“Predicting Student Dropout & Success: A Data-Driven Approach to Academic Retention”

by Kalyani Wasave

Agenda

- Motivation Overview
- Data EDA
- Data Processing
- Model building and accuracy
- Conclusion

Motivation: Why this

Understanding Student's Success & Dropout Rates:

Learn about the various factors contributing to student success and dropout rates to improve educational outcomes

Practical Skill Development:

Learn data analysis strategies by finding different patterns in the dataset Learn data cleaning, training, modeling and predicting the outcome

Real-world

Implement data analysis to identify, predict and solve academic problems

Improve various factors affecting the student's academic performance and decisions.

Dataset Overview

Data Source

The dataset was created in a project that aims to contribute to the reduction of academic dropout and failure in higher education. This dataset is supported by program SATDAP

Data Attribute

The data refers to records of students enrolled between academic years 2008/09–2018/2019 and from different undergraduate degrees

Data Structure

The dataset is in a tabular format, with each row representing a student and the columns containing the various attributes collected about them.

Data Quality

The data has been already went through rigorous data preprocessing to handle data from anomalies, unexplainable outliers, and missing values.

Data Columns

Financial data

- Scholarship holder
- Tuition fees upto date Debtor
- Parent's occupation
- Inflation rate
- GDP
- Unemployment rate

Demographics data

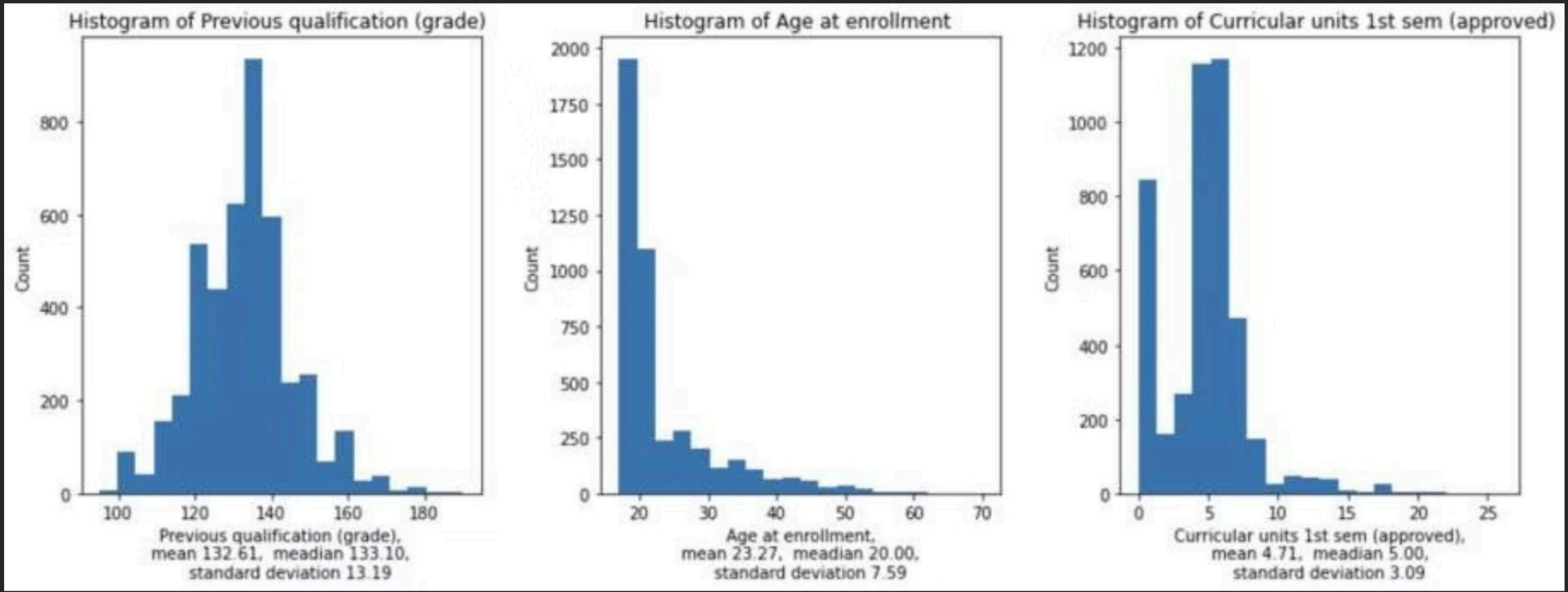
- Age at enrollment
- Gender
- Marital status
- International
- Daytime/Evening
- Nationality
- Parent's qualification
- Educational Special needs

