University of Nairobi

School of Computing and Informatics

USE OF DATA MINING TO DETECT FRAUD HEALTH INSURANCE CLAIMS

BY

Sharifa Rigga Mambo

P56/P/7861/2002

Supervisor

Christopher A. Moturi

May 2019

A Research Project Report Submitted in Partial Fulfillment of the Requirements for the Award of the Master of Science in Information Systems Degree of the University of Nairobi

DECLARATION

The research project presented in this project report is my original work and has not been presented in any other institution. Reference is hereby made from works of other researchers that may have more insight into this project.

Sharifa Rigga Mambo: _____ Date: _____
(P56/P/7861/2002)

This project report has been submitted in partial fulfillment of the requirement of the Master of Science Degree in Information Systems of the University of Nairobi with my approval as the University supervisor

Christopher A Moturi: _____ Date: _____
Director
ICT Centre

ABSTRACT

Background

The extent, possibility, and complexity of the healthcare industry has attracted widespread fraud that has an impact on the economy. The fraudulent activities not only contribute to the problem of rising healthcare expenditure but also affect the health of patients. The challenge in the current fraud detection systems mostly lies in realizing the burden of the money lost and the unusual behavior areas.

Problem

Despite putting up various technologies and strategies to fight fraud such as planned, targeted audits, and random audits, whistle blowing, and biometric systems, fraud in claims have continued to be a challenge in most of the health insurance providers in Kenya.

Purpose

This research tried to analyze the appropriateness of data mining techniques in detecting fraudulent health insurance claims.

Method

To achieve this, classification models were used to guide the entire knowledge discovery process. Classification algorithms i.e. Naïve Bayes, Decision tree and K-Nearest Neighbor were used to build predictive models.

Findings

Several experiments were conducted and the resulting models shows that the Naïve Bayes works well in detecting fraud in claims with 91.790% classification accuracy and 74.12% testing hit rate.

Value of Study

A prototype was developed based on the rules extracted from the Naïve Bayes model which, if adopted, will save costs by detecting fraud as it is committed.

Conclusion

Fraud detection in health insurance companies, is much needed in developed as well as undeveloped countries so as to help reduce loss of money and in return improve service delivery to patients.

Key Words: Healthcare, Health Insurance Claim, Fraud, Naïve Bayes classification, Data Mining

TABLE OF CONTENTS

List of Figures

List of Tables

Acronyms

KDD – Knowledge Discovery Databases

NHIF National Health Insurance Fund

IRA - Insurance Regulatory Authority

HIS - Health Information Systems

HMIS - Hospital Management Information Systems

MDGs - Millennium Development Goals

K-NN - K-Nearest-Neighbor

CRISP-DM - Cross-Industry Standard Process for Data Mining

WEKA - Waikato Environment for Knowledge Analysis

CHAPTER ONE
INTRODUCTION

This chapter covers the background, problem statement, objectives, research questions, scope and significance of the study.

1.1 Background

The size and dimension of corruption impacting health is immense. Corruption impacts worldwide health results thus causing monetary waste and dare health repercussions. Methods of corruption and their repercussion on the global health exist and can be felt worldwide in public and private sectors and in undeveloped and developed country setting and threaten future investment (Mackey & Liang, 2012). In Kenya fraud and abuse in medical claims has manifested the health insurance companies in Kenya in the last years and in return ripping off the government. Taxpayer-funded programs such as National Hospital Insurance Funds (www.nhif.or.ke) are among the biggest victims. Huge amounts of money are lost each year from fraudulent claims. 143 instances fraud in medical insurance were reported in 2012 in which a total of Ksh 253.6 million was lost. Out of that only Ksh5.2 million recovered (HealthInsurance Fraud SurveyReport, 2013). This makes healthcare fraud one of Kenya's largest taxpayer rip-offs. Kenya's medical insurance sector claims loss ratio is 5% higher than other insurance category. (Alniz, 2018). Health insurance fraud as described by the National Health-Care Anti-Fraud Association is the "purposeful submittance of untrue claims to private or public (tax funded) health-insurance programs". This can be anybody i.e. commercial enterprise, untrustworthy provider or payer employees can embezzle healthcare payments for personal benefits or gains. Providers can also give claims for prescriptions or medications that were never given or were only partly rendered. Dishonest health-care providers may also force patients to undergo unwanted medications just for them to increase the charges to the payers (Kyriakakis, 2015). Recently, ways of handling electronic claims have gradually been established to automatedly carry out audits and analysis of claim data. These systems are modeled for determining areas that need special attention like incorrect and insufficient data input, matching claims and medically non-covered services. Despite these systems being used to detect fraud, their fraud discovery abilities are usually minimal since the discovery mainly depends on pre-defined procedures specified by the fraud field experts. (Pawar, 2016).

Data-mining can be used, where based on previous knowledge and experience of data we can relate things happening around us or sometimes even detect something out of the data at hand

(Mukherjee S et al, 2015). Prompt prevention and detection of fraud will go a long way by providing significant cost reductions to insurance companies there by reduction to the rising cost of healthcare.

1.2 Statement of the Problem

Despite putting up different technologies and strategies to fight fraud such as planned and targeted audits, random audits, whistle blowing, computerized systems and biometric systems fraud in claims have continued to be a challenge. And it remains to be a limiting factor in the delivery of quality healthcare services. The fraud hit rate is still high in most health insurance companies because the various computerized systems which are expected to help detect undesired behavior still relies on experts' experiences in selecting statistically significant features to detect the fraudulent claims.

1.3 Objectives of the study

1.3.1 Main objective

The main objective of this study was to develop a data mining prototype to detect fraudulent health-insurance claims

1.3.2 The specific objectives

1. To identify fraudulent activities perpetrated in health-insurance claims.
2. To explore data mining techniques that could be used to manage fraudulent activities in health-insurance
3. To develop and test a data mining prototype for detecting fraudulent health insurance claims.

1.4 Research questions

a) What types of fraudulent activities are perpetrated against health insurance claims?
b) How is data mining used in health-insurance?
c) What data mining techniques could be used to discover fraud in health insurance claims?
d) How can a data-mining prototype be used to detect fraudulent health-insurance claims?

1.5 Significance

The research will add value to the healthcare fraud domain by giving insights and recommendations on how to design, develop and adopt fraud prediction/detection systems using data mining methodologies. This will prompt early discovery of fraudulent billings in health insurance companies while at the same time minimizing the wrong positives in choosing of providers for audit. The aim is to reduce costs in healthcare thus in return enable better health care that is accessible. When fully applied, the prototype will enable health insurers to identify the perpetrators of fraud and subject them to the legal process.

1.6 Scope

The study used medical claims obtained online (https://www.cms.gov) covering the year 2010. This dataset was chosen because it was the most complete dataset with fewer missing fields.

1.7 Assumption

The research exists within a well governed health domain and as such makes several basic assumptions. Key is that the claims dataset used adhered to the International health standards and ICD coding.

CHAPTER TWO

LITERATUREREVIEW

This chapter outlines various research studies on fraud discovery using data mining techniques. The chapter forms a basis for conceptual framework for this research study.

2.1 Health Insurance in Kenya

Health insurance is a group insurance method, which allows people to give installment payments or taxes so as secure their selves from paying expected or unanticipated healthcare costs (Patil, 2012).The total risk of healthcare costs is estimated in health insurance, and then a regular finance arrangement (like a monthly payment or an annual payment) is developed and adopted. This routine finance structure ensures that the money contributed is available to cater for any health benefits that are indicated in the agreement cover of the insurance. Mostly, the healthcare payment is managed by a primary institution, which can be a governmental institution, a non-governmental institution or a nonprofit organization that is managing a health plan. The uptake of health-insurance cover among Kenyans is rising owing to increased availability and affordability of health insurance covers. Currently there are two types of health insurance in Kenya; Government sponsored health-insurance through the NHIF and private health insurance (KHF, 2016).

The Government sponsored NHIF gives health-insurance protection to citizens in the formal and informal sectors of work. Citizens in the formal bracket of work are required to remit an earnings rated monthly contributions that are automatically debited from their monthly salaries and taken to the NHIF fund by their employing company. The formal monthly contributions vary from Ksh 150 which is the for least income earners (monthly income of Ksh 6000 and less) to Ksh 1700 for the most income earners (monthly income 100,000 and above). Informal sector citizens are required to pay an optional monthly standard rate of Ksh 600 (Munge, Mulupi & Chuma, 2015).Private health insurance covers are offered by private insurance companies such as Britam, UAP, AAR insurance, Eagle Africa, heritage, first assurance, jubilee insurance, AON and many other insurance companies. Still with fewer figures, the private health Insurance has grown over the years In 2010 there were about 600,000 citizens covered by private insurance cover and the number had risen to up to 1.5 million in 2016 (KH F, 2016**)**.

Together both public and private health insurance coverabout8million Kenyans, 20%oftheKenyan entirepopulation. Toexpress the magnitudeofthesetwo health insurance programs, in2016thespendinglevelsfor health insurance is estimatedtoroughly40billion. 20% of the health insurance claims presented are always fraudulent (Alniz, 2018).

2.2 Fraud in Health Insurance

Over the years healthcare- insurance companies receive millions of claims from healthcare givers or providers. Out of the millions 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 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) 2016 report, general insurance business suffered a Ksh 2.1 billion in losses, with claims rising by 11.8% from Ksh 49.05 billion in 2015 to hit 54.86 billion in 2016. Corruption and fraud have been recognized as a major barrier to health care (Frankfurter &Cuervo, 2017). For instance, in the USA, healthcare insurance fraud is over 30 billion dollars annually and the situation in developing countries like India is not that different (Rawte&Anuradha, 2015).

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 to an insurance company is low. Professional fraud on the other side is mostly performed by organized groups of individuals with many false identities mostly targeting a number of 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 possible 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 need in orders to detect deception and misuse or corruption in the healthcare setting. A number of the health insurance systems in place depend on experienced humans to physically scrutinize insurance claims for establishment of the suspicious ones. This makes the system development and claim reviewing processes very tedious especially when dealing with a 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 audit and review of healthcare-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 curatively not covered services. All those systems are limited as they are applied to find or discover only specific categories of fraud.

