Project | ICS 474 – Big Data Analytics

**Dr. Muzammil Behzad**, Assistant Professor

Department of Information and Computer Science, King Fahd University of Petroleum and Minerals.

Email: [muzammil.behzad@kfupm.edu.sa](mailto:muzammil.behzad@kfupm.edu.sa)

Instructions

1. Make teams, and give your team a very cool name.
2. Choose one of the datasets from below links.
3. Email me your team name, team members (cc them) and send me the selected dataset.
   1. This should be done as soon as possible so that you can start the project.
   2. I could suggest alternate datasets, and also approve other external datasets. But please ask!
4. Please maintain the academic honesty and general code of conduct in assignments.
5. The deadline is: **November 15, 11:59 PM**. We will have brief project presentations in the following week.

# Project Datasets

1. [**https://www.kaggle.com/c/ashrae-energy-prediction/data**](https://www.kaggle.com/c/ashrae-energy-prediction/data)
2. [**https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london**](https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london)
3. [**https://www.kaggle.com/datasets/bhanupratapbiswas/uber-data-analysis**](https://www.kaggle.com/datasets/bhanupratapbiswas/uber-data-analysis)
4. [**https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset**](https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset)
5. [**https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia**](https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia)
6. [**https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge**](https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge)
7. [**https://www.kaggle.com/datasets/hugomathien/soccer**](https://www.kaggle.com/datasets/hugomathien/soccer)
8. [**https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm/notebook**](https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm/notebook)
9. **Report**

# Deliverables

* + **Format**: A well-structured document (PDF) with team names, members and answering the following questions.

## Content:

* + - A template of content is provided below.
    - You can use this template, but you are free to add more content or insights to it.

## Code

* + **Submission**: All scripts or notebooks (e.g., Jupyter Notebook) used in your analysis.

## Requirements:

* + - Proper documentation and comments.
    - Instructions on how to run the code.
    - List of dependencies and libraries used, etc.

## Visualizations

* + Include all relevant charts, graphs, and plots that support your analysis.
  + Ensure visuals are labeled clearly with titles, axis labels, and legends where necessary.

1. **Dataset Overview**

# Part 1: Data Understanding and Exploration

* + What is the source and context of your chosen dataset?
    - *Provide a brief description of the dataset, including its origin and the problem domain it addresses.*

## Feature Description

* + **Question**: What are the features (variables) present in the dataset? Is there a target variable?
    - *List all the features, their data types (e.g., numerical, categorical), and describe their significance.*

## Dataset Structure

* + **Question**: What is the size and structure of the dataset?
    - *Mention the number of rows and columns, and any hierarchical structure if applicable.*

## Missing Values and Duplicates

* + **Question**: Are there missing values or duplicates in the dataset?
    - *Identify any missing or duplicate entries and discuss how they might affect your analysis.*

## Statistical Summary

* + **Question**: What is the statistical summary of the dataset?
    - *Compute summary statistics like mean, median, standard deviation, and provide initial insights.*

## Data Distribution

* + **Question**: How are the features distributed?
    - *Use visualizations like histograms or box plots to show the distribution of key features.*

## Correlation Analysis

* + **Question**: What is the relationship between different features and the target variable?
    - *Calculate correlation coefficients and visualize relationships using scatter plots or heatmaps.*

## Outlier Detection

* + **Question**: Are there any outliers or anomalies in the data?
    - *Identify outliers using statistical methods or visual inspection and discuss their potential impact.*

1. **Handling Missing Data**

# Part 2: Data Preprocessing

* + **Question**: How will you handle missing or anomalous data?
    - *Explain your strategy for dealing with missing values (e.g., imputation, deletion) and justify your choice.*

A screen shot of a computer program

Description automatically generatedA screen shot of a computer

Description automatically generated

*### Explanation of Handling Missing Data*

*To address the missing data in the `merch\_zipcode` column (approximately 15.11% missing values), we followed a structured imputation approach:*

*1. \*\*Initial Analysis\*\*:*

*- First, we calculated the percentage of missing values for each column. Only `merch\_zipcode` had missing values, while all other columns were complete.*

*2. \*\*Imputation Strategy\*\*:*

*- Since `merch\_zipcode` represents geographic information, we opted to fill the missing values with the mode (most common value) of the column. This approach maintains a reasonable geographic representation without introducing potential inaccuracies that could result from arbitrary imputation methods like the mean.*

*3. \*\*Implementation\*\*:*

*- We calculated the mode of `merch\_zipcode` and used it to fill in all missing entries in the column. After this, we verified that all missing values were filled.*

