**DETECTING CLAIMS FRAUD IN INSURANCE USING MACHINE LEARNING TECHNIQUES**

A report submitted in partial fulfillment of the requirements for the award of

The B. Tech Degree

By

Sai Spandana Echambadi

(20EG105142)

Kodidala Koushik Kumar

(20EG105125)

Shaik Mohammed Kaif

(20EG105145)

Marri Nithin Reddy

(20EG105153)

**Under the Guidance of**

Mr. D Ramana Kumar

 Assistant Professor

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**ANURAG UNIVERSITY**

**VENKATAPUR – 500088**

**TELANGANA**

**YEAR 2023-24**

**DECLARATION**

I Here declare that the Report entitled “DETECTION OF CLAIMS FRAUDS IN INSURANCE USING MACHINE LEARNING TECHNIQUES” submitted for the award of Bachelor of technology Degree is my original work and the Report has not formed the basis for the award of any degree, diploma, associate ship or fellowship of similar other titles. It has not been submitted to any other University or Institution for the award of any degree or diploma.

Place: Anurag University, Hyderabad

K.Koushik Kumar

(20EG105125)

Sai Spandana Echambadi

(20EG105142)

Shaik Md. Kaif

(20EG105145)

Marri Nithin Reddy

(20EG105153)

**CERTIFICATE**

This is to certify that the Report / dissertation entitled “DETECTION OF CLAIMS FRAUDS IN INSURANCE USING MACHINE LEARNING TECHNIQUES” that is being submitted by Kodidala Koushik Kumar (20EG05125), Sai Spandana Echambadi (20EG105142), Shaik Mohammed Kaif (20EG105145), Marri Nithin Reddy (20EG105153) in partial fulfilment for the award of B. Tech in Computer Science and Engineering to the Anurag University is a record of bonafide work carried out by him under our guidance and supervision.

The results embodied in this Report have not been submitted to any other university or Institute for the award of any degree or diploma

.

**head of the department**

**signature of supervisor**

Mr. D. Ramana Kumar

**(ASSISTANT PROFFESOR)**

**ACKNOWLEDGMENT**

We would like to express our sincere thanks and deep sense of gratitude to project supervisor **Mr. D Ramana Kumar** (Assistant Professor) for his constant encouragement and inspiring guidance without which this project could not have been completed. His critical reviews and constructive comments improved our grasp of the subject and steered to the fruitful completion of the work. His/her patience, guidance and encouragement made this project possible.

We would like express my special thanks to **Dr. V. Vijaya Kumar**, Dean School of Engineering, Anurag University, for their encouragement and timely support in

our B. Tech program.

We would like acknowledge our sincere gratitude for the support extended by **Dr. G. Vishnu Murthy**, Dean, Dept. of CSE, Anurag University. We also express my deep sense of gratitude to **Dr**. **V V S S S Balaram,** Academic Coordinator, **Dr. Pallam Ravi**, Project In-Charge, **Mr. Prabhakar Raju** Project Coordinator and Project Review committee members, whose research expertise and commitment to the highest standards continuously motivated me during the crucial stage our project work.

**ABSTRACT**

The insurance industries consist of more than thousand companies worldwide and they collect premium worthy of more than one trillion dollars every year. When a person or entity makes false insurance claims in order to obtain compensation or benefits to which they are not entitled is known as an insurance fraud.

The total cost of an insurance fraud is estimated to be more than forty billion dollars. So, detection of an insurance fraud is a challenging problem for the insurance sector. The traditional approach for fraud detection is based on developing heuristics around the fraud indicator. The auto\vehicle insurance fraud is the most prominent type of insurance fraud, which can be propagated by false accident claim. In this paper, the main focus shall be on detecting the auto\vehicle fraud by using, machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall.

To address this problem, first extensive research should be made to check out what has been applied and what the most promising solution using machine learning and data analytics is out there. After learning, then applying and building upon the findings of the research we propose a model that can flag these suspicious fraudulent claims for the insurance companies to help them out in saving money and time and helping them become more efficient in reacting to these fraudulent claims.

Keywords –Machine learning (ML), Logistic Regression (LG), Random Forest (RF).

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **FIGURE NAME** | **PAGE NO.** |
| Fig 1 | Insurance Fraud Stats | 1 |
| Fig 2 | Claims Fraud Process | 3 |
| Fig 3 | Cases of Insurance fraud in Real-Life | 5 |
| Fig 4 | Random Forest | 6 |
| Fig 5 | Logistic Regression | 7 |
| Fig 6 | Data Visualization-Types of Charts | 9 |
| Fig 7 | Importing Libraries | 10 |
| Fig 8 | GUI Code | 10 |
| Fig 9 | Created GUI | 10 |
| Fig 10 | Data Modelling: 21 attributes | 11 |
| Fig 11 | Data Modelling :Logistic Regression | 11 |
| Fig 12 | Data Modelling :Random Forest | 12 |
| Fig 13 | Implementation using the dataset. | 12 |
| Fig 14 | Output: Not A Fraud | 14 |
| Fig 15 | Output: It’s A Fraud | 14 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE NO.** | **TABLE NAME** | **PAGE NO.** |
| 1 | Literature Review | 4 |
| 2 | Accuracy Score Table | 13 |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO.** | **CONTENT** | **PAGE NO.** |
| 1 | Introduction   * 1. Project Introduction   2. Scope   3. Project Overview   4. Objectives | 1 |
| 2 | Literature Review  2.1 Background  2.2 Existing System | 4 |
| 3 | Proposed Methodology | 5 |
| 4 | Implementation and Results  4.1 Language Used  4.2 Steps Involved in Implementation  4.2.1 Constructing Dataset  4.2.2 Adding Libraries  4.2.3 Data Preprocessing  4.2.4 Model Creation  4.2.5 Data Visualization  4.3 Sample Code | 6 |
| 5 | Testing: Test Cases and Results | 13 |
| 6 | Summary  6.1 Conclusion  6.2 Recommendation  6.2 Future Scope | 15 |
| 7 | References | 16 |

