

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology degree in Information Technology

By

SRIRAM J (Reg.No - 40120092)

SUDHARSAN P (Reg.No - 40120094)

DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

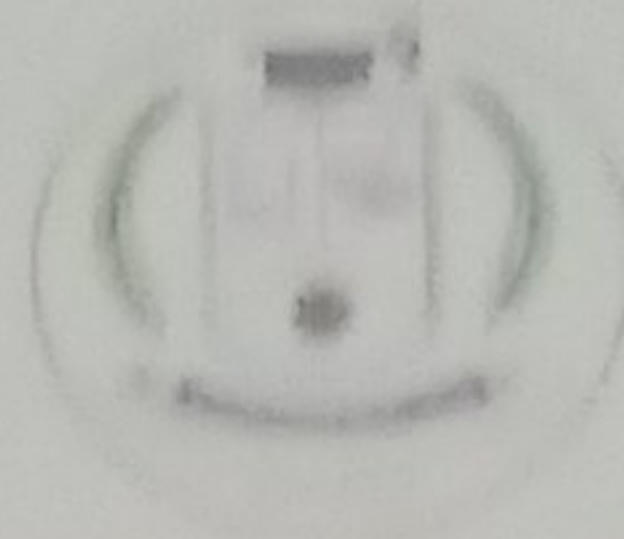
Category-1 University by UGC

Accredited with Grade “A++” by NAAC | 12B Status by UGC |

Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600119

APRIL - 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

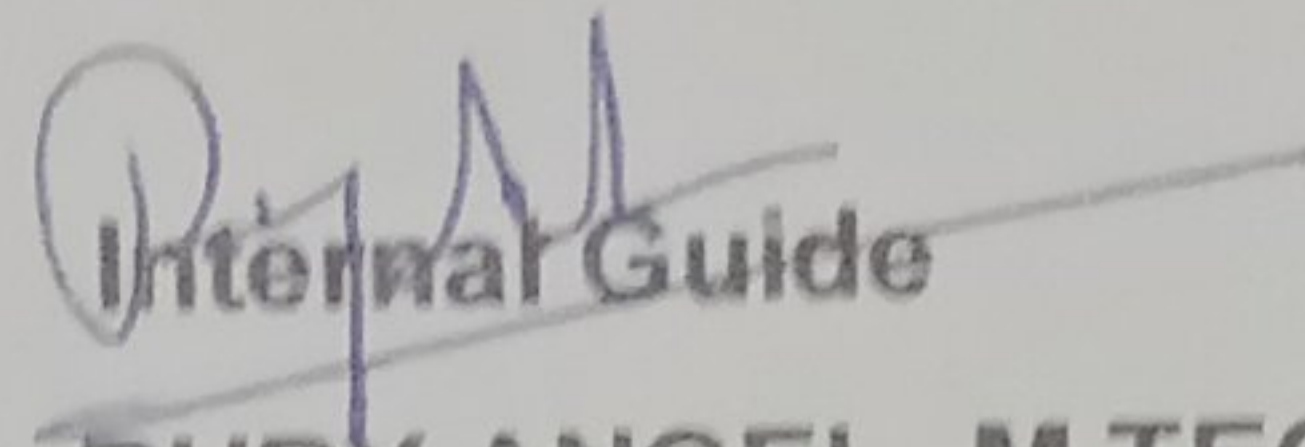
Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

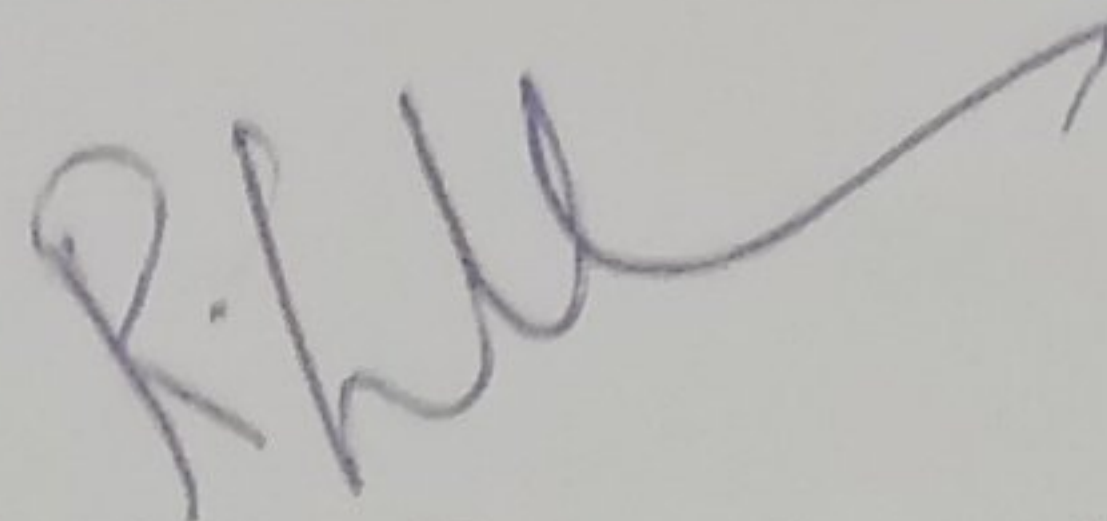
BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SRIRAM J (40120092)** and **SUDHARSAN P(40120094)** who carried out the Project Phase-2 entitled "**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING**" under my supervision from January 2024 to April 2024.



Internal Guide

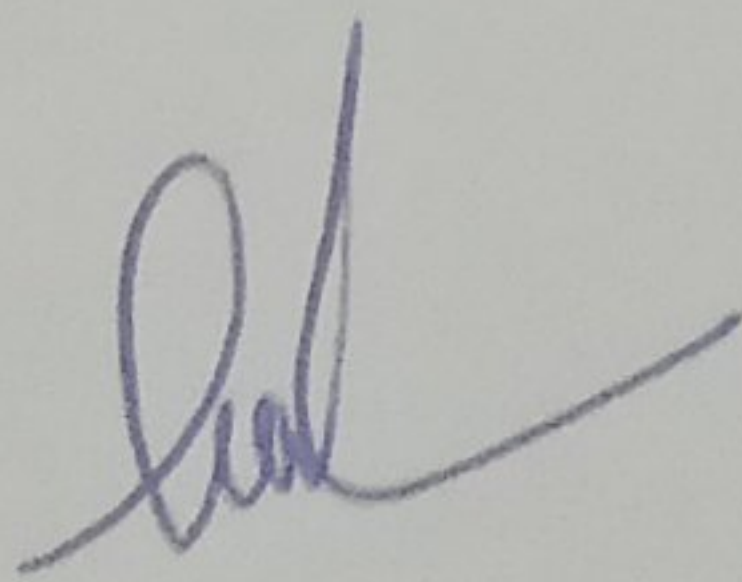
Ms.T.G. RUBY ANGEL, M.TECH.,



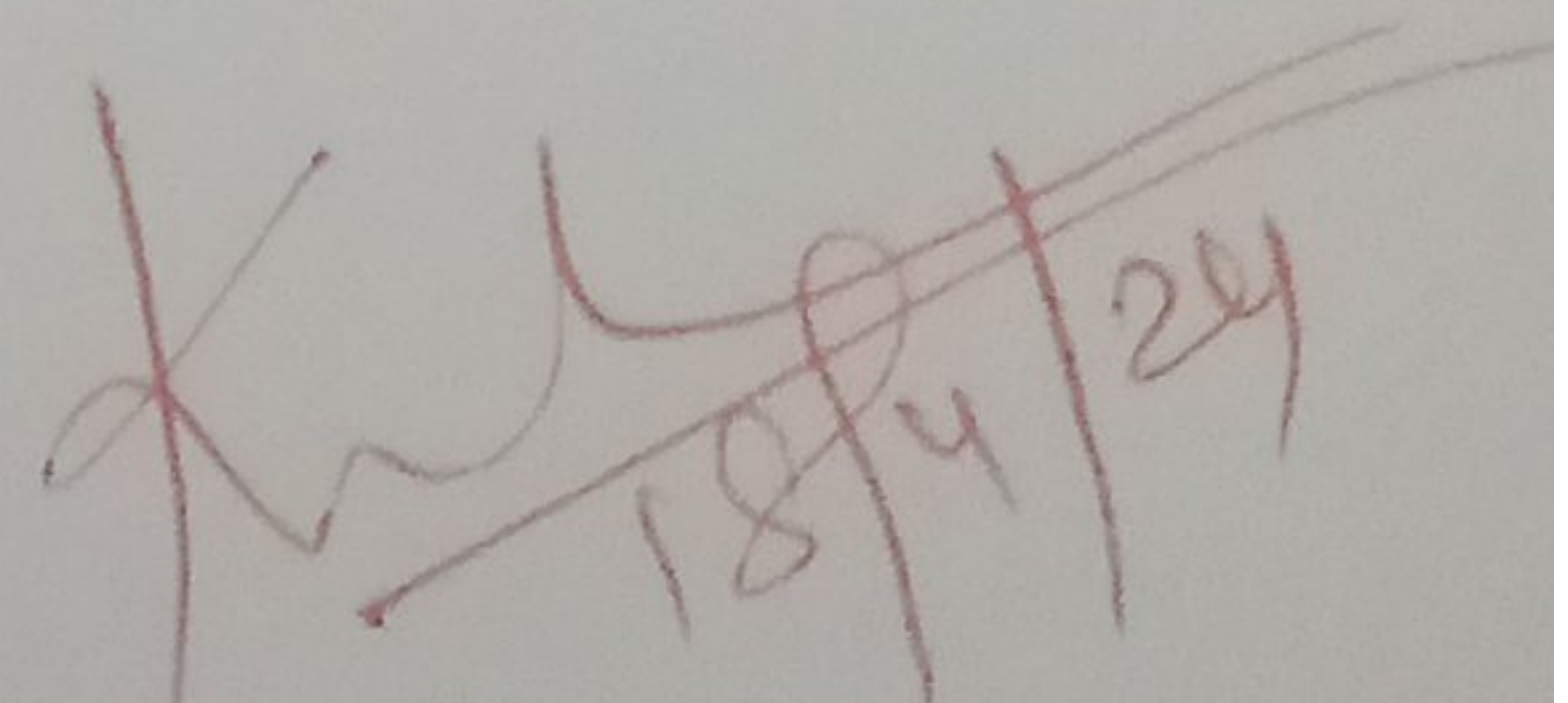
Head of the Department

Dr.R. SUBHASHINI, M.E., Ph.D.

Submitted for Viva voce Examination held on 18/04/24



Internal Examiner



18/4/24

External Examiner

DECLARATION

I, **SUDHARSAN P(40120094)** hereby declare that the Project Phase-2 Report entitled "**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING**" done by me under the guidance of **MS. T.G. Ruby Angel, M.TECH.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in **Information Technology**.

DATE: 18/04/2024

p. Sudharsan

SIGNATURE OF CANDIDATE

PLACE: Chennai

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. T. Sasikala M.E., Ph.D., Dean, School of Computing, Dr. R.Subhashini M.E., Ph.D., Head of the Department of Information technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Ms.T.G.Ruby Angel, M.TECH., for her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Information Technology who were helpful in many ways for the completion of the project.