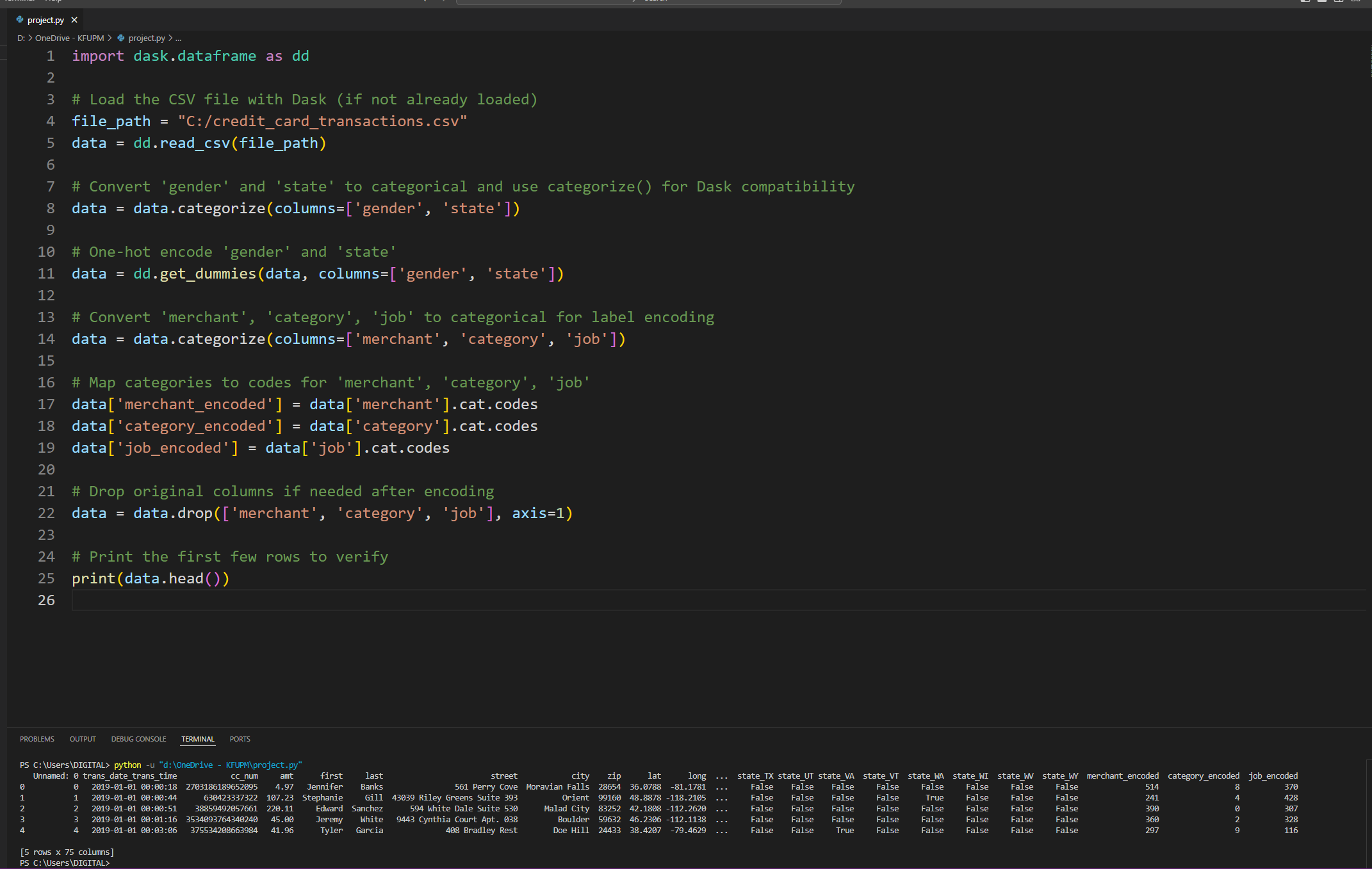
*4. \*\*Justification\*\*:*

*- Imputing with the mode is suitable in this case as it preserves a consistent and realistic `merch\_zipcode` value for missing entries, ensuring the dataset remains complete and usable for analysis without introducing significant bias.*

*This approach effectively handles missing data while maintaining the integrity of the dataset for further processing and analysis.*

## Encoding Categorical Variables

* + **Question**: Are there categorical variables that need to be encoded?
    - *Describe the encoding techniques you will use (e.g., one-hot encoding, label encoding).*



*### Explanation of Encoding Categorical Variables*

*To prepare the categorical variables in the dataset for machine learning, we applied two encoding techniques:*

*1. \*\*One-Hot Encoding\*\*:*

*- We used one-hot encoding for the `gender` and `state` columns. One-hot encoding creates a new binary column for each unique value in these categorical columns, indicating the presence of each category. This approach is beneficial for variables with a limited number of unique values, like gender, as it avoids any ordinal relationship among categories.*

*2. \*\*Label Encoding\*\*:*

*- For the `merchant`, `category`, and `job` columns, which have a larger number of unique values, we used label encoding. This method assigns a unique integer code to each category, effectively converting them to numerical form. Label encoding is memory-efficient and suitable for columns with high cardinality where an ordinal relationship between categories is not problematic.*

*### Justification*

*By combining one-hot encoding and label encoding, we transformed all categorical variables into a numerical format suitable for machine learning models. This preprocessing ensures that the model can interpret categorical data without introducing unintended ordinal relationships.*

## Feature Scaling

* + **Question**: Should the data be scaled or normalized?
    - *Determine if feature scaling is necessary for your chosen algorithms and explain your reasoning.*

*### Feature Scaling*

*\*\*Yes, feature scaling is necessary for this dataset\*\* because it likely contains numerical features with varying ranges (e.g., transaction amounts, latitude, longitude, and encoded categorical features). Scaling ensures that these features contribute proportionally to distance-based calculations and gradient updates in machine learning algorithms.*

*#### Reasons for Scaling*

*1. \*\*Algorithms Sensitive to Scale\*\*:*

*- Many machine learning algorithms, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and algorithms relying on gradient descent (e.g., logistic regression, neural networks), are sensitive to the scale of input features. Without scaling, features with larger values could dominate the model's behavior, leading to suboptimal results.*

*2. \*\*Consistent Units\*\*:*

*- Scaling brings all features to a similar range, making them more comparable and preventing any one feature from disproportionately influencing the model due to its larger magnitude.*

