Predicting Student Dropout and Academic Success

Patricia Götz Lana Kabbani Noémie Glaus
Estela Gonzalez Vizcarra

2025-12-20

1. - Introduction

Student retention and **academic success** are crucial challenges for higher education institutions worldwide. Recent international observations show **rising university dropout** trends across multiple regions, including Australia and the United States (Sokolova, 2025). Looking closer at Europe, recent data from the German Center for Higher Education Research and Science Studies (2022), show that **almost 30% of bachelor's students in Germany**

leave university **without graduating** (Hachmeister & Berghoff, 2024). **In Portugal**, which is the focus of our analysis, recent data by Statistics Portugal reveal that a considerable portion of **young adults (16.8%) aged from 15 to 34 have dropped out at least one level of education during their academic path** (Europe-Data.com, 2025). Moreover, among those who dropped out, over more than half **(50.8%) did not complete their tertiary studies**, highlighting that higher education represents a **critical point of disengagement** (Europe-Data.com, 2025). These figures underline the **seriousness of dropouts in higher education** and the reinforced **need for universities to rely on data-driven insights** to identify at-risk students and to **design early intervention strategies**. We chose this topic because **predicting student dropout** not only helps **optimize institutional resources** but also supports students in **achieving their academic goals**. Understanding the **factors that influence academic success**, such as **socio-economic background, previous academic performance, or family situation**, can improve educational policies and personalized support systems. This subject is particularly meaningful in data science, as it allows us to combine analytical and predictive methods to **better understand and prevent student dropout**.

1.1 - Project Goals

The **main objective** of this project is to **identify the factors that influence students to drop out, stay enrolled, or graduate from higher education**. The dataset provides detailed information on each student's academic performance, socioeconomic background, and demographic profile, offering a **comprehensive view of the variables that shape educational outcomes**. By the end of our analysis, we seek to **identify the most significant combinations of academic and personal factors that influence student success**. First, our analysis will focus on **academic performance**, examining how variables such as admission grades, semester evaluations, and course results relate to final outcomes. For instance, we will analyze whether **early academic performance** can serve as a **reliable predictor of future dropout risk**. We will then explore the **influence of socioeconomic and personal factors**, including parental education, occupation, and financial situation, to understand their **impact on academic achievement**. Lastly, the dataset will be used to build and evaluate **classification models that predict students' academic status** (**Dropout, Enrolled**, or **Graduate**). In summary, this study combines exploratory analysis, visualization, and predictive modeling to generate **actionable insights that help universities detect at-risk students early and strengthen academic success**.

---

2 - Related Work

As students' dropout is a **major challenge** in higher education, it represents a well-established area of research that has widely been studied in the literature over the years. Previous research

articles have helped us acquire information about the topic, including the **methodological approaches** used to address the different research questions.

One relevant study titled *"Predicting Students' Academic Success and Dropout Using Supervised Machine Learning"* (Arora et al., 2024) investigates the prediction of student academic success using **supervised machine learning classification models**. Throughout the paper, the authors compare multiple classification algorithms such as **Decision Trees**, **Random Forest** or **Logistic Regression** to assess their ability to predict student outcomes on student data. Their results show that these models are in fact an effective tool for identifying students at risk of dropping out and thus highlights the relevance of formulating this issue as a classification task.

Other articles emphasize the importance of constructing **robust predictive models**, but also the role of **feature selection**. In particular a recent paper by *Anaíle Mendes Rabelo* and *Luis Enrique Zarate* (2024) (Anaíle Mendes Rabelo, 2024) demonstrates how combining academic performance indicators with contextual variables such as course selection, improves the reliability of dropout prediction models.

In addition, an article published in 2022 named *"Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms"* (Oqaidi et al., 2022) focuses on the overall **analytical pipeline** used in educational **data mining**, from **data preprocessing** to **model evaluation**. The authors underline that data quality and preprocessing decisions play a **key role in model performance**.

Overall, these research articles guided our methodological choices, particularly our use of a classification framework, our focus on feature selection and our structured analytical process, with additional emphasis on **exploratory data analysis** and **model interpretability**.

3 - Research Questions

- I. How do academic performance indicators and study conditions influence students' likelihood of graduation or dropout?
- II. What is the impact of demographic and socioeconomic background on students' probability of dropping out?
  - a. To what extent do financial factors (debtor status, scholarship holder) affect student retention?
- III. Can we accurately predict a student's final status (Dropout, Enrolled, or Graduate) based on their demographic, socioeconomic, and academic characteristics. Which are the most relevant among them?

- a. Which features category, academic (grades, units), socioeconomic (debt, scholarship) or demographic (age, gender) contribute the most in predicting students' dropout?

4 - Data

4.1 - Data Sourcing

The dataset is publicly available on UCI Machine Learning Repository and was created from multiple databases of higher education institutions in Portugal. It is related to enrolled students in different undergraduate programs and shows **how different demographic, socioeconomic and academic factors are related to the dropout**. Since the data has already been collected and can be directly downloaded from UCI MLR - Predict Students' Dropout and Academic Success - [Accessed on 20th October] , there is no need to collect more data via webscraping or APIs.

4.2 - Data Description

The dataset, containing data from a Portuguese higher education institution, is provided as a CSV file, approximately 520 KB in size, and contains detailed information about students'demographic, academic and socio-economic characteristics. It includes 4424 student records and 37 variables (features). After reviewing the dataset variables, we removed two irrelevant ones, resulting in **35 relevant variables selected for analysis**.

4.2.1 - Data Loading

```
Dataset shape: (4424, 37)
```

4.2.2 - Variable Selection

We selected 35 relevant variables for analysis:

```
Selected 35 variables
```

4.2.3 - Selected Variable Descriptions

Variable	Description	Type
Marital Status	Student marital status	Categorical
Application order	Application preference order	Categorical
Course	Course taken by student	Categorical
Daytime/evening attendance	Attendance type (daytime or evening)	Categorical
Previous qualification	Type of previous qualification	Categorical
Previous qualification (grade)	Grade of previous qualification	Numerical (Continuous)
Nacionality	Student nationality	Categorical
Mother's qualification	Educational qualification of mother	Categorical
Father's qualification	Educational qualification of father	Categorical
Mother's occupation	Occupation of mother	Categorical
Father's occupation	Occupation of father	Categorical
Admission grade	Admission grade to the program	Numerical (Continuous)
Educational special needs	Whether student has special educational needs	Binary
Gender	Student gender	Binary
Scholarship holder	Whether student is scholarship holder	Binary
Age at enrollment	Age of student at enrollment	Numerical (Discrete)
Displaced	Whether student is displaced from home	Binary
Debtor	Whether student is a debtor	Binary
International	Whether student is international	Binary
Curricular units 1st sem (credited)	Credited units in 1st semester	Numerical (Discrete)
Curricular units 1st sem (enrolled)	Enrolled units in 1st semester	Numerical (Discrete)
Curricular units 1st sem (evaluations)	Number of evaluations in 1st semester	Numerical (Discrete)
Curricular units 1st sem (approved)	Approved units in 1st semester	Numerical (Discrete)
Curricular units 1st sem (grade)	Average grade in 1st semester	Numerical (Continuous)
Curricular units 1st sem (without evaluations)	Units without evaluations in 1st semester	Numerical (Discrete)
Curricular units 2nd sem (credited)	Credited units in 2nd semester	Numerical (Discrete)

Variable	Description	Type
Curricular units 2nd sem (enrolled)	Enrolled units in 2nd semester	Numerical (Discrete)
Curricular units 2nd sem (evaluations)	Number of evaluations in 2nd semester	Numerical (Discrete)
Curricular units 2nd sem (approved)	Approved units in 2nd semester	Numerical (Discrete)
Curricular units 2nd sem (grade)	Average grade in 2nd semester	Numerical (Continuous)
Curricular units 2nd sem (without evaluations)	Units without evaluations in 2nd semester	Numerical (Discrete)
Unemployment rate	Unemployment rate at time of enrollment	Numerical (Continuous)
Inflation rate	Inflation rate at time of enrollment	Numerical (Continuous)
GDP	GDP at time of enrollment	Numerical (Continuous)
Target	Student status (Dropout, Enrolled, or Graduate)	Categorical

Through this step, we didn't encounter any difficult challenges. The dataset was already clean and encoded, so we didn't need to perform variable merging, one-hot encoding or ordinal encoding. We only had to convert categorical variables into readable labels to facilitate our visualization analysis.

4.2.4 - Preprocessing (Data Cleaning and Wrangling)

One of the most important steps in our project is data cleaning and wrangling. After running the code to check for missing values and undefined numerical data, we found that the dataset contains no missing values, no mistakes and no data entry mistakes.

The dataset was already encoded, and we removed "Application mode" and "Tuition fees up to date" variables because they are not relevant to our research questions. Therefore we dropped two columns from the dataset. Ensuring that the numeric columns are numeric, categorical variables such as "Gender", "Debtor", "Displaced" , "Daytime/Evening attendance" were translated to readable string labels for analysis. Although we had a well-structured and clean dataset, our main challenge was to determine the reliability of our dataset. We verified if there were any missing values, spotting mistakes, and determined irrelevant variables for our analysis.

We pursue our cleaning work with the conversion of the categorical variables. Therefore, the reliable dataset was ready to be analyzed.

```
 No missing values found!

Shape after cleaning: (4424, 35)
Missing values: 0
Shape after cleaning: (4424, 35)
Missing values: 0
```

Although we had a well-structured and clean dataset, our main challenge was to determine the reliability of our dataset. We verified if there were any missing values, spotting mistakes, and determined irrelevant variables for our analysis. We pursue our cleaning work with the conversion of the categorical variables. Therefore, the reliable dataset was ready to be analyzed.

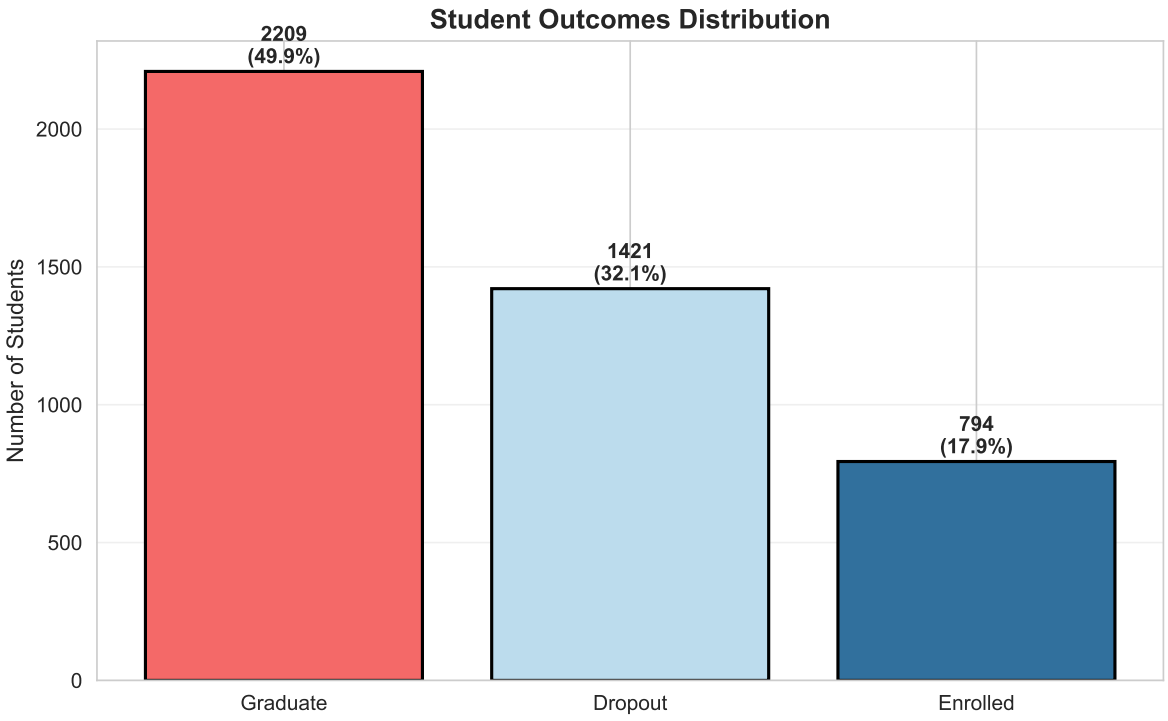
---

5 - Exploratory Data Analysis (EDA)

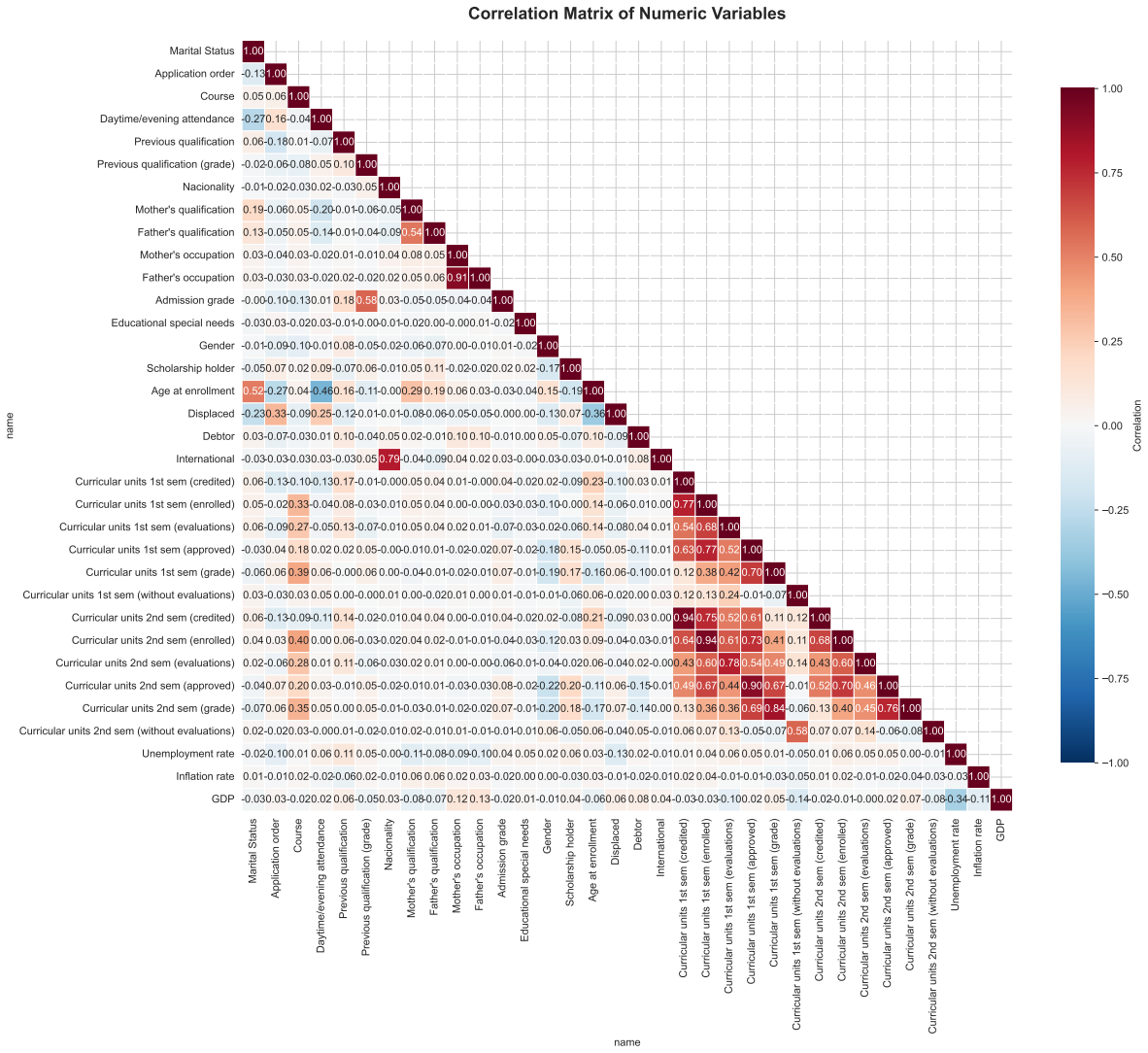
In this section, we explore the dataset to understand the main characteristics of the variables and how they relate to student outcomes (Dropout, Enrolled, Graduate). The goal of the EDA is to identify patterns, detect anomalies, and determine which features are most informative for predicting dropout.

