The Factors that Influence a Student

to Drop Out of College

GROUP 2

Team Members:

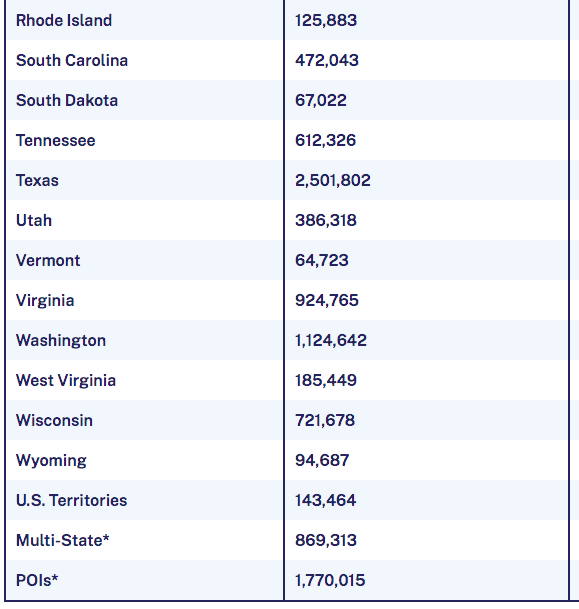
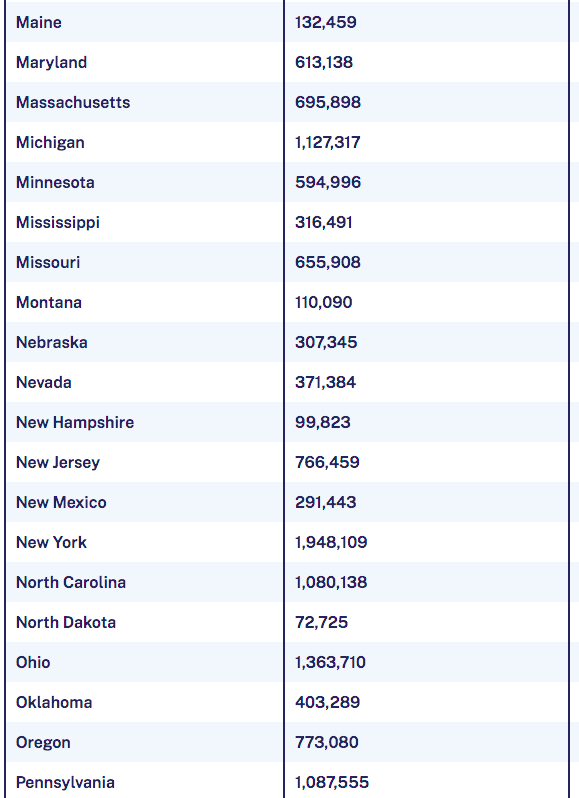
Mikaela Olaya, Sara Gassew, Hailey York, Jeff Sreca, and Alec Carpenter

Course: INFO 3237

Date Completed: 4/24/2023

# **1.0** **Executive Summary: [10 points]**

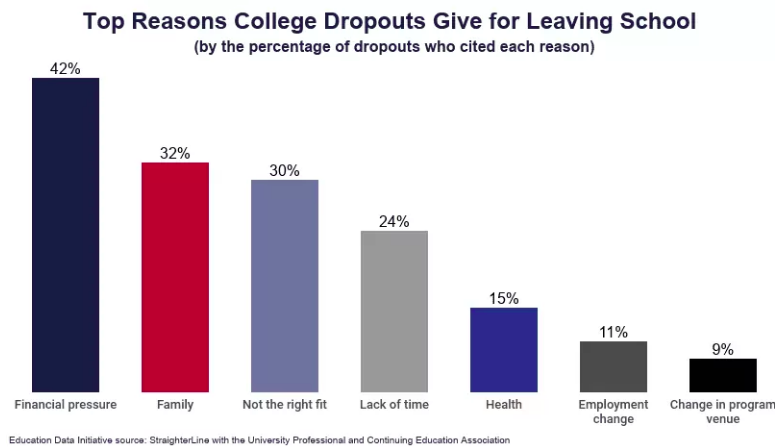
## 1.1 Business Problem Statement



Above is information found from the Education Data Initiative. In this data, Hanson highlights the total number of college dropouts per state as of July 2022. College dropout rates are an everlasting problem with a large portion of college dropouts being under the age of 35. This is a problem on a large scale because many college dropouts impact the overall economy with the increasing debt accrued that they are less likely to be able to repay due to their higher rates of unemployment.

