USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

Paulo Cortez and Alice Silva
Dep. Information Systems/Algoritmi R&D Centre
University of Minho
4800-058 Guimarães, PORTUGAL
Email: pcortez@dsi.uminho.pt, alicegsilva@gmail.com

KEYWORDS
Business Intelligence in Education, Classification and Regression, Decision Trees, Random Forest

ABSTRACT
Although the educational level of the Portuguese population has improved in the last decades, the statistics keep Portugal at Europe's tail end due to its high student failure rates. In particular, lack of success in the core classes of Mathematics and the Portuguese language is extremely serious. On the other hand, the fields of Business Intelligence (BI)/Data Mining (DM), which aim at extracting high-level knowledge from raw data, offer interesting automated tools that can aid the education domain. The present work intends to approach student achievement in secondary education using BI/DM techniques. Recent real-world data (e.g. student grades, demographic, social and school related features) was collected by using school reports and questionnaires. The two core classes (i.e. Mathematics and Portuguese) were modeled under binary/five-level classification and regression tasks. Also, four DM models (i.e. Decision Trees, Random Forest, Neural Networks and Support Vector Machines) and three input selections (e.g. with and without previous grades) were tested. The results show that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption). As a direct outcome of this research, more efficient student prediction tools can be be developed, improving the quality of education and enhancing school resource management.

INTRODUCTION

Education is a key factor for achieving a long-term economic progress. During the last decades, the Portuguese educational level has improved. However, the statistics keep the Portugal at Europe's tail end due to its high student failure and dropping out rates. For example, in 2006 the early school leaving rate in Portugal was 40% for 18 to 24 year olds, while the European Union average value was just 15% (Eurostat 2007). In particular, failure in the core classes of Mathematics and Portuguese (the native language) is extremely serious, since they provide fundamental knowledge for the success in the remaining school subjects (e.g. physics or history).

On the other hand, the interest in Business Intelligence (BI)/Data Mining (DM) (Turban et al. 2007), arose due to the advances of Information Technology, leading to an exponential growth of business and organizational databases. All this data holds valuable information, such as trends and patterns, which can be used to improve decision making and optimize success. Yet, human experts are limited and may overlook important details. Hence, the alternative is to use automated tools to analyze the raw data and extract interesting high-level information for the decision-maker.

The education arena offers a fertile ground for BI applications, since there are multiple sources of data (e.g. traditional databases, online web pages) and diverse interest groups (e.g. students, teachers, administrators or alumni) (Ma et al. 2000). For instance, there are several interesting questions for this domain that could be answered using BI/DM techniques (Luan 2002, Minaei-Bidgoli et al. 2003): Who are the students taking most credit hours? Who is likely to return for more classes? What type of courses can be offered to attract more students? What are the main reasons for student transfers? Is is possible to predict student performance? What are the factors that affect student achievement? This paper will focus in the last two questions. Modeling student performance is an important tool for both educators and students, since it can help a better understanding of this phenomenon and ultimately improve it. For instance, school professionals could perform corrective measures for weak students (e.g. remedial classes).

In effect, several studies have addressed similar topics. Ma et al. (2000) applied a DM approach based in Association Rules in order to select weak tertiary school students of Singapore for remedial classes. The input variables included demographic attributes (e.g. sex, region) and school performance over the past years and the proposed solution outperformed the traditional allocation procedure. In 2003 (Minaei-Bidgoli et al. 2003),

online student grades from the Michigan State University were modeled using three classification approaches (i.e. binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). The database included 227 samples with online features (e.g. number of corrected answers or tries for homework) and the best results were obtained by a classifier ensemble (e.g. Decision Tree and Neural Network) with accuracy rates of 94% (binary), 72% (3-classes) and 62% (9-classes). Kotsiantis et al. (2004) applied several DM algorithms to predict the performance of computer science students from an university distance learning program. For each student, several demographic (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment) were used as inputs of a binary pass/fail classifier. The best solution was obtained by a Naive Bayes method with an accuracy of 74%. Also, it was found that past school grades have a much higher impact than demographic variables. More recently, Pardos et al. (2006) collected data from an online tutoring system regarding USA 8th grade Math tests. The authors adopted a regression approach, where the aim was to predict the math test score based on individual skills. The authors used Bayesian Networks and the best result was an predictive error of 15%.

