National College of Ireland

Project Submission Sheet

Student Name:	Stamatios Karvounis		
Student ID:	x18197051		
Programme:	PGDDA_JAN23_O - Postgraduate Diploma in Science in Data Analytics	**Year:**	2023
Module:	Domain Application of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	01/12/2023		
Project Title:	Project Design CA Predictive Model in Car Insurance Dataset		
Word Count:	5032		

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 references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Stamatios Karvounis

Date: 01/12/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

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

AI Acknowledgement Supplement

Domain Application of Predictive Analytics

Predictive Model in Car Insurance Dataset

Your Name/Student Number	Course	Date
Stamatios Karvounis	Domain Application of Predictive Analytics	01/12/2023

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

CA2 Domain Applications of Predictive Analytics

Car Insurance Fraud Detection

Stamatios Karvounis
School of Computing
National College of Ireland
Dublin, Ireland
x18197051@student.ncirl.ie

***Abstract*—This project dives spotting fraud in vehicle insurance—a huge concern for insurance companies. The unique aspects of insurance fraud are explored. By hunting down and using machine learning tools dedicated to fraud detection, the goal of this project is to make fraud detection systems better at catching culprits. By identifying and implementing machine learning techniques suited to this domain, the project aims to enhance the accuracy and efficiency of fraud detection systems. The findings encompass quantitative metrics, including accuracy, precision, recall, and F1-score, which are interpreted with respect to their direct impact on reducing fraudulent claims and improving decision-making within the vehicle insurance sector. This research contributes to the development of tailored fraud detection solutions, with a focus on practical applications and business value**

Keywords— Predictive Analytics, Classification, Fraud Detection, Car Insuranc)

I. INTRODUCTION

Insurance fraud is increasingly recognized as a significant and growing problem within the global insurance industry. The financial implications of fraudulent claims are substantial, with annual estimates surpassing $40 billion. This not only threatens the profitability of insurers but also undermines the integrity of the entire insurance system (Roy & George, 2017). In European countries, it is estimated that between 5% and 10% of total annual indemnities paid in non-life insurance are due to fraudulent claims. In Ireland, the financial burden of insurance fraud on companies is estimated at around €200 million each year. These costs, borne ultimately by honest policyholders, highlight the real impact of fraudulent claims on the broader insurance system (Insurance Ireland, 2020).

The consequences of insurance fraud are extensive. The financial health of insurers is put at risk, and premium rates for honest policyholders are driven up. The fundamental principle of insurance, which is the solidarity among policyholders, is compromised by fraud. If left unchecked, insurance fraud has the potential to cause significant disruptions in the industry and create adverse effects on social and economic structures (Viaene and Dedene, 2004).

In light of these challenges, insurers have been exploring various strategies, including the implementation of automated detection systems and machine learning techniques. Traditional methods of fraud detection, which are based on the creation of heuristics from known fraud indicators, have shown limitations (Roy & George, 2017). The evolving nature of fraud schemes necessitates more adaptable, data-driven, and intelligent detection tools.

Research and technological advancements in addressing the complexities of insurance fraud have seen a considerable increase, as noted by Nian et al (2016). This includes the application of data-driven initiatives and artificial intelligence. These approaches focus on analyzing and modeling the complex relationships between fraud indicators and suspicion of fraud, aiming to improve fraud detection with tools that are semi-automatic, understandable, and accountable.

II. RESEARCH AND INVESTIGATION INTO APPLICABLE TECHNIQUES

A. Data Sources and Characteristics

In the context of vehicle insurance fraud detection, a variety of data sources are utilized. These include transaction logs, which record all transactions related to insurance claims and payments. These logs are critical in identifying patterns that may suggest fraudulent activity. User profiles provide another rich source of data. They contain information about policyholders, including their claim history, which can be analyzed for inconsistencies or unusual patterns.

