Data analysis aimed at identifying key factors for student dropout and academic success in higher education

***Abstract*—This study examines factors influencing student dropout and graduation rates, focusing on demographic, socioeconomic, and academic variables. Younger students and females showed higher graduation rates, while nationality had no significant impact on dropout rates, challenging assumptions about international students. Socioeconomic factors, such as financial stress and scholarships, were strongly linked to academic outcomes, with scholarships boosting graduation rates. Academic performance indicators, like prior qualifications and early grades, had limited predictive power, highlighting the complex interplay of factors in student success. The findings underscore the need for early interventions and tailored support to improve outcomes.**

***Index Terms*—Education, Student Performance, Dropout rates, Academic success, Student retention.**

I. Introduction

Student retention and graduation rates are critical indicators of educational success, influencing both institutional performance and individual career outcomes. Understanding the factors that contribute to student dropout and graduation outcomes is essential for developing effective strategies to support academic achievement and reduce dropout rates. This study explores the influence of demographic, socioeconomic, and academic factors on student success, providing a comprehensive analysis of patterns and correlations across various factors.

By examining a large dataset and employing statistical analyses, this research identifies key trends and relationships that can inform institutional policies aimed at improving student support services. The findings highlight the importance of targeted interventions to address disparities and promote equitable educational opportunities for all students.

II. Literature Review

Research on student dropout rates and performance in higher education spans both traditional and online learning environments. Rovai's [1] study highlights that students often leave online programs after initial courses due to personal, work-related, and program-specific factors. Demographic attributes like age, gender, and nationality were not significant predictors of dropout rates, aligning with broader higher education research. Instead, personal circumstances (e.g., financial issues, family obligations) and program-related challenges (e.g., workload stress, lack of interaction) were key contributors to attrition. Existing studies often focus on specific contexts, limiting generalizability. This research addresses such gaps by examining a wider range of demographic, socioeconomic, and academic factors across diverse educational settings. Including variables like nationality, cultural barriers, and prior qualifications, it offers a more comprehensive analysis. With a larger sample size, it also provides robust insights for policymakers and institutions to enhance student retention and success rates.

A comprehensive study conducted in Australia analyzed the relationship between mental health and student dropout [2]. The research utilized administrative data from 652,139 domestic undergraduate students who commenced their studies between 2012 and 2015. The findings revealed that students who received mental health treatment prior to starting university were 1.77 times more likely to drop out compared to those who did not receive such treatment. Interestingly, this effect remained consistent across different demographic and academic segments of the student population, suggesting that mental health issues have a uniform impact on dropout rates regardless of gender, socioeconomic status, or type of academic program.

Another study, focused on Spanish university students, investigated the sociodemographic, academic, and psychosocial factors influencing thoughts about dropping out [3] . This research employed quantitative analysis on a sample of 759 students, utilizing various instruments such as Likert scales to measure academic self-efficacy, help-seeking behavior, and coping strategies. The results highlighted that academic performance, academic self-efficacy, and planning were the main predictors of dropout thoughts. Notably, students with diverse sexual orientations (bisexual and homosexual) reported more frequent thoughts about dropping out, while factors such as gender, age, and origin (local or migrant) were not significant predictors. These studies contribute valuable insights to our understanding of student retention in higher education. They emphasize the importance of mental health support, academic self-efficacy, and inclusive environments in reducing dropout rates. Furthermore, they provide methodological examples for analyzing large datasets and applying statistical models to identify key factors influencing student persistence in higher education.

Kehm et al. [4] conducted a comprehensive systematic review of university dropout rates across Europe, synthesizing findings from multiple empirical studies to identify key determinants of student persistence. Their analysis highlights that institutional factors, including the quality of academic support

services, faculty engagement, and campus resources, significantly impact students' decisions to remain enrolled. Additionally, student motivation and academic integration—such as involvement in extracurricular activities, peer support networks, and faculty interactions—play crucial roles in retention. The study underscores the necessity for universities to implement targeted interventions, such as early warning systems, mentorship programs, and curriculum flexibility, to support at-risk students and reduce dropout rates across diverse educational systems.

Srairi [5] examined dropout patterns in Tunisian universities, offering insights into the interplay between economic conditions, institutional quality, and student retention. The study found that students from lower socioeconomic backgrounds faced a higher likelihood of dropping out, primarily due to financial instability, limited access to academic resources, and insufficient institutional support. Additionally, the research highlighted that universities with high student–staff ratios and lower faculty qualifications exhibited higher dropout rates, suggesting that faculty engagement and institutional resources are critical to student success. These findings emphasize the broader socioeconomic challenges affecting student retention in non-Western educational settings, underscoring the need for more equitable funding models, improved faculty training, and enhanced student support services to bridge educational disparities.

Nurmalitasari et al. [6] explored dropout trends in Indonesian private universities, identifying financial constraints, academic dissatisfaction, and personal circumstances as primary determinants of student attrition. Their study revealed that students struggling with tuition fees, limited access to scholarships, and part-time work obligations were more likely to discontinue their studies. Additionally, academic dissatisfaction—stemming from perceived low teaching quality, lack of career guidance, and inadequate course relevance—contributed to disengagement and eventual dropout. The research also highlighted that personal circumstances, such as family responsibilities, mental health challenges, and lack of peer support, played a significant role in shaping students' educational trajectories. These findings reinforce the urgent need for policy interventions that address both academic and non-academic barriers, including expanded financial aid programs, curriculum enhancements, and mental health support initiatives, to improve student retention and success rates in developing countries.

These studies contribute valuable insights to our understanding of student retention in higher education. They emphasize the importance of mental health support, academic self-efficacy, and inclusive environments in reducing dropout rates. Furthermore, they provide methodological examples for analyzing large datasets and applying statistical models to identify key factors influencing student persistence in higher education.

III. Methodology

The dataset originates from a higher education institution in Portugal [7], compiled from several disjoint databases containing comprehensive student information across various undergraduate programs including agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset comprises 4,424 student records with 36 features. To focus solely ondropout and graduate outcomes, the dataset was filtered to exclude 'Enrolled' cases, with this the number of student records went down to 3630. This study focuses on three main categories: demographic factors, socioeconomic factors and academic background. The dataset underwent rigorous preprocessing to ensure data quality: Removal of anomalies and unexplainable outliers, treatment of missing values resulting in a complete dataset, standardization of educational metrics, categorical encoding for non-numeric variables. The dataset was structured as a two-category classification problem tracking student outcomes: Dropout and Graduate. Additional variables were created for hypothesis testing, including age grouping. Ages were categorized as follows: 18-20, 21-25, 26-30, 31-40, 41-50, and 51-60 years. Independent variables (Predictors) that were used: Debtor Status (debtor or non-debtor), Scholarship Holder (has a scholarship or not), Previous Qualification (Grade), Age at Enrollment, Financial Stress Indicators (e.g., debtor status), Semestar 1 and 2 Grades, Age Group, Gender (Male or Female), Nationality (Various nationalities mapped by codes). Dependent variable that was used: Dropout or graduate. Statistical methods used include descriptive statistics, Chi-Square Test of Independence, correlation analysis, ANOVA (Analysis of Variance), and data visualization. For data analysis and statistical testing, the following Python libraries were used: pandas, matplotlib, seaborn, scipy. The dataset exhibits a notable class imbalance, with one category being significantly more represented than others.

IV. Results

In this section, we present the results of our analysis on factors influencing student dropout and graduation status. The data was explored in three categories: Demographic Factors, Socioeconomic Factors, and Academic Factors. We hypothesized that older students demonstrate lower academic success and are more likely to drop out compared to younger students. The bar plot displays the distribution of dropout and graduate outcomes across age groups. The majority of both dropouts and graduates belong to the youngest age group (18-20). However, the number of graduates is significantly higher than dropouts in this group. As age increases, dropout rates become more consistent, while the number of graduates decreases significantly (Fig. 1).