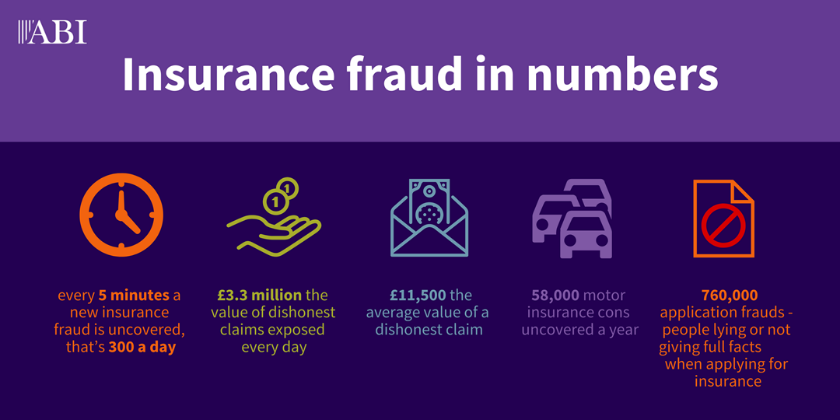
**INTRODUCTION**

**1.1 Introduction**

**Insurance Fraud:**

Fraud occurs when someone knowingly lies to obtain a benefit or advantage to which they are not otherwise entitled, or someone knowingly denies a benefit that is due and to which someone is entitled. According to the law, the crime of insurance fraud can be prosecuted when:

1. The suspect had the intent to defraud. Insurance fraud is a "specific" intent crime. This means a prosecutor must prove that the person involved knowingly committed an act to defraud.
2. An act is completed. Simply making a misrepresentation (written or oral) to an insurer with knowledge that is untrue is sufficient.
3. The act and intent must come together. One without the other is not a crime.
4. Actual monetary loss is not necessary as long as the suspect has committed an act and had the intent to commit the crime.



*Figure 1*

**Types Of Fraud:**

1. Fake policies.
2. Fraud by intermediaries.
3. Premium diversion.
4. Identity theft.
5. Claims Fraud
6. Interest-free loan promise.
7. Premium refund for a lapsed policy.
8. Job assurance.
9. Misleading investment returns.

**Why Claims Fraud Detection:**

The consequences of not detecting insurance fraud are far-reaching. First, the fraud costs are passed on to consumers, who pay higher premiums. Insurance companies lose money and may go out of business if they cannot recoup their losses from fraudulent claims.

Failing to detect fraud is a crime under state law, which means that if they knowingly allow someone who has committed fraud on their policy or allow them access to the claim process, then they could be held liable for any damages caused by this person's actions or inaction (i.e., filing false claims).The consequences can include:

1. Non-payment of claims
2. The insurer seeking costs incurred
3. Cancellation of the insurance policy
4. Reporting the case to the police for further investigation
5. As a result, they could not obtain insurance and other financial services.
6. Prosecution and a custodial sentence

**1.2 Project Scope**

Due to the real-time processing of vast volumes of data using machine learning algorithms, fraudulent claims can be identified and flagged considerably more quickly than conventional techniques. Machine learning algorithms can examine data from many different sources and spot trends that can point to fraud. This results in fewer false positives and more accurate fraud detection.

Insurance companies may save a lot of money if fraudulent claims are caught early. Insurance firms may identify and stop fraudulent claims before they are paid out by utilizing machine learning algorithms, which can result in considerable cost savings. The whole customer experience may be enhanced by insurance firms by identifying and avoiding false claims. Fraud is less likely to cause valid claims to be delayed or refused, which can increase customer satisfaction. The demands of the insurance firm may be met by scaling up or down the machine learning algorithms. Machine learning algorithms can handle the increased burden as data volume increases without the need for extra resources.

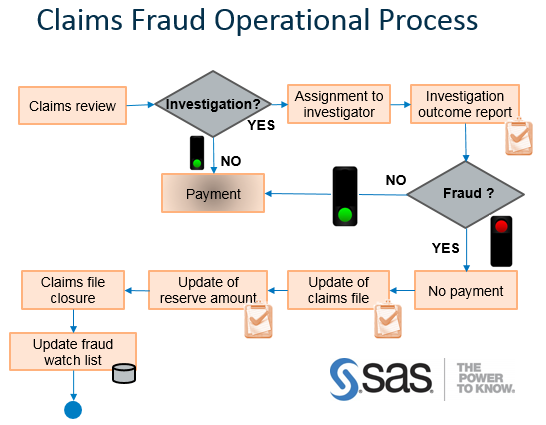
**1.3 Project Overview**

The goal of this project is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims. This type of problems is known as imbalanced class classification.Several models were tested with different methods of handling imbalance datasets. The top models were also fitted and tested with different ensembles.

Prior to modeling, the data was clean and exploratory data analysis was conducted. After which, the data was pre-processed for the modeling. After modeling, the models were evaluated, and the best fitted model was selected using the accuracy and precision value. The performance of the final fitted model was discussed in further details and its top features were displayed. The project concluded by reiterated the importance of the research and what had been done and finally, with some limitations.

