P2 Dataset visualisation

Visualising the Student Dropout Journey in Higher Education

R.M.A. van Nee (r.m.a.van.nee@student.vu.nl) V. Li (v.li@student.vu.nl)

1. Introduction

In higher education, the high rate of student dropouts is a persistent, complicated problem (Tinto, 1993). The causes and predictors of student dropout have been the subject of a wide amount of study (Bean, 1980; Cabrera, Nora, & Castaneda, 1993; Realinho, et al, 2022), but a thorough understanding of the complex decision-making process that underlies student dropout is still needed. We suggest that the development of dropout intentions is a multi-stage, complicated decision-making process.

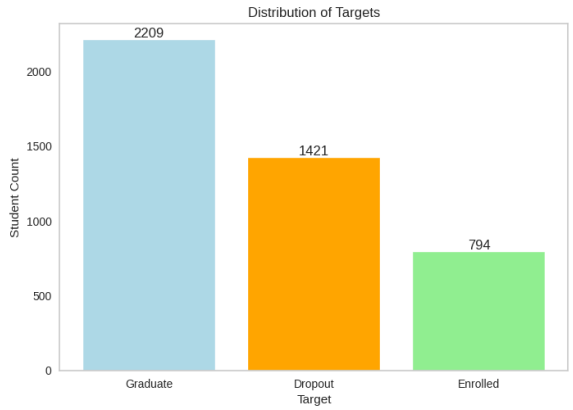
There are many factors that can determine the academic success or the dropout of an undergraduate student. These factors include demographic, socioeconomic factors, academic performance and personal characteristics of the student. All of these factors have a bearing on how students develop and complete their studies. The objective of this study is to determine the most important factors that lead to academic success or to the decision to drop out of school. By visualising the dropout rates and the factors that influence it, we hope to provide a more detailed knowledge of the process and dive further into the complexities of the dropout phenomena.

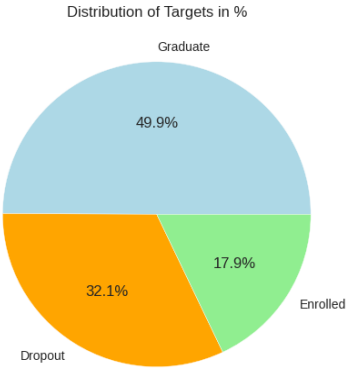
We found the dataset on Kaggle, a renowned platform for data science competitions and collaborations. The dataset is called “Predicting Student Dropout and Academic Success” and it was created by Realinho, V., Machado, J., Baptista, L., Martins, M.V. in 2022. The dataset offers an in-depth look into the trends and markers linked to student outcomes, with an emphasis on instances of academic achievement and situations that result in dropout.

2. Dataset overview

We use an existing dataset of N = 4424 students (average enrollment age of 23.3 years; 2868 female; 1556 male) with 35 variables to visually examine this notion (Realinho, et al 2022). The dataset provides a thorough summary of all the variables that might have an impact on how many students dropout of school. Data sources include internal and external information from the university, giving a comprehensive and all-inclusive view of student experiences. The dataset spans a ten-year period, from the academic years 2008–2009 to 2018–2019. Information is gathered from a wide range of undergraduate degrees, including those in agronomy, design, education, nursing, journalism, management, social service, and technology. This diversity of academic specialties enables us to understand the peculiarities of various fields of study and their possible impact on the dropout procedure. The lack of missing values in this dataset makes it special and guarantees the validity of our analysis.

This dataset is an amazing tool for acquiring nuanced insights on the complex issue of student dropout due to its completeness and comprehensiveness. We hope that our study will help to clarify the intricate interactions between variables that result in two distinct types of dropout: complete withdrawal and change in the academic field.





