CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

BY

MALATHI K

(17PCS014)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Education for Women, (Deemed to be University),

Coimbatore-641043

April 2019

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

BY

MALATHI K

(17PCS014)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Education for Women, (Deemed to be University),

Coimbatore-641043

April 2019

Signature of the Head of the Department **Signature of the Supervisor**

Viva Voce Examination Held on _____

Signature of the Examiners

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere thanks to **God Almighty**, for his constant love and grace that he has showered upon me.

I am very grateful to **Shri, Dr.P.R.KrishnaKumar, Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for his support and encouragement during the course of my project.

I heartily thank **Dr. (Mrs.) Premavathy Vijayan, M.Sc., M.Ed., Dip.Spl.Edn., M.Phil, Ph.D., Vice Chancellor** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing the facilities to do the project.

I express my humble gratitude to **Dr. (Mrs.) S Kowsalya, M.Sc., M.Phil, Ph.D., Registrar,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all the facilities necessary for the project.

I am also thankful to **Dr. (Mrs.) K.Udaya Chandrika M.Sc., M.Phil., Ph.D., Dean,** School of Physical Sciences & Computational Sciences of our university, for granting the facility required.

I wish to place on record my deep sense of gratitude to **Dr. (Mrs.) V.Radha, M.Sc., PGDOR, PGDCA, B.Ed., M.Phil., Ph.D., Professor** and **Head, Department of Computer Science,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all the facilities to complete the project.

I express my honourable thanks to my project coordinator **Dr. (Mrs.) G.Padmavathi, M.Sc., M.Phil., Ph.D., Professor, Department of Computer Science,** for her kind advice and knowledgeable suggestions which helped me to complete my project successfully.

I owe great deal of gratitude to my esteemed guide **Dr. (Mrs.) V.Radha, M.Sc., PGDOR, PGDCA, B.Ed., M.Phil., Ph.D., Professor** and **Head, Department of Computer Science,** for imparting the tremendous assistance and well-timed support for triumph of my project.

Finally, I take pride to thank my parents and those who helped me directly or indirectly for carrying out this work.

ABSTRACT

ABSTRACT

The research work entitled as **"Credit Card Fraud Detection Using Machine Learning Techniques"**. The machine learning techniques are used to detect the suspicious transactions made by fraudsters with unauthorized credit card and the performance evaluation is done.

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations and finance industry. Due to rapid advancement in the electronic commerce technology, the use of credit cards has dramatically increased and it caused an explosion in the credit card fraud. As credit card becomes the most popular mode for payment for both online and regular purchase.

In the transaction of credit card, the fraudulent transactions are generated in new ways by the fraudsters. Machine learning algorithms such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression are used to detect the fraudulent transactions and the performance of the techniques are evaluated based on the accuracy, precision, recall or sensitivity, area under curve and receiver operating characteristic.

Performance of the machine learning techniques are calculated to evaluate which machine learning algorithm detects the suspicious transactions accurately when compared to other algorithm. Performances are evaluated based on the accuracy, precision, recall or sensitivity, area under curve and receiver operating characteristic.

CONTENTS

TABLE OF CONTENT

INTRODUCTION

1. INTRODUCTION

Machine learning is an application of Artificial Intelligence (AI) that provides the system to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. These programs or algorithms are designed in a way that they learn and improve over time when exposed to new data. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithm is trained using a training data set to create a model. When new input data is introduced to the machine learning algorithm, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the machine learning algorithm is deployed. If the accuracy is not acceptable, the machine learning algorithm is trained again and again with an augmented training data set.

Machine learning algorithms are often categorized as

- Supervised Learning
- Unsupervised Learning
- Reinforcement learning

Supervised learning is the data mining task of inferring a function from labeled training data. The training data consist of a set of training examples and each example is a pair consisting of an input object and a desired output value. Supervised machine learning techniques attempt to find out the relationship between input attributes (independent variable) and a target attribute (dependent variable). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The goal of unsupervised learning algorithm is to determine the hidden patterns or grouping in data from unlabeled data. It is mostly used in exploratory data analysis. One of the defining characters of unsupervised learning is that both input and output are not known. Unsupervised learning algorithms can perform more complex processing tasks than supervised learning systems.

Reinforcement learning is a type of dynamic programming that trains algorithms using a system of reward and punishment. A reinforcement learning algorithm learns by interacting with its environment. The algorithm learns without intervention from a human by maximizing its reward and minimizing its penalty. Reinforcement learning is used in operations research, information theory, game theory, control theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. The process involved in machine learning techniques is shown in the below Figure 1.1.

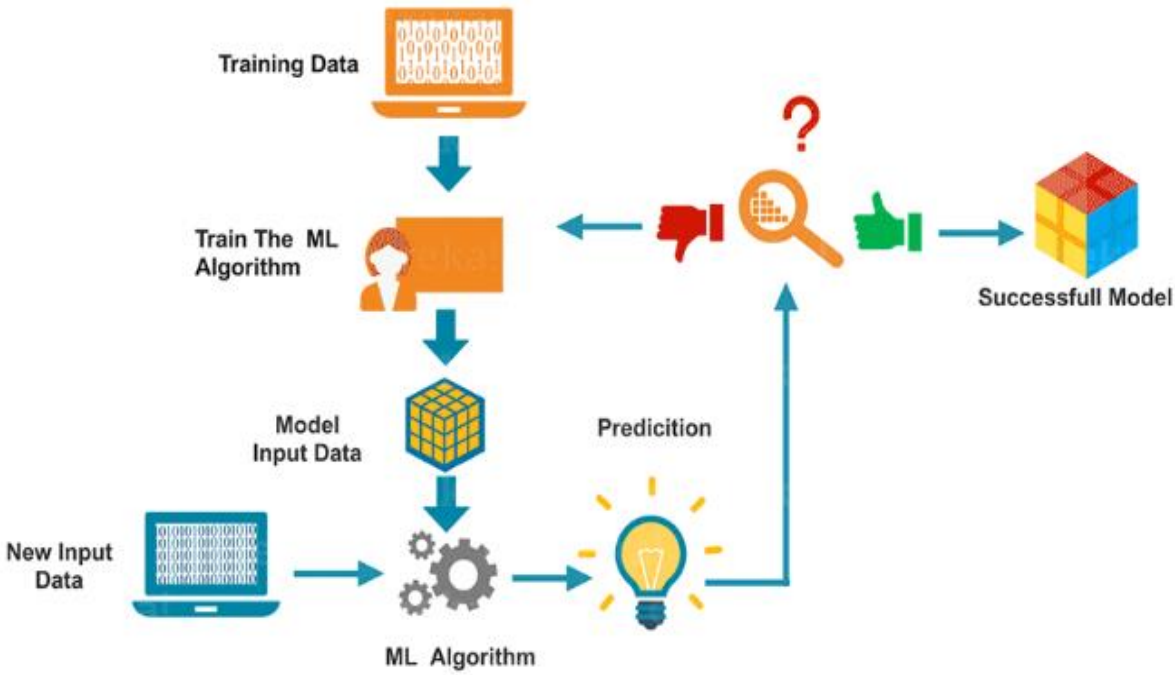


FIGURE 1.1 Machine Learning Process

SUPERVISED MACHINE LEARNING

Supervised learning is a method used to enable machines to classify objects, problems or situations based on related data fed into the machines. It adapts the model to reproduce outputs known from a training set. **Supervised learning algorithm** analyzes the training data and produces an inferred function. The goal of supervised learning algorithm is to predict Y as accurately as possible when given new examples X.

Supervised learning can be carried out using two categories

- Classification
- Regression

Classification techniques are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. It is the process of predicting the class of given data points. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (Y). The main goal of a classification problem is to identify the class. A classification technique includes K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Logistic Regression and Neural Networks.

Regression is a statistical technique to determine the linear relationship between two or more variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (predictors).

Credit card fraud is an unauthorized usage of card, unusual transaction behavior, or transactions on an inactive card. The Credit card usage is increasing day by day for both regular purchases as well as online and the fraudulent transactions are generated in new ways by the fraudsters. The supervised learning algorithms such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression are used for detecting the fraudulent transactions.

1.1 PROBLEM DEFINITION

The problem is to detect the suspicious transactions using machine learning techniques like K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression. The performance of the classification techniques is evaluated based on accuracy, sensitivity or recall, precision, area under curve and receiver operating characteristic.

SYSTEM STUDY

2. SYSTEM STUDY

2.1 LITERATURE REVIEW

In literature most of the works have been concentrated on detecting the credit card fraud transactions using supervised machine learning techniques.

1) Maes 2002 tried Artificial Neural Networks (ANN) and Bayesian Belief Networks (BBN) on a real dataset obtained from Europay International. Their experiment showed that the Bayesian Belief networks outperforms ANN in terms of classification accuracy and training time.
2) Sahin & Duman, 2011 applied decision trees and support vector machines (SVM) on a dataset obtained from a real world national bank's credit card data warehouses. They found out that decision trees outperform SVM in solving the problem.
3) Haung, 2013 developed two models based on logistic regression and SVM. IT was found that logistic regression outperforms SVM.
4) Ehramikar, 2000 developed a fraud detection model based on the decision trees and he founded that decision tress suffer from under fitting problem in case of imbalanced data set (case of fraud detection dataset).

2.2 ABOUT THE SOFTWARE

In this research work, Anaconda Distribution (Spyder) is used as a front end and Microsoft Excel as a back end.

2.2.1 PYTHON PROGRAMMING

Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is a powerful multi-purpose programming language and was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. Python is extremely attractive in the field of Rapid Application Development because it offers high-level built in data structures, combined with dynamic typing and dynamic binding options.

Python is simple, easy to learn since it requires a unique syntax focuses on readability and reduces the cost of program maintenance and development. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and

the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. It has a wide range of applications from web development (Django and Bottle), scientific and mathematical computing.

The Features of Python Programming

1) **Simple and easy to learn**

   Python has few keywords, simple structure and a clearly defined syntax.

2) **Easy to read**

   Python code is more clearly defined and visible to the eyes.

3) **Easy to maintain**

   Python's source code is fairly easy-to-maintain.

4) **Interactive Mode**

   Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

5) **Portable**

   Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

6) **GUI Programming**

   Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems such as Windows MFC, Macintosh and X Window system of UNIX.