In this work, we will analyze recent real-world data from two Portuguese secondary schools. Two different sources were used: mark reports and questionnaires. Since the former contained scarce information (i.e. only the grades and number of absences were available), it was complemented with the latter, which allowed the collection of several demographic, social and school related attributes (e.g. student's age, alcohol consumption, mother's education). The aim is to predict student achievement and if possible to identify the key variables that affect educational success/failure. The two core classes (i.e. Mathematics and Portuguese) will be modeled under three DM goals:

i) binary classification (pass/fail);
ii) classification with five levels (from I very good or excellent to V - insufficient); and
iii) regression, with a numeric output that ranges between zero (0%) and twenty (100%).

For each of these approaches, three input setups (e.g. with and without the school period grades) and four DM algorithms (e.g. Decision Trees, Random Forest) will be tested. Moreover, an explanatory analysis will be performed over the best models, in order to identify the most relevant features.

MATERIALS AND METHODS

Student Data

In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. There are several courses (e.g. Sciences and Technologies, Visual Arts) that share core subjects such as the Portuguese Language and Mathematics. Like several other countries (e.g. France or Venezuela), a 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, students are evaluated in three periods and the last evaluation (G3 of Table 1) corresponds to the final grade.

This study will consider data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. Although there has been a trend for an increase of Information Technology investment from the Government, the majority of the Portuguese public school information systems are very poor, relying mostly on paper sheets (which was the current case). Hence, the database was built from two sources: school reports, based on paper sheets and including few attributes (i.e. the three period grades and number of school absences); and questionnaires, used to complement the previous information. We designed the latter with closed questions (i.e. with predefined options) related to several demographic (e.g. mother's education, family income), social/emotional (e.g. alcohol consumption) (Pritchard and Wilson 2003) and school related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get a feedback. The final version contained 37 questions in a single A4 sheet and it was answered in class by 788 students. Latter, 111 answers were discarded due to lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes.

During the preprocessing stage, some features were discarded due to the lack of discriminative value. For instance, few respondents answered about their family income (probably due to privacy issues), while almost 100% of the students live with their parents and have a personal computer at home. The remaining attributes are shown in Table 1, where the last four rows denote the variables taken from the school reports.

Data Mining Models

Classification and regression are two important DM goals. Both require a supervised learning, where a model is adjusted to a dataset made up of $k \in \{1, ..., N\}$ examples, each mapping an input vector $(x_1^k, \ldots, x_I^k)$ to a given target y_k. The main difference is set in terms of the output representation, (i.e. discrete for classification and continuous for regression). In classification, models are often evaluated using the Percentage of Correct Classifications (PCC), while in regression the Root

Table 1: The preprocessed student related variables

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
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 $1 \leq n < 3$, else 4)
schoolsup	extra educational school 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 (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)
goout	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)
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)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Mean Squared (RMSE) is a popular metric (Witten and Frank 2005). A high PCC (i.e. near 100%) suggests a good classifier, while a regressor should present a low global error (i.e. RMSE close to zero). These metrics can be computed using the equations:

$$\begin{aligned} \Phi(i) &= \begin{cases} 1 & , \text{ if } y_i = \widehat{y}_i \\ 0 & , \text{ else} \end{cases} \\ PCC &= \textstyle\sum_{i=1}^{N} \Phi(i)/N \times 100\,(\%) \\ RMSE &= \sqrt{\textstyle\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2/N} \end{aligned} \tag{1}$$

where $\widehat{y}_i$ denotes the predicted value for the i-th example.
In this work, the Mathematics and Portuguese grades (i.e. G3 of Table 1) will be modeled using three supervised approaches:

1. **Bin**ary classification – *pass* if G3$\geq$10, else *fail*;
2. **5-L**evel classification – based on the Erasmus[1] grade conversion system (Table 2);
3. **Reg**ression – the G3 value (numeric output between 0 and 20).

