Prediction of Academic Performance of Engineering Students by Using Data Mining Techniques

Swati Verma, Rakesh Kumar Yadav, and Kuldeep Kholiya

***Abstract*—In the current age, students' academic performance deterioration is a very crucial problem in engineering education. Prediction of low-performing students at an early stage is important so that their faculties and administration could provide timely support. The present study attempts to perform this prediction task at the entry-time with the help of four single supervised educational data mining algorithms, namely: Decision tree, Naïve Bayes, k-Nearest Neighbor, and Support Vector Machine along with an ensemble method called "Random Forest". These classifiers have been applied to a students' dataset of an Indian Engineering College, having four categories of parameters viz., student's background, academic, social, and psychological parameters. Different libraries of Python programming language such as Pandas, Seaborn, Scikit-learn, and Scipy were used for analysis, visualization, classification, and statistics computation, respectively. The present study shows that among all of the five algorithms, Naïve Bayes gives the highest accuracy with 89%, and finally to improve the results, a model is proposed in which three Naïve Bayes classifiers were integrated with the help of 'Bagging'. The achieved accuracy with the proposed model was 91%, with the highest recall and highest precision for identifying low performers.**

***Index Terms*—Chi-square test, classification, educational data mining, students' academic performance.**

I. Introduction

In this era of technology, engineering education plays a very important role in the growth and betterment of the nation. There are thousands of engineering institutions in which every year, million of students are admitted, but many of them are either dropout or do not get an engineering degree timely. Thus, there is a requirement to identify the parameters that diminish the performance of the engineering students and find a suitable educational data mining technique for predicting low- performing students at an early stage so that necessary support could be given to help them.

There are numerous kinds of parameters that may affect the performance of the engineering students. These parameters include background parameters (gender, medium/language of study, living location, category, annual family income, parent occupation, parent qualification, etc.), academic parameters (10th standard marks, 12th standard marks, JEE Rank, etc.), social parameters (use of the internet, food habit, outing with friends, etc.), and psychological parameters (motivation to join engineering course, homesickness, interest in the study, etc.). At the present time, there is a large amount of educational information that exists in every educational institute. This information can be utilized for identifying low-performing students by using the educational data mining techniques that support various methods of machine learning, statistics, database systems, etc., and then after examining this information decisions may be concluded so that timely support could be given. There are several techniques of educational data mining such as classification, clustering, association rule mining, etc. which can be applied to students' datasets for identifying low-performing students. In order to identify low performers, previous authors [1]-[15] have proposed and analyzed the use of Educational Data Mining (EDM) techniques during the course but in the present study, emphasis is given to identifying students likely to fail from the beginning of the course, so that timely help may be provided to the needy students. The present study has three important research objectives: i) to analyze the effect of various input parameters on the first-semester grade of the engineering students for finding out influential input parameters, ii) to identify the best performing data mining technique by comparing five supervised educational data mining techniques (classification techniques), viz., Decision tree, Naïve Bayes, k-Nearest Neighbor, Random Forest, and Support Vector Machine in predicting students' performance for the present dataset, and iii) to propose a model for enhancing the accuracy in predicting low performers by integrating best performing single data mining techniques.

The present paper is organized as follows: Section II presents a brief literature review, the method used in the present study is conferred in Section III, the obtained results are discussed in Section IV, and in Section V the conclusion and future work are given.

II. Related Works

There are various studies performed by different authors to predict students' academic performance by using supervised (classification) as well as unsupervised (clustering and association rule mining) educational data mining techniques. These techniques were applied to students' datasets containing various features related to their background, academic performance, social behavior, psychology, etc.

Buldu and Üçgün [1] used the Apriori algorithm to find out the association rules between the courses in which students failed. Another team consisting of Bhardwaj and Pal [2] applied the Bayes classification model to the 300 BCA

Manuscript received May 18, 2022; revised June 27, 2022.

Swati Verma and Rakesh Kumar Yadav are with the Department of Computer Science and Engineering, IFTM University Moradabad, Uttar Pradesh, India (e-mail: mgsswati@gmail.com, rkyiftmuniversity@yahoo.com).

Kuldeep Kholiya is with the Department of Applied Science, B.T. Kumaon Institute of Technology, Dwarahat, Uttarakhand, India (e-mail: kuldeep_phy1@rediffmail.com).

