AN ENSEMBLE FEATURE SELECTION MODEL WITH MACHINE LEARNING MODEL FOR DETECTION OF FRAUDULENT MOTOR VEHICLE INSURANCE CLAIMS

Anthony Mwiti Wambu Bsc (Computer Science)

J57/37307/2017

Sign……..……..Date……..

Department of Computing and Information Sciences

A research project submitted in partial fulfillment of the requirements for the award of the degree of Master of Science in Computer Science in the School of Pure and Applied Sciences of Kenyatta University

Supervisor

Dr. Eric Araka
Department of Computing and Information Sciences
Kenyatta University

Signature…….. Date…….

May 2025

DECLARATION

I declare that this research proposal is my original work and has not been presented in any other university/institution for consideration of any certification. This research proposal has beencomplemented by referenced sources duly acknowledged. Where text, graphics pictures, figures, or tables have been borrowed from other sources, including the internet, these are specifically accredited and references cited using the current APA system and in accordance with anti- plagiarism regulations.

Signature…….. Date……

Anthony Mwiti Wambu

Computing and Information Science

Supervisor's declaration: This research proposal has been submitted for appraisal with my approval as University Supervisor.

Signature…….. Date……

Dr. Eric Araka

TABLE OF CONTENTS

DECLARATION ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF FIGURES ..... **vii**

LIST OF TABLES ..... **viii**

ABSTRACT ..... **ix**

LIST OF ABBREVIATIONS AND ACRONYMS ..... **x**

CHAPTER 1: INTRODUCTION **1**

1.0 Introduction 1

1.1 Background of the study 1

1.2 Statement of the Problem 3

1.3 Objectives 4

1.3.1 General Objective 4

1.3.2 Specific Objectives of the study 4

1.4 Research Questions 4

1.5 Justification 5

1.6 Significance of the study 5

1.7 Scope 6

1.8 Limitations 6

CHAPTER 2: LITERATURE REVIEW **7**

2.0 Introduction 7

2.1 Feature Selection Techniques 7

2.1.1 Wrapper methods 8

2.1.2 Filter methods 8

2.1.3 Embedded methods 8

2.1.4 Ensemble methods 9

2.1.5 Information Gain 10

2.1.6 Gain Ratio 12

2.1.7 Chi- 2 13

2.2 Machine Learning Techniques 13

2.2.1 Decision Trees 14

2.2.2 K- Nearest Neighbor (KNN) 15

2.2.3 Support Vector Machine (SVM) 17

2.2.4 Naïve Bayes (NB) Algorithm 18

2.3 Related Work 19

2.4 Research Gaps 23

2.5 Conceptual Model 24

CHAPTER 3: METHODOLOGY 32

3.0 Introduction 32

3.1 The Research Design 32

3.1.1 CRISP-DM Methodology 33

3.1.1.1 Understanding the business 34

3.1.1.2 Data assessment 35

3.1.1.3 Preparation of the data 39

3.1.1.4 Modelling 44

3.1.1.5 Evaluation 45

3.1.1.6 Deployment 47

CHAPTER 4: RESULTS AND ISCUSSIONS 48

4.0 Introduction:..... 48

4.1 Data Exploratory Analysis 48

4.2 Multiple Feature selection model evaluation 49

4.2.1 Mutual information 49

4.2.2Gain ratio 51

4.2.3 Chi-Square 53

4.2.3.1 Final set of features..... 55

4.3 Machine Learning Model Evaluation 57

4.4 Performance Evaluation and Results 58

4.4.1 Performance evaluation using feature selected dataset..... 58

4.4.2 Performance evaluation using full dataset 60

4.5 Fraudulent Vehicle Claims Detection System 62

4.6 Study Discussions 63

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS..... 66

5.0 Introduction..... 66

5.1 Findings Summary..... 66

5.2 Study Conclusion..... 67

5.3 Achievements of the research 68

5.4 Limitations of the study 70

5.5 Recommendations..... 70

5.6 Future Work..... 71

REFERENCES..... 72

APPENDICES..... 79

Appendix I: Gantt chart 79

Appendix II: Project Budget 80

Appendix III: Project Code .....**Error! Bookmark not defined.**

Appendix IV: Research Project Proposal Approval 81

LIST OF FIGURES

Figure 2.1: Machine learning algorithms classification (Fatima et al., 2020)..... 14
Figure 2.2: A Decision Tree (Hegde et al., 2021). 15
Figure 2.3: K-Nearest Neighbor (Hegde et al., 2021)..... 16
Figure 2.4: SVM analysis (Hegde et al., 2021)..... 17
Figure 2.5: Conceptual model 24
Figure 3.1: Methodology Diagram- CRISP DM..... 34
Figure 3.2: CSV file extract of motor vehicle insurance claims dataset. 35
Figure 3.3: Motor vehicle insurance claims dataset, data distribution. 36
Figure 3.4: Datatypes for the dataset..... 37
Figure 4.1: Feature selection with information gain 49
Figure 4.2: Feature selection with gain ratio 51
Figure 4.3: Feature selection with chi-square..... 53
Figure 4.4: Comparison of Features selected by IG, GR and chi-2 feature selection methods..... 56
Figure 4.5: Fraudulent motor vehicle insurance claims detection system user interface 62
Figure 4.6 : Fraudulent motor vehicle insurance claims detection system output file 63

LIST OF TABLES

Table 3.1 : Number of Columns of dataset with Null Values 38
Table 4.1 : Feature selected Dataset model's Evaluation Report 58
Table 4.2 : Full dataset model's Evaluation Report 60

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. Despite the insurance industry having extensive motor vehicle policy data and claims information, fraudulent claims remain a significant challenge in motor vehicle insurance. Proper analysis of this data can result in development of more efficient methods for identifying fraudulent claims. The challenge lies on how to extract valuable insights and knowledge from this data. This is because insurance datasets inherently include noisy features or low-quality subsets of data. This study used feature selection techniques to select relevant features from motor vehicle insurance claim dataset. The selected features were then used in training machine learning model. The machine learning model consisted of multiple machine learning algorithms whose individual prediction results were combined by use of a voting method. This helped to improve classification performance. Machine learning model's performance with feature selected dataset and with full dataset was then evaluated using recall, precision and F1-score. The results indicated that the model trained with feature selected dataset performed better than the model trained with full dataset attaining higher values in recall, precision and F1-score. This indicated improved capability in minimizing false negative and improved overall effectiveness in fraud detection. For feature work the model developed for detecting fraudulent motor vehicle insurance claims can be enhanced by integrating machine learning techniques with nature-inspired optimization algorithms. This will help in better handling of extensive datasets and result to development of more rapid and effective models for identifying false claims.

Keywords: Data Mining, Machine Learning, Feature Selection, ensemble multiple filter feature selection method,SMOTE.

LIST OF ABBREVIATIONS AND ACRONYMS

FS	Feature Selection
ML	Machine Learning
DT	Decision Tree
GR	Gain Ratio
CRISP-DM	Cross Industry Standard Process for Data Mining.
IG	Information gain
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
SMOTE	Synthetic Minority Oversampling Technique.

CHAPTER 1: INTRODUCTION

1.0 Introduction

This chapter covers the background of the study, problem statement, objectives of the study, research questions, justification, scope, and limitations. The chapter begins 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 is identified in the problem statement, which is further explored through objectives of the study, research questions, and a discussion of the study's scope and limitations.

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 thanwhat's rightfully due (Baesens et al., 2021a). 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 contributes to national economies in a major way (Roy & George, 2017). In Kenya, for instance,the Insurance Regulatory Authority reported KES 195.2 billion worth of gross written premium in 2022, with annual premium growth being a consistent trend (*Insurance Industry Quarterly Claims Statistics for the Period*, 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 such as actuarial work, analysis of customer behavior and in detection of fraud (Firdaus et al., 2021). Actuaries often employ domain-specific modelsdue to data complexities involved. General ML methods are mostly used for detecting fraud and analyzing the customer behavior. This allows for the adaptation of advancements from other sectors (Subudhi & Panigrahi, 2018).

Quality of training datasets significantly impacts the performance of supervised ML models. Insurance datasets often contain redundant and irrelevant attributes which tend to hindermodel 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 FS methods can improve model performance. By grouping weak features and identifying those that have strong association with the output variable, effectiveness of the models can be enhanced (Guyon & Elisseeff, n.d.).

This study employed an ensemble approach, utilizing multiple FS techniques and multiple ML techniques. Ensemble feature selection method entailed combining filter feature selection method, that is, IG, GR, and chi-square, to identify important features for use by the multiple machine learning algorithms. The output from these multiple machine learning

algorithms were combined using a voting algorithm which grouped the insurance claims as either fraudulent or legitimate.

1.2 Statement of the Problem

Detecting fraud in motor vehicle insurance claims remains a big challenge to insurance companies globally. This results in significant financial losses and reputational harm to these companies (Patil, 2023). Traditional methods of detecting fraud are not able to accurately identify these fraudulentclaims due to evolving fraud techniques and complex data patterns (Aslam et al., 2022). To address the issue, ML-based approaches have been adopted. However, their effectiveness is impacted by the complexity and noise within insurance data. This emphasizes the need for carrying out feature selection. 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). In order to enhance the accuracy of ML models in identifying fraudulent claims various feature selection strategies have been employed. These FS techniques can be filter-based, wrapper, or embedded methods (Piao & Ryu, 2017). However, each approach has its limitations, including challenges related to feature interactions, computational complexity, and dependency on specific machinelearning 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 study's primary aim was to design, develop and test a model for detecting whether a given motor vehicle insurance claim is fraudulent.

1.3.2 Specific Objectives of the study

1. To establish the FS techniques that can be used to come up with features that can be employed to build ML models for detecting vehicle insurance claims that are not genuine.
2. To explore ML techniques that are currently used detect fraudulent insurance claims.
3. To create and implement an ensemble FS model with MLthat can be used for identifying vehicle insurance claims that are not genuine.
4. To evaluate the working of the ensemble FS model with ML algorithms that can be employed to identify vehicle insurance claims that are not genuine.

1.4 Research Questions

1. Which FS techniques that can be used to identify features for building ML models for detecting fraudulent motor vehicle insurance claims?
2. Which ML techniques that are currently used detect fraudulent insurance claims?
3. How can an ensemble FS model with ML that can be used todetect fraudulent motor vehicle insurance claims be developed and implemented?

4. How effective is an ensemble feature selection model with machine learning in detectingfraudulent 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 FS techniques and multiple ML algorithms. At its core, the model acts as an accurate data-driven decision-making tool that will help the insurance sector, motor vehicle domain, to evaluate the authenticity of the claims. The model has the capability of revolutionizing strategies for detecting fraud in the motor vehicle insurance sector due to its accuracy, efficiency and transparency. This can result to an insurance ecosystem that is more secure and trustworthy hence safeguarding the financial interests of theinsurance companies as well as offering satisfaction to the policyholders.

1.6 Significance of the study

Fraud in motor vehicle insurance filings cause insurance companies huge financial burden. This makes the insurance companies have strained resources and hence increase cost of insurance premiums. By use of ensemble FS and multiple ML algorithms, this study provides a platform for detection of fraudulent motor vehicle insurance claims more effectively and accurately. This will help the motor vehicle insurance companies make better data driven decisions, reducing operational costs and be able to process claims more effectively hence benefiting both the insurance companies and the policy holders.

1.7 Scope

This study employed an ensemble of multiple FS techniques to minimize the quantity of dataset features by discarding features that are noisy and irrelevant. To enable detection of fraudulent claims in motor vehicle insurance, various ML techniques were used. The predictive outputs from the ML algorithms were combined using a voting algorithm in order to come up with final output. The study made use of online available datasets for motor vehicleinsurance claims fraud from Kaggle dataset (www.kaggle.com).

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 for model training. 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 covers a review of existing work in relation to fraudulent motor vehicle insurance claims detection. First, it explores various featureselection techniques their pros and cons, how they were used to build an ensemble FSmodel. Second, it explores ML techniques that were employed to effectively identify fraudulent insurance claims. Then the chapter reviews related work by other researchers and identifies existing approaches and methodologies used in similar studies. Finally, it shows research gap by discussing the limitations of existing studies and justifies the need for the proposed model. This chapter forms a basis of conceptual framework for the research.

2.1 Feature Selection (FS) Techniques

FS refers to the process within machine learning where a small set of variables that are most relevant is selected a dataset. This serves to eliminate irrelevant and redundant features, thereby mitigating overfitting, enhancing interpretability, and reducing computational complexity ofmodels (Cai et al., 2018). When selecting a FS technique various aspects are considered, such as, the nature of the problem, characteristics of dataset, and the ML algorithm employed. An effective technique should prioritize simplicity, interpretability and accuracy of themodel (Taha et al., 2022a). FS methods can be grouped in different categories depending on how they interact with ML algorithm and how they evaluate the features. These categories can be, filter methods, wrapper methods, embedded methods, or ensemble methods.

2.1.1 Wrapper methods

Wrapper methods refer to a category of FS methods that work by training and assessing a model with various feature subsets and then the one that achieves the best performance is selected (Piao & Ryu, 2017). They are dependent of the ML algorithm, which may be supervised or unsupervised (Taha et al., 2022a). Wrapper methods have high performance measures but they take too long to run (Y. Wang et al., 2022). They are also restricted to a specific learning algorithm. Forwardselection, backward elimination and Bi-directional elimination are examples of wrapper methods (Njoh-Paul, n.d.).

2.1.2 Filter methods

Filter methods are a category of FS techniques that utilize statistical measureslike IG, GR, chi-square or correlation to rank features according to their relevance in determining the research goal (Y. Wang et al., 2022).

Filter methods are not dependent on specific ML algorithm, unlike the wrapper methods, since 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 (Dr.K.K.Savitha, 2023).

2.1.3 Embedded methods

Embedded methods represent a class of FS techniques that integrate FS into the model training process (J. Wang et al., 2019). These methods work by modifying the algorithm in use to incorporate FS in both the model training and optimization processes. Embedded methods integrate aspects of 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., 2022a). Examples of embedded feature selection methods are random forest, gradient boosting, and DT. These are tree-based ML algorithms (Pes, 2020).

2.1.4 Ensemble methods

These is a category of FS methods which works by combining individual FS techniques to collectively come up with subset of features to be used by machine learning (Duboue, 2020). This helps to address the limitation of the individual methods while simultaneously leveraging their strengths. The ensemble methods result in improved feature selection which leads to enhanced effectiveness and efficiency of the ML algorithm (J. Wang et al., 2019). Ensemble methods are grouped into different categories based on how they work, as explained below:

2.1.4.1 Stability selection

Stability selection is an ensemble FS method which operates by applying a FS 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 (Kuhn & Johnson, 2019).

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 every repetition a new feature is included in the subset based on its individual selection performance. The features selected most frequently acrossthe iterations forms the final subset (Bolón-Canedo & Alonso-Betanzos, 2018).

2.1.4.3 Voting-Based Ensembles

A voting-based is an ensemble FS technique which works by combining the decisions of multiple individual FS methods using a voting mechanism. The features with most votes are considered important and they are selected form the final subset (Galli, 2020).

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 FS techniques to form the final subset of features (Dong & Liu, 2018).

2.1.4.5 Genetic Algorithms

Genetic algorithms are ensemble feature selection methods which work by treating feature subsets as individuals in a population. In order to come up with final subset of features, selection, crossover and mutation operations are perfomed on the feature subsets (Brownlee, 2020).

This research used an ensemble multiple filter FS techniques that combined output of IG algorithm, GR and chi-square in order to achieve their combined capability in selecting features for use in machine learning.

2.1.5 Information Gain

This is a filter-based FS technique which operates on information theory principle. The information theory operates by minimizing the uncertainty in identifying the class attributes when feature value is not known.

Uncertainty associated with each feature in determining the output is obtained by computing the entropy value of the distribution (Awan et al., 2019).

In a case of an attribute x the entropy value is derived as below:

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

H(X): entropy of the random attribute X

$P(x_i)$: probability of event xi occurring.

Σ: denotes the sum over all possible events xi

log2: represents the base 2 logarithm.

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

$$H(X|Y) = - \Sigma [P(y_j) * \Sigma [P (x_{i\,j}| y_j) * \log2(P (x_{ij} | y_j))]]$$

H(X|Y): conditional entropy of random variable X given random variable Y

$P(y_j)$: probability of event y_j occurring for random variable Y.

$P(x_{ij} | y_j)$: conditional probability of event x_{ij} occurring for random variable X given event y_j of random variable (Y).

The conditional entropy H(X|Y) helps to know how much uncertainty remains in the distribution of Xwhen Y is known. That is, helps to measures the average amount of information needed to determine the valueof X given the value of Y.

Information gain can be derived from the conditional entropy of (X) given (Y), that is, H(X|Y) as shown below:

Information gain (IG) = H(X) - H(X|Y)

H(X): entropy of random variable X prior to observing variable Y.

H(X|Y): conditional entropy of random variable X after considering variable Y.

In this case the enhancement in predictive power gained by considering the value of variable Y when predicting the outcomes of variable X, is the Information gain. This results to more informed and accurate predictions (He et al., 2022).

2.1.6 Gain Ratio

Use result to biasness towards feature with large diversity values. In order to counterbalance the biasness, this research used gain ratio. GR value increases when data is uniformly distributed across and it decreases when the data is concentrated inone branch of the attribute. Gain ratio considers quantity and size of branches, hence accommodating the inherent information within the dataset. The inherent information of a given feature is obtained by evaluated entropy of it's distribution (Duboue, 2020).

To obtain GR of given feature X with a feature value Y the following equation can be used:

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) = $- \Sigma [(jS_{ij} / jS_j) * \log2(jS_{ij} / jS_j)]$

jS_j: number of possible outcomes of feature X can take

S_{ij}: number of actual outcomes of feature X (Bolón-Canedo et al., 2014).

2.1.7 Chi- 2

Chi-2 is a statistical metric used to assess the degree of independence or correlation 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 chi-2 score indicates that the variable is strongly associated with the output class in a statistically significant way (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-2:

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

Where:

χ^2 is the chi-2 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*, 2023)

2.2 Machine Learning Techniques

ML is a category of AI which enables computer systems to learn from data, improve performance, and formulate decisions without requiring specific programming (Hegde et al., 2021). ML can be categorize as supervised learning, unsupervised learning, semi-supervised learning, or reinforcement learning (Kuhn & Johnson, 2013). Supervised

learning involves training and testing ML models with data that has been data. This helps to achieve accurate prediction through training with labeled datasets (Breiman et al., 2017). In unsupervised learning, ML algorithms are trained and tested using unlabeled data (Molnar, 2020), while in reinforcement learning machine learning algorithms learns decision- making strategies through interaction with the environment (Mohamad & Tasir, 2013). ML encompasses various models, algorithms, and learning systems (Tuggener et al., 2019). ML algorithms enable autonomous learning and decision-making (Hegde et al., 2021). Data mining, utilizing ML techniques, has widespread applications across various domains (Kuhn & Johnson, 2013).

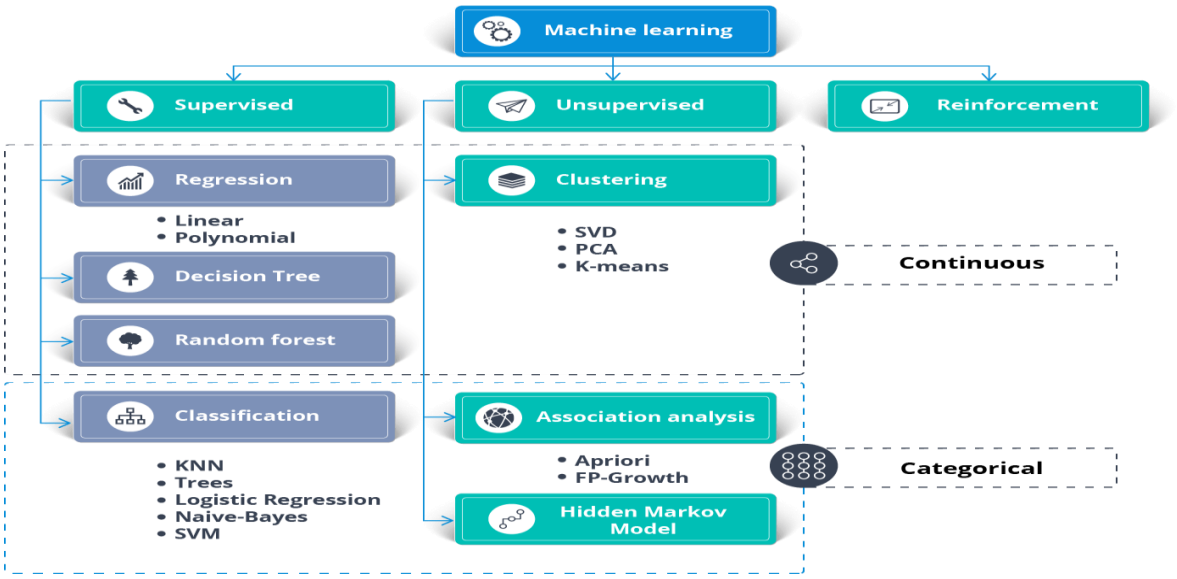


Figure 2.1: ML classification (Fatima et al., 2020)

2.2.1 Decision Trees

As a supervised ML predictive model. In a DT, internal nodes represent tests on an attributes or features, each branch reflects a possible outcome of the test, and leaf nodes reflects predicted class label or decision, as highlighted by (Witten et al., 2016). The tree is

formed through iterative dividing of dataset into smaller subsets according to the attribute that is most informative, a process that continues until all data is classified.

DT algorithms are advantageous due to their interpretability, simplicity of implementation, and flexibility in handling both categorical and numerical data types. However, they are prone to overfitting, especially when the model becomes excessively complex or when the training data is noisy, as noted by(Breiman et al., 2017).

Various algorithms are utilized to construct DT, including ID3, C4.5, and CART. These algorithms differ in their criteria for choosing the most suitable attribute to split on and addressing missing data (Molnar, 2020).

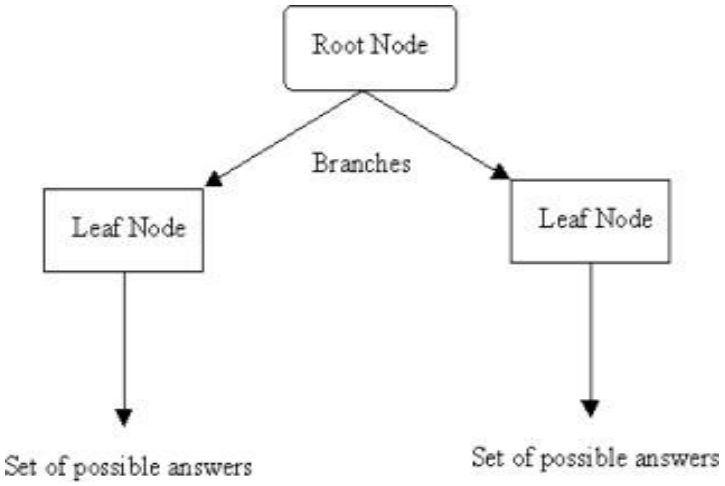


Figure 2.2: A DT (Hegde et al., 2021).

2.2.2 K- Nearest Neighbor (KNN)

KNN is a supervised ML algorithm that classifies an unseen data point by examining the k closest data points from the training set. Parameter 'k' shows the count of neighbors to take into consideration. The new data point is assigned the class that appears most often among its

k nearest neighbors (Nicosia et al., 2020).

In KNN regression, predictions for new data points are made by locating the k closest neighbors in the training set and calculating the average of their associated output values. KNN has great simplicity and flexibility, and it can handle different types of decision boundaries. However, its effectiveness depends on the distance metric used to compare data points and the appropriate tuning of the hyper-parameter k. Choosing the right value for k helps is important since it helps to achieve a balance between the model's bias and variance. (Witten et al., 2016). Additionally, KNN can encounter computational challenges when dealing with large datasets because it requires to examine all training instances to identify the k closest neighbors for each new instance (Patnaik et al., 2017).

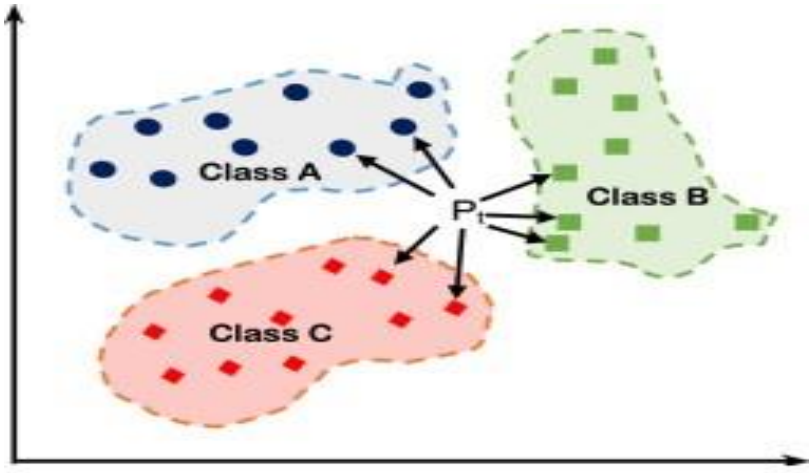


Figure 2.3: KNN (Hegde et al., 2021)

2.2.3 Support Vector Machine (SVM)

The SVM is a supervised ML technique that is used to handle classification and regression tasks. It's most effective in handling binary and multi-class tasks. The algorithm works by coming up with a decision boundary, known as a hyperplane, that distinguishes data points belonging to different classes. SVM then targets to maximize the margin between this decision boundary and the closest points of each class, which are called support vectors (Baesens et al., 2021b).

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).

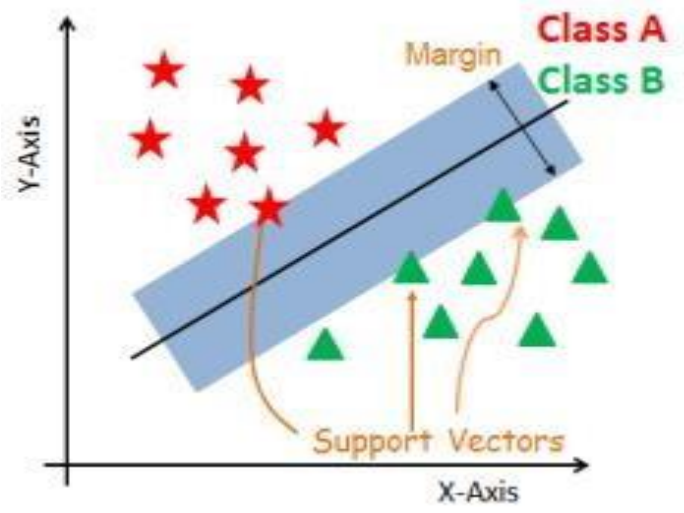


Figure 2.4: SVM analysis (Hegde et al., 2021)

2.2.4 Naïve Bayes (NB) Algorithm

Naive Bayes is a ML algorithm that works on the basis of Bayes' theorem. It 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 acquires knowledge of the probability distributions of both classes and features from the training dataset.

In the prediction phase, it computes the probability of each class for a given feature set and selects the class with the highest probability (Witten et al., 2016).

Considering X and Y as random variables,

P(Y) is prior probability of Y,

P(Y|X) is the posteriorprobability 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, ML techniques are widely applied due to data availability in terms of diversity and quantity and also due to the fact that factors that determinethe suitable reserve rate are too complicated to be modelled using a linear function (Taha et al., 2022a).

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's frequency. Claim severity is typically modeled using gamma distributed, while claim's frequency follows Poisson distribution (Itri et al., 2019).

(Al-Hashedi & Magalingam, 2021) explored Generalized Additive Models (GAM) which are more superior than GLMs in calculation of non-linear relationships and hence could perform betterin 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 insurancedatasets 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 chainladder 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 are also applicable to a wider variety of data andto reserving claims on an individual basis (Severino & Peng, 2021).

In order to detect insurance fraud, ML techniques are employed. In most cases, identifying insurance fraud is seen as a classification challenge that needs supervised learning models to assess whether 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 (Bellatreche et al., 2021). Deep learning, text mining, and unsupervised learning were offered as potential solutions by (Belhadji et al., 2000) to the issue of class imbalances. For the purpose of discovering motor vehicle insurance claims that are not genuine, (Moon et al., 2019) suggested a model thatcombines text mining with deep learning.

Verma et al. (2017) introduced a model for detecting fraud in health sector insurance claims. The model employed three mining methods that is, Association Rule Mining which analyzes data correlations to identify frequent patterns, K-Means Clustering which enhances

outlier detection, reduces time complexity, and increases performance in exposing insurance claim frauds. Fraudulent behavior in the study was classified as either period-based anomalies, disease-based anomalies or claim related anomalies.

The effectiveness of ML algorithms depends on how pertinent the chosen features are. Several research have used FS methods to pick a subset of featuresfrom the main collection of features. This aids in making the machine algorithm perform morequickly and precisely (Taha et al., 2022a). It can be difficult to choose the appropriate featuresfor 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-2. The model then employed the decision tree classifier called C 4.5 and a Bayesian network to identify fraudulentinsurance claims. Despite the model's accuracy being the same, the findings demonstrated that FS approaches enhanced the model's overall efficiency.

Sarkar et al. (2018) created a model that applied supervised inductive learning methodology to identify insurance fraud. The model made use of ensemble and monolithic FS methods. The model used IG, gain ratio, and Group Method for Data Handling (GMDH) to rank features during the pre-processing stage. SVM, 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 (Kelleher et al., 2015) as a way of removing irrelevant feature for an insurance fraud detection model. The model used neuro tree to achieved higher accuracy.

So as 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 (Corea, 2017).Wrapper, filter, and clustering algorithms are combined in the model's operation to determinethe multiple-weight of each feature.

A study done by (Ürgenç et al., 2022) proposed use of filter feature selection methods for usein 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 relevantinformation 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 FS techniques are swift compared to other FS techniques (Taha et al., 2022b). They

rank features separately based on how useful they are for predictingthe output class (Awan et al., 2019). Contrary to earlier suggested methods, this study used anensemble multiple filter FS method that puts together the results from the IG algorithm, GR, and chi-2 to form a final set of features that were used by ML 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 techniquesand feature selection techniques.

1. How to do effective feature selection in data preprocessing for fraudulent motor insuranceclaims 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 Model

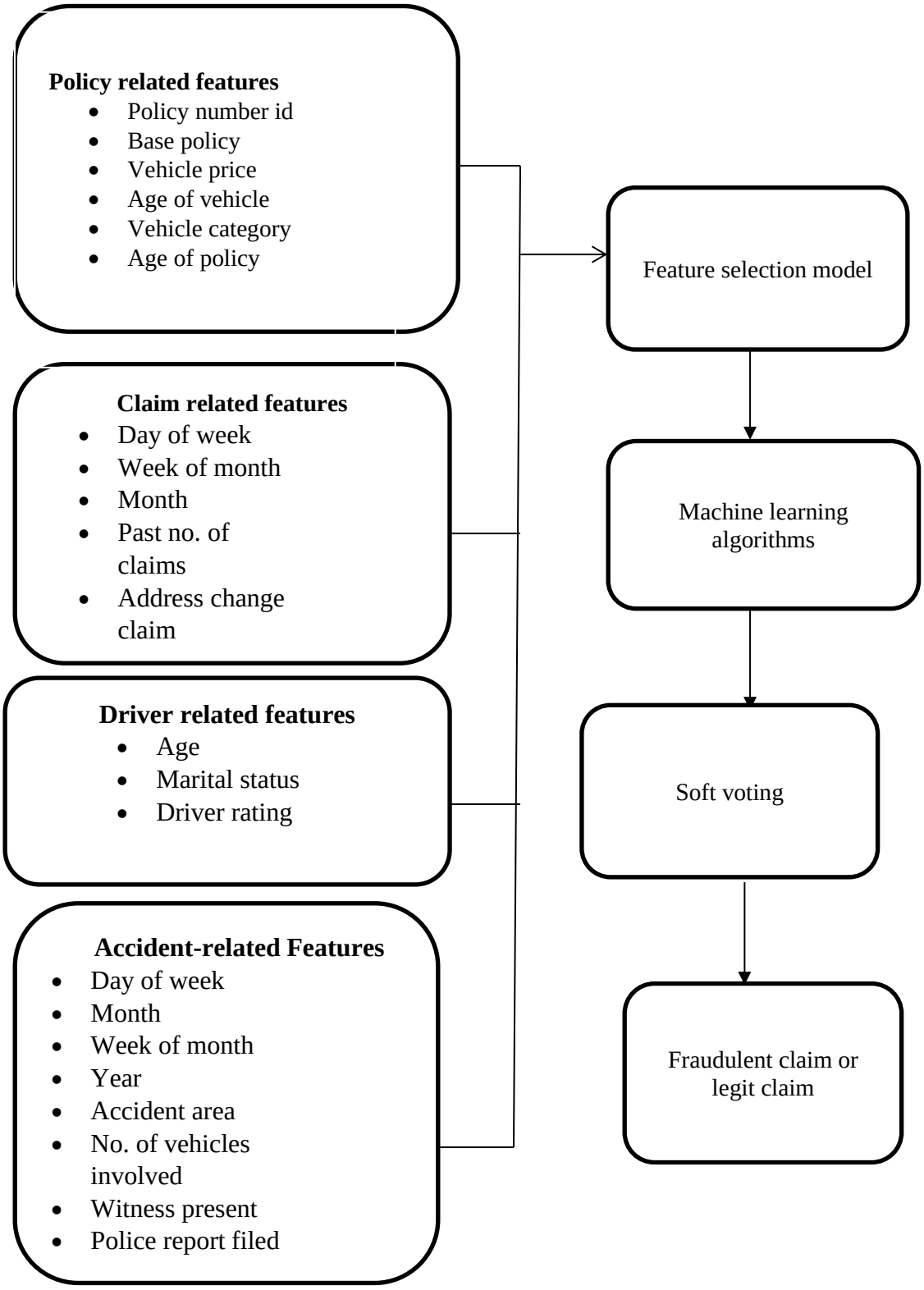


Figure 2.5: Conceptual model

CHAPTER 3: METHODOLOGY

3.0 Introduction

In this chapter, the research methodology is outlined, which involved employing ensemble multiple filter FS techniques and multiple ML algorithms and the procedures that were followed so as to build an effective model that could detect fraudulent motor vehicle insurance claims.

3.1 The Research Design

Research design serves as functions as a structure that offers direction to a researcher on how to conduct the research work and achieve the project's objectives. A research design is selected based on the area of study, research objectives, availability of data and tools (Firdaus et al., 2021).

This study employed a mixed methodology approach whereby a number of design methods were used. This research used quantitative experimental research design in gathering, model training, evaluation, and analysis. The process involved use of numerical data. The experimental research design used aimed at identifying relevant features for use by ML algorithms in identifying fraudulent motor vehicle insurance claims.

This study used CRISP-DM methodology so as to achieve all the goals of the research and be able to deploy the final model.

3.1.1 CRISP-DM Methodology

This study employed CRISP-DM methodology. This methodology is highly regarded in the areas of data mining and data analysis, due to its adaptability, and its comprehensive approach to data mining project management. CRISP-DM 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 stages ofa data mining project, detailing the tasks associated with each phase, and illustrating the interconnections of these tasks, hence offering a holistic view of the data mining life cycle.

CRISP-DM methodology consists six 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. Understanding the business – This phase aims to grasp the needs for the business
2. Data assessment – The phase aims to identify the necessary data and assess its sufficiency.
3. Preparing the data – This deals with organization of data for modelling.
4. Modelling – This phase deals with modelling techniques and how they are applied in theproject.
5. Evaluation – This phase deals with model evaluation to check whether it meets thebusiness objectives.
6. Deployment – This phase deals with how the results are accessed.

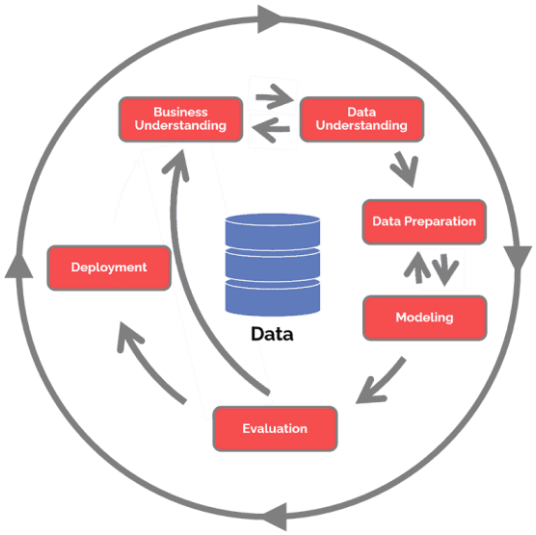


Figure 3.1: Methodology Diagram- CRISP DM

During this phase, various information sources, including secondary sources, were used to gain insights into the issue of fraudulent claims in the motor vehicle insurance sector.

3.1.1.1 Understanding the business

In this phase various information sources such as secondary sources were utilized to enable understand deeply the problem of false claims in motor vehicle insurance industry. Various references such as regional/global onlinepublications, books and journals focusing on ML approaches for uncovering fraud in insurance claims were used. Through analysis of these secondary sources, it became apparent that there has been anotable rise in fraudulent motor vehicle insurance claims. This has ed to substantial financial losses within the industry.

This presented a pressing need for a system capable of swiftly identifying false claims within the motor vehicle insurance industry.

3.1.1.2 Data assessment

During 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 confidentiality of theinformation in it.

The quality of the dataset was assessed and variables were sourced from the dataset to assist in building the model. Below is an extract of the dataset obtained which is in csv file:

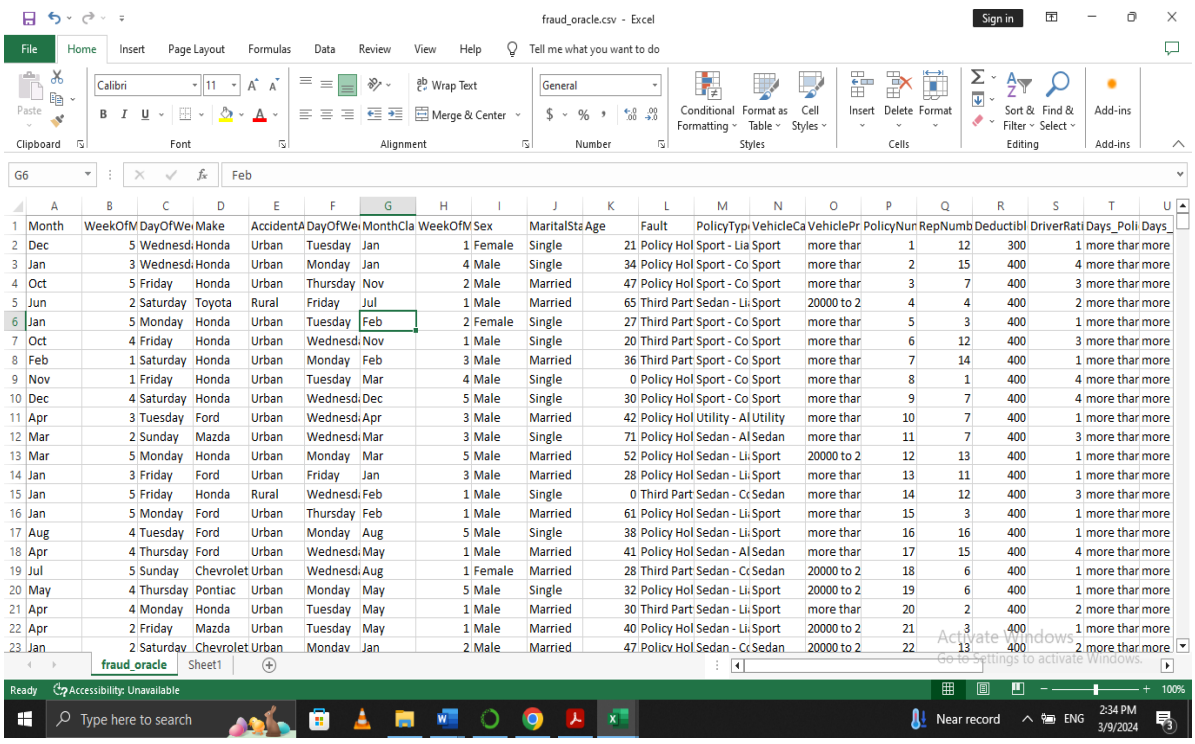


Figure 3.2: CSV file extract of motor vehicle insurance claims dataset.

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

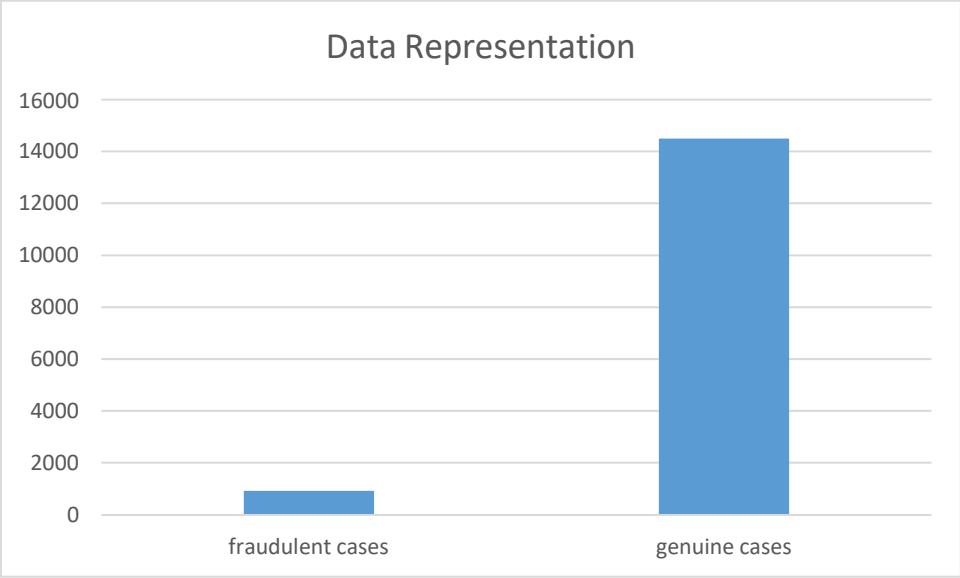


Figure 3.3: Motor vehicle insurance claims dataset, data distribution.

The dataset was made up of a total of 15,420 rows and 33 columns. 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 FS model formed the independent variable while ‘FraudFound_P’ label formed the dependent variable.

The independent and dependent variable were later utilized to train and test ML algorithms so as to detect false claims in motor vehicle insurance.

Dataset used had datatypes as detailed 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 3.4: Datatypes 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 missingdata, driver rating had 92 rows missing data while the rest of the attributes had complete data.

Table 3.1 : Number of Columns of dataset with Null Values

Attributes	Numberof 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

3.1.1.3 Preparation of the data

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 ML techniques. In order to improve data quality for better performance of ML techniques in identifying false claims in motor insurance, the study started by first preparing the data. 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. Cleaning up the data: This step focused on removing anomalies from the data and addressing data gaps, aiming to enhance quality and data consistency.

ii. Data Transformation: This involved transforming the data to formats that are usable by themachine 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 processenhances efficiency by reducing the volume of data while preserving its informational content.

3.1.1.3.1 Cleaning up the data

The process of preparing the data started by identifying and removing duplicate records, followed by handling missing data values. Duplicate data values were manually removed using python skip function. Python's fillna () method was utilized to replace null values with specified alternatives.

The below pseudo code was used:

- *REMOVE duplicate records from the data.*
- *REPLACE empty values with null.*
- *For each column with missing values:*
  Find the most common value.
  Replace missing values with the most common value.
  - Columns: collision_type, property_damage, police_report_available

3.1.1.3.2 Data transformation

Data transformation process involved converting data formats into machine-interpretable formats suitable for ML classifiers. Textual data was converted into integer values, as text cannot be processed directly ML classifiers. 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., 2008)., refers to information arranged into groups with a finite set of possible values.

Python function from 'Scikit-learn library' was employed to perform this conversion, as shown in the pseudo code below:

- *IDENTIFY all columns with text values (categorical variables).*
- *CREATE a copy of the dataset.*
- *For each categorical column:*
- *a. Convert the text values to numeric labels*

3.1.1.3.3 Data Integration

As part of data integration defining the target variable was done. Then it was excluded from the remaining features as it served as the dependent variable while the rest of features represented independent variables. The below pseudo code was used in data integration:

- *SET target_variable TO 'FraudFound_P'*
- *SET y TO the column in the dataset corresponding to target_variable*
- *SET X TO the dataset with the target_variable column removed*
- *REMOVE 'PolicyNumber' column from X*

3.1.1.3.4 Data reduction:

So as to increase the model's performance and lower computational overhead, this research used ensemble multiple filter FS method to select only features that were most important from the dataset. Information gain, GR and chi-square were used to rankfeatures from the original dataset, based on their importance in distinguishing fraudulent and non-fraudulent motor vehicle insurance claims.

3.1.1.3.4.1 Information gain:

IG algorithm works by calculating the mutual information in the dataset. Mutual information quantifies the statistical dependency between variables by evaluating the reduction in uncertainty of one variable given knowledge of another variable. This measure indicates how much knowing the value of a feature reduces the uncertainty about the target label (Guyon et al., 2008). In this case it quantified the association between each feature and the target label (fraudulent or genuine insurance claim). The algorithm then used mutual information scores calculated for every feature in the dataset to arrange the features

based on their mutual information scores.

3.1.1.3.4.2 Gain ratio:

Gain ratio metric is used in FS to rank the significance of features according to their ability to predict the target variable. It helps in deciding which features are most informativefor building predictive models while accounting for the intrinsic complexity of the features (Breiman et al., 2017).

Gain ratio value is computed by dividing the IG by the intrinsic value of the feature. IG quantifies the reduction of uncertainty of the target variable when the data is divided according to a specific feature (Witten et al., 2016).

Intrinsic value 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 variable in the dataset, they were then 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-2 worked by computing chi-2 score and corresponding p-value for every feature in the dataset. The chi-square score quantifies the extent of association between a categorical feature and the target label. ‘p-value’ indicates the likelihood of observing the association by chance alone (Breiman et al., 2017). The computed chi-square scores and p- values were used to evaluatethe degree of correlation between each attribute and the target outcome.

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-2 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 IG, GR and chi-square FS 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 subsetstop k features were selected 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 from the final feature set, after doing feature selection, was grouped into training and testing sets using a 7:3 split ratio, where 70% of the data was utilized for training and the rest 30% for testing purposes. For purpose of model's performance evaluation, the full dataset, before feature selection, was also divided into 7:3 split ratio. This splitting ratio ensured that there was sufficient data for training the model while reserving a separate portion for evaluating the model's performance on new, unseen data (Witten et al., 2016). The training set was used to train the models, while the testing set remained untouched during the training process and was only utilized to evaluate the model's generalization ability.

3.1.1.4 Modelling

In this phase, a model was developed using the following data mining techniques: DT, Naïve Bayes, SVM, and KNN. These are supervised data miningtechniques. The prediction results from individual algorithms were combined by a voting method. This helped to enhance the performance of model in predicting illegitimate 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 ofeach prediction from each algorithm and picking the prediction with the highest probability.

The split dataset obtained from feature selection model was used in training and testing the model. The full dataset was split for purpose of model's performance evaluation.

To tacle the issue of class imbalance, SMOTE was utilized on both training sets for the complete dataset, before FS, and the feature selected dataset. SMOTE creates artificial samples for the underrepresented class, effectively equalizing the classes distribution

To fully utilize the available data for model training and testing, k-fold cross validation method was implemented for every classifier. This provided a robust way of assessing the ML models. It involves splitting the dataset into K subsets, where each subset takes turns as the validation set while the remaining K-1 subsets are used for training. This method

helps reduce to bias by ensuring every data point appears in both training and testing sets (Molnar, 2020). In this research, the dataset was split into 10 segments (K=10) to enhance the model. Initially, the first fold was allocated for testing purposes, while the remaining folds were allocated for training purposes. Subsequently, each fold took turns asthe testing set while ensuring comprehensive evaluation across all 10 folds.

3.1.1.4.1 Experiment Environment

For modeling purposes, this study employed Jupyter Notebook, a widely-used interactive computing environment. Jupyter Notebook allows users to create and share documents that include live code, equations, visualizations, and descriptive text. Supporting multiple programminglanguages, 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 the model's effectiveness, alongside establishing the risk threshold. The model's performance evaluation was done using metrics such as the confusion matrix, classification accuracy, and classification report that included recall, precision, and F-1 score. This evaluation was conducted with the complete dataset and also with feature-selected dataset.

3.1.1.5.1 Confusion Matrix

According to (Bhowmik, 2008), this metric is used to assess how effectively a ML algorithm performs with respect to the target class. To derive the classification metrics aforementioned, the following values were initially calculated using a confusion matrix:

- True Positives (TP) – This represented the number of fraudulent vehicle insurance claims that were correctly identified
- False Negatives (FN) – This represented the number of fraudulent vehicle insurance claims that were missed.
- False Positives (FP) – This represented the number of legitimate vehicle insurance claims that were wrongly flagged as fraudulent.
- True Negative (TN) – This represented the number of legitimate vehicle insurance claims that were accurately recognized as non-fraudulent.

3.1.1.5.2 The Accuracy

As per (Bhowmik, 2008), accuracy is obtained by the ratio of True Positives (correctly identified observations) to the total number of all observations. This summation includes True Positives, False Positives, False Negatives and True Negatives.

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions Made)

3.1.1.5.3 Precision

Precision measures how many of the predicted positive cases are actually correct, calculated as true positives divided by the sum of true and false positives (Bhowmik, 2008).

Precision = TP / (TP + FP)

3.1.1.5.4 Recall

This metric represents the ratio of correctly 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).

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

3.1.1.5.5 F-1 Score

F1 Score is a metric that harmonizes precision and recall into a single value, 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 FS techniques and multiple ML classifiers. This model demonstrated a high-level performance in predicting and accurately classifying fraudulent vehicle insurance claims.

Implementing this model has the capacity to greatly boost the sustained financial gains and customer satisfaction of insurance firms.

CHAPTER 4: RESULTS AND ISCUSSIONS

4.0 Introduction:

In this chapter, the findings of the research are presented, focusing on the utilization of multiple filter FS techniques and various ML algorithms, that is, DT,Naive Bayes, KNN, and SVM for predicting illegitimate motor vehicle insurance claims. This study aimed at evaluating how effective these methods are in accurately detecting fraudulent insurance claims.

4.1 Data Exploratory Analysis

Primary data was not available since insurance companies were not willing to provide dataset onvehicle 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 fordetecting fraudulent claims. With fraudulent claims being a minority class, there was a risk of biased model predictions favoring the non-fraudulent cases which forms the dominant class. To deal with the issues arising from class imbalance, 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. Specified values were used replace the missing data by use of fillna() method in python.

4.2 Multiple Feature selection model evaluation

Multiple filter FS model was used to select the relevant features from the dataset for use by the multiple ML 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 employed.

4.2.1 Mutual information

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

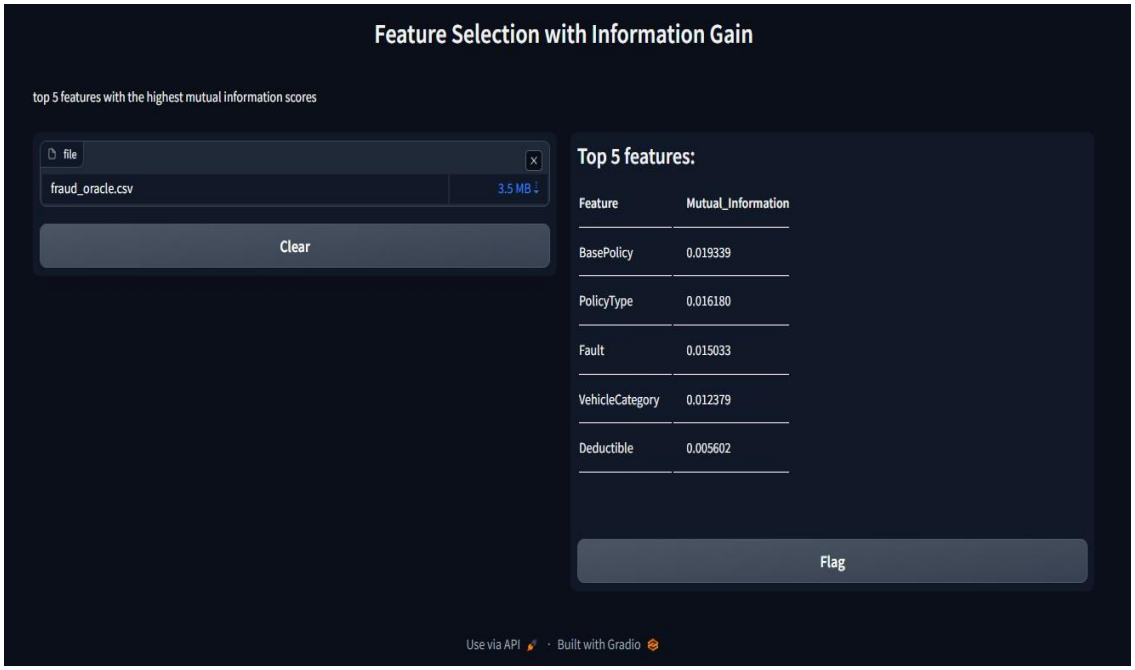


Figure 4.1: FS with information gain

These mutual information scores shown in the figure quantifies how much each feature contributedto predicting the fraudulent insurance claims variable. This measured how much knowing the valueof a feature reduced the uncertainty about the target variable (Liu & Motoda, 2012). Hence offeringvaluable 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 targetvariable. Higher scores indicated that the feature provided more information about the target variable.

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

Following closely behind, ‘PolicyType’ demonstrated a substantial mutual information score of 0.016180. This indicate that it is significance in distinguishing between illegitimate and legitimate claims based on the insurance policy type. 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 asignificant predictor of fraudulent claims. The category or type of vehicle insured may provide important information in determining 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

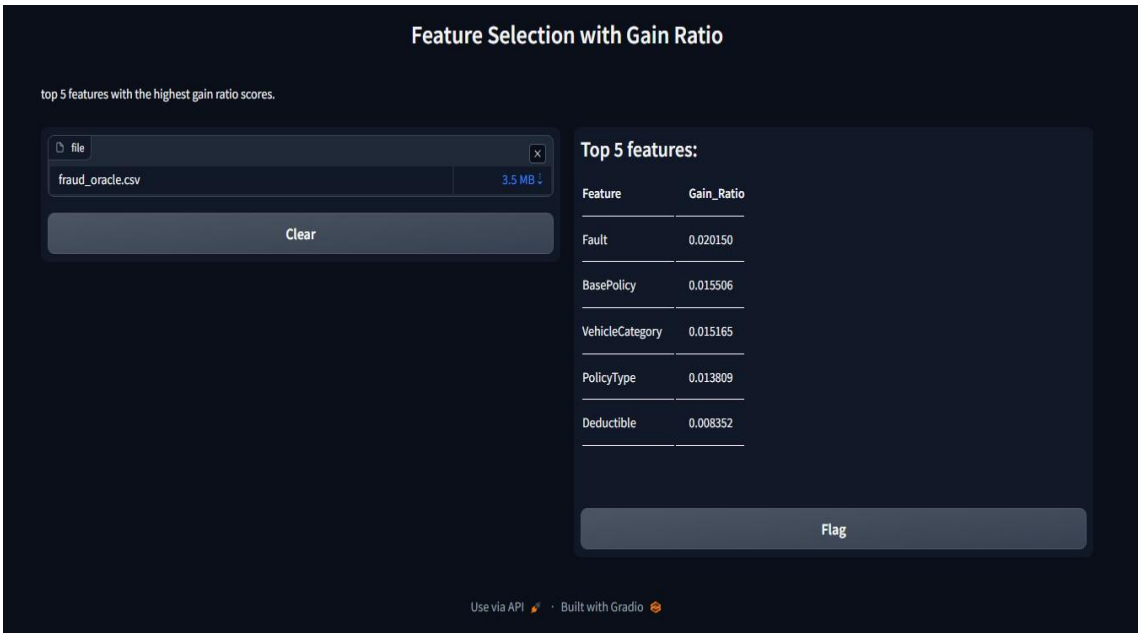


Figure 4.2: Feature selection with gain ratio

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.GR considers 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 impactful 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 kind 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 critical 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 influencethe 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

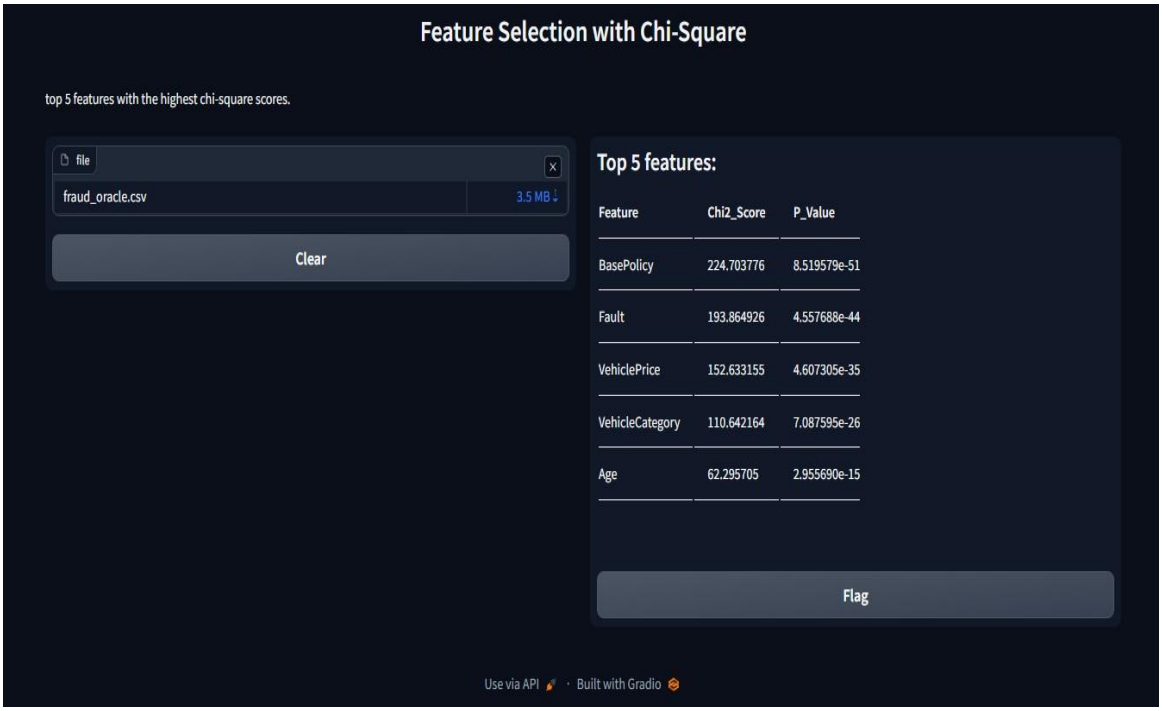


Figure 4.3: Feature selection with chi-square

Chi-2 is statistical tool employed to measure the strength of the association between two categorical variables (Liu & Motoda, 2012). Chi- square assessed the association between each feature and the target label in the dataset by calculating chi-2 scores and corresponding p-values. As 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.

‘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 scoreof 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 toshift 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 cost or value of the insured vehicle. High-valuevehicles 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 categoryof the insured vehicle, such as sedan, SUV, or luxury vehicle. Different vehicle categories may beassociated 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 factorthat can influence insurance claim patterns. Younger claimants may be more inclined to engage inrisky 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 drivingbehavior but could also be targets for insurance scams due to perceived vulnerabilities.

4.2.4 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 ML 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 ratiofeature selection methods. This indicated the that policy terms and deductible amounts had a lot ofrelevance 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 motor vehicle insurance claims that are not genuine.

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.

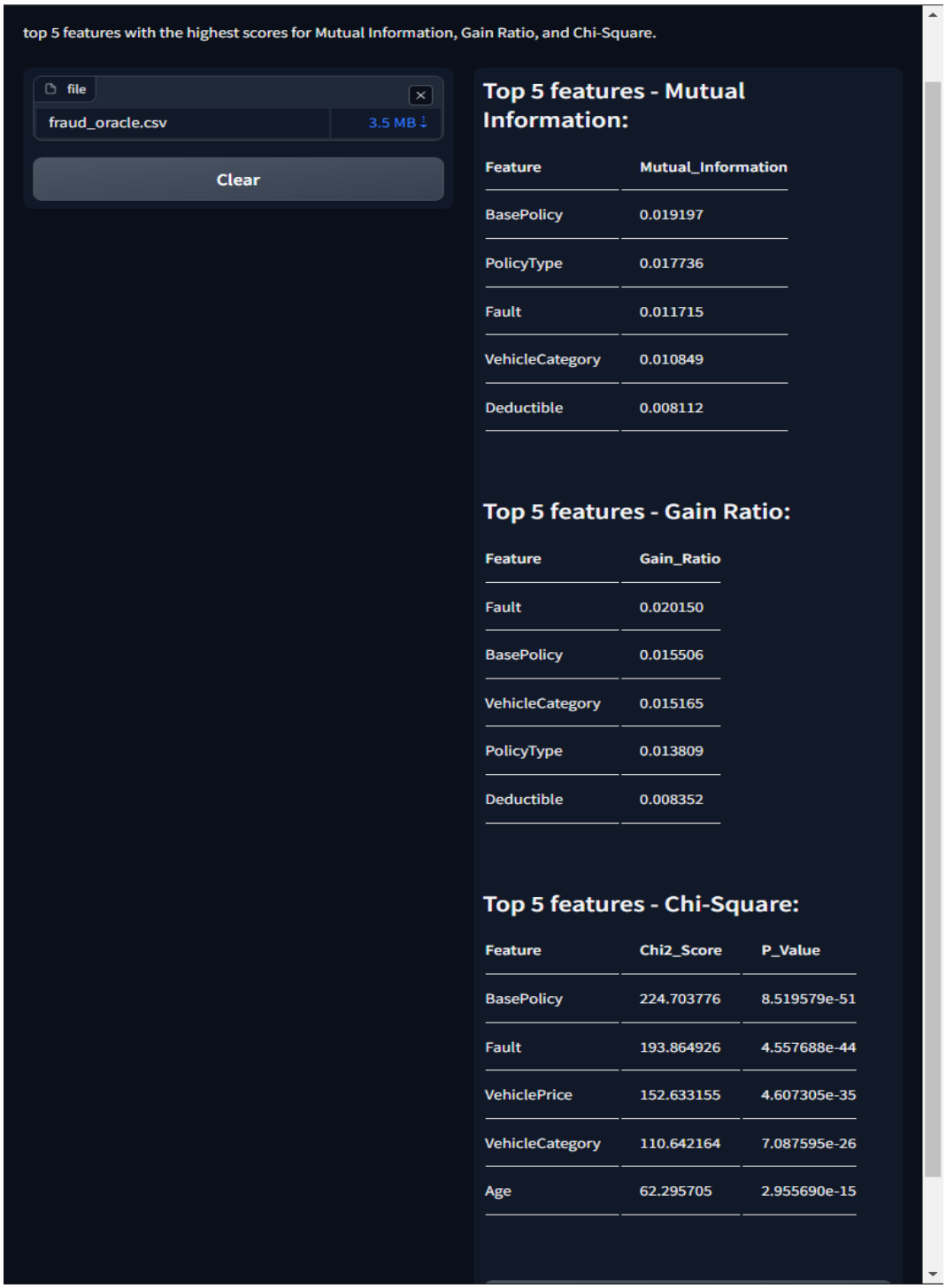


Figure 4.4: Comparison of Features selected by IG, GR and chi-2 FS methods.

4.3 Machine Learning Model Evaluation

The split dataset obtained from the feature selection model was used in the ML algorithms in the ML model. The ML model used DT, Naïve Bayes, SVM, and KNN algorithms. The results from these machine learning algorithms were combined by use soft voting algorithm for the purposes of obtaining the final prediction.

An analysis of ML model classification was performed. The final prediction from the soft voting model using full dataset and also using feature selected dataset, was also analyzed. This was done to evaluate the model's effectiveness and efficiency in discoveringfraudulent motor vehicle insurance claims.

The execution time for machine learning model with feature selected dataset was approximately 7 minutes and 41 seconds. With full dataset the ML model took approximately 10 minutes and 15 seconds to execute.

After the model was trained and evaluated, an evaluation was done. Calculation of metrics such as accuracy, recall, precision, and F-1 score was done so as to compare the model's working using both the feature selected dataset and the full dataset.

The study made use of a confusion matrix to assess the performance of the ML models to predict the target variable. It is constructed by computing True positive denoted as TP, which represents the positive cases that are accurately classified, True Negative denoted as TN, which represents the negative instances that are accurately classified, False positives denoted as FP, which represents negative cases that are not classified correctly and FN which stands for false negative, representing positive instances that are not correctly identified (Bellatreche et al., 2021).

For the purposes of understanding the model's effectiveness metrics such as precision, recall, F1 score, and accuracy were also used.

Precision quantifies the accuracy of positive classifications by dividing the number of true positives by total of all positive predictions made by the classifier. Recall evaluates how well a classifier identifies all actual all positive instances by calculating the ratio of true positive predictions to the total number of actual positive instances in a dataset. A high recall value reflects a low incidence of false negative. This imply that the model can accurately recognize majority of positive cases (Goldberg, 1989).

The F1 score acts as a composite measure, that harmonizes recall and precision. It offers a single value that strikes a balance between these two evaluation measures (Witten et al., 2016).

4.4 Performance Evaluation and Results

4.4.1 Performance evaluation using feature selected dataset

Table 4.1: Feature selected Dataset model's Evaluation Report

Model Type	TN	FP	FN	TP	Precision	Recall	F1 Score	Accuracy
DT	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
Naïve 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

From the above performance assessment of the ML model using feature selected dataset, Decision Tree algorithm had 0.638 Precision. This indicated that out of all the cases predicted as positive by the algorithm, approximately 63.8% were true positive.It had a Recall of 0.134 which means it correctly identified approximately 13.4% of all actual positive cases. It had 0.221 F1 score. This indicated that the algorithm achieved a good balance between precision and recall. The algorithm had 0.724 accuracy. This indicated that the final accuracy of the DT model was 72.4%.

K-Nearest Neighbors had a 0.145 Precision. This indicated that only a small fraction of the predicted positive cases was truly positive. KNN had a recall of 0.240 which indicated that it identified correctly 24% of all real positive instances. The 0.180 F1 score indicated that KNN's balance between precision and recall was lower compared to DT. Despite precision and recall values being low, KNN achieved a high accuracy of 92%.

Naive Bayes had 0.775 Precision. Among all the algorithms, this was the highest. This indicated that out of all positive cases, it correctlyidentified 77.5% true positive cases. However, its recallwas quite low at 11.6%, which indicated that it was not able to detect a big number of actual positive cases. The naïve bayes algorithm had 0.202 F1 Score. This indicated that the balance between precision and recall was moderate. Despite high precision, Naive Bayes had an accuracy of 63.9%, which is due to its low recall.

SVM had a moderate precision of 0.472, which indicated that, out of all positive cases it correctly identified 47.2% of positive instances. It had a low recall at 11%. This indicated thatit missed many actual positive instances. Its F1 Score was 0.178, which suggested its

trade-off between precision and recall was similar to that of Decision Tree. SVM attained 79.9%, which was relatively higher than Naive Bayes algorithm.

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

4.4.2 Performance evaluation using full dataset

Table 4.2: Full dataset model's Evaluation Report

Model Type	TN	FP	FN	TP	Precision	Recall	F1 Score	Accuracy
DT	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
Naïve 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

From the above performance assessment of the ML model using the full dataset, the DT algorithm had 0.260 Precision. This indicated approximately 26% of all the cases predicted as positive by the algorithm, were actual positive. The algorithm 0.164 Recall. This indicated that it correctly identified approximately 16.4% of all instances that were actually positive. The F1 Score was 0.193. This indicated that DT algorithm had a good trade-off between precision and recall. The algorithm's accuracy was 0.893. This suggested that the general perfomance accuracy of the DT algorithm was 89.3%.

K-Nearest Neighbors (KNN) had a Precision of 0.432, which means that approximately

43.2% ofthe instances it identified as positive were actually positive. Recall for KNN was 0.081, indicatingit correctly identified 8.1% of all real positive cases. 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 attained a moderate accuracy of 67.6%.

Naive Bayes had 0.632 Precision. Among all the models this was highest precision score. This indicated that the algorithm correctly identified63.2% of all true positive cases. However, its recall was relatively low at 0.119, meaning it was not able to predict correctly a significant number of real positive cases. The algorithm had 0.200 F1 Score. This indicated that it had a moderate balance between precision andrecall. 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 indicated that it correctly identified 44.9% of positive cases out of all instances predicted as positive. It had unusually high recall score of 1.128. This indicated that it had more false positives than true positives. The F1 Score of 0.199 suggested that it had a trade-offbetween precision and recall similar to the DT algorithm. SVM attained 65.2% accuracy, which is relatively lower compared to Naive Bayes.

The final prediction model had a 0.014 Precision and 0.167 recall score. This indicated that it correctly identified 16.7% cases of all cases that were actually positive. A 0.026 F1 Score 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 ML model that employed DT, KNN, NaïveBayes and SVM, ML algorithms. The output of the individual ML algorithms were combined by use soft voting method so as to obtain the final prediction.

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

The user interface offered an option to upload an input file in csv format. This was then used by FSmodel and the ML model for training, testing the model and making prediction for detection of motor vehicle insurance claims that are not genuine. Output of the fraudulent motor vehicle insurance claims detection system is a csv file which is saved in the users pc.

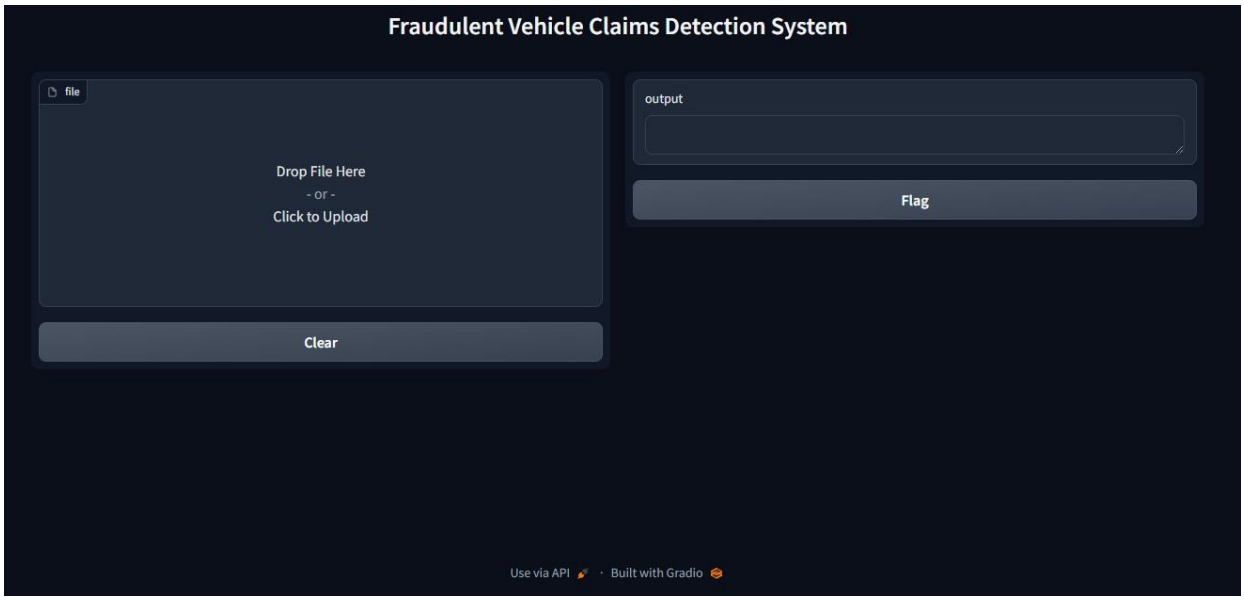


Figure 4.5: Fraudulent motor vehicle insurance claims detection system user interface

The output file had columns for comparison of the prediction made by the ML model anda 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 4.6: Fraudulent motor vehicle insurance claims detection system output file

4.6 Study Discussions

Fraudulent activities in motor vehicle insurance poses significant financial risks to insurers and policyholders alike. The detection of such fraudulent claims is critical in preserving the trustworthiness and reliability of insurance systems while minimizing financial losses.

This study's findings shows that when ensemble FS techniques, that is, IG, GR and chi-2 are applied the performance of the ML model in detecting fraud in motor vehicle insurance claims, is improved. This extends some of the methodologies discussed in the literature.

The ensemble FS approach resulted in reduction of noise and irrelevant attributes. This resulted in enhanced classification by the machine learning model. These results align with those in the assessment done by (Belhadji et al., 2000) and (Sarkar et al., 2018) which

found that using IG and chi-2 as FS techniques greatly improved classification efficiency without losing the accuracy.

The ensemble approach used in this study combines multiple filter feature selection techniques which is in line with recommendations by (Corea, 2017) to use hybrid feature selection as a strategy for better feature ranking in order to achieve feature relevance.