Academic data

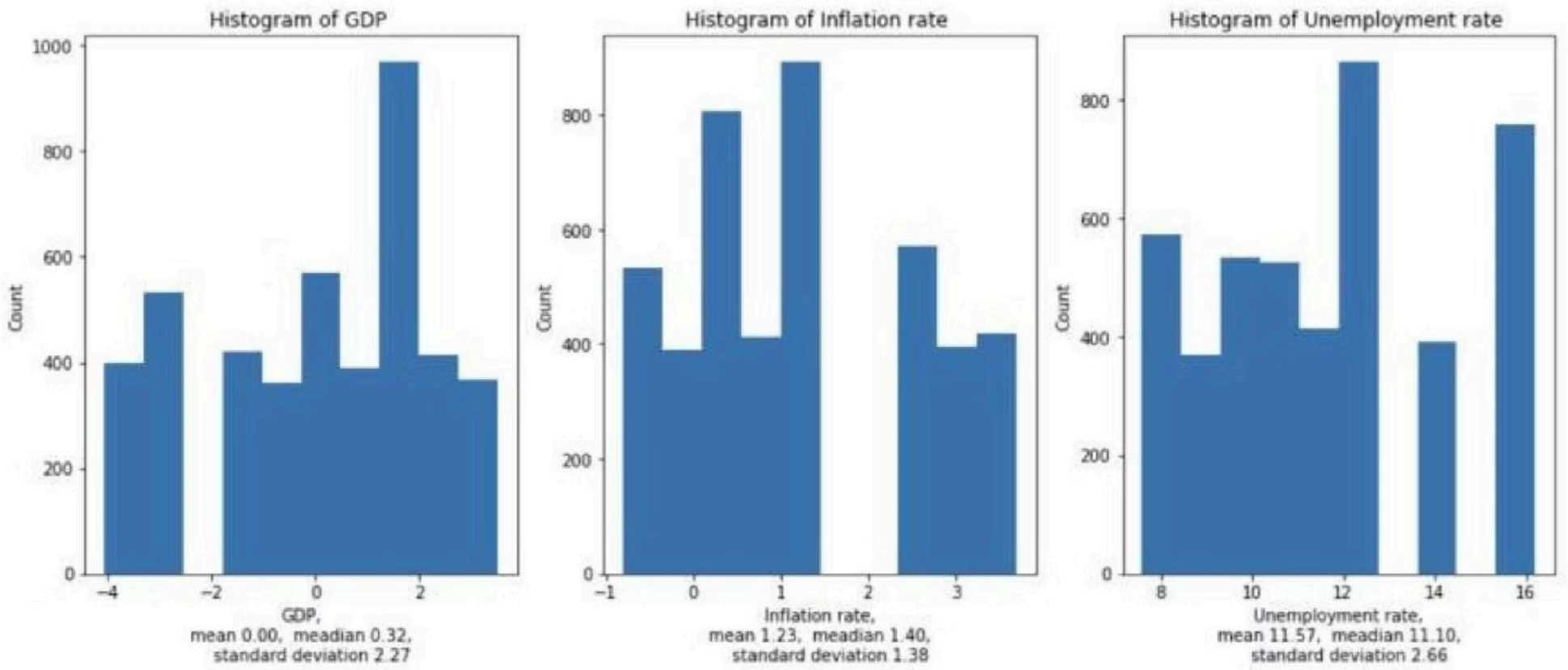
- Application mode Application
- Order Previous qualification
- Admission grade Curricular units
- 1st & 2nd sem Evaluations
- (credited | enrolled |
- approved | grade)

