Advanced Fraud Detection: Leveraging K-SMOTEENN and Stacking Ensemble to Tackle Data Imbalance and Extract Insights

Authors

G. Lakpathi [1], Ms. Chirra Pushpalatha [2], Ms. Divya Sood Tagala [3], Mr. Gudepu Shashi Vardhan [4]

[1] Assistant Professor, Dept. of Computer Science & Engineering, Guru Nanak Institute of Technology, India

[234] scholars, Dept. of Computer Science & Engineering, Guru Nanak Institute of Technology, India

ABSTRACT

This design presents a robust system for detecting credit card fraud using a mongrel model that combines Random Forest with K- means SMOTEENN, a fashion specifically designed to attack class imbalance. The approach intelligently creates synthetic nonage samples and removes noise to insure a cleaner, more balanced dataset. By integrating Random Forest — a model known for its rigidity to complex data and resolvable AI(LIME), the system becomes both accurate and interpretable. This combination not only enhances vaticination trustability but also offers translucency, making it a strong result for fraud discovery in finance.

KEYWORDS

Fraud Detection, Random Forest, K- SMOTEENN, Data Imbalance, Stacking Ensemble, resolvable AI, LIME

1. INTRODUCTION

The system introduces a new fraud discovery fashion by incorporating Random Forest with the K- means SMOTEENN testing strategy. This integration addresses disposed datasets while limiting overfitting. also, it leverages LIME — a tool in resolvable AI — to give accessible perceptivity into how prognostications are made. This combination strengthens both the model's delicacy and interpretability, making it well- suited for practical use in fiscal fraud forestallment.

2. LITERATURE REVIEW

Title: Deep Learning Ensemble with Resampling for Credit Card Fraud

Authors: Ibomoiye Domor Mienye, Yanxia Sun(2023)

Summary: The paper explores a mongrel of deep literacy and resampling styles to address severe data imbalance, perfecting literacy from rare fraud cases through ensemble modeling.

Title: Ensemble Learning & Data Augmentation for Imbalanced Problems

Authors: Azal Khan, Omkar Chaudhari, Rohitash Chandra(2023)

Summary: This study highlights that introductory resampling ways like SMOTE can outperform advanced models like GANs, offering practical, low- cost results to ameliorate performance on imbalanced datasets.

Title: Soft Voting Ensemble for Fraud Detection

Authors: Mimusa Azim Mim, Nazia Majadi, Peal Mazumder(2024)

Summary: This exploration presents a soft voting ensemble system that effectively deals with imbalanced fraud data, delivering high perfection and recall in real- world sale discovery.

3. METHODOLOGY

MODULE NAMES

1. Organizing Data
2. Comprehending Trends
3. Formatting Information
4. Performing Computations
5. Aligning Model Predictions
6. Execution Quality
7. Determining Future Scenarios

MODULES EXPLANATION

1. **Organizing Data**
Transaction data is gathered and gutted to ensure quality and thickness for farther processing.
2. **Comprehending Trends**
Exploratory analysis identifies unusual patterns that may gesture fraudulent conditioning.
3. **Formatting Information**
Data preprocessing includes normalization, garbling, and point selection to enhance model performance.
4. **Performing Computations**
The dataset is balanced using K- means SMOTEENN, also reused using a Random Forest model to make prophetic capabilities.
5. **Aligning Model Predictions**
Predictions are meliorated using hyperparameter tuning and interpreted using LIME to insure explainability.
6. **Execution Quality**
Model delicacy is tested with criteria similar as AUC, recall, and perfection to confirm its trustability.
7. **Determining Future Scenarios**
The system is designed to acclimatize over time, streamlining itself with new data to stay effective in dynamic fraud surroundings.

4. EXISTING SYSTEM

K- Nearest Neighbors(KNN) is a traditional algorithm that compares new deals to their closest known exemplifications. Although simple, it suffers in large or imbalanced datasets, frequently misclassifying nonage class cases. Its reliance on distance- grounded comparison also leads to inefficiencies in high- dimensional data spaces.