7) **Scalable**

   Python provides a better structure, support for large programs than shell scripting.

8) **Object-Oriented Language**

   Python supports object oriented language and concepts of classes and objects.

9) **Large Standard Library**

Python has a large and broad library and provides rich set of module and functions for Rapid Application Development.

Apart from the above-mentioned features, python has a big list of good features, few are listed below

- It supports functional and structure programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA and Java.

Benefits of Python

- Extensive Support Libraries
- Integration Feature
- Open source and community development
- Easy to learn
- User-friendly data structures
- Presence of third-party modules
- Improved Programmer's Productivity
- Embeddable
- IOT Opportunities

Applications of Python

- GUI Based Desktop Applications
- Web Frameworks and Applications
- Enterprise and Business Applications
- Operating Systems
- Language Development

2.2.2 FRONT END: SPYDER

Spyder, the scientific python development environment is a free Integrated Development Environment (IDE) included with anaconda. The name Spyder derives from "Scientific Python Development Environment" (SPYDER). It is a powerful scientific environment written in Python. Spyder features a unique combination of the advanced editing, analysis, debugging and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection and beautiful visualization capabilities of a scientific package. Spyder offers built-in integration with many popular scientific packages, including NumPy, SciPy, Pandas, IPython, QtConsole, Matplotlib, and SymPy. The Spyder was created and developed by Pierre Raybaut in 2009, since 2012 Spyder has been maintained and continuously improved by a team of scientific python developers and the community. Spyder can also be used as a PyQt5 extension library, allowing developers to build upon its functionality and embed its component, such s the interactive console in their own PyQt software.

Core building blocks of powerful IDE

- Editor

  Work efficiently in a multi-language editor with a function/class browser. Code analysis tools, automatic code completion, horizontal/vertical splitting.

- IPython Console

  Harness the power of as many IPython consoles within the flexibility of a full GUI Interface.

- Variable Explorer

  Interact and modify variables on the fly. Plot a histogram or time series. Edit a data frame or NumPy array.

- Profiler

  Find and eliminate bottlenecks to unchain code's performance.

- Debugger

  Trace each step of code's execution interactively.

2.2.3 BACKEND: EXCEL

Microsoft Excel is a software program produced by Microsoft Corp that allows users to organize, format and calculate data with formulas using a spreadsheet system. This software is a part of the Microsoft office suite and is compatible with other applications in the office suite. Excel is a commercial spreadsheet application produced and distributed by Microsoft for Microsoft Windows and Mac OS X. It features the ability to perform basic calculations, use graphing tools, create pivot tables and create macro programming language. Excel has the same basic features as every spreadsheet, which use a collection of cells arranged into rows and columns to organize data manipulation. They also display data as charts, histograms and line graphs.

The Features of Microsoft Excel are

- ✓ Multi-Threading Recalculation (MTR) for commonly used functions
- ✓ Improved pivot tables
- ✓ More conditional formatting options
- ✓ Additional image editing capabilities
- ✓ In-cell charts called spark lines
- ✓ Ability to preview before pasting
- ✓ Office 2010 backstage feature for document-related tasks
- ✓ Ability to customize the Ribbon
- ✓ Many new formulas, most highly specialized to improve accuracy

Advantages of MS-Excel

- ✓ Analyzing and storing data
- ✓ Excel tools make your work easier
- ✓ Data recovery and spreadsheet
- ✓ Mathematical formulas of MS Excel make things easier
- ✓ Helps businessmen in developing future strategy

Disadvantages of MS-Excel

- ✓ Excel is vulnerable to change
- ✓ Excel is difficult to troubleshoot or test
- ✓ Excel is obstructive to regulatory compliance

METHODOLOGY

3. METHODOLOGY

3.1 OVERVIEW OF THE RESEARCH WORK

A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictors for the response to new data. It uses classification and regression techniques to develop predictive models. Classification can be performed on structured or unstructured data. A classification technique includes K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Logistic Regression and Neural Networks. The supervised learning algorithms such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression are used to detect the fraudulent transactions and the performance evaluation is done.

SUPPORT VECTOR MACHINE

Support vector machine is a supervised machine learning algorithm used for both classification and regression. Support vector machine is a discriminative classifier formally defined by a separating hyper plane. Support vectors are simply the co-ordinates of individual observation. Support vector machine best segregates the two classes.

A hyper plane is a line that splits the input variable space. It is selected to best separate the points in the input variable space by their class, either 0 or 1. The algorithm plot each data item as a point in n-dimensional space where n is the number of features and perform classification by finding the hyper-plane that differentiate the two classes.

$$B0 + (B1 * X1) + (B2 * X2) = 0 \qquad \text{……. (1)}$$

From the above equation (1) the coefficients (B1 and B2) determine the slope of the line and the intercept (B0) are found by the learning algorithm, and X1 and X2 are the two input variables.

K-NEAREST NEIGHBOR

K-Nearest Neighbors (KNN) is one of the simplest algorithm used in Machine Learning for regression and classification problem. KNN algorithms use a data and classify new data points based on a similarity measures (e.g. distance function). Classification is done

by a majority vote to its neighbors. The data is assigned to the class which has the most nearest neighbors. The output can be calculated as the class with the highest frequency from the K-most similar instances.

Suppose the value of K is 3. The KNN algorithm starts by calculating the distance of point X from all the points. It then finds the 3 nearest points with least distance to point X. The final step of the KNN algorithm is to assign new data points to the class to which majority of the nearest data points belong.

NAÏVE BAYES

Naïve Bayes classifier is the supervised machine learning algorithm that uses the Bayes Theorem. It is a binary and multi-class classification problem based on Bayes' Theorem with an assumption of independence among predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is easy to build and useful for very large datasets. Naïve Bayes is known to outperform even highly sophisticated classification methods.

Bayes Theorem works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, calculate the probability of an event using its prior knowledge.

Bayes theorem provides a way of calculating posterior probability P (c|x) from P(c), P(x) and P (x|c).

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)} \qquad \ldots\ldots\ldots\ldots (2)$$

LOGISTIC REGRESSION

Logistic regression is one of the most popular machine learning algorithms for binary classification. It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.

3.2 METHODOLOGY DIAGRAM

The overall methodology is presented in the below Figure 3.1.

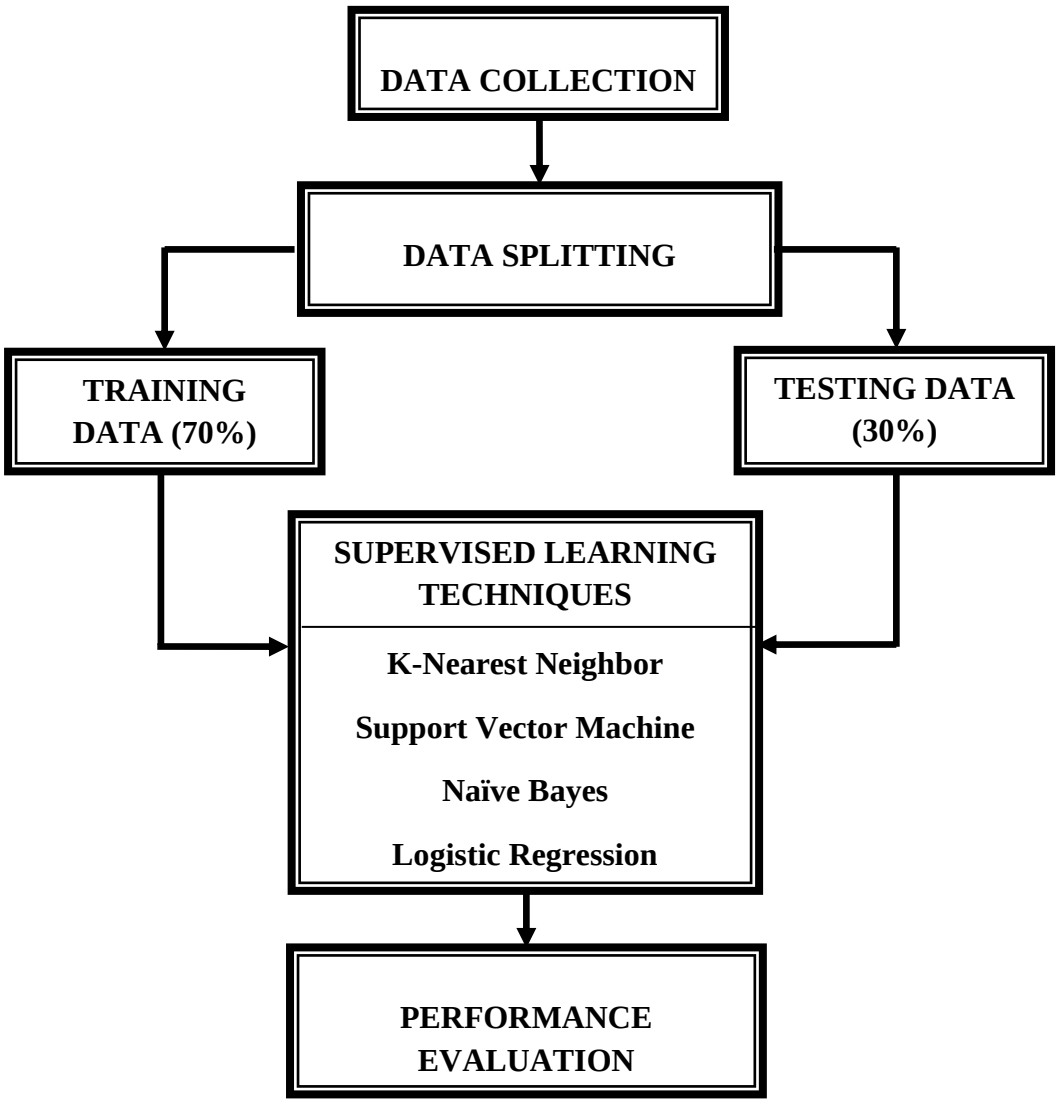


FIGURE 3.1 OVERALL METHODOLOGY

3.3 DATASET DESCRIPTION

For this work credit card fraud detection dataset is collected from Kaggle, contains transactions made by credit cards in September 2013 by European cardholders that occurred in two days and have 284,807 transactions.

