UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

APPLICATION OF MACHINE LEARNING TO DETECT FRAUDULENT MATERNAL MEDICAL CLAIMS

ONYANGO NYAKENO MARGARET

P52/38304/2020

Supervisor

Dr. Evans A.K Miriti

The Research Project Submitted in Partial Fulfillment of the Requirements for the Award of the Master of Science in Computational Intelligence at the University of Nairobi

AUGUST, 2022

DECLARATION

The research project & prototype here presented is originally my work and has not been presented in any other institution of higher learning . References have been made from the works of researchers that have been used .

Margaret Nyakeno Onyango:  Date: 26-08-2022

(P52/38304/2020)

This research project & prototype has been submitted for a partial fulfillment of the Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University supervisor

Dr. Evans K. Miriti :  Date: 26-08-2022

ACKNOWLEDGEMENTS

First and foremost, I thank the Almighty God for granting me knowledge, strength, and good health as I tirelessly wrote and researched on this project. My heartfelt gratitude goes to my supervisor Dr. Evans K. Miriti for his professional guidance which ensured the successful completion of this proposal. Lastly, my sincere gratitude goes to my lovely family for their consistent and continuous support

ABSTRACT

Fraudulent activities have caused great losses in the healthcare industry all over the world. Different methods such as upcoding, billing for services not rendered and many more are ways fraudulent activities occur. Traditional methods of fraud detection such as auditing and rule-based programming are no longer efficient due to the increase of data and complexity in the billing process of medical claims. There is a great need for new optimized methods to assist in fraud detection. Data mining and Machine learning are optimized methods that can be used to improve the sector.

The objectives of the research were to train a machine learning-based model which detects a health insurance claim considered fraudulent, identify features in insurance claims that can be used for fraud detection, identify appropriate machine algorithms models to use for fraud detection, compare the performance of different machine learning algorithms and implement a prototype for detecting fraudulent health insurance claims

The research explored the use of different machine learning methods to be able to detect fraud. The method used was the CRISP-DM process. The data went through stages of data collection, where data was collected from an insurance company which is based in Kenya, data preprocessing and transformation to ensure the data was clean, Training where the data was trained using different models, and lastly evaluation where a comparison analysis was done fo based on the performance of each model.

The results gotten from the benchmark and performance evaluation showed that the gradient boosting classifier performed the best with an accuracy of 90.0% and AUC of 95.0%. The other models that performed well included the random forest with an accuracy of 90% and ANN with an accuracy of 88.0%. The model that performed poorly was the Logistic Regression with an accuracy of 59% and Naive Bayes with an accuracy of 47%. The gradient boosting tree classifier model was then used to develop a prototype.

TABLE OF CONTENT

List of Figures

List of Tables

Acronyms

NHIF- National Health Insurance Fund

UHC - Universal Health Care

KDD – Knowledge Discovery Databases

LEIE - List of Excluded Individual or Entities database

IRA - Insurance Regulatory Authority

HIS - Health Information Systems

NN - Neural Network

SVM - Support Vector Machine

ANN -Artificial Neural Network

MLP -Multilayer Perceptron Neural Network

GBC - Gradient Boosting Classifier

CHAPTER ONE
INTRODUCTION

1.1 Background

Healthcare fraud is a type of white-collar crime which occurs when healthcare claims are filed dishonestly to gain profit. Many organizations across the world have lost a lot of money due to healthcare fraud and corruption. Annually, expenditure in healthcare increases rapidly in different countries. Globally, approximately 10% of gross domestic product is spent on healthcare. There are many sources of inefficiency such as fraud and abuse and thus up to 10% of that money is wasted (Joudaki et al., 2016).

In the US, it is estimated that fraud in health insurance costs annually, 80 billion US dollars which are approximately 3% of the national healthcare expenditure (Yao et al., 2014). In the year 2020, about 330 fraud offenders were charged in court. With the increasing losses in the sector, The National Healthcare Anti-Fraud Association has been formed to research in the area of healthcare fraud.

In China, the Chinese Insurance Regulatory Commission estimates that fraud cases lead to losses of about 10-30% of the total income (Yao et al., 2014). In response to the escalating issue, the government proposed to build an anti-fraud system in 2012. In 2021, the government introduced new regulations regarding health insurance fraud. The new law has increased penalties regarding fraudulent acts and has placed a penalty of five times the amount of fraud. Also, the fraudulent activity may be subjected to a 3 to 12 months suspension of compensation by the fund.

In South Africa, an investigation firm Qhubeka Forensic Services indicated that the health system in South Africa loses between half a billion to 1 billion US dollars in healthcare fraud. In our country Kenya, according to the NHIF report ("Strides Towards Universal Health Coverage For All Kenyans," 2018), NHIF loses to a tune of 10 billion Kenyan shillings every year in false medical claims. Between January and February 2021, the NHIF almost lost 27 million in only 15

Service providers are using many methods to defraud healthcare systems both in the private and public sectors. Some of these activities Include; Billing for services that were not performed, upcoding claims, exaggerating medical illness, receiving kickbacks, and phantom billing

Fraud in healthcare is now perceived as a serious social concern. It is a problem for Insurance companies and the governments and thus there is a great need for more effective detection and prevention methods.

1.2 Problem Statement

Traditional methods of detecting healthcare fraud such as whistleblowing, planned audits, Statistical methods, and pattern matching are time-consuming and not effective. This has led to organizations losing up to 4 to 5% of their revenue due to fraud. This has become a limiting factor in the delivery of affordable premiums in insurance and quality healthcare to the insurance subscribers. Therefore, there is a great need to have automatic fraud detection systems in place.

1.3 Objectives

1.3.1 Main Objective

To develop and test a machine learning-based model for detection of fraudulent health insurance claims

1.3.2 Specific Objectives

1. To identify features in insurance claims that can be used for fraud detection
2. To identify appropriate machine algorithms that can be used for fraud detection
3. To implement a prototype for detecting fraudulent health insurance claims

1.4 Research Questions

1. What features in health insurance claims can be used for fraud detection?
2. Which machine learning algorithms are most appropriate for the identification of fraudulent claims?
3. What is the performance of the machine learning algorithm selected for use in the detection of fraudulent medical claims?

1.5 Significance

This research will contribute to the domain of fraud in insurance and specifically healthcare fraud. It will give recommendations on which machine learning algorithms are appropriate to use when developing a system to detect fraudulent claims. When the number of losses are minimized, patients will experience better and more affordable healthcare. When the prototype is implemented, it will enable insurance to detect fraudulent claims before payments and the perpetrators can be charged.

1.6 Scope

The study will use data from a local insurance company based in Kenya. The data will have maternity claims records which include both fraudulent and non-fraudulent claims.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Healthcare is a primary basic need that everyone around the world needs. The United Nations Sustainable development goal 3 is ensuring healthy lives and promoting well-being. In Kenya Vision 2030 aims to transform Kenya into a newly industrializing, middle-income country providing a high quality of life to all its citizens by 2030, affordable healthcare must be achieved. (GOK, 2008) Good health, in turn, increases the capacity of people to be productive and in turn boosts economic growth within the country. The biggest issue lies in the affordability of healthcare and the continuous increase in cost over time. Good healthcare thus is a form of luxury rather than a basic need. Health insurance was introduced to make the cost of healthcare services affordable (Yao et al, 2014).

Insurance is a device that is used to provide financial compensation in case of a misfortune. These payments are made from a bank of accumulated contributions of different people participating in a scheme. Health Insurance thus pays for medical or surgical expenses incurred during treatment. The insurance may pay the expenses directly to the service provider or may refund an insured member for medical expenses incurred during treatment (Switlick et al., 2015). Over time, the health insurance sector has greatly been impacted by fraudulent activities to gain profit. This has a great impact on the health systems since it causes great inefficiency.

In this chapter, to gain a contextual knowledge of health insurance, we discuss the history of health. We will then look at types of health insurance. Thirdly we will look at how the health insurance ecosystem works, healthcare frauds, healthcare systems in Kenya, discuss machine learning algorithms used for classification, conceptual framework, related works, and the research gap.

2.2 History of Health Insurance

1883 was a defining year for health insurance in Germany. A sickness insurance law was passed and Industrial employers were expected to pay for injury and illness insurance for their workers.

The employees contributed to a sickness fund through their wages and employers also contributed an equal amount (Butticè, 2019). Other countries and companies in Europe began having insurance cover as well. In that time up to 1900, the cost of medical treatment was very low. Most of the doctor visits took place in the home of the patient. The charges for treatment were mostly determined through negotiation (Rosenberg, 1987). Also, there was no growth in the field of medicine and mostly traditional medicine was used to solve illness and accidents at that time. The adoption of medical insurance was very low since there was no perceived value of medical insurance (Gorman, 2006).

In the mid-1920's the modern hospital began as a place of treatment, learning, and improving medical advancement. Consequently, the profession created policies and responded to these improvements by increasing the knowledge base of physicians, training, division of labor, and increased medical specialization. Physicians increased their charges and their decisions were now based on the cost of learning, time is taken to learn the skills, specializations, administrative charges, and increased competition (Starr, 2018).

Another great driver for the growth of the healthcare industry was the world wars. The field had a drastic growth using experience that was gained from the treatment of soldiers. New medications came into the market such as antibiotics, treatment of diseases such as cancers began and maternal health became a hospital event. (Obodoekwe Nnaemeka, 2017) . Due to the major advancements, people began to spend more money on healthcare and with time the people with low-income wages were unable to afford some treatments. When it came to medical emergencies people across all wages were unable to afford treatment. Then the need for insurance began to be on the rise.

In the US, after the world war, there was a great need for labor. In order to compete appropriately, companies raised salary offerings to attract employees. The government came in and regulated the salary range, and then passed a law that would allow companies to offer medical insurance. With the government wrapped in, health insurance was introduced as a corporate package and members contributed to a medical scheme.

Among the first organizations to offer hospital treatment funds was the Blue Cross and Blue Shield company. They gave a treatment plan to their subscribers who contributed to a pool of funds which ensured they were covered in case of an emergency. Over the years, governments especially in Europe began public health insurance commonly known as Universal Health Care whose aim was to provide affordable treatment to their citizens. This was mostly funded by taxes and contributions from members of the scheme (Gorman, 2006).

2.3 Health Insurance Categories

Healthcare insurance can be classified into two major types. Public-based health insurance is a subsidized insurance policy that is paid for by the government and sometimes may require subscribers to pay as well. Private-based health insurance is where the subscriber entirely pays for the premiums.

2.3.1 Public Health Insurance

Most commonly known as National health insurance in many different parts of the world, this is health insurance managed by the governments. The subscribers have part or all their medical costs covered by the Insurance. NHI mostly takes care of primary healthcare. Funding the insurance varies depending on the program or scheme and the country. Some are fully financed by the government through tax while others the population is expected to contribute a certain amount of money to be a beneficiary of the fund.

2.3.2 Private Health Insurance

Private health insurance is run by private companies in the Insurance business. The insurance sells their product to a subscriber who then pays for a subscription. Most organizations pay a certain amount to the insurance carrier and the employees pay a subsidized amount. The government may also pay indirectly by subsidizing tax paid by the insurance organizations. Private medical insurance offers a wide range of benefits to the subscriber such as optical care, dental care, mental care, long-term illness cover, maternal cover, and many more. This is based on a scheme chosen.

2.4 How Health Insurance Works

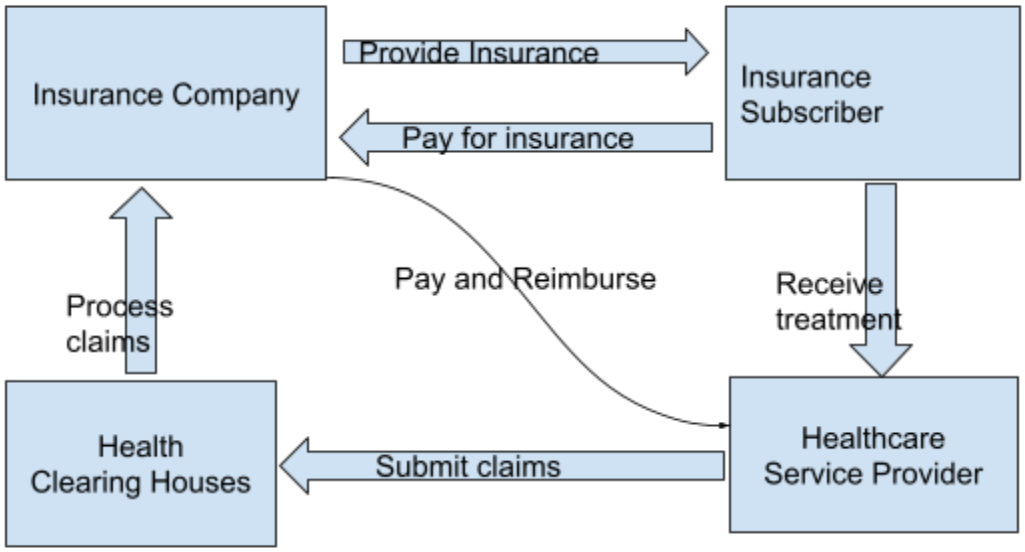


Figure 2.1 - Illustration of the health insurance ecosystem

2.4.1 Health Insurance Companies

They are also known as insurance carriers. These are the organizations that sell insurance to customers. They receive money from insurance subscriptions or premiums. They also pay for services to the health providers once their customers are treated.

2.4.2 Health Insurance Subscriber

The subscriber is the person who pays for insurance premiums. This can be based on employment, where the employer pays the premiums, both employer and employee contribute towards the premiums, and also may include individuals who opt to pay for their premiums. The subscribers have the benefit of enrolling dependents who may also use the insurance. Depending on the contributions, the subscribers can get different levels of care which can range from basic health care to comprehensive health coverage

2.4.3 Health Service Provider

A healthcare service provider is an entity either individual health professionals or a health facility who are accredited to practice healthcare under existing laws. They include hospitals, doctors, clinicians, pharmacists, laboratories, ambulance companies, and many more. They offer services to the subscriber. They then fill a claim to the insurance company which once approved, payment of the service rendered is made.

