Using Machine Learning Models to Predict Student Academic Success Rates

For “Introduction to Machine Learning” – 242PYT305

By: Riley O’Rourke, Blue Alday, and Andrew Richardson

Date: 10/26/2024

# 1.0 Introduction

Academic institutions, advisors, teachers, and parents, to name several entities, all have vested interests in maximizing academic success rates of their respective students and children. Society as a whole directly benefits from increased educational levels: quality of life improves, technological advances enhance the way we communicate and travel, and advancements in medicine increase life expectancies. As such, research into student academic success is of vital importance so we can understand barriers to success, take corrective action, and implement positive changes. This study aims to provide critical data and insight into the variables that most significantly impact student academic success rates.

**The primary objective of this study is to use the dataset provided herein to train, test, and employ Machine Learning (ML) models to predict the “enrolled” student academic success rate.**

This report is directly coupled to a Jupyter Notebook file that contains the source code for the analysis (refer to Appendix D). and must be considered mandatorily supplemental to this written report. Many of the images, figures, and supporting data used are directly pulled from the source code. Therefore, this report, in many cases, summarizes the in-depth analysis performed in the code.

# 2.0 Description of the dataset

Table 1 below lists a summary of the dataset that will be used in the ML models. The data was captured from the following source: <https://www.kaggle.com/datasets/syedfaizanalii/predict-students-dropout-and-academic-success/data>. The raw data can also be accessed in Appendix B. A complete column mapping codex can be found in Appendix C.

|  |  |  |
| --- | --- | --- |
| **Data Column #'s** | **Summary of Data Descriptions** | **Data Category** |
| 0 | Marital status of student's parent(s) | Categorical |
| 1 - 2 | Student application information | Categorical |
| 3 | Student degree/coursework category | Categorical |
| 4 | Student attendance (daytime vs. evening) | Boolean |
| 5 - 6 | Student previous education level and performance (i.e. grade(s)) | Categorical |
| 7 | Student nationality | Categorical |
| 8 - 11 | Mother's and/or Father's education level and occupation data | Categorical |
| 12 | Student admission grade | Numeric values |
| 13 | Was the student displaced? | Boolean |
| 14 | Does the student have educational special needs? | Boolean |
| 15 | Did the student go into debt to pay for education? | Boolean |
| 16 | Are the students' tuition fees up to date? | Boolean |
| 17 | Student gender (M/F) | Boolean |
| 18 | Is the student a scholarship holder? | Boolean |
| 19 | Student age at enrollment | Numeric values |
| 20 | Is the student attending school internationally? | Boolean |
| 21 - 32 | Curricular units by semester - credited, enrolled, # of evaluations, and grade average | Numeric values |
| 33 | Unemployment rate | Percentage |
| 34 | Inflation rate | Percentage |
| 35 | Gross Domestic Product (GDP) of country where attending school | Numeric values |
| 36 | Target - Did the student dropout, enroll, or graduate the course? | Categorical |

**Table 1.** Summary of Dataset

Note: Rows in yellow form the early hypotheses for academic success prediction

The dataset consists of 4,424 entries containing 37 columns. Column indices 0-35 contain variables such as student parental education levels and occupations, student and societal macro-economical data, and various demographical information to name a few. Column 36 contains the target outcome data that will be used as the target data for training and testing the ML models.

The 4,424 target outcomes are summarized below in Table 2.

|  |  |
| --- | --- |
| **Target Outcomes** | **Value Counts** |
| "Graduate" | 2209 |
| "Dropout" | 1421 |
| "Enrolled" | 794 |
| Sub-Total | 4424 |

**Table 2.** Value Count Summary of Target Outcomes

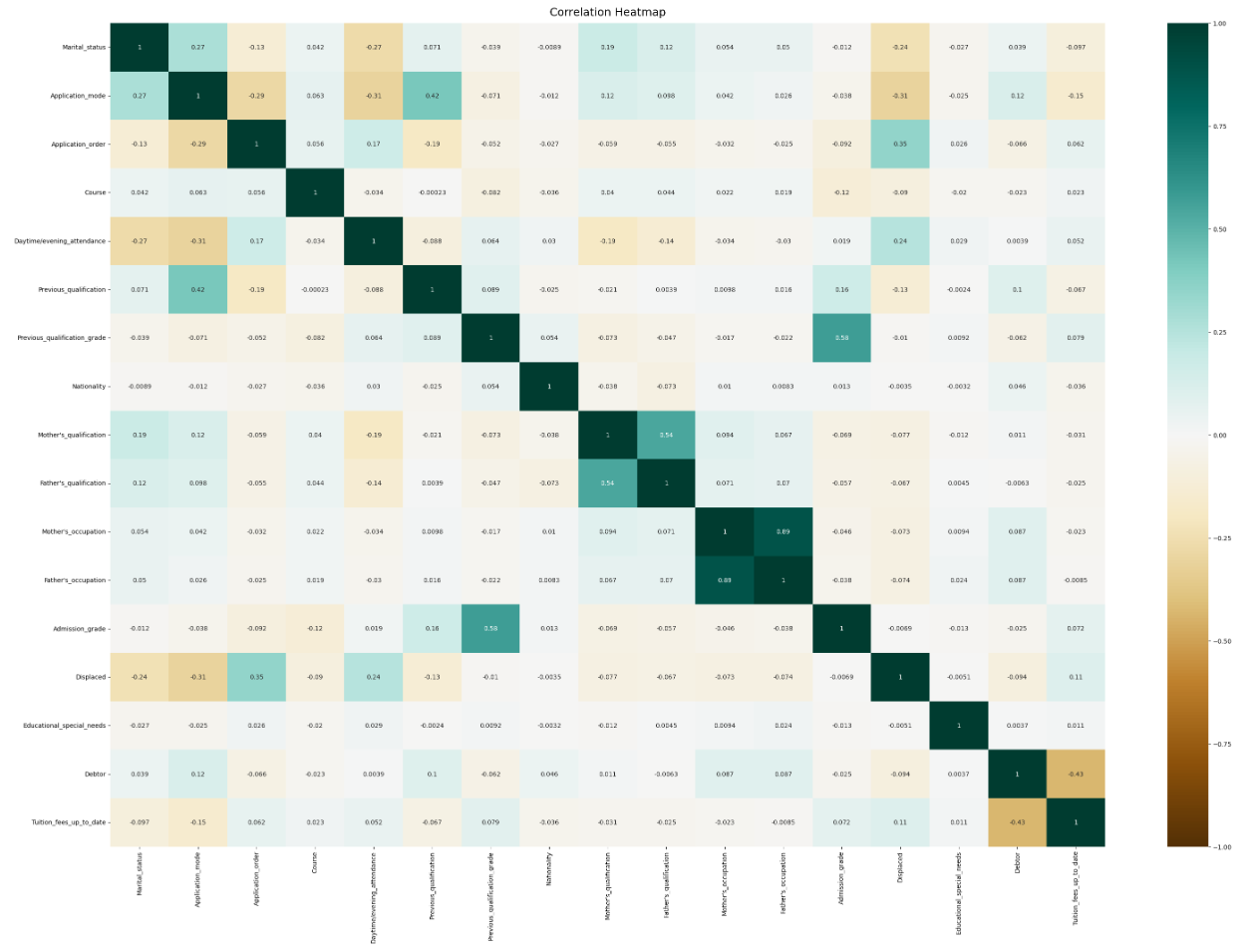
It is noted that there are *three* categorical values for target outcomes in this dataset: “Dropout”, “Enrolled”, and “Graduate”. The “Dropout” and “Graduate” categories are straight-forward to interpret as academic “failure” or “success”. However, the “Enrolled” category is not a definitive categorical state. Therefore, it would make sense to use the “Dropout/Graduate” categories as our failure/success criteria to train and test our ML model(s), and then use the trained ML model(s) to predict the future success/failure rates of the currently “Enrolled” students. This will be the primary objective of the study.