**1.4 Objectives**

The main objective of the project is to detect fraud in real-time. An insurance fraud detection solution helps insurance companies identity fraud in the insurance process (application to claims) and consequently plan actions to mitigate them. They may utilize sophisticated analytic tools such as network analysis or text analysis. In the long run, an insurance fraud detection system's purpose is to improve the overall claims ratio of an insurance company by reducing refund in case of fraudulent claims



*Figure 2*

**LITERATURE REVIEW**

**2.1 Background**

* [**Jörn Debener**](https://onlinelibrary.wiley.com/authored-by/Debener/J%C3%B6rn)**,**[**Volker Heinke**](https://onlinelibrary.wiley.com/authored-by/Heinke/Volker)**,**[**Johannes Kriebel**](https://onlinelibrary.wiley.com/authored-by/Kriebel/Johannes) **:**

This study has considered the application of isolation forests, a convenient but effective unsupervised learning algorithm that can be used for fraud detection, and XGBoost, a fast and effective current supervised machine learning algorithm. We further compared these approaches to neural-network-based and clustering-based fraud detection algorithms. We have considered how the supervised and unsupervised learning methods perform in terms of identifying insurance fraud in a large proprietary data set and in terms of identifying insurance fraud in a field experiment. Our results generally emphasize the usefulness of unsupervised learning for insurance companies (particularly when no labeled data are available).

|  |  |  |  |
| --- | --- | --- | --- |
| **Author(s)** | **Method** | **Advantages** | **Disadvantages** |
| Belhadji, E., G. Dionne, and F. Tarkhani | Decision Trees | Model effective in estimating the probability of fraud. | The accuracy rate for large data sets was low. |
| Kajia muller | Comprehensive analysis using Logistic Regression | A difference between soft frauds and hard frauds could be observed. | Wrong inputs or misinterpretation leads in low accuracy |
| Tennyson, S. & Salsas-Forn, P, | Auditing Approach | Reveals the area where maximum fraud takes place | High Risk of using unauthenticated data. |
| Pérez, J. M, Muguerz J, Arbelaitz, O., Gurrutxaga, I., & Martín, J. I., 2005 | Consolidation Trees classification trees | Identification of different types of risks and errors takes place | Decoding of every input is not possible |

**2.2 Existing System**

There are various AI methods with which frauds can be detected.

1) To classify, combine cluster data, and segment, using data mining that can find rules in data and be able to highlight specific patterns, including those related to fraud.

2) Professional programs to detect fraud in the form of laws.

3)To automatize determining factors of false claims, ML techniques are employed

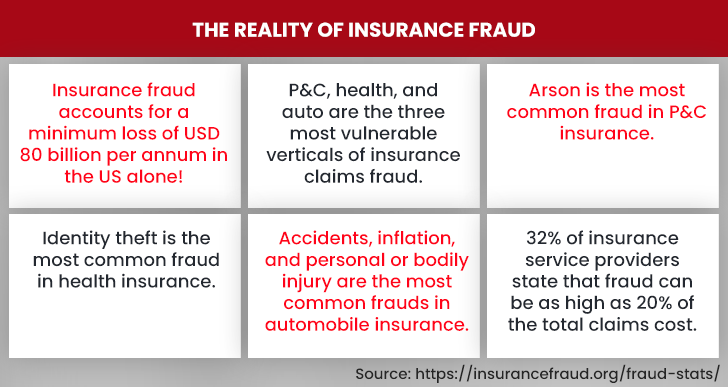
**PROPOSED METHODOLOGY**

**3.1 Proposed System**

We have a set of rules and anomalies for the creation of the raw data. The rules and anomalies are depended upon a specific set of attributes. This paper focuses on fraud claims of auto/vehicle-related to insurance. The raw data consists of details of the policy and claims. The raw data is transformed into a dataset and is fed into the classification algorithms like Random Forest, and Logistic Regression.

The transformed data set contains of different sets of attributes. By using these attributes, we can check every detail regarded to the claim. In case of normal claims, the gap claim occurrence date and claim report date is less than seven days, where the whole document is submitted with proof. In case the gap between policy effective date and claim occurrence, date is less than five days, also check the claims on same vehicle during all policy periods.

In the Logistic Regression algorithm each attribute is separately trained and tested. The random forest algorithm helps us validate or predict the data. By using confusion matrix, we can calculate performance of each algorithm. Confusion matrix provides information on false positive (fp), false negative (fn), true positive (tp), true negative (tn) which is used for assessing the percentage of test data which has been correctly labelled by the classifier. With the help of this information, the accuracy, precision, and recall is calculated.



*Figure 3*

**IMPLEMENTATION AND RESULTS**

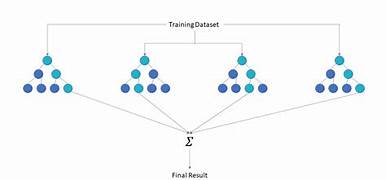
**4.1 Languages Used**

1. **Python**: Python is a high-level, interpreted programming language that Guido van Rossum developed and was made available in 1991. Since then, it has grown in popularity and is now one of the most widely used programming languages, with uses ranging from web development to data analysis and scientific computing. Python's simplicity and usability are among its most important characteristics. The syntax is simple and straightforward, making the language easy to learn and use. It is also designed to be readable and intuitive. With a sizable standard library that offers a wide range of built-in functions and modules for a number of applications, Python is also incredibly adaptable. Because Python is an interpreted language, code can be run without first needing to be compiled.

2. **Machine Learning**: In the quickly expanding subject of machine learning, algorithms and models are created that can recognise patterns in data and predict outcomes without explicit programming. Python is a well-known programming language with a large community of tools and libraries that makes it ideal for machine learning. Python offers a variety of potent machine learning libraries and frameworks, such as NumPy, SciPy, Pandas, Scikit-learn, and TensorFlow. In terms of data preprocessing, feature extraction, model selection, and model evaluation, these technologies offer a wide range of functionalities. With support for massive, multi-dimensional arrays and matrices as well as a variety of mathematical operations, NumPy is a well-known Python library for numerical computing. A library called SciPy, which is based on NumPy, adds support for operations like optimisation, integration, and interpolation in scientific computing.

