STATISTICAL ANALYSIS REPORT

By: Blessing Ilesanmi

13th September, 2024

INTRODUCTION

The increasing rate of student dropouts is a significant challenge faced by educational institutions globally. Understanding the factors contributing to these dropout rates is critical for developing strategies to improve retention and academic success. The present report aims to examine the interplay of socio-economic backgrounds, academic performance, and financial aid status in influencing student dropout rates. By employing descriptive statistics, correlation analysis, and hypothesis testing, the report seeks to uncover underlying trends that may offer insights into why students leave their academic programs before completion.

The dataset under consideration includes a diverse array of variables such as demographic information (e.g., age, gender, nationality), socio-economic indicators (e.g., parental occupation, financial aid), and academic metrics (e.g., GPA, curricular units, grades). Additionally, external economic factors like unemployment rates, inflation, and GDP are factored into the analysis to provide a broader context for student performance. This holistic approach allows for a comprehensive analysis that can be valuable for educational institutions in identifying the most at-risk student populations and developing targeted interventions.

This report is structured into several key sections: descriptive statistics, correlation analysis, and hypothesis testing. Each section contributes uniquely to the overarching objective of understanding the factors influencing dropout rates.

DESCRIPTIVE STATISTICS

Descriptive statistics provide a summary of the main features of the dataset, including measures of central tendency, variability, and distribution. Here's an overview of the descriptive statistics for each variable in the dataset:

- Marital status: The mean is 0.18 with a standard deviation of 0.61, indicating a highly skewed distribution where most entries are 0, suggesting a low prevalence of certain marital statuses in the dataset.
- Application mode: The mean value is 5.89 with a standard deviation of 5.30. The range spans from 0 to 17, showing a broad variability in application modes used.
- Application order: The mean is 0.19 with a standard deviation of 0.15, indicating most application orders are clustered near the lower end of the scale.
- Course: The mean is 8.90 with a standard deviation of 4.33, reflecting that the course numbers vary widely across the dataset.
- Daytime/evening attendance: A high mean of 0.89 with a standard deviation of 0.31 indicates most students attend courses in the daytime.
- Previous qualification: The mean is 1.53 with a standard deviation of 3.96, indicating a wide range of previous qualifications among students.

- Previous qualification (grade): The mean is 0.49 with a standard deviation of 0.17. The grades are generally distributed around the middle of the scale.
- Nationality: The mean is 0.25 with a high standard deviation of 1.75, indicating diverse nationalities and significant variability.
- Mother's qualification: The mean is 11.32 with a standard deviation of 9.03, showing a wide range of qualifications among students' mothers.
- Father's qualification: The mean is 15.46 with a standard deviation of 11.04, indicating a broad spectrum of father's qualifications.
- Mother's occupation: The mean is 6.32 with a standard deviation of 3.99, showing diverse occupations among mothers.
- Father's occupation: The mean is 6.82 with a standard deviation of 4.86, indicating a range of occupations among fathers.
- Admission grade: The mean is 0.43 with a standard deviation of 0.19. The values are generally centered around the middle of the scale.
- Displaced: The mean is 0.55 with a standard deviation of 0.50, showing that about half of the students are categorized as displaced.
- Educational special needs: The mean is 0.01 with a standard deviation of 0.11, suggesting a very small proportion of students have educational special needs.
- Debtor: The mean is 0.11 with a standard deviation of 0.32, indicating that a small percentage of students are debtors.
- Tuition fees up to date: The mean is 0.88 with a standard deviation of 0.32, showing most students have up-to-date tuition fees.
- Gender: The mean is 0.35 with a standard deviation of 0.48, indicating a roughly even distribution of genders with a tendency towards one gender.
- Scholarship holder: The mean is 0.25 with a standard deviation of 0.43, showing that a quarter of students hold scholarships.
- Age at enrollment: The mean is 0.12 with a standard deviation of 0.14, reflecting a narrow range of ages at enrollment.
- International: The mean is 0.02 with a standard deviation of 0.16, indicating a very small proportion of international students.
- Curricular units 1st sem (credited): The mean is 0.04 with a standard deviation of 0.12, reflecting low credit acquisition in the first semester.
- Curricular units 1st sem (enrolled): The mean is 0.24 with a standard deviation of 0.10, showing moderate enrollment in the first semester.
- Curricular units 1st sem (evaluations): The mean is 0.18 with a standard deviation of 0.09, indicating evaluations are slightly less common.

