Educational Data Mining: preliminary results at University of Porto

Pedro Strecht[1], João Mendes Moreira[2], Carlos Soares[3]

[1]Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, pstrecht@fe.up.pt
[2]LIAAD-INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, jmoreira@fe.up.pt
[3]INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, csoares@fe.up.pt

Keywords
Educational data mining, decision trees, prediction of approval.

1. ABSTRACT

Predicting the success or failure of a student in a course or a program is a problem that has recently been addressed using data mining techniques. However, the literature shows that there is still no consensus on what is the best set of variables that may lead to accurate models. Moreover, the problem is quite complex and appears to be very dependent on the data set used (Kabakchieva, 2013). It is related to student attrition, a research area of Educational Data Mining. This paper presents the results of preliminary experiments in this research area at the University of Porto. The experiments were carried out using automated approaches on administrative data. Although the number of courses is small, initial conclusions were drawn.

2. INTRODUCTION

The University of Porto,[1] formally founded in 1911, is currently the largest education and research institution in Portugal. It consists of 14 faculties, 1 business school, 30 000 students, 2 000 teachers and researchers along with 1 700 administrative staff. Since the early nineties, the University manages its academic data using University Information Systems (UIS). These are designed to store data and support the activity of the institution (not only the pedagogical process) and present this information in various ways.

One of the most serious problems higher education institutions have to deal with is related to students failing to reach the goals of educational process activities which represent a real threat to both the institution and the students themselves. The growing adoption of UIS has allowed research to move towards approaches based on automatic knowledge discovery from databases with educational data. Consequently, over the past ten years there has been an increase on research using data mining techniques in the discovery of the causes of phenomena such as student attrition.

This paper, motivated by this problem, presents an approach to predict the success or failure of a student enrolled in a course, a typical Educational Data Mining (EDM) task. Besides predicting the outcome, we also aim to understand the reasons behind it, which prompts us to use the decision tree algorithm. A set of candidate predictor variables were selected from socio-demographic and academic data from a small set of courses of the University in order to draw some initial conclusions that point out research questions and show directions for future research. These preliminary results enabled us to identify the variables with more discriminative power when the goal is to predict the success of a student in passing a specific course.

The remainder of this paper is structured as follows. Section 3 presents the research area of Educational Data Mining. Section 4 presents an illustrative example of an EDM task describing a

[1] www.up.pt

methodology, analysis of results and discussion. Section 5 presents the main conclusions and Section 6 some possible paths for future work.

3. EDUCATIONAL DATA MINING

University Information Systems have been used worldwide for about thirty years in higher education institutions. These systems have occupied a central role as key resource in administration settings covering the management of their students as well as supporting a variety of processes. As they reached maturity, institutions increasingly became dependent on their use on a daily basis, and managers realized that a great wealth of information was being accumulated over the years. Like in other industry areas, decision support systems for education started to explore this data, with proper presentation tools.

The growth of this kind of data gave rise to the term “educational big data”, which encompasses pedagogical data, referring to data collected from teaching and learning environments and administrative data which is related to academic information (enrollments, grades). Other types of data can also be considered relevant, as is the case of social network interaction. In Bienkowski’s perspective of educational big data (Bienkowski, Feng, & Means, 2012), presented in Figure 1, the field has two subareas: Learning Analytics (LA) and Educational Data Mining (EDM). Although related, both have had different origins and settled as distinct research areas.

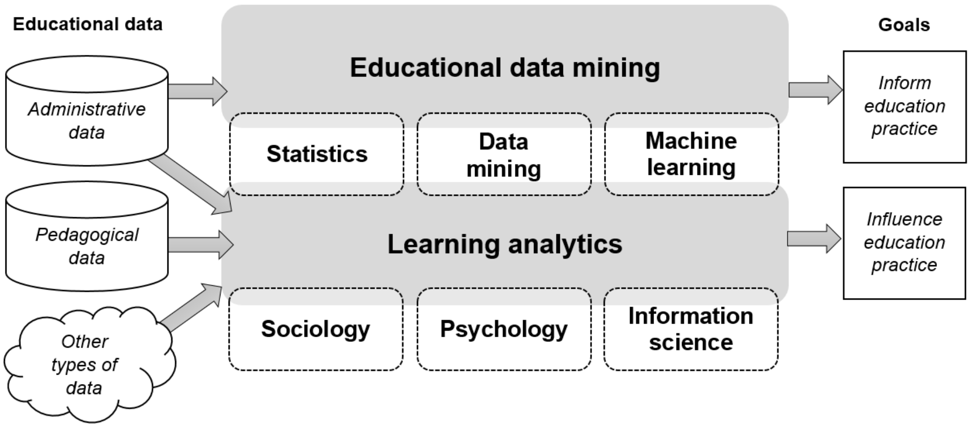

Figure 1 – Educational big data

Learning Analytics is an emerging field in which sophisticated analytic tools are used to improve learning and education (Elias, 2011). It is related to the interpretation of a large pool of academic data produced in the context of teaching activity in order to assess academic progress, predict future performance, and identify potential problems concerning the students. The data is obtained from actions taken by students, such as completing assignments and taking exams, and inferred actions, including online social interactions, extracurricular activities, posts on discussion boards and others that are not directly associated to the educational progress of a student. From these data, descriptive models can be created to assist faculty members and school staff to interpret the behavior of students. The learning objective of the analysis is to allow teachers and schools to adapt educational opportunities to the level of need and ability of each student (Johnson, Smith, Willis, Levine, & Haywood, 2011).

Educational Data Mining is also an emerging field concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn (Baker & Yacef, 2009). Currently the major research areas and trends are: student retention and attrition, improvement of student models, personal learning environments and recommender systems, course management systems improvement, studying pedagogical support, educational theories, easier use of knowledge extraction tools, systematization of both methods and data and finally integration with the e-learning system (Romero & Ventura, 2007).

Although LA and EDM share many characteristics and have some similar goals and interests, there are also key differences that can distinguish the two communities. These are presented in Table 1, in which 1-4 are described by Romero and Ventura (Romero & Ventura, 2013) and 5-6 are described by Bienkowski et al (Bienkowski et al., 2012).

Table 1 – Key differences between LA and EDM

Differences	Learning Analytics	Educational Data Mining
1. Techniques	Statistics, visualization, social network analysis, sentiment analysis, influence analytics, discourse analysis, concept analysis, and sense-making models	Classification, clustering, Bayesian modeling, relationship mining and discovery with models
2. Origins	Semantic Web, intelligent curriculum, and systemic interventions	Educational software, student modeling, and predicting course outcomes
3. Emphasis	Description of data and results	Description and comparison of the data mining techniques used
4. Type of discovery	Leveraging human judgment	Automated discovery
5. Data used	Pedagogical, administrative and other types of data	Mostly administrative data
6. Goals	Influence education practice	Inform education practice

Bienkowski et al. (Bienkowski et al., 2012) presented an exhaustive report on how EDM and LA can enhance teaching and learning, namely:

- a focus on usability and impact of presenting learning data to instructors;
- development of decision support systems and recommendation systems that minimize instructor intervention;
- development of tools for protecting individual privacy while still advancing educational data mining;
- development of models that can be used in multiple contexts.

4. AN ILLUSTRATIVE EXAMPLE OF AN EDM TASK

To illustrate how data mining can be used in educational contexts we describe a preliminary study on the problem of predicting the student success or failure in courses. In the next sections, we describe in detail the system architecture, experimental setups and the results.

4.1. System architecture

To carry out the experiments, a system was developed with the architecture presented in Figure 2. The main goal is to obtain an accurate model for the prediction of success or failure of a student in a course. Given the sensitive nature of the task, it is also important that the model is interpretable. Therefore, we decided to generate decision tree classifiers using the C5.0 algorithm by Ross Quinlan (Quinlan, 1998).

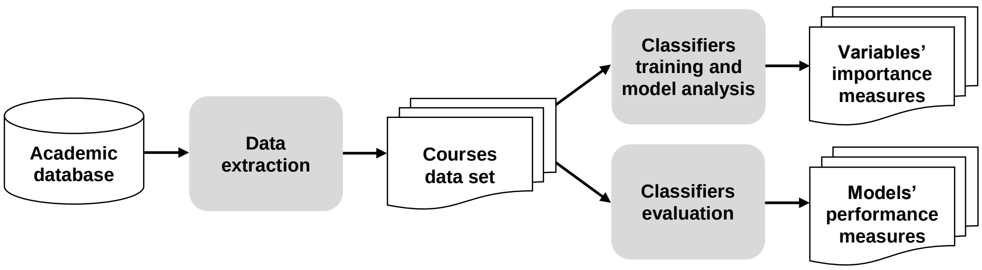

Figure 2 - System architecture

The system architecture includes three distinct processes. In the first process, variables are selected and extracted from academic data and gathered in a data set for each course in a format suitable for classifier training (Section 4.2). After this, a second process generates decision tree models for each course in the data set and then analyzes them in order to determine which are the most important variables (Section 4.3). In parallel, a third process evaluates each model and calculates predictive performance measures (Section 4.4).