Target Variables: Dropout, Enrolled & Graduate

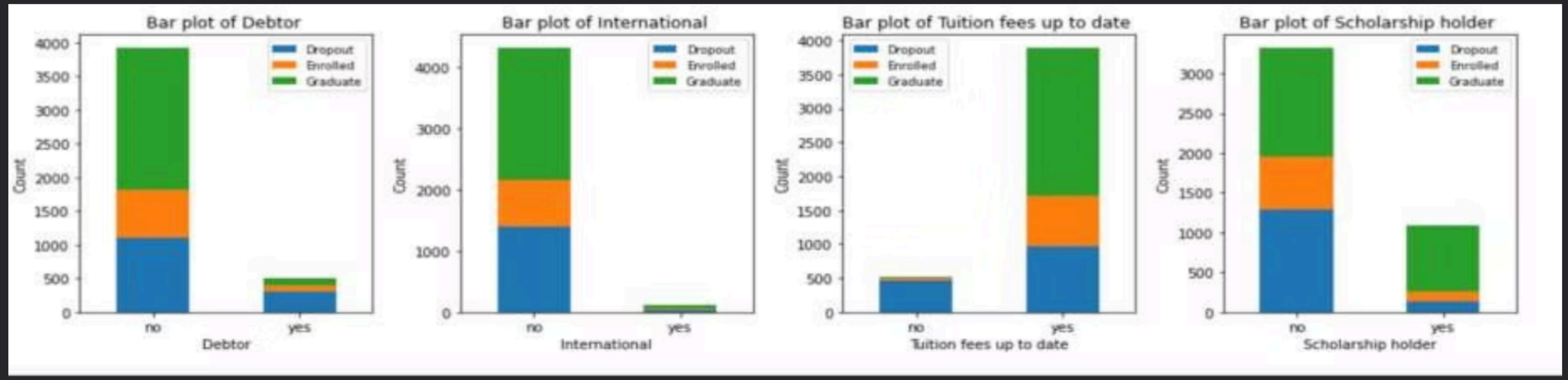
Histograms of Numeric

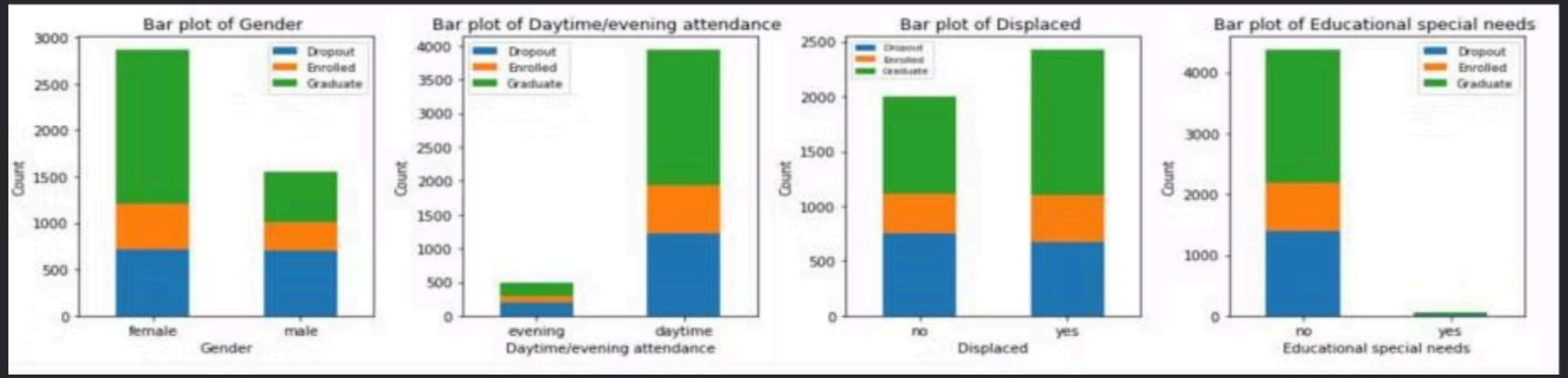


Histograms of Numeric variables

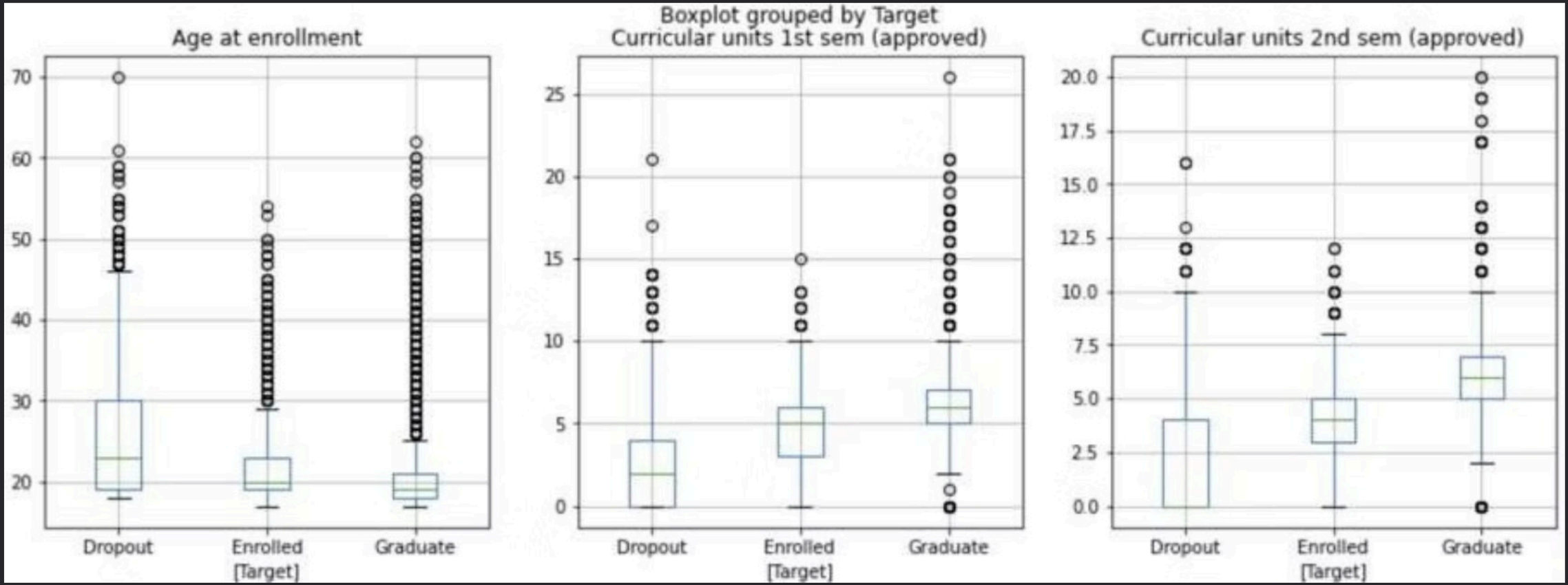


Bar Plot of Categorical

