KENYATTA UNIVERSITY

MOTOR INSURANCE FRAUD DETECTION USING NAÏVE BAYES CLASSIFIER

MUTWARUHIU AGNES	I162/0553/2018
GITONGA VINCENT	I162s/14192/2018
KAMAU DANIEL	I162/0537/2018
KIHARA EDMUNDS	I162/0538/2018
KIPKOECH EMMACULATE	I162/0555/2018

A research proposal submitted in partial fulfilment of the degree of Bachelor of Actuarial Science of Kenyatta University.

Department of Mathematics and Actuarial Science.

October 2022.

DECLARATION

We, the undersigned, declare that this project is our original work and has not been presented for an award of certificate, diploma, or degree in any university whatsoever.

Sign________________________Date....../....../.........
Mutwaruhiu Agnes

Sign ________________________Date....../....../.........
Gitonga Vincent

Sign________________________ Date....../....../.........
Kamau Daniel

Sign________________________ Date....../....../.........
Kihara Edmunds

Sign________________________ Date....../....../.........
Kipkoech Emmaculate

I confirm that this research project has been submitted for examination by students under my supervision.

Project Supervisor

Sign________________________ Date....../....../.........
Dr. Ananda Kube
Lecturer Kenyatta University.

DEDICATION

This project is dedicated to our parents, instructors, and colleagues for the support during the course of this study.

ACKNOWLEDGMENT

First, we wish to express sincere gratitude to our supervisor Dr. Ananda Kube who provided us with the needed guidance, encouragements, and motivation. It was truly a pleasure working under his supervision. Thanks are extended to our family, friends and colleagues who took time to support us during the course of this study.

ABBREVIATIONS AND ACRONYMS

IFIU – Insurance Fraud Investigation Unit

IRA – Insurance Regulatory Authority

AKI – Association of Kenyan Insurers

K-NN – K- Nearest Neighbor

SVM- Support Vector Machine

WHO – World Health Organization?

NHCAA- National Healthcare Anti-Fraud Association

ABSTRACT

Insurance fraud is a deception that is perpetrated by or against an insurance agent or company for financial gain. It is a significant and costly problem for both the insured and the insurers. Between the year 2018 and 2019, IRA released a report stating that insurance fraud had risen 24% with an estimated loss of Ksh 52.87 million. Amidst the covid-19 pandemic, IRA established that there was an increase of insurance fraud from Ksh 19.2 million to Ksh 28.4 million. Motor insurance fraud and healthcare fraud are common types of insurance fraud. These types of insurance fraud usually occur when the information is concealed or misrepresented with the intention of getting benefits. In fact, the NHCAA estimated that a conservative approximate of $68 billion is lost each year in healthcare fraud. This study focuses on detecting motor insurance fraudulent claims using Naïve bayes classifier, a supervised machine learning algorithm. Naïve bayes classifier simplifies the problem by assuming that each attribute contributes equally to the prediction regardless of their relative significance. In 2019, Tope conducted a study where he used Naïve bayes classifier to detect spam emails. He concluded that the algorithm used, produced accurate results classifying the spam and the legitimate emails. In conclusion, this study attempts to develop a model using Naïve bayes classifier, test the properties of the model and fit a motor insurance claims dataset into the model so that the model detects the fraudulent claims in the dataset.

CONTENTS

List of Tables

CHAPTER 1

1 INTRODUCTION

1.1 Background of the Study

Detection of insurance fraud is a great concern for many insurance companies in Kenya. However, this has not been addressed in depth. A lot of money is lost each year in insurance companies because of fraud. Bismart insurance limited (2019) estimated that 30% to 40% of their motor insurance claims were fraudulent. The insurance fraud investigation unit (IFIU) detected 83 insurance fraud cases in 2019 worth ksh 386.34 million.

Insurance Regulatory Authority (2020) observed that the covid-19 Pandemic affected insurers' day in day operations hence insurers were forced to alter their original plans to digitally up scale in the coming years. They also reported that insurance fraud rose from ksh19.2 million to 28.4 million amidst the covid-19 pandemic. The authority also concluded that the losses incurred due to the fraud particularly in motor and health care insurance have strained the insurance industry.

Furthermore, the insurance fraud investigation unit found an increase in fraud cases from 83 in 2019 to 127 in 2020. In 2019, AKI launched a mobile application for filing claims to reduce motor insurance fraud and cases where drivers would connive with the authorities to change the state of the accident.

An investigation done by the Coalition Against Insurance Fraud (2006) revealed that the United States lost roughly 80 billion dollars to insurance fraud. Insurance fraud is not only committed in the auto insurance sector, but it also happens in other insurance sectors. Feldman (2001) stated that one of the main reasons why health care fraud is common is that almost all the parties involved gets the benefits. They do this by intentionally not paying claims and clearing them from their systems, denying and cancelling coverage, and the blatant underpayment to physicians and hospitals beneath what are normal fees for care they provide.

Insurance companies in Kenya have not been able to address such issues and are looking for ways to detect fraud in insurance claims. Some of the methods that motor insurance companies in Kenya have been using to detect fraud in insurance include the use of services of some intelligence officers who have been trained to look for areas in which fraud can and have been committed. Jubilee insurance has established a full-fledged forensic unit with highly trained investigators to deal with this menace. They go through previous claims and search for any suspicious patterns as used by a risk assessor. The insurance company uses a claims manual that is filled with all the terms and conditions on how to settle the claim. These manuals are revised from time to time due to changes in trends and techniques used by fraudsters. Insurance Regulatory Authority (2011) set up an insurance fraud investigation unit to help curb insurance fraud in Kenya.

Insurance companies also conduct a physical inspection of the area, the vehicle, and details of the accident. They have tools like GPS which can assist in verifying the location. Association of

Kenya insurers (2018) rolled out an integrated Motor Database System (IMIDS) that enabled the insurance sector to save over forty million in 2019. These methods, however, are not highly effective and do not provide insurance companies with enough information they require to solve the problem. However, there is no denying that this database has been useful in detecting fraud.

In the recent past, methods of handling claims have progressively been exhibited to automatically conduct audits and claim data analysis. The system has been modelled to determine the areas with insufficient data input and claims that match. Pawar (2016) established the ability of these systems to discover fraud is typically minimal because it mainly relies on preestablished procedures described by the fraud experts.   Based on previous experience and his knowledge of data, Mukherjee *et al*. (2015) concluded that we can connect things occurring around us or sometimes even detect something out of the data at hand. He suggested that prompt detection and prevention of fraud will go a long way by providing outstanding cost reductions to insurers there by curtailment to the increasing cost of premium.

Helbock (2018) argued that most car accident fraud occurs when we have the staged accidents, here the victim of the accident and the driver are accomplices. In larger schemes the staging involves other players such as witnesses and insurance investigators whereby the value of the vehicle that caused the accident and the one involved is exaggerated and so is the cost to repair the damages which means the Insurance will have to write off the vehicle and the value of the vehicles is then paid. The money is used to repair the vehicles and the cycle can start all over again.

In addition, Car owners report a small accident, and the insurance company compensates for an estimate of what the damages, which often is usually on the higher end. Helbock (2018) also established that car damages are the most generic form of insurance fraud, and some commit them without knowing. Another instance of car fraud occurs in fake car thefts, The stolen car insurance swindle is used in two ways. The first way is when the owner strips down the car and sells it for body parts and report it as stolen. The body parts shop tends to be in on the scheme. The second way is the owner may decide to hide the car and declare it as stolen.

According to Ngosiah (2012), Public Service Vehicles (PSV) industry is one of most common victims of fraudulent activities. Njenga & Osiemo (2013) also observed that vehicles that have no faults were fitted with defective panels to deceive that they engaged in an accident so that the insured can file a claim. Moreover, Sybase (2012) argued that many of these incidents occurred where fraud was stage-managed using expensive cars which may cause fake deaths claims to be lodged and vehicles written off. However, studies have been done on several models used to detect fraud. This project will utilize the naïve bayes classifier which according to Clifton *et al.* (2010), is faster compared to other complex algorithms such as ANNs in classification. Tope (2019) also used naïve bayes classifier to filter out spam emails and concluded there was a higher accuracy and low error rate compared to other machine learning algorithms therefore this is an ideal method for fraud detection in insurance claims.

