Predicting Students' Dropout and Academic Success via Classification

Betul Mescioglu
betulmescioglu@lewisu.edu
DATA-51000-001, Summer 2023
Data Mining and Analytics
Lewis University

I. Introduction

The dataset was obtained from a higher education institution in Portugal by M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho, who also wrote the paper "Early prediction of student's performance in higher education: a case study"(2021) using this dataset [1][2]. The funding was provided by the program SATDAP - Capacitação da Administração Pública. The purpose of creating this dataset was to contribute to the reduction of academic dropout and failure in higher education by using machine learning techniques. The dataset includes information on academic path, demographics, and socioeconomic factors related to the students who enrolled in undergraduate degrees, such as education, nursing, and journalism at the time, as well as the students' academic performance at the end of the first and second semesters. The 'Target' values were 'Dropout', 'Enrolled' and 'Graduate'. By using the provided features and employing classification techniques, I classified the data into these three categories. This is important as the institution can identify the students who are on the "dropout path" and intervene earlier to ensure they have all the necessary means to continue their education.

The future sections of this report describe the dataset, applied pre-processing methods, applied classification algorithms and their results and the conclusion.

II. Data Description

The dataset consists of 36 features, a "Target" column and 4424 student records as presented in Table IV in Appendix. Features describe each student in terms of their demographic, socioeconomic status, and academic standing. Some of the demographic features are nationality, gender, age, and whether they are international student or not. Some of the socioeconomic features are marital status, father's occupation, mother's occupation and whether tuition fees are up to date or not. Some of the academic standing features are grade of previous degree, admission grade and grades of the first and second semesters and how many curricular credits the student acquired. All the values in the dataset are numeric except the "Target" column where values were labeled as 'Dropout', 'Enrolled', and 'Graduate'. I used all the features in my analysis.

The frequency distribution of all the features and the "Target" column can be seen in Figure 1. We observe that most of the features are discrete; and continuous features, such as grades columns exhibit a Gaussian distribution pattern. "Target" values are encoded as "Graduate" => 2, "Enrolled" => 1, and "Dropout" => 0. The graduation rate among students is nearly three times higher than the enrollment rate and almost twice as high as the dropout rate.

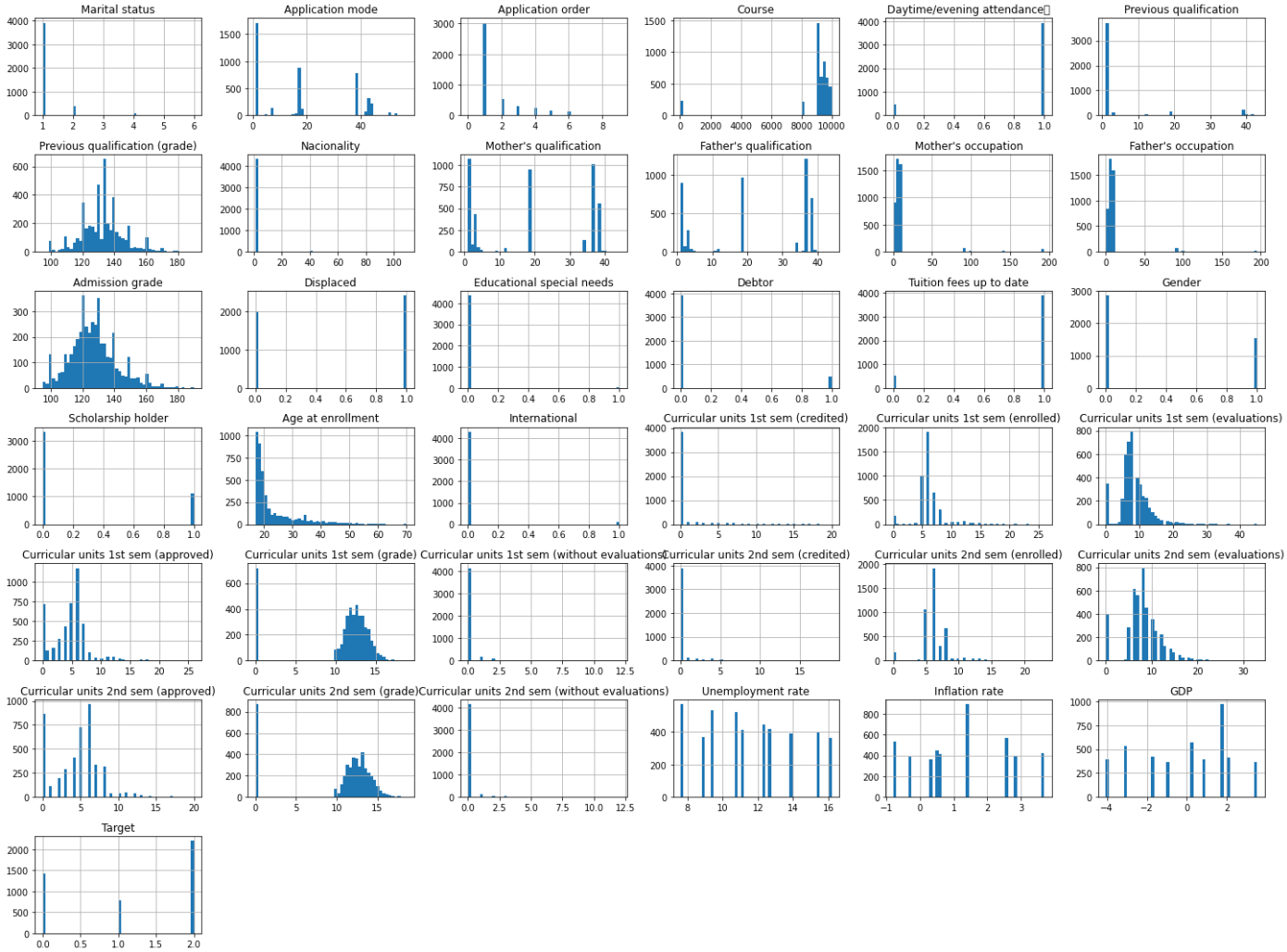

Fig 1. Frequency Distribution of the features and the “Target” column

In Table I, the columns that exhibit high correlation with “Target” column are presented. These features consist of grades and credits students acquired in the second and first semesters, tuition and scholarship information among others. Second semester’s correlation is higher than others, as finishing the second the semester is a good indicator of how invested students are in the program.

TABLE I. COLUMNS THAT ARE HIGHLY CORRELATED WITH TARGET

	Target Correlation
Curricular units 2nd sem (approved)	0.62
Curricular units 2nd sem (grade)	0.57
Curricular units 1st sem (approved)	0.53
Curricular units 1st sem (grade)	0.49
Tuition fees up to date	0.41
Scholarship holder	0.30
Age at enrollment	0.24
Debtor	0.24

	Target Correlation
Gender	0.23
Application mode	0.22
Curricular units 2nd sem (enrolled)	0.18
Curricular units 1st sem (enrolled)	0.16
Admission grade	0.12
Displaced	0.11
Previous qualification (grade)	0.10

III. Methodology

Before applying any classification model, I needed to preprocess the data. The authors who provided the data had conducted thorough preprocessing to handle anomalies, unexplainable outliers, and missing values. Therefore, the preprocessing steps I performed were limited to encoding the "Target" values, as mentioned earlier, scaling the data using a Standard Scaler, and splitting the data into training, validation, and test sets. Twenty percent of the data was allocated as the test set, another twenty percent as the validation set, and the remaining portion was used as the training set. The training set comprised 2831 instances, the validation set had 708 instances, and the testing set had 885 instances. All hyperparameter tuning was conducted on the training data and validated using the validation data. Finally, the tuned models were applied to the test set to evaluate their performance.

Four classification algorithms were employed: K Nearest Neighbors (KNN), Support Vector Machines (SVM), SGDClassifier (utilized as a Logistic Regression Classifier), and Random Forests. For each model, manual cross-validation was performed, and the average and standard deviation of the accuracy scores were computed to assess the presence of overfitting. Furthermore, grid search cross-validation was conducted to identify the best hyperparameters. Subsequently, the models were tuned using these optimal hyperparameters and applied to the test sets. Finally, statistical measures such as confusion matrices, and accuracy scores were generated for the validation and test sets. Additionally, the features used in the decision-making process were plotted for each model.

IV. Results and Discussion

This Section is divided into three parts:

A. Cross Validation Results

B. Best Hyperparameters and Most Relevant Features

C. Training, Validation, and Test Results along with Test Confusion Matrices

A. Cross Validation Results:

The results of cross-validation, conducted with 5 folds on the training and validation data, are presented in Table II. Among the evaluated classifiers, KNN, SGDClassifier, and Random Forest exhibit similar standard deviations, indicating a comparable range of fold accuracies. However, KNN has a significantly lower average accuracy compared to SGDClassifier and Random Forest. Random Forest achieves the highest accuracy at 78%. On the other hand, SVM shows the lowest standard deviation and the second-highest accuracy. This implies that SVM generalizes better than the other models with a relatively good accuracy. Once the models are tuned, we will reevaluate their performances.

TABLE II. CROSS VALIDATION RESULTS

	KNN	SVM	SGD	Random Forest
Average Accuracy	0.70	0.76	0.73	0.78
Standard Deviation	0.012	0.0049	0.0125	0.0111
95% confidence interval	[0.68 0.71]	[0.75 0.77]	[0.72 0.75]	[0.77 0.79]

B. Best Hyperparameters and the Most Relevant Features:

a. KNN:

KNN considers the distance between data points to make classifications. Distance measures, such as “Euclidean” or “Manhattan” can be employed. The algorithm measures the distance between a given point and the k nearest data points and determines the majority class among them and identifies the data point as belonging to that class. KNN’s “n_neighbors” parameter corresponds to k.

KNN’s “n_neighbors” parameter was tuned by determining the accuracy scores for training and validation sets for n values ranging from 1 to 25. From the results presented in Figure 2, n=8 and n=18 gave the best results.

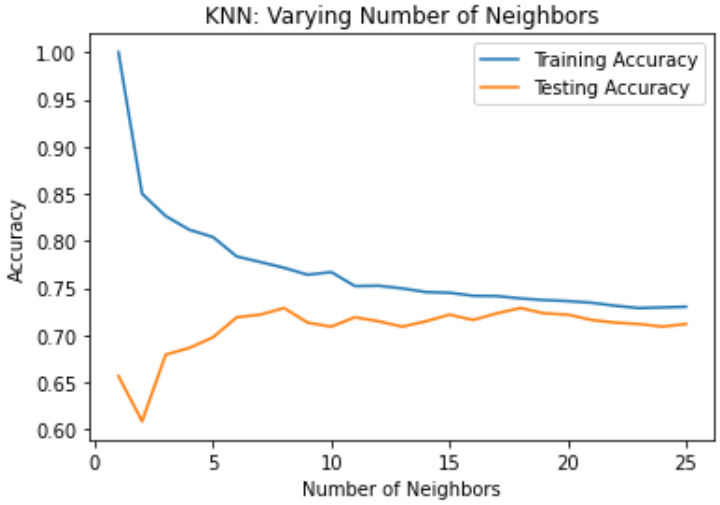

Figure 2. Training and Validation accuracies for number of neighbors ranging from 1 to 25.

From the grid search, we determined that the best parameters are n=18 and metric = Manhattan.

The top 10 most relevant features for KNN are Curricular units 2nd sem (approved), Tuition fees up to date, Debtor, Curricular units 2nd sem (evaluations), Scholarship holder, Curricular units 1st sem (approved), Curricular units 2nd sem (grade), Curricular units 2nd sem (credited), Curricular units 2nd sem (without evaluations), Application mode.

