Supervised Learning

Students' Success and Dropout

Artificial Intelligence

Group 19_1D

Henrique Pinho - up201805000

João Lopes - up201805078

Luís Marques - up201104354

Specification

The goal is to learn how to classify samples according to the subject being studied, this means that using some learning methods about some specific data we are able to learn the "pattern" in order to predict the outcome.

In our project, the objective is to predict if a student will graduate or dropout.

Tools and Algorithms

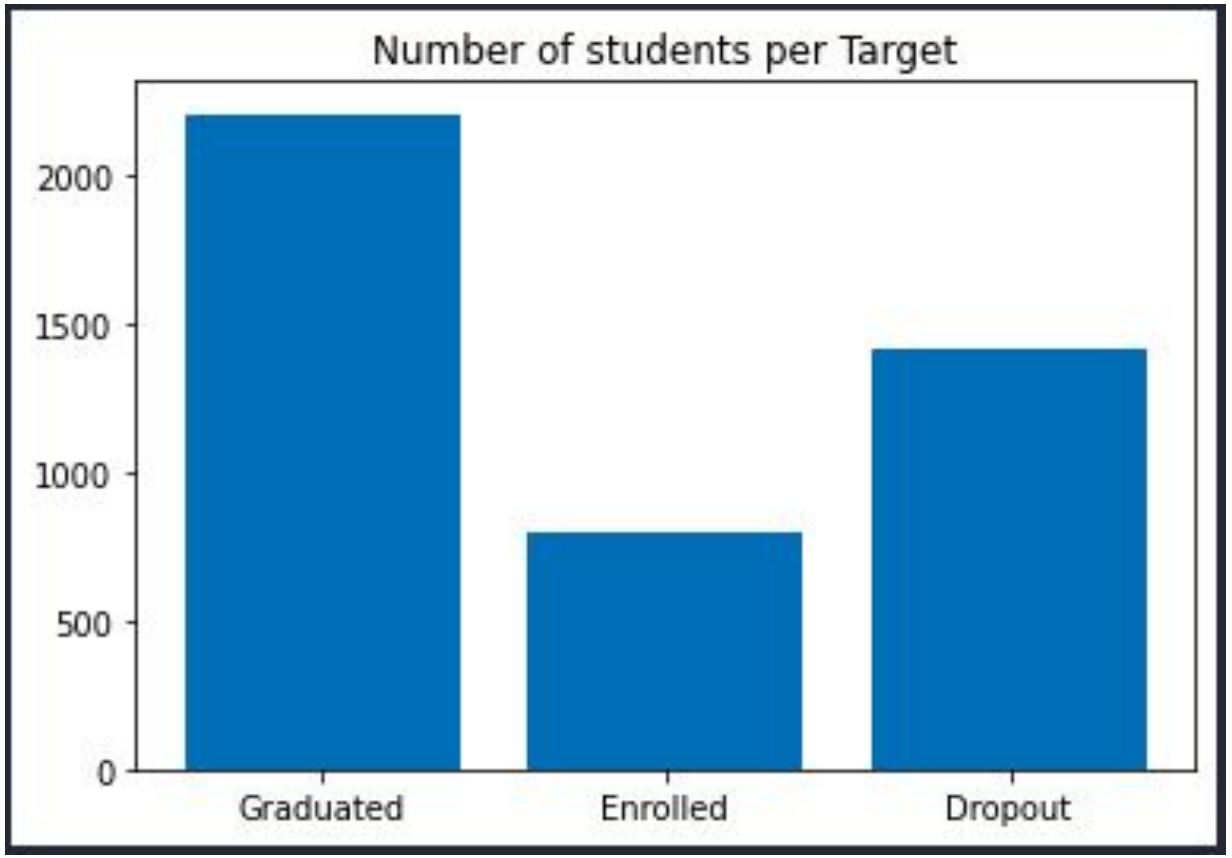
Programming language: Python3 and notebook

Libraries such as pandas, numpy, sklearn, seaborn, matplotlib, …

Classification algorithms to be implemented will be at least three of these: Decision Tree, Neural Networks, K-NN, SVM.

Learning algorithms using appropriate evaluation metrics.

Data Distribution



Pre-Processing

There’s no missing values.

Since the project is focus on success and dropout of the students we decide to not include the enrolled data.

To balance the dataset we applied a stratified sampling to the graduate data.