1.2 Problem Statement

Financial fraud is progressively becoming a fundamental problem in the insurance and the financial sector. The Insurance and Regulatory Authority (IRA) released a report which showed that the level of insurance fraud rose by 24% between the year 2018 and 2019. A lot of money is lost each year in the insurance industry mainly because of insurance fraud. In fact, the insurance fraud investigation unit (2019) detected eighty-three fraud cases worth 386.34 million Kenyan shillings. The Insurance Regulatory Authority (2020) further estimated that insurance fraud increased from ksh19.2 million to 28.4 million amidst the covid-19 pandemic. Healthcare fraud is a common in the insurance sector. It occurs when information is concealed or misrepresented with the intention of getting the health care benefits. In fact, the National Health Care Anti-Fraud Association (2007) established that a conservative estimate of $68 billion is lost each year due to fraud in healthcare.

This study proposes an approach to motor insurance fraud detection that uses supervised machine learning algorithm called naïve Bayes, a generic classification algorithm. Naïve Bayes classifier simplifies the problem by assuming that each feature contributes equally to the prediction regardless of their relative levels of significance. This method assumes that, in general, if someone commits a crime once, it is more likely than not they will commit it again. According to a study conducted by Tope (2019), this algorithm produced accurate results in detecting the spam emails. The classification rules of the email messages were used to train the model. Therefore, this project attempts to achieve more accurate results when detecting the fraudulent claims.

1.3 Objectives of the Study

1.3.1 General Objectives

The main objective of this study is to develop a Naïve bayes model that accepts motor insurance claims data and detects the fraudulent claims.

1.3.2 Specific Objectives

i. To develop a model that detects fraudulent claims ii.

ii. To assess the properties of the developed model

iii. To fit the insurance claims dataset into the developed model

1.4 Significance of the Study

This study will advance the motor fraud detection system minimizing the cost of paying off fraudulent claims and the cost of investigation. In addition, it will improve the efficiency of handling claims as it will be able to differentiate between valid and fraudulent claims. Therefore, the claim unit will settle the valid claims in appropriate time resulting in consumer satisfaction. The model can easily be incorporated into existing system of an insurance company. This study will also be helpful in actuarial work in pricing premiums and educating the underwriting unit on the vital information they should acquire when writing new business. We invite scholars and institutions to use this study as a basis of research for further studies.

CHAPTER 2

2 LITERATURE REVIEW

Albrecht and Zimbelman (2011) defined fraud as perpetrating deception with the intention of gaining benefits by misrepresenting the truth in an unfair way. According to Tseng and Kuo (2014), motor insurance is more susceptible to fraud compared to other lines of business. Fraud incidents may include concealing or misrepresenting facts to lodge a claim, deliberately staging and exaggerating an accident's losses and backdating a policy. Such behaviors are often motivated by factors such as high inflation, moral hazard, and extremely high interest rates.

A study conducted by Burri *et al.* (2019) provided a systematic description of machine learning algorithms and how they could be applied in the insurance industry. The study also gave an in-depth analysis of predictive models and their application in understanding claim costs. During the process of introducing diverse ways of implementing machine learning in the insurance setup, the research faced a lot of challenges such as the inability to obtain the right data, lack of data security and a higher requirement for training. The study also gave a little attention to the challenges of obtaining suitable data and ensuring the security of the obtained data when applying machine learning in insurance.

Palacio (2019) conducted a study on regression and statistical methods. The study focused on detection of fraud using semi-supervised machine learning. The method made use of the mini-batch K-Means and stratified 5-Fold cross-validation. The method produced good and reliable results; however, it faced some limitations in that it had to be calibrated each time predictors changed which meant that the model had to be retrained through dynamic learning. The study further failed to show that using random indicators produced a standard fraud detection model that could be shared across the country. The calculations made with data from the industry produced the parameters that were used. This made the model irrelevant to insurers who may have wanted to use it for their benefit.

Zhou *et al.* (2011) conducted a study and concluded that most fraud detection systems use at least one supervised learning method and employ unsupervised and semi-supervised methods also. Their study showed that those techniques could be used alone or as a combination to build more accurate classifiers and that those methods were relatively better in fraud detection and even credit scoring. The authors observed that fraud-related data was not abundant enough for the investigators to assess and train their models and that complex financial scenarios were impossible to represent. Further, they mentioned that detection of fraud and data-mining-based credit scoring were subject to the same classification-related issues such as parameter selection and hyper parameter tuning. However, they explained that for fraud detection to constantly evolve, it had to depend on the in which it is applied.

Ongati & Lawrence (2019) conducted research showing that the adaptation of decentralized training of models in the medical industry was an effective method for disease diagnosis while at the same time the privacy of customer's data. They discussed an algorithm to train a shared using distributed deep neutral networks. Their study adopted an approach of sharing learnt representation instead of sharing raw data. Their model, however, faced challenges in having a central coordinator who could break the whole network when down.

The studies by Liu *et al.* (2019) assumed that domains would have the same number of samples. That assumption however was not practical across most industries like insurance which was a challenge in their research. The study also required that all participating parties be online during the training which was also a major limitation to their research since different entities could encounter different complexities to be online during the training. They also faced the problem of tightly coupled dependencies. Also, the limitation of prediction due to the partial models that were stored on client nodes.

According to a study done by Sadiq (2019), he published an article called "Analysis on credit card fraud Detection Methods," which was talking about Bayesian networks. In this technique, two Bayesian networks are constructed. One network model the behavior of a legitimate user while the other models the behavior of the fraud. The way these two networks are set up is different as the fraud network uses expert knowledge while the legitimate user network uses data from the nonfraudulent user for its set up.

The aim of developing this network is to determine the conditional probabilities using an acyclic graph. This Bayesian networks, however, have limitations as there is difficulty in the conversion of expert knowledge into probability distributions.

Cooper (2003) proposed a methodology to be used by insurance companies to detect frauds committed by providers and patients. The researcher adopted a supervised learning algorithm called the decision tree to be used in detecting fraud in the health sector. The algorithm was broken down into stages and the first three stages were aimed at detecting abnormalities among services, providers and claim amounts and the fourth stage was used to integrate the information derived from the three stages to form a risk measure which formed the basis upon which a decision tree method was used to compute risk threshold values. Comparison of the risk value obtained in stage four with the risk threshold value enabled them to decide whether a claim is fraudulent or not. This method however had a challenge as in the real world, insurance data was difficult to interpret because of the challenge of overly complex decision tress with rules.

Permsirivisarn (2001) did a study on fraudulent claim patterns exhibited by four groups of individuals, Insurance company employees, the beneficiaries, hospital personnel, and police officers. The study revealed that a total of seventy patterns was discovered from the 4 groups. It went further to suggest that improvement should be done on both the individual group and general matters. His study was the first to take a deeper look at patterns of the insured in relation to making fraudulent claims. This proved to be extremely useful to other researchers, the claims office and other related insurance service sectors.

Kamdee (2005) studied the distinct types of motor vehicle insurance, the different defraud methods of insured a well as obstacles in the law enforcement when implementing the criminal and concerning nonlife defraud penalty. His study showed that fault according to the criminal code concerning the non-life insurance defraud penalty cannot be considered as unlawful as this provision is only applicable with the main non-life insurance, which is the Motor Insurance type 1.

Kosaisook (2007) studied the problems of fraud in non-life insurance business of Thailand to identify the problem arising when enforcing the law by comparing it with measures of guidelines from other countries. And, see how provision on non-life insurance fraud do create efficiency on law enforcement and benefits associated with it. His study suggested that Thailand should have specific provisions enforced for the case of non-life insurance fraud, which involves combining into one law the three types of fraud so that law enforcement on this matter is more efficient.

