CSCE 5320.002

Group 6 – Blessy Kuriakose
May 8, 2023

Project Final

Topic: “Predict students' dropout and academic success”

Team Members: Blessy Kuriakose

Link to Code: https://colab.research.google.com/drive/1Jp3tpbkfYz3n1UHeLslYDA3O1dJVsCXP?usp=share_link

Part I

The dataset used for this topic was accessed via *Kaggle* [1]. The data was originally published on *Zenodo* [2] and is noted to contain information on undergraduates of differing degrees from an unspecified higher education institution, at the end of the normal duration of their coursework. It is presented as a three-class prediction problem with 34 features, both categorical and numerical, and a target column that shows whether the student has dropped out, is still enrolled, or has graduated.

Features of the dataset include:

Marital status: “The marital status of the student. (Categorical)”

Application mode: “The method of application used by the student. (Categorical)”

Application order: “The order in which the student applied. (Numerical)”

Course: “The course taken by the student. (Categorical)”

Daytime/evening attendance: “Whether the student attends classes during the day or in the evening. (Categorical)”

Previous qualification: “The qualification obtained by the student before enrolling in higher education. (Categorical)”

Nacionality: “The nationality of the student. (Categorical)”

Mother's qualification: “The qualification of the student's mother. (Categorical)”

Father's qualification: “The qualification of the student's father. (Categorical)”

Mother's occupation: “The occupation of the student's mother. (Categorical)”

Father's occupation: “The occupation of the student's father. (Categorical)”

Displaced: “Whether the student is a displaced person. (Categorical)”

Educational special needs: “Whether the student has any special educational needs. (Categorical)”

Debtor: “Whether the student is a debtor. (Categorical)”

Tuition fees up to date: “Whether the student's tuition fees are up to date. (Categorical)”

Gender: “The gender of the student. (Categorical)”

Scholarship holder: “Whether the student is a scholarship holder. (Categorical)”

Age at enrollment: “The age of the student at the time of enrollment. (Numerical)”

International: “Whether the student is an international student. (Categorical)”

Curricular units 1st sem (credited): “The number of curricular units credited by the student in the first semester. (Numerical)”

Curricular units 1st sem (enrolled): “The number of curricular units enrolled by the student in the first semester. (Numerical)”

Curricular units 1st sem (evaluations): “The number of curricular units evaluated by the student in the first semester. (Numerical)”

Curricular units 1st sem (approved): “The number of curricular units approved by the student in the first semester. (Numerical)”

Curricular units 1st sem (grade): (Numerical)

Curricular units 1st sem (without evaluations): The number of curricular units unevaluated by the student in the first semester. (Numerical)

Curricular units 2nd sem (credited): The number of curricular units credited by the student in the second semester. (Numerical)

Curricular units 2nd sem (enrolled): The number of curricular units enrolled by the student in the second semester. (Numerical)

Curricular units 2nd sem (evaluations): The number of curricular units evaluated by the student in the second semester. (Numerical)

Curricular units 2nd sem (approved): The number of curricular units approved by the student in the second semester. (Numerical)

Curricular units 2nd sem (grade): (Numerical)

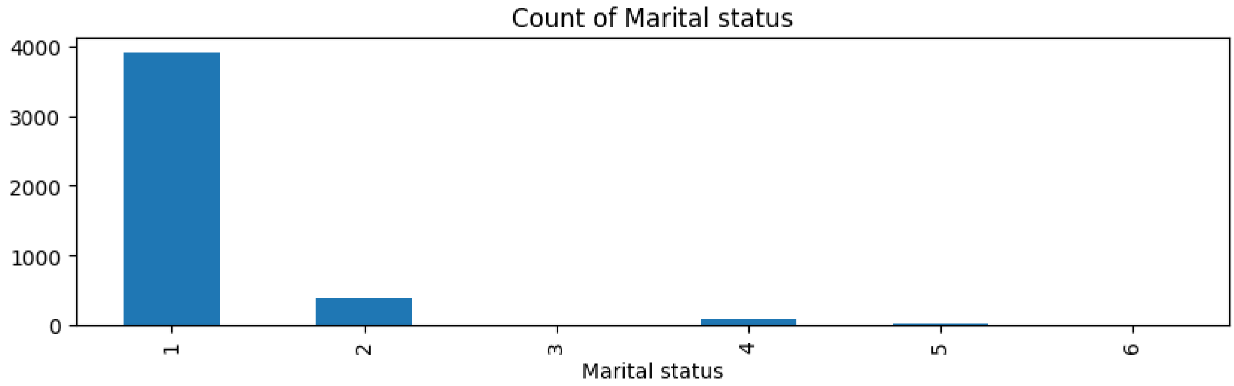
Curricular units 2nd sem (without evaluations): The number of curricular units unevaluated by the student in the second semester. (Numerical)

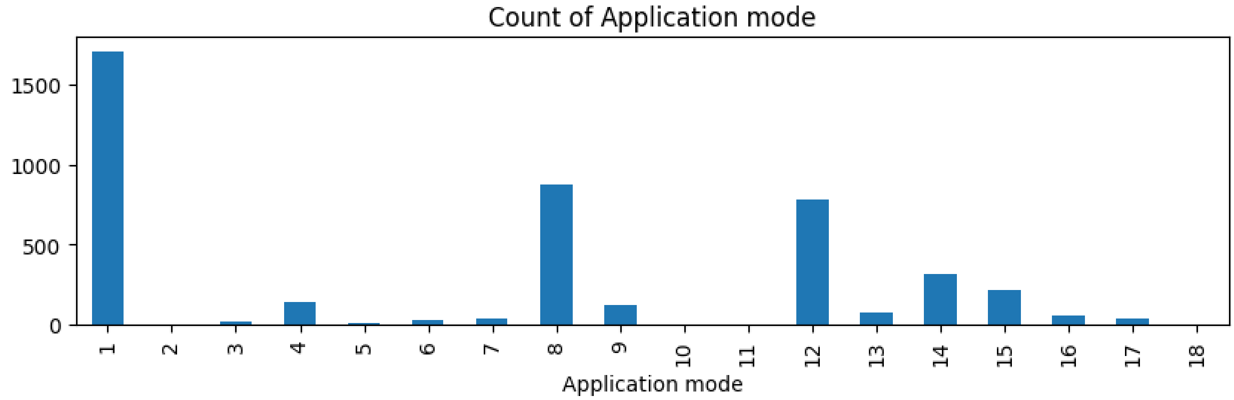
Unemployment rate: (Numerical)

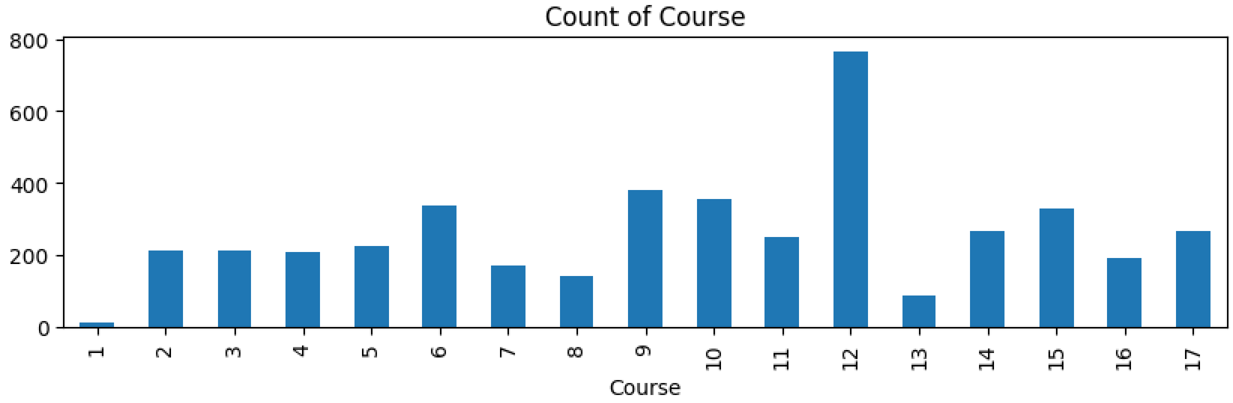
Inflation rate: (Numerical)