Credit card fraud detection dataset is highly unbalanced. It consists of 284,807 instances and 31 attributes. It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, V3,…..V28 are the principle components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transactions. Feature 'Amount' is the transaction amount, this feature can be used for example-dependent-cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. In this research work the dataset is divided into training data (70%) and testing data (30%).

Attributes

- Time
- Principal Component Analysis (PCA) V1 – V28
- Amount
- Class

3.4 MODULES

This project is designed with four major modules. They are

- **DATA COLLECTION**
- **DATA SPLITTING (Training and Testing Data)**
- **SUPERVISED MACHINE LEARNING TECHNIQUES**
  - K-Nearest Neighbor
  - Naïve Bayes
  - Support Vector Machine
  - Logistic Regression
- **PERFORMANCE EVALUATION**

3.5 MODULE DESCRIPTION

3.5.1 DATA COLLECTION

The data is collected from Kaggle website and it contains transactions made by credit cards in September 2013 by European cardholders that occurred in two days and have 284,807 transactions.

The dataset is skewed and highly unbalanced. It contains only numerical input variables which are the result of a PCA transformation. The dataset is divided into training data (70%) and testing data (30%). Features V1, V2, V3, V4, V5, V6, V7, V8, V9,…..V28 are the principle components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transactions. Feature 'Amount' is the transaction amount, this feature can be used for example-dependent-cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud transaction and 0 otherwise.

3.5.2 DATA SPLITTING

Machine learning algorithm works in two stages

- Training Data
- Testing Data

The dataset is divided into training and evaluation subsets, usually with the ratio of 70-80% for training and 20-30% for evaluation. Machine learning algorithm uses 70% of the input data for the training data source and the remaining 30% of the input data for the evaluation data source.

```
20 #Split Dataset into Training and Testing dataset
21 from sklearn.model_selection import train_test_split
22
23 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state
```

FIGURE 3.2 – DATA SPLITTING

From the above Figure 3.2, it is implied that 30% of data are taken for testing and the remaining 70% of data are taken for training using train_test_split.

3.5.3 SUPERVISED MACHINE LEARNING TECHNIQUES

Machine learning is an application of Artificial Intelligence (AI) that provides the system to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning techniques include supervised learning and unsupervised learning techniques.

A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictors for the response to new data. It uses classification and regression techniques to develop predictive models. It is one of the most powerful engines that enable AI systems to make business decisions faster and more accurately than humans. It includes both classification and regression.

CLASSIFICATION

Classification techniques are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. It is the process of predicting the class of given data points. A classification technique in supervised machine learning includes K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression and Neural Networks.

3.5.3.1 SUPPORT VECTOR MACHINE

Support vector machine is a supervised machine learning algorithm used for both classification and regression. Support vector machine is a discriminative classifier formally defined by a separating hyper plane. A hyper plane is a line that splits the input variable space. It is selected to best separate the points in the input variable space by their class, either 0 or 1. The algorithm plot each data item as a point in n-dimensional space where n is the number of features and perform classification by finding the hyper-plane that differentiate the two classes. To separate the two classes of data points, there are many possible hyper planes that could be chosen. The objective is to find a plane that has the maximum margin, therefore

the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyper planes are decision boundaries that help classify the data points. Data points falling on either side of the hyper plane can be attributed to different classes. Also, the dimension of the hyper plane depends upon the number of features. If the number of input features is 2, then the hyper plane is just a line. If the number of input features is 3, then the hyper plane becomes a two dimensional plane and it is presented in Figure 3.3. It becomes difficult to imagine when the number of features exceeds 3.

Support vectors are data points that are closer to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyper plane.

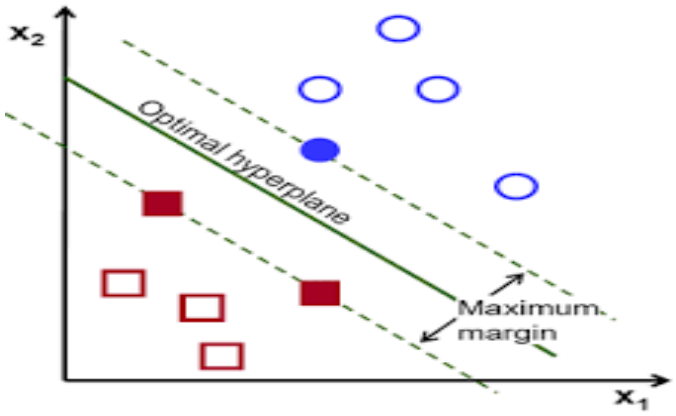


FIGURE 3.3 – Support Vector Machine using Hyper Plane

The learning of the hyper plane in linear SVM is done by transforming the problem using some linear algebra. The equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows

F (x) = B (0) + sum (ai * (x, xi)) ……. (3)

The above equation (3) is used to calculate the inner products of a new input vector (x) with all support vectors in training data. The coefficients B and ai (for each input) must be estimated from the training data by the learning algorithm.

3.5.3.2 K-NEAREST NEIGHBOR

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithm calculates the distance of a new data point to all other training data points. The distance can be of any type (Example: Euclidean or Manhattan distance). Classification is done by a majority vote to its neighbors, where K can be any integer. Then, it assigns the data point to the class to which the majority of the K data points belong.

Suppose the value of K is 3. The KNN algorithm starts by calculating the distance of point X from all the points. It then finds the 3 nearest points with least distance to point X which is presented in the below Figure 3.4.

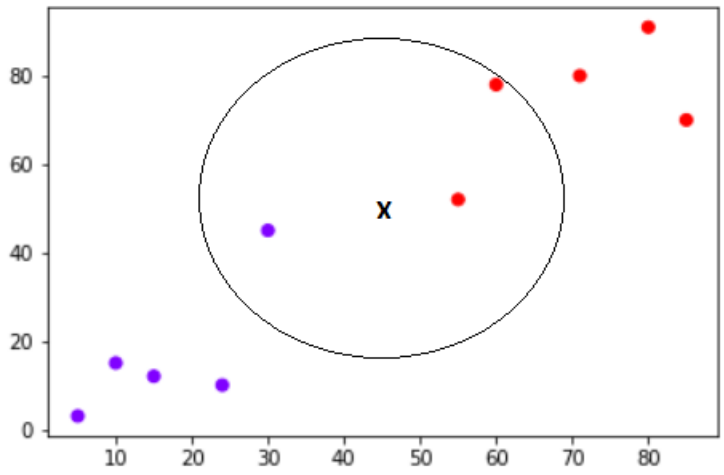


FIGURE 3.4 – K-Nearest Neighbor

The final step of the KNN algorithm is to assign new data points to the class to which majority of the nearest data points belong.

DIFFERENT WAYS TO COMPUTE DISTANCE

1) Euclidean Distance

The Euclidean is often the default distance used to find the ‘k closest points’ of a particular sample point.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad \text{..... (4)}$$

2) Manhattan Distance

Manhattan Distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates.

$$\sqrt{\sum_{i=1}^{k}|x_i - y_i|} \quad \text{..... (5)}$$

3) Minkowski Distance

Minkowski Distance is a generalized metric form of Euclidean and Manhattan distance. It is used for distance similarity of vector.

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \text{..... (6)}$$

3.5.3.3 LOGISTIC REGRESSION

Logistic regression is a supervised classification algorithm. In a classification problem, the target variable or output, Y can take only discrete values for given set of features or inputs, X.

It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

Logistic Regression returns the probability of binary dependent variable that is predicted from the independent variable of dataset that is logistic regression predict the probability of an outcome which has two values either zero or one, yes or no and false or true. Logistic regression has similarities to linear regression but as in linear regression a straight line is obtained, logistic regression shows a curve.

Logistic regression can be classified as

- Binomial

  Target variable can have only two possible types.
- Multinomial

  Target variable can have three or more possible types which are not ordered.

- Ordinal

  It deals with target variable with ordered categories.

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.

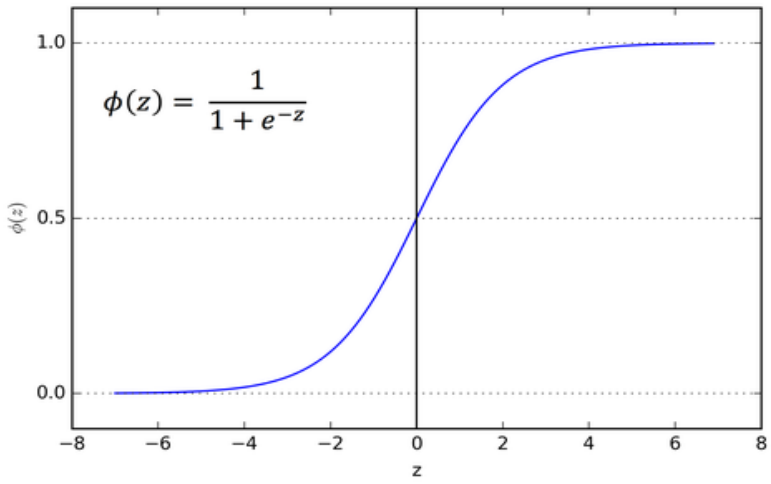


FIGURE 3.5 – Logistic Regression with sigmoid curve

From the above Figure 3.5, it is evident that, it is an S-shaped curve that can take any real-value number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value) …….. (7)

From the above equation (2), where ‘e’ is the base of the natural logarithms (Euler’s number or the EXP () function in your spreadsheet) and value is the actual numerical value.

3.5.3.4 NAÏVE BAYES

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in wide variety of classification tasks, which assumes that features are statistically independent. The theorem relies on the naïve assumption that input variables are independent of each other, i.e. there is no way to know anything about variables given an additional variable.

BAYES THEOREM

Bayes Theorem provides a way of calculating the posterior probability, P (H|E) from P (H), P (E) and P (E|H). Naïve Bayes classifier assumes that the effect of the value of a predictor (E) on a given class (H) is independent of the values of other predictors. This assumption is called class conditional independence. Bayes Theorem works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, calculate the probability of an event using its prior knowledge.

Formula for Calculating Conditional Probability

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \qquad \ldots\ldots\ldots\ldots (8)$$

Where

P - The symbol to denote probability.