# 3.0 Exploration of Data and Early Theses

The following five (5) questions, associated w/ the five (5) highlighted yellow rows depicted in Table 1, are the authors early selections of significant predictor variables:

1. Marital status [column 0]
   1. How is student marital status correlated with academic success rates?
      1. Hypothesis: It is expected that single students outperform students in marriages, or students from divided households due to a variety of external stressors and factors/obligations (children, economic obligations like mortgages, car payments, etc.)
2. International status [column 20]
   1. How does a student’s international vs. national/domestic status correlate with academic success rates?
      1. Hypothesis: It is expected that domestic students outperform international students due to lack of barriers such as language barriers, physical proximity to familial/friendship support groups, etc.
3. Father’s & Mother’s qualifications (i.e. educational background) [columns 8 & 9]
   1. How do student parental education levels correlate with the students’ academic success rates?
      1. Hypothesis: Student academic success rate rise w/ increased levels of parental education
4. Unemployment rate [column 33]
   1. How does economic unemployment rate correlate to student academic success rates?
      1. Hypothesis: It is expected that as unemployment rates rise, academic success rates decrease.
5. Father & Mother’s occupation [columns 10 & 11]
   1. How do student parental occupations correlate with academic success rates?
      1. Hypothesis: Occupation type is directly correlated w/ academic success rates
      2. Note: This variable is an early candidate for feature engineering w/ parental qualifications mentioned in hypothesis #3.

## 3.1 Data Correlation, Multicollinearity, and Eigenvalues



**Figure 1.** Variable Correlation Heatmap

“In statistics, multicollinearity or collinearity is a situation where the predictors in a regression model are linearly dependent” [1]. This can occur when there are redundant variables in the dataset, or when multiple variables appear to be dependent/coupled (i.e. the change of one variable inducing a change in another). Because of this correlation, it can be difficult to identify, or isolate, which of the variable(s) is truly driving the change of the output. In short, multicollinearity is an undesirable trait for building ML models, so additional steps must be taken to engineer, reduce, or eliminate such patterns in the model training set.

One approach to detecting the presence of multicollinearity is to calculate eigenvalues for the matrixed data. Eigenvalues, derived from eigenvectors, indicate the sensitivity/stability of a variable to produce predictable outcomes. Though no timeless rules exist for applying thresholds to the values of the eigenvalues, there are several Rules of Thumb (RoT) [2] that will be applied herein:

1. Small eigenvalues = instability in regression coefficient = high multicollinearity = undesirable
   1. Typically, values less than 1 are considered “small”
2. Large eigenvalues = stable regression coefficients = low multicollinearity = desirable
   1. Typically, values greater than 1 are considered “large”
3. If one or more of the eigenvalues is much less than the others, this is an indicator of multicollinearity in the data

A summary of the eigenvalue calculations can be observed in the source code. **For the distribution of eigenvalues in this dataset, columns (i.e. variables) with eigenvalues > 0.78 will be used.** This reduces the columns (or variables) of the dataset from 37 to 17.

The next step involves plotting a heatmap to observe variable correlations within this reduced dataset. Figure 1 above provides a heatmap of the variable correlation matrix. In short, its objective is to visualize variables that have high correlation coefficients, indicated by regions of dark blue and/or dark brown coloring. Such variables would have negative outcomes on training ML models.

There are a few regions where the data still appears to be correlated. Therefore, further in-depth analysis will be performed to detect which variables have high correlation coefficients, and therefore which columns should be dropped from each respective dataset, as a unique DataFrame will be created and used for further analysis of each of the five (5) hypotheses/variables.

For each of the five (5) questions posed in section 3.0, correlation coefficients will be calculated. The rationale for compartmentalizing correlation coefficients is provided below in Table 3, and a summary of that analysis is provided in Appendix A *Correlation Coefficient Table – Results.*

|  |  |  |
| --- | --- | --- |
| **Label** | **Correlation Coefficient Range** | **Keep Associated Variable (Column) in Dataset?** |
| Low | X < 0.3 | Yes |
| Medium | 0.3 > X > 0.7 | Yesnote1 |
| High | X > 0.7 | No |
| Note: 1.) Further analysis may be required. | | |

**Table 3.** Correlation Coefficient Rationale

## 3.2 Results of Data Exploration and Hypothesis Insights

### 3.2.1 Marital Status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Marital\_status Encoded Value** | **Marital\_status** | **Graduate\_Count** | **Dropout\_Count** | **Success\_Rate** |
| **0** | **Single** | 2015 | 1184 | **0.63** |
| 1 | Married | 148 | 179 | 0.45 |
| 2 | Widower | 1 | 1 | 0.5 |
| 3 | Divorced | 33 | 42 | 0.44 |
| 4 | Facto\_Union | 11 | 11 | 0.5 |
| 5 | Legally\_Separated | 1 | 4 | 0.2 |

**Table 4.** Marital Status vs. Academic Success Rates

It can be observed from Table 4 that, *as predicted* in hypothesis #1, single students had the highest academic success rates of all marriage categories. An interactive Plotly stacked-bar chart is also visually provided in the Jupyter Notebook.

