National College of Ireland

A Comparative Study of Machine Learning Algorithms for Vehicle Insurance Fraud Detection

MSc Research Project
MSc in Data Analytics

Shruthi Manthena
Student ID: x23235853@student.ncirl.ie

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shruthi Manthena
Student ID:	x23235853@student.ncirl.ie
Programme:	MSc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	12/12/2024
Project Title:	A Comparative Study of Machine Learning Algorithms for Vehicle Insurance Fraud Detection
Word Count:	7026
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature	Shruthi Manthena
Date:	25th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study of Machine Learning Algorithms for Vehicle Insurance Fraud Detection

Shruthi Manthena
x23235853@student.ncirl.ie

Abstract

Vehicle insurance fraud detection is the biggest problem for insurance organizations, because scammers cause great losses and insurance is becoming more expensive for loyal customers. Traditional methods cannot scale well, read poorly and cannot adequately address large volumes of complex data. This study explores the application of machine learning models for predicting insurance claims and fraud detection across three distinct datasets: Claim Fraud Identification within the Insurance Industry, Automobile Insurance Information and Automobile Insurance Claim Forecasting. To get started data cleaning and preprocessing techniques are applied, which involve prominent missing values, encoding critical categorical features and scaling the features if required. Model selection methodologies were used to improve feature selection. Five machine learning algorithms namely Decision Tree, K- Nearest Neighbors (KNN), Light Gradient Boosting Machine (Light GBM), Random Forest and Support Vector Classifier (SVC) were used with each dataset to classify insurance outcomes. To evaluate model efficiency, accuracy together with F1 score, precision and recall were calculated. For Insurance Fraud Claim Detection set, it was determined that Light GBM has the best performance with F1 score of 0.612 and a precision of 0.6. In Car Insurance Data case, Light GBM was the best with an accuracy of 0.841 and F1 score of 0.744. Last but not least, the Light GBM model showed 100% accuracy, F1 score, precision and recall in the vehicle insurance claim prediction dataset to become a model of choice once more across the three simulations. The findings presented in this study clearly illustrate the prospects of machine learning algorithms, especially Light GBM for improving the effectiveness of insurance claim prediction and fraud detection.

Keywords - Vehicle Insurance Fraud Detection, Machine Learning, Fraudulent Claims, Insurance Data, SVC (Support Vector Classifier)

1 Introduction

1.1 Background

One of the biggest problems identified in the global insurance industry is the increasing levels of vehicle insurance fraud, which in turn cause enormous losses Macedo et al. (2021). People fake their ability to pay to influence the cost of insurance contracts and exert an economic burden on insurance firms and genuine policy holders. Insurance car fraud can be pulling off in so many ways such as crashes and fakes O'Brien (2021), an exaggerated bodywork cost, an exaggerated injury claim as well as a fake damage report.

The possibility of detecting such fraud has been more challenging because of the extended complexity and variety of frauds. Conventionally, insurance firms used to undertake insurance fraud investigations manually, which are not only slow, expensive, and susceptible to human interference. Another important type of fraud which has not yet been solved to the extent possible is credit card frauds Mienye and Jere (2024). Automated fraud detection has become a popular solution to this problem due to the availability of large data sets and reduction in machine learning algorithms. By using machine learning models a company can analyze a large number of records and find out that some of them are suspicious. Using machine learning techniques such as decision trees, support vector machines (SVM), AdaBoost and Random Forests insurance providers can easily identify cases of fraud. However, the performance of these models imposed on the quality of the data set, feature selection and the tuning of the algorithms.

1.2 Aim of the study

The aim of this study is threefold: this research seeks to conduct a systematic review and benchmark various machine learning techniques for approach to vehicle insurance fraud detection based on the multiple datasets. Misleading performances in vehicle insurance enormously affect insurance firms hence implying but increased amount to customers, organizational monetary loses. This study analyzes three distinct datasets: irrespectively of being smaller than others, the "insurance_claims.csv "database contains 1000 records; the "Car_Insurance_Claim.csv "contains 10000 records; and the "car_InsuranceClaim.csv "database contains 10300 records to determine the features that characterize fraudulent and genuine claims. Data exploration is followed by exploratory data analysis (EDA) in order to define critical attributes such as the claim amount, preceding accident records, coverage positions, and the attributes of the claim submitters. In this research, several state-of-the-art machine learning techniques such as decision tree, random forest, LightGBM, SVC and k-NN will be employed to identify fraudulent claims accurately. Also, the study aims to supplement the existing research by enhancing feature selection and the general fine-tuning of the models such as LightGBM, and Decision Trees finding that they performed worse as they were preceded by the feature selection conducted in the earlier research.

1.3 Research Question

There are two research questions of this study which are as follows:

1. How do different machine learning classifiers (Random Forest, Light GBM, KNN, SVC, and Decision Tree) compare in terms of performance metrics and which classifier offers the best trade-off between model complexity and predictive performance?

2. What is the impact of feature engineering, such as encoding categorical variables and handling missing data on the performance and robustness of the machine learning models used in vehicle insurance fraud detection?

2 Related Work

2.1 Traditional Approaches to Insurance Fraud Detection

Traditional techniques for insurance fraud monitoring mainly rely on confining and decision rules to detect fraudulent cases. This section indicates that rule-based expert systems and threshold-based systems have proven significant; these systems incorporate rules regarding likely fraudulent activities.

The study given by Baumann (2021) focuses on the modern problems and issues connected with identifying potentially fraudulent claims for insurance organizations, the problems that grow in extent and complexity with the times fraudulent subjects use to avoid truthful settlement. In general, the earlier traditional rule-based expert systems are well used to indicating suspicious activities with specified rules, however, these systems often detach rules from each other, which does not reflect that rules have certain relationships with each other. Among the difficulties that were experienced in this research was how to define the relationships among the rules and at the same time ensure that the system is interpretable and efficient. Furthermore, effective optimization formulate by genetic algorithms need to converge to the correct solution of the problem, as well as ensuring that problem of fraudulent behavior is captured well by the parameter setting required. The outcomes revealed for implementing this enhanced system show that implementing a comprehensive fraud system was highly possible and achieved better rates of fraud detection as compared to rule based techniques.

In the research work discussed by Liu et al. (2020) the author underlines the paramount significance of detecting automobile insurance frauds for minimizing the losses of insurance organizations. Despite the fact that lots of organizations call on the expertise in order to identify frauds the use of such expert knowledge is rather naive and results in misjudgments. To solve this problem, this paper develops a new Evidential Reasoning (ER) rule which incorporate outside information including experience-based factors and prior fraud risk assessments. All components are combined conjunctively and weights are used which are optimized through a MAKER framework for data-driven inferential modeling. One major difficulty, which was encountered during the work, is the possibility of overfitting caused by the inclusion of historical changes in the data, which could make the model less comprehensible and more difficult to use. The findings, derived from the application of the real insurance claims data base, showed that the proposed method retains all the benefits of the expert systems and adapts itself to the evolution of fraud behaviour by tracing the trends in the weights of the experience-based factors.

The study presented in Ahmed et al. (2021) discusses the arising problem of digital fraud which has become a large concern for both consumers and the financial sector due to the expansion of internet banking systems. Realizing that most approaches to fraud detection work after the attempt and is costly, this study therefore turns its focus towards improving on fraud deterrent capacities which are proactive in nature and seek to forestall fraud attempts as soon as they are observed. The solution provides for the development of an Intimation Rule Based (IRB) alert generation algorithm which categorizes the alerts in terms of their importance levels so as to increase the effectiveness in responding to the alerts. A problem that was faced during the course of this research was how best to

incorporate domain knowledge into this rule-based system and at the same time design the layer for flexibility to adapt to new forms of fraud. The findings clearly show that the use of the IRB alerts enhances a more organized and efficient warning system on fraud and hence greatly enhance the preventions by the financial institutions. This model does not only increase the capacities to counter fake attempts, but it also places institutions in a better standing regarding futuristic threats in the field of cybersphere.

