Predicting College Student Dropouts with Machine Learning

By Hanah Kim
Affiliation: Leonia High School

Abstract

In the U.S., 33% of college students do not graduate, with 24% dropping out in their first year [2]. Various factors contribute to these high dropout rates, with finances as a significant one: some sources estimate that as high as 51% of college students drop out because of a lack of funds, a statistic that is confirmed by the machine-learning models used in this experiment [6]. In this study the neural network and random forest classifier machine learning models are used to predict whether a student will graduate or drop out. The “Predict students dropout, academic success” dataset was used [4]. This dataset provides various pieces of information about students, including gender, age of enrollment, and dropout status. The random forest classifier had an accuracy of 92.05% and the neural network had an accuracy of 91.71%, with academic and finance related factors contributing most to the dropout rate predictions. The ability to predict college student dropout status using machine learning is crucial, as it allows colleges to recognize students at risk of dropping out, enabling institutions to provide additional resources.

Introduction

In the U.S. 33% of college students do not graduate, with 24% of college students dropping out in their first year [2]. It is estimated that 51% of college students cite financial difficulty as a cause of their drop out [6]. In this study a random forest classifier and neural network are used to analyze 34 factors that can be used to identify potential dropouts, such as approved credits, grade average, and tuition fees. By analyzing these factors with AI, colleges can better identify students likely to drop out.

Methods

In this study both a random forest classifier and neural network are used to predict whether a student would drop out or not. A random forest classifier is a machine learning model made up of decision trees, estimators that each make yes/no predictions based on the inputted information from a dataset, splitting and classifying the data. The random forest classifier averages these separate predictions to make a more accurate prediction. A neural network is a machine learning model that processes data, loosely inspired by how the human brain processes information. A neural network consists of an input layer that takes in the raw data, hidden layer(s) that transform the data, and an output layer that produces a final prediction. For both random forests and neural networks the user must specify the hyperparameters that control certain aspects of the model’s structure. Significant hyperparameters in this study include the use of “log_loss,” a type of loss that calculates the cross-entropy between the model’s predictions and that of the ground truth [5], and “adam,” an optimization algorithm used neural networks that modifies the parameters using stochastic gradient descent [3].

Explaining the Dataset

This study used the dataset “Predict students dropout, academic success,” sponsored by SATDAP-Capacitação da Administração Pública, Portugal. Information regarding number of curricular units taken, age at enrollment, and other factors affecting the probability of dropping out are included for the 4424 students in the dataset [4]. These 34 factors are listed alongside

the “Target” column, which classify students as either “Graduate,” “Dropout,” or “Enrolled.” The dataset is not balanced, as 50% of the students are graduates, 32% are dropouts, and only 18% are enrolled [4].

In the dataset each row represents one student and their relevant information.

Marital status	Application mode	Application order	Course	Daytime/evening attendance
1	8	5	2	1
1	6	1	11	1
1	1	5	5	1
1	8	2	15	1
2	12	1	3	0

Figure 1: Snippet of the dataset’s first 5 columns.
Marital status: 1 - single; 2 - married
Application mode: How students applied and were admitted, each number representing a different application method
Application order: (between 0 - first choice; and 9 - last choice)
Course: Course taken by the student, with each number representing a different course
Daytime/evening attendance: 0 - evening; 1 - daytime

Each column is a factor that influences the likelihood of dropout, except for the “Target” column that classifies a student as either a “Graduate” or a “Dropout.”

Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
0	10.8	1.4	1.74	Dropout
0	13.9	-0.3	0.79	Graduate
0	10.8	1.4	1.74	Dropout
0	9.4	-0.8	-3.12	Graduate
0	13.9	-0.3	0.79	Graduate

Figure 2: Snippet of the dataset’s last 5 columns.
Curricular units 2nd sem (without evaluations): Number of courses without evaluations in the 2nd semester
Unemployment rate: Unemployment rate of region (%)
Inflation rate: Inflation rate of region (%)
GDP: GDP of region
Target: Lists the student as either a “Dropout,” “Graduate,” or “Enrolled”

Experimental Setup

The dataset was imported to Google Colab and then modified. All rows containing the word “Enrolled” in the “Target” column were excluded from the dataset because all enrolled students

will either graduate or drop out eventually, causing them to resemble the target they will eventually become, making classification difficult.

Two modified datasets were created: “X,” which contains all the feature data, and “y,” which contains all the labels. To create the modified dataset “X,” the columns: "Marital status,” “Application mode,” “Course,” “Nationality,” “Mother’s occupation,” and “Father’s occupation” are removed because of their arbitrary assignment of values. For instance, in the“Mother’s occupation” column the occupation “Health professionals” is arbitrarily represented with the number 123, and “cleaning workers” is represented by 191. These columns are removed to prevent future conflicts in the random forest classifier; because the random forest classifier splits data into a group that is less than a certain value and a group that is greater than that value; when numerical values are arbitrary these splits do not function properly. The “Target” column was removed from the “X” dataset to create a new dataset that only contains factors testable by a random forest classifier and a neural network. This new dataset was then used to train and test the random forest classifier and neural network models.

The dataset “y” was created to contain only the “Target” column, which was used to test the accuracy of the models.

Results

The neural network was able to accurately predict whether a student would graduate or drop out 91.74% of the time. The optimal hyperparameters of the neural network were found to be:

```
params = { 'hidden_layer_sizes' : [4, 3, 2],
                    'activation' : 'tanh', 'solver' : 'adam',
                    'alpha' : 0.0, 'batch_size' : 10,
                    'random_state' : 1,'tol' : 0.0001,
                    'nesterovs_momentum' : True,
                    'learning_rate' : 'constant',
                    'learning_rate_init' : 0.01,
                    'max_iter' : 1000, 'shuffle' : True,
                    'n_iter_no_change' : 90, 'verbose' :
False }
```

Switching the solver from the default “sgd” solver to “adam” and changing the “hidden_layer_sizes” from [10] to [4, 3, 2] greatly increased the accuracy of the neural network.

The confusion matrix for the neural network is shown in Figure 3.

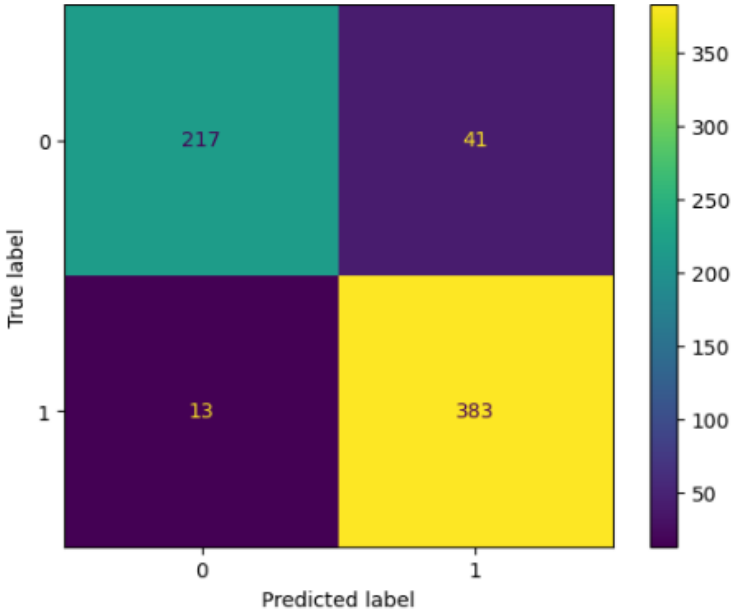

Figure 3: Confusion matrix for the neural network. The 0 label represents dropouts and the 1 label represents graduates.

As shown by the confusion matrix, the neural network was able to correctly guess “Dropout” 81.11% of the time and correctly guess “Graduate” 96.61% of the time. The discrepancy in accuracy may be explained by the different number of graduates and dropouts exposed to the neural network: the modified dataset was 60.55% “Graduate” and 39.45% “Dropout.” The neural network was trained more on “Graduate” data, a possible reason for its ability to better correctly guess “Graduate” than “Dropout.”

The random forest classifier was able to accurately predict whether a student would graduate or drop out 92.05% of the time.

The random forest classifier’s hyperparameters were left as their default values except for “n-estimators,” which were changed from 100 to 10000, and “criterion,” which was changed from “gini” to “log_loss.” This increased the accuracy of the random forest classifier by 2.75%, from 89.30% to 92.05%. The final random forest parameters were as follows:

(n_estimators = 10000, *, criterion = ‘log_loss’, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 'sqrt', max_leaf_nodes = None, min_impurity_decrease = 0.0, bootstrap = True, oob_score = False, n_jobs = None, random_state = None, verbose = 0, warm_start = False, class_weight = None, ccp_alpha = 0.0, max_samples = None)

The confusion matrix for the random forest classifier is shown in Figure 4.

