Algorithm for Predicting Educational Performance in Portuguese Schools

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

It does not come as a surprise that education is of high importance. Not only can it empower an individual to do what they please and be successful, it also helps one potentially benefit society at large. It should then follow that improving education is of high importance as well. There are several ways to aid the education of people. The one way we are exploring primarily is how we can predict student performance. The ability to proficiently predict student performance would naturally allow us to find patterns in high and low performing students. Using these patterns, we can administer corrective measures towards lower performing students more feasibly. More specifically, we may be able to isolate what factors affect student performance, and perhaps more directly, we can simply apply a trained machine learning algorithm to find out which students may need academic assistance. That being said, a highly pertinent question would be what specific machine learning algorithm may be best suited for predicting student performance. Our study aims to help answer this question. Typically, data cleaning and adjusting items like hyperparameters make up a bulk of machine learning development. So, by performing these steps and comparing the results of various machine learning models, we can find which machine learning algorithm works best given our particular set of features. At the very least, we hope to provide a machine learning algorithm that can help predict student performance using the features that were provided to us in our chosen data set. (SHORT DESCRIPTION OF THE PROBLEM) yadda yadda yadda. (CONTRIBUTION) Our contribution to this line of inquiry is . Therefore, In this study we cleaned the data and applied several machine learning algorithms to see which one was best at predicting the final grade of the students. The machine learning algorithms we used were k-nearest neighbors (KNN), Decision Tree (DT), linear regression (LD), support vector machines (SVM), and artificial neural networks (ANN).

# DESCRIPTION OF THE DATASET

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# RESULTS

Five different supervised machine-learning algorithms were used to evaluate the training and testing datasets. These machine-learning algorithms were k-nearest neighbors (KNN), decision tree (DT), logistic regression (LR), support vector machines (SVM), and multilayer perceptron (). These algorithms are already publicly available. They were obtained through scikit learn’s public Python package.

Our results are tabulated in the following page.

## K-Nearest Neighbors Results

TABLE 1. MATH CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.78 | 0.75 | 0.76 | 28 |
| **C (Sufficient)** | 0.64 | 0.70 | 0.67 | 46 |
| **F (Poor)** | 0.76 | 0.70 | 0.73 | 37 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.71 | 111 |
| **Macro avg** | 0.73 | 0.72 | 0.72 | 111 |
| **Weighted avg** | 0.72 | 0.71 | 0.71 | 111 |

TABLE 2. PORTUGUESE CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.67 | 0.72 | 0.70 | 54 |
| **C (Sufficient)** | 0.73 | 0.76 | 0.75 | 100 |
| **F (Poor)** | 0.70 | 0.50 | 0.58 | 28 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.71 | 182 |
| **Macro avg** | 0.70 | 0.66 | 0.67 | 182 |
| **Weighted avg** | 0.71 | 0.71 | 0.71 | 182 |

## Decision Tree Results:

TABLE 3. MATH CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.75 | 0.75 | 0.75 | 28 |
| **C (Sufficient)** | 0.63 | 0.59 | 0.61 | 46 |
| **F (Poor)** | 0.68 | 0.73 | 0.70 | 37 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.68 | 111 |
| **Macro avg** | 0.68 | 0.69 | 0.69 | 111 |
| **Weighted avg** | 0.67 | 0.68 | 0.67 | 111 |

TABLE 4. PORTUGUESE CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.64 | 0.70 | 0.67 | 54 |
| **C (Sufficient)** | 0.76 | 0.66 | 0.71 | 100 |
| **F (Poor)** | 0.61 | 0.79 | 0.69 | 28 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.69 | 182 |
| **Macro avg** | 0.67 | 0.72 | 0.69 | 182 |
| **Weighted avg** | 0.70 | 0.69 | 0.69 | 182 |

## Logistic Regression:

TABLE 5. MATH CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.81 | 0.79 | 0.80 | 28 |
| **C (Sufficient)** | 0.69 | 0.67 | 0.68 | 46 |
| **F (Poor)** | 0.74 | 0.78 | 0.76 | 37 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.74 | 111 |
| **Macro avg** | 0.75 | 0.75 | 0.75 | 111 |
| **Weighted avg** | 0.74 | 0.74 | 0.74 | 111 |

TABLE 8. PORTUGUESE CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.73 | 0.69 | 0.70 | 54 |
| **C (Sufficient)** | 0.75 | 0.79 | 0.77 | 100 |
| **F (Poor)** | 0.73 | 0.68 | 0.70 | 28 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.74 | 182 |
| **Macro avg** | 0.74 | 0.72 | 0.73 | 182 |
| **Weighted avg** | 0.74 | 0.74 | 0.74 | 182 |

## Support Vector Machines:

TABLE 7. MATH CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.83 | 0.71 | 0.77 | 28 |
| **C (Sufficient)** | 0.66 | 0.72 | 0.69 | 46 |
| **F (Poor)** | 0.76 | 0.76 | 0.76 | 37 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.73 | 111 |
| **Macro avg** | 0.75 | 0.73 | 0.74 | 111 |
| **Weighted avg** | 0.74 | 0.73 | 0.73 | 111 |

TABLE 8. PORTUGUESE CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.76 | 0.59 | 0.67 | 54 |
| **C (Sufficient)** | 0.68 | 0.89 | 0.77 | 100 |
| **F (Poor)** | 0.89 | 0.29 | 0.43 | 28 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.71 | 182 |
| **Macro avg** | 0.78 | 0.59 | 0.62 | 182 |
| **Weighted avg** | 0.74 | 0.71 | 0.69 | 182 |

## Multilayer Perceptron:

TABLE 9. MATH CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.76 | 0.79 | 0.77 | 28 |
| **C (Sufficient)** | 0.61 | 0.67 | 0.64 | 46 |
| **F (Poor)** | 0.74 | 0.62 | 0.68 | 37 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.68 | 111 |
| **Macro avg** | 0.70 | 0.69 | 0.70 | 111 |
| **Weighted avg** | 0.69 | 0.68 | 0.69 | 111 |

TABLE 10. PORTUGUESE CLASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **B/A (Good)** | 0.60 | 0.59 | 0.60 | 54 |
| **C (Sufficient)** | 0.69 | 0.72 | 0.70 | 100 |
| **F (Poor)** | 0.67 | 0.57 | 0.62 | 28 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.66 | 182 |
| **Macro avg** | 0.65 | 0.63 | 0.64 | 182 |
| **Weighted avg** | 0.66 | 0.66 | 0.66 | 182 |

# CONCLUSIONS

Using these five algorithms, good results were given back with no accuracy below 66%. The Portuguese class dataset under the Support Vector Machines algorithm is the only set of data that seems unreliable. The recall on the “F (Poor)” label was only 29%. This is probably due to the lower sample size of 28 in comparison to the “C (Sufficient)” label which has a sample size of 100.

The algorithm that had the highest accuracy on both the math class and Portuguese class datasets, was the logistic regression dataset. Despite that, the k-nearest neighbors, decision tree, and multilayer perceptron algorithms are also all viable to use when studying student performance using similar features. We recommend, though, that if you can use logistic