Historical data is also invaluable. It offers insights into past claims and fraud incidents, helping to identify trends and patterns that might indicate fraudulent behavior. Additionally, external data sources, such as public records, police reports, and data from other insurance companies, can be integrated to provide a more comprehensive view. These external sources can help in cross-verifying the information provided by claimants and in identifying potential fraud rings. (Benedek, Ciumas, & Nagy, 2022)

B. Domain Applications

Logistic Regression

Regression Analysis, particularly Logistic Regression, serves as a valuable tool for fraud detection techniques estimating the likelihood of a particular claim being fraudulent, based on a set of predictor variables. Logistic Regression is especially suitable for binary classification problems, where the outcome is dichotomous, such as determining whether an insurance claim is fraudulent (yes or no). Logistic Regression predicts the probability of occurrence of an event by fitting data to a logistic curve. It essentially models the relationship between a dependent binary variable and one or more independent variables.

In insurance fraud detection, the dependent variable is fraudulent status of a claim, while the independent variables could include numerous aspects of the claim, such as the amount of the claim, the history of the claimant, the nature of the incident, time factors, and other relevant details. Logistic Regression assesses how these variables affect the odds of a claim being fraudulent.

Logistic Regression models are relatively simple and computationally efficient, making them suitable for scenarios where quick fraud detection is crucial. They can also be easily updated with new data, which is essential in the dynamic environment of insurance fraud where patterns of fraudulent behavior can change rapidly. Overall, Logistic Regression is a robust and straightforward method for fraud detection, offering clear interpretability and efficiency, which makes it a

popular choice among data analysts and fraud investigators in the insurance industry (Itoo, Meenakshi, & Singh, 2021).

Decision Trees

Decision Trees is a fundamental machine learning technique used extensively in the area of fraud detection, including in the insurance industry. A Decision Tree works by splitting a dataset into smaller and smaller subsets based on different criteria. Each node of the tree represents a decision point where the data is divided, and the branches of the tree represent the outcomes of these decisions. At the end of the branches are the leaves, where the final decisions or predictions are made.

Decision Trees analyze various attributes of claims, such as the amount claimed, the claim history of the policyholder, the type of incident, etc. One of the key advantages of using Decision Trees is their simplicity and interpretability. The decision-making process is transparent and easy to follow, which is crucial in settings where understanding the reasoning behind a model's prediction is important. However, a notable limitation of a single Decision Tree is its tendency to overfit the training data, which can reduce its effectiveness on new data.

Random Forests

Random Forests elevates the concept of Decision Trees to a more advanced level. A Random Forest is an ensemble technique that combines the predictions from multiple Decision Trees to produce a more accurate and stable result. Each tree in a Random Forest is built from a sample drawn with replacement (bootstrap sample) from the training set. Moreover, when splitting a node during the construction of the tree, a random subset of the features is considered for splitting.

Random Forests offer several advantages. It works well at handling large datasets with numerous variables and can manage missing data effectively (even though in the current dataset there are no missing values). The ensemble nature of Random Forests also helps in mitigating the overfitting problem seen with individual Decision Trees. The aggregation of predictions from multiple trees typically leads to improved accuracy and generalization, making Random Forests a powerful tool for identifying fraudulent insurance claims. Despite their complexity compared to single Decision Trees, Random Forests still provide relatively interpretable results, allowing fraud analysts to understand and trust their predictions. (Xuan et al 2018).

Support Vector Machines (SVM)

Support Vector Machines (SVM) play a significant role in classification tasks, including fraud detection in insurance. At its essence, SVM is a powerful algorithm that excels in finding the optimal boundary between different classes of data. SVMs is known for their effectiveness in high-dimensional spaces, a typical characteristic of insurance data. Insurance datasets often involve numerous variables, such as the amount of the claim, the history of the claimant, the nature of the incident, policy details, and more. SVM's ability to handle this high dimensionality makes it a robust choice for fraud detection.

One of the key strengths of SVMs in this context is their versatility. This adaptability allows SVMs to perform well in various scenarios, from simple linear separations to complex, non-linear distributions.