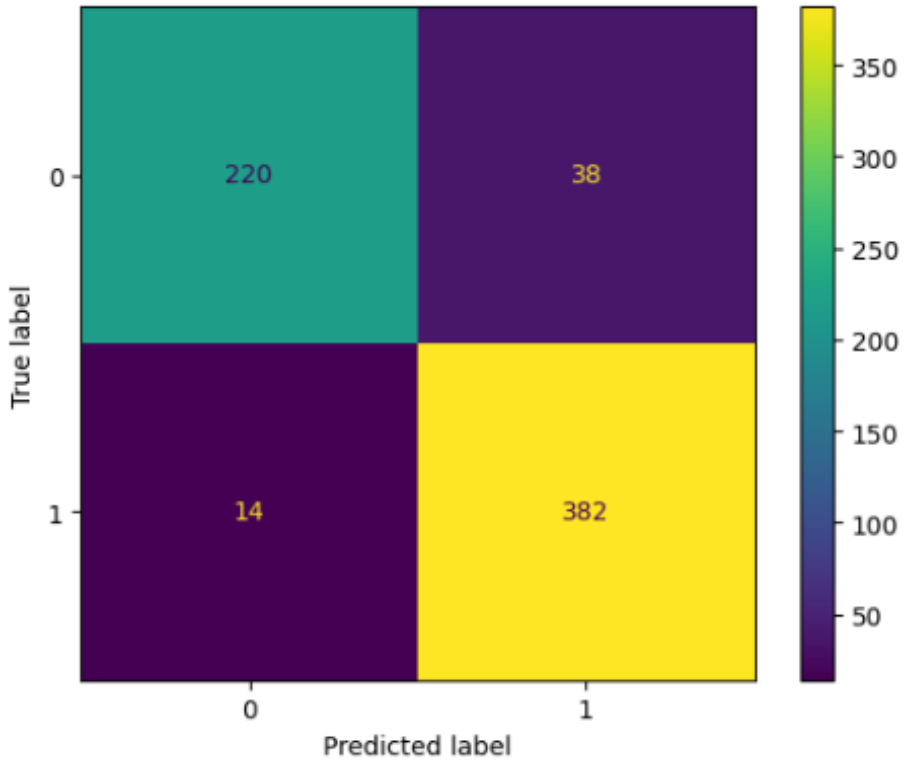

Figure 4: Confusion matrix for the random forest classifier. The 0 label represents dropouts and the 1 label represents graduates.

As shown by the confusion matrix, the random forest classifier was able to correctly guess “Dropout” 82.73% of the time and correctly guess “Graduate” 96.63% of the time. Similar to the neural network, the discrepancy in accuracy can be explained by the fact that the random forest classifier was trained with more “Graduate” data than “Dropout” data, so the model was better at correctly guessing “Graduate.”

Figures 5 -7 show examples of the decision trees created by the program.

The trees can be interpreted as follows: the color orange represents dropping out, and the color blue represents graduating. The greater the color saturation is, the greater the homogeneity of the data is. The percentage next to “samples” represents the percentage of samples that have followed the decisions leading to that node in the tree.
The “value” is a quantitative explanation of the color saturation/homogeneity of the data. The first number in the interval represents the percentage of “dropouts,” and the second number represents the percentage of “graduates” present in the data pool.

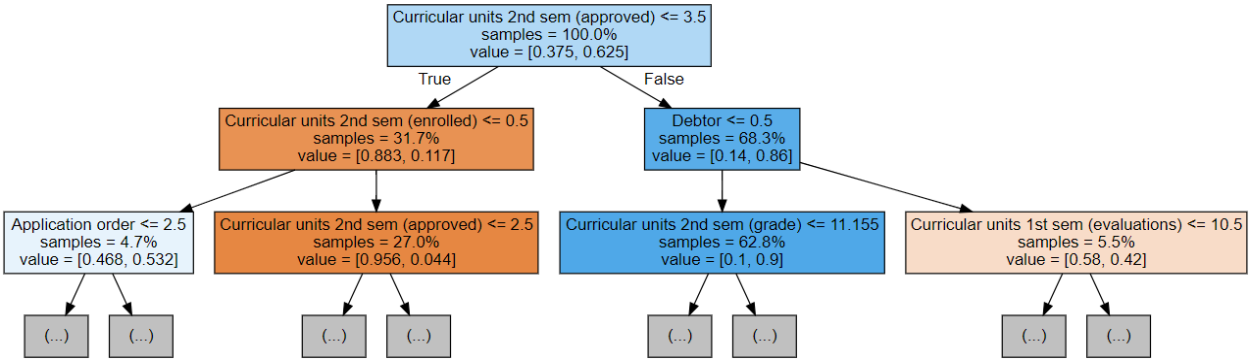

Figure 5: Example decision tree from the random forest classifier.
Only the results of the first 2 decisions are shown.
Shows the strength of “Curricular units 2nd sem (approved)”