Chantachit (2008) studied meaning and concept on economic crime. His study found that most of automobile insurance defraud which are considered under the economic crime for instance the actions in work administration, are usually by authority or those individuals who hold high ranking positions.

Furthermore, the Association of British Insurers (2009) conducted research on General Insurance Claim Fraud. The study revealed that the insurance industry detects more of the fraud that is attempted. This has been attributed to improvements in data sharing through the Insurance Fraud bureau together with a more focused approach to fraud detection across insurers which have contributed to the improvements.

SBD (2011) studied the status of Car Theft and Insurance in Thailand. This study suggested that the Thailand Insurance sector has been registering rising costs which are driven up by the size and quantity of claims which are related to accident repair. From this study the result showed that from the Fraudulent Insurance Claims, actual vehicle theft accounts for only 3% of the insurance claims in Thailand. Several factors attributed to the rise of cost of insurance and a high amount of fraudulent insurance claims appears to be one of the factors.

Chuleekorn (2015) drafted an article on the role of using Data Analytics for fraud detection in insurance. There are many techniques that are used to perform analytics to detect insurance fraud. From this study she states five methods popularly used: Investigation on Business Rules, Anomaly Detection, Predictive Modelling/ Advanced social network Analysis and Text mining. Based on these concepts mentioned by the Thai researcher, she found that the criminal code is more focused on insurance fraud.

Busch (2008) published a book on healthcare fraud which mainly focused on Detection and auditing. In this book, fraud is defined as an intentional and knowing execution of scheme to swindle the benefits of the medical care program. The National Healthcare Anti-Fraud Association (NHCAA) defined fraud as intentional misrepresentation that an entity does knowing that it could result in some unapproved benefits to the individual or according to BlueCross (2016), to the entity

Kowshalya and Nandini (2018) used a synthetic dataset in their case study to predict the fraudulent claims in automobile insurance.in this thesis, mock insurance claim dataset is created based on the previous case studies since the real dataset is not easily accessible. In the mock dataset, two categories were established, one for accident claim and the other for theft claims. Classification algorithms such as J48, Random Forest and Naïve Bayes are used to predict the fraudulent claims. To enhance the accuracy of the model, preprocessing of data is performed.

Narvaez *et al.* (2019) proposed a study to predict motor insurance claims using telematics data. She compared the predictability of both XGBoost and logistic regression to see which one was more accurate in predicting the claim amount. In this study, she used telematics data which has more vehicle information. She concluded that without adjusting the model, XGBoost performed better than Logistic regression in training sample but relatively poor in testing sample. However, with adjustments, both XGBoost and Logistic regression predictive performance was the same. Clearly, XGBoost needs to be adjusted to achieve a higher predictive performance.

In another study, Verbelen *et al.* (2018) unraveled the predictive power in pricing car insurance using telematics data. In this paper the predictive power of telematics data is observed using the Belgian telematics dataset. Before, data that was provided by the insured was used to price premiums but in this article, a black box is integrated into the policyholder's vehicle to keep track of the policyholder's driving behavior for instance the speed. This project was main focused on the young adults, and it seeks to discover more variables that influence the predictive capability.

Furthermore, Pozzolo (2010) compared different data mining techniques used to predict the insurance claim amount. In this research, the prediction is fully dependent on the vehicle of the customer. The insurance claim amount is predicted using Naïve Bayes, Random Forest, K Nearest Neighbors (KNN), Decision Tree Support Vector Machine (SVM), Unsupervised Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) data mining techniques. All these methods were compared against each other.

Akomea *et al.* (2016) observed the effects and the level of insurance fraud on the performance of insurance companies in Ghana. The study examined the measures and causes that can be integrated to curb fraud in insurance. The study focused on both secondary and primary data collected from almost forty insurance companies that were analyzed using regression analysis. This method’s results concluded that there was a negative and meaningful relationship between claims and insurance fraud on the performance of insurance firms in Ghana. The study further examined that some of the main causes of insurance fraud in insurance firms were, poor training of insurance brokers, employees' salary, misrepresentation of documents and poor internal controls. The study recommended the need to sufficiently train insurance brokers, to establish effective internal controls and integrate efficient technology.

Moronge and Kuria (2014) presented the effects of insurance fraud control system on the growth of the insurance sector in Kenya. The study established a descriptive survey and collected data from risk managers of almost fifty insurance companies in Kenya. The census method was used. Data was collected using questionnaires and the outcome showed that 40% of the insurance claims paid off by insurance companies were fraudulent. It was also discovered that the supervisor does not help the in controlling the fraudulent activities and that there was no effective technology that would aid in curbing these fraudulent activities. The study suggested that the supervisor should employ fraud control systems when settling claims.

Previous works by Dhieb *et al.* (2019) proves that you can train fraud detection models and still preserve the client's privacy. In their study, they suggest an extreme gradient boosting machine learning algorithm to detect fraud. The research uses categorization to convert most of the attributes of the customer's claims into a binary format. The research also uses generalization to change the low-level data to high level data in order to hide the sensitive data such as the client's information. The model had 99.25% accuracy however since the data is centralized, it is biased to the insurers providing the data. Therefore, the model does not represent the real-life fraud incidents in the industry.

Tesfaye (2017) conducted a study on Nyala Insurance where he attempted to develop a predictive model for insurance risk assessment that specifically detected behavior patterns to identify the risky customers in motor insurance. In the study he used insured vehicles and customers' information and used the neural network and decision tree to develop the predictive model. The dataset has 1,056 records of which 10% of it was used for testing and 90% for training. Using decision tree, 95.69% of the data set was classified accurately, and the classification accuracy for high, medium, and minimal risk groups were 92.96%, 94.12%, and 98.15% respectively whereas the neural network model classified the high-risk policies with accuracy of 92.24%, medium-risk policies with 76.47% accuracy and low-risk policies were classified with 98.15% accuracy. In the study a pattern was found between the two models that some policies that were incorrectly classified by decision tree were correctly classified by neural networks, and vice versa.

Therefore, he concluded that a hybrid of the two models may lead to a better classification accuracy.

In this project, we will develop a model that will detect motor insurance fraud by assessing the behavior pattern of the customer and identifying the fraudulent claims. We are going to use Naïve bayes classifier to develop this model. The insurance claim dataset to be used will be obtained from Kaggle. Properties of the model will be assessed to assess its ability to predict the fraudulent claims. Lastly, we will fit the insurance claims data into the developed model and document the results.

CHAPTER 3

3 RESEARCH METHODOLOGY

3.1 Introduction

The naïve bayes classifiers is a probabilistic classifier that can be used in several classification tasks this means that naïve bayes predicts based on the probability of an object. It is referred to as 'naïve' because it assumes that the attributes that go into the model are independent hence impractical to real data. Nevertheless, it is one of the simplest yet powerful classification algorithms which helps in building fast machine learning models that can make quick predictions.

3.2 Naïve Bayes

Let $\mathbf{X} = (X_1, \dots , X_n)$ be a random vector where each random variable X_i represents a feature of a claim i.e., previous number of claims, age of policy holder, the presence of a police report, etc. and can be discrete or continuous. In addition, let $\mathbf{x}$ represent an observation of the random vector $\mathbf{X}$ that can also be considered as an input of a model. Furthermore, let t represent the target variable. An observation can belong to one of the K classes, denoted by the class labels C_0, C_1, …,C_{K-1} . C denotes a random variable that represents a class label C_k.The aim is to develop a classifier that allocates correct observations to their correct classes using a classification rule.

The posterior probability is denoted by $P(C = C_k|\mathbf{X} = \mathbf{x})$.

In this study, the goal is to classify fraudulent and legitimate claims correctly. Therefore, to make model more ideal for detecting fraud, we will reduce the scenario to a two-class problem, K=2, where the target variable t will correspond to the non-fraudulent class C_0 when t=0 and fraudulent class C_1 when t=1. Moreover, we are going to assume that our dataset has m observations. A portion of this dataset that contains M observations will be referred to as the training set and it will be used to fit the model to the dataset. The effectiveness of the model will be assessed using the remaining m-M observations, these observations are referred to as the testing set.

