SETSCI Conference Proceedings,
4 (1): 130-133, 2019

ISAS 2019
3rd International Symposium on Innovative Approaches in Scientific Studies
Section: Engineering and Natural Sciences
April 19 – 21, 2019, Ankara, Turkey

Student Graduation Prediction with Data Mining: Amasya University Distance Education Application and Research Center Sample

Osman KAYHAN[1*+], Yavuz ÜNAL [2] and Ahmet SAĞLAM [3]

[1]Institute of Science / Technology and Innovation Department, Amasya University, Amasya, Turkey
[2]Computer Engineering/Technology Faculty, Amasya University, Amasya, Turkey
[3]Computer Programming/Merzifon Vocational School, Amasya University, Amasya, Turkey
[*]Corresponding author: osmankayhan@msn.com
[+]Speaker: osmankayhan@msn.com
Presentation/Paper Type: Oral / Full Paper

Abstract **–**Correct orientation of students during education is very crucial for preventing future failures. Students who cannot graduate at the time may cause negative impacts on the family and the country's economy as well as the decrease of the young labor force. This situation requires the studies to be held concerning the students who cannot graduate in time. Analyzing the educational data related to the students is included in this study. Considering the accumulation of large number of educational data in higher education institutions, it becomes more important to analyze these data with various methods. In data mining, which is one of the methods used to analyze the data, estimation, classification and clustering methods are benefited. In this study, inferences were tried to be made about whether the students who enrolled to Amasya University Distance Education Child Development, Medical Documentation and Secretariat Associated Degree Programs in 2016-2017 will be able to graduate on time or not. Algorithms such as Decision Tree, Naïve Bayes, Support Vector Machines and Random Forest have been used for estimation. The best estimation success was reached with the decision tree algorithm. The comparative study of the estimation classification of algorithms is included in the study.

Keywords – Data Mining, Decision Trees, Distance Education Application and Research Center Educational Data Mining

I. INTRODUCTION

It is observed that the developments in information and data communication technologies are developing at a great pace. Innovations also follow these technologies day by day. With these innovations, the decline in the prices of technology-based products is another important issue that attracts attention. These developments can provide users to have faster, more capable and more useful information technology products easily.

Every process is recorded in digital media with the convenience provided by information technologies and the increasing availability of these technologies in our lives. Increase of data records in very large proportions, development of automatic data collection centers, widespread use of business intelligence, classification of data qualitatively and quantitatively, increase in gene technologies, widespread use of large data with cloud technology, increase in data in commercial, social and economic areas, increasing the importance of the concepts of cheapening and customer satisfaction are the elements that require the use of data mining. [1]

Data mining is the process of discovering significant new correlations, patterns and trends by eliminating the stacks of data stored in storage environments using statistical and mathematical techniques together with pattern recognition technologies.[2]

In distance education institutions, courses, course times, student information, academic staff information, grade point averages, graduation degrees, etc. data are strategic data. The discovery of meaningful information hidden by processing these raw data will enable the educational institutions to take some measures to improve the quality of education. Statistical methods are not always enough to analyze this data and reveal meaningful information. In such cases, data mining methods are used to analyze and process data.[3]

Many researchers in the field of data mining in education have done studies on various subjects. In this study by Ayık, Özdemir and Yavuz [4], the link between the high school types and diploma grades of Atatürk University formal education students and the faculties they have enrolled are examined by using data mining techniques. The aim of this course is to determine the importance of high school type diploma grades on the gained faculty. According to this study, it is determined that the effect of high school type on the acquisition of a desired university is very crucial and the diploma grade is equally important. In the study conducted by Bresfelean and colleagues [5], it was aimed to determine the student status by using data clustering and classification methods in data mining and to prevent and eliminate student failures by revealing the reasons for the failure of the course. In a study carried out by Can [6], an end-of-term course evaluation questionnaire was applied to students of a public education institution. The questionnaire was applied to 5820 university students. Logistic regression and decision tree algorithms have been created by using SPSS Clementine and WEKA programs and the results obtained are interpreted.. According to the results of the analysis, it is seen that the success of the course is related to the improvement of the student's professional development and the development of world view.. In a study carried out by

Şengür and Tekin [7], 127 students who completed the Computer Education and Instructional Technology Program of Fırat University in 2011 were evaluated with the estimation of their graduation grades by using two of the VM methods of the end-of-term grades of 49 culture and vocational courses that they were responsible for during their undergraduate education.

In this study, year-end grades of culture and vocational courses of associate degree students enrolled in Medical Documentation and Secretariat program of Amasya University Distance Education Child Development program in 2016-2017 were used. Decision Tree, Naïve Bayes, Support vector machines and random forest algorithms were used to estimate whether they could graduate or not. The main aim of using four different algorithms in the study is to create a warning mechanism by predicting whether the student can graduate or not in the early stage and to warn the students who are successful in advance. In this study, the best estimation performance was determined by the decision tree method.

II. MATERIALS AND METHOD

Data mining methods were used to reveal the meaningful information that was hidden from the data used in this study. Data mining can be defined as the determination of rules that will enable future forecasting by processing raw data. [8]. Another definition of data mining is that the generic name given to the process of extracting meaningful information that is not apparent from the available data. Data mining can be used to estimate data from unknown sample data at hand. [2] In this research, it is tried to make inferences about whether students who are enrolled in Amasya University Distance Education Child Development program and Medical Documentation and Secretariat program in 2016-2017 will be able to graduate in time by using estimation from data mining methods. Algorithms such as Decision Tree, Naïve Bayes, Support vector machines and random forest were used for estimation. The results are given in Chapter 3.

A. Decision Tree

Hark [9] is one of the most commonly used algorithms for decision trees in which each node is a test on a property, each branch is the output of this test, each leaf refers to the node, and the top node is called the stem node. The flowchart made by Can [6] in a manner like a tree and the nodes on the decision tree represent a test on a property, and each branch represents the result of the test operation. Hark [9] these are the most widely used decision trees models:

- Chi square automatic interaction detector,
- Regression trees with classification,
- Interferences of Decision Trees
- C 4. Is same with 5

B. Naïve Bayes

Bilekdemir[10] this algorithm is a VM model based on the philosophy of calculating the effects of each criterion on the probability, and the chances of a result in this algorithm are multiplied by the chances of all factors affecting that outcome to realize that result. Taşdemir[11] In the Naive Bayes classification, the analysis of the contributions of the independent features is made and the possibility of a condition is determined. Algorithm test data is used to label unknown classes as a result. When Bayes classifiers are performed in huge databases and compared to artificial nets and decision-making networks, they can provide great benefit.

C. Support Vector Machines

DVM, which forms an N-dimensional hyperplane, optimally divides the data into 2 categories. It is closely related to artificial neural networks and uses a sigmoid kernel function and has a two-layer feeder artificial neural network.[12] The crucial feature of DVM is that it minimizes the mean error frame on the data set. One of the basic assumptions of the DVM is that all samples in the training set are distributed independently and similarly. [13]

DVM has four core functions that are widely used. These functions: [14]

- Linear Function,
- Polynomial Function
- Sigmoid Function,
- Radial Based Function

D. Random Forest

This method, which consists of many decision trees, is also known as the forest classifier. Classification or regression trees can be established by RF method and clustering can be done. [15]. If the class variable in the data set is categorical, the classification, if it is continuous, regression trees are installed. Each decision tree in the forest is created by selecting random samples from the original data set with bootstrap technique and selecting the number of random variables from all variables in each decision node. In this method trees are created with the CART algorithm and trees are not chopped. The CART algorithm decides from which variable the data set is to be divided into branches, using the information gain.

III. DISCUSSION

Data mining is an interesting tool to discover interesting results from a large amount of data. In this study, the data of the associate degree students who enrolled to the Medical Documentation and Secretariat program and the Distance Education Child Development program of Amasya University in 2016-2017 were used. Year-end pass grades were calculated by using the midterm, final and make-up exams of each student and their success status was examined.

Data Preparation

The raw data were first obtained from the study. The data in separate tables are combined into a single table. The missing data with the data that are not needed in the analysis were removed.

While determining the graduation status of the students and the students who enrolled in the 2016/2017 fall semester were graduated on time or not the students who took the courses in the 2018/2019 academic year were determined to be the students who would not graduate in time.

This table includes the course names, gender, age, lecture notes, and the information in the last column about the students whether they can graduate or not. In the study, the attributes

used in the estimation of the students' graduation on time are given in Table 1.

Variable No	Abbreviation	Description
1	Gender	Male- Female
2	Age	18-45
3	Course names	0-100
4	Graduate status	0-1

Evaluation of Classifiers

Sensitivity (SEN), determinism (SPE) and accuracy (ACC) were used to measure the performance of classifier algorithms. These evaluation criteria are derived from the responses of the classifier to the test data. That is, the calculations were held according to the TP (the number of respondents who answered correctly when the recorded answer was correct), TN (the number of responders incorrectly answered when the registered answer was incorrect), FP (the number of respondents answered correctly when the registered answer was incorrect), FN (the registered answer was correct, while Number of respondents are calculated according to the parameters) [16]:

$$SEN = \frac{TP}{TP + FN}$$

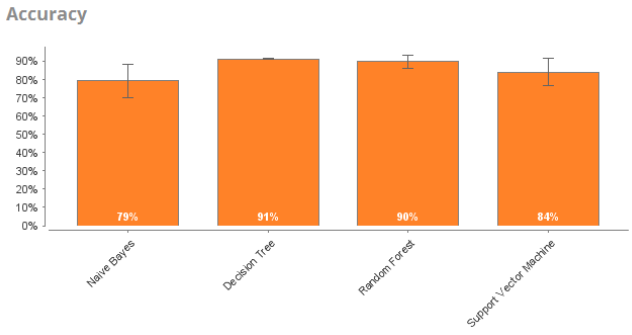
$$SPE = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. CONCLUSION

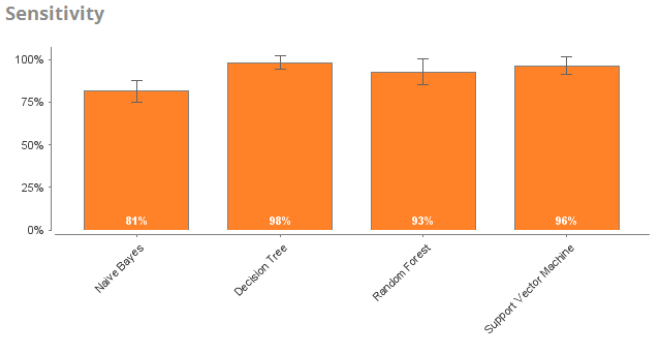
RapidMiner program was used for data analysis. First we have imported the data set to RapidMiner. Naïve bayes, Decision tree, Random Forest and Support Vector Machine algorithms were used for the students' graduation status prediction. The performing principles of these algorithms were explained in the Material and Method section. Accuracy values of 4 different algorithms for 785 student data are given in Figure 1.

Figure 1. Comparission of Different Prediction Model



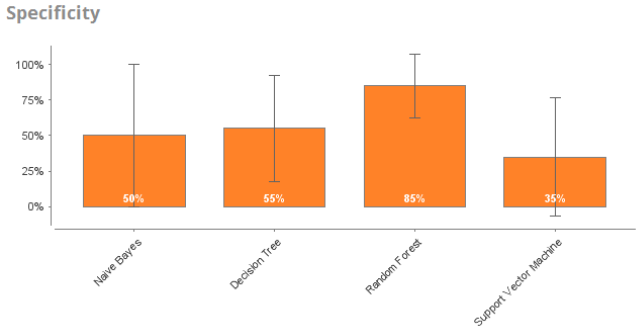
When Figure 1 is examined, it is seen that the best accuracy value is obtained by Decision Tree algorithm with 91%. This algorithm was followed by 90% Random forest, 84% support vector machine and 79% Naïve bayes algorithms respectively. The sensitivity ratio of the analysis is given in Figure 2.

Figure 2: Sensitivity Value



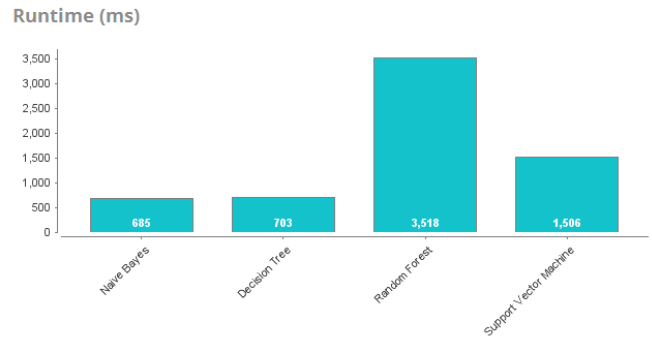
When Figure 2 is examined, it is seen that the highest sensitivity value is obtained by Decision Tree algorithm with 98%. This algorithm was followed by 96% support vector machine, 93% Random forest, and Naïve bayes algorithms with 81%. The specificity value of the study is given in Figure 3.

Figure 3. Specificity Value



According to Figure 3, it is seen that the highest specificity value is obtained by the random forest algorithm with 85%. This algorithm was followed by 55% Decision Tree, 50% Naïve bayes and 35% support vector machine algorithms respectively. Figure 4 shows the runtimes of the algorithms.

Figure 4: Comparative Working Times of Algorithms (Runtime)



According to the analyses held Naïve bayes algorithm with 685 ms has the shortest Runtime period. This is followed by Decision Tree with 703 ms, Support Vector Machine with 1.506ms and Random Forest with 3,518 ms.

In case of failure of distance education students' drop-out of school is more common than formal education students. In this study, graduation status of distance education pre-graduate

students was estimated by machine learning algorithms. Thus, the students who are likely to not graduate in time will be given the necessary warnings by their institutions and the students will be graduated on time and the dropout rate of the students will be reduced.

According to the analysis made, the best estimate and high accuracy rate was gained from decision trees algorithm. In future studies, higher accuracy rates can be obtained by using different algorithms.

ACKNOWLEDGMENT

This paper is derived from Amasya University, Institute of Science, Technology and Innovation Master Student Osman Kayhan's Data Mining and Student Graduation Status Estimation: Amasya University UZEM Example titled thesis.

REFERENCES

[1] G. Silahtaroğlu, *Veri Madenciliği Kavram ve Algoritmaları*, İstanbul: Papatya Yayıncılık 22, 2016.

[2] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği," İÜ İşletme Fakültesi Dergisi, 29(1), 1-22, 2000

[3] J. Luan, "Data mining, knowledge management in higher education, potential applications,", 42nd Associate of Institutional Research International Conference, Toronto, Canada, 2002. Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[4] Y. Z. Ayık, A. Özdemir, and U. Yavuz, "Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi,". Atatürk Üniversitesi Sosyal Bilimler Dergisi, 10 (2), 441-454, 2007.

[5] V. P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C. A. Comes, "Determining students' academic failure profile founded on data mining methods," In Information Technology Interfaces, International Conference, Dubrovnik, 23- 26 Haziran, 317-322, 2008.

[6] Ş. Can, "Veri madenciliği ve eğitim sektöründe bir uygulama," (Yüksek Lisans Tezi). Manisa: Celal Bayar Üniversitesi Sosyal Bilimler Enstitüsü, 2017.

[7] D. Şengür, and A. Tekin, "Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini," Internatıonal Journal Of Informatıcs Technologıes, 6(3): 7-16, 2013.

[8] M. Karabatak, "Özellik seçimi, sınıflama ve öngörü uygulamalarına yönelik birliktelik kuralı çıkarımı ve yazılım geliştirilmesi," F.Ü. Fen Bilimleri Enstitüsü, Doktora Tezi, Elazığ, 116s., 2008.

[9] C. Hark, "Öğrencilerin akıllı tahtaya ilişkin tutumlarının incelenmesine yönelik bir veri madenciliği uygulaması," *(Yüksek Lisans tezi).* Fırat Üniversitesi Eğitim Bilimleri Enstitüsü Bilgisayar ve Öğretim Teknolojileri Eğitimi, Elazığ, 2013.

[10] G. Bilekdemir, "Veri madenciliği tekniklerini kullanarak üretim süresi tahmini ve bir uygulama*," (Yüksek Lisans Tezi*). Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Üretim Yönetimi Ve Endüstri İşletmeleri Yüksek Lisans Tezi, İzmir, 2010.

[11] M. Taşdemir, "Veri madenciliği: öğrenci başarısına etki eden faktörlerin regresyon analizi ile tespiti," (Yüksek Lisans Tezi), 2012.

[12] S. Haykin. "*Neural Networks: A Comprehensive Foundation*," [Elektronik Sürüm], Prentice Hall Inc, New Jersey, 1999.

[13] O. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," [Kurşun deliklerini destek vektör makineleri ile siniflandirma], IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Rewiews, 32(4): 440, 2002

[14] E. Yakut, B. Elmas, and S. Yavuz, "Yapay sinir ağları ve destek vektör makineleri yöntemleriyle borsa endeksi tahmini," Süleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 19(1), 2014

[15] M. Akman, Y. Genç, and H. Ankarali, "Random forests yöntemi ve sağlık alanında bir uygulama," Türkiye Klinikleri Journal of Biostatistics, 3(1), 36-48, 2011

[16] Y. İşler, A. Narin, "WEKA Yazılımında k-Ortalama algoritması kullanılarak konjestif kalp yetmezliği hastalarının teşhisi," SDÜ Teknik Bilimler Dergisi, c. 2, s. 4, ss. 21-29, 2012.