A PROJECT REPORT

on

“Comparative Analysis of Machine Learning Models for Credit Card Fraud Detection”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN INFORMATION TECHNOLOGY

BY

MUKTESH MISHRA (21052258)

UNDER THE GUIDANCE OF
Dr. Debanjan Pathak

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
April

ABSTRACT

Credit card fraud poses a significant challenge in financial services, resulting in substantial financial losses annually. Despite its prevalence, there's a paucity of research due to confidentiality constraints surrounding real-world credit card data. This paper addresses this gap by employing logistic regression, decision tree, random forest, and support vector machine (SVM) algorithms for credit card fraud detection. Evaluation of model efficiency involves the utilization of publicly available credit card data sets, alongside analysis of real-world data obtained from a financial institution. Additionally, noise is deliberately introduced into the data samples to further gauge the algorithms' robustness. Encouraging experimental results underscore the effectiveness of the algorithms, showcasing notable accuracy rates in identifying fraudulent transactions within credit card data.

This research endeavors to contribute to the ongoing efforts in combating credit card fraud through a comprehensive comparative analysis of machine learning models. By leveraging logistic regression, decision tree, random forest, and SVM algorithms, the study not only enhances understanding of algorithm performance but also provides insights into the practical applicability of these techniques in real-world scenarios. The findings underscore the potential of the majority voting method as a promising tool for fraud detection in credit card transactions, thereby offering valuable implications for financial institutions and stakeholders in mitigating the adverse impacts of fraudulent activities

Keywords: Credit card, fraud detection, Classification models, Logistic regression, Decision tree, Random forest, Support Vector Machine

Contents

1	Introduction		4
2	Basic Concepts		5
	2.1	Machine Learning Algorithm	5
	2.2	Performance Metrics	6
3	Problem Statement		9
	3.1	System Design	9
4	Exploring the dataset		10
5	Performance of the data set on various classification algorithms and its analysis		12
	5.1	Logistic Regression	12
	5.2	Decision Tree classifier	12
	5.3	Random Forest Classifier	13
	5.4	Support Vector Machine	13
6	Conclusion and Future Scope		14
	6.1	Conclusion ………………………..	14
	6.2	Future Scope ……………………….	14
References			15

Chapter 1

Introduction

Credit card fraud continues to be a significant financial burden, costing billions of dollars annually. Traditional methods, reliant on rules and manual review, struggle to keep pace with evolving fraudulent activities. Additionally, the lack of research on real-world data due to confidentiality restrictions hinders the development of more effective solutions.

This project aims to address these shortcomings by leveraging machine learning's power. By employing advanced data analytics techniques, we aim to develop a more robust and accurate fraud detection system. We will compare prominent machine learning models like logistic regression, decision trees, random forests, and support vector machines to identify the most effective approach for credit card fraud detection.

This project's findings can significantly impact the financial services industry. By identifying the most effective machine learning model, we can empower financial institutions to proactively prevent fraudulent activities, ultimately enhancing the security of credit card transactions in the digital age.

Chapter 2

Basic Concepts

The tools and techniques used in this project are describes as follows:

2.1. Machine Learning Algorithms

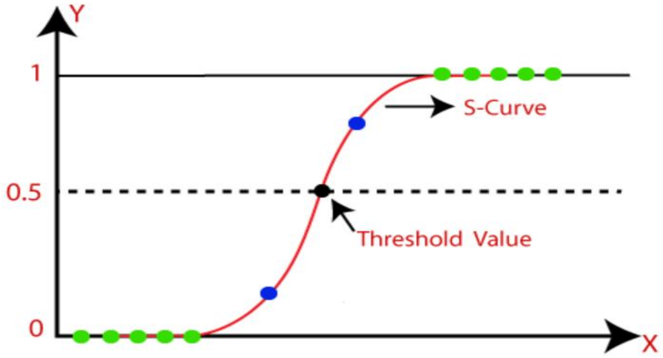
Machine learning algorithms serve as the backbone of credit card fraud detection systems, enabling automated analysis of transaction data to identify potentially fraudulent activities. In this subsection, we discuss the key machine learning algorithms utilized in this project:

2.1.1. Logistic Regression:

Logistic regression is the supervised learning algorithm, which is used to **predict the categorical variables or discrete values**. It can be used for the classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc.

