Predict Students' Dropout and Academic Success


Alex Valerio


Eduardo Aleman


Irena Mehic


Nathan Beathea-Martinez


Ning Gao

Introduction

Higher education institutions record a significant amount of data about their students, representing a considerable potential to generate information, knowledge, and monitoring. Both school dropout and educational failure in higher education are an obstacle to economic growth, employment, competitiveness, and productivity, directly impacting the lives of students and their families, higher education institutions, and society as a whole.

Purpose

By using machine learning techniques to identify students at risk at an early stage of their academic path, so that strategies to support them can be put into place.

The Dataset

- This dataset is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal; founded in UC Irvine Machine Learning Repository
- The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters.
- The data is used to build classification models to predict students' dropout and academic success. The problem is formulated as a three category classification task, in which there is a strong imbalance towards one of the classes.

Dataset Characteristics	*Tabular*
Associated Tasks	*Classification*
# Instances	*4424*
# Attributes	*36*

Data Description

Macroeconomic
GDP *Numeric/ Continuous*
Inflation rate *Numeric/ Continuous*
Unemployment rate *Numeric/ Continuous*

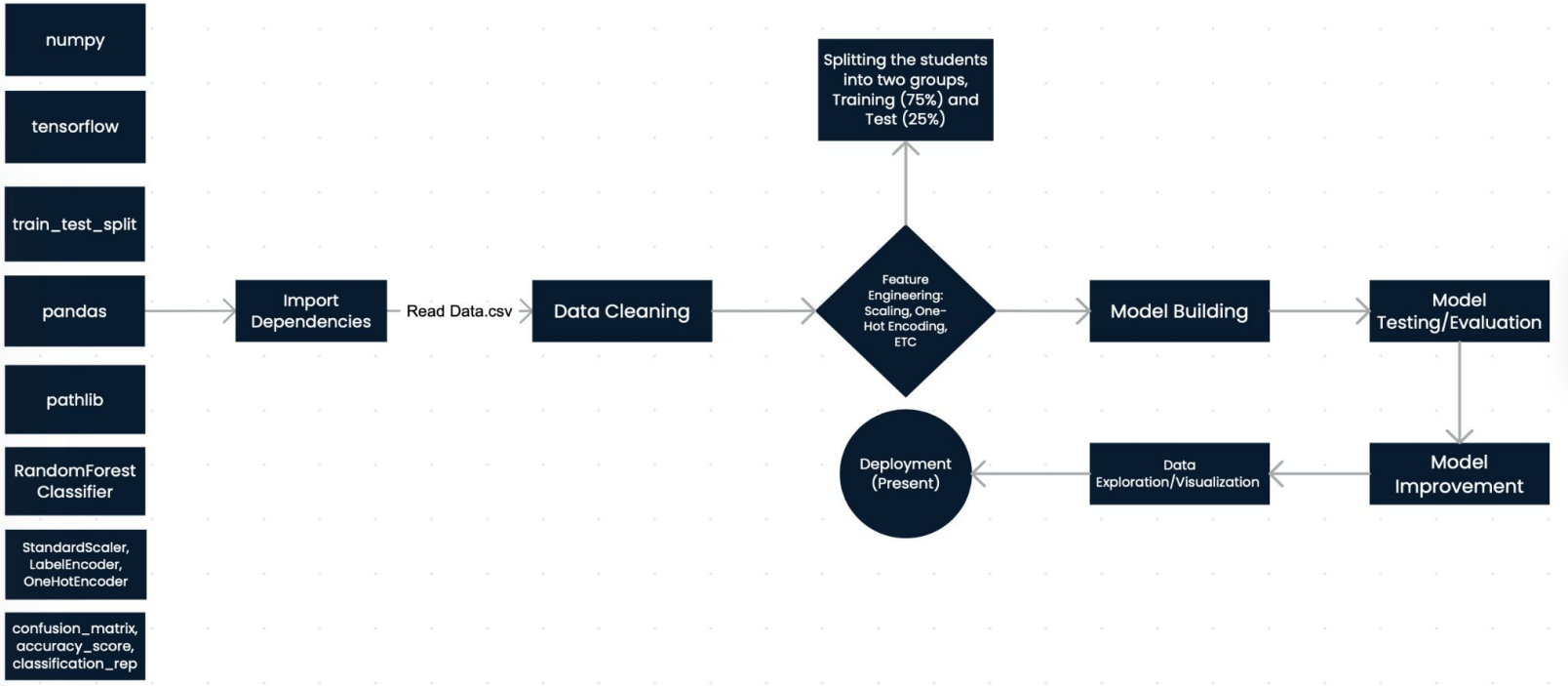
Demographic
Marital status *Numeric/ Discrete*
Nationality *Numeric/ Discrete*
Gender *Numeric/ Binary*
Age at enrollment *Numeric/ Discrete*
Displaced *Numeric/ Binary*
International *Numeric/ Binary*