There is another study by Hancock et al. (2022) highlights the need to choose the right output probability threshold for classifiers especially in situations where there is a large volume of samples in the classes of interest. This research also points out a weakness in previous work in terms of the systematic way to determine the value of threshold and attempts to establish a quantitative approach to analyze a classification by choosing the value of the threshold according to the user-specified constraints. One of the greatest issues to contend with throughout this study was the question of what constituted the best decision, given that a decision improving one aspect would detract from another. The results show that these thresholds selected for these rates is always higher than the default threshold of 0.5 which shows that a system specific best threshold yields better classification performance.

The challenge of evaluating testing goals with reduced testing capabilities is a problem that the study presented by Landin et al. (2021) aimed at solving via a dynamic threshold-based methodology for test limits detection. The proposed solution uses ML to identify divergent yield points and send notifications to testers when these points fall beyond the optimal values established using data. Such flexibility allows engineers, for the first time, to measure the effect that changes in production have on product performance, and take preventative action to avoid failures. One of the main problem was that the dynamic thresholds had to be easily incorporated into existing testing frameworks.

Different approaches employed and outcomes meant that the five studies can easily be distinguished from each other. Baumann (2021) and Liu et al. (2020) are both centered on enhancing traditional rule-based systems to detect fraud patterns; however, where Baumann (2021) uses association rule mining and genetic algorithms to discover relations between rules, Liu et al. (2020) applies an Evidential Reasoning (ER) rule, which considers experience-based factors that emerge during fraud detection to act dynamically. Liu et al. (2020) method here also outweighs that of Baumann (2021) in that whereas this provides static rule relationships that cannot be altered, Liu et al. (2020) and Ahmed et al. (2021) also pay attention to dynamic fraud detection and prevention but Ahmed et al. (2021) addresses practical internet fraud prevention proactively on digital platforms with the Intimation Rule-Based (IRB) alert model. Both Ahmed et al. (2021) and Hancock et al. (2022) consider threshold-based optimizations. Lastly, Hancock et al. (2022) and Landin et al. (2021) used machine learning for optimization, while the former is specifically on the dynamic threshold in telecommunications testing rather than fraud detection.

2.2 Machine Learning Approaches to Insurance Fraud Detection

There is one study which is given by Prasasti et al. (2020) put forward an accurate automobile insurance fraud detection prediction model for the elimination of costly and ineffective detection systems. This is a critical issue for all machine learning algorithms, especially for the new employee of the automobile insurance company in Indonesia where the instances of fraudulent claims are abysmally low compared to actual claims. In an effort to confront this problem, the researchers use SMOTE together with undersampling techniques. Here the proposed classifiers are MLP, C4.5 Decision Tree, and Random forest. The study shows that the Random Forest classifier has the highest level of credibility among others, therefore, has the highest accuracy level of 98.5

The paper by Salmi and Atif (2021) offers a methodology of using data mining to aid in detecting fraudulent claims on the side of automobile insurance industry, due to escalating fraud cases that have resulted to numerous losses for firms. The first important problem stated is the class imbalance problem of the dataset raising the need to use advanced sampling techniques for the purpose of effective identification of fraudulent claims. In response, the researchers work with two sampling techniques: SMOTE and ROSE, as well as two subsets of features to be used: 23 and 5. The models employed for analysis are Random Forests and Logistic Regression, with a validation process involving a 75:25% split of the actual data set of automobile insurance claims. These findings show that using models based on the reduced set of features contributes increased performance, especially in the Recall for the fraudulent claims, which has achieved the Random Forest Recall rate of 95.24%.

Nabrawi and Alanazi (2023) developed a healthcare fraud detection model used to detect fraudulent claims in health insurance in Saudi Arabia using supervised machine learning and deep learning and algorithms such as Random forest, Logistic regression, artificial neural network etc. The study adapted a real-world imbalanced dataset from three healthcare providers and balanced using SMOT technique for the model development and was used Boruta technique for removing irrelevant features. The first problem solved was related to effective fraud identification in conditions of significant fluctuations in the characteristics of healthcare claims and a relatively small amount of data. Validation metrics used to check the model performance based on which random forest classifier has higher classification accuracy 98.21% than other models for recall, f1-score and AUC.Chararteristic found most important to predict fraud is policy type, education and age. Logistic regression and ANN were equally good with accuracies of 80.36% and 94.64% respectively.

Another study given by Nalluri et al. (2023) suggested a healthcare fraud detection, which aims to determine on factors alleged to be causing medical insurance fraud and to identify the best machine learning technique for its detection. Because of these difficulties, common statistical techniques that may adequately manage medical data are not sufficient and the study therefore employed two unpublished datasets that maybe help identify new patterns. The best model to use for the fraud detection was also identified using four machine learning models, namely SVM, DT, RF, and MLP. The first problem was that of model robustness and efficacy in response to multiple and intricate

data kinds. The Decision Tree model identified 19 significant factors of medical insurance fraud, grouped into categories: health care advocates, cost reimbursement, HCPCS, and beneficiaries information. These results could help insurance management in designing efficient automated auditing systems for reducing fraud thereby Building credibility with these experiments to support the insurance management.

At last Hancock and Khoshgoftaar (2021) suggested a study using Machine Learning techniques, namely Gradient Boosting trees; CatBoost and LightGBM to identify frauds in Medicare Claims; an application of AI in effort to curb the healthcare costs skyrocketed by fraudulently inflated claims. The work centered on coping with huge numerical datasets with numerous feminine features to differentiate fraud. This work was the first to use CatBoost and LightGBM for the Medicare fraud detection and highlighted the issue with encoding categorical variables in Medicare claims. As for the analysis of the outcomes, it was found out that the CatBoost model was the winner, outcompeting LightGBM with mean AUC of 0.77452 which differed significantly from LightGBM with mean AUC of 0.76132 at 0.05% level of significance/1% confidence level. When an extra categorical feature was introduced (Healthcare provider state), CatBoost had an AUC of 0.88245 and, again, outperformed LightGBM with an AUC of 0.85137. The outcomes prove that CatBoost is more effective when it comes to inspecting large categorical features in Medicare fraud detection.

The studies of Prasasti et al. (2020) and Salmi and Atif (2021) are selected because both papers deal with automobile insurance fraud detection, but the authors use different methods and samples. With classifiers, such as MLP, C4.5 Decision Tree, SMOTE and undersampling, Prasati et al. successfully introduce high accuracy of 98.5% with the Random Forest classifier. On the other hand, Salmi and Atif (2021) use enhanced sample selection methods— SMOTE and ROSE, with lesser features set, and has applied Random Forest that performs best again in Recall (95.24%). This raises questions about the effects of feature selection and ways of sampling on the performance of the models. Transiting to the heath care industry, Nabrawi and Alanazi (2023) and Nalluri et al. (2023) both deal in fraud detection but in different perspectives. Nabrawi and Alanazi (2023) consider the theme of health insurance claims based on the classification of supervised and deep learning algorithms. Finally, Hancock and Khoshgoftaar (2021) analyse Medicare fraud detection, applying Gradient Boosting methods (CatBoost and LightGBM).

Unlike previous studies that primarily focus on logistic regression (LR) and rely on basic machine learning models with limited feature selection, my research enhances the feature selection process and fine-tunes advanced models such as LightGBM and Decision Trees for improved fraud detection. Previous work often overlooked the critical role that selecting the right features plays in boosting model performance. In contrast, my approach utilizes multiple datasets—three in total—that analyze key factors and employs advanced techniques like SVM and k-NN to develop a more robust solution for detecting fraud in vehicle insurance claims.

3 Methodology

This section provides a thorough summary of the research methodology. This part provides a technical overview of the process and lists all the steps required to carry out this research successfully. To ensure quality of the data, Methodology part starts with pre-processing, which comprises data cleaning, data normalisation and missing value identification. Feature engineering and feature selection are used to find and extract additional features from the current features in order to improve the model's performance.

