

A

PROJECT REPORT ON

Credit Card Fraud Detection By Using ML and Python

Batch-March-August-24

PG-DBDA

UNDER THE GUIDANCE OF

Mrs. Priti Bhardwaj

**PRESENTED BY**

240320525007 Lokesh Singh

240320525011 Shubham Gupta

At

**CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING, NIODA**

**CERTIFICATE**

This is to certify that ‘Lokesh Singh’ and ‘Shubham Gupta’ of PG-DBDA has submitted project report entitled “Credit Card Fraud Detection Using Machine Learning And Python” in partial fulfilment for award of PG-DBDA from CDAC, NOIDA in session March-August-24.

Mr. Ravi Payal Mrs. Priti Bhardwaj

H.O.D Project Guide

**ACKNOWLEDGEMENT**

The project ”Credit Card Fraud Detection” was a great learning experience for us and we are submitting this work to Advanced Computing Training School (C-DAC, NOIDA).

We are very glad to mention the name of Mrs. Priti Bhardwaj for her valuable guidance to work on this project.

We are highly grateful to Mr. Ravi Payal, HOD, CDAC, Noida for her guidance and support whenever necessary during the course of our journey to acquire PG- Diploma in Big Data Analytics (PG-DBDA) through CDAC, NOIDA.

Our heartfelt thanks go to CDAC-Faculty ( PG-DBDA) who gave us all the required support and kind coordination to provide all the necessities to complete the project and throughout the course up to the last day of the course.

FROM:-

Lokesh Singh 240320525007

Shubham Gupta 240320525011

**ABSTRACT**

Credit card fraud is a significant problem, with billions of dollars lost each year. Machine learning can be used to detect credit card fraud by identifying patterns that are indicative of fraudulent transactions. Credit card fraud refers to the physical loss of a credit card or the loss of sensitive credit card information. Many machine- learning algorithms can be used for detection. This project proposes to develop a machine-learning model to detect credit card fraud. The model will be trained on a dataset of historical credit card transactions and evaluated on a holdout dataset of unseen transactions.

**INDEX**

Certificate……………………………………………………………..ii

Ackonwledgement…………………………………………………….iii

Abstract ……………………………………………………………….iv

1. Introduction ………………………………………………………....1

2. Data collection ……………………………………………………...2

3. Data Preprocessing …………………………………………………2

4. Describe Dataset and visualization ………………………………….3

4.1 Analysis ………………………………………………………...3

4.2 Imbalanced data ………………………………………………..5

4.3 Class Imbalance Handling in Machine Learning……………….5

5. Modelling Implementation ………………………………………….8

5.1 Train And Test Dataset ………………………………………...8

6. Model Evalution and Comparison …………………………………..9

6.1 Logistic Regression …………………………………………….9

6.2 Random Forest …………………………………………………12

6.3 KNN …………………………………………………………...16

6.4 SVM …………………………………………………………….17

7. Evaluation Result ……………………………………………………20

Conclusion ……………………………………………………………..21

Reference ………………...……………………………………………..22

**1.INTRODUCTION**

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future.In other words, Credit Card Fraud can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.

Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting.

This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated.This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course time.

Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent.

The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

**2.DATA COLLECTION**

The dataset was retrieved from an open-source website, Kaggle.com. It con- tains data on transactions made in 2013 by European credit card users in two days only. Thedataset consists of 31 attributes and 284,808 rows. Twenty-eight attributes are numeric variables that, due to the confidentiality and privacy of the customers, have been transformed using PCA transformation; the three remaining attributes are ”Time”, which contains the elapsed seconds between the first and other transactions of each Attribute, ”Amount” is the amount of each transaction, and the final attribute “Class” which contains binary variableswhere “1” is a case of fraudulent transaction, and “0” is not as case of fraudulent transaction.

Dataset : h[ttps://www.kaggle.com/datasets/mlg-ulb/creditcardfraud](http://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

**3.Data Preprocessing**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

**4.Describe dataset and data visualization**

The first figure below shows the structure of the dataset where all attributes are shown, with their type, in addition to a glimpse of the variables within each Attribute; as shown at the end of the figure, the Class type is integer, which I needed to change to factor and identify the 0 as Not Fraud and the one as Fraud to ease the process of creating the modeland obtain visualizations.The second figure shows the class distribution; the red bar, which contains 284,315variables, represents the non-fraudulent transactions, and the blue bar, with 492 variables, represents the fraudulent transactions***.***