Figure 1 plots the respective histograms.
Several DM algorithms, each one with its own purposes and capabilities, have been proposed for classification and regression tasks. The Decision Tree (DT) is a

[1] European exchange programme that enables student exchange in 31 countries.

Table 2: The five-level classification system

Country	**I** (excellent/very good)	**II** (good)	**III** (satisfactory)	**IV** (sufficient)	**V** (fail)
Portugal/France	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

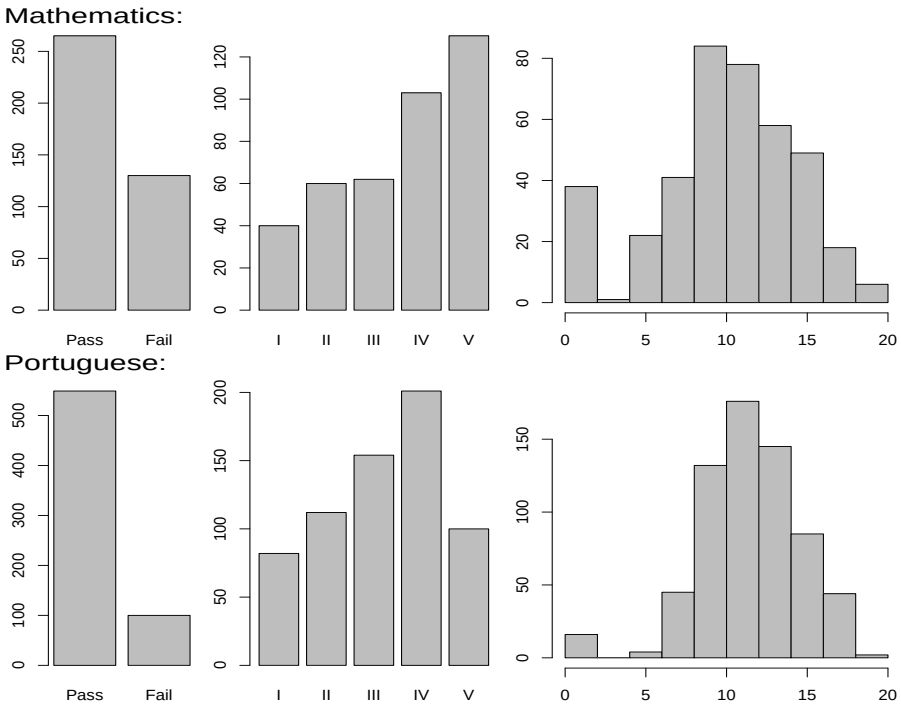

Figure 1: Histograms for the output variables (binary/5-level classification and regression)

branching structure that represents a set of rules, distinguishing values in a hierarchical form (Breiman et al. 1984). This representation can translated into a set of IF-THEN rules, which are easy to understand by humans. The Random Forest (RF) (Breiman 2001) is an ensemble of T unpruned DT. Each tree is based in a random feature selection from bootstrap training samples and the RF predictions are built by averaging the outputs of the T trees. The RF is more difficult to interpret when compared with the single DT, although it is still possible to provide explanatory knowledge in terms of its input variable relevance. Nonlinear functions, such as Neural Networks (NN) and Support Vector Machines (SVM), have also been proposed for DM tasks (Hastie et al. 2001), obtaining better results when a high nonlinearity is present. In this work, the NN model is based in the popular multilayer perceptron, with one hidden layer with H hidden nodes, while the SVM will use a gaussian kernel with one hyperparameter (γ). It should be noted that NN and SVM use model representations that are difficult to understand by humans. Also, NN and SVM are more affected by irrelevant inputs than the DT/RF algorithms, since the latter explicitly perform an internal feature selection.

Computational Environment

All experiments reported in this study were conducted using the **RMiner** (Cortez In press), an open source library for the **R** environment that facilitates the use of DM techniques (Figure 2). **R** is a free and high-level matrix programming language with a powerful suite of tools for statistical and data analysis (R Development Core Team 2006). It can run in multiple platforms (e.g. *Windows*, *MacOS* or *Linux*) and new features can be added by creating packages.

