Forecasting students’ success in an open university

George Kostopoulos*
Educational Software Development Laboratory (ESDLab),
Department of Mathematics,
University of Patras, Greece
Email: kostg@sch.gr
*Corresponding author

Sotiris Kotsiantis
Educational Software Development Laboratory (ESDLab),
Department of Mathematics,
University of Patras, Greece
and
Hellenic Open University,
Parodos Aristotelous 18, 26335 Patras, Greece
Email: sotos@math.upatras.gr

Christos Pierrakeas
Technological Educational Institute of Western Greece,
Meg. Alexandrou 1, 26334 Patras, Greece
and
Hellenic Open University,
Parodos Aristotelous 18, 26335 Patras, Greece
Email: pierrakeas@eap.gr

Giannis Koutsonikos
Technological Educational Institute of Western Greece,
Meg. Alexandrou 1, 26334 Patras, Greece
Email: gkoutson@gmail.com

George A. Gravvanis
Department of Electrical and Computer Engineering,
School of Engineering,
Democritus University of Thrace,
University Campus, Greece
and
Hellenic Open University,
Parodos Aristotelous 18, 26335 Patras, Greece
Email: ggravvan@ee.duth.gr
Abstract: In recent years, students’ performance prediction has been identified as one of the most essential and challenging research topics for educational institutions. The necessity for exploitation and analysis of data originating from several educational contexts has led to a widespread implementation of familiar machine learning methods trying to effectively analyse students’ academic behaviour and predict their performance. The early detection of low performers is of major importance for open universities seeking to decrease dropout ratios, improve educational outcomes and provide high quality education. This paper introduces an ensemble of classification and regression algorithms for predicting students’ performance in a distance web-based course. Several state-of-the-art machine learning methods have also been applied to compare the efficiency of our method. A plethora of experiments have been conducted for this purpose, using data provided by the Hellenic Open University. The proposed ensemble combines classification and regression rules and is as accurate as the powerful ensembles, while the produced model remains comprehensive. In addition, a prototype software support tool has been designed and it simulates the presented ensemble.

Keywords: educational data mining; EDM; performance prediction; classification; regression; ensemble; distance learning; learning technology; open university.


Biographical notes: George Kostopoulos is a PhD student at the Department of Mathematics at the University of Patras. He is a mathematician and holds a Master in Computer Science from the University of Patras, Greece. He has some publications with enough citations. His research interests are in the field of learning analytics and educational data mining.

Sotiris Kotsiantis is a Lecturer in the Department of Mathematics at the University of Patras. He is a mathematician and holds a Master as well as a PhD in Computer Science from the University of Patras, Greece. He has a lot of publications with numerous citations. His research interests are in the field of data mining, machine learning and educational data mining.

Christos Pierrakeas holds a BSc in Mathematics (1986) and a PhD in Medical Informatics (1994) from the University of Patras, Greece. He also holds two Postgraduate Certificates in Open and Distance Learning and Adults Education from the Hellenic Open University (HOU). He is an Assistant Professor at the Technological Educational Institute (TEI) of Western Greece in the Business Administration Department and Tutor at the HOU in the Department of Informatics. His research interests include AI and ICT applications in education, semantic modelling, design, development and evaluation of educational material and educational processes. He has co-authored three books and over 50 papers in international journals and conferences (https://scholar.google.gr/citations?user=d5ybCLIAAAAAJ&hl=en).

Giannis Koutsonikos received his Diploma from the Computer Engineering and Informatics Department (1985) and his PhD in Computer Communication Networks (1991), both in the University of Patras, Greece. He holds two Postgraduate Certificates at the Hellenic Open University in Open and Distance Learning and Adult Education. He is an Assistant Professor at the Business Administration Department – Technological Educational Institute of
Western Greece. His research interests include computer communication networks, internet services and applications, e-learning and e-business. He is the author of more than 25 scientific papers in these domains. He has also participated in more than 30 research and development projects.

George A. Gravvanis is a Professor at the Department of Electrical and Computer Engineering of Democritus University of Thrace. He received Bachelor of Science with Joint Honours in Science from University of Salford (Manchester, UK) in 1982, degree of Mathematics from National and Kapodistrian University Athens in 1986, Master of Science in Numerical of Analysis from Brunel University (London, UK) in 1984 and Doctoral degree (PhD) in Computing Sciences from the Department of Informatics of Athens University of Economics and Business in 1992. His research interests include computational methods, mathematical modelling and applications, and parallel computations.

