Predicting Student Success

A Data Mining Project

Data Science Department
University of Colorado Boulder

ABSTRACT

This project explores the application of data mining and machine learning techniques to predict student dropout rates and academic success. By analyzing this dataset from the UCI Machine Learning Repository, this project aimed to identify key factors that influence student outcomes and develop predictive models. The results show that a model like Random Forest outperforms others, highlighting the importance of more specific features in determining student success. This research has significant implications for educational institutions seeking to improve retention rates and graduation rates, as well as tailoring interventions for students at risk of dropping out.

1. Introduction

Student dropout rates pose quite a challenge to educational institutions as it impacts the futures of students outside of school as well as affecting institutional resources. This study uses data mining techniques to analyze factors contributing to dropout rates and academic success. The findings aim to provide actionable insight for educators, school administration, and policymakers, using predictive analytics to enhance student support.

2. Data Collection and Preprocessing

2.1 Data Sources

The data set used for this project is from the UC Irvine Machine Learning Repository, entitled “Predict Students’ Dropout and Academic Success.” It was created from a Portuguese higher education institution to study factors contributing to student success. It includes various features such as students’ demographic information, academic performance, and social factors that may influence the likelihood of graduating or dropping out.

2.2 Data Preprocessing Steps

Data preprocessing involved several steps to prepare the dataset for analysis:

- Handling Missing Values: Identified and filled or removed missing entries as appropriate
- Encoding Categorical Variables: Converted categorical features into numerical format using one-hot-encoding and mapping
- Determining Predictor Relevance: Utilized the Chi-Squared Test to identify columns of statistical significance (p-value < 0.05) to keep for modeling
- Correlation Matrix: Calculated and visualized correlation matrix to see features with high correlation for removal

3. Methodology

3.1 Models

Two machine learning algorithms were employed. Decision trees utilize a tree-like structure to model decisions and their consequences. Each internal node represents a feature (attribute), each branch represents a decision rule, and each leaf node represents an outcome or class label. Random forest is an ensemble method that constructs multiple decision trees and combines their outputs to improve metrics like accuracy and reduce overfitting.

3.2 Hyperparameter Tuning

Hyperparameter tuning was done with Grid Search (GridSearchCV) to optimize model parameters. These parameters were systematically tested for each model in order to identify the combination that yields the best performance.

3.3 Cross Validation

K-fold cross-validation was also done in conjunction with Grid Search (GridSearchCV) to ensure model robustness. It allows for the assessment of model performance across different subsets of the dataset, aiming to reduce overfitting.

4. Evaluation and Results

4.1 Decision Trees Performance

Metric	Value
Accuracy	0.6845
Precision	0.6206
Recall	0.5717
F1 Score	0.5774

Figure 1: Table showing performance metrics for Decision Trees

Classification Report				
	Precision	Recall	F1 Score	Support
0: Dropout	0.80	0.62	0.70	472
1: Graduated	0.68	0.90	0.78	663
2: Enrolled	0.37	0.19	0.25	238
Accuracy			0.68	1328
Macro avg	0.62	0.57	0.58	1328
Weighted avg	0.67	0.68	0.66	1328

Figure 2: Table showing the classification report for Decision Trees

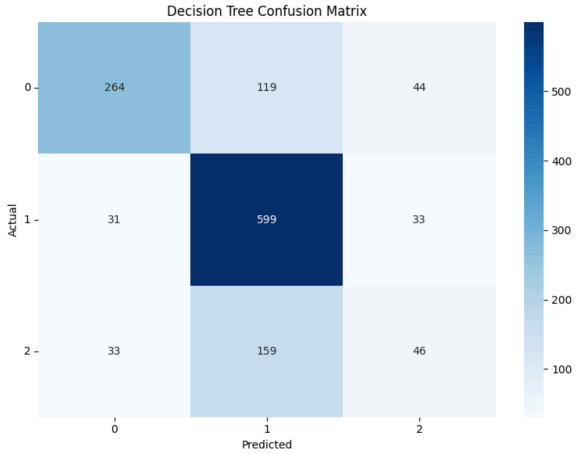

Figure 3: Decision Tree confusion matrix

Overall, the Decision Trees model achieved an accuracy of 68.45%, precision of 62.06%, recall of 57.17%, and F1 score of 57.74%. By also looking at the classification report, it is evident that the model struggles to correctly identify currently enrolled students and its effects are seen in overall model performance. Precision for Dropout students (Class 0) is decently high at 80%, meaning the model is correct 80% of the time for this group of students. For Graduated students (Class 1), the recall is 90%, meaning the model excels at identifying graduates and captures 90% of actual graduates.