3.2.1 Estimation

In this part, we will assume that the variables $X_1,\ldots,Xn$ are conditionally independent of each other. Under this assumption, Bayes' theorem can be applied to find

$$P(C_k|\mathbf{X} = \mathbf{x}) = \frac{q(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(\mathbf{x})}$$

The training aspect of the Naïve Bayes classifier corresponds to the estimation of maximum likelihood of the necessary parameters. In this case, the parameters required are the univariate marginal class probabilities for each observation, the prior probabilities for each class and each attribute of the training set D. The structure of the class probabilities is dependent on the type of variables forming the training set. Since the available data comprises of only discrete variables, it is sufficient to consider the cases when the variables are categorical, i.e., where a variable $X_i \in \{1, \ldots, N\}$.

Consider the case when a variable X_i is a categorical variable such that it can take $N > 2$ possible values, i.e., $X_i \in \{1, \ldots, N\}$,then, the probability that the observed variable x_i takes a value n from the N possible values conditioned on the class, is represented as

$$P(X_i = n|t) = \pi_{it}^{(n)}$$

Such that

$$\sum\nolimits_{n=1}^{N} P(X_i = n|t) = 1$$

Suppose that there are $y_2 \leq y$ categorical variables in a certain training set $D = \{(\mathbf{x}^{(m)}, t^{(m)})\}_{m=1}^{M}$, where *n* is the total number of attributes in the set. Given the target variable *t*, each observation is independently and identically distributed so that the likelihood becomes:

$$L = \prod_{m=1}^{M} P(\mathbf{x}^{(p)}|t^{(p)}) = \prod_{m=1}^{M}\prod_{i=1}^{n_2}\prod_{n=1}^{N}\prod_{t=0}^{1} (\pi_{it}^{(n)})^{\mathbb{I}\{x_i^{(m)}=n\}\mathbb{I}\{t^{(m)}=t\}}$$

where $\mathbb{I}$ is the indicator function, which equals to 1 if the predicate holds, otherwise, 0.

The log-likelihood is

$$l = \sum_{m=1}^{M}\sum_{i=1}^{n_2}\sum_{n=1}^{N}\sum_{t=0}^{1} \mathbb{I}\{x_i^{(m)} = n\}\mathbb{I}\{t^{(m)} = t\} \ln \pi_{it}^{(n)}$$

To get the required estimates for $\pi_{it}{}^{(n)}$, we will use the Lagrange multiplier method as suggested by Barber (2010) to check that probabilities add up to one such that

$$l = \sum_{m=1}^{M}\sum_{i1}^{n_2}\sum_{n=1}^{N}\sum_{t=0}^{1} \mathbb{I}\{x_i^{(m)} = n\}\mathbb{I}\{t^{(m)} = t\} \ln \pi_{it}^{(n)} + \sum_{t=0}^{1}\sum_{i=1}^{n_2} \lambda_{it}(1 - \sum_{n=1}^{N} \pi_{it}^{(n)})$$

Differentiating w.r.t $\pi_{it}{}^{(n)}$ and equating to zero, we get

$$\lambda_{it} = \sum_{m=1}^{M} \frac{\mathbb{I}\{x_i^{(m)} = n\}\mathbb{I}\{t^{(m)} = t\}}{\pi_{it}^{(n)}}$$

Thus, the maximum likelihood

$$\hat{\pi}_{it}^{(n)} = \frac{\sum_{m} \mathbb{I}\{x_i^{(m)} = n\}\mathbb{I}\{t^{(m)} = t\}}{\sum_{n\prime,m\prime} \mathbb{I}\{x_i^{(m\prime)} = n'\}\mathbb{I}\{t^{(m\prime)} = t\}}$$

which coincides with the number of times the categorical variable X_i takes the n*th* value of the *N* possible values for a given class.

Besides, it can be proved that for binary attributes, $x_i^{(m)} \in (0,1)$, we have the estimator for π_{it} as:

$$\hat{\pi}_{it} = \hat{P}(X_i = 1|t) = \frac{\sum_{m} \mathbb{I}\{x_i^{(m)} = 1, t^{(m)} = t\}}{\sum_{m}\{\mathbb{I}\{x_i^{(m)} = 0, t^{(m)} = t\} + \mathbb{I}\{x_i^{(m)} = 1, t^{(m)} = t\}\}}$$

which correlates to the number of times $X_i = 1$ for a certain value of *t* divided by the total number of observations in the class C_k. The estimator for the prior probabilities is as follows:

$$\hat{q}_k = \frac{M_k}{M}$$

which also correlates with the ratio of the total number of observations in the training set *M* and the number of observations in class C_k, represented by M_k. Practically, this is determined by counting the number of examples from each class and observing the target value $t^{(n)}$ for each observation, therefore the target variable is employed in the derivation. Consider the case when X_i is a binary variable. Let the observations of the training dataset have a value of 0 for a specific property for the class C_0. This suggests that the probability is conditional on the class i.e., $P(X_i = 1|C_0) = 0$. When integrating this probability into the Naïve Bayes model, this would suggest that whenever a new observation **x** has a value of 1 for that attribute, then the probability of the observation belonging to class C_0 is 0, while the probability of the observation belonging to C_1 is 1. This prediction is likely to be inaccurate. Similarly, for the case when the attribute X_i can take 2 or more possible values. The more possible values the attribute takes, the more likely an issue can occur since certain possibilities are more likely to never occur in the training set.

The most suitable way of dealing with this issue is referred to as the Laplacian estimator or Laplace correction. It is assumed that for the training set D, the number of observations M is large enough such that adding a small number of cases for each level of an attribute to both classes would be insignificant. Let $b \in$ M denote the number of cases included in an attribute class pair. It is significant to note that b cases shall be included in each possible value of the attribute, therefore for binary attributes, $2b$ must be added to the denominator such that the Laplace estimator becomes:

$$\hat{\pi}_{it}^{L} = \frac{\sum_{m} \mathbb{I}\{x_i^{(m)} = 1, t^{(m)} = t\} + b}{\sum_{n} \{\mathbb{I}\{\, x_i^{(m)} = 0, t^{(m)} = t\} + \mathbb{I}\{x_i^{(m)} = 1, t^{(m)} = t\}\} + 2b}$$

For categorical attributes, recall that there are $N > 2$ possible values such that Nb shall be added to the denominator. Hence the Laplace estimator becomes:

$$\hat{\pi}_{it}^{(n)L} = \frac{\sum_{m} \mathbb{I}\{x_i^{(m)} = n\}\mathbb{I}\{t^{(m)} = t\} + b}{\sum_{n',m'} \mathbb{I}\{x_i^{(m')} = n'\}\mathbb{I}\{t^{(m')} = t\} + Nb}$$

Normally, b is taken to be 1 such that there is at least one count for each attribute-class pair in the training set. This e of obtaining eliminates the possibility of getting zero values for the marginal probabilities. Since under the Naïve Bayes assumption the marginal probabilities are multiplied, this would also destroy the other marginal probabilities, so it is an issue that needs to be dealt with, especially when the distribution of a variable is skewed.

