An Enhanced Predictive Approach for Students' Performance

Mohamed Farouk Yacoub[1], Huda Amin Maghawry[2],
Nivin A Helal[3], Tarek F. Gharib[5]
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

Sebastián Ventura[4]
Department of Computer
Sciences and Numerical Analysis
University of Cordoba
Spain

***Abstract*—Applying data mining for improving the outcomes of the educational process has become one of the most significant areas of research. The most important corner stone in the educational process is students' performance. Therefore, early prediction of students' performance aims to assist at-risk students by providing appropriate and early support and intervention. The objective of this paper is to propose an enhanced predictive model for students' performance prediction. Selecting the most important features is a crucial indicator for the academic institutions to make an appropriate intervention to help students with poor performance and the top influencing features were selected in feature selection step besides the dimensionality reduction and build an efficient predictive model. DB-Scan clustering technique is applied to enhance the proposed predictive model performance in the preprocessing step. Various classification techniques are used such as Decision Tree, Logistic regression, Naive Bayes, Random Forest, and Multilayer Perceptron. Moreover ensemble method is used to solve the trade-off between the bias and the variance and there are two proposed ensemble methods through the experiments to be compared. The proposed model is an ensemble classifier of Multilayer Perceptron, Decision Tree, and Random Forest classifiers. The proposed model achieves an accuracy of 83.16%.**

***Keywords*—Educational data mining; students' performance; classification; feature selection; machine learning**

I. Introduction

Educational systems contain massive data about students' behavior, enrollment of students, results, and attendance that could be analyzed to improve outcomes of the educational process [1]. Therefore, Educational Data Mining (EDM) has become a necessity to discover knowledge that helps decision makers to improve the educational process [2].EDM can reduce the drop-out rate by providing the academic institutions with knowledge to be able to develop appropriate strategies. It can also provide a timely decision to help students who are vulnerable to failure. Data mining deals with the educational field to identify and evaluate several important learning indicators from the data [3].

Students' performance prediction is one of the most challenging and interesting topics of EDM research [4]. It helps instructors to track their students' performance to identify those at-risk [12]. Research on students' performance prediction is useful to identify the features, behaviors and hidden relations that affect the students' performance [6][11].Many organizations, such as Accreditation Council for Business Schools and Programs (ACBSP) and Accreditation Board for Engineering and Technology (ABET) evaluate the educational programs quality based on the learning outcomes [7]. There are two research perspectives, the first one is to find the features that affect the performance of students. The second one is to find an effective methodology to predict students' performance [4]. It is very crucial to apply the feature selection to discover high influence features that need to be improved the dropout rate and enhance student performance [5]. Machine learning is applied to discover hidden patterns and the relations between the features in addition to the prediction of at-risk students [20].

The massive growth of the educational data gives the education institutes the opportunity to apply data mining techniques to extract useful and hidden information for predicting students' academic performance effectively [22]. Therefore, the objective of this paper is to propose an enhanced predictive model for accurate prediction of students' performance. Various machine learning techniques were experimented. For enhancing the predictive model, we applied DB-Scan clustering technique and feature selection approach. Experimental results proved the effectiveness of the proposed model. The following sections of the paper are organized as follows: Section II presents the related work in page 1 and 2, Section III introduces the methodology in page 2, experiments and results are discussed in Section IV in page 2, 3, and 4, and finally conclusion is stated in Section V in page 5.

II. Related Work

Predicting learning outcomes and especially students' performance became very important in the overall learning and educational process. Francis et al. achieved an accuracy of 75.47% using a hybrid data mining technique to predict weak students [8]. (Pojon, 2017) compared various machine learning techniques such as Decision Tree, Linear Regression, and Naïve Bayes classifiers [13]. The experiments were conducted on two datasets. Experiments have shown that models with clustering and classification techniques achieve better results in predicting students' performance. The accuracy achieved using classification and regression trees (CART) algorithm were of 93% and 78%.