DISADVANTAGES OF EXISTING SYSTEM

- **Computationally Precious:** These systems demand high processing power, making them hamstrung for large- scale or real- time fraud discovery.

- **Sensitive to Inapplicable Features:** Inapplicable inputs can mislead the model, reducing its delicacy in relating fraud.

- **Poor Performance with Imbalanced Data:** They struggle to accurately identify rare fraudulent cases when legitimate transactions They frequently misclassify rare fraudulent cases due to the dominance of licit deals.

- **Difficulty in High-Dimensional Spaces:** Performance drops when handling datasets with numerous features, as distance criteria come less meaningful.
- **Requires careful selection of the' K' parameter:** The model's delicacy heavily depends on choosing the right value of' K', which is frequently non-trivial.

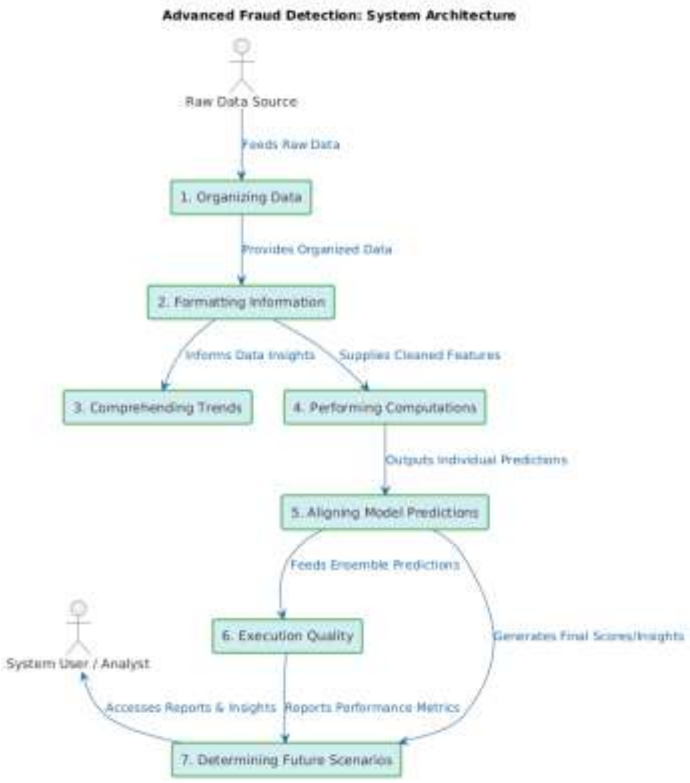
5. PROPOSED SYSTEM

Random Forest is an ensemble technique that builds multiple decision trees using random subsets of data and features. For credit card fraud detection, it combines tree predictions through majority voting, enhancing accuracy and reducing overfitting. Its robustness to noise, outliers, and imbalanced data makes it well-suited for fraud detection. Coupled with K-means SMOTEENN for balancing classes, Random Forest effectively learns patterns from rare fraudulent transactions while providing interpretability through feature importance insights.

ADVANTAGES OF PROPOSED SYSTEM

- **High Accuracy:** The proposed system achieves superior precision in identifying credit card fraud.
- **Robust to Overfitting:** It maintains strong performance on new, unseen data, avoiding excessive fitting to training data.
- **Handles Large Datasets Well:** The system is designed to efficiently process and learn from extensive volumes of transaction data.
- **Feature Importance:** It can identify which specific data characteristics are most influential in determining fraud.
- **Handles Both Regression and Classification:** The system is versatile enough to manage tasks involving both continuous value prediction and categorization.

