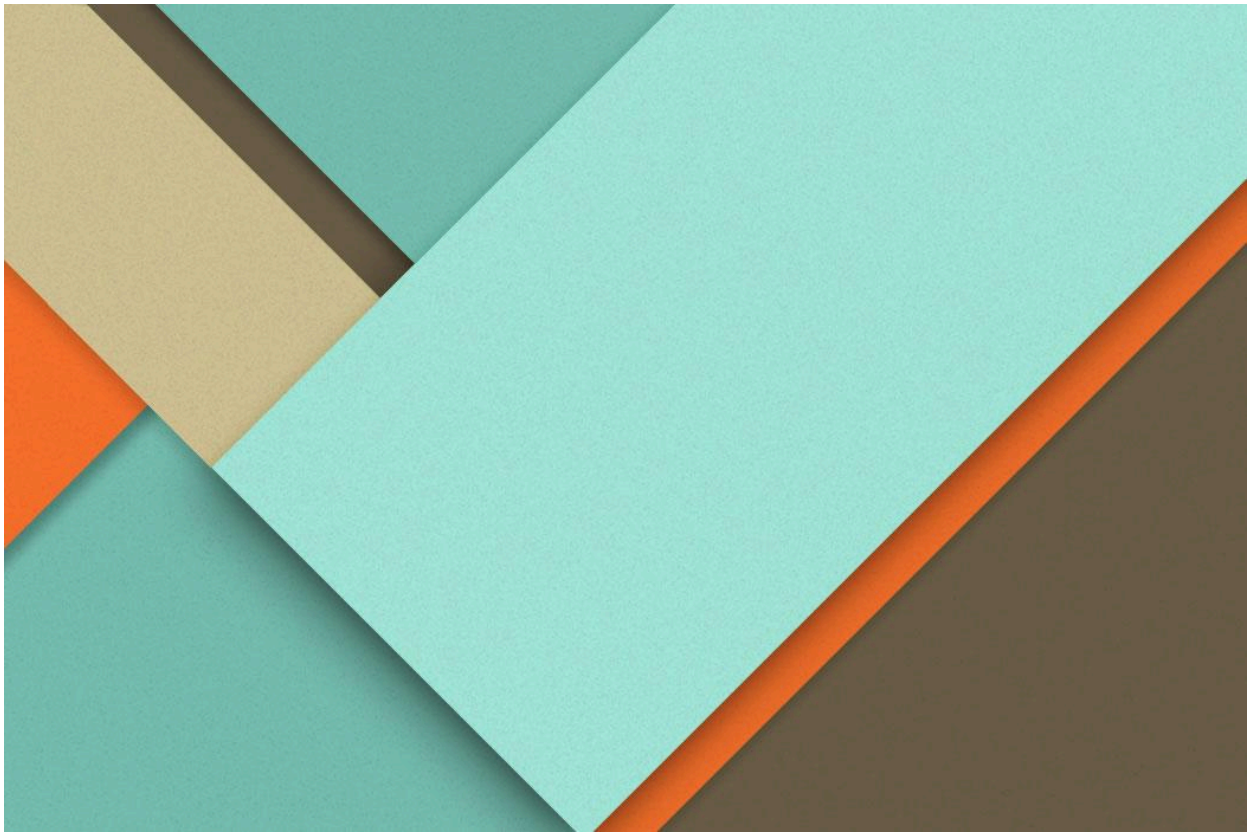


Student Dropout and Academic Success Prediction

12.03.2024

Hemanth Rayudu
NUID : 002273160
Under the Guidance of: Professor Hong Pan, Ph.D.

Overview

This project aims to analyze and predict undergraduate students' academic outcomes, focusing on dropout, continued enrollment, and graduation. Using machine learning and statistical techniques, the project identifies key factors influencing student success and provides actionable insights to reduce dropout rates.

Goals

1. Identify the primary factors influencing student academic outcomes.
2. Build predictive models for classifying student statuses into "Dropout," "Enrolled," and "Graduate."
3. Validate the relevance of academic, demographic, and socio-economic features through statistical methods and visualizations.