One of the largest challenges the world faces due to college dropouts is that there are essentially less educated people within the workforce. By preventing college dropouts, the gap in skills will be reduced. College graduates also have increased potential for competitive compensation amongst employers. An increase in college graduates will benefit the economy and society as a whole as it has the potential to decrease poverty, unemployment, and reliance on other government assistance programs.

## 1.2 Business Goal



According to a 2019 article, one in three students that enroll in college courses will never earn a degree. Our data set provides the insight of students that are enrolled in undergraduate degree programs within higher education institutions. Our goal is to determine what factors most likely influenced these former students to drop out. Using this data we can determine student performance by the end of each semester using variables like enrolled, credited and their grades earned. We can use variables like unemployment rate, GDP, and inflation rate per demographic that will help us determine the economic impact that played a role in the students dropping out. We aim to be able to establish which of these factors have the largest impact so that programs can be put in place across universities in order to prevent college dropouts through additional financial or academic assistance.

## 1.3 Data Profile

We obtained the data set from Kaggle.com which is a subsidiary of Google. This is a community where data scientists use these resources and tools to help them achieve their data science goals. The website allows individuals to find or publish data sets, explore and help build certain models in a web-based environment. The dataset that we obtained will be used to help us understand why students are dropping out of college. This data uses many factors like demographic, social-economic along with overall academic performance.The data contains 4424 observations, 35 variables and does not include any missing values. The target variable included in our dataset is labeled target and includes three options: enrolled, dropout, and graduated. For the purposes of our model we will be using dropout and graduated to perform decision tree analysis.

## 1.4 Results

With using decision trees, specifically random forests, we were able to get an accuracy of 87.88% with using an ntree of 100. This means that there are 100 iterations to the random forest and that amount gave us the highest accuracy out of all the trees that were used. That same random forest has a total of 310 true positives and 647 true negatives reported, with 49 false positives and 83 false negatives. The reported sensitivity is 92.96% and a specificity of 78.88%.

**2.0** **Project Report**

## 2.1 Introduction [10 points]

The increase in students dropping out of school has been a long-standing concern in education. It not only impacts the individual student's future prospects but also affects the overall performance of an institution. In order to see change, it needs to be determined what factors increase a student's probability to drop out. Data mining can play a crucial role in identifying the factors that contribute to student dropout rates and figuring out the most effective strategies to address them.

We are determined to analyze a variety of data sources such as student demographics, academic performance, unemployment rate, and social and economic factors. Data mining can uncover patterns and correlations that can help predict which students are at the highest risk of dropping out. This information can be used to develop a multitude of help and outreach programs for these students, such as tutoring, counseling, help programs, or financial assistance.

Data mining can provide valuable insights that can improve the educational experience for students and help reduce dropout rates. By understanding the factors contributing to student dropout rates, educators can develop more effective policies and interventions that help students stay on track and achieve their academic goals.

## 2.2 Background [10 points]

Various data mining methods have previously been used to gain insight on the topic of student academic performance and dropout rates. Everything from logistic regressions to hybrid deep learning algorithms have been used effectively over the years, but decision tree-based models have been the most common by far. Most of the older research papers on this subject made use of classification and regression trees (CART). For example, a case study done by Dekker et al. in 2009 used various types of decision trees, and their CART model ended up having the highest accuracy of 81%. They concluded that simpler classifiers gave a useful result with accuracies between 75 and 80%, which was hard to beat with other more sophisticated models.

When using a decision tree model, we must keep in mind that the variables we decide to include in our model will greatly affect the overall accuracy, and the Kovačić case study in 2012 is a testament to this. His models exclusively included socio-demographic variables (age, gender, ethnicity, education, work status, and disability), and left out important factors like academic achievement, financial aid, courses completed, etc. so his CART model had an accuracy of only 60.5%. He concluded that demographics alone were not statistically significant in predicting student success and dropout rates.

If we run into a similar problem where our original decision tree accuracy is too low, we may resort to using a random forest model. A random forest will help improve accuracy because it combines multiple decision trees into a single model, and that way we can include more significant factors. But we will still need to be careful with our variable selection so we can ensure that we have clean data (no duplicates or redundant data entries).

