Data Preprocessing Project Report

Introduction

This report presents the preprocessing steps performed on two real-world datasets to ensure they are machine learning-ready. The work was divided into three parts: data augmentation, dataset merging, and consistency checks.

Part 1: Data Augmentation (Geofrey Tumwesigye)

Objectives:

- Expand the dataset using synthetic data.
- Handle missing values efficiently.
- Apply transformations to improve data quality.

Methods Used:

1. **Missing Values Handling:** Used median imputation for numerical features.
2. **Synthetic Data Generation:**
    - Applied **random noise** to transaction amounts.
    - Used **SMOTE** to balance the dataset.
    - Duplicated transactions with added perturbations.
3. **Feature Transformations:**
    - Applied **log transformation** to normalize skewed data.
    - Introduced new synthetic transactions.

Results:

- The dataset was expanded from **150** rows to **650** rows.
- The cleaned and augmented dataset was saved as `customer_transactions_augmented.csv`.

Part 2: Merging Datasets (Serge Kamanzi)

Objectives:

- Merge customer transactions with social media profiles.
- Ensuring consistency in customer IDs using an intermediate mapping.
- Create new features for better insights.

Methods Used:

1. **ID Mapping:** Used `id_mapping.csv` to link old and new customer IDs.
2. **Dataset Merging:** Combined three datasets using left joins.
3. **Feature Engineering:**
    - Created a **Customer Engagement Score** based on transaction history and social media activity.
    - Computed **moving averages** for transactions.

  - Applied **TF-IDF vectorization** to analyze customer reviews.

Results:

- Successfully merged datasets while resolving ID inconsistencies.
- Generated a clean dataset with **new behavioral features**.
- Saved as `final_customer_data_5.csv`.

Part 3: Data Consistency & Quality Checks (Willy Kalisa)

Objectives:

- Validate data integrity and remove inconsistencies.
- Summarize key statistical distributions.
- Select the most relevant features for modeling.

Methods Used:

1. **Data Integrity Checks:**
   - Removed duplicate records (**52 duplicates found and removed**).
   - Ensured categorical values were correctly mapped.
   - Verified customer transactions match valid social profiles.
2. **Statistical Summarization:**
   - Generated **describe()** reports for numerical features.
   - Plotted **transaction amount distributions** before and after augmentation.
3. **Feature Selection:**
   - Applied a **correlation heatmap** to identify highly correlated features.
   - Used **SelectKBest** to extract the **top 10 features**.

Results:

- Data quality was significantly improved.
- The final dataset, `final_dataset_ready_5.csv`, was prepared for machine learning.

Key Insights & Challenges

Insights Gained:

- Data augmentation significantly improved dataset size and quality.
- Merging transitive datasets required precise mapping strategies.
- Feature engineering helped enhance predictive potential.

Challenges Faced:

- **Handling missing values** in multiple datasets required different imputation techniques.
- **Ensuring consistency** across datasets with different ID formats was challenging.
- **Feature selection** required careful analysis to retain only meaningful attributes.

Conclusion

This project successfully transformed raw datasets into structured, clean, and machine learning-ready datasets. The preprocessing pipeline ensures improved data quality, consistency, and usability for future modeling.

Final Outputs:

- `customer_transactions_augmented.csv`
- `final_customer_data_5.csv`
- `final_dataset_ready_5.csv`