djdmchm

by TURNITIN REPORT

Submission date: 14-Apr-2021 01:27PM (UTC-0400)

Submission ID: 1559169249

File name: ICA-CIS4035-N.docx (674.67K)

Word count: 2725

Character count: 15588

Case Study Report on Using Python and Machine Learning Methods

Name of the Student:

Name of the University:

Word Count: 2045

Abstract

The proposed project is based on case studies selected as credit card data. The data set was taken from the Kaggle dataset library and the prediction was made to detect fraud using various machine learning and deep learning methods such as Logistic Regression, Multi-layer Perception, random forest as well as Support Vector Machine methods made using Python programming language per Confusion Matrix, ROC curve, and Precision-Recall curve for a particular path are shown. Performing a critical analysis using a different method of the forecasting process. Not only that PCA analysis in case studies and presenting different perspectives such as confusion matrix, curve subconscious, confusion matrix, and ROC and AOC curve have been added to provide integrated analysis to draw a summary to predict credit card fraud.

Contents

Abstract 1

Introduction 3

Data Exploration and Features Selection 3

- Challenges in Credit Card Fraud Detection 5

Experiments 5

- PCA Analysis 6
- Logistic Regression Analysis 7
- Multilayer Perception 8
- Random Forest 9
- Support Vector Machine 10

Results 11

- PCA Analysis 11
- Logistic Regression 12
- Multilayer Perception 14
- Random Forest 16
- SVM 18

Discussion, Future Work, and Conclusion 20

Reference 21

Introduction

The informational indexes utilized in the investigation taken from the European Cardholder contained 284,807 exchanges. In the present monetary environment, Mastercard use has gotten very typical. These cards permit the client to pay enormous amounts of cash without conveying huge amounts of cash. They have changed the installment technique for nothing and made any type of installment for the buyer. This electronic installment technique is helpful yet accompanies its dangers. With a developing number of clients, Mastercard misrepresentation is likewise expanding at a similar rate (Rout, M., 2021). Individual Credit Card data might be wrongfully gathered and might be utilized for false exchanges. Some AI calculations can be utilized to gather information to address this issue. Extortion can be characterized as intentional misdirection to make a benefit, particularly cash. It is a misinterpretation that is expanding everyday events (Sailusha et al., 2020). There has been a critical expansion in the utilization of electronic installment strategies, for example, credit and bank cards and this has prompted an increment in Mastercard extortion. These cards can be utilized both on the web and disconnected strategies to make installments. On account of the online installment mode, the card should not have to be truly shown. In such cases, the card information is regularly assaulted by programmers or cybercriminals. These types of misrepresentations bring about the deficiency of millions every year. To conquer this obstruction, numerous calculations are as yet being researched and studied. Different techniques are utilized to track down the best answer for this issue (Husejinovic, A., 2020).

Data Exploration and Features Selection

The objective is to create the best classifier for credit card fraud detection. To do it, various comparison of classification models from different methods will be applied in analysis with the help of Python and Jupiter Notebook:

- PCA Analysis
- Logistic regression
- Multilayer Perception
- Bagging or Random Forest
- SVM or Support Vector Machine

The databases contain Credit Card exchanges in September 2013 by European cardholders. This database provides a two-day transaction, in which we have 492 deceptions of 284,807 exchanges. The database is very closed, the best class cheats represent 0.172%, all equal.

It contains only the mathematical information that is the result of the change in PCA. Unfortunately, due to privacy issues, we are unable to provide initial highlights and additional basic data about the information. In summarizing, we will sort out common issues, for example, unequal data, continuous working conditions, and highlight the difficulty of compiling almost every experimental work experience, and separate common ways to deal with it. The problem of unequal databases occurs because the value of the actual exchange is much higher than the deceptive one when using the right part composition than the open hand as the key points are obtained in restricted businesses, and the use of prominent design techniques and data enhancement is very important (Khatri et al., 2020). Likewise, adjusting the identification framework in permanent cases is a test because the exchange rate for a prepaid card is very high. It contains only the mathematical information that is the result of the change in PCA. Shockingly, due to privacy issues, the first highlights and additional basic data about the information provided. Variable names are V1, V2…, V28 are the main components obtained by PCA, the only ones that are not modified by PCA are "Time" as well as "Amount", and there are omitted from the model before starting analysis.

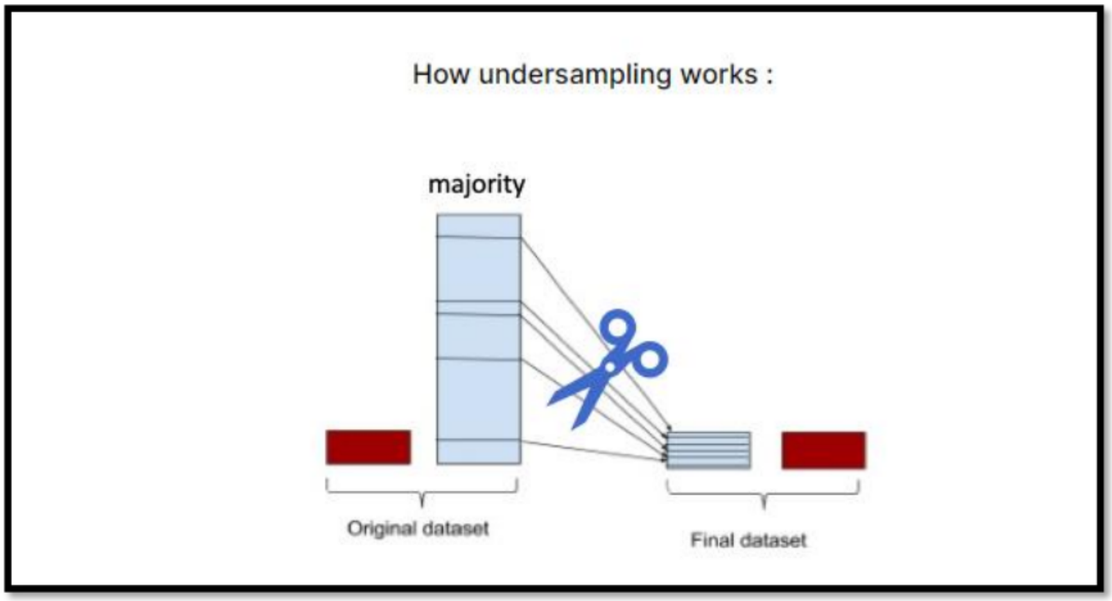


Figure: Showing How Under sampling works by imbalance learn python package

Source: (Web)

Challenges in Credit Card Fraud Detection

There are various difficulties faced by the techniques of over-recognition by various experts. Forgiveness of specially produced results ensures that any method of misdiagnosis has the potential to address the associated problems.