CHAPTER 4

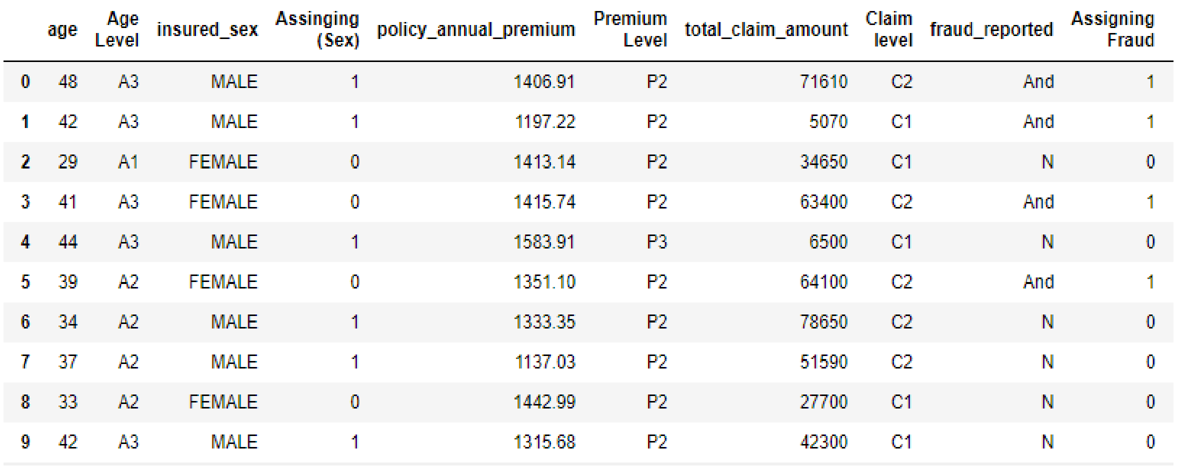
4 DATA ANALYSIS AND DATA REPRESENTATION

4.1 Data set

The dataset was obtained from Kaggle. It comprises of 1000 Datasets and 40 categorical variables, and one target variable fraudulent claims reported. The dataset was randomly divided into the testing and training dataset. The training data set consists of 700 claims while the test data set consists of 300 claims. We identified that some columns had many, over 900 distinct categories and therefore they cannot be converted to numeric. These columns include the policy number, policy bind date, insured location, incident hour of the day, insured zip, and the incident date. Four categorical variables were taken and one target variable to test our hypothesis. The four categorical variables are: Age of the insured, Insured sex, Policy annual premium and Total claim amount. The target variable is Fraud reported.

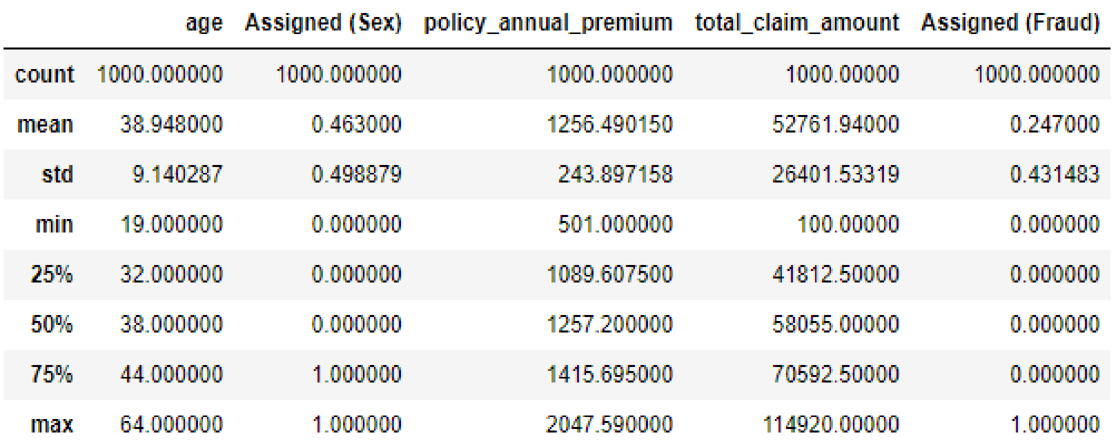
4.2 Basic Exploratory Data Analysis (EDA)

Under Exploratory Data Analysis the data was analyzed with the goal of understanding the mean, standard deviation, minimum and maximum, interquartile ranges and the correlation coefficients of the attributes chosen. The graphs of how the data behaves was also visualized.



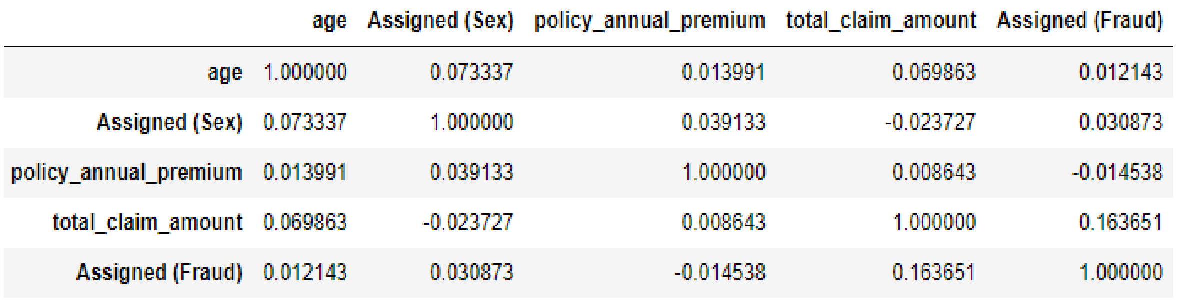
	age	Age Level	insured_sex	Assinging (Sex)	policy_annual_premium	Premium Level	total_claim_amount	Claim level	fraud_reported	Assigning Fraud
0	48	A3	MALE	1	1406.91	P2	71610	C2	And	1
1	42	A3	MALE	1	1197.22	P2	5070	C1	And	1
2	29	A1	FEMALE	0	1413.14	P2	34650	C1	N	0
3	41	A3	FEMALE	0	1415.74	P2	63400	C2	And	1
4	44	A3	MALE	1	1583.91	P3	6500	C1	N	0
5	39	A2	FEMALE	0	1351.10	P2	64100	C2	And	1
6	34	A2	MALE	1	1333.35	P2	78650	C2	N	0
7	37	A2	MALE	1	1137.03	P2	51590	C2	N	0
8	33	A2	FEMALE	0	1442.99	P2	27700	C1	N	0
9	42	A3	MALE	1	1315.68	P2	42300	C1	N	0

Table 1 First 10 rows of Data.



	age	Assigned (Sex)	policy_annual_premium	total_claim_amount	Assigned (Fraud)
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000
mean	38.948000	0.463000	1256.490150	52761.94000	0.247000
std	9.140287	0.498879	243.897158	26401.53319	0.431483
min	19.000000	0.000000	501.000000	100.00000	0.000000
25%	32.000000	0.000000	1089.607500	41812.50000	0.000000
50%	38.000000	0.000000	1257.200000	58055.00000	0.000000
75%	44.000000	1.000000	1415.695000	70592.50000	0.000000
max	64.000000	1.000000	2047.590000	114920.00000	1.000000

Table 2 Count, mean, standard deviation, minimum and maximum, and interquartile ranges.



	age	Assigned (Sex)	policy_annual_premium	total_claim_amount	Assigned (Fraud)
age	1.000000	0.073337	0.013991	0.069863	0.012143
Assigned (Sex)	0.073337	1.000000	0.039133	-0.023727	0.030873
policy_annual_premium	0.013991	0.039133	1.000000	0.008643	-0.014538
total_claim_amount	0.069863	-0.023727	0.008643	1.000000	0.163651
Assigned (Fraud)	0.012143	0.030873	-0.014538	0.163651	1.000000

