UNIVERSITY OF NAIROBI

DEPARTMENT OF COMPUTER SCIENCE & INFORMATICS

Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning

By

Ambrose Muchangi Njeru
P58/63269/2011

Supervisor: Dr. Evans A.K. Miriti

Research Project Report Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of Master of Science in Computer Science, Department of Computer Science & Informatics, University of Nairobi

November 2022

DECLARATION

Student

This research project report is entirely my original work, and I certify that it has not been submitted to any institution or the University of Nairobi in exchange for a degree or other academic award, to the best of my knowledge. Every piece of literature, resource, and reference utilized in the development of this study is properly cited and listed.

SIGNATURE  **DATE 10th November 2022**

Ambrose Muchangi Njeru

Registration Number: P58/63269/2011

Supervisor

This research project report has been submitted for assessment in partial fulfillment of the requirements for the award of the Degree of Master of Science in Computer Science of the University of Nairobi, with my approval as the University Supervisor.

SIGNATURE  **DATE 10th November 2022**

Dr. Evans A.K. Miriti

ABSTRACT

One of the biggest and most pervasive issues facing the insurance sector is the filing of false insurance claims by customers. Insurance firms incur significant financial losses due to pricey fraudulent claims. Concerns from stakeholders and observers have been raised about insurance fraud, which continues to be a major concern for insurers and customers who pay the expenses through insurance premiums. Understanding the institution processes and operationalization of ICT in fraud detection is the first step in implementing the appropriate corrective actions. However, the procedure is time and money consuming because personally reviewing all insurance claims filed with insurance companies has become challenging.

Given the prevalent issue of fraud in vehicle insurance claims, the manual approach to identifying fraudulent claims has been problematic because it is time-consuming and inaccurate. One of the various ways that researchers have tested is machine learning algorithms, which have demonstrated promising performance and enhanced accuracy in detecting fraudulent vehicle insurance claims. This study evaluated a range of ML algorithms, including AdaBoost, XGBoost NB, SVM, LR, DT, ANN, and RF, to discern between real and fraudulent automobile claims. Additionally, a machine learning-powered web-based system to predict and categorize vehicle insurance claims as either genuine or fraudulent was developed. The system was based on the machine learning classifier with the highest levels of prediction performance and classification accuracy.

The AdaBoost and XGBoost classifiers outperformed the other models with both imbalanced and balanced data because they had the highest classification accuracy of 84.5%. The LR classifier performed poorly since it had the lowest classification accuracy for both unbalanced and balanced data. The ANN classifier performed better with unbalanced data than it did with balanced data. The final finding was that all eight classifiers could only be used on smaller datasets.

ACKNOWLEDGEMENTS

Firstly, I wish to express my gratitude to the All-mighty God for endowing me with the strength and ability to pursue my studies and undertake this research.

Secondly, my sincere gratitude goes to my lovely wife and our children for all their support, encouragement, and tutelage throughout the entirety of my studies. My family, my lecturers, and my close friends all unconditionally gave me unwavering support and contributed significantly to the completion of my study endeavour.

My profound gratitude goes out to my direct supervisor, Dr. Evans Miriti, for his unwavering encouragement, steadfast support, considerate guidance, and valuable assistance with my research. I appreciate the chances provided to me for professional development and the expansion of my knowledge, experiences, and abilities. Finally, I sincerely thank all the individuals, both inside and outside of the University of Nairobi, who helped me complete my project by lending me their direct or indirect support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

ACL	Audit Command Language.
AdaBoost	Adaptive Boosting.
ADASYN	Adaptive Oversampling Technique.
ANN	Artificial Neural Network.
CHAID	Chi Square Automatic Interaction Detection.
CRISP-DM	CRoss Industry Standard Process for Data Mining.
CSV	Comma-Separated Values.
DT	Decision Tree.
GBM	Gradient Boosting Machines.
GLM	Generalized Linear Models.
ICT	Information Communication Technology.
LDA	Latent Dirichlet Allocation.
LMT	Logistic Model Tree.
LR	Logistic Regression.
MCC	Matthews's Correlation Coefficient.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
NB	Naïve Bayes.
NBU	Naïve Bayes Updatable.
RF	Random Forest.
RT	Random Tree.
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.
XGBoost	Extreme Gradient Boosting.

CHAPTER 1: INTRODUCTION

1.1 Background

This section defines the problem statement, establishes the study's main purpose and specific objectives, and presents the research questions. The chapter's conclusion emphasizes the importance of the study.

Information and communication technology (ICT) has continuously asserted itself as the architect of systems in recent decades by connecting markets, enterprises, governments, and individuals. This connection has reduced distances and given the globe a multidimensional aspect, improving task administration and servicing, and enabling real-time combat against ethics and fraud. The ICT revolution of the twenty-first century not only influences commercial developments and organics, but it also forecasts and defines social interaction, culture, and behaviour at many levels. Susceptibility to fraud prevention is one of the organizational, individual, and behavioural aspects that ICT has significantly altered in the business environment. Organizations like insurance companies have made significant investments in ICT to improve their capacity for information processing as part of this conflict and given them the advantage in identifying, handling, and reporting fraud-related situations.

False insurance claims filed by clients are one of the most frequent and chronic concerns confronting the insurance industry. Gill et al. (2005) defines insurance fraud as "knowingly making a fraudulent claim, inflating a claim, adding extra items to a claim, or being in any other way dishonest with the intent of collecting more than genuine entitlement." This rationale is based on dishonest, wilful, or fraudulent concealment, which leads in fraudulent claimant or policyholder illegal financial advantage. Insurance companies suffer huge financial losses because of costly fraudulent claims. As a result, it is critical to distinguish between genuine and fraudulent claims.

Insurance fraud continues to be a big concern for insurers and clientele who pay the expenses incurred through insurance premiums, according to a report by the Coalition Against Insurance Fraud (2016), raising concerns among stakeholders and observers. Understanding the procedures involved in implementing ICT and operationalizing it for fraud detection is the first step in adopting the appropriate corrective steps. Organizations like insurance companies have made significant investments in ICT to improve their capacity for information processing as part of this fight. With the use of this information, they now have the advantage in identifying, handling, and reporting fraud-related situations. Nevertheless, reviewing every insurance claim submitted to the insurance

companies, has become quite challenging, rendering the process expensive in terms of time and money invested.

Fraud in the insurance sector can be seen from four angles, according to International Association of Insurance Supervisors (2011) and Frimpong (2016). The first of them is internal fraud, which refers to an insurance employee defrauding the insurance firm either on their own or in collaboration with other parties either inside or internationally. The second type of fraud is false claim, in which the insured party provides false information to obtain payment or wrongful coverage. Thirdly, insurance intermediary fraud, in which insurance intermediaries conspire with one another or act alone to defraud the policyholder or insurer, and fourthly, insurer fraud, in which the insurer defrauds the policyholder through unfair policies, payment premiums, and compensation schemes.

Considering the widespread challenge of fraud in vehicle insurance claims, the manual approach for identifying fraudulent claims is problematic because it is slow and imprecise. Hence, Machine Learning (ML) approaches can be used to detect fraudulent vehicle insurance claims effectively due to their superior performance and improved predictive accuracy.

1.2 Problem Statement

Fraud has long been a significant concern and one of the most serious problems facing organizations due to the catastrophic effects. According to Gedela and Karthikeyan (2022), fraud is any act aimed at defrauding another party financially. Sybase (2012) emphasizes that steps should be taken to allow fraud detection as a first line of defence since it recognizes the financial cost and cultural consequences of the problem. The Kenyan insurance sector is well established, according to Association of Kenya Insurers (2020) and ranks first in Sub-Saharan Africa with a high growth rate, (African Insurance Organization, 2018). This has made a significant contribution to the market's readiness for adoption and attraction of foreign investment. However, holding such a prestigious position comes with a lot of challenges, chief among them being fraud and competition.

According to the Insurance Regulatory Authority (2021), the insurance industry is notoriously hesitant to evolve, especially when it comes to using new technologies to combat the alarming issue of fraud. They present numerous explanations for this, such as a lack of funding, the belief that things should be done the way they have always been done, and overstretched resources. Despite this, the insurance sector must act quickly to stay ahead of the growing fraud rates to safeguard both itself and policyholders. The Authority also points out that due to an increase in complaints and rising fraud,

costs associated with fraud investigations and tribunals are anticipated to reach tens of millions of dollars yearly.

Motor vehicle insurance fraud is a serious vice that has contributed to the collapse of several insurance companies and continues to present a substantial challenge to the insurance industry. According to the Association of Kenya Insurers (2020), automobile insurance is one of the most difficult products for Kenyan insurance companies to sell since they suffer significant technical losses, which amount to 68.92% for private vehicles and 60.72% for commercial vehicles. This means in other words, for every KShs 100 in premiums received by the insurer, KShs 68.92 and KShs 60.72 are used to settle insurance claims, respectively. The issue is exacerbated by the significant costs associated with the investigations done to confirm the claim's validity, which account for 44.16 percent of overall costs. This implies that the insurer loses KShs 13.08 and KShs 4.88 in net premium revenue, respectively. Most of these losses are attributable to fraudulent insurance claims.

Additionally, according to statistics from the Insurance Regulatory Authority (2021), 35 percent of insurance claims were fraudulent, with motor vehicle insurance claims leading the way and registering the greatest loss percentages in the sector. The fraudulent automobile insurance claims entail someone engaging in a variety of unethical behaviours to obtain a favourable conclusion from the insurance providers. These acts range from fabricating accidents, making false insurance claims, fabricating details for a real insurance claim, and misrepresenting an incident's cause and relevant players (Subudhi, et.al, 2018). As a result of the rise in fraudulent vehicle insurance claims, insurance companies are devoting more time and resources to the detection of these claims. The employment of conventional methods allows some to go unnoticed. As the economy recovers, an increase in fraud claims will raise overall insurance costs, making the issue of fraudulent insurance claims a key concern for both the government and insurance companies.

1.3 Main Objective

The primary objective of this project was to investigate how machine learning algorithms can leverage features extracted from vehicle insurance claim datasets to aid in the detection of fraudulent vehicle insurance claims. Following this investigation, a novel system to predict and categorize vehicle insurance claims as either genuine or fraudulent was developed.

1.4 Specific Objectives

1. Characterise fraudulent insurance claims in the context of vehicle insurance domain.

2. Identify features that could be utilized to train machine learning models to recognize fraudulent vehicle insurance claims.
3. Evaluate the performance of several machine learning models for detecting fraudulent vehicle insurance claims using a balanced and imbalanced dataset.
4. Develop a system that categorises vehicle insurance claims as either genuine or fraudulent using the best performing machine learning classifier.

1.5 Study Significance

This study is timely in that it offers a mechanism for developing a system by using the top-performing machine learning algorithm to identify fraudulent vehicle insurance claims. As the number of fraudulent insurance claims rises and their detection becomes a difficult problem on a global scale, fraud in the insurance industry is becoming an increasing concern. By guaranteeing quality and stability, this will assist insurance businesses in showcasing their exceptional claim administration, which will have a significant impact on their revenue and client's satisfaction. Additionally, the study will broaden the area of machine learning investigation into the identification of fraudulent vehicle insurance claims in the Kenyan insurance sector.

CHAPTER 2: LITERATURE REVIEW

The chapter begins by examining the general aspects of automobile insurance in Kenya before moving on to a discussion of the manual methods that have been used to identify fraudulent insurance claims. The discussion then moves on to the automation of systems for identifying fraudulent insurance claims before wrapping up with a discussion of methods for identifying such claims that make use of machine learning and deep learning.

2.1 Vehicle Insurance in Kenya

According to Mark and Liam (2021), "vehicle insurance" is "a contract under which the insurer assumes the risk of any loss experienced by the owner or operator of a vehicle as a result of property damage or individuals as a result of an accident." They went on to explain that there are many different sorts of motor vehicle insurance that are particular in character, ranging from the legal principles that underpin them to the kinds of risks they cover. The Insurance Regulatory Authority is the entity obliged by Kenyan legislation under Cap 486 of the Insurance Act passed in 1988 to regulate the country's insurance industry, according to Insurance Regulatory Authority (2021).

According to the Association of Kenya Insurers (2020), gross premiums for non-life insurance were KShs 132.70 billion in 2020. The motor vehicle and medical insurance sectors together accounted for approximately 67.14% of these gross premiums, with motor vehicle insurance holding the lion's share at 33.71%. Kenya's motor vehicle insurance market consists of both private and commercial coverage. According to statistics provided by the Insurance Regulatory Authority (2021), insurance fraud cost insurance companies KShs 258.4 million in 2020, up from KShs 19.2 million in 2019, a 13.4-fold increase. Additionally, the Insurance Fraud Investigation Unit (IFIU) received 127 reports of identified fraud insurance claim cases in 2020 as opposed to 83 cases in 2019, demonstrating a significant increase in fraudulent cases. They pointed out that the vehicle insurance area was the most targeted, with 39 vehicle insurance claims comprising 30.7% of the 127 insurance fraud instances reported in that year. Due to widespread fraud in the market, the vehicle insurance sector has seen enormous losses as a result. These report by Insurance Regulatory Authority (2021) demonstrated losses incurred from insuring both private and commercial vehicles nearly doubled, leading to the largest underwriting losses for vehicle insurers in more than 20 years. From KShs 1.8 billion in 2020 to KShs 6.34 billion in 2021, the underwriting loss increased by a factor of five,

2.2 Fraud Detection in Vehicle Insurance Sector

According to the Association of Kenya Insurers (2020), fraudulent claims account for 25% of all insurance industry claims, with motor vehicle insurance fraud instances being the most common. Fraud is a nebulous indicator of one's business, personal, and social ethics. Simha and Satyanarayan (2016) define fraud as an intentional deceit carried out through information concealment and misrepresentation with the objective to harm another party's interests by furthering one's own interests. According to the International Association of Insurance Supervisors (2011), fraud is defined as any action or inaction that is designed to benefit the perpetrator of the hoax or any other party dishonestly. Insurance fraud covers a wide range of behaviours, including wilful misrepresentation of facts, careless insurer management, abuse of a trust account or responsibility, and concealment or destruction of evidence that is relevant to financial transactions, communications, and insurance contracts (Ernst & Young, 2011).

There are numerous ways that fraud is perpetrated in the insurance industry. Vieane and Dedene (2015) concluded from their study that fraud in the vehicle insurance sector can take the form of impersonating legitimate claimants, forging insurance claim documents, misappropriating claimant money, or manipulating the system by an insurance company employee to compensate people who are undeserving. The definition by Derri (2002) regards fraud as an illegal act that entails obtaining financial benefit through the misrepresentation of an actual position for monetary gain. This definition considers the various forms that fraudulent insurance claims might take. The fraudulent acts of insurance claims are best described by this description since the fundamental motivation behind fraud is to profit financially from the transaction by recovering the lost asset and making up for the loss.

The Coalition Against Insurance Fraud (2016) defines fraud detection as the use of systems and observations to identify insurance-based false information, transactions, and intentions. The alleged act may or may not be concealed. The insurer is typically the perpetrator of fraud acts that are not concealed by unfair claims management and policy practices. Due to the appearance of normalcy around them, such scams are challenging to spot. Claims fraud, which typically involves unfair and deceptive representations of facts, is the most prevalent type of fraud in concealing (Mathenge, 2016).

2.3 Manual Fraud Detection Approaches

According to the Association of Kenya Insurers (2020), ideally, insurance companies employ insurance agents who can analyse and evaluate each claim and determine whether it is genuine or

fraudulent. However, because of how time- and money-consuming this process is, finding and paying for the insurance agents would involve examining all the numerous claims that are filed every day, which is just not viable. Insurance agents occasionally make use of information pertaining to submitted automobile insurance claims before attempting to organize the claims and waiting for the investigation result. The agent determines whether the claim is legitimate or fraudulent based on the data gathered and the investigation report. Insurance Regulatory Authority (2021) mandates that insurance firms and car insurance holders gather specified data at the site of accidents to aid in the processing of vehicle insurance claims. These details include the drivers' names, residences, insurance coverage and certificates from other drivers, vehicle registration numbers, and the year, make, and model of the car that was involved in the collision.

A supervisor in the insurance firm reviews the claims logged based on the facts supplied and uncovered to manually identify fraudulent auto insurance claims. Scores are determined using indicators against a checklist depending on the specifics of the damaged vehicle components. If the scores are high, an investigator is given a case to investigate the damaged vehicle once the scores are generated. The investigator then delivers an investigation report to the person in charge. If the investigation report is positive, the claim is thought to be true; if it is negative, the claim is thought to be false (Association of Kenya Insurers, 2020).

These approaches provide several difficulties due to their heavy reliance on physical interventions, which could result in:

i. Approaches that rely on human knowledge, which is composed of a limited set of well-known parameters, while being aware that other decisions may be affected by other factors.
ii. The approaches' inability to comprehend context-specific correlations between parameters that might not accurately reflect the overall situation.
iii. The manual model needs to be calibrated on a regular basis to account for changing behaviour and to ensure that it conforms to the findings of the investigations. It takes skill to accomplish this calibration manually, which is difficult.

2.4 Automation of Fraud Detection Systems

Insurance businesses have purchased computerized systems thanks to the adoption of ICT, which have helped them increase the efficiency of their daily business operations. The technology used, together with the availability of internal and external data, are all important factors in fraud detection.

According to the Association of Kenya Insurers (2020), one method of detecting claim fraud is to use rule-based expert systems to match every claim by analysing a list of predetermined business indicators or rules using "if-then" statements and sending out alerts when certain conditions are met to mark the claims suspicious. The association further adds that the automation of fraud schemes detection inside the insurance framework has been examining situations such "if within a short period of time, there has been multiple claims made from a different area, then submit the account for manual assessment." Systems have been developed considering these principles and improved based on decades of manual experience in reviewing and analysing fraud data. Additionally, many of these criteria are also designed to allow for further examination of the atypical claim transaction behaviour.

Rule-based systems, according to Baumann (2021), are those that use human-made rules to store, sort, and manipulate data to help identify fraudulent insurance claims. Even though the systems are complex, they can be built using algorithmically observable signals, demanding labelled data or reliable expert assessments. The use of algorithms in rule-based systems involves running several scenarios for fraud detection that are manually designed by fraud analysts. They also mentioned that, on average, the systems apply around 300 distinct rules to approve a claim. These systems are overly simplistic, necessitating manual addition or modification of cases, and they are incapable of detecting implicit relationships. Furthermore, rule-based systems usually use outdated technology that is incapable of handling the real-time data streams required in the digital environment.

According to a report by the Coalition Against Insurance Fraud (2016), 81% of insurance companies utilize automated methods to spot fraudulent claims. Red flags would be raised by these systems when a claim seemed strange or caused concern. When claims are made, there are several elements that might raise red flags. For instance, a claim could be filed soon after a change or increase in coverage, or the insured might ask their agent hypothetical questions concerning coverage in the case of a loss that is extremely like the one for which the claim is being filed. Moon et al. (2019) investigated the use of rule-based methods to evaluate various fraud detection scenarios. They discovered that these systems make use of manually created algorithms by fraud analysts. Since circumstances that fail to recognize the underlying linkages must be manually adjusted, rule-based systems are simple to implement.

Owusu-Oware et al., (2018) investigated the application of biometric technologies in Ghana to prevent national health insurance fraud. They used an integrated system of social and technical systems to combat health insurance fraud, including online enrolment and verification of members at the point of receiving health services using biometrics, the use of e-claims as complementary

technologies, and utilizing physicians to evaluate service providers' claims as operational strategies. Their research revealed that the use of biometric membership enrolment online removed numerous identities while the use of biometric service delivery point verification reduced potential for fraudulent insurance claims. Like audit trails in information systems security, where user modifications to a database were recorded and used as proof for systems or forensic audit, the biometric verification kit generated a special health facility attendance code. The study demonstrated how claims for "ghost patients" might be filtered out by comparing provider claims to evidence attendance codes. The study also showed that the adoption of biometric technologies contributed to a decrease in patient billing fraud.