1. Unequal details: Unequal. This means that there are not many issues of Credit Card robbery. This will make it difficult for you to identify fraudulent transactions and be difficult to verify.
2. During the perception cycle, each misalignment error has significance for the other. Wrong classification of standard exchanges will not create many problems compared to thinking of non-standard exchanges that are as well done as normal exchanges. Since the problem of ambiguity can be eliminated or reversed in some continuation.
3. Spreading information: Sometimes real or virtual transactions are considered fraudulent or improper exchanges in some way around. This will create a problem distinguishing real and different exchanges. The biggest test is considered by each Mastercard organization.

Experiments

To achieve the highest level of commitment and closeness to investigations, PCA assessments have been performed on a case-by-case basis. Similarly, Logistic Regression, Multilayer Perceptron, and unconventional forest counts, SVM, and PCA have never separated true authentication among all AI strategies (Trivedi et al., 2020).

PCA Analysis

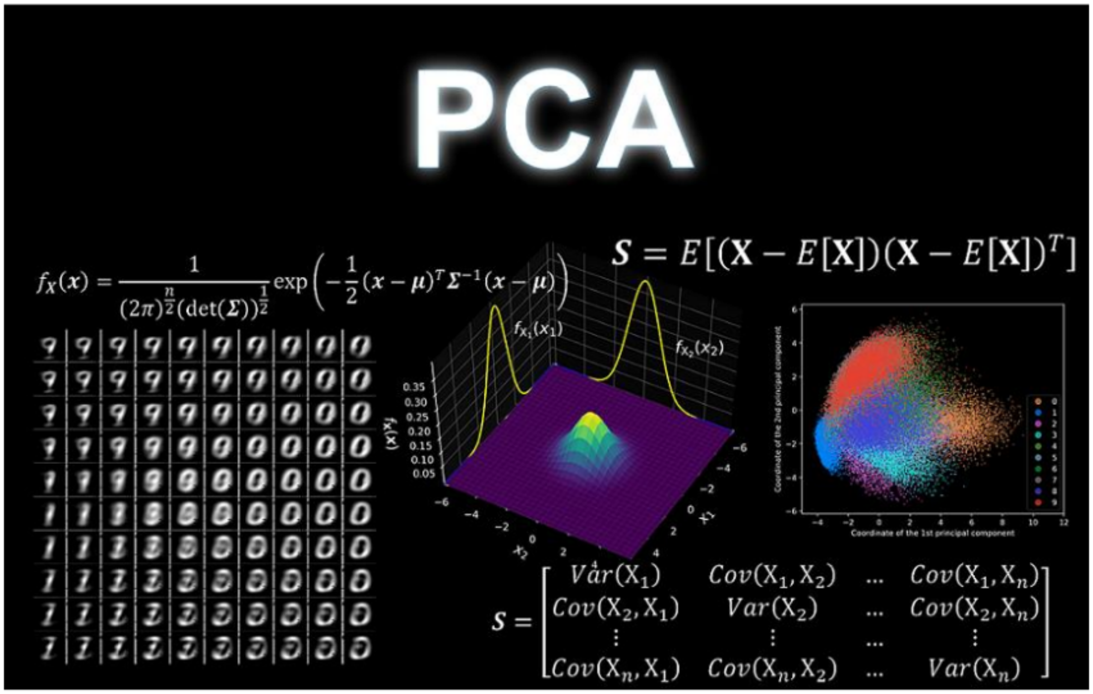

PCA
$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det(\Sigma))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
$S = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$
$f_{X_1}(x_1)$
$f_{X_2}(x_2)$
$f_X(x)$
0.35
0.30
0.25
0.20
0.15
0.10
0.05
x_1
x_2
Coordinate of the 2nd principal component
Coordinate of the 1st principal component
$S = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & \dots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Var(X_n) \end{bmatrix}$

Figure: Showing How PCA works

Source: (Web)

This piece of the examination uncovered that a mix of PCA and LR is conceivable and compelling in recognizing extortion on Credit Cards (Agarwal et al., 2020). The introduction appears to recognize false data from normal data. Additionally, a vital method to restrict the degree of recreation has likewise been created. Regularly, their area innovation is utilized to help set this breaking point since it straightforwardly influences the precision and memory of exchanging said by (Cynthia et al., 2020).

Logistic Regression Analysis

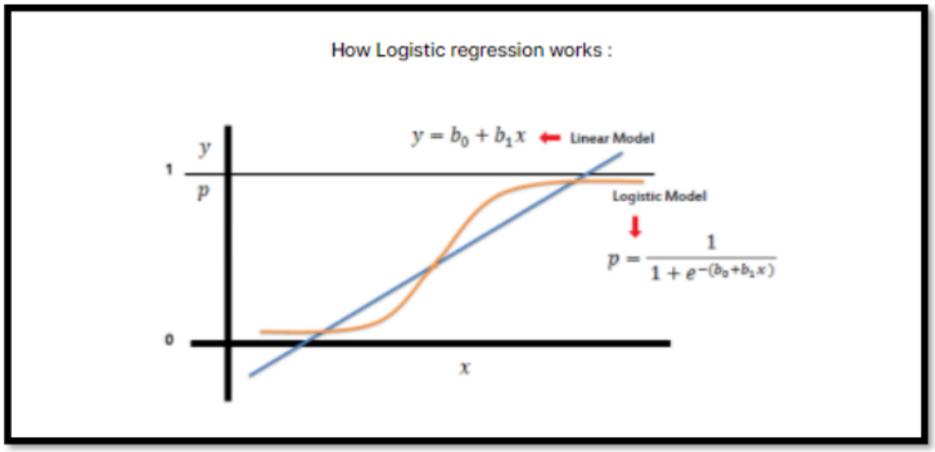

How Logistic regression works :
$y = b_0 + b_1 x$
Linear Model
y
1
p
Logistic Model
$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$
0
x

Figure: Showing How Logistic works

Source: (Web)

In Logistic Regression, the input values (X) are assembled consecutively utilizing loads or coefficient esteems to anticipate the predicted value (y) (Hussein et al., 2020). The principal contrast from the inversion of the line is that the mimicked yield esteem is a parallel worth (0 or 1) as different to a mathematical value. Resource recovery is a clear interaction, however, the hypothesis is changed over utilizing a resource work as per (Ata, O. also, Hazim, L., 2020).

