Report (PDF) Detailing the Following:

A. Description of Design Choices:

1. **Model Choice**:
    - The **RandomForestClassifier** was chosen for its robustness, scalability, and ability to handle high-dimensional datasets. It can model complex relationships and handle both categorical and numerical features effectively. Additionally, RandomForest is relatively resilient to overfitting compared to other algorithms and has some capacity to handle imbalanced data, although further methods are required for more extreme imbalance.
2. **Preprocessing**:
    - **Data Cleaning**: The data cleaning process involved standardization of features using `StandardScaler` to ensure that features are on a similar scale, which can improve the performance of many machine learning models.
    - **Imbalanced Data**: As the dataset is highly imbalanced (fraudulent transactions are much rarer than legitimate ones), this part would benefit from additional techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** or adjusting **class weights** within the RandomForest model to address the imbalance. This could be a part of future improvements if not yet completed.
    - **Feature Engineering**: Feature engineering was minimally required for this dataset since many features were pre-transformed using **PCA**. However, further feature extraction or transformation may be considered for model improvement.
3. **Performance Metrics**:
    - The evaluation was done using **accuracy**, **precision**, **recall**, and **F1-score**. However, due to the imbalanced nature of the data, **precision** and **recall** are more informative than accuracy. The **confusion matrix** was used to visually assess the model's performance on predicting fraudulent and non-fraudulent transactions.

B. Performance Evaluation of the Model:

1. **Model Performance**:
    - The RandomForestClassifier was evaluated based on the test data using various performance metrics:
        - **Accuracy**: (Model's overall accuracy score)
        - **Precision**: (Model's ability to correctly identify fraud from all identified fraud cases)
        - **Recall**: (Model's ability to detect fraud from all fraud cases)
        - **F1-Score**: (The harmonic mean of precision and recall, useful for evaluating imbalanced datasets)

        - **Confusion Matrix**: A matrix to display the performance in terms of True Positives, False Positives, True Negatives, and False Negatives.
2. **Graphs**:
    - The performance was also visualized using:
        - **Confusion Matrix**: To understand the true and false positives/negatives.
        - **ROC-AUC Curve**: To evaluate the trade-offs between true positive rate and false positive rate at different classification thresholds.
3. **Overfitting**:
    - There was an assessment of how well the model generalizes to unseen data. Comparing training and testing results suggested whether there was **overfitting** (i.e., the model performs well on training data but poorly on unseen data).

C. Discussion of Future Work:

1. **Improving Data Imbalance**:
    - While class weights or SMOTE can address data imbalance to an extent, other techniques such as **ensemble methods** or **anomaly detection algorithms** may offer better solutions for highly imbalanced datasets like this one.
2. **Model Tuning**:
    - Future improvements could involve fine-tuning the model's hyperparameters using techniques such as **Grid Search** or **Randomized Search Cross-Validation** to further optimize the RandomForest model.
3. **Other Algorithms**:
    - Experimentation with other machine learning models could provide a better understanding of what works best for this problem. **XGBoost**, **LightGBM**, or even a **Logistic Regression** baseline could be compared to assess their effectiveness in handling imbalanced data and overall performance.
4. **Feature Engineering**:
    - Future iterations of this project could explore more **advanced feature engineering** techniques to extract more useful features from the existing data or create new features that improve the model's prediction power.