Moreover, SVMs are known for their accuracy and efficiency, particularly in cases where the number of dimensions exceeds the number of samples, which is a common scenario in insurance fraud detection. However, SVMs do have their limitations. They can be computationally intensive, especially with large datasets, and the choice of the kernel and tuning of its parameters can be challenging, requiring careful consideration and expertise. SVMs offer a powerful and flexible approach to classifying data in the insurance fraud detection domain. Their ability to effectively handle high-dimensional data and their adaptability to various types of data distributions make them a valuable tool in the detection and prevention of insurance fraud (Singh, Gupta, Rastogi, Chandel, & Ahmad, 2012).

Gradient Boosting

Gradient Boosting is a highly effective machine learning technique widely used in classification tasks, including fraud detection in the insurance sector. Gradient Boosting constructs a predictive model in a stage-wise fashion, where each new model incrementally improves upon the previous ones. In the context of insurance fraud detection, Gradient Boosting offers significant advantages. It starts by building a simple model, usually a decision tree, and then progressively adds more models, each correcting the errors made by the previous ones.

One of the key strengths of Gradient Boosting in fraud detection is its ability to handle various types of data, including non-linear relationships. It can effectively work with both numerical and categorical data, making it versatile for the diverse and complex datasets typically found in insurance claims. Gradient Boosting is also known for its high level of accuracy. By combining multiple weak models, it often achieves superior performance compared to individual models or even other ensemble methods. This accuracy is particularly valuable in fraud detection, where the cost of misclassifying a legitimate claim as fraudulent (false positive) or failing to identify a fraudulent claim (false negative) can be very high.

Gradient Boosting is a powerful technique suitable for the complexities of insurance fraud detection. Its ability to progressively improve makes it a valuable tool for identifying fraudulent activity in insurance claims. Despite its complexity and computational demands, its high accuracy and effectiveness often make it a preferred choice in the fight against insurance fraud (Hancock & Khoshgoftaar, 2021)

III. IMPLEMENTATION OF THE SELECTED TECHNIQUE

A. A Implementation Steps

Data Collection

The dataset titled "Vehicle Insurance Fraud Detection," which was retrieved by “Kaggle’ website (Kaggle, 2021), represents a rich variety of diverse data aimed at identifying fraudulent activities in the vehicle insurance sector. This dataset includes broad aspects of insurance claims, policies, and personal information, along with vehicle-specific details and legal considerations.

Key Components of the Dataset

1. Claim Information: Iincludes time-related details such as the month, week of the month, and day of the week when the claim was made and subsequently filed.

2. Vehicle Information: Details such as the make of the vehicle, its category, price, and age are included.

3. Policy Information: Critical information about the insurance policy itself, including the type of policy, policy number, representative number, deductible amount, driver rating, age of the policyholder, and the base policy.

4. Personal Information: Attributes like the sex, marital status, and age of the policyholder are gathered. Personal demographics can sometimes correlate with claim patterns and are thus valuable in fraud detection analysis.

5. Claim and Policy History: Historical data such as the fault in past incidents and the number of previous claims are recorded.

6. Legal and Investigative Aspects: Legal factors including whether a police report was filed, the presence of witnesses, the type of agent involved, the number of supplements, the number of cars, and any address changes related to the claim are also part of the dataset.

The dataset is structured for binary classification, with the target variable "FraudFound_P" indicating the presence (1) or absence (0) of fraud in a claim. It consists of 15,421 instances, available in a CSV format, providing a solid foundation for developing and testing fraud detection models. The wide-ranging attributes included in the dataset are crucial for a detailed analysis, enabling the identification of potential fraudulent claims in the vehicle insurance sector.

Data cleaning

Upon thorough examination, it was observed that the dataset was remarkably well-maintained and there were no missing values and it was determined that the dataset did not require the typical data cleaning steps such as handling missing values. However, a critical step undertaken was the identification and handling of outliers. Outliers can significantly affect the performance of machine learning models, especially in complex tasks like fraud detection. Once identified by conducting boxplot analysis, the next step was to handle these outliers effectively. The IQR method was used for this purpose. By employing boxplot analysis and the IQR method, the dataset was cleansed of outliers, ensuring that the data used to train and test the machine learning models was of high quality and representative of typical cases.