Multilayer Perception

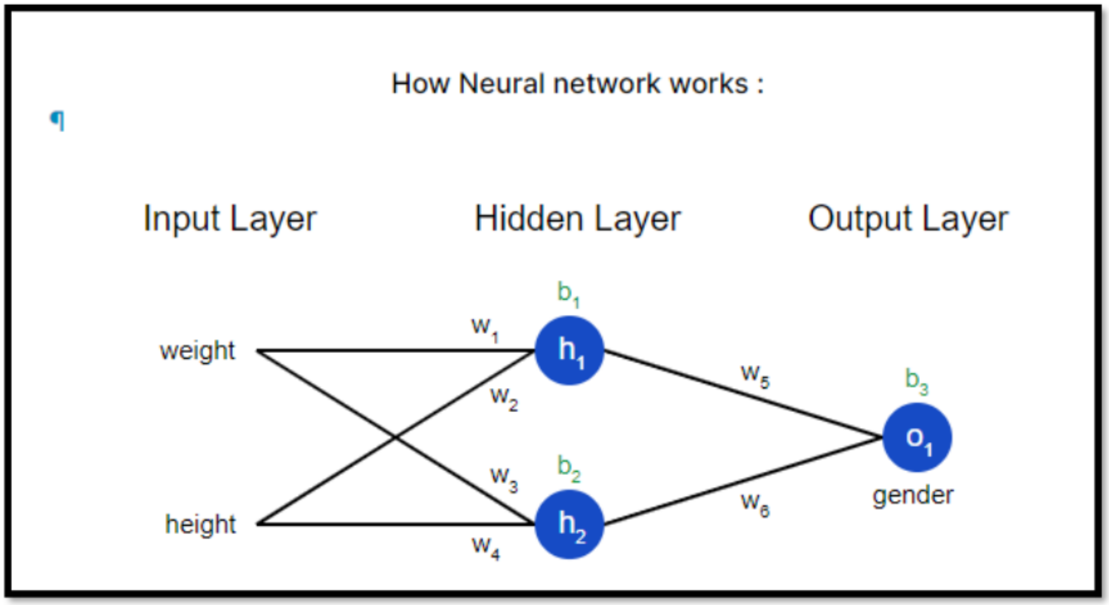


Figure: Showing How multilayer perception works

Source: (Web)

Layers of the neural organization are comprised of nodes. A node consolidates contribution from information with a bunch of coefficients and inclination, which increments or diminishes the info, along these lines giving the info esteem comparable to the capacity the calculation endeavors to peruse as indicated by (Bagga et al., 2020). These information-weighted items are summed up and the sum is sent through a capacity called node initiation, tracking down that the sign ought to be consistently evolved through the organization to influence the eventual outcome, say, the activity of detachment. At the point when signs pass, the neuron is "activated."

Random Forest

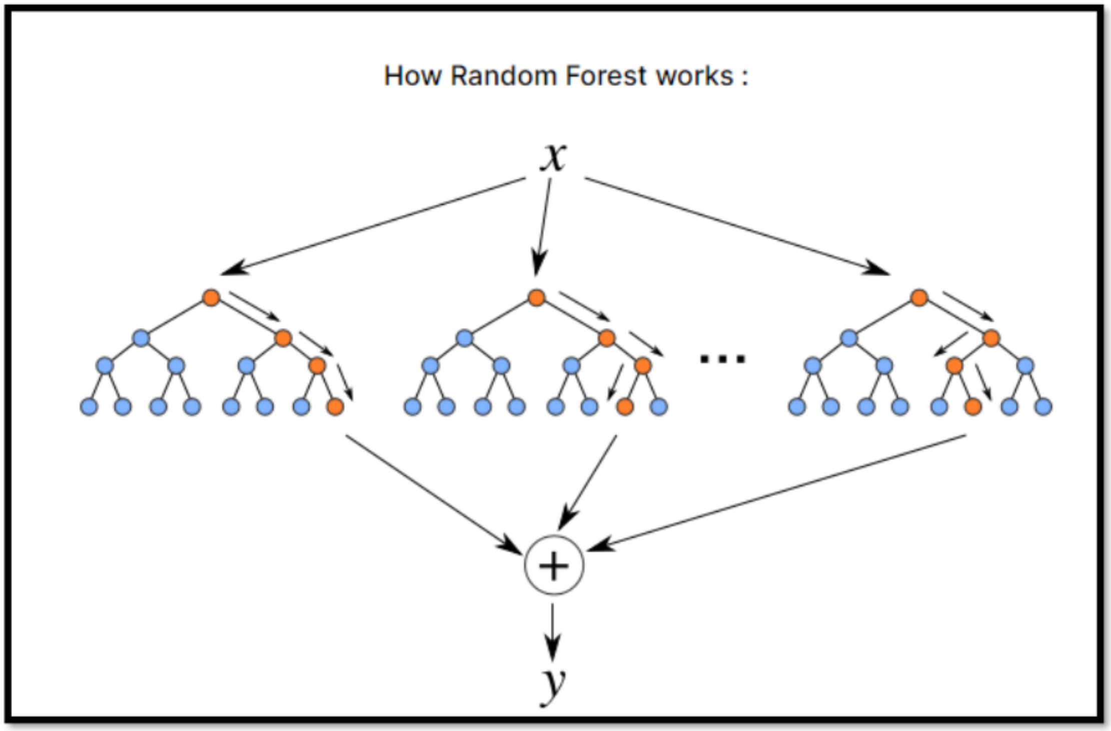


Figure: Showing How Random forest works

Source: (Web)

"Many unrelated species (trees) that generally act as a team of advisers will strike any image of the individual". The random forest contains many optional trees that act as a blanket. Each tree in a random forest reflects the expectations of the category and the category with the most votes became the prediction of our model said by (Riffi et al., 2020).