6. SYSTEM ARCHITECTURE


Advanced Fraud Detection: System Architecture
Raw Data Source
Feeds Raw Data
1. Organizing Data
Provides Organized Data
2. Formatting Information
Informs Data Insights
Supplies Cleaned Features
3. Comprehending Trends
4. Performing Computations
Outputs Individual Predictions
5. Aligning Model Predictions
Feeds Ensemble Predictions
Generates Final Scores/Insights
6. Execution Quality
System User / Analyst
Accesses Reports & Insights
Reports Performance Metrics
7. Determining Future Scenarios

This architecture illustrates a step-by-step fraud detection pipeline. It begins with organizing raw data, followed by formatting and preprocessing. Trends are analyzed and machine learning computations are performed. Predictions are aligned and evaluated for quality. Final results and insights are shared with analysts, enabling informed decisions and system improvement through continuous monitoring and scenario forecasting.

7. IMPLEMENTATION

The model was evaluated using a public credit card dataset with a severe class imbalance. The system starts with a homepage offering options for user registration and login. After logging in, users are prompted to enter transaction details such as ‘step,’ ‘type,’ ‘amount,’ ‘oldbalanceOrg,’ ‘newbalanceOrig,’ ‘oldbalanceDest,’ and ‘newbalanceDest,’ which correspond to features in the credit card fraud dataset.

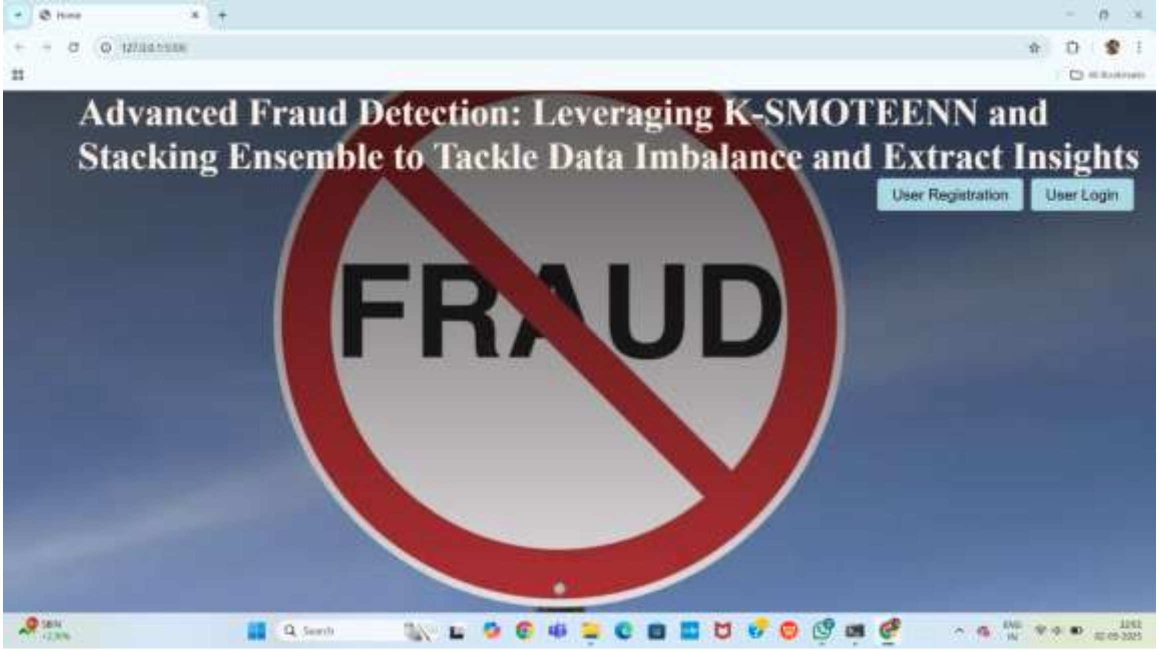
Once the user inputs the details and clicks ‘Predict,’ the system processes the data through the trained model to determine if the transaction is fraudulent or legitimate. The prediction result is then displayed on a new page, which also provides an option to perform another prediction.