These along with less relevant features can be seen in Figure 3:

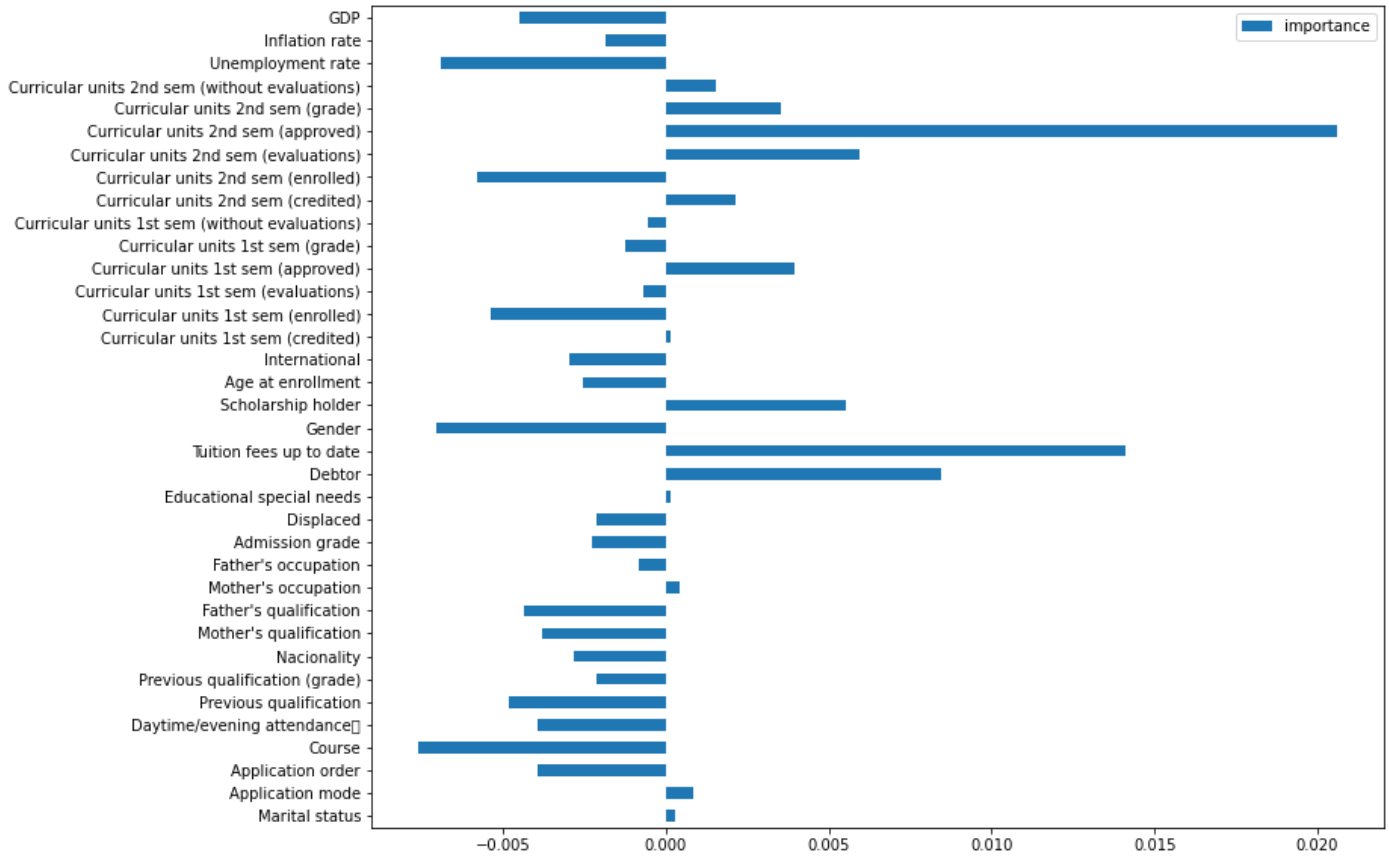

Figure 3. The most relevant features for KNN.

b. SVM:

SVM aims to find the optimal decision boundary that maximizes the margin between different classes. The data points that lie closest to that boundary are called support vectors and these are used in defining the boundary. A regularization parameter (C) is used to establish a balance between maximizing the margin and minimizing the classification error. Gamma parameter defines how each data point will influence the shape of the decision boundary.

Hyperparameters “C” and “gamma” were tuned. In figure 4, the accuracy scores for varying C and gamma values can be seen:

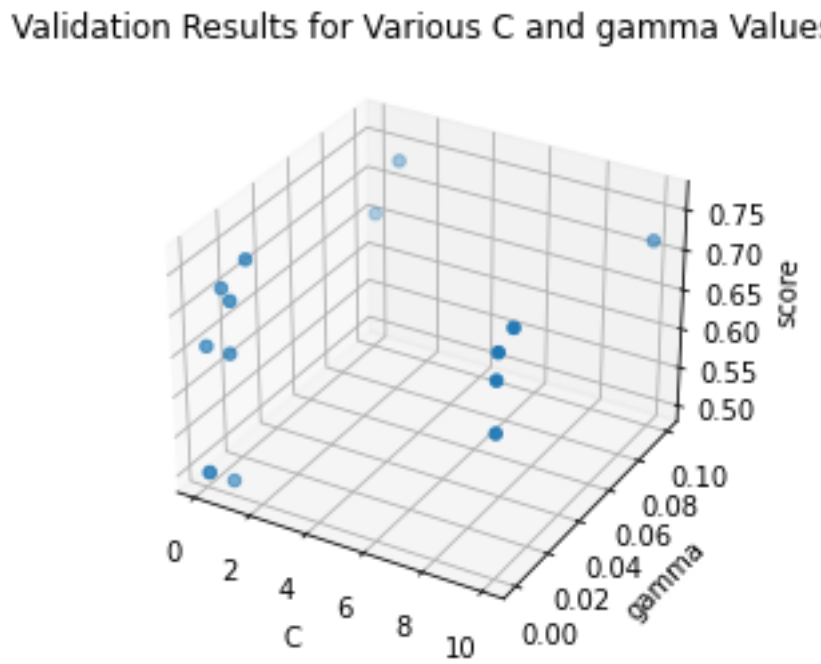

Figure 4. Accuracy Scores for varying C and gamma.

Best parameters were found to be 'C': 10, and 'gamma': 0.01 through grid search.

