Student Dropout Prediction Project: Detailed EDA and Preliminary Analysis Report

Introduction

Student dropout, graduation, and continued enrollment represent significant educational outcomes impacting both individual futures and broader societal well-being. Dropouts experience diminished economic opportunities, while graduates often enjoy better employment prospects. Accurately predicting these outcomes—dropout, enrolled, and graduate—enables institutions to intervene proactively, supporting at-risk students.

This project aims to predict student outcomes by identifying patterns and key predictors within a dataset of 4424 records featuring 35 demographic, academic, and socioeconomic variables, including attendance rates, GPA, gender, family income, and parental education.

Feature 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: 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)
- Educational 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 unit’s 1st sem (credited): The number of curricular units credited by the student in the first semester. (Numerical)
- Curricular unit’s 1st sem (enrolled): The number of curricular units enrolled by the student in the first semester. (Numerical)
- Curricular unit’s 1st sem (evaluations): The number of curricular units evaluated by the student in the first semester. (Numerical)
- Curricular unit’s 1st sem (approved): The number of curricular units approved by the student in the first semester. (Numerical)

Libraries used in the notebook.

1. **Import important packages**

- import pandas as pd
- import numpy as np
- import seaborn as sns
- import matplotlib.pyplot as plt

2. **Data Preprocessing and EDA**

- from sklearn.preprocessing import OrdinalEncoder
- from scipy.stats import chi2_contingency

Data Cleaning

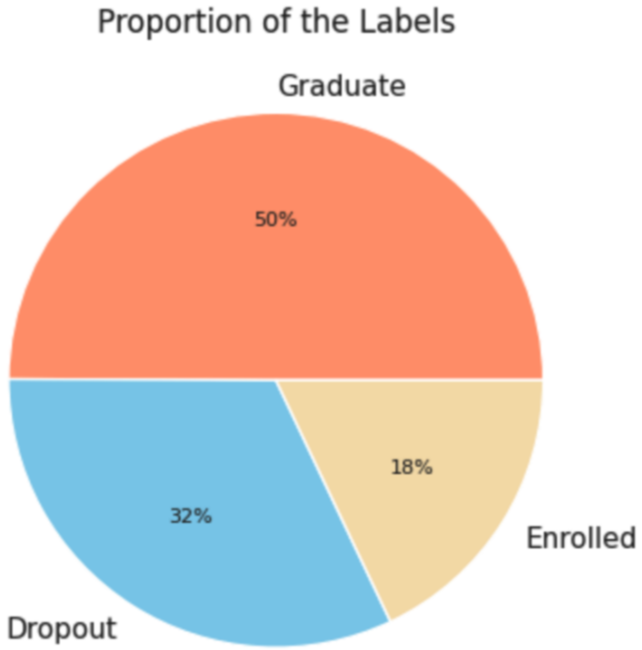
- Renamed columns to remove whitespace and parentheses for consistency.
- Handled whitespace and formatting inconsistencies in categorical variables.
- Converted categorical variables (e.g., Gender, School_Type) to the 'category' data type, optimizing memory and facilitating accurate analysis.

Exploratory Data Analysis (EDA)

Overall Approach

I systematically explored relationships between features and the multivariate target variable (dropout, enrolled, graduate) using visualizations and statistical tests to uncover significant predictors.