Additionally, the result page includes a ‘Performance Analysis’ button that displays a pie chart showing the distribution of legitimate (0) and fraudulent (1) transactions, giving users insight into the dataset’s composition and the model’s effectiveness.

The independent variables in the dataset represent transaction attributes, while the dependent variable, ‘isFraud,’ indicates whether a transaction is legitimate or fraudulent. This setup allows the system to accurately predict fraud based on input transaction features.

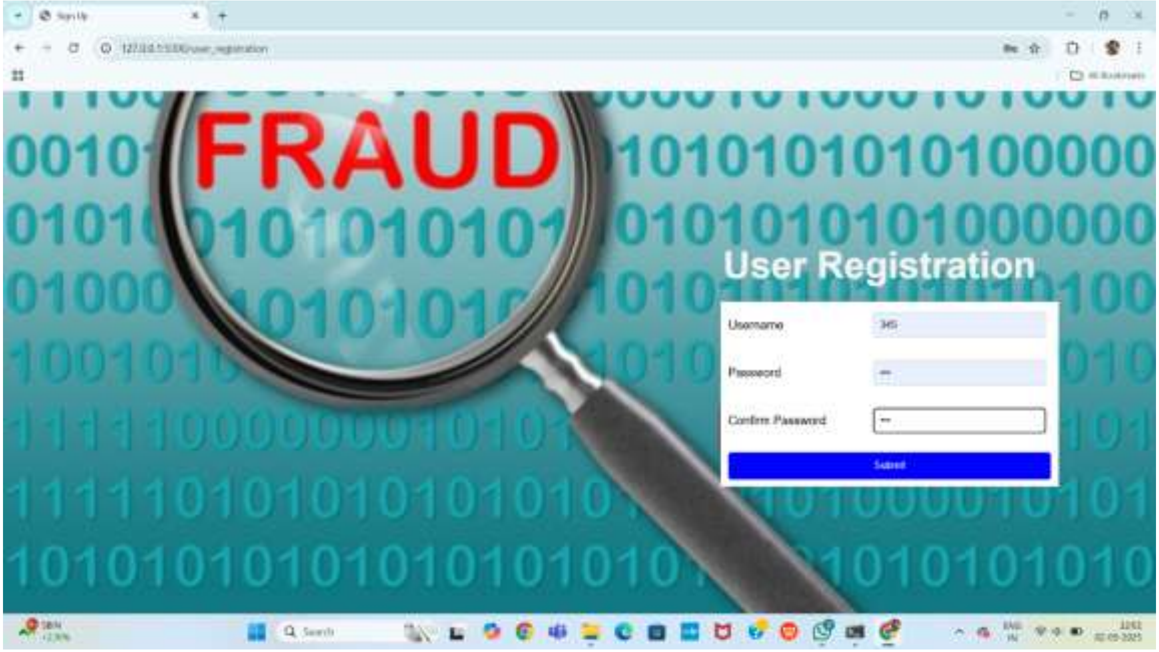
8. EXPERIMENTAL RESULTS

RESULTS



Advanced Fraud Detection: Leveraging K-SMOTEENN and Stacking Ensemble to Tackle Data Imbalance and Extract Insights

