

Received 5 February 2023, accepted 21 February 2023, date of publication 1 March 2023, date of current version 22 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250702

RESEARCH ARTICLE

Early Predicting of Students Performance in Higher Education

ESSA ALHAZMI^{ID} AND ABDULLAH SHENEAMER^{ID}

Faculty of Computer Science and Information Technology, Jazan University, Jazan 45142, Saudi Arabia

Corresponding author: Abdullah Sheneamer (asheneamer@jazanu.edu.sa)

This work was supported by Jazan University under Grant W41-055.

ABSTRACT Students learning performance is one of the core components for assessing any educational systems. Students performance is very crucial in tackling issues of learning process and one of the important matters to measure learning outcomes. The ability to use data knowledge to improve education systems has led to the development of the field of research known as educational data mining (EDM). EDM is the creation of techniques to investigate data gathered from educational settings, allowing for a more thorough and accurate understanding of students and the improvement of educational outcomes for them. The use of machine learning (ML) technology has increased significantly in recent years. Researchers and teachers can use the measurements of success, failure, dropout, and more provided by the discipline of data mining in education to predict and simulate education processes. Therefore, this work presents an analysis of students performance using data mining methods. The paper presents both clustering and classification techniques to identify the impact of students performance at early stage with on the GPA. For the clustering technique, the paper uses dimensionality reduction mechanism by T-SNE algorithm with various factors at early stage such as admission scores and first level courses, academic achievement tests (AAT) and general aptitude tests (GAT) in order to explore the relationship between these factors and GPA's. For the classification technique, the paper presents experiments on different machine learning models on predicting student performance at early stages using different features including courses' grades and admission tests' scores. We use different assessment metrics to evaluate the quality of the models. The results suggest that educational systems can mitigate the risks of students failures at the early stages.

INDEX TERMS Graph mining, students' performance prediction, student academic performance, early prediction, data mining.

I. INTRODUCTION

Education is an important element and plays a significant role in our society. Information and communication technology has affected many fields of research, specifically in the education field. For example, as we have seen in many countries used various e-Learning environments [1] due to the recent pandemic COVID19.

A higher education institution considers the academic performance of students as one of the most important issues regarding presenting quality education to its students [2], [3].

Understanding the significant factors in students performance at early stage of their education is complex. Various

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu^{ID}.

effective tools have been used to overcome the students' performance challenges in academia [2], [3], [4], [5], [6]. However, these tools may not be easy to generalized in all circumstances of education. In the recent years, with the advances of the application of technologies to forecasting students' performance, there are still gaps to be filled in order to analysis and improve the accuracy of students performance using new features and data mining methods and present both clustering and classification techniques to identify the impact of students performance at early stage with on the GPA.

The learning process includes a lot of student performance. Identifying students who are more likely to have poor academic success in the future requires making predictions about student performance. If the data has been transformed into

knowledge, it may be useful and used in predictions. As a result, the information might improve the quality of education and learning and help students in achieving their academic objectives. Data mining techniques are used in the study area known as educational data mining (EDM) to analyze information derived from educational backgrounds [7]. EDM implementation also aids in the planning of strategies for raising student performance. As a result, it will improve teaching and learning and the students' experience in the educational institution.

Academic success is important because it is strongly linked to the positive outcomes we value. One of the academic success factors is the academic students performance in the college or university. The cumulative academic achievement for each student still indicates the success of every college or university. Also, the other factors we can use in analysis and predict the academic students performance are aptitude test, GPA of secondary school and the name of the school which the student graduated from. We believe that the performance of the students in the first year in college can be used as a factor to predict the performance of student in the rest of years of his/her studies. These factors lead to early remedy for students and take actions to improve student performance Artificial intelligence techniques have been applied on educational data to reveal the significant reasons behind student performance. The contributions of the paper are as follows:

- We propose a framework for predicting students performance using student's academic performance and his/her social relationships features.
- We use admission scores, his/her first level courses scores and academic achievement test (AAT) and general aptitude test (GAT).
- We explore a new way of using admission, his/her first level courses scores, and AAT and GAT by tSNE dimensionality reduction. To the best of our knowledge this attempt is a first of its kind to use features from both admission scores and his/her first level courses scores to early predict student's performance using machine learning.
- We also explore a new way of using increasing the threshold of relocating which is to compute the absolute difference between a grade and following grade after or before.
- We use a state-of-the-art classification models to evaluate the effectiveness of our proposed idea.

We organize the paper as follows: Section II provides the literature related to students performance prediction techniques used in the field of Education. Section III provides details of the used dataset includes data characterization and correlations. Section IV provides the research methodology followed by the paper. We evaluate and analysis our proposed method and report findings in Section V. Finally, we conclude our work in Section VI.

II. RELATED WORK

Conijn et al. [8] gave a summary of the theoretical justifications for the research on learning analytics and the common predictors that have been applied in recent studies. The next step is to analyze 17 blended courses with 4,989 students in a single institution using the Moodle LMS. Here, they forecasted student performance using multi-level and standard regressions. Unfortunately, it is challenging to draw broad conclusions regarding the online behavior of potential students at risk because of differences between course data. To forecast student performance and enhance teaching and learning, more theoretical justification and data sources are required.

