USING MACHINE LEARNING CLASSIFIERS IN A STUDENT HABITS AND DEMOGRAPHICS VS PERFORMANCE DATASET

Rafael Marques Braga

University of Miami
rxm403@miami.edu

ABSTRACT

This paper will focus on the application of three main machine learning models in a student performance dataset. It will focus on the decision tree, neural network, and K-nearest neighbor algorithms. The data set studied includes attributes such as students' demographics, habits and past school performance. The classification algorithms studied attempt to mine data from the dataset, being able to accurately classify the student on a pass/fail basis for a portuguese class purely based on the attributes studied.

Index Terms— Data Mining, KNN, Neural Network, Decision Tree, Student performance

1. INTRODUCTION

Portugal's public education system lies on the bottom of EU's public education system due to its high failing rate among its students. [1] Even though it has drastically improved in the past decades, Portugal still has its lower passing rate than average, being an area for improvement for the teacher's and professors in the country.

A hypothesis can be created that students' demographics, such as where they live (rural or urban areas) and where they go to school, their habits, such as average study time, average absences in past years of school and how much free time they have, can have an impact on their school performance and whether they end up passing or failing a high school class due to their every day patterns.

A dataset was formed and will be studied to see if such as pattern can be found to judge the student's performance based on their demographics and habits. If such pattern is found, this can bring high school teachers the prediction of which students will pass or fail their class and allow the educators to put these students on alert to make sure they can better succeed in their classes.

To look for such patterns in the students lives, a dataset will be studied and analyzed using three main data mining algorithms: Decision trees, K-nearest neighbors (KNN) and Neural Networks. By the end of the study, the accuracy of these models will be compared to reach the conclusion of which, if any, algorithm is a best fit to analyze this data. If a high accuracy is reached, teachers can easily use data mining models and data acquired from surveys in the start of semesters to have a higher passing rate in their classes.

2. THE DATASET AND ITS PRE-PROCESSING

The original dataset used contains 30 attributes listed on table 1 along with their descriptions for 2 classes, Portuguese and Mathematics [2]. However, not all attributes were used for the algorithms studied due to their low impact on adding useful data to the model. Different attributes were studied on each model with the final goal of achieving the highest accuracy possible as a result. All the results listed are also the results of the Portuguese dataset since this dataset contained more students and were able to provide a better learning rate for the algorithm and therefore better results.

Most of the data in the model happened to be recorded in a nominal way, mostly being binary results such as "yes" or "no". These were all transformed into numeric data by assigning numeric values (dummy variables) to each attribute. For example, yes or no attributes were assigned a 1 and a 0 for their attributes respectively. The data was also normalized using a min-max normalization, so all the data was represented from [0, 1].

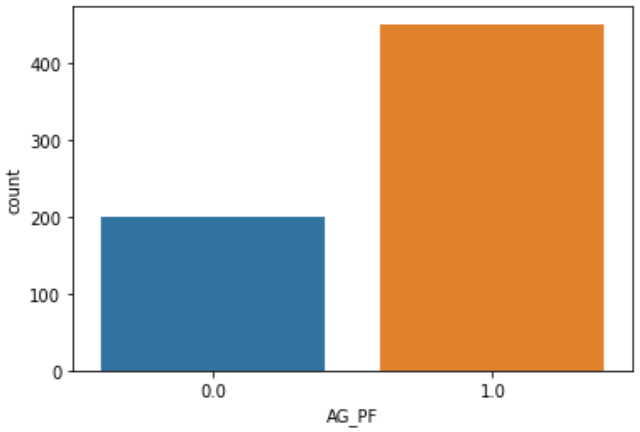
As for the results, in the original dataset, 3 grades were presented, G1, G2 and G3. For the classification algorithm, G1, G2 and G3 were averaged out and then classified into a failing/passing rate. According to the Portuguese education grading system, if a student receives a grade higher than 10 (out of 20) as their final grade, the student will pass the subject [3]. A one was assigned as a passing grade and a 0 was assigned as failing the class for the dataset. This was created as a new column ("AG_PF" --> AverageGrade_PassFail). The columns G1, G2 and G3 were then deleted from the data set.

Table 1. Data attributes

Attribute	Description
School	Student's school (binary: gabriel pereira or mousinho da silveira)
Sex	Student's sex (binary: "f" - female or "m" - male)
Age	Student's age (numeric: from 15 to 22)
Address	Student's home address type (binary: "u" - urban or "r" - rural)
Famsize	Family size (binary: "le3" - less or equal to 3 or "gt3" - greater than 3)
Pstatus	Parent's cohabitation status (binary: "t" - living together or "a" - apart)
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	Mother's job (nominal: "teacher", "health" care related, civil "services" (e.g., Administrative or police), "at_home" or "other")
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fjob	Father's job (nominal: "teacher", "health" care related, civil "services" (e.g., Administrative or police), "at_home" or "other")
Guardian	Student's guardian (nominal: "mother", "father" or "other")
Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Reason	Reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
Traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. To 1 hour, or 4 - >1 hour)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
Schoolsup	Extra educational support (binary: yes or no)
Famsup	Family educational support (binary: yes or no)
Activities	Extra-curricular activities (binary: yes or no)
Paidclass	Extra paid classes within the course subject (math or Portuguese) (binary: yes or no)
Internet	Internet access at home (binary: yes or no)
Nursery	Attended nursery school (binary: yes or no)
Higher	Wants to take higher education (binary: yes or no)
Romantic	With a romantic relationship (binary: yes or no)
Freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
Gout	Going out with friends (numeric: from 1 - very low to 5 - very high)
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Health	Current health status (numeric: from 1 - very bad to 5 - very good)
Absences	Number of school absences (numeric: from 0 to 93)

It is also good to notice that the dataset used is a bit unbalanced. Looking at the dataset, there are around 450 passing students in total and 199 failing students (graph 1). This means that if an algorithm were to guess that all students would pass the subject, it would be accurate around 69% of the time.