ABSTRACT

A credit card is issued by a bank or financial services company that allows cardholders to borrow funds with which to pay for goods and services with merchants that accept cards for payment. Nowadays as everything is made cyber so there is a chance of misuse of cards and the account holder can lose the money so it is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. This type of problems can be solved through data science by applying machine learning techniques. It deals with modelling of the dataset using machine learning with Credit Card Fraud Detection. In machine learning the main key is the data so modelling the past credit card transactions with the data of the ones that turned out to be fraud. The built model is then used to recognize whether a new transaction is fraudulent or not. The objective is to classify whether the fraud had happened or not. The first step involves analysing and pre-processing data and then applying machine learning algorithm on the credit card dataset and find the parameters of the algorithm and calculate their performance metrics. The main aim of the paper is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyze the past transaction details of the customers and extract the behavioural patterns. Where cardholders are clustered into different groups based on their transaction amount. Then using sliding window strategy [1], to aggregate the transaction made by the cardholders from different groups so that the behavioural pattern of the groups can be extracted respectively. The main aim of the paper is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyze the past transaction details of the customers and extract the behavioural patterns. Where cardholders are clustered into different groups based on their transaction amount. Then using sliding window strategy to aggregate the transaction made by the cardholders from different groups so that the behavioural pattern of the groups can be extracted respectively.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
1	INTRODUCTION	1
2	LITERATURE SURVEY	4
3	METHODOLOGY	8
	3.1 DATA SOURCE	8
	3.2 SYSTEM REQUIREMENTS	8
	3.3 DOMAIN DESCRIPTION	9
	3.4 DATA MINING	10
	3.5 TYPES OF DATA MINING	10
	3.6 PYTHON PACKAGES	11
4	PROJECT DESIGN	12
	4.1 LIST OF MODULES	12
	4.2 MODULE DESCRIPTION	13
5	RESULT AND DISCUSSION	22
6	CONCLUSION	23
	6.1 FUTURE ENHANCEMENT	24
	REFERENCES	25
	APPENDIX	
	A.SOURCE CODE	27
	B.SCREENSHOTS	30
	C.PUBLICATION	33

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
4.1	DATASET STRUCTURE	13
4.2	CLASS DISTRIBUTION	14
4.3	CORRELATION	14
4.4	ATTRIBUTES OF CLASSIFICATION	15
4.5	KNN CLASSIFIER	16
4.6	KNN CONFUSION MATRIX	16
4.7	KNN CLASSIFIER	17
4.8	KNN CONFUSION MATRIX	17
4.9	NAÏVE BAYES CLASSIFIER	18
4.10	NAÏVE BAYES CONFUSION MATRIX	18
4.11	LOGICAL REGRESSION CLASSIFIER	19
4.12	LOGICAL REGRESSION CONFUSION MATRIX	19
4.13	SVM CLASSIFIER	20
4.14	SYSTEM ARCHITECTURE	21

CHAPTER 1

INTRODUCTION

With the increase of people using credit cards in their daily lives, credit card companies should take special care in the security and safety of the customers. According to (Credit card statistics 2021) the number of people using credit cards around the world was 2.8 billion in 2019, in addition 70% of those users own a single card at least. Reports of Credit card fraud in the US rose by 44.7% from 271,927 in 2019 to 393,207 reports in 2020. There are two kinds of credit card fraud, the first one is by having a credit card account opened under your name by an identity thief, reports of this fraudulent behaviour increased 48% from 2019 to 2020. The second type is by an identity thief uses an existing account that you created, and it's usually done by stealing the information of the credit card, reports on this type of fraud increased 9% from 2019 to 2020 (Daly, 2021). Those statistics caught my attention as the numbers are increasing drastically and rapidly throughout the years, which gave me the motive to try to resolve the issue analytically by using different machine learning methods to detect the credit card fraudulent transactions within numerous transactions. The main aim of this project is the detection of credit card fraudulent transactions, as it's important to figure out the fraudulent transactions so that customers don't get charged for the purchase of products that they didn't buy. The detection of the credit card fraudulent transactions will be performed with multiple ML techniques then a comparison will be made between the outcomes and results of each technique to find the best and most suited model in the detection of the credit card transaction that are fraudulent, graphs and numbers will be provided as well. In addition, exploring previous literatures and different techniques used to distinguish the fraud within a dataset. Credit card fraud detection employs a multifaceted approach to safeguard financial transactions. Machine learning algorithms, including neural networks, decision trees, and ensemble methods, are adept at learning from historical transaction data to recognize fraud-related patterns. Predictive analytics harness historical data and statistical techniques to gauge the likelihood of a transaction being fraudulent. Real-time monitoring systems swiftly identify and thwart fraudulent activities as they unfold, ensuring prompt intervention.

Furthermore, biometric authentication offers an additional layer of security by utilizing traits like fingerprints and facial recognition, thereby mitigating the risk of unauthorized transactions. Behavioral analytics delves into spending habits, transaction frequencies, and behavioral trends to pinpoint irregular activities. Augmenting transaction data with external information, such as device data and geolocation, through data enrichment techniques, enhances the accuracy of fraud identification. This comprehensive arsenal of methods collaborates to fortify credit card fraud management, providing a dynamic defense against evolving fraudulent tactics.

For instance, a machine learning model trained on a dataset of legitimate and fraudulent transactions can learn to differentiate between normal spending patterns and unusual activities. Let's say a user typically makes transactions within a certain geographic region and at specific times of the day. If suddenly, there are multiple transactions from different countries or during unusual hours, the machine learning model can flag these transactions as potentially fraudulent. Credit card fraud detection is the collective term for the policies, tools, methodologies, and practices that credit card companies and financial institutions take to combat identity fraud and stop fraudulent transactions.

In recent years, as the amount of data has exploded and the number of payment card transactions has skyrocketed, credit fraud detection has become largely digitized and automated.. Most modern solutions leverage artificial intelligence (AI) and machine learning (ML) to manage data analysis, predictive modeling, decision-making, fraud alerts and remediation activity that occur when individual instances of credit card fraud are detected. When a transaction falls outside the scope of normal activity, the anomaly detection tool will then alert the card issuer and, in some cases, the user. Depending on the transaction details and risk score assigned to the action, these fraud detection systems may flag the purchase for review or put a hold on the transaction until the user verifies their activity.

It is important to note that the most advanced algorithms are based on the individual user's behaviors and transaction history. Therefore, the model could make exceptions and allow transactions that are usually considered high-risk for the wider user base.

The goal is to develop a machine learning model for Credit Card Fraud Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

They proposed a method and named it as Information-Utilization Method INUM it was first designed and the accuracy and convergence of an information vector generated by INUM are analyzed. The novelty of INUM is illustrated by comparing it with other methods. Two D-vectors (i.e., feature subsets) a and b , where A_i is the i th feature in a data set, are dissimilar in decision space, but correspond to the same O-vector y in objective space. Assume that only a is provided to decision-makers, but a becomes inapplicable due to an accident or other reasons (e.g., difficulty to extract from the data set). Then, decision-makers are in trouble. On the other hand, if all two feature subsets are provided to them, they can have other choices to serve their best interest. In other words, obtaining more equivalent D-vectors in the decision space can provide more chances for decisionmakers to ensure that their interests are best served. Therefore, it is of great significance and importance to solve MMOPs with a good Pareto front approximation and also the largest number of D-vectors given each O-vector.

The proposed model is to build a classification model to classify whether its fraud or not. The dataset of previous credit card cases are collected where it is used to make the machine to learn about the problem. The first step for involves the analysis of data where each and every column is analyzed and the necessary measurements are taken for missing values and other forms of data. Outliers and other values which are not much impact is dealt. Then preprocessed data is used to build the classification model where the data will be split into two parts one is for training and remaining data for testing purpose.

CHAPTER 2

LITERATURE SURVEY

Review of Literature Survey

[1]Title : Credit Card Fraud Detection using Machine Learning: A Systematic

Literature Review

Author: Harish Paruchuri

Year: 2017

General

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them .Companies want to give more and more facilities to their customers. One of these facilities is the online mode of buying goods. The customers now can buy the required goods online but this is also an opportunity for criminals to do frauds. The criminals can theft the information of any cardholder and use it for online purchases until the cardholder contacts the bank to block the card. This paper shows the different algorithms of machine learning that are used for detecting this kind of transaction. The research shows the CCF is the major issue of financial sector that is

increasing with the passage of time. More and more companies are moving towards the online mode that allows the customers to make online transactions. This is an opportunity for criminals to theft the information or cards of other persons to make online transactions. The most popular techniques that are used to theft credit card information are phishing and Trojan. So a fraud detection system is needed to detect such activities.

[2]Title : A Research on Credit Card Fraudulent Detection System

Author: Devika S P, Nisarga K S, Gagana P Rao, Chandini S B, Rajkumar

Year : 2019

Nowadays credit card is more popular among the private and public employees. By using the credit card, the users purchase the consumable durable products in online, also transferring the amount from one account to other. The fraudster is detecting the details of the behaviour user transaction and doing the illegal activities with the card by phishing, Trojan virus, etc. The fraudulent may threaten the users on their sensitive information. In this paper, we have discussed various methods of detecting and controlling the fraudulent activities. This will be helpful to improve the security for card transaction in future. Credit card fraudulent activities which are faced by the people is one of the major issues. Due to these fraudulent activities, many credit card users are losing their money and their sensitive information. In this paper, we have discussed the different fraudulent detection and controlling techniques in credit card and also it will be helpful to improve the security from the fraudsters in future to avoid the illegal activities.

[3]Title : An Efficient Techniques for Fraudulent detection in Credit Card Dataset: A

Comprehensive study

Author: Akanksha Bansal and Hitendra Garg

Year : 2021

Increasing trends of financial transactions through credit cards also invite fraud activities that involve the loss of billions of dollars globally.

It is also been observed that fraudulent transactions have increased by 35% from 2018. A huge amount of transaction data is available to analyse the fraud detection activities that require analysis of behaviour/abnormalities in the transaction dataset to detect and ignore the undesirable action of the suspected person. The proposed paper lists a compressive summary of various techniques for the classification of fraud transactions from the various datasets to alert the user for such transactions. In the last decades, online transactions are growing rapidly and the most common tool for financial transactions. The increasing growth of online transactions also increases threats. Therefore, in keeping in mind the security issue, nature, an anomaly in the credit card transaction, the proposed work represents the summary of various strategies applied to identify the abnormal transaction in the dataset of credit card transaction datasets. This dataset contains a mix of normal and fraud transactions; this proposed work classifies and summarizes the various classification methods to classify the transactions using various Machine Learning-based classifiers. The efficiency of the method depends on the dataset and classifier used. The proposed summary will be beneficial to the banker, credit card user, and researcher to analyse to prevent credit card frauds. The future scope of this credit card fraud detection is to explore the things in each and every associations and banks to live safe and happily life. The data must be balanced in each place and we are getting the best results.

[4]Title : A Review On Credit Card Fraud Detection Using Machine Learning

Author: Suresh K Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil

Year : 2019

In recent years credit card fraud has become one of the growing problem. A large financial loss has greatly affected individual person using credit card and also the

merchants and banks. Machine learning is considered as one of the most successful technique to identify the fraud. This paper reviews different fraud detection techniques using machine learning and compare them using performance measure like accuracy, precision and specificity. The paper also proposes a FDS which uses supervised Random Forest algorithm. With this proposed system the accuracy of detecting fraud in credit card is increased. Further, the proposed system use learning to rank approach to rank the alert and also effectively addresses the problem concept drift in fraud detection. This paper has reviewed various machine learning algorithm detect fraud in credit card transaction. The performances of all this techniques are examined based on accuracy, precision and specificity metrics. We have selected supervised learning technique Random Forest to classify the alert as fraudulent or authorized. This classifier will be trained using feedback and delayed supervised sample. Next it will aggregate each probability to detect alerts. Further we proposed learning to rank approach where alert will be ranked based on priority. The suggested method will be able to solve the class imbalance and concept drift problem. Future work will include applying semi-supervised learning methods for classification of alert in FDS

[5]Title : Credit Card Fraud Detection and Prevention using Machine Learning

Author: S. Abinayaa, H. Sangeetha, R. A. Karthikeyan, K. Saran Sriram, D. Piyush

Year : 2020

This research focused mainly on detecting credit card fraud in real world. We must collect the credit card data sets initially for qualified data set. Then provide queries on the user's credit card to test the data set. After random forest algorithm classification method using the already evaluated data set and providing current data set[1]. Finally, the accuracy of the results data is optimised.

