Group 18 Report: Data Preprocessing (Formative 2)

1. Steps Taken in Preprocessing

a) Data Loading

- The dataset customer_transactions.csv was loaded into a Pandas DataFrame.
- An initial exploration of the data was conducted using head(), info(), and describe() to understand its structure, data types, and basic statistics.

b) Handling Missing Values

- Missing values were identified in the customer_rating column.
- To address this: **Mean, Mode, Median & KNNImputer imputation** was applied to the customer_rating column in different cells.
- After imputation, there were no remaining missing values in the dataset.

c) Data Type Conversion and Cleaning

- Data types were checked to ensure compatibility with machine learning algorithms.
- purchase_date was parsed and stored as a datetime object for future processing.
- Non-numeric columns like product_category were encoded later during feature engineering for compatibility.

d) Exploratory Data Analysis (EDA)

- Statistical summaries were generated using describe() to identify potential anomalies.
- purchase_amount and customer_rating were plotted using histograms and boxplots.
- Outliers in purchase_amount were visually detected, indicating potential skewness.

e) Feature Encoding

- Categorical features like product_category were encoded using OneHotEncoder and LabelEncoder.
- A ColumnTransformer pipeline was applied to standardize preprocessing for categorical and numerical features.

2. Summary of Key Insights Found During Preprocessing

- The customer_rating column had 10 missing values, which were successfully handled through the 4 imputations.
- The purchase_amount column exhibited skewness and potential outliers, warranting further transformation.
- The dataset contained a class imbalance in product_category, which was later addressed using SMOTE.
- Proper encoding and handling of categorical variables were necessary to prepare the data for machine learning models.

3. Challenges Faced and How They Were Solved

a) Merging Three Different Notebooks

- Each team member worked in separate notebooks. Clear role allocation and stepwise merging were used to integrate these notebooks into a cohesive single file.

b) The Discrete Nature of the Tasks

- One couldn't proceed until the previous task was completed. A shared timeline was maintained for smooth hand-offs between tasks.

c) Creating Custom Sentiment Formulas

- Coming up with sentiment formulas manually was complex. Different approaches was tested until the satisfactory formula was gotten.

d) Choosing a Method for Feature Selection

- Selecting an appropriate feature selection method required weighing complexity. SelectKBest was used because of its simplicity and straightforwardness.

4. Video Presentation = https://youtu.be/6AM59Pw5AsQ

5. Group Members

- **Eunice Adewusi:** Performed data augmentation and handled missing values (Part 1).
- **Christian Mutabazi:** Managed dataset merging and engineered new features (Part 2).
- **Theodora Omunizua:** Conducted data consistency checks, statistical analysis, and feature selection (Part 3).