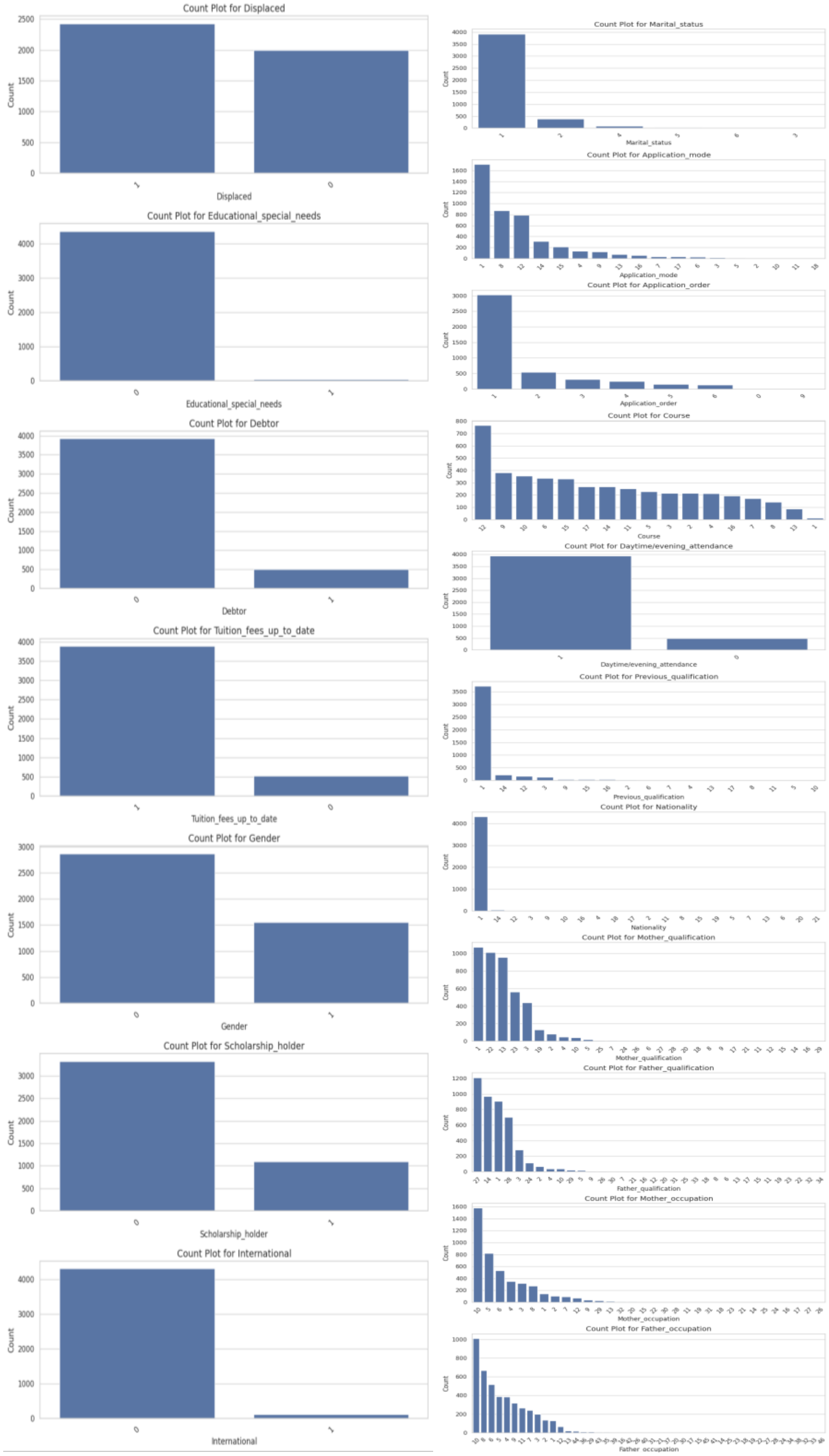
Proportion of the Labels



From the pie chart above we can see that the data is imbalanced: with about 50% of the labels are 'Graduate', 32% are 'Dropout', and 18% are 'Enrolled'. The labels are encoded as ordinal data -- 0 represents 'Dropout', 1 represents 'Enrolled', and 2 represents 'Graduate' -- since most classification models only handle numeric values.

Analysis of Categorical Features

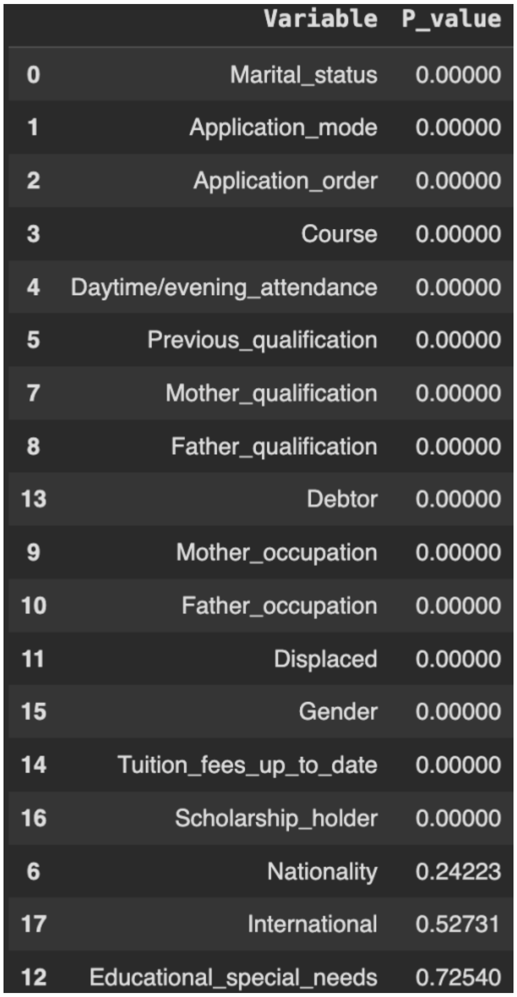
Univariate Analysis - Count plots were generated for categorical features like 'Gender', 'School_Type', and 'Parent_education_level'. Etc.



Interpretation:

- Marital Status is heavily skewed toward one group, which may reduce its impact unless the minority classes have distinct dropout patterns.
- Application Mode and Order show clear preference for a few categories. These could signal motivation or institutional processes and might be useful predictors.
- Course shows a broad distribution, which is promising—it could help the model differentiate dropout risk across programs.
- Daytime/Evening Attendance is mostly daytime, but evening students might represent a unique group (e.g., working students), possibly influencing outcomes.
- Previous Qualification and Nationality have low variability, especially nationality, which may limit their predictive power.
- Parental Education and Occupation have high cardinality and long tails. They might capture socioeconomic factors if grouped smartly.
- Displaced, Special Needs, Debtor, and Tuition Fees are binary and imbalanced, but they likely capture key risk factors—especially financial-related ones.
- Gender, Scholarship Holder, and International Status are slightly imbalanced. Their usefulness will depend on how they interact with other features.
- Overall, I'll revisit these after initial model runs to evaluate their importance and decide if any features need transformation, grouping, or exclusion.

Bivariate Analysis (Chi-Square Test) - Chi-Square tests were conducted to determine associations between categorical features and the three outcomes:
(Insert Chi-Square test results table here)
Interpretation:



	Variable	P_value
0	Marital_status	0.00000
1	Application_mode	0.00000
2	Application_order	0.00000
3	Course	0.00000
4	Daytime/evening_attendance	0.00000
5	Previous_qualification	0.00000
7	Mother_qualification	0.00000
8	Father_qualification	0.00000
13	Debtor	0.00000
9	Mother_occupation	0.00000
10	Father_occupation	0.00000
11	Displaced	0.00000
15	Gender	0.00000
14	Tuition_fees_up_to_date	0.00000
16	Scholarship_holder	0.00000
6	Nationality	0.24223
17	International	0.52731
12	Educational_special_needs	0.72540

Most of the p-values are close to zero, except for three variables ('Nationality', 'International', 'Educational_special_needs') with very high p-values (0.24, 0.53, 0.73), indicating that no statistically significant association between these three features and the label. They are excluded from model.

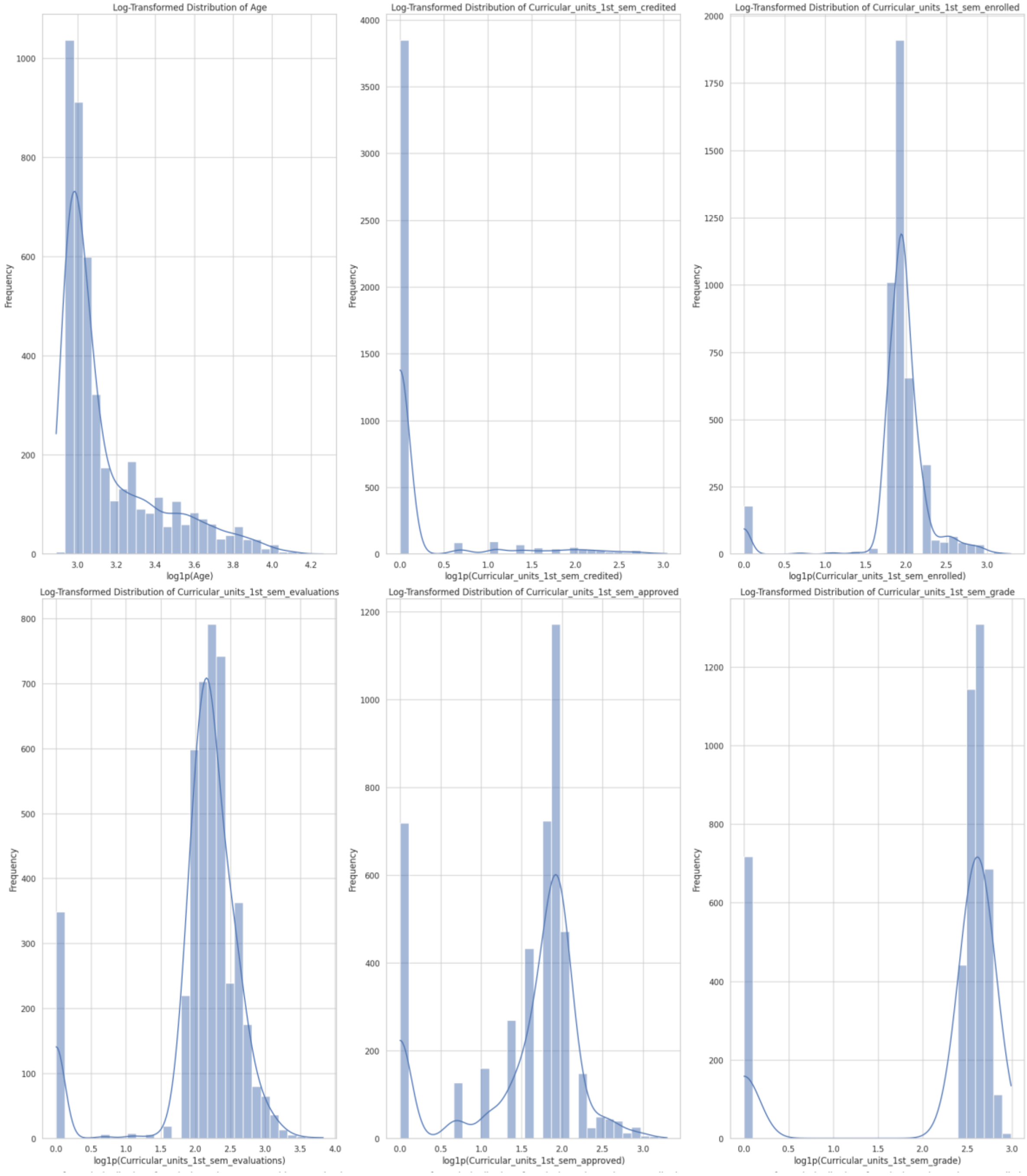
Analysis of Numerical Features

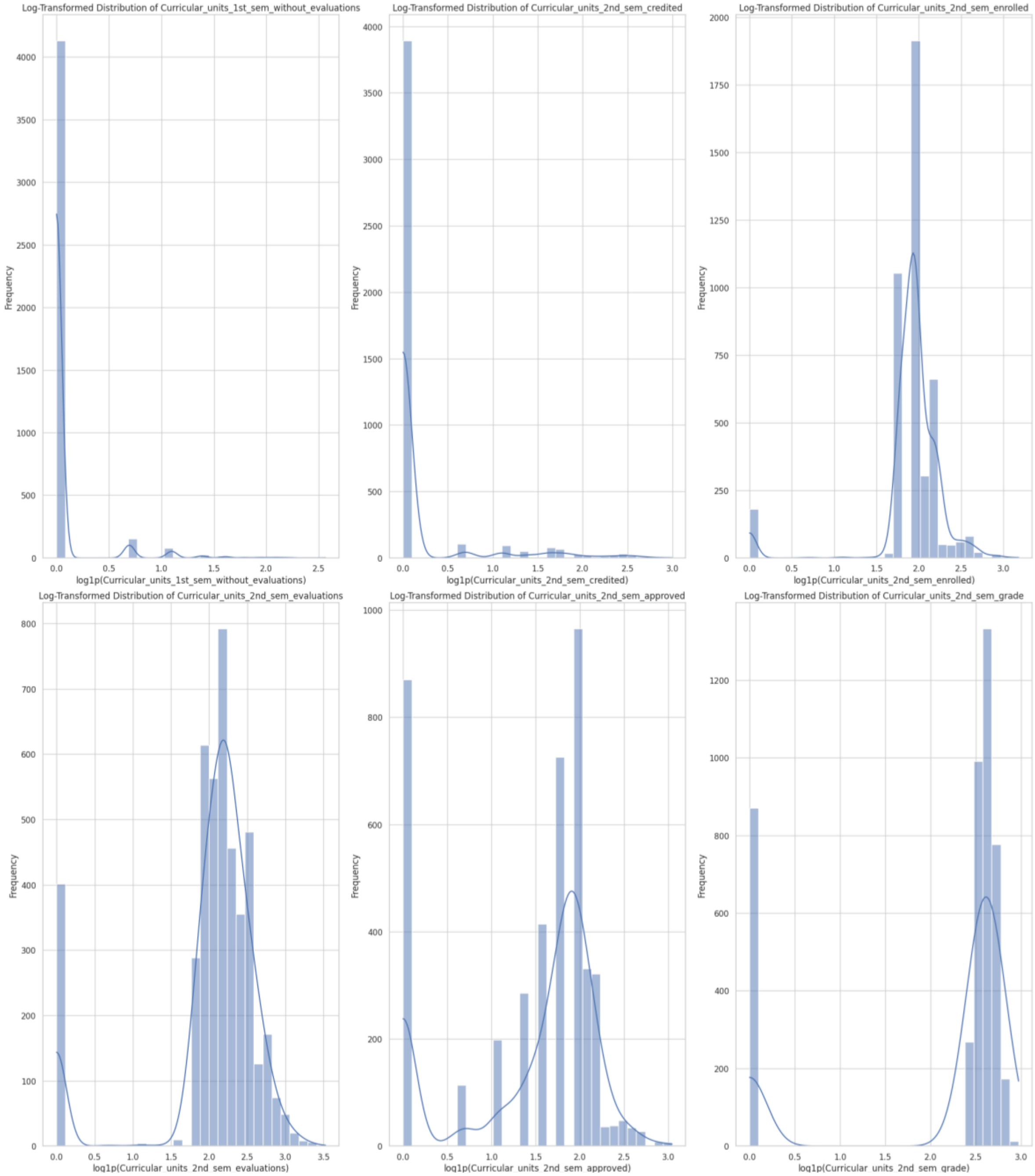
Estimates of Location and Variability –
https://colab.research.google.com/drive/1wYYqiduWaAtsj4tIQ3XHqQgSl8U3n1cd#scrollTo=gezQNhXN6Zew&line=1&uniqifier=1

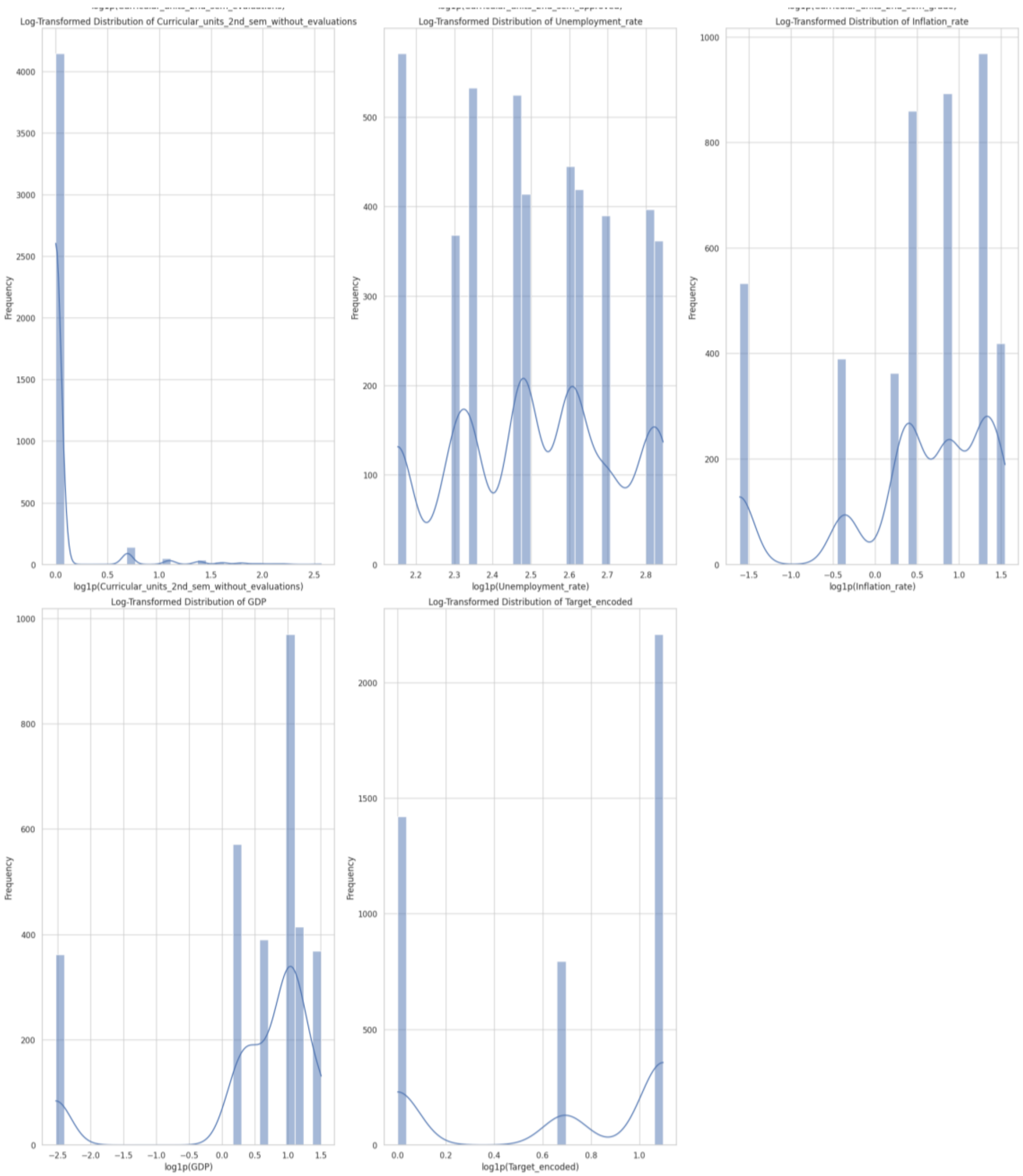
- Looking at Age, Curricular_units_1st_sem_credited and Curricular_units_1st_sem_enrolled, its evident that there are outliers in the data.
- Age - 75% quartile is up-to 25 and max is 70 years of age, which is a huge jump.

- Curricular_units_1st_sem_credited - 75% quartile is up-to 0 and max is at 20 credits, which is a huge jump. Similarly, it is same with Curricular_units_1st_sem_enrolled.

Univariate Analysis - Histograms to visualize the distribution of numerical feature.




Log-Transformed Distribution of Curricular_units_1st_sem_without_evaluations
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.5
1.0
1.5
2.0
2.5
log1p(Curricular_units_1st_sem_without_evaluations)
Log-Transformed Distribution of Curricular_units_2nd_sem_credited
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log1p(Curricular_units_2nd_sem_credited)
Log-Transformed Distribution of Curricular_units_2nd_sem_enrolled
Frequency
0
250
500
750
1000
1250
1500
1750
2000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log1p(Curricular_units_2nd_sem_enrolled)
Log-Transformed Distribution of Curricular_units_2nd_sem_evaluations
Frequency
0
100
200
300
400
500
600
700
800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
log1p(Curricular_units_2nd_sem_evaluations)
Log-Transformed Distribution of Curricular_units_2nd_sem_approved
Frequency
0
200
400
600
800
1000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log1p(Curricular_units_2nd_sem_approved)
Log-Transformed Distribution of Curricular_units_2nd_sem_grade
Frequency
0
200
400
600
800
1000
1200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log1p(Curricular_units_2nd_sem_grade)



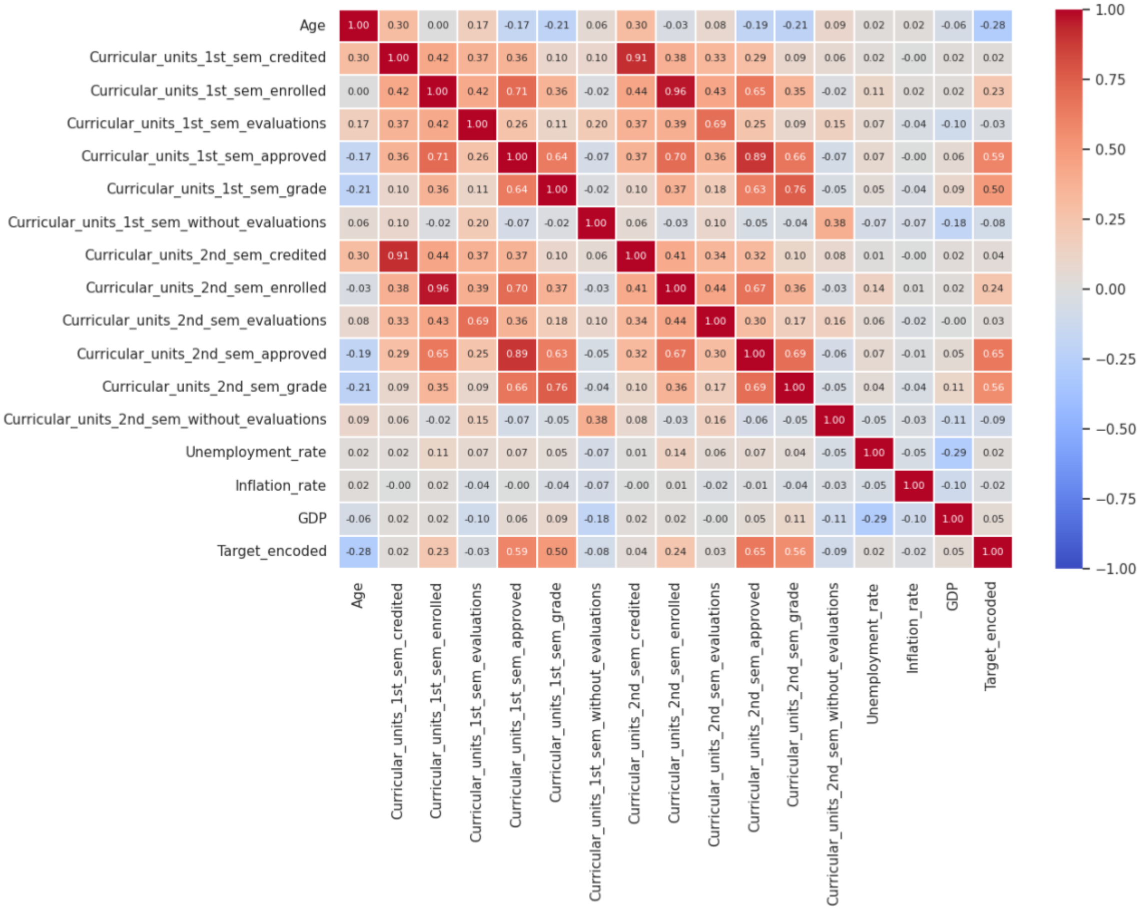
Interpretation:

Looking at these log-transformed histograms, I can tell that a lot of the data was originally skewed — probably with long tails or a lot of zeros. After applying the log transformation, the shapes of the distributions look a lot more normalized or at least smoothed out.

- Age: has a right skew originally, but the log transformation helped compress the extreme values. Now, I can clearly see that most students fall into a narrow age band.

- Curricular Units (1st & 2nd Sem): For things like curricular_units_1st_sem_approved, grade, evaluations, and similar for the 2nd semester, I see clear peaks and groupings. The log scale helped reveal subtleties, like small groups with low or zero completions or grades. Some of the variables still have sharp peaks — maybe a lot of students scored zeros or maxed out the units.
- Unemployment Rate & Inflation: These were categorical or rounded values originally, so even after log-transforming, they still look like bar charts. They seem to cluster around specific national benchmarks, which might reflect broader economic conditions.

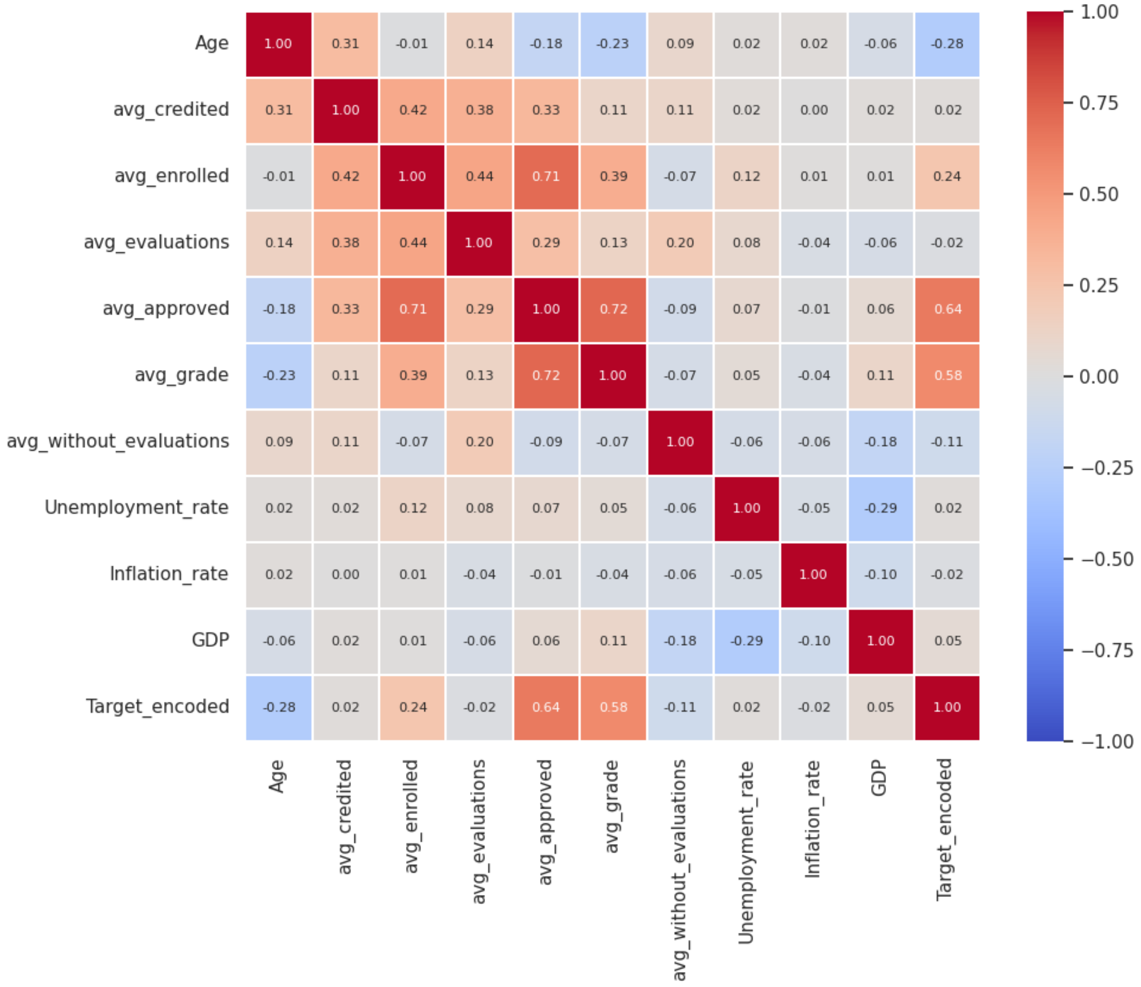
Bivariate Analysis - Spearman's rank correlation between numerical features and the Target



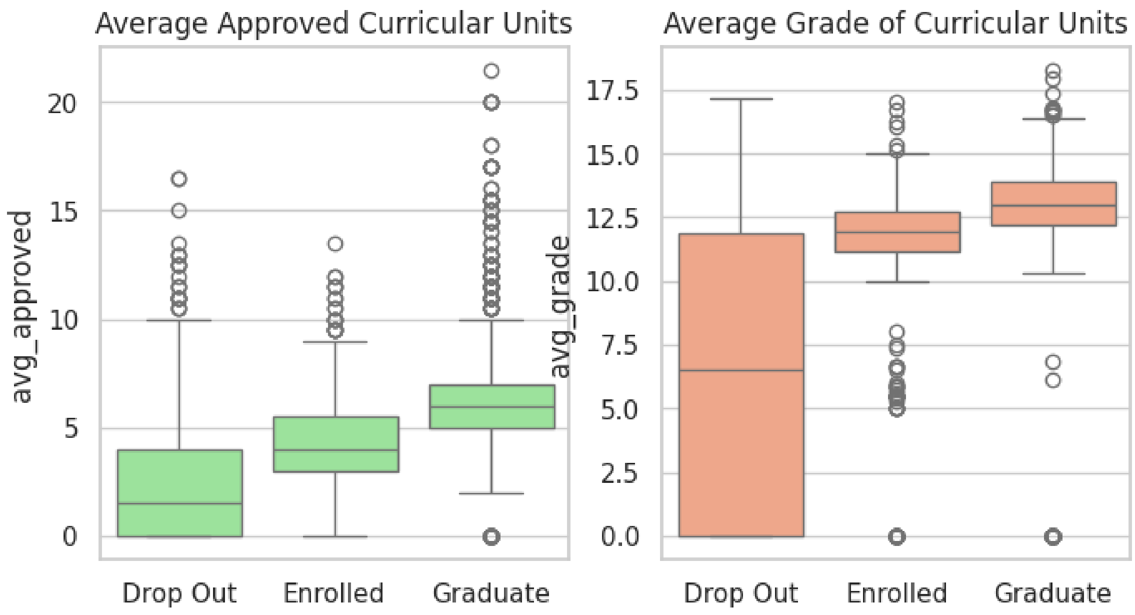
Interpretation:

- As can be seen from the heat map, there are four features ('Curricular_units_2nd_sem_approved', 'Curricular_units_2nd_sem_grade', 'Curricular_units_1st_sem_approved', 'Curricular_units_1st_sem_grade') that have relatively high and positive correlations with the label, while some have very low correlations(e.g., 'Unemployment_rate', 'Inflation_rate')

- The heat map also reveals multicollinearity among the features related to curricular units. These features represent students' academic performance at the end of the first and second semesters. I will aggregate them to get the average value between the two semesters.4. Numerical Feature Analysis



- The new correlation matrix above shows that 'curri_avg_approved' and 'curri_avg_grade' still have a relatively high correlation with the labels ('Target_encoded'), while 'curri_avg_credited' and 'curri_avg_evaluations', along with 'the macroeconomic data ('Unemployment_rate', 'Inflation_rate'), have very low correlations, all between -0.02 and 0.02. I will exclude these four features.
- The multicollinearity still exists among the academic data. I'll take it into account when selecting the models.
- Let's check how 'curri_avg_approved' and 'curri_avg_grade' are associated with students' situation at the end of the normal duration of the course.



It's not surprising that 'Graduate' is associated with more approved curricular units and higher grades. However, there are some instances of a 0 value for average grade and average approved curricular units in the 'Graduate' class.

Data Preparation for Modeling

Feature Aggregation:

- Academic performance data from two semesters were aggregated into features such as `avg_credited`, `avg_enrolled` to enhance predictive accuracy as shown above.

Re-evaluation of Correlations:

- Post-aggregation correlations demonstrated stronger associations with the multivariate outcomes, affirming the aggregation approach as shown above in heatmap.

Outlier Removal:

- Statistical criteria identified and removed outliers, ensuring data integrity and model effectiveness.

Feature Selection:

- Irrelevant or redundant features were dropped to streamline the dataset for modeling:
- The final dataset focused on highly predictive variables, ensuring robust modeling.

Discussions and Insight summary

Class Imbalance:

- The dataset clearly shows a significant class imbalance: approximately 50% Graduates, 32% Dropouts, and 18% currently Enrolled.
- Class imbalance may bias predictive models towards the majority class, potentially reducing prediction accuracy for minority classes (particularly "Enrolled").

Categorical Features:

- Several categorical features like Parental Education, Course, and Daytime/Evening Attendance appear promising due to clear variations across classes.
- Low variability in Nationality, International status, and Educational special needs justify their exclusion based on Chi-Square tests.

Numerical Features and Transformations:

- Log transformations effectively normalized skewed distributions (e.g., Age, Curricular units approved), enhancing interpretability and likely predictive accuracy.
- Strong correlations identified (Curricular_units_approved, Curricular_units_grade) are meaningful and expectedly crucial for modeling dropout or graduation outcomes.
- Macroeconomic indicators (Unemployment rate, Inflation rate) showed negligible correlations and rightly excluded.

Multicollinearity:

- Multicollinearity is observed among academically related features (e.g., curricular units). Aggregating these features is appropriate and improves correlation with outcomes.

Outliers:

- Outlier removal enhanced data integrity. Continued vigilance for outliers, particularly post-aggregation, remains vital.

Next Steps for Further Analysis/Model Development:

Model Development and Evaluation:

Start building baseline models such as:

- Logistic Regression (with regularization)
- Decision Trees
- Random Forest
- Gradient Boosting (XGBoost, CatBoost, LightGBM)

Evaluate models rigorously using metrics suitable for imbalanced datasets (e.g., F1-score, AUC-ROC, Precision-Recall curves).

Model Validation:

Implement cross-validation (e.g., k-fold cross-validation) and hyperparameter tuning to ensure model robustness.

Team Member	Contributions
Balram Iyengar	Led the data cleaning process, including renaming columns, handling inconsistencies, encoding categorical variables, and optimizing memory usage. Performed detailed exploratory data analysis (EDA) using visualizations and statistical tests to uncover significant predictors. Also worked on interpreting categorical and numerical feature distributions.
Dhanush	Conducted bivariate analysis using Chi-Square tests for categorical variables and Spearman's correlation for numerical features. Identified important features and excluded irrelevant ones (e.g., nationality, macroeconomic indicators). Analyzed class imbalance and provided insights on data distribution and transformations.
Maneesh Rao	Focused on preparing data for modeling, including outlier removal, feature aggregation (e.g., averaging curricular units performance), and re-evaluating correlations. Contributed to discussing multicollinearity handling and designed the next steps for model development, including selection of baseline models and validation strategy.