Socioeconomic
Mother's qualification *Numeric/ Discrete*
Father's qualification *Numeric/ Discrete*
Mother's occupation *Numeric/ Discrete*
Father's occupation *Numeric/ Discrete*
Educational special needs *Numeric/ Binary*
Debtor *Numeric/ Binary*
Tuition fees up to date *Numeric/ Binary*
Scholarship holder *Numeric/ Binary*

Academic path
Application mode *Numeric/ Discrete*
Application order *Numeric/ Ordinal*
Admission grade *Numeric/ Continuous*
Course *Numeric/ Discrete*
Daytime/evening attendance *Numeric/ Binary*
Previous qualification *Numeric/ Discrete*
Previous qualification (grade) *Numeric/ Continuous*

Academic performance
Curricular units 1st sem (credited) *Numeric/ Discrete*
Curricular units 1st sem (enrolled) *Numeric/ Discrete*
Curricular units 1st sem (evaluations) *Numeric/ Discrete*
Curricular units 1st sem (approved) *Numeric/ Discrete*
Curricular units 1st sem (grade) *Numeric/ Discrete*
Curricular units 1st sem (without evaluations) *Numeric/ Discrete*
Curricular units 2nd sem (credited) *Numeric/ Discrete*
Curricular units 2nd sem (enrolled) *Numeric/ Discrete*
Curricular units 2nd sem (evaluations) *Numeric/ Discrete*