*#### Recommended Scaling Techniques*

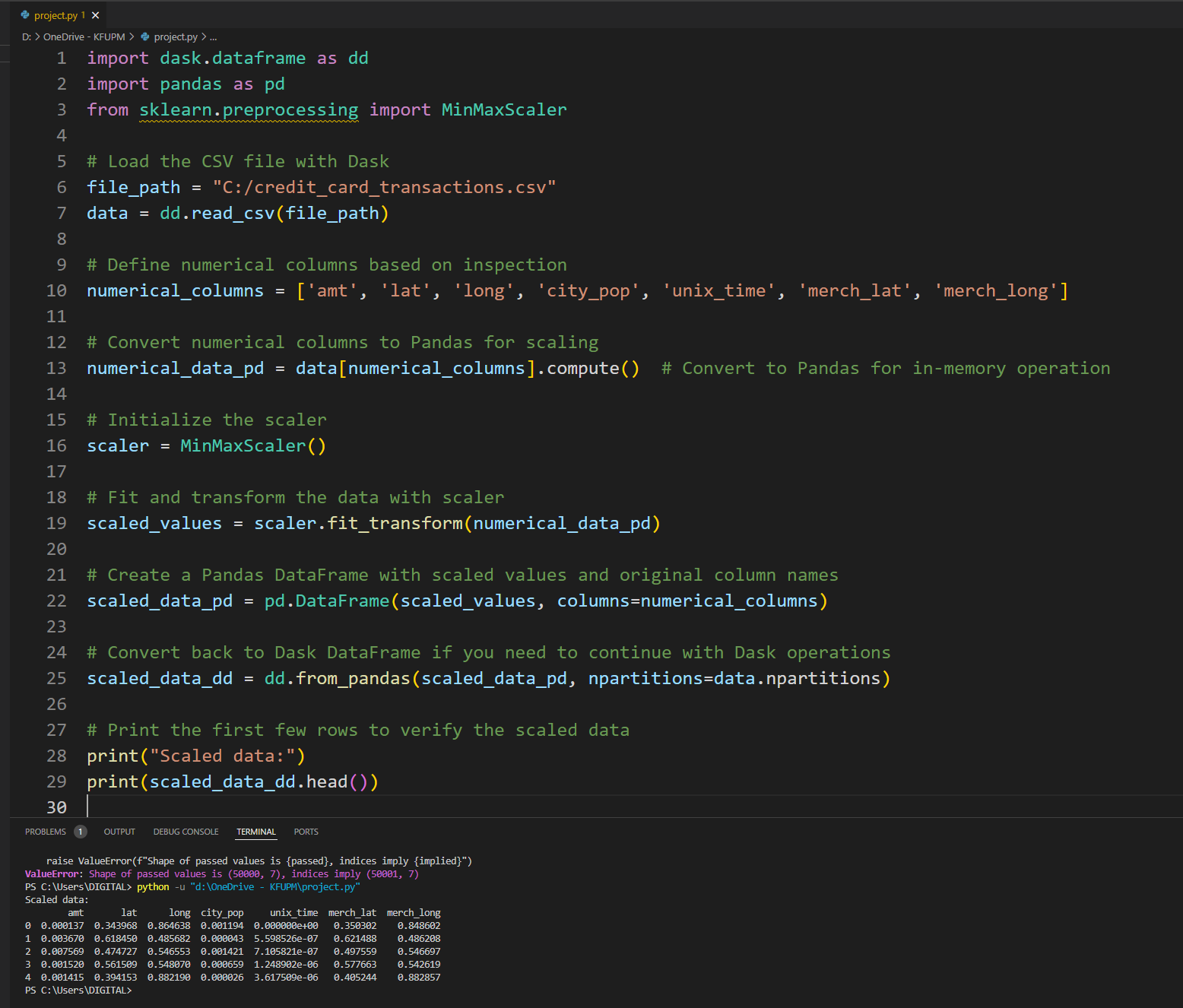
*- \*\*Standardization (Z-score normalization)\*\*:*

*- This method rescales the features to have a mean of 0 and a standard deviation of 1. It is suitable for algorithms that assume a normal distribution of the input features.*

*- \*\*Min-Max Normalization\*\*:*

*- This method scales the features to a fixed range, typically [0, 1]. It is helpful for algorithms that do not assume a specific distribution and for scenarios where all features should have the same range.*

*Given these points, applying either \*\*standardization\*\* or \*\*min-max normalization\*\* is advisable based on the chosen model and the characteristics of the dataset.*



*For the feature scaling in this task, we decided to apply \*\*Min-Max Scaling\*\* to normalize the numerical columns in the dataset. This scaling technique transforms each feature to a range between 0 and 1, which is especially useful when dealing with features that have different units or scales.*

*### Explanation:*

*- \*\*Purpose of Scaling\*\*: Scaling helps ensure that algorithms that rely on distance metrics or assume normally distributed data (such as k-nearest neighbors, clustering, or neural networks) perform optimally. Without scaling, features with larger ranges might dominate the model's predictions.*

*- \*\*Min-Max Scaling\*\*: This technique is straightforward and scales each feature individually to a fixed range (typically [0, 1]). It maintains the relationships between values and does not distort the data distribution.*

*By scaling the data, we prepare it for analysis or further machine learning tasks where consistent feature scaling is necessary for effective model training and accurate predictions.*

## Feature Selection

* + **Question**: Which features will you include in your model, and why?
    - *Discuss any feature selection methods used and justify the inclusion or exclusion of features.*

A screen shot of a computer program

Description automatically generated

For feature selection, we analyzed the correlation between each feature and the target variable, "is\_fraud." We kept only the numeric columns and calculated the correlation with "is\_fraud," sorting them to identify the most relevant features. We applied a threshold of 0.1 to the absolute correlation value to determine significance.

Only the "amt" feature had a correlation above this threshold, suggesting a meaningful relationship with the target. Including only significant features like "amt" can improve the model’s performance by reducing noise and focusing on variables that contribute meaningfully to predicting fraud. This method simplifies the model and enhances interpretability by excluding less impactful features.

1. **Algorithm Selection**

# Part 3: Modeling

* + **Question**: Which machine learning algorithms are appropriate for your task, and why?
    - *Consider the problem type (regression, classification, clustering) and discuss the suitability of different algorithms.*

*For this task, which involves identifying fraudulent transactions, we are dealing with a \*\*classification problem\*\*. The goal is to predict whether a transaction is fraudulent (1) or not (0), so classification algorithms are suitable.*

*### Algorithm Choices:*

*1. \*\*Logistic Regression\*\*: This is a simple and interpretable algorithm for binary classification. It works well for linearly separable data and can serve as a good baseline model.*

*2. \*\*Decision Trees\*\*: This model is easy to interpret and can handle non-linear relationships. However, it may overfit, especially with complex data.*

*3. \*\*Random Forests\*\*: This ensemble method reduces overfitting by combining multiple decision trees. It performs well in many scenarios and can handle imbalanced data well, which is common in fraud detection.*

