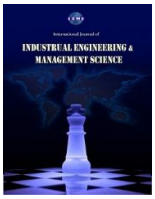
University of Hormozgan

International Journal of Industrial Engineering & Management Science

journal homepage: www.ijiems.com



A hybrid of random forest and SVM for predicting student performance

Sima Sayadi [a], Mohammad Kazem Sayadi [b,*]

[a] *ICT Research Institute, Tehran, Iran*
[b] *Meybod University, Yazd, Iran*

ARTICLE INFO

Article history:
Received: 2023-02-02
Received in revised form: 2023-03-03
Accepted: 2023-03-04

Keywords:
Educational data mining
Machine learning
Students' failure
Students' progress
Dropout rate

ABSTRACT

Identifying at risk students is a crucial task in universities to decrease the dropout rate. This causes optimal use of resources and facilitates the decision making. So improving models for identifying significant factors would be beneficial to the students, university and society. The present work proposes educational data mining techniques and identifies methods which can help increasing the algorithms' performances. In this research, using random forest, with 0.94 accuracy, total grade point average of previous semesters, admission quota and number of conditional semesters (failed courses) were identified as the most important factors effect on students' failure or success. Comparing the performance of SVM, LR, LDA, CART, KNN, NB algorithms using the most important features as input, showed that SVM classifier acts best between them and reducing the variables causes developing the algorithms' performance. Also tuning the parameters on python program helped getting a desirable result for predicting at risk students.

1. Introduction

Warning students about their poor performance would be generative and constructive if the university officials seek solutions to help them cope with the complicated circumstances and get rid of continuing their life's complex conditions. So, they have to detect the main reasons which cause the low performance of the students. Mining student's data such as historical grades, demographical features and student's characteristics is a popular method which can help them specifying the factors which cause the unpleasant outcome. It is called Educational Data Mining (EDM) (Ingale et al., 2021). Astin (1971) have recognized the dropout epithet as a misuse of explaining the student's condition in the literature when the student doesn't yield and performs inefficiency. Withdrawal is another terms used for abandoning education before completion (Agrusti, 2019).

Reducing the number of features in data set increases the accuracy and reduces the obtained model's complexity (Ramaswami & Bhaskaran, 2009). Feature selection is the way to achieve this goal. This prevents over fitting of the model by identifying irrelevant (noisy) features and ranks the most important features (Sokhey, 2019).

The aim of this research is to discover the reasons make the University students make full use of their capacity and capability and help reaching their competency level. To fulfill this target, the data mining random forest and SVM methods are being used. Using random forest, the most important features are identified. Also performance of algorithms will be assessed and compared using the most important features and the whole data set.

This research is the first research being done in a university to identify the reason of students drop out and to help them and the university to reach a high level of quality in education and also to identify the best performance methods for mining the university students' data. The analysis of this research is done using machine learning packages such as scikit-learn and pandas.

* Corresponding author.
E-mail address: MK.Sayadi@itrc.ac.ir

2. Theoretical foundations and Literature review

Theoretical foundations:

Before starting to review the literature, it is needed to define/explain some terms:

Data Mining (DM): DM is an automatic or semi-automatic process which discovers regular patterns from raw data. Predicting, clustering, classification, association and estimation tasks, helps in making decisions and solving organizations' problems (Stock, 2021).

Machine learning: Machine learning teaches a computer how to learn and comprehend the given data. The learning methods can be supervised or unsupervised method (Kaufman, 2012):

- Supervised learning: DM models which have a variable as label of classification, such as decision tree or support vector machine. So it fundamentally is another name for classification.
- Unsupervised learning: DM models which haven't any label. It is another name for clustering. Regression is an unsupervised learning method.

KNN (K-Nearest Neighbors): makes predictions using some function, such as average of response values of the similar observations (Myatt & Johnson, 2006).

Support vector machine (SVM): is a supervised learning method which is used for both classification and regression tasks. But naturally is improved for classifying numeric data into two classes (Huang et al., 2006).