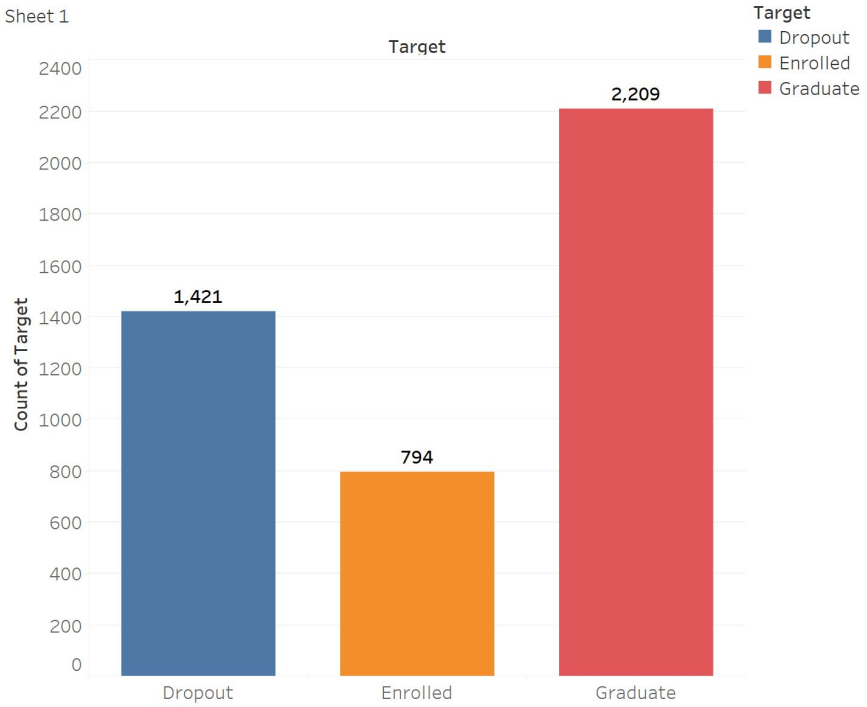
Methodology Roadmap



Data Exploration

Target

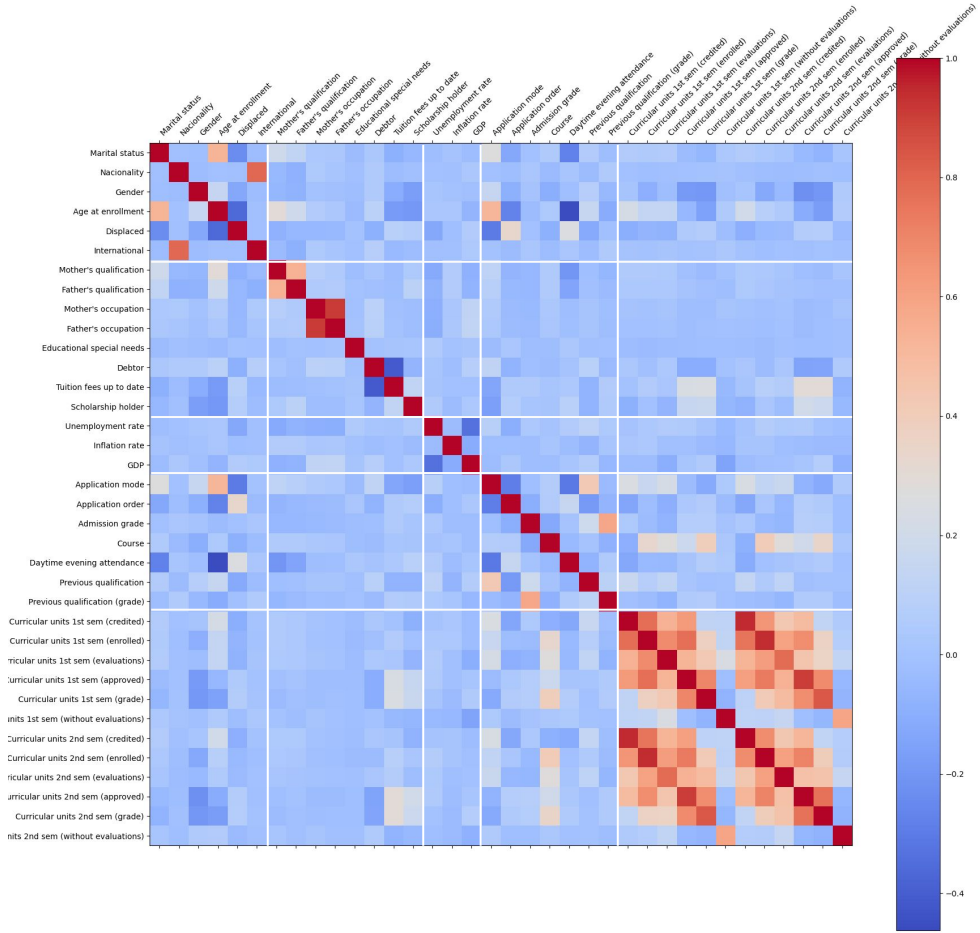
- The problem is formulated as a three category classification task, in which there is a strong imbalance towards one of the classes.
- This might result in a high prediction accuracy driven by the majority class at the expense of a poor performance of the minority class.
- At the data-level approach, a sampling technique can be applied.
- At the algorithm-level approach, a machine learning algorithm that already incorporates balancing steps must be used.



Data Exploration

Analysis

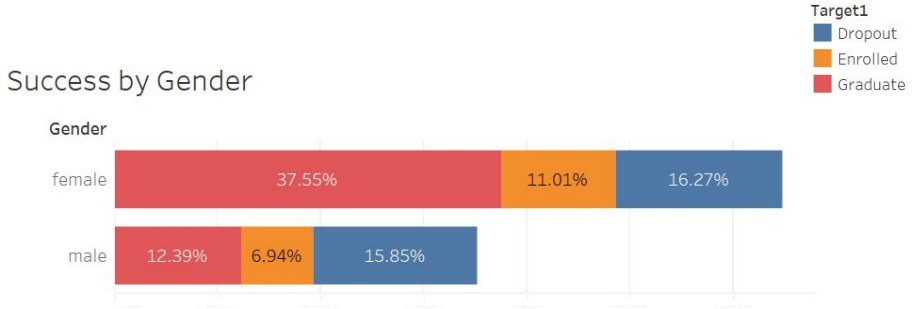
- Collinearity (or multi-collinearity) may be an issue that must be considered in some types of problems.
- The analysis of the heatmap using the Pearson correlation coefficient, shows that there are some pairs of features having high correlation coefficients, which increases multi-collinearity in the dataset.
- The collinearity is strongest within the same group of features, but we can also find higher values of correlation between groups.