Halal et al. [9] proposed the concept of using heterogeneity to create better prediction models. Four popular machine learning algorithms—JRip, sequential minimum optimization, C4.5, and Naive-Bayes—were used to create these models. The results of the experiment showed that employing student subpopulations to predict academic performance is both successful and promising. Additionally, it demonstrated that both rule-based and tree-based algorithms offered more clear explanations, making them more effective in an educational environment [9]. This study did not consider the combined features corresponding to a particular module.

Xu et al. [10] proposed a technique utilizing three popular machine learning algorithms to predict student performance using Internet usage data. From the real Internet usage data of 4000 students, they collected, computed, and normalized parameters like online time, Internet traffic volume, and connection frequency. Their findings indicated that it is possible to predict students' academic success using data on Internet usage.

Dien et al. [11] proposed a method to predict student performance using several deep learning methods. Twenty one features from their model are used as convolutional neural network input. The dataset was obtained from the information system of a multidisciplinary university in Vietnam, and experimental findings on the dataset revealed successful prediction. This work used traditional and simple features to build a student's performance prediction model for predicting student performance.

Giannakas et al. [1] proposed a binary classification framework based on deep neural networks (DNN), and the most crucial features that affected the result were extracted. Different activation functions (Sigmoid, ReLu, and Tanh) and optimizers were used to evaluate the framework (Adagrad and Adadelta). The experimental findings indicate that, when the Adadelta and Adagrad optimizers were utilized, the prediction accuracy of the framework was 76.73% and 82.39%, respectively. The learning performance was 80.76% and 86.57% overall.

Ha et al. [12] used various machine learning algorithms to determine the final grade average of students using a variety of criteria, including personal characteristics, university entrance scores, a gap year, and their first- and

second-year academic performance. The dataset that was used was obtained from the university's student management information system and a survey of graduates from three different years. The findings demonstrated a connection between the factors and students' academic achievement.

Rastrollo-Guerrero et al. [13] studied and analyzed almost 70 papers to show different modern techniques applied for predicting students' performance. They have studied current research on predicting student behavior in a class environment. They came to the conclusion that there is a substantial tendency to predict university student performance from the analysis of these papers. Due to its ability to deliver accurate and trustworthy outcomes, supervised learning has become a popular strategy for predicting students' behavior. On the other hand, due to the low accuracy of predicting students' conduct in the circumstances analyzed, unsupervised learning is a technique that is unattractive to researchers.

Mustafa Agaoglu [14] explains how classifier models are created using four different classification techniques: decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis. Using performance criteria for accuracy, precision, recall, and specificity, their results are compared over a given dataset composed of student replies to an actual course evaluation questionnaire.

In order to determine the most crucial factors for ensuring the academic performance of engineering students, Gonzalez-Nucamendi et al. [15] describe the determination of student profiles based on the constructs of multiple intelligences and on learning and affective techniques.

Alshanqiti and Namoun [16] proposed hybrid regression model that optimizes the prediction accuracy of student academic performance, an optimized multi-label classifier that predicts the qualitative values for the influence of various factors associated with the obtained student performance and applied combining three dynamically weighted techniques, namely collaborative filtering, fuzzy set rules, and Lasso linear regression. They need to claim their approach generality to confirm the reliability of their model. Also, they need using real academic datasets.

In Mengash [17], this study's objective is to assist higher education institutions in making good admissions decisions by predicting applicants' academic potential before granting admission. They used certain pre-admission criteria (high school grade average, Scholastic Achievement Admission Test score, and General Aptitude Test score).

Raihan et al. [18] gathered data from several universities which has 32 unique variables and 1000 instances. This data was used in this study to identify the contributing aspects to the undergraduate students' performance. It is a proven truth that passing rates are falling as the literacy rate rises daily. Academic performance of students is deteriorating daily. The goal of this study was to pinpoint the variables that differentiate pupils who perform well academically from those who are not making it.

In order to predict a student's placement in the IT business based on their academic achievement in class ten, class twelve, graduation, and backlog to date in graduation, Maurya et al. [19] have presented a few supervised machine learning classifiers.

Asalm et al. [20] proposed an approach to predict the student's grades using the mathematics and Portuguese language course grades data set and applied Deep learning model. The model of this work is only validated using two datasets and need to be validated on other large size balanced datasets.

Ahmed et al. [21] proposed an approach to predict university students' performance in final exams using the algorithm (GBDT) which is a machine learning technology used for regression, classification, and ranking tasks, and is part of the Boosting method family.

Liu and Niu [6] proposed a new approach which is called the Multi-Agent System (MAS). It is used to propose an Agent-based Modeling Feature Selection(ABMFS) model, and the selected feature subset effectively removes the features that are irrelevant to the prediction results. Then, they applied the Deep Learning techniques to construct a Convolutional Neural Network (CNN) based structure to predict student performance.

Evangelista and Sy [22] provided a method for improving the performance prediction of several single classification algorithms by employing them as base classifiers of heterogeneous ensembles and homogeneous ensembles (bagging and boosting) (voting and stacking). Their model needs to perform optimization techniques to find out the algorithm parameters and configuration.

