Proceedings of the

2016 Sixth International Conference on Advanced Computing & Communication Technologies

3 – 4 September 2016
India

Edited by

Dr. Harish Mittal

Organizing Committee

General and Financial Chair

J. P. Mittal, *President, RGES, INDIA Former-Principal, SPGOI Rohtak, Vaish College Rohtak*

Program Chair

Harish Mittal, *Director, Sat Priya Group of Institutions, Rohtak*

Organizing Secretary

Sandeep Tayal, *Associate Prof, VCE, Rohtak*

Ankit Bansal, *Associate Prof, VCE, Rohtak*

Mukesh Singla, *Associate Prof, MSIET,Kalanaur*

Anish Mittal, *Lead IT Network Analyst, Concentrix Daksh India Pvt Ltd*

Coordinators

Amit Bansal, *Scientist, NIC, Jhajjar*

Munish Gupta, *EM, RGES, INDIA*

Technical Programme Committee

Sartaj Sahni, *University of Florida - Computer & Information Science & Engineering Department, USA*

Alfredo Vaccaro, *University of Sannio - Department of Enginnering, Piazza Roma , Benevento, Italy*

Krzysztof Galkowski, *ICCE, University of Zielona Góra, Poland*

Izzat M Alsmadi, *Yarmouk University, Jordan*

Sampson Asare, *University of Botswana, Gaborone, Botswana*

S. P. Khatkar, *UTD, M D University, Rohtak*

Pradeep Kumar Bhatia, *Guru Jambheshwar University of Science and Technology, Hisar*

Sunil Kumar Khatri, *AIIT, Amity University, Noida, India*

V. V. R. Raman, *ACHS, Asmara, Eritrea*

Olumuyiwa Akinrole Fadugba, *Covenant University, Nigeria*

Sefiu Taiwu Oloruntoyin, *Emmanuel Alayande College of Education, Nigeria*

Dan Randall, *Adjunct Prof., American Sentinel University, USA*

A. Clementking, *King Khalid University, Saudi Arabia*

Veena T. Nandi, *Majan College, University College, Ruwi, Muscat, Oman*

Nirbhay Chaubey, *ISTAR, Gujarat Technological University (GTU), Gujrat*

Poonam Bansal, *MSIT,New Delhi*

A. V. Senthil Kumar, *Hindusthan College of Arts and Science, Coimbatore*

S. Hariharan, *Pavendar Bharathidasan College of Engg. & Tech. Tiruchirapalli*

Pankaj Gupta, *Vaish College of Engineering, Rohtak*

Bhavesh R. Javani, *MEFGOI, Rajkot, Gujrat*

Shyam Akashe, *ITM,Gwalior*

N. S.Murthi Sarma, *Sreenidhi Institute of Sc & Tech, Yamnampet, Hyderabad*
Vijay Nehra , *Bhagat Phool Singh Mahila Vishwavidyalaya, Khanpur Kalan, Sonepat, India*
Rajeshwar Singh, *Doaba Group of Colleges, Nawanshar*
O. P. Sangwan, *GJUSÁT, Hisar*
Mukesh Kumar, *TITS Bhiwani, Haryana*
Anand Nayyar, *KCLIMT, Jalandhar, Punjab*
Deepak Goyal, *Vaish College of Engineering, Rohtak*
Deepak Gupta, *Director, Gurin Technologies Pvt Ltd, INDIA*
Sunil Maggu, *Vaish College of Engineering, Rohtak*
Satish Sood, *Dronacharya College of Education, Kangra, H.P.*
Umesh Gupta, *Vaish College of Engineering, Rohtak*
Amit Kant pandit, *Shri Mata Vaishno Devi University Katra JK*
Geeta R. B, *GMRIT, Rajam, Andhra Pradesh*
Piyush Javeria, *Techno India NJR Institute of Technology, Udaipur*
Dr. P. Sanjeevikumar, *VIT University, Chennai*

