Predict Students' Dropout and Academic Success w/

Data Wrangling, Visual Displays, and Hypothesis Testing

Daniel Song

Abstract and Introduction

The Student Dropout and Academic Success data set is a comprehensive data set on student profile and academic performance in higher education. It aims to contribute to the reduction of academic dropout and failure. The data includes information known at the time of student enrollment—academic path, demographics, and social-economic factors. Features like admission grades and semester grades might have more weight than something like a mother's or father's occupation/qualification because grades are heavily influenced by behaviors such as student attendance, study habits, and participation. These are within the student's control and can change academic outcomes more immediately than socioeconomic factors. However, if social class differs by a wide margin, this can affect the students' efforts to try hard in the courses. For this problem, we used visuals displays to identify students at risk at an early stage of their academic path so that strategies to support them can be put into place. We then performed hypothesis testing to conclude our analysis and weight the factors.

Materials

- The Student Success and Dropout dataset is from the UCI Machine Learning Repository: Predict Students' Dropout and Academic Success - UCI Machine Learning Repository
- The dataset consists of 4,421 rows of student data and 37 columns. It includes information about demographics (e.g., marital status, nationality, parental education, and occupation), academic performance (e.g., admission grades, first and second-semester grades, curricular unit evaluations and approvals), and socioeconomic indicators (e.g., unemployment rate, inflation rate, GDP). The target variable, "Target", indicates whether a student graduated, dropped out, or is still enrolled.

Methods

Data Preparation

- Missing values and duplicates were removed from the dataset using na.omit() and unique().
- Columns were standardized for consistency using make.names().
- Numeric columns were selected for further analysis.

Data Visualization

- A correlation heatmap was generated to explore relationships between numeric variables using ggplot2.
- A histogram was created to analyze the distribution of admission grades.
- A box plot was used to compare first-semester grades across target groups (Dropout, Enrolled, Graduate).
- A scatterplot visualized the relationship between first-semester and second-semester grades.

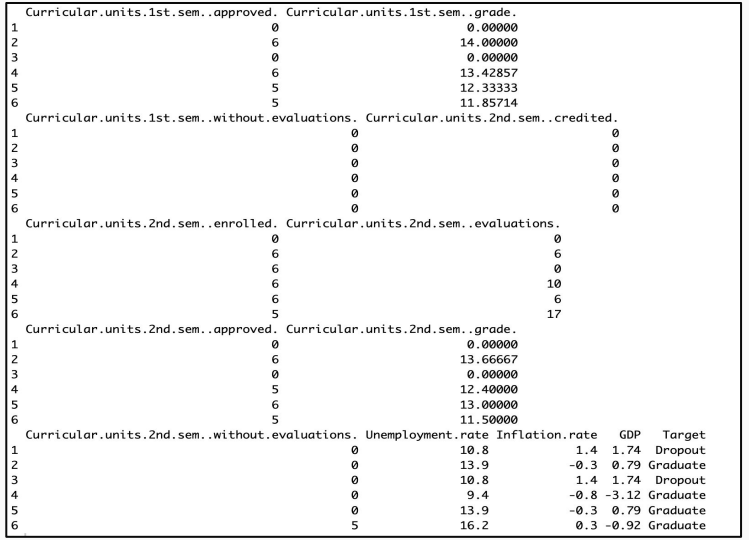
Methods continued...

Hypothesis Testing

- A t-test was conducted to test for significant differences in admission grades between "Enrolled" and "Dropout" groups.
- Graduate was not used here because these grades depend on current grades used or previous but recent admission grades.
- p-values and confidence intervals were calculated to evaluate statistical significance.

Parts of the Data

- From this output, we can see that each line represents many attributes that are associated with a student.
- Each attribute of the data is numerical. Some numbers are unique in the sense that categorical variables would be better suited
    - For example, in “Mother’s occupation,” the arrow pointing to number “5” would mean that the mother works as “Personal Services, Security and Safety Workers and Sellers.”



	Curricular.units.1st.sem..approved.	Curricular.units.1st.sem..grade.
1	0	0.00000
2	6	14.00000
3	0	0.00000
4	6	13.42857
5	5	12.33333
6	5	11.85714

	Curricular.units.1st.sem..without.evaluations.	Curricular.units.2nd.sem..credited.
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0

	Curricular.units.2nd.sem..enrolled.	Curricular.units.2nd.sem..evaluations.
1	0	0
2	6	6
3	6	0
4	6	10
5	6	6
6	5	17

	Curricular.units.2nd.sem..approved.	Curricular.units.2nd.sem..grade.
1	0	0.00000
2	6	13.66667
3	0	0.00000
4	5	12.40000
5	6	13.00000
6	5	11.50000

	Curricular.units.2nd.sem..without.evaluations.	Unemployment.rate	Inflation.rate	GDP	Target
1	0	10.8	1.4	1.74	Dropout
2	0	13.9	-0.3	0.79	Graduate
3	0	10.8	1.4	1.74	Dropout
4	0	9.4	-0.8	-3.12	Graduate
5	0	13.9	-0.3	0.79	Graduate
6	5	16.2	0.3	-0.92	Graduate

	Marital.status	Application.mode	Application.order	Course	Daytime.evening.attendance.
1	1	17	5	171	1
2	1	15	1	9254	1
3	1	1	5	9070	1
4	1	17	2	9773	1
5	2	39	1	8014	0
6	2	39	1	9991	0

	Previous.qualification	Previous.qualification..grade.	Nacionality	Mother.s.qualification
1	1	122.0	1	19
2	1	160.0	1	1
3	1	122.0	1	37
4	1	122.0	1	38
5	1	100.0	1	37
6	19	133.1	1	37

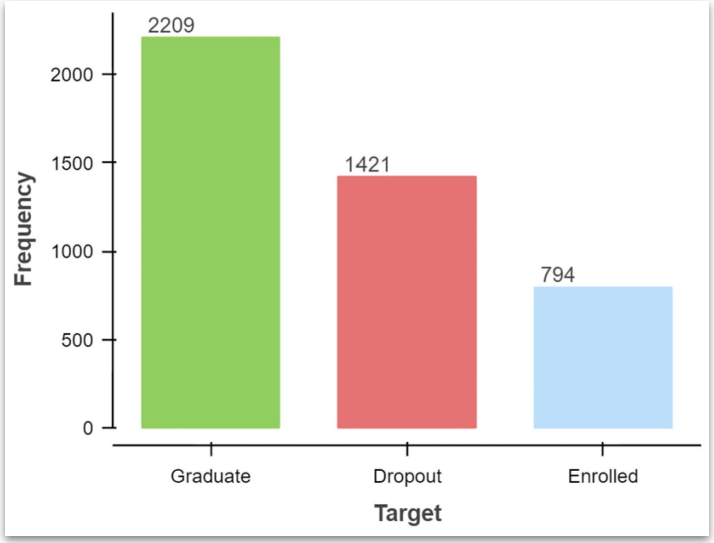
	Father.s.qualification	Mother.s.occupation	Father.s.occupation	Admission.grade	Displaced
1	12	5	9	127.3	1
2	3	3	3	142.5	1
3	37	9	9	124.8	1
4	37	5	3	119.6	1
5	38	9	9	141.5	0
6	37	9	7	114.8	0

