International Journal of Advanced Scientific Innovation Volume 05 Issue 10, October 2023
ISSN: 2582-8436

Vehicle Insurance Fraud Detection Using Machine Learning

H Ranjitha[*2], Joshni PS[#3],Vaishnavi CS
Visvesvaraya Technological University, Karnataka, India

** Corresponding Author(s),Email(s): rajithah08@gmail.com*

Abstract **-** There are thousands of companies in the insurance industry globally, and collect premiums totaling more than $1 trillioneachyear.
Insurance fraud occurs when a person or organisation submitsa fake insurance claim in an effort to collect money or benefitsto which they are not legally entitled. An insurance fraud is thoughtto have a total financial impact of over $40 billion. Deterring insurance fraud is thus a difficult issue for the insurance sector. The established method for detecting fraud is focused on creat-ing heuristics around fraud indicators. The most prevalent form of insurance fraud is auto fraud, which is accomplished by filing false accident claims. This essay focuses on finding auto-vehicle fraud.

Keyword—Machine Learning,Insurance,Fraud detection,Auto insurance ,claim records.

I. INTRODUCTION:

The insurance sector is now embracing efficient fraud control.Some people defraud businesses in order to receive payment, while others pay premiums. Hard insurance fraud and soft insurance fraud . Hard insurance fraud is described as the deliberate fabrication of an accident. Soft insurance fraud occurs when a person files a legitimate insurance claim but falsifies a portion of it. Customer satisfaction will increase if acompany has a solid fraud detection and prevention management system. The higher satisfaction will result in lower loss adjustment costs. A person who commits insurance fraud is engaging in improper behaviour in an effort to gain favourable treatment from the insurance provider. \cite{r1} Soft insurance fraud and hard insurance fraud are the two categories of fraud. Soft scams are more prevalent and include policyholders exaggerating valid claims. They are also known as opportunistic con artists. An intentional loss, such as the theft of a car or the setting on fire of goods covered by an insurance policy, is considered a hard fraud.

The goal of the research is to create a model that can spot fraud in auto insurance. The difficulty with machine learning fraud detection is that it is much less common than legal insurance claims. Imbalanced class classification is the issue.\cite{r2}To provide a real time application for detecting the fraud claims.By helping the insurance companies from identifying such fraud claims and prevent themselves from loss.Analytical science is a burgeoning area. the creation of new models, software tools, and procedures, and their evaluation.This auto insurance detection is very helpful for to find weather checked system is genuine and fraud. We can get accurate results by comparing the performance of machine learning algorithm with other algorithm. We can able to identify fraud insurance easily. Using this model we can quickly identify the fraud insurance.Results indicate a substantial relationship between effective elements in fraud that is similar. easy Nave Bayes to find fraud in data from vehicle insurance. And vehicle insurance fraud detection used to find weather the checked system is genuine or fraud.An insurance claim is a formal request to the insurance industry asking insurance company for the claim amount based on the terms and conditions of insurance policy. The fraudulent claims in vehicle insurance are mostly done by faking accidents.\cite{r3} Fake claims in the insurance domain make the insurer to take more time for recovering. When the fraud claims increase, overall cost of insurance is increased to settle the financial health. So the fraudulent claims are a serious problem to the insurance companies and to the government. This paper proposes a solution for the problem of detecting fraud vehicle insurance claims by using Machine Learning techniques and Naïve Bayes algorithm. Usage of Naïve Bayes algorithm increases the percentage of right prediction .

II. Literature Review

Artificial intelligence (AI) in the form of machine learning enables computers to learn without having to be explicitly programmed. The development of computer programmes that can adapt to new data is the main goal of machine learning. In this post, we'll go through the fundamentals of machine learning and look at how a straightforward machine learning algorithm is implemented in Python.