3.1 Libraries Imported

The packages applied in this study include Pandas and Numpy libraries for data manipulation and computation. Matplotlib and Seaborn are used for creating plots and visualization during Exploratory data analysis. For Feature Scaling we need Label Encoder from sklearn. They are preprocessing to convert categorical data into numerical data. Random Forest Classifier, K-Nearest Neighbors Classifier, SVC, Decision Tree Classifier, Light GBM Classifier: Machine learning models for training and evaluating for the modules.

3.2 Data Cleaning

Data Cleaning for all three datasets involved dealing with the following issues; missing values, data types of result and categorical variable encoding. In Dataset 1, the missing values were imputed by assigning the mean values for numerical variables such as age, policy_annual_premium and capital gains. Outlier values and anomalous observations in each of the columns were also first searched for and removed if found. Other variables that were categorical like incident_type, insured_sex and policy_state were further encoded using label encoding to convert them into machinereadable formats. The same was applied to Dataset 2 where in columns such as credit_score and annual_mileage, missing values were replaced with the mean values. Other features such as age, vehicle_year and income were also preprocessed and some of these features were converted into right format. This was preceded by the removal of dollar signs and commas to get correct float data types for income and home_val the missing values were also mean imputed. The categorical variables including gender, education and vehicle_type were encoded to machine understandable form by using label encoding. In the 3rd Dataset, the following lines of columns had missing data; the age, YOJ which stands for Years on Job, occupation and car_age. The variable income and home_val were cleaned in a similar way as above and then converted into float data type and for the missing values, we used mean imputations. Education, car_use, and occupation are categorical variables, which were transformed into label encoding.

3.3 Data Preprocessing

Data preprocessing for all three datasets involved some major steps of data preparation before feeding the data into the model. In Dataset 1, missing values were treated, features coded and normalization was done to the numerical features to make them comparable. Such items as age, policy_annual_premium, and capitalgains, were normalized, using features such as standard scaling. Preprocessing of Dataset 2 was also the same, in

which gender, education and vehicle type features were also encoded to convert them into numerical form. The missing values in numerical features credit_score, annual_mileage were also imputed with the mean for data standardization, and other numerical variables were scaled with MinMax feature scaling for effective computations for the machines. Similarly, income has special characters such as dollar and comma which were removed and converted to float to analyze home_val as well. Besides, much identical to Datasets 1 and 2, this dataset also had respondents with missing responses in different questions like age, YOJ, Occupation, and car_age and hence imputed median values. The education and car_use categorical data were encoded using label encoding so that they are in appropriate for machine learning algorithms. Moreover, distributed dependent variables income and home_val were preprocessed for removing symbols and commas and converted into floattype data and missing data were filled up with mean values.

3.4 Dataset Description

This section gives brief information on the three datasets which has been used in this study and each of them is focused on insurance related predictions. Section 3.4.1: Insurance Fraud Claims Detection, in this dataset has 1000 rows, based on customer behaviour, claim history and demographic information aims at determining fraudulent claim. It draws attention to issues that are relevant for fraud detection in such comparatively small data and to the need for increasing accuracy and recall rates of the predictions. Section 3.4.2: Car Insurance Data is a detailed description of 10,000 records with 19 characteristics: 18 companies log raw data with a binary class for claim or no-claim status. This section stresses how the offered dataset relates to realistic car insurance claims, offering ways to approach and calculate the results of claims successfully. Section 3.4.3: Vehicle Insurance Claim Prediction refers to a dataset of 10300 rows aiming at predicting claims with high level of accuracy (98-99%), While less focus is given to the practical features of the dataset to make realistic validation and testing of models. The characteristics subsections provided in this paper explicate the nature of the datasets, the challenges likely to be encountered and their perceived relevance by this study.

3.4.1: Insurance Fraud Claims Detection

This dataset contains 1,000 rows and focuses on detecting fraudulent claims in the insurance domain. It comprises features relevant to customer behavior, claim history, and demographic details, providing an essential foundation for training machine learning models to differentiate between fraudulent and legitimate claims.

3.4.2: Car Insurance Data

This dataset includes 10,000 rows and captures insurance claims related to cars, reflecting real-world scenarios with 19 features. Among these, 18 represent logs recorded by the company, while the target variable indicates whether a customer filed a claim (1) or not (0). The dataset offers a rich mix of numerical and categorical data, requiring robust preprocessing and feature engineering to extract meaningful patterns.

3.4.3: Vehicle Insurance Claim Prediction

This dataset consists of 10,300 rows and is designed to predict insurance claims for vehicles with exceptional accuracy, achieving up to 98-99% performance in prior evaluations. The features represent practical, real-world characteristics, facilitating a comprehensive exploration of claim prediction scenarios.

3.5 Feature Engineering, Label Encoding and Data Splitting

In this study, feature engineering was done to improve the chosen models based on their performance from existing data. All the categorical variables of the three datasets including (Insurance Fraud Claims Detection, Car Insurance Data, and Vehicle Insurance Claim Prediction) were encoded to numerical utilizing label encoding. The data sets were then further divided into test data and training data with 70:30 ratio respectively. This way we make sure that the model learns from most of the data then tested on new unseen data to check if it can generalize well.

3.6 EDA for Insurance Fraud Claims Detection-Dataset1

This section is going to explain all Exploratory Data Analysis for insurance fraud claim detection.

3.6.1 Age Distribution : As shown in figure 1, the age of the vehicle insurance claimants is also shown, which is very much represented in the age group 20-60.

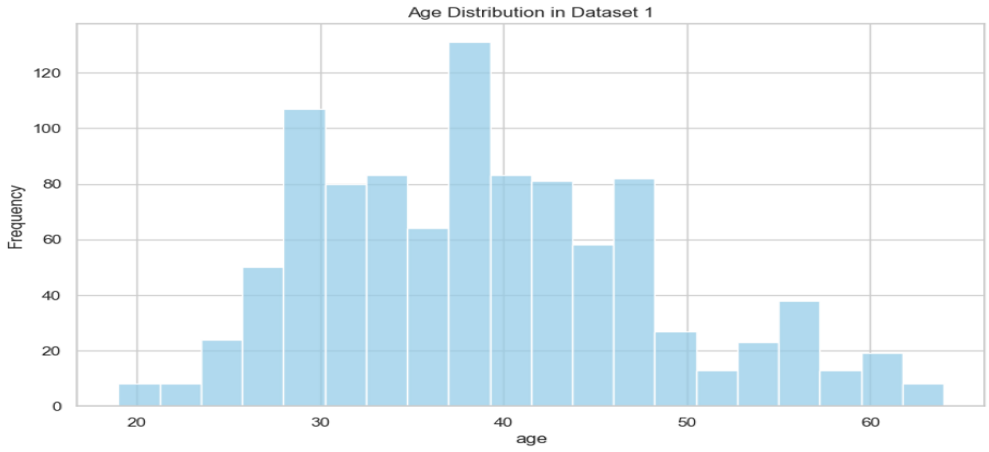

Figure 1: Age Distribution in Insurance Fraud Claims Detect

3.6.2 Boxplot Distribution : The boxplot in figure 2 represents the distribution of vehicle insurance claim amounts which is 55,000 on average. The box shows that claim values are between 45,000 & 70,000; this shows what percentage of normal claims look like.

3.6.3 Correlation Matrix : From the results presented in the correlation matrix below

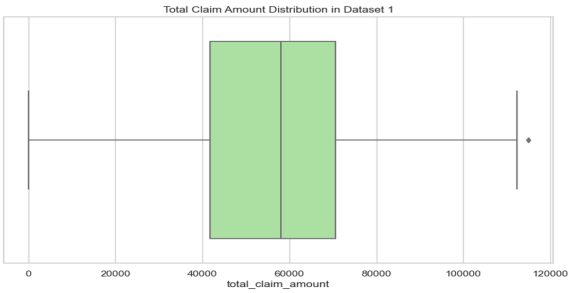

Figure 2: Distribution of Insurance Fraud Claims Detect

in figure 3, one can deduce certain important relations existing between the insurance claims variables. Immediately by the diagonal there are perfect auto-correlations equal to 1.00, while the other elements are reduced and range from weak to moderate correlations. Most importantly, coefficient of determination falls between 0.0001 and 0.0029, which entails that the majority of the examined linear relationships are neither strong nor moderate.