The top 10 most relevant features for SVM were Curricular units 2nd sem (approved), Curricular units 1st sem (approved), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (grade), Tuition fees up to date, Curricular units 1st sem (enrolled), Course, Age at enrollment, Curricular units 1st sem (credited), Unemployment rate as can be seen in Figure 5:

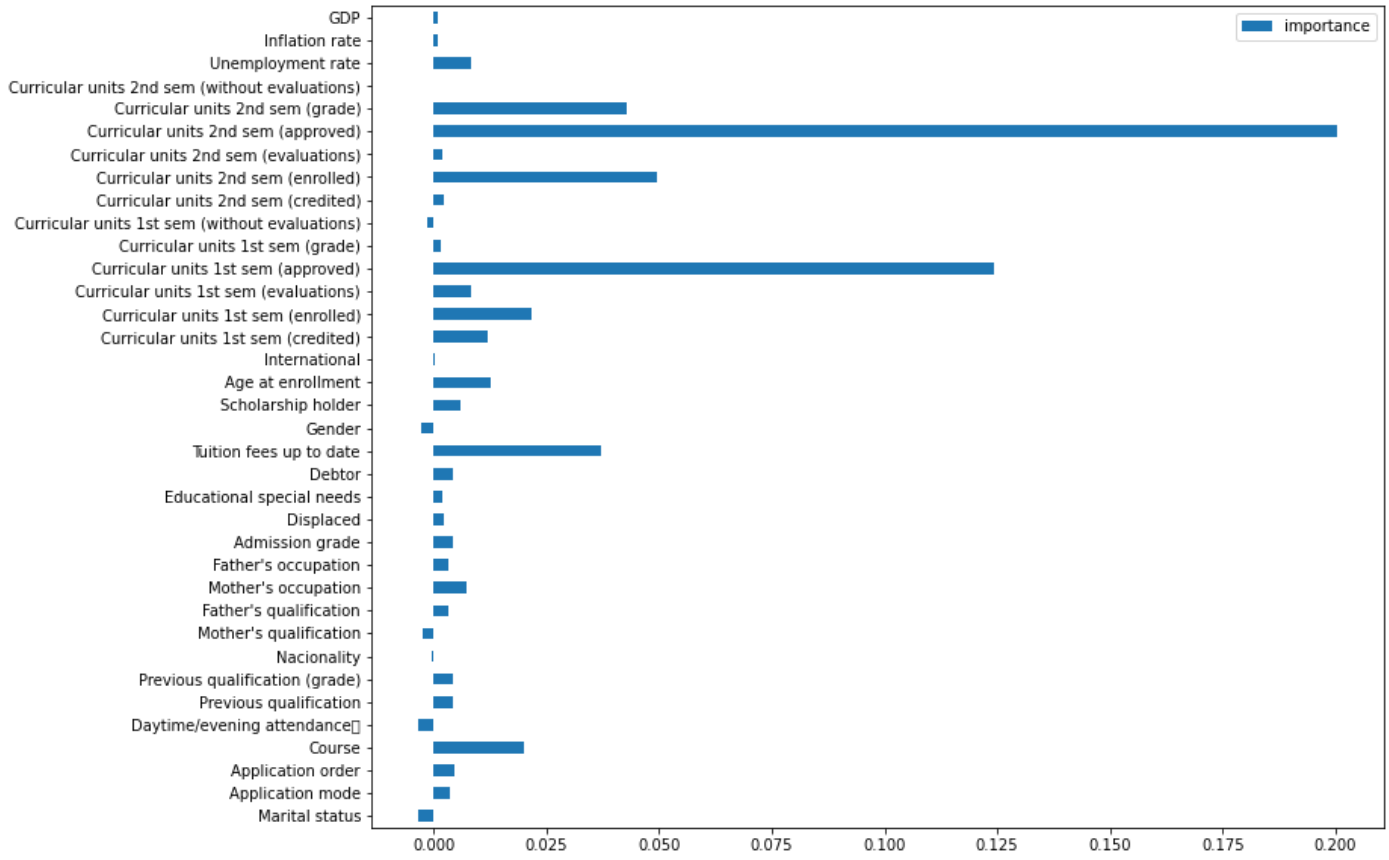

Figure 5. The most relevant features for SVM

c. SGD Classifier:

SGDClassifer acts like a linear SVM when using the hinge loss function. When using the log loss function, it behaves like Logistic Regression Classifier. Since I already performed SVM, I will use log loss function and run SGDClassifier as a Logistic Regression Classifier.

The model adjusts its parameters by minimizing the logistic loss using stochastic gradient descent optimization. The logistic loss measure the difference between the predicted probabilities and the true labels of the training instances. The model keeps updating the parameters until it reaches optimal values that minimizes the logistic loss. The model makes its decision by comparing the probabilities assigned to each class. If the probability of a certain class is higher than a predetermined threshold, the model predicts that the data point belongs to that certain class.

Alpha determines the strength of the regularization. Higher alpha provides stronger regularization. It helps finding the best trade-off between fitting the training data closely (overfitting) and finding a simpler model with better generalization.

The grid search found the best parameters as alpha=0.01 and penalty=l2. In figure 6 the accuracy scores for varying alpha values and different penalties (l1, l2) can be seen:

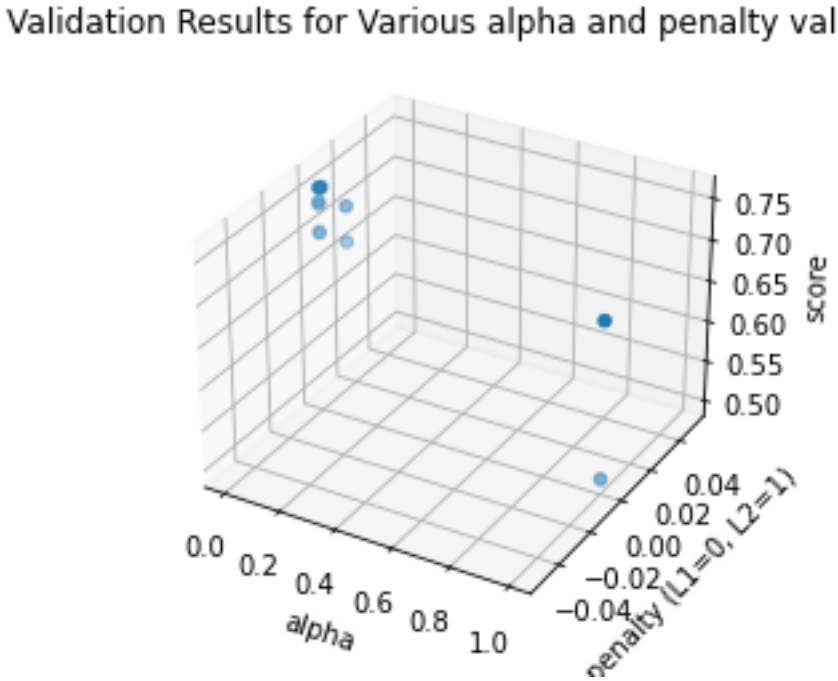

Figure 6. Accuracy scores for varying alpha and penalties.

