879 | | 5 | 6.00 | 23.40 | | | 34.40 |
| [illegible] | " | 9 | [illegible] | [illegible] | [illegible] | 03 | P336 | 14797 | | 5 | — | 46.45 | | | 51.45 |
| [illegible] | " | 39 | [illegible] | [illegible] | [illegible] | [illegible] | 96P | 04618 | | 5 | 35 | | | | 40.00 |
| [illegible] | " | 28 | ወ | [illegible] | [illegible] | 04 | 196 | 14802 | | 5 | — | 32.00 | | | 37.00 |
| [illegible] | " | 48 | [illegible] | [illegible] | [illegible] | 01 | 712 | 01207 | | 5 | — | 138.00 | | | 143.00 |
| [illegible] | " | 45 | [illegible] | [illegible] | [illegible] | [illegible] | P514 | 14785 | | 5 | — | 57.60 | | | 62.60 |
| [illegible] | " | 9 | ወ | [illegible] | [illegible] | [illegible] | P817 | 14783 | | 7 | 6 | 46.45 | | | 59.45 |
| [illegible] | " | 63 | ወ | [illegible] | [illegible] | [illegible] | P911 | 14784 | | 7 | — | 184.80 | | | 191.80 |
| [illegible] | " | 44 | [illegible] | [illegible] | [illegible] | [illegible] | 71 | 05103 | | 5 | — | 12.95 | | | 17.95 |
| [illegible] | " | 18 | ወ | [illegible] | [illegible] | [illegible] | 311 | 14805 | | 5 | — | 19.00 | | | 24.00 |
| [illegible] | " | 2 | ወ | [illegible] | [illegible] | 02 | P613 | 14806 | | 5 | 6 | 23.40 | | | 34.40 |
| [illegible] | " | 16 | ወ | [illegible] | [illegible] | 04 | 136P | 14807 | | 5 | 8 | 1.60 | | | 14.60 |
| [illegible] | " | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 877 | [illegible] | | | | | | | |