SC1015 MINI PROJECT

Predicting how well students can cope in university.

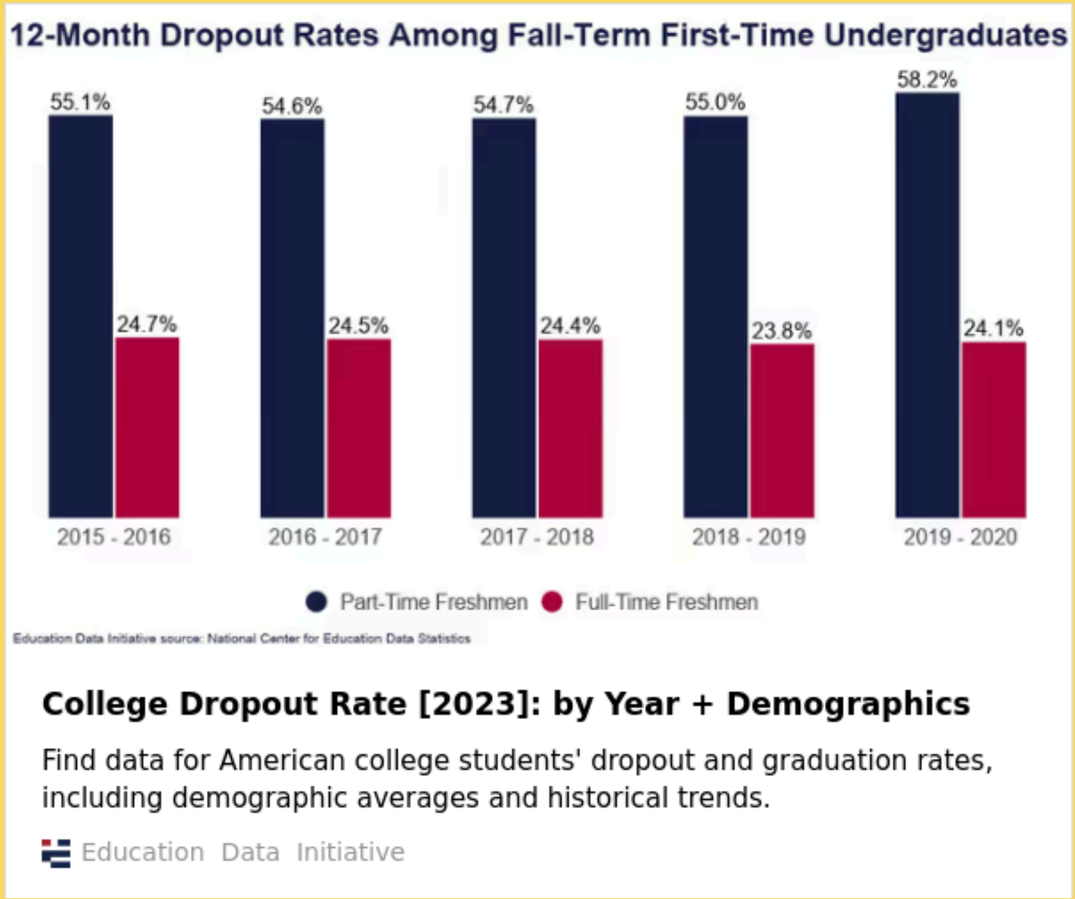
FDAA Group 8

Yeo Ruizhi Carwyn
Kent Karsten Pangestu
James Ang Wen Jie

Table of Contents

Problem

Problem



Students cannot cope in University

College dropout rates indicate that up to 32.9% of undergraduates do not complete their degree program.

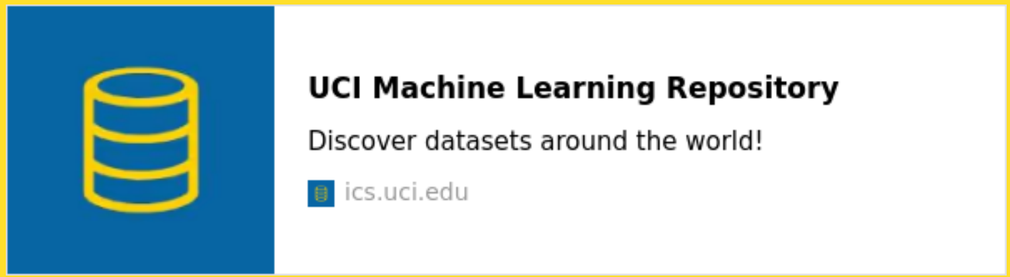
- First-time undergraduate first-year students have a 12-month dropout rate of 24.1%.
- Among first-time bachelor’s degree seekers, 25.7% ultimately drop out; among all undergraduate students, up to 40% drop out.

Problem Statement

Dropout Prediction: Can we classify students into those likely to drop out vs. those who will continue their education?

Academic Success Prediction: Can we predict the GPA, graduation likelihood, or other academic success indicators at the end of their course or semester?

Dataset



“ Predict Students' Dropout and Academic Success”,
Based on a paper by By Mónica V. Martins, Daniel Tolledo, Jorge Machado, Luís M. T. Baptista, and Valentim Realinho. 2021.

DATA
CLEANING

VARIABLES REMOVED

'MARITAL STATUS',
'APPLICATION MODE',
'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)',
'UNEMPLOYMENT RATE',
'INFLATION RATE',
'GDP',
'DISPLACED',
'EDUCATIONAL SPECIAL NEEDS',
'DAYTIME/EVENING ATTENDANCE '

VARIABLES REMOVED

```
# Variables we removed because we think they are irrelevant to our problem:
Columns_to_remove = ['Marital status', 'Application mode',
'Curricular units 1st sem (credited)', 'Curricular units 1st sem (enrolled)', 'Curricul
'Curricular units 1st sem (without evaluations)','Curricular units 2nd sem (credited)',
'Curricular units 2nd sem (evaluations)','Curricular units 2nd sem (without evaluations
'GDP', 'Displaced', 'Educational special needs', 'Daytime/evening attendance    ']

data.drop(Columns_to_remove, axis=1, inplace=True)

# Streamline the 'Nationality' variable to only local students (i.e. Nationality == 1)
data = data.loc[data['Nacionality'] == 1].copy()
data.drop(columns=['Nacionality', 'International'], axis=1, inplace=True)
```

VARIABLES RECATEGORISED

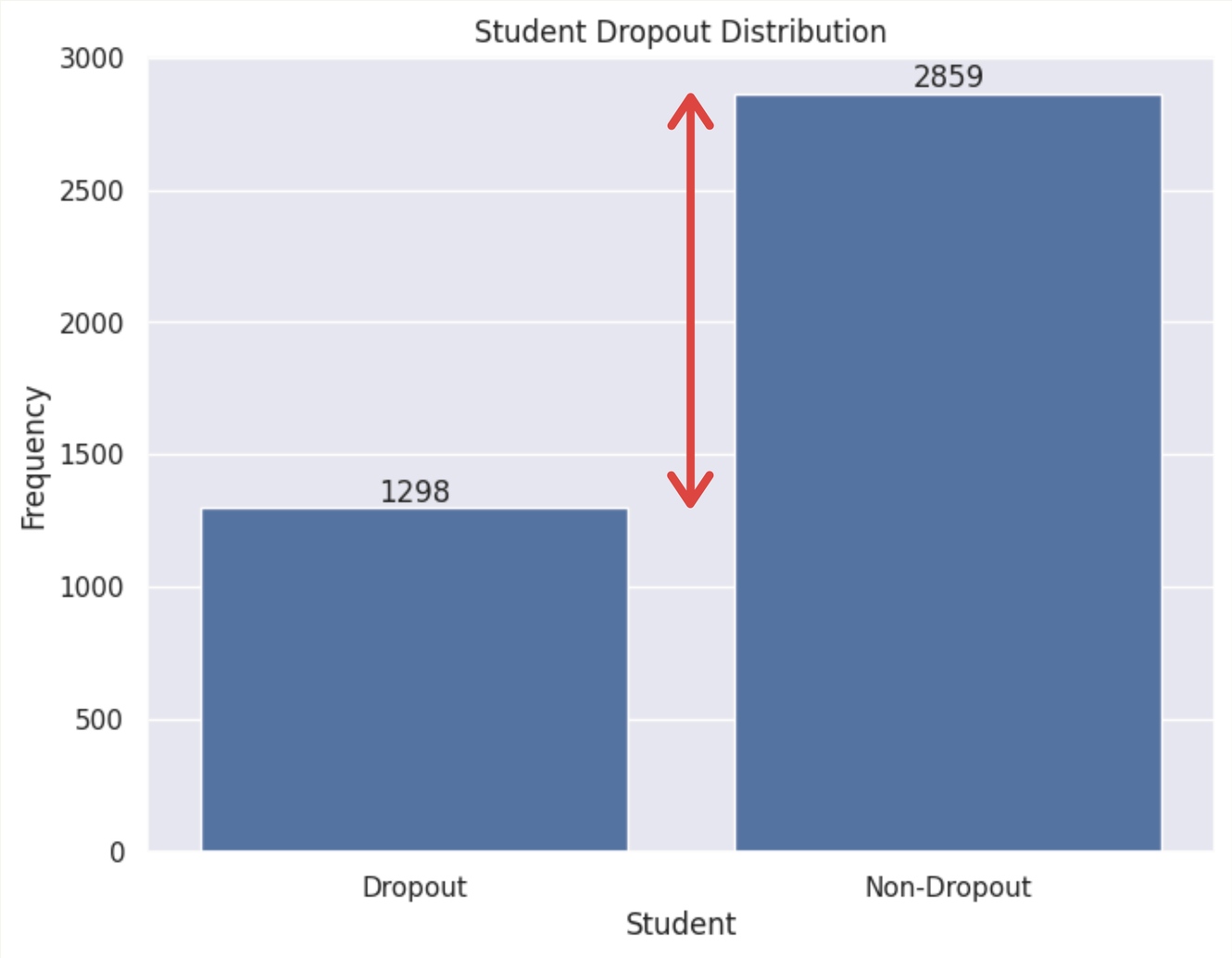
'TARGET'
'COURSE'
"MOTHER'S OCCUPATION"
"FATHER'S OCCUPATION"
"MOTHER'S QUALIFICATION"
"FATHER'S QUALIFICATION"
"PREVIOUS QUALIFICATION"

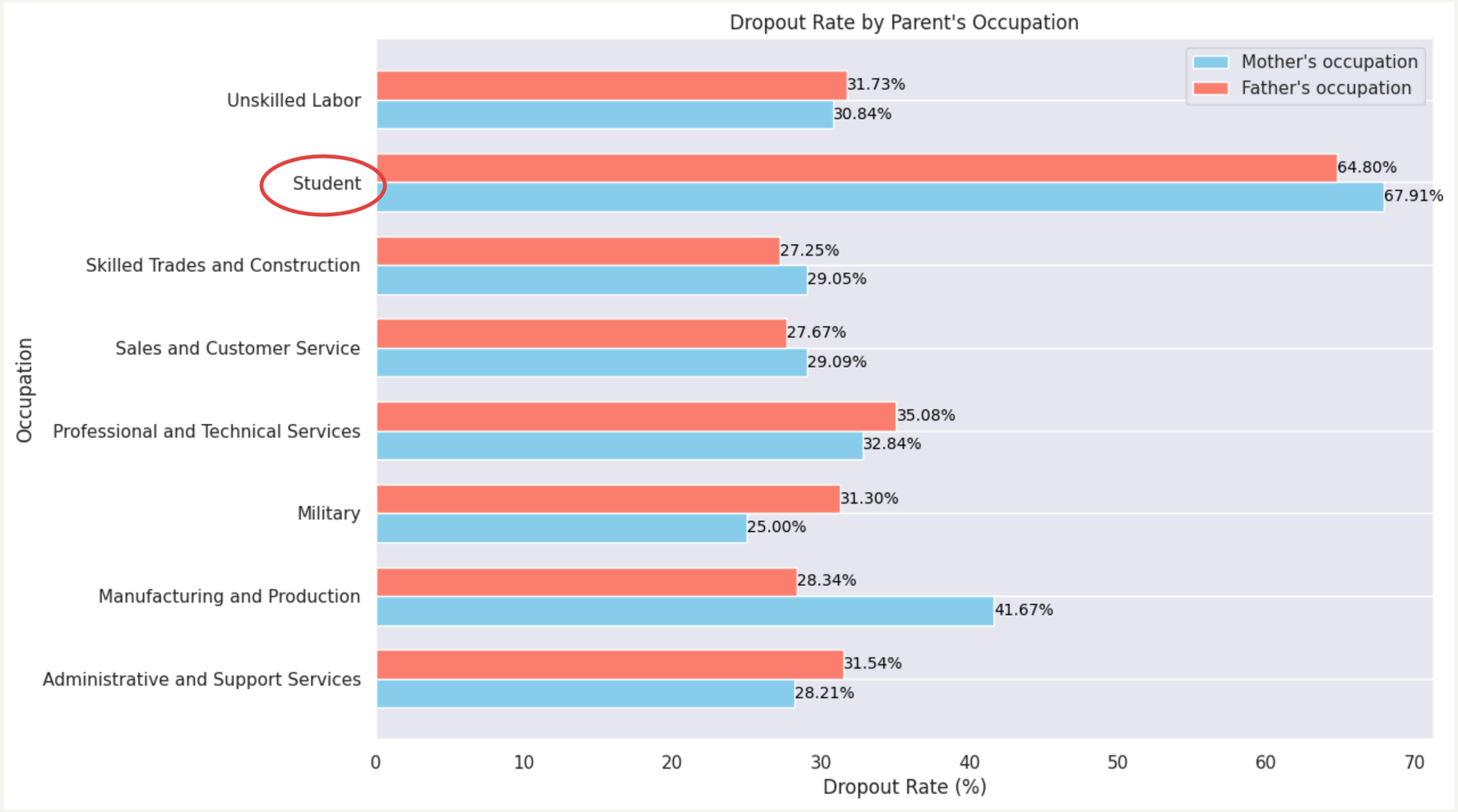
```python
# We categorised the level of educations from 0 to 5, where:
def clean_Education(code):
    if code in [34, 35, 36]:
        return 0 #'Uneducated'
    elif code in [37, 38]:
        return 1 #'Primary Education'
    elif code in [1, 10, 11, 12, 14, 18, 19, 26, 27, 29, 30, 39]:
        return 2 #'Secondary/High School'
    elif code in [22]:
        return 3 #'Tertiary Education'
    elif code in [2, 40]:
        return 4 #'Bachelor's'
    elif code in [3, 4, 5, 41, 42, 43, 44]:
        return 5 #'Above Bachelor's'
    else:
        return -1 #'Others'

def clean_previous_qualification(code):
    if code in [1, 9, 10, 12, 14, 15, 19, 38]:
        return 2 #'Secondary/High School'
    elif code in [39]:
        return 3 #'Tertiary Education'
    elif code in [2, 40]:
        return 4 #'Bachelor's'
    elif code in [3, 4, 5, 6, 42, 43]:
        return 5 #'Above Bachelor's'
    else:
        return -1 #'Others'
```

