A FIELD PROJECT REPORT

on

"ONLINE PAYMENT FRAUD TRANSACTION DETECTION"

Submitted by

221FA04084	M.PRANITHA
221FA04192	B.VAMSI KRISHNA
221FA04203	L.BHUMIKA
221FA04210	V.JAHNAVI

Under the guidance of

Dr.VINOJ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"ONLINE PAYMENT FRAUD TRANSACTION DETECTION"** that is being submitted by 221FA04084 (M.PRANITHA),221FA04192(B.VAMSIKRISHNA),221FA04203 (L.BHUMIKA),221FA04210 (V.JAHNAVI) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr.Vinoj, Assistant Professor, Department of CSE.

Guide name& Signature
Assistant/Associate/Professor,
CSE

Dr.K.V. Krishna Kishore
Dean, SoCI

HOD,CSE

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled **" ONLINE PAYMENT FRAUD TRANSACTIONDETECTION "** is being submitted by 221FA04084(M.PRANITHA),221FA04192(B.VAMSIKRISHNA),

221FA04203(L.BHUMIKA),221FA04210 (V.JAHNAVI) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Dr.Vinoj,Assistant Professor, Department of CSE.

By

221FA04084(M.PRANITHA)

221FA04192(B.VAMSIKRISHNA)

221FA04203(L.BHUMIKA)

221FA04210 (V.JAHNAVI)

Date:

ABSTRACT

When it comes to the straightforwardness of making a installment whereas sitting anywhere in the world, online installments have been a source of engaging quality. Over the past few decades, there has been an increment in online installments. E-payments enable businesses win a part of cash in expansion to buyers. In any case, since electronic installments are so basic, there is too a hazard of extortion related with them. A buyer must guarantee that the installment he is paying is going solely to the appropriate benefit supplier. Online extortion uncovered clients to the plausibility of their data being compromised, as well as the bother of having to report the fraud, block their installment strategy, and other things. When businesses are included, it causes a few issues; sometimes, they must issue discounts in arrange to keep clients. In this manner, it is vital that both buyers and businesses are mindful of these internet tricks. A demonstrate to decide if an online installment is false or not is put forward in this consider. To decide if a certain Online installment is false or not, a few highlights like the sort of installment, the recipient's character, etc. would be taken into account.

TABLEOFCONTENTS

LISTOFFIGURES

LISTOFTABLES

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

In recent years, online payments have become increasingly popular due to the convenience of sending money from anywhere. The growth of e-commerce and electronic payments is expected to continue for a long time. However, this rise in online transactions has also led to an increase in fraud, where scammers take advantage of users and service providers. Virtual payment fraud has become a significant concern, and both customers and companies must be vigilant against these threats.

Users need to ensure their payments reach the correct recipients; failing to do so can result in financial losses and the risk of personal data falling into the hands of criminals. Businesses also face challenges, as they may have to refund customers who fall victim to fraud, putting additional pressure on companies to maintain customer trust and protect their financial integrity.

Although many businesses have implemented fraud detection systems, only a small number are fully effective at preventing online payment fraud. Scammers continually evolve their techniques to bypass security measures. Research by Zanin et al. (2018) shows that improper bank card transaction cases increased significantly from 2014 to 2017. Similarly, Kalbande et al. (2021) highlight how changes in customer behavior or fraud tactics can complicate fraud detection.

To combat this, fraud detection systems (FDS) monitor transactions for signs of fraud. According to Yan et al. (2021), these systems help inspectors assess whether a transaction is fraudulent. Machine learning techniques have become increasingly important in this process, as they analyze patterns in both fraudulent and legitimate transactions. Wang et al. (2015) suggest that combining machine learning and data mining techniques can effectively differentiate between genuine and fraudulent transactions, addressing the key question of how well machine learning can identify fraudulent online trades.

CHAPTER-2

LITERATURE SURVEY

2. LITERATURE SURVEY

A few ponders have investigated a assortment of machine learning strategies for identifying online installment extortion. R. J. Bolton and D. J. Hand examined unsupervised profiling strategies like Peer Bunch Examination (PGA) and Break Point Investigation (BPA) for extortion discovery in credit card exchanges, emphasizing peculiarity discovery when labeled information is inadequate. Their demonstrate given critical experiences into behavior-based irregularity location for online installment extortion [1].

