UCI Student Performance Data Set Classification Problem

João Calhau

January 22, 2018

Abstract

Portugal's population as always been in the end of the list of Europe's educational level due to it's high student failure rates and lack of success in it's core classes of Mathematics and Portuguese. Education is key when talking about achieving success and achieving long-term economic progress. This paper, aims to extract high-level knowledge from raw, real-world data, that was collected by using school reports and questionnaires, through Data Mining techniques. [5]
The tool this paper is going to focus on is Weka, and in particular the classifiers that are going to be used are among, Random Forest, Decision Trees, Neural Networks and Support Vector Machines (although we're also gonna test some of the more basic classifiers like One Rule and Naive Bayes).

Contents

1 Introduction

The chosen data set was the one referring to a study about Student Performance in Portuguese schools [5]. Although Portugal has been evolving in terms of education it's still in the end of the list of Europe's educational level due to it's high student failure rates and lack of success in it's core classes of Mathematics and Portuguese. For example, in 2015 Portugal had a total of 49,829 total graduates, while Germany had a total of 340,052 (Eurostat 2015), this shows a difference in over 600%.

This is a matter of utmost importance given the fact that the most failed classes are the core classes of Portuguese and Mathematics which are the classes that provide the most fundamental knowledge for the remainder of school subjects. What we are trying to accomplish here is find a correlation between all the attributes in the data set that could negatively influence our academic path.

2 Data

In this data we're going to analyze real-world data from two Portuguese secondary schools. This data comes from two different sources, mark reports and questionnaires [5] The data was divided into two data sets, a data set containing answers belonging to Portuguese students and a data set containing answers that belong to Mathematics students.

Both the data sets are composed by 33 attributes, which are the following:

Table 1: Data set attributes

Attribute	Description	Domain
sex	Student's sex	Binary: Male or Female
age	Student's age	Numeric: From 15 to 22
school	Student's school	Binary: Gabriel Pereira or Mousinho da Silveira
address	Student's home adress	Binary: Urban or Rural
Pstatus	Parent's cohabitation status	Binary: Living together or Apart
Medu	Mother's education	Numeric: From 0 to 4
Mjob	Mother's job	Nominal
Fedu	Father's education	Numeric: From 0 to 4
Fjob	Father's Job	Nominal
guardian	Student's Guardian	Nominal: Mother, Father or Other
famsize	Family Size	Binary: ≤ 3 or > 3
famrel	Quality of family relationships	Numeric: From 1 to 5
reason	Reason to choose this school	Nominal
traveltime	Home to school travel time	Numeric: From 1 to 4
studytime	Weekly study time	Numeric: From 1 to 4
failures	Number of past class failures	Numeric: From 1 to 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	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 to 5
goout	Going out with friends	Numeric: From 1 to 5
Walc	Weekend alcohol consumption	Numeric: From 1 to 5
Dalc	Workday alcohol consumption	Numeric: From 1 to 5
health	Current health status	Numeric: From 1 to 5
absences	Number of school absences	Numeric: From 0 to 93
G1	First period grade	Numeric: From 0 to 20
G2	Second period grade	Numeric: From 0 to 20
G3	Final grade	Numeric: From 0 to 20

3 Algorithms/Proposed Solutions

The majority of the classification techniques used in this work were Decision Tree based (REPTree, J48 and RandomForest) the other one is a Bayesian Belief Network type of classification. There were more algorithms applied to these data sets, but their results were inconclusive.

3.1 Decision Trees

3.1.1 REPTree

This classifier is a fast decision tree learner, it builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with Back Fitting). It Only sorts values for numeric attributes once and missing values are dealt with by splitting the corresponding instances into pieces. [4]

For the Portuguese data set, when applied the REPTree classifier, through the leaves of the pruned tree became apparent that, for the grade to be over 9, the attribute weekend alcohol consumption should be very low (in the data that corresponds to a 1), for the grade to be over 12, the attribute going out with friends should be between 1 and 3 and, for the grade to be over 15, the attribute free time after school should be bellow 3.
For the Mathematics data set, however, what became apparent was that, to have a grade over 9, the only attribute that matters is the number of absences to be lower than 7.

3.1.2 J48

This classifier generates a pruned (or unpruned) C4.5 decision tree. C4.5 is an algorithm used to generate decision trees that are to be used for classification. [2]

Through the appliance of this classifier to the data sets, what we learned was that the number of correctly classified instances was much lower, but the tree build time was much lower, about 2 seconds lower. But through the leaves of the pruned tree, we couldn't come to any conclusion.

3.1.3 RandomForest

This classification algorithm is an ensemble learning method for classification, regression and other tasks, that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classifications or mean prediction of the individual trees. [3]

When applied to the data sets what we learned was that, although the correctly classified instances was higher, the time it took to generate all the unpruned trees was about 10 seconds longer, as because this algorithm generates so many trees during the training set, it becomes too difficult to extract any worthwhile data.

3.2 Bayesian Belief Network

This classifier uses various search algorithms and quality measures. The base classifier provides data structures (network structure, conditional probability distributions, etc) and facilities common to Bayes Network learning algorithms like K2 and B. [1]

Through the application of this classifier we learned that the final grade of the class is directly correlated with every other attribute, because the final grade node as at least one connection to all the other attributes. We learned this, but at the same time, the quantity of correctly classified instances went down when compared to the three classifiers above.

4 Conclusions

From all this data we can conclude that the biggest risk factors associated with Student Performance in Portuguese classes are Weekend Alcohol Consumption, which should be kept the lowest possible or none at all, if we want to go out with friends we should do it, but not too much, we should keep it at a medium, we should have free time, but not too much, and in the case of Mathematics classes, the overall attendance should be kept at least 86 classes attended, if not more. Regarding this last factor, I think it is due to all the themes being connected, and if we skip one, it is going to leave us seriously behind, while in Portuguese that is not so much the case.

References

[1] Weka bayesnet classifier. http://weka.sourceforge.net/doc.stable/weka/classifiers/bayes/BayesNet.html. Accessed: 2018-01-22.

[2] Weka j48 classifier. http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/J48.html. Accessed: 2018-01-22.

[3] Weka randomforest classifier. http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/RandomForest.html. Accessed: 2018-01-22.

[4] Weka reptree classifier. http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/REPTree.html. Accessed: 2018-01-22.

[5] Paulo Cortez and Alice Silva. Using Data Mining To Predict Secondary School Student Performance. *5th Annual Future Business Technology Conference*, 2003(2000):pp. 5–12, 2008.