**Insurance Claim Fraud Detection**

**Introduction**

1. **Background Information**

In today's materialistic world, everyone seeks to protect their possessions in various ways. Recent years have seen a surge in cyberattacks, posing significant challenges to businesses and individuals alike. Notable incidents, such as the Equifax breach in 2017, exposed the personal data of millions, highlighting the critical need for cyber insurance. Organizations now consider cyber insurance essential for safeguarding against the financial and reputational damage resulting from cyber incidents. This need for protection underscores the fundamental principle of the insurance industry: people are willing to pay to mitigate the risk of unforeseen losses.

In the United States alone, the insurance industry is valued at $1.28 trillion, but it suffers from annual losses of at least $80 billion due to fraud. This significant fraud leads insurance companies to increase policy costs, making them less competitive. Consequently, the minimum threshold for policy payments rises as companies adjust prices to counteract fraud-related losses.

**Insurance**

Insurance is a financial arrangement that provides protection against various types of risks, such as property damage, illness, death, or legal liability. By paying a premium, policyholders transfer the financial burden of potential losses to the insurance company, which agrees to compensate them according to the terms of the policy. This arrangement helps individuals and businesses manage uncertainty and protect their financial well-being.

**Types of Insurance**

There are several types of insurance, each designed to cover specific risks:

1. Health Insurance: Covers medical expenses for illnesses, injuries, and other health-related issues. Policies may include coverage for doctor visits, hospital stays, medications, and preventive care.
2. Life Insurance: Provides financial support to beneficiaries in the event of the policyholder’s death. It can help cover funeral expenses, pay off debts, and provide income replacement for the deceased’s dependents.
3. Auto Insurance: Protects against financial losses from car accidents, theft, and other vehicle-related incidents. It typically includes liability coverage, collision coverage, and comprehensive coverage.
4. Homeowners Insurance: Covers damage to a home and its contents due to events like fire, theft, and natural disasters. It may also provide liability coverage if someone is injured on the property.
5. Business Insurance: Offers various types of coverage for businesses, including property insurance, liability insurance, and workers' compensation. It helps protect against financial losses from lawsuits, property damage, and employee injuries.
6. Travel Insurance: Provides coverage for unexpected events that occur during travel, such as trip cancellations, medical emergencies, and lost luggage

**Importance of Insurance**

Insurance is vital as it mitigates financial risk by providing compensation for losses, thereby offering financial stability and peace of mind. It encourages risk-taking by acting as a safety net, supports economic stability by spreading risk, and ensures legal and regulatory compliance for individuals and businesses.

**Understanding Insurance Fraud**

Insurance fraud is an intentional act of deceiving an insurance provider to receive benefits or payouts that are not rightfully owed. This type of fraud can be committed by individuals, groups, or organizations and can occur in various forms across different types of insurance, including health, auto, life, and property insurance. The impact of insurance fraud is substantial, leading to increased premiums for consumers and financial losses for insurance companies.

**Types of Insurance Fraud**

1. Hard Fraud: This involves deliberate planning or staging of a loss, such as setting fire to one's property or staging a car accident, with the intent to file a fraudulent claim.
2. Soft Fraud: Also known as opportunistic fraud, this occurs when a legitimate claim is exaggerated to obtain a higher payout. For example, a person might inflate the value of stolen items in a burglary claim.
3. Claim Padding: Adding non-existent damages to a legitimate claim, such as claiming more extensive repairs than necessary in an auto accident.
4. False Claims: Submitting claims for injuries or damages that never occurred, often using fake documents or staged incidents.
5. Premium Fraud: Misrepresenting information on an insurance application to obtain lower premiums, such as underreporting the number of drivers in a household or the usage of a vehicle.

**2. Problem Statement**

The primary objective of this article is to develop a robust model capable of accurately identifying fraudulent auto insurance claims. Through rigorous testing of multiple algorithms, we aim to create an optimized model that meets the specific needs of insurance companies. This model is designed to be both efficient in processing large datasets and sophisticated enough to achieve a high success rate in fraud detection. Our goal is to present insurance companies with a tailored solution that integrates seamlessly into their existing systems, enhancing their ability to combat fraudulent activities effectively.

**3. Methodology Overview**

For this project, we began with extensive market research to gain a deep understanding of insurance, insurance fraud, and the impact of fraudulent claims on insurance companies. Our focus was specifically on auto insurance claim fraud data. We meticulously analyzed the dataset features to determine how to enhance and modify the data without compromising the goal of identifying suspicious claims. This process involved cleaning the data, removing unnecessary values, and creating new attributes through data integration methods.

Subsequently, we conducted Exploratory Data Analysis (EDA) and Data Visualization to uncover patterns and insights. This was followed by Pre-Processing and Feature Engineering to prepare the data for model training. We then evaluated various machine learning models to identify the best performing one. The data was passed through these models, their performance was assessed, and predictions were made accordingly.

**Tools and Libraries Used:**

1. **Jupyter Notebook:** For writing and running code.
2. **Python Libraries:**
   * **Pandas:** Data manipulation and analysis.
   * **NumPy:** Numerical computing.
   * **Seaborn and Matplotlib:** Data visualization.
   * **Warnings:** Managing warning messages.
   * **Datetime:** Date and time manipulation.
   * **Sklearn (scikit-learn):** Machine learning library, including modules such as preprocessing, metrics, tree, ensemble, svm, neighbors, model\_selection.
   * **Statsmodels:** Statistical models, specifically outliers\_influence.
   * **Imbalanced-learn (imblearn):** Handling imbalanced datasets, particularly over\_sampling.
   * **Joblib:** Serialization and deserialization of Python objects.
   * **Models:**
     + RandomForestClassifier and ExtraTreesClassifier for ensemble learning.
     + LogisticRegression for regression analysis.
     + Various other classifiers for comparison and performance evaluation.

**Project Description**

1. **Data Collection**

The first important step is to collect data. After formulating the business problem, it is important to understand the data sources. The dataset used in this project is dataset link provided by Flip Robot Technology and is available on GitHub.