According to the above study, adopting biometric technology had no appreciable impact on reducing fraud incidents in the insurance industry. Even when a possible fraudulent claim has been identified, sometimes the technologies available only allow for rudimentary analysis with poor accuracy, which prompts an insurance agent to launch a more thorough inquiry. As a result of the difficulties with human methods, insurers have begun to use machine learning methods. Additionally, because of information technology advancements, insurance fraud detection has been automated further with a combination of data mining, potent analytical algorithms, and expert expertise on the mined data to produce insightful information (Dull, 2014). Audit Command Language (ACL) data analytics and IDEA management are a few examples of the other technologies that insurance firms can employ to spot fraud (Moore, 2016).

2.5 Insurance Fraud Detection using Machine Learning

Researchers are primarily looking for novel, efficient, and effective strategies that can be used to predict and analyse claim content using deep learning and ML algorithms to identify between genuine and fraudulent insurance claims. Because of technological advancement on many scales, a range of new fraudulent behaviours have emerged.

Burri et al. (2019) examined the performance of the NBU, NB, Multi-Layer Perceptron, J48, RT, RF and LMT machine learning algorithms to predict fraudulent claims. Despite various barriers to the adoption of machine learning to categorize claims and difficulties in its implementation, the researchers' research demonstrated how crucial it was for insurance companies to embrace machine learning technologies to detect fraudulent insurance claims. On the 1240 samples of the dataset, they applied 8 attributes or features. Additionally, they found that the LMT and Random Forest algorithms outperformed the other ones utilized during their investigation. All three metrics—precision, recall,

and F1 Score—were 100%. Precision, recall, and F1 Score were all at 81.6%, with Naïve Bayes Updatable and Naïve Bayes achieving the poorest results. In their study, they also noted the following challenges: Firstly, there was imbalance and a lack of representativeness in the training datasets, which led to bias. Secondly, insurers had trouble providing pertinent data for the development of machine learning algorithms. Thirdly, because funding requirements can change, it can be challenging to estimate advantages that machine learning could bring to a project. Lastly, the enormous data sets that the machine learning algorithms tapped into raised the security risk for insurers, leading to data leaks and security breaches that made insurers nervous. Unfortunately, in their study they did not address the noted challenges and did not reveal the features they used.

Sunita Mall et al. (2018), created a system using 46,175 commercial vehicle claim data from an Indian motor insurance company to anticipate the behaviour of fraudulent customers. The model required finding important fraud triggers and predicting the likelihood of each claim being approved as well as rejected using the statistical techniques like LR and CHAID, which uses a decision tree approach. Utilizing the triggers found in an existing algorithm, the strategies were applied to forecast the fraudulent clients' behaviour. To extract features, they considered elements like the insurer's information, the type of vehicle, the vehicle's maximum tonnage, the insurance branch's code, the insured amount, the paid loss, the claim's specifics, the vehicle's age, etc. They discovered that, to classify fraudulent claims, Seats/Tonnage, No Claim Bonus, Type of Vehicle, Gross Written Premium, Sum Insured, Discounts, State Similarity, and Previous Insurance Details formed a 1% importance level, whereas Branch Code and Risk Types formed a 5% important level. Variables like the Channel code and Reporting delay were more important in generating fraudulent claims features, considerably assisting in the prediction of false claims, with a significance level of 10%. They concluded that while variables like Branch code, Type of vehicle, Seats/Tonnage, and Previous Insurance Details would be used to classify the likelihood of receiving a claim that is genuine, and variables like Branch code, Third Party Flag, State Similarity, Gross Written Premium, and Reporting Delay would be used to classify the likelihood of receiving a claim that is fraudulent.

Dhieb et al. (2019) employed XGBoost technique to create a system that could recognize and categorize various types of motor vehicle insurance claims as genuine or fraudulent. They extracted relevant machine learning classification features using data cleaning, exploration, and extraction techniques, and then they evaluated the algorithm's performance using a range of metrics. Their classifier was trained, validated, and tested using a dataset of more than 64,000 claims divided into eight classes based on three categories of motor vehicle fraud claims: "Invalid kind of loss," "No premium but has claim," and "Fraudulent claim amount." They started by creating a fictitious dataset

to evaluate the effectiveness of the suggested fraud detector and classifier. Finally, they used accuracy, precision, recall, and F1-score measures to compare the performance of their proposed technique to those of other machine learning approaches such as k-nearest Neighbor, NB, and DT. XGBoost outscored the other machine learning methods considerably, with an accuracy of 99.25% as opposed to 86.99% for Decision Tree, 52.06% for Naïve Bayes, and 42.70% for K-Nearest Neighbor. Training and evaluation, however, took more time. Unfortunately, they did not reveal the features they used in their research.

Machine learning-based fraud detection regarding credit cards was the subject of research by Awoyemi et al. (2017). In their discussion, they discovered that the constant change in the profiles of fraudulent and normal behaviour and the frequent unbalanced nature of the data sets used made fraud detection difficult. As a result, the dataset sampling, variable selection, and fraud detection techniques' effectiveness were all significantly impacted. The study investigated and evaluated the effectiveness of fraud detection strategies such as NB, k-nearest Neighbor, and LR. They assessed the performance of the three algorithms in terms of sensitivity, accuracy, specificity, and precision, as well as MCC and balanced cataloguing rate. They also utilized a hybrid sampling strategy to resolve the dataset's imbalance. Naïve Bayes performed the best, attaining an accuracy level of 97.92%, while Logistic Regression and k-nearest Neighbor earned accuracy levels of 54.86% and 97.69%, respectively. However, the researchers did not explore the idea of applying a feature selection approach.

Gedela, B., and Karthikeyan, P. R. (2022) created a system that used the AdaBoost classifier to recognize fraudulent credit card transactions for each bank that issues credit cards. Researchers evaluated the proposed technique against NB, LR, ANN, and DT algorithms to determine its effectiveness. A training dataset with 2,27,845 transactions making up 80% of the dataset and a testing dataset with 56,962 transactions making up the remaining 20% were both used instead of the entire 2,84,807 transactions. The dataset contained 2,84,807 transactions, of which 492 were fraud transactions. Accuracy, sensitivity, specificity, precision, and f1-score metrics were generated to assess the performance of the algorithms. AdaBoost, NB, LR, ANN, and DT detection accuracy rates were 99.43%, 90.93%, 95.35%, 94.81%, and 94.81%, respectively. The f1-score of the AdaBoost algorithm was 99.48%. The qualitative analysis discovered that their proposed AdaBoost algorithm outperformed the NB, LR, ANN, and DT algorithms in detecting credit card fraud.

In their study, Bauder et al. (2017) focused on detecting Medicare fraud by comparing four performance evaluation systems, including class imbalance reduction using the 80-20 under-sampling

and over-sampling methodologies, and hybrid, unsupervised, and supervised machine learning algorithms. They built their features based on the behaviour that known fraudulent and non-fraudulent providers exhibit, as well as the risk that is spread through geographic collocation, or shared addresses. Then, using the feature matching technique in the data processing and clean up exercise, they eliminated features that were not present over all five years of their study. Several factors were employed in the study, including the provider's unique provider identification number, the kind of medical provider, the provider's gender, and the fraud classifications from the dataset. They concluded that all machine learning techniques performed badly after employing the oversampling strategy and the supervised approaches outperformed unsupervised or hybrid approaches. The researchers also concluded that using Balanced Accuracy (BA) to evaluate the model's performance across all of the ML algorithms used was unreliable because it was unable to appropriately reflect the more accurate fluctuations observed in the other metrics.

Subudhi, et al. (2018) developed a cutting-edge approach for identifying bogus claims in the field of vehicle insurance. They made use of a dataset that had one class attribute of legitimate and fraudulent labels in addition to 32 insurance-related variables. The dataset included 15,420 insurance claims that were made in the United States during 1994 and 1996. The dataset contained 14,497 legitimate claims and 923 fraudulent claims, resulting in an imbalance of 94% to 6%. The researchers used ADASYN to address the data imbalance on minority class occurrences (fraudulent claims). By raising the number of occurrences in the false claims class to 12,230, a virtually balanced data set was produced. The study used 10-fold cross validation to test the effectiveness of the classifiers after separating anomalous data from regular records using machine learning techniques such as SVM, DT, and MLP. The three algorithms' accuracy, specificity, and sensitivity were evaluated using both balanced and imbalanced datasets (after applying ADASYN). Multi-Layer Perceptron and Decision Tree classifiers both performed poorly on the skewed dataset; however, Support Vector Machine performance was not much affected. On the balanced dataset, decision trees and support vector machines both generated successful classification outcomes.

To identifying fraudulent auto insurance claims, Wang et al. (2018) developed a deep learning method that made use of text analytics and the LDA technique. The LDA technique was used in their analysis to extract text features from text images of an accident while concealing accident text descriptions that were present in the claims. The data that was acquired, which included text and conventional numeric features, was then used to train the deep neural networks. The results showed that neural networks outperformed SVM and RF when it came to identifying fraudulent vehicle insurance claims.

Finally, Randhawa et al. (2018) used machine learning approaches to develop a system for detecting credit card fraud. In addition to a publicly available credit card dataset, a real-world credit card dataset from a financial institution was examined. They began by investigating prominent credit card fraud detection algorithms such as RF, SVM, and LR. The algorithms' performance was evaluated using data under-sampling, and RF outperformed SVM and LR. They then used hybrid methods such as AdaBoost and majority voting algorithms. The MCC metric was chosen as a performance indicator because it considers both genuine and false positive and negative outcomes. With an MCC score of 0.823, the majority voting algorithm performed the best. Noise ranging from 10% to 30% was injected into the data samples used to evaluate the hybrid models to test their robustness. The majority voting technique produced the best MCC score of 0.942 for a 30% increase in noise. This has demonstrated that the majority voting method performs effectively in noisy environments and has high accuracy rates for detecting credit card fraud.

2.6 Machine Learning Classifiers for Vehicle Insurance Fraud Detection

A machine learning classifier, according to Tang et al. (2016), is an algorithm that categorizes data automatically into one or more sets of categories. The provided data is used to train the classifiers to the necessary level of accuracy. They went on to explain that implementing machine learning involves approximately transferring a mapping function (f) from discrete input variables (X) to output variables (Y). Burri et al, (2019), described machine learning classification as a process where machine learning algorithms are needed to learn how to categorize input samples from a problem area. This study examined the following machine learning classification algorithms: XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR to ascertain how effectively and efficiently fraudulent vehicle insurance claims can be identified.

2.6.1 Naïve Bayes (NB) Classifier

Under the straightforward premise that a probabilistic machine learning model's attributes are conditionally independent, the NB method is implemented on the Bayes' theorem (Tang et al., 2016). According to Lewis (1998) and Tang et al. (2016), NB is a straightforward algorithm to develop and performs best when there are classification features that are unrelated to one another. The Naïve Bayes algorithm was used by Bhavna and Sheetal (2019) in their work to identify vehicle insurance fraud. According to them, the method was simple to use, quick to detect fraud, required less training data, and was highly scalable. According to their investigation, the algorithm provided excellent accuracy levels of 89.6% with a short execution time.

2.6.2 Decision Tree (DT) Classifier

According to Asiri (2018), the Decision Tree approach generates classification or regression models that are visualized as flowcharts. The methodology employs an if-else structure to train the rules from the dataset before creating a tree structure. The DT algorithm's tree represents each rule as a node, each leaf corresponds to a class, and the rules are learned one at a time using the training data. The DT algorithms are used to regulate the construction of decision trees by selecting the best discriminatory splitting rule at each non-terminal node and limiting the number of terminal nodes in the DT (Gepp, et al., 2012). By creating too many branches in the decision tree classifier, over-fitting can be easily accomplished, and abnormalities may also be reflected, leading to very poor performance on the unseen data. Gepp et al. (2012) used a real-world car insurance fraud dataset from the United States to give a comparative examination of fraud prediction. They compared decision trees, survival analyses, and artificial neural networks (ANNs) as part of their study. Because it generates the "if-then" rules, they concluded that the DT classifier was the best computational data mining tool for identifying fraud in the insurance industry because the classifier was 75.7% accurate in predicting if a claim was false.

2.6.3 Logistic Regression (LR) Classifier

This is a supervised machine learning method that builds a prediction model for a binary response variable using specified judgments (Jason, 2016). The maximum-likelihood estimation approach and the cause-and-effect relationship are used by the classifier to estimate the training data. The pre-processing stages are applied to manage filthy data to guarantee a high degree of detection accuracy.

In their study, Moon et al. (2019) used the logistic regression algorithm to carry out a variable importance ranking procedure using the 10-fold cross-validation technique across 20 iterations. The learning set was divided into 10 segments, which were then split into nine segments for the training set and one segment for the logistic regression model's validation set at random. The 10-fold CV was repeated 20 times, and the variable selection approach for each model was stepwise selection, yielding 200 logistic regression models. The model's accuracy was 87.1%, with a sensitivity/specificity imbalance of 62.4%/93.1%. Additionally, Wilson (2009) used a logistic regression classifier in his research to find fraud in the automotive insurance industry. With a focus on both legitimate and fraudulent claims data, he examined in his model the various circumstances and strategies employed by insured individuals to cheat insurance firms. At 59.2% accuracy, the model predicted whether new

claims would be marked as false. Additionally, the model anticipated that legitimate claims would be 81.6%, with an overall projection of 70.4%.

2.6.4 Random Forest (RF) Classifier

RF classifier is a classification and regression supervised machine learning technique. It uses the divide-and-conquer approach to create decision trees on randomly selected data samples, and when each tree produces a forecast, the best alternative is picked via voting (Breiman, 2007). The classifier passes through four stages to perform optimally: It randomly takes samples from a specified dataset, builds a decision tree from that sample, and outputs a prediction result for each decision tree. Finally, as shown in figure 1, each anticipated outcome is given a vote, and the final prediction is determined based on the forecast result that receives the most votes.

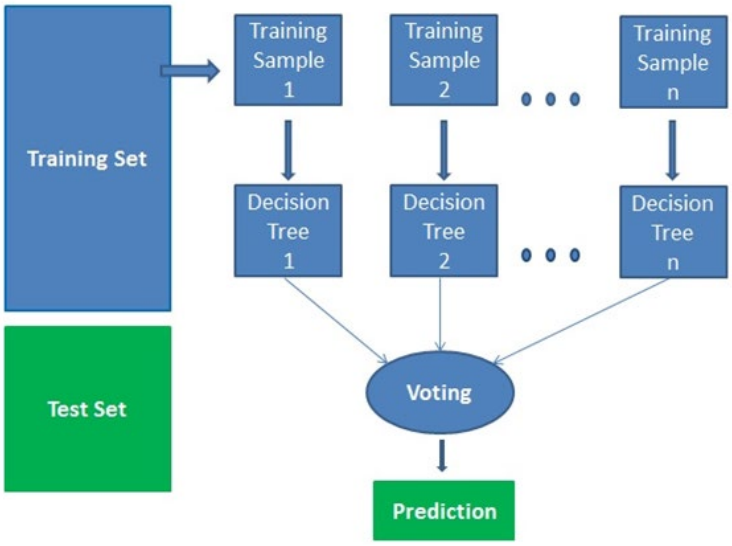


Figure 1: Random Forest Classifier Stages.

DeBarr and Wechsler (2013) used a dataset from the auto insurance industry to demonstrate how a RF classifier can be used to identify fraud. In their study, they employed ensemble learning to account for different data distributions and original and reputation variables to describe insurance claims. For both the original characteristics and the reputation features in each Random Forest model, they built a total of 2,000 decision trees using the balanced stratified random sampling method. They examined the efficacy of the model based on the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), precision, recall, and F1 Score measures, with cost savings serving as the primary metric. The RF classifier attained ROC/AUC scores of 82.0% based on the reputation features approach, as opposed to 73.8% for the original features approach.

2.6.5 Support Vector Machine (SVM) Classifier

This is a supervised machine learning approach that can be employed for classification as well as regression. To categorize high-dimensional data, the classifier employs hyperplanes. A hyperplane is a subspace that has one dimension less than the feature space. When a data point is plotted in an n-dimensional space (where n is the number of features), it is assumed that the input data can be linearly separated in a geometric space, with the value of each feature being the value of a specific coordinate (Joachims, 2001). The adaptive synthetic sampling method was used by Charles et al. (2020) to reduce imbalances in the dataset and identify fraudulent motor insurance claims. They subsequently classified the claim instances using the SVM classifier, and they contrasted the unbalanced datasets with alternative approaches. The model provided up to 93% accuracy.

2.6.6 Adaptive Boosting (AdaBoost) Classifier

AdaBoost, which stands for Adaptive Boosting, is a is an ensemble learning classifier that is used in conjunction with other classifiers to improve their performance for binary classification. The training error is reduced by giving a weak classifier with a high accuracy additional weight by passing a coefficient. The MCC measure is used by the classifier to assess the problem's quality. A value of +1 denotes a flawless prediction, whereas a result of -1 denotes complete disagreement. The outputs are combined using a weighted sum, which represents the combined output of the boosted classifier (Randhawa et al., 2019). The three most important variables in this classifier are the number of estimators, the base estimator, and the learning rate. The number of estimators indicates how many weak learners must be trained. The base estimator, a weak learner, is used to train the model. AdaBoost modifies the weak learners' weights based on their learning rate in favour of incorrectly categorized data samples. Despite its vulnerability to noise and outliers, AdaBoost can improve the specific outcomes of various approaches if the classifier performance is not random.

Mishra (2021) employed the AdaBoost classifier to detect fraud utilizing data from a large data mining and fraud detection research collaboration between Worldline and the Machine Learning group of the Universite Libre de Bruxelles. The dataset contained the timing and number of transactions done by European cardholders. In his investigation, he used random forest as the basis estimator. He emphasized that there were various numbers of estimators for each observation. Utilizing 200 estimators and a learning rate of 0.01, the classifier had an accuracy rate of 92.48%. While the learning rate does not greatly affect the AdaBoost Classifier's accuracy, he concluded that it does improve the model's stability.

2.6.7 Extreme Gradient Boosting (XGBoost) Classifier

XGBoost is an ensemble learning technique that use Gradient Boosted decision trees to improve model speed and performance. It produces a forecast by combining the results of previous learners. To create decision trees sequentially, all independent variables are assigned weights before being fed into the decision tree to forecast results. The variables are then fed into the second decision tree, which has a higher weight for variables predicted wrongly by the first tree. These several classifiers/predictors are then integrated to form a strong and accurate model. It can be used to handle problems including regression, classification, ranking, and custom prediction (Dimitrakopoulos et al., 2018).

To obtain correct fraud insurance claims quickly, Shah et al. (2021) created an automated fraud detection application framework based on the XGBoost classifier. To clean, validate, and extract the pertinent features from the dataset, data analysis techniques such as, data insertion, clustering, data pre-processing, and data validation are utilized. They also used other machine learning classifiers and determined that the XGBoost classifier outperformed LR, SVM, and RF with the best precision accuracy (94%), based on precision, recall, F1-score, and ROC Area metrics.

2.6.8 Artificial Neural Networks (ANN) Algorithm

ANNs are a deep learning concept based on human brain with the interconnection of neurons being the same as the interconnection of nodes in an artificial neural network. With the use of many computational layers, this approach uses cognitive computing to create machines that can self-learn to perform data mining, pattern identification of authorized behaviour, and natural language processing (Graupe, 2016). ANNs may distinguish between fraudulent and genuine transactions more precisely than other machine learning algorithms during the data training process because they use cognitive computing and learn from patterns of authorized behaviour. As a result, ANNs digest data faster and operate in real time.