So as to attain efficacious fraud detection, several researchers have endeavored to come up with complicated or advanced anti-fraud techniques which include predictive modelling, data analytics, automated red business rules, geographic data mapping and link analysis (Kizito, 2016). Predictive modelling 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 amidst the claims, transactions and people thus helping tie up distinctive components or actors and then establishes the degree of the connection between the parties thereby giving out information that may be used to define aspects that point to any possible fraudulent activity. Automated red business rules are procures that may be incorporated into an insurers root or main Information systems. These procedures are very fundamental in assisting to foresee certain types of doubtful claim activity in relation to past fraud through the recognition of inconsistencies or non-uniformity during the claims processing. Geo-mapping may assist insurers assess risk and uncover anomalies during underwriting and processing of claims. With Geo-mapping an insurer can ascertain that an incident truly happened where the person alleges it did. (Kizito, 2016).

In the developed counties, laws have been regulated for violation of health insurance. HIPAA conditions that health-insurance fraud and abuse is a federal lawbreaking crime that can have substantial punishment attached to it. Persons that are found guilty of healthcare-insurance fraud are entitled fir a sentence of up to ten (10) years in a national prison coupled with a considerable monetary fines.

2.3 Fraudulent Activities in Health Insurance Claims

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 really provided, carrying out medically wasteful or unneeded services so as to generate insurance compensations and falsifying a patients examination so as to validate tests, surgeries or any other medical routines that are not curatively 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; untrue claims, fluctuating or erratic invoicing of codes, exaggerated invoicing products and facility services, providing of wasteful curative services, replicated claims, exempted benefits and products costed as insured allowances or payments and claims from illegitimate service providers.

In the global scenario a survey on healthcare fraud investigations conducted by the Health Insurance Association of America private insurers revealed that a large number of healthcare fraud undertaking is affiliated to diagnosis (43%) and 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 and asking, giving and receiving of kickback.(HCFA, 2015).

2.4 Role of ICT in Health Insurance Fraud

With the conventional approaches of health-insurance deception and corruption discovery, a couple of auditors manage heaps of healthcare claims forms. As a matter of fact, 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 providers exploits. (Rashidian et al, 2012). The approach works to a certain degree if the claims are relatively small. Despite how, the prevailing threat with the expert assessment and analysis is that, not all the can be put to investigation as a person can only be able to analyze a few claims to identify any fraud in them. It has thus proved very difficult to use human power in the detection of health insurance fraud. (Swapnil, 2018). This scenario is up to now prevalent in many middle income and low income countries. (Copeland et al, 2012).

ICT plays a paramount part preventing and detecting fraud. It enables the monitoring and tracking of claims and makes it harder for fraudsters to falsify claims. With ICT one can have a 360 degree view and understanding of the entire process. Insurers have come to learn that no

stand-alone technology technique is adequate. Incorporation of technologies and techniques is necessary to discover unscrupulous and organized fraud. (SAS report survey, 2016). Health insurance has witnessed tremendous changes linked with the developments in ICT over the years. The emerging use of computerized systems and the many electronic health records has brought about arising prospects for better discovery and uncovering of abuse in health-insurance. Developments artificial intelligence and machine learning takes interest to programmed ways of detecting fraud. Consolidating programmed approaches and mathematical knowledge to a recent rising multifaceted field of science or knowldege know as Knowledge Discovery from Databases (KDD).

Biometric technologies, which are ICT-based human identification systems have strengthen efforts in fighting fraud in the health insurance (Owusuet al, 2018). The biometric shapes may vary and can be palm-prints, iris-scans, voice recognition, finger-prints, gait and face recognition. (Rawlson, 2015). Biometric technology can fight fraud in health insurance by: eliminating multiple identities and fake IDs through, eliminate impersonation and ghost patients by use biometric verification at the claim system and reduce fraudulent provider bills. Algorithms in data analytics technology can as well envisage a possible fraud and in return an expert examine them. In such a setting the process becomes simplified and can make use the experience of an expertise. (Swapnil, 2018). Block chain technology can also remove common roots of fraud in health-insurance industry by shifting health-insurance claims onto unmodifiable ledger. (Cloudia, 2018).

GPS functionality, Mobile devices and social media engagement are other forms of ICT that have greatly have an impact on the processing of health-insurance claims by companies and policy assessment by insurance agents. (Khristy, 2017).

From the literature, it shows that in order to reduce health-insurance fraud, there is need for the bringing into play of a combined solution of social and technical techniques consisting of online biometric registration of persons and authentication at the place of service health delivery; utilizing interdependent technologies like e-claims; and well-functioning policies like the use of practitioners in scrutinizing service providers' claims.

2.5 Health Information Systems in Kenya

Health Information Systems describes any system that epitomize, holds, administers, controls or disseminate data and information related to health of persons or the processes of the institution that operate within the health sector (Nutley, 2012). HIS is entrusted with the obligation of gathering routine data from various sources, collating, analyzing and

disseminating health information to all stakeholders' confirmative based policy making. The information that is distributed is used for developing and administration of health services and agendas. The need for attestation so as to achieve the Millennium Development Goals (MDGs), associated with the deepening need for bidirectional and polygonal donors to show and prove their grants towards health development has in a way created a significant demand for health information in Kenya. (MoH, 2015).

Hospital Management Information Systems (HMIS) is a vital unit of a governmental health information system and actions needed of hospital management to assist information creation and generation in Kenya are formulated in specific policy documents. (Measure, 2017**)** explains that a serious functional Health Information System (HIS) is responsible for giving superior type of information to be utilized for policy formulation at all the ranks of the health organization or system. ICT and Health are gradually becoming more closely related in Kenya. Kenya has shown to be pace setter in innovative ICT solutions in conventional and also in the health sector. (KHF, 2016).

2.6 Use of Data Mining in Fraud Detection

Borse&Maitre (2015) presents a brief insight about data mining, information of healthcare insurance frauds and highlights of the advantages of data mining technique (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 diving the data mining technologies into supervised and unsupervised (Phua et al 2010). From literature as viewed and assessed by (Travaille et al, 2011) in regards to the efforts in applying supervised and unsupervised data mining technologies in the health-care sector shows that, supervised classification is notably useful in identifying complex fraud strategies when the aspects describing those strategies are familiar, and training sets can be validated while unsupervised technologies are important in detecting possibly current strategies and unscrupulous actions to deal with the ever changing character of fraud offenders that employ new procedures and systems.

2.6.1 Supervised Data Mining Methods

Supervised techniques are normally used for grouping and projection objectives which include traditional statistical methods like support vector machine (SVM), neural networks, Bayesian networks discriminant analysis and regression analysis. These supervised approaches need certainty in the true or accurate grouping of the records (Rashidian et al., 2012). Decision tree,

genetic algorithms, Support Vector Machine and neural networks are examples of supervised data mining methodologies that have been applied to expose abuse and fraud in healthcare . (Hyunjung et al, 2012), (Kirlidog&Asuk, 2012).

2.6.2 Unsupervised Data Mining Methods

The unsupervised data mining techniques usually evaluate ones claim's 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 Apriori algorithm. It is able to remove patterns and association rules amongst records, recognize irregular record(s) or a bunch of a more or less equal records. Clustering, outlier detection (Capelleveen, 2013) and association rules are a few unsupervised data mining models that have been adapted to expose abuse and fraud in healthcare (Liu, 2014).

2.6.3 Data Mining Classification techniques

Classification data mining models can be linear or non-linear. Linear models are mathematical approaches or patterns take on linear association amongst the constants of the parameters that are being studied (webdocs.cs.ualberta.ca). Defined as "The rough calculation of a discriminant function or recession function using a hyper plane". It can widely be maximized by using easy methods or approaches, but does not satisfactory predicate a lot of the real domain challenges. Another category of classification designs or models i.e. non-linear models defined as "non-linear projecting models are mathematical models which do not take on or accept linear interconnection amongst the constants of the parameters that are being studied" (statsoft.com). In spite of their mathematical straightforwardness and stability, linear models have clear possible difficulty. The definite step or procedure may not be straight and such a presumption presents unrealisable incline into the predictings.

2.6.3.1 Naïve Bayes

The Naïve Bayesian classifier is a clear chancy classifier centered on implementing Bayesian theorem (from Bayesian statistics) with firm (naïve) objectivity expectations. Bayes theorem is used by the Naïve Bayes algorithm for it classification. This is by calculating the possibilities of the feature values for each of the classification group and using the possibilities to guess the class of the undefined instances (McCallum & Nigam, 1998). The

Naïve Bayes classifier method is specifically adapted if the dimension of entered data higher. Naïve Bayesian classifier supposes that the elements or attributes characters are hypothetically independent and there consists no dependence associations amongst the attributes. This make Naïve Bayes the most accurate classifier when the presumption holds true (Han, Kamber&Pei, 2012)**.** The merit of Naïve Bayes classifier is, it only needs a limited quantity of the training data to compute the means' and variances' of the parameters required for classification. Being that independent parameters are hypothesized, only the variances' of the parameters within all the labels require to be affected and not the whole covariance matrix.

2.6.3.2 K-Nearest-Neighbor

The K-Nearest Neighbor classifier describes each ordered pair as a data point in a d-structural space where d is the count of the attributes. In such an arrangement, all the tuples are saved and reproduced during the learning stage. If a new tuple with an unknown class is given, the K-NN classifier analyses its closeness or nearness with the K-nearest training tuples and allocates the class of K-NN classifier with the most vote or distance weighted vote to the new tuple. Euclidean distance, Manhattan distance, simple matching coefficient, Jaccard similarity coefficient, cosine similarity, and correlation coefficient are few of the commonly used proximity measures for getting the nearest neighbor. A wide number of applications are present for K-NN such as pattern recognition. Image analysis, cluster analysis, prediction and economic forecasting.

