# Predicting Insurance Fraud with Machine Learning

Oluwaseun Oluwasola Mustapha*

September 19, 2022

## Contents



*Mathematical Sciences Department, University of Essex, United Kingdom (e-mail: om21222@essex.ac.uk).

**Abstract**

Insurance fraud is when an insured, claimant, or company intentionally makes a false or misleading claim in order to profit financially. Insurance applicants, policyholders, third-party claimants, or professionals like insurance companies or agents that offer such services can all perpetrate insurance fraud at various times along the insurance life-cycle. This study examines the efficiency and confirmability of the most popular machine learning algorithms for fraud prediction. Product recommendation, Medical Diagnosis, Image Detection amongst others in recent years have all greatly benefited from the current machine learning algorithms which has helped in the improvement of medical advancements and public safety.

In this paper, we perform an analysis using 5 different algorithms, which are Random-Forest (RF), Decision Tree (DT), Ada Boost, K-Nearest Neighbor (KNN) and XG Boost to detect the Insurance fraud and our findings draws a conclusion that Decision Tree produces the highest accuracy of 85% when compared against other techniques. Experts and risk analysts can make future insurance guidelines and policies based on the resulting benefits and drawbacks of the models in this study.

# 1 Introduction

As a result of substantial information available in this current era, a lot of people are beginning to invest in various Insurance policies. With increasing involvements in insurance policies also comes increasing insurance frauds. Insurance fraud is an Intentional deception conducted against or on behalf of an insurance business or agent with the goal of making a profit. Over the years a sizeable amount of money has been moved across the insurance industry as it is a quite profitable one. Insurance fraud also accounts for a significant amount of the costs incurred by insurance companies because it reduces their earnings and has a long-term impact on their pricing policies. A few million dollars were misappropriated each year through insurance fraud. For example, In 2017, the Australian Insurance Fraud Bureau found false claims worth $280 million [Insurance Fraud Bureau of Australia (2021)]. Additionally, 44814 fraud claims totalling EUR 214 million were discovered in France in 2013 with the amount increasing to EUR 500 million in 2018. [French Agency for the Fight against Insurance Fraud (2021)]. Statistics from the United States Coalition Against Insurance Fraud show that insurance fraud represents 17% of the total compensation paid out by insurance companies, with an estimated annual value of $80 billion. [S.Jordon. (2016)] and according to the Insurance Bureau of Canada (IBC), car insurance fraud totalled more than 542 million Canadian dollars in 2007. Chinese insurance officials went further to report that an average of RMB 35 billion annually in 2011 was lost to insurance fraud, accounting for around 20% of all insurance company payments. Finally, it is estimated that insurance fraud costs developing nations $600 million annually. Therefore, insurance fraud is an international problem that harms the nation and the community.

# 2 Literature Review

Since the development of artificial intelligence theory, machine learning techniques have been widely used for fraud detection. Itri, Mohamed, Mohammed, and Omar (2019) developed a fresh strategy to raise fraud prediction accuracy. Ten machine learning fraud prediction algorithms were evaluated for effectiveness and confirmability. They took advantage of auto insurance claim data. The study showed that Random Forest outperformed all other algorithms in predicting fraud. A method of analysis called inductive reasoning serves as the foundation for machine learning techniques, which derives conclusions from data patterns without making assumptions about the functional form of things like probability distributions or linearity. Peng (2020) discuss how this flexibility necessitates extra caution to control the models' balance between generalisation ability and complexity because different models or even minor changes to their hyperparameters can have significant effects on the performance of prediction when applied to same dataset.

Fraud detection is one of the major applications of machine learning in business administration and finance. From the viewpoint of a decision-maker, this topic is extremely relevant because decision support systems that assist risk analysts in predicting fraudsters directly affect a company's financial performance. This topic has been investigated by a big number of researchers in recent years, as discussed in papers like Awoyemi, Adetunmbi, and Oluwadare (2017) , Ngai, Hu, Wong, Chen, and Sun (2011) , Raghavan and El Gayar (2019) , Waghade and Karandikar (2018)

Sheshasaayee and Thomas (2018) discuss the benefits of using machine learning methods to perform such tasks, especially concerning the most salient features of fraudsters, while illustrating the main challenges faced by risk and fraud analysts in developing fraud identification mechanisms and decision rules in light of the fact that the presence of fraud entails significant profit losses for the insurance sector. In a similar vein, Dal Pozzolo, Caelen, Le Borgne, Waterschoot, and Bontempi (2014) discussed the complexity involved in developing a data-driven fraud detection algorithm, highlighting typical issues such as the highly unbalanced class dis-

tributions, non-stationary distribution of the data, a lack of readily accessible microdata due to confidentiality issues and an ongoing large stream of new transactions. In order to address the issues mentioned, the writers assessed three machine learning models' ability to predict outcomes (the random forest, the neural network and support vector machine) utilizing a dataset derived from actual credit card transactions. They also examined the overall effects of update periodicity, the use of balancing techniques, and the retention of older observations in the training dataset. For all training approaches, the results showed the random forest model continuously outperforming support vector machines and neural networks. In addition, models that were updated with fresh data more frequently also performed better, suggesting how fast the distribution of fraud can change in intervals. In terms of the problem of imbalanced classes, methods of balancing were applied to increase how they perform over "static" dataset which was unbalanced, in which the random forest performed worst. Finally, compared to keeping the dataset balanced, the method of removing earlier observations showed a marginally lesser benefit.

