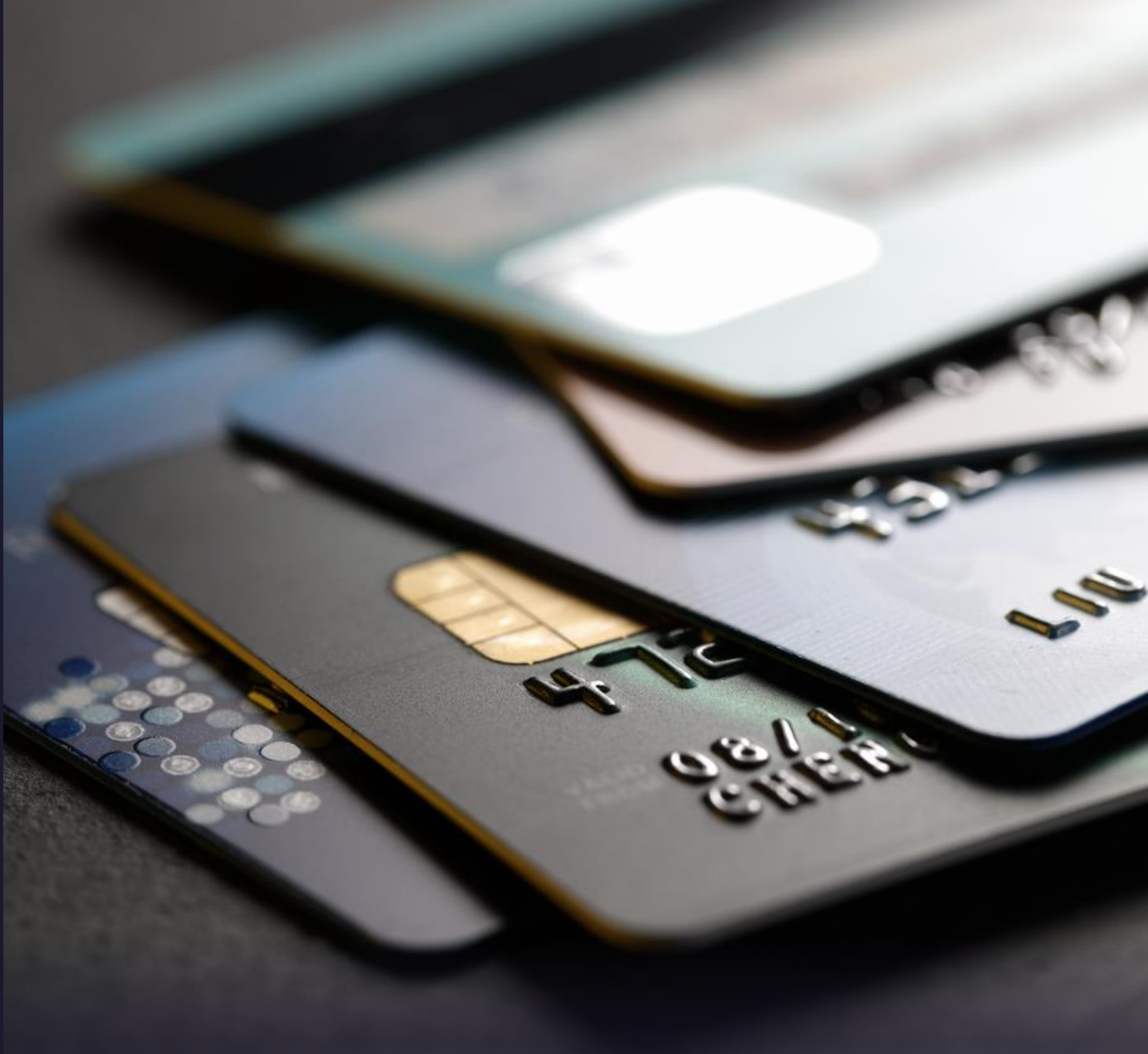
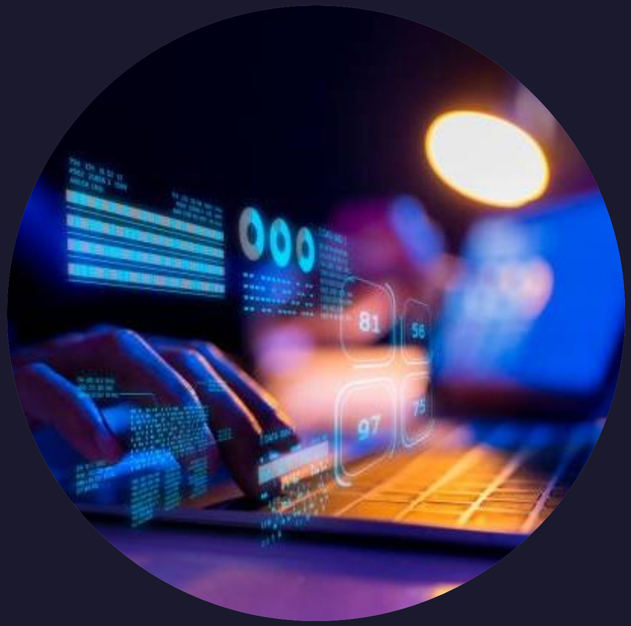
Credit Card Fraud Detection



