

**MSc in Business Analytics**

**Module code – BEMM457**

**Topics in Business Analytics**

**Student Id: 740095163**

Credit card Fraud Detection



# Table of Contents

[List of Figures 3](#_bookmark0)

1. Introduction 4
2. Dataset Description………………………………………………………………………6
3. Exploratory Data Analysis (EDA)………………………………………………………7

3.1 Class Distribution Analysis………………………………………………………....7

3.2 Correlation Heatmap……………………………………………………………….8

3.3 Transaction Amount Analysis……………………………………………………....8

3.4 Outliers…………………………………………………………………………….....9

1. Models for Predictive Analysis…………………….………………………………......10

4.1 Random Forest Classifier………………………………………………………….10

4.2 Gradient Boosting (XG Boost) Classifier…………………………………………11

1. Comparing Model Performance………….…………………….…………………….12
2. Data Ethics in Fraud Detection………………………………………………………13
3. Recommendations…………………………………..………………………………...14
4. Conclusion…………………………………………………………………………….15

References………………………………………………………………………………………..16

# List of Figures

Figure 1: Image representing a credit card fraud detection ……………………………...5

Figure 2: Class Distribution analysis………………. …………………………………....7

Figure 3: Correlation Heatmap……………………………...…………………………..8

Figure 4: Transaction amount analysis……………….…………………………………..9

Figure 5: Outliers in Features ………………………………..…………………………9

Figure 6:The AUC-ROC Curve for the Random Forest Classifier……………………..10

Figure 7: The AUC-ROC Curve for the Gradient Boosting Classifier…………………11

Figure 8: Comparison of Random Forest and XGBoost Models .……………...……..12

Figure 9: Data Ethics in FraudDetection ………………………………………….…13

# Introduction

Credit card fraud has evolved from a regional issue to a worldwide catastrophe, posing a significant risk to the assets of financial institutions and individuals alike. Certain highly sophisticated tactics have been developed in an effort to scam consumers of digital payments and electronic commerce in particular, which rapidly outpace more traditional methods of detection. In 2023, fraud in the financial industry cost billions of dollars and put businesses and customers in danger . The increasing difficulty highlights the pressing need for stronger and more flexible solutions that can safeguard transactions and contribute to the restoration of confidence in digital financial systems. Beyond these direct financial losses, fraudulent credit card transactions reduce customer confidence, disrupt financial plans of individuals, and even the corporate reputation of organizations. It is a cat-and-mouse game, with fraudsters coming up with new means to exploit vulnerabilities continuously; it becomes increasingly important for a financial institution to step up security without compromising user convenience. The rule-based systems of old have, in fact, proven rather inefficient for fraud patterns that are dynamic and ever-evolving. This calls for more sophisticated techniques, especially machine learning, which can analyze a host of data, identify complex patterns, and predict fraudulent behavior in real time. It offers a whole new approach to fraud detection: speed, scalability, and accuracy all combined.

*Figure 1: Image representing credit card fraud detection.*

**Problem Statement:**

Because fraudulent transactions are so few compared to authorized ones, detecting credit card fraud is a crucial but difficult undertaking. Building predictive models that reliably detect fraudulent transactions without producing an excessive number of false positives is challenging due to these infrequent occurrences. In addition to putting a load on customer support teams, a high number of false positives might drive away real customers, resulting in discontent and possibly lost revenue.

The aim of this study is to develop a robust predictive framework that can assist in determining if a particular transaction is fraudulent or not by utilizing data from previous transactions. The dynamic and ever-evolving nature of fraud trends, the necessity for real-time choices, and data imbalance are some of the actual issues with fraud detection that the model addresses. By employing advanced machine learning techniques, the study aspires to enhance detection accuracy while minimizing false alarms, ultimately improving operational efficiency and customer experience.

**Significance of Fraud Detection:** Credit card fraud has its impacts on customers and financial institutions, too, apart from the monetary loss. More frequent fraud incidents result in damage to an organization's brand, increased operational costs, and reduced consumer trust. Consumers go through emotional stress, disruptions in their financial planning, and reduced trust in digital payment systems.

