Milestone #3 for Group 4 | ISC4242

Andres Machado, Bret Geyer, Jackson Small, Jadan Colon, and Thomas Tibbetts

Description of Data:

Broadly speaking, this project's focus is predicting the academic success of university students. We are working with a tabular dataset which lists 37 attributes for each of 4,424 students at the Polytechnic Institute of Portalegre, and contains contains 18 categorical features which encode information such as the student's program, marital status, application mode, and their parents' level of education. Some of the quantitative variables encode the student's age and numerical evaluations (0-100) of their admission profile and previous qualifications, and other quantitative variables which detail the number of curricular units for which the student was enrolled, approved, and credited during their first two semesters. The dataset's target variable is a categorical column which designates each student as one of the following:

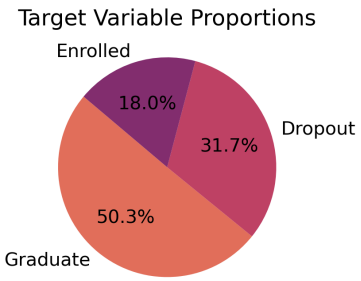
- **Dropout**; Indicating that the student dropped out by the end of the program's normal term. Where 1421 students classify as Dropouts.
- **Enrolled**; Indicating that the student was still enrolled (not graduated) at the end of the normal term of the program. Where 794 students classify as Enrolled.
- **Graduated**; Indicating the student graduated by the end of the normal term of the program. Where 2,209 students classify as Graduated.

Data Exploration:

One of the dataset's issues is containing many categorical features, each of which consists of many different levels. For example, the "Application mode" variable alone has 18 levels, some of which apply to fewer than 10 students. If all of these categorical features were one-hot encoded, they would add hundreds of sparse columns to the dataset and create unnecessary bulk during the data exploration. For this reason, it was decided that all categorical feature levels representing less than 2% of the dataset would be binned into one level called "Other". This significantly reduced the unnecessary complexity in the dataset.

While checking the data, we also noticed that there were 180 students listed as having enrolled in 0 curricular units during both their first and second semester. This may be attributed to faulty data collection or other unusual circumstances. For the purpose of predicting student success, we considered these records to be highly anomalous and decided to remove them from the dataset. After this removal, the dataset contained 4,244 records.

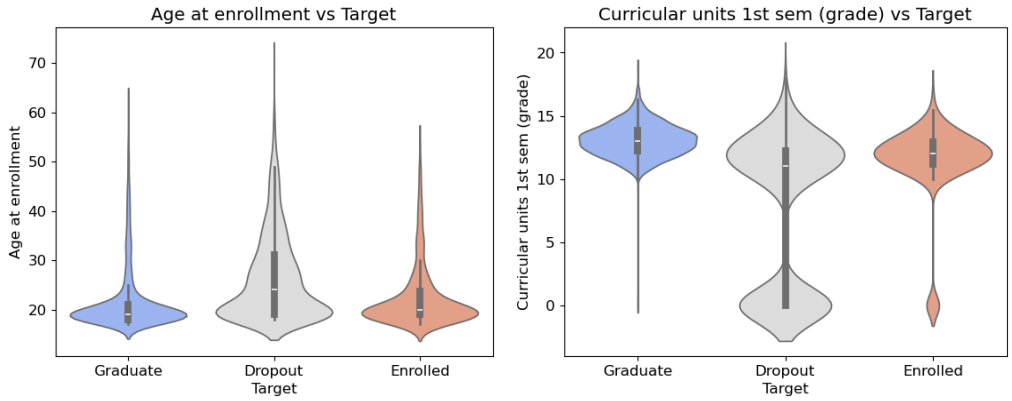
It's important to note that the target classes are imbalanced. About 50% of the students are graduates, 18% are enrolled, and 31.7% are dropouts. This will inform our modeling and experimental design. To account for the unbalanced classes, it may be necessary to use methods like Stratified K-Folds or SMOTE.



