Predicting Academic Failure: Estimating Semester Grade Point Average with Data Mining Methods

Mustafa Yağcı (mustafayagci06@gmail.com)
Ahi Evran University https://orcid.org/0000-0003-2911-3909
Yusuf Ziya Olpak
Ahi Evran Universitesi Egitim Fakultesi

Research Article

Keywords: Data mining, Machine learning, Predict of academic performance, Educational data mining, Early warning systems

Posted Date: July 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-698647/v1

Abstract

This study proposes a new model to analyze the grade point averages (GPAs) in the previous semester using data mining algorithms and to predict the final GPAs that students may receive in the following semesters in three gradually expanding categories (department, faculty, and university). The performances of the Random Forest, Linear Regression, Support Vector Machines, and k-Nearest Neighbors algorithms, which are among the data mining algorithms, were calculated and compared to estimate the GPAs of the students at the end of the semester. This study focused on three parameters. The first was to predict academic performance with a single independent variable. The second was to compare the performance indicators of four algorithms. The third was to compare the predictions made in three different categories. All algorithms applied correctly classified the samples at rates varying between 92% and 94%. The proposed model correctly estimated students' grade point averages at the end of the semester with an average deviation of 0.28 points over a 4 with a single variable. Students with a high risk of failure can be determined in advance by estimating their final grade point averages.

Introduction

Data Mining (DM) is the process of extracting new orientations and new patterns from big data using different classification algorithms (Baker & Yacef, 2009). In other words, it is the discovery of useful information from big data. On the other hand, educational data mining (EDM) "develops and adapts statistical, machine-learning and data-mining methods to study educational data generated basically by students and instructors" (Calvet Liñán & Juan Pérez, 2015, p.100). EDM develops new methods to reveal meaningful sections, original structures, and hidden patterns in the data obtained from educational environments. The main purpose of EDM is to extract information from educational data to support decision-making for educational issues (Calvet Liñán & Juan Pérez, 2015). It involves the collection and interpretation of data produced by learners in order to assess academic performance, predict future performance, and identify current issues, etc. In short, EDM includes the automatic extraction of previously unknown, hidden, meaningful, and beneficial patterns from the huge amount of data.

Various DM algorithms have been successfully applied to identify students at risk of failure in terms of academic performance (e.g., Hu, Lo, & Shih, 2014). Both students and teachers benefit from the knowledge discovered through the use of these algorithms, which serve to improve learning/teaching processes (Akçapınar, Altun, & Aşkar, 2019). In today's education systems, too much data, including demographic data and students' academic grades, are stored in electronic environments. This data can be obtained from various learning management systems (LMS) and student information systems (SIS). The rapid increase in the amount of educational data available can contribute to improving students' learning outcomes (e.g., Shorfuzzaman et al., 2019; Viberg et al., 2018).

EDM provides statistical information to identify the relationship between relevant stakeholders and optimize the learning environment by discovering hidden patterns in the educational environment (Fernandes et al., 2019). For instance, some studies on EDM have compared e-learning systems (e.g.,

Lara et al., 2014), some have classified educational data (e.g., Chakraborty et al., 2016), and others have tried to estimate student performance (e.g., Fernandes et al., 2019). Thus, corrective strategies and pedagogical methods can then be developed by identifying both successful and at-risk students (Casquero et al., 2016; Fidalgo-Blanco et al., 2015).

Ahmad and Shahzadi (2018) developed a model with machine learning methods to identify students who may potentially fail academically. They determined students' learning skills, study habits, and academic interaction characteristics to be independent variables. The model had an 85% success rate. Cruz-Jesus et al. (2020) attempted to predict students' academic performance using their demographic characteristics. The k-nearest neighbors (kNN), logistic regression, random forest (RF), and support vector machines (SVM) algorithms correctly estimated the academic performance of 65% of students. Furthermore, Fernandes et al. (2019) developed a model using the demographic characteristics of the students and their achievement scores for in-term activities. Moreover, Musso et al. (2020) developed a machine learning model based on students' socio-economic characteristics and academic performance.

Waheed et al. (2020) developed a new machine learning model using students' interactions in LMS from a different perspective. According to the researchers, who stated that the model they developed made predictions that were 85% accurate, students who had been browsing previous courses online were more successful. Xu et al. (2019) examined the relationship between the internet usage characteristics and academic performance of university students. The model they developed predicted the performance of the students with a high accuracy rate. Burgos et al. (2018) examined the relationship between the academic performance of the students in past semesters and their academic performance in the following semesters.

In summary, EDM provides an early estimation of the probability of situations such as students dropping out of university or showing decreased interest in a course, analyzes internal factors affecting their performance, and uses statistical techniques to measure students' performance. Various DM methods can be used to predict students' performance, identify slow learners, and identify who is going to leave university (e.g., Hardman, Paucar-Caceres, & Fielding, 2013; Kaur, Singh, & Josan, 2015). In this context, early prediction is a relatively new phenomenon that uses assessment methods to support students by proposing appropriate corrective strategies and policies in this field (Akçapınar et al., 2019; Waheed et al., 2020).

Researchers aiming to predict academic performance and retention have applied a range of techniques, including, neural networks, decision trees, logit, probit, and regression (Nandeshwar, Menzies, & Nelson, 2011). However, most recent studies have adopted RF (e.g., Hung et al., 2020), genetic programming (e.g., Pillay, 2020), and Naïve Bayes (e.g., Sutoyo & Almaarif, 2020) algorithms. When the literature in this area was examined, it was found that a wide variety of variables were used in the studies:

• Various digital traces that students leave on the internet (browsing, time spent watching courses, attendance percentage) (e.g., Fernandes et al., 2019; Waheed et al., 2020; Xu et al., 2019),

• The demographic characteristics of students (gender, age, economic status, number of courses attended, internet access, etc.) (e.g., Aydemir, 2017; Bernacki et al., 2020; Cruz-Jesus et al., 2020; García-González & Skrita, 2019; Rebai, Yahia, & Essid, 2020; Rizvi, Rienties, & Ahmed, 2019),

• Learning skills, study approaches, study habits (e.g., Ahmad & Shahzadi, 2018),

• Learning strategies, perception of social support, motivation, health, academic performance characteristics (e.g., Costa-Mendes et al., 2020; Musso et al., 2020; Kılınç, 2015; Gök, 2017),

• Homework, projects, quizzes (e.g., Kardaş & Güvenir, 2020).

It is observed that the classification accuracy rate varies between 70% and 95% in almost all models developed in such studies. However, the collection and processing of such a variety of data takes a lot of time and requires expert knowledge. Furthermore, Hoffait and Schyns (2017) stated that socio-economic data (e.g., parents' educational level and occupation) is unnecessary and that it is difficult to collect so much data. Besides, these demographic or socio-economic data may not always provide the right ideas for how to prevent failure (Bernacki et al., 2020). In this context, this study aimed to develop a new model that can analyze the grade point averages (GPAs) of the previous semester with data mining methods and predict the final GPAs in the following semesters in three categories (department, faculty, and university). The dataset was divided into these three categories. Thus, the performance of the developed model could be evaluated by the group. For this general purpose, it was determined which algorithms from the DM had the highest performance. This will contribute to the development of pedagogical interventions and new policies that will contribute to students' academic development. In this way, the number of students who have the potential to fail can be reduced with the assessments made at the end of each academic term.

Methods

2.1. Dataset

Educational institutions regularly store data about students electronically. This data can be of a wide variety and volume, from the demographic characteristics of the students to their academic performance. In this study, the data were obtained from the SIS, which holds all the student records of a state university in Turkey. Department of primary education students' records were selected as the dataset for the department category, faculty of education students' records were selected as the dataset for the faculty category, and a total of 5649 records of the students who were enrolled in the fall and spring semester of the 2017–2018 academic year were selected as the dataset for the university category. The dataset was divided into three categories to evaluate the significance and consistency of the performance of the model in different groups. In other words, the dataset was grouped to determine the performance of the model in three categories: Department, faculty, and university as a whole. The distribution of the students by the academic unit is given in Table 1.

The GPA at the end of the fall semester of the 2017–2018 academic year was determined as an independent variable, and the GPA at the end of the spring semester of the 2017–2018 academic year was determined as the dependent variable. The model, developed based on the students' GPA of the students in the fall semester, estimates the GPA for the spring semester. In other words, it was examined at what level the academic performance of the student in the fall semester explained their potential academic performance in the spring semester. There are approximately five months in which to perform any corrective activities for students who may have the potential to fail according to their estimated GPA.

2.2. Data preparation

Data preparation is the process of making the data ready for use. It is the conversion of raw data into processable and noise-free data. For this purpose, a total of 1080 records with incomplete values were deleted (e.g., the records of students who attended the classes in the fall semester but did not attend in the spring semester or who canceled their registration). Furthermore, the dataset obtained from SIS contained the grades for the students' midterm, final, and make-up exams for each course. Each of these grades was recorded as a row. These entries, which consisted of many lines for each student, were grouped on a semester basis and averaged. Then, the midterm, final, and make-up exam grades, which consisted of lines, were transformed into columns and turned into features.

2.3. Applying the algorithms

After the data identification and collection, the development phase of the model was started. For this purpose, DM algorithms were applied. In this context, linear regression (LR), RF, SVM, and kNN were applied to predict students' academic performance, similar to earlier studies (e.g., Akçapınar et al., 2019; Cruz-Jesus et al., 2020; Zabriskie et al., 2019). Thus, students who had the potential to fail and who were likely to drop out from the course/university were identified. In the faculty and university categories, 70% of the data were distributed as training data, 30% as test data, in the department category 95% as training data and 5% as test data. Table 2 shows the distribution of training and test data according to the categories.

Results

The entire experimental stage was carried out with Orange software (Ratra & Gulia, 2020). The data included the 2017–2018 fall and spring semester GPAs of 426 students studying in the department of primary education, 2,379 students studying in the faculty of education, and 5,649 students studying at the university. Since each observation in the dataset could be represented with a sufficient number of training data samples, no dataset imbalance occurred in the pre-processing phase. Fall semester GPAs was the independent variable in the design of the model. The variable to be explained was the spring semester GPAs. Table 3 shows the model variables.

In Table 4, the values in the 2017–2018 Spring column are the actual values. The values in the LR, RF, SVM, and kNN columns are the values estimated by the relevant model. For example, the spring semester

GPA of the student numbered std1 in the department category was 3.05. The predicted values of the LR, RF, SVM, and kNN models were 2.86, 2.97, 2.91, and 2.87, respectively. As can be seen from the first example, the models made accurate predictions with a deviation of about 0.28 points.

DM methods analyze the data measured and predict the results of samples in similar situations. Two types of these methods are regression and classification algorithms. While a regression algorithms continuously predict values, the classification algorithms predict categorical values. As a result, the main difference is that the output variable is numerical (or continuous) for regression and categorical (or discrete) for classification. That is, the independent variable was a continuous variable. Therefore, the accuracy of the prediction results was measured by regression metrics. The estimated values for the GPAs were evaluated using four different metrics (Coefficient of Determination–CoD, Mean Absolute Error–MAE, Mean Squared Error–MSE, and Root Mean Square Error-RMSE) (Botchkarev, 2018; Botchkarev 2019; Willmott & Matsuura, 2005). The higher the accuracy coefficient of a DM model, the closer the estimated values are to the actual values. MSE, RMSE, and MAE values are the error measures of the model. Low values mean that the model shows high performance. In this study, the models' performances were evaluated with MSE, RMSE, MAE, and R^2 metrics. Correlation coefficient (R^2) of 1.00 indicates a perfect positive relationship; -1.00 is a perfectly negative relationship; 0.00 indicates that there is no relationship. If the absolute value of the correlation coefficient is 0.70-1.00 there is a high-level of relationship, 0.70 – 0.30 there is a medium-level relationship, and 0.30-0.00 there is a low-level of relationship (Büyüköztürk, 2008, p.32). Table 5 shows the results of the analysis regarding the estimation of the students' final GPAs.

In the department category, the kNN algorithm gave the highest R^2 (0.775) value. According to this finding, there was a very high-level correlation between the data predicted in the department category and the actual data. Furthermore, the MAE value (0.250), the actual value was correctly predicted with a deviation of 0.250 points up or down. As a result, the kNN algorithm was approximately 94% correct in their classifications of the samples.

In the faculty category, the LR algorithm gave the highest R^2 (0.543) value. According to this finding, there was a very medium-level correlation between the data predicted in the faculty category and the actual data. Moreover, the MAE value (0.296), the actual value was correctly predicted with a deviation of 0.296 points up or down. As a result, the LR algorithm was approximately 93% correct in their classifications of the samples.

In the university category, the LR and SVM algorithms gave the highest R^2 (0.723) value. According to this finding, there was a very high-level correlation between the data predicted in the university category and the actual data. In addition, the MAE value (0.315), the actual value was correctly predicted with a deviation of 0.315 points up or down. As a result, the LR and SVM algorithms were approximately 92% correct in their classifications of the samples.

Discussion And Conclusion

This study proposed a new model based on DM algorithms to identify students who have the potential to fail and who may be likely to drop out the university. This new model analyzes the students' GPAs from the previous semester with DM algorithms and predicts the GPAs they may receive in the following semesters in three categories (department, faculty, and university). The performance of the developed model was evaluated on the basis of these categories. In addition, the performance indicators of four algorithms (LR, RF, SVM, and kNN) were compared. In short, this study focused on three parameters. The first was to predict academic performance with a single independent variable. The second was to compare the performance indicators of four algorithms. The third was to compare the predictions made in three different categories.

With regard to the LR, RF, SVM, and kNN algorithms in the university categories and RF and kNN algorithms in the department categories, there was an high-level correlation between the students' GPAs for the previous semester and their GPAs of the following semester. In addition to the high performance indicators of the algorithms, the fact that predictions were made using only one variable indicates the originality of the study. The findings allow it to be stated that the students' GPAs for the previous semester explained the GPAs they would receive for the following semester at a high-level.

In the faculty category, the LR, RF, SVM, and kNN algorithms demonstrated a medium-level correlation between the students' GPAs for the previous semester and their GPAs for the following semester. Although there was an high-level correlation in the department and university categories, the reason for the medium-level of correlation in the faculty category can be explained by the fact that the high number of departments in the faculty category (eight departments). This is because of the placement scores while the placement score was very high in some of the departments in the faculty of education, in others it was very low.

As of the date of this study, no studies have been found in which the GPAs for the following semester/academic period have been predicted based on the previous semester GPAs using a single variable. Therefore the results of the research were compared with studies that tried to predict students' academic performance based on various demographic and socio-economic variables. Hoffait and Schyns (2017) developed a new model with DM methods to identify students at high risk of academic failure based on their various demographic characteristics. They compared the performance indicators of the Logistic Regression, Artificial Neural Network (ANN), and RF algorithms. They were able to predict students at high risk of academic failure with 90% accuracy. Waheed et al. (2020) used deep learning models to identify students who were at risk of poor academic performance and had the potential to leave the course. They developed a model with a total of 54 student behavioral characteristics in the LMS along with the demographic characteristics of the students. The model had an average of 88% accuracy in making correct classification, and it was claimed that the results obtained will contribute to decision-making processes. Similarly, Xu et al. (2019) examined the relationship between students' internet usage behaviours and academic performance through machine learning methods. In like manner, Bernacki et al. (2020) tried to predict students' academic performance based on the digital traces they left in an LMS. They had a 75% success rate in predicting which students would need to repeat the course. Ahmad and

Shahzadi (2018) determined the relationship between students' study habits, learning skills, academic interaction, and academic performance through machine learning methods. The model they proposed made had a predictive accuracy of 85%. As a result, a high-level of relationship was found between these variables, and they argued that machine learning methods will contribute to the development of educational management. Machine learning methods have thus had very successful results in determining the relationship between students' demographic and socio-economic characteristics and academic performance (Cruz-Jesus et al., 2020; Costa-Mendes et al., 2020). However, the prediction model in all of these studies was established with a large number of independent variables.

The model proposed here accurately estimated the GPAs of the students at the end of the semester with an average deviation of seven points out of a hundred with a single variable. By estimating the final GPAs, students who are at risk of failure or who are at risk of drop out can be identified. So, education and training authorities can be given opportunities to implement corrective actions for these students. Modules that predict academic performance with DM methods can also be added to the LMS. It will thus be possible to make the most accurate predictions automatically and quickly. In short, teaching-learning processes can be managed more effectively and more efficiently thanks to the predictions for academic performance made by DM methods. Timely and targeted individual interventions can be ensured.