### 3.2.2 International Status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **International Encoded Value** | **Domestic v. International** | **Graduates** | **Dropouts** | **Success\_Rate** |
| 0 | Domestic | 2155 | 1389 | 0.61 |
| **1** | **International** | 54 | 32 | **0.63** |

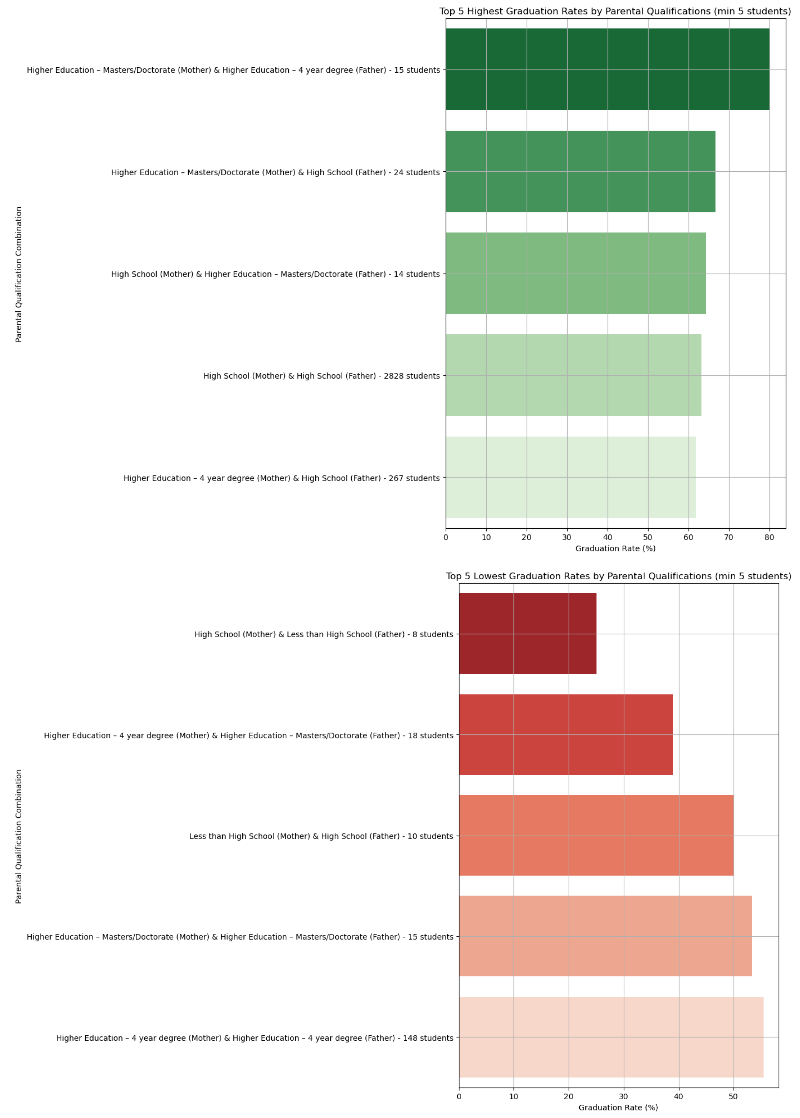
**Table 5.** Student Internation Status vs. Academic Success Rate

It can be observed from Table 5 that, in *contrast* to hypothesis #2, international students had higher academic success rates compared to domestic students, although the margin was small. An interactive Plotly stacked-bar chart is also visually provided in the Jupyter Notebook.

### 3.2.3 Father’s/Mother’s Qualifications (Education Levels)

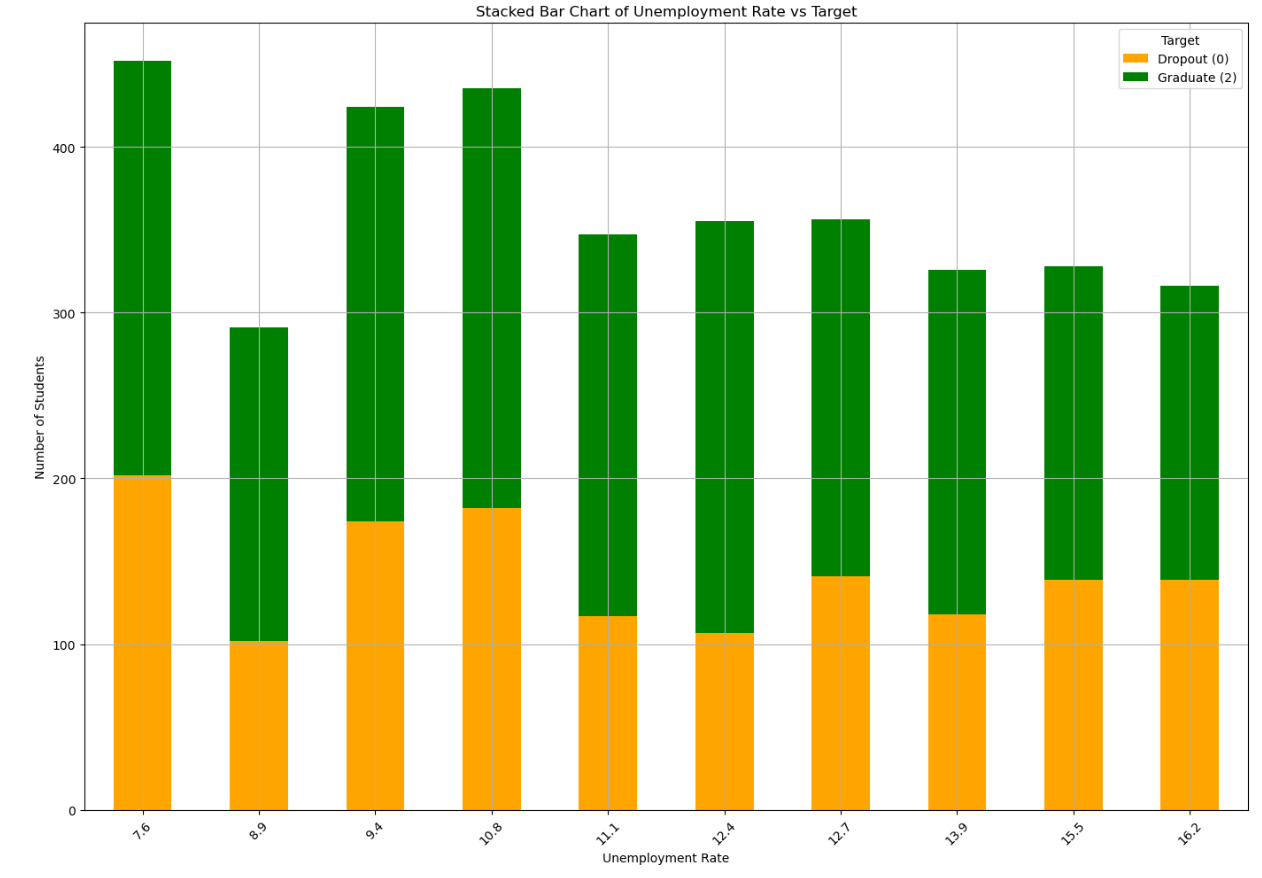
In hypothesis #3, it was predicted that as parental education levels increase, so do the students academic success rates. This was *positively observed* in the following figures:

1. Heatmaps were generated to look at student dropout and graduation rates spanning increasing levels of parental education. However, because the number of students within each category are poorly distributed, with a large number of students in the “less-than-High-School” category and very few in categories like “Higher Education – Masters/Doctorate”, a direct comparison cannot be inferred.
2. Therefore, additional plots were generated to compare *graduation rates*, or success rate per student, to normalize and compare the data. The results are provided below in Figure 2.

**Figure 2.** Comparison of Graduation Rates by Parental Education Levels

### 3.2.4 Unemployment Rate

In hypotheses #4, it was predicted that as economic Unemployment Rates (URs) rose, so did student failure rates. This *could not be positively confirmed* in the overall trend shown in Figure 3 below. Student dropout rates seemed somewhat randomized as URs increased. The trendline of dropout rates was globally marginally-positive, with localized decreases in dropout rates at the 11.1% and 12.4% unemployment rates.

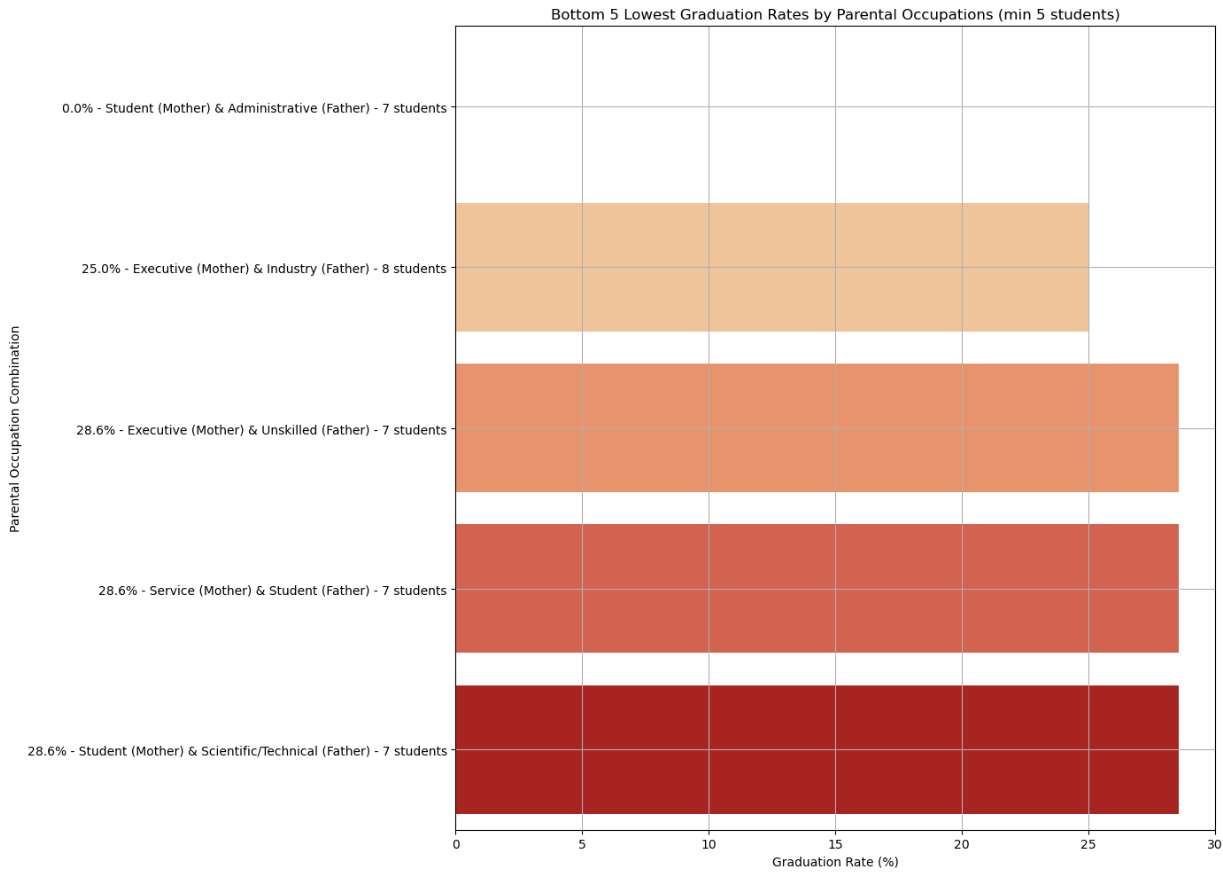
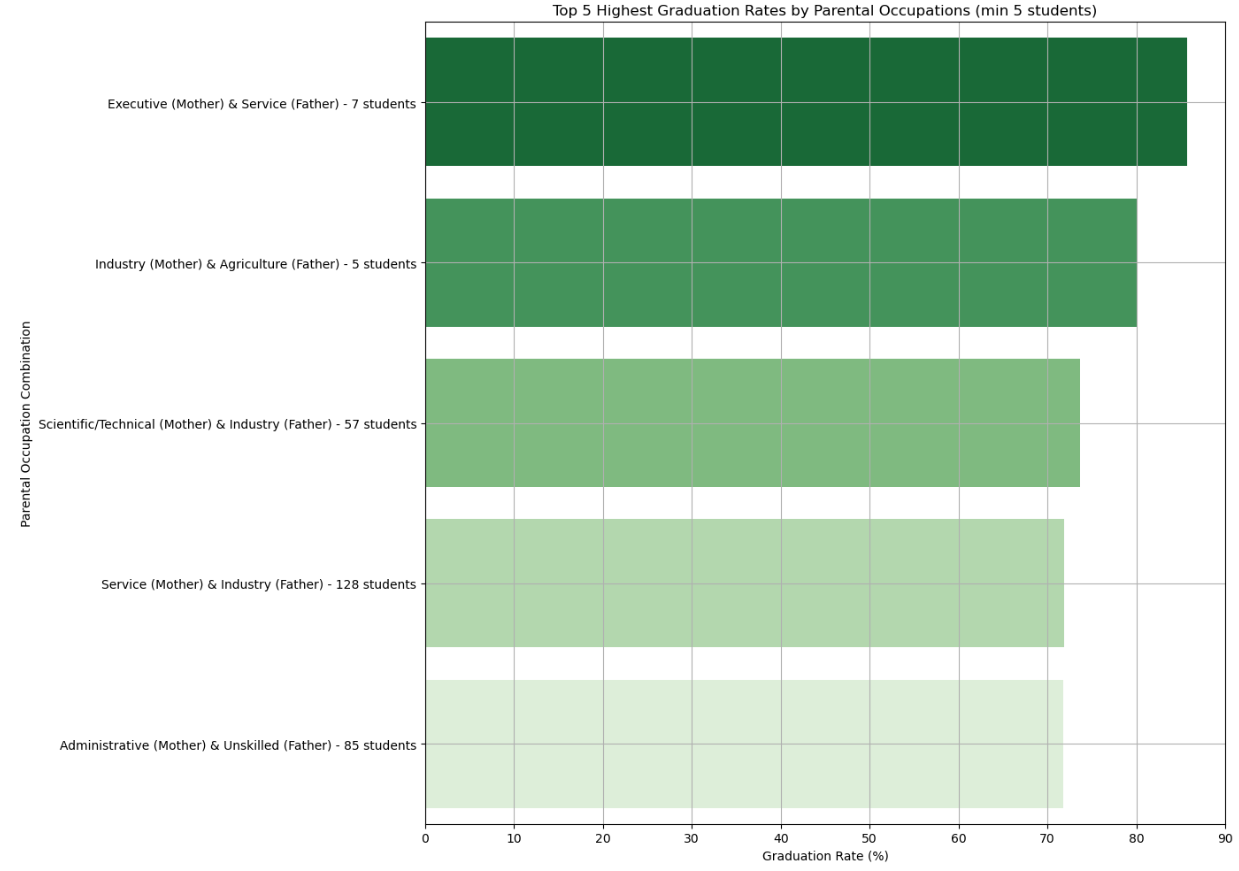
**Figure 3.** Unemployment Rate vs. Academic Success Rates