Table 3 The correlation coefficient of the attributes

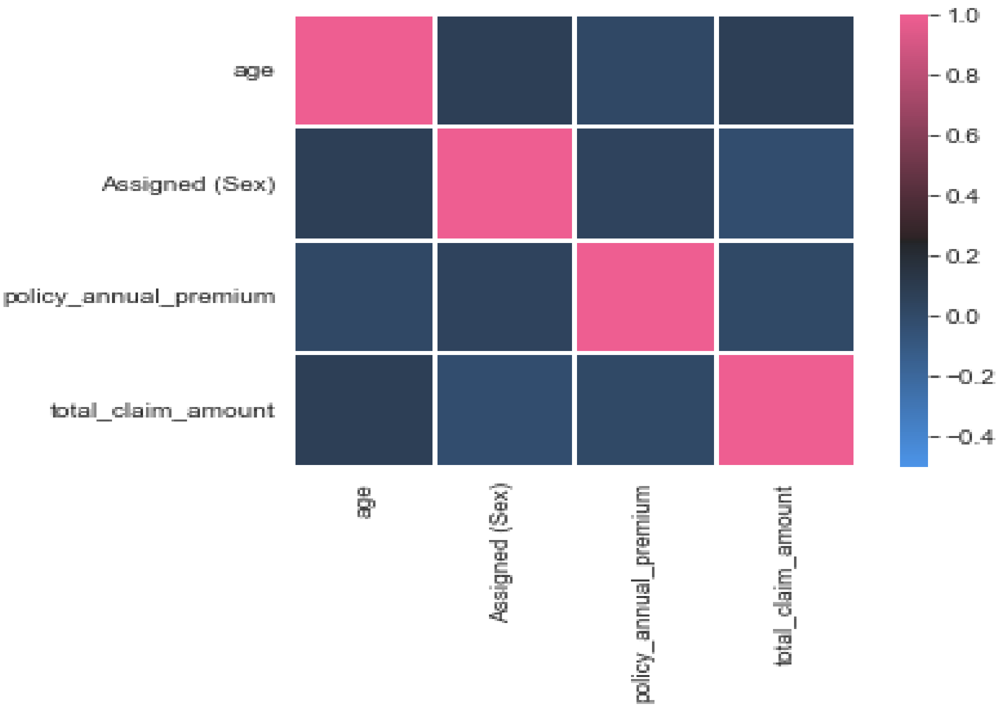


Figure 1 ***Pearson's correlation coefficient heat map.***

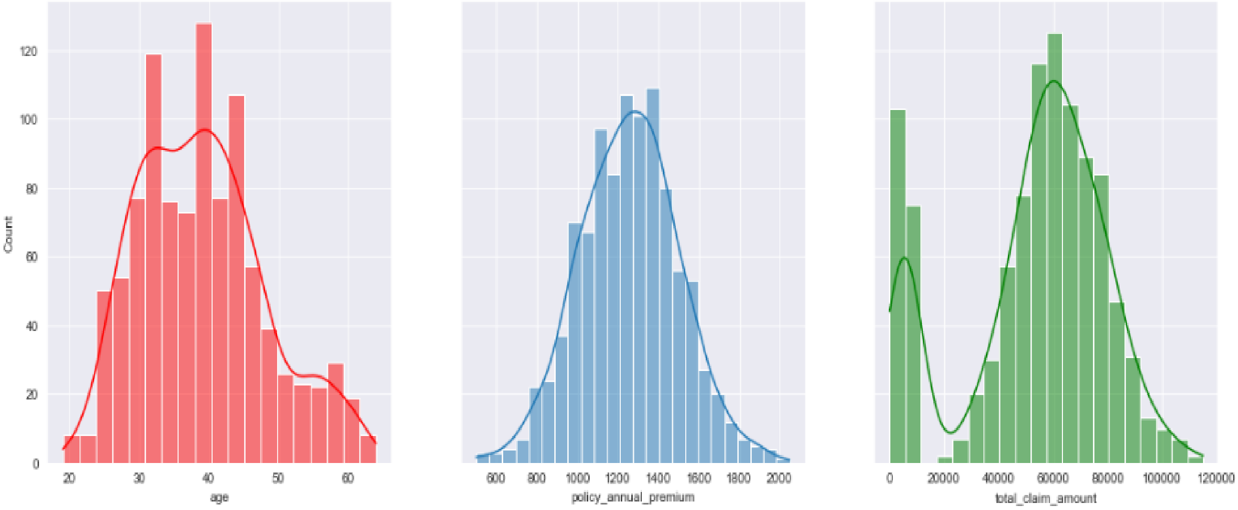


Figure 2 Histogram of attributes against count.

4.3 Fitting the Model

The data was split into testing set and training set. The testing set had 300 datasets and the training set has 700 datasets. To fit the model prior and conditional probabilities are needed. This will be done in the training set category. To get posterior probability we multiply prior probability by likelihood function, in our case the conditional probabilities. The probabilities are as follows.

Row Labels	Count of fraud reported
NO	76.71%
YES	23.29%
Grand Total	**100.00%**

Table 4 Prior probabilities

Count of fraud reported	**Column Labels**		
Row Labels	**NO**	**YES**	**Grand Total**
FEMALE	53.63%	52.15%	53.29%
MALE	46.37%	47.85%	46.71%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Table 5 Insured sex conditional probabilities

Count of Age Level	**Column Labels**		
Row Labels	**NO**	**YES**	**Grand Total**
A1	15.27%	13.50%	14.86%
A2	40.60%	40.49%	40.57%
A3	33.80%	28.83%	31.29%
A4	8.80%	12.88%	10.43%
A5	2.42%	3.88%	2.75%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Table 6 Age Level Conditional probabilities

Age levels were categorized as follows:

20 years – 29 years = A1.

30 years – 39 years = A2.

40 years – 49 years = A3.

50 years – 59 years = A4.

60 years – 69 years = A5.

Count of Premium Level	Column Labels		
Row Labels	NO	YES	Grand Total
P1	14.34%	20.25%	15.71%
P2	69.09%	63.19%	67.71%
P3	16.57%	16.56%	16.57%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Table 7 Premium Level Conditional probabilities

Premium levels were categorized as follows:

Kshs 500 – Kshs 999.

Kshs 1000 – Kshs 1499.

Kshs 1500 – Kshs 1999.

Count of Claim Level	Column Labels		
Row Labels	N	Y	Grand Total
C1	36.87%	25.15%	34.14%
C2	62.01%	74.23%	64.86%
C3	1.12%	0.61%	1.00%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Table 8 Claim Level Conditional probabilities

Claim Levels were categorized as follows:

Kshs 0 – Kshs 49,999.

Kshs 50,000 – Kshs 99,999.

Kshs 100,000 – Kshs 149,999.

4.4 Prediction

The model was fitted on the training dataset to get the posterior distribution. The tables above demonstrate the counts and subsequent probabilities of a few selected variables in the dataset. There were 537 legitimate claims and 163 fraudulent claims in the training dataset.

Given the simulated data, we estimate the prior probability as

P(YES)=163/700=0.23

P(NO) =537/700=0.77

These are the prior probabilities for p(fraudulent) and p(legitimate) respectively.

We use these values to classify a new claim. Suppose we want to classify X = (Age = 37 **(A2),** Male, Premium = Kshs 1086.48 **(P2)**, Claim = Kshs 77,440 **(C2)**). using the associated probabilities of Age, Male, Premium Limit, and Claim Limit, we obtain the following estimates:

P(X |legitimate) = P(NO) * P(A2|NO)*P(MALE|NO)*P(P2|NO)*P(C2|NO)

= 77%*40.60%*46.37%*69.09%*62.01%

= 0.06210.

P(X |fraudulent) = P(YES)*P(A2|YES)*P(MALE|YES)*P(P2|YES)*P(C2|YES)

= 23%*40.49%*47.85%*63.19%*74.23%

= 0.0209.

We wish to normalize to get the probabilities.

P(X |legitimate) = 0.06210/(0.06210 + 0.0209)

= 0.75 = 75%

P(X |fraudulent) = 0.0209/(0.06210 + 0.0209)

= 0.25 = 25%

Cleary, based on these probabilities, we can classify the new claim as a legitimate claim because it has a higher probability. Observations are treated independently therefore adding the redundant ones decreases its predictive power. In order to relax this conditional independence, we add the derived observations that have been created from combining the existing observations.

CHAPTER 5

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

In this study, we describe the Naive Bayes approach and the justifications for its application. We concentrate on creating an efficient classifier to categorize automobile insurance claims in order to provide effective services to customers while keeping the company’s stability in the market environment. Based on the findings, the performance of the naïve bayes model was effective with an approximate accuracy of 80%. This model can be included into a decision support system for insurance in identifying fraudulent claims, although it still needs to be compared to other methods to see which is the best method.