The lower performance for Enrolled students (Class 2) is a significant drawback and negatively affects performance metrics. This may indicate that further turning, feature engineering, or even more data collection can help improve prediction of this important class.

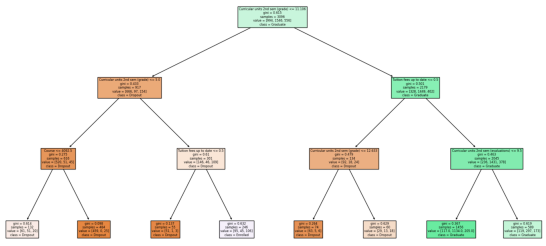

Figure 4: Visualization of the created decision tree

By looking at the decision tree itself, we can see the features that contribute to the classification process of

whether a student drops out or graduates. They are as follows:

- Curricular units 2nd sem (grade): Grade average in the 2nd semester (between 0 and 20)
- Curricular units 2nd sem (evaluations): Number of evaluations to curricular units in the 2nd semester
- Course: Course number identifier
- Tuition fees up to date: Student's tuition fees are paid in full

4.2 Random Forest Performance

Metric	Value
Accuracy	0.7078
Precision	0.6197
Recall	0.5866
F1 Score	0.5790

Figure 5: Table showing performance metrics for Random Forest

Classification Report				
	Precision	**Recall**	**F1 Score**	**Support**
0: Dropout	0.76	0.72	0.74	472
1: Graduated	0.72	0.91	0.80	663
2: Enrolled	0.39	0.13	0.20	238
Accuracy			0.71	1328
Macro avg	0.62	0.59	0.58	1328
Weighted avg	0.67	0.71	0.67	1328

Figure 6: Table showing the classification report for Random Forest

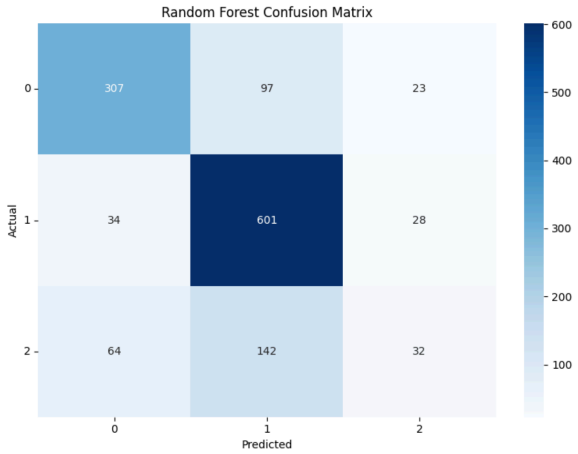

Figure 6: Random Forest confusion matrix

Overall, the Random Forest model achieved an accuracy of 70.78%, precision of 61.97%, recall of 58.66%, and F1 score of 57.90%. By looking at the classification report, this model struggles to correctly identify currently enrolled students (similar to the Decision Trees Model) and it affects overall model performance. The Random Forest model predicts Dropouts (Class 0) with decent precision of 76% and excels at capturing graduates with a recall of 91% percent.

Like the Decision Trees model, the Random Forest model's poor capturing of the Enrolled (Class 2) category poses a significant drawback. This signifies that further model refinement or the addition of more features could be beneficial to improve predictions for the Enrolled class.

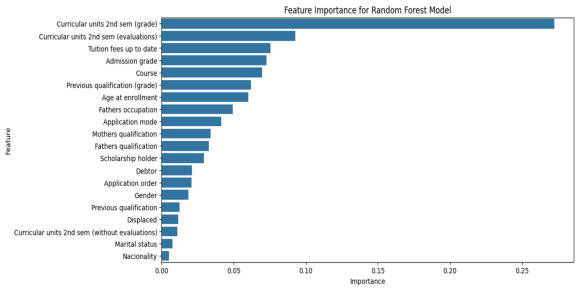

Figure 7: Bar graph of feature importance

Above is a graph generated to calculate and show feature importance for the random forest model. A list of the top five most important features are as follows:

- Curricular units 2nd sem (grade): Grade average in the 2nd semester (between 0 and 20)

- Curricular units 2nd sem (evaluations): Number of evaluations to curricular units in the 2nd semester
- Tuition fees up to date: Student's tuition fees are paid in full
- Admission grade: Admission grade (between 0 and 200)
- Course: Course number identifier

5. Model Interpretation

The analysis of feature importance for both models revealed the variables that significantly contribute to predicting student outcomes such as Curricular Units 2nd Sem (Grade), Curricular Units 2nd Sem (Evaluations), Tuition Fees Up to Date, Admission Grade, and Course. This suggests that academic performance, financial stability, and the type of course in which a student is enrolled influences their likelihood of graduating or dropping out.