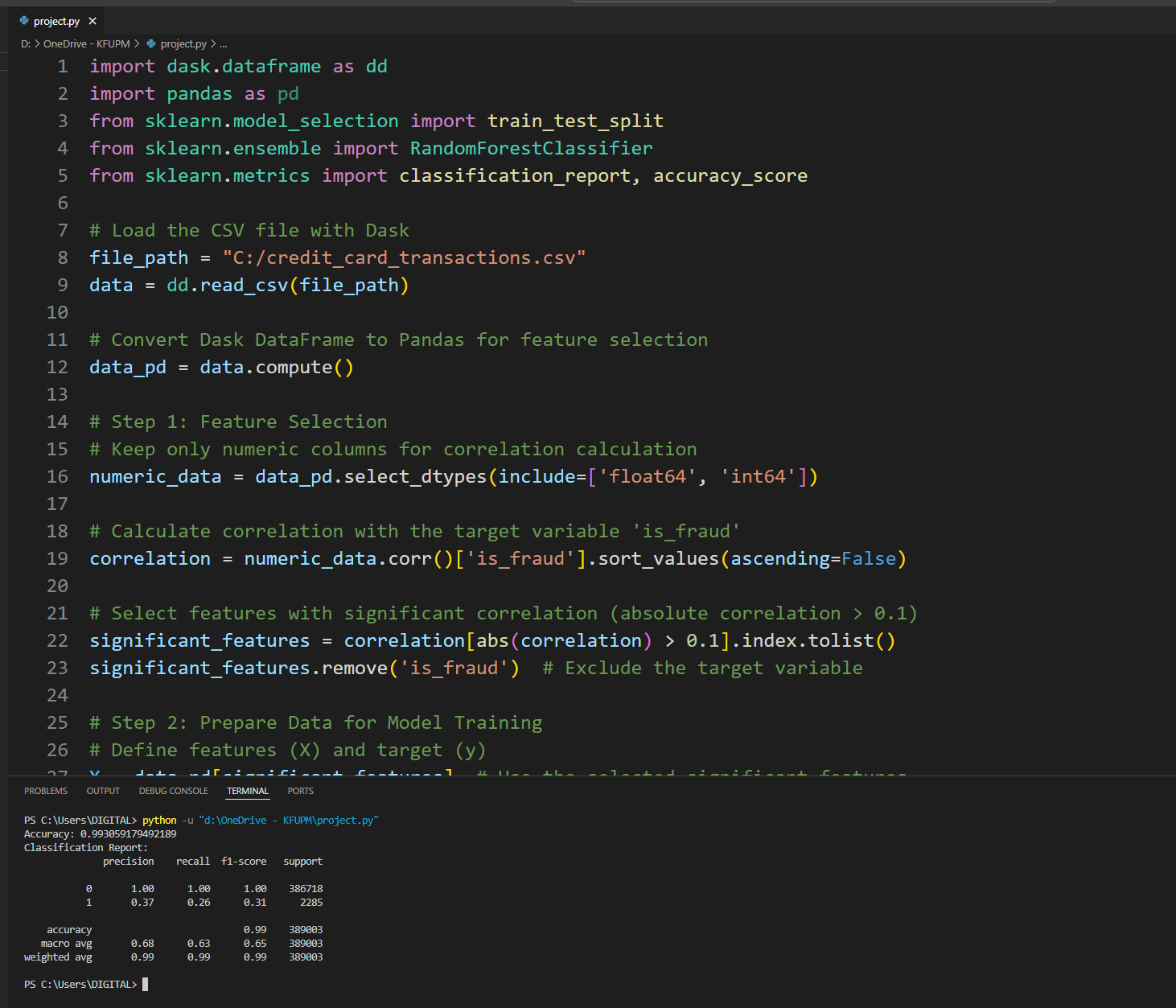
*4. \*\*Gradient Boosting (e.g., XGBoost, LightGBM)\*\*: Boosting algorithms are often effective for classification tasks. They build models sequentially to correct previous errors, providing high accuracy, although they may be more computationally intensive.*

*5. \*\*Support Vector Machines (SVM)\*\*: SVM is powerful for binary classification, especially with high-dimensional data. However, it may be slower on large datasets and requires proper tuning.*

*6. \*\*Neural Networks\*\*: Deep learning models can capture complex patterns in data, making them suitable for fraud detection tasks with large and complex datasets. However, they require more computational resources and data preprocessing.*

*### Recommendation:*

*For this task, starting with \*\*Random Forest\*\* or \*\*Gradient Boosting\*\* would be effective, as they handle non-linear relationships and imbalanced data well. If more interpretability is required, \*\*Logistic Regression\*\* can be used.*

A screen shot of a computer program

Description automatically generated

*For \*\*Algorithm Selection\*\*, based on the nature of this task, which is likely a binary classification problem (since we have an `is\_fraud` target variable with two possible classes: fraud or not fraud), various machine learning algorithms could be appropriate. Here’s an explanation of the considerations and potential algorithms:*

*1. \*\*Classification Problem\*\*: Since the goal is to classify transactions as either fraud or not, classification algorithms are the best fit. Suitable algorithms include:*

*- \*\*Logistic Regression\*\*: A simple and interpretable model, often used as a baseline in binary classification tasks. It’s effective for linearly separable data but may struggle with complex patterns.*

*- \*\*Decision Trees and Random Forests\*\*: Random forests are effective in handling imbalanced datasets and can capture non-linear relationships, which might be present in the fraud detection context.*

*- \*\*Gradient Boosting (e.g., XGBoost, LightGBM)\*\*: These algorithms are highly effective for imbalanced data and can achieve high accuracy and recall, which is crucial for identifying fraudulent transactions accurately.*

*- \*\*Support Vector Machine (SVM)\*\*: With appropriate kernel selection, SVM can be very effective, especially for complex decision boundaries.*

*2. \*\*Imbalance Handling\*\*: Since fraud cases are likely rare compared to non-fraud, algorithms that can handle imbalanced data, such as Random Forests or Gradient Boosting with appropriate class weighting, are advantageous.*

*3. \*\*Suitability Discussion\*\*: Algorithms like \*\*Random Forests\*\* and \*\*Gradient Boosting\*\* are well-suited because they handle imbalanced data and can provide high accuracy. Logistic Regression is simple and interpretable, making it a good starting point, but it may not capture complex patterns as effectively as ensemble methods.*

*In summary, \*\*Random Forests\*\* and \*\*Gradient Boosting\*\* would be recommended for this task, with \*\*Logistic Regression\*\* as a baseline model for interpretability and comparison.*