4.2. Data extraction

The first process aims to extract a data set from the database maintained by the University Information System, developed in-house in 1992. The system has been evolving and currently also encompasses front-end web portals available in all 14 faculties. The academic database consists of an Oracle central repository and stores a large amount of data on students, program syllabuses, courses, academic acts and assorted data related to a variety of sub-processes of the pedagogical process. The database stores data from the early eighties. Although there have been changes to the schema, the data is essentially reliable. The extracted data set encompasses a selection of candidate variables presented in Table 2.

Table 2 – Predictor's variables

Variable	Remarks	Type
Age	age of student at the date of enrollment in course	Numerical
Sex	male | female	Nominal
Marital status	single | married | divorced | widower | unmarried partner	Nominal
Nationality	first nationality of student (country acronym)	Nominal
Displaced from place of residence	whether the student formerly lived outside the Porto district	Boolean
Scholarship	whether the student has a scholarship	Boolean
Special needs	whether the student has disabilities (visual, hearing or motor impairments)	Boolean
Type of admission	type of application contest: regular | transfer | over 23 | holder of degree | …	Nominal
Type of student	regular | mobility | extraordinary	Nominal
Status of student	ordinary | employed | athlete | …	Nominal
Years of enrollment	number of academic years the student has been enrolled in previously	Numerical
Delayed courses	number of courses the student should have completed	Numerical
Type of dedication	full-time | part-time	Nominal
Debt situation	whether there are fees due	Boolean

To use the best possible data in the first set of experiments, we limited the courses analyzed in the following way:

- for data quality reasons, only courses from the school year 2012/2013 with the largest number of students enrolled were considered;
- to account for the variability regarding scientific areas, 8 courses from different faculties were selected.

These criteria led to the set of courses presented in Table 3.

Table 3 – Selected courses

Course	Program	Faculty	Academic year	Semester	Number of students
Economic History	Economics	Economics	1st	2nd	656
Organic Chemistry II	Pharmaceutical Sciences	Pharmacy	1st	2nd	562
Neuroanatomy	Medicine	Medicine	2nd	1st	542
Marketing	Economics	Economics	2nd	1st	519
Anatomy I	Medicine	Medicine	1st	1st	518
Anatomy II	Medicine	Medicine	1st	2nd	477
Mathematics II	Economics	Economics	1st	2nd	476
Introduction to Linear Signals and Systems	Electrical and Computers Engineering	Engineering	2nd	1st	475

Table 4 presents a sample of data for the course Mathematics II.

Table 4 – Data set sample for the course Mathematics II

Age	Sex	Marital status	Nationality	Displaced	Scholarship	Special needs	Type of admission	Type of student	Status of student	Years of enrollment	Delayed courses	Type of dedication	Debt situation	Approval
18	m	s	pt	y	n	n	r	r	o	0	0	f	n	n
32	m	m	pt	n	n	n	tcs	r	o	8	12	p	n	n
18	f	s	pt	y	n	n	r	r	o	0	0	f	n	y
18	m	s	pt	n	n	n	r	r	o	0	0	f	n	y
22	m	s	br	n	n	n	to	r	o	1	0	f	n	y

4.3. Classifier training and model analysis

The goal of the second process is to generate (train) decision tree classifiers for each course and then analyze each model in order to find out which are the variables used for prediction and some measure of importance. This process is carried out for each course, according to the experimental setup presented in Figure 3.