HOME PAGE

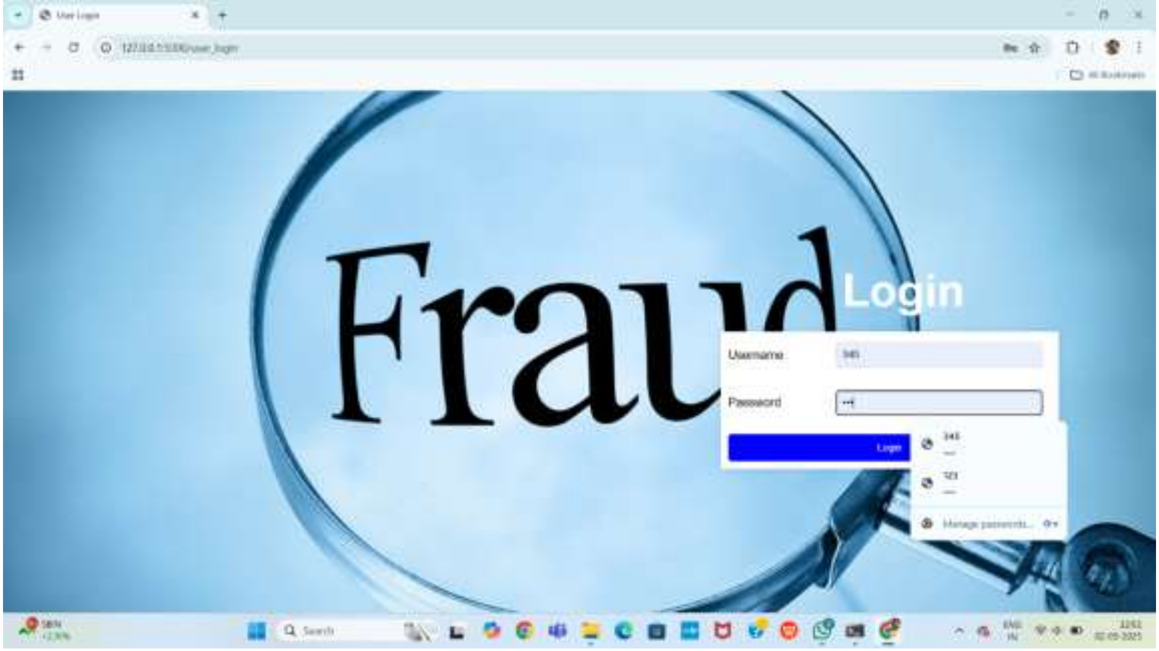


Username 345

Password •••

Confirm Password •••

USER REGISTRATION PAGE



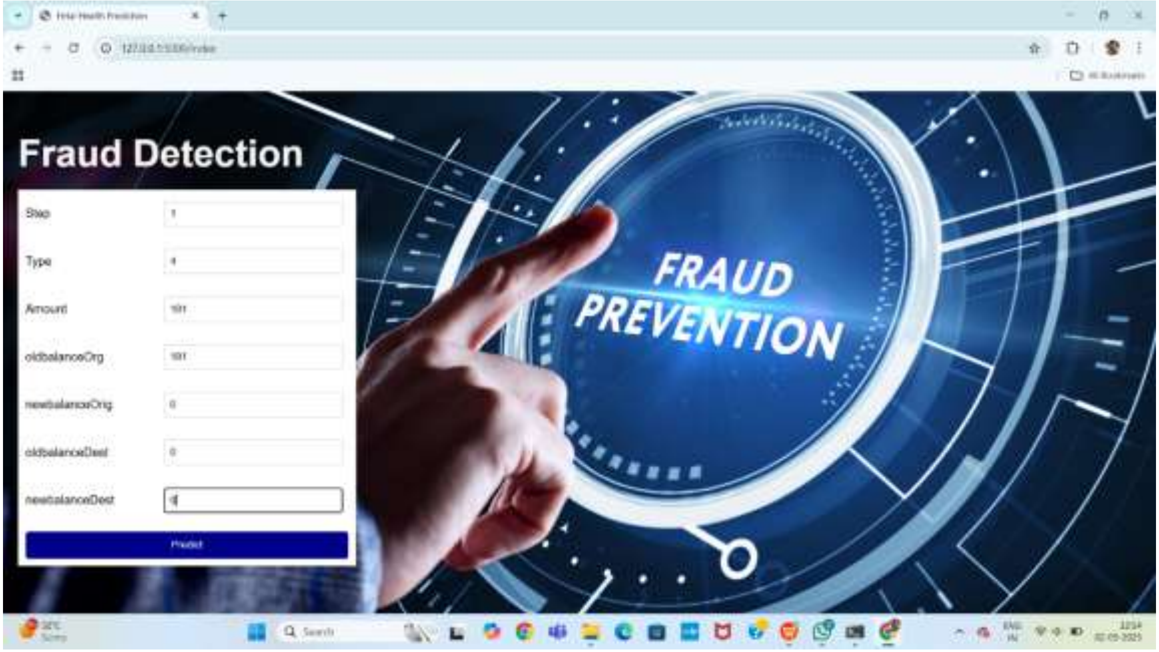
Username 345