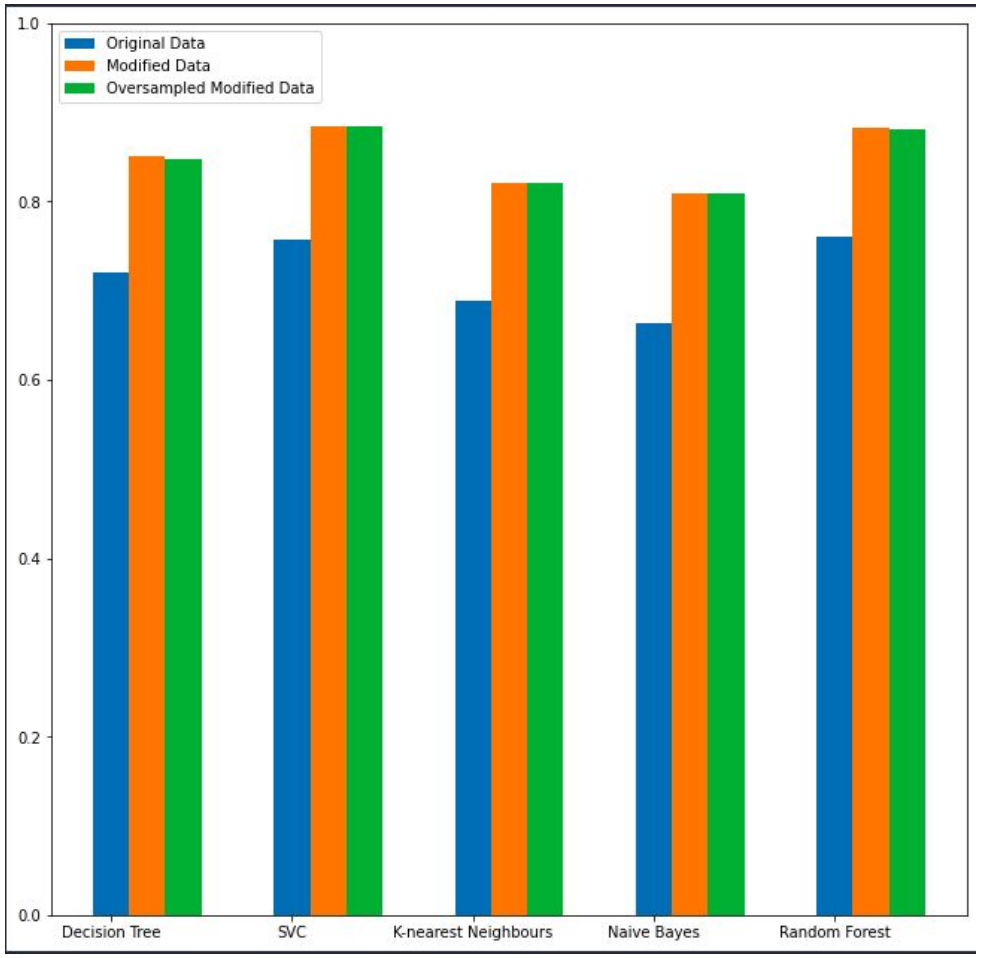
We analysed the correlation of features by constructing a correlation matrix and dropping the most correlated ones, more than 90%.

Work implemented

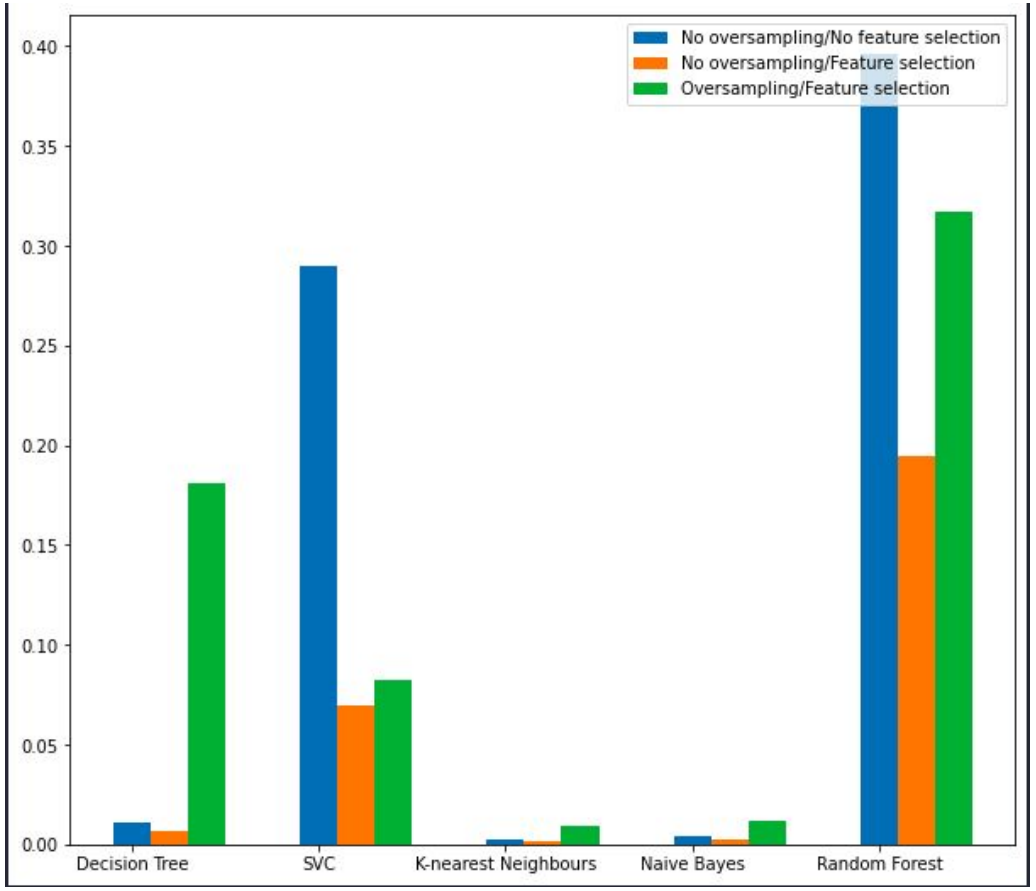
For this project we will be using multiple learning methods, as the project statement refers to.

We have decided to implement Decision Trees, KNN, SVM, RandomForest and Naive Bayes.

Model Comparison



Best score



Time to Train

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

References

Until now, we have only used the notebook from the classes:
https://moodle.up.pt/mod/resource/view.php?id=137121

Remove columns with higher correlation
https://stackoverflow.com/questions/29294983/how-to-calculate-correlation-between-all-columns-and-remove-highly-correlated-on