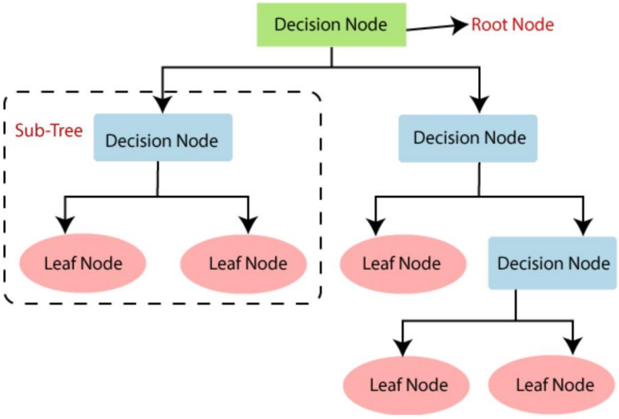
Logistic regression is similar to the linear regression except how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and used to predict the discrete values.

Instead of fitting the best fit line, it forms an S-shaped curve that lies between 0 and 1. The S-shaped curve is also known as a logistic function that uses the concept of the threshold. Any value above the threshold will tend to 1, and below the threshold will tend to 0.



2.1.2. Decision Tree Classifier:

A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems. It can work with both categorical variables and continuous variables. It shows a tree-like structure that includes nodes and branches, and starts with the root node that expand on further branches till the leaf node. The **internal node** is used to represent the **features of the dataset, branches show the decision rules,** and **leaf nodes represent the outcome of the problem.**

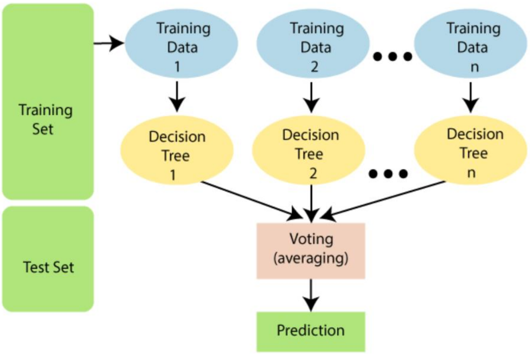


2.1.3. Random Forest Classifier:

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

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

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

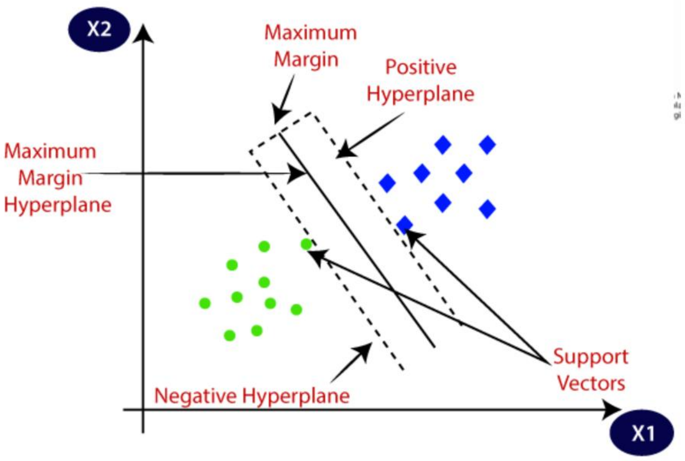


2.1.3. Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



2.2. Performance Metrics :

In a classification problem, the category or classes of data is identified based on training data. The model learns from the given dataset and then classifies the new data into classes or groups based on the training. It predicts class labels as the output, such as *Yes or No, 0 or 1, Spam or Not Spam,* etc. To evaluate the performance of a classification model, different metrics are used, and some of them are as follows:

2.2.1. Accuracy:

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.

It can be formulated as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions}$$

2.2.2. Confusion Matrix:

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.

In general, the table is divided into four terminologies, which are as follows:

1. **True Positive(TP):** In this case, the prediction outcome is true, and it is true in reality, also.
2. True Negative(TN): in this case, the prediction outcome is false, and it is false in reality, also.
3. False Positive(FP): In this case, prediction outcomes are true, but they are false in actuality.
4. False Negative(FN): In this case, predictions are false, and they are true in actuality.