5.2 Recommendations

Finding a dataset was the study's first challenge; even publicly accessible datasets in the field of fraud detection are uncommon and challenging to come by. For future studies, the creation of a Kenyan fraud detection dataset is advised. Furthermore, only the Naive Bayes classifier was used to create the model in this project so that future research could compare the effectiveness of the models by using other classification algorithms, such as random forest.

CHAPTER 6

6 REFERENCES

Akomea-Frimpong, I., Andoh, C., Akomea-Frimpong, A., & Dwomoh-Okudzeto, Y. (2019). Control of fraud on mobile money services in Ghana: an exploratory study. *Journal of Money Laundering Control*, 22(2), 300-317.

Albrecht., C. C., Zimbelman M. F., W.S. Albercht dan C. O. Albrecht (2011). *Forensic Accounting Fourth Edition. AS: South-Western Cengage Learning*

Barber, P. R., Ameer-Beg, S. M., Pathmananthan, S., Rowley, M., & Coolen, A. C. C. (2010). A Bayesian method for single molecule, fluorescence burst analysis. *Biomedical optics express,* 1(4), 1148-1158.

Burri, R. D., Burri, R., Bojja, R. R., & Buruga, S. (2019). Insurance claim analysis using machine learning algorithms. *International Journal of Innovative Technology and Exploring Engineering*, 8(6S4), 577-82.

Chantachit K., (2008). Economic Crime : *Case study on the Automobile Insurance Defraud*. *Master of Law. Bangkok: Graduate School, Dhurakij Pundit University.*

Chuleekorn, T. (2015). The role of using Data Analytics to detect fraud in insurance. *The Insurance Journal. 127: 9-14*

Clifton, D. A., Niehaus, K. E., Charlton, P., & Colopy, G. W. (2015). Health informatics via machine learning for the clinical management of patients. *Yearbook of medical informatics*, *24*(01), 38-43.

Cooper, C. (2003). Turning information into action. *Computer Associates: The Software That Manages eBusiness, Report* .

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access, 8, 58546-58558.*

Feldman, R. (2001). An economic explanation for fraud and abuse in public medical care programs. *The Journal of Legal Studies*, 30(S2), *569-577.*

Helbock M. (2018). *The Law Office of Melinda J. Helbock, A.P.C, Accident/Injury*

Kamdee A., (2005). Non-Life Insurance (Property and Liability Insurance) Defraud: *A case study on Automobile Insurance Defraud by the Insured. Master of Law. Bangkok: Graduate School, Dhurakij Pundit University.*

Kosaisook A., (2007). Fraud in Non-Life Insurance Business. *Master of Law. Bangkok: Graduate School, Chulalongkorn University*

Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1338-1343). IEEE.

Liu, Y., Yang, Q., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(2), 1-19.

Kuria, J. T., & Moronge, M. (2014). Effect of fraud control mechanisms of the growth of insurance companies in Kenya. *International Journal of Innovative Social & Science Education Research, 2*(1), 26-39.

Mukherjee, A., Kumar, A., & McGinnis, J. (2015). Who Wants to Be an E-tailpreneur? Experiences from an Electronic Retailing Course. *Marketing Education Review, 25(2), 117-128.*

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, *7*(2), 70.

Ngosiah (2012). Effect of Fraud Control Mechanisms on the Growth of Insurance Companies in Kenya: *International Journal of Innovative Social & Science Education Research, 2(1), 26-39.*

Njenga, N., Osiemo M. (2013). Effect of fraud risk management on organization performance: A case of deposit-taking microfinance institutions in Kenya. *International Journal of Social Sciences and Entrepreneurship, 1(7), 490-507.*

Ongati, F., Muchemi, D., & Lawrence, E. (2019). Big Data Intelligence Using Distributed Deep Neural Networks. *arXiv preprint arXiv:1909.02873.*

Palacio, S. M. (2019). Abnormal pattern prediction: Detecting fraudulent insurance property claims with semi-supervised machine-learning. *Data Science Journal*, *18*(1).

Permsirivisarn Pol. Col., (2001). The survey of Automobile Insurance Patterns. *Retrieved from* *http://kukr.lib.ku.ac.th/db/BKN/search_detail/result/9036*

Pozzolo, A. D. (2010). *Comparison of Data Mining Techniques for Insurance Claim Prediction* (Doctoral dissertation, Thesis of Università degli Studi di Bologna, Academic Year 2010/2011. Electronic copy available at: https://dalpozz. github. io/static/pdf/Claim_prediction. pdf KDD'10).

Sadiq, A. S., Faris, H., Ala'M, A. Z., Mirjalili, S., & Ghafoor, K. Z. (2019). Fraud detection model based on multi-verse features extraction approach for smart city applications. *In Smart cities cybersecurity and privacy (pp. 241-251). Elsevier*

SBD (2011). The preview of Thailand-Car Theft and Insurance, *SBD Secure Car Research in United Kingdom*

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, 2(1): pp. (26-39).*

Tesfaye, F. (2017). *The Opportunities and Challenges of Life Insurance Growth in Ethiopia* (Doctoral dissertation, st. mary's University).

Tope, M. Email Spam Detection using Naive Bayes Classifier.

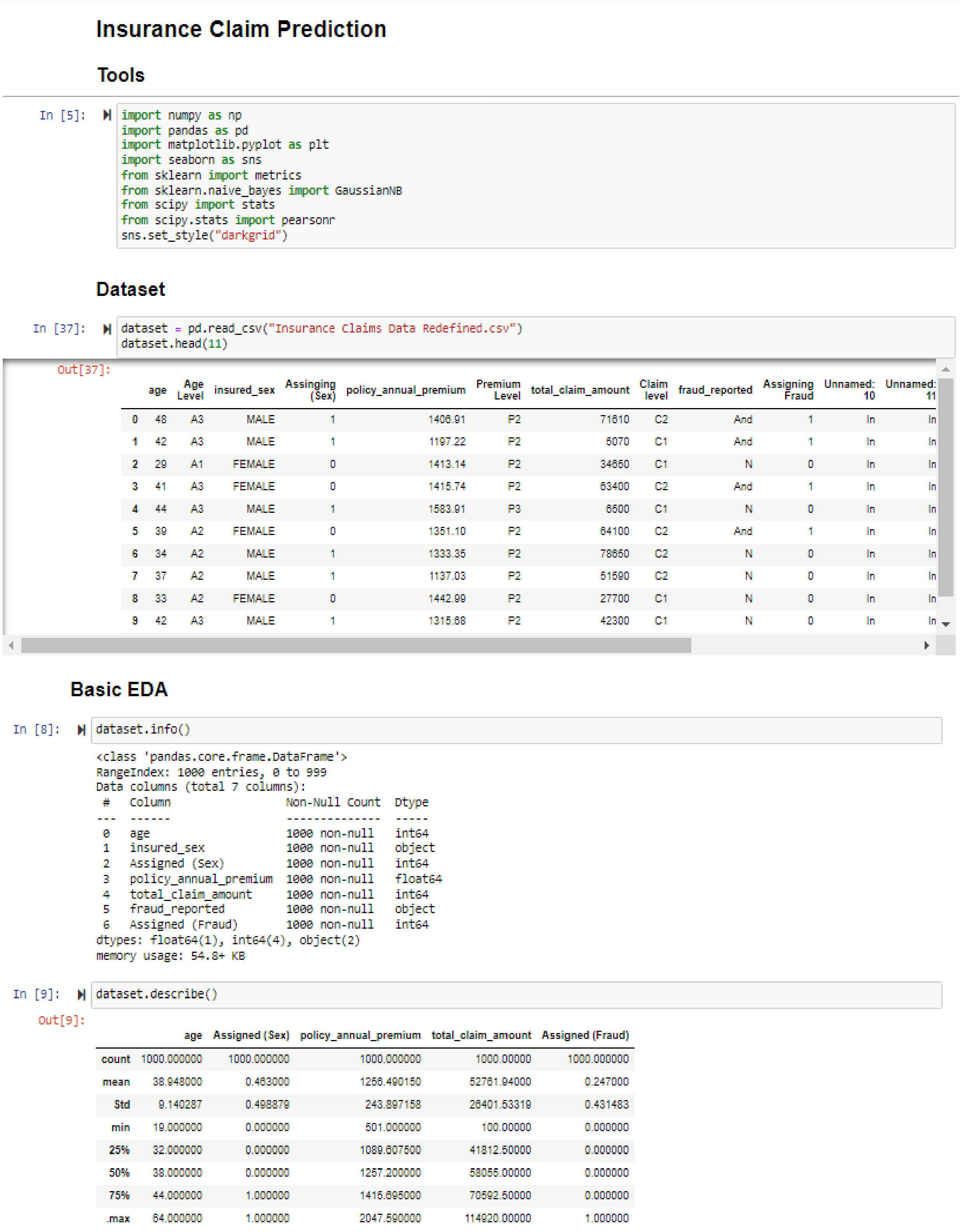
Tseng, L. M., & Kuo, C. L. (2014). Customers' attitudes toward insurance frauds: An application of Adams' equity theory. *International Journal of Social Economics*.

