# An Ensemble Feature Selection Model with Machine Learning Model for Detection of Fraudulent Motor Vehicle Insurance Claims.

**Anthony Mwiti Wambu & Dr. Erik Araka**

Department Of Computing and Information Sciences

School of Pure and Applied Sciences, Kenyatta University, Nairobi, KENYA

**ABSTRACT**

Insurance companies are continuously inventing new competitive insurance products in order to enlarge their market share. This has continuously created opportunities for insurance fraud as well. In motor vehicle insurance, fraudulent claims continue to be a big challenge despite the insurance industry having vast amounts of motor vehicle insurance policy data and claim's data. Proper analysis of this data can help to develop a more efficient way of detecting fraudulent claims. The challenge is how to extract insightful information and knowledge from this data since by their nature insurance datasets contain noisy features or subsets of data which are of poor quality. In order to achieve an effective machine learning model, one needs to choose the right set of features of data in the pre-processing step. Including noisy and less important features has proven to affect the performance of most of the existing machine learning models being used in the insurance companies. With aid of proper and effective feature selection techniques machine learning models that uses only relevant features of data can be developed in order to detect fraudulent insurance claims effectively. These models can be employed in insurance industry to aid in detecting fraudulent motor vehicle insurance claims. This will result in reduction of loss adjustment expenses and also improve customer satisfaction. Although there exist several other methods and ways of data preprocessing this study employed an ensemble multiple filter feature selection method. This study involved development of motor vehicle fraud detection model using multiple algorithms whose results were combined by use of a voting method.

**Keywords:** Data Mining, Machine Learning, Feature Selection, Decision Tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, ensemble multiple filter feature selection method, information Gain (IG), Voting, SMOTE

# CHAPTER 1

## INTRODUCTION

### 1.o. Introduction

This chapter covered the background of the study, statement of the problem, the research objectives, the research questions, the justification, the scope of the study and limitations of the study. The chapter started with background of the study which stipulates the current state of fraudulent insurance claims detection more so motor vehicle insurance claims and the various techniques in place. The research gap was then identified in the problem statement which was then discussed in the research objectives, research questions, justification of the study, the scope and limitations of the research.

### 1.1.Background of the study

Fraud presents a significant obstacle for insurance firms. It involves actions like filing fictitious claims, exaggerating claims, or incorporating false elements with the intent of obtaining more than what's rightfully due (Baesens et al., 2021). Fraud can occur through deliberate acts or planned omissions, resulting in profits for perpetrators and losses for victims (Subudhi & Panigrahi, 2018). With over a thousand companies globally and trillions in collected premiums, the insurance industry plays a crucial role in national economies (Roy & George, 2017). In Kenya, for instance, the Insurance Regulatory Authority reported a gross written premium of KES 195.2 billion in 2022, with annual premium growth being a consistent trend (*Quarter 4 2022 Industry Release 06-03-2023-1.Pdf*, n.d.).

Motor vehicle insurance policies establish a contract between insurers and vehicle owners, with insurers assuming the risk of any losses incurred due to accidents (Aslam et al., 2022). Fraudulent motor vehicle insurance claims involve illicit attempts to gain financial advantages through false information (Subudhi & Panigrahi, 2018). The insurance sector heavily relies on data analysis, with data mining being widely used, especially in areas like actuarial work, fraud detection, and customer behavior analysis (Firdaus et al., 2021). Actuaries often employ domain-specific models due to data complexities involved. General machine learning methods are mostly used for fraud detection and customer behavior analysis. This allows for the adaptation of advancements from other sectors (Subudhi & Panigrahi, 2018).

The quality of training datasets significantly impacts the performance of supervised learning models. Insurance datasets often contains redundant and irrelevant attributes which tend to hinder model performance. Thus, feature selection before model development is crucial to eliminate low-influence attributes. This

enhances prediction accuracy for various insurance processes such as fraud detection, policy pricing, and customer retention prediction (Roy & George, 2017). Recent studies indicate that combining feature selection methods can enhance model performance by identifying weak features that become significant when grouped and determining features highly correlated with the output class (Guyon & Elisseeff, n.d.).

This study employed an ensemble approach utilizing multiple feature selection techniques and multiple machine learning techniques. Ensemble feature selection method entailed combining filter feature selection method, that is, information gain, gain ratio, and chi-square, to identify important features for use by the multiple machine learning algorithms. The output from these multiple machine learning algorithms was aggregated using a voting algorithm to classify motor vehicle insurance claims as either fraudulent or legitimate.

### 1.2.Statement of the Problem

Detecting fraudulent motor vehicle insurance claims is a significant challenge for insurance companies worldwide, leading to substantial financial losses and reputational damage (Patil, 2023). Traditional fraud detection methods are not able to accurately identify these fraudulent claims due to evolving fraud techniques and complex data patterns (Aslam et al., 2022). To address this issue, machine learning algorithms have been employed, but their effectiveness depends on the complexity and noise within insurance data, making feature selection crucial. The complexity of motor vehicle insurance data arises from diversity in claim types, policyholders, and vehicles, along with potential data collection errors (Taha et al., 2022a). To enhance the accuracy of machine learning models in identifying fraudulent claims, various feature selection strategies, such as filter-based, wrapper, and embedded methods, have been explored (Piao & Ryu, 2017). However, each approach has its limitations, including challenges related to feature interactions, computational complexity, and dependency on specific machine learning techniques (Awan et al., 2019).

This research employed ensemble multiple filter feature selection techniques to overcome these challenges. By leveraging the strengths of individual feature selection methods, ensemble filter feature selection creates a comprehensive model that carefully selects relevant features for machine learning algorithms, hence improving efficiency, robustness, and the ability to detect fraudulent motor vehicle insurance claims effectively.

## 1.3.Objectives

### 1.3.1. General Objective

The main objective of this study was to design, develop and test a model for detecting whether a given motor vehicle insurance claim is fraudulent. This will help insurance companies save on the revenue that could have been used to compensate fraudulent claims.

### 1.3.2. Specific Objectives

1. To investigate the feature selection techniques that can be used to identify features that can be used to build machine learning models for detecting fraudulent motor vehicle insurance claims.
2. To explore machine learning techniques that are currently used detect fraudulent insurance claims.
3. To design and implement an ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims.
4. To evaluate the performance of the ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims.

## 1.4.Research Questions

a) Which feature selection techniques that can be used to identify features for building machine learning models for detecting fraudulent motor vehicle insurance claims?

b) Which machine learning techniques that are currently used detect fraudulent insurance claims?

c) How can an ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims be developed and implemented?

d) How effective is an ensemble feature selection model with machine learning in detecting fraudulent motor vehicle insurance claims?

## 1.5.Justification

This research contributes to the insurance industry fraudulent claim detection domain, by giving insightful recommendations on how to detect fraudulent claims presented to them more accurately, efficiently and in a transparent manner by use of a model that uses ensemble feature selection techniques and multiple machine learning algorithms. At its core, the model acts as an accurate data-driven decision-making tool that helps the

motor vehicle insurance companies to assess the authenticity of the insurance claims. The model analyzes a wide range of features related to motor vehicle insurance claims, policyholders, vehicles and circumstances. The model then systematically evaluates the likelihood of fraud and provide actionable insights that can be used in decision making. The model has the potential of revolutionizing fraud detection strategies in the motor vehicle insurance industry due to its accuracy, efficiency and transparency. This contributes to a more secure and trustworthy insurance ecosystem that safeguards the financial interests of the insurance companies as well as offering satisfaction to the policyholders.

### 1.6.Scope

This study used an ensemble multiple feature selection technique to reduce the size of the feature space by removing features that are noisy and not relevant. Then various machine learning techniques were applied in order to detect fraud in motor vehicle insurance claims. The results from the machine learning algorithms used were combined using a voting algorithm in order to come up with final output. The study made use of online available datasets for motor vehicle insurance claims fraud from Kaggle dataset (www.kaggle.com).

### 1.7.Hypothesis

The hypothesis of this research was ensemble feature selection techniques and machine learning algorithms can be effectively applied to improve the accuracy and efficiency of fraudulent motor vehicle insurance claim detection

### 1.8.Limitations

1. It was difficult to get the metadata of the dataset that could assist in getting more insights on the data being used to train a model. This is because most insurance companies were not willing to give access of the data that they hold, because of its sensitivity.
2. Most of the datasets available had unknown sources therefore unable to verify the originality of the data.

# CHAPTER 2

## LITERATURE REVIEW

### 2.0. Introduction

This chapter gave a review of existing work in relation to motor vehicle insurance claims, feature selection techniques their pros and cons, how they were used to build an ensemble feature selection model and machine learning techniques that were used to detect fraudulent insurance claims. The chapter formed a basis of conceptual framework for the research.

### 2.1.Feature Selection Techniques

Feature selection refers to the process within machine learning where a subset of variables is chosen from a larger set. This serves to eliminate irrelevant and redundant features, thereby mitigating overfitting, enhancing interpretability, and reducing computational complexity of models (Cai et al., 2018). The selection of a feature selection technique depends on factors such as the nature of the problem, characteristics of the dataset, and the machine learning algorithm employed. An effective technique should prioritize simplicity, accuracy, and interpretability of the model (Taha et al., 2022b). Feature selection methods can be classified as filter methods, wrapper methods, embedded methods, or ensemble methods.

#### 2.1.1. Wrapper methods

Wrapper methods are a category of feature selection methods that work by training and evaluating a model with various subsets of features and then the one that achieves the optimal performance is selected (Piao & Ryu, 2017). They are dependent of the learning algorithm, which may be supervised or unsupervised (Taha et al., 2022b). Wrapper methods have high performance measures but they take too long to run (Wang et al., 2019). They are also restricted to a specific learning algorithm. Among the commonly used techniques within wrapper methods are Forward selection, backward elimination and Bi-directional elimination (Njoh-Paul, n.d.).

### 2.1.2. Filter methods

Filter methods represents a class of feature selection techniques that leverages statistical measures like information gain, gain ratio, chi-square or correlation to rank features according to their relevance to the target variable or study objective (Wang et al., 2019).

Unlike the wrapper methods, filter methods are independent of the specific machine learning algorithm as they are applied prior to classification. While filter methods may not surpass wrapper methods in performance, they are extensively employed due to their high scalability, rapid execution, and suitability for high-dimensional data (Tuv, n.d.).

### 2.1.3. Embedded methods

Embedded methods constitute a category of feature selection techniques that operate by selecting features during the model training process (Wang et al., 2019). These methods modify the algorithm used to incorporate feature selection into both the model training and optimization processes. Embedded methods combine elements from both filter and wrapper methods (Guyon & Elisseeff, n.d.). Unlike wrapper methods, embedded methods do not iterate the learning algorithm, making them more efficient, although they typically do not surpass wrapper methods in performance (Taha et al., 2022b). Examples of embedded feature selection methods include tree-based machine learning algorithms such as random forest, gradient boosting, and decision trees (Pes, 2020).

### 2.1.4. Ensemble methods

Ensemble feature selection methods is a category of feature selection methods which works by combining individual feature selection techniques to collectively identify subset of features to be used in machine learning (Tuv, n.d.). This help to overcome the limitation of the individual methods and at the same time leverage on the strengths of the methods. The ensemble methods result in improved feature selection which leads to enhanced effectiveness and efficiency of the machine learning algorithm (Wang et al., 2019). Ensemble methods can be grouped into categories explained below, depending on the way they work:

#### 2.1.4.1. Stability selection

Stability selection is an ensemble feature selection method which works by applying a feature selection technique multiple times in order to create different subset of features. Features are consistently chosen across the subsets by aggregating the sections across iterations. The chosen features are considered to be more stable and are retained (Tuv, n.d.).

#### 2.1.4.2. Recursive Feature Addition

Recursive feature addition is an ensemble feature selection method which works by applying different feature selection methods iteratively and at each iteration a new feature is added to the subset based on its individual selection performance. The features selected most frequently across the iterations forms the final subset (Tuv, n.d.).

#### 2.1.4.3. Voting-Based Ensembles

A voting-based is an ensemble feature selectin method which combines the decisions of multiple individual feature selection methods by a voting mechanism. The features with most votes are considered relevant and selected for the final subset (Tuv, n.d.).

#### 2.1.4.4. Meta- Learning Approaches

Meta-learning approaches are ensemble feature selection techniques that involves training a meta-learner that combines output of feature selection techniques. The meta-learner weighs relevance of features from various feature selection techniques to form the final feature subset (Tuv, n.d.).

#### 2.1.4.5. Genetic Algorithms

Genetic algorithms are ensemble feature selection methods which work by treating feature subsets as individuals in a population. Selection, crossover and mutation operations are done on the subsets of features in order to come with the final subset (Tuv, n.d.).

This research used an ensemble multiple filter feature selection techniques that combined output of information gain algorithm, gain ratio and chi-square in order to harness their combined capability to select features for use in machine learning.

### 2.1.5. Information Gain

Information gain (IG) is a filter feature selection method employed to identify relevant features from a pool of features. Rooted in information theory, it operates by diminishing the uncertainty linked with identifying the class attribute when the feature value is unknown. This method computes the entropy value of the distribution to gauge the uncertainty associated with each feature in determining the output class (Awan et al., 2019). The entropy value of an attribute x can be defined as:

$H(X) = - \Sigma [P(xi) * \log2(P(xi))]$

H(X) is the entropy of the random attribute X, P(xi) represents the probability of event xi occurring, Σ denotes the sum over all possible events xi and log2 represents the base 2 logarithm.

