Leveraging Educational Data Mining: XGBoost and Random Forest for Predicting Student Achievement

Arash Khosravi[1], Ahmad Azarnik[2]
Faculty of Engineering, Mahallat Institute of Higher Education, Mahallat, Iran
[1]Khosravi.280@gmail.com, [2]ahmad.azarnik@gmail.com

Corresponding author email: ahmad.azarnick@gmail.com

***Abstract*—** Universities and educational institutions are accumulating and storing substantial amounts of data that include the personal and educational information of students. There is an ongoing debate regarding the most crucial factors for predicting students' academic achievement, as well as determining the most suitable algorithm to employ. Furthermore, if these results are achieved, administrators need to develop better planning strategies. Educational Data Mining (EDM) is a technique used to extract specific data types from an educational system, aiding in a comprehensive understanding of students and the system itself. EDM involves transforming raw data obtained from training systems into valuable data that can facilitate data-driven decision-making. In comparison to other fields, the development of data mining and analysis in education has been relatively slow. However, mining educational data on the web presents unique challenges due to specific characteristics of the data. Although various data types possess sequential aspects, the distribution of training data over time exhibits remarkable properties. In this research, we want to find out whether alternative machine learning models, in addition to random forest, can perform comparable or even better in predicting students' academic achievement, therefore, we propose a method that utilizes XGBoost and Random Forest algorithms to identify the significant factors influencing prediction accuracy.

***Keywords*—Academic Performance, Educational Data Mining, Machine Learning, Random Forest, XGBoost.**

1. Introduction

The environment of universities and educational institutions broadly includes three types of items, namely professors, students, and the educational environment. The interaction between these three groups produces extensive data derived from both personal and educational information. Institutions point to the fact that mass data distillation requires a more advanced set of algorithms, which leads to the emergence of educational data mining. Academic data mining is a process in which raw data from educational systems is converted into useful information that can potentially have a greater impact on research and educational practice. Researchers have traditionally used data mining methods such as classification, clustering, rule extraction, communication, and text extraction in academic texts. Academic data mining is an interdisciplinary field of research and a field between education, computer science, statistics, and data mining. The relationship between machine learning, computer-based learning analysis, and training is shown in Figure 1. Humans can vote and check small data, but if the amount of data is large, a person cannot assess all data and extract results, this is where data mining can come in handy, and it can analyze massive volumes of data. Data mining is not only an active and young topic in computer science but also is known in many fields involving academic data mining.

Much has been done in the field of academic data mining, but one of the most important issues in this field is that universities and educational institutions have goals and perspectives for themselves, and in line with these perspectives and goals, they present their educational and research activities. Romero and S. Ventura reviewed articles in the field of academic data mining between 1995 and 2005, found that the use of academic data mining transforms traditional educational institutions into web-based learning management systems and intelligent learning systems [1][2].

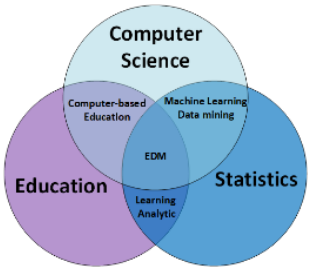

Figure 1 Educational data mining (EDM) [3,4]

Data mining techniques are more widespread in the field of education. In many sectors of higher education, these techniques find the potential impact on the learning process and moving towards new academic courses. Educational data mining and learning analysis are two specific areas used to illustrate the application of data mining. Reports and work with digital data to improve the educational process and academic data mining can shape existing teaching and learning patterns or provide new solutions to problems [5].

In general, the most important benefits of research in the field of academic data mining can be summarized as follows

- To transform traditional educational institutions into web-based learning management systems and intelligent learning systems.
- To study online courses.
- To forecast methods to develop student models.
- To spend money and resources of the university and institute on talented students based on forecasts.

Data used in this research is from a collection of databases of an educational and research institute. This institute is under the

supervision of the Ministry of Science and proposes three types of face-to-face, part-time, and virtual education. It is done directly through the university entrance exam and the assessment organization, including part-time and full-time. In this research, we use face-to-face training data of undergraduate students whose records are from 2010 to 2015.

In this research, we want to find out whether alternative machine learning models, in addition to random forest, can perform comparable or even better in predicting students' academic achievement.

2. Related Works

To determine the subject of academic data mining, many research papers have been published till now. These statistics and information have been extracted from the Google Scholar article search database.

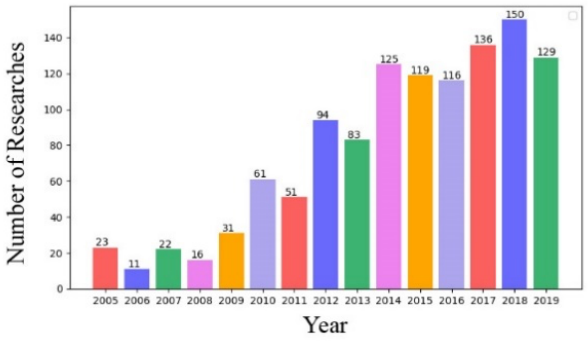

Figure 2 Research in the field of data mining education based on the content

Figure. 2 indicates that during the last decade the number of research papers on Data mining in education has increased. Aldowah et al. [5] stated that there are some data mining techniques that have been taken from various dimensions, and they have divided these dimensions into four main elements:

- Computer-supported learning analytics (CSLA)
- Computer-supported predictive analytics (CSPA)
- Computer-supported behavioral analytics (CSBA)
- Computer-supported visualization analytics (CSVA) [6]

Injadat et al. predict students' final grades, to identify students who may need help later, the data of this research were extracted from the e-learning system. Their database includes records of students who completed their first year of undergraduate studies at the University of Genoa [7]. The objective, which is the total grade point average (GPA), was classified into two categories: good, with scores between 60 and 100, and bad, with scores smaller than or equal to 59. SVM-RBF, NB, K-NN, MLP, and RF algorithms were used for this research. Also, they introduced the top three models in terms of Gini index RF, NB, and k-NN. Asif et al. [8] asked three questions: Question 1; so far, several classifications have been produced in the university to predict students' grades. In all these classifications, only the educational characteristics of the students, such as the grades accepted from the final high school exams are standardized and none of the social, demographic, or economic characteristics are considered. Is this prediction possible only through these educational features? Question 2: Can courses be identified as high-level or low-level as an indicator? Question 3: Is it possible to identify the usual improvements in students' performance and relate them to index courses? To conduct their research, they combined three of the five approaches introduced by [8], including prediction, clustering, relationship mining, discovery within models, and distillation of data for human judgment. The possible answer to the first question shows that only using high school grades of a four-year course for students be predicted. The answer to the second question included four special periods using the decision tree. For the third and fourth questions, students are divided into two main groups: one group of students in the high-performance group and the other group in the low-performance group [8]. Rodrigues et al. added two questions: First, what are the main perspectives and trends in the field of academic data mining for e-learning treatment? And second, what are the potential research topics to consider in assessing e-learning?[9] Slater et al. discuss the importance of familiarity with several tools, then present a toolkit for analyzing data from learning analysis research and academic data mining research [10]. Fernandes [11] provided two sets of characteristics related to education: the first set is characteristics of the students' school and the second set of characteristics of the students themselves. Education and schools of the capital were conducted, the characteristics were school area, shift, classroom environment, student, gender, age, city, neighborhood, grade, absence, and total grade point average. They used the CRISP-DM (The Cross Industry Standard Process for Data Mining) [12] method to perform data mining operations, the operation of which is shown in Figure 3. By performing data mining operations using the Gradient Boost Machine algorithm on the above features, the most important features in 2015 and 2016 were introduced as follows: Student neighborhood, School, Age, City, School district, and Gender. As a result of their research, they stated that the features of the first group, such as neighborhood and suitable school, are more important for students' success. In another research, Maggor et al. stated the preprocessing stages of online educational data and the awareness of researchers and educational policymakers. They also minimise irrelevant data and errors in data analysis were reported by researchers. They organized their surveys into four sequential steps: data collection, data interpretation, database creation, and data organization. In the first stage, they found that the main challenge begins in the data collection stage, the descriptive article begins with the raw data, and in cases where researchers receive the processed data, they need to know the pre-processing stages to evaluate the data. In the second stage, the emphasis is on the need to carefully examine the characteristics of the data, researchers expanded the list of characteristics by the objectives of specific research and available data. The third step is to establish a database emphasizing compliance with EU general data protection regulations. The last step is to organize the data. In this step, the data is filtered and integrated from different sources [13].

