Predicting Students' Academic Performance Through Machine Learning Classifiers: A Study Employing the Naive Bayes Classifier (NBC)

Xin ZHENG[1], Conghui LI[2]

College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China[1]

College of Management and Economics, North China University of Science and Technology, Tangshan 063210, China[2]

***Abstract*—Modern universities must strategically analyze and manage student performance, utilizing knowledge discovery and data mining to extract valuable insights and enhance efficiency. Educational Data Mining (EDM) is a theory-oriented approach in academic settings that integrates computational methods to improve academic performance and faculty management. Machine learning algorithms are essential for knowledge discovery, enabling accurate performance prediction and early student identification, with classification being a widely applied method in predicting student performance based on various traits. Utilizing the Naive Bayes classifier (NBC) model, this research predicts student performance by harnessing the robust capabilities inherent in this classification tool. To bolster both efficiency and accuracy, the model integrates two optimization algorithms, namely Jellyfish Search Optimizer (JSO) and Artificial Rabbits Optimization (ARO). This underscores the research's commitment to employing cutting-edge machine learning and algorithms inspired by nature to achieve heightened precision in predicting student performance through the refinement of decision-making and prediction quality. To classify and predict G1 and G3 grades and evaluate students' performance in this study, a comprehensive analysis of the information pertaining to 395 students has been conducted. The results indicate that in predicting G1, the NBAR model, with an F1_Score of 0.882, performed almost 1.03% better than the NBJS model, which had an F1_Score of 0.873. In G3 prediction, the NBAR model outperformed the NBJS model with F1_Score values of 0.893 and 0.884, respectively.**

***Keywords*—Machine learning; Naive Bayes Classifier; Artificial Rabbits Optimization; Jellyfish Search Optimizer; student performance**

I. Introduction

Education is the foundational pillar for any nation or society, embodying a crucial element that provides guidance, societal status, extensive knowledge, and avenues for exploration [1–3]. Modern universities must analyze performance, identify uniqueness, and develop a strategic plan. Management should prioritize understanding admitted students' diverse characteristics. In the competitive academic landscape, excellence in student performance is crucial for higher learning institutions [4–6]. Knowledge discovery (KD) involves extracting meaningful, unknown, and potentially valuable information from extensive databases. Data mining (DM) is crucial for educational data analysis, offering various methods for this purpose [7, 8]. The substantial amount of student data in databases exceeds the human capacity for manual analysis, necessitating automated techniques [9]. This includes creating early warning systems to reduce costs, save time, and optimize resources [10, 11]. Educational Data Mining (EDM) is a theory-oriented DM approach employed in academic and educational settings. Its primary goal is to create computational methods that integrate theory and data. The objective of EDM is to improve and support academic performance among students and graduates and to enhance the management of faculty information within educational institutions [12–14].

Machine learning (ML) algorithms serve as indispensable tools for knowledge discovery, playing a pivotal role in various applications. One of their crucial functions lies in accurate performance prediction, a capability that proves instrumental in identifying struggling students early. By leveraging these algorithms, educational institutions can proactively address academic challenges, fostering a more supportive and responsive learning environment [15]. Various machine learning methods, including classification [16], prediction [17, 18], and clustering [19], are continuously evolving and expanding the scope of data mining.

Classification, the most common and effective data mining approach for categorizing and predicting values, also applies to EDM [20]. In the context of student performance prediction, classification refers to grouping or classifying pupils based on specific traits or features. These traits include past academic achievement, demographic information, socioeconomic background, study habits, and other pertinent data. Patterns and links within the data are detected using classification algorithms, allowing predictions about future student performance to be generated [21–24]. For instance, using the ICRM classifier, Marquez-Vera et al. [25] tackled the intricate task of predicting student failure in academic contexts by utilizing a genetic programming algorithm and diverse DM approaches. Employing real data from 670 high school students in Zacatecas, Mexico, the research addressed challenges such as high dimensionality and imbalanced datasets. It strategically selected influential attributes, rebalances data, and incorporated cost-sensitive classification methods. Additionally, the study introduced a genetic programming model, comparing its interpretability and accuracy with other white box techniques. The ultimate goal was to identify the most effective approach for enhancing classification accuracy, particularly in predicting students at risk of failure. The findings contributed valuable insights to predictive modeling in education, offering nuanced perspectives on factors influencing student outcomes and guiding targeted interventions for improved educational support. Hu and Song [26] focused on the analysis and evaluation of student achievement as a crucial aspect of teaching and school management. The objective was to scientifically assess academic performance, providing accurate insights for teachers and enabling students to understand their learning situation. The research utilized the XGBoost algorithm for classifying and evaluating student performance through statistical analysis of basic data. A performance evaluation model was established, taking into account curriculum relevance by statistically compiling student performance data. The subjective and objective structural entropy weight method was employed to classify characteristic importance results, offering insights into relevant courses. Moreover, the XGBoost method was used to predict grades for unfinished courses based on completed course results. The study aimed to comprehensively, objectively, and reasonably evaluate students' learning situations, contributing valuable information for teaching management and improvement strategies. Kabakchieva et al. [27] aimed the outcomes of a data mining research conducted at a prestigious Bulgarian university. The primary objective was to showcase the substantial potential of data mining applications in university management, particularly in optimizing enrollment campaigns and attracting high-caliber students. The research focused on developing data mining models to predict student performance, utilizing personal, pre-university, and university-performance characteristics. The dataset encompassed information about students admitted to the university over three consecutive years. Various well-known data mining classification algorithms, including a rule learner, a decision tree classifier, a neural network, and a Nearest Neighbour classifier, were applied and their performances were analyzed and compared. The study aimed to contribute insights for more effective university management by employing data mining techniques to predict and understand student performance. By presenting a model for predicting poor academic performance among first-year students, Tamasiri et al. [28] aimed to address the challenging task of predicting student attrition, particularly dealing with class-imbalanced data common in the realm of student retention. The study contrasted four widely used classification techniques logistic regression, decision trees, neural networks, and support vector machines with three alternative data balancing strategies: over-sampling, under-sampling, and synthetic minority over-sampling ($SMOTE$). The research aimed to retain overall excellent classification performance while improving predicting accuracy for the minority class using large-scale institutional student data from 2005 to 2011. Based on the 10-fold holdout sample, the support vector machine and SMOTE data-balancing approach produced the best classification result, with an overall accuracy of 90.24% for the minority class, the three data-balancing strategies increased prediction accuracy. Additionally, sensitivity analyses identified crucial variables for accurately predicting student attrition, suggesting the potential application of these models to reduce student dropout rates by accurately identifying at-risk students. Marbouti et al. [29] utilized predictive modeling methods to early identify at-risk students in courses employing standards-based grading. The goal of the study was to modify at-risk prediction models to take use of standards-based grading, which offers advantages over traditional score-based grading in education. Prediction approaches were limited to using the course instructors' access to performance data from the previous semester. The study prioritized minimizing false negatives (type II errors) in identifying at-risk students without significantly increasing false positives (type I errors). To enhance generalizability and accuracy, a feature selection method was applied to reduce the number of variables in each model. Among the seven tested modeling methods, the Naive Bayes (NB) Classifier model and an Ensemble model showed the most promising results, contributing insights for more effective educational interventions in identifying at-risk students.

While various classification algorithms have received considerable attention in recent studies, Naive Bayes Classifier (NBC) has been comparatively less explored. This research introduces and evaluates NBC alongside two hybrid models optimized using Jellyfish Search Optimizer (JSO) and Artificial Rabbits Optimization (ARO). The study comprehensively assesses their estimation capabilities by training 70% of the models on literature-derived input parameters and testing the remaining 30%, enabling comparisons with other models and evaluations of enhanced versions of a single model. The examination involves statistical metrics in two distinct phases and categorizing students into four grade classes, providing a thorough comparative analysis. Ultimately, the study identifies the optimal model for understanding and anticipating students' academic performance, emphasizing NBC's adaptability, uncertainty estimation, and interpretability when integrated with optimization algorithms to enhance predictive accuracy and improve educational outcomes. In the following Section II outlines the methods, procedures, and details of your research, encompassing data collection, experimental design, participants, materials, and any statistical or computational methods employed. Moreover, the Experimental Design or Data Collection subsection provides a detailed account, including variables, controls, and procedures implemented. Dataset overview is given in Section III. Section IV presents study findings using tables, graphs, or figures, incorporating both descriptive and inferential statistical analyses. In Section V, results are interpreted in the context of addressing implications, limitations, and potential future research directions. Finally Section VI summarizes main findings, emphasizing their significance, broader implications, and suggesting potential applications or areas for further investigation.

II. METHODOLOGY AND STRATEGY

A. Naive Bayes Classifier (NBC)

The NBC is a probabilistic classifier that utilizes Bayes' theorem under the assumption of high independence. This algorithm was formulated by Thomas Bayes, a British scientist. The theoretical foundation of NBC revolves around predicting future opportunities by drawing on past experiences [30]. Assume a category of documents $D = \{d_1, d_2, \dots, d_n\}$ and m

potential classes $C = \{c_1, c_2, \dots, c_m\}$. Where $W = \{w_1, w_2, \dots, w_s\}$ represent the collection of distinct terms present in at least one of the documents in D. The subsequent formula can be used to calculate the probability that a document d belongs to class c [31].

$$\mathrm{P}\,(c|d) = \frac{\mathrm{P}\,(d|c)\,\mathrm{P(c)}}{\mathrm{P(d)}} \tag{1}$$

The denominator of Eq. (1) is often omitted in Maximum A Posteriori (MAP) calculation for parametric numerical problems because $P(d)$ is a constant for the known data set size. In the context of a NBs model, it accepts that each term or word, w_k, independently occurs in a document given the class c . Consequently, Eq. (1) is simplified to reflect this assumption:

$$P(c|d) \propto P(c) \prod_{\mathrm{k}=1}^{n_d} \left[P(w_k \,|\, c)\right]^{t_k} \tag{2}$$

In the provided context, n_d denotes the count of single words in file d, and t_k represents the frequency of each word w_k. To address concerns regarding floating−point underflow, an alternative equation is utilized:

$$\log P(c|d) \propto \log P(c) + \sum_{k=1}^{n_d} \left[t_k \log P(w_k \,|\, c)\right] \tag{3}$$

The classification of document d is determined as the class c^* that maximizes the logarithm of $P(c|d)$ in Eq. (3).

$$c^* = argmax_c \in \mathbb{C} \left\{\log P(c) + \sum_{k=1}^{n_d} \left[t_k \log P(w_k \,|\, c)\right]\right\} \tag{4}$$

In the application of the Naive Bayes classifier (NBC), $P(c)$ and $P(w_k \,|\, c)$ can derive estimations as follows:

$$\hat{P}(c) = \frac{N_c}{N} \, and \; \hat{P}(w_k|c) = \frac{N_{w_k}}{\sum_{w_i \in \mathbb{W}} N_{w_i}} \tag{5}$$

Where N signifies the total document count, N_c denotes the quantity of records in class c, and N_{w_i} represents the word's frequency w_i in class c. Utilizing these approximations, the computation of the expression on the right−hand side of Eq. (4) essentially becomes a counting challenge.

B. Jellyfish Search Optimizer (JSO)

One of the contemporary swarm-based metaheuristics is the JSO, introduced by Chou and Truong in 2021 [32]. The JSO algorithm emulates the foraging behaviour of jellyfish in search of oceanic sustenance [33].

1) Mathematical model: The JSO algorithm adheres to three ideal principles:

a) Marine flow: To locate and feed on smaller planktonic organisms, jellyfish utilize Eq. (6) to detect the direction of ocean currents.

$$\vec{O} = X' - \beta \times M \times r(0,1) \tag{6}$$

The direction of the ocean current is denoted as $\vec{O}$, where β ($\beta > 0$) represents the length distribution coefficient of $\vec{O}$. The current best location of the jellyfish swarm is denoted as X', and M signifies the mean location of all jellyfish.

The new location of each jellyfish can be defined as follows:

$$X_i(t+1) = X_i(t) + r(0,1) \times \vec{O} \tag{7}$$

Following the adjustment of each jellyfish's position, the current location of the jellyfish is chosen as a preferred destination, potentially representing a position with increased accessibility to food sources.

b) Blooming of jellyfish: Jellyfish within a bloom display two types of motion: passive and active. The subsequent section introduces the mathematical models that characterize these waves:

$$\text{Passive waves:} \quad X_i(t+1) = X_i(t) + \lambda \times r(0,1) \times (w_b - L_b) \tag{8}$$

λ ($\lambda > 0$) signifies a coefficient associated with the extent of passive waves. w_b and L_b denote the lower and upper bounds of the search space, correspondingly.

$$\text{Active waves:} \quad X_i(t+1) = X_i(t) + r(0,1) \times \vec{D} \tag{9}$$

which

$$\vec{D} = \begin{cases} X_i(t) - X_j(t) & if \quad g(X_i) < \; g(X_j) \\ X_j(t) - X_i(t) & if \quad g(X_i) \geq \; g(X_j) \end{cases} \tag{10}$$

The functions $g(X_i) and\ g(X_j)$ represent the objective purpose values corresponding to jellyfish $i\ and\ j$, correspondingly.

c) Temporal regulation mechanism: Within this framework, a mechanism for time control is utilized to regulate both the motion of jellyfish within the bloom and their navigation toward ocean currents. The temporal control function is denoted as:

$$T(t) = \left|\left(1 - \frac{t}{MaxIter}\right) \times (2 \times r(0,1) - 1\right| \tag{11}$$

where, t shows the time index given as the iteration quantity and $MaxIter$ represents the iteration max number. t denotes the time index, representing the iteration number, and $MaxIter$ signifies the maximum number of iterations.

2) Population initialization: In this optimizer, the initial population is generated using the Logistic map.

$$X_{i+1} = \upsilon X_i(1 - X_i), \qquad 0 \leq X_0 \leq 1 \tag{12}$$

X_i and X_0 represent the positional values for the i_{th} jellyfish and a randomly selected place, respectively. It is set to 4 in all testing.

a) Boundary handling mechanism: If a jellyfish surpasses the confines of the specified search space, it will be realigned within those limits, according to Eq. (13).

$$\begin{cases} X'_{i,d} = \left(X_{i,d} - W_{b,d}\right) + L_{b,d} & if \quad X_{i,d} > \; W_{b,d} \\ X'_{i,d} = \left(X_{i,d} - L_{b,d}\right) + W_{b,d} & if \quad X_{i,d} < \; W_{b,d} \end{cases} \tag{13}$$

$X_{i,d}$ and $X'_{i,d}$ represent the current and updated location of the d_{th} dimension of the i_{th} jellyfish, respectively. $W_{b,d}$ *and* $L_{b,d}$ symbolizes the lower and upper bounds of the d_{th} dimension within the search space.

The flowchart illustrating the JSO process is shown in Fig. 1.

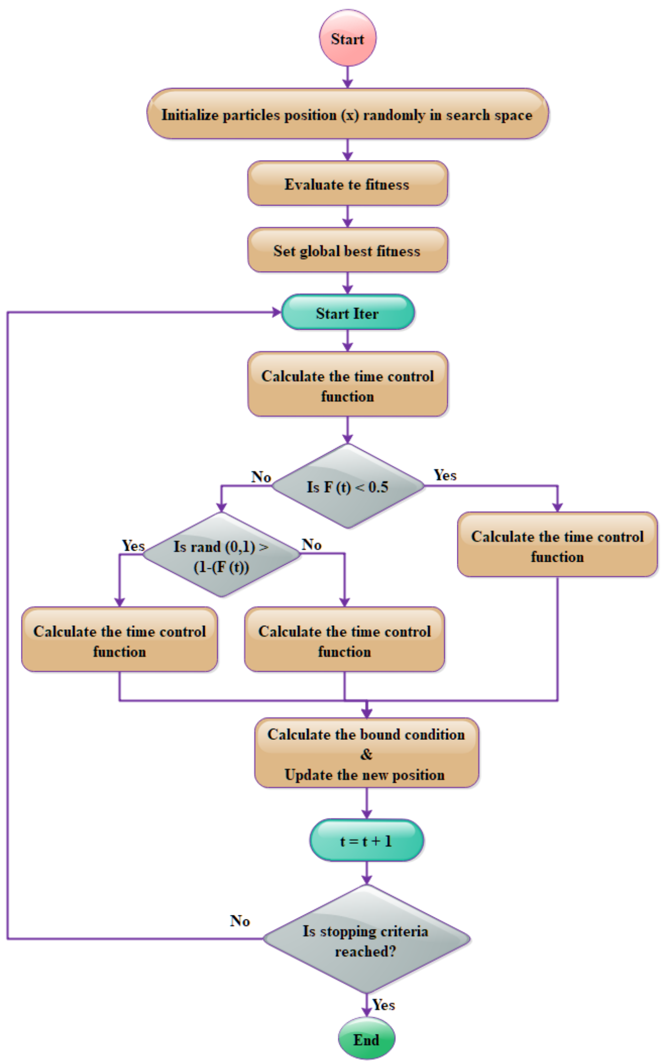


Fig. 1. Flowchart of JSO.

C. Artificial Rabbits Optimization (ARO)

Rabbits' adoption of survival strategies within their natural environment has significantly influenced the formulation of ARO. These strategies are designed to effectively counteract predators and optimize the rabbit's ability to evade surveillance. ARO, in its design, integrates the rabbit's dual strategies of hunting and hiding, along with its adept energy management techniques, to seamlessly transition between these adaptive approaches [34].

1) Detour foraging: Rabbits commonly employ a circuitous method during their quest for food, emphasizing distant food sources while overlooking nearby ones. Consider an ARO scenario in which each rabbit in a group has its territory complete with caves and grass. Serendipitous encounters with each other's feeding areas are common. A mathematical framework is presented that captures rabbits' departure search actions:

$$\vec{B}_i(t+1) = x_j(t) + S \times (x_i(t) - x_j(t)) + w(0.5 \times (0.05 + r_1)) \times m_1, \tag{14}$$

$$i, j = 1, \dots, n \ and \ j \neq 1$$

$$S = M \times v \tag{15}$$

$$M = (e - e^{\left(\frac{t-1}{I}\right)^2}) \times \sin(2\pi r_2) \tag{16}$$

$$v(y) = \begin{cases} 1 & if \quad y = f(1) \\ 0 & \quad else \end{cases} \quad k = 1, \dots, d \ and \ l = 1, \dots, \lceil r_3 \times d \rceil \tag{17}$$

$$f = p(d) \tag{18}$$

$$m_1 = N(0,1) \tag{19}$$

The population size of rabbits is denoted as n, and the dimensions of the problem are denoted as d.

$\vec{B}_i(t+1)$ signifies the standard normal distribution describes the distribution of the i_th rabbit's location at times $t + 1$ and $n1$. T represents the max number of iterations. $x_i(t)$ signifies the place of the i_{th} rabbit at time t. A random permutation of numbers between 1 and d is produced by the function p.

w functions as a tool inside the algorithm, promoting the diverse collection of components from the traveler to introduce variety into the process. r_1, r_2, and r_3 depict random numbers within the (0,1) range. The variable s denotes the run length, indicating the pace of development during reroute scavenging.

2) Random hiding: Rabbits tend to randomly select a burrow to seek shelter, a behaviour crucial for their survival. The mathematical model elucidating this performance is articulated through the equations presented below. The formulation for the j_{th}burrow of the i_{th} rabbit is expressed as:

$$\vec{B}_i(t+1) = x_i(t) + N \times f \times \vec{x}_i(t), \quad i, j = 1, \dots, n \ and \ j \neq 1 \tag{20}$$

$$D = \frac{I - t + 1}{I} \times r_4 \tag{21}$$

$$m_2 = N(0,1) \tag{22}$$

$$f(y) = \begin{cases} 1 & if \quad y = g(1) \\ 0 & \quad else \end{cases} \quad k = 1, \dots, d \tag{23}$$

$$\vec{R}_{i,r}(t) = \vec{x}_i(t) + N \times f \times \vec{x}_i(t) \tag{24}$$

N embodies the concealment parameter, undergoing a linear reduction from 1 to $\frac{1}{I}$ throughout an iterative process that incorporates random perturbations.

In the eventual scenario where either detour foraging strategies or random hiding methods are employed, the update of the i_{th} rabbit's location observes to the Eq. (25):

$$\vec{x}_i(t+1) = \begin{cases} \vec{x}_i(t) & g(\vec{x}_i(t)) \le g\left(\vec{B}_i(t+1)\right) \\ \vec{B}_i(t+1) & g(\vec{x}_i(t)) > g\left(\vec{B}_i(t+1)\right) \end{cases} \quad (25)$$

3) Energy shrinks: While the rabbit persists in its cyclic behaviour of detouring to find food and intermittently hiding at random, its energy level regularly diminishes. Hence, the integration of energy factors becomes critical within the ARO framework:

$$E(t) = 4\left(1 - \frac{t}{I}\right) ln \frac{1}{r} \quad (26)$$

Fig. 2 illustrates the flowchart of ARO, and Algorithm 1 provides its pseudo-code.

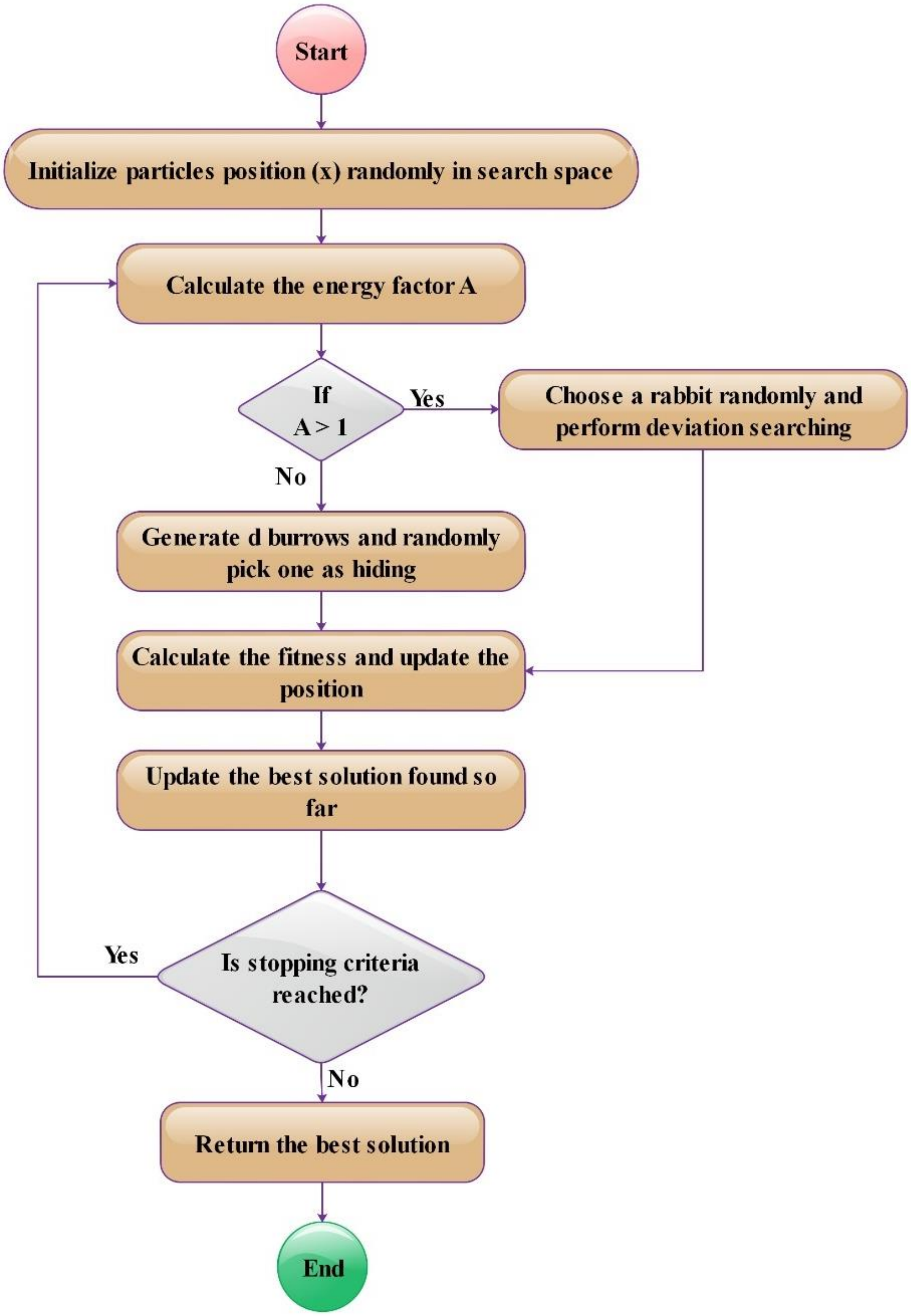


Fig. 2. Flowchart of ARO.

Algorithm 1: Pseudo-Code of ARO Algorithm

Randomly initialize a set of rabbits. X_i (solutions) and evaluate their fitness
While the stop criterion is not satisfied, do
for each individual X_i do
Compute the energy factor A using Eq. (26).
if $A > 1$
Choose a rabbit randomly from other individuals.
Compute R using Eqs. $(15) - (19)$.
Perform detour foraging using Eq. (14).
Compute the fitness Fit_i.
Upgrade the position of the current individual using Eq. (25).
else
Generate d burrows and randomly pick one as hiding using Eq. (24).
Perform random hiding using Eq. (20).
Compute the fitness Fit_i.
Update the position of the individual using Eq. (25).
end if
Upgrade the best solution found so far X_{best}
end for
end while
return X_{best}

III. Dataset Overview

A. Data Preparation

Data mining is a strategic business process, systematically delving into vast datasets to unearth significant patterns and rules that contribute valuable insights [35]. Classification and regression are pivotal objectives in data mining, playing crucial roles in extracting valuable insights from complex datasets by identifying meaningful patterns and relationships [36]. The primary distinction lies in the output representation, with classification producing discrete results and regression generating continuous outcomes. Evaluation metrics also differ, as classification models are often assessed using the percentage of correct classifications; while regression commonly employs the root mean squared metric [37].

This study aims to develop a robust method for accurately assessing students' academic performance and contextual factors. The dataset undergoes essential preprocessing, including transforming text into numerical values. Attributes are selected to describe performance based on individual information and academic conditions, utilizing two questionnaire methods and academic histories. The dataset encompasses a varied array of variables that have the potential to influence students' academic outcomes. The dataset incorporates information, including students' school, gender, age, home address, family size, parental cohabitation status, as well as details about the education and occupations of both parents. The dataset also includes information on factors influencing school choice, such as proximity, reputation, course preference, and others. It covers details about the student's guardian, weekly study time, travel time to school, past class failures, participation in educational support, family educational support, involvement in paid classes and extracurricular activities, internet access at home, aspirations for higher education, nursery school attendance, engagement in romantic relationships, family relationship quality, current health status, socializing with friends, weekday and weekend alcohol consumption, free time after school, and the number of school absences. These diverse input variables, encompassing nominal, numeric, and binary data types, provide a comprehensive and varied source of information for the study. In addition to these inherent traits, three supplementary variables, namely, $G2$, $G1$, and $G3$, depict students' grades across three assessment periods throughout their academic journey, spanning from zero (indicating the lowest grade) to 20 (representing the highest grade attainable). To classify their scores, a segmentation was applied, dividing them into four categories: 0–12 denoting Poor performance, 12–14 indicating Acceptable, 14–16 representing Good, and 16–20 reserved for Excellent academic achievements.

In this research, a correlation matrix encompassing the examined input and output variables is depicted in Fig. 3. Parents' educational background, particularly the mother, exerted the most positive influence on students' scores in G1 and G3, with the father's education also demonstrating effectiveness. As anticipated, family support and aspirations for higher education had positive effects on outcomes, while the influence of prior student failures was negative. Additionally, there was a noticeable gender effect on scores in G1 and G3.

B. Evaluation of Models' Applicability

In classification problems, Accuracy is a widely used metric that gauges overall model performance based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). While Accuracy is common, it has limitations in imbalanced data situations, favoring the majority class and providing limited insights. Three additional metrics, Recall, Precision, and F1-Score, address this. Recall evaluates a model's ability to correctly identify all relevant instances within a specific class, which is crucial for reducing False Negatives. Precision assesses the accuracy of positive predictions, reducing False Positives. The F1-Score combines Precision and Recall, offering a balanced assessment, especially valuable in imbalanced data scenarios.

These metrics, outlined through mathematical formulas Eq. (27) to Eq. (30), collaboratively contribute to a more comprehensive grasp of a classification model's effectiveness [38]. Particularly valuable in tackling class imbalances, they empower researchers and data analysts to make well-informed decisions and adjustments, enhancing model performance in challenging scenarios involving imbalanced data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$Precision = \frac{TP}{TP + FP} \tag{28}$$

$$Recall = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{29}$$

$$F1\_score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{30}$$

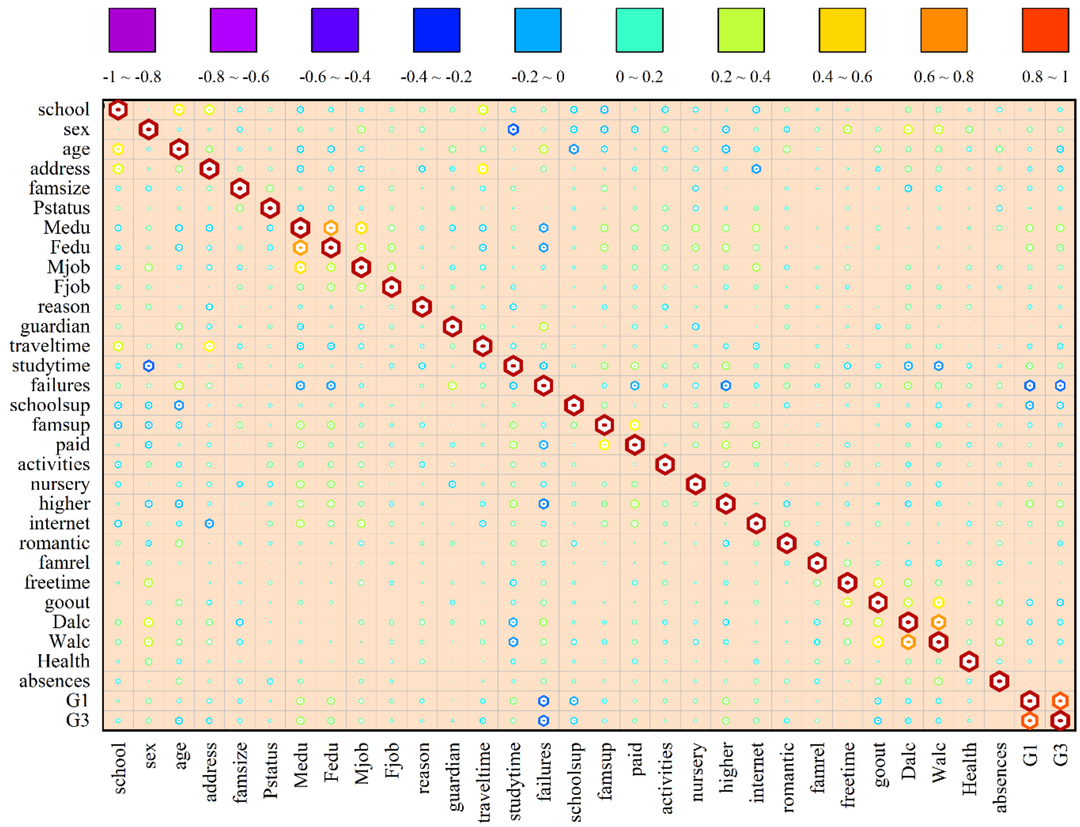


Fig. 3. Correlation matrix for the input and output variables.

IV. RESULTS

To improve the accuracy of the NBC model in predicting G1 and G3, this study employed two optimization algorithms, ARO and JSO. The dataset was divided, with 70% allocated for the training phase and the remaining 30% for thorough testing, enabling a comprehensive assessment of predictive capabilities. The data underwent processing after a detailed evaluation of the models' classification during training and testing, involving 395 students and grounded in their test results (specifically G1 and G3 values).

The primary objective involved fine-tuning and optimizing model parameters through these algorithms. To assess the convergence of these optimization methods, a convergence curve, depicted in Fig. 4, tracked accuracy over 200 iterations. The convergence rate of the NBAR model in G1 and G3 is similar, with a noticeable shift to a linear pattern around the 150th iteration in the convergence process. In contrast, the NBJS model, which predicts both G1 and G3 metrics, is the optimal model, achieving superior accuracy levels before the 150th iteration.

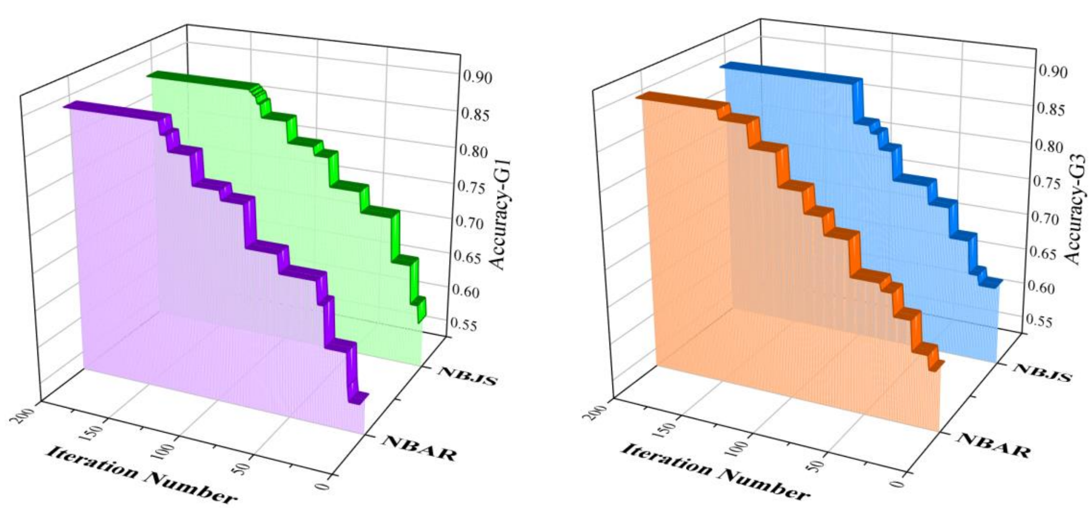


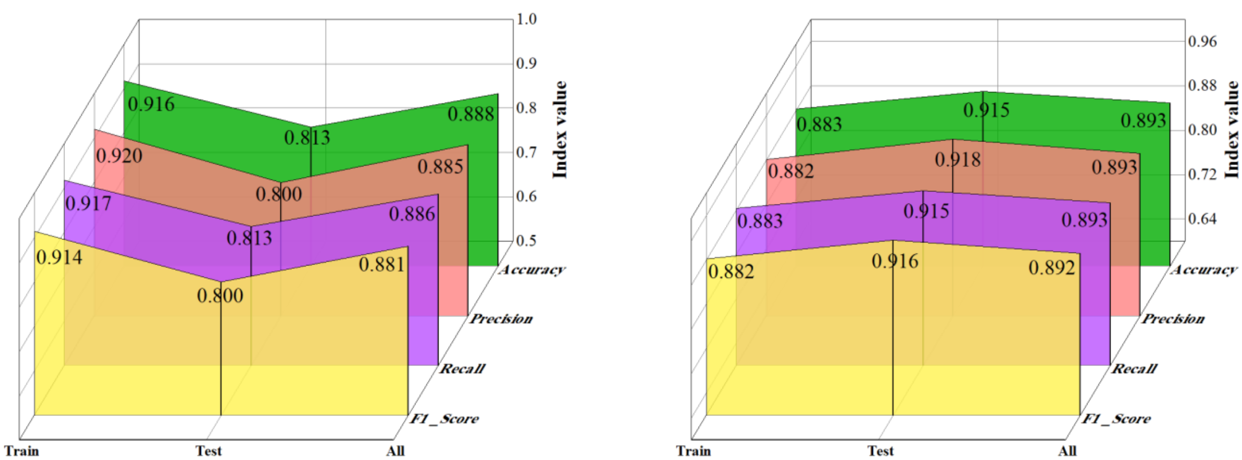
Fig. 4. Convergence of hybrid models based on ribbon plot.

This section assesses how each model contributes to predicting students' academic performance based on the G1 and G3 grades. Table I presents Accuracy, Recall, Precision, and F1-score measures for the training and testing phases across all models. Notably, the NBAR model demonstrated superior performance, exhibiting higher metric values. Specifically, in both G1 and G3, the NBAR model achieved maximum metric values with Accuracy=0.889, Precision = 0.885, Recall = 0.886, and F1-score = 0.882 for G1, and Accuracy= 0.894, Precision= 0.893, Recall= 0.894, and F1-score= 0.893 for G3, respectively. Additionally, the waterfall plot in Fig. 5 provides a visual assessment of the performance of the presented models.

The students were classified into four distinct groups based on their scores: Poor (0 to 12), Acceptable (12 to 14), Good (14 to 16), and Excellent (16 to 20). In terms of Precision within G1 estimation, shown in Table II, the NBJS model showcases superior performance, achieving values of 0.923 and 0.769 in the Excellent and Good categories, respectively. Conversely, for the Acceptable and Poor groups, the NBAR model demonstrates Precision values of 0.880 and 0.912, respectively. As indicated in Table III for G3 precision values, the NBAR model demonstrates superior performance in the Excellent and Acceptable groups with values of 0.917 and 0.768, respectively. Conversely, the NBJS model, with values of 0.836 and 0.957, is more suitable for the Good and Poor groups.

TABLE I. RESULT OF PRESENTED MODELS

	Model	phase	Index values			
			Accuracy	*Precision*	*Recall*	*F1 _Score*
G1	NBC	Train	0.884	0.882	0.885	0.881
		Test	0.822	0.810	0.822	0.811
		All	0.866	0.861	0.866	0.861
	NBAR	Train	0.917	0.921	0.917	0.915
		Test	0.814	0.801	0.814	0.801
		All	0.889	0.885	0.886	0.882
	NBJS	Train	0.899	0.900	0.899	0.894
		Test	0.831	0.825	0.831	0.821
		All	0.878	0.877	0.879	0.873
G3	NBC	Train	0.866	0.866	0.866	0.865
		Test	0.898	0.905	0.898	0.900
		All	0.876	0.878	0.876	0.876
	NBAR	Train	0.883	0.882	0.883	0.882
		Test	0.915	0.918	0.915	0.916
		All	0.894	0.893	0.894	0.893
	NBJS	Train	0.870	0.869	0.870	0.869
		Test	0.915	0.922	0.915	0.917
		All	0.884	0.885	0.884	0.884



NBAR-G1 NBAR-G3

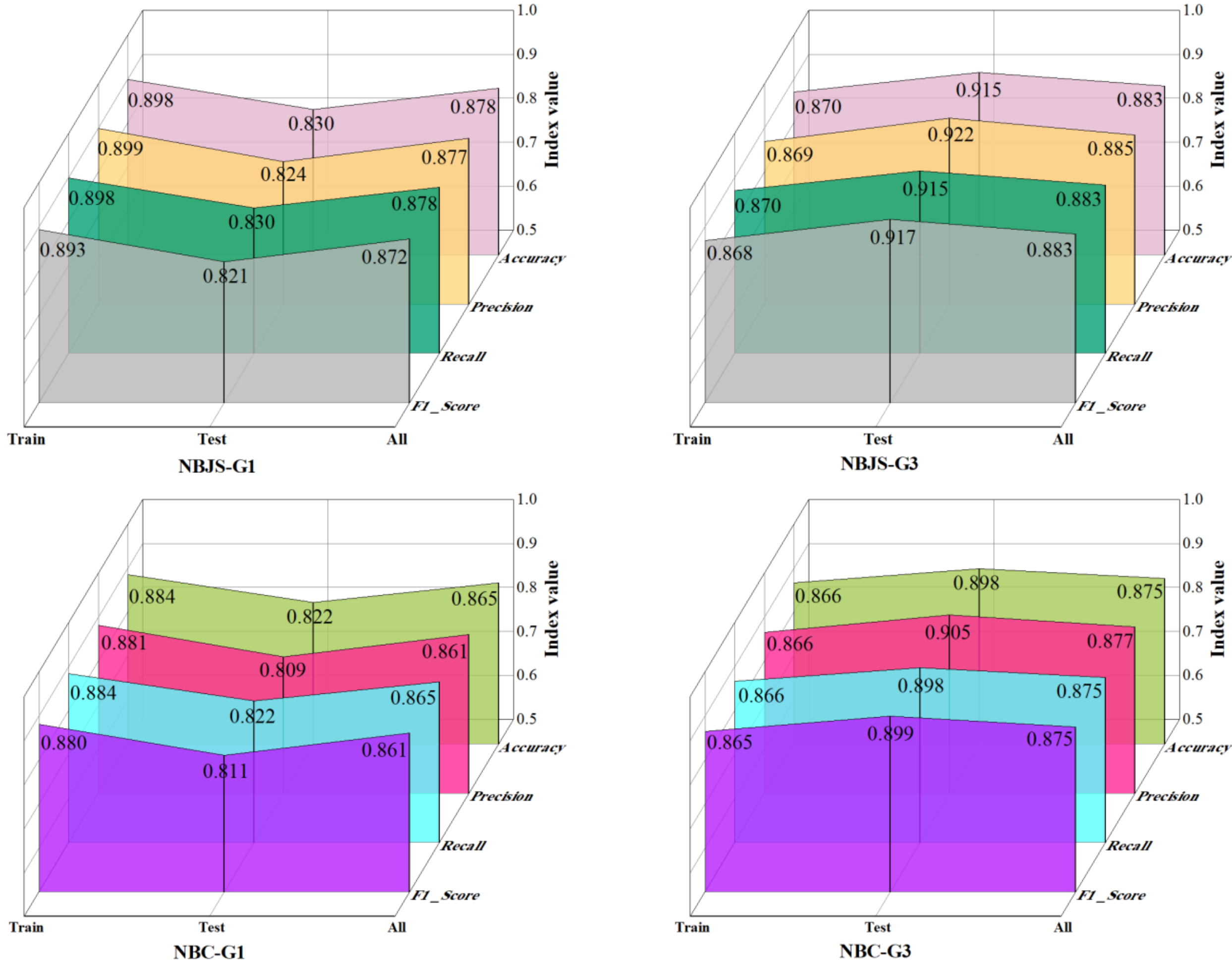


Fig. 5. Waterfall plot utilized to assess the performance of the presented models.

TABLE II. ASSESSMENT METRICS FOR THE PERFORMANCE OF THE GENERATED MODELS DERIVED FROM G1 PREDICTION

Model	Grade	Index values		
		Precision	*Recall*	*F1-score*
NBC	Excellent	0.875	0.854	0.864
	Good	0.745	0.704	0.724
	Acceptable	0.800	0.647	0.715
	Poor	0.904	0.970	0.936
NBAR	Excellent	0.897	0.854	0.875
	Good	0.768	0.796	0.782
	Acceptable	0.880	0.647	0.746
	Poor	0.912	0.983	0.946
NBJS	Excellent	0.923	0.878	0.900
	Good	0.769	0.741	0.755
	Acceptable	0.875	0.617	0.724
	Poor	0.895	0.987	0.939

TABLE III. ASSESSMENT METRICS FOR THE PERFORMANCE OF THE GENERATED MODELS DERIVED FROM G3 PREDICTION

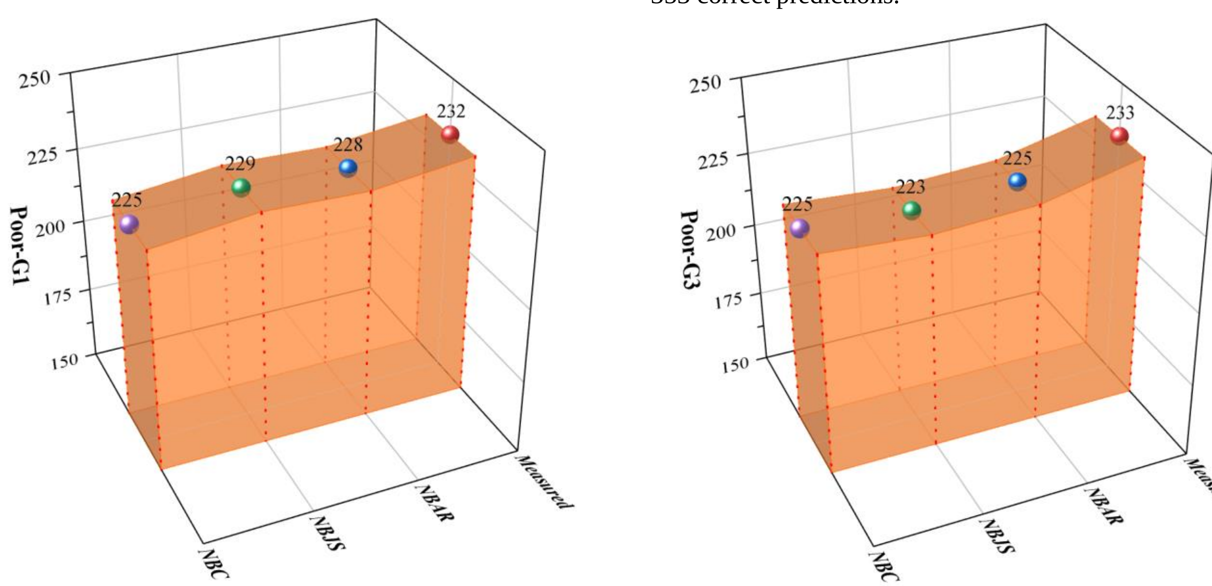
Model	Grade	Index values		
		Precision	***Recall***	***F1-score***
NBC	Excellent	0.909	0.750	0.822
	Good	0.772	0.733	0.752
	Acceptable	0.691	0.758	0.723
	Poor	0.949	0.966	0.957
NBAR	Excellent	0.917	0.825	0.868
	Good	0.788	0.867	0.825
	Acceptable	0.768	0.694	0.729
	Poor	0.949	0.966	0.957
NBJS	Excellent	0.833	0.750	0.790
	Good	0.836	0.767	0.800
	Acceptable	0.761	0.871	0.812
	Poor	0.957	0.957	0.957

In Fig. 6, 3D walls illustrate a detailed comparison between predicted and measured values, presenting the distribution of students across categories for G1 and G3. Individual graphs for each category (Poor, Acceptable, Good, and Excellent) are included. It is noteworthy that the institute's report mentions a total of 395 students. The subsequent sections undertake a comprehensive assessment of the models' classification effectiveness.

As per the chart, the data for G1 reveals that 232 individuals fall into the Poor category, 68 individuals in the Acceptable category, 54 individuals in the Good category, and 41 individuals in the Excellent category. Notably, the NBJS model is the most effective classifier for the Poor and Excellent segments, demonstrating accurate predictions. Conversely, in the Acceptable and Good groups, the NBAR model outperforms, exhibiting superior performance.

As depicted in the G3 graph, recorded figures for the Poor, Acceptable, Good, and Excellent categories were 233, 62, 60, and 40 students, respectively. An exception arises as the single NBC model, and NBAR exhibit superior performance in the Poor class. Moreover, the NBAR model consistently demonstrates superior performance in the Excellent and Good classes. However, the NBJS model outperforms others in the Acceptable class, while the NBAR model exhibits a noticeable performance drop.

Valuable information regarding the precise classification of students and instances of misclassifications can be extracted from the confusion matrix illustrated in Fig. 7. In G1 estimation, the NBAR model accurately classified a total of 350 students, encompassing 35 Excellent, 43 Good, 44 Acceptable, and 228 Poor students, in their respective grades, while 45 students were misclassified. In contrast, the NBJS model had 48 misclassifications, indicating a 6.25% difference in G1 between the two hybrid models, with the NBAR model exhibiting superior performance. In G3, the NBAR model achieved precise categorization for 353 students, accurately placing them in their respective grades, with only 42 students being misclassified. Similarly, the NBJS model also achieved 353 correct predictions.



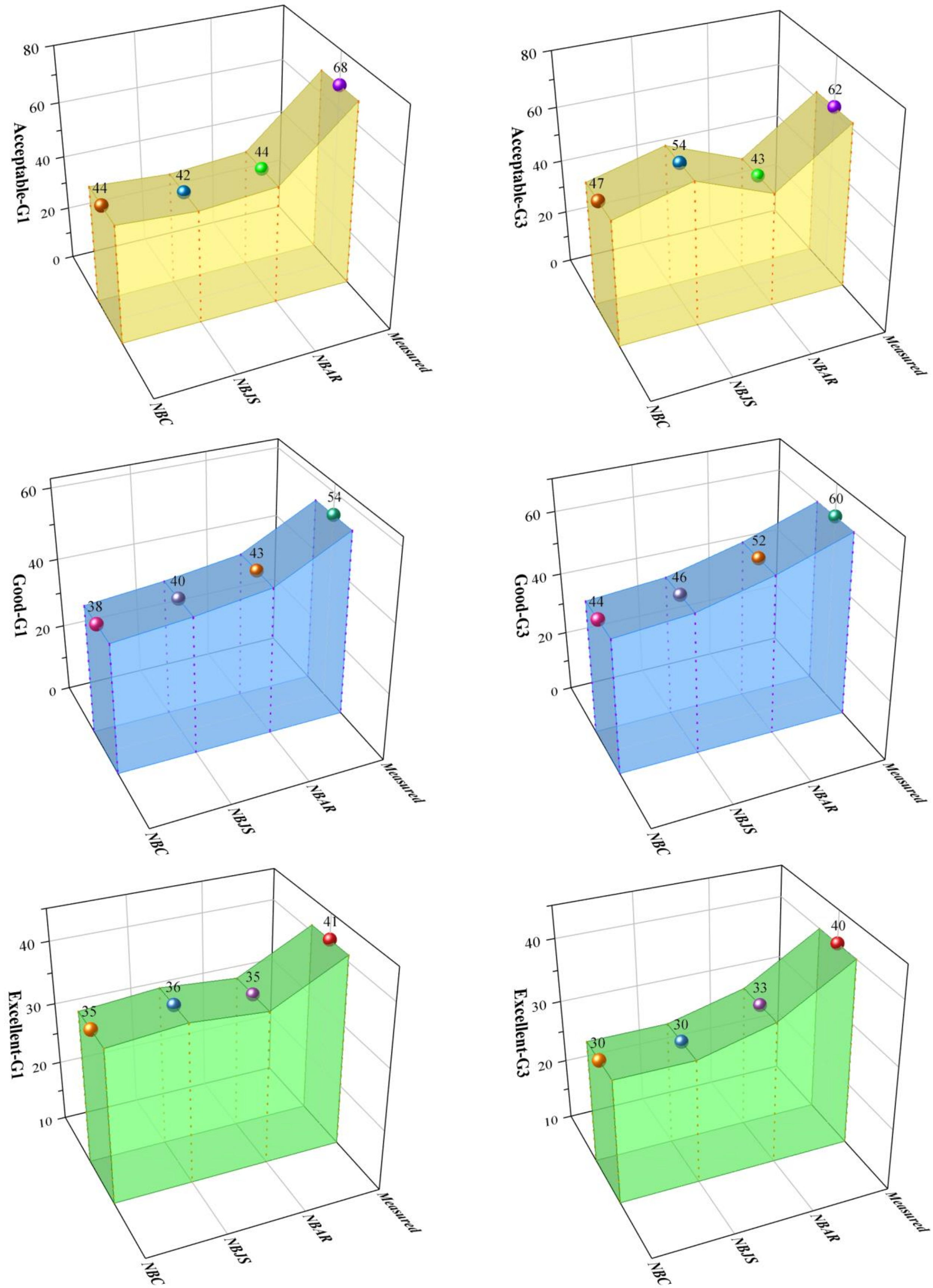


Fig. 6. 3D walls for the comparison between predicted and measured values.

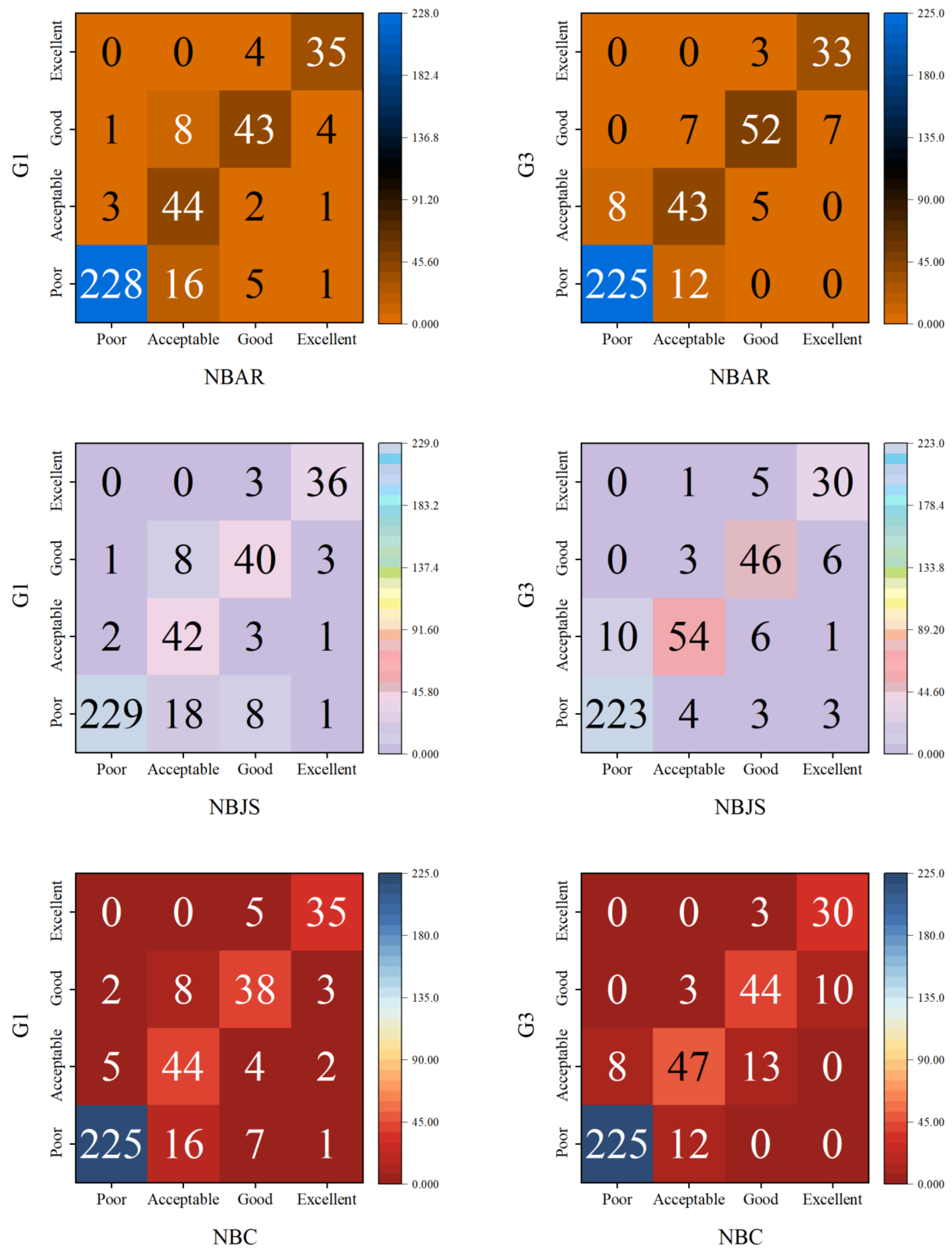


Fig. 7. Confusion matrix illustrating the accuracy of the model.

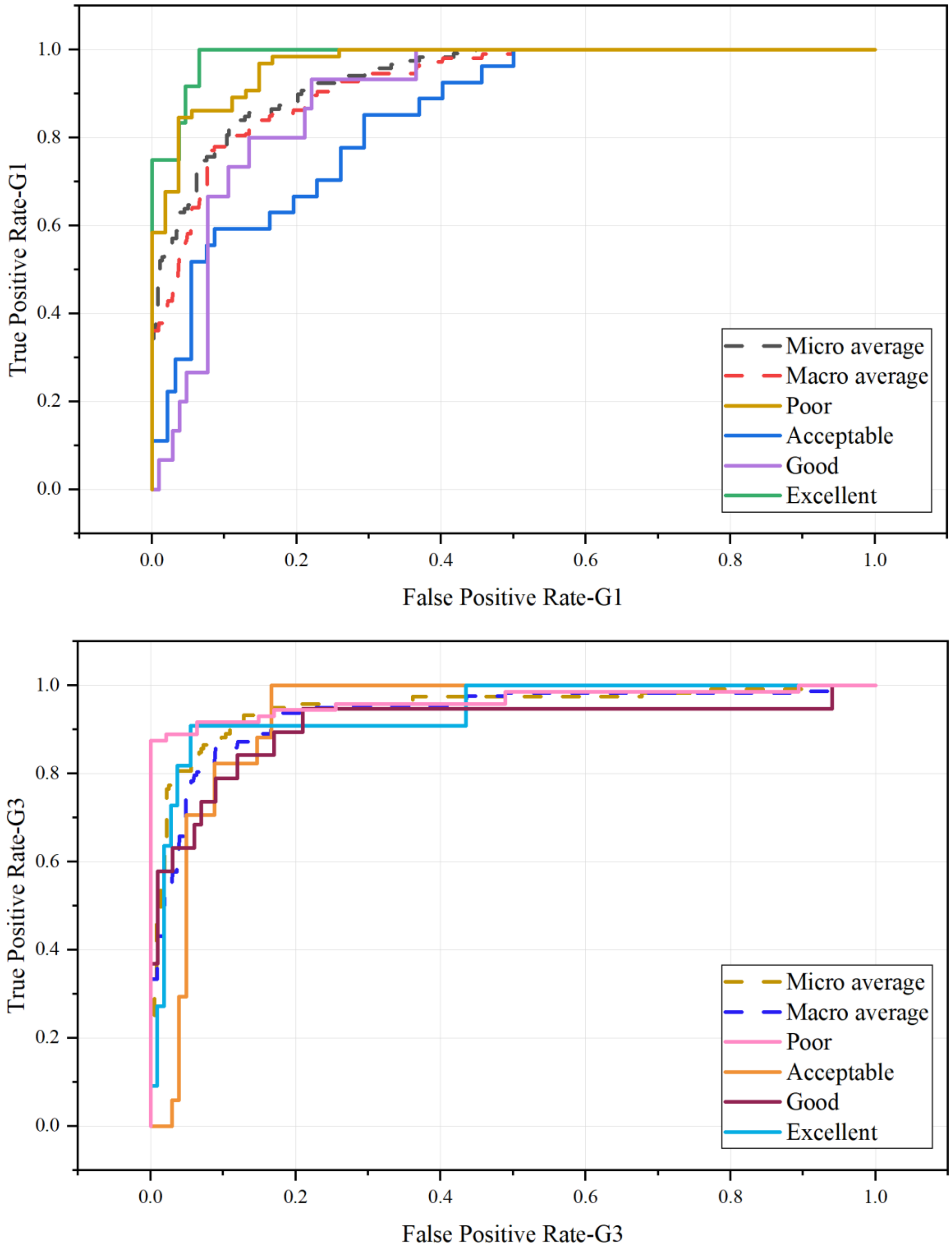


Fig. 8. Result of ROC curve.

The Receiver Operating Characteristics (ROC) curve is crucial for assessing classification algorithms. This curve evaluates the model's performance by graphing TP rates against FP rates. A test exhibiting perfect discrimination would manifest as an ROC plot traverse the upper left corner, indicating both 100% sensitivity and 100% specificity. The ROC curve analysis in Fig. 8 shows that, the NBC emerges as the best overall classifier in G1 prediction, particularly for the poor and excellent classes, as evidenced by its proximity to 1 on the ROC curve. In the framework of G3, there is no discernible trend for comparing the performance of the models; however, there is a relative improvement in predicting the Poor group.

V. Discussion

The study has limitations that should be considered. Firstly, its focus on a specific cohort of 395 students may restrict the generalizability of the findings to a broader student population. Additionally, the evaluation predominantly revolves around quantitative metrics, such as Accuracy, Recall, Precision, and F1-score, potentially overlooking qualitative nuances in student academic performance. The exclusive emphasis on specific grading criteria (G1 and G3) raises questions about the adaptability of the proposed methodology to different grading systems or academic contexts. Despite these limitations, the study represents a notable advancement in predictive modeling for education.

In addition, Future studies in the realm of predictive modeling in education could explore diverse paths to advance the field. Key areas for investigation include assessing the generalizability of hybrid models across different educational settings and student populations, incorporating qualitative factors to provide a more comprehensive understanding of academic performance, and evaluating the adaptability of the proposed methodology to various grading systems. Longitudinal analyses and the integration of real-time data offer opportunities for dynamic predictions and a deeper exploration of academic trajectories. Comparative studies with other predictive models, ethical considerations, and impact assessments on implementation in educational institutions are also important avenues for further research. Addressing these aspects will contribute to refining predictive models and enhancing their practical application in education.

In addition, Table IV is shown to compare the accuracy of the developed best model with other published papers.

TABLE IV. Comparison between the Presented and Published Papers

No.	Paper	Model	Accuracy
1	Al-Radaideh et al. [39]	DTC	87.9%
2	Bichkar and R. R. Kabra [40]	DTC	69.94%
3	Carlos et al. [41]	ADTree	97.3%
4	Kabakchieva [42]	DTC	72.74%
5	Nguyen and Peter [43]	DTC	82%
6	Edin Osmanbegovic et al. [44]	NBC	76.65%
7	Present study	NBAR	89.4%

VI. Conclusion

This study underscores the vital role of data-driven predictive models in education, emphasizing the need to consider qualitative and quantitative factors in forecasting and assessing student academic performance. The research introduces innovative hybrid models that integrate the Naive Bayes classifier (NBC) with optimization algorithms, namely Jellyfish Search Optimizer (JSO) and Artificial Rabbits Optimization (ARO). The study showcases a cutting-edge methodology demonstrating how the precision and effectiveness of predictive models can be enhanced through advanced machine learning and optimization algorithms. The thorough assessment using essential metrics such as Accuracy, Precision, Recall, and F1-score underscores these meta-heuristic algorithms' capability to optimize classification results. This study focuses on classifying grades G1 and G3 for a cohort of 395 students. In predicting G1, the NBAR model demonstrated superior performance compared to the NBJS model based on the F1-score criterion, surpassing it by approximately 1.03% and outperforming the NBC model by 2.39%. Regarding Recall, this advantage amounted to 0.18% and 2.83%, respectively. Furthermore, in the G3 forecast, the NBAR model exhibited better performance in the F1-score criterion, approximately 1.01% better than the NBJS model and 1.1% better than the NBC model. This superiority in Recall is represented by percentages of 1.01% and 2.02%, respectively. This research represents a significant advancement in predictive modelling within the field of education, presenting promising avenues for improving the precision and efficiency of evaluating academic performance.

References

[1] T. PanduRanga Vital, B.G. Lakshmi, H. Swapna Rekha, M. DhanaLakshmi, Student Performance Analysis with Using Statistical and Cluster Studies, in: Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018, Springer, 2019: pp. 743–757.

[2] A. Behr, M. Giese, H.D. Teguim Kamdjou, K. Theune, Motives for dropping out from higher education—An analysis of bachelor's degree students in Germany, Eur J Educ 56 (2021) 325–343.

[3] D. Buenaño-Fernandez, W. Villegas-CH, S. Luján-Mora, The use of tools of data mining to decision making in engineering education—A systematic mapping study, Computer Applications in Engineering Education 27 (2019) 744–758.

[4] B.A. Pereira, A. Pai, C. Fernandes, A comparative analysis of decision tree algorithms for predicting student's performance, Int J Eng Sci 7 (2017) 10489–10492.

[5] A. Shanthini, G. Vinodhini, R.M. Chandrasekaran, Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers., J. Comput. Sci. 14 (2018) 654–662.

[6] M. Bucos, B. Drăgulescu, Predicting student success using data generated in traditional educational environments, TEM Journal 7 (2018) 617.

[7] J. Xu, Y. Han, D. Marcu, M. Van Der Schaar, Progressive prediction of student performance in college programs, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[8] X. Liu, Y. Ding, H. Tang, F. Xiao, A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data, Energy Build 231 (2021) 110601.

[9] W. Villegas-Ch, S. Luján-Mora, Analysis of data mining techniques applied to LMS for personalized education, in: 2017 IEEE World Engineering Education Conference (EDUNINE), IEEE, 2017: pp. 85–89.

[10] F. Ünal, Data mining for student performance prediction in education, Data Mining-Methods, Applications and Systems 28 (2020) 423–432.

[11] L. Masangu, A. Jadhav, R. Ajoodha, Predicting student academic performance using data mining techniques, Advances in Science, Technology and Engineering Systems Journal 6 (2021) 153–163.

[12] L.D. Yulianto, A. Triayudi, I.D. Sholihati, Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4. 5: Implementation Educational Data Mining For Analysis of Student Performance Prediction w, Jurnal Mantik 4 (2020) 441–451.

[13] S. Batool, J. Rashid, M.W. Nisar, J. Kim, H.-Y. Kwon, A. Hussain, Educational data mining to predict students' academic performance: A survey study, Educ Inf Technol (Dordr) 28 (2023) 905–971.

[14] B. Mehboob, Predicting student performance and risk analysis by using data mining approach, (2023).

[15] B. Sadaghat, G.G. Tejani, S. Kumar, Predict the Maximum Dry Density of soil based on Individual and Hybrid Methods of Machine Learning, Advances in Engineering and Intelligence Systems 2 (2023).

[16] A.P. Alfiani, F.A. Wulandari, Mapping student's performance based on data mining approach (a case study), Agriculture and Agricultural Science Procedia 3 (2015) 173–177.

[17] T. Mahboob, S. Irfan, A. Karamat, A machine learning approach for student assessment in E-learning using Quinlan's C4. 5, Naive Bayes and Random Forest algorithms, in: 2016 19th International Multi-Topic Conference (INMIC), IEEE, 2016: pp. 1–8.

[18] M. Pandey, S. Taruna, Towards the integration of multiple classifier pertaining to the Student's performance prediction, Perspect Sci (Neth) 8 (2016) 364–366.

[19] Purnawansyah, Haviluddin, K-Means clustering implementation in network traffic activities, in: 2016 International Conference on Computational Intelligence and Cybernetics, IEEE, 2016: pp. 51–54.

[20] Y.K. Saheed, T.O. Oladele, A.O. Akanni, W.M. Ibrahim, Student performance prediction based on data mining classification techniques, Nigerian Journal of Technology 37 (2018) 1087–1091.

[21] G. Deeva, J. De Smedt, C. Saint-Pierre, R. Weber, J. De Weerdt, Predicting student performance using sequence classification with time-based windows, Expert Syst Appl 209 (2022) 118182.

[22] A.D. Kumar, R.P. Selvam, V. Palanisamy, Hybrid classification algorithms for predicting student performance, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021: pp. 1074–1079.

[23] E.H. Yossy, Y. Heryadi, Comparison of data mining classification algorithms for student performance, in: 2019 IEEE International Conference on Engineering, Technology and Education (TALE), IEEE, 2019: pp. 1–4.

[24] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, K. Phasinam, Classification and prediction of student performance data using various machine learning algorithms, Mater Today Proc 80 (2023) 3782–3785.

[25] C. Márquez-Vera, A. Cano, C. Romero, S. Ventura, Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data, Applied Intelligence 38 (2013) 315–330.

[26] T. Hu, T. Song, Research on XGboost academic forecasting and analysis modelling, in: J Phys Conf Ser, IOP Publishing, 2019: p. 12091.

[27] D. Kabakchieva, Student performance prediction by using data mining classification algorithms, International Journal of Computer Science and Management Research 1 (2012) 686–690.

[28] D. Thammasiri, D. Delen, P. Meesad, N. Kasap, A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, Expert Syst Appl 41 (2014) 321–330.

[29] F. Marbouti, H.A. Diefes-Dux, K. Madhavan, Models for early prediction of at-risk students in a course using standards-based grading, Comput Educ 103 (2016) 1–15.

[30] N. Dengen, E. Budiman, M. Wati, U. Hairah, Student Academic Evaluation using Naïve Bayes Classifier Algorithm, in: 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), IEEE, 2018: pp. 104–107.

[31] H. Sibyan, J. Svajlenka, H. Hermawan, N. Faqih, A.N. Arrizqi, Thermal Comfort Prediction Accuracy with Machine Learning between Regression Analysis and Naïve Bayes Classifier, Sustainability 14 (2022) 15663.

[32] J.-S. Chou, D.-N. Truong, A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean, Appl Math Comput 389 (2021) 125535.

[33] A. Alam, P. Verma, M. Tariq, A. Sarwar, B. Alamri, N. Zahra, S. Urooj, Jellyfish search optimization algorithm for mpp tracking of pv system, Sustainability 13 (2021) 11736.

[34] L. Wang, Q. Cao, Z. Zhang, S. Mirjalili, W. Zhao, Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems, Eng Appl Artif Intell 114 (2022) 105082.

[35] G.S. Linoff, M.J.A. Berry, Data mining techniques: for marketing, sales, and customer relationship management, John Wiley & Sons, 2011.

[36] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer, 2009.

[37] I.H. Witten, E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, Acm Sigmod Record 31 (2002) 76–77.

[38] X. Luo, Efficient English text classification using selected machine learning techniques, Alexandria Engineering Journal 60 (2021) 3401–3409.

[39] Q.A. Al-Radaideh, A. Al Ananbeh, E. Al-Shawakfa, A classification model for predicting the suitable study track for school students, Int. J. Res. Rev. Appl. Sci 8 (2011) 247–252.

[40] R.R. Kabra, R.S. Bichkar, Performance prediction of engineering students using decision trees, Int J Comput Appl 36 (2011) 8–12.

[41] C. Márquez-Vera, A. Cano, C. Romero, S. Ventura, Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data, Applied Intelligence 38 (2013) 315–330.

[42] D. Kabakchieva, Student performance prediction by using data mining classification algorithms, International Journal of Computer Science and Management Research 1 (2012) 686–690.

[43] N.T. Nghe, P. Janecek, P. Haddawy, A comparative analysis of techniques for predicting academic performance, in: 2007 37th Annual Frontiers in Education Conference-Global Engineering: Knowledge without Borders, Opportunities without Passports, IEEE, 2007: pp. T2G-7.

[44] E. Osmanbegovic, M. Suljic, Data mining approach for predicting student performance, Economic Review: Journal of Economics and Business 10 (2012) 3–12.