CHAPTER 3

METHODOLOGY

In order to accomplish the objective and goal of the project which is to find the most suited model to detect credit card fraud several steps need to be taken. Finding the most suited data and preparing/preprocessing are the first and second steps, after making sure that the data is ready the modelling phase starts, where 4 models are created, K-Nearest Neighbour (KNN) , Naïve Bayes, SVM and the last one is Logistic Regression. In the KNN model two Ks were chosen K=3 and K=7. All models were created in both R and Weka programs except SVM which was created in Weka only, in addition all visualizations are taken from both applications.

3.1 Data Source

The dataset was retrieved from an open-source website, Kaggle.com. it contains data of transactions that were made in 2013 by credit card users in Europe, in two days only. The dataset consists of 31 attributes, 284,808 rows. 28 attributes are numeric variables that due to 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 variables where “1” is a case of fraudulent transaction, and “0” is not as case of fraudulent transaction.

3.2 SYSTEM REQUIREMENTS

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware. The minimal hardware requirements are as follows

- | | |
|-------------------|--------------|
| 1.Processor | : Pentium IV |
| 2.RAM | : 8 GB |
| 3.Processor | : 2.4 GHz |
| 4.Main Memory | : 8GB RAM |
| 5.Hard Disk Drive | : 1tb |

Software requirements deals with defining resource requirements and prerequisites that needs to be installed on a computer to provide functioning of an application. The minimal software requirements are as follows,

- | | |
|------------------|--------------|
| Front end | : python |
| Dataset | : csv |
| IDE | : anaconda |
| Operating System | : Windows 10 |

3.3 DOMAIN DESCRIPTION DATA MINING

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology. The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc. Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis. Knowledge extraction, data/pattern analysis, information harvesting, etc .Data mining is a big area of data sciences, which aims to discover patterns and features in data, often large data sets. It includes regression, classification, clustering, detection of anomaly, and others. It also includes preprocessing, validation, summarization, and ultimately the making sense of the data sets.

3.4 Data Mining Techniques

Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

Regression

Regression analysis is the data mining method of identifying and analysing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

3.5 Types of Data Mining

Data mining has two primary processes: supervised and unsupervised learning.

SUPERVISED LEARNING

The goal of supervised learning is prediction or classification. The easiest way to conceptualize this process is to look for a single output variable. A process is considered supervised learning if the goal of the model is to predict the value of 13 an observation. One example is spam filters, which use supervised learning to classify incoming emails as unwanted content and automatically remove these messages from your inbox.

UNSUPERVISED LEARNING

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labelled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns from the huge amount of data .

3.6 USED PYTHON PACKAGES:

SKlearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NUMPY:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

PANDAS:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

MATPLOTLIB:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

CHAPTER 4

PROJECT DESIGN AND IMPLEMENTATION

In proposed System, we are applying Random forest algorithm, Bernoulli Naive Bayes Model and Logistic Regression algorithm for classification of the credit card dataset. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of over fitting to their training set. ,A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting. We propose a Machine learning model to detect fraudulent credit card activities in online financial transactions. Analysing fake transactions manually is impracticable due to vast amounts of data and its complexity. However, adequately given informative features, could make it is possible using Machine Learning. This hypothesis will be explored in the paper. To classify fraudulent and legitimate credit card transaction by supervised learning Algorithm such as Random forest. To help us to get awareness about the fraudulent and without loss of any financially.

4.1 LIST OF MODULES

Data Pre-processing1

Data Preparing

Data Modeling

Data Analysis of Visualization

Evaluation and deployment

4.2 MODULE DESCRIPTION:

DATA PRE-PROCESSING :

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to finetune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a timeconsuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

```
'data.frame': 284807 obs. of 31 variables:
 $ Time : num 0 0 1 1 2 2 4 7 7 9 ...
 $ V1 : num -1.36 1.192 -1.358 -0.966 -1.158 ...
 $ V2 : num -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
 $ V3 : num 2.536 0.166 1.773 1.793 1.549 ...
 $ V4 : num 1.378 0.448 0.38 -0.863 0.403 ...
 $ V5 : num -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
 $ V6 : num 0.4624 -0.0824 1.8005 1.2472 0.0959 ...
 $ V7 : num 0.2396 -0.0788 0.7915 0.2376 0.5929 ...
 $ V8 : num 0.0987 0.0851 0.2477 0.3774 -0.2705 ...
 $ V9 : num 0.364 -0.255 -1.515 -1.387 0.818 ...
 $ V10 : num 0.0908 -0.167 0.2076 -0.055 0.7531 ...
 $ V11 : num -0.552 1.613 0.625 -0.226 -0.823 ...
 $ V12 : num -0.6178 1.0652 0.0661 0.1782 0.5382 ...
 $ V13 : num -0.991 0.489 0.717 0.508 1.346 ...
 $ V14 : num -0.311 -0.144 -0.166 -0.288 -1.12 ...
 $ V15 : num 1.468 0.636 2.346 -0.631 0.175 ...
 $ V16 : num -0.47 0.464 -2.89 -1.06 -0.451 ...
 $ V17 : num 0.208 -0.115 1.11 -0.684 -0.237 ...
 $ V18 : num 0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
 $ V19 : num 0.404 -0.146 -2.262 -1.233 0.803 ...
 $ V20 : num 0.2514 -0.0691 0.525 -0.208 0.4085 ...
 $ V21 : num -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
 $ V22 : num 0.27784 -0.63867 0.77168 0.00527 0.79828 ...
 $ V23 : num -0.11 0.101 0.909 -0.19 -0.137 ...
 $ V24 : num 0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
 $ V25 : num 0.129 0.167 -0.328 0.647 -0.206 ...
 $ V26 : num -0.189 0.126 -0.139 -0.222 0.502 ...
 $ V27 : num 0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
 $ V28 : num -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
 $ Amount: num 149.62 2.69 378.66 123.5 69.99 ...
 $ Class : int 0 0 0 0 0 0 0 0 0 0 ...
```

FIGURE 4.1 Dataset Structure

The second figure shows the distribution of the class, the red bar which contains 284,315 variables represents the non-fraudulent transactions, and the blue bar with 492 variables represents the fraudulent transactions.

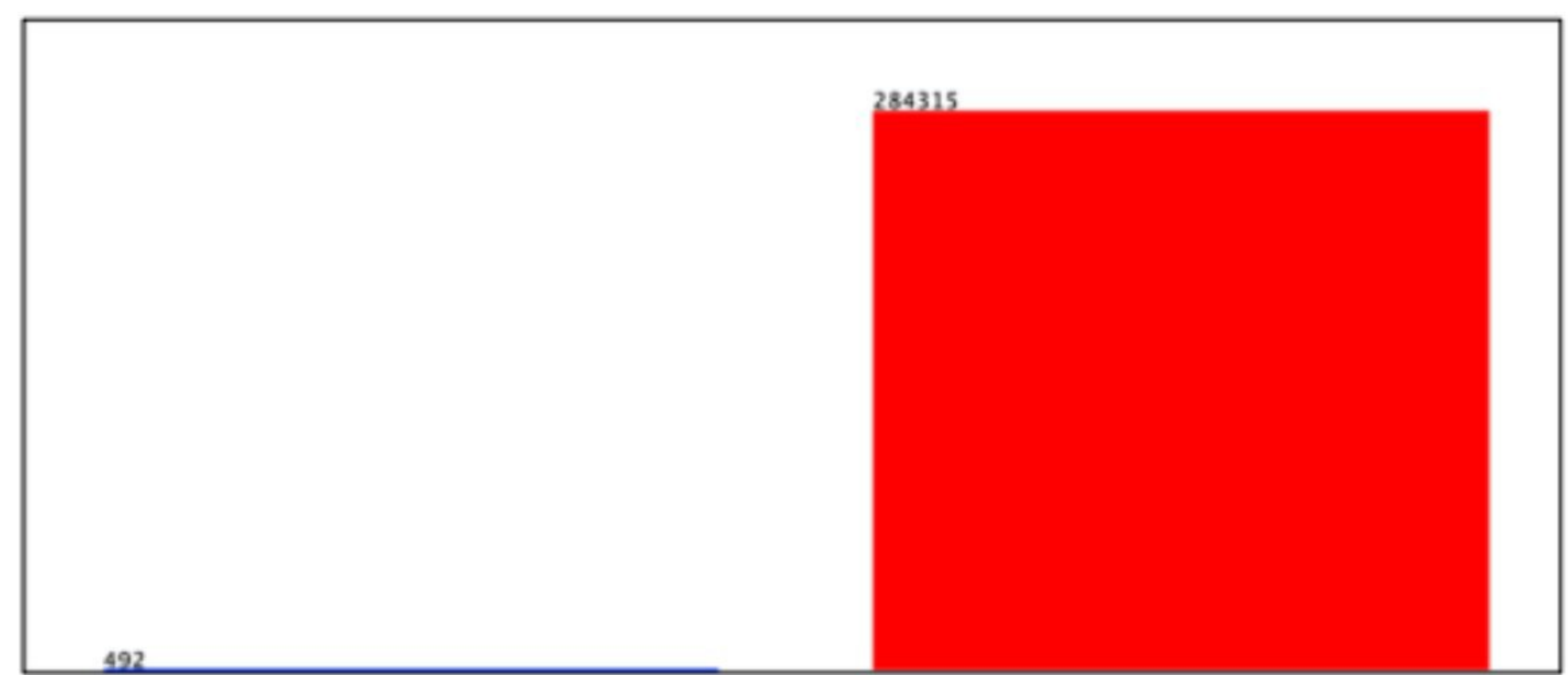


FIGURE 4.2 Class distribution

The correlations between all the of the attributes within the dataset are presented in the figure below