A neural network-based method incorporating information from millions of claims records with more than 20 variables was proposed by Jalali (2020). Among the variables used were factors relating to the individual, such as age, age at claim, and gender. Claims characteristics included claims history (many claims), minor or major claims, minor or major claims via one or more products, and claim amount. Claims characteristics also included information about the hospital, such as the reason for admission, the length of stay, and institution's state. They used three distinct algorithms – GLM,

GBM, and ANNs—to evaluate their model to the best-performing model. The test results showed that the GBM model did not perform satisfactorily since the GLM model could only identify 26% of the false claims and was unable to correctly identify any fraudulent claims. The ANN beat the other two algorithms, identifying about 53% of the fraudulent claims.

2.7 Research Gap

A comprehensive amount of work has been done to develop fraud detection systems within the insurance industry, and several researchers have sought to advance numerous studies that address the problem of predicting fraud, according to the literature review. Due to the sensitivity and privacy of the data some Kenyan insurance companies possess, they only have access to the dataset for themselves. As a result, it is challenging to analyse, develop, and test machine learning algorithms in the field of vehicle insurance, which focuses on the fraudulent claims made to them. To address the problem of fraudulent claims using machine learning algorithms and build a dataset from the insurance firms, research is therefore required that largely focuses on the Kenyan vehicle insurance industry. From the research previously mentioned, it was seen that Burri et al. (2019), Dhieb et al. (2019), and Ayowemi et al. (2017) did not disclose the features they employed in their investigations.

2.8 Proposed System Description

The goal of this project is to detect fraudulent vehicle insurance claims as soon as they are submitted and before they are processed by the insurance company. Eight machine learning classifiers, including XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR, were trained using features taken from the acquired dataset to achieve the suggested solution. These classifiers were selected based on relevant research that showed encouraging outcomes when tested for accuracy. The classification outcomes from the eight classifiers were analysed and assessed to choose the best-performing and most accurate algorithm to use in the development of a system that can recognize and classify vehicle insurance claims as either legitimate or fraudulent.

Based on the characteristics of a vehicle insurance policy, a model for the identification of fraudulent vehicle insurance claims was created, and its performance and prediction accuracy assessed using training and testing datasets. Our input variables were divided into insurance policy details and insurance claim details to match the target fraudulent status of the vehicle insurance claims. Some of the details captured to meet the two variables were the name of customer, sex, age of policy holder vehicle category, vehicle make, age of the vehicle, sum insured, the insurance cover details, policy

start and end date, date the claim was logged, the date of when the incident occurred, place the incident occurred, police report etc. Figure 2 shows the proposed model diagram for unbalanced and balanced datasets and figure 3 shows the proposed machine learning-powered web-based system for detecting fraudulent vehicle claims.

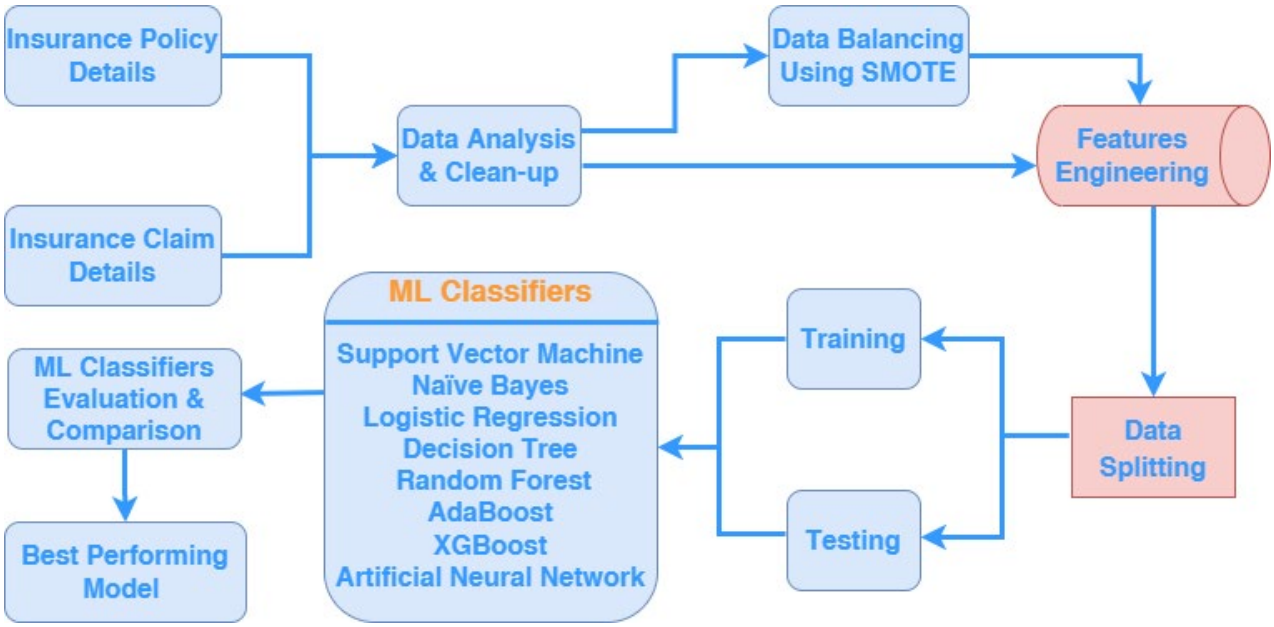


Figure 2: Proposed Model Diagram with Unbalanced and Balanced Datasets.

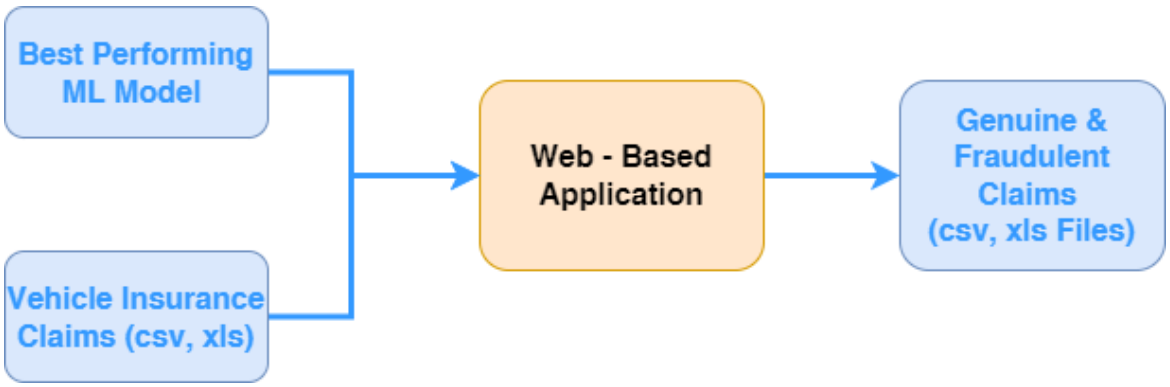


Figure 3: Proposed Machine Learning-Powered Web-Based System.

CHAPTER 3: RESEARCH METHODOLOGY

The research on using machine learning classifiers to identify fraudulent vehicle insurance claims is the primary focus of this chapter. The CRISP-DM methodology, data collection and analysis, model creation, and model evaluation metrics are all covered.

3.1 Introduction

The fundamental goal of the research is to achieve the objectives outlined in the introduction section. Before developing the machine learning model, the claims content was analysed to identify relevant features. Vehicle insurance data was collected as part of the research process to better comprehend the data structure and extract the necessary features to train the machine learning classifiers. To satisfy the study's objectives, the best-performing and most accurate machine learning classifier that can predict and categorize vehicle insurance claims as genuine or fraudulent was discovered utilizing the CRISP-DM methodology.

3.2 CRISP-DM Methodology

The CRISP-DM methodology was employed for this study due of its widespread use in data analysis and mining, flexibility, and extensive backtracking. CRISP-DM, is a 1966 invention that organizes, plans, and executes data mining (machine learning) operations (Rodrigues, 2020). It is a process model that outlines the normal project phases, tasks connected to each phase, and relationships between these tasks while also providing an overview of the data mining life cycle. The technique is used to conceive a data mining project and consists of six successive steps. Depending on the requirements of the developers, iterations might be introduced. The phases are as follows and as depicted by figure 4 below:

1. Business Understanding – What does the business need?
2. Data Understanding – What data do we have / need? Is it clean?
3. Data Preparation – How do we organize the data for modelling?
4. Modelling – What modelling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How results accessed?

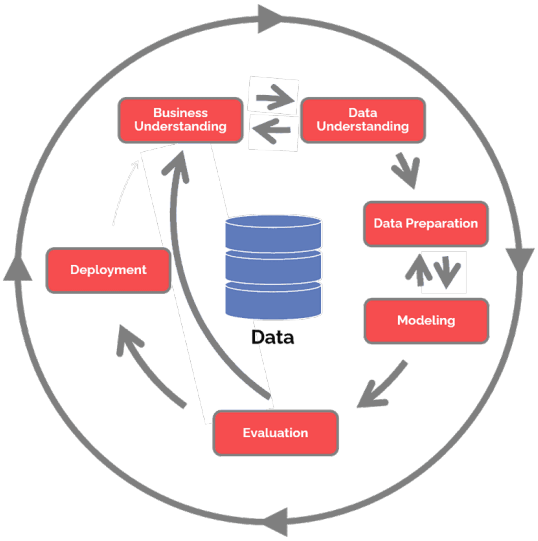


Figure 4: CRISP-DM Methodology Diagram

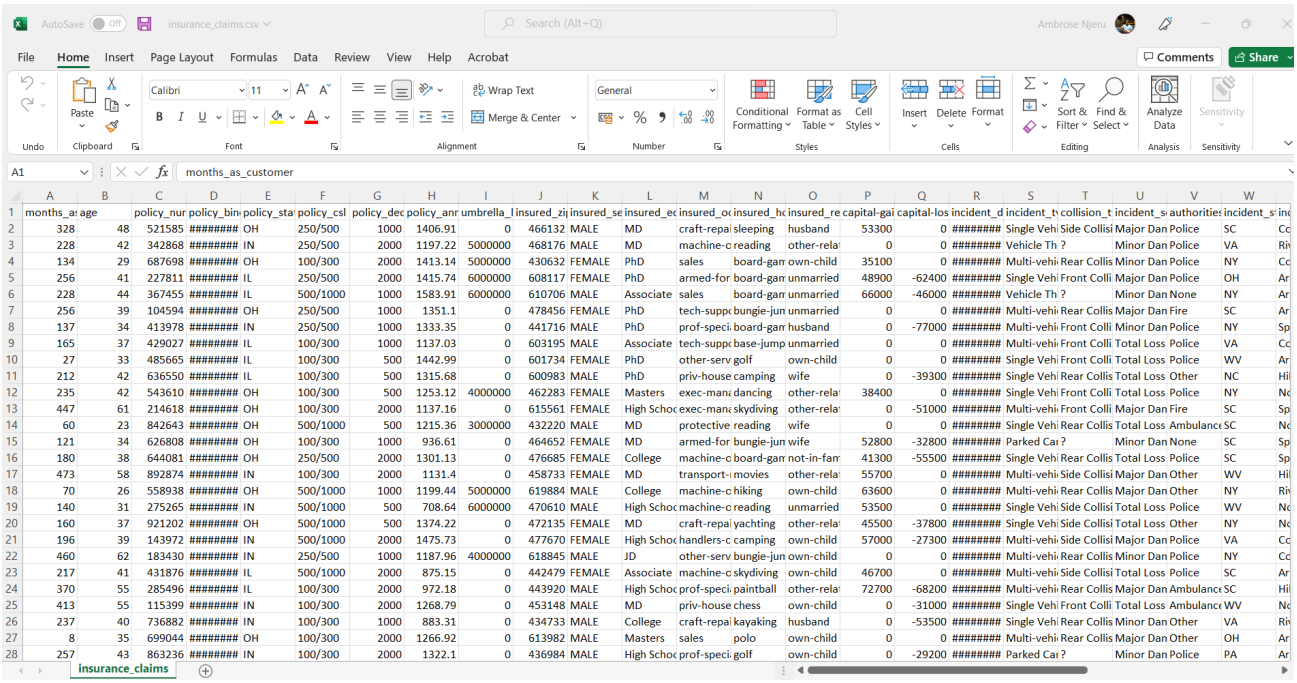
3.2.1 Business Understanding

Both primary and secondary sources were employed in this phase to comprehend the issue of fraudulent vehicle insurance claims. As secondary sources, we used books, journals, regional and worldwide online papers on machine learning with a focus on detecting insurance claims fraud. The researcher concentrated on the study's goals and objectives, which are related to fraudulent activity committed in connection with vehicle insurance claims. This made it easier to choose the best methods for gathering data and conducting the research. We obtained a motor vehicle insurance proposal form and a motor vehicle insurance claim form from Britam Insurance Company to better comprehend the important information gathered and acquired by insurance firms, as indicated in appendices 3 and 4, respectively. In the area of vehicle insurance, there has been an upsurge in fraudulent insurance claims, which has resulted in large financial losses. Therefore, a system that can identify fraudulent insurance claims in real time is needed for the vehicle insurance sector.

3.2.2 Data Understanding

In this step, we started by obtaining relevant data for the study, then familiarized ourselves with the data, assessed its quality, gained a fundamental understanding of the data, and extracted variables from the dataset to aid in the model construction. By contacting Britam Insurance Kenya, Co-operative Insurance Company (CIC), Jubilee Car Insurance, APA Car Insurance, and Heritage Insurance Company, we attempted to gather a dataset focusing on Kenyan insurance providers for this study. Unfortunately, were unable to secure a dataset owing to the sensitive nature and privacy

of the information it included. However, as can be seen in the excerpt in figure 5 below, we obtained an online CSV file on vehicle insurance claims dataset from Kaggle (2018).



months_as	age	policy_nur	policy_bin	policy_sta	policy_csl	policy_dec	policy_ann	umbrella_l	insured_zi	insured_se	insured_ed	insured_oc	insured_hc	insured_re	capital-gai	capital-los	incident_d	incident_t	collision_t	incident_s	authorities	incident_s
328	48	521585	########	OH	250/500	1000	1406.91	0	466132	MALE	MD	craft-repai	sleeping	husband	53300	0	########	Single Vehi	Side Collisi	Major Dan	Police	SC
228	42	342868	########	IN	250/500	2000	1197.22	5000000	468176	MALE	MD	machine-c	reading	other-rela	0	0	########	Vehicle Th	?	Minor Dan	Police	VA
134	29	687698	########	OH	100/300	2000	1413.14	5000000	430632	FEMALE	PhD	sales	board-gan	own-child	35100	0	########	Multi-vehi	Rear Collis	Minor Dan	Police	NY
256	41	227811	########	IL	250/500	2000	1415.74	6000000	608117	FEMALE	PhD	armed-for	board-gan	unmarried	48900	-62400	########	Single Vehi	Front Colli	Major Dan	Police	OH
228	44	367455	########	IL	500/1000	1000	1583.91	6000000	610706	MALE	Associate	sales	board-gan	unmarried	66000	-46000	########	Vehicle Th	?	Minor Dan	None	NY
256	39	104594	########	OH	250/500	1000	1351.1	0	478456	FEMALE	PhD	tech-suppo	bungie-jun	unmarried	0	0	########	Multi-vehi	Rear Collis	Major Dan	Fire	SC
137	34	413978	########	IN	250/500	1000	1333.35	0	441716	MALE	PhD	prof-speci	board-gan	husband	0	-77000	########	Multi-vehi	Front Colli	Minor Dan	Police	NY
165	37	429027	########	IL	100/300	1000	1137.03	0	603195	MALE	Associate	tech-suppo	base-jump	unmarried	0	0	########	Multi-vehi	Front Colli	Total Loss	Police	VA
27	33	485665	########	IL	100/300	500	1442.99	0	601734	FEMALE	PhD	other-serv	golf	own-child	0	0	########	Single Vehi	Front Colli	Total Loss	Police	WV
212	42	636550	########	IL	100/300	500	1315.68	0	600983	MALE	PhD	priv-house	camping	wife	0	-39300	########	Single Vehi	Rear Collis	Total Loss	Other	NC
235	42	543610	########	OH	100/300	500	1253.12	4000000	462283	FEMALE	Masters	exec-mana	dancing	other-rela	38400	0	########	Single Vehi	Front Colli	Total Loss	Police	NY
447	61	214618	########	OH	100/300	2000	1137.16	0	615561	FEMALE	High Schoo	exec-mana	skydiving	other-rela	0	-51000	########	Multi-vehi	Front Colli	Major Dan	Fire	SC
60	23	842643	########	OH	500/1000	500	1215.36	3000000	432220	MALE	MD	protective	reading	wife	0	0	########	Single Vehi	Rear Collis	Total Loss	Ambulance	SC
121	34	626808	########	OH	100/300	1000	936.61	0	464652	FEMALE	MD	armed-for	bungie-jun	wife	52800	-32800	########	Parked Car	?	Minor Dan	None	SC
180	38	644081	########	OH	250/500	2000	1301.13	0	476685	FEMALE	College	machine-c	board-gan	not-in-fam	41300	-55500	########	Single Vehi	Rear Collis	Total Loss	Police	SC
473	58	892874	########	IN	100/300	2000	1131.4	0	458733	FEMALE	MD	transport-i	movies	other-rela	55700	0	########	Multi-vehi	Side Collisi	Major Dan	Other	WV
70	26	558938	########	OH	500/1000	1000	1199.44	5000000	619884	MALE	College	machine-c	hiking	own-child	63600	0	########	Multi-vehi	Rear Collis	Major Dan	Other	NY
140	31	275265	########	IN	500/1000	500	708.64	6000000	470610	MALE	High Schoo	machine-c	reading	unmarried	53500	0	########	Single Vehi	Side Collisi	Total Loss	Police	WV
160	37	921202	########	OH	500/1000	500	1374.22	0	472135	FEMALE	MD	craft-repai	yachting	other-rela	45500	-37800	########	Single Vehi	Side Collisi	Total Loss	Other	NY
196	39	143972	########	IN	500/1000	2000	1475.73	0	477670	FEMALE	High Schoo	handlers-c	camping	own-child	57000	-27300	########	Multi-vehi	Side Collisi	Major Dan	Police	VA
460	62	183430	########	IN	250/500	1000	1187.96	4000000	618845	MALE	JD	other-serv	bungie-jun	own-child	0	0	########	Multi-vehi	Rear Collis	Minor Dan	Police	NY
217	41	431876	########	IL	500/1000	2000	875.15	0	442479	FEMALE	Associate	machine-c	skydiving	own-child	46700	0	########	Multi-vehi	Side Collisi	Total Loss	Police	SC
370	55	285496	########	IL	100/300	2000	972.18	0	443920	MALE	High Schoo	prof-speci	paintball	other-rela	72700	-68200	########	Multi-vehi	Rear Collis	Major Dan	Ambulance	SC
413	55	115399	########	IN	100/300	2000	1268.79	0	453148	MALE	MD	priv-house	chess	own-child	0	-31000	########	Single Vehi	Front Colli	Total Loss	Ambulance	WV
237	40	736882	########	IN	100/300	1000	883.31	0	434733	MALE	College	craft-repai	kayaking	husband	0	-53500	########	Single Vehi	Rear Collis	Minor Dan	Other	VA
8	35	699044	########	OH	100/300	2000	1266.92	0	613982	MALE	Masters	sales	polo	own-child	0	0	########	Multi-vehi	Rear Collis	Major Dan	Other	OH
257	43	863236	########	IN	100/300	2000	1322.1	0	436984	MALE	High Schoo	prof-speci	golf	own-child	0	-29200	########	Parked Car	?	Minor Dan	Police	PA

Figure 5: Vehicle Insurance Claims CSV File Extract.