**4.1Analysis**

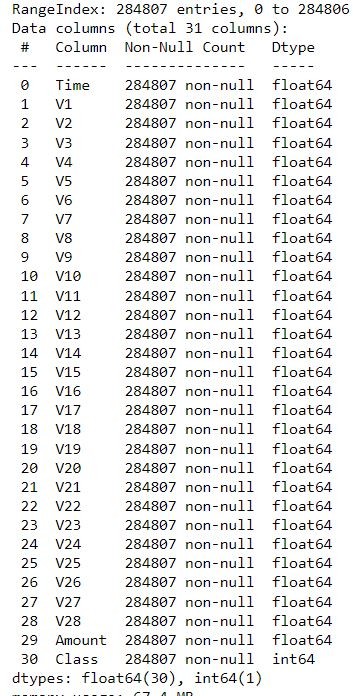


Fig-4.1. Data Structure

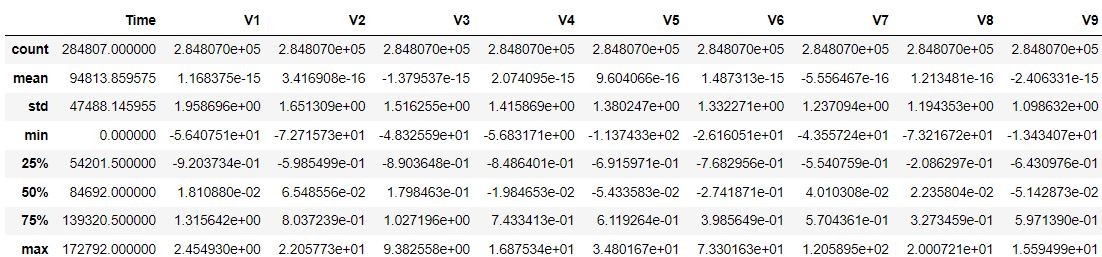
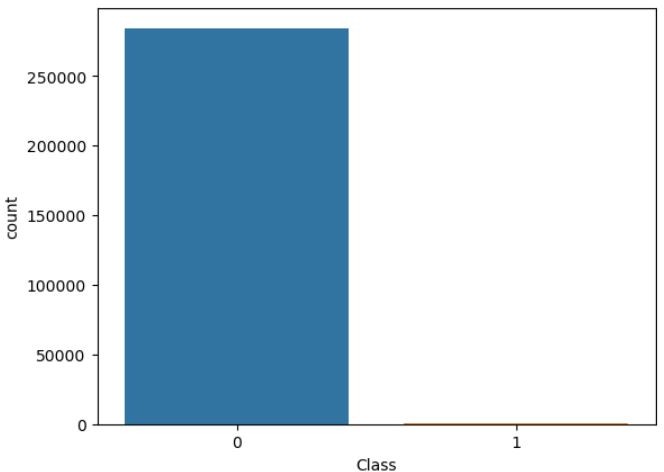
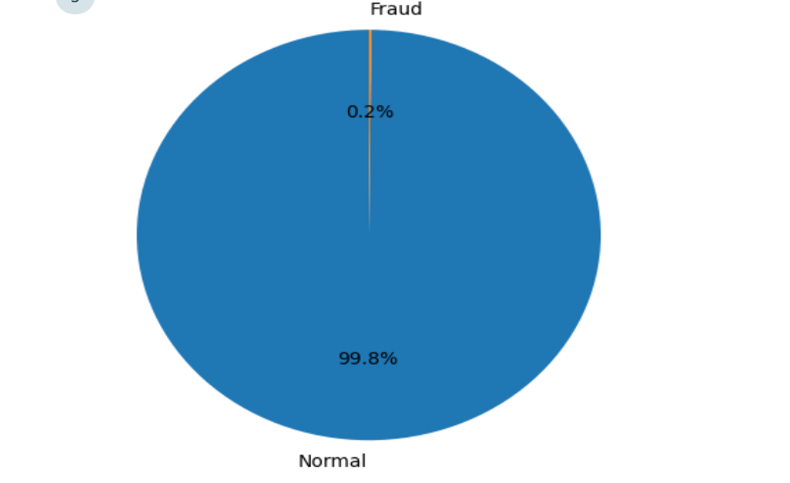


