ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 3 (July - Sept. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

Improving Secondary School Student Performance using Data Mining Techniques

[I]Ahmed Ashraf, [II]Hazem M. El-Bakry, [III]Yehia El-Mashad, [IV]Samir Abd-Elrazik, [V]Mohammed El-Desouky

[I,II,IV]Information Systems Dept., Faculty of Computer and Information Sciences, Mansoura University, Mansoura, EGYPT

[V]Computer science Dept., Faculty of Computer and Information Sciences, Mansoura University, Mansoura, EGYPT

[III]Delta University, EGYPT

Abstract

This paper aims to improve students' performance in secondary e-learning systems by employing data mining (DM) models. Data is collected using online school tests, reports and quizzes. This paper applies SVM with accuracy 89%, Decision tree with accuracy 89%, M5-Rules with Root mean squared error equal to 1.4621 and Linear Regression with Root mean squared error equal to 2.0017, 3.0089 and 3.6057.The results show that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption). As a direct outcome of this research, more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.

Keywords

Data Mining in Education, Classification and Regression, Decision Trees

I. Introduction

We have to learn in anywhere by any way. E-Learning is the use of technology to enable learners to learn anytime and anywhere. E-learning has created new markets for education is far from the traditional trend of education depending on the tremendous revolution of information technology. E-learning is an education via the Internet, network, or standalone computer (Dr.P.Nagarajan. Dr.G.Wiselin Jiji, 2010) (Devajit Mahanta, Majidul Ahmed, 2012). Usual e-learning systems offer knowledge and evaluation for learners but our system aiming to ensure that students received educational content correctly by monitoring his handling of the course content and communicate directly with him to direct it to the correct way. On the other hand, the interest Data Mining (DM) (Turban et al. 2007), arose due to the advances of Information System, leading to an exponential growth of educational and organizational systems. All of this data holds actual information, Such as student grades, demographic, social and school related features that can be used to support Making decisions and achieving perfect goals. Because of the limited human ability to predict, we need an alternative tool to analyze a big data and extract useful information for the decision-maker. Education is the best area for applying data mining application applications, since there are multiple sources of data such as traditional databases, online web pages and process elements like students, teachers and administrators (Ma et al. 2000). There are many questions concerning the field of e-learning that we can answer with techniques using Data Mining techniques (Luan 2002, Menai-Pedgoli et al., 2003): Who are students who have a credit hours system? Who are the students who need to study the course again? How to increase the number of our students? How to handle system's errors? How to prevent student converting to another eLearning system? What are the methods of predicting student performance? How to improve the student's performance and achievement? The points of this paper are prediction and improving, predict the student's performance and improve his achievement. Educational Data Mining (EDM) is to present useful information that can be used by educational software Developers, students, teachers, parents, and other educational researchers. Data Mining was born to extract knowledge from e-learning systems through the information available in databases generated by their users. Data Mining was born to extract knowledge from e-learning systems through the information available in databases generated by their users [21-51].

II. Materials And Methods

It is known that the secondary education consists of 3 years. Most of the students belong to the public and free education system. There are several courses (e.g. Sciences, Historical and Geometric) that share core subjects such as the mother tongue of the state and Mathematics. Often a 20-point grading scale is used, Starting from o and ending with 20. During the school year, students are evaluated in three periods (G1, G2, and G3) and the last evaluation (G3) as shown in Table 1. Our data collected during the 2015-2016 school year from public school's eLearning website. There has been a trend for an increase of Information Tech in learning we developed website to make e learning system instead of sheets system. The paper system has many disadvantages such as lack of data and lack of credibility.The database was built from three dimensions data registration on our website questionnaires and test results for each year. We designed our site to get special data divided into four sections grades (i.e. Test results), related feature (e.g. School name, number of past class failures and absence), social (i.e. quality of family relationships, alcohol consumption) (Pritchard and Wilson 2003) and demographic (e.g. mother's education, father's education, mother's and father's job) which affect student performance. During the preprocessing step some variables had to be ignored due to the lack of discriminative value.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Table 1: The preprocessed student related variables