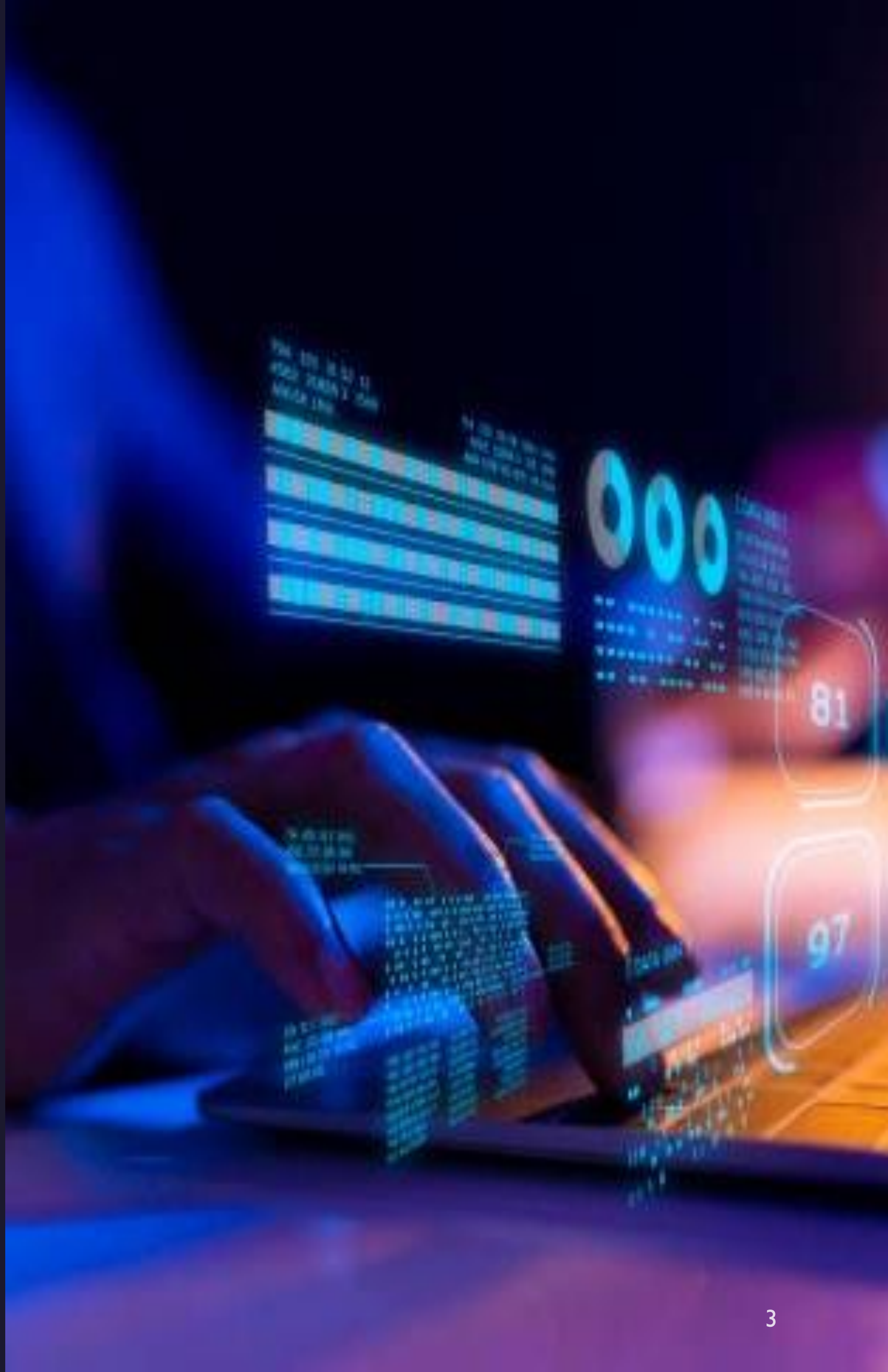
Team Members

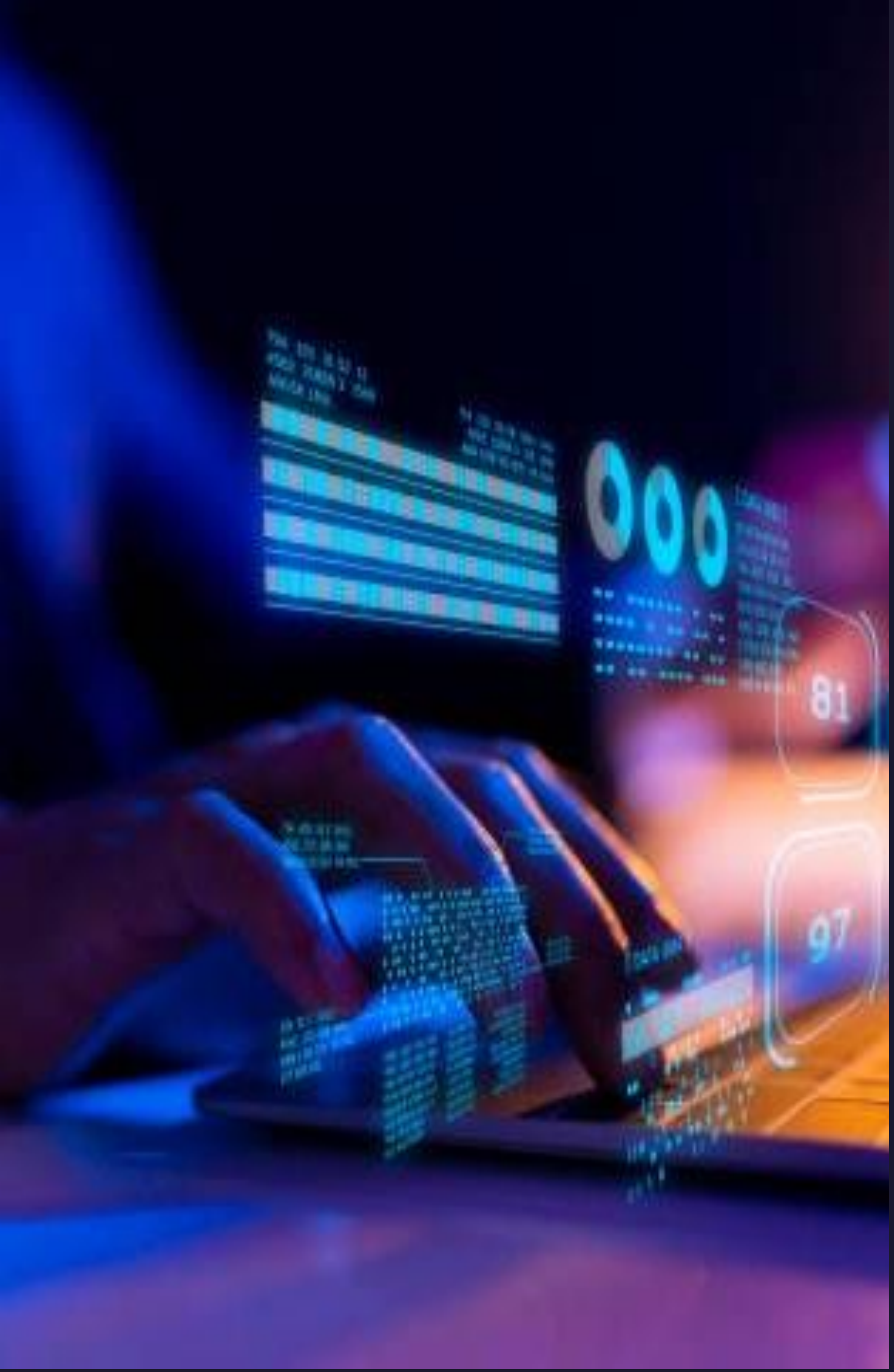
- IT21070594 – Zainab M.Z
- IT21004568 – Nuha M.N
- IT21013850 – Rashida M.S.F
- IT21006166 – Bishirhafi F.S.M.T