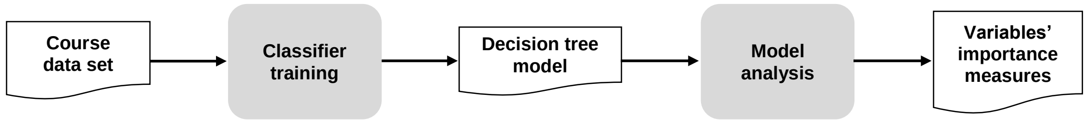

Figure 3 – Experimental setup for classifier training and model analysis (for each course)

4.3.1. Classifier training

Classification is a data analyzing task that extracts models discriminating observations belonging to different classes. Such models, called classifiers, predict categorical class labels. It is a two-step process consisting of a learning step, where the model is constructed, and a classification step, where it is actually used to predict an outcome for a given observation (Han, Kamber, & Pei, 2011). Classifiers obtained from decision tree algorithms have the characteristic of not requiring previous domain knowledge or heavy parameter tuning making them appropriate not only for prediction but also for exploratory data analysis. The tree-like representation of knowledge presents itself as intuitive, making models that are usually interpretable by humans. In this study, students are classified as either having passed (positive instances) or failed (negative instances) a course.

The algorithm is executed separately for each course data set. The output is a decision tree model, such as the one presented in Figure 4, for the course Mathematics II.

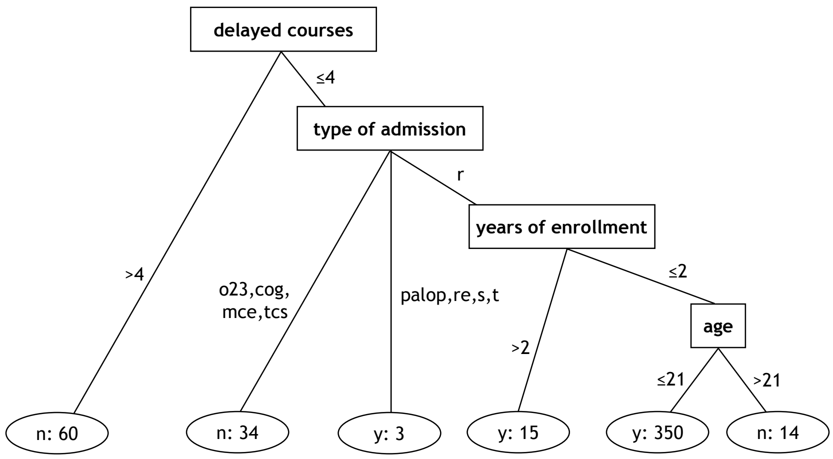

Figure 4 – Example of decision tree for the course Mathematics II

4.3.2. Model analysis

Decision trees can be interpreted by reading them down from the root to the leaf nodes. Each node in the tree specifies a test on one variable, and each branch descending from that node corresponds to one of the possible values for this variable.

A student is classified as having passed (*y* leaf node) or failed (*n* leaf node) at the course by starting at the root node of the tree, testing the variable in this node, then moving down the tree branch according to the outcome of the test. This process is then repeated for the sub tree rooted at the new node (Mitchell, 1997) and so on, until a leaf is reached. In the example presented in Figure 4, the number of students classified under each node is shown. This value is necessary to determine the importance of each variable in the model as discussed next.

The C5.0 algorithm for decision tree generation measures the importance of variable by determining the percentage of training set samples that fall into the leaf nodes. The variable used for the first split automatically has an importance measurement of 100 percent (Kuhn, Weston, Coulter, & Quinlan, 2013). Equation (1) shows that variable importance *p* is the proportion of examples under analysis in the node where *p* is tested concerning the total number of examples in the training data.

$$I_p = \frac{\#examples\ in\ node}{total\ examples} \qquad (1)$$

For example, in the course Mathematics II, the total number of students is 476, and from Figure 4, we can derive that:

$$I_{age} = \frac{\#examples_{>21} + \#examples_{\leq 21}}{total\ examples} = \frac{14+350}{476} = \frac{364}{476} = 76.4\%$$
$$I_{years\ of\ enrollment} = \frac{\#examples_{\leq 2} + \#examples_{>2}}{total\ examples} = \frac{364+15}{476} = \frac{379}{476} = 79.6\%$$
$$I_{type\ of\ admission} = \frac{\#examples_{=r} + \#examples_{=\{palop,re,s,t\}} + \#examples_{=\{o23,cog,mce,tcd\}}}{total\ examples} = \frac{379+3+34}{476} = \frac{416}{476} = 87.4\%$$
$$I_{delayed\ courses} = \frac{\#examples_{\leq 4} + \#examples_{>4}}{total\ examples} = \frac{416+60}{476} = \frac{476}{476} = 100{,}0\%$$

4.3.3. Preliminary results

Table 5 presents the results for each course: *#P* is the number of variables actually used in the model and then the values of the importance measure for each selected variable. Of all 13 variables, only 5 are used in the decision trees for the selected courses, namely, age, type of admission, status of the student, years of enrollment and delayed courses. It is also worthwhile noting that, although only 8 courses are being included in this study, the decision trees vary substantially from one to another. This hampers the ability to elect a definite set of variables. Additionally, in the 3 courses where only one variable is used this variable is never the same: type of admission for Economic History, years of enrollment for Marketing and delayed courses for Anatomy I.