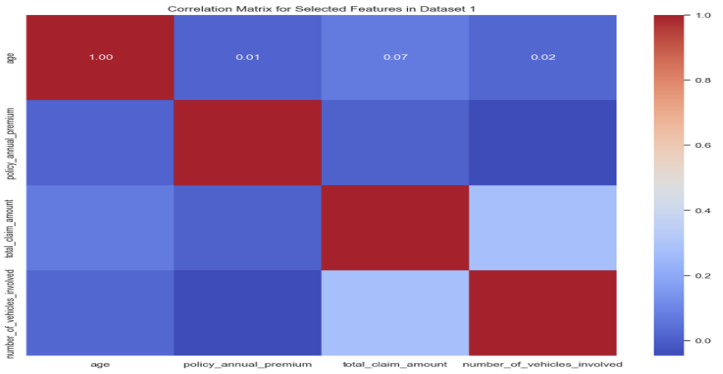

Figure 3: Correlation Matrix for Insurance Fraud Claims Detect

3.6.4 Missing Values : In the horizontal bar chart missing data patterns presented in the vehicle insurance dataset are presented in figure 4, with reference the 'authorities_contacted' field.

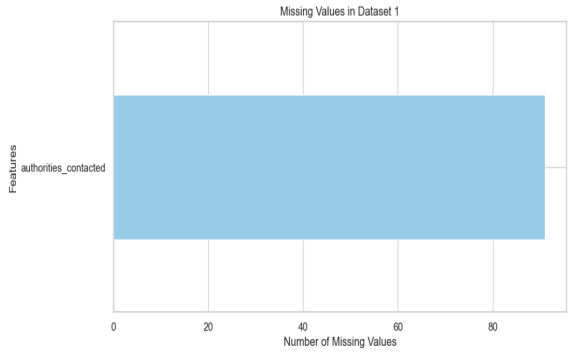

Figure 4: Analysis of Missing Values in Insurance Fraud Claims Detect

3.6.5 Bar Chart Distribution : The horizontal bar chart shows four broad classes of vehicle type involved in the accidents reflected in the insurance claims dataset in figure 5.

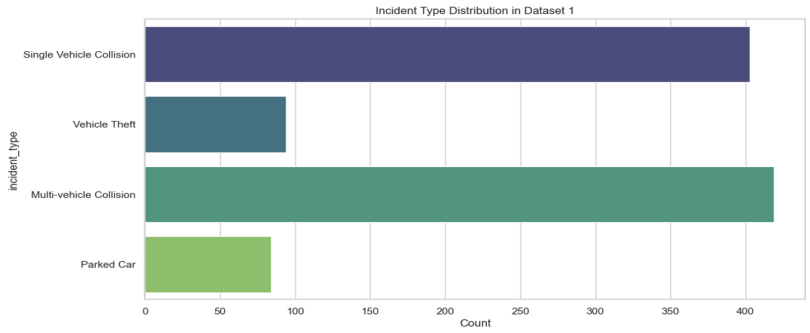

Figure 5: Distribution of Insurance Fraud Claims Detect

The above horizontal bar chart represents the claims frequency for vehicles insurance claims with respect to severity level in figure 6. For minor damage claims account for nearly 350, whereas major damage accounts for about 300. Total loss is the most frequent category, accounting for about 275 occurrences, with trivial damage being the least claim, with approximately 100 occurrences.

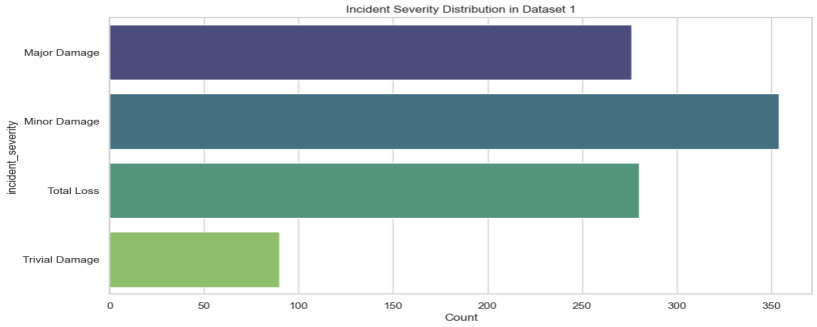

Figure 6: Insurance Fraud Claims Detect

3.6.6 Distribution of Fraud Cases : The below given bar chart shows that how the data regarding the fraud reporting about the vehicle insurance into two categories in figure 7. Analyzing the total claims distribution we have about 750 (87%) as non-fraudulent (N), and 250 (13%) as fraudulent (Y).

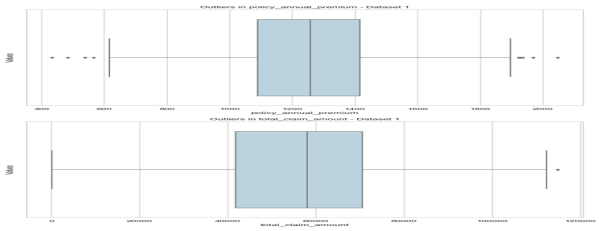

Figure 7: Outlier Analysis of Premium and Claim Amounts in Insurance Fraud Claims Detect

3.7 EDA for Car Insurance Data-Dataset2

3.7.1 Annual Mileage Distribution : The histogram shows how driver records are distributed according to an annual mileage and it can be observed that the results are multi-modal in figure 8.

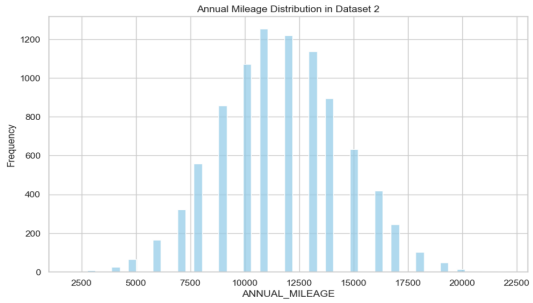

Figure 8: Distribution of Annual Vehicle Mileage for Car Insurance Data

3.7.2 Speeding Violations Distribution : As mentioned earlier, the distribution of speeding violations in the current bar chart is definitely positively skewed in figure 9.

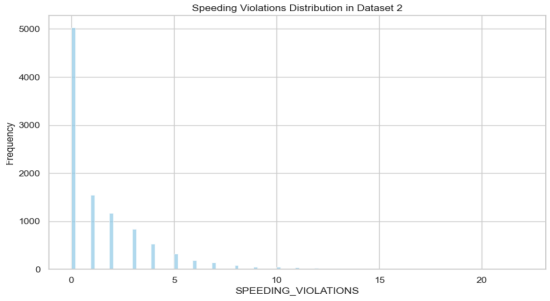

Figure 9: Distribution of Speeding Violations for Car Insurance Data

3.7.3: Correlation Matrix : The correlation matrix below provides significant results including negative and significant association (-0.28) between age, and the annual miles driven implying that the older people are more likely to use few miles annually in figure 10.

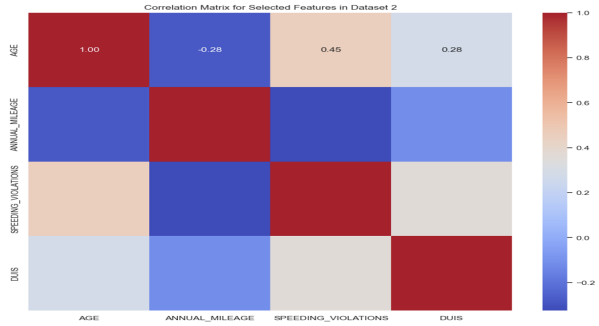

Figure 10: Correlation Matrix for Car Insurance Data

3.7.4: Bar Graph of Missing Values : This horizontal bar chart in figure 11 shows how much of the dataset is missing values as per two important variables in Dataset 2.