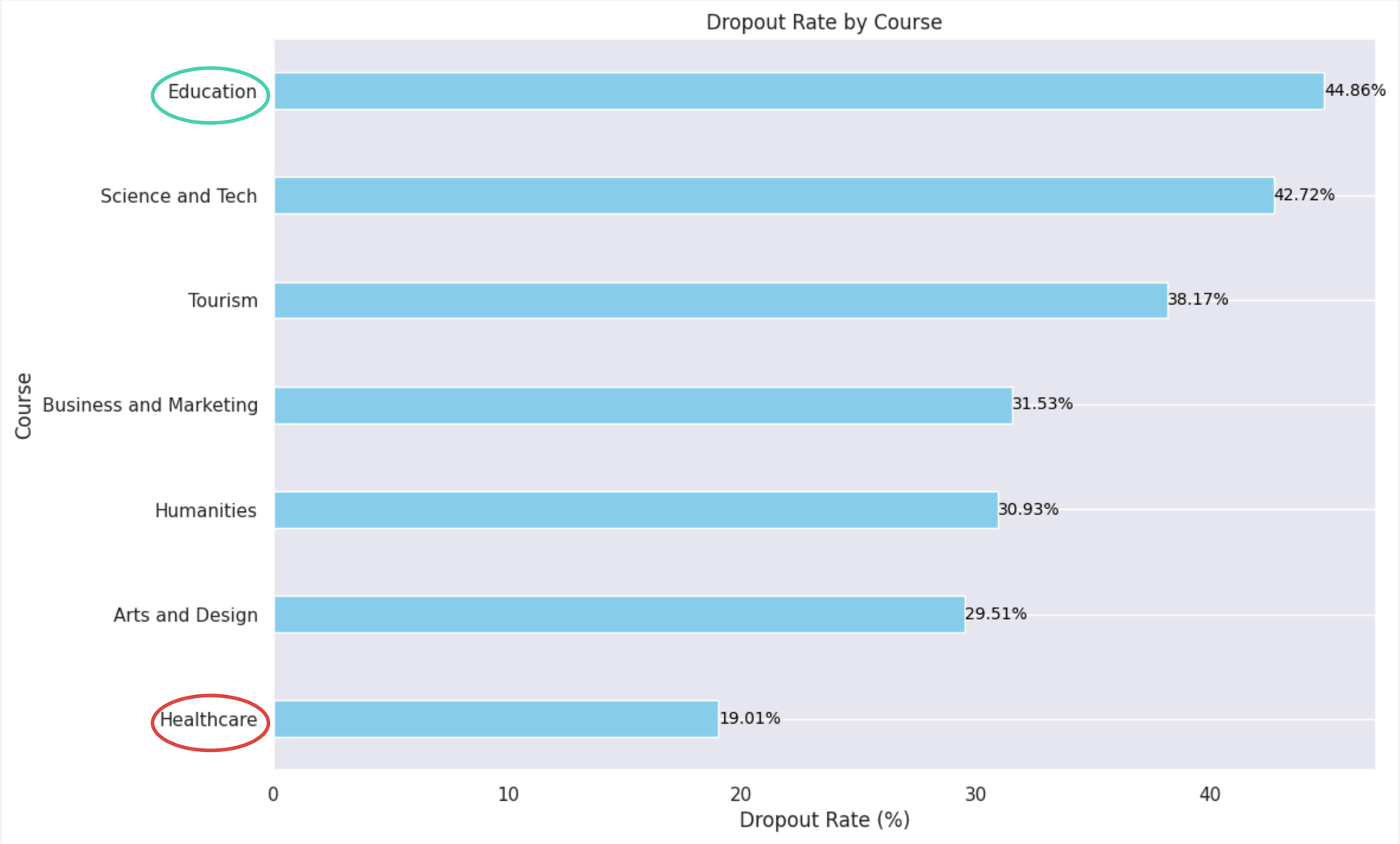
```
<class 'pandas.core.frame.DataFrame'>
Index: 4157 entries, 0 to 4422
Data columns (total 19 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   Application order                       4157 non-null   int64  
 1   Course                                  4157 non-null   object 
 2   Previous qualification                  4157 non-null   int64  
 3   Previous qualification (grade)          4157 non-null   float64
 4   Mother's qualification                  4157 non-null   int64  
 5   Father's qualification                  4157 non-null   int64  
 6   Mother's occupation                     4157 non-null   object 
 7   Father's occupation                     4157 non-null   object 
 8   Admission grade                         4157 non-null   float64
 9   Debtor                                  4157 non-null   int64  
 10  Tuition fees up to date                 4157 non-null   int64  
 11  Gender                                  4157 non-null   int64  
 12  Scholarship holder                      4157 non-null   int64  
 13  Age at enrollment                       4157 non-null   int64  
 14  Curricular units 1st sem (approved)     4157 non-null   int64  
 15  Curricular units 1st sem (grade)        4157 non-null   float64
 16  Curricular units 2nd sem (approved)     4157 non-null   int64  
 17  Curricular units 2nd sem (grade)        4157 non-null   float64
 18  Target                                  4157 non-null   object 
dtypes: float64(4), int64(11), object(4)
memory usage: 649.5+ KB
```