doi: 10.18178/ijiet.2022.12.11.1734

students' dataset and found that living location, medium of teaching, and senior secondary grade were the most influential factors that affected the student's division or performance. Further, after implementing five classifiers such as decision tree (J48), k-Nearest Neighbor, Bayesian classifiers (Naïve Bayes, Bayes Net), and rule learners (OneR, JRip), Kabakchieva [3] found that the decision tree (J48) classifier has the highest prediction accuracy. Ajay and Saurabh [4] applied three classification algorithms, viz., Iterative Dichotomiser 3 (ID3), C4.5, and Bagging to students' data for predicting their academic performance and concluded that ID3 has the highest classification accuracy of 78% and the lowest average error of 0.16. Moreover, Huang and Fang [5] performed a comparison among four classifiers, namely: Multiple Linear Regression, Multilayer Perceptron Network Model, Radial Basis Function Network Model, and Support Vector Machine to the student's academic information (CGPA, grades in four pre-requisite courses, and scores in three dynamics midterm exams) for predicting student performance in the engineering dynamics course. In their study, the Support Vector Machine had the highest average prediction accuracy of 64%. Another team consisting of Amrieh*et al.* [6] implemented Artificial Neural Network, Naïve Bayes, and Decision Tree as well as Bagging, Boosting, and Random Forest to the students' demographic, academic, and behavioral features. In their study, students' behavioral characteristics significantly affected their academic success. Another team, Hamoud *et al.* [7], applied Naïve Bayes and Bayes Net on the 161 students' dataset and concluded that Naïve Bayes performed better than the Bayes Net for the prediction of students' performance.

Asif *et al.* [8] applied a decision tree as well as a clustering technique to a dataset of 210 undergraduate students, which comprises pre-admission marks and all four-year subjects' marks for analyzing students' progress and found that the pre-university marks and subjects' marks in the first and second years had an impact on students' final year marks. Furthermore, Costa *et al.* [9] performed a comparison of the effectiveness of different educational data mining techniques to predict students' performance in introductory programming courses and concluded that the support vector machine outperformed. Pavithra *et al.* [10] implemented Decision Tree, Naïve Bayes, Multilayer Perceptron, and Rep Tree on the 127 student dataset of Sree Saraswathy Thyagaraja College to predict their ability to get a job and found Naïve Bayes to be outperformed. Other authors, Gray and Perkins [11], predicted students' performance as early as week 3 with the help of 1-NN, C4.5 Trees with Leave One Out Cross-Validation. Adekitan and Salau [12] predicted the final cumulative grade point average (CGPA) by using the previous three years' grade point average (GPA) and six data mining algorithms viz., Probabilistic Neural Network, Random Forest, Decision Tree, Naïve Bayes, Tree Ensemble, and Logistics Regression, and found the highest accuracy of 89.15% for Logistic Regression. Dinh Thi Ha *et al.* [13] applied OneR, PART, Random Tree, J48, Random Forest, MLP, SVM, and Naive Bayes to students' background and academic attributes for the prediction of the final GPA of students and observed that MLP and Naïve Bayes had the highest accuracy of 86.19%. Another research group, Tomasevic *et al.* [14], has compared supervised data mining techniques for the prediction of students' examination performance. Recently, to predict and analyze student performance, Dixit *et al.* [15] used the Case-Based Reasoning Knowledge Base System (CBR-KBS) model that was suitable for choosing the best performers to get a job.

In past studies, ensemble models that integrate several base models were also used to improve the accuracy of the results. Ashraf *et al.* [16] implemented base classifiers and an ensemble classifier, namely "StackingC". In their study, the ensemble classifier achieved better results than the base classifiers. Moreover, Injadat *et al.* [17] implemented an ensemble-based model on two different datasets, and their experimental results show that the proposed ensemble model accomplished high accuracy and a low false-positive rate at all stages for both datasets. Furthermore, Asselman *et al.* [18] also used three ensemble models, viz., Random Forest, AdaBoost, and XGBoost, to enhance the prediction accuracy of students' performance and found that XGBoost had achieved the highest accuracy.

M. Yagci [19] compared the performance of Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayes, and k-Nearest Neighbor for the prediction of final exam grade by using a dataset of 1854 students having midterm exam grades, departmental data, and faculty data. In their study, Random Forest outperformed with 74.6% accuracy. A. Hussain *et al.* [20] implemented a hybrid model containing Decision Trees and Support Vector Machines to predict students' academic performance and identify factors that contribute to their academic performance. H. Yuliansyah [21] presented a prediction model for students' on-time graduation using the C4.5 algorithm by considering four features, namely the department, GPA, English score, and age. The prediction performance result achieved 90% accuracy using 300 testing data.

Goundar *et al.* [22] applied Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine algorithms to build predictive models to determine whether a student will pass or fail the course. The results concluded that the Random Forest algorithm had superior predictive performance capability.

Previous research has concluded that students' academic performance can be predicted by using different input parameters (such as academic, background, social, behavioral, and psychological features) and different supervised (such as classification) and unsupervised (such as association and clustering) techniques. Further, the result of classification depends upon the dataset and the educational data mining technique [23].

Although in most of the previous research, students' academic performance was predicted during the course, the present study attempts to predict academic performance at the entry-time or just after admission to the institute.

III. Research Methodology

The main goal of the present study is to compare the performance of the educational data mining techniques for the early identification of poor students and finally propose a model to enhance prediction accuracy, recall, and precision.