Wang and Xu (2018) used support vector machine (SVM), random forest, and deep neural networks as the evaluated models to analyse the explanations of automobile accidents in a bid to forecast frauds for automobile insurance claims. All three models achieved an F1 Score greater than 75%. On the other hand, Roy and George (2017) used Naive Bayes and Random forest classifiers in identifying vehicle claims fraud, and they discovered that Random forest outperformed Naive Bayes. Similar to this, Yao, Zhang, and Wang (2018) proposed a financial fraud detection technique that combines feature selection with machine learning classification methods. PCA(Principal Component Analysis) and Gradient Boosting Technique(XGBoost) were utilized to start with high-dimensional data and find the key elements of information. A few machine learning models were employed, with the random forest having the greatest out-of-sample performance.

On the other hand, Eshghi and Kargari (2019) argued that unsupervised methods like clustering and outlier detection may not be sufficient for difficult fraud detection tasks and suggested a framework with Multi-Criteria Decision Analysis and intuitionistic fuzzy sets to take the impact of behavioural uncertainties into account when modelling the likelihood that a banking transaction is fraudulent. In a similar vein, Carcillo et al. (2021) argued in favour of combining supervised and unsupervised learning techniques for credit card fraud detection in order to more effectively adapt to changes in consumer behaviour and fraudsters' capacity to create original fraud patterns. Based on clustering analysis, the authors created outlier scores for various granularities, applied them to a real-world dataset, and reported an improvement in detection effectiveness.

In a recent study, [Kim, Baik, and Cho (2016)] suggested a 'multi-class algorithm' to detect fraud intention in financial mis-statements using MetaCost (Domingos, 1999) to account for the unbalanced classes, using asymmetric misclassification costs; and [Varmedja, Karanovic, Sladojevic, Arsenovic, and Anderla (2019)] used SMOTE (Synthetic Minority Oversampling Technique) to balance the training data together with Neural Network, Naive Bayes, Random Forest and Logistic Regression as machine learning algorithms.

Xu, Wang, Zhang, and Yang (2011) suggests a random rough subspace neural network ensemble-based method for detecting insurance fraud. In order to create a set of reductions that can maintain stable data information, this method starts with a crude set reduction. Next, a subset of reductions is assembled by selecting reductions at random. Then, a neural network classifier is trained using each of the chosen reductions on the insurance data. The qualified neural network classifiers are then combined using ensemble techniques. Additionally, the effectiveness and efficiency of the suggested strategy are examined using a real-world vehicle insurance situation. The findings of their research demonstrate that a random rough subspace dependent neural network ensemble technique can detect fraudulent insurance claims more quickly and accurately, making it a viable tool for detecting insurance fraud.

Therefore, it is important to examine which machine learning models can more effectively identify fraud trends and precisely anticipate future offences using real-world data and machine learning techniques. In addition, given the wide range of frauds that can occur, each of which has its own characteristics and method of operation [Gottschalk (2010)], this study's concentration is on frauds where the consumer is the perpetrator, restricted to policy claims of automobile insurance. The fact that the data for this study were obtained from a significant insurance provider adds to our understanding of the relative significance of the database elements, which is crucial for evaluating insurance policies in the real world.

# 3 PROPOSED MODEL (Data Collection, Exploration & Pre-Processing

Data overview, pre-processing, applying models, and performance evaluation are the primary steps of the suggested model. Every phase of the suggested model is crucial and improves its effectiveness. The proposed model for detecting insurance fraud in this work is displayed in Figure 1 below.

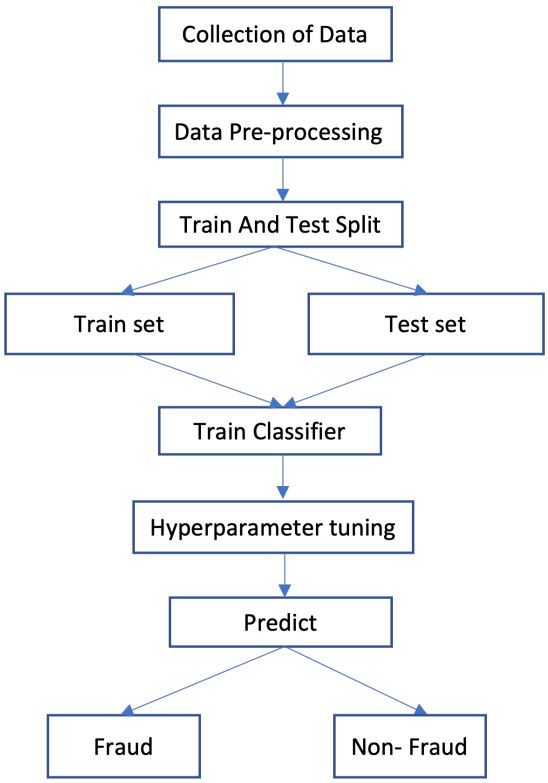


Figure 1: Suggested Model Of Fraud Insurance Detection.

### 3.0.1 Data Overview

The dataset used in this study was gotten from Kaggle here: (`https://www.kaggle.com/code/niteshyadav310`
The dataset comprises 1000 automobile incidents and auto-insurance claims from the geographical locations of Indiana, Illinois and Ohio during the period of January 1st 2015 to March 1st 2015. A total of 39 variables constitutes the data set and each of these features are explained in Table 1.