Data Transformation

Techniques such as one-hot encoding and label encoding were employed which involves creating new binary (0 or 1) columns for each category of the variable. For example, there are three vehicle categories (sport, sedan and utility), three new columns are created, each representing one make, with a 1 or 0 indicating the presence or absence of that category. For ordinal categorical data, where the categories have a logical order, label encoding was used. This step is critical in bridging the gap between the qualitative nuances of the data and the quantitative demands of machine learning techniques.

Handling Imbalanced Data

A significant imbalance in the distribution of the target variable, "FraudFound_P," was identified. The dataset comprised 15,420 instances, out of which 14,497 were labeled as '0' (no fraud) and only 923 were labeled as '1' (fraud). Such an imbalance in the class distribution poses a challenge for machine learning models, as they can become biased towards the majority class, leading to poor performance in detecting the minority class.

To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized. SMOTE is an advanced oversampling method that generates synthetic samples for the minority class. This technique works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The application of SMOTE effectively balanced the dataset by augmenting the minority class with these synthetic samples, thereby reducing the model's bias towards the majority class. The use of SMOTE ensured that the predictive models developed later in the project would not only focus on the majority class but also give due importance to the minority class, which is vital for effective fraud detection in vehicle insurance.

Data Splitting

The last step was to divide the dataset into training and testing sets, guided by the 80/20 rule, a concept rooted in the Pareto Principle. This principle, commonly used in various business and economic contexts, posits that roughly 80% of consequences come from 20% of causes. Applying this concept to the data splitting process, the dataset was partitioned in such a manner that 80% of the data was used for training the machine learning models, while the remaining 20% was reserved for testing and evaluating their performance.

Using the 80/20 split ensures a substantial amount of data for training purposes, which is particularly important given the complexity of fraud detection tasks. At the same time, it retains a significant portion of the data for validation purposes, allowing for a robust evaluation of the models' predictive capabilities. This balance is essential to prevent issues like overfitting, where a model might perform exceptionally well on the training data but poorly on new, unseen data. The application of the Pareto Principle provides a structured approach to partitioning the dataset, ensuring that both the training and testing phases are adequately catered to.

IV. Findings and Business Value

A. Quantitative Results

Logistic Regression

Accuracy (0.755): achieves an accuracy of 0.755, indicating that it correctly classifies transactions as fraudulent or non-fraudulent around 75.5% of the time.

Precision (0 - Non-Fraud) (0.964): high precision for non-fraudulent transactions, meaning that when it predicts a transaction as non-fraudulent, it's correct 96.4% of the time.

Precision (1 - Fraud) (0.146): However, the precision for fraud cases is quite low. When it predicts a transaction as fraudulent, it's correct only 14.6% of the time.

Recall (0) (0.766): Recall for non-fraudulent transactions is reasonably good, suggesting that the model captures most non-fraudulent cases.

Recall (1) (0.584): Recall for fraud cases is moderate, indicating that the model detects 58.4% of actual fraud cases.

F1-Score (Class 0) (0.854): The F1-Score for non-fraudulent transactions is relatively high, implying a balance between precision and recall for non-fraud cases.

Logistic Regression shows an ability to correctly classify non-fraudulent transactions but struggles with fraud detection. It has a high false positive rate for fraud cases.

Decision Tree

Accuracy (0.858): achieves an accuracy of 0.858

Precision (0) (0.950): It has high precision for non-fraudulent transactions, indicating a low false positive rate for non-fraud cases.

Recall (0) (0.896): Recall for non-fraudulent transactions is good, implying that the model captures most non-fraud cases.

F1-Score (0) (0.922): The F1-Score for non-fraudulent transactions is high, showing a good balance between precision and recall.

Decision Tree has a good balance between precision and recall for non-fraud cases but falls short in fraud detection. It suffers from a high false positive rate for fraud cases.

Random Forest

Accuracy (0.894): Random Forest achieves a higher accuracy than both Logistic Regression and Decision Tree.

Precision (0) (0.945): Like Decision Tree, it has high precision for non-fraudulent transactions.

Precision (1) (0.187): Precision for fraud cases is improved compared to Decision Tree but is still relatively low.

Recall (0) (0.941): Recall for non-fraudulent transactions is high, indicating good coverage of non-fraud cases.

