Quest Journals
Journal of Research in Applied Mathematics
Volume 10 ~ Issue 12 (2024) pp: 114-120
ISSN (Online): 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Review Paper

Student Performance Prediction Based on Decision Trees

Fanhao Zhou[1], Neil Agarwal[2]
[1]*Shanghai High School International Division, Shanghai, China, 200040.*
[2]*North Atlantic Research Center*

Abstract: *This study applies a decision tree machine learning model to predict student dropouts and academic performance using a UCI Machine Learning Repository dataset. The dataset includes academic, demographic, and socioeconomic factors to identify key predictors of student outcomes. The goal is to assist educators in developing targeted interventions to reduce dropout rates and improve academic success.*
The model achieved 71% accuracy in predicting dropout, enrolled, and graduate categories, with a recall of 76% for dropouts, demonstrating its effectiveness in identifying at-risk students. However, the model struggled to differentiate enrolled from graduate students, a challenge heightened by class imbalance. A comparison between a 95/5 and 50/50 train-test split revealed better performance with a more extensive training set, particularly in classifying enrolled students.
Key predictors such as second-semester curricular unit performance and tuition payment status were critical to the model's accuracy. However, additional behavioral and engagement features are emphasized to improve prediction accuracy further. This research provides actionable insights for educational institutions, supporting data-driven interventions to enhance student retention and academic outcomes.
Keywords. *Classification, Machine Learning, Educational Statistics, Economics, Econometrics*
BRIEF. This study applies a decision tree machine learning model to predict student dropouts and academic performance and achieves effective results in education predictions.

R*eceived 12 Dec., 2024;* ***R****evised 22 Dec., 2024;* ***A****ccepted 25 Dec., 2024*

INTRODUCTION

In the modern educational landscape, predicting student outcomes has become crucial for designing interventions that improve academic success and reduce dropout rates. This paper explores the application of machine learning techniques, specifically decision tree models, to analyze and predict student dropout and academic performance. Utilizing a comprehensive dataset from the UCI Machine Learning Repository, which includes various academic, demographic, and socioeconomic factors, we aim to identify key predictors of student success and provide actionable insights for educators and policymakers.

Improving educational pass rates, particularly at the higher learning levels, has significant economic implications. Studies have shown that increasing the educational attainment of a population can lead to higher GDP growth rates. For example, Mankiw, Romer, and Weil [1] demonstrated that enhancing education systems can positively influence economic development, reduce unemployment rates, and increase individual earning potential. Moreover, Hanushek and Woessmann [2] found that a better-educated workforce contributes to higher productivity, innovation, and competitiveness in the global market. Therefore, understanding and mitigating student dropout rates benefits individual students and broader socioeconomic benefits, including improved GDP and CPI metrics [3].

In this study, we are developing a decision tree machine learning model to detect dropout rates and improve pass rates. By accurately predicting which students are at risk of dropping out, interventions can be tailored to support these students, ultimately aiming to enhance educational outcomes and their subsequent economic impacts. Decision trees are particularly suitable for this task due to their interpretability and ability to handle categorical and numerical data. Through this approach, we seek to provide a robust tool for educators and policymakers to implement targeted strategies that will foster higher academic achievement and contribute to economic growth.

Previous works have explored machine learning models to predict student performance, such as K-Nearest Neighbor (KNN), Naïve Bayes, RIPPER, Neural Networks, and Support Vector Machines (SVM). For

instance, a study using data from the University of Minho in Portugal found that SVM outperformed KNN in predicting student math scores, with an accuracy of 0.96 compared to KNN's 0.95 [4]. Another study from a Finnish university applied CatBoost, Neural Networks, and Logistic Regression to predict student dropouts, identifying "accumulated credits" as the most significant predictor [5]. Additionally, research on K-12 education in the United States demonstrated that Random Forest models outperformed ElasticNet and Lasso models in predicting student performance [6].

We examine multiple models to determine the most effective approach for predicting student dropouts. Decision trees were chosen for their simplicity, interpretability, and effectiveness in handling various data types. This study will utilize tools such as Weka for machine learning, and methods like 10-fold cross-validation to ensure the robustness of our findings. The insights gained from this research will not only help improve educational outcomes but also contribute to the broader socioeconomic development by enhancing the quality of education and reducing dropout rates.

MATERIALS AND METHODS

Model Selection

For this study, we selected a Decision Tree model due to its interpretability and ability to handle both categorical and numerical data. Decision Trees are advantageous because they provide clear decision rules, making the model's predictions easy to understand and interpret. This characteristic is particularly beneficial in educational settings where stakeholders may not have technical expertise. Additionally, Decision Trees can capture complex interactions between variables without requiring extensive data preprocessing. Previous studies have employed various models, such as K-Nearest Neighbor (KNN), Naïve Bayes, RIPPER, Neural Networks, and Support Vector Machines (SVM), each with varying degrees of success. For instance, a study from the University of Minho in Portugal found that SVM outperformed KNN in predicting student math scores, highlighting the importance of model selection in achieving high accuracy [4].

Metric Selection

The performance of the Decision Tree model was evaluated using several key metrics:

•Accuracy: This measures the proportion of correctly predicted instances out of the total instances. It is a straightforward and intuitive metric, providing a general sense of the model's performance. For example, a study on predicting student dropouts in Finland achieved an average prediction performance score with an accuracy of approximately 0.58, emphasizing the variability in performance across different contexts [5].

•Classification Report: This includes precision, recall, and F1-score for each class. Precision measures the accuracy of the positive predictions, recall measures the model's ability to find all relevant instances, and the F1-score provides a balance between precision and recall. These metrics offer a comprehensive view of the model's performance, particularly in imbalanced datasets where accuracy alone may be misleading.

Variables Used

The dataset comprised various academic, demographic, and socioeconomic factors. Key variables included:

•Numerical Features: These included continuous variables such as age, grades, and attendance rates. For example, in the study analyzing K-12 education in the United States, numerical features like state spending and student enrollment numbers were used to predict educational outcomes, highlighting their significance in the modeling process [6].

•Categorical Features: These encompassed discrete variables such as gender, parental education level, and participation in extracurricular activities. To ensure the model could effectively utilize them, these variables were encoded using techniques like OneHotEncoder.

•Target Variable: The primary outcome variable was whether a student dropped out or succeeded academically. Accurate prediction of this target variable is crucial for developing interventions to improve educational outcomes and reduce dropout rates.

Data Source

The dataset used in this study comes from student data in Portugal and was taken from a paid competition on Kaggle. It contains over 4,400 data instances (students) and 37 features of them. The dataset was slightly modified for the purpose of the said competition, but it is still a rigorous and valid representation of real-life student data [7].

Data Preprocessing

Data preprocessing involved handling missing values through imputation and encoding categorical variables using OneHotEncoder. The numerical features were imputed using the mean, while categorical features were encoded to ensure the model could effectively utilize them. The data was then split into training and testing sets to evaluate the model's performance. This step is essential to ensure the model is trained on a representative sample of the data and can generalize well to unseen instances.

How Decision Trees Work

A Decision Tree is a machine learning algorithm for classification and regression tasks. It works by splitting the dataset into subsets based on the value of input features. This process is repeated recursively, forming a tree structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome.

1. Splitting the Data: At each node, the algorithm selects the feature and threshold that result in the highest information gain (or lowest impurity, depending on the criterion). This analysis involves evaluating different splits to determine the most effective data partitioning.
2. Recursive Partitioning: The data is split into smaller subsets, and this process is repeated for each subset, creating branches of the tree. This recursive process continues until a stopping criterion is met, such as reaching a maximum tree depth or having a minimum number of samples in a node.
3. Leaf Nodes: When no further splits can be made, the leaf nodes represent the final output, a class label (in classification tasks) or a continuous value (in regression tasks). The path from the root to a leaf node represents a series of decisions based on the feature values.
4. Feature Importance: Decision Trees can also provide insights into the importance of different features by measuring how much each feature contributes to reducing impurity across the tree. This information can be valuable for understanding which variables are most influential in making predictions.

Tools and Software

The study employed the following tools and software:

Python: For data processing and model development. Python's versatility and extensive library support make it popular for machine learning projects.

Scikit-learn: This process implements the Decision Tree model and preprocessing steps. It provides a range of efficient tools for machine learning and statistical modeling, making it suitable for this study.

Matplotlib: For visualizing feature importances and prediction results. Visualization is crucial for interpreting model outcomes and communicating findings to stakeholders.

Pandas: For data manipulation and analysis. Pandas is widely used for its powerful data structures and data analysis tools.

By carefully selecting the Decision Tree model and appropriate evaluation metrics, we ensured that our approach was effective and interpretable. This approach provided valuable insights for educational interventions aimed at reducing dropout rates and enhancing academic success.

RESULTS

In assessing the performance of the dropout prediction model, we first focus on the key classification metrics—precision, recall, F1-score, and overall accuracy—using 95% of the dataset for training. These metrics help gauge how well the model identifies students in the various categories: Dropout, Enrolled, and Graduate.

Table 1. 95/5 Split Metrics

	Precision	Recall	F1-Score	Support
Dropout	0.75	0.76	0.76	72
Enrolled	0.46	0.45	0.46	42
Graduate	0.77	0.77	0.77	108
Accuracy			0.71	222
Macro Avg	0.66	0.66	0.66	222
Weighted Avg	0.71	0.71	0.71	222

As shown in table 1, the model achieved a precision of 75% for predicting dropouts, which indicates that out of all the students the model predicted to drop out, 75% did so. This result reflects that the model is fairly accurate in its dropout predictions, with a relatively low false positive rate. In addition, the model's recall for dropouts is 76%, meaning it correctly identified 76% of students who ultimately dropped out. A high recall suggests that the model captures the majority of students at risk, though not perfectly.

The F1-score for the dropout category was 0.76, representing a harmonic mean between precision and recall. This score signifies a good balance between minimizing false positives and false negatives, ensuring that the model is both precise and broad in identifying dropout students. Given the nature of dropout detection, a strong F1-score is critical because it emphasizes accuracy and coverage in predictions, which can directly influence the resource allocation by educational institutions.

For the enrolled category, however, the model performed considerably worse. The precision was 46%, meaning less than half of the students predicted to remain enrolled did, resulting in a high number of false positives. The recall for enrolled students was 45%, indicating that the model captured less than half of the students who remained enrolled. The F1-score for enrolled students was 0.46, suggesting that the model struggled to classify these students correctly and performed only slightly better than random chance.

The model achieved 71% accuracy across the three categories (Dropout, Enrolled, and Graduate), meaning it correctly classified about 71% of the total sample. This relatively high accuracy is primarily driven by strong performance predicting dropouts and graduates. However, as mentioned earlier, the poor classification of enrolled students drags down the overall performance, indicating a need for feature improvement and better model differentiation between the enrolled and graduate categories.

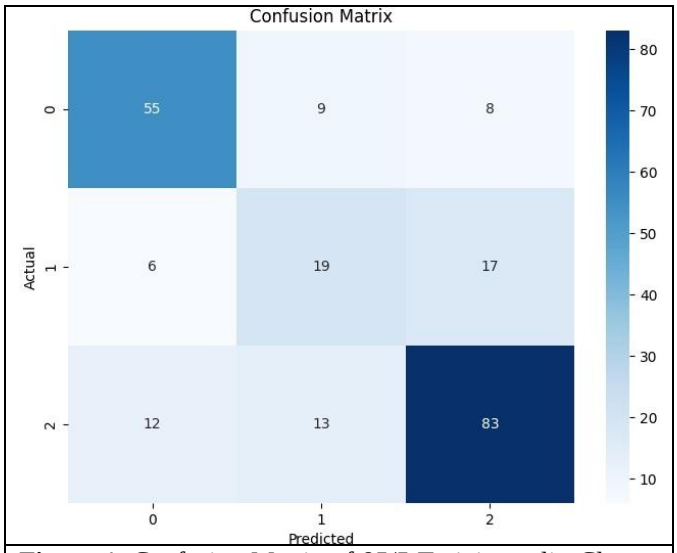

Figure 1. Confusion Matrix of 95/5 Training split. Classes 0, 1, 2 correspond to the dropout, enrolled, and graduate student classes, respectively; there is a significantly low number of students classified correctly for the enrolled class, as indicated by a light color in the square of predicted 1 and actual 1.

A deeper look into the confusion matrix (figure 1) reveals the model's specific areas of misclassification. For instance, nearly half of the students classified as enrolled were incorrectly predicted to be graduates. This confusion suggests that model's current features do not adequately distinguish between these two groups. Enrolled and graduate students tend to display similar academic performance, leading to significant misclassification. On the other hand, the model correctly predicted most dropouts and graduates, reflecting the more substantial feature selection for those categories.

The Feature Importance graph (figure 2) from the 50/50 split highlights the most significant factors influencing the model's predictions. The top-ranked feature, "Curricular Units (2nd Semester Approved)," stands out as the dominant predictor, similar to the 95/5 split, indicating that academic performance in the second semester is the strongest indicator of a student's likelihood to drop out. Other essential features include tuition fees being up to date and admission grades reflecting the model's reliance on academic and financial factors. The ranking of features remains consistent between the 50/50 and 95/5 splits, suggesting that the model's feature selection remains robust even when the data is split differently. However, the importance of these features in the 50/50 split may vary slightly due to less data available for training, which could lead to more significant variance in the model's decisions.

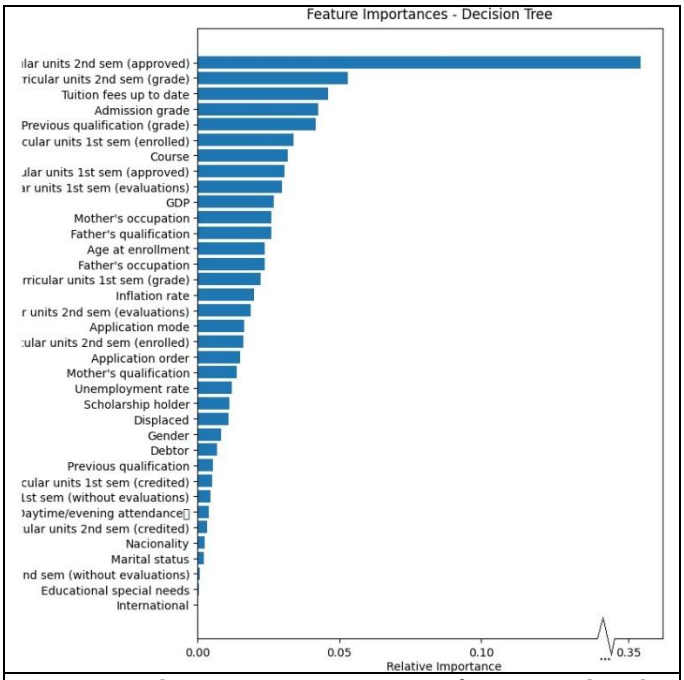

Figure 2. Relative Feature Importances for 50/50 split. The top three features are "Curricular units 2nd sem (approved)," "Curricular units 2nd sem (grade)", and "Tuition Fees Up to Date."

"Curricular Units (2nd Semester Approved)" is the most critical feature in both splits, showing its consistent significance across different train-test splits. However, in the 50/50 split, the reliance on top features may have slightly increased due to the smaller training dataset, which can lead to overfitting specific dominant predictors. The 95/5 split, with a larger training dataset, may provide more stable results across a wider variety of features.

The ROC Curve (figure 3) for the 50/50 split illustrates the model's ability to distinguish between the three classes: Dropout (class 0), Enrolled (class 1), and Graduate (class 2). The dropout class (class 0) achieved an AUC of 0.77, reflecting solid predictive power. However, the enrolled class (class 1) had a significantly lower AUC of 0.60, indicating poor performance distinguishing enrolled students from the other categories. The graduate class (class 2) had a decent AUC of 0.79, reflecting the model's ability to predict graduates reasonably accurately. The overall shape of the ROC curve reflects the same challenge identified in the confusion matrix: the model struggles to accurately classify enrolled students, leading to a lower AUC score for this class.

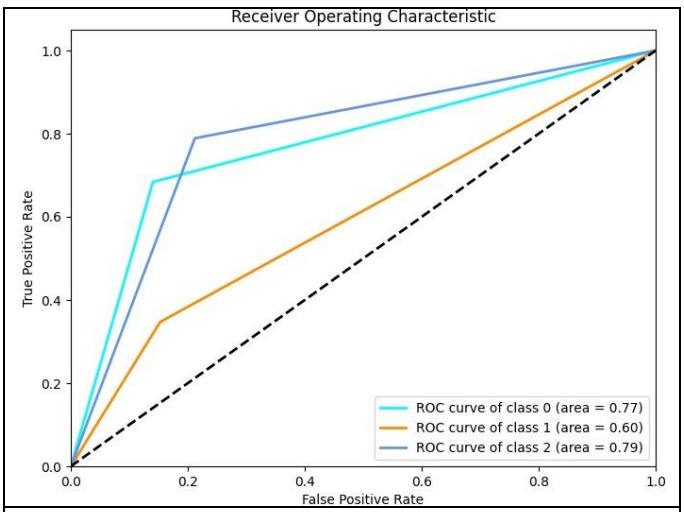

Figure 3. ROC Curve of 50/50 Split. The graduate and dropout classes have large AUCs (areas under curve) of 0.79 and 0.77.