Fig-4.2. Data Description



Fig-4.3. Class Count

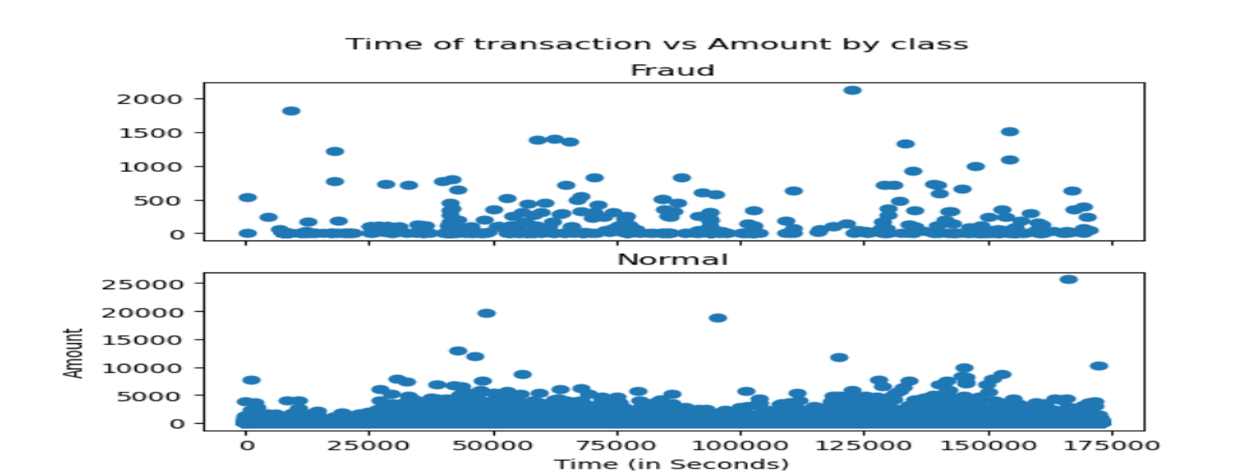
 Fig- 4.4. Pie Chart

Fig-4.5. Scatter plot

**4.2. Imbalanced Data:**

As you see we got the imbalanced data in the above slides now lets know what is imbalanced data and how to tackle it

* Datasets where the distribution of observations in the target class is uneven. In other words, one class label has a significantly higher number of observations, while the other has a notably lower count.

**4.2.1. Problem with Handling Imbalanced Data for Classification**

* Algorithms may get biased towards the majority class and thus tend to predict output as the majority class.
* Minority class observations look like noise to the model and are ignored by the model.
* Imbalanced dataset gives misleading accuracy score.

**4.3. Class Imbalance Handling in Machine Learning:**

Using Random Under-Sampling

* When observations from the majority class are eliminated until the majority and minority classes are balanced, this is known as undersampling.
* Undersampling has advantages when working with large datasets, especially ones with millions of rows, but there is a risk that important information will be lost during the removal process.

Data Sampling

Build a sample dataset containg similar distribution of normal transction and fraudlent Transactions

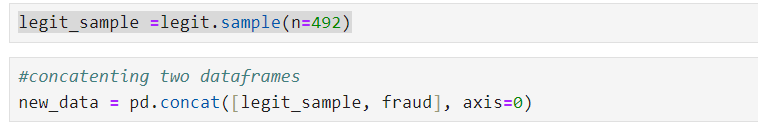
Number of fraudulent transactions = 492 & Legit Transactions = 284315

Fig- 4.6. Balanced Dataset

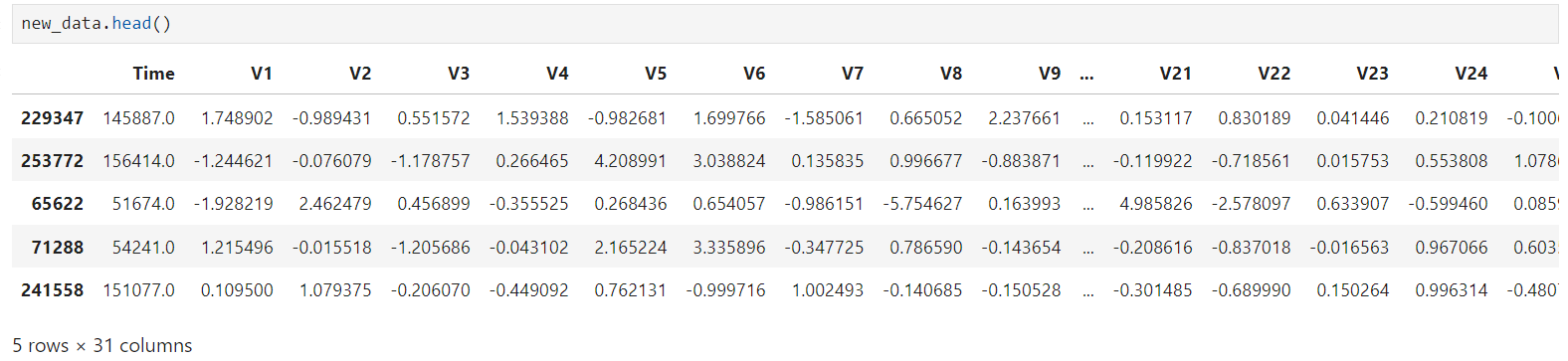


Fig- 4.7. Dataset

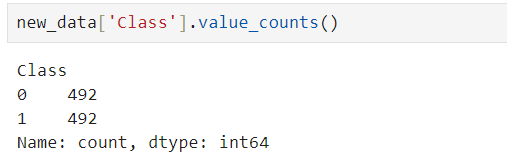
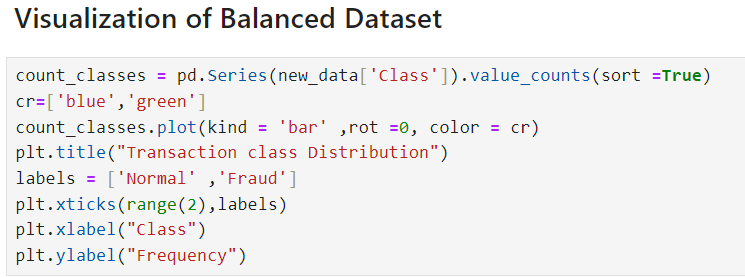