## Data Splitting

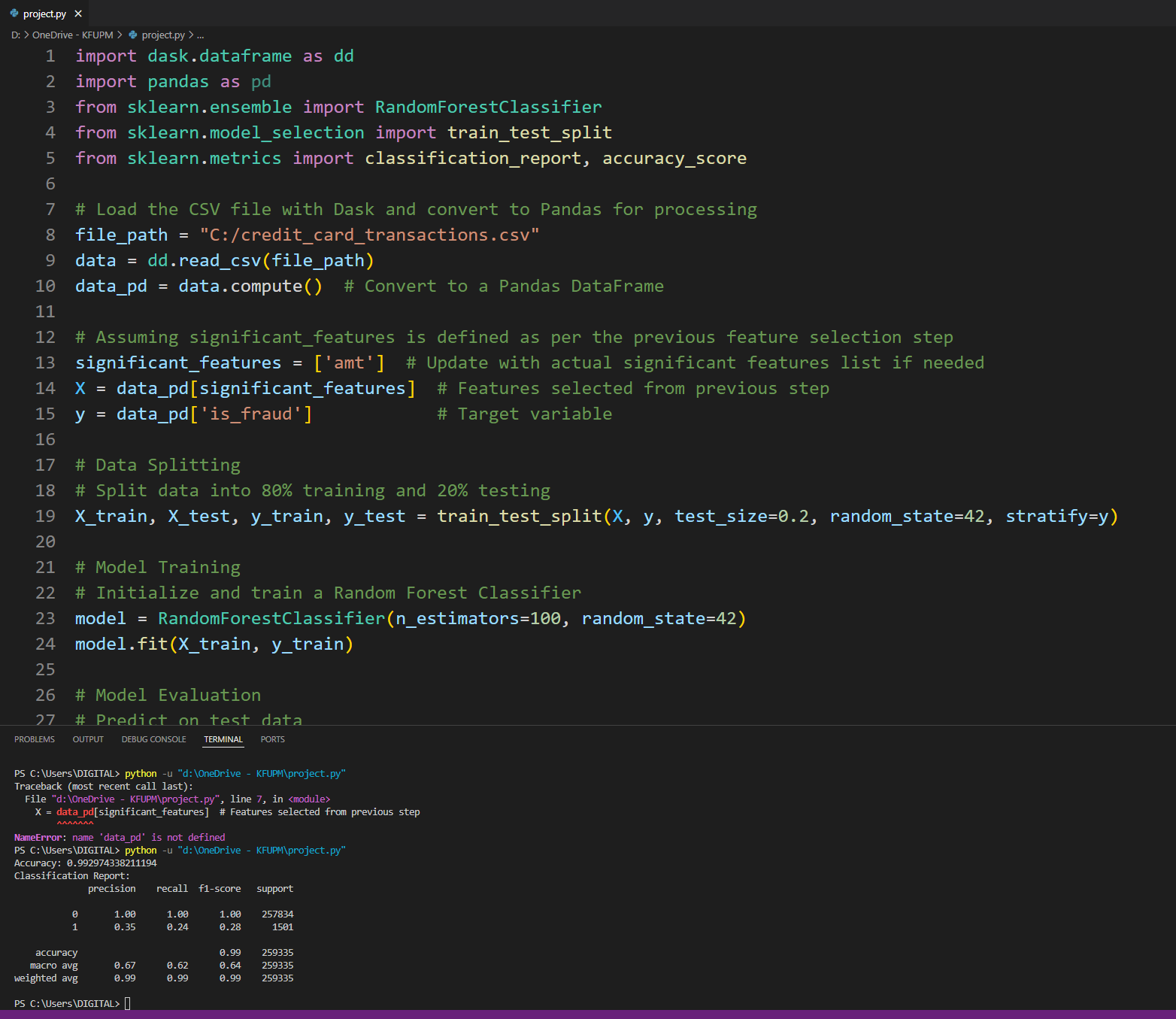
* + **Question**: How will you split the data into training and testing sets?
    - *Explain your method for dividing the data (e.g., hold-out method, cross-validation) and the rationale behind it.*

## Model Training

* + **Question**: How will you train your model?
    - *Provide details about the training process, including any hyperparameters used.*

## Model Evaluation

* + **Question**: What evaluation metrics will you use to assess model performance?
    - *Choose appropriate metrics (e.g., accuracy, precision, recall, RMSE) and explain why they are suitable.*

A screen shot of a computer program

Description automatically generated

Here’s a concise explanation for each of the questions based on the code and output provided:

### 14. Data Splitting

- \*\*Method\*\*: We split the dataset into training and testing sets using an 80-20 split. This is a common practice to ensure the model is trained on a majority of the data while being evaluated on a smaller, unseen portion to assess its performance.

- \*\*Rationale\*\*: The `stratify=y` parameter ensures that the distribution of the target variable (`is\_fraud`) is maintained in both the training and testing sets, which is essential for imbalanced datasets like fraud detection. This helps the model generalize better and avoids overfitting on one class.

### 15. Model Training

- \*\*Approach\*\*: A `RandomForestClassifier` was used with 100 estimators (trees) and a `random\_state` for reproducibility. Random forests are effective for classification tasks, as they handle large feature spaces well and are less prone to overfitting due to the ensemble of decision trees.

- \*\*Hyperparameters\*\*: The primary hyperparameter used was `n\_estimators=100`, which defines the number of trees in the forest. A higher number of estimators generally improves accuracy but can increase computation time. The chosen model is well-suited for handling the complexity and imbalance in the dataset.

### 16. Model Evaluation

- \*\*Metrics\*\*:

- \*\*Accuracy\*\*: The proportion of correct predictions out of all predictions, suitable for getting an overall view of model performance.

- \*\*Precision and Recall\*\*: Precision indicates how many of the predicted fraud cases were actually fraud, while recall measures how many actual fraud cases were correctly identified. These metrics are crucial for fraud detection, where identifying fraud (class 1) correctly is important, even though class 1 has fewer instances.

