Will you Dropout or not?

Group 6

Adam G Johnson
Bolutife Samad Idowu
Bryan Manuel Pedroza
Jason T Quach
Katherine Dolores Hernandez

December 5th 2025

Introduction

Description of Data

This project uses the Predict Students' Dropout and Academic Success dataset from the UCI Machine Learning Repository, a real world dataset collected from a higher education institution. It contains 4,424 student records and 36 features describing demographics, academic background, socioeconomic factors, and early academic performance. The dataset was chosen because it provides a comprehensive view of student characteristics known at the time of enrollment, which makes it well suited for building models that support early intervention strategies.

The core task is framed as a three class classification problem predicting whether a student will drop out or not which aligns directly with real institutional challenges. Universities increasingly rely on data driven approaches to identify at risk students early and allocate support resources effectively. This dataset is particularly valuable because it reflects an authentic class imbalance seen in real academic settings and allows for exploration of techniques used to handle uneven class distributions.

In this project, we apply tree based classification methods, including decision trees with pruning, bagging, and random forests, to model student outcomes and assess which factors are most strongly associated with dropout, persistence, and graduation. Overall, this dataset was

selected because it combines realistic student information with a socially impactful prediction task, allowing us to use machine learning methods to better understand and potentially reduce dropout rates in higher education.

Can we accurately predict whether a student will dropout or not at the time of admission?

Dataset Variables

Variable Name	Description / Possible Values (with Type)	Variable name	Description / Possible Values (with Type)
Marital Status	*(Integer)* — 1 = Single; 2 = Married; 3 = Widower; 4 = Divorced; 5 = Facto Union; 6 = Legally Separated	**Age at Enrollment**	*(Integer)* — Age of the student at first enrollment
Application Mode	*(Integer)* — Type of admission (e.g., 1 = 1st Phase General, 5 = Special Contingent Azores, 17 = 2nd Phase General, 39 = Over 23 Years Old, 51 = Change of Institution/Course, etc.)	**International**	*(Integer / Binary)* — 1 = International student; 0 = Domestic
Application Order	*(Integer)* — Rank of the application (0 = first choice to 9 = last choice)	**Gender**	*(Integer / Binary)* — 0 = Female; 1 = Male
Course	*(Integer)* — Study program code (e.g., 33 = Biofuel Tech, 171 = Animation & Multimedia Design, 9003 = Agronomy)	**Scholarship Holder**	*(Integer / Binary)* — 1 = Yes; 0 = No
Daytime/Evening Attendance	*(Integer)* — 1 = Daytime student; 0 = Evening student	**Unemployment Rate**	*(Real)* — National/regional unemployment rate at enrollment (%)
Previous Qualification	*(Integer)* — Highest prior education (1 = Secondary, 2 = Bachelor's, 3 = Master's, 4 = Doctorate, etc.)	**Inflation Rate**	*(Real)* — Inflation rate at enrollment (%)
Previous Qualification Grade	*(Real)* — Numeric grade of prior qualification (0 – 200 scale)	**GDP**	*(Real)* — GDP value or growth rate at enrollment
Nationality	*(Integer)* — Nationality code (e.g., 1 = Portuguese, 6 = Spanish, 41 = Brazilian, etc.)	**Target (Outcome)**	*(Categorical)* — Student status: **Dropout**, **Enrolled**, or **Graduate**
Mother's Qualification	*(Integer)* — Mother's education level (same coding as Previous Qualification)	**Educational Special Needs**	*(Integer / Binary)* — 1 = Yes; 0 = No
Father's Qualification	*(Integer)* — Father's education level (same coding as Previous Qualification)	**Displaced**	*(Integer / Binary)* — 1 = Lives away from family home; 0 = Lives with family

Mother's Occupation	*(Integer)* — Occupation category code for mother	**Father's Occupation**	*(Integer)* — Occupation category code for father

Data Preprocessing

Before fitting any models, we cleaned and restructured the dataset so it would work smoothly with our tree-based methods. First, we read in the original UCI file and converted the three class outcome variables (Dropout, Enrolled, Graduate) into a binary factor, where "Dropout" is one class and "Enrolled, Graduate" combines the other two. This effectively allowed us to predict Dropout vs No Dropout. We then marked a set of demographic and background variables, such as marital status, application mode, course, nationality, the mother's education and occupation, and scholarship status as categorical so that R would treat them as factors in the models. To avoid leaking information from the future, we removed all first and second semester curricular unit variables and the billing related variables Tuition, fees, up.to.date, and Debtor, since students who drop out early would naturally have very different values in those columns, and the model could "cheat" by using that. After this, we used 80/20 train test splits with a fixed random seed as the basis for our decision tree, bagging, random forest, and boosting models, all of which were trained on this processed version of the data.