P (H|E) - The probability of event H (hypothesis) occurring given that E (evidence) has occurred. This is also known as posterior probability.

P (E|H) - The probability of event E (evidence) occurring given that H (hypothesis) has occurred.

P (H) – The probability of event E (evidence) occurring.

P (E) – The probability of event H (hypothesis) occurring.

The representation of Naive Bayes is probabilities

1) Class probabilities - The probabilities of each class in the training dataset.
2) Conditional probabilities – The conditional probabilities of each input value given each class value.

In binary classification the probability of an instance belonging to class 1 would be calculated as

$$P(\text{class}=1) = \frac{\text{count (class=1)}}{\text{(count (class=0) + count (class=1))}} \quad \ldots\ldots\ldots (9)$$

TYPES OF NAIVE BAYES

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bernoulli Naive Bayes

GAUSSIAN NAIVE BAYES

When attribute values are continuous, an assumption is made that the values associated with each class are distributed according to Gaussian and it is presented in the below Figure 3.6.

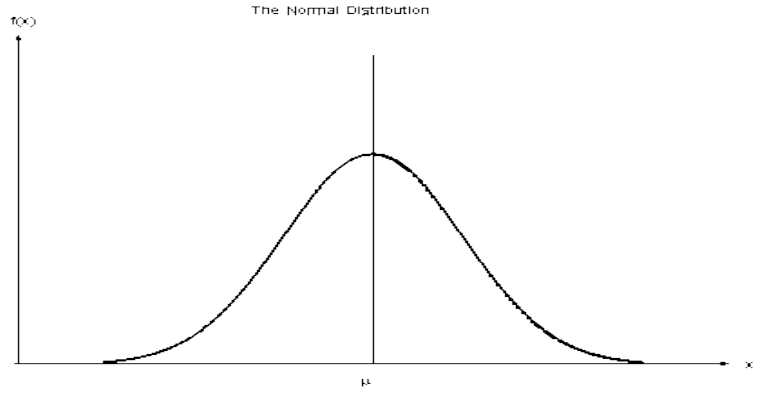


FIGURE 3.6 – GAUSSIAN DISTRIBUTION

In data, an attribute say "x" contains continuous data. Segment the data by the class and then compute mean and variance of each class.

$$p(x_i|y_i) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}} \quad \ldots\ldots\ldots (10)$$

EXPERIMENTAL RESULTS AND DISCUSSIONS

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 PERFORMANCE EVALUATION

The metrics used to evaluate the performance of different machine learning algorithms are

- Confusion matrix
- Accuracy
- Precision
- Recall
- F-Measure
- roc_auc_score

CONFUSION MATRIX

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model, which is presented in Figure 4.1. It is used for Classification problem where the output can be of two or more types of classes.

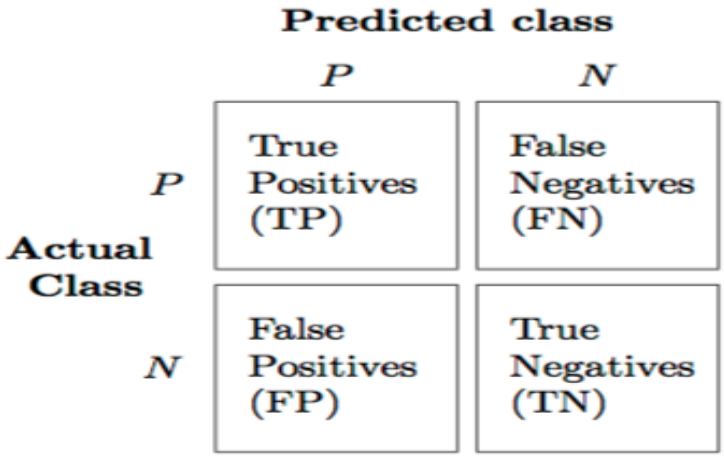


FIGURE 4.1 Confusion Matrix

True Positives (TP)

True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True)

True Negatives (TN)

True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False)

False Positives (FP)

False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True).

False Negatives (FN)

False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False).

ACCURACY

Accuracy in the classification problem is the ratio of number of correct predictions to the total number of input samples.

$$\textbf{Accuracy} = \frac{\text{TRUE POSITIVES + TRUE NEGATIVES}}{\text{TOTAL NUMBER OF SAMPLES}} \quad \ldots\ldots\ldots (11)$$

PRECISION

It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\textbf{Precision} = \frac{\text{TRUE POSITIVES}}{\text{TRUE POSITIVES + FALSE POSITIVES}} \quad \ldots\ldots\ldots\ldots (12)$$

RECALL OR SENSITIVITY

It is the number of correct positive results divided by the number of *all* relevant samples (all samples that should have been identified as positive).

$$\textbf{Recall} = \frac{\text{TRUE POSITIVES}}{\text{TRUE POSITIVES + FALSE NEGATIVES}} \quad \text{..... (13)}$$

F-MEASURE

F-Measure is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tries to find the balance between precision and recall. F-Measure combines precision and recall. It is also known as harmonic mean. The function F assumes values in the interval [0, 1]. It is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. Also the value of harmonic mean is high when both recall and precision are high.

$$\textbf{F-Measure} = \frac{2*(\text{PRECISION}*\text{RECALL})}{(\text{PRECISION}+\text{RECALL})} \quad \text{..... (14)}$$

ROC_AUC_SCORE

Compute Area under the Receiver Operating Characteristic. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC curve is used for visual comparison of classification models which shows the trade-off between the true positive rate and the false positive rate. The area under the ROC curve is a measure of the accuracy of the model. When a model is closer to the diagonal, it is less accurate and the model with perfect accuracy will have an area of 1.0

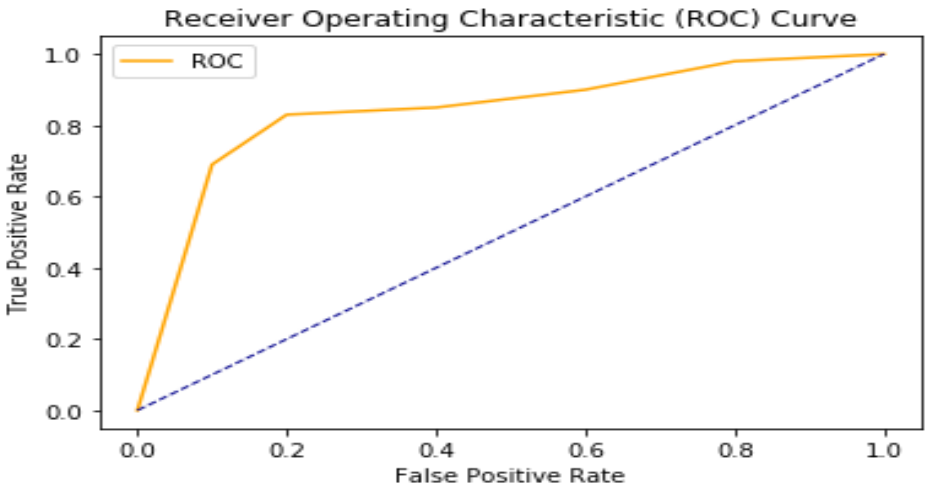


FIGURE 4.2 ROC Curve

From the above Figure 4.2, it is evident that the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. ROC curves are appropriate when the observations are balanced between each class, whereas precision recall curves are appropriate for imbalanced datasets.

TABLE 4.1 PERFORMANCE EVALUATION

METRICS / ALGORITHMS	ROC_AUC SCORE (%)	ACCURACY (%)	PRECISION (%)	RECALL (%)
KNN	54	99	91	7
LOGISTIC REGRESSION	76	99	75	52
SVM	70	99	40	61
NAÏVE BAYES	82	99	13	65

From the above table 4.1 the performance evaluation of the supervised machine learning techniques such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression are done based on the area under curve and receiver operating characteristic. The Naïve Bayes algorithm detects the fraudulent transactions with 82% and performs well when compared to other supervised learning algorithms such as K-Nearest Neighbor, Support Vector Machine and Logistic Regression.

The below Figure 4.3 shows the performance evaluation of supervised machine learning algorithms such as K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression based on Receiver Operating Characteristic and Area Under Curve. The Naïve Bayes algorithm performs well when compared to other supervised machine learning algorithms like K-Nearest Neighbor, Support Vector Machine and Logistic Regression.

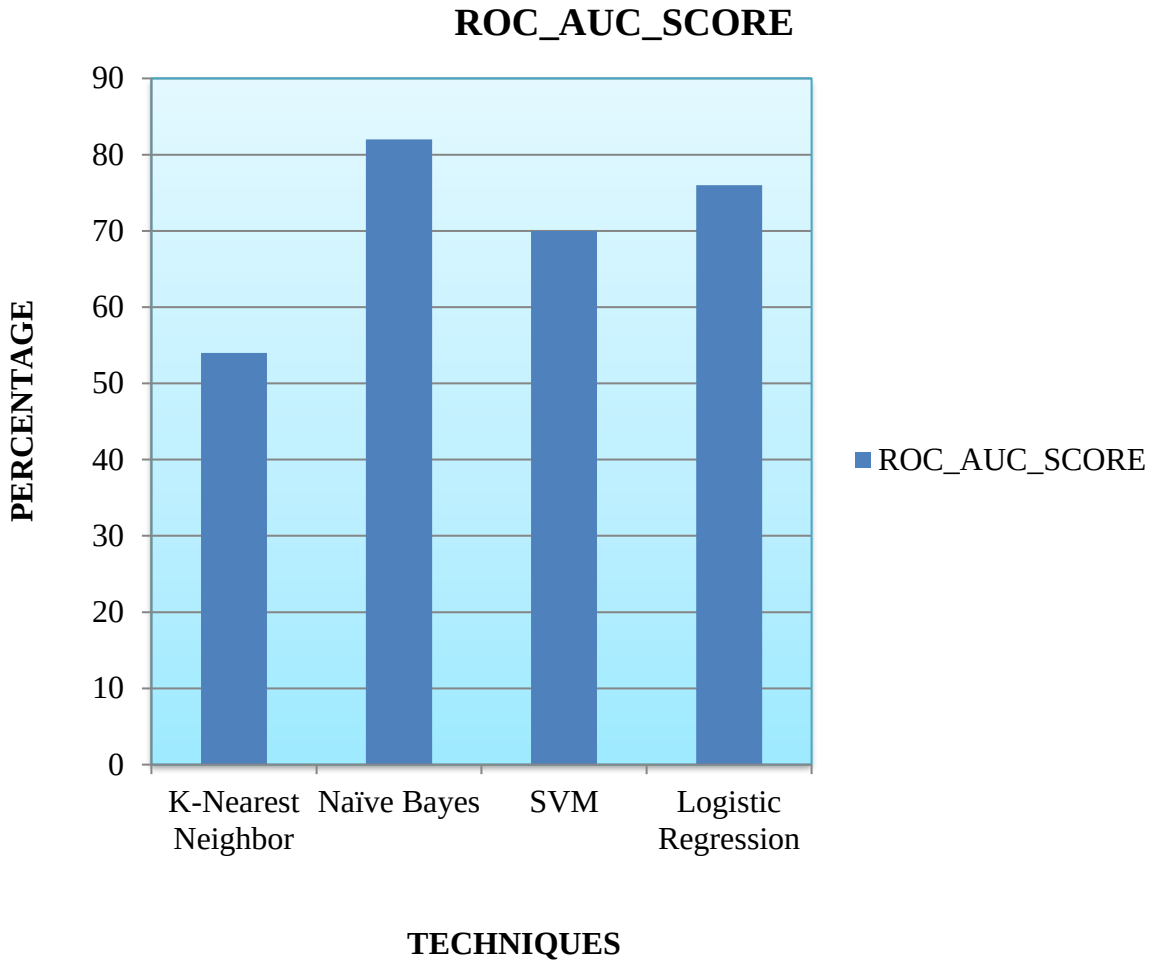


FIGURE 4.3 – Performance Evaluation of Supervised Machine Learning Techniques

CONCLUSION

5. CONCLUSION

This research work concentrated in detecting the suspicious credit card transactions using supervised machine learning algorithms like K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Logistic Regression and the performance of these techniques are evaluated based on the area under curve and receiver operating characteristic. The results of roc_auc_score for K-Nearest Neighbor (54%), Naïve Bayes(82%), Support Vector Machine(70%) and Logistic Regression(82%). The Naïve Bayes algorithm detects the fraudulent transactions and performs well compared to the K-Nearest Neighbor, Support Vector Machine and Logistic regression.

SCOPE FOR FUTURE ENHANCEMENT

6. SCOPE FOR FUTURE ENHANCEMENT

- This research work has covered almost all the requirements. Further requirements and improvements can easily be done.
- The future work study will attempt to explore more credit card fraud detection using real time data.
- AE (Auto-Encoder) and RBM (Restricted Boltzmann Machine) are the two deep learning algorithms that use real credit card fraud transactions with a huge amount of data and make more accurate AUC for receiver operator characteristics.

BIBLIOGRAPHY

7. BIBLIOGRAPHY

JOURNAL REFERENCES

1. Lakshmi SVSS , Selvani Deepthi Kavila, (2018), “Machine Learning For Credit Card Fraud Detection System”, Department of CSE, Anil Neerukonda Institute Of Technology And Sciences(A), Visakhapatnam-531162,India, International Journal of Applied Engineering Research ,Volume 13, Number 24, ISSN 0973-4562.
2. Sahin.Y & Duman.E (2011), “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines”, Hong Kong, China, International Multi Conference of Engineers and Computer Scientists, Volume 01, ISSN 2078-0958.
3. Navanshu Khare and Saad Yunus Sait, (2018), “Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models”, International Journal of Pure and Applied Mathematics, Volume: 118, Number 20, ISSN 1314-3395.
4. G.Suresh ,R.Justin Raj, (2018) “A Study on Credit Card Fraud Detection using Data MiningTechniques”,International Journal of Data Mining Techniques and Applicatio ns, Volume 07, Issue 01, Page Number 21-24, ISSN 2278-2419.
5. B.Pushpalatha, C.Willson Joseph, (2017), “Credit Card Fraud Detection Based on the Transaction by Using Data mining Techniques”, International Journal of Innovative Research in Computer and Communication Engineering, Volume 05, Issue 02, ISSN 2320-9801.
6. Shivakumar Swamy N, Sanjeev C. Lingareddy, (2014), “Fraud Detection using Data Mining Techniques”, Department of CSE, International Journal of Innovations in Engineering and Technology (IJIET), Volume 04, Issue 01, ISSN 2319-1058.

WEBSITES

1. https://www.researchgate.net/publication/326986162_Credit_Card_Fraud_Det ection_Using_Machine_Learning_As_Data_Mining_Technique
2. https://www.academia.edu/36810759/Machine_Learning_Approaches_for_Cr edit_Card_FraudDetection
3. https://acadpubl.eu/hub/2018-118-21/articles/21b/90
4. https://www.ijcaonline.org/archives/volume140/number5/24594-2016909316
5. https://www.3pillarglobal.com/insights/credit-card-fraud-detection

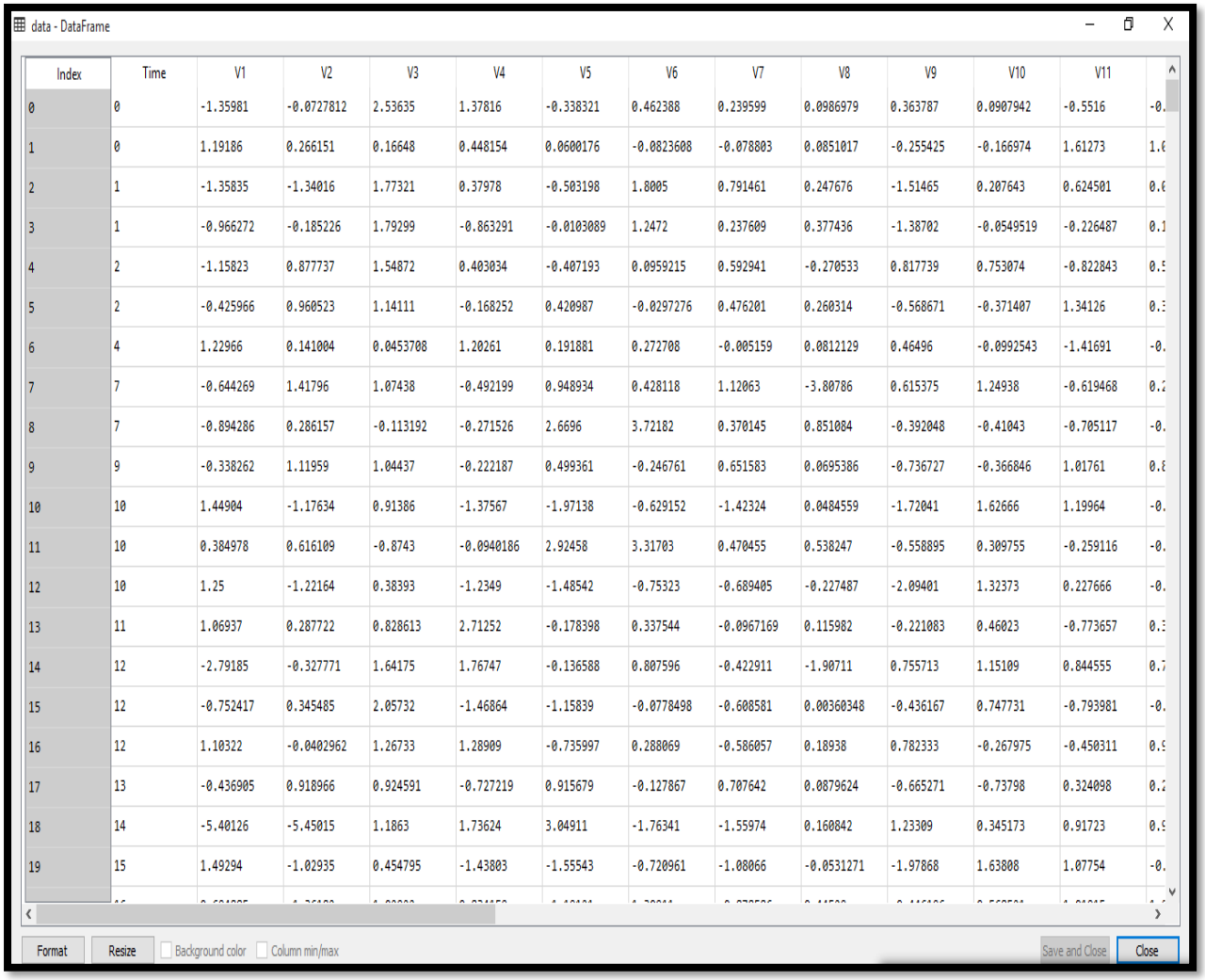
APPENDIX

8. APPENDIX

8.1 SCREEN SHOTS

8.1.1 DATA COLLECTION

The source of the dataset is from Kaggle and it contains transactions made by credit cards in September 2013 by European cardholders that occurred in two days and have 284,807 transactions. Credit card fraud detection dataset consists of 284,807 instances and 31 attributes. It contains only numerical input variables which are the result of a PCA transformation, time elapsed between transactions, transaction amount and the feature class is the response variable.