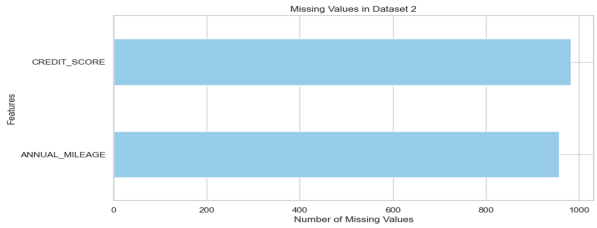

Figure 11: Analysis of Missing Data Points in for Car Insurance Data

3.7.5: Bar Chart Distribution : On the below horizontal bar chart, the distribution of vehicles in the Dataset 2 is presented and the disparities in vehicle classes are quite evident in figure 12.

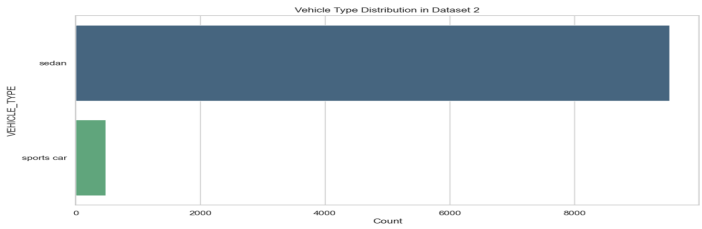

Figure 12: Distribution Analysis of Vehicle Types in for Car Insurance Data

3.7.6: Bar Chart of Distribution of Fraud Cases : Based on the bar chart below, it is illustrated in figure 13 by the light blue bar equal to 6,800; the orange bar corresponds to 3,100 fraudulent cases. This means that there is clearly something closer to 70-30 substantive split.

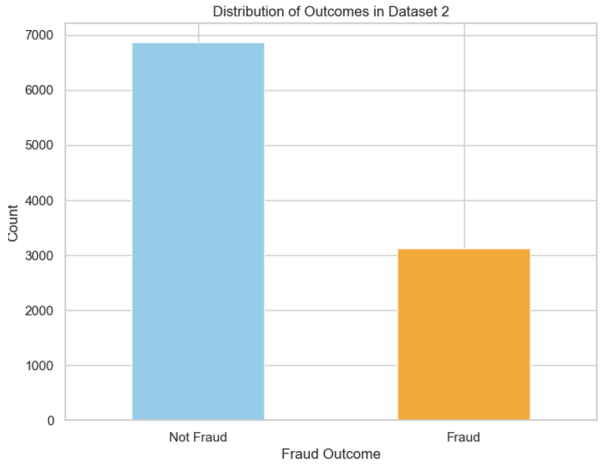

Figure 13: Fraud Case Distribution for Car Insurance Data

3.8 EDA for Vehicle Insurance Claim Prediction-Dataset3

3.8.1 Age Distribution : From the histogram below about 850 separate policyholders are at the age of 50 at which the highest frequency is observed in dataset 3 in figure 14, while there is another peak at the age of 40 and includes around 650 people. The distribution ranges from 20 to 80 years of policies to the policyholders; most particularly 35-60 years.

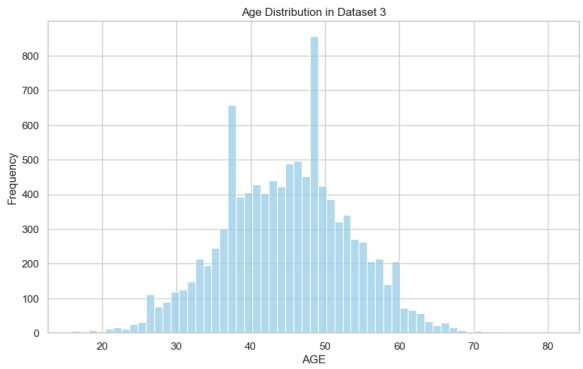

Figure 14: Age Demographics Analysis for Vehicle Insurance Claim Prediction

3.8.2 Correlation Matrix : The correlation matrix in figure 15 shows various and interesting relations between important insurance corresponding variables. There is still a slight tendency for older customers to own older vehicles, with the coefficient for AGE and CAR_AGE being 0.18 (in an attempt at positivity, this is a weak positive customer correlation).

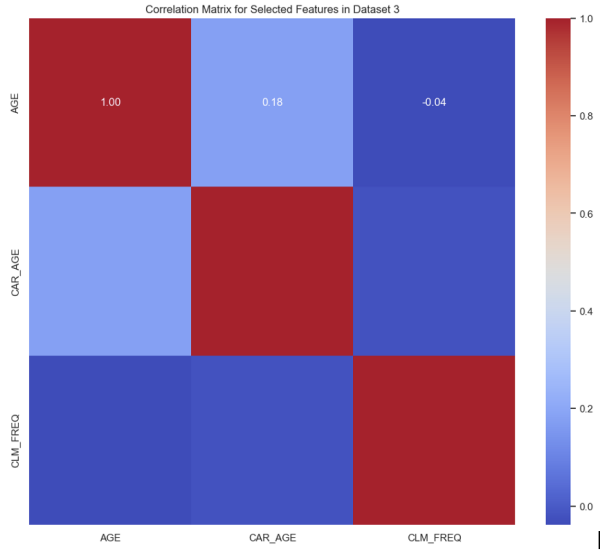

Figure 15: Correlation Analysis of Age Demographics and Customer Behavior Metrics

3.8.4 Bar Chart Missing Values : This horizontal bar chart in figure 16 shows the pattern of missing data in six important variables in Dataset 3. Among all entries, Occupation has the most number of missing entries, around 600, trailed cogently by car age.

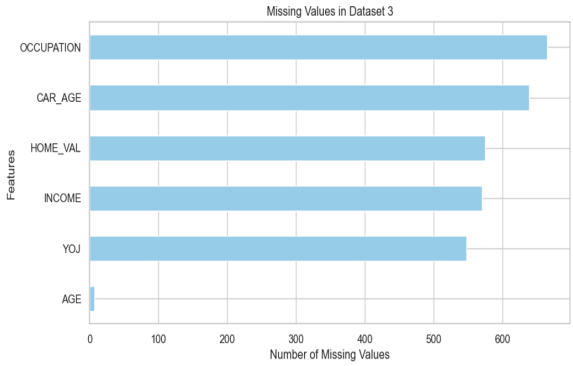

Figure 16: Analysis of Missing Values Distribution for Vehicle Insurance Claim Prediction

Therefore, home value, income and YOJ each has approximately 500-550 missing values.

3.8.5 Educational Attainment Distribution : This graph is the horizontal bar graph showing the proportion of education in Dataset 3 in figure 17. Holders of high school diploma and bachelors degree make up the largest proportion of the customers for the product with over 3,000 each.

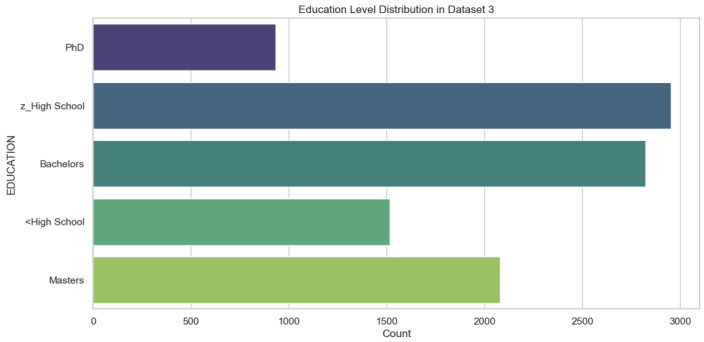

Figure 17: Educational Attainment Distribution for Vehicle Insurance Claim Prediction

In figure 18, the largest group is in the blue-collar employee category with more than 2000 customers, the second largest is in the clerical category with about 1700 customers and the third largest the professionals group with approximately 1500 customers.