2.6.3.3 Decision Tree

A Decision tree classifier is a tree design that comprises of the root, non-leaf nodes and leaf nodes. Each non leaf node represents an analysis on an attribute, each branch denotes the outcome of the analysis and each leaf node carries a class babel (Han, Kamber& Pei , 2012).The key conception of decision tree is to split the data repetitively into subcategories such that each subcategory comprises of almost identical states of intended variables (Gosain & Kumar, 2009). To select the diving formula, an attribute selection criterion is chosen in that it properly divides a given dataset. A few common decision tree algorithms include CART, C4.5 and ID3 that use Information gain, Gini Index and Gain Ratio as their attribute sampling standard measure respectively.

2.7 Fraud Detection Data Mining Research

A number of papers and researches relating to fraud discovery using data mining technologies were reviewed and discussed (Inês, 2017; Richard & Taghil, 2018; Joudaki et al., 2015 and Wei et al, 2013). Several investigators or academicians have devoted a considerable amount of work in exploring and investigating fraud detection models using data mining. Below are some of researches on fraud detection

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 effectiveness of applying machine learning with random under sampling to identify Medicare fraud. They employed random under sampling building four class arrangements. The results revealed that C4.5 decision tree and logistic regression learners gives 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 club. In FDiBC, based upon each score obtained by customer members of bank club, 14 features are derived, and then based on all the scores of each customer member, five sliding time and scores window-based feature vectors are proposed. For generating training and test dataset from the obtained scores of fraudster and common customers in the customers' club system of a bank, a positive and a negative label are used, respectively. After generating the training dataset, learning is performed through two approaches: 1) clustering and binary classification with the OCSVM method for positive data, i.e. fraudster customers, and 2) multi-class classification including SVM, C4.5, KNN, and Naïve Bayes methods. The results obtained reveal that FDiBC has the ability to detect fraud with 78% accuracy, and thus can be used in practice.

Inês (2017) created an artificial neural network which learns with insurance data and evolve continuously over time, anticipating fraudulent behaviors or actors, and contribute to institutions 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 in reference to 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 was verified and assessed in order to get one accepted technique to be adopted by the 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 at 74.58%.

Dharani & Shoba (2015) suggested a data mining technique or approach for discovering fraudulent prescriptions in a big prescription database, they 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 in regard to 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 dicovery of healthcare fraud. According to the outcomes, ID3 has the highest accuracy of 100% with J48 having the lowest accuracy of 96.7213%.

Joudaki et al., (2015) implemented a data mining methodology to a sizable health-insurance institution dataset 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 remorseless behavior. The outcomes also determined 2% of physicians as fraud culprits. Discriminant 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) have suggested a geo-location clustering design that's analyses the geo-location particulars of Medicare/Medicaid recipients and service givers 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 doubtfulness in levy or tax claims. In his research, he first applied the clustering algorithm to the dataset 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 J48 decision tree algorithm displayed the highest classification accuracy of 99.98%. The model was assessed with 2200 testing dataset and recorded a prediction accuracy of 97.19%

Wei et al, (2013) proposed a novel algorithm, to effectively extract variance patterns and determine unscrupulous 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 benchmarking fraud unearthing systems integrating domain knowledge and conventional fraud discovery techniques.

Eyad, (2012**)** developed an intelligent model that predicts and select suspicious water bills for discovering of fraudulent undertakings. The model improves the discovery hit rate of 1-10% ad hoc manual discovery to 80% intelligent discovery. Support Vector Classification technique (SVC) was applied in to uncover the irregular customers' load profile operations.

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

Ogwueleka, (2011) developed a neural network (NN) design for the credit 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 finest groupings of each record to its related group. The receiver operating curve (ROC) for credit 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.

2.8 Financial Fraud Detection

18 firms were used as a study reference by academicians from U.S.A and China in order to study and analyze data mining techniques. They recommended a general data mining financial fraud detection framework as is depicted in Figure 2.1

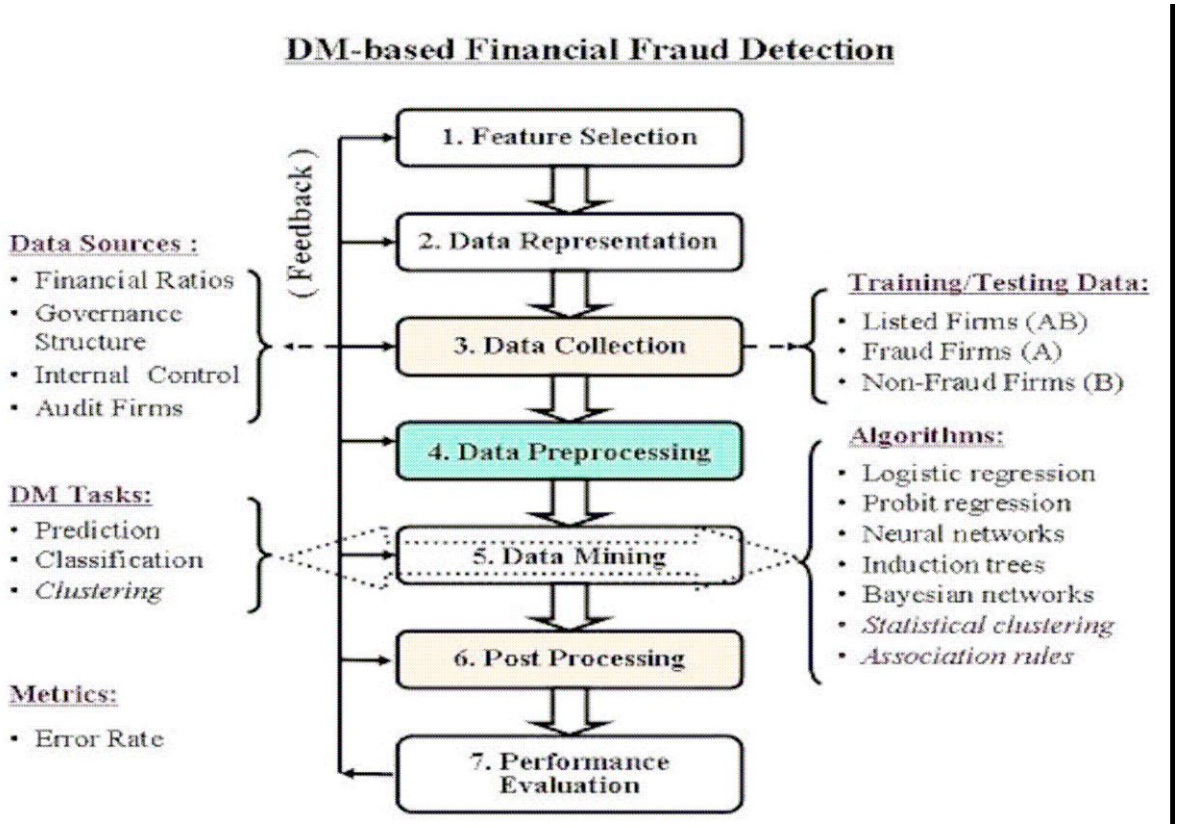

Figure 2.1: General Data mining Financial Fraud Detection Framework (Chang and C.-J. Lin, 2003)

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter covers research philosophy, research design, data collection and analysis methods, conceptual framework.

3.1 Research philosophy

In this study, we analyzed different data mining methodologies used for fraud discovery. Thus, the philosophy of positivism was adopted for collecting creditable and most appropriate data for the analysis and fit for the study area.

3.2 Research Design

Research design is considered as the design in which the research is conducted and contains the gathering, assessment or evaluation and analysis of data (Akhtar, 2016). Quantitative experimental research was used in this study.

3.3 Source of data

The data that was used for this research study was obtained from center for Medicare and medical services website on (https://www.cms.gov).The data had met the required standards in order for it to be usable. The dataset availed contains both outpatient and inpatient claims records for the period 2008 to 2010. The data was obtained in excel files format (Fig.3.1). We were unable to obtain Kenyan data from health insurance companies (both public and private) due to security and privacy and reasons.

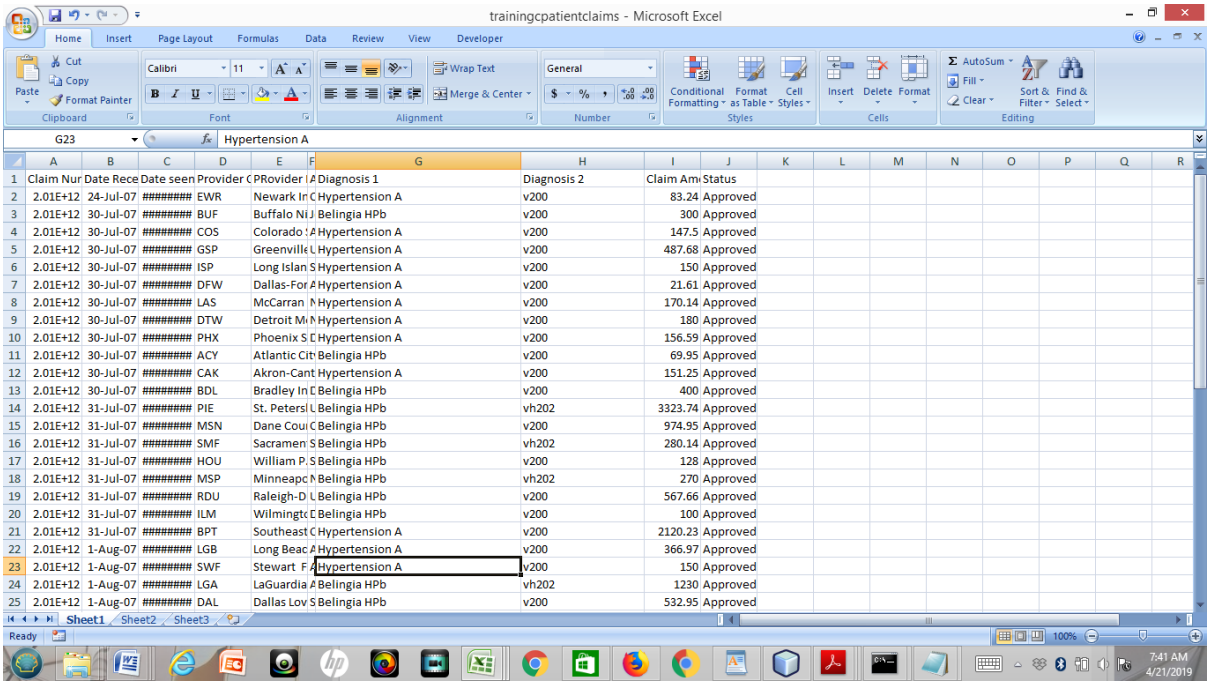

Figure 3.1: Claims Excel file

3.4 CRISP-DM Overview

CRISP-DM methodology was used for this research. Several data mining methodologies i.e. KDD and SEMMA do exist. CRISP-DM was adopted because of its adoption in data mining, flexibility and great deal of backtracking. CRISP-DM's prevalence went up from 42% in 2007 research to 43% in 2014 succeeding in being the utmost accepted data mining methodology (Gregory, 2014).