Dataset Description

The **Student Dropout Prediction Dataset** contains **36 features** describing various academic, demographic, and socio-economic characteristics of undergraduate students, with a total of **4,424 rows**.

Features:

I. Academic Attributes:
- **Grades:** Admission grades, semester grades, and curricular unit evaluations (1st and 2nd semester).
- **Credits:** Curricular units credited, enrolled, and approved.
- **Performance Metrics:** Academic success indicators like grades and approvals.

II. Demographic Attributes:
- **Personal Information:** Age, gender, nationality, marital status.
- **Parental Information:** Parents' qualifications and occupations.

III. Socio-Economic Attributes:
- **Financial:** Tuition fee payment status, scholarship holding status.
- **External Factors:** Unemployment rate, inflation rate, and GDP.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan.

Data Preprocessing

I. Handling Missing Values

No missing values detected in the dataset.

II. Feature Engineering:

Categorical variables were encoded, and numerical variables were normalized where necessary.

III. Outlier Removal:

Outliers identified using Z-scores and visualized through boxplots.

IV. Class Balancing:

Applied SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset with 2,209 samples for each class.

Exploratory Data Analysis (EDA)

I. Insights from Visualizations:

- Over 90% of students pay tuition fees on time.
- Female students comprise more than 60% of the dataset.
- Most students enrolled between ages 18-20.
- Boxplots showed differences in semester grades among student categories.
- Heatmaps revealed strong correlations among semester grades and approved units.

II. Key Findings:

- Students with higher admission grades (above 140) have higher graduation rates.
- Semester grades significantly influence academic success.

Feature Selection and Dimensionality Reduction

SelectKBest Result:

Top predictors include semester grades, tuition fee status, and scholarship holding status.

PCA Results: Top 10 principal components explained over 99.9% of variance, ensuring efficient dimensionality reduction.

Model Building and Evaluation

Models Used:

- Decision Tree
- Random Forest
- Logistic Regression
- Support Vector Machine (SVM)

Best Performing Model:

Random Forest Classifier

- Accuracy: 83%
- F1 Score: 84%

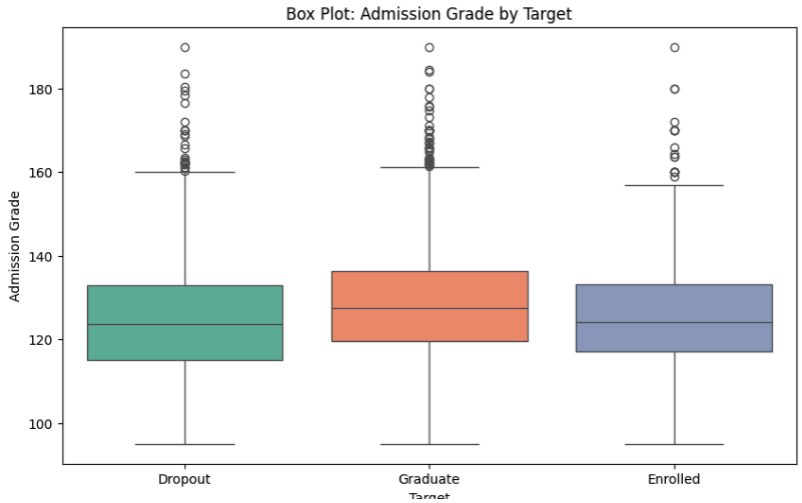
Cross-Validation:

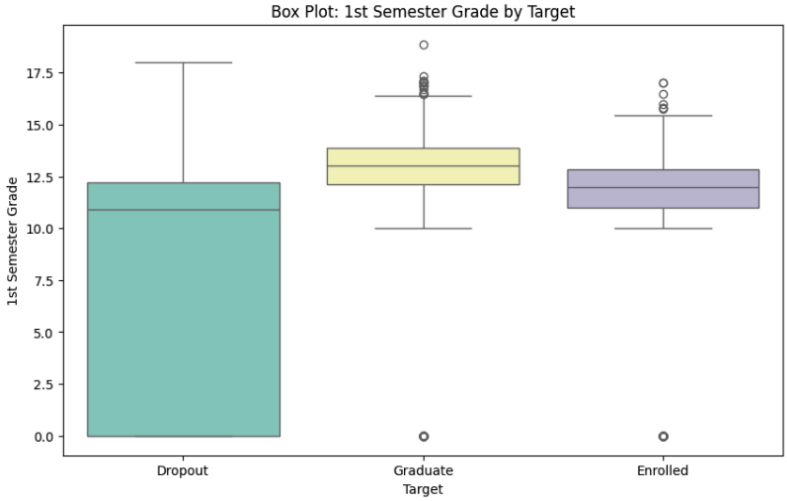
Ensured model reliability with a mean accuracy of 82.5%.

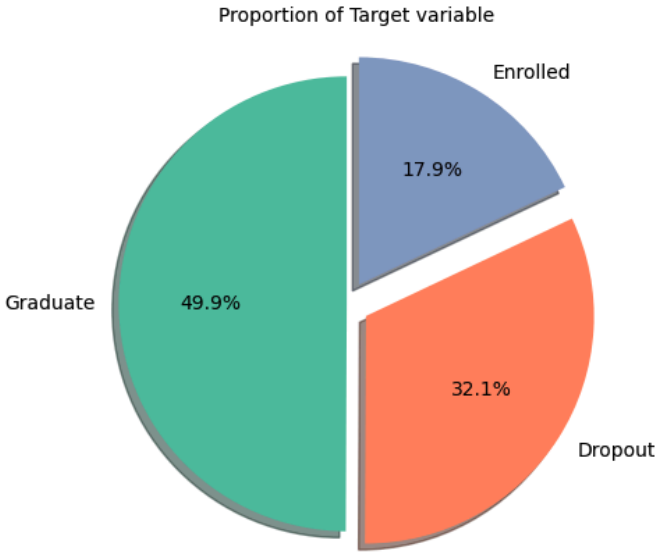
Confusion Matrix:

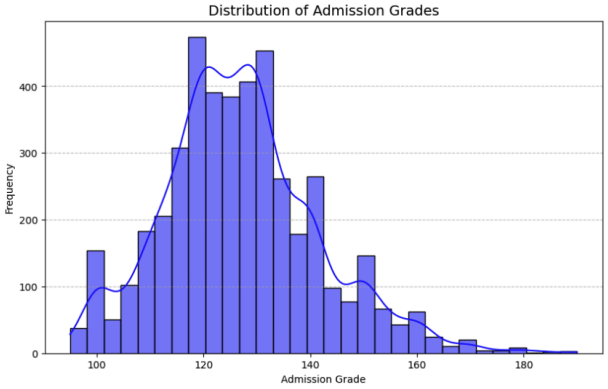
Visualized classification performance with balanced accuracy across classes.

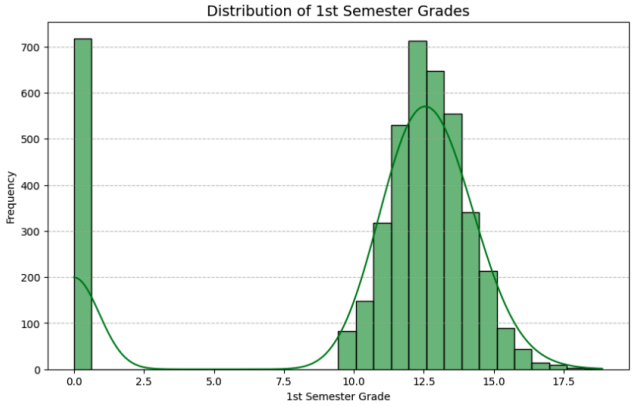
Results:

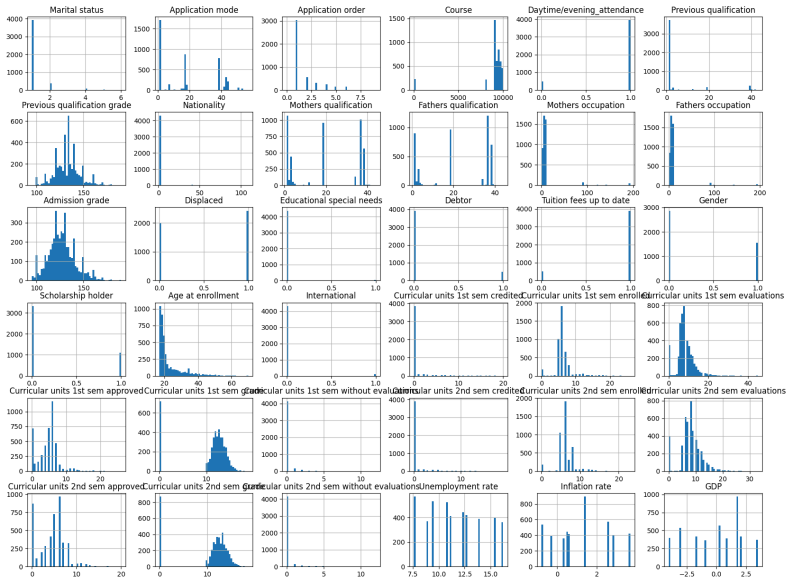

Box Plot: Admission Grade by Target
Admission Grade
180
160
140
120
100
Dropout
Graduate
Enrolled

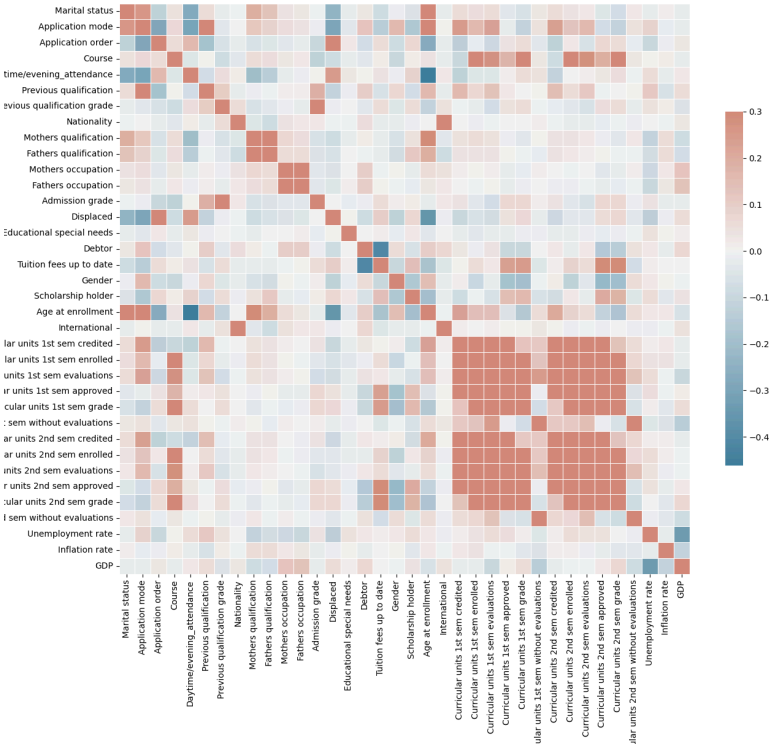

Box Plot: 1st Semester Grade by Target
1st Semester Grade
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

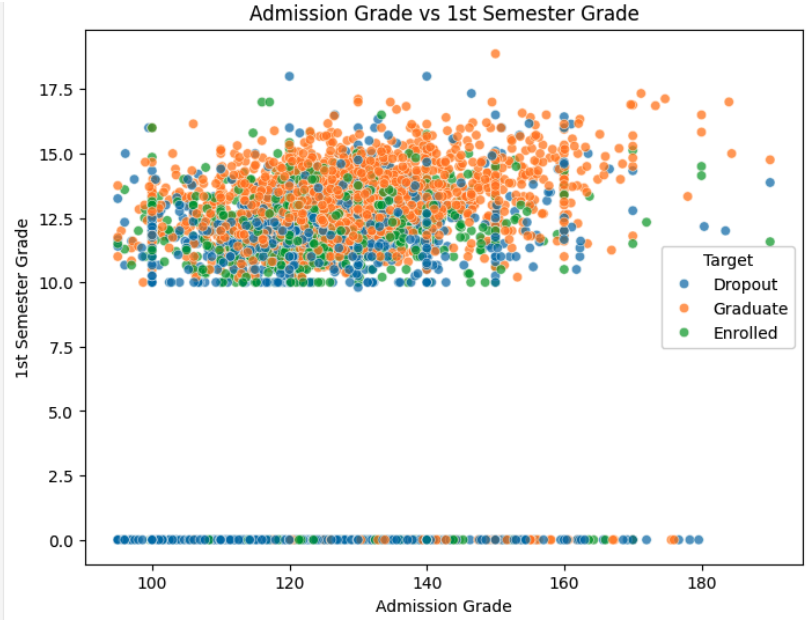

Proportion of Target variable
Enrolled
17.9%
Graduate
49.9%
32.1%
Dropout

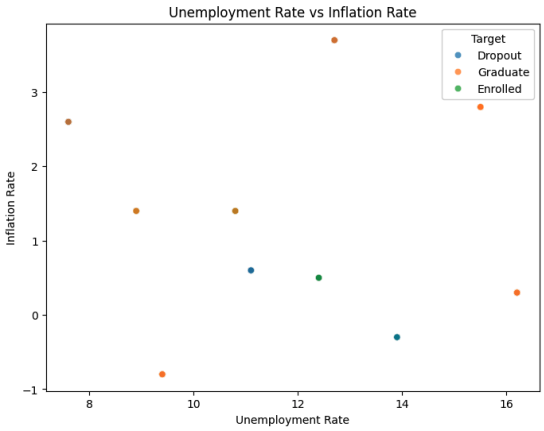

Distribution of Admission Grades
Frequency
Admission Grade

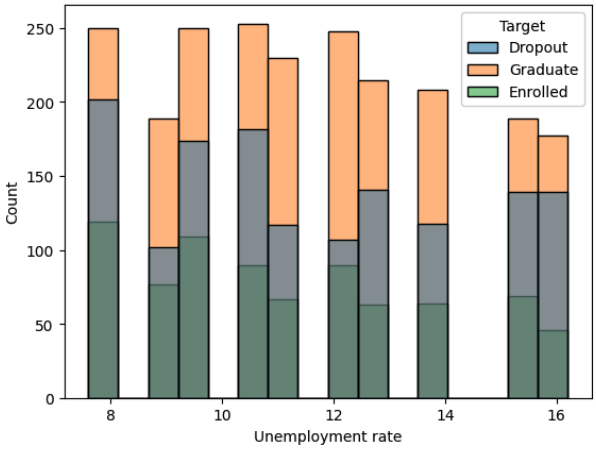

Distribution of 1st Semester Grades
Frequency
1st Semester Grade

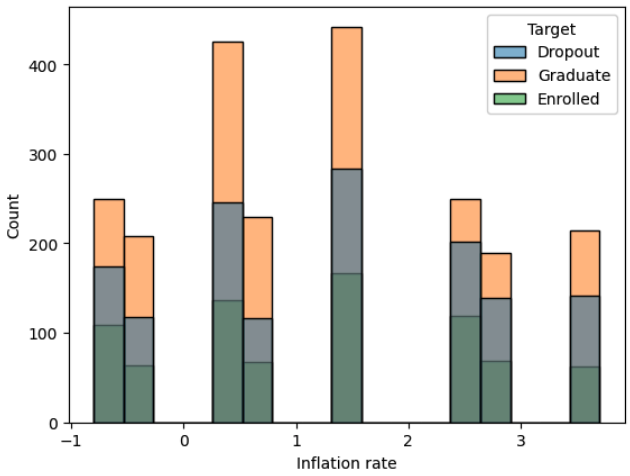

Marital status
Application mode
Application order
Course
Daytime/evening_attendance
Previous qualification
Previous qualification grade
Nationality
Mothers qualification
Fathers qualification
Mothers occupation
Fathers occupation
Admission grade
Displaced
Educational special needs
Debtor
Tuition fees 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

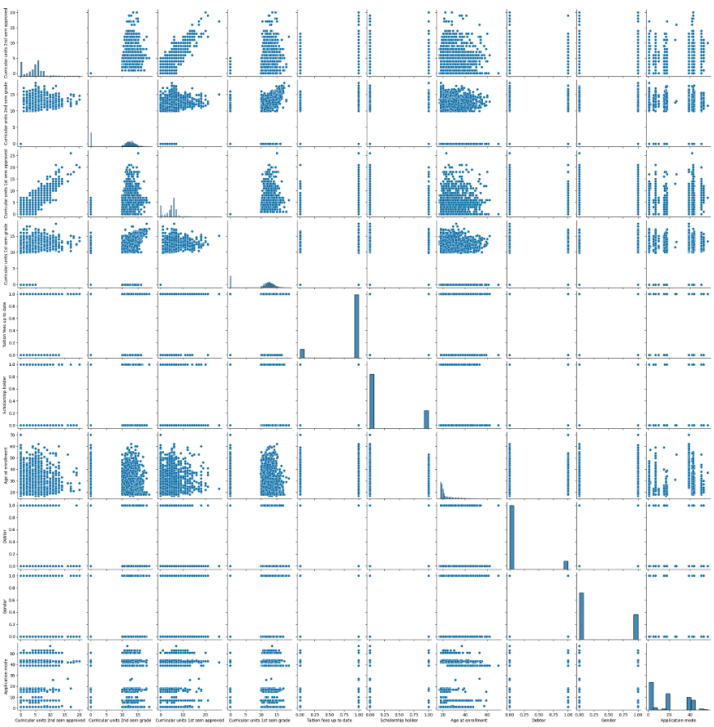

Marital status
Application mode
Application order
Course
time/evening_attendance
Previous qualification
evious qualification grade
Nationality
Mothers qualification
Fathers qualification
Mothers occupation
Fathers occupation
Admission grade
Displaced
Educational special needs
Debtor
Tuition fees up to date
Gender
Scholarship holder
Age at enrollment
International
lar units 1st sem credited
lar units 1st sem enrolled
units 1st sem evaluations
r units 1st sem approved
cular units 1st sem grade
: sem without evaluations
ar units 2nd sem credited
ar units 2nd sem enrolled
nits 2nd sem evaluations
r units 2nd sem approved
:ular units 2nd sem grade
l sem without evaluations
Unemployment rate
Inflation rate
GDP
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4

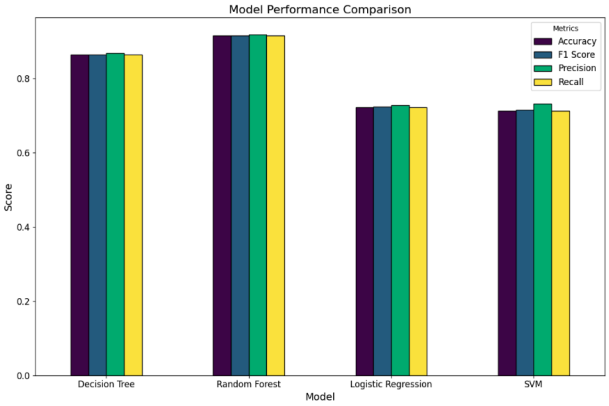

Admission Grade vs 1st Semester Grade
1st Semester Grade
Admission Grade
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
100
120
140
160
180
Target
Dropout
Graduate
Enrolled