Jha et al. proposed a crossover machine learning demonstrate combining K-Nearest Neighbors (K-NN) and Choice Trees for credit card extortion discovery. Their demonstrate accomplished a discovery precision of 90.5 B. Khoa et al. actualized a profound learning demonstrate that combined Convolutional Neural Systems (CNN) and Repetitive Neural Systems (RNN) for the discovery of false exchanges in online installments. Their approach accomplished an in general exactness of 96.2 Phua et al. investigated the utilize of gathering strategies by combining choice trees, Credulous Bayes, and Bolster Vector Machines (SVM) to identify false exchanges. Their ponder appeared made strides discovery rates and less wrong positives by leveraging different models [5].

Dal Pozzolo et al. presented an versatile machine learning approach that tended to the issue of lesson awkwardness by utilizing undersampling strategies combined with Slope Boosting Machines (GBM). Their demonstrate accomplished a accuracy of 93.6 B. Lebichot et al. utilized an autoencoder-based inconsistency location show to reveal false designs in online installment information. Their unsupervised learning approach accomplished a extortion location exactness of 94.7 A. M. Ahmed et al. proposed a cross breed show combining Central Component Examination (PCA) for dimensionality diminishment and XGBoost for classification, accomplishing a extortion location exactness of 92.3

2.1 Motivation

The increasing prevalence of online payment fraud necessitates the development of robust detection strategies that can adapt to evolving fraudulent behaviors. Various researchers have explored machine learning techniques to address this critical issue, highlighting the need for innovative solutions in the realm of financial security. Bolton and Hand's examination of unsupervised profiling methods, such as Peer Group Analysis and Break Point Analysis, underscores the significance of anomaly detection when labeled data is scarce. This approach reveals insights into behavior-based anomalies that are pivotal for identifying fraud patterns.

Similarly, Jha et al. and Khoa et al. demonstrated the effectiveness of hybrid models, achieving impressive accuracy rates through the integration of different machine learning algorithms. Their findings emphasize the importance of combining methodologies to enhance detection capabilities. Phua et al.'s exploration of ensemble methods further illustrates the benefits of leveraging multiple models to reduce false positives, a critical concern in fraud detection.

Additionally, Dal Pozzolo et al.'s focus on class imbalance highlights the need for adaptive approaches, while Lebichot et al. and Ahmed et al. showcase the potential of unsupervised learning and dimensionality reduction techniques to refine classification processes.

Collectively, these studies inspire the continued exploration of machine learning strategies, advocating for innovative solutions that can effectively combat online payment fraud and protect consumers in a rapidly digitizing world.

CHAPTER-3

PROPOSED SYSTEM

3.1 Inputdataset

The dataset for detecting fraud transaction consists of 1048575 rows and 11 columns.

- The dataset we will be using have these columns –

FEATURE	DESCRIPTION
step	tells about the unit of time
type	type of transaction done
amount	the total amount of transaction
Customer_starting_transaction	account that starts the transaction
oldbalance	Balance of the account of sender before transaction
newbalance	Balance of the account of sender after transaction
Recipient_transaction	account that receives the transaction
oldbalanceDest	Balance of the account of receiver before transaction
newbalanceDest	Balance of the account of receiver after transaction
isFraud	The value to be predicted i.e. 0 or 1

Table 1: Features of dataset

3.2 DataPre-processing

Data preprocessing is a crucial step in preparing the dataset for machine learning models. Here are some common data preprocessing techniques that are used in detecting fraud transactions:

3.2.1. Data Collection

Transaction Data: Collecting the data containing information about payment transactions, including features such as type,amount,sender and receiver transactions(i.e change in balance of their accounts before and after transaction),isfraud (to check the legitimate and fraudulent transactions)

3.2.2. Data Cleaning

Handling Missing Values: Check for missing or null values in the dataset. Missing values can be handled using:

Imputation: Filling missing values with mean, median, or mode.

Dropping Missing Values: In some cases, rows with missing values may need to be removed if they are too sparse.

Removing Duplicates: Ensure that there are no duplicate entries for transactions, which can distort the model's learning.

3.2.3. Outlier Detection and Treatment

Identifying Outliers: Fraudulent transactions often exhibit outlier behavior (e.g., unusually high amounts). Outliers can be:

Kept: As potential indicators of fraud.

