Formative 2 - Data Preprocessing Assignment

Summary Report

Members:
- MAHAMAT AHAMAT; Adediwura Emmanuel; Henriette Cyiza; Sabir Walid Abdurahman

Preprocessing Steps

This assignment involved enhancing two real-world datasets by handling missing values, merging data with transitive properties, and engineering new features to improve machine learning readiness. The workflow included:

1. **Data Augmentation** – Synthetic data generation, applying noise to numerical values, balancing with SMOTE, and log transformation.
2. **Dataset Merging** – Linking customer transactions with social media profiles using an ID mapping file and resolving inconsistencies.
3. **Data Consistency Checks** – Identifying duplicates, validating categorical values, and ensuring transaction-social profile alignment.

Key Insights

- Data augmentation improved dataset balance and distribution.
- Merging via transitive relationships required careful ID resolution.
- Feature engineering enhanced predictive potential.
- Statistical summaries provided clarity on data distribution and integrity.

Challenges & Solutions

- **Missing Values**: Solved using mean/mode imputation and predictive modeling.
- **Merging Complexity**: Addressed by ensuring correct ID mapping and resolving many-to-one relationships.
- **Feature Selection**: Used correlation analysis and selection algorithms to refine features.

Final Outcome

The preprocessed dataset is structured, balanced, and ready for machine learning applications, ensuring better model performance and accuracy.

Github repo link:
https://github.com/MAHAMAT263/Data_Preprocessing_Assignment_for_ML_Pipeline.git

Video demo link:

https://drive.google.com/file/d/1WUPsZo2qEQBJtbxUvmldjcVH18XmEKKM/view?usp=sharing