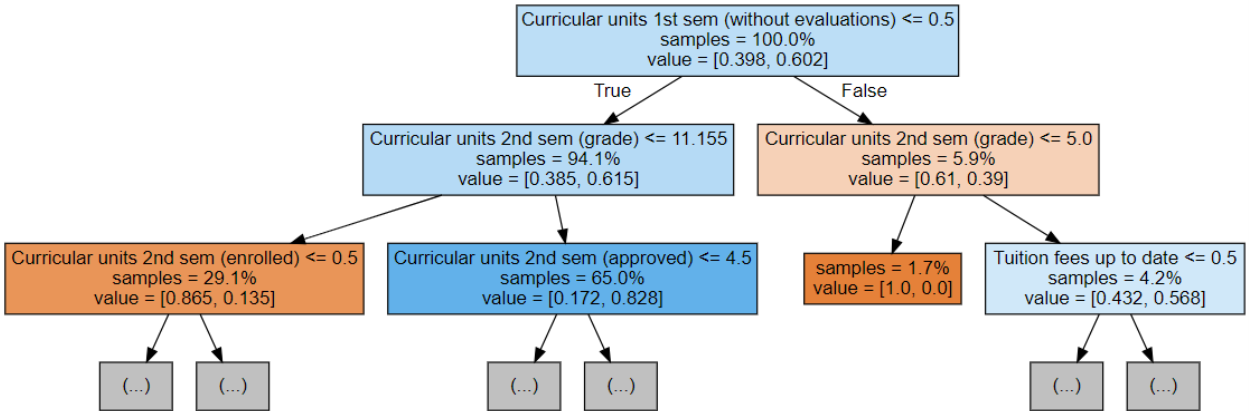

Figure 6: Example decision tree from the random forest classifier.
Only the results of the first 2 decisions are shown.
Shows the strength of “Curricular units 2nd sem (grade)” and the weakness of “Curricular units 1st sem (without evaluations)”

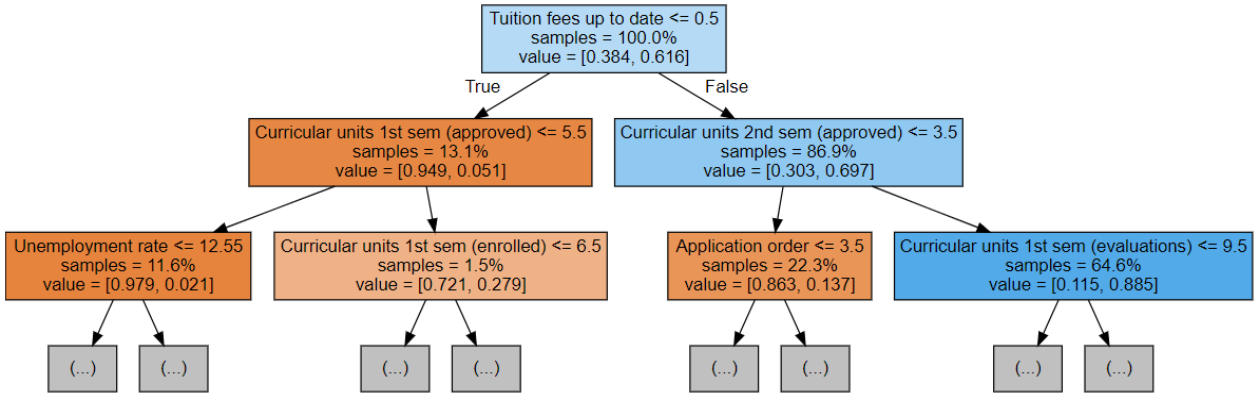

Figure 7: Example decision tree from the random forest classifier.
Only the results of the first 2 decisions are shown.
Shows the strength of “Tuition fees up to date”