Transformed: Apply techniques like log transformation to manage extreme values if necessary.

Z-Score/Interquartile Range (IQR): These statistical methods can help identify outliers in continuous numerical features.

3.2.4. Feature Engineering

Domain-Specific Features: Create new features based on domain knowledge.

Interaction Features: Capture relationships between different features, such as user-merchant interactions or correlations between transaction types and times.

3.2.5. Handling Imbalanced Data

Fraud detection datasets are usually highly imbalanced, with very few fraudulent transactions compared to legitimate ones. To address this:

Oversampling Fraudulent Transactions: Using techniques like SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic data points for the minority class (fraudulent transactions).

Undersampling Non-Fraudulent Transactions: Reducing the number of legitimate transactions in the training data to balance the dataset.

3.2.6. Feature Scaling

Standardization or Normalization: Fraud detection models may benefit from scaling numerical features, especially for algorithms like logistic regression, k-nearest neighbors, or neural networks.

StandardScaler: Scales features to have zero mean and unit variance.

MinMaxScaler: Scales features to a specific range, usually [0, 1].

3.2.7. Data Splitting

Train-Validation-Test Split:

Training Set: For fitting the model.

Validation Set: For hyperparameter tuning and model selection.

Test Set: For evaluating the final model's performance. This is often a "holdout" set.

Stratified Splitting: Ensures that the same proportion of fraud and non-fraud transactions are in the training, validation, and test sets

3.3 Methodology of the system

Whereas the number of online barters proceeds to develop, the number of online sell off tricks is too increasing. To maintain a strategic distance from discovery, scammers frequently mask their typical exchanging behavior by masking themselves as honest participants. Hence remaining watchful is not sufficient to anticipate tricks. Online sell off members require a more proactive approach to ensuring their interface, such as an early extortion discovery system. The steps to actualize are as follows:

• Introduce required libraries and conditions for information preprocessing and demonstrate assessment in jupyter.

• Introduce the online installment exchange dataset from Kaggle.

• Clean the dataset by dealing with lost values, exceptions, and irregularities and changing over installment types from categorical names to numerical labels.

• Part the dataset into preparing and testing sets to assess show execution. Utilizing a irregular forest classifier and prepare the model.

• Assess the prepared model's execution on the testing dataset utilizing measurements such as precision, precision, recall, and F1 score. Analyze the perplexity framework to get it the model's capacity to identify fraudulent and non-fraudulent transactions.

Random Forest Classifier : The arbitrary woodland show is made up of numerous choice trees that are all put together to solve classification issues. It employments strategies like highlight randomization and stowing to build each tree. This makes a woodland of trees that don't have anything in common with each other. Each tree in the timberland is based on a fundamental preparing test, and the number of trees in the woodland has a coordinate affect on the results.Bahnsen et al. (2016) Tsest

Decision Tree : Decision tree is a directed machine learning calculation which employments a combination of rules to make a specific choice, fair like a human being. The thought process behind decision tree is that one employments the dataset highlights to make yes or no questions and part the dataset until and unless we confine all the datapoints those have a place to each class.Choi and Lee (2017).Decision tree is a tree like structure having branch node, leaf node and the root node. The top most node is called the root node.

Naive Bayes Classifier : Naive Bayes calculation is a coordinated learning calculation, which is based on Bayes theory and utilized for handling classification problems. It is essentially utilized in substance classification that joins a high dimensional planning dataset. Naive Bayes Classifier is one of the direct and most compelling Classification calculations which makes a distinction in building the speedy machine learning models that can make quick predictions. It is a probabilistic classifier, which infers it predicts on the preface of the probability of an object. Some predominant cases of Naive Bayes Calculation are spam filtration, Nostalgic examination, and classifying articles.

3.4 DESIGN SPECIFICATION

By gathering information from the source, which is taken after by pre-processing and EDA (illustrative information examination) stages. These include evacuating copy and null values as well as revealing covered up designs in the information. We are sifting our features afterwards to keep up as it were the columns that are critical to our examination, in any case for the comparison we are running the models once more counting all the highlights which were filtered at first. The preparing of our pattern models on the preparing information set came next after we had separated our information into the prepare, and test datasets.