Problem Definition

- Surge in credit card fraud poses challenges for financial institutions.
- Growing credit card use amplifies fraud risks, causing financial loss and reputation damage.
- Using data analysis and machine learning models can be built to distinguish fraudulent and non-fraudulent transactions.
- The most accurate model can then be deployed to enhance security and profitability.
- Understanding fraud parameters enables effective prevention, reducing losses for all.





Project Aims

In credit card fraud detection, the primary goals include:

- Minimizing financial losses
- Optimizing resource allocation
- Early detection
- Preventing last-minute frauds
- Client communication
- Reducing false positives
- Enhancing customer experience
- Reducing fraud records.

Dataset Description

- The dataset was sourced from DataCamp.
- Compiled from various banks, and partially cleaned.
- Contains credit card transaction information, customer and merchant details, purchase amounts, and fraud indicators.
- Goal is to build a cautious predictive model to protect customers and prevent financial losses.
- To streamline processing, a 10% subset with balanced classes (33,960 records) was created from the original 339,607 records.

Technologies Used:

- Jupyter Notebook
- Github
- Python
- Streamlit

Preprocessing techniques

1 Exploration and Understanding

- Data Exploration, understanding column values, correlation, and unique categories.

2 Data Cleaning

- Removed irrelevant columns, dataset was checked for null values and duplicate rows.

3 Feature Engineering

- Converted DOB to Age and encoded categorical columns using binary encoding.

4 Formatting Data

- Renamed columns.

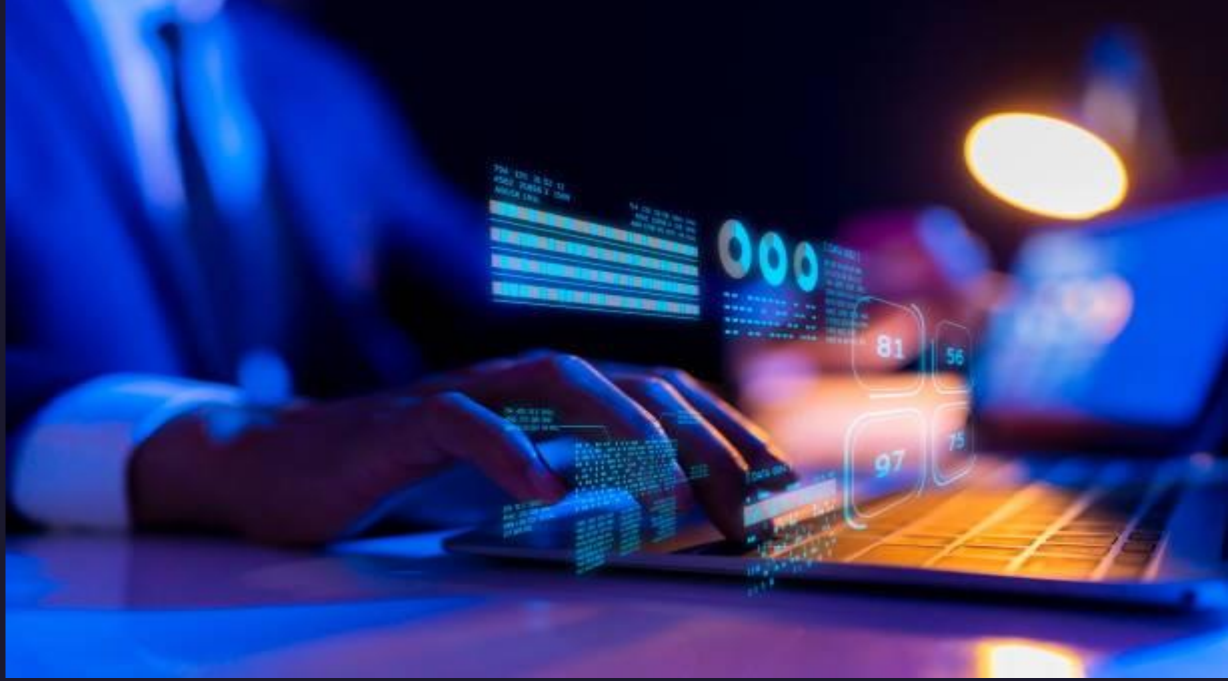
5 Feature selection

- Defined relevant feature columns.