Link:- [https://raw.githubusercontent.com/FlipRoboTechnologies/ML\_- Datasets/main/Insurance%20Claim%20Fraud%20Detection/Automobile\_insurance\_fraud.csv](https://raw.githubusercontent.com/FlipRoboTechnologies/ML_-%20%20%20%20%20%20%20%20%20%20Datasets/main/Insurance%20Claim%20Fraud%20Detection/Automobile_insurance_fraud.csv)

Python Code used for load data in jupyter Nootbook :-

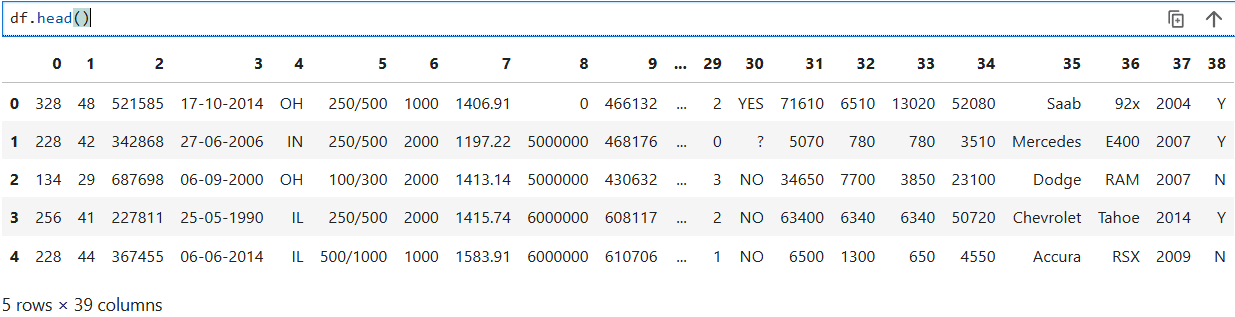
df= pd.read\_csv('https://raw.githubusercontent.com/FlipRoboTechnologies/ML\_- Datasets/main/Insurance%20Claim%20Fraud%20Detection/Automobile\_insurance\_fraud.csv', header=None)

1. **Dataset** **Overview**

The dataset contains 1000 records and 39 attributes, including both numerical and categorical data. The main goal is to classify whether a claim is fraudulent or not. The 1000 records include information about auto insurance policy claims. Some attributes may not be relevant for running the models and algorithms and are used only to identify individuals or entities involved in the claims. These attributes can be modified or integrated as needed for data analysis. Feature included in the following are:

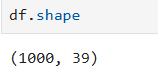
"months\_as\_customer","age","policy\_number","policy\_bind\_date","policy\_state","policy\_csl", "policy\_deductable", "policy\_annual\_premium","umbrella\_limit", "insured\_zip","insured\_sex", "insured\_education\_level", "insured\_occupation", "insured\_hobbies", "insured\_relationship", "capital\_gains", "capital\_loss", "incident\_date", "incident\_type", "collision\_type", "incident\_severity", "authorities\_contacted","incident\_state", "incident\_city","incident\_location", "incident\_hour\_of\_the\_day", "number\_of\_vehicles\_involved", "property\_damage", "bodily\_injuries", "witnesses", "police\_report\_available", "total\_claim\_amount", "injury\_claim", "property\_claim", "vehicle\_claim", "auto\_make", "auto\_model", "auto\_year" , \_c39 and "fraud\_reported"Data

**Showing first five rows**



37 columns represent the independent variables, while one column serves as the target variable. Given that the target variable is categorical, we will employ classification techniques for this project.

**Displaying the shape of the dataset shows how many rows and columns it contains.**



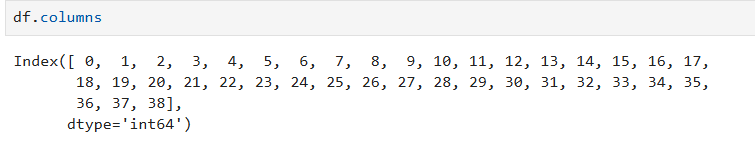
1. **Data Pre-Processing**

Pre-processing is a critical step in machine learning that can significantly enhance the quality of information, thereby facilitating the extraction of meaningful insights. It involves preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models. In essence, data preprocessing is an information mining technique that transforms raw data into a comprehensible and structured format. Once raw data is collected, it must be organized to be ready for further processing and analysis.

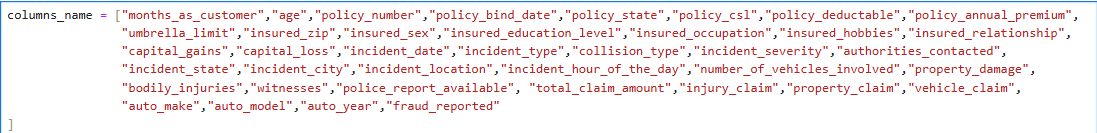
1. **Data Cleaning and Exploratory Data Analysis**

Data cleaning is an essential step in preparing data for analysis. It ensures that the dataset is free from errors that could affect the results. Common issues like duplicate records, missing values, or inconsistent data need to be fixed. In this step, we remove bad data and fill in any missing information. The aim is to tidy up the dataset by getting rid of unnecessary information and adding any missing pieces, ensuring that the data is accurate and ready for analysis.

**First , we check column names and find that a column name is missing.**



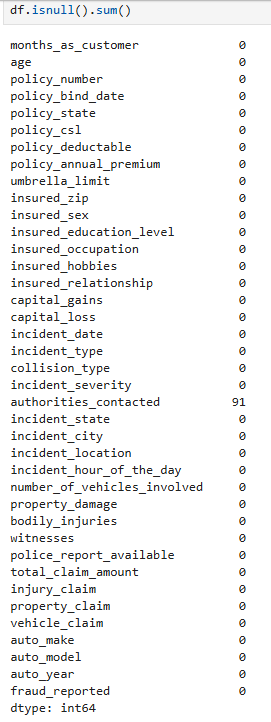
**We then create a list to define new column names.**



**Replace the old ones with the new names**

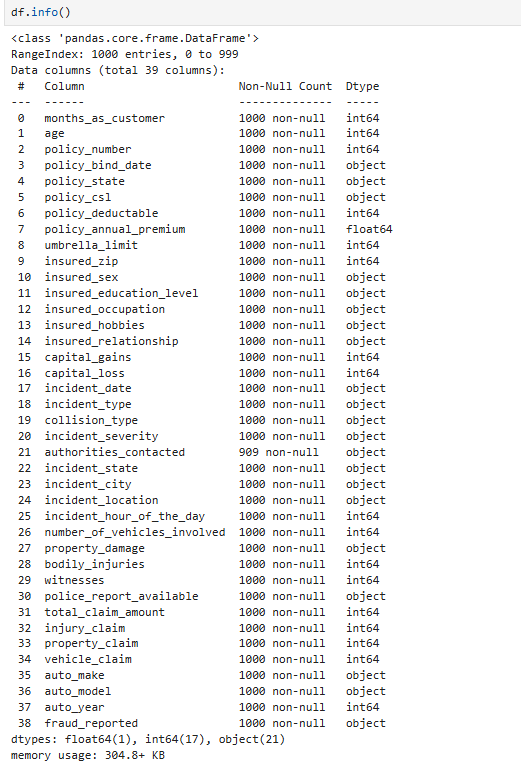


**After that We check Null values**



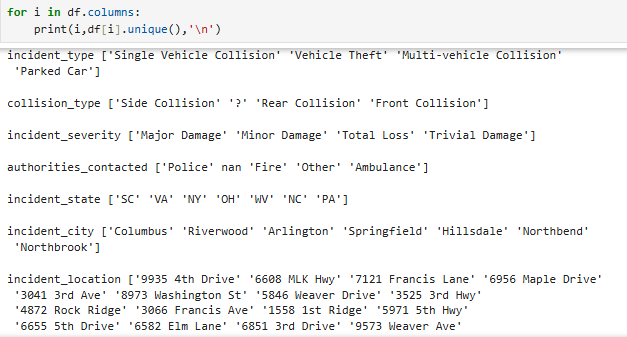
There are 91 missing values in the data set with column name ‘authorities\_contached’, so handling null values is a very important step in the data cleaning process. Because it helps to deal with the problems that appear during the subsequent procedures. There are several ways to handle null values and missing values. We can remove all records from the data set or impute missing values using mean, median, or regression methods.