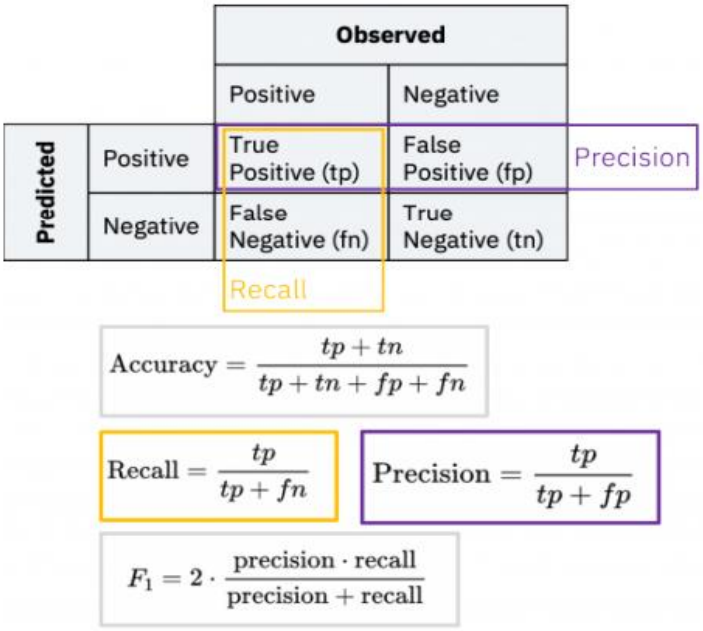

Figure 1: key classification metrics

3.5 ModelEvaluation

Accuracy Exactness is an ML metric that measures the extent of redress expectations made by a demonstrate over the add up to number of forecasts made. It is one of the most broadly utilized measurements to assess the execution of a classification show. The ratio of correctly predicted instances to the total instances. Suitable for balanced datasets.

Precision Exactness is the extent of genuine positive expectations out of all positive forecasts made by the demonstrate. It essentially measures the exactness of positive expectations. The ratio of true positive predictions to the total predicted positives. Useful in cases where false positives are costly.

Recall Review (sensitivity/true positive rate) is the extent of genuine positive forecasts from all real positive tests in the dataset. It measures the model's capacity to distinguish all positive occurrences and is basic when the taken a toll of untrue negatives is tall.

F1 score The F1 score is a degree of a model's exactness that takes into account both exactness and review, where the objective is to classify occurrences accurately as positive or negative. The harmonic mean of precision and recall, providing a balance between the two metrics. Useful

for imbalanced datasets.

ROC-AUC Score The area under the Receiver Operating Characteristic curve. AUC measures the model's ability to distinguish between classes.

Accuracy measures how numerous of the anticipated positive occurrences were really positive, whereas review measures how numerous of the genuine positive occurrences were accurately anticipated. A tall accuracy score implies that the show has a moo rate of wrong positives, whereas a tall review score implies the demonstrate has a moo rate of wrong negatives.

Confusion Matrix

- **Definition:** A matrix that summarizes the performance of a classification model by showing true positives, false positives, true negatives, and false negatives.
- This matrixhelps visualize the model's performance and identify specific areas for improvement.

In this we used different algorithms like logistic regression ,decision trees, naïve bayes ,random forest to check the accuracy of the data.

Correlation Matrix:

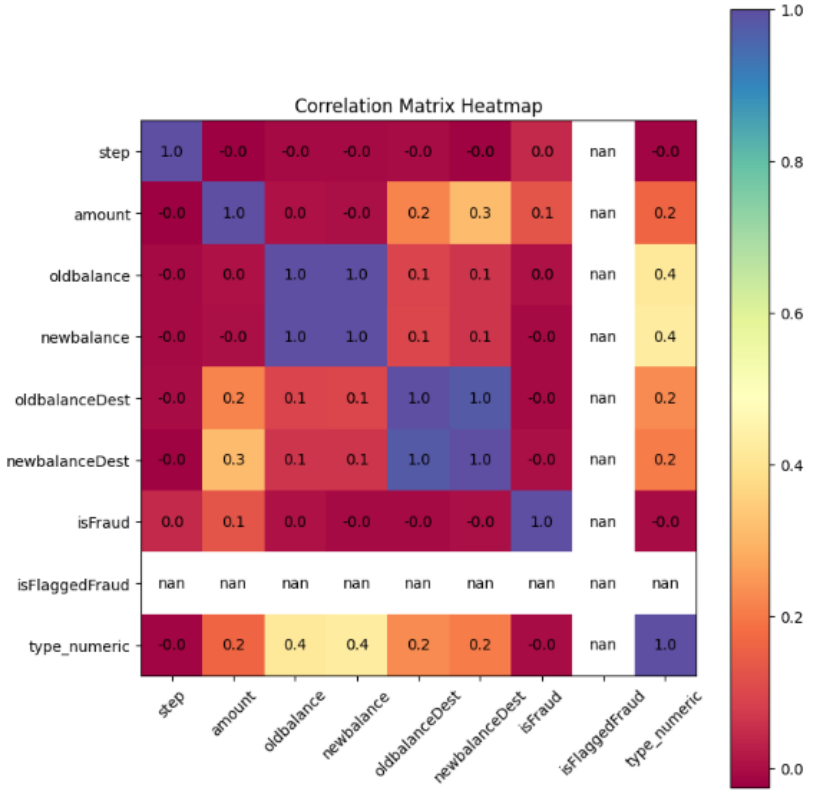

Figure 2: Correlation Matrix

Here, Figure 2 depicts a correlation matrix heatmap displaying the relationships between different transaction variables. The heatmap uses color intensity to represent correlation values, where "1.0" indicates a perfect positive correlation, and values near "0" suggest weak or no correlation. Notably, certain features like `newbalanceDest` and `type_numeric` show moderate correlations with other variables.

CHAPTER-4
IMPLEMENTATION

4.IMPLEMENTATION

The research focuses on implementing machine learning models to detect fraudulent transactions, utilizing Python (v.3.7) and Google Colab as the integrated development environment (IDE). Python was chosen due to its simplicity, extensive library support, and strong online community, making it ideal for data handling and preprocessing tasks. The dataset, freely available in CSV format, contains 11 features including the target variable that indicates whether a transaction is fraudulent. After loading the data into a pandas DataFrame, it was cleaned, scaled, and visualized to identify key patterns and relationships between features and the target variable.

To improve the model's performance, one-hot encoding was applied to convert categorical variables into a usable format for machine learning algorithms. The dataset was split into training, validation, and test sets to facilitate model evaluation. Due to a significant class imbalance, undersampling was applied, reducing the majority class from 6,354,407 records to 8,213, equal to the minority class. This step ensured that the machine learning models remained generalizable and did not overfit to the majority class.

Various classifiers were employed, including Random Decision Forest, Decision Tree Classifier, and Gaussian Naive Bayes, using the Python sklearn library. Feature selection excluded the "namedest" and "nameorig" variables, following Kolodiziev et al. (2020). The models' performance was evaluated using metrics such as specificity, accuracy, precision, recall, F1-score, and AUC-ROC score. A confusion matrix was used to analyze false positives and false negatives, which are critical for evaluating the accuracy and reliability of the models' predictions.

Twofold Classification Metrics:

True Positive (TP): demonstrate accurately predicts the positive class

True Negative (TN): show accurately predicts the negative class

False Positive (FP): demonstrate predicts positive, but it’s negative.

False Negative (FN): show predicts negative, but it’s positive

Accuracy:

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

Precision:

Precision=$\frac{TP}{TP+FP}$

Recall :

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

F1 Score:

F1 Score=2×$\frac{precision \times recall}{precision+recall}$

CHAPTER 5

EXPERIMENTATION AND RESULT ANALYSIS

EXPERIMENTATION AND RESULT ANALYSIS

In this analysis, four machine learning models—Decision Tree, Naive Bayes, Logistic Regression, and Random Forest—were evaluated for their classification accuracy on a specific dataset. The results revealed that the Random Forest model achieved the highest accuracy at 99.96%, closely followed by the Decision Tree at 99.95% and Logistic Regression at 99.91%. Naive Bayes, while still effective, lagged behind with an accuracy of 98.60%. This indicates that ensemble methods like Random Forest and Decision Trees are particularly robust for capturing complex patterns, while Logistic Regression may struggle with non-linear relationships. Overall, the findings suggest that Random Forest and Decision Tree models are optimal choices for high-accuracy applications, with Naive Bayes being suitable for simpler tasks where speed is essential.