2.4.4 Insurance Clearing Houses

These are the companies or hubs that function as intermediaries between health service providers and insurance companies. They majorly handle data cleaning of the claims. A claim is a request for payment benefits received after a service is rendered. They process the claims by performing scrubbing, checking, and editing out errors in the claim. This process ensures that claims are correct and reduces the risks of errors in a claim. This helps to fasten the payment process.

2.5 Health Insurance in Kenya

Many developing countries have made great developments to improve healthcare. The Kenyan government has set policies to try and improve the system. The National Health Insurance Fund is a state parastatal that was created as a department in the Ministry of Health in 1996 in an act of parliament. Over the years it has adjusted to changes and several reconstructions have been made. The core mandate of NHIF is to provide medical insurance to cover all its members and declared dependents who are the spouse and children

NHIF is mandatory for Kenyans in the formal sector and optional for those who are in the informal sector. For the civil servants and disciplined forces, the insurance gives a comprehensive cover for both inpatient and outpatient and also other specialized medical services such as optical care, dentistry, mental health, ambulance services, and many more. For those in the private sector, the insurance covers only primary healthcare (Abuya et al., 2015).

NHIF is funded by both government and premiums. Citizens who work in the formal sector are required to contribute monthly and these earnings are deducted from their monthly salary and remitted directly to NHIF. The contribution varies based on salary and may range from Ksh 150

to Ksh. 1700. Those who are in the informal sector, are required to pay Ksh.500 per month to NHIF or they make a one-off yearly payment of Ksh.6000.

Private health insurance is offered by licensed insurance companies that operate in Kenya. They are regulated by the Insurance Regulatory Board to ensure efficiency. This type of insurance is mainly purchased by higher-income employees. Most established companies also pay for their staff a private health insurance cover. The rate of growth in this sector is increasing.

Kenya is defining a new health financing strategy that will lead to universal health coverage.UHC is among the big four strategic pillars declared by the ruling party in Kenya between 2017 -2022. This will see a major improvement in the health sector. The major goal is to ensure all Kenyans have access to quality and affordable medical health coverage.

NHIF was mandated by the government and ministry of health to implement universal health coverage. This was a major problem to other stakeholders due to allegations of corruption, fraud, and financial instability of the institution. The World bank performed an extensive collaboration, research, and dialogue with NHIF and MOH which resulted in the approval that NHIF had the capability to implement UHC in Kenya (Mwaura, et al., 2015).

The healthcare system in Kenya can be placed into three major categories. First is the public healthcare providers, who depend mostly on the government for financing. Secondly, we have the private and for-profit healthcare providers who work with both the government, private insurance companies, and the community to pay for services rendered. Thirdly, we have private-non profit-based healthcare providers who include faith-based and mission hospitals that are majorly funded by donors but also charge subsidized fees to the community. The figure below represents how healthcare providers are distributed in terms of categories.

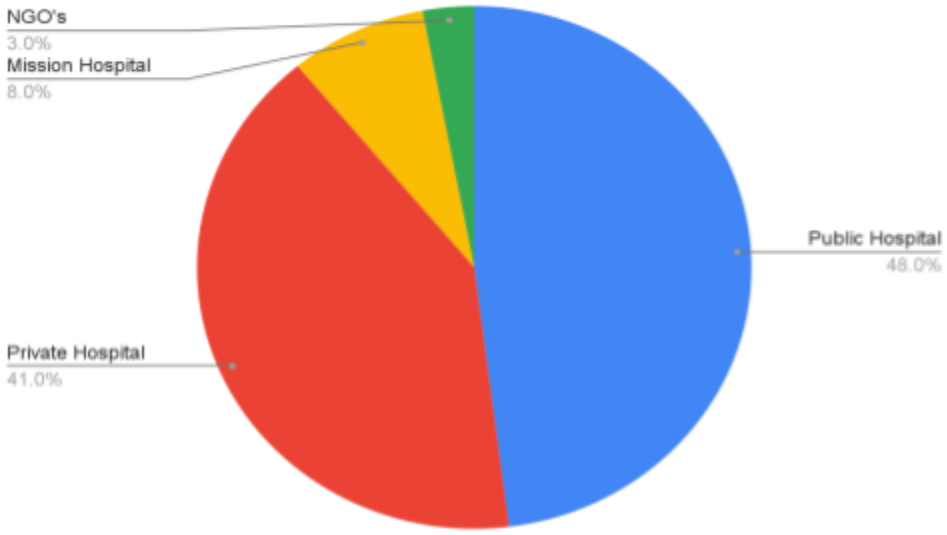


Figure 2.2 - Healthcare facilities distribution in kenya

In Kenya, approximately 20% of the population is privately reported to have a form of insurance cover. Of this population, 89% have been insured by NHIF which is the National Health Insurance Fund(Dutta et al., 2018). The numbers are on the rise due to the affordability of health insurance. The two main types of medical cover taken in Kenya include the government-sponsored NHIF cover and the private sector health insurance covers, offered by insurance companies licensed in Kenya and regulated by the Insurance Regulatory Board.

2.6 Health Insurance Fraud

Fraud has many facets in the insurance industry. This includes health insurance fraud, health insurance waste, and health insurance abuse. Health insurance fraud occurs when a provider or subscriber lies intentionally to gain profit. Healthcare waste is when the health services are used carelessly and irresponsibly and healthcare abuse occurs when required best practices are not followed causing unnecessary expensive treatment.

Health Insurance abuse takes place when the service provider fails to use standard and recommended business practices to the subscriber. Examples of this include performing unnecessary services to a subscriber, performing low-quality services to the subscriber, and improper conduct in filing claims. Abuse and fraud have similar characteristics but in abuse, the investigating party is not able to establish if the act was committed with intentions of making profit (Pies, 2017).

Healthcare waste mainly occurs when unnecessary medical services are provided. It includes practices that are indirectly or directly related to the cost of the medical service. The major difference between fraud and waste is that there is no motive behind the waste. It is unintentional. When trying to classify waste, you have to prove that the actions leading to excess billing were not intentional.

Healthcare fraud occurs when an insured or healthcare provider provides misleading information to the insurance carrier to achieve profit. Some examples of Fraud include Billing for services that were not performed, upcoding claims, exaggerating medical illness, receiving kickbacks, and phantom billing.

2.7 Types of Fraud in Healthcare Systems

In figure 2.1, we looked at the health insurance ecosystem. Fraudulent activities may occur at different levels of the ecosystem.

2.7.1 Healthcare Service Providers Fraud

This occurs when the service provider exploits the system to gain profit. Healthcare service provider fraud takes different forms which include upcoding, unbundling, falsifying medical records, unwarranted procedures, and impersonations.

Upcoding happens when the medical provider bills for services that have not been rendered. The aim is to inflate the total amount of money a patient owes and this cost is pushed to the insurance provider. Unbundling is the billing of multiple procedures for a group of procedures that are

billed as a single item. The service provider will claim for each treatment stage yet the service should be billed as one item. Breaking down the payments will result in a higher bill amount.

Falsifying medical records of patients to justify unwanted procedures, expensive medicines, and irrelevant treatment. This can make the patient's subsequent treatment and diagnosis to be wrong. Impersonation may happen where treatment of an insured person is done but the billing is insured to a person covered by the insurance plan.

2.7.2 Health Subscriber Fraud

The health subscriber may participate in fraud directly or indirectly. Some of these include falsifying information, impersonation, and conspiracy fraud. A subscriber may supply false information about their details to the insurance to obtain better premiums and thus increase their rates and services.

Conspiracy fraud is when a subscriber conspires intentionally with a provider to claim for services not rendered. This may include the subscriber getting some kickbacks as a benefit.
The third is impersonation; this is where a subscriber uses another person's details to get services from a provider. This person is not listed in the list of dependencies. This itself is a crime not only of fraud but also identity theft. Using another person’s membership details illegally is known as a form of identity theft (Piper, 2017).

2.8 Impact of Healthcare Fraud

When service providers perform unnecessary procedures on patients, their decisions could be fatal. Many tests may increase the likelihood of a false diagnosis (Luther, 2020). Other effects include misdiagnosis and overdiagnosing. This may affect the patient's medical records. Rectifying these records may be difficult due to bureaucracy in the field. Future referencing to these medical records may cause fatal effects to the patient. This may even lead to the death of patients.

Insurance inflation is another major impact of fraud. Fraud has greatly cost insurance companies hundreds of millions of dollars over the years (Luther, 2020). These losses need to be shouldered

by someone. For government-based insurance, the taxpayers take the burden. When most of the money is lost, the other sectors of health are under-resourced. This makes the public sector hospitals overburdened and reduces the quality of healthcare offered by the provider. For the private sector, the customers who are the insurance subscribers bear the burden. The insurance providers are in business and in case they see more losses in fraud, they increase the insurance premiums.

Identity theft is another major impact of fraud. Medical insurance companies have been a major target of this. Data contained in the medical records is very personal. They may include credit card details, id details, addresses, and personal medical records. If this information lands in the wrong hands it can be exploited. Medical issues that may arise from identity theft can have serious repercussions. Someone's insurance may be depleted by a thief, and lack when they need to use it. mixed medical records such as blood groups, allergies, and previous illness may cause consequences during emergency cases.

2.9 Machine Learning Algorithms

2.9.1 Logistic Regression

Logistic Regression is a classification algorithm. Logistic regression models the relationship between independent variables and one or more independent variables. The algorithm is based on the following properties: linear property in the independent variables and log odds, little or no multicollinearity among the independent variables, and independence of observations and errors. The logistic regression utilizes the sigmoid function to map the predictions to probabilities. mathematically the formula of the sigmoid function is represented as:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Where S(x) is the sigmoid function and e is Euler's number represented as $e = \sum_{n=0}^{\infty} \frac{1}{n!}$. The sigmoid curve gives a probability of a class or target prediction which lies in the interval between 0 and 1

The sigmoid curve is represented by the graph below

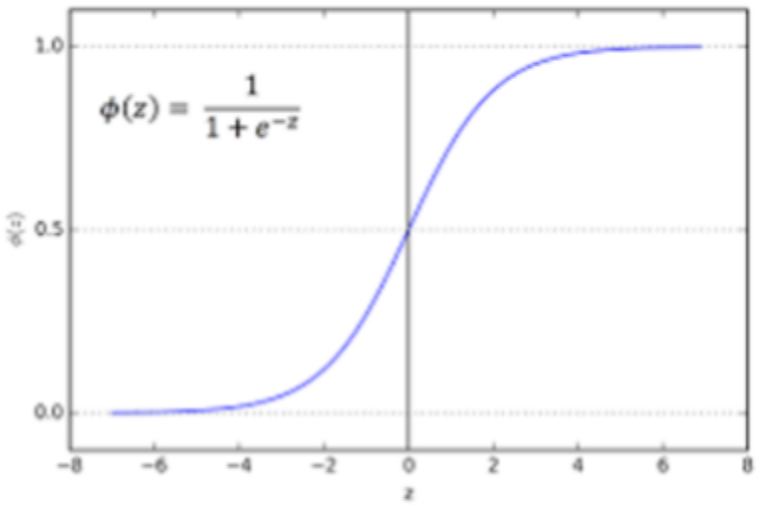


Figure 2.3 Sigmoid Curve

The advantage of logistic regression is that first, it does not make any assumptions about the distribution of the classes or target variable. Secondly, it can easily extend to multivariate logistic regression that predicts multiple classes. Lastly is that the model is less prone to overfitting in low dimensional datasets and thus more accurate in prediction when working with such data. One disadvantage of the model is that it requires minimal or no multicollinearity between the independent variables. Secondly, logistic regression needs that the independent variables are linearly related to the log odds

$$\text{logit}(p) = \log(\frac{p}{1-p})$$

The major drawback is the assumption of linearity among the dependent and independent variables.

2.9.2 Naive Bayes

Naive Bayes is a supervised learning algorithm that is based on the Bayes theorem. The Bayes theorem states that: The conditional probability of an event, X, given the occurrence of another

event, Y, is equal to the product of the likelihood of Y given X and the probability of X.

Mathematically, the Bayes theorem is represented as :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where A and B are events

The probability of B represented as P(B) is not equal to 0.

P(A) represents the probability of A

P(A|B) - the probability of event A occurring given event B is true

P(B|A) - the probability of event B occurring given that event A is true

The algorithm is based on the assumption that the predictor variables are independent and equal. The independence assumption is not always true but often works well during implementation. Some pros of the Naive Bayes algorithm include: it performs better when handling categorical variables as input as compared to numerical variables. Secondly, when the assumption of independence holds for the set of data being predicted, the classifier performs better as compared to other algorithms. Lastly, the algorithm works well in a multi-class prediction problem. A major disadvantage of the algorithm is the assumption of independent predictors.

2.9.3 Random Forests

Random forest is a supervised learning algorithm. It serves both problems of classification and regression . The algorithm is based on the concept of ensemble learning. This is the process by which multiple models are combined to solve a problem and improve the performance of the model(cite). The random forest is a group of decision trees placed together and they utilize the bagging method. Bagging is a method of combining machine learning models ; thereafter improving the result ovarally. It takes decisions from the various trees and based on the majority result chosen, it predicts the final decision. The diagram below shows how it works

Figure 2.4 Random Forest Algorithm

Some advantages of Random forest include: it is very good at handling large datasets with volumes of dimensionality. Secondly, it enhances the accuracy of a model by reducing the variable and avoiding overfitting. Lastly, it works well with both conscious and categorical data. A major drawback of the model is that it suffers interpretability. This fails to determine the significance of each variable in the dataset.

2.9.4 Gradient Boosting Tree Classifier

Gradient boosting algorithm borrows from the concept of boosting. Boosting is a technique that is used to convert low or weak-performing learners to become better ((Obodoekwe Nnaemeka, 2017). Boosting saw great success in the application of Adaptive Boosting or commonly known as the AdaBoost. An explanation of the Adaboost algorithm will aid in understanding how boosting works.

This begins by training the first decision tree which has weights assigned to it. An evaluation is done for the tree. The observations that are difficult to classify are improved by adding more weight to the respective instances. The weighted data is then used to grow the second decision

tree. The tree is evaluated, a new tree is grown to improve on the weakness of the previous trees. new weak learners are added sequentially to focus on training, To ensure the model is improved where the previous trees had difficulty on classification. The final ensemble model is thus the weighted sum of previous predictions.