Index	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	
0	0	-1.35981	-0.0727812	2.53635	1.37816	-0.338321	0.462388	0.239599	0.0986979	0.363787	0.0907942	-0.5516	-0.
1	0	1.19186	0.266151	0.16648	0.448154	0.0600176	-0.0823608	-0.078803	0.0851017	-0.255425	-0.166974	1.61273	1.(
2	1	-1.35835	-1.34016	1.77321	0.37978	-0.503198	1.8005	0.791461	0.247676	-1.51465	0.207643	0.624501	0.(
3	1	-0.966272	-0.185226	1.79299	-0.863291	-0.0103089	1.2472	0.237609	0.377436	-1.38702	-0.0549519	-0.226487	0.1
4	2	-1.15823	0.877737	1.54872	0.403034	-0.407193	0.0959215	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.5
5	2	-0.425966	0.960523	1.14111	-0.168252	0.420987	-0.0297276	0.476201	0.260314	-0.568671	-0.371407	1.34126	0.3
6	4	1.22966	0.141004	0.0453708	1.20261	0.191881	0.272708	-0.005159	0.0812129	0.46496	-0.0992543	-1.41691	-0.
7	7	-0.644269	1.41796	1.07438	-0.492199	0.948934	0.428118	1.12063	-3.80786	0.615375	1.24938	-0.619468	0.2
8	7	-0.894286	0.286157	-0.113192	-0.271526	2.6696	3.72182	0.370145	0.851084	-0.392048	-0.41043	-0.705117	-0.
9	9	-0.338262	1.11959	1.04437	-0.222187	0.499361	-0.246761	0.651583	0.0695386	-0.736727	-0.366846	1.01761	0.8
10	10	1.44904	-1.17634	0.91386	-1.37567	-1.97138	-0.629152	-1.42324	0.0484559	-1.72041	1.62666	1.19964	-0.
11	10	0.384978	0.616109	-0.8743	-0.0940186	2.92458	3.31703	0.470455	0.538247	-0.558895	0.309755	-0.259116	-0.
12	10	1.25	-1.22164	0.38393	-1.2349	-1.48542	-0.75323	-0.689405	-0.227487	-2.09401	1.32373	0.227666	-0.
13	11	1.06937	0.287722	0.828613	2.71252	-0.178398	0.337544	-0.0967169	0.115982	-0.221083	0.46023	-0.773657	0.3
14	12	-2.79185	-0.327771	1.64175	1.76747	-0.136588	0.807596	-0.422911	-1.90711	0.755713	1.15109	0.844555	0.7
15	12	-0.752417	0.345485	2.05732	-1.46864	-1.15839	-0.0778498	-0.608581	0.00360348	-0.436167	0.747731	-0.793981	-0.
16	12	1.10322	-0.0402962	1.26733	1.28909	-0.735997	0.288069	-0.586057	0.18938	0.782333	-0.267975	-0.450311	0.9
17	13	-0.436905	0.918966	0.924591	-0.727219	0.915679	-0.127867	0.707642	0.0879624	-0.665271	-0.73798	0.324098	0.2
18	14	-5.40126	-5.45015	1.1863	1.73624	3.04911	-1.76341	-1.55974	0.160842	1.23309	0.345173	0.91723	0.9
19	15	1.49294	-1.02935	0.454795	-1.43803	-1.55543	-0.720961	-1.08066	-0.0531271	-1.97868	1.63808	1.07754	-0.

FIGURE 8.1 – Loading the Data

8.1.2 DATASET SPLITTING

The dataset is divided into training and testing data with 30% of the data is given for testing and the remaining 70% for training the data using supervised machine learning techniques to detect the fraudulent transactions.

FIGURE 8.2 – Splitting Training and Testing Data

8.1.3 SUPERVISED LEARNING TECHNIQUES

Supervised learning techniques include

- K-Nearest Neighbor
- Naïve Bayes
- Support Vector Machine
- Logistic Regression