1 Introduction

Over the last two decades, e-learning has become an innovative way of teaching and learning. Therefore, many universities have focused on open and distance learning by offering flexible and high quality online undergraduate and postgraduate courses. Students have the opportunity to attend distance courses in various scientific fields according to their personal needs and requirements studying at their place and at any time. Unfortunately, the successful completion of a distance learning course is influenced by a number of students’ characteristics (Lykourentzou et al., 2009). Demographic characteristics of students (i.e., age, gender and ethnicity), social environment, family obligations, job commitments, pre-university knowledge background and university performance are just a few of the factors affecting students’ academic performance (Mashiloane and Mchunu, 2013). The identification of students at risk of failing is of utmost importance for open universities seeking to improve educational outcomes and provide high quality education. Therefore, it is important for tutors to understand the students’ learning behaviour and recognise low performers as soon as possible during the academic year. Thus, developing appropriate learning strategies and supporting them with personalised guidance and additional educational material could reduce students’ failure ratios and increase retention rates (Simpson, 2006).

The growing need for exploitation and analysis of data originating from educational databases has resulted in the development of educational data mining (EDM). EDM is an emerging subfield of data mining dealing with the development of methods used to analyse students’ academic behaviour and predict their performance (Romero and Ventura, 2010). The prediction of students’ academic performance has turned into an essential and challenging topic in the educational field and is regarded as one of the most interesting and well-studied aspects of EDM. Classification and regression are familiar machine learning methods that have been successfully applied to the educational field to detect students’ failure or dropout rates (Silva and Fonseca, 2017).

Classification is one of the most frequently studied problems in machine learning. According to this task, a learning model is trained on a set of labelled examples with a view to predicting the output class of an example from a finite set of predefined classes. The most frequently studied classification problem is considered to be the binary
classification where the output variable $y \in \{0, 1\}$, while in multiclass classification $y \in \{0, 1, 2, \ldots, n\}$. The classification task has been widely used to predict students’ performance in the final examinations of a course with $y \in \{\text{pass, fail}\}$, to predict students’ dropout in universities with $y \in \{\text{dropout, no dropout}\}$ or to predict students’ grade (i.e., $y \in \{\text{excellent, very good, good, average, poor}\}$ or $y \in \{A, B, C, D, E, F\}$). Decision trees, decision rules, neural networks (NNs), naïve Bayes (NB), random forest, instance-based learning algorithms and support vector machines (SVMs) are some popular classification methods (Kotsiantis et al., 2007).

On the other hand, regression is used in the case of a continuous output variable, for example when the grade $y \in [0, 10]$. One of the most popular regression methods used in EDM is linear regression (Baker and Inventado, 2014). Various studies deal with the implementation of classification techniques in the educational field in contrast to regression, which is considered to be a more general case (Rwebangira and Lafferty, 2009). It seems that the real valued type of the output variable in regression results in an inherent difficulty in applying classification algorithms for predicting its value (Zhou and Li, 2007).

In this paper, an ensemble of classification and regression algorithms combining classification and regression rules is proposed to estimate students’ academic success in a web-based university course. Several ‘pre-university’ and ‘performance’ attributes of students are used to train the predictive model. The produced ensemble is finally used to predict the performance (pass/fail) and the grade (5–10) of an individual student in the final course examination. Our approach shows significantly better performance than other supervised approaches confirming that ensembles often perform better than the individual classifiers they are created from (Dietterich, 2000). The analysis of the prediction performance indicates that the proposed ensemble schema performs really well, while, at the same time, the percentage of correctly classified instances is increased as new performance attributes are added during the academic year. Moreover, we investigate how early such a prediction could be made, so as to provide a timely intervention and enhance students’ performance. A plethora of experiments were conducted for this purpose using familiar supervised algorithms, concluding that combining classification and regression rules is really beneficial for identifying low performers before the middle of the academic year.

The rest of the paper is structured as follows: In Section 2, we present a short review and recent studies concerning the prediction of students’ performance by applying several data mining methods. A description of the dataset, as well as the corresponding attributes is presented in Section 3. The experimental procedure and the respective results are analysed and presented in Section 4. The proposed ensemble method is introduced in Section 5, while a prototype software tool implementing the proposed method is set out in Section 6. Finally, conclusive remarks and future research directions are given in the last section.