Methods & Results

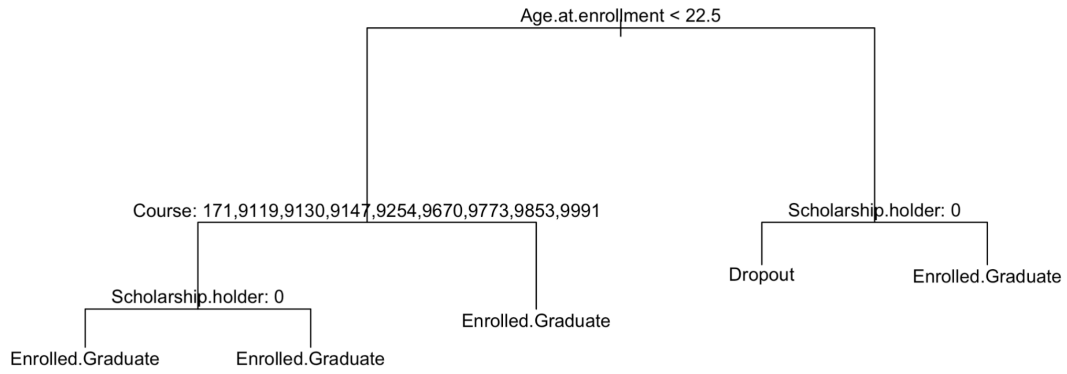
Decision Tree & Pruning (Katherine Hernandez and Jason Quach)

Decision trees are generally just more fun to use and explore! But practically, they are easy to read, follow, and interpret. They also easily find factors that have the most influence. They are also better for more complex datasets that are not better represented by a linear model. However, because of their simplistic nature, they tend to overfit their training data as a result. Luckily, with methods like pruning, we can avoid overfitting.

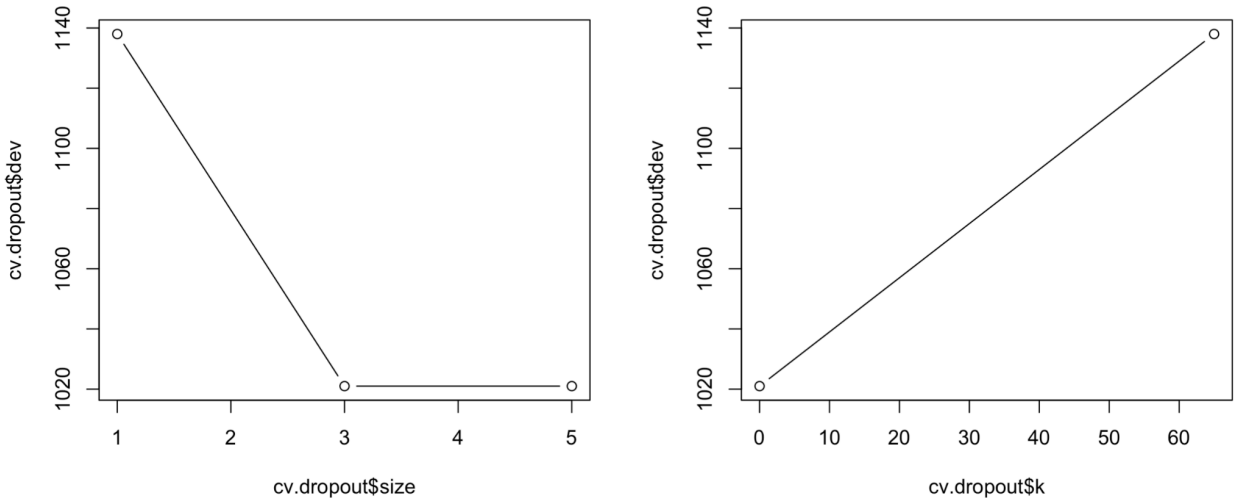
Dropout ~ X1 + X2 + X3 + … + Xp.

We also went and dropped a fair number of columns from the dataset, as some of the columns skewed the data when they had an implicit answer. For example, students who dropped out in the first semester would be ineligible to enroll in the second semester, and therefore would not have any credits. So initially, the decision tree's first split was whether the student had been approved for 2nd semester credits or not. Before, the test error rate would range from 12%-16%, and after removing such biased columns, it increased to 26%. While

making our model less accurate, it would be unfair to encourage higher accuracy on an answer that would be obvious. Below is the unpruned tree.

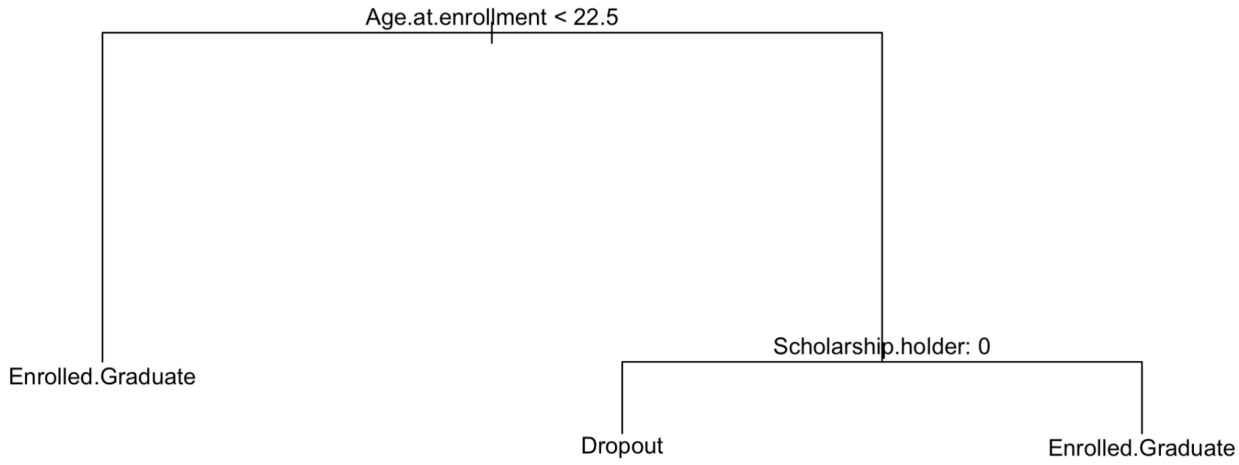


There are a total of 5 terminal trees, and only 3 variables were actually used, which are Age. at.enrollment, Course, and Scholarship.Holder. The average test error rate was 0.2829379, with the most important variable being the age at enrollment.



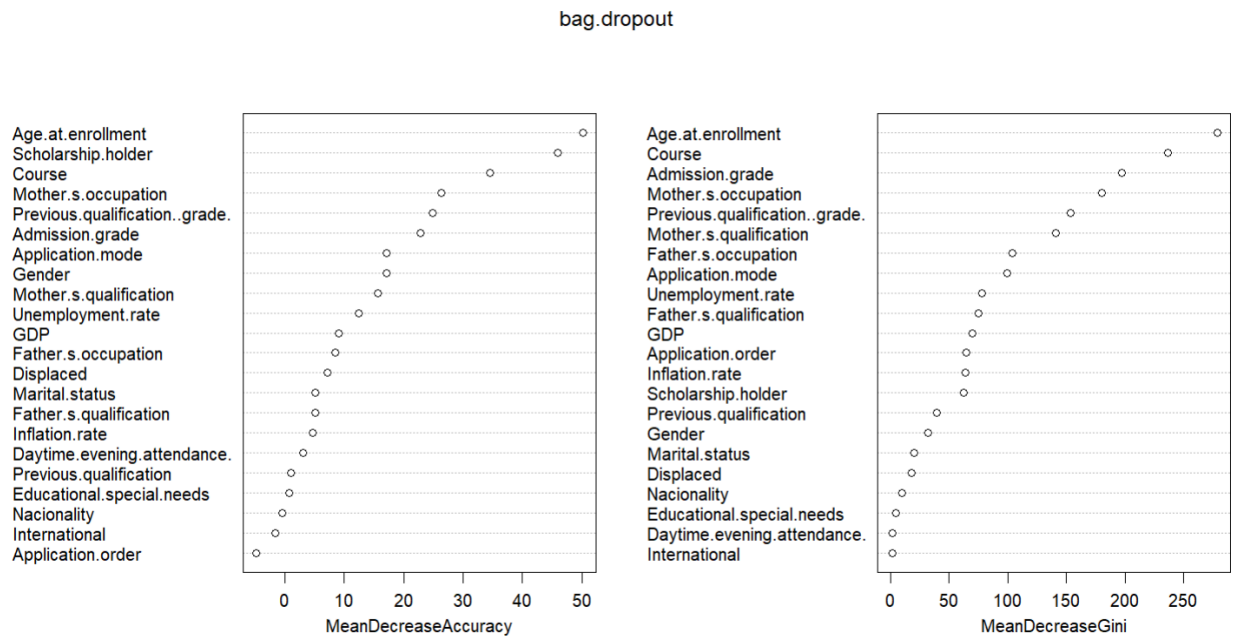
With the nature of how decision trees are built, the tree on its own is not smart enough to stop whenever it notices some repetitive splits. It is inherently greedy and can create any split with any minute difference. This is also a symptom of overfitting the data, as it will try to take care of differences that do not actually matter for our predictions. So, in our case, at node 4 handling courses, if a student is enrolled in any of those courses, it then splits into node 8 on whether they hold a scholarship or not. Regardless of whether they do or not, they will be determined to have not dropped out. Therefore, there is no real reason for there to be a

question about whether the student holds a scholarship or not, in the context of determining if they have dropped out or not. Since there were two best splits with the same deviance, 5 and 3, we go with the lower number for simplicity since the deviance doesn't significantly drop after.



Here, there are only 3 terminal nodes and an average test error rate of 0.2815819. This is actually less than the test error rate of the unpruned tree, which may be a sign that the old tree was overfitting. The only variables used in the construction of this tree are the age of enrollment and whether they are a scholarship holder or not. The pruned model tells us that there is one pathway where it is predicted that the student will drop out. If the age at enrollment is greater than 22.5, and they are not a scholarship holder, then our pruned tree predicts that they will drop out. For clarity, 0 is the indicator of no scholarship in terms of the Scholarship holder variable. So, if a student is under the age of 22.5, they have likely not dropped out. Students who are over the age of 22.5 and have a scholarship are also predicted not to have dropped out.

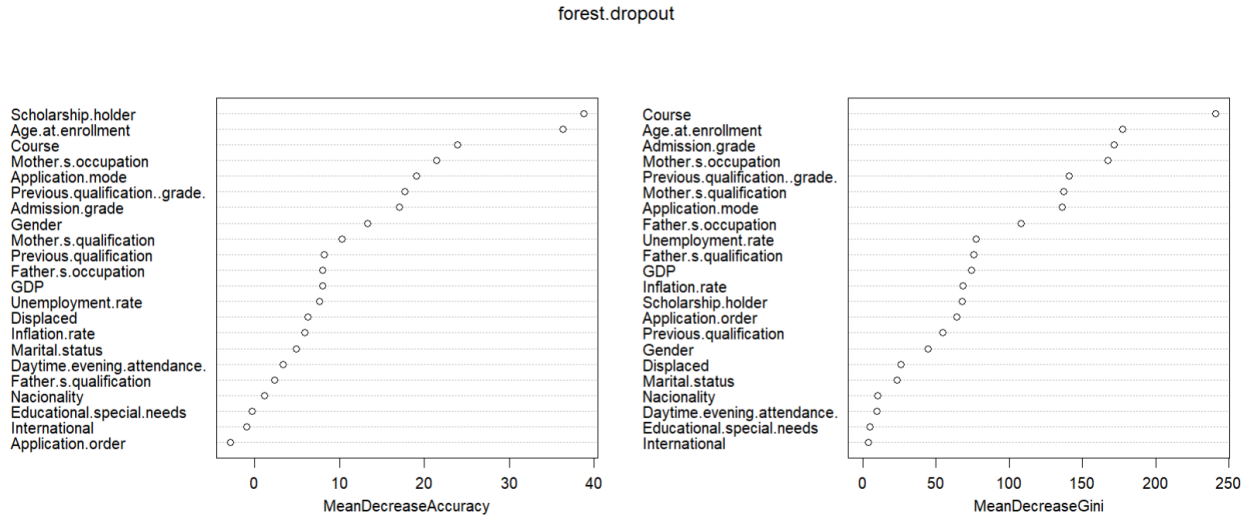
Bagging (Adam Johnson, Bryan Pedroza)



Dropout ~ X1 + X2 + X3 + … + Xp.

