Predicting Student Dropout and Academic Success

Exploratory Data Analysis

Patricia Gotz, Lana Kabbani, Estela Gonzalez Vizcarra, Noémie Glaus

2025-11-16

I. Introduction

This analysis examines data from a Portuguese higher education institution to identify factors that contribute to student dropout and academic success. The dataset contains information on 4,424 students enrolled across various undergraduate programs.

I.I - Background and Motivation

Student retention and academic success are crucial issues for higher education institutions. Universities increasingly 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, because we can combine analytical and predictive methods to better understand and prevent student dropout.

I.II - 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.

I.III - 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? 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?

Data Loading

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

Variable Selection

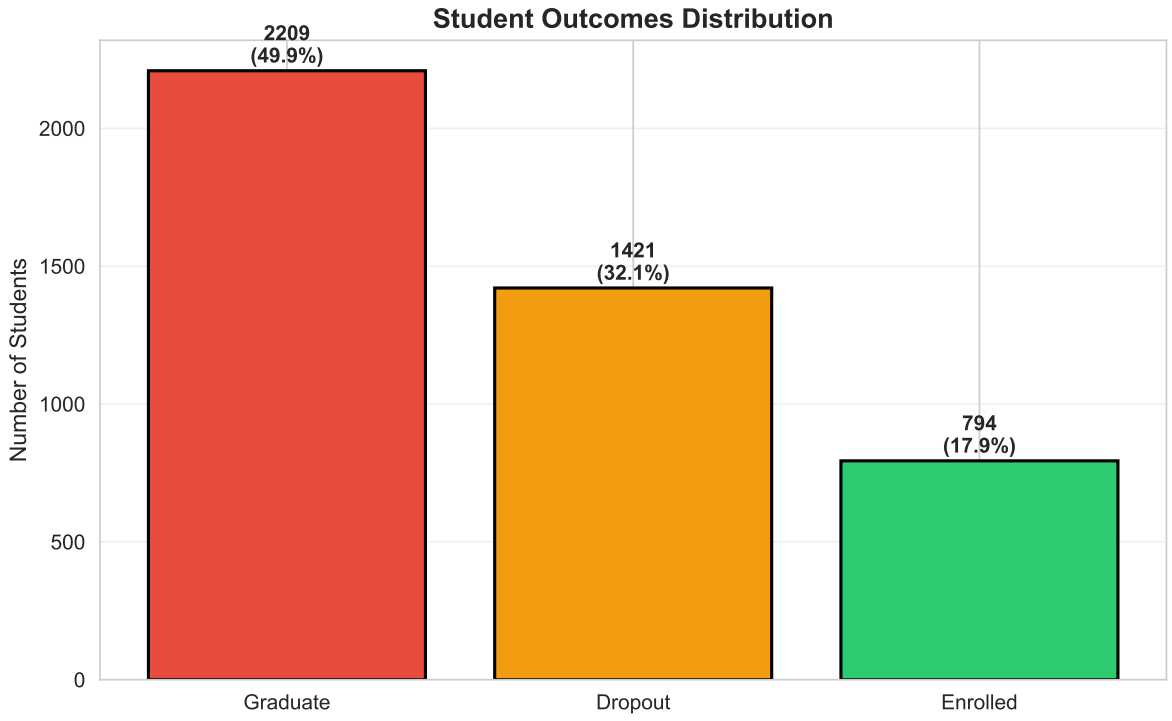
We selected 35 relevant variables for analysis:

```
Selected 35 variables
```

Data Cleaning

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

Target Variable

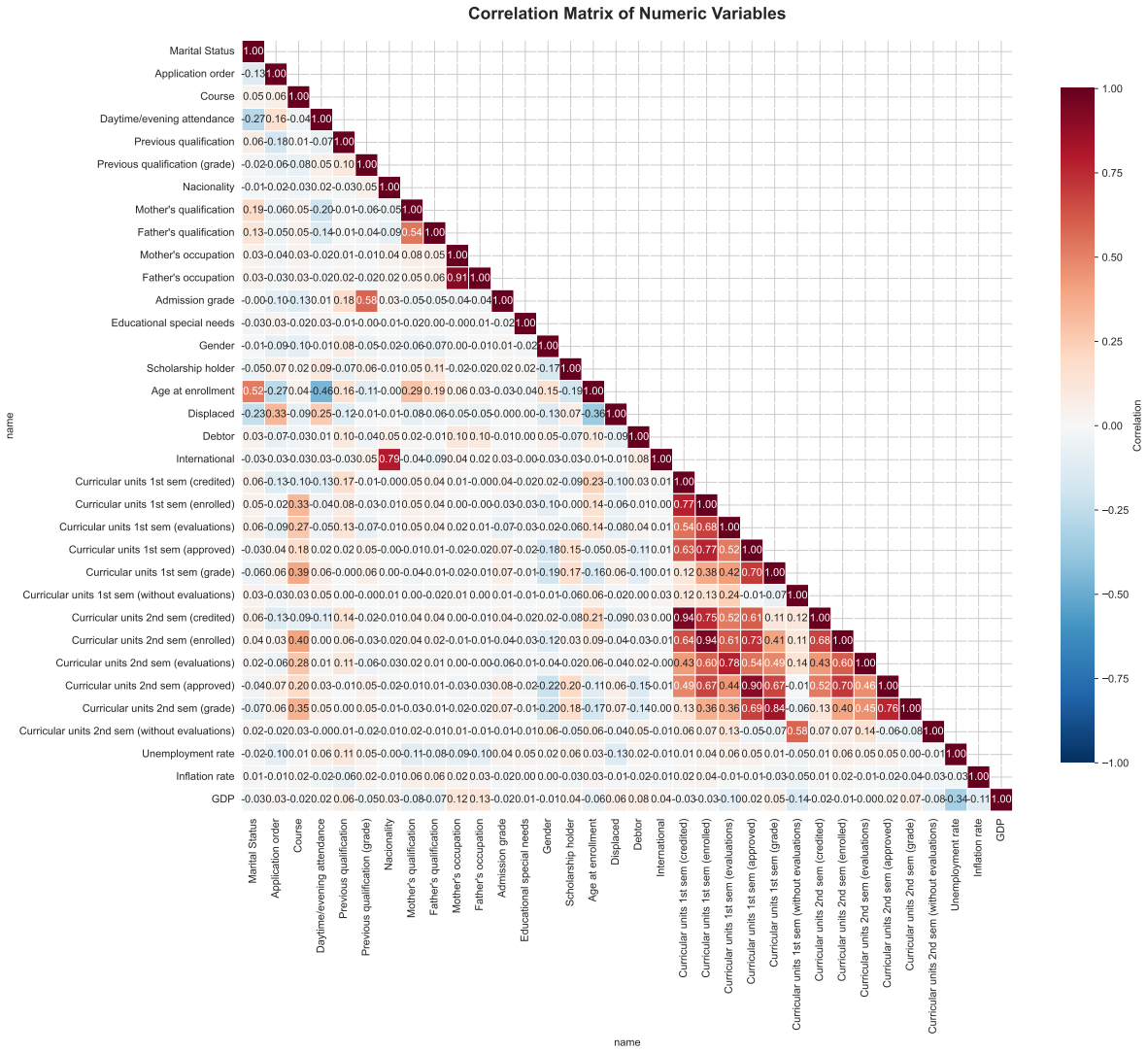


Descriptive Statistics

name	count	mean	std	min	25%	50%	75%
Marital Status	4424.0	1.18	0.61	1.00	1.00	1.00	1.00
Application order	4424.0	1.73	1.31	0.00	1.00	1.00	2.00
Course	4424.0	8856.64	2063.57	33.00	9085.00	9238.00	9556
Daytime/evening attendance	4424.0	0.89	0.31	0.00	1.00	1.00	1.00
Previous qualification	4424.0	4.58	10.22	1.00	1.00	1.00	1.00
Previous qualification (grade)	4424.0	132.61	13.19	95.00	125.00	133.10	140.0
Nacionality	4424.0	1.87	6.91	1.00	1.00	1.00	1.00
Mother's qualification	4424.0	19.56	15.60	1.00	2.00	19.00	37.00
Father's qualification	4424.0	22.28	15.34	1.00	3.00	19.00	37.00
Mother's occupation	4424.0	10.96	26.42	0.00	4.00	5.00	9.00
Father's occupation	4424.0	11.03	25.26	0.00	4.00	7.00	9.00

name	count	mean	std	min	25%	50%	75%
Admission grade	4424.0	126.98	14.48	95.00	117.90	126.10	134.8
Educational special needs	4424.0	0.01	0.11	0.00	0.00	0.00	0.00
Gender	4424.0	0.35	0.48	0.00	0.00	0.00	1.00
Scholarship holder	4424.0	0.25	0.43	0.00	0.00	0.00	0.00
Age at enrollment	4424.0	23.27	7.59	17.00	19.00	20.00	25.00
Displaced	4424.0	0.55	0.50	0.00	0.00	1.00	1.00
Debtor	4424.0	0.11	0.32	0.00	0.00	0.00	0.00
International	4424.0	0.02	0.16	0.00	0.00	0.00	0.00
Curricular units 1st sem (credited)	4424.0	0.71	2.36	0.00	0.00	0.00	0.00
Curricular units 1st sem (enrolled)	4424.0	6.27	2.48	0.00	5.00	6.00	7.00
Curricular units 1st sem (evaluations)	4424.0	8.30	4.18	0.00	6.00	8.00	10.00
Curricular units 1st sem (approved)	4424.0	4.71	3.09	0.00	3.00	5.00	6.00
Curricular units 1st sem (grade)	4424.0	10.64	4.84	0.00	11.00	12.29	13.40
Curricular units 1st sem (without evaluations)	4424.0	0.14	0.69	0.00	0.00	0.00	0.00
Curricular units 2nd sem (credited)	4424.0	0.54	1.92	0.00	0.00	0.00	0.00
Curricular units 2nd sem (enrolled)	4424.0	6.23	2.20	0.00	5.00	6.00	7.00
Curricular units 2nd sem (evaluations)	4424.0	8.06	3.95	0.00	6.00	8.00	10.00
Curricular units 2nd sem (approved)	4424.0	4.44	3.01	0.00	2.00	5.00	6.00
Curricular units 2nd sem (grade)	4424.0	10.23	5.21	0.00	10.75	12.20	13.33
Curricular units 2nd sem (without evaluations)	4424.0	0.15	0.75	0.00	0.00	0.00	0.00
Unemployment rate	4424.0	11.57	2.66	7.60	9.40	11.10	13.90
Inflation rate	4424.0	1.23	1.38	-0.80	0.30	1.40	2.60
GDP	4424.0	0.00	2.27	-4.06	-1.70	0.32	1.79

Correlation Analysis



Based on this correlation analysis, we identified several highly correlated variable pairs that suggest multicollinearity. We excluded 8 redundant semester variables that were highly correlated with other metrics.

Feature Selection

```
Removed 8 highly correlated variables
```

```
Remaining variables: 27
```

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.

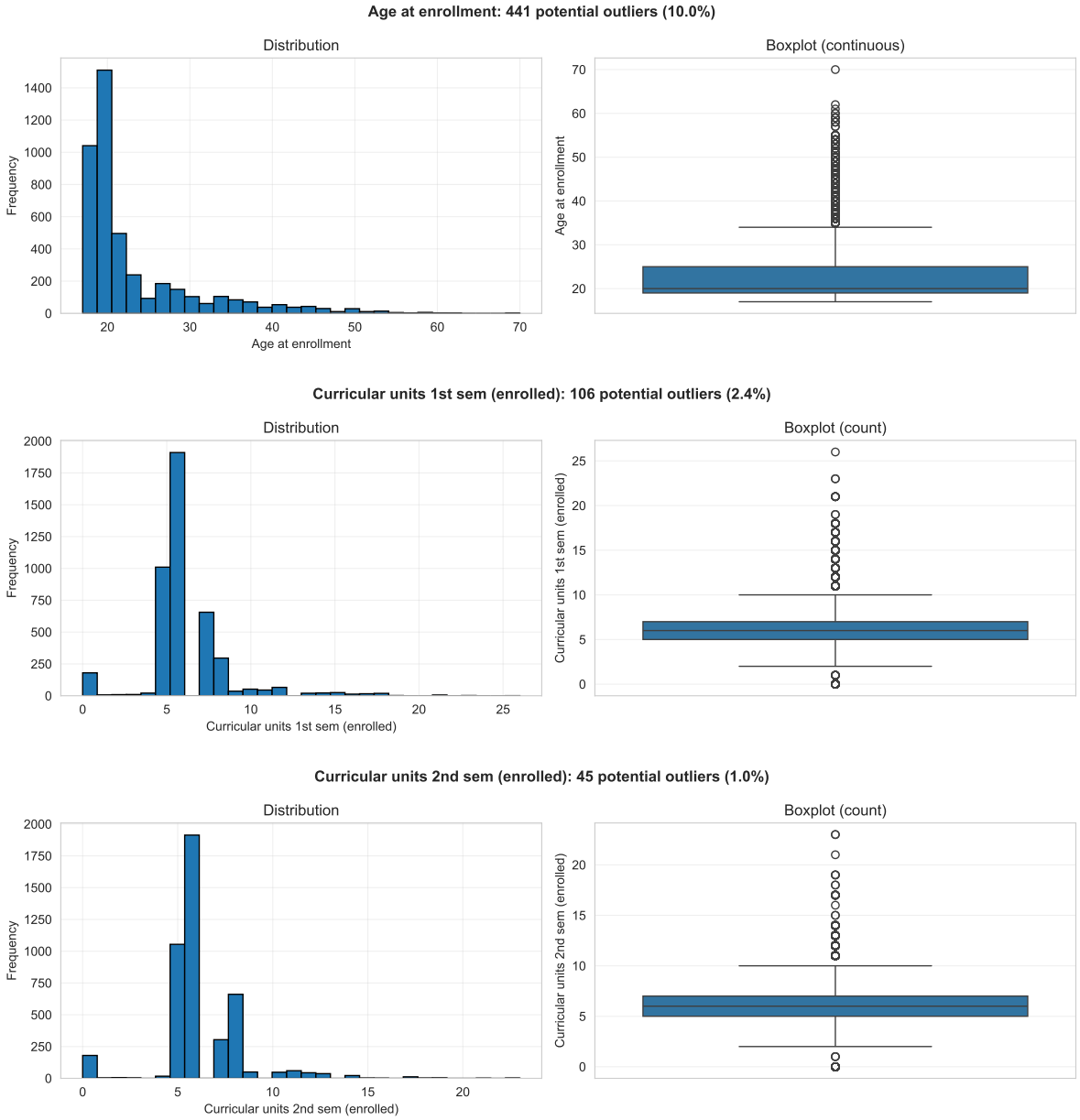
This approach ensures that outlier detection is contextually appropriate for each variable type, reducing false positives while identifying genuine data quality issues.

```
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)]
 - International: 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 scale):
 - Previous qualification (grade): range=[95.0, 190.0], invalid: 0 (0.0%)
 - Admission grade: range=[95.0, 190.0], invalid: 0 (0.0%)
 - Curricular units 1st sem (grade): range=[0.0, 18.9], invalid: 0 (0.0%)
 - Curricular units 2nd sem (grade): range=[0.0, 18.6], invalid: 0 (0.0%)

Count variables (3×IQR threshold):
 - Curricular units 1st sem (enrolled): extreme outliers: 106 (2.4%)
 - Curricular units 2nd sem (enrolled): extreme outliers: 45 (1.0%)

Continuous variables (1.5×IQR threshold):
 - Age at enrollment: outliers: 441 (10.0%)
 - Unemployment rate: outliers: 0 (0.0%)
 - Inflation rate: outliers: 0 (0.0%)
 - GDP: outliers: 0 (0.0%)
```

Outlier Summary

```
Detected Issues:
```



Feature Importance Analysis

Methodology

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:

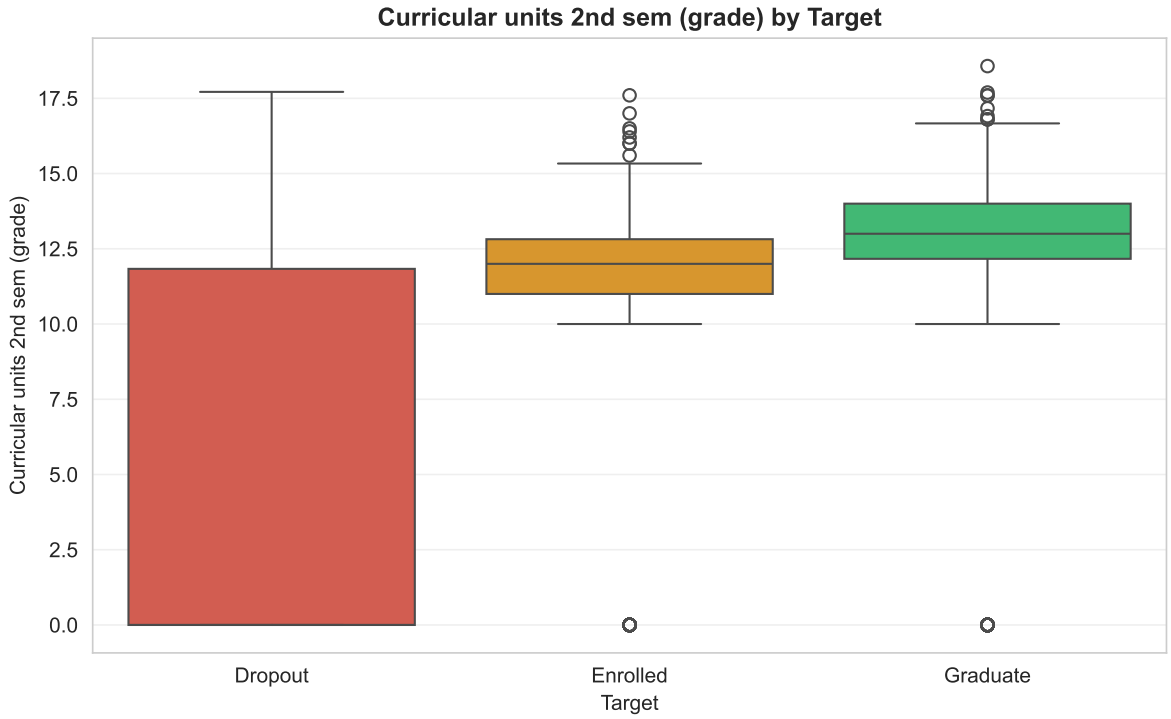
- **p-value**: Statistical significance of differences between groups (= 0.05)
- **Eta-squared (²)**: Effect size measure representing the proportion of variance explained by the target variable (ranges from 0 to 1, where higher values indicate stronger association)

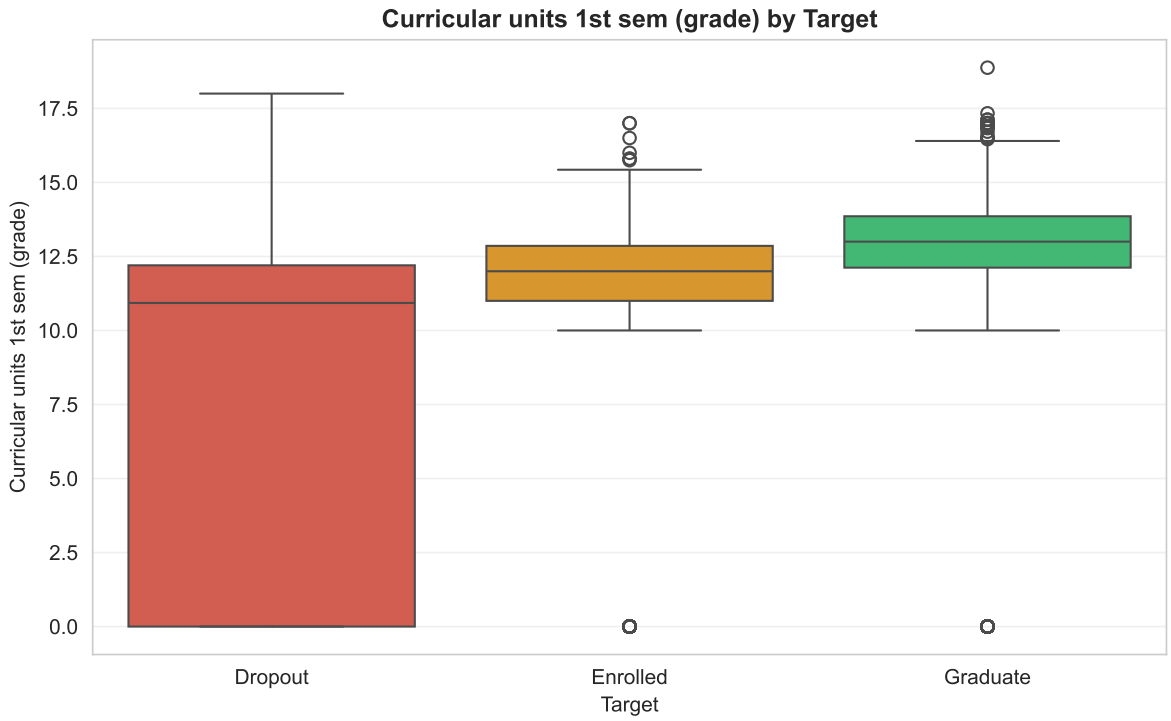
Variables with p-value < 0.05 are considered significantly associated with student outcomes and may be strong predictors in classification models.

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

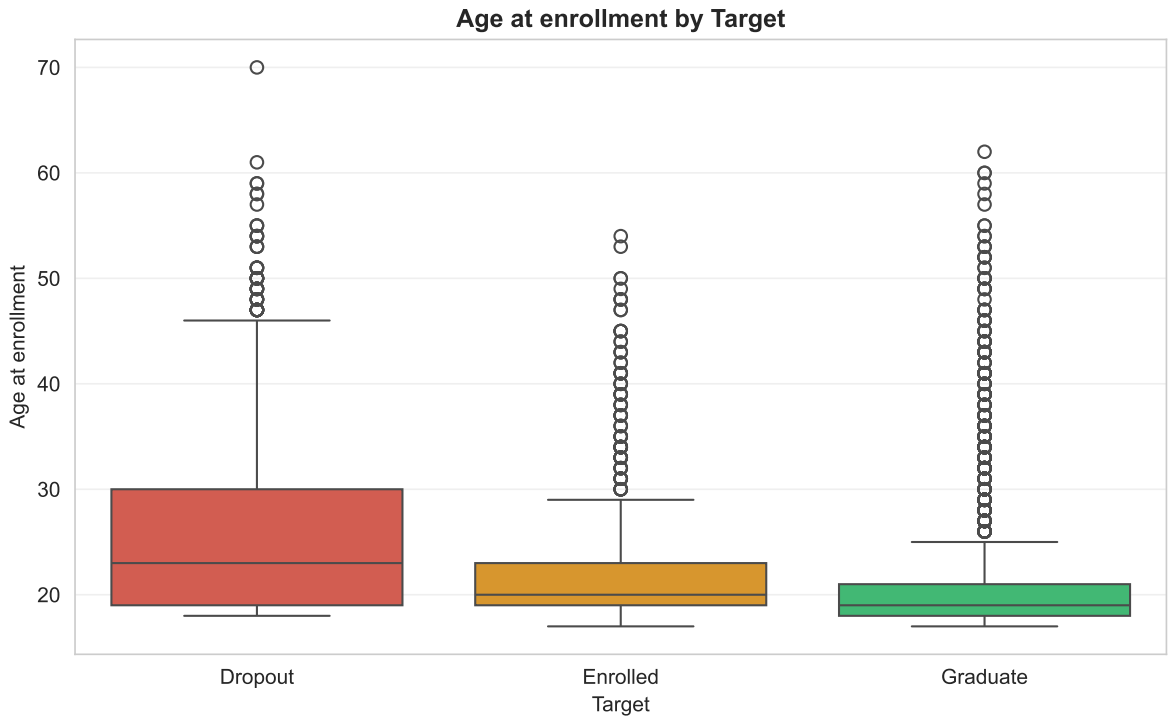
	p_value	eta_sq	significant
Curricular units 2nd sem (grade)	0.000000e+00	0.339086	True
Curricular units 1st sem (grade)	2.803052e-269	0.244020	True
Scholarship holder	4.436825e-94	0.092663	True
Age at enrollment	1.138849e-65	0.065412	True
Debtor	1.018223e-58	0.058620	True
Gender	9.950346e-53	0.052727	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

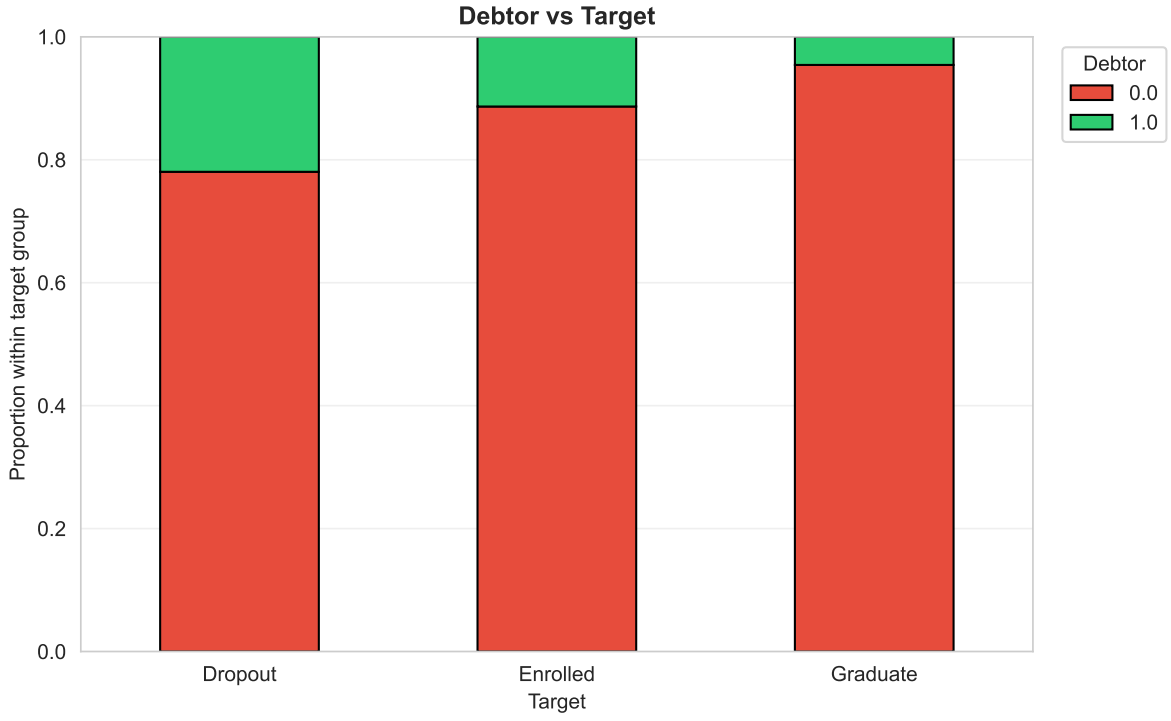
Top Predictive Variables

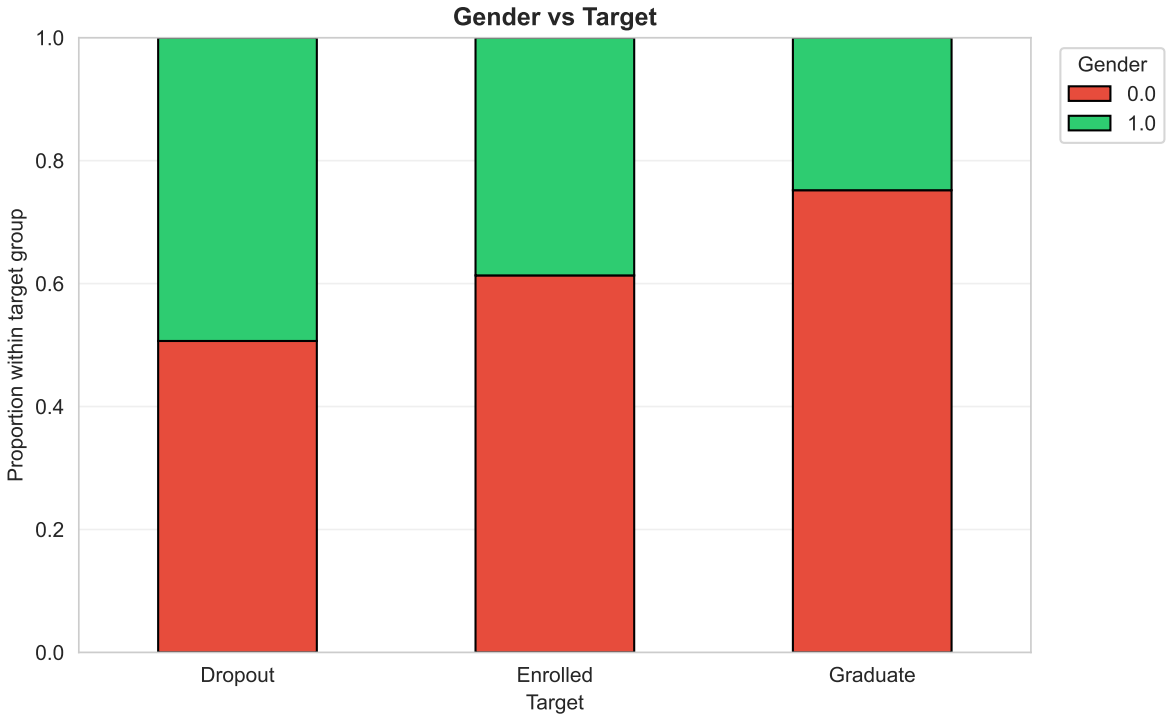


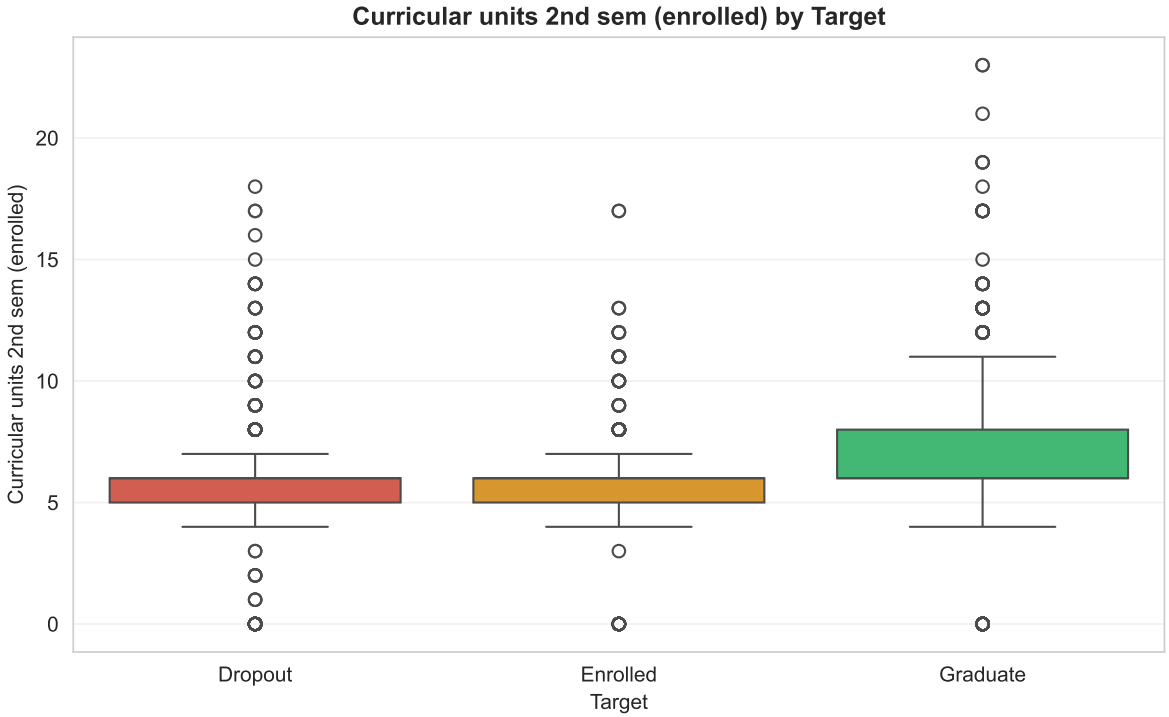

Curricular units 1st sem (grade) by Target
Curricular units 1st sem (grade)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Enrolled
Graduate
Target

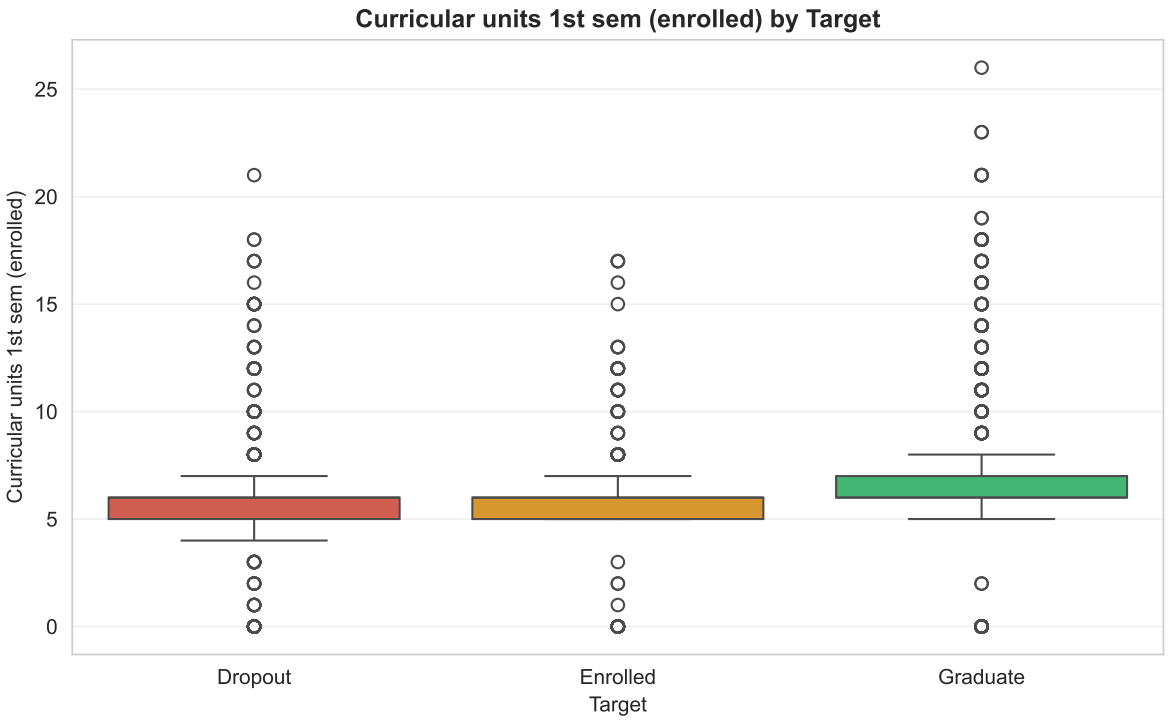

Scholarship holder vs Target
Proportion within target group
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Enrolled
Graduate
Target
Scholarship holder
0.0
1.0

