DATA319_2024Summer_Final_Project_Team1

July 26, 2024

Data 319, Summer 2024

Washington State University

Final Report

Students: Jason Cross, Zachary Davis, Abagail Ponsetto, Shae Sandland, Natalie Toledo

0.0.1 Introduction

This project can potentially impact student success by evaluating college students' dropout rate and academic performance and help determine what attributes influence that outcome. The original data, sourced from a university in Portugal, may provide valuable insights that could be applicable to the United States and other contexts. This study aims to determine whether the features in the data set can predict whether a student will drop out, graduate, or enroll.

0.0.2 Approach

The study first focused on which variables might be statistically significant in determining the outcome. Then, the data were analyzed for their relationships to the response variable (dependent variable) and to each other. Once the statistically significant variable(s) were identified using one model, different models were constructed to determine the best model for predicting student graduation, dropout, or enrollment. Further analysis was conducted using visualization methods and tools available in Python. Research Question: The pivotal question our team will address is to identify the variables that significantly influence a student's decision to drop out, enroll or graduate. Based on the observed data, we aim to leverage these insights to predict whether a student will dropout, enroll, or graduate. Because the question we are attempting to answer is a classification problem, we intend to use the following methods to analyze and model the data: Logistic Regression, K-Nearest, and SVM. Logistic regression is a good model for predicting the probability of an outcome and estimating relationships between a dependent variable and one or more independent variables. Additionally, logistic regression can be ideal for this type of analysis because we are predicting a categorical variable (did a student drop out, enroll, or graduate). K-Nearest Neighbor is a supervised learning prediction tool typically used for classification problems like student dropout prediction. Because the dataset has many variables, some of which seem similar, KNN might help determine the similarity between variables. SVM (Support Vector Machine) is a standard supervised learning method for classification and regression problems. It is flexible and is used for linear and nonlinear problems.

Number of observations(rows): 4424
Number of features(columns): 35

Total Null: 0
Total NA: 0

Describe/Look for outliers:

	Marital status	Application mode	Application order	Course	\
count	4424.000000	4424.000000	4424.000000	4424.000000	
mean	1.178571	6.886980	1.727848	9.899186	
std	0.605747	5.298964	1.313793	4.331792	
min	1.000000	1.000000	0.000000	1.000000	
25%	1.000000	1.000000	1.000000	6.000000	
50%	1.000000	8.000000	1.000000	10.000000	
75%	1.000000	12.000000	2.000000	13.000000	
max	6.000000	18.000000	9.000000	17.000000	

	Daytime/evening attendance	Previous qualification	Nacionality	\
count	4424.000000	4424.000000	4424.000000	
mean	0.890823	2.531420	1.254521	
std	0.311897	3.963707	1.748447	
min	0.000000	1.000000	1.000000	
25%	1.000000	1.000000	1.000000	
50%	1.000000	1.000000	1.000000	
75%	1.000000	1.000000	1.000000	
max	1.000000	17.000000	21.000000	

	Mother's qualification	Father's qualification	Mother's occupation	\
count	4424.000000	4424.000000	4424.000000	
mean	12.322107	16.455244	7.317812	
std	9.026251	11.044800	3.997828	
min	1.000000	1.000000	1.000000	
25%	2.000000	3.000000	5.000000	
50%	13.000000	14.000000	6.000000	
75%	22.000000	27.000000	10.000000	
max	29.000000	34.000000	32.000000	

	…	Curricular units 1st sem (without evaluations)	\
count	…	4424.000000	
mean	…	0.137658	
std	…	0.690880	
min	…	0.000000	
25%	…	0.000000	
50%	…	0.000000	
75%	…	0.000000	
max	…	12.000000	

	Curricular units 2nd sem (credited)	\
count	4424.000000	
mean	0.541817	
std	1.918546	
min	0.000000	
25%	0.000000	
50%	0.000000	
75%	0.000000	
max	19.000000	

	Curricular units 2nd sem (enrolled)	\
count	4424.000000	
mean	6.232143	
std	2.195951	
min	0.000000	
25%	5.000000	
50%	6.000000	
75%	7.000000	
max	23.000000	

	Curricular units 2nd sem (evaluations)	\
count	4424.000000	
mean	8.063291	
std	3.947951	
min	0.000000	
25%	6.000000	
50%	8.000000	
75%	10.000000	
max	33.000000	

	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	\
count	4424.000000	4424.000000	
mean	4.435805	10.230206	
std	3.014764	5.210808	
min	0.000000	0.000000	
25%	2.000000	10.750000	
50%	5.000000	12.200000	
75%	6.000000	13.333333	
max	20.000000	18.571429	

	Curricular units 2nd sem (without evaluations)	Unemployment rate	\
count	4424.000000	4424.000000	
mean	0.150316	11.566139	
std	0.753774	2.663850	
min	0.000000	7.600000	
25%	0.000000	9.400000	
50%	0.000000	11.100000	
75%	0.000000	13.900000	

```
max                                          12.000000          16.200000
```

```
       Inflation rate          GDP
count     4424.000000  4424.000000
mean         1.228029     0.001969
std          1.382711     2.269935
min         -0.800000    -4.060000
25%          0.300000    -1.700000
50%          1.400000     0.320000
75%          2.600000     1.790000
max          3.700000     3.510000

[8 rows x 34 columns]
```

0.0.3 Data Description

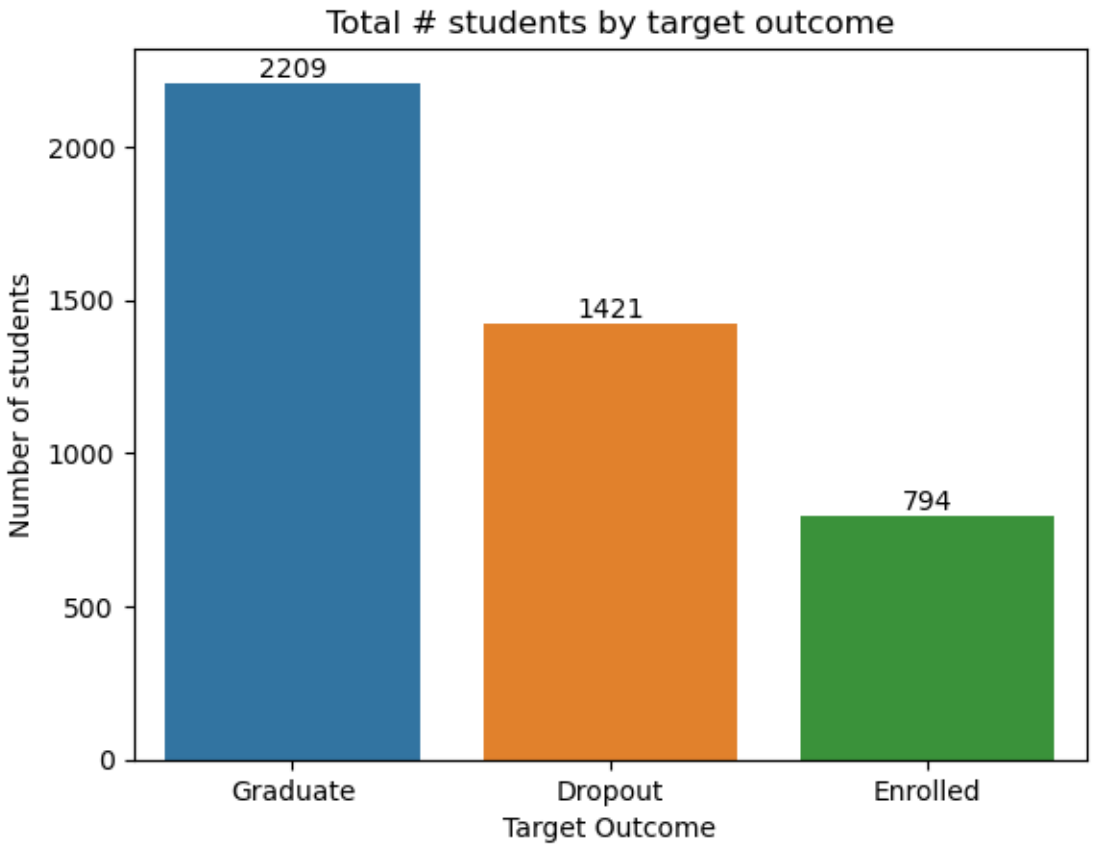
The data set is taken from (https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention) and consists of 4424 rows(observations) and 35 columns(variables/features). Each observation of the dataset represents a student.