5.1 - Target Variable

We begin by examining the distribution of the target variable. The three student outcomes (Dropout, Enrolled, and Graduate) are highly imbalanced, with Graduates representing the largest group, followed by Dropouts, and a smaller proportion of Enrolled students.


Student Outcomes Distribution
2209
(49.9%)
1421
(32.1%)
794
(17.9%)
Number of Students
0
500
1000
1500
2000
Graduate
Dropout
Enrolled

5.2 - Correlation Analysis



Based on our correlation analysis, we identified several moderately and highly correlated variable pairs that indicate multicollinearity. A high correlation between international and nationality students can be observed, therefore we choose to remove the variable *international*, since it won't be as relevant as the *nacionality* variable. We can see that the variables *father's occupation* and *mother's occupation* are highly correlated, but in this case the correlation reflects social structure. They represent two distinct individuals and two potentially different socioeconomic effects. Same thing applies for *mother's qualification* and *father's qualification*. Although the variables *Curricular units 1st sem (enrolled) Curricular units 2nd sem (enrolled)* and *Curricular units 1st sem (grade) Curricular units 2nd sem (grade)* are respectively highly correlated, we keep them because they provide performance progression across different time

periods, which is relevant for predicting dropout. Therefore, we excluded 8 redundant semester variables and one nationality variable.

5.3 - Feature Selection

```
Removed 9 highly correlated variables
Remaining variables: 26
```

5.4 - Outlier Detection

We implemented a type-aware outlier detection strategy that applies different methods based on the nature of each variable:

Binary variables (e.g., Gender, Scholarship holder): Outlier detection was skipped entirely, as these variables only contain two valid values (0/1).

Nominal categorical variables (e.g., Course, Nationality): No outlier detection applied, as these represent distinct categories without natural ordering. We only reported the number of unique categories present.

Ordinal categorical variables (e.g., qualifications, occupations): We reported the number of levels but did not apply outlier detection, as these represent ordered categories rather than continuous measurements.

Grade variables (0-200 scale): We checked for values outside the valid range (0-200). According to the dataset documentation, grades in the Portuguese system can range from 0 to 200.

Count variables (e.g., enrolled courses): We used a more lenient threshold of 3×IQR (Interquartile Range) rather than the standard 1.5×IQR, as count variables naturally exhibit right-skewed distributions where high values may represent legitimate cases (e.g., students enrolling in many courses).

Continuous variables (e.g., Age, GDP, Unemployment rate): We applied the standard Tukey method with 1.5×IQR threshold to identify potential outliers: values below Q1 - 1.5×IQR or above Q3 + 1.5×IQR.

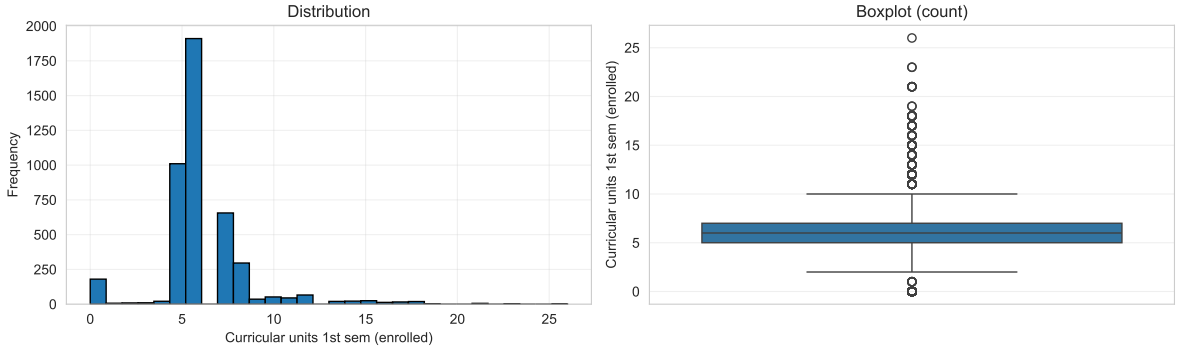
This approach ensures that outlier detection is contextually appropriate for each variable type, reducing false positives while identifying genuine data quality issues.

```
Binary variables (skipping outlier detection):
 - Daytime/evening attendance: values = [np.float64(0.0), np.float64(1.0)]
 - Educational special needs: values = [np.float64(0.0), np.float64(1.0)]
 - Gender: values = [np.float64(0.0), np.float64(1.0)]
 - Scholarship holder: values = [np.float64(0.0), np.float64(1.0)]
 - Displaced: values = [np.float64(0.0), np.float64(1.0)]
 - Debtor: values = [np.float64(0.0), np.float64(1.0)]

Nominal Categorical (no natural order):
 - Course: 17 categories
 - Nacionality: 21 categories

Ordinal Categorical (meaningful order):
 - Marital Status: 6 levels
 - Application order: 8 levels
 - Previous qualification: 17 levels
 - Mother's qualification: 29 levels
 - Father's qualification: 34 levels
 - Mother's occupation: 32 levels
 - Father's occupation: 46 levels

Grade variables (0-200 range + Z-score > 3):
 - Previous qualification (grade): 0 out-of-range + 21 extreme (Z>3) = 21 total (0.5%)
 - Admission grade: 0 out-of-range + 22 extreme (Z>3) = 22 total (0.5%)
 - Curricular units 1st sem (grade): 0 out-of-range + 0 extreme (Z>3) = 0 total (0.0%)
 - Curricular units 2nd sem (grade): 0 out-of-range + 0 extreme (Z>3) = 0 total (0.0%)

Count variables (Z-score > 3):
 - Curricular units 1st sem (enrolled): extreme values: 106 (2.4%)
 - Curricular units 2nd sem (enrolled): extreme values: 82 (1.9%)

Continuous variables (Z-score > 3):
 - Age at enrollment: extreme values: 101 (2.3%)
 - Unemployment rate: extreme values: 0 (0.0%)
 - Inflation rate: extreme values: 0 (0.0%)
 - GDP: extreme values: 0 (0.0%)
```

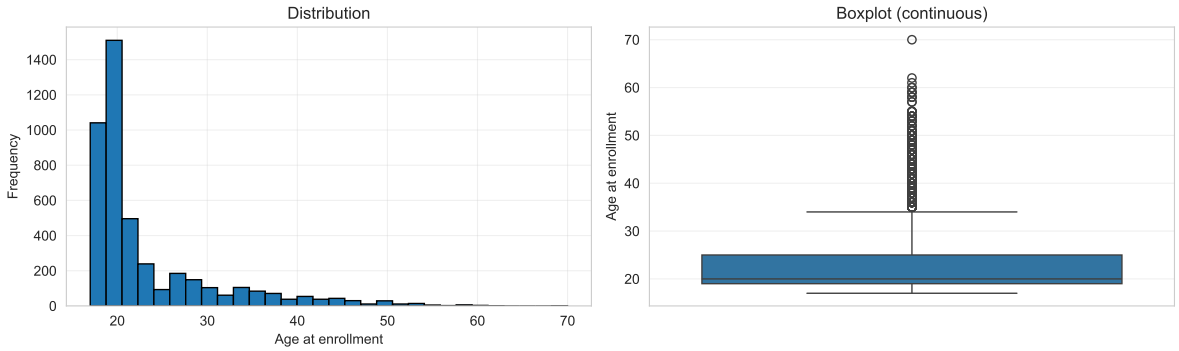
5.4.1 - Outlier Summary

```
Detected Issues:
```

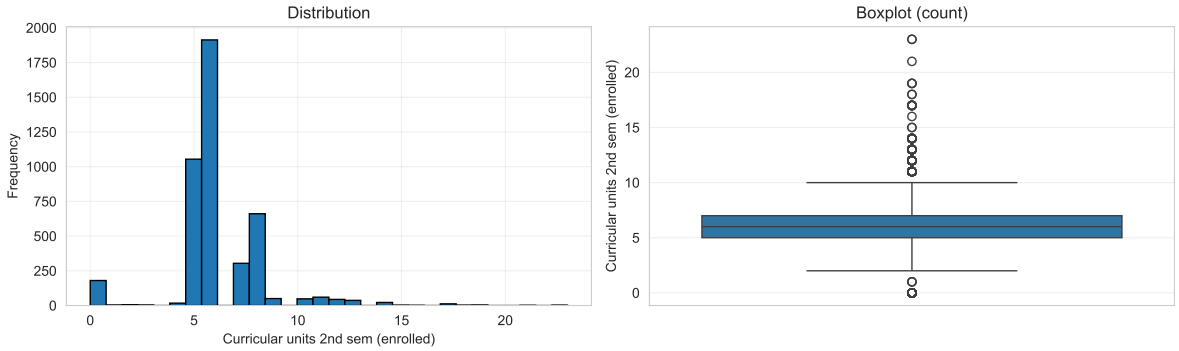
Curricular units 1st sem (enrolled): 106 potential outliers (2.4%)

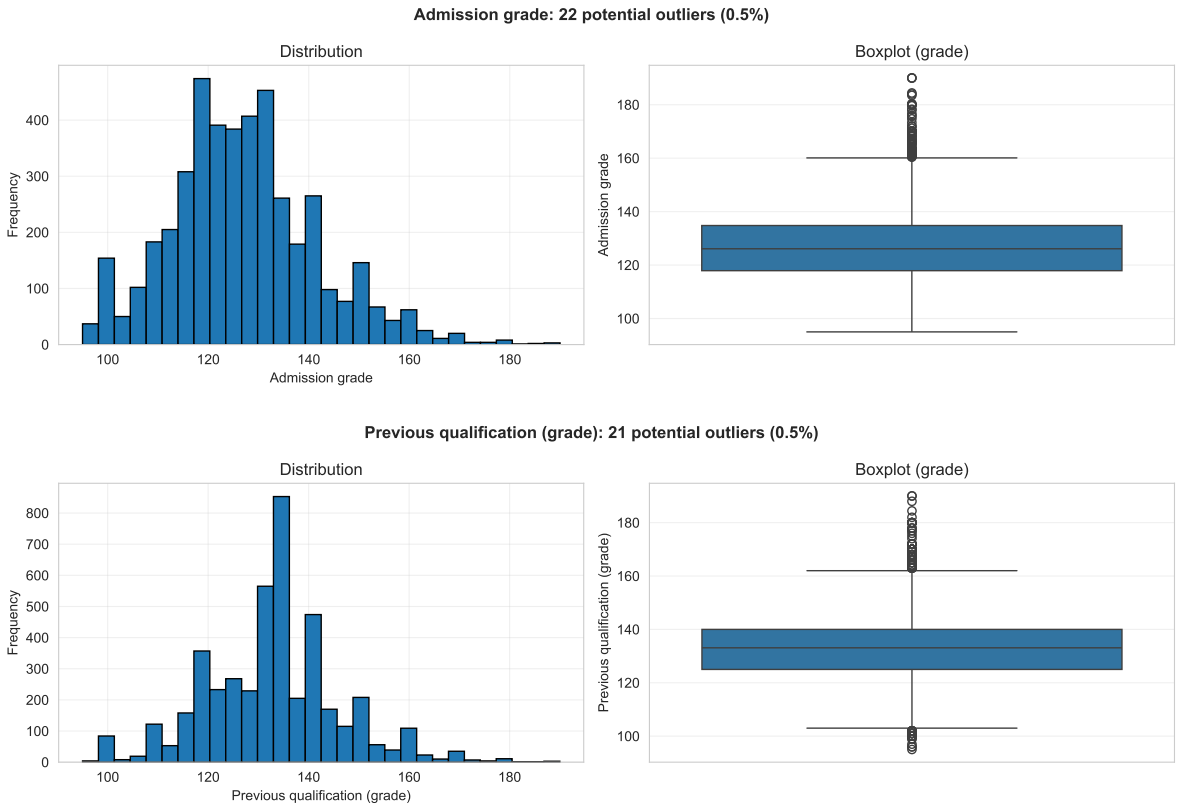


Age at enrollment: 101 potential outliers (2.3%)



Curricular units 2nd sem (enrolled): 82 potential outliers (1.9%)





We identified outliers in five different variables. *Curricular units 1st semester (enrolled)* and *Curricular units 2nd semester (enrolled)*, which represent the **number of courses students register for each semester**. We observed **106 potential outliers** in the **first** curricular semester and **82** in the **second semester**. Since the average course load is usually 5 to 6 classes, students taking a **much higher** or **lower number of courses** are naturally flagged as **outliers**.

In the first semester, the **highest value** reaches **26 classes**. Although this is an **ambitious workload**, it remains **possible**. Several situations could explain such a high number: for instance, a student trying to complete their degree quickly, or a student retaking courses after previous failures. These cases can reflect **meaningful academic behaviours**, so removing them would risk losing useful information.

For the second semester, the **maximum value** is around **20 classes**, leading to **similar conclusions**. In both semesters, we also observe students enrolled in **zero courses**, which appears as an **extreme value** as well. This may correspond to students who completed most of their required courses earlier, or students taking a temporary break while still being officially enrolled. These profiles are **still relevant** and **should be included**.

In this context, **these extreme values are not problematic**. On the contrary, they may help us understand whether taking unusually many, or unusually few, courses has an impact

on college dropout. For this reason, we decided not to remove or cap these observations.

The next variable with detected outliers is *Age at enrollment* , for which **101 potential outliers** were identified. Since the **average age at enrollment** is around **20 years old**, students beginning their studies at 40 or 50 naturally appear as **unusual cases**. The oldest student is 70 years old, which, while rare, is no need for concern regarding methodology within our analysis. Being 70 years old is no different regarding being classified as a student and this data point **should be included**. These values represent **real and meaningful student profiles**, such as mature students or individuals returning to education after a long break. Excluding them would remove **important diversity** from the dataset and limit our understanding of the different types of students who may or may not drop out. For this reason, **we chose not to remove or limit the age-related outliers.**

Finally, outliers were also detected in *Admission grade* (22 cases) and *Previous qualification grade* (21 cases). **These extreme values** reflect either **exceptionally high academic performance** or, conversely, unusually **low grades**. Since these cases may provide **insights into how prior academic achievement relates to dropout behavior**, removing or capping them would not be appropriate. We therefore **opted to retain all outliers in these grade variables** as well.

Based on our **research questions**, we conclude that **removing these outliers would not benefit our analysis**, as they do not represent errors but rather uncommon yet meaningful observations. **Retaining them allows us to capture the full diversity of student profiles** and provides a **more accurate understanding of the factors that may influence college dropout**.

6 - Feature Importance Analysis

6.1 - Methodology (ANOVA)

We used **one-way ANOVA (Analysis of Variance)*** to identify which numeric variables show significant differences across the three target groups (Dropout, Enrolled, Graduate). For each variable, we calculated:

- **p-value**: Statistical significance of differences between groups (= 0.05)
- **Eta-squared (²)**: Effect size measure representing the proportion of variance explained by the target variable (ranges from 0 to 1, where higher values indicate stronger association)

Variables with **p-value $<$ 0.05** are considered **significantly** associated with student outcomes and may be **strong predictors** in classification models.

```
Significant variables (p < 0.05): 21
```

	p_value	eta_sq	significant
Curricular units 2nd sem (grade)	0.000000e+00	0.339086	True
Curricular units 1st sem (grade)	2.803052e-269	0.244020	True
Scholarship holder	4.436825e-94	0.092663	True
Age at enrollment	1.138849e-65	0.065412	True
Debtor	1.018223e-58	0.058620	True
Gender	9.950346e-53	0.052727	True
Curricular units 2nd sem (enrolled)	5.244430e-33	0.033066	True
Curricular units 1st sem (enrolled)	3.272852e-26	0.026197	True
Admission grade	4.380466e-16	0.015871	True
Displaced	2.425582e-13	0.013055	True
Previous qualification (grade)	1.077783e-12	0.012389	True
Marital Status	2.662987e-09	0.008892	True
Application order	2.955293e-09	0.008845	True
Daytime/evening attendance	5.534625e-07	0.006496	True
Mother's qualification	2.800636e-06	0.005767	True

When evaluating feature importance, both statistical significance (p-value) and practical significance (effect size) must be considered. With large sample sizes, even trivial differences can reach statistical significance, making **effect size interpretation essential**.

**Effect Size (²)* measures the proportion of variance in student outcomes explained by each feature, with interpretations:

- **² = 0.01 (1%)** : Small effect
- **² = 0.06 (6%)** : Medium effect
- **² = 0.14 (14%)** : Large effect

For example, marital status as a highly significant **p-value (2.66e-0)**but explains ***less than 1% of variance (² = 0.009)**, indicating negligible practical importance. In contrast, **Curricular units 2nd sem (grade)** explains **34% of variance (² = 0.339)**, representing a large and meaningful effect. When identifying important predictors, the features with **larger effect sizes** should be **prioritized** rather than relying only on p-values.

6.2 - Top Predictive Variables

6.2.1 - Acadamic Performance Indicators

Our exploratory analysis shows relationships between academic performance measures and student outcomes (Dropout, Enrolled, Graduate). Several patterns emerge across **admission grades**, **semester performance**, and **course load**.

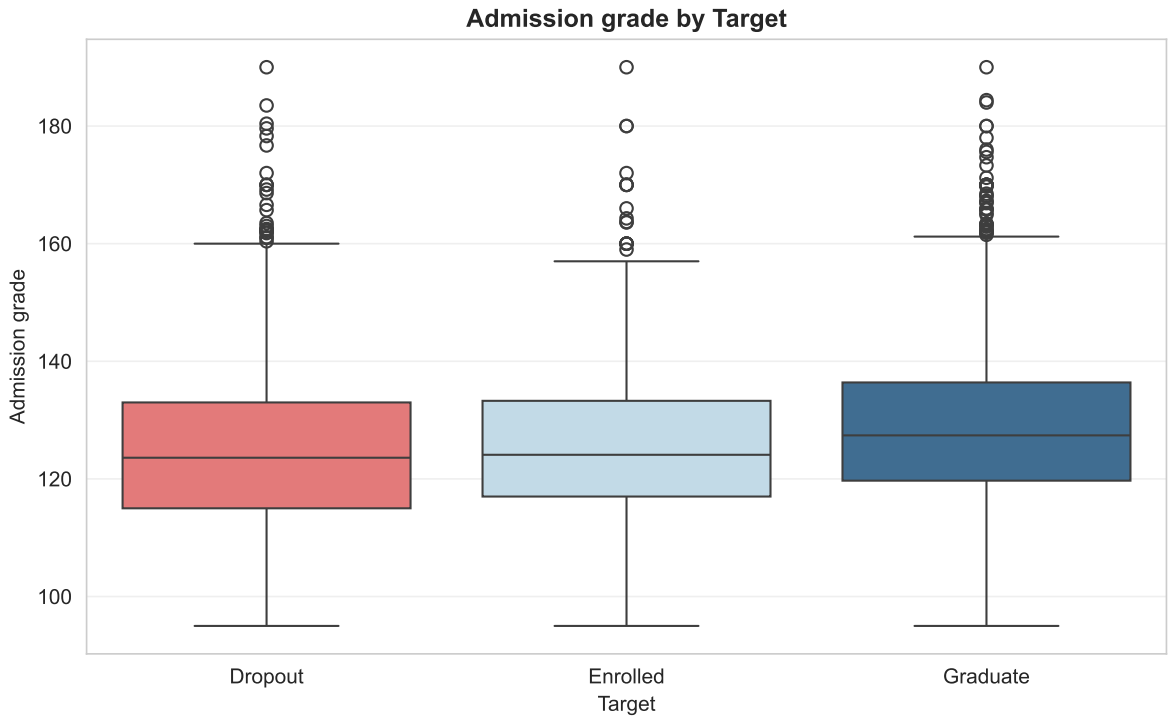


Figure 1: Admission grade by student outcome

In **Figure 1**, we observe the distribution of the admission grade across the three categories (Dropout, Enrolled Target, and Graduate). **Dropout students* have an** average admission grade** of around **122**, with several **outliers** reaching above **160. Enrolled Target students** show a very similar average grade to Dropout students, but with **fewer extreme values**. **Graduate students** display a slightly higher average admission grade, around **125**, and similarly present a few outliers above **160.** Overall, the three groups show **comparable distributions**, with considerable overlap in their admission grades. Graduate students tend to have a marginally higher average, which may suggest that stronger academic preparation is associated with a greater likelihood of graduating. However, the presence of high admission grades in both the Dropout and Graduate categories indicates that good grades alone do not fully determine academic outcomes. In other words, while **admission grade** may play a **role**, it is **not** a **decisive predictor** of whether a student will graduate or drop out.

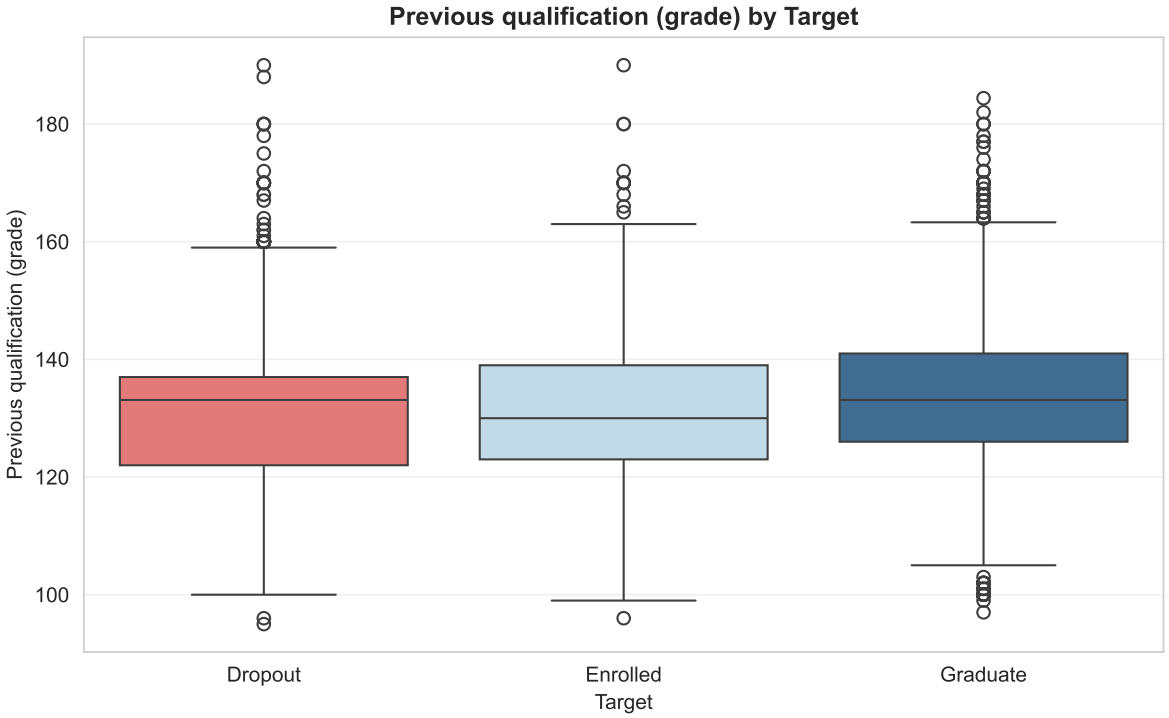


Figure 2: Previous qualification grade by student outcome

Figure 2 shows the distribution of the **Previous qualification (grade)** across the three target students which are Dropout, Enrolled, and Graduate. All three boxplots display similar characteristics, with medians around **130-133**. The minimum and maximum values are also comparable from **100 to 165**. The three groups have multiple outliers at both lower and upper extremes of the grade distribution, dropout and graduates are the one that show more **extreme values**.

For the interpretation, as the distributions and medians are quite similar this suggests that previous qualification grade is not a strong predictor of students' performance. Interestingly the **Dropout group's median** is quite **high** which indicates that students who drop out have not necessarily lower prior grades than those who graduate or stay enrolled.The outliers indicate that in each category there are both very high and very low grades, which suggests that there are other factors **beyond academic performance**.

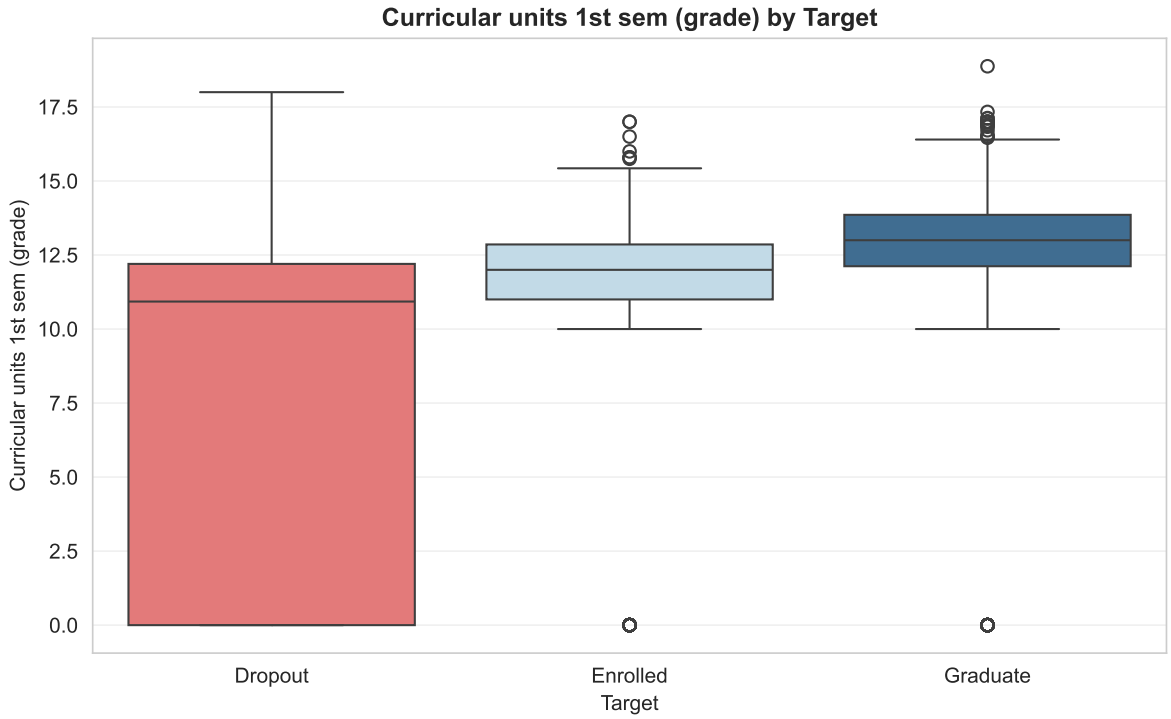


Figure 3: First semester grade by student outcome

In **Figure 3**, we see the first-semester grades for the three target groups: Dropout, Enrolled Target, and Graduate. Dropout students show a wide range of grades. Enrolled Target students have grades around a median of **12.5**, with **moderate spread**. Graduate students have the highest median, around **13.5**, and a **tighter distribution**. For interpretation, the wide spread of Dropout students suggests that leaving the program is not only due to low grades. Enrolled Target students show **average performance**, indicating steady progress but not full completion. Graduate students perform consistently better, suggesting that higher and more stable first-semester grades are associated with graduation.

