Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach

Ahmet TEKIN*

Suggested Citation:

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research* 54, 207-226.

Abstract

Problem Statement: There has recently been interest in educational databases containing a variety of valuable but sometimes hidden data that can be used to help less successful students to improve their academic performance. The extraction of hidden information from these databases often implements aspects of the educational data mining (EDM) theory, which aims to study available data in order to shed light on more valuable, hidden information. Data clustering, classification, and regression methods such as *K*-means clustering, neural networks (NN), extreme learning machine (ELM), and support vector machines (SVM) can be used for to predict aspects of the educational data. EDM outputs can ultimately identify which students will need additional help to improve their grade point averages (GPAs) at graduation.

Purpose of Study: This study aims to implement several prediction techniques in data mining to assist educational institutions with predicting their students' GPAs at graduation. If students are predicted to have low GPAs at graduation, then extra efforts can be made to improve their academic performance and, in turn, GPAs.

Methods: NN, SVM, and ELM algorithms are applied to data of computer education and instructional technology students to predict their GPAs at graduation.

Findings and Results: A comparative analysis of the results indicates that the SVM technique yielded more accurate predictions at a rate of 97.98%. By contrast, the ELM method yielded the second most accurate prediction rate (94.92%) evaluated based on the criterion of correlation coefficient. NN reported the least accurate prediction rate (93.76%).

Conclusions and Recommendations: The use of data mining methodologies has recently expanded for a variety of educational purposes. The

*Dr. Fırat University, Turkey, e_mail: atekin@firat.edu.tr

assessment of students' needs, dropout liability, performance, and placement test improvement are some important emerging data mining applications in education. Since educational institutions have several seemingly unsolvable domain-related problems, this study's results reveal that EDM can assist with how educational institutions analyze and solve these problems. Furthermore, ensemble models can be used to obtain improved results, while feature selection algorithms can be used to reduce the computational complexity of the prediction methods.

Keywords: GPA prediction, educational data mining, prediction methods, higher education

Data mining is the process of extracting important patterns from a given database and is therefore a valuable tool for converting data into usable information. Data mining has a wide range of applications in different areas, including marketing, banking, educational research, surveillance, telecommunications fraud detection, and scientific discovery (Han & Kamber, 2008). More specifically, data mining can discover hidden information to inform decision-making in various domains. The education system is one of these domains in which the primary concern is the evaluation and, in turn, enhancement of educational organizations.

Institutions of higher learning such as universities are at the core of educational systems in which extensive research and development is performed in a competitive environment. The prerequisite mission of these institutions is to generate, collect, and share knowledge. Specifically, universities commonly require knowledge mined from past and current data sets that, once mined, can be used for representing and delivering information to university administrators for monitoring conditions and taking action to resolve problems.

A growing volume of data is currently stored in educational databases that contain various hidden information that can help to improve the academic performance of students. Educational data mining is thus used to study available data and extract the hidden information for subsequent processes. This hidden information can be used in several educational processes such as predicting course enrollment, estimating student dropout rate (Yukselturk, Ozekes, & Turel, 2014), detecting abnormal values in the result sheets of students, and predicting student performance. Several prediction techniques can be used to help the educational institutions to predict their students' grade point averages (GPAs) at graduation. If this prediction output indicates that a student will have a low GPA, then extra efforts can be made to improve the student's academic performance and, in turn, his or her GPA at graduation. In this context, neural networks (NN), support vector machines (SVM), and extreme learning machine (ELM) algorithms can be applied to such data, and the comparative analysis of results can indicate that which students should receive extra academic help.

Since data mining techniques can be used to identify student performance trends, many researchers and authors have investigated EDM. In this study, a literature

review concerning the EDM was conducted to better understand the importance of EDM's applications in higher education, especially regarding the improvement of student performance.

Bharadwaj and Pal (2011a) used EDM to evaluate student performance among 300 students from five different colleges who were enrolled in an undergraduate computer application course. The study employed a Bayesian classification scheme of 17 attributes, of which student performance on a senior secondary exam, residence, various habits, family's annual income, and family status were shown to be important parameters for academic performance. In a subsequent study, Bharadwaj and Pal (2011b) constructed a new data set with the attributes of a student attendance and test, seminar, and assignment scores in order to predict academic performance. Meanwhile, Ramaswami and Bhaskaran (2009) compared various feature selection methods for obtaining the best feature combination for improving prediction accuracy. Their data set included several interesting features such as student vision, eating habits, and family attributes. More recently, Sen, Uçar, and Delen (2012) used various data mining models to predict secondary education placement test results. They investigated sensitivity analysis identifying the most important predictors and also demonstrated that compared to NN, SVM, and logistic regression models, the C5 decision tree algorithm was the best predictor. A similar work was earlier proposed by Kovacic (2010), who used EDM to identify the extent to which enrollment data could be used to predict student academic performance. For this purpose, CHAID and CART algorithms were used on a dataset of student enrollment of information system students at the Open Polytechnic of New Zealand. Among other studies, Ben–Zadok, Hershkovitz, Mintz, and Nachmias (2009) presented a student warning scheme that uses student data to analyze learning behavior and warn them of risk before their final exams. Al-Radaideh, Al-Shawakfa, and Al-Najjar (2006) used data mining methods to analyze student academic data and improve the quality of the higher educational system. Feng, Beck, Heffernan, and Koedinger (2008) conducted a study to predict the standardized tests scores of students in middle and high schools that used a regression model with 25 variables. Kobrin, Camara, and Milewski (2002) studied student SAT scores and high-school grades within several diverse student bodies and ultimately determined three groups. While the first group comprised students with no significant variance in grades or test scores, the second group contained students whose SAT scores were significantly better than their grades would have otherwise suggested. Finally, the third group consisted of students whose SAT scores were abnormally low compared to their high-school performance and, interesting, was represented by women and minority students more heavily than the other two groups. An unsupervised *k*-means clustering algorithm was proposed by Shaeela, Tasleem, Ahsan, and Khan (2010) to predict student's learning activities; results suggested that the outputs could be helpful for both instructors and students. A similar work was conducted by Erdoğan and Timor (2005), who proposed the *k*-means algorithm to identify student characteristics of 722 students at Maltepe University; the study sought a probable relationship between the university entrance exam results and other academic achievements. Luan (2002) proposed a data mining application in which the