For this purpose, the dataset of 383 students' belonging to seven different engineering branches, viz., Computer Science & Engineering, Electronics and Communication Engineering, Chemical Engineering, Mechanical Engineering, Electrical Engineering, Civil Engineering, and Biochemical Engineering of Bipin Tripathi Kumaon Institute of Technology, Dwarahat (Almora), India, was used. The Dataset contains the following parameters: i) Background parameters (Gender, Category, Number of siblings, Status of parent, Father's highest qualification, Mother's highest qualification, Father's occupation, Mother's occupation, Annual family income, Living location, and Medium/Language of the previous study), ii) Academic parameters (10^{th} standard grade, 12^{th} standard grade, JEE rank, self-study time), iii) Social parameters (Participation in extra-curricular activities, Have good friends in your batch), iv) Psychological parameters (Motivation to join the course, Health-issue, Homesickness). The data was collected online by using outsourced technology and then cross-verified for background and academic parameters with the institute database as well. A detailed description of the student-related parameters is provided in Table I, and the part of the dataset is shown in Table II.

To make analysis and classification easy and efficient, data is preprocessed by feature selection. In the present study, as all the variables in the dataset are categorical, the chi-square feature selection technique is used to find out the influential parameters that affect the first-semester grade. The p-values from the chi-square technique have been calculated with the help of the chi2 method of the sklearn.feature_selection library in Python. The p-values less than 0.01 show the significant relationship between first-semester grades and the selected categorical variable. This procedure is used to choose the right subset of parameters so that the accuracy and training of the model can be improved.

To predict students' academic performance, five supervised data mining algorithms, namely: Decision Tree, Naïve Bayes, k-Nearest Neighbors (k-NN), Random Forest, and Support Vector Machine (SVM) were used on the preprocessed data. All the classifiers were implemented with the help of the Scikit-learn tool in Python. To evaluate the performance of a classifier, the confusion matrix is a very beneficial tool and it is shown in table III for two classes. With the help of the confusion matrix, the following metrics are evaluated-

1) *Accuracy:* It is defined as the percentage of test data that is correctly predicted by the model.

$$Acc\ (M) = (TP+TN) / (P+N)$$

2) *Precision:* Precision is the percentage of tuples that are actually positive out of the total number of tuples that are predicted as positive.

$$Precision = TP/(TP+FP) = TP/P'$$

3) *Recall: I*t is the percentage of positive tuples that are correctly classified by the model.

$$Recall = TP/P$$

Here, *TP* denotes the positive tuples that were correctly classified, *TN* is negative tuples that were correctly classified, *FP* is negative tuples that were incorrectly classified as positive, *FN* is positive tuples that were incorrectly classified as negative, and *P* and *N* are the total number of actual positive tuples and the total number of actual negative tuples, respectively, while *P'* and *N'* represents total number of tuples that are predicted as positive and total number of tuples that are predicted as negative, respectively.

After finding the best-performing data mining technique with the help of accuracy, precision, and recall, an ensemble model was formed by integrating the best-performing classifiers with the help of 'Bagging'.

The Schematic of the proposed methodology is depicted in Fig. 1.

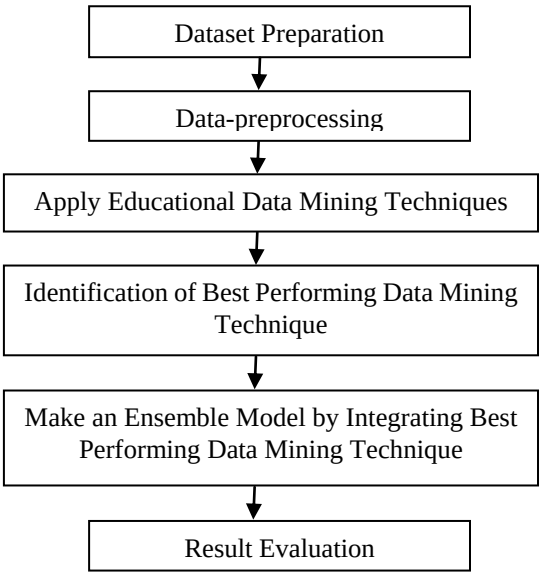

Fig. 1. Framework of the proposed work methodology.

IV. Result and Discussion

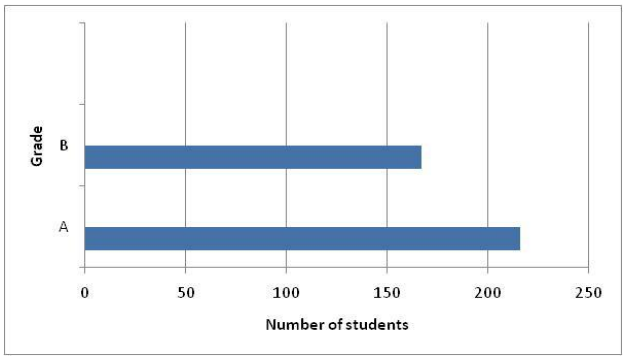

Fig. 2. Distribution of the students' first-semester grade.