K-NEAREST NEIGHBOR

FIGURE 8.3 - K-Nearest Neighbor

PERFORMANCE EVALUATION OF K-NEAREST NEIGHBOR

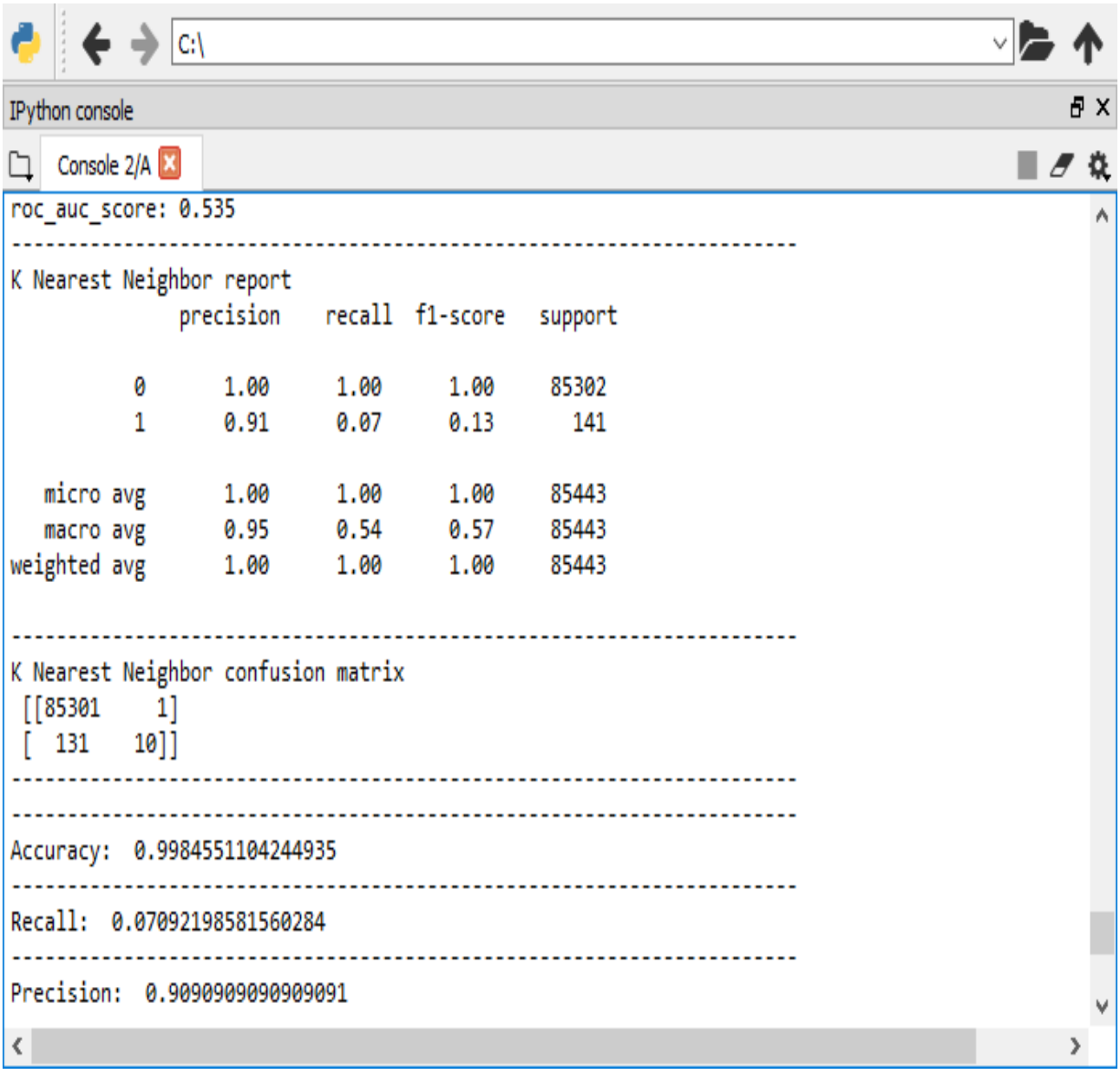


FIGURE 8.4 – Performance of K-Nearest Neighbor

NAÏVE BAYES

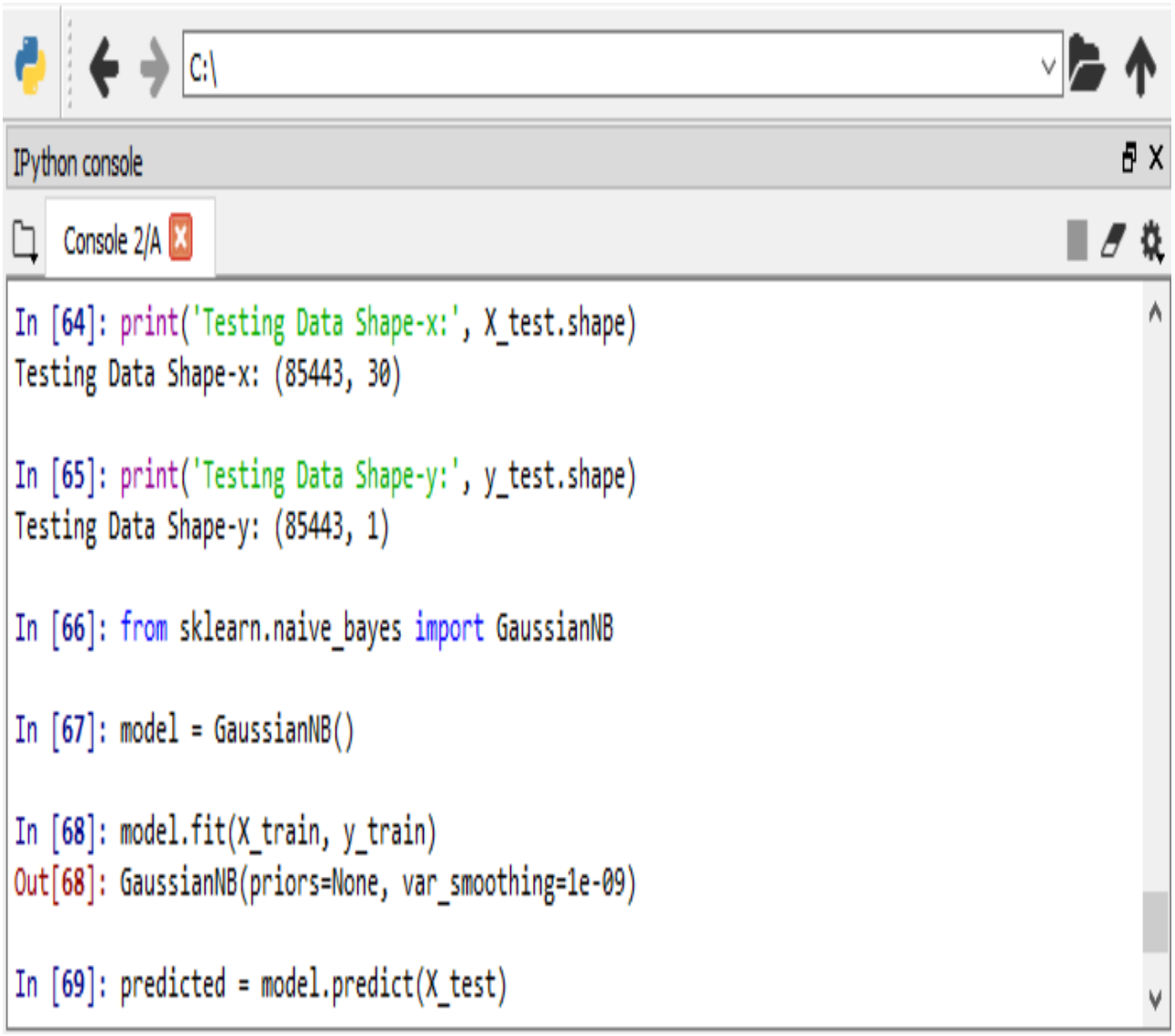


FIGURE 8.5 - Naïve Bayes

PERFORMANCE EVALUATION OF NAÏVE BAYES

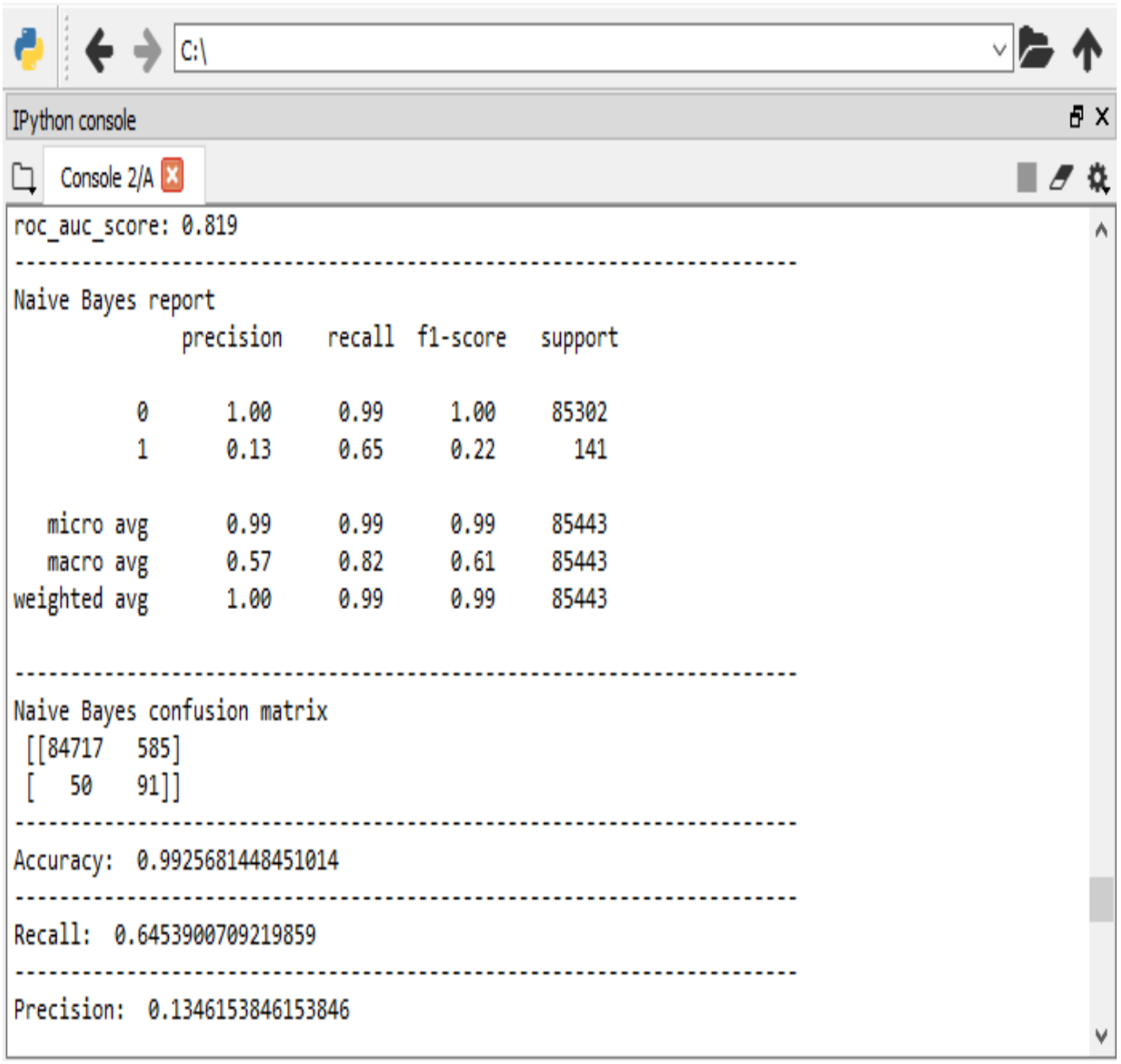


FIGURE 8.6 – Performance of Naïve Bayes

SUPPORT VECTOR MACHINE