Bagging bootstraps many samples on the same main dataset, then fits a decision tree to each bootstrapped sample. When running the model, we run the observation through each decision tree and take an average or majority (the case here, since its classification) vote for the response. The bagging model also gave us a variable-importance plot, which helped us see which features the ensemble actually relied on when predicting dropout. In both the MeanDecreaseAccuracy and MeanDecreaseGini charts, a few variables stood out immediately. Age at enrollment was by far the most influential predictor, showing the biggest drop in accuracy when removed and the largest contribution to node purity across the trees. This matches what we saw earlier with the decision tree, but bagging reinforces that age consistently carries strong signal. After that, variables such as Course, Scholarship Holder, Admission Grade, and the parents' education/occupation levels showed up as moderately important. On the opposite end, features like Nationality, International status, Marital status, and Educational special needs contributed very little to prediction, removing them barely changed the model's accuracy. Overall, the plots confirmed that the bagged model relies primarily on a handful of academic and background-related factors, while many of the smaller demographic variables had minimal predictive value. This lines up well with what we'd expect: early academic indicators tend to say much more about a student's likelihood of dropping out than minor demographic details. Overall, bagging had an error rate of 0.2633898, which is better than the single decision trees.

Random Forest (Adam Johnson, Bryan Pedroza)



Dropout ~ X1 + X2 + X3 + … + Xp.

Random Forests is the same idea as bagging except that when we fit the decision trees, for each split we only consider p/3 or sqrt(p) (the case here since its classification) predictors. This helps to reduce bias with a bit of an increase in variance. The test confusion matrix below summarizes how well the Random Forest model classified each outcome. The model correctly predicted a large portion of the Enrolled/Graduate class, but misclassified many Dropout cases. This is expected due to the strong class imbalance in the dataset, where far fewer students drop out than remain enrolled.

RF Confusion Matrix	Dropout	Enrolled Graduate
Dropout	87	34
Enrolled Graduate	199	565

To further improve predictive performance beyond the decision tree and bagged model, we fit a Random Forest classifier using 500 trees and four randomly selected predictors at each split. The model achieves an OOB error rate of 0.2633 and an average test set error rate of

0.2625989, giving it the best overall performance among our tree based models (slightly ahead of bagging).

The Random Forest variable importance measures provide insight into which predictors most strongly influenced classification. Course was one of the most important predictors overall (MDA = 26.03, Gini = 200.04), indicating substantial variation in dropout likelihood across academic programs. Other highly influential variables include Age at enrollment (MDA = 30.96, Gini = 135.66), Scholarship holder status (MDA = 35.76, Gini = 56.88), Mother's occupation (MDA = 15.53, Gini = 130.01), Admission grade (MDA = 17.95, Gini = 110.85), and Mother's qualification (MDA = 12.46, Gini = 115.23). These results highlight the combined importance of program characteristics, academic preparedness, and family background in predicting dropout. Economic context variables such as GDP, unemployment rate, and inflation rate had moderate importance, while variables such as nationality, educational special needs, and international status showed minimal contribution.

Overall, the Random Forest model provided the strongest predictive accuracy and the clearest feature importance structure, making it the most effective model for identifying key factors related to student dropout in the analysis.

Conclusion

Across all the tree based models we applied, we were able to consistently determine the most influential factors across the differing models. Age at enrollment, course of study, and scholarship status emerged as the strongest predictors, while variables such as nationality, international status, marital status, and educational special needs had little predictive value. These patterns suggest that early academic preparation plays a much larger role in dropout risk than minor demographic characteristics, along with the difficulty associated with certain courses. This could support evidence that students who take difficult courses may be more likely to fail out of their major. Students who earn scholarships, especially merit-based, may have experience performing well academically that can be applied to complete their education.

The random forest model achieved the best accuracy and produced the clearest importance rankings, although all models had more difficulty identifying students who dropped out compared to those who did not. However, from the perspective of a model that can accurately predict if a student is enrolled/graduated, the random forest model reigned superior. With needing to mitigate the cheat code of students who had made it to the second semester, a

dataset that fairly considers these possibilities in advance could prove to create a more accurate prediction on whether students drop out or not.

Bibliography

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