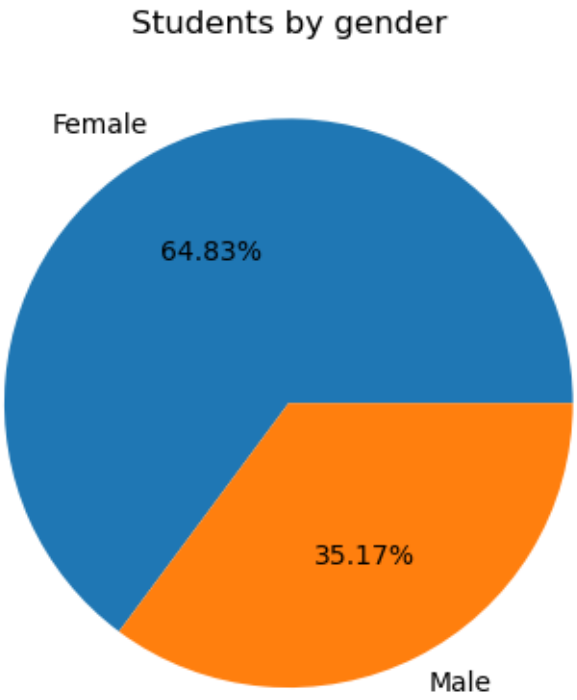
This study will use 'Target' as the response variable (y). The 'Target' variable has three values: 'Enrolled,' 'Graduate,' or 'Dropout,' each indicating the student's action after taking a course (the course is aptly saved in a variable called 'Course').

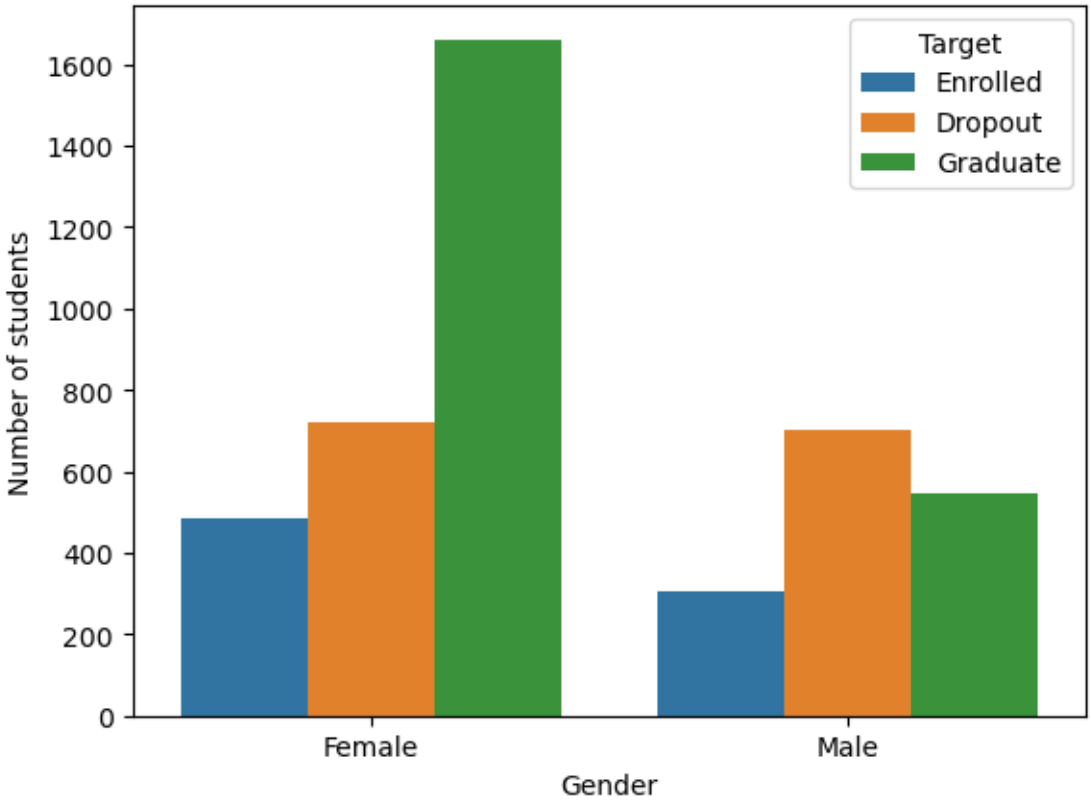
It's worth noting that the initial data check has confirmed the absence of null or NA values for any columns in the dataset. This means there is no need to spend time cleaning or massaging the data to account for missing or problematic data.An initial review of the data with the describe() function shows that there are probably no major outliers. However, the mean for Curricular units 2nd sem (without evaluations) and for GDP, when compared to their maximum values, do seem far off. Unless these features emerge as significant or important, we can probably ignore them for now.

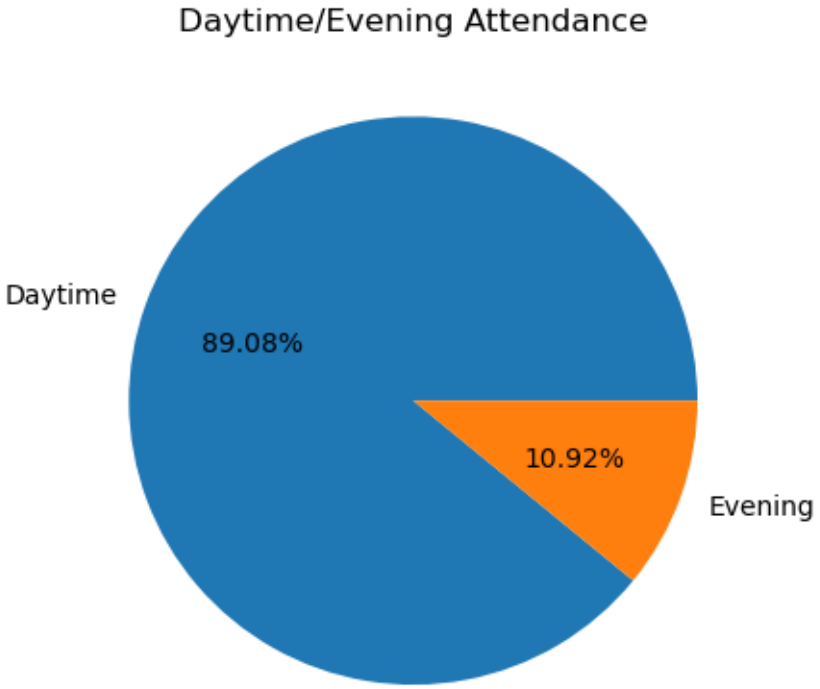
The data is loosely categorized as socioeconomic(e.g., mother's qualification, mother's occupation, debtor, tuition fees up to date), demographic(e.g., marital status, nationality, gender, age at enrollment, international), economic(e.g., unemployment rate, inflation rate, GDP), and academic (e.g., application order, course, Curriculum units).