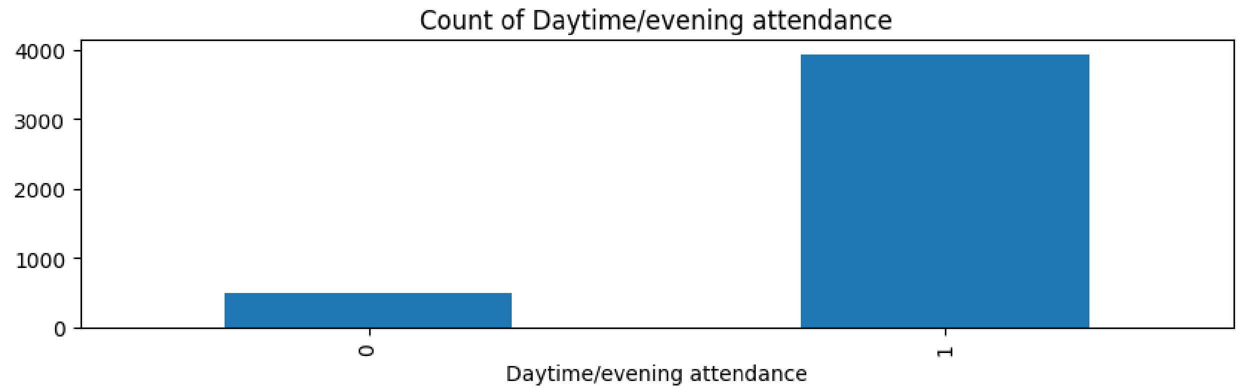
GDP: gross domestic product (Numerical)

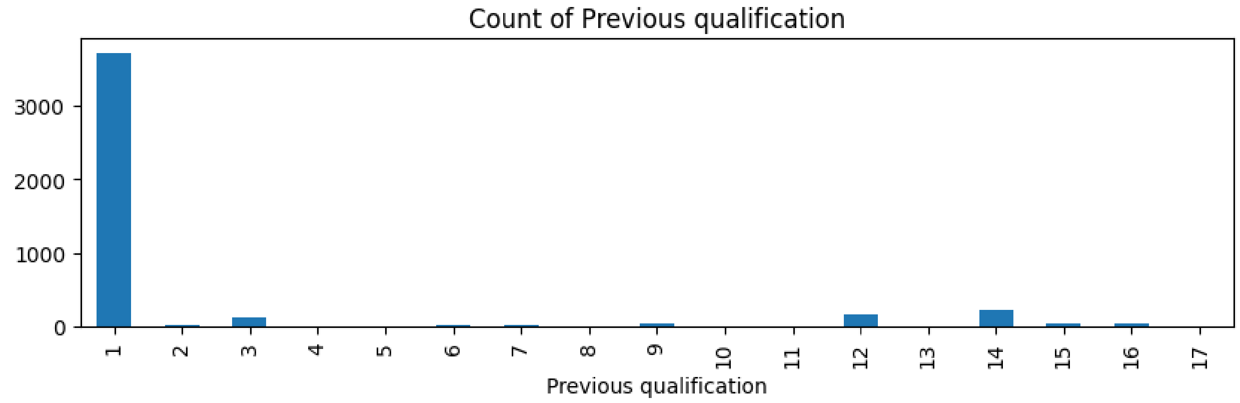
Histograms for the categorical features are shown in the figures below:

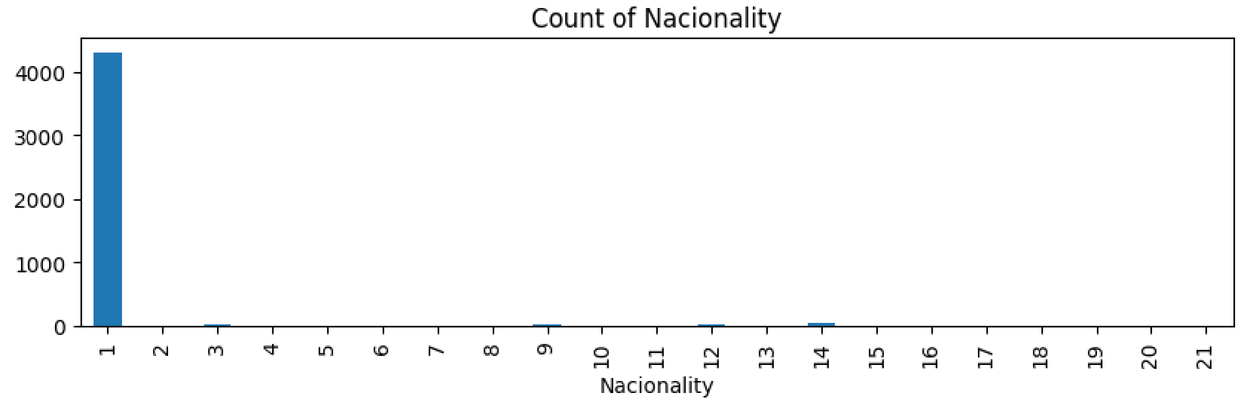


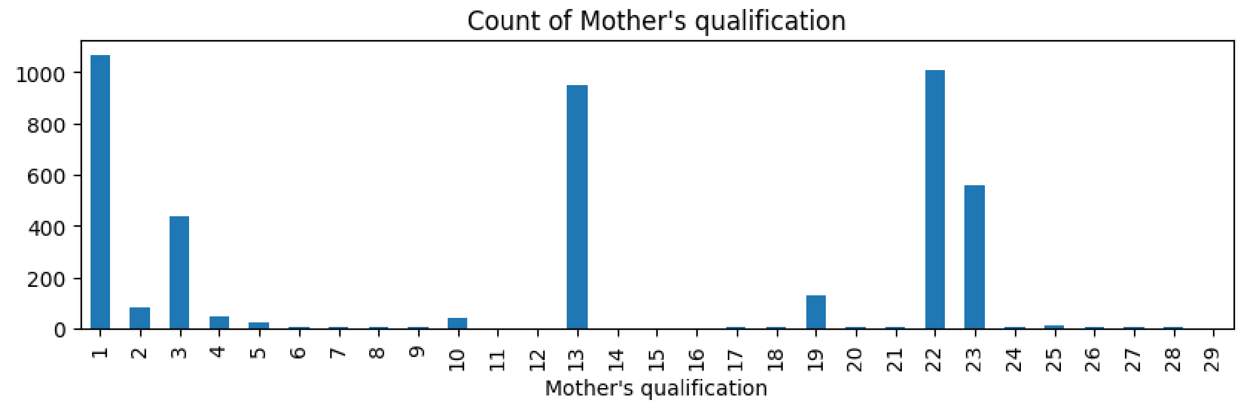


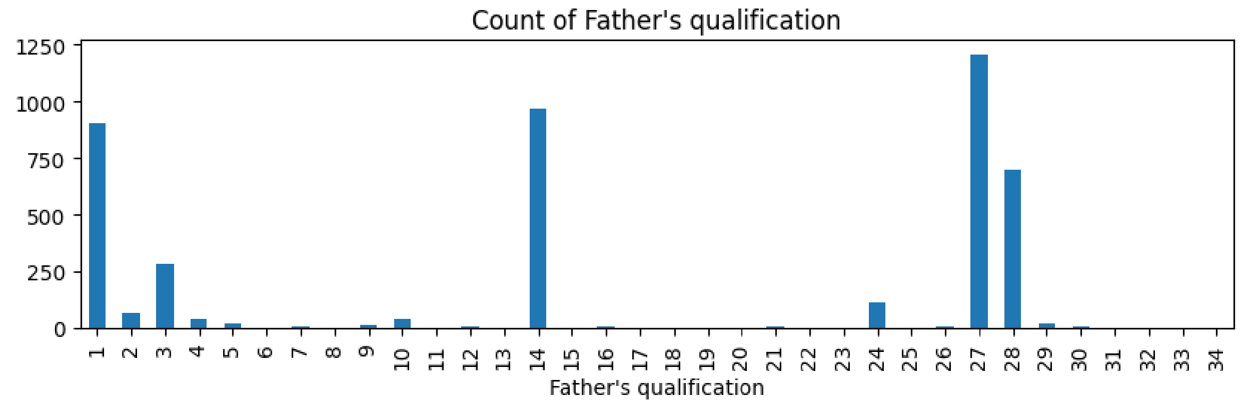


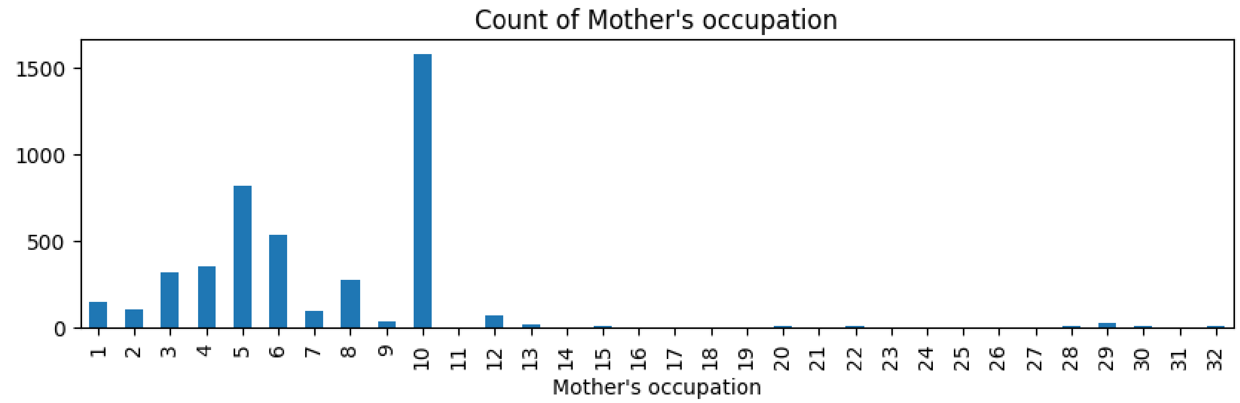

Count of Daytime/evening attendance
4000
3000
2000
1000
0
0
1
Daytime/evening attendance

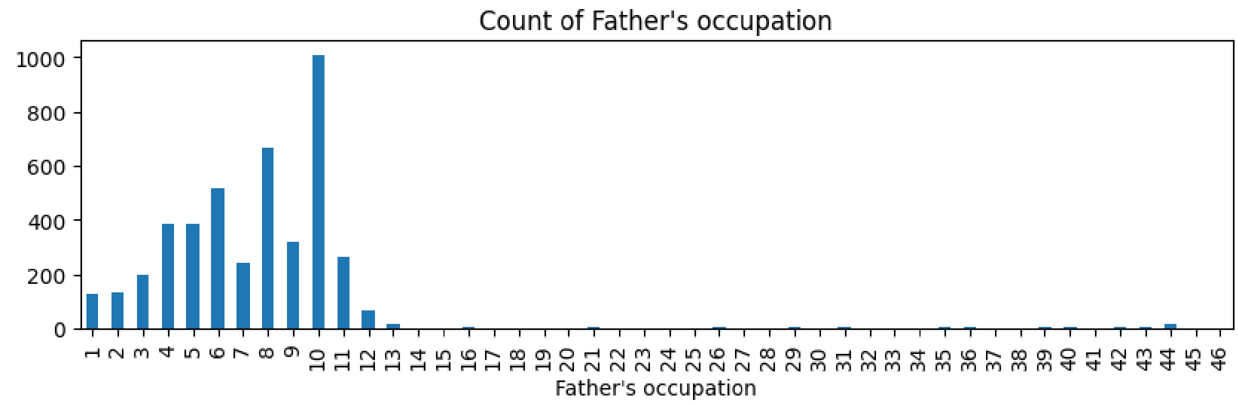

Count of Previous qualification
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Previous qualification

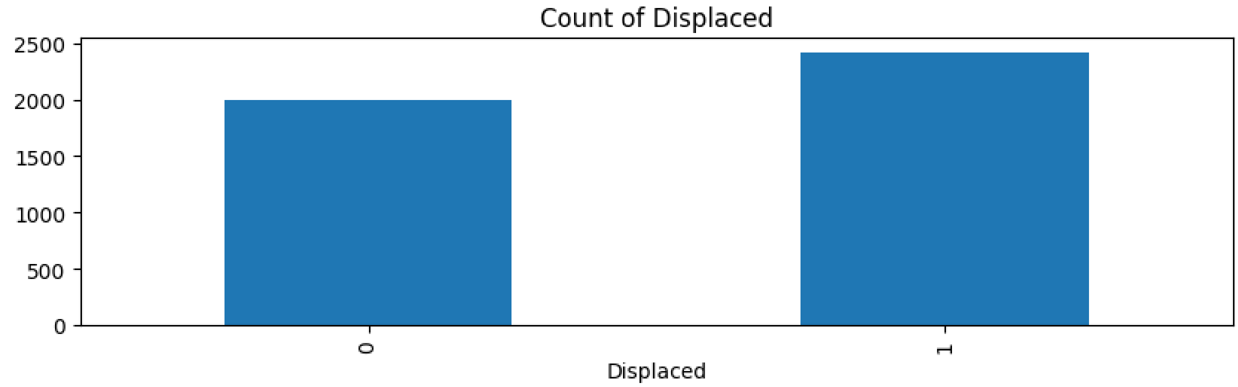

Count of Nacionality
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Nacionality

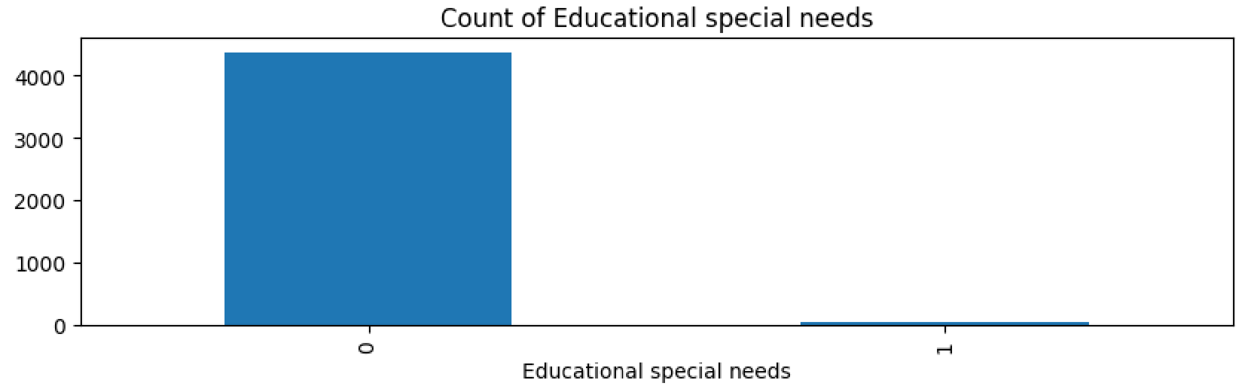

Count of Mother's qualification
0
200
400
600
800
1000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Mother's qualification

