INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Internet Manufacturing and Services

ISSN online: 1751-6056 - ISSN print: 1751-6048
https://www.inderscience.com/ijims

Utilising a Gaussian process classifier integrating with meta-heuristic optimisers to predict and classify performance systems

Kaifeng Huang, Chun Wang

DOI: 10.1504/IJIMS.2025.10074231

Article History:

Received:	14 March 2025
Last revised:	20 May 2025
Accepted:	01 June 2025
Published online:	28 October 2025

Utilising a Gaussian process classifier integrating with meta-heuristic optimisers to predict and classify performance systems

Kaifeng Huang*

College of Electronic and Commerce,
Luoyang Normal University,
Luoyang, 471934, China
Email: hkfly2003@163.com
*Corresponding author

Chun Wang

Changchun Cigarette Factory,
Jilin Tobacco Industry Co., Ltd.,
Changchun, 130000, China
Email: wangchuncy@sina.com

Abstract: This study pioneers academic achievement prediction with a powerful Gaussian process classifier (GPC) model. Advanced optimisation methods like PVSA and smell agent optimisation improve the model's prediction power. These algorithms use machine learning (ML) and bioinspired methods to improve forecasting and decision-making. The main goal is to accurately predict students' comprehensive performance, which improves educational outcomes, especially in higher education, where it helps strategic decision-making and reduce dropout rates. To fully examine the input variables and determine how each component affected student academic performance, machine learning (ML) approaches were used. The research carefully evaluates varied educational datasets using machine learning (ML) to reduce dimensionality. Proactive educators can make data-driven decisions to boost academic performance. By categorising people by their intrinsic strengths and reducing failure rates, the study hopes to improve education. Predictive modelling, especially machine learning (ML), helps the academic community proactively address issues, improving learning environments and student results. The SAO-optimised GPC model outperformed the PVSA-optimised model with 88 correct predictions, as seen by its bigger Area under the receiