ISSN: 2773-7055
Proceedings of 3rd International Research Conference
Institute of Technology, University of Moratuwa
December 18, 2024

A PREDICTIVE ANALYSIS OF STUDENT DROPOUTS IN IT HIGHER EDUCATION PROGRAMMES

U.G.N. Kumari*
Faculty of Information Technology, University of Moratuwa, Sri Lanka
niranjala23@yahoo.com*

ABSTRACT: The study primarily aims at identifying the key attributes that contribute to student dropouts in Information Communication Technology (ICT) courses offered by Higher Education Institutes, a significant issue in educational data mining. It seeks to explore the distinct factors influencing dropout rates that have been underexplored in existing literature. Data was collected from five batches of students enrolled in an Information Technology course at a government tertiary education institute in Sri Lanka. The collected data underwent pre-processing and feature selection was carried out using the Correlation-Based Feature Selection (CFS) to pinpoint subsets of attributes closely linked to dropout outcomes. Mostly used classification algorithms were evaluated based on their performance using confusion matrix metrics. Therefore, this study trains a set of classification models, namely Decision Tree, K-Nearest Neighbor, Naïve Bayes, and Rule-Based approaches that attained an accuracy of over 83.17% in defining strong associations between dropout factors and dropout status which is known as "Yes" and "No". J48 Decision Tree was the topmost algorithm for this dataset, and the predictive modeling of student profiles was done using the same. The model's performance was validated using a new dataset sourced from institutional records. The dropout prediction application was implemented using the Java WEKA API and achieved 92.61% accuracy in predicting student dropouts in ICT higher education in all educational streams. By uncovering strong relationships between dropout factors and dropout status, the study highlights key influences, with the most significant factors being perceived course quality, previous academic qualifications, previous ICT experience, Ordinary Level results, and English proficiency level in the Sri Lankan context. This model can be utilized to predictively analyze student dropouts in ICT higher education, allowing early identification of at-risk students and facilitating targeted intervention strategies.

Keywords: classification algorithms, data mining, dropouts

1. INTRODUCTION

Student academic performance is vital for educational institutions, especially in identifying at-risk students for timely interventions. In higher education, dropout rates pose a significant challenge, affecting student success and institutional reputation. Although existing studies have examined dropout predictors, a gap remains in understanding the unique factors influencing dropout rates within ICT programs. Prior research in educational data mining (EDM) has used techniques like regression, clustering, and classification to reveal academic performance patterns. However, these often generalized findings fail to capture the distinct characteristics of ICT education, which requires specific skills and learning styles. Moreover, factors such as academic background, socioeconomic status, and prior ICT experience are underexplored in dropout prediction, particularly in datasets spanning multiple years and cohorts. This study addresses these gaps by focusing on ICT programs in Sri Lanka's tertiary education, specifically through classification techniques in data mining to identify key dropout attributes. Using data from the Higher National Diploma in Information Technology at the Sri Lanka Institute of Advanced Technological Education, the research evaluates classification algorithms to determine the most effective dropout predictor. This approach tailors the predictive model to ICT students' unique profiles, moving beyond the "one-size-fits-all" method. By focusing on ICT-specific attributes, this research aims to help educational institutions mitigate dropout rates and foster a more adaptive educational environment.

1.1. Related Work

Educational predictive analysis uses data from schools and online platforms, applying data mining techniques like classification algorithms to reveal learning patterns and relationships (Lee & Chung, 2019). Educational Data Mining (EDM) employs classification and prediction to interpret student behaviour, converting large datasets into actionable insights. Known as Knowledge Discovery in Databases (KDD), DM identifies academic performance factors, with algorithms like Neural

Networks, Decision Trees, SVM, and Naïve Bayes widely used for accurate predictions(Jalota & Agrawal, 2019). Studies show that factors such as demographics, income, motivation, mental health, institutional support, and curriculum relevance affect dropout rates, highlighting the importance of multidimensional analysis in addressing dropout issues(Ortiz-Lozano et al., 2020).

1.2. Theory and Technology Use

This study combines feature selection, classification techniques, and cross-validation to optimize predictive accuracy in data analysis and machine learning. Using Correlation-Based Feature Selection (CFS) and the WEKA platform, the research leverages both supervised and unsupervised learning to handle data complexity effectively. The WEKA tool, implemented in Java, supports a wide range of machine learning algorithms, providing a strong foundation for model development and analysis. This integrated approach ensures a robust framework for extracting meaningful insights from data, balancing theoretical principles with advanced technological tools(Kemper et al., 2020).

2. METHODOLOGY

This study investigates the effectiveness of classification algorithms in data mining to predict student dropouts in Information Technology (IT) higher education. It aims to address four critical questions: identifying dropout factors through correlation-based feature selection, determining the most relevant factors, selecting the best classification algorithm, and evaluating the accuracy of prediction tools like the Decision Tree Algorithm. The goal is to develop a reliable predictive model for dropout rates in IT programs, ensuring that the evaluation of dropout data is timely, dependable, and adaptable. The research utilizes a training dataset containing instances with a target class attribute, essential for constructing a stable model. The accuracy of the model is validated against a separate test dataset. The study anticipates first identifying key factors contributing to student dropouts through a literature review, followed by using correlation-based feature selection to highlight the most significant factors(Yağcı, 2022). Various algorithms will be tested on collected datasets to determine the most accurate classification technique for constructing the predictive analysis model, and the performance will be measured.