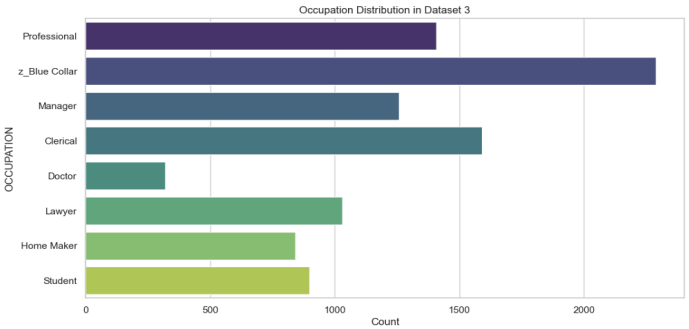

Figure 18: Occupational Demographics for Vehicle Insurance Claim Prediction

3.8.6 Fraud vs Non-Fraud Cases : The bar chart used below helps to depict the fraud distribution in claims of the third dataset in figure 19. In real, the blue bar highlights a definite claim case (about 7500 cases of frauds), while the orange bar reveals the fraud cases (about 2700 cases).

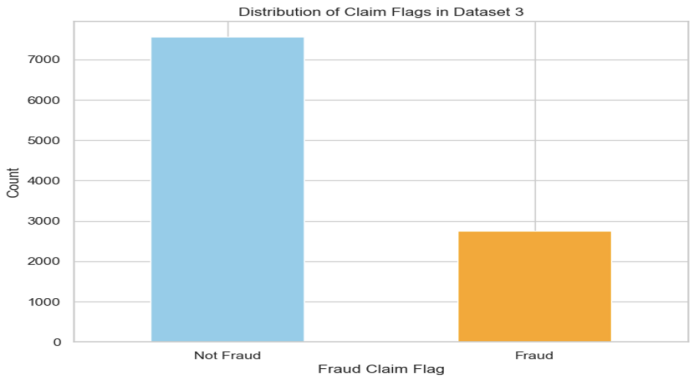

Figure 19: Binary Classification of Insurance Claims: Fraud vs Non-Fraud Cases

4 Design Specification

This design specification is going to show and explain proposed workflow diagram of the study in figure 20 which elucidates the primary activities involved in design specification for a project. The first phase involves dataset collection, which categorizes data into segments with different applications, primary insurance fraud claims detection, car insurance, and vehicle insurance claim prediction. The gathered data is subsequently passed through data cleaning process to make them suitable for use. Jupyter Notebook has been used in this study for python programming.

Data cleaning is done next, and the process moves to data preprocessing and transformation from which the features are extracted and done feature engineering. This stage involves selection and operation of variables which are directly fed into the set models of machine learning. The third step, feature engineering is sensible since it has a direct bearing on the performances of the models.

The next step is the process of data splitting, or the partitioning of the dataset between training and testing. This partitioning makes it possible to build and test the machine learning models as can be seen below. One of the most important steps in the model is called model training during which using the mentioned above algorithms like Random Forest, Light GBM, KNN, etc., is applied to the training data. During this testing, the obtained data is used to evaluate the models formed during the model construction phase.

It defines the use of Support Vector Classifier (SVC) and Decision Tree as also part of the development of the model and the decision-making in the workflow. The last activity of the workflow is the model evaluation in which the overall effectiveness of the trained models is established, and the optimal and best performing one is implemented.

The following is the workflow diagram of proposed design specification that forms a well-

structured and systematized approach to perform data processing feature engineering and model development and evaluation for the project showing in figure 20.

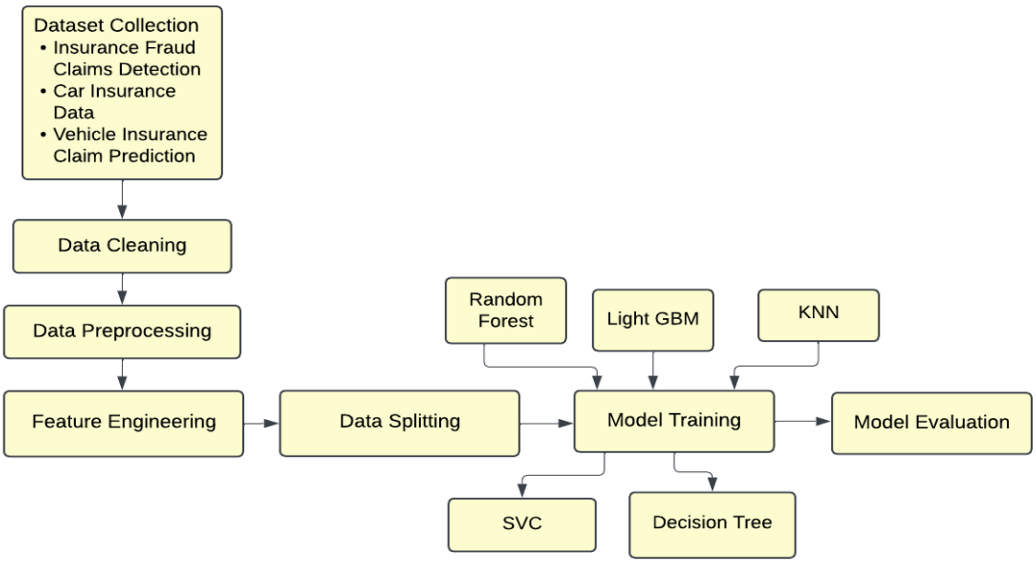

Figure 20: Workflow Diagram

5 Implementation

5.1 Decision Tree Implementation

The Decision Tree is a model which belongs to the supervised learning algorithm used for classification problems only. This approach generates a tree-like structure where each node is a decision depending on the test features leading to class labels at terminals of the tree. In this study, initialization of the Decision Tree classifier was done using DecisionTreeClassifier() in the sklearn toolkit. The classifier was trained by using the training data as X_train and y_train. The model was also measured using different measures such as accuracy, F1 score, precision, recall and confusion matrix. The performance was demonstrated to be convincing relative to the number of trees, depths, and to its handling of numerical as well as categorical variables, but overemphasized, tree depth could lead to overfitting the model.

5.2 KNN Implementation

K-Nearest Neighbors (KNN) is a simple and powerful classification algorithm that labels a data point based on the majority class of its 'K' closest neighbours in feature space. KNN classifier was done using the code from sklearn, with the parameter of neighbours (K) set to the optimal value. In the training phase, it determines the distances of the test sample from all the training samples and uses the majority label. KNN provides good performance in cases, where the number of instances is small and classes are well separable However, its weak point is high sensitivity to the number of instances and the necessity to store all the training instances for the further usage of the algorithm.

5.3 Light GBM Implementation

Light GBM is a Light Gradient Boosting Machine that is a developed gradient boosting algorithm with a major focus on efficiency and performance, particularly for the Light GBM very large data sets. It is related to decision trees and constructs a model through a process of training weak learners (tree) to reduce the loss function. For this implementation, the LGBMClassifier() from the lightgbm library was adopted. This model was developed based on the preprocessed training data with the appropriate hyperparameters including numbers of boosting rounds and learning rate opted. LightGBM has one major advantage over other models concerning larger datasets with more features because it can work with categorical features directly and, therefore, does not require additional memory. Furthermore, LightGBM was able to provide interpretation and feature importance plot that enable one to see which features play a crucial role in fraud detection. By and large, LightGBM had higher predictive accuracy as compared to other models and therefore is the best model in this study.

5.4 Random Forest Implementation

Random forest is another type of learner that involves the use of many decision trees to earn a better result and minimize over-training. The trees in the forest are trained on a bootstrap sample of the data and use a bootstrap sample of the features in splitting at each node, which results in a number of different models that are strong together. This model was implemented using the RandomForestClassifier() from sklearn. The proposed model's performance was assessed using the training data and checked on the testing dataset. Random Forest gives less variance in itself as each tree is built and is considered to give less overfitting than a single tree. It also gives the procedures for feature importance measures to enable the model to learn the vital indicators for the identification of fraudulent transactions.

5.5 SVC Implementation

Support Vector Classifier (SVC) is a very capable supervised learning algorithm that is used in cases of binary and multiclass classification. It achieves its task by maximizing the margin, that is, the distance between different classes in the feature space. The SVC () model from sklearn was used and trained on the preprocessed dataset. However, one of the limitations encountered in the use of SVC was its dependence on the type of kernel used as well as the other hyperparameters.