Figure 4 reveals distinct patterns across the three groups. The Dropout category displays the **widest range of performance**. Enrolled students demonstrate moderate variability with a median near **12 units**. Graduates show the tightest distribution and highest median at approximately **13 units**. By the second semester, the gaps between groups widen. Many

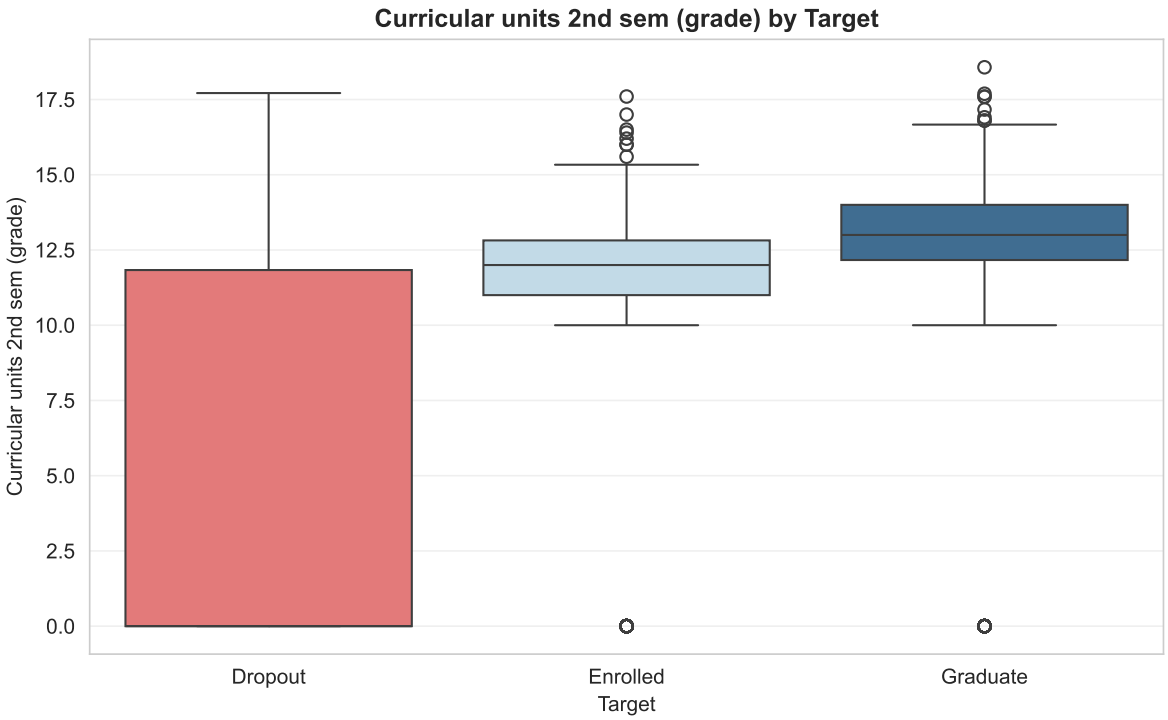


Figure 4: Boxplot of Grades by Target

dropouts completed few or no units (the distribution starts at 0), indicating this is likely when they left the program. Graduates continued performing well with consistent results around **13 units**. Enrolled students fell somewhere in between with decent but **mixed performance**. The second semester appears to be a turning point where struggling students drop out while successful students keep their momentum.

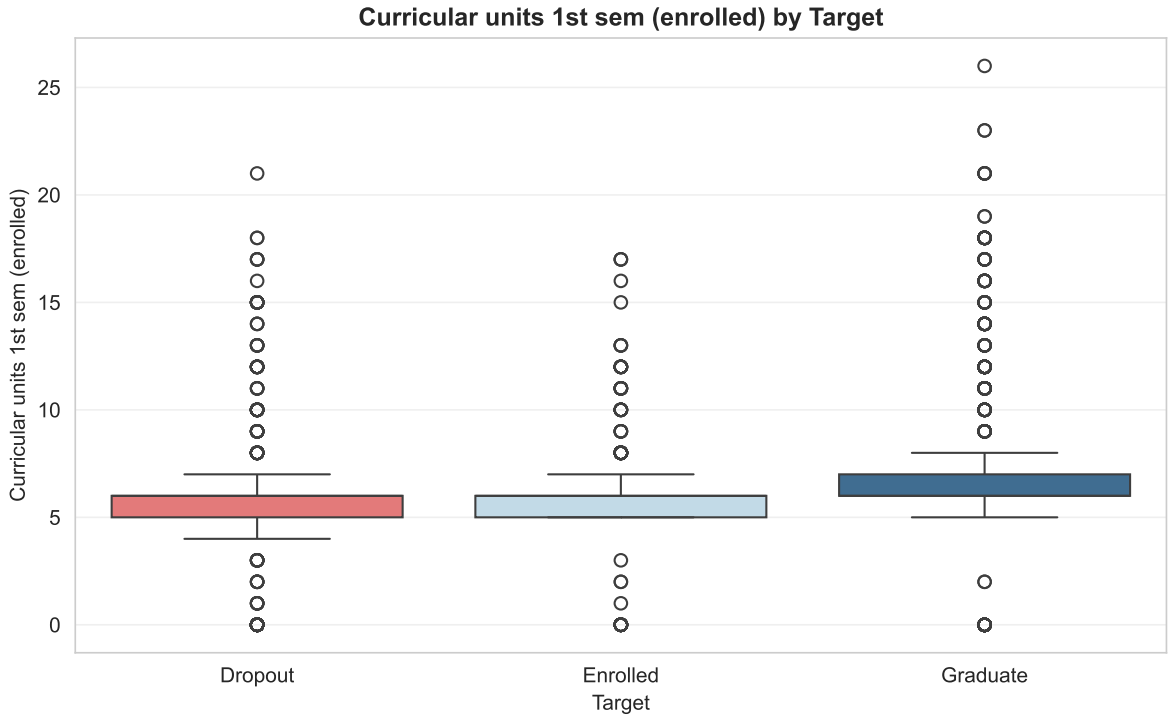


Figure 5: First semester enrollment by student outcome

Figure 5 demonstrates the relationship between **Curricular units 1st Sem (enrolled)** and the three Target outcomes (**Dropout, Enrolled, Graduate**). All three groups show similar box positions with medians around **5-6 units**. Dropouts and Enrolled students have nearly **identical distributions**, while Graduates have a slightly higher box position. All groups show **numerous outliers**, particularly on the upper end, with some students enrolling in **15-26 units**. **Figure 5** reveals that the number of courses taken is not a factor influencing different outcomes, since all groups show **similar enrollment patterns**. Many high outliers appear across all groups, suggesting that ambitious enrollment is common regardless of eventual outcome.

As shown in **Figure 6**, Dropout and Enrolled students have **similar distributions** with their boxes positioned in the lower range. Graduate students show a noticeably **higher box position** and **a wider spread**. All three groups display numerous outliers, particularly on the upper end. Like the 1st semester enrollment patterns, the **2nd semester** shows that

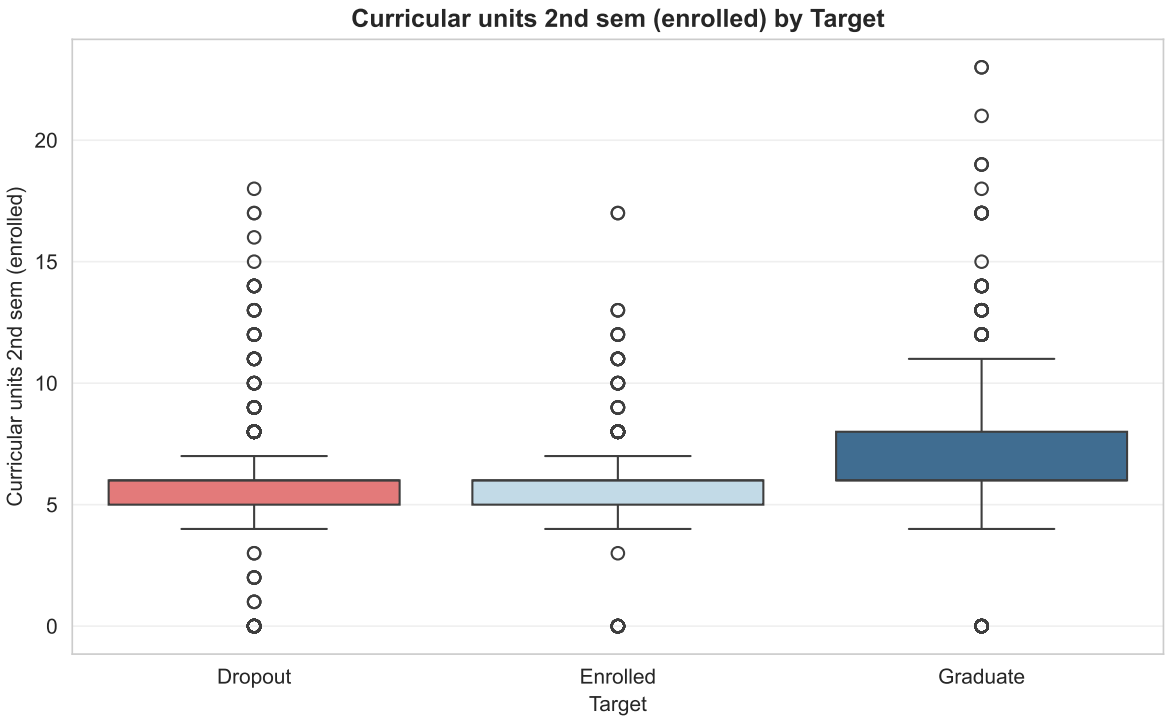


Figure 6: Second semester enrollment by student outcome

graduates tend to **enroll in slightly more courses**, though the differences remain modest. The similar enrollment behavior between dropouts and enrolled students suggests that course load decisions in the 2nd semester don't strongly differentiate these groups - the key difference lies in completion rates rather than enrollment ambitions.

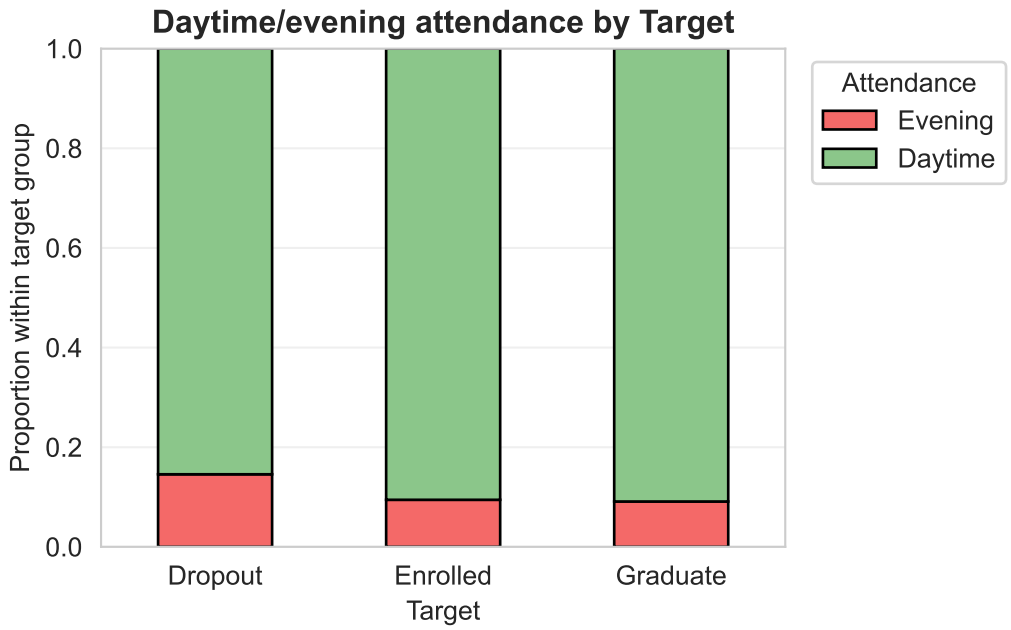


Figure 7: Daytime/evening attendance by student outcome

Figure 7 shows the proportion of daytime and evening attendance within the three groups (Dropout, Enrolled, Graduate). Daytime attendance dominates across all three groups, representing approximately **85-90% of students**. However, Dropout students show a slightly higher proportion of evening attendance (**around 15%**) compared to Enrolled and Graduate students (**around 10%**). This small difference might indicate that evening students face additional challenges, though the similarity across all groups suggests attendance timing is not a primary driver of dropout rates.

Figure 8 shows the **Application order by the three target students**. All three groups show similar distributions and are positioned in the lower range. The medians are approximately **1.5-2** for all categories. The **upper whisker** is similar for the three groups, reaching **3** and the **lower whisker** is at **0** for Graduates and around 1 for Dropout and Enrolled. There are numerous outliers that are at **4, 5 and 6, and even 9** for Enrolled category, indicating that some students applied as their **4th, 5th, 6th and 9th** choice. Regarding the interpretation, as the **distribution is similar** in the three categories this implies that the application order has not a strong relationship with students' success. Most students have applied to this institution as their first or second choice, suggesting that institutional preferences do not really predict if a student will drop out, stay enrolled or graduate. We can also confirm that, as the

Figure 8: Application order by student outcome

outliers are similar, the application order is **not** a **meaningful predictor of a student's performance**.

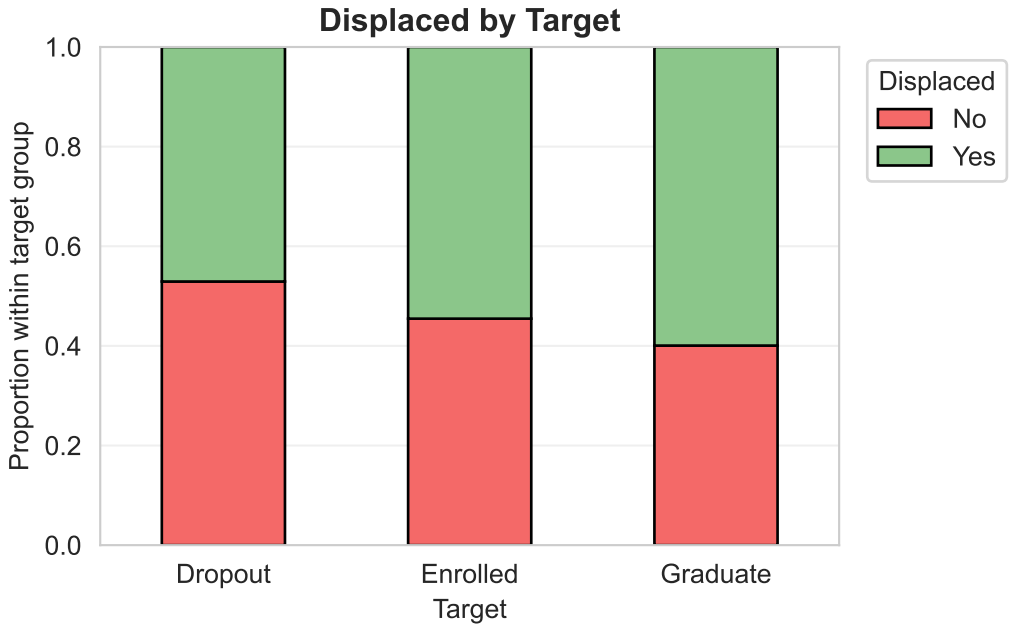


Figure 9: Displaced status by student outcome

Figure 9 shows the **proportion of displaced students** (those who moved or changed residence) across the three target groups. Dropout students have the highest proportion of non-displaced students at around **53%**. Enrolled students show about **45% non-displaced**. Graduate students have the lowest at approximately **40% non-displaced**, meaning **60% of graduates relocated**. The pattern shows that students who relocated for their studies were more likely to graduate. This could be because moving demonstrates **stronger commitment** to education, or because staying home means dealing with work, family responsibilities, or other obligations that interfere with studying. Dropouts were the least likely to have relocated, suggesting that remaining in their original environment may have made it harder to focus on academics.

6.2.2 - Key Findings for Academic Performance and Study Conditions

Graduates have **higher admission grades** and **previous qualification grades** compared to dropouts, though the differences are **relatively small**. This demonstrates that prior academic preparation shows **limited predictive power**.

First-semester grades are the **strongest predictor* of students' performance. Students who drop out show dramatically lower grades (many between** 0-5**), while**

graduates consistently have higher grades (median around 12**).** First semester performance** is therefore a **critical warning signal** for identifying at-risk students.

Graduates tend to enroll in more courses in the first semester (median around **6-7**) compared to those who drop out (median around **5-6**), this may reflect a **stronger initial academic engagement**, even though this difference remains **small**.

Daytime/evening attendance suggests an observable difference and proves to be an important predictor. Evening students show higher drop out rates, around **15%** of dropouts compared to **10%** for graduates. This reflects additional challenges faced by students who must balance work, or family responsibilities with their studies.

Students who are displaced have higher graduation rates (**60%** of graduates vs around **48%** of dropouts). This **counter-intuitive pattern** reflects that relocating for studies may reflect **stronger commitment** or **independence**.

6.2.3 - Demographic & Socioeconomic Background

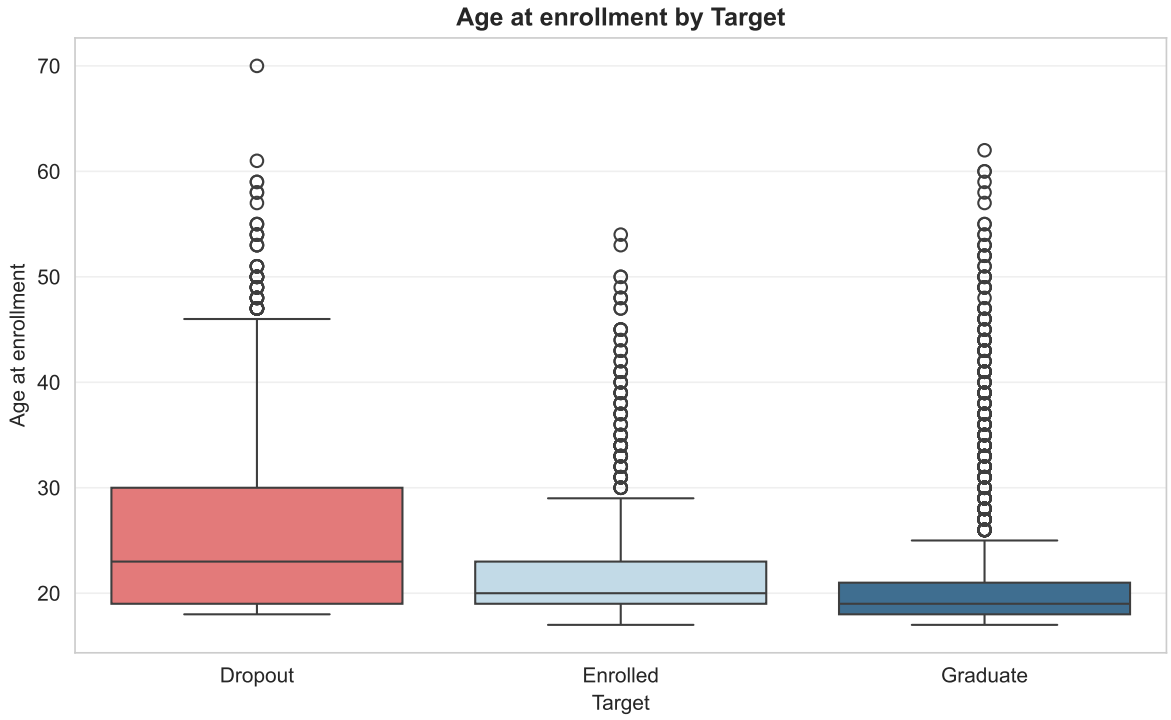


Figure 10: Age at enrollment by student outcome

Figure 10 demonstrates the **relationship between Age at enrollment and the three students outcomes**. The Dropout group has the **highest median age**, which is approximately

23 and the **widest interquartile range**. The Enrolled group has a median age around **20-21**, while the Graduate group shows the lowest median age at around **19**. It is also the narrowest. The three groups contain numerous outliers, showing particularly older students from late 30th to 70 years old.

This suggests that age at enrollment is a **significant predictor of students' performance**. We see that students who enroll at a younger age are more likely to graduate, while older students face more risk of dropping out. This can be caused by several factors, such as the fact that younger students may have fewer external responsibilities compared to older students that may deal with multiple commitments that can interfere with their studies. The wider distribution dropout's group shows that students can occur at any age. However, older students do successfully graduate, showing that age does not determine success alone.

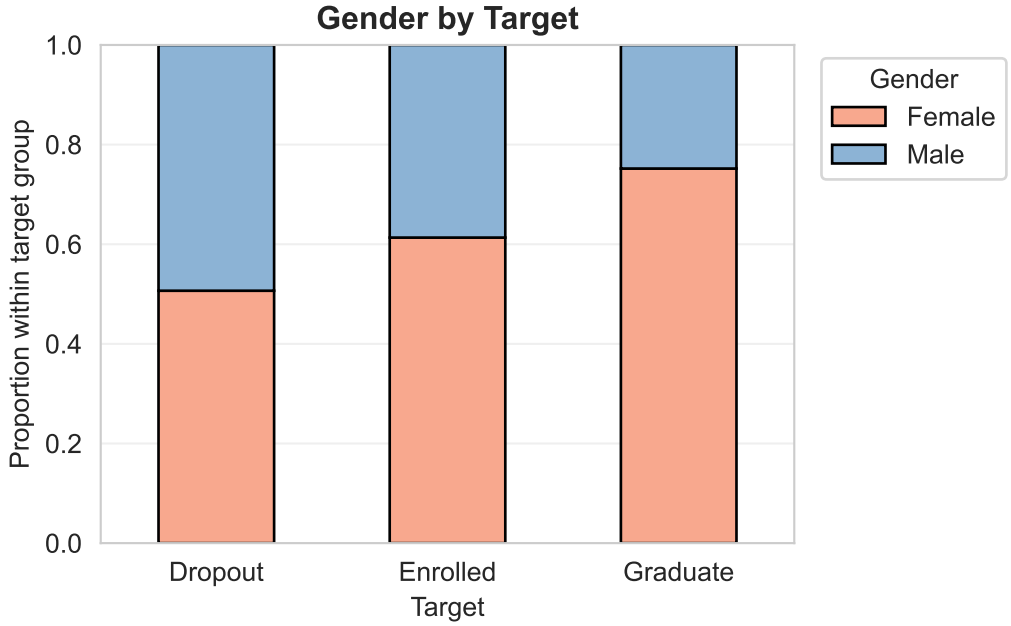


Figure 11: Gender by student outcome

Figure 11 shows the **proportion of genders** within each of the three target groups (Dropout, Enrolled Target, Graduate). In the Dropout group, the proportion of male and female students is almost **equal**. In contrast, both the **Enrolled Target** and **Graduate groups** have a **higher proportion of female students** than male students.

This suggests that female students tend to persist and complete their studies at **higher rates** than male students. Male students appear slightly more likely to interrupt or drop out of their programs, which may contribute to the lower proportions observed in the Enrolled Target and Graduate groups.

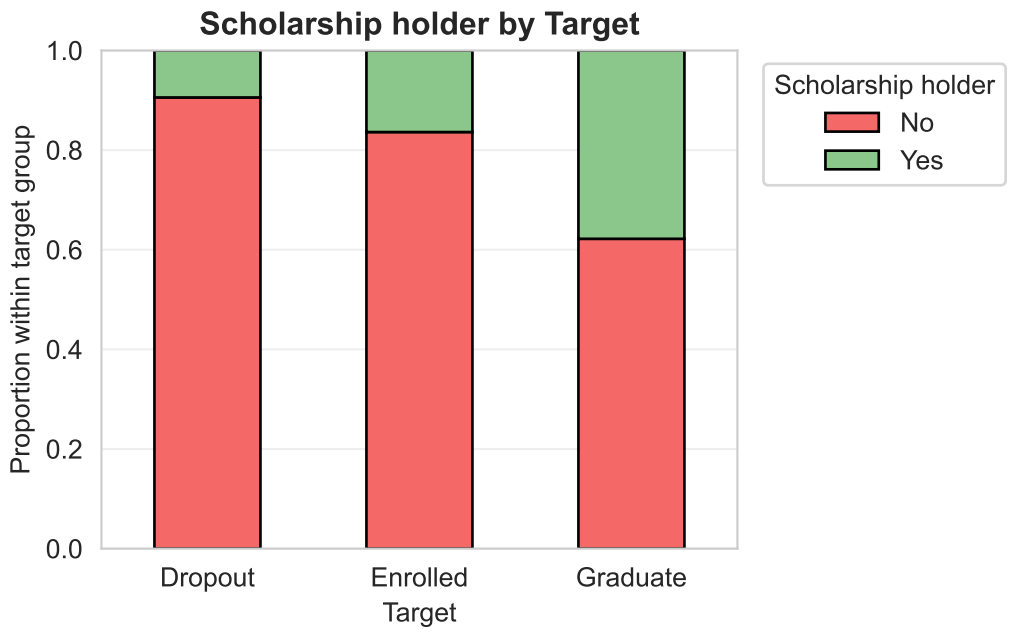


Figure 12: Scholarship holder status by student outcome

Figure 12 shows the **proportion of scholarship holders** within each target group (Dropout, Enrolled Target, Graduate). In both the Dropout and Enrolled Target groups, the vast majority of students do not receive a scholarship, with only a small proportion being scholarship holders. In contrast, the Graduate group contains a noticeably **higher proportion of scholarship recipients**. This suggests that students who receive a scholarship may be more likely to graduate than those who do not. Scholarships often reduce financial pressure and provide support that may help students remain enrolled and complete their studies. Conversely, students without scholarships seem more represented among dropouts and ongoing enrollments.

The **Figure 13** shows that the **dropout group** has the **largest proportion of students who are debtors**. Enrolled students still include some debtors, but the proportion is **noticeably smaller**. In the graduate group, almost all students have no debt, with only a very small fraction appearing as debtors.

This trend suggests that having debt is more common among students who end up dropping out, hinting that financial pressure may contribute to early departure. Conversely, students without debt seem more likely to remain enrolled and reach graduation.

Figure 14 shows the **distribution of marital status across the three target groups** using a jittered scatter plot. While **marital status** achieved **statistical significance** in the **ANOVA test** ($\mathbf{p < 0.001}$), its **effect size** is **negligible** ($^{2} = \mathbf{0.009}$), explaining **less than 1%** of the variance in student outcomes. This **small effect** is evident in the plot, where points align in nearly identical horizontal bands for each category, indicating that the marital status

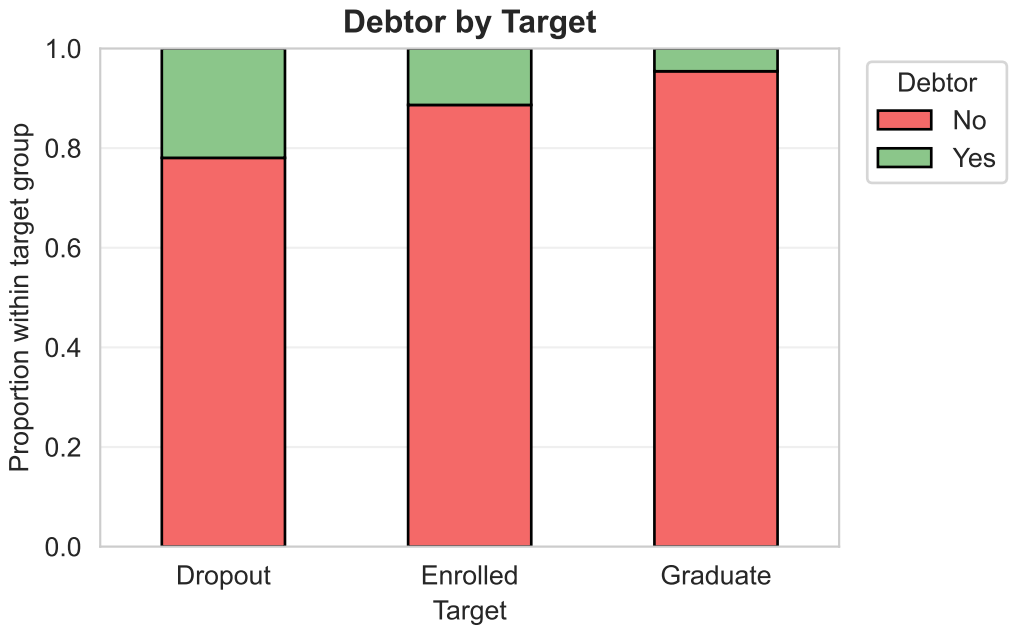


Figure 13: Debtor status by student outcome

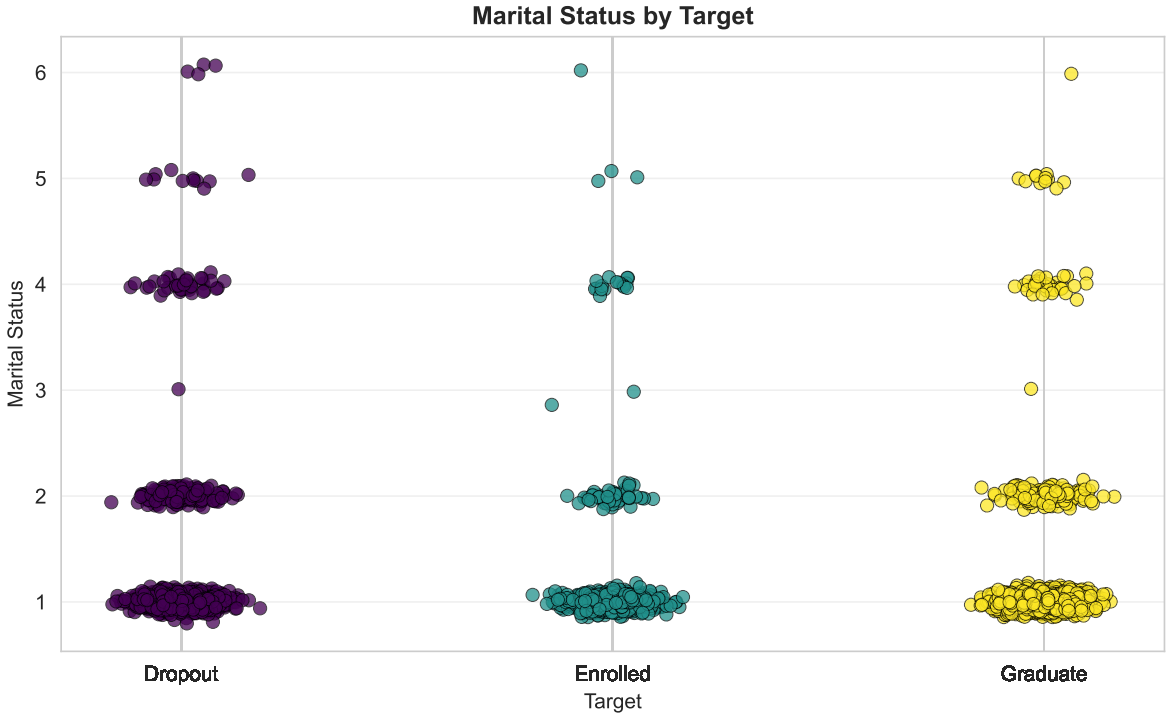


Figure 14: Marital status by student outcome

profiles are essentially the same among Dropout, Enrolled, and Graduate students. The visual clearly shows that the **vast majority of students** are **single**, which is expected given the typical age range of university students. The less common marital statuses married, divorced, widower, facto union, and legally separated—appear only sporadically and are spread evenly across the three groups, further confirming the lack of meaningful differences.

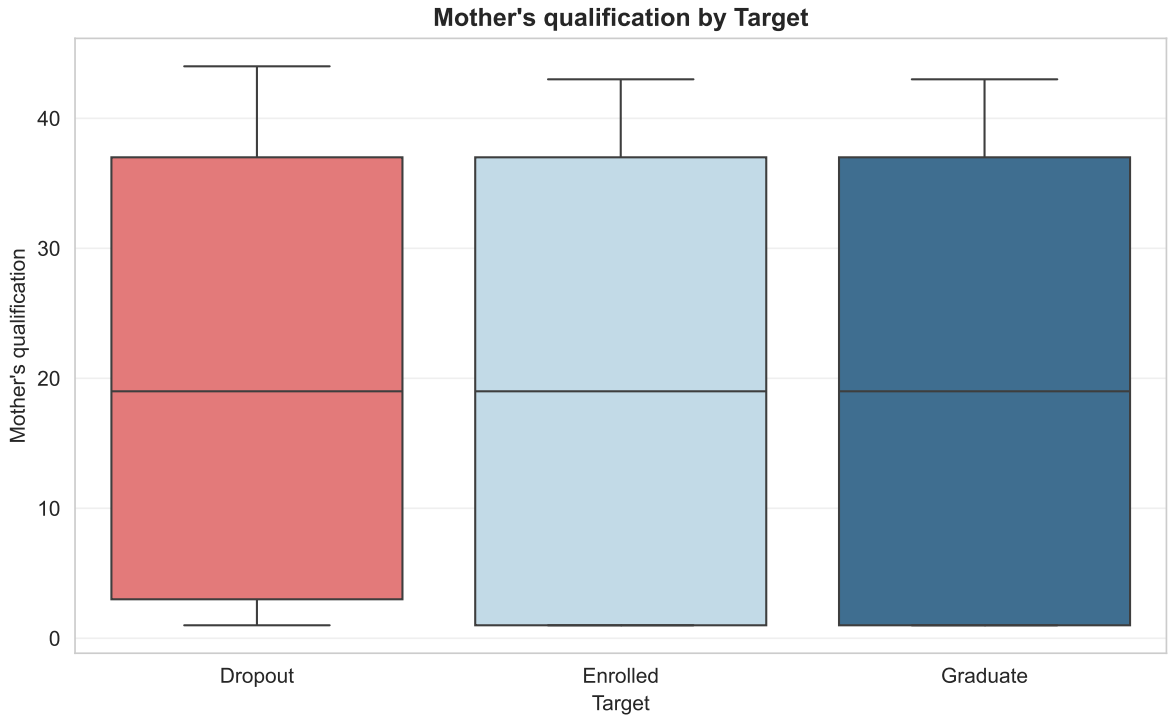


Figure 15: Mother's qualification by student outcome

Figure 15 shows the **Mother's qualification** across the three target groups. Enrolled and Graduate **distributions** are **identical**. The **three groups** display **nearly identical distributions** with all median at approximately **19**. However, the **Dropout group** shows a **slightly wider interquartile range**, extending lower compared to the two other groups. They reach **similar limits** and there are **no outliers**.

While the three distributions are very similar, the dropout category shows slightly fewer students with mothers who have very low qualifications, but this is minimal. The overall similarity indicates that **a mother's qualification has little influence** on whether students complete their studies. This is also supported by the **small effect size**.

6.2.4 - Key Findings for Demographic & Socioeconomic Background

Younger students are **more likely to graduate** while **older students** face **higher dropout risk**. The **dropout group** also shows the**widest age variation**. This suggests that **older students may face competing life responsibilities** that interfere with their studies. **Female students graduate at slightly higher rates**, making up a larger proportion of both enrolled and graduate groups compared to dropouts (50-50 split).

Graduates have a noticeably higher proportion of **scholarship recipients** compared to **dropouts** and **enrolled students**, most of whom do not receive scholarships. This suggests that **financial support may help students complete their studies**. **Dropouts have the largest proportion of students in debt**, while enrolled students include some debtors but fewer than dropouts. Graduates have almost no debt, indicating that financial pressure is strongly associated with dropout, whereas **financial stability aligns with graduation**.

All three groups show **nearly identical distributions for marital status**, with outliers appearing equally across categories. This suggests that **marital status has little to no relationship with student outcomes** and holds no predictive value.

The groups also display nearly identical distributions in terms of age, with medians around 19. The dropout group shows a slightly wider spread extending lower, but the difference is minimal. Finally, **mother's qualification** appears to have **little to no influence** on whether students complete their studies.

7 - Predictive Modelling

In this section, we begin building a predictive model aimed at understanding which factors are most strongly associated with student dropout. Our goal is not only to classify students into the three outcome categories (Dropout, Enrolled, Graduate), but mainly to identify which variables contribute most to the **risk of dropping out**.

We train a **Random Forest classifier** as a first baseline model. This allows us to evaluate predictive performance and obtain a first indication of which features may be important. To further interpret and validate these results, we use **LIME explanations**, both at the individual level (example students) and globally across multiple samples.

This modelling part is therefore an exploratory step toward identifying meaningful patterns, influential academic or demographic factors, and evaluating which features could be most relevant for predicting student success or failure. Later, these insights can be refined and made more specific to dropout prediction.

7.1 - Model Performance

Overall Accuracy: 0.671 (67.1%)

The Random Forest classifier achieved an overall accuracy of 67.3%, correctly predicting students' academic performance in approximately two-thirds of cases. However, accuracy by itself is not sufficient, and it is necessary to examine the model's performance for each class.

Performance by Class

	precision	recall	f1-score	support
Dropout	0.774	0.651	0.707	284
Enrolled	0.340	0.459	0.390	159
Graduate	0.780	0.760	0.770	442

Global Performance

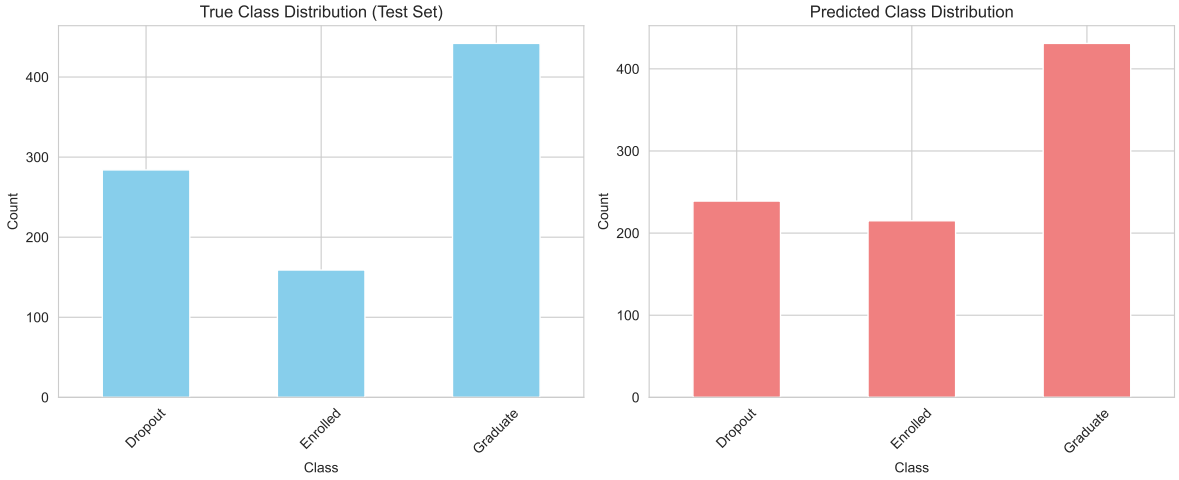
	precision	recall	f1-score	support
accuracy	0.671	0.671	0.671	0
macro avg	0.631	0.624	0.623	885
weighted avg	0.699	0.671	0.682	885

Performance varies significantly across the three outcomes. The model performs best at identifying Dropout and Graduate students, with F1-scores of **0.707** and **0.773** respectively. However, it struggles with the Enrolled category, achieving only an F1-score of **0.392**. This means that the model has difficulty distinguishing between currently enrolled students and those who will drop out or graduate.

The confusion matrix reveals specific prediction patterns. The model correctly identifies 185 out of 284 dropouts (65.1% recall) and 338 out of 442 graduates (76.5% recall). The struggle with Enrolled students is evident: only **73** out of **159** (45.9%) are correctly classified, with many being misclassified as either future graduates (53 cases) or potential dropouts (33 cases). This reflects the inherent difficulty in predicting outcomes for students still in progress, their final status remains uncertain until they complete their program or drop out.

Handling class imbalance. With a smaller number of Enrolled students in the dataset (159 vs 284 Dropouts and 442 Graduates), the model has **less training data to learn patterns** for this group. To address this imbalance, we used the `class_weight='balanced'` parameter in the Random Forest model, which automatically adjusts weights inversely proportional to the

frequency of classes. This ensures the model gives equal attention to all three classes during training rather than being biased toward the majority class. While this helps mitigate the imbalance, the challenge remains: enrolled students represent an **"in-between"** state that is harder to characterize than the final outcomes of dropping out or graduating.

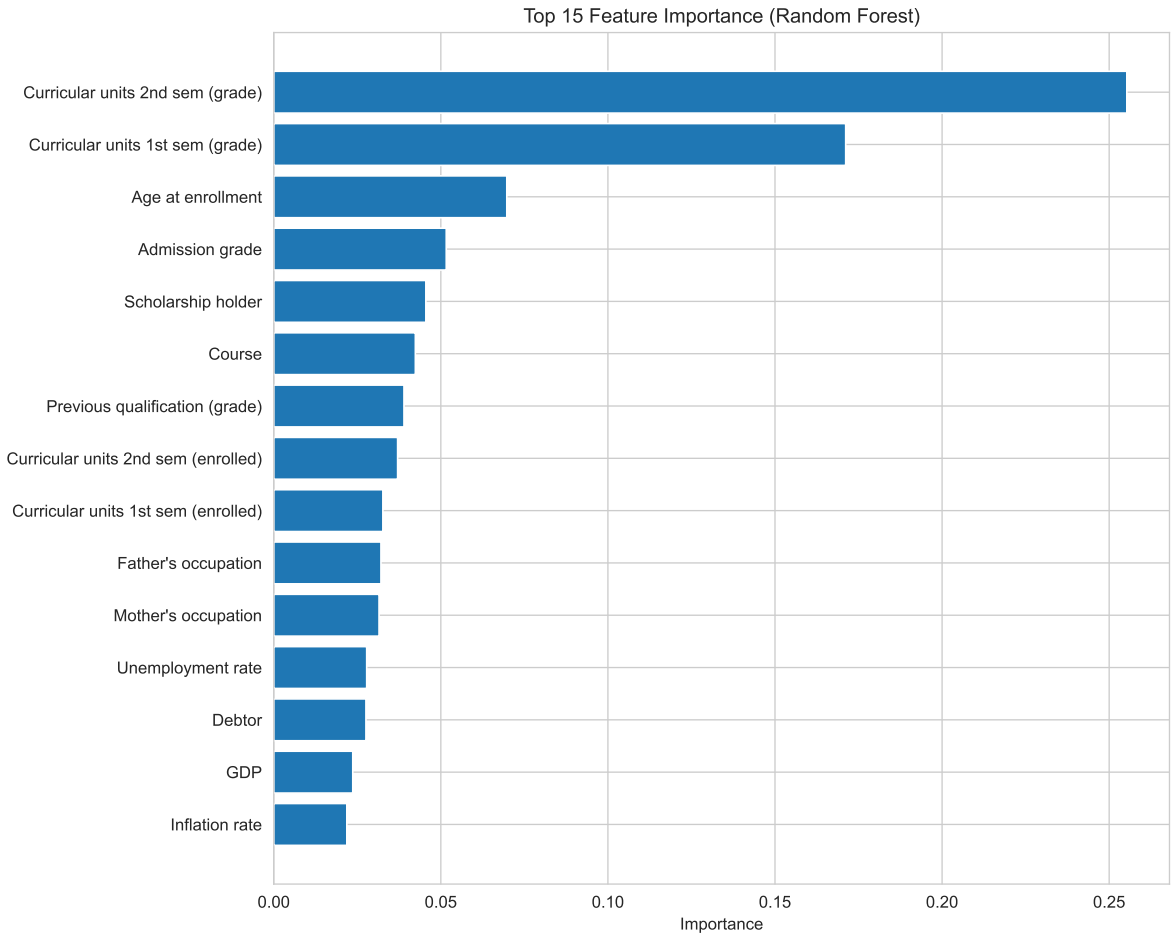


Predicted distribution. The last figure shows the predicted distribution of classes after applying the model. Despite the class imbalance in the original data, the model with balanced class weights manages to capture quite well the real distribution, proving that the balancing strategy is effective.

7.2 - Random Forest Feature Importance

Random Forest models calculate **feature importance** by measuring how much each feature contributes to reducing prediction error across all decision trees in the forest. Features that consistently lead to better splits and more accurate predictions receive **higher importance scores**. It measures each variable's contribution to prediction while considering all other features, capturing interactions and non-linear relationships.

7.2.1-Top 15 Most Important Features (Random Forest)



Even though Random Forest differs from ANOVA by evaluating variables in combination rather than in isolation, results obtained in both cases show that **academic performance variables** (*1st and 2nd semester grade*) are dominant.

Variables like financial or administrative factors showed weak effects in ANOVA but appear among the top 15 Random Forest features because they interact with academic variables to improve predictions. Therefore, these variables reveal conditional effects such as financial stress that can increase academic risk when performance is already low. Thus, these interactions enhance both predictive accuracy and the interpretation of underlying patterns in the data.