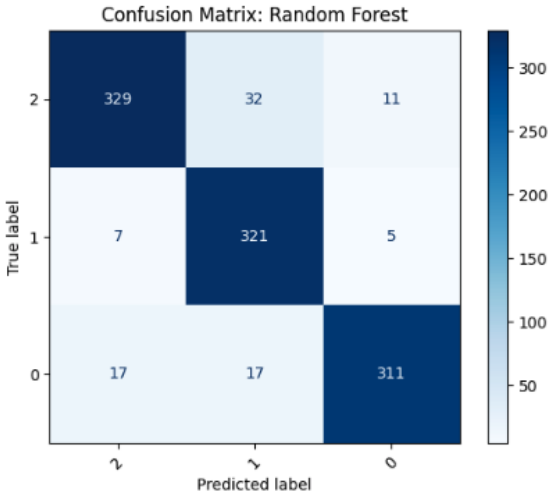

Unemployment Rate vs Inflation Rate
Target
Dropout
Graduate
Enrolled
Inflation Rate
Unemployment Rate

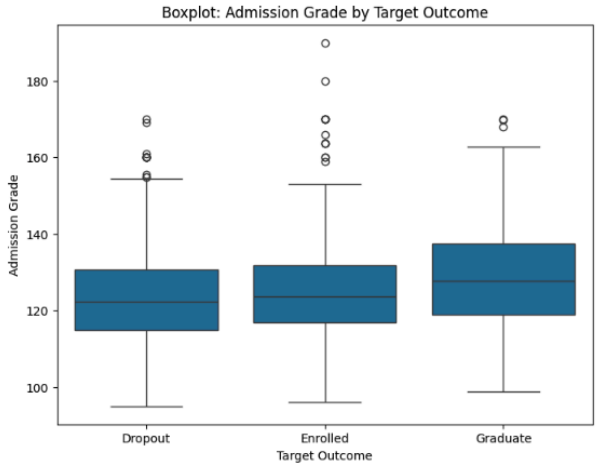

Target
Dropout
Graduate
Enrolled
Count
Unemployment rate

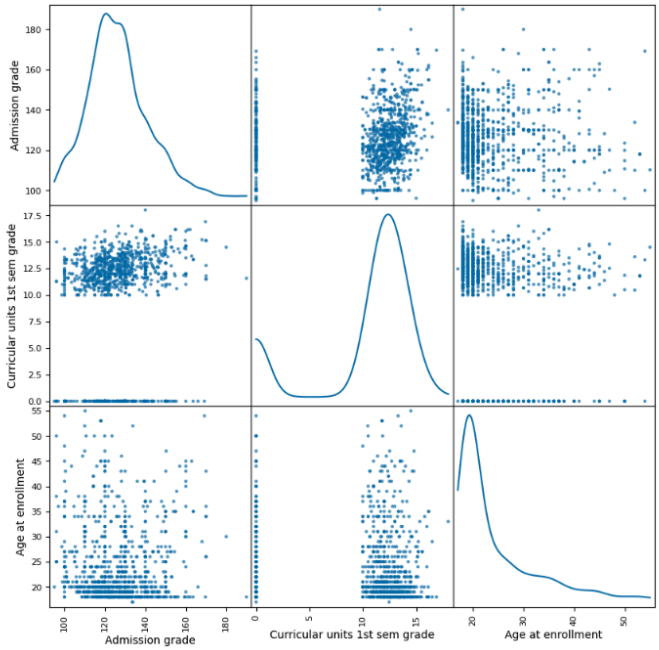

Target
Dropout
Graduate
Enrolled
Count
Inflation rate

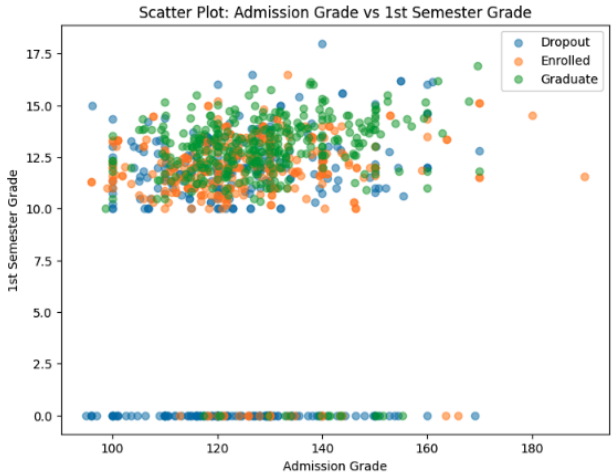


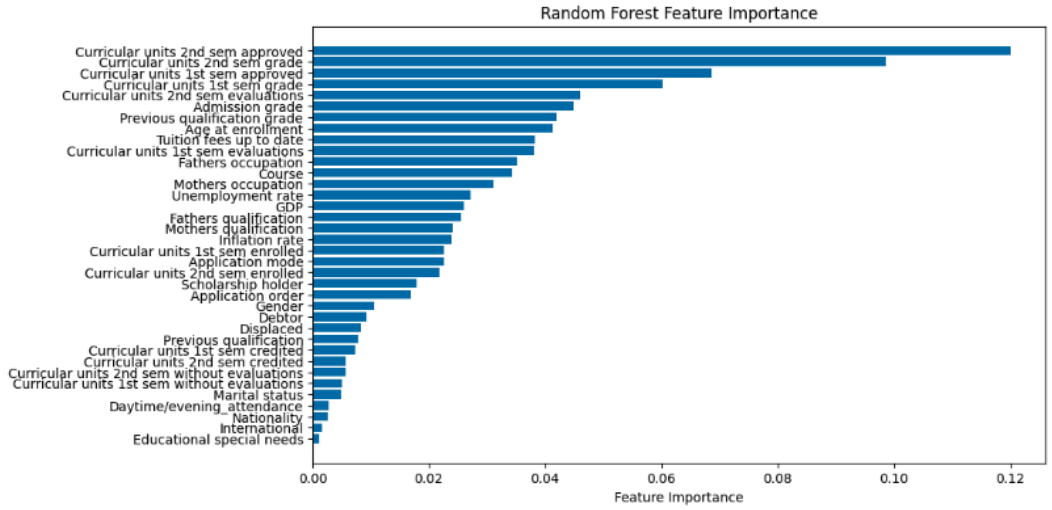

Model Performance Comparison
Metrics
Accuracy
F1 Score
Precision
Recall
Score
0.0
0.2
0.4
0.6
0.8
Decision Tree
Random Forest
Logistic Regression
SVM
Model


Confusion Matrix: Random Forest
True label
2
1
0
329
32
11
7
321
5
17
17
311
300
250
200
150
100
50
2
1
0
Predicted label


Boxplot: Admission Grade by Target Outcome
Admission Grade
100
120
140
160
180
Dropout
Enrolled
Graduate
Target Outcome


Admission grade
Curricular units 1st sem grade
Age at enrollment
Admission grade
Curricular units 1st sem grade
Age at enrollment





Conclusion:

The analysis identified key factors influencing student outcomes, with "Curricular units 2nd sem approved" (Score: 2110.25) and "1st sem grade" (Score: 973.41) being the most significant predictors. Random Forest achieved the highest performance, with an accuracy of 83% and F1 Score of 84%, effectively classifying student statuses. SMOTE oversampling balanced the dataset (2209 samples per class), improving model reliability. Visualizations highlighted that students with admission grades above 140 were more likely to graduate, while those below 120 had higher dropout rates. These findings emphasize the importance of academic and financial support to enhance graduation rates.

Video URL link:
https://drive.google.com/file/d/1d5nn4OAMdlBxWsfsbZdBDrbXGoYRBg1Y/view?usp=sharing