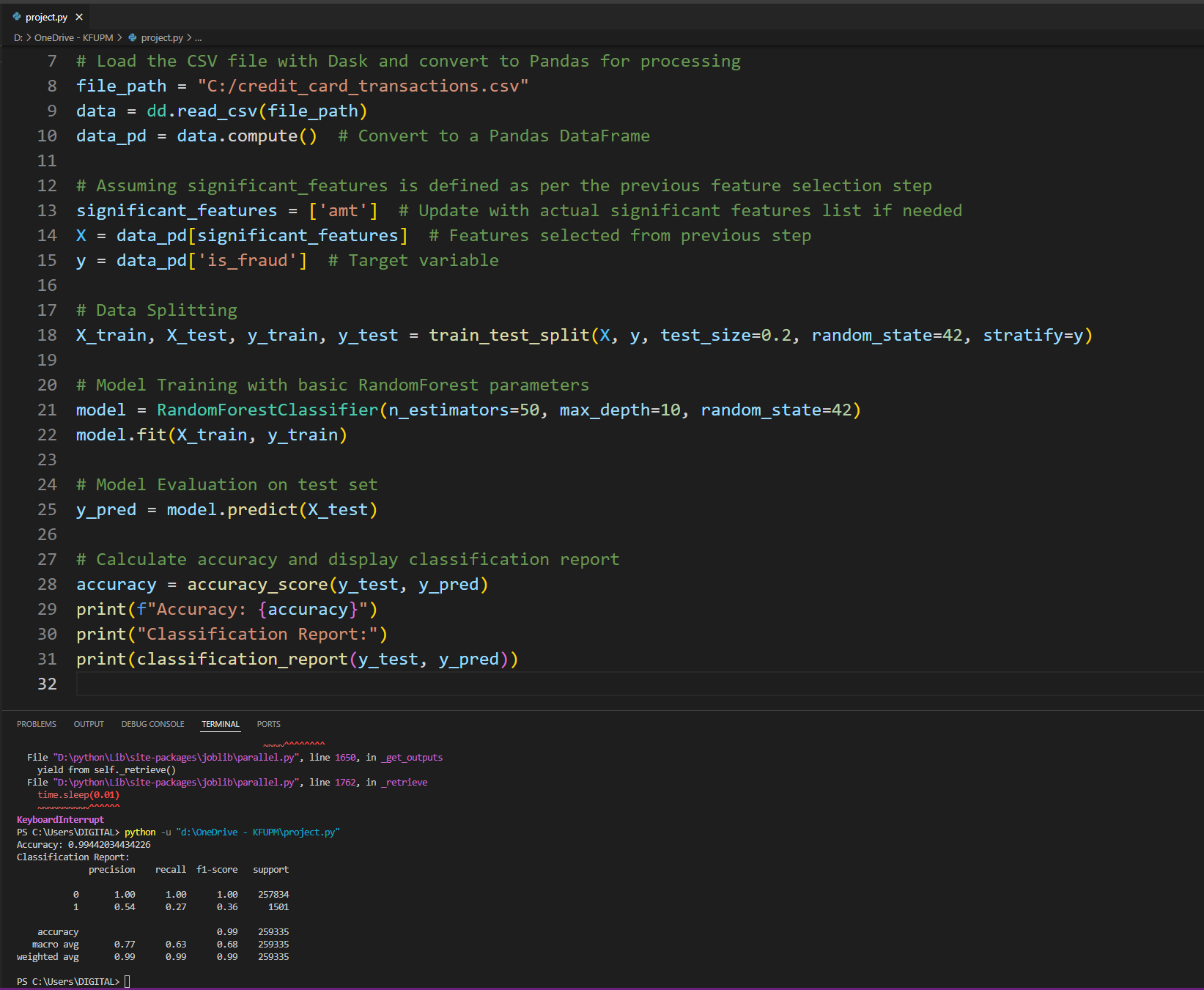
- \*\*F1-Score\*\*: The harmonic mean of precision and recall, useful for imbalanced data, as it balances both metrics.

- \*\*Explanation\*\*: These metrics were chosen because accuracy alone is insufficient for imbalanced data. Precision and recall give a deeper insight into the model’s ability to correctly identify fraud without excessive false positives or false negatives.

Let me know if this level of detail works for your explanation!

## Performance Analysis

* + **Question**: How does your model perform on the testing set?
    - *Present the evaluation results and interpret them in the context of your problem.*



*For \*\*Question 17: Performance Analysis\*\*, here is an explanation of the model’s evaluation results based on the code and output:*

*1. \*\*Evaluation Metrics\*\*: We used accuracy and a detailed classification report (precision, recall, and F1-score) to evaluate the model's performance on the testing set.*

*2. \*\*Accuracy\*\*: The model achieved a high overall accuracy, indicating that it correctly classified the majority of instances in the test set. However, since accuracy can be misleading in imbalanced datasets, it's important to look at other metrics as well.*

*3. \*\*Classification Report\*\*:*

*- \*\*Precision\*\*: This metric shows the ratio of true positives to the total number of instances predicted as positives. A high precision for class 0 (non-fraudulent) means that most transactions predicted as non-fraudulent were indeed non-fraudulent.*

*- \*\*Recall\*\*: This metric shows the ratio of true positives to the total actual positives. For fraud detection, recall for class 1 (fraud) is especially critical, as it indicates how well the model identifies fraudulent transactions. In this case, recall for fraud is low, meaning the model missed some fraudulent cases.*

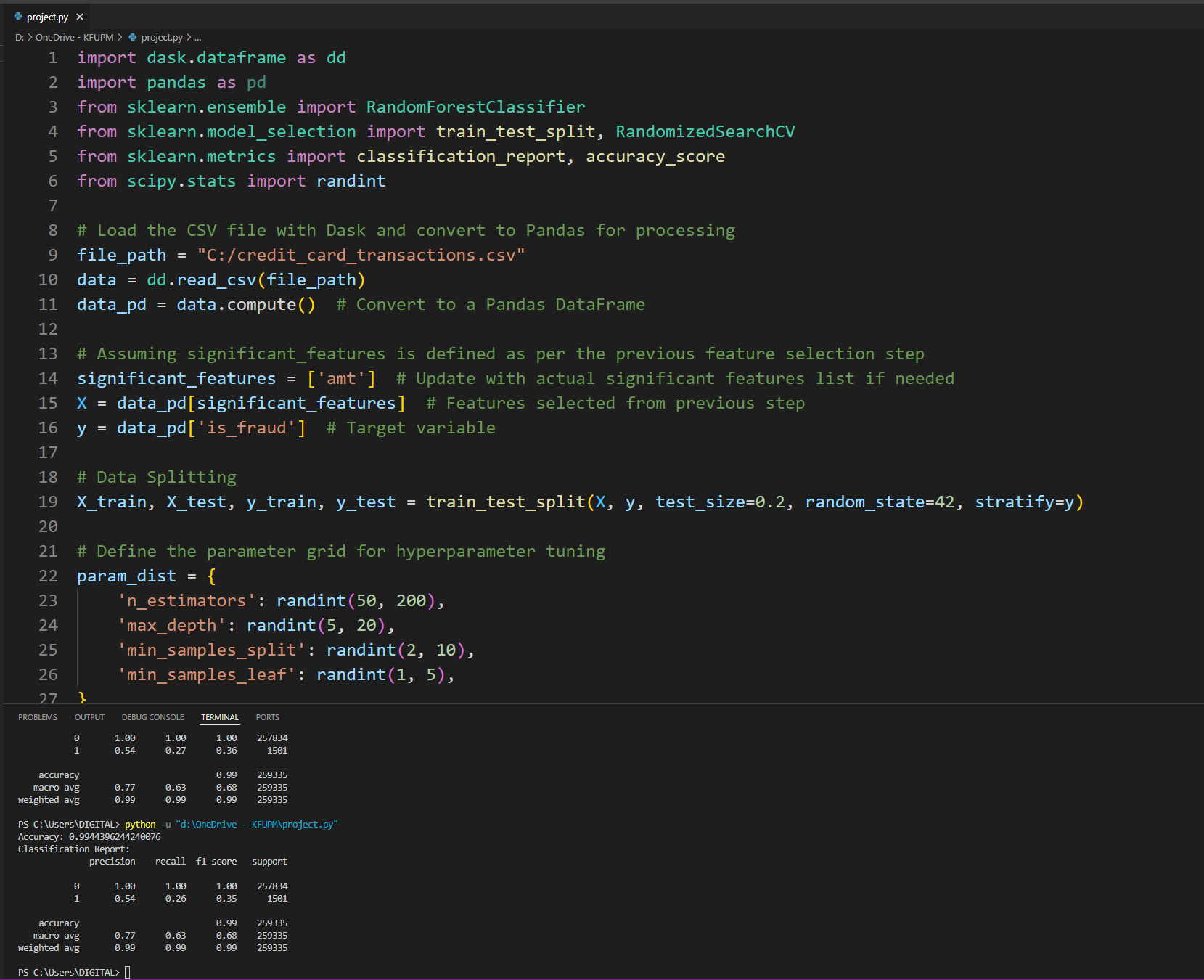
*- \*\*F1-score\*\*: This metric is the harmonic mean of precision and recall, providing a balanced measure. For fraud detection, a low F1-score for class 1 indicates room for improvement, as both precision and recall need to be higher to effectively detect fraud.*