The Gradient Boosting tree works similarly to the Adaboost, with a difference of using gradient loss function. A loss function is a function that calculates the distance between the expected output or prediction and the current output. Gradient boosting involves three major stages. First, a weak learner, specifically the decision tree which is used to make predictions. Secondly a loss function, and this is dependent on the problem being solved whether a regression problem: which may use squared error, or a classification problem which may use logarithmic loss. Lastly, To minimize the loss function, an additive model adds onto the weak learners .

.

2.9.5 Support Vector Machine

The (SVM) , Support Vector Machine is a supervised machine learning algorithm. It is used regression problems and also for classification problems. The algorithms goal is to create both a line in the data, which is a decision boundary which can data into its specific n-dimensional space. When a new data point is introduced it can easily categorize it into its specific class.
It can be used for both classification or regression problems. The diagram below shows the details of a SVM Classifier.

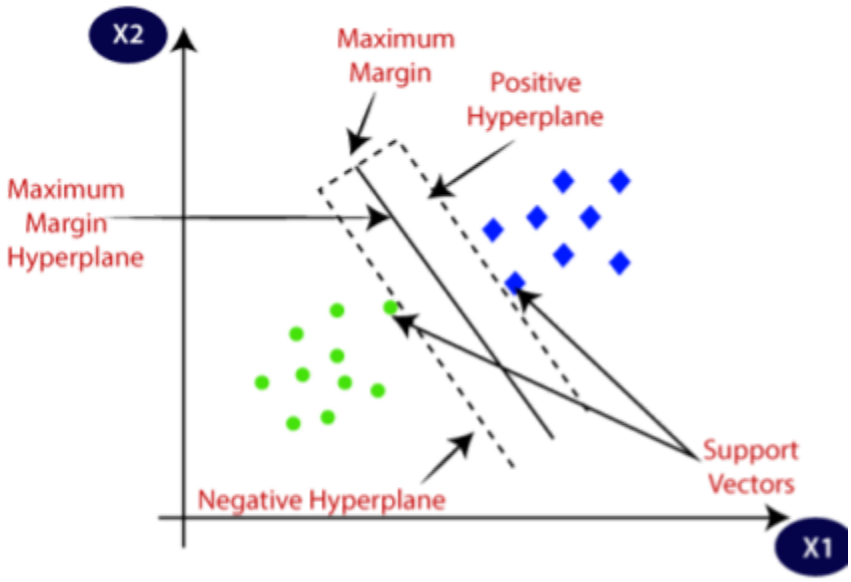
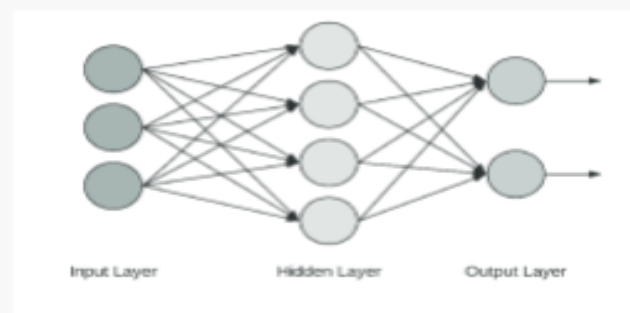


Figure 2.4 Support Vector Machine

2.9.6 Artificial Neural Network

Artificial Neural network commonly known as ANN was created based on how the human brain works. In the brain, we have nerve cells which are the primary units of the brain. They receive sensor input from the outside world, they then process this information and provide an output that may be picked as input in the next neuron and then the final output/interpretation is given. The neural network architecture is represented in the diagram below:



The neural network receives input data using the input layer, it can receive data as patterns or images as vectors. The input values are then multiplied by corresponding weights. Weights are a representation of the strength of interconnections between neurons. If the weighted sum gives a zero, a bias b is added to ensure the output is not a zero. This leads to the hidden layer. These layers transform data and recognize the patterns. These layers are interconnected. Lastly is the output layer which contains a response to patterns recognized by the algorithm. There are several types of neural networks which include the Feedforward neural network, recurrent neural network, convolutional neural network, modular neural network, and Radial basis function neural network. The ANN are gaining a lot of popularity because of their robust nature and high accuracy levels. A major disadvantage is the hardware dependence required to use neural networks and requires a lot of data to create good models.

2.10 Related Work

According to Yao (2014), who worked with a dataset of medical expense insurance in China, they proposed a discrete choice model to be used to identify factors of fraudulent claims. A discrete choice model shows the probability that a particular alternative event will occur, with the probability expressed as a function of variables observed that relate to the alternatives and the event. The research showed that the following factors are significant to detect medical insurance fraud. This includes the total cost of healthcare, a hospital's qualification, policyholder's status, claim duration, and file duration. The research provided a significant contribution by giving a better understanding of predictive factors for healthcare insurance in china. On the other hand, this cannot be generalized to all areas because of the difference between socialism and the economics of a region.

Obodoekwe (2017) study explored a data mining methodology and utilized a knowledge discovery approach in the pipeline. He performed his experiments with different machine learning models and his best performing model was the GBT Classifier. it had had an accuracy of 92%, weighted precision, and recall of 93%, and the F1-score was 92%. The ROC-Area under the curve was exemplary at 97%. The data he used was data from the medicare database which is a US-based insurance company. This research cannot be generalized as well.

Joudaki (2016), did a study that identified indicators of healthcare fraud. They worked with a dataset of health insurance organizations in the private sector based in Iran. A data mining approach with cluster analysis and discriminant analysis was used. They were able to identify Thirteen indicators of healthcare fraud. Using cluster analysis 54% of general physicians were suspects of abusive behavior and 2% of physicians were flagged as suspects of fraud. Discriminant Analysis shows that the thirteen indicators developed performed well in detecting physicians who were suspected of fraud at 98% and abuse at 85%. Their approach was great in clustering and grouping claims and abuse. This would ensure auditing is done to the suspect groups rather than all the physicians.

Thornton et al. (2013) builds on Sparrow's fraud type classification. It Maps to a set of multidimensional data models and analyses the techniques which are important at each level of fraud detection. The research used Medicaid data in dental services. This insurance is located in the US and is used by the majority of low-income citizens. The model used 3 univariate machine learning methods: Linear regression, time series, and box plot. One multivariate method, in which clustering was used. The study was able to successfully predict 17 records correctly as fraudulent claims among 360 records.

A study by Liou et al. (2008), used the diabetic patient claim form details submitted to National Health Insurance of Taiwan. This dataset was used for both training and testing. Three machine learning algorithms were used which include: Logistic regression, decision tree, and Neural network. The Decision tree performed very well in this study with an accuracy of 99%.

2.11 Research Gap

From the studies discussed in the previous sections, adequate comparison analysis of the machine learning algorithms has not been done. Different researchers choose to use different algorithms thus a need to second the performance of algorithms based on a different dataset. Secondly, no study or research has been contextualized to the Kenyan industry of healthcare insurance. Most insurance companies use rule-based analysis to detect fraud which is not accurate and efficient. Thus there is a great need to work on local data.

2.12 The Process Model

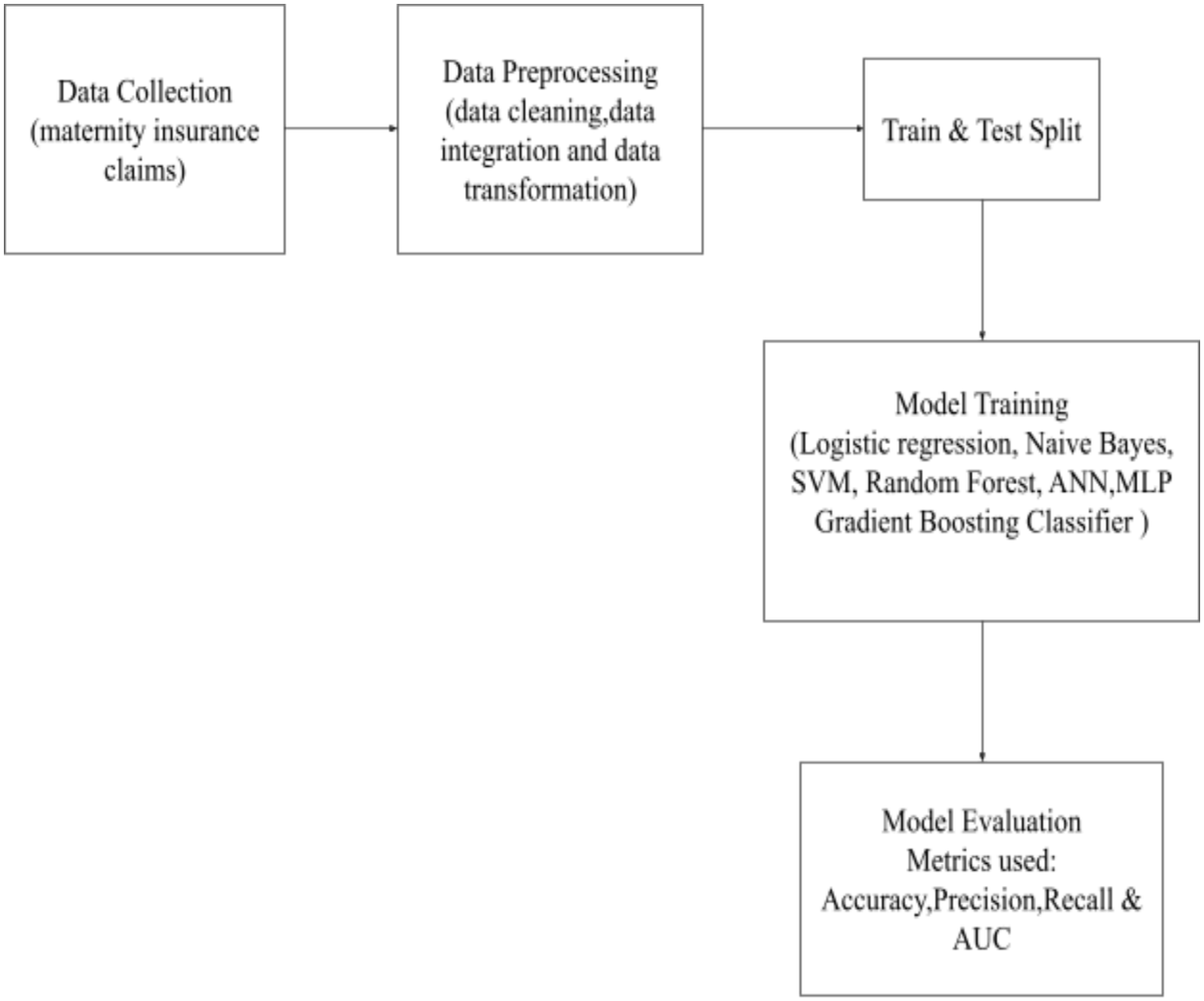


Figure 2.3 - The Process Model

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

In every research, there is a need to establish the method in which the research is conducted. This section covers the research design, research paradigm used, the analysis methods, and tools used to achieve the set of objectives.

3.2 Research Design

Research design is considered the overall strategy which is used to integrate the parts of the study logically to ensure that the problem of research is well addressed (Wanjohi, 2014).

The study used a Quantitative Research design. This emphasizes the objective of measurement, statistical and numerical analysis of data. The study aims to identify the best machine learning model to be used for fraud detection in healthcare claims. The data collected will be transformed into numerical form and passed through different machine learning models.

We also used experimentation as a research method. We took the exploratory approach when building the machine learning models. We experimented with several machine learning models in order to determine which among them was the best.

The study utilizes the CRISP-DM model to go through the iterative steps.

3.3 Overview of CRISP-DM

Cross-Industry Standard Process for Data Mining, CRISP-DM was created in 1996 by a group of organizations working in the data-mining field through an initiative sponsored by the European Commission. The methodology is a framework used for planning and managing projects. We have chosen this model due to its high level of flexibility and ability to perform regular iterations. It consists of six steps which are discussed below:

Figure 3.1 shows the CRISP-DM process diagrammatically.

Figure 3.1 : CRISP -DM Process ,Teichmann (2020)

3.3.1 Business Understanding

Understanding the problem domain is very important. We used both primary and secondary sources of literature review to understand the problem. I had a meeting with some stakeholders from the firm that provided data and got to understand the issues and challenges they face and the kind of solution they needed. This assisted me to select the appropriate techniques to be used in my research.