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 67(5), 1275-1304.*

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision support systems, 50(3), 570-575.*

APPENDIX

Python code used to analyze data



Insurance Claim Prediction

Tools

In [5]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import metrics
from sklearn.naive_bayes import GaussianNB
from scipy import stats
from scipy.stats import pearsonr
sns.set_style("darkgrid")
```

Dataset

In [37]:
```
dataset = pd.read_csv("Insurance Claims Data Redefined.csv")
dataset.head(11)
```

Out[37]:

	age	Age Level	insured_sex	Assinging (Sex)	policy_annual_premium	Premium Level	total_claim_amount	Claim level	fraud_reported	Assigning Fraud	Unnamed: 10	Unnamed: 11
0	48	A3	MALE	1	1406.91	P2	71610	C2	And	1	In	In
1	42	A3	MALE	1	1197.22	P2	5070	C1	And	1	In	In
2	29	A1	FEMALE	0	1413.14	P2	34650	C1	N	0	In	In
3	41	A3	FEMALE	0	1415.74	P2	63400	C2	And	1	In	In
4	44	A3	MALE	1	1583.91	P3	6500	C1	N	0	In	In
5	39	A2	FEMALE	0	1351.10	P2	64100	C2	And	1	In	In
6	34	A2	MALE	1	1333.35	P2	78650	C2	N	0	In	In
7	37	A2	MALE	1	1137.03	P2	51590	C2	N	0	In	In
8	33	A2	FEMALE	0	1442.99	P2	27700	C1	N	0	In	In
9	42	A3	MALE	1	1315.68	P2	42300	C1	N	0	In	In

Basic EDA

In [8]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   age                    1000 non-null   int64
 1   insured_sex            1000 non-null   object
 2   Assigned (Sex)         1000 non-null   int64
 3   policy_annual_premium  1000 non-null   float64
 4   total_claim_amount     1000 non-null   int64
 5   fraud_reported         1000 non-null   object
 6   Assigned (Fraud)       1000 non-null   int64
dtypes: float64(1), int64(4), object(2)
memory usage: 54.8+ KB
```

In [9]:
```
dataset.describe()
```

Out[9]:

	age	Assigned (Sex)	policy_annual_premium	total_claim_amount	Assigned (Fraud)
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000
mean	38.948000	0.463000	1256.490150	52761.94000	0.247000
Std	9.140287	0.498879	243.897158	26401.53319	0.431483
min	19.000000	0.000000	501.000000	100.00000	0.000000
25%	32.000000	0.000000	1089.607500	41812.50000	0.000000
50%	38.000000	0.000000	1257.200000	58055.00000	0.000000
75%	44.000000	1.000000	1415.695000	70592.50000	0.000000
.max	64.000000	1.000000	2047.590000	114920.00000	1.000000

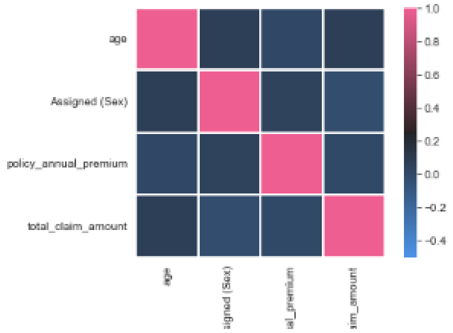
```
In [33]:  dataset.corr()
```

Out[33]:

	age	Assinging (Sex)	policy_annual_premium	total_claim_amount	Assigning Fraud	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
age	1.000000	0.073337	0.013991	0.069863	0.012143	In	In	In	In
Assinging (Sex)	0.073337	1.000000	0.039133	-0.023727	0.030873	In	In	In	In
policy_annual_premium	0.013991	0.039133	1.000000	0.008643	-0.014538	In	In	In	In
total_claim_amount	0.069863	-0.023727	0.008643	1.000000	0.163651	In	In	In	In
Assigning Fraud	0.012143	0.030873	-0.014538	0.163651	1.000000	In	In	In	In
Unnamed: 10	In	In	In	In	In	In	In	In	In
Unnamed: 11	In	In	In	In	In	In	In	In	In
Unnamed: 12	In	In	In	In	In	In	In	In	In
Unnamed: 13	In	In	In	In	In	In	In	In	In

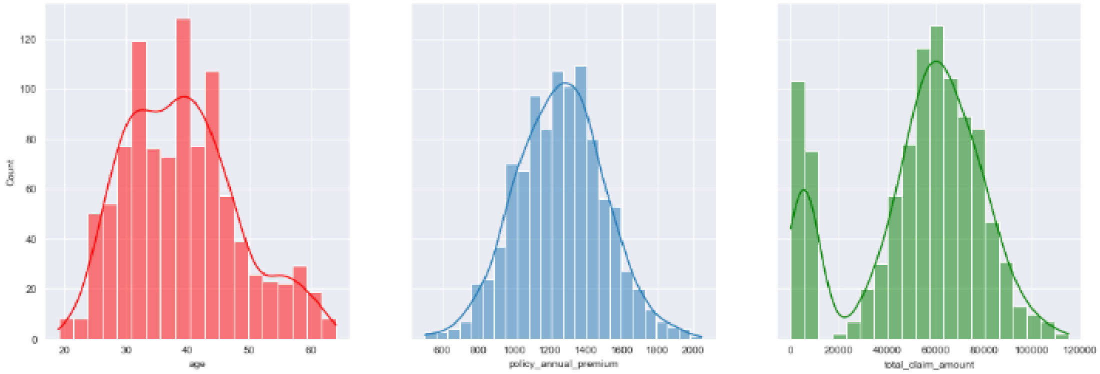
```
In [12]:  corr = dataset.iloc[:,:-1].corr(method="pearson")
          cmap = sns.diverging_palette(250,354,80,60,center='dark',as_cmap=True)
          sns.heatmap(corr, vmax=1, vmin=-.5, cmap=cmap, square=True, linewidths=.2)
          #sns.heatmap(data.corr(), cmap=cmap, vmin =-1, vmax=1, annot=True);

Out[12]:  <AxesSubplot:>
```



```
In [32]:  fig, axes = plt.subplots(1, 3, figsize=(18, 6), sharey=True)
          sns.histplot(dataset, ax=axes[0], x="age", kde=True, color='r')
          sns.histplot(dataset, ax=axes[1], x="policy_annual_premium", kde=True)
          sns.histplot(dataset, ax=axes[2], x="total_claim_amount", kde=True, color='g')

Out[32]:  <AxesSubplot:xlabel='total_claim_amount', ylabel='Count'>
```



```
In [15]:  # Providng the input and output data
          # Importing numpy library
          import numpy as np
```

```
In [16]:  # The input data (X, regressor)
          x = np.array(dataset["age"],).reshape(-1,1)
```