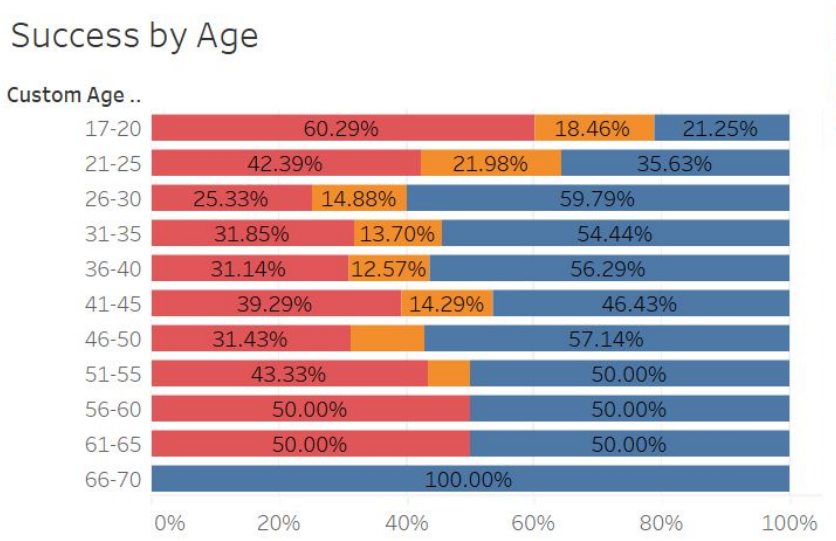


Data Exploration

Analysis

- Analysis student success possibility for each feature attribute:
  - shows that the most successful students are female students with 37.55% graduating while only 12.39% of male students successfully graduate
  - The most successful age range for graduating is 17-20 years old at 60.29%
  - Total of 1,556 male students and 2,868 female students in the data set