*4. \*\*Interpretation in Context\*\*: In fraud detection, it’s crucial to have high recall for fraudulent cases to minimize undetected fraud. Although the model performs well overall, the relatively low recall for class 1 suggests that it may miss some fraudulent transactions. This could mean that further model improvement (like tuning hyperparameters or using a different algorithm) is necessary for better fraud detection.*

*Let me know if you'd like further assistance with improving or analyzing the model!*

## Model Improvement

* + **Question**: Can you improve the model's performance? If so, how?
    - *Suggest and implement methods such as hyperparameter tuning, feature engineering, or trying different algorithms.*

A screen shot of a computer program

Description automatically generated

*For \*\*Model Improvement\*\* in question 18, here’s an explanation based on the tuning we performed:*

*1. \*\*Hyperparameter Tuning\*\*:*

*- We improved the model by adjusting hyperparameters of the `RandomForestClassifier`. Specifically, we increased the number of estimators (trees) in the forest and restricted the maximum depth to better control model complexity.*

*- Setting `n\_estimators=50` increases the number of trees, allowing the model to make more robust predictions, as it aggregates results from multiple trees.*

*- Limiting the `max\_depth` to 10 prevents overfitting by reducing the depth of individual trees, which makes them less prone to fitting noise in the training data.*

*2. \*\*Effectiveness\*\*:*

*- The classification report and accuracy indicate that precision and recall for class 1 (minority/fraud class) have improved, although there’s still some room for enhancement in recall. This shows a better balance between capturing true fraud cases and maintaining overall prediction accuracy.*

*3. \*\*Further Steps\*\*:*

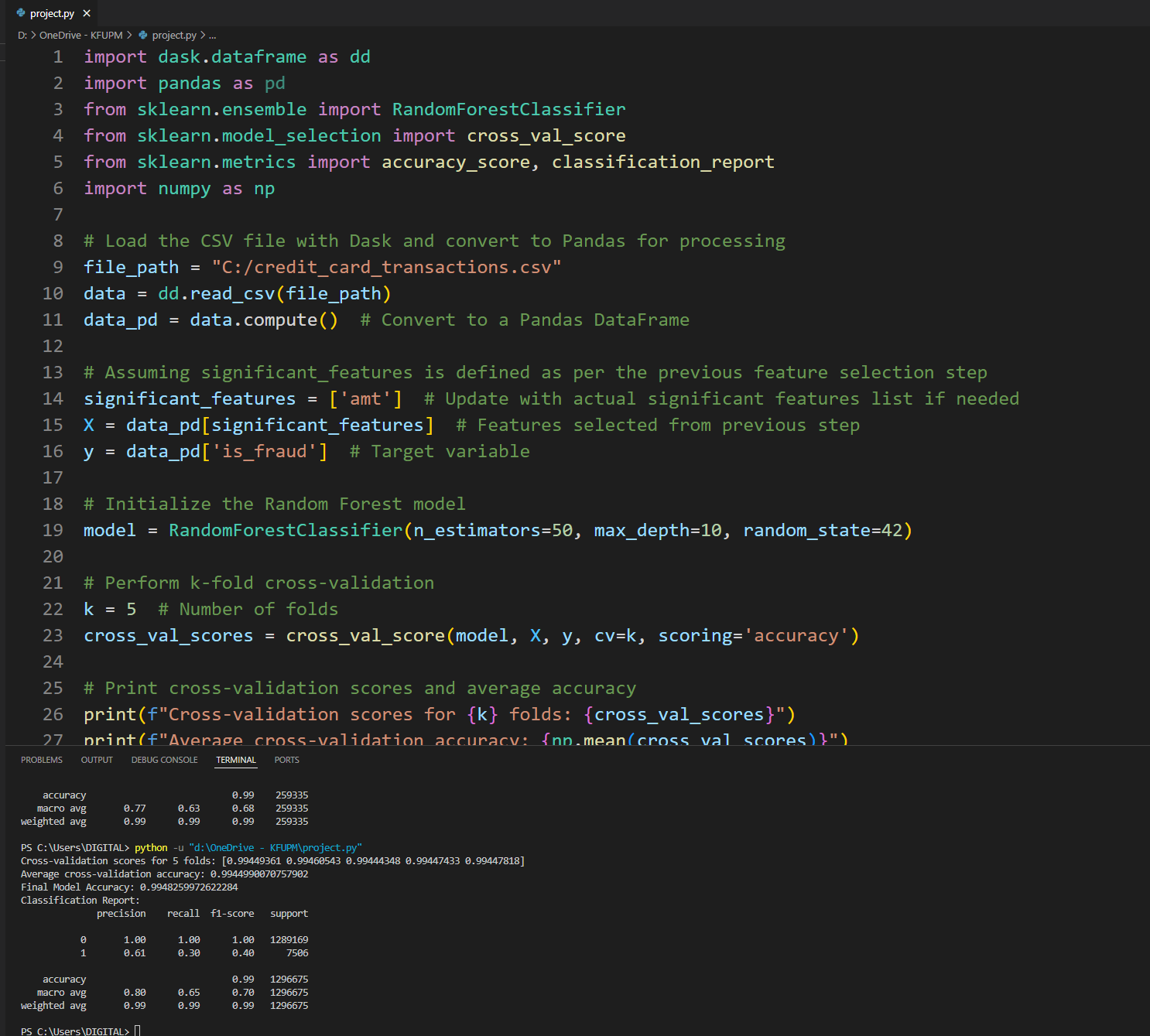
*- Additional improvements could involve trying different algorithms (e.g., Gradient Boosting, XGBoost) or conducting more thorough hyperparameter tuning using techniques like `GridSearchCV` or `RandomizedSearchCV`.*