Ensemble method: a method which increases overall accuracy by learning and combining a series of individual (base) classifier models. Bagging, boosting, and random forests are popular ensemble methods (Kaufman, 2012).

Random forest classifier: an ensemble method which potentially helps in increasing accuracy and avoiding over fitting (Yang, 2019). It produces a large number of decision trees following the Classification and Regression Trees (CART) methodology. In each tree, the training samples are randomly selected with replacement and for a set of M features, in every tree of the forest, a subset of m (m is less than M) attributes are chosen and on each node the best variable is chosen for splitting. Accepted rules obtain from the majority votes gained from ensemble trees. Random forests can give an estimate of ranking the variable importance (Kaufman, 2012).

Linear Discriminant Analysis (LDA): LDA is a classification machine learning algorithm comes from the British statistician, Fisher. It works by calculating summary statistics for the input features by class label, such as the mean and standard deviation by the maximization of the gap between classes and within-class variances minimization.

Logistic Regression (LR): LR fits a line to a dataset using a defined threshold or after finding equilibrium of sensitivity and specificity. Then it returns the probability that a new sample belongs to one of the two classes according to its location with respect to the line. So in this algorithm the target variable must be binary (Myatt & Johnson, 2006).

Naive Bayse (NB): The NB algorithm's function is to assess the probability of all estimated models and find the maximum probability to assign it to the new input. This method is more suitable for large data sets. Also samples are required to be conditionally independent. It follows the Bayes' theorem which has its own formula. This method is popular for WEB text (Myatt & Johnson, 2006).

Features: In the context of our paper, student's characteristics or specifications, for example their school marks, employment, marital status and their city of residence, are definitions/terms called as variables, features and attributes respectively by statisticians, data miners and computer scientists.

Feature selection: feature selection, a preprocessing technique which eliminates the irrelevant features, is a process consisting of 4 steps; at first a feature subset is selected using a certain strategy. In the second step, this subset is evaluated and compared with another unselected features. If any feature is evaluated as a better property, one of the most important ones will be omitted and it takes its place. This process is repeated until a predefined stopping criterion is satisfied (Ramaswami & Bhaskaran, 2009). Feature selection and parameter optimization (tuning parameters according to the accuracy of test set) improve the prediction accuracy. Using both of them always doesn't give a good result. Feature selection gives better results than other (Damopolii et al., 2021, Duch et al., 2003).

Accuracy: is the ratio of well-classified elements to the sum of all elements (Lamrini, 2020).

Literature review

Damopolii et al. (2021) conducted a study to predict the students of Indonesia's public junior high schools' grades using 30 variables such as entrance grades, gender, religion, parents' education and occupation, etc. with 168 records of student data. They used decision tree, K-Nearest Neighbor and Naive Bayes classifiers without applying feature selection and parameter optimization and compared the accuracy results after using these two methods. In their paper, the best accuracy was achieved by K-Nearest Neighbor with feature selection and parameter optimization.

Sivakumar (2016) developed the improved decision tree to predict students' dropout. They realized that family problem, home sickness, campus environment, adjustment problem in hostel, participate in extra curriculum activity, and residence are the most important reasons of students' dropout.

Alturki et al. (2021) used C4.5, Simple CART, LAD Tree, Naïve Bayes and Bayes Net with AD Tree, and Random Forest methods to predict students' academic achievement. They collected 300 students' records and compared the performance of the algorithms. Random Forest performed the best compared to the tree-based models. GPA of previous semesters and previous GPA from secondary school identified as the most effective factors on students' performance. The results are similar with the Osmanbegović & Suljic (2012) and Pal & Pal (2013) in which the excellent grades in high school showed a positive relationship with the grades in university.