2.2.3. Precision:

The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})}$$

2.2.4. Recall:

It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

The formula for calculating Recall is given below:

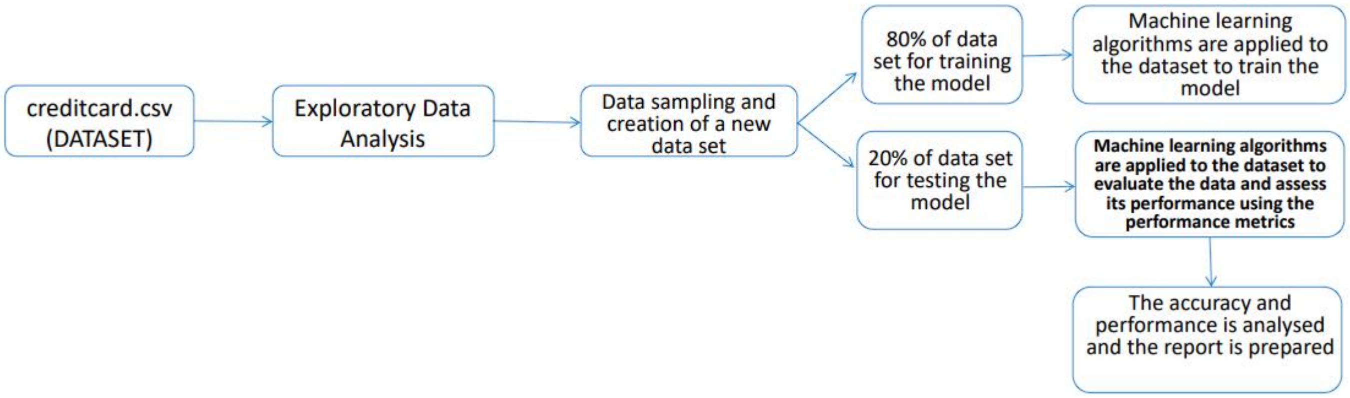
$$\textbf{Recall=}\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

Chapter 3

Problem Statement

To improve accuracy and efficiency in identifying fraudulent transactions, machine learning algorithms for credit card fraud detection are being developed and evaluated. This project tackles issues including imbalanced datasets, the lack of study on real-world credit card data, and the shortcomings of rule-based systems. The goal is to determine the best methods for enhancing fraud detection systems by contrasting algorithms such as logistic regression, decision tree, random forest, and SVM.

3.1 Project Planning and Design:


creditcard.csv (DATASET)
Exploratory Data Analysis
Data sampling and creation of a new data set
80% of data set for training the model
20% of data set for testing the model
Machine learning algorithms are applied to the dataset to train the model
Machine learning algorithms are applied to the dataset to evaluate the data and assess its performance using the performance metrics
The accuracy and performance is analysed and the report is prepared

Data Collection: Gathering the information needed to train the model is the first stage. The dataset we have taken is "crreditcard.csv"

Analyzing exploratory data (EDA): Knowing what our data contains is crucial after we have it. The goal of this procedure, known as exploratory data analysis (EDA), is to find patterns, trends, and possible flaws in the data.

Data preprocessing: Cleaning and preparing the data for model training. This entails dealing with missing numbers, locating and eliminating outliers, and formatting the data so that the model can comprehend it.

Data splitting: The data will be divided into two sets - a training set and a testing set. The model is trained on the training set, and its performance is assessed on the testing set. We have considered, 20% for testing utilization and 80% is used for training.

Model Training: This is the application of the machine learning algorithm. The machine learning algorithms receive the training data and use it to create a model by learning from the data. After that, predictions based on fresh data can be made using this model.

Model Evaluation: It's critical to assess the model's performance on the testing set following training. This aids in evaluating how effectively the model extrapolates to unknown data. Metrics of performance include accuracy, precision, recall, F1 Score, and confusion matrix.

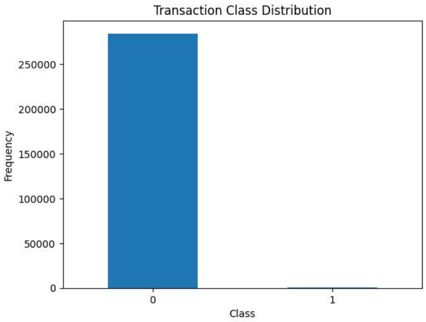
Creation of reports: Ultimately, a report summarising the outcomes of the machine learning procedure is created. Analysis of the results is done and we reach effective conclusions.