The dataset was distributed with data representing about 247 fraudulent claims, which made up 24.7% of the data, and 753 genuine claims, which made up 75.3% of the data, as seen in the bar graph in figure 6 below.

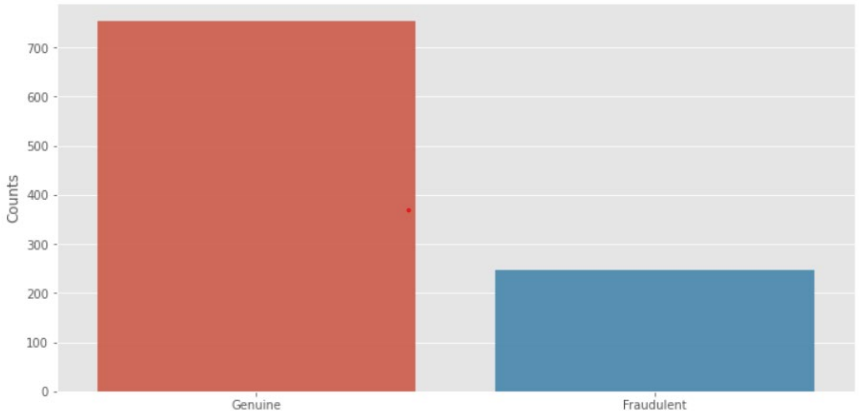


Figure 6: Vehicle Insurance Claims Distribution.

A total of 1000 rows and 39 columns made up the dataset, which also included the following input variables as denoted in figure 7 below: age, policy_number, policy_bind_date, policy_state, policy_deductible, policy_annual_premium, months_as_customer_insured, zip_code, insured_sex, insured_level_of_education, insured_job, insured_pastimes, insured_relationship, covered_capital_gains, insured_capital_loss, occurrence_date, occurrence_type, collision_type,

incident_severity, authorities_contacted, occurrence_state, occurrence_city, occurrence_location, occurrence_hour_of_the_day, vehicle_make, vehicle_model, witnesses, occurrence_bodily_injuries, occurrence_number_of_vehicles_involved, ,occurrence_police_report_available, occurrence_total_claim_amount, occurrence_property_damage, vehicle_year_of_manufacture and label to denote a claim as either genuine or fraudulent. Some of the input variables were later used to generate the features that were used to train the eight models used in this study, while others were eliminated for failing to reach the predetermined threshold.

```
Data columns (total 39 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
 6   policy_deductable            1000 non-null   int64
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64
 9   insured_zip                  1000 non-null   int64
 10  insured_sex                  1000 non-null   object
 11  insured_education_level      1000 non-null   object
 12  insured_occupation           1000 non-null   object
 13  insured_hobbies              1000 non-null   object
 14  insured_relationship         1000 non-null   object
 15  capital-gains                1000 non-null   int64
 16  capital-loss                 1000 non-null   int64
 17  incident_date                1000 non-null   object
 18  incident_type                1000 non-null   object
 19  collision_type               822 non-null    object
 20  incident_severity            1000 non-null   object
 21  authorities_contacted        1000 non-null   object
 22  incident_state               1000 non-null   object
 23  incident_city                1000 non-null   object
 24  incident_location            1000 non-null   object
 25  incident_hour_of_the_day     1000 non-null   int64
 26  number_of_vehicles_involved  1000 non-null   int64
 27  property_damage              640 non-null    object
 28  bodily_injuries              1000 non-null   int64
 29  witnesses                    1000 non-null   int64
 30  police_report_available      657 non-null    object
 31  total_claim_amount           1000 non-null   int64
 32  injury_claim                 1000 non-null   int64
 33  property_claim               1000 non-null   int64
 34  vehicle_claim                1000 non-null   int64
 35  auto_make                    1000 non-null   object
 36  auto_model                   1000 non-null   object
 37  auto_year                    1000 non-null   int64
 38  fraud_reported               1000 non-null   object
```

Figure 7: Dataset Columns Showing Input Variables

The dataset characteristics are summarized in table 1 below.

Number of Claims	1000
Number of Attributes	39
Categorical Attributes	24

Genuine Claims	753
Fraudulent Claims	247
Fraudulent Claims Incidence Rate	24.7%

Table 1: Dataset Features

The extracted dataset was also not entirely clean because several input variables had null values, as can be seen in figure 8 below, where collision_type variable was missing 178 values, property_damage was missing 360 values, and police_report_available variable was missing 343 values. However, no missing values were found in the data for the other input variables.

```
months_as_customer            0
age                           0
policy_number                 0
policy_bind_date              0
policy_state                  0
policy_csl                    0
policy_deductable             0
policy_annual_premium         0
umbrella_limit                0
insured_zip                   0
insured_sex                   0
insured_education_level       0
insured_occupation            0
insured_hobbies               0
insured_relationship          0
capital-gains                 0
capital-loss                  0
incident_date                 0
incident_type                 0
collision_type              178
incident_severity             0
authorities_contacted         0
incident_state                0
incident_city                 0
incident_location             0
incident_hour_of_the_day      0
number_of_vehicles_involved   0
property_damage             360
bodily_injuries               0
witnesses                     0
police_report_available     343
total_claim_amount            0
injury_claim                  0
property_claim                0
vehicle_claim                 0
auto_make                     0
auto_model                    0
auto_year                     0
fraud_reported                0
```

Figure 8: Dataset Columns Showing Null Values

3.2.3 Data Preparation

Because the dataset was acquired in raw format, pre-processing was required to generate high-quality features that would be presented to the ML classifiers. To analyse and choose quality features, this study employs a classical exploratory approach to data analysis. For machine learning to produce accurate and insightful results, data pre-processing is a crucial step. The reliability of the outcomes is inversely correlated with data quality. Real-world datasets are imperfect, inconsistent, and noisy in nature. Data pre-processing improves the data quality by addressing the gaps in the data, reducing noise, and addressing inconsistencies. Data preparation, according to Pandey (2019), entails cleaning, integrating, transforming, and reducing data to eliminate any duplicate or irrelevant data, leaving just the bits that provide valuable information to aid in establishing an efficient and effective classification. The stages in the procedure are as follows:

i. Data cleaning which aims to remove outliers from the dataset and impute missing values.
ii. Application of data transformation techniques like normalization. For instance, normalization may increase the precision and effectiveness of distance-based mining algorithms.
iii. Data integration, which combines data from several sources into one data warehouse.
iv. Data reduction, which involves removing redundant features from the data to lower its size. Techniques for feature extraction and feature selection can be used.

3.2.3.1 Data Clean-up

The data preparation process stated by checking for duplicate records and missing values. The missing values were then replaced with specified values using the fillna() python method in the dataset. Figure 9 below demonstrates the checking of duplicate and null values and replacement of null values with specifies values.

```
#Checking for duplicate claims
df.drop_duplicates(inplace = True)
df.shape
```

```
#We replace missing values with np.nan
df.replace('?', np.nan, inplace = True)
```

Handling missing values

```
df['collision_type'] = df['collision_type'].fillna(df['collision_type'].mode()[0])
df['property_damage'] = df['property_damage'].fillna(df['property_damage'].mode()[0])
df['police_report_available'] = df['police_report_available'].fillna(df['police_report_available'].mode()[0])
```

Figure 9: Checking and Filling Null Values

The dependent variable, fraud reported, was used as the starting point for exploratory data analysis. Heatmaps were created for variables with at least a 0.3 Pearson's correlation coefficient, including

the dependent variable, to better visualize the input variables within the dataset, aid in directing attention to areas of data visualizations that matter the most, and examine the relationships between them. The Pearson's correlation coefficient, which measures the linear relationship between two sets of data, is the ratio of the standard deviations of two continuous variables. Because the result is always between -1 and 1, it is effectively a normalized measurement of covariance (Statistics Solutions, 2022).

The heatmap analysis in figure 10 below demonstrates a strong correlation between month_as_customer and age, with a correlation of 0.92. This is most likely because people get vehicle insurance when they own a car, and because the time measure simply increases with age, therefore the "age" variable was dropped. The total_claim variable was also dropped because it was discovered that there was a strong correlation between the total_claim_amount, injury_claim, property_claim, and vehicle_claim variables. Additionally, to avoid redundancy, several of the data variables with high correlation were dropped. Afterwards, it was noticed there seemed not to be any multicollinearity issues, other from the possibility that all the claims are correlated and the total claims have been considered. On the other hand, the other claims offer a level of granularity that is not otherwise covered by total claims. As a result, these variables were retained.

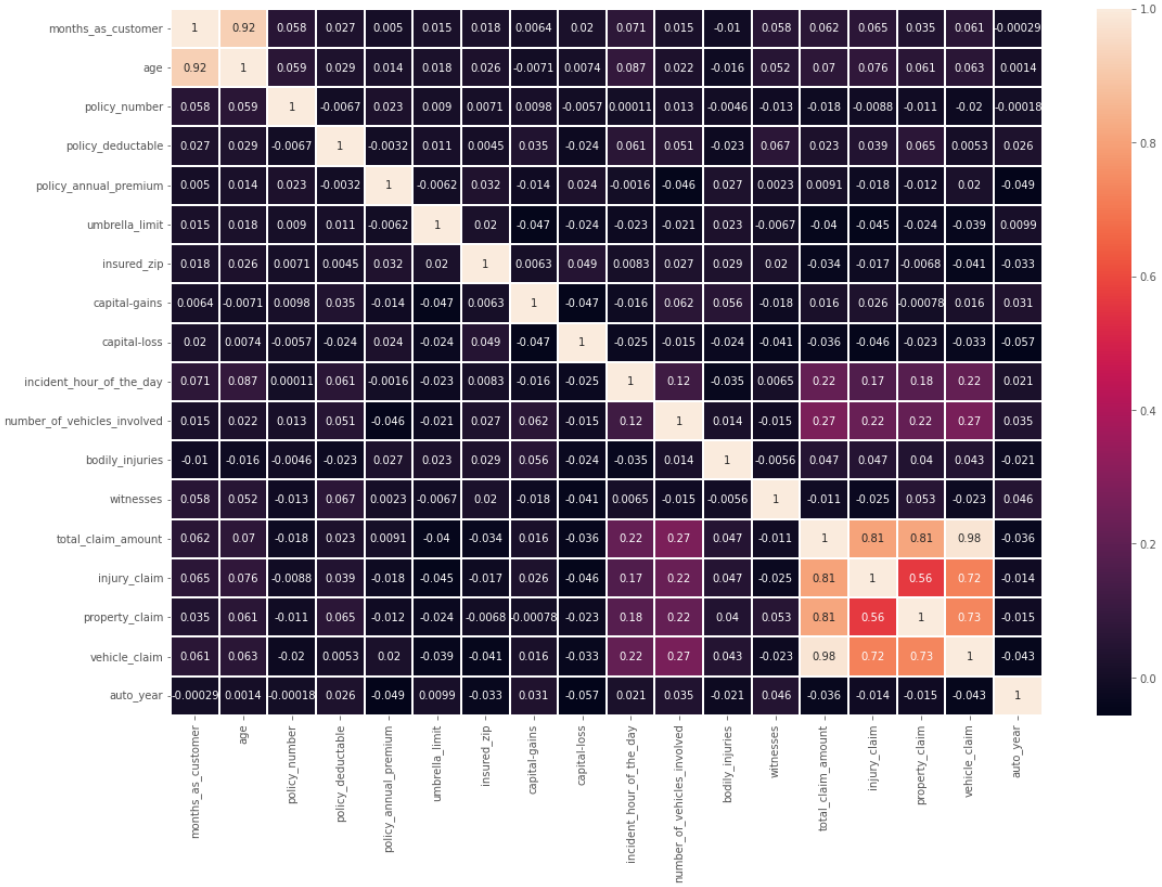


Figure 10: Correlation Heatmap Among Data Variables

Additional analysis was performed on the remaining data variables by identifying the unique values to obtain the reputable features to be utilized for classification, in addition to employing multicollinearity to eliminate some features as indicated on the heatmap in figure 10 above. A feature was eliminated if it had many unique values because there wasn't anything to be learned from them. Figure 11 below illustrates the number of unique values present in the remaining data variables. As a result, the following features were removed: policy_number, policy_bind_date, policy_state, insured_zip, incident_location, incident_date, incident_state, incident_city, insured_hobbies, auto_make, auto_model and auto_year.

months_as_customer	391
age	46
policy_number	1000
policy_bind_date	951
policy_state	3
policy_csl	3
policy_deductable	3
policy_annual_premium	991
umbrella_limit	11
insured_zip	995
insured_sex	2
insured_education_level	7
insured_occupation	14
insured_hobbies	20
insured_relationship	6
capital-gains	338
capital-loss	354
incident_date	60
incident_type	4
collision_type	3
incident_severity	4
authorities_contacted	5
incident_state	7
incident_city	7
incident_location	1000
incident_hour_of_the_day	24
number_of_vehicles_involved	4
property_damage	2
bodily_injuries	3
witnesses	4
police_report_available	2
total_claim_amount	763
injury_claim	638
property_claim	626
vehicle_claim	726
auto_make	14
auto_model	39
auto_year	21
fraud_reported	2

Figure 11: Unique Values Present in the Data Variables

3.2.3.2 Data Transformation

Data was transformed into formats that machine learning classifiers could interpret. For instance, text values must be converted into integer values since machine learning classifiers cannot interpret text values. We converted the categorical data into integer format to enable categorical data encoding, which enables categorical values to be fed into different models. This improved the predictions of our models. For the models to use the data with converted categorical values to produce and enhance the predictions, Verma (2021) defines categorical data encoding as the process of turning categorical data into integer format. He continued by defining categorical data as information that has been obtained and is organized into groups and has a limited number of possible values. The dataset's categorical data was extracted for conversion, and each column's unique values printed. Policy_csl, insured_sex, insured_education_level, insured_occupation, insured_relationship, incident_type, collision_type, incident_severity, authorities_contacted, property_damage, and police_report_available were the columns extracted. Figure 12 and 13 displays the unique values of the retrieved categorical data columns and the converted categorical data into integer format respectively.

```
policy_csl:
['250/500' '100/300' '500/1000']

insured_sex:
['MALE' 'FEMALE']

insured_education_level:
['MD' 'PhD' 'Associate' 'Masters' 'High School' 'College' 'JD']

insured_occupation:
['craft-repair' 'machine-op-inspct' 'sales' 'armed-forces' 'tech-support'
 'prof-specialty' 'other-service' 'priv-house-serv' 'exec-managerial'
 'protective-serv' 'transport-moving' 'handlers-cleaners' 'adm-clerical'
 'farming-fishing']

insured_relationship:
['husband' 'other-relative' 'own-child' 'unmarried' 'wife' 'not-in-family']

incident_type:
['Single Vehicle Collision' 'Vehicle Theft' 'Multi-vehicle Collision'
 'Parked Car']

collision_type:
['Side Collision' 'Rear Collision' 'Front Collision']

incident_severity:
['Major Damage' 'Minor Damage' 'Total Loss' 'Trivial Damage']

authorities_contacted:
['Police' 'None' 'Fire' 'Other' 'Ambulance']

property_damage:
['YES' 'NO']

police_report_available:
['YES' 'NO']
```

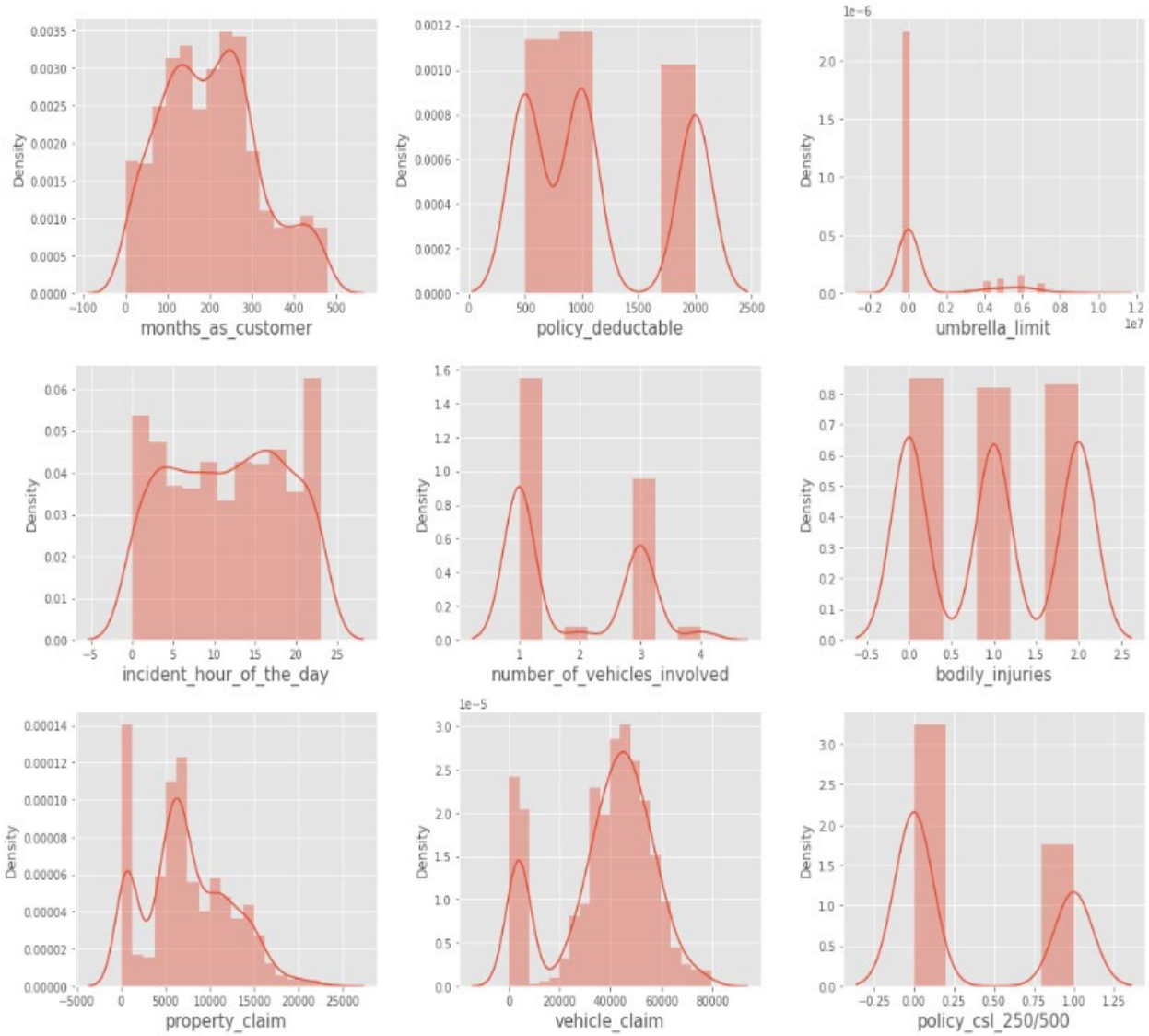
Figure 12: Categorical Data Columns Unique Values

	policy_csl_250/500	policy_csl_500/1000	insured_sex_MALE	insured_education_level_College	insured_education_level_High School
0	1	0	1	0	0
1	1	0	1	0	0
2	0	0	0	0	0

Figure 13: Converted Categorical Data Columns into Integer Values

3.2.3.3 Data Integration

To create the final dataset that would be utilized for both training and testing the various models, the columns that included numerical values were also extracted and combined with the converted numerical values from the categorical data. Figure 14 displays the distribution plot to show the variation in the final dataset's data distribution.