The **RMiner** library[2] presents a set of coherent functions (e.g. `mining`, `saveMining`) for classification and regression tasks (Cortez In press). In particular, the library uses the **rpart** (DT), **randomForest** (RF), **nnet** (NN) and **kernlab** (SVM) packages. As an example, the following **RMiner/R** code was used during the DT predictive experiments of the portuguese binary classi-

[2]http://www.dsi.uminho.pt/~pcortez/R/rminer.zip

fication:

```
source("rminer.R") # read the RMiner library
# load the data:
data=read.table("a-por-bin.dat",sep=";",header=T)
K=c("kfold",10) # 10-fold cross-validation
# execute 10 runs of a DT classification:
DT=mining(y~.,data,model="dt",Runs=20,method=K)
# show mean classification error on test set:
print(mean(DT$error))
# save the results (predictions, ...):
saveMining(DT,"DT-results")
```

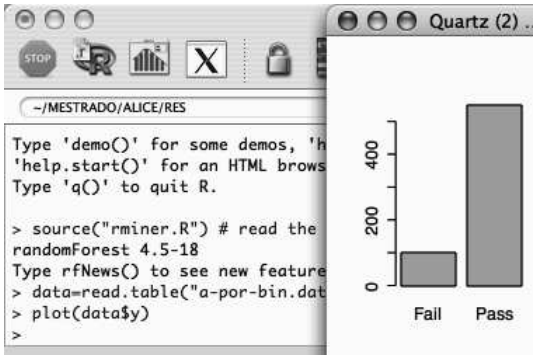

Figure 2: MacOS example of the R/RMiner tool.

RESULTS

Predictive Performance

Before fitting the models, some preprocessing was required by the NN and SVM models. The nominal[3] variables (e.g **Mjob**) were transformed into a *1-of-C* encoding and all attributes were standardized to a zero mean and one standard deviation (Hastie et al. 2001). Next, the DM models were fitted. The DT node split was adjusted for the reduction of the sum of squares. Regarding the remaining methods, the default parameters were adopted for the RF (e.g. $T = 500$), NN (e.g. $E = 100$ epochs of the BFGS algorithm) and SVM (e.g. Sequential Minimal Optimization algorithm). Also, the NN and SVM hyperparameters were optimized using an internal grid search (i.e. using only training data) where $H \in \{0, 2, 4, 6, 8\}$ and $\gamma \in \{2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$.
Since we suspected that the G1 and G2 grades would have a high impact, three input configurations were tested for each DM model:

- **A** - with all variables from Table 1 except G3 (the output);
- **B** - similar to **A** but without G2 (the second period grade); and
- **C** - similar to **B** but without G1 (the first period grade).

[3]Discrete with more than two non-ordered values.

To access the predictive performances, 20 runs of a 10-fold cross-validation (Hastie et al. 2001) (in a total of 200 simulations) were applied to each configuration. Under such scheme, for a given run the data is randomly divided in 10 subsets of equal size. Sequentially, one different subset is tested (with 10% of the data) and the remaining data used to fit the DM technique. At the end of this process, the evaluated test set contains the whole dataset, although 10 variations of the same DM model are used to create the predictions.
As a baseline comparison, a naive predictor (NV) will also be tested. For the **A** setup, this model is equal to the second period grade (G2 or its binary/5-level versions). When the second grade is not available (**B** setup), the first period grade is used (or its binary/5-level variants). If no evaluation is present (**C** setup) then the most common class (for classification tasks) or the average output value (regression) is returned.
The test set results are shown in Tables 3 to 5 in terms of the mean and respective t-student 95% confidence intervals (Flexer 1996). As expected, the **A** setup achieves the best results. The predictive performance decreases when the second period grade is not known (**B**) and the wost results are obtained when no student scores are used (**C**). Using only the last available evaluation (NV method for the first two input setups) is the best option for the Mathematics classification goals (both binary and 5-level) and Portuguese regression under the **A** input selection. This suggests that in these cases the non evaluation inputs are of no use. However, the scenario changes for the remaining experiments. The RF is the best choice in 8 cases, followed by the DT, which obtains 4 best results. In general, the nonlinear function methods (NN and SVM) are outperformed by the tree based ones. This behavior may be explained by a high number of irrelevant inputs, as shown in the next section.
As an example of the quality of the predictions, Figure 3 shows the confusion matrices for the DT algorithm, Portuguese class and **A** setup. In both binary and 5-level classification, the majority of the values are in/near the matrix diagonal, revealing a good fit. The figure also presents the RF scatter plot (predicted versus the observed values) for the Mathematics regression approach. The RF predictions are close to the diagonal line (which denotes the perfect forecast) within most of the output range (i.e. $[5, 20]$).