7.3 - LIME Explanations (Individual Examples)

Showing one representative example from each class

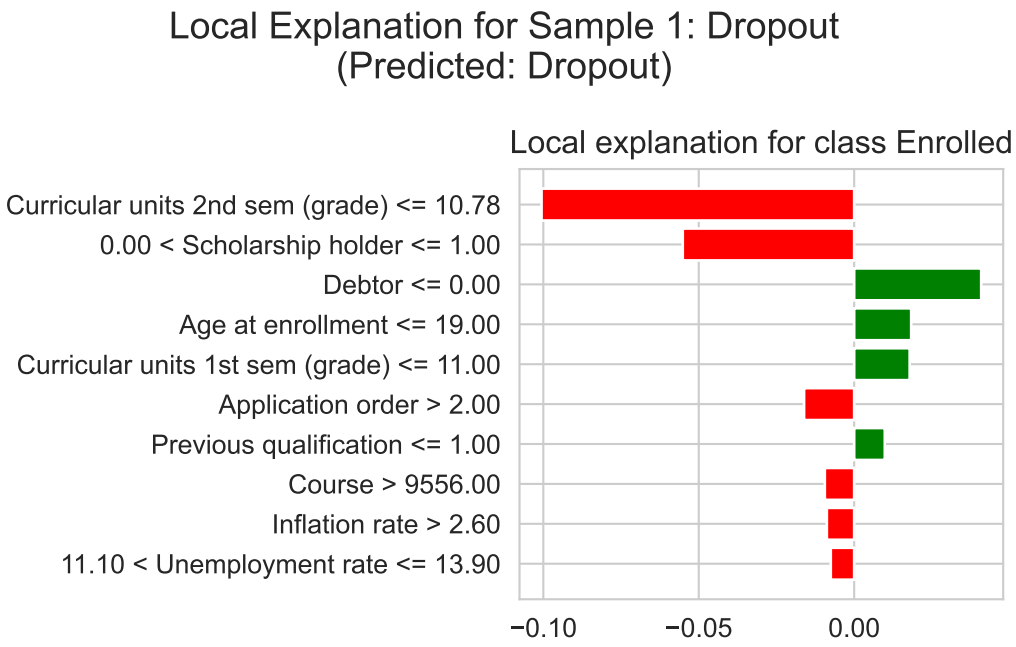
LIME (Local Interpretable Model-agnostic Explanations) helps us understand **why the model made a specific prediction for an individual student**. Unlike feature importance which shows what matters globally across all predictions, LIME reveals which features drove the decision for this particular case. The tables below show the **top 10 features that most influenced this prediction**. Positive weights push the prediction toward the predicted class, while negative weights push against it. The magnitude indicates the strength of influence.

Sample 1: Dropout

	Metric	Value
0	True Label	Dropout
1	Predicted Label	Dropout
2	Dropout Prob	0.834
3	Enrolled Prob	0.108
4	Graduate Prob	0.058

Top 10 features influencing this prediction:

	Feature	Weight
0	Curricular units 2nd sem (grade) <= 10.78	-0.101
1	0.00 < Scholarship holder <= 1.00	-0.055
2	Debtor <= 0.00	0.041
3	Age at enrollment <= 19.00	0.018
4	Curricular units 1st sem (grade) <= 11.00	0.018
5	Application order > 2.00	-0.016
6	Previous qualification <= 1.00	0.010
7	Course > 9556.00	-0.009
8	Inflation rate > 2.60	-0.009
9	11.10 < Unemployment rate <= 13.90	-0.008



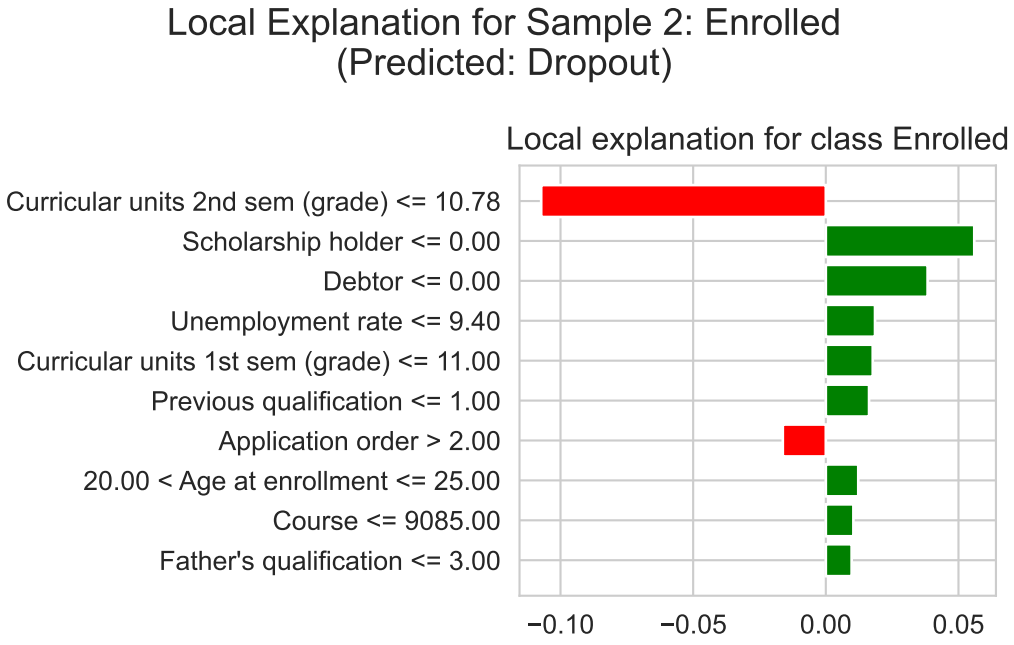
Interpretation: This student was correctly identified as at risk of dropping out. The dominant factor is their **poor academic performance**, with a second semester grade of 10.78 or below strongly pushing toward dropout (large negative weight). Additionally, lacking **scholarship support** increases dropout risk. Protective factors such as not being in debt, younger age (19), and decent first semester grades (11) provide some counterbalance, are not enough to overcome poor results in the second semester. This highlights how academic difficulties, especially in later semesters, are key indicators of the risk of dropping out.

Sample 2: Enrolled

	Metric	Value
0	True Label	Enrolled
1	Predicted Label	Dropout
2	Dropout Prob	0.677
3	Enrolled Prob	0.279
4	Graduate Prob	0.044

Top 10 features influencing this prediction:

	Feature	Weight
0	Curricular units 2nd sem (grade) <= 10.78	-0.107
1	Scholarship holder <= 0.00	0.056
2	Debtor <= 0.00	0.038
3	Unemployment rate <= 9.40	0.019
4	Curricular units 1st sem (grade) <= 11.00	0.018
5	Previous qualification <= 1.00	0.016
6	Application order > 2.00	-0.016
7	20.00 < Age at enrollment <= 25.00	0.012
8	Course <= 9085.00	0.010
9	Father's qualification <= 3.00	0.010



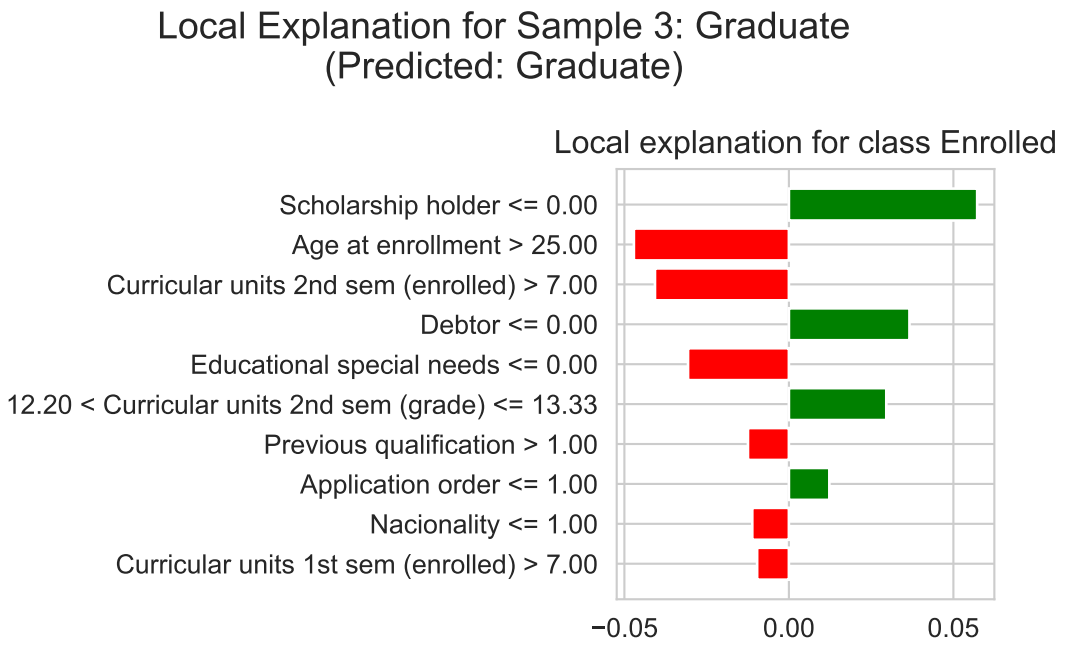
Interpretation: This enrolled student presents a highly **mixed profile** that makes prediction challenging. **Poor second semester grades** (10.78) strongly push toward dropout, while positive factors like having **no scholarship (paradoxically protective here), not being in debt, low unemployment rate, and moderate academic performance in the first semester push toward remaining enrolled or graduating**. The model shows **significant uncertainty**, with multiple features pulling in different directions. This reflects the inherent difficulty in predicting outcomes for students still in progress, they're in a transitional state where their trajectory could go either way.

Sample 3: Graduate

	Metric	Value
0	True Label	Graduate
1	Predicted Label	Graduate
2	Dropout Prob	0.280
3	Enrolled Prob	0.231
4	Graduate Prob	0.489

Top 10 features influencing this prediction:

	Feature	Weight
0	Scholarship holder <= 0.00	0.057
1	Age at enrollment > 25.00	-0.047
2	Curricular units 2nd sem (enrolled) > 7.00	-0.041
3	Debtor <= 0.00	0.037
4	Educational special needs <= 0.00	-0.031
5	12.20 < Curricular units 2nd sem (grade) <= 13.33	0.030
6	Previous qualification > 1.00	-0.012
7	Application order <= 1.00	0.012
8	Nacionality <= 1.00	-0.011
9	Curricular units 1st sem (enrolled) > 7.00	-0.010



Interpretation: This student shows **mixed but ultimately positive indicators** of academic success. While poor second semester grades (10.78) push against graduation, several protective factors dominate: having **no scholarship, not being in debt, low unemployment rate, being in a "traditional" age range (20-25), and having moderate previous qualifications** all strongly predict graduation. The combination of financial stability (no debt) and adequate academic preparation outweighs the concerning second semester performance, allowing the model to confidently predict graduation. This demonstrates that graduation is driven by multiple factors beyond just grades alone.

7.4 - Global LIME Feature Importance

ADD A SMALL TRANSITION

Top 15 Most Important Features (LIME Global)

	feature	lime_importance
21	Curricular units 2nd sem (grade)	0.0637
14	Scholarship holder	0.0546
17	Debtor	0.0363
19	Curricular units 1st sem (grade)	0.0284
15	Age at enrollment	0.0245

	feature	lime_importance
20	Curricular units 2nd sem (enrolled)	0.0162
1	Application order	0.0124
22	Unemployment rate	0.0099
2	Course	0.0094
12	Educational special needs	0.0090
18	Curricular units 1st sem (enrolled)	0.0090
11	Admission grade	0.0071
4	Previous qualification	0.0069
6	Nacionality	0.0066
8	Father's qualification	0.0062

Global LIME Feature Importance: When aggregating LIME explanations across all three classes (Dropout, Enrolled, and Graduate), we obtain a **global view** of which features most consistently **influence the model's predictions regardless of outcome**. The ranking remains similar to the dropout-specific analysis, with **second semester grades** (0.0583) again dominating as the single most important predictor. **Scholarship holder status** (0.0527) and **debtor status** (0.0369) maintain their strong positions, reinforcing that financial factors and academic performance are universally important across all prediction scenarios.

The consistency between dropout-specific and global feature importance suggests that the same fundamental factors drive all student outcomes, just in different directions. Strong academic performance and financial stability predict graduation, while their absence predicts dropout. Enrolled students fall somewhere in between, showing mixed patterns of these key indicators. This global perspective validates our earlier findings and demonstrates that interventions targeting academic support and financial aid would benefit students across all outcome categories, not just those at risk of dropping out.

Where does the Model struggle?

Confusion patterns in LIME sample:

	True Class	Pred: Dropout	Pred: Enrolled	Pred: Graduate
0	Dropout	185 (65%)	57 (20%)	42 (15%)
1	Enrolled	33 (21%)	73 (46%)	53 (33%)
2	Graduate	21 (5%)	85 (19%)	336 (76%)

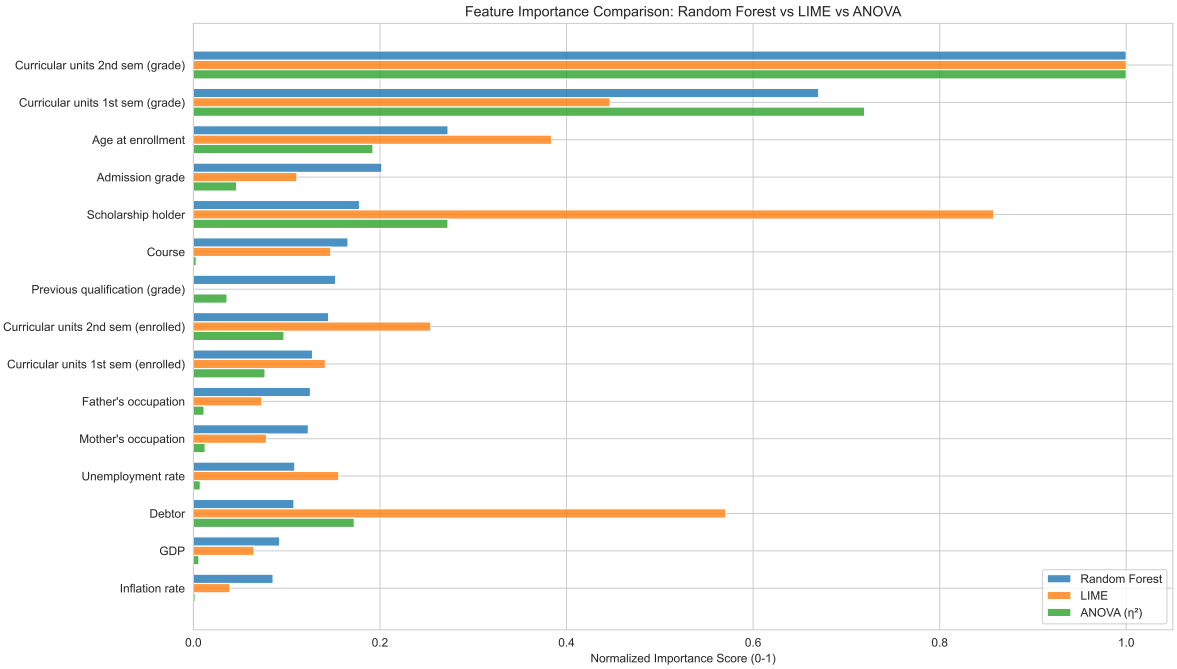
Key Issues:

- **Dropout** students are often misclassified as **Enrolled** (57/284 cases, 20%)

- **Enrolled** students are often misclassified as **Graduate** (53/159 cases, 33%)

Due to the class imbalance the model can struggle, since most of the cases people are Graduated, there is a bias towards that group in the data. Hence, it is more likely to predict that for cases when it is hesitant.