A chi-square test yielded a chi-square value of $\chi^2 = 414.45$, degrees of freedom $df = 5$, and a p-value < 0.01. Since the p-value is extremely low, the null hypothesis is rejected, indicating a statistically significant relationship between age and educational outcomes (dropout vs. graduate). While exploring the role of age in this study, we also hypothesized that younger

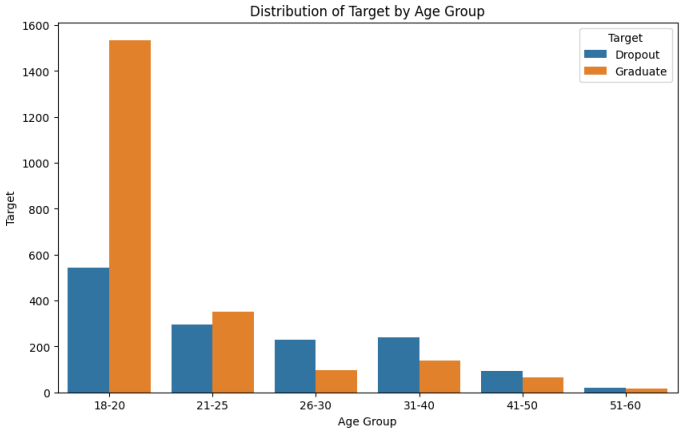

Fig. 1. Visualization of distribution of Target (Graduate marked with orange and Dropout marked with blue) by Age Group.

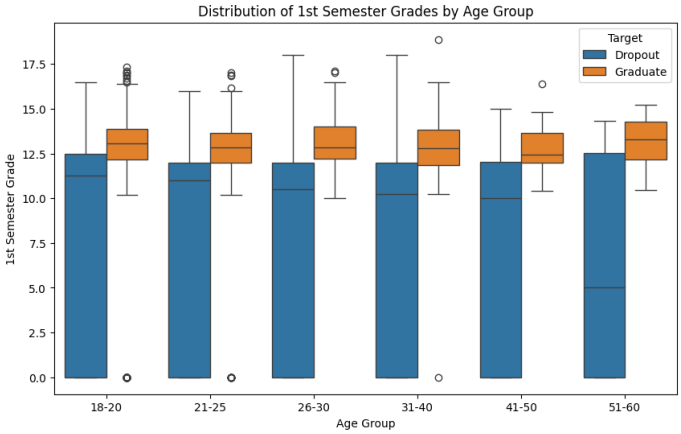

Fig. 3. Box plot showing the Distribution of 1st Semestar Grades by Age Group (Graduate marked with orange and Dropout marked with blue)

students are more likely to graduate, while older students show lower success rates. A positive correlation of 0.84 between first and second semester grades indicates consistent academic performance across semesters. Conversely, weak negative correlations of -0.16 and -0.17 between age at enrollment and grades suggest older students may experience slightly lower academic performance (Fig. 2).

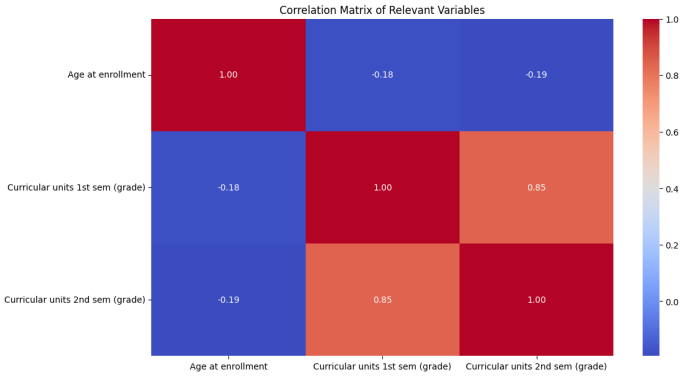

Fig. 2. Correlation Matrix of Relevant Variables (High correlation makred with red and low-correlation marked with blue)

The visualizations (Fig. 3) and (Fig. 4) present grade distributions across age groups, showing that graduates consistently outperform dropouts. Median grades for graduates range between 12.5 and 14.0, while dropout grades cluster around 10.0-12.0. The grade distribution patterns remain stable between the first and second semesters, reflecting consistent academic performance throughout the year. Notable patterns include slightly higher median grades for graduates aged 51-60, higher grade variability in younger age groups (18-20, 21-25), and more consistent performance in middle-aged groups (26-40). Outliers include exceptionally high grades around 17.5 and low performance near 0.0.

The ANOVA test results for both semesters showed statistically significant differences in grade distributions across age groups. For semester 1, the F-statistic was 30.65 (p-value: 9.48e-31), and for semester 2, the F-statistic was 40.64 (p-value: 5.30e-41), indicating significant differences across age groups. This supports the hypothesis that age plays a crucial role in academic success, with younger students potentially outperforming older students.

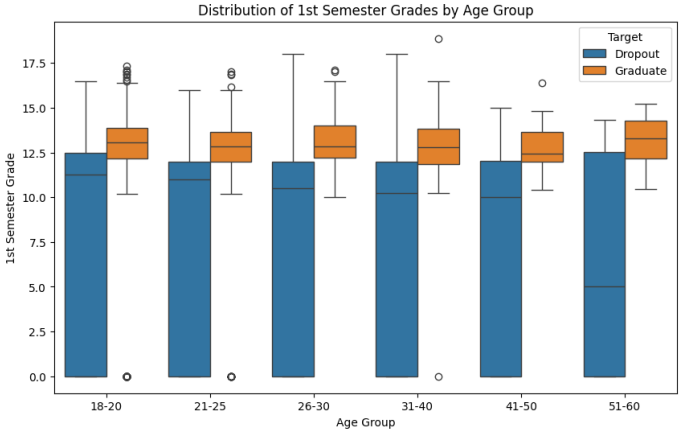

Fig. 4. Box plot showing the Distribution of 2nd Semestar Grades by Age Group (Graduate marked with orange and Dropout marked with blue)

The study moves on to the field of age-related performance variations across different genders. The hypothesis here states that males are more likely to drop out compared to females. The analysis showed that out of the total sample, 2,209 students graduated (60.9%), while 1,421 dropped out (39.1%). Gender-specific analysis revealed that males had a dropout rate of 56.1%, while females had a dropout rate of 30.2%. (Fig. 5)

A chi-square test yielded a chi-square value of $\chi^2 = 229.35$, degrees of freedom $df = 1$, and a p-value < 0.0001, leading to the rejection of the null hypothesis. This indicates a significant relationship between gender and academic outcomes, suggesting the need for gender-specific academic interventions.

Exploring the nationality influence on dropout rates, we hypothesized that foreign students face higher dropout rates due to cultural and language barriers. The data included

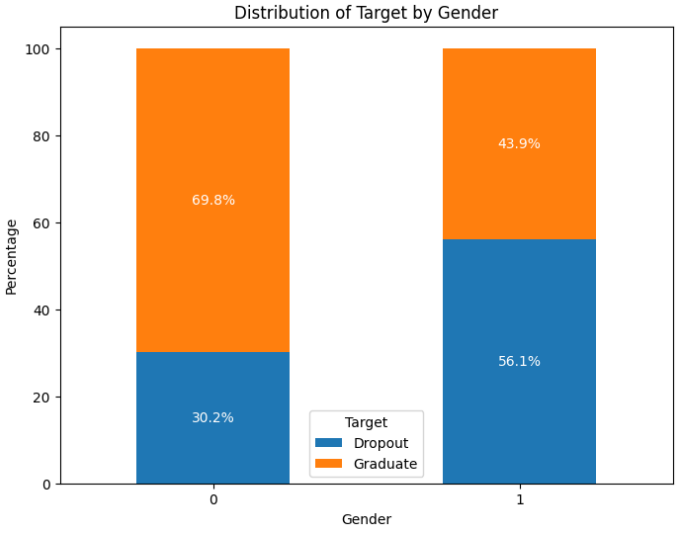

Fig. 5. Distribution of Target by Gender

19 nationalities, with Portuguese students showing a 39.2% dropout rate and a 60.8% graduation rate. Several nationalities, including German, Mozambican, Romanian, English, Dutch, and Italian students, had perfect graduation rates. Conversely, Colombian, Russian, Moldovan, Lithuanian, and Angolan students experienced 100% dropout rates. (Fig. 6)

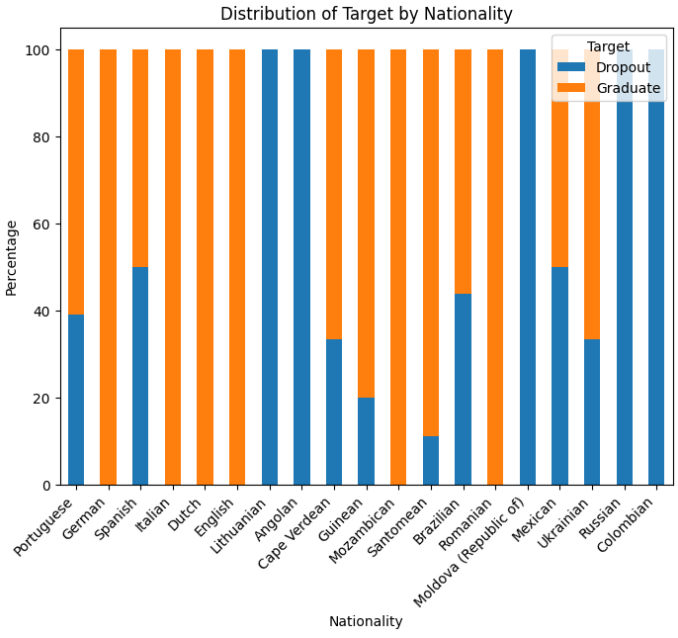

Fig. 6. Distribution of Target by Nationality

The chi-square test yielded a chi-square value of $\chi^2 = 19.85$, degrees of freedom $df = 18$, and a p-value < 0.341, indicating no statistically significant relationship between nationality and academic outcomes.

Socioeconomic factors like Debtor Status and Dropout Rates were also considered in this study. It was hypothesized that students with debtor status are less likely to graduate due to financial stress. The bar chart (Fig. 7) shows that students with debtor status had a dropout rate approximately 119% higher than non-debtors.

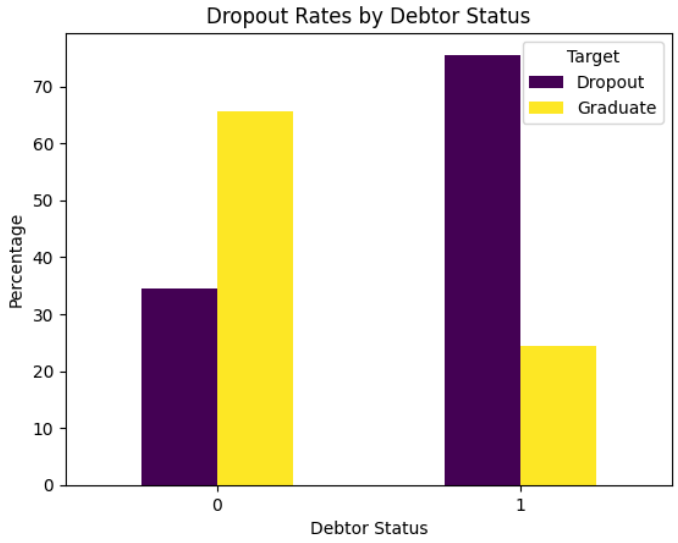

Fig. 7. Dropout Rates by Debtor Status

A chi-square test yielded a chi-square value of $\chi^2 = 257.46$, degrees of freedom $df = 1$, and a p-value < 0.01, rejecting the null hypothesis and confirming a strong association between debtor status and graduation outcomes.

In the area of Socioeconomi factors this study also looks at the relationship between Scholarship Status and Graduation Rates. We hypothesized that students receiving scholarships are more likely to graduate due to reduced financial stress. The bar chart (Fig. 8) shows that among non-scholarship students, 48.35% dropped out, while 51.65% graduated. Among scholarship recipients, only 13.83% dropped out, while 86.17% graduated.

A chi-square test yielded a chi-square value of $\chi^2 = 354.22$, degrees of freedom $df = 1$, and a p-value < 0.0001, indicating a significant association between scholarships and graduation outcomes. Scholarship recipients showed a significantly lower dropout rate compared to non-recipients.

The final area of analysis of this study is Previous Qualifications and Academic Performance. We hypothesized that students with higher previous qualifications achieve better results. The scatter plot (Fig. 9) shows a weak positive correlation (r = 0.077, p $= 0.05$) between previous qualification grades and first-semester performance. Graduates tend to cluster at higher grades, but the correlation remains weak, suggesting other factors influence graduation outcomes.

Linear regression analysis further revealed a slope coefficient of 0.029 and an R^2 value of 0.006, indicating that only 0.6% of the variance in curricular unit grades can be explained by previous qualifications. This suggests previous academic performance is a poor predictor of first-semester success.