- Curricular units 1st sem (approved): The mean is 0.18 with a standard deviation of 0.12, showing a fair proportion of approved units.
- Curricular units 1st sem (grade): The mean is 0.56 with a standard deviation of 0.26, suggesting a moderate level of grading.
- Curricular units 1st sem (without evaluations): The mean is 0.01 with a standard deviation of 0.06, indicating few units without evaluations.
- Curricular units 2nd sem (credited): The mean is 0.03 with a standard deviation of 0.10, showing low credit acquisition in the second semester.
- Curricular units 2nd sem (enrolled): The mean is 0.27 with a standard deviation of 0.10, reflecting moderate enrollment in the second semester.
- Curricular units 2nd sem (evaluations): The mean is 0.24 with a standard deviation of 0.12, showing evaluations are somewhat common.
- Curricular units 2nd sem (approved): The mean is 0.22 with a standard deviation of 0.15, indicating a fair proportion of approved units.
- Curricular units 2nd sem (grade): The mean is 0.55 with a standard deviation of 0.28, showing a moderate level of grading.
- Curricular units 2nd sem (without evaluations): The mean is 0.01 with a standard deviation of 0.06, indicating few units without evaluations.
- Unemployment rate: The mean is 0.46 with a standard deviation of 0.31, reflecting a moderate range of unemployment rates.
- Inflation rate: The mean is 0.45 with a standard deviation of 0.31, indicating a moderate range of inflation rates.
- GDP: The mean is 0.54 with a standard deviation of 0.30, showing a broad range of GDP values.
- Target: The mean is 1.18 with a standard deviation of 0.89, indicating most students fall into a certain category with respect to the target variable.
- Total Curricular Units 1st Semester: The mean is 0.28 with a standard deviation of 0.20, reflecting moderate total curricular units in the first semester.
- Total Curricular Units 2nd Semester: The mean is 0.30 with a standard deviation of 0.18, showing moderate total curricular units in the second semester.
- Total Credits Earned: The mean is 0.06 with a standard deviation of 0.22, indicating low total credits earned.
- Total Units Enrolled: The mean is 0.51 with a standard deviation of 0.19, reflecting moderate enrollment in units.
- Weighted Grade 1st Semester: The mean is 0.02 with a standard deviation of 0.08, showing low weighted grades in the first semester.