### 3.2.5 Father’s/Mother’s Occupations

In hypothesis #5, it was predicted that parental occupation type was directly proportional to student academic success rates. Parental occupation data was combined and analyzed. Figures were created and are provided in Figure 4 below. The following insights were obtained:

1. The highest graduation rate (~85%) was seen with a mother in the executive category and father in the service category
   1. Interestingly, with the mother’s occupation held constant, low graduation rates were observed when the father had an industry job (25% success rate) or was unskilled (28.6% success rate).
2. The lowest graduation rate (0%) was seen when the mother was a student and the father had an administrative position.
3. 7 out of the 10 occupation combinations had less than 10 students in each category. It is suggested that additional data be captured (of at least 30 students in each respective category) so sample distributions can normalize.

**Figure 4.** Parental Occupation vs. Graduation (Success) Rates



# 4.0 Machine Learning Model Approach

The two classification modeling approaches used for this study are the Decision Tree (DT) and Random Forest (RF) model classifiers. Classification models are being pursued over regression models due to the bimodal nature of the target data.

For each of the five (5) hypotheses proposed, the models will be constructed according to the following procedure:

1. One (1) DT model and one (1) RF model will be built for each hypothesis
   1. 80% of the dataset will be used for model training, and the remaining 20% will be used for model testing
2. Each DT and RF models’ accuracy will be assessed and tuned according to the following procedure:
   1. Build the model using default parameters to baseline model performance
   2. Calculate Feature Importances (FI) to quantify features most significantly impacting model accuracy
      1. If FI improves model accuracy it will be optimized using GridSearch
      2. If FI does not improve model accuracy, GridSearch will be performed on the default parameter dataset.
   3. Execute GridSearch
      1. Perform hyperparameter tuning to visualize model accuracy improvements
      2. If results are marginally improved, step 2.c.i may not be pursued.
   4. The same “random\_state” parameter will be used for each respective DT and RF models for each hypothesis category during the tuning stage
3. Confusion matrices will be developed for each model, totaling 30 confusion matrices
   1. (3 DTs + 3 RFs) \* 5 variables = 30 figures.
   2. This will allow the reader to easily observe the model accuracy progression during steps 2a through 2c.
4. Variable sets for the DTs and RFs will be summarized at the end of each respective section for follow-on analysis of results
5. Success will be determined if model accuracy > 70% is observed.

# 5.0 Evaluation of Data and Model Performance

## 5.1 Overview of K-Fold Model Accuracy Analysis

Optimal models attempt to reduce both bias and variance, although both cannot be simultaneously minimized, so there is a natural trade-off that must occur. For this study, the K-Fold Cross Validation Method (herein referenced as “K-Fold”) will be employed to balance or reduce the high variance and underfitting limitations of other data validation methods. A detailed guide to K-Fold CVM employment can be found in the Sci-kit Learn documentation [3].

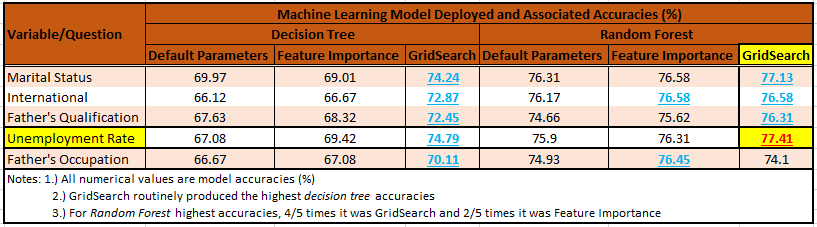
Model performance evaluation will be performed according to the following procedure:

1. Each of the Decision Tree and Random Forest model classifiers which yielded the highest accuracy will be run again using K-Fold.
   1. The DTs will use 20-fold values
   2. The RFs will use 10-fold values
2. This will be performed on each of the hypotheses.