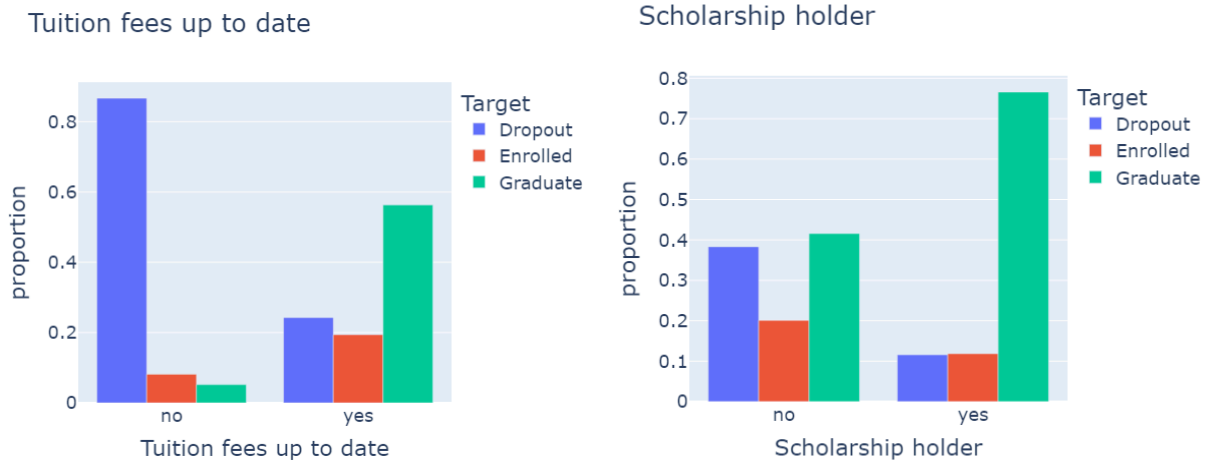
Data Visualization:

To understand the features that influence the students final outcome, we visualized key features in relation to the target variable. Each of the following plots helped us gain a deeper knowledge of the demographics, financial, and academic patterns that distinguish students from graduates, dropouts, and currently enrolled.

We discovered that students who became dropouts had a heavily right-skewed distribution of age of enrollment as compared to the other classes. We also noticed that within the first semester grades variable, a clear cluster of dropouts had grades centered around 0. In contrast, those students who graduated displayed a more balanced and higher spread of grades.



The “Tuition fees up to date” variable is a binary feature tracking whether the student had fully paid their tuition fees at the time of data collection. In the chart below, it can be seen that a high proportion of students who didn’t keep up with their tuition fees became dropouts. This finding makes us realize how much of an impact financial hardship can play in student retention. Additionally, it was found that students who held scholarships were much more likely to graduate than those who did not. Overall, these visualizations capture risk factors associated with students who dropped out, graduated or are currently enrolled, but also highlight how early academic performance and financial support are needed for students' success.



Feature Selection:

When starting the feature selection, it was essential to consider how to handle each main variable data type-categorical/nominal and numerical. For each of these data types, feature selection was treated differently; for categorical features, we use Cramer's V statistic, which utilises the Chi-squared statistic and some transformation to provide a value between 0 and 1 for a pair of categorical features. Alternatively, for numerical variables, we used Pearson's Correlation Coefficient and the SelectKBest algorithm with the ANOVA F-stat scoring function.

One observation worth noting was a tremendous amount of multicollinearity between 1st-semester features and their 2nd-semester counterparts. To generalize our features as much as possible, we opted to keep the 1st-semester features, which should result in minimal changes in results. We also dropped features that were not clearly defined when we collected the data.

Of the 17 categorical features the dataset provided, 14 were selected. However, out of the 19 numerical features the dataset provided, 7 were selected; meaning, we began with 36 features, but settled on 21 once selection was complete.

Revised Project Statement:

We attempt to use data collected from university students to predict whether they will be graduates, dropouts, or still enrolled at the end of a typical program term. Some variables of particular interest are the student's age at enrollment, scholarship status, and early grades. We will use classification algorithms with considerations for reducing class imbalance. The objective is to identify students at risk of failing to graduate, so that they can be prioritized for assistance.

Preliminary Modeling:

We established a baseline of evaluation metrics for several different classification techniques, each using 5-fold stratified cross-validation on the dataset. The models were initialized without tuning of hyperparameters. The results are summarized in the table below:

Classification Algorithm	Avg. Precision			Avg. Recall			Avg. F1-Score		
	Grad.	Enr.	Drop.	Grad.	Enr.	Drop.	Grad.	Enr.	Drop.
Logistic regression	0.76	0.54	0.77	0.94	0.18	0.78	0.84	0.27	0.78
Naive Bayes (Gaussian)	0.74	0.28	0.73	0.69	0.49	0.51	0.71	0.36	0.59
K-nearest neighbors	0.71	0.38	0.74	0.85	0.26	0.64	0.78	0.31	0.69
Perceptron	0.74	0.40	0.77	0.92	0.17	0.71	0.82	0.23	0.74
Support vector machine	0.76	0.50	0.80	0.94	0.25	0.73	0.84	0.33	0.77

While evaluating the baseline models, we noticed that the "Enrolled" class had the lowest F1-score across all algorithms. This could be due to it being the smallest minority class in the data. Further modeling may benefit from using SMOTE to resample and balance the classes.