Support Vector Machine

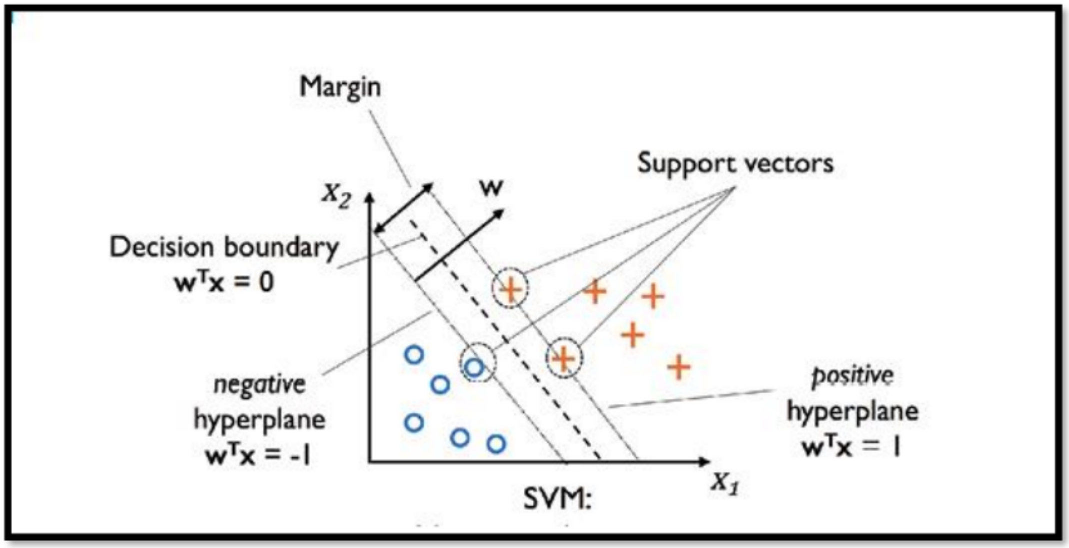


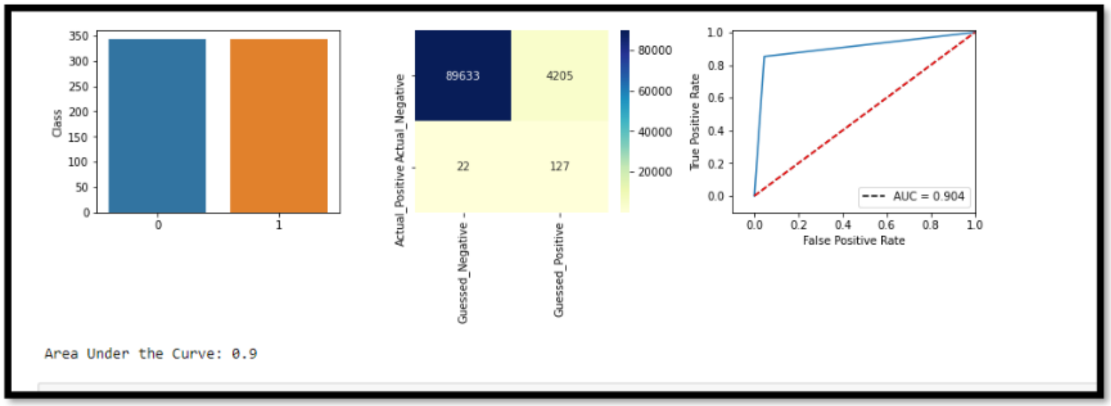
Figure: Showing How SVM works

Source: (Web)

SVM Classifier utilizes a cycle called a portion stunt to change over information and dependent on this alteration it tracks down a total limit (hyper-plane) between potential results. The vector support machines center just around the most troublesome focuses to recognize, and a few classifiers overlook all focuses as per (Sudha, C. and, Akila, D., 2021).

Results

PCA Analysis



```
Area Under the Curve: 0.9
```

Figure: Showing PCA analysis Result

Source: (Analysis)

By using the meanings and standard deviations of the reconstruction school we can set the appropriate limit. After that, I prefer to set the limit to say +2 * std. With this, 94.2% of the surveys we conducted were able to detect a 57% fraudulent estimate and an average value of Area Under the Curve of 0.94. There were 3 equal folders per person out of 12, a total of 36, and the matrix came next:

	Precision	**Recall**	**f1-score**	**Support**
For 0	1.00	0.99	1.00	93838
For 1	0.14	0.89	0.25	149
Accuracy			0.99	93987
Macro Avg. Value	0.57	0.94	0.62	93987
Weighted Avg. Value	1.00	0.99	0.99	93987

PCA analysis result of detecting fraud is shown as following:

	principal_component_1	principal_component_2	Class
0	0.125725	-0.479107	0
1	0.386067	0.531411	0

2	0.324796	-0.044122	0
3	0.995606	1.076000	0
4	0.113791	0.677320	0

[0.10000819 0.10000457]

After training the data sample again fitting 3 folds for each of 12 candidates, totaling 36 fits

	Precision	**Recall**	**F1-score**	**Support**
For 0	1.00	0.96	0.98	93838
For 1	0.03	0.85	0.06	149
Accuracy			0.96	93987
Macro Avg. value	0.51	0.90	0.52	93987
Weighted Avg. value	1.00	0.96	0.98	93987

Based on the processed data in PCA, achieved good accuracy of 96 % on training and 99 % on the independent test set.

Logistic Regression

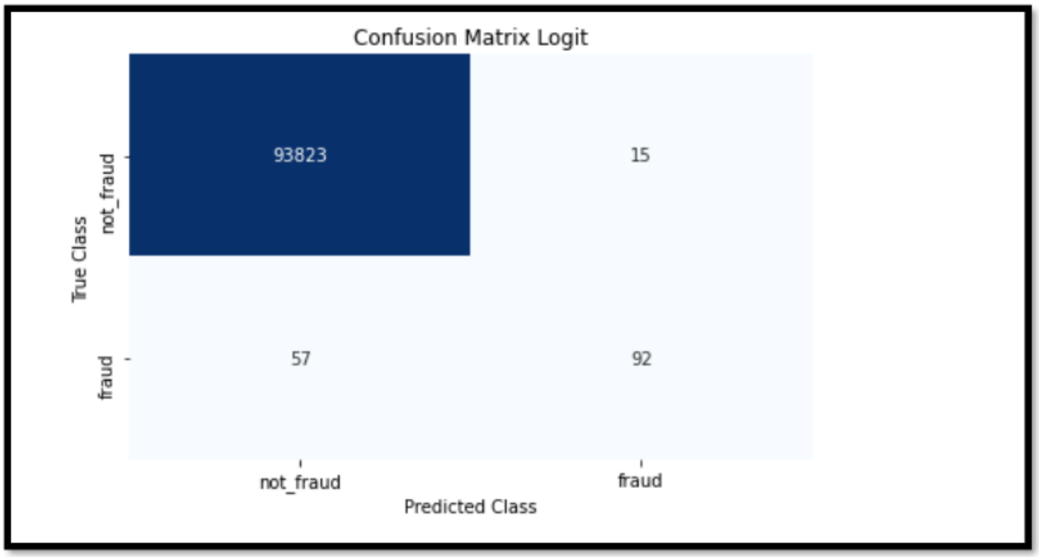


Figure: Showing Confusion Matrix for Logistic Regression

Source: (Analysis)

After successfully applied the Logistic regression algorithm on the data set trained to model the gained result is as following:

Accuracy of Logistic Regression method: 0.9992339366082543

Precision value of Logistic Regression method: 0.8598130841121495

Recall value of accuracy Logistic Regression method: 0.6174496644295302

F1 Score value Logistic Regression method: 0.7187499999999999

AUC value for Logistic Regression method: 0.979455740587341

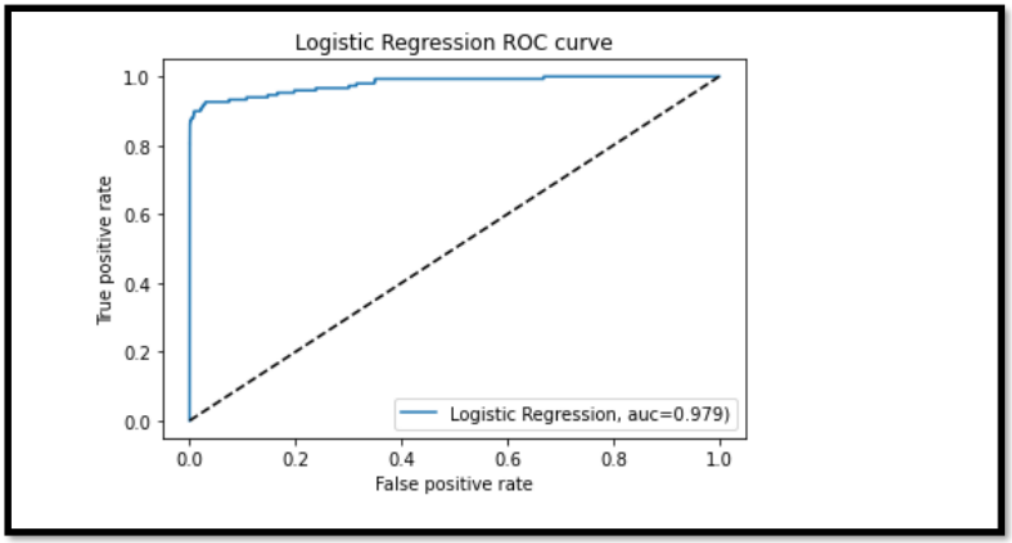


Figure: Showing ROC Curve for Logistic Regression

Source: (Analysis)

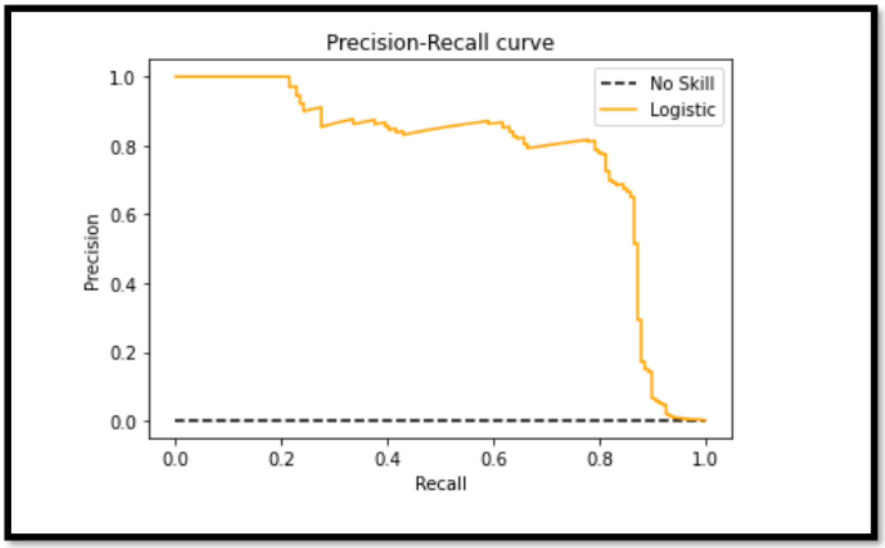


Figure: Showing Precision-Recall Curve for Logistic Regression

Source: (Analysis)

Multilayer Perception

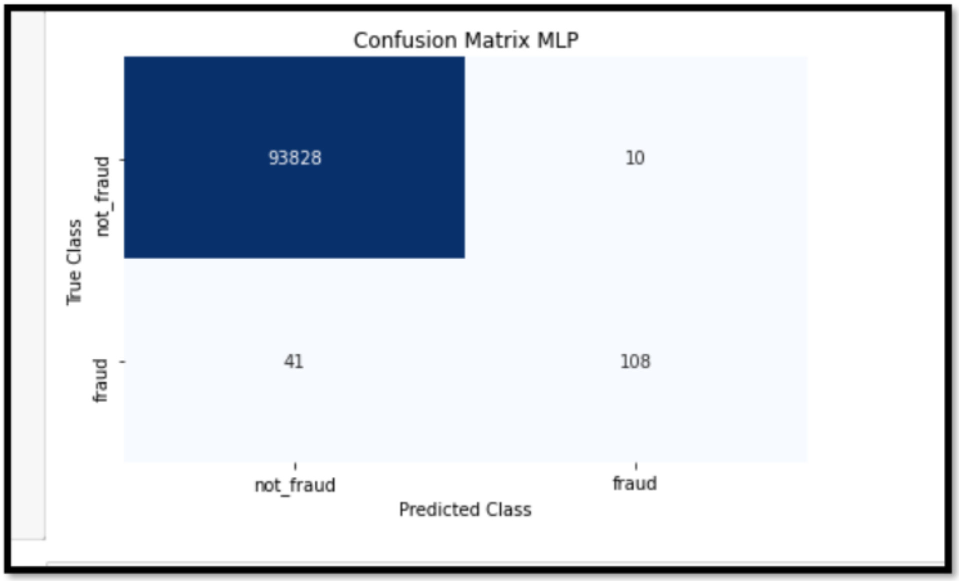


Figure: Showing Confusion Matrix for Multilayer Perception

Source: (Analysis)

The gained score values of multilayer perception are as following:

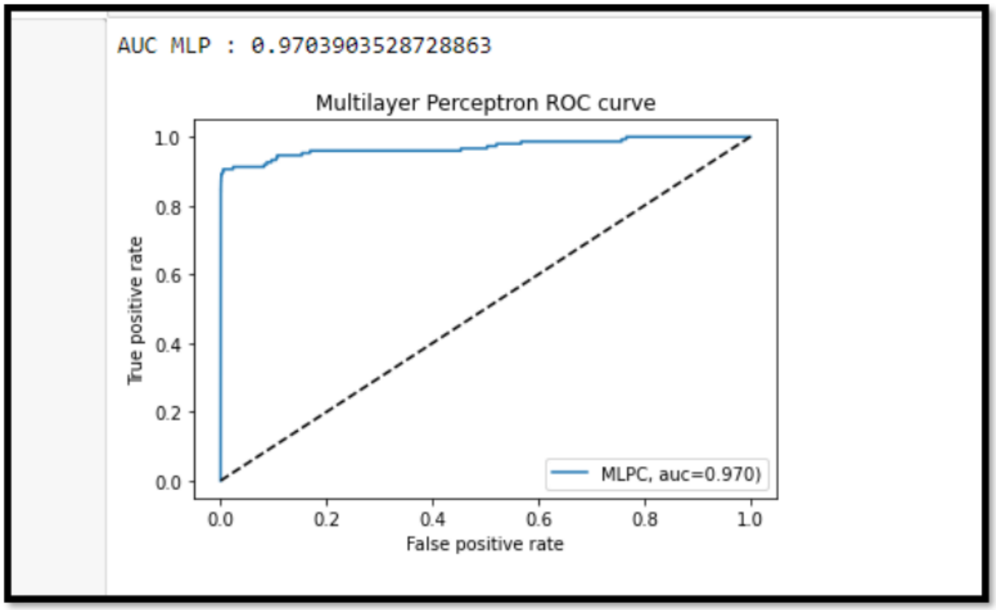
Accuracy MLP: 0.9994573717641801

Precision MLP: 0.9152542372881356

Recall MLP: 0.7248322147651006

F1 Score MLP: 0.8089887640449437

AUC MLP: 0.9703903528728863



```
AUC MLP : 0.9703903528728863
```

Figure: Showing ROC Curve for Multilayer Perception

Source: (Analysis)

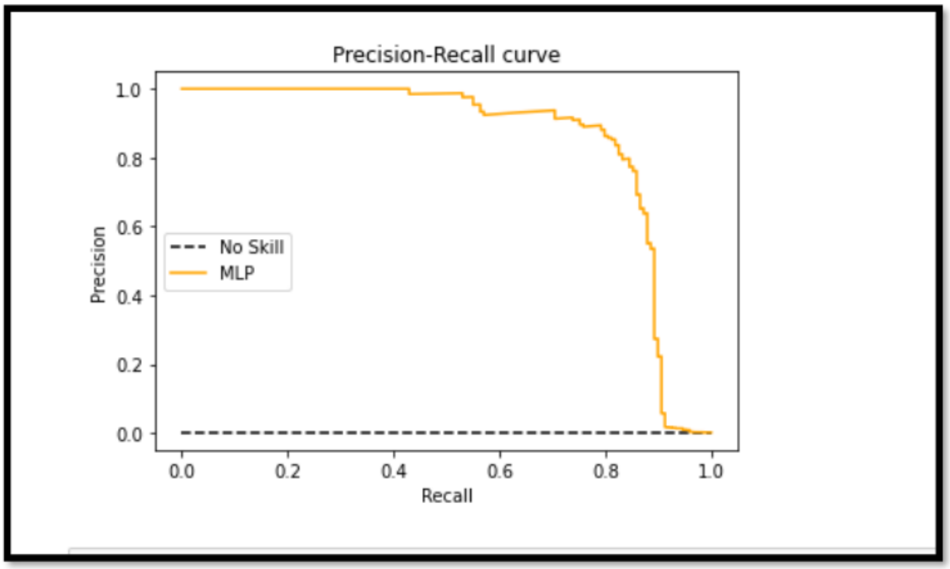


Figure: Showing Precision-Recall Curve for Multilayer Perception

Source: (Analysis)

Random Forest

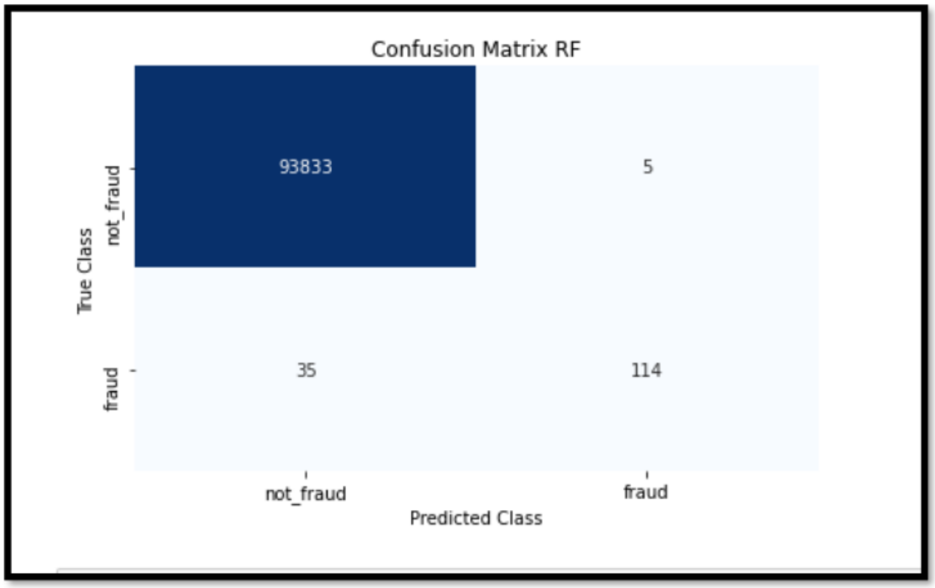


Figure: Showing Confusion Matrix for Random Forest

Source: (Analysis)

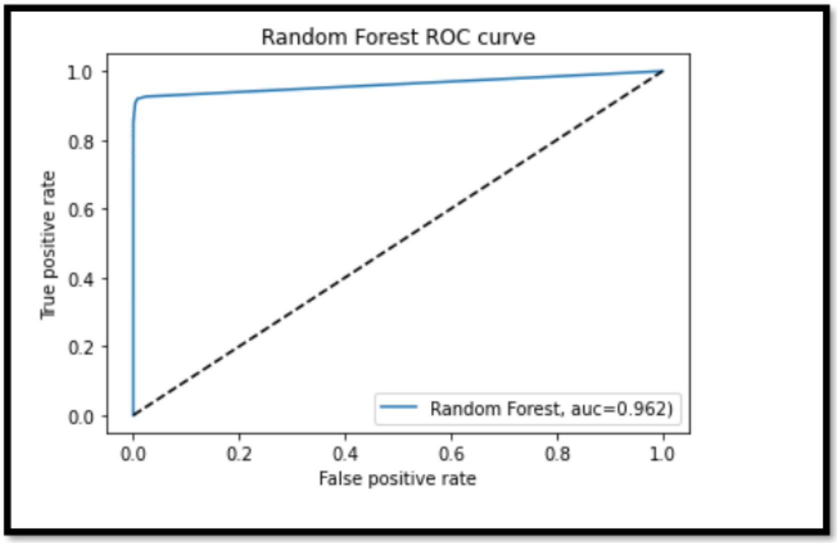


Figure: Showing ROC Curve for Random Forest

Source: (Analysis)