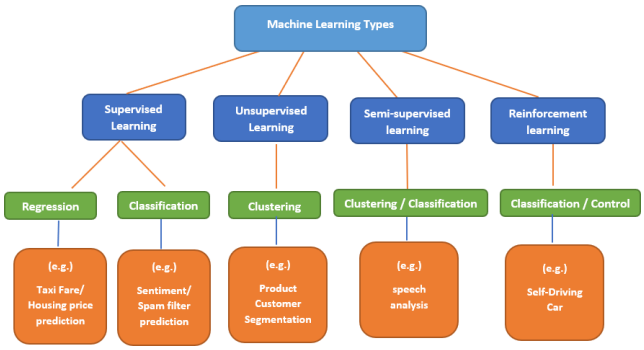
Machine learning, as the name implies, is the process by which computers learn automatically without beingexplicitly programmed or receiving direct human assistance. The first step in the machine learning process is to provide them with high-quality data, after which the computers are trained by creating different machine learning models utilising the data and various methods. What kind of data wehave and how we want to use it determine the algorithm we use.

There are 4 types of Machine Learning, they are Supervised Machine Learning, Unsupervised Machine Learning, Semi-Supervised Machine Learning, Reinforcement Learning. Supervised Machine Learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output. The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then wetest the model by providing a sample data to check whetherit is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is thesame as when a student learns things in the supervision of the teacher. The example of supervised learning is spamfiltering. Supervised learning can be grouped further in two categories of algorithms Classification and Regression.

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar pattrns .In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms Clustering and Association.

Semi-supervised learning is an important category that lies between the Supervised and Unsupervised machine learning. Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for the corporate purpose, it may have few labels.

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.



Machine Learning Algorithms

KNN(K Nearest Neighbors): K-Nearest Neighbors is one of the simplest Machine Learning algorithms based on Supervised learning technique. KNN is also a lazyalgorithm (as opposed to an eager algorithm). this means is that it does not use the training data points to do any generalization .In other words, there is no explicit training phase or it is very minimal. In other words, there is no explicit training phase or it is very minimal. This also means that the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data. To be more exact, all (or most) the training data is needed during the testing phase. KNN can be used for classification — the output is a class membership (predicts a class — a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. It can also be used for regression — output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbors.

Naive Bayes: The naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and used class levels represented as feature values or vectors of predictors for classification. Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these

features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

Decision Tree: Decision Tree in machine learning is a partof classification algorithm which also provides solutions to the regression problems using the classification rule(starting from the root to the leaf node).Its structure is like the flowchat where each of the internal nodes represents the test on a feature (e.g., whether the random number is greater than a number or not), each leaf node is used to represent theclass label(results that need to be computed after taking all the decisions) and the branches represents conjunction conjunctions of features that lead to the class labels. Decision Tree in Machine Learning has got a wide field in the modern world.There are a lot of algorithms in ML whichis utilized in our day-to-day life. One of the important algorithms is the Decision Tree used for classification and a solution for regression problems. As it is a predictive model, Decision Tree Analysis is done via an algorithmic approach where a data set is split into subsets as per conditions. The name itself says it is a tree-like model in the form of if-then- else statements. The deeper is the tree and more are the nodes, the better is the model.

III. RELATED WORK

Machine learning analysis of vehicle insurance data to identify fraud. 2020 Sai Pranavi et al. In the world, there are over a million insurance businesses that deal with a lot of data. The most prevalent type of fraud in enterprises acrossall sectors is now insurance fraud. Fraud detection is crucial in the modern world since the insurance sector is one of the fastest-growing sectors. Constructing a model that can identify motor insurance fraud is the aim of this study. The difficulty with machine learning for fraud detection is that fraudulent insurance claims are far less frequent than legitimate ones. KNN and Random Forest algorithms.Choose Random Forest, K-Nearest Neighbors, and K Best. initially checks to see if the claim request has any potential for fraud. Verify automobile registration information from the public directory if fraud is suspected. Fraud must be addressed since it is a significant issue in today's society. Create systems that can detect fraud in the provided data to tackle these issues. These systems are created utilising a variety of machine learning methods, including neural networks, naïve Boyes, KNN, and Random Forest. Itinvolves foreseeing fraud. These methods are then contrastedusing five criteria from various angles.

