Chapter

Data Mining for Student Performance Prediction in Education

Ferda Ünal

Abstract

The ability to predict the performance tendency of students is very important to improve their teaching skills. It has become a valuable knowledge that can be used for different purposes; for example, a strategic plan can be applied for the development of a quality education. This paper proposes the application of data mining techniques to predict the final grades of students based on their historical data. In the experimental studies, three well-known data mining techniques (decision tree, random forest, and naive Bayes) were employed on two educational datasets related to mathematics lesson and Portuguese language lesson. The results showed the effectiveness of data mining learning techniques when predicting the performances of students.

Keywords: data mining, student performance prediction, classification

1. Introduction

Recently, online systems in education have increased, and student digital data has come to big data size. This makes possible to draw rules and predictions about the students by processing educational data with data mining techniques. All kinds of information about the student's socioeconomic environment, learning environment, or course notes can be used for prediction, which affect the success or failure of a student.

In this study, the successes of the students at the end of the semester are estimated by using the student data obtained from secondary education of two Portuguese schools. The aim of this study is to predict the students' final grades to support the educators to take precautions for the children at risk. A number of data preprocessing processes were applied to increase the accuracy rate of the prediction model. A wrapper method for feature subset selection was applied to find the optimal subset of features. After that, three popular data mining algorithms (decision tree, random forest, and naive Bayes) were used and compared in terms of classification accuracy rate. In addition, this study also investigates the effects of two different grade categorizations on data mining: five-level grade categorization and binary grade categorization.

The remainder of this paper is organized as follows. In Section 2, the previous studies in this field are mentioned. In Section 3, the methods used in this study are

briefly explained to provide a comprehensive understanding of the research concepts. In Section 4, experimental studies are presented with dataset description, data preprocessing, and experimental result subtitles. Finally, conclusion and the direction for future research are given in Section 5.

2. Related work

Predicting students' academic performance is one of the main topics of educational data mining [1, 2]. With the advancement of technology, technological investments in the field of education have increased. Along with technological developments, e-Learning platforms such as web-based online learning and multimedia technologies have evolved, and both learning costs have decreased, and time and space limitations have been eliminated [3]. The increase of online course trainings and the increase of online transactions and interactive transactions in schools led to the increase of digital data in this field. Costa (2017) emphasized the data about the failure rate of the students; the educators were concerned and raised important questions about the failure prediction [4].

Estimating students' performances becomes more difficult because of the large volume of data in training databases [5]. Descriptive statistical analysis can be effectively used to provide the basic descriptive information of a given set of data [6]. However, this alone is not always enough. To inform the instructors and students early, students may be able to identify early, using estimated modeling methods [7]. It is useful to classify university students according to their potential academic performance in order to increase success rates and to manage the resources well [8]. The large growth of electronic data from universities leads to an increase in the need to obtain meaningful information from these large amounts of data [9]. By using data mining techniques on education data, it is possible to improve the quality of education processes [10].

Until now, data mining algorithms have been applied on various different educational fields such as engineering education [11], physical education [12], and English language education [13]. Some studies have focused on high school students [14], while some of them have interested in higher education [15]. Whereas some data mining studies have focused on the prediction of student performance [16], some studies have investigated the instructor performance [17].

3. Method

The increase in digitalization caused us to have plenty of data in every field. Having too much data is getting worth if we know how to use it. *Data mining* aims to access knowledge from data using various machine learning techniques. With data mining, it becomes possible to establish the relationships between the data and make accurate predictions for the future. One of the application areas of data mining is education. *Data mining in education* is the field that allows us to make predictions about the future by examining the data obtained so far in the field of education by using machine learning techniques. There are basically three data mining methods: *classification*, *clustering*, and *association rule mining*. In this study, we focus on the classification task.

The methods to be used in data mining may differ depending on the field of study and the nature of the data we have. In this study, three well-known

classification algorithms (decision tree, random forest, and naive Bayes) were employed on the educational datasets to predict the final grades of students.