Training on feature selected dataset resulted in faster execution and higher accuracy as compared to using full dataset. This supports findings by (Taha et al., 2022a) that use appropriate feature selection technique increases execution speed and precision of the machine learning models. This also supported findings by (Ürgenç et al., 2022) which demonstrated that feature selection techniques are scalable and efficient especially in high dimensional datasets.

This study's findings shows that use of ensemble classifiers resulted in improved accuracy .This supports the findings drawn by (Moon et al., 2019) and (Severino & Peng, 2021)which demonstrated that tree based and ensemble models are more suitable in capturing non-linear and complex interactions of features in datasets.

The study further strengthened the model's predictive power by use of soft voting method to combine the predictions of the individual classifiers. This resulted in improved accuracy of the final prediction since the final prediction resulted from a consensus of multiple classifiers. This also helped to mitigate limitations of individual classifiers hence resulting in improved performance as earlier indicated in the performance table. This approach is supported by (Sarkar et al., 2018) in the literature, which explores how ensemble classifiers reduces bias and variance hence resulting in a more accurate prediction.

Like many real-world insurance datasets, the dataset used in this research suffered from class imbalance, due low number of the fraudulent claims. This class is a challenge as documented by (Bellatreche et al., 2021) in the literature. The ensemble FS and machine learning model used SMOTE, as earlier explained, to overcome the issue.

The Ensemble feature selection model with ML Model represents a major progress in curbing fraudulent motor vehicle insurance claims. Its robust performance, couple with its practical applicability, positions it as a valuable tool for insurers seeking to protect their assets and maintain the trust of their policyholders.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

Detection of fraudulent motor vehicle insurance claims presents a significant challenge for insurers, hence necessitating adoption of advanced technologies and methodologies. The primary objective of this study was to create, implement and evaluate a model capable of detecting whether a given motorvehicle insurance claim is fraudulent using ensemble FS model and ensemble ML model. A novel web-based application that uses ensemble FS techniques and ensemble ML model was developed from of this research. This system effectively classifies motor vehicle insurance claims as either legitimate or illegitimate. This chapter summarizes the study's findings for developing an Ensemble FS Model coupled with multiple ML algorithms for detecting motor vehicle insurance claims that are not genuine. Recommendations are provided for future research and industry implementationbased on this study's insights.

5.1 Findings Summary

This research employed an Ensemble FS Model in conjunction with MLalgorithms, that is, DT, KNN, Naïve Bayes, and 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 the various FS techniques and ensemble modeling resulted in improved overall

performance of the fraud detection system, henceproviding insurers with a reliable tool for identifying claims that are not genuine.

5.2 Study Conclusion

The research was conducted with the main goal being designing, developing and evaluating a ML learning model for detection of claims that are fraudulent in motor vehicle insurance industry.

The study started by identifying FS techniques that can be used to come up with features for use in building ML models for detecting vehicle insurance claims that are not genuine. Three widely used filter-based FS techniques that is, IG, GR and chi-square were identified and evaluated. The feature selection techniques were selected due to their effectiveness in ranking and selecting features independently of any specific machine learning algorithm. This enhanced generalization and reduced model complexity.

The study proceeded to explore ML techniques that are commonly employed in fraud detection. Several ML algorithms were explored, that is, naïve bayes, KNN, DT, and SVM. This is because they have different strengths when dealing with classification tasks with imbalanced data.

So as to achieve the third objective of the study which was to create and implement an ensemble FS model with ML that can be used for identifying vehicle insurance claims that are not genuine, the study developed an ensemble FS model. The FS model combined the outputs of IG, GR, and chi-2 to come up with a set of features for use by the ensemble ML model. The ensemble ML model was further implemented use of soft voting method. This helped to combine the individual classifiers hence improving the overall prediction

performance.

The study evaluated the working of the ensemble FS and ML model for detection of fraudulent motor vehicle insurance claims. This helped to achieve the fourth objective of the study. The results displayed that the model trained on feature selected dataset achieved better performance than the one trained on full dataset in terms of precision, accuracy and processing time. Decision tree achieved high precision. This is because it has high ability to handle categorical data effectively. KNN achieved low precision due its sensitivity class imbalance. SVM is good in handling high dimensional dataset(Liu & Motoda, 2012). In this study it was less effective due to non-linear separation of the dataset. Naïve bayes achieved a high precision and a low recall which indicated a bias towards the majority class. The ensemble modelling approach helped to achieve greater performance as compared to individual classifiers, highlighting the importance of combining diverse algorithms to improve fraud detection capabilities. This research work supports ongoing initiatives to address insurance fraud while taking care of the interests of both insurers and policyholders.

5.3 Achievements of the research

This research's principal aimwas to design and implement an ensemble FS model with ML that for identifying claims that are fraudulent in motor vehicle insurance domain. Following that, a model that uses ensemble FS techniques for selecting features for use by ML techniques to identify and classify insurance claims as either authentic or fraudulent was created. The model's performance was evaluated.

The final result was aweb- based system that takes an input of the insurance claim data

inform csv file and gives an csvoutput file that indicates whether the claims are authentic or fraudulent.

The research aimed to explore FS techniques for identifying relevant features that can be employed to build ML models to detect claims that are fraudulent in motor vehicle insurance domain, investigate existing ML methods that are currently being used to identify fraudulent insurance claims, design and implement an ensemble FS modelwith ML for detecting fraud in motor vehicle insurance domain and lastly assessing the performance of this ensemble model. Objectives of this study were successfully achieved by first understanding the insurance industry's operations, specifically the motor vehicle insurance segment. Various sources, including data from insurance companies, wereused 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 employed to take of the issues resulting from data imbalance. During data preprocessing phase, the research developed an ensemble FS model that leveraged IG, GR and chi- square FS techniques. This FS model selected a set of relevant features that were usedby the ML model to identify claims that are not genuine in motor vehicle insurance domain.

The study trained ML classifiers on 70% of the feature-selected dataset, the remaining 30% was reserved for testing the classifiers. This process was repeated using the full dataset. This aimed at assessing how FS impacts the working of the ML model. A performance analysis of themodel using both feature-selected data and the complete dataset was done.

As earlier explained in the chapter for results and discussion, the performance results indicated that final prediction obtained from individual classifiers through soft voting was better with the feature selected dataset. The machine learning model also exhibited shorter execution time with the feature selected dataset. The study resulted in development of a model that uses ensemble FS techniques and ensemble ML model that can classify a claim as legitimate or illegitimate. Lastly, all the objectives of this study were successfully achieved.

5.4 Study Limitations

Despite the achievements of the research, some challenges were also 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.

5.5 Recommendations

Drawing from our findings, the study proposes the following for future research and industry implementation:

1. Integration of Advanced Feature Selection Techniques: Advanced FS methods should be explored so as to improve identification of relevant features for use in ML models to detect claims that are fraudulent in motor vehicle insurance industry.
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 feedbackfrom insurers and stakeholders in order to adopt evolving fraud schemes.

4. Collaboration and Knowledge Sharing: There should be a collaboration and knowledge sharingamong insurers, researchers, and regulatory agencies to collectively tackle the challenges posed by insurance fraud and enhance industry-wide resilience.

5.6 Future Work

Detection of fraudulent vehicle insurance claims remains a major challenge within the insuranceindustry. For future work, this study suggests enhancing the system by integrating ML techniques with nature-inspired optimization algorithms. These algorithms draw inspiration from natural phenomena and offer diverse approaches for solving problems (Rathore et al., 2020). This integration addresses the limitation of ML algorithms in handling extensive datasets. This will result to development of the more rapid and effective models for identifying false claims. This integration will also help in automatic and dynamic discovery of the most pertinent features for use in ML so as to detect and classify fraudulent insurance claims. This will result to improved classification accuracy.

In addition, future research could investigate the utilization datasets that spans for multiple years hence containing more data.

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), Article 6. https://doi.org/10.1371/journal.pone.0218760

Baesens, B., Höppner, S., & Verdonck, T. (2021a). Data engineering for fraud detection. *Decision Support Systems*, *150*, 113492. https://doi.org/10.1016/j.dss.2021.113492

Baesens, B., Höppner, S., & Verdonck, T. (2021b). 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), Article 4. 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., & Alonso-Betanzos, A. (2018). *Recent Advances in Ensembles for Feature Selection*. Springer.

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.

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. 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

Corea, F. (2017). *Artificial Intelligence and Exponential Technologies: Business Models Evolution and New Investment Opportunities*. Springer International Publishing.

Dong, G., & Liu, H. (2018). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

Dr.K.K.Savitha. (2023). *Machine Learning based Feature Selection and Classification Techniques for Big Data Applications*. SK Research Group of Companies.

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. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3016715

Feature selection by chi-squared. (2023, August 20). ResearchGate. 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), Article 1. https://doi.org/10.36088/manazhim.v3i1.903

Galli, S. (2020). *Python Feature Engineering Cookbook: Over 70 recipes for creating, engineering, and transforming features to build machine learning models*. Packt Publishing.

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* (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.

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

Insurance Industry Quarterly Claims Statistics for the Period. (n.d.). INSURANCE

REGULATORY AUTHORITY.

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

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

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

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

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), Article 5. 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), Article 10. 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

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

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

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), Article 3. 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), Article 6. 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

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

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

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), Article 7.

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), Article 2. https://doi.org/10.1080/21642583.2019.1620658

Wang, Y., Yu, W., Teng, P., Liu, G., & Xiang, D. (2022). A Detection Method for Abnormal Transactions in E-Commerce Based on Extended Data Flow Conformance Checking. *Wireless Communications and Mobile Computing*, *2022*, 1–14.

https://doi.org/10.1155/2022/4434714

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

APPENDICES

Appendix I: Gantt chart

	July			August				**September**				**October**			
	Week			**Week**				**Week**				**Week**			
ACTIVITY	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Topic Selection															
Writing Proposal															
Presentation of the proposal															
Analysis of the Requirement															
Designing the system															
Implementation															
Testing the system															
Documentation															
Project Presentation															

Appendix II: Project Budget

NO	ITEMS	AMOUNT (KSHS)
1.	Purchase of Laptop	35,000.00
2.	Internet charges	5,000.00
3.	Transport Cost	6,000.00
4.	Stationery	4,000.00
6.	Printing, photocopying and binding	3,000.00
7.	Publication of research paper	8,000.00
	TOTAL	**61,000.00**

Appendix III: Project Code

```
import gradio as gr
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import mutual_info_classif, chi2
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import VotingClassifier

# --- Helper function to load and prepare data ---
def load_and_prepare_data(csv_file):
    """
    Load dataset from CSV, separate target and features, encode categorical variables.
    Drops 'PolicyNumber' column if present.
    """
    df = pd.read_csv(csv_file.name)
    target_col = 'FraudFound_P'

    y = df[target_col]
    X = df.drop(columns=[target_col], errors='ignore')

    if 'PolicyNumber' in X.columns:
        X = X.drop(columns=['PolicyNumber'])

    # Encode categorical columns using pandas factorize for variety
```

```
    categorical_columns = X.select_dtypes(include='object').columns
    for col in categorical_columns:
        X[col], _ = pd.factorize(X[col])

    return X, y, df

# --- Information Gain (Mutual Information) based feature selection ---
def select_features_info_gain(csv_file):
    X, y, _ = load_and_prepare_data(csv_file)

    np.random.seed(42)
    mi_scores = mutual_info_classif(X, y)

    feature_scores = pd.DataFrame({
        'Feature': X.columns,
        'Information_Gain': mi_scores
    }).sort_values(by='Information_Gain', ascending=False)

    top5_features = feature_scores.head(5)
    return top5_features.to_html(index=False)

# --- Gain Ratio based feature selection ---
def entropy(labels):
    """Calculate entropy for an array of labels."""
    values, counts = np.unique(labels, return_counts=True)
    probs = counts / counts.sum()
    return -np.sum(probs * np.log2(probs + 1e-9))  # small epsilon for stability

def gain_ratio_features(csv_file):
    X, y, _ = load_and_prepare_data(csv_file)
    num_samples, num_features = X.shape
```

```
    total_entropy = entropy(y.values)
    gain_ratios = []

    for col in X.columns:
        feature_values = X[col].values
        unique_vals, counts = np.unique(feature_values, return_counts=True)

        split_entropy = 0
        intrinsic_value = 0

        for val, count in zip(unique_vals, counts):
            subset_labels = y[feature_values == val]
            ent = entropy(subset_labels)
            weight = count / num_samples
            split_entropy += weight * ent
            intrinsic_value -= weight * np.log2(weight + 1e-9)

        info_gain = total_entropy - split_entropy
        gain_ratio = info_gain / intrinsic_value if intrinsic_value > 0 else 0
        gain_ratios.append(gain_ratio)

    gain_ratio_df = pd.DataFrame({
        'Feature': X.columns,
        'Gain_Ratio': gain_ratios
    }).sort_values(by='Gain_Ratio', ascending=False)

    return gain_ratio_df.head(5).to_html(index=False)

# --- Chi-Square based feature selection ---
def select_features_chi_square(csv_file):
```

```
    X, y, _ = load_and_prepare_data(csv_file)

    chi_scores, p_values = chi2(X, y)
    chi_df = pd.DataFrame({
        'Feature': X.columns,
        'Chi2_Score': chi_scores,
        'P_Value': p_values
    }).sort_values(by='Chi2_Score', ascending=False)

    top5_chi = chi_df.head(5)
    return top5_chi.to_html(index=False)

# --- Model training with ensemble voting and predictions ---
def train_models_and_predict(csv_file):
    X, y, full_df = load_and_prepare_data(csv_file)

    # Retain PolicyNumber from original dataframe for final results
    policy_numbers = full_df.loc[y.index, 'PolicyNumber'] if 'PolicyNumber' in
full_df.columns else None

    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.3, random_state=42, stratify=y
    )

    # Balance classes with SMOTE on training data only
    sm = SMOTE(random_state=42)
    X_train_bal, y_train_bal = sm.fit_resample(X_train, y_train)

    # Initialize classifiers
    dtree = DecisionTreeClassifier(random_state=42)
```

```
naive_bayes = GaussianNB()
knn = KNeighborsClassifier()
svm = SVC(probability=True, random_state=42)

# Fit individual classifiers
dtree.fit(X_train_bal, y_train_bal)
naive_bayes.fit(X_train_bal, y_train_bal)
knn.fit(X_train_bal, y_train_bal)
svm.fit(X_train_bal, y_train_bal)

# Predict on test set
preds_dtree = dtree.predict(X_test)
preds_nb = naive_bayes.predict(X_test)
preds_knn = knn.predict(X_test)
preds_svm = svm.predict(X_test)

# Ensemble voting classifier with soft voting
ensemble = VotingClassifier(
    estimators=[
        ('dtree', dtree), ('nb', naive_bayes),
        ('knn', knn), ('svm', svm)
    ],
    voting='soft'
)
ensemble.fit(X_train_bal, y_train_bal)
preds_ensemble = ensemble.predict(X_test)

# Map numeric predictions to string labels
label_map = {0: 'genuine', 1: 'fraudulent'}

results = pd.DataFrame({
```

```
        'PolicyNumber': policy_numbers.loc[X_test.index].reset_index(drop=True) if
policy_numbers is not None else 'N/A',
        'Actual_Label': y_test.reset_index(drop=True).map(label_map),
        'DecisionTree_Pred': pd.Series(preds_dtree).map(label_map),
        'NaiveBayes_Pred': pd.Series(preds_nb).map(label_map),
        'KNN_Pred': pd.Series(preds_knn).map(label_map),
        'SVM_Pred': pd.Series(preds_svm).map(label_map),
        'Ensemble_Pred': pd.Series(preds_ensemble).map(label_map)
    })

    # Save results
    results.to_csv('fraud_detection_predictions.csv', index=False)
    return "Model training complete! Predictions saved as 'fraud_detection_predictions.csv'."

# --- Gradio Interface with tabs for each task ---
iface = gr.Blocks()

with iface:
    gr.Markdown("# Fraud Detection Feature Selection and Modeling")

    with gr.Tabs():
        with gr.TabItem("Information Gain Feature Selection"):
            csv_input_ig = gr.File(label="Upload CSV file")
            ig_output = gr.HTML()
            ig_button = gr.Button("Select Top Features")
            ig_button.click(fn=select_features_info_gain, inputs=csv_input_ig,
outputs=ig_output)

        with gr.TabItem("Gain Ratio Feature Selection"):
            csv_input_gr = gr.File(label="Upload CSV file")
            gr_output = gr.HTML()
```

```
        gr_button = gr.Button("Select Top Features")
        gr_button.click(fn=gain_ratio_features, inputs=csv_input_gr, outputs=gr_output)

    with gr.TabItem("Chi-Square Feature Selection"):
        csv_input_chi = gr.File(label="Upload CSV file")
        chi_output = gr.HTML()
        chi_button = gr.Button("Select Top Features")
        chi_button.click(fn=select_features_chi_square, inputs=csv_input_chi,
outputs=chi_output)

    with gr.TabItem("Train Models and Predict"):
        csv_input_train = gr.File(label="Upload CSV file")
        train_output = gr.Textbox()
        train_button = gr.Button("Train and Predict")
        train_button.click(fn=train_models_and_predict, inputs=csv_input_train,
outputs=train_output)

iface.launch()
```

Appendix IV: Proposal Approval For The Research Project

KENYATTA UNIVERSITY
GRADUATE SCHOOL

E-mail: dean-graduate@ku.ac.ke
Website: www.ku.ac.ke
P.O. Box 43844, 00100
NAIROBI, KENYA
Tel. 810901 Ext. 4150

Internal Memo

FROM: Executive Dean, Graduate School
DATE: 27th March, 2024

TO: Anthony Mwiti Wambua
C/o Computing & Info. Science
REF: J57/37307/2017

SUBJECT: APPROVAL OF RESEARCH PROJECT PROPOSAL

This is to inform you that Graduate School Board at its meeting of 13th March, 2024 approved your Research Project Proposal for the M.sc Degree Entitled, **"An Ensemble Feature Selection Model with Machine Learning Model for Detection of Fraudulent Motor Vehicle Insurance Claims."**

You may now proceed with your Data Collection, Subject to Clearance with Director General, National Commission for Science, Technology and Innovation.

As you embark on your data collection, please note that you will be required to submit to Graduate School completed Supervision Tracking and progress report Forms per semester. The Forms are available at the University's Website under Graduate School webpage downloads.

Also, please ensure that you publish article(s) from your project before submitting it to Graduate School for examination as per the Commission for University Education and Kenyatta University guidelines.

Thank you.

ANNE EL MWANIKI
FOR: EXECUTIVE DEAN, GRADUATE SCHOOL

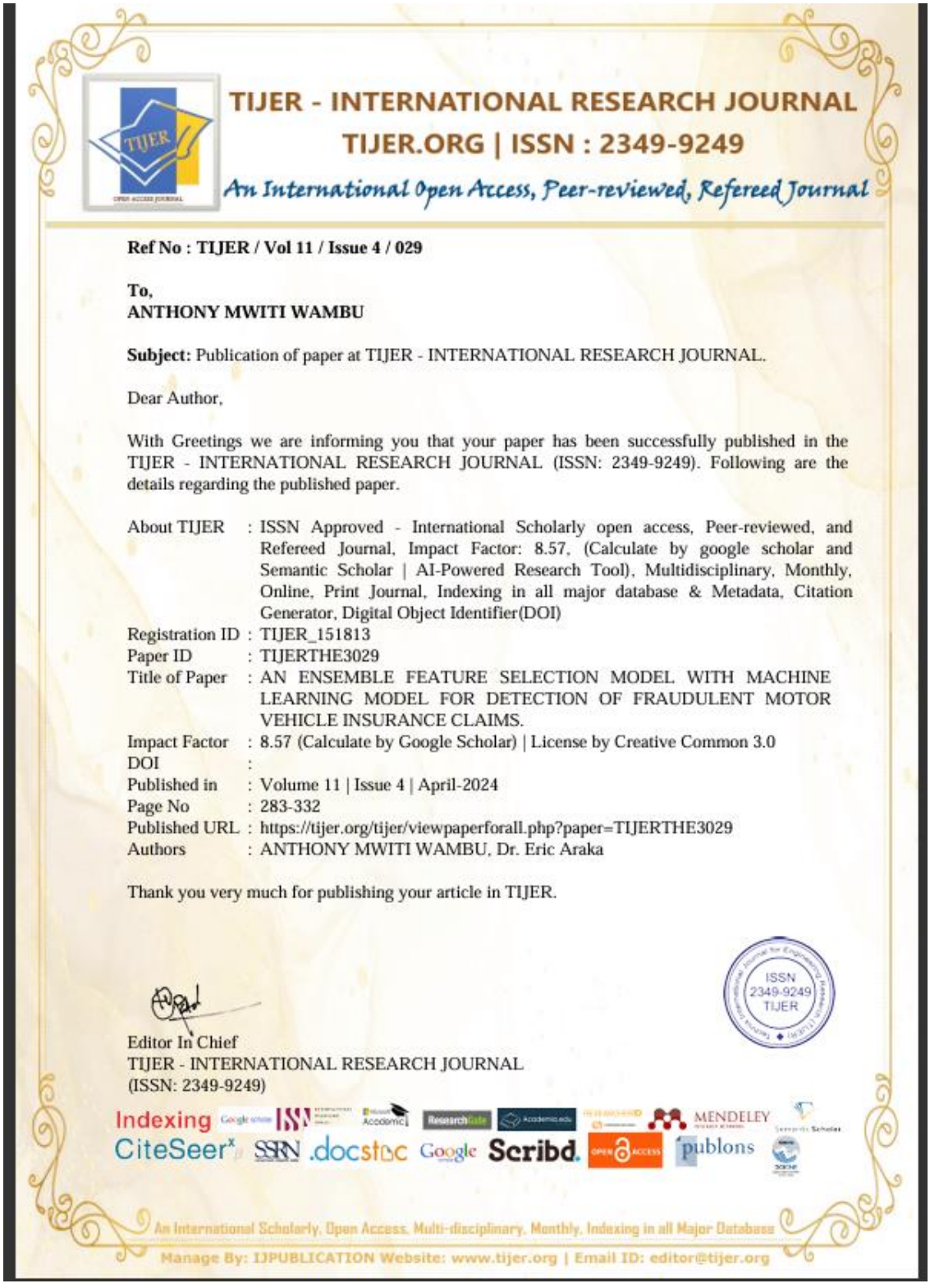
c.c. Chairman, Computing and Information Science

Supervisors:

1. Dr. Eric Araka
   C/o Department of Computing and Information Science
   Kenyatta University

AM/mo

Appendix V: Research Project Published Paper



TIJER - INTERNATIONAL RESEARCH JOURNAL
TIJER.ORG | ISSN : 2349-9249
An International Open Access, Peer-reviewed, Refereed Journal

Ref No : TIJER / Vol 11 / Issue 4 / 029

To,
ANTHONY MWITI WAMBU

Subject: Publication of paper at TIJER - INTERNATIONAL RESEARCH JOURNAL.

Dear Author,

With Greetings we are informing you that your paper has been successfully published in the TIJER - INTERNATIONAL RESEARCH JOURNAL (ISSN: 2349-9249). Following are the details regarding the published paper.

About TIJER	: ISSN Approved - International Scholarly open access, Peer-reviewed, and Refereed Journal, Impact Factor: 8.57, (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool), Multidisciplinary, Monthly, Online, Print Journal, Indexing in all major database & Metadata, Citation Generator, Digital Object Identifier(DOI)
Registration ID	: TIJER_151813
Paper ID	: TIJERTHE3029
Title of Paper	: AN ENSEMBLE FEATURE SELECTION MODEL WITH MACHINE LEARNING MODEL FOR DETECTION OF FRAUDULENT MOTOR VEHICLE INSURANCE CLAIMS.
Impact Factor	: 8.57 (Calculate by Google Scholar) | License by Creative Common 3.0
DOI	:
Published in	: Volume 11 | Issue 4 | April-2024
Page No	: 283-332
Published URL	: https://tijer.org/tijer/viewpaperforall.php?paper=TIJERTHE3029
Authors	: ANTHONY MWITI WAMBU, Dr. Eric Araka

Thank you very much for publishing your article in TIJER.

ISSN 2349-9249 TIJER

Editor In Chief
TIJER - INTERNATIONAL RESEARCH JOURNAL
(ISSN: 2349-9249)

Indexing CiteSeer SSRN .docstoc Google Scribd. publons MENDELEY

An International Scholarly, Open Access, Multi-disciplinary, Monthly, Indexing in all Major Database
Manage By: IJPUBLICATION Website: www.tijer.org | Email ID: editor@tijer.org