Efficient fraud detection is actually about the stability and integrity of the financial ecosystem. It also helps to instill confidence in the digital economy while protecting businesses and customers from financial loss. The most important challenge would then be to develop a predictive model that balances the seamless experience of users with the accuracy of fraud detection. This project aims to enhance fraud detection by using machine learning algorithms together with data-driven insights. The results will aid in the creation of safe, reliable, and effective systems that will enable Institutions and customers to conduct digital transactions with assurance.

1. **Dataset Description**

The dataset comprises 568,630 credit card transaction records, each labeled as either legitimate (Class 0) or fraudulent (Class 1). The data includes 28 anonymized features (V1–V28) derived using Principal Component Analysis (PCA) to protect sensitive information, alongside two additional features: the transaction amount and a unique identifier. This anonymization ensures privacy while retaining predictive value. The dataset reflects real-world conditions, where fraudulent transactions are rare, presenting a significant challenge for classification models due to the highly imbalanced class distribution.

Fraudulent behaviors evolve rapidly, necessitating adaptable and robust predictive models. The large volume of transactions and the high dimensionality of the data provide a substantial opportunity for applying advanced machine learning techniques to improve detection. Sourced from the Kaggle repository, this dataset has become a benchmark for fraud detection research, offering insights into combating fraud while maintaining data security and operational efficiency.

The dataset used for this project was obtained from Kaggle, specifically the **Credit Card Fraud Detection Dataset 2023** repository by Nelgiriya (2023), available at <https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023>."

1. **Exploratory Data Analysis (EDA)**

EDA, or exploratory data analysis, has been widely used to gain insight into the structure of a dataset, spot trends, and identify important insights within it. In addition, it helps identify irregularities and forms the basis for selecting appropriate machine learning models.

**3.1 Class Distribution Analysis**

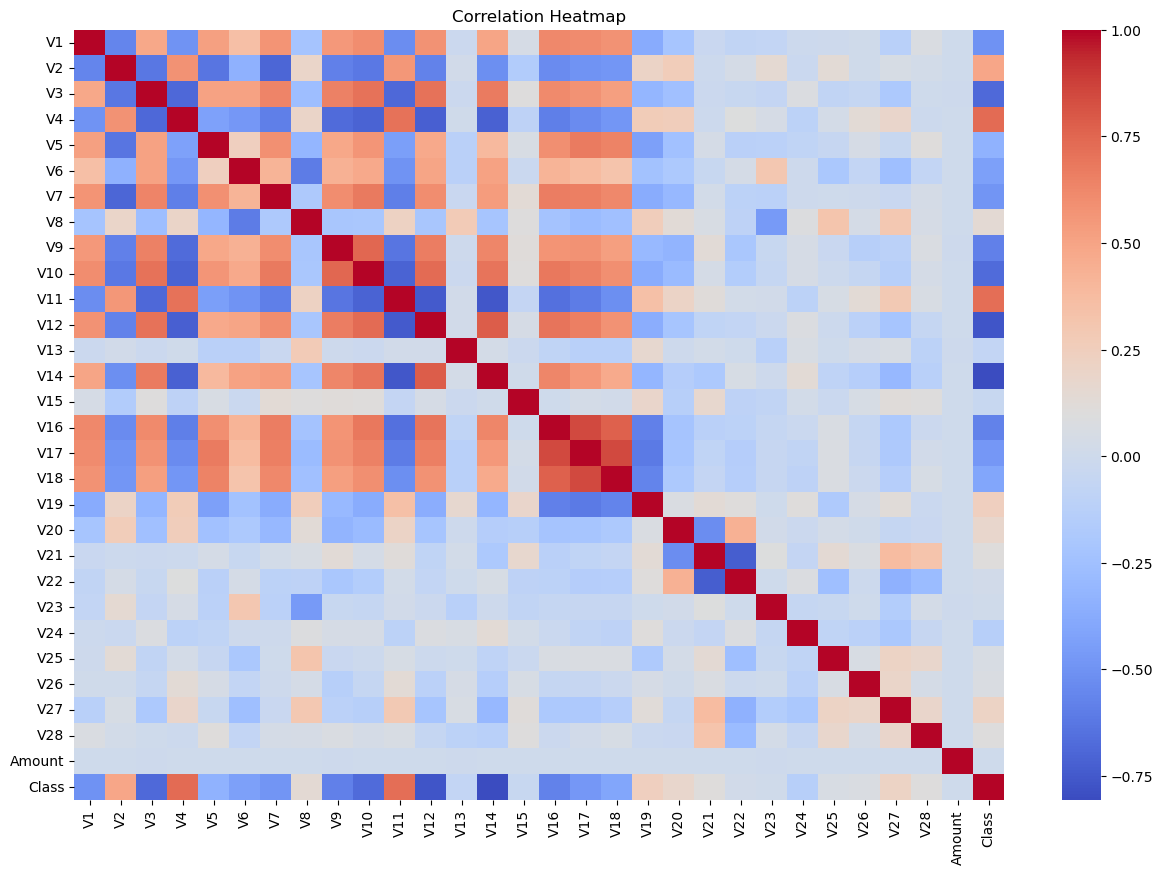
Class distribution analysis highlights the balance or imbalance in the dataset between legitimate (Class 0) and fraudulent transactions (Class 1). The bar chart shows a balanced distribution of transactions, with 50% legitimate (Class 0) and 50% fraudulent (Class 1). This balanced dataset eliminates the need for rebalancing techniques, ensuring unbiased model training and fair evaluation using standard metrics like accuracy and F1-score. Such a balance provides an ideal foundation for building robust fraud detection models.