3.3.2 Data Collection

The data was provided by an Insurance company in Kenya. Due to the sensitivity of data, a non-disclosure agreement was signed to ensure the data and information gained was safe. During the collection, a meeting that included some stakeholders discussed which major attributes contributed to a claim being fraudulent on not. This then directed on which variables should be fetched from the database. The data was then queried from the database of the insurance company and saved as a CSV file. It was shared offline

The data included claim records of maternal healthcare from different service providers. The dataset contained a total of 84,260 entries with 19 columns. The following diagram shows a representation of the dataset.

```
In [5]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84260 entries, 0 to 84259
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   member_number       84260 non-null  object
 1   gender              84260 non-null  object
 2   cover_group         84260 non-null  object
 3   county_name         84260 non-null  object
 4   county_code         84260 non-null  int64
 5   claim_id            83827 non-null  object
 6   provider            84260 non-null  object
 7   hospital_level      84260 non-null  object
 8   category_name       84260 non-null  object
 9   admission_date      84260 non-null  object
 10  discharge_date      81448 non-null  object
 11  no_of_days          81448 non-null  float64
 12  option_name         84260 non-null  object
 13  desease_code        84248 non-null  object
 14  desease_name        84260 non-null  object
 15  practitioner_number 75076 non-null  object
 16  total_claim_amount  84260 non-null  float64
 17  bill_amount         84260 non-null  float64
 18  paid                13616 non-null  object
dtypes: float64(3), int64(1), object(15)
memory usage: 12.2+ MB
```

Figure 3.2 Description of the Dataset

3.3.3 Data Understanding

In this phase, we familiarized ourselves with the data that was collected. We checked the quality of data, identified major and important features from the dataset, and identified the interesting subsets and subgroups of the data. We also did data Analysis just to get familiar with the data and features of the dataset. Here are a few findings we got from the data analysis done

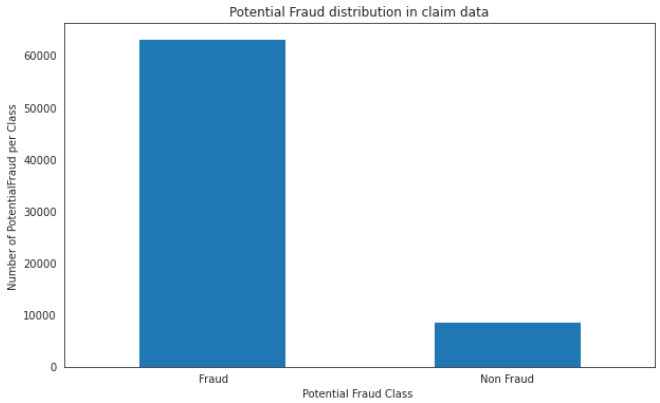


Figure 3.3 Distribution of Claim Records

In figure 3.3 We can note that we had more possible fraud data as compared to non-fraudulent data

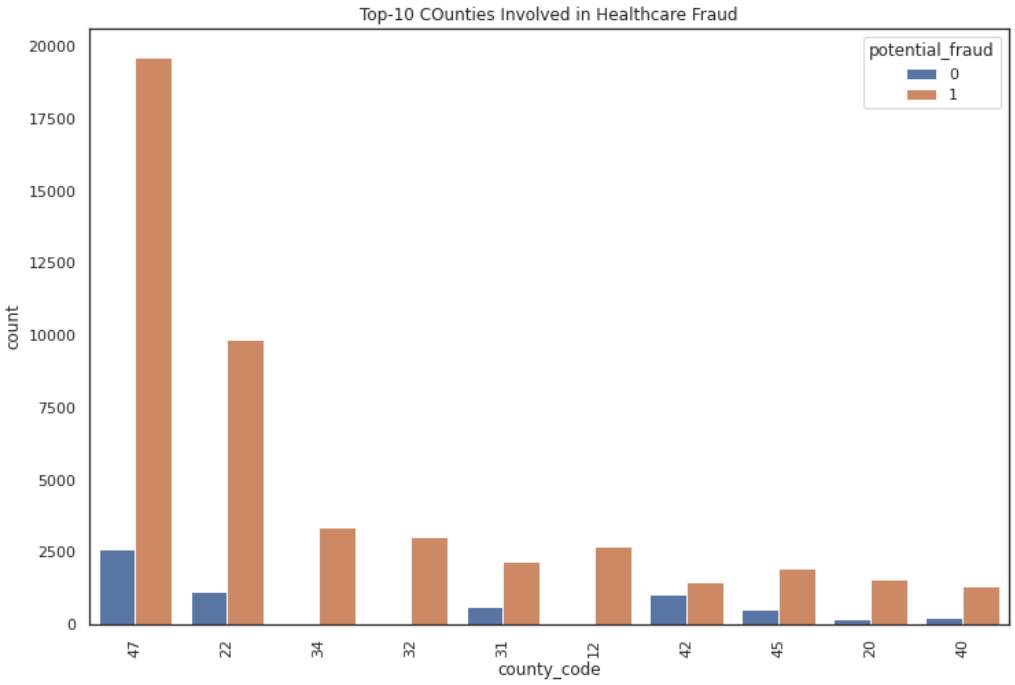


Figure 3.4 Top 5 Counties Involved in Fraud

In figure 3.4, we can note that county 47, which is Nairobi county, is the highest in fraud cases followed by county 22 which is Kiambu county.

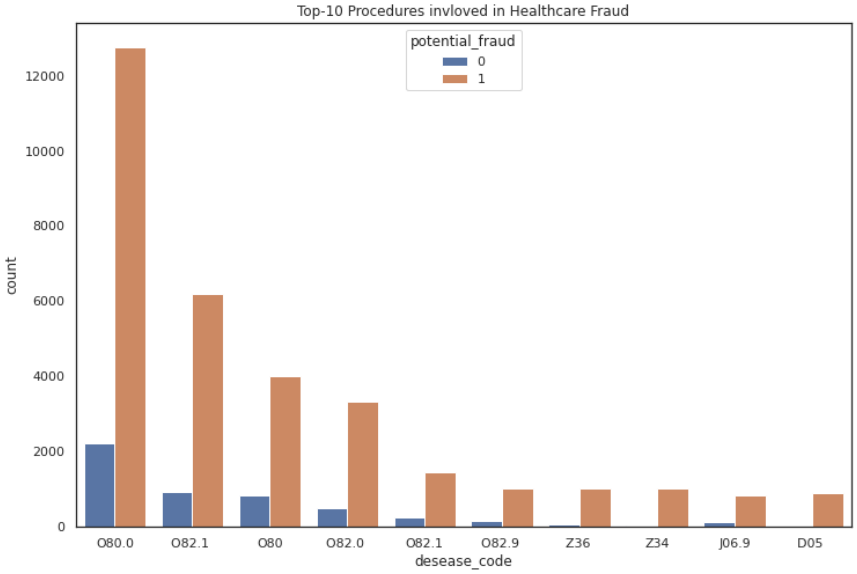


Figure 3.5 Top Procedures Involved in Fraud

Figure 3.4 shows the top procedures involved in fraud. Here we can see procedure 080, was highly involved in fraudulent cases

Using the random forest, we were able to plot a graph of the feature importance of the various variables in the dataset. So as the value tends to be close to 1, the more influence it has on the outcome of the prediction of the class. The most important feature was identified as the bill amount. Below is a graph that shows the top 5 important features

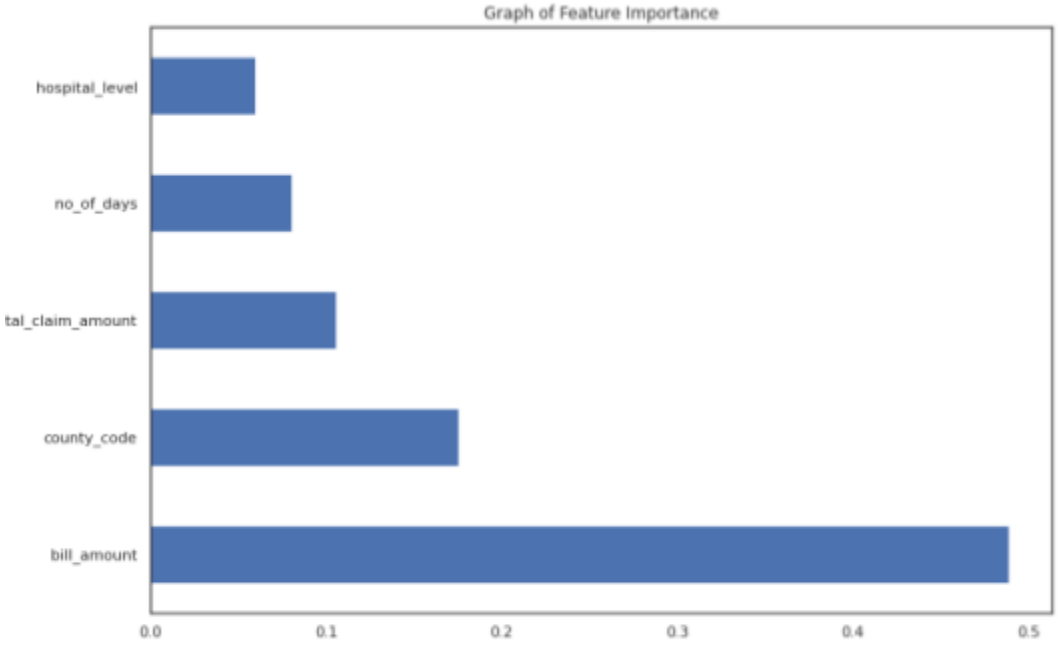


Figure 3.6 Graph of feature importance

3.3.4 Data Preparation

In this phase, the dataset was cleaned, integrated, transformed, and formatted. we performed the below steps;

1. The data that was collected was classified into two major categories , Paid claims and unpaid claims. The paid claims were considered non-fraudulent claims and the unpaid claims were considered as potential fraudulent claims. A python function was used to create the labels for the dataset.
2. Secondly, the dataset contained some missing columns which we dropped since some of the key columns were missing. The remaining final dataset had 50,450 entries which were sufficient for training.
3. Thirdly, The categorical columns were converted to numerical using one-hot encoding, target encoding, and dictionary encoding based on the categories available.
4. We normalized our dataset using a Scaler preprocessing class called the MinMaxScaler

method.

3.3.5 Model Development

In this phase, several machine learning algorithms were used to create models. The models included; Logistic regression, Random forest, Naive Bayes, Support Vector Machine, Gradient Boosting Classifier, and Neural Networks. The development of these models was done using python programming.

3.3.6 Evaluation

In this phase, we checked how the different machine learning models performed. This assisted us to highlight the strengths and weaknesses of our models. The criteria we used were based on Performance metrics. This included: Confusion matrix, ROC-Area under Curve, Precision, Recall, F1 Score, Specificity, Sensitivity, and Accuracy. The subsections below explain more about the evaluation matrix listed above.

3.3.6.1 Confusion matrix

A confusion matrix is a two by two binary classification matrix with actual values on one axis and predicted on another. The matrix contains the following metrics :

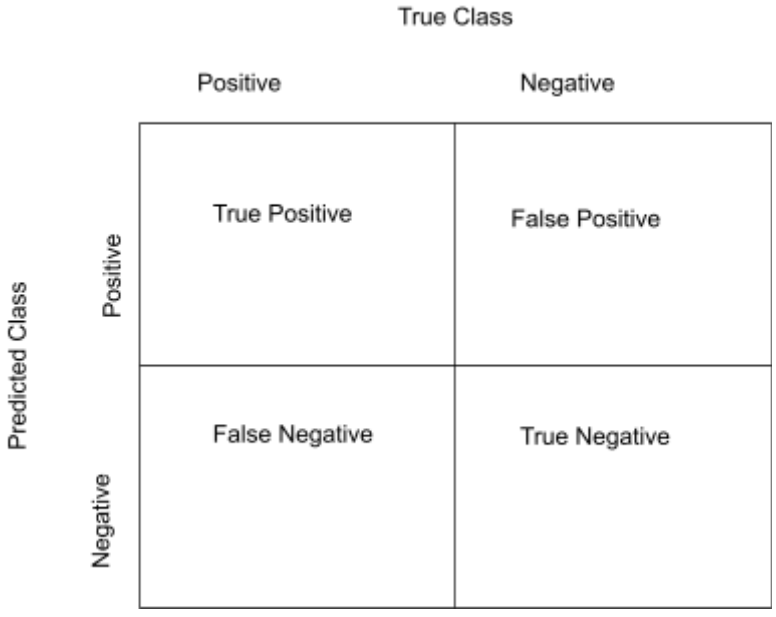


Figure 3.7 : Confusion Matrix

True Positive values are a measure of the class which is predicted as true and the actual category class is true. True Negative is a measure of the class that is predicted as false and the actual category is false. False Positive is a measure of the class that is predicted as true and the actual category is false. False-negative is a measure of the class which is predicted as false, but the actual category is true.

3.3.6.2 Classification Report

Precision is the ratio of correctly predicted positive observations over the predicted total positive observations. The question that this metric answer is of all claims which are classified as fraudulent, how many are fraudulent

The recall is defined as the ratio of correctly predicted positive observations to all observations in the actual class. The question recall answers are: Of all the claims that are not fraudulent, how many did we label correctly?

F1 Score is defined as the weighted average of Precision and Recall.

Accuracy is defined as the number of classifications a model predicts correctly divided by the total number of predictions.

3.3.6.3 ROC - Area under the curve

ROC is the representation for Receiver Operating Characteristics. This is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests. Area Under the Curve which can be represented as AUC is the ability of a model classifier to distinguish among negative and positive classes. A higher AUC represents a higher performance of the model at distinguishing between the positive and negative classes.

3.3.7 Deployment

Deployment is the process where a machine learning model is integrated into the production environment to be able to make decisions based on data fed. To be able to utilize the model, we

developed a prototype to be used to predict fraudulent claims. We used python Django to develop the prototype.

3.4 Prototyping

Prototyping is the process of developing and creating a working replica of a system or product. Our research implemented the best-performing machine learning model to be used to detect a fraudulent claim. The prototype elements are discussed below:

3.4.1 System Architecture

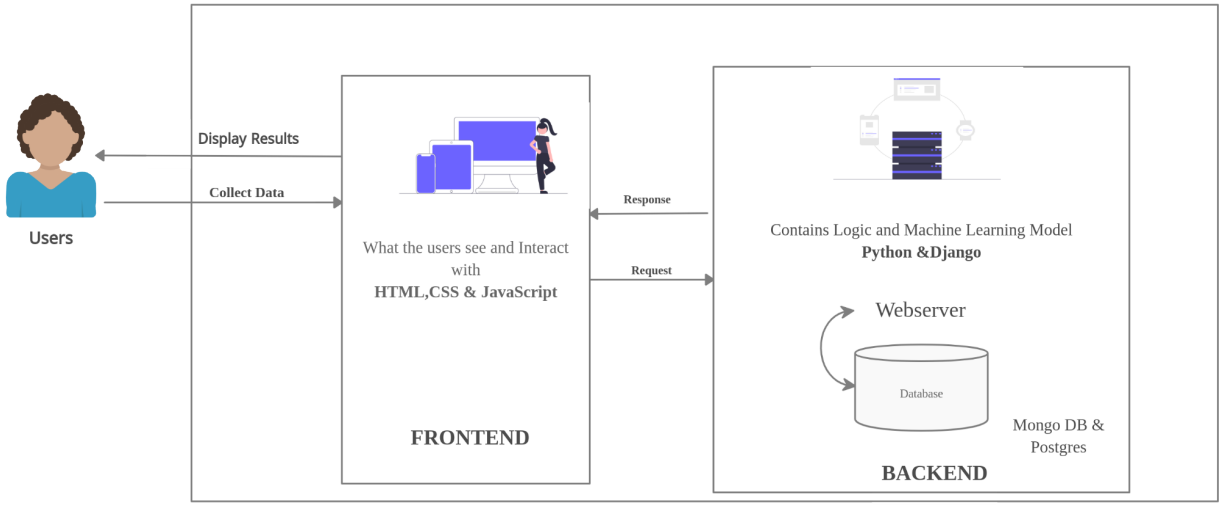


Figure 3.8 System Architecture

3.4.1.1 Users

The users collect data using the frontend and receive responses from the model

3.4.1.2 Frontend

This is the access point where the users interact with the system. It has a user input section, where the users enter claim data, and the user output where the user will know the claim status;

whether fraudulent or not. It also has buttons that can be used to flag claims. HTML, CSS, and JavaScript are the languages used to develop.