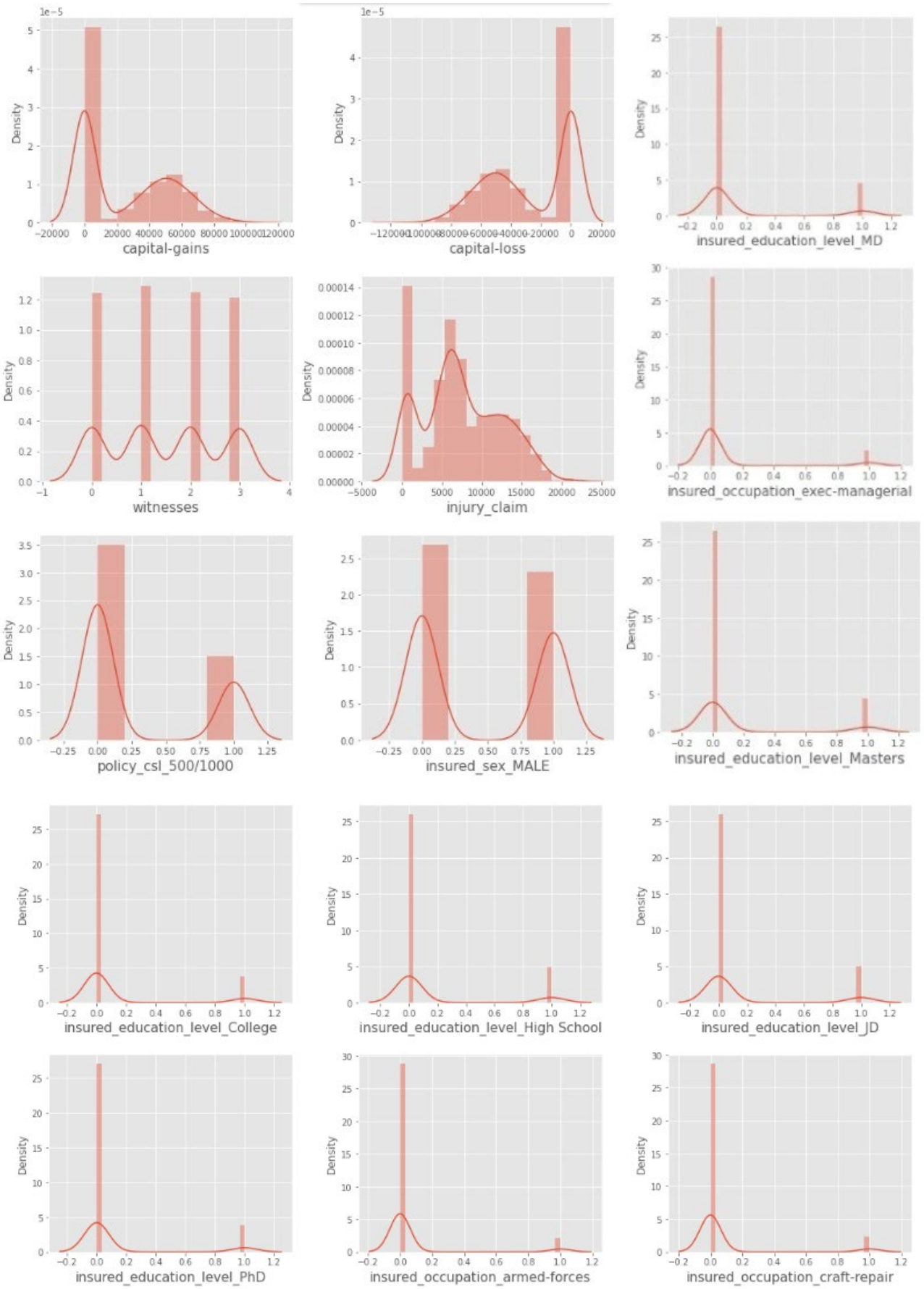


Figure 14: Final Dataset Data Distribution Plot

3.2.3.4 Feature Selection

We performed an analysis on the data to identify anomalies and outliers before splitting the dataset into training and testing sets. As a result, we scaled the numerical columns that had the outliers and deemed the data suitable for training. According to Tang et al. (2016), an outlier is an observation that deviates significantly from other values in a sample drawn at random from a population, almost as if the data were produced differently, or the potential for a data collecting error. The Inter Quantile Range (IQR) was utilized in this study to identify outliers. According to Tang et al. (2016), when values are sorted from lowest to highest, the IQR describes the median 50% of those values as shown in figure 15 below. The median (middle value) of the lower and upper half of the data is found first before calculating the IQR. These numbers are in the first quartile (Q1) and third quartile (Q3). The difference between Q3 and Q1 is the IQR. Some numerical columns were found to contain outliers; thus, scaling was applied to them.

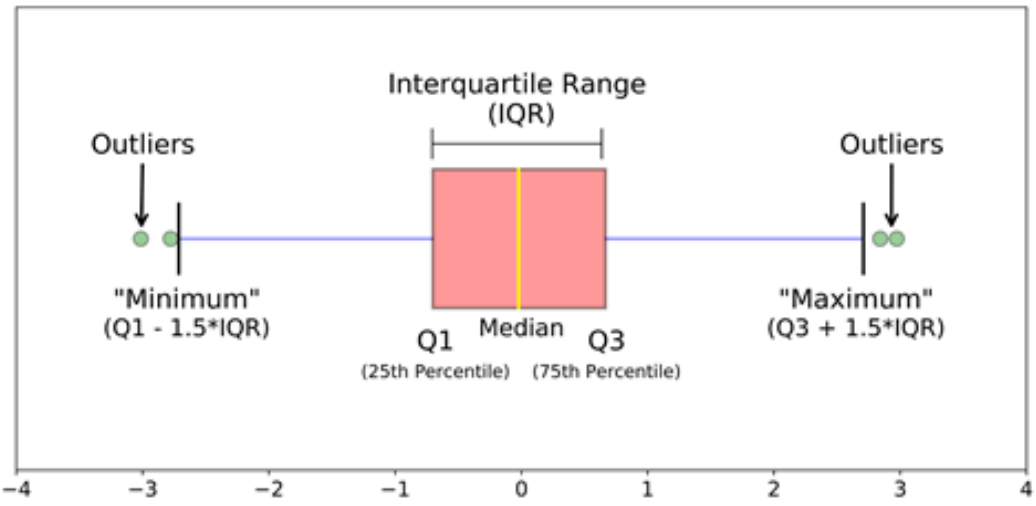


Figure 15: Inter Quantile Range Calculation Graph

The remaining dataset also included several attributes derived from the two variables that were identified for this study, such as information about insurance policies and information about insurance claims that matched the targeted level of fraud in the vehicle insurance claims. The following features as shown in figure 16 below were maintained for machine learning models classification as they satisfied the two variables identified for this study: months_as_customer, policy_csl, policy_deductable, policy_annual_premium, umbrella_limit, insured_sex, insured_education_level, insured_occupation, insured_relationship, capital-gains, property_damage, bodily_injuries, witnesses, incident_hour_of_the_day, number_of_vehicles_involved, injury_claim, property_claim, vehicle_claim and fraud_reported.

```
Data columns (total 25 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   policy_csl                   1000 non-null   object
 2   policy_deductable            1000 non-null   int64
 3   policy_annual_premium        1000 non-null   float64
 4   umbrella_limit               1000 non-null   int64
 5   insured_sex                  1000 non-null   object
 6   insured_education_level      1000 non-null   object
 7   insured_occupation           1000 non-null   object
 8   insured_relationship         1000 non-null   object
 9   capital-gains                1000 non-null   int64
 10  capital-loss                 1000 non-null   int64
 11  incident_type                1000 non-null   object
 12  collision_type               1000 non-null   object
 13  incident_severity            1000 non-null   object
 14  authorities_contacted        1000 non-null   object
 15  incident_hour_of_the_day     1000 non-null   int64
 16  number_of_vehicles_involved  1000 non-null   int64
 17  property_damage              1000 non-null   object
 18  bodily_injuries              1000 non-null   int64
 19  witnesses                    1000 non-null   int64
 20  police_report_available      1000 non-null   object
 21  injury_claim                 1000 non-null   int64
 22  property_claim               1000 non-null   int64
 23  vehicle_claim                1000 non-null   int64
 24  fraud_reported               1000 non-null   object
```

Figure 16: Features Maintained for Classification

3.2.4 Modelling

For this study, eight models - XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR - were trained and tested to identify the algorithm that performed the best with unbalanced and balanced datasets. The eight models were selected based on prior research that produced promising results when accuracy tests were conducted. The claims dataset was split into two parts: 20% for testing the classifier's prediction and 80% for training the classifiers. To ascertain which classifier performed the best, the accuracy of the results for each classifier employed was obtained.

After using the unbalanced dataset to train and test the classifiers, the dataset was then split into genuine and fraudulent class samples and the majority and minority class instances in the dataset were then identified. The classifiers were then put to the test in the process of balancing the data using the oversampling with SMOTE method. Subudhi et al. (2018) found that SMOTE chooses a point at random from the minority class to find its k-nearest neighbours. The synthetic points are inserted between the chosen point and its neighbours. Data was again divided into training and testing sets at 80% and 20%, respectively, after the data had been balanced.

To maximize the utilization of the data available for both training and testing the models, the K-Fold Cross Validation technique was employed to train all the classifiers. K-Fold Cross Validation is a statistical method for evaluating machine learning models that divides a given dataset into K folds, each of which serves as a testing set at some point. The K-Fold cross-validation technique yields less biased models because every data point from the original dataset appears in both the training and testing sets (Krishni, 2018). The dataset used in this study was divided into 10-fold cross validation (K=10) to fine-tune our models, with the first fold used in the first iteration for testing the models and the remaining folds for training them. The second fold functioned as the testing set in this iteration, while the remaining folds served as the training set. This procedure was repeated until all 10 folds were utilized as the testing set.

3.2.4.1 Experiment Environment

The study employed Google Colaboratory Notebook, popularly known as Colab, for modeling purposes (Google Inc., 2017). The Google notebook provides simple data sharing, allowing programmers to write and run Python in their browsers with no setup fees and free Graphics Processing Unit (GPU) access.

3.2.5 Evaluation

After all the classifiers had been trained with both balanced and unbalanced datasets, the model's performances were evaluated using the test data to see if they could categorize claims as genuine or fraudulent. An analysis of the performance categorization indicators was done in this study to gauge the model's efficiency and effectiveness, as well as establish their risk threshold. Confusion matrix, classification accuracy, classification report based on recall, precision, and F-1 score were the metrics employed. This found which classifier had the best levels of prediction performance and classification accuracy.

3.2.5.1 Confusion Matrix

A confusion matrix, according to Parab (2020), is a performance classification metric used to assess a machine learning algorithm's performance based on target classes. To determine the classification metrics above, the following values were first computed using a confusion matrix:

- True Positives (TP) – The amount of fraudulent vehicle insurance claims that were discovered.
- False Negatives (FN) - The amount of fraudulent vehicle insurance claims that went undiscovered.
- False Positives (FP) - The number of genuine vehicle insurance claims that were incorrectly categorized as fraudulent.
- True Negative (TN) - The proportion of genuine vehicle insurance claims that were not flagged as fraudulent.

3.2.5.2 Accuracy

According to Parab (2020), accuracy is the proportion of accurately predicted observations (True Positives) to all the input observations (sum of True Positives, False Positives, False Negatives, True Negatives).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions\ (TP + TN)}{Total\ Number\ of\ Predictions\ Made\ (TP + TN + FP + FN)}$$

3.2.5.3 Precision

A measure of precision is the proportion of accurately predicted positive samples (also known as True Positives) to the total number of predicted positive samples (sum of True Positives and False Positives) (Parab, 2020).

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Postives(FP)}$$

3.2.5.4 Recall

This is the proportion of correctly predicted positive samples (True Positives) to all samples in the actual class (sum of True Positives and False Negatives).

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives(FN)}$$

3.2.5.5 F-1 Score

F1 Score is the weighted average between precision and recall. The formula used to compute it is:

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

3.2.6 Deployment

An effective, novel system, based on the machine learning classifier with the highest levels of prediction performance and classification accuracy was developed to identify fraudulent vehicle insurance claims. The long-term profitability and consumer satisfaction of insurance businesses will benefit greatly from this.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Introduction

The key goal of this chapter is to present the research findings and to investigate how machine learning algorithms can leverage features extracted from vehicle insurance claim datasets to aid in the identification of fraudulent vehicle insurance claims with both balanced and imbalanced datasets. The best fraudulent detection results were determined via a comparative investigation of eight classification models, namely XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR.

4.2 Data Exploratory Analysis

We were unable to acquire a dataset on vehicle insurance claims after contacting a few Kenyan insurance companies due to the private and sensitive nature of the data it included. As a result, we collected data for this study from Kaggle (2018). 1,000 vehicle insurance claims made up the dataset, of which genuine claims made up 75.3% of the data and fraudulent claims made up 24.7% of the data. The dataset was unbalanced since it was heavily skewed in favour of genuine claims. By using a heatmap to examine for correlations between the data variables, exploratory data analysis was performed on the dependent and independent variables. This aided in cleaning up the dataset by getting rid of the variables that had a high degree of association with one another. The unnecessary features were dropped after the data exploratory analysis. The dataset also contained ccategorical data that was transformed into integer format to create features formats the machine learning classifiers could interpret hence improve the predictions of our models.

4.3 Machine Learning Classifier's Evaluation

An analysis of the following machine learning classification classifiers - XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR - was performed to assess how effectively and efficiently fraudulent vehicle insurance claims might be discovered. Using both unbalanced and balanced datasets, the classifiers were trained and evaluated to determine which model performed the best. AdaBoost and XGBoost classifiers were seen to execute considerably slow during training with balanced dataset, lasting approximately 2 minutes and 10 seconds, compared to the other models, which executed quickly, taking less than 7 seconds on average. AdaBoost and XGBoost similarly took a while to run on an unbalanced dataset, lasting about 1 minute, 35 seconds.

4.4 Performance Evaluation and Results

For the purposes of this study, eight classification models were constructed and trained using the selected features and a dataset divided into 80% for training and 20% for testing the classifiers. Following the training of all classifiers, the model's performance was evaluated, and the classification report of each classifier was calculated based on accuracy, recall, precision, and F-1 score to determine which classifier performed the best.

A confusion matrix is a classification performance indicator that is used to assess the effectiveness of a machine learning algorithm based on target classes. It is formed by generating TP, which are correctly classified positive claims, TN, which are correctly classified negative claims, FP, which are correctly classified negative claims but are positive claims, and FN, which are correctly classified negative claims but are positive claims. We further evaluated the confusion matrices for the imbalanced and balanced datasets to better comprehend the data results.

We also investigated other variables including as precision, recall, F1 score, and accuracy to acquire a deeper understanding of the effectiveness of our models. Precision is the percentage of class members among all those who were expected to be class members who were accurately identified. Recall is the percentage of all members of a class who were correctly predicted to be a member of that class. F1 score is a measure that combines recall and precision into a single metric. If precision and recall are both high, the F1 score will be high. Finally, accuracy of the models describes how well they performed and is calculated as the ratio of correct predictions to total predictions.

The results are shown in tables 2 and 3 below.

Classifier	TPs	FPs	TNs	FNs	Precision	Recall	F1 Score	Accuracy
SVM	45	115	36	4	0.28	0.92	0.43	0.685
Naïve Bayes	13	27	121	39	0.33	0.25	0.28	0.680
Logistic Regression	49	137	14	0	0.26	1.00	0.41	0.260
Decision Tree	32	38	110	20	0.46	0.62	0.52	0.635
Random Forest	24	14	134	28	0.63	0.46	0.53	0.735
AdaBoost	36	20	128	16	0.64	0.69	0.67	0.845
XGBoost	36	20	128	16	0.64	0.69	0.67	0.845
ANN	0	0	158	42	0.0	0.0	0.0	0.765

Table 2: Unbalanced Dataset Evaluation Report

The results in the confusion matrix above show that ANN performed best with unbalanced data since it had the highest percentage of true negatives in comparison to other classifiers. Random Forest

followed ANN closely, then followed by AdaBoost and XGBoost. Because it had the lowest percentage of true negatives and achieved the highest numbers of false positives, indicating that the classifier misclassified the fraudulent claims as genuine claims among the classifiers, the confusion matrix also showed that the Logistic Regression classifier performed poorly. As fraudulent claims were intrinsically unbalanced in the dataset used for this study, this suggests that AdaBoost, XGBoost and ANN performs well while Logistic Regression performs poorly on unbalanced data.

The models displayed different levels of performance on the input dataset, as can be seen from the findings in table 2 above. The F1 score was also used to rate the models according to how well they performed. The AdaBoost classifier was found to be superior to the other models by having the highest F1 score.

The AdaBoost and XGBoost classifiers performed the best out of all the classifiers with unbalanced data, achieving an accuracy rate of 84.5%, according to the results of the classification accuracy performance tests conducted on the eight classifiers. ANN classifier came in second at 76.5%. Accuracy rates of 73.5%, 68.5%, 68.0% and 63.5% were attained by Random Forest, SVM, Naïve Bayes and Decision Tree respectively. Logistic Regression came a distance last with an accuracy rate of 26.0% indicating it was not a good classifier to detect fraudulent insurance claims.

The dataset was then balanced using the oversampling with SMOTE method and all the classifiers retrained and retested, and results are shown in table 3 below.

Classifier	**TPs**	**FPs**	**TNs**	**FNs**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
SVM	40	32	124	106	0.58	0.73	0.64	0.513
Naïve Bayes	36	36	119	111	0.50	0.24	0.33	0.513
Logistic Regression	146	147	8	1	0.50	0.99	0.66	0.510
Decision Tree	82	45	110	65	0.65	0.56	0.60	0.636
Random Forest	110	46	138	37	0.87	0.75	0.80	0.821
AdaBoost	126	19	136	21	0.87	0.86	0.86	0.868
XGBoost	126	19	136	21	0.87	0.86	0.86	0.868
ANN	147	155	0	0	0.49	1.00	0.65	0.487

Table 3: Balanced Dataset Evaluation Report

Results on balanced data from the above confusion matrix revealed that ANN scored poorly because there were no true negative values, which indicated that they had completely failed to identify any

fraudulent claims. The AdaBoost, XGBoost, and Random Forest classifiers scored the highest percentages of true negatives, highlighting how proficiently the classifiers were able to distinguish between genuine and fraudulent claims. Despite having the lowest percentage of true negatives and the highest amounts of false positives after ANN, the Logistic Regression classifier still performed the badly on both unbalanced and balanced data. The AdaBoost, XGBoost, and Random Forest classifiers obtained the highest F1 scores making them superior to the other models.

It was further noted AdaBoost, XGBoost, Random Forest and Logistic Regression classifiers improved their predictive accuracy rates while SVM and Naïve Bayes classifiers dropped. Decision Tree classifier maintained an accuracy rating of 63.5%, showing the classifier was not affected by the balancing of the data. Of great concern was ANN which dropped considerably, portraying that it was not suited for to identifying fraudulent insurance claims using balanced data. The AdaBoost and XGBoost classifiers performed the best, achieving an accuracy rate of 86.8%, followed closely by Random Forest classifier at 82.1%.

4.5 Fraudulent Vehicle Claims Detection System

Based on the above-mentioned results, the AdaBoost or XGBoost classifier was chosen as the model to be utilized with the Web-based application to identify fraudulent vehicle insurance claims. Our web-based application was developed using the Streamlit Framework. According to Patil and Lokesha (2022), Streamlit is an open-source, Python-based platform for developing and deploying interactive data science dashboards and machine learning models on web applications. Figure 17 below displays a screen image of a web application that prompts the user to submit a csv or excel file containing the claims and the model for identifying legitimate and fraudulent vehicle insurance claims.

Figure 17: Web-Based Application Screen Image.