## 2.3 Data [15 points]

This data consists of information about students enrolled in undergraduate degrees. The dataset includes demographic data, social-economic factors and academic performance information about the students. This information can be used to predict whether a student may or may not drop out. The data contains 4424 observations and 35 variables and does not include any missing values.

The variables from that dataset we used are:

Marital status: The marital status of the student. (Categorical)

Course: The course taken by the student. (Categorical)

Daytime/evening attendance: Whether the student attends classes during the day or in the evening. (Categorical)

Previous qualification: The qualification obtained by the student before enrolling in higher education. (Categorical)

Nationality: The nationality of the student. (Categorical)

Mother's qualification: The qualification of the student's mother. (Categorical)

Father's qualification: The qualification of the student's father. (Categorical)

Mother's occupation: The occupation of the student's mother. (Categorical)

Father's occupation: The occupation of the student's father. (Categorical)

Displaced: Whether the student is a displaced person. (Categorical)

Educational special needs: Whether the student has any special educational needs. (Categorical)

Debtor: Whether the student is a debtor. (Categorical)

Tuition fees up to date: Whether the student's tuition fees are up to date. (Categorical)

Gender: The gender of the student. (Categorical)

Scholarship holder: Whether the student is a scholarship holder. (Categorical)

Age at enrollment: The age of the student at the time of enrollment. (Numerical)

International: Whether the student is an international student. (Categorical)

Curricular units 1st sem (credited): The number of curricular units credited by the student in the first semester. (Numerical)

Curricular units 1st sem (enrolled): The number of curricular units enrolled by the student in the first semester. (Numerical)

Curricular Units 1st sem (grade): The grade the Student received at the end of the first semester. (Numerical)

Curricular units 2nd sem (credited): The number of curricular units credited by the student in the second semester. (Numerical)

Curricular units 2nd sem (enrolled): The number of curricular units enrolled by the student in the second semester. (Numerical)

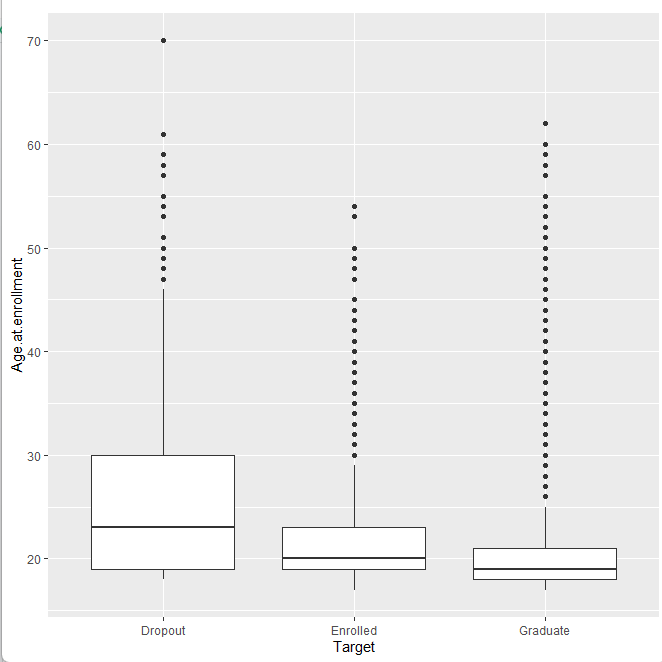
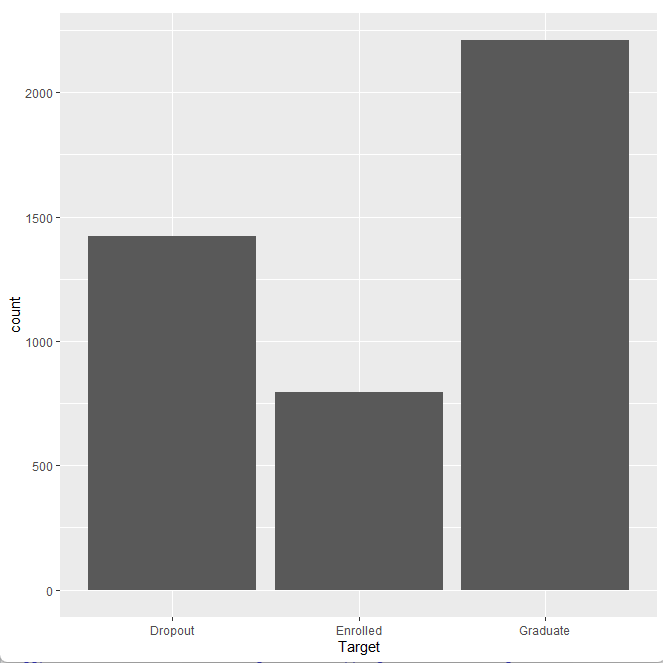
Curricular units 2nd sem (grade): The grade the student received at the end of the second semester. (Numerical).