**Random Forest**:

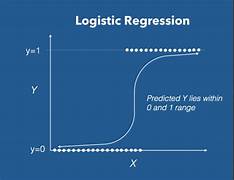
The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we’ll come back to later.



*Figure 4*

**Logistic Regression**:

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.



*Figure 5*

**4.2 Steps involved in Implementation**

**4.2.1 Constructing Dataset**

The Dataset consists of 32 different attributes to analyse the claims fraud. The 32 attributes are individually assessed using data visualization to understand the percentage of Fraud that can detected using a particular attribute.

The initial Dataset was collected from [www.Kaggle.com](http://www.Kaggle.com) and was named carclaims.csv.

32 attributes are:

'Month', 'WeekOfMonth', 'DayOfWeek', 'Make', 'AccidentArea', 'DayOfWeekClaimed', 'MonthClaimed', 'WeekOfMonthClaimed', 'Sex', 'MaritalStatus', 'Age', 'Fault', 'PolicyType', 'VehicleCategory', 'VehiclePrice', 'FraudFound\_P', 'PolicyNumber', 'RepNumber', 'Deductible', 'DriverRating', 'Days\_Policy\_Accident', 'Days\_Policy\_Claim', 'PastNumberOfClaims', 'AgeOfVehicle', 'AgeOfPolicyHolder', 'PoliceReportFiled', 'WitnessPresent', 'AgentType', 'NumberOfSuppliments', 'AddressChange\_Claim', 'NumberOfCars', 'Year', 'BasePolicy'.

**4.2.2 Importing Libraries**

In this project, we're utilizing a variety of libraries, including Matplotlib, NumPy, SK-Learn and Pandas is used to extract the chat and transform them into data frames.

**Matplotlib**: Matplotlib is a Python module for data visualization. It is used to deliver dynamic, animated, and interactive visualizations in Python. With Matplotlib, you can create a wide range of visualizations, including line plots, scatter plots, bar plots, error bars, histograms, bar charts, pie charts, box plots, and more.

**Pandas**: An open-source Python library for data analysis and manipulation is called Pandas. It offers tools for data analysis and working with structured data that are simple to use. Series, a one-dimensional labelled array, and DataFrame, a two-dimensional labelled array, are the two primary data structures in Pandas. For activities including data wrangling, data cleansing, data exploration, and data analysis, Pandas is frequently employed. As well as handling missing data, reshaping and pivoting data, and merging and joining datasets, it offers functionality for these tasks. In a variety of disciplines, including finance, economics, social sciences, and more, Pandas is a popular choice for data analysis.

**4.2.3 Data Pre-Processing**

The user inputs the Exported WhatsApp chat text file to the system. This text file consists of raw data with date, usernames, message, media data and group notification messages. The main objective of this data processing is to split the data into different categories and represent the whole set of data separately. We use the data frames to represent the data where each column consists of a particular set of data. The stop words are removed from the messages before inputting the words into the machine learning model for the emotion detection. The words other than the stop words which help in detecting the emotion of the individual user are recorded into a text file.

**3.2.4 Model Creation**

The machine learning model is created using the dataset of the words and the emotions. We use the Multinomial Naive Bayes theorem to build the model which is trained on 70% data of the dataset and is tested on the remaining 30% of the data. This model takes the text file created and runs the naive bayes algorithm in order to predict the emotion from the trained data and later predicts the overall emotion of the user and outputs the results to the user in the form of pie chart.

**Data Rescaling :**

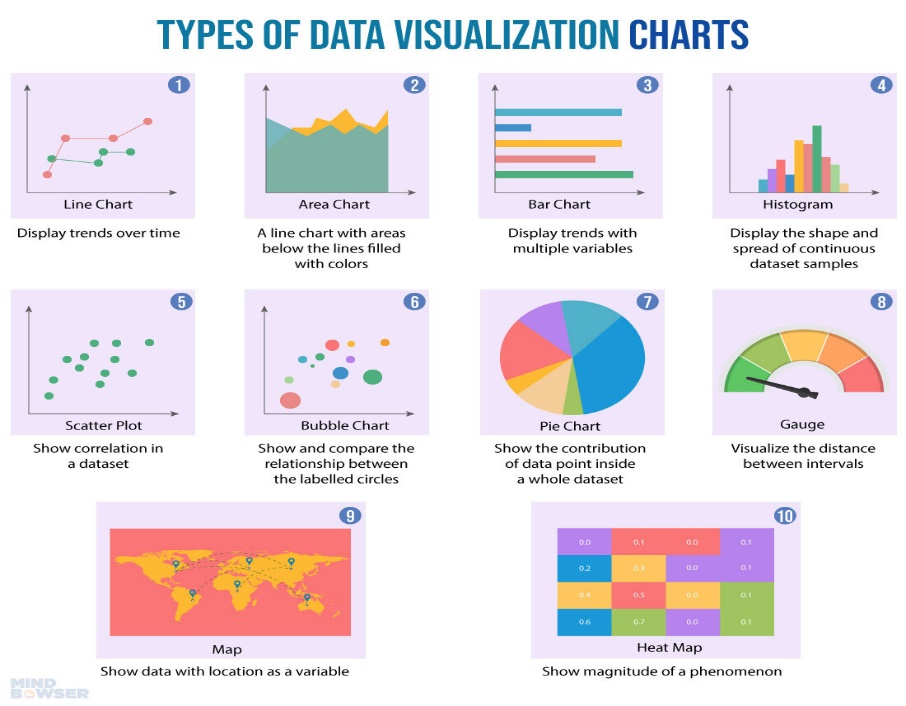
Your pre-processed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume. Many machine learning methods expect or are more effective if the data attributes have the same scale. Two popular data scaling methods are normalization and standardization. There are 2 main categories of Data Rescaling

1.Upscaling

2.Downscaling

**4.2.5 Data Visualization**

The data which is analysed such as the busiest days, most active users, count of messages, count of media messages, top active users, emotion of the individual etc are given as a result to the users in the form of graphs and charts.