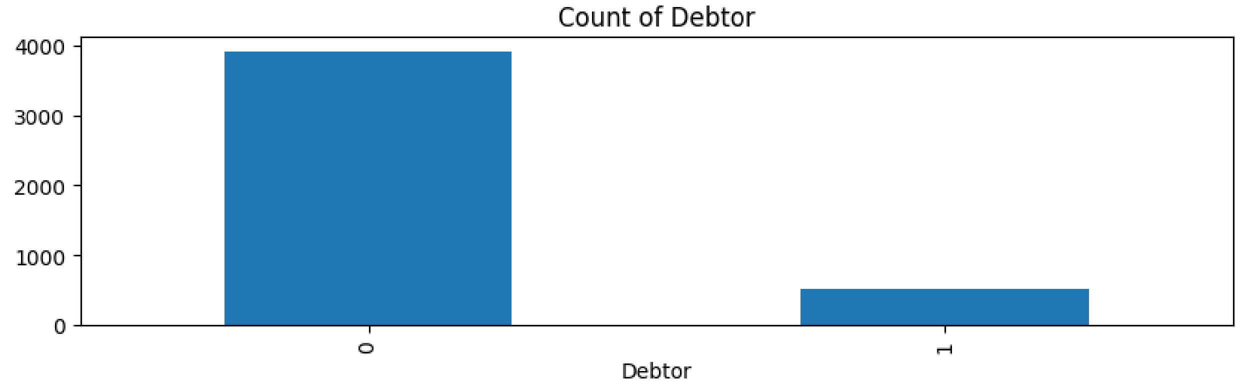

Count of Father's qualification
0
250
500
750
1000
1250
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Father's qualification

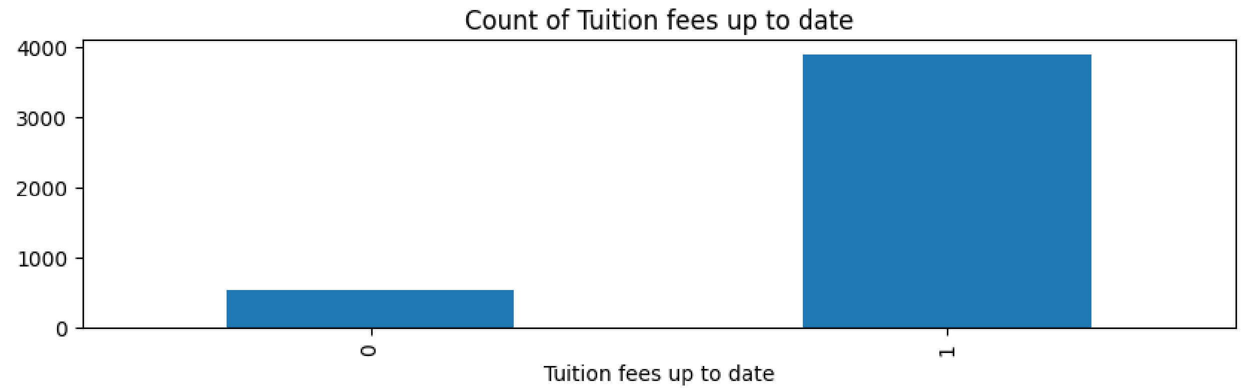

Count of Mother's occupation
0
500
1000
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Mother's occupation

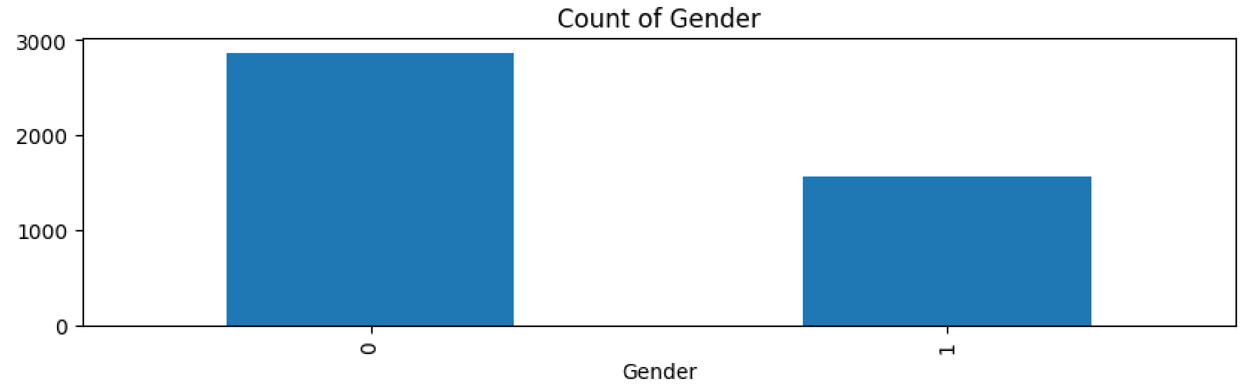

Count of Father's occupation
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Father's occupation

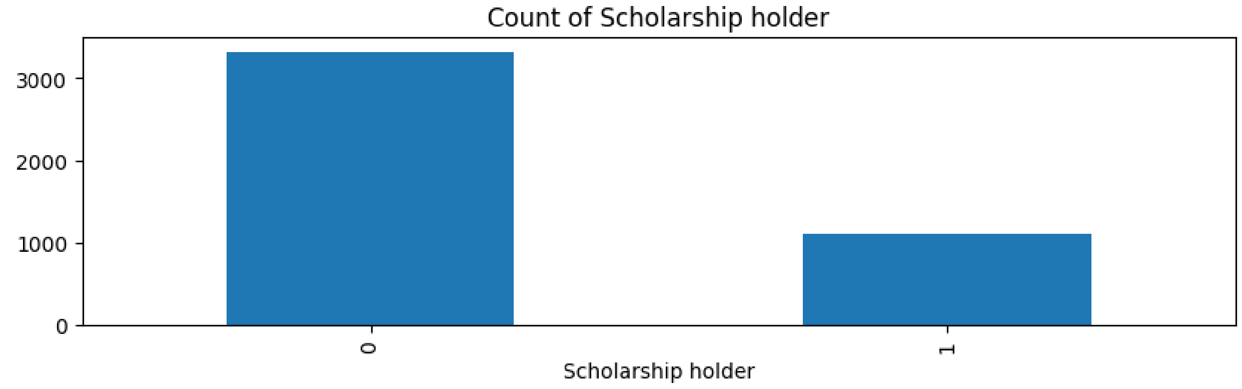

Count of Displaced
2500
2000
1500
1000
500
0
0
1
Displaced

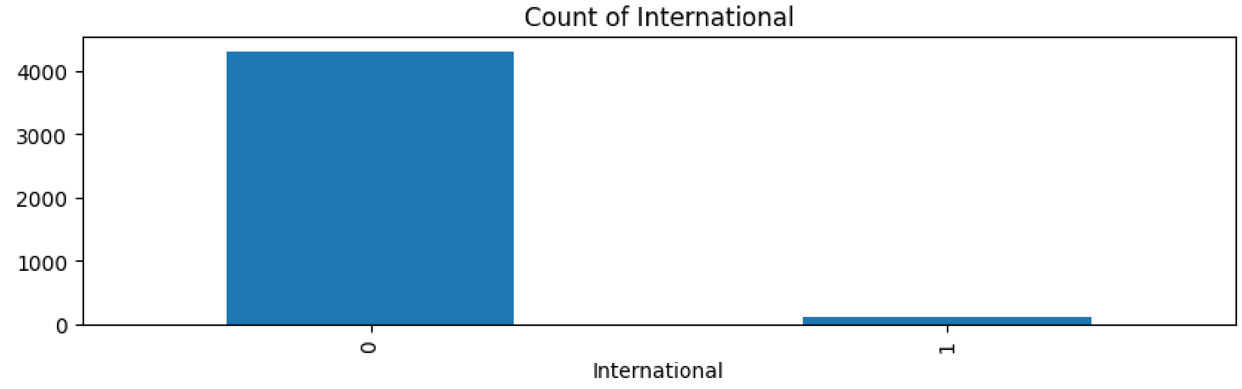

Count of Educational special needs
4000
3000
2000
1000
0
0
1
Educational special needs