6 Evaluation

6.1 Case Study 1: Insurance Fraud Claim Detection with Machine Learning Approaches

In this particular work, five machine learning models were tested on the Insurance Fraud Claim Detection dataset to determine which model is the most effective at detecting fraudulent claims. For distinguishing genuine from fraudulent transactions, Decision Tree model performed at 0.79 accuracy with moderate F1 score of 0.59, precision of 0.54, and recall of 0.65. The classifier used was the K-Nearest Neighbors (KNN) which gave only

0.71 accuracy for classification but poorly performed in fraud detection with the F1 Score of only 0.12, precision of 0.2 and recall of 0.08 meaning it lacks enough robustness for this task. For fraud detection, LightGBM was the best model giving an accuracy of 0.81, F1 Score 0.60, precision 0.58 and recall 0.62. Even by using the Random Forest model, the accuracy achieved was 0.75, the F1 Score was relatively low – only 0.35, with moderate precision – 0.43, and potentially low recall – 0.30. On the other hand, Support Vector Classifier (SVC) had low accuracy of 0.77 that had totally nil level of fraud detection in F1 Score, precision & recall. From this comparative analysis, LightGBM emerges as the best model used in insurance fraud identification to present a stable solution as shown in Table 1.

Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Decision Tree	0.79	0.59	0.54	0.65
KNN	0.71	0.12	0.2	0.08
Light GBM (best model)	**0.81**	**0.60**	**0.58**	**0.62**
Random Forest	0.75	0.35	0.43	0.30
SVC	0.77	0.00	0.00	0.00

Table 1: Performance Comparison of Machine Learning Models on Insurance Fraud Claim Detection Dataset

Figure 21 depicts the evaluation results of performance measures for different created models aimed at the Insurance Fraud Claim Detection dataset. The model which has the highest Accuracy and F1 Score is the Random Forest, demonstrating the best ability of correctly classifying the insurance frauds claims. LightGBM is also high in performance with high Accuracy, Precision, and Recall. However, lower outcome of Accuracy and F1 Score can be observed in the KNN model. One can note that high Accuracy is reached by the SVC model, but F1 Score and Recall are not as high, whereas in the Decision Tree model, all the examined metrics are optimal.

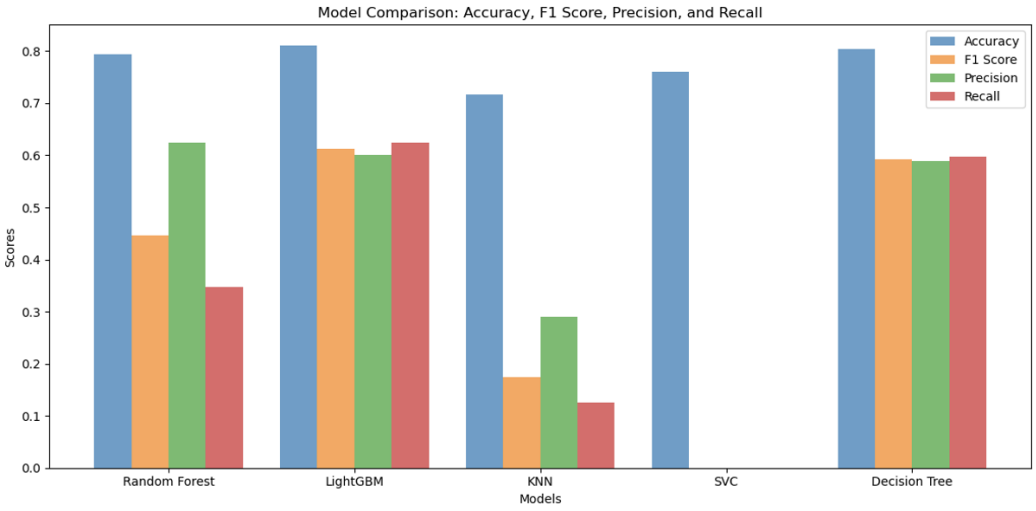

Figure 21: Model Comparison - Insurance Fraud Claim Detection Dataset

6.2 Case Study 2: Car Insurance Data with Machine Learning Approaches

In this case study, we apply several machine learning algorithms on the Car Insurance Data to predict insurance claims. The Decision Tree model aimed at insurance claims classification had an accuracy of 0.788 and an F1 score 0.67, with a precision of 0.66 and recall of 0.68. The KNN model itself did moderately bad, by returning the accuracy of 0.639, F1 = 0.27, the precision of 0.37 and the recall of 0.21, which mean that is was not very good at identifying the fraudulent claims. On the other hand, the Light GBM model was seen to have the highest accuracy of 0.841, F-score of 0.74, precision of 0.75, recall of 0.74 for identifying the fraud. The Random Forest model was almost equally good with an accuracy of 0.83, F1 score of 0.72, precision of 0.74 and recall of 0.71.

Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Decision Tree	0.788	0.67	0.66	0.68
KNN	0.639	0.27	0.37	0.21
Light GBM (best model)	**0.841**	**0.74**	**0.75**	**0.74**
Random Forest	0.83	0.72	0.74	0.71
SVC	0.686	0.0	0.0	0.0

Table 2: Performance Comparison of Machine Learning Models on Car Insurance Dataset

In Figure 22, an analysis of the Car Insurance Data is provided in terms of performance, with the various corresponding machine learning models indicated. Random forest is the best-performing model that reveals maximum values in Accuracy as well as F1 score for classifying car insurance data. The LightGBM model also possesses acceptable accuracy, a high value of Precision, and a high value of Recall. However, the KNN model is comparatively lower in having an Accuracy and F1 Score. The SVC model attains a superior Accuracy compared to the F1-Score and Recall whereas, the Decision Tree model yields a more rightful scores across the criteria.

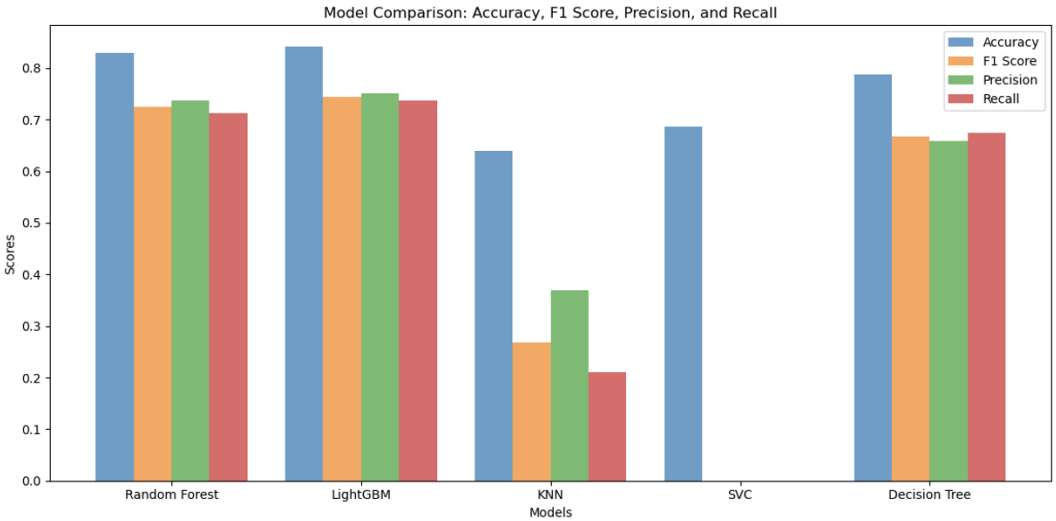

Figure 22: Model Comparison - Car Insurance Data

6.3 Case Study 3: Vehicle Insurance Claim Prediction with Machine Learning Approaches

In the proposed case study, the Vehicle Insurance Claim Prediction, the Light GBM was found to be the best model as it obtained an accuracy of 1.0, F1 score of 1.0, the precision of 1.0, as well as the recall value of 1.0. The same is true about the Decision Tree model that had a high accuracy of 0.999, the F1 score of 0.999, and perfect precision of 1.0 though it was slightly low in recall 0.998. Random Forest tops it a little behind with the accuracy of 0.999 F1 score of 0.998 and with the precision of 1.0 but has slightly lower recall of 0.996. KNN lacked better performance hence recorded an accuracy of 0.679 in F1 and a record of an undesirable figure of (0.137) in recall. SVC was the worst algorithm with an accuracy of 0.728 as well as an insignificant F1 score, precision as well as recall.

Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Decision Tree	0.9994	0.9988	1.0	0.9976
KNN	0.6787	0.1881	0.3003	0.1369
Light GBM (best model)	**1.0**	**1.0**	**1.0**	**1.0**
Random Forest	0.9990	0.9982	1.0	0.9964
SVC	0.7282	0.0	0.0	0.0

Table 3: Performance Comparison of Machine Learning Models on Vehicle Insurance Claim Prediction Dataset

In Figure 23 a comparison of the performance indicators of the different machine learning models, which were implemented on the dataset called Vehicle Insurance Claim Prediction, is shown. The Random Forest model has the highest accuracy, F1-score, precision, and recall all of which suggest the model has high capability to predict claims for vehicle insurance with high accuracy. The LightGBM model, too, presents the good results that prove high Accuracy, Precision, and Recall. However, KNN model identifies less Accuracy, F1 Score and Recall comparatively. The SVC model's Accuracy is good, but the F1 Score and Recall are comparatively low and although the Decision Tree model's overalls performance is worse, it is much more balanced.

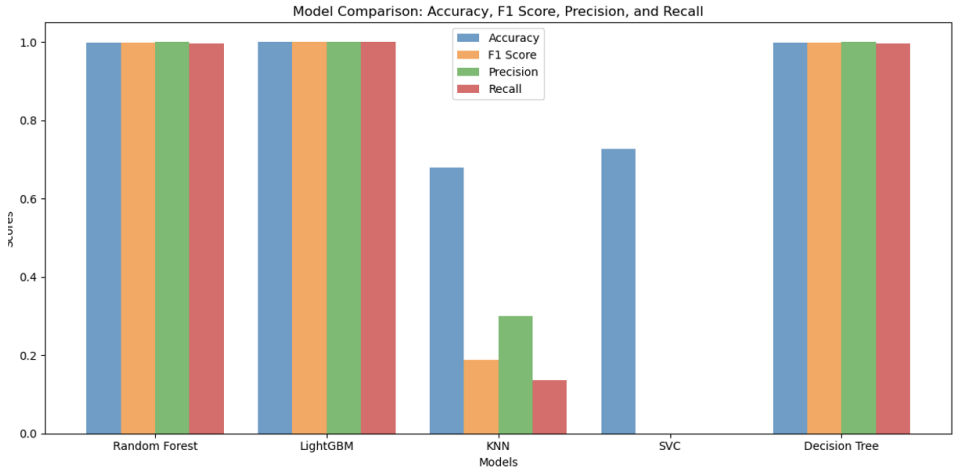

Figure 23: Model Comparison - Vehicle Insurance Claim Prediction

6.4 Discussion

Comparing the outcome of the machine learning models on the three datasets offers important lessons on the capabilities and utility of the techniques. While studying Insurance Fraud Claim Detection, LightGBM works best yielding the highest accuracy 0.81 and favourable F1 score 0.60 for imbalanced data set and capturing complex pattern. In Car Insurance Data, LightGBM prevailed again as it had an improved accuracy of 0.84 and F1 score of 0.744 testifying to its capability on handling large voluminous and varied data. Random Forests also did a good job which is consistent with this model's overall applicability. In Vehicle Insurance Claim Prediction LightGBM achieved 100% accuracy score and F1 score of 1 which means it can be used for datasets having clear cut features. Although Decision Tree and Random Forests performed well at the benchmark, KNN and SVC were the worst models that performed poorly in every evaluation with high variance, skewed datasets, correlated features or high interaction. These results serve to underscore the need to select sophisticated decision-making algorithms such as LightGBM for insurance-related predictions and their scalability. More future work can be done to improve the prediction by using the combination of HWL and other methods.

7 Conclusion and Future Work

7.1 Conclusions

This study analyzed the performance of multiple machine learning models—Decision Tree, KNN, LightGBM, Random Forest, and SVC—on three diverse insurance datasets: insurance fraud claim detection features, car insurance data and vehicle insurance claim prediction. LightGBM demonstrated the highest accuracy in all datasets because of its ability to solve high-dimensional data, imbalanced classes, and interactions, achieving 84.1% of accuracy for Car Insurance Data and a perfect accurate result with being 100% for Vehicle Insurance Claim Prediction. Nevertheless, furthering the results of some models contaminated low accuracy, for instance KNN and SVC in the frame of class imbalance and non-linear interactions. These results support the use of ensemble-based techniques, and suggest directions for further development of model selection procedures and preprocessing.

7.2 Limitations and Future Works

However, several limitations exist in this study, which would have to be addressed in future work despite achieving successful results in implementing insurance prediction models using machine learning. First, the problem of having imbalanced class distribution in the datasets affected the performance of some models a situation that can be rectified by use of SMOTE (Synthetic Minority Oversampling Technique). Second, hyperparameter tuning was limited, and more research could be conducted in the future for various level of hyperparameter tuning like Bayes or Grid Search. Third, new architectures of deep learning models, including neural networks or transformers, could be investigated on the strategy of enhancing the performance of traditional machine learning for analyzing intricate data patterns. Furthermore, blend of models such as stack or blend models could enhance accuracy, integration of external datasets with other features and real data might also improve generalization as a result. Finally, for post-deployment interpretation of the

models, explainability techniques like SHAP or LIME must be used to enhance usability of the models in decision-making for insurance industry.

References

Ahmed, M., Ansar, K., Muckley, C. B., Khan, A., Anjum, A. and Talha, M. (2021). A semantic rule based digital fraud detection, *PeerJ Computer Science* **7**: e649.

Baumann, M. (2021). Improving a rule-based fraud detection system with classification based on association rule mining.

Hancock, J., Johnson, J. M. and Khoshgoftaar, T. M. (2022). A comparative approach to threshold optimization for classifying imbalanced data, *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, IEEE, pp. 135–142.

Hancock, J. T. and Khoshgoftaar, T. M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection, *SN Computer Science* **2**(4): 268.

Landin, C., Liu, J. and Tahvili, S. (2021). A dynamic threshold based approach fordetecting the test limits, *The Sixteenth International Conference on Software Engineering Advances ICSEA 2021*, p. 81.

Liu, X., Yang, J.-B., Xu, D.-L., Derrick, K., Stubbs, C. and Stockdale, M. (2020). Automobile insurance fraud detection using the evidential reasoning approach and data-driven inferential modelling, *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, pp. 1–7.

Macedo, A. M., Cardoso, C. V., Neto, J. S. M. et al. (2021). Car insurance fraud: the role of vehicle repair workshops, *International journal of law, crime and justice* **65**: 100456.

Mienye, I. D. and Jere, N. (2024). Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions, *IEEE Access* .

Nabrawi, E. and Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning, *Risks* **11**(9): 160.

Nalluri, V., Chang, J.-R., Chen, L.-S. and Chen, J.-C. (2023). Building prediction models and discovering important factors of health insurance fraud using machine learning methods, *Journal of Ambient Intelligence and Humanized Computing* **14**(7): 9607–9619.

O'Brien, S. (2021). *The criminal act of committing insurance fraud: The challenges facing insurers when detecting and preventing insurance fraud.*, PhD thesis, Dublin Business School.

Prasasti, I. M. N., Dhini, A. and Laoh, E. (2020). Automobile insurance fraud detection using supervised classifiers, *2020 International Workshop on Big Data and Information Security (IWBIS)*, IEEE, pp. 47–52.

Salmi, M. and Atif, D. (2021). Using a data mining approach to detect automobile insurance fraud, *International Conference on Soft Computing and Pattern Recognition*, Springer, pp. 55–66.