The top 10 most relevant features are Curricular units 1st sem (approved), Tuition fees up to date, Curricular units 1st sem (enrolled), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Scholarship holder, Course, Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Curricular units 2nd sem (evaluations). Figure 7 shows all features:

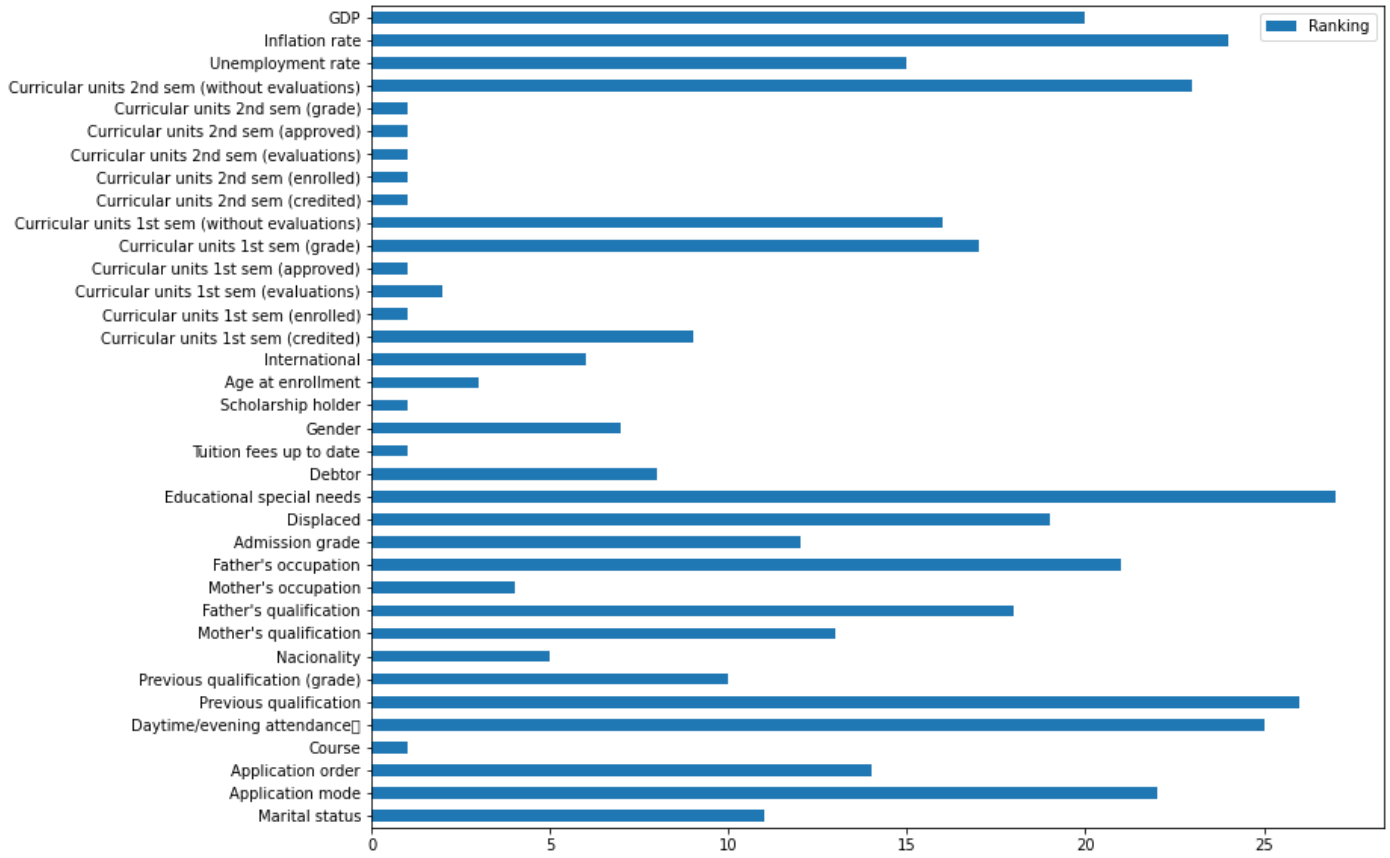

Figure 7. The most relevant features for SGDClassifier (Logistic Regression Classifier)

d. Random Forest

The final classifier I applied was Random Forest. Random Forest employs multiple decision trees to make its decisions. Each decision tree is constructed using a random subset of the training data. During prediction, each tree independently classifies the input and the final prediction is determined by majority voting of the individual tree predictions. The randomness helps to reduce overfitting and improve the generalization. Since it harnesses the ensemble approach, it is a very powerful model.

Grid search found the best hyperparameters to be 'max_depth': 12, 'min_samples_leaf': 2, 'n_estimators': 160.