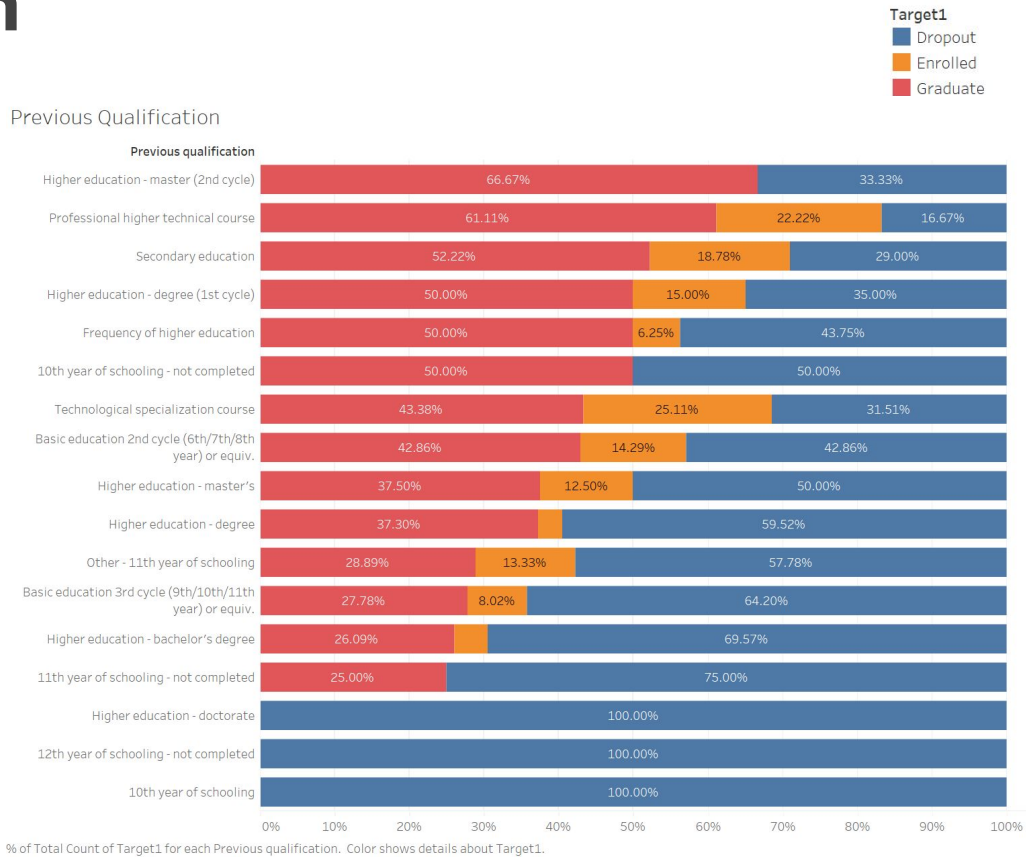




Data Exploration

Analysis

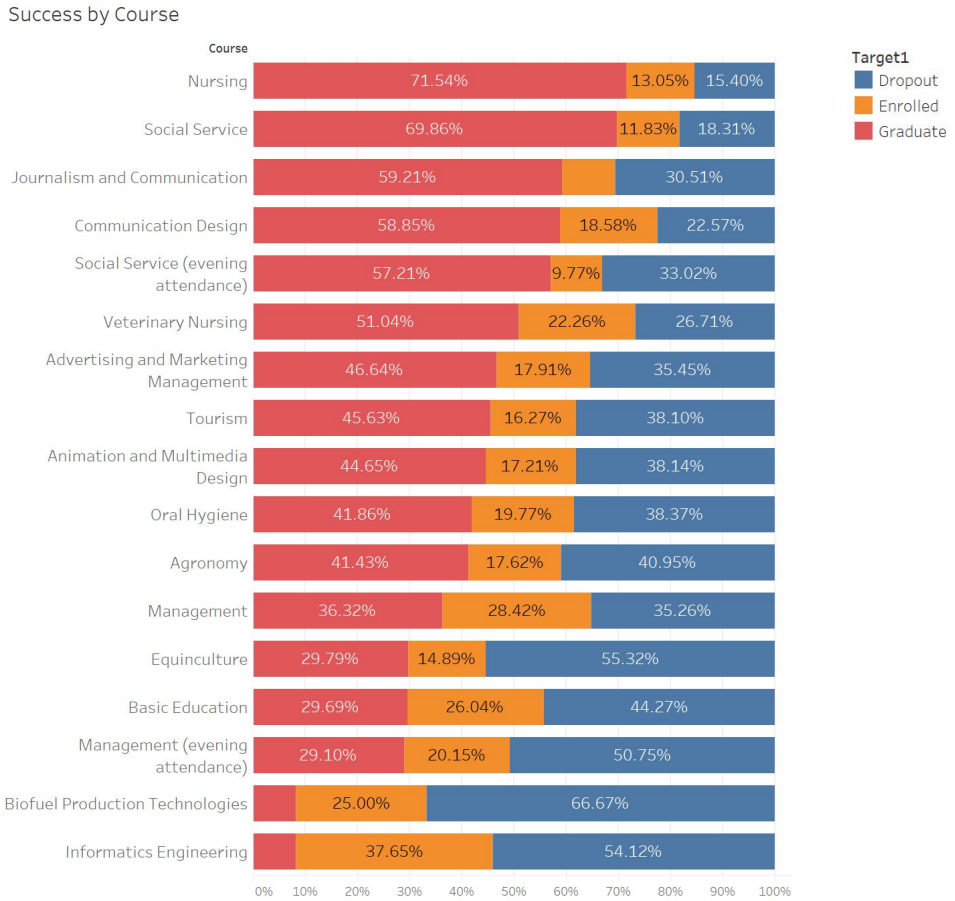
- The students with the highest rate of graduation at 66.7% had a previous qualification of a master's degree (2nd Cycle)



Data Exploration

Analysis

- Analysis student success possibility for each feature attribute:
  - Student's success rates vary a lot by courses;
  - Nursing and Social Service has the top 2 graduation rates, which are over 70%;
  - Biofuel Production Tech and Informatics Engineering are looking at very low graduation rates.



Data Exploration

Feature Attributes

- Histogram of feature attribute values to see central tendency:
  - Demographic - most of the instance are at younger age, from the same country, and a reasonable distribution between genders.
  - Socioeconomic - looking at similar level of financial level while wide range of parents' education background

Data Normalization

- **One-Hot Encoding**

  - Enabled the algorithms to process categorical data, and prevented misinterpretation of the categorical features relationships.

- **Scaling Features**

  - Helped to ensure that features with different scales did not skew the algorithm's performance, allowing it to make accurate predictions.

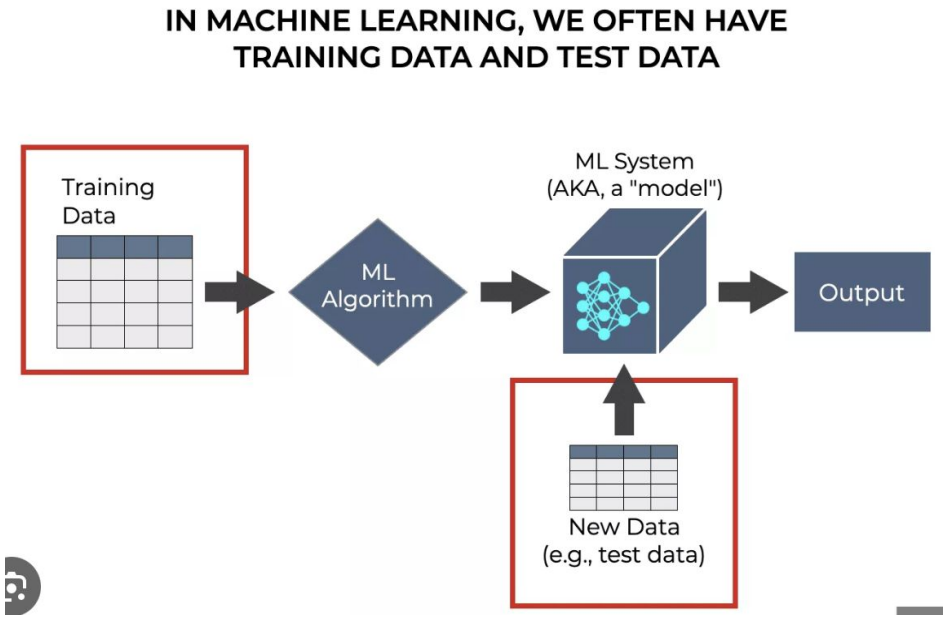
Tuition fees up to date_yes	Gender_female	Gender_male	Scholarship holder_no	Scholarship holder_yes	International_no	International_yes	Target_Dropout	Target_Enrolled	Target_Gr
1	0	1	1	0	1	0	1	0	
0	0	1	1	0	1	0	0	0	
0	0	1	1	0	1	0	1	0	
1	1	0	1	0	1	0	0	0	
1	1	0	1	0	1	0	0	0	
...	...	...	...	...	...	...	...	...	
1	0	1	1	0	1	0	0	0	
0	1	0	1	0	0	1	1	0	
1	1	0	0	1	1	0	1	0	
1	1	0	0	1	1	0	0	0	

	Application order	Previous qualification (grade)	Admission grade	Age at enrollment	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)	Curricu units 2 s
0	2.490896	-0.804841	0.022229	-0.430363	-0.300813	-2.528560	-1.986068	-1.521257	-2.197102	-0.199273	-0.282442	-2.8383
1	-0.554068	2.076819	1.071926	-0.562168	-0.300813	-0.109105	-0.550192	0.418050	0.693599	-0.199273	-0.282442	-0.1057
2	2.490896	-0.804841	-0.150419	-0.562168	-0.300813	-0.109105	-1.986068	-1.521257	-2.197102	-0.199273	-0.282442	-0.1057
3	0.207173	-0.804841	-0.509526	-0.430363	-0.300813	-0.109105	-0.071567	0.418050	0.575611	-0.199273	-0.282442	-0.1057
4	-0.554068	-2.473171	1.002867	2.864765	-0.300813	-0.109105	0.167746	0.094832	0.349468	-0.199273	-0.282442	-0.1057

Data Normalization Pt. 2

- **Split into Training/Test Set**

- Splitting data into training and testing sets was crucial to evaluate the model's performance. The training set is used for model learning, while the testing set assesses how well the model is performing, ensuring its reliability and effectiveness.



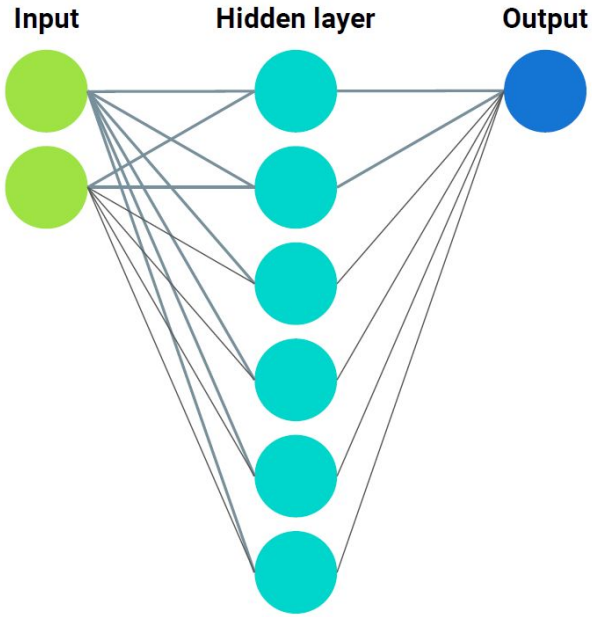
Logistic Regression - Initial Model

- This was our first attempt to create a model
  - Includes all raw data to predict a target of either **“Graduate”** or “**Dropout”**
  - Some specifics from the model: random_state=9, max_iter=1000, solver='saga'