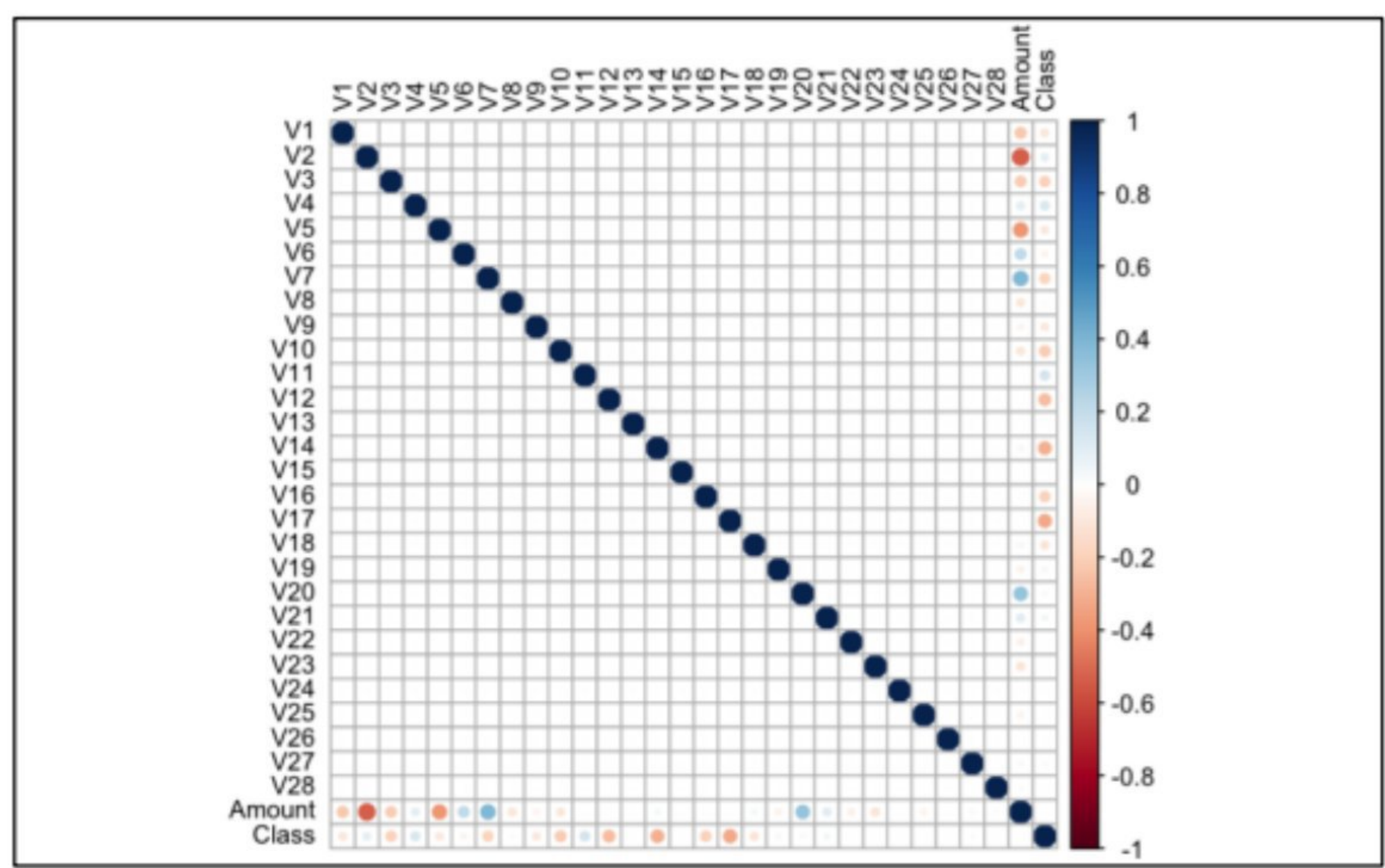


FIGURE 4.3 Correlation

Figure 4 below shows attribute 18 the attribute with the most credit card fraudulent transactions, the blue line represents the variable 1 which is the fraudulent transactions.

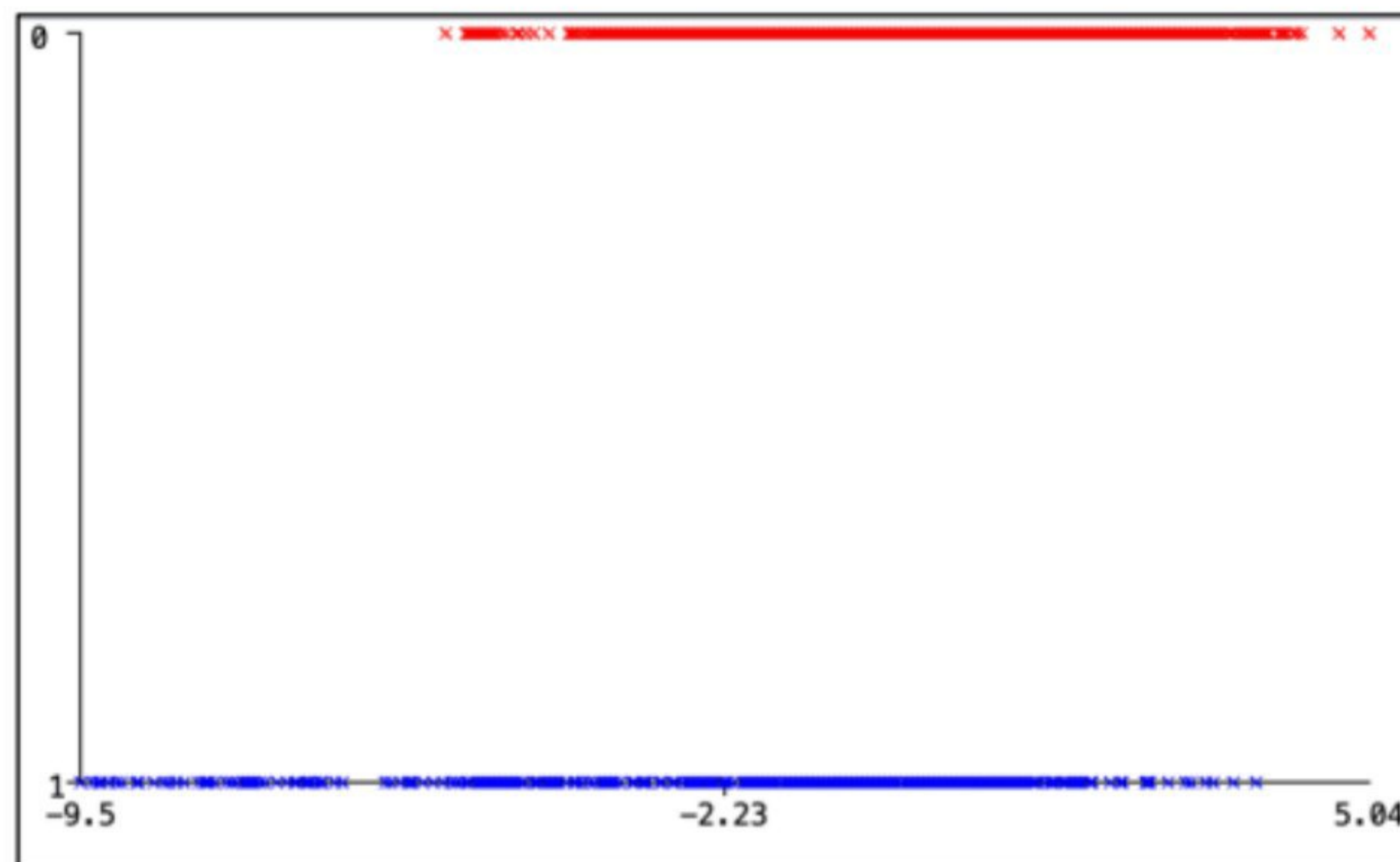


FIGURE 4.4 Attributes of classification

DATA VALIDATION/ CLEANING/PREPARING PROCESS :

Importing the library packages with loading given dataset. To analysing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Data Modeling

After making sure that the data is ready to get modeled the four models were created using both Weka and R. the model SVM was created using Weka only, as for KNN, Logistic Regression and NaïveBayes they were created using R and Weka.

KNN

The K-Nearest Neighbour algorithm (KNN) is a supervised ML technique that can be applied in both scenario instances, classification instances along with regression instances.To figure the best KNN model two Ks where used K=3 and K=7, both are presented with figures from both Weka and R.During the making of the KNN model, I decided to create two models where K=3 and K=7. Figure 5 shows the model created in R, the model scored an accuracy of 99.83% and managed to correctly identify 91,719 transactions and missed 155. As for the Weka program the model scored 99.94% for the accuracy and miss-classified 52 transactions. As there are different accuracies the average of the accuracies is 99.89%.

