Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: XXXX-XXXX / Volume 1 Issue 1 January 2024 / Page No: 07-14
Paper Id: IJMSM-V1I1P102 / Doi:

Research Article

Machine Learning for the Identification of Credit Card Fraud

[1]Safrin S, [2]Madhu S

[1,2]Department of Computer Science Mother Teresa Women's University, Kodaikanal.

Received: 29 December 2023 Revised: 16 January 2024 Accepted: 28 January 2024 Published: 01 February 2024

Abstract - *The most prevalent problem in the world right now is identifying credit card theft. This can be explained by the rise in online transactions and e-commerce platforms. The most frequent ways that credit card fraud happens are when a card is lost or stolen and used without permission, or even when the cardholder uses their personal information for illicit purposes. The modern world has several credit card problems. To detect fraudulent behavior, the credit card fraud detection system was developed. The initiative aims to investigate machine learning techniques. In this research, we propose to detect fraudulent transactions by using the Kaggle dataset. We fit the dataset to Random Forest to ascertain whether a transaction is fraudulent or not. Finally, we compared the performance of XGBoost and LightGBM.*

Keywords - *Fraudulent transactions, Fraud prevention, Feature engineering, Fraud detection systems, Data mining, Fraud patterns, Predictive modeling.*

I. INTRODUCTION

A. About Machine Learning

In the fields of artificial intelligence (AI) and computer science, machine learning simulates human learning processes and gradually increases the accuracy of those simulations through the use of data and algorithms. A relatively recent field of research called "machine learning" makes it possible for computers to learn on their own by utilizing historical data. Using a range of methods, machine learning creates mathematical models and forecasts outcomes based on past experience or knowledge. These days, common uses include Facebook auto-tagging, image and audio recognition, email filtering, recommender systems, and more. An introduction to the field and a variety of machine learning methods, including supervised, unsupervised, and reinforcement learning, are covered in this course on machine learning.

B. Project Overview

To prevent their customers' accounts from being impacted and charged for items they did not purchase, credit card firms must be able to discern between legitimate and bogus transactions with their credit cards. To reduce losses, all credit card issuers must have fraud detection systems in place. Fraud creates large financial losses for many financial companies and institutions because criminals are always coming up with new ways to break the law and commit crimes. The volume and speed of online transactions for clients that engage in e-commerce have increased recently. Fraudsters usually use a number of methods to obtain credit card information and quickly transfer large amounts of money.

Even with a large dataset, it delivered decent output and was quick to train, despite its outstanding performance. Because it is lighter and faster, we have chosen to utilize it as our classifier. In this project, we use the Random Forest algorithm's Decision Tree (DT) techniques to construct a fraud detection system. Furthermore, the data processing that was carried out has a connection to feature engineering. The hidden knowledge within the data is found by means of cleaning, correcting, extracting, and selecting several data attributes. Our goal is to extract the varied information between the future Decision Trees (DT) model in order to more effectively detect fraudulent transactions, distinct user behavior patterns, and fraud activity patterns in the data.

C. Architecture of Credit Card Fraud Detection

As society moves toward becoming cashless, credit card payments are growing in popularity. However, it's crucial to keep in mind that debit card theft is the most prevalent type of identity theft crime. Classification is one

of the primary goals of machine learning algorithms. Each purchase made with a credit card generates some data that can be utilized to create a classifier using machine learning methods. Time and money can be saved by almost instantaneously identifying forged transactions with the use of such a classifier in real-time. This post will utilize a random forest classifier to forecast the likelihood of an exchange.