The vehicle insurance claims were accurately classified as either genuine or fraudulent claims as shown in the screen image in figure 18 after uploading a csv file containing the vehicle insurance claims and the saved AdaBoost or XGBoost models to the web application. A csv file was then generated that the user could download showing the results produced, as shown in the csv file extract in figure 19 below.

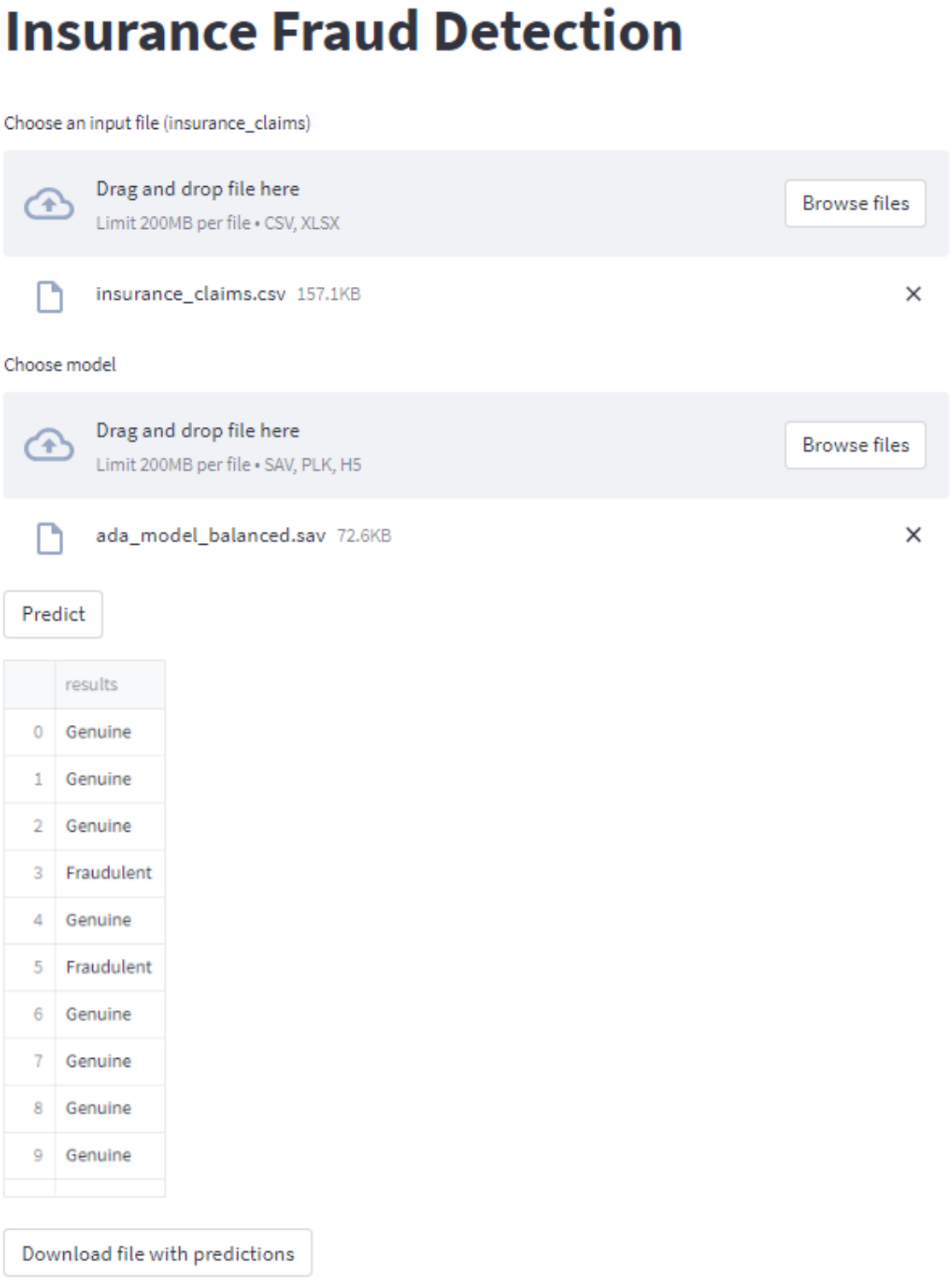
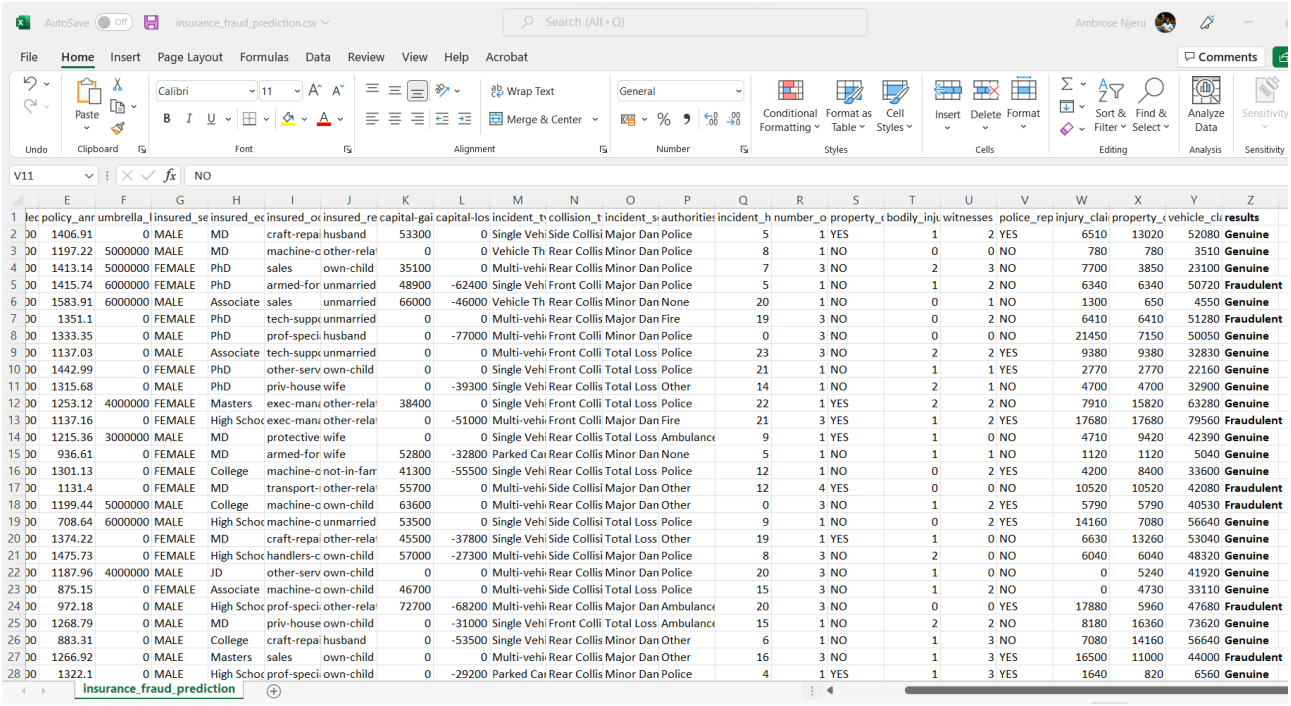


Figure 18: Categorized Vehicle Insurance Claims



	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	lec	policy_ann	umbrella_l	insured_se	insured_ed	insured_oc	insured_re	capital-gai	capital-los	incident_t	collision_t	incident_s	authorities	incident_h	number_o	property_d	bodily_inju	witnesses	police_rep	injury_clai	property_d	vehicle_cla	results
2	00	1406.91	0	MALE	MD	craft-repa	husband	53300	0	Single Vehi	Side Collisi	Major Dan	Police	5	1	YES	1	2	YES	6510	13020	52080	Genuine
3	00	1197.22	5000000	MALE	MD	machine-c	other-rela	0	0	Vehicle Th	Rear Collis	Minor Dan	Police	8	1	NO	0	0	NO	780	780	3510	Genuine
4	00	1413.14	5000000	FEMALE	PhD	sales	own-child	35100	0	Multi-vehi	Rear Collis	Minor Dan	Police	7	3	NO	2	3	NO	7700	3850	23100	Genuine
5	00	1415.74	6000000	FEMALE	PhD	armed-for	unmarried	48900	-62400	Single Vehi	Front Colli	Major Dan	Police	5	1	NO	1	2	NO	6340	6340	50720	Fraudulent
6	00	1583.91	6000000	MALE	Associate	sales	unmarried	66000	-46000	Vehicle Th	Rear Collis	Minor Dan	None	20	1	NO	0	1	NO	1300	650	4550	Genuine
7	00	1351.1	0	FEMALE	PhD	tech-supp	unmarried	0	0	Multi-vehi	Rear Collis	Major Dan	Fire	19	3	NO	0	2	NO	6410	6410	51280	Fraudulent
8	00	1333.35	0	MALE	PhD	prof-speci	husband	0	-77000	Multi-vehi	Front Colli	Minor Dan	Police	0	3	NO	0	0	NO	21450	7150	50050	Genuine
9	00	1137.03	0	MALE	Associate	tech-supp	unmarried	0	0	Multi-vehi	Front Colli	Total Loss	Police	23	3	NO	2	2	YES	9380	9380	32830	Genuine
10	00	1442.99	0	FEMALE	PhD	other-serv	own-child	0	0	Single Vehi	Front Colli	Total Loss	Police	21	1	NO	1	1	YES	2770	2770	22160	Genuine
11	00	1315.68	0	MALE	PhD	priv-house	wife	0	-39300	Single Vehi	Rear Collis	Total Loss	Other	14	1	NO	2	1	NO	4700	4700	32900	Genuine
12	00	1253.12	4000000	FEMALE	Masters	exec-mana	other-rela	38400	0	Single Vehi	Front Colli	Total Loss	Police	22	1	YES	2	2	NO	7910	15820	63280	Genuine
13	00	1137.16	0	FEMALE	High Schoo	exec-mana	other-rela	0	-51000	Multi-vehi	Front Colli	Major Dan	Fire	21	3	YES	1	2	YES	17680	17680	79560	Fraudulent
14	00	1215.36	3000000	MALE	MD	protective	wife	0	0	Single Vehi	Rear Collis	Total Loss	Ambulance	9	1	YES	1	0	NO	4710	9420	42390	Genuine
15	00	936.61	0	FEMALE	MD	armed-for	wife	52800	-32800	Parked Car	Rear Collis	Minor Dan	None	5	1	NO	1	1	NO	1120	1120	5040	Genuine
16	00	1301.13	0	FEMALE	College	machine-c	not-in-fam	41300	-55500	Single Vehi	Rear Collis	Total Loss	Police	12	1	NO	0	2	YES	4200	8400	33600	Genuine
17	00	1131.4	0	FEMALE	MD	transport-l	other-rela	55700	0	Multi-vehi	Side Collisi	Major Dan	Other	12	4	YES	0	0	NO	10520	10520	42080	Fraudulent
18	00	1199.44	5000000	MALE	College	machine-c	own-child	63600	0	Multi-vehi	Rear Collis	Major Dan	Other	0	3	NO	1	2	YES	5790	5790	40530	Fraudulent
19	00	708.64	6000000	MALE	High Schoo	machine-c	unmarried	53500	0	Single Vehi	Side Collisi	Total Loss	Police	9	1	NO	0	2	YES	14160	7080	56640	Genuine
20	00	1374.22	0	FEMALE	MD	craft-repa	other-rela	45500	-37800	Single Vehi	Side Collisi	Total Loss	Other	19	1	YES	1	0	NO	6630	13260	53040	Genuine
21	00	1475.73	0	FEMALE	High Schoo	handlers-c	own-child	57000	-27300	Multi-vehi	Side Collisi	Major Dan	Police	8	3	NO	2	0	NO	6040	6040	48320	Genuine
22	00	1187.96	4000000	MALE	JD	other-serv	own-child	0	0	Multi-vehi	Rear Collis	Minor Dan	Police	20	3	NO	1	0	NO	0	5240	41920	Genuine
23	00	875.15	0	FEMALE	Associate	machine-c	own-child	46700	0	Multi-vehi	Side Collisi	Total Loss	Police	15	3	NO	1	2	NO	0	4730	33110	Genuine
24	00	972.18	0	MALE	High Schoo	prof-speci	other-rela	72700	-68200	Multi-vehi	Rear Collis	Major Dan	Ambulance	20	3	NO	0	0	YES	17880	5960	47680	Fraudulent
25	00	1268.79	0	MALE	MD	priv-house	own-child	0	-31000	Single Vehi	Front Colli	Total Loss	Ambulance	15	1	NO	2	2	NO	8180	16360	73620	Genuine
26	00	883.31	0	MALE	College	craft-repa	husband	0	-53500	Single Vehi	Rear Collis	Minor Dan	Other	6	1	NO	1	3	NO	7080	14160	56640	Genuine
27	00	1266.92	0	MALE	Masters	sales	own-child	0	0	Multi-vehi	Rear Collis	Major Dan	Other	16	3	NO	1	3	YES	16500	11000	44000	Fraudulent
28	00	1322.1	0	MALE	High Schoo	prof-speci	own-child	0	-29200	Parked Car	Rear Collis	Minor Dan	Police	4	1	YES	1	3	YES	1640	820	6560	Genuine

Figure 19: Generated CSV File of Categorized Vehicle Insurance Claims.

4.6 Study Discussions

According to the literature review, XGBoost outperformed LR, SVM, and RF classifiers (Shah et al. ,2021). This study, in which the XGBoost classifier outperformed these classifiers, recommended a technique that is now supported by their research. Our study also showed that when classifying vehicle insurance claims using balanced data, ANN is not the best classifier to use. This is supported by Jalali (2020), who discovered that their neural network-based model underperformed since it couldn't reliably detect any fraudulent claims, as was the case with our research. Finally, this study supports Sunita Mall et al. (2018) research, in which characteristics were derived based on information about the insurer's information, the type of vehicle, the vehicle's maximum tonnage, the insurance branch's code, the insured amount, the paid loss, the claim's specifics, the vehicle's age, etc. Similar elements were also collected from our research to generate our features.

The dataset utilized in this study was also used in the research of Gondalia et al. (2022), who balanced the dataset using the ADASYN sampling technique and then used the Random Forest classifier to classify the provided cases of vehicle insurance claims as fraudulent or genuine. This investigation confirms their findings in that the accuracy of the models was impaired and performed badly because of the imbalanced dataset. As a result, to improve model accuracy, the dataset must be balanced. Punith, (2021) used the same dataset in their research. They used the SMOTE methodology to balance the data and identified five classifiers to perform claim classification. These were LR, DT, RF, NB

and XGBoost. The RF classifier outperformed all other models in their experiment, with an accuracy score of 85.39% and an F1 score of 85.20% with balanced data. This validated our analysis because the RF classifier achieved an accuracy score of 82.1% and an F1score of 80%, even though the AdaBoost and XGBoost classifiers performed the best using the identical dataset and data balancing technique. Their research had certain drawbacks as well. For instance, it was limited by a small data sample size, just like our experiment, because models are more stable with larger datasets.

Finally, Chew (2020) created a model to detect vehicle insurance fraud using the same dataset as in this study. They used the F1 score as a measure of model performance on imbalanced data with a focus on the minority class, because accuracy is not a good measure of success when the dataset is imbalanced. The classifiers used in their investigation were LR, K-nearest Neighbors, RF, XGBoost, and AdaBoost. Their study revealed that the XGBoost classifier performed the best with the imbalanced dataset, with an F1 score of 0.72, which is corroborated by our findings that XGBoost and AdaBoos performed the best, with an F1 score of 0.67.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

Several researchers have developed and applied machine learning models to identify fraudulent insurance claims. We were capable of demonstrating how machine and deep learning technologies can be used to address the troublesome and chronic problem of fraudulent insurance claims in the vehicle insurance industry through our research. The study also demonstrates how stakeholders in the insurance industry may cooperate to create a model that benefits everyone, as opposed to just their own business, and how doing so can help insurance risk personnel identify fraudulent insurance claims while reducing the losses incurred by the insurance companies.

The main goal of this study was to investigate how machine learning algorithms may be utilized to detect fraudulent claims using features extracted from vehicle insurance claim databases. Thereafter, a novel web-based application was created out of this research to assist the insurance industry in identifying and classifying vehicle insurance claims as genuine or fraudulent based on the best performing machine learning algorithm.

5.2 Summary of Findings

The AdaBoost and XGBoost classifiers outperformed the other models with both unbalanced and balanced data because they had the highest classification accuracy of 84.5%. This allowed us to utilize any of them for the web application. Additionally, it was observed that with the two models having accuracy levels of 80% and above on both unbalanced and balanced data, indicated that the features set retrieved mostly matched the characteristics of fraudulent claims. With both unbalanced and balanced data, the Logistic Regression classifier performed terribly because it had the lowest classification accuracy. In contrast to balanced data, the ANN classifier performed better with unbalanced data. Finally, it was discovered that all eight classifiers could only be employed on smaller datasets. This was because none of the classifiers could manage handling huge datasets without crashing the Colab graphics processing unit.

5.3 Study Conclusion

Due to the ICT revolution of the twenty-first century, people are increasingly interacting differently with insurance companies while lodging insurance claims. Unfortunately, the increased use of ICT

has resulted in an increase in the number of fraudulent claims submitted. Fraudsters regularly create novel, impenetrable techniques to evade the built-in fraud detection systems. This has inspired and pushed researchers to seek out novel, workable solutions to the problem at hand, which has resulted in the invention of machine learning techniques, which are currently the subject of intense research by academics and business professionals (Ayowemi et al, 2017).

To counter the growing problem of lodging fraudulent insurance claims, technological advancement has resulted in research, development, and application of methods for detecting these claims. Unfortunately, despite the various strategies used to combat this prevalent problem, attackers are using new tactics and becoming more skilled at creating plans that allow them to get around the measures taken, and it is escalating into a serious threat to security, organizations, and the global economy. As a result, I have attempted to close the gap in the literature by presenting in this research project a practical and efficient web-based system that uses the best machine learning technique to classify insurance claims as either genuine or fraudulent.

The AdaBoost and XGBoost classifiers, which had an accuracy rate of 86.7% out of the eight machine learning classifiers that were evaluated, performed better using balanced data and using unbalanced data with an accuracy rate of 84.5%. It was noted with data balancing, higher accuracy rates were generated hence more accurate models created.

5.4 Study Achievements

The study's primary objective was to investigate how features extracted from vehicle insurance claim dataset could be used by machine learning algorithms to help detect fraudulent vehicle insurance claims. Following that, a novel system for predicting and categorizing vehicle insurance claims as genuine or fraudulent was developed. This was accomplished by training and testing a model utilizing seven ML classifiers and one deep learning classifier on features taken from the vehicle insurance dataset utilized in this study. A performance evaluation of all models was performed, and a web-based system was developed to obtain vehicle insurance claims files and the best performing classifier, in this case XGBoost or AdaBoost, to predict and categorize vehicle insurance claims as either genuine or fraudulent.

The study's specific objectives were to characterize fraudulent insurance claims in the context of the vehicle insurance domain, identify features that could be used to train ML models to recognize fraudulent vehicle insurance claims, evaluate the performance of different ML models for identifying

fraudulent vehicle insurance claims using balanced and unbalanced datasets, and finally develop a system to categorize vehicle insurance claims as either genuine or fraudulent using the best performing machine learning classifier. All of this was achieved by first attempting to comprehend the insurance industry business, with a particular emphasis on the vehicle insurance sector. Several sources were used, as well as information from insurance companies. After obtaining relevant data for the study, we were able to analyse its quality, develop a foundational understanding of the data, and extract variables from the dataset to aid in model construction. Because the information was gathered in raw format, a classical exploratory approach to data analysis was used as a data preparation pre-processing method to build high-quality features that would be supplied to the machine learning classifiers. This aided us in characterizing fraudulent insurance claims in the context of the vehicle insurance domain and identifying features utilized to train the models.