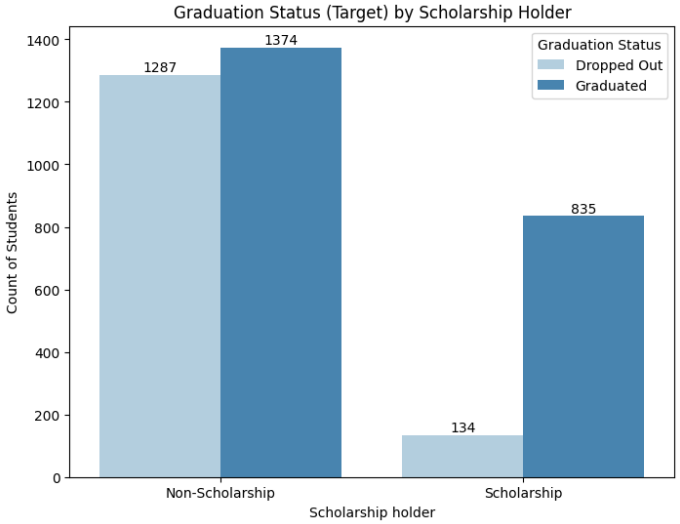

Fig. 8. Graduation Status (Target) by Scholarship Holder

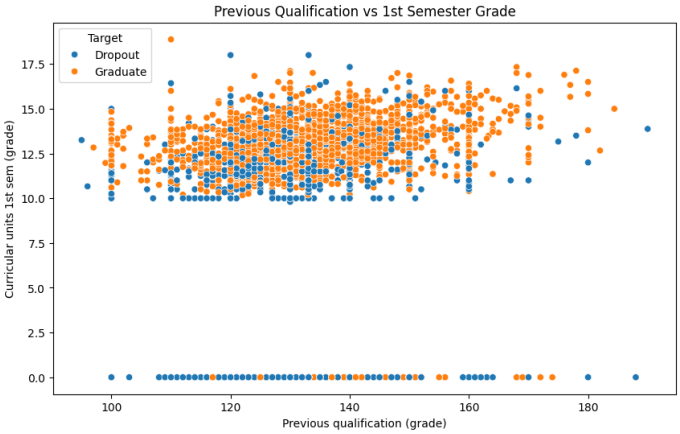

Fig. 9.

The scatter plot (Fig. 10) displays the relationship between previous qualification grades (x-axis) and curricular units 1st semester grades (y-axis), differentiated by student outcomes (dropouts and graduates). The linear regression analysis reveals a very weak positive relationship between these variables, with a slope coefficient of 0.029 and y-intercept of 6.615. This indicates that for each one-point increase in previous qualification grade, the curricular units grade increases by only 0.029 points. The extremely low R-squared value of 0.006 suggests that only 0.6% of the variance in curricular units grades can be explained by previous qualification grades. This remarkably weak coefficient of determination indicates that previous academic performance is a poor predictor of first-semester performance in this dataset.

The visualization (Fig. 10) shows considerable scatter around the regression line, with both graduates and dropouts displaying similar patterns of dispersion. The clustering of points appears relatively uniform across the range of previous qualification grades, though there is a notable presence of zero

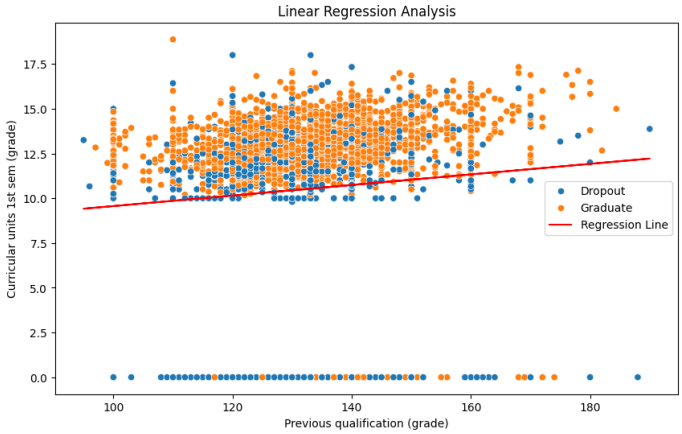

Fig. 10.

grades across all previous qualification levels.