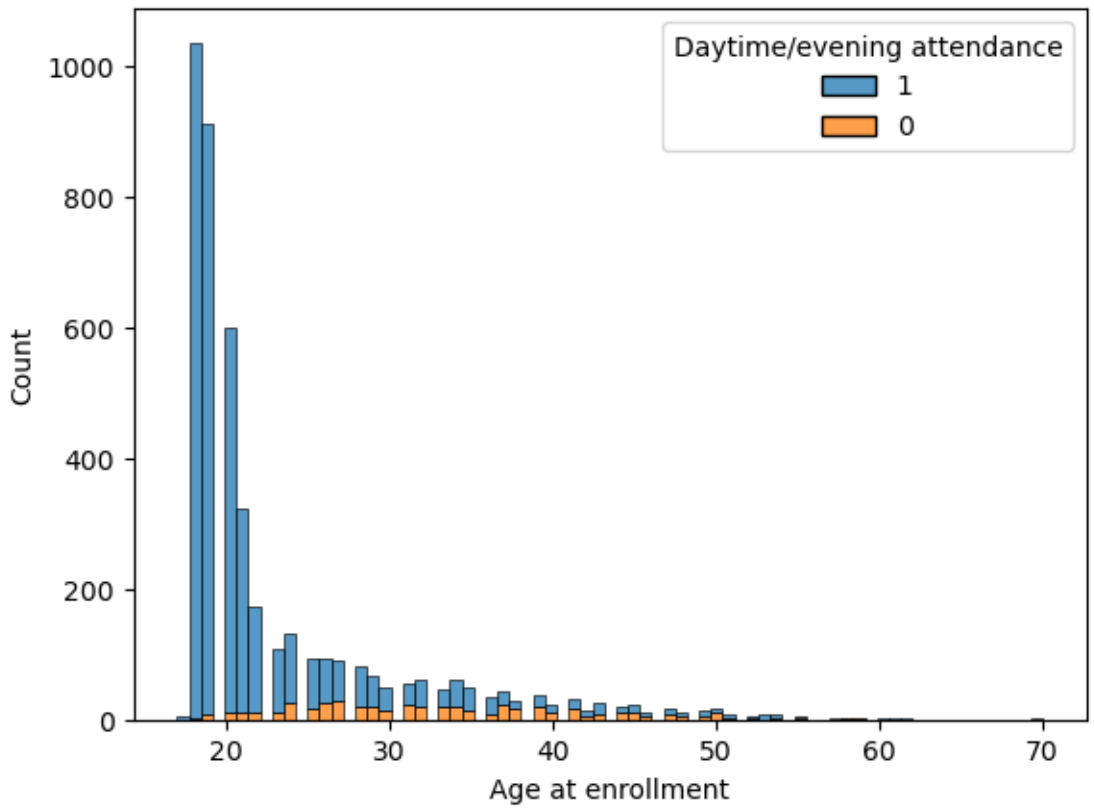
Given that the 'Target' variable is a qualitative/classification variable with 3 values, a multiple logistic regression model makes sense. This model is well suited to handle its complexity and will be instrumental in our analysis.


Total # students by target outcome
2209
1421
794
2000
1500
1000
500
0
Number of students
Graduate
Dropout
Enrolled
Target Outcome


Students by gender
Female
64.83%
35.17%
Male


Target
Enrolled
Dropout
Graduate
Number of students
0
200
400
600
800
1000
1200
1400
1600
Female
Male
Gender