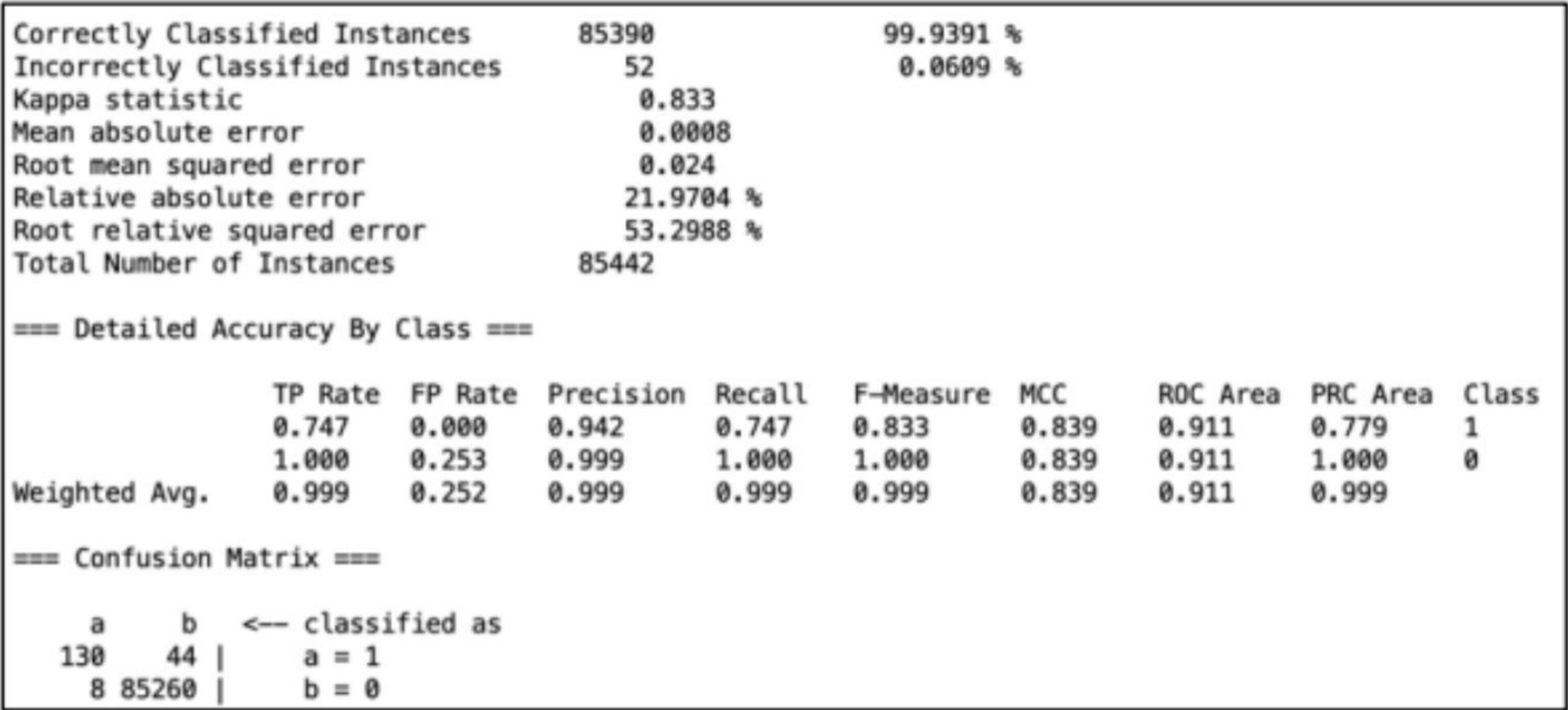


FIGURE 4.5 KNN classification

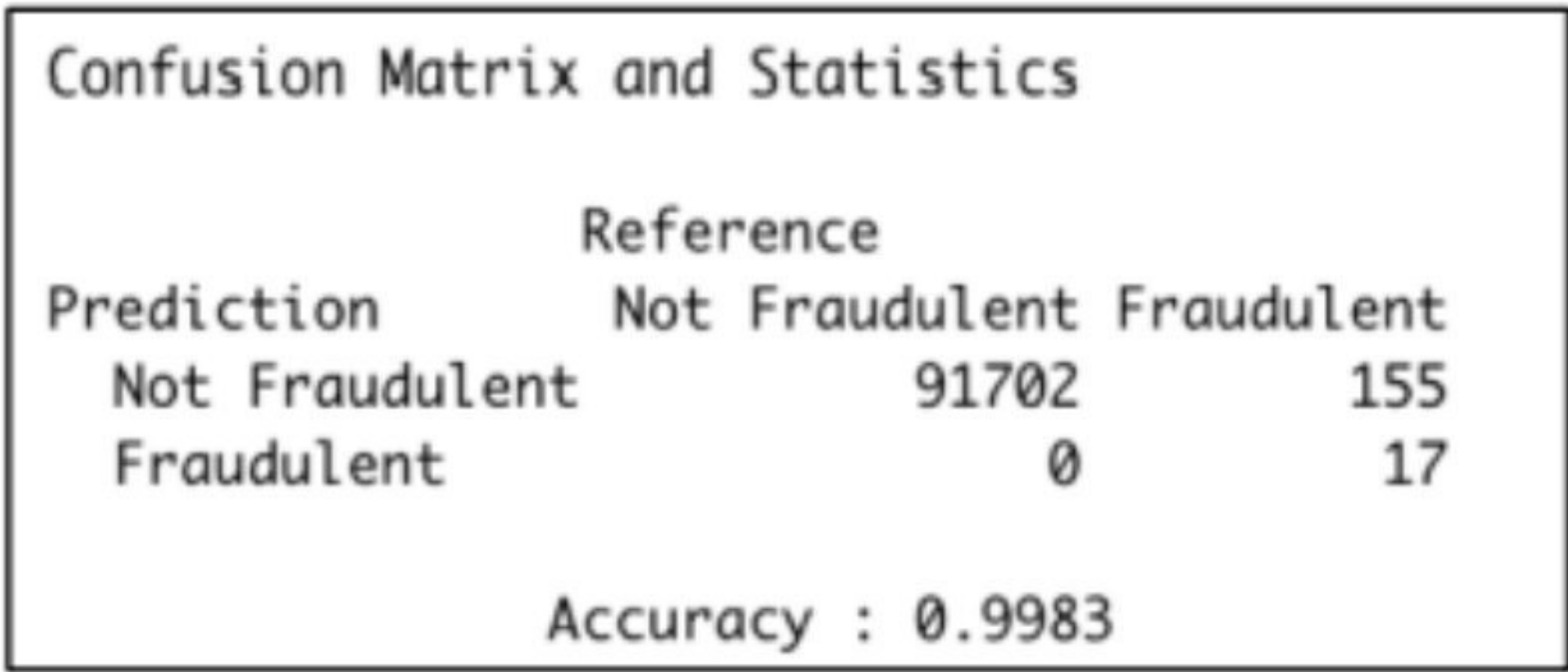


FIGURE 4.6 KNN confusion matrix

K = 7

There was a slight decrease in the accuracy in the model created in R (Figure 6) as it scored 99.82% when K is 7, and the model miss classified 166 fraudulent transactions as nonfraudulent. As for Weka (Figure 7) the accuracy is the same as K=3 99.94% with 52 misclassified transactions, the only difference is within the classifications. The average of the accuracies is 99.88%

Correctly Classified Instances	85390	99.9391 %							
Incorrectly Classified Instances	52	0.0609 %							
Kappa statistic	0.8351								
Mean absolute error	0.0008								
Root mean squared error	0.0235								
Relative absolute error	22.256 %								
Root relative squared error	52.1008 %								
Total Number of Instances	85442								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.759	0.000	0.930	0.759	0.835	0.839	0.917	0.795	1
	1.000	0.241	1.000	1.000	1.000	0.839	0.917	1.000	0
Weighted Avg.	0.999	0.241	0.999	0.999	0.999	0.839	0.917	0.999	
=== Confusion Matrix ===									
a	b	<-- classified as							
132	42	a = 1							
10	85258	b = 0							

FIGURE 4.7 KNN CLASSIFIER

Confusion Matrix and Statistics		
Prediction	Reference	
	Not Fraudulent	Fraudulent
Not Fraudulent	91702	166
Fraudulent	0	6
Accuracy : 0.9982		

FIGURE 4.8 CONFUSION MATRIX

Naïve Bayes

Naïve Bayes is a classification algorithm that consider the being of a certain trait within a class is unrelated to the being of any different feature, the main use of it is for clustering and classifications, depending on the conditional probability of happening. The second model created by R is Naïve Bayes, figure 9 shows the performance of the model, it scored an accuracy of 97.77% and misclassified a total of 2,051 transactions, 33 fraudulent as nonfraudulent and 2018 nonfraudulent as fraudulent. There is a slight difference in the accuracy of the Naïve bayes model created within Weka as its 97.73% and the misclassification instances are 1,938.

Correctly Classified Instances	83504	97.7318 %							
Incorrectly Classified Instances	1938	2.2682 %							
Kappa statistic	0.1292								
Mean absolute error	0.0227								
Root mean squared error	0.1491								
Relative absolute error	626.539 %								
Root relative squared error	330.6127 %								
Total Number of Instances	85442								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.851	0.022	0.072	0.851	0.132	0.243	0.968	0.091	1
	0.978	0.149	1.000	0.978	0.989	0.243	0.964	1.000	0
Weighted Avg.	0.977	0.149	0.998	0.977	0.987	0.243	0.964	0.998	
=== Confusion Matrix ===									
a	b	<-- classified as							
148	26	a = 1							
1912	83356	b = 0							

FIGURE 4.9 NAÏVE BAYES CLASSIFIER

Confusion Matrix and Statistics		
Prediction	Reference	
	Not Fraudulent	Fraudulent
Not Fraudulent	89684	33
Fraudulent	2018	139
Accuracy : 0.9777		

FIGURE 4.10 CONFUSION MATRIX

Logistic Regression

Logistic Regression model is statical model where evaluations are formed of the connection among dependent qualitative variable (binary or binomial logistic regression) or variable with three values or higher (multinomial logistic regression) and one independent explanatory variable or higher whether qualitative or quantitative,.The last model created using both R and Weka is Logistic Regression, the model managed to score and accuracy of 99.92% in R (figure 11) with 70 misclassified instances, while it scored 99.91% in Weka with 77 misclassified instances as presented in figure 10.

Correctly Classified Instances	85365	99.9099 %							
Incorrectly Classified Instances	77	0.0901 %							
Kappa statistic	0.7256								
Mean absolute error	0.0014								
Root mean squared error	0.0276								
Relative absolute error	38.6516 %								
Root relative squared error	61.1796 %								
Total Number of Instances	85442								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.586	0.000	0.953	0.586	0.726	0.747	0.976	0.831	1
	1.000	0.414	0.999	1.000	1.000	0.747	0.976	1.000	0
Weighted Avg.	0.999	0.413	0.999	0.999	0.999	0.747	0.976	1.000	
=== Confusion Matrix ===									
a	b	<-- classified as							
102	72	a = 1							
5	85263	b = 0							

FIGURE 4.11 Logical regression classifier

	FALSE	TRUE
0	91693	9
1	61	111
[1]	99.92381	

FIGURE 4.12 confusion matrix

Support Vector Machine

Support Vector machine is a supervised ML technique with connected learning algorithms which inspect data used for both classification and regression analyses, it also performs linear classification, additionally to non-linear classification by creating margins between the classes, which are created in such a fashion that the space between the margin and the classes is maximum which minimizes the error of the

classification. Finally, the model Support Vector Machine as show in figure 12 managed to score 99.94% for the accuracy and misclassified 51 instances.

Correctly Classified Instances	85391	99.9403 %
Incorrectly Classified Instances	51	0.0597 %
Kappa statistic	0.8388	
Mean absolute error	0.0006	
Root mean squared error	0.0244	
Relative absolute error	16.4433 %	
Root relative squared error	54.1917 %	
Total Number of Instances	85442	
=== Detailed Accuracy By Class ===		
	TP Rate	FP Rate
	0.764	0.000
	1.000	0.236
Weighted Avg.	0.999	0.235
	Precision	Recall
	0.930	0.764
	1.000	1.000
	0.999	0.999
	F-Measure	MCC
	0.839	0.843
	1.000	0.843
	0.999	0.843
	ROC Area	PRC Area
	0.882	0.711
	0.882	1.000
	0.882	0.999
	Class	
	1	
	0	
=== Confusion Matrix ===		
a	b	← classified as
133	41	a = 1
10	85258	b = 0

FIGURE 4.13 SVM CLASSIFIER

EXPLORATION DATA ANALYSIS OF VISUALIZATION:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

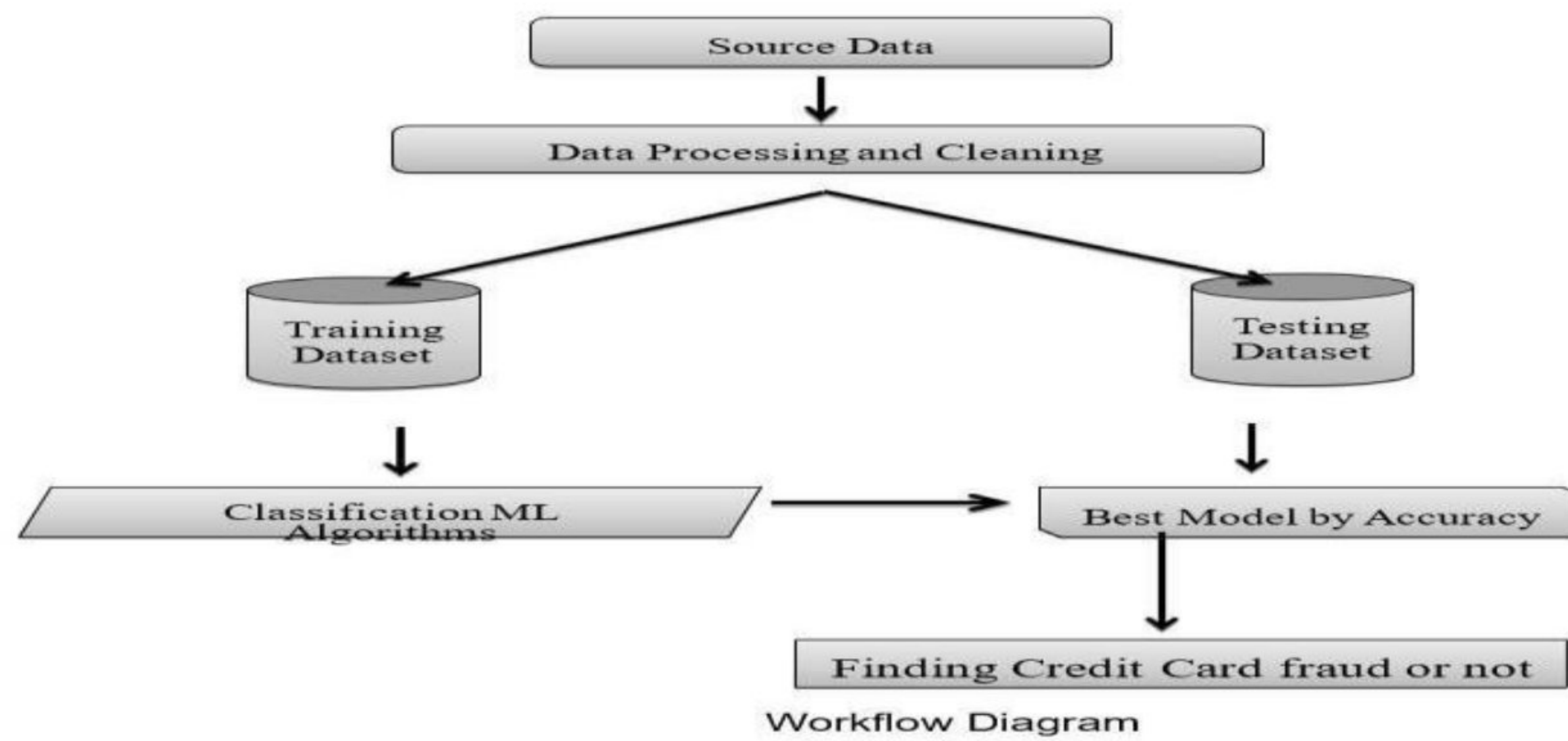


FIGURE 4.14 SYSTEM ARCHITECTURE

Evaluation and Deployment

The last stage of the model is the evaluation and deployment stage, as presented in table 2 below all models are being compared to each other to figure the best model in identifying fraudulent credit card transactions. Accuracy is the overall number of instances that are predicted correctly, accuracies are represented by confusion matrix where it showed the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive represents the transactions that are fraudulent and was correctly classified by the model as fraudulent. True Negative represents the not fraudulent transactions that were correctly predicted by the model as Not fraudulent. The third rating is False positive which represents the transaction that are fraudulent but was misclassified as not fraudulent. And finally False Negative which are the not fraudulent transactions that were identified as fraudulent.

CHAPTER 5

RESULTS AND DISCUSSIONS

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The table above shows all the components to calculate an accuracy of a model which is displayed in the below equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Model		Accuracy
KNN	K = 3	99.89%
	K = 3	
	K = 7	99.88%
	K = 7	
Naïve Bayes	Naïve Bayes	97.76%
	Naïve Bayes	
Logistic Regression	Logistic Regression	99.92%
	Logistic Regression	
Support Vector Machine	SVM	99.94%

Table shows all of the accuracies of all the models that were created in the project, all models performed well in detecting fraudulent transactions and managed to score high accuracies. Out of all the models the model that scored the best is Support Vector Machine as its accuracy is 99.94%, the second best is Logistic Regression, then in third place is KNN as both Ks scored similar accuracies, and the model that scored the lowest accuracy out of all models is Naïve Bayes with a score of 97.76%.

CHAPTER 6

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of credit card fraud or not. The main objective of this project was to find the most suited model in credit card fraud detection in terms of the machine learning techniques chosen for the project, and it was met by building the four models and finding the accuracies of them all, the best model in terms of accuracies is Support Vector Machine which scored 99.94% with only 51 misclassified instances. I believe that using the model will help in decreasing the amount of credit card fraud and increase the customers satisfaction as it will provide them with better experience in addition to feeling secure. Since human verification is often necessary and the payment processing time is limited, the second task manages the largest volume of transactions. In the literature, fraud detection challenges and algorithms performance are widely studied but the very formulation of the problem is never disrupted: it aims at predicting if a transaction is fraudulent based on its characteristics and the past transactions of the cardholder. Yet, in posterior detection, verification often takes days, so new payments on the card become available before a decision is taken. This is our motivation to propose a new paradigm: posterior fraud detection with "future" information. We start by providing evidence of the on-time availability of subsequent transactions, usable as extra context to improve detection. We then design a Bidirectional LSTM to make use of these transactions. On a real-world dataset with over 30 million transactions, it achieves higher performance than a regular LSTM, which is the state-of-the-art classifier for fraud detection that only uses the past context. We also introduce new metrics to show that the proposal catches more frauds, more compromised cards, and based on their earliest frauds. We believe that future works on this new paradigm will have a significant impact on the detection of compromised cards.

6.1 FUTURE ENHANCEMENTS

Detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

REFERENCES:

- [1] Q. Wu, M. Zhou, Q. Zhu, Y. Xia, and J. Wen, —MOELS: Multiobjective evolutionary list scheduling for cloud workflows,|| *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 166– 176, Jan. 2020.
- [2] L. Huang, M. Zhou, and K. Hao, —Non-dominated immune-endocrine short feedback algorithm for multi-robot maritime patrolling,|| *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 362–373, Jan. 2020.
- [3] X. Wang, K. Xing, C.-B. Yan, and M. Zhou, —A novel MOEA/D for multiobjective scheduling of flexible manufacturing systems,|| *Complexity*, vol. 2019, pp. 1–14, Jun. 2019.
- [4] J. J. Liang, S. T. Ma, B. Y. Qu, and B. Niu, —Strategy adaptive memetic crowding differential evolution for multimodal optimization,|| in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2012, pp. 1–7.
- [5] M. M. H. Ellabaan and Y. S. Ong, —Valley-adaptive clearing scheme for multimodal optimization evolutionary search,|| in *Proc. 9th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2009, pp. 1–6.
- [6] X. Li, —Efficient differential evolution using speciation for multimodal function optimization,|| in *Proc. Conf. Genetic Evol. Comput. - GECCO*, Jun. 2005, pp. 873–880.
- [7] Y. Feng *et al.*, —Target disassembly sequencing and scheme evaluation for CNC machine tools using improved multiobjective ant colony algorithm and fuzzy integral,|| *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 12, pp. 2438–2451, Dec. 2019.
- [8] L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, —Two-level master– slave RFID networks planning via hybrid multi objective artificial bee colony optimizer,|| *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 5, pp. 861– 880, May 2019.
- [9] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, —A network reduction-based multi objective evolutionary algorithm for community detection in large-scale complex networks,|| *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 703–716, Feb. 2020.

APPENDIX

Source Code

```
#import library packages import pandas as p import matplotlib.pyplot as plt import  
seaborn as s import numpy as n
```

```
In [ ]:
```

```
import warnings
```

```
warnings.filterwarnings('ignore')
```

```
In [ ]:
```

```
#Load given dataset  
data = p.read_csv('creditcard.csv')
```

```
In [ ]:
```

```
del data['TransactionDate'] del data['Merchant_id']
```

```
In [ ]:
```

```
df=data.dropna()
```

```
In [ ]:
```

```
from sklearn.preprocessing import LabelEncoder
```

```
var_mod = ['Is_declined', 'isForeignTransaction', 'isHighRiskCountry', 'isFradulent'] le  
= LabelEncoder() for i in var_mod:
```

```
    df[i] = le.fit_transform(df[i]).astype(int)
```

```
In [ ]:
```

```
#preprocessing, split test and dataset, split response variable  
X = df.drop(labels='isFradulent', axis=1)
```

```
#Response variable
```

```
y = df.loc[:, 'isFradulent']
```

```
In [ ]:
```

```
""We'll use a test size of 30%. We also stratify the split on the response variable, which  
is very important to do because there are so few fraudulent transactions""
```



```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=1,
stratify=y) print("Number of training dataset: ", len(X_train)) print("Number of test
dataset: ", len(X_test))
```

```
print("Total number of dataset: ", len(X_train)+len(X_test))
```

In []:

```
#According to the cross-validated MCC scores, the random forest is the bestperforming
model, so now let's evaluate its performance on the test set.
```

```
from sklearn.metrics import confusion_matrix, classification_report,
```

```
matthews_corrcoef, cohen_kappa_score, accuracy_score, average_precision_score,
roc_auc_score
```

In []:

```
from sklearn.metrics import accuracy_score, confusion_matrix from
sklearn.linear_model import LogisticRegression from sklearn.model_selection import
cross_val_score
```

```
logR= LogisticRegression()
```

```
logR.fit(X_train,y_train)
```

```
predictLR = logR.predict(X_test)
```

```
print("")
```

```
print('Classification report of Logistic Regression Results:') print("")
```

```
print(classification_report(y_test,predictLR))
```

```
print("")
```

```
cm1=confusion_matrix(y_test,predictLR)
```

```
print('Confusion Matrix result of Logistic Regression is:\n',cm1) print("")
```

```
sensitivity1 = cm1[0,0]/(cm1[0,0]+cm1[0,1])
```

```
print('Sensitivity : ', sensitivity1 ) print("")
```



```
specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])
```

```
print('Specificity : ', specificity1) print("")
```

```
accuracy = cross_val_score(logR, X, y, scoring='accuracy')
```

```
print('Cross validation test results of accuracy:') print(accuracy)
```

```
#get the mean of each fold print("")
```

```
print("Accuracy result of Logistic Regression is:",accuracy.mean() * 100)
```

```
LR=accuracy.mean() * 100
```

In []:

```
def graph(): import matplotlib.pyplot as plt    data=[LR]    alg="Logistic Regression"  
plt.figure(figsize=(5,5))    b=plt.bar(alg,data,color="c")
```

```
    plt.title("Accuracy comparison of Earth Quake",fontsize=15)  
plt.legend(b,data,fontsize=9)
```

In []:

```
graph()
```

In []:

```
TP = cm1[0][0]
```

```
FP = cm1[1][0]
```

```
FN = cm1[1][1] TN = cm1[0][1] print("True Positive :",TP) print("True Negative :",TN)  
print("False Positive :",FP) print("False Negative :",FN)
```

```
print("")
```

```
TPR = TP/(TP+FN)
```

```
TNR = TN/(TN+FP)
```

```
FPR = FP/(FP+TN) FNR = FN/(TP+FN)
```

```
print("True Positive Rate :",TPR) print("True Negative Rate :",TNR) print("False  
Positive Rate :",FPR)
```

```
print("False Negative Rate :",FNR) print("")
```

```
PPV = TP/(TP+FP) NPV = TN/(TN+FN)
```



```
print("Positive Predictive Value :",PPV) print("Negative predictive value :",NPV)
```

In []:

```
def plot_confusion_matrix(cm1, title='Confusion matrix-Logistic_Regression',  
cmap=plt.cm.Blues):
```

```
    target_names=['Predict','Actual']
```

```
    plt.imshow(cm1, interpolation='nearest', cmap=cmap)    plt.title(title)
```

```
    plt.colorbar()
```

```
    tick_marks = np.arange(len(target_names))    plt.xticks(tick_marks, target_names,  
rotation=45)    plt.yticks(tick_marks, target_names)
```

```
    plt.tight_layout()    plt.ylabel('True label')
```

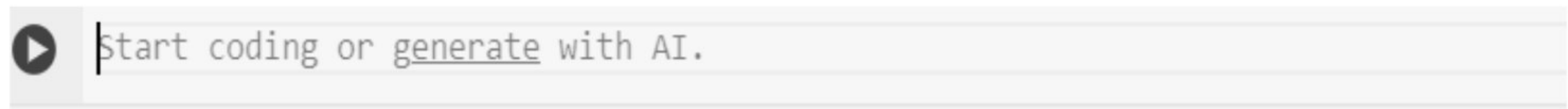
```
    plt.xlabel('Predicted label')
```

```
cm1=confusion_matrix(y_test, predictLR)    print('Confusion matrix-  
Logistic_Regression:') print(cm1)
```

```
plot_confusion_matrix(cm1)
```

In []:

PROJECT SCREENSHOTS



importing dependencies

```
[ ] import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Importing dependencies

```
[ ] #loading the data
cd = pd.read_csv ('/content/creditcard.csv')

[ ] # first five rows of the data set
cd.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153

5 rows x 31 columns

Columns

```
# dataset information
cd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180351 entries, 0 to 180350
Data columns (total 31 columns):
 #   Column      Non-Null Count  Dtype
---  ---
 0   Time        180351 non-null float64
 1   V1          180351 non-null float64
 2   V2          180351 non-null float64
 3   V3          180351 non-null float64
 4   V4          180351 non-null float64
 5   V5          180351 non-null float64
 6   V6          180351 non-null float64
 7   V7          180351 non-null float64
 8   V8          180351 non-null float64
 9   V9          180350 non-null float64
10  V10         180350 non-null float64
11  V11         180350 non-null float64
12  V12         180350 non-null float64
13  V13         180350 non-null float64
14  V14         180350 non-null float64
15  V15         180350 non-null float64
16  V16         180350 non-null float64
17  V17         180350 non-null float64
18  V18         180350 non-null float64
19  V19         180350 non-null float64
20  V20         180350 non-null float64
21  V21         180350 non-null float64
22  V22         180350 non-null float64
23  V23         180350 non-null float64
24  V24         180350 non-null float64
```

Dataset Information


```
# dataset information
cd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180351 entries, 0 to 180350
Data columns (total 31 columns):
 #   Column  Non-Null Count  Dtype  
---  -
 0   Time    180351 non-null  float64
 1   V1       180351 non-null  float64
 2   V2       180351 non-null  float64
 3   V3       180351 non-null  float64
 4   V4       180351 non-null  float64
 5   V5       180351 non-null  float64
 6   V6       180351 non-null  float64
 7   V7       180351 non-null  float64
 8   V8       180351 non-null  float64
 9   V9       180350 non-null  float64
10  V10      180350 non-null  float64
11  V11      180350 non-null  float64
```