Advisory Committee

Virivada Venkata Raghava Raman, *Asmara College of Health Sciences, Eritrea, Africa*
Sartaj Sahni, *University of Florida - Computer & Information Science & Engineering Department, USA*
Alfredo Vaccaro, *University of Sannio - Department of Engineering, Piazza Roma , Benevento, Italy*
Krzysztof Galkowski, *ICCE, University of Zielona Góra, Poland*
Belhocine Ali, *University of Sciences and Technology of Oran, Algeria*
Dharminder Kumar, *Dean FET, Guru Jambeshwar University of Sc. & Tech., Hisar*
Nirmal Dayanand Raju, *Majan College, Ruwi, Muscat, Oman*
Samadhiya Durgesh, *Chung Hua University, Hsinchu, Taiwan*
R. S. Chhillar, *Deptt. of Computer Applications, M.D. University, Rohtak*
Pradeep Bhatia, *Deptt. of CSE, GJUS&T, Hisar*
Rahul Rishi, *Director UIET, MDU, Rohtak*
A. K. Garg, *Chairman, Deptt of ECE, DCRUST, Murthal*
Vinay Goyal, *Director, HIT , Asaudha*
Yudhvir Singh, *Deptt of CSE, GJUS&T, Hisar*
Saurabh Mukherjee, *Banasthali University, Rajasthan*
Ashvine Kumar, *HIM, Sonepat*
S. K. Bansal, *Govt. College of Engg. & Tech., Bikaner(Raj.)*
Saurabh Dutta, *Dr. B. C. Roy Engineering College, Durgapur*
P. Sanjeevikumar, *School of Electrical Engineering, VIT University, Chennai, India*
N. Rajkumar, *Sri Ramakrishna Engineering College, Vattamalaipalayam, Coimbatore, India*

Predicting and Analyzing Students' Performance: An Educational Data Mining Approach

Musa Peker
Department of Information Systems Engineering, Faculty of Technology, Mugla Sitki Kocman University, Mugla, Turkey
musa@mu.edu.tr

***Abstract*—Identification of factors affecting the success performance of students is an issue which has been studied extensively. In this study, the predicting and analyzing of student success in secondary school was conducted using data mining algorithms. Main subject targeted in this study is exploring the factors that affect student success significantly. For this purpose, determination of the factors which affect students' performance more is targeted with using feature selection algorithms. In the classification stage, four classification algorithms have been used. Data set consists of data collected from two public schools in the Alentejo region of Portugal. This data set includes information about student grades, demographic, social features and the school. In this study, students' performance in mathematics has been evaluated.**

Keywords*—*educational data mining; feature selection; classification; data mining.

I. Introduction

The increase in the amount of data has revealed the subject of how to benefit from these data. The determination of the relevant information will be difficult in the assessment performed by conventional methods. Depending on the technology, data mining method has been widely used in recent years for the knowledge discovery from large data sets [1].

Usage area of data mining is quite large. Data mining can be used in almost every area where data is obtained. Health, industry, marketing, banking and education are the main areas in which data mining is heavily used [1]. Today, information related to education is stored in the database and especially, information such as personal information, grades, courses failed and courses completed successfully is stored in large databases. It may be possible to identify problems in the field of education and improving the quality of education with the determination of the significant and important information from these data sets.

Some applications within the scope of educational data mining in the literature are as follows: Kotsiantis et al. [2] have used data mining techniques to predict performances of computer science students in a distance learning program of a university. Various demographic features and performance characteristics for each student have been given to the classifier as input. The best solution has been achieved with 74% accuracy using the Naive Bayes method. Also, it has been found that the past school grades have a much greater effect than demographic variables. In the study carried out by Halees [3], student behaviors have been tried to be evaluated with data mining and improving students' performance has been aimed according to the results achieved. For this purpose, personal, academic and e-learning system related records of 151 students have been used in Gaza Islamic University in 2007-2008 academic years. Similarly, in the study carried out by Bresfelean et al. [4] student profiles have been tried to be revealed using classification and data aggregation methods in data mining and improving of student successes have been aimed with determining the causes of academic failure. Gaafar and Khanmis [5] have applied different data mining methods to the database created using data obtained from different databases. In this study, modeling of two different student profiles (as successful and unsuccessful) has been aimed. In the study carried out by Zhang et al. [6], it has been investigated that how data mining helps students under risk and how obtained results can be adapted to the students. In the study carried out by Bırtıl [7], the questionnaire applied to determine the reasons for failure of students have been analyzed by using clustering method of data mining methods. In this study, it is aimed to identify the factors for students being unsuccessful. In the study carried out by Gulen and Ozdemir [8], data has been obtained by applying Academic Self-Concept Scale and Recreational Survey to 12 and over year-old gifted students. Apriori association rule algorithm has been used for the knowledge discovery from obtained data.

In this study, a feature selection-based data mining approach has been used for the estimation and evaluation of student success. In the study, real-world data obtained from two secondary schools in Portugal has been used. Portugal is a country that attracts attention in terms of education. Educational level of the population in Portugal has increased significantly in recent years. However, due to the high failure rates, Portugal is in the last row in Europe in terms of education. Significant decrease in success is seen especially in mathematics courses. The main purpose of this study is to predict student success and to identify important variables that affect educational success / failure. In this context, experiments have been performed on the data related to mathematic class.

II. Materials And Methods

A. Dataset

In this study, data collected from two public schools in the Alentejo region of Portugal during the 2005-2006 academic

2016 Sixth International Conference on Advanced Computing & Communication Technologies (ACCT-2016)
Copyright © 2016 ACCT 2016 Organisers :: *Published by* Research Publishing
ISBN: 978-981-11-0783-2 :: doi:10.3850/978-981-11-0783-2_395

year has been considered [9]. In Portugal, secondary education lasts 3 three years after 9-year basic education. After that, higher education begins. Students often take advantage of the free public education system. 20-point grade scale used for the evaluation of student success where 0 is the lowest grade and 20 is the best grade. Students are evaluated in three different periods during the academic year and final evaluation corresponds to the final grade (Table I - G3). Data includes student grades, demographic, social and school-related features. Features and descriptions in the data set are presented in Table I [9].

TABLE I. FEATURES AND DESCRIPTIONS

ID	Feature	Description (Domain)
1	school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
2	sex	student's sex (binary: female or male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: urban or rural)
5	famsize	family size (binary: ≤ 3 or > 3)
6	Pstatus	parent's cohabitation status (binary: living together or apart)
7	Medu	mother's education (numeric: from 0 to 4[a])
8	Fedu	father's education (numeric: from 0 to 4[a])
9	Mjob	mother's job (nominal[b])
10	Fjob	father's job (nominal[b])
11	reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
12	guardian	student's guardian (nominal: mother, father or other)
13	traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
14	studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
15	failures	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
16	schoolsup	extra educational school support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paidclass	extra paid classes (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
25	freetime	free time after school (numeric: from 1 – very low to 5 – very high)
26	goout	going out with friends (numeric: from 1 – very low to 5 – very high)
27	Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
28	Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
29	health	current health status (numeric: from 1 – very bad to 5 – very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20)

[a] 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

[b] teacher, health care related, civil services (e.g. administrative or police), at home or other.

B. ReliefF Feature Selection Algorithm

Relief method performs the feature selection process using the distance information between samples in the data set [10]. In this method, a model is created depending on proximity of a sample in the data set to other samples in same class and depending on sample's distance with different classes. The algorithm originally developed for the two-class problems has been adapted to the multi-class problems as ReliefF [11]. $W[A]$ weight for an A feature in a R sample is calculated as below when nearest sample in its class is H and nearest samples in C number of classes that it's not belong are $M(C)$,

$$W[A] = W[A] - \mathrm{diff}(A, R, H)/m + \sum_{C \neq \mathrm{class}(R)} \left[P(c)\,x\,\mathrm{diff}(A, R, M(C))\right]/m \quad (1)$$

Here, m is normalization coefficient, $diff$ is the difference function between two samples.

C. Decision Trees

Decision tree is a rule-based algorithm commonly used in the solution of classification problems. The reason for the widespread preference of this method is the rules which are used to create the tree structure are understandable and simple. In this method, multi-stage or sequential approach is used in carrying out the classification process. In this study, C4.5 algorithm has been preferred among the decision tree algorithms.

D. Feed-Forward Neural Networks

Artificial neural networks (ANN) is a mathematical system that has been developed taking into account the working principle of human brain. In this system, there are several processing units connected to each other in weighted form. These processing units receive signals from other neurons, unify them, transform them and reveal a numeric result. In this study, feed-forward neural network which is among the neural network models has been used.

E. Support Vector Machines

SVM is a machine learning method based on statistical learning theory. This method is a classification and regression method which easily classifies data sets (with the help of the kernel functions that it uses) that difficult to be classified (linear or nonlinear) [12]. The method determines the largest margin among many possible linear function to distinguish data which is linearly classifiable. It also transfers data which cannot be linearly classified to higher-dimensional space by using kernel functions and finds multiple planes with the greatest margin.

F. Kernel Based Extreme Learning Machine

ELM is a learning algorithm developed for a feed-forward neural network with one hidden layer [13]. Unlike the gradient-based feed-forward networks, input weights and threshold values are randomly generated in ELM. In addition, analytical methods are used in the calculation of the output weights. In this way, the learning process is expedited. Huang et al. [14] proposed a kernel-based method to improve the generalization capabilities of ELM. Many kernel functions such as linear, polynomial and radial basis functions can be used in kernel-based ELM. In this study, polynomial kernel function has been preferred.

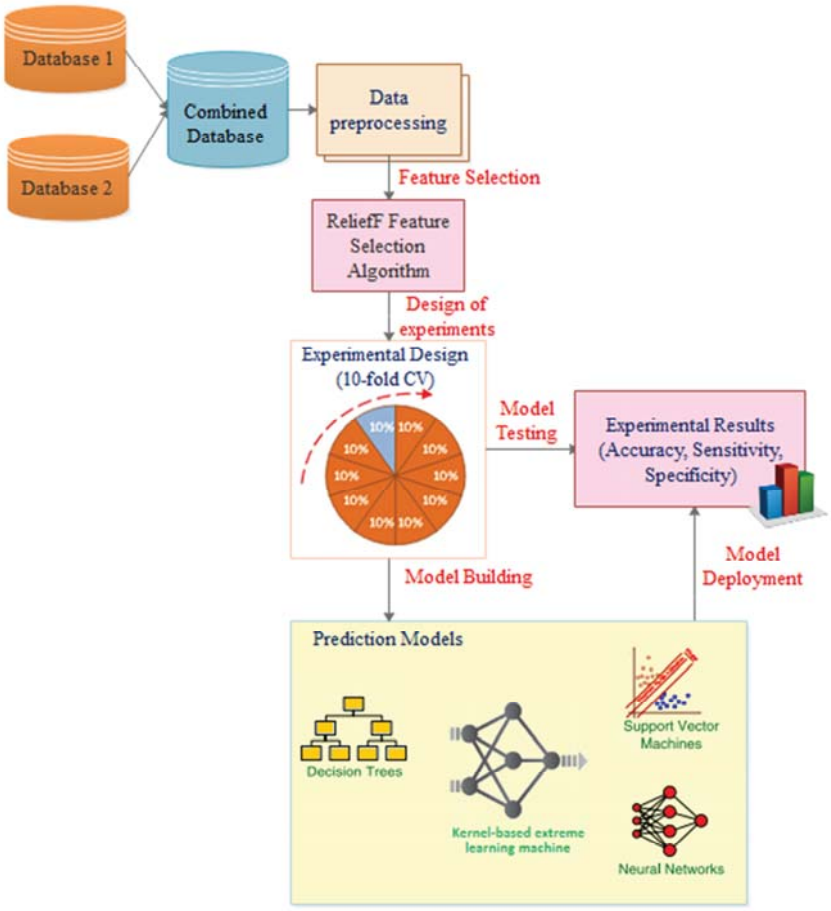

Fig. 1. A graphical depiction of the methodology followed in this study.

III. Experimental Results And Discussions

The block diagram of the proposed method has been presented in Fig. 1. First, feature set has been submitted to the system. In next stage, normalization process has been applied (between 0-1) to the data to improve the performance of classification processes. Data has been reduced to 0-1 with normalization method.

After normalization process, ReliefF algorithm has been used to determine efficient features. The sorting of efficient features obtained by applying ReliefF algorithm is as follows: f_{32}, f_{31}, f_2, f_9, f_{28}, f_7, f_{18}, f_{15}, f_{14}, f_4, f_{16}, f_{17}, f_{27}, f_6, f_{21}, f_5, f_{22}, f_{29}, f_{30}, f_8, f_3, f_1, f_{26}, f_{11}, f_{24}, f_{25}, f_{19}, f_{23}, f_{13}, f_{20}, f_{12}, f_{10}. Here, f_3, represents second feature according to Table I, "age". Effect ratios of some features are given in Fig. 2.

In this study, Erasmus grade conversion system based 5-levels classification approach has been used to estimate student success. Accordingly, grades are converted into 5 levels corresponding to particular ranges. This conversion table is presented in Table II.

TABLE II. The five level classification

Five level	I (excellent)	II (good)	III (satisfactory)	IV (sufficient)	V (fail)
0-20 grade scale	16-20	14-15	12-13	10-11	0-9

The histogram graph which shows the distribution of grades with respect to the number of students in class is presented in Fig. 3.

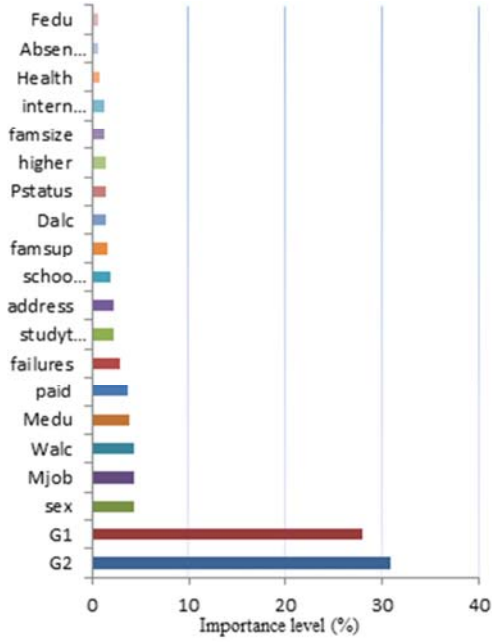

Fig. 2. Effect ratio of the features

Changes in the classification accuracy with the increase in the number of features selected by ReliefF algorithm are shown in Fig. 4. In general, it is observed that the best results are obtained with KELM algorithm. The highest classification accuracy has been obtained by classifying first 9 features by KELM algorithm. The effect of first two features on the results is quite high. These features are midterm grades (G1 and G2 grades) of the students. The most successful result has been obtained with C4.5 decision tree algorithm after KELM algorithm.

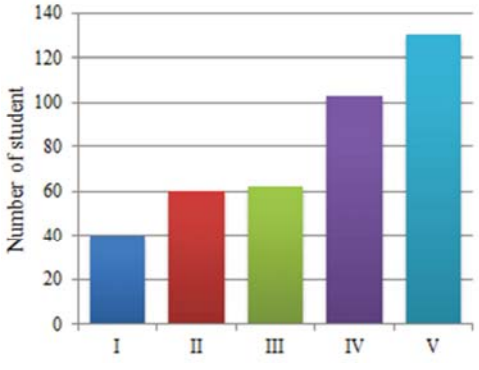

Fig. 3. Histograms for the output variables (5-level classification)