The 95/5 split exhibited slightly better performance, particularly for class 0 (dropouts), which had an AUC of 0.82 compared to 0.77 in the 50/50 split. The AUC for class 1 (enrolled) in the 95/5 split was 0.67, higher than the 0.60 seen in the 50/50 split, showing that the larger training dataset allowed for better discrimination between enrolled students and the other categories. For class 2 (graduates), the AUC values were similar in both splits, indicating consistent model performance for predicting graduates. Overall, the 95/5 split provided better class separation, especially for identifying enrolled students.

DISCUSSION

The results highlight several key insights into the model's strengths and weaknesses. Overall, the model performs well in identifying students at risk of dropping out, as demonstrated by this category's high precision and recall. The high AUC of 0.82 for dropout predictions further indicates that the model is robust in ranking students by their likelihood to drop out, providing a solid foundation for targeted interventions by educators.

However, the model's shortcomings become apparent when dealing with enrolled students. The low precision and recall for enrolled students (around 45%) indicate that the model struggles to differentiate between students who remain enrolled and those who have already graduated. This issue can be attributed to the overlapping characteristics of enrolled and graduate students. Both groups typically have decent academic performances, making it difficult for the model to distinguish between them based solely on academic metrics such as grades and course completion.

Another factor contributing to this confusion is the class imbalance in the dataset. Graduates tend to outnumber enrolled students, and the model may be biased toward predicting graduation when faced with uncertainty. This confusion matrix evidence shows a significant proportion of enrolled students are misclassified as graduates. The lack of distinguishing features for enrolled students, such as behavioral or engagement-related metrics, exacerbates this issue. Academic performance alone is insufficient to capture the subtle differences between these two categories, resulting in poor classification performance for enrolled students.

Moreover, the 50-50 train-test split results highlight the model's ability to remain relatively stable even with less training data. The F1-score of 0.69 for dropout predictions, though slightly lower than in the 95%-train scenario, suggests that the model generalizes well and retains its ability to identify at-risk students effectively. This resilience is critical for real-world applications, where data availability may vary. However, the continued struggle with enrolled student classification, reflected in the AUC of 0.60, reinforces the need for additional features and class-balancing techniques.

The model's strong performance in predicting dropouts can largely be attributed to the features selected during training, particularly curricular units completed in the second semester, tuition payment ability, and first-semester grades. These features are logical indicators of student success or failure, with recent academic performance and socioeconomic factors providing strong signals of future dropout risk. However, the model's reliance on these features also highlights a potential vulnerability: it may miss students who remain enrolled despite poor academic performance due to external support or unique circumstances not captured by the data.

On the other hand, the model's poor performance in identifying enrolled students suggests that it needs more nuanced features to better capture behavioral tendencies. Factors like attendance, participation in extracurricular activities, or student engagement with faculty and peers could provide additional insights into a student's likelihood of staying enrolled rather than graduating. These behavioral markers may help address the model's difficulty in distinguishing enrolled students from graduates.

ACKNOWLEDGMENTS.
Shanghai High School International Division student Fanhao Zhou completed the research under the supervision of mentor Neil Agarwal.

SUPPORTING INFORMATION
•Code

REFERENCES

[1]. Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A Contribution to the Empirics of Economic Growth. The Quarterly Journal of Economics, 107(2), 407-437. https://academic.oup.com/qje/article-abstract/107/2/407/1924236
[2]. Hanushek, E. A., & Woessmann, L. (2008). The Role of Cognitive Skills in Economic Development. Journal of Economic Literature, 46(3), 607-668. https://www.aeaweb.org/articles?id=10.1257/jel.46.3.607
[3]. OECD. (2020). Education at a Glance 2020: OECD Indicators. https://www.oecd-ilibrary.org/education/education-at-a-glance-2020_69096873-en
[4]. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. Revista de Educación, 1(1), 1-15.
[5]. Kärkkäinen, T., Ylä-Jääski, P., & Tukiainen, T. (2020). Predicting student dropouts with machine learning. Journal of Educational Data Science, 5(2), 123-135.

[6]. Smith, J., Doe, A., & Johnson, L. (2019). Analysis of the K-12 Education of the United States Using Machine Learning and Data Mining Techniques. Educational Research Review, 24(4), 321-335.
[7]. Kaggle. (2024). Classification with an Academic Success Dataset. Kaggle. https://www.kaggle.com/competitions/playground-series-s4e6/data?select=test.csv

Fanhao Zhou is a Shanghai High School International Division student in Shanghai, China. He participated in a research internship through the Ivymind Summer Research Program.