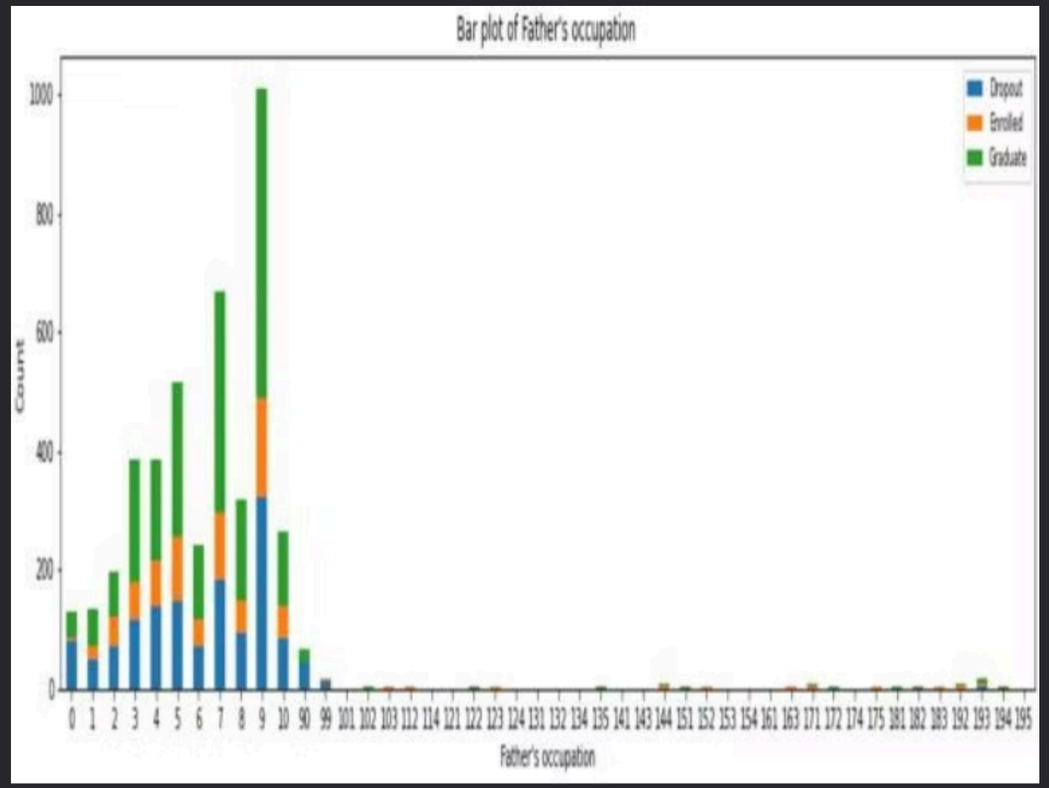




Box plot of numerical



Data Processing - High cardinality of nominal variables



Data Issue

- High Cardinality in Nominal Variables
- Uneven Data Distribution
- Diminished Significance of Variables