The dataset comprises theoretical and empirical factors impacting student performance, including socio-demographic indicators (age, geographical location, parents' education, and income), educational factors (performance in O/L and A/L examinations, prior ICT education, and entrance exam scores), as well as institutional, psychological, and social integration factors. Data preprocessing is crucial, involving evaluation of data quality and integrity. Missing values are addressed through mean and mode imputation for continuous and categorical variables, respectively, to maintain dataset integrity. Continuous variables like English scores and income are normalized to a common scale (0 to 1), preventing larger range variables from skewing results, particularly in distance-based algorithms like K-Nearest Neighbor. Categorical variables are encoded using one-hot encoding to ensure compatibility with machine learning algorithms, avoiding biased interpretations that can arise from inherent ordering in categorical data. Outliers in continuous variables are managed using z-score analysis, capping values beyond three standard deviations to preserve the majority of the dataset while reducing noise.

After preprocessing, the dataset is prepared for data mining techniques. The feature selection process identifies the most useful inputs for analysis by eliminating irrelevant attributes. Based on previous studies, several classifiers are tested, including Decision Tree (J48), K-Nearest Neighbor (Lazy-IBK), Naïve Bayes, and Rule-Based ZeroR. Test datasets with record sizes of 500 to 1000 are created to assess changes in accuracy for each algorithm. Classification accuracy is evaluated using the test data, and once satisfactory, classification rules can be applied to new data. The findings reveal that

the J48 Decision Tree and K-Nearest Neighbor classifiers are the top performers, with the J48 Decision Tree achieving 100% accuracy in predicting student dropouts. Consequently, it is selected as the preferred model for application. To achieve high accuracy and mitigate over-fitting, the study employs cross-validation, pruning techniques for the Decision Tree classifier, and adjustments to reduce noise sensitivity and enhance generalization in the K-Nearest Neighbor classifier(Matzavela & Alepis, 2021). The predictive application is developed using the Java WEKA API, effectively uncovering hidden patterns in the training dataset. These derived prediction patterns are integrated into a functional application designed to predict student success, showcasing the study's contribution to improving dropout prediction in IT higher education. Overall, this research demonstrates the potential of classification algorithms to identify at-risk students, facilitating targeted interventions to enhance student retention and success in tertiary level IT courses.

3. RESULTS AND DISCUSSION

This study generated the following four classification models: Decision Tree-J48, Naïve Bayes, K-Nearest Neighbor-Lazy-IBK, and Rule-Based-ZeroR. Testing was then done on a dataset with a number of 2000 instances divided into several subsets with 500, 600, 700, 800, 900, and 1000 instances to test changes in accuracy as the instances increase for each of the selected algorithms: Naïve Bayes, Lazy-IBK, J48, and ZeroR. The study compares four classifiers for predicting dropouts in ICT higher education, focusing on accuracy, precision, and recall. Both the Decision Tree (J48) and K-Nearest (lazy-IBK) classifiers achieved perfect scores (100%) in accuracy, precision, and recall, outperforming Naïve Bayes and ZeroR, the latter scoring lowest. Key factors influencing dropout include income, demographics, institutional support, psychological aspects (like self-efficacy and motivation), and social integration (peer support and cultural fit). J48 stands out as the most effective model due to its user-friendly, interpretable structure, making it ideal for accurately assessing dropout risks (Burgos et al., 2018). The Confusion Matrix in Table 1 illustrates the performance of each classifier, detailing the numbers of true positives, false positives, true negatives, and false negatives encountered during evaluation.

Table 1. Confusion Matrix for Model Evaluation

Classifier	True Positives (TP)	False Positives (FP)	True Negatives (TN)	False Negatives (FN)
J48	100	0	0	0
lazy-IBK	100	0	0	0
Naïve Bayes	91	8	27	5
ZeroR	0	0	83	17

4. CONCLUSION

This study effectively uses the Decision Tree method to predict student dropouts in ICT higher education, achieving 100% accuracy in identifying key relationships between dropout factors and prediction outcomes. By training classification techniques like Decision Tree, K-Nearest Neighbor, Naïve Bayes, and Rule-Based models, the research demonstrates over 83% accuracy, with J48 and K-Nearest Neighbor performing best. Critical dropout factors include academic performance, prior ICT experience, socio-economic status, motivation, and institutional characteristics. A proposed predictive tool based on these findings could help educators assess dropout risks at admission, enabling targeted interventions. However, the study’s reliance on a single institutional dataset may limit its generalizability to other ICT contexts. Future improvements could include refining predictors with advanced feature selection and applying ensemble methods, such as Random Forests, to enhance model resilience. Further research into association and rule mining may uncover additional patterns, broadening dropout intervention strategies for diverse educational settings.

5. REFERENCES

Jalota, C., & Agrawal, R. (2019). Analysis of Educational Data Mining using Classification. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 243–247. https://doi.org/10.1109/COMITCon.2019.8862214

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland)*, *9*(15). https://doi.org/10.3390/app9153093

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. *Computers and Education: Artificial Intelligence*, *2*, 100035. https://doi.org/10.1016/j.caeai.2021.100035

Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., & Casadesús-Fa, M. (2020). University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innovations in Education and Teaching International*, *57*(1), 74–85. https://doi.org/10.1080/14703297.2018.1502090

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z