Fig- 4.7. Balance class



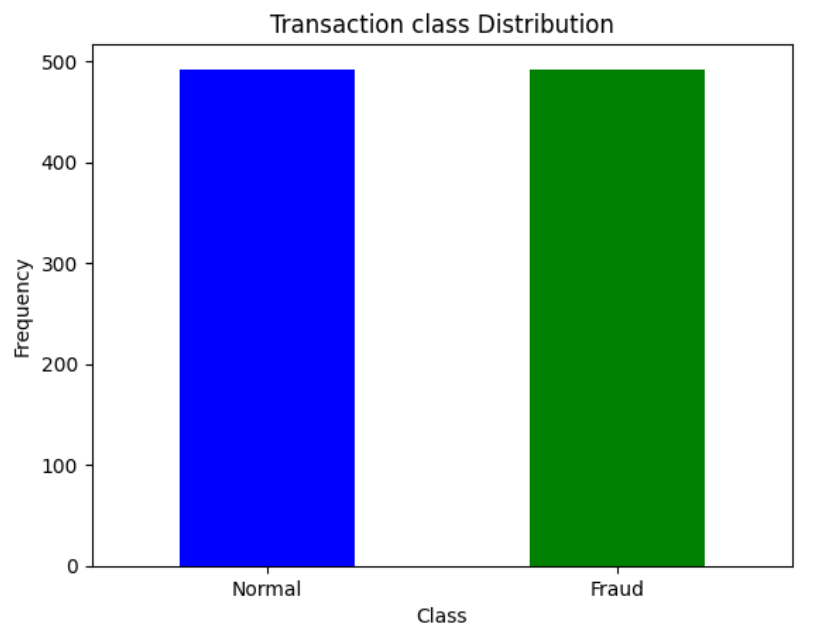


Fig-4.8. Visualization of Balance class

**5.Modelling implementation**

When implementing a model, start simple. Most of the work in ML is on the data side, so getting a full pipeline running for a complex model is harder than iterating on the model itself. After setting up your data pipeline and implementing a simple model that uses a few features, you can iterate on creating a better model.

Simple models provide a good baseline, even if you don't end up launching them. In fact, using a simple model is probably better than you think. Starting simple helps you determine whether or not a complex model is even justified.

**5.1 Train And Test Dataset**

**5.1.1.TRAIN DATASET:**

* The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.
* The training data varies depending on whether we are using Supervised Learning or Unsupervised Learning Algorithms.

**5.1.2. TEST DATASET:**

* Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset.
* The test dataset is another subset of original data, which is independent of the training dataset.
* Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

Fig-5.1. import train\_test\_split

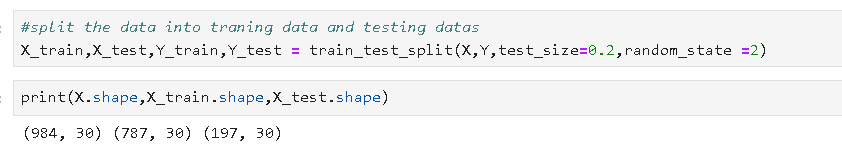


Fig-5.2. Train and Test dataset

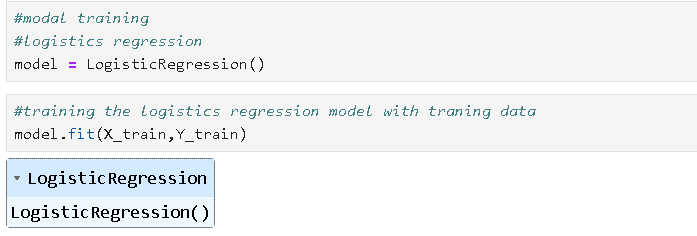
**6.Model evaluation and comparison**

**6.1.Logistic Regression**-

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.