3.1 Naive Bayes

Naive Bayes classifiers are a family of algorithms. These classifiers are based on Bayes' Theorem, which finds the possibility of a new event based on previously occurring events. Each classification is independent of one another but has a common principle.

3.2 Decision tree

A decision tree uses a tree like graph. Decision trees are like flowchart but not noncyclic. The tree consists of nodes and branches. Nodes and branches are arranged in a row. Root node is on the top of a tree and represents the entire dataset. Entropy is calculated when determining nodes in a tree. It models decisions with efficacy, results, and resource costs. In this study, decision tree technique is preferred because it is easy to understand and interpret.

3.3 Random forest

Random forest is an ensemble learning algorithm. It is a supervised classification method. It consists of randomly generated many decision trees. The established forest is formed by the decision trees community trained by the bagging method, which is one of the ensemble methods. Random forest creates multiple decision trees and combines them to achieve a more accuracy rates and stable prediction.

Figure 1 illustrates the workflow of data mining model for classification. In the first step, feature selection algorithms are applied on the educational data. Next, classification algorithms are used to build a good model which can accurately map inputs to desired outputs. The model evaluation phase provides feedback to the feature selection and learning phases for adjustment to improve classification performance. Once a model is built, then, in the second phase, it is used to predict label of new student data.

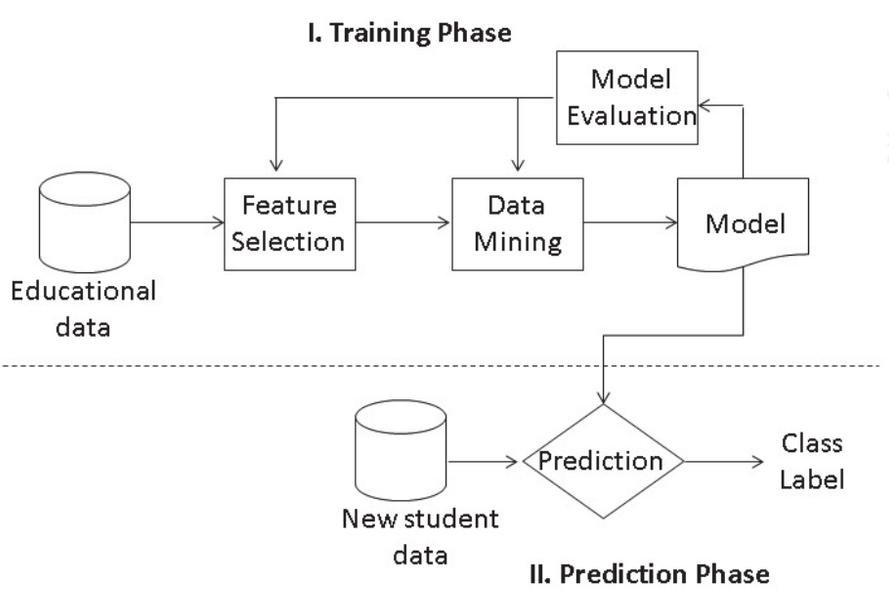

Figure 1.
Flowchart of the data mining model.

4. Experimental studies

In this study, the feature subset selection and classification operations were conducted by using WEKA open-source data mining software [18]. In each experiment, 10-fold cross-validation was performed to evaluate the classification models. The classification accuracy of the algorithm for the test dataset was measured as given in Eq. 1:

$$\text{accuracy}(T) = \frac{\sum_{i=1}^{|T|} \text{eval}(t_i)}{|T|} \qquad \text{eval}(t) = \begin{cases} 1, \text{if } \text{classify}(t) = c \\ 0, \text{otherwise} \end{cases} \tag{1}$$

where T is a test set that consists of a set of data items to be classified; c is the actual class of the item t, where $t \in T$; and *classify(t)* returns the classification output of t by the algorithm.

4.1 Dataset description

In this study, two publically available datasets [19] were used to predict student performances. Both datasets were collected from secondary education of two Portuguese schools. Dataset attributes are about student grades and social, demographic, and school-related features. All data were obtained from school reports and questionnaires. The first dataset has information regarding the performances of students in Mathematics lesson, and the other one has student data taken from Portuguese language lesson. Both datasets have 33 attributes as shown in **Table 1**.

4.2 Data preprocessing

In the raw dataset, the final grade is in the range of 0–20 as with many European countries, where 0 is the worst grade and 20 is the best score. Since the final grade of the students is in the form of integer, the predicted class should be in the form of categorical values, the data needed to be transformed to categories according to a grading policy. In the study, we used and compared two different grading systems: five-level grading and binary grading systems.

We first categorized the final grade in five groups. These ranges are defined based on the Erasmus system. As shown in **Table 2**, the range 0–9 refers to grade F, which is the worst grade and corresponds to "fail" label. The others (10–11, 12–13, 14–15, and 16–20) correspond to D (sufficient), C (satisfactory), B (good), and A (excellent/very good) class labels, respectively.

To compare the results, we also categorized the final grade as "passed" and "fail." As shown in **Table 3**, the range of 0–9 corresponds to F, and it means "fail"; the range of 10–20 refers to A, B, C, and D, and it means "pass."

4.3 Experimental results

As a preprocessing operation, the final grade attribute was categorized according to two different grading systems, before classification. As a result, we have created two versions of both datasets. Both math and Portuguese datasets were available in both five-level and binary grading versions. Hence, we have the chance to compare the results of these versions.

In the first experiment, three algorithms [decision tree (J48), random forest, and naive Bayes] were compared on the five-level grading version and binary version of

Feature	Description	Type	Values
Sex	The gender of the student	Binary	Female or male
Age	The age of the student	Numeric	From 15 to 22
School	The school of the student	Binary	GP (Gabriel Pereira) or MS (Mousinho da Silveira)
Address	Home address type of student	Binary	Urban or rural
Pstatus	Cohabitation status of student's parent	Binary	Living together or apart
Medu	Education of student's mother	Numeric	From 0 to 4
Mjob	Job of student's mother	Nominal	Teacher, health, services, at home, others
Fedu	Education of student's father	Numeric	From 0 to 4
Fjob	Job of student's father	Nominal	Teacher, health, services, at home, others
Guardian	Guardian of student	Nominal	Mother, father, or otherd
Famsize	Size of family	Binary	"LE3" (less or equal to 3) or "GT3" (greater than 3)
Famrel	Quality of family relationships	Numeric	From 1 very bad to 5 excellent
Reason	Reason of choosing this school	Nominal	Close to home, school reputation, course preference, or others
Travel time	Travel time of home to school	Numeric	1–<15 min., 2–15 to 30 min., 3–30 min. to 1 hour, or 4–>1 hour
Study time	Study time of a week	Numeric	–< 2 hours, 2–2 to 5 hours, 3–5 to 10 hours or 4– > 10 hours
Failures	Number of past class failures	Numeric	n if 1 < =n < 3, else 4
Schoolsup	Extra educational school support	Binary	Yes or no
Famsup	Family educational support	Binary	Yes or no
Activities	Extracurricular activities	Binary	Yes or no
Paid class	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
Free time	Free time after school	Numeric	From 1 (very low) to 5 (very high)
Go out	Going out with friends	Numeric	From 1 (very low) to 5 (very high)
Walc	Alcohol consumption of weekend	Numeric	From 1 (very low) to 5 (very high)
Dalc	Alcohol consumption of workday	Numeric	From 1 (very low) to 5 (very high)
Health	Status of current health	Numeric	From 1 (very low) to 5 (very high)
Absences	Number of school absences	Numeric	From 0 to 93
G1	Grade of first period	Numeric	From 0 to 20
G2	Grade of second period	Numeric	From 0 to 20
G3	Grade of final period	Numeric	From 0 to 20

Table 1.
The main characteristics of the dataset.

1	2	3	4	5
Excellent/very good	Good	Satisfactory	Sufficient	Fail
16–20	14–15	12–13	10–11	0–9
A	B	C	D	F

Table 2.
Five-level grading categories.

Pass	Fail
10–20	0–9
A, B, C, D	F

Table 3.
Binary fail/pass category.

the Portuguese dataset. As shown in **Table 4**, the best performance for the five-level grading version for this dataset was obtained with an accuracy rate of 73.50% with the random forest algorithm. However, this accuracy rate was increased with binary grading version of this dataset. In the dataset, where the final grade is categorized in binary form (passing or failing), the accuracy rate was increased to 93.07%.

The performances of three classification algorithms on mathematics datasets (five-level and binary label dataset versions) are shown in **Table 5**. The best results for five-level grading version were obtained with the decision tree (J48) algorithm with an accuracy rate of 73.42%. The best accuracy rate 91.39% for binary dataset version was obtained with the random forest ensemble method.

As a second experiment, we made all comparisons after dataset preprocessing, in other terms, after feature subset selection. Hence, the most appropriate attributes were selected by using wrapper subset method to increase the accuracy rates.

One of the important steps to create a good model is attribute selection. This operation can be done in two ways: first, select relevant attributes, and second, remove redundant or irrelevant attributes. Attribute selection is made to create a

Algorithm	Five-level grading	Binary grading (P/F)
Decision tree (J48)	67.80%	91.37%
Random forest	**73.50%**	**93.07%**
Naive Bayes	68.26%	88.44%

(accuracy values, bold – best model).

Table 4.
Classification accuracy rates for the Portuguese lesson dataset.

Mathematics	Five-level grading	Binary grading (P/F)
Decision tree (J48)	**73.42%**	89.11%
Random forest	71.14%	**91.39%**
Naive Bayes	70.38%	86.33%

(accuracy values, bold – best model).

Table 5.
Classification accuracy rates for the mathematics lesson dataset.

simple model, to create a model that is easier to interpret, and to find out which features are more important for the results. Attribute selection can be done using filters and wrapper methods. In this study, we use the wrapper method, because it generally produces better results. This method has a recursive structure. The process starts with selecting a subset and induces the algorithm on that subset. Then evaluation is made according to the success of the model. There are two options in this assessment. The first option returns to the top to select a new subset, the second option uses the currently selected subset.

In **Table 6**, the accuracy rates were compared before and after the attribute selection process for the Portuguese dataset for five-level grade version. Thanks to the wrapper subset method, the accuracy rate of the J48 algorithm has increased from 67.80 to 74.88% with the selected attributes. This accuracy rate increased from 68.26 to 72.57% for naive Bayes algorithm. For the random forest method where we get the best accuracy results, the accuracy rate has increased from 73.50 to 77.20%.

In **Table 7**, the accuracy rates were compared before and after the attribute selection process for the mathematics dataset for five-level grading version. In this dataset, attribute selection significantly increased our accuracy. Here, unlike Portuguese language dataset, the best jump was obtained with J48 algorithm and search forward technique in wrapper method. In this way, the accuracy rate increased from 73.42 to 79.49%. A close result was obtained with the search backward technique and accuracy increased from 73.42 to 78.23%. Through this way, naive Bayes and random forest methods also increased significantly. This method increased the

Feature selection	**Wrapper subset (J48)**	**Wrapper subset (naive Bayes)**	**Wrapper subset (random forest)**
Before	67.80%	68.26%	73.50%
After	74.88%	72.57%	**77.20**%
Selected features	Search backward: age, famsize, Mjob, schoolsup, paid, internet, go out, health, G1, G2	Search backward: travel time, romantic, free time, health, G1, G2	School, Travel time, G2

The obtained classification accuracy rates for the Portuguese lesson dataset with five-level grading system. (accuracy values, bold – best model).