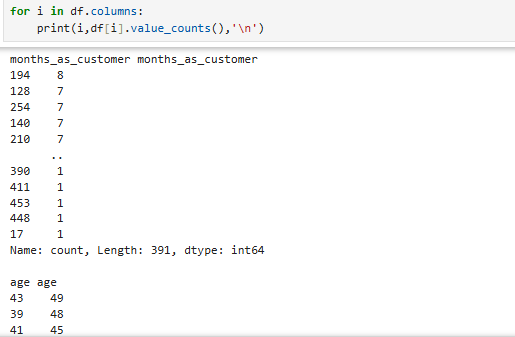
**Before filling missing value we will check the information about the data .**



The info function show that, In the Insurance claim dataset, there are 39 features, with 18 being numerical and the remaining being categorical. The dataset includes three types of data: 'float', 'int', and 'object'. The date columns are not in date format.

**We will check unique values and count of values**





The unique function shows the unique values in each column and observe that the 'collision type', property\_damage’ and 'police report available' columns contain the special character '?'.

**We'll replace these special characters with Nan values and then fill those Nan values with the mode of the data**.



**Now we will be handling nan values**

1. Filling Nan values with mode



1. Filling nan values with ‘Others’ character



The 'authorities\_contacted' column contains unique values such as 'Police' ,'Fire', 'Other ‘and 'Ambulance', with some null values. To handle missing data, Nan values in this column we will replace Nan values with 'other'



1. **Data Visualization**

Visualization is crucial in data analysis because it simplifies complex data, enhances understanding, and facilitates easy comparisons. It improves communication by making data insights accessible to all stakeholders, supporting better decision-making. Additionally, visualization can reveal hidden patterns and engage the audience effectively.

1. **Univariate Analysis:** This involves examining each variable individually to understand its distribution, central tendency, and variability.
2. **Bivariate Analysis:** This examines the relationship between two variables.
3. **Multivariate Analysis:** This involves exploring the relationships between three or more variables simultaneously.

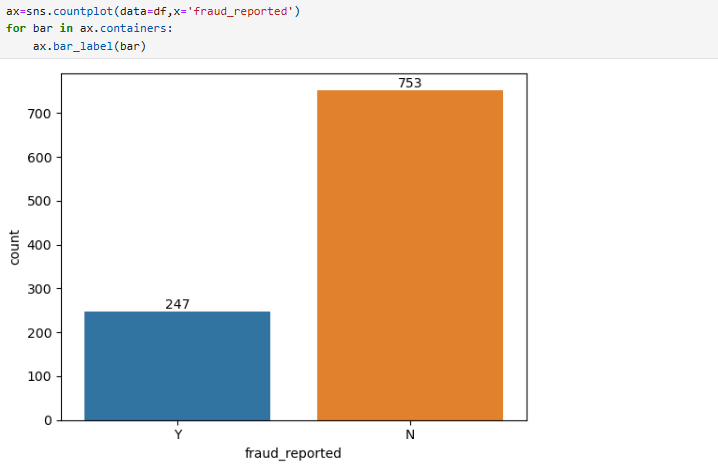


Figure 1. Plot of fraud\_report against count

In Figure 1, the bar plot generated using seaborn's countplot() function illustrates the distribution of fraud reports categorized by the 'fraud\_reported' column, which serves as our target variable with values 'YES' and 'NO'. The plot highlights a significant class imbalance, where 'YES' (indicating fraud reports) substantially outnumber 'NO', emphasizing an imbalance issue within the dataset.

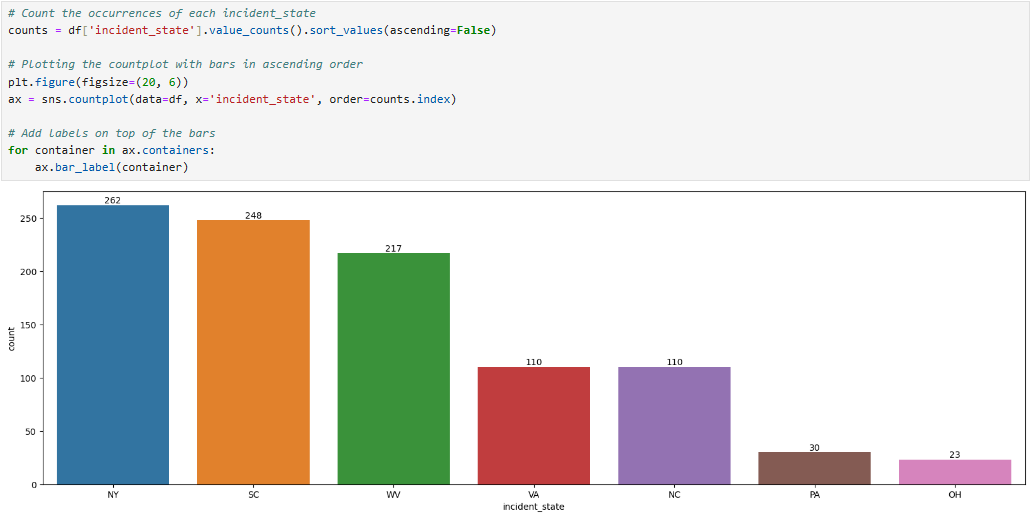


Figure 2. Plot of incident state against count Figure 2., In the analysis, the countplot visually represents the distribution of fraud reports across incident states. Notably, New York (NY) and South Carolina (SC) emerge with the highest incidence rates, while West Virginia (WV), Virginia (VA), and North Carolina (NC) also exhibit significant numbers of incidents. Conversely, Ohio (OH) and Pennsylvania (PA) show the lowest reported incidents.