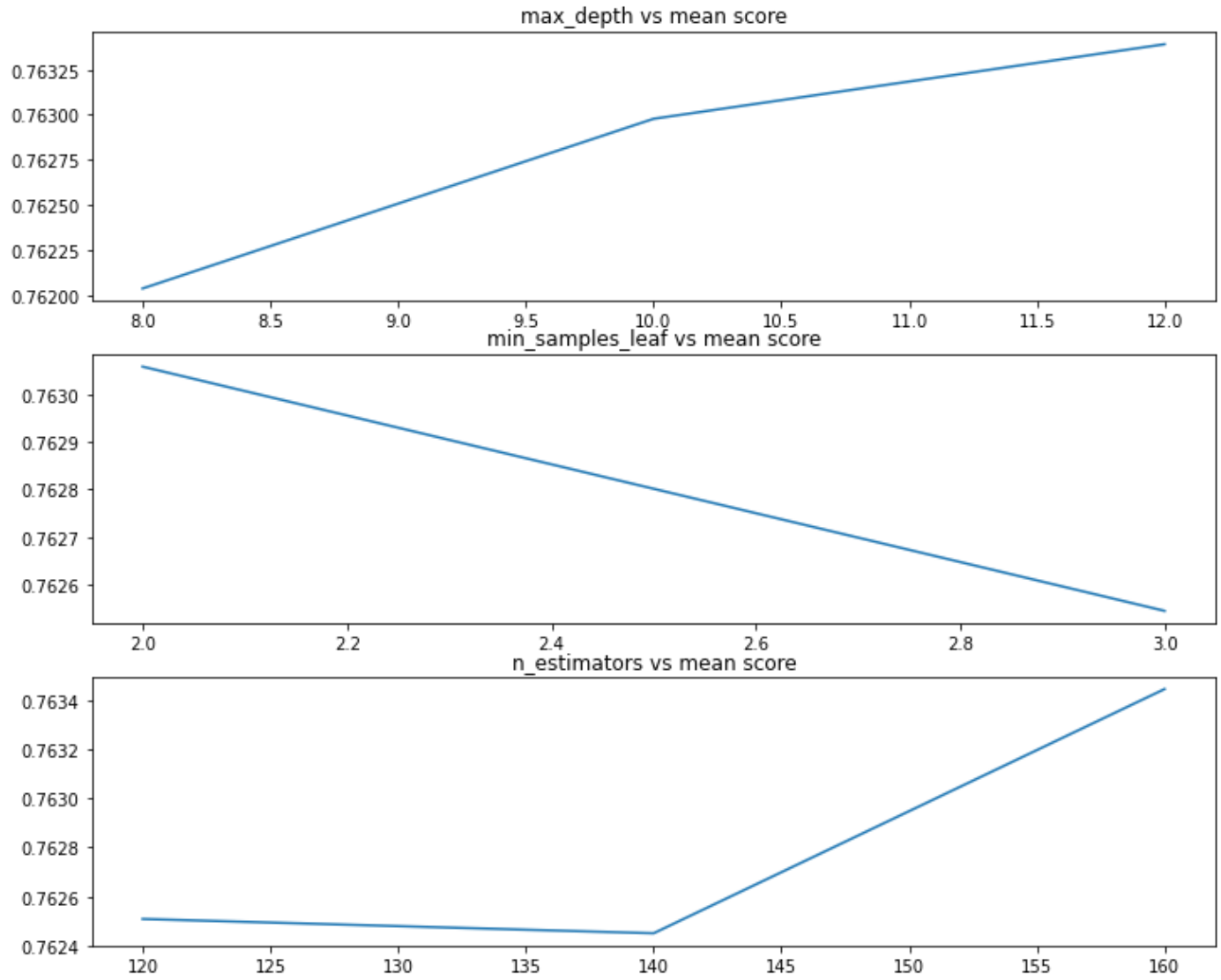

Figure 8. For each hyperparameter, the average accuracy scores are shown. Max_depth:[8, 10, 12], min_samples_leaf:[2,3] and n_estimators:[120, 140, 160]

The top 10 most relevant features for Random Forest were Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Curricular units 1st sem (approved), Curricular units 1st sem (grade), Tuition fees up to date, Curricular units 2nd sem (evaluations), Age at enrollment, Admission grade, Curricular units 1st sem (evaluations), Previous qualification (grade).

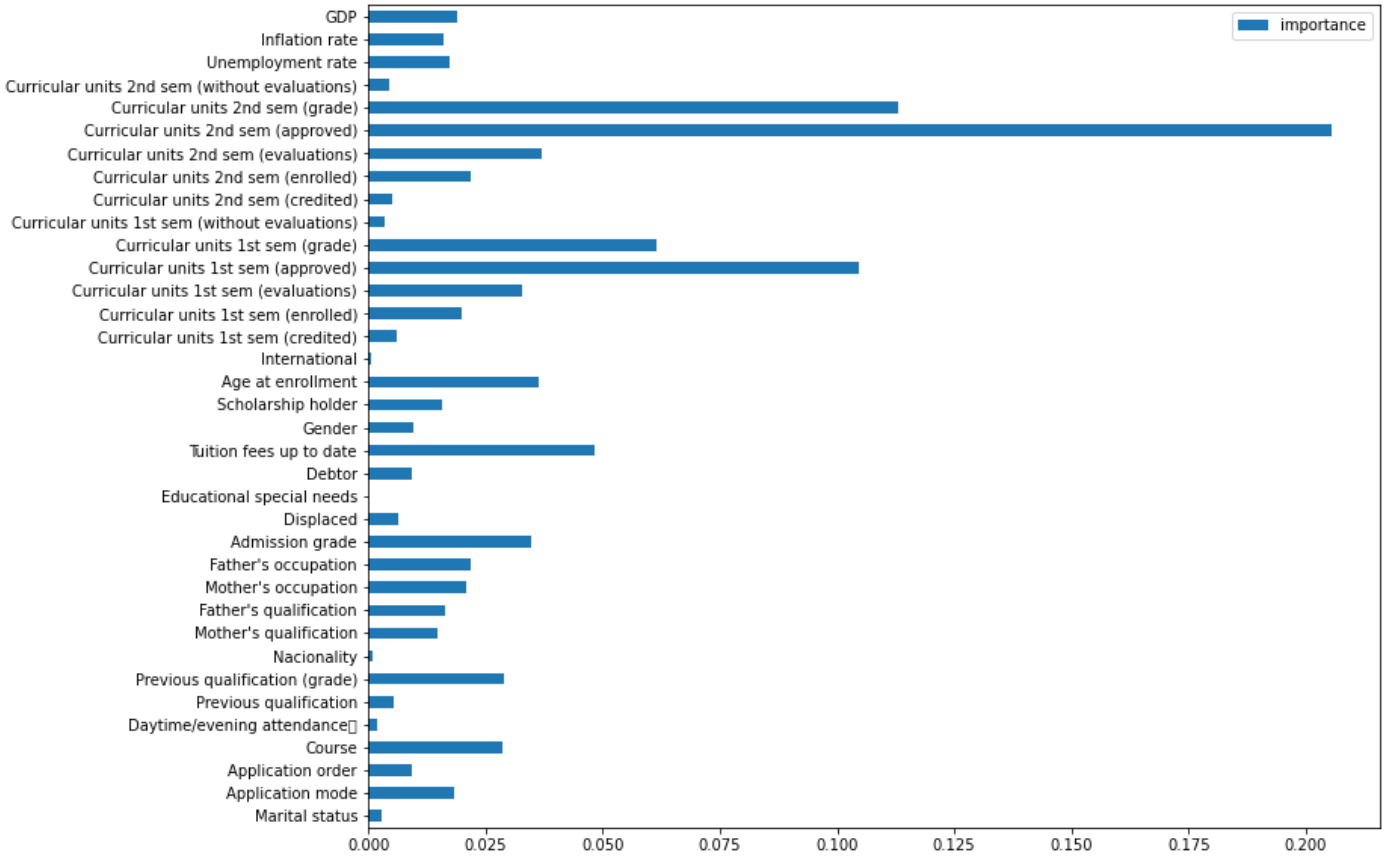

Figure 9. The most relevant features for Random Forest Classifier

Some features occurred in the most relevant top 15 for all classifiers. These are: Curricular units 2nd sem (approved), Tuition fees up to date, Curricular units 1st sem (approved), Curricular units 2nd sem (grade), Curricular units 1st sem (evaluations), Mother's Occupation.

