Improving the Prediction of Student Performance by Integrating a Random Forest Classifier with Meta-Heuristic Optimization Algorithms

Chao Ma
Academic Affairs Office of Jiangsu University, Zhenjiang 223001, Jiangsu Province, China

***Abstract*—Anticipating student performance in higher education is crucial for informed decision-making and the reduction of dropout rates. This study focuses on the intricate analysis of diverse educational datasets using machine learning, particularly emphasizing dimensionality reduction. The aim is to empower educators with data-driven insights, enabling timely interventions for academic improvement. By categorizing individuals based on their inherent aptitudes, the study seeks to mitigate failure rates and enhance the overall educational experience. The integration of predictive modeling, particularly employing the robust Random Forest Classifier (RFC), allows the academic community to proactively address challenges and foster a supportive learning environment, thereby improving student outcomes. To bolster predictive capabilities, the study adopts the RFC model and enhances its efficacy through advanced optimization algorithms, specifically Electric Charged Particles Optimization (ECPO) and Artificial Rabbits Optimization (ARO). These sophisticated algorithms are strategically integrated to refine decision-making processes and enhance predictive precision. Furthermore, the analysis of the input variables has been conducted to assess their individual impact on student performance. This analysis can help institutions identify and address areas for improvement in their management practices. The study's commitment to leveraging state-of-the-art machine learning and bio-inspired algorithms underscores its dedication to achieving precise and resilient predictions of the performance of 4424 students, ultimately contributing to the advancement of educational outcomes. The research outcomes highlight the superiority of the RFEC model, optimized through ECPO for RFC, in aligning with actual measured values, affirming its efficacy in predictive accuracy.**

***Keywords*—Classification; student performance; machine learning; Random Forest Classifier; Electric Charged Particles Optimization; Artificial Rabbits Optimization**

I. Introduction

Success in higher education is crucial for employment, social equity, and economic development. Addressing dropout rates stands out as a significant challenge for higher education institutions aiming to enhance their success. The definition of dropout lacks universal acceptance, leading to variations in the reported proportion of students leaving, influenced by differing definitions, calculation methods, and data sources [1]. Research often analyzes dropouts by considering the timing of the event, distinguishing between early and late dropouts [2]. Comparing dropout rates across institutions becomes challenging due to discrepancies in reporting practices [3]. Consequently, the diverse definitions and reporting variations contribute to the complexity of understanding and addressing the dropout issue in higher education [4].

In the domain of higher education research, student dropout is precisely defined as a distinctive manifestation of attrition, delineating individuals who disengage from the higher education system without acquiring a (first) degree and fail to complete their academic pursuits after that. This narrow conceptualization has gained prominence in scholarly investigations, as evidenced by studies such as those conducted by Schröder-Gronostay and Daniel, Ziegele, and Heublein, Schmelzer, and Sommer [5–7]. Consequently, alterations in degree programs or fields of study, interruptions in academic pursuits, and changes in institutions are categorized as different forms of attrition. Various methods exist for gauging the frequency of student dropout, with the most effective being statistical tracking of course progression, wherein the investigation status of each student is documented every semester [8–10].

As students' progress through multiple semesters in their academic programs, their evaluation occurs on a semester or term basis. The final academic status, whether at graduation or in a subsequent semester, is inherently influenced by preceding semesters. This pattern allows for the prediction of future semester performance based on historical academic data. Contemporary advancements in this predictive process leverage various Data Mining (DM) tools and techniques, particularly within the domain of Educational Data Mining (EDM) [11–13]. EDM focuses on the prediction of Student Academic Performance (SAP) [14] and often employs predictive models generated by DM tools. These models play a vital part in facilitating SAP prediction, enabling the monitoring of students' academic progress. This, in turn, assists in determining strategic interventions for both students and other education stakeholders [15–17].

II. Literature Review

The exploration, modeling, and prediction of student performance and academic progression have garnered substantial research attention in recent decades, as evidenced by an influx of scholarly contributions [18–20]. While early works in this domain trace back to the '70s and '80s, the contemporary surge in data availability from educational institutions, coupled with the ascent of data science, has ushered in novel research avenues [21–25]. Also, recent research related to this study's target, exemplified by Jayaprakash et al. delved comprehensively into the intricate factors shaping students' academic accomplishments and their applicability in identifying students at risk. This study innovatively introduced an upgraded Random Forest classifier, striving for heightened accuracy in classification and prediction when juxtaposed with alternative algorithms like Naive Bayes, Bagging, Boosting, and the conventional Random Forest [26]. In alignment with this, Batool et al. employed the Random Forest classification model to anticipate students' final exam outcomes, leveraging publicly available datasets with diverse demographic features. The assessment incorporated meticulous methodologies such as hold-out and cross-validation [27]. Chen and Zhai's investigation took a multifaceted approach, employing three task-oriented educational datasets and implementing seven parameter-optimized machine learning methods for diverse performance prediction tasks [28]. Additionally, Asselmen et al. concentrated on the effectiveness of Ensemble Learning methods, proposing an innovative Predictive Feature Analytics (PFA) approach grounded in various models (Random Forest, XGBoost and AdaBoost,) to augment predictive accuracy in performance of students. The proposed models underwent rigorous evaluation across three distinct datasets [29].

Harnessing the capabilities of machine learning (ML) models to predict student dropout, enrollment, and graduation brings numerous advantages to both students and educational institutions. These models empower educators to accurately identify individuals in danger of dropping out, allowing them to create targeted support strategies that improve the likelihood of a successful post-graduation path. In this research, the recently developed Random Forest Classifier (RFC) method was applied to identify crucial factors influencing dropout, enrollment, and graduation outcomes. The RFC model underwent optimization using two distinct optimizers, Electric Charged Particles, and Artificial Rabbits, aimed at improving its overall performance. A subset of data was utilized from existing scientific articles during the training phase. After training, the model's effectiveness was assessed by testing it with separate data. Ultimately, the model that demonstrated optimal performance, surpassing the predefined benchmark ratio denoted as the actual measured value, was identified as the most adept in predictive capacities.

The research utilizes the RFC for predictive modeling and integrates advanced optimization algorithms, Electric Charged Particles Optimization (ECPO) and Artificial Rabbits Optimization (ARO). ECPO and ARO were chosen for their superior ability to navigate complex search spaces and avoid local optima, ensuring more accurate predictions. Additionally, the analysis of input variables helps identify areas for improvement in management practices. By leveraging state-of-the-art machine learning and bio-inspired algorithms, the study aims to achieve precise predictions for student performance, ultimately advancing educational outcomes.

The paper is structured as follows. Literature review is given in Section II. The detailed explanation of the model is given, and the meta-heuristic techniques employed are covered in Section III. In addition, the description of the dataset and its processing are covered in depth in this section. The created models' performance assesses in Section IV. In Section V and VI, the conclusion and future works shows the summary of paper based on results and description.

III. Materials and Methodology

A. Random Forest Classifier (RFC)

The RF is a supervised ML algorithm tailored for classification and prediction tasks, highlighting its prowess in classification. In this method, the term "forest" denotes a collection of numerous decision trees, and the model's robustness grows as more trees are added. Utilizing diverse data samples, the RFC method constructs individual decision trees. When faced with the challenge of predicting the class for new data points, each tree independently provides its prediction, thereby playing a role in the overall decision-making process. The culmination of this process involves identifying the most effective solution through a voting mechanism, with each decision tree contributing a vote for an input vector (x). The final prediction (C_{rf}^{B}) is determined through a majority vote. Functioning as an ensemble method, this model leverages the power of multiple uncorrelated models (trees) working together to surpass the performance of a single model. By adopting this collaborative approach, errors are mitigated, and overall accuracy is improved, as a range of diverse decision trees collectively contribute to the ultimate prediction.

In shaping decision trees, crucial considerations involve the selection of attributes and pruning techniques. Among these, the Gini Index method holds prominence as a commonly favored approach for attribute selection within RFC [30]. This method gauges the impurity of attributes concerning their respective classes. The assessment involves measuring impurity by randomly selecting a sample case from the training set and predicting its class as C_i. This informed attribute selection is articulated through the following equation, where $(F(C_i,T)/(|T|)$ signifies the probability that a chosen case belongs to the class C_i [31].

$$\sum\sum_{j\neq i}(F(C_i,T)/(|T|)(F(C_j,T)/(|T|) \qquad (1)$$

When establishing a prediction model with RFC, it is imperative to define two key parameters: the number of trees and each tree node's assigned input variables. RFC is composed of N decision trees (with N being user-defined), and these trees collaboratively contribute their votes to ascertain the class of new data points, relying on their predictions [32].

The framework associated with RFC is displayed in Fig. 1.

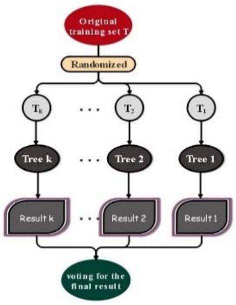


Fig. 1. Flowchart of the RFC.

B. Electric Charged Particles Optimization (ECPO)

Drawing inspiration from the interactions of electric-charged particles (ECPs), the ECPO functions as a population-based algorithm. It incorporates several internal parameters, each serving a specific purpose. The total number of ECPs is denoted as nECP, as well as nECPI represents the number of interacting ECPs. Additionally, naECP denotes the archive pool's size, and MaxITER indicates the maximum number of iterations.

One crucial aspect is nECPI, determining the number of particles engaged in interactions using a unique strategy. During these interactions, when two particles come into contact, a distinctive dynamic unfolds. The worst-performing particle repels the best one, while simultaneously, the best-performing particle attracts the worst one. This interplay within the ECPO framework contributes to its optimization process.

Algorithm 1, encapsulating the pseudo-code of the ECPO algorithm [33]:

ALGORITHM 1. PSEUDO-CODE OF ECPO OPTIMIZER

```
Input objective function, Problem Size (dimension of a problem), nECP,
nECPI, Strategy, naECP, and MaxITER
Output $ECP_{best}$
**Initialization** ()
For Iter=1: MaxITER
Selection ()
Interaction ()
BoundsCheck ()
Diversification ()
PopulationUpdate ()
end for
```

1) Initialization: Commencing with the era of nECP-charged particles in the search space, the ECPO, like other population-based metaheuristics, sees the sorting of these charged particles from the finest to the worst. The charged particles in this ECPO version are generated randomly utilizing an ordinary dispersion method. Nevertheless, the implementation of any other strategy for creating the initial ECPs can be comfortably carried out by the user.

2) Archive pool: Alongside the generated population, an archive pool, denoted as archiveECP and of a predetermined size naECP, is established and populated with the best ECPs. The role of this archive is to retain only the finest ECPs, as will be detailed later. The archive ECP is updated at the conclusion of each cycle.

3) Selection: Selecting the appropriate ECPs is a critical step that significantly influences the algorithm's functionality and the results of subsequent phases. In the ECPO algorithm, a random set of charged particles, denoted as nECPI, is selected from the population. Subsequently, these particles are arranged in order from the worst to the best. The chosen particles undergo the interaction phase in accordance with the specified plan.

4) Interaction: As previously noted, not all ECPs engage in communication with each other; only a selected subset, determined by nECPI, participates in this phase. In this stage, the chosen nECPI particles interact with each other in diverse ways, as specified in the plan. For instance, consider a scenario where nECPI = 3 (this applies to any other value of nECPI as well). These particles are arranged from the best to the worst and denoted as. The particles denoted as ECP_1, ECP_2, and ECP_3 are arranged from the best to the worst. The overall best particle is represented as ECP_{best}.

- Strategy 1

In the initial strategy, communication occurs among the chosen ECP utilizing only the best overall ECP, denoted as ECP_{best}, and one other ECP at a time. In this specific scenario where three ECPs are involved in the interaction, each ECP generates two new particles labeled $ECP_{i\ new\ 1}$ and $ECP_{i\ new\ 2}$ (where i represents the index of the chosen ECP).

- For ECP_1 :

Initially, it is simultaneously influenced by ECP_2 and ECP_{best} to transition to $ECP_{1\ new\ 1}$. Subsequently, ECP_1 is influenced concurrently by ECP_3 and ECP_{best} to transition to $ECP_{1\ new\ 2}$. The consequent force required to move ECP_1 to $ECP_{1\ new\ 1}$ is given by:

$$F = F_{b1} + F_{21} \tag{2}$$

In this context, F_{b1} denotes the force exerted by ECP_{best}, and F_{21} represents the force exerted by ECP_2 on $ECP_{1.}$.

Here is how these two forces are expressed:

$$F_{b1} = \beta \times (\, ECP_{best} - ECP_1\,) \tag{3}$$

$$F_{21} = \beta \times (ECP_1 - ECP_2\,) \tag{4}$$

The random number β can be generated utilizing various distributions.

The forces can be expressed to indicate that ECP_{best} attracts ECP_1 (since ECP_{best} is superior to ECP_1), while ECP_2 repels ECP_1 (as ECP_2) is inferior to ECP_1).

Therefore, the cumulative force driving ECP_1 to transition to $ECP_{1\ new\ 1}$ is determined by:

$$\begin{aligned} ECP_{1\ new\ 1} &= ECP_1 + F \\ &= ECP_1 + F_{b1} + F_{21} \\ &= ECP_1 + \beta \times (\, ECP_{best} - ECP_1\,) + \beta \times (ECP_1 - ECP_2\,) \end{aligned} \tag{5}$$

Similarly, the total force propelling ECP_1 to transition to $ECP_{1\ new\ 2}$ can be expressed as:

$$\begin{aligned} ECP_{1\ new\ 2} &= ECP_1 + F \\ &= ECP_1 + F_{b1} + F_{31} \\ &= ECP_1 + \beta \times (\, ECP_{best} - ECP_1\,) + \beta \times (ECP_1 - ECP_3\,) \end{aligned} \tag{6}$$

- For ECP_2 :

Initially, it is concurrently influenced by ECP_1 and ECP_{best} to transition to $ECP_{2\ new\ 1}$. Following that, ECP_2 is simultaneously influenced by ECP_3 and ECP_{best} to move to $ECP_{2\ new\ 1}$. The forces acting on ECP_2 share the same expressions as Eq. (5) as well as Eq. (6).

- For ECP_3 :

The third particle experiences a dual influence, initially from ECP_1 and ECP_{best}, leading to its transition to $ECP_{3\ new\ 1}$. Subsequently, ECP_3 is simultaneously affected by ECP_2 and ECP_{best}, directing its movement to $ECP_{3\ new\ 2}$.

- Strategy 2

In the second strategy, ECP_{best} is not connected with the remaining ECPs, and the interaction is carried out on the selected ECP using all the remaining interacting ECPs. Subsequently, in the outlined scenario where there are three interacting ECPs, one new ECP is generated by each ECP, referred to as $ECP_{i\ new}$ (where i denotes the index of the chosen ECP).