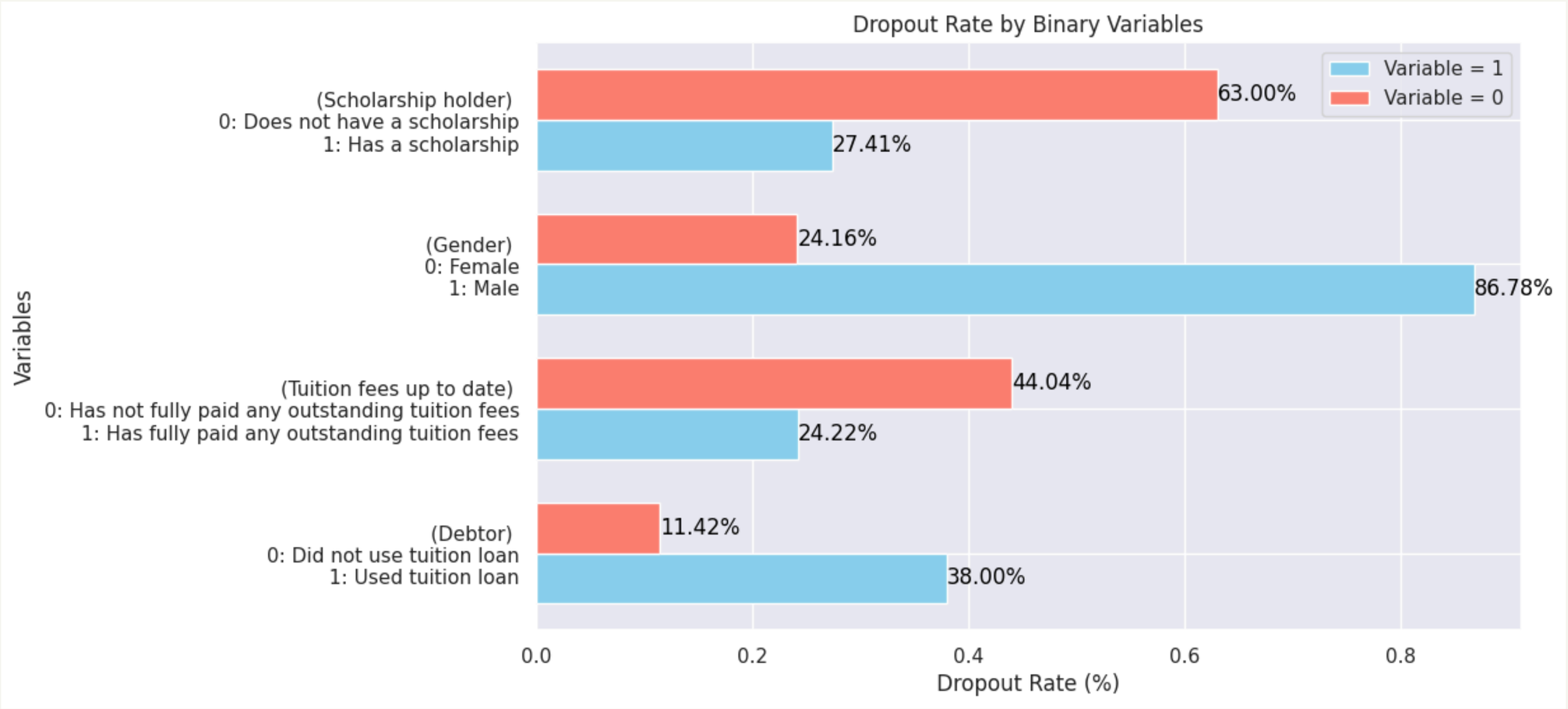
EXPLORATORY DATA ANALYSIS

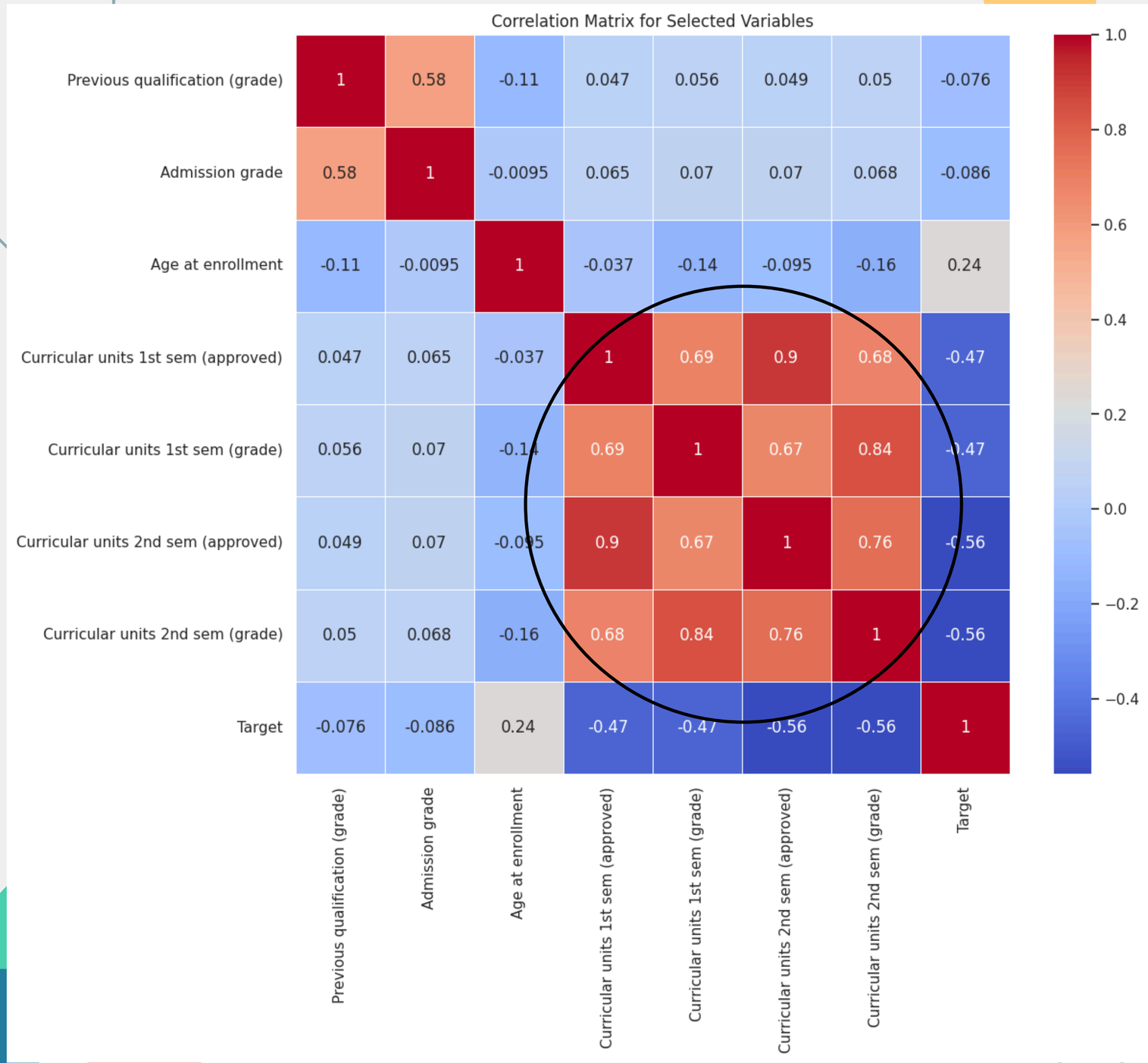
Dropout vs Non Dropout

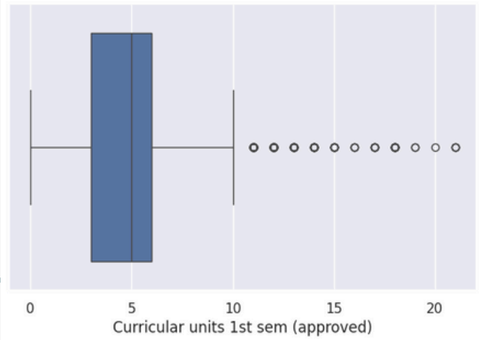


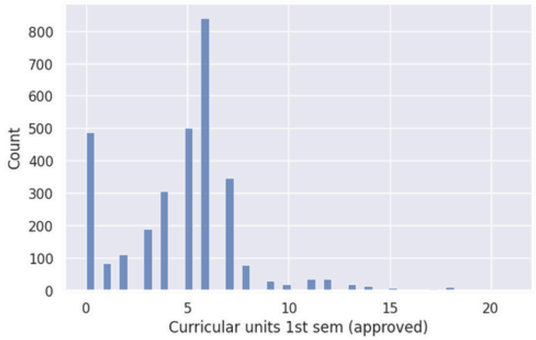

Dropout Rate by Parent's Occupation
Mother's occupation
Father's occupation
Occupation
Unskilled Labor
Student
Skilled Trades and Construction
Sales and Customer Service
Professional and Technical Services
Military
Manufacturing and Production
Administrative and Support Services
31.73%
30.84%
64.80%
67.91%
27.25%
29.05%
27.67%
29.09%
35.08%
32.84%
31.30%
25.00%
28.34%
41.67%
31.54%
28.21%
0
10
20
30
40
50
60
70
Dropout Rate (%)

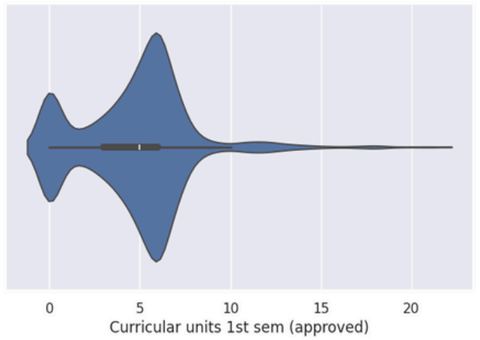

Dropout Rate by Course
Education
Science and Tech
Tourism
Business and Marketing
Humanities
Arts and Design
Healthcare
44.86%
42.72%
38.17%
31.53%
30.93%
29.51%
19.01%
Course
0
10
20
30
40
Dropout Rate (%)

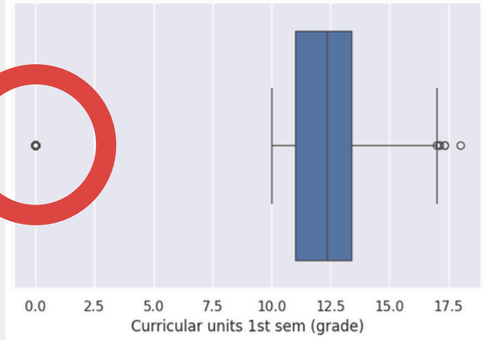

Dropout Rate by Binary Variables
Variable = 1
Variable = 0
(Scholarship holder)
0: Does not have a scholarship
1: Has a scholarship
63.00%
27.41%
(Gender)
0: Female
1: Male
24.16%
86.78%
(Tuition fees up to date)
0: Has not fully paid any outstanding tuition fees
1: Has fully paid any outstanding tuition fees
44.04%
24.22%
(Debtor)
0: Did not use tuition loan
1: Used tuition loan
11.42%
38.00%
Variables
0.0
0.2
0.4
0.6
0.8
Dropout Rate (%)

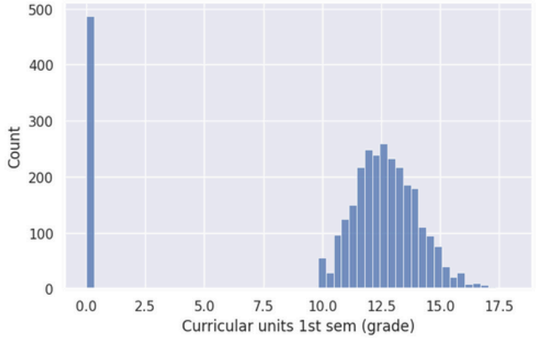

Correlation Matrix for Selected Variables
Previous qualification (grade)
Admission grade
Age at enrollment
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (grade)
Target
1 0.58 -0.11 0.047 0.056 0.049 0.05 -0.076
0.58 1 -0.0095 0.065 0.07 0.07 0.068 -0.086
-0.11 -0.0095 1 -0.037 -0.14 -0.095 -0.16 0.24
0.047 0.065 -0.037 1 0.69 0.9 0.68 -0.47
0.056 0.07 -0.14 0.69 1 0.67 0.84 -0.47
0.049 0.07 -0.095 0.9 0.67 1 0.76 -0.56
0.05 0.068 -0.16 0.68 0.84 0.76 1 -0.56
-0.076 -0.086 0.24 -0.47 -0.47 -0.56 -0.56 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

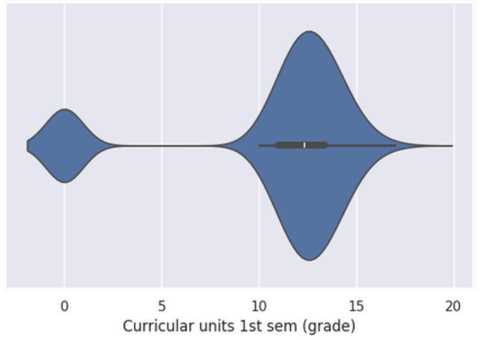

0
5
10
15
20
Curricular units 1st sem (approved)

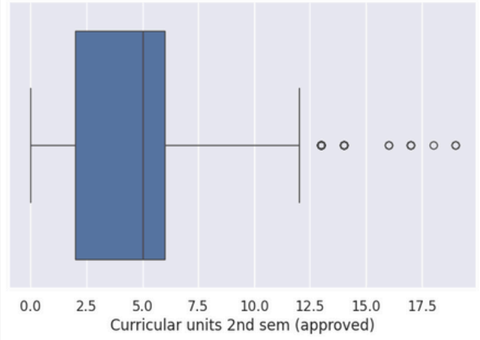

Count
800
700
600
500
400
300
200
100
0
0
5
10
15
20
Curricular units 1st sem (approved)

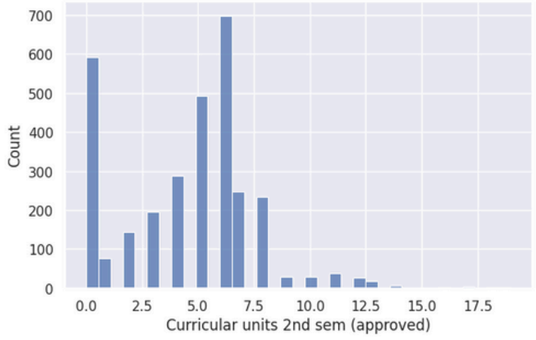

0
5
10
15
20
Curricular units 1st sem (approved)

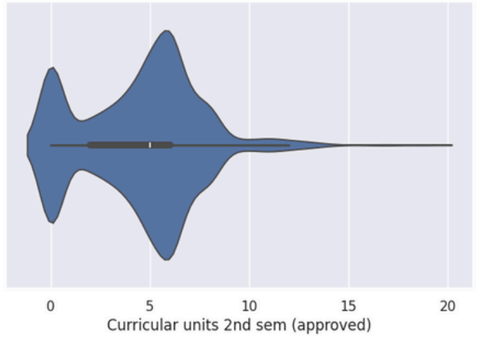

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 1st sem (grade)

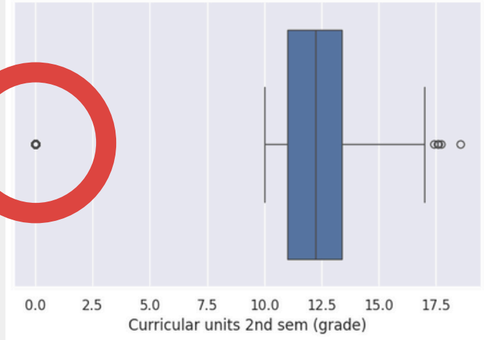

Count
500
400
300
200
100
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 1st sem (grade)

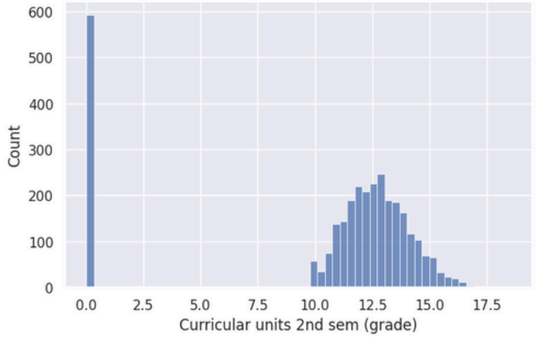

0
5
10
15
20
Curricular units 1st sem (grade)

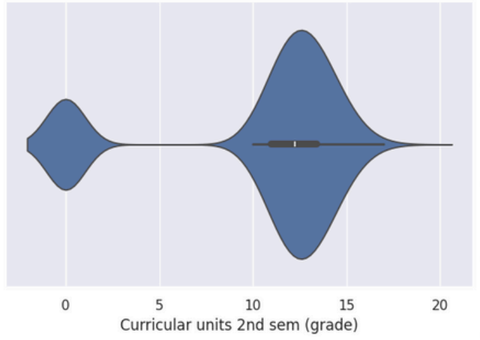

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 2nd sem (approved)