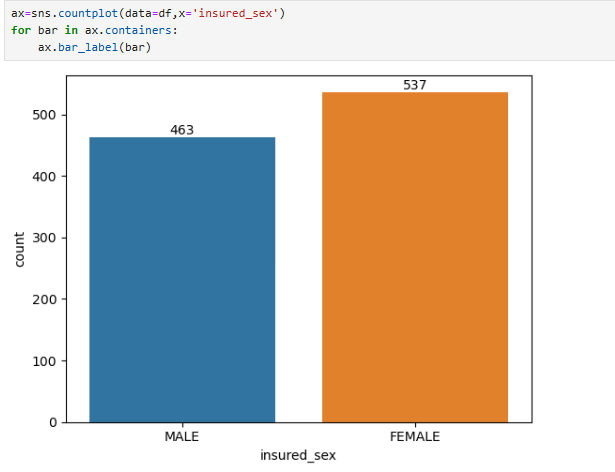


Figure 3. Plot of insured sex against count

In Figure 3, the countplot using seaborn's function displays the distribution of insurance claims by gender, with the 'insured\_sex' column on the x-axis. The data reveals that females have a higher frequency of insurance claims compared to males , suggesting that females may be more proactive in filing insurance claims than males. This observation could reflect broader trends in insurance claim behavior influenced by factors such as risk perception or coverage needs across different demographics.

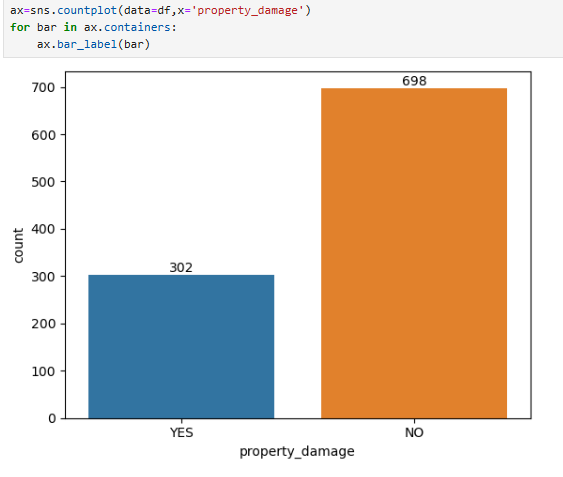


Figure 4. Plot of property damage against count

In Figure 4, the countplot using seaborn's function displays the distribution of insurance claims of property damage , with the 'property\_damage' column on the x-axis. A visual representation that property damage is less frequently reported compared to incidents with no property damage.

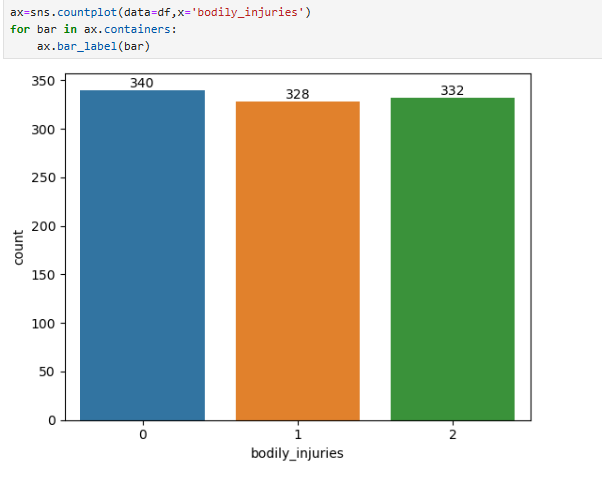


Figure 5. Plot of bodily injuries against count

In Figure 5, the countplot using seaborn's function illustrates the distribution of insurance claims categorized by bodily injuries, with the 'bodily\_injuries' column plotted on the x-axis. The plot visually depicts a balanced distribution across the three categories (0, 1, 2). Specifically, there are slightly more claims with no bodily injuries compared to those with one or two bodily injuries, indicating a relatively even distribution across these injury categories in the dataset.

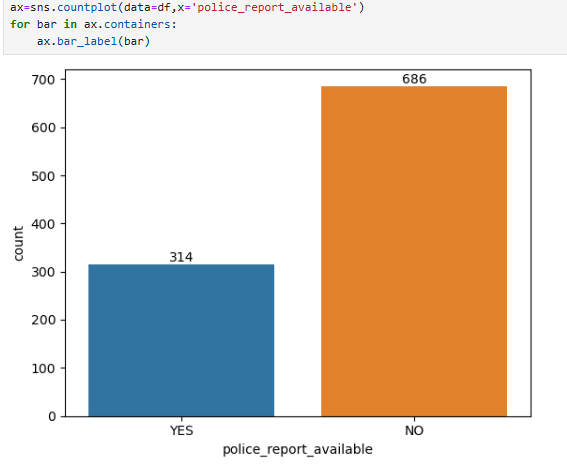


Figure 6. Plot of police report available against count

In Figure 6, the countplot using seaborn's function depicts the distribution of insurance claims categorized by the availability of police reports, with the 'police\_report\_available' column plotted on the x-axis. The plot visually highlights that a significant portion of insurance claim fraud cases lack accompanying police reports. This absence can impede thorough investigation and verification processes essential for detecting and adjudicating fraudulent claims effectively.