*Figure 6*

**4.3 Sample Code**

**4.3.1. Importing Required Libraries:**

**Import re**: The functions in this module allow you to determine whether a given string matches a given regular expression.

**Import os**: Python offers a mechanism to communicate with the underlying operating system using its "os" module. It offers tools for managing processes, environment variables, and many other things in addition to working with files and directories.

**Import pandas as pd**: This library provides data structures to deal with different types of data.

**Import matplotlib.pyplot as plt**: This library helps in data visualization like plotting graphs, bar charts.

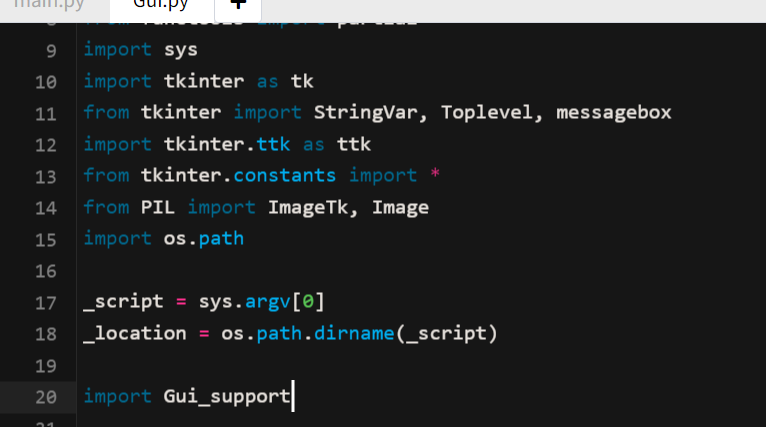
**Import numpy as np**: This library provides function to deal with arrays in python