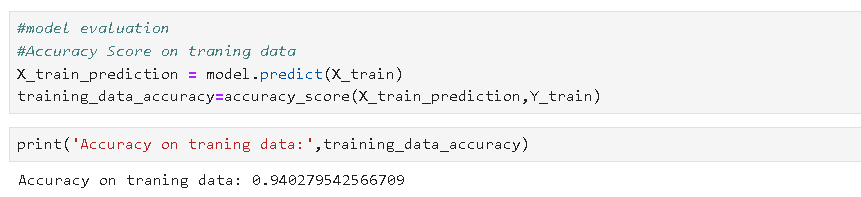
Fig-6.1.1. Model Logistics Regression

Fig-6.1.2. Model Evaluation & Accuracy On Traning Data

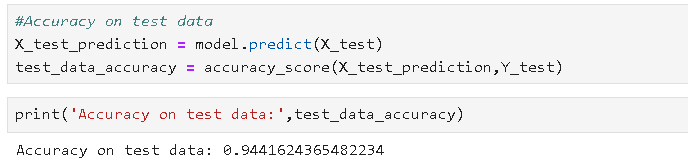


Fig-6.1.3. Model Evaluation & Accuracy On TestData

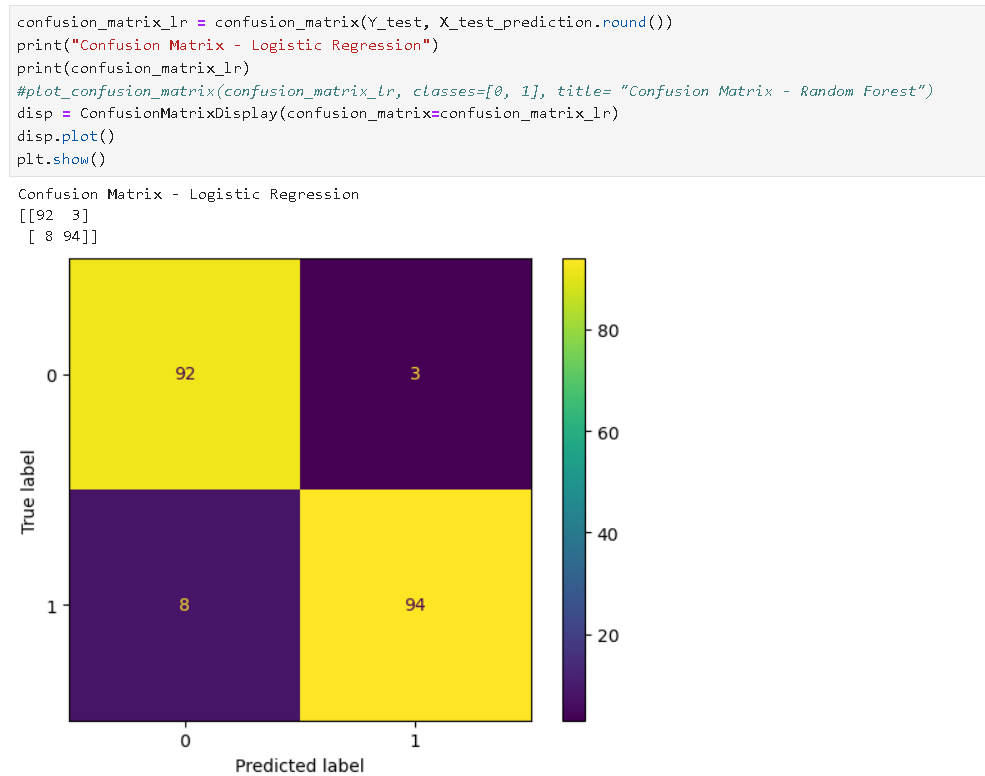


Fig- 6.1.4. Confusion Matrix for Logistic Regression

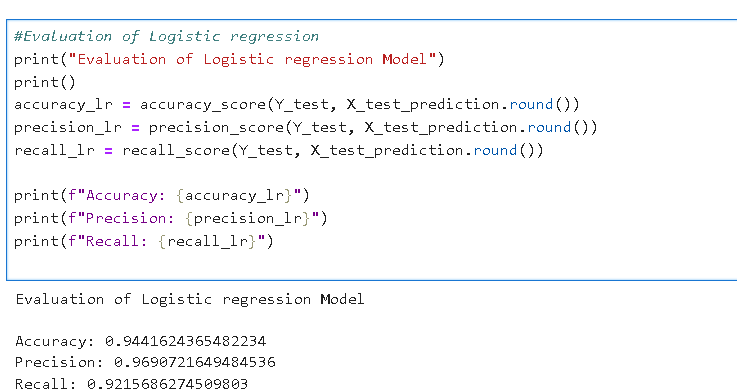


Fig-6.1.5. Model Evaluation of Logistic Regression Model

# 6.2.Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

# IMG_256

Fig-6.2.1. Random Forest

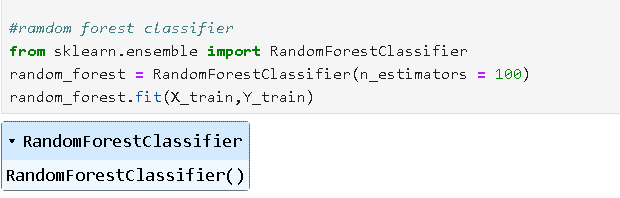


Fig-6.2.2. Random Forest Classifier

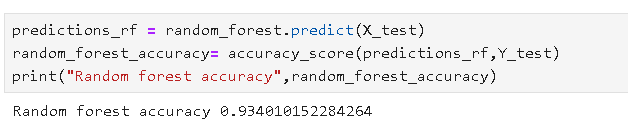