C. Training, Validation, and Test Results along with Test Confusion Matrices:

From the results presented in Table III, we observe that KNN's accuracy is the lowest in training, validation and test sets. However, it does not overfit as much as Random Forest and SVM do. The least overfitting and the most accurate model in testing is SGD Classifier which behaves like a Logistic Regression Classifier. It identifies data points belonging to "Graduate" and "Dropout" classes more accurately than any other model. However, for "Enrolled" class, it does not perform as well. Considering the fact that none of the models do exceptionally well in this group, we can conclude that SGD Classifier is the best performing model among all.

TABLE III. TRAINING, VALIDATION AND TEST RESULTS

	KNN	**SVM**	**SGD**	**Random Forest**
Training Accuracy	0.75	0.85	0.77	0.94
Validation Accuracy	0.72	0.77	0.78	0.79
Testing Accuracy	0.71	0.74	0.76	0.76
F1 Score Dropout (0)	0.71	0.74	0.78	0.77
F1 Score Enrolled (1)	0.30	0.37	0.28	0.41
F1 Score Graduate (2)	0.80	0.83	0.85	0.84

Based on the presented heat maps below, we arrive at a similar conclusion: SGD identifies a higher percentage of "Dropout" and "Graduate" data points correctly, with accuracies of 25% and 47% respectively. However, it struggles significantly in classifying the "Enrolled" class. If the primary objective is to identify students who will dropout, the SGD Classifier would be the best option. On the other hand, if the goal is to identify students who will graduate, KNN actually achieves a higher accuracy for that class.