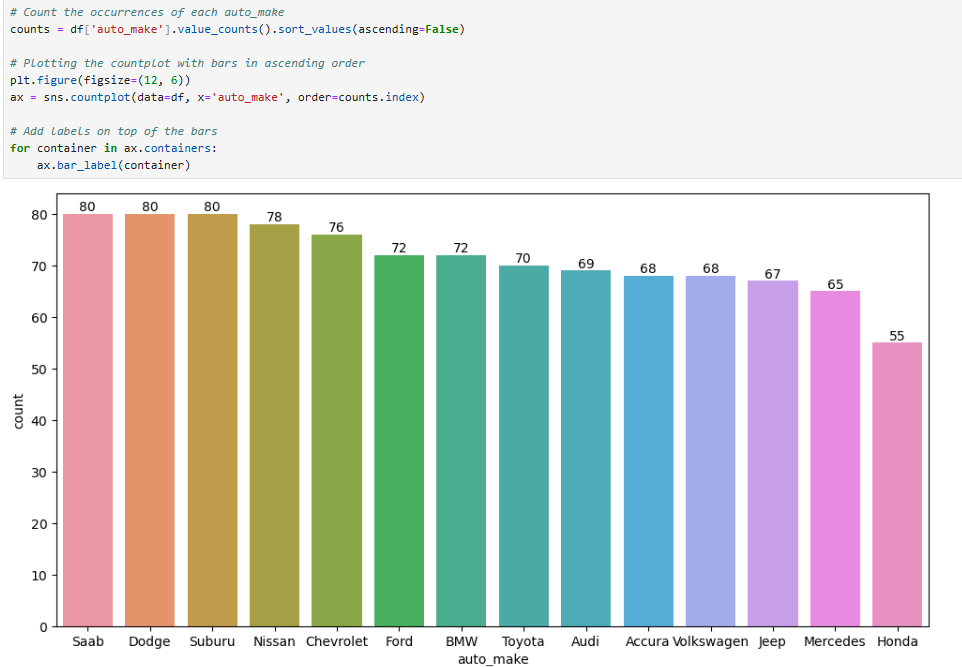
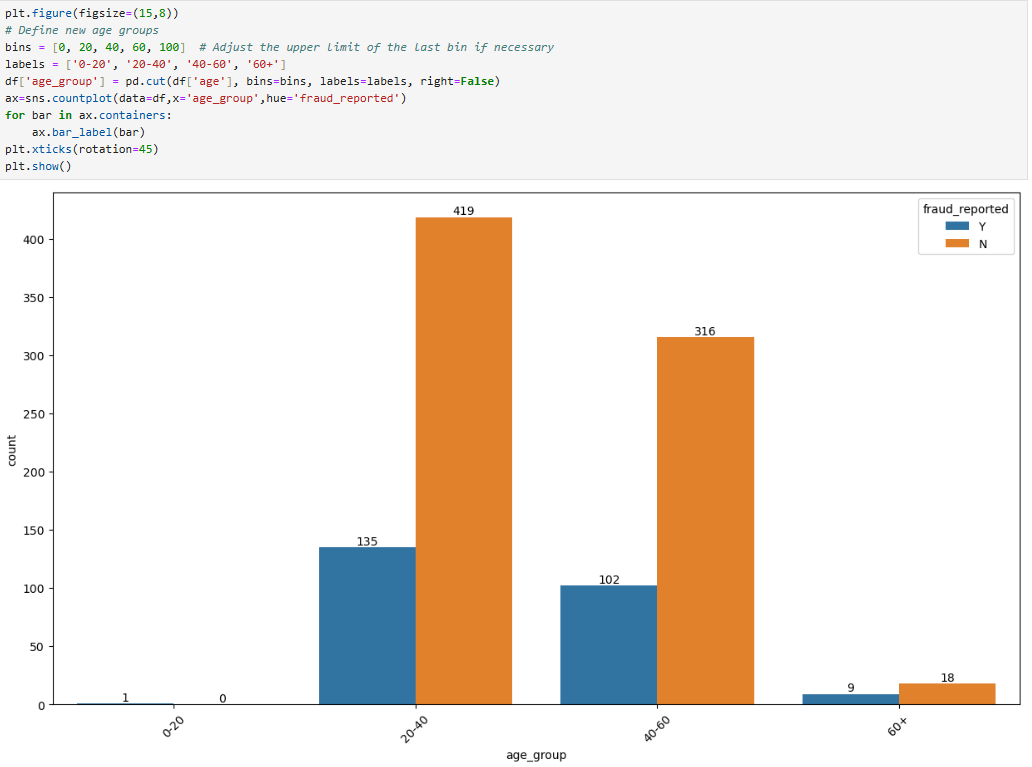


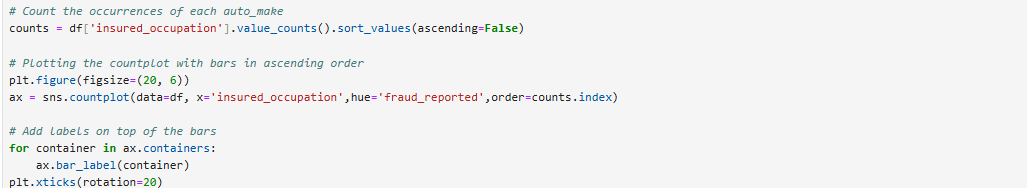
Figure 7.Plot of Auto make against count

In Figure 7, the countplot using seaborn's function displays the distribution of insurance claims by auto make, with notable findings:

* Saab, Dodge, and Subaru exhibit the highest fraud report counts (80 each), suggesting frequent involvement in fraud cases.
* Nissan, Chevrolet, and Ford also show substantial counts, indicating moderate involvement in fraud cases.
* BMW, Toyota, and Audi have notable fraud report counts.
* Honda has the lowest count (55), indicating it is the least frequently involved in fraud reports among the analysed makes.

 Figure 8. Plot of age group against count

In Figure 8, the countplot using seaborn's function categorizes insurance claims by age group ('Age group' on x-axis). It shows that the 20-40 and 40-60 age groups have significant instances of reported fraud alongside high volumes of non-fraudulent claims. In contrast, the 0-20 and 60+ age groups exhibit fewer fraudulent claims, with the latter slightly higher in non-fraudulent claims.



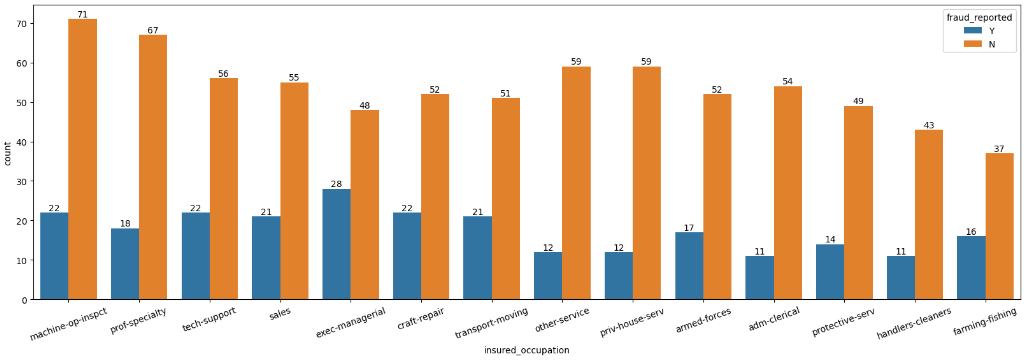


Figure 9. Plot of insured occupation against count

In Figure 9, the countplot using seaborn's function shows the distribution of insurance claims by insured occupation . The data reveals that Professional Specialty and Machine Operators/Inspectors have the highest fraud incidence suggesting these occupations are frequently involved in fraud reports. Moderate fraud incidence is observed in occupations like Tech Support, Sales, Executive Managerial, and Craft Repair. Conversely, Protective Service, Handlers/Cleaners, and Farming/Fishing have the lowest counts, indicating these occupations are less frequently associated with fraud reports.