Unemployment Rate: The current unemployment rate. (Numerical)

Inflation Rate: The current inflation rate. (Numerical)

**Target:** Dropout or Graduate. (Categorical)

The bar chart below shows the general distribution of the amount of students who dropout, who are enrolled, and who graduate.



The boxplot below shows the distribution of students who dropout, are enrolled, or who graduate based on the age that they are at the time of enrollment.

## 2.4 Method [20 points]

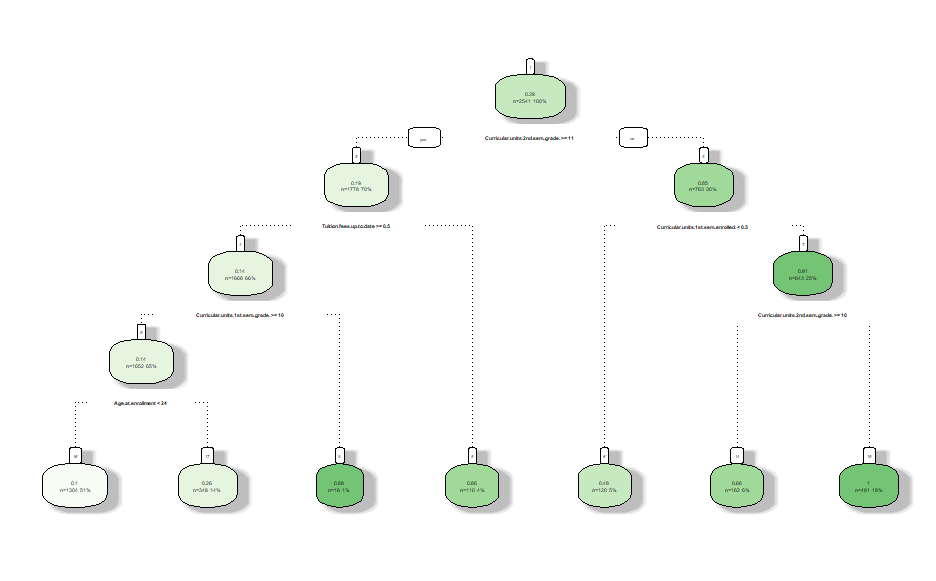
As previously stated, the original dataset comprised 35 variables and 4424 records with no missing variables. Upon further inspection of the dataset, it was determined that we would take out eight variables as they served no purpose. The variables that were dropped are as follows:

* Application mode
* Application order
* Curricular units 1st sem (evaluations)
* Curricular units 1st sem(approved)
* Curricular units 1st sem (without evaluations)
* Curricular units 2nd sem (evaluations)
* Curricular units 2nd sem(approved)
* Curricular units 2nd sem (without evaluations)

After removing the variables we changed the target variable to binary by getting rid of the enrolled option, keeping only graduate and dropout records. If a record contained the word “dropout” it would receive a 1 and if it contained “graduate” it would receive a 0. This altered our dataset by decreasing the number of records from 4424 to 3631.

It was decided that we would use decision trees for our model analysis as there are many variables in our dataset which could lead to many possible outcomes. This means that a decision tree would most accurately show us which factors affect dropout rates more, and how each individual decision can change the possibility of graduating or dropping out of college.

We created our first decision tree before moving on to perform random forest to try to improve the accuracy of our model.



We decided to run three different random forest models containing three different ntree values; 10, 50 and 100. After running each random forest we were able to view the most important variables in the models and create a confusion matrix to assess the accuracy of each model.