3.4.1.2 Backend

The backend contains the Web Server. The Model, the Logic, and files are hosted here. This is coded in Python & Django and the database is also hosted here

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This section highlights the findings of our research. it presents the performance metric results of the varying machine learning methods used. it shows the validity and viability of the model and draws appropriate conclusions

4.2 Performance Metrics

4.2.1 Logistic Regression

The Area Under the Curve was noticeably low with a value of 73.0%. The sensitivity was at 58.6% while the specificity was at 67.0%.

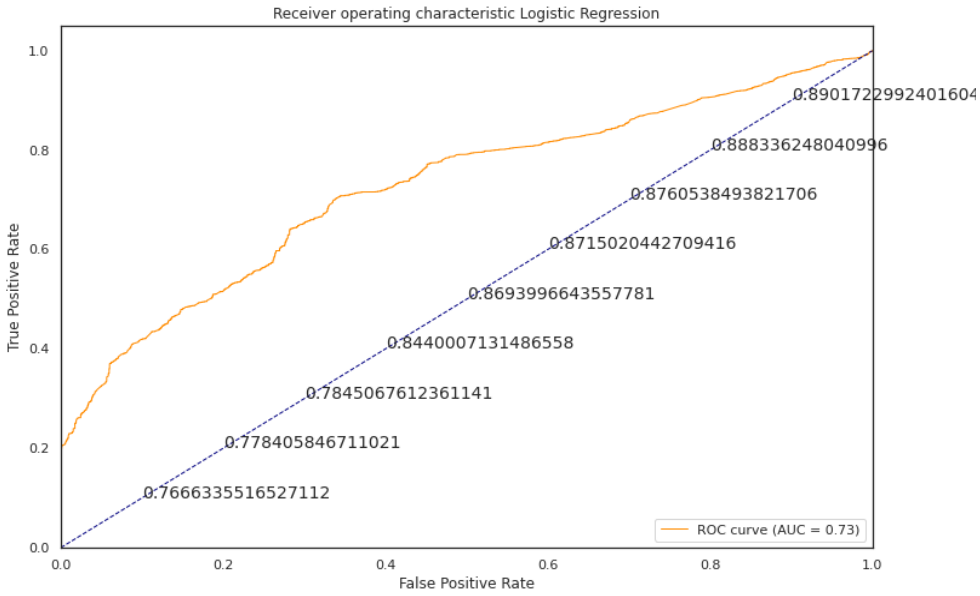


Figure 4.1 AUC Logistic Regression

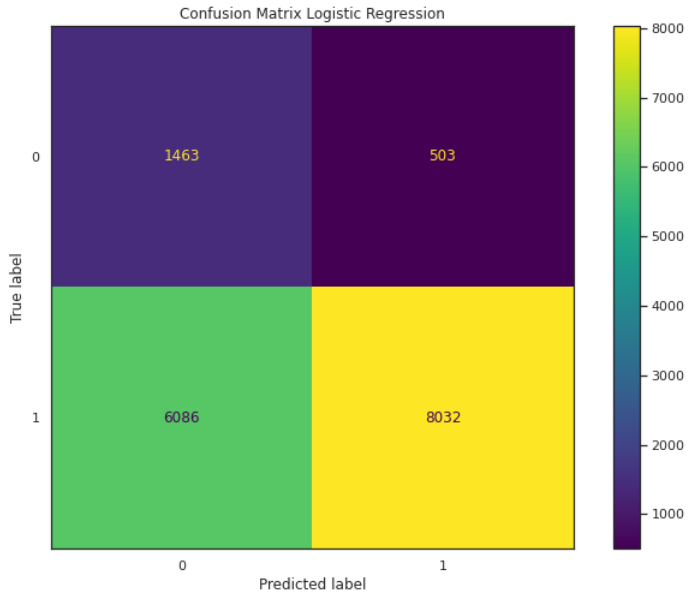


Figure 4.2 The Confusion Matrix for Logistic Regression

4.2.2 Naive Bayes Classifier

The area under the curve was also noticeably lower as compared to Logistic regression, with a score of 68%. The specificity was 65.425, and the sensitivity was 34.65%. Naive Bayes, we achieved a weighted precision of 0.76, a weighted recall of 0.77, and a weighted F1-score of 0.77. The accuracy of the model was 76.65%. The Naive Bayes classifier predicted 5339 fraudulent claims correctly and 1606 non-fraudulent claims correctly. The classifier had false negatives of 360 claims. The classifier tends to predict fraudulent claims correctly.

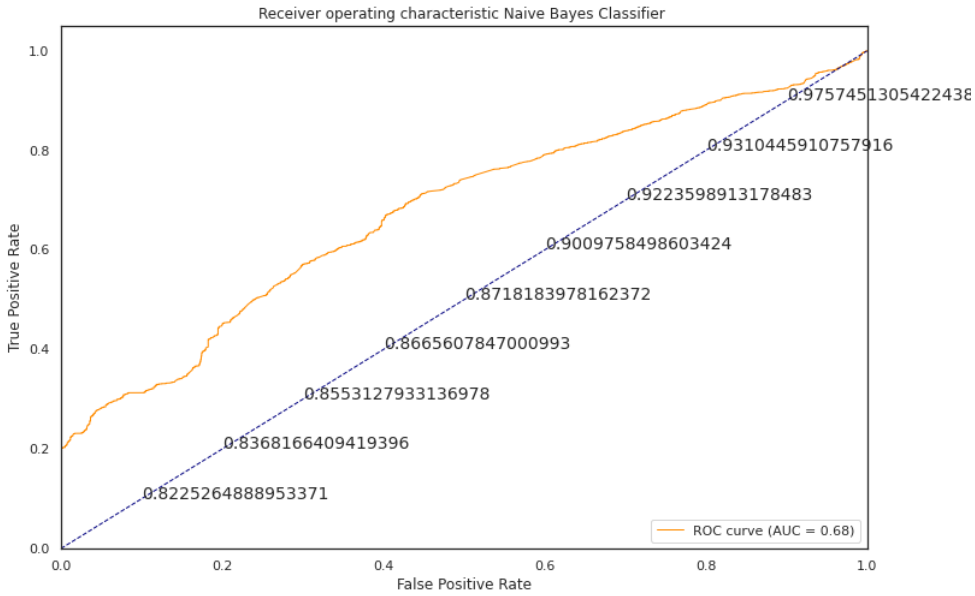


Figure 4.3 AUC Naive bayes

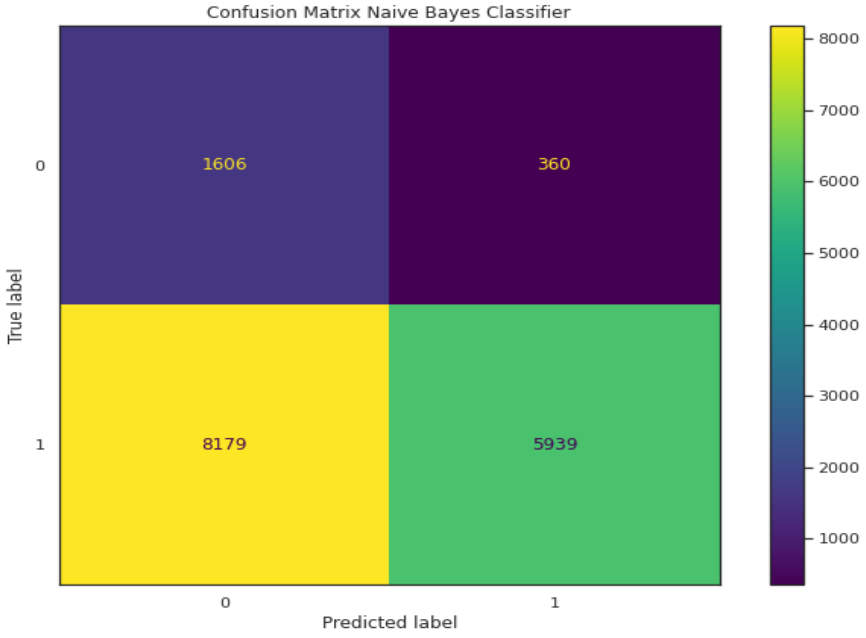


Figure 4.4 Confusion Matrix for Naive Bayes Classifier

4.2.3 Support Vector Machine

The support vector machine performed noticeable low as compared to logistic regression and Naive Bayes. It had a ROC - area under the curve of 66.0%. The specificity of the model was 99.97% and sensitivity of 0.2746%.

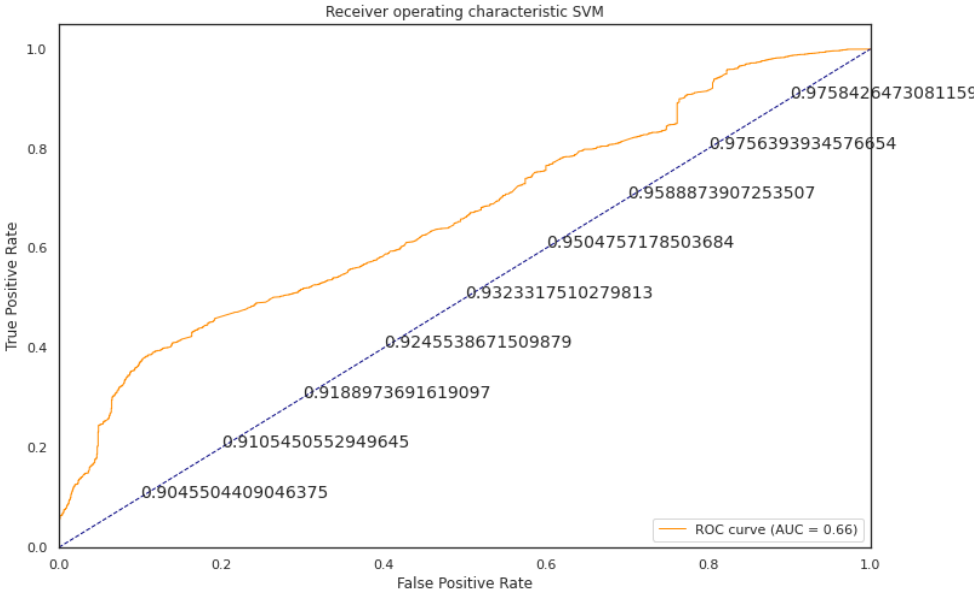


Figure 4.5 AUC Support Vector Machine

The weighted precision was 0. 88, the recall is 0.88 and the f1-score is 0.84. The general accuracy of the model is 88.07%. The model correctly predicted 14112 fraudulent claims. The model tends to predict fraudulent claims better. On the contrary, a and has a very difficult time predicting non-fraudulent claims. 54 records were only predicted correctly.

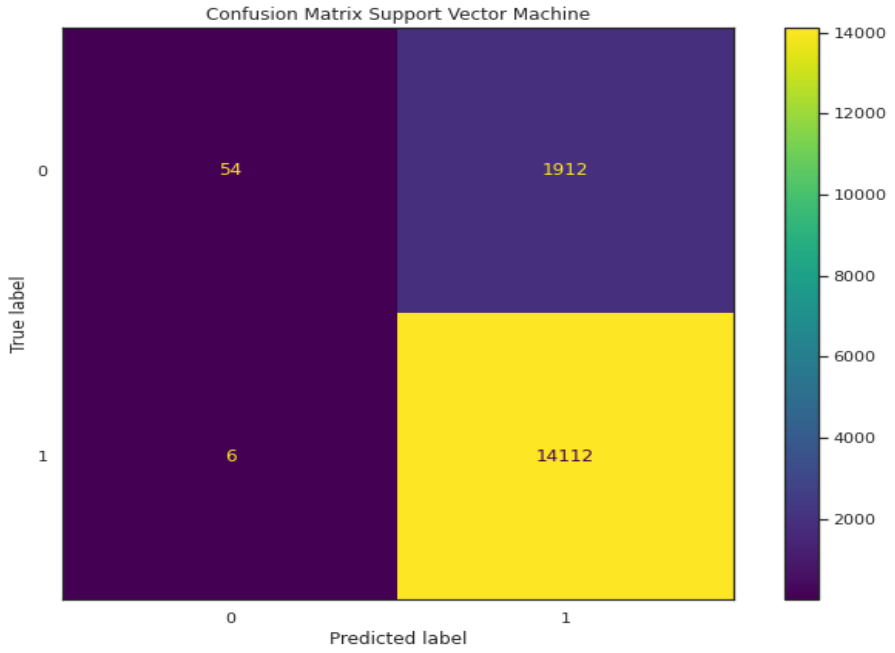


Figure 4.6 Confusion Matrix for Naive Bayes Classifier

4.2.3 Random Forest Classifier

The random forest classifier performed better than the previous three models. The ROC curve area under the curve was 90.0%. The model had a Sensitivity of 51.23% and a Specificity of 96.89%.

```
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)
```

```
{'n_estimators': [100, 200, 300, 400, 500, 600, 700, 800, 900], 'max_features': ['auto', 'sqrt'], 'max_depth': [1
0, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2,
4], 'bootstrap': [True, False]}
```

Figure 4.7 Hyper Parameter Tuning Random Forest

We also performed hyperparameter tuning using the RandomizedSearchCV. Figure 4.7 shows the parameters we experimented with the best combination of parameters are listed in the table below