V. Discussion

The study confirmed several key hypotheses regarding factors influencing student dropout and graduation rates in higher education. Age was found to be a significant factor, with younger students demonstrating higher graduation rates. The hypothesis that males are more likely to drop out was also accepted, with males exhibiting a substantially higher dropout rate compared to females. Notably, the hypothesis regarding differences between local and international students was rejected, as nationality showed no significant impact on dropout rates. Socioeconomic factors played a crucial role, with the hypothesis that students with debtor status are less likely to graduate being accepted. Similarly, the hypothesis that scholarship recipients are more likely to graduate was supported, with scholarship holders showing significantly lower dropout rates compared to non-recipients. However, the hypothesis regarding the predictive power of previous qualifications on academic performance was not strongly supported, as only a weak positive correlation was found between previous qualification grades and first-semester performance.

These findings have broader implications for higher education policies and practices. The strong influence of financial factors suggests that institutions should prioritize expanding financial aid programs and developing strategies to mitigate economic stress among students. The gender disparity in dropout rates calls for further investigation into the underlying causes and the development of targeted support systems for male students. Additionally, the weak correlation between previous qualifications and academic performance challenges traditional admission criteria and highlights the need for more comprehensive evaluation methods. Future research could explore psychological factors, such as motivation and resilience, and their impact on student success. Furthermore, longitudinal studies could provide insights into how these factors evolve over a student's academic career, potentially revealing critical intervention points. By addressing these multifaceted aspects, institutions can create more inclusive and supportive learning

environments, ultimately improving retention rates and academic outcomes for diverse student populations

VI. Limitations

This study, while providing valuable insights into factors influencing student dropout and graduation rates, has several limitations. The dataset, originating from a single higher education institution in Portugal, may not be fully representative of global trends in higher education. The focus on undergraduate programs in specific fields limits the generalizability of findings to other academic levels or disciplines. The study's reliance on quantitative data fails to capture qualitative aspects of student experiences, such as motivation, social integration, or personal circumstances, which could significantly impact academic outcomes. Additionally, the analysis of nationality's impact on dropout rates was limited by the small sample sizes for some nationalities, potentially affecting the statistical significance of these findings. The weak correlation between previous qualifications and academic performance suggests that important predictors of student success may not have been captured in the available data. Furthermore, the study's cross-sectional nature does not account for changes in student circumstances or institutional policies over time, which could influence dropout and graduation rates. Future research should address these limitations by incorporating qualitative methods, expanding the scope to multiple institutions and countries, and conducting longitudinal studies to better understand the complex dynamics of student retention and success in higher education.

VII. Conclusion

The research topic provides valuable insights into the factors influencing student success and dropout rates. The findings suggest that addressing financial barriers, demographic risks, and academic challenges collectively can create a more supportive educational environment. Further research is recommended to explore additional predictors, such as psychological factors and institutional practices, to gain a comprehensive understanding of student retention and success. The results underscore the need for multifaceted support strategies to address student dropout rates.

References

[1] A. P. Rovai, "In search of higher persistence rates in distance education online programs," *The Internet and Higher Education*, vol. 6, no. 1, pp. 1–16, 2003.

[2] T. Zajac, F. Perales, W. Tomaszewski, N. Xiang, and S. R. Zubrick, "Student mental health and dropout from higher education: an analysis of australian administrative data," *Higher Education*, vol. 87, pp. 325–343, Feb 2023.

[3] S. Martín-Arbós, E. Castarlenas, F. Morales-Vives, and J. M. Dueñas, "Students' thoughts about dropping out: Sociodemographic factors and the role of academic help-seeking," *Social Psychology of Education*, Mar 2024.

[4] B. M. Kehm, M. R. Larsen, and H. B. Sommersel, "Student dropout from universities in europe: A review of empirical literature," *Hungarian Educational Research Journal*, vol. 9, no. 2, pp. 147–164, 2019.

[5] S. Srairi, "An analysis of factors affecting student dropout: The case of tunisian universities," *Research in Higher Education*, vol. 31, no. 2, pp. 183–198, 2022.

[6] N. Nurmalitasari, Z. A. Long, and M. F. M. Noor, "Factors influencing dropout students in higher education," *Education Research International*, vol. 2023, p. Article ID 7704142, 2023.

[7] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early prediction of student's performance in higher education: a case study," in *Trends and Applications in Information Systems and Technologies*, ser. Advances in Intelligent Systems and Computing. Springer, 2021, vol. 1.