After our data processing, we dived into the demographics of the students who have graduated, dropped out or still enrolled at the end of the normal candidature. We found out that 2209 of the students, making up 49.9% of the population, managed to graduate on time. On the other hand, the number of people who dropped out was 1421 students at 32.1%, which was higher than those 794 students who are still enrolled, which make up the remaining 17.9%. This is expected, a significant fraction of students fail to complete their degrees within the typical candidacy term due to a variety of reasons, as literature frequently highlights, including financial hardships, academic strain, or personal concerns (Tinto, 2012). It is unexpected due to the relatively high dropout rate, which highlights how serious the issue of student dropouts is and how it persists despite numerous efforts and programs to assist student retention and timely completion in higher education.

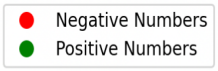
The table below shows a diverse set of attributes with its description encompassing students' personal, educational, and socioeconomic backgrounds. Its creation is likely driven by the objective of studying and predicting the factors influencing students' outcomes in higher education, particularly focusing on dropout rates.

Predict students' dropout and academic success

Attribute	**Description**
Marital status	The marital status of the student. (Categorical)
Application mode	The method of application used by the student. (Categorical)
Application order	The order in which the student applied. (Numerical)
Course	The course taken by the student. (Categorical)
Daytime/evening attendance	Whether the student attends classes during the day or in the evening. (Categorical)
Previous qualification	The qualification obtained by the student before enrolling in higher education. (Categorical)
Nationality (Typed as Nationality in the dataset)	The nationality of the student. (Categorical)

Mother’s qualification	The qualification of the student's mother. (Categorical)
Father’s qualification	The qualification of the student's father. (Categorical)
Mother’s occupation	The occupation of the student's mother. (Categorical)
Father’s occupation	The occupation of the student's father. (Categorical)
Displaced	Whether the student is a displaced person. (Categorical)
Education special needs	Whether the student has any special educational needs. (Categorical)
Debtor	Whether the student is a debtor. (Categorical)
Tuition fees up to date	Whether the student's tuition fees are up to date. (Categorical)
Gender	The gender of the student. (Categorical)
Scholarship holder	Whether the student is a scholarship holder. (Categorical)
Age at enrollment	The age of the student at the time of enrollment. (Numerical)
International	Whether the student is an international student. (Categorical)
Curricular units 1st sem (credited)	The number of curricular units credited by the student in the first semester. (Numerical)
Curricular units 1st sem (enrolled)	The number of curricular units enrolled by the student in the first semester. (Numerical)
Curricular units 1st sem (evaluations)	The number of curricular units evaluated by the student in the first semester. (Numerical)
Curricular units 1st sem (approved)	The number of curricular units approved by the student in the first semester. (Numerical)
Curricular units 1st sem (grade)	The number of curricular units graded by the student in the first semester. (Numerical)
Curricular units 1st sem (without evaluations)	The number of curricular units evaluations by the student in the first semester. (Numerical)
Curricular units 2nd sem (credited)	The number of curricular units credited by the student in the second semester. (Numerical)
Curricular units 2nd sem (enrolled)	The number of curricular units enrolled by the student in the second semester. (Numerical)
Curricular units 2nd sem (evaluations)	The number of curricular units evaluated by the student in the second semester. (Numerical)

Curricular units 2nd sem (approved)	The number of curricular units approved by the student in the second semester. (Numerical)
Curricular units 2nd sem (grade)	The number of curricular units graded by the student in the second semester. (Numerical)
Curricular units 2nd sem (without evaluations)	The number of curricular units evaluations by the student in the second semester. (Numerical)
Unemployment rate	Unemployment rate of the student's country
Inflation rate	Inflation rate of the student's country
GDP	GDP rate of the student's country
Target	The category it is classified under at the end of the normal duration of the course. (Categorical)



Attribute	Minimum	Maximum	Average	Common Values	Missing Percentage
Application mode	1.00	18.00	6.89	1.00	0.00
Course	1.00	17.00	9.90	12.00	0.00
Age at enrollment	17.00	70.00	23.27	18.00	0.00
Curricular units 1st sem (enrolled)	0.00	26.00	6.27	6.00	0.00
Curricular units 2nd sem (enrolled)	0.00	23.00	6.23	6.00	0.00
Curricular units 1st sem (grade)	0.00	18.88	10.64	0.00	0.00
Curricular units 2nd sem (grade)	0.00	18.57	10.23	0.00	0.00
Unemployment rate	7.60	16.20	11.57	7.60	0.00
GDP	-4.06	3.51	0.00	0.32	0.00
Inflation rate	-0.80	3.70	1.23	1.40	0.00

High dropout rates among students are a pressing concern for educational institutions and policymakers. To tackle this issue effectively, it's essential to identify the multifaceted factors that contribute to dropout rates. One of the key attributes we examine is "Application Mode." This attribute represents the various methods students use to apply for educational programs or services. Understanding application modes is vital because they reveal user behaviour and preferences during the application process. Analysing these modes can uncover patterns that affect students' decisions to drop out. For instance, certain application modes associated with higher dropout rates may indicate challenges in the application experience, prompting the need for improvements.

"Course Information" is another critical aspect of our analysis. This includes specific details about the courses offered within an educational institution. Analysing individual courses and their associated dropout rates is crucial for assessing curriculum effectiveness and tracking student progress. Identifying courses with higher dropout rates allows institutions to target areas that may require adjustments in curriculum design or additional support services. This data-driven approach ensures that resources are directed where they are most needed, ultimately enhancing student retention. The "Age at Enrollment" attribute provides valuable demographic insights. Understanding the age

distribution of enrolled students is essential for tailoring interventions. Younger and older students may face different challenges leading to dropout. By identifying age-related patterns in dropout rates, institutions can develop targeted interventions to address the unique needs of different age groups.

We also consider attributes related to academic progress, specifically "Curricular Units Enrolled" and "Curricular Units Grades." These attributes offer insights into a student's course load and performance in specific semesters. Analysing enrollment patterns helps identify when students are more likely to drop out, shedding light on the impact of course load on dropout rates. Furthermore, incorporating economic indicators, such as the "Unemployment Rate," "GDP (Gross Domestic Product)," and "Inflation Rate," into our analysis provides a broader context. For instance, high unemployment rates can create financial stress for families, potentially impacting students' ability to continue their education. A strong economy, reflected in a high GDP, may provide more opportunities for students and reduce financial barriers to education.

3. Charts

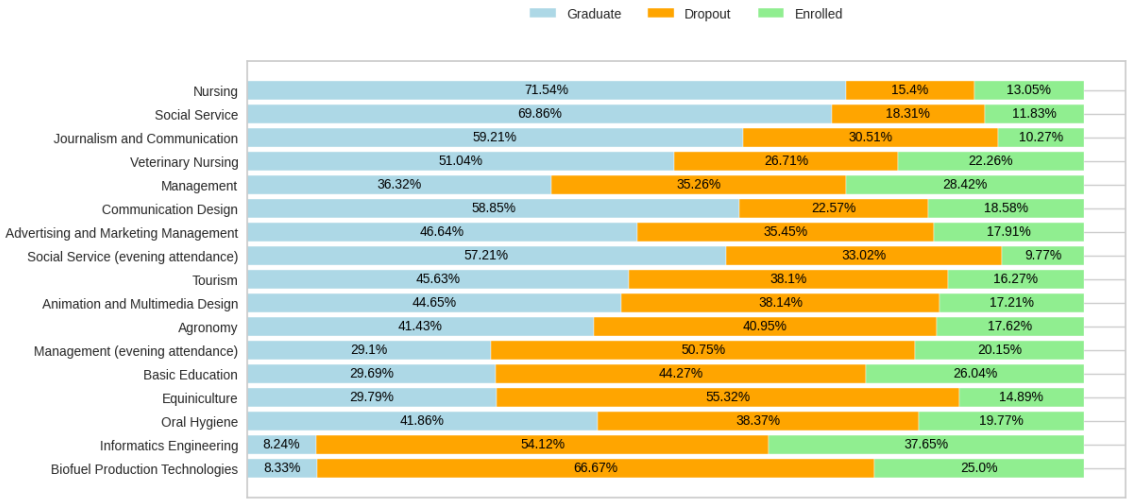


Figure 1. Distribution of the student target based on Course in percent.