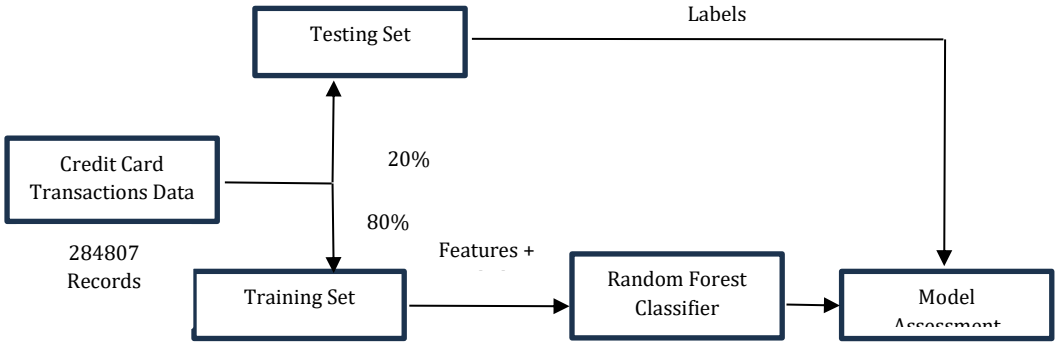

Figure 1. The Credit Card Fraud Detection Architecture

II. LITERATURE SURVEY

Ashutosh Kumar Singh, Jiendra Kumar, Nishtha Sakhuja, Pooja Tiwari, and Simran Mehta [1], As the world rapidly swings toward digitization and cashless transactions become the norm, credit card usage has increased. the vastly increased amount of fraud associated with it. It is imperative to conduct an investigation and distinguish between transactions that are fraudulent and those that are not. We offer a comprehensive review of every method this study employed to detect credit card fraud. Among these methods are hidden markov models, neural networks, and Bayesian belief networks.

XinYing Chew, Wai Peng Wong, Bahari Belaton, Khai Wah Khaw, and Esraa Faisal Malik [2], Over time, financial crimes have harmed financial institutions more and more. Numerous to name However, these approaches do have substantial limitations because several hybrid procedures for a given dataset were not investigated further. To detect dishonest behavior, this research Initially, the hybrid models were created to detect credit card fraud using cutting edge machine learning methods. Secondly, hybrid methodologies were employed in their creation.

Harish Paruchuri [3], Companies want to offer more and more conveniences to their customers. Among these comforts is the opportunity to buy goods online. Nowadays, customers can buy the required supplies online, but there is a risk of fraud by unscrupulous individuals. Any cardholder's details can be stolen by fraudsters and used for online purchases until the cardholder tells the bank to stop the card. The many machine learning algorithms that are utilized to recognize this kind of transaction are presented in this study.

K.M. Anil Kumar, [4], Research on data-driven social security fraud detection is somewhat lacking. Social projects are becoming increasingly popular in developing countries like India. The main objectives of social programs are to reduce, however fraud and deceit cause a sizable portion of the eligible population to be turned away from the program.

Drs. Johannes Jurgovsky and Yvan Lucas [5], issue for the that handles electronic payments. We look at the specifics of data-driven credit card fraud detection in this survey, along with a range of machine learning algorithms, to handle each of its complex issues and identify fraudulent transactions created fraudulently on behalf of the legitimate card holder. Specifically, we first introduce a commonly used identification problem, along with the dataset and its properties, selected metrics, and various methods to deal with imbalanced datasets. These are the first inquiries in any credit card fraud inquiry.

Sonam S. Patil, Rashmi, Chetan J. Awati, and Suresh K. Shirgave [6], Credit card fraud is one issue that has gotten worse recently. A large financial loss has had a substantial impact on credit card users, banks, and retailers. Among the finest methods for spotting fraud is believed to be machine learning. This study evaluates various machine learning-based fraud detection methods using performance metrics such as specificity, accuracy, and precision. Additionally, the study proposes an FDS that employs the supervised Random Forest method. The recommended approach improves the accuracy of credit card fraud detection.

Yanxia Sun, Zenghui Wang, and Emmanuel Ileberi [7], Recent developments in electronic payment and e-commerce have led to a surge in financial fraud events, including credit card fraud. Thus, it is essential to put in place mechanisms that can identify theft with credit cards. When using machine learning to detect theft with credit cards, the features of the scams are crucial and need to be carefully selected.