Table 6.
Before and after feature selection with five-level grading system

Feature selection	**Wrapper subset (J48)**	**Wrapper subset (J48)**	**Wrapper subset (naive Bayes)**	**Wrapper subset (random forest)**
Before	73.42%	73.42%	70.38%	71.14%
After	78.23%	**79.49**%	74.18%	78.99%
Selectedfeatures	Search backward: age, pstatus, Medu, Fedu, Fjob, failures, schoolsup, paid, activities, famrel, Dalc, Walc, G2	Search forward: sex, Mjob, Walc, G2	Famsize, Medu, Fjob, activities, higher, romantic, free time, G2	Famsize, Fedu, schoolsup, paid, activities, higher, romantic, Walc, absences, G1, G2

The obtained classification accuracy rates for the mathematics lesson dataset with five-level grading system. (accuracy values, bold – best model).

Table 7.
Before and after feature selection with binary grading system.

Feature selection	Wrapper subset (J48)	Wrapper subset (naive Bayes)	Wrapper subset (random forest)
Before	91.37%	88.44%	93.07%
After	91.99%	89.68%	**93.22**%
Selected Features	School, age, address, Medu, Fjob, travel time, study time, schoolsup, nursery, higher, famrel, free time, G1, G2	Sex, age, Pstatus, Fedu, Mjob, Fjob, reason, failures, famsup, paid, higher, Internet, romantic, go out, health, absences, G1, G2	School, sex, age, address, famsize, Pstatus, Medu, Mjob, Fjob, reason, guardian, travel time, study time, failures, schoolsup, famsup, paid, activities, higher, Internet, romantic, famrel, free time, go out, Dalc, Walc, health, absences, G1, G2

The obtained classification accuracy rates for the Portuguese lesson dataset with binary grading system. (accuracy values, bold – best model).

Table 8.
Before and after feature selection.

Feature selection	Wrapper subset (J48)	Wrapper subset (naive Bayes)	Wrapper subset (random forest)
Before	89.11%	86.33%	91.39%
After	90.89%	89.11%	**93.67**%
Selected features	School, age, address, Medu, Fedu, guardian, failures, schoolsup, famsup, Internet, romantic, famrel, free time, G1, G2	Sex, age, Pstatus, Fedu, Mjob, Fjob, reason, failures, famsup, paid, higher, Internet, romantic, go out, health, absences, G1, G2	Address, famsize, Fedu, Mjob, Fjob, reason, guardian, study time, schoolsup, higher, famrel, go out, absences, G2

The obtained classification accuracy rates for the mathematics lesson dataset with binary grading system. (accuracy values, bold – best model).

Table 9.
Before and after feature selection.

accuracy rate of naive Bayes method from 70.38 to 74.18%. Random forest result is increased from 71.14 to 78.99%. These results show that attribute selection with this wrapper subset method also works in this dataset.

In **Table 8**, the results of the wrapper attribute selection method before and after the application to the Portuguese binary version are compared. There was no significant increase in accuracy. The best results were obtained with random forest. The best jump was experienced by the naive Bayes method but did not reach the random forest value. Naive Bayes result has risen from 88.44 to 89.68%. Random forest maintained the high accuracy achieved before the attribute selection and increased from 93.07 to 93.22%.

After successful results in five-level grade versions, we tried the same attribute selection method in binary label version dataset. **Table 9** shows the accuracy values before and after the wrapper attribute selection for the mathematical binary dataset version. Because the accuracy of the binary version is already high, the jump is less than the five-level grades. But again, there is a nice increase in accuracy. The accuracy rate of the J48 algorithm was increased from 89.11 to 90.89%, while the naive Bayes result was increased from 86.33 to 89.11%. As with the mathematics five-level grade dataset, the best results were obtained with random forest in binary label dataset. Accuracy rate increased from 91.39 to 93.67%.

As a result, it can be possible to say that accuracy rates have changed positively in all trials using wrapper subset attribute selection method.

5. Conclusion and future work

This paper proposes the application of data mining techniques to predict the final grades of students based on their historical data. Three well-known classification techniques (decision tree, random forest, and naive Bayes) were compared in terms of accuracy rates. Wrapper feature subset selection method was used to improve the classification performance. Preprocessing operations on the dataset, categorizing the final grade field into five and two groups, increased the percentage of accurate estimates in the classification. The wrapper attribute selection method in all algorithms has led to a noticeable increase in accuracy rate. Overall, better accuracy rates were achieved with the binary class method for both mathematics and Portuguese dataset.

In the future, different feature selection methods can be used. In addition, different classification algorithms can also be utilized on the datasets.

Author details

Ferda Ünal
The Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, Turkey

*Address all correspondence to: ferda.balci@ceng.deu.edu.tr

References

[1] Fan Y, Liu Y, Chen H, Ma J. Data mining-based design and implementation of college physical education performance management and analysis system. International Journal of Emerging Technologies in Learning. 2019;**14**(06):87-97

[2] Guruler H, Istanbullu A. Modeling student performance in higher education using data mining. Studies in Computational Intelligence. 2014;**524**: 105-124

[3] Hu YH, Lo CL, Shih SP. Developing early warning systems to predict students' online learning performance. Computers in Human Behavior. 2014; **36**:469-478

[4] Costa EB, Fonseca B, Santana MA, de Araújo FF, Rego J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior. 2017;**73**: 247-256

[5] Shahiri AM, Husain W. A review on predicting student's performance using data mining techniques. Procedia Computer Science. 2015;**72**:414-422

[6] Fernandes E, Holanda M, Victorino M, Borges V, Carvalho R, Van Erven G. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research. 2019;**94**:335-343

[7] Marbouti F, Diefes-Dux HA, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers in Education. 2016;**103**:1-15

[8] Miguéis VL, Freitas A, Garcia PJ, Silva A. Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems. 2018;**115**:36-51

[9] Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. Computers in Education. 2017;**113**:177-194

[10] Rodrigues MW, Isotani S, Zárate LE. Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics. 2018;**35**(6): 1701-1717

[11] Buenano-Fernandez D, Villegas-CH W, Lujan-Mora S. The use of tools of data mining to decision making in engineering education—A systematic mapping study. Computer Applications in Engineering Education. 2019;**27**(3): 744-758

[12] Zhu S. Research on data mining of education technical ability training for physical education students based on Apriori algorithm. Cluster Computing. 2019;**22**(6):14811-14818

[13] Lu M. Predicting college students English performance using education data mining. Journal of Computational and Theoretical Nanoscience. 2017; **14**(1):225-229

[14] Marquez-Vera C, Cano A, Romero C, Noaman AYM, Mousa FH, Ventura S. Early dropout prediction using data mining: A case study with high school students. Expert Systems. 2016;**33**(1):107-124

[15] Amjad Abu S, Al-Emran M, Shaalan K. Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. Technology, Knowledge and Learning. 2019;**24**(4): 567-598

[16] Fujita H. Neural-fuzzy with representative sets for prediction of student performance. Applied Intelligence. 2019;**49**(1):172-187

[17] Agaoglu M. Predicting instructor performance using data mining techniques in higher education. IEEE Access. 2016;**4**:2379-2387

[18] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD explorations newsletter. 2009

[19] Cortez P, Silva A. Using data mining to predict secondary school student performance. In: Brito A, Teixeira J, editors. Proceedings of 5th Annual Future Business Technology Conference. tpPorto: EUROSIS-ETI; 2018. pp. 5-12