From Figure 1 we can identify that Nursing has the highest graduation rate of 71.54% and the lowest is Biofuel Production Technologies at 8.33%. The dropout rate seems to be different from what we observed earlier, and Biofuel Production Technologies has the highest dropout rate of 66.67% followed by Equiniculture at 55.32%. The enrolled rate seems to be different from the earlier observation as well, with Informatics Engineering topping the chart with 37.65%. Nursing which seemed to have a high enrollment in fact has one of the lowest rates.

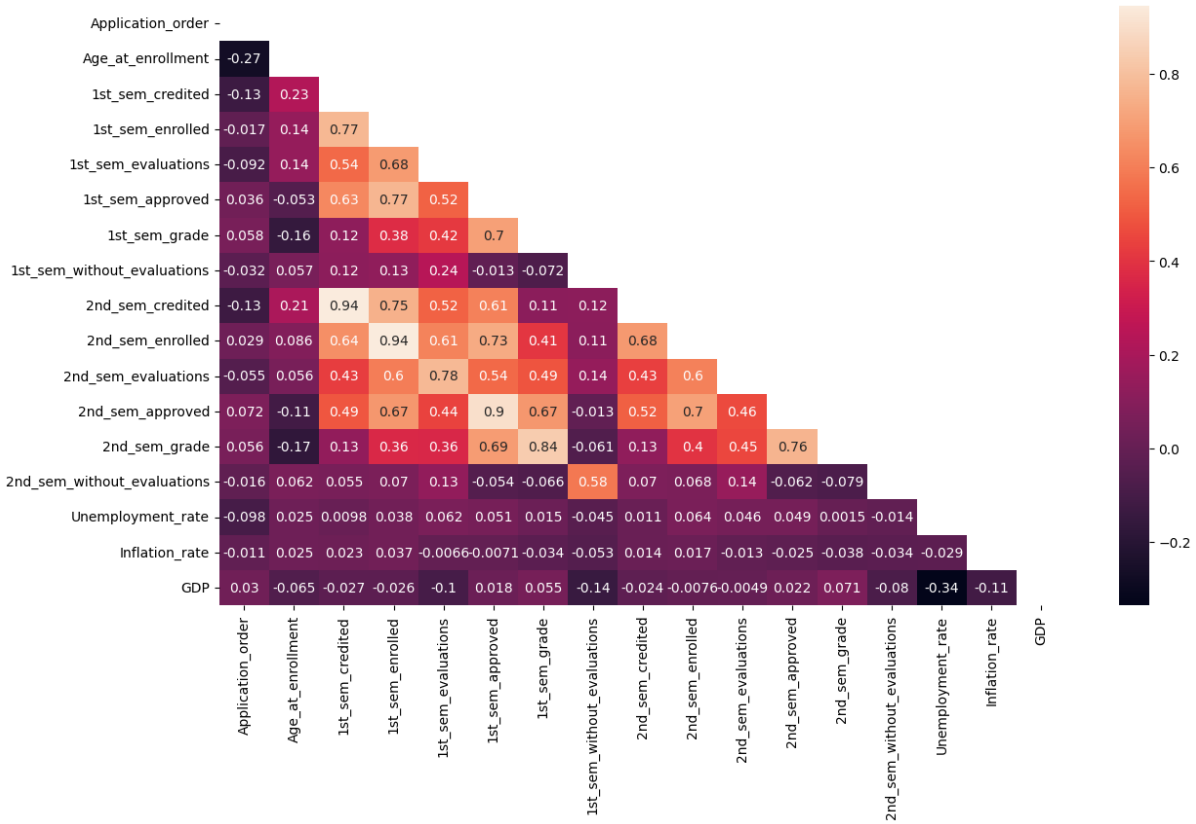


Figure 2. Correlation Between Numerical Attributes

The dataset primarily comprises categorical attributes; nonetheless, it also encompasses a handful of numerical attributes that seem associated with academic achievement. Consequently, to analyse and scrutinise these attributes further, a heatmap (Figure 2) was employed as a correlation visualisation technique. From this above heatmap, a few attributes turned out to be correlated with each other significantly. Those are '1st_sem_enrolled', '1st_sem_credited', '1st_sem_evaluations', '1st_sem_approved', '2nd_sem_credited', '2nd_sem_enrolled', '2nd_sem_evaluations', '2nd_sem_approved', '1st_sem_without_evaluations', and '2nd_sem_without_evaluations', which have had over 0.5 correlation score more than once time.