Password •••

345

123

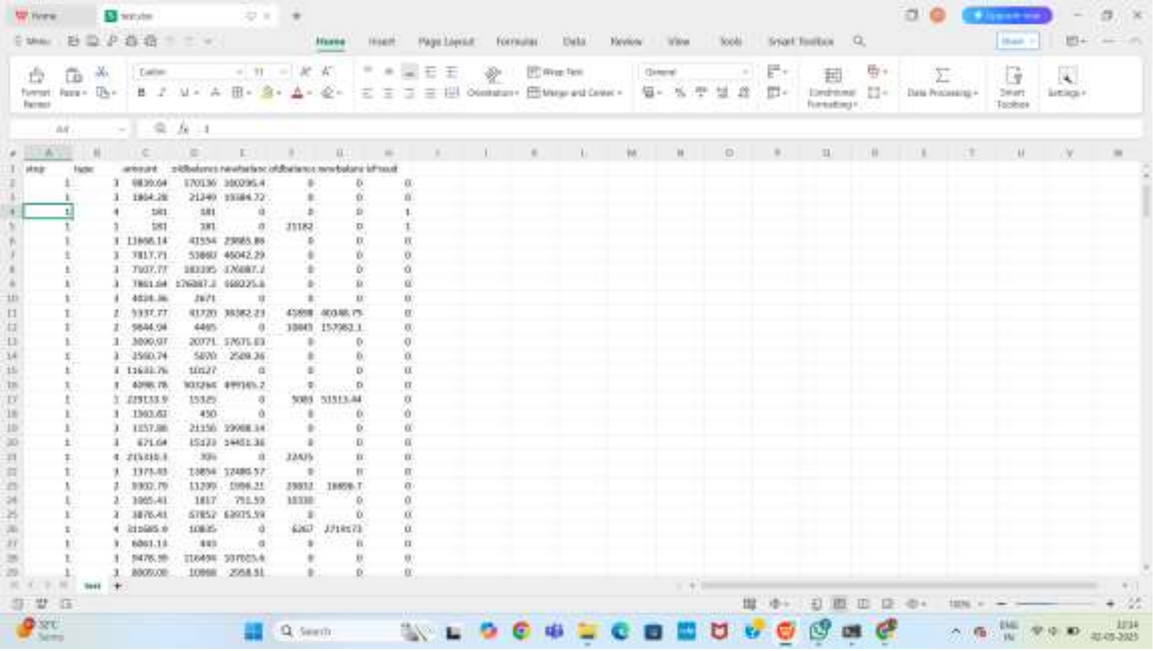
LOGIN PAGE



Fraud Detection

Step	1
Type	4
Amount	181
oldbalanceOrg	181
newbalanceOrig	0
oldbalanceDest	0
newbalanceDest	0

