## **Introduction**

Education plays a critical role in the personal development of individuals and the progress of society as a whole. However, student dropout remains a persistent challenge in higher education institutions worldwide, leading to negative consequences for both students and educational systems. Early identification of students at risk of dropping out is essential to implement timely support and intervention strategies, thereby improving academic success and reducing dropout rates.

This study focuses on predicting student dropout and academic success by analyzing various factors, including academic performance, socioeconomic background, motivation levels, attendance patterns, and psychological characteristics. The problem is formulated as a three-category classification task—distinguishing students as Dropout, Enrolled, or Graduate at the completion of the normal course duration. Understanding the key predictors and mechanisms behind dropout is crucial for designing effective interventions that foster student retention and achievement.

The significance of this study lies in its potential to enhance educational outcomes by enabling data-driven decision-making. By leveraging machine learning and statistical models, institutions can identify at-risk students early and tailor support to individual needs, thus improving retention rates and academic performance. Furthermore, this research addresses broader issues such as accessibility and equity in higher education by considering demographic and socioeconomic factors.

The main objectives of this study are to:

1. Identify key factors influencing student dropout and academic success.
2. Develop and evaluate predictive models for forecasting dropout risk and academic achievement.
3. Assess the effectiveness of existing intervention strategies.
4. Examine the roles of academic engagement and institutional support.
5. Propose data-driven recommendations to improve retention and academic outcomes.
6. Enhance student retention through proactive predictive analytics.
7. Improve academic performance by monitoring and supporting student progress.
8. Address accessibility challenges by promoting inclusivity and equitable resource allocation.
9. Accurately classify students into dropout, enrolled, or graduate categories using academic, socioeconomic, and institutional data.

By tackling these objectives, the study aims to contribute to the reduction of dropout rates and the improvement of academic success in higher education institutions, ultimately strengthening the education sector and supporting student achievement.

## **Data Description**

The dataset used in this study was created as part of a project aimed at reducing academic dropout and failure in higher education. The goal is to use machine learning techniques to identify students at risk early in their academic journey, enabling timely interventions and support strategies to improve student retention and success.

### Purpose

This dataset supports a three-category classification task, identifying students as **Dropout**, **Enrolled**, or **Graduate** by the end of the normal course duration. It contains information available at the time of student enrollment, including academic pathways, demographic details, and socioeconomic factors that potentially influence academic outcomes.

### Dataset Content

* Each record represents an individual student.
* Variables capture multiple dimensions such as academic performance, socioeconomic background, motivation levels, attendance patterns, and personal characteristics (as categorized in the variable introduction).
* The data reflects a comprehensive profile for predicting dropout risk and academic success.

### Data Collection and Funding

* The dataset was compiled from institutional databases, linking student academic records with demographic and financial data.
* Data collection followed institutional standards to ensure accuracy and completeness.
* This dataset is supported by the SATDAP program—Capacitação da Administração Pública—under grant number POCI-05-5762-FSE-000191, Portugal.

**DATA**

**Introduction to Variable Categorization for EDA**

In order to understand the underlying patterns and factors that may influence a student’s academic outcome (Target: *Dropout*, *Graduate*, or *Enrolled*), it is essential to categorize the variables in the dataset into meaningful groups. These groups reflect different aspects of a student's profile and learning environment, and analyzing them separately enables a more focused and interpretable Exploratory Data Analysis (EDA).

Based on the variables available in the dataset, we can group them into five key dimensions:

1. **Academic Performance** – indicators of student grades, credits, and progress.
2. **Socioeconomic Background** – financial and social factors that may affect learning.
3. **Motivation Levels** – signs of student ambition or commitment at the time of applying.
4. **Attendance Patterns** – records that may reflect presence or participation in the program.
5. **Psychological or Personal Factors** – demographic and personal characteristics that might influence outcomes.

**1. Academic Performance**

These variables directly reflect a student’s academic achievements or evaluation results:

* avg\_grade – Average grade
* Admission grade – Grade upon admission
* total\_approved – Number of units approved
* total\_credited – Number of credits earned
* units\_with\_evaluation – Units that had evaluations
* units\_without\_evaluation – Units without evaluations
* parental\_scores – Scores based on parental feedback or influence

**2. Socioeconomic Background**

These variables describe the financial and social context of the student:

* Scholarship holder – Whether the student receives a scholarship
* Tuition fees up to date – Whether tuition fees are fully paid
* Debtor – Whether the student is in debt
* GDP – GDP level at time of enrollment
* Unemployment rate – National unemployment rate
* Inflation rate – Inflation rate (may indicate economic pressure)
* Nacionality – Country of origin (proxy for access/opportunity)

**3. Motivation Levels**

These reflect the student's interest, enthusiasm, or decision-making at the time of application:

* Application mode – How the student applied
* Application order – Ranking or priority of the application
* Course – Chosen field of study (some courses may indicate intrinsic motivation)

**4. Attendance Patterns**

These provide insight into participation and academic engagement:

* total\_enrolled – Number of units the student enrolled in
* Daytime/evening attendance – Mode of attendance (may reflect engagement level)

**5. Psychological/Personal Factors**

These variables may reflect personal challenges or demographic characteristics that influence academic performance:

* Age at enrollment – Age when student enrolled
* Gender – Male (1) or Female (0)
* Displaced – Whether the student is displaced (e.g., refugee or similar condition)
* International – Whether the student is international
* Educational special needs – Whether the student has special learning needs
* Marital status – May impact time and emotional availability