Classification report

	precision	recall	f1-score	support
Dropout	0.94	0.84	0.89	1129
Graduate	0.90	0.96	0.93	1775
accuracy			0.92	2904
macro avg	0.92	0.90	0.91	2904
weighted avg	0.92	0.92	0.91	2904

Machine Learning Model 1 - Neural Network



- 2 hidden layers
  - hidden_nodes_layer1 = 8
  - hidden_nodes_layer2 = 5
- Output layer

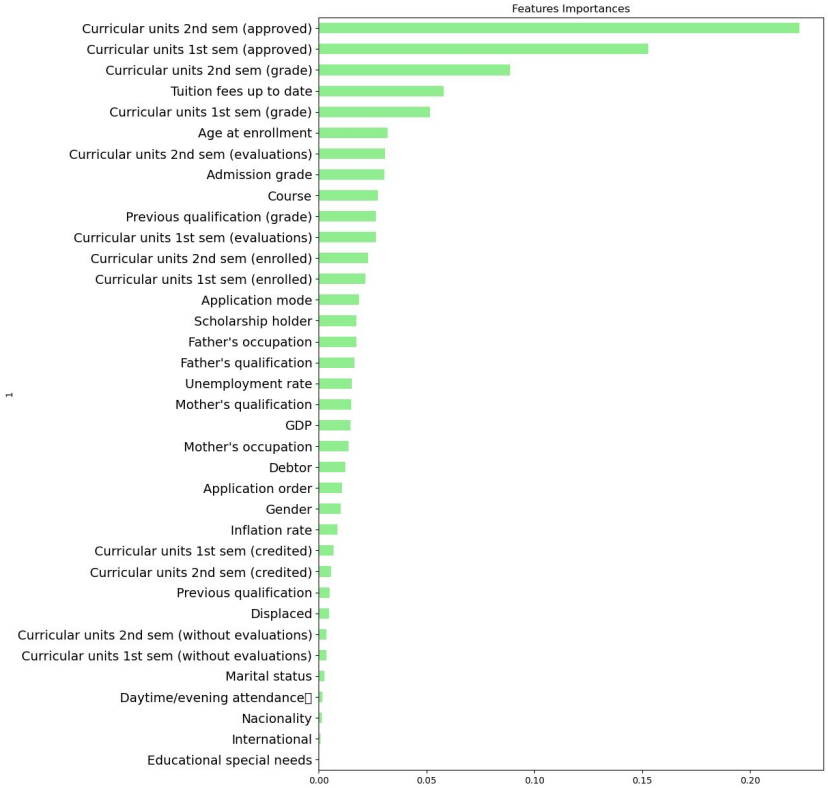
Loss: 0.3132462203502655

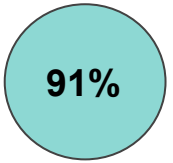
Accuracy: 0.8564566373825073

```
28/28 [==============================] - 0s 1ms/step
              precision    recall  f1-score   support

           0       0.47      0.27      0.35       161
           1       0.85      0.93      0.89       724

    accuracy                           0.81       885
   macro avg       0.66      0.60      0.62       885
weighted avg       0.78      0.81      0.79       885
```

Machine Learning Model 2 - Random Forest

Rationale





Accuracy Score

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	315	62
Actual 1	17	514

Accuracy Score : 0.9129955947136564
Classification Report

	precision	recall	f1-score	support
0	0.95	0.84	0.89	377
1	0.89	0.97	0.93	531
accuracy			0.91	908
macro avg	0.92	0.90	0.91	908
weighted avg	0.92	0.91	0.91	908

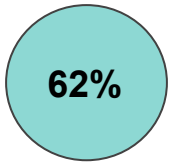
Confusion Matrix

ML Logistic Regression - (Normalized Variables)

Used 7 variables that looked to be normally distributed.

Variables Used:

- Admission grade
- Previous Qualification (grade)
- Unemployment Rate
- GDP
- Inflation Rate
- Fathers Qualifications
- Mothers Qualifications



Accuracy Score

Confusion Matrix:

		Actual	
		0	1
Predicted	0	42	316
	1	25	525

	precision	recall	f1-score	support
0	0.63	0.12	0.20	358
1	0.62	0.95	0.75	550
accuracy			0.62	908
macro avg	0.63	0.54	0.48	908
weighted avg	0.63	0.62	0.54	908

Logistic Regression (One-Hot Encoding & Standard Scaler)

Optimizing all of the columns:

One-Hot Encoding:

- Converting categorical data into a binary format

Standard Scaler:

- Scaling numerical data

Accuracy Score

Confusion Matrix

	precision	recall	f1-score	support
0	0.91	0.79	0.84	358
1	0.87	0.95	0.91	550
accuracy			0.89	908
macro avg	0.89	0.87	0.88	908
weighted avg	0.89	0.89	0.88	908

	precision	recall	f1-score	support
0	0.94	0.36	0.52	301
1	0.69	0.98	0.81	442
accuracy			0.73	743
macro avg	0.82	0.67	0.66	743
weighted avg	0.79	0.73	0.69	743

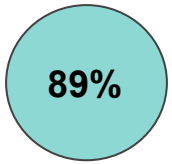
Logistic Regression (Random Forest)

Optimization:

- Leveraged the top 10 featured importance variables from the Random Forest Model.