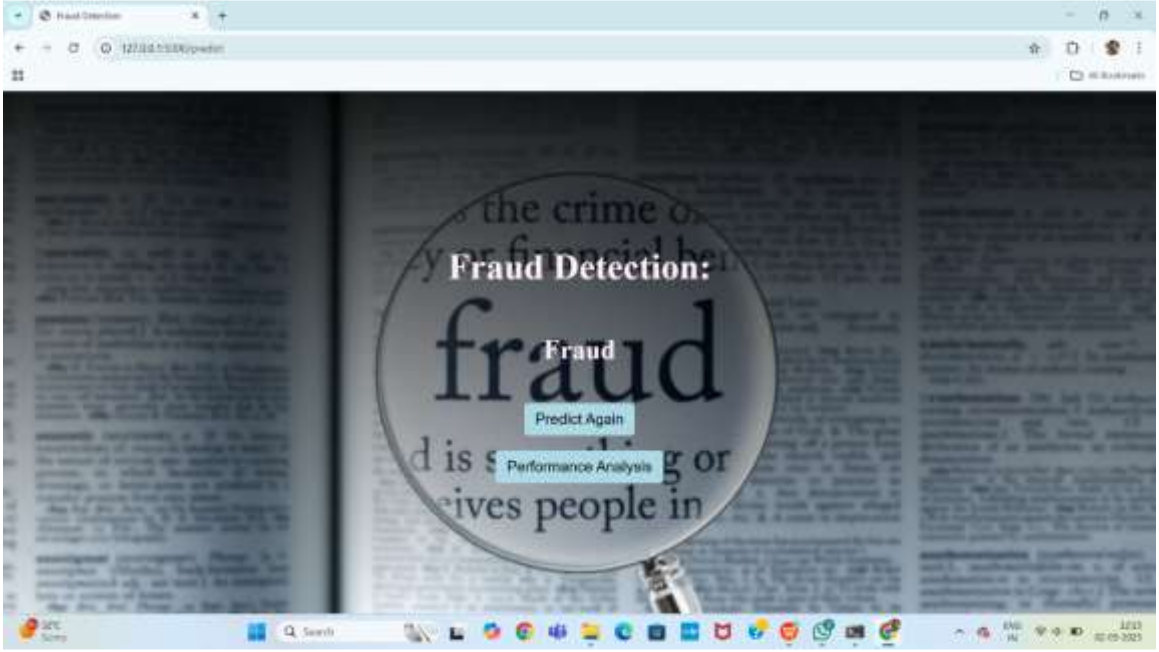
ENTERING DETAILS



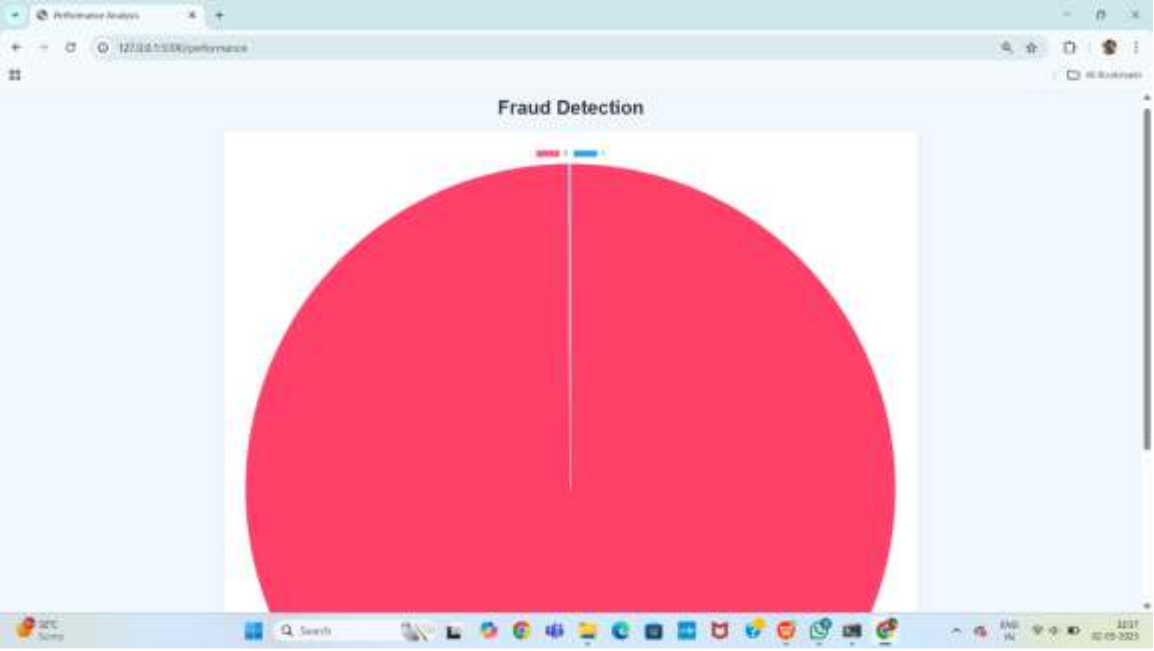
A4 fx 1

step	type	amount	oldbalanc	newbalanc	oldbalanc	newbalanc	isFraud
1	3	9839.64	170136	160296.4	0	0	0
1	3	1864.28	21249	19384.72	0	0	0
1	4	181	181	0	0	0	1
1	1	181	181	0	21182	0	1
1	3	11668.14	41554	29885.86	0	0	0
1	3	7817.71	53860	46042.29	0	0	0
1	3	7107.77	183195	176087.2	0	0	0
1	3	7861.64	176087.2	168225.6	0	0	0
1	3	4024.36	2671	0	0	0	0
1	2	5337.77	41720	36382.23	41898	40348.79	0
1	2	9644.94	4465	0	10845	157982.1	0
1	3	3099.97	20771	17671.03	0	0	0
1	3	2560.74	5070	2509.26	0	0	0
1	3	11633.76	10127	0	0	0	0
1	3	4098.78	503264	499165.2	0	0	0
1	1	229133.9	15325	0	5083	51513.44	0
1	3	1563.82	450	0	0	0	0
1	3	1157.86	21156	19998.14	0	0	0
1	3	671.64	15123	14451.36	0	0	0
1	4	215310.3	705	0	22425	0	0
1	3	1373.43	13854	12480.57	0	0	0
1	2	9302.79	11299	1996.21	29832	16896.7	0
1	2	1065.41	1817	751.59	10330	0	0
1	3	3876.41	67852	63975.59	0	0	0
1	4	311685.9	10835	0	6267	2719173	0
1	3	6061.13	443	0	0	0	0
1	3	9478.39	116494	107015.6	0	0	0
1	3	8009.09	10968	2958.91	0	0	0

DATA STORED IN DATABASE



RESULT PAGE



PERFORMANCE ANALYSIS OF THE RESULT

9. CONCLUSION AND FUTURE WORK