2 Related work

EDM deals with the development and implementation of data mining techniques in educational data (Romero and Ventura, 2010). Various studies have analysed these techniques to identify students that are prone to failure or dropout and predict their
performance. A comprehensive survey of EDM papers published from 1995 to 2005 is presented by Romero and Ventura (2007). Well-known data mining techniques such as classification, clustering, pattern mining, association rules, text mining, outlier detection, statistics and visualisation applied to traditional educational systems, adaptive and intelligent web-based educational systems, learning management systems (LMSs) and web-based courses are explicitly presented and outlined in their survey. In addition to the aforementioned researchers, Baker and Yacef (2009) have included in their survey 64 papers published at the 2008 and 2009 EDM conferences. Moreover, Romero and Ventura (2010) grouped the published EDM papers until 2009 into ten categories: analysis and visualisation of data, provision of feedback for supporting instructors, recommendations for students, prediction of students’ performance, student modelling, detection of undesirable student behaviours, grouping of students, social network analysis, development of concept maps, courseware constructing and course planning and scheduling. Furthermore, Mohamad and Tasir (2013) reviewed nine studies from 2004 to 2012 concentrating on the use of data mining methods in an online learning environment. Classification, clustering, prediction, association rule analysis and sequential pattern were applied to these studies for assessing the learning performance of students participating in several e-learning environments. Following the line of researchers delving into the EDM techniques efficiency, Peña-Ayala (2014) analysed representative EDM papers presented in conferences and workshops, published in journals and chapter books from 2010 to 2013, organising them in seven categories: Student modelling, student behaviour modelling, student performance modelling, assessment, student support and feedback, curriculum and EDM tools. Recently, Romero and Ventura (2017) made a new attempt to describe four issues in massive open online courses (MOOCs) that had been solved by using educational data science (EDS) methods:

a. analysing students’ interactions
b. predicting students at risk of dropout
c. grading, assessing and providing feedback to students
d. using adaptive learning systems and making recommendations to students.

In a very recent study, Costa et al. (2017) examined the effectiveness of four common supervised techniques for early prediction of students that are likely to fail in two introductory programming courses in a Brazilian university. The first dataset concerned 262 students attending a 10 weeks distance course and the second one concerned 161 students attending a 16 weeks on-campus course. Both datasets contained several demographic, social and university attributes. The experimental results showed that EDM methods are effective in the early identification of low performers, while in addition the effectiveness was improved after data pre-processing and fine-tuning of algorithms.

Based on the fact that students’ performance in an online course is highly related to their engagement in several course activities, Shelton et al. (2017) examined the hypothesis that students with similar profiles and learning patterns will have similar performance outcomes. So, they studied the impact of students’ interaction within a LMS. Demographic and dynamic course data regarding students attending a 16-week semester course were gathered, while traditional and time-series clustering were used by applying four familiar supervised methods. The results demonstrated that the frequency

Based on the fact that students’ performance in an online course is highly related to their engagement in several course activities, Shelton et al. (2017) examined the hypothesis that students with similar profiles and learning patterns will have similar performance outcomes. So, they studied the impact of students’ interaction within a LMS. Demographic and dynamic course data regarding students attending a 16-week semester course were gathered, while traditional and time-series clustering were used by applying four familiar supervised methods. The results demonstrated that the frequency
Forecasting students’ success in an open university

of interaction in a course is a better indicator of students’ success than the total amount of interaction, whereas students at risk of failure could be identified after the 10th week of the course.

To sum up, several predictive models based on data mining methods have been developed over recent years extracting knowledge from educational data to gain an understanding on the academic behaviour of students and predict their performance. However, only few of the related studies utilise ensemble methods to classify students into two groups: {pass, fail} or {dropout, no dropout}. In a very recent study, Amrigh et al. (2016) applied familiar ensemble methods to classify students into three classes: {high, medium, low}. The results verified that ensemble methods outperform common supervised methods.

Motivated by the abovementioned study, in this paper an attempt is made to put forward an ensemble of classification and regression algorithms combining both classification and regression rules to predict students’ grade in the final examinations of a distance learning course. To the best of our knowledge, there is no study dealing with the implementation of a two level ensemble method of this kind to predict students’ performance in distance learning.

3 Research questions and dataset description

This study focuses on the following three research questions:

1. Is it possible to combine classification and regression rules to forecast students’ success in a distance university course?
2. How efficient is the proposed ensemble method compared to familiar machine learning methods in terms of classification accuracy?
3. Is the timely prediction of students’ failure in the final examinations achievable?

3.1 Data selection

The dataset used for this research (available here) was provided by the Hellenic Open University (HOU). The HOU is the only Greek State University offering undergraduate and postgraduate courses through distance learning. At the undergraduate level, students are required to successfully complete 12 course modules, while they are allowed to register for a maximum of three modules per academic year. The successful completion of a course module requires the submission of four written assignments and a grade equal to or higher than five in the final examination (ten grade scale). Students can undertake the final examination of the module only if they have successfully completed the written assignments with a total score of 20 or more. Moreover, students may participate in five optional four-hour contact sessions with their tutors during the academic year.