Descriptive Knowledge

Here the goal is not to infer about the predictive capabilities of each model, as measured in the previous section, but to give a simple description that summarizes the best DM models. Thus, the whole dataset will be used in the descriptive experiments. Furthermore, only the DT/RF algorithms will be considered, since it is more easy to extract knowledge from these models and they

Table 3: Binary classification results (PCC values, in %; underline – best model; **bold** – best within the input setup)

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**91.9**$^{\dagger}{}_{\pm 0.0}$	88.3$_{\pm 0.7}$	86.3$_{\pm 0.6}$	90.7$_{\pm 0.3}$	91.2$_{\pm 0.2}$	89.7$_{\pm 0.0}$	90.7$_{\pm 0.5}$	91.4$_{\pm 0.2}$	**93.0**$^{\dagger}{}_{\pm 0.3}$	92.6$_{\pm 0.1}$
B	**83.8**$^{\dagger}{}_{\pm 0.0}$	81.3$_{\pm 0.5}$	80.5$_{\pm 0.5}$	83.1$_{\pm 0.5}$	83.0$_{\pm 0.4}$	87.5$_{\pm 0.0}$	87.6$_{\pm 0.4}$	88.0$_{\pm 0.3}$	88.4$_{\pm 0.3}$	**90.1**$^{\dagger}{}_{\pm 0.2}$
C	67.1$_{\pm 0.0}$	66.3$_{\pm 1.0}$	**70.6**$^{*}{}_{\pm 0.4}$	65.3$_{\pm 0.8}$	70.5$_{\pm 0.5}$	84.6$_{\pm 0.0}$	83.4$_{\pm 0.5}$	84.8$_{\pm 0.3}$	84.4$_{\pm 0.4}$	**85.0**$^{*}{}_{\pm 0.2}$

† – statistical significance under pairwise comparisons with other methods.
∗ – statistical significance under a pairwise comparison with NV.

Table 4: Five-level classification results (PCC values, in %; underline – best model; **bold** – best within the input setup)

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**78.5**$^{\dagger}{}_{\pm 0.0}$	60.3$_{\pm 1.6}$	59.6$_{\pm 0.9}$	76.7$_{\pm 0.4}$	72.4$_{\pm 0.4}$	72.9$_{\pm 0.0}$	65.1$_{\pm 0.9}$	64.5$_{\pm 0.6}$	**76.1**$_{\pm 0.1}$	73.5$_{\pm 0.2}$
B	**60.5**$^{\dagger}{}_{\pm 0.0}$	49.8$_{\pm 1.2}$	47.9$_{\pm 0.7}$	57.5$_{\pm 0.8}$	52.7$_{\pm 0.6}$	58.7$_{\pm 0.0}$	52.0$_{\pm 0.6}$	51.7$_{\pm 0.6}$	**62.9**$_{\pm 0.2}$	55.3$_{\pm 0.4}$
C	32.9$_{\pm 0.0}$	30.4$_{\pm 1.0}$	31.0$_{\pm 0.7}$	31.5$_{\pm 0.6}$	**33.5**$_{\pm 0.6}$	31.0$_{\pm 0.0}$	33.7$_{\pm 0.6}$	34.9$_{\pm 0.5}$	32.8$_{\pm 0.6}$	**36.7**$^{\dagger}{}_{\pm 0.6}$

† – statistical significance under pairwise comparisons with other methods.