CRISP-DM stands for Cross Industry Standard Process for Data Mining. It was built by a consortium of data mining agents and organizations through an initiative sponsored by the European Commission.

The CRISP-DM methodology depicts a data mining method as a 6 phase cycle (see figure below) where the ordering of the phases is not fixed. CRISP-DM phases as adopted by the researcher are

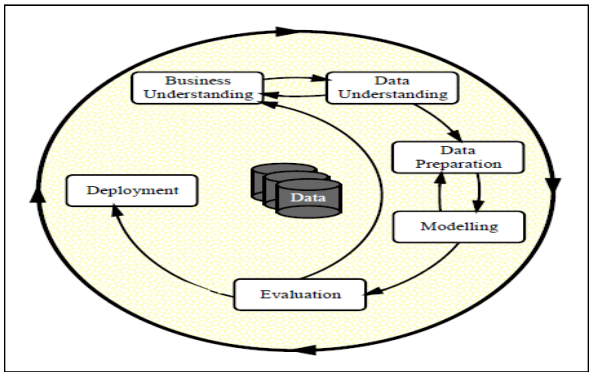

Figure3.2: CRISP-DM process (Source: Chapman et al. 2000)

3.4.1 Business Understanding

In this phase, in order to understand the problem primary and secondary sources were used. Secondary sources included books, local and international articles on data mining and specific to the area of fraud detection. The researcher focused on familiarizing the project goals and necessities from the health insurance point of view in terms of fraudulent activities perpetrated in health insurance claims. This helped to select the appropriate techniques for the research. The knowledge was then translated into a data-mining example description and an initial plan outlined to attain the goals.

3.4.2 Data Understanding

The data understanding phase started first with collection of the data and proceeded with processes of data familiarization, identifying the quality of the data and whether they are useful for the specific problem area or not, discovering initial understandings into the data and eventually discovering interesting subgroups from the data.

3.4.3 Data Preparation

In this phase the appropriate datasets perfect for the research was chosen. It was then cleaned, integrated, and formatted in order for it to be fit for use. Data cleansing activities done to the data included removing duplicate records, correcting noisy data, filling missing values by using estimates, removing irrelevant attributes and records i.e. attributes and records which are out of interest of the data mining problem. The cleaned data with the above techniques was further processed for dimensionality and numerocity reduction. For dimensionality reduction purpose the WEKA attribute selection (GainRatioAttributeEval attribute evaluator together with Ranker search method) is used. Finally the input data set for running the classification algorithms as training and test data was generated.

3.4.4 Modeling

Under this phase, a number of modeling technologies were picked and applied and their variables calibrated to the ideal values. Several data mining techniques exist for solving a data mining problem. K-Nearest Neighbor, decision tree and Naïve Bayes were applied in this research.

3.4.5 Evaluation

The models which were captured from the modeling phase were evaluated. Testing the classification accuracy was done by using the training set mode. The algorithms were tested on test data set to see how many of the test set are classified as true positive and false positive. The performances of the algorithms were also evaluated by percentage of classification and time parameter: the time taken to build the model.

3.4.6 Deployment

Finally how to use the discovered knowledge is seen by developing a prototype to test some claim for fraud. For development of the prototype Java platform was used. The environment was selected because the researcher is familiar with it.

3.5 Data Analysis Tools and Presentation

WEKA was adopted for mining of the data. The WEKA tool was chosen because of its functionality (support for many algorithms), familiarity of the researcher with the software and ease of use (has a graphical user interface) and runs on almost any platform.

WEKA (Waikato Environment for Knowledge Analysis) is open-source and free machine learning software that is used to extract data. The WEKA platform consists of a package of visualization tools and algorithms for data exploration and projective designing or modelling as well as a menu driven design for easy access. (HallMet al, 2009). Different algorithms are supported by WEKA i.e. classification, regression, decision trees and clustering. This tool enables users to instantly try out and contrast independent machine learning techniques on new data sets. Its flexible, extendable framework allows complicated data mining procedures to be modelled from the vast collection of base learning algorithm and tools given. (Hall M et al, 2009).

3.6 Prototype Analysis and Design

Since this is a prototype and not a fully functioning system, many areas were not implemented. Below is the research prototype context

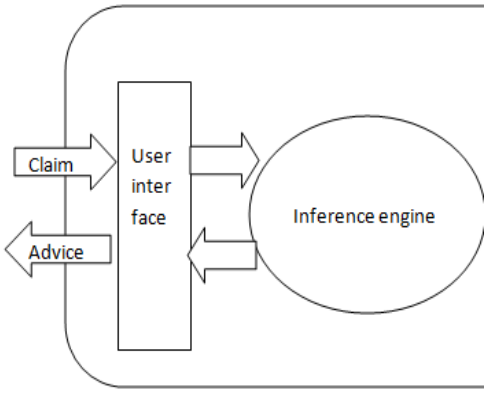

Figure 3.3: Prototype context diagram

3.6.1 Key System Prototype Elements

The system consists of these three main constituent elements in order to accomplish its functions.

3.6.1.1 Input

This is the starting point of the system and is provided by the user. In our case, this consists of the input variables of a patient claim form.

3.6.1.2 Transaction data

This is the actual data that the system will process. This includes a WEKA file with rules to identify abnormalities on the claims fields that are entered.

3.6.1.3 Output

The output of our system will be the notification of weather a claim(s) is fraudulent or not.

3.6.2 System Design & Architecture

The prototype consists of two major components:

3.6.2.1 Logic

This is the most important component of the prototype as it implements the chosen algorithm.

3.6.2.2 User Interface

This is where the user of the system interfaces with the system. The details of the claim will be entered into the system to know its status. The user will then be presented with an answer as to whether the claim is fraudulent or not.

3.6.3 Prototype Implementation

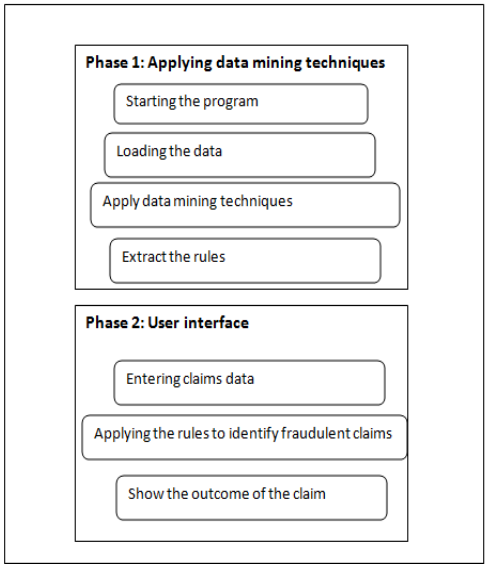

Figure 3.4: Prototype implementation diagram

CHAPTER FOUR

RESULTS AND DISCUSSION

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

4.1 Data Preparation, Filtering and Selection

The original claims dataset was classified in different groups i.e. approved, denied, and cancelled, 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 dataset provided us with 20 features (attributes) as follows: Beneficiary Code, Claim ID, Claim Line Segment, Claims start date, Claims end date, Provider Institution, Primary-Payer Claim paid amount, Claim- Payment Amount Claim, Attending Physician – National Provider Identifier Number. Operating Physician – National Provider Identifier Number, Inpatient admission date, Claim _Admitting Diagnosis Code, Beneficiary Inpatient Deductible Amount, Claim Utilization Day Count, Claim Diagnosis Related Group Code, Claim-Diagnosis Code1 – Claim-Diagnosis Code10, Claim-Procedure Code1 – Claim-Procedure Code6 and Revenue Center HCFA Common Procedure Coding System.

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 disease, in addition to status, we used diagnosis as the control variable in our analysis. So, the features that were considered for preprocessing step were; Claim ID, Claims start date, CLM_THRU_DT, Provider Institution, Attending Physician, Claim Payment Amount, Claim Diagnosis Code 1 – Claim Diagnosis Code2, Status

Later the claims dataset was classified according to the principal diagnosis. The five most common diagnoses in the dataset were taken. After claims filtering and selection, only 23,583 claim records remained from the initial 99, 456 claims. Even though a large number of claims were removed after filtering conditions, the amount of claims remaining was more than

sufficient for classification on the different classifiers. The dataset was trained using the following data mining technologies i.e. Naïve Bayes, K-Nearest Neighbor and decision tree.

4.2 Classifiers Performance

This is a process in knowledge finding process in which the real processes data mining or model building takes place to extract novel patterns hidden in the dataset. The training dataset was classified into approved and denied claims. Claims with the denied status were taken to be suspicious. Therefore classification methods were used to build the fraudulent predictive models. The K-Nearest, 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 analytics over time.
- They need reasonably little trouble from users in the preparation of data to get around the scale variations amidst the parameters.

10 fold cross-validation and the percentage –split (66%) classification setups were applied in performing the experiments of training the models. 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, 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 models are recorded and their effectiveness is compared in grouping new instances of records. A different test dataset was applied to test the effectiveness of the classification models. The models build by the three algorithms were then compared.6 experiments were done in three different scenarios. The data set was run on each algorithm in to two instances in order to establish the best model. The 10 fold and 66% percentage split were used on each algorithm.

4.2.1. J48 Decision Tree Model Building

J48 is a classification algorithm which 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 J48 data mining algorithm is run in different scenarios. J48 has different parameters (like confidenceFactor, minNumObj, reducedErrorPruning, Unpruned etc) that have initial default values and depending on data the values can be changed so that the classification accuracy could be

increased. In this research the algorithm was used with the default values and also tested by changing these values to see the changes in the classification accuracy.

At first, the classification model was modelled 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.

Table 4.1: Default variables for J48 Tree 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

Experiment I

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 tress of 17 leaves and a tree size 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 4.2: J48 10 folds using default values confusion matrix

	True No	**True Yes**	**Class Precision**
Pred No	13987	1941	59.3%
Pred Yes	2254	2812	80.1%
Class recall	87.8%	53.5%	
Accuracy	76.5547%		
Classification error	1.325%		

As evident from the confusion matrix of the experiment 1, J48 Decision tree algorithm recorded an accuracy of 76.554%.The classifier classified 15987 normal claims correctly while 2254 claims were misclassified as claims with anomaly. Out of 4753 claims found with anomaly 1941 claims were misclassified as normal customers thus giving a class recall of

53.5%. The probability of misclassification is approximately 1.325% as given by classification error. The class precision is 59.3% for prediction 'No' and 80.1% for prediction 'Yes'.

Experiment 2

This experiment was done by modifying the default testing option (the 10 fold cross validation). A percentage split was applied that divided the dataset into training and testing data in this learning scheme. This parameter was applied so as to get the efficiency of the learning scheme by incrementing the proportion of the testing dataset to see if it can attain greater classification accuracy than the initial experimentation. The default value of percentage split (66%) was used in this experiment. Table (4.3) below shows the confusion matrix.

Table 4.3: J48 with 66% percentage split Confusion matrix

	True No	**True Yes**	**Class Precision**
Pred No	5702	761	60.4%
Pred Yes	1007	894	79.4%
Class recall	78.8%	53.5%	
Accuracy	76.4921%		
Classification error	1.302%		

As evident from the confusion matrix of the experiment 2, J48 Decision tree recorded an accuracy of 76.4921%.The classifier classified 5702 normal claims correctly while 1007 claims were misclassified as claims with anomaly. Out of 1655 claims found with anomaly 761 claims were misclassified as normal customers thus giving a class recall of 53.5%.The probability of misclassification is approximately 1.320% as given by classification error. The class precision is 60.4% for prediction 'No' and 79.4% for prediction 'Yes'.

4.2.2 Naïve Bayes Model Building

Experiment 3

Naïve Bayes was the second 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 drew from the presumption of conditional independence amidst the attributes.

WEKA software framework was used to model the Naive Bayes model using the Naïve Bayes simple algorithm. The 10 fold cross validation and the percentage split with 66% for training

were employed. Table (4.4) below shows the confusion matrix for the first experiment of the Naïve Bayes simple algorithm model with the default 10 fold cross validation test option.

Table 4.4: Naïve Bayes 10 folds Confusion matrix

	True No	**True Yes**	**Class Precision**
Pred No	16159	1756	75.8%
Pred Yes	180	5488	98.9%
Class recall	90.2%	96.8%	
Accuracy	91.7907%		
Classification error	1.089%		

As evident from the confusion matrix of the above experiment, Naïve Bayes simple algorithm recorded an accuracy of 91.790%. The classifier classified 16159 normal claims correctly while 180 claims were misclassified as claims with anomaly. Out of 8018 claims found with anomaly 1756 claims were misclassified as normal customers thus giving a class recall of 96.8%.The probability of misclassification is approximately 1.089% as given by classification error. The class precision is 75.8% for prediction 'No' and 98.9% for prediction 'Yes'.

Experiment 4

The Naïve Bayes model building second experiment was done using the Naïve Bayes simple algorithm with the 66% training and testing percentage split test option. Table (4.5) below shows the confusion matrix for the second experiment of the Naïve Bayes simple algorithm model with 66% percentage split.

Table 4.5: Naïve Bayes with 66% percentage split Confusion matrix

	True No	**True Yes**	**Class Precision**
Pred No	5549	579	75.9%
Pred Yes	70	1820	98.8%
Class recall	90.6%	96.3%	
Accuracy	91.9057%		
Classification error	1.107%		

As evident from the confusion matrix of the experiment above, Naïve Bayes classifier recorded an accuracy of 91.9057%. The classifier classified 5549 approved claims correctly while 70 claims were misclassified as claims with anomaly. Out of 2399 claims found with anomaly

579 claims were misclassified as normal claims thus giving a class recall of 96.3%.The the probability of misclassification is approximately 1.107% as given by classification error. The class precision is 75.9% for prediction 'No' and 98.8% for prediction 'Yes'

4.2.3 K-Nearest Neighbor Model Building

Experiment 5

The third data mining method applied for the classification was the K-Nearest Neighbor. It applied the K-Nearest Neighbor algorithm in building the model. The 10 fold cross validation set by default and the percentage split with 66% for training and testing of the model were used. Table (4.6) below shows the K-Nearest confusion matrix with the default 10 fold cross validation test option.

Table 4.6: K-Nearest 10 folds Confusion matrix

	True No	**True Yes**	**Class Precision**
Pred No	16037	1878	54.0%
Pred Yes	2605	3063	86.0%
Class recall	89.5%	54.0%	
Accuracy	80.9905%		
Classification error	2.25%		

As evident from the confusion matrix of the experiment 5, the K-Nearest Neighbor classifier scored an accuracy of 80.9905%. The classifier classified16037 approved claims correctly while 2605 claims were misclassified as claims with anomaly. Out of 2399 claims found with anomaly 1878 claims were misclassified as normal claims thus giving a class recall of 54.0%.The the probability of misclassification is approximately 2.254% as given by classification error. The class precision is 54.0% for prediction 'No' and 86.0% for prediction 'Yes'

Experiment 6

The second experiment K-Nearest model modelling was done applying the K-Nearest Neighbor with the 66% training and testing percentage split test option. Table (4.7) below shows the K-Nearest confusion matrix with 66% percentage split.

Table 4.7: K-Nearest with 66% percentage split Confusion matrix

	True No	**True Yes**	**Class Precision**

Pred No	5524	604	61.7%
Pred Yes	915	975	85.8%
Class recall	90.1%	51.6%	
Accuracy	81.651%		
Classification error	2.254%		

The results above denote that K-Nearest Neighbor classifier got81.651%best accuracy score. The classifier classified 5524 approved claims correctly while 915 claims were misclassified as claims with anomaly. Out of 1579claimsfound with anomaly995 claims were misclassified as normal claims thus giving a class recall of 51.6%.The the probability of misclassification is approximately 2.254% as given by classification error. The class precision is 61.7% for prediction 'No' and 85.8% for prediction 'Yes'.

4.3 Comparison of the Classification Models

The second objective of this research was to explore an appropriate data mining algorithm for the problem of detecting doubtful health-insurance claims. From what the researcher found in literatures, the algorithms selected to perform well in fraud cases are K Nearest, Naïve Bayes and decision tree algorithms. Experiments were done to verify these classification algorithms on Medicare patient claim dataset. It was needed to compare the three algorithms applied on the same training dataset so as to get the approriate algorithm to develop the prototype. The comparison based on classification accuracy and effectiveness of the three classification models is summarized and depicted in Table (4.8). It is shown from below the Naïve Bayes classifier gives the best results.

Table 4.8: Comparison of classification models

Classification Model	Correctly classified	Misclassified	Better classifier
K-Nearest			
10 Folds	19100	4483	**Naïve Bayes**
66% split	6499	1519	
Decision Tree			
10 Folds	15978	7605	
66% split	12946	5771	
Naïve Bayes			
10 Folds	21,647	1936	
66% split	649	3294	

4.4 Model and Prototype Design and Development

The methodology shown in figure 4.1 was adopted to design and develop the model for detecting and predicting fraud in the claims. Naïve Bayes, The data was trained using the following data mining classification models; Naïve Bayes, K-Nearest Neighbor and decision tree. The unlabeled data was tested and the output file generated classifying claims as either normal or fraudulent. The prototype was developed using the Java platform and WEKA software was used for model development. The computer used for training and testing was HP core i3 with 4.0GB RAM capacity.