Data regarding students that attended the ‘introduction to informatics’ module of the ‘computer science’ course were collected for a period of three years’ time (2008–2010). A total of 3882 instances form the dataset, and each instance is characterised by the values of 17 attributes.
3.2 Dataset attributes

For the purpose of our study, the dataset was divided into three attribute groups, named ‘pre-university’, ‘Contact session’ and ‘performance’ group. The ‘pre-university’ attributes group contains the following students’ information: gender, age, marital status, number of children, computer knowledge, computer usage at work, type of employment and new student (Table 1). Previous education in the field of computer science as well as employment in the information technology field was taken into consideration. A student who had attended a seminar or an introductory computer course of at least 100 hours was characterised as having computer knowledge. Studies have shown that computer knowledge is crucial for students attending online courses (Simpson, 2006). Moreover, a student who used simple software programs at work was characterised as a ‘junior’ user, while a student who worked as a programmer was characterised as a ‘senior’ user. Additionally, a student was characterised as ‘no’ user if his occupation had no association with computer usage. Furthermore, a first year student was characterised as a new one, with an attribute value ‘yes’, while an attribute value ‘no’ corresponded to a student that had failed to pass the module during the previous year and retook the examination. The above mentioned ‘pre-university’ attributes are also referred to as time-invariant attributes and they form the core of several studies for predicting students’ performance and dropout rates in higher education. It is worth noting that many studies indicate that such attributes have a material impact on students’ academic success (Navarro and Shoemaker, 2000), which is also confirmed by the experimental results of our study.

Table 1 Pre-university attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male, female</td>
</tr>
<tr>
<td>Age</td>
<td>24–46</td>
</tr>
<tr>
<td>Marital status</td>
<td>Single, married, widowed, divorced</td>
</tr>
<tr>
<td>Number of children</td>
<td>0, 1, 2, 3, &gt; 3</td>
</tr>
<tr>
<td>Computer knowledge</td>
<td>No, yes</td>
</tr>
<tr>
<td>Employment associated with computers</td>
<td>No, senior, junior</td>
</tr>
<tr>
<td>Type of employment</td>
<td>No, part-time, full-time, over-time</td>
</tr>
<tr>
<td>New student</td>
<td>No, yes</td>
</tr>
</tbody>
</table>

The ‘contact session’ group includes five attributes (OCSi, i = 1, 2, 3, 4, 5) related to students’ absence or presence in the five optional contact sessions, which are held at weekends throughout the academic year and before the scheduled written assignments submission deadline. These face-to-face meetings are useful and beneficial for students and offer opportunities for interaction with other students, communication with tutors and clarification of misunderstandings (Tsagari, 2013). Finally, the ‘performance’ group represents four attributes (WRIi, i = 1, 2, 3, 4) corresponding to students’ grades in the four written assignments. At least three written assignments are compulsory and must be submitted online by the due date mentioned in the module study plan. Tutors evaluate written assignments according to the 10-grade scale and an average grade equal to or greater than five is required for a student to attend the final examinations of the module.
For the ‘final examination grade’ output attribute two familiar supervised learning approaches were used:

- Binary classification with two output classes: {pass, fail}. The value ‘pass’ corresponds to a student who successfully completed the module with a grade equal to or greater than five in the final examinations.
- Regression for continuous output (final grade between 5 and 10).

It should be noted that the ‘contact session’ and ‘performance’ attributes are added consecutively during the academic year and that they affect the timely prediction of students’ performance, as demonstrated at the experimental results and presented in the following section.

3.3 Performance measures

The performance of the learning algorithms applied to our study was evaluated using two popular statistical metrics: accuracy and mean absolute error (MAE). Accuracy is one of the most frequently used measures for assessing the overall effectiveness of a classification algorithm (Sokolova et al., 2006). For binary classification problems, accuracy corresponds to the percentage of correctly classified instances and is defined as follows:

\[ \text{Accuracy} = \frac{TP + TN}{n} \]  

1

\( TP \) a student that passed and is classified as passed
\( TN \) a student that did not pass and is classified as did not pass
\( N \) number of instances (students).

MAE is considered to be a very effective metric for measuring the average error of a regression model (Willmott and Matsuura, 2005) by showing how much the predictions deviate from the true probability (Ferri et al., 2009) and is defined as follows:

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \]  

\( y_i \) the true value of instance \( x_i \)
\( \hat{y}_i \) the predicted value of instance \( x_i \).

4 Experimental setup and results

The experiments were conducted following ten distinct steps. The 1st step included all the ‘pre-university’ attributes, while in the next steps, the ‘contact session’ and ‘performance’ attributes were added successively. Specifically, in the 2nd step, the attribute OCS1 (presence or absence in the first optional contact session) was added to the previous group of attributes, while in the 3rd step the attribute WRI1 regarding the grade of the first written assignment was added. The procedure was repeated until all
variables described in Section 2 were used as shown in Figure 1. Furthermore, student
data were collected in 40 groups associated with the corresponding tutors for the next
academic year (2010–2011) and each one of these 40 groups was used to estimate the
classification accuracy of the implementing classifiers within these groups. The testing
phase steps were in correspondence with the training phase steps.