Phauk & Okazaki (2021) conducted a study aimed to develop the accuracy of K-Nearest Neighbor, Hybrid C5.0 and Hybrid RF, Improved Deep Belief Network (IDBN) method and integrated the models into a web-based system for predicting the high school students' performance. Data were consisted of Parents' educational levels, Parents' occupational status, class environment, curriculum, students' interest and motivation and etc. The Hybrid C5.0 and Hybrid RF comparatively performed better than other models. They used 5 feature selection methods

and compared the gained accuracy and RMSE (Root mean squared error).

Jain et al. (2021) implemented Apriori data mining techniques to analyze and predict the correlation between science subjects' grades and the students' gender. They found strong relationships between subjects as well as subjects with gender.

Altabrawee et al. (2019) conducted a study of 161 student records with 20 attributes categorizing to personal and life style, studying style, family related, educational environment satisfaction, and student's grades. Computer Grades-Course, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency, which is a life style attribute, were found as the most predictive factors. Applying Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, Naïve Bayes model had the lowest ROC index equal to 0.697 and the lowest accuracy equal to 66.52. Artificial Neural Network accuracy was 77.04 and ROC index was 0.807.

Saa (2016) conducted a research in Ajman University of Science and Technology. A data set with 270 records and attributes such as student's gender, Living Location, Total Family, Monthly Income, Previous Semester GPA, etc. The applied data mining models were Naïve Bayes, C4.5, ID3, CART, CHAID Decision Tree methods that CART had the best accuracy of 40%.

3. Research Methodology

The research methodology consists of 3 sub-sections. In the first part, the questionnaire and the statistical population are explained, and in the second part, data analysis methods are discussed. In part 3, data mining methods are detailed.

3.1. Questionnaire and statistical population

By using literature, the most important independent features is specified. These features are shown in second column of Table 1. The dependent variable, X [26], previous semester's total GPA, between "0 to 20", is continuous and nominal class variable should be used in classification models. So, after discretizing the continuous variable, the grades "10 to 18" are assigned to class A and "18 to 20" is assigned to class B.

In this research the statistical population consists of all the second semester of 2020 – 2021 students, except the freshmen. To determine the sample size, the Cochran's formula was used with the following assumption:

$$n = \frac{NZ^2_{\alpha/2}P(1-P)}{\varepsilon^2(N-1) + Z^2_{\alpha/2}P(1-P)}$$

Equation1. Cochran's formula for determining the sample size

By setting 95% (confidence level), $P = 0.5$, $\alpha = 0.05$ (error level), $\varepsilon = 0.085$ (maximum estimation error), $N = 1511$ (population size), the number of samples gained by Cochran's formula is 122. But after distributing 130 questionnaires and collecting them, 128 of them were filled correctly and used in this research. Discretizing continuous variables is shown in table 1.

Table 1. Features and the way of encoding and decoding

Numbering	Feature	Encoding
X[0]	City of residence	Native=0, Non-native=1
X[1]	Father's education	Illiterate = 0, elementary = 1, middle school = 2, high school = 3, university = 4
X[2]	Economic status	poor = 0, medium = 1, good = 2
X[3]	Distance from residence to university	0-50 km = 0, 50-200 km = 1 , more than 200 km = 2
X[4]	Employment	Unemployed=0, Employed=1
X[5]	Dependence on family	Low = 0, medium = 1, high = 2
X[6]	Being idealistic and ignoring existing facts	Low = 0, medium = 1, high = 2
X[7]	Parental discord and family instability	Yes=1, No=0
X[8]	Disputes with parents or other family members	Yes=1, No=0
X[9]	Admission quota	Special quota = 0, region 1 = 1, region 2 = 2, region 3 = 3
X[10]	Total average of diploma	0-14.99=0, 15-17.99=1, 18-20=2
X[11]	Total grade point average of previous semesters	0-14.99=0, 15-17.99=1, 18-20=2
X[12]	Number of conditional semesters	0, 1, 2
X[13]	Number of semesters of participation in extracurricular activities	0,1,…,8
X[14]	Appropriateness of consulting services	Poor = 0, medium = 1, excellent = 2
X[15]	Support and educational, research, welfare facilities	Poor = 0, medium = 1, excellent = 2
X[16]	Engage with fringe issues and associations	Yes=1, No=0
X[17]	How to choose the field at the entrance exam	acceptance in the field=0, family pressure=1, Motivation=2
X[18]	Interest in the field	Low = 0, medium = 1, high = 2
X[19]	Educational quality and curriculum of the faculty	Poor = 0, medium = 1, excellent = 2
X[20]	The classroom environment	Poor = 0, medium = 1, excellent = 2
X[21]	Educational rules and curriculum of the university	Poor = 0, medium = 1, excellent = 2
X[22]	Communication with classmates	Low = 0, medium = 1, high = 2
X[23]	Evaluate the students ' performance by professors	Poor = 0, medium = 1, excellent = 2
X[24]	University's Physical Environment	Poor = 0, medium = 1, excellent = 2
X[25]	Field of study	Humanities = 0, Engineering and Technology = 1, Theology and Quran Sciences = 2

3.2. Data analysis method

After collecting the data and saving them in excel file and preprocessing, it was ready for splitting into test and train data.

- **Data preprocessing**

After removing duplicates and filling the blanks and discretizing, the data was turned into the nominal so as to be practical for python language algorithm.

- **Determining the volume of train and test data**

20% percentage of data was given to the algorithm as test set in order to examine quality of it.

- **Determining parameters to prevent over fitting**

For each model the algorithm's parameters are adjusted to the optimal amount. For SVM algorithm the parameters C and gamma, were tuned into the best values.

3.3. Data mining methods

In this paper random forest algorithm as a feature selection method was used for ranking the features. Ensemble models usually are more accurate than the individual models which are made by them [22].

After that we made a SVM model for categorizing the data. Since there is no best method of data mining for all the settings [14], and to receive a good result from models, used variables and DM methods are important [22], so we calculated the accuracy metrics of a number of the other data mining methods (LR, LDA, KNN, KART, NB, SVM) before and after reducing the data, and compared them. The process is shown in figure 1.

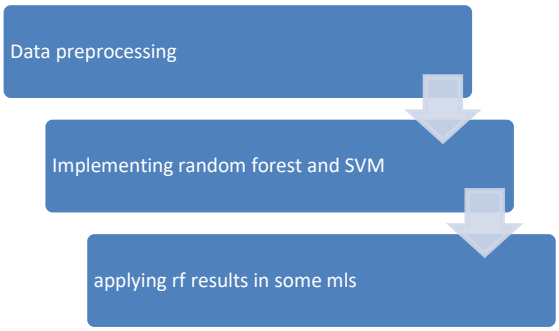


Figure 1. The process which this article follows

4. Implementing data mining methods

4.1. Random forest

There are two ways of categorizing feature selection methods. First: 1. univariate, 2. multivariate. Second: 1. filter method, 2. wrapper method which uses the classifier to evaluate the score of features or feature subsets, 3. embedded method in which feature selection is done during the training of classifier such as random forest (Ramaswami & Bhaskaran, 2009). This method can give an estimate of ranking the variable importance by creating a large number of decision trees. Then important features are used in another method to compare the performance.

Also identifying students at risk of drop out is more crucial. As this class has fewer samples in data set, there is a class imbalance problem. To handle this problem, there are some tips like oversampling, under sampling, threshold moving, and ensemble techniques (Kaufman, 2012). So, this is another advantages of random forest.

Using random forest here, the accuracy for this method is 94.26% which considers as a good performance.

As we can see from the feature importance gained from random forest, shown in table 2, X [11] (Total grade point average of previous semesters) [14], admission quota (X [9]) and number of conditional semesters (failed courses) [23] (X [12]) are found among the most important variables.

Table 2. Importance of the features

Variable	Importance	Variable	Importance
11	0.78	9	0.03
12	0.02	1	0.01
2	0.01	3	0.01
5	0.01	6	0.01
8	0.01	10	0.01
13	0.01	14	0.01
15	0.01	17	0.01
18	0.01	19	0.01
21	0.01	22	0.01
23	0.01	24	0.01
25	0.01	0	0.0
4	0.0	7	0.0
16	0.0	20	0.0

Fig 2 shows the scatter-matrix of all variables and the 3 most important ones and the two-case target variable. This matrix is used when there are more than two variables and is used for showing the relation between two variables. Using python programming tools, each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart (in one and two dimensions) [24].

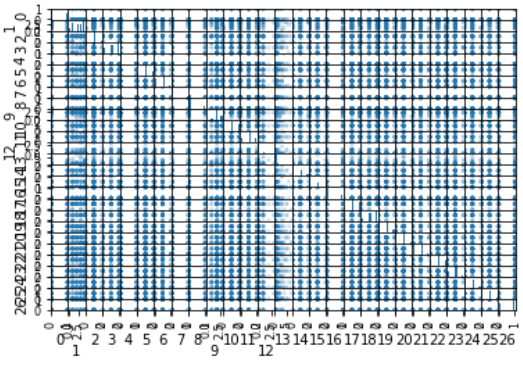


Figure ۲. Plot matrix (pairs/correlation plot) of all attributes

4.2. SVM

Using all 26 features and 20% of the data set for testing and other 80% for training, the SVM model (Fig 3) was made. The figure visualizes the support vectors, which are the border points in categorizing, with overall accuracy of 0.73 and finds a line separating the educational data shown in Table 3. The second target variable (1), which is the low performance, is more important. Because specifying the low performance students, for taking necessary actions, is more essential. As the first target variable (0), upper level students, has more samples, more accurate results are gained and the overall accuracy has increased. This model visualizes a categorization of data with 0.73 accuracy which is an optimized accuracy.

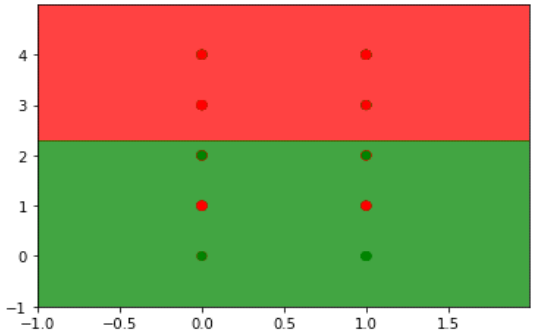


Figure ۳. The SVM model categorized the data into 2 target classes

Table 3 shows the different standards for measuring the performance of model. Precision is criterion for measuring the percent of ''yes'' answers (dropped out students) which are recognized truly. Recall shows the delicacy of algorithm for specifying dropped out students (yes answers).

Table ۳. Model Performance

	precision	f1-score	recall
0	0.57	0.50	0.53
1	0.79	0.83	0.81
accuracy	0.73		

In the next section, some other DM algorithms with different size of features will be implemented to compare their performance.

5. Implementing other DM methods and Comparing the Results

After applying some machine learning models, LR, LDA, CART, KNN, NB, and SVM on our data the accuracies in Table 4 have acquired. As we see the Accuracy of the SVM method is more than other ones.

Table 4. The results of implementing some other models

Method	Accuracy (26 attributes)	Accuracy (3 attributes)
LR	0.706364	0.754545
LDA	0.696364	0.754545
KNN	0.639091	0.704545
CART	0.674545	0.723636
NB	0.609091	0.763636
SVM	0.715455	0.763636

Fig 4 demonstrates the diagram of gained accuracy using the main data set as income. By this diagram, models' accuracies can be compared easily.