**Import TkInter**



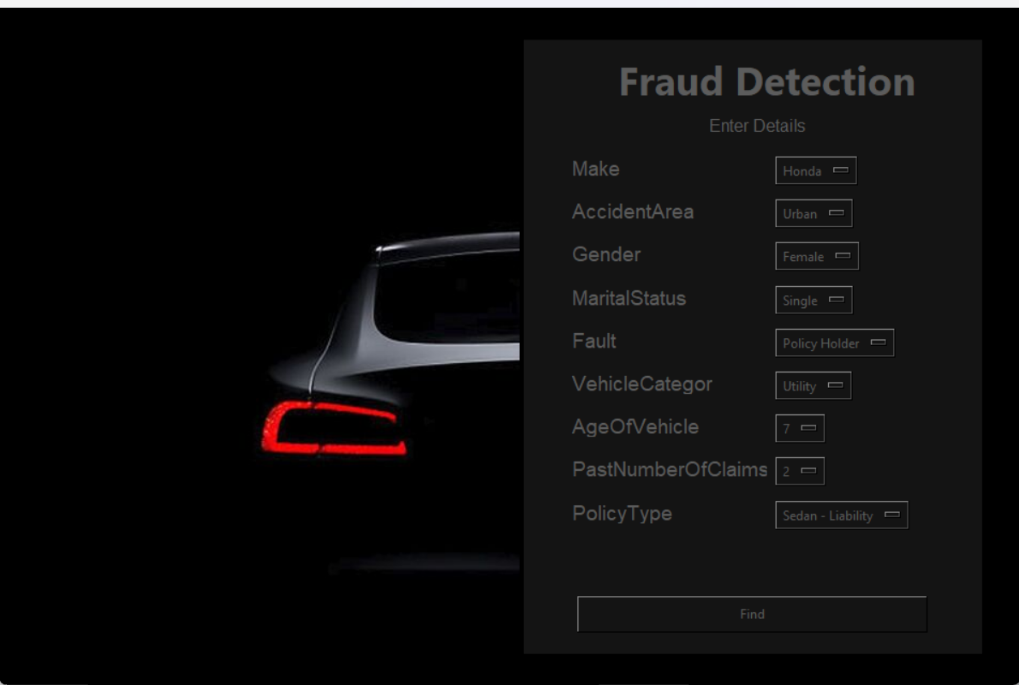
*Figure 7*

**4.3.2. Creating a GUI Interface:**



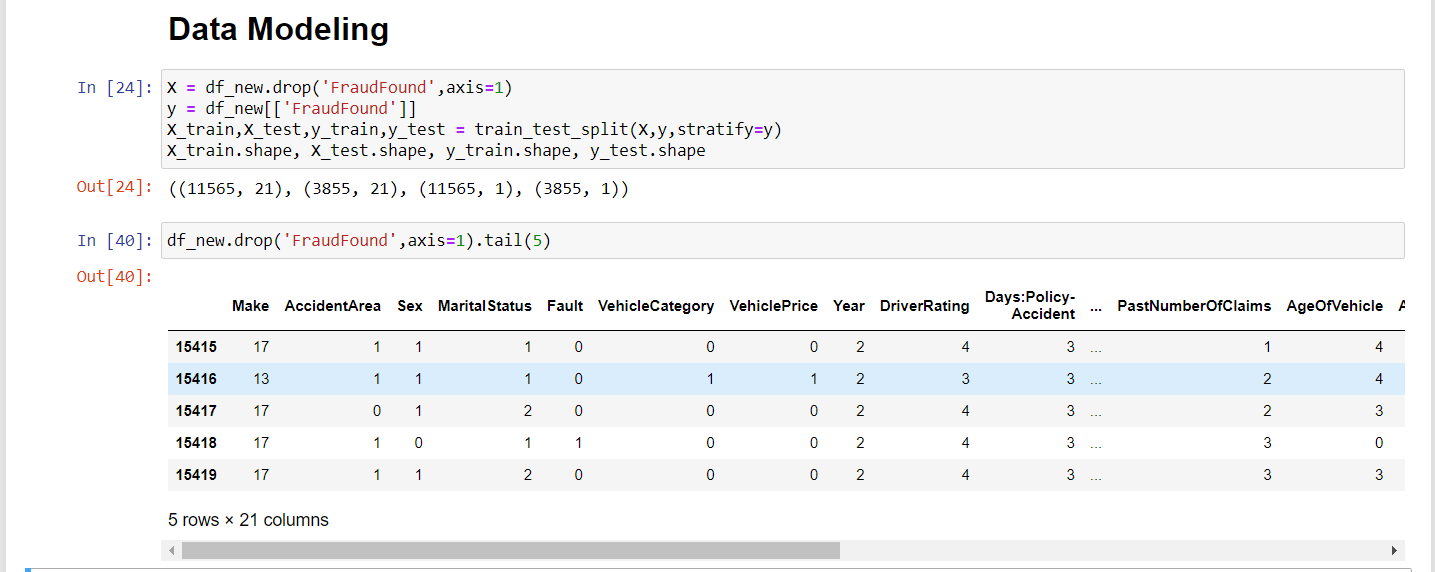
*Figure 8*

GUI Created

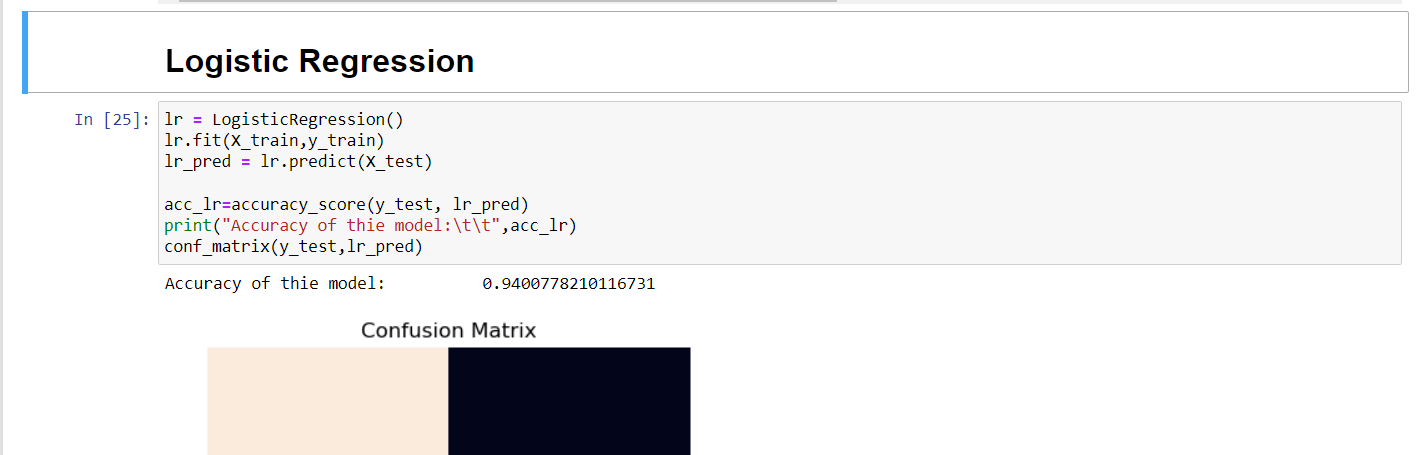


*Figure 9*

**4.3.3.Modelling The data: Train and Test**



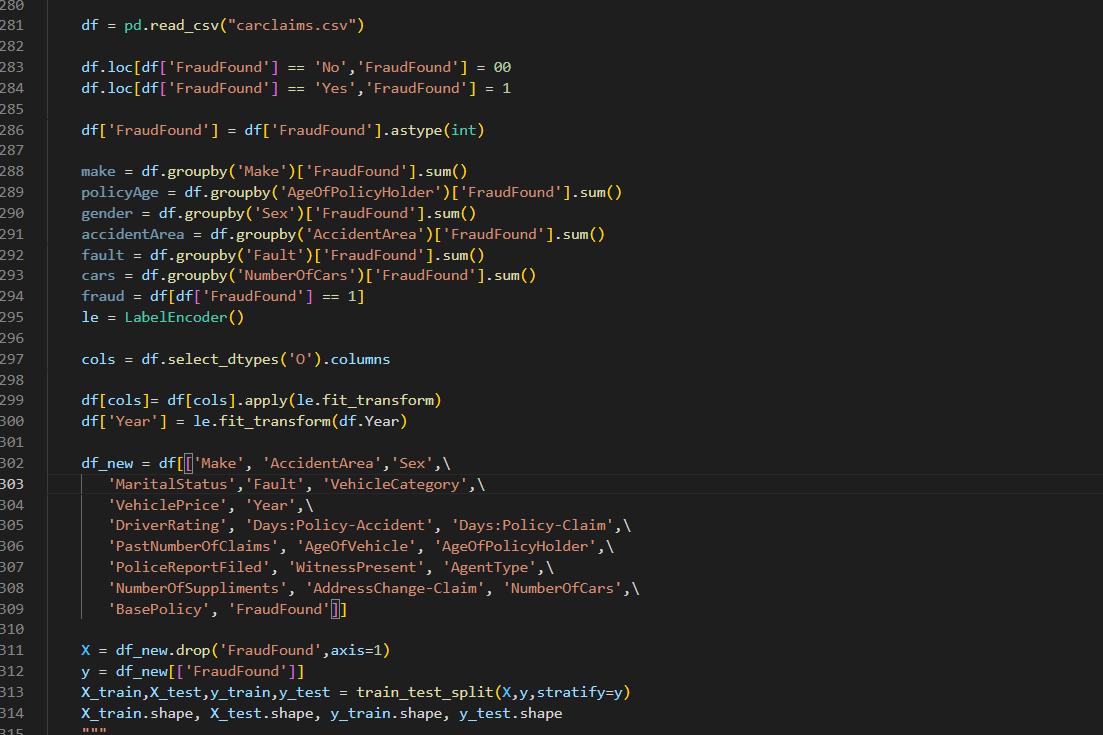
*Figure 10*



*Figure 11*



*Figure 12*



*Figure 13*

**TESTING : TEST CASES AND RESULTS**

1.The Accuracy of Logistic Regression in this model is: 0.9400778210116731

2.The Accuracy of Random Forest Classifier in this model is: 0.9374837872892348

**The dataset is extremely imbalanced and will not give accurate information. This can be overcome by one of the two ways :**

Upscale the Minority Class (FraudFound ==1) in this case

Downscale the Majority Class (FraudFound ==0) in this case