**Figure 1** Attributes used in the experiments steps (see online version for colours)

### 4.1 First phase of experiments

In the first phase of experiments, we applied a set of six classification algorithms and a
set of three ensemble learning methods. The first set consisted of familiar algorithms
corresponding to different classification methods, such as NNs, NB, SVMs, decision
trees and rule-based methods. More specifically we used:

- The C4.5 decision tree (Quinlan, 1993), a widely used classification algorithm
categorising instances to a predefined set of classes according to their attribute
values from the root of a tree down to a leaf.

- The reduced error pruning tree (REPTree) classifier, a fast and effective decision tree
learner based on the information gain or variance (Witten and Frank, 2005).

- The 3-nearest neighbour (3-NN) instance based learner (Aha, 1997), which uses a
distance function to predict the output class of an instance.

- The radial basis function (RBF) classifier, a representative NNs algorithm
(Lippmann, 1989) in which a new example is classified by computing its Euclidean
distance to a set of N centres chosen by unsupervised methods (Wettschereck and
Dietterich, 1992).

- The NB statistical learning classifier, a very effective classification algorithm,
representative of the Bayesian networks (John and Langley, 1995) based on Bayes’
theorem.

- The sequential minimisation optimisation (SMO) algorithm (Platt, 1999), a simple,
fast and efficient algorithm based on SVMs.
The second set consists of three ensemble learning methods: AdaBoost, LogitBoost and rotation forest. Ensemble learning or committee-based learning or learning with multiple classifier systems, concerns the generation and training of a set of classifiers for solving the same problem (Zhou, 2009). There is a growing conviction that a combination of multiple learners is more effective than a single learner, particularly in the case that the component learners are as accurate and diverse as possible. It should be underlined that ensemble learning method has become increasingly popular in the last decade with appreciable results. Bagging and boosting constitute two very effective and popular ensemble learning methods (Opitz and Maclin, 1999). Bagging or bootstrap aggregating is a method for generating multiple learning sets from which an aggregated predictor is built (Breiman, 1996). On the other hand, boosting builds an ensemble of classifiers by changing the distribution of the training data based on the accuracy of the previous learners. AdaBoost (Freund and Schapire, 1996) and LogitBoost (Friedman et al., 2000) are known boosting algorithms. Rodriguez et al. (2006) proposed rotation forest, a powerful ensemble algorithm, in which diversity is promoted by using principal component analysis (PCA) for each base classifier.

The free available source codes (Witten et al., 2011) for the abovementioned algorithms have been used in our study. The classification results of the experiments (Table 2) show that the percentage of correctly classified instances ranges from 69.56% in the initial step (based only on demographic characteristics of students) and exceeds 86% in the final step. It is concluded that the classification accuracy is quite similar for the machine learning methods used for comparative reasons. The ensembles are a bit more accurate, but their weakness is the decreased understandability. With the usage of multiple learners it is more difficult to comprehend the underlying reasoning behind a decision. Thus, for comprehensive reasons, the classification rules are more easily understood. The produced classification rules in every step resulting from the application of the REPTree algorithm (with maximum tree depth = 2) are presented in Table 3, showing the influence of the following attributes in predicting students’ performance: employment, new student, OCS2, OCS4, OCS5, WRI1, 2, and 3.

### Table 2

<table>
<thead>
<tr>
<th>Step</th>
<th>NB</th>
<th>RBF</th>
<th>SMO</th>
<th>3NN</th>
<th>C4.5</th>
<th>REPTree</th>
<th>AdaBoost</th>
<th>LogitBoost</th>
<th>Rotation forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>69.56</td>
<td>70.03</td>
<td>70.07</td>
<td>69.84</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
</tr>
<tr>
<td>2nd</td>
<td>62.04</td>
<td>69.98</td>
<td>70.07</td>
<td>69.28</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
<td>70.07</td>
</tr>
<tr>
<td>3rd</td>
<td>70.13</td>
<td>72.55</td>
<td>72.79</td>
<td>71.55</td>
<td>73.98</td>
<td>73.98</td>
<td>75.04</td>
<td>75.72</td>
<td>75.87</td>
</tr>
<tr>
<td>4th</td>
<td>72.57</td>
<td>74.12</td>
<td>77.94</td>
<td>73.87</td>
<td>74.12</td>
<td>74.12</td>
<td>76.67</td>
<td>77.32</td>
<td>77.63</td>
</tr>
<tr>
<td>5th</td>
<td>77.06</td>
<td>77</td>
<td>80.03</td>
<td>72.6</td>
<td>74.87</td>
<td>74.87</td>
<td>81.43</td>
<td>82.32</td>
<td>81.36</td>
</tr>
<tr>
<td>6th</td>
<td>78.46</td>
<td>77.32</td>
<td>81</td>
<td>73.17</td>
<td>75.99</td>
<td>75.99</td>
<td>81.76</td>
<td>82.51</td>
<td>82.1</td>
</tr>
<tr>
<td>7th</td>
<td>78.62</td>
<td>79.49</td>
<td>80.81</td>
<td>70.09</td>
<td>76.02</td>
<td>76.02</td>
<td>83.28</td>
<td>83.47</td>
<td>82.9</td>
</tr>
<tr>
<td>8th</td>
<td>80.69</td>
<td>79.65</td>
<td>81.49</td>
<td>73.74</td>
<td>77.05</td>
<td>77.05</td>
<td>83.35</td>
<td>83.91</td>
<td>83.53</td>
</tr>
<tr>
<td>9th</td>
<td>80.52</td>
<td>82.04</td>
<td>81.5</td>
<td>68.9</td>
<td>76.98</td>
<td>76.98</td>
<td>83.68</td>
<td>85.27</td>
<td>84.91</td>
</tr>
<tr>
<td>10th</td>
<td>81.44</td>
<td>82.52</td>
<td>81.9</td>
<td>74.25</td>
<td>78.12</td>
<td>78.12</td>
<td>84.74</td>
<td>86.48</td>
<td>85.84</td>
</tr>
</tbody>
</table>
Table 3  The produced classification rules

<table>
<thead>
<tr>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>If Employment = No then Pass else Fail (accuracy = 70.35%)</td>
</tr>
<tr>
<td>If WRI1 &gt; 8.75 and NewStudent = yes then Pass else Fail (accuracy = 75.85%)</td>
</tr>
<tr>
<td>If WRI1 &gt; 8.75 and OCS2 = present then Pass else Fail (accuracy = 77.06%)</td>
</tr>
<tr>
<td>If WRI2 &gt; 6.75 and NewStudent = yes then Pass else Fail (accuracy = 79.80%)</td>
</tr>
<tr>
<td>If WRI3 &gt; 6.75 and NewStudent = yes then Pass else Fail (accuracy = 81.74%)</td>
</tr>
<tr>
<td>If WRI3 &gt; 1.25 and OCS4 = present then Pass else Fail (accuracy = 81.87%)</td>
</tr>
<tr>
<td>If (Average(WRIi) &gt; 7.5 and NewStudent = yes) OR (OCS5 = present) then Pass else Fail (accuracy = 85.04%)</td>
</tr>
</tbody>
</table>

4.2 Second phase of experiments

In the second phase of experiments, we applied a set of five regression methods to predict students’ grade (0–10) in the final course examination. These methods are:

- linear regression (LR) (Weisberg, 2005)
- the M5’ algorithm, a machine learning method for inducing trees of regression models (Wang and Witten, 1996)
- the M5’ rules algorithm (Holmes et al., 1999) for inducing simple and accurate rule sets and propositional regression rules (Frank et al., 1998) from model trees
- the RBF networks, a type of artificial NNs for application to regression problems (Du and Swamy, 2014)
- the 3-NN instance based learner (Aha, 1997).

Table 4  Regression results (MAE)

<table>
<thead>
<tr>
<th>Step</th>
<th>M5’</th>
<th>M5’ rules</th>
<th>LR</th>
<th>RBF</th>
<th>3NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1.20</td>
<td>1.20</td>
<td>1.18</td>
<td>1.18</td>
<td>1.18</td>
</tr>
<tr>
<td>2nd</td>
<td>1.18</td>
<td>1.18</td>
<td>1.15</td>
<td>1.14</td>
<td>1.12</td>
</tr>
<tr>
<td>3rd</td>
<td>0.91</td>
<td>0.90</td>
<td>1.09</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>4th</td>
<td>0.84</td>
<td>0.84</td>
<td>1.04</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td>5th</td>
<td>0.70</td>
<td>0.69</td>
<td>0.99</td>
<td>0.80</td>
<td>0.88</td>
</tr>
<tr>
<td>6th</td>
<td>0.69</td>
<td>0.68</td>
<td>0.95</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>7th</td>
<td>0.65</td>
<td>0.65</td>
<td>0.92</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td>8th</td>
<td>0.64</td>
<td>0.64</td>
<td>0.89</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>9th</td>
<td>0.55</td>
<td>0.56</td>
<td>0.89</td>
<td>0.69</td>
<td>0.80</td>
</tr>
<tr>
<td>10th</td>
<td>0.55</td>
<td>0.56</td>
<td>0.86</td>
<td>0.68</td>
<td>0.73</td>
</tr>
</tbody>
</table>
The MAE is calculated for the abovementioned regression algorithms in each step of the experiments. The results are presented in Table 4. A progressive reduction of the MAE is noted as the ‘contact sessions’ and ‘performance’ attributes are added during the academic year. Moreover, the M5’ and M5’ rules algorithms prevail with a MAE value decreasing from 1.20 (1st step) to 0.55 (10th step).

5 The proposed method

Our contribution in this paper is a combination of the REPTree classification algorithm (with maximum tree depth = 2) and the M5’ rules regression algorithm, which is based on the following schema (Table 5).

Table 5 Proposed schema

<table>
<thead>
<tr>
<th>Rules</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IF Prediction of REPTree Algorithm = FAIL and Prediction of M5’ Rules Regression Algorithm &lt; 6 then Final Prediction = FAIL</td>
<td></td>
</tr>
<tr>
<td>IF Prediction of REPTree Algorithm = FAIL and Prediction of M5’ Rules Regression Algorithm &gt;= 6 then Final Prediction = 5</td>
<td></td>
</tr>
<tr>
<td>IF Prediction of REPTree Algorithm = PASS and Prediction of M5’ Rules Regression Algorithm &gt;= 6 then Final Prediction = Prediction of Regression Algorithm</td>
<td></td>
</tr>
<tr>
<td>IF Prediction of REPTree Algorithm = PASS and Prediction of M5’ Rules Regression Algorithm &lt; 6 then Final Prediction = 5</td>
<td></td>
</tr>
</tbody>
</table>

The accuracy and MAE values for the proposed ensemble in each step of the experiments for the particular data are presented in Table 6. It is clear that the accuracy is constantly increasing while ‘contact session’ and ‘performance’ attributes are gradually added, with an accuracy measure ranging from 70.07% to 87.20%. At the same time, the MAE value is steadily decreasing from 1.18 in the 1st step to 0.55 in the final 10th step.

Table 6 Accuracy (%) and MAE of the proposed schema

<table>
<thead>
<tr>
<th>Step</th>
<th>Accuracy</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>70.07</td>
<td>1.18</td>
</tr>
<tr>
<td>2nd</td>
<td>70.07</td>
<td>1.18</td>
</tr>
<tr>
<td>3rd</td>
<td>76.20</td>
<td>0.90</td>
</tr>
<tr>
<td>4th</td>
<td>77.72</td>
<td>0.84</td>
</tr>
<tr>
<td>5th</td>
<td>82.25</td>
<td>0.69</td>
</tr>
<tr>
<td>6th</td>
<td>82.69</td>
<td>0.68</td>
</tr>
<tr>
<td>7th</td>
<td>83.56</td>
<td>0.65</td>
</tr>
<tr>
<td>8th</td>
<td>83.87</td>
<td>0.64</td>
</tr>
<tr>
<td>9th</td>
<td>85.19</td>
<td>0.55</td>
</tr>
<tr>
<td>10th</td>
<td>87.20</td>
<td>0.55</td>
</tr>
</tbody>
</table>
The performance of the classification methods used in our study and the proposed ensemble is evaluated and the existence of statistically significant differences among them is examined by using the Friedman aligned ranks non-parametric test (Hodges and Lehmann, 1962). The null hypothesis (Ho) that the mean of the results of the proposed ensemble compared to any other method is equal is rejected (p-value = 0.000001 with a significance level of 0.05). According to the test results (Table 7), the methods are ranking from the best performer to the lower one (Holm, 1979). Therefore, the proposed ensemble takes precedence over all other methods, followed by LogitBoost, rotation forest and AdaBoost.

Table 7  Friedman aligned ranks test (classification algorithms vs. proposed algorithm)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rank test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed algorithm</td>
<td>19.7</td>
</tr>
<tr>
<td>Logitboost</td>
<td>21.2</td>
</tr>
<tr>
<td>Rotation forest</td>
<td>24.7</td>
</tr>
<tr>
<td>Ada boost</td>
<td>28.4</td>
</tr>
<tr>
<td>SMO</td>
<td>43.5</td>
</tr>
<tr>
<td>RBF</td>
<td>61.5</td>
</tr>
<tr>
<td>NB</td>
<td>71.8</td>
</tr>
<tr>
<td>C4.5</td>
<td>74.45</td>
</tr>
<tr>
<td>Ripper</td>
<td>74.45</td>
</tr>
<tr>
<td>3-NN</td>
<td>85.3</td>
</tr>
</tbody>
</table>

We evaluate the performance of the regression algorithms and the proposed ensemble and examine all the pairwise differences among them using the Friedman non-parametric test (Friedman, 1937). The null hypothesis (Ho) that the mean of the results of two or more algorithms is the same is rejected (p-value = 0.00072 with a significance level of 0.05). According to the test results (Table 8), the algorithms are ranking from the best performer to the lower one. Therefore, the proposed algorithm takes precedence over all others, followed by M5’ rules and M5’.