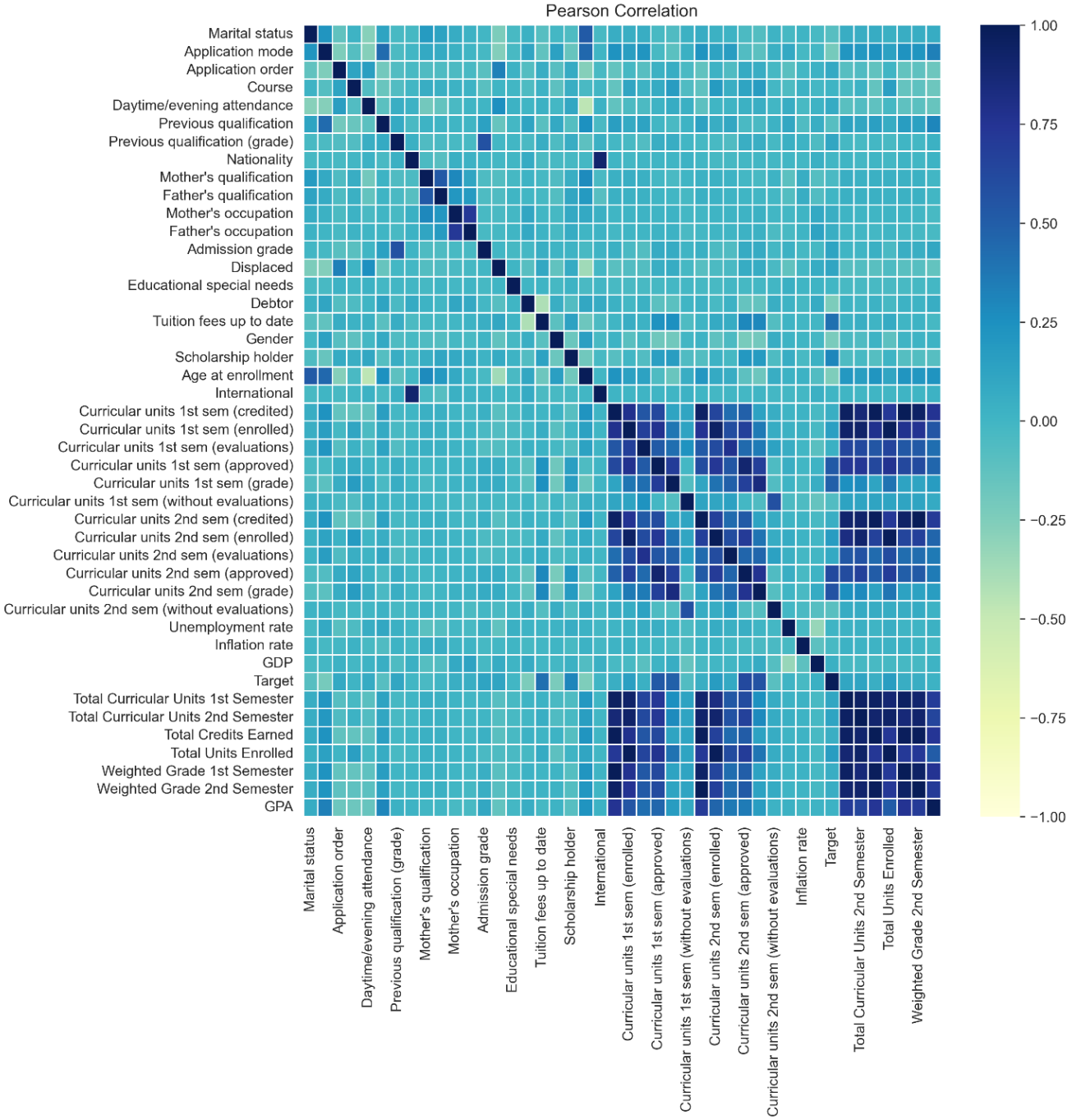
- Weighted Grade 2nd Semester: The mean is 0.02 with a standard deviation of 0.07, reflecting low weighted grades in the second semester.
- GPA: The mean is 0.09 with a standard deviation of 0.23, indicating a generally low GPA across the dataset.

The descriptive statistics reveal a diverse student population with varying academic backgrounds, demographics, and performance. Students come from different countries and have diverse parental education and occupations. While most students attend daytime classes and have up-to-date tuition fees, a significant portion are displaced and have average grades. Economic indicators like unemployment, inflation, and GDP also vary within the dataset. Overall, the students in this dataset demonstrate a range of characteristics and performance levels.

CORRELATION MATRIX HEATMAP

The correlation matrix heatmap shows the pairwise Pearson correlation coefficients between numeric variables in the dataset. The heatmap provides a visual representation of the Pearson correlation coefficients between various variables related to student performance. The colour scale ranges from blue to yellow, indicating negative to positive correlations, respectively.

- **Strong Positive Correlations:**
    - **Target variable (GPA) and Total Credits Earned:** This suggests that students with a higher number of credits earned tend to have higher GPAs.
    - **Total Curricular Units 1st Semester and Total Curricular Units 2nd Semester:** These variables are highly correlated, indicating that students who enroll in more units in the first semester are likely to also enroll in more units in the second semester.
    - **Weighted Grade 1st Semester and Weighted Grade 2nd Semester:** Similarly, students who perform well in the first semester are more likely to perform well in the second semester.
    - **Curricular Units 1st Sem (approved) and Curricular Units 2nd Sem (approved):** Students who successfully complete units in the first semester are more likely to successfully complete units in the second semester.