Hyperparameter	Value
n_estimators	400
min_samples_split	5
min_samples_leaf	1
max_features	auto
max_depth	20
bootstrap	False

Table 4.1 Best Hyperparameters for Random Forest

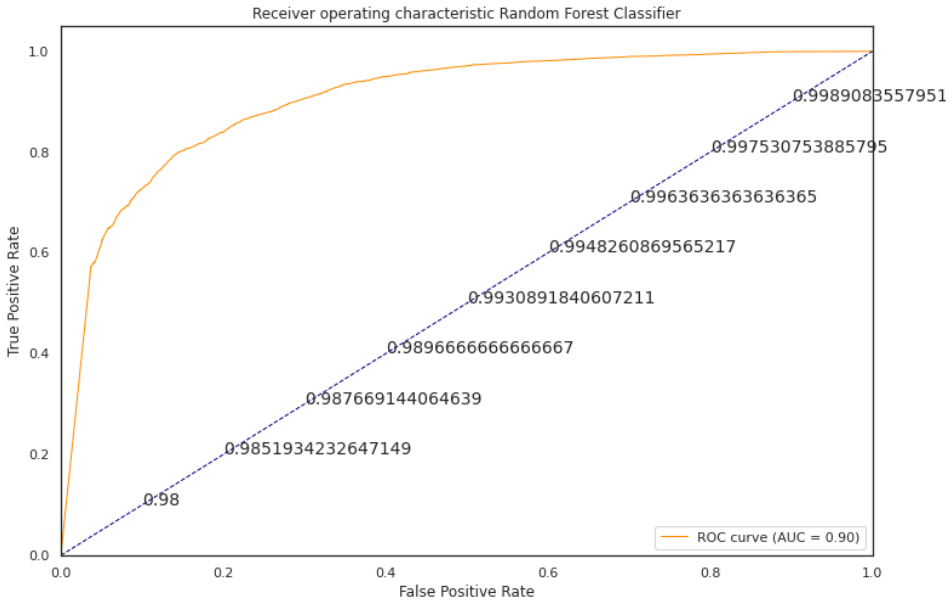


Figure 4.8 ROC Random Forest

The general accuracy of the model was 91.27%. The weighted average for precision is 0.91, the recall was 0.91, and the feed-forward f1-score of 0.91 remained cross-entropy the same. The model leaned towards predicting fraudulent claims better. A total of 13,676 fraudulent claims were predicted correctly and1005 non-fraudulent claims were predicted correctly. This so far is the highest number of correctly classified claims.

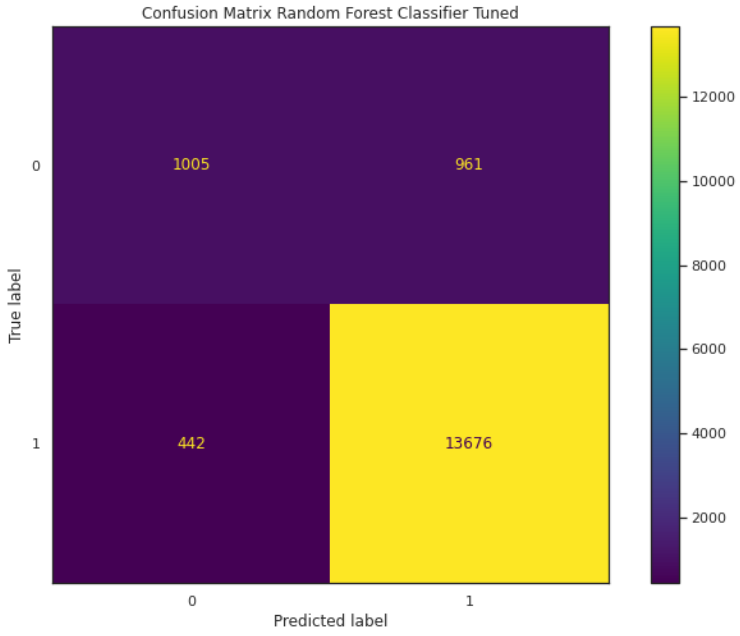


Figure 4.9 Random Forest Tuned Confusion Matrix

4.2.4 Artificial Neural Network

The feed-forward neural network underperformed against the gradient-boosting classifier. We used two neural networks algorithms. the Sequential model and the MLP classifier. The sequential model was run through various epochs. The model used the SGD optimizer and binary cross-entropy as the loss function. GridsearchCV was used for hyperparameter tuning to get the best combination of epochs and batch size. The table below shows the best parameters from the tuning of the SGD model.

Hyperparameter	Value
activation	softmax
batch_size	335
dropout	0.4361
dropout_rate	0.2307
epochs	44
learning_rate	0.4260
neurons	31
normalization	0.3376
optimizer	sgd

Table 4.2 Best Hyperparameters for Neural Network Sequential

The sequential model had an accuracy of 87.87% and a test loss of 0.3715. The MLP classifier performed better than the sequential model,the general accuracy of the model was 88.30%. ROC area under the curve of 100%. The specificity was 99.04% and sensitivity was 0.95% and The diagram below shows the ROC of the MLP model

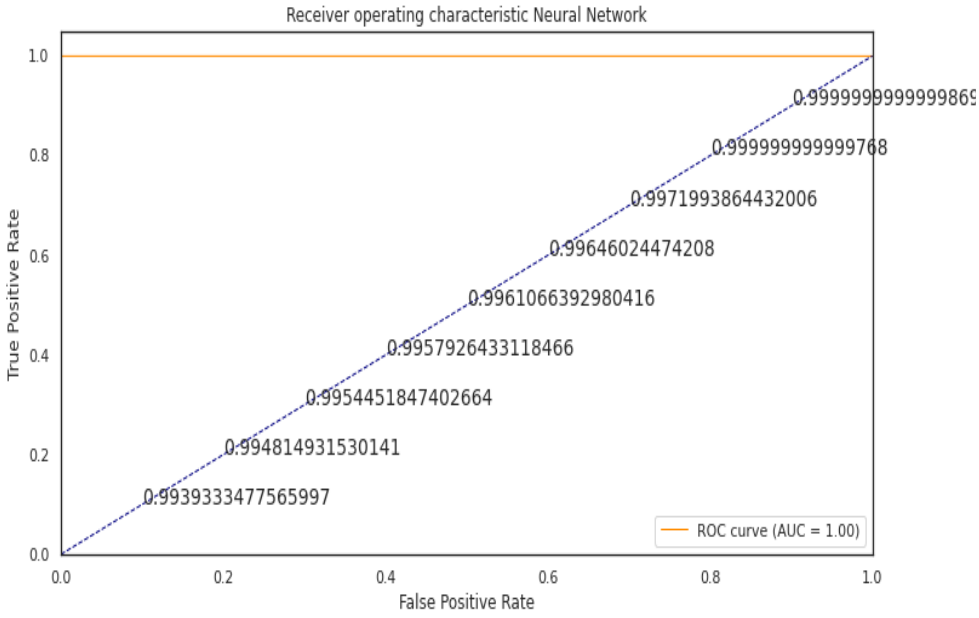


Figure 4.10 RUC MLP Neural Network

The weighted precision was 0.86, the weighted recall was 0.88 and the weights average of the f1-score was 0.85. The neural network predicted 13915 fraudulent claims correctly and 200 non-fraudulent claims correctly. The model leaned towards predicting fraudulent claims better.

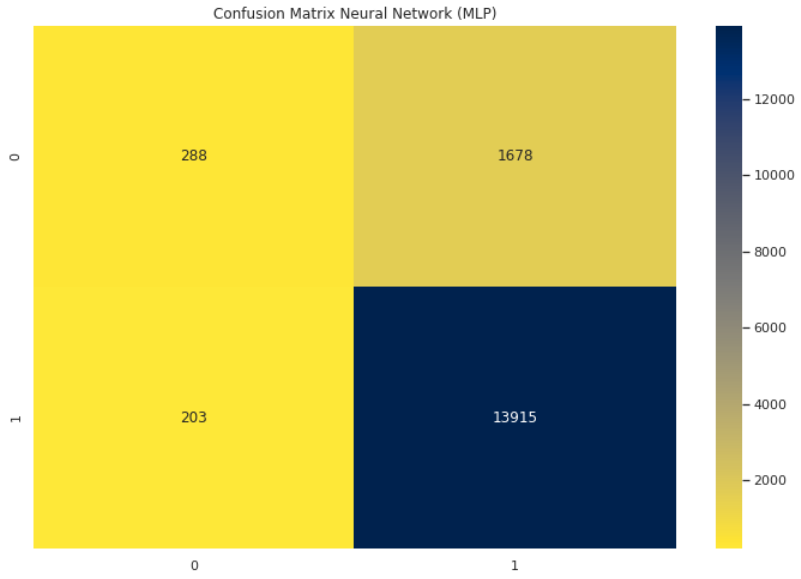


Figure 4.11 MLP Neural Network Confusion Matrix

4.2.5 Gradient Boosted Tree Classifier

The gradient boosting classifier performed the best among all the models. The model is a very powerful learner. We used the GridsearchCV for parameter tuning to ensure the model was efficient. The parameters that we experimented with are the learning rate (0.1,0.3,0.7,0.9) , the maximum_depth of (4,5,7,8) and n_estimators (2,3,4,5,6,7,8,9,11,13). The best parameters chosen are indicated in the table below

Hyperparameter	Value
learning_rate	0.7
max_depth	8
n_estimators	11

Table 4.3 Hyperparameters for Gradient boosting classifier

Figure 4.12 represents this below

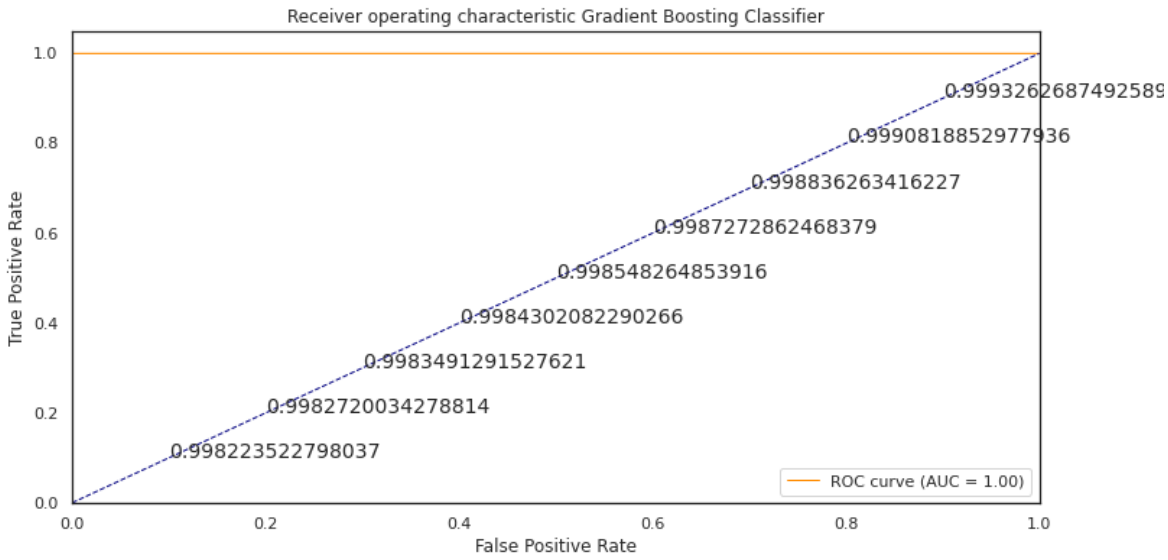


Figure 4.12 ROC Gradient Boosting Classifier

The weighted average for precision was 0.89, recall 0.90, f1-score of 0.89. The model recorded a sensitivity of 36.26% and specificity of 97.78%. The classifier was able to predict 13805 fraudulent claims correctly and 713 non-fraudulent claims well. The classifier performed averagely okay on both ends of the classification. It will be ideal for detecting possible fraudulent claims.

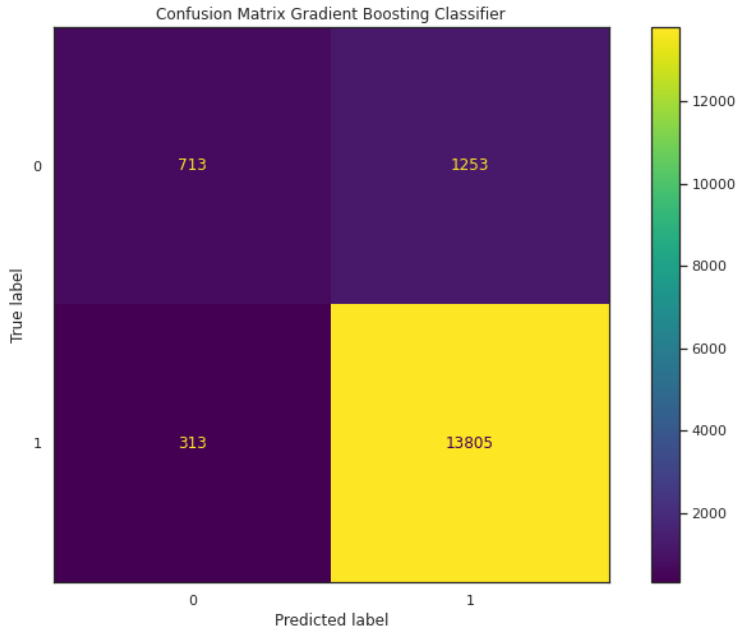


Figure 4.12 GBC Confusion Matrix

Based on the results, we can see that we have a low rate of false positives and low false negatives. This means that we can identify threats more and thus we can pass on a few false alarms. The specificity of the model was quite low. All the models struggled to classify the non-fraudulent claims.