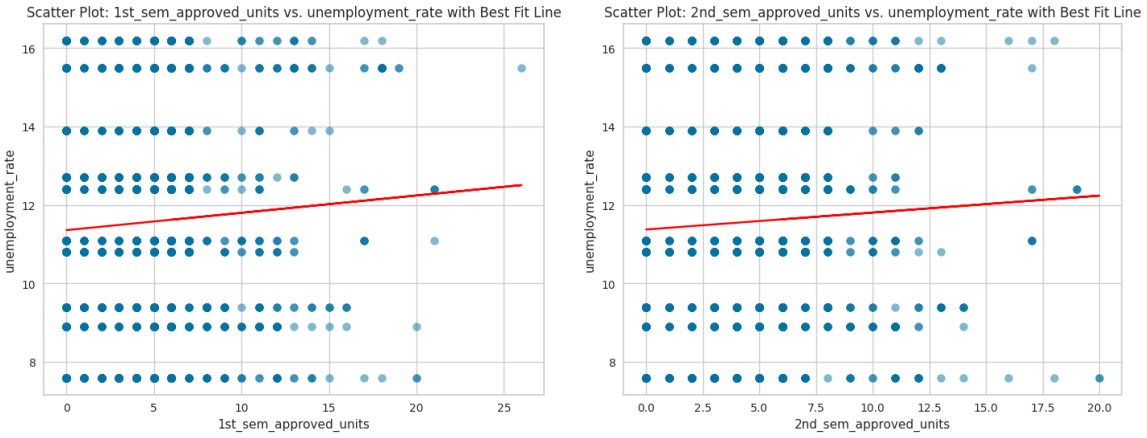


Figure 3. a) Approved units for the 1st semester; b) Approved units for the 2nd semester.

We analysed the connection between the unemployment rate and the number of approved units for the 1st and 2nd semesters, taking macroeconomic factors into account. The results showed a remarkably

low correlation of 0.05 between these variables, indicating the possibility of an additional factor influencing the relationship. It is crucial to consider other variables that could impact both the unemployment rate and the number of approved units. One potential explanation for the weak correlation is that the unemployment rate is influenced by various factors such as government policies, economic cycles, and industry-specific dynamics, while the number of approved units for the two semesters may be influenced by individual student choices, academic factors, and institutional policies. Therefore, the weak correlation implies that factors other than the number of approved units for the semesters are likely to have a stronger influence on the unemployment rate.

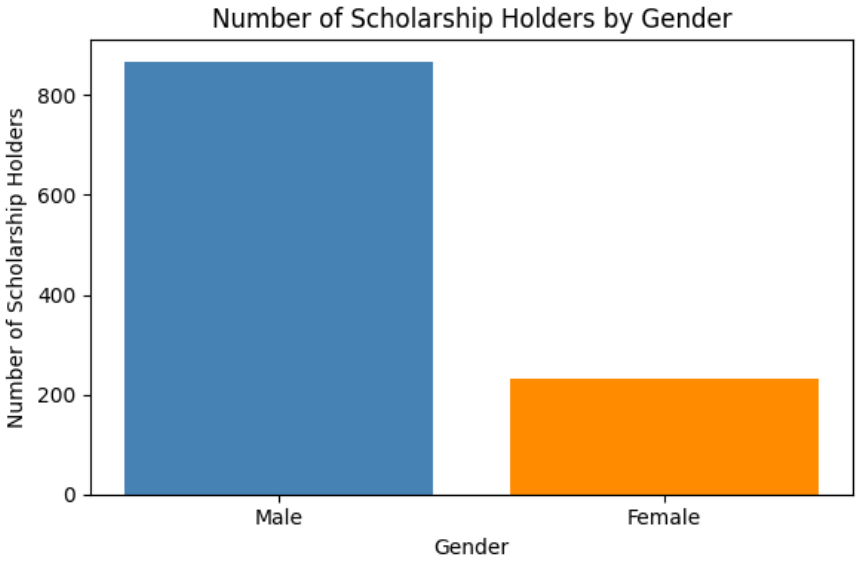


Figure 4. Number of Scholarship Holders by Gender

Figure 4 presents a visual depiction of gender-based disparities in scholarship holder status, where "Male" and "Female" are the categories under examination. The height of each bar corresponds to the count of scholarship recipients within each gender group. This graphical representation underscores the presence of a noteworthy gender disparity in scholarship awards. Specifically, it becomes evident that the "Male" category boasts a significantly higher number of scholarship holders compared to the "Female" category. This disparity in scholarship access poses a concerning challenge, potentially resulting in elevated dropout rates among female students. The adverse effects of this disparity are driven by financial constraints that hinder their academic pursuits and limit their access to educational opportunities, ultimately diminishing their prospects for success. Addressing these gender-based disparities in scholarship distribution is essential to promoting equitable access to education and reducing the adverse consequences, such as elevated dropout rates, that can result from such disparities.

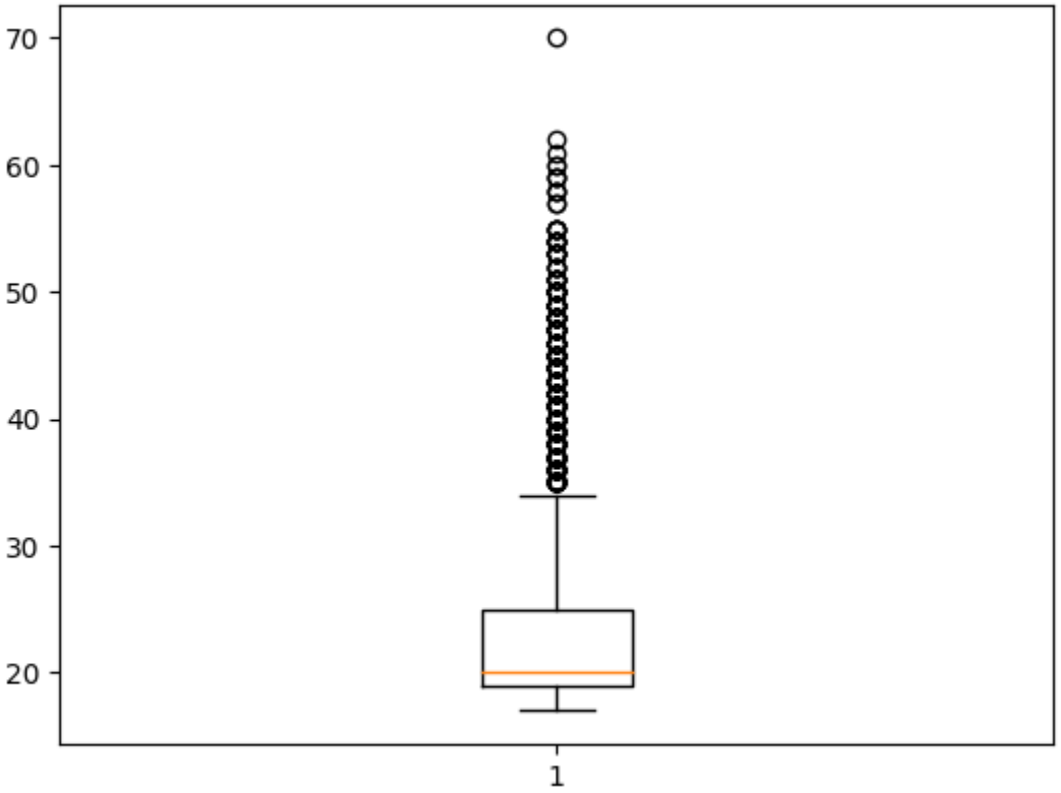


Figure 5. Box and Whisker showing the distribution of the age at the point of enrollment

In Figure 5 we can see the distribution of the age of the students. The box plot shows that the median age is clearly on the younger side, notably around the age of 22, indicating that a sizable fraction of the students are in their early 20s. The age of students' outlier data, which includes students as elderly as 70, also offers intriguing directions for further research. Further study should look at the special demands and problems of older students and how these affect dropout rates in light of the growing trend of lifelong learning and the rising average age of university students.

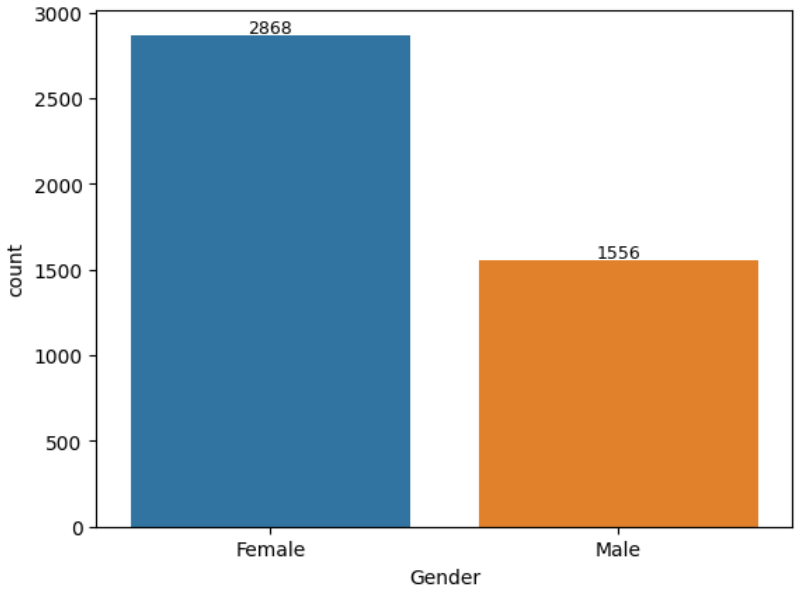


Figure 6. Female to Male Gender ratio

Figure 6 was an addition to our visualisations. There are 2868 female students and 1556 male students.This obvious distinction draws attention to a serious gender imbalance within the student body, as female students outnumber male pupils by approximately two to one. This disparity raises crucial queries about gender patterns in enrollment, course selection, and other connected aspects of the institution.

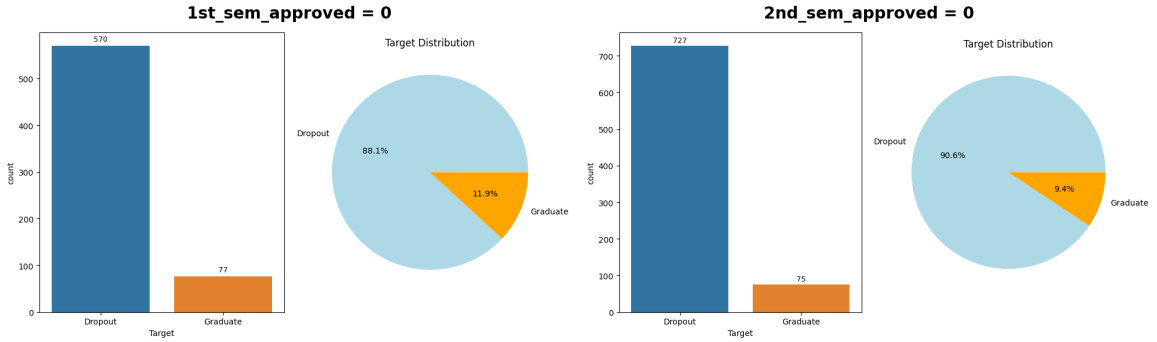


Figure 7. a) 1st semester approved = 0; b) 2nd semester approved = 0.