Daytime/Evening Attendance
Daytime
89.08%
10.92%
Evening

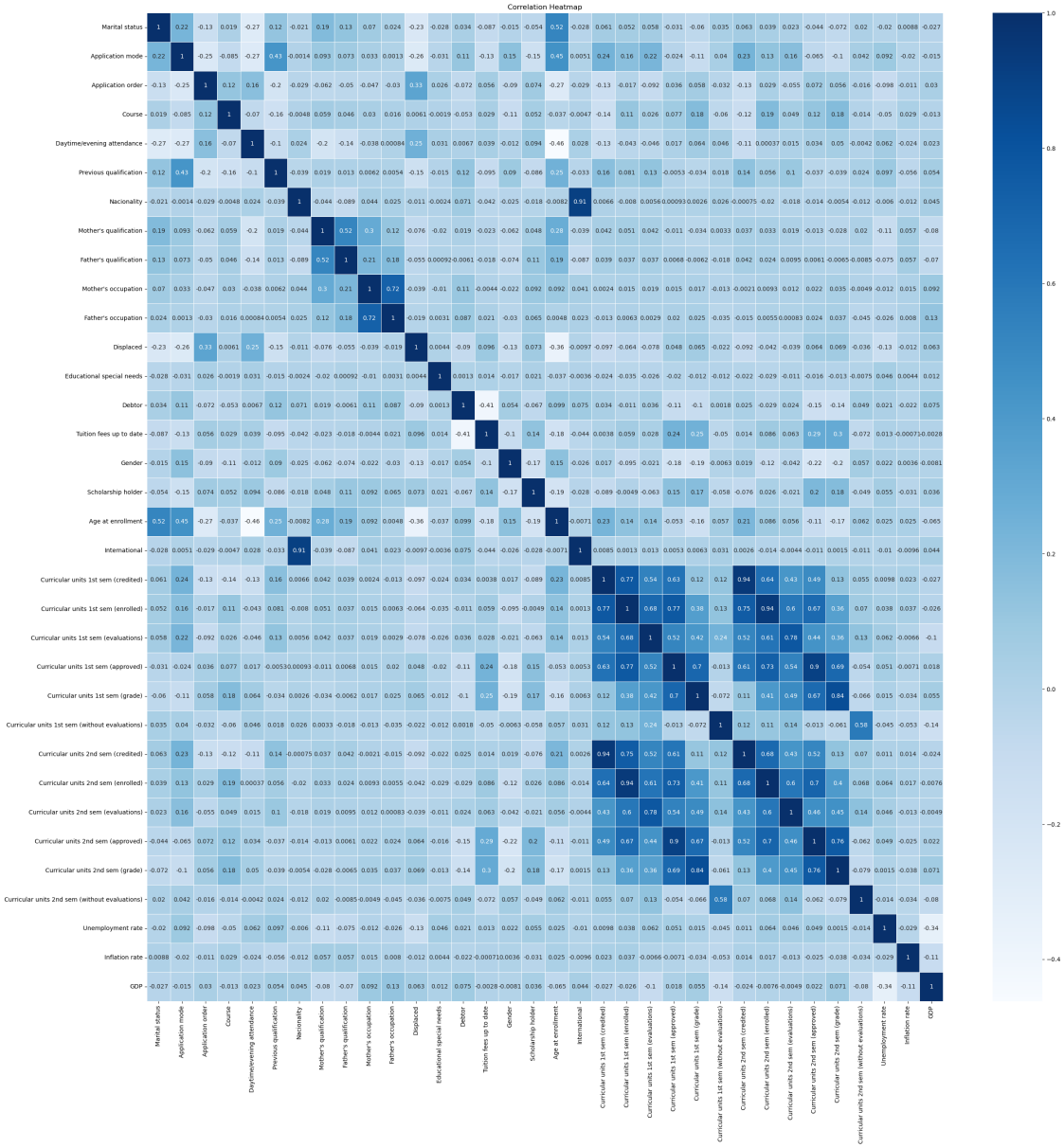


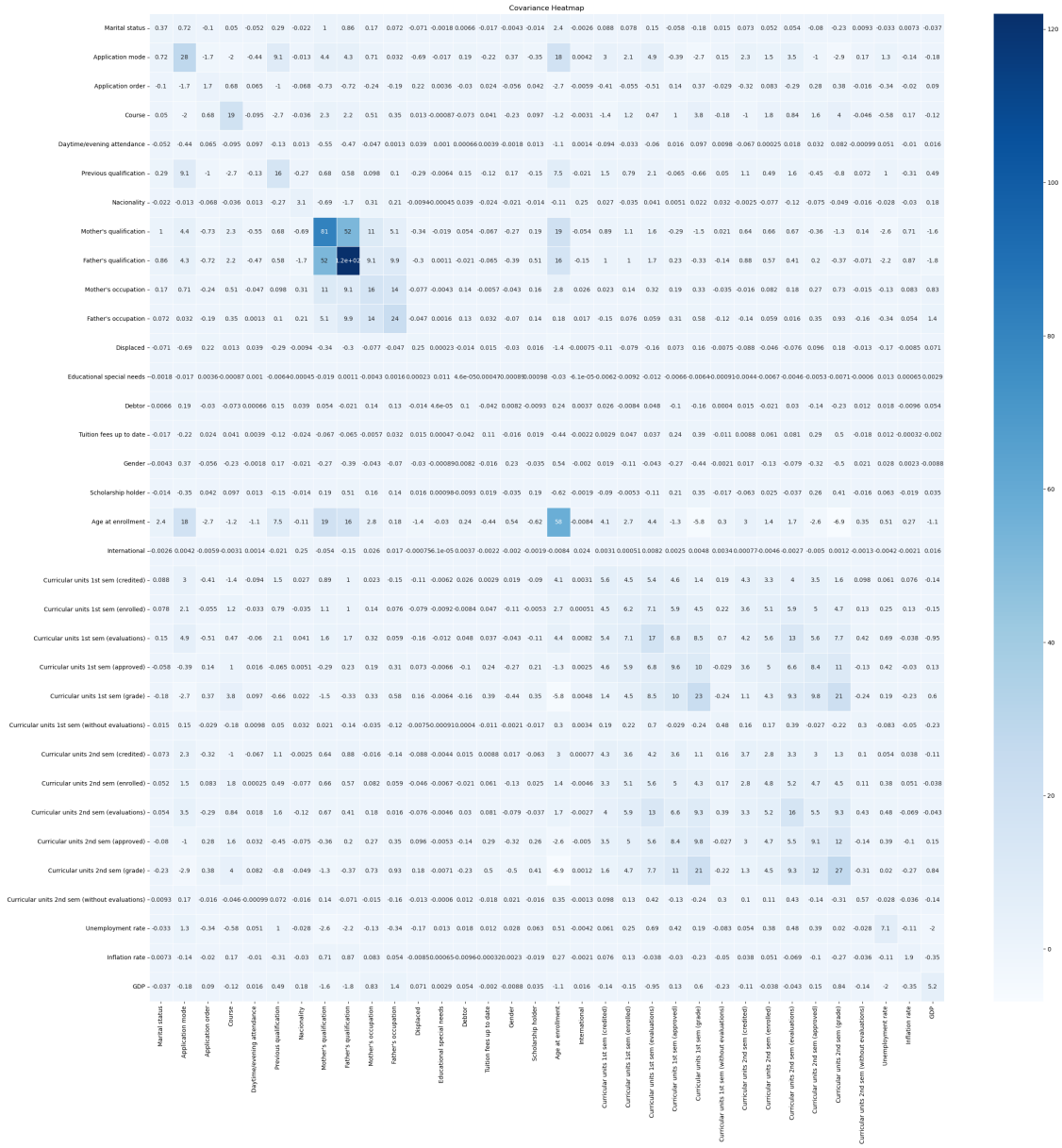
Observations from Visualizations Some interesting relations and observations can be made regarding age at enrollment, daytime/evening attendance, gender, and whether a dropout, graduate, or enrollment outcome is involved.

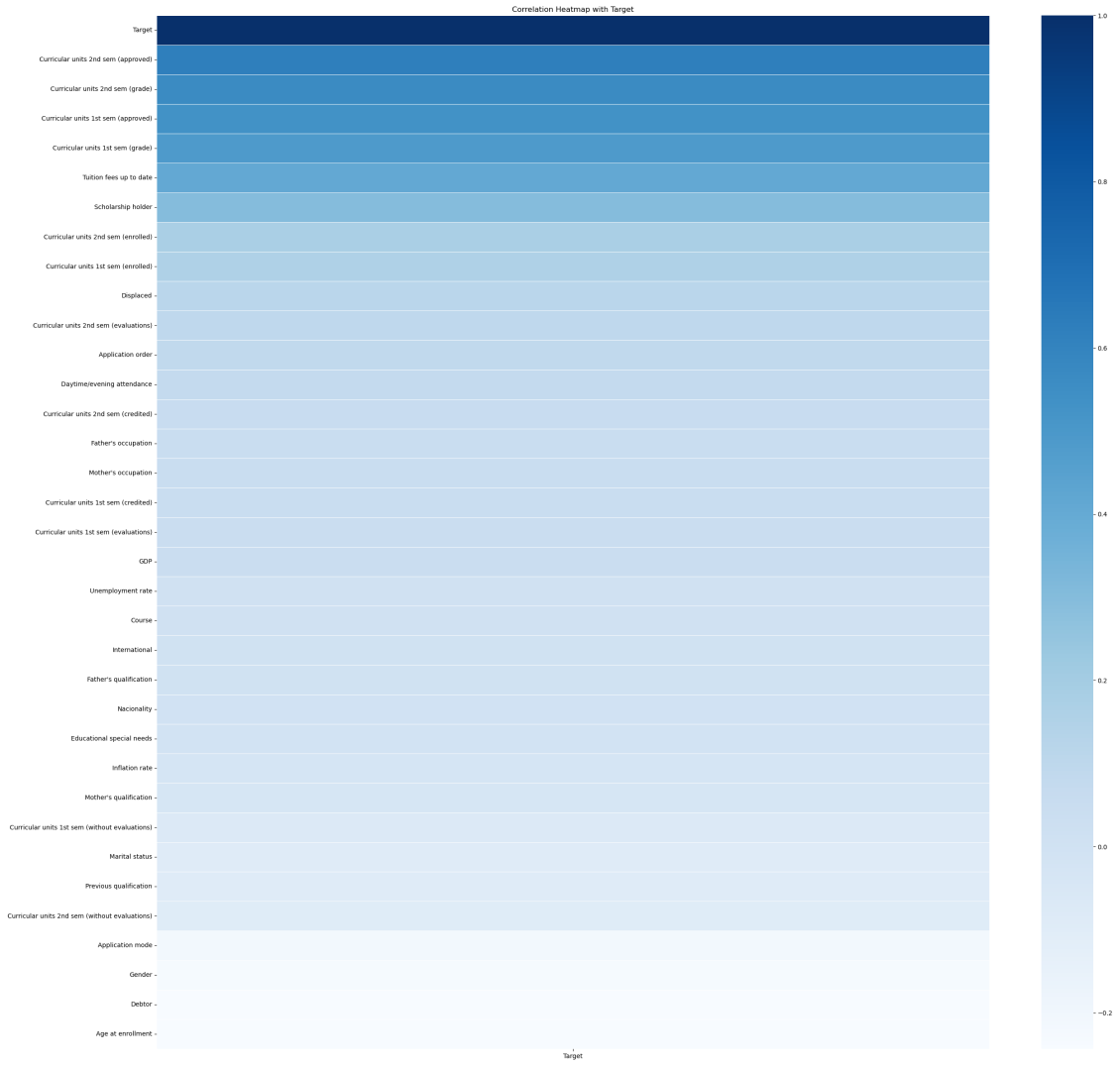
Females make up the majority of students, over 64%, so more females are expected to graduate. But if we look at the entirety of the data, we see that proportionally, more males drop out and fewer males graduate than females.

In the traditional college experience, students attended either in the day or evening. The data shows that older students, typically those 26 and above, have a higher number of students attending in the evening. While this observation doesn't change our model, it's crucial to be aware of as it offers valuable insights into student behavior, keeping us informed and enlightened.

- 1 = daytime, 0 = evening


Correlation Heatmap
Marital status
Application mode
Application order
Course
Daytime/evening attendance
Previous qualification
Nacionality
Mother's qualification
Father's qualification
Mother's occupation
Father's occupation
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
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Covariance Heatmap
Marital status
Application mode
Application order
Course
Daytime/evening attendance
Previous qualification
Nacionality
Mother's qualification
Father's qualification
Mother's occupation
Father's occupation
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
120
100
80
60
40
20
0



Observations from correlations and covariances matrices and heatmaps Recall that one (1) is a perfect positive correlation between features and is indicated by the darker the blue. The lighter the blue, the more the features have a negative correlation.

The covariance matrix supports the idea that mother and father qualifications are positively correlated. This is not surprising, as people tend to marry and date people of the same educational level. We should consider removing one of the pairs. There's also a correlation between the nationality and whether a student is international. There is not a surprise with these being very similiar.

Marital status and Age at enrollment also seem to correlate, which makes sense. People tend to get married older.

And there is also collinearity between the 1st and 2nd-semester features.

Positive:

1) Marital status & Age at enrollment

2) Application order & Course
3) Nacionality & International
4) Mother's qualifications & Father's qualifications
5) Mother's occupation & Father's occupation
6) GDP & Inflation rate.