The expectation is that the K-fold mean accuracy will be within one (1) standard deviation from the models’ maximum accuracy, found previously in Section 4.0. Further, it is expected that standard deviations will be low, indicating low variability, due to the extensive data cleaning and Exploratory Data Analysis (EDA) steps described in Section 3.0.

## 5.2 Summary of Model Performance

Table 6 below summarizes the model accuracies obtained in Sections 4.0 and 5.0 It can be observed that for each variable, a model performance of 70% or greater was achieved (as highlighted by the blue text). The highest performing model was the Unemployment Rate Random Forest Model using GridSearch, with a peak model accuracy of 77.41%. Therefore, this model will be used for the prediction phase of the study in Section 6.0

**Table 6.** Model Performance Accuracy Table

K-Fold results are tabulated below in Table 7. The expectation that all model maximum accuracies would be within one (1) standard deviation of mean accuracies (calculated per each K-fold iteration) was true in all cases except for the Marital Status Decision Tree Model. Thus, it can be concluded that this expectation was 90% correct. Further explanation to improve this is explored in Section 6.0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Model Employed** | **K-Fold CVM Results** | | | **Max Accuracy w/n std. dev?** |
| **Mean Accuracy** | **Std Accuracy** | **Max Accuracy** |
| Marital Status DT | 69.69 | 4.22 | 74.24 | 4.55 |
| Marital Status RF | 75.87 | 2.39 | 77.13 | 1.26 |
| International DT | 73.69 | 3.36 | 72.87 | 0.82 |
| International RF | 76.14 | 2.35 | 76.58 | 0.44 |
| Father's Qualification DT | 73.02 | 4.48 | 72.45 | 0.57 |
| Father's Qualification RF | 76.5 | 2.44 | 76.31 | 0.19 |
| Unemployment Rate DT | 73.71 | 4.77 | 74.79 | 1.08 |
| Unemployment Rate RF | 76.31 | 2.23 | 77.41 | 1.1 |
| Father's Occupation DT | 68.81 | 5.08 | 70.11 | 1.3 |
| Father's Occupation RF | 76.12 | 2.39 | 76.45 | 0.33 |
| Note: 1.) DT = Decision Tree  2.) RF = Random Forest | | | | |
|  |

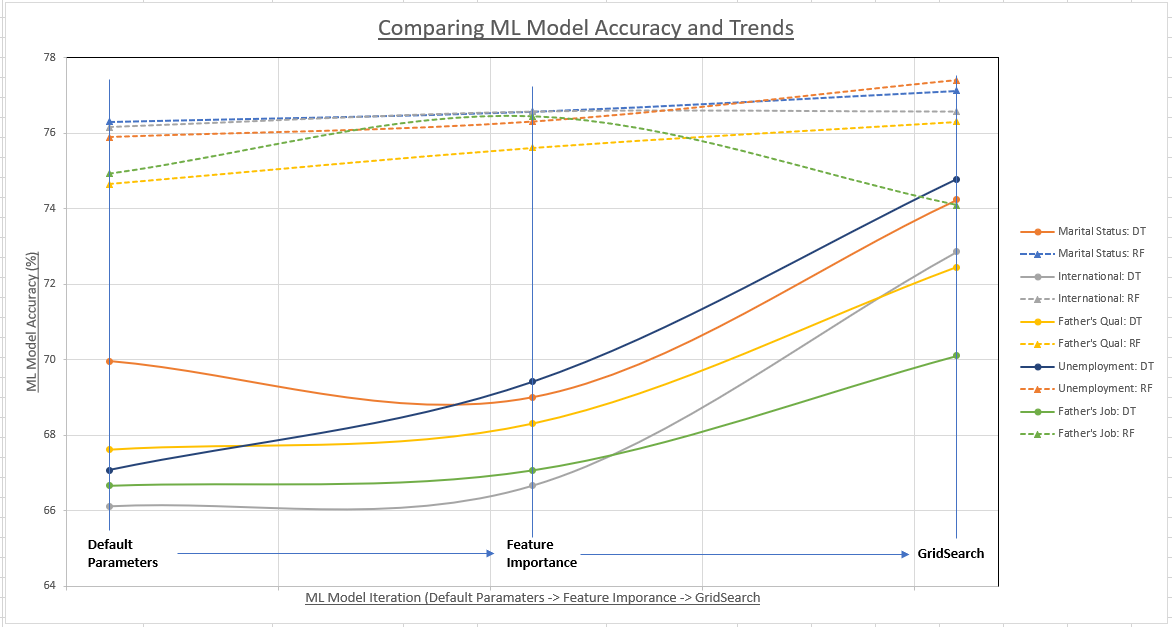
**Table 7.** K-Fold Accuracies and Statistics

# 6.0 Summary and Enrolled Student Predictions

## 6.1 Model Summary

Model comparisons and insights are provided below and graphically presented in Figure 5 after the summary points:

1. For the decision tree models, the general trend was that, in terms of model accuracy, GridSearch > Feature Importance > default parameters. This was true in all cases, as indicated by the positive slopes of each decision tree trendline, except for the marital\_status model. In this singular case, Feature Importance produced the lowest accuracy, however the GridSearch function still produced the highest overall accuracy.
2. Similarly, the Random Forest models had the same general trend. However, in the singular case of the Father's Occupation model, it can be seen that GridSearch returned the lowest accuracy.
3. In all 30 models we produced 3 Decision Trees and 3 Random Forests for each of the 5 hypothesized variables (3\_dts + 3\_rfs) x 5\_variables = 30\_models. 90% of the time, the GridSearch models returned the highest model accuracy.
4. In 15/15 cases, the RF models outperformed the DT models, with an average model accuracy increase of 6.57%
5. Early on, the decision to remove the "enrolled" students from the dataset also significantly reduced noise in the data and improved all model accuracies by >10%
6. Working with smaller datasets after applying eigenvalues and retaining only high-value coefficients greatly reduced multicollinearity

**Figure 5.** Comparing ML Model Accuracy and Trends

## 6.2 Enrolled Student Prediction

As identified ins Section 4.0 and 5.0, the Unemployment Rate GridSearch Random Forest model yielded the highest classification accuracy (of 77.41%). Therefore, this model was used to predict the academic success rate of the currently enrolled students. The enrolled student data was passed into the RF prediction model and results were obtained.