Table 5 – Variables' importance measure for each course

Course	#P	I_p												
		Age	Sex	Marital status	Nationality	Displaced	Scholarship	Special needs	Type of admission	Type of student	Status of student	Years of enrollment	Delayed courses	Type of dedication	Debt situation
Economic History	1	-	-	-	-	-	-	-	100.0	-	-	-	-	-	-
Organic Chemistry II	3	-	-	-	-	-	-	-	11.5	-	-	72.1	100.0	-	-
Neuroanatomy	2	-	-	-	-	-	-	-	-	-	-	11.8	100.0	-	-
Marketing	1	-	-	-	-	-	-	-	-	-	-	100.0	-	-	-
Anatomy I	1	-	-	-	-	-	-	-	-	-	-	-	100.0	-	-
Anatomy II	4	-	-	-	-	-	-	-	36.7	-	20.9	18.4	100.0	-	-
Mathematics II	4	76.4	-	-	-	-	-	-	87.4	-	-	79.6	100.0	-	-
Introduction to Linear Signals and Systems	3	100.0	-	-	-	-	-	-	93.1	-	-	-	83.9	-	-

Table 6 presents a summary of variable importance by counting the number of courses in which each variable is used and separating them by importance groups.

Table 6 – Summary of variable importance

Variable	# of courses	Importance group						
		0-29%	30-49%	50-69%	70-79%	80-89%	90-99%	100%
Age	2	0	0	0	1	0	0	1
Type of admission	5	1	1	0	0	1	1	1
Status of student	1	1	0	0	0	0	0	0
Years of enrollment	5	2	0	0	2	0	0	1
Delayed courses	6	0	0	0	0	1	0	5

The importance group of 100% shows the number of courses in which the variable is the most important one. From Table 6 we conclude that, despite the variability of the decision trees, delayed courses is the most frequent important variable (5 out of 8 models), and simultaneously the most used variable overall, being present in 6 models. None of the other variables exhibit the same pattern. Age, type of admission and years of enrollment are the most important variables in only one model. In the analyzed courses, status of the student is hardly used, as it shows up in only one model in the lowest importance group. The type of admission and the years of enrollment are quite frequent appearing in 5 models, but not consistently, as they have different measures of importance.

4.4. Classifier evaluation

4.4.1 Evaluation measures

The goal of the third process is to assess the predictive performance of each classifier. For that purpose, it is necessary to determine performance measures. The predictions made by a classifier can be described in a confusion matrix (Han, Kamber, & Pei, 2011) such as the one shown in Table 7. It consists of four entries:

- True positives (*TP*) are the number of students correctly classified as having passed the course;
- False positives (*FP*) are the number of students incorrectly classified as having passed the course;
- True negatives (*TN*) are the number of students correctly classified as having failed the course;
- False negatives (*FN*) are the number of students incorrectly classified as having failed the course.

Table 7 – Confusion matrix

	Predicted positive	Predicted negative
Actual positive	*TP*	*FN*
Actual negative	*FP*	*TN*

From these entries, three evaluation measures can be defined:

- Precision (*P*) is the number of students correctly classified as having passed divided by the total number of students predicted as having passed (Eq. 2);
- Recall (*R*) is the number of students correctly classified as having passed divided by the total number of students that actually passed (Eq. 3);
- *F*1 combines both precision and recall with equal weights into a single measure (Eq. 4).

$$P = \frac{TP}{TP + FP} \quad (2) \qquad R = \frac{TP}{TP + FN} \quad (3) \qquad F1 = \frac{2PR}{P + R} \quad (4)$$