Variables Used:

- Course
- Previous qualification (grade)
- Admission grade
- Tuition fees up to date
- Age at enrollment
- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 2nd sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)



Accuracy Score

Confusion Matrix:

		Actual	
		0	1
Predicted	0	282	76
	1	28	522

	precision	recall	f1-score	support
0	0.91	0.79	0.84	358
1	0.87	0.95	0.91	550
accuracy			0.89	908
macro avg	0.89	0.87	0.88	908
weighted avg	0.89	0.89	0.88	908

Application : Prediction

Use the best performance Machine Learning model to predict those students - whose original target status are “Enroll” - their possibility to graduate, in order to offer opportunities to students and thus avoid dropping out.

The Code:

	Marital st	Applicatic	Applicatic	Course	Daytime/e	Previous c	Previous c	Curricular	Curricular	Unemploy	Inflation r	GDP	Target	Prediction
16	single	3rd phase	1	Social Ser	daytime	Secondary	137	11	0	10	1	1	Enrolled	Graduate
19	single	1st phase	1	Basic Educ	daytime	Secondary	140	13	0	16	0	0	Enrolled	Graduate
21	single	3rd phase	4	Oral Hygie	daytime	Secondary	127	11	0	12	3	-1	Enrolled	Graduate
25	single	1st phase	1	Social Ser	daytime	Secondary	151	11	0	7	2	0	Enrolled	Graduate
27	single	1st phase	1	Veterinar	daytime	Secondary	138	13	0	9	0	-3	Enrolled	Graduate
29	single	2nd phase	2	Nursing	daytime	Secondary	127	13	0	16	0	0	Enrolled	Graduate
30	single	Technolog	1	Agronomy	daytime	Technolog	150	10	0	16	0	0	Enrolled	Graduate
32	single	1st phase	1	Managem	daytime	Secondary	143	13	0	8	1	3	Enrolled	Graduate
45	single	2nd phase	1	Managem	evening	Secondary	154	14	1	7	2	0	Enrolled	Graduate
51	single	1st phase	1	Nursing	daytime	Secondary	139	11	0	9	0	-3	Enrolled	Graduate
52	single	1st phase	1	Journalisn	daytime	Secondary	127	11	0	9	0	-3	Enrolled	Graduate
59	single	1st phase	3	Animatior	daytime	Secondary	125	0	0	7	2	0	Enrolled	Graduate
62	single	2nd phase	3	Animatior	daytime	Secondary	133	0	0	10	1	1	Enrolled	Graduate
63	single	2nd phase	1	Communi	daytime	Secondary	127	10	0	10	1	1	Enrolled	Graduate
64	single	Change of	1	Communi	daytime	Secondary	116	0	0	12	3	-1	Enrolled	Dropout
69	single	Over 23 ye	1	Tourism	daytime	Secondary	160	10	0	12	3	-1	Enrolled	Graduate
70	single	2nd phase	1	Managem	daytime	Secondary	148	0	0	7	2	0	Enrolled	Dropout
83	single	2nd phase	1	Tourism	daytime	Secondary	120	11	0	12	3	-1	Enrolled	Graduate
88	single	1st phase	6	Managem	daytime	Secondary	143	13	0	11	0	2	Enrolled	Graduate

Breakdown of the Predicted Instance:

Distribution

Prediction	
Dropout	16
Graduate	778

Course	Gender	Daytime/ev..	Debtor	
Advertising and Marketin..	male	daytime	no	1
Agronomy	male	daytime	no	2
			yes	1
Animation and Multimedi..	male	daytime	no	1
Communication Design	male	daytime	no	1
Informatics Engineering	male	daytime	no	1
			yes	3
Management	male	daytime	no	2
			yes	1
Management (evening att..	male	evening	no	1
Nursing	female	daytime	no	1
Veterinary Nursing	male	daytime	yes	1

Questions?

Thank you!