Student Dropout Prediction

Data Preprocessing Report

Description of Data Cleaning Steps:

1. Column Name Cleaning:

- Any tabs or whitespaces in column names were removed to standardize them.

2. Handling Missing Values:

- Missing values were handled by either dropping or filling them. The default method was dropping missing values, though the option to fill missing values with a specific value was available.

3. Outlier Detection and Treatment:

- The Interquartile Range (IQR) method was applied to detect outliers in numerical columns such as 'Admission grade,' 'Curricular units 1st sem (grade),' and 'Curricular units 2nd sem (grade).' These outliers were then capped at the 1st and 99th percentiles.

Summary of Data Quality Issues:

Missing Values: There were missing values in the dataset that were either dropped or filled depending on the selected method.

Outliers: Outliers in admission grades and curricular unit grades were identified and capped using the IQR method.

Justification for Chosen Data Transformation Methods:

Handling Missing Values: Dropping missing values was chosen to ensure data consistency for the analysis, though filling missing values can be justified when domain knowledge provides an appropriate fill strategy.

Outlier Capping: Capping outliers using the 1st and 99th percentiles ensured that extreme values did not overly influence the statistical models.

Scaling: Numerical features were scaled using the `StandardScaler`, which standardizes the data by removing the mean and scaling to unit variance. This is crucial when using machine learning algorithms sensitive to feature scaling.

Statistical Analysis Report

Descriptive Statistics for All Variables:

- Descriptive statistics for the numerical variables, such as the mean, standard deviation, minimum, and maximum values, were generated. The key statistics for the numerical columns are summarized:

  - Admission grade: Provided detailed summary statistics.

  - Curricular units (1st and 2nd sem) grades: Descriptive analysis showed the distribution of student performance across these metrics.

Correlation Matrix Heatmap:

- A correlation matrix heatmap was generated for all numerical variables. Strong correlations, if present, were highlighted to help assess multicollinearity or associations between variables.

Results and Interpretation of Hypothesis Tests:

T-Test for Admission Grade (Dropout vs Graduate): A t-test was conducted to compare the admission grades of students who dropped out and those who graduated. The test result showed a significant difference:

- T-statistic: -7.858

- P-value: \(5.10 \times 10^{-15}\)

Interpretation: Since the p-value is very small, we reject the null hypothesis, indicating that the admission grades of students who dropped out significantly differ from those who graduated.

Chi-square Test for Gender vs Dropout: A chi-square test was performed to examine the relationship between gender and dropout status. The result was:

- Chi-square statistic: 233.27
- P-value: \(2.22 \times 10^{-51}\)

Interpretation: The very small p-value suggests a significant association between gender and dropout status, meaning gender has an influence on whether a student is likely to drop out.

Data loaded successfully!

Show raw data

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality
0	1	17	5	171	1	1	122	1
1	1	15	1	9,254	1	1	160	1
2	1	1	5	9,070	1	1	122	1
3	1	17	2	9,773	1	1	122	1
4	2	39	1	8,014	0	1	100	1

Show initial data exploration

Dataset Shape:

Rows: 4424, Columns: 37

Data Types:

	0
Curricular units 2nd sem (credi	int64
Curricular units 2nd sem (enrol	int64
Curricular units 2nd sem (evalu	int64
Curricular units 2nd sem (appr	int64
Curricular units 2nd sem (grade	float64
Curricular units 2nd sem (withc	int64
Unemployment rate	float64
Inflation rate	float64
GDP	float64
Target	object

Summary Statistics:

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacion
count	4,424	4,424	4,424	4,424	4,424	4,424	4,424	4
mean	1.1786	18.6691	1.7278	8,856.6426	0.8908	4.5778	132.6133	1.8
std	0.6057	17.4847	1.3138	2,063.5664	0.3119	10.2166	13.1883	6.9
min	1	1	0	33	0	1	95	
25%	1	1	1	9,085	1	1	125	
50%	1	17	1	9,238	1	1	133.1	
75%	1	39	2	9,556	1	1	140	
max	6	57	9	9,991	1	43	190	

☑ Check for missing values

Missing Values per Column:

	0
empty	

☑ Show outliers

Outliers in Admission grade:

	0
21	776
22	1,002
23	1,037
24	1,073
25	1,126
26	1,177
27	1,215
28	1,254
29	1,283
30	1,503

Outliers in Curricular units 1st sem (grade):

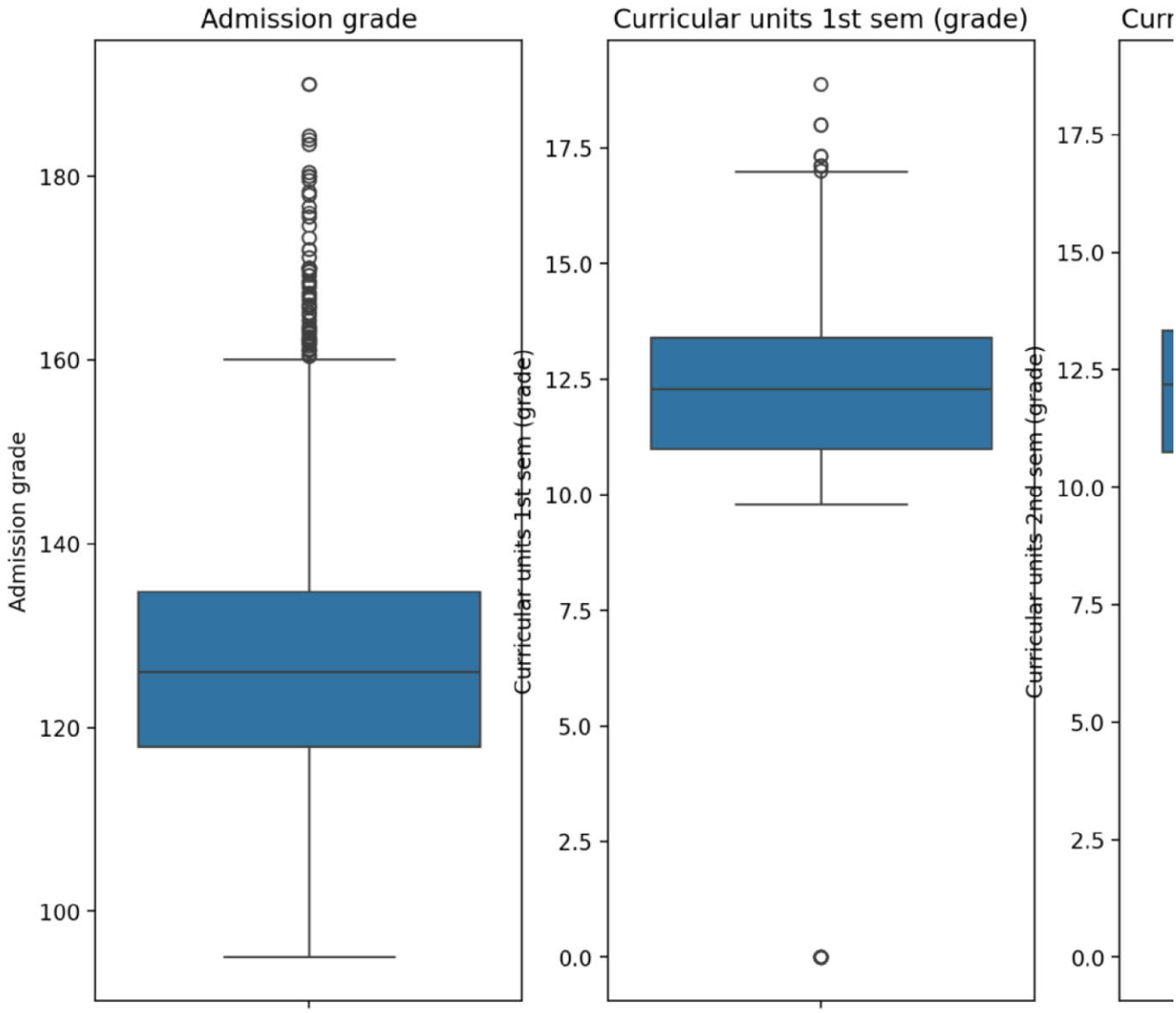
	0
0	0
1	2
2	7
3	12
4	20
5	35
6	36
7	44
8	56
9	59

Outliers in Curricular units 2nd sem (grade):

	0
0	0
1	2
2	7
3	12
4	15
5	20
6	36
7	40
8	44
9	56

☑ Show box plots for numerical

columns Box plots for numerical columns:



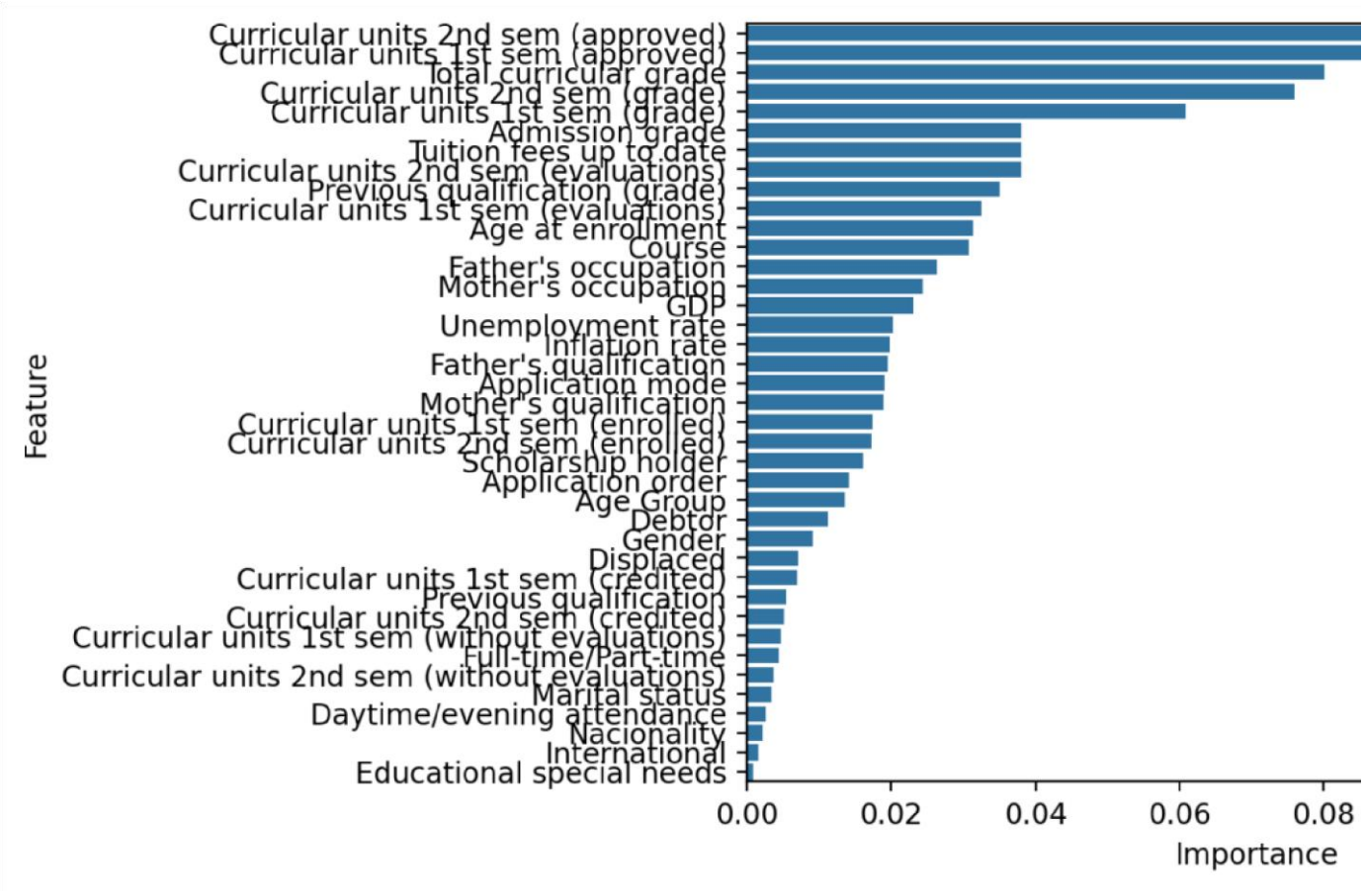
Show preprocessed data

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality
0	1	17	5	171	1	1	122	1
1	1	15	1	9,254	1	1	160	1
2	1	1	5	9,070	1	1	122	1
3	1	17	2	9,773	1	1	122	1
4	2	39	1	8,014	0	1	100	1

Show feature importance

Feature Importance using RandomForest:

	Feature	Importance
30	Curricular units 2nd sem (approved)	0.1346
24	Curricular units 1st sem (approved)	0.0863
37	Total curricular grade	0.0801
31	Curricular units 2nd sem (grade)	0.076
25	Curricular units 1st sem (grade)	0.0608
12	Admission grade	0.0381
16	Tuition fees up to date	0.038
29	Curricular units 2nd sem (evaluations)	0.038
6	Previous qualification (grade)	0.0351
23	Curricular units 1st sem (evaluations)	0.0325

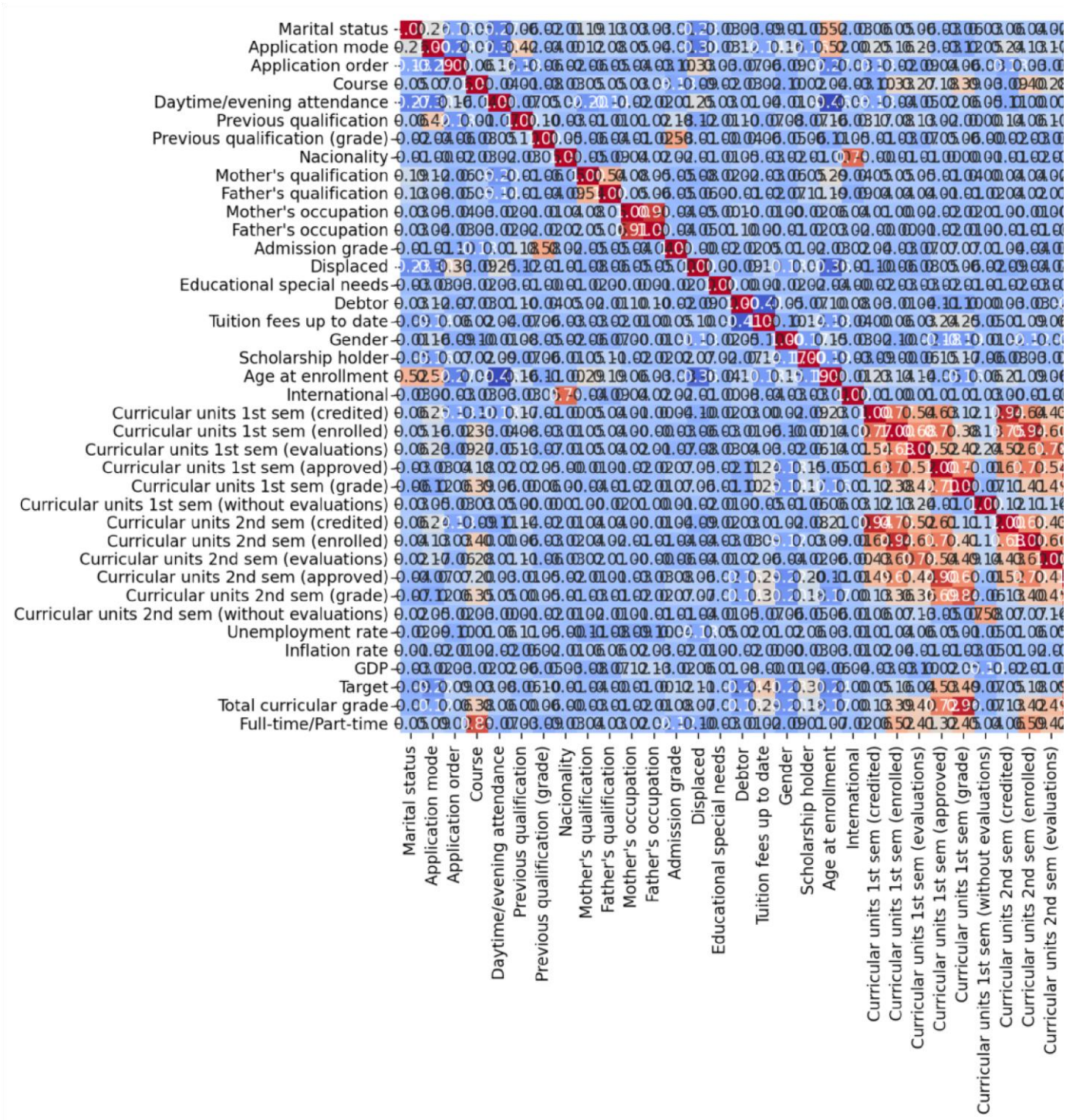


☑ Show descriptive statistics

Numerical Descriptive Statistics:

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacion
count	4,424	4,424	4,424	4,424	4,424	4,424	4,424	4
mean	1.1786	18.6691	1.7278	8,856.6426	0.8908	4.5778	132.6133	1.8
std	0.6057	17.4847	1.3138	2,063.5664	0.3119	10.2166	13.1883	6.9
min	1	1	0	33	0	1	95	
25%	1	1	1	9,085	1	1	125	
50%	1	17	1	9,238	1	1	133.1	
75%	1	39	2	9,556	1	1	140	
max	6	57	9	9,991	1	43	190	

☑ Show correlation matrix heatmap



☑ Show t-test results for Admission Grade (Dropout vs Graduate)

T-test results: T-statistic: 0.834913824313304, P-value: 0.4038563525431528

☑ Show chi-square test results for Gender vs Dropout

Chi-square test results for Gender vs Dropout: Chi-square stat: 233.26643249623856, P-value: 2.2224795668092454e-51

Enter details for prediction:

Admission Grade

Insights, Recommendations, and Conclusions

Insights

- The dataset contains 4,424 rows and 37 columns.
- Outliers were detected in key numerical variables such as Admission grade and semester grades.

The T-test showed no significant difference between dropouts and graduates in Admission grades.

Gender shows a significant impact on dropout rates according to the Chi-square test.

Recommendations

- Investigate missing values and handle them appropriately to improve model performance.
- Consider outlier treatment techniques like transformations to avoid skewing model predictions.

Enhance feature engineering and explore additional models (e.g., XGBoost) for better accuracy.

Address dataset imbalance (if present) to improve dropout predictions, especially related to gender.

Conclusions

- Admission grades alone may not be a strong predictor of dropouts.
- Gender has a notable influence on dropout rates, indicating a need for gender-targeted interventions.

The current model is a good baseline, but improvements can be made with additional data processing and model tuning.