6 Handling imbalance

- Using SMOTE minority class was oversampled to handle the imbalance.

Models Implemented



Implementation done using cross-validation techniques to split data into equal k-subsets. (k – number of subsets). One validation set and the rest used for training. For each fold, the evaluation metrics is calculated.

1. Random Forest

- Supervised learning algorithm used for classification and regression.
- Combines multiple decision trees to make predictions and selects the best outcome through voting.
- Larger number of trees in the forest leads to higher accuracy, making it a flexible and effective choice for various tasks.

2. Logistics Regression

- Ideal for binary classification tasks like identifying fraudulent transactions (0 for non-fraudulent, 1 for fraudulent).
- Efficient, easy to implement, and offers interpretable results
- Providing probabilities of events rather than just classifications.

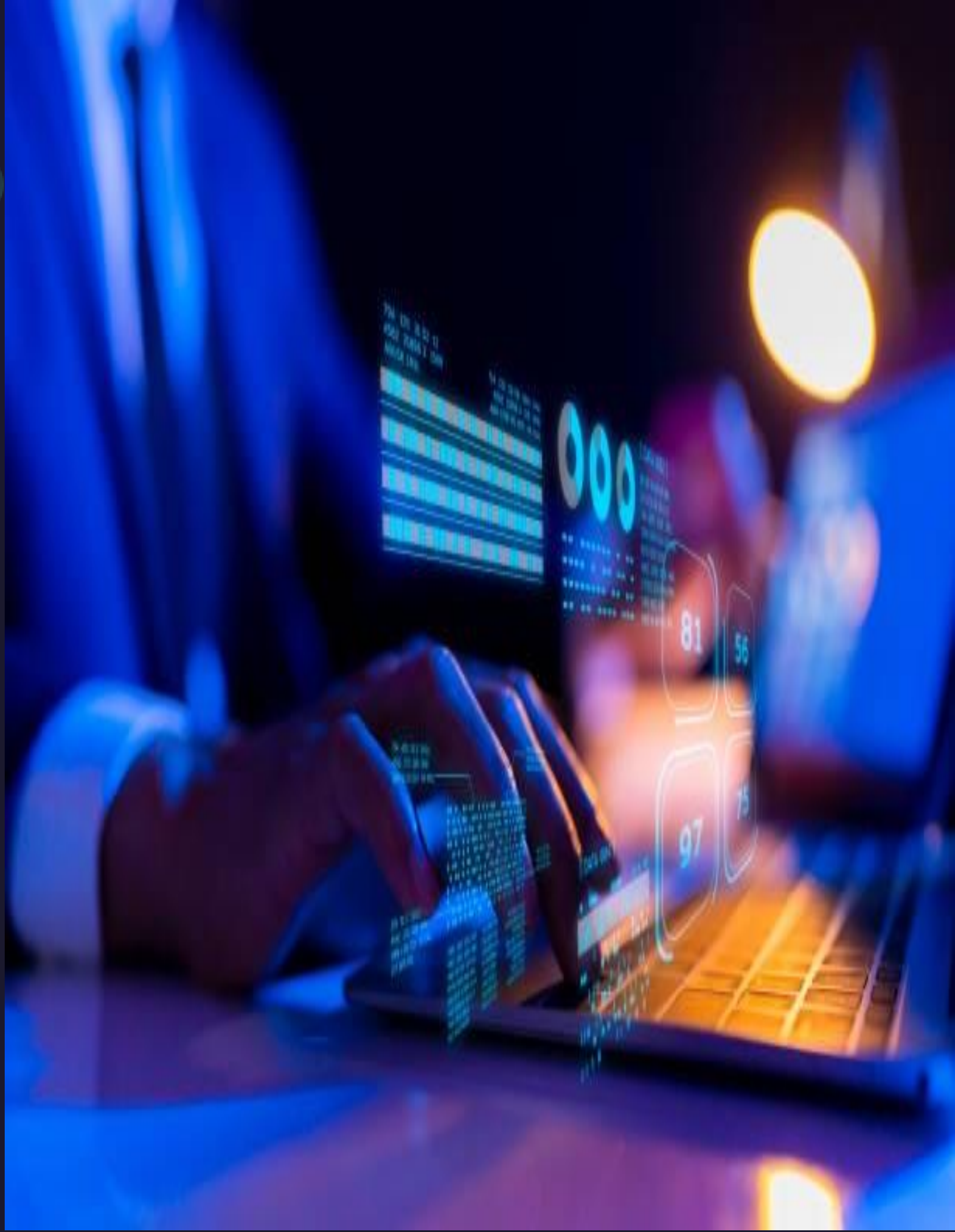
Models Implemented

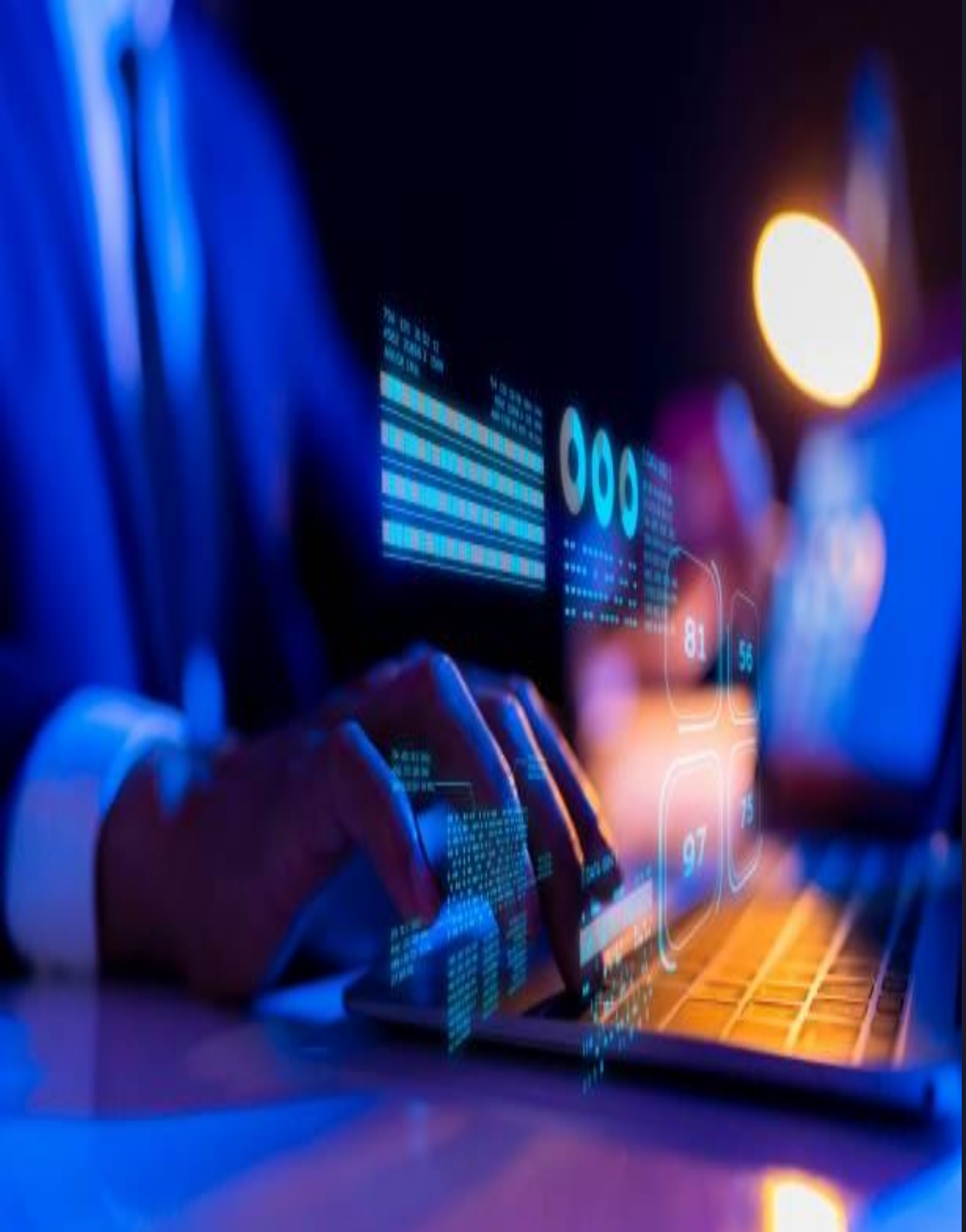
3. Naïve Bayes

- Fast, effective in classifying test data
- Ability to provide estimated probabilities for predictions.
- Assumes normal distribution within classes and feature independence, making it a practical choice for this context.