4.3 Comparison of Algorithms Used

In this section, we will compare how these machine learning algorithms compare to each other. The table below shows the comparison table representation of how the algorithms performed

ML Algorithm	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	AUC	Sensitivity	Specificity
Logistic Regression	0.59	0.85	0.59	0.66	0.73	0.74	0.57
Naive Bayes	0.47	0.85	0.47	0.54	0.68	0.81	0.42
Support Vector Machine	0.88	0.88	0.88	0.84	0.66	0.027	0.99
Random Forest	0.91	0.90	0.91	0.90	0.90	0.52	0.97
Gradient Boosting Classifier	0.90	0.89	0.90	0.89	1.0	0.36	0.97
Neural Network	0.88	0.77	0.88	0.82	0.89	0.0	0.1
Multilayer Perceptron NN	0.88	0.86	0.88	0.85	1.0	0.12	0.99

Table 4.4 : Comparison of Algorithms

Based on the AUC, we can see that the Neural network model, gradient boosting classifier, and random forests performed better. The scores were 1.00, 1.00, and 0.88 respectively. The support vector machine, Naive Bayes, and logistic regression performed poorly in terms of the AUC. The scores were 0.68,0.67 and 0.66. This shows that the models had a difficult time classifying the claims.

Based on the weighted precision, recall, and f1-score the Random Forest and Gradient boosting classifier performed well. They had values of 0.89 and above. This made them better-performing models.

In terms of accuracy, the Gradient Boosting Classifier and Random forest were high with scores of 0.91 and 0.90 respectively.

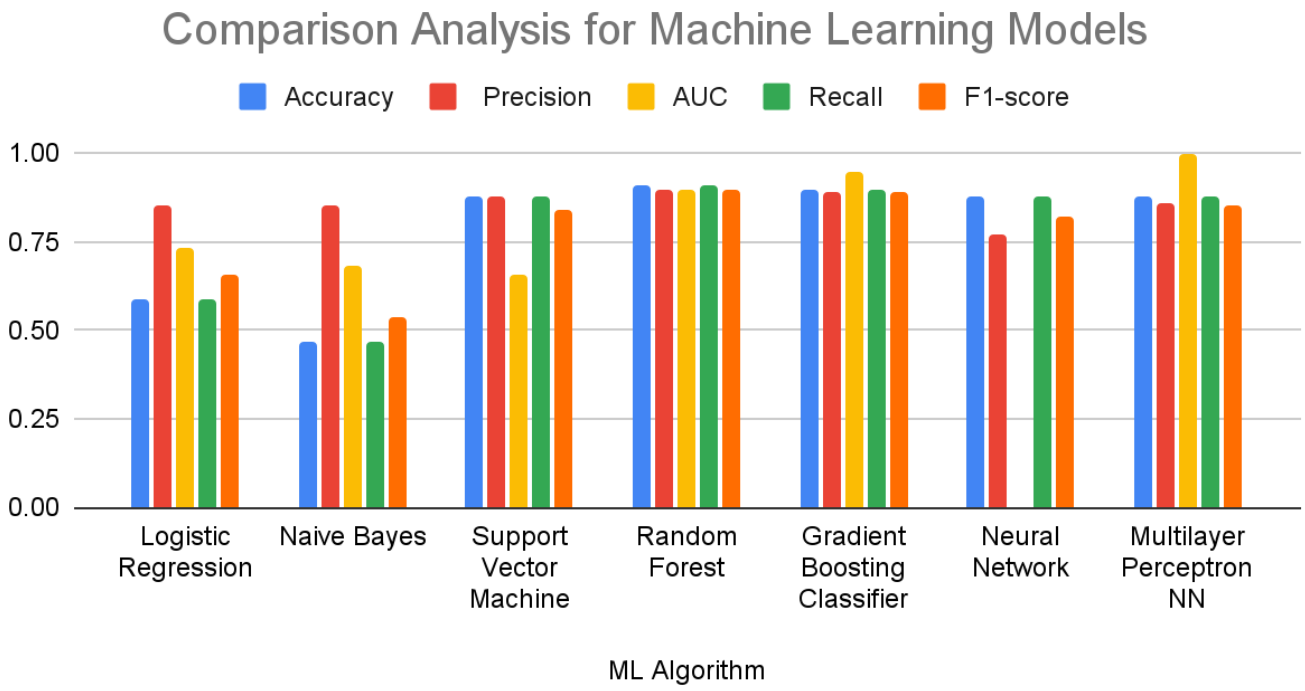


Figure 4.13 Graphical Comparison Analysis

Based on the four parameters in graph 4.13, Gradient Boosting Classifier, and Random Forest were the models that performed better than the others.

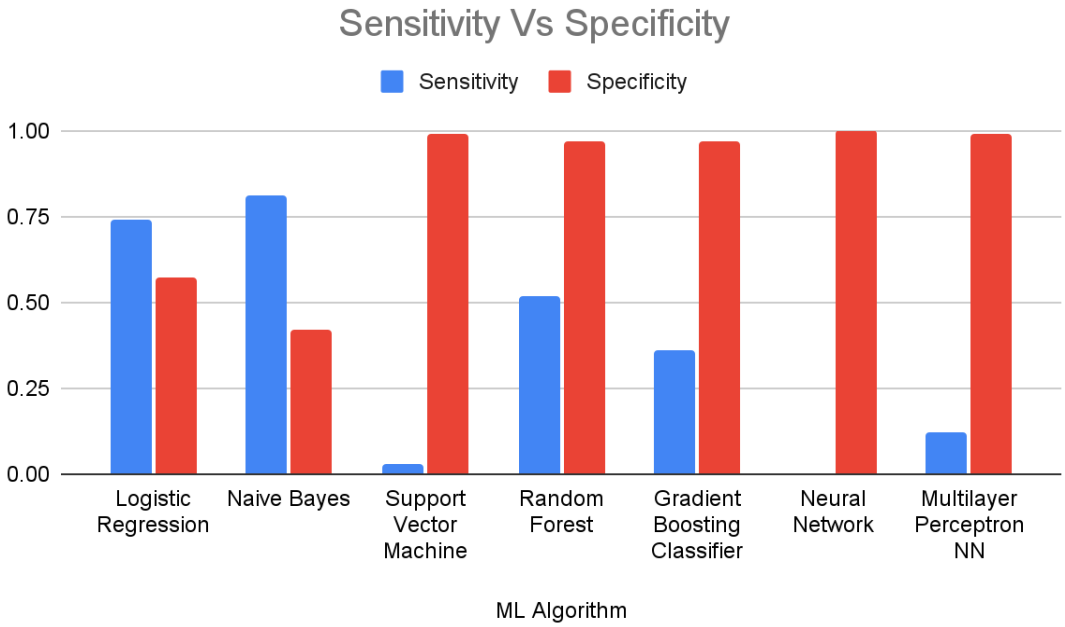


Figure 4.14 Graphical Comparison Analysis Sensitivity and Specifity

Generally, All the models had a very low sensitivity as compared to specificity. The neural network had a very high specificity of 100%. This means the models were able to identify fraudulent claims correctly. The downside was that they had a very low chance of detecting non-fraudulent claims due to the sensitivity rate. The Gradient Boosting Classifier performed had a specificity score of 97.78% and a sensitivity score of 36.26%. This was a better performance as compared to other algorithms as indicated in figure 4.13

In conclusion, looking at all our benchmarking parameters, the model that performed well in the classification of our dataset was the Gradient Boosting Tree Classifier. This is the model that was chosen to be deployed in our prototype.

4.4 The Prototype

The prototype is a python based application that runs on the Django framework. The system launches with a home page that requests users to enter in key details of a claim. The Figure below 4.14 shows the launch page.

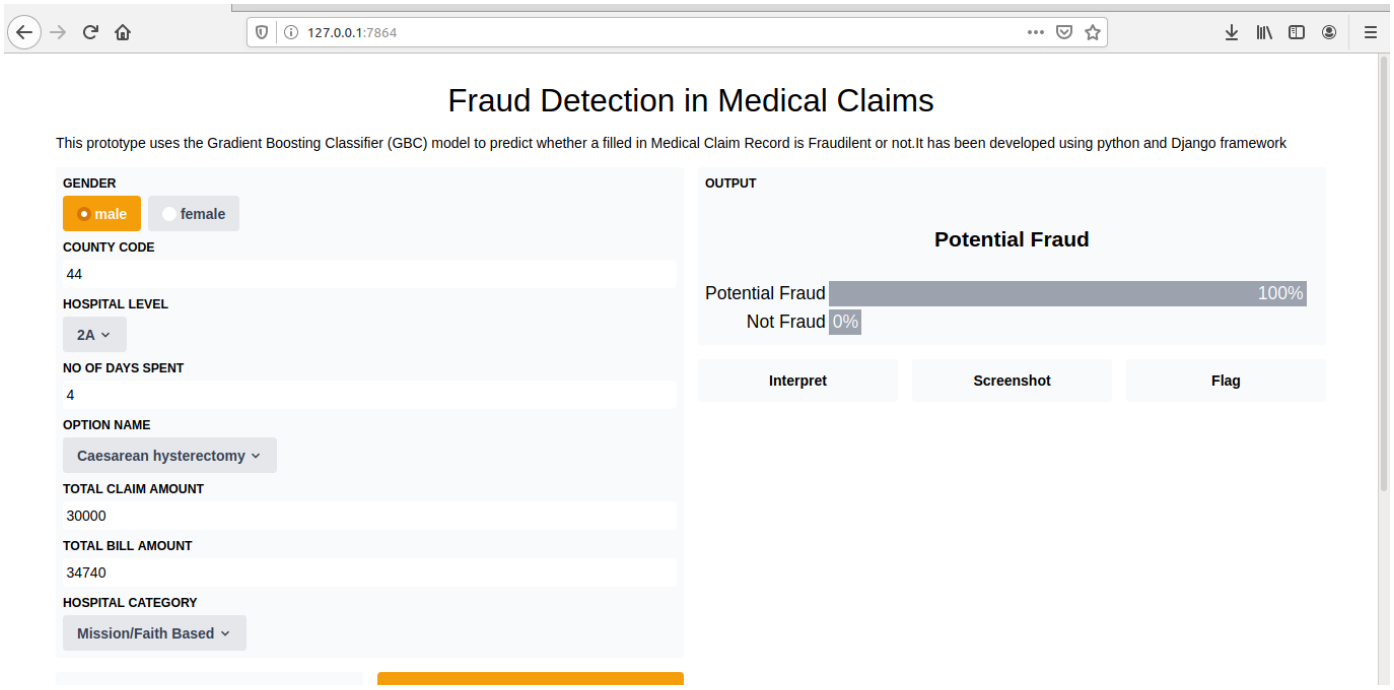


Figure 4.15 Launch Page

The user then is required to pass details and click the submit button. This will pass the details through the saved machine learning model which is the Gradient Boosting Classifier which will perform a prediction and indicate whether it is potential fraud or not with percentage probability. Figure 4.14 shows a prediction of a fraudulent case using the system.

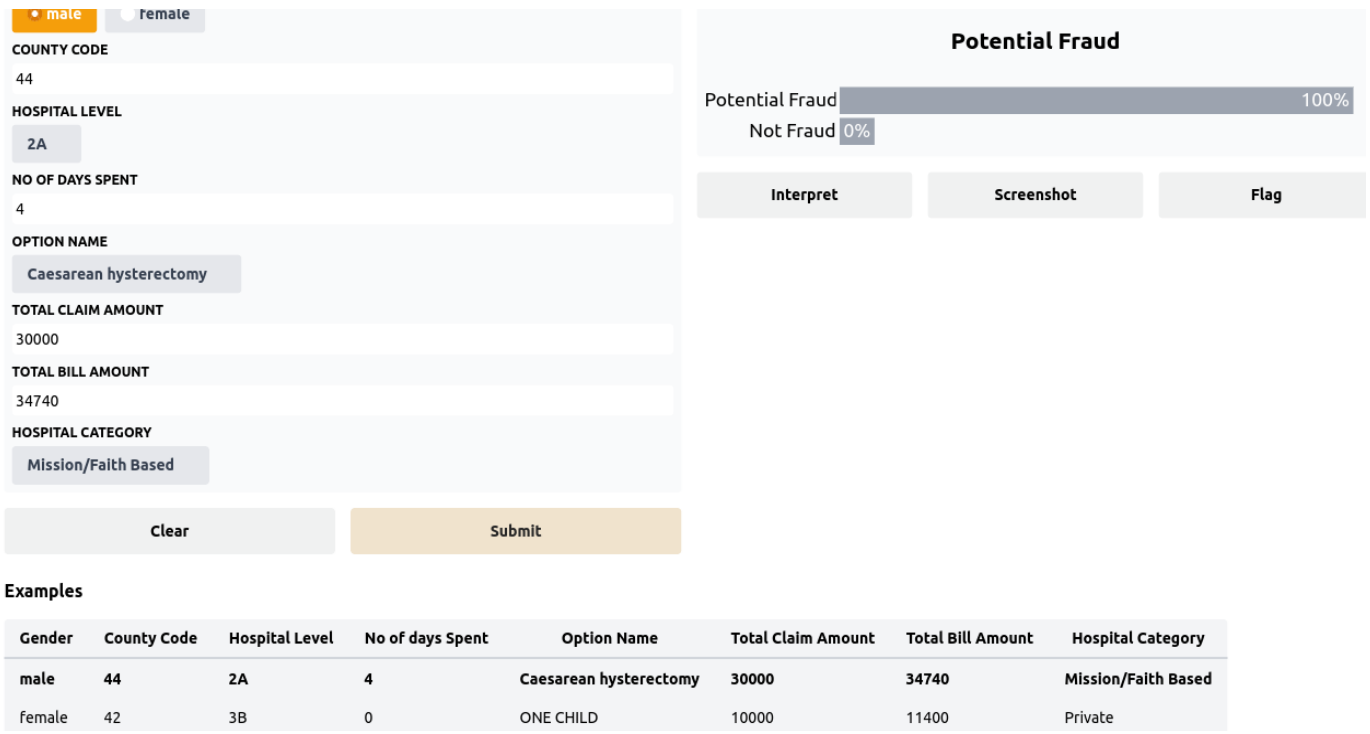


Figure 4.16 Fraud Detection Output

The system then contains buttons that can be used to flag a particular claim and screenshot results. The Flagged Claims are saved in a local file as a CSV. The code is found in the appendix section of our project

4.5 Conclusion

This chapter brought out a clear understanding of the strengths and weaknesses of different algorithms by using different benchmarking parameters and performance measures discussed in chapter three. The Logistic Regression, Naive Bayes, and Support Vector Machine performed poorly based on the performance metrics used. The Random Forest, Neural Network using MLP classifier and Gradient Boosting Tree Classifier were the best performing models in that order.