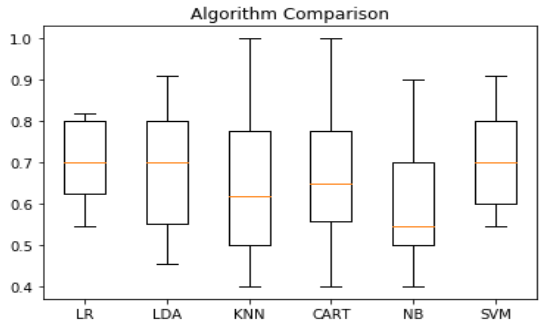


Figure 4. Accuracy of different methods with main data

Fig 5 shows the diagram of gained accuracy using the outcome of random forest (first 3 attributes) in hybrid of other methods.

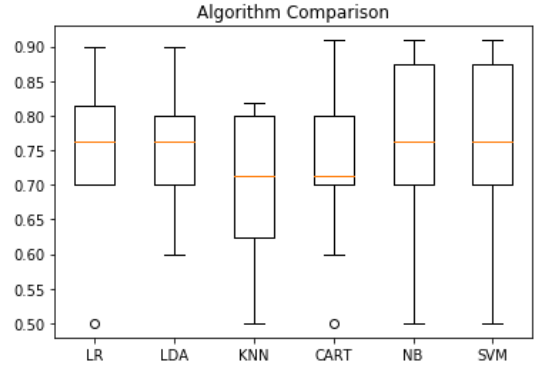


Figure 5. The accuracy with 3 attributes

As we can see, after applying the reduced data set (3 most important features), the accuracy has significantly increased.

A model with a higher accuracy (ACC), means a model with better performance [14] and in contrast to ACC, the smaller the value of RMSE, the model predicts better [18]. So in this research, without feature selection but using parameter setting, SVM method worked better, after that LR, LDA, CART, KNN, and NB are less well-performed respectively.

But after applying only the 3 most important features, NB with a significant improvement (0.16) and SVM with 0.05 improvement, were in the same place of performance (0.76), after that the accuracy of LR and LDA, with about 0.05 reached at 0.75, and CART with a 0.72 accuracy (0.05 improvement) performed better than KNN with 0.70 accuracy (0.07 improvement). Comparing all the models, the random forest giving accuracy equal to 0.94 performed better and then SVM with reduced data have performed better.

6. Conclusion

The main objectives of this study was to evaluating the performance of random forest and SVM algorithms and the effects of applying random forest's outcomes on DM algorithms and to identify factors that influence students' performance and helping the students to enhance their ability, skills and quality of life. Ensemble random forest showed that among 26 features, 4 of them with importance = 0 are irrelevant, 21 of them with importance = 0.03 are relevant and 3 of them are the most relevant features which were used in SVM method and another DM methods. The best accuracy was achieved by random forest, 0.94. After that, SVM gave the best accuracy, 0.77, among all the algorithms. The features identified by random forest, as important features, were admission quota and number of conditional semesters (failed courses) which means that the students with special quota have more chance for getting better grades in university and also students should try more to study their lessons so as not to fail or drop out. Also students can refer to the university counselor to find their problems and obstacles which cause their low performance [25]. So university counselor can play a specific role in students' prosperity and success. The most important variable identified by the random forest was GPA which was expected to be chosen. So in future researches it is suggested to omit this variable from main data, so that another important factors will be selected and the quality and usefulness of research will increase. University managers should prepare facilities by getting help from charities or allocating more budgets for buildings and helping the students to enjoy their free time. It helps the students to become more energetic and it can bring spiritual and physical health.

References

[1] Ingale, Nilesh V. "Survey on prediction system for student academic performance using educational data mining." Turkish Journal of Computer and Mathematics Education (TURCOMAT) vol. 12, no. 13, 363-369, 2021.

[2] Astin, A. W. "Predicting academic performance in college: Selectivity data for 2300 American colleges," Free Press, 1971.

[3] Agrusti F, Bonavolontà G, Mezzini M. "University dropout prediction through educational data mining techniques: A systematic review," Journal of e-learning and knowledge society, vol. 15, no. 3, 161-82, 2019.