The entropy of attribute X after observing value of another attribute Y can be defined as:

$H(X|Y) = - \Sigma [P(yj) * \Sigma [P (xi\ j|\ yj) * \log2(P (xij\ |\ yj))]]$

H(X|Y) is the conditional entropy of random variable X given random variable Y,

P(yj) represents the probability of event yj occurring for random variable Y.

P(xij | yj) represents the conditional probability of event xij occurring for random variable X given event yj of random variable Y.

The outer sum Σ is taken over all possible events yj. The inner sum Σ is taken over all possible events xij for each event yj. log2 represents the base 2 logarithm.

The conditional entropy H(X|Y) quantifies how much uncertainty remains in the distribution of X when Y is known. It measures the average amount of information required to determine the value of X given the value of Y. If observing Y reduces the uncertainty of X, the conditional entropy will be lower than the unconditional entropy of X.

Information gain from the conditional entropy H(X|Y) can be calculated as:

$\text{Information Gain} = H(X) - H(X|Y)$

H(X) is the entropy of random variable X before observing variable Y.

H(X|Y) is the conditional entropy of random variable X after observing variable Y.

Information gain from the conditional entropy measures the improvement in predictive power gained by considering the value of variable Y when predicting the outcomes of variable X. It reflects how much uncertainty is reduced by observing Y, leading to more informed and accurate predictions (He et al., 2022).

### 2.2.2. Gain Ratio

In this investigation, Gain Ratio was utilized to counterbalance the bias of information gain toward features with significant diversity values. Gain ratio exhibits a heightened value when data is uniformly distributed among branches, but it yields a diminished value when data is concentrated in one branch of the attribute. By considering both the quantity and size of branches, gain ratio adapts information gain, thereby accommodating the inherent information within the dataset. The inherent information of a specific feature is evaluated by examining the entropy distribution of that feature (Duboue, 2020).

Gain ratio of given feature X with a feature value Y can be calculated as:

Gain Ratio(X, Y) = Information Gain(X, Y) / Split Information(X, Y)

Where the intrinsic value or inherent value X can be calculated as:

Intrinsic Value(x) = - Σ [(jSij / jSj) * log2(jSij / jSj)]

jSj is the number of possible outcomes of feature X can take while Sij is the number of actual outcome of feature X (Bolón-Canedo et al., 2014).

### 2.2.3. Chi- square

Chi-square is a statistical metric employed to evaluate the independence between two variables, typically regarding the target or output class. Initially assuming independence between features and the output class, it computes scores to assess this assumption. A high score indicates a notable dependence between the variable and the output class (Bolón-Canedo et al., 2014). Chi- square is calculated using a contingency table and from the table below formula is used to calculate value of the chi-square:

$\chi^2 = \Sigma ((O - E)^2 / E)$

Where:

$\chi^2$ is the chi-square statistic.

Σ represents the sum over all cells in the frequency table.

O is the observed frequency in a specific cell.

E is the expected frequency in the same cell under the assumption of independence.(*Feature Selection by Chi-Squared*, n.d.)

### 2.2.Machine Learning Techniques

Machine learning, a branch of artificial intelligence, allows computer systems to learn from data, improve performance, and make decisions without explicit programming (Hegde et al., 2021). It involves supervised, unsupervised, semi-supervised, and reinforcement learning methods (Kuhn & Johnson, 2013). In supervised learning, labeled input data enables accurate prediction through training with labeled datasets (Breiman et al., 2017), commonly used for classification and regression tasks (Witten et al., 2016). Unsupervised learning identifies patterns in unlabeled data, revealing hidden relationships (Molnar, 2020), while reinforcement learning learns decision-making strategies through interaction with the environment (Mohamad & Tasir, 2013).

Machine learning encompasses various models, algorithms, and systems for learning (Tuggener et al., 2019). Machine learning algorithms, such as Decision Trees, K-Nearest Neighbor, Support Vector Machines (SVM), and Naïve Bayes, enable autonomous learning and decision-making (Hegde et al., 2021). Data mining, utilizing machine learning techniques, has widespread applications across domains (Kuhn & Johnson, 2013).

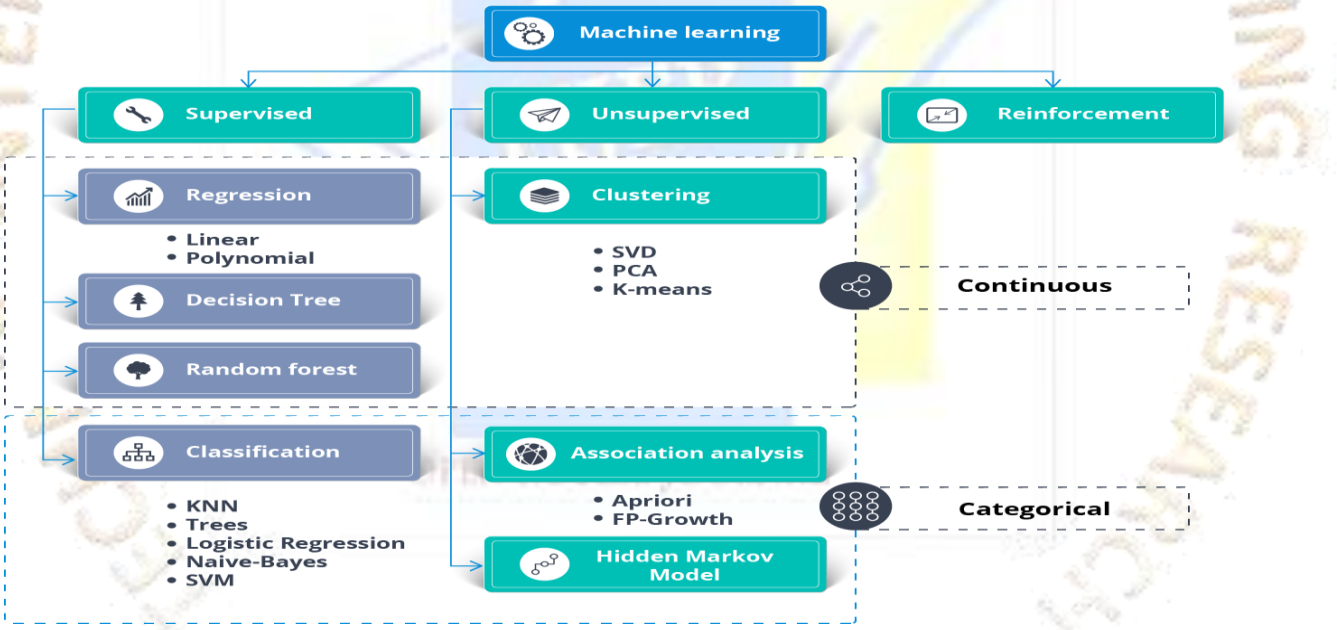



*Figure 1: Machine learning algorithms classification (Fatima et al., 2020)*

#### 2.2.1. Decision Trees

This is a supervised machine learning predictive model. In a decision tree, each internal node represents a test on an attribute or feature, with each branch depicting the outcome of the test, and each leaf node indicating a class label or decision, as highlighted by (Witten et al., 2016). The tree is constructed through iterative splitting

of data into subsets based on the most informative attribute, a process that continues until all data has been classified.

Decision trees offer several advantages, such as interpretability, ease of use, and the ability to handle both categorical and numerical data. However, they are susceptible to overfitting, particularly when the tree becomes overly complex or when the training data is noisy, as noted by (Breiman et al., 2017)

Various algorithms are utilized to construct decision trees, including ID3, C4.5, and CART. These algorithms diverge in their methods for selecting the best attribute to split on and addressing missing data (Molnar, 2020).

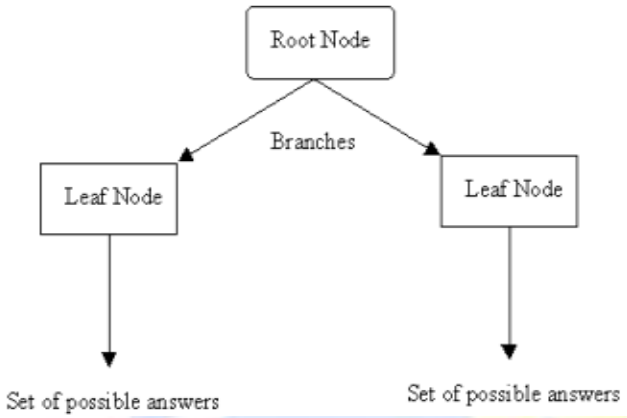



*Figure 2*: A Decision Tree (Hegde et al., 2021).

### 2.2.2. K- Nearest Neighbor (KNN)

In KNN classification, the algorithm assigns a class label to a new instance by examining the class labels of its k-nearest neighbors in the training data. The parameter k determines the number of neighbors considered, and the predicted label for the new instance is the most frequent class label among its k-nearest neighbors (Nicosia et al., 2020).

For KNN regression, the algorithm predicts a continuous output value for a new instance by averaging the output values of its k-nearest neighbors in the training data. KNN is appreciated for its simplicity and adaptability, accommodating various decision boundary types. However, its performance depends on the choice of distance metric for measuring instance similarity and the selection of the hyper-parameter k. Finding the right value for k is crucial to balance bias and variance in the model (Witten et al., 2016). Additionally, KNN can encounter computational challenges with large datasets because it needs to search the entire training data to identify the k-nearest neighbors for each new instance (Patnaik et al., 2017).

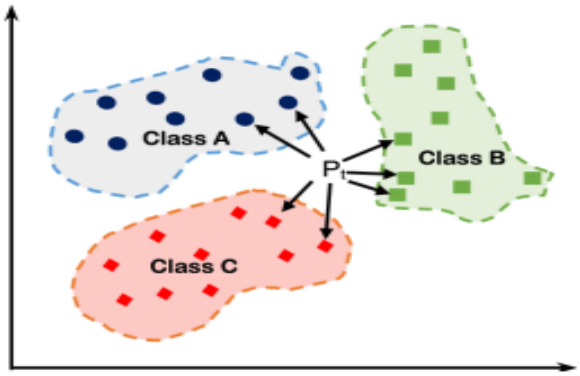



*Figure 3: K-Nearest Neighbor (Hegde et al., 2021)*

### 2.2.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning algorithm employed for classification and regression tasks, applicable to both binary and multi-class scenarios. Its primary aim is to determine a hyperplane, or decision boundary, effectively segregating data points of different classes. The SVM algorithm strives to maximize the margin, the distance between the hyperplane and the nearest points from each class, known as support vectors (Goldberg, 1989).

SVM demonstrates the ability to handle non-linear classification challenges by utilizing a kernel trick. A kernel function transforms the input data into a higher-dimensional space, facilitating the establishment of a linear separation boundary. Commonly used kernel functions include linear, polynomial, and radial basis function (RBF) kernels (Brownlee, 2016).

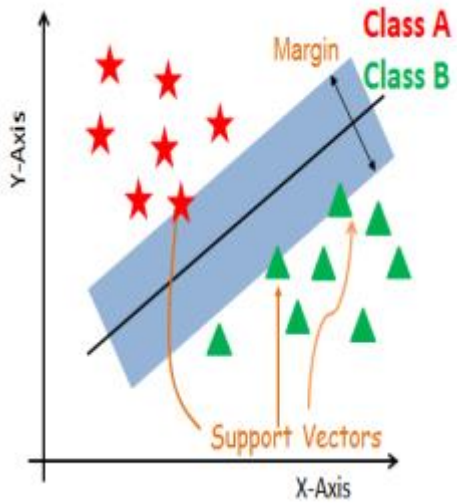



*Figure 4: Support Vector Machine analysis (Hegde et al., 2021)*

### 2.2.4. Naïve Bayes (NB) Algorithm

Naive Bayes, a machine learning algorithm grounded in Bayes' theorem, is utilized for classification and prediction tasks (Sarkar et al., 2018). It functions as a probabilistic model, assuming that features within a class are independent. During training, the algorithm learns the probability distributions of classes and features from the training set. In the prediction phase, it calculates the probability of each class for a given feature set and selects the class with the highest probability (Witten et al., 2016).

Given X and Y are random variables, P(Y) is prior probability of Y, P(Y|X) is the posterior probability of Y, P(X|Y) will be the class conditional probability obtained as:

$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)}$ (Honghong & Lili, 2017)

## 2.3.Related Work

In insurance industry machine learning is mainly applied in actuarial tasks. In insurance ratemaking and reserving, machine learning techniques are widely applied due to data availability in terms of diversity and quantity and also due to the fact that factors that determine the suitable rate of reserve are too complex to be modelled in a linear function (Taha et al., 2022c).

Generalized linear models (GLM) which is a traditional method is still being used along with Gamma or Poisson distribution models for ratemaking. Ratemaking requires calculation of claim severity and claim frequency. Claim severity is gamma distributed, while claim frequency if Poisson distributed(Itri et al., 2019). (Hassan & Abraham, n.d.) explored Generalized Additive Models (GAM) which are more superior than GLMs in calculation of non-linear relationships and hence could perform better in ratemaking tasks. Neural networks have been recently explored by (Al-Hashedi & Magalingam, 2021) for ratemaking tasks and they have proved to be better in modelling non-linear relationships compared to GAM and GLMs. However, Neural networks to work they require large datasets. Due to confidentiality of the insurance data publicly available insurance datasets are few and they contain scarce data. This has impaired exploration of neural networks (Vosseler, 2022).