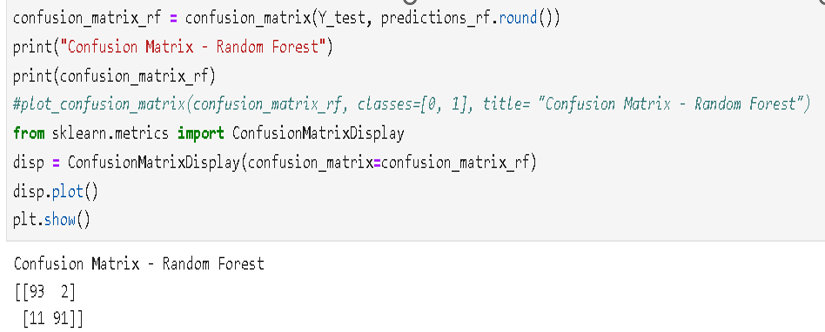
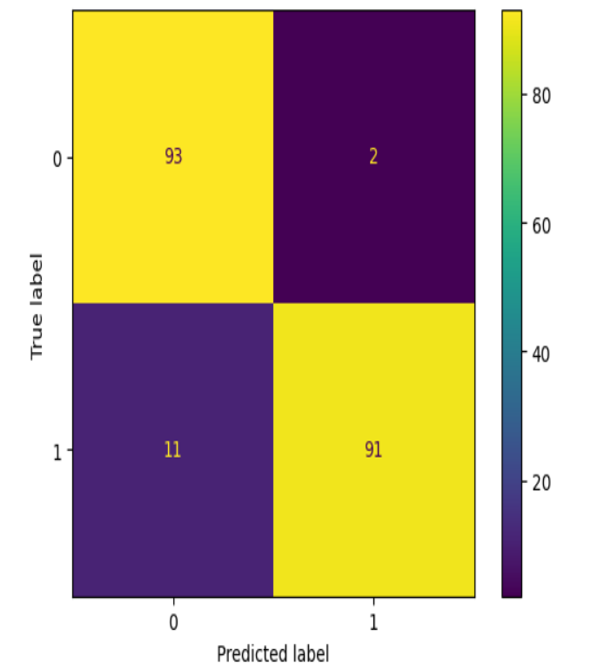


Fig-6.2.3. Random Forest Accuracy

****

****

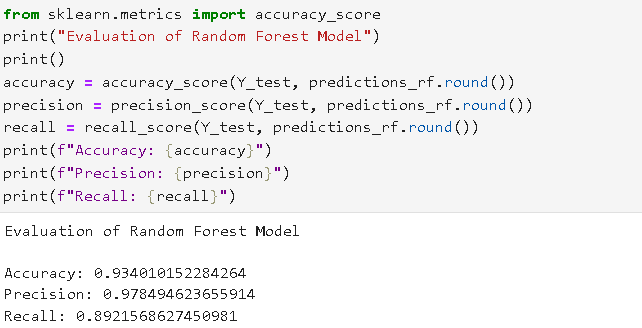
Fig-6.2.4. Confusion Matrix for Random Forest

Fig-6.2.5. Model Evaluation of Random Forest Model

**6.3. KNN Algorithm**

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

The KNN algorithm can be used for either regression or classification problems, it is typically used as a classification algorithm.

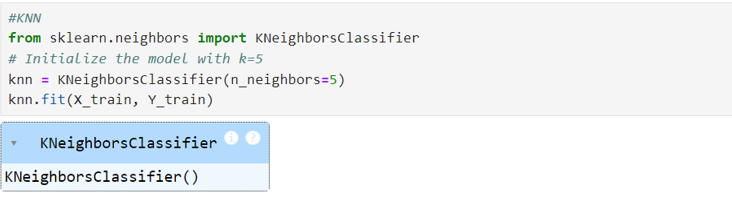
****

Fig-6.3.1 KNN

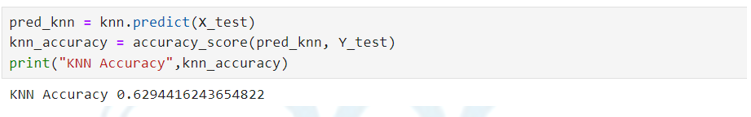
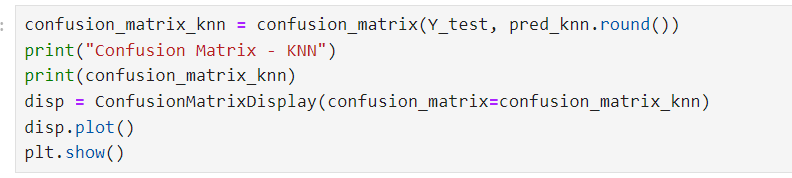
****