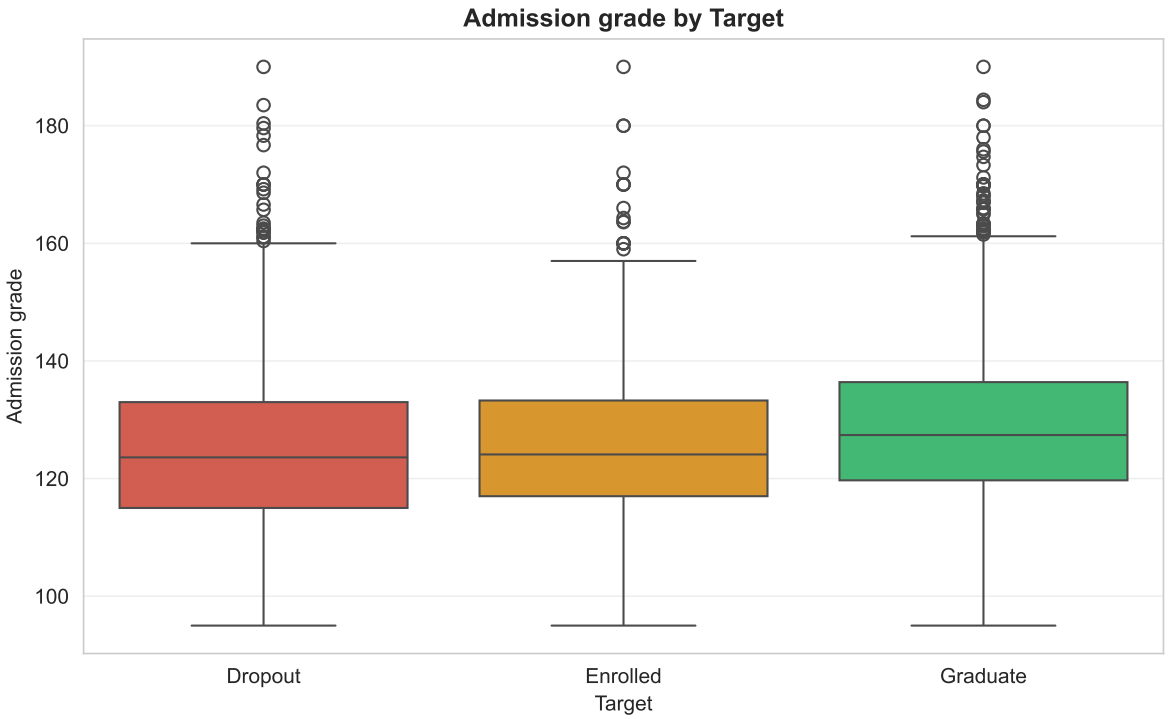

Age at enrollment by Target
Age at enrollment
70
60
50
40
30
20
Dropout
Enrolled
Graduate
Target

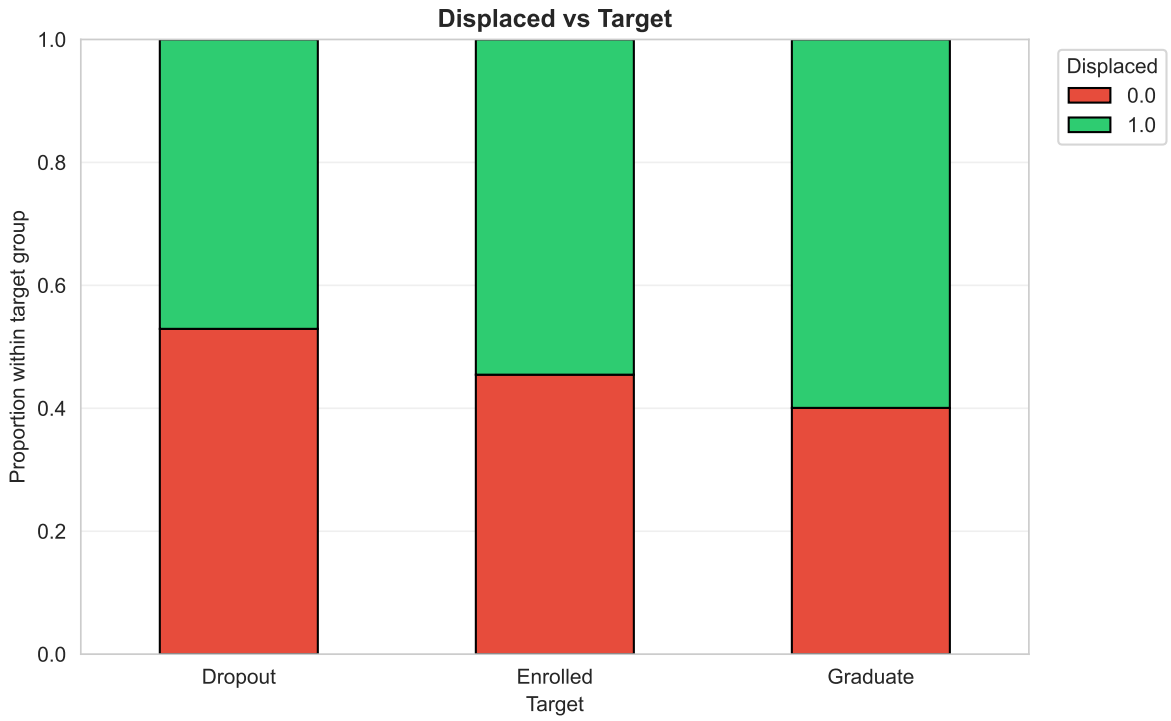

Debtor vs Target
Proportion within target group
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Enrolled
Graduate
Target
Debtor
0.0
1.0


Gender vs Target
Proportion within target group
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Enrolled
Graduate
Target
Gender
0.0
1.0


Curricular units 2nd sem (enrolled) by Target
Curricular units 2nd sem (enrolled)
20
15
10
5
0
Dropout
Enrolled
Graduate
Target


Curricular units 1st sem (enrolled) by Target
Curricular units 1st sem (enrolled)
25
20
15
10
5
0
Dropout
Enrolled
Graduate
Target


Admission grade by Target
Admission grade
180
160
140
120
100
Dropout
Enrolled
Graduate
Target



Key Findings: Academic Performance and Study Conditions

Academic Performance Impact: - Students who graduate have significantly higher admission grades (mean: X) compared to dropouts (mean: Y) - First semester grades show the strongest association with outcomes (^{2} = X), suggesting early academic performance is a critical indicator - Approved course units in semester 1 differentiate graduates from dropouts more than enrollment numbers

Study Conditions Impact: - Daytime students show X% higher graduation rates compared to evening students - Application mode significantly affects outcomes (^{2} = X, p $<$ 0.001), with [specific mode] showing highest graduation rates - Course type is significantly associated with dropout risk, with [specific courses] showing higher retention