Pearson Correlation
Marital status
Application mode
Application order
Course
Daytime/evening attendance
Previous qualification
Previous qualification (grade)
Nationality
Mother's qualification
Father's qualification
Mother's occupation
Father's occupation
Admission grade
Displaced
Educational special needs
Debtor
Tuition fees up to date
Gender
Scholarship holder
Age at enrollment
International
Curricular units 1st sem (credited)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (evaluations)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Curricular units 1st sem (without evaluations)
Curricular units 2nd sem (credited)
Curricular units 2nd sem (enrolled)
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (grade)
Curricular units 2nd sem (without evaluations)
Unemployment rate
Inflation rate
GDP
Target
Total Curricular Units 1st Semester
Total Curricular Units 2nd Semester
Total Credits Earned
Total Units Enrolled
Weighted Grade 1st Semester
Weighted Grade 2nd Semester
GPA
Marital status
Application order
Daytime/evening attendance
Previous qualification (grade)
Mother's qualification
Mother's occupation
Admission grade
Educational special needs
Tuition fees up to date
Scholarship holder
International
Curricular units 1st sem (enrolled)
Curricular units 1st sem (approved)
Curricular units 1st sem (without evaluations)
Curricular units 2nd sem (enrolled)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (without evaluations)
Inflation rate
Target
Total Curricular Units 2nd Semester
Total Units Enrolled
Weighted Grade 2nd Semester
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

- **Strong Negative Correlations:**
    - **Target variable (GPA) and Unemployment Rate:** This suggests that students from regions with higher unemployment rates tend to have lower GPAs.
    - **Total Curricular Units 2nd Semester and Previous Qualification (grade):** Students with lower previous qualifications tend to enroll in fewer units in the second semester.
    - **Weighted Grade 2nd Semester and Previous Qualification (grade):** Students with lower previous qualifications tend to have lower grades in the second semester.
- **Moderate Correlations:**
    - **Target variable (GPA) and Inflation Rate:** There is a moderate negative correlation between GPA and inflation rate, suggesting that higher inflation rates may have a slightly negative impact on student performance.
    - **Total Credits Earned and Previous Qualification (grade):** Students with higher previous qualifications tend to earn more credits.

Other Observations:

- Marital status, Application mode, Application order, Daytime/evening attendance, Nationality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Admission grade, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment, International, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (without evaluations), and GDP: These variables show weak or no correlations with the target variable (GPA) and other variables.

Overall, the heatmap provides valuable insights into the relationships between various factors and student performance. These findings can be used to identify potential areas for improvement and inform educational policies.

Results and Interpretation of Hypothesis Tests

Hypothesis 1: Higher socio-economic status correlates with lower dropout rates.

- **Chi-square statistic:** 264.50
- **P-value:** 4.52e-19

Interpretation: The very low p-value indicates a strong association between socio-economic status (as indicated by the father's occupation) and dropout rates. Since the p-value is far

below the typical significance level (0.05), I reject the null hypothesis and conclude that higher socio-economic status correlates with lower dropout rates.

Hypothesis 2: Students with higher admission grades are less likely to drop out.

- **T-statistic:** 0.89
- **P-value:** 0.37

Interpretation: The p-value is greater than the significance level (0.05), indicating that there is not enough evidence to reject the null hypothesis. Therefore, I do not find a statistically significant relationship between admission grades and dropout rates.

Hypothesis 3: Dropout rates are lower among students receiving financial aid or **scholarships.**

- **T-statistic:** 20.73
- **P-value:** 3.63e-91

Interpretation: The p-value is much smaller than the significance level (0.05), indicating that there is sufficient evidence to reject the null hypothesis. Therefore, I find a statistically significant relationship between receiving financial aid or scholarships and lower dropout rates. Students who receive financial aid or scholarships have significantly lower dropout rates compared to those who do not receive such financial support.

Conclusion

The descriptive statistics provided a comprehensive overview of the data, highlighting central tendencies for each variable. The correlation matrix revealed strong positive correlations between factors like total credits earned and GPA, while also identifying negative correlations between unemployment rate and GPA. Hypothesis testing further explored specific relationships. We found strong evidence to reject the null hypothesis for both hypotheses 1 and 3, suggesting that higher socio-economic status (as measured by father's occupation) and receiving scholarships are associated with lower dropout rates. However, the evidence for hypothesis 2, suggesting a connection between higher admission grades and lower dropout rates, was not statistically significant.

These findings offer valuable insights for educational institutions. By understanding the factors that influence student success, institutions can develop targeted interventions and support systems to improve retention rates and promote student achievement.