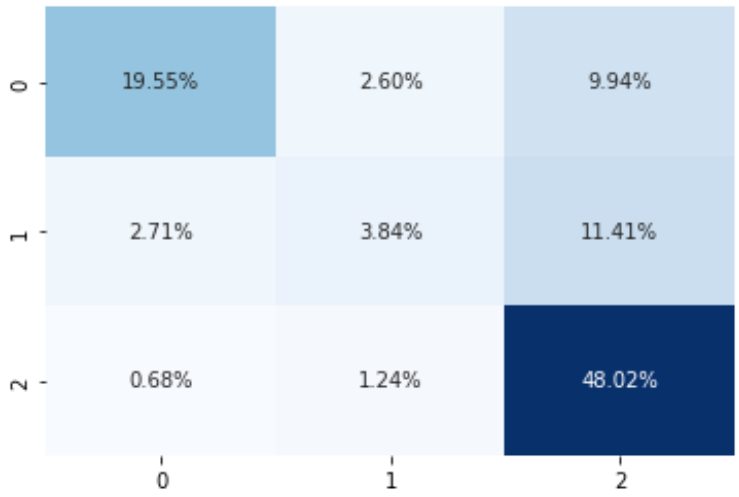

Figure 10. Heat map of test results for KNN

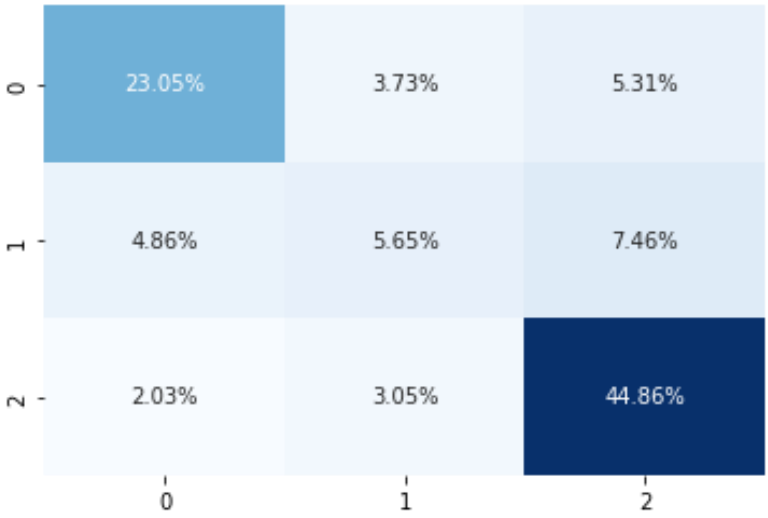

Figure 11. Heat map of test results of SVM

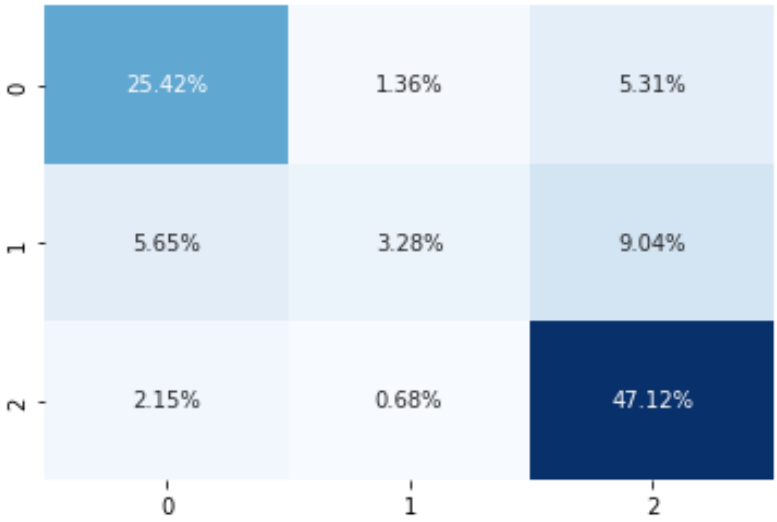

Figure 12. Heat map of test results of SGDClassifier

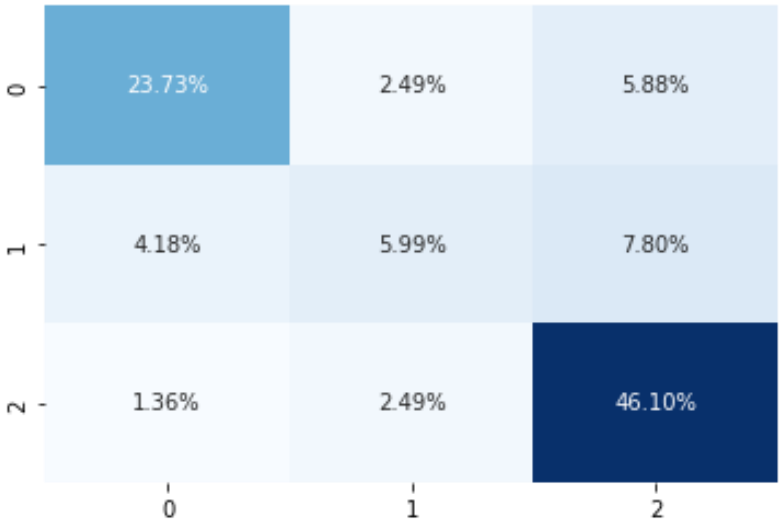

Figure 13. Heat map of test results of Random Forest.

V. Conclusions

In this paper, I employed four classification models to predict student outcomes, including dropout, graduation, and enrollment. I conducted cross-validation on the combined training and validation datasets for each model. Subsequently, I utilized grid search to identify the optimal hyperparameters for each model and fine-tuned them accordingly. I also identified the most influential features used by the models to make their predictions. Finally, I applied the tuned models to unseen data.

However, despite Random Forest yielding a high training accuracy of 94%, its performance declined when tested on validation and testing data. This indicates that Random Forest is prone to overfitting and is the most affected model among all. SVM also exhibits overfitting tendencies, although to a lesser extent compared to Random Forest. On the other hand, KNN provides the lowest accuracy among the models. The most generalizable model, with the highest testing accuracy, appears to be SGD Classifier, which behaves similarly to a Logistic Regression Classifier. While its testing accuracy of 76% is not exceptional, it still outperforms all other models.

All models achieved relatively satisfactory results with the "Dropout" and "Graduate" classes. However, they encountered challenges when classifying instances within the "Enrolled" class. This discrepancy may be attributed to the fact that the "Enrolled" class is not at the same level of exclusivity as the "Dropout" and "Graduate" classes. Dropout and graduation are mutually exclusive events, meaning a student who graduated did not drop out, and vice versa. However, the same exclusivity does not hold true for the "Enrolled" class. A student in the "Enrolled" class can either drop out or graduate, while students who dropped out or graduated

were once enrolled. If the dataset were divided into two classes, specifically "Graduate" and "Dropout," I believe the models would have achieved significantly higher accuracy.

VI. Appendix

TABLE IV. Dataset attributes and target

Attribute	Type	Example Value	Description
Marital status	Numeric (integer)	1	1 – single 2 – married 3 – widower 4 – divorced 5 – facto union 6 – legally separated
Application mode	Numeric (integer)	1	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)
Application order	Numeric (integer)	0	Application order (between 0 - first choice; and 9 last choice)
Course	Numeric (integer)	33	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)
Daytime/evening attendance	Numeric (integer)	1	1 – daytime 0 - evening
Previous qualification	Numeric (float)	1	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)
Previous qualification (grade)	Numeric (integer)	100	Grade of previous qualification (between 0 and 200)
Nacionality	Numeric (integer)	1	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
Mother's qualification	Numeric (integer)	1	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

Attribute	Type	Example Value	Description
			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)
Father's qualification	Numeric (integer)	1	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)
Mother's occupation	Numeric (integer)	0	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 -

Attribute	Type	Example Value	Description
			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
Father's occupation	Numeric (integer)	0	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
Admission grade	Numeric (float)	100	Admission grade (between 0 and 200)
Displaced	Numeric (integer)	1	1 – yes 0 – no
Educational special needs	Numeric (integer)	1	1 – yes 0 – no
Debtor	Numeric (integer)	1	1 – yes 0 – no

Attribute	Type	Example Value	Description
Tuition fees up to date	Numeric (integer)	1	1 – yes 0 – no
Gender	Numeric (integer)	1	1 – male 0 – female
Scholarship holder	Numeric (integer)	1	1 – yes 0 – no
Age at enrollment	Numeric (integer)	20	Age of student at enrollment
International	Numeric (integer)	1	1 – yes 0 – no
Curricular units 1st sem (credited)	Numeric (integer)	1	Number of curricular units credited in the 1st semester
Curricular units 1st sem (enrolled)	Numeric (integer)	0	Number of curricular units enrolled in the 1st semester
Curricular units 1st sem (evaluations)	Numeric (integer)	0	Number of evaluations to curricular units in the 1st semester
Curricular units 1st sem (approved)	Numeric (integer)	0	Number of curricular units approved in the 1st semester
Curricular units 1st sem (grade)	Numeric (float)	0	Grade average in the 1st semester (between 0 and 20)
Curricular units 1st sem (without evaluations)	Numeric (integer)	0	Number of curricular units without evaluations in the 1st semester
Curricular units 2nd sem (credited)	Numeric (integer)	0	Number of curricular units credited in the 2nd semester
Curricular units 2nd sem (enrolled)	Numeric (integer)	0	Number of curricular units enrolled in the 2nd semester
Curricular units 2nd sem (evaluations)	Numeric (integer)	0	Number of evaluations to curricular units in the 2nd semester
Curricular units 2nd sem (approved)	Numeric (integer)	0	Number of curricular units approved in the 2nd semester
Curricular units 2nd sem (grade)	Numeric (float)	0	Grade average in the 2nd semester (between 0 and 20)
Curricular units 2nd sem (without evaluations)	Numeric (integer)	0	Number of curricular units without evaluations in the 1st semester
Unemployment rate	Numeric (float)	10.80	Unemployment rate (%)
Inflation rate	Numeric (float)	1.40	Inflation rate (%)
GDP	Numeric (float)	1.74	GDP
Target	Ordinal	'Dropout'	Target. The problem is formulated as a three categories classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course

REFERENCES

[1] M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho. (2021) "Early prediction of student's performance in higher education: a case study" Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16

[2] Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89. [Accessed: June 14, 2023]