Quy et al. [5] assess seven well-liked group fairness metrics for issues predicting student performance. On five educational datasets, they run trials with four traditional machine learning models and two fairness-aware machine learning techniques. Their study is only limited to public schools not student academics in universities

To predict undergraduate students' final test grades, Yağcı [4] proposed a prediction model. The final exam grades were divided into four groups using a comparison of a variety of machine learning methods, including RF, SVM, LR, NB, ANN, and k-nearest neighbor (KNN). These groups were "32.5," "32.5-55," "55-77.5," and " 77.5." The grades of 1854 students who took Turkish Language-I were taken into account in the proposed model using data that was gathered from a Turkish state university. The department, faculty, and midterm exam scores from the student's tree characteristics have been used to forecast the final exam grades. In comparison to other classification models, RF and ANN performed the best, classifying final exam grades with an area under the curve (AUC) of 86% and 74% accuracy, respectively.

III. DATASET DESCRIPTION

In this study, we use a students records' dataset for the five consecutive years. The dataset provides more than 275 thousands records for almost five thousands students including their identification, instructors, gender, sections,

program, admission criteria and time, graduation time and GPA, courses' details and their grades. This dataset is being used in this paper to understand the factors that influence the academic performance of the students.

A. DATA CHARACTERIZATION

From this dataset, we explored various attributes that support our task in quantifying and prediction the students performance in early stage.

By analyzing admission criteria, Figure 1 shows the density distribution function of the admission criteria. The admission criteria include the academic achievement test, the general aptitude test, and the final grade of high school. These scores are calculated in a weighted ratio to satisfy the admission requirements. The academic achievement test scores are normally distributed showing the mean and median of 66 and 65, consecutively. Similarly, the general aptitude test scores show mean and median of 68 and 69, consecutively. In contrast, the distribution of the final grade of high school scores are left-skewed with mean and median 92 and 94, consecutively.

When comparing the admission criteria with respect to gender during the five years, Figure 2 shows the slight variations between them. It is clearly noticed that the male scores are more stable than female scores in all the admission criteria.

In addition to the admission criteria, figures 3 and 4 describe the most courses are the obstacles for the students. It is clearly noticed that Math courses are the most challenging courses. The English courses are considered the second challenging courses for the students.

IV. PROPOSED METHOD

A. PREPARING THE DATA

In this phase, we use a sample of students' academic records at our CS college to explore the significant factors in students performance. Our target is to present the data of students records and features in early stage to predict their performance in the final stage (final accumulative GPA).

B. FEATURES USED

We use admission scores of the students, gender and all first-level required courses scores as features. We also mentioned all features we used in Table 1. We find these features influence the students' academic success. We attempt to predict and improve student performance by utilization of using these mentioned features.

C. T-SNE ALGORITHM

t-SNE is a non-linear dimensionality reduction used for exploring high-dimensional data. It stands for t-distributed Stochastic Neighbor Embedding [23]. We use t-SNE algorithm with various factors such as GAT and AAT for analyzing and exploring the relationship between these factors and GPA of students.

D. MACHINE LEARNING ALGORITHMS

We train and test our classification models using five machine learning algorithms. These include a recently published classification algorithm such as Xgboost [24], and others common such as Logistic Recognition [28], Support Vector Machine (SVM) [25], K nearest neighbor (KNN) [26] and Random Forest (RF) [27]. Xgboost is short for (eXtreme Gradient Boosting). Gradient boost defines an objective function that contains two parts: training loss and regularization [24]. We compare these classification algorithms in their ability to detect performance students show that supervised machine learning methods using our novel features perform much better than using just traditional features.

V. EXPERIMENTS

A. ASSESSMENT DESIGN

We perform experiments with varying numbers of features (admission scores, admission scores and gender, and admission scores with all first-level courses scores) features. The intention behind such experiments is to show the significance of our proposed features in achieving better accuracy, and that it is not by chance. This further establishes the fact that our features are crucial in deriving high accuracy detection results.

We build our model for predicting the student performance. First, we train the model with appropriate samples. To prepare training samples we extract student records from the dataset. We extract admission scores of the students, gender and all first-level required courses scores features as discussed earlier. We follow the same steps for prediction of unknown samples.