satisfaction of students at institutions of higher education was measured according to various student characteristics. Vranić, Pintar, and Skoćır (2007) investigated the use of data mining for improving various aspects of educational quality for students in specific courses as target audiences in academic environments. A final grade prediction system was also proposed by Minaei–Bidgoli, Kashy, Kortmeyer, and Punch (2003), who used several features obtained from logged data in an educational web-based system. In this study, it was shown that a genetic algorithm-based ensemble model yielded improvements against a single model. The accuracy improvement ranged from 10–12%. Kotsiantis, Patriarcheas, and Xenos (2010) developed an ensemble model for predicting student performance in a distance learning system for which an incremental version of Naive Bayes, 1-NN, and WINNOW algorithms were combined by way of voting.

Method

Data mining, which is considered as an interdisciplinary aspect of computer science, is a computation-based pattern search process for large datasets. It involves several methods such as artificial intelligence, machine learning, statistics, and database systems. Predicting student GPAs is one application within the domain of education, though it requires that several parameters be considered. According to Sen at al. (2012), the effective prediction of student academic performance requires a prediction model that includes all personal, social, psychological, and other environmental variables.

Research Design

In this study, the GPAs of student in computer education and instructional technology at the end of their first-, second-, and third-year courses were used to predict their GPAs at graduation. Although the nature of this study can be defined as descriptive study, a data mining methodology was followed that involves data preparation, the creation of the prediction model, and the evaluation of the created model. A schematic illustration is depicted in Figure 1. A common data preparation step involves data preprocessing, data cleaning, and data transforming, all of which use several algorithms. Model creation requires generating a wide range of models using analytical techniques, and selecting the appropriate modeling technique and subsequent selected model parameters is necessary for optimal performance. To this end, several artificial intelligence models can be tuned to model the investigated system. Model evaluation involves assessing the validity and utility of the models against each other and against the goals of the study.

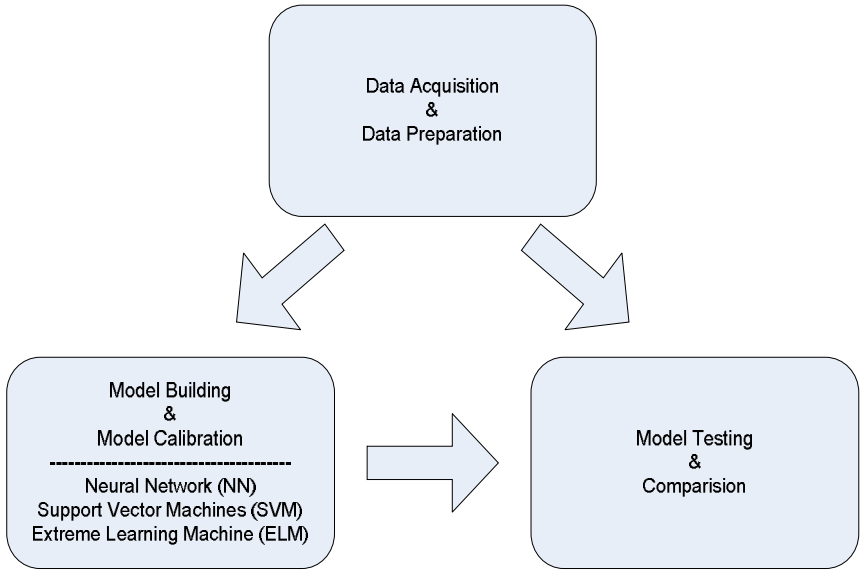


*Figure 1.*Illustration of the proposed methodology.

Research Sample

The dataset used in this study was acquired from Student Affairs of Fırat University in Turkey. The data consisted of 127 unique undergraduate student records of students of the Department of Computer Education and Instructional Technology enrolled in the university either from 2006 to 2010 or from 2007 to 2011.

Research Instruments and Procedures

Datasets contain scores of students' 49 vocational and cultural courses that are required to be successfully passed by students prior to their graduation. A student's score in a course was calculated by using the relative evaluation system, the principles of which were determined by the university senate. Exams were evaluated out of 100 points possible. Course scores were calculated according to the contribution of the mid-term exam (40%) and the general or supplementary exam (60%). As a result, the calculated course scores were converted to success coefficients shown in Table 1.

Table 1.
Course Score Conversion Table.

Course Score	*Success Coefficients*
AA	4.0
BA	3.5
BB	3.0
CB	2.5
CC	2.0
DC	1.5
DD	1.0
FF	0.0

A 127 × 49 data matrix was thus constructed in which the rows show the students and the columns show the lesson scores for subsequent purposes of prediction. The GPA of a student was calculated with the weighted mean of scores of all 4 years.

Data Analysis

In this study, NN, SVM, and ELM classification methods were used separately to predict the student's GPA, and the methods were compared to one another. These prediction methods were chosen due to their superior capability in classification problems.

Neural networks (NN). NNs are biologically inspired mathematical methods capable of modeling extremely complex nonlinear functions (Haykin, 2008). Any NN is composed of an interconnected group of artificial neurons, the information of which is processed by using a connectionist approach (Guldemir & Sengur, 2007). NNs can be used for the purposes of both classification and regression. Moreover, NNs can model the nonlinear relationship of dependent variable with independent variables completely based on data without any statistical assumptions. A multilayer perceptron (MLP) is the most popular NN structure; it uses a monitored learning method.

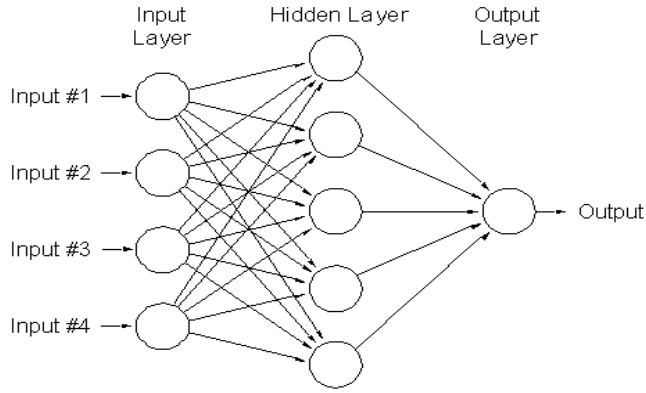
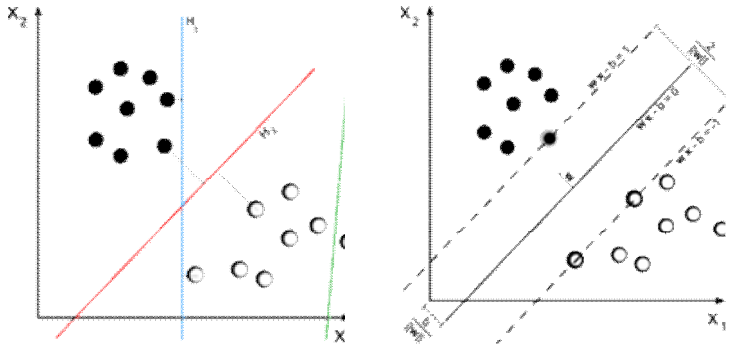


Figure 2. NN architecture.

The MLP structure was used in this study with back-propagation type supervised-learning algorithm. MLP is capable of modeling input–output relationships for the purposes of classification or regression. The NN architecture used in this study is shown in Figure 2, which shows the architecture's input layer, a hidden layer, and an output layer. All neurons were connected to each other from the input to output layer. The input layer neuron number was selected as the number of variables in the input dataset. During the iterative experimentation process, the number of the hidden layers and the number of neurons in each hidden layers were determined.

Support vector machines (SVMs). SVMs are in the family of generalized linear models, which perform the tasks of classification and regression by using the linear combination of features derived from the variables (Suykens & Vandewalle, 1999). In SVM methodology, input data are transformed to a high dimensional feature space, since after this transformation the input data become more manageable compared to the original input space (Esen, Ozgen, Esen, &Sengur, 2009). In addition, SVMs aim to separate training data into many classes by using the mathematical definition of a hyperplane. Figure 3 shows a two-class classification problem. Before classifying new data, SVMs identify the best hyperplane for modeling the training dataset.



(a) (b)

Figure 3. (a) A hyperplane separating bi-class data; (b) Best hyperplane and margins. (http://en.wikipedia.org/wiki/Support_vector_machine)

Extreme learning machine (ELM). In most applications, NNs are trained with finite datasets, and the input and hidden layer weights must be tuned during iterations. Huang, Zhu, and Siew (2006) recently proposed a single hidden layer feed for an NN model with extremely fast learning algorithms. This new method was called ELM.ELM does not have any tunable parameters and is a one-pass procedure. In other words, there are no iterations in the training of the network; once the input weights are assigned randomly, weights between the hidden and output layer are calculated with Moore–Penrose pseudo inverse. Once the input layer part of the ELM

is processed, the rest of the network can be considered a least squares problem, the solution of which facilitates training the network.

k-Fold Cross Validation. The *k*-fold cross validation test was used in this study. In the *k*-fold cross validation procedure, the input dataset was randomly divided into *k* subsets, each of which was considered to have approximately equal data points. *k*-fold cross validation method was repeated *k* times (Esen, Inalli, Sengur, & Esen, 2008a; Esen, Inalli, Sengur, and Esen, 2008b), and at each time, one of the *k* subsets was used as the test set while the other subsets (*k-1*) were used to form the training dataset. Thus, every data point appeared in a test set only once and in a training set *k-1* times.

Evaluation Criteria. For the objective evaluation of the investigated methods, several statistical methods—namely, the root-mean squared (RMS), the coefficient of multiple determinations (R^2), and the coefficient of variation (COV)—were used (Bulut & Kan, 2012; Esen et al., 2008a; Esen et al., 2008b). The RMS was defined as:

$$RMS = \sqrt{\frac{\sum_{m=1}^{n} (y_{pre,m} - t_{pre,m})^2}{n}}, \quad (1)$$

The correlation coefficient (R) and COV in percent were used in evaluating the methods. The related definitions were:

$$R = \frac{\sum_{m=1}^{n} ((y_{pre,m} - y_{mea})(t_{pre,m} - t_{mea}))}{\sqrt{\sum_{m=1}^{n} (y_{pre,m} - y_{mea})^2 \sum_{m=1}^{n} (t_{pre,m} - t_{mea})^2}} \quad (2)$$

$$\mathrm{cov} = \frac{RMS}{|t_{mea}|} 100 \quad (3)$$

In the above equations, n signifies the number of data points, y_{pre} signifies the predicted value, and t_{pre} signifies the actual dataset. t_{mea} and y_{mea} are the mean values of the measured and predicted data points, respectively.

Results

The main aim of this work was to predict well in advance the students' GPAs at graduation in order to reveal whether a student tends to have a GPA so that extra efforts can be made to improve the student's academic performance and, in turn, improve his or her GPA.

Two different scenarios were investigated in this study. The students' GPAs from the first 2 years scores (i.e., scores of 24 courses) comprised the first scenario. Students' GPAs from the first 3 years were used for prediction in the second scenario, the GPA of which includes a total of 38 course scores.

As aforementioned, several data mining prediction tools were applied in the study. All parameters for each prediction method were tuned according to these extensive experiments. For the NN prediction method, one hidden-layered NN model with tangent sigmoid activation function was constructed. The input layer had 24 and 38 neurons according to the two scenarios, while the output layer had one neuron. The hidden-layer neuron number was determined according to the neuron number of the input and output layers. For the hidden layers of the two different NN topologies, 25 and 39 neurons were used. The linear activation function was chosen for the output layer. The scaled conjugate gradient algorithm was also used to reveal the best performance in the experiments.

To find the optimal parameters for SVM, a search mechanism in the 2-D gamma versus sigma plane was examined to obtain optimal gamma and sigma values. The resultant gamma and sigma values were 75.1138 and 5.3491, respectively. The radial basis function kernel was selected, which yielded the best performance in the experiments.

For the ELM structure, the sigmoid activation function in the hidden layer was selected, and the number of neurons in the hidden layer was set to 25.

According to the 5-fold cross validation, the training dataset contained approximately 100 samples, while the test dataset contained about 25. These values were not fixed; when divided for 5-fold validation, the whole dataset (127 samples) included one fold containing 27 samples or two folds containing 26. The prediction results of the three modeling methods for the first scenario are presented in Table 2. The results presented in Table 2 are the 5-fold cross validation results.

Table 2.

Prediction Results for All Classification Methods.

Prediction Method	***RMSE***	***Correlation coefficient***	***COV***
NN	0.2068	0.8494	7.4005
SVM	0.1146	0.9306	4.0997
ELM	0.1200	0.9241	4.2939

As the results indicate, all three prediction methods performed reasonably well in predicting the student GPA at graduation. Among the three model types, SVM algorithms produced the most accurate prediction results, in which 93.06% correlation coefficient, 0.1146 RMSE, and 4.0997 COV values were obtained with the 5-fold cross validation test. The second most accurate results were obtained with the ELM method. The recorded correlation coefficient, RMSE, and COV values were 92.41%, 0.1200, and 4.2939, respectively.

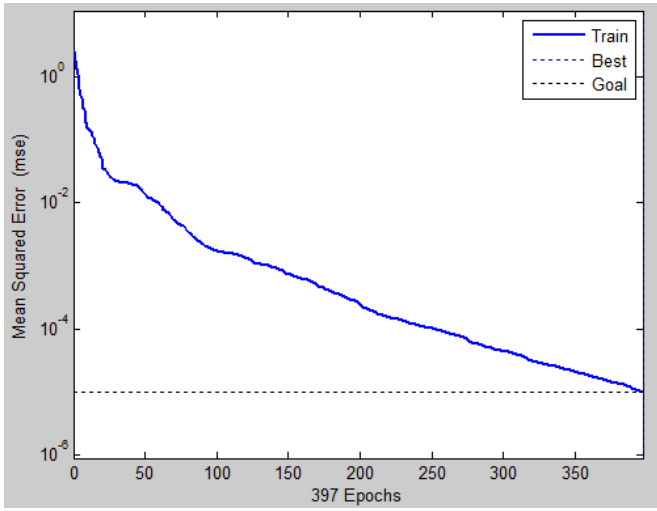


Figure 4. Training performance of the NN model.

The least accurate results were obtained by the NN method, in which the lowest correlation coefficient (84.94%) and highest RMSE (0.2068) and COV (7.4005) values were recorded. Other than these quantitative performance evaluation results, prediction results were qualitatively evaluated by plotting both the predicted and actual GPAs shown in Figures 5-11. The training performance of the NN model is illustrated in Figure 4. The NN model could learn the first scenario for 397 epochs.

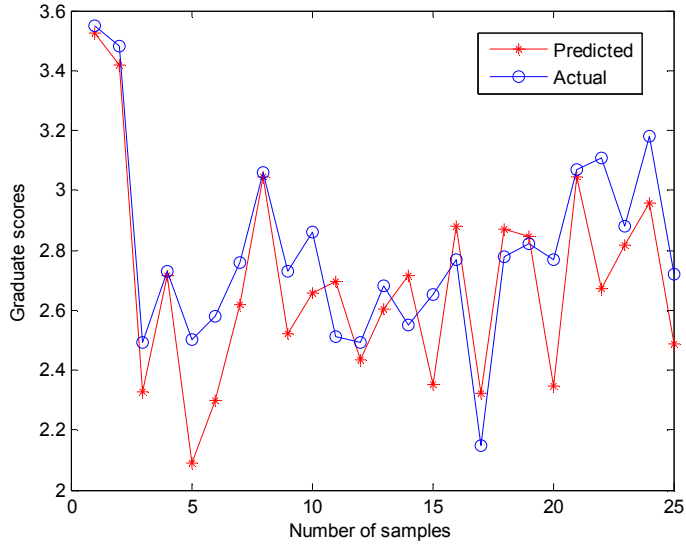


Figure 5. Prediction results of the NN method.

Figure 5 shows the NN predictions for 25 test samples. Except test samples 5, 6, 20, and 22, predictions were close to the actual samples.

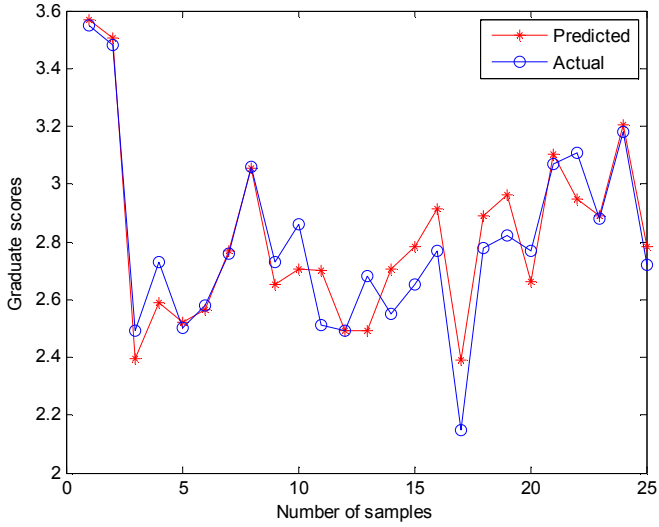


Figure 6. Prediction results of the SVM method.

Figure 6 illustrates the SWM predictions for the same test samples. Since only a few samples were not enough close to the actual samples, it was clear that SWM prediction was more reliable than NN predictions.

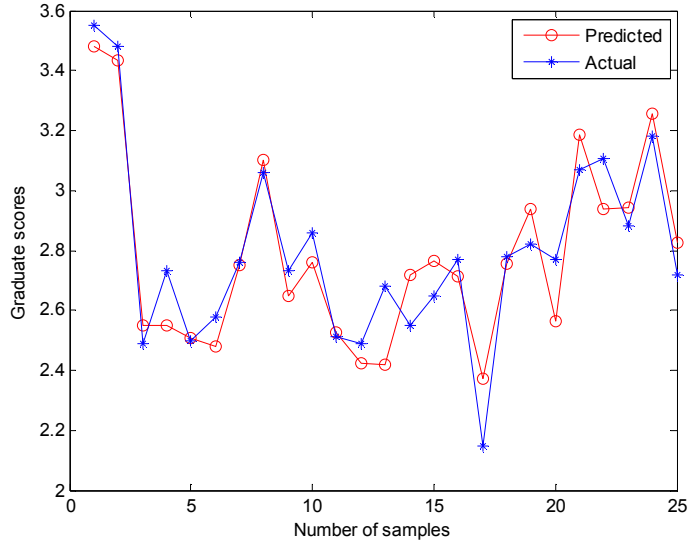


Figure 7. Prediction results of the ELM method.

Figure 7 illustrates the ELM results. Similar to the quantitative results, the ELM prediction illustration also indicates results close to the actual results.

The experiments were repeated for the second scenario, in which GPAs from the 3 years of coursework were used to predict the GPAs of students at graduation. As aforementioned, 38 course scores were used to predict the students' GPAs at graduation with the NN, SVM, and ELM methods, respectively. Similar to the first scenario, the model parameters were tuned with the experiments and the best results were obtained by using the same parameters of the first scenario. However, the hidden-layer neurons of the NN and ELM methods were changed.

The prediction results of the three modeling methods for second scenario are presented in Table 3. The results presented in Table 3 are the average performance results of the 5-fold cross validation results.

Table 3.

Prediction Results for All for Classification Methods for The Second Scenario.

Prediction method	***RMSE***	***Correlation coefficient***	***COV***
NN	0.1329	0.9376	4.7547
SVM	0.0708	0.9798	2.5334
ELM	0.1010	0.9492	3.6136

As the results indicate, all three prediction methods performed reasonably well in predicting student GPAs at graduation. Among the three model types, SVM algorithms again produced the most accurate prediction results with 97.98 % correlation coefficient, 0.0708 RMSE, and 2.5334 COV values were obtained with 5-fold cross validation test. The second most accurate results were obtained by the ELM method with correlation coefficient, RMSE, and COV values of 94.92%, 0.1010, and 3.6136, respectively.

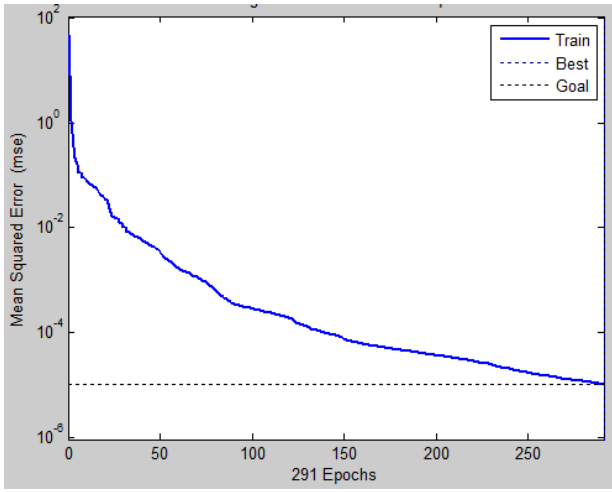


Figure 8. Training performance of the NN model for the second scenario.

The least accurate results were obtained by the NN method, for which the lowest correlation coefficient (93.76%) and highest RMSE (0.1329) and COV (4.7547) values were recorded.

Other than these quantitative performance evaluation results, the prediction results were qualitatively evaluated by plotting both the predicted and actual GPAs at graduation, as shown in Figures 9–11.

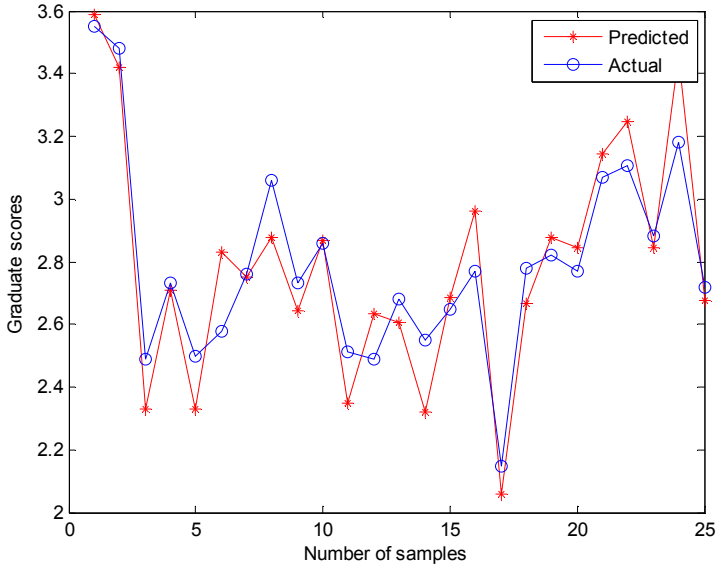


Figure 9. Prediction results of NN method for the second scenario.

Figures 9–11 show the predictions according to the NN, SVM, and ELM methods, respectively, while Figure 8 shows the training performance of the NN model.

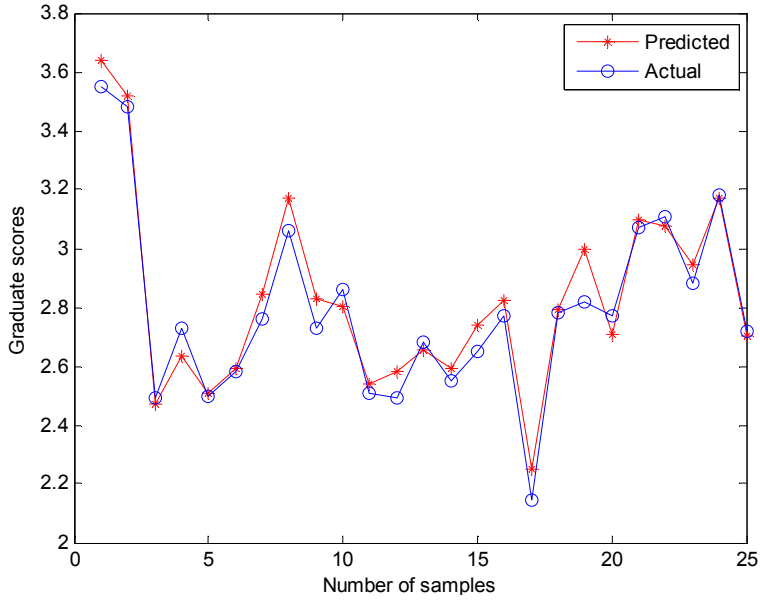


Figure 10. Prediction results of SVM method for the second scenario.

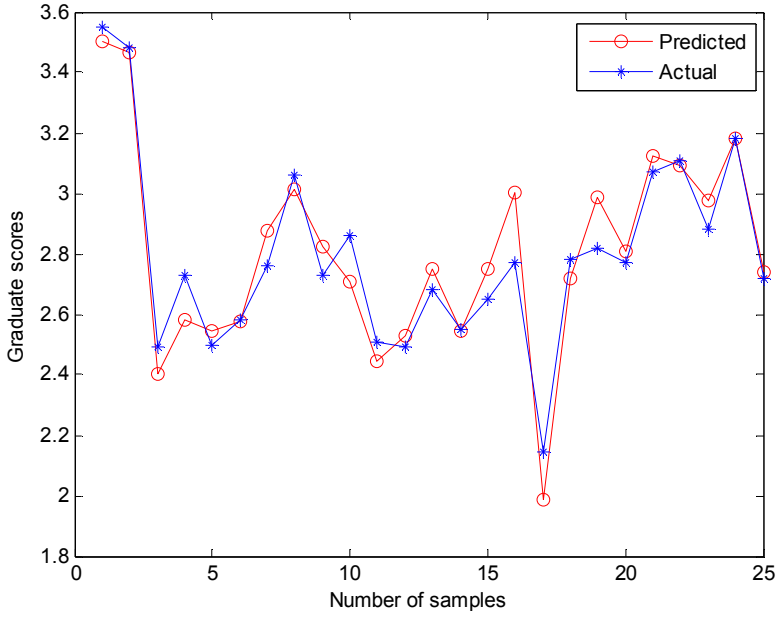


Figure 11. Prediction results of ELM method for the second scenario.

It is worth comparing the obtained results for the first and second scenarios. Upon examining the quantitative and qualitative results tabulated for all investigated methods, it is clear that there was improvement due to the augmentation of the number of course scores used as input for the prediction models.

Discussion and Conclusion

Educational research has taken advantage of data mining to a lesser extent than other domains, including banking, marketing, and healthcare. However, the current pace of applying data mining methods in the domain has increased for a variety of educational purposes. Assessing student need, detecting dropout rates, analyzing student academic performance, improving placement test scores are some of the important emerging data mining applications. As awareness of the capabilities of data mining increases in educational institutions, researchers will be able to identify as well as analyze and solve seemingly unsolvable domain-related problems.

As this study illustrated, several data mining prediction techniques can accurately predict student GPA at graduation well in advance, which can identify students needing extra help to improve their academic performance and, in turn, their GPAs at graduation. This study's sample consisted of 127 student GPAs that were measured and evaluated with tools assumed to be reliable and valid. In this context, three prediction methods—NN, SVM, and ELM—were employed to estimate student GPAs at graduation for the study of two scenarios. The first scenario was designed to estimate the GPAs at graduation of students according to their GPAs of coursework completed during their first 2 years of study. The results are as following; NN obtains 84.94%, SVM obtains 93.06% and ELM obtains 92.41% accuracy rates respectively. In the second scenario, the GPAs after the first 3 years of coursework were used as data. As cross validation results indicate, SVM techniques offered more accurate predictions than ELM and NN. The obtained results are 97.98%, 94.92% and 93.76 respectively. Moreover, for all methods, results indicated that GPAs after the first 3 years of coursework demonstrated improved predictions for all methods. ,

These results are in line with the previous results related to classifier algorithms. For example, Sen et al. (2012) used various data mining models to predict secondary education placement test results. The authors reached that C5 decision tree algorithm was the best predictor with 95% accuracy, SVM with an accuracy of 91% and NN with an accuracy of 89%. In addition, logistic regression models came out to be the least accurate with accuracy of 82%. In another study, Kotsiantis et al. (2010) developed an ensemble model for predicting student performance in a distance learning system for which an incremental version of Naive Bayes, 1-NN, and WINNOW algorithms were combined by way of voting. The obtained accuracy was 73% in the initial forecast. In addition, Minaei–Bidgoli et al. (2003) used combination of the multiple classifiers for predicting the final grade of the students. Without

genetic algorithm (GA), the authors reached the accuracy of 86.8% for 2-classes, 70.9% for 3-classes and 51.0% for 9-classes respectively. Moreover, with GA, the authors reached the accuracy of 94.09% for 2-classes, 72.13% for 3-classes and 62.25% for 9-classes respectively. Another study which was carried out by Kovacic (2010), used EDM to identify the extent to which enrollment data could be used to predict student academic performance. For this purpose, CHAID and CART algorithms were used and 59.4% and 60.5% accuracy rates were indicated.

In conclusion, as can be seen from the above comparisons, instead of choosing one classification method, researchers have proposed the use of ensemble models or combining multiple classifiers to achieve more robust prediction results. Moreover, feature selection algorithms have been employed to reduce the computational complexity of prediction methods.

References

AI-Radaideh, Q. A., AI-Shawakfa, E. W., and AI-Najjar, M. I. (2006).*Mining student data using decision trees.* International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.

Ben-Zadok G., Hershkovitz, A., Mintz, R. andNachmias, R.(2009). Examining online learning processes based on log files analysis: a case study. *Research, Reflection and Innovations in Integrating ICT in Education.*

Bharadwaj, B.K. and Pal, S. (2011a). Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 9(4), 136-140.

Bharadwaj, B.K. and Pal, S. (2011b). Mining Educational Data to Analyze Students' Performance. *International Journal of Advance Computer Science and Applications (IJACSA)*, 2(6), 63-69.

Bulut, O.,& Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari- Eurasian Journal of Educational Research,* 49, 61-80.

Erdoğan, Ş., Timor, M. (2005). A Data Mining Application in a Student Database. *Havacılık ve Uzay Dergisi.* 2(2), 57-64.

Esen, H., Inalli, M., Sengur, A., Esen, M.(2008a). Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing. *Int. J. Thermal Sciences*, 47(4), 431-41.

Esen, H., Inalli, M., Sengur, A., Esen, M. (2008b). Modelling a ground-coupled heat pump system by a support vector machines. *Renewable Energy*, 33(8), 1814-1823.

Esen, H., Ozgen, F., Esen, M. andSengur, A. (2009). Modelling of a new solar air heater through least-squares support vector machines. *Expert Systems with Applications*, 36(7), 10673-10682.

Feng, M., Beck, J., Heffernan, N., &Koedinger, K. (2008). *Can an intelligent tutoring system predict math proficiency as well as a standardized test?* In Baker & Beck (Eds.), Proceedings of the 1st international conference on education data mining, 107–116, Montreal, CA.

Guldemir, H, Şengür, A. (2007). Online Modulation Recognition of Analog Communication Signals using Neural Network. *Expert Systems with Applications*, 33 (1).

Han, J. Kamber, M. (2008). *Data Mining: concepts and techniques*. 2nd Edition, Morgan Kaufmann publishers.

Haykin, S. (2008). *Neural networks and learning machines* (3rd ed.)., New Jersey: Prentice Hall.

http://en.wikipedia.org/wiki/Support_vector_machine

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. (2006). Extreme Learning Machine: Theory and Applications, *Neurocomputing*, vol. 70, 489-501.

Kotsiantis, S. B., Patriarcheas, K., Xenos, M. N. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowl.-Based Syst.* 23(6), 529-535.

Kovacic, Z. J. (2010). *Early prediction of student success: Mining student enrollment data.* Proceedings of Informing Science & IT Education Conference.

Luan, J. (2002). *Data Mining, Knowledge Management in Higher Education, Potential Applications.* 42nd Associate of Institutional Research International Conference (Toronto,Canada: 2002), 1.

Milewski, G. B., Camara, W. J., &Kobrin, J. L. (2002). Students with discrepant high school GPA and SAT scores. *College Board Research.*

Minaei-Bidgoli, B., Kashy, D. A., Kortmeyer, G., & Punch, W. F. (2003). *Predicting student performance: An application of data mining methods with an educational web-based system.* In The proceedings of the 33rd ASEE/IEEE frontiers in education conference, Boulder, CO.

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *Journal of Computing*, 1(1), 7–11.

Sen, B.,Ucar, E. and Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, 39, 9468–9476.

Shaeela A., Tasleem M., Ahsan Raza S., Khan. M. I. (2010). Data mining model for higher education system. *Europen Journal of Scientific Research*, 43(1), 24-29.

Suykens, Johan A. K., Vandewalle, Joos P. L. (1999). Least squares support vector machine classifiers, Neural Processing Letters, 9(3), 293–300.

Vranić, M., Pintar, D., Skoćır, Z. (2007). *The Use of Data Mining in Education Environment*, ConTEL 2007 (Zagrep 13-15 June 2007), 243.

Yukselturk, E., Ozekes, S., Turel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program, *European Journal of Open, Distance and e-Learning*, 17(1), 118-133.

Öğrencinin Mezuniyet Notunun Erken Tahmini: Bir Veri Madenciliği Yaklaşımı

Atıf

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research* 54, 207-226.

Özet

Problem Durumu: Son zamanlarda, eğitim kurumlarının veri tabanlarında depolanan veriler giderek artmakta ve bu verilere büyük bir ilgi bulunmaktadır. Eğitim kurumlarındaki öğrenciler, dersler, akademik ve idari personel, yönetim sistemleri vb. veriler stratejik verilerdir. Stratejik verilerin çözümlenerek anlamlı bilgilerin ortaya çıkarılması, eğitim kurumlarının birtakım tedbirler alarak, eğitimdeki kaliteyi artırmasını sağlayacaktır. Eğitim kurumları daha çok öğrenci ve mezunların yol haritalarını tahmin etmeye odaklanmalıdır. Eğitimsel veri madenciliği, eğitim alanında mevcut verileri incelemek ve ondan gizli bilgiyi ortaya çıkarmak için kullanılır. Veriyi çözümlemek ve anlamlı bilgileri ortaya çıkarmada istatistiki yöntemler her zaman kullanışlı olmayabilmektedir. Bu durumlarda verileri işlemek ve çözümlemek için veri madenciliği yöntemleri kullanılmaktadır. Yapay Sinir Ağları, Destek Vektör Makineleri ve benzeri sınıflama ve regresyon yöntemleri, eğitim verilerinde tahmin amaçlı kullanılabilir. Bu tahmin, akademik başarısı zayıf öğrencilerin belirlemesinde ve onların başarılarının artırılmasında yardımcı olacaktır.

Araştırmanın Amacı: Bu çalışmanın amacı zamanında mezun olamayacak veya düşük bir ortalama ile mezun olabilecek başarısız öğrencilerin önceden tespit edilerek, mezun olabilecek bir seviyeye getirmek veya daha yüksek bir ortalama ile mezun olmalarına yardımcı olabilmektir. Bu amaçla, veri madenciliğinde kullanılan bazı tahmin teknikleri, eğitim kurumlarına yardımcı olmak üzere, öğrencilerin mezuniyet notlarının tahmininde kullanılmıştır. Bu tahmin, bir öğrencinin düşük bir lisans

ortalaması ile mezun olacağını bildirirse, o zaman öğrencinin başarısının artırılması için ekstra çaba gösterilecektir.

Araştırmanın Yöntemi: Veri madenciliğinde kümeleme, birliktelik kuralları ve sınıflandırma yöntemleri ile veriler analiz edilmektedir. Literatürde en çok kullanılan sınıflandırma yöntemleri arasında Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (DVM) ve Ekstrem Öğrenme Makinesi (EÖM) algoritmaları bulunmaktadır. Bu çalışmada Bilgisayar ve Öğretim Teknolojileri Eğitimi bölümü öğrencilerinin lisans mezuniyet notlarının tahmininde bu sınıflandırma yöntemleri kullanılmıştır. Öğrencilerin mezun olması için gerekli olan 49 adet mesleki ve kültürel ders, veri kümesinin özniteliklerini oluşturmaktadır. Veri kümesi oluşturulurken 127 öğrencinin ders notları göz önüne alınmıştır. Böylece 127x49'lik bir veri matrisi elde edilmiştir. Çalışmada iki farklı uygulama gerçekleştirilmiştir. Bunların ilkinde, öğrencilerin sadece ilk iki yılda aldıkları yılsonu notları göz önüne alınmıştır. Böylece toplam 24 adet dersin yılsonu notlarından, öğrencilerin mezuniyet notları YSA, DVM ve EÖM ile tahmin edilmiştir. İkinci uygulamada ise öğrencilerin ilk üç yılsonunda almış oldukları 38 adet dersin yılsonu notları kullanılarak, öğrencilerin mezuniyet notları YSA, DVM ve EÖM sınıflandırma yöntemleri ile tahmin edilmiştir. Gerçekleştirilen bilgisayar benzetimlerinde 5 katlı çapraz geçerlilik kullanılmıştır. Böylece, kullanılan sınıflandırma yöntemlerinde eğitim için yaklaşık 101 örnek ve test için de 26 örnek kullanılmıştır.

Araştırmanın Bulguları: Her iki uygulama için de gerçekleştirilen karşılaştırmalı analizler DVM tekniğinin en iyi sonuçları ürettiğini göstermiştir. DVM tekniğinin başarımı birinci uygulama için %93.06 ve ikinci uygulama için ise % 97.98'dir. Diğer taraftan EÖM ikinci en iyi tahmin başarımını göstermiştir. Korelasyon katsayısı değerlendirme kriterine göre birinci uygulama için %92.41 ve ikinci uygulama için ise % 94.92'lik bir başarım kaydedilmiştir. En kötü tahmin performansı YSA tarafından elde edilmiştir. Buradaki başarım birinci uygulama için %84.94 ve ikinci uygulama için ise % 93.76'dır.

Araştırmanın Sonuçları ve Önerileri: Günümüzde, veri madenciliği yöntemlerinin eğitim amaçlı kullanımı hızla artmaktadır. Öğrenci ihtiyaç değerlendirmesi, öğrencilerin okuldan ayrılma tahmini ve öğrencinin performans analizi, eğitim kurumları için önemli veri madenciliği uygulamalarından bazılarıdır. Eğitim kurumlarının çözülemeyecek gibi görülen bazı problemlerin analizinde ve çözümünde, veri madenciliği yetenekleri önemli rehberlik hizmeti verebilecektir. Gerçekleştirilmiş çalışmanın sonuçları incelendiğinde, kullanılan veri madenciliği yöntemlerinden elde edilen başarımlara göre, ikinci sınıf sonunda öğrencilerin mezuniyet notları en düşük %84.94 doğruluk ile tahmin edilmektedir. Böylece ikinci sınıfın sonundan itibaren, öğrencilere verilecek rehberlik hizmetleri ile öğrenci başarıları artırılabilecektir. Şöyle ki; öğrencilere etkili çalışma becerileri öğretilecek, derslerin teorik yapısı yanında uygulamalar yapmaya teşvik edilecek ve ders/ödev/projelerinin mutlaka zamanında yapılarak teslim edilmesi konusunda uyarılabilecektir. Bu ve benzeri yönlendirmeler ile öğrencinin lisans mezuniyet notunun yükseltilmesi sağlanabilecektir.

Son zamanlarda, veri madenciliği yöntemlerini kullanan tahmin uygulamalarında, tek bir tahmin yöntemi kullanmak yerine, daha iyi bir başarım için topluluk modelleri (ensemble model) veya birkaç farklı sınıflandırıcının kombinasyonu şeklindeki yapılar bir hayli dikkat çekmektedir. Diğer bir ifade ile çok sayıda sınıflandırma sonuçlarının çoğunluk oylaması (majority voting) ve ortalama gibi yöntemler ile birleştirilmesi dayanıklı tahmin yapıları oluşturabilmektedir. Bu gibi yapılar, ileriki çalışmalar da kullanılabilir. Diğer taraftan veri boyutlarının yüksek olması nedeniyle ortaya çıkan hesaplama yükü ağırlığı önemli bir problem olarak ortaya çıkmaktadır. Bu problemin çözümü için de özellik seçimi (feature selection) algoritmaları, yine ileriki çalışmalarda kullanılabilir.

Anahtar Sözcükler: Mezuniyet notu tahmini, eğitimsel veri madenciliği, tahmin metodları, yükseköğretim.