Count of Debtor
4000
3000
2000
1000
0
0
1
Debtor


Count of Tuition fees up to date
4000
3000
2000
1000
0
0
1
Tuition fees up to date

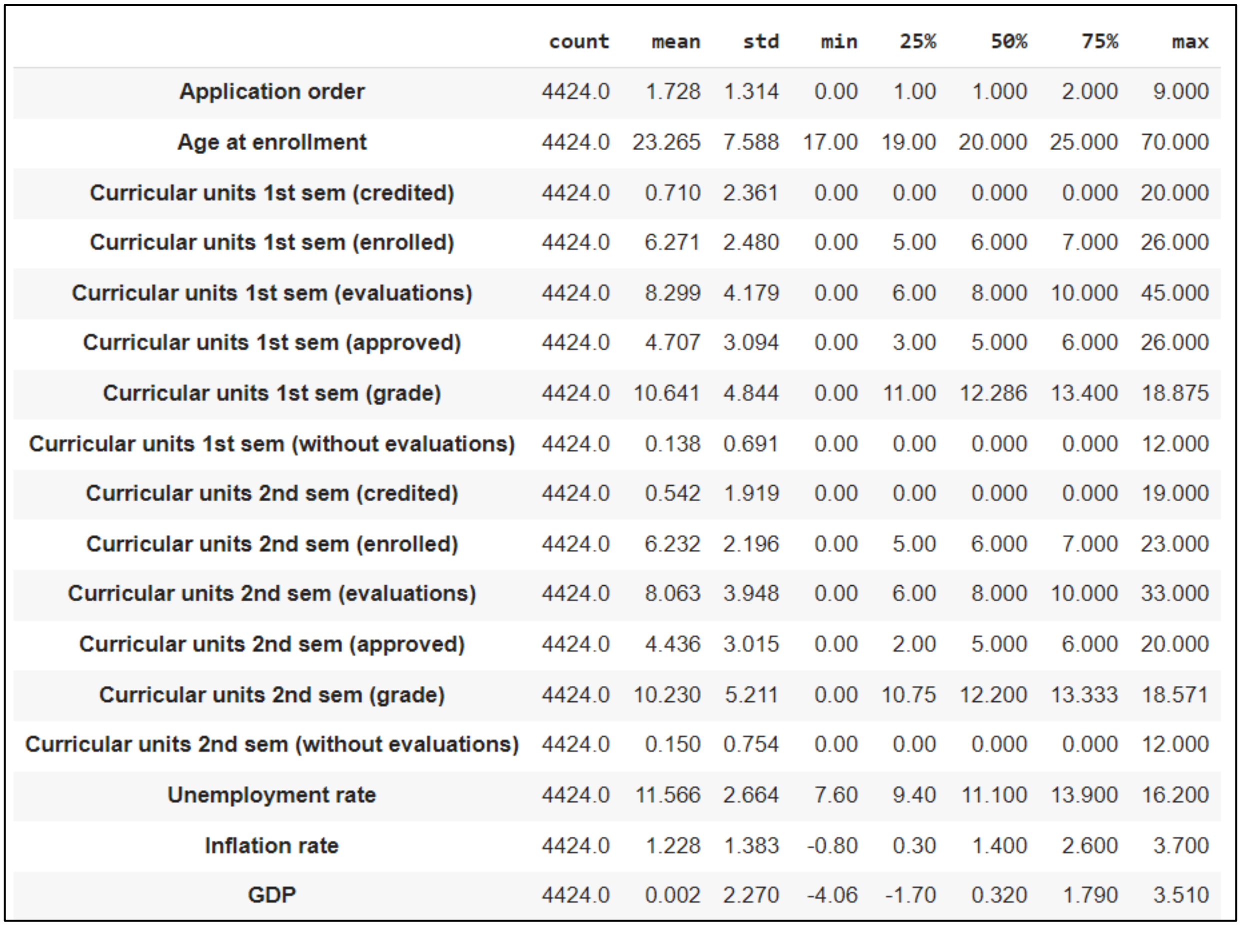

Count of Gender
3000
2000
1000
0
0
1
Gender


Count of Scholarship holder
3000
2000
1000
0
0
1
Scholarship holder


Count of International
4000
3000
2000
1000
0
0
1
International

No information was found regarding what the categories for each of these features are. The data comes pre-encoded as integers. There are several blaring imbalances visible in the histograms. Features containing many categories show strong imbalances as well.

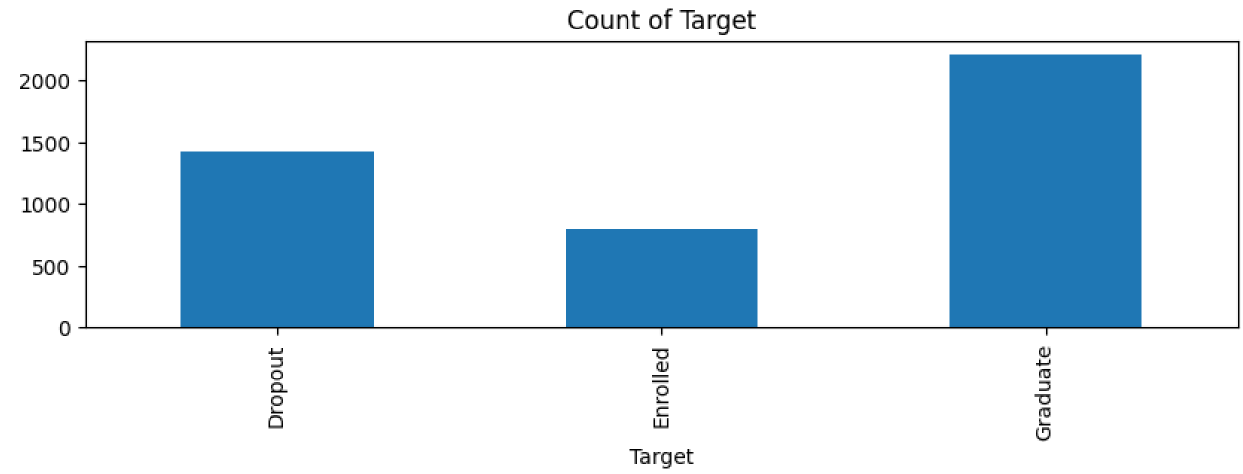
Exploratory statistics for the numerical features can be found in the table below:



	count	mean	std	min	25%	50%	75%	max
Application order	4424.0	1.728	1.314	0.00	1.00	1.000	2.000	9.000
Age at enrollment	4424.0	23.265	7.588	17.00	19.00	20.000	25.000	70.000
Curricular units 1st sem (credited)	4424.0	0.710	2.361	0.00	0.00	0.000	0.000	20.000
Curricular units 1st sem (enrolled)	4424.0	6.271	2.480	0.00	5.00	6.000	7.000	26.000
Curricular units 1st sem (evaluations)	4424.0	8.299	4.179	0.00	6.00	8.000	10.000	45.000
Curricular units 1st sem (approved)	4424.0	4.707	3.094	0.00	3.00	5.000	6.000	26.000
Curricular units 1st sem (grade)	4424.0	10.641	4.844	0.00	11.00	12.286	13.400	18.875
Curricular units 1st sem (without evaluations)	4424.0	0.138	0.691	0.00	0.00	0.000	0.000	12.000
Curricular units 2nd sem (credited)	4424.0	0.542	1.919	0.00	0.00	0.000	0.000	19.000
Curricular units 2nd sem (enrolled)	4424.0	6.232	2.196	0.00	5.00	6.000	7.000	23.000
Curricular units 2nd sem (evaluations)	4424.0	8.063	3.948	0.00	6.00	8.000	10.000	33.000
Curricular units 2nd sem (approved)	4424.0	4.436	3.015	0.00	2.00	5.000	6.000	20.000
Curricular units 2nd sem (grade)	4424.0	10.230	5.211	0.00	10.75	12.200	13.333	18.571
Curricular units 2nd sem (without evaluations)	4424.0	0.150	0.754	0.00	0.00	0.000	0.000	12.000
Unemployment rate	4424.0	11.566	2.664	7.60	9.40	11.100	13.900	16.200
Inflation rate	4424.0	1.228	1.383	-0.80	0.30	1.400	2.600	3.700
GDP	4424.0	0.002	2.270	-4.06	-1.70	0.320	1.790	3.510

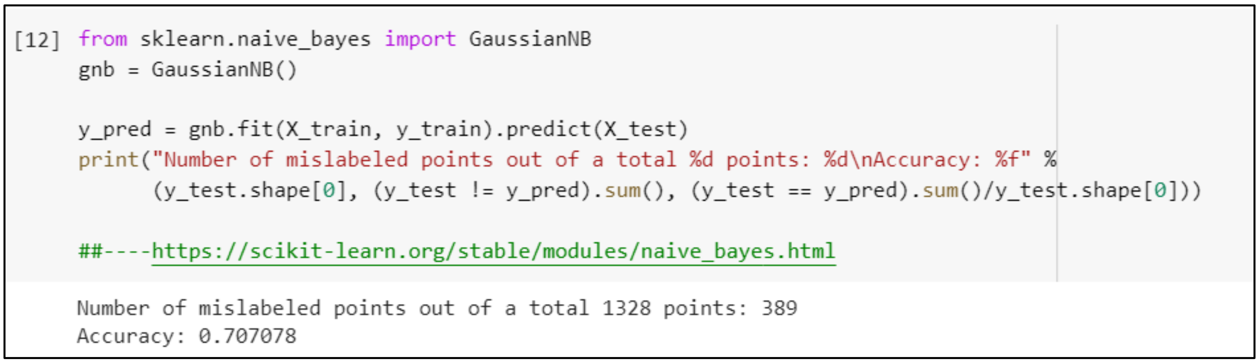
The “curricular units” features may need to be further looked in to, particularly the types that are “credited” or “without evaluations”; the quartiles are showing 0, but the max is 12.

We aim to use the features to predict the column named ***Target*** which tells you if the individual has Dropped out from (*‘Dropout’*), is still enrolled in(*‘Enrolled’*), or has graduated from (*‘Graduate’*). Below is a histogram showing the spread of the data between these categories.



The following figures show screenshots of very basic models trained with 70% of the dataset and tested with the remaining 30%:

Naïve Bayes:



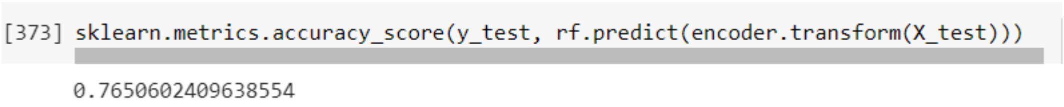
```
[12] from sklearn.naive_bayes import GaussianNB
     gnb = GaussianNB()

     y_pred = gnb.fit(X_train, y_train).predict(X_test)
     print("Number of mislabeled points out of a total %d points: %d\nAccuracy: %f" %
           (y_test.shape[0], (y_test != y_pred).sum(), (y_test == y_pred).sum()/y_test.shape[0]))

     ##----https://scikit-learn.org/stable/modules/naive_bayes.html

     Number of mislabeled points out of a total 1328 points: 389
     Accuracy: 0.707078
```

Random Forest:

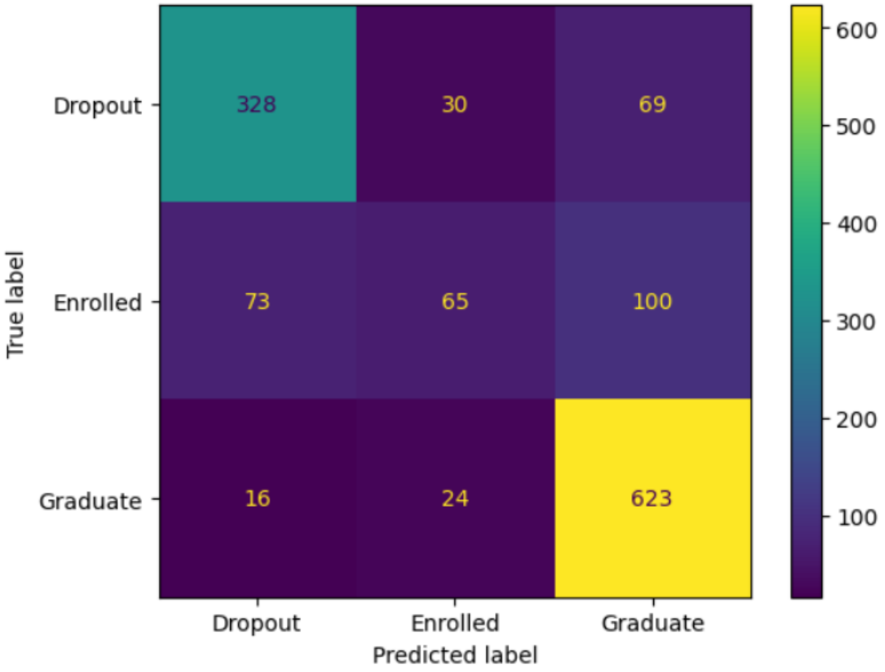


```
[373] sklearn.metrics.accuracy_score(y_test, rf.predict(encoder.transform(X_test)))

      0.7650602409638554
```

Part II

Here is where we left off in Part I with the RF model, which is the model type we will be continuing with for the remainder of the project:

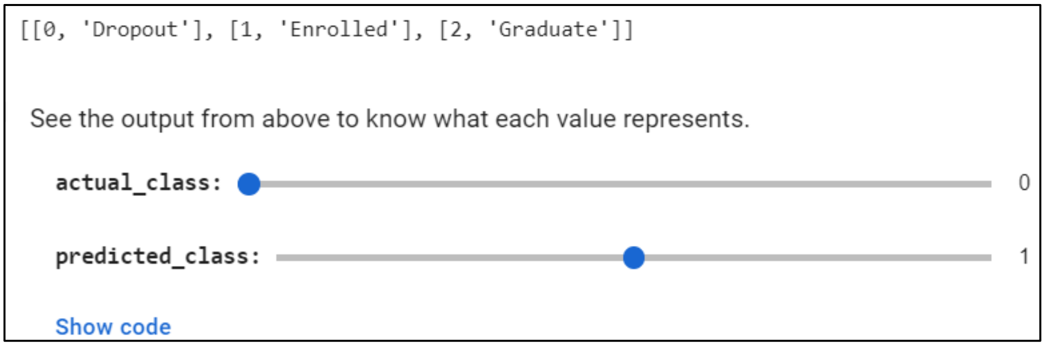


As we can see, the model has an extremely hard time predicting the “Enrolled” class – in fact, most of the time, the prediction for “Enrolled” is “Graduate”. It is very possible that the class imbalance is influencing this outcome, since the “Graduate” class is greater in quantity within the dataset. At another level, this disparity is not surprising since “Enrolled” is a more ambiguous state compared to “Graduate” or “Dropout”. It is a state which has not yet ended in a conclusion of one or the other (referring to “Graduate” or “Dropout”).