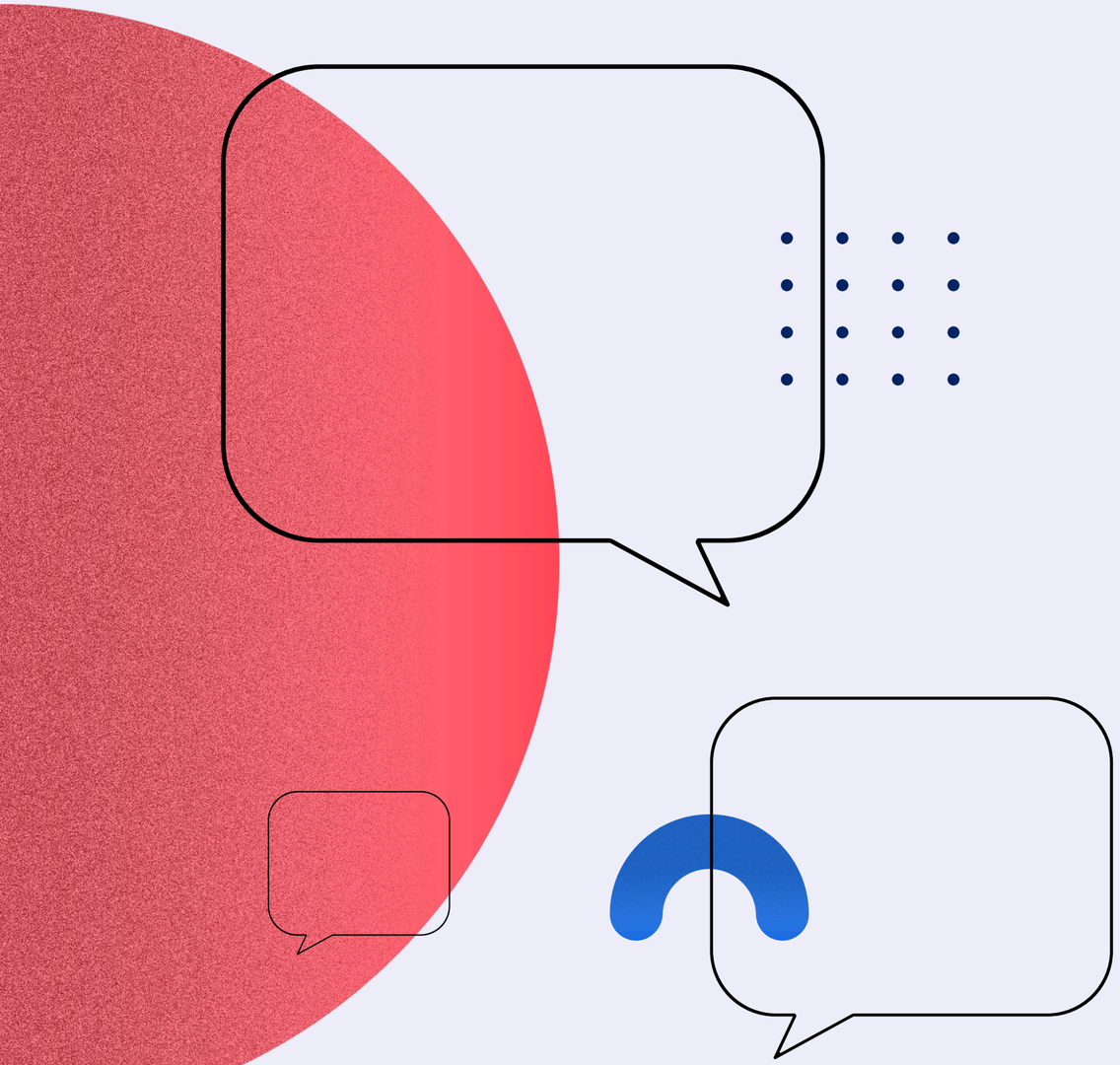

Count
700
600
500
400
300
200
100
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 2nd sem (approved)


0
5
10
15
20
Curricular units 2nd sem (approved)


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 2nd sem (grade)

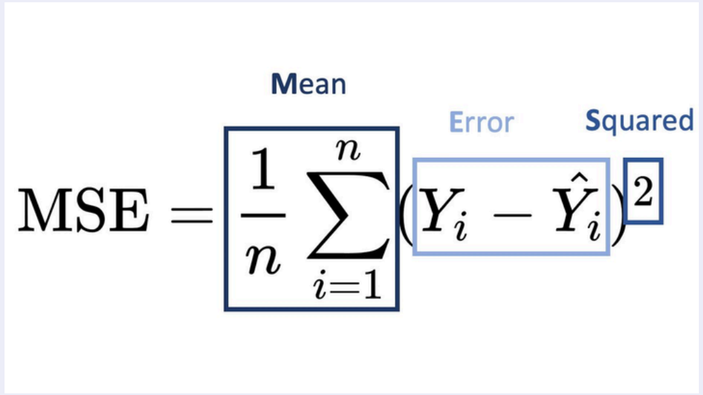

Count
600
500
400
300
200
100
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 2nd sem (grade)


0
5
10
15
20
Curricular units 2nd sem (grade)



MACHINE LEARNING

LINEAR REGRESSION

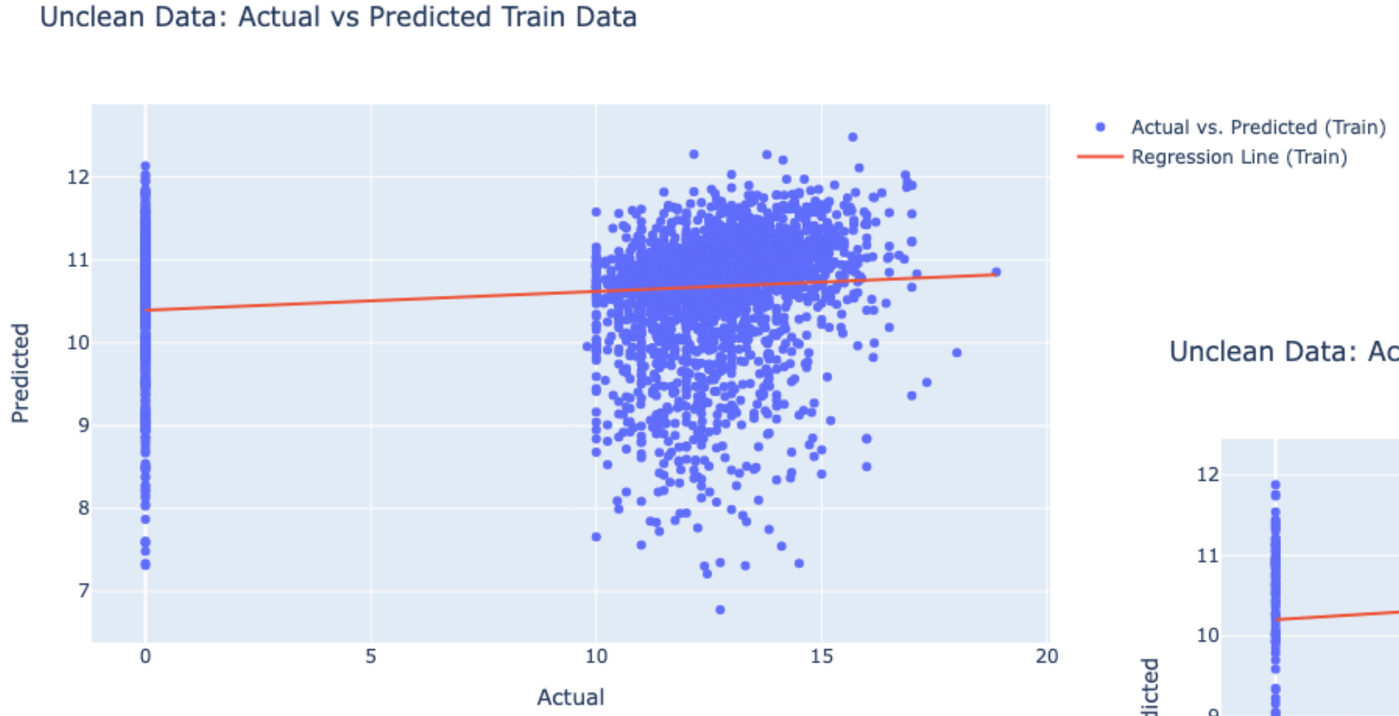


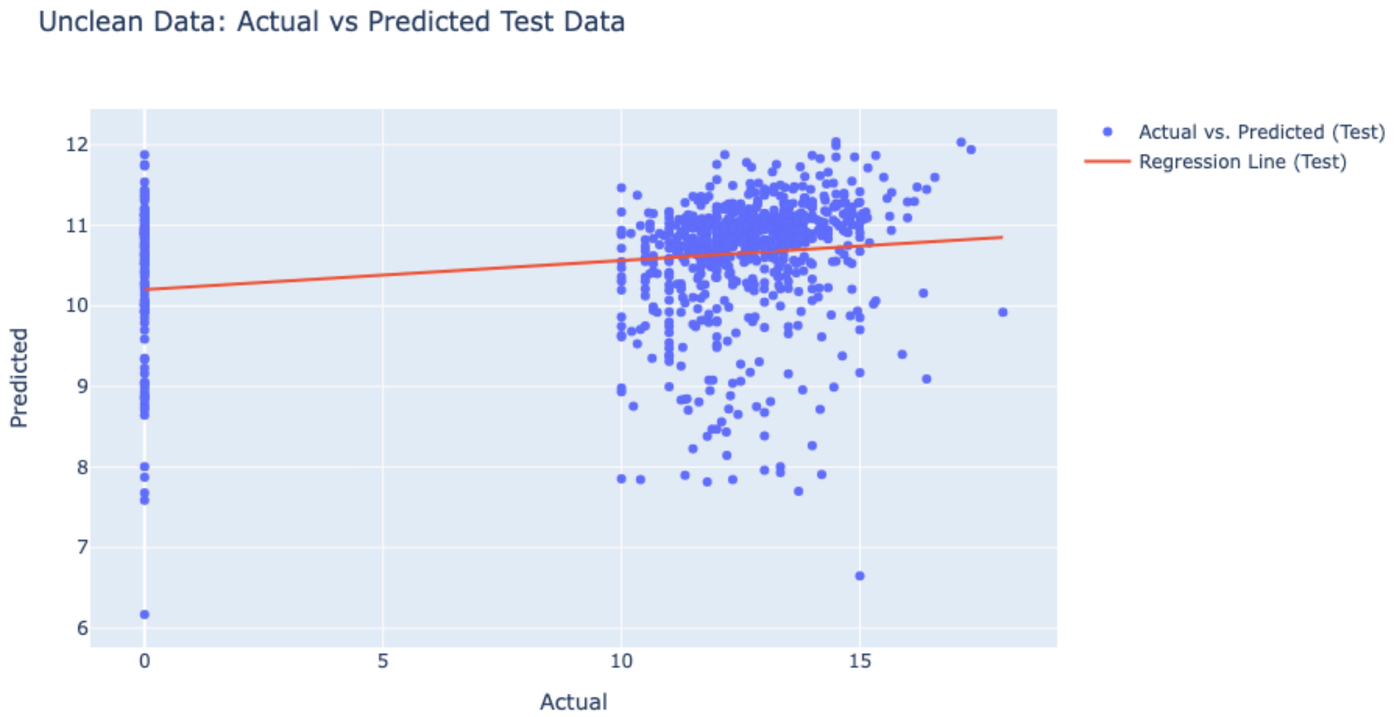
The Response Variables will be the Curricular units 1st sem (grade) and Curricular units 2nd sem (grade)

The Predictor Variables will be the variables that have floating point values:
Previous qualification, Age at enrollment, Admission grade, Previous qualification (grade)

We will be using MSE as a metric for our regression models.

LINEAR REGRESSION

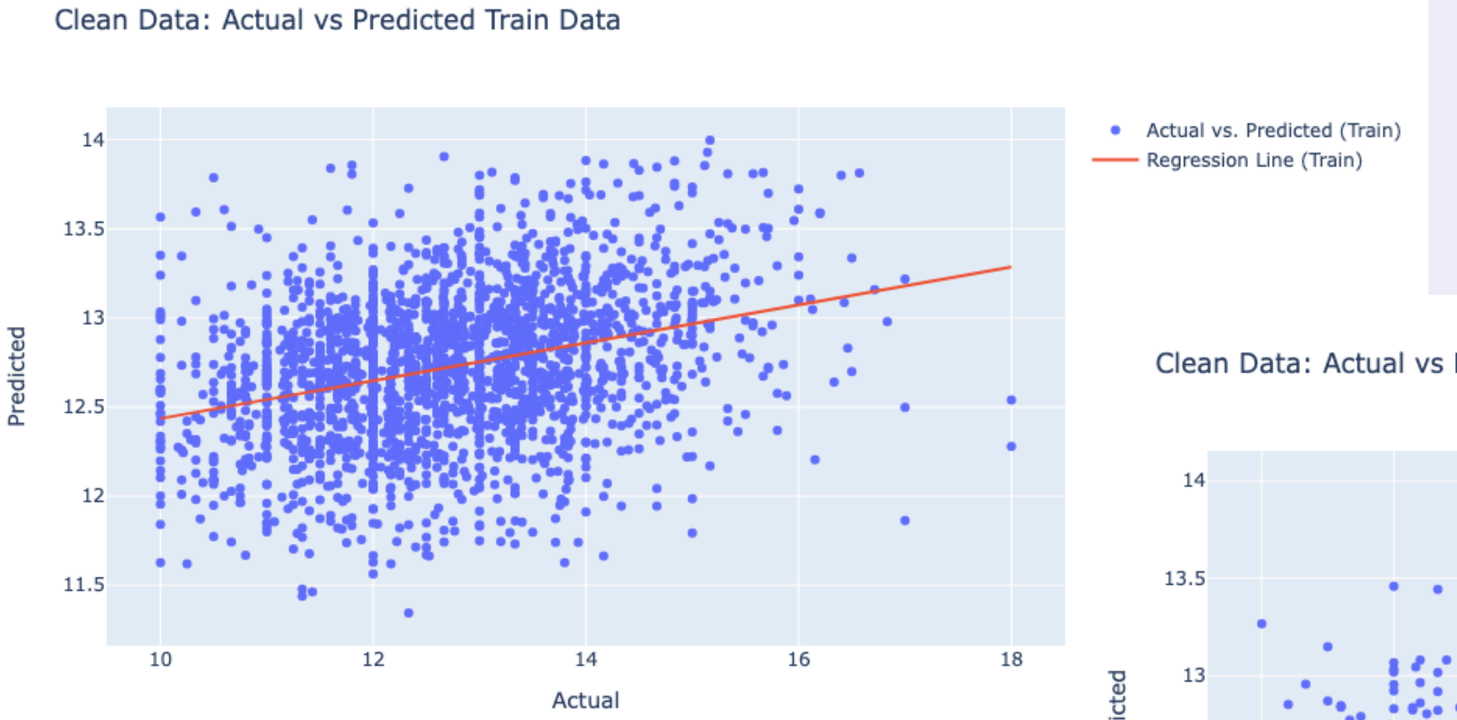


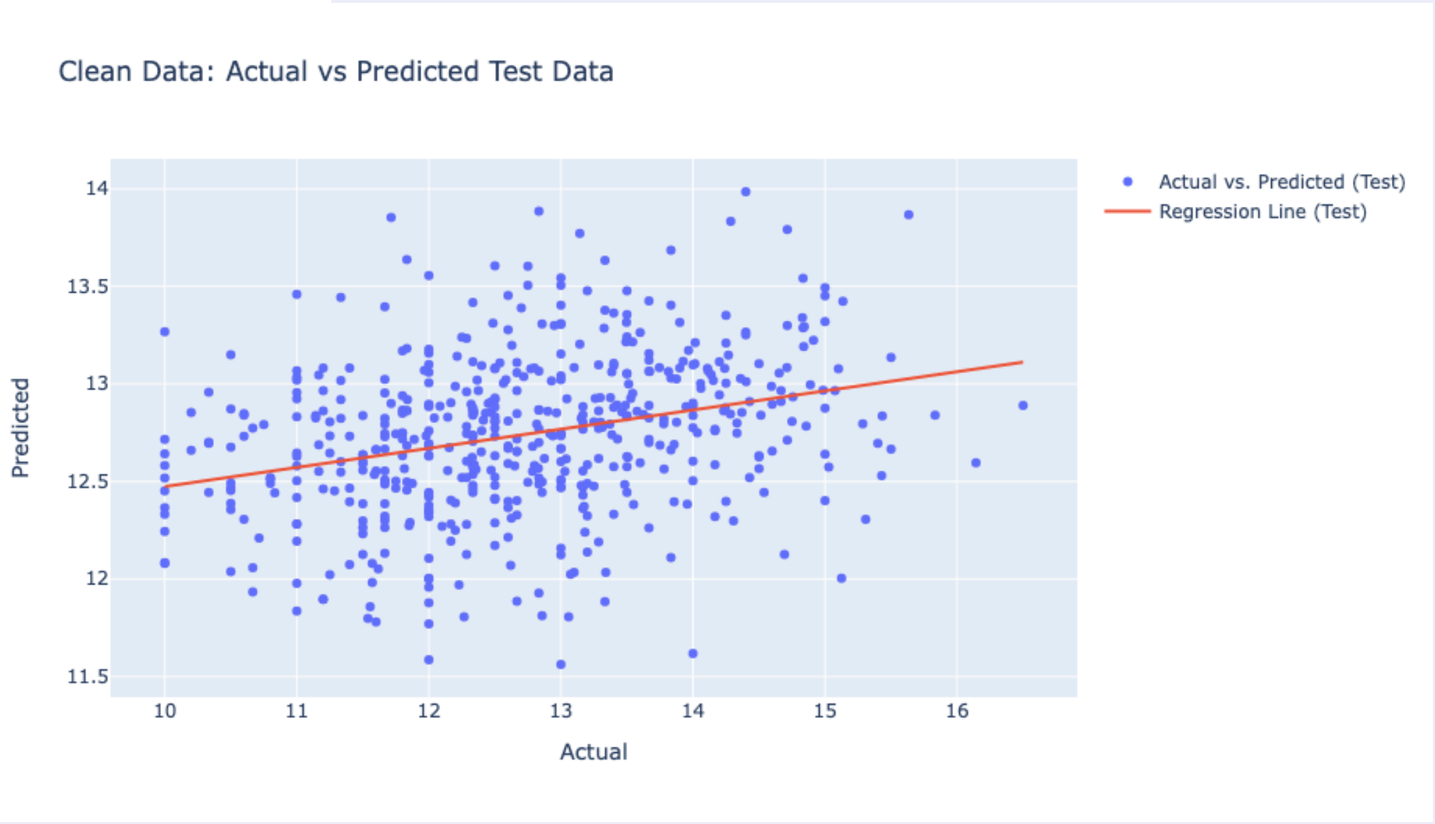


CLEAN DATA

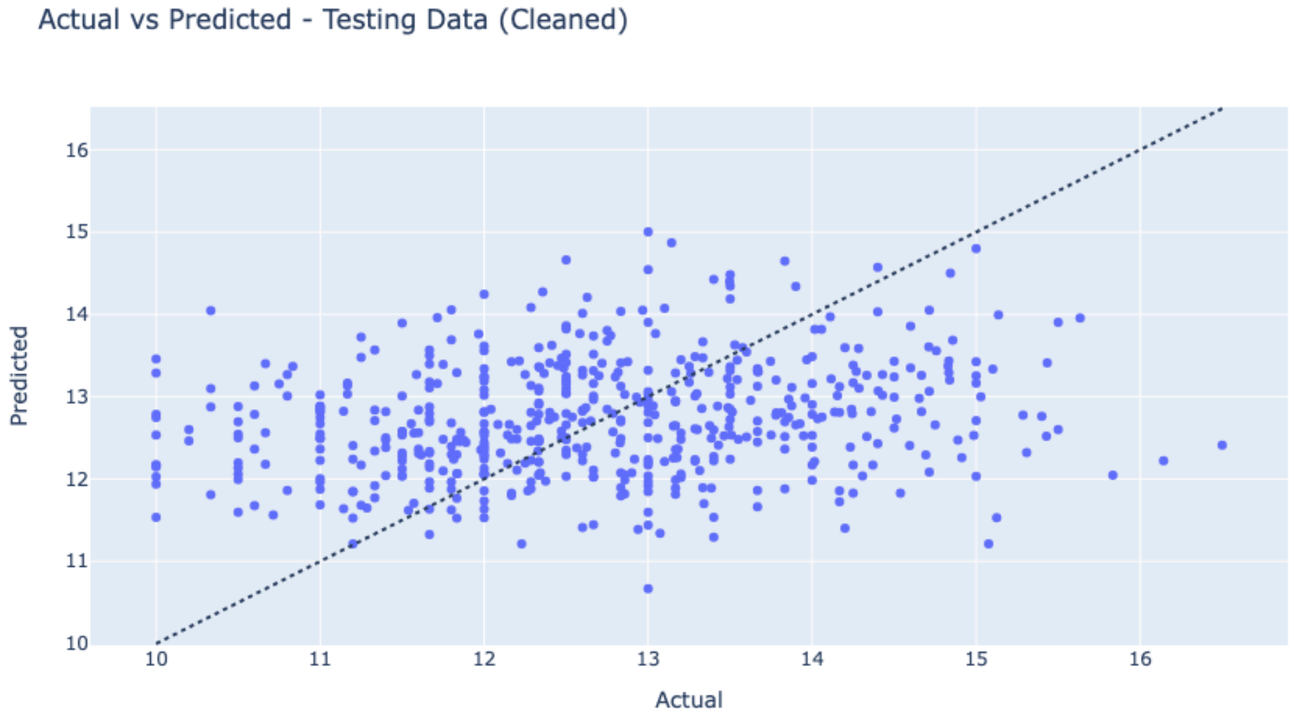
```
# Removing outliers
Q1 = X.quantile(0.25)
Q3 = X.quantile(0.75)
IQR = Q3 - Q1
X_clean = X[~((X < (Q1 - 1.5 * IQR)) | (X > (Q3 + 1.5 * IQR))).any(axis=1)]
y_clean = y.loc[X_clean.index]

# Remove all actual scores equal to 0 (as observed from the unclean data, it is obvious that they are outliers to be r
y_clean = y_clean[(y_clean["Curricular units 1st sem (grade)"] != 0) & (y_clean["Curricular units 2nd sem (grade)"] !=
X_clean = X_clean.loc[y_clean.index]
```

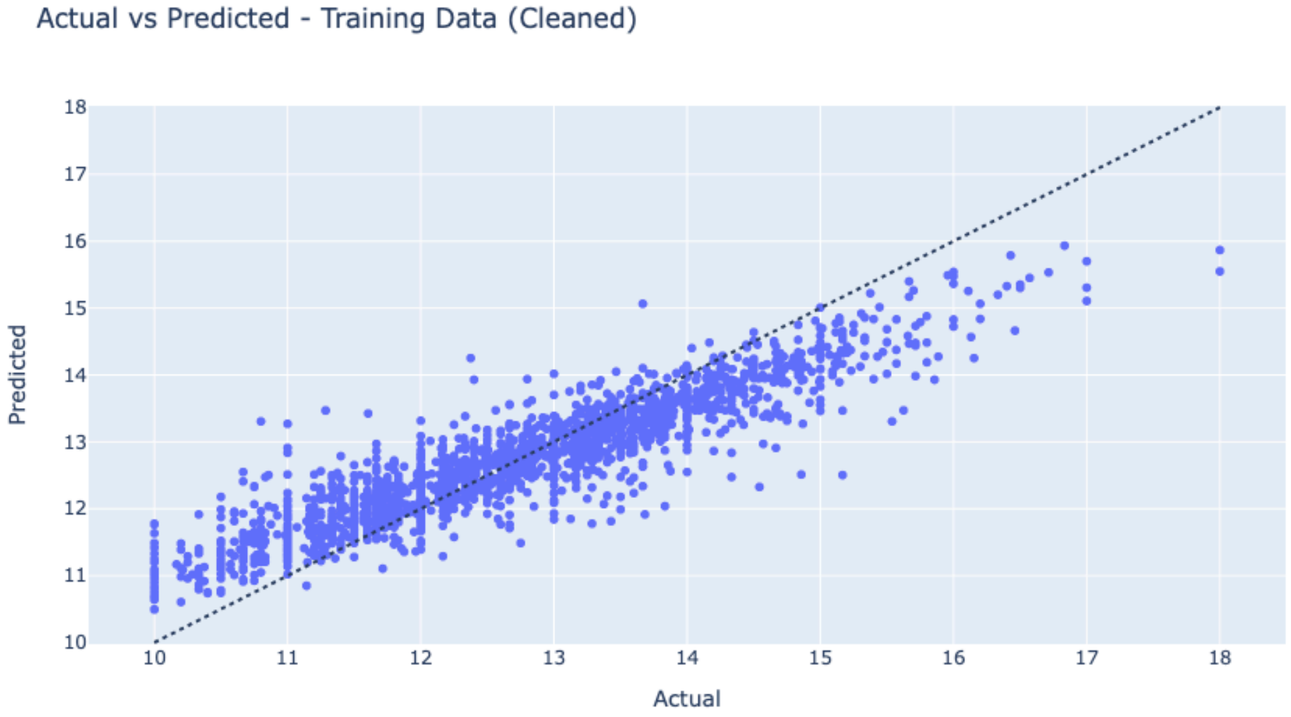
CLEAN DATA


Clean Data: Actual vs Predicted Train Data
Predicted
Actual
Actual vs. Predicted (Train)
Regression Line (Train)
14
13.5
13
12.5
12
11.5
10
12
14
16
18