*Figure 2: Class Distribution*

**3.2 Correlation Heatmap**

The relationships between the dataset's features are graphically represented by a correlation heatmap. With values ranging from -1 (strong negative correlation) to 1 (strong positive correlation), each cell in the heatmap represents the direction and intensity of the connection between two variables. High correlation features could be a sign of redundancy, which can be fixed by dimensionality reduction or feature selection. The correlation heatmap shows minimal correlations between features due to PCA transformation (V1–V28), ensuring independent contributions to models. The Amount feature also has weak correlations, indicating its independence. This structure minimizes multicollinearity, making the dataset ideal for predictive modelling**.**

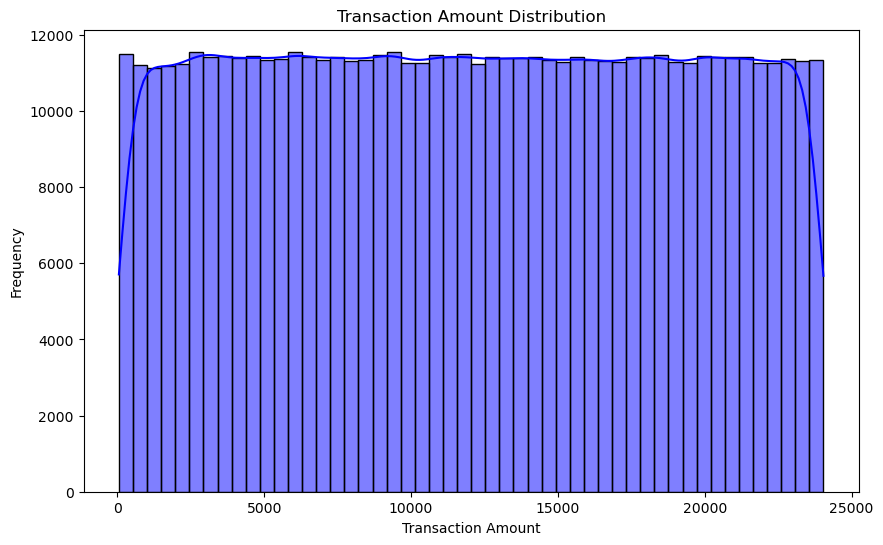


*Figure 3: Correlation Heatmap*

* 1. **Transaction Amount Analysis**

Transaction amount analysis explores the distribution of monetary values in the dataset, helping to identify trends or anomalies. Fraudulent transactions often exhibit distinct patterns in transaction amounts compared to legitimate ones. Analyzing these patterns provides valuable insights into the behavior of fraudulent activities**.**