5.1 Explanation of Important Features

5.1.1 Curricular Units 2nd Sem (Grade)

This feature represents academic performance in the second semester. High importance suggests that a student's grade during this period strongly predicts their overall success or failure. Poor academic performance could correlate with a higher likelihood of dropping out.

5.1.2 Curricular Units 2nd Sem (Evaluations)

Similar to grades, this feature represents the number of evaluations given to a student throughout the semester (like exams, projects, or assessments). A higher number of evaluations can reflect the academic engagement of a student and shows how often they receive feedback. A high number of evaluations can lead to more frequent assessment of student performance and understanding of the content. This can help put in place timely interventions for those that struggle. This feature is necessary for identifying at-risk students and improving their learning outcomes.

5.1.3 Tuition Fees Up to Date

This feature indicates whether students are current on their tuition payments and sheds light on their financial stability. Being up to date with tuition feels could correlate with a student's ability to focus on their studies without being stressed about financial issues. It could also signify stresses at home as well with parental and financial support. If a student's tuition is not paid, they are forced to drop out, possibly until the next academic session. Such a delay could cause students to not want to return to school. Focusing on this feature shows that educational institutions should offer ways to support students facing financial challenges.

5.1.4 Admission Grade

Similar to the earlier mentioned feature of grades, this feature of admission grade reflects the initial academic ability of the student prior to enrollment. It can suggest that students that enter this school with higher grades are more likely to academically succeed, stay enrolled, and eventually graduate. It also indicates that initial academic ability can predict future performance and retention. This signifies to institutions that a selective admission process can be vital to promoting student success.

5.1.5 Course

The courses taken by students can significantly impact their academic experience. Some courses may have varying levels of difficulty, support structures, teachers, evaluations, and types of engagement. This means that certain courses could have higher dropout rates or graduation rates based on those factors. Course importance implies that certain courses may be better at promoting student success and retention due to their curriculum development or resource allocation.

5.2 Implications of Feature Performance

5.2.1 Targeted Interventions

An understanding of academic performance and student evaluations can lead to more focused academic initiatives designed to support student success. Programs that implement things like tutoring or mentoring could especially help lower-performing students excel.

5.2.2 Financial Aid Policies

Recognizing the importance of tuition payment status can guide educational institutions to strengthen financial aid programs and better allocation of

financial resources. This can ensure that financial barriers do not hinder student academic success.

5.2.3 Admission Strategies

An emphasis on student grades at time of admission suggests that rigorous admission standards may contribute to higher graduation rates. It can lead to changes in the admission process and criteria, as well as also figure out support strategies for incoming students that struggle academically.

5.2.4 Curriculum Development

Further insight into specific courses can better shape curriculum design. This can allow educational institutions to enhance programs that are already yielding great student outcomes and improve those that do not.

6. Conclusion

6.1 Overall Findings

This project's aim was to use data mining to analyze student outcomes and use machine learning techniques such as Decision Trees and Random Forest to identify certain attributes that attribute highly to student dropout and graduation rates. The results indicated that both models provided valuable insight, but the Random Forest model outperformed Decision Tree with its overall accuracy and robustness. Key features were identified that affect student outcomes, emphasizing the importance of academic performance and financial stability. These insights can help educational institutions create targeted interventions to support students while improving retention and graduation rates.

6.2 Limitations and Future Improvements

Both models were limited mainly because of an imbalance in class labels. There was not enough data for currently enrolled students, therefore both models had a hard time predicting for it. Both models tended to fit the majority classes of Graduate and Dropout, perhaps overfitting for them and not predicting strongly enough for Enrolled. This could be fixed in the future with more collected data or some techniques that help address class imbalance.

REFERENCES

[1] Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. 2021. Predict Students' Dropout and Academic Success [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5MC89.

[2] GeeksforGeeks. n.d. ML | Chi-Square Test for Feature Selection. Retrieved December 10, 2024, from https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.