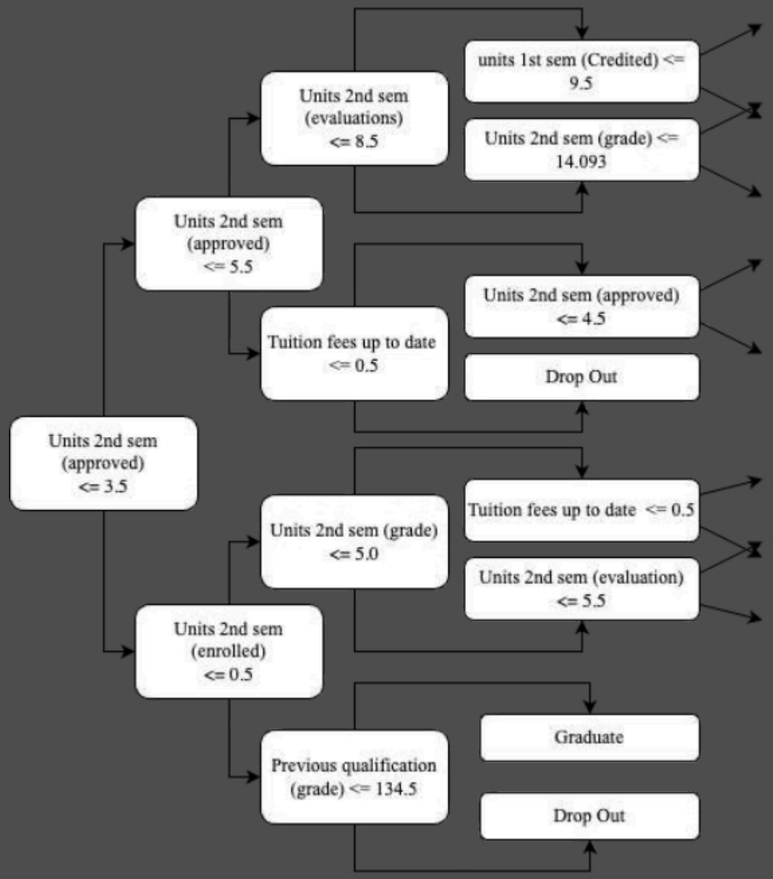
Proposed Solution:

- Examine the frequency of each category in categorical variables
- Retain categories that encompass at least 90% of the data
- Focus on categories with high frequency

Data Processing - Numerical

- Model:
  - Decision Tree and Random Forest.
- Robustness to Data Properties:
  - Any data Distribution
  - Handling Missing Values
  - Normalization and Scaling
- Data Processing:
  - Removing low frequency data from high cardinal categorical variables.
  - No Other Data Processing Done for Model Building

Decision Tree



Decision Tree with dept -5 (Figure on shows first 3 layers)

Simple and interpretable machine learning model, structured like a tree, handles complex data; prone to overfitting, mitigated with parameter tuning or ensemble methods.

Observations:

- Curricular units 2nd sem(approved) holds high importance in student outcomes
- Other Important variables: - Tuition fee upto date - Previous Qualification
- Strong academic performance significantly increases the likelihood of graduation.
- Financially stable students with normal and poor academic performance tend to opt for enrollment in alternative courses.
- Students with poor academic performance and financial instability have high probability of dropping out from course.

Decision Tree Performance

Training Data

Precision	Recall	F1 Score	Observation
0.87	0.67	0.76	935
0.52	0.31	0.39	525
0.75	0.96	0.84	1504