AlEmad, Meera [8], This program employs machine learning techniques to identify fraud with credit cards in order to prevent unauthorized users from accessing clients' accounts without authorization. Given the increasing global prevalence of credit card fraud, action needs to be taken to put an end to fraudsters. Limiting those kinds of operations would benefit the clients since their money would be recovered and credited to their accounts, helping to meet the project's primary goal of not charging them for goods or services they did not acquire.

Mohammad Sanaullah Chowdhury, Aapan Mutsuddy, Ibrahim Mohammad Sayem, and Abrar Hayat Nadim [9], There are more fraudulent cases everywhere due to the credit card system's quick and simple transactions. Algorithms for machine learning have been used to detect fraudulent transactions. The two main problems with fraud identification are: (1) the behavior of both legitimate and fraudulent entities is always changing; and (2) highly skewed datasets. The system's implementation was rife with machine learning techniques, variable extraction, and dataset sampling.

Subhash, K R Sumana [10], When thieves obtain cash advances from a credit card account that belongs to someone else, it's referred to as credit card fraud. Using the user's current accounts, opening an unknown credit card account in their name, physical credit card theft, account numbers, or PINs are some ways that this might occur.

Shazli Meraj and Mohammed Azhan [11], Any action that is intended to harm another party's finances is considered fraud. The expansion of digital currency in several nations has coincided with an increase in associated fraud. Every year, banks and credit card firms lose billions of dollars to these kinds of fraudulent activities, which severely reduces their earnings and has an adverse effect on the employment of numerous employees.

Om Shantanu Rajora, Dong-Lin Li, Chandan Jha, and Mukesh Prasad [12], This study compares ten different machine learning algorithms according to how well they work in a credit card detection application. Classification algorithms and ensemble learning are the two recognized kinds of machine learning techniques. Each category contains five different algorithms.

III. REQUIREMENTS SPECIFICATION

A. Functional Requirements

Product features are what are called functional requirements. Functional requirements include every feature that should be included in any development.

B. Non-Functional Requirements

Non-functional requirements specify the client's expectations for product design, security, accessibility, dependability, and performance.

a. Design Restrictions: Python must be used for project creation, and Windows must be used for project execution. We are implementing the Python code on Google Colab.

b. Reliability: No product malfunction should cause an interruption to an operation.

c. Availability: The program is available for use whenever you'd want.

d. Security has to be the top priority for any software that stores user-sensitive data.

Maintainability: The software administrator should be able to manage the data.

e. Portability: The project ought to function on any Windows OS.

C. Feasibility Study

In this stage, the viability of the project is examined, and a business proposal with a basic project plan and some cost estimates is presented. System analysis requires a feasibility assessment of the proposed system.
These are the three main factors that feasibility analysis considers. They are.

1. Possibility from an Economic Standpoint
2. Practicality in Technology
3. Practicality in Society

i). Possibility from an Economic Standpoint

Even with today's technology advances, a system that the business implements and uses must be a prudent investment. The system's economic feasibility is assessed by comparing its development expenses with the new system's potential benefits. The profits should equal or exceed the expenses. Profitability of the system is achievable. There is no need for any extra hardware or software.

ii). Practicality in Technology

The purpose of this study is to evaluate the system's technological requirements, or technical feasibility. The systems that are created shouldn't put too much strain on the technical resources that are available. The availability of technological resources will consequently be in high demand. As such, the client will need to follow specific guidelines [8]. Because implementing the intended system will only require minimal or nonexistent adjustments, it must have the fewest requirements possible.

iii). Practicality in Society

Finding out how much users accept the system is the aim of the study. This entails instructing the user on the proper usage of the technology. The user must view the system as required rather than as something to be afraid of. The methods employed to educate and acquaint people with the system will dictate the degree of adoption among users. As the final user of the system, it is imperative to enhance his self-assurance regarding the Register Module to enable him to provide constructive feedback.