Lau et al. built a prediction model for students' performance using neural networks technique. The model contains two hidden-layers and the output layer with 11 features as

input. The prediction model achieved precision of 84.8% [10]. Xing et al. conducted a prediction model using deep learning approach based on the perspective of predicting temporal drop-out for improving the online learning. The proposed model achieved an accuracy of 90.8% and 96.1%, after the 1st week and the 7th week, respectively [21]. Divyabharathi et al. constructed a predictive model using Naïve Bayes classification technique to detect and prevent performance academic risk based on the students' data, and they achieved an accuracy of 94% [22]. Raihana et al. applied classification based on the academic performance as well as the quality of life that were psychological health, physical health, social relationship, environment, and the overall life quality. They used support vector machine (SVM) algorithm and achieved an accuracy of 73.33%. The experiments revealed that students, who achieved good academic performance, were psychologically healthy, physically healthy, have good social relationship, were in a good environment and have good overall life quality [23]. Uzel et al. experimented various machine learning predictors such as Decision Tree, Naïve Bayes, Random Forest, Multilayer Perceptron, and Ensemble method. They also applied Apriori technique to discover the hidden patterns from the data. They achieved the accuracy of 80.6% for the classification by the voting method classifier[19].
Sana et al. showed that there are many features and factors that affect the final students' performance significantly such as the number of student's absence days and the involvement of parents with students in the learning process. The accuracy of 78.1% was achieved by their approach using Artificial Neural Network (ANN) algorithm with highly ranked features [18].

Comparative results of some related works are shown in Table I.

TABLE I. COMPARATIVE RESULTS OF RELATED WORK

Ref.	Algorithm	Accuracy	Dataset
[8]	Hybrid Data Mining Approach	75.47%	480 records and 16 features
[10]	Neural Networks	84.8%	1000 records
[13]	CART	78%	480 records and 16 features
[18]	ANN	78.1%	480 records and 16 features
[19]	Voting Classifier	80.6%	480 records and 16 features
[21]	Deep Learning	96.1%	3617 records and 13 features
[22]	Naïve Bayes	94%	500 records and 8 features
[23]	SVM	73.33%	60 records and 26 features

There are many studies that focus either on finding hidden patterns, discovering relation between features, or enhancing the accuracy. However there is a lack in making a comprehensive work that combines enhancing the preprocessing step, finding the top influencing features, and proposing a predictive model to enhance the performance based on different measures to check the model's stability from different perspectives and show the variations of results to get insights from the experiments. Therefore the performance of prediction models in the previous studies needs to be enhanced to help at-risk students to improve their performance and help the academic institutions to make an appropriate intervention to assist the students with poor performance. Therefore, in this study we propose a predictive model to face these limitations and improve students' performance prediction.

III. METHODOLOGY

The purpose of this paper is to propose a predictive model for students' performance prediction. This is achieved by exploring various classification techniques, besides the ensemble method that solves the trade-off between the bias and variance, to investigate which one would achieve the best performance. Moreover, DB-Scan was used in preprocessing for outlier detection and features selection was used to enhance the predictive model of students' performance. The proposed model is shown in Fig. 1. The following subsections describe the proposed model.

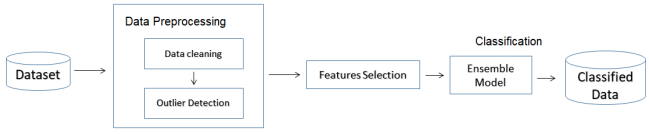

Fig. 1. The Proposed Model.

A. Preprocessing

Pre-processing is an essential process for any data set. It includes data cleaning and transformation. We pre-processed the dataset through three steps. Firstly, data was converted from nominal to numerical values. Secondly, some features were reshaped to be within a certain range using standardization method. Finally, DB-Scan clustering methodology was applied for outlier detection, as it had a great efficiency in [16].

B. Feature Selection

Feature selection aims to select the most important and influencing features in the dataset. Also, it is very important for dimensions' reduction before implementing the prediction and classification methods. It works by selecting the best features that contribute most to the target variable based on univariate statistical tests.

We used the SelectKBest technique [15]. SelectKBest technique selects the first k features with the highest score values based on the Chi-Square test, for comparing the actual and predicted results, as a score function [14] using equation (1).

$$X^2 = \sum(Oi–Ei)^2/Ei \quad (1)$$

where Oi and Ei are the actual and expected values, respectively.

C. Classification

For classification, we designed two different Ensemble models. One consists of Multilayer Perceptron (MLP), Random Forest (RF) and Decision Tree (DT), the other consists of RF, Logistic regression (LR) and DT.

IV. EXPERIMENTS AND RESULTS

To evaluate how the proposed models will perform in predicting the students' performance, several evaluation measures were used on a Learning Management System (LMS) dataset through several experiments as follows in page 3:

A. Dataset

The dataset is collected from a Learning Management System (LMS) [24]. It contains 480 records of students' data in various educational levels with 16 features. The features were categorized as follows:

• Academic features: Section id, Semester, Educational stages, Viewing announcements, Grade levels, Topic, Discussion groups, Visited resources, Raised hand, and Student absence days.

• Personal features: Gender, Parent responsible for student, Parent answering survey, and Parent school satisfaction.

• Demographic features: Nationality and Place of birth.

We pre-processed the dataset through two steps. Firstly, we converted the data from nominal to numerical values for the features: Section Id, Semester, Educational stages, Grade levels, Topic, Discussion groups, Gender, Parent responsible for student, Parent answering survey, Parent school satisfaction, Nationality, Place of birth, and Student absence days.

Secondly, the following features: Viewing announcements, Discussion, Visited resources, and Raised hand, were reshaped to be within a certain range using standardization method. Fig. 2 shows a sample of the dataset's features and instances.

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction	StudentAbsenceDays
470	0	11	5	1	5	0	8	0	0	0.81	0.868687	0.877551	0.428163	0	1	0
471	0	11	5	1	5	0	8	1	0	0.78	0.828283	0.795918	0.530612	0	1	0
472	0	11	11	1	5	0	11	0	0	0.80	0.878788	0.755102	0.663673	0	1	0
473	0	11	11	1	5	0	11	1	0	0.85	0.888889	0.806122	0.704082	0	1	0
474	1	5	5	1	5	0	10	0	0	0.02	0.070707	0.040816	0.071429	1	0	1
475	1	5	5	1	5	0	10	1	0	0.05	0.040404	0.051020	0.071429	1	0	1
476	1	5	5	1	5	0	11	0	0	0.50	0.777778	0.142857	0.275510	1	0	0
477	1	5	5	1	5	0	11	1	0	0.55	0.747475	0.255102	0.285714	1	0	0
478	1	5	5	1	5	0	8	0	0	0.30	0.171717	0.142857	0.571429	1	0	1
479	1	5	5	1	5	0	8	1	0	0.35	0.141414	0.234694	0.622449	1	0	1

Fig. 2. Sample of the Dataset Features and Instances.

B. Experiment Environment

Anaconda Navigator [17] was used to simplify the packages management and deployment. The implementation language was Python using pandas, numpy, and sklearn libraries. The experiments were performed on a machine with 2.60 GHz processor, 16 GB memory, and Windows 10 64-bit operating system. All the experiments were performed using the 10-fold cross validation [9].

C. Evaluation Measures

The accuracy measure is used for the evaluation of each model using equation (2).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

It indicates a proportion of correctly predicted observation to the total observations, where: TP = True positive, FP = False positive, TN = True negative and FN = False negative.

Moreover, Precision, Recall, and F1-Score measures are used for the evaluation of each model using equations (3), (4), and (5) respectively.

$$\text{Precision} = (TP) / (TP + FP) \quad (3)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (4)$$

$$\text{F1-Score} = 2 \text{ X } (\text{Precision X Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

D. Experiments

According to dataset features, the performance indicator was the success levels that were classified to three categories:

• Low-Level contains 127 of the data instances.
• Middle-Level contains 211 of the data instances.
• High-Level contains 142 of the data instances.

To evaluate our ensemble methods, several experiments were conducted. First, we studied the effect of using the DB-Scan as a preprocessing step as shown in the methodology section.