F1-Score (0) (0.943): The F1-Score for non-fraudulent transactions is high, reflecting a good balance between precision and recall.

Decision Tree maintains high precision for non-fraud cases but struggles with recall for fraud cases.

Support Vector Machine (SVM)

Accuracy (0.827): SVM achieves a good overall accuracy.

Precision (0) (0.953): It has high precision for non-fraudulent transactions, indicating a low false positive rate for non-fraud cases.

Precision (1) (0.154): Precision for fraud cases is low, meaning it has a high false positive rate for fraud detection.

Recall (0) (0.857): Recall for non-fraudulent transactions is decent, capturing most non-fraud cases.

Recall (1) (0.381): Recall for fraud cases is higher compared to Logistic Regression and Decision Tree, suggesting a better ability to detect fraud.

F1-Score (0) (0.903): The F1-Score for non-fraudulent transactions is relatively high, showing a balance between precision and recall.

SVM achieves a good balance between precision and recall for non-fraud cases and shows improved recall for fraud cases compared to some other models.

Gradient Boosting

Accuracy (0.803): Gradient Boosting provides a decent overall accuracy.

Precision (0) (0.960): It has the highest precision for non-fraudulent transactions among all models, indicating a low false positive rate for non-fraud cases.

Recall (0) (0.824): Recall for non-fraudulent transactions is moderate, capturing most non-fraud cases.

Recall (1) (0.492): Recall for fraud cases is the highest among all models, indicating its ability to detect a significant portion of actual fraud cases.

F1-Score (0) (0.886): The F1-Score for non-fraudulent transactions is relatively high, showing a balance between precision and recall.

F1-Score (1) (0.242): The F1-Score for fraud cases, while the highest among all models, still indicates a need for improvement in fraud detection.

B. Business Value

Interpreting the business value of the findings from the five models involves understanding how these models' performance translates into practical, real-world benefits for insurance companies.

Logistic Regression

This model's high accuracy in identifying non-fraudulent transactions is valuable for minimizing unnecessary investigations into legitimate claims, which can save time and resources. However, its low precision for fraud detection indicates a higher number of false positives in fraud cases, potentially leading to customer dissatisfaction and increased operational costs.

Logistic Regression model is ideal for fast, preliminary assessment of claims due to its high accuracy with non-fraudulent transactions. However, its low precision for fraud detection could lead to a high number of false positives. This means the model may incorrectly flag legitimate claims as fraudulent, potentially increasing customer dissatisfaction and the workload for manual review teams. Insurance companies might use Logistic Regression for initial claim filtering, but they should be prepared for additional layers of verification.

An example of using this model is that an insurance company uses Logistic Regression for initial claim assessment. While it efficiently processes a high volume of claims, identifying most legitimate claims correctly, it flags an unusual number of claims as suspicious. After manual review, many of these flagged claims are found to be legitimate, leading to increased workload for the review team and frustration among customers who experience delays.

Decision Tree

The Decision Tree model shows a good balance in identifying non-fraudulent cases but needs improvement in detecting fraud. Its tendency to false positives in fraud detection could result in wasted resources on investigating legitimate claims. Decision Tree model offers a transparent and interpretable model, making it easier to understand and explain why certain claims are flagged as potentially fraudulent. However, its lower precision in fraud detection could lead to a substantial number of false alarms, which may strain resources.

A possible example of applying the Decision Tree model is when it implemented to understand common patterns in fraudulent claims. The model identifies that a high frequency

of claims involving certain vehicle makes and in specific geographic locations are often fraudulent. While this insight is valuable, the model also flags a significant number of legitimate claims in these categories as fraudulent, leading to resource-intensive manual reviews. This example shows the utility of Decision Trees in hypothesis generation, but also highlights the need for caution due to the potential for false positives.

Random Forest

Random Forest's high overall accuracy and precision in identifying non-fraudulent cases make it a strong tool for effectively screening claims. Its moderate performance in fraud detection, however, suggests potential missed opportunities in identifying fraudulent activities. Random Forest can be an effective tool in automating parts of the claim assessment process. Its high accuracy and precision for non-fraudulent cases allow it to reliably filter out legitimate claims. However, its lower recall for fraudulent claims means it might miss some instances of fraud, necessitating.

