1. Data Cleaning

Overview

Data cleaning is an essential step in data preprocessing that involves handling missing values, ensuring data consistency, and preparing the dataset for further analysis and modelling. This step ensures that the data is in a suitable format for machine learning algorithms to process effectively.

Steps

1. **Loading the Dataset**:
    - The dataset is loaded from a CSV file using the pandas library. This allows for easy manipulation and analysis of the data.
2. **Inspecting the Dataset**:
    - The structure of the dataset is examined using methods such as `.info()` and `.describe()`. These methods provide insights into the data types, the presence of null values, and basic statistics of numerical columns.
3. **Handling Missing Values**:
    - Missing values in the 'Memo/Description' column are filled with a placeholder value ('No Description'). This ensures that there are no null values in this column, which could cause issues during data analysis and modelling.
    - Erroneous entries in the 'Split' column (e.g., '-Split-') are replaced with `NaN` to handle them appropriately.
4. **Correcting Erroneous Data**:
    - Erroneous data entries in specific columns are identified and corrected. For instance, the 'Split' column's incorrect values are standardized to ensure consistency.
5. **Converting Data Types**:
    - The 'Date' column is converted to a datetime format to facilitate date-based feature engineering.
    - Columns such as 'Amount' and 'Balance' are ensured to be in float format to allow for numerical operations and statistical analysis.
6. **Dropping Unnecessary Columns**:
    - The 'Memo/Description' column is dropped as it is not needed for the model and contains non-numeric data that could complicate the analysis.

---

2. Exploratory Data Analysis (EDA)

Overview

EDA is a crucial step in understanding the underlying patterns and distributions in the data. This involves visualizing the data to identify trends, outliers, and relationships between variables.

Steps

1. **Distribution Analysis**:
    - Histograms are used to visualize the distribution of the 'Amount' and 'Balance' columns. This helps in understanding the central tendency, spread, and skewness of these numerical variables.
2. **Visualizing the 'Split' Column**:
    - A count plot is created to visualize the distribution of values in the 'Split' column. This helps in understanding the frequency of different categories within this column.

---

3. Feature Engineering

Overview

Feature engineering involves creating new features from the existing data to improve the performance of machine learning models. This step can include extracting information from date columns, encoding categorical variables, and more.

Steps

1. **Extracting Date Features**:
    - New features such as 'Day', 'Month', and 'Year' are extracted from the 'Date' column. This allows the model to capture seasonal patterns and other time-based trends.
2. **Dropping the Original 'Date' Column**:
    - The original 'Date' column is dropped after extracting the necessary features to avoid redundancy.
3. **Encoding Categorical Variables**:

- Categorical variables such as 'Transaction Type' and 'Name' are encoded using one-hot encoding. This converts categorical values into a binary format that can be used by machine learning algorithms.

---

4. Model Development

Overview

Model development involves splitting the data into training and testing sets, standardizing the features, choosing a machine learning model, and evaluating its performance.

Steps

1. **Splitting the Data**:
   - The dataset is split into features (X) and target (y). The features include all columns except the target variable 'Split'.
   - Any rows with missing values in the target column are removed to ensure a clean dataset for training and testing.
2. **Standardizing the Features**:
   - The features are standardized to have a mean of 0 and a standard deviation of 1. This step is essential for many machine learning algorithms to perform optimally.
3. **Choosing and Training the Model**:
   - A Random Forest Classifier is selected as the machine learning model. This model is trained on the training set.
4. **Evaluating the Model**:
   - The model's performance is evaluated using metrics such as classification report and accuracy score. These metrics provide insights into the model's precision, recall, F1-score, and overall accuracy.

---

5. Documentation

Data Cleaning

- **Missing Values**: Missing values in the 'Memo/Description' column are filled with 'No Description'. Erroneous entries in the 'Split' column are corrected.
- **Data Types**: The 'Date' column is converted to datetime, and 'Amount' and 'Balance' columns are ensured to be floats.
- **Column Removal**: The 'Memo/Description' column is dropped.

Exploratory Data Analysis (EDA)

- **Distribution Analysis**: Histograms for the 'Amount' and 'Balance' columns.
- **Split Column Visualization**: Count plot for the 'Split' column.

Feature Engineering

- **New Features**: Day, Month, and Year extracted from the 'Date' column.
- **Encoding**: Categorical variables are one-hot encoded.

Model Development

- **Data Split**: The dataset is split into training and testing sets.
- **Standardization**: Features are standardized.
- **Model Selection**: Random Forest Classifier is chosen and trained.
- **Evaluation**: Model performance is evaluated using classification report and accuracy score.