Chapter 4

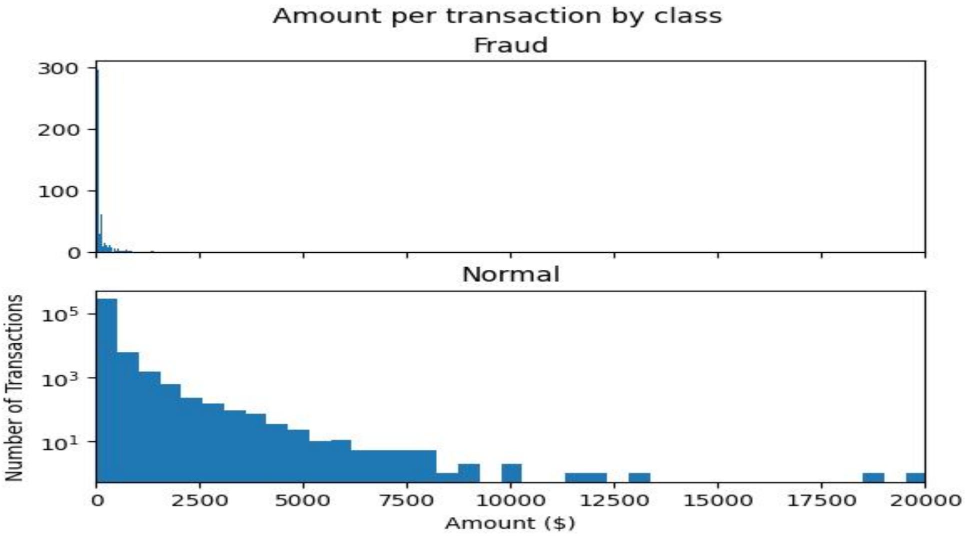
Exploring the Data Set

Shape of the Data set: (284807, 31)

Number of Legit cases and Fraud cases in the data set: (284315, 31) Legit and (492, 31) fraud

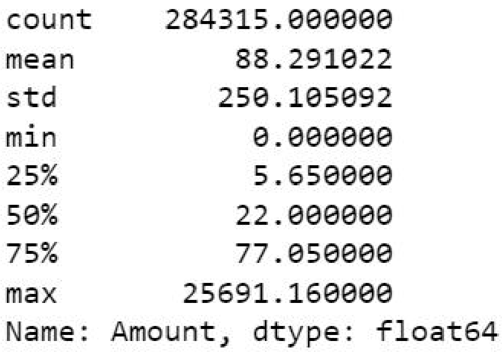


Here, 0 indicates the number of legit Transactions and 1 indicates the number of fraud transactions.



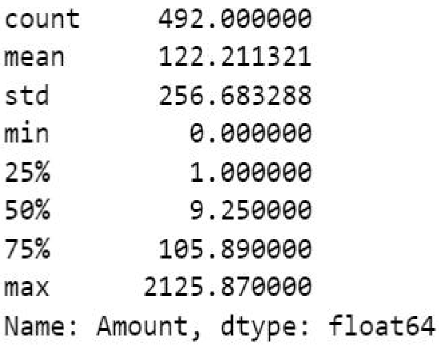
The detailed description of the fraud and Legit transactions are:

Legit:

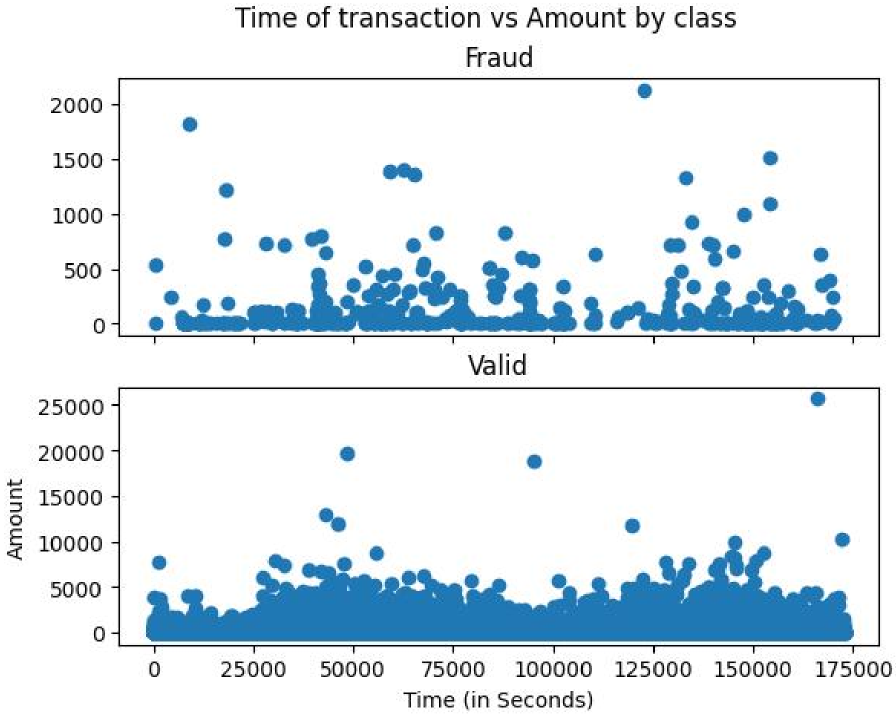


```
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

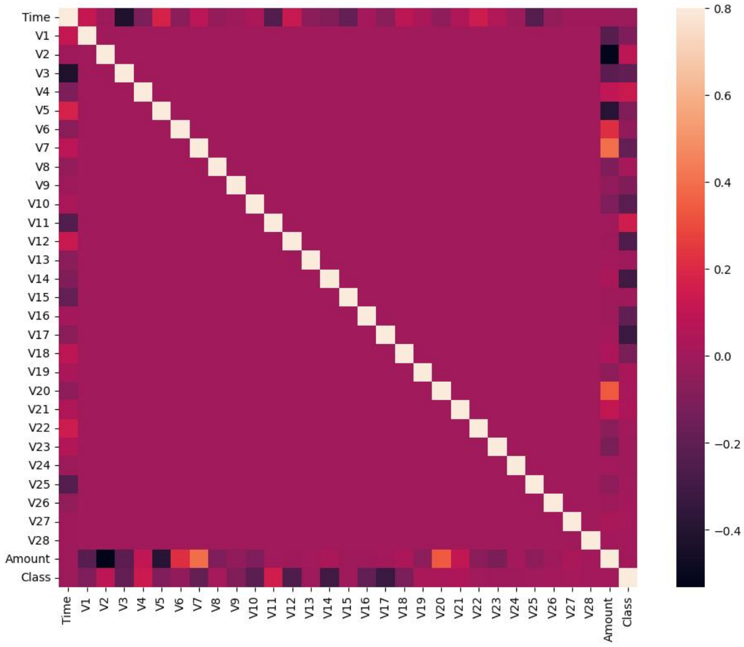
Fraud:



```
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```



This graph shows the amount of the transactions and time of the transactions according to the legit or fraud transactions out of the entire data set.

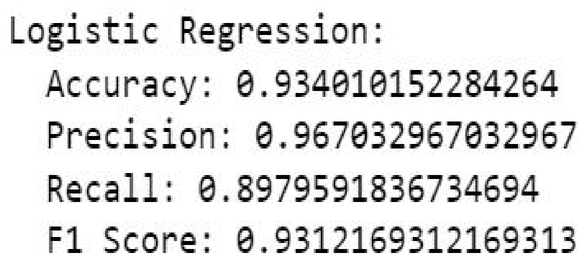


The above representation is the corelational heat-map of the entire data set.

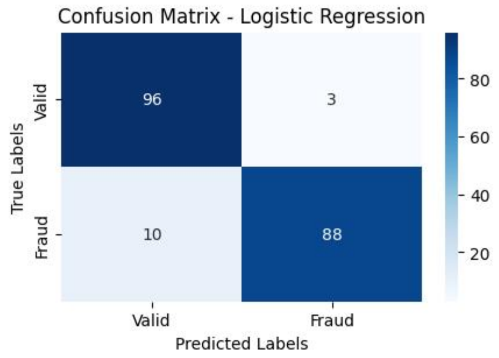
Chapter 5

Performance of the data set on various classification algorithms and its analysis

5.1: Logistic Regression:

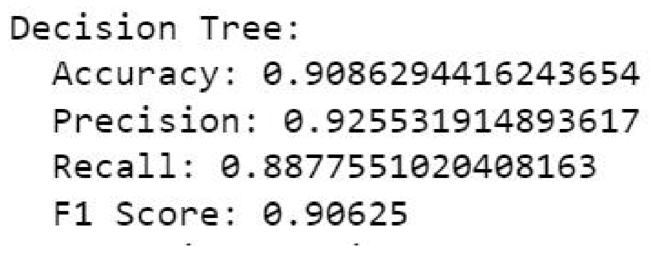


```
Logistic Regression:
  Accuracy: 0.934010152284264
  Precision: 0.967032967032967
  Recall: 0.8979591836734694
  F1 Score: 0.9312169312169313
```