IV. ALGORITHM DESIGN

Random Forest is one of the most popular and widely used algorithms among data scientists. Random forest is a well-liked supervised machine learning technique for regression and classification problems.One of the most important characteristics of the Random Forest Algorithm is its capacity to handle data sets containing both continuous variables, which are used in regression, and categorical variables, which are used in classification. It performs better in tasks involving classification and regression. Using random selections of the attributes and training data, numerous decision trees are built in a random forest. The model's ability to generalize is enhanced by this volatility, which also lessens overfitting. = decision tree in the forest. Each decision tree in the forest makes a prediction. For the procedure to work, a random subset of the features and data must first be selected. A decision tree is then built using this subset of the features and data. The process is done multiple times to create a forest of deciding trees.

A. Working of Random Forest Algorithm

Step 1: In the Random Forest model, a subset of qualities and data points are used to form each decision tree. Stated simply, a data collection consisting of k records is picked to yield m characteristics and n random records.
Step 2: A distinct decision tree is built for each sample.
Step 3: Every decision tree will yield a result.
Step 4: The techniques employed to evaluate the final outcome are averaging, regression, and classification-based majority voting.

V. DESIGN OF THE SYSTEM

The process of defining a system's components, architecture, interfaces, modules, and data that comply with standards is known as systems design. It is the process of identifying, creating, and designing systems to satisfy the specific needs and requirements of an organization or company.

A. Workflow Diagram

Figure 2 illustrates the proposed credit card fraud detection process. We are collecting credit card data. Then you may need to preprocess the dataset by cleaning the data, handling missing values, and applying machine learning-based classifier algorithms to calculate performance, i.e., whether credit card fraud may happen or not.

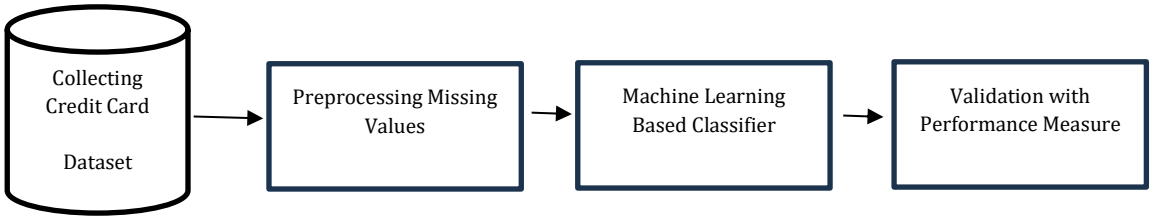

Figure 2. Workflow Diagram

B. Activity Diagram

Another crucial UML diagram for illuminating the system's dynamic components is the activity diagram. A flow graphic that shows how one activity leads to another is what an activity diagram basically is. One could refer to the action as a system operation. As a result, the control flow is switched between operations. This flow could occur branching, simultaneously, or sequentially. Activity diagrams use various elements, such as joins and forks, to manage various forms of flow control. The system's activity diagram displayed in Figure 3.