Attribute	Description (Domain)
student's sex	(binary: female or male)
student's age	(numeric: from 15 to 22)
student's school	(binary: Gabriel Pereira or Mousinho da Silveira)
student's student's	address type (binary: urban or rural)
parent's cohabitation status	(binary: living together or apart)
mother's education	(numeric: from 0 to 4a)
mother's job	(nominal)
father's education	(numeric: from 0 to 4a)
Father's job	(nominal)
student's guardian	(nominal: mother, father or other)
family size	(binary: _ 3 or > 3)
Family relationships	quality of relationships (numeric: from 1 – very bad to 5 – excellent)
reason to choose this school	(nominal: close to home, school reputation, course preference or other)
traveltime	(numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
weekly study time	(numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
number of past class failures	(numeric: n if 1 _ n < 3, else 4)
extra educational school support	(binary: yes or no)
family educational support	(binary: yes or no)
extra-curricular activities	(binary: yes or no)
extra paid classes	(binary: yes or no)
Internet access at home	(binary: yes or no)
attended nursery school	(binary: yes or no)
wants to take higher education	(binary: yes or no)
with a romantic relationship	(binary: yes or no)
free time after school	(numeric: from 1 – very low to 5 – very high)
going out with friends	(numeric: from 1 – very low to 5 – very high)
weekend alcohol consumption	(numeric: from 1 – very low to 5 – very high)
workday alcohol consumption	(numeric: from 1 – very low to 5 – very high)
current health status	(numeric: from 1 – very bad to 5 – very good)
number of school absences	(numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

III. Data Mining Models

Classification and Regression Tree Analysis, CART, is a simple yet powerful analytic tool that helps determine the most "important" (based on explanatory power) variables in a particular dataset. The main difference is set in terms of the output representation, (i.e. discrete for classification and continuous for regression). In classification, models are often evaluated using the Percentage of Correct Classifications (PCC), while in regression the Root Mean Squared (RMSE) is a popular metric (Witten and Frank 2005). A high PCC (i.e. near 100%) suggests a good classifier, while regression should present a low global error (i.e. RMSE close to zero). These metrics can be computed using the equations:

$$\Phi(i) = \begin{cases} 1 & , \quad if\ y_i = \hat{y}_i \\ 0 & , \quad else \end{cases} \qquad (1)$$

$$\mathrm{PCC} = \sum_{i=1}^{N} \frac{\Phi(i)}{N} \times 100(\%) \qquad (2)$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} (y_i + \hat{y}_i)^2 / N} \qquad (3)$$

Where yi denotes the predicted value for the i- th example.

The dataset of student grades (i.e. G3 of Table 1) records is tested and applied on various classification algorithms using WEKA an Open source tool such as:

- Linear Regression (finding the best-fitting straight line through the point)
- SVM Support Vector Machines
- Decision tree algorithm J48

- M5-Rules Algorithm

There are many data mining algorithms have been proposed for classification and regression tasks, each one have its own goal

A. Decision tree algorithm J48

The most important feature in Decision tree (DT) is the use of a tree structure as simple representation for a set of rules that distinguish values hierarchically ((Breiman et al.1984).

It has proved to be useful tools for the classification, description, and generalization of data (Ihsan A. Kareem 1, Mehdi G. Duaimi, 2014). Decision Tree Algorithm is to discover the behavior of the attributes (Gaganjot Kaur, Amit Chhabra, 2014). Tree classification algorithm is used to make the prediction of the target variable and understood the critical distribution of the data is easily (Nadali, A; Kakhky, E.N.; Nosratabadi, 2011). The basic algorithm for learning decision trees is: starting with default data, select attribute or value along dimension that gives "best" split ,create child nodes based on split and each child using child data until a stopping criterion is reached: all examples have same class or the amount of data is too small or the tree is too large. The C4.5 algorithm for building decision trees is implemented in Weka as a classifier called J48. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = s1, s2... of already classified samples. Each sample si = x1, x2... is a vector where x1, x2... represent attributes or features of the sample. The training data is augmented with a vector C = c1, c2... where c1, c2... represent the class to which each sample belongs.

1. confusion matrix

A confusion matrix (an error matrix) is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix) (Stehman, Stephen V.1997). This matrix used in the field of machine learning and specifically the problem of statistical classification. In the table layout rows and columns have a role, each row represents the instances in an actual class and each column of the matrix represents the instances in a predicted class (Powers, David M .2011).

Condition positive (P)
The number of real positive cases in the data
Condition negatives (N)
The number of real negative cases in the data
True positive (TP)
Eqv. With hit
True negative (TN)
Eqv. With correct rejection
False positive (FP)
Eqv. With false alarm, Type I error
False negative (FN)
Eqv. With miss, Type II error

Sensitivity, recall, hit rate, or true positive rate (TPR)

4. $$\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (4)$$

Specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (5)$$

Precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (6)$$

Negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \quad (7)$$

Miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN+TP}} = 1\text{- TPR} \quad (8)$$

Fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} = 1 - \mathrm{TNR} \quad (9)$$

False discovery rate (FDR)

$$\mathrm{DR} = \frac{\mathrm{FP}}{\mathrm{FP+TP}} = 1\text{- PPV} \quad (10)$$

False omission rate (FOR)

$$\mathrm{FOR} = \frac{FN}{FN+TN} = 1\text{- NPV} \quad (11)$$

Accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{P+N}} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (12)$$

2. pruning

One of the most important techniques in machine learning is a pruning which used to remove a section of the tree to reduce the size of decision trees. A tree with a lot of branching is a big problem so we need to reach to the optimal size by removing the nodes that do not provide important information (Trevor Hastie, Robert Tibshirani, and Jerome Friedman .2001)

B. Linear Regression

Regression analysis considers one of the best statistical analysis tools (K. Hrona, P. Filzmoserb and K. Thompsonc 2009). To explain the relationships between a dependent variable (either categorical or continuous) and a set of independent variables we have to use regression analysis, taking into consideration that variable based on a sample from a given community (Cody S. Ding, 2006). The general regression model found in any basic statistics test can be written as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . \beta_n x_n + \varepsilon_i \quad (13)$$

Where β_0 is intercept; β_k is the regression slope or coefficient for a given independent variable k, and ε_i is error term for individuals i. Equation 1 has one key feature. It assumes that all individuals are drawn from a single population with common population parameters.

C. SMO

SMO solves the SVM QP problem by decomposing it into QP sub-problems and solving the smallest possible optimization problem, involving two Lagrange multipliers, at each step (John C. Platt. 1998). QP problem (Maximize/Minimize) is a Quadratic Objective Function subject to a Set of Linear Constraints.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

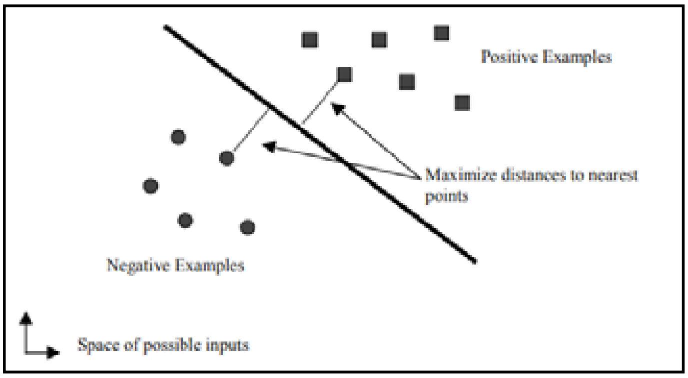

Fig. 1 : A linear Support Vector Machine

$u = \vec{w} \cdot \vec{x} - b$, As shown in the figure 1 in the linear case, the margin is defined by the distance of the hyper plane to the nearest of the positive and negative examples. The formula for the output of a linear SVM is

Where w is the normal vector to the hyper plane and x is the input vector. The separating hyper plane is the plane u=0. The nearest points lie on the planes u = ±1. The margin m is thus

$$M = \frac{1}{||w||_2} \tag{14}$$

D. M5-Rules Algorithm

M5 algorithm is one of the common methods for generating rules from the trees. The M5 builds regression trees whose leaves are consist of multivariate linear models, and the nodes of the tree are chosen over the attribute that maximizes the expected error reduction as a function of the standard deviation of output parameter (Dolado et al., 2007). The association rule, and the classification rule are the only two rule algorithm type (Angelina Njegus1, Vanja Nikolie2, Verka Jovanovi , 2015). The important of of using Rule algorithms in data mining application : It offer simple and clear results, active in undirected data mining ,deal with huge amount of data ,using a simple computation . Create M5 tree in three steps: generates a regression tree using the training data, and calculates a linear model for each node of the tree generated, tries to simplify the regression tree deleting the nodes of the linear models whose attributes do not increase the error and reduces the size of the tree without reducing the accuracy.

IV. Experimental Work And Results

We have performed classification using J48 decision tree algorithms (Sequential Minimal Optimization), Linear Regression and M5-Rules Algorithm m on mathematical dataset in weka tool.

1. Results for classification using J48

J48 is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. PASS or FAIL.

Confusion Matrix

a	b	<-- classified as
102	28 |	a = Fail
15	250 |	b = Pass

For above confusion matrix, true positives for class a='Fail' is 102 while false positives is 28 whereas, for class b=Pass, true positives is 250 and false positives is 15 i.e. diagonal elements of matrix 102+250 =352 represents the correct instances classified and other elements 15+28 = 43 represents the incorrect instances.

Hence,
TP rate for class a = 0.815
FP rate for class a = 0.072
TP rate for class b = 0.928
FP rate for class b = 0.185
Average TP rate = 0.891
Average FP rate = 0.147
Precision = diagonal element/sum of relevant column
Precision for class a = 102/ (102+15) = 0.568
Precision for class b = 250/ (250+28) = 0.899
F-measures = 2*precision*recall/ (precision + recall)
F-measure for class a = 2*.872*.785/ (.872+.785) =.826
F-measure for class b= 2*.899*.943/ (.899+.943) = .921
Accuracy (ACC):

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

$$ACC = \frac{102+250}{102+250+15+28} = 0.891$$

Decision tree:

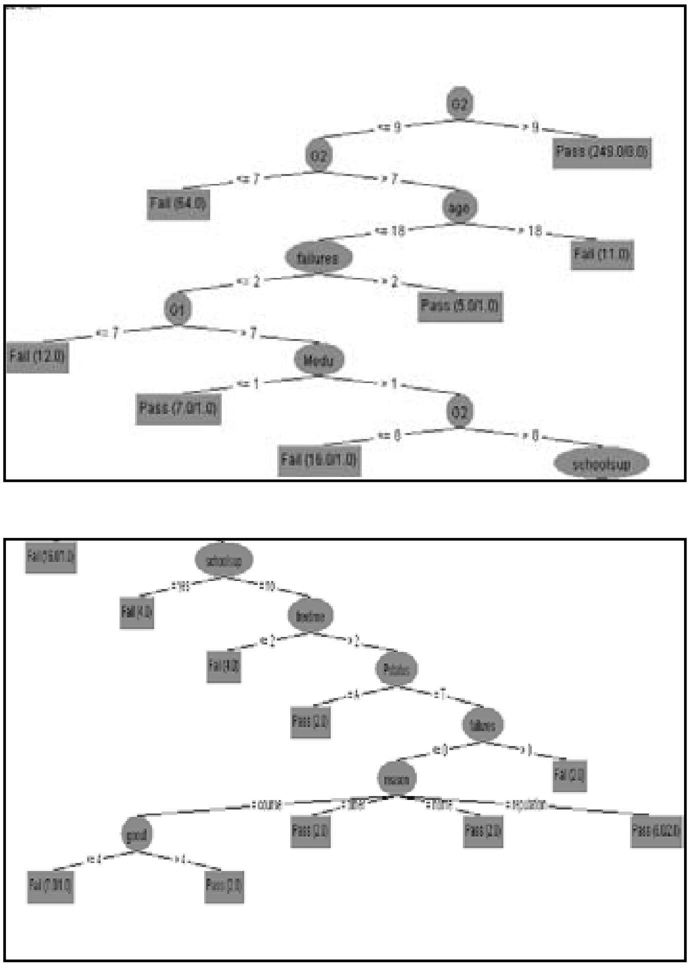

Fig. 2 : Decision tree for student-mat.csv

There are some interesting rules that can be extracted from these trees, for instance:
1. if G1< 10 ^ G2=9 ^ Mjob 2 {teacher, other} then pass;
2. if G1< 10 ^ G2=9 ^ Mjob 2 {home, health, civil services} then fail;
3. if 12 _G2< 14 ^ goout> 1 then III;

4. if 12 _G2< 14 ^ goout= 1 then II;
5. if 11 <G1_ 13 ^ absences< 7 then 13;
6. if 11 <G1_ 13 ^ absences_ 7 then 11;

2. Results for classification using SVM

SVM is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. PASS or FAIL.

Confusion Matrix
a b <-- classified as
106 24 | a = Fail
19 246 | b = Pass

For above confusion matrix, true positives for class a='Fail' is 106 while false positives is 24 whereas, for class b=Pass, true positives is 246 and false positives is 19 i.e. diagonal elements of matrix 106+246 =352 represents the correct instances classified and other elements 24+19 = 33 represents the incorrect instances.

Hence,
TP rate for class a =0.815
FP rate for class a = 0.072
TP rate for class b =0.928
FP rate for class b = 0.185
Average TP rate = 0.891
Average FP rate = 0.163
Precision = diagonal element/sum of relevant column
Precision for class a = 0.872
Precision for class b = 0.899
F-measures = 2*precision*recall/ (precision + recall)
F-measure for class a =2*.848*.815/ (.848+.815) = .831
F-measure for class b = 2*.911*.928/ (.911+928) = .920
Accuracy (ACC):)

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$ACC = \frac{106+246}{106+246+19+24} = 0.891$$

3. Results for classification using Linear Regression

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

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

To access the predictive performances, 20 runs of a 10- fold cross-validation (Hastie et al. 2001) (in a total of 200 simulations) were applied to each configuration. Under such scheme, for a given run the data is randomly divided in 10 subsets of equal size. Sequentially, one different subset is tested (with 10% of the data) and the remaining data used to fit the DM technique. At the end of this process, the evaluated test set contains the whole dataset, although 10 variations of the same DM model are used to create the predictions.

Regression results

A. Linear Regression Model

G3 =
-0.53 school +0.2568 * age - 0.4192 * Fjob + 0.5391 * Fjob - 0.2845 * activities + 0.3167 * romantic + 0.4022 * famrel + 0.1355 * Walc + 0.0474 * absences + 0.1687 * G1 + 0.9718 * G2 + 0.7893

=== Cross-validation ===

Correlation coefficient	0.8996
Mean absolute error	1.3059
Root mean squared error	2.0017
Relative absolute error	37.952 %
Root relative squared error	43.5862 %
Total Number of Instances	395

B. Linear Regression Model

G1 =
0.8726 * sex + 0.2222 * Fedu + 0.5466 * Mjob + 0.8306 * Mjob - 1.2776 * Mjobth + 1.8858 * Mjob + 1.0861 * Fjob + 0.9411 * Fjob + 0.6673 * studytime - 1.2542 * failures + 2.0633 * schoolsup +0.9542 * famsup + 1.2838 * higher +0.2502 * freetime - 0.4451 * goout - 0.216 * health + 6.1228

=== Cross-validation ===

Correlation coefficient	0.4434
Mean absolute error	2.4809
Root mean squared error	3.0089
Relative absolute error	89.8577 %
Root relative squared error	90.5317 %
Total Number of Instances	395

C. Linear Regression Model

G2 =
0.8958 * sex - 0.2264 * age + 0.687 * famsize + 0.3074 * Medu + 0.9718 * Mjob - 1.5888 * Mjob + 2.2183 * Mjob +1.4052 * Fjob - 0.9341 * guardian - 0.4422 * traveltime +0.5564 * studytime - 1.4591 * failures + 1.4613 * schoolsup +0.8055 * famsup +0.6754 * internet +0.7634 * romantic - 0.4796 * goout - 0.2518 * health + 13.2063

=== Cross-validation ===

Correlation coefficient	0.3476
Mean absolute error	2.7981
Root mean squared error	3.6057
Relative absolute error	94.7091 %
Root relative squared error	95.4984 %
Total Number of Instances	395

4. Results for classification using M5Rules

M5 pruned model rules
(Using smoothed linear models) :
Number of Rules: 5

Rule: 1

IF
G2 > 10.5
THEN

G3 =
0.0439 * age + 0.201 * Mjob - 0.1484 * traveltime - 0.0236 * activities + 0.022 * romantic + 0.0256 * famrel - 0.1054 * Walc + 0.003 * absences + 0.0124 * G1 + 1.041 * G2 - 1.0993 [203/15.592%]

Rule: 2
IF
absences > 1
G2 > 7.5
THEN

G3 =
- 0.038 * age - 0.2381 * famsize + 0.046 * Pstatus - 0.0259 * Medu - 0.1884 * Fedu - 0.0369 * Mjob + 0.039 * Mjob - 0.0618 * Fjob + 0.0326 * reason + 0.0461 * schoolsup + 0.0745 * romantic + 0.0912 * famrel + 0.0314 * Walc + 0.0074 * absences + 0.199 * G1 + 0.8447 * G2 + 0.389 [97/20.84%]

Rule: 3
IF
absences <= 1
G2 > 6.5
THEN

G3 =
-1.4169 * age - 1.3755 * reason - 0.3079 * activities + 0.0343 * absences + 1.1151 * G2 + 20.2863 [38/87.272%]

Rule: 4
IF
absences > 1
THEN

G3 =
0.3999 * age - 0.2253 * Medu + 0.6979 * Mjob - 0.2776 * failures + 0.0455 * absences + 0.7721 * G2 - 5.2655 [34/21.977%]

Rule: 5

G3 =
+ 0 [23]
=== Cross-validation ===

Correlation coefficient	0.9476
Mean absolute error	0.9001
Root mean squared error	1.4621
Relative absolute error	26.1591 %
Root relative squared error	31.8378 %
Total Number of Instances	395

V. Conclusion

In this paper, we have addressed the prediction Mathematics student grades in secondary depending on previous student grades in first or/and second year and other attributes. Four data mining methods were tested (SVM, Decision tree M5-Rules and Linear Regression). The output results prove that it is possible to achieve a high predictive accuracy in Decision tree and SVM methods and low Root mean squared error in M5-Rules and Linear Regression methods ,provided that the first and/or second school period grades are known.

References

[1]. *Turban E. ; Sharda R.; Aronson J.; and King D., 2007. Business Intelligence, A Managerial Approach. Prentice-Hall.*
[2]. *Dr.P.Nagarajan. Dr.G.Wiselin Jiji, ONLINE EDUCATIONAL SYSTEM (e- learning) ,International Journal of u- and e-Service, Science and Technology Vol. 3, No. 4, December, 2010.*
[3]. *Devajit Mahanta, Majidul Ahmed, E-Learning Objectives, Methodologies, Tools and its Limitation, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-1, December 2012*
[4]. *Ma Y.; Liu B.; Wong C.; Yu P.; and Lee S., 2000. Targeting the right students using data mining. In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.*
[5]. *Luan J., 2002. Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 113, 17–36.*
[6]. *Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., 2003. Predicting student performance: an application of data mining methods with an educa- tional web-based system. In Proc. of IEEE Frontiers in Education. Colorado, USA, 13–18.*
[7]. *Pritchard M. and Wilson S, 2003. Using Emotional and Social Factors To Predict Student Success. Journal of College Student Development, 44, no. 1, 18–28.*
[8]. *K. Hrona, P. Filzmoserb and K. Thompsonc. 2009, Linear regression with compositional explanatory variables, Journal of Applied Statistics Vol. 00, No. 00, January*
[9]. *Cody S. Ding, 2006, Using Regression Mixture Analysis in Educational Research, Volume 11 Number 11, December 2006*
[10]. *Breiman L.; Friedman J.; Ohlsen R.; and Stone C., 1984. Classification and Regression Trees. Wadsworth, Monterey, CA.*
[11]. *Ihsan A. Kareem 1, Mehdi G. Duaimi, 2014, Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization, IJCSMC, Vol. 3, and Issue. 6, June 2014, pg.176 – 183*
[12]. *Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.161,165, 8- 10 April 2011*
[13]. *Gaganjot Kaur, Amit Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014*
[14]. *Powers, David M .2011, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Technical Report SIE-07-001*
[15]. *Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. Volume 62, Issue 1, October*

1997

[16]. Trevor Hastie, Robert Tibshirani, and Jerome Friedman .2001, the Elements of Statistical Learning. Springer, pp. 269-272

[17]. John C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14 April 21, 1998

[18]. Dolado, J.J., Rodriguez, D., Riquelme, J., Ferrer-Troyano,F., & Cuadrado, J.J. (2007). A Two Stage Zone Regression Method for Global Characterization of a Project Database. Available at: http://www.cc.uah.es/drg/b/DolRodetAlIdea07.pdf (Last accessed 16-2-2017).

[19]. Angelina Njegus1 , Vanja Nikolie2 , Verka Jovanovi , 2015 , Angelina Njegus1 , Vanja Nikolie2 , Verka Jovanovi , 2015 THE SELECTION OF OPTIMAL DATA MINING METHOD FOR SMALL-SIZED HOTELS, International Scientific Conference of IT and Business-Related Research

[20]. Hastie T.; Tibshirani R.; and Friedman J., 2001. The Elements of Statistical Learning: Data Mining, Infer- ence, and Prediction. Springer-Verlag, NY, USA.

[21]. H. El-Bakry and M. Hamada, Enhancing Adaptive Mobile E-Learning by using Cloud Computing, In "Active and Collaborative Learning: Practices, Problems and Prospects", M. Hamada, (Ed.), pp. 139-165, NOVA Science Publisher, ISBN: 978-1629480664, Dec. 2013.

[22]. A. M. Riad, Hazem M. El-Bakry, and Samir M. Abd El-razek, "Virtual Investigation of Patients for Medical E-Learning," International Journal of Computer Science and Information Security, vol. 10, no. 7, August, 2012, pp. 6-10.

[23]. Samir M. Abd El-Razek, Waeil F. Abd El-Wahed and, Hazem M. El-Bakry, "MUVES: A Virtual Environment System for Medical Case Based Learning," International Journal of Computer Science and Network Security, vol. 10, no. 9, September 2010, pp. 159-163.

[24]. Mohammed Yaqub, A. M. Raid, and Hazem M. El Bakry, "Online Social Learning Suggestion System Based On Student Behavior," International Journal of Information Science and Intelligent System, vol. 4, No. 1, Jan. 2015, pp. 43-64.

[25]. Ahmad Ashraf, Hazem M. El Bakry, Samir M. Abd El-razek, and Yehia El-Mashad "Handling big data in e-learning," International Journal of Advanced Research in Computer Science & Technology, vol. 3, issue 1, Jan.-March 2015, pp. 47-51.

[26]. M. Yaqub, A. M. Raid and H. El-Bakry, "IMPROVING E-LEARNING PERFORMANCE THROUGH SOCIAL COMMUNICATIONS," International Journal of Intelligent Computing and Information Science, Vol. 14, No.4, October 2014, pp. 73-85.

[27]. Mohammed Yaqub, A. M. Raid, H. M. El-Bakry, Haitham A. EL-Ghareeb "Online Social Learning Suggestion System Based On Student Behavior," International Journal of Advance Foundation and Research in Computer, vol. 1, issue 10, October 2014, pp. 7-18.

[28]. Mohammed Khaleel, Hazem M. El Bakry, and Ahmed A. Saleh, "Developing E-learning Services Based on Cache Strategy and Cloud Computing," International Journal of Information Science and Intelligent System, vol. 3, No. 4, October 2014, pp. 45-52.

[29]. Mohammed Yaqub, A. M. Raid, and H. M. El-Bakry, "Open Social Learning Network," International Journal of Computer Engineering and Intelligent Systems, vol. 5, No. 7, July 2014, pp. 1-12.

[30]. Ahmed A. Saleh and Hazem M. El-Bakry, "Adaptive E-Learning Based on Learner's Styles," Bulletin of Electrical Engineering and Informatics, vol. 2, No. 4, December 2013, pp. 240-251.

[31]. Ahmed A. Saleh and Hazem M. El-Bakry, " A Novel Adaptive Mobile E-Learning Model Based on Developed Learner's Styles," Bulletin of Electrical Engineering and Informatics, vol. 2, No. 2, June 2013, pp. 141-157.

[32]. Ahmed A. Saleh and Hazem M. El-Bakry, "A Novel Adaptive Mobile E-Learning Model," International Journal of Computer Applications, vol. 63, No. 14, Feburary 2013, pp. 12-25.

[33]. N. S. Abu El-Ala, W. A. Awad and H. M. El-Bakry, " Cloud Computing for Solving E-Learning Problems," International Journal of Advanced Computer Science and Applications, vol. 3, No. 12, December 2012, pp. 135-137.

[34]. Hazem El-Bakry, and Mohamed Hamada, "Adaptive E-Learning for Data Encoding and Computer Networks based on Learner's Styles," International Journal of Computer Networks and Security, vol. 22, issue 12, December 2012, pp. 333-342.

[35]. A. M. Riad, Hazem M. El-Bakry, and Samir M. Abd El-razek, " Design of Virtual Environment for Medical E-Learning," Global Advanced Research Journal of Engineering, Technology and Innovation, vol. 1, no. 6, September 2012, pp. 138-147.

[36]. A. M. Riad, Hazem M. El-Bakry, and Samir M. Abd El-razek, "Simulation of Medical Consultation on Virtual Patient," International Journal of Computational Linguistics and Natural Language Processing, vol. 1, issue 3, October 2012, pp. 60-64.

[37]. Ahmed A. Saleh, Hazem M. El-Bakry, and Taghreed T. Asfour, "Design of Adaptive E-Learning for Logic Operations," International Journal of Education and Information Technologies, issue 2, vol. 4, 2010, pp. 49-56.

[38]. Hazem M. El-Bakry, and Nikos Mastorakis, "Realization of E-University for Distance Learning," WSEAS Transactions on Computers, issue 1, vol. 8, Jan. 2009, pp. 48-62.

