Algorithm for Predicting Educational Performance in Portuguese Schools

Karan Swansi  
*School of Computing  
Southern Adventist University*Collegedale, US  
kswansi@southern.edy

Samuel Rivera  
School of Computing  
*Southern Adventist University*Collegedale, US  
srivera@southern.edy

*Abstract*— Education is an important part of our lives today. Thus, it stands to reason that being able to better educate is a desirable outcome. Using machine learning, we can find better ways to predict student performance. Machine learning was applied to datasets collected from two Portuguese schools. We used five machine learning algorithms and found four to generally be accurate. The best accuracy score was found logistic regression. This can be useful because it can help educators act for students at risk of poor academic performance, and perhaps future researchers may be able to determine factors that seem to affect academic performance the most*.*

Keywords—machine learning, logistic regression, student performance, education

# Introduction

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. This should assist future researchers when trying to predict educational outcomes with a similar or identical 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.

# DESCRIPTION OF THE DATASET

The particular data set we used was found online. It was of two Portuguese schools and contained around 30 attributes such as the student grades, school, demographic, social, and school related features. The data was collected using school reports and questionnaires. Overall, there were 649 instances in the data set. With grades from both math and Portuguese classes.

The data engineering process was relatively straightforward because no instance contained any faulty values or had any missing data.

We proceeded to encode categorical data such as sex, parental cohabitation status, parents’ education level, etc. as integers to make training the algorithms feasible.

However, we initially obtained poor results. After inspecting the data, we noticed that some of our target values (the grades “A”, “B”, “C”, and “F”) had a huge imbalance in the number of instances of certain categories that we were trying to predict. To be precise, letter grades “B”, “C”, and “F” were in significantly greater number than the “As. Noticing that this could affect the quality of our machine learning algorithms, we grouped some labels together. “F” letter grades were aggregated into a category that we called “Poor”, “C” letter grades were “Sufficient”, and “B” and “A” letter grades were “Good”. This way we achieved a much higher performance across our algorithms.

Finally, we experimented with how much of the data set was going towards our training data set and testing data set. We found that 72% of the data set going towards training and 28% of the data set going towards testing produced the best possible results.

# 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 (MLP). 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

These five algorithms yielded high accuracy; every algorithm had an accuracy above 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 algorithm. 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, to use logistic regression because it provides the highest accuracy. Machine learning algorithms are not deterministic, so despite them only being around 70% accurate, they can still be used to sufficiently predict student performance.