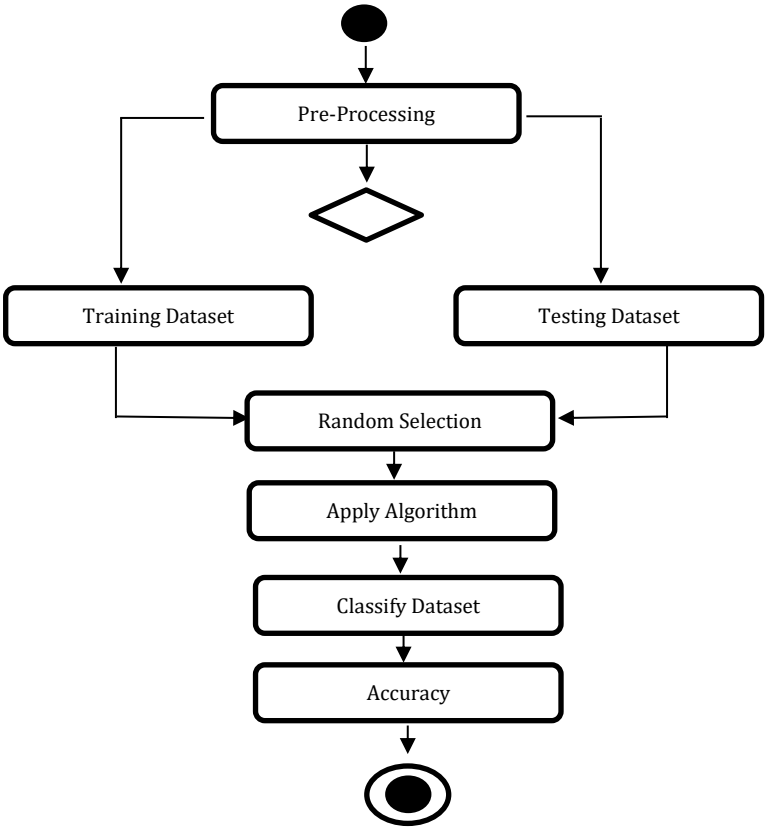

Figure 3. Activity Diagram

C. Use Case Diagram

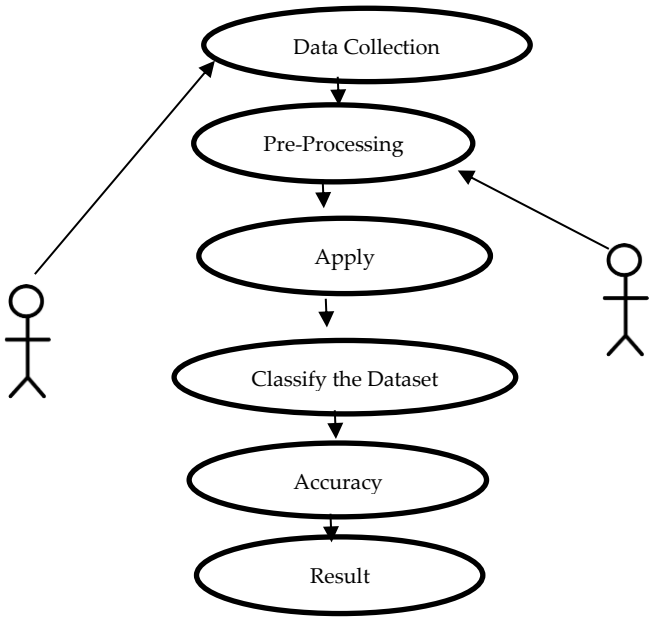

Figure 4. Use case Diagram

VI. System Implementation

A. Modules Description

The system defined as the four modules; they are.
i. Dataset Collection
ii. Data Pre-processing
iii. Implementation of Random Forest Algorithm
iv. Evaluation

i). Dataset Collection

You can obtain credit card transaction data from banks, payment processors, and credit card providers, among other sources. This dataset was derived from the Kaggle the dataset's original features, which comprise 29 numerical input variables, are not made public due to confidentiality issues. The result of PCA transformation is the dataset.

ii). Data Pre-Processing

Before training your model, you might need to preprocess the dataset by encoding categorical variables, adding missing values, and cleaning it up. Separate the dataset. One way to show the correlation between multiple parameters in a data frame visually is with a heatmap. When choosing characteristics or getting ready to prepare data, it may be useful to identify the variables that have the strongest relationships.

iii). Implementation of Random Forest Algorithm

Train the random forest model: The preprocessed training data is used to train the random forest model. Several decision trees are constructed to train the model using a subset of the features and training data. The goal of any decision tree training procedure is to lower the data's level of impurity, as measured by the Gini index.