Hence, after applying the algorithm of RF gained value scores are:

Accuracy RF: 0.999574409226808

Precision RF: 0.957983193277311

Recall RF: 0.7651006711409396

F1 Score RF: 0.8507462686567164

AUC Random Forest: 0.9619061824526661

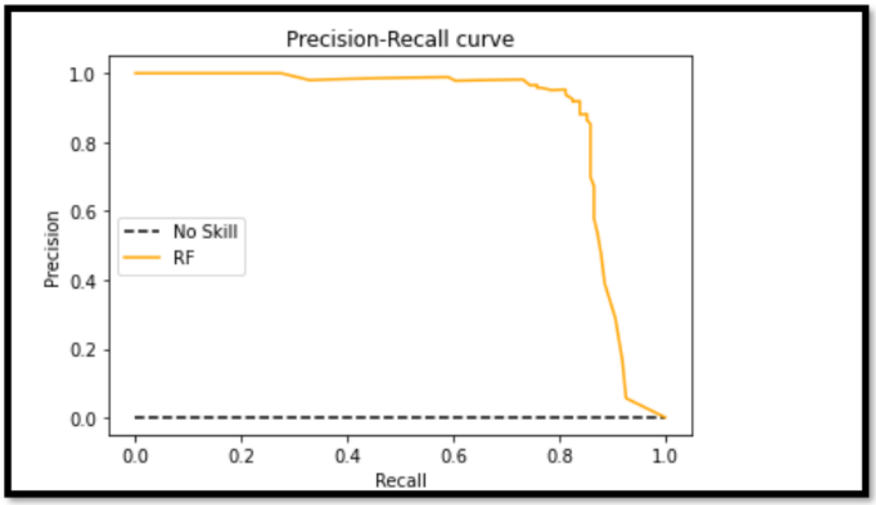
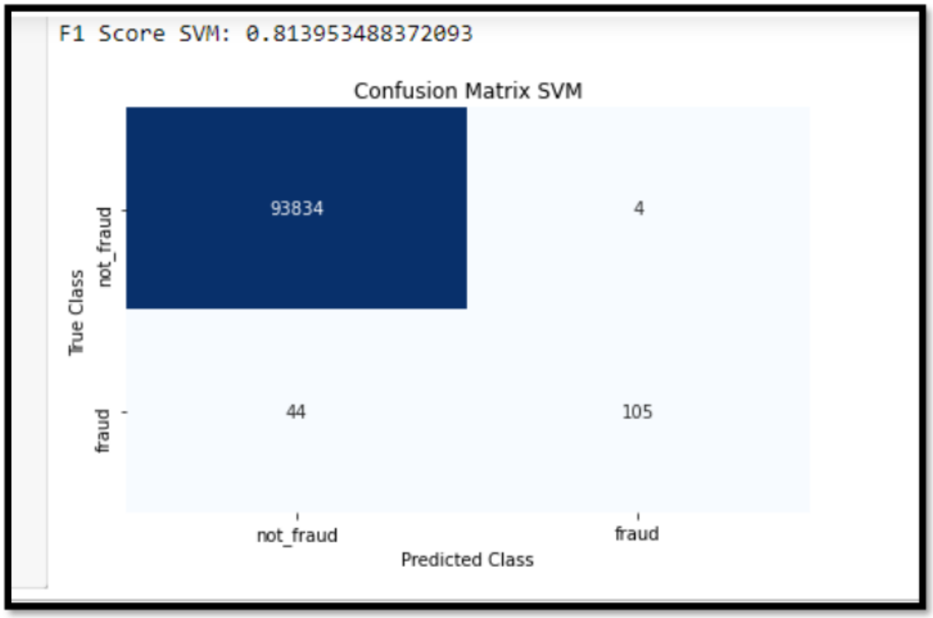


Figure: Showing Precision-Recall Curve for Random Forest

Source: (Analysis)

SVM



```
F1 Score SVM: 0.813953488372093
```

Figure: Showing Confusion Matrix for SVM

Source: (Analysis)

After applying this algorithm, the gained score value of prediction are as follows:

Accuracy SVM: 0.9994892910721696

Precision SVM: 0.963302752293578

Recall SVM: 0.7046979865771812

F1 Score SVM: 0.813953488372093

AUC SVM: 0.954083440388698

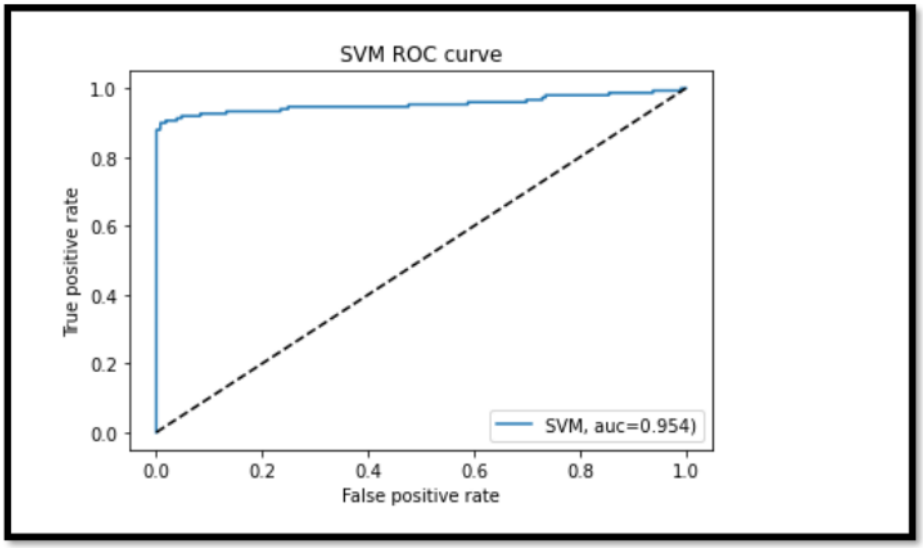


Figure: Showing ROC Curve for SVM

Source: (Analysis)