The importance of each factor is given in Table 1:

Curricular units 2nd sem (approved)	0.202
Curricular units 2nd sem (grade)	0.126
Curricular units 1st sem (approved)	0.123
Curricular units 1st sem (grade)	0.079
Tuition fees up to date	0.056
Age at enrollment	0.045
Curricular units 2nd sem (evaluations)	0.039
Curricular units 1st sem (evaluations)	0.034
Curricular units 1st sem (enrolled)	0.030
Curricular units 2nd sem (enrolled)	0.030
Father's qualification	0.025
Mother's qualification	0.024
GDP	0.024
Unemployment rate	0.023
Scholarship holder	0.022
Inflation rate	0.021
Application order	0.018
Debtor	0.017
Gender	0.014
Displaced	0.009
Curricular units 1st sem (credited)	0.009
Curricular units 2nd sem (credited)	0.007
Previous qualification	0.007
Curricular units 1st sem (without evaluations)	0.005
Curricular units 2nd sem (without evaluations)	0.004

Daytime/evening attendance	0.003
International	0.002
Educational special needs	0.001

Table 1: Relative feature importances from the random forest classifier, in descending order, rounded to three decimal places.

Discussion

As shown by Table 1, the most significant four factors were related to academics: “Curricular units 2nd sem (approved)—number of credits approved in the 2nd semester,” “Curricular units 2nd sem (grade)—grade average in the 2nd semester (between 0 and 20), ” “Curricular units 1st sem (approved)—number of credits approved in the 1st semester,” and “Curricular units 1st sem (grade)—grade average in the 1st semester (between 0 and 20).” Perhaps these are the most significant factors because they indicate a student’s academic performance, which directly shows if a student is eligible for graduation. The fifth most important factor was related to a student’s finances: “Tuition fees up to date.” Although it was not found to be the factor with the most impact, it is important to note that it is the most significant factor not related to academics. The factors with the least impact on the model “Education special needs” and “International” have almost no impact on the random forest classifier, although this could be because there are very few international students and people with special needs in this study. The performance of the neural network and random forest classifier were similar, although the random forest classifier was 1.62% more accurate when predicting true dropouts while the neural network was 0.02% more accurate when predicting true graduates. The random forest classifier had an overall accuracy of 92.05%, making it more accurate than the neural network with an overall accuracy of 91.74%. The neural network’s accuracy could increase—potentially surpassing the random forest classifier’s accuracy—if more data is included in future studies.

It is important to acknowledge the possible ethical implications of using AI to predict student dropouts. Using AI in this way would require the collection of sensitive personal data such as academic records, and could lead to a decrease in trust between students and teachers [1]. There is also the possibility of false predictions due to racial and gender biases, causing professors to mistakenly target the wrong students.

Conclusion

Machine learning models can be used to better predict which students are likely to drop out, allowing colleges to be made aware of students who may need assistance.

For future studies, machine learning algorithms such as K-nearest neighbors and more combinations of hyperparameters can be tested to create a more accurate model. To increase the accuracy of the models more data should be obtained, especially for students categorized as “Dropout.” Additionally, the scope of the study should be increased—data from universities in different countries should be used to account for biases in the sample. It would be ideal to get more students classified as “International” or “Educational special needs” to account for the low number of these types of students in the current dataset.

Works Cited

1. Chiaramonte, Francesco. “Preventing Dropout: How AI Is Shaping the Future of Education.” *Francesco Chiaramonte*, 6 Mar. 2023, https://fchiaramonte.com/ai-to-prevent-dropout/.
2. “College Dropout Rate [2023]: By Year + Demographics.” *Education Data Initiative*, https://educationdata.org/college-dropout-rates. Accessed 3 Sept. 2023.
3. Gupta, Ayush. “A Comprehensive Guide on Optimizers in Deep Learning.” Analytics Vidhya (blog), October 7, 2021. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/.
4. *Predict Students Dropout, Academic Success*. https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success. Accessed 3 Sept. 2023.
5. scikit-learn. “Sklearn.Metrics.Log_loss.” Accessed September 4, 2023. https://scikit-learn/stable/modules/generated/sklearn.metrics.log_loss.html.
6. Scipioni, Jade. “51% of College Students Dropped out of School Due to Costs, Study Finds.” FOXBusiness, 9 Jan. 2018, https://www.foxbusiness.com/features/51-of-college-students-dropped-out-of-school-due-to-costs-study-finds.