Data Preprocessing Summary Report

Introduction

This report summarizes the data preprocessing steps undertaken, key insights derived, and challenges encountered during the process. The goal was to prepare a high-quality dataset for machine learning by handling missing values, merging datasets, and performing feature engineering.

Steps Taken in Preprocessing

1. **Handling Missing Values**
    - Identified missing data in numerical and categorical features.
    - Applied mean imputation for numerical fields and mode imputation for categorical fields.
    - Used predictive imputation techniques where applicable.
2. **Synthetic Data Generation**
    - Augmented data using SMOTE for underrepresented classes.
    - Generated additional samples based on statistical distributions to balance the dataset.
3. **Dataset Merging & ID Mapping**
    - Merged multiple datasets using a transitive ID mapping file.
    - Ensured consistency by verifying relational integrity across tables.
4. **Feature Engineering & Transformation**
    - Created new features from existing columns to enhance predictive power.
    - Normalized numerical features using Min-Max scaling.
    - Applied one-hot encoding and label encoding to categorical variables.
5. **Consistency Checks & Quality Assurance**
    - Detected and removed duplicate records.
    - Standardized column names and ensured uniform data formats.
    - Performed outlier detection using Z-score analysis and visualizations.

Key Insights

- Data augmentation improved class balance, leading to better model training potential.
- Normalization significantly reduced feature skewness, improving data distribution.
- Feature engineering enhanced interpretability and correlation with target variables.

Challenges and Solutions

1. **High Percentage of Missing Data**
    - Solution: Used a hybrid approach of imputation and synthetic data generation.
2. **Merging Large Datasets with ID Mapping Issues**
    - Solution: Cleaned and standardized identifiers before merging, ensuring alignment.

Conclusion

Through rigorous preprocessing, the dataset is now clean, consistent, and optimized for modeling. The structured approach to handling missing data, feature engineering, and quality assurance has enhanced data usability, setting a solid foundation for machine learning applications.