ABSTRACT

Student dropout is a major challenge for higher education institutions, as it affects both the quality of education and the social and economic development of society. Machine learning (ML) techniques can offer valuable insights into the factors and patterns that influence student retention and success, and enable early intervention strategies to prevent dropout. However, one of the main difficulties in applying ML to student dropout prediction is the class imbalance problem, which occurs when the number of students who drop out is much lower than the number of students who persist. This problem can cause bias towards the majority class and fail to capture the characteristics of the minority class. To address this issue, this work proposes a novel approach that combines synthetic minority over-sampling technique (SMOTE) and XGBoost to predict student dropout in higher education using class-imbalanced datasets. SMOTE is a technique that generates synthetic samples of the minority class to balance the dataset, while XGBoost is a robust and efficient ensemble method that can handle high-dimensional and noisy data. The proposed approach is applied to a dataset of undergraduate students from a Portuguese Polytechnic University, enrolled between 2009 and 2017. The dataset contains academic, demographic, and socioeconomic information at different stages of the first academic year.