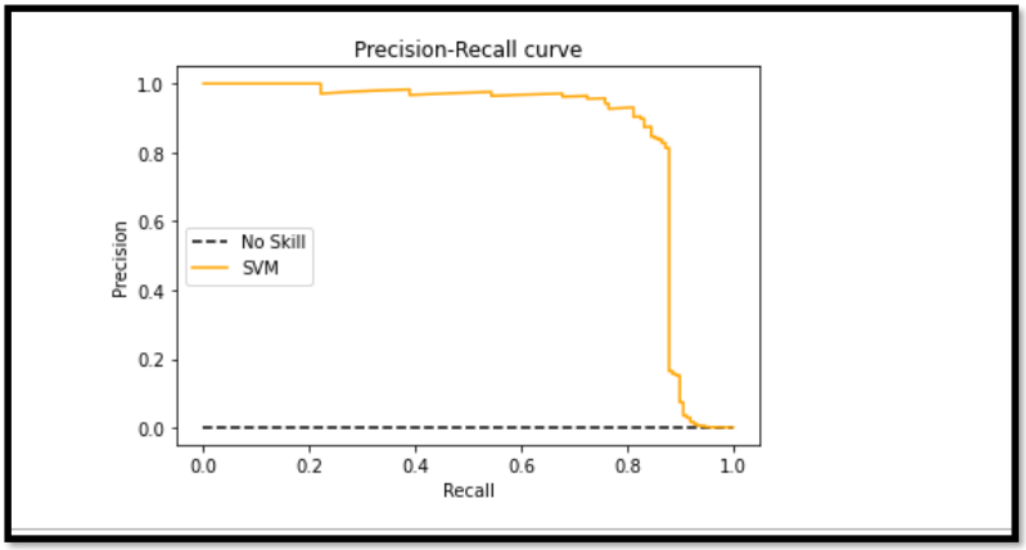


Figure: Showing Precision-Recall Curve for SVM

Source: (Analysis)

Discussion, Future Work, and Conclusion

To figure or recognize Credit Card extortion, an alternate AI calculation has been utilized to test the most ideal approaches to anticipate exactness levels in the wake of preparing an alternate calculation with a similar informational index. Promptly, all insightful outcomes are attracted to finish the most ideal outcomes and extortion discovery has been made. AUC esteem score results for every strategy, PCA with 90.40% in the wake of preparing, calculated with 97.95% after preparing, MLP with 97.08% in the wake of preparing, arbitrary woodland with 96.19% exactness in the wake of preparing, and the SVM technique has 95.40% precision in the wake of preparing. Subsequently, we can reason that MLP is the most ideal approach to anticipate misrepresentation for this situation study, or there are numerous calculations in AI in future work that can work in other unwinding strategies to test informational collection. We can likewise test neural organization-based multi-facet insight utilizing Cameras and Tensorflow and future. The precision and F1 school score on PCA show 99 and 94% separately. While the Logistic Regression investigation shows a 99.92% exactness and an F1 rating of 71.87%. By examination, Multilayer Perception has an exactness of 99.94% and the F1 Score shows 80.89%. For arbitrary woods or Bagging, the precision of the street is 99.96% and the F1 Score is 85.07%. At long last, SVM technique exactness is 99.95% and

F1 Score esteem scores are 81.39%. Contrasted with the previously mentioned impact factor, PCA has a more exact degree of precision and a score of F1 contrasted with other scientific techniques. In this manner, correlations are attracted to assess the presence of classifiers in the data set. What's more, execution measures are determined to quantify model execution. Administered and Unsupervised calculations have additionally been acquainted with recognizing fakes on inconsistent information. An assortment of test techniques is utilized to test the presentation of each model and the chart is intended to address the exactness of each model. With future work, the spotlight can be put on rebuilding techniques to lessen information base tendency. The charge card has become a need for everybody nowadays and along these lines, extortion likewise happens rapidly. A brilliant extortion identification program by thinking about their exchange history so you understand what the misrepresentation is or is valid. This paper records numerous procedures proposed to decrease or keep away from charge card misrepresentation. Also, the benefits and weaknesses of the techniques have been exhibited and looked at. Be that as it may, there is a need to improve methodologies for better results to keep away from future deceitful exercises.

Reference

Agarwal, A., Rana, A., Gupta, K., and Verma, N., 2020, June. A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 443-448). IEEE.

Ata, O. and Hazim, L., 2020. Comparative Analysis of Different Distributions Dataset by Using Data Mining Techniques on Credit Card Fraud Detection. *Tehnički vjesnik*, *27*(2), pp.618-626.

Bagga, S., Goyal, A., Gupta, N., and Goyal, A., 2020. Credit Card Fraud Detection using Pipeline and Ensemble Learning. *Procedia Computer Science*, *173*, pp.104-112.

Cynthia, P.C. and George, S.T., 2021. An Outlier Detection Approach on Credit Card Fraud Detection Using Machine Learning: A Comparative Analysis on Supervised and Unsupervised Learning. In *Intelligence in Big Data Technologies—Beyond the Hype* (pp. 125-135). Springer, Singapore.

Husejinovic, A., 2020. Credit card fraud detection using naive Bayesian and c4. 5 decision tree classifiers. *Husejinovic, A.(2020). Credit card fraud detection using naive Bayesian and C*, *4*, pp.1-5.

Hussein, A.S., Khairy, R.S., Najeeb, S.M.M. and ALRikabi, H.T., 2021. Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression. *International Journal of Interactive Mobile Technologies*, *15*(5).

Khatri, S., Arora, A. and Agrawal, A.P., 2020, January. Supervised machine learning algorithms for credit card fraud detection: a comparison. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 680-683). IEEE.

Riffi, J., Mahraz, M.A., El Yahyaouy, A. and Tairi, H., 2020, June. Credit Card Fraud Detection Based on Multilayer Perceptron and Extreme Learning Machine Architectures. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-5). IEEE.

Rout, M., 2021. Analysis and comparison of credit card fraud detection using machine learning. In *Advances in electronics, communication, and computing* (pp. 33-40). Springer, Singapore.

Sailusha, R., Gnaneswar, V., Ramesh, R. and Rao, G.R., 2020, May. Credit Card Fraud Detection Using Machine Learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1264-1270). IEEE.

Sudha, C. and Akila, D., 2021, January. Credit Card Fraud Detection System based on Operational & Transaction features using SVM and Random Forest Classifiers. In *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)* (pp. 133-138). IEEE.

Trivedi, N.K., Simaiya, S., Lilhore, U.K. and Sharma, S.K., 2020. An efficient credit card fraud detection model based on machine learning methods. *International Journal of Advanced Science and Technology*, *29*(5), pp.3414-3424.

djdmchm

ORIGINALITY REPORT

4%	4%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pastebin.com Internet Source	2%
2	Submitted to University of Technology, Sydney Student Paper	1%
3	Submitted to Harrisburg University of Science and Technology Student Paper	1%
4	Walid A. Hanafy, Alfredo Pina, Sameh A. Salem. "Machine Learning Approach for Photovoltaic Panels Cleanliness Detection", 2019 15th International Computer Engineering Conference (ICENCO), 2019 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		