In the insurance industry, the chain ladder method is typically used for reservation duties. Matrix calculations are used to calculate claims data that has gathered over time using the chain ladder method. A stochastic approach is used to estimate the final reserve amount from the total claims data. The value of the insurance

reserve for the claims is lastly predicted using a stochastic regression model. The stochastic models perform fairly well on large portfolio claims, but they are unable to handle the shifting dynamics that give rise to a claim (Raghavan & Gayar, 2019). These models employ aggregated data and they cannot be used for individual claim level reservations since they are unable to use information about people or small groups (Aslam et al., 2022). Recently, as they are not dependent on historical data, machine learning approaches including SVM, neural networks, deep learning, and tree-based techniques are being used for reserving jobs. These methods can also be applied to a wider variety of data and to reserving claims on an individual basis (Severino & Peng, 2021).

In order to detect insurance fraud, machine learning techniques are used. In most cases, identifying insurance fraud is seen as a classification challenge that needs supervised learning models to determine if a claim is genuine or false (Taha et al., 2022a). Insurance claim data requires manual labeling since it lacks data that has been flagged as fraudulent. Errors and inconsistent labeling are possible when labeling by hand. Due to the rarity of false insurance claims, the majority of insurance claim databases suffer from class imbalances(Vajiram & Senthil, n.d.). Deep learning, text mining, and unsupervised learning were offered as potential solutions by (Belhadji et al., 2000) to the issue of class imbalances. In order to identify fraudulent motor vehicle insurance claims, (Moon et al., 2019) suggested a model that combines text mining with deep learning.

(Verma et al., 2017) introduced a model for detecting fraud in health insurance claims. The model used three mining methods that is, Association Rule Mining that works on the aspect of correlation of data analysis to find frequent patterns, K-Means Clustering that increases detection of outlier, reduces complexity of time, and increases performance in exposing insurance claim frauds. The study classified fraudulent behavior as either period-based disease-based anomalies or claim anomalies.

The effectiveness of machine learning algorithms depends on how pertinent the chosen features are. Several research have used feature selection methods to pick a subset of features from the main collection of features. This aids in making the machine algorithm perform more quickly and precisely(Taha et al., 2022c). It can be difficult to choose the appropriate features for machine learning. To address this issue, several solutions have been put up.

(Belhadji et al., 2000) created a model that extracts a subset of features from a collection of insurance claim data using the filter selection method of IG and chi-squared. The model then employed a decision tree classifier

called C 4.5 and a Bayesian network to identify fraudulent insurance claims. Despite the model's accuracy being the same, the findings showed that the feature selection approaches utilized enhanced the model's overall efficiency.

In order to identify insurance fraud, (Sarker, 2021) created a model that applied supervised inductive learning methodology. The model made use of ensemble and monolithic feature selection methods. The model used IG, gain ratio, and group method for data handling (GMDH) to rank features during the pre-processing stage. Support vector machine (SVM), decision tree (DT), and simulated annealing were used by (Moon et al., 2019) to suggest an insurance reserve (SA). SVM and SA choose the best features, increasing the model's accuracy. A model that used gradual feature removal method from an insurance dataset was proposed by (Patil, 2023). The feature removal was done prior to combining machine learning algorithms such as SVM, ant colony and cluster method in order to develop an insurance rate making model.

A wrapper method was proposed by (S et al., 2021) as a way of removing irrelevant feature for an -insurance fraud detection model. The model used neuro tree to achieved higher accuracy.

In order to discover crucial insurance data aspects for use in insurance ratemaking, a model that employs a multi-measure multi-weight ranking technique was proposed by (Kgare, n.d.). Wrapper, filter, and clustering algorithms are combined in the model's operation to determine the multiple-weight of each feature.

A study done by (Ürgenç et al., 2022) proposed use of filter feature selection methods for use in insurance crime detection. The study noted filter methods are widely used due to their scalability and unlike the wrapper methods filter selection methods are not dependent on machine learning algorithms, and unlike the embedded methods the filter methods are swifter.

It is clear from the aforementioned overview of the literature on feature selection that, regardless of the feature selection technique utilized, the recommended strategies eliminate noise and irrelevant features from the dataset by detecting linked features. It is also noteworthy that the suggested feature selection methods find features that carry particular relevant information about the output class and eliminate the features with little to no information. The literature also demonstrates that several characteristics that could be poor when considered separately can become strong when combined. The literature also reveals that some features which might be weak as individual becomes strong when combined.

Filter feature selection techniques are swift compared to other feature selection techniques (Taha et al., 2022a). They rank features separately based on how useful they are for predicting the output class (Awan et al., 2019).

Contrary to earlier suggested methods, this study used an ensemble multiple filter feature selection method that combines the results of the information gain algorithm, gain ratio, and chi-square to form a final set of features that were used by machine learning algorithms to predict fraudulent claims in motor vehicle insurance.

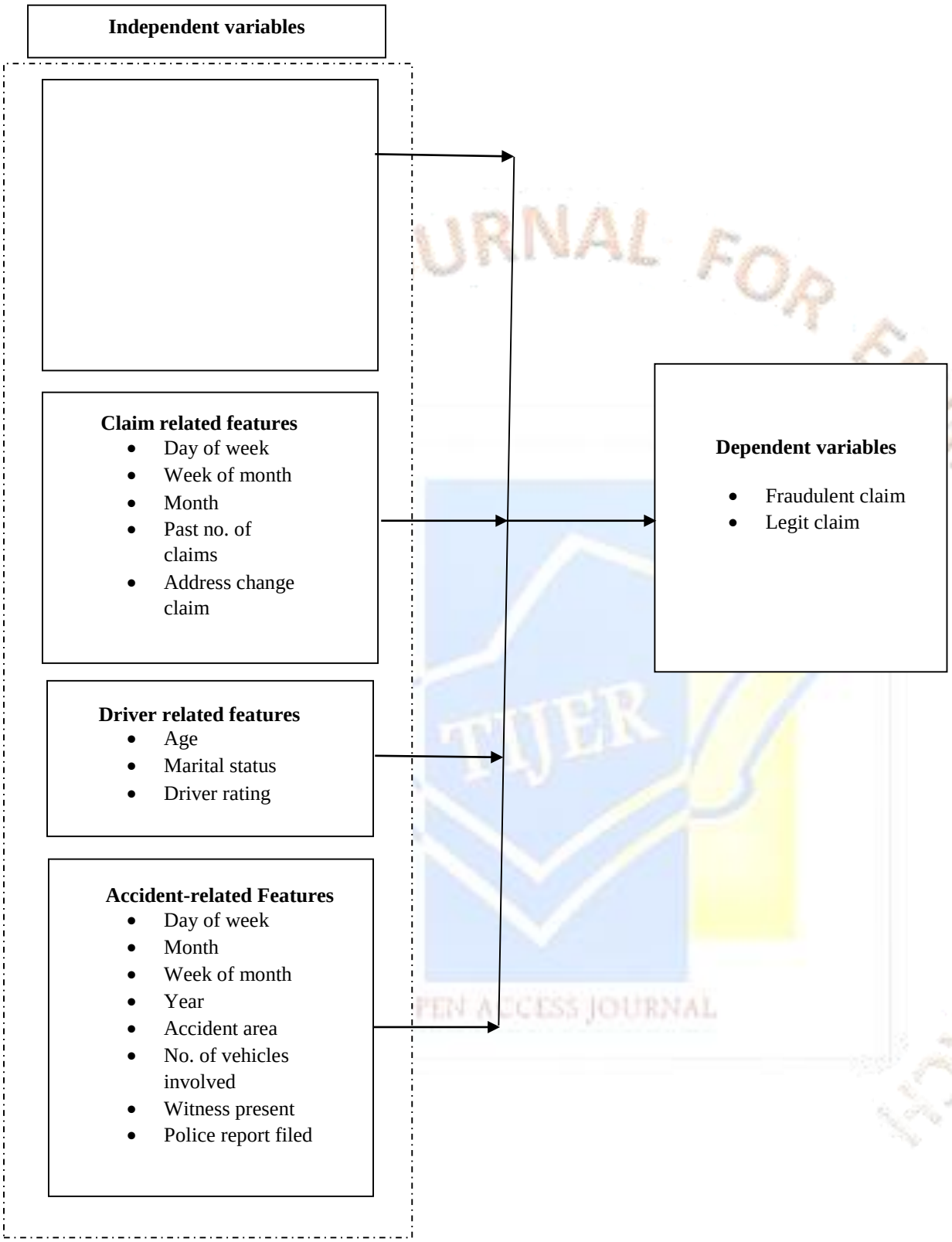
### 2.4.Research Gaps

Considering the proposed models and suggested improvements discussed from great related work, some gaps in the literature were found in relation to fraud detection using data mining techniques and feature selection techniques.

1. How to do effective feature selection in data preprocessing for fraudulent motor insurance claims detection.
2. How missing data were handled while training classification models.
3. A need to tweak the machine learning methods to enhance their accuracy in detecting fraudulent claims.

## 2.5.Conceptual Framework

*Figure 5:conceptual framework*



# CHAPTER 3

## METHODOLOGY

### 3.0. Introduction

This chapter describes how this research work was carried out using ensembled multiple filter feature selection techniques and multiple machine learning algorithms and the procedures that were followed in order to come up with a model that could detect fraudulent motor vehicle insurance claims.

### 3.1.The Research Design

Research design is a framework that guides a researcher on how to carry out the research work and achieve objectives of the project. A research design is selected based on the area of research, research objectives, availability of data and tools(Firdaus et al., 2021).

This study used quantitative experimental research in gathering, model training, evaluation, and analysis. The process involved use of numerical data.

The study used correlational research design, which analyzed relation between dependent and independent variables without manipulating them. The design also defined power of relationship between variables that would help in detection of fraudulent claims in motor vehicle insurance claims.

The study used CRISP-DM methodology in order to achieve all the objectives of the research and be able to deploy the final model.

#### 3.1.1. CRISP-DM Methodology

This study used CRISP-DM methodology. CRISP-DM methodology is widely accepted in the field of data analysis and mining due to its adaptability, and its comprehensive approach to data mining project management. CRISP-DM, an acronym for Cross-Industry Standard Process for Data Mining, was introduced in 1996. It facilitates the organization, planning, and execution of data mining (machine learning) operations(Nielsen et al., 2020). It outlines the standard phases of a data mining project, the tasks associated with each phase, and the interconnections between these tasks, hence offering a holistic view of the data mining life cycle.

CRISP-DM methodology comprises six sequential steps designed to guide the completion of a data mining project effectively. These steps ensure thorough coverage of all aspects of the project, from initial data exploration to model deployment and maintenance. The phases are as follows:

1. Business Understanding – This phase seeks to understand what the business need
2. Data Understanding – This phase seeks to understand which data is needed or is available and whether its lean.
3. Data Preparation – This phase deals with organization of data for modelling.
4. Modelling – This phase deals with modelling techniques and how they are applied in the project.
5. Evaluation – This phase deals with model evaluation to check whether it meets the business objectives.
6. Deployment – This phase deals with how the results are accessed.

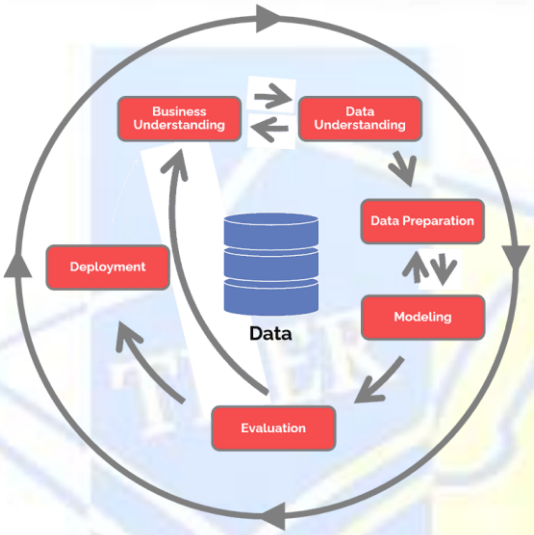



*Figure 6 CRISP-DM Methodology Diagram*

### 3.1.1.1. Business Understanding

Secondary sources were utilized to gain insights into the problem of fraudulent motor vehicle insurance claims. Various references such as books, journals, regional, and global online publications focusing on machine learning techniques for detecting insurance claims fraud were used. Through the analysis of these secondary sources, it became apparent that there has been a notable rise in fraudulent motor vehicle insurance claims. This has resulted to substantial financial losses within the industry.

This presented a pressing need for a system capable of identifying fraudulent insurance claims in real-time within the motor vehicle insurance sector.

### 3.1.1.2. Data Understanding

In this phase, an online dataset site (www.kaggle.com) was used to provide the target population for the study. The dataset obtained had features captured for a motor vehicle insurance claim.

Primary motor vehicle insurance claim data was not available. This is because insurances companies were not wiling provide the data due to its sensitive nature and privacy of the information in it.

The quality of the dataset was assessed and variables were extracted from the dataset in order to aid in model construction. Below is an extract of the dataset obtained which is in csv file:

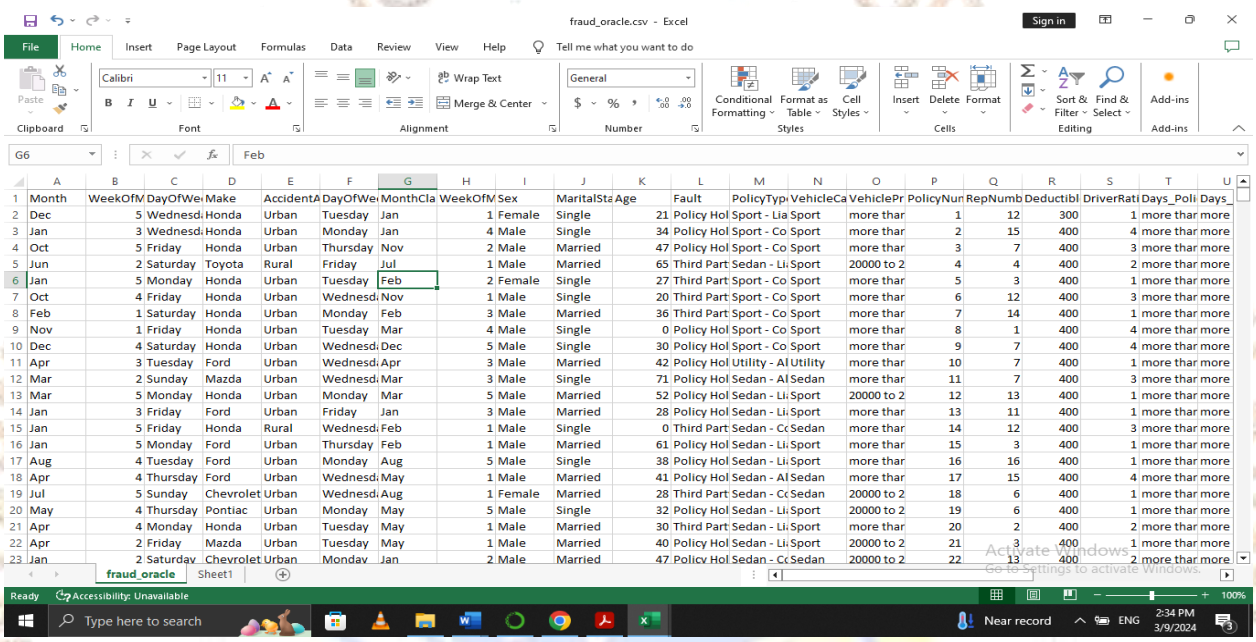



*Figure 7 motor vehicle insurance claims dataset csv file extract.*

The total number of instances in the dataset were 15,420. The dataset was distribution with 923 fraudulent claims which made 6 % of the data while the remaining 14,497 were genuine claims which made 94% of the total data, as seen in the bar graph in the graph below:

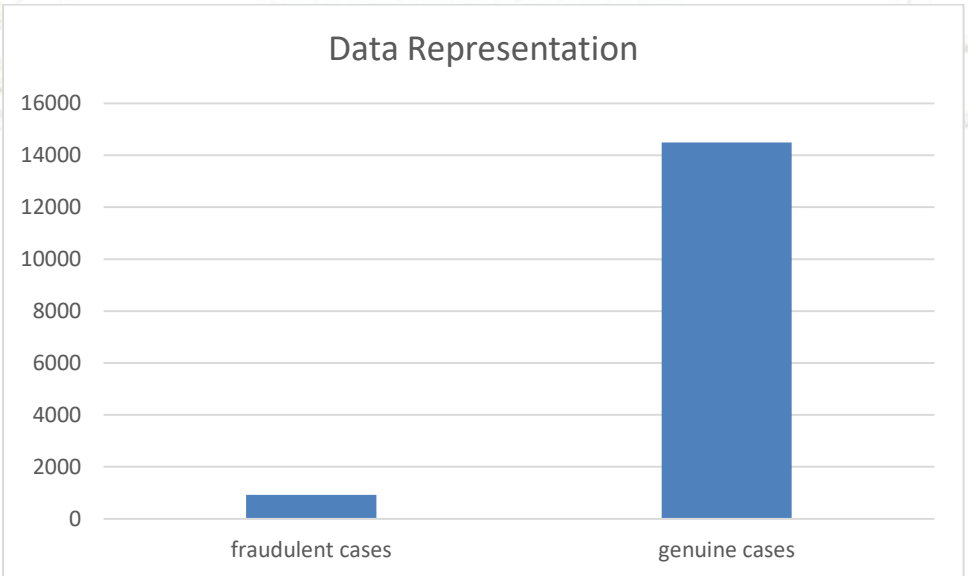



*Figure 8 motor vehicle insurance claims dataset, data distribution.*

A total of 15,420 rows and 33 columns made up the dataset. The columns had the following labels:

Month, WeekOfMonth, DayOfWeek, Make, AccidentArea, DayOfWeekClaimed, MonthClaimed, WeekOfMonthClaimed, Sex, MaritalStatus, Age, Fault, PolicyType, VehicleCategory, VehiclePrice, PolicyNumber, RepNumber, Deductible, DriverRating, Days_Policy_Accident, Days_Policy_Claim, PastNumberOfClaims, AgeOfVehicle, AgeOfPolicyHolder, PoliceReportFiled, WitnessPresent, AgentType, NumberOfSuppliments, AddressChange_Claim, NumberOfCars, Year, BasePolicy, FraudFound_P

All the labels listed above except FraudFound_P label were used as input variables for feature selection model. The resulting set features from the feature selection model formed the independent variable while FraudFound_P label formed the dependent variable.

The independent and dependent variable were later used in machine learning models training and testing in order to detect fraudulent motor vehicle insurance claims.

The dataset had datatypes as shown in the figure below:

```
Month                  object
WeekOfMonth             int64
DayOfWeek              object
Make                   object
AccidentArea           object
DayOfWeekClaimed       object
MonthClaimed           object
WeekOfMonthClaimed      int64
Sex                    object
MaritalStatus          object
Age                     int64
Fault                  object
PolicyType             object
VehicleCategory        object
VehiclePrice           object
PolicyNumber            int64
RepNumber               int64
Deductible              int64
DriverRating            int64
Days_Policy_Accident   object
Days_Policy_Claim      object
PastNumberOfClaims     object
AgeOfVehicle           object
AgeOfPolicyHolder      object
PoliceReportFiled      object
WitnessPresent         object
AgentType              object
NumberOfSuppliments    object
AddressChange_Claim    object
NumberOfCars           object
Year                    int64
BasePolicy             object
FraudFound_P            int64
dtype: object
```

*Figure 9datatypes for the dataset*

The raw data obtained from the online data set was not entirely clean. As shown in the table below, age attribute had 9 rows that had missing data, Deductible attribute had 79 rows missing data, driver rating had 92 rows missing data while the rest of the attributes had complete data.

| Attributes | Number of blank cells |
|---|---|
| Month | 0 |
| WeekOfMonth | 0 |
| DayOfWeek | 0 |
| Make | 0 |
| AccidentArea | 0 |
| DayOfWeekClaimed | 0 |
| MonthClaimed | 0 |
| WeekOfMonthClaimed | 0 |
| Sex | 0 |
| MaritalStatus | 0 |
| Age | 7 |
| Fault | 0 |
| PolicyType | 0 |
| VehicleCategory | 0 |
| VehiclePrice | 0 |
| PolicyNumber | 0 |
| RepNumber | 0 |
| Deductible | 79 |
| DriverRating | 92 |
| Days_Policy_Accident | 0 |
| Days_Policy_Claim | 0 |
| PastNumberOfClaims | 0 |
| AgeOfVehicle | 0 |
| AgeOfPolicyHolder | 0 |
| PoliceReportFiled | 0 |
| WitnessPresent | 0 |
| AgentType | 0 |
| NumberOfSuppliments | 0 |
| AddressChange_Claim | 0 |
| NumberOfCars | 0 |
| Year | 0 |
| BasePolicy | 0 |
| FraudFound_P | 0 |

*Table 1 : Dataset Columns Showing Null Values*

### 3.1.1.3. Data preparation

The data from the online dataset was in raw format hence may contain anomalies, incorrect values or missing values which may compromise its quality and lower performance of data mining techniques. In order to improve data quality for better performance of data mining techniques in predicting fraudulent motor vehicle insurance claims, the study prepared the data first. According to (Nicosia et al., 2020) data preparation entails

removing duplicates, correcting noisy data and handling missing feature values. Data preparation entailed carrying out the below key steps:

i. Data Cleaning: This step focused on removing outliers from the dataset and imputing missing values, aiming to enhance data quality and consistency.

ii. Data Transformation: This involved transforming the data to formats that are usable by the machine learning model.

iii. Data Integration: This involved consolidation of data to facilitate comprehensive analysis.

iv. Data Reduction: Redundant data was eliminated during the data reduction phase. This process enhances efficiency by reducing the volume of data while preserving its informational content.

#### 3.1.1.3.1. Data Clean-up

The data preparation process began with an initial check for duplicate records and missing values. Subsequently, missing values were addressed by replacing them with specified values using the `fillna()` method in Python. Below is a snippet of the Python code employed for detecting duplicate and null values, and then replacing null values with specific 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 10:Python code for Checking and Filling Null Values*

#### 3.1.1.3.2. Data transformation

The data transformation process involved converting data formats into machine-interpretable formats suitable for machine learning classifiers. Textual data was converted into integer values, as machine learning classifiers cannot process text directly. Categorical data was transformed into integer format to facilitate categorical data encoding, making it usable in machine learning. Categorical data encoding, as defined by (Guyon et al., 2008), is the conversion of categorical data into integer representation. Categorical data, according to Guyon et al., refers to information organized into groups with a finite number of possible values.

The Label Encoder function from the Scikit-learn library was employed to perform this conversion, as demonstrated in the Python code snippet below:

```
# Encoding categorical variables
cat_cols = X.select_dtypes(include=['object']).columns
le = LabelEncoder()
X_encoded = X.copy()
for col in cat_cols:
    X_encoded[col] = le.fit_transform(X[col])
```

*Figure 11: Python code for encoding variable in the dataset*

#### 3.1.1.3.3. Data Integration

As part data integration defining the target variable was done. Then it was excluded from the rest of the features since it was the dependent variable and the rest of the features were the independent variables. The below python code was used:

```
# Define the target variable
target_variable = 'FraudFound_P'

# Separating the target variable and features
y = df[target_variable]
X = df.drop(columns=[target_variable])

# Excluding 'PolicyNumber' column from features
X = X.drop(columns=['PolicyNumber'])
```

*Figure 12: Python code for defining the target and separating it from the rest of the features.*

#### 3.1.1.3.4. Data reduction:

This research used ensemble multiple filter feature selection method to select only features that are most important from the dataset. Information gain, gain ratio and chi-square were used to rank features of the original dataset.

##### 3.1.1.3.4.1. Information gain:

Information gain algorithm works by calculating the mutual information in the dataset. Mutual information measures the amount of information obtained about one variable through to the other variable. This measure indicates how much knowing the value of a feature reduces the uncertainty about the target variable(Guyon et

al., 2008). In this case it if quantified the association between each feature and the target variable (fraudulent or genuine insurance claim). After calculating the mutual information for each feature in the dataset, the algorithm ranked the features based on their mutual information scores.

#### 3.1.1.3.4.2. Gain ratio:

Gain ratio is a metric that is used in feature selection to rank the importance of features based on their ability to predict the target variable. It helps in deciding which features are most informative for building predictive models while accounting for the intrinsic complexity of the features(Breiman et al., 2017).

Gain ratio value is calculated by dividing the information gain by the intrinsic information of the feature.

Information gain measures the reduction in entropy (or uncertainty) of the target variable when the data is split based on a particular feature(Witten et al., 2016).

Intrinsic information of a feature is related to the entropy or uncertainty associated with the feature itself. Features with many distinct values or categories might have higher intrinsic information compared to features with fewer values(Guru et al., 2018).

After computing the gain ratio for each feature in the dataset, they were ranked based on their respective gain ratios. Features with higher gain ratios were considered more relevant and informative for prediction of the target variable.

#### 3.1.1.3.4.3. Chi- square

Chi-square worked by computing chi-square score and corresponding p-value for each feature in the dataset. The chi-square score measures the extent of association between a categorical feature and the target variable. The p-value indicates the probability of observing the association by chance alone(Breiman et al., 2017). The computed chi-square scores and p- values were used to evaluate the strength of link between each feature and the target variable. Features with higher chi-square and lower p-values are considered to have stronger link with the target variable, hence they have more potential in predicting the target variable. After computing chi-square scores and p-values for all the features, they were then ranked based on their respective scores.

#### 3.1.1.3.4.4. Final Feature Set

The ranked top 5 features form information gain, gain ratio and chi-square feature selection methods formed mutually exclusive subsets of features. These subsets of features contained the most important features with respect to the feature selection method used. From the three subsets top k features were selected in order to form the final set of features. The value of k in this study was 5.

#### 3.1.1.3.4.5. Data Splitting

The dataset obtained after feature selection was divided into training and test sets using a 70:30 split ratio, where 70% of the data was allocated for training purposes and 30% for testing. For the purpose of model's performance evaluation, the full dataset, before feature selection, was also divided into 70:30 split ratio. This splitting ratio ensured that there was sufficient data for model training while also keeping a side a separate portion for testing the model's performance on unseen data (Witten et al., 2016). The training set was used to train the models, while the test set remained untouched during the training process and was only utilized to assess the model's generalization ability.

### 3.1.1.4. Modelling

In this phase, a model was developed using the following data mining techniques: Decision Tree, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor. These are supervised data mining techniques. The prediction results from individual algorithms were combined by a voting method. This helped to improve the performance of model in predicting fraudulent motor vehicle insurance claims. Use of voting method helped to reduce overfitting which can arise when the individual machine learning algorithms captures noise when learning the training data. overfitting makes the machine learning algorithm perform well on training data but perform poorly with new data(Liu & Motoda, 2012).

Voting method in this research is soft voting method. This entailed combining the probabilities of each prediction from each algorithm and picking the prediction with the highest probability.

The split dataset obtained from feature selection model was used to train and test the model. The full dataset was split for purpose of model's performance evaluation.

In order to deal with the issue of class imbalance the Synthetic Minority Over-sampling Technique (SMOTE) was employed on training sets for the full dataset and the feature selected dataset.

In order to optimize the utilization of available data for both model training and testing, k-fold cross validation technique was implemented across all classifiers. K-fold cross validation is a statistical approach for assessing machine learning models, wherein a dataset is divided into K folds, each taking turns as a testing set. This method helps reduce bias in models as every data point appears in both training and testing sets (Molnar, 2020). For this study, the dataset was portioned into 10 folds (K=10) to refine the model. Initially, the first fold was used as the testing set while the remaining folds were allocated for training. Subsequently, the second fold served as the testing set while the rest functioned as the training set for that iteration. This process continued until all 10 folds were employed as the testing set, ensuring comprehensive evaluation.

#### 3.1.1.4.1. Experiment Environment

For modeling purposes, this study employed Jupyter Notebook, a widely-used interactive computing environment. Jupyter Notebook enables users to develop and share documents featuring live code, equations, visualizations, and narrative text. Supporting multiple programming languages, including Python, it offers a versatile platform for conducting data analysis, machine learning, and other computational tasks.

### 3.1.1.5. Evaluation

An evaluation of performance categorization indicators was conducted to assess the efficiency and effectiveness of the model, alongside establishing the risk threshold. Metrics such as the confusion matrix, classification accuracy, and classification report based on recall, precision, and F-1 score were utilized. This analysis of the model's performance was conducted using both the full dataset and also with feature selected dataset.

#### 3.1.1.5.1. Confusion Matrix

A confusion matrix, according to (Bhowmik, 2008), is a classification performance metric utilized to evaluate the effectiveness of a machine learning algorithm concerning target classes. To derive the classification metrics mentioned above, the following values were initially calculated using a confusion matrix:

- True Positives (TP) – The number of detected fraudulent vehicle insurance claims.
- False Negatives (FN) - The number of fraudulent vehicle insurance claims that remained undetected.
- False Positives (FP) - The count of legitimate vehicle insurance claims that were mistakenly

classified as fraudulent.

- True Negative (TN) - The ratio of legitimate vehicle insurance claims that were not identified as fraudulent.

#### 3.1.1.5.2. Accuracy

According to (Bhowmik, 2008), Accuracy is computed as the proportion of True Positives (correctly predicted observations) to all input observations. (sum of True Positives, False Positives, False Negatives, TrueNegatives).

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

#### 3.1.1.5.3. Precision

Precision is calculated as the ratio of True Positives (correctly predicted positive samples) to the total number of predicted positive samples. (sum of True Positives and False Positives) (Bhowmik, 2008).

Precision = TP / (TP + FP)

#### 3.1.1.5.4. Recall

This metric represents the ratio of accurately predicted positive samples (True Positives) to the total number of samples in the corresponding actual positive class. (sum of True Positives and False Negatives)(Bhowmik, 2008).

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

#### 3.1.1.5.5. F-1 Score

The F1 Score is a measure that combines precision and recall into a single metric. It is calculated using the following formula:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

(Bhowmik, 2008)

#### 3.1.1.6. Deployment

This study developed an effective and innovative model, that uses multiple filter feature selection techniques and multiple machine learning classifiers. This model demonstrated a high level of prediction performance and classification accuracy in identifying fraudulent vehicle insurance claims.

The adoption of this model has the potential to significantly enhance the long-term profitability and customer satisfaction of insurance enterprises.

## CHAPTER 4:

## RESULTS AND DISCUSSIONS

### 4.0. Introduction:

This chapter aims to present the findings of the research, focusing on the utilization of multiple filter feature selection techniques and various machine learning algorithms, that is, Decision Trees, Naive Bayes, k-Nearest Neighbors (KNN), and Support Vector Machines (SVM) for predicting fraudulent motor vehicle insurance claims. The investigation seeks to determine the efficacy of these methods in accurately identifying fraudulent claims within the insurance domain.

### 4.1. Data Exploratory Analysis

Primary data was not available since insurance companies were not willing to provide dataset on vehicle insurance claims due to the confidentiality and sensitive nature of the data it included. As a result, this study used online available dataset from (www.kaggle.com) was used to provide the target population.

The dataset obtained had features captured for a motor vehicle insurance claim. It had a total of 15,420 rows and 33 columns. The dataset was distribution with 923 fraudulent claims which made 6 % of the data while the remaining 14,497 were genuine claims which made 94% of the total data.

The significant disparity in class distribution posed challenges for developing a robust model for detecting fraudulent claims. With fraudulent claims being a minority class, there was a risk of biased model predictions favoring the majority class.

To mitigate the effects of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE is a resampling technique that generates synthetic samples for the minority class, thereby

balancing the dataset and enabling more effective learning from the minority class instances(Houari et al., 2014).

The dataset was not entirely clean. It had some missing data which had to be filled up before being utilized. The missingvalues were replaced with specified values using the fillna() python method.

### 4.2. Multiple Feature selection model evaluation

Multiple filter feature selection model was used to select the relevant features from the dataset for use by the multiple machine learning model. The model employed information gain, gain ration and chi-square to come with a subset top 5 features as per the feature selection technique used.

#### 4.2.1. Mutual information

By use of information gain feature selection technique, the top features that were selected from the dataset are BasePolicy, PolicyType, Fault, VehicleCategory, and Deductible. They had mutual information score as shown in the figure below.

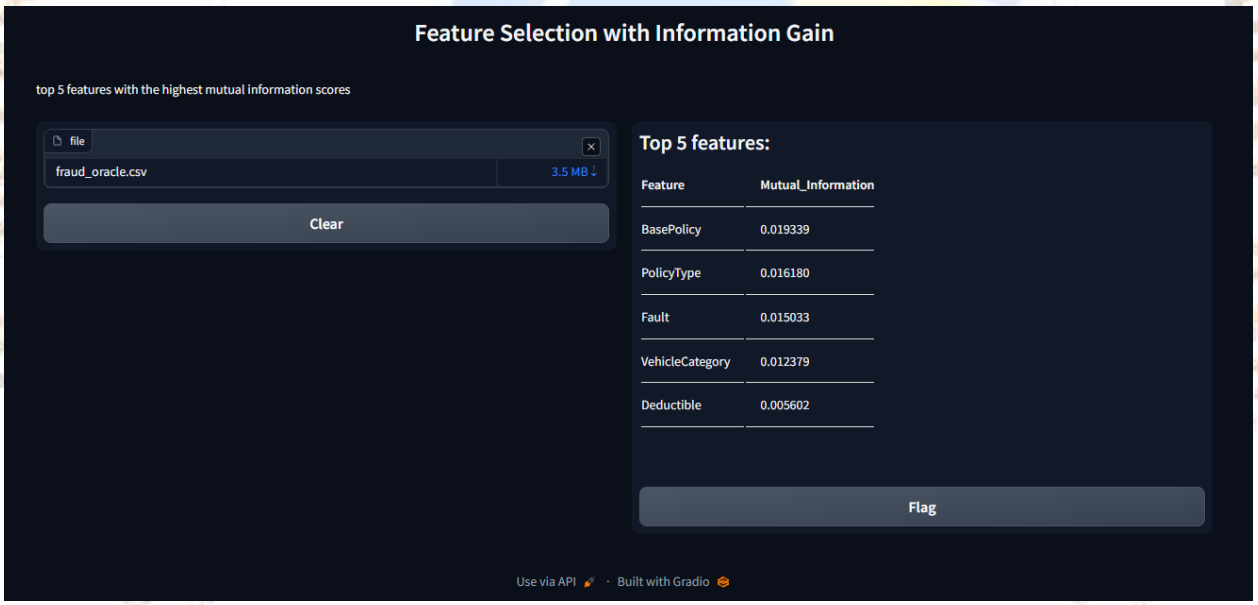



*Figure 13: Feature selection with information gain*

These mutual information scores shown in the figure quantifies how much each feature contributed to predicting the fraudulent insurance claims variable. This measured how much knowing the value of a feature reduced the uncertainty about the target variable(Liu & Motoda, 2012). Hence offering valuable insights into the significance of each feature in identifying fraudulent insurance claims.

Mutual information scores were ranked based on their predictive power or relevance to the target variable. Higher scores indicated that the feature provided more information about the target variable.

The ‘BasePolicy’ feature emerged as the most informative feature in predicting fraudulent claims, with a high Mutual Information score of 0.019339. This suggests that the type of insurance policy held by the claimant plays a pivotal role in determining the likelihood of fraudulent insurance claim. Different policy types may entail varying levels of risk or coverage, thereby influencing the propensity for fraudulent behavior.

Following closely behind, ‘PolicyType’ demonstrated a substantial mutual information score of 0.016180, indicating its significance in distinguishing between fraudulent and non-fraudulent claims based on the type of insurance policy. The specific terms and conditions associated with different policy types may influence the incentives and motivations for fraudulent activities.

The ‘Fault’ feature ranked third in terms of mutual information score (0.015033), suggesting its importance in predicting fraudulent claims. Whether the claim involves fault on the part of the insured party could serve as a crucial indicator of potential fraudulent claim. Claims involving disputed fault or contentious circumstances may warrant closer scrutiny for fraudulent intent.

With a mutual information score of 0.012379, the ‘VehicleCategory’ feature emerged as a significant predictor of fraudulent claims. The category or type of vehicle insured may provide valuable insights into the risk profile of the claimant and the likelihood of fraudulent behavior associated with certain vehicle types.

‘Deductible’ feature demonstrated relevance in predicting fraudulent claims, with a score of 0.005602. The deductible amount specified in the insurance policy could influence the financial incentives for engaging in fraudulent activities.

### 4.2.2. Gain ratio

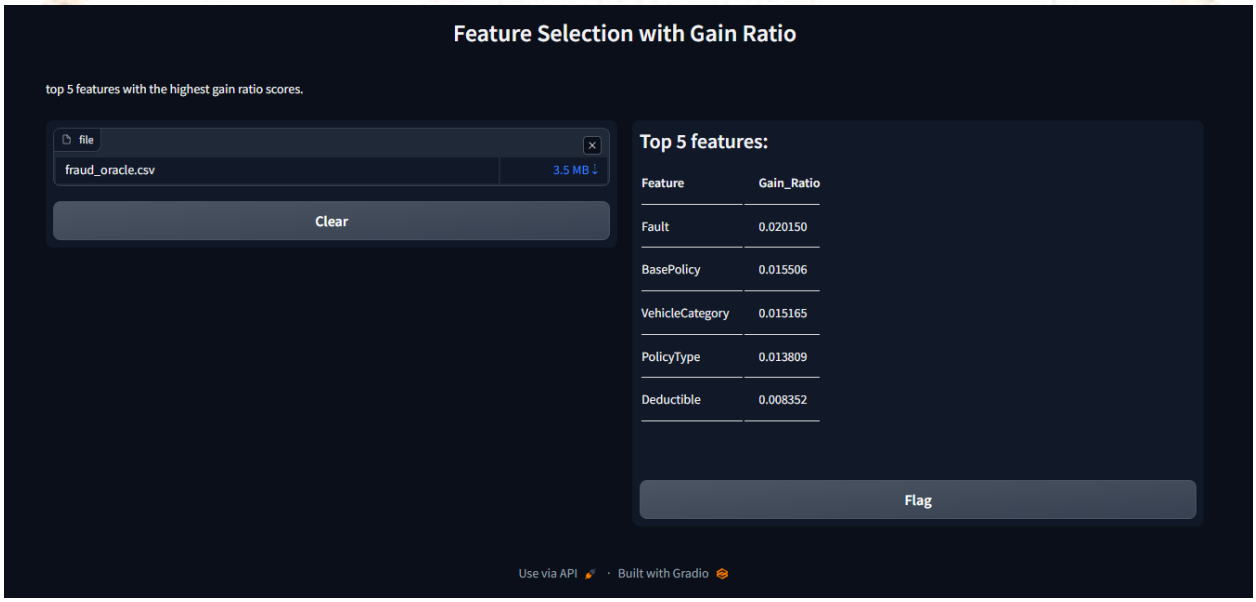



*Figure 14:Feature selection with gain ratio*

Gain Ratio scores provided a quantitative measure of a feature's predictive power relative to the intrinsic information in the dataset(Guyon et al., 2008). A higher Gain Ratio indicated that the feature effectively discriminates between classes, making it more valuable for classification tasks. Gain Ratio takes into account both the purity of the splits produced by the feature and the intrinsic information of the classes, offering a balanced assessment of feature relevance(Liu & Motoda, 2012).

From the dataset 'Fault' , 'BasePolicy', 'VehicleCategory', 'PolicyType', 'Deductible' emerged as the top predictors, ranked by their Gain Ratio scores as indicated in the figure above.

'Fault' emerged as the most influential feature in predicting fraudulent insurance claims, with a high Gain Ratio of 0.020150. This suggests that the attribution of fault in motor vehicle incidents holds substantial predictive power regarding the likelihood of fraudulent behavior. Claims involving disputed fault or ambiguous circumstances may warrant heightened scrutiny for potential fraudulence.

'BasePolicy' feature had gain ratio of 0.015506, indicating its importance in distinguishing fraudulent from non-fraudulent claims. The type of insurance policy held by the claimant serves as a significant predictor of fraudulence, with different policy types potentially associated with varying levels of risk or coverage.

The 'VehicleCategory' feature ranked third in terms of Gain Ratio (0.015165), underscoring its relevance in predicting fraudulent claims. The category or type of vehicle insured provides valuable insights into the risk profile of the claimant and the likelihood of fraudulent behavior associated with specific vehicle types.

'PolicyType' feature emerged as another critical predictor of fraudulent claims with gain ratio of 0.013809. The specific terms and conditions associated with different policy types may influence the incentives and motivations for engaging in fraudulent activities.

'Deductible' feature with a gain ratio of 0.008352 also emerged as major feature in predicting fraudulent motor insurance claim. The deductible amount specified in the insurance policy can impact the financial incentives for fraudulent behavior, warranting consideration in predictive modeling efforts.

4.2.3. Chi-Square

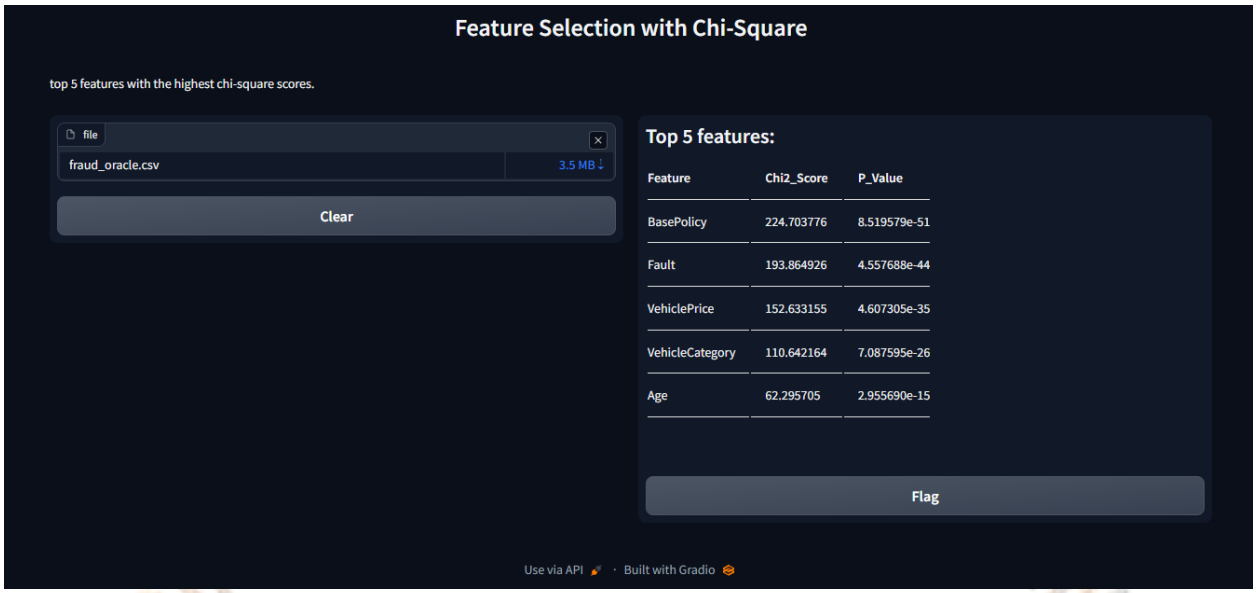



*Figure 15:Feature selection with chi-square*

Chi-square is a statistical measure that is used in feature selection. It works by calculating the significance of the association between two categorical variables(Liu & Motoda, 2012). Chi- square assessed the relationship between each feature and the target variable in the dataset by calculating chi-square scores and corresponding p-values. A shown in the diagram above 'VehiclePrice', 'PastNumberOfClaims', 'BasePolicy', 'Make' and 'Fault' were selected and ranked as the top 5 features with the highest chi-square values.

The 'BasePolicy' feature had the highest chi-square of 224.703776 and the lowest p-value. This feature represented the type of insurance policy held by the claimant. The high Chi2 score and significantly low p-value indicated that different policy types have distinct impacts on the likelihood of fraudulent behavior. For instance, comprehensive policies might incentivize fraudulent claims due to higher coverage.

'fault' feature indicated presence or absence of fault in an insurance claim. It had chi-square score of 193.864926 hence indicating its importance in predicting the target feature. Claims involving disputed fault or unclear circumstances may indicate potential fraud, as claimants may attempt to shift blame to avoid penalties or claim benefits to which they are not entitled.

'VehiclePrice' feature had chi-square score of 152.633155 which shows it's a major feature in determining the target variable. It represented the price or value of the insured vehicle. High-value vehicles may attract fraudulent activities, such as staged accidents or theft, to maximize insurance payouts. Conversely, lower-value vehicles might be targeted for insurance fraud due to their perceived lower risk of detection.

'VehicleCategory' feature had chi-square score of 110.642164. It represented the type or category of the insured vehicle, such as sedan, SUV, or luxury vehicle. Different vehicle categories may be associated with varying levels of risk and susceptibility to fraudulent activities. For instance, luxury vehicles might be targeted for theft or vandalism, while commercial vehicles may be involved in staged accidents for fraudulent claims.

The 'Age' feature had chi-square score of 62.295705. Age of the claimant is a demographic factor that can influence insurance claim patterns. Younger claimants may be more inclined to engage in risky behaviors, such as speeding or reckless driving, leading to a higher likelihood of accidents and subsequent fraudulent claims. Conversely, older claimants may exhibit more cautious driving behavior but could also be targets for insurance scams due to perceived vulnerabilities.

**4.2.3.1. Final set of features**

The top 5 features identified through three different methods, that is Mutual Information, Gain Ratio, and Chi-Square were combined to form the final set of features for use in machine learning model. However, some features were selected by more than feature selection method as top 5 features, as shown in the feature comparison figure below.

'BasePolicy' feature consistently appeared as a top feature across all three feature selection methods. This indicated it's a strong feature for predicting fraudulent insurance claims. This suggests that the type of motor vehicle insurance policy held by the claimant significantly influences the likelihood of fraudulence.

'Fault' and 'VehicleCategory' features were also consistently identified as important features by all the three feature selection methods. This indicated that presence or absence of fault in a claim and the category of the insured vehicle provide valuable information for detecting fraudulent activities in motor vehicle insurance claim.

'PolicyType' and 'Deductible' features were selected by both the information gain and gain ratio feature selection methods. This indicated the that policy terms and deductible amounts had a lot of relevance in predicting fraudulent motor vehicle insurance claims.

'VehiclePrice' and 'Age' features were selected by chi-square as part of the top 5 features with that can help to predict fraudulent motor vehicle insurance.

The final set of features had the following features: BasePolicy, Fault, VehicleCategory, PolicyType, VehiclePrice, Age, and Deductible. These were the features chosen based on their significance across the three feature selection methods.

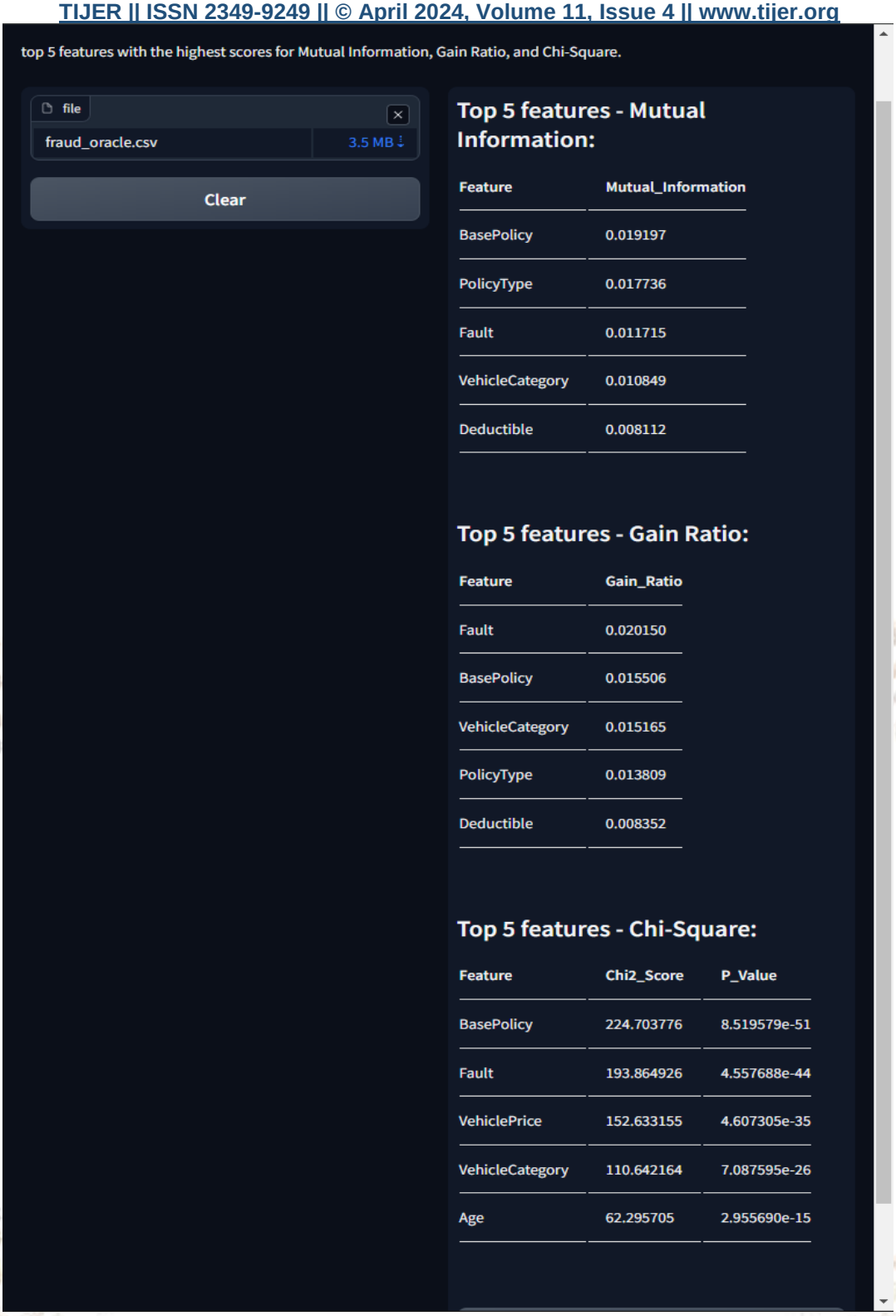



*Figure 16: comparison of Features selected by information gain, gain ratio and chi-square feature selection methods.*

### 4.3. Machine Learning Model Evaluation

The split dataset obtained from the feature selection model was used in the machine learning model. The machine learning model used Decision Tree, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor machine learning techniques. The results from these machine learning algorithms were combined by use soft voting in order to come up with the final prediction.

The full dataset was used to evaluate the performance of machine learning model.

An analysis of the machine learning model classification was performed. The resulting final prediction from the soft voting model using full dataset and also using feature selected dataset, was also analyzed. This was in order to assess the effectiveness and efficiency of the model in discovering fraudulent motor vehicle insurance claims. The execution time for machine learning model with feature selected dataset was approximately 7 minutes and 41 seconds. The execution time for machine learning model with the full dataset was approximately 10 minutes and 15 seconds.

After the model was trained and tested, the model's performance was evaluated, and a report of the model's classification with feature selected dataset and with full dataset was calculated based on accuracy, recall, precision, and F-1 score.

A confusion matrix was employed to evaluate the performance of the machine learning model based on the target variable. It is constructed by computing TP (true positives), representing correctly classified positive instances, TN (true negatives), indicating correctly classified negative instances, FP (false positives), denoting negative instances incorrectly classified as positive, and FN (false negatives), signifying positive instances incorrectly classified as negative. (Bellatreche et al., 2021).

Other metrics such as precision, recall, F1 score, and accuracy were also used to evaluate and understand the effectiveness of the model.

Precision measures the accuracy of positive predictions by calculating the proportion of true positive predictions among all positive predictions made by the classifier. Recall assesses the classifier's capability to identify all positive instances by calculating the proportion of true positive predictions among all actual positive instances in the dataset. A high recall value suggests a low false negative rate, indicating that the model can correctly identify most positive instances(Goldberg, 1989). The F1 score is a combined measure of precision and recall. It provides a single metric that balances between precision and recall. Accuracy measures the overall correctness of the model's predictions by calculating the proportion of correctly classified instances among all instances in the dataset(Witten et al., 2016).

### 4.4. Performance Evaluation and Results

#### 4.4.1. Performance evaluation using feature selected dataset

| Model Type | True Negatives (TN) | False Positives (FP) | False Negatives (FN) | True Positives | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | 3160 | 103 | 1181 | 182 | 0.638 | 0.134 | 0.221 | 0.724 |
| KNN | 4211 | 244 | 130 | 41 | 0.145 | 0.240 | 0.180 | 0.920 |
| Naive Bayes | 2665 | 64 | 1676 | 221 | 0.775 | 0.116 | 0.202 | 0.639 |
| SVM | 3585 | 192 | 756 | 93 | 0.472 | 0.110 | 0.178 | 0.799 |
| Final Prediction | 4340 | 284 | 1 | 1 | 0.015 | 0.999 | 0.030 | 0.938 |

*Table 2 : feature selected Dataset model's Evaluation Report*

From the above performance evaluation of the machine learning model using feature selected dataset, Decision Tree algorithm had a of Precision of 0.638. This indicates that out of all the instances predicted as positive by the Decision Tree, approximately 63.8% were actually positive. It had a Recall of 0.134 which means it correctly identified approximately 13.4% of all actual positive instances. It had F1 Score of 0.221 which is the harmonic mean of precision and recall. This indicates Decision Tree had a balance between precision and recall. The Accuracy for decision tree was 0.724 which indicated the overall accuracy of the Decision Tree model was 72.4%.

K-Nearest Neighbors had a Precision of 0.145 which means that only a small percentage of the instances it identified as positive are actually positive. Recall for KNN was 0.240 which means that it correctly identified 24% of all actual positive instances. The F1 score of 0.180, indicated that KNN's balance between precision and recall was lower compared to the Decision Tree. Despite lower precision and recall, KNN achieved a high accuracy of 92%.

Naive Bayes had a Precision of 0.775. This was the highest precision among all models, correctly identifying 77.5% of positive instances out of all predicted positive instances. However, its recall was quite low at 11.6%, meaning it missed a significant number of actual positive instances. F1 Score of 0.202 for Naive Bayes indicates that it had a moderate balance between precision and recall. Despite high precision, Naive Bayes had an accuracy of 63.9%, which might be due to its low recall.

SVM had a moderate precision of 0.472, which indicates that it correctly identified 47.2% of positive instances out of all predicted positive instances. Its recall was quite low at 11%, indicating it missed many actual positive instances. F1 Score of 0.178 suggests that it had a balance between precision and recall similar to the Decision Tree. SVM achieved an accuracy of 79.9%, which is relatively higher compared to Naive Bayes.

The Final Prediction had a Precision of 0.015. It had a very high recall of 0.999, meaning it correctly identified nearly all actual positive instances. It had a F1 Score of 0.030. The model had a high accuracy of 93.8%.

**4.4.2. Performance evaluation using full dataset dataset**

| Model Type | True Negatives (TN) | False Positives (FP) | False Negatives (FN) | True Positives | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | 3934 | 211 | 407 | 74 | 0.260 | 0.164 | 0.193 | 0.893 |
| KNN | 2949 | 162 | 1392 | 123 | 0.432 | 0.081 | 0.136 | 0.676 |
| Naive Bayes | 3008 | 105 | 1333 | 180 | 0.632 | 0.119 | 0.200 | 0.697 |
| SVM | 2584 | 157 | 1757 | 128 | 0.449 | 1.128 | 0.199 | 0.652 |
| Final Prediction | 4321 | 281 | 20 | 4 | 0.014 | 0.167 | 0.026 | 0.938 |

*Table 3 : Full dataset model's Evaluation Report*

From the above performance evaluation of the machine learning model using the full dataset, the Decision Tree algorithm had a Precision of 0.260. This indicates that out of all the instances predicted as positive by the Decision Tree, approximately 26.0% were actually positive. It had a Recall of 0.164, which means it correctly identified approximately 16.4% of all actual positive instances. It had an F1 Score of 0.193, which is the harmonic mean of precision and recall. This indicates the Decision Tree had a balance between precision and recall. The Accuracy for the Decision Tree was 0.893, which indicated the overall accuracy of the Decision Tree model was 89.3%.

K-Nearest Neighbors (KNN) had a Precision of 0.432, which means that approximately 43.2% of the instances it identified as positive were actually positive. Recall for KNN was 0.081, indicating it correctly identified 8.1% of all actual positive instances. The F1 score of 0.136 indicated that KNN's balance between precision and recall was lower compared to the Decision Tree. Despite lower precision and recall, KNN achieved a moderate accuracy of 67.6%.

Naive Bayes had a Precision of 0.632, the highest precision among all models, correctly identifying 63.2% of positive instances out of all predicted positive instances. However, its recall was relatively low at 0.119, meaning it missed a significant number of actual positive instances. The F1 Score of 0.200 for Naive Bayes indicates that it had a moderate balance between precision and recall. Despite high precision, Naive Bayes had an accuracy of 69.7%, which might be due to its relatively low recall.

SVM had a Precision of 0.449, which indicates that it correctly identified 44.9% of positive instances out of all predicted positive instances. Its recall was unusually high at 1.128, indicating it had more false positives than true positives. The F1 Score of 0.199 suggested that it had a balance between precision and recall similar to the Decision Tree. SVM achieved an accuracy of 65.2%, which is relatively lower compared to Naive Bayes.

The final prediction model had a Precision of 0.014. It had a recall of 0.167, meaning it correctly identified 16.7% of all actual positive instances. The F1 Score of 0.026 indicated that the model's balance between precision and recall was very low. However, the model had a high accuracy of 93.8%, suggesting that its overall performance was good despite the low precision and recall.

**4.5. Fraudulent Vehicle Claims Detection System**

Feature selection model was used along with machine model that employed decision tree, KNN, Naïve Bayes and SVM. From the output of these machine learning models the final prediction was obtained by use of soft voting model.

Gradio which is a Python library was used create the user interface for the fraudulent motor vehicle insurance claims detection system. Gradio is written using python and leverages web technologies such as HTML, CSS, and JavaScript to create interactive user interfaces that can be accessed via web browsers(Nicosia et al., 2020).

The user interface offered an option for uploading an input csv file that was used by feature selection model and the machine learning model as shown in the figure below. The output of the fraudulent motor vehicle insurance claims detection system is a csv file which is saved in the users pc.

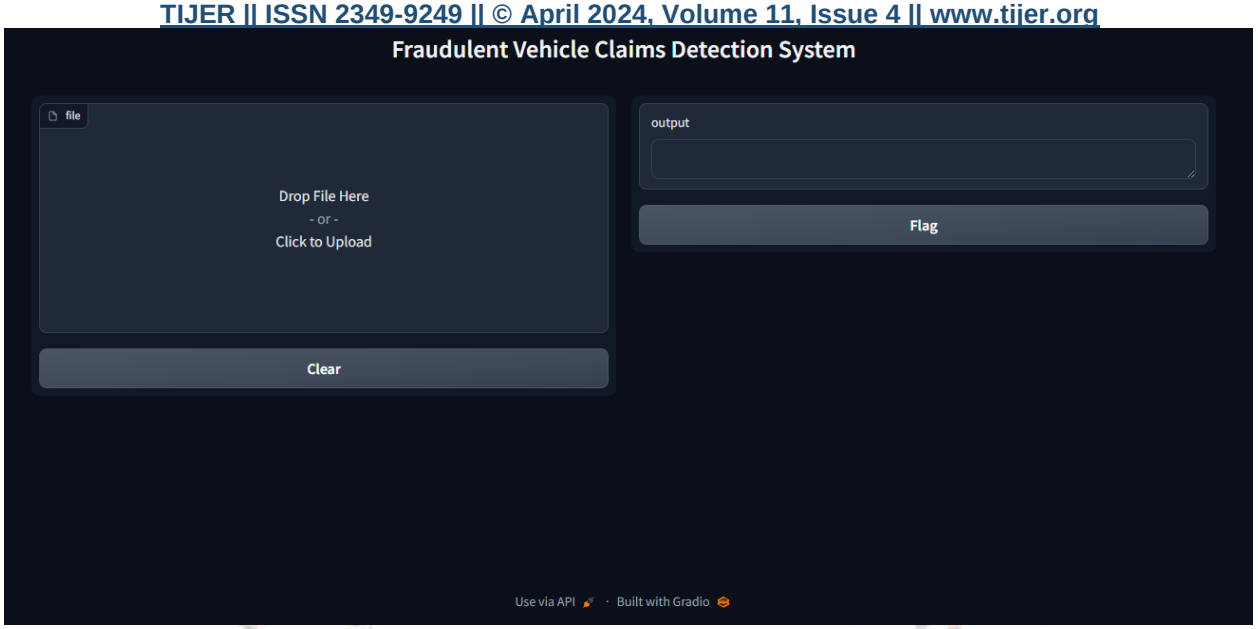



*Figure 17: fraudulent motor vehicle insurance claims detection system user interface*

The output file had columns for comparison of the prediction made by the machine learning model and a column with the final prediction as shown in the figure below

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | PolicyNumber | FraudFound_P | DecisionTree_Predic | KNN_Prediction | NaiveBayes_Predict | SVM_Prediction | Final_Prediction | |
| 2 | 1 | genuine | fraudulent | genuine | fraudulent | genuine | genuine | |
| 3 | 4 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 4 | 9 | genuine | genuine | genuine | fraudulent | genuine | genuine | |
| 5 | 15 | genuine | fraudulent | fraudulent | genuine | fraudulent | genuine | |
| 6 | 16 | genuine | fraudulent | fraudulent | fraudulent | genuine | genuine | |
| 7 | 18 | genuine | fraudulent | genuine | fraudulent | genuine | genuine | |
| 8 | 20 | genuine | genuine | genuine | fraudulent | fraudulent | genuine | |
| 9 | 28 | genuine | genuine | genuine | genuine | fraudulent | genuine | |
| 10 | 32 | genuine | fraudulent | genuine | fraudulent | fraudulent | genuine | |
| 11 | 34 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 12 | 36 | fraudulent | fraudulent | genuine | fraudulent | genuine | genuine | |
| 13 | 37 | genuine | genuine | fraudulent | fraudulent | fraudulent | genuine | |
| 14 | 40 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 15 | 42 | genuine | fraudulent | genuine | fraudulent | genuine | genuine | |
| 16 | 44 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 17 | 47 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 18 | 48 | genuine | genuine | genuine | genuine | fraudulent | genuine | |
| 19 | 51 | genuine | fraudulent | genuine | fraudulent | genuine | genuine | |
| 20 | 60 | genuine | fraudulent | genuine | fraudulent | fraudulent | genuine | |
| 21 | 62 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 22 | 64 | genuine | genuine | genuine | genuine | genuine | genuine | |
| 23 | 69 | genuine | genuine | genuine | fraudulent | genuine | genuine | |

*Figure 18: fraudulent motor vehicle insurance claims detection system output file*

## 4.6. STUDY DISCUSSIONS

Fraudulent activities in motor vehicle insurance pose significant financial risks to insurers and policyholders alike. Detecting such fraudulent claims is crucial for maintaining the integrity of insurance systems and preventing financial losses.

According to the literature review, (Tuggener et al., 2019) and (Aslam et al., 2022) emphasized the importance of accurate fraud detection methodologies, highlighted the challenges associated with deep neural networks and data availability. (Verma et al., 2017) proposed a comprehensive approach integrating association rule mining, clustering, and outlier detection for health insurance fraud detection.

(Belhadji et al., 2000) introduced a filter selection method using information gain and chi-squared, while (Sarker, 2021) proposed a supervised inductive learning approach utilizing information gain and gain ratio. (Moon et al., 2019) employed Support Vector Machine (SVM) and Decision Tree (DT) for feature selection in insurance reserve modeling.

This study used an ensemble feature selection model and ensemble machine learning model tailored specifically for fraudulent motor vehicle insurance claims. It employed multiple feature selection techniques, including information gain, gain ratio, and chi-square, to identify relevant features from the dataset. The selected features were then used to train a machine learning algorithms for motor vehicle fraud detection.

The ensemble feature selection approach extends the methodologies discussed in the literature by combining various feature selection techniques, thereby leveraging the strengths of each approach to improve overall performance. By leveraging ensemble techniques, the model was able to identify relevant features that capture unique patterns indicative of fraudulent behavior in motor vehicle insurance claims.

As earlier indicated in the performance evaluation table, the model attained a significant improved accuracy and precision with feature selected dataset as compared to full dataset.

As earlier demonstrated, when using feature selected data, the execution of the machine learning model was faster.

The ensemble approach further strengthened the predictive power of the model by combining the predictions of individual classifiers using soft voting. This resulted in a final prediction that reflected the consensus of multiple models. This ensemble strategy mitigated the limitations of individual classifiers and had an overall improved performance which resulted from leveraging the strengths of each model.

The Ensemble Feature Selection Model with Machine Learning Model represents a significant advancement in the fight against fraudulent motor vehicle insurance claims. Its robust performance, coupled with its practical applicability, positions it as a valuable tool for insurers seeking to safeguard their assets and uphold the trust of their policyholders.

# CHAPTER 5

## CONCLUSION AND RECOMMENDATIONS

### 5.1. Introduction

The detection of fraudulent motor vehicle insurance claims presents a significant challenge for insurers, necessitating the adoption of advanced technologies and methodologies. The main objective of this study was to design, develop and test a model for detecting whether a given motor vehicle insurance claim is fraudulent using ensemble feature selection model and ensemble machine learning model. A novel web-based application that uses ensemble feature selection techniques and ensemble machine learning model was developed from of this research. This system was able to classify motor vehicle insurance claims as genuine or fraudulent. This chapter provides a summary of the study findings on developing an Ensemble Feature Selection Model coupled with multiple Machine Learning Models for the detection of fraudulent motor vehicle insurance claims. Recommendations are provided for future research and industry implementation based on the study insights.

### 5.2. Summary of Findings

The study employed an Ensemble Feature Selection Model in conjunction with machine learning algorithms, including Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and Support Vector Machine (SVM), to detect fraudulent motor vehicle insurance claims. The output of these individual models was integrated using a soft voting approach to obtain the final prediction.

The study revealed that the ensemble approach outperformed individual models, with a higher accuracy rate and improved robustness in fraud detection. Notably, the combination of feature selection and ensemble modeling enhanced the overall performance of the fraud detection system, providing insurers with a reliable tool for identifying fraudulent claims.

### 5.3. Study Conclusion

The research has demonstrated the effectiveness of integrating feature selection techniques with machine learning models for the detection of fraudulent motor vehicle insurance claims. By leveraging ensemble modeling, superior performance was achieved as compared to individual classifiers, highlighting the importance of combining diverse algorithms to enhance fraud detection capabilities. The study contributes to the ongoing efforts to combat insurance fraud and protect the interests of insurers and policyholders alike.

### 5.4. Study Achievements

The study's main objective was to design and implement an ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims. Following that, a model that uses ensemble feature selection techniques for selecting features for use by machine learning techniques to predict and categorize motor vehicle insurance claims as genuine or fraudulent was developed. The performance of the model was evaluated. The final result was a web- based system that takes an input of the insurance claim data inform csv file and gives an csv output file that indicates whether the claims are genuine or fraudulent.

The study's specific objectives were, to investigate the feature selection techniques that can be used to identify features that can be used to build machine learning models for detecting fraudulent motor vehicle insurance claims, to explore machine learning techniques that are currently used detect fraudulent insurance claims, to design and implement an ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims, and finally to evaluate the performance of the ensemble feature selection model with machine learning that can be used to detect fraudulent motor vehicle insurance claims. The objectives were successfully achieved by first understanding the insurance industry's operations, specifically the motor vehicle insurance segment. Various sources, including data from insurance companies, were used to provide required information for the study. After getting the relevant data, an assessment of data quality was conducted. Since the information was collected in its raw form, a comprehensive data exploratory method was employed during the data preparation phase. It entailed filling up the missing data. SMOTE was used to deal with the issue of imbalanced data. As part data pre-processing an ensemble feature selection model that used information gain, gain ratio and chi- square was developed. This model selected a set of relevant features that were used by the machine learning model to predict fraudulent motor vehicle insurance claims. Using the feature selected dataset, the Machine learning classifiers were trained on 70% and tested on 30% of the data. Additionally, the machine learning classifiers were trained on 70% and tested on 30% of the full dataset. This was in order to evaluate the effect feature selection on the performance of machine learning model. Evaluation of the performance of the machine learning model with feature selected data and with full dataset was done. As explained in chapter 4, the performance results revealed that final prediction obtained from individual classifiers through soft voting was better with the feature selected dataset. The execution time the machine learning model was shorter with the feature selected dataset. The study resulted in development

of a model that uses ensemble feature selection techniques and ensemble machine learning model in order to classify a claim as genuine or fraudulent. Finally, all the objectives outlined in this study were entirely met.

### 5.5. Study Limitations

Despite the achievements of the study, there some challenges that were encountered. Due to privacy concerns the insurance companies were not willing to give the insurance claims data for the study. This data could have been more comprehensive and hence more beneficial to the study.

The computational complexity of feature selection and ensemble modeling techniques required substantial computational resources, posing challenges for scalability.

### 5.6. Recommendations

Based on our findings, the study recommends the following for future research and industry implementation:

1. Integration of Advanced Feature Selection Techniques: Advanced feature selection methods should be explored in order to improve identification of relevant features for use in machine learning models to detect fraudulent motor vehicle insurance claims should be done.

2. Optimization of Ensemble Models: Ensemble models should be further optimized to enhance scalability and efficiency. This will enable real-time fraud detection in large-scale insurance datasets.

3. Continuous Evaluation and Improvement: There should be a continuous evaluation and improvement framework of fraud detection systems. This framework should incorporate feedback from insurers and stakeholders in order to adopt evolving fraud schemes.

4. Collaboration and Knowledge Sharing: There should be a collaboration and knowledge sharing among insurers, researchers, and regulatory agencies to collectively address the challenges of insurance fraud and enhance industry-wide resilience.

### 5.7. Future Work

Detecting fraudulent motor vehicle insurance claims poses a significant challenge for the insurance industry. For the future work this study proposes enhancement this system by integrating machine learning approaches with nature- inspired optimization algorithms. Nature-inspired algorithms entails a variety of problem-solving approaches derived from natural phenomena(Rathore et al., 2020). This integration addresses the limitation of machine learning algorithms in handling extensive datasets. This will result to development of the more rapid and effective models for identifying false claims. Use of nature-inspired optimization algorithms will result to

automatic and dynamic discovery of the most pertinent features for use in machine learning in order to identify fraudulent insurance claims. This will result in more accurate classification.

Additionally, future research may also explore use of larger datasets spanning multiple years.

## REFERENCES

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402

Aslam, M.-F., Hunjra, Dr. A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance Fraud Detection: Evidence from Artificial Intelligence and Machine Learning. *Research in International Business and Finance*, *62*, 101744. https://doi.org/10.1016/j.ribaf.2022.101744

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., Chow, B. J., & Dwivedi, G. (2019). Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLOS ONE*, *14*(6), e0218760. https://doi.org/10.1371/journal.pone.0218760

Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems*, *150*. https://doi.org/10.1016/j.dss.2021.113492

Belhadji, E. B., Dionne, G., & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *25*(4), 517–538. https://doi.org/10.1111/1468-0440.00080

Bellatreche, L., Goyal, V., Fujita, H., Mondal, A., & Reddy, P. K. (2021). *Big Data Analytics: 8th International Conference, BDA 2020, Sonepat, India, December 15–18, 2020, Proceedings*. Springer Nature.

Bhowmik, R. (2008). Data Mining Techniques in Fraud Detection. *Journal of Digital Forensics, Security and Law*. https://doi.org/10.15394/jdfsl.2008.1040

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2014). Data classification using an ensemble of filters. *Neurocomputing*, *135*, 13–20. https://doi.org/10.1016/j.neucom.2013.03.067

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification And Regression Trees* (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Brownlee, J. (2016). *Machine Learning Mastery With Weka: Analyze Data, Develop Models, and Work Through Projects*. Machine Learning Mastery.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press.

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, *8*, 150360–150376. https://doi.org/10.1109/ACCESS.2020.3016715

*Feature selection by chi-squared*. (n.d.). ResearchGate. Retrieved August 20, 2023, from https://www.researchgate.net/figure/Feature-selection-by-chi-squared_tbl2_364083534

Firdaus, F., Zulfadilla, Z., & Caniago, F. (2021). Research Methodology: Types in the New Perspective. *MANAZHIM*, *3*(1), 1–16. https://doi.org/10.36088/manazhim.v3i1.903

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

Guru, D. S., Suhil, M., Pavithra, S. K., & Priya, G. R. (2018). Ensemble of Feature Selection Methods for Text Classification: An Analytical Study. In A. Abraham, P. Kr. Muhuri, A. K. Muda, & N. Gandhi (Eds.), *Intelligent Systems Design and Applications* (Vol. 736, pp. 337–349). Springer International Publishing. https://doi.org/10.1007/978-3-319-76348-4_33

Guyon, I., & Elisseeff, A. (n.d.). *An Introduction to Variable and Feature Selection*.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature Extraction: Foundations and Applications*. Springer.

Hassan, A. K. I., & Abraham, A. (n.d.). *Modeling Insurance Fraud Detection Using Ensemble Combining Classification*.

He, T., Baik, J. M., Kato, C., Yang, H., Fan, Z., Cham, J., & Zhang, L. (2022). Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data. *Frontiers in Genetics*, *13*, 821832. https://doi.org/10.3389/fgene.2022.821832

Hegde, R., V, A., Madival, S., S, S., & U, S. (2021). *A Review on Data Mining and Machine Learning Methods for Student Scholarship Prediction*. 923–927. https://doi.org/10.1109/ICCMC51019.2021.9418376

Honghong, S., & Lili, H. (2017). A Binary Approximate Naive Bayesian Classification Algorithm Based on SOM Neural Network Clustering. *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*, 1344–1347. https://doi.org/10.1109/ICCSEC.2017.8446854

Houari, R., Bounceur, A., Tari, A. K., & Kecha, M. T. (2014). Handling Missing Data Problems with Sampling Methods. *2014 International Conference on Advanced Networking Distributed Systems and Applications*, 99–104. https://doi.org/10.1109/INDS.2014.25

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–4. https://doi.org/10.1109/ICDS47004.2019.8942277

Kgare, M. (n.d.). *Predicting Lapse Rate in Life Insurance Using Machine Learning Algorithms*.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer Science & Business Media.

Liu, H., & Motoda, H. (2012). *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Springer Science & Business Media.

Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. *Procedia - Social and Behavioral Sciences*, *97*, 320–324. https://doi.org/10.1016/j.sbspro.2013.10.240

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.

Moon, H., Pu, Y., & Ceglia, C. (2019). A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters*, *9*(6), Article 6. https://doi.org/10.4236/tel.2019.96120

Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G., & Sciacca, V. (2020). *Machine Learning, Optimization, and Data Science: 5th International Conference, LOD 2019, Siena, Italy, September 10–13, 2019, Proceedings*. Springer Nature.

Nielsen, J. P., Asimit, A., & Kyriakou, I. (2020). *Machine Learning in Insurance*. MDPI.

Njoh-Paul, I. (n.d.). *A Comparative Study of Ensemble Techniques and Individual Classifiers in Predicting Insurance Claim*.

Patil, V. (2023). Fraud Detection and Analysis for Insurance Claim Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *11*(5), 5559–5565. https://doi.org/10.22214/ijraset.2023.52875

Patnaik, S., Yang, X.-S., & Nakamatsu, K. (2017). *Nature-Inspired Computing and Optimization: Theory and Applications*. Springer.

Pes, B. (2020). Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. *Neural Computing and Applications*, *32*(10), 5951–5973. https://doi.org/10.1007/s00521-019-04082-3

Piao, Y., & Ryu, K. H. (2017). A Hybrid Feature Selection Method Based on Symmetrical Uncertainty and Support Vector Machine for High-Dimensional Data Classification. In N. T. Nguyen, S. Tojo, L. M. Nguyen, & B. Trawiński (Eds.), *Intelligent Information and Database Systems* (Vol. 10191, pp. 721–727). Springer International Publishing. https://doi.org/10.1007/978-3-319-54472-4_67

*Quarter 4 2022 Industry Release 06-03-2023-1.pdf*. (n.d.). Retrieved July 20, 2023, from https://www.ira.go.ke/images/Q4/Quarter%204%202022%20Industry%20Release%2006-03-2023-1.pdf

Raghavan, P., & Gayar, N. E. (2019). Fraud Detection using Machine Learning and Deep Learning. *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 334–339. https://doi.org/10.1109/ICCIKE47802.2019.9004231

Rathore, V. S., Dey, N., Piuri, V., Babo, R., Polkowski, Z., & Tavares, J. M. R. S. (2020). *Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2020*. Springer Nature.

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 1–6. https://doi.org/10.1109/ICCPCT.2017.8074258

S, T. K., Deep, U., Shoiab, S., Atif, S., Bhatnagar, T., & T, R. (2021). Insurance Fraud Detection Using Machine Learning. *International Journal of Advanced Information and Communication Technology*, 1–4. https://doi.org/10.46532/ijaict-2020210101

Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-3207-1

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074. https://doi.org/10.1016/j.mlwa.2021.100074

Subudhi, S., & Panigrahi, S. (2018). Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques: *International Journal of Rough Sets and Data Analysis*, *5*(3), 1–20. https://doi.org/10.4018/IJRSDA.2018070101

Taha, A., Cosgrave, B., & Mckeever, S. (2022a). Using Feature Selection with Machine Learning for Generation of Insurance Insights. *Applied Sciences*, *12*(6), 3209. https://doi.org/10.3390/app12063209

Taha, A., Cosgrave, B., & Mckeever, S. (2022b). Using Feature Selection with Machine Learning for Generation of Insurance Insights. *Applied Sciences*, *12*(6), Article 6. https://doi.org/10.3390/app12063209

Taha, A., Cosgrave, B., & Mckeever, S. (2022c). Using Feature Selection with Machine Learning for Generation of Insurance Insights. *Applied Sciences*, *12*(6), 3209. https://doi.org/10.3390/app12063209

Tuggener, L., Amirian, M., Rombach, K., Lorwald, S., Varlet, A., Westermann, C., & Stadelmann, T. (2019). Automated Machine Learning in Practice: State of the Art and Recent Results. *2019 6th Swiss Conference on Data Science (SDS)*, 31–36. https://doi.org/10.1109/SDS.2019.00-11

Tuv, E. (n.d.). *Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination*.

Ürgenç, S., Kaplan, H., & Pehlivanl, A. Ç. (2022). *Fraud Detection with Machine Learning in Property Insurance Policy Requests*.

Vajiram, J., & Senthil, N. (n.d.). *Correlating Medi- Claim Service by Deep Learning Neural Networks.*

Verma, A., Taneja, A., & Arora, A. (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. *2017 Tenth International Conference on Contemporary Computing (IC3)*, 1–7. https://doi.org/10.1109/IC3.2017.8284299

Vosseler, A. (2022). Unsupervised Insurance Fraud Prediction Based on Anomaly Detector Ensembles. *Risks*, *10*(7), 1–20.

Wang, J., Xu, J., Zhao, C., Peng, Y., & Wang, H. (2019). An ensemble feature selection method for high-dimensional data based on sort aggregation. *Systems Science & Control Engineering*, *7*(2), 32–39. https://doi.org/10.1080/21642583.2019.1620658

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science.