ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)

Available Online at www.ijetcr.org

Volume 11; Issue 01; January-February; 2023; Page No. 36-43

Prediction of Student Academic Performance using ML

Garima Bhatia, Dr. Vishnu Sharma

M.Tech. Scholar in Poornima University, Jaipur

Professor in Poornima University, Jaipur

Conflicts of interest: Nil

Corresponding author: Garima Bhatia

Abstract

It is now possible to forecast student academic success and explore the underlying links in educational data using educational data mining. This study proposes a new model that uses algorithms for machine learning, using the outcomes of their midterm exam scores as the primary data, to estimate university students' final test marks. The predictive capabilities of the supervised classification methods random forests, nearest neighbor, vector support machines, regression models, Naive Bayes, & k-nearest neighbor were calculated and compared in order to forecast the students' final exam marks. The academic performance ratings of 1854 students who registered in Turkish Language-I at a public Turkish college or university for the fall 2019–2020 academic year were included in the dataset. According to the findings, the proposed model has a system of classification that between 70 and 75 percent. Only three distinct argument types were used to make the predictions: faculty data, departmental data, and midterm test grades. These data-driven studies play a critical role in developing a framework for learning analysis in graduate school influencing the decision-making processes. The study also identifies the most efficient machine learning techniques and contributes to the early prognosis of pupils who are likely to fail.

Keywords: Machine learning, Predicting achievement, Educational data mining, Early warning systems, Learning analytics, Early warning systems.

Introduction

Recently, there has been a lot of interest in the use of data mining techniques when it comes to education. Data discovery is done through data mining (DM). It is the study of how to extract new, possibly helpful information or significant findings from large amounts of information (Witten et al., 2011). It also seeks to discover novel patterns and trends from vast data sets by utilizing various categorization techniques (Baker & Inventado, 2014).

Traditional DM techniques are used in EDM (educational data mining) to address issues with schooling (Baker & Yacef, 2009; cited in Fernandes et al., 2019).

EDM refers to the application of DM techniques to information about students' education academic records, exam results, engagement in class and how often questions are asked by students. EDM has grown into a helpful instrument over the past few years for academic performance prediction, uncovering interesting patterns in data sources, and improving the atmosphere for learning and instruction.

Literature

Using DM techniques, Asif et al. (2017) concentrated on two facets of undergraduate students' performance. Predicting students' academic success at the completion of a four-

year programme of study is the first component. The second is to assess students' growth and incorporate the findings into predictions. He split the class into groups with poor and high academic achievement. In order to provide timely warnings, support struggling students, and provide guidance and He has found that it is essential for educators to focus on a small number of courses displaying exceptionally outstanding or bad performance in order to provide opportunities to high-performing students.

Cruz-Jesus et al. (2020) used 16 demographic variables, including age, gender, attendance in class, internet access, ownership of a computer, and how many classes you've taken, to predict student academic achievement. The use of support vector machines, k-nearest neighbors, and random forest were all able to accurately predict students' academic achievement, with accuracies ranging from 50 to 81%.

According to Xu et al. (2019), there is a correlation between university students' internet usage habits and their academic success. To forecast students' success, he applied machine learning techniques. He suggested an approach that was quite good at forecasting students' academic success. The findings indicated that aspects of Internet traffic volume were negatively connected with academic achievement, while features of Internet connection frequency were associated with academic achievement in a favorable way. Also, he came to the conclusion that elements of internet usage had a profound effect on pupils' academic achievement. The question of whether the learning management system's log records alone would be adequate to forecast accomplishment was pursued by Bernacki et al. in 2020. He arrived to the conclusions that the behavior patterns prediction model correctly identified 75% of people who were going to have to repeat a course. He added that this strategy would allow for the identification and support of students who might struggle in the upcoming semesters. For learners that were more probable to fail, Burgos et al. (2018) devised a system that forecasted the achievement grades that the learners may obtain in the next semesters. In comparison to prior years, he discovered a 14% decrease in the number of failing students.

According to the literature evaluation, it is essential to raise educational standards by foreseeing kids' academic performance and helping those who are at danger. Academic achievement was predicted using a variety of factors, including various digital footprints that students had left online (browsing, class duration, percentage of engagement) (Fernandes et al., 2019; Rubin et al., 2010; Waheed et al., 2020; Xu et al., 2019) demographics of students, including gender, age, economic position, number of courses taken, access to the internet, etc. (Bernacki et al., 2020; Rizvi et al., 2019; Garca-González & Skrita, 2019). Almost all of the trials' models had prediction accuracy ranging from 70 to 95%.Yet, gathering and processing such a wide range of data demands both a lot of time and specialized skills. Similar to this, Hoffait and Schyns (2017) argued that gathering so many data points is challenging and that socioeconomic data are superfluous. However, these socioeconomic or demographic statistics might not necessarily provide the best insight into how to avoid failure (Bernacki et al., 2020).

Method

The specifics of the dataset, preprocessing methods, and machine learning algorithms used in this work are covered in this section.

Dataset

All available student data is routinely stored in electronic form by educational institutions. Databases are used to store and process data. This information might range from student demographics to academic accomplishments and can be of many different forms and volumes. The Student Information System (SIS), which houses all student records at a State University in

Turkey, provided the data for this study. The dataset for these records includes the midterm and final exam grades, faculty, and department information for 1854 students who took Turkish Language-I during the 2019–2020 autumn semester. The distribution of students by academic unit is shown in Table 2. Moreover, file 1 (a supplementary file) contains the dataset.

Table 2: The dataset

Academic unit	Number of Students
Faculty of Education	404
Faculty of Arts and Sciences	319
Faculty of Health Sciences	296
Faculty of Economics and Administrative Sciences	221
School of Physical Education and Sports	192
Faculty of Engineering and Architecture	116
School of Physical Therapy and Rehabilitation	92
Faculty of Islamic Sciences	88
Faculty of Agriculture	68
Faculty of Fine Arts	30
Vocational School of Applied Sciences	28
Total Number of Students	1854

Establishing DM model and implementation of algorithm

The following models were used to predict students' academic performance: RF, LR, NN, SVM, KNN and NB. Tenfold cross validation was used to assess the prediction accuracy. The DM procedure has two basic objectives. Making forecasts by examining the database's data is the initial goal (predictive model). The second one describes actions (descriptive model). Predictive models use data whose outcomes are already known to build a model. Then, for datasets with unknown results, the result values are projected using this model. Descriptive models use identified patterns in the current data to guide decision-making.

Models that can accurately predict output values based on existing input data must be created using a statistical approach. Yet when a guided optimization problem is given, machine learning algorithms automatically create a model that matches the inputs with the predicted target values. Indicators from the confusion matrix were used to gauge the model's effectiveness. It is clear from the literature that no single classifier consistently produces the best predictions. Investigating which classifiers are more researched for the analyzed data is therefore important (Asif et al., 2017).

Experiments and results

Orange machine learning software was used for the entire trial phase. Orange is a robust and user-friendly component-based DM programming solution for both seasoned and novice data scientists. Data analysis in Orange is accomplished by stacking widgets into workflows. Every widget has a data retrieval, data preparation, data visualization, modelling, or evaluation job. A workflow is a sequence of steps that will be taken on the platform to complete a certain task. Charts for comprehensive data analysis can be produced by merging several elements in a workflow. The intended workflow diagram is shown in Figure 1.

The Turkish Language-I course was taken by 1854 students during the 2019–2020 Fall Semester. The dataset comprised midterm test scores, final exam grades, faculty, and department information. The Extra file 1 contains the complete dataset. Some of the dataset can be seen in Table 3.

The midterm and final exam grades were classified according to the equal width discretization model after setting the model variables. The criteria used to convert midterm and final grades to categorical format are shown in Table 5.

Table 7 The Confusion matrix

		Predicted	
		Positive (1)	Negative (0)
Actual	Positive (1)	TP	FP
	Negative (0)	FN	TN

Confusion matrix

The confusion matrix displays the dataset's current state as well as the number of accurate and inaccurate model predictions. The confusion matrix is displayed in Table 7. The ratio of successfully categorized examples to wrongly classified instances serves as a measure of the model's performance. The columns reflect the model's estimation, and the rows display the actual sample counts in the test set.

The number of instances that were successfully identified is shown in Table 6 under true positive (TP) and true negative (TN). False positive (FP) displays instances that are anticipated to be 1 (positive) when they should be 0 (negative). False negative (FN) indicates the proportion of cases that are projected to be 0 (negative) but are really in class 1. (Positive).

The confusion matrix for the Radiofrequency method is displayed in Table 8. The main diagonal of the confusion matrix, which has four by four dimensions, displays the proportion of properly predicted cases, whereas the matrix elements other than the main diagonal display the proportion of incorrectly predicted instances.