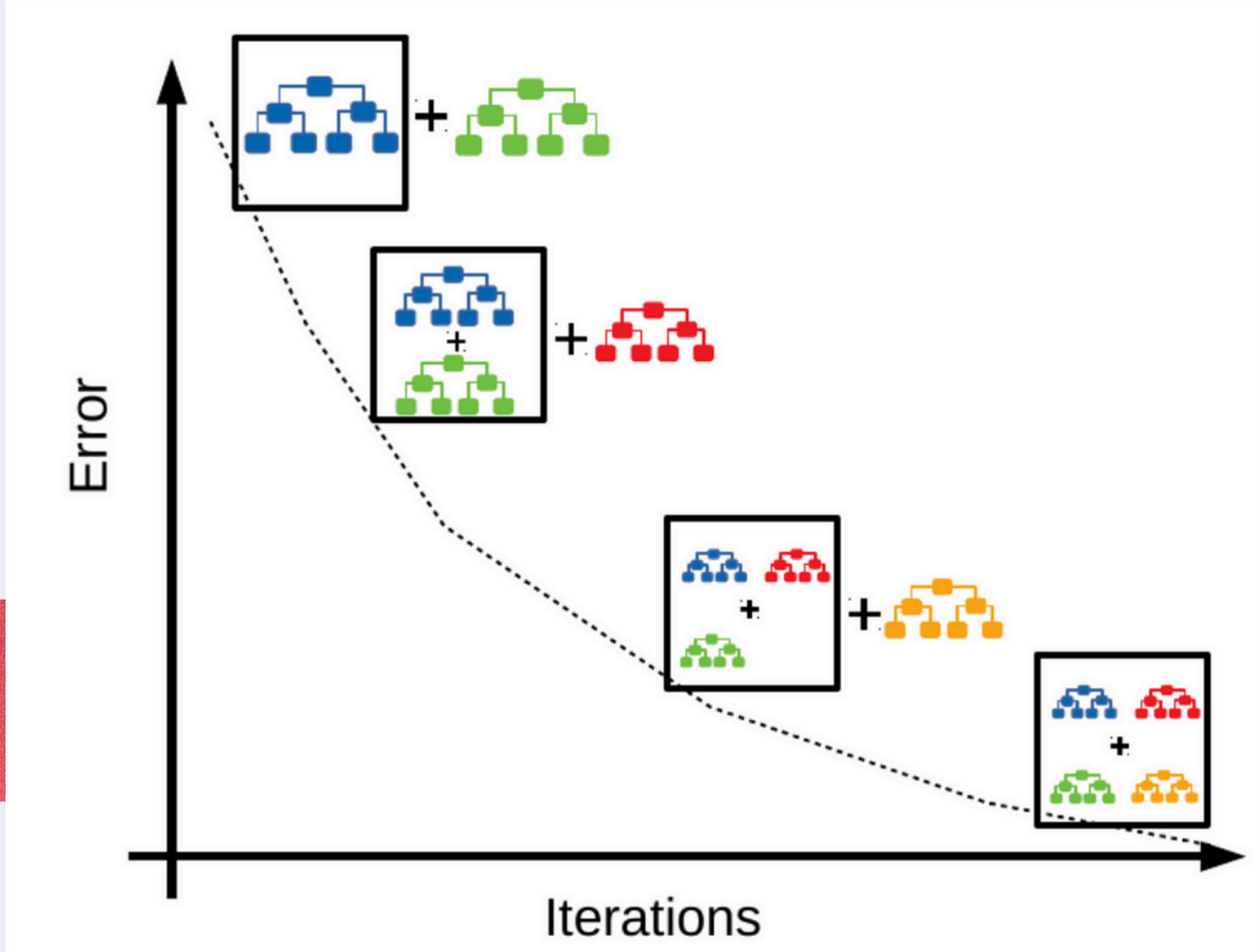

Clean Data: Actual vs Predicted Test Data
Predicted
Actual
Actual vs. Predicted (Test)
Regression Line (Test)
14
13.5
13
12.5
12
11.5
10
11
12
13
14
15
16

RANDOM FOREST REGRESSOR





GRADIENT BOOSTING

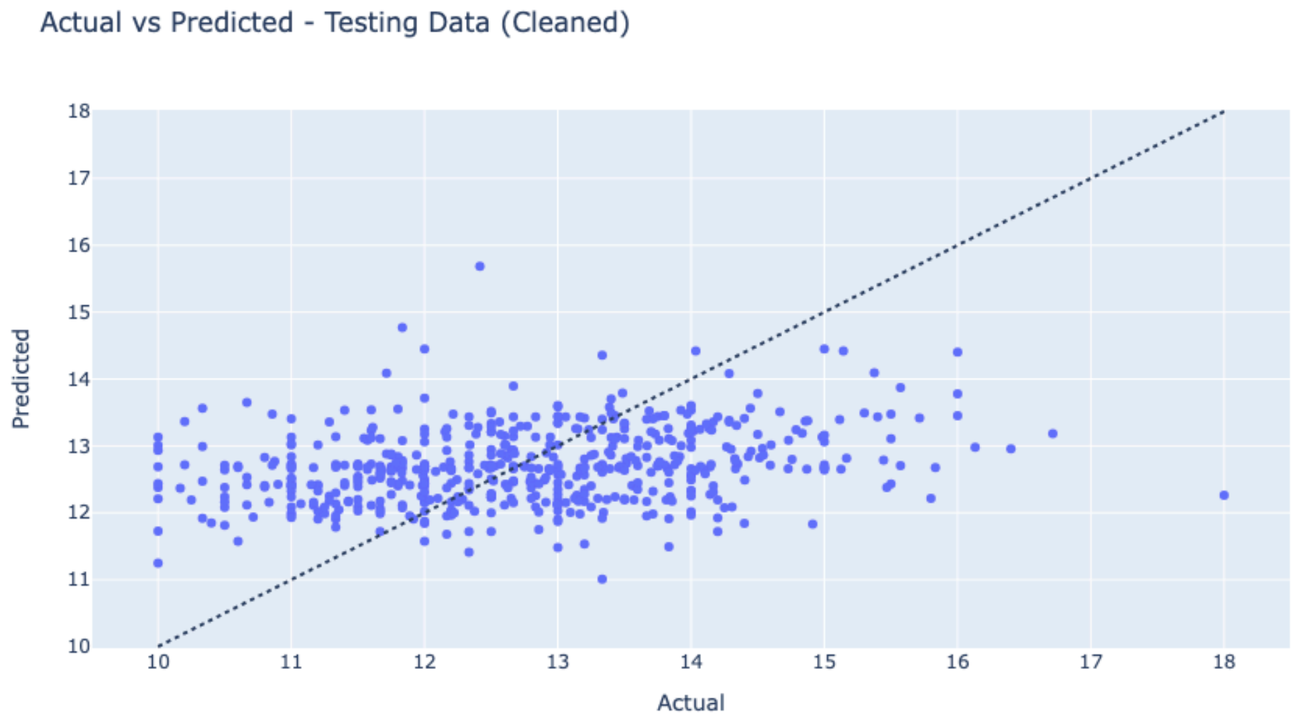


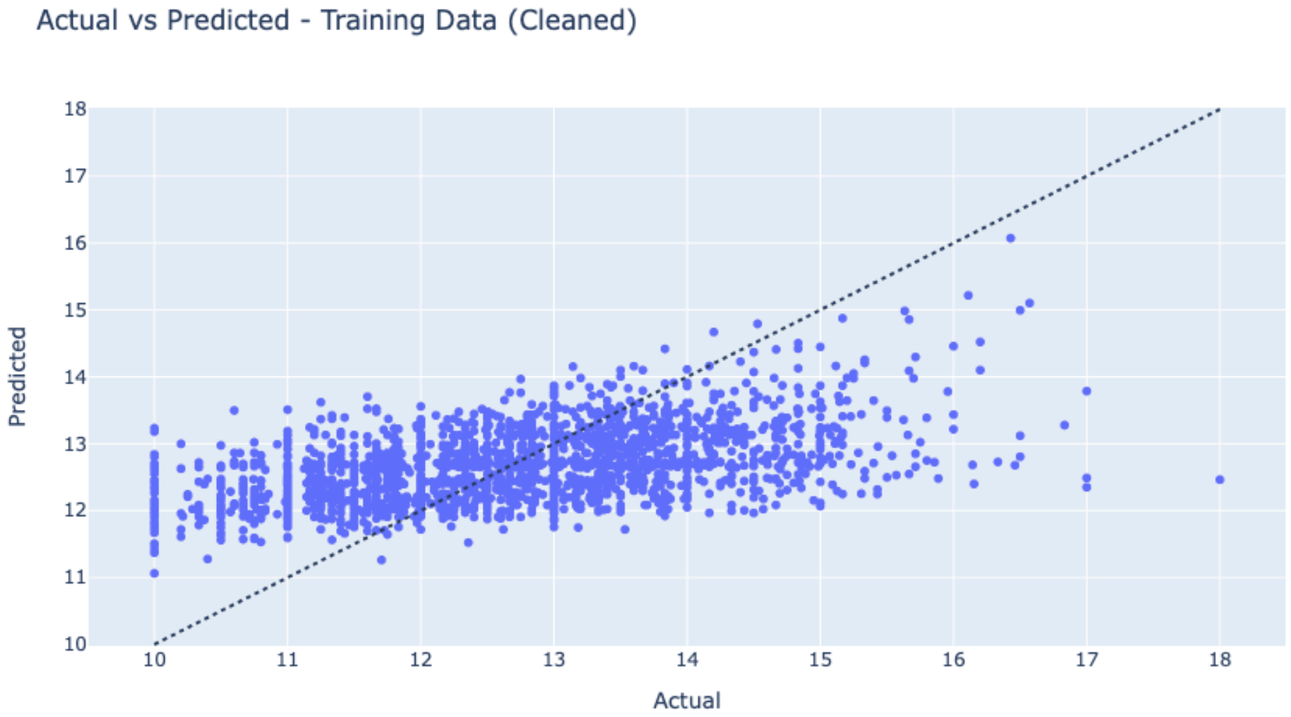
Building a series of weak learners (usually decision trees) sequentially, with each new learner correcting the errors made by the previous ones.

It fits a sequence of models to the data, where each new model focuses on the examples that were poorly predicted by the previous models

Gradient boosting typically produces highly accurate predictions compared to other machine learning algorithms.

GRADIENT BOOSTING





Best Model

LINEAR REGRESSION

Linear Regression performed the best with a lower average MSE when performing K-Fold Cross Validation.

Linear Relationships

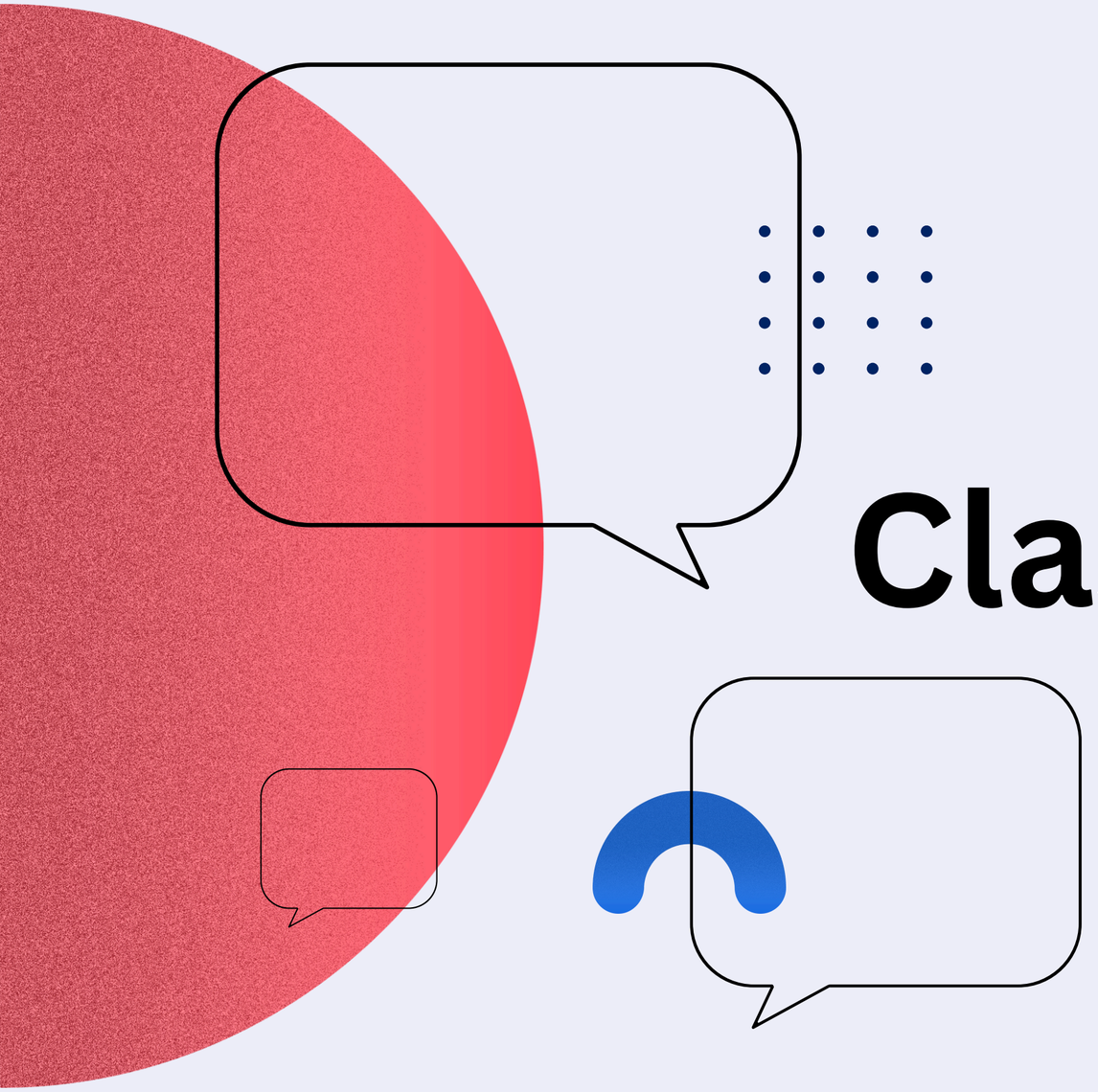
If the relationship between the features and the target variable is approximately linear, a linear regression model may provide a better fit