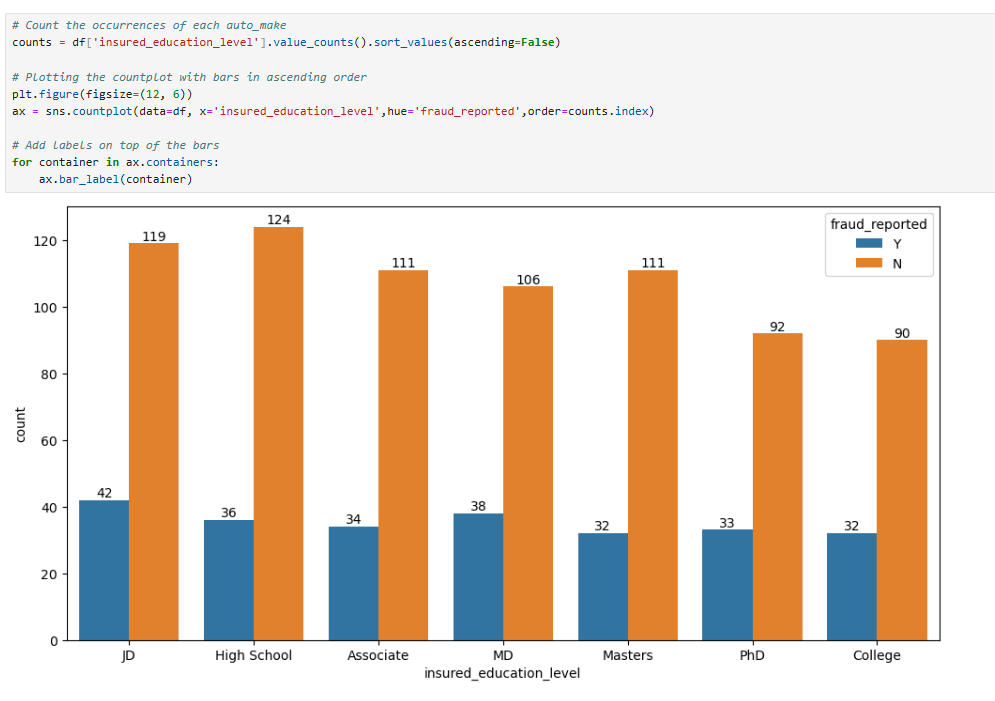
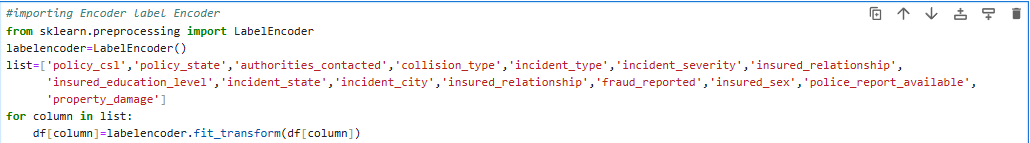


Figure 10. Plot insured education level against count In Figure 10, the countplot using seaborn's function displays the distribution of insurance claims categorized by insured education. The visual representation shows that across all education levels, non-fraudulent claims ('N') significantly outnumber fraudulent claims ('Y'). The highest incidence of reported fraud is among individuals with a JD degree, followed by those with an MD. High school graduates have the highest number of non-fraudulent claims, closely followed by those with an Associate degree and a Master's degree.

1. **Data Transformation**

In order to build a predictive model by using machine learning it is important to have all the input and output variables in numeric format not in categorical format as we know machines only understand numeric values, so we have to convert all the categorical variables into numeric to fit and access the model.



From sklearn.preprocessing label encoder is used to apply label encoding to categorical data and converted into numerical form.

1. **Correlation**

When predicting the output using supervised machine learning techniques, managing the dataset size is crucial. With small datasets, we can easily identify predictors or variables. However, as data volume increases, using all the data can negatively impact model accuracy and computational efficiency. To address this, we explore the concept of correlation to understand the relationships between dependent and independent features, allowing us to select the most important features for prediction. This selective approach enhances model performance and optimizes resource usage.





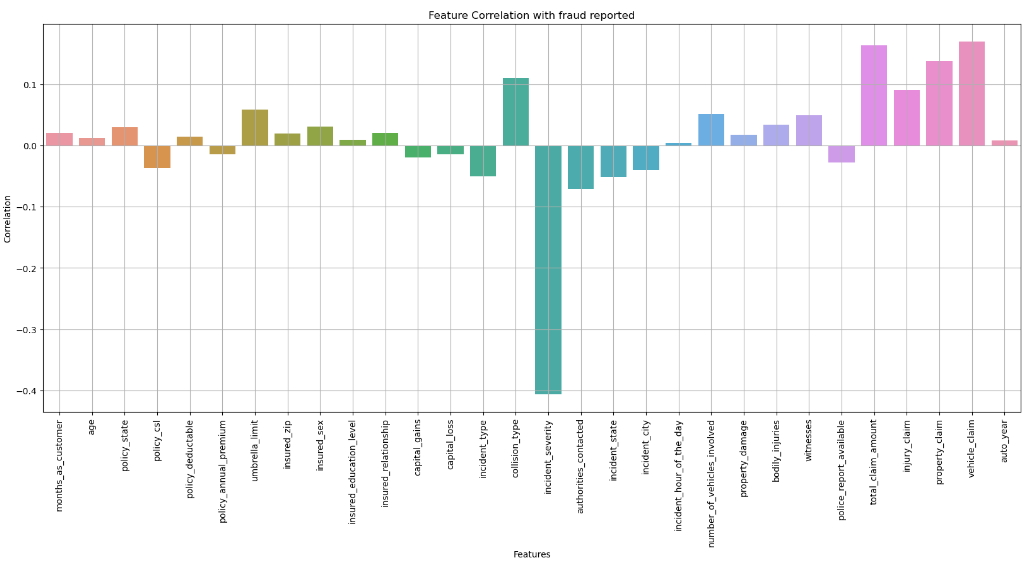


Figure 11. Feature Correlation with fraud report

Figure 11. show , The correlation plot illustrates the relationships between various features and fraud reports. Notably, `incident\_type` and `collision\_type` show the strongest negative correlations with fraud reports, while features like `vehicle\_claim`, `total\_claim\_amount`, and `injury\_claim` have positive correlations. These insights help identify key predictors for building an accurate fraud detection model.

1. **Build & training** **model**

Building and training a model involves selecting an appropriate machine learning algorithm and preparing the dataset by splitting it into training and testing sets. The model is trained on the training set to learn patterns and relationships within the data. After training, the model's performance is evaluated on the test set to assess its accuracy and generalizability. This process helps ensure that the model can make reliable predictions on new, unseen data

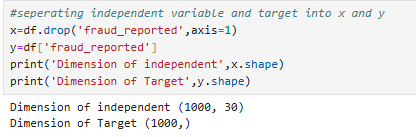


Figure 12. Separating independent and target variable

1. **Feature Scaling**

Feature scaling, using tools like StandardScaler, standardizes data by removing the mean and scaling to unit variance. This step is essential for many machine learning algorithms to ensure that all features contribute equally and improve model performance and convergence.

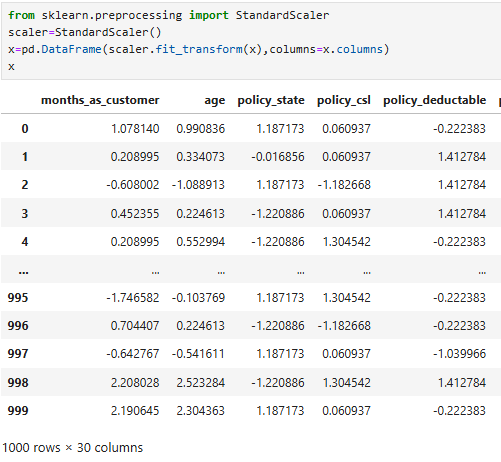


Figure 12. scaling by using standard scaler

Figure 12., We have scaled the data using the standard scaler method to ensure that all features contribute equally to the analysis, thus overcoming the issue of bias caused by varying scales of the features .

1. **Variance Inflation Factor**

Variance Inflation Factor (VIF) measures the degree of multicollinearity in a set of predictor variables within a regression model. High VIF values indicate that a predictor is highly collinear with other predictors, which can negatively impact model stability and interpretability. Reducing multicollinearity by removing or combining variables with high VIF improves model performance.

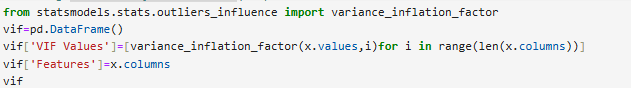


Figure 13. VIF

Figure 13., From statsmodels.stats.outliers\_influence using variance\_inflation\_factor for checking VIF score for multicollinearity in your data .

1. **Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining as much variance as possible. It identifies the directions (principal components) that capture the maximum variance in the data, allowing for simplified visualization and analysis of complex datasets. This method is widely used in exploratory data analysis and feature extraction across various fields, including data science, image processing, and signal processing.



Figure 14. PCA

Figure 14., From sklearn.decomposition using PCA method to manage bias in the data by transforming the original features into a set of orthogonal components, which helps in reducing dimensionality and capturing the most important variance in the data.

1. **Oversampling**

Oversampling is a technique used in machine learning to handle class imbalance by increasing the number of instances in the minority class. This helps in improving model performance by ensuring that the training dataset has a balanced distribution of classes, thereby reducing bias towards the majority class and enhancing the model's ability to correctly predict the minority class. Common methods include Random Oversampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling).

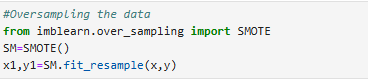
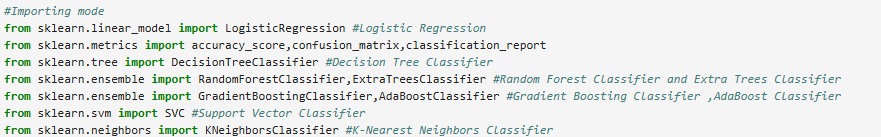


Figure 15. Oversampling

Figure 15., We have used SMOTE methods from imblearn.over\_sampling for overcome the class imbalance problem

1. **Model Selection**

Model selection refers to the process of choosing the best machine learning algorithm or statistical model that fits a given dataset and problem. It involves evaluating multiple models based on metrics like accuracy, precision, recall, or cross-validation scores to determine which model generalizes best to unseen data. Techniques such as grid search, cross-validation, and hyperparameter tuning are commonly used to optimize model performance and select the most suitable one for a specific task**.**

****

1. Logistic Regression:

Logistic Regression is a linear model used for binary classification tasks, predicting the probability that an instance belongs to a particular class based on input features.

2. Decision Tree Classifier:

Decision Tree Classifier partitions data into subsets based on features, creating a tree-like model of decisions. It handles both numerical and categorical data, making it suitable for various classification tasks.

1. Random Forest Classifier and Extra Trees Classifier:

These ensemble methods combine multiple decision trees trained on different data subsets. Random Forest averages predictions to improve generalization, while Extra Trees selects splits randomly, both mitigating overfittings.

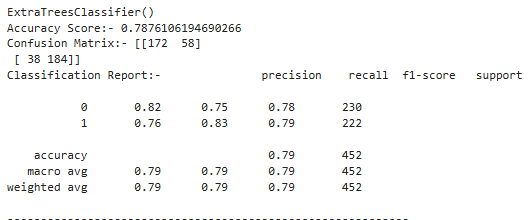
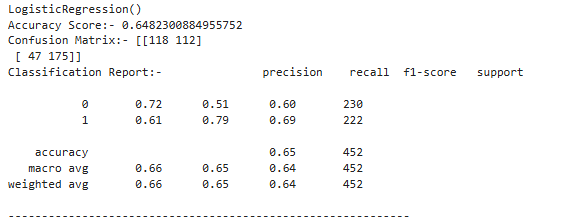
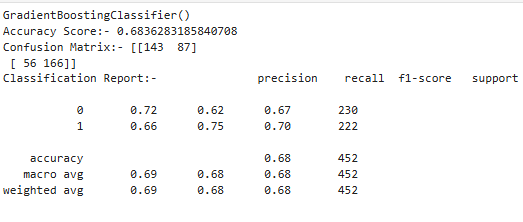
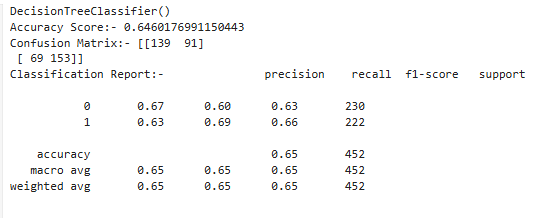
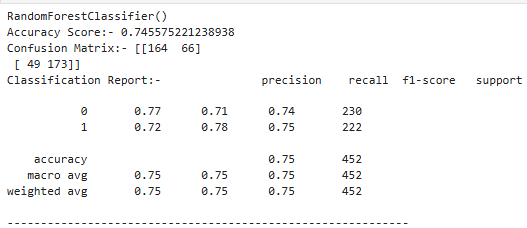
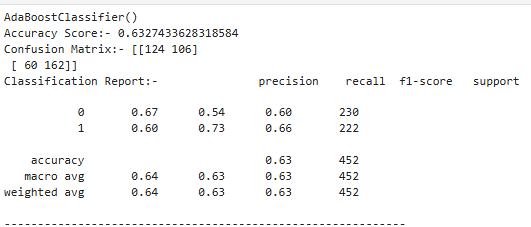
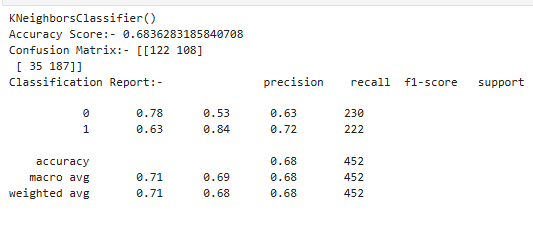
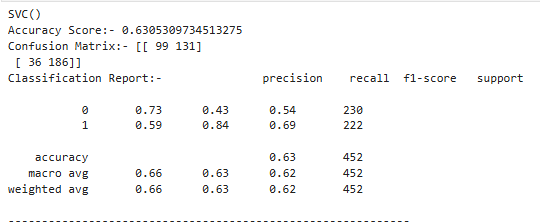
1. Gradient Boosting Classifier and AdaBoost Classifier:

Gradient Boosting and AdaBoost build models sequentially, correcting errors from previous models to enhance accuracy, particularly effective for challenging instances.