Fig-6.3.2 Accuracy of KNN

****

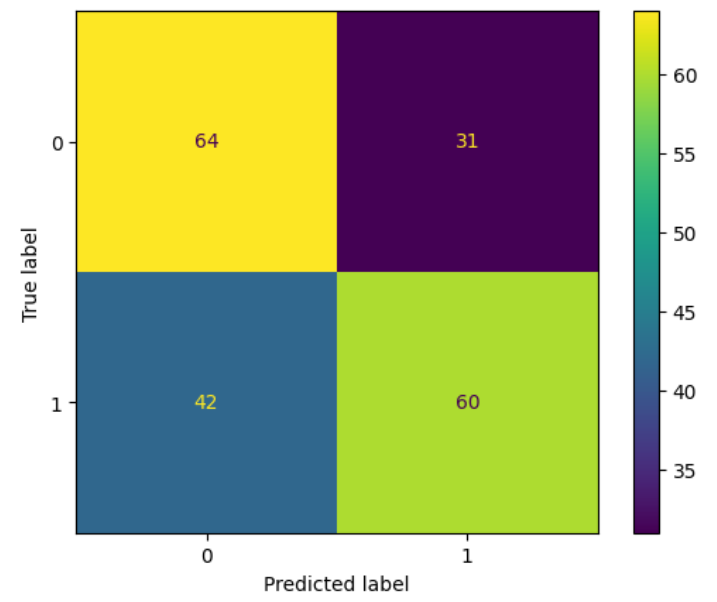
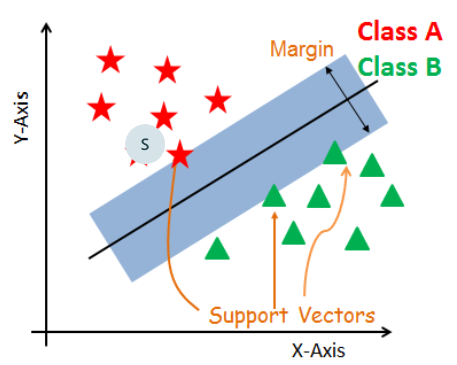
****

Fig-6.3.3. Confusion Matrix for KNN

**6.4. SVM**

A support vector machine (SVM) is a  algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

SVMs are commonly used within classification problems. They distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closest data points of opposite classes.