The models were also analyzed in terms of resource metrics, robustness, and characteristics of data used. The models created used minimum resources and were able to run on local servers without any problem

CHAPTER FIVE
CONCLUSION

5.1 Introduction

This chapter will give a general conclusion of our research. It will also check on how our set objectives were met, a critique of our machine learning models, and give reference of future work.

5.2 Achievements

The overall objective of the study was to create, train and test a machine learning-based mode which is able to detect fraudulent claims in healthcare . The objective was met by designing, implementing, and deploying the gradient boosting tree classifier model which can detect fraudulent claims.

The first objective was to identify features in insurance claims that can be used for fraud detection. The study made use of data from a local insurance company based in Kenya. From the database, key variables were chosen and feature engineering was done based on the different variables selected. Using the random forest algorithm, we were able to identify the feature importance of the different variables listed thereby solving the first objective.

The second objective was to identify appropriate machine algorithms to use for fraud detection. From the literature review, we were able to look at research that has been carried out and which machine learning algorithms were used. From that, we were able to select the following algorithms: Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, Gradient Boosting Tree Classifier, and Neural Networks. The algorithms were applied to the dataset and results were derived from it thereby meeting the second objective.

The third objective was to compare the performance of different machine learning algorithms. A comparison analysis was done based on the following performance metrics. Accuracy of the model, average weighted Precision, average weighted recall, average weighted f1-score, the area under the curve, specificity, and sensitivity. Other key parameters that were used include;

resource metrics, robustness, and characteristics of data used. The Logistic regression, Naive Bayes, and support vector machine performed poorly based on the performance metrics used while the Random forest, neural network using MLP classifier, and gradient boosting tree classifier were the best forming models in that order. This thereby met the third objective.

The fourth objective was to implement an prototype that can be easily used to mark a claim as fraudulent or not. This was achieved by implementing and deploying the gradient boosting classifier model. The prototype contains a user interface, where the user gives details of a claim and it predicts whether the claim is fraudulent or not.

5.3 Limitations of the Study

The dataset that was used for the study was imbalanced.it contained more fraudulent claims as compared to non-fraudulent data. This made the models biased during prediction. Also, the data used for this study involved only maternity claim records thus limiting our solution to this sector only.

5.4 Recommendation for future work

During the study, getting locally available data that is correctly labeled was a major problem. Research can be done using unsupervised learning or semi-supervised learning to cater for data that is unlabeled which will be a great solution in the field.

Our study used classical machine learning algorithms and an artificial neural network. Further research can be done using deep learning to see how the model would perform in the detection of fraud in healthcare.

REFERENCES

Abuya, T., Maina, T., & Chuma, J. (2015, February 12). Historical account of the national health insurance formulation in Kenya: experiences from the past decade. *BMC Health Services Research*.

Butticè, C. (2019). *Universal Health Care*. Greenwood Publishing Group.

Department Of Health And Human Services, Centers For Medicare & Medicaid Services. (2016). *Medicare Fraud & Abuse: Prevention, Detection, and Reporting*. Medicare Learning Network.

Dunning, T., & Friedman, E. (2018). Rendezvous Architectures (S. Sakr & A. Zomaya, Eds.). *Springer International Publishing*. https://doi.org/10.1007/978-3-319-63962-8_199-1

Dutta, A., Maina, T., Ginivan, M., & Koseki, S. (2018). Kenya Health Financing System Assessment. *Time to Pick the Best Path*.

Gorman, L. (2006). The History Of HealthCare Costs And Health Insurance. *Wisconsin Policy Research Institute*, *19*(10).

Joudaki, H., Rashidian, A., Bidgoli, B. M., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2016). Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study. *International Journal of Health Policy and Management*, *5*(3), 165-172.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2016). Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study. *International Journal of Health Policy Management*, *5*(3), 165–172.

Luther. (2020, July 23). Health insurance fraud and its impact on the healthcare system. *Pacific Prime*. https://www.pacificprime.com/blog/healthcare-system-fraud-impacts.html

Mwaura,, R. N., BArasa, E., Ramana, G., Coarasa, J., & Khama, R. (2015). The Path to Universal Health Coverage in Kenya: Repositioning the Role of the National Hospital Insurance Fund. *IFC SmartLessons; World Bank, Washington, DC.*

Oates, B. J. (2006). *Researching Information Systems and Computing*. SAGE Publications.

Obodoekwe Nnaemeka, F. C. (2017, August 22). A Model for the Automated Detection of Fraudulent Healthcare Claims using Data Mining Methods. *Doctoral Thesis / Master's Dissertation*.

Patil, K. S., & Godbole, P.A. (2018, October 31). A Survey on Machine Learning Techniques for Insurance Fraud Prediction. *Helix Scientific Explorer*, *8*(6), 4358- 4363. https://www.researchgate.net/publication/329448645

Pies, H. E. (2017). *"Control of Fraud and Abuse in Medicare and Medicaid*, *American Journal of Law & Medicine*(3). http://scholar.google.com/scholar?hl=en&lr=&q=info:f_HBBkf-6KUJ:scholar.google.com/&output=viewport&pg=1.

Piper, C. (2017, August). popular health care provider fraud schemes. https://www.acfe.com/article.aspx?id=4294976280

Rosenberg, C. E. (1987). *The Care of Strangers: The Rise of America's Hospital System*. New York: Basic Books. https://repository.library.georgetown.edu/handle/10822/1034768

Starr, P. (2018). *The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry*. NY: Basic Books.

Strides Towards Universal Health Coverage For All Kenyans. (2018). *NHIF Performance Report*.

Switlick, K., Wang, H., Ortiz, C., Connor, C., & Zurita, B. (2015). *Africa Health Insurance Handbook—How To Make It Work* (C. Connor & H. Wang, Eds.). Abt Associates Inc.

Thornton, D., Mueller, R. M., Schoutsen, P., & Hillegersberg, J. V. (2013). Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection. *Procedia Technology*, *9*, 1252 – 1264.

Wanjohi, A. M. (2014). *Social Research Methods Series: Proposal Writing Guide*. KENPRO Publications.

Yao, Y., Sun, Q., & Lin, S. (2014, June 15th). *Detection of Health Insurance Fraud with Discrete Choice Model: Evidence from Medical Expense Insurance in China*. SSRN. https://ssrn.com/abstract=2459343

Yu, H., & Dick, A. W. (2012). Impacts of Rising Health Care Costs on Families with Employment-Based Private Insurance: A National Analysis with State Fixed Effects.

Health Service Research, *47*(5).

F.-M. Liou, Y.-C. Tang and J.-Y. Chen, “Detecting hospital fraud and claim abuse through diabetic outpatient services,” Health Care Manage Science, vol. 11, pp. 353-358, 2008.

S. Rajasekar, P. Philominathan, and V. Chinnathambi, “RESEARCH METHODOLOGY,” Cornell University, New York, 2013.

Government of Kenya (2008) Kenya Vision 2030: A Globally Competitive and Prosperous Kenya. National Economic and Social Council (NESC), Nairobi.

APPENDIX

BETWEEN: .IMITED, hereinafter referred to as "the data owner";

AND: **MARGARET NYAKENO ONYANGO** hereinafter referred to as "the prospective applicant";

Together "the Parties"

WHEREAS THE PARTIES CONFIRM THAT:

The prospective applicant is seeking to refer to data that the data owner owns;

The prospective applicant is seeking to do so for the purpose of software development and analysis ("the Purpose");

The data owner is under an obligation in certain circumstances to share data with the prospective applicant and may in any event choose to do so regardless of that obligation;

The Parties are entering into data sharing negotiations;

A non-disclosure agreement is necessary to reassure the Parties that the use to which any information exchanged or otherwise disclosed during the negotiations will be limited to the legitimate purpose as established in the software development and analysis;

THE PARTIES HAVE THEREFORE AGREED AS FOLLOWS:

1. Disclosure of Information

    a. A Party may disclose to the other Party information with a view to negotiating the sharing of data for a purpose under the software development and analysis ("the Purpose"). The Parties agree that the terms and conditions set forth in this Agreement shall govern any such disclosure of information. All information disclosed by a Party or by Affiliates of a Party to the other Party or its respective Affiliates orally, electronically, writing or by any other means during the data sharing negotiations shall be considered as confidential unless expressly stated otherwise by the disclosing Party. All such confidential information shall be referred to hereinafter as "information". Information shall also include the identity of the Parties, the contents of this agreement and the fact that they have entered into this Agreement.

IN WITNESS WHEREOF, the parties hereto, by their duly authorized representatives, have executed this Agreement:

LIMITED
For and on behalf of:

Name JOSEPH MIXON

Signature

Date 27/09/2021

International Business Machines Ltd.
P.O. Box 46667 - 00100 GPO
NAIROBI - KENYA
Tel: (+254-20) 2102710-13

MARGARET NYAKENO ONYANGO
For and on behalf of:

Name Margaret Nyakeno

Signature

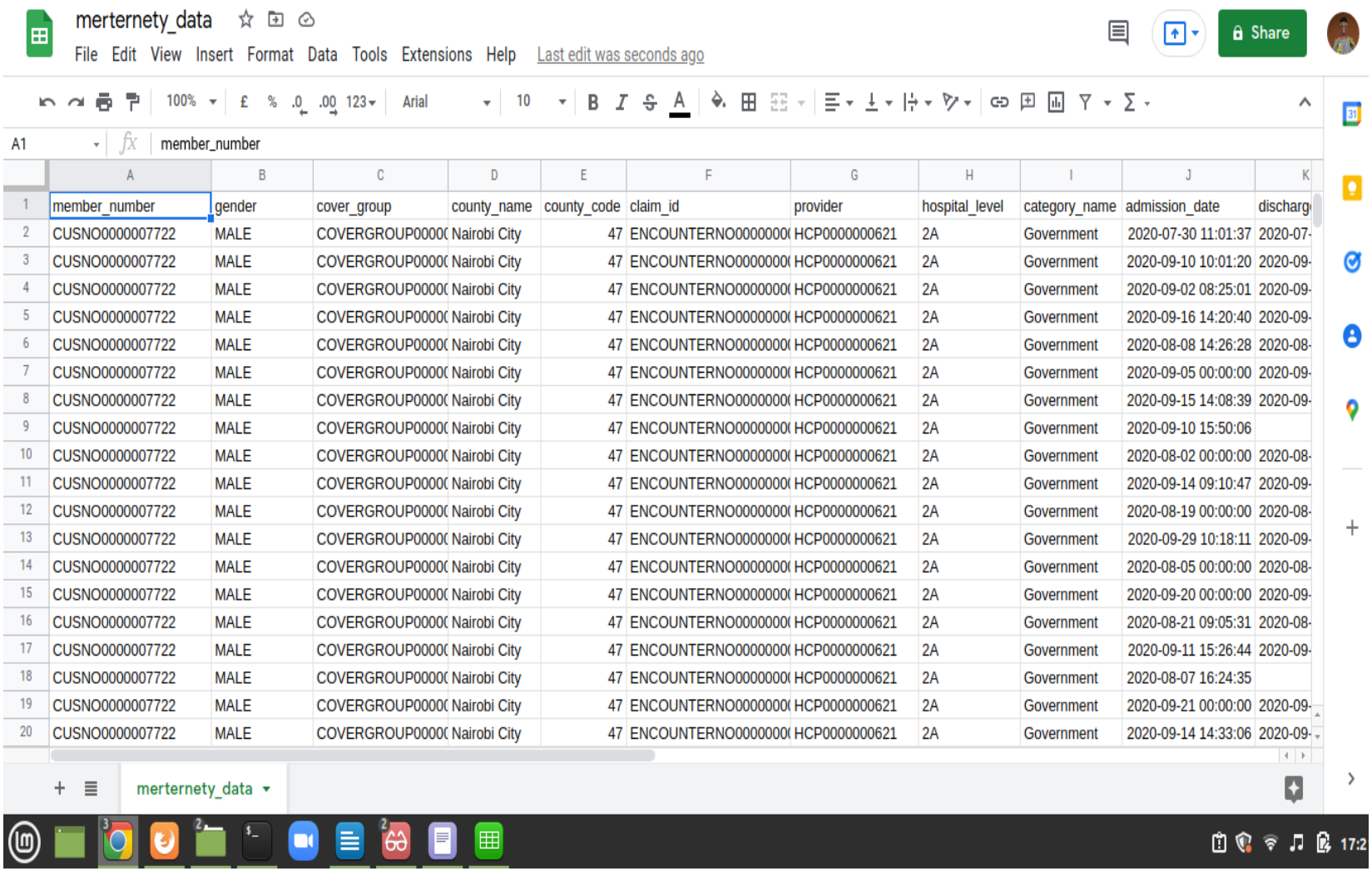
Date 27/09/2021

Witness:

Name Nixon Thuo

Signature

Date 27/09/2021



merternety_data

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

Share

A1 member_number

	A	B	C	D	E	F	G	H	I	J	K
1	member_number	gender	cover_group	county_name	county_code	claim_id	provider	hospital_level	category_name	admission_date	discharg
2	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-07-30 11:01:37	2020-07-
3	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-10 10:01:20	2020-09-
4	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-02 08:25:01	2020-09-
5	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-16 14:20:40	2020-09-
6	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-08 14:26:28	2020-08-
7	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-05 00:00:00	2020-09-
8	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-15 14:08:39	2020-09-
9	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-10 15:50:06	
10	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-02 00:00:00	2020-08-
11	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-14 09:10:47	2020-09-
12	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-19 00:00:00	2020-08-
13	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-29 10:18:11	2020-09-
14	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-05 00:00:00	2020-08-
15	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-20 00:00:00	2020-09-
16	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-21 09:05:31	2020-08-
17	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-11 15:26:44	2020-09-
18	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-08-07 16:24:35	
19	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-21 00:00:00	2020-09-
20	CUSNO0000007722	MALE	COVERGROUP0000(	Nairobi City	47	ENCOUNTERNO0000000(	HCP0000000621	2A	Government	2020-09-14 14:33:06	2020-09-

merternety_data