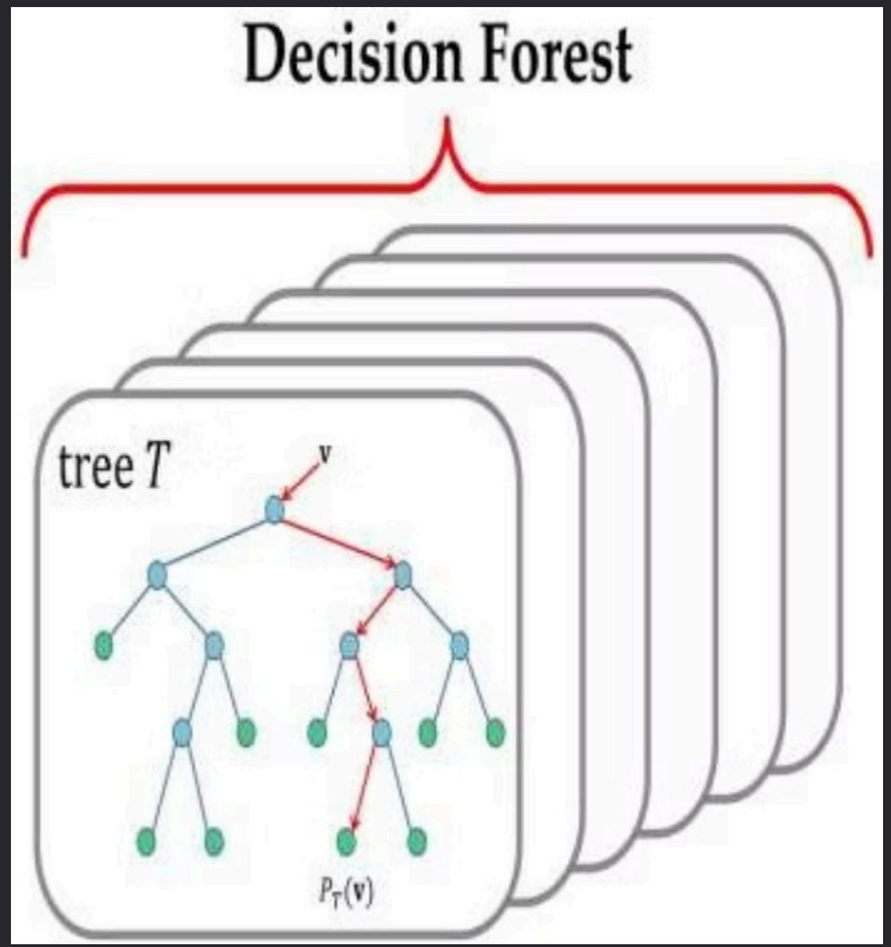
Test Data

Precision	Recall	F1 Score	Observation
0.88	0.64	0.74	486
0.49	0.29	0.37	269
0.72	0.96	0.82	707

Observations

- F1 scores for training & testing is similar. Model is performing good with Dropout and Graduate classifications and struggling to predict enrolled students.
- For Graduate category, Model seems to have high Recall but less Precision, it means some of the enrolled student might be getting predicted as Graduated students.
- For Dropout category, Recall is less and Precision is high, it means some of the Drop out might be getting predicted as Enrolled students. Because of 3rd and 4th points, Enrolled has both low recall and low precisions rate.

Random Forest



- In Decision Tree, whenever we were increasing the complexity of tree, Test accuracy was dropping, limiting our test F1 score to 0.71.
- To handle it, we decided to train the model in random forest, as it is an ensemble of multiple tree even if we increase the complexity of tree it will have minimal impact on overfitting and can give chance to increase the testing accuracy.
- Hyperparameter of random forest:
  - Number of trees -500
  - max sample -0.5

Random Forest Performance Metrics

Training Data Accuracy

Metrics	Precision	Recall	F1 Score	Observations
Drop Out	0.98	0.96	0.97	935
Enrolled	1.0	0.90	0.95	525
Graduate	0.95	1.0	0.98	1504
Accuracy	0.98	0.95	0.97	2964

Test Data Accuracy

Metrics	Precision	Recall	F1 Score	Observations
Drop Out	0.81	0.77	0.79	486
Enrolled	0.56	0.28	0.37	269
Graduate	0.76	0.94	0.84	707

Observations

- Test accuracy improved to 0.74 compared to decision tree.
- Train accuracy is high compared to Test accuracy, there are still some chance of improvement in Test accuracy.

Conclusion

- Financial factors and academic factors are playing a crucial role in decision of student being graduated or dropped out.
- Predicting enrolled student is hard with current data
- Decision tree and Random forest are performing good to classify dropped out and graduate students.
- More complex algorithms such as Support Vector Machine or Neural Network can be used to improve the performance.
- In conclusion, identifying the student who are struggling with academics and having financial burden at early stage of college and guiding them can help us to reduce the drop out students.