It can also be observed that the model seems to predict “Graduate” at a higher likelihood than “Dropout” which *may*, again, be due to the higher quantity of data corresponding to this class. This favoritism could be occurring due to brute probability, where the outcome is always more likely to be the class of higher quantity due to disproportional outcomes in the training set, or better distinguishment, whereby the model ‘knows’ this class better by having more data of this class’s type to reference, or train on. (My guess is the former. A balanced set can be tried to confirm.)

From here, we explore further into the Random Forest (RF) model, using LIME analysis. The graphs for the lime analysis are partially presented within a table shown further below. A link to the code file can be found here or at the top of this report. Once opened, the analyzed graphs (including those not presented in the table) can be viewed by a simple run.

Below is a screenshot of the form that can be used to run a LIME explanation for a sample of each of the cases present in the confusion matrix above.



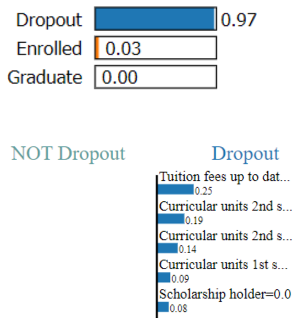
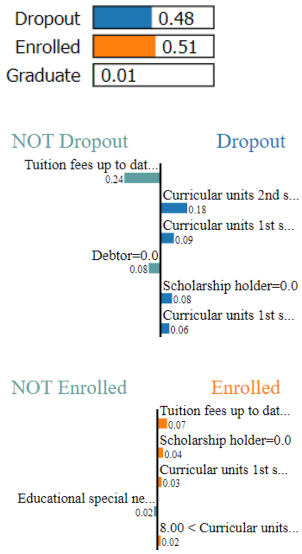
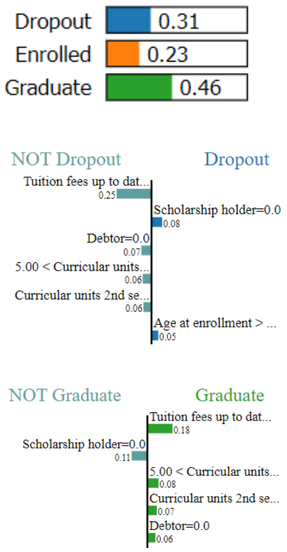
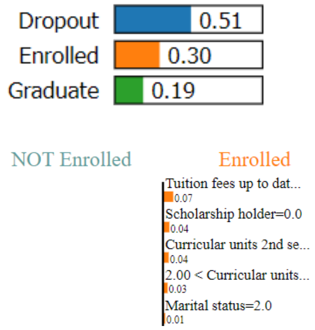
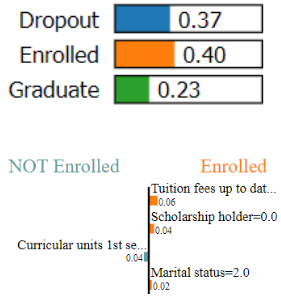
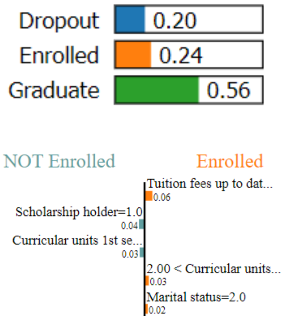
```
[[0, 'Dropout'], [1, 'Enrolled'], [2, 'Graduate']]
```

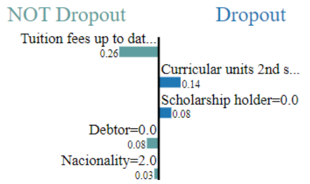
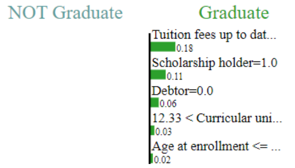
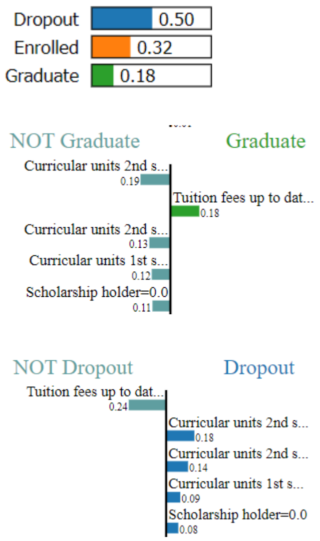
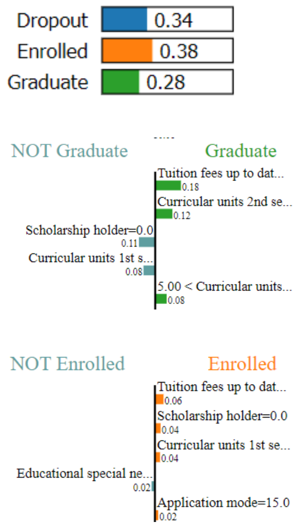
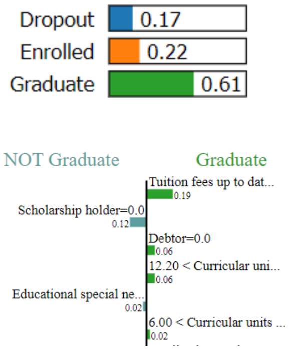
See the output from above to know what each value represents.

actual_class: 0

predicted_class: 1

A summary of the findings for several sample are presented in the table below. (Note: These sample cases are the first of their kind when going down the testing set from the second row of testing data. If any code is changed, the sample sets are subject to change, and the summary below may not be accurate to the full extent.)

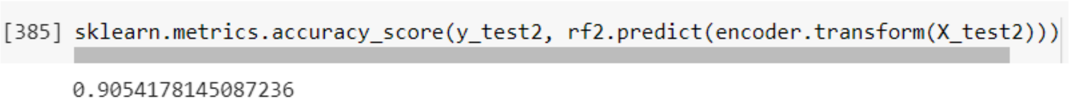
<table>
<tr>
<td rowspan="2"><b>Actual</b></td>
<td><i>Dropout</i></td>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td><i>Enrolled</i></td>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
</table>

				
	Graduate			
		Dropout	*Enrolled*	*Graduate*
		Predicted		

Part III

We continue by creating a second RF model, but this time we take out all "Enrolled" class and just work with the other two classes. We find that we get higher accuracy, which is expected going by the confusion matrix seen in the previous RF version (3-class) –a lot of the inaccuracy was in the prediction of the "Enrolled" class.