Table 5: Regression results (RMSE values; underline – best model; **bold** – best within the input setup)

I.	Mathematics					Portuguese				
S.	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	2.01$_{\pm 0.00}$	2.05$_{\pm 0.02}$	2.09$_{\pm 0.02}$	1.94$_{\pm 0.04}$	**1.75**$^{\dagger}{}_{\pm 0.01}$	**1.32**$_{\pm 0.00}$	1.36$_{\pm 0.04}$	1.35$_{\pm 0.01}$	1.46$_{\pm 0.03}$	**1.32**$_{\pm 0.00}$
B	2.80$_{\pm 0.00}$	2.82$_{\pm 0.02}$	2.90$_{\pm 0.02}$	2.67$_{\pm 0.04}$	**2.46**$^{\dagger}{}_{\pm 0.01}$	1.89$_{\pm 0.00}$	1.88$_{\pm 0.02}$	1.87$_{\pm 0.01}$	**1.78**$^{*}{}_{\pm 0.03}$	1.79$_{\pm 0.01}$
C	4.59$_{\pm 0.00}$	4.41$_{\pm 0.03}$	4.37$_{\pm 0.03}$	4.46$_{\pm 0.04}$	**3.90**$^{\dagger}{}_{\pm 0.01}$	3.23$_{\pm 0.00}$	2.79$_{\pm 0.02}$	2.76$_{\pm 0.02}$	2.93$_{\pm 0.02}$	**2.67**$^{\dagger}{}_{\pm 0.01}$

† – statistical significance under pairwise comparisons with other methods.
∗ – statistical significance under a pairwise comparison with NV.

achieved in general the best predictive performances.

Table 6 presents the relative importance (in percentage) of each input variable as measured by the RF algorithm (Breiman 2001). To clarify the analysis, only the five most relevant results are shown in the table. These five variables present an overall impact that ranges from 43% to 77%, which indicates that there is a high number of irrelevant inputs. As expected, the student evaluations have a high impact (from 23% to 46%) in the models. For instance, G2 is the most important feature for the **A** input selection example, while G1 is highly relevant for the **B** setup. Also, the number of past failures, which is related with previous student performance, is the most important factor when no student scores are available. Nevertheless, there are other relevant factors, such as school related (e.g. number of absences, extra school support or travel time), demographic (e.g. the mother's job) and social (e.g. going out with friends, alcohol consumption) variables.

Figure 4 plots the best four DT. Again, the student grades are the most relevant features, appearing at the root of the trees, and only a small number (2 to 5) of the inputs considered are used. There are some interesting rules that can be extracted from these trees, for instance:

1. if G1$<$ 10 $\wedge$ G2=9 $\wedge$ Mjob $\in$ {teacher, other} then *pass*;
2. if G1$<$ 10 $\wedge$ G2=9 $\wedge$ Mjob $\in$ {home, health, civil services} then *fail*;
3. if 12 $\leq$G2$<$ 14 $\wedge$ goout$>$ 1 then III;
4. if 12 $\leq$G2$<$ 14 $\wedge$ goout= 1 then II;
5. if 11 $<$G1$\leq$ 13 $\wedge$ absences$<$ 7 then 13;
6. if 11 $<$G1$\leq$ 13 $\wedge$ absences$\geq$ 7 then 11;

These rules show the influence of the mother's job (rules 1–2), going out with friends (rules 3–4) and the number of absences (rules 5–6).

CONCLUSIONS

Education is a crucial element in our society. Business Intelligence (BI)/Data Mining (DM) techniques, which

Table 6: Relative importance of the input variables for the best RF models

Setup	Relative Importance
C-Mat-Bin	failures: 21.8%, absenses: 9.4%, schoolsup: 7.0%, goout: 6.5%, higher: 6.4%
B-Por-Bin	G1: 22.8%, failures: 14.4%, higher: 11.9%, school: 8.1%, Mjob: 4.1%
C-Por-Bin	failures: 16.8%, school: 13.2%, higher: 13.1%, traveltime: 5.9, famrel: 5.7%
C-Mat-5L	failures: 18.3%, schoolsup: 9.5%, sex: 5.7%, absences: 5.6%, Medu: 4.5%
C-Por-5L	failures: 16.8%, higher: 9.9%, school: 9.3%, schoolsup: 6.9%, Walc: 6.6%
A-Mat-Reg	G2: 30.5%, absenses: 20.6%, G1: 15.4%, failures: 6.7%, age: 4.2%
B-Mat-Reg	G1: 42.2%, absenses: 18.6%, failures: 8.9%, age: 3.3%, schoolsup: 3.2%
C-Mat-Reg	failures: 19.7%, absenses: 18.9%, schoolsup: 8.3%, higher: 5.4%, Mjob:4.2%
C-Por-Reg	failures: 20.7%, higher: 11.4%, schoolsup: 6.9%, school: 6.7%, Medu:5.6%