Negative:

1) age at enrollment & day evening attendance
2) unemployment rate & gdp
3) debtor and tuition fees up to date

0.0.4 Correlation heatmap

According to the correlation heatmap, the most significant features most impacting Target are: Curricular units 2nd sem(grade), Curricular units 2nd sem(approved), Curricular unit 1st sem(grade), Curricular units 1st sem(approved), and Tuition fees up to date.

	Features	VIF
0	Marital status	6.764345
1	Application mode	4.275637
2	Application order	3.286051
3	Course	7.648223
4	Daytime/evening attendance	10.433699
5	Previous qualification	1.844585
6	Nacionality	8.823306
7	Mother's qualification	4.634543
8	Father's qualification	4.678143
9	Mother's occupation	10.380200
10	Father's occupation	7.986871
11	Displaced	2.732772
12	Educational special needs	1.020361
13	Debtor	1.403031
14	Tuition fees up to date	9.999712
15	Gender	1.754392
16	Scholarship holder	1.568309
17	Age at enrollment	19.109473
18	International	6.021480
19	Curricular units 1st sem (credited)	17.317823
20	Curricular units 1st sem (enrolled)	172.380705
21	Curricular units 1st sem (evaluations)	19.683509
22	Curricular units 1st sem (approved)	42.141554
23	Curricular units 1st sem (grade)	28.782252
24	Curricular units 1st sem (without evaluations)	1.779192
25	Curricular units 2nd sem (credited)	13.513508
26	Curricular units 2nd sem (enrolled)	153.388956
27	Curricular units 2nd sem (evaluations)	17.429092
28	Curricular units 2nd sem (approved)	33.025656
29	Curricular units 2nd sem (grade)	26.823782

```
30  Curricular units 2nd sem (without evaluations)    1.651570
31                                Unemployment rate   19.855342
32                                   Inflation rate    1.841532
33                                              GDP    1.253082
```

Reduced features vif:

```
                                          Features        VIF
0                                  Marital status   6.739522
1                                Application mode   4.224294
2                               Application order   3.282680
3                                          Course   7.520452
4                       Daytime/evening attendance  10.238237
5                           Previous qualification   1.833495
6                                      Nacionality   8.803166
7                           Mother's qualification   4.625634
8                           Father's qualification   4.670326
9                              Mother's occupation  10.369576
10                             Father's occupation   7.978196
11                                       Displaced   2.727518
12                       Educational special needs   1.019999
13                                          Debtor   1.398030
14                        Tuition fees up to date   9.985343
15                                          Gender   1.745775
16                             Scholarship holder   1.559058
17                              Age at enrollment  18.724126
18                                  International   6.008329
19              Curricular units 2nd sem (credited)   3.046375
20              Curricular units 2nd sem (enrolled)  34.179802
21          Curricular units 2nd sem (evaluations)  10.388889
22              Curricular units 2nd sem (approved)  19.425847
23                 Curricular units 2nd sem (grade)  18.306633
24  Curricular units 2nd sem (without evaluations)   1.112968
25                                Unemployment rate  19.572348
26                                   Inflation rate   1.820974
27                                              GDP   1.204703
```

Further Reduced features vif:

```
                              Features       VIF
0                              Course  4.382174
1                  Mother's occupation  3.495713
2                              Debtor  1.172203
3                              Gender  1.449146
4                   Scholarship holder  1.480446
5                        International  1.033233
6  Curricular units 2nd sem (approved)  4.616898
7  Curricular units 2nd sem (credited)  1.659619
```

0.0.5 VIF Interpretation (Variance Inflation Factor)

VIF (Variance Inflation Factor) quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. Multicollinearity is when a correlation between multiple independent variables in a multiple regression model can negatively impact the results. A VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity.

We used VIF for the logistics regression model to assist with feature reduction.

A VIF value of 1 indicates no collinearity, values between 1 and 5 suggest moderate collinearity, and values above 5 (sometimes 10) indicate high collinearity, which could be problematic.

We can see that there is a lot of collinearity with the Curricular units. We removed the 1st semester units and re-analyzed the data. The VIF values still went down, but not below 5.

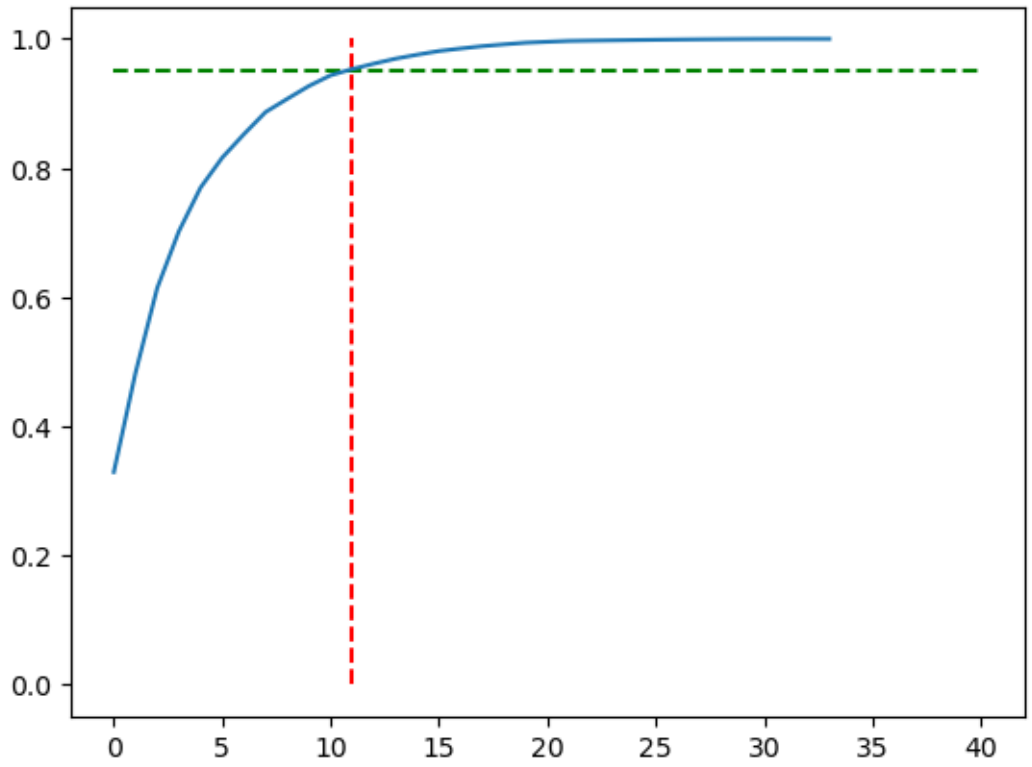
The VIF shows us that we could probably start with these initial variables.

```
Features        VIF
```

0 Course 4.382174 1 Mother's occupation 3.495713 2 Debtor 1.172203 3 Gender 1.449146 4 Scholarship holder 1.480446 5 International 1.033233 6 Curricular units 2nd sem (approved) 4.616898 7 Curricular units 2nd sem (credited) 1.659619

More variables can be added to see how they impact the model accuracy.

```
[<matplotlib.lines.Line2D at 0x2c75ff62120>]
```



0.0.6 PCA Results:

The PCA results indicate that we can reduce the number of features to 11 and retain 95% of the information. We will attempt to reduce the variables more while trying to maintain similary accuracy.

```
Logistic Regression Accuracy, all features:
 0.7559322033898305
              precision    recall  f1-score   support

     Dropout       0.82      0.78      0.80       316
    Enrolled       0.49      0.25      0.33       151
    Graduate       0.76      0.92      0.83       418

    accuracy                           0.76       885
   macro avg       0.69      0.65      0.65       885
weighted avg       0.73      0.76      0.74       885


Confustion Matrix:
 [[248  19  49]
 [ 42  38  71]
 [ 14  21 383]]
```

0.0.7 Logistic Regression Observations - All Features

The logistic regression model using all features has a 75.6% accuracy rate.

The model also performs well (better) for "Dropout" with a precision of 82% and recall of 78%. The model did not perform well for "Enrolled" with a precision of 49% and recall of 25%. The model performs moderately for the "Graduate" class with a precision of 76%, and 92% recall.