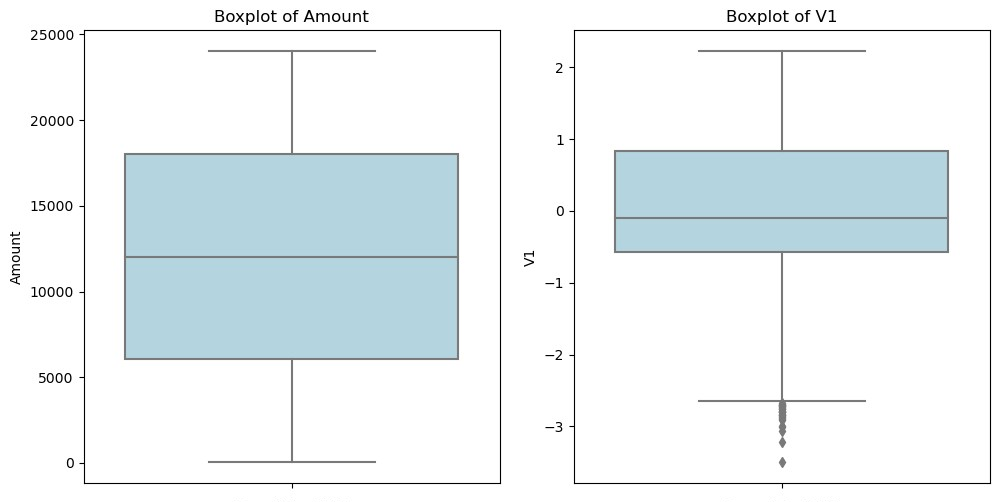
The Amount feature shows a right-skewed distribution, with most transactions being of low value. Fraudulent transactions occur across all ranges, indicating no direct correlation between transaction value and fraud likelihood. This suggests Amount can contribute to fraud detection but requires further analysis alongside other features.



*Figure 4: Transaction Amount Analysis*

**3.4 Outliers**

The boxplots for Amount and V1 reveal the presence of outliers, especially in the Amount feature due to high-value transactions. These outliers are typical in financial datasets and should be carefully handled to avoid skewing model predictions. The feature V1 shows fewer outliers, indicating more stable behaviour.



*Figure 5: Outliers in Features*

1. **Models for Predictive Analysis**

# By using past data to create precise classifications, predictive models are essential for identifying fraudulent transactions. In this study, robust models that can differentiate between fraudulent transactions (Class 1) and authentic ones (Class 0) are constructed using machine learning algorithms. Because of their complementing strengths, the following models were selected:

# • Random Forest Classifier: renowned for its resilience, interpretability, and capacity to manage intricate data interactions.

# • Gradient Boosting (XGBoost): An advanced boosting method that works well with unbalanced data and is very successful for tabular datasets.

# These models were trained on the dataset using a systematic process of data preprocessing, feature scaling, and hyperparameter tuning. Their performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure a comprehensive assessment of their effectiveness in fraud detection.

# 4.1 Random Forest Classifier

# Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. In this study, it achieved an accuracy of 95% and an AUC-ROC score of 0.99, showcasing its strong ability to distinguish between fraudulent and legitimate transactions. The model delivered 99% precision for fraudulent transactions, with minimal false positives (327), but its recall of 91% indicates some fraud cases (5,054) were missed. For legitimate transactions, it achieved 92% precision and 99% recall, ensuring most genuine transactions were correctly identified. The balanced F1-scores of 95% validate the model's robustness, making it effective for fraud detection.

*Figure 6: The AUC-ROC Curve for the Random Forest Classifier*

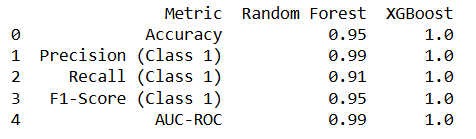
# 4.2 Gradient Boosting (XGBoost) Classsifier

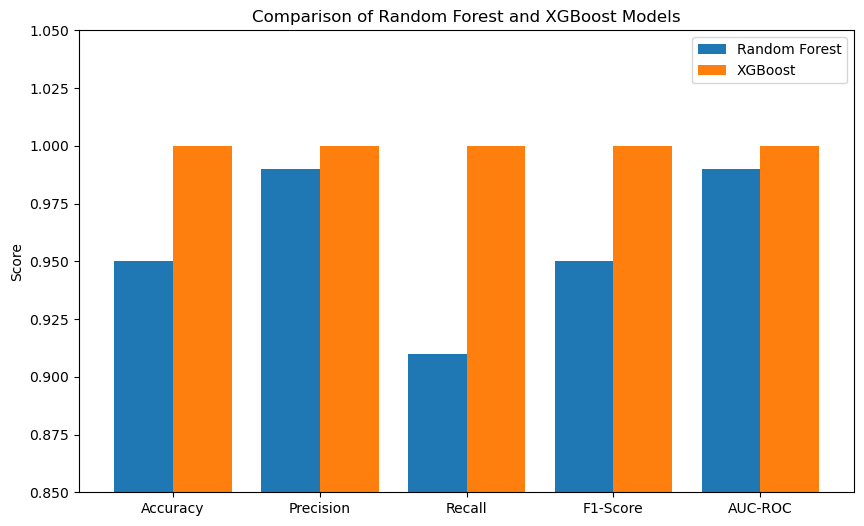
# A potent machine learning technique called XGBoost (Extreme Gradient Boosting) creates decision trees one after the other while improving accuracy by fixing past mistakes. XGBoost performed exceptionally well in this investigation, achieving 100% accuracy and a flawless AUC-ROC score of 1.00. With only 33 false positives and 0 false negatives, the model showed perfect precision, recall, and F1-scores for both authentic and fraudulent transactions. These findings are further supported by the confusion matrix, which demonstrates XGBoost's remarkable class distinction capabilities and makes it a very trustworthy model for fraud detection.