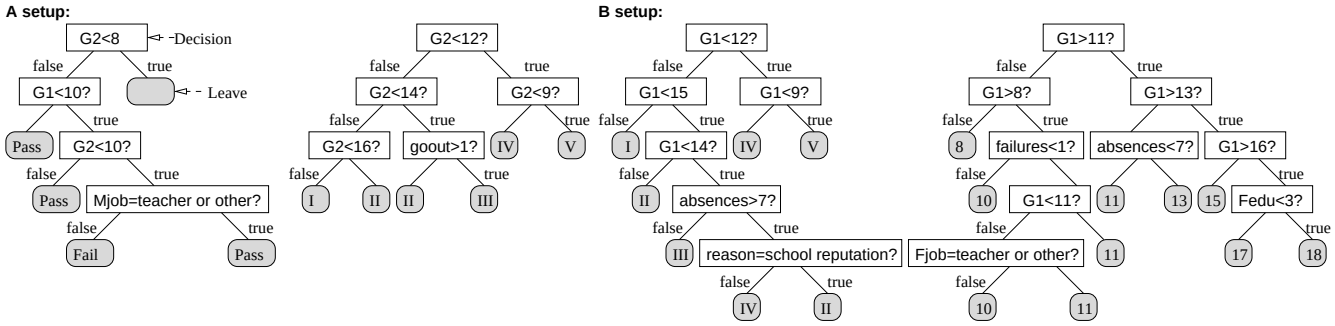

Figure 4: Examples of Decision Trees (A-Por-Bin, A-Por-5L, B-Por-5L and B-Por-Reg)

allow a high level extraction of knowledge from raw data, offer interesting possibilities for the education domain. In particular, several studies have used BI/DM methods to improve the quality of education and enhance school resource management.

In this paper, we have addressed the prediction of secondary student grades of two core classes (Mathematics and Portuguese) by using past school grades (first and second periods), demographic, social and other school related data. Three different DM goals (i.e. binary/5-level classification and regression) and four DM methods, i.e. Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM), were tested. Also, distinct input selections (e.g. with or without past grades) were explored. The obtained results reveal that it is possible to achieve a high predictive accuracy, provided that the first and/or second school period grades are known. This confirms the conclusion found in (Kotsiantis et al. 2004): student achievement is highly affected by previous performances. Nevertheless, an analysis to knowledge provided by the best predictive models has shown that, in some cases, there are other relevant features, such as: school related (e.g. number of absences, reason to choose school, extra educational school support), demographic (e.g. student's age, parent's job and education) and social (e.g. going out with friends, alcohol consumption) variables.

This study was based on an off-line learning, since the DM techniques were applied after the data was collected. However, there is a potential for an automatic on-line learning environment, by using a student prediction engine as part of a school management support system. This will allow the collection of additional features (e.g. grades from previous school years) and also to obtain a valuable feedback from the school professionals. Furthermore, we intent to enlarge the experiments to more schools and school years, in order to enrich the student databases. Automatic feature selection methods (e.g. filtering or wrapper) (Witten and Frank 2005) will also be explored, since only a small portion of the input variables considered seem to be relevant. In particular, this is expected to benefit the nonlinear function methods (e.g. NN and SVM), which are more sensitive to irrelevant inputs. More research is also needed (e.g. sociological studies) in order to understand why and how some variables (e.g. reason to choose school, parent's job or alcohol consumption) affect student performance.

ACKNOWLEDGMENTS

We wish to thank the students and school professionals who voluntarily participated in this study.