Because of this reason we need to upscale and downscale the data.

|  |  |  |
| --- | --- | --- |
| **s.no** | **Model** | **Accuracy Score** |
| 0 | Logistic Regression | 0.940078 |
| 1 | Upscale Logistic Regression | 0.740378 |
| 2 | Downscale Logistic Regression | 0.688312 |
| 3 | Random Forest Classifier | 0.937484 |
| 4 | Upscale Random Forest Classifier | 0.986343 |
| 5 | Downscale Random Forest Classifier | 0.742424 |

Hence based on the accuracy values we will be implementing upscale RandomForest Classifier on the Data considering the 21 arguments.

After implementation in a Sample Case Function:

Make- Ford

Accident Area- Urban

Gender- Male

Marital Status- Married

Fault- Policy Holder

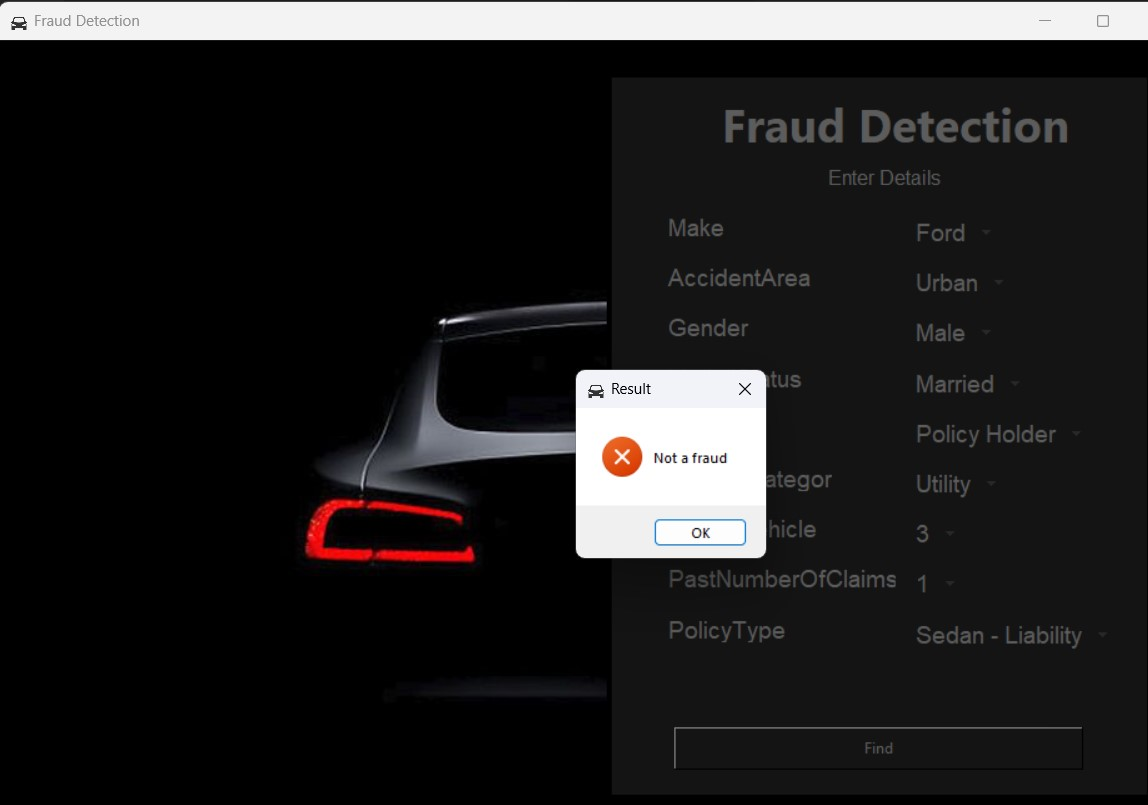
Vehicle Category- Utility

Age of the Vehicle- 3

Policy Type- Sedan-liability

Past Number of Claims- 1

Output:

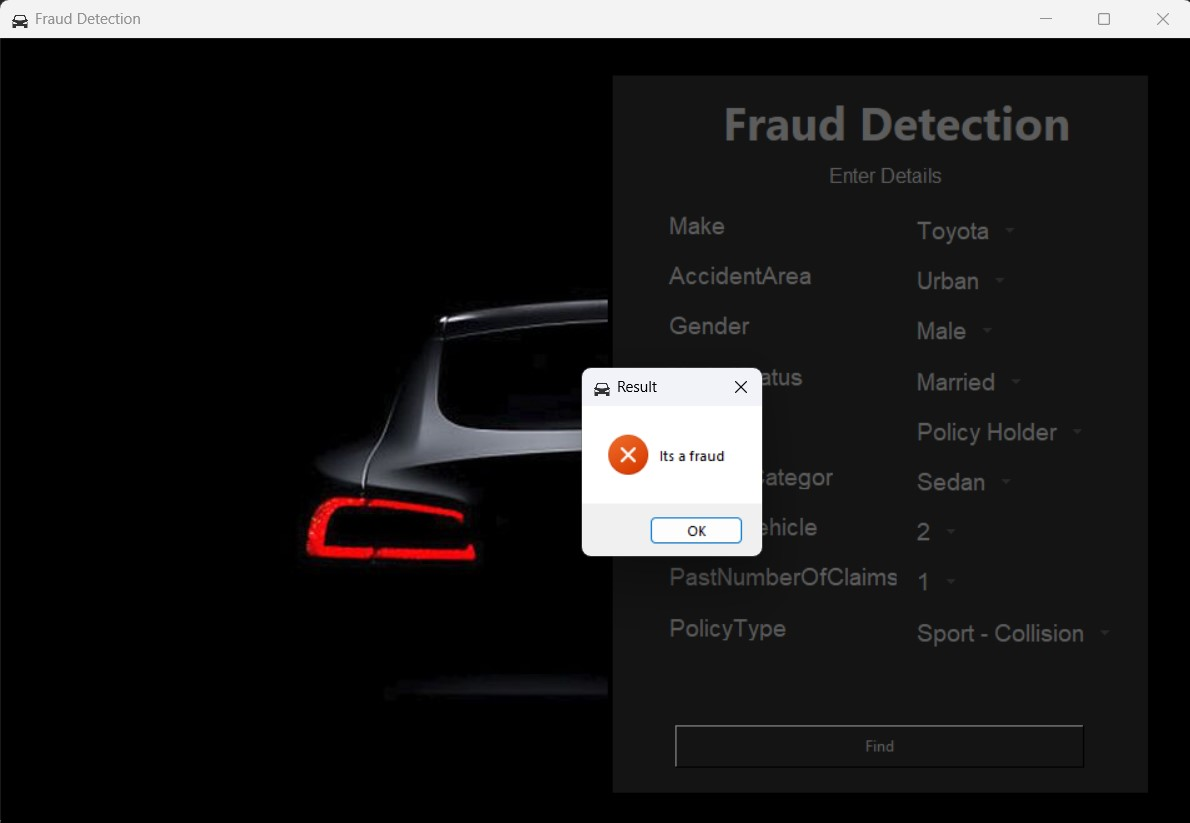


*Figure 14*

In this case the values inserted match the pattern of NOT A FRAUD

After making certain changes in the attribute values as shown below-

It classifies under the category of FRAUD and a pop-up is shown



*Figure 15*

**SUMMARY**

**5.1. Conclusion:**

As the different countries around the world evolve into a more economical-based one, stimulating their economy is the goal. To fight these fraudsters and money launderers was quite a complex task before the era of machine learning but thanks to machine learning and AI we are able to fight these kinds of attacks. The proposed solution can be used in insurance companies to find out if a certain insurance claim made is a fraud or not. The model was designed after testing multiple algorithms to come up with the best model that will detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

**5.2. Recommendations:**

The dataset which we used for building this insurance predictive model was taken from Kaggle and it was the data between 1994 to 1996. It will be good for us if we collect the new dataset of the past 2 to 5 years To determine whether the suggested solution would perform well when compared to other datasets that might serve as imitations, testing random combinations or a predetermined set of parameters is advised or to test the model to a similar type of dataset. surroundings that are different from the environment created after data cleansing. Reducing the number of characteristics is advised to cut further computational costs. The study was done by the Machine learning supervised techniques, which are used to build insurance claims predictive models. We have used Random Forest and Logistic Regression amongst all these algorithms Random Forest performed exponentially well on the dataset.

**5.3. Future Work:**

In order to compare the effectiveness of machine learning and deep learning methodologies, future research should focus on attempting to use an advanced or recently obtained dataset. Additionally, it is advised to utilize a different dataset in light of the fact that the one being used is unbalanced. Additional evaluation should be done to determine feature relevance across various datasets that may or may not have similar characteristics in order to develop a much more universal method to feature selection and focus. Because this research has been done by using all features in the future, we will do the feature selection to measure the variance between the total and selected features.

**REFERENCES**

[1] Belhadji, E., G. Dionne, and F. Tarkhani, “A Model for the Detection of Geneva

Papers on Risk and Insurance Theory”, 25: 517-538, may 2012

[2] Crocker, K. J., and S. Tennyson,” Insurance Fraud and Optimal Claims Settlement Strategies: An Empirical Investigation of Liability Insurance Settlements” The Journal of Law and Economics, 45(2), april 2010

[3] Kajia muller, “The Identification of Insurance Fraud – an Empirical Analysis Working papers on Risk Management and Insurance” no: 137, June 2013.

[4]Tennyson, S. & Salsas-Forn, P, “Claims Auditing in Automobile Insurance - Fraud Detection and Deterrence Objectives”, The Journal of Risk and Insurance, Vol. 69, No. 3, pp. 289-308,sep 2009

[5]Bhowmik, R., “Detecting Auto Insurance Fraud by Data Mining Techniques”, Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, april 2011