****

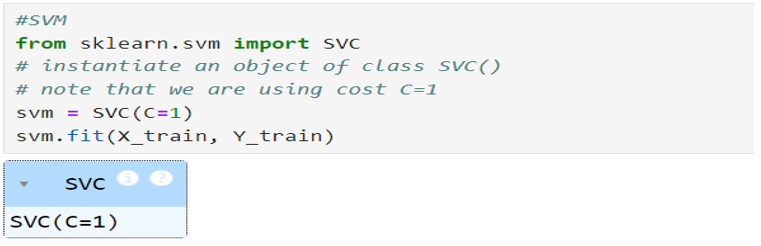
**** Fig- 6.4.1. SVM

Fig-6.4.2. SVM Model

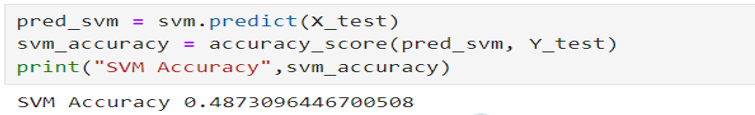
****

Fig-6.4.3. Accuracy of SVM

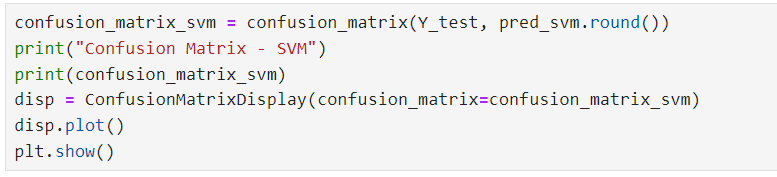
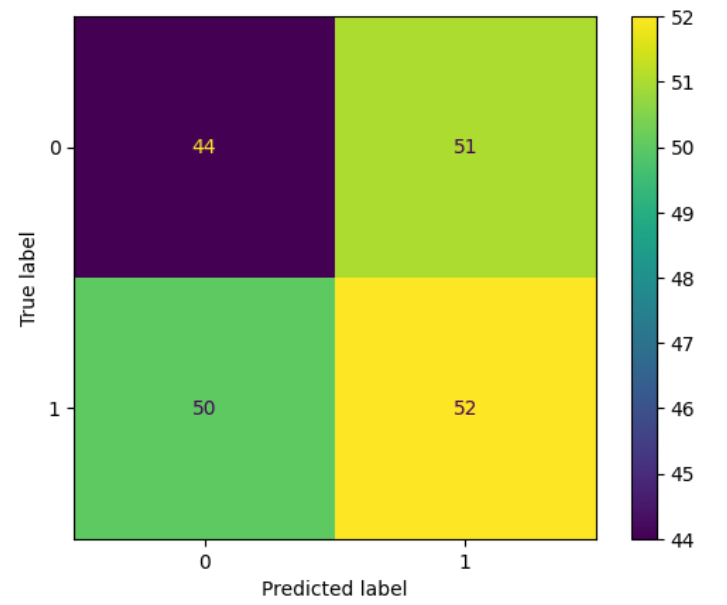
****

Fig-6.4.4.Confusion Matrix for SVM

**7.Evaluation Result**

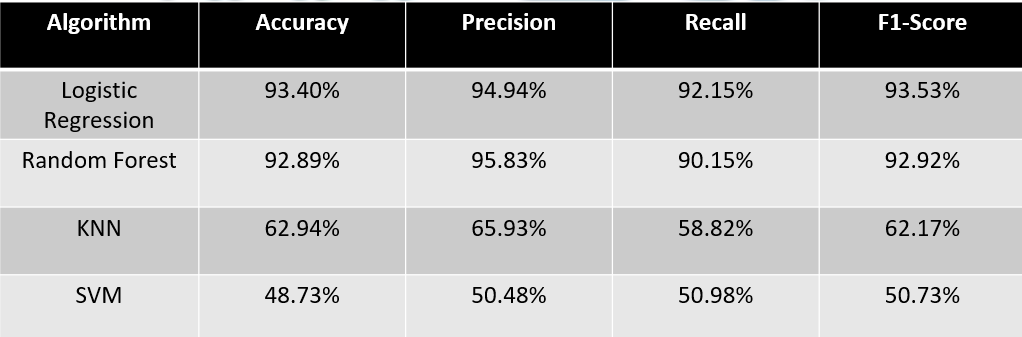
****

Fig-7.Table of Evaluation result

**Conclusion**

In conclusion, the main objective of this project was to find the most suited model for credit card fraud detection in terms of the machine learning techniques chosen for the project. It was met by building the two models and finding the accuracies of them all the best model in terms of accuracies is Logistic Regression, which scored 93.4% the amount of credit card fraud and increase the customer’s satisfaction as it will provide them with a better experience and feeling secure.

**References**

[1] Z. et al, “Analysis on credit card fraud detection techniques: Based on certain

design criteria.” https://research.ijcaonline.org/volume52/number3/pxc3881538.pdf, 2012. Accessed: 26-oct-2023.

[2] . A. N. O. Alenzi, H. Z., “Fraud detection in credit cards using logistic regression.”

https://thesai.org/Publications/ViewPaper?Volume=11&Issue=12&Code=IJACSA&SerialNo=65, 2020. Accessed: 26-oct-2023.

[3] S. A. A. S. . S. S. D. Maniraj, S. P., “Credit card fraud detection using machine learning and data science.” https://doi.org/10.17577/ijertv8is090031,2019. Accessed: 25-oct-2023.

[4] . D. R. Dheepa, V., “Analysis on credit card fraud detection techniques: Based on certain design behaviour-based credit card fraud detection using support vector machines criteria.” https://doi.org/10.21917/ijsc.2012.0061, 2012. Accessed:26-oct-2023.

[5] . P. M. Malini, N., “Analysis of credit card fraud identification techniques based on

knn and outlier detection.” https://doi.org/10.1109/aeeicb.2017.7972424, 2017. Accessed: 26-oct-2023.

[6] T. K. V. B. . M. B. Maes, S., “Credit card fraud detection using bayesian and neural networks.” https://www.ijert.org/research/ credit-card-fraud-detection-using-machine-learning-and-data-science-IJERTV8IS090031.pdf, 2002. Accessed: 23-oct-2023.

[7] N. S. . J. S. Jain, Y., “A comparative analysis of various credit card fraud detection

techniques,” 2019. Accessed: 25-oct-2023.

[8] P. S. . K. S. Dighe, D., “Detection of credit card fraud transactions using machine

learning algorithms and neural networks.” https://doi.org/10.1109/iccubea. 2018.8697799, 2018. Accessed: 26-oct-2023