- For ECP_1 :

ECP_1 is simultaneously influenced by ECP_2 as well as ECP_3, inducing movement to $ECP_{1\ new}$. The resulting force required to transition ECP_1 to $ECP_{1\ new}$ is expressed by:

$$F = F_{21} + F_{31} \tag{7}$$

F_{31} represents the force exerted by ECP_3 on ECP_1, and F_{21} is the force exerted by ECP_2 on ECP_1. Therefore, the cumulative force propelling ECP_1 to transition to $ECP_{1\ new}$ is determined by:

$$ECP_{1\ new} = ECP_1 + F_1$$

$$= ECP_1 + F_{21} + F_{31}$$

$$= ECP_1 + \beta \times (ECP_1 - ECP_2) + \beta \times (ECP_1 - ECP_3)(8)$$

- For ECP_2:

The second particle, ECP_2, is concurrently influenced by the first and third particles (ECP_1 and ECP_3), resulting in movement to $ECP_{2\ new}$. The resulting force required to transition ECP_1 to $ECP_{1\ new}$ is expressed by:

$$F = F_{12} + F_{32} \tag{9}$$

F_{12} is the force exerted by ECP_1 on ECP_2, and F_{32} is the force exerted by ECP_3 on ECP_2. The overall force propelling ECP_2 to transition to $ECP_{2\ new}$ is determined by Eq. (8).

- For ECP_3 :

ECP_3 is simultaneously influenced by ECP_1 and ECP_2 with the force expressed as:

$$F = F_{13} + F_{23} \tag{10}$$

F_{13} is the force exerted by ECP_1 on ECP_3, and F_{23} is the force exerted by ECP_2 on ECP_3. The detailed expressions for F_{13} and F_{23} are akin to the expressions in Eq. (9). Consequently, $ECP_{3\ new}$ will transition to $ECP_{3\ new}$.

- Strategy 3

In the third strategy, a combination of the first and second strategies is applied to generate new ECPs. Consequently, for the illustrated scenario where nECPI = 3, nine new ECPs are generated, with 6 resulting from strategy 1 and 3 from strategy 2. The equations previously described are applicable in this context.

At the conclusion of the interaction phase, a set of ECPs is termed new-ECP, and its size remains consistent with the original ECP population. This remains true regardless of the chosen nECPI or the strategy utilized. In simpler terms, if the process begins with 30 particles, the population size remains 30 particles after the interaction phase, irrespective of the strategy or the number of particles involved in the interaction.

In the final phase of ECPO, the newly generated ECPs are subject to bounds checks to ensure they fall within the defined search space. If any ECPs are found to exist outside these bounds, adjustments are made accordingly. Subsequently, a subset of the newly created ECPs undergoes expansion based on the probability of diversification (Pd). The diversity operator, integral to ECPO, incorporates information from both the new ECP population (newECP) and the existing archive pool (archiveECP).

Following the diversification phase, the population is updated by aligning the new population with the previously established archive pool. The best nECP particles, ranked from 1 to nECO, shape the updated population. This refined population then undergoes the same procedure as explained earlier for another cycle.

In terms of termination, the current version of ECPO concludes after iterating MAXIter times, utilizing the various phases described above. However, users retain the flexibility to terminate the process differently if desired.

C. Artificial Rabbits Optimization (ARO)

ARO, or Adaptive Rabbit Optimization, draws its inspiration from the resourceful survival techniques employed by rabbits in their natural surroundings. These techniques, intricately designed to outwit predators and ensure effective evasion, form the foundation of ARO. The algorithm assimilates the foraging and hiding strategies inherent in rabbits, along with their adept energy management, creating a dynamic framework that seamlessly transitions between these strategic modes [34].

1) Detour foraging: During the quest for sustenance, a distinctive detour foraging strategy is observed in rabbits, with a focus on distant food sources and a potential oversight of nearby ones. Within the ARO framework, a community of rabbits is envisioned, each possessing its designated territory comprising burrows and grass. Encounters among these rabbits at each other's foraging sites occur randomly. In this scenario, a mathematical model is presented to articulate the deviation search behavior demonstrated by rabbits.

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

$$i, j = 1, \ldots, n \; and \; j \neq 1 \tag{11}$$

$$S = M \times v \tag{12}$$

$$M = (e - e^{\left(\frac{t-1}{I}\right)^2}) \times \sin(2\pi r_2) \tag{13}$$

$$v(y) = \begin{cases} 1 & if \quad y = f(1) \\ 0 & else \end{cases} \quad k = 1, \ldots, d \; and \; l = 1, \ldots, \lceil r_3 \times d \rceil \tag{14}$$

$$f = p(d) \tag{15}$$

$$m_1 = N(0,1) \tag{16}$$

In the given framework, n denotes the quantity of rabbits within the population, while d represents the dimension of the problem. The position of the $i-th$ rabbit at time t + 1 is denoted by $\vec{B}_i(t+1)$. The variable n_1 is distributed according to the standard normal distribution. T signifies the maximum number of iterations, and $x_i(t)$ denotes the position of the $i-th$ rabbit at time t . The variable p generates a random rearrangement (permutation) of integers ranging from 1 to d. Additionally, w is a mapping tool within the algorithm, facilitating the random selection of elements from the explorer to introduce variation in the search process. The random numbers r_1, r_2, r_3 fall within the range of (0, 1). Lastly, S is introduced to represent the run length, symbolizing the speed of movement during detour foraging in the algorithm. This comprehensive set of parameters and variables collectively defines the key components and dynamics of the rabbit optimization algorithm.

2) Random hiding: To secure their survival, rabbits demonstrate a proclivity for randomly selecting one of their burrows as a shelter. The mathematical model capturing this stochastic shelter-seeking behavior is articulated through the following equations. The formulation of the $j-th$ burrow of the $i-th$ rabbit is expressed as:

$$\vec{B}_i(t+1) = x_i(t) + N \times f \times \vec{x}_i(t), \;\; i,j = 1,\dots,n \; and \; j \neq 1 \tag{17}$$

$$D = \frac{I-t+1}{I} \times r_4 \tag{18}$$

$$m_2 = N(0,1) \tag{19}$$

$$f(y) = \begin{cases} 1 & if \;\; y = g(1) \\ 0 & \;\;\;\;\; else \end{cases} \;\; k = 1,\dots,d \tag{20}$$

$$\vec{R}_{i,r}(t) = \vec{x}_i(t) + \; N \times f \times \vec{x}_i(t) \tag{21}$$

The parameter of hiding, denoted as N, undergoes linear reduction throughout the iteration process from 1 to $\frac{1}{I}$ with the incorporation of a random perturbation.

In the final stages of implementing either the random hiding or detour foraging strategies, the update to the position of the $i-th$ rabbit adheres to the formula outlined in Eq. (22):

$$\vec{x}_i(t+1) = \begin{cases} \vec{x}_i(t) & g(\vec{x}_i(t)) \leq g\left(\vec{B}_i(t+1)\right) \\ \vec{B}_i(t+1) & g(\vec{x}_i(t)) > g\left(\vec{B}_i(t+1)\right) \end{cases} \tag{22}$$

3) Energy shrinks: Due to the recurrent cycles of detour foraging and random hiding, the energy level of the rabbits gradually diminishes. Therefore, the incorporation of an energy factor becomes crucial within the ARO framework:

$$E(t) = 4\left(1 - \frac{t}{I}\right) ln\frac{1}{r} \tag{23}$$

The algorithmic steps for ARO in Algorithm 2, are displayed as Pseudo-code form as well as Fig. 2 shows the flowchart of ARO.

ALGORITHM 2: PSEUDO-CODE OF ARO ALGORITHM

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

D. Data Collection

The primary objective of this investigation is to create a robust framework for the precise assessment of student's academic achievements, taking into account contextual nuances. A crucial step in this process involves the thorough preprocessing of the dataset, wherein textual data undergoes conversion into numerical values. This transformation forms the foundation for the application of ML algorithms and advanced statistical methodologies, facilitating a comprehensive analysis of the dataset. The diverse variables within the dataset are systematically categorized, ensuring a structured approach to understanding and predicting academic performance. This strategic approach aims to enhance the accuracy and effectiveness of the assessment, providing valuable insights into the complex landscape of students' academic achievements.

The research incorporates a comprehensive set of inputs to explore various dimensions influencing students' academic performance.

- The Student Demographics category covers details such as Marital Status, Nationality, and Gender, including additional factors like being a Displaced Candidate and the age at enrollment.
- Parental Information delves into the educational qualifications and occupations of both mother and father, providing insights into the familial context.
- Financial and Support Information offers a comprehensive view of students' financial backgrounds, including their academic fee situation, scholarship status, and potential debt.
- Economic Indicators introduce external factors like the Academic Unemployment Rate, Educational Inflation Rate, and GDP, providing contextual economic insights.
- Enrollment Information examines the mode and order of application, as well as the specialized field of study chosen by students.

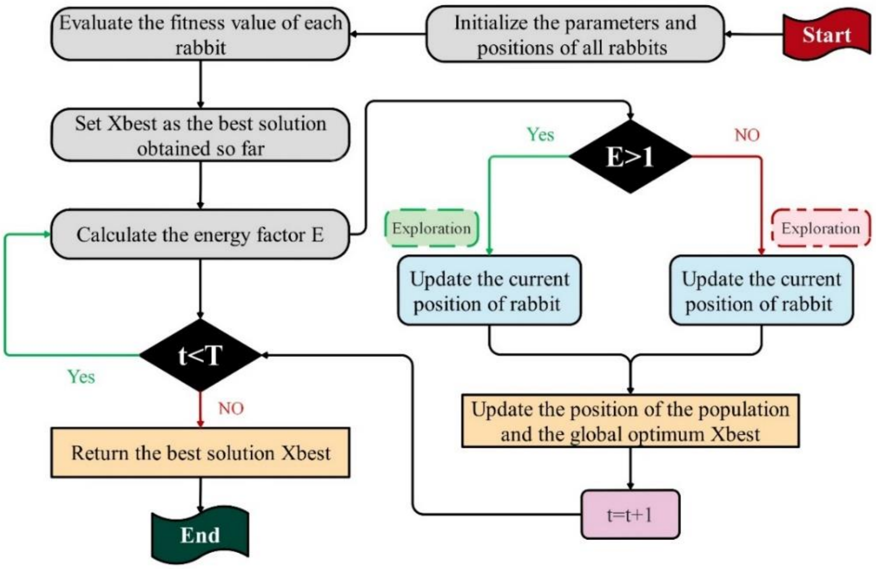

Start
Initialize the parameters and positions of all rabbits
Evaluate the fitness value of each rabbit
Set Xbest as the best solution obtained so far
Calculate the energy factor E
E>1
Yes
NO
Exploration
Exploration
Update the current position of rabbit
Update the current position of rabbit
Update the position of the population and the global optimum Xbest
t=t+1
t<T
Yes
NO
Return the best solution Xbest
End

Fig. 2. Flowchart of ARO.

- Finally, Academic Performance metrics include attendance regimes, past educational credentials, and a comprehensive breakdown of curricular units, encompassing enrollment, evaluation, approval, grading, and units without evaluation.

This multifaceted approach ensures a nuanced analysis, considering diverse aspects that collectively contribute to the intricate landscape of student academic performance [17].

In Fig. 3, the visual representation intricately displays the impact of inputs on each other and, crucially, on the target variable. The color spectrum, from white (positive impact) to purple (negative impact), guides the discernment of dynamics at play. Parameters show a self-reinforcing nature, evident from bold white along the main diameter. The final line outlining each input's impact on the target reveals crucial insights. Inputs like Tuition Fees Up to Date, Scholarship Recipient, Curricular Units (evaluations), Curricular Units (approved), and Curricular Units (grade) emerge as influential with substantial positive impact.

Conversely, Debtor, Gender, and Age at Enrollment show predominantly negative impacts. The remaining parameters in pale colors signify minimal influence, underlining limited significance in the predictive context. This visual analysis offers a nuanced perspective on dataset relationships, guiding focus toward the most impactful variables for predictive modeling.

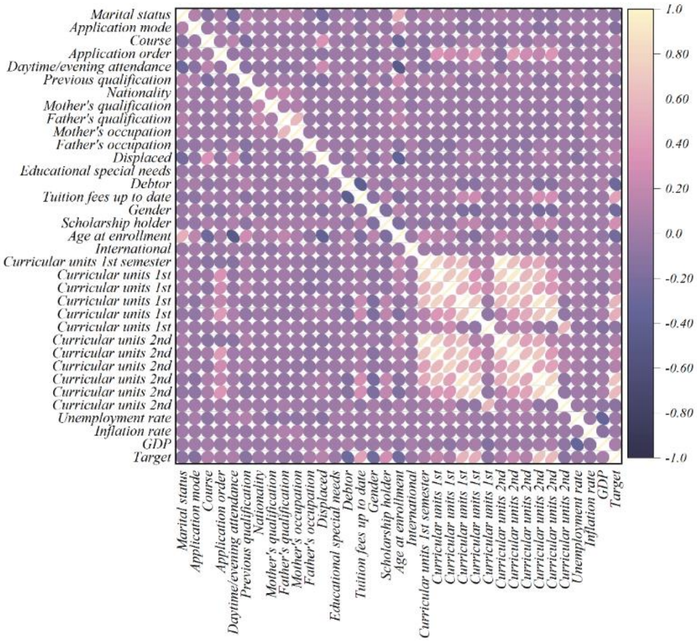


Fig. 3. Correlation matrix for the input and output variables.

E. Hyperparameter

Table I presents the hyperparameters used for the different developed models in this study: RFC, RFAR, and RFEC. These hyperparameters significantly influence the performance and efficiency of each model.

TABLE I. RESULT OF HYPERPARAMETERS FOR THE DEVELOPED MODELS

Models	n_estimators	max_depth	min_samples_split	min_samples_leaf
RFC	20	10	2	1
RFAR	123	1311	2	1
RFEC	56	64	2	1

IV. RESULTS

A. Model Applicability Assessment

In assessing classification problems, Accuracy is a commonly employed metric to gauge a model's overall performance. This metric relies on four essential elements: True Positives (Tp) for accurate positive predictions, True Negatives (Tn) signifying precise negative predictions, False Positives (Fp) representing inaccurate positive predictions, and False Negatives (Fn) indicating incorrect negative predictions. However, the utility of Accuracy diminishes in situations involving imbalanced data, where it tends to favor the majority class, limiting its interpretability. To address this drawback, three additional evaluation metrics—namely Recall, Precision, and F1-Score—are frequently utilized. These metrics provide a more nuanced understanding of a model's performance, particularly in the presence of imbalanced class distributions. Presented through mathematical equations, typically numbered from 24 to 27, these metrics collaboratively contribute to a refined and comprehensive assessment of the effectiveness of a classification model.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \tag{24}$$

$$Precision = \frac{Tp}{Tp+Fp} \tag{25}$$

$$Recall = TpR = \frac{Tp}{P} = \frac{Tp}{Tp+Fn} \tag{26}$$

$$F1\_score = \frac{2\times Recall \times Precision}{Recall+Precision} \tag{27}$$

B. Convergence Results

The utilization of a convergence diagram is a prevalent practice in scientific discourse, employed to visually elucidate the progression of convergence or optimization inherent in a model or algorithm across successive iterations. This methodology finds frequent application in diverse domains, including but not limited to machine learning, optimization techniques, and computational science. In the context of this article, the convergence diagram, as illustrated in Fig. 4, functions as a tool for juxtaposing the convergence trajectories of the two optimized iterations of the RFC model, namely RFEC and RFAR. A discerning analysis of the diagram reveals that, during the initial iterations, the RFEC model attains a superior convergence level compared to the RFAR model. Notably, the RFEC model sustains this superiority throughout subsequent iterations, culminating in its establishment as the preeminent model. This visual representation effectively communicates the distinctive convergence dynamics of the two optimized models, substantiating the designation of the RFEC model as the optimal choice within the scope of this study.

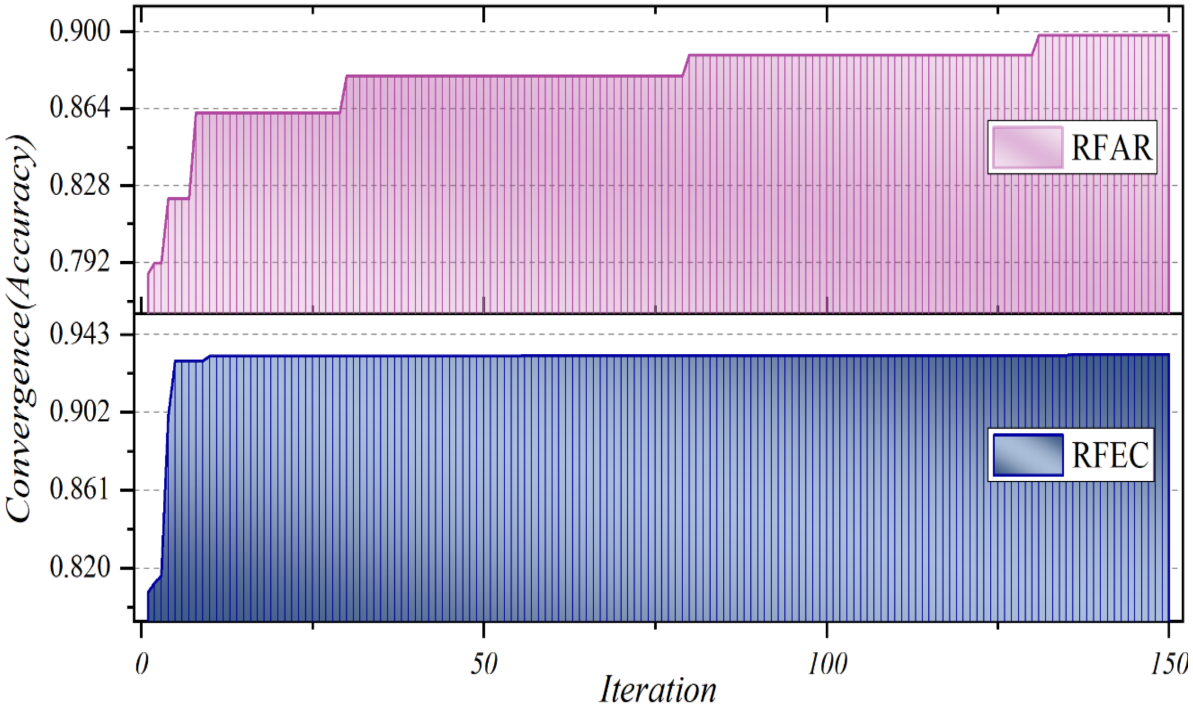


Fig. 4. Line plot for convergence of hybrid models.

C. Hyperparameter

D. Comparing Results of Predictive Models

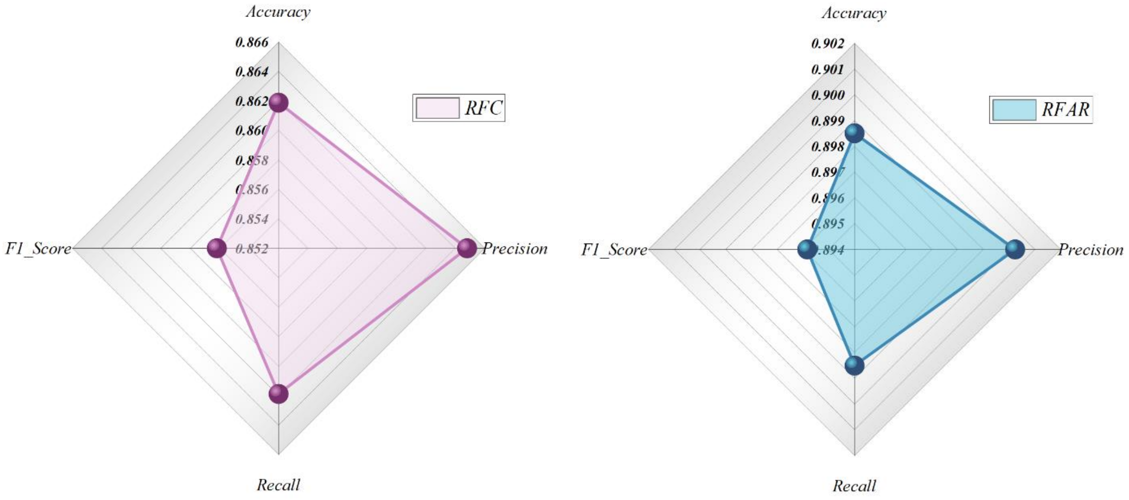
Table II encapsulates a comprehensive compilation of outcomes emanating from the formulated RFC models, facilitating a nuanced comprehension of their performances. Concurrently, Fig. 5 employs a radar plot to present an evaluative comparison among these models. The objective is to discern the model that exhibits the highest precision in predictions when contrasted with real-world outcomes. A substantial proportion of the dataset undergoes rigorous training, and the remaining values are meticulously subjected to testing. The results, spanning the entire dataset, are systematically documented. The pivotal parameter for model assessment lies in all datasets, graphically elucidated in Fig. 5. The evaluation of accuracy is conducted across three distinct phases: Train, Test, and All. In the training phase, the RFEC model emerges as the frontrunner, boasting an accuracy of 0.9997, outpacing the RFC model at 0.9060 and the RFAR model at 0.8949. Transitioning to the testing phase, the RFAR model excels with an accuracy of 0.8985, surpassing the RFC model at 0.7589 and the RFEC model at 0.7566. Remarkably, in the comprehensive all data groups, the RFEC model ascends to the zenith with an accuracy of 0.9326, followed by the RFAR model at 0.8985 and the RFC model at 0.8619. The visual representation encapsulated in Fig. 5 distinctly underscores the discernible superiority of the RFEC model in predictive accuracy, affirming its prominence among the models considered.

E. Classification Outcomes

Table III provides a detailed breakdown of Precision, Recall, and F1-score metrics concerning the classification of 4424 students based on their academic performance. These tabulated metrics offer valuable insights, illuminating the model's precision in positive predictions, ability to accurately identify true positives, and overall effectiveness in classifying students based on their academic achievements. The precision values reflect the accuracy of the model in making positive predictions, while recall signifies the model's ability to capture true positives. Additionally, the F1-score offers a comprehensive measure that balances precision and recall, providing a holistic assessment of the model's performance in classifying students across various academic performance categories. These tables play a pivotal role in the comprehensive evaluation of the model's effectiveness in handling diverse aspects of academic performance prediction.

TABLE II. RESULT OF DEVELOPED MODELS FOR RFC

Phase	Index values	Models		
		RFC	RFAR	RFEC
Train	Accuracy	0.9060	0.8949	0.9997
	Precision	0.9123	0.8971	0.9997
	Recall	0.9060	0.8949	0.9997
	F1-score	0.9031	0.8924	0.9997
Test	Accuracy	0.7589	0.8985	0.7566
	Precision	0.7496	0.8999	0.7495
	Recall	0.7589	0.8985	0.7566
	F1-score	0.7446	0.8963	0.7449
All	Accuracy	0.8619	0.8985	0.9326
	Precision	0.8648	0.9002	0.9258
	Recall	0.8619	0.8985	0.9268
	F1-score	0.8562	0.8958	0.9258



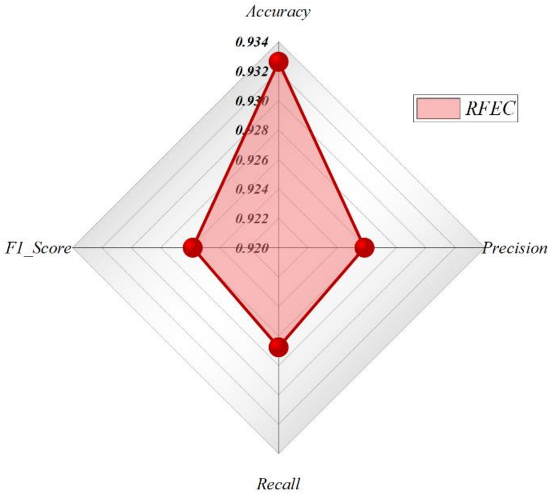


Fig. 5. Radar plot for achievement of developed models based on evaluators.

1) Precision: In this evaluative index, a meticulous examination of each model's performance across distinct categories elucidates the RFEC model's prominence. Notably, its proximity to the numerical value 1 stands out conspicuously, surpassing the comparative performance of other models. This observation underscores the RFEC model's superior precision and effectiveness in positive predictions within the evaluated groupings.

2) Recall: In this evaluation index, uniform performance is observed across all models in the graduate group, registering an identical score of 0.98. Contrarily, within the dropout and enrollment categories, the singular RFC model exhibited inferior performance compared to other models. Further scrutiny reveals notable distinctions between the two optimized models: the RFEC model outperforms the RFAR model by 14.08% in the enrollment group and 4.59% in the dropout group. These discernible variations underscore the efficacy of optimization, particularly emphasizing the superior predictive capabilities of the RFEC model in specific academic performance categories.

3) F1-score: In this performance index, the RFEC model emerges as the most adept, achieving a commendable accuracy rate of 93% in dropout predictions, 85% in enrollment, and 95% in graduation. These results position the RFEC model as the standout performer, showcasing its efficacy in accurately predicting students' academic outcomes across diverse performance categories.

In Fig. 6, a comprehensive examination is conducted, contrasting the predictive performance of models against actual measured values. Notably, within the dropout category, the RFEC model exhibits a superior level, closely aligning with the measured values, indicating accurate predictions. Similarly, in the enrollment grouping, the RFEC model demonstrates predictions that closely resemble reality. However, in the graduate category, the RFAR model outperforms other models. In summary, the RFEC model displays superior performance compared to its counterparts, and while its performance in the graduate group may appear suboptimal, its overall predictive accuracy in other categories warrants commendation.

TABLE III. EVALUATION INDEXES OF THE PERFORMANCE OF DEVELOPED MODELS

Model	Situation	Index values		
		Precision	Recall	F1-score
RFC	Dropout	0.93	0.83	0.88
	Enrolled	0.82	0.59	0.69
	Graduate	0.83	0.98	0.9
RFAR	Dropout	0.95	0.87	0.91
	Enrolled	0.87	0.71	0.78
	Graduate	0.88	0.98	0.93
RFEC	Dropout	0.96	0.91	0.93
	Enrolled	0.89	0.81	0.85
	Graduate	0.92	0.98	0.95

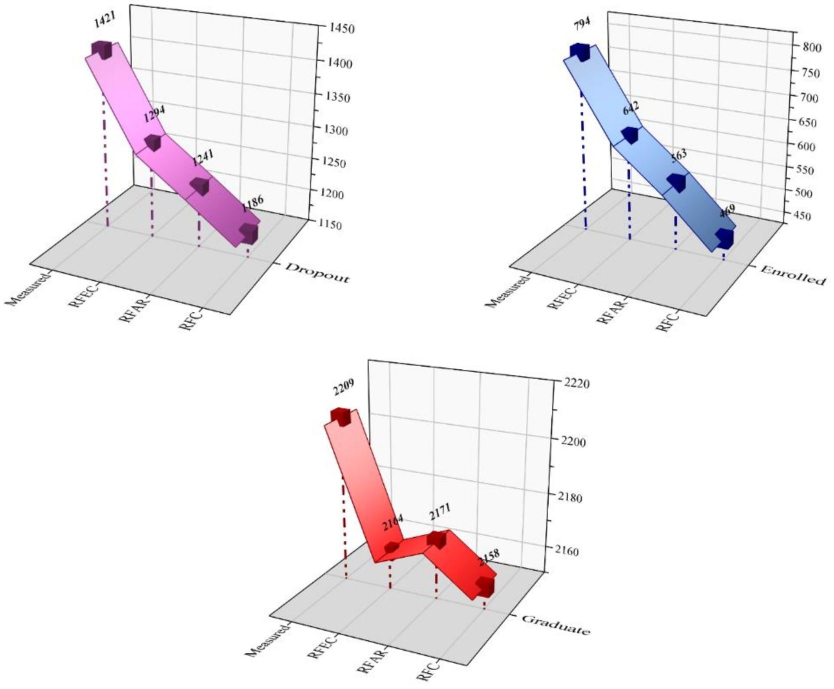


Fig. 6. 3D Ribbon plot for the comparison between the measured and predicted values.

In Fig. 7, a detailed representation of the confusion matrix unveils the accurate classification of students and instances of misclassifications. Within the RFEC model, a meticulous examination reveals the accurate classification of 4100 out of 4424 students across various academic grades. Specifically, within these classifications, 1294 students were accurately identified in the Dropout category, 642 in the Enrolled category, and 2164 in the Graduate category. However, the model exhibited 324 instances of misclassification. In contrast, the RFAR model displayed 449 misclassifications, while the conventional RFC model accurately misclassified 611 students. This comprehensive breakdown provides valuable insights into the models' performance, aiding in a nuanced understanding of their strengths and limitations in accurately categorizing students into their respective classes.

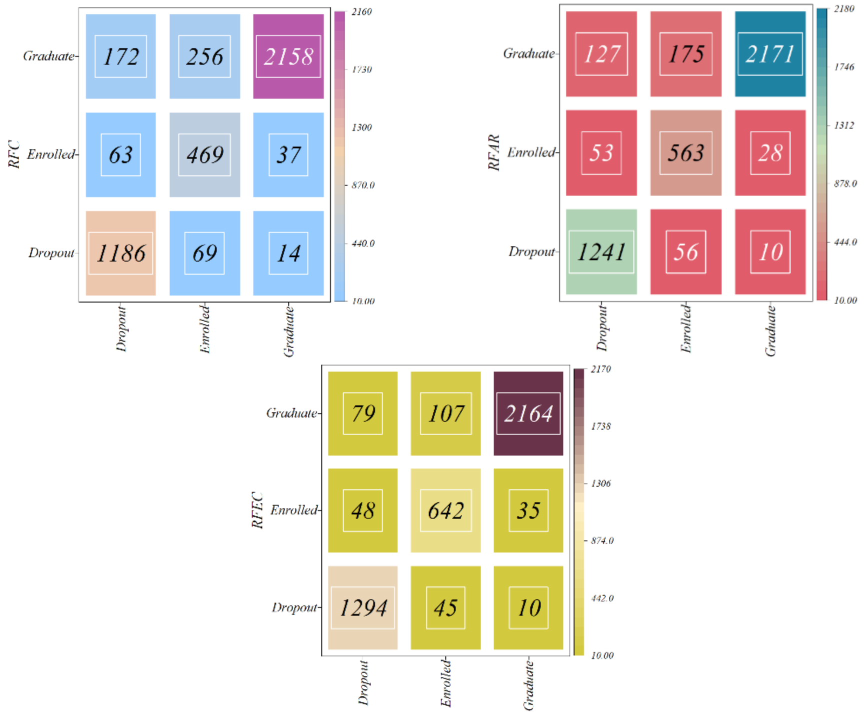


Fig. 7. Confusion matrix for each models' accuracy.

In Fig. 8, a comprehensive analysis of model performance is facilitated through the presentation of two Receiver Operating Characteristic (ROC) charts strategically overlaid for enhanced comparison. The ROC charts visually depict the trade-off between true positive and false positive rates across different classification thresholds. The optimal model in this context is identified by a larger area under the curve (AUC), denoting superior discriminatory power. The meticulous examination of the presented ROC charts leads to the unequivocal identification of the RFEC model as the most efficacious among its counterparts. This determination is substantiated by the model's early attainment of the number 1 true positive rate, signifying its prompt and accurate identification of positive instances. Additionally, a discernible concentration of the ROC curve below the graph further underscores the RFEC model's exceptional performance. These nuanced observations collectively position the RFEC model as the optimal choice, demonstrating a superior ability to balance true positive and false positive rates and thereby affirming its efficacy in classification tasks.

F. Analysing Input Variables

Fig. 9 presents the impact of the presented input variables on the performance of the students. Student success in education is a complex issue with many contributing factors, including the ability to pay tuition fees. Input variables, such as student demographics, home environment, and school resources, play a crucial role in determining student financial well-being and their ability to meet tuition obligations. Understanding how these factors influence tuition fee payment is essential for developing effective interventions to improve student outcomes and promote financial equity in education

Student demographics, such as socioeconomic status, gender, race, and ethnicity, are among the most significant input variables influencing tuition fee payment. Studies have consistently shown that students from low-income families tends to struggle more with tuition payments than their peers from higher-income households. This financial burden is often exacerbated by disparities in access to scholarships and financial aid, leaving students from low-income backgrounds at a greater risk of tuition delinquency or default.

Gender also plays a role in tuition fee payment, with boys generally facing more financial challenges than girls. This disparity may stem from differences in employment opportunities, access to financial resources, and cultural expectations. Race and ethnicity are also associated with variations in tuition fee payment. For instance, African American and Hispanic students tend to have lower rates of tuition payment compliance than White and Asian students. These payment disparities are likely due to a combination of factors, including historical discrimination, unequal access to financial aid, and disparities in parental education.

The home environment is another critical input variable that influences tuition fee payment. Parental involvement in education is particularly important, as it has been shown to positively impact student financial literacy and budgeting skills. Parents who actively support their children's financial education, such as teaching them about saving, budgeting, and the importance of paying bills, can help students make informed decisions about their tuition obligations. Additionally, a supportive and nurturing home environment, free from conflict and stress, can foster a positive financial mindset that promotes responsible fiscal behaviour.

School resources, such as financial aid counseling services and tuition assistance programs, play a crucial role in helping students meet their tuition obligations. Schools that provide comprehensive financial literacy education and accessible financial aid resources can empower students to make informed decisions about their financial aid options and better manage their tuition payments. Additionally, schools with strong partnerships with community organizations and financial institutions can expand access to scholarships, work-study programs, and other forms of financial assistance that can alleviate the burden of tuition payments.

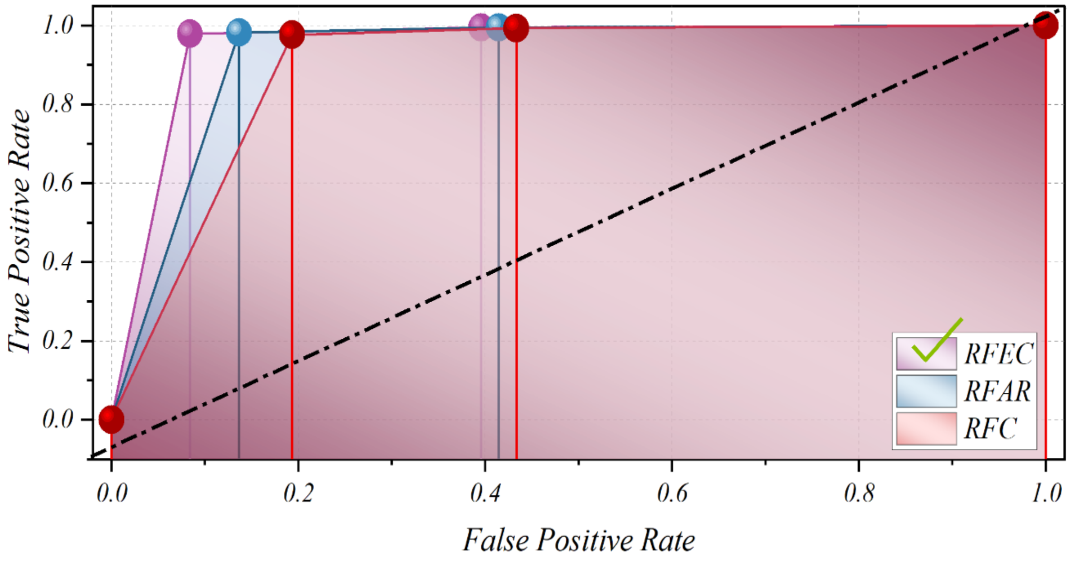


Fig. 8. ROC curve for developed models.

The impact of input variables on tuition fee payment has significant implications for educational practice. It is crucial to recognize that financial well-being is not solely determined by individual effort or ability; it is also shaped by factors beyond a student's control. Educators, policymakers, and community leaders must work together to address the inequities that exist in education and create a more equitable financial environment for all students. This includes providing targeted financial literacy education and support services to students from low-income backgrounds, promoting parental involvement in financial education, and expanding access to scholarships, work-study programs, and other forms of financial assistance.

Input variables, such as student demographics, home environment, and school resources, play a critical role in determining a student's ability to pay tuition fees. By understanding the impact of these variables, educators, policymakers, and community leaders can develop effective interventions to improve student financial well-being, promote financial equity in education, and ensure that all students have the opportunity to reach their full potential.

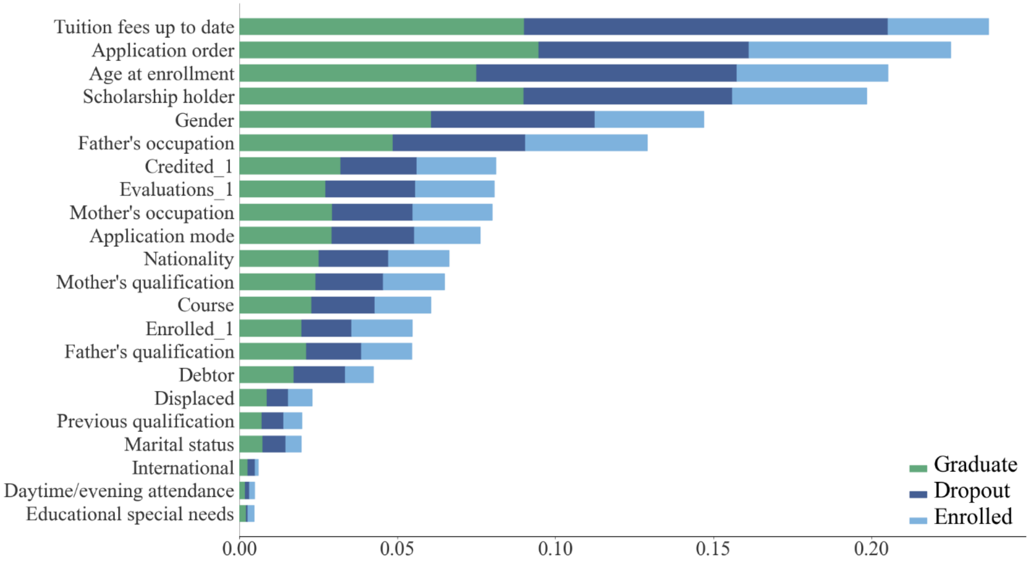


Fig. 9. The SHAP sensitivity analysis of the models.

G. Discussion

Table IV compares the accuracy of the present study's RFEC model with several published models. The RFEC model achieved the highest accuracy at 92.58%, significantly outperforming others. The superior performance of the RFEC model is attributed to the integration of advanced optimization algorithms (ECPO and ARO), which enhance the model's ability to handle complex data. This demonstrates the effectiveness of using sophisticated machine learning techniques and optimization to improve predictive accuracy in educational research.

TABLE IV. PRESENT MODEL EVALUATION WITH PUBLISHED STUDIES

Study	Developed Models	Accuracy
Present study	RFEC	92.58%
Kabakchieva [35]	DTC	72.74%
Bichkar and R. R. Kabra [36]	DTC	69.94%
Nguyen and Peter [37]	DTC	82%
Edin Osmanbegovic et al. [38]	NBC	76.65%

V. CONCLUSION

This research has strategically applied predictive data mining modeling, specifically employing the potent Random Forest Classifier (RFC), to address challenges within the academic domain proactively. The primary aim is to empower educators with the capacity to intervene timely, thereby enhancing academic trajectories, reducing failure rates, elevating the overall educational experience, and fostering an environment conducive to improved student outcomes. The single RFC model exhibited suboptimal performance in predictive modeling. To enhance its efficacy, two optimization algorithms, Electric Charged Particles Optimization (ECPO) and Artificial Rabbits Optimization (ARO), were incorporated. This integration led to the creation of two new optimized models, namely RFEC and RFAR. The utilization of these two optimizers represents a pioneering initiative in the domain of student performance forecasting, signifying a noteworthy advancement for future research and applications in this field. This article conducts an analysis and prediction of information data about 4424 students based on their previous enrollment, graduation, and dropout records. Additionally, a comparative assessment of each model's results against the actual measured values is performed to ascertain the optimal predictive model. The outcomes, accompanied by pertinent tables and figures, indicate that the RFEC model exhibits the smallest deviation, approximately 7.32%, in contrast to the actual measured values. This stands in contrast to the RFAR-optimized model, which demonstrates a higher difference of about 10.14%, and the RFC single model, showcasing a more substantial difference of approximately 13.81% from the measured values.

VI. Future Study

Future research should focus on expanding data sources to include socio-economic backgrounds and extracurricular activities for a comprehensive understanding of student performance. Collaborating with interdisciplinary experts can enrich analyses by considering individual characteristics, social dynamics, and institutional practices. Longitudinal studies are essential for tracking academic trajectories and developing proactive intervention strategies. Validating predictive models across diverse contexts and populations is crucial for ensuring scalability and effectiveness. Ethical guidelines must be prioritized for transparent and accountable deployment of predictive analytics. Exploring emerging technologies like artificial intelligence offers opportunities to enhance personalized learning experiences. By addressing these areas, future studies can contribute to advancing predictive analytics in higher education and fostering a more inclusive learning environment.

References

[1] Behr A, Giese M, Teguim Kamdjou HD, Theune K. Motives for dropping out from higher education—An analysis of bachelor's degree students in Germany. Eur J Educ 2021;56:325–43.

[2] Kehm BM, Larsen MR, Sommersel HB. Student dropout from universities in Europe: A review of empirical literature. Hungarian Educational Research Journal 2019;9:147–64.

[3] Atchley W, Wingenbach G, Akers C. Comparison of course completion and student performance through online and traditional courses. International Review of Research in Open and Distributed Learning 2013;14:104–16.

[4] Al-Shehri H, Al-Qarni A, Al-Saati L, Batoaq A, Badukhen H, Alrashed S, et al. Student performance prediction using support vector machine and k-nearest neighbor. 2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE), IEEE; 2017, p. 1–4.

[5] Heublein U, Richter J, Schmelzer R, Sommer D. Die Entwicklung der Schwund-und Studienabbruchquoten an den deutschen Hochschulen. HIS: Forum Hochschule, vol. 3, 2012, p. 7.

[6] Schröder-Gronostay M, Daniel HD. Studienerfolg und Studienabbruch: Beiträge aus Forschung und Praxis. Luchterhand; 1999.

[7] Ziegele F. Grundlagen der Analyse von Studienabbrüchen: Erfassung, Bewertung und Maßnahmen. Beiträge Zur Hochschulforschung 1997;19:435–54.

[8] Asif R, Hina S, Haque SI. Predicting student academic performance using data mining methods. Int J Comput Sci Netw Secur 2017;17:187–91.

[9] Altaher A, BaRukab O. Prediction of student's academic performance based on adaptive neuro-fuzzy inference. International Journal of Computer Science and Network Security (IJCSNS) 2017;17:165.

[10] Hamoud A, Hashim AS, Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence 2018;5:26–31.

[11] Pyle D. Data preparation for data mining. morgan kaufmann; 1999.

[12] Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. Ieee Access 2017;5:15991–6005.

[13] Baker RS, Inventado PS. Educational Data Mining and Learning Analytics 2018.

[14] Yunita A, Santoso HB, Hasibuan ZA. Deep Learning for Predicting Students' Academic Performance. 2019 Fourth International Conference on Informatics and Computing (ICIC), IEEE; 2019, p. 1–6.

[15] Ameen AO, Alarape MA, Adewole KS. Students' academic performance and dropout predictions: A review. Malaysian Journal of Computing 2019;4:278–303.

[16] Shaleena KP, Paul S. Data mining techniques for predicting student performance. 2015 IEEE international conference on engineering and technology (ICETECH), IEEE; 2015, p. 1–3.

[17] Kannan R, Abarna KTM, Vairachilai S. Student Academic Performance prognosticative Using optimized Hybrid Machine Learning Algorithms 2023.

[18] Brezavšček A, Bach MP, Baggia A. Markov analysis of students' performance and academic progress in higher education. Organizacija 2017;50:83–95.

[19] Spady WG. Dropouts from higher education: An interdisciplinary review and synthesis. Interchange 1970;1:64–85.

[20] Bean JP. Dropouts and turnover: The synthesis and test of a causal model of student attrition. Res High Educ 1980;12:155–87.

[21] Márquez-Vera C, Morales CR, Soto SV. Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje 2013;8:7–14.

[22] Thammasiri D, Delen D, Meesad P, Kasap N. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Syst Appl 2014;41:321–30.

[23] Yukselturk E, Ozekes S, Türel YK. Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and e-Learning 2014;17:118–33.

[24] Spady WG. Dropouts from higher education: An interdisciplinary review and synthesis. Interchange 1970;1:64–85.

[25] Bean JP. Dropouts and turnover: The synthesis and test of a causal model of student attrition. Res High Educ 1980;12:155–87.

[26] Yunita A, Santoso HB, Hasibuan ZA. Deep Learning for Predicting Students' Academic Performance. 2019 Fourth International Conference on Informatics and Computing (ICIC), IEEE; 2019, p. 1–6.

[27] Batool S, Rashid J, Nisar MW, Kim J, Kwon H-Y, Hussain A. Educational data mining to predict students' academic performance: A survey study. Educ Inf Technol (Dordr) 2023;28:905–71.

[28] Chen Y, Zhai L. A comparative study on student performance prediction using machine learning. Educ Inf Technol (Dordr) 2023:1–19.

[29] Asselman A, Khaldi M, Aammou S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments 2023;31:3360–79.

[30] Liu C, White M, Newell G. Measuring the accuracy of species distribution models: a review. Proceedings 18th World IMACs/MODSIM Congress. Cairns, Australia, vol. 4241, 2009, p. 4247.

[31] Ghosh SK, Janan F. Prediction of student's performance using random forest classifier. Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore, 2021, p. 7–11.

[32] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[33] Bouchekara H. Electric Charged Particles Optimization and its application to the optimal design of a circular antenna array. Artif Intell Rev 2021;54:1767–802.

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

[35] Kabakchieva D. Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research 2012;1:686–90.

[36] Kabra RR, Bichkar RS. Performance prediction of engineering students using decision trees. Int J Comput Appl 2011;36:8–12.

[37] Nghe NT, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance. 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports, IEEE; 2007, p. T2G-7.

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