A classification report is a performance evaluation metric in machine learning. It is used to show the accuracy, recall, score for F1, and support of your trained classification model. Accuracy is one metric used to evaluate the model's overall performance. It establishes the percentage of precise forecasts the model produced. Precision is the percentage of accurately predicted cases that are successful.

given as: True Positives/ (True Positives + False Positives)

The recall measures how many actual positive cases our model could correctly anticipate.

given as: True Positives/ (True Positives + Negatives)

iv). Implementation of Random Forest Algorithm

Assess using the preprocessed testing data, evaluate the Random Forest model's performance. Numerous performance indicators, of the true positive, false positive, true negative, and false negative rates of the model can also be obtained using a confusion matrix, which is a table that compares the expected and actual class labels. It's a simple method to determine the number of true positives, true negatives, false positives, and false negatives, as well as the effectiveness of the model. The evaluation provided support for two methods. Their names are Stratified Fold and Repeated Fold.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

In a nutshell, this study's primary goal was to determine which of the four machine learning techniques used would result in the most accurate model for detecting credit card fraud. This was achieved by building the models and determining the corresponding accuracy levels. In comparison to the XGBoost and LightGBM Model on the same dataset, which produced ROC-AUS scores of 0.88, Recall of 0.27 and F1-scores of 0.41 and ROC-AUS scores of 0.95, Recall of 0.83 and F1-scores of 0.45, the model performed better and demonstrated effective handling of imbalanced cases in credit card fraud, displaying a ROC-AUS score of 0.93, Recall of 0.85, and F1-score of 0.91.

B. Future Work

Artificial Intelligence and Machine Learning: By allowing fraud detection systems to pick up on and adjust to novel patterns of fraudulent conduct, the application of AI and ML approaches can improve fraud detection systems. While unsupervised learning can reveal patterns and behaviors that were previously unknown, deep

learning algorithms and neural networks can assist in identifying and categorizing fraudulent transaction patterns.

Advanced Data Visualization: Techniques for advanced data visualization, like network analysis and geographic mapping, can be used to see and show patterns of fraudulent behavior that conventional data analysis methods would miss.

Biometric Authentication: By using techniques like fingerprint scanning or facial recognition, biometric authentication can add another degree of protection and help stop illegal access to credit card accounts.