Maggor et al. made three main recommendations for addressing the challenges: collaboration, automation, and interpretation

[13]. "Educational institutions often have limited control over the type and format of their data. To solve this problem, the software and policies of administrators must be updated. Successful entry into educational data mining requires the cooperation of people from different departments of the university and educational institutions." Their automated processes are as follows: Automating various technical aspects of data preprocessing can increase the volume of data-driven studies, and automation can also help develop policies to better design education by creating user-friendly and reliable reports. Maggor et al. also commented on the interpretation as follows: interpretation is a major problem in understanding the present data; several variables can be misleading, such as estimating system time consumed or the number of online active users [13]. To address this issue, Martínez-Abad et al. endorsed the method used in previous papers to identify schools with high or low effectiveness [14]. J48, Naïve Bayes, Random tree, and KNN algorithms were used for data mining [15]. In this method, several algorithms are selected from among the algorithms and finally voting on them is done.

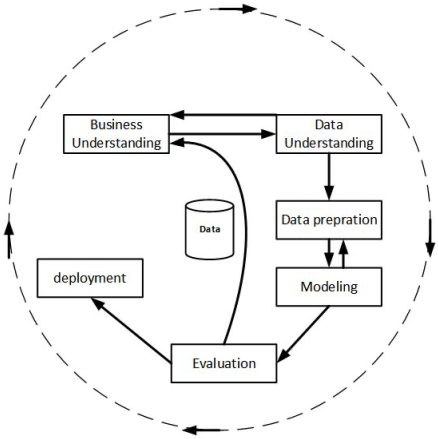

Figure 3 CRISP-DM process [12]

3. Methodology

In this research, a combination of classification methods has been used, and Figure 4 illustrates our general research method. The data of this research have been extracted from several educational records. Extracted data are preprocessed and include several steps, and finally, our data include 89 features. The label considered in this study is TotalAve. The reason for examining this label is to predict students' GPA and to identify the factors that have the greatest impact on students' GPA, or in other words, to identify the most important factors affecting students' academic achievement. This database includes all undergraduate students whose entrance is in the period 2015 to 2019. In the following, we explain methodology more in detail.

3.1. Preprocessing

The following steps are performed to pre-process the data:

- Checking and fixing outdated data.
- Changing the target attribute from a continuous variable to a discrete variable.
- Reduce features using the appropriate algorithm.
- Data normalisation.

Changing the data from continuous to discrete causes the algorithm to face a smaller range of data and thus the complexity of the machine learning model will be reduced. In this regard, the total grade point average has been converted to four A, B, C, D, and E intervals shown in Table 1.

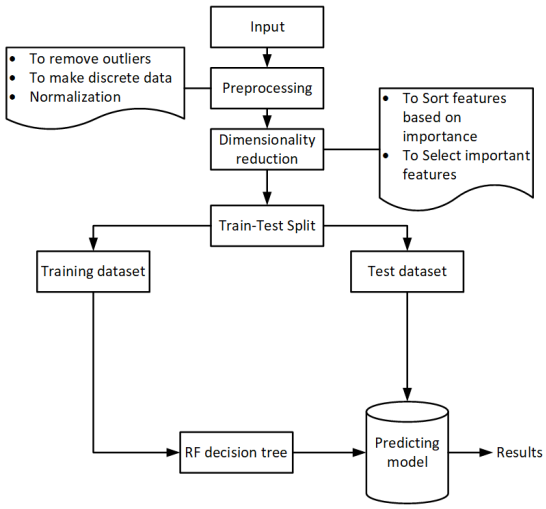

Figure 4 Proposed framework

Table 1intervals

18-20	16-17.99	14-15.99	12-13.99	<12
A	B	C	D	E

3.2. Algorithms

The number of algorithms used in the ensemble method must be odd and greater than the number of tags, so in this case where the number of tags is three, we explain in next section, then the number of algorithms used must be five. The following Machine Learning algorithms are used in this research: CART decision tree, RF decision tree, SVM, Gradient Boosting, and XGBoosting.

3.2. Dimensionality reduction

Due to the large number of features, it is necessary to select important features to perform data mining operations with better models and accuracy, as some features may not have much effect on the results. One of the methods to reduce the dimensions and select the most important features is XGBoost algorithm. XGBoost can be used to reduce the feature noise and reduce the dimensions by increasing the boosting and average gain.

For each attribute, the algorithm calculated the three values of weight, gain, and cover, and then sorted them based on their averages, the results of which are shown in Figure 5. We assume that the minimum acceptable score for the bachelor's degree is 12 and the semester grade point average should not be less than 14, so our tag contains A, B, and C items. We consider accuracy to select the most important properties from 0.0408, so that all properties whose mean value is higher than 0.0408

are pushed into the algorithm and the output of four metrics Accuracy, F1-score, Recall, and Precision are computed. The trend continued until the amount of accuracy increased and then decreased. Therefore, this peak point is selected, and the properties whose mean value is greater than the peak are selected. Results are shown in Appendix I, Table 1. Therefore, the selected features include all the properties whose value is more than 0.1173, the list of these properties is shown in Table 3. The top five features in the tables are Avg5, Avg2, Avg4, Avg3, and Avg1 respectively, and ranged from .6659 to 0.461. Then we use these important features for training step (Figure 5 and Table 2).

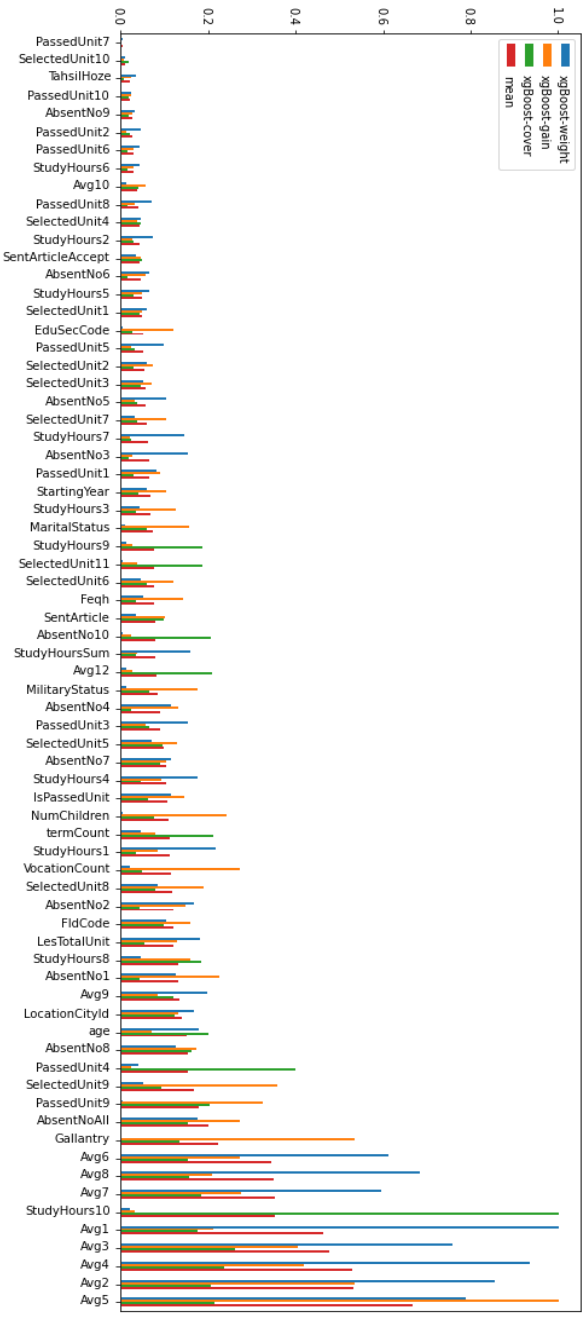

Figure 5 Features' importance

Table 2 The mean of top 22 features

No	Name	Explanation	mean
1	Avg5	GPA of Term 5	0.6659
2	Avg2	GPA of Term 2	0.5297
3	Avg4	GPA of Term 4	0.5272
4	Avg3	GPA of Term 3	0.4731
5	Avg1	GPA of Term 1	0.461
6	StudyHours10	hours of studying in semester 10	0.3495
7	Avg7	GPA of Term 7	0.3492
8	Avg8	GPA of Term 8	0.3475
9	Avg6	GPA of Term 6	0.3431
10	Gallantry	Sacrifice point	0.2215

4. Results and Analysis

In this part, we explain the results of our proposed method and other well-known algorithms and finally, we make a comparison between them (Table 3).

Table 3 The result of different algorithms

Models / Metrics	CART	RF	SVM	Gradient Boosting	XGBoosting	Voting
Accuracy	0.8203	0.9152	0.8508	0.9016	0.8983	0.9118
F1-score	0.8183	0.9058	0.8288	0.8993	0.8957	0.9065
Recall	0.8203	0.9152	0.8508	0.9016	0.8983	0.9118
Precision	0.8192	0.9188	0.8599	0.8998	0.8982	0.9146
Number of Correctly classified	242	270	251	258	265	269
Number of misclassified	53	25	44	37	30	26

The metrics results are shown in the Table 3. Among six models RF has the highest accuracy with a value of 0.9152 compared to SVM which is the lowest at 0.8508. Voting accounts for 0.9065 in the F1-score, which is the highest, while the figure for CART is the lowest. For Recall, as could be predicted, RF is the highest and CART contributes the least result. The highest value of precision is the result of RF by 0.9188, while CART with 0.8192 is the lowest.

Figure 6a shows that, in group A, 111 records are correctly classified, and 14 records are misclassified. Group B shows better results, and 124 records are appropriately recognised. 7 out of 15 C's records are classified correctly and 8 records are assigned to group B. In Figure 6b, in group A, 119 records are correctly classified, and six records are misclassified. 147 records are appropriately recognised for group B. Only four records of C are classified correctly, and 11 records are misclassified. According to Figure 6c, 110 and 141 records are correctly classified in A and B respectively, however, no record is correctly classified in group c. Figure 6d shows that, in group A, 117 records are correctly classified, and eight records are misclassified. Group B shows better results, and 141 records are appropriately recognized. Eight out of 15 C's records are classified correctly, and seven records are assigned to group B. In Figure 6e, in group A, 117 records are correctly classified,

and 8 records are misclassified. Group B shows better results, and 140 records are appropriately recognized. Nearly half of C's records are classified correctly and the rest are misclassified. Regarding to Figure 6f, for A, 118 records are correctly classified, and 5 records are misclassified. Group B shows better results, and 145 records are appropriately recognized. In C, only 6 records are classified correctly.

By considering Table 1, these results state that the number of correctly classified records in class C is low. It may stem from the number of available records with label C, and we could not argue that these algorithms perform poorly, because the results of A and B are satisfactory. Of the five algorithms used, the RF method based on all criteria, has more accuracy. The order of accuracy of the algorithms used in general is as follows: 1) RF method, 2)Voting, 3)Gradient Boosting algorithm, 4)XGBoosting algorithm, 5) SVM, and 6) CART.

By identifying important and influential features in students' academic achievement, decision-makers in university settings can, by trying to remove obstacles, pave the way for their students' progress and ultimately their university progress. To achieve better results, these items should be considered.

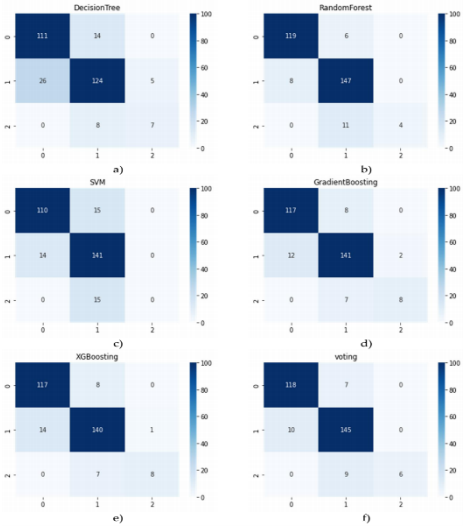

Figure 6 Confusion Matrices. a) CART, b) RF, c) SVM, d) GradientBoosting, e) XGBoosting, f) Voting

5. Discussion

In this research we investigated the effectiveness of various machine learning algorithms for predicting student academic achievement. Here's a breakdown of the key findings and their implications:

5.1. Model Performance

Random Forest (RF) achieved the highest overall accuracy (0.9152), F1-score (0.9065), recall, and precision compared to other models like SVM, Voting, Gradient Boosting, XGBoosting, and CART. While all algorithms performed well for groups A and B, Group C exhibited lower accuracy. This suggests a potential class imbalance issue, where there might be fewer data points for Class C compared to A and B. Considering this class imbalance, the overall performance of the models cannot be solely judged based on Class C results.

5.2. Feature Importance

The study acknowledges that the research did not consider identifying the most important features influencing student achievement.

5.3. Implications and Future Work

RF emerged as the most effective model for predicting academic achievement in this study. However, exploring other algorithms or ensemble methods like combining RF with XGBoosting could be valuable for further improvement.

Addressing class imbalance through techniques like oversampling or under-sampling the majority class might enhance the model's ability to predict Class C outcomes more accurately. Moreover, identifying the most impactful features through techniques like feature selection or feature importance analysis could provide valuable insights into factors influencing student success. This knowledge can be used by universities to develop targeted interventions and support systems for students.

Overall, the study demonstrates the potential of machine learning for predicting student academic achievement. Future research should focus on addressing class imbalance, identifying key features, and exploring more sophisticated models or ensemble methods to achieve even better prediction accuracy and actionable insights.

6. CONCLUSION

Educational data mining (EDM) is a field of research related to the application of data mining, machine learning, and statistics to information generated in educational settings (e.g., universities and intelligent educational systems). At a high level, the discipline seeks to develop and improve methods for exploring this data, which often has multiple levels of meaningful hierarchies, to uncover new insights into how people learn in such settings. We propose a method to predict academic achievements by XGBoost and Random Forest. Then, we use the pre-processed data to train random forest. In the end, a comparison, including accuracy, recall, F1-score, and precision, indicates that our model outperforms other algorithms.

References

[1] S. M. Dol and P. M. Jawandhiya, "Classification technique and its combination with clustering and association rule mining in educational data mining—A survey," *Eng. Appl. Artif. Intell.*, vol. 122, p. 106071, 2023.

[2] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 905–971, 2023.

[3] N. Bošnjaković and I. \DJur\djević Babić, "Systematic review on educational data mining in educational gamification," *Technol. Knowl. Learn.*, pp. 1–18, 2023.

[4] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2013, doi: 10.1002/widm.1075.

[5] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*. 2019. doi: 10.1016/j.tele.2019.01.007.

[6] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, 2007, doi: 10.1016/j.eswa.2006.04.005.

[7] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Syst.*, 2020, doi: 10.1016/j.knosys.2020.105992.

[8] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, 2017, doi: 10.1016/j.compedu.2017.05.007.

[9] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational Data Mining: A review of evaluation process in the e-learning," *Telematics and Informatics*. 2018. doi: 10.1016/j.tele.2018.04.015.

[10] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review," *J. Educ. Behav. Stat.*, 2017, doi: 10.3102/1076998616666808.

[11] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, 2019, doi: 10.1016/j.jbusres.2018.02.012.

[12] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step," *ASHA Present.*, 2000.

[13] Y. Feldman-Maggor, S. Barhoom, R. Blonder, and I. Tuvi-Arad, "Behind the scenes of educational data mining," *Educ. Inf. Technol.*, 2021, doi: 10.1007/s10639-020-10309-x.

[14] F. Martínez-Abad, A. Gamazo, and M. J. Rodríguez-Conde, "Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment," *Stud. Educ. Eval.*, 2020, doi: 10.1016/j.stueduc.2020.100875.

[15] M. Ashraf, M. Zaman, and M. Ahmed, "An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches," in *Procedia Computer Science*, 2020. doi: 10.1016/j.procs.2020.03.358.

Appendix I

Table 1. Score of ten important features

Score of feature	0.0408	0.0542	0.0666	0.0747	0.082	0.1007	0.1104	0.1173	0.1196	0.1296
Number of records	61	52	45	39	35	30	26	22	20	19
Accuracy	0.8915	0.8983	0.9050	0.9016	0.9186	0.9118	0.9220	0.9254	0.9254	0.9118
F1-score	0.8836	0.8879	0.9067	0.8893	0.9146	0.9027	0.9186	0.9149	0.9241	0.9067
Recall	0.8915	0.8983	0.9050	0.9016	0.9186	0.9118	0.9220	0.9254	0.9254	0.9118
Precision	0.8930	0.9039	0.9093	0.8980	0.9162	0.9044	0.9188	0.9334	0.9276	0.9030