After collecting the dataset of 383 students, the data is preprocessed before classification. Data were collected using outsourced technology, with all questions being objective type and mandatory, ensuring that the data was free of noise and complete for all 383 students. Thus, there was no need for data cleaning and all 383 records were used in the present study. Another thing that may be done for data preprocessing is data balancing. In the case of an imbalanced dataset, the classifier assigns each new object to the majority class only, so the accuracy measure is insufficient to determine the performance of the classifier[24]-[26]. Li and Sun [27] stated

in their study that a dataset is called imbalanced if the percentage of minority class is less than 35%. But in the present study, the number of students having Grade A is 216 (56.4%) and having Grade B is 167 (43.6%) hence, the dataset is approximately balanced. The distribution of the students' first-semester grades is shown in Fig. 2.

TABLE I: STUDENTS RELATED ACADEMIC VARIABLE

Attribute Category	Attribute	Possible Values
Background Parameters	Gender	{Male, Female}
	Category	{General, OBC, SC, ST}
	Number of siblings	{None, One, Two or above}
	Status of parent	{Living together, Living apart}
	Father's highest qualification	{None, Primary or upper primary, Secondary or higher secondary, Graduate, Post Graduate or above}
	Mother's highest qualification	{None, Primary or upper primary, Secondary or higher secondary, Graduate, Post Graduate or above}
	Father's occupation	{None, Own business, Private job, Government Job}
	Mother's occupation	{None, Own business, Private job, Government Job}
	Annual family income	{0-100000, 100001-250000, 250001-500000, Above 500000}
	Living location	{Rural area, Urban area}
	Medium/language of previous study	{English, Hindi}
Academic Parameters	10^{th} standard %	{Below 60%, Between 60% to 75%, Above 75%}
	12^{th} standard %	{Below 60%, Between 60% to 75%, Above 75%}
	Entrance exam/JEE Rank	{Below 100000, Between 100000 to 250000, Between 250001 to 500000, Above 500000}
	Average Self-Study Time	{Less than 1 hour, 1-2 hours, More than 2 hours}
	First Semester Grade (Target Variable)	{A (≥ 70%), B (<70%)}
Social Parameters	Participation in Extra-Curricular Activities	{Yes, No}
	Have good friends in your batch	{Yes, No}
Psychological Parameters	Motivation to join course	{Do not know, Other's motivation, Self-motivation}
	Health Issue	{Yes, No}
	Homesickness	{Yes, No}

TABLE II: PART OF THE ORIGINAL DATASET, WITH 383 RECORDS AND 20 INPUT ATTRIBUTES

Category	Father's Qualification	Mother's Qualification	Medium	10^{th} standard %	12^{th} standard %
OBC	Post graduate or higher	Post graduate or higher	Hindi	Above 75%	Above 75%
General	Graduate	Post graduate or higher	English	Above 75%	Between 60% to 75%
SC	Secondary or higher secondary	Primary or upper primary	English	Above 75%	Above 75%
General	Primary or upper primary	Primary or upper primary	English	Between 60% to 75%	Between 60% to 75%
General	Secondary or higher secondary	Graduate	English	Above 75%	Above 75%
OBC	Graduate	Post graduate or higher	English	Above 75%	Above 75%
OBC	Graduate	Secondary or higher secondary	Hindi	Between 60% to 75%	Above 75%
General	Secondary or higher secondary	Graduate	English	Above 75%	Above 75%
General	Secondary or higher secondary	Secondary or higher secondary	English	Above 75%	Above 75%
OBC	Post graduate or higher	None	Hindi	Between 60% to 75%	Between 60% to 75%
General	Secondary or higher secondary	Secondary or higher secondary	English	Above 75%	Above 75%

TABLE III: CONFUSION MATRIX

Actual/Predicted	C1	C2	Total
C1	True Positive (TP)	False Negative (FN)	P
C2	False Positive (FP)	True Negative (TN)	N
Total	P'	N'	

To study the effects of various parameters on the first-semester grades of engineering students, p-values for different attributes have been calculated by using the chi-square feature-selection technique.

The calculated p-values for background, academic, psychological and social attributes are shown in Tables IV, V, VI, and VII, respectively. From these tables, it may be concluded that past academic parameters such as the percentage of 10^{th} & 12^{th} standards have the most significant effect on the academic performance of engineering students as the p-value for the attributes comes out to be 7.689982e-37 and 1.203198e-26, respectively. Among the background parameters, calculated p values for Medium/language of the previous study, Mother's highest qualification, Category, Father's highest qualification comes out to be 2.511737e-10, 0.0002620, 0.0003097 and 0.0006837, respectively so these parameters also affect students' academic achievement. But in the present study, psychological and social attributes have not shown any significant effect on academic performance.

TABLE IV: CALCULATED P-VALUES FOR BACKGROUND ATTRIBUTES

Attribute	Calculated p-value	Attribute	Calculated p-value
Gender	0.1044184	Father's occupation	0.9634918
Category	0.0003097	Mother's occupation	0.5139445
Number of siblings	0.6885947	Annual family income	0.9218397
Status of parent	0.5375286	Living location	0.1536680
Father's highest qualification	0.0006837	Medium/language of previous study	2.511737e-10
Mother's highest qualification	0.0002620		

TABLE V: CALCULATED P-VALUES FOR ACADEMIC ATTRIBUTES

Attribute	Calculated p-value	Attribute	Calculated p-value
10th standard %	7.689982e-37	Entrance exam/JEE Rank	0.4732423
12th standard %	1.203198e-26	Average Self-Study Time	0.0333741

TABLE VI: CALCULATED P-VALUES FOR PSYCHOLOGICAL ATTRIBUTES

Attribute	Calculated p-value	Attribute	Calculated p-value
Motivation to join course	0.9729823	Homesickness	0.5541018
Health Issue	0.5451448		

TABLE VII: CALCULATED P-VALUES FOR SOCIAL ATTRIBUTES

Attribute	Calculated p-value	Attribute	Calculated p-value
Participation in Extra-Curricular Activities	0.4032153	Have good friends in your batch	0.7137615

For the sake of simplicity and better understanding, calculated p-values for all the variables are shown in Fig. 3.

Thus, in the present study, among 20 variables, only 6 most influential parameters, viz., father's highest qualification, category, mother's highest qualification, medium/language of the previous study, 12th standard %, and 10th standard % were selected as input parameters for each model to evaluate performance. The effect of these six attributes on the first-semester grade of the students is shown in Fig. 4. From this figure, it may be concluded that students having more than 75% in the 10th standard, or 12th standard, or general category have a higher probability to achieve grade 'A' while students having medium/language of the previous study other than English or less qualified parents have a lower probability of achieving grade 'A'.

After making a suitable subset with these six input parameters, all the different supervised classifiers were applied to the hold-out method that considers the 80:20 percentage train-test split ratios. To make the testing dataset to be approximately equal to the original dataset in the class distribution, stratified sampling was used. During the classification, we performed the hyper parameter tuning of educational data mining techniques using the GridSearchCV algorithm to achieve the best results.

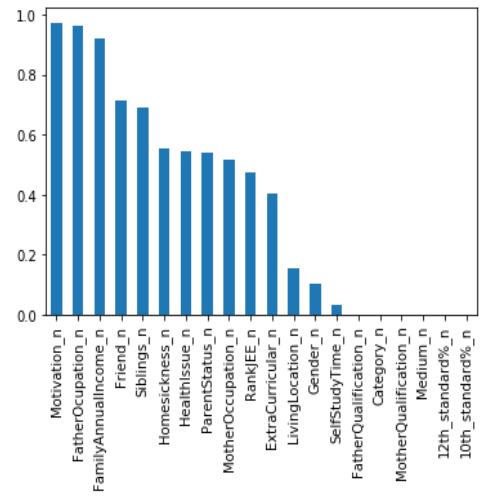

Fig. 3. Input parameters and their corresponding p-values.

A. Result Using Decision Tree Algorithm

In the present study, CART (Classification & Regression Tree) classification algorithm was implemented on the dataset, which chooses "Gini" as the attribute selection criteria, and the results of the classification are presented in Table VIII. It is inferred from Table VIII that CART has correctly classified about 85.71% dataset with max_depth=3 and max_leaf_nodes=8 as the passing parameters, which are set with the help of the GridSearchCV algorithm. The results from the table reveal that recall and precision are high for class A, but recall is not good enough for class B.

TABLE VIII: CONFUSION MATRIX FOR DECISION TREE

Actual/Predicted	A	B	Precision
A	40	3	0.83
B	8	26	0.90
Recall	0.93	0.76	

B. Result Using Random Forest Algorithm

The classification results after implementing Random Forest classifier on the dataset are presented in Table IX. From the table, it was found that the Random Forest classifier gave about 79.22% accuracy with n_estimator=7 and random_state=1 as passing parameters that were set again with the help of the GridSearchCV algorithm. It may also be noted from the table that the results obtained from the Random Forest algorithm are lower than that of the Decision Tree algorithm.

TABLE IX: CONFUSION MATRIX FOR RANDOM FOREST

Actual/Predicted	A	B	Precision
A	35	8	0.81
B	8	26	0.76
Recall	0.81	0.76	

C. Result Using Naïve Bayes Algorithm

After the implementation of a Bayesian classifier namely Naïve Bayes on the dataset, the results are presented in Table X. The Table found that Naïve Bayes classifier correctly classifies about 89.61%. Moreover, the results from the table reveal that precision and recall are high for both classes so Naïve Bayes can identify low performers and good performers efficiently.

TABLE X: CONFUSION MATRIX FOR NAÏVE BAYES

Actual/Predicted	A	B	Precision
A	40	3	0.89
B	5	29	0.91
Recall	0.93	0.85	

D. Result Using k-Nearest Neighbors (k-NN) Algorithm

TABLE XI: CONFUSION MATRIX FOR K-NN

Actual/Predicted	A	B	Precision
A	39	4	0.83
B	9	25	0.87
Recall	0.91	0.76	

The result obtained by using the k-NN algorithm is given in Table XI. From this confusion matrix, it was found that 84.42% accuracy was achieved with n_neighbors=19 as a passing parameter to the k-Neighbors classifier. The table shows that precision and recall were less compared to Naïve

Bayes for both the classes.

E. Result Using Support Vector Machine (SVM) Algorithm

SVM was also implemented on the dataset and 81.81% accuracy was achieved with C=1 and kernel='rbf'. These parameters are achieved using GridSearchCV. The results of SVM using the hold-out method are shown in Table XII. It is shown that SVM achieved lower recall and lower precision for class B than Naïve Bayes so it cannot recognize poor performers as efficiently as Naïve Bayes.

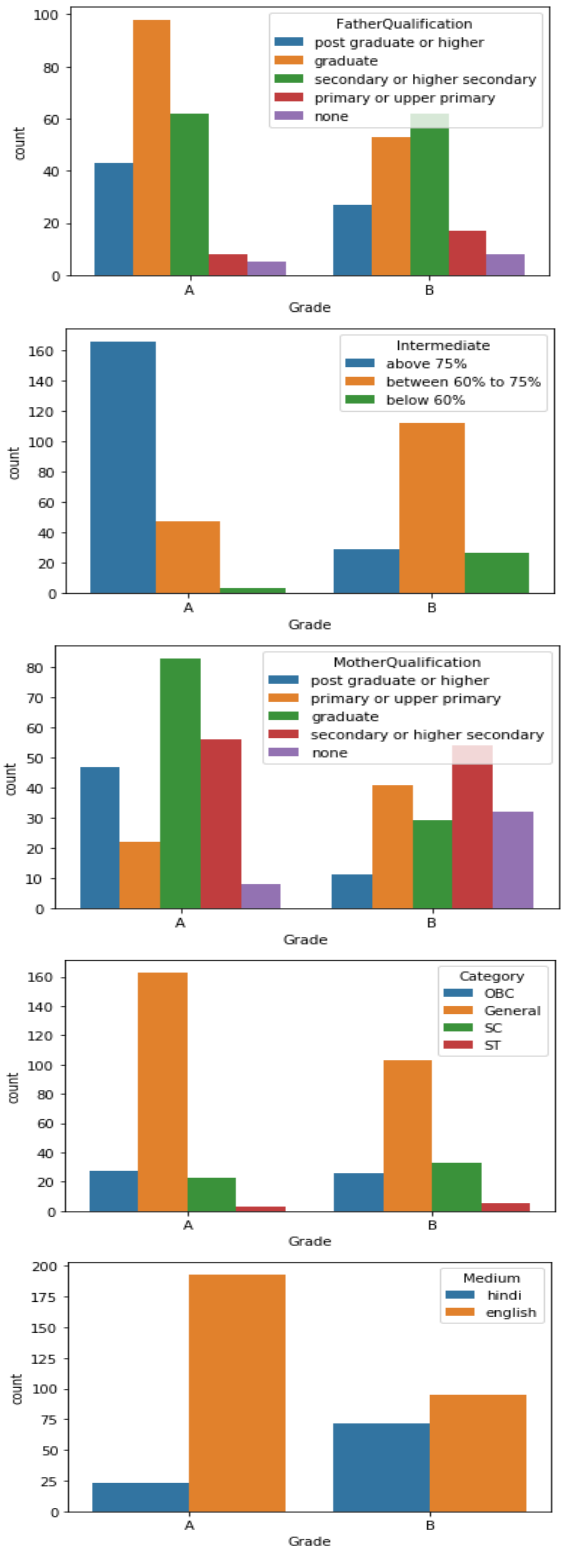

Fig. 4. Effect of most influential attributes on the final grade of the students.

TABLE XII: CONFUSION MATRIX FOR SVM

Actual/Predicted	A	B	Precision
A	37	6	0.82
B	8	26	0.81
Recall	0.86	0.76	

F. Result Using Proposed Model

Among all the five classification algorithms, Naïve Bayes achieved the highest accuracy, recall, and precision for predicting poor performers. So, to further improve the results, three Naïve Bayes classifiers were integrated with the help of Bagging and about 91% accuracy was achieved. The results of the proposed model are shown in Table XIII. It is shown that the proposed ensemble classifier achieved better results than any other classifier used in the present study. Its recall and precision for class B are also high, so it can recognize poor performers efficiently.

TABLE XIII: CONFUSION MATRIX FOR PROPOSED MODEL

Actual/Predicted	A	B	Precision
A	40	3	0.91
B	4	30	0.91
Recall	0.93	0.88	

It is pertinent to mention here that other combinations of single classifiers have also been implemented, but the achieved accuracy was not more than the accuracy of a single Naïve Bayes classifier, i.e., 89%. Further, during integration with the number of Naïve Bayes estimators more than 3, the achieved accuracy was not increased for the present dataset. The results obtained for accuracy, precision, and recall with different classification algorithms are shown in Fig. 5. It could be seen from Fig. 5 that the proposed model performed very well in comparison with all the other classifiers and achieved the highest accuracy (91%). Further, the proposed model has the highest precision (91%) and the highest recall (88%) for class B (poor performers), i.e., it identifies low performers efficiently, which was the main goal of the present work.

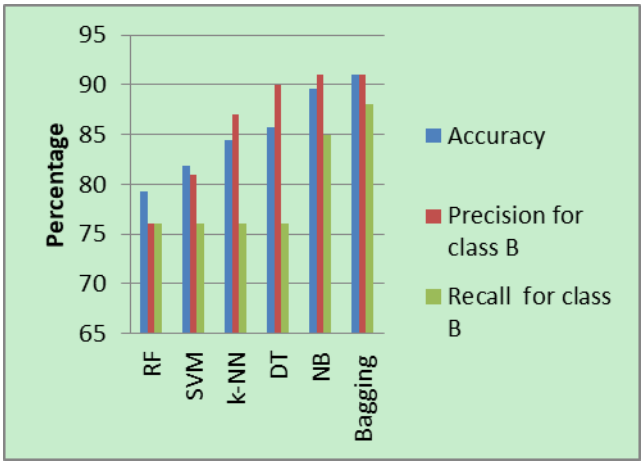

Fig. 5. Comparison of classifiers.

V. CONCLUSION AND FUTURE WORK

From the present study, it may be concluded that using the chi-square feature selection technique, only six features, viz., 10th standard percentage, 12th standard percentage, medium mothers' highest qualification, category, and fathers' highest

qualification have a significantly high impact on student's future academic performance. From these results, the following conclusions may be drawn:

1) Parents' education especially mothers' education, affects the academic performance of the students, so to increase the success percentage in education all over the country, more emphasis should be given to women's education, as women's literacy rate is lower in many parts of the world.
2) At the starting stage of the course, there is a need for extra classes or bridge courses in English as the students having mediums other than English are low performers. Also, there is a lack of engineering books in local languages, so students also find the problem in understanding concepts. Thus, promoting the authors to write engineering books in the local languages can make it easy for students to understand the concepts in their local languages.
3) As the 10^{th} standard percentage and the 12^{th} standard percentage affect students' success, it is justifiable to give some weightage to the 10^{th} standard and 12^{th} standard marks for admission to higher education, and they can also be used as a criterion for giving admission in the situations like the Covid-19 pandemic when it is difficult to conduct an entrance exam.
4) Indian society is divided into different categories, and in general, some categories are socially and economically weaker. So, there is a need of special support for the students belonging to the socially and economically weaker category.

Further, this study applied five classifiers, viz., Decision Tree, k-NN, SVM, Random Forest, and Naïve Bayes for predicting low performers at an entry level, and their efficiency was evaluated and compared. All the classifiers were able to predict low performers, and among all, the Naïve Bayes classifier gave the highest accuracy of about 89 %. Moreover, an ensemble-based model is proposed and implemented to improve the results by integrating three Naïve Bayes classifiers. The present study may conclude that the proposed model achieved the highest accuracy, precision, and recall for predicting poor-performing students. The present study shows that the ensemble model performs better than the single base classifier and is consistent with the findings of previous researchers [28]-[30].

The limitations of this study are that the size of the dataset was limited and belongs to only one institute. So, to study the effects of different attributes on the students' academic performance in-depth and in general perspective, the present work points out the need for a combined study with a large sample size of different territories students. The present study was also limited to only engineering degree students, but it could be extended to all higher education courses.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Swati Verma conducted the research, analyzed the data, proposed the methodology, and wrote the initial draft; Rakesh Kumar Yadav supervised the research and modified the initial draft; Kuldeep Kholiya has written the final version of the manuscript. All authors had approved the final version.

Acknowledgment

The authors of this article would like to thank all participants who provided data for this work.

References

[1] A. Buldu, and K. Üçgün, "Data mining application on students' data," *Procedia Social and Behavioral Sciences*, vol. 2, pp. 5251–5259, 2010.
[2] B. K. Bhardwaj and S. Pal, "Data mining: A prediction for performance improvement using classification," *International Journal of Computer Science and Information Security*, vol. 9, no. 4, pp. 136-140, 2011.
[3] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 61-72, 2013.
[4] A. K. Pal and S. Pal, "Analysis and mining of educational data for predicting the performance of students," *International Journal of Electronics Communication and Computer Engineering*, vol. 4, no. 5, pp. 1560-1565, 2013.
[5] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers & Education*, vol. 61, pp. 133–145, 2013.
[6] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict Students' academic performance using Ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119-136, 2016.
[7] A. K. Hamoud, A. M. Humadi, W. A. Awadh, and A. S. Hashim, "Students' success prediction based on bayes algorithm," *International Journal of Computer Application*, vol. 178, no. 7, pp. 6-12, 2017.
[8] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177-194, 2017.
[9] E. B. Costa, B. Fonseca, M. A. Santana, F. Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming course," *Computers in Human Behavior*, vol. 73, pp. 247-256, 2017.
[10] A. Pavithra and S. Dhanaraj, "Prediction accuracy on academic performance of students using different data mining algorithms with influencing factors," *International Journal of Scientific Research in Computer Science Applications and Management Studies*, vol. 7, no. 5, pp. 1-7, 2018.
[11] C.C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, 2019.
[12] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, pp. 1-20, 2019.
[13] D. T. Ha, C. N. Giap, P. T. T. Loan, and N.T. L. Huong, "An empirical study for student academic performance prediction using machine learning techniques," *International Journal of Computer Science and Information Security*, vol. 18, no. 3, pp. 21-28, 2020.
[14] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, pp. 1-18, 2020.
[15] P. Dixit, H. Nagar, and S. Dixit, "Student performance prediction using case based reasoning knowledge base system (CBR-KBS) based data mining," *International Journal of Information and Education Technology*, vol. 12, no. 1, pp. 30-35, 2022.
[16] M. Ashraf, M. Zaman, and M. Ahmed, "Using ensemble StackingC method and base classifiers to ameliorate prediction accuracy of pedagogical data," *Procedia Computer Science*, vol. 132, pp. 1021-1040, 2018.
[17] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, pp. 1-16, 2020.
[18] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, pp. 1-20, 2021.
[19] M. Yagci, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, pp. 1-19, 2022.

[20] A. Hussain, M. Khan, and K. Ullah, "Student's performance prediction model and affecting factors using classification techniques," *Education and Information Technology*, pp. 1-8, 2022.
[21] H. Yuliansyah, R. A. Imaniati, A. Wirasto, and M. Wibowo, "Predicting students graduate on time using C4.5 algorithm," *Journal of Information Systems Engineering and Business Intelligence*, vol. 7, no. 1, pp. 67-73, 2021.
[22] S. Goundar, A. Deb, G. Lal, and M. Naseem, "Using online student interactions to predict performance in a first-year computing science course," *Technology, Pedagogy and Education*, 2022.
[23] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, pp. 1-27, 2021.
[24] R.Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting student's performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899-67911, 2020.
[25] A. Ghavidel, R. Ghousi, and A. Atashi, "An ensemble data mining approach to discover medical patterns and provide a system to predict the mortality in the ICU of cardiac surgery based on stacking machine learning method," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1-11, 2022.
[26] C. W. Teoh, S. B. Ho, K. S. Dollmat, and C. H. Tan, "Ensemble-learning techniques for predicting student performance on video-based learning," *International Journal of Information and Education Technology*, pp. 1-5, 2022.
[27] H. Li and J. Sun, "Forecasting business failure: The use of nearest-neighbor, support vector and correcting imbalanced samples — Evidence from the Chinese hotel industry," *Tourism Management*, vol. 33, no. 3, pp. 622-634, 2012.
[28] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471-1483, 2020.
[29] M. Ragab, A. M. K. A. Aal, A. O. Jifri, and N. F. Omran, "Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1-9, 2021.
[30] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of undergraduate student's study completion status using missforest imputation in random forest and XGBoost models," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 13, no. 1, pp. 53-62, 2022.

Swati Verma was born on June 11, 1984 in Uttarakhand, India. She has obtained B. Tech. and M. Tech. degree from G.B.P.U.A. & T., Pantnagar, India and Banasthali Vidyapith, Rajasthan, India, respectively, in computer science and engineering. Currently, she is assistant professor, Department of Computer Science and Engineering, B.T.K.I.T., Dwarahat, India and pursuing the Ph.D. from IFTM University Moradabad, Uttar Pradesh, India. She has published many research papers in different reputed journals and conferences. Her research interest includes data mining, machine learning, soft computing and artificial intelligence.

Rakesh Kumar Yadav completed the bachelor of technology from Uttar Pradesh Technical University, UP, the master of technology from Singhaniya University, Rajasthan and the Ph.D. from IFTM University, UP. He is currently working as an assistant professor in the Department of Computer Science & Engineering, IFTM University, UP. He has published more than 30 research papers in reputed international journals and conferences. He has been reviewed many research papers of journals and conferences. He has published many patents, written many books and examination series also. His main research work focuses on machine learning, biometric, image processing, computer vision, soft computing and artificial intelligence.

Kuldeep Kholiya was born on November 15, 1982 in Uttarakhand, India. He has obtained Ph.D. degree from G.B.P.U.A. & T., Pantnagar, with major subject physics and minor subject computer science. He is currently working as an assistant professor in the Department of Applied Science, B.T.K.I.T., Dwarahat, India and having several academic and administrative responsibilities such as, associate dean academics, associate examination controller etc. He has published more than 25 research papers in reputed international journals and conferences. His research interest includes nano materials, phase transition, condensed matter theory and educational data mining.