A comparison of our first and second ensemble classifiers' performance results before and after Applying DB-Scan in terms of accuracy, Precision, Recall, and F1-Score measures is shown in Table II.

TABLE II. PERFORMANCE OF OUR FIRST AND SECOND ENSEMBLE METHODS BEFORE AND AFTER DB-SCAN

Method	Accuracy	Precision	Recall	F1-Score
Our First Ensemble (MLP, RF, DT) (Before DB-Scan)	79.17%	79.2%	81.3%	80.1%
Our First Ensemble (MLP, RF, DT) (After DB-Scan)	83.16%	84.1%	83.3%	83.6%
Our Second Ensemble (RF, DT, LR) (Before DB-Scan)	78.125%	78.2%	80.6%	79.2%
Our Second Ensemble (RF, DT, LR) (After DB-Scan)	80.0%	80.7%	80.8%	80.7%

A comparison of classifiers' performance showing the evaluation results in terms of accuracy, Precision, Recall, and F1-Score measures after applying the DB-Scan algorithm as well as the evaluation of these classifiers using all the data without applying the DB-Scan clustering technique, and theses comparison's results are shown in Table II.

Second experiment as shown in Table III is designed to evaluate our two ensemble models comparable to the Naïve Bayes, DT and MLP models. Evaluation is done using the accuracy, precision, recall and F1-Score.

In addition to assess the accuracy of our proposed models, we compared our work with the results of [18]. We compared our two ensemble models using the given evaluation measures.

The experiments in table IV show that our first ensemble model achieved better performance in terms of different evaluation measures than [18] and our second ensemble model.

TABLE III. PERFORMANCE OF OUR FIRST AND SECOND ENSEMBLE METHODS AS WELL AS NAÏVE BAYES, DECISION TREE, AND MLP CLASSIFIERS

Method	Accuracy	Precision	Recall	F1-Score
Naive Bayes	81.05%	81.2%	83.1%	81.6%
Decision Tree	77.89%	78.7%	78.0%	78.2%
MLP	80.0%	81.2%	80.0%	80.5%
Our First Ensemble (MLP, RF, DT)	83.16%	84.1%	83.3%	83.6%
Our Second Ensemble (RF, DT, LR)	80.0%	80.7%	80.8%	80.7%

TABLE IV. PERFORMANCE OF OUR FIRST AND SECOND ENSEMBLE METHODS BESIDES [18]

Method	Accuracy	Precision	Recall	F1-Score
[18]	78.1%	78.6%	78.9%	78.8%
Our First Ensemble (MLP, RF, DT)	83.16%	84.1%	83.3%	83.6%
Our Second Ensemble (RF, DT, LR)	80.0%	80.7%	80.8%	80.7%

Finally, to study the feature selection technique, the SelectKBest technique is applied based on chi-square test for feature selection approach for selecting the most influencing features to be involved in the model building phase. It works by selecting the best features that contribute most to the target variable based on univariate statistical tests.

The experiments were implemented using the top 10 and 5 features besides all the features from the dataset to be used in the classification.

Table V shows a comparison of models' performance in terms of accuracy measure using all the features of dataset, top 10, and top 5 features of the dataset correlated to the target.

TABLE V. PERFORMANCE ACCURACY WITH VARIOUS NUMBER OF SELECTED BEST FEATURES

Method	All Features	Top 10 Features	Top 5 Features
Naive Bayes	81.05%	80.0%	64.21%
Decision Tree	77.89%	68.42%	66.32%
MLP	80.0%	81.05%	68.42%
Our First Ensemble (MLP, RF, DT)	83.16%	78.95%	65.26%
Our Second Ensemble (RF, DT, LR)	80.0%	75.79%	65.26%

According to Fig. 3, the accuracy corresponding to each number of top K features is measured and plotted for selecting the best number of features in feature selection stage and the best accuracy can be achieved by using only the top correlated 9 features to the target that is the accuracy when we used all the features.

Table VI shows a comparison of models' building time in milliseconds using all the features of dataset, top 10, and top five features of the dataset correlated to the target. The results showed that the difference between models' building time is very small. Therefore building models using all the features

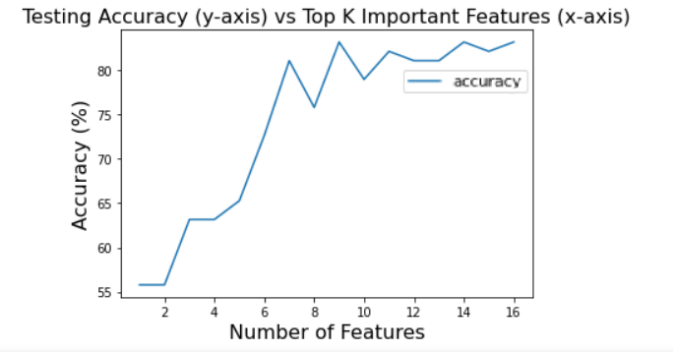

Fig. 3. Features' Selection Graph.

is the best choice. Moreover Fig. 4 shows the bar-chart graph representation of the results of Table VI.

TABLE VI. COMPARISON BETWEEN MODELS' BUILDING TIME IN MILLISECONDS WITH VARIOUS NUMBER OF SELECTED BEST FEATURES

Method	All Features	Top 10 Features	Top 5 Features
Naive Bayes	8.026	7.947	6.952
Decision Tree	10.121	9.091	8.285
MLP	145.065	162.463	155.802
Our First Ensemble (MLP, RF, DT)	217.56	231.273	207.702
Our Second Ensemble (RF, DT, LR)	127.004	142.911	110.764

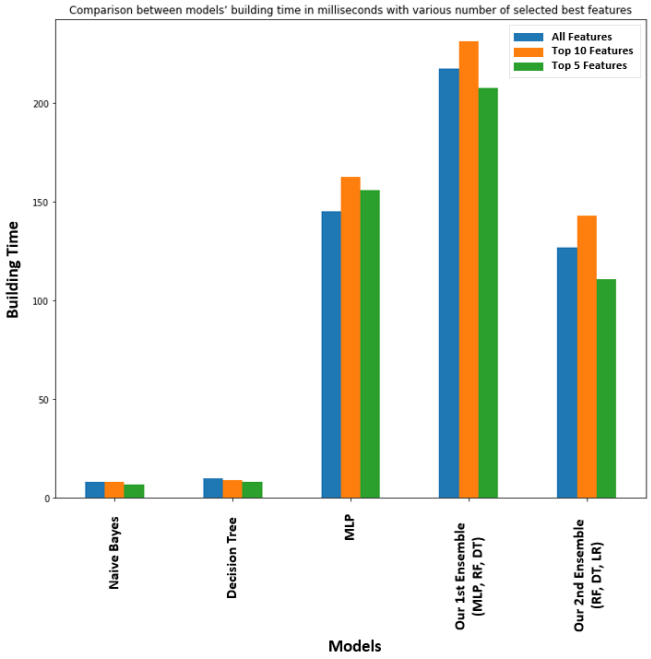

Fig. 4. Models' Building Time Graph.

V. CONCLUSION

One of the essential objectives of educational data mining is to accurately predict students who are vulnerable to drop-

out for providing them with more support and the suitable intervention. EDM helps the academic institutions to make an appropriate intervention to assist those students enhance their performance. In this paper, we proposed an enhanced predictive model for students' performance to improve the prediction accuracy. We applied various machine learning techniques for predicting the students' performance. Additionally, DB-Scan clustering algorithm and feature selection steps have been exploited, for choosing the significant features. Our first ensemble method has achieved an accuracy of 83.16%, 78.95%, and 65.26% using all the features, the top 10 influencing features, and the top 5 influencing features, respectively. The proposed predictive model outperformed previous work using the same dataset from the learning management system. Applying DB-Scan clustering technique as a preprocessing step has a great effect on enhancing the predictive model performance and the distribution of results as seen in the confusion matrix of each predictive model. For future work, we intend to apply the proposed predictive approach to various datasets, experiment different feature selection techniques, and implement alternatives for DB-Scan clustering technique.