Healthcare Financial Fraud Detection Using Deep Learning and Machine Learning Methods. 11 September 2021. Abolfazl Mehbodniya and others. One of the well-known industries where a lot of data may be gathered on finances and health is the healthcare sector. Due to the widespread use of credit cards in the healthcare industry and the ongoing development of electronic payments, credit card fraud monitoring has been difficult financially for the various service providers. Healthcare is one of the numerousindustries where credit cards are becoming increasingly common. Credit cards have improved the convenience and accessibility of online transactions. However, fraudulenttransactions cause a significant loss of cash each year that might rise in the upcoming year. Dataset, Sequential Model, KNN, Logistic Regression, Naive Bayes Classifier. Dataset, Sequential Model, KNN, Logistic Regression, Naive Bayes Classifier. The UCI Machine Learning Repository is the dataset's original source. The dataset contains information- related transactions that were carried out using credit cards asthe default payment method by various Taiwanese clients. There are probably six main data mining approaches used to compare the accuracy. This article presents a research on the use of deep learning and machine learning approaches to detect credit card fraud. The several common models, including Sequential Model, Decision Tree, Random Forest, and Naive Bayes, are described and put to the test using real data.

A Fraud Detection Use-Case for the Vehicle Insurance Industry: Detecting Vehicle Insurance Claim Fraud. Mohammed, Dilkhaz Y. 2021 December.Since the beginning of insurance, there has been an increase in both the amount and frequency of insurance fraud events. Conspiring to create fraudulent or inflated claims about property damage or personal injuries as a result of an accident is known as vehicle insurance fraud. The goal of this research is to create a model that can identify fraudulent vehicle insurance. fraud problems with machine learning. The five main classificationalgorithms are Naive Bayes, Bayes Net, J48, Random Forest,and Random Tree. Finally, other scholars could change this study by employing alternative techniques. Data preprocessing techniques including variable transformation, missing value handling, data aggregation, sampling, and discretization. The Oracle Dataset provided the dataset. Both information about the container and information about the insurance policy are included in the dataset. It also includes the accident data that was utilised to support the claims.Fraud analytics is a rapidly expanding field. The goal ofresearch is to create, create, and evaluate new models, software tools, and processes to help in the battle against terrorism.

Detection of auto insurance fraud. July 2020 Kavya Priya M L, et al. Auto insurance fraud detection is a highly valuable tool for the insurance business and for identifying fraud.Here, we'll employ a few approaches and strategies to address this problem. The sector faces a serious problem withinsurance fraud. It's challenging to spot fraud claims. The vehicle insurance sector is in a unique position to benefit from machine learning in this situation. The three modules and three actors that make up this suggested system are: Admin: The person in charge of overseeing the whole

application.The person in charge of the branch who accepts insurance claimsPublic: The person who submits an insurance claim.Algorithm of Naive Bayes.In this instance, the dataset comprises customer and insurance information as well as details on the insurance claim procedure that are used to identify fraud. Necessary Bayes algorithm. Detecting insurance fraud is a difficult endeavour since it has become so prevalent in recent years. The proposed approach attempts to provide a system that can assist in accurately identifying potential frauds. This car insurance fraud detection tool is particularly useful for determining if a system check is "GENUINE" or "FRAUD." Therefore, this will be useful for locating fraud insurances and accuracy.

