DATA PREPROCESSING REPORT

Group 17

Contributors:

- **John Akech** – Part 1: Data Cleaning & Handling Missing Values
- **Kuir Juach Kuir Thuch** – Part 2: Data Augmentation & Merging Datasets
- **Geu Aguto Garang** – Part 3: Feature Engineering & Data Quality Checks

---

1. Overview

The goal of this project was to refine and prepare datasets for machine learning by performing cleaning, augmentation, merging, and feature engineering. This process involved handling missing values, resolving inconsistencies, generating synthetic data, and ensuring high-quality data for predictive modeling.

---

2. Steps in Preprocessing

Part 1: Data Cleaning & Handling Missing Values (John Akech)

- **Identified Missing Values:** Found gaps in key columns such as `customer_rating`, `purchase_date`, and `review_sentiment`.
- **Imputation:**
  - Numerical values were filled using the median.
  - Categorical values were replaced with the most frequent category.
- **Data Type Adjustments:**
  - Converted `purchase_date` to datetime format.
  - Extracted year, month, and day for time-based analysis.
- **Encoding Categorical Variables:**
  - One-hot encoding was applied to categorical variables like `product_category` for machine learning compatibility.

Part 2: Data Augmentation & Merging Datasets (Kuir Juach Kuir Thuch)

- **Synthetic Data Generation:**
    - Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance `customer_rating`.
    - Introduced random noise to numerical features such as `purchase_amount` to enhance data diversity.
    - Saved the augmented dataset as `customer_transactions_augmented.csv`.
- **Merging Datasets:**
    - Merged `customer_transactions_augmented.csv` with `social_profiles.csv` using `id_mapping.csv`.
    - **Conflict Resolution:**
        - Aggregated duplicate rows by taking the mean for numerical values.
        - Used the most frequent category for categorical values.
    - The final merged dataset was saved as `final_customer_data_group17.csv`.

Part 3: Feature Engineering & Data Quality Checks (Geu Aguto Garang)

- **Feature Engineering:**
    - **Behavioral Features:** Created `moving_avg_purchase` (rolling transaction average) and `customer_engagement_score` based on engagement metrics.
    - **Text-Based Features:** Applied **TF-IDF vectorization** to convert `review_sentiment` into meaningful numerical features.
- **Data Quality Checks:**
    - Identified and removed duplicate entries.
    - Verified that all transactions were linked to valid social profiles.
    - Generated descriptive statistics to detect trends and anomalies.

---

3. Key Insights from Preprocessing

- **Purchase Amount Skewness:**
    - Applied transformations to normalize the slightly skewed `purchase_amount` distribution.
- **Feature Correlations:**
    - A heatmap revealed a strong relationship between `purchase_amount` and `customer_engagement_score`, highlighting potential multicollinearity.
- **Feature Importance:**

    - Used **SelectKBest** to identify the top 10 most impactful features, including `purchase_amount`, `customer_rating`, and `engagement_score`.
- **Impact of Data Augmentation:**
    - Synthetic data improved class diversity, particularly in `customer_rating`.

---

4. Challenges & Solutions

Missing Target Values:

- **Issue:** Missing values in `customer_engagement_score` caused errors during training.
- **Solution:** Dropped affected rows and applied imputation where necessary.

Inconsistent Column Names:

- **Issue:** Mismatched names (`customer_id_new` vs. `customer_id_legacy`) complicated merging.
- **Solution:** Standardized naming conventions using `id_mapping.csv`.

Class Imbalance in Ratings:

- **Issue:** Imbalanced `customer_rating` led to biased model predictions.
- **Solution:** Applied **SMOTE** to generate synthetic samples for underrepresented classes.

Handling Text Data:

- **Issue:** Unstructured text in `review_sentiment` was incompatible with numerical models.
- **Solution:** Applied **TF-IDF vectorization** to extract sentiment patterns.

---

5. Conclusion

Through careful preprocessing, we transformed raw data into a structured and machine-learning-ready format. By addressing missing values, balancing classes, merging datasets, and engineering features, we ensured the dataset was optimized for predictive modeling. Challenges were effectively tackled with robust techniques, resulting in the final dataset, `final_dataset_ready_group17.csv`, ready for advanced analytics and model development.