B. EXPERIMENTING WITH ADMISSION SCORES FEATURES

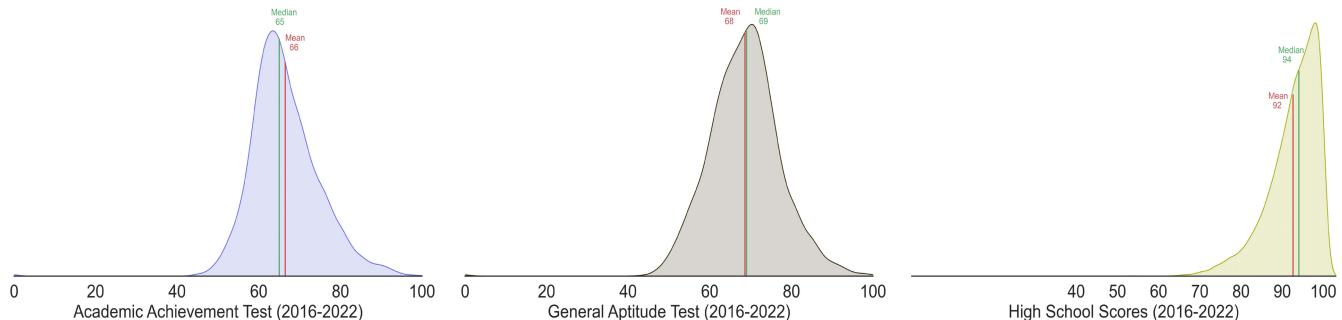
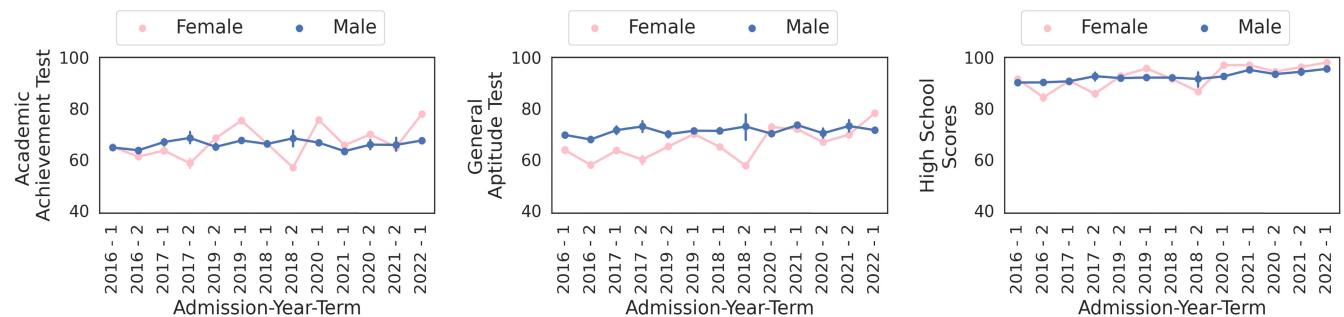
In this experiment, we used only examples of admission scores for training and testing. Our primary goal is to show student performance accuracy. We used six classifiers such as Logistic Recognition, Random forest, KNN, SVM, GNB, and XGB. However, we showed the best performed classifiers such as Random forest, SVM, and GNB. In Figure 6 and Table 2, the accuracy of admission score features results beat the model of admission scores and gender features combination.

C. EXPERIMENTING WITH ADMISSION SCORES AND GENDER FEATURES

In this experiment, we combined two different types of feature which are admission scores and gender features to observe if the gender features impact and produce better accuracy performance using only admission scores as features. Unfortunately, we noticed that the accuracy performance is reduced in all used classifiers as shown in Figure 7 and Table 3.

TABLE 1. Dataset description.

Record	Description
Student ID	Student identification number (numeric: unique number)
Sex	student's sex(binary: female or male)
Campus name	building of a college which student is related to (nominal: a college and the related building is situated)
Name of the course	A course title (nominal: course title which a student is taking)
Course level	The number by which a course is designated indicates the level of the course (numeric: identified by three to four digits)
Section number	The course section number (numeric: identified by one to four digits)
Lecturer ID	Lecturer identification number (numeric: unique number)
Semester year	Three semesters per academic year (numeric: Year+number of semester 1,2, 3)
Major	student's major (nominal: CS, IT, CNET)
Grade	Student's grade in a specific course (numeric: 0-100)
Student status	(nominal: regular, graduate, discontinuous, discontinuous)
Admission year	The semester that student's admitted to his/her college (numeric: year+number of semester 1,2, 3)
Graduate year	The semester that the student is expected to graduate (numeric: year+number of semester 1,2, 3)
Expected graduate year	student's graduated from his/her college (numeric: year+number of semester 1,2, 3)
Secondary school GPA	Student's secondary school GPA (numeric: 0-100%)
General Aptitude Test (GAT) Scores	This test is equivalent to the Arabic version of GAT and is based on mathematical and verbal skills (numeric: from 0 to 100)
Academic Achievements Test Scores	Analyzing achievement tests applied by the Ministry of Education on students in Math, Science, and Arabic subjects(numeric: from 0 to 100)
School name	Name of secondary school that the students graduated from (nominal: name of the secondary school)
Student GPA	Student's GPA when he/she graduated from the college (numerical: from 1 to 5)
Student semester GPA	Last semester of student's GPA (numerical: from 1 to 5)

**FIGURE 1.** Density distribution functions of the admission criteria including the scores of the academic achievement test, the general aptitude test, and the final grade of high school.**FIGURE 2.** Gender comparison in admission criteria over admission years and terms including the scores of the academic achievement test, the general aptitude test, and the final grade of high school.