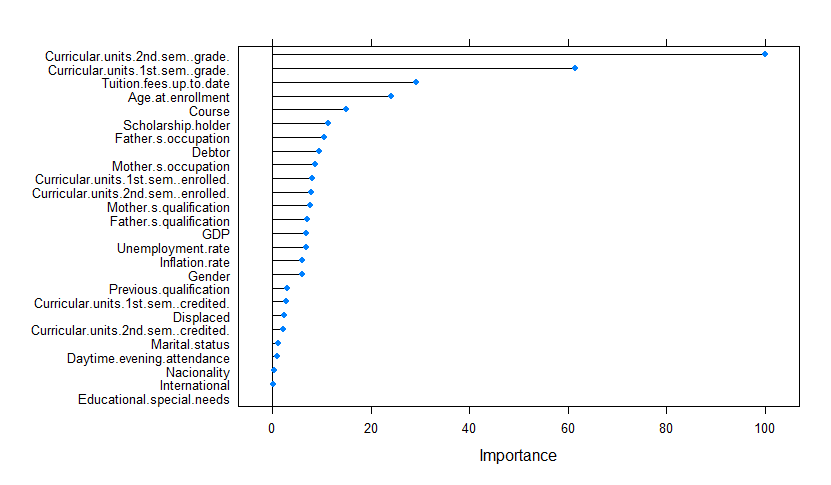
## 2.5 Results [10 points]

The random forest model with 100 iterations had the highest accuracy. The accuracy of our final model was 87.88%. From here we were able to view which variables had the greatest effect on whether a student is likely to drop out. The variable that has the greatest impact on students is their second semester grade, followed by their first semester grade and then whether their tuition fees are up to date. Other socioeconomic and demographic variables had much less of an impact than expected. It is clear from the analysis that to improve one's likelihood to graduate the main focus should be on your academic performance.

When comparing our results to our original theories it can still be seen that age does have a great impact on academic performance. Before running random forests, we created plots that showed how age affected dropout rates as we thought this variable would have the greatest impact. After running the random forests we can see that while age is not the most important variable, it is certainly one of top variables that influence drop out rates.

When comparing our results to the studies conducted by other researchers we have come up with very similar results with a higher accuracy. In another study done on the same dataset it was found that the most important variables were Curricular units at 2nd sem(approved), Grade 2nd sem, Curricular units at 1st sem(approved), Grade 1st sem, and Tuition fees up to date. The prediction accuracy of their model was 76%. When comparing the two results it is clear that academic performance and staying up to data with tuition fees are the most accurate predictors.

Based on the results the biggest surprise was that Educational.special.needs and International did not play any role to students dropping out of school. This shows that the students with disabilities or have traveled to different countries to attend college are being accommodated properly to treat their needs, guiding them to graduation and properly preparing them for the future.



## 2.6 Conclusion [5 points]

College is undoubtedly one of the most challenging and complex times in someone's life, but has become more necessary in today's society. It is important to pinpoint which factors could increase a students probability of dropping out so we can create a plan to combat the dropout rates.

By using decision trees, specifically random forests, we were not only able to pinpoint what variables influence someone's decision to drop out of college but also see the possible combinations of factors that lead to a drop out outcome. This means that a decision tree would *most accurately* show us these factors and how each individual decision can change the possibility of graduating or dropping out of college. It was found that rather than socioeconomic factors having the most influence, how many credits the student takes, monetary issues, and age have the biggest effect on a student dropping out.

**3.**0 **Appendix [No points but mandatory]**

Hanson, Melanie. “College Dropout Rates” EducationData.org, June 17, 2022,

<https://educationdata.org/college-dropout-rates>

Ameen et al. (2019)

<https://www.researchgate.net/publication/340406248_STUDENTS'_ACADEMIC_PERFORMANCE_AND_DROPOUT_PREDICTION>

Dekker et al. (2009)

<https://files.eric.ed.gov/fulltext/ED539082.pdf>

Kovačić (2012)

<https://repository.openpolytechnic.ac.nz/bitstream/handle/11072/1486/Kovacic%20~%20Predicting%20student%20success%20by%20mining%20enrolment%20data.pdf?sequence=1&isAllowed=y>

Leonhardt, D., & Chinoy, S. (2019, May 23). *The College Dropout Crisis*. The New York Times. Retrieved from https://www.nytimes.com/interactive/2019/05/23/opinion/sunday/college-graduation-rates-ranking.html

Whistle, W. (2019). *Ripple effect: The Cost of the College Dropout Rate – Third way*. Third Way. Retrieved from https://www.thirdway.org/report/ripple-effect-the-cost-of-the-college-dropout-rate