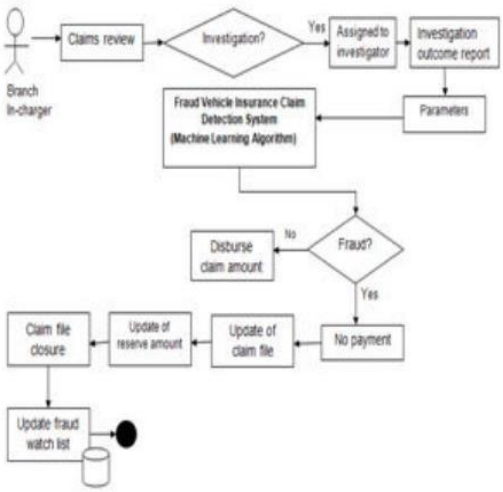
IV. SYSTEM ANALYSIS

4.1 EXISTING SYSTEM

The insurance business faces a difficult difficulty with the detection of insurance fraud. When someone purposely fabricate an accident, that is considered hard insurance fraud. Soft insurance fraud occurs when a person makes alegitimate insurance claim but falsifies a portion of it. When someone purposely fabricate an accident, that is considered hard insurance fraud. Soft insurance fraud occurs when a person makes a legitimate insurance claim but falsifies aportion of it. The proposed method allows users to register on the website and send the first information report or complaint against a specific circumstance or person, unlike the current system, which only allows crimes to be recorded and increases the workload for authorised users. Excel spreadsheets, macros, and algorithms were initially used in the attempt to find fraud—yuck! The procedure, known as data matching, might occasionally be as simple as manually searching a spreadsheet for matches. What exactly does data matching entail? It occurs when evidence of the claimant's prior involvement in a claim is shown. That was the main area of fraud detection for a very long period. Due to the quantity of physical work required, this process was cumbersome, sluggish, and prone to mistakes (we called these false positives). They are required to capture transactional history data, an assistance with feature creation for basic or sophisticated machine learning models. more personnel takes a lot of time. consumes a significant amount of paper. manual computation is required. Higher authorities have no specific responsibility. The system has to be automated better in order to get around all these restrictions and improve functioning accuracy.

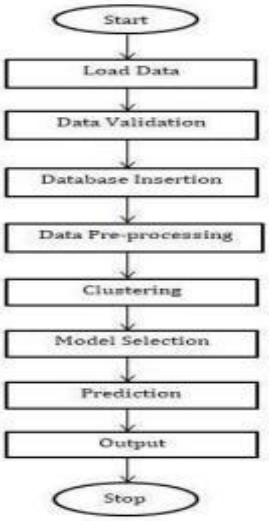
3.2 PROPOSED SYSTEM

The proposed solution uses real-time technology.This system is intended for the insurance sector.predicts claims of insurance fraud.also uses machine learning techniques to forecast fraud claims.uses categorization criteria for fraud prediction, such as naive Bayes, KNN, or ID3 algorithms.uses historical claim data for processing and predicting claim fraud.The proposed system is created with Python.It is a reasonably straightforward, widely accepted form of technology.It is a real-time application that is excellent for the insurance sector.It is very important and reduces considerable loss by identifying fraudulent claims.better determine risk.Detecting fraud more quickly.AI is more capable of identifying.Fraud analytics significantly speed up the process for insurers.



V. METHODOLOGY

It entails researching the theories and ideas that underpin the procedures employed in your industry in order to create a strategy that is in line with your goals. Methods are the particular instruments and practises you employ to gather and examine data (for example, experiments, surveys, and statistical tests).

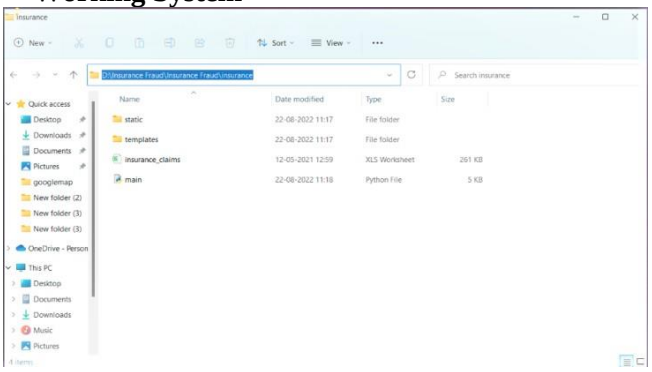


The model's flow is depicted in Fig. 5.1, where the dataset is in csv (comma separated values) format. The XGBoost (eXtreme Gradient Boosting) technique is used for classification. The accuracy of recognition has increased dramatically as a result of the use of K-NN and decision trees in the recognition job. Starting with fig, we must load the data in order to begin the process and determine whether the data is genuine or not. After that, add the data to the database. Data Preprocessing involves changing or removing the data. Model selection is the process of choosing the best model from a pool of candidates based on performance criteria. The output of an algorithm that has been trained on a dataset for a certain result is known as a prediction.
Once a specific output is obtained, the procedure is terminated.