### 3.0.2 Data Pre-processing

Amongst the most important process in Machine Learning happens to be the pre-processing of Data. This is simply transforming data from its raw form into a format which can be understood by machine learning models. Errors and Missing values are almost impossible not to be encountered in Datasets and this is what we try to manage in this phase [Peng (2020)]; For instance, the outliers in the dataset were detected, we also tried to categorize correctly the Numeric and Categorical data before finally proceeding to utilizing the Machine Learning Models for analysis of our processed dataset. This are briefly illustrated below:


Figure 2: Visualizing outliers

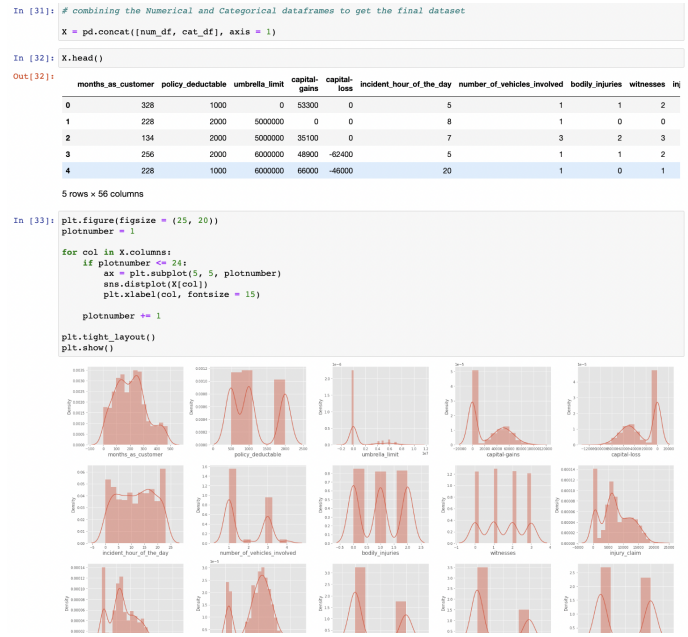


Figure 3: Encoding Categorical and Numerical data

| @*8R@ height**Variable** | **Description** |
|---|---|
| months_as_customer | The duration in months as a customer |
| age | The ages of the insured |
| policy_number | The policy numbers of the insured |
| policy_bind_date | The time period between the effective date of coverage and when the policy or endorsement is is |
| policy_state | The state in which the policy was issued |
| policy_csl | The maximum amount that covers any combination of injuries or property damage in an inciden |
| policy_deductable | The amount of money that you are responsible for paying toward an insured loss. |
| policy_annual_premium | The amount of Regular Premium payable by the Policyholder in a Policy Year |
| umbrella_limit | Provision of excess limits and additional excess coverage over the normal limits and coverage of |
| insured_zip | Zip code of insured |
| insured_sex | Gender of insured |
| insured_education_level | Education level of the insured |
| insured_occupation | Occupation of the insured |
| insured_hobbies | Hobbies of the insured |
| insured_relationship | The marital/relationship status of the insured |
| capital-gains | When calculating the gain of an item, any insurance compensation is the deemed proceeds |
| capital-loss | When calculating the loss of an item, any insurance compensation is the deemed proceeds |
| incident_date | For an automobile wreck or other accident, the date of incident is the date of the accident |
| incident_type | The Incident Type is the actual situation found upon arrival to the scene |
| collision_type | The type of collision that happened |
| incident_severity | How severe the incident was |
| authorities_contacted | The contacted authorities the incident was reported to |
| incident_state | The state in which the incident occured |
| incident_city | The city in which the incident occured |
| incident_location | The exact location of the incident |
| incident_hour_of_the_day | Hour of the day in which the incident occured |
| number_of_vehicles_involved | The number of vehicles involved in the accident |
| property_damage | What were the damaged properties in the accident |
| bodily_injuries | Record of body injuries sustained in accident |
| witnesses | Who were those that witnessed the accident |
| police_report_available | Is there an available police report about accident |
| total_claim_amount | The total amount claimed by insured |
| injury_claim | The amount claimed on injury by insured |
| property_claim | The amount claimed on properties by insured |
| vehicle_claim | The amount claimed on vehicle by insured |
| auto_make | The maker of the automobile |
| auto_model | The model of the automobile |
| auto_year | The year automobile was produced |
| fraud_reported | Was fraud reported or not |
| _c39 | No description for this variable as cells are blank |

Table 1: Description of Variable

# 4 MACHINE LEARNING ALGORITHMS

The different machine learning algorithms used in this study are the Random-Forest (RF), Decision Tree Classifiers (DTC), AdaBoost, K-Nearest Neighbor (KNN) and XG-Boost (XG) which we will do a brief overview of each in details below.

- **Decision Tree Classifiers:** Like the name implies, a decision tree is one that helps in making decisions. It does this with the help of a root node, an internal node, several branches and finally a decision node. The decision tree can be likened to factorization where a larger chunk is broken down into smaller/simpler chunks for easier interpretation. Several fields have adopted the Decision tree, one of which is largely the medical diagnosis. This is illustrated in Figure 4

- **Random Forest:** One of the most used supervised learning techniques In most cases, it produces precise results with little to no information planning, showing, or modelling required. Decision trees from past segments are used in Random Forests. More specifically, Random Forests are collections of decision trees that produce better forecast accuracy. It is essentially a lot of decision trees, which is why it is termed as a "forest." The basic idea is to create distinct decision trees based on the independent subsets of the dataset. The optimal split on n factors is determined at each node by randomly selecting one from the list of capabilities. An illustration of the Random Forest algorithm is seen in Figure 5

- **Adaboost Classifier:** AdaBoost which is also referred to as Adaptive Boosting, an ensemble classifier which combines two or more weaker classifiers in a bid to achieve a stronger accuracy classifier. The Adaboost Algorithm simply adds predictors gradually to the pair/group of classifiers (ensemble) in a bid to make it better.One major disadvantage of the Adaboost is the fact that each of the predictors have to wait for each other in other to be trained which increases the processing time. An illustration is seen in Figure 6