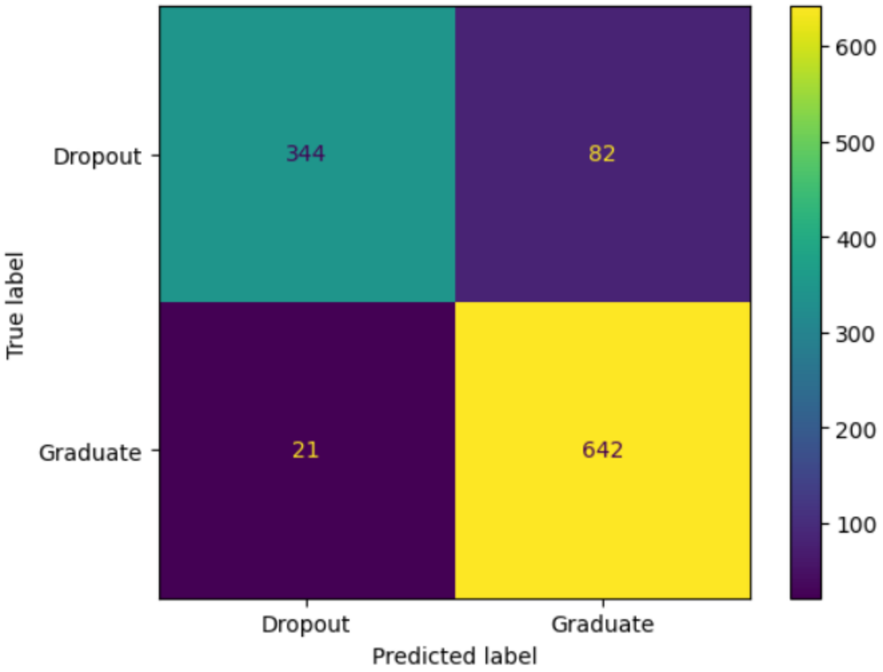
Random Forest (2-Class model):



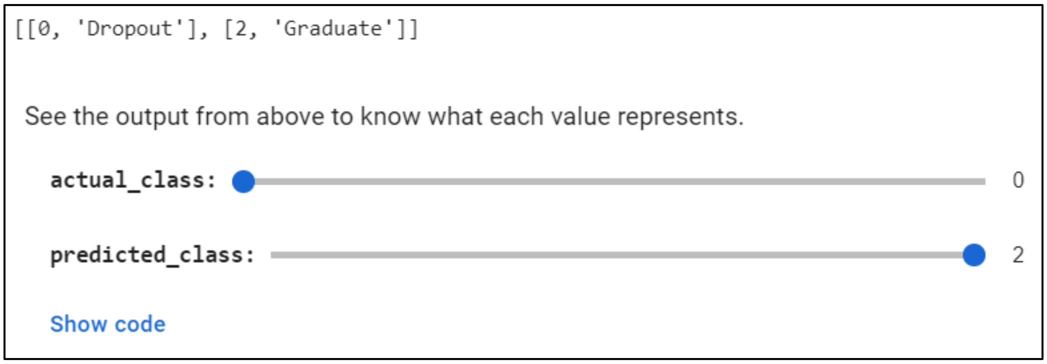
```
[385] sklearn.metrics.accuracy_score(y_test2, rf2.predict(encoder.transform(X_test2)))

      0.9054178145087236
```

The confusion matrix for this model is given below:



A LIME analysis is done again on a sample set of the testing data. Once again, the associated graphs can be found in the code file and run using forms, as follows:



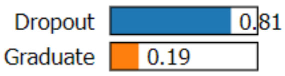
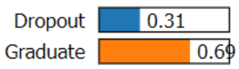
```
[[0, 'Dropout'], [2, 'Graduate']]
```

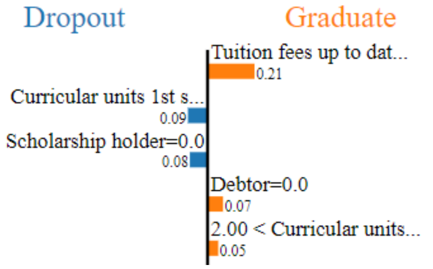
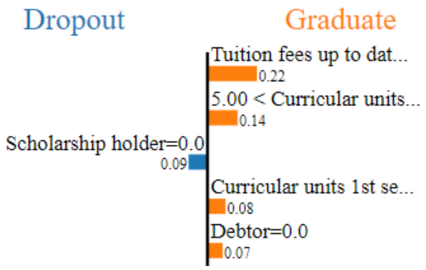
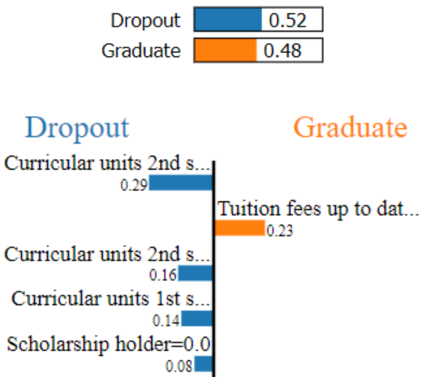
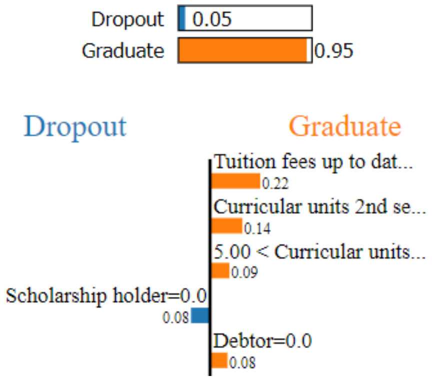
See the output from above to know what each value represents.

actual_class: 0

predicted_class: 2

The table below gives the notable findings for the samples:

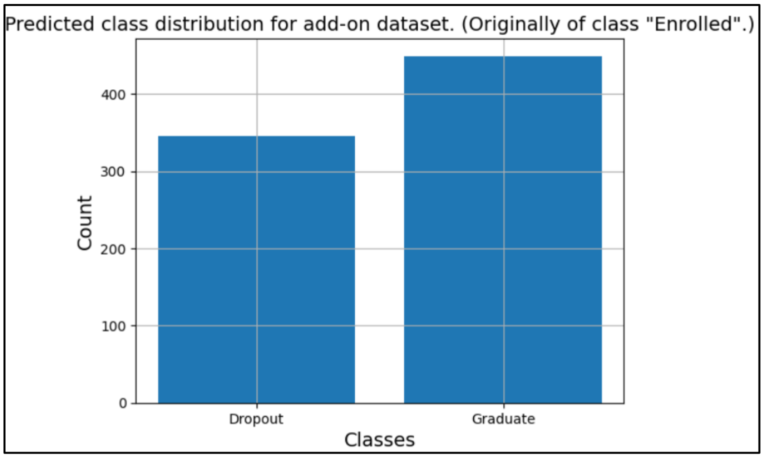
Actual	*Dropout*		

<table>
<tr><td rowspan="2"></td><td></td><td>
</td><td>
</td></tr>
<tr><td><i>Graduate</i></td><td>
</td><td>
</td></tr>
<tr><td colspan="2"></td><td><i>Dropout</i></td><td><i>Graduate</i></td></tr>
<tr><td colspan="2"></td><td colspan="2"><b>Predicted</b></td></tr>
</table>

The correct predictions happen with quite a margin of difference, particularly for the “Graduate” class. The samples belonging to the “Graduate” class had the greatest disparity, with the correct prediction being with a very high margin of difference (the model is very confident), and the incorrect prediction being a close match.

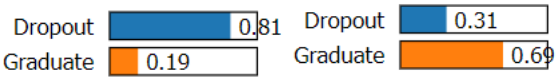
Extras

A few other steps were taken as an experiment; however, the findings were not very interesting. The second RF was utilized to predict the class of the data belonging to the “Enrolled” class. The following bar chart shows the distribution predicted from the “Enrolled” class:

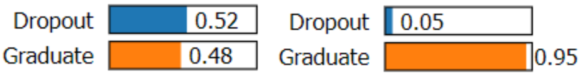


After this a new RF was trained with this data included for training. The max accuracy was still 90% (as would be expected since roughly 10% of those predictions are likely incorrect).

Actual class = Dropout:



Actual class = Graduate:



These are no different from the prior run (; the original RF with 2-classes).

References

[1] "The Devastator". (2023). Predict students' dropout and academic success, Version 2. Retrieved March 11, 2023 from https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

[2] Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins. (2021). Predict students' dropout and academic success (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5777340