[4] Ramaswami, M. and Bhaskaran, R.,"A study on feature selection techniques in educational data mining," arXiv preprint arXiv:0912.3924, 2009.

[5] Sokkhey P, Navy S, Tong L, Okazaki T., "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia". IEIE Transactions on Smart Processing and Computing. vol. 9, no. 3, 217-219, 2020.

[6] Mining, W. I. D., "Data mining: Concepts and techniques". Morgan Kaufinann, vol. 10, 559-569, 2006.

[7] Cheng, K. C., Huang, M. J., Fu, C. K., Wang, K. H., Wang, H. M., & Lin, L. H., "Establishing a multiple-criteria decision-making model for stock investment decisions using data mining techniques". Sustainability, vol. 13, no. 6, 3100, 2021.

[8] Myatt, G. J., & Johnson, W. P., "Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications," John Wiley & Sons, 2009.

[9] Huang, T. M., Kecman, V., & Kopriva, I. (2006). "Kernel based algorithms for mining huge data sets," Vol. 1, Heidelberg: Springer.

[10] Yang, X. S., "Introduction to algorithms for data mining and machine learning". Academic press, 2019.

[11] Damopolii, W. W., Priyasadie, N., & Zahra, A. "Educational Data Mining in Predicting Student Final Grades," International Journal, vol. 10, no. 1, 2021.

[12] W. Duch, T. Winiarski, J. Biesiada, J, and A. Ka,chel, "Feature Ranking, Selection and Discretization," Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), pp. 251 – 254, 2003.

[13] Lamrini, B., "Contribution to Decision Tree Induction with Python: A Review," Data Mining-Methods, Applications and Systems, 2020.

[14] Alturki, S., Alturki, N., & Stuckenschmidt, H., "Using educational data mining to predict students' academic performance for applying early interventions," Journal of Information Technology Education: JITE. Innovations in Practice: IIP, vol. 20, 121-137, 2021.

[15] Osmanbegovic, E., & Suljic, M., "Data mining approach for predicting student performance," Economic Review: Journal of Economics and Business, vol. 10, no. 1, 3-12, 2012.

[16] Pal, A. K., & Pal, S., "Analysis and mining of educational data for predicting the performance of students," International Journal of Electronics Communication and Computer Engineering, vol. 4, no. 5, 1560-1565, 2013.

[17] Sivakumar, S., Venkataraman, S., & Selvaraj, R., "Predictive modelling of student dropout indicators in educational data mining using improved decision tree," Indian Journal of Science and Technology, vol. 9, no. 4, 1-5, 2016.

[18] Phauk, S., & Okazaki, T., "Integration of Educational Data Mining Models to a Web-Based Support System for Predicting High School Student Performance," International Journal of Computer and Information Engineering, vol. 15, no. 2, 131-144, 2021.

[19] Jain, R., Bekuma, Y., Pandey, B., & Gothania, J., "A Data Mining Based Approach to Evaluate Assessment Performances of Graduating Students of Schools," PSYCHOLOGY AND EDUCATION, vol. 58, no. 2, 3708-3720, 2021.

[20] Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q., "Predicting students' performance using machine learning techniques," Journal of university of babylon for pure and applied sciences, vol. 27, no. 1, 194-205, 2019.

[21] Saa, A. A., "Educational data mining & students' performance prediction," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[22] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K., "Models for early prediction of at-risk students in a course using

standards-based grading," Computers & Education, vol. 103, 1-15, 2016.

[23] Kabakchieva, D., "Predicting student performance by using data mining methods for classification," Cybernetics and information technologies, vol. 13, no. 1, 61-72, 2013.

[24] Manu, G. P. S., "Classifying educational data using support vector machines: A supervised data mining technique," Indian Journal of Science and Technology, vol. 9, no. 34, 2016.

[25] Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. S., "Mining educational data to predict students' performance," Education and Information Technologies, 1-17, 2021.