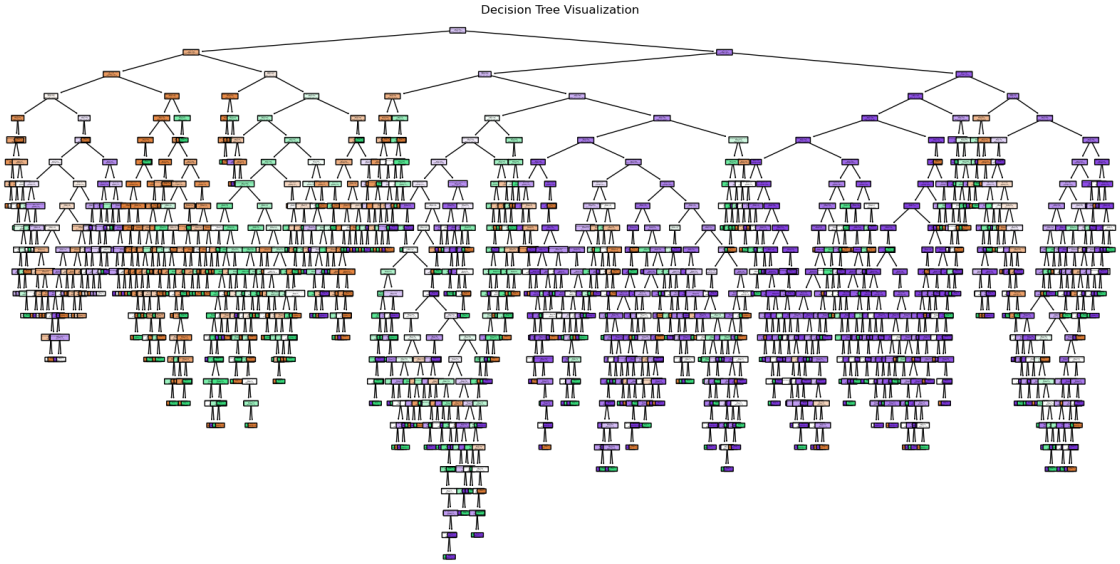
```
Training set size: 3539
Testing set size: 885
Logistics Regression Accuracy, reduced features:
 0.7491525423728813
              precision    recall  f1-score   support

     Dropout       0.81      0.77      0.79       316
    Enrolled       0.48      0.28      0.35       151
    Graduate       0.76      0.91      0.83       418

    accuracy                           0.75       885
   macro avg       0.68      0.65      0.66       885
weighted avg       0.73      0.75      0.73       885


Confustion Matrix:
 [[242  28  46]
 [ 34  42  75]
```

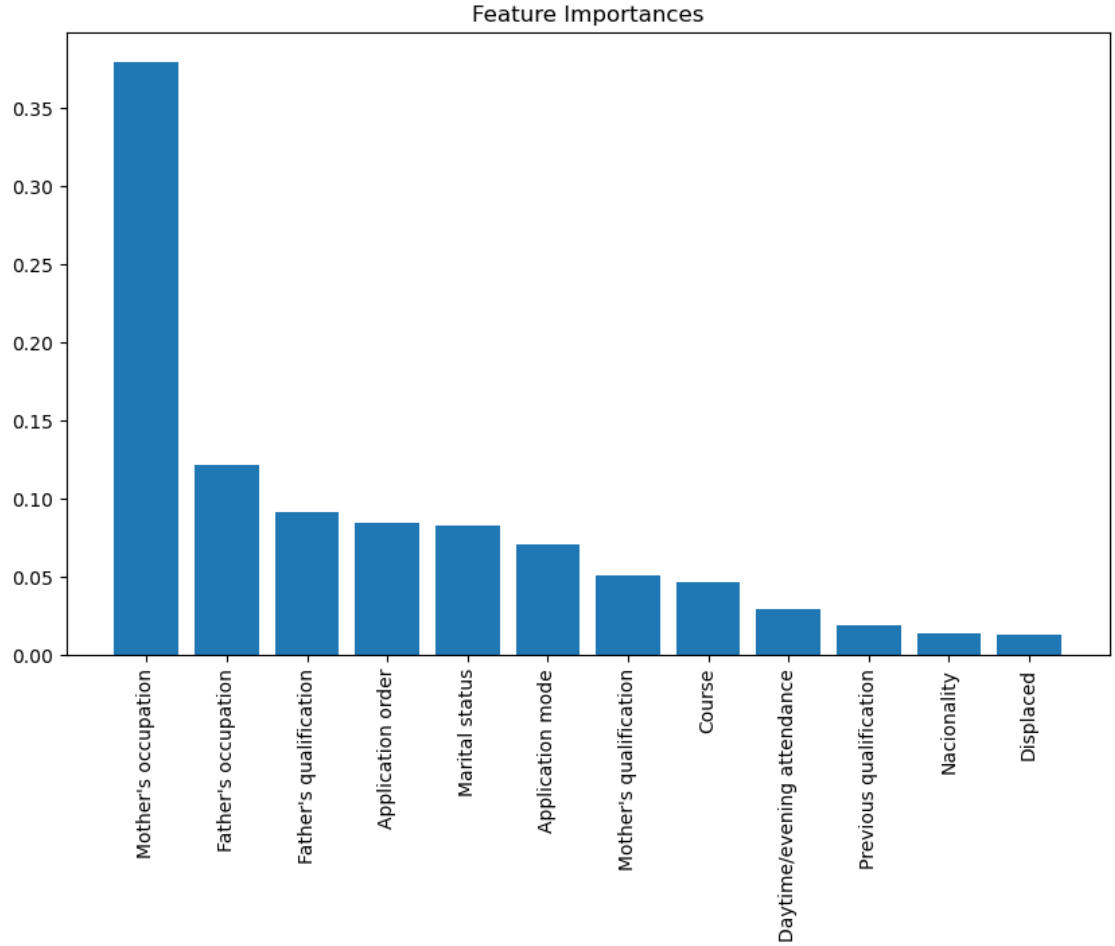
```
 [ 22  17 379]]
```

0.0.8 Logisitics Regression Observations - Reduced Features

The logistic regression model using a subset of the features has a 74.9% accuracy rate. The model also performs well (better) for "Dropout" with a precision of 81% and recall of 77%. The model did not perform well for "Enrolled" with a precision of 48% and recall of 28%. The model performs moderately for the "Graduate" class with a precision of 76%, and 91% recall.

```
Decision Tree Accuracy, all features:  0.6768361581920904
Classification Report
              precision    recall  f1-score   support

     Dropout       0.74      0.65      0.69       316
    Enrolled       0.35      0.38      0.37       151
    Graduate       0.76      0.81      0.78       418

    accuracy                           0.68       885
   macro avg       0.62      0.61      0.61       885
weighted avg       0.68      0.68      0.68       885

Confusion Matrix:
 [[204  60  52]
 [ 38  58  55]
 [ 33  48 337]]

Decision Tree Accuracy, reduced features:  0.6734463276836158
Classification Report
              precision    recall  f1-score   support

     Dropout       0.74      0.67      0.71       316
    Enrolled       0.31      0.37      0.34       151
    Graduate       0.78      0.78      0.78       418

    accuracy                           0.67       885
   macro avg       0.61      0.61      0.61       885
weighted avg       0.69      0.67      0.68       885

Confusion Matrix:
 [[213  60  43]
 [ 46  56  49]
 [ 27  64 327]]
```


Decision Tree Visualization


Feature Importances
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mother's occupation
Father's occupation
Father's qualification
Application order
Marital status
Application mode
Mother's qualification
Course
Daytime/evening attendance
Previous qualification
Nacionality
Displaced

0.0.9 Decision Tree Observations: Reduced Features

The decision tree model with reduced features has an overall accuracy of 67.3% The model also performs well (better) for "Dropout" with a precision of 74% and recall of 67%. The model did not perform well for "Enrolled" with a precision of 31% and recall of 37%. The model performs moderately for the "Graduate" class with a precision of 78%, and 78% recall.

```
Support Vector Machine Accuracy, all features:  0.7288135593220338
Classification Report
              precision    recall  f1-score   support

     Dropout       0.83      0.67      0.74       316
    Enrolled       0.51      0.27      0.35       151
    Graduate       0.71      0.94      0.81       418

    accuracy                           0.73       885
   macro avg       0.68      0.63      0.64       885
weighted avg       0.72      0.73      0.71       885

Confusion Matrix:
 [[212  28  76]
 [ 28  41  82]
 [ 14  12 392]]
```