Avoidance of Overfitting

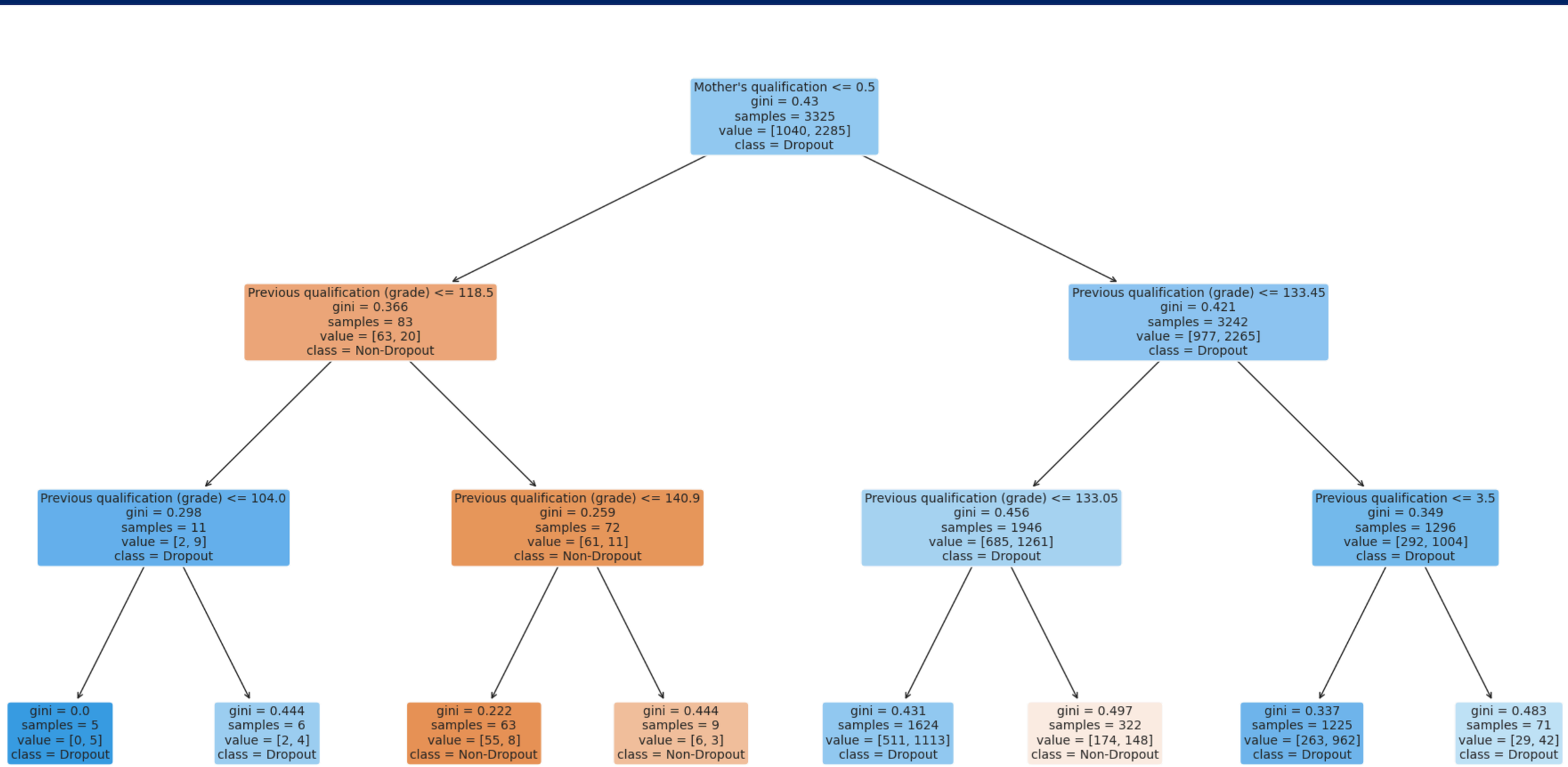
If there is limited training data or a risk of overfitting, a simpler linear regression model might generalize better.

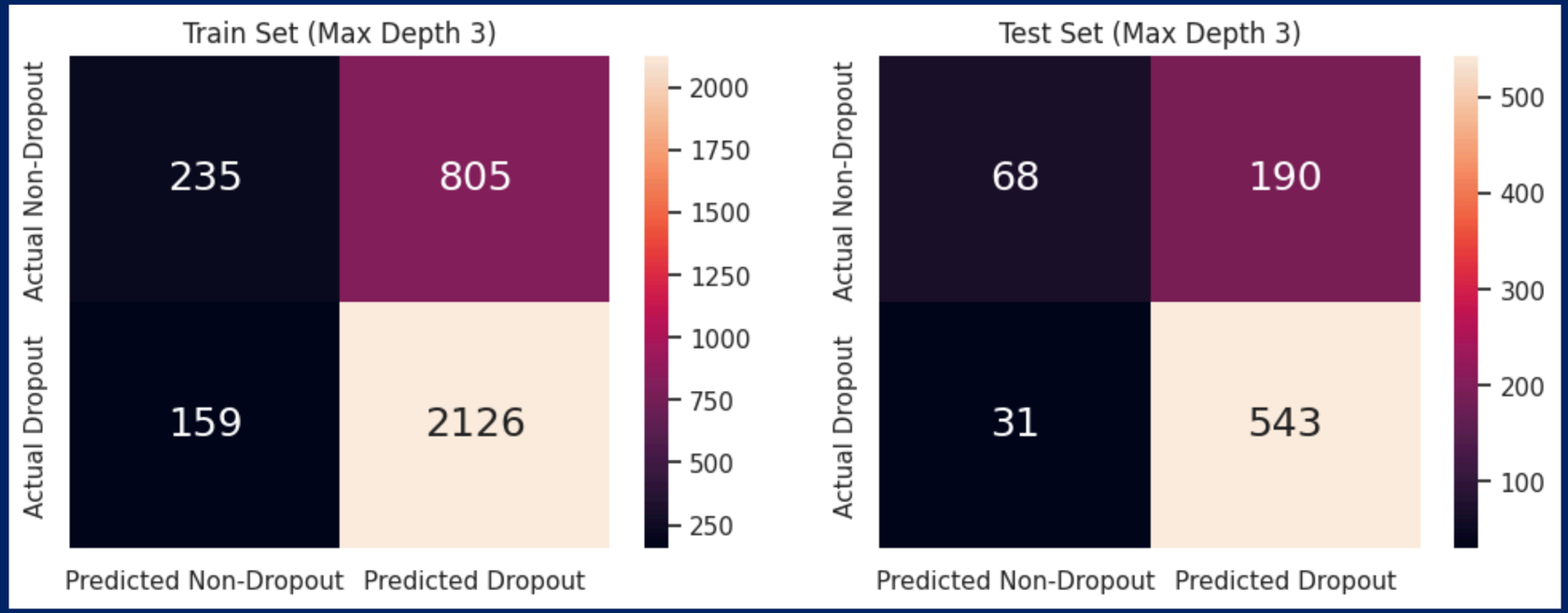
Low-Dimensional Data

Tree-based models may be prone to overfitting in such cases.