	Pass	Fail
Pass	**480**	69
Fail	12	**88**

	I	II	III	IV	V
I	**44**	33	4	1	0
II	2	**54**	42	14	0
III	0	19	**84**	48	3
IV	0	0	24	**111**	66
V	0	0	0	1	**88**

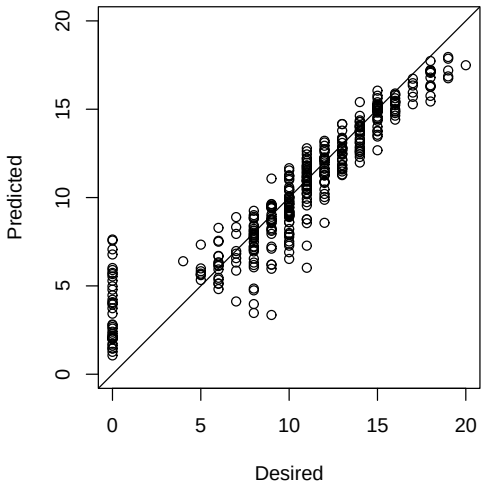

Figure 3: Example of the predictions on the test set: DT confusion matrices for A-Por-Bin and A-Por-5L setups (top); and RF scatter plot for A-Mat-Reg (bottom)

REFERENCES

Breiman L., 2001. *Random Forests. Machine Learning*, 45, no. 1, 5–32.

Breiman L.; Friedman J.; Ohlsen R.; and Stone C., 1984. *Classification and Regression Trees.* Wadsworth, Monterey, CA.

Cortez P., In press. *RMiner: Data Mining with Neural Networks and Support Vector Machines using R.* In R. Rajesh (Ed.), *Introduction to Advanced Scientific Softwares and Toolboxes.*

Eurostat, 2007. *Early school-leavers.* http://epp.eurostat.ec.europa.eu/.

Flexer A., 1996. *Statistical Evaluation of Neural Networks Experiments: Minimum Requirements and Current Practice.* In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research.* Vienna, Austria, vol. 2, 1005–1008.

Hastie T.; Tibshirani R.; and Friedman J., 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, NY, USA.

Kotsiantis S.; Pierrakeas C.; and Pintelas P., 2004. *Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence (AAI)*, 18, no. 5, 411–426.

Luan J., 2002. *Data Mining and Its Applications in Higher Education. New Directions for Institutional Research*, 113, 17–36.

Ma Y.; Liu B.; Wong C.; Yu P.; and Lee S., 2000. *Targeting the right students using data mining.* In *Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Boston, USA, 457–464.

Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., 2003. *Predicting student performance: an application of data mining methods with an educational web-based system.* In *Proc. of IEEE Frontiers in Education.* Colorado, USA, 13–18.

Pardos Z.; Heffernan N.; Anderson B.; and Heffernan C., 2006. *Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks.* In *Proc. of 8th Int. Conf. on Intelligent Tutoring Systems.* Taiwan.

Pritchard M. and Wilson S., 2003. *Using Emotional and Social Factors To Predict Student Success. Journal of College Student Development*, 44, no. 1, 18–28.

R Development Core Team, 2006. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, http://www.R-project.org.

Turban E.; Sharda R.; Aronson J.; and King D., 2007. *Business Intelligence, A Managerial Approach.* Prentice-Hall.

Witten I. and Frank E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, CA.

BIBLIOGRAPHIE

PAULO CORTEZ was born in Braga, Portugal and went to the University of Minho where he obtained an Eng. Degree (1995), a M.Sc. (1998) and a Ph.D (2002) in Computer Science. In 2001, he became Lecturer at the Dep. of Information Systems and Researcher at the Algoritmi R&D Centre. He participated in 7 R&D projects (principal investigator in 2) and he is co-author of more than fifty publications in int. conferences and journals. Research interests: Business Intelligence/Data Mining; Neural and Evolutionary Computation; Forecasting. Web-page: http://www.dsi.uminho.pt/˜pcortez

ALICE SILVA studied Informations Systems, obtaining a Degree (Bragança Polytechnic Inst., 2003) and a M.Sc. (U. Minho, 2007). Currently, she teaches Informatics at the secondary school of Arcozelo, Portugal.