Group 15 - Machine Learning Data Preprocessing Assignment Report

Group Members:
1. Audry Ashleen Chivanga

2. Humphrey Nyahoja

3. Dieudonne Ngum

Repository Link:
[GitHub Repository]

Preprocessing Steps Taken

We began by cleaning the `customer_transactions.csv` dataset, handling missing data (using mean imputation for `customer_rating`). To balance the dataset, we applied SMOTE for synthetic data generation, log transformations, and expanded data with new synthetic transactions. Next, we merged this with `customer_social_profiles.csv` using `id_mapping.csv`, resolving ID conflicts and engineering new features like the **Customer Engagement Score** and predictive behavioral metrics. Finally, we ensured data consistency by removing duplicates, validating categorical values, and performing feature selection for machine learning readiness.

Challenges Faced & Solutions

We faced several challenges:

- **Missing Data:** We imputed missing values using the column mean, ensuring no loss of data.
- **Merging Datasets with Different IDs:** We resolved this using `id_mapping.csv` to correctly map customer IDs across datasets.
- **Feature Engineering Complexity:** Creating meaningful features was challenging but solved by combining domain knowledge and statistical methods.
- **Data Consistency Issues:** We performed rigorous validation checks, including removing duplicates and ensuring all transactions were linked to valid profiles.

Deliverables

- `customer_transactions_augmented.csv`
- `final_customer_data_group15.csv`
- `Final_dataset_ready_group15.csv`
- `All notebooks in notebook folder`
- This summary report
- A video presentation our README.md explaining the preprocessing steps and every team member’s contribution.