Table 8  Friedman test (regression algorithms vs. proposed algorithm)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rank test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed algorithm</td>
<td>2.15</td>
</tr>
<tr>
<td>M5’ rules</td>
<td>2.75</td>
</tr>
<tr>
<td>M5’</td>
<td>2.95</td>
</tr>
<tr>
<td>RBF</td>
<td>3.85</td>
</tr>
<tr>
<td>3-NN</td>
<td>3.95</td>
</tr>
<tr>
<td>LR</td>
<td>5.35</td>
</tr>
</tbody>
</table>

Finally, a timeline prognosis is presented in Figure 2 showing the accuracy percentage in predicting students at risk of failing during the academic year. It can be easily observed that low performers can be identified with an accuracy of 82.25% in the middle of the academic year. Hence, educators are able to apply early intervention strategies.
A prototype software support tool has been designed by implementing the presented ensemble (Figure 3) and may be used by tutors for the prediction of students’ performance. The software tool consists of a two column attribute table and a bottom tab for the prediction of the final grade value. It should be mentioned that the prototype software support tool can be used for the input of an even more informative dataset and retrain the presented ensemble for more accurate results.

6 Implementation of the proposed ensemble

A prototype software support tool has been designed by implementing the presented ensemble (Figure 3) and may be used by tutors for the prediction of students’ performance. The software tool consists of a two column attribute table and a bottom tab for the prediction of the final grade value. It should be mentioned that the prototype software support tool can be used for the input of an even more informative dataset and retrain the presented ensemble for more accurate results.

Figure 2 Timeline prognosis (see online version for colours)

Figure 3 Screenshot of the prediction tool (see online version for colours)
7 Conclusions

In recent years, an increased emphasis on students’ performance prediction is observed. Various significant studies demonstrate that machine learning can be an efficient tool for open universities so as to estimate students’ success in distance learning courses with sufficient accuracy. This study introduces an ensemble of machine learning algorithms for estimating students’ academic performance in distance courses at higher education level. Ensemble methods have been effectively applied with great success to other scientific fields and often perform better than the individual classifiers they are created from (Dietterich, 2000). The proposed ensemble combines classification and regression rules for the prediction of students’ academic performance (pass/fail) and the grade (5–10) in the final examination of a web-based course.

The suggested ensemble was compared with known classification and regression methods and achieved excellent results. A plethora of experiments were conducted using a dataset provided by the HOU. The experimental results illustrate the advantage of the proposed method, since it is more accurate than the powerful ensembles, confirming the superiority of ensemble methods. The proposed method outperforms several classification and regression methods with an accuracy ranging from 70.07% at the initial step, based only on demographic characteristics of the students, and exceeds 87% before the final examination. The experimental results indicate that students at risk of failing could be identified before the middle of the academic year (5th step of experiments) with an accuracy of 82.25%. The early identification of low performers is of utmost importance for educational institutions targeting to offer high quality education, to reduce dropout ratios and to increase retention rates. Supplementary lessons, additional learning material and personalised guidance to students that are likely to fail may motivate them and enhance their performance.

For future work, an interesting aspect would be the implementation of semi-supervised learning (SSL) and active learning (AL) techniques for predicting students’ performance and dropout rates in educational institutions. These methods constitute the two primary paradigms for learning with labelled and unlabeled data. Several SSL methods were thoroughly studied in Kostopoulos et al. (2015), confirming their superiority in comparison with familiar supervised techniques for predicting students’ academic performance. According to a recent study, AL methods perform really well for the early identification of low performance students in distance courses at higher education level (Kostopoulos et al., 2017). SSL and AL exploit a limited set of labelled data together with a large set of unlabeled ones and achieve the same or even better level of accuracy as supervised methods in the whole dataset. Moreover, the lack of labelled data shows that the application of SSL and AL methods to identify students at risk in sufficient time is extremely useful for educators.

Combining SSL and AL in the educational field is a challenging topic for future research, since these methods have effectively been applied to other domains as well. In this context, a combination of semi-supervised methods and ensemble learning may boost the prediction’s performance, which is an interesting topic to be studied in the future. Zhou (2009) showed that SSL and ensemble learning may be beneficial to one other, since unlabeled data are in abundance and can be fully exploited by a combination of diverse classifiers.
References