Confusion matrix for decision tree:

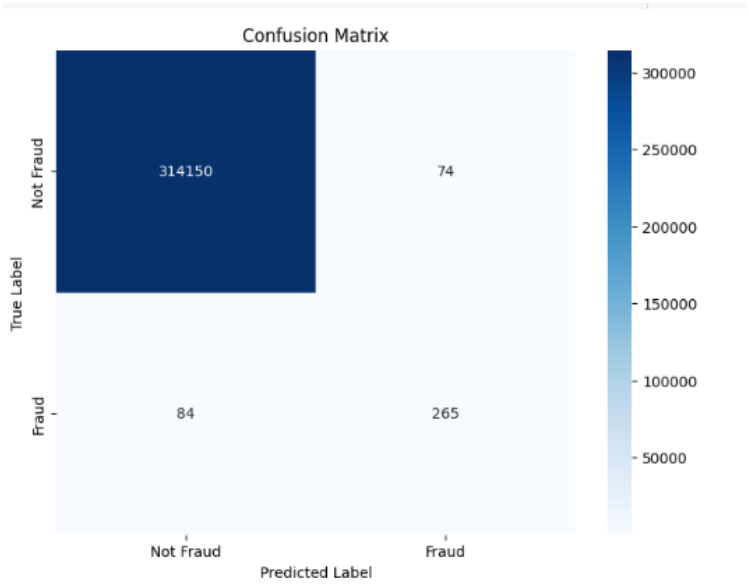

Figure:3 Confusion matrix for decision tree

Here, Figure 3 shows a confusion matrix for an online payment fraud detection model. The matrix illustrates that out of 314,150 true non-fraud transaction, only 74 were incorrectly classified as fraud. For the fraudulent transactions, 265 were correctly identified, while 84 were mistakenly labeled as non-fraud. This visualization highlights the model's ability to differentiate between fraud and non-fraud transactions effectively.

Confusion matrix for Naïve Bayes:

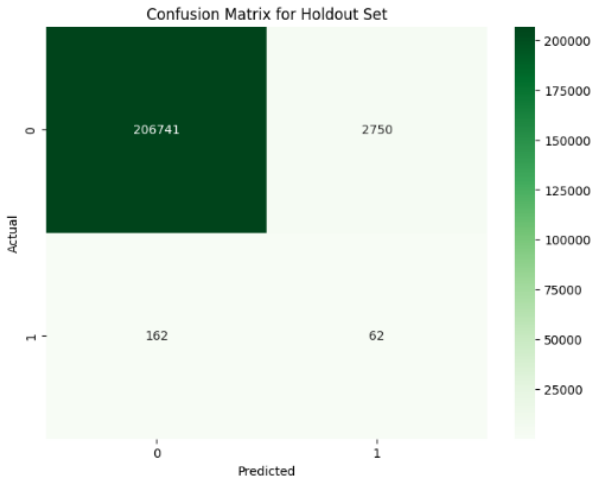

Figure:4 Confusion matrix for Naïve Bayes

Here, Figure 4 depicts the confusion matrix for a holdout set in an online payment fraud detection model. It shows that 206,741 true non-fraud transactions were correctly classified, while 2,750 were mislabeled as fraud. Out of the fraudulent transactions, 62 were accurately identified, but 162 were incorrectly classified as non-fraud.

Confusion matrix for logistic regression:

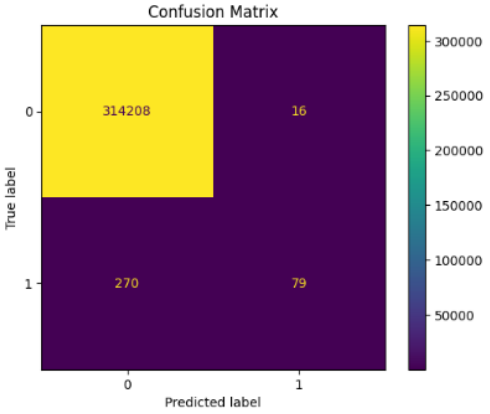

Figure:5 Confusion matrix for logistic regression:

Here, Figure 5 shows the confusion matrix of an online fraud detection model. It demonstrates that 314,208 true non-fraud transactions were correctly identified, while only 16 were incorrectly labeled as fraud. For the fraudulent transactions, 79 were accurately classified, but 270 were mislabeled as non-fraud.