**Final prediction results of the model:**

**Of the 794 currently enrolled students, the study predicts that 76.95% (611 out of 794) students will graduate in the future, and 23.05% (183 out of 794) will dropout.**

# 7.0 References

[1] https://en.wikipedia.org/wiki/Multicollinearity

[2] https://hex.tech/blog/detecting-and-remedying-multicollinearity

[3] https://scikit-learn.org/stable/modules/cross\_validation.html

# Appendix A: Correlation Coefficient Table - Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Hypothesis #** | **Variable/Data Column** | **# of Correlation Coefficients in Respective Range** | | | **Dropped Column** | **Dropped Column Correlation Coefficient** | **Notes** |
| **High (X > 0.7)** | **Medium (0.7 < X < 0.3)** | **Low (X < 0.3)** |
| 1 | Marital Status | 0 | 0 | 17 | - | - | - |
| 2 | International | 1 | 0 | 16 | Nationality | 0.797 | - |
| 3 | Mother's Qualification | 0 | 1 | 16 | - | - | The mother's/father's qualification variables returned identical results. Therefore, only the father's quals will be used going forward. Mother's qualification coefficient in the "Medium" range = 0.544, but will be kept in dataset |
| Father's Qualification | 0 | 1 | 16 | - | - |
| 4 | Unemployment Rate | 0 | 0 | 17 | - | - | - |
| 5 | Mother's Occupation | 1 | 0 | 16 | - | - | The mother's/father's occupation variables returned similar results, and in each case was the highly correlated coefficient for the other. Therefore, the father's occupation will be used, and the mothers' occupation will be dropped from the respective DataFrame |
| Father's Ocucpation | 1 | 0 | 16 | Mother's Occupation | 0.919 |