D. EXPERIMENTING WITH ADMISSION SCORES AND ALL FIRST-LEVEL ENGLISH COURSES SCORES FEATURES

We also conducted an experiment to explore the accuracy performance of a combination of admission scores and

all first-level English courses features model as shown in Figure 8 and Table 4. Its results are better than the results of model using only admission scores or combination of admission scores and gender features. We conclude that a

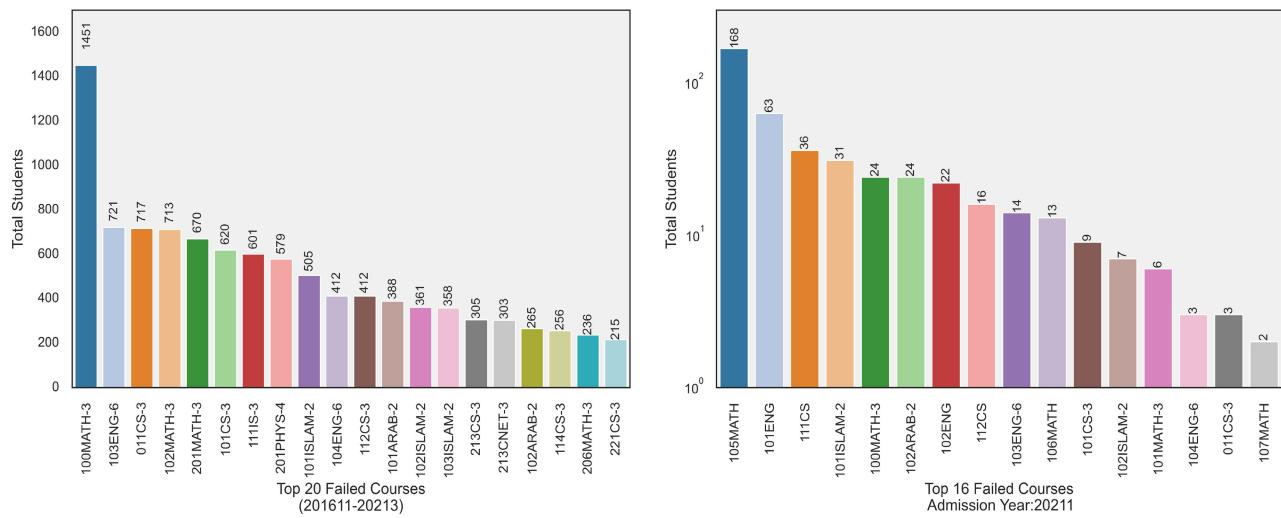


FIGURE 3. Total failed students in courses (left:during the five years, right: in 2021).

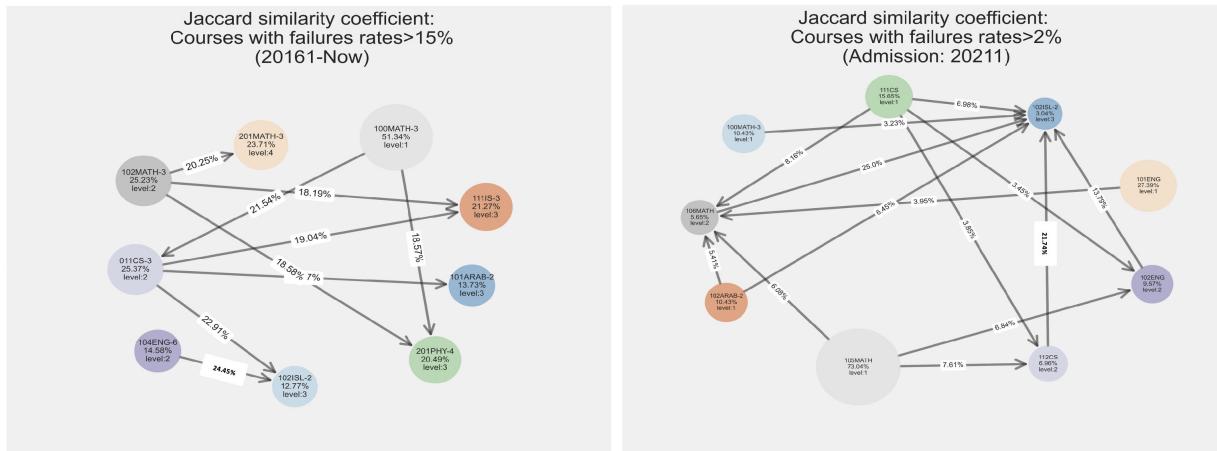


FIGURE 4. Correlation networks between the failed courses (edge direction denotes next level course) calculated by the Jaccard similarity index where similar students who failed in courses are the set.

model uses first-level English courses scores or first-level courses scores in general is better than a model uses student's characteristics information.

E. EXPERIMENTING WITH ADMISSION SCORES AND ALL FIRST-LEVEL MATH SCORES FEATURES

We also conducted an experiment for a model uses admission scores and all first-level Math courses features as shown in Figure 9 and Table 5. We noticed its accuracy performance beats all other combination features models. However, it is defeated by the model that uses combination of admission scores and all first-level computer science scores features and also combination of all various features.

F. EXPERIMENTING WITH ADMISSION SCORES AND ALL FIRST-LEVEL CS SCORES FEATURES

We also conducted an experiment for a model uses admission scores and all first-level CS courses features as shown in

Figure 10 and Table 6. All accuracy performance results beat all other combination features models. However, it is defeated by the model that uses combination of all various features.

G. EXPERIMENTING WITH ADMISSION SCORES FEATURES AND ALL FIRST-LEVEL COURSES SCORES FEATURES

Here we combine two different types of feature which are admission scores and all first-level courses scores features to assess the importance and report the results on the dataset in Figure 11 and Table 7. We observe that a admission scores and all first-level courses scores features combination beats all other combinations of features types. It produces better results than admission scores and gender features combination for all sizes of data. A admission scores and all first-level courses scores features combination works better in comparison to a admission scores and gender features combination, because gender features only have two values

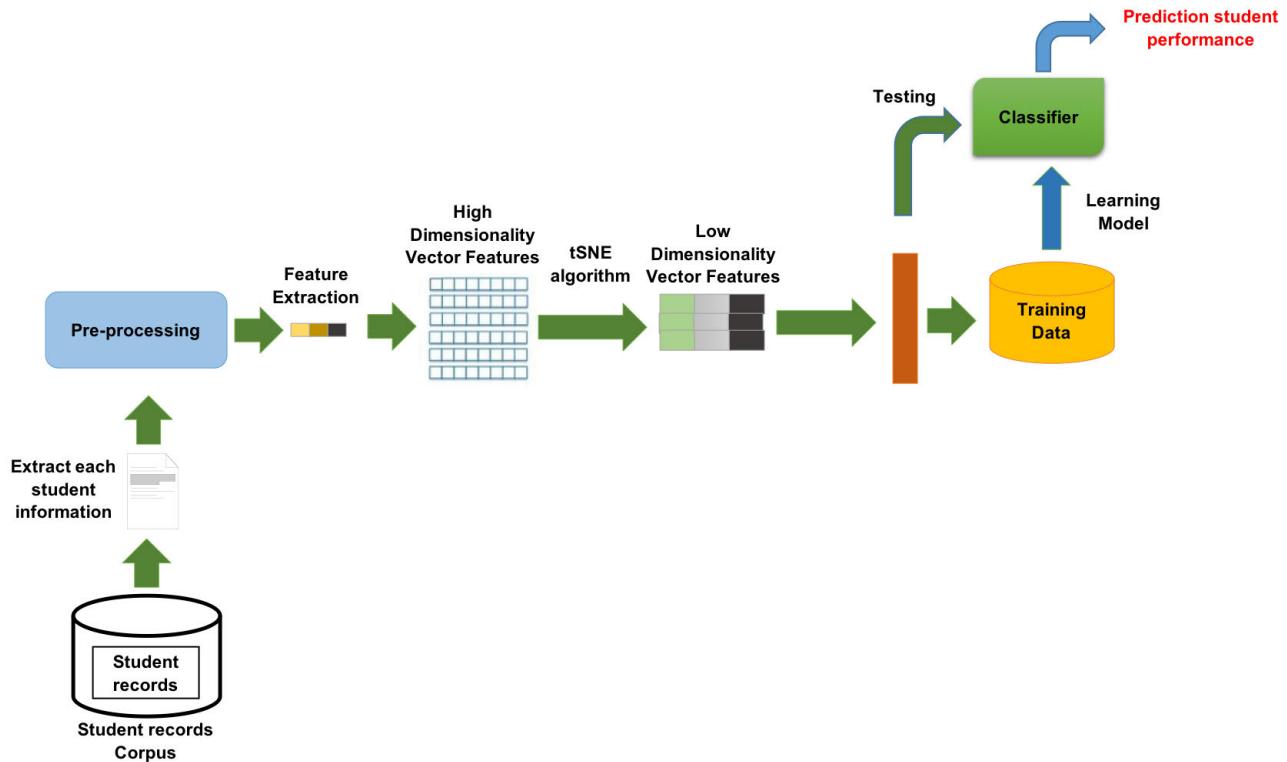


FIGURE 5. Workflow of the proposed student performance prediction framework.

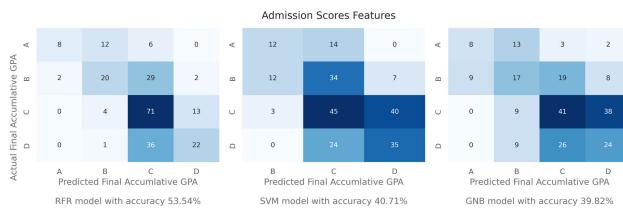


FIGURE 6. Admission scores features.

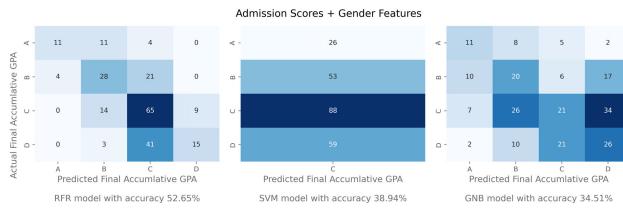


FIGURE 7. Admission scores and gender features.

either male or female. It effects on the accuracy performance of the classifiers.

H. EXPERIMENTING WITH DIMESIONALITY REDUCTION BY T-SNE

Because it is difficult to visualize data with more than two Dimensions, an essential task that involves the analysis and interpretation of high-dimensional data sets. In order to improve visualization, the t-SNE algorithm [23] is used for dimensionality reduction. This algorithm is an effective

nonlinear dimensionality reduction technique for visualizing data sets containing hundreds or even thousands of dimensions in 2D and 3D maps. We use 2D maps in this work. We use t-SNE to a lower-dimensional space to make the data easier for analyzing and visualizing. t-SNE reduced our data into 2D therefore, we can see the relationship between the classes.

I. PERFORMANCE MODELS AFTER ROUNDING GRADES

We perform six different kinds of experiments with varying combination of features (admission score, admission scores+gender, admission scores+first level Eng courses scores, admission scores+first level Math courses scores, admission scores+first level CS courses scores, and admission scores+first level courses (Eng,Math, CS scores) features as shown in tables 2, 3, 4, 5, 6, and 7. The intention behind such experiments is to show the significance of our features after rounding grades of students with different scales in achieving better accuracy. We specifically round the GPA to the closest decimal point on different ranges including 0.10, 0.30, 0.50 before measuring the accuracy of the model based on the GPA label. This method simplifies the assessment of different results of the model performance and avoids difficulties when applying errors-based metrics that might be hard to interpret.

We compare the performance of our method with different common machine learning classifiers. Tables 2 and 3 show comparison of our results with different classifiers based on

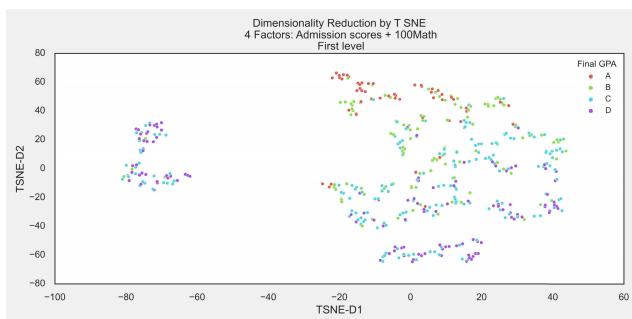


FIGURE 12. Admission scores and 100 Math using dimensionality reduction by tSNE.

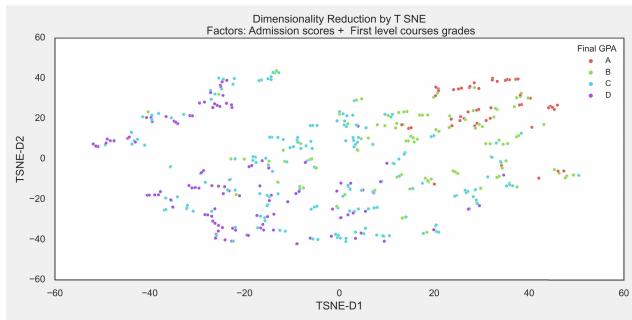


FIGURE 13. Admission scores and first level courses grades using dimensionality reduction by tSNE.

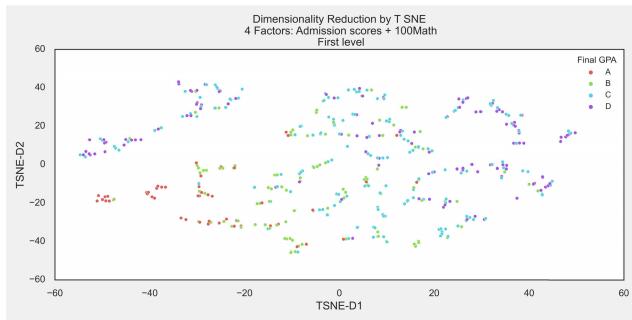


FIGURE 14. Admission scores and English using dimensionality reduction by tSNE.

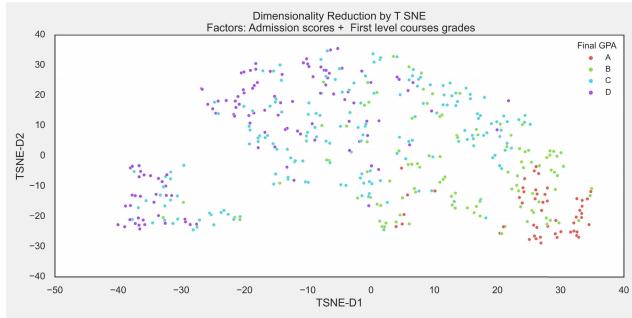


FIGURE 15. Admission scores and All first level courses grades using dimensionality reduction by tSNE.

$$Precision(P) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (3)$$

$$F1-score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

In the above equations, TP, TN, FN, and FP are true positives, true negatives, false positives, and false negatives. We use F1-score as the primary performance indicator to evaluate all the classifier models used in our experiments. F1 score is a single metric that combines both precision and recall. The precision or True Positive Rate (TPR) is a way to look at the accuracy of positive predictions performed by a classifier and can be defined as follows: precision = $TP/(TP + FP)$ where the True Positives (TP) is the number of true positives, i.e., correct prediction of a positive sample, and the False Positives (FP), i.e., the wrong prediction. But precision is used with another parameter called recall. Recall is defined by $TP/(TP + FN)$.

We have also used a confusion matrix table to study the performance of classifiers. The confusion matrix is a table with rows and columns that report false positives, false negatives, true positives, and true negatives. This allows a more detailed analysis than the mere proportion of correct classifications, i.e., prediction accuracy.

VI. CONCLUSION

The student performance is a vital issue. It is difficult to deal with this issue. This paper presented an analysis of the results data mining research to develop models of students' performance prediction. Our paper showed the use of machine learning algorithms to be better understand efficiency of the algorithms with data dimensionality reduction by T-SNE. It uses four factors such as admission scores and first level courses, academic achievement test (AAT) and general aptitude test (GAT). In the future, we would like to use deep learning architectures to construct the prediction and improve performance. It can be combined non-academic features with academic features.

REFERENCES

- [1] F. Giannakas, C. Troussas, I. Voyatzis, and C. Sgouropoulou, "A deep learning classification framework for early prediction of team-based academic performance," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107355.
- [2] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student academic performance prediction model using decision tree and fuzzy genetic algorithm," *Proc. Technol.*, vol. 25, pp. 326–332, Jan. 2016.
- [3] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *J. Med. Syst.*, vol. 43, no. 6, pp. 1–15, Jun. 2019.
- [4] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, pp. 1–19, Dec. 2022.
- [5] T. Le Quy, T. H. Nguyen, G. Frieg, and E. Ntouts, "Evaluation of group fairness measures in Student performance prediction problems," 2022, *arXiv:2208.10625*.
- [6] X. Liu and L. Niu, "A student performance predication approach based on multi-agent system and deep learning," in *Proc. IEEE Int. Conf. Eng., Technol. Educ. (TALE)*, Dec. 2021, pp. 681–688.
- [7] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, pp. 135–146, Jul. 2007.
- [8] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, Jan./Mar. 2017.

- [9] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowl.-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.
- [10] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *J. Comput. Hum. Behav.*, vol. 98, pp. 166–173, Sep. 2019.
- [11] T. T. Dien, S. H. Luu, N. Thanh-Hai, and N. Thai-Nghe, "Deep learning with data transformation and factor analysis for student performance prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 1–11, 2020.
- [12] D. T. Ha, P. T. T. Loan, C. N. Giap, and N. T. L. Huong, "An empirical study for student academic performance prediction using machine learning techniques," *Int. J. Comput. Sci. Inf. Secur. (IJCSIS)*, vol. 18, no. 3, pp. 1–8, 2020.
- [13] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Appl. Sci.*, vol. 10, no. 3, p. 1042, 2020.
- [14] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016.
- [15] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, R. M. G. García-Castelán, and D. Escobar-Castillejos, "The prediction of academic performance using engineering student's profiles," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107288.
- [16] A. Alshanti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020.
- [17] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.
- [18] M. Raihan, M. T. Islam, P. Ghosh, J. H. Angon, M. M. Hassan, and F. Farzana, "A machine learning approach to identify the correlation and association among the students' educational behavior," in *Proc. Int. Conf. Comput. Advancements*, Jan. 2020, pp. 1–6.
- [19] L. S. Maurya, M. S. Hussain, and S. Singh, "Developing classifiers through machine learning algorithms for student placement prediction based on academic performance," *Appl. Artif. Intell.*, vol. 35, no. 6, pp. 403–420, May 2021.
- [20] N. Aslam, I. Khan, L. Alamri, and R. Almuslim, "An improved early student's academic performance prediction using deep learning," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 16, no. 12, pp. 108–122, 2021.
- [21] D. M. Ahmed, A. M. Abdulazeez, D. Q. Zeebaree, and F. Y. H. Ahmed, "Predicting university's students performance based on machine learning techniques," in *Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (I2CACIS)*, Jun. 2021, pp. 276–281.
- [22] E. D. L. Evangelista and B. D. Sy, "An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 12, no. 5, p. 5226, Oct. 2022.
- [23] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
- [24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
- [26] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
- [27] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2007.
- [28] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson Education, 2002.



ESSA ALHAZMI is currently an Assistant Professor in computer science program with Jazan University, Saudi Arabia, where he is the Vice Dean of Research and Development with the College of Computer Science and Information Technology. He focuses on quantifying and modeling large-scale data using artificial intelligence algorithms. He has contributed and participated in academia and the research community by publishing and reviewing papers in different conferences and journals. His research interests include different areas in data science, social computing, and networks science.



ABDULLAH SHENEAMER received the B.Sc. degree in computer science from King Abdulaziz University, Saudi Arabia, in 2008, and the M.Sc. and Ph.D. degrees in computer science from the University of Colorado at Colorado Springs, USA, in 2012 and 2017, respectively. He is currently an Associate Professor of computer science with Jazan University, Saudi Arabia, where he was the former Vice Dean of the Faculty of Computer Science and Information Technology. His current work focuses on software clone and refactoring software clone, malware detection, and code obfuscation detection using machine learning approaches. He has published several papers in reputed international journals and conferences. His research interests include data mining, machine learning, software engineering, and malware analysis. He had reviewed several articles in reputed journals, such as IEEE ACCESS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, *Information Sciences* (Elsevier), and *Frontiers of Computer Science*. He is a Senior Meta Reviewer of IEEE International Conference on Machine Learning and Applications.