5. Support Vector Classifier (SVC):

SVC finds an optimal hyperplane to separate classes in high-dimensional spaces. It excels in complex tasks and performs well with smaller to medium-sized datasets.

6. K-Nearest Neighbors Classifier (KNN):

 KNN predicts class based on majority voting from its nearest neighbors, suitable for both binary and multi-class classification tasks with irregular decision boundaries.

1. **Cross Validation Score**

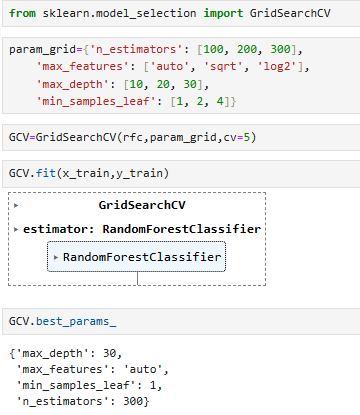
Cross-validation score is a technique used to evaluate the performance of a machine learning model by partitioning the dataset into subsets. The model is trained on some subsets and validated on the remaining ones, repeated multiple times to ensure robustness. In this project, cross-validation was applied using accuracy\_score as the evaluation metric to assess the reliability and generalizability of the models. The result is below :-

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine learning model** | **Accuracy Score** | Cross Validation Score mean value | Difference of cross validation |
| 1. LogisticRegression | 64.82% | 0.75 | -10.17 |
| 1. ExtraTreesClassifier | 78.76% | 0.70 | 8.06 |
| 1. DecisionTreeClassifier | 64.60% | 0.63 | 1.10 |
| 1. GradientBoostingClassifier | 68.36% | 0.71 | -3.53 |
| 1. AdaBoostClassifier | 63.27% | 0.73 | -10.32 |
| 1. RandomForestClassifier | 74.55% | 0.71 | 2.63 |
| 1. KNeighborsClassifier | 68.36% | 0.71 | -3.33 |
| 1. SVC | 63.05% | 0.75 | -12.24 |

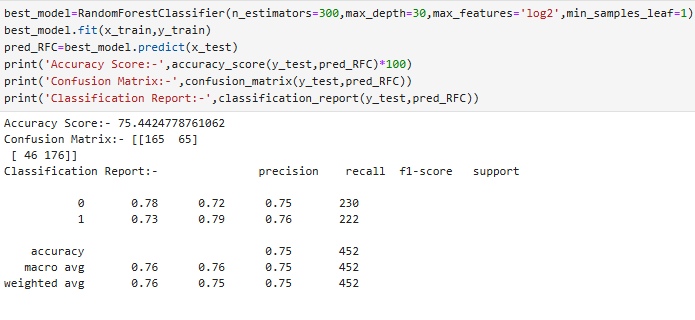
The RandomForestClassifier seems to perform relatively well among the models tested, as it has a high mean cross-validation score and a small difference between accuracy score and cross-validation score. RandomForestClassifier is our best fitting and best performing model.

1. **Hyper Parameter Tuning**

GridSearchCV is a method from scikit-learn used to improve model performance by exhaustively searching for the best hyperparameters over a specified parameter grid. This approach helps in fine-tuning the model by finding the optimal combination of hyperparameters.



GridSearchCV evaluates different combinations of n\_estimators, max\_features, max\_depth, and min\_samples\_leaf to find the optimal set of hyperparameters that maximizes the accuracy of the RandomForestClassifier.



1. **AUC ROC curve and Precision-Recall Curve**

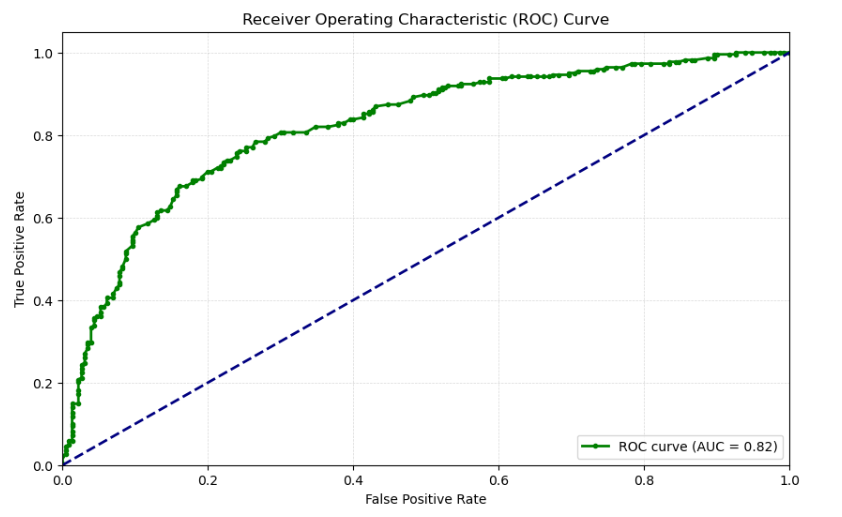


Figure 16. ROC Curve

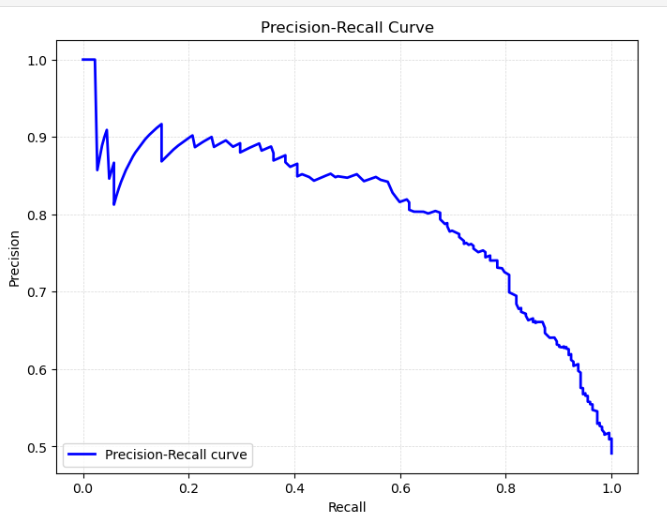


Figure 17. Precision-Recall Curve

**Conclusion**

As countries worldwide evolve towards more economically driven systems, stimulating their economies becomes a central goal. Combating fraudsters and money launderers was a complex task before the advent of machine learning. However, thanks to advancements in machine learning and AI, we are now better equipped to tackle such issues. The proposed solution can be applied by insurance companies to determine whether a particular insurance claim is fraudulent. This model was developed after testing multiple algorithms to identify the best one for detecting fraudulent claims. It is intended as a pitch to insurance companies to encourage the development of a tailored model suited to their specific systems. The model should be straightforward enough to handle large datasets, yet sophisticated enough to achieve a high success rate.