- **K-Nearest Neighbor (KNN):** K-Nearest Neighbor regressor is employed when there is no presumption on the frequency distribution of the relationship between the predictor variables. By averaging certain features in the same area, this is achieved. K-Nearest Neighbor has been found to be the best classification technique, even though this study doesn't totally agree. K is calculated using the square root of the total amount of data in the training data set. However, it hasn't always been like this. The yield in the KNN classification problem would be a class to which the data model has a spot and which is predicted by the majority vote of the k nearest neighbours. The property estimation, which is typically a mean estimation of the k nearest neighbours, would be the yield in the regression problem. The Euclidean Distance can be used to resolve the nearest neighbour division:

$$EuclideanDistance = \sqrt{\sum_{i=1}^{N}(q_i - p_i)^2; pandqarethepointsinn - space} \tag{1}$$

  The calculation's expected accuracy depends heavily on the estimation of k. Since each instance in the preparation set currently carries a bigger weight throughout the decision process, smaller predictions of k will likely result in lower precision, especially in datasets with a lot of noise. The calculation's display is reduced as k's estimation values increase. Additionally, if the estimation value is too high, the model may overfit, weakening the distinction between class bounds and causing once more decreased exactness or accuracy. As a general methodology, it is advised to select k by applying the formula:

$$k = \sqrt{n} \tag{2}$$

  We take a look at the algorithm for K-Nearest Neighbor in Figure 7

- **XG-Boost:** The XGBoost (Extreme Gradient Boosting) method adds models into an ensemble of classifiers iterating through in a cycle. It is a repetitive method of using initial ensembles in generating several predictions for individual observations, it then proceeds to fit a new model which it adds to the ensemble. The XGBoost is greatly affected by some parameters such as the 'n-estimators' and 'learning-rate' which in turn can influence the training speed and also the accuracy speed. It also happens to be one of the top ranked machine learning techniques because of it's enablement of the parallel tree boosting. Illustrated in Figure 8 is the XG-Boost algorithm

### 4.1 HYPERPARAMETER TUNING

Let's liking hyperparameter tuning to the tuning of a guitar or a drum set for better output. Hyperparameters are variables whose values control learning and specify the model parameter values that a learning algorithm will finally learn. Hyperparameter tuning simply put is finding a set of ideal hyperparameter values for a learning algorithm and using this improved algorithm on every given data set. We are going to be using the GridSearchCV for the tuning the parameters in this project.

## 5 EVALUATION METHOD

Methods of evaluation are crucial for model comparison and best model selection. Considering that they are evaluating the effectiveness of classifiers (Hossin (2015). Since bias can sometimes be introduced for a majority class in classification problems due to imbalanced data, accuracy alone is not always dependable(Ganganwar (2012)(Hanafy and Ming (2021). Car insurance claims are a prime example of imbalanced data because majority of policyholders don't engage in fraud. Therefore, if accuracy is the only criterion, there will be bias against a fraud class. Therefore, several measurement techniques are employed, including F1-score, sensitivity, accuracy, and region under the curve (AUC). AUC might be a better metric to employ if results need to balance sensitivity and specificity, especially when there is an imbalanced class distribution.

$$SENSITIVITY = \frac{TP}{(TP + FN)} \tag{3}$$

$$ACCURACY = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

$$SPECIFICITY = TN(FP + TN) \tag{5}$$

$$PRECISION = \frac{TP}{(TP + FP)} \tag{6}$$

$$F - MEASURE = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{7}$$

In this formula, TP stands for the number of true positives, FP stands for false positives, TN stands for true negatives, and FN stands for false negatives. A greater Accuracy

rating indicates a better overall performance of the forecast. Accuracy is a measure of the percentage of predictions that are accurate. Sensitivity has to do with the accuracy of fraud claims detection. The ability to accurately identify legal claims is referred to as specificity. Precision is the measure of the importance of the projected positives. And The F1 score is the precision and sensitivity harmonic average. The overall classifier performance metric is AUC (Wu and Flach (2005). It is employed to evaluate the model's overall performance.