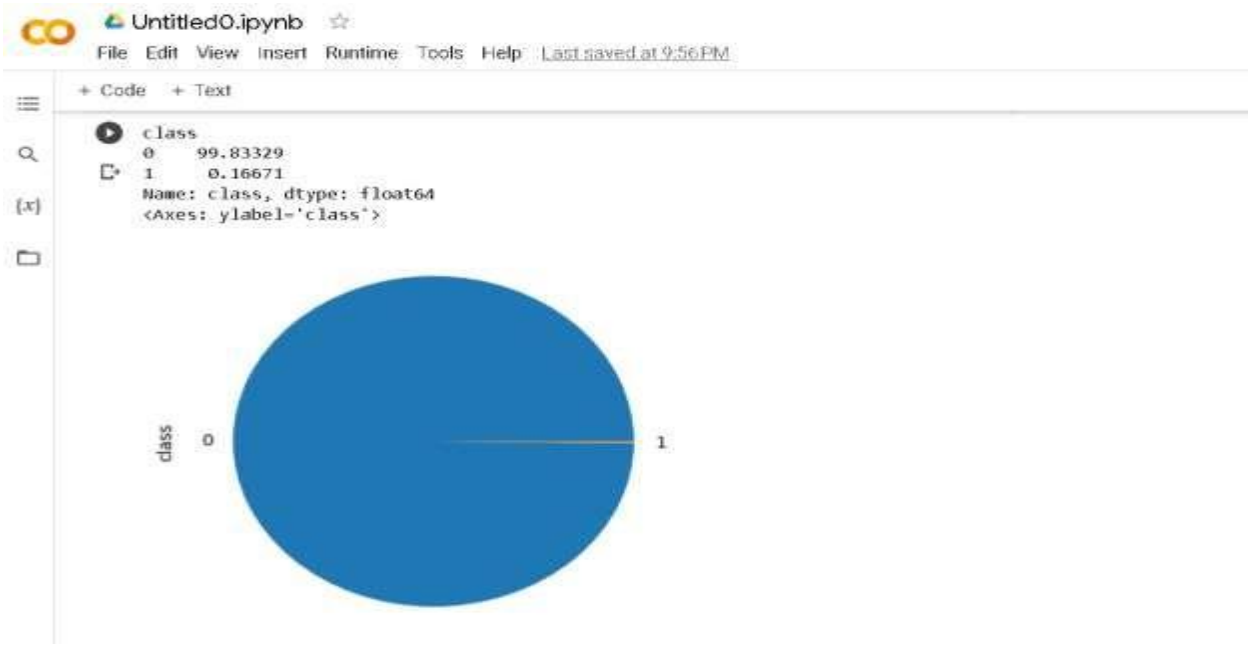

Figure 5. Checking the class distribution (0 and 1) of the target variable in percentage