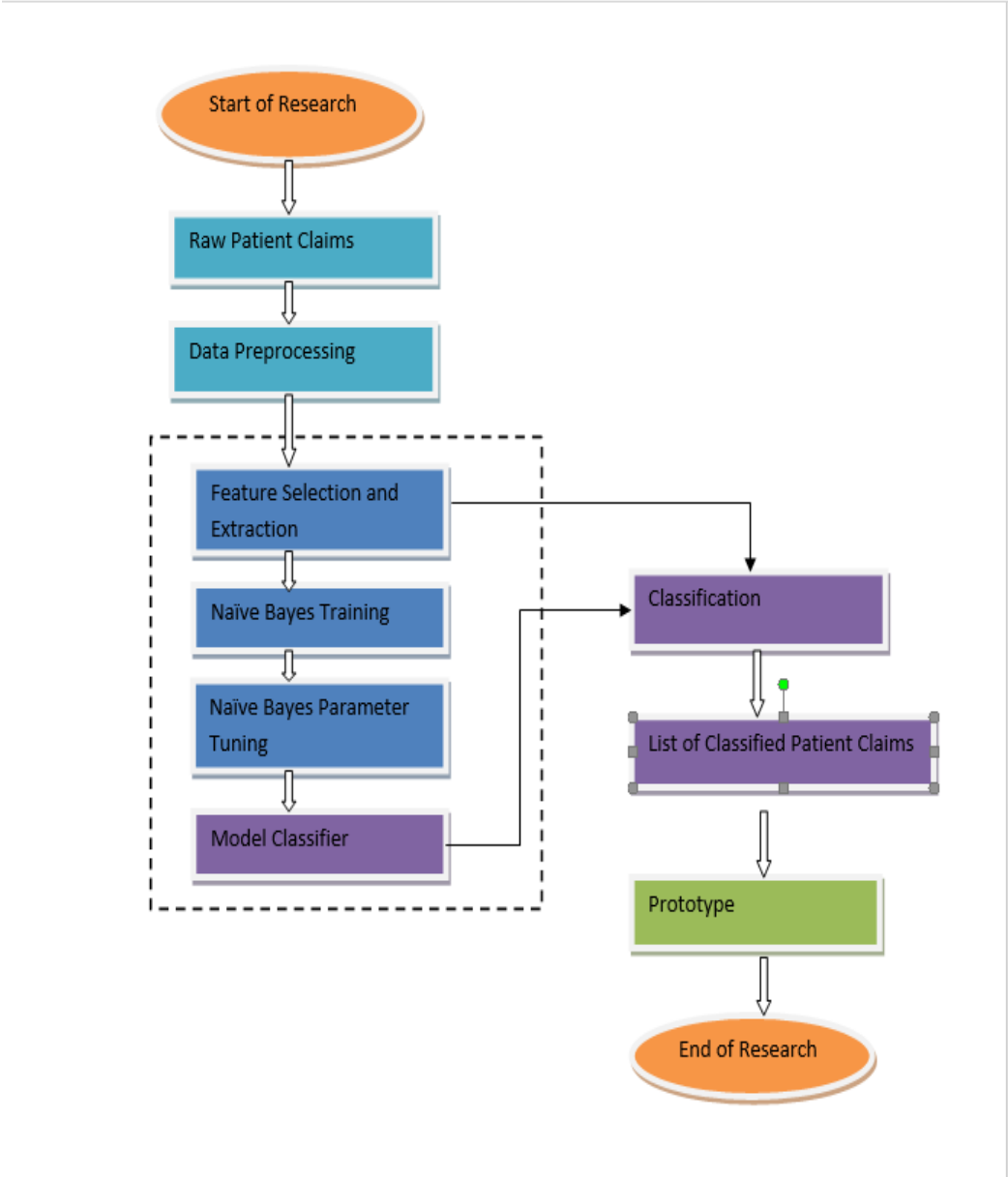

Fig 4.1: Proposed framework for detection of fraud in claims flowchart

4.4.1 Naïve Bayes Model

4.4.1.1 Classification Engine Development

In developing the classification engine, the Naïve Bayes model was the major emphasis of the research. Developing the classification engine entailed; claims attributes assessment and analysis, training and developing the Naïve Bayes classifier, Naïve Bayes variables adjusting and lastly testing the Naïve Bayes model.

4.4.1.2 Claims attribute analysis

Ten (10) attributes of the claims were analyzed and used, so as to develop the Naïve Bayes model design for the reason of training. For this study, a 2-class Naïve Bayes classifier was applied to depict the two categories of claims. The claims dataset were thus analyzed and classified into two different categories (fraudulent or normal claims) according to their behavior of providers and amount paid per patient diagnosis.

From the 23583 claims dataset used, 5667 were indicated as denied meaning there were some anomalies in them. Claims tagged approved and claims tagged denied formed the backbone for the development of the Naïve Bayes model.

4.4.1.3 Naïve Bayes Development

After the claims dataset analysis, a fresh dataset was built containing 17915 approved claims and 5668 denied claims to be used in the development phase of the model and later be used to train the Naïve Bayes classifier so as to build a Naïve Bayes model. Developing the Naïve Bayes model consisted of; parameter optimization, and Naïve Bayes training and testing.

4.4.1.4. Naïve Bayes Model Training and Testing

After replacing missing attributes and normalization, all the 23583 claims were trained so as to design the Naïve Bayes model (classifier). So as to establish unbiased evaluation between classifiers, similar number of model inputs were selected for the experiment. WEKA version 8.3.4 was used to train and test the classifier. In order to group the claims as normal or abnormal, 10-fold cross validation was applied to test and assess the classifier. 91.7907% accuracy was achieved by the model during training. And during testing a 72.12% hit rate was achieved with the supplied test dataset as is shown in the WEKA Figure (4.2) and Figure (4.3).

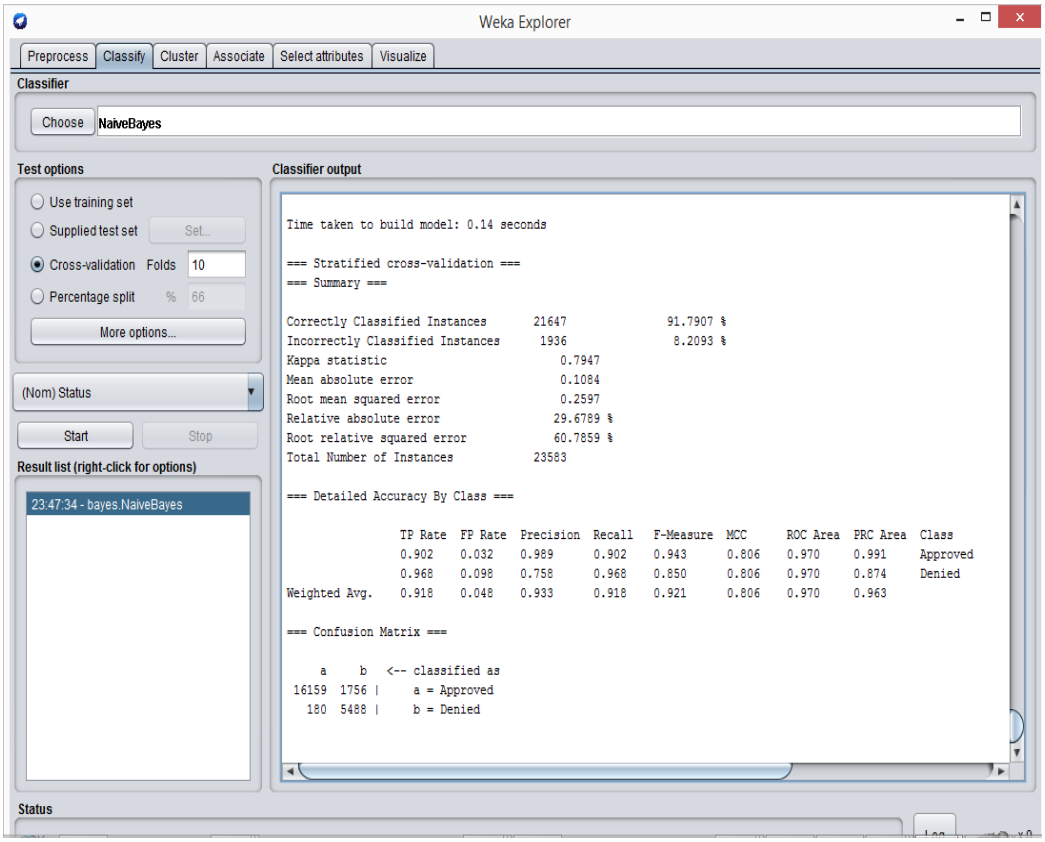

Figure 4.2: Naïve Bayes model training results

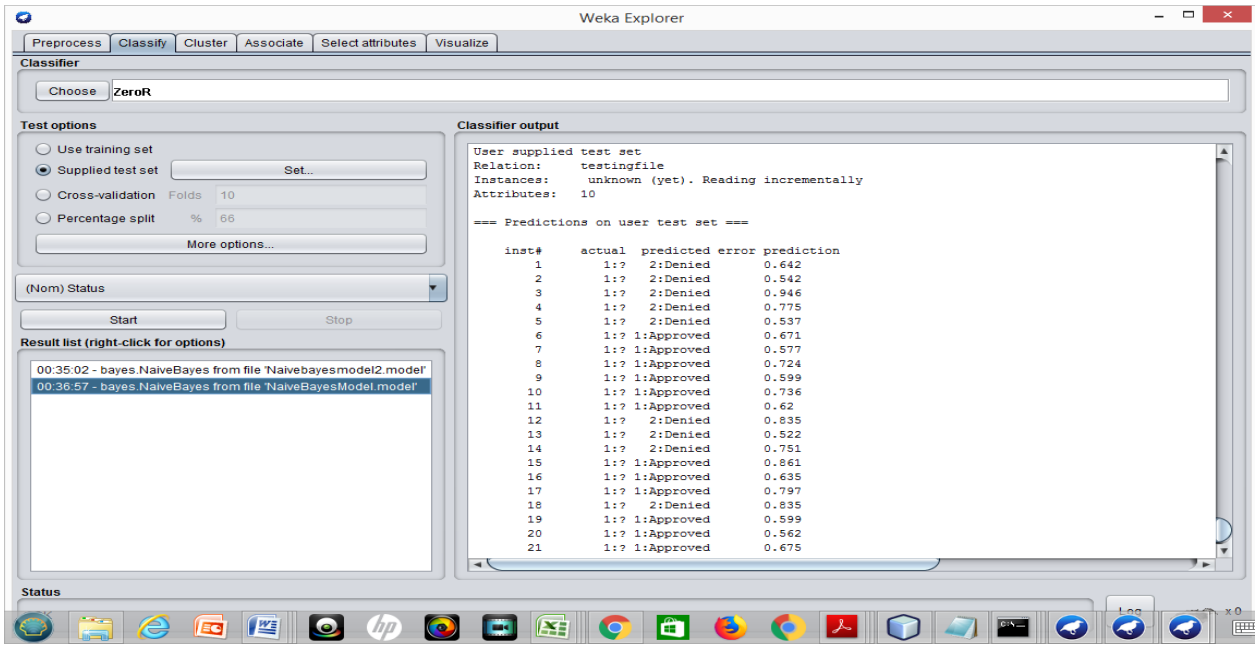

Figure 4.3: Naïve Bayes model testing results

4.4.2 The Prototype

Prototype development was needed to show that the data mining model developed could be deployed. Healthcare organizations can organize data warehousing to make use of data mining model. But for this research purpose a simple Java application was developed. The system load by asking the user to provide a training dataset for the claims he wants to check. A model is then build using the dataset provided. On clicking the predict button, he should provide the claims that he want to check for fraud. An output is then presented on whether the claims are fraudulent or not.