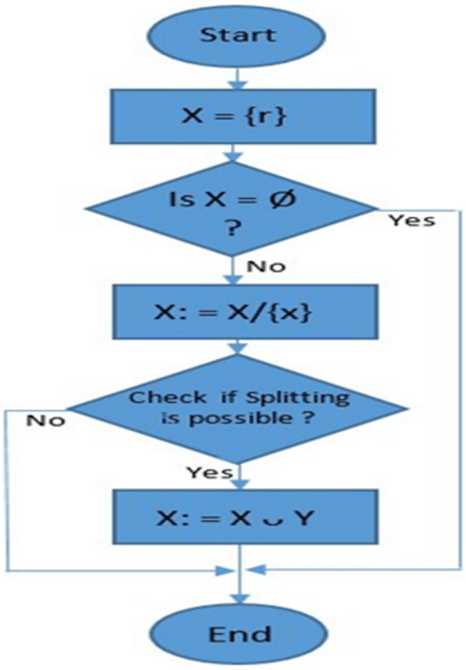


Figure 4: Flowchart of Decision Tree Classifiers

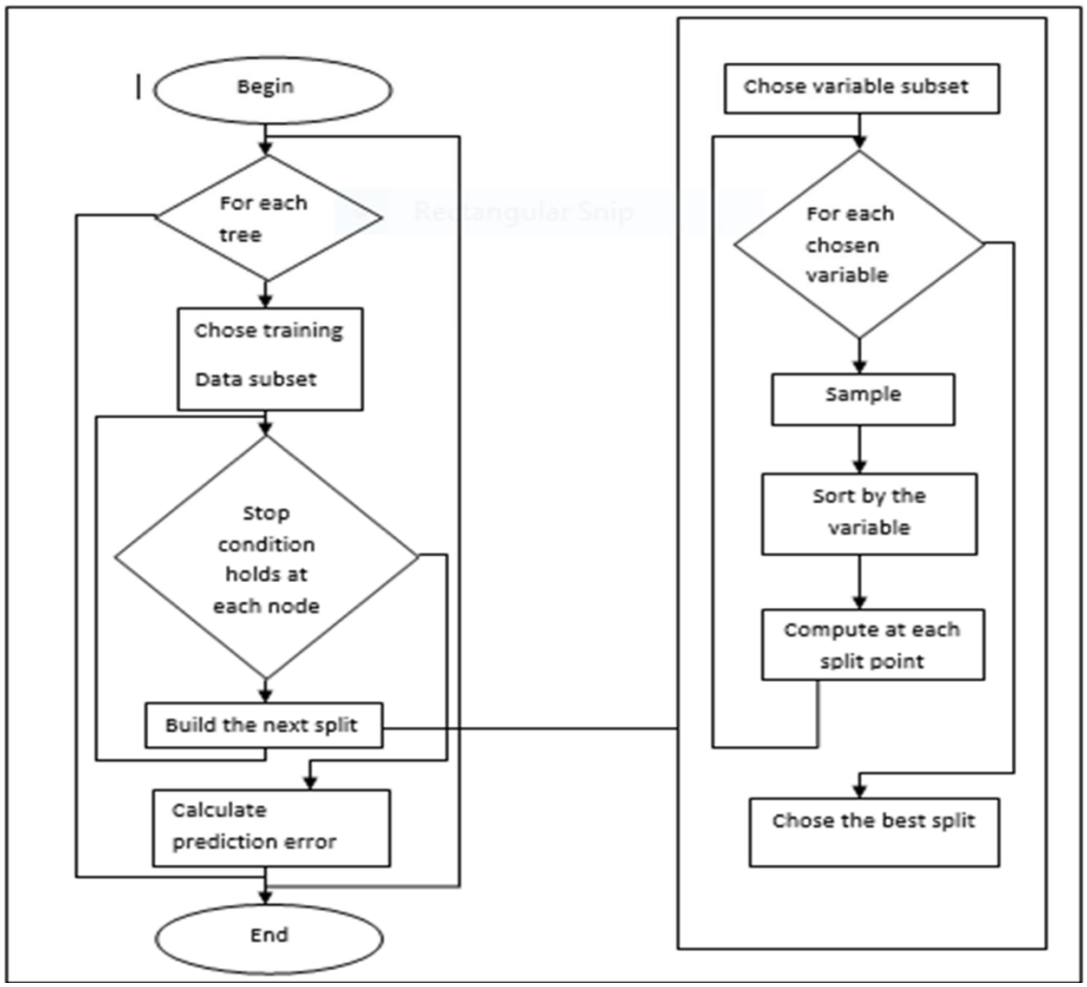


Figure 5: Flowchart of Random Forest Algorithm

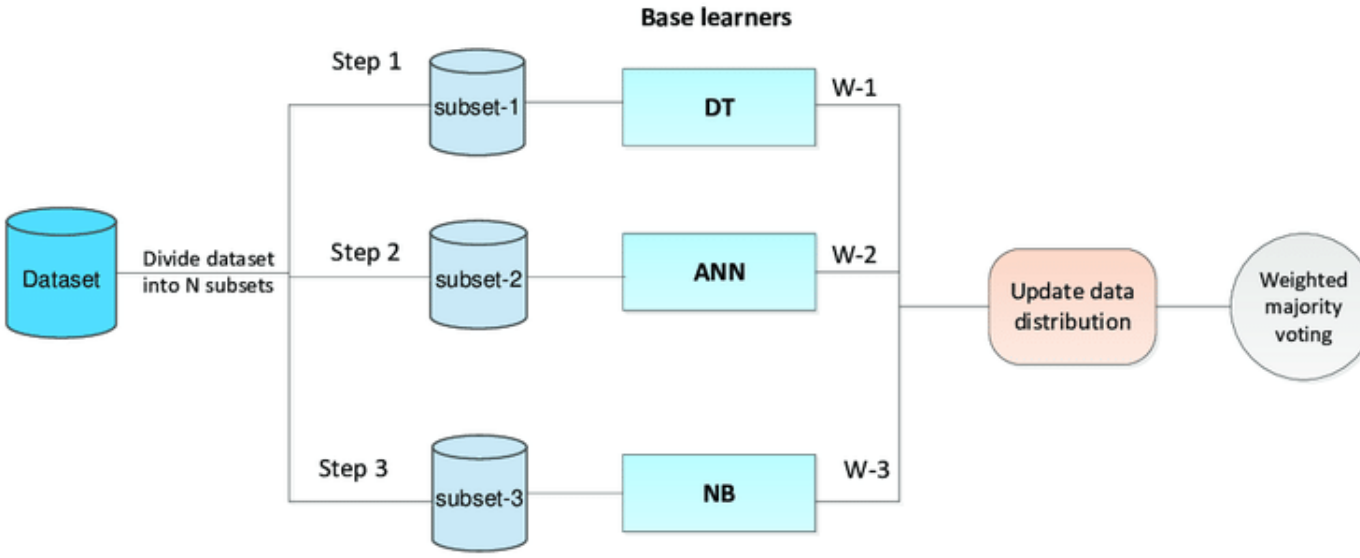


Figure 6: AdaBoost Flowchart

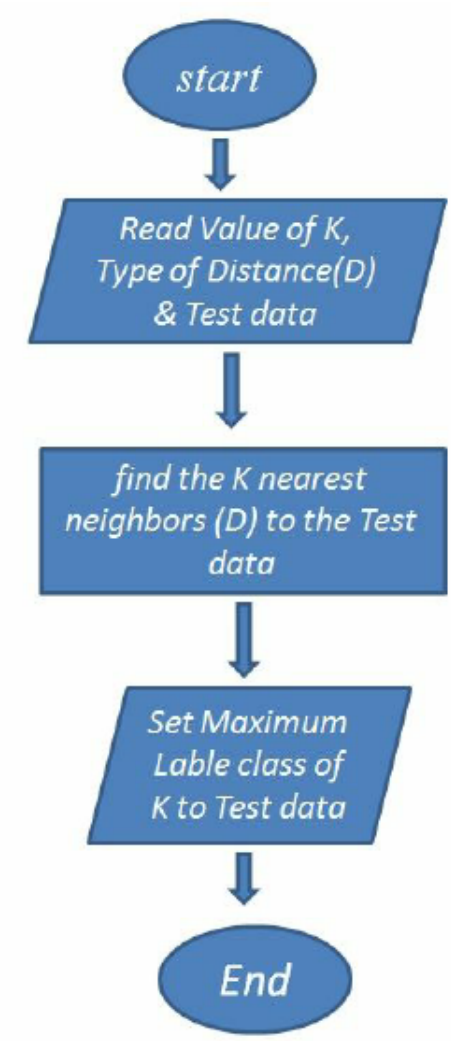


Figure 7: Flowchart of K-Nearest Neighbor Algorithm

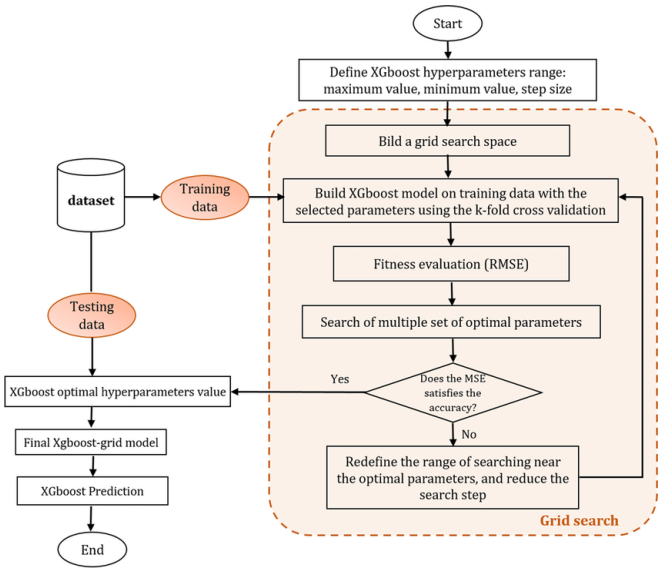


Figure 8: Flowchart of XG-Boost

# 6 RESULTS

Fraud must be addressed because it is a significant issue in today's society. We can create systems that can detect fraud in the supplied data to overcome these issues. These systems are created utilising a variety of machine learning methods, including neural networks, naive Bayes, KNN, and random forests. In this paper, we've spoken about different machine learning (ML) techniques, how they work in systems, and how good they are at predicting fraud. These methods are then contrasted using five criteria from various angles. Additionally, we randomly divide the data and utilise 80% in training whilst we use 20% for testing so as to assess how well the machine learning algorithms perform on fraud discrimination. We use the training data to train the machine learning algorithms, and the learned model is then used to forecast whether or not the examples in the test data are fraudulent. Six assessment techniques are used to assess the performance of the models on the testing data: accuracy, sensitivity, specificity, AUC, precision, and F1-score. Additionally, we set each machine learning model to its peak performance in order to make the comparisons as fair as feasible.