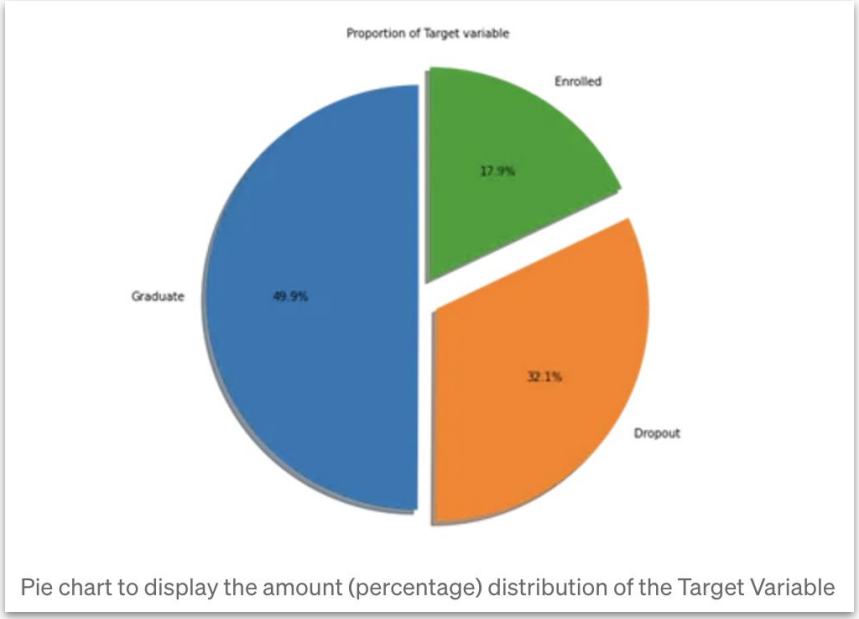
	Educational.special.needs	Debtor	Tuition.fees.up.to.date	Gender	Scholarship.holder
1	0	0	1	1	0
2	0	0	0	1	0
3	0	0	0	1	0
4	0	0	1	0	0
5	0	0	1	0	0
6	0	1	1	1	0

	Age.at.enrollment	International	Curricular.units.1st.sem..credited.
1	20	0	0
2	19	0	0
3	19	0	0
4	20	0	0
5	45	0	0
6	50	0	0

	Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..evaluations.
1	0	0
2	6	6
3	6	0
4	6	8
5	6	9
6	5	10

Pie chart and Histogram of all students' "Target Status"





Pie chart to display the amount (percentage) distribution of the Target Variable

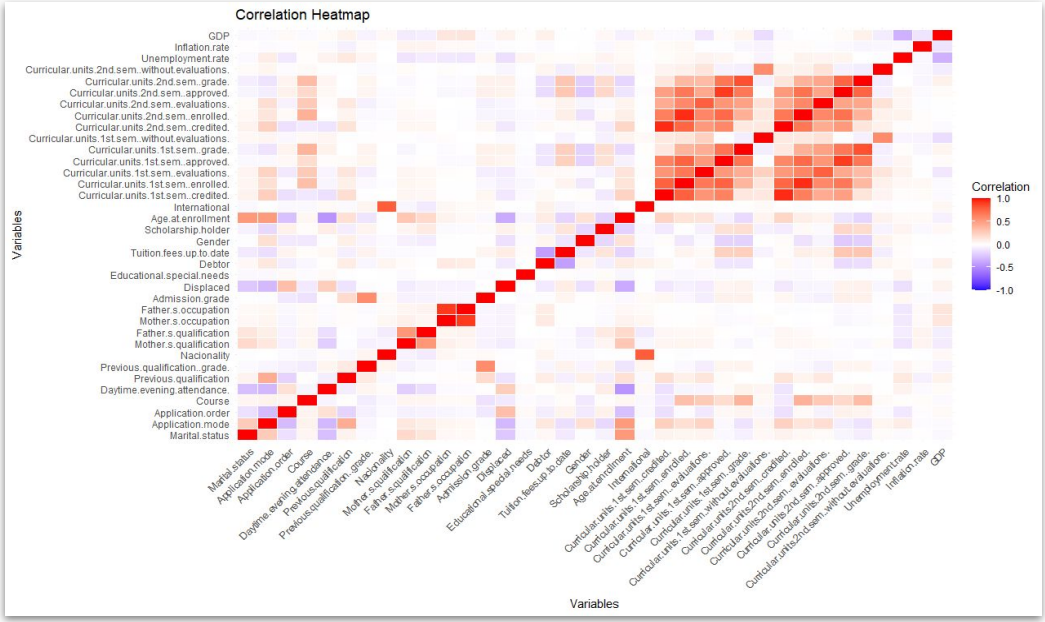
In this general pie chart, around 49.9% of the students are graduated, 32.1% students are dropout, and 17.1% of the students are enrolled in some other course.

The histogram shows the distribution of student records among the three target categories considered for academic success.

Results and Analysis

We will divide the data set into three factors: Academic, Socio-Economic, and Demographic

Correlation Heatmap Across All Variables



This heatmap highlights the correlations between all numeric variables in the dataset. Darker red squares indicate strong positive correlations, while blue areas represent negative correlations. It provides an overview of how variables such as socioeconomic factors and academic performance are interrelated.

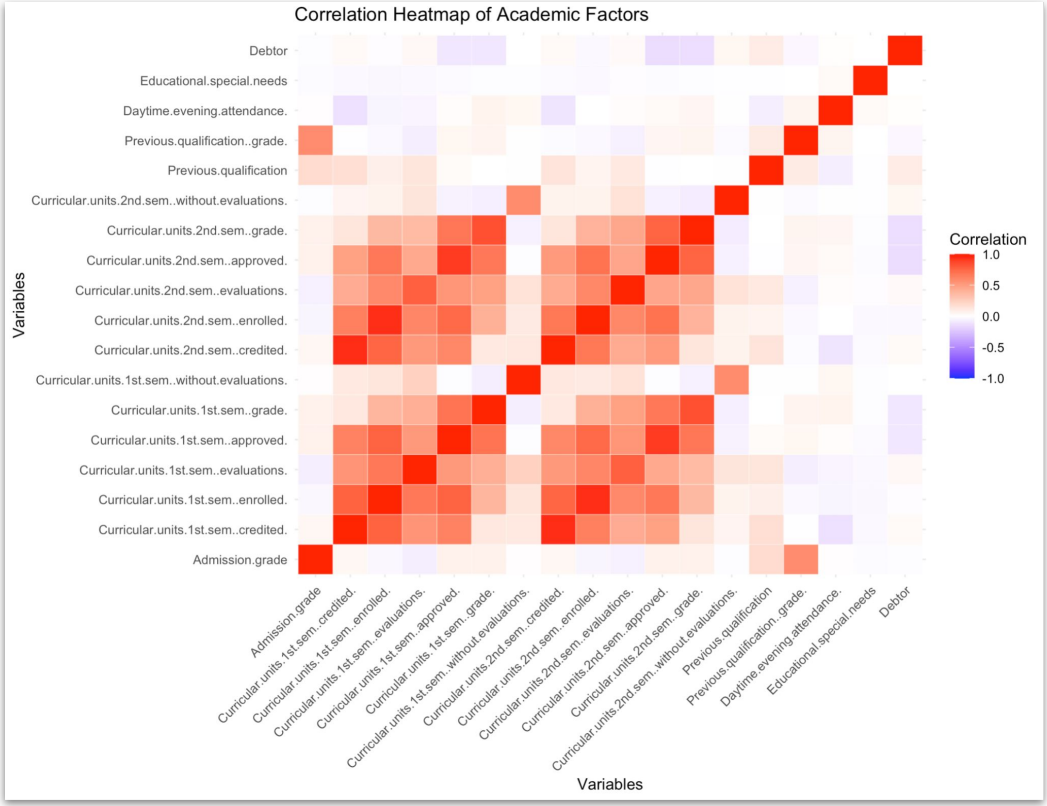
Academic Factors

High Correlations:

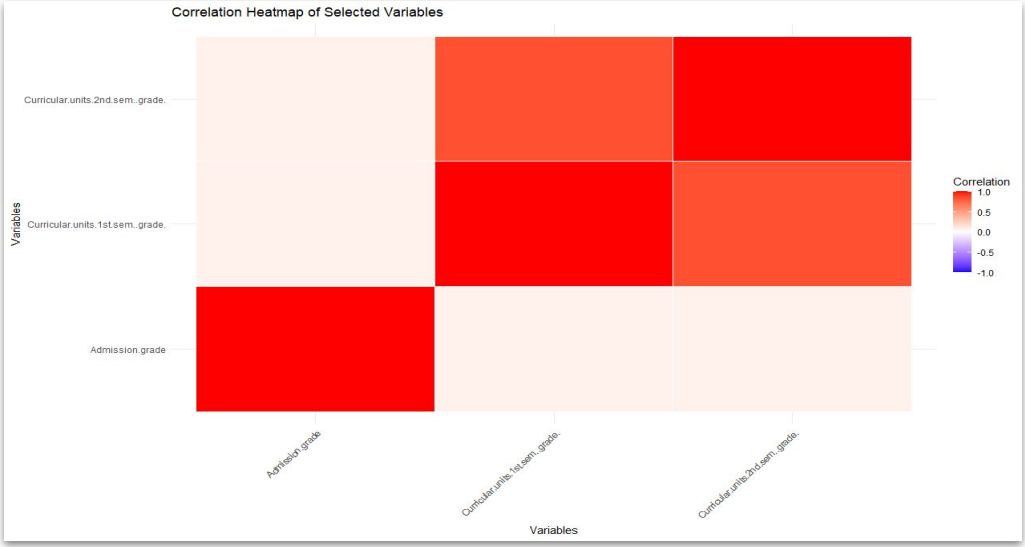
- "Curricular variables" (e.g., grades, credits, evaluations) for the first and second semesters show strong positive correlations, suggesting interdependence.
- Admission grade also correlates with "Curricular Variables"

Negative correlations:

- Debtor and Educational special needs show weak correlations with other variables, which indicates that these may be less related to academic performance.

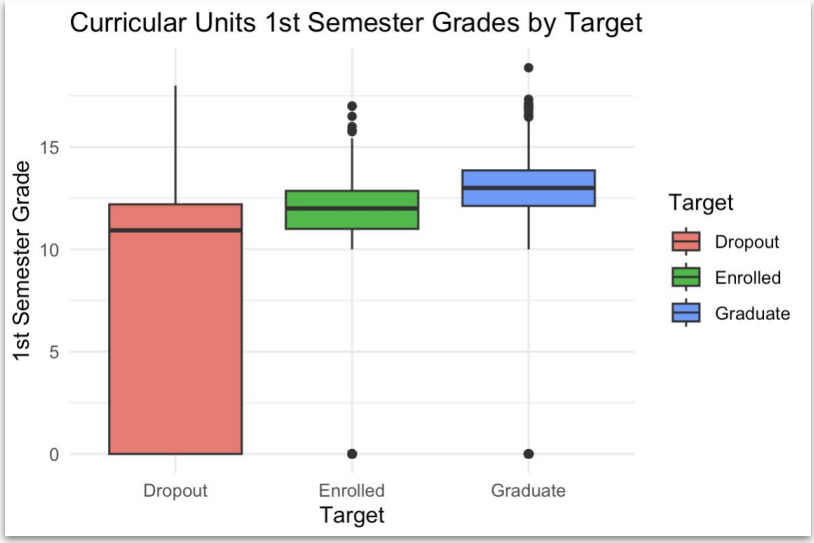


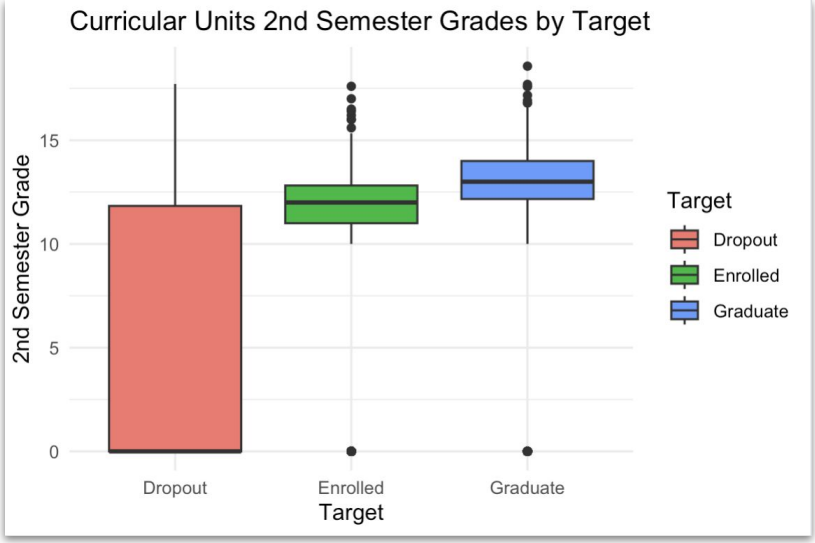
1st and 2nd Semester Grade Heatmap



The heatmap shows the correlations between admission grade, first-semester grade, and second-semester grade. Strong positive correlations (in red) are observed between these variables, indicating that higher admission grades are generally associated with better semester grades.

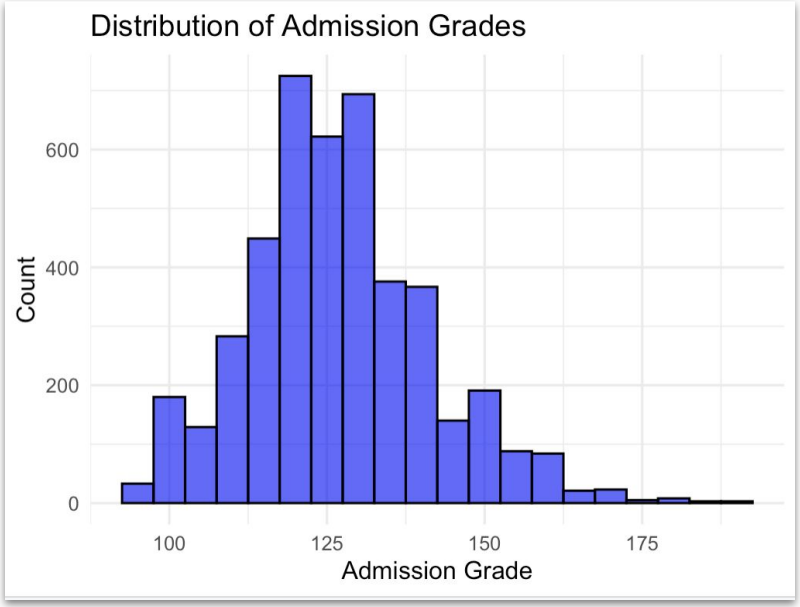
Semester Grades Box Plots





The box plots display the first-semester and second-semester grades categorized by student outcomes (Dropout, Enrolled, Graduate). Graduates tend to have higher and less variable grades, while dropouts show a wider spread and lower median grades in both semesters.

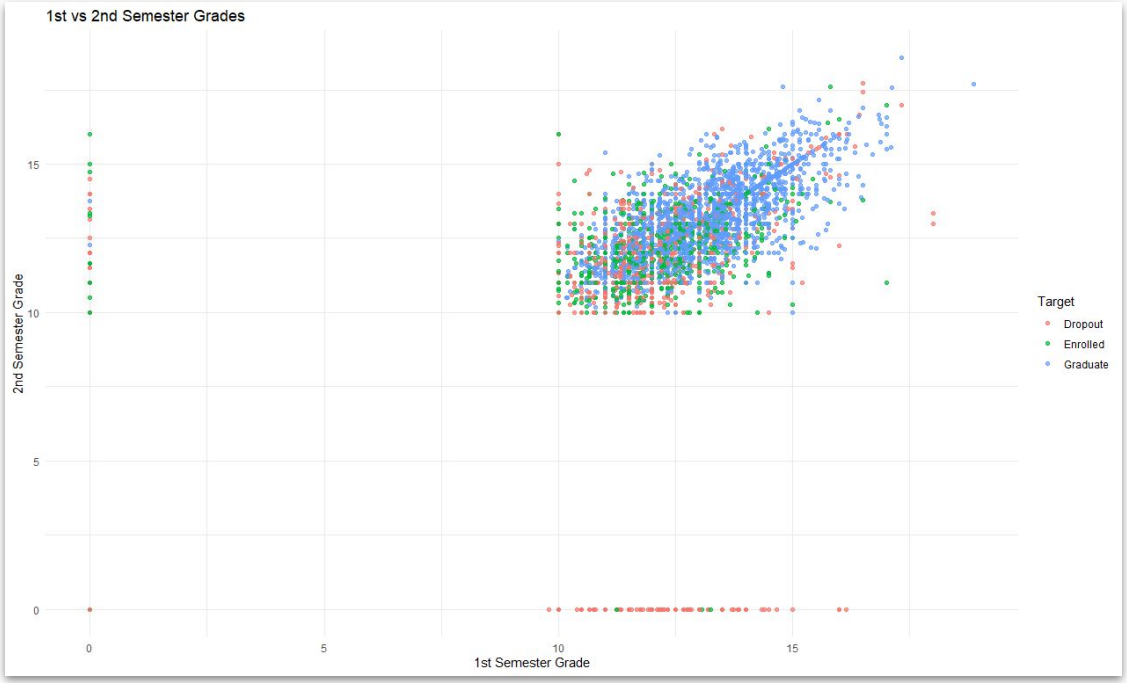
Histogram of Students' Admission Grades



The histogram reveals that most students cluster around admission grades of 120 to 140, with fewer students at the extremes. This distribution appears slightly skewed to the left, indicating that lower admission grades are less common.

1st and 2nd Semester Grades Scatterplot

- The scatterplot shows a clear positive relationship between first and second-semester grades, with points colored by student outcomes (Dropout, Enrolled, Graduate).
- Graduates (blue) cluster in the upper-right corner, indicating consistently high grades in both semesters.
- Enrolled students (green) are more spread out, suggesting mixed performance.
- Dropouts (red) are concentrated near the lower left, showing lower grades in both semesters.
- The plot emphasizes that strong first-semester grades are a good predictor of success in the second semester.



From this point, we can conclude that those students who consistently show **high grades** each semester are likely to be academically **successful**. Therefore, the **dropout rate** will be much **lower**. We will now consider the other two factors that affect the students who dropped out of university.

Socio-Economic Factors

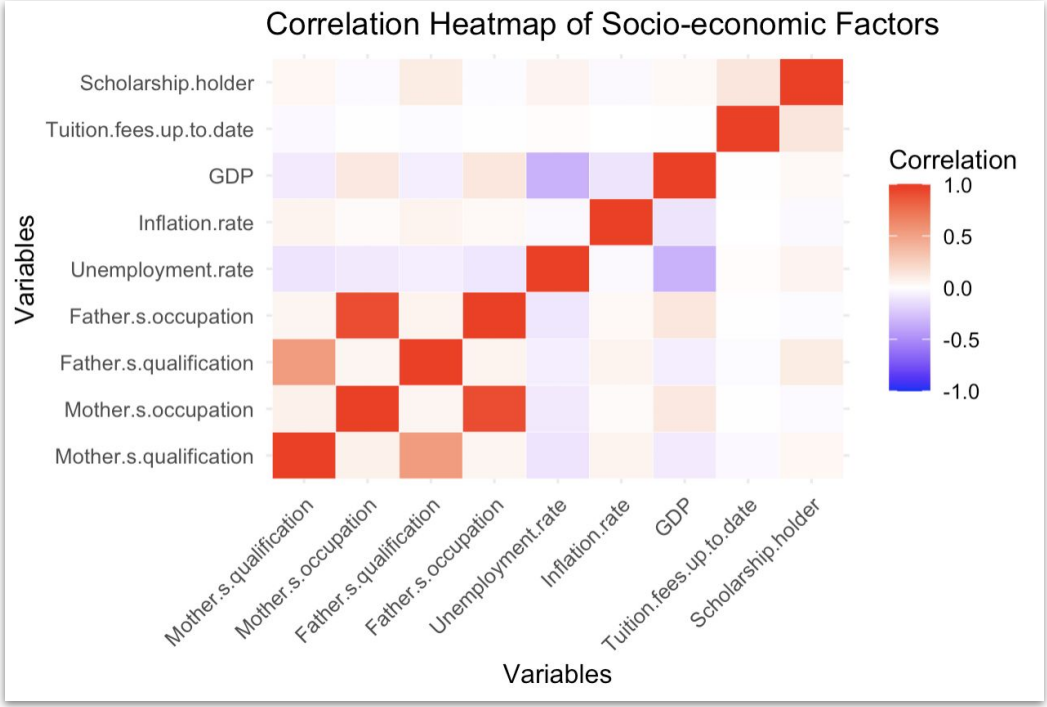
High Positive Correlations

- Father's and Mother's occupation/qualification implies that the parent's job type is closely related to their education level.

Negative Correlations

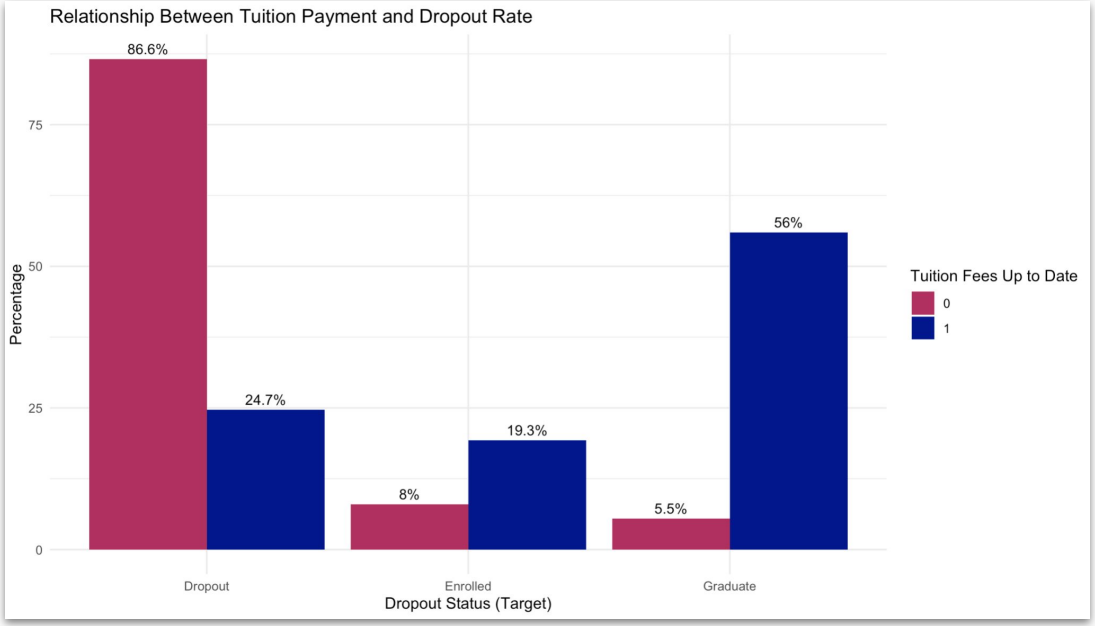
- Unemployment rate and GDP
- These variables align well with the economic principle that economic growth or higher GDP usually leads to lower unemployment rates.

Keeping up to date with tuition fees is also important and can affect the current students from either dropping out or staying enrolled.



Histogram on the relationship between Tuition Payment and Dropout Rate

- The histogram was created by setting 0 for those who didn't pay their tuition up to date out from school and 1 for those who did pay their tuition up to date. It focuses on the dropout rate while also showing the financial statuses for enrolled and graduate students.
- In this plot, many students who have paid off their tuition fees have not dropped out of school. In other target statuses, the dropout rate is lower. This shows that students' academic performance is significantly impacted by financial support, and that flexible late payment policies and larger scholarship programs might be effective in lowering dropout rates.



Percentages in each target variable do not add up to 100% because it is not normalized across all groups for a single target.

This is just one feature of the Socio-Economic factor where the dropout rate shows greatly for students with a lower background statuses. All other features somewhat behave the same way.

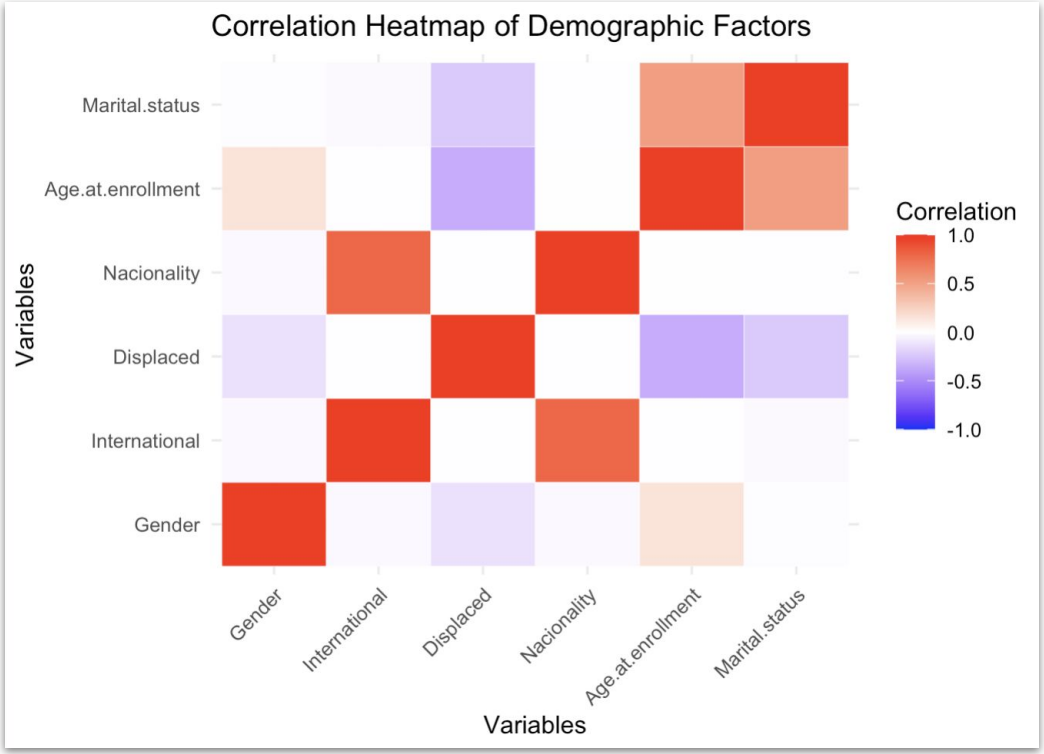
Demographic Factors

Highly Correlated Variables:

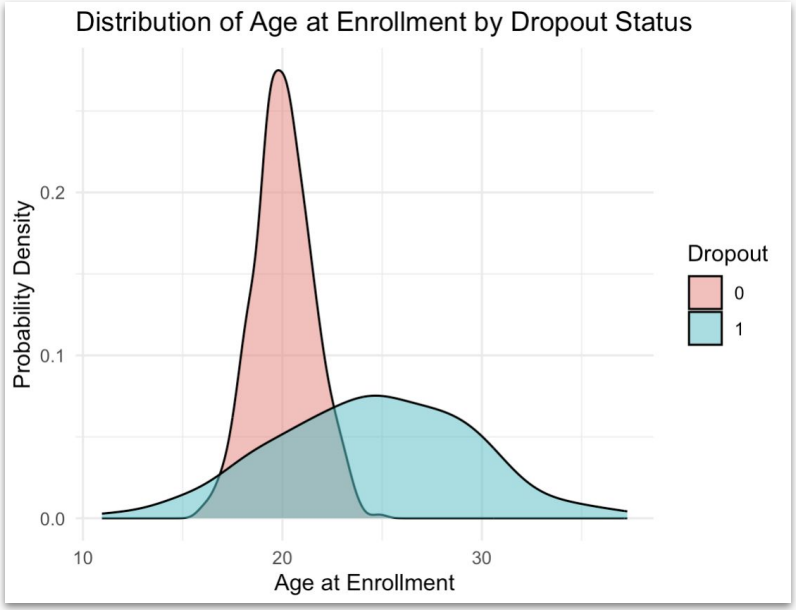
- Nationality and International: International status is strongly tied to Nationality. They likely provide overlapping information.
- Marital status and Age at enrollment

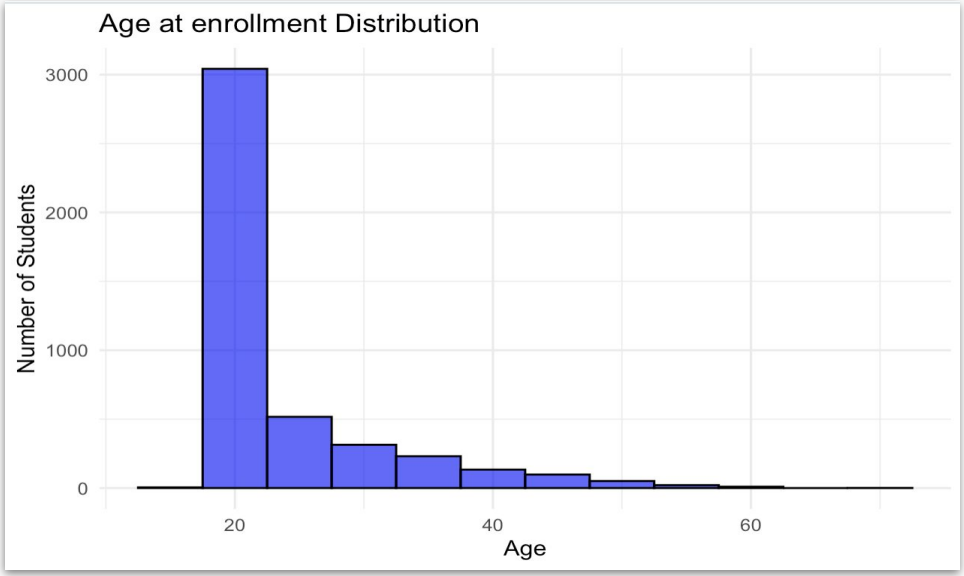
Negative Correlation Variables

- Age at enrollment and displaced: Displacement is usually influenced by socio-political or economic factors rather than the age at which a person enrolls in an institution.
- Gender and Displaced
- Marital status and Displaced



Age at enrollment Histogram and Relationship with dropout rates



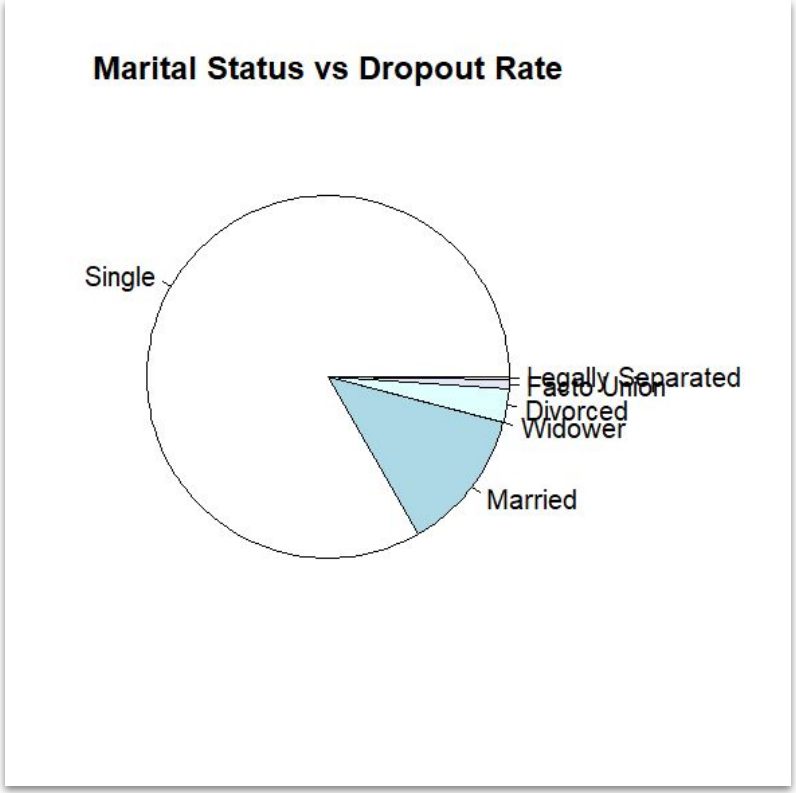


The histogram shows the distribution of students ages. Most instances of students are around 20 years of age.

In the KDE plot, the pink graph show graduates, where the majority of them are in their 20s, show that they are least likely to drop out. On the other hand, the graph representing dropouts (green) remains moderate, even among those in their 30s. That means that students who are older at enrollment are more likely to drop out.

Marital Status vs Dropout Rate

This chart visualizes the relationship between marital status and dropout rate. Each slice represents a marital status category, labeled with both the category name and the number of dropouts. The chart highlights how dropout rates vary across different marital statuses, providing insights into potential patterns or demographic influences affecting student retention.



Hypothesis Testing (Two Sample t-test) with select variables from each factor: Academic, Socio-Economic, Demographic

Academic Factors

The p-values for the 1st and 2nd semester grades are very small, which indicates strong evidence against the null hypothesis. This suggests that academic performance in the first and second semester can be an important differentiator between students who drop out and those who remain enrolled.

The large t-statistics for the 1st and 2nd semester grades indicate a considerable effect size, meaning that the difference in grades between the two groups is statistically significant and meaningful.

The lack of a significant difference in admission grades suggests that pre-enrollment academic performance is not a strong predictor of student success or dropout.

```
        Two Sample t-test

data:  Admission.grade by Target
t = -0.88189, df = 2213, p-value = 0.3779
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -1.8468222  0.7010389
sample estimates:
 mean in group Dropout mean in group Enrolled
              124.9614               125.5343
```

```
        Two Sample t-test

data:  Curricular.units.1st.sem..grade. by Target
t = -16.448, df = 2213, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -4.329850 -3.407353
sample estimates:
 mean in group Dropout mean in group Enrolled
              7.256656              11.125257
```

```
        Two Sample t-test

data:  Curricular.units.2nd.sem..grade. by Target
t = -21.994, df = 2213, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -5.683271 -4.752780
sample estimates:
 mean in group Dropout mean in group Enrolled
              5.899339              11.117364
```

The select variables include Admission grade, 1st semester grade, and 2nd semester grade. The target variables were either Dropout or Enrolled.

Socio-Economic Factors

Tuition Fees Up to Date by Target

The t-test results indicate a significant difference in the mean tuition fees between the two groups. The low p-value and the t-statistic leads to the rejection of the null hypothesis where there is minimal difference in means. The mean tuition fees for the Enrolled group (0.9471) are significantly higher than for the Dropout group (0.6784). This suggests that enrolled students tend to have higher tuition fees up to date.

The analysis shows a considerable difference in the mean of the mother's and father's occupation variable between the two target groups. The t-statistic and the p-value allows us to reject the null hypothesis. There is also significant difference in the CI interval. The mean for the Enrolled group is higher than for the Dropout group, suggesting that students who are enrolled have mothers and fathers with higher occupational status.

```
        Two Sample t-test

data:  Tuition.fees.up.to.date by Target
t = -15.254, df = 2213, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -0.3032530 -0.2341626
sample estimates:
 mean in group Dropout mean in group Enrolled
             0.6783955              0.9471033
```

```
        Two Sample t-test

data:  Mother.s.occupation by Target
t = -3.7449, df = 2213, p-value = 0.0001851
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -7.015377 -2.193198
sample estimates:
 mean in group Dropout mean in group Enrolled
              10.11612               14.72040
```

```
        Two Sample t-test

data:  Father.s.occupation by Target
t = -3.6675, df = 2213, p-value = 0.0002507
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 -6.625322 -2.008655
sample estimates:
 mean in group Dropout mean in group Enrolled
              10.14145               14.45844
```

The select variables include Tuition fees up to date, Mother's occupation, and Father's occupation. The same target variables were used.

Demographic Factors

We reject the null hypothesis since both p-values for Age at enrollment and Marital status are well below 0.05.

Students who drop out tend to enroll at a later age compared to those who remain enrolled. This could indicate that older students face unique challenges that may contribute to higher dropout rates.

Age at Enrollment: Older students are more likely to drop out compared to those who remain enrolled, possibly due to outside responsibilities or challenges associated with returning to education later in life.

Marital Status: Students with higher marital status values are slightly more likely to drop out, potentially due to family or relationship duties.

```
        Two Sample t-test

data:  Age.at.enrollment by Target
t = 10.534, df = 2213, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 3.011126 4.388770
sample estimates:
 mean in group Dropout mean in group Enrolled
              26.06897               22.36902
```

```
        Two Sample t-test

data:  Marital.status by Target
t = 3.6516, df = 2213, p-value = 0.0002666
alternative hypothesis: true difference in means between group Dropout and group Enrolled is not equal to 0
95 percent confidence interval:
 0.05032052 0.16706107
sample estimates:
 mean in group Dropout mean in group Enrolled
              1.261084               1.152393
```

The select variables include Age at enrollment and Marital Status. The same target variables were used. From our visual displays, gender, nationality, and similar variables didn't really have a big impact on the dropout rate of the students currently school.

Discussion

Key Points in the dropout rate trend:

- The dropout rate increases with age (especially in the 30s).
- Marital status is closely correlated to the number of dropouts. Being married, divorced, or widows increases the dropout rate.
- Having the ability to pay tuition up to date is linked to the number of dropouts.
- There is minimal relationship between the number of dropouts and the unemployment rate/GDP
- Scholarship holders are more likely not to drop out since they have a “real” reason to stay in school.
- Students with good grades are naturally at lower risk of dropping out.
- Socioeconomic status plays a part in whether students would remain enrolled while paying the tuition given by their parents or drop out due to halted payments or financial plans.

Discussion continued..

- Being able to recognize a student's past performance is a great indicator of academic success/failure. If this performance is poor, tutoring programs and accommodations should be offered while being alert of their grades and personal environment around their family.
- Although academic factors like grades and demographic factors like nationality and age do affect student's success or failure, the dropout rate is more affected by socioeconomic factors and the student's personal background.
- To reduce the number of dropouts, colleges and universities should provide financial support and flexible educational environments for students facing financial difficulties, as these weigh heavily on the their efforts to do well in school.

We will now implement a **random forest** (RF) model to determine the **important** variables and other **predictions** and **errors**.

RF confusion matrix, OOB rate, and RMSE

We use the out-of-bag-evaluation (OOB evaluation) to evaluate the quality of the model.

Here the OOB rate is 22.27% which is reasonable for this kind of dataset, though there is room for improvement.