VI. RESULTS AND DISCUSSIONS

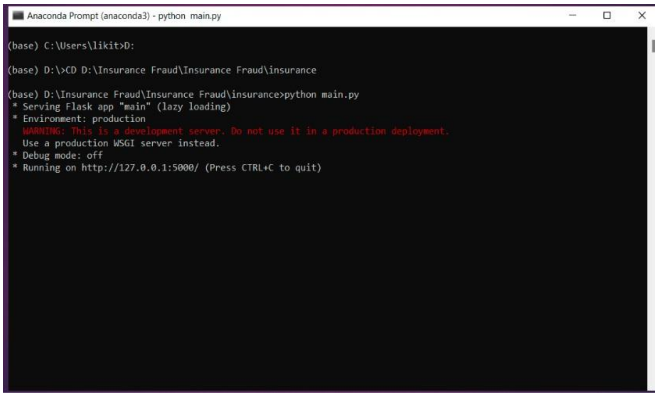
Here, we'll explore the project's outcomes and explain how they came about. The operating system, the Anaconda Prompt, the browser, and eventually themain page of our project are all demonstrated in this.

Working System



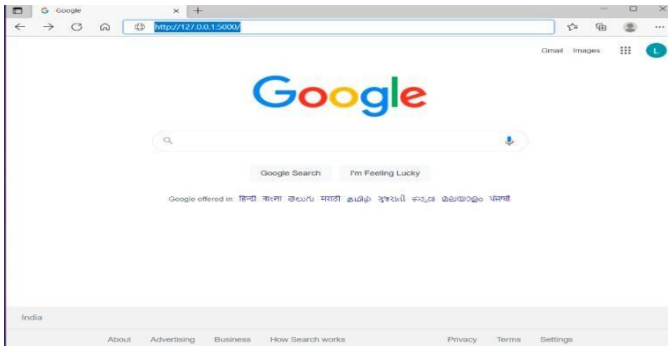
In Fig 6.1 shows that the working system and here inserted that the dataset of the vehicle insurance. In this we should copy the file path.

Anaconda Prompt



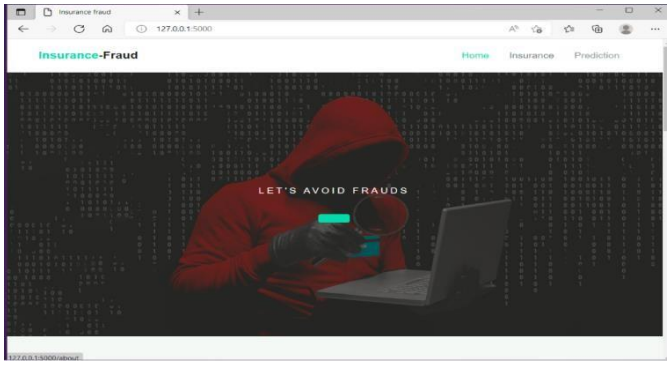
In Fig 5.2 We have to paste the file path in Anaconda Prompt with a command (python main.py) and run. Later, copy the IP Address.

Browser



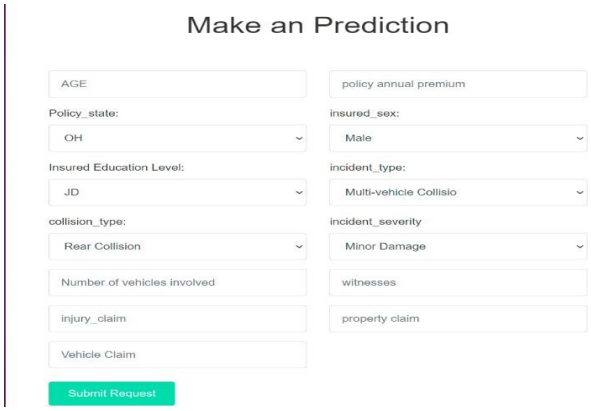
In Fig 6.3 we can open any browser like google, chrome etc, and paste the IP Address.

Output



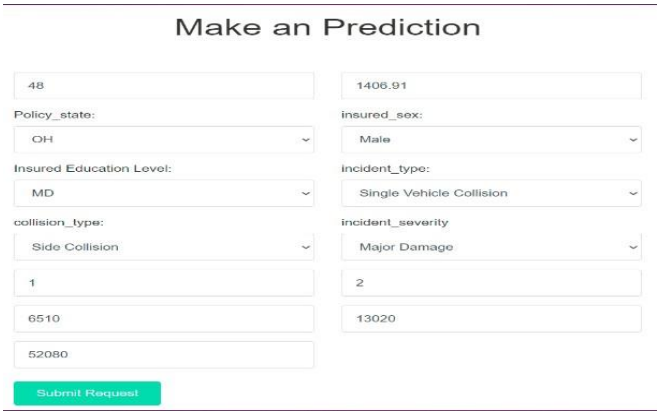
In Fig 6.4 shows that the main page of our project**.**