To enhance claim processing efficiency, an insurance company employs a Random Forest model. The model excels in identifying genuine claims with high accuracy. However, it also misses some fraudulent claims, which are only caught during random audits. This situation reflects the model's trade-off between high precision for non-fraudulent cases and lower recall for fraudulent cases, necessitating a mix of automated and manual oversight.

Support Vector Machine (SVM)

SVM's balanced performance in precision and recall for non-fraud cases, along with improved recall for fraud detection, indicates its potential as a reliable screening tool. The improved ability to detect fraud cases, even with some false positives, can be valuable in reducing fraud losses. SVM's balanced performance makes it a strong candidate for integrated systems where both types of classifications (fraudulent and non-fraudulent) are important. It can help reduce the workload on manual review teams by accurately identifying many non-fraudulent cases. However, its lower precision in fraud detection could still result in false positives, requiring careful consideration in customer-facing processes to avoid damaging customer trust.

For instance, an insurance firm using an SVM model into their claim processing system where the model performs well in differentiating between fraudulent and non-fraudulent claims with a fair balance. However, it occasionally marks some complex, but legitimate, claims as fraudulent, leading to some customer complaints about the claim handling process.

Gradient Boosting

With the highest recall for fraud cases among the models, Gradient Boosting is particularly adept at identifying a larger portion of actual fraud cases. This capability is extremely beneficial in mitigating financial losses due to fraud.

The model excels in identifying a higher proportion of actual fraud cases (high recall), making it invaluable in minimizing financial losses due to fraud. However, its tendency to produce false positives (lower precision) can lead to inefficiencies in fraud investigation processes and potential harm to customer relations. Gradient Boosting can be strategically deployed where catching the highest possible number of fraudulent claims is prioritized, perhaps in high-risk segments.

For example, it can be used by a company in order to detect as many fraudulent claims as possible. The model successfully uncovers a higher proportion of fraud cases, including some sophisticated fraud scenarios that other models miss. However, this comes at the cost of mistakenly classifying some complex but legitimate high-value claims as fraud, leading to an increased burden on the fraud investigation team and some dissatisfied high-value customers

Overall Business Implications

Each model offers distinct advantages and areas for improvement. For an insurance company, the choice of a model or a combination of models should align with its specific operational priorities, whether it's maximizing accuracy, reducing false positives, or effectively identifying as many fraudulent cases as possible. The ultimate goal is to create a fraud detection system that not only protects the company's financial interests but also upholds customer satisfaction and trust.

REFERENCES

[1] Benedek, B., Ciumas, C., & Nagy, B. Z. (2022). Automobile insurance fraud detection in the age of big data–a systematic and comprehensive literature review. Journal of Financial Regulation and Compliance, 30(4), 503-523.

[2] Cody, C., Ford, V., & Siraj, A. (2015, December). Decision tree learning for fraud detection in consumer energy consumption. In 2015 IEEE 14th international conference on machine learning and applications (ICMLA) (pp. 1175-1179). IEEE.

[3] Hancock, J. T., & Khoshgoftaar, T. M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection. SN Computer Science, 2(4), 268.

[4] Insurance Ireland. (2020, May 28). insuranceireland. Retrieved from https://www.insuranceireland.eu/

[5] Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13, 1503-1511.

[6] Kaggle. (2021). Kaggle. Retrieved from https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection

[7] Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58-75.

[8] Roy, R., & George, K. T. (2017, April). Detecting insurance claims fraud using machine learning techniques. In 2017 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-6). IEEE.

[9] Singh, G., Gupta, R., Rastogi, A., Chandel, M. D., & Ahmad, R. (2012). A machine learning approach for detection of fraud based on svm. International Journal of Scientific Engineering and Technology, 1(3), 192-196.

[10] Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research, 176(1), 565-583.

[11] Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. Expert systems with applications, 29(3), 653-666.

[12] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In 2018 IEEE 15th international conference on networking, sensing and control (ICNSC) (pp. 1-6). IEEE.