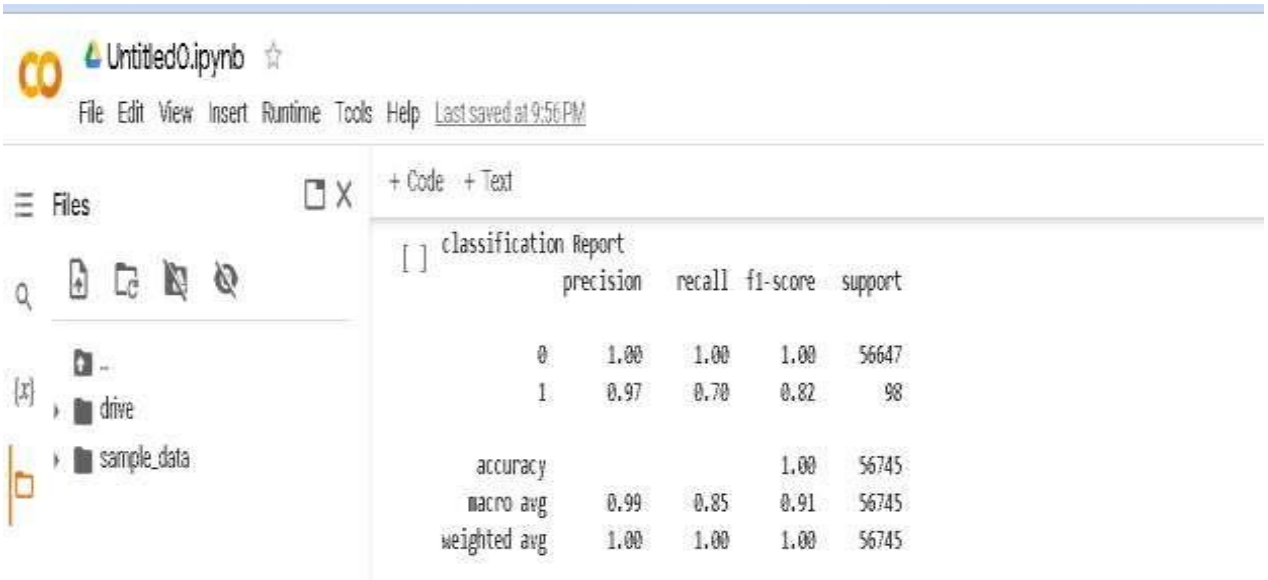

Figure 6. To show plot confusionmatrix in non-fraudalent and fraudulent

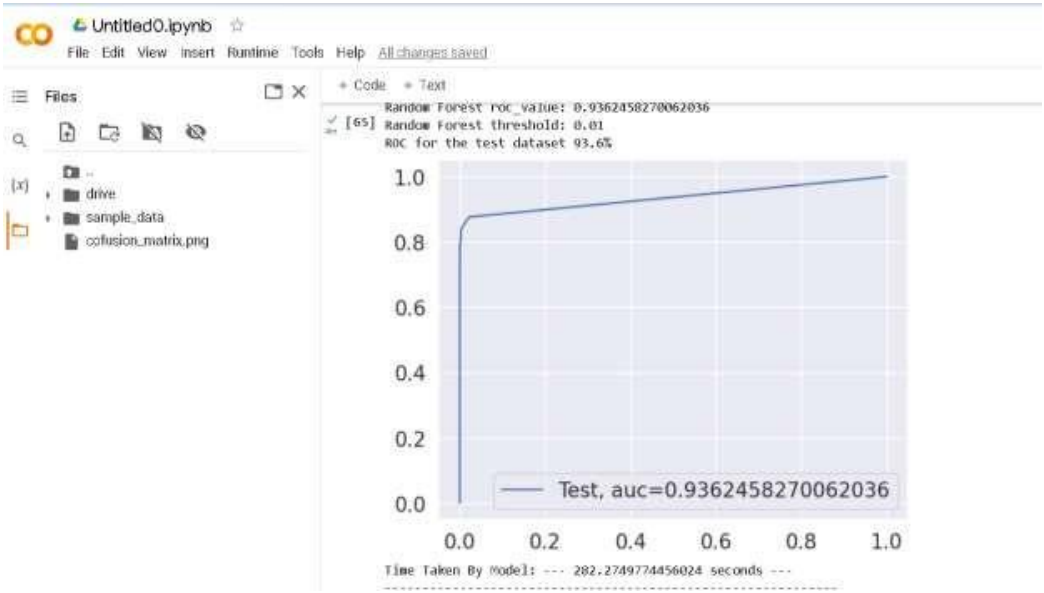

Figure 7. To calculate precision, recall and f1-score

VIII. REFERENCES

1. Esraa Faisal Malik et al., "Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture," *Mathematics*, vol. 10, no. 9, pp. 1-16, 2022. Google Scholar | Publisher Link
2. Pooja Tiwar et al., "Credit Card Fraud Detection Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 1264-1270, 2021. Google Scholar | Publisher Link
3. Harish Paruchuri, "Credit Card Fraud Detection Using Machine Learning: A Systematic Literature Review," *ABC Journal of Advanced Research*, vol. 6, no. 2, pp. 1-8, 2017. Google Scholar | Publisher Link
4. K.M. Anil Kumar et al., "Detection of False Income Level Claims Using Machine Learning," *International Journal of Modern Education and Computer Science*, vol. 14, no. 1, pp. 1-13, 2022. Google Scholar | Publisher Link
5. Yvan Lucas, and Johannes Jurgovsky, "Credit Card Fraud Detection Using Machine Learning: A Survey," *Arxiv*, pp. 1-31, 2020. Google Scholar | Publisher Link
6. Suresh K Shirgave et al., "A Review on Credit Card Fraud Detection Using Machine Learning," *International Journal of Scientific & Technology Research*, vol. 10, no. 8, pp. 1-4, 2019. Google Scholar | Publisher Link
7. Emmanuel Ileberi, Yanxia Sun, and Zenghui Wang, "A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection," *Journal of Big Data*, vol. 9, pp. 1-17, 2022. Google Scholar | Publisher Link
8. Meera AlEmad, "*Credit Card Fraud Detection Using Machine Learning*," Rochester Institute of Technology RIT Scholar Works, pp. 1-34, 2022. Google Scholar | Publisher Link
9. Abrar Hayat Nadim et al., "Analysis of Machine Learning Techniques for Credit Card Fraud Detection," *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Taipei, Taiwan, pp. 42-47, 2019. Google Scholar | Publisher Link
10. Salomi Hurriya Anjum, and Geeta Patil, "Cheat Detection for Credit Cards Using Artificial Intelligence," *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*, Vijaypur, India, pp. 1-6, 2022. Google Scholar | Publisher Link
11. Mohammed Azhan, Shazli Meraj, "Credit Card Fraud Detection using Machine Learning and Deep Learning Techniques," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, pp. 514-518, 2020. Google Scholar | Publisher Link
12. Shantanu Rajora et al., "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, pp. 1958-1963, 2018. Google Scholar | Publisher Link