Data type values

```
[ ] # distribution of legit and fraud transactions
cd['Class'].value_counts()
```

```
0.0    179986
1.0      364
Name: Class, dtype: int64
```

Distribution

Classification report of Logistic Regression Results:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	789
1	0.97	0.96	0.97	134
accuracy			0.99	923
macro avg	0.98	0.98	0.98	923
weighted avg	0.99	0.99	0.99	923

Confusion Matrix result of Logistic Regression is:

```
[[785  4]
 [  5 129]]
```

Sensitivity : 0.9949302915082383

Specificity : 0.9626865671641791

Cross validation test results of accuracy:

```
[0.9902439 0.98211382 0.9804878 0.98211382 0.98699187]
```

Accuracy result of Logistic Regression is: 98.4390243902439

J. Sriram
Department of IT
Sathyabama Institute of Science
and Technology Chennai, India
jothisriram02gmail.com

P. Sudharsan
Department of IT
Sathyabama Institute of Science
and Technology Chennai, India
itsfanofsudhar@gmail.com

T.G. Ruby angel
Department of IT
Sathyabama Institute of Science
and Technology Chennai, India
tgrubyangel@gmail.com

Abstract: Credit card usage has become widespread for online transactions due to its efficiency and user-friendly features. However, the surge in credit card usage has also led to an increase in the potential for misuse, resulting in significant financial losses for both cardholders and financial institutions. The primary objective of this research study is to develop effective methods for detecting credit card fraud, taking into account various factors such as the availability of public data, the challenge of high-class imbalance in the data, the evolving nature of fraud, and the issue of high false alarm rates. The research acknowledges the challenges posed by the imbalance in the data (where legitimate transactions significantly outnumber fraudulent ones), the dynamic nature of fraud, and the need to minimize false alarms. The proposed machine learning methods aim to enhance the accuracy and efficiency of credit card fraud detection, contributing to the overall security of online financial transactions.

Keywords : Artificial Intelligence, Machine Learning, Data science, Natural Language Processing, Python Programming Language, Graphical User Interface.

I. Introduction

Credit card usage, while convenient, is not universally advantageous and can lead to substantial financial losses for individuals. The escalating prevalence of credit card fraud is a growing concern, particularly in the digital age where internet transactions are becoming increasingly common. The surge in online purchases, fuel by the widespread use of credit and debit cards, has created an environment ripe for various types of fraudulent activities. Given this scenario, credit card companies face the imperative task of detecting and preventing fraudulent transactions. The spectrum of fraudulent activities spans a multitude of situations and methods, posing a risk to anyone, anywhere engaging in financial transactions. To address this challenge, the incorporation of Data Science and machine learning has become instrumental in identifying and mitigating fraudulent transactions. These advanced technologies play a crucial role in scrutinizing and analyzing patterns within transaction data, aiding in the early detection of activities that indicate potential fraud. By leveraging the power of machine learning, financial institutions can make informed decisions regarding the approval or denial of credit card applications, thereby enhancing the security of financial transactions and protecting customers from unauthorized purchases they did not make.

In essence, the integration of Data Science and machine learning serves as a proactive measure to combat the rising tide of credit card fraud. It empowers financial institutions to stay ahead of fraudulent activities, providing a layer of defence that contributes to making digital transactions more secure and reliable for consumers.

II. Literature Review

A literature review is a comprehensive body of text that aims to examine and present the key aspects of current knowledge or methodological approaches related to a specific topic. It primarily relies on secondary sources, discussing information published within a particular subject area and sometimes within a defined time frame. The primary objective is to update the reader on the existing literature pertaining to a subject, serving as a foundation for subsequent goals such as suggesting potential future research directions. It typically precedes a research proposal and can range from a straightforward summary of sources to a more intricate combination of summary and synthesis.

1.] Title: Credit Card Fraud Detection Using Machine Learning: A Comprehensive Literature Review

Author: Harish Paruchuri

Year: 2017

In the quest to provide enhanced services to customers, companies have embraced the convenience of online purchasing. However, this convenience also presents an opportunity for criminals to engage in fraudulent activities. Criminals may exploit vulnerabilities to steal the information of

cardholders, subsequently using it for unauthorized online transactions until the cardholder discovers the fraud and contacts the bank to block the card. This paper delves into various machine learning algorithms utilized for the detection of such fraudulent transactions. The research highlights Credit Card Fraud (CCF) as a significant issue within the financial sector, escalating over time.

2.] Title: Research on Credit Card Fraud Detection Systems

Author: Devika S P, Nisarga K S, Gagana P Rao, Chandini S B, Rajkumar N

Year: 2019

In contemporary times, credit cards have gained widespread popularity among both private and public employees, serving as a means to make online purchases of consumable durable products and facilitating fund transfers between accounts. However, the prevalence of fraudsters exploiting user transaction behaviours for illegal activities, such as phishing and Trojan viruses, poses a significant threat to users' sensitive information. This paper addresses the escalating issue of credit card fraudulent activities, which has become a major concern for credit card users. The fraudulent activities not only result in financial losses but also jeopardize the security of users' sensitive information.

3.] Title: Review of Credit Card Fraud Detection Using Machine Learning

Author: Suresh K Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil

Year: 2019

In recent years, credit card fraud has emerged as a significant and growing problem, leading to substantial financial losses for individuals, merchants, and banks alike. Recognizing the potential of machine learning as a successful technique in fraud identification, this paper provides a comprehensive review of various fraud detection methods employing machine learning algorithms. The evaluation of these methods is conducted using key performance measures such as accuracy, precision, and specificity. The paper introduces a Fraud Detection System (FDS) that leverages the supervised Random Forest algorithm, aiming to enhance the accuracy of fraud detection in credit card transactions. The proposed system incorporates a learning-to-rank approach to prioritize alerts, effectively addressing issues related to concept drift in fraud detection. By utilizing this proposed system, the accuracy of identifying fraudulent activities in credit card transactions is notably improved.

4.] Title: Credit Card Fraud Detection and Prevention Using Machine Learning

Author: S. Abinayaa, H. Sangeetha, R. A. Karthikeyan, K. Saran Sriram, D. Piyush

Year: 2020

This research is primarily centered on the real-world detection of credit card fraud. The initial step involves the collection of credit card datasets, ensuring their qualification for the study. Subsequently, queries are applied to the user's credit card to test the dataset. The research employs the Random Forest algorithm as a classification method, utilizing both previously evaluated datasets and the current dataset. The final step involves optimizing the accuracy of the results data.

III. Proposed Methodology

In our proposed system, we are employing three different algorithms—Random Forest, Bernoulli Naive Bayes Model, and Logistic Regression—for the classification of credit card datasets. Specifically, Random Forest is a versatile algorithm designed for both classification and regression tasks.

In our proposed machine learning model, the primary objective is to identify fraudulent activities within online financial transactions involving credit cards. The manual analysis of fake transactions is deemed impractical due to the vast volume of data and its inherent complexity. By leveraging the Random Forest algorithm, among others, our model aims to efficiently analyze and classify transactions, providing a practical solution for the detection of fraudulent credit card activities in the realm of online financial transactions.

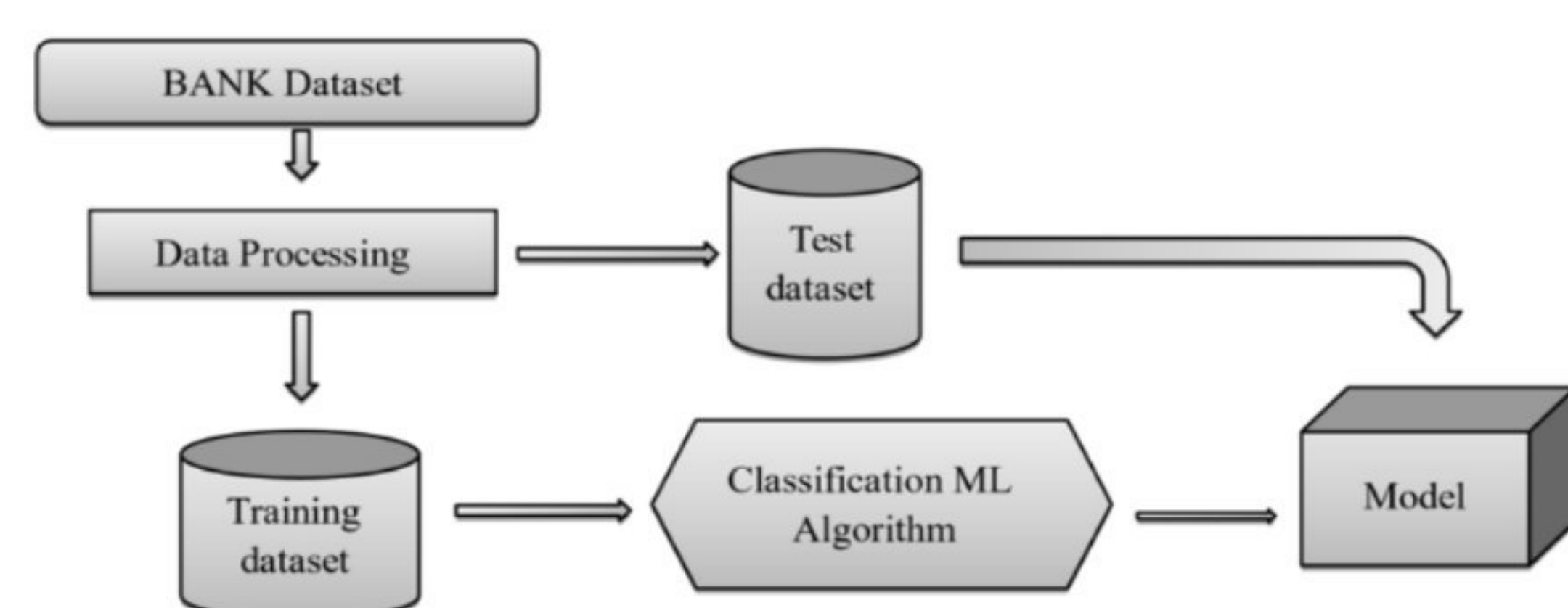


Fig 1 Block Diagram of our Proposed method

Module 1: Pre-processing

Validation techniques in machine learning serve the purpose of determining the error rate of a Machine Learning (ML) model, providing an estimate that closely mirrors the true error rate of the dataset. In cases where the dataset is extensive enough to be representative of the entire population, validation techniques may be deemed unnecessary. The validation process involves addressing various aspects,

including identifying missing values, handling duplicate entries, and examining data types such as whether variables are of the float or integer type. Ultimately, validation techniques contribute to enhancing the robustness and generalization capability of machine learning models, making them more reliable in real-world applications.

Module 2: Data Validation

The initial phase involves importing library packages and loading the provided dataset. Following this, a thorough analysis of variables is undertaken, including the identification of data shape, data types, and an assessment of missing values and duplicate entries. Additionally, a validation dataset is set aside, serving as a sample of data not utilized during model training. This dataset aids in estimating the model's skill while fine-tuning its hyperparameters and refining procedures for optimal model evaluation.. Through these processes, the dataset is refined and prepared to ensure accurate and meaningful insights during subsequent stages of the analytical pipeline.

Module 3: Data Analysis

The process begins with importing library packages and loading the provided dataset. A comprehensive examination of variables follows, involving the assessment of data shape, types, and a thorough investigation of missing values and duplicate entries. Concurrently, a validation dataset is set aside, acting as a sample excluded from model training. This dataset proves valuable for estimating model skill during the refinement of hyperparameters and evaluation procedures.