4.4.2 Estimation procedure

For evaluation of each decision tree classifier the *k*-fold cross-validation method was used (Stone, 1974) with stratified sampling (Kohavi, 1995). In *k*-fold cross-validation, students are divided and grouped into *k* mutually exclusive subsets D_1, D_2, ..., D_k, each of approximately equal size. As in every course, the distribution of positive and negative instances (the two categories of classification) is not balanced, it is necessary to ensure that the distribution of students in each fold respect these proportions. For this purpose, we used the stratified sampling technique to assign students to each fold. In general, stratified 10-fold cross validation is recommended for estimating accuracy due to its relatively low bias and variance (Han, Kamber, & Pei, 2011). The tasks of training and testing are therefore performed 10 times. At each iteration *i*, the D_i fold is used as the test set and the remaining data partitions are used for training. The model obtained is applied on the test data and *F*1 is used to assess its performance. For the overall estimate, the average of all 10 iterations of *F*1 is calculated. This process is carried out for each course, according to the experimental setup presented in Figure 5.

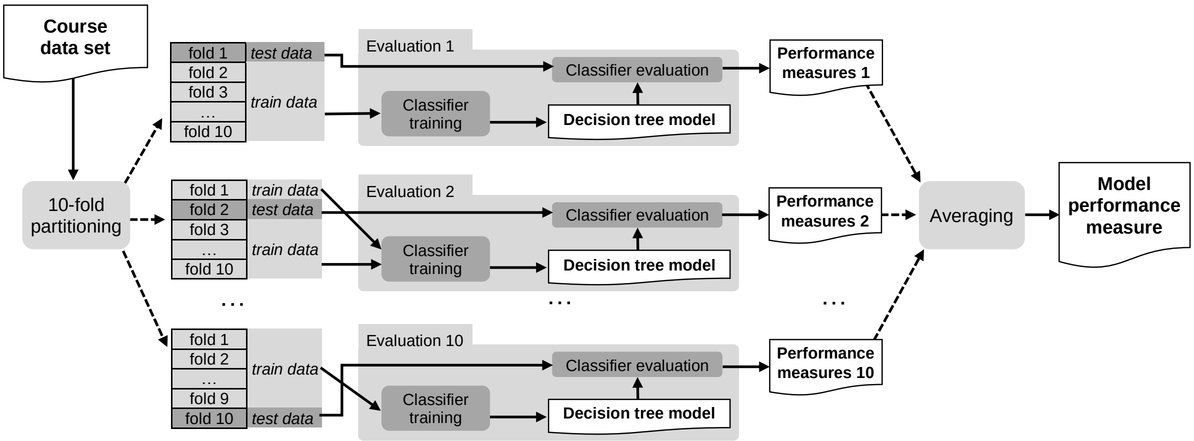

Figure 5 – Experimental setup for classifier evaluation (for each course)

4.4.3 Performance results

A total of 10 experiments were carried out for each course. Different results are obtained because the data is randomly shuffled before every 10-fold cross validation procedure. The results are shown in Table 8, with the number of examples (students), the category distribution of examples in each category and the average *F*1 with standard deviation for each course.

Table 8 – Model performance for each course

Course	Number of examples	Category distribution (%)		*F*1 (avg ± std.dev)
		y	*n*	
Economic History	656	72	28	0.8328 ± 0.0033
Organic Chemistry II	562	21	79	0.1039 ± 0.0306
Neuroanatomy	542	94	6	0.9664 ± 0.0014
Marketing	519	90	10	0.9559 ± 0.0016
Anatomy I	518	73	27	0.8557 ± 0.0028
Anatomy II	477	73	27	0.8420 ± 0.0044
Mathematics II	476	61	39	0.7863 ± 0.0053
Introduction to Linear Signals and Systems	475	55	45	0.7079 ± 0.0999

Overall, the results in terms of *F*1 in Table 8 show that the performance of the models is quite acceptable. The only exception is the course Organic Chemistry II with very low performance which, at this point, still needs to be investigated. However, it may be related to the fact that the number of positive examples (*y* category) is significantly lower than the negative ones (*n* category).

It is worthwhile noting that there is no apparent relationship between the number of variables used in the model (*#P* in Table 5) and its performance.

5. CONCLUSIONS

This paper describes a typical task of educational data mining that uses administrative data from the University of Porto. It presents an architecture for the prediction of success or failure of a student in a selection of 8 courses across different faculties. For each student enrolled in those courses, the value of a set of predictor variables was collected. In addition to the predictive ability it was also important to understand the causes leading to success or failure in each course. For this reason, we used the decision tree algorithm, which generates models that are typically interpretable by humans.