# Appendix B: Raw Dataset



# Appendix C: Column Mapping Codex for Dataset



|  |  |  |
| --- | --- | --- |
| **Column #** | **Variable Name** | **Encoded Variable Value - Associated Text-Based Description** |
| 0 | Marital status | 1 – single  2 – married  3 – widower  4 – divorced  5 – facto union  6 – legally separated |
| 1 | Application mode | 1 - 1st phase - general contingent  2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island)  7 - Holders of other higher courses  10 - Ordinance No. 854-B/99 15 - International student (bachelor)  16 - 1st phase - special contingent (Madeira Island)  17 - 2nd phase - general contingent  18 - 3rd phase - general contingent  26 - Ordinance No. 533-A/99, item b2) (Different Plan)  27 - Ordinance No. 533-A/99, item b3 (Other Institution)  39 - Over 23 years old  42 - Transfer  43 - Change of course  44 - Technological specialization diploma holders  51 - Change of institution/course  53 - Short cycle diploma holders  57 - Change of institution/course (International) |
| 2 | Application order | Application order (between 0 - first choice; and 9 last choice) |
| 3 | Course | 33 - Biofuel Production Technologies  171 - Animation and Multimedia Design  8014 - Social Service (evening attendance)  9003 - Agronomy  9070 - Communication Design  9085 - Veterinary Nursing  9119 - Informatics Engineering  9130 - Equinculture  9147 - Management  9238 - Social Service  9254 - Tourism  9500 - Nursing  9556 - Oral Hygiene  9670 - Advertising and Marketing Management  9773 - Journalism and Communication  9853 - Basic Education  9991 - Management (evening attendance) |
| 4 | Daytime/evening attendance | 1 – daytime  0 - evening |
| 5 | Previous qualification | 1 - Secondary education  2 - Higher education - bachelor's degree  3 - Higher education - degree  4 - Higher education - master's  5 - Higher education - doctorate  6 - Frequency of higher education  9 - 12th year of schooling - not completed  10 - 11th year of schooling - not completed  12 - Other - 11th year of schooling  14 - 10th year of schooling  15 - 10th year of schooling - not completed  19 - Basic education 3rd cycle (9th/10th/11th year) or equiv.  38 - Basic education 2nd cycle (6th/7th/8th year) or equiv.  39 - Technological specialization course  40 - Higher education - degree (1st cycle)  42 - Professional higher technical course  43 - Higher education - master (2nd cycle) |
| 6 | Previous qualification (grade) | Grade of previous qualification (between 0 and 200) |
| 7 | Nationality | 1 - Portuguese 2 - German 6 - Spanish 11 - Italian 13 - Dutch 14 - English 17 - Lithuanian 21 - Angolan 22 - Cape Verdean 24 - Guinean 25 - Mozambican 26 - Santomean 32 - Turkish 41 - Brazilian 62 - Romanian 100 - Moldova (Republic of) 101 - Mexican 103 - Ukrainian 105 - Russian 108 - Cuban 109 - Colombian |
| 8 | Mother's qualification | 1 - Secondary Education - 12th Year of Schooling or Eq.  2 - Higher Education - Bachelor's Degree  3 - Higher Education - Degree  4 - Higher Education - Master's  5 - Higher Education - Doctorate  6 - Frequency of Higher Education  9 - 12th Year of Schooling - Not Completed  10 - 11th Year of Schooling - Not Completed  11 - 7th Year (Old)  12 - Other - 11th Year of Schooling  14 - 10th Year of Schooling  18 - General commerce course  19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.  22 - Technical-professional course  26 - 7th year of schooling  27 - 2nd cycle of the general high school course  29 - 9th Year of Schooling - Not Completed  30 - 8th year of schooling  34 - Unknown  35 - Can't read or write  36 - Can read without having a 4th year of schooling  37 - Basic education 1st cycle (4th/5th year) or equiv.  38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.  39 - Technological specialization course  40 - Higher education - degree (1st cycle)  41 - Specialized higher studies course  42 - Professional higher technical course  43 - Higher Education - Master (2nd cycle)  44 - Higher Education - Doctorate (3rd cycle) |
| 9 | Father's qualification | 1 - Secondary Education - 12th Year of Schooling or Eq.  2 - Higher Education - Bachelor's Degree  3 - Higher Education - Degree  4 - Higher Education - Master's  5 - Higher Education - Doctorate  6 - Frequency of Higher Education  9 - 12th Year of Schooling - Not Completed  10 - 11th Year of Schooling - Not Completed  11 - 7th Year (Old)  12 - Other - 11th Year of Schooling  13 - 2nd year complementary high school course  14 - 10th Year of Schooling  18 - General commerce course  19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.  20 - Complementary High School Course  22 - Technical-professional course  25 - Complementary High School Course - not concluded  26 - 7th year of schooling  27 - 2nd cycle of the general high school course  29 - 9th Year of Schooling - Not Completed  30 - 8th year of schooling  31 - General Course of Administration and Commerce  33 - Supplementary Accounting and Administration  34 - Unknown  35 - Can't read or write  36 - Can read without having a 4th year of schooling  37 - Basic education 1st cycle (4th/5th year) or equiv.  38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.  39 - Technological specialization course  40 - Higher education - degree (1st cycle)  41 - Specialized higher studies course  42 - Professional higher technical course  43 - Higher Education - Master (2nd cycle)  44 - Higher Education - Doctorate (3rd cycle) |
| 10 | Mother's occupation | 0 - Student  1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers  2 - Specialists in Intellectual and Scientific Activities  3 - Intermediate Level Technicians and Professions  4 - Administrative staff  5 - Personal Services, Security and Safety Workers and Sellers  6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry  7 - Skilled Workers in Industry, Construction and Craftsmen  8 - Installation and Machine Operators and Assembly Workers  9 - Unskilled Workers  10 - Armed Forces Professions 90 - Other Situation  99 - (blank)  122 - Health professionals  123 - teachers  125 - Specialists in information and communication technologies (ICT)  131 - Intermediate level science and engineering technicians and professions  132 - Technicians and professionals, of intermediate level of health  134 - Intermediate level technicians from legal, social, sports, cultural and similar services  141 - Office workers, secretaries in general and data processing operators  143 - Data, accounting, statistical, financial services and registry-related operators  144 - Other administrative support staff  151 - personal service workers  152 - sellers  153 - Personal care workers and the like  171 - Skilled construction workers and the like, except electricians  173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like  175 - Workers in food processing, woodworking, clothing and other industries and crafts  191 - cleaning workers  192 - Unskilled workers in agriculture, animal production, fisheries and forestry  193 - Unskilled workers in extractive industry, construction, manufacturing and transport  194 - Meal preparation assistants |
| 11 | Father's occupation | 0 - Student  1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities  3 - Intermediate Level Technicians and Professions  4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers  6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry  7 - Skilled Workers in Industry, Construction and Craftsmen  8 - Installation and Machine Operators and Assembly Workers  9 - Unskilled Workers  10 - Armed Forces Professions  90 - Other Situation  99 - (blank)  101 - Armed Forces Officers  102 - Armed Forces Sergeants  103 - Other Armed Forces personnel  112 - Directors of administrative and commercial services  114 - Hotel, catering, trade and other services directors  121 - Specialists in the physical sciences, mathematics, engineering and related techniques  122 - Health professionals  123 - teachers  124 - Specialists in finance, accounting, administrative organization, public and commercial relations  131 - Intermediate level science and engineering technicians and professions  132 - Technicians and professionals, of intermediate level of health  134 - Intermediate level technicians from legal, social, sports, cultural and similar services  135 - Information and communication technology technicians  141 - Office workers, secretaries in general and data processing operators  143 - Data, accounting, statistical, financial services and registry-related operators  144 - Other administrative support staff  151 - personal service workers  152 - sellers  153 - Personal care workers and the like  154 - Protection and security services personnel  161 - Market-oriented farmers and skilled agricultural and animal production workers  163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence  171 - Skilled construction workers and the like, except electricians  172 - Skilled workers in metallurgy, metalworking and similar  174 - Skilled workers in electricity and electronics  175 - Workers in food processing, woodworking, clothing and other industries and crafts  181 - Fixed plant and machine operators  182 - assembly workers  183 - Vehicle drivers and mobile equipment operators  192 - Unskilled workers in agriculture, animal production, fisheries and forestry  193 - Unskilled workers in extractive industry, construction, manufacturing and transport  194 - Meal preparation assistants  195 - Street vendors (except food) and street service providers |
| 12 | Admission grade | Admission grade (between 0 and 200) |
| 13 | Displaced | 1 – yes  0 – no |
| 14 | Educational special needs | 1 – yes  0 – no |
| 15 | Debtor | 1 – yes  0 – no |
| 16 | Tuition fees up to date | 1 – yes  0 – no |
| 17 | Gender | 1 – male  0 – female |
| 18 | Scholarship holder | 1 – yes  0 – no |
| 19 | Age at enrollment | Age of student at enrollment |
| 20 | International | 1 – yes  0 – no |
| 21 | Curricular units 1st sem (credited) | Number of curricular units credited in the 1st semester |
| 22 | Curricular units 1st sem (enrolled) | Number of curricular units enrolled in the 1st semester |
| 23 | Curricular units 1st sem (evaluations) | Number of evaluations to curricular units in the 1st semester |
| 24 | Curricular units 1st sem (approved) | Number of curricular units approved in the 1st semester |
| 25 | Curricular units 1st sem (grade) | Grade average in the 1st semester (between 0 and 20) |
| 26 | Curricular units 1st sem (without evaluations) | Number of curricular units without evalutions in the 1st semester |
| 27 | Curricular units 2nd sem (credited) | Number of curricular units credited in the 2nd semester |
| 28 | Curricular units 2nd sem (enrolled) | Number of curricular units enrolled in the 2nd semester |
| 29 | Curricular units 2nd sem (evaluations) | Number of evaluations to curricular units in the 2nd semester |
| 30 | Curricular units 2nd sem (approved) | Number of curricular units approved in the 2nd semester |
| 31 | Curricular units 2nd sem (grade) | Grade average in the 2nd semester (between 0 and 20) |
| 32 | Curricular units 2nd sem (without evaluations) | Number of curricular units without evaluations in the 1st semester |
| 33 | Unemployment rate | Unemployment rate (%) |
| 34 | Inflation rate | Inflation rate (%) |
| 35 | GDP | GDP |
| 36 | Target | 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 |

# Appendix D: Source Code