```
In [35]: # splitting data into training set and test set

         from sklearn.model_selection import train_test_split

         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)
```

```
In [36]: X_train.head()
```

Out[36]:

| | months_as_customer | policy_deductable | umbrella_limit | capital-gains | capital-loss | incident_hour_of_the_day | number_of_vehicles_involved | bodily_injuries | witnesses |
|---|---|---|---|---|---|---|---|---|---|
| **218** | 328 | 500 | 0 | 24800 | 0 | 0 | 1 | 2 | 3 |
| **96** | 325 | 1000 | 0 | 61500 | 0 | 11 | 1 | 0 | 3 |
| **301** | 107 | 2000 | 5000000 | 20000 | -82700 | 21 | 3 | 1 | 2 |
| **724** | 284 | 500 | 0 | 48200 | 0 | 12 | 3 | 0 | 0 |
| **63** | 215 | 500 | 0 | 0 | -49000 | 20 | 3 | 2 | 2 |

5 rows × 56 columns

```
In [37]: num_df = X_train[['months_as_customer', 'policy_deductable', 'umbrella_limit',
                'capital-gains', 'capital-loss', 'incident_hour_of_the_day',
                'number_of_vehicles_involved', 'bodily_injuries', 'witnesses', 'injury_claim', 'property_claim',
                'vehicle_claim']]
```

```
In [38]: # Scaling the numeric values in the dataset

         from sklearn.preprocessing import StandardScaler

         scaler = StandardScaler()
         scaled_data = scaler.fit_transform(num_df)
```

```
In [39]: scaled_num_df = pd.DataFrame(data = scaled_data, columns = num_df.columns, index = X_train.index)
         scaled_num_df.head()
```