[39]. Mohammed Yaqub, A. M. Raid, H. M. El-Bakry, and Nikos Mastorakis, "Improving E-Learning Performance Through Social Communications," Proc. of the 11th International Conference on Educational Technologies (EDUTE '15) Dubai, United Arab Emirates, February 22-24, 2015, pp. 129-137.

[40]. A. M. Riad, Hazem M. El-Bakry, and Samir M. Abd El-razek, " Design and Evaluation of Medical E-Learning Model," the 3rd WSEAS International Conference on Bioscience and Bioinformatics (ICBB '12), Montreux, Switzerland, December 29-31, 2012, pp. 265-272.

[41]. Ahmad Ashraf, Hazem M. El Bakry, Samir M. Abd El-razek, Yehia El-Mashad, and Nikos Mastorakis, "Enhancing Big Data Processing in Educational Systems," Proc. of the 11th International Conference on Educational Technologies (EDUTE '15) Dubai, United Arab Emirates, February 22-24, 2015, pp. 176-181.

[42]. Hazem M. El-Bakry, Ahmed A. Saleh, Taghreed T. Asfourand, and Nikos Mastorakis, "A New Adaptive E-Learning Model Based on Learner's Styles," Proc. of 13th WSEAS Int. Conf. on MATHEMATICAL and COMPUTATIONAL METHODS

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

in SCIENCE and ENGINEERING (MACMESE '11), Catania, Sicily, Italy, November 3-5, 2011, pp. 440-448.

[43]. *Samir M. Abd El-Razek, Hazem M. El-Bakry, and Wael F. Abd El-Wahed, "Design of Virtual Environment for Medical E-Learning," Proc. of ICI 10 – 10th International Conference on Information at Delta University for Science and Technology, Gamassa, EGYPT, December 4-6, 2010, pp.*

[44]. *Ahmed A. Saleh, Hazem M. El-Bakry, Taghreed T. Asfour and Nikos Mastorakis "Adaptive E-Learning Framework for Digital Design," Proc. of 9th WSEAS International Conference on Telecommunications and Informatics, Italy, May 29-31, 2010, pp. 176-182.*

[45]. *Ahmed A. Saleh, Hazem M. El-Bakry, Taghreed T. Asfour and Nikos Mastorakis "Adaptive E-Learning Tools for Numbering Systems," Proc. of 9th WSEAS International Conference on Applications of Computer Engineering (ACE'10), Penang, Malaysia, March 23-25, 2010, pp. 293-298.*

[46]. *Hazem M. El-Bakry, and Nikos Mastorakis "Activation of Informal Learning with E-learning Technology " Proc. of EDUCATION and EDUCATIONAL TECHNOLOGY CONFERENCE 2009 (EDU'09), University of Genova, Genova, Italy, October 17-19, 2009, pp. 245 -247.*

[47]. *Ahmed A. Saleh, Hazem M. El-Bakry, Mahmud Atta Alla and Nikos Mastorakis "ON THE WAY TO MODELING: A CASE STUDY IN MANAGING MODEL TRANSFORMATIONS IN MDA," Proc. of 9th WSEAS International Conference on DISTANCE LEARNING and WEB ENGINEERING (DIWEB '09), Budapest, Hungry, September 3-5, 2009, pp. 89-103.*

[48]. *Hazem M. El-Bakry, and Nikos Mastorakis "Modular Networks for Active E-learning," Proc. of 9th WSEAS International Conference on Applied Informatics and Communications (AIC'09), Moscow, Russia, August 26-28, 2009, pp. 373-382.*

[49]. *Hazem M. El-Bakry, and Nikos Mastorakis "E-Learning and Management Information Systems For E-Universities," Proc. of 13th WSEAS International Conference on Computers, Rodos, Greece, July 22-25, 2009, pp. 566-574.*

[50]. *Hazem M. El-Bakry, and Nikos Mastorakis "Advanced Technology for E-Learning Development," Proc. of Recent Advances in Applied Mathematics and Computational and Information Sciences, Houston, USA, April 30-May 2, 2009, pp. 501-522.*

[51]. *Hazem M. El-Bakry, and Nikos Mastorakis, " Realization of E-University for Distance Learning, " 8th WSEAS International Conference on DISTANCE LEARNING and WEB ENGINEERING (DIWEB '08), Santander, Cantabria, Spain, September 23-25, 2008, pp. 17-31.*