0.0.10 Support Vector Model Observations: All Features

The SVM model with all features has an overall accuracy of 72.9 The model also performs well (better) for "Dropout" with a precision of 83% and recall of 67%. The model did not perform well for "Enrolled" with a precision of 51% and recall of 27%. The model performs moderately for the "Graduate" class with a precision of 71%, and 94% recall.

```
Support Vector Machine Accuracy, reduced features:  0.7129943502824859
Classification Report
              precision    recall  f1-score   support

     Dropout       0.81      0.66      0.73       316
    Enrolled       0.38      0.23      0.29       151
    Graduate       0.72      0.92      0.81       418

    accuracy                           0.71       885
   macro avg       0.64      0.61      0.61       885
weighted avg       0.70      0.71      0.69       885

Confusion Matrix:
 [[210  44  62]
 [ 30  35  86]
 [ 18  14 386]]
```

0.0.11 Support Vector Model Observations: Reduced Features

The SVM model with all features has an overall accuracy of 71.3 The model also performs well (better) for "Dropout" with a precision of 81% and recall of 66%. The model did not perform well for "Enrolled" with a precision of 38% and recall of 23%. The model performs moderately for the "Graduate" class with a precision of 72%, and 92% recall.

```
Logistic Regression Accuracy, all features: 0.7559322033898305
Logistics Regression Accuracy, reduced features: 0.7491525423728813

Decision Tree Accuracy, all features:  0.6768361581920904
Decision Tree Accuracy, reduced features:  0.6734463276836158

Support Vector Machine Accuracy, all features:  0.7288135593220338
Support Vector Machine Accuracy, reduced features:  0.7129943502824859
```

0.0.12 Summary and Ethics of Analysis

The logistics regression model performed better than the Decision Tree and Support Vector Machine models. We also noticed that reducing the number of features from 35 to 12 provided the same accuracy as having all the features. Since there was collinearity with some features, removing one of the collinear pairs ensures that we measure individual regression. The PCA analysis confirms the feature selection predicting that 11 features can explain 95%.

The following features seemed to have the most impact on predicting whether a student would graduate, enroll, or drop out.

"Age at enrollment" = Age of student at enrollment "Course" = multiple courses listed, which we did not explore "Mother's occupation" = multiple occupations listed, which we did not explore "Tuition fees up to date" = 1 – yes 0 – no "Gender" = 1 – male 0 – female "Scholarship holder" = 1 – yes 0 – no "Curricular units 2nd sem (credited)" = Number of curricular units credited in the 2nd semester "Curricular units 2nd sem (enrolled)" = Number of curricular units enrolled in the 2nd semester "Curricular units 2nd sem (evaluations)" = Number of evaluations to curricular units in the 2nd semester "Curricular units 2nd sem (approved)" = Number of curricular units approved in the 2nd semester "Curricular units 2nd sem (grade)" = Grade average in the 2nd semester (between 0 and 20) "Curricular units 2nd sem (without evaluations)" Number of curricular units without evaluations in the 1st semester

Most notably, when we remove "Curricular units 2nd sem (approved)," the model's accuracy declines by 10 points, possibly indicating how important it is for an advisor to approve courses.

We could summarize that a parent's occupation plays a role in a student's success in college. This could imply that parents who have professional jobs and require a college degree might be better able to support their children, emotionally and financially, through college. But, further analysis is necessary. Additionally, the gender and age of the students seem to play a role. The reasons behind this probably vary widely, and further analysis is needed. Also, there seems to be a financial element in that scholarship holders and whether tuition is paid influence outcomes. Finally, the second-semester courses, as one would expect, influence the outcome. This could imply that involving advisors to help with picking classes could help boost positive outcomes.

There are some ethical considerations in using these data. We must be mindful of the features we

exclude and include for analysis, particularly those around gender, age, race, and other protected status. Before any conclusions are made, careful analytical analysis should be conducted. Making assumptions or determinations based on single data sets or one model could have ethical implications. Finally, when researching protected status data, it would be advised to have representation from those protected groups either on the team or serving in an advisory capacity.

0.0.13 Future Study

There was collinearity between the mother's and father's qualifications and occupation, so the Mother's information was kept in the model. There was also collinearity between nationality and internationality, so International was kept. If we had time, keeping the other pair and re-running the analysis to see if the other feature provided worse or better prediction accuracy would be interesting.

Further exploring data around gender might be useful and interesting in terms of support for genders and looking into what underlying variables might contribute more to a person's education in relation to gender.

It would be interesting to conduct a similar study in which online courses were included and/or substituted for day/evening courses. With students more adept at excelling in online schooling, what variables/features impact graduation, enrollment, and dropout might change in this more modern learning environment.

0.0.14 References

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning, An Introduction to Statistical Learning: with Applications in Python. Springer.

Leskovec, J., Anand Rajaraman, & Jeffrey David Ullman. (2015). Mining of massive datasets. Cambridge University Press.

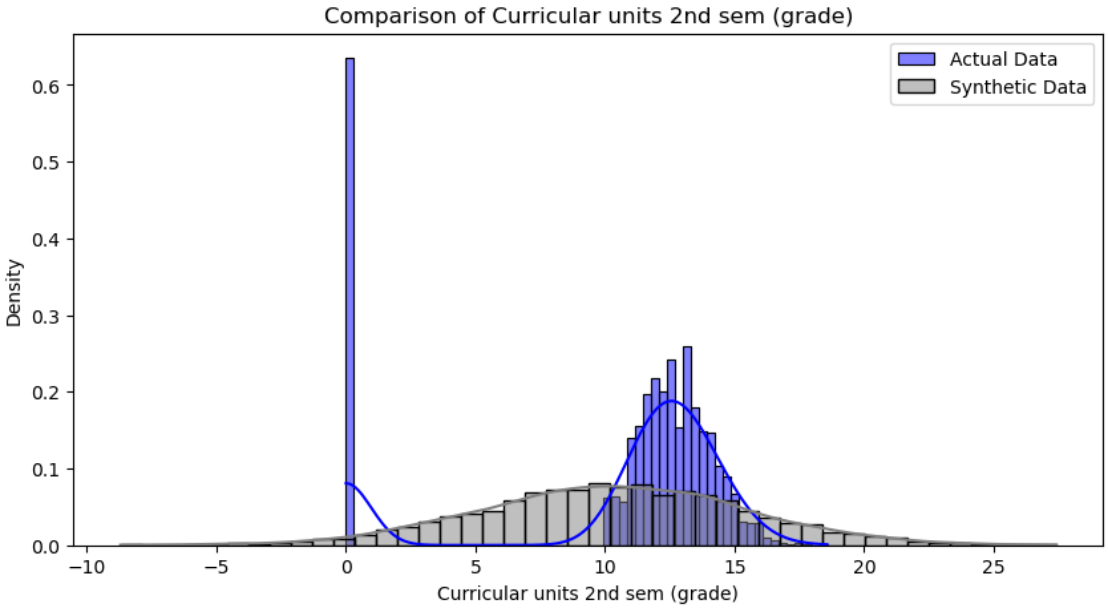
seaborn. (2012). seaborn: statistical data visualization — seaborn 0.9.0 documentation. Pydata.org. https://seaborn.pydata.org/