Out[39]:

| | months_as_customer | policy_deductable | umbrella_limit | capital-gains | capital-loss | incident_hour_of_the_day | number_of_vehicles_involved | bodily_injuries | witness |
|---|---|---|---|---|---|---|---|---|---|
| **218** | 1.048359 | -1.021808 | -0.479902 | -0.028710 | 0.954892 | -1.663058 | -0.815084 | 1.217594 | 1.3842 |
| **96** | 1.022382 | -0.195993 | -0.479902 | 1.287503 | 0.954892 | -0.082387 | -0.815084 | -1.182394 | 1.3842 |
| **301** | -0.865337 | 1.455636 | 1.720136 | -0.200858 | -1.991191 | 1.354588 | 1.156894 | 0.017600 | 0.4759 |
| **724** | 0.667352 | -1.021808 | -0.479902 | 0.810510 | 0.954892 | 0.061311 | 1.156894 | -1.182394 | -1.3406 |
| **63** | 0.069863 | -1.021808 | -0.479902 | -0.918141 | -0.790671 | 1.210890 | 1.156894 | 1.217594 | 0.4759 |

```
In [40]: X_train.drop(columns = scaled_num_df.columns, inplace = True)
```

```
In [41]: X_train = pd.concat([scaled_num_df, X_train], axis = 1)
```

Figure 9: Train-Test Split

# 7 CONCLUSION

After performing this study, the DecisionTree Classifier turned out to be the most optimal with an 85% accuracy. The model predicts 39 true positives from 54 positive cases, 173 true negatives from 196 cases, 23 false positives from 196 positive cases and 15 false negatives from 54 cases. An f1 score of 85% is obtained.

An auto insurance fraud detection model has been developed as part of this research. The model can lessen losses for insurance firms in this way. The difficulty with machine learning for fraud detection is that fraudulent insurance claims are far less frequent than legitimate ones.

In this study, six different classifiers were employed: Decision Tree Classifier, K-nearest Neighbors, Random Forest Classifier, ADA-Boost Algorithm, XG-Boost and the Voting Classifier. These six classifiers were used together with the hyperparameter adjustment to address imbalance problems.

A Decision Tree model with a score of 0.85 for F1 and a 0.90 ROC AUC turned out best and final fitted model. The model worked perfectly. The model had the highest ROC AUC and F1 scores when compared to the other models. In conclusion, the model proved highly accurate at differentiating between legitimate and fraudulent claims.

```
In [67]: # accuracy_score, confusion_matrix and classification_report

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

dtc_train_acc = accuracy_score(y_train, dtc.predict(X_train))
dtc_test_acc = accuracy_score(y_test, y_pred)

print(f"Training accuracy of Decision Tree is : {dtc_train_acc}")
print(f"Test accuracy of Decision Tree is : {dtc_test_acc}")

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

#confusion matrix
matrix2=confusion_matrix(y_test, y_pred)
plt.figure(figsize = (8,4))
#visualise confusion matrix
sns.heatmap(matrix2 , annot = True, cmap="Greens")
```

```
Training accuracy of Decision Tree is : 0.8013333333333333
Test accuracy of Decision Tree is : 0.848
[[173  23]
 [ 15  39]]
              precision    recall  f1-score   support

           N       0.92      0.88      0.90       196
           Y       0.63      0.72      0.67        54

    accuracy                           0.85       250
   macro avg       0.77      0.80      0.79       250
weighted avg       0.86      0.85      0.85       250
```

```
Out[67]: <AxesSubplot:>
```

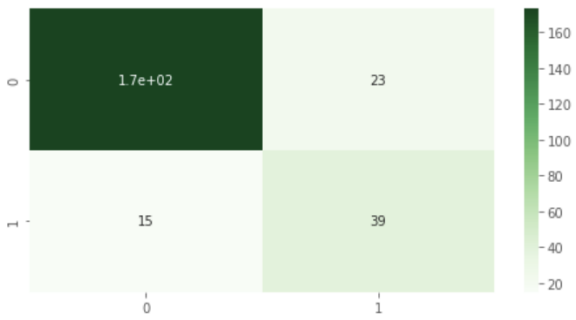


Figure 10: Confusion Matrix of Decision Tree

## Models Comparison

```
In [66]: models = pd.DataFrame({
             'Model' : ['KNN', 'Decision Tree', 'Random Forest','Ada Boost', 'XgBoost', 'Voting Classifier'],
             'Score' : [knn_test_acc, dtc_test_acc, rand_clf_test_acc, ada_test_acc, xgb_test_acc, vc_test_acc
         })


         models.sort_values(by = 'Score', ascending = False)
```

Out[66]:

| | Model | Score |
|---|---|---|
| 1 | Decision Tree | 0.848 |
| 5 | Voting Classifier | 0.848 |
| 0 | KNN | 0.784 |
| 2 | Random Forest | 0.768 |
| 3 | Ada Boost | 0.736 |
| 4 | XgBoost | 0.736 |

```
In [68]: px.bar(data_frame = models, x = 'Score', y = 'Model', color = 'Score', template = 'plotly_dark',
                title = 'Models Comparison')
```

