Data Preprocessing Report for School Drop

Educational institutions need accurate predictive models to identify students at risk of dropping out. This project aims to develop and optimize predictive models to forecast student dropout rates and recommend interventions to improve retention and student success.

Dataset Overview

Dataset Description: the dataset contains 4424 row and 37 columns

Target variable: the target variable is categorical representing the status of students:

Dropout: students who did not complete their education

Enrolled: students who are still in school

Graduated: students who completed their education

Data Quality Check

Missing values: there are no missing values in the dataset

Duplicate values: there are no duplicate values in the dataset

Outliers detection: outliers where detected using boxplot and z-score

Feature Engineering

Encoding categorical variables:

One-hot encoding: this techniques was used to convert categorical variables to numerical, the target out containing Dropout, Enrolled, and Graduated because the categories where unordered

Handling imbalanced data

Class weights: for algorithms such as random forest, and logistic regression, class weight was added to balance assign a higher penalty for misclassifying the minority class i.e. enrolled and dropout

Data Transformation

Scaling of features

Standard scalar: