A Scalable Machine Learning-based Ensemble Approach to Enhance the Prediction Accuracy for Identifying Students at-Risk

Swati Verma[1], Rakesh Kumar Yadav[2], Kuldeep Kholiya[3]
Research Scholor, IFTM University Moradabad, Uttar Pradesh, India[1]
Assistant Professor, IFTM University Moradabad, Uttar Pradesh, India[2]
Assistant Professor, B.T. Kumaon Institute of Technology, Dwaraht, Uttarakhand, India[3]

***Abstract*—Among the educational data mining problems, the early prediction of the students' academic performance is the most important task, so that timely and requisite support may be provided to the needy students. Machine learning techniques may be used as an important tool for predicting low-performers in educational institutions. In the present paper, five single-supervised machine learning techniques have been used, including Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression. To analyze the effect of an imbalanced dataset, the performance of these algorithms has been checked with and without various resampling methods such as Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, SVM-SMOTE, and Adaptive Synthetic (ADASYN). The Random hold-out method and GridSearchCV were used as model validation techniques and hyper-parameter tuning respectively. The results of the present study indicated that Logistic Regression is the best performing classifier with every balanced dataset generated using all of the four resampling techniques and also achieved the highest accuracy of 94.54% with SMOTE. Furthermore, to improve the prediction results and to make the model scalable, the most suitable classifier was integrated with the help of bagging, and a well-accepted accuracy of 95.45% was achieved.**

Keywords—Educational data mining; resampling methods; feature selection technique; machine learning; imbalanced data

I. INTRODUCTION

Due to the digitization and use of technology in the educational field, there is a large amount of educational data. Educational Data Mining helps to analyze and extract useful information, such as selecting the factors that affect the students' performance, predicting students' performance, etc., from a large amount of educational data. As students or youths are the future of any nation, predicting the success rate of students in their academic area is a very important and beneficial task. This may be achieved with the help of educational data mining, which utilizes various machine learning techniques.

Although the field of Educational Data Mining (EDM) is old and its definition was given by Fayyad et al. [1] in 1996, EDM emerged as a convincing research area after the establishment of the annual International Conference on Educational Data Mining and the Journal of Educational Data Mining in 2008 [2]. After that, Baker [3] identified the application of data mining in education to discover models for predicting students' performance by using the methods of prediction, clustering, relationship mining, and discovery with models. Among the applications of EDM, detecting student failure at an early stage has been an appealing research topic for researchers due to its social impact. The prediction of the students at risk of being dropouts from an institute or school becomes difficult due to the large number of factors that may influence the academic performance of the students. Thus, it is quite important to predict low-performing students at an early stage with higher accuracy, along with the important factors that may affect their performance.

To achieve this goal, the present study has three important research objectives: (i) to identify the influential features by using a filter-based feature selection technique. (ii) to identify the best performing classifier by comparing various single-supervised machine learning techniques, viz., decision trees, Naïve Bayes, k-Nearest Neighbor, Logistic Regression, and Support Vector Machine with various resampling techniques such as random oversampling, SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN. (iii) to enhance the prediction rate of the students at-risk by using an ensemble model that integrates the most suitable data mining technique.

In rest of the paper, the work related to the present study is given in section II. The methodology used in the present work is explained in Section III. In Section IV, the obtained results are analyzed and discussed. Finally, the conclusion and future work are given in Section V.

II. RELATED WORK

In the past, various review studies have been performed on educational data mining [4, 5], and many researchers have worked on identifying the factors that deteriorate the academic performance of students. Ahmed et al. [6] selected nine attributes such as department, attendance, high school degree, mid-term marks, student participation, lab test grades, assignment scores, seminar performance, and homework to predict the final grade and generate the rules set by the Decision Tree. Tomasevic et al. [7] have compared the performance of several data mining algorithms using past student performance, student engagement, and student demographic data. They concluded that students' engagement and past performance data have a significant influence, while demographic attributes have a slight impact, on students' performance. Further, Verma and Yadav [8] used the cross-tabulation method and the chi-square test to analyze the effects of different attributes such as background, academic, social, and psychological characteristics on students' academic performance. In their finding, it was concluded that students' academic and background attributes were the most influential factors that may affect students' grades.

With knowledge of the factors that influence the students' performance, predictions can be made with the help of data mining algorithms to identify students at risk. To analyze students' performance, Asif et al. [9] implemented decision tree and clustering technique on a dataset of 210 students that contained pre-admission marks and all subjects' marks and found that the pre-university marks and subjects' marks in the first and second years had an impact on students' final year marks. Hamoud et al. [10] applied Bayesian classifiers, namely Naïve Bayes and Bayes Net, to the dataset of 161 students and found that Naïve Bayes outperformed for predicting the students' performance. Costa et al. [11] 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. Moreover, Ha et al. [12] implemented rule-based learners, neural-based learners, and statistical-based learners (Naïve Bayes, and Support Vector Machine) on students' datasets, which consist of personal and past academic information, to predict students' performance. In their experiment, neural-based learners and Naïve Bayes achieved the highest accuracy of 86.19%.

A suitable approach towards feature selection and handling imbalanced class problems may enhance the prediction accuracy of machine learning models. Thammasiri et al. [13] compared random oversampling and SMOTE balancing methods along with four popular data mining models: logistic regression, decision trees, neural network, and support vector machine to assess the students' performance. In their results, Support Vector Machine (SVM) achieved the highest accuracy of 90.24% with SMOTE. Mueen et al. [14] applied Naïve Bayes, Neural Network, and Decision Tree to students' data having their general, academic, and forum-related variables along with feature selection and SMOTE oversampling method to solve the imbalanced data problem and found Naïve Bayes to be outperformed with 86% accuracy. Ghorbani and Ghousi [15] used and compared different resampling methods, viz., Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek, by evaluating the performance of the various classifiers, and Random Forest obtained the highest accuracy of 81.27% with SVM-SMOTE. Further, Ghavidel et al. [16] solved the problem of imbalanced data by using a combination of the SVM-SMOTE (an over-sampling technique) and Edited-Nearest-Neighbor (an under-sampling technique) while predicting disease mortality. Recently, Desiani et al [17] applied k-Nearest Neighbor (k-NN), Artificial Neural Network (ANN), and C4.5 to students' educational background records along with SMOTE to make the dataset balanced, and that balanced dataset increased the accuracy of prediction, and for k-NN the maximum achieved accuracy was 83.71%.

Another aspect that enhances the prediction accuracy is the appropriate use of ensemble models. Teoh et al. [18] used feature selection and SMOTE oversampling techniques and then applied various ensemble machine learning methods, namely stacking, boosting, and bagging. In their findings, AdaBoost has achieved a maximum accuracy of more than 90%.

Although there are several studies to predict the students' academic performance, the study which considers all categories of variables, i.e., background, academic, social, and psychological, and predicts students at-risk at an early stage with adequate accuracy is lacking. Also, a single classifier-based prediction is not suitable from one perspective to another. Moreover, a classifier giving the highest prediction accuracy for a particular dataset may not be valid for a different dataset. Thus, the aim of the present study is to identify low performers at an early stage with a higher prediction rate by using a scalable approach.

III. METHODOLOGY

The main objective of the present paper is to predict the academic performance of students with higher accuracy. To achieve this goal, the different single supervised machine learning algorithms were applied with and without data balancing, and finally, by comparing the results, a model was constructed to enhance the prediction accuracy. The methodology applied in the present work may be given as follows:

- Dataset preparation.
- Data preprocessing including data transformation, feature selection, and data balancing.
- Identification of the best classification technique by comparing the results of classification models when applied to the preprocessed data.
- Make a scalable ensemble model with the help of the best classification technique.
- Result evaluation of the proposed ensemble model.

The workflow of the proposed methodology is given in Fig. 1.

A. Dataset

To make the data versatile, it is collected from the two different engineering colleges situated in different regions (the north and south of India). In the present paper, the sample size comprises 550 engineering students from two different engineering colleges in India, i.e., Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand, and Cochin University of Science & Technology, Trivandrum, Kerala. The dataset includes information regarding background, past academic, social, and psychological factors with 30 different attributes, of which three attributes (roll-number, name, and branch) are used for identification purposes only and do not play any role in the prediction of low-performers. So, only 27 attributes were used for the present work, with first semester GPA as the output variable. For these attributes, data was collected online with the help of a multiplechoice questionnaire created via outsourced technology, i.e., Google Form. As the aim of the paper is to identify the students having the highest risk of dropping out of college, the information about the output attribute for the dataset is divided only into two categories, i.e., low performers and high performers, based on the first-semester grade point of the students.

B. Data Preprocessing

Before applying any machine learning model to the dataset, data should be preprocessed so that any machine learning model can be performed efficiently. In the present study, the dataset is complete and free from noise, so there is no need to handle missing data and outliers. To preprocess the data, data transformation, feature selection, and data balancing have been performed.

1) Data transformation: In the present study, all the features were categorical except students' GPA as it was initially in numerical form. So, GPA was generalized into categorical values, i.e., "class A (high performer)" and "class B (low performer)". Finally, these categorical variables were encoded into the suitable format of machine learning models.

2) Feature selection: Feature selection is an important part of the students' performance prediction model for two main reasons:

- The main purpose of the prediction of students' academic performance is to provide timely support to the low-performing students in the area where they are lacking. Only after identifying the attributes that have a significant impact on the output variable, i.e., students' academic performance, suitable corrective measures may be taken to provide support to the low-performing students.
- With the help of feature selection, irrelevant attributes may be removed from the data without losing reliability in classification. Thus, the dimensionality reduction raises the processing speed, and hence the classifier can learn faster.

There are three main feature selection techniques: manual selection based on pedagogical theories or expert experience; filter-based selection; and wrapper feature selection [19]. In the present study, as all the attributes were categorical, a filter-based feature selection technique, namely "chi-square", was used by which p-values were calculated for each attribute [8]. The attributes having a p-value of less than 0.01 show a highly significant correlation with the student's grades.

3) Data balancing: Data balancing is an important part of preprocessing step by which class distribution have to make equal so that classifier do not assign every new sample to the majority class only. In the present study the distribution of "class A" and "class B" is shown in Fig. 2. From the figure, it may be revealed that the dataset contained more samples from "class A" (66%) than the "class B" (34%). Previous study [20] shows that if the percentage of minority class is less than 35% of dataset then it is called imbalanced and hence the dataset of present study is imbalanced to some extent. There are mainly three types of re-sampling techniques i.e., over-sampling, under-sampling, and hybrid-sampling [15] that may be used to balance the dataset. Due to the limited size of dataset, in the present study, only over-sampling techniques i.e., Synthetic Minority Oversampling Technique (SMOTE) [21], Borderline SMOTE [22], SVM-SMOTE [23], and ADASYN [24] were used and compared.

C. Machnie Learning Techniques

There are different types of classification machine learning models that may be used to predict the students' academic performance. In the present study, five single supervised machine learning models have been applied, including Decision Tree [25], Naïve Bayes [9, 26], k-Nearest-Neighbor [27], Support Vector Machine [28], and Logistic Regression [29]. To achieve the best performance of these machine learning models, the passing parameters for these models were set with the help of an algorithm called "GridSearchCV" which gives the best combination of passing parameters [30]. These combinations of passing parameters are listed in Table I.

TABLE I. CLASSIFICATION MODELS AND THEIR PASSING PARAMETERS

Machine learning model	Passing parameters
Decision Tree	Criterion="gini", max_depth=4, max_leaf_node=8
Naïve Bayes	No parameter
k-Nearest Neighbor	n_neighbor=21
Support Vector Machine	c=2, kernel="rbf"
Logistic Regression	No parameter

D. Model Validation and Result Evaluation

Model validation is used to check the effectiveness of the model across independent datasets. In the present study, the random hold-out method was used for model validation, in which 80% of the data was for training purposes and 20% of the data was reserved for testing purposes.

Furthermore, the performance of all the machine learning techniques was evaluated in terms of accuracy, precision, recall, and f1-score. These performance metrics are given as follows:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ number\ of\ samples} \tag{1}$$

$$Precision = \frac{True\ Positive}{Total\ classes\ predicted\ as\ positive} \tag{2}$$

$$Recall = \frac{True\ Positive}{Total\ number\ of\ actual\ positive\ classes} \tag{3}$$

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

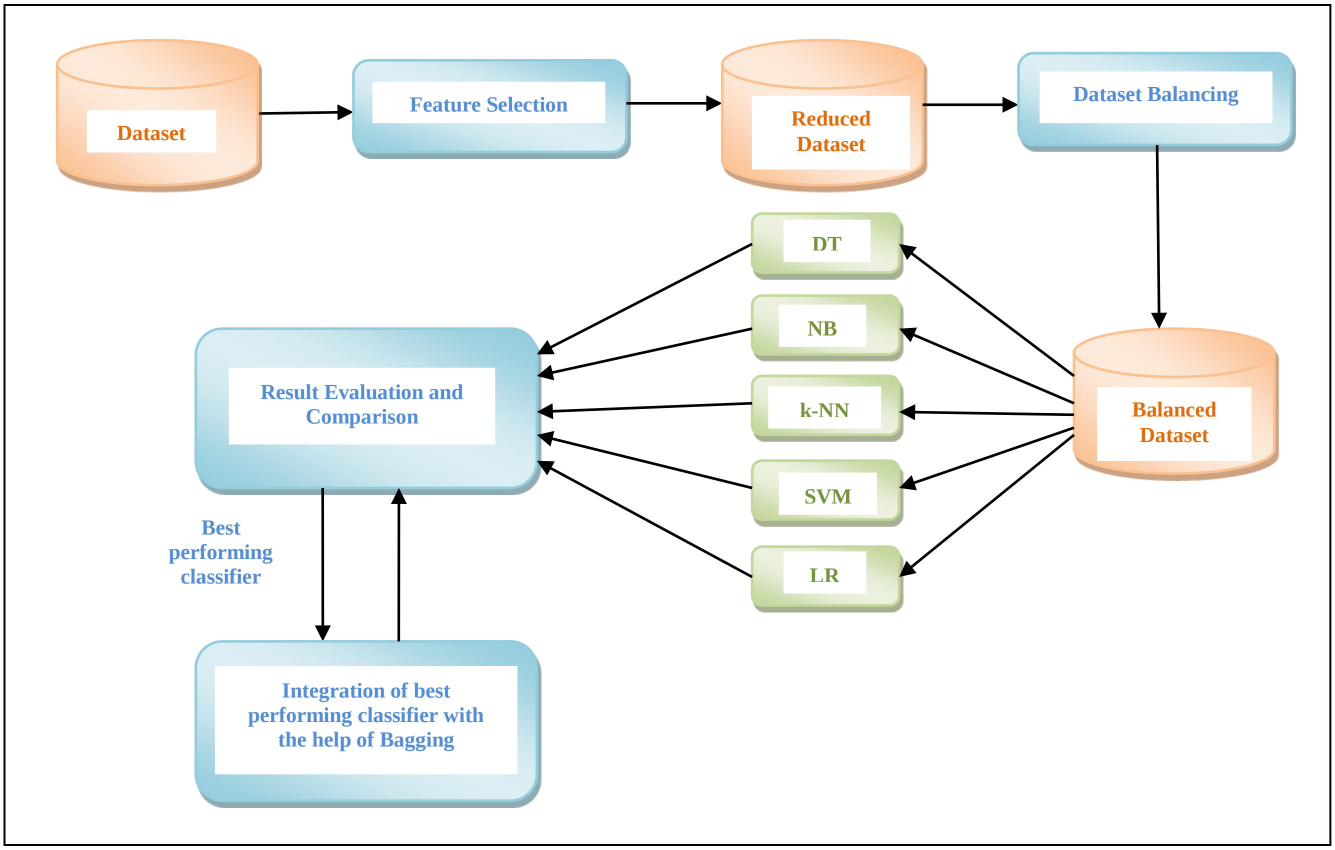
Dataset
Feature Selection
Reduced Dataset
Dataset Balancing
DT
NB
k-NN
SVM
LR
Balanced Dataset
Result Evaluation and Comparison
Best performing classifier
Integration of best performing classifier with the help of Bagging

Fig. 1. Framework of Proposed Methodology.

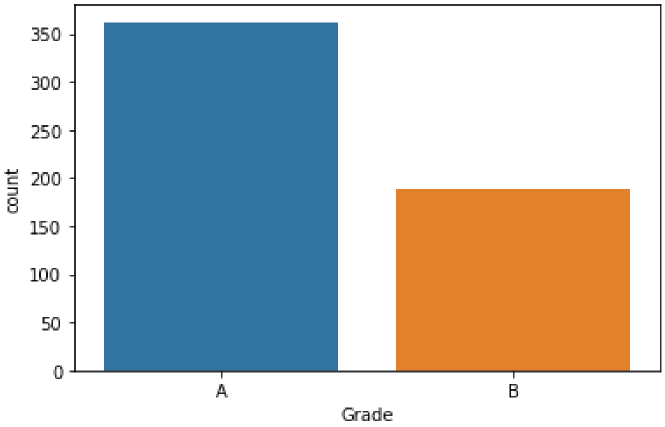

Fig. 2. Distributions of the Grades.

E. Construction of Ensemble based Classifier

In most of the previous studies [18, 31–38], it was shown that the ensemble model gives a higher prediction accuracy, so, to enhance the prediction accuracy, an ensemble model was constructed in the present study. For this, the best performing classifier was selected along with its suitable resampling method, after comparing the results of different single machine learning algorithms with balanced dataset. Finally, in order to make an ensemble classifier, the three best-performing classifiers were integrated with the help of bootstrap aggregation.

IV. RESULT AND DISCUSSION

In the present work, the whole experiment was done with the help of different libraries such as Pandas, Seaborn, and Scikit-learn of the Python programming language, which is a very powerful and user-friendly language for data scientists. The first aspect of the present work is to find out the influential attributes and to reduce the dimensionality with the help of a filter-based feature selection technique. For this purpose, the p-values were calculated for different attributes using the chi2 method of the sklearn.feature_selection library of Python programming and are shown in Table II. From this table, it is depicted that after applying the feature selection technique, the following 11 features are selected as influential features that affect students' academic performance: percentage in 10th standard, percentage in 12th standard, confidence, mathematics % in 12th standard, punctuality, curiosity, medium/language of previous study, category, father's highest qualification, mother's highest qualification, and mental stress.

After selecting the most influential attributes, Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression algorithms have been applied to the dataset, which contains only the 11 selected most influential attributes. The results obtained for accuracy, precision, recall, and f1-score of these algorithms are represented in Table III.

TABLE II. STUDENTS' RELATED INPUT FEATURES AND THEIR CORRESPONDING P-VALUES

Attribute Category	Attribute	p-value
Background Attributes	Gender	.0304
	Category	8.1425e-05
	Number of Siblings	.5330
	Status of Parent	.4112
	Father's Highest Qualification	.0001
	Mother's Highest Qualification	.0027
	Father's Occupation	.8812
	Mother's Occupation	.8034
	Annual Family Income	.2393
	Living Location	.1042
	Medium/Language of Previous Study	4.4161e-06
Academic Attributes	Percentage in 10th standard	1.0815e-42
	Percentage in 12th standard	1.3151e-35
	Entrance Exam/JEE Rank	.1319
	Average Self-Study Time	.0407
	Mathematics % in 12th standard	2.9478e-29
Social Attributes	Participation in Extra-Curricular Activities	.4782
	Whether have Friends	.9547
Psychological Attributes	Motivation to Join Course	.9281
	Mental stress	.0033
	Homesickness	.2046
	Personality	.1333
	Adaptability	.4372
	Confidence	6.5301e-33
	Curiosity	2.0818e-09
	Punctuality	5.1669e-14

TABLE III. RESULTS OF THE CLASSIFIERS ON IMBALANCED DATASET

Classifier	Accuracy (in %)	Recall		Precision		F1-score	
		A	*B*	*A*	*B*	*A*	*B*
Decision Tree	91.81	0.99	0.79	0.90	0.97	0.94	0.87
Naïve Bayes	88.18	0.89	0.87	0.93	0.80	0.91	0.84
k-Nearest Neighbor	89.09	0.94	0.79	0.89	0.88	0.92	0.83
Support Vector Machine	90.90	0.99	0.76	0.89	0.97	0.93	0.85
Logistics Regression	92.72	0.99	0.82	0.91	0.97	0.95	0.89

From Table III, it may be observed that the highest accuracy, i.e., 92.72%, was achieved with Logistic Regression. In terms of recall and precision for classes A and B, no single algorithm can be declared best. This is because precision and recall for classes A and B are not the highest for the same algorithm. For example, in Naïve Bayes recall and precision for class B and class A is highest, respectively, but recall for class A and precision for class B is lowest. In such situations, the f1-score may be taken as an evaluation criterion, as the f1-score is the harmonic mean of precision and recall. Logistic Regression has achieved the highest accuracy and highest f1-score for both classes 'A' and 'B', and hence it may be considered the best performing algorithm with the imbalanced dataset. The dataset of the present study was imbalanced, and hence four resampling techniques (SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN) have been used, and the performance of all the classifiers was evaluated with the balanced dataset.

The performances of different models with the different resampling methods are shown in Table IV. From Table IV, it may be noted that the accuracy of the models, except for Logistic Regression, was not significantly improved when applied to the balanced dataset. This may be because of the fact that, in the case of balanced data, all the algorithms considered both the classes "A" and "B" with equal weightage. So, it may be concluded that although in the case of balanced datasets, the accuracy of every classifier is not increasing; the prediction accuracy may now be trustable and sufficient to measure the model's performance. The performances of various classifiers using the resampling methods SMOTE, Borderline SMOTE, SVM-SMOT, and ADASYN are shown in Fig. 3-6 respectively. From these figures, it may be observed that Logistic Regression outperformed all the classifiers in every balanced dataset generated with all the four resampling techniques, and the highest accuracy of 94.54% and the highest F1-score were achieved when SMOTE was considered as a resampling method.

TABLE IV. RESULTS OF THE CLASSIFIERS ON BALANCED DATASET

Classifier	Evaluation Metric		SMOTE	Borderline SMOTE	SVM- SMOTE	ADASYN
Decision Tree	Accuracy (in %)		89.09	88.18	88.18	91.81
	Recall	A	0.89	0.86	0.89	0.92
		B	0.89	0.92	0.87	0.92
	Precision	A	0.94	0.95	0.93	0.96
		B	0.81	0.78	0.80	0.85
	F1-score	A	0.91	0.91	0.91	0.94
		B	0.85	0.84	0.84	0.89
Naïve Bayes	Accuracy (in %)		80.90	83.63	86.36	83.63
	Recall	A	0.86	0.83	0.85	0.83
		B	0.71	0.84	0.89	0.84
	Precision	A	0.85	0.91	0.94	0.91
		B	0.73	0.73	0.76	0.73
	F1-score	A	0.86	0.87	0.89	0.87
		B	0.72	0.78	0.82	0.78
k-Nearest Neighbor	Accuracy (in %)		85.45	82.72	83.63	81.81
	Recall	A	0.86	0.79	0.79	0.78
		B	0.84	0.89	0.92	0.89
	Precision	A	0.91	0.93	0.95	0.93
		B	0.76	0.69	0.70	0.68
	F1-score	A	0.89	0.86	0.86	0.85
		B	0.80	0.78	0.80	0.77
Support Vector Machine	Accuracy (in %)		90.90	90.00	89.09	90.90
	Recall	A	0.96	0.92	0.92	0.94
		B	0.82	0.87	0.84	0.84
	Precision	A	0.91	0.93	0.92	0.92
		B	0.91	0.85	0.84	0.89
	F1-score	A	0.93	0.92	0.92	0.93
		B	0.86	0.86	0.84	0.86
Logistics Regression	Accuracy (in %)		94.54	90.90	91.81	93.63
	Recall	A	0.99	0.93	0.94	0.97
		B	0.87	0.87	0.87	0.87
	Precision	A	0.93	0.93	0.93	0.93
		B	0.97	0.87	0.89	0.94
	F1-score	A	0.96	0.93	0.94	0.95
		B	0.92	0.87	0.88	0.90

TABLE V. RESULTS OF THE PROPOSED MODEL

Classifier	Evaluation Metric		Imbalanced dataset	SMOTE	Borderline SMOTE	SVM- SMOTE	ADASYN
Proposed Model	Accuracy (in %)		93.63	95.45	93.63	93.63	94.54
	Recall	A	0.99	0.99	0.96	0.96	0.97
		B	0.84	0.89	0.89	0.89	0.89
	Precision	A	0.92	0.95	0.95	0.95	0.95
		B	0.97	0.97	0.92	0.92	0.94
	F1-score	A	0.95	0.97	0.95	0.95	0.96
		B	0.90	0.93	0.91	0.91	0.92

Finally, after evaluating the performance of all classifiers, the best performing classifier, namely Logistic Regression, was chosen to create the ensemble model in order to improve prediction accuracy. In order to make the ensemble model, three Logistic Regression classifiers were integrated with the help of bagging. The result of the proposed integrated model is shown in Table V. The proposed model has achieved the highest accuracy of 95.45%, the highest prediction rate for low performers, and the highest f1-score for both classes while using SMOTE. It is pertinent to mention here that the accuracy of the proposed model increased by 1.82% after using the resampling technique SMOTE, while in the study of Desiani et al., the average accuracy was increased by 20.13%. The possible reason may be that the dataset used in the present study has a small sample size and was not highly imbalanced. In the case of a large sample size, the number of students at risk will be significantly lower, and hence, in such situations of highly imbalanced data, the present model may be quite useful.

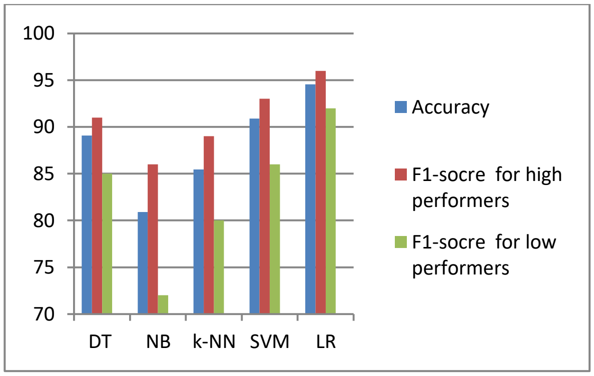

Fig. 3. Performance of Different Classifiers with SMOTE.

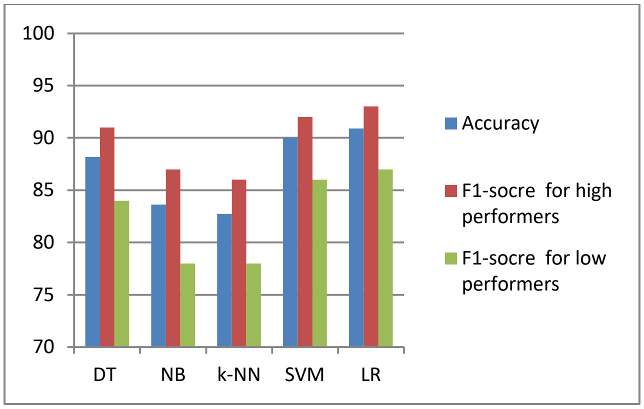

Fig. 4. Performance of Different Classifiers with Borderline SMOTE.

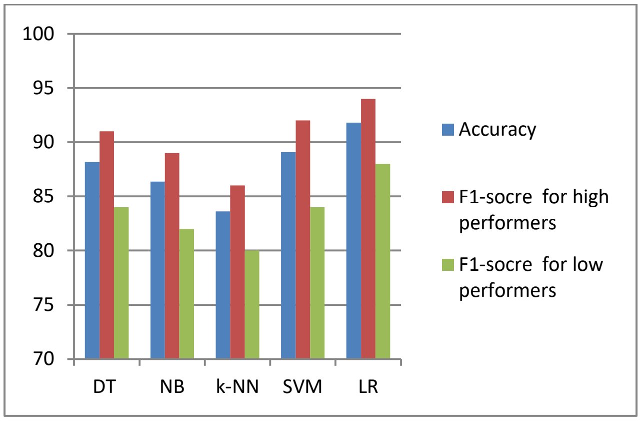

Fig. 5. Performance of Different Classifiers with SVM-SMOTE.

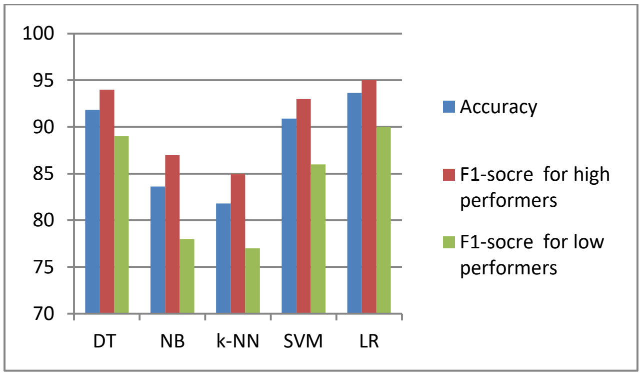

Fig. 6. Performance of Different Classifiers with ADASYN.

The highest prediction accuracy achieved in the present study is 95.45%, which is greater than most of the previous studies [12-18]. Along with the enhanced prediction accuracy, the main advantage of the present work is that the methodology proposed in the present study is scalable from one context to the other.

V. Conclusion and Future Work

From the present work, it may be concluded that students' past academic performance (10th standard %, 12th standard %, and Math's % in the 12th standard), their background (category, parents' qualification, and medium of the previous study), and their psychological features (mental stress, confidence, curiosity, and punctuality) were the relevant attributes. Thus, to increase the academic performance of the students, these factors may be considered as the focus points.

In the present study, all the used classifiers were able to predict students' outcomes with reasonable accuracy of more than 80%. Among all the used classifiers, Logistic Regression was the best performing algorithm with a balanced as well as an imbalanced dataset. Further, the accuracy and prediction rate for identifying low performers as well as for high performers were improved when the Logistic Regression was applied to the balanced dataset. The prediction accuracy was further enhanced with the use of an ensemble classifier in which three Logistic Regression classifiers (because of its highest performance) were integrated with the help of bootstrap aggregation. The proposed integrated model has achieved the highest accuracy of 95.45% and the highest precision and recall for low performers with the balanced dataset formulated with the help of the resampling technique SMOTE.

It should be noted that with different datasets, the different classifiers may give the highest prediction accuracy, and hence there is a need for the methodology to be scalable for every situation. Thus, the main advantage of the present approach is its scalability for different datasets. Further, this study may also be applied to the different domains of data mining and machine learning applications for enhancing prediction accuracy. The limitation of the present study is that the examined dataset has a small sample size and slightly imbalanced data, so in the future, the proposed methodology should be used with large sample sizes and highly imbalanced data for the prediction of students' academic performance.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI magazine, Vol. 17, pp. 37-53, 1996.

[2] R. S. Baker, "Educational Data Mining: An Advance for Intelligent Systems in Education", IEEE Intelligent systems, 29, 78–82, 2010.

[3] R. S. J. D. Baker. "Data Mining for Education", International Encyclopaedia of Education, 3rd edition, Vol. 7, pp. 112-118, 2010.

[4] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques", Educational Sciences, Vol. 11, No. 9, pp. 1-27, 2021.

[5] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance", Engineering Reports, pp. 1-23, 2021.

[6] A. Ahmed, and I. Elaraby, "Data mining: A prediction for student's performance using classification method", World Journal of Computer Application and Technology, Vol. 2, pp. 43-47, 2014.

[7] 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.

[8] S. Verma, and R. K. Yadav, "Effect of Different Attributes on the Academic Performance of Engineering Students", ICATMRI, pp. 1-4, 2020.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition", Expert Syst. Appl., Vol. 41, No. 2, pp. 321-330, 2014.

[14] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques", Int. J. Mod. Educ. Comput. Sci., Vol. 8, No. 11, p. 36, 2016.

[15] 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.

[16] 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.

[17] A. Desiani, S. Yahdin, and A. Kartikasari, "Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment", International Journal of Artificial Intelligence, Vol. 10, No. 2, pp. 346-354, 2021.

[18] 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, Vol. 12, No. 8, pp. 741-745, 2022.

[19] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications", 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1200-1205, 2015.

[20] 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.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Reserach, Vol. 16, pp. 341-378, 2002.

[22] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new oversampling method in imbalanced data sets learning", Proc. Int. Conf. Intell. Comput. Berlin, Germany: Springer, pp. 878-887, 2005.

[23] Y. Tang, Y. Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification ", IEEE Trans. Syst., Man, Cybern, B. Cybern., Vol. 39, No. 1, pp. 281-288, 2009.

[24] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", IEEE World Congress on Computational Intelligence, pp. 1322-1328, 2008.

[25] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An Introduction to decision tree modeling", Journal of Chemometrics: A Journal of the Chemometrics Society, Vol. 18, No. 6, pp. 275-285, 2004.

[26] I. Rish, "An empirical study of the naive Bayes classifier", IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, No. 2, pp. 41-46, 2001.

[27] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, "The prediction of scholarship recipients in higher education using k-Nearest Neighbor algorithm", IOP conference series: material science and engineering, Vol. 434, No. 1, 2018.

[28] D. A. Pisner, and D. M.Schnyer, "Support vector machine in Machine learning", Machine learning Academic Press, pp. 101-121, 2020.

[29] S. F. Costa, and M. M. Diniz, "Application of logistic regression to predict the failure of students in subject of a mathematics undergraduate course", Education and Information Technology, pp. 1-7, 2022.

[30] S. Jeganathan, A. R. Lakshminarayan, and N. Ramchandran, "Predicting Academic Performance of Immigrant Students Using XGBoost Regressor", International Journal of Information Technology and Web Engineering, Vol. 17, No. 1, pp. 1-19, 2022.

[31] 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.

[32] 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.

[33] 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.

[34] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm", Interactive Learning ironments, pp. 1-20, 2021.

[35] 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.

[36] 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.

[37] 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.

[38] S. Begum, and S. S. Padmannavar, "Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction", International Journal of Intelligent Engineering & Systems, Vol. 15, No. 2, pp. 316-328, 2022.