4. Support Vector Machine

- Versatile model used for binary classification tasks.
- Excels in complex, high-dimensional scenarios by finding the best hyperplane to separate classes, known as the "support vector."





Models Implemented

5. K-Nearest Neighbors (KNN)

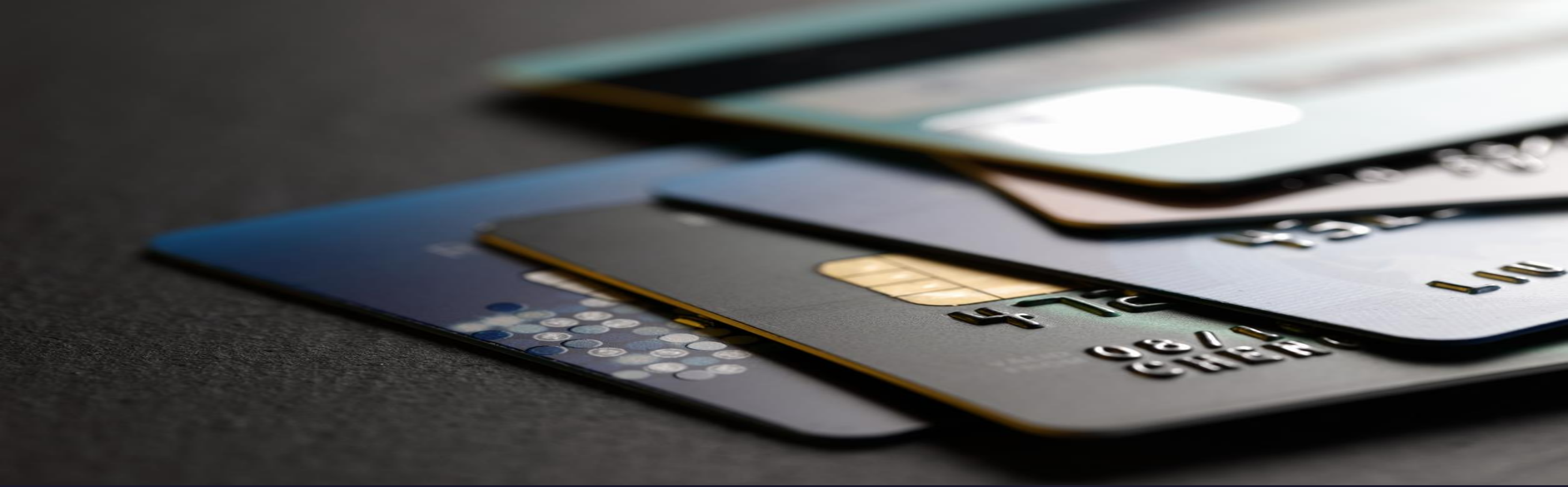
- For classification and regression tasks.
- Relying on data point similarity to make predictions.
- Uses a distance metric, like Euclidean distance, to determine proximity between data points.
- The 'K' parameter sets number of nearest neighbors to consider. In classification, it assigns labels based on majority votes.

Best Model

The confusion matrices were observed and analyzed while other measures of accuracy, such as the F1 score was also compared to choose the best model

Random Forest

- Has a higher precision, recall and accuracy combination in comparison to the rest of the models.
- Predicted the model outputs correctly whether a transaction was fraudulent or not.



Conclusion

Achieved its objectives by deploying a functional solution accessible via URL. Implemented and selected the best classification model based on accuracy, simplifying the user experience. However, some challenges included a lack of strongly correlated features, location-specific validity, model processing time, and dealing with dataset imbalance. To improve, we could gather global data and utilize more powerful computers for application development.

Thank You