

Title: Improving Predictive Analytics for Student Dropout: A Comprehensive Analysis and Model Evaluation

Author: Wahid Sabbir; Md Abdullah-Al-Kafi; Arafat Sahin Afridi; Md. Sadekur Rahman; Mousumi Karmakar

Abstract: This research project uses careful data preparation and machine learning model assessment to provide an in-depth analysis of a dataset of students in college or university. The first analysis looks at goal value distributions, economic variables, and student counts by gender. The handling of outliers, feature selection, and class imbalance are all addressed by further filtering. Using ROC curves to highlight classification strength, the study assesses several classifiers, including XGBoost, Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. With the greatest AUC of 0.99, Random Forest remarkably shows excellent predictive power, closely followed by XGBoost at 0.98. XGBoost performs exceptionally well on testing and training datasets. The findings contribute valuable insights into predictive modeling for understanding and predicting student outcomes, emphasizing the potential to enhance educational support systems. This integrated approach, combining exploratory data analysis and machine learning techniques, establishes a robust framework for future research in educational data mining and predictive analytics.

Keywords: Predictive Analytics, Random Forest, XGBoost, ROC Curve, Data Preprocessing, Class Imbalance

DOI: <https://doi.org/10.23919/INDIACom61295.2024.10498520>

<https://ieeexplore.ieee.org/abstract/document/10498520>