In classification tasks, performance measures, such as the *F*1 measure, allow the assessment of the quality of the predictions and the comparison between different algorithms. However, it is also necessary to analyze the decision trees themselves to judge the importance of each variable. Results show that out of the 13 initial candidate variables, only 5 are actually used in the models. Nevertheless, the decision trees are quite different from one another. With the exception of one case, the model performance is quite acceptable overall.

Delayed courses is the most important variable, as it is the one that is most frequently used as the root node in the tree. An interesting question is if this pattern would hold if more courses were used, or even considering all courses of the University. Since the number of courses is small, it is not possible to generalize conclusions. However, as the architecture of the system is developed, it is possible to increase the number of courses and start correlating results with characteristics of the courses themselves, such as the scientific area or academic year.

As a final remark, this study shows that it is possible to draw conclusions on prediction of student performance using data mining techniques from a set of small experiments on academic data.

6. FUTURE WORK

The study described in this paper will proceed by consolidating and extending the work already carried out. Concerning the model analysis task we propose to understand the reasons for the variability of variables in each course and furthermore, to study alternatives to combine the decision trees into a single consensual tree or a small set of trees that represent general knowledge about the success/failure behavior of students across all the University. To the best of our knowledge, there is only one approach to this problem, which is based on a mathematical perspective (Gorbunov & Lyubetsky, 2011). The main goal would be to develop a data mining approach for combining a large number of decision trees into a smaller one in order to increase the interpretability of the knowledge represented by those trees. The choice of decision trees is motivated by their interpretability. Although our focus is in Educational Data Mining, we believe that such an approach will be interesting for different areas of application.

7. REFERENCES

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. *Department of Education's (ED) Office of Educational Technology* (pp. 1–57). Washington, D.C.

Elias, T. (2011). *Learning analytics: Definitions, processes and potential*. Retrieved from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Gorbunov, K. Y., & Lyubetsky, V. a. (2011). The tree nearest on average to a given set of trees. *Problems of Information Transmission*, *47*(3), 274–288. doi:10.1134/S0032946011030069

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). *The 2011 Horizon Report*. Austin, Texas: The New Media Consortium. Retrieved from http://net.educause.edu/ir/library/pdf/hr2011.pdf

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, *13*(1), 61–72. doi:10.2478/cait-2013-0006

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Conference on AI (IJCAI)* (pp. 1137–1145). San Mateo, CA: Morgan Kaufmann. Retrieved from http://frostiebek.free.fr/docs/Machine Learning/validation-1.pdf

Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2013). C5.0 Decision Trees and Rule-Based Models. Retrieved from http://cran.r-project.org/web/packages/C50/C50.pdf

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Quinlan, R. (1998). C5.0: An Informal Tutorial. Retrieved from http://www.rulequest.com/see5-unix.html

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. doi:10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. doi:10.1002/widm.1075

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, *36*(2), 111–147.

8. AUTHORS' BIOGRAPHIES

Pedro Strecht received his M.Sc. degree in Informatics and Computing Engineering at the Faculty of Engineering of the University of Porto (FEUP). He is Software Developer in the technical team of the University of Porto information system, where he collaborates with requirements elicitation, database design, and development. Currently he is also a PhD student in Informatics Engineering in FEUP, with a research focus in education data mining approaches for strategies to thwart students' attrition in higher education.

João Mendes Moreira received his Ph.D. degree in Engineering Sciences at the Faculty of Engineering of the University of Porto (FEUP). He is an Assistant Professor at the Department of Informatics Engineering, FEUP, and a researcher at LIAAD (Laboratory for Artificial Intelligence and Decision Support at INESC TEC). His research focuses on predictive data mining and its applications.

Carlos Soares received a B.Sc. ("licenciatura") degree in Systems Engineering and Informatics from the Universidade do Minho, Portugal. He received his M.Sc. degree in Artificial Intelligence and his Ph.D. in Computer Science from Universidade do Porto, Portugal. He is now an associate professor at the Faculty of Engineering of the University of Porto and a researcher at INESC TEC. His research interests are in machine learning, data mining and optimization using meta-heuristics. Carlos is particularly interested in applications in management and industrial engineering. He is the (co-)author of many papers and books, most notably in metalearning. Carlos has participated in the organization of many events, most notably KDD 2009 and ECMLPKDD 2012. He was awarded PremeIA (Merit and Excellence Award in AI) by the Portuguese AI Association (APPIA) in 2009.