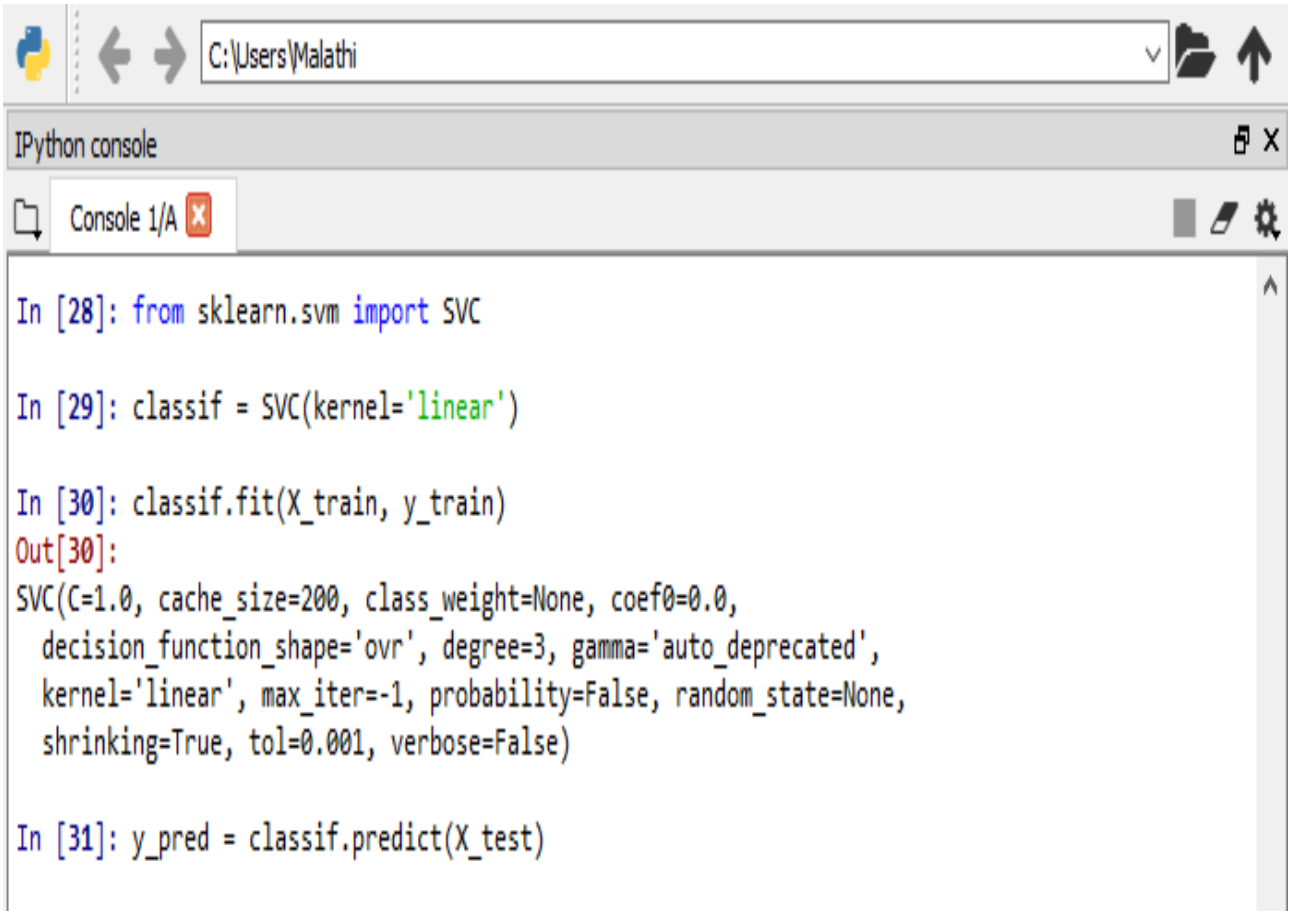


FIGURE 8.7 - Support Vector Machine

PERFORMANCE EVALUATION OF SUPPORT VECTOR MACHINE

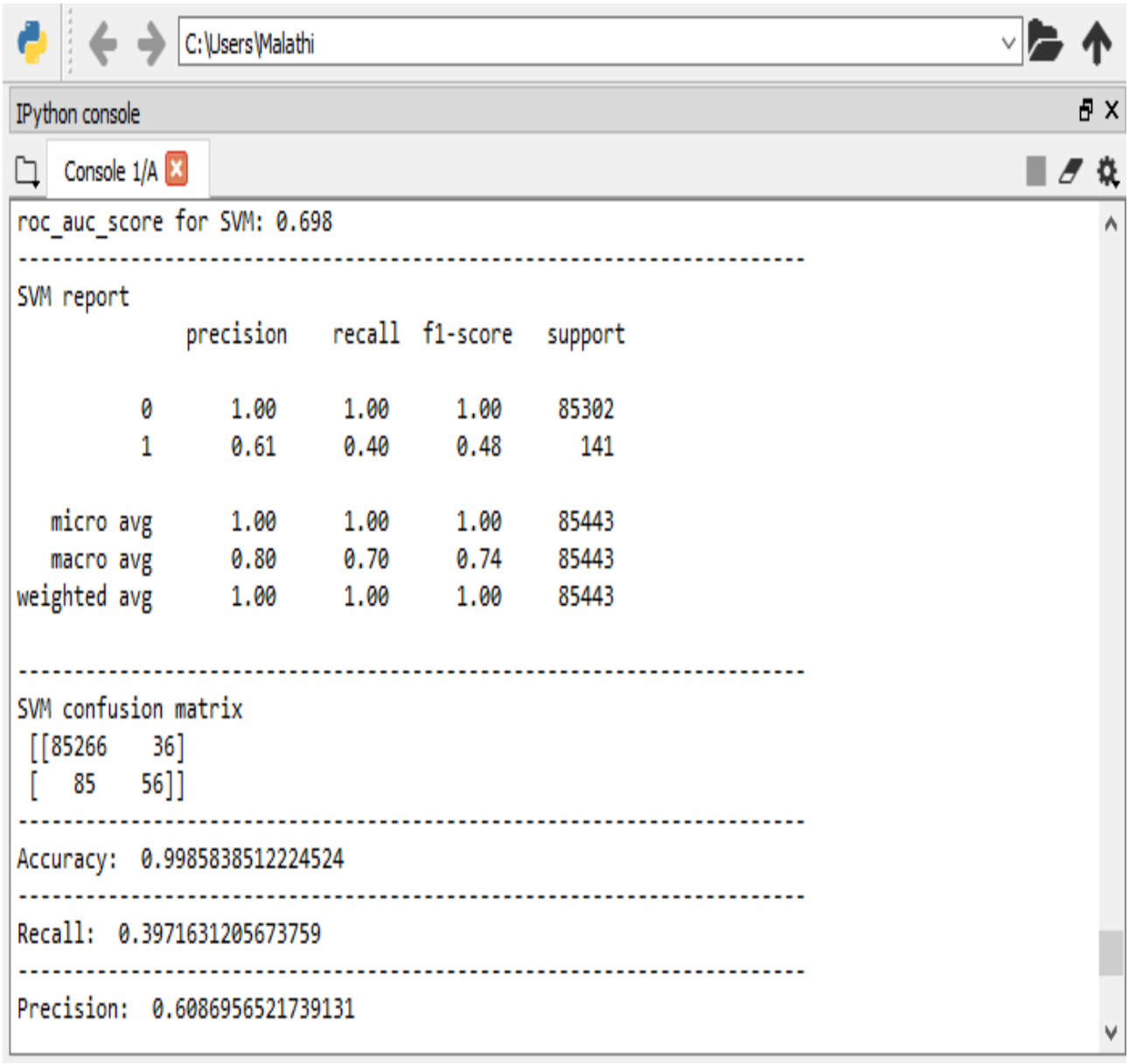


FIGURE 8.8 – Performance of Support Vector Machine

LOGISTIC REGRESSION

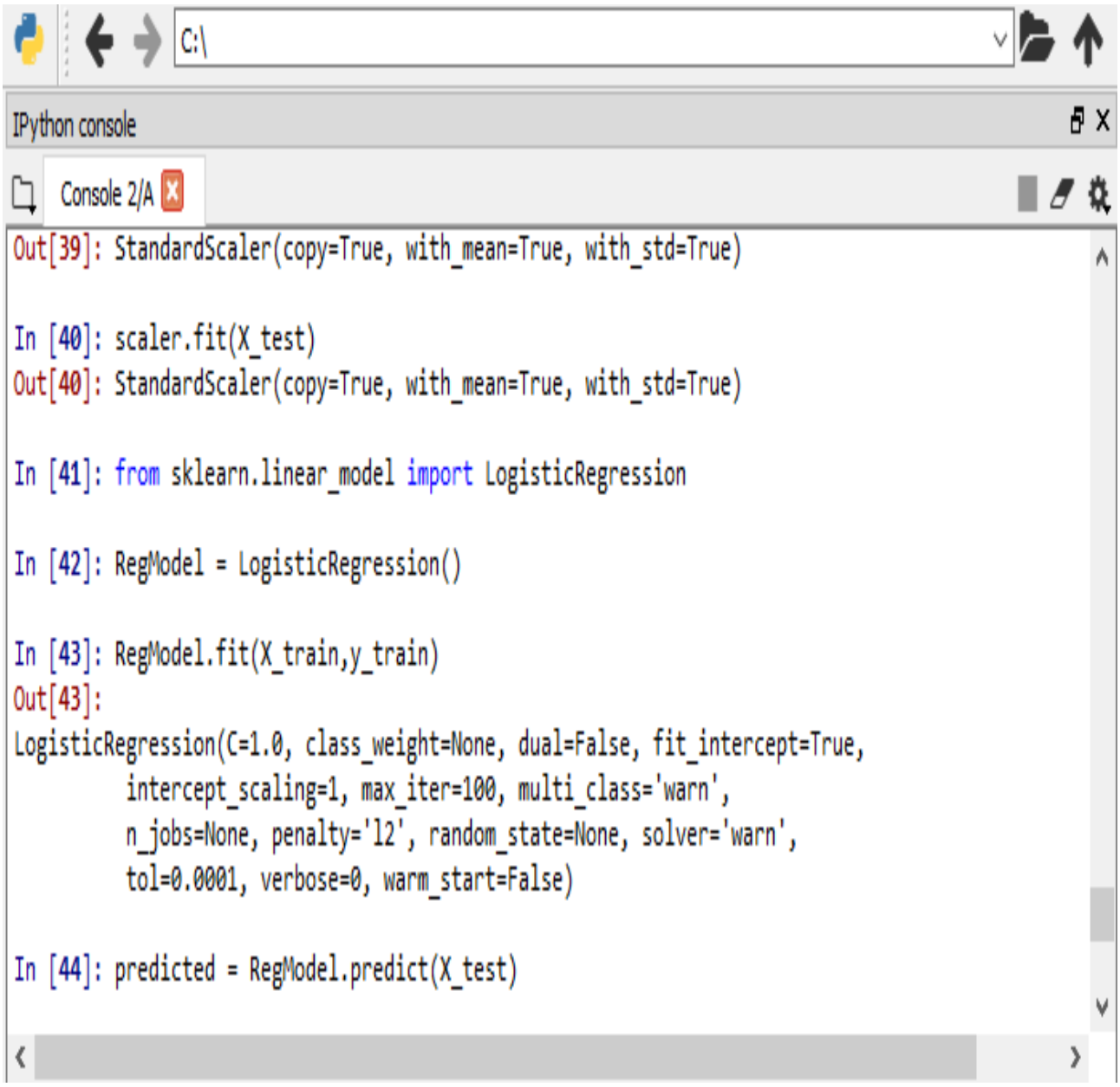


FIGURE 8.9 – Logistic Regression

PERFORMANCE EVALUATION OF LOGISTIC REGRESSION

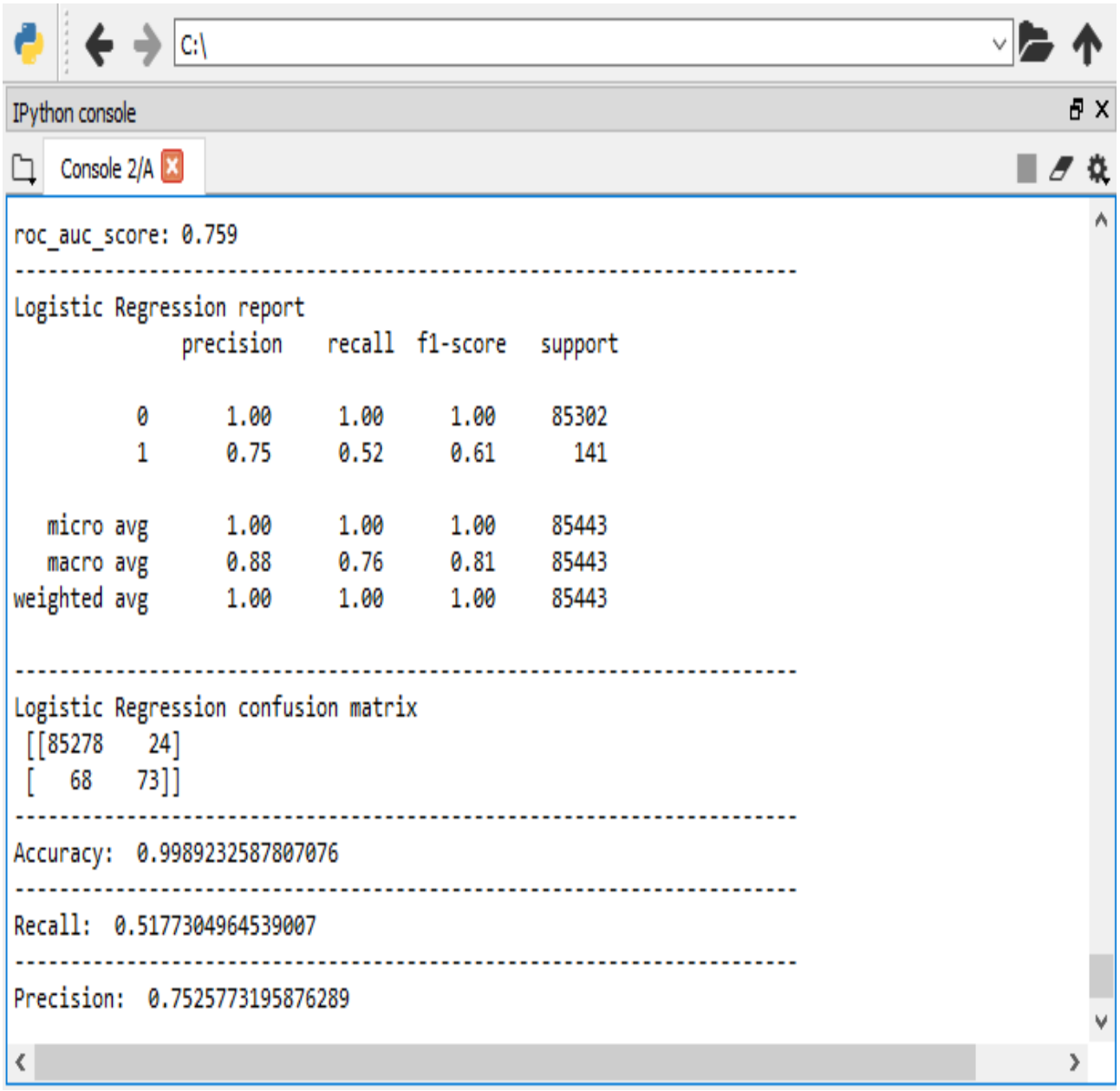


FIGURE 8.10 – Performance of Logistic Regression