This system could be integrated with the operational system and after experts encode the model in the operational system they will only fetch the claim information to this system through database link functionality. The user interface is only for loading the training dataset and predicting for fraud in another test data based on the rules set by the model build. The interface is shown in Fig.4.4, 4.5 and 4.6. And the source code is attached as an appendix.

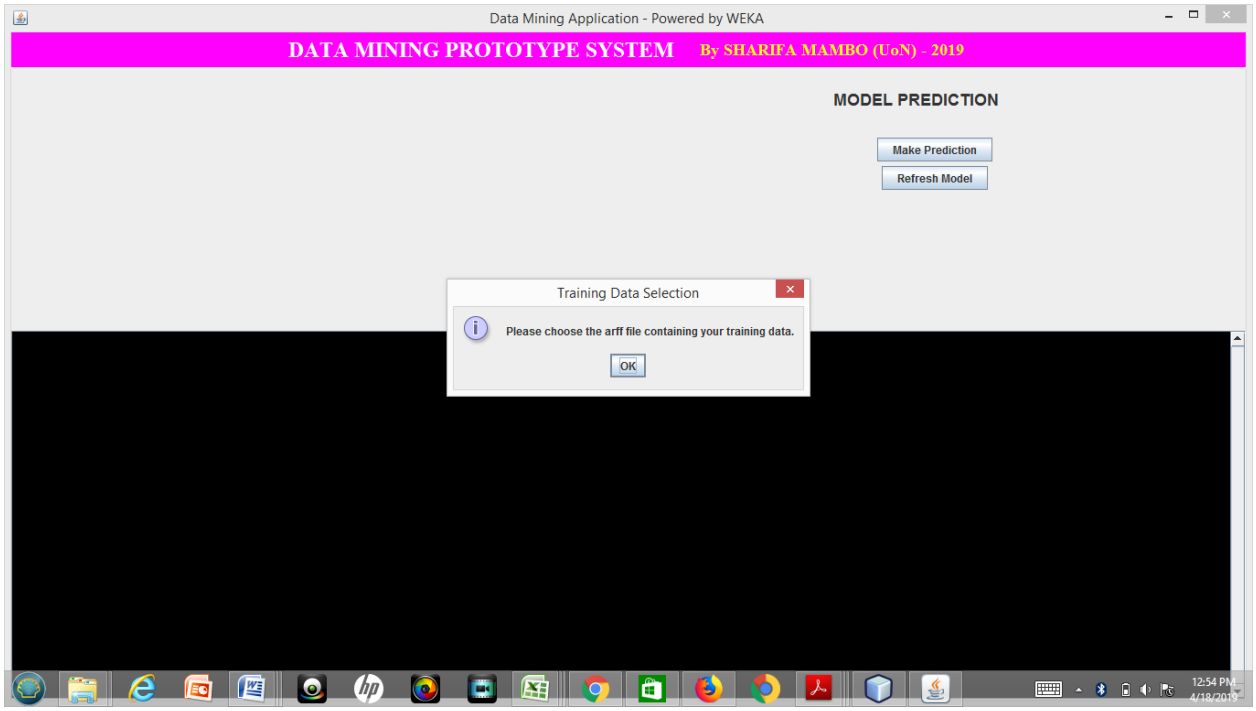

Fig 4.4: Prototype home page

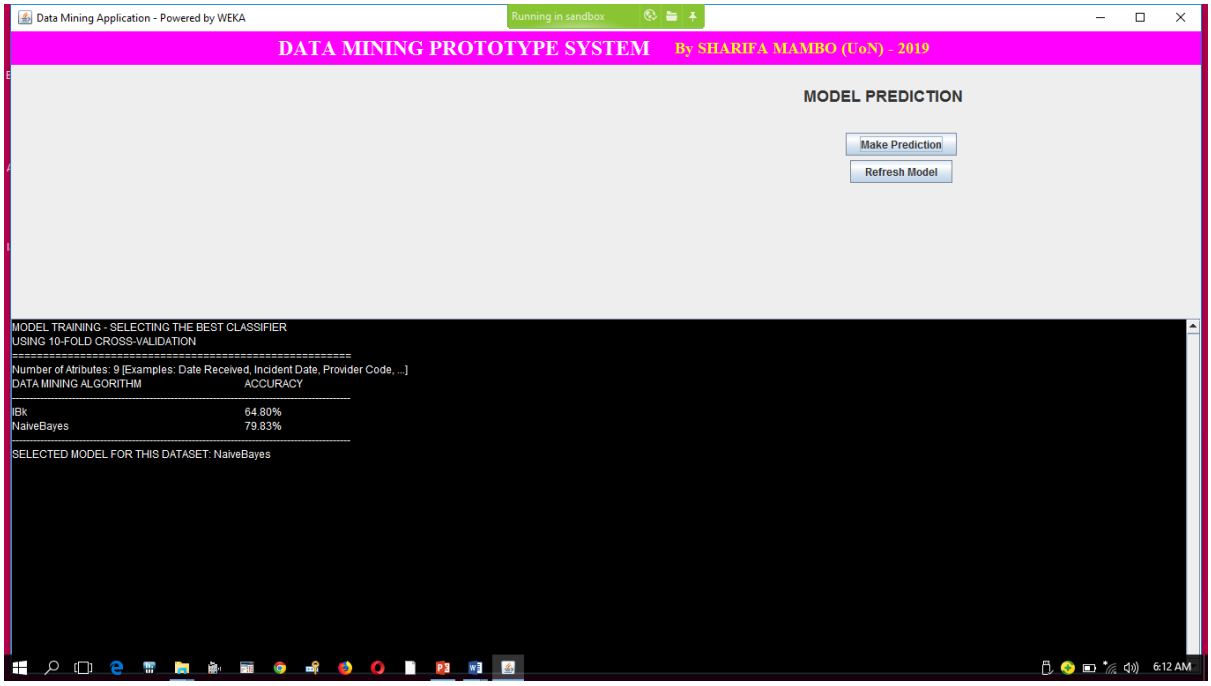

Fig 4.5: Model Building

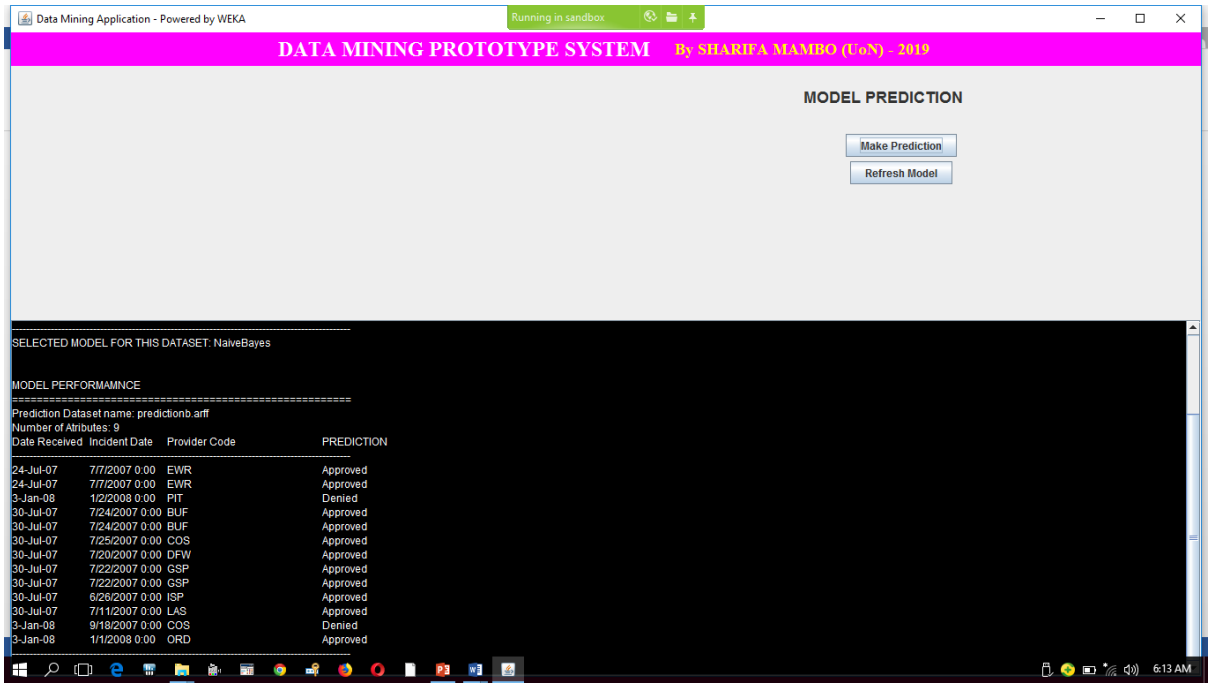

Fig 4.6: Predicting results

CHAPTER FIVE
CONCLUSION AND RECOMENNDATIONS

This chapter a present summary of research findings, conclusion, recommendations and suggestion for further work

5.1 Summary of Research Findings

Objective No 1: To identify the fraudulent activities perpetrated in health insurance claims.

The research established that the main the fraudulent activities perpetrated in health insurance claims are: billing for more overpriced facilities or services than those that were truly administered, carrying out medically services that are not required, false claims, and forging a patients' diagnosis to explain some tests.

Objective No 2: To explore data mining techniques that could be used to manage fraudulent activities in health-insurance

In this research study, the Naïve Bayes, K-Nearest and decision tree data mining algorithms or techniques were investigated and used. The techniques were chosen because they have been widely and successfully applied in predictive analytics over time and they need a comparatively little effort from users in preparing the data to solve scale contrasts between parameters.

From the result analysis of the tested models, the Naïve Bayes technique emerged as the best model in comparisons with other data mining techniques in terms of accuracy percentage.

Objective No 3: To develop and test a data mining prototype for detecting fraudulent health insurance claims

The desired fraud detection prototype was developed based on a Naïve Bayes WEKA model and verified using some test data that was not applied or used while training the model. We believe that when fully developed and used, this prototype will be able to help detect fraudulent claims in health insurance.