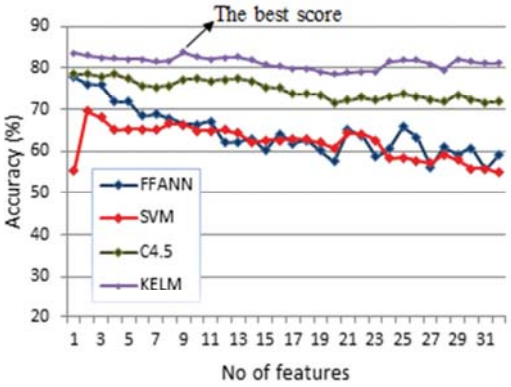

Fig. 4. Changes in the classification accuracy with the increase in the number of features selected by ReliefF algorithm

IV. Conclusion

In this study, a study was conducted for the evaluation and estimation of student success by using data mining techniques. The results obtained are presented below.

- The effect of features selected by feature selection algorithm on the performance has been positive and more according to the use of all the features. Student success has been estimated with high accuracy with only nine features.
- The study offers a different solution in terms of both identification of the most effective features among the features that are used and identification effective classification algorithm.
- The effect of mid-term grades has been high in the estimation of the end of year success of the student.
- The 9 features which give best results with KELM algorithm are as follows, respectively: G2, G1, student's sex, mother's job (Mjob), weekend alcohol consumption (Walc), Mother's education (Medu), extra paid classes (paid), number of past class failures (failures), weekly study time (studytime). It can be said that these features significantly affect student success.
- According to the obtained results, the factor least affecting student success is Father's Job.

References

[1] J. Han and M. Kamber, Data Mining: Concept and Techniques, USA: Academic Press, 2001.

[2] S. Kotsiantis, C. Pierrakeas and P. Pintelas, "Prediction of student's performance in distance learning using machine learning techniques", Applied Artificial Intelligence, vol. 18, no. 5, pp. 411-426, 2004.

[3] A. Halees, "Mining students data to analyze learning behavior: A case study", The International Arab Conference of Information Technology, 2008.

[4] P. Bresfelean, M. Bresfelean and N. Ghisoiu, “Determining students’ academic failure profile founded on data mining methods”, Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, pp. 23-26, 2008.

[5] L. Gaafar and M. Khamis, “Applications of data mining for educational decision support”, Proceedings of the 2009 Industrial Engineering Research Conference, pp. 228-233, 2009.

[6] Y. Zhang, S. Oussena, T. Clark and H. Kim, “Use data mining to improve student retention in higher education: A Case Study”, Proceedings of the 12th International Conference on Enterprise Information Systems, pp. 190-197, 2010.

[7] F.S. Bırtıl, “Analysis of girls vocational high school students’ academic failure causes with data mining techniques”, Master's thesis, Afyon Kocatepe University, Turkey, 2012.

[8] O. Gulen and S. Ozdemir, “Analysis of gifted students’ interest areas using data mining techniques”, Journal of Gifted Education Research, vol. 1, no. 3, pp. 215-226, 2013.

[9] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance”, In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference, pp. 5-12, 2008.

[10] K. Kira and L. Rendell, “The feature selection problem: Traditional methods and a new algorithm”. The Tenth National Conference on Artificial Intelligence (AAAI-92), 1992.

[11] I. Kononenko et al., “Overcoming the myopia of inductive learning algorithms with RELIEFF”, Applied Intelligence, vol. 7, no. 1, pp. 39-55, 1997.

[12] C. Cortes and V. Vapnik, “Support-vector networks”, Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[13] G.B. Huang, Q.Y. Zhu and C.K. Siew, “Extreme learning machine: theory and applications”, Neurocomputing, vol. 70, no. 1, pp. 489-501, 2006.

[14] G.B. Huang, H. Zhou, X. Ding and R. Zhang, “Extreme learning machine for regression and multiclass classification”, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, vol. 42, no. 2, pp. 513–529, 2012.