According to Table 8, 84.9% of students whose final grades actually exceeded 77.5, 71.2% of students whose grades fell between 57.5 and 77.5, 65.4% of students whose grades fell between 32.5 and 55, and 60% of students whose grades were below 32.5 were accurately anticipated. Tables 9, 10, 11, 12, and 13 display the confusion matrices of different approaches.

Classification accuracy (CA) is the proportion of correctly predicted events (TP + TN) to all instances (TP + TN + FP + FN).

Precision:

Precision is defined as the ratio of the number of accurately categorized positive occurrences to the total number of instances that are anticipated to be positive. Obtains a value between [0 .1].

$$Precision = \frac{TP}{TP + FP}$$

Recall:

Recall is defined as the proportion of positive examples that are correctly classified to all cases whose real class is positive. The real positive rate is another name for the recall. Obtains a number in the [0.1] range.

Table 8 Confusion matrix of the RF algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**60%**	3.8%	1.2%	0.6%	38
	32.5–55	26.7%	**65.4%**	9.5%	0.8%	154
	55–77.5	10.0%	30.8%	**71.2%**	13.6%	1016
	≥77.5	3.3%	0.0%	18.1%	**84.9%**	646
	Sum	30	26	1320	478	1854

Table 3 Part of the dataset consist of 1854 rows

stdID	Midterm	Final	Faculty	Department
std1	60	68	Faculty of Economics and Administrative Sciences	Political Science and Public Administration
std2	34	67	School of Physical Education and Sports	Coaching Education

Table 5 Categorical criteria

Category	Criteria
1	grade<32.5
2	32.5<=grade<55
3	55<=grade<77.5
4	grade>=77.5

Table 12 Confusion matrix of the NB algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**40.0%**	9.5%	0.9%	0.0%	38
	32.5–55	18.2%	**42.9%**	9.4%	1.2%	154
	55–77.5	18.2%	42.9%	**70.4%**	19.3%	1016
	≥77.5	23.6%	4.8%	19.2%	**79.5%**	646
	Sum	55	42	1270	487	1854

Table 4 The model of variables

Features	Target variable	Meta Attributes
Midterm	Final	stdID
Faculty		
Department		

Table 9 Confusion matrix of the NN algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**64%**	9.7%	1.2%	0.6%	38
	32.5–55	24%	**61.3%**	9.6%	1.0%	154
	55–77.5	12.0%	25.8%	**71.8%**	14.9%	1016
	≥77.5	0.0%	3.2%	17.4%	**83.5%**	646
	Sum	25	31	1296	502	1854

Table 10 Confusion matrix of the SVM algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**68.8%**	14.3%	1.6%	0.6%	38
	32.5–55	31.2%	**52.4%**	9.9%	0.9%	154
	55–77.5	0.0%	14.3%	**70.1%**	14.3%	1016
	≥77.5	0.0%	19.0%	18.4%	**84.2%**	646
	Sum	16	21	1349	468	1854

Table 11 Confusion matrix of the LR algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**56.0%**	8.3%	1.5%	0.8%	38
	32.5–55	24.0%	**41.7%**	10.3%	1.7%	154
	55–77.5	4.0%	25.0%	**70.0%**	20.1%	1016
	≥77.5	16.0%	25.0%	18.1%	**77.4%**	646
	Sum	25	12	1295	522	1854

Table 13 Confusion matrix of the kNN algorithm

		Predicted				
		<32.5	32.5–55	55–77.5	≥77.5	Sum
Actual	<32.5	**50.0%**	2.6%	1.1%	0.5%	38
	32.5–55	30.0%	**31.3%**	8.9%	1.5%	154
	55–77.5	15.0%	55.7%	**72.9%**	24.9%	1016
	≥77.5	5.0%	10.4%	17.1%	**73.1%**	646
	Sum	40	115	1089	610	1854

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

F-Criterion (F1): Precision and recall have the opposite relationships. For more precise and sensitive findings, the geometric mean of both parameters is calculated. The term for this is the F-criterion.

$$F\text{-}Criterion = \frac{2 \times Duyarlilik \times Kesinlik}{Duyarlilik + Kesinlik}$$

Table 14 AUC, CA, F1, precision and recall values of the models

Model	(AUC)	Classification accuracy (CA)	F1	Precision	Recall
Random Forest	0.860	0.746	0.721	0.752	0.746
Neural Network	0.863	0.746	0.723	0.748	0.746
SVM	0.804	0.735	0.704	0.735	0.735
Logistic Regression	0.826	0.717	0.685	0.700	0.717
Naïve Bayes	0.810	0.713	0.692	0.706	0.713
kNN	0.810	0.699	0.694	0.691	0.699

Discussion and conclusion

This study suggests a new machine learning-based model to forecast undergraduate students' final exam grades using the results of their midterm exam grades as the primary data. To forecast the students' final exam marks, the performances of the machine learning algorithms Random Forests, nearest neighbor, support vector machines, Logistic Regression, Naive Bayes, and k-nearest neighbor were calculated and compared. Two parameters were the main focus of this investigation. Prediction of academic performance based on prior achievement grades was the initial parameter. The second involved contrasting machine learning algorithm performance measures.

The findings indicate that the proposed model had a classification accuracy of between 70 and 75 percent. This conclusion implies that the midterm exam scores of students are a significant predictor to be utilized in forecasting their final exam scores. Algorithms with a very high accuracy rate that can be used to forecast students' final test scores include RF, NN, SVM, LR, NB, and kNN. Additionally, just three types of characteristics were used to make the predictions: faculty data, department data, and midterm test grades. The findings of this study were compared to studies that used various demographic and socioeconomic variables to predict students' academic achievement grades. A model suggested by Hofait and Schyns (2017) makes use of students' prior academic success.

The academic performance of the pupils was predicted by Waheed et al. (2020) based on their regional and demographic features. He discovered that it significantly affects how well students succeed academically. He was 85% accurate in predicting whether the kids would succeed or fail. Internet usage data can distinguish between and predict students' academic achievement, according to Xu et al. (2019). The academic success of pupils was predicted by Costa-Mendes et al. (2020), Cruz-Jesus et al. (2020), and Costa-Mendes et al. (2020) in consideration of socioeconomic data, income, age, occupation, and markers of cultural level. Similar to this, Babi (2017) used artificial neural networks, classification trees, and support vector machine algorithms to predict students' achievement with an accuracy of 65% to 100%.

The accuracy of the proposed model in predicting students' final exam marks was 73%.

As a result, many predictors, algorithms, and techniques were used to forecast pupils' academic performances. The outcomes demonstrate that it is possible to forecast pupils' academic achievement using machine learning algorithms. More crucially, the forecast was limited to the department, professor, and midterm grade. The findings of this study can help teachers identify pupils who are academically motivated above or below average early on. In the future, for instance, as Babi (2017) notes, they can pair students with low academic drive with students with high academic motivation to boost group or project work. This will increase the students' motivation and assure their active engagement in the learning process. Such data-driven investigations ought to aid higher education in developing a framework for learning analytics and assist to decision-making procedures.

References

1. Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artifcial neural network. Bulletin of Education and Research, 40(3), 157–164.
2. Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE Access, 8, 203827–203844. https://doi.org/10.1109/access.2020.3036572
3. Arias Ortiz, E., & Dehon, C. (2013). Roads to success in the Belgian French Community's higher education system: predictors of dropout and degree completion at the Université Libre de Bruxelles. Research in Higher Education, 54(6), 693–723. https://doi.org/10.1007/s11162-013-9290-y
4. Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers and Education, 113, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007
5. Aydemir, B. (2017). Predicting academic success of vocational high school students using data mining methods graduate. [Unpublished master's thesis]. Pamukkale University Institute of Science.
6. Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. Croatian Operational Research Review, 8(2), 443–461. https://doi.org/10.17535/crorr.2017.0028
7. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. Learning analytics (pp. 61–75). Springer.
8. Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.
9. Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158(August), 103999. https://doi.org/10.1016/j.compedu.2020.103999
10. 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
11. Capuano, N., & Toti, D. (2019). Experimentation of a smart learning system for law based on knowledge discovery and cognitive computing. Computers in Human Behavior, 92, 459–467. https://doi.org/10.1016/j.chb.2018.03.034
12. 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
13. Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. In Proceedings of 2nd IEEE international conference on engineering and technology, ICETECH 2016, March (pp.

431–436).
14. 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, 26, 1527–1547.
15. Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artifcial intelligence methods to assess academic achievement in public high schools of a European Union country. Heliyon.
16. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. DecisionSupport Systems, 49(4), 498–506.
17. Delen, D. (2011). Predicting student attrition with data mining methods. Journal of College Student Retention: Research, Theory and Practice, 13(1), 17–35.
18. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research,94(February 2018), 335–343.