Credit card fraud detection is critical for financial security, and this research presents a robust system combining Random Forest with K-means SMOTEENN to balance classes and reduce noise. Using Explainable AI like LIME improves transparency and trust, enhancing detection accuracy while addressing overfitting. The system effectively handles complex transaction patterns, offering valuable protection for financial institutions.

Future improvements include real-time analytics for instant fraud detection, deep learning models like LSTM for capturing intricate patterns, and unsupervised methods to detect new fraud trends. Incorporating blockchain, adaptive learning, hybrid models, and cloud deployment will further enhance security, scalability, and interpretability.

REFERENCES

[1] Mienye, I. D., & Sun, Y. (2023). A Deep Learning Ensemble with Data Resampling for Credit Card Fraud Detection. IEEE Access.
[2] Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A Review of Ensemble Learning and Data Augmentation Models. Expert Systems with Applications.
[3] Ni, L., Li, J., Xu, H., & Wang, X. (2023). Fraud Feature Boosting Mechanism for Credit Card Fraud Detection.
[4] Mim, M. A., Majadi, N., & Mazumder, P. (2024). A Soft Voting Ensemble Learning Approach for Credit Card Fraud Detection.
[5] Mienye, I. D., & Jere, N. R. (2024). Deep Learning for Credit Card Fraud Detection: A Review.