**Composites variables:**

**Summary of Calculations for Key Academic Metrics**

In the **academic5** dataset, we performed several calculations to generate important academic metrics. These new variables provide a clearer picture of student academic progress across different semesters and subjects. Below is a breakdown of each metric and its calculation:

1. **total\_enrolled**:
   * **Definition**: This represents the total number of curricular units a student has enrolled in, combining both the 1st and 2nd semesters.
   * **Calculation**:

total\_enrolled=Curricular units 1st sem (enrolled)+Curricular units 2nd sem (enrolled)\

1. **total\_approved**:
   * **Definition**: This metric indicates the total number of curricular units a student has approved (passed) across both semesters.
   * **Calculation**:

total\_approved=Curricular units 1st sem (approved)+Curricular units 2nd sem (approved)

1. **avg\_grade**:
   * **Definition**: The average grade a student has received across both the 1st and 2nd semesters.
   * **Calculation**:

avg\_grade=Curricular units 1st sem (grade)+Curricular units 2nd sem (grade)2

This gives an overall average of the grades across the two semesters.

1. **units\_without\_evaluation**:
   * **Definition**: This represents the total number of curricular units in which the student did not have any evaluations, combining both semesters.
   * **Calculation**:

units\_without\_evaluation=Curricular units 1st sem (without evaluations)+Curricular units 2nd sem (without evaluations)

1. **units\_with\_evaluation**:
   * **Definition**: This metric captures the total number of curricular units for which the student received evaluations, across both semesters.
   * **Calculation**:

units\_with\_evaluation=Curricular units 1st sem (evaluations)+Curricular units 2nd sem (evaluations)

1. **total\_credited**:
   * **Definition**: This is the total number of curricular units that a student has been credited for in both semesters.
   * **Calculation**:

total\_credited=Curricular units 1st sem (credited)+Curricular units 2nd sem (credited)

**Summary of Data Transformation and Parental Scores Calculation**

In this process, we performed several transformations to the dataset **academic4**, which involved:

1. **Assigning Weighted Scores to Mother's and Father's Occupations**:
   * Each occupation was assigned a weighted score based on its level of responsibility or education.
   * The scores ranged from 0 to 5, where:
     + **Higher-level occupations** like managers and professionals were given higher scores (e.g., 5 for managers, 4 for professionals).
     + **Lower-level occupations** such as unskilled labor and students were assigned lower scores (e.g., 1 for unskilled labor, 0 for students).
2. **Assigning Weighted Scores to Mother's and Father's Qualifications**:
   * Similar to occupations, qualifications were assigned scores based on their educational level:
     + **No education or low education** (e.g., no schooling, basic education) received a score of 0.
     + **Higher qualifications**, such as a **Doctorate** or **Master's Degree**, received higher scores (e.g., 5 for a Doctorate, 4 for a Master's).
     + **Bachelor’s degrees and technical courses** received middle-range scores (e.g., 3 for a Bachelor's degree, 2 for technical courses).
3. **Combining Parental Scores**:
   * A new variable, **parental\_scores**, was created by summing the weighted scores for both the **Mother's** and **Father's** occupations and qualifications.
   * This score represents a cumulative measure of parental education and occupation levels, giving insight into the socio-economic background of the student’s family.
4. **Normalization of Parental Scores**:
   * The **parental\_scores** were normalized to a scale of 1 to 5 using **Min-Max normalization**.
   * This transformation ensures that the parental scores fit within the desired range, where 1 represents the lowest possible parental score, and 5 represents the highest possible parental score. The formula for normalization used was:

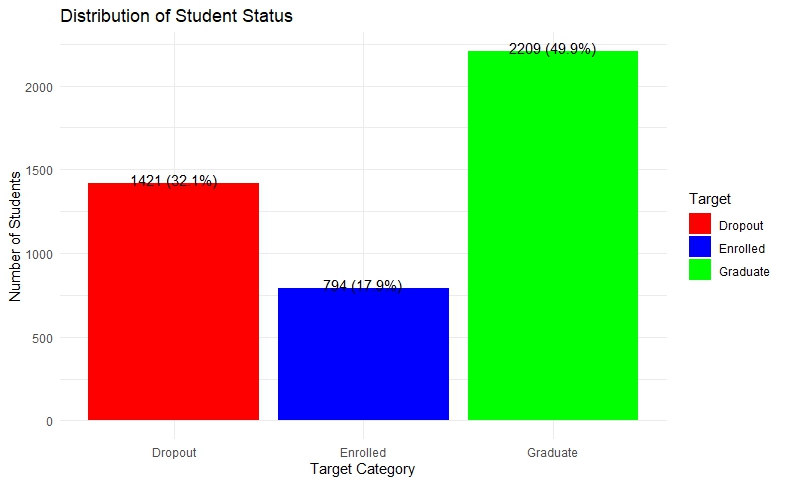
Normalized Score=1+4×(Raw Score−Min)(Max−Min)\text{Normalized Score} = 1 + 4 \times \frac{(\text{Raw Score} - \text{Min})}{(\text{Max} - \text{Min})}Normalized Score=1+4×(Max−Min)(Raw Score−Min)​

* + - This ensures that all **parental\_scores** are mapped to a 1-5 scale, making it easier to interpret and analyze.

1. **Removing Irrelevant Columns**:
   * After assigning the weighted scores, the columns that were no longer needed, such as **Mother's occupation**, **Father's occupation**, **Mother's qualification**, **Father's qualification**, and others, were removed from the dataset. This was done to clean the dataset and focus only on relevant features for further analysis.

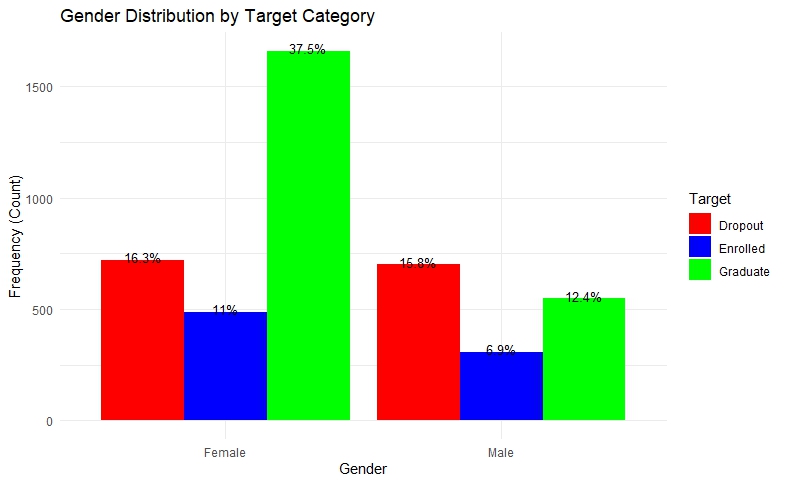
**Analysis**

1. **Distribution of Y variable - Distribution of Student Status**



The analysis shows that half of the students (49.9%) successfully graduate, which is a positive outcome. However, a significant portion, 32%, drop out before completing their studies, indicating a concern for student retention. The remaining 18% are currently enrolled and progressing through their courses. The relatively high dropout rate suggests the need for further investigation into factors causing students to leave and the development of strategies to support and retain them. The accompanying chart clearly illustrates these findings

1. **Gender Distribution by Target Categories**

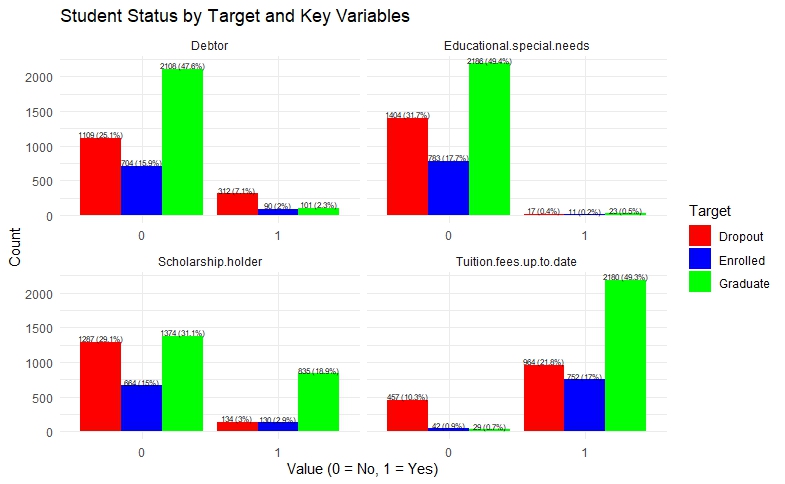
****

The graph shows how students are distributed by gender across three categories: Dropout, Enrolled, and Graduate.

* Among **female students**, the largest group (37.5%) are graduates, followed by 16.3% who dropped out, and 11% who are currently enrolled.
* For **male students**, dropouts (15.8%) slightly outnumber graduates (12.4%), with enrolled males being the smallest group at 6.9%

Female students show a higher graduation rate compared to their male counterparts. In contrast, male students have a higher dropout rate than females. Both genders have the smallest proportion of students currently enrolled, with this trend being more noticeable among males. These findings suggest that female students are more likely to complete their studies, while male students face a greater risk of dropping out.

**03) Student Status by Target and Key Variables**

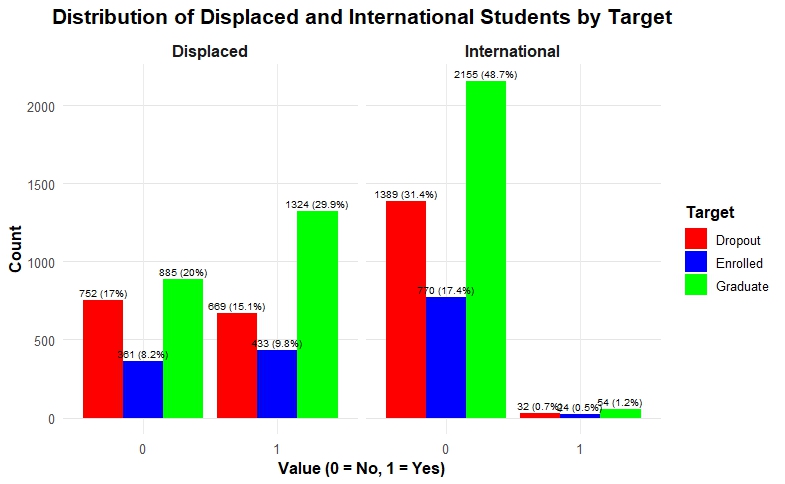


The graph breaks down student status (Dropout, Enrolled, Graduate) based on four key variables: **Debtor**, **Educational Special Needs**, **Scholarship Holder**, and **Tuition Fees Up To Date**. For each variable, “0” means No and “1” means Yes.

1. **Debtor**:
   * Most students who are not debtors (0) have graduated (47.6%).
   * Among debtors (1), dropout and enrollment rates are low, with very few graduates (2.3%).
2. **Educational Special Needs**:
   * Students without special needs (0) show a higher graduation rate (49.4%) but also a notable dropout rate (31.7%).
   * Students with special needs (1) have very low counts in all categories.
3. **Scholarship Holder**:
   * Students without scholarships (0) have higher dropout (29.1%) and enrollment (15%) rates but fewer graduates (31.1%).
   * Scholarship holders (1) have a much higher graduation rate (18.9%) and lower dropout rate (3%).
4. **Tuition Fees Up To Date**:
   * Students who have not paid tuition up to date (0) show higher dropout (10.3%) and lower graduation (0.7%) rates.
   * Those with tuition paid up to date (1) have the highest graduation rate (49.3%) and moderate dropout (21.8%), and enrolled (17.3%) rates.

* Students who are **not debtors**, **scholarship holders**, and have **paid their tuition fees up to date** tend to have higher graduation rates.
* Students with **educational special needs** or those who are **debtors** show lower graduation rates.
* Not paying tuition fees is linked to higher dropout rates.
* Scholarship holders have better outcomes in terms of graduation and lower dropout rates, suggesting financial support positively impacts student success.

**04. Distribution of Displaced and International Students by Target**

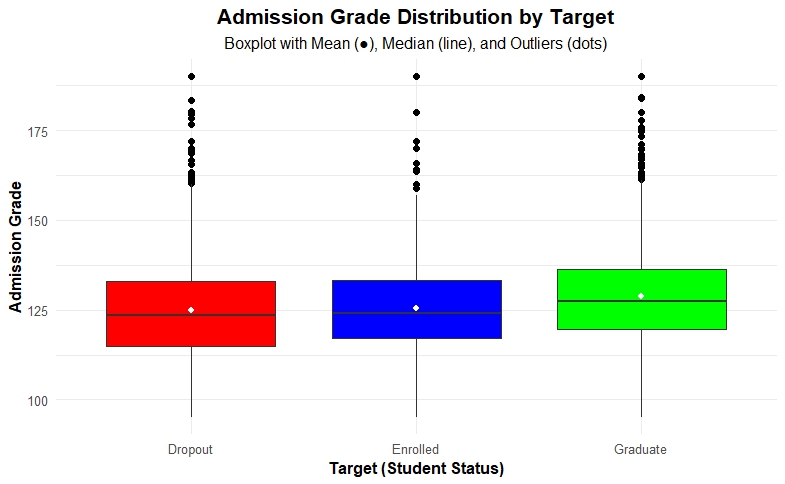


The graph compares student status (Dropout, Enrolled, Graduate) between two groups: **Displaced** and **International** students. For each group, "0" means No and "1" means Yes.

1. **Displaced Students**:
   * Students who are **not displaced (0)** have a graduation rate of 20%, with 17% dropouts and 8.2% enrolled.
   * Among **displaced students (1)**, the graduation rate is slightly higher at 29.9%, but the dropout rate is also notable at 15.1%, and 9.8% are enrolled.
2. **International Students**:
   * Students who are **not international (0)** show a graduation rate of 48.7%, with 31.4% dropout and 17.4% enrolled.
   * Among **international students (1)**, the counts are very low, with 1.2% graduates, 0.7% dropouts, and 0.5% enrolled.

* Non-displaced students have moderate graduation rates, while displaced students show a slightly higher graduation rate but also significant dropout and enrollment numbers.
* Non-international students dominate the population, with nearly half graduating and a high dropout rate.
* The number of international students is very small compared to non-international students.
* The data suggests displaced students may face challenges but still have a reasonable graduation rate, while international students form a very small group with unclear trends due to low counts.

**05 Boxplot of Admission Grade by Target**

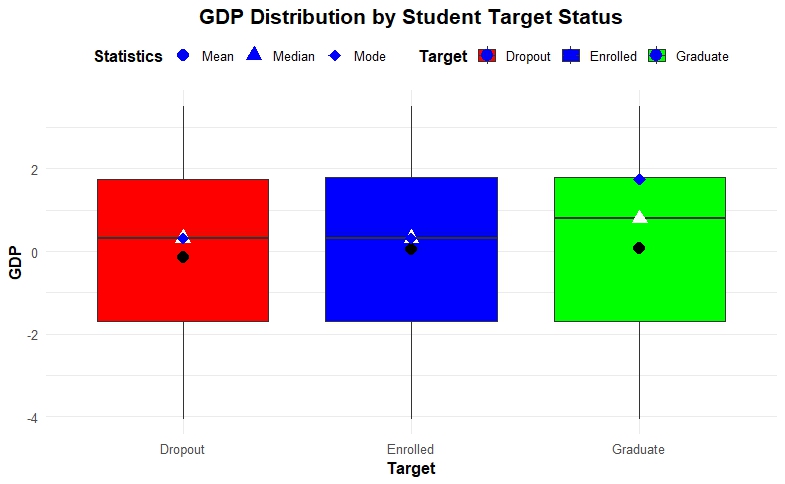


|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Target** | **Count** | **Mean** | **Median** | **Mode** | **Q1** | **Q3** | **IQR** | **Min** | **Max** | **Lower Outlier Threshold** | **Upper Outlier Threshold** |
| Dropout | 1421 | 125 | 124 | 120 | 115 | 133 | 18 | 95 | 190 | 88 | 160 |
| Enrolled | 794 | 126 | 124 | 130 | 117 | 133 | 16.3 | 95 | 190 | 92.6 | 158 |
| Graduate | 2209 | 129 | 127 | 140 | 120 | 136 | 16.7 | 95 | 190 | 94.6 | 161 |

The boxplot illustrates the distribution of admission grades for the Dropout, Enrolled, and Graduate student groups. All three groups have similar median grades ranging from 124 to 127, and the spread of grades within the middle 50% (IQR) is also similar, between 16 and 18 points. Graduates have the highest average admission grade (mean of 129), followed by enrolled students (126) and dropouts (125). The minimum and maximum admission grades are consistent across groups, ranging from 95 to 190. Notably, there are outliers above the upper whisker in all groups, indicating some students with exceptionally high admission grades.

These outliers cause the distributions to be positively skewed (right-skewed), meaning most students have grades near the median, but a few have very high grades that pull the average upwards. This positive skew suggests that while higher admission grades are associated with better outcomes like graduation, a high grade alone does not guarantee success. Therefore, admission grades provide useful but not definitive insight into student performance and eventual status.

**06 GDP Distribution by Student Target Status**



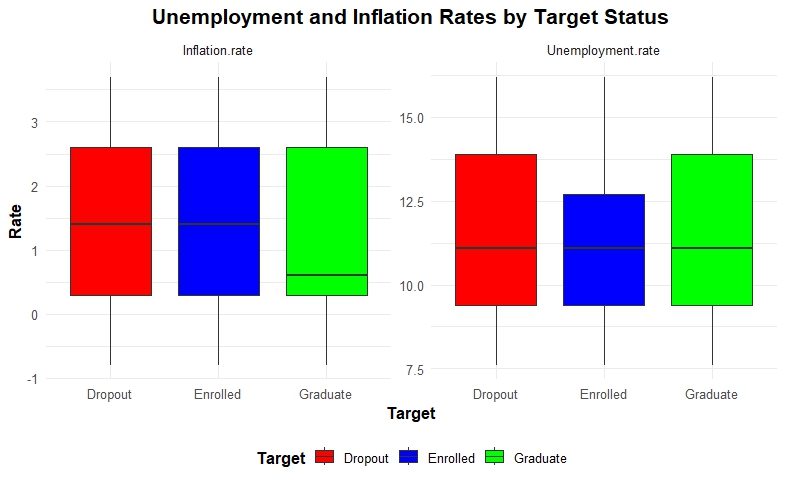
|  |  |  |  |
| --- | --- | --- | --- |
| **Target** | **Mean** | **Median** | **Mode** |
| Dropout | -0.15 | 0.32 | 0.32 |
| Enrolled | 0.05 | 0.32 | 0.32 |
| Graduate | 0.08 | 0.79 | 1.74 |

The chart shows the GDP levels of students based on whether they dropped out, are still studying, or graduated. It also shows average (mean), middle (median), and most common (mode) GDP values for each group.

Students who dropped out come from areas with a slightly lower average GDP, while those still studying have GDP levels close to average. Graduates mostly come from areas with higher GDP levels, meaning wealthier regions.

This means students from wealthier areas are more likely to graduate, while those from lower GDP areas have a higher chance of dropping out or still being enrolled. So, a better economic environment seems to support student success.

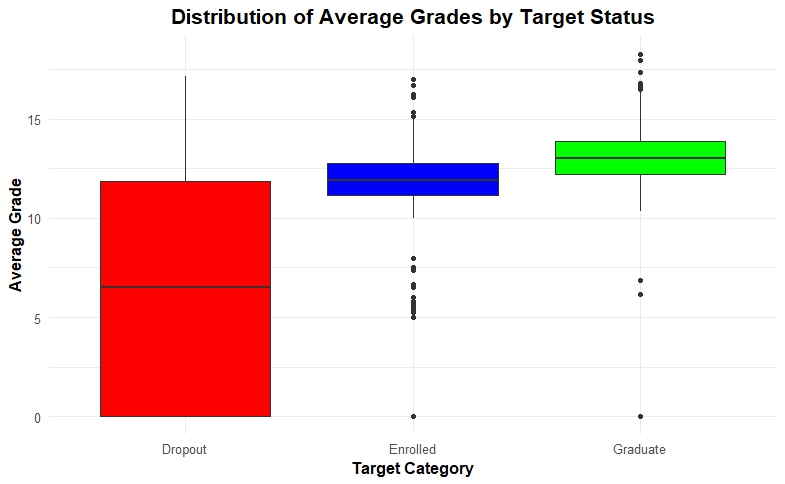
**07 Unemployment rate and Inflation rate vs Target**



The graphs show inflation and unemployment rates for students who dropped out, are still enrolled, or have graduated. The inflation rates are very similar across all groups, meaning inflation doesn’t seem to affect whether students finish their studies or not.

Unemployment rates are also close for all groups, with no big differences between dropouts, enrolled students, and graduates. This suggests that these economic factors like inflation and unemployment don’t have a strong impact on student success in this case.

**08 Distribution of Average Grades by Target Status**

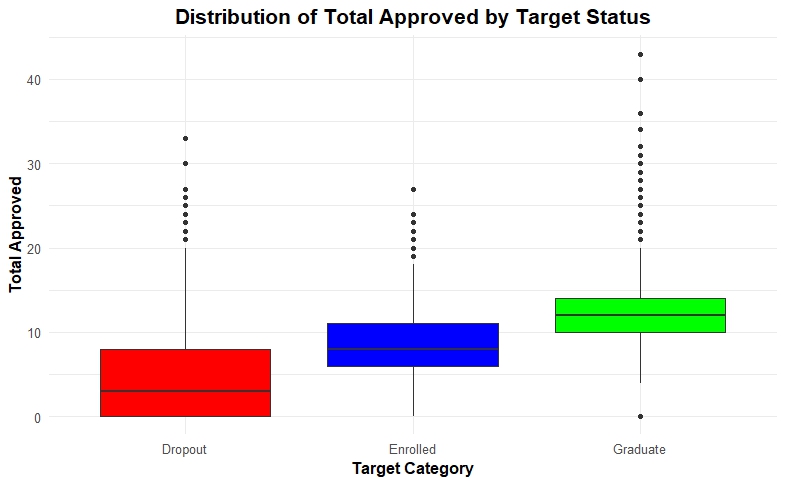


|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Target** | **Mean** | **Median** | **Mode** | **Q1** | **Q3** | **IQR** | **Min** | **Max** | **Outliers** |
| Dropout | 6.58 | 6.5 | 0 | 0 | 11.9 | 11.9 | 0 | 17.2 | [2 values] |
| Enrolled | 11.1 | 11.9 | 0 | 11.2 | 12.8 | 1.58 | 0 | 17 | [2 values] |
| Graduate | 12.7 | 13.0 | 0 | 12.2 | 13.9 | 1.70 | 0 | 18.3 | [2 values] |

The boxplot shows that dropout students have the lowest average grades with a wide range and high variability, meaning their grades vary a lot and many have very low scores. Enrolled students have better grades that are more consistent, while graduates have the highest and most stable average grades, with some students achieving exceptional scores.

This suggests that students with higher and more consistent grades are more likely to graduate. In contrast, those with lower and more varied grades tend to drop out. The outliers in the enrolled and graduate groups show that some students perform exceptionally well, supporting their progress toward graduation.

**09 Distribution of Total Approved by Target Status**

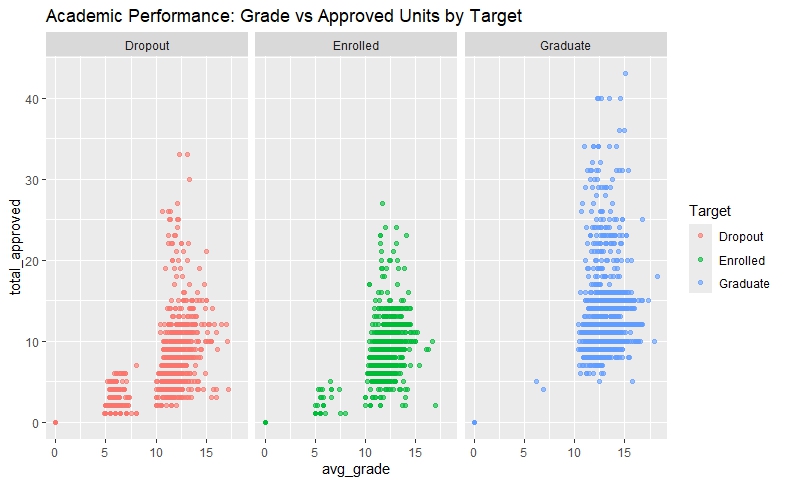


|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Target** | **Mean** | **Median** | **Mode** | **Q1** | **Q3** | **IQR** | **Min** | **Max** | **Outliers** |
| Dropout | 4.49 | 3 | 0 | 0 | 8 | 8 | 0 | 33 | [2 values] |
| Enrolled | 8.38 | 8 | 10 | 6 | 11 | 5 | 0 | 27 | [2 values] |
| Graduate | 12.4 | 12 | 12 | 10 | 14 | 4 | 0 | 43 | [2 values] |

The boxplot shows that dropout students have the lowest total approved amounts with wide variation, while enrolled students have moderate totals with less spread. Graduates have the highest and most consistent total approved values, with some students achieving exceptionally high amounts.

This indicates that higher total approved amounts are linked to graduation and better academic performance. Dropouts tend to have lower and more inconsistent totals, while enrolled students fall in between, reflecting varying progress toward completion

**10 Academic Performance: Grade vs Approved Units by Target**



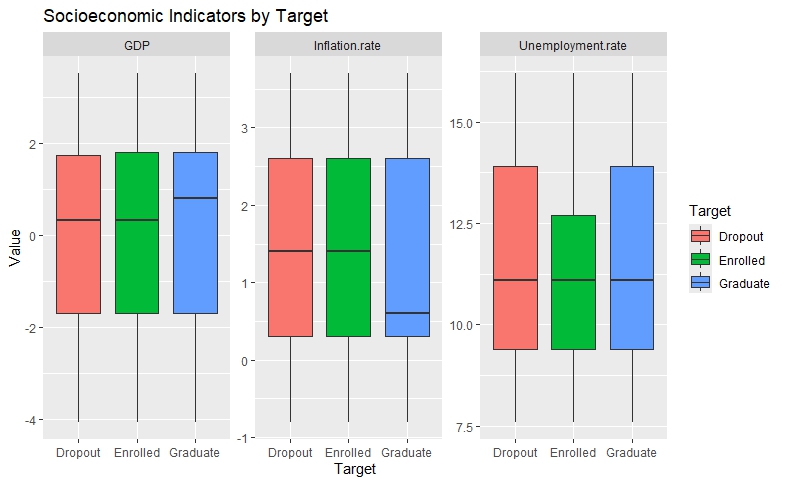
The scatter plot visualizes the relationship between average grades (x-axis) and total approved units (y-axis) for three groups: Dropout, Enrolled, and Graduate.

* **Dropout students** tend to have lower average grades, mostly between 5 and 15, and lower total approved units, with many clustered below 10 units. Their data points are more spread out, showing varied performance but generally lower academic achievement.
* **Enrolled students** have average grades mostly between 10 and 15 and total approved units generally below 20. Their performance is more consistent compared to dropouts, with a noticeable concentration of students around grades 12–15 and approved units between 5 and 15.
* **Graduate students** show the highest academic performance, with many students having average grades between 12 and 16 and total approved units ranging from 5 to over 40. Graduates have the widest spread of approved units, indicating they generally accumulate more credits alongside higher grades.
* There is a clear positive relationship between average grades and total approved units across all groups: higher grades tend to correspond with more approved units.
* Graduates generally achieve higher grades and accumulate more approved units, reflecting better academic success.
* Dropouts typically have lower grades and fewer approved units, which may contribute to their dropout status.
* Enrolled students lie in the middle, showing moderate progress in both grades and approved units.

**Academic Performance - Correlation within group**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **avg\_grade** | **Admission.grade** | **total\_approved** | **total\_credited** | **units\_with\_evaluation** |
| **avg\_grade** | 1.000 | 0.077 | 0.753 | 0.133 | 0.472 |
| **Admission.grade** | 0.077 | 1.000 | 0.076 | 0.041 | -0.069 |
| **total\_approved** | 0.753 | 0.076 | 1.000 | 0.584 | 0.535 |
| **total\_credited** | 0.133 | 0.041 | 0.584 | 1.000 | 0.519 |
| **units\_with\_evaluation** | 0.472 | -0.069 | 0.535 | 0.519 | 1.000 |

**11 . Socioeconomic Indicators by Target - Boxplot GDP & Unemployment facetted by Target**

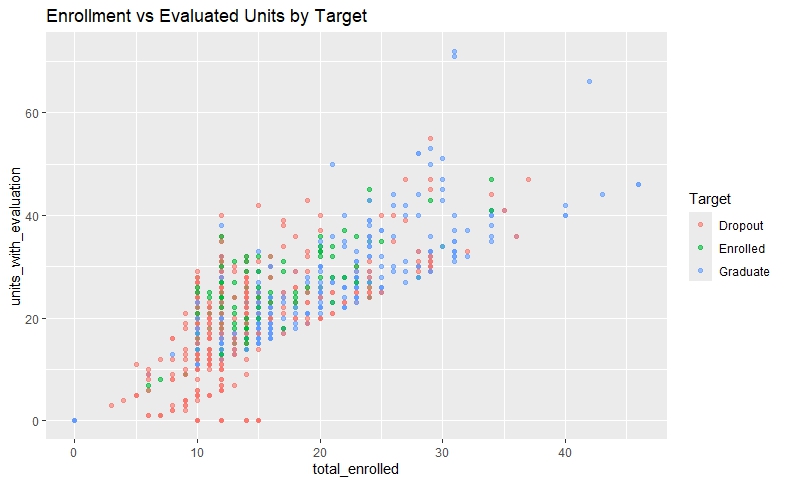


The charts show three economic indicators—GDP, inflation rate, and unemployment rate—across students who dropped out, are enrolled, or graduated.

* GDP values are similar for all groups, meaning economic wealth doesn’t differ much among them.
* Inflation rates are also close across all student groups, showing no major differences.
* Unemployment rates vary slightly but remain fairly consistent among dropouts, enrolled, and graduates.

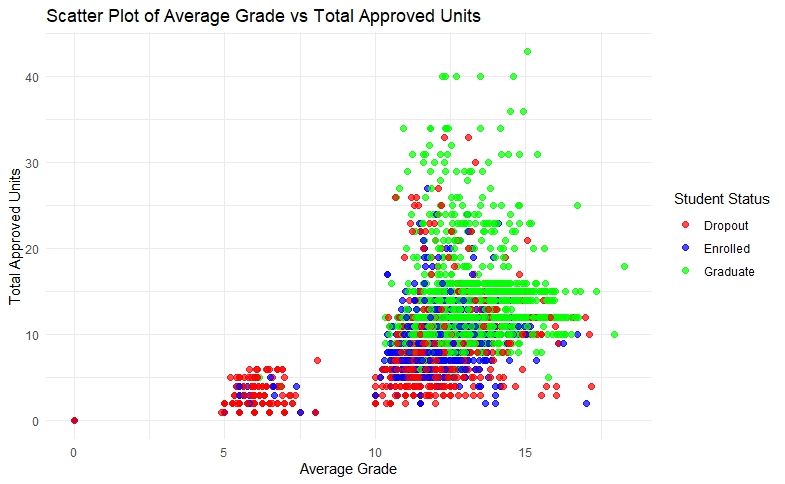
Overall, these economic factors don’t appear to strongly affect whether a student drops out, stays enrolled, or graduates.