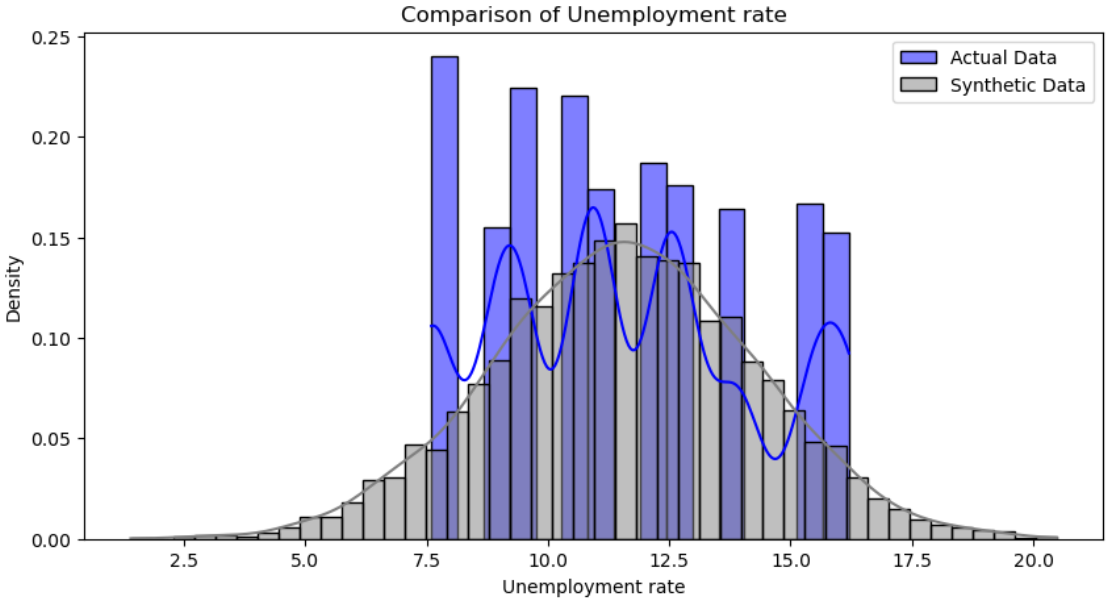
Scikit-learn. (2019). scikit-learn: machine learning in Python. Scikit-Learn.org. https://scikit-learn.org/stable/

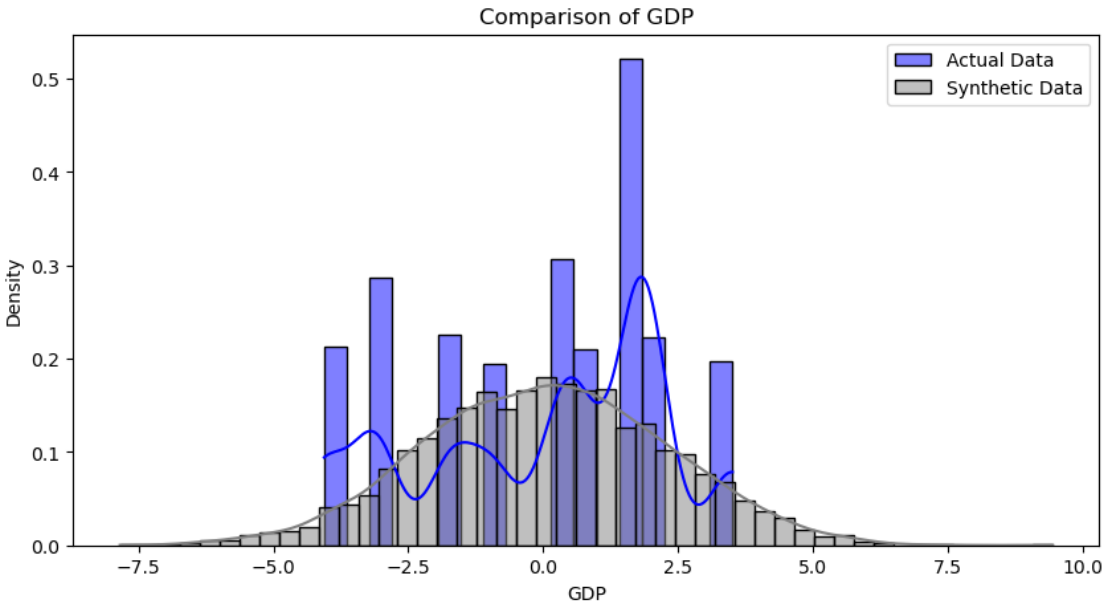
SciPy. (2020). SciPy.org — SciPy.org. Scipy.org. https://scipy.org/

0.0.15 Appendix A

Conduct Multivariate Normal Distribution We will use GDP, Unemployment rate, and inflation rate for comparison with a multivariate normal distribution. The selected columns will be tested for normality and a similarly-sized set of points contructed from a multivariate normal distribution with parameters and Σ matched to those of the chosen columns.


Comparison of Curricular units 2nd sem (grade)
Actual Data
Synthetic Data
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−10
−5
0
5
10
15
20
25
Curricular units 2nd sem (grade)


Comparison of Unemployment rate
Actual Data
Synthetic Data
Density
0.00
0.05
0.10
0.15
0.20
0.25
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Unemployment rate



0.0.16 Observations from Normality Test

The normality tests showed significant deviations from normality for the three columns. The visual comparisons further confirmed that the synthetic data generated from a multivariate normal distribution did not accurately capture the characteristics of the actual data, highlighting the importance of using appropriate distribution assumptions for modeling real-world data.

Curricular Units 2nd Sem (Grade): The actual data shows a right-skewed distribution with a peak about 0. The synthetic data shows a bell-shaped curve and does not show the skewness observed in the actual data.

Unemployment Rate: The actual data displays multiple peaks and appears multimodal. The synthetic data shows a unimodal distribution and does not show the multimodal nature of the actual data.

GDP: The actual data has a more varied distribution with extended tails. The synthetic data shows a more constrained normal distribution and does not show the actual data's variability.

0.0.17 Appendix B

Column Names and Descriptions

1. Marital status: The marital status of the student. (Categorical)
2. Application mode: The method of application used by the student. (Categorical)
3. Application order: The order in which the student applied. (Numerical)
4. Course: The course taken by the student. (Categorical)
5. Daytime/evening attendance: Whether the student attends classes during the day or in the evening. (Categorical)
6. Previous qualification: The qualification obtained by the student before enrolling in higher education. (Categorical)

7. Nacionality: The nationality of the student. (Categorical)
8. Mother's qualification: The qualification of the student's mother. (Categorical)
9. Father's qualification: The qualification of the student's father. (Categorical)
10. Mother's occupation: The occupation of the student's mother. (Categorical)
11. Father's occupation: The occupation of the student's father. (Categorical)
12. Displaced: Whether the student is a displaced person. (Categorical)
13. Educational special needs: Whether the student has any special educational needs. (Categorical)
14. Debtor: Whether the student is a debtor. (Categorical)
15. Tuition fees up to date: Whether the student's tuition fees are up to date. (Categorical)
16. Gender: The gender of the student. (Categorical)
17. Scholarship holder: Whether the student is a scholarship holder. (Categorical)
18. Age at enrollment: The age of the student at the time of enrollment. (Numerical)
19. International: Whether the student is an international student. (Categorical)
20. Curricular units 1st sem (credited): The number of curricular units credited by the student in the first semester. (Numerical)
21. Curricular units 1st sem (enrolled): The number of curricular units enrolled by the student in the first semester. (Numerical)
22. Curricular units 1st sem (evaluations): The number of curricular units evaluated by the student in the first semester. (Numerical)
23. Curricular units 1st sem (approved): The number of curricular units approved by the student in the first semester. (Numerical)
24. Curricular units 1st sem (grade): Grade average in the 2nd semester (between 0 and 20)
25. Curricular units 1st sem (without evaluations):
26. Curricular units 2nd sem (credited): The number of curricular units credited by the student in the second semester. (Numerical)
27. Curricular units 2nd sem (enrolled): The number of curricular units enrolled by the student in the second semester. (Numerical)
28. Curricular units 2nd sem (evaluations): The number of curricular units evaluated by the student in the second semester. (Numerical)
29. Curricular units 2nd sem (approved): The number of curricular units approved by the student in the second semester. (Numerical)
30. Curricular units 2nd sem (grade):Grade average in the 2nd semester (between 0 and 20)
31. Curricular units 2nd sem (without evaluations):
32. Unemployment rate: Unemployment rate (%)
33. Inflation rate: Inflation rate %
34. GDP: Gross Domestic Product
35. Target: The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course.