Graph 1. Data Distribution



3. DECISION TREES

The first algorithm implemented was a decision tree. This algorithm was done using the scikit-learn library. This is a machine learning library used in python. After randomizing the data and splitting the dataset into its training and testing datasets using a 80:20 split for each respectively, the decision tree was formed and trained using the DecisionTreeClassifier model.

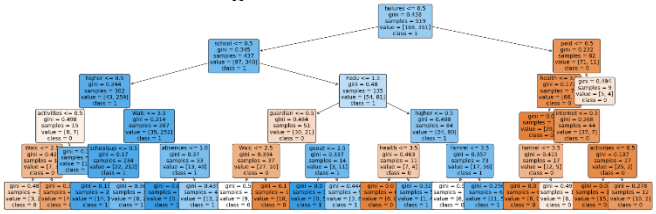
After multiple trials, a maximum depth of 5 was defined to be the best one, reaching the best results.

This model had an average accuracy of around 76% after multiple tries of different attributes being used and different randomized training and testing sets being created. The highest accuracy score using a data tree was of 79.23% using 27 of the 30 attributes given in the dataset.

The attributes which were given the most weight and most used in the data tree in decreasing order were Failure, school, higher, activities and Walc.

It is important to note that this model never reached below 70% accuracy unless very few attributes were used. This means that this model is always better than purely guessing the students as all passing.

Figure 1. Decision Tree



4. K-NEAREST NEIGHBOR CLASSIFIER

The second algorithm to be used was the KNN classifier. This algorithm was implemented from scratch using MATLAB. This algorithm also contained Tomek Links removal to try to remove outliers from the training data set. This removes outliers when there is an instance in the middle of a opposite majority class. This cleans the data and improves the accuracy of the algorithm. The pseudocode of the KNN model used is shown in Figure 2.

Figure 2: KNN Algorithm Pseudocode

Algorithm 1: Brute force kNN Algorithm

Input : $\mathcal{Q}$, a set query points and $\mathcal{R}$, a set of reference point;
Output: A list of k *reference* points for each *query* point;

```
foreach query point q ∈ Q do
    compute distances between q and all r ∈ R;
    sort the computed distances;
    select k-nearest reference points corresponding to k smallest distances;
```

Using the KNN classifier a maximum accuracy of 77% was reached. This was reached the same training and testing set as used in the decision tree. In the final result, 35 examples were removed from the training set.

The final confusion matrix for the KNN algorithm was the following.

Table 2. Confusion Matrix for KNN

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	91 (TP)	22 (FP)
	Negative	9 (FN)	8 (TN)

It also gave out results for recall being around 92%. This is good as the algorithm only predicted a few students as failing the class when they were actually passing.

When looking at precision however, the result is a bit lower (around 81%). In this case, it is beneficial for the precision to be higher as this is what shows the prediction of students as passing when they were failing the class.

5. NEURAL NETWORKS

A neural network was also built to continue the study using this dataset. To build the neural network, pytorch was used in python. Google colab was also used in this model to provide the training for the model. Since neural networks are harder to train, Google provides the free use of their GPUs to train models faster and efficiently.

The final model used a bit different dataset, only using the binary data. Only using the binary data (binary attributes) provided better results. A theory is because, even though all attributes were normalized between 0 and 1, the binary attributes would provide a higher weight to the data since they are separated further apart more often (they can only be 0 or 1 instead of anything in between). Therefore, these were always the most important attributes.

Since every different training would provide a different accuracy for the model, the average accuracy for the NN algorithm was in between 76%. However, some results had an accuracy as low as 72% depending how the model was trained, and as high as 79% accuracy.

The neural networks were mostly created and studied using 5 hidden layers, 9 epochs and a learning rate of 0.002. This, however, was changed a lot to fine tune the algorithm and try to get the best possible result.

Any number of epochs higher than 4 would not change much the result as the average loss of data would decline and stay more or less constant after the 4th epoch. The learning rate did not seem to change much the results as well, as long as the learning rate did not change much from 0.002 (±0.001).

6. CONCLUSION & FURTHER STUDIES

After studying the data set provided and trying it with multiple algorithms, the average result of around 78% accuracy was reached. The results summarized in one table can be found in table 3.

Table 3. Accuracy Results

Accuracy of the Different Algorithms		
KNN	Decision Tree	Neural Network
77%	79%	77%

Even though all algorithms did perform better than average, as all of them did about 10% better than purely guessing as all students as passing the class, using this data set, it proves that students habits, demographics and past performance cannot fully determine their future performance in class when looking at the Portuguese school system.

When other algorithms were quickly examined as well using MATLAB's ClassificationLearner function, none of the algorithms trained were also able to surpass the 80% threshold, further implying that either there was not enough data for a data mining model to be useful or there were not enough patterns in the data. This included other algorithms such as SVMs, Ensemble and Kernel.

There can be multiple reasons why this is true, the main one is just because not enough patterns were found to form an accurate learning model on the students.

To continue with this study, further data can be gathered, from multiple classes and thousands of students. Some of the attributes can also be changed as they would prove to be more useful. Attributes such as grades from past classes and how enjoyable the subject is from the students can be attributes which make a huge impact to the algorithm as this has a higher correlation of how a student does in class when compared to attributes like the student's age or parent's job.

7. REFERENCES

[1] "The Study.EU Country Ranking 2018 for International Students" Study.eu. https://www.study.eu/article/the-study-eu-country-ranking-2018-for-international-students

[2] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[3] "The Portuguese grading system". Study in Europe. https://www.studyineurope.eu/study-in-portugal/grades