Insurance Fraud Prediction



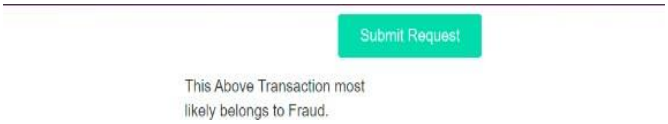
In Fig 6.5 Fill Details and Submit Request

Fill Details and Submit Request



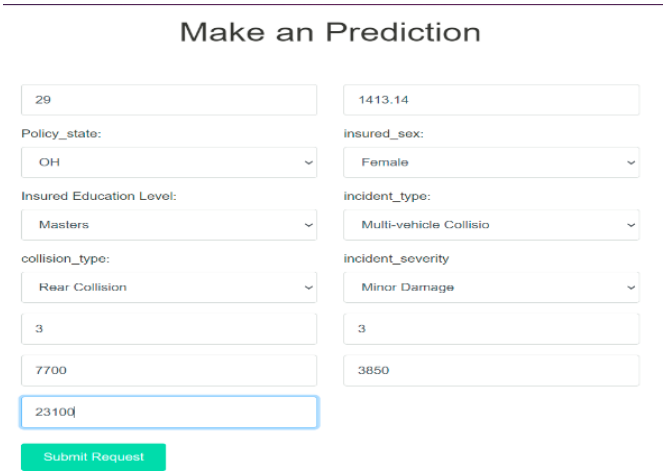
In Fig 6.6 we should fill the details and submit the request.

If Fraud , It Shows



In fig 6.7 after filling the details if it is fraud it shows the result as “This Above Transaction most likely belongs to fraud”.

Fill Details and Submit Request



In Fig 6.8 we should fill the details and submit the request.

If Not Fraud, It Shows

In fig 6.9 after filling the details if it is not a fraud it shows the result as “This Above Transaction most likely belongs to not fraud”.

VII. CONCLUSIONS

Insurance claim fraud has been a problem for this business from the start, making it a difficult process to discover. The proposed approach attempts to provide a system that can assist in accurately identifying potential frauds. The suggested approach determines if the insurance being claimed is "FRAUD" or "GENUINE". This makes it easier for insurance firms to identify scams quickly and accurately. Since we are aware that frauds cost businesses money, our project can determine if a certain person is committing fraud or not.

VIII. REFERENCES

[1] https://www.law.cornell.edu/wex/insurance fraud
[2] https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3
[3] Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud.DOI: 10.1109/ICDSBA.2018.00104
[4] C. Phua, D. Alahakoon, and V. Lee, “Minority report in fraud detection:classification of skewed data,” Acm sigkdd explorations newsletter,vol. 6, no. 1, pp. 50–59, 2004.
[5] https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML
[6] Types of Machine Learning – Javatpoint
[7] Supervised Machine Learning Algorithms 2 Types of Learning Algorithm (educba.com)
[8] Unsupervised Machine learning – Javatpoint
[9] Semi-Supervised Machine Learning Algorithms HackerNoon
[10] reinforcement machine learning algorithms - Search (bing.com)
[11] K-Nearest Neighbor(KNN) Algorithm for Machine Learning – Javatpoint
[12] Naive Bayes Algorithm Discover the Naive Bayes Algorithm (educba.com)
[13] Decision Tree in Machine Learning Split creation and Building a Tree(educba.com)
[14] (99+) Analysis of Vehicle Insurance Data to Detect Fraud using Machine Learning IJRASET Publication - Academia.edu
[15] (PDF) Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques (researchgate.net)
[16] (PDF) Detection of Vehicle Insurance Claim Fraud A Fraud Detection Use-Case for the Vehicle Insurance Industry (researchgate.net)
[17] (PDF) Automobile Insurance Fraud Detection (researchgate.net)
[18] Auto Insurance Fraud Detection – IJARCCE
[19] https://ijesc.org/
[20] http://127.0.0.1:5000/