# 

*Figure 7: The AUC-ROC Curve for the Gradient Boosting Classifier*

1. **Comparing Model Performances**

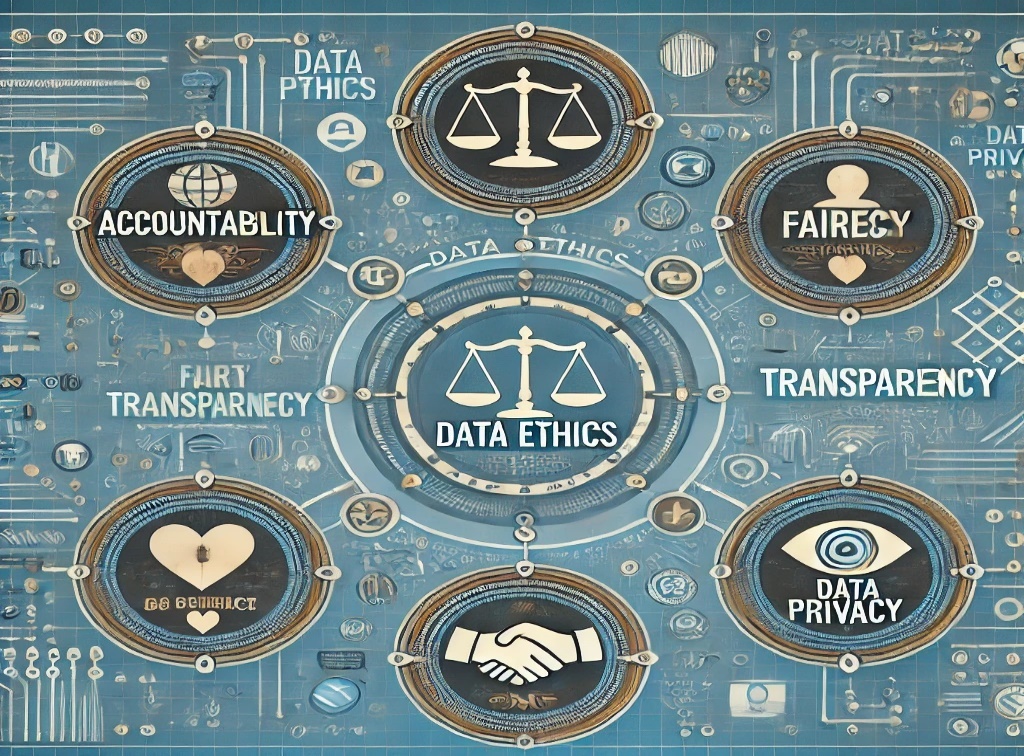
The models in Random Forest and XGBoost did exceptionally well with fraud detection, but again, XGBoost beat the performance of Random Forest across all metrics. XGBoost was faultless with correct categorization, accuracy, precision, recall at 100%, and an AUC-ROC of 1.00.Random Forest, on the other hand, missed 5,054 fraud cases while having a little lower recall for the fraudulent class (91%), despite achieving a high accuracy of 95% and an AUC-ROC score of 0.99. While Random Forest is effective, XGBoost's perfect recall and minimal false positives make it the superior choice for this dataset.