We will look at the important attribute = Curricular units 2nd sem (approved) and 1st sem (approved) that was found in the feature importance plot. We can observe that as their circular units for 1st and 2nd sem respectively are 0, they are more likely to have dropped out of the university course at the end of their normal candidature.

4. Problem description

Student dropout rates are one of the most important issues facing educational institutions across the world. High dropout rates can have detrimental effects on students, educational institutions, and society as a whole. In order to improve educational performance and overall student wellbeing, institutions should take preventative action by identifying which students are likely to leave school. With the dataset we have a valuable tool to tackle this issue. For instance, based on a variety of inputs, such as demographic data, academic background, and engagement indicators, a model can be created to predict a student's risk of dropping out. This prediction is more complex than just the binary outcomes of "dropout" or "not dropout." It may further divide the risk into three categories (high, medium, and low), enabling various responses. Within educational institutions, educational administrators and guidance counsellors are the main target group for using this prediction model. These people are the best candidates for making use of the model's data-driven insights since they are at the forefront of dealing with student welfare and academic success. Educational administrators can use the predictions in various impactful ways: timely interventions, i.e. administrators can start timely and suitable support mechanisms, like tutoring, counselling, or additional academic resources, by identifying at-risk students early in their educational path; personalised support, i.e. more individualised intervention tactics can be used to meet the needs and difficulties of each student by taking into account the variables that affect their risk of dropping out; policy formulation, i.e. model findings may help institutions develop more comprehensive institutional policies and initiatives that will improve student performance and retention.

5. Discussion and References

As an alternative we have chosen two datasets called “Predict students dropout, academic success” and the US Department of Education dataset. The first one was found on Kaggle and the data is acquired from higher education educational institutions showing a wide range of undergraduate degrees, including technology, agronomy, design, teaching, nursing, and media. The dataset is used to

create classification models to predict student success and dropout rates. Despite being extensive, the dataset shows a clear bias in favour of one class, which might make it difficult to develop impartial prediction algorithms. In the case with the US Department of Education – National Center for Education Statistics Dataset, the size is very comprehensive, with national-level data from various educational institutions. Scope is broad, covering various aspects of education from pre-K to post-secondary, including enrollment, academic achievement, and dropout rates. The quality of the dataset is great since the federal dataset ensures high reliability and accuracy. However, the "Predicting Student Dropout and Academic Success" dataset was selected based on the criteria because of its specific focus on dropout and academic success measures. The particular factors in the dataset, such as demographic information, academic records, and engagement measures, are in line with our prediction objectives and might provide our model with more detailed and pertinent insights.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in Higher Education, 12(**2**), 155-187. DOI: https://doi.org/10.1007/BF00976194

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. The Journal of Higher Education, 64(**2**), 123-139. DOI: 10.1080/00221546.1993.11778419

Kasworm, C. E. (1980). The Older Student as an Undergraduate. *Adult Education*, *31*(1), 30–47. DOI: https://doi.org/10.1177/074171368003100103

Little, Brenda and Tang, Win-Yee (2008). Age differences in graduate employment across Europe. HigherEducation Funding Coucil for England, Bristol, UK

Phan, M., De Caigny, A., and Coussement, K. (2023). A decision support framework to incorporate textual data for early student dropout prediction in higher education. Decision Support Systems, 168. DOI: https://doi.org/10.1016/j.dss.2023.113940

Realinho, V., Machado, J., Baptista, L., Martins, M.V. (2022). Predicting Student Dropout and Academic Success. Data, 7, 146. DOI: https://doi.org/10.3390/data7110146

Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). University of Chicago Press.

Tinto, V. (2012). Completing College: Rethinking Institutional Action. University of Chicago Press.