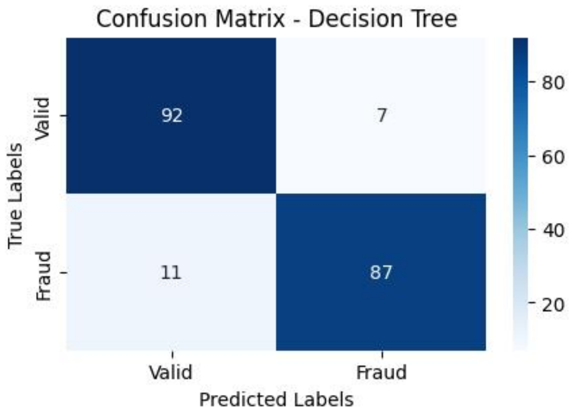


Logistic Regression achieved a high accuracy of 93.40%. It shows excellent precision (96.70%), indicating a low false positive rate, and decent recall (89.80%), suggesting a good ability to detect positive instances. The F1 score of 93.12% indicates a balance between precision and recall. The confusion matrix further illustrates that the model has correctly classified most instances, with a low number of false positives and false negatives.

5.2: Decision Tree Classifier:

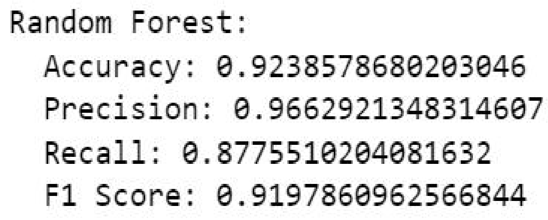


```
Decision Tree:
  Accuracy: 0.9086294416243654
  Precision: 0.925531914893617
  Recall: 0.8877551020408163
  F1 Score: 0.90625
```

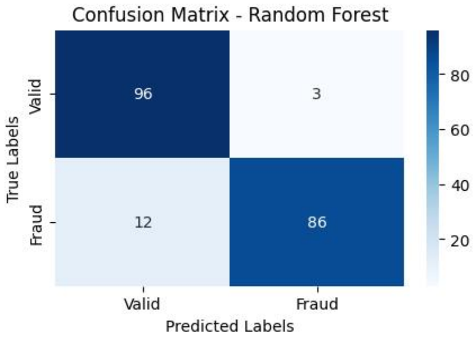


The Decision Tree model achieved an accuracy of 90.86%. It shows good precision (92.55%) and recall (88.76%), with an F1 score of 90.62%. The confusion matrix indicates that the model has made some misclassifications, particularly with false negatives, where it failed to identify positive instances correctly.

5.3: Random Forest Classifier:

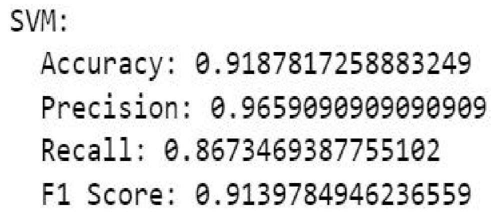


```
Random Forest:
  Accuracy: 0.9238578680203046
  Precision: 0.9662921348314607
  Recall: 0.8775510204081632
  F1 Score: 0.9197860962566844
```

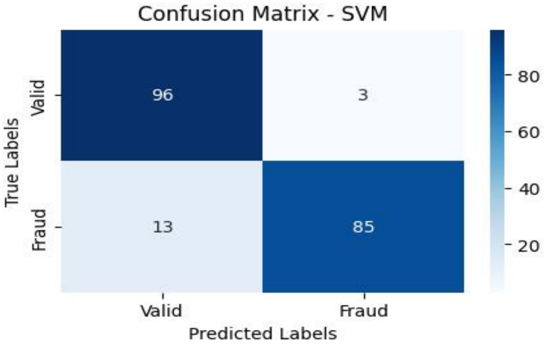


Random Forest achieved an accuracy of 92.39%. It exhibits high precision (96.63%) and moderate recall (87.76%), resulting in an F1 score of 91.98%. The confusion matrix demonstrates that the model performs well in correctly classifying most instances, although there are some false negatives.