*Figure 8: Comparison of Random Forest and XGBoost Models*

1. **Data Ethics in Fraud Detection**

**Accountability ensures organizations take responsibility for the outcomes of fraud detection models. False positives can inconvenience customers, while false negatives might allow fraud to occur. To address this, institutions must implement human oversight for flagged transactions, maintain audit trails, and establish protocols for correcting errors. Regular model evaluations are essential to ensure accuracy and reliability over time.**

**Fairness is vital to avoid discriminatory outcomes in fraud detection. To avoid sustaining injustices, models must be trained on a variety of objective datasets. Unintended biases can be reduced with the use of fairness-aware algorithms and ongoing monitoring, guaranteeing that all client groups receive fair treatment. This not only fosters trust but also aligns with ethical and regulatory standards.**

**Transparency and data privacy are critical for building trust in fraud detection systems. To ensure that stakeholders comprehend the process, models should clearly explain decisions made using techniques such as SHAP or LIME. Organizations must simultaneously secure sensitive consumer data by utilizing anonymization techniques, adhering to laws like GDPR, and putting strong security measures in place to stop data breaches.**

*Figure 9: Data Ethics in Fraud Detection*

1. **Recommendations**

This project offered insightful information about using machine learning models to detect fraud. I discovered how crucial it is to manage unbalanced datasets and assess models using measures such as AUC-ROC, F1-score, precision, and recall. The application of models like Random Forest and XGBoost demonstrated both their advantages and disadvantages, with XGBoost standing out as the best model because of its remarkable performance and capacity to manage intricate patterns. The difficulties of implementing fraud detection systems were brought to light by issues including managing training time for big datasets, optimizing hyperparameters, and guaranteeing ethical considerations like openness and fairness. I suggest using XGBoost in real-time fraud detection systems, optimizing it for low latency, and periodically retraining it to adjust to changing fraud patterns in order to improve future implementations. Incorporating explainability tools, such as SHAP, can improve stakeholder trust, while monitoring for biases ensures fairness and reliability in decision-making. These strategies can further enhance the effectiveness and scalability of fraud detection models.

1. **Conclusion**

This project offered insightful information about using machine learning models to detect fraud. I came to understand the importance of handling imbalanced datasets and evaluating any model on AUC-ROC, F1-score, precision, and recall. How to use models such as Random Forest and XGBoost, the advantages and disadvantages of each, and how XGBoost is the best model in this dataset, as it has a very impressive performance and can handle complex patterns. The difficulties of implementing fraud detection systems were brought to light by issues including managing training time for big datasets, optimizing hyperparameters, and guaranteeing ethical considerations like openness and fairness. I suggest using XGBoost in real-time fraud detection systems, optimizing it for low latency, and periodically retraining it to adjust to changing fraud patterns in order to improve future implementations.

Machine learning models were effectively used in this study to identify fraudulent transactions, highlighting their vital role in combating financial fraud. After Random Forest and XGBoost were compared, XGBoost was shown to be the better model, with an AUC-ROC score of 1.00 and flawless accuracy of 100%. These outcomes demonstrate its remarkable capacity to minimize errors while differentiating between authentic and fraudulent transactions. In order to create reliable fraud detection algorithms, the research also underlined the significance of ethical factors including accountability, fairness, openness, and data privacy.

The difficulties of managing unbalanced datasets, maximizing model performance, and guaranteeing moral implementation in practical applications are some of the important realizations made. The recommendations provided, such as periodic retraining, real-time implementation, and the use of explainability tools, ensure the system remains reliable, adaptable, and fair. By leveraging advanced models like XGBoost, organizations can effectively combat fraud, safeguard financial systems.

1. **References**

1. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>

2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>

3. Scikit-learn developers. (n.d.). Random Forest documentation. *Scikit-learn User Guide*. Retrieved from <https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. Retrieved from <https://arxiv.org/abs/1705.07874>

5. GDPR.eu. (n.d.). General Data Protection Regulation (GDPR) compliance guidelines. Retrieved from <https://gdpr.eu>

6. Python Software Foundation. (n.d.). Matplotlib: Visualization with Python. Retrieved from <https://matplotlib.org>

7. Kaggle. (n.d.). Credit card fraud detection dataset. *Kaggle Datasets*. Retrieved from <https://www.kaggle.com/mlg-ulb/creditcardfraud>

8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. <https://doi.org/10.1145/2939672.2939778>

9.Fraud Detection Dataset <https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023>