7.5 - Feature Importance Comparison: Random Forest, LIME, and ANOVA



Comparing Three Approaches to Feature Importance:

This comparison reveals how different methods assess feature importance:

- **Random Forest (blue)**: Global feature importance based on how much each feature reduces prediction error across all decision trees. Captures complex interactions and non-linear relationships.
- **LIME (orange)**: Local feature importance averaged across 100 samples from each class. Explains individual predictions by measuring how changes in feature values affect model outputs locally.
- **ANOVA ^{2} (green)**: Statistical effect size measuring the proportion of variance each feature explains in the target variable. Captures linear relationships and univariate associations.

Key Observations: All three methods agree that ***2nd semester grades*** and ***1st semester grades*** are the most important predictors, validating academic performance as the dominant factor. However, they diverge on other features: - ***Scholarship holder*** and ***Debtor*** rank highly in LIME (practical prediction influence) but lower in Random Forest, suggesting these features work through interactions rather than independently. - **ANOVA** identifies strong linear relationships (high ^{2} for grades) but may underestimate features that work through complex interactions. - **Random Forest** balances both direct effects and feature interactions, providing a comprehensive view of predictive power in the classification context.

8 - Conclusion

This project aimed to understand the factors driving student dropout and academic success. The objective was also to assess whether these outcomes can be predicted accurately, using demographic, socioeconomic, and academic information. Therefore, we were able to answer our research questions through a combination of exploratory data analysis and predictive modeling techniques.

1. **How students' academic indicators influence their likelihood of graduating or dropping out.**

The analysis indicates that **academic performance indicators** are the **strongest determinant of students' final academic outcomes**. Second-semester grades constitute the **most decisive factor** in distinguishing between students who drop out, remain enrolled, or graduate, as reflected by a very large **ANOVA effect size** ($^2 = 0.339$). **First-semester grades** are used as a **critical warning signal** for identifying students in difficulty. In contrast, indicators of prior academic preparation and course enrollment intensity show **lower significance**, which suggests that success at the university level relies more heavily on student's adaptation and performance within the academic environment than on their background before university.

Study conditions contribute to outcomes in a more nuanced way. Students attending evening programs face slightly higher dropout risk, likely due to external commitments, while students who relocate for their studies appear more likely to graduate, possibly reflecting stronger educational commitment or fewer competing obligations.

Overall, these results point to a **critical transition** between the **first and second semesters**. While early academic difficulties can be identified shortly after enrollment, **second-semester performance** has a **major role** in determining whether they rebound or lose interest. From a practical perspective, this emphasizes the **importance of early academic monitoring** and **targeted interventions** during the first year to reduce dropout risk and promote student success.

2. **What is the impact of demographic and socioeconomic background on students' probability of dropping out?**

a. **To what extent do financial factors (debtor status, scholarship holder) affect student retention?**

While **demographic and socioeconomic factors** have an **impact on student outcomes**, it is not as significant compared to academic performance. A clear pattern is observed for *Age at enrollment*: younger students are more likely to graduate, while older students face a higher risk of dropout (^{2} 0.065). However, age alone does not determine success. We also observe gender differences, with **female students** being **more likely** to **remain enrolled** and **graduate**, despite the effect size being moderate.

Financial factors also show a **clear correlation** with **student outcomes**. Dropping out indeed correlates with indebtedness. Students with debt are more likely to drop out, whereas those who graduate are less concerned about it. These findings support the **LIME analysis**, which shows that **debt and scholarships** are two of the **most important non-academic factors**: financial issues increase the risk of dropping out, while financial support improves retention.

Considering other demographic variables such as marital status , application order and parental background, their **effect sizes** are **very small**, their distributions are similar across outcome groups and model-based explanations consistently rank them among the least important predictors. Therefore, although they are **statistically significant**, their effects are **negligible** and do not meaningfully explain differences in student outcomes.

The findings indicate that **socioeconomic factors**, such as financial stress (*debt*) and financial support (*scholarships*), have a **significant impact** on student retention. Whereas demographic characteristics such as age and gender play a secondary role compared to academic performance and financial stability.

3. **Can we accurately predict a student's final status (Dropout, Enrolled, or Graduate), and which characteristics are most relevant?**

    a. **Which features category, academic, socioeconomic or demographic contribute the most in predicting students' dropout?**

While the previous research questions focused on identifying and interpreting the individual effects of academic, socioeconomic, and demographic factors on student dropout, the final step in the analysis evaluates these factors jointly within a predictive modeling framework. This shift from explanation to prediction assesses how well a machine-learning model can classify students' final academic status and which categories of variables contribute most to its performance.

The results indicate that predicting students' final academic status is feasible, though subject to important limitations. The **Random Forest model** performs **well** for students with clearly defined outcomes, particularly graduates and dropouts, while currently enrolled students are harder to classify, reflecting the inherent uncertainty of trajectories that are still in progress rather than shortcomings of the model itself.

Regarding **predictor relevance**, a clear hierarchy emerges consistently across all methods. **Academic features** dominate predictive performance, accounting for approximately **60–70%** of **explanatory power**, with second semester grades and first semester grades ranking **highest** across **ANOVA**, **Random Forest importance**, and **LIME explanations**. **Socioeconomic factors** form the **second most influential** category (around **20–30%**), with scholarship holding and debtor status standing out as key non-academic predictors that condition students' ability to sustain academic performance. Demographic characteristics contribute more modestly, with age at enrollment being the most relevant, while contextual and macroeconomic variables show minimal predictive value. These findings confirm that **student outcomes** can be predicted with **moderate accuracy**, primarily driven by academic performance and reinforced by financial stability.

Taken together, this analysis highlights a clear hierarchy in the factors shaping student dropout and academic success. **Academic performance** stands out as the **most influential determinant**, followed by financial conditions, while demographic characteristics play a more limited supporting role and contextual variables show little direct impact. Across all methods, **semester grades** and especially **second-semester performance**, consistently emerge as the **strongest indicators** of students' final outcomes.

Rather than being a simple correlation, **academic performance** appears to be the channel through which other factors take effect. Financial pressure, age-related responsibilities, and study conditions influence students' capacity to perform academically, which in turn directly affects their likelihood of persisting or dropping out.

The practical implications are clear. **Effective dropout prevention** should focus on **early academic monitoring**, combined with **targeted financial support**, particularly during the first year and before second-semester outcomes become decisive. Encouragingly, the **most influential factors** identified are also those that institutions can directly address through **academic support services**, **financial aid policies**, and **early warning systems**.

Overall, this project demonstrates that student dropout is **neither inevitable nor primarily driven by factors beyond institutional control**. With timely, **data-driven interventions** targeting the right factors at the right moments, higher education institutions can meaningfully improve **student retention and academic success**.

9 - Appendix

This following analysis is used to validate the robustness of our results, while the main body focuses on the full dataset to reflect real student outcome distributions.

9.1 - LIME Analysis - Class-Wise Aggregation

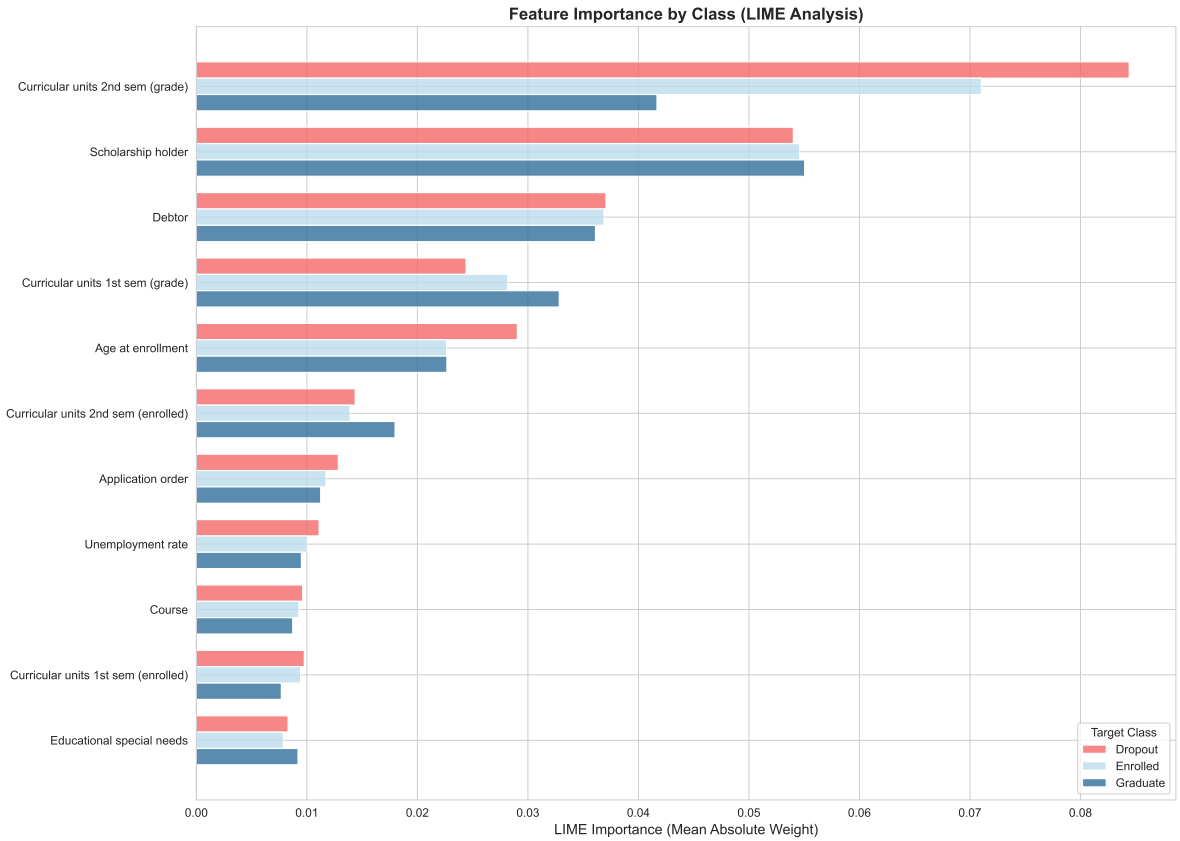
Top 15 Features for Predicting: Dropout

Based on 100 samples

	feature	mean_importance	std_importance	appearances
21	Curricular units 2nd sem (grade)	0.0844	0.0305	100
14	Scholarship holder	0.0540	0.0025	100
17	Debtor	0.0371	0.0031	100
15	Age at enrollment	0.0290	0.0169	100
19	Curricular units 1st sem (grade)	0.0244	0.0148	100
20	Curricular units 2nd sem (enrolled)	0.0144	0.0123	100
1	Application order	0.0128	0.0040	100
22	Unemployment rate	0.0111	0.0050	100
18	Curricular units 1st sem (enrolled)	0.0098	0.0049	100
2	Course	0.0096	0.0025	100
12	Educational special needs	0.0083	0.0058	100
4	Previous qualification	0.0070	0.0052	200
11	Admission grade	0.0069	0.0028	100
6	Nacionality	0.0065	0.0053	100
8	Father's qualification	0.0058	0.0036	100

Analysis of Dropout Risk Factors: The aggregated LIME analysis across 100 dropout samples reveals the most influential features driving dropout predictions. Second semester grades emerge as the dominant predictor with a mean importance of 0.0852, significantly higher than any other feature. This is followed by scholarship holder status (0.0525) and debtor status (0.0370), highlighting how financial pressures compound academic struggles. Age at enrollment (0.0283) and first semester grades (0.0234) also play important roles. Notably, all top features appear in all 100 samples, demonstrating their consistent relevance across different dropout cases. The relatively low standard deviations (especially for scholarship holder and debtor status) indicate these features have stable, predictable effects, making them reliable indicators for early intervention systems targeting at-risk students.

9.2 - Feature Importance Comparison Across Classes



Derived from aggregated local LIME, this figure shows which variables influence the most the decision of the model for each predictive class. For the Dropout class, the features with the strongest influence are *Curricular units 2nd sem (grade)*, *Scholarship holder*, *Debtor* and *Age* at enrollment. This indicates that the model frequently relies on these variables when producing predictions classified as Dropout.

For the Enrolled class, Curricular units 2nd sem (grade), Scholarship holder, Debtor and Curricular units 1st sem (grade) are found to have the highest contributions, indicating a greater reliance upon academic performance and enrollment-related variables in the model decision process.

Finally, for the Graduate class, Scholarship holder, Curricular units 2nd sem (grade), Debtor and Curricular units 1st sem (enrolled) are the most influential features. Overall, the results show that several features are consistently influential across all classes, with Age at enrollment being particularly distinctive for the Dropout class compared to the other classes where it appears with slightly lower influence. These contributions reflect aggregated local decision patterns of the model rather than global feature importance or causal relationships.

Conversely, some features show lower aggregated LIME contributions across all predicted classes, suggesting that the model relies less frequently on these variables in its local decisions, although this doesn't mean they are unimportant in certain individual cases. Across all classes, the least influential features are Course, Educational special needs, Unemployment rate and Curricular units 1st sem (enrolled).

Overall, these results indicate that the importance of the academic and financial individual characteristics overcomes the importance of contextual characteristics in the predictions. The prevalence of grades and financial variables, like Scholarship holder or Debtor, indicates that the characteristics of individual students have been taken into consideration. On the other hand, environmental characteristics, like Course or Unemployment rate, have less influence over the predictions.

References

Anaíle Mendes Rabelo, L. E. Z. (2024). A model for predicting dropout of higher education students. *KeAI Chinese Roots Global Impact*. https://www.sciencedirect.com/science/article/pii/S2666764924000341

Arora, D. et al. (2024). *Predicting students' academic success and dropout using supervised machine learning*. https://www.researchgate.net/profile/Divvyam-Arora-3/publication/384055745_Predicting_Students_Academic_Success_and_Dropout_Using_Supervised_Machine_Learning/links/66e7a314f84dd1716cf2ddf9/Predicting-Students-Academic-Success-and-Dropout-Using-Supervised-Machine-Learning.pdf

Europe-Data.com. (2025). *One in six young people in portugal have dropped out of education*. Europe-Data.com. https://europe-data.com/one-in-six-young-people-in-portugal-have-dropped-out-of-education/

Hachmeister, C.-D., & Berghoff, S. (2024). *German universities intensify measures to prevent student drop-out*. CHE Centre for Higher Education. https://www.che.de/en/2024/german-universities-intensify-measures-to-prevent-student-drop-out/

Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(18), 103–117. https://doi.org/10.3991/ijet.v17i18.25567

Sokolova, T. (2025). *Dropout rates in universities worldwide: Trends and reasons*. educations.com. https://www.educations.com/higher-education-news/rising-dropout-rates-in-universities-worldwide-reasons-and-solutions