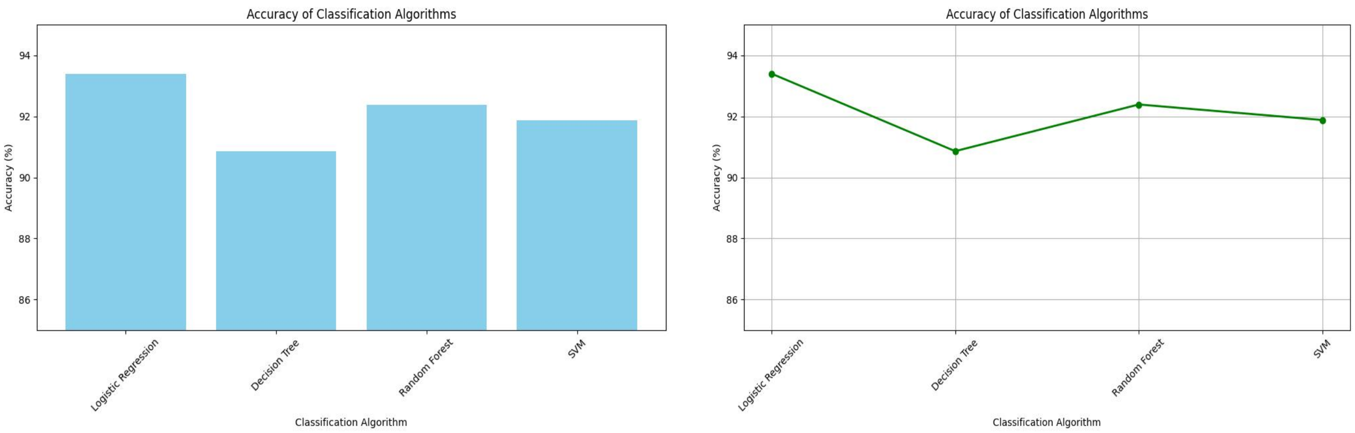
5.4: Support Vector Machine:



```
SVM:
  Accuracy: 0.9187817258883249
  Precision: 0.9659090909090909
  Recall: 0.8673469387755102
  F1 Score: 0.9139784946236559
```



The SVM model achieved an accuracy of 91.88%. It demonstrates high precision (96.59%) and moderate recall (86.73%), resulting in an F1 score of 91.40%. The confusion matrix reveals that the model has accurately classified most instances, with a slightly higher number of false negatives compared to false positives.



Visual representations of the accuracies of different algorithms on the credit card fraud transactions data set.

Chapter 6

Conclusion and Future Scope

Overall, all models performed well in classifying instances with high accuracies ranging from 90.86% to 93.40%. Logistic Regression, Random Forest, and SVM exhibited higher precision and recall compared to Decision Tree. However, Decision Tree still provided a decent balance between precision and recall. Random Forest achieved the highest precision, while Logistic Regression had the highest recall. Further analysis, such as feature importance and model interpretability, could provide additional insights into model performance and help in selecting the most appropriate model for deployment.

Looking forward within the realm of "Comparative Analysis of Machine Learning Models for Credit Card Fraud Detection," several avenues beckon for further exploration and refinement. Firstly, delving deeper into feature engineering tailored specifically for fraud detection holds promise. Techniques such as anomaly detection algorithms, behavior-based features, or transactional patterns analysis can offer insights into fraudulent activities that might not be captured by traditional features alone. By integrating domain knowledge and leveraging advanced feature engineering methods, we can enhance the models' ability to detect subtle fraud patterns, leading to more effective fraud detection systems.

Moreover, advancing model optimization and ensemble techniques tailored to the nuances of credit card fraud detection is essential. Exploring novel approaches such as semi-supervised learning, where the models learn from both labeled and unlabeled data, or incorporating temporal aspects into the modeling process to capture evolving fraud patterns over time, can further enhance detection capabilities. Additionally, investigating ensemble methods specifically designed for fraud detection, such as combining anomaly detection algorithms with supervised classifiers or utilizing model stacking techniques to leverage the strengths of different models, can lead to more robust and reliable fraud detection systems. By focusing on these avenues, we can propel the field of credit card fraud detection forward, paving the way for more sophisticated and effective fraud detection solutions in the financial sector.

References

[1]https://www.javatpoint.com/logistic-regression-in-machine-learning

[2] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[3] https://www.javatpoint.com/confusion-matrix-in-machine-learning

[4] An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine, ALTYEB ALTAHER TAHA AND SHARAF JAMEEL MALEBARY

[5] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/