In conclusion, although this research uses various predictors, different algorithms, and different approachs to determine the student's GPAs, the results are consistent with previous research and confirm that DM methods can create an effective model for predicting student academic performance. The LR, RF, SVM, and kNN algorithms had high-performance rates. It was also observed that these algorithms can be applied to the categories of department, faculty, and university as a whole. It can be said that such data-driven studies can make very important contributions to decision-making processes. It is, however, also necessary to support students, manage decision-making processes, and develop corrective strategies to ensure students' attendance.

In the present study aimed to analyze the students' GPAs in the previous semester using data mining methods and to predict the final GPAs that students may receive in the following semesters, the data of students at a state university in Turkey. Therefore, students of different education levels can be studied in future research. Furthermore, future studies can be planned by taking into account the various individual differences that affect students' academic performance. Moreover, similar studies can be conducted in different countries. Thus, the situation in different cultures can be compared.

Declarations

Funding

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research, 40*(3), 157–164.
2. Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16. https://doi.org/10.1186/s41239-019-0172-z
3. Aydemir, B. (2017). *Veri madenciliği yöntemleri kullanarak meslek yüksekokulu öğrencilerinin akademik başarı tahmini [Predicting academic success of vocational high school students using data mining methods]* [Master's Thesis]. Pamukkale University, Denizli, Turkey. http://hdl.handle.net/11499/2464
4. Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009 : A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-16. https://doi.org/10.5281/zenodo.3554657
5. Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*. https://doi.org/10.1016/j.compedu.2020.103999
6. Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. Retrieved from *http://www.gsrc.ca/metrics_typology2018.pdf* at 15 February 2021.
7. Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge & Management*, *14*.
8. Burgos, C., Campanario, M. L., De, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance : A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *66*(2018), 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005
9. Büyüköztürk, Ş. (2008). *Sosyal bilimler için veri analizi el kitabı. Ankara: PegemA Yayıncılık* (9th ed., p. 201). Ankara: PegemA.
10. Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *RUSC. Universities and Knowledge Society Journal, 12*(3), 98–112. https://doi.org/10.7238/rusc.v12i3.2515
11. Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: A case study in higher education using learning analytics approach. *Interactive Learning Environments*, *24*(1), 49–67. https://doi.org/10.1080/10494820.2013.817441

12. Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*, *March*, 431–436. https://doi.org/10.1109/ICETECH.2016.7569290
13. Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2020). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*. https://doi.org/10.1007/s10639-020-10316-y
14. Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, *6*(6). https://doi.org/10.1016/j.heliyon.2020.e04081
15. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining : Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012
16. Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using learning analytics to improve teamwork assessment. *Computers in Human Behavior*, *47*, 149–156. https://doi.org/10.1016/j.chb.2014.11.050
17. García-González, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. *Psychology, Society and Education*, *11*(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056
18. Gök, M. (2017). Makine öğrenmesi̇ yöntemleri̇ ile akadem̈ik başarının tahmin edilmesi. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, *5*(3), 139–148.
19. Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science*, *30*(2), 194–203. https://doi.org/10.1002/sres.2130
20. Hoffait, A., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*(2017), 1–11. https://doi.org/10.1016/j.dss.2017.05.003
21. Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469–478. https://doi.org/10.1016/j.chb.2014.04.002
22. Hung, H.-C., Liu, I.-F., Liang, C.-T., & Su, Y.-S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry, 12*(2). https://doi.org/10.3390/sym12020213
23. Kardaş, K., & Güvenir, A. (2020). Kısa sınavların , ödevlerin ve projelerin dönem sonu sınavına olan etkilerinin farklı makine öğrenmesi teknikleri ile araştırılması. *EMO Bilgisayar Dergisi*, *10*(1), 22–29.
24. Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, *57*, 500–508. https://doi.org/10.1016/j.procs.2015.07.372

25. Kılınç, Ç. (2015). *Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi [Examining the effects on university student success by data mining techniques]* [Master's Thesis]. Eskişehir Osmangazi University, Turkey. http://hdl.handle.net/11684/1256
26. Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. *Computers and Education*, *72*, 23–36. https://doi.org/10.1016/j.compedu.2013.10.009
27. Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*, *80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7
28. Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, *38*(12), 14984–14996. https://doi.org/10.1016/j.eswa.2011.05.048
29. Pillay, N. (2020). The impact of genetic programming in education. *Genetic Programming and Evolvable Machines, 21*, 87-97. https://doi.org/10.1007/s10710-019-09362-4
30. Ratra, R., & Gulia, P. (2020). Experimental evaluation of open source data mining tools (WEKA and Orange). *International Journal of Engineering Trends and Technology, 68*(8), 30-35. https://doi.org/10.14445/22315381/IJETT-V68I8P206S
31. Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*. https://doi.org/10.1016/j.seps.2019.06.009
32. Rizvi, S., Rienties, B., & Ahmed, S. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32–47. https://doi.org/10.1016/j.compedu.2019.04.001
33. Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., & Alamri, A. (2019). Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Computers in Human Behavior*, *92*, 578–588. https://doi.org/10.1016/j.chb.2018.07.002
34. Sutoyo, E., & Almaarif, A. (2020). Educational data mining for predicting student graduation using the naïve bayes classifier algorithm. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4*(1), 95-101. https://doi.org/10.29207/resti.v4i1.1502
35. Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027
36. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189
37. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

38. Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173. https://doi.org/10.1016/j.chb.2019.04.015

39. Zabriskie, C., Yang, J., DeVore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research, 15*(2). https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

Tables

Table 1
The distribution of the students by the academic unit

Academic unit		Number of students
Faculty of Economics and Administrative Sciences		1,464
Faculty of Education	Science Education	281
	Social Sciences Education	289
	Turkish Education	261
	Primary Education	426
	Computer Education and Instructional Technology	148
	Elementary Mathematics Education	193
	Psychological Counseling and Guidance	573
	Early Childhood Education	208
Faculty of Arts and Science		1,487
Faculty of Agriculture		319
Total		5,649

Table 2
The distribution of training and test data according to the categories

Academic unit	Training data	Test data	Dataset
Department of Primary Education	405	21	426
Faculty of Education	1,666	713	2,379
University	3,955	1,694	5,649

Table 3

Model variables

Features	Target variable	Meta attributes
2017–2018 Fall Semester	2017–2018 Spring Semester	stdID

Table 4

Probabilities and final decisions of prediction models

	stdId	LR	RF	SVM	kNN	2017–2018 Spring	2017–2018 Fall
Department	std1	2.86	2.97	2.91	2.87	3.05	2.53
	std2	3.01	3.07	3.04	2.95	2.52	2.76
	std3	2.92	2.68	2.97	2.53	2.92	2.63
	std4	3.27	3.17	3.27	3.21	3.43	3.17
	std5	3.32	3.31	3.31	3.23	2.94	3.24
Faculty	std1	2.98	3.01	3.00	3.07	3.27	2.86
	std2	2.44	2.40	2.48	2.41	1.85	2.22
	std3	2.74	2.81	2.77	2.85	2.88	2.58
	std4	3.87	3.65	3.86	3.68	4.00	3.92
	std5	3.40	3.35	3.40	3.40	3.44	3.36
University	std1	2.62	2.54	2.64	2.70	2.84	2.54
	std2	1.61	1.69	1.63	1.58	0.26	1.46
	std3	3.43	3.43	3.45	3.48	3.33	3.40
	std4	1.51	1.34	1.53	1.48	1.38	1.35
	std5	3.88	3.43	3.90	3.71	3.09	3.88

Table 5
LR, RF, SVM, and kNN models' performance criteria

	Model	MSE	RMSE	MAE	R^2
Department	LR	0.154	0.392	0.295	0.665
	RF	0.114	0.338	0.274	0.752
	SVM	0.175	0.418	0.298	0.619
	kNN	0.103	0.321	0.250	0.775
Faculty	LR	0.154	0.393	0.296	0.543
	RF	0.158	0.398	0.300	0.530
	SVM	0.154	0.393	0.295	0.542
	kNN	0.157	0.396	0.299	0.534
University	LR	0.171	0.413	0.315	0.723
	RF	0.178	0.422	0.320	0.712
	SVM	0.171	0.414	0.315	0.723
	kNN	0.176	0.419	0.319	0.715

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- dataset.xlsx