Module 4: Modeling

The implementation of the models was executed using Weka for all four algorithms, while in the R programming environment, we specifically implemented KNN and Naïve Bayes. This dual implementation approach allows us to gauge the consistency of results across different platforms and programming languages. In the subsequent sections of the paper, we will delve into the detailed results obtained from each model, utilizing a variety of performance metrics. Furthermore, visualizations will be presented, shedding light on the decision boundaries, classification outcomes, and other pertinent aspects of model behaviour

Module 5: Evaluation and Deployment

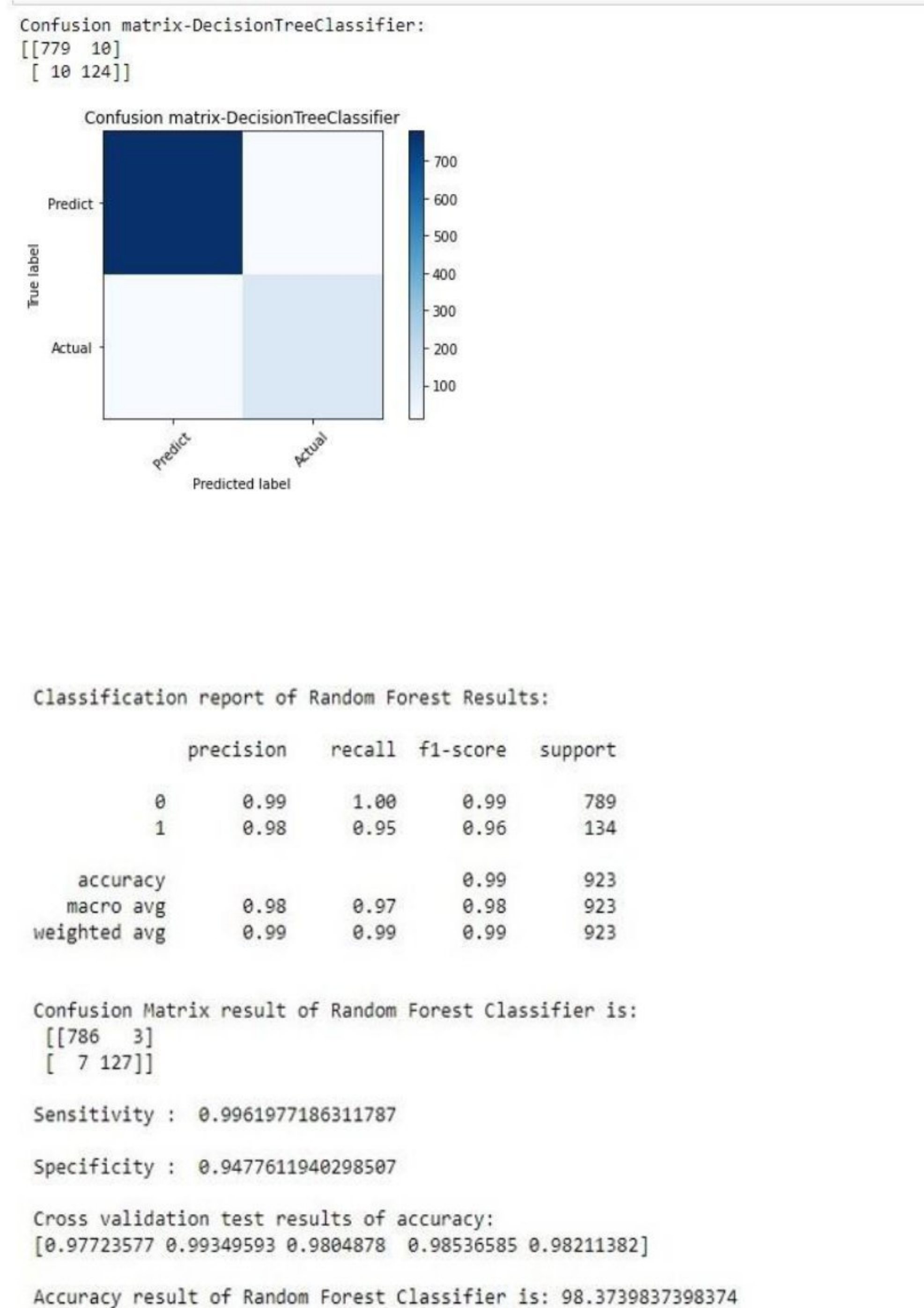
The final phase will show evaluations of the models by presenting their efficiency, the accuracies of the models will be presented in addition to any comment observed, to find the best and most suited model for detecting the fraud transactions made by credit card.

V. Results and Discussion

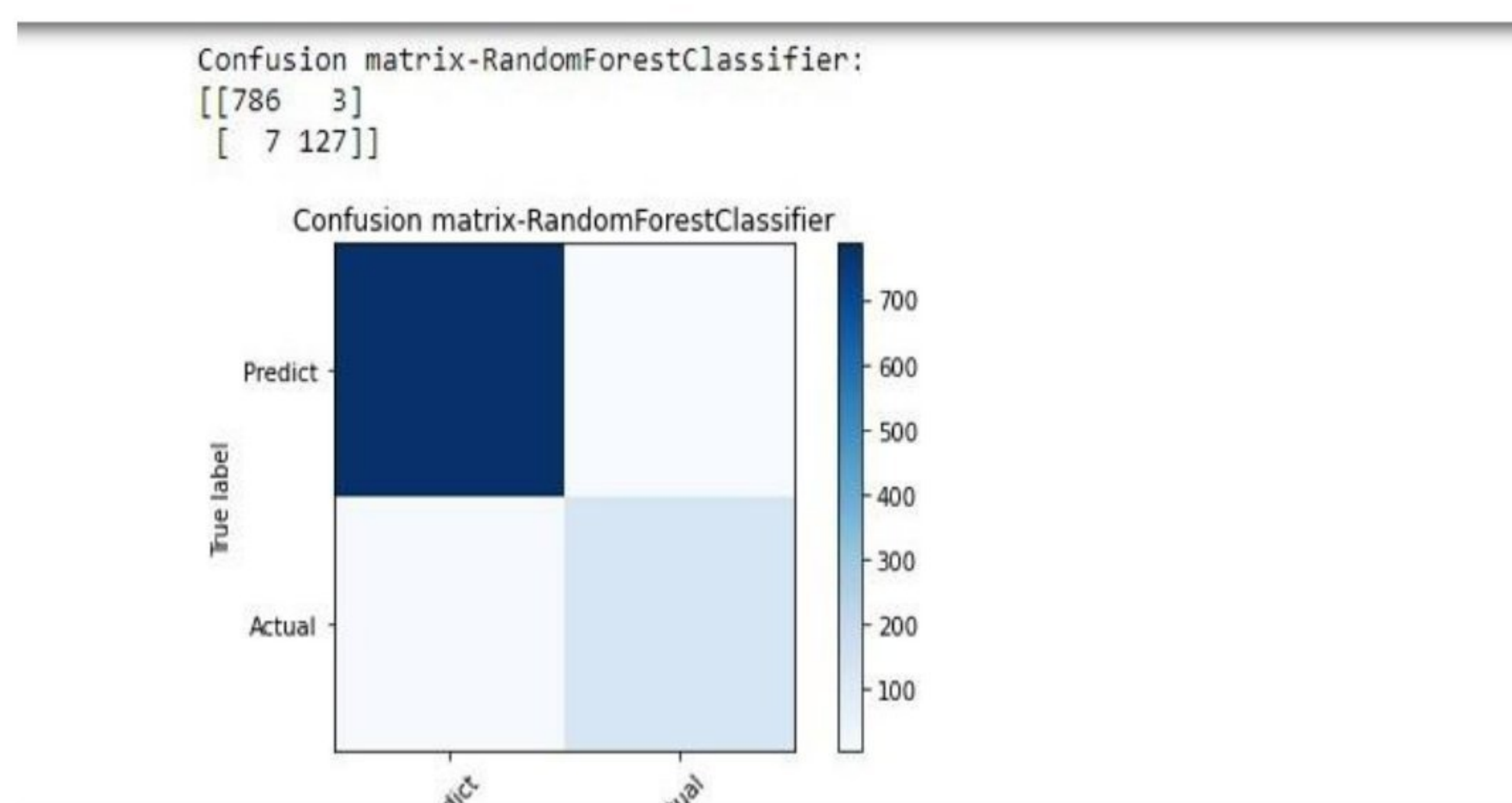
The Results and Discussion section serves as the reflective narrative of our system's response to the proposed methodology.

Classification report of Decision Tree Results:				
	precision	recall	f1-score	support
0	0.99	0.99	0.99	789
1	0.93	0.93	0.93	134
accuracy			0.98	923
macro avg	0.96	0.96	0.96	923
weighted avg	0.98	0.98	0.98	923
Confusion Matrix result of Decision Tree Classifier is:				
[[779 10]				
[10 124]]				
Sensitivity : 0.9873257287705957				
Specificity : 0.9253731343283582				
Cross validation test results of accuracy:				
[0.98373984 0.9804878 0.97723577 0.98373984 0.97398374]				
Accuracy result of Decision Tree Classifier is: 97.98373983739837				

Decision Tree



Random forest



VI. Conclusion

The analytical journey initiated with meticulous data cleaning and processing, addressing aspects such as handling missing values and ensuring data quality. Following this preparatory phase, we delved into exploratory analysis to gain insights into the dataset's characteristics and uncover potential patterns. Subsequently, the focus shifted towards model building and evaluation, where various machine learning models were employed to predict the occurrence of credit card fraud. Throughout this process, the objective was to achieve the highest accuracy on the public test set, signifying the model's proficiency in making accurate predictions. The ultimate goal of this application lies in its ability to discern whether a given credit card transaction is fraudulent or not. By optimizing for accuracy, the model aims to provide a reliable and effective tool for predicting the likelihood of credit card fraud, contributing to enhanced security measures in financial transactions.

VII. References

- 1].Q. Wu, M. Zhou, Q. Zhu, Y. Xia, and J. Ven, —MOELS: Multi objective evolutionary list scheduling for cloud workflows,|| *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 166–176, Jan. 2020.
- 2] L. Huang, M. Zhou, and K. Hao, —Non-dominated immune-endocrine short feedback algorithm for multi-robot maritime patrolling,|| *IEEE Trans.*
- [3] X. Wang, K. Xing, C.-B. Yan, and M. Zhou, —A novel MOEA/D for multiobjective scheduling of flexible

manufacturing systems,|| *Complexity*, vol. 2019, pp. 1–14, Jun. 2019.

[4] J. J. Liang, S. T. Ma, B. Y. Qu, and B. Niu, —Strategy adaptative memetic crowding differential evolution for multimodal optimization,|| in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2012, pp. 1–7.

[5] M. M. H. Ellabaan and Y. S. Ong, —Valley-adaptive clearing scheme for multimodal optimization evolutionary search,|| in *Proc. 9th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2009, pp. 1–6.

[6] X. Li, —Efficient differential evolution using speciation for multimodal function optimization,|| in *Proc. Conf. Genetic Evol. Comput. - GECCO*, Jun. 2005, pp. 873–880.

[7] Y. Feng *et al.*, —Target disassembly sequencing and scheme evaluation for CNC machine tools using improved multiobjective ant colony algorithm and fuzzy integral,|| *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 12, pp. 2438–2451, Dec. 2019.

[8] L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, —Two-level master-slave RFID networks planning via hybrid multiobjective artificial bee colony optimizer,|| *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 5, pp. 861–880, May 2019.

[9] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, —A network reduction-based multiobjective evolutionary algorithm for community detection in large-scale complex networks,|| *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 703–716, Feb. 2020.