REFERENCES

[1] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining", IEEE Access, IEEE Xplore, 2017.

[2] A. Narvekar, V. Menezes, O. P. Angle, M. Lotlekar, S. Naik, and A. Purohit, "Students performance prediction using data mining techniques", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020.

[3] A. S. Olaniyi, S. Yakub, H. Moshood, S. Ibrahim, and A. Nathaniel, "Student's performance analysis using decision tree algorithms", Anale: Seria Informatică. Computer Science Series, 2017.

[4] A. Khan, S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies", Educ Inf Technol 26, 205–240, Springer, 2020.

[5] A. U. Khasanah, H. Harwati, "A comparative study to predict student's performance using educational data mining techniques", Institute of Physics Conference Series: Materials Science and Engineering, Volume 215, 2017.

[6] A. Namoun, A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: a systematic literature review", Applied Sciences, 11(1), 237, 2021.

[7] A. Rajak, A. K. Shrivastava, and D. P. Shrivastava, "Automating outcome based education for the attainment of course and program outcomes", Fifth HCT Information Technology Trends (ITT), pp. 373–376, 2018.

[8] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach", Journal of Medical Systems, 2019.

[9] C. Goutte and J. Larsen, "Optimal cross-validation split ratio: experimental investigation", International Conference on Artificial Neural Networks, pp. 681-686, 1998.

[10] E. T. Lau, L. Sun, Q. Yang, "Modelling, prediction and classification of student academic performance using artificial neural networks", SN Applied Sciences, 2019.

[11] L. M. Zohair, "Prediction of student's performance by modelling small dataset size", International Journal of Educational Technology in Higher Education, 16-27, 2019.

[12] M. Swathi, K. L. Soujanya, R. Suhasini, "Review on predicting student performance", International Conference on Communications and Cyber Physical Engineering (pp.1323-1330), Springer, 2020.

[13] M. Pojon and J. Laurikkala, "Using machine learning to predict student performance (Master's thesis)", University of Tampere, Finland, 2017.

[14] R. Miller and D. Siegmund, "Maximally selected chi square statistics", Biometrics, pp. 1011-1016, 1982.

[15] R. Nair and A. Bhagat, "Feature selection method to improve the accuracy of classification algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019.

[16] R. Rupasi, T. Jwalitha, R. Sudarshan, T. Manasa, and G. Gundabatini, "Analysis of educational data mining using machine learning algorithms", Journal of Xi'an University of Architecture & Technology, 2021.

[17] R. Phelps, M. Krasnicki, R. A. Rutenbar, L. R. Carley, and J. R. Hellums, "Anaconda: robust synthesis of analog circuits via stochastic pattern search", IEEE, 1999.

[18] Sana, I. F. Siddiqui, and Q. A. Arain, "Analyzing students' academic performance through educational data mining", 3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición Especial, pp. 402–421, 2019.

[19] V. N. Uzel, S. S. Turgut, and S. Ayşe, "Prediction of students' academic success using data mining methods", 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2018.

[20] W. F. Yaacob, N. M. Sobri, S. A. Nasir, N. D. Norshahidi, W. Z. Husin, and K. Shaalan, "Predicting student drop-out in higher institution using data mining techniques Institute of Physics Conference Series: Materials Science and Engineering, 2020.

[21] W. Xing and D. Du, "Dropout prediction in moocs: using deep learning for personalized intervention", Journal of Educational Computing Research, 2019.

[22] Y. Divyabharathi and P. Someswari "A framework for student academic performance using naive bayes classification technique", Journal of Advancement in Engineering and Technology, 2018.

[23] Z. Raihana and A. M. Nabilah, "Classification of students based on quality of life and academic performance by using support vector machine", Journal of Academia UiTM Negeri Sembilan Vol. 6, pp. 45-52, 2018.

[24] Cortez, P. (2008) [Students' Academic Performance Dataset]. https://www.kaggle.com/aljarah/xAPI-Edu-Data