5.2 Conclusion

Fraud detection is a field which can never rest; fraudsters do exist and always will device for new ways to perform fraud. Data mining discovers patterns not quite visible in data to convey some knowledge. Fraud detection in health insurance companies in both developed and undeveloped countries is much needed as inability to discover fraudulent claims put them at risk of losing a lot of money and in return affecting the service delivery to patients. Developing

a mechanism to predict fraud is considered an achievement in the health insurance organizations.

The results of this study demonstrated how Naïve Bayes can be applied to groups of data under investigation to detect any abnormal behavior in some of the data. This research did not dwell on only one particular technique to build the model but did a comparison of other techniques on the same dataset to identify the positive and negative aspects of each. Such comparison allowed for the identification of the best techniques to provide a more robust fraud detection system.

Our proposed prototype is dynamic and flexible enough to enable a user train the model with different datasets as per requirement; It is built on the best two classification model thus picks the best for the anomaly detection and lastly its main methodology is flexible to many other sectors in health-care and likely in other industries

Traditional claims fraud detection by most health insurance companies gives a detection hitrate of 3%. Our developed fraud detection system will guarantee any health insurance company a detection hitrate of 60-70%.This will benefit the health insurance companies not only in improving its handling of fraud in claims, but will also register tremendous savings if the system concept is put into use.

Though this prototype is done for academic research purpose and it is an initial work regarding the health insurance fraud detection area, it has positive results to be implemented for the intended purpose and even adopted by other companies with similar problem of fraud.

5.3 Research Limitations

Limitations to the research include: Using only three common data mining techniques for training: Naïve Bayes, K-Nearest Neighbor and decision tree ; Bigger datasets were taking long to build thus the researcher used a sub dataset; and building a prototype that is based on a model and not an expert system. The prototype accepts data in ARFF format only.

5.4 Recommendations

It is proper to carry out the experiments with very large training and testing datasets and also make several trials to come out with better explicit classifiers

The prototype can be customized and extended appropriately to accommodate and be adopted for commercial use by health insurance companies.

5.5 Future Work

We recommend the implementation of a full working expert system to detect fraud in claims based on the rules provided in the future.

References

Akhtar, D. M. I. (2016). Research Design. *Available at SSRN 2862445*.

Alniz, P. (2018).”Why fraud is a major threat to your SME’s health costs (and how to beat it!)” Linkedin 26th Sep. 2018, https://www.linkedin.com/pulse/why-fraud-major-threat-your-smes-health-costs-how-beat-alniz-popat.

Apima, M (2018). Insurance Fraud; The pointers and safeguards. Aki Journal 16 Retrieved from https://www.akinsure.com/images/journal/AKI-JOURNAL-JUNE-2018WEB.pdf

Cloudia, K. (2018). How Block chain Could Disrupt Insurance? Future Internet 10, 20

Copeland, L., Edberg, D., Panorska, A.K., Wendel, J. (2012). Applying Business Intelligence Concepts to Medicaid Claim Fraud Detection. *Journal of Information Systems Applied Research*, 5(1) pp 51-61. http://jisar.org/2012-5/ ISSN: 1946-1836. (A preliminary version appears in The Proceedings of CONISAR 2011)

Dharani, S. & Shoba, S.A. (2015). Identifying the Fraud Detection in Health Care System Using Data Mining. International Research Journal of Engineering and Technology, 2(9) 2395-0072

Eyad, H. S. (2012). A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG (Master’s thesis). Retrieved from http://library.iugaza.edu.ps/thesis/106989.pdf

Frankfurter, C. &Cuervo, L.G. (2017). E-Government as tool to advance health. Global Journal of Medicine and Public Health. 5. http://www.gjmedph.com/uploads/VP1-Vo5No6.pdf.

Gosain A, Kumar A. (2009) Analysis of health care data using different data mining techniques. International Conference on Intelligent Agent & Multi-Agent Systems. p. 1-6. crossref.

Gregory, P. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques, Waltham, MA. *Morgan Kaufman Publishers*, *10*, 978-1.

Hasheminejad, S. M., &Salimi, Z. (2018). FDiBC: a novel fraud detection method in bank club based on sliding time and scores window. *Journal of AI and Data Mining*, *6*(1), 219-231.

Khristy, H. (2017, December). How technology impacts the insurance sector. XprimmMagazine Retrieved from http://www.xprimm.com/How-technology-impacts-the-insurance-sector-articol-117,163-9818.htm

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect health care fraud and abuse: a review of literature. *Global journal of health science*, *7*(1), 194.

Shin, H., Park, H., Lee, J., &Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, *39*(8), 7441-7450.

Hyunjung, S., Hayoung, P., Junwoo, L.&Won, C. J. (2012). A scoring model to detect abusive billing patterns in health insurance claims. 39(8).

Inês, O. (2017). Application of Neural Networks to the Detection of Fraud in Workers' (Master's Thesis).Retrieved from https://run.unl.pt/bitstream/10362/32561/1/TEGI0404.pdf

Gee, J. & Button, M. (2015). The Financial Cost of Fraud 2015 What the latest data from around the world shows. Retrieved from

https://www.pkf.com/media/31640/PKF-The-financial-cost-of-fraud-2015.pdf

Kenya National Health Accounts FY2015/16 Technical Report 2016. Retrieved from ResearchGate website https://www.researchgate.net/publication/321864804_Kenya_National_Health_Accounts_FY_201516

Kizito, M. (2016). Technology is How We Win Insurance Fraud Battle [Blog Post].Retrieved from https://www.cio.co.ke/technology-is-how-we-win-the-battle-against-insurance-fraud/

Kyriakakis, A (2015) "The Missing Victims of Health Care Fraud," Utah Law Review: Vol. 2015 : No. 3 , Article 2

Legotlo, T. &Mutezo, A (2018).Understanding the types of fraud in claims to South African medical schemes. SamJ Research S*Afr Med J* 2018; 108(4):299-303.

Lelenguiya, J. K. (2015). An Intelligent Model for Detecting Fraud in the Non-Technical Loss of Commercial Power: Case of Kenya Power-Ruiru Area (Master's Thesis). Retrieve from http://erepository.uonbi.ac.ke/bitstream/handle/11295/90122/Lelenguiya_An%20intelligent%20model%20for%20detecting%20fraud%20in%20the%20non%20technical%20loss%20of%20commercial%20power?sequence=3&isAllowed=y

McCallum A, Nigam K. A comparison of event models for Naïve Bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752; 1998. p. 41–8.)

Mackey, T. K., & Liang, B. A. (2012). Combating healthcare corruption and fraud with improved global health governance. BMC international health and human rights, 12(1), 23.

Mamo, D. (2013). Application of Data Mining Technology to Support Fraud Protection: The Case of Ethiopian Revenue and Custom Authority (Master's thesis). Retrieved from http://etd.aau.edu.et/bitstream/handle/123456789/14247/Daniel%20Mamo.pdf?

Measure Evaluation (2017). "How Kenya Monitors Health Information System Performance Findings from a Case Study". Retrieved from https://www.measureevaluation.org/resources/publications/fs-17-232

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

MoH (2015). Kenya Health Sector Monitoring and Evaluation Framework: - Accelerating attainment of universal coverage. Nairobi, Kenya.

Munge, K., Mulupi, S., Chuma, J. (2015). A critical analysis of the purchasing arrangements in Kenya: the case of the National Hospital Insurance Fund, Private and Community-based health insurance.

BorseNikita, B., &MaitreNeeta, M. (2015). Health Care Insurance Fraud Detection: A Data Mining Perspective. IJACCCS 1(2)

Nutley, T. (2012). Improving data use in decision making: an intervention to strengthen health systems. Chapel Hill, NC: MEASURE Evaluation, Carolina Population Center. 12(73)

Ogwueleka, F. N. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology*, *6*(3), 311-322.

Owusu, O. O., Effah, J., &Boateng, R. (2018). Biometric Technology for Fighting Fraud in National Health Insurance: Ghana's Experience.

Pal, R., & Pal, S. (2015). Application of Data Mining Techniques in Health Fraud Detection. *International Journal of Engineering Research and General Science*, *3*(5), 129-137.

Patil, D.Y (2012). Health Insurance: "Identifying Awareness Preferences, and Buying Pattern In Mumbai." (Master's Thesis)

Phua, C., Lee, V., Smith, K., &Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Pawar P. (2016) Review on Data Mining Techniques for Fraud Detection in Health Insurance International Journal on Emerging Trends in Technology (IJETT) 3(2)

Liu, Q. (2014). *The application of exploratory data analysis in auditing* (Doctoral dissertation, Rutgers University-Graduate School-Newark).

Rashidian, A., Joudaki, H., &Vian, T. (2012). No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature. *PloS one*, *7*(8), e41988.

Rawlson, O. K. (2015). Biometrics and Healthcare. Biometric Update Special Report. Retrieved from https://www.biometricupdate.com/201502/special-report-biometrics-in-healthcare

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.

Mukherjee, S., Shaw, R., Haldar, N., &Changdar S (2015). A Survey of Data Mining Applications and Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 20

SAS (2016). The State of Insurance Fraud Technology A study of insurer use, strategies and plans for anti-fraud technology. Coalition against Fraud. Retrieved from

https://www.insurancefraud.org/.../State_of_Insurance_Fraud_Technology2016.pdf

Subelj, L., Furlan, S., &Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, *38*(1), 1039-1052.

Swapnil, S. (2018, December 9). Role of Analytics in Insurance Claim Fraud Detection [Blog Post]. Retrieved from https://www.mindtree.com/blog/role-analytics-insurance-claim-fraud-detection

Travaille, P., Müller, R. M., Thornton, D., & Van Hillegersberg, J. (2011, August). Electronic Fraud Detection in the US Medicaid Healthcare Program: Lessons Learned from other Industries. In *AMCIS*.

Wafula, L., Orto, A &Mageto, S. (2014) Application of The K-Means Clustering Algorithm In Medical Claims Fraud / Abuse Detection. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)* Web Site: www.ijaiem.org Email: editor@ijaiem.org3(7) 2319 – 4847

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, *16*(4), 449-475.