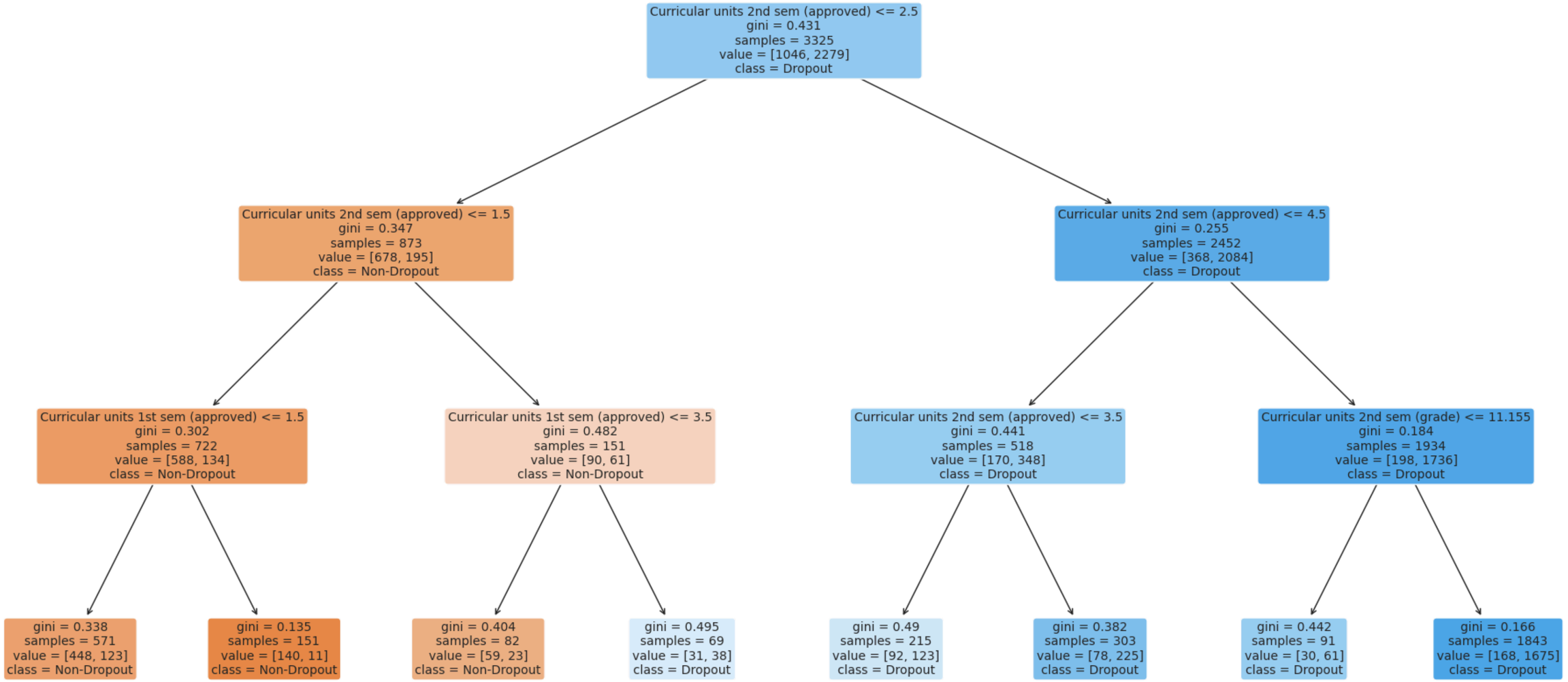


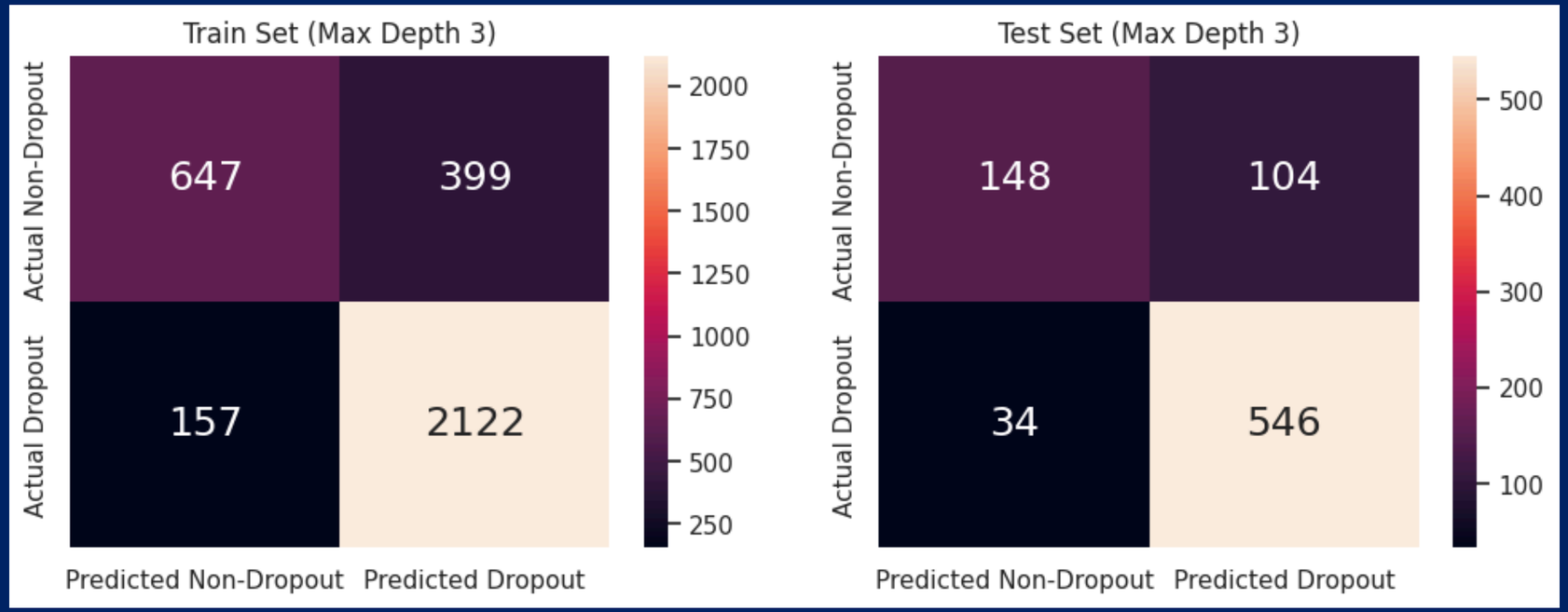
Classification Model


Mother's qualification <= 0.5
gini = 0.43
samples = 3325
value = [1040, 2285]
class = Dropout
Previous qualification (grade) <= 118.5
gini = 0.366
samples = 83
value = [63, 20]
class = Non-Dropout
Previous qualification (grade) <= 133.45
gini = 0.421
samples = 3242
value = [977, 2265]
class = Dropout
Previous qualification (grade) <= 104.0
gini = 0.298
samples = 11
value = [2, 9]
class = Dropout
Previous qualification (grade) <= 140.9
gini = 0.259
samples = 72
value = [61, 11]
class = Non-Dropout
Previous qualification (grade) <= 133.05
gini = 0.456
samples = 1946
value = [685, 1261]
class = Dropout
Previous qualification <= 3.5
gini = 0.349
samples = 1296
value = [292, 1004]
class = Dropout
gini = 0.0
samples = 5
value = [0, 5]
class = Dropout
gini = 0.444
samples = 6
value = [2, 4]
class = Dropout
gini = 0.222
samples = 63
value = [55, 8]
class = Non-Dropout
gini = 0.444
samples = 9
value = [6, 3]
class = Non-Dropout
gini = 0.431
samples = 1624
value = [511, 1113]
class = Dropout
gini = 0.497
samples = 322
value = [174, 148]
class = Non-Dropout
gini = 0.337
samples = 1225
value = [263, 962]
class = Dropout
gini = 0.483
samples = 71
value = [29, 42]
class = Dropout

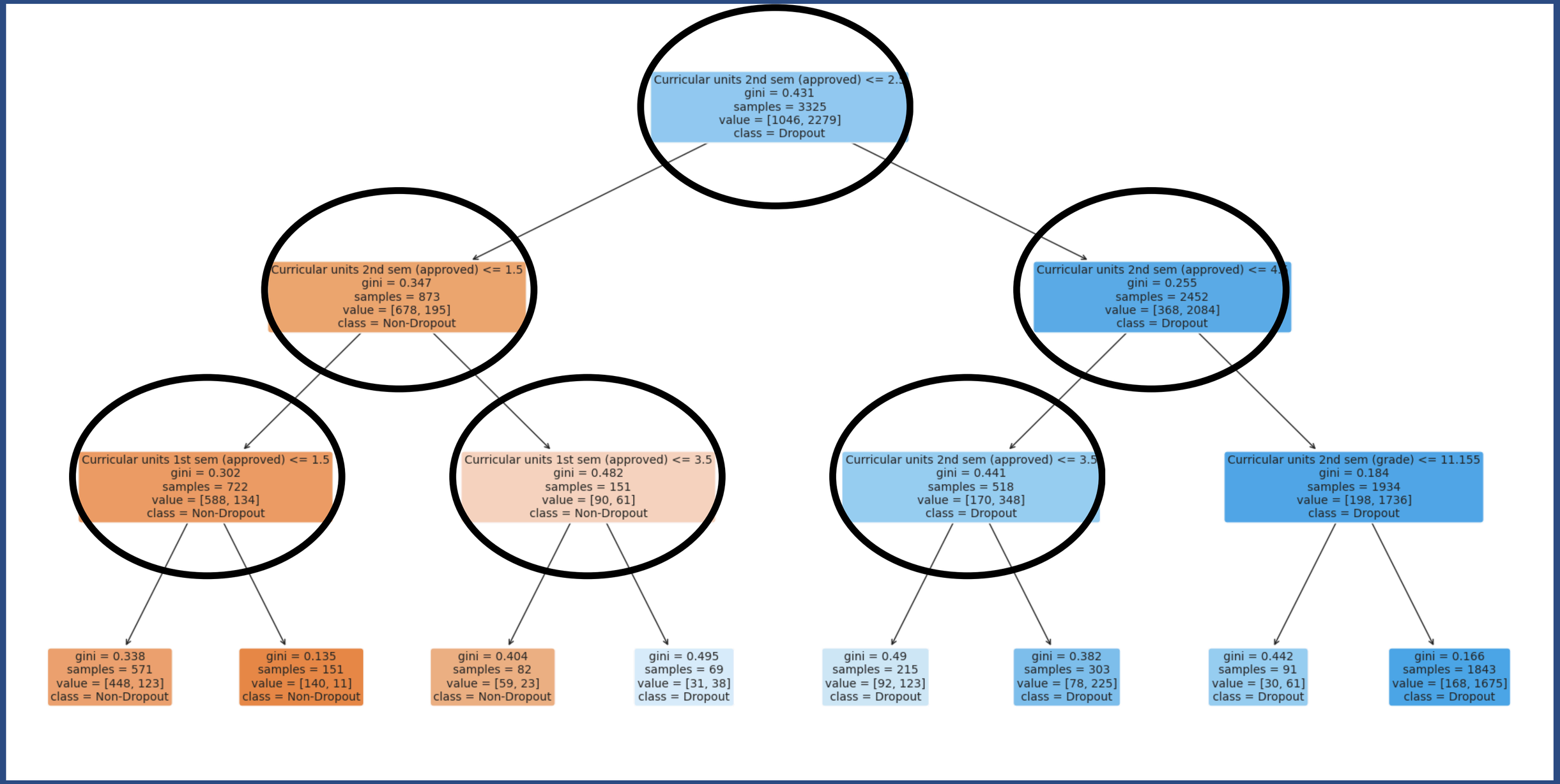



```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.7100751879699249
F1 Score                        : 0.3277545327754533
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.734375
F1 Score                        : 0.380952380952381
```


Curricular units 2nd sem (approved) <= 2.5
gini = 0.431
samples = 3325
value = [1046, 2279]
class = Dropout
Curricular units 2nd sem (approved) <= 1.5
gini = 0.347
samples = 873
value = [678, 195]
class = Non-Dropout
Curricular units 2nd sem (approved) <= 4.5
gini = 0.255
samples = 2452
value = [368, 2084]
class = Dropout
Curricular units 1st sem (approved) <= 1.5
gini = 0.302
samples = 722
value = [588, 134]
class = Non-Dropout
Curricular units 1st sem (approved) <= 3.5
gini = 0.482
samples = 151
value = [90, 61]
class = Non-Dropout
Curricular units 2nd sem (approved) <= 3.5
gini = 0.441
samples = 518
value = [170, 348]
class = Dropout
Curricular units 2nd sem (grade) <= 11.155
gini = 0.184
samples = 1934
value = [198, 1736]
class = Dropout
gini = 0.338
samples = 571
value = [448, 123]
class = Non-Dropout
gini = 0.135
samples = 151
value = [140, 11]
class = Non-Dropout
gini = 0.404
samples = 82
value = [59, 23]
class = Non-Dropout
gini = 0.495
samples = 69
value = [31, 38]
class = Dropout
gini = 0.49
samples = 215
value = [92, 123]
class = Dropout
gini = 0.382
samples = 303
value = [78, 225]
class = Dropout
gini = 0.442
samples = 91
value = [30, 61]
class = Dropout
gini = 0.166
samples = 1843
value = [168, 1675]
class = Dropout




```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.8327819548872181
F1 Score                        : 0.6994594594594594
Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.8341346153846154
F1 Score                        : 0.6820276497695853
```


Curricular units 2nd sem (approved) <= 2.5
gini = 0.431
samples = 3325
value = [1046, 2279]
class = Dropout
Curricular units 2nd sem (approved) <= 1.5
gini = 0.347
samples = 873
value = [678, 195]
class = Non-Dropout
Curricular units 2nd sem (approved) <= 4.5
gini = 0.255
samples = 2452
value = [368, 2084]
class = Dropout
Curricular units 1st sem (approved) <= 1.5
gini = 0.302
samples = 722
value = [588, 134]
class = Non-Dropout
Curricular units 1st sem (approved) <= 3.5
gini = 0.482
samples = 151
value = [90, 61]
class = Non-Dropout
Curricular units 2nd sem (approved) <= 3.5
gini = 0.441
samples = 518
value = [170, 348]
class = Dropout
Curricular units 2nd sem (grade) <= 11.155
gini = 0.184
samples = 1934
value = [198, 1736]
class = Dropout
gini = 0.338
samples = 571
value = [448, 123]
class = Non-Dropout
gini = 0.135
samples = 151
value = [140, 11]
class = Non-Dropout
gini = 0.404
samples = 82
value = [59, 23]
class = Non-Dropout
gini = 0.495
samples = 69
value = [31, 38]
class = Dropout
gini = 0.49
samples = 215
value = [92, 123]
class = Dropout
gini = 0.382
samples = 303
value = [78, 225]
class = Dropout
gini = 0.442
samples = 91
value = [30, 61]
class = Dropout
gini = 0.166
samples = 1843
value = [168, 1675]
class = Dropout

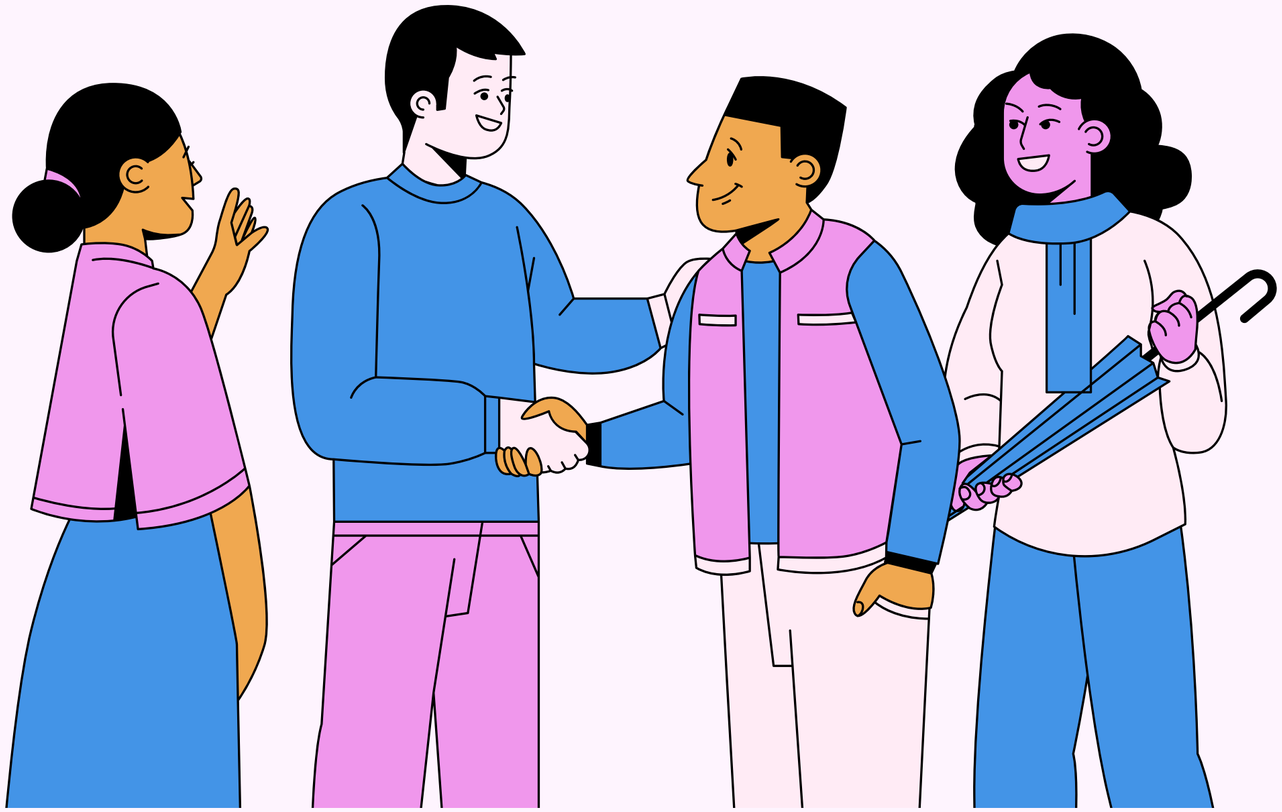
PROJECT OUTCOME

-Predictions using Regression Models

-Predictions using Classification Models

CONCLUSION

Helping aid for Universities/Teachers to identify students to administer help to reduce dropout rates



THANK YOU