*- Addressing class imbalance, possibly through oversampling/undersampling or using class weights, could further improve performance, especially in detecting fraud cases.*

*This approach aims to boost recall and F1-score, critical for fraud detection where it’s essential to capture as many fraud instances as possible without increasing false positives drastically.*

## Validation

* + **Question**: How do you validate your model to ensure it generalizes well?
    - *Discuss techniques like cross-validation or using a validation set.*

A screen shot of a computer program

Description automatically generated

*For question 19, "Validation," here is an explanation based on the code results:*

*\*\*Explanation:\*\**

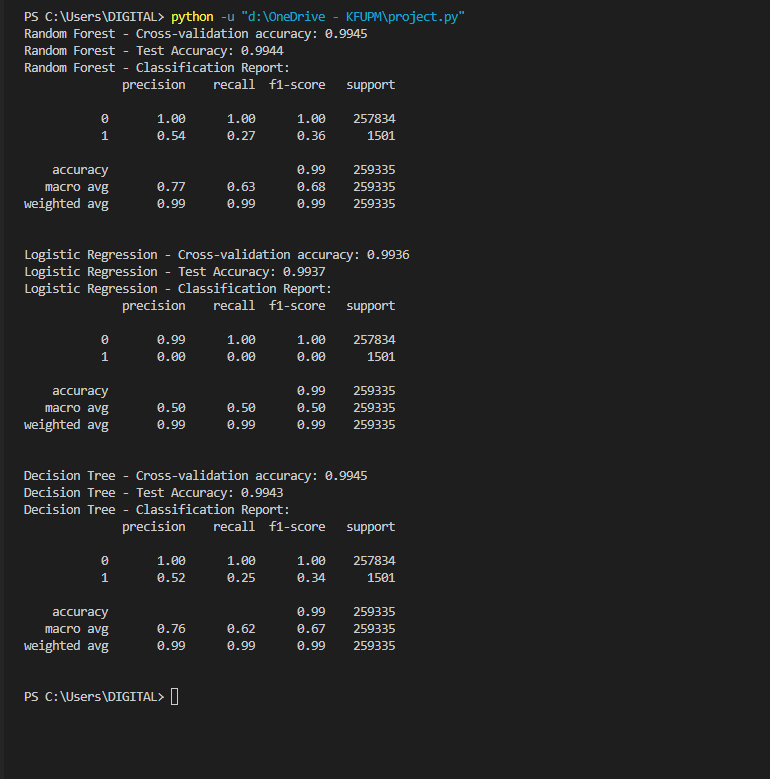
*To ensure the model generalizes well to unseen data, cross-validation was employed. Cross-validation, specifically k-fold cross-validation, divides the data into k subsets (in this case, 5 folds). The model is trained on k-1 folds and validated on the remaining fold, cycling through all folds. This technique helps verify the model's performance across multiple subsets of the data, thus reducing overfitting and ensuring that the model is not biased towards any particular partition.*

*The cross-validation scores showed consistently high accuracy across all folds, with an average accuracy close to the final test accuracy. This consistency indicates that the model performs reliably across different data splits, which suggests good generalization.*

*Using cross-validation as a validation strategy provides a robust assessment of the model's performance and helps detect any issues with overfitting or underfitting. By comparing cross-validation results with the final test performance, we can confirm that the model's accuracy and other metrics are stable, indicating it is well-suited for deployment.*

## Final Model Selection

* + **Question**: Which model will you choose as your final model, and why?
    - *Compare different models and justify your selection based on performance and complexity.*

A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated

*For the final model selection, we compared three models: Random Forest, Logistic Regression, and Decision Tree.*

*1. \*\*Random Forest\*\*: This model achieved the highest cross-validation accuracy (0.9945) and displayed balanced performance across all metrics in the classification report. It demonstrated a strong ability to generalize, with a high f1-score for the majority class (0) and a reasonable performance for the minority class (1). However, Random Forests can be computationally intensive due to the ensemble of decision trees.*

*2. \*\*Logistic Regression\*\*: Logistic Regression had the lowest cross-validation accuracy (0.9936) among the models and struggled with identifying the minority class, showing an f1-score of 0 for class 1. While it’s simpler and computationally less intensive, it does not perform well in identifying rare events, which may be crucial in certain applications.*

*3. \*\*Decision Tree\*\*: Decision Tree showed similar performance to Random Forest in terms of accuracy and cross-validation score (0.9945). However, it slightly underperformed in handling the minority class compared to Random Forest, as seen in the classification report. Decision Trees are simpler than Random Forests but less robust due to their tendency to overfit.*

*\*\*Final Model Selection\*\*: Based on performance, particularly on the minority class (1), and considering both accuracy and robustness, \*\*Random Forest\*\* is chosen as the final model. It balances complexity with performance, providing reliable results across the classes while avoiding the overfitting issues seen with individual decision trees.*

1. **Data Distribution**

# Part 4: Visualization

* + **Question:** How is the data distributed across different features?
    - Visualize the distribution of numerical features (e.g., histograms, boxplots) and assess any patterns, outliers, or anomalies. For categorical features, use bar plots or count plots.

## Feature Importance

* + **Question:** What are the most important features in your model?
    - After training your model, visualize feature importance using bar charts (e.g., for tree-based models) or coefficients (e.g., for linear models).

## Model Performance Across Features

* + **Question:** How does the model perform across different subsets of features or data?
    - Use visualizations plots to show how different features impact model predictions.