Machine learning classifiers were trained on 80% of the data and tested on 20%. The performance results for each classifier used, were generated to determine which classifier performed the best. This helped us choose the best performing model utilizing imbalanced data and after balancing the same data with the SMOTE approach. As explained in chapter 4, the performance results revealed the best performing model, which was later utilized to classify vehicle insurance claims as genuine or fraudulent using the web-based application that was also developed. Finally, all the objectives outlined in this study were entirely met.

5.5 Study Limitations

Finding a dataset with the Kenyan insurance industry as its focus was the significant hurdle to this study's accomplishment. For the purposes of conducting this research, we contacted Britam Insurance Kenya, Co-operative Insurance Company (CIC), Jubilee Car Insurance, APA Car Insurance, and Heritage Insurance Company; however, due to the sensitivity and privacy of the data they hold, they were unable to provide us with any information.

Additionally, only smaller datasets could be employed on all the eight classifiers used because none of the classifiers could manage handling huge datasets without damaging the Colab graphics processing unit.

5.6 Study Recommendations

More features found in fraudulent claims should be included to the study's recommendations to make them more effective at identifying malware that might be downloaded onto users' devices via

insurance claims. This will increase the effectiveness of the suggested solution and broaden the classification of fraudulent claims. Additionally, to develop better explicit classifiers, it is also appropriate to conduct trials using very large training and testing datasets. Finally, the proposed system can be modified, enhanced, and appropriately expanded to accommodate the vehicle insurance industry and be adopted for commercial use.

5.7 Future Work Suggestions

For the insurance industry, detecting fraudulent claims is a significant challenge, hence and it is proposed that this system be improved by combining machine learning approaches with nature-inspired optimization algorithms. Nature-inspired algorithms are a collection of unique problem-solving methodologies and approaches that are drawn from the behaviour of natural situations (Xin-She, 2014). By bridging the gap created by machine learning algorithms' inability to handle massive datasets, models for identifying false claims will be developed more rapidly and effectively when paired with nature-inspired optimization algorithms. Another benefit of combining machine learning algorithms with nature-inspired optimization algorithms is that a powerful technique can be developed to find the best features of fraudulent insurance claims automatically and dynamically with very high classification accuracy. Finally, future research may also look towards acquiring larger datasets spanning multiple years.

REFERENCES

African Insurance Organization, (2018). Africa Insurance Barometer 2018; Market Survey. Cameroon, Douala.

Asiri, S., (11 June, 2018). Machine Learning Classifiers. Decision Trees. Available at: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623.

Association of Kenya Insurers, (2020). Insurance Industry Annual Report. Nairobi.

Awoyemi, J.O., Adetunmbi, A.O., & Oluwadare, S.A., (2017). Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis. In: Proceedings IEEE International Conference Computing Networking Informatics, ICCNI 2017, pp. 1–9.

Bauder, R..A. & Khoshgoftaar, T.M., (2017). Medicare Fraud Detection Using Machine Learning Methods. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 858-865.

Baumann, M., (2021). Improving a Rule-based Fraud Detection Systemwith Classification Based on Association Rule Mining. Available at: https://www.researchgate.net/publication/349244021_Improving_a_Rule-based_Fraud_Detection_System_with_Classification_Based_on_Association_Rule_Mining.

Bhavna, B., & Sheetal, K., (2019). Naïve Classification Approach for Insurance Fraud Prediction. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5.

Breiman, L. & Cutler, A., (2007). Random Forests Classification Description. Department Of Statistics Homepage. Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Burri, R.D., Burri, R., Bojja, R.R., & Buruga, S.R. (2019). Insurance Claim Analysis Using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering Vol, Issue 6, Special Issue 4, pp.577-582.

Charles, M., Ahmed, A., & Thokozani, S. (2020). Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method. 61st International

Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), pp. 1-5.

Chew, I., (2020). For Real? Auto Insurance Fraud Claim Detection with Machine Learning. Published in Towards Data Science. . Available at: https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3.

Coalition Against Insurance Fraud (2016). The State of Insurance Fraud Technology: A Study of Insurer, Strategies and Plans for Ant-Fraud Technology.

Cortes, C., & Vapnik, V., (1995). Support-Vector Networks. Machine Learning, vol. 20,no. 3. pp. 273–297.

DeBarr, D., & Wechsler, H. (2013). Fraud Detection Using Reputation Features , SVMs , and Random Forests. Available at: http://worldcomp-proceedings.com/proc/p2013/DMI8055.pdf.

Derrig, R.A. (2002), Insurance Fraud. Journal of Risk and Insurance, 69(3), pp.271-287.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations. 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp.1-5.

Dimitrakopoulos, G.N., Vrahatis, A.G., Plagianakos, V., Sgarbas, K., (2018). Pathway Analysis Using XGBoost Classification in Biomedical Data. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp 1–6.

Dull, R. (2014). What Gets Monitored Gets Detected. Journal of Accountancy, Feature Fraud/Technology. Available at: http://www.journalofaccountancy.com/issues/2014/feb/20137694.html.

Ernst & Young (2011). Fraud Insurance on the Rise. India Survey 2010-2011.

Frimpong, I., A. (2016). Causes, Effects and Deterrence of Insurance Fraud: Evidence from Ghana. MPHIL Thesis, University of Ghana.

Gedela, B., & Karthikeyan, P. R. (2022). Credit Card Fraud Detection using AdaBoost Algorithm in Comparison with Various Machine Learning Algorithms to Measure Accuracy, Sensitivity,

Specificity, Precision and F-score. International Conference on Business Analytics for Technology and Security (ICBATS), 2022, pp. 1-6.

Gill, K. M., Woolley, A., & Gill, M. (2005). Insurance Fraud: The Business as a Victim? Crime at Work. Palgrave Macmillan, London, pp. 73–82.

Gepp, A., Wilson, J.H., Kumar, K., & Bhattacharya, S. (2012). A Comparative Analysis of Decision Trees Vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. Journal of Data Science, Vol. 10, pp. 537-556.

Gondalia, D., Gurav, O., Joshi, A., Joshi, A., & Selvan, S., (2022). Automobile Insurance Claim Fraud Detection. International Research Journal of Engineering and Technology (IRJET), Vol.09 Issue. 01, pp. 906 – 909.

Google Inc. (2017). What is Colaboratory? Available at: https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=5fCEDCU_qrC0.

Graupe, D., (2016). Deep Learning Neural Networks. University of Illinois, Chicago, USA: World Scientific Publishing Company.

IBM & SPSS Modeler (n.d.). Using Data Mining Detect Insurance Fraud: Improve Accuracy and Minimize Loss. IBM Software Business Analytics.

International Association of Insurance Supervisors (2011): Application Paper on Deterring, Preventing, Detecting, Reporting and Remedying Fraud In Insurance. Available at: https://www.iaisweb.org/uploads/2022/01/Application_paper_on_fraud_in_insurance.pdf.pdf.

Insurance Regulatory Authority (2021). Insurance Industry Annual Report for the Year 2020. Available at: https://www.ira.go.ke/images/annual_2020/INSURANCE-INDUSTRY-ANNUAL-REPORT1.pdf.

Jalali, B., (2020). Detecting Fraudulent Claims – A Machine Learning Approach. Gen Re, Cologne. Available at: https://www.genre.com/knowledge/publications/ri20-1-en.html.

Jamal, D., (2017). Multicollinearity and Regression Analysis. Journal of Physics: Conference Series. 949. 012009.

Jason, B., (April 1, 2016). Logistic Regression for Machine Learning. Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Joachims, T., (2001). A Statistical Learning Model of Text Classification with Support Vector Machine. SIGIR'01, Proceedings of International Conference on Research and Development in Information Retrieval. New Orleans, LA, USA: ACM. pp. 128–136.

Kamau, J., N. (2015). The Effect of Forensic Accounting Services and Fraud Prevention in the Insurance Companies of Kenya. MBA Research Project, University of Nairobi.

Krishni. (2018). K-Fold Cross Validation. Available at: https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833. [Accessed on: 15 September, 2022]

Lewis D. D., (1998). Naïve Bayes at Forty: The Independence Assumption In Information Retrieval. In European Conference on Machine Learning. Springer. pp. 4–15.

Mark, A.,C & Liam, G., (2021) Automobile Insurance Fraud Detection, Communications in Statistics: Case Studies, Data Analysis and Applications, pp. 520-535.

Mathenge, M., N., (2016). Effects of Internal Audit Functions on Fraud Detection in Insurance Companies in Kenya. MBA Research Project, University of Nairobi.

Mishra, A. (2021). Fraud Detection: A Study of AdaBoost Classifier and K-Means Clustering. Available at SSRN: https://ssrn.com/abstract=3789879.

Moon, H., Pu, Y. & Ceglia, C. (2019) A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. Theoretical Economics Letters, 9, pp1886-1900.

Moore, F., O. (2016). Information Systems and Fraud Detection. Term Paper, Northcentral University.

Owusu-Oware, E., Effah, J., & Boateng, R., (2018). Biometric Technology for Fighting Fraud in National Health Insurance: Ghana's Experience. Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018.

Pandey, P., (2019). Machine Learning Data Preprocessing: Concepts. Available at: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825.

Parab, R., (2020). Performance Evaluation Metrics for Machine Learning Models with Python Code. Available at: https://medium.com/swlh/performance-evaluation-metrics-for-machine-learning-models-ad0dd480d5af.

Patil, S., and Lokesha, V., (2022). Live Twitter Sentiment Analysis Using Streamlit Framework. Available at SSRN: https://ssrn.com/abstract=4119949 or http://dx.doi.org/10.2139/ssrn.4119949.

Punith, (2021). Insurance Claims - Fraud Detection Using Machine Learning. Published in Geek Culture. Available at: https://medium.com/geekculture/insurance-claims-fraud-detection-using-machine-learning-78f04913097.

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P. & Nandi, A. K., (2018). Credit Card Fraud Detection Using AdaBoost and Majority Voting. IEEE Access, vol. 6, pp. 14277-14284.

Rodrigues, I., (2020). CRISP-DM Methodology Leader in Data Mining and Big Data. Available at: https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781.

Simha, A., & Satyanarayan, S. (2016). Straight From the Horse's Mouth: Auditors' on Fraud Detection and Prevention, Roles of Technology, and White-Collars Getting Splattered with Red! Journal of Accounting & Finance Vol. 16(1), pp. 26-44.

Shah, B. (2018). Auto Insurance Claims Data. Available at: https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data/metadata. [Downloaded: 15 June, 2022].

Shah, S., Phadke, S., Koli, P., & Sharma, S., (2021). Insurance Fraud Detection using Machine Learning. International Research Journal of Engineering and Technology (IRJET). Vol: 08 Issue: 04.

Soni, R., R., & Soni, N., (2013). An Investigative Study of Banking Cyber Frauds with Special Reference to Private and Public Sector Banks, Research Journal of Management Sciences, Vol. 2(7), pp. 22-27.

Statistics Solutions (2022). Directory of Statistical Analyses, Pearson's Correlation Coefficient. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/.

Subudhi, S. & Panigrahi, S. (2018). Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud. 2nd International Conference on Data Science and Business Analytics (ICDSBA), pp. 528-531.

Sunita, M., Prasun, G., & Parita, S. (2018). Management of Fraud: Case of an Indian Insurance Company. Accounting and Finance Research, Vol 7, No 3.

Sybase, (2012). Fraud is a Significant and Costly Problem for both Policyholders and Insurance Companies in the Insurance Sector. International Journal of Innovative Social & Science Education Research., Vol. 2, No.1., pp. 26-39.

Tang, B., Kay, S., & He, H., (2016). Toward Optimal Feature Selection in Naïve Bayes For Text Categorization. IEEE Transactions on Knowledge and Data Engineering, Vol.28, No.9. pp. 2508-2521.

Wang, Y., & Xu, W. (2018). Leveraging Deep Learning with LDA-based Text Analytics to Detect Automobile Insurance Fraud. Decision Support Systems., Vol. 105, pp. 87-95.

Wilson, J.H. (2009). An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression. Journal of finance and Accountancy. Available at: https://www.researchgate.net/publication/253116638_An_Analytical_Approach_To_Detecting_Insurance_Fraud_Using_Logistic_Regression/citations.

Verma, Y (2021). A Complete Guide to Categorical Data Encoding. Analytics India Magazine. Available at: https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/.

Viaene, S. & Dedene, G. (2015), Insurance Fraud: Issues and Challenges, Geneva Papers on Risk and Insurance-Issues and Practice., Vol. 29, No.2., pp. 313-333.

Xin-She, Y. (2014). Nature-Inspired Optimization Algorithms.

APPENDICES

Appendix 1: Project Budget

No:	Item	Amount (KShs)
1.	HP Elite Book 850G8 i7 4.8GHz, 16GB RAM and 512 GB SSD Laptop	194,000.00
2.	Google Colaboratory - Python Development - Open Source	0.00
3.	Web Development - Streamlit Python Platform - Open Source	0.00
4.	Vehicle Insurance Claims Dataset Collection – Online Download	0.00
5.	Documentation Photocopying, Printing and Binding	17,000.00
6.	Airtime	1,000.00
7.	Transport	8,000.00
Total		**220,000.00**

Table 4: Project Budget

Appendix 2: Project Schedule

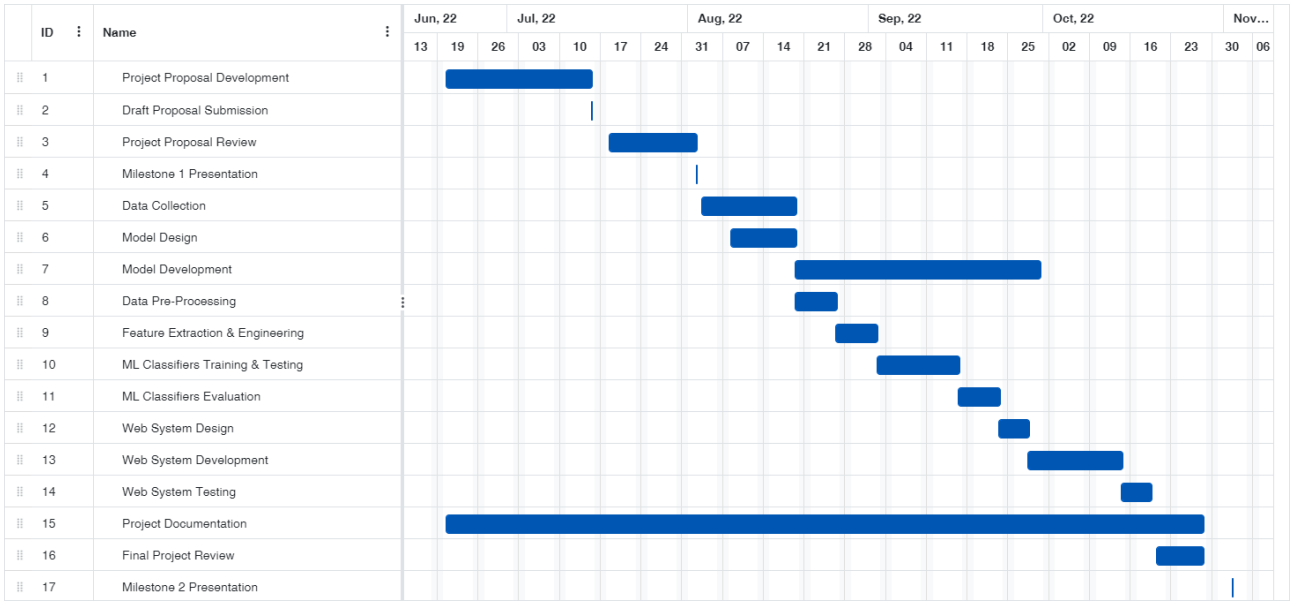


Figure 20: Project Schedule

Appendix 3: Motor Vehicle Insurance Proposal Form

Britam
With you every step of the way

Combined Application / Proposal Form

SECTION A

1. CUSTOMER INFORMATION

Full name: (Mr., Mrs, Miss, Dr., Prof., Eng., Pst., Other……..)

Nationality:

Identity document: National ID ☐ Passport ☐

National ID / Passport number:

Date of birth: D D M M Y Y Y Y

Marital status: Married ☐ Single ☐ Other _____

Gender Male ☐ Female ☐

Address Code: Town: _____

Physical Residential Address

PIN No. Mobile/Tel:

Email Address:

Nature of Business / Occupation / Profession _____

Source Of Funds _____

Are You a Politically Exposed Person? YES ☐ NO ☐

If yes, give brief details below: _____

SECTION B

PROPOSAL FORM FOR MOTOR PRIVATE INSURANCE

1. TYPE OF COVER

Tick as required

1. COMPREHENSIVE ☐ THIRD PARTY FIRE AND THEFT ☐ THIRD PARTY ONLY ☐

Registered Letters and Numbers	Make	Type of Body	Cubic capacity	Date of manuf.	Engine and Chassis Number	Seating capacity (inc. Driver)	Proposer's Estimate of the present value (inc. Accessories)

Windscreen Estimated Value _____

Radio Cassette Estimated Value _____ Logbook No. _____

Period of Insurance: From _____ To _____

2. THE DRIVER (Owner)

(a) How long have you been driving a Motor Vehicle? _____

(b) Do you, or any other person, who to your knowledge will drive, Suffer from defective hearing or from any physical infirmity?

_____

(c) Have you, or any other person, who to your knowledge will drive been convicted of any offence in connection with driving of any motor vehicle?

_____

(d) Date of issue of permanent driving license in Kenya and its expiry date _____

3. USE OF VEHICLE

(a) Will the vehicle(s) be used solely for social, domestic or pleasure purpose and by the insured and in connection with the insured's business and Profession?

YES ☐ NO ☐

If NO, Explain _____

(b) Will the vehicle(s) be used for Hire and Reward?

YES ☐ NO ☐

If YES explain _____

4. PREVIOUS EXPERIENCE

(a) Are you now, or have you been insured in respect of any motor vehicle? If so, please state name of Company or Underwriter

_____

(B) Has any Company or underwriter ever:-

(i) Declined your proposal? YES ☐ NO ☐

(ii) Required an increased premium or imposed special terms? YES ☐ NO ☐

(iii) Repudiated any claim? YES ☐ NO ☐

(iv) Cancelled your policy? YES ☐ NO ☐

(v) Refused to renew your policy? YES ☐ NO ☐

(c) Have you suffered any accidents or losses in connection with any motor vehicles owned? Or driven by you and/or by any other person who will regularly drive the vehicle(s) now proposed for Insurance ? If so, please give brief details _____

(d) Give record of accidents and or losses during the past three years with any motor vehicle owned and driven by you whether insured or uninsecured including any claims outstanding.

_____

TOTAL NUMBER OF ACCIDENTS AND LOSSES

Year	Total No. of Motor Vehicle owned by proposer	Total No. of Accidents and Losses		Damage to proposer's Motor Vehicles		Third party		Others	
				No.	Amount Shs.	No.	Amount Shs.	No.	Amount Shs.
			Paid Outstanding						
			Paid Outstanding						

5. ADDITIONAL BENEFITS

The following extensions are available on payment of additional premium

Please tick as appropriate.