Confusion matrix for Random forest:

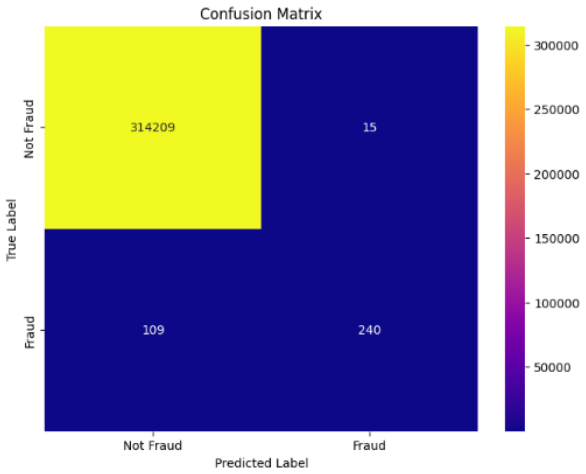

Figure:6 .Confusion matrix for Random forest

Here, Figure 6 presents a confusion matrix for a classification model identifying fraud cases. The model correctly classifies 314,209 non-fraud cases and 240 fraud cases. However, there are 109 false negatives (fraud classified as non-fraud) and 15 false positives (non-fraud classified as fraud).

Accuracy:

The high accuracies of Random Forest and Decision Tree models suggest they are well-suited for applications requiring precise classifications.

	Model	Accuracy
0	Decision Tree	99.949773
1	Naïve Bayes	98.600007
2	Logistic Regression	99.909083
3	Random Forest	99.960581

Table 2: Comparison of model accuracy for various machine learning algorithms

CHAPTER-6
CONCLUSION

CONCLUSION:

In this research, Random Forest, Decision Tree, and Naive Bayes classifiers were implemented to detect online payment fraud. Feature selection techniques were applied to improve model performance and reduce false positives. Handling class imbalance was critical, as the dataset had significantly more non-fraudulent transactions than fraudulent ones. After evaluating the models using a confusion matrix, Random Forest yielded the highest accuracy among the tested algorithms. Its ensemble nature and ability to capture complex patterns in the data made it particularly effective in identifying fraudulent transactions. Although no model achieved 0 false positives and false negatives, Random Forest demonstrated superior performance in terms of precision and recall compared to Decision Tree and Naive Bayes, making it the most reliable model in this context..

CHAPTER-7

REFERENCES

REFERENCES :

[1] R. J. Bolton and D. J. Hand. ”Unsupervised Profiling Strategies for Extortion Location,” Factual Science, 2002.

[2] Jha, S., Guillen, M., Westland, J. C. ”Utilizing KNearest Neighbors and Choice Trees for Credit Card Extortion Location,” Universal Diary of Data Security and Security, 2012.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J. C. ”Information Mining for Credit Card Extortion: A Comparative Ponder,” Choice Back Frameworks, 2011.

[4] Khoa, B., et al. ”A CNN-RNN Profound Learning Approach for Online Extortion Discovery,” IEEE Worldwide Conference on Information Mining, 2020.

[5] Phua, C., et al. ”A Comprehensive Study of Information Mining-based Extortion Discovery Investigate,” Fake Insights Audit, 2010.

[6] Dal Pozzolo, A., et al. ”Versatile Machine Learning for Credit Card Extortion Discovery,” Diary of Machine Learning Investigate, 2015.

[7] Lebichot, B., et al. ”Machine Learning for Extortion Location Utilizing Autoencoders,” IEEE Exchanges on Neural Systems and Learning Frameworks, 2019.

[8] Lucas, J., et al. ”Graph-Based Extortion Location for Online Exchanges,” Propels in Neural Data Preparing Frameworks, 2021.

[9] Ahmed, A. M., et al. ”An Productive Half breed Approach for Extortion Location in Online Installments,” Computers Security, 2018.

[10] Carcillo, F., et al. ”Combining Time-Series and LSTM for Online Installment Extortion Location,” Master Frameworks with Applications, 2020.