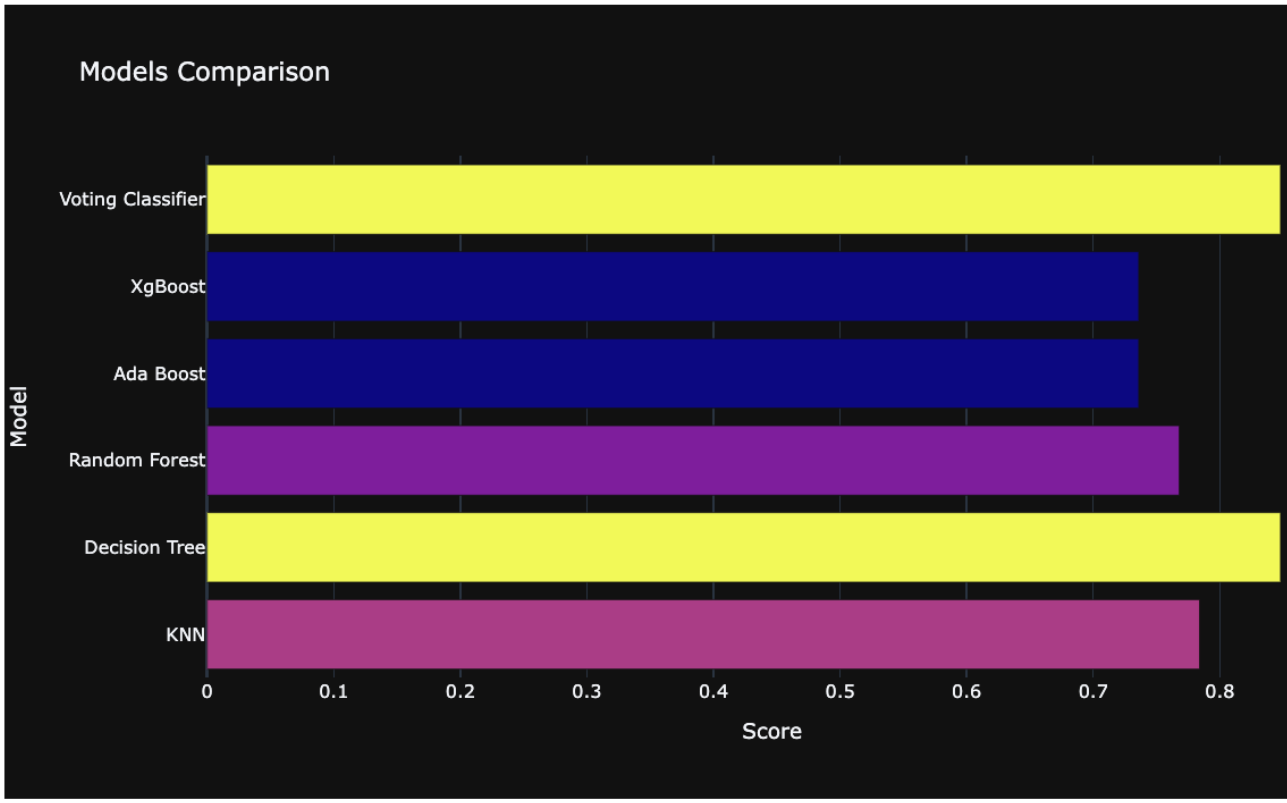


Figure 11: Models Comparison Visualization

The study has several restrictions. The first limitation of this study is the limited sample size. Larger data sets increase the stability of statistical models. Because it includes a larger fraction of the actual population, it also generalises better.

## References


Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (iccni)* (pp. 1–9).

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, *557*, 317–331.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, *41*(10), 4915–4928.

Eshghi, A., & Kargari, M. (2019). Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty. *Expert Systems with Applications*, *121*, 382–392.

French Agency for the Fight against Insurance Fraud, A. (2021). Fight against fraud conference. *ALFA Homepage*, https://www.alfa.asso.fr/.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

Gottschalk, P. (2010). Categories of financial crime. *Journal of financial crime*.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, *9*(2), 42.

Hossin, M. S., Mohammad. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process 5: 1*.

Insurance Fraud Bureau of Australia, I. (2021). Insurance fraud. *Insurance Fraud Bureau of Australia Homepage*, https://ifba.org.au.

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In *2019 third international conference on intelligent computing in data sciences (icds)* (pp. 1–4).

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert systems with applications*, *62*, 32–43.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559–569.

Peng, . N., Y. (2020). *An empirical overview of nonlinearity and overfitting in machine learning using covid-19 data*. Chaos, Solitons and Fractals, Article 110055.

Raghavan, P., & El Gayar, N. (2019). Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (iccike)* (pp. 334–339).

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. In *2017 international conference on circuit, power and computing technologies (iccpct)* (pp. 1–6).

Sheshasaayee, A., & Thomas, S. S. (2018). Usage of r programming in data analytics with implications on insurance fraud detection. In *International conference on intelligent data communication technologies and internet of things* (pp. 416–421).

S.Jordon. (2016). Insurance fraud: 'its all over the place' and you should care about it officials say. *Omaha World-Herald*, https://omaha.com/business/.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit card fraud detection-machine learning methods. In *2019 18th international symposium infoteh-jahorina (infoteh)* (pp. 1–5).

Waghade, S. S., & Karandikar, A. M. (2018). A comprehensive study of healthcare fraud detection based on machine learning. *International Journal of Applied Engineering Research*, *13*(6), 4175–4178.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95.

Wu, S., & Flach, P. (2005). A scored auc metric for classifier evaluation and selection. In *Second workshop on roc analysis in ml, bonn, germany.*

Xu, W., Wang, S., Zhang, D., & Yang, B. (2011). Random rough subspace based neural network ensemble for insurance fraud detection. In *2011 fourth international joint conference on computational sciences and optimization* (pp. 1276–1280).

Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. In *2018 international conference on artificial intelligence and big data (icaibd)* (pp. 57–61).