The RMSE was 0.6544 which is also reasonable and just outside the optimal range.

```
Call:
 randomForest(formula = Target ~ ., data = train_data, importance = TRUE,      ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 6

        OOB estimate of  error rate: 22.27%
Confusion matrix:
         Dropout Enrolled Graduate class.error
Dropout      873       97      167  0.23218997
Enrolled     140      238      257  0.62519685
Graduate      49       78     1640  0.07187323
```

Confusion Matrix Analysis

Statistics by Target Status:

Dropout:

- The model correctly predicted 873 instances as "Dropout."
- 97 instances of "Enrolled" and 167 instances of "Graduate" were misclassified as "Dropout."
- Class error: 23.22%, indicating moderate accuracy for this class.

Enrolled:

- Only 238 instances were correctly predicted as "Enrolled."
- 140 and 257 instances were misclassified as "Dropout" and "Graduate," respectively.
- Class error: 62.52%, showing poor accuracy for this class.

Graduate:

- The model performed very well for this class, correctly predicting 1640 instances.
- Only 49 and 78 instances were misclassified as "Dropout" and "Enrolled," respectively.
- Class error: 7.18%, indicating high accuracy for this class.

Model Accuracy

The p-value is less than 0.05 so the accuracy improvement is statistically significant. The model correctly predicted 76.61% of the classes.

No Information Rate (NIR): 0.4994, representing the accuracy by always predicting the majority class. The model accuracy is significantly greater than NIR which suggest that the model is learning patterns from the dataset and outperforming a naive baseline.

```
Overall Statistics

               Accuracy : 0.7661
                 95% CI : (0.7368, 0.7936)
    No Information Rate : 0.4994
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6036

 Mcnemar's Test P-Value : 4.087e-10

Statistics by Class:

                     Class: Dropout Class: Enrolled Class: Graduate
Sensitivity                  0.7500         0.35220          0.9253
Specificity                  0.9235         0.93526          0.7427
Pos Pred Value               0.8224         0.54369          0.7820
Neg Pred Value               0.8866         0.86829          0.9088
Prevalence                   0.3209         0.17966          0.4994
Detection Rate               0.2407         0.06328          0.4621
Detection Prevalence         0.2927         0.11638          0.5910
Balanced Accuracy            0.8367         0.64373          0.8340
```

Model Accuracy Results and Analysis

Statistics by Target Status:

Dropout

- **Sensitivity (Recall):** 0.7500, meaning 75% of actual dropouts were correctly identified.
- **Specificity:** 0.9235, showing the model is 92.35% correct in identifying non-dropouts.
- **Balanced Accuracy:** 0.8367, reflecting a good balance between sensitivity and specificity.

Enrolled

- **Sensitivity:** 0.3522, meaning only 35.22% of enrolled students were correctly identified (low recall).
- **Specificity:** 0.9353, indicating a high ability to identify non-enrolled cases.
- **Balanced Accuracy:** 0.6437, highlighting the challenge in predicting this class.

Graduate

- **Sensitivity:** 0.9253, meaning the model successfully identified 92.53% of graduates.
- **Specificity:** 0.7427, slightly lower, showing a challenge in differentiating graduates from others.
- **Balanced Accuracy:** 0.8340, showing good performance.

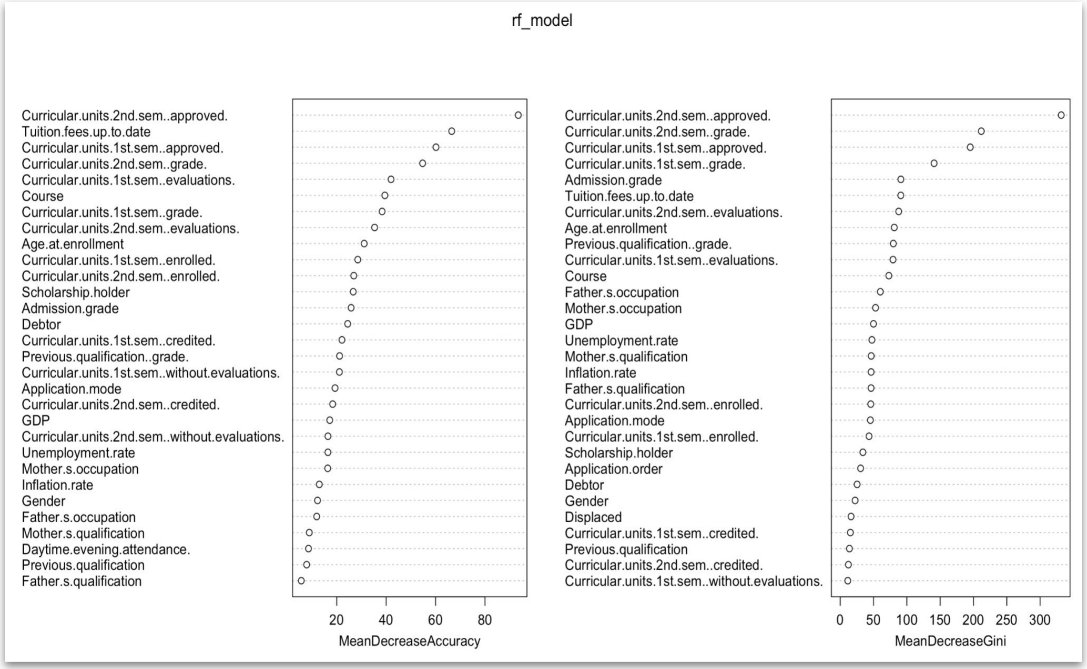
RF Plots

Higher the value of the mean decrease accuracy or mean decrease gini score, higher the importance of the variable.

Mean Decrease Accuracy - How much the model accuracy decreases if we drop that variable.

Mean Decrease Gini - Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees.

The most important feature is the curricular units 2nd semester approved grades which refers to the final grades that have been officially accepted for all the courses a student has taken during their second semester of a curriculum. Eventually, these are the grades that will count towards their overall academic record for that semester. The 2nd semester grades stand out as a high outlier compared to the rest of the features.



RF plot continued... and Discussion

- Based on the mean decrease gini and mean decrease accuracy scores, previous qualification was consistently lower than other variables and deemed to be the least important feature to predicting success and dropouts. This includes qualification like a student's high school diploma or GED, Advanced Placement credits, or other institution scores.
- Compared to the discussion made previously about socio-economic factors being the most heavily affected, it was the approved semester grades and evaluations that showed most importance.
- Strategies in improving student grades would help to lower dropout rates. This can include creating more tutoring programs, interactive session among professors, teacher assistants or students, and effective study groups.

Literature Cited/Acknowledgements

Predict Students Dropout or Academic Success

Prediction For Success/Failure

ML and Data Visuals

College-Dropout-Rates

ChatGPT was used for formatting and styling

Appendices

Data Visuals Code: Link

Random Forest Code: Link