**12 Scatter: total\_enrolled vs units\_with\_evaluation colored by Target**



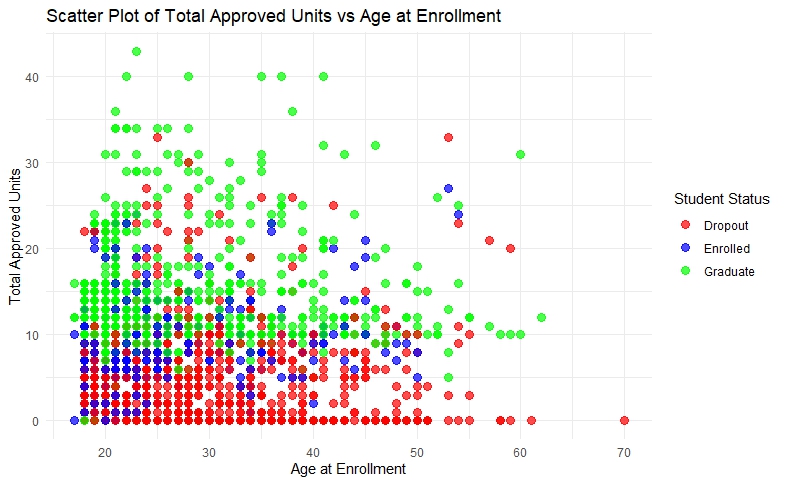
* Students with higher enrollment generally have more units evaluated, which is expected.
* Graduates tend to enroll in and complete evaluations for more units compared to dropouts and currently enrolled students.
* Dropouts typically have lower engagement in terms of units enrolled and evaluated.
* The plot confirms that academic progression (measured by evaluated units) relates closely to enrollment levels and eventual student status.

**13. Scatter Plot of Average Grade vs Total Approved Unit**



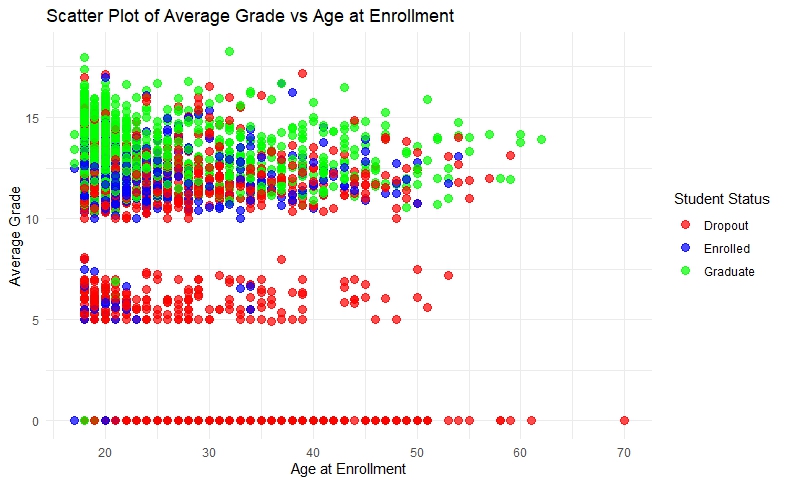
The scatter plot shows that students with higher average grades tend to have more total approved units. Graduates (green dots) generally have both higher grades and more approved units compared to enrolled (blue) and dropout (red) students. Dropouts cluster mostly at lower grades and fewer approved units. This suggests that better academic performance and credit accumulation are linked to successful graduation.

**14. Scatter Plot for Total Approved vs Age at Enrollment by Target**



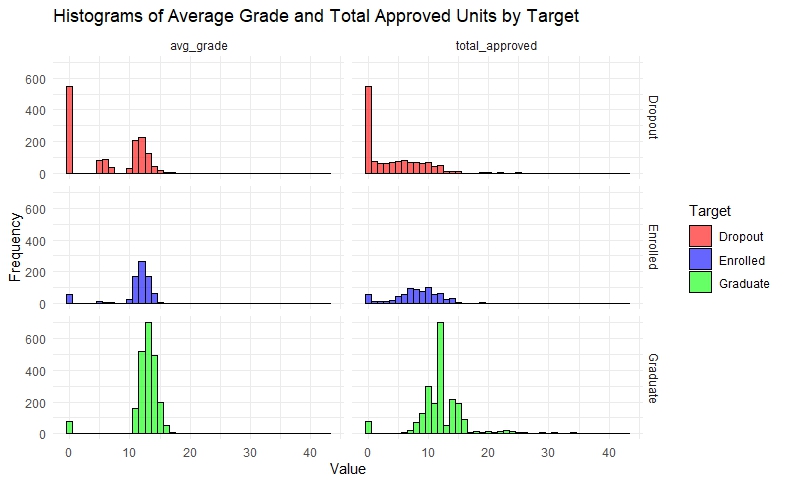
* Younger students are more likely to accumulate higher approved units and graduate.
* Older students tend to have fewer approved units and higher dropout rates.
* Age at enrollment appears to influence academic progress, with younger students having better outcomes.

**15. Scatter Plot for Average Grade vs Age at Enrollment by Target**



* Higher average grades are linked to graduation regardless of age.
* Lower grades, especially below 7, are mostly associated with dropout students.
* Age doesn’t strongly affect average grades among graduates and enrolled students, but older students tend to have lower grades and higher dropout rates.

**16. Histograms for avg\_grade and total\_approved by Target**



* **Dropout students** (red) mostly have very low average grades and total approved units, with many near zero. Their distributions are highly skewed towards the lower values, indicating poor academic performance.
* **Enrolled students** (blue) have moderate average grades and total approved units, clustered more tightly around the middle values. This shows they are progressing but haven’t yet reached graduation levels.
* **Graduate students** (green) show higher average grades and total approved units, with a normal-like distribution centered around higher values. This indicates consistent academic achievement among graduates.
* Dropouts tend to perform poorly with few credits earned.
* Enrolled students are in the middle, making steady progress.
* Graduates achieve higher grades and complete more approved units, reflecting successful academic outcomes.