(i) Political Violence and Terrorism YES ☐ NO ☐

(ii) Excess Protector YES ☐ NO ☐

(iii) Courtesy Car YES ☐ NO ☐

(iv) Is Duty Paid on the vehicle YES ☐ NO ☐

For Duty free vehicles:

1. The maximum indemnity payable by the insurer under this policy will be the sum insured less the applicable excess(es).
2. The estimate of value declared by the Insured for purpose of this insurance excludes the customs duty and other levies.
3. In the event of loss and/or damage arising out of an accident to the vehicle insured by the within mentioned policy and giving rise to a claim thereunder the amount of Company's liability shall be limited to the labour charges for repairs in Kenya plus the duty free costs of spare parts.
4. The insurer will maintain the right of selecting an assessor in case of damage to the motor vehicle and I can only use garages that are on the approved panel of the insurer.
5. In the event of a claim being treated as a total loss and/or constructive total loss, the onus of payment of duty if any levied by the Customs Authorities shall be on the Insured and the claim if any under the policy would be payable only after all formalities relating thereto are completed by the insured. Under no circumstances would the company be responsible for payment of any customs duty.

SECTION C

PROPOSAL FORM FOR MOTOR COMMERCIAL INSURANCE

1. TYPE OF COVER

Tick as required

1. COMPREHENSIVE ☐ THIRD PARTY FIRE AND THEFT ☐ THIRD PARTY ONLY ☐

Registered Letters and Numbers	Make	Type of Body	Cubic capacity	Date of manuf.	Engine and Chassis Number	Seating capacity (inc. Driver)	Proposer's Estimate of the present value (inc. Accessories)

Windscreen Estimated Value _____

Radio Cassette Estimated Value _____ Logbook No. _____

Period of Insurance: From _____ To _____

2. PREVIOUS EXPERIENCE

(a) Are you now, or have you been insured in respect of any motor vehicle? If so, please state name of Company or Underwriter

_____

(B) Has any Company or underwriter ever:-

(i) Declined your proposal? YES ☐ NO ☐

(ii) Required an increased premium or imposed special terms? YES ☐ NO ☐

(iii) Repudiated any claim? YES ☐ NO ☐

(iv) Cancelled your policy? YES ☐ NO ☐

(v) Refused to renew your policy? YES ☐ NO ☐

(c) Have you suffered any accidents or losses in connection with any motor vehicles owned? Or driven by you and/or by any other person who will regularly drive the vehicle(s) now proposed for Insurance? If so, please give brief details _____

(d) Give record of accidents and or losses during the past three years with any motor vehicle owned and driven by you whether insured or uninsecured including any claims outstanding.

_____

TOTAL NUMBER OF ACCIDENTS AND LOSSES

Year	Total No. of Motor Vehicle owned by proposer	Total No. of Accidents and Losses		Damage to proposer's Motor Vehicles		Third party		Others	
				No.	Amount Shs.	No.	Amount Shs.	No.	Amount Shs.
			Paid Outstanding						
			Paid Outstanding						

4. USE OF VEHICLE

(a) Give full particulars of all purposes for which Vehicle will be used

_____

_____

(b) If used for Carriage of insured's own goods, define the goods and their general nature?

_____

(c) Do you undertake carriage for other persons? YES ☐ NO ☐

(d). If any Passengers Carried:

(i) Are the passengers carried for hire or reward? YES ☐ NO ☐

(e) Will the vehicle be used in connection with the motor trade? YES ☐ NO ☐

(f) Will the vehicle be let to carry goods on hire? YES ☐ NO ☐

(g) Where is the vehicle garaged at night? (Street and City)

_____

5. THE DRIVER (Owner)

a) State the owner of the motor vehicle and in whose name it is registered _____

b) State the occupation / business or profession of the insured _____

(c) How long have you been driving a Motor Vehicle? _____

(d) Do you, or any other person, who to your knowledge will drive. Suffer from defective hearing or from any physical infirmity?

_____

(e) Have you, or any other person, who to your knowledge will drive been convicted of any offence in connection with driving of any motor vehicle?

_____

(f) Will anyone drive the car except yourself? _____

6. ADDITIONAL BENEFITS

The following extensions are available on payment of additional premium

Please tick as appropriate.

(i) Political Violence and Terrorism Risks YES ☐ NO ☐

(ii) Excess Protector YES ☐ NO ☐

SECTION D

PROPOSAL FORM FOR DOMESTIC PACKAGE INSURANCE

BUILDINGS

The following questions constitute a part of this proposal and must be answered fully.

1. Of what material is the dwelling constructed:

   (a) Walls _____

   (b) Roof _____

2. How many storeys has the dwelling? _____

3. How are the outbuildings (if any) constructed:

   (a) Walls? _____

   (b) Roof? _____

4. Is any business, profession or trade carried on in any portion of the premises of which the dwelling forms a part? If so, give particulars

   _____

5. Is the dwelling:

   (a) a private dwelling house/or _____

   (b) a self - contained flat with separate entrance exclusively under your control? _____

6. Do you own the premises? _____

7. Is the dwelling solely in your occupation? _____

8. If not solely in your occupation, do you let apartments or receive boarders? _____

9. Will the dwelling be left without an inhabitant for more than seven consecutive days? _____

   If so, state to what extent? _____

10. Are the Buildings in a good state of repair and will they be so maintained? _____

11. Has any Company or insurer, in respect of any of the contingencies to which the proposal applies?

   (a) Declined to insure you? _____

   (b) Required special terms? _____

   (c) Cancelled or refused to renew your Insurance? _____

   (d) Increased your premium at renewal? _____

12. Have you ever sustained any loss in respect of the contingencies to which the proposal applies? If so,

   Please give particulars _____

13. Are all the windows of the dwelling protected by iron bars? _____

14. (a) Do you employ a house servant? _____

   (b) Do you employ guards/watchmen? _____

15. Period of Insurance: From _____ To _____

PROPERTY TO BE INSURED

	OPTION 1	OPTION 2	OPTION 3
Total Value of all Buildings(Home Owner)	Up to KES 5,000,000/	From KES 5,000,001/ - to 7,500,000/-	From KES 7,500,001/ - to 15,000,000/-
Total Value of Your Portable Items & Household Contents	Up to KES 500,000/	From KES 500,001/ - to 1,000,000/-	From KES 1,000,001/ - to 1,500,000/-
Maximum Compensation for any Single Portable Items & Household Contents	KES 75,000/-	KES 100,000/-	KES 150,000/-
Number of Domestic Servants Covered (WIBA)	1	2	3
Occupiers/Owners Liability/	Kes 500,000/	Kes 1,000,000/	Kes 1,500,000/
Policy Excesses (Contents & All Risk)	10% OF CLAIM AMOUNT MINIMUM KES 5,000/-	10% OF CLAIM AMOUNT MINIMUM KES 5,000/-	10% OF CLAIM AMOUNT MINIMUM KES 5,000/-
Policy Excesses (WIBA)	**KES 5,000/ - EXCEPT FUNERAL EXPENSES**	**KES 5,000/ - EXCEPT FUNERAL EXPENSES**	**KES 5,000/ - EXCEPT FUNERAL EXPENSES**
Premiums inclusive of levies and stamp duty	**7,950** INCLUDING BUILDING	**13,626** INCLUDING BUILDING	**23,621** INCLUDING BUILDING
	3,681 EXCLUDING BUILDING	**7,222** EXCLUDING BUILDING	**10,813** EXCLUDING BUILDING

Indicate Selected Cover Option _____ Premium Amount (Kshs) _____

SECTION E

PROPOSAL FORM FOR PERSONAL ACCIDENT INSURANCE (With Political Violence and Terrorism extension)

1. Have you previously held a Personal Accident YES ☐ NO ☐
2. If yes, name of the insurer _____
3. Are you in good state of health and free from physical and mental defects or infirmity to the best of the proposer's knowledge and belief?
   YES ☐ NO ☐
4. If Not, Please give details _____
   _____
5. Give particulars of all accidents which you have suffered during the last three years.
   _____
   _____
   _____

Next of Kin:

Name _____ ID Number _____ Cell Phone Number _____

Relationship _____ Beneficiary _____

Name _____ ID Number _____ Cell Phone Number _____

Relationship _____

6. Britam PA Schedule and Benefits

	Option A	Option B	Option C	Option D	Option E	Option F	Option G	Option H	Option I
Death	250,000	500,000	8 00,000	1,000,000	2,000,000	3,000,000	5,000,000	8,000,000	10,000,000
Permanent total disability	250,000	500,000	800,000	1,000,000	2,000,000	3,000,000	5,000,000	8,000,000	10,000,000
Temporary total disability	1,500	5,000	8,000	10,000	12,500	15,000	30,000	40,000	50,000
Medical expenses	50,000	70,000	100,000	150,000	200,000	250,000	500,000	800,000	1,000,000
Funeral expenses	10,000	50,000	60,000	70,000	80,000	90,000	100,000	120,000	150,000
Annual premium per person including levies and stamp duty	**1,282**	1,773	2,682	3,591	5,605	**9,407**	13,144	20,180	25,228

Name of Spouse _____ **ID / Passport Number** _____ **Pin Number** _____

Cell Phone Number_____ **Date of Birth** _____ **Occupation** _____

Indicate Selected Cover Option For Insured _____ **Premium Amount (Kshs)** _____

Indicate Selected Cover Option For Spouse _____ **Premium Amount (Kshs)** _____

PLAN OF BENEFITS PER CHILD (BELOW 18 YEARS)

BENEFIT	PLAN 1	PLAN 2	PLAN 3	PLAN 4	PLAN 5
Accidental death	50,000	75,000	100,000	150,000	200,000
Permanent disabilities	50,000	100,000	200,000	400,000	500,000
Accidental dental treatment	10,000	10,000	10,000	10,000	10,000
Accident medical expenses	40,000	60,000	70,000	100,000	150,000
Artificial appliances	25,000	30,000	35,000	40,000	50,000
Funeral cover	20,000	20,000	20,000	20,000	20,000
Annual premium per child inclusive of levies and stamp duty	367	500	623	879	1,180

Child full name _____ **Date of birth** _____

Indicate Selected Cover Option For Child _____ **Premium Amount (Kshs)** _____

SECTION F

DETAILS OF THE INTERMEDIARY

1. Name of The Intermediary _____

2. Cell phone number of the intermediary _____

3. Email Address of the Intermediary _____

4. Intermediary Signature _____ Date _____

PREMIUM PAYMENT

Premium payment are to be paid directly to the company through the below noted options:

- Mpesa pay bill number – 111555 and the account number should either be your **car registration number** for Motor Vehicle Insurance or **National Identification (ID)** number for Domestic package and Personal Accident policies.
- Personal / Corporate Cheques
- Visa Cards
- Bank deposit / EFTs

Bank – Equity Bank Limited
Account Name – BRITAM
Branch – Community Corporate
Account Number – 0180293047296

NB : PLEASE NOTE THAT COMPANY SHALL NOT BE HELD LIABLE FOR ANY PREMIUM FUNDS PAID TO AN INDIVIDUAL WHETHER THIRD PARTY OR BRITAM STAFF.ALL PAYMENTS MUST BE ACCOMPANIED BY A RECEIPT.

DECLARATION

I/We desire to insure with the Britam General Insurance Company Kenya Limited, the motor vehicle or vehicles described in the above and I/We hereby warrant that the above statements and particulars are true, and I /We have not suppressed, mispresented or mis-stated any material fact and I/We agree that the declarations shall be the basis of the contract between Me/Us and the Company. I/We further agree that if this proposal in any particular is filled by any other person, such person shall be deemed my/our agent and not the agent of the Company. I/We further declare that I/We have read and understood all Particulars entered herein and I/We have signed this after verifying the same to be true and complete in all respects.

Insured name and signature _____ Date _____

No liability (except for the period stated in the Insurer's Official Cover Note) is undertaken until the Proposal is accepted by the Insurer and the premium paid.

DOCUMENTS TO ATTACH

1. Customer's Copy of National ID or Passport
2. Customer's copy of KRA PIN
3. Copy of Motor Vehicle Log Book (For Motor Vehicle Insurance).

Appendix 4: Motor Vehicle Insurance Claim Form

Britam
With you every step of the way

MOTOR ACCIDENT REPORT FORM

IMPORTANT NOTICE
ALL QUESTIONS ON THIS FORM MUST BE ANSWERED

(1) No liability under the policy is admitted by issue of this form.
(2) Neither owner nor driver must admit fault or liability.
(3) Do not answer communication about this accident, but send them to the insurers for consideration.
(4) Repairs must not be authorized without prior authority of the Insurers.

POLICY HOLDER	Name: _____ Telephone: _____ Address: _____ Business/Occupation: _____
POLICY	Number _____ Expiry Date _____ Name of Hire purchase or Finance Company _____
VEHICLE	Make & Model: _____ HP/CC: _____ Year of Manufacture: _____ Reg. No of Vehicle: _____ Carrying Capacity: _____ Reg. No of Trailer: _____ Carrying Capacity: _____ **Attach a copy of the Logbook and Driving Licence**
USE	State the exact purpose for which the vehicle was being used at the time of the accident: _____ _____
COMMERCIAL VEHICLES	Description of goods being carried: _____ Name of owner of goods _____ Was trailer attached _____ Weight of load on (a) _____ vehicle (b) Trailer's _____
DRIVER	Name: _____ Occupation: _____ Date of Birth: _____ Address: _____ Tel No: _____ Is he employed by you? _____ How long has he been in your service? _____ Was he driving with your permission? _____ How long has he been driving motor vehicles? _____ Was he in anyway to blame for the accident? _____ Did he admit liability? _____ Has he had any previous accident? If so, how many, and approximate date(s) _____ Has he any conviction for any offence in connection with any motor vehicle of any charges pending? _____ If so, give details including dates: _____ Does he hold a full or provisional licence to drive the vehicle? _____ If full, state exact date, driving test first passed: _____ Licence No.:_____ Does he own a motor vehicle? _____ If so give name and address of Insurer Driver's Policy No.:_____
ACCIDENT	Date:_____Time:_____AM/PM:_____Place:_____ Type of road surface:_____Visibility:_____Wet or Dry?_____ What lights were showing on your vehicle?_____ What warning did your driver give?_____ Estimated speed before accident:_____Weather Conditions:_____ Did Police take particulars?_____ If so, give Constable's No. and Station_____ To which police station was the accident reported?_____ **Attach copy of Notice of Intended Prosecution if any.**

Britam
With you every step of the way

MOTOR ACCIDENT REPORT FORM

IMPORTANT NOTICE
ALL QUESTIONS ON THIS FORM MUST BE ANSWERED

(1) No liability under the policy is admitted by issue of this form.
(2) Neither owner nor driver must admit fault or liability.
(3) Do not answer communication about this accident, but send them to the insurers for consideration.
(4) Repairs must not be authorized without prior authority of the Insurers.

POLICY HOLDER	Name: _____ Telephone: _____ Address: _____ Business/Occupation: _____
POLICY	Number _____ Expiry Date _____ Name of Hire purchase or Finance Company _____
VEHICLE	Make & Model: _____ HP/CC: _____ Year of Manufacture: _____ Reg. No of Vehicle: _____ Carrying Capacity: _____ Reg. No of Trailer: _____ Carrying Capacity: _____ **Attach a copy of the Logbook and Driving Licence**
USE	State the exact purpose for which the vehicle was being used at the time of the accident: _____ _____
COMMERCIAL VEHICLES	Description of goods being carried: _____ Name of owner of goods _____ Was trailer attached _____ Weight of load on (a) _____ vehicle (b) Trailer's _____
DRIVER	Name: _____ Occupation: _____ Date of Birth: _____ Address: _____ Tel No: _____ Is he employed by you? _____ How long has he been in your service? _____ Was he driving with your permission? _____ How long has he been driving motor vehicles? _____ Was he in anyway to blame for the accident? _____ Did he admit liability? _____ Has he had any previous accident? If so, how many, and approximate date(s) _____ Has he any conviction for any offence in connection with any motor vehicle of any charges pending? _____ If so, give details including dates: _____ Does he hold a full or provisional licence to drive the vehicle? _____ If full, state exact date, driving test first passed: _____ Licence No.:_____ Does he own a motor vehicle? _____ If so give name and address of Insurer Driver's Policy No.:_____
ACCIDENT	Date:_____Time:_____AM/PM:_____Place:_____ Type of road surface:_____Visibility:_____Wet or Dry?_____ What lights were showing on your vehicle?_____ What warning did your driver give?_____ Estimated speed before accident:_____Weather Conditions:_____ Did Police take particulars?_____ If so, give Constable's No. and Station_____ To which police station was the accident reported?_____ **Attach copy of Notice of Intended Prosecution if any.**

Britam
With you every step of the way

This part MUST be filled by the **insured,** the **driver** and any witness listed in the claim form. Reproduce this part where necessary.

STATEMENT FORM	
NAME:	ID NO:
POSTAL CONTACTS:	TEL. CONTACTS:
CELL NO:	EMAIL:
DATE OF ACCIDENT:	DATE RECORDED:
PLACE:	TIME:
CLASS OF PERSON: (Insured, Driver, Witness)	

Declaration: I declare the foregoing particulars to be a true account of the accident herein.

SIGNATURE …….. DATE:……..

Appendix 5: Models Training Source Code

```
#Program to detect fraudulent vehicle insurance claims using Machine Learning
#The necessary imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import joblib
from imblearn.over_sampling import SMOTE
import warnings
warnings.filterwarnings('ignore')
plt.style.use('ggplot')
smote=SMOTE()
#Reading of the insurance claims csv file
df = pd.read_csv('insurance_claims.csv')
#Display of the first 10 records in the dataset
df.head(n=10)
#Display of the number of claims in the dataset
df.shape
#Distribution of genuine and fraudulent claims in the dataset
df.groupby(['fraud_reported'])['fraud_reported'].count()
#Checking for duplicate claims
df.drop_duplicates(inplace = True)
df.shape
#Description of the data
df.describe()
df.info()
#Data Pre-processing
#Columns with missing values have been filled
df.isna().sum()
#Returning the number of unique values for each column
df.nunique()
#Checking for multicollinearity
plt.figure(figsize = (18, 12))
corr = df.corr()
mask = np.triu(np.ones_like(corr, dtype = bool))
sns.heatmap(data = corr, mask = mask, annot = True, fmt = '.2g', linewidth = 1)
plt.show()
#Splitting data into 80% training set and 20% testing set and balancing the data using SMOTE()
from sklearn.model_selection import train_test_split
#Checking whether is balanced or imbalanced
balance_data = True
#Data balancing using SMOTE()
if(balance_data):
    X,y=smote.fit_resample(X,y)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20)
X_train.head()
#Classification of the Models
#Imports on the models
```

```
from sklearn.svm import SVC #Support Vector Machine Model
from sklearn.naive_bayes import GaussianNB #Naive Bayes Model
from sklearn.linear_model import LogisticRegression #Logistic Regression Model
from sklearn.tree import DecisionTreeClassifier #Decision Tree Model
from sklearn.ensemble import RandomForestClassifier #Random Forest Model
from sklearn.ensemble import AdaBoostClassifier #AdaBoost Model
from xgboost import XGBClassifier #XGBoost Model
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

Appendix 6: Web Application Source Code

```
#The necessary imports
import pandas as pd
import numpy as np
import joblib
import sklearn
import streamlit as st
def convert_df(df):
   return df.to_csv().encode('utf-8')
st.title("Insurance Fraud Detection")
test_file = st.file_uploader("Choose a test file", type=["csv","xlsx"])
test_model=st.file_uploader("Choose model", type=["sav","plk"])
# If button is pressed
if st.button("Predict"):
   df = pd.read_csv(test_file)
   #Load the model from disk
   loaded_model = joblib.load(test_model)
   result=loaded_model.predict(X)
   result = pd.DataFrame(result, columns = ['results'])
   result["results"].replace(['Y', 'N'],
                  ["Fraudulent", "Genuine"], inplace=True)
   st.write(result)
   df["results"]=result
   download_csv = convert_df(df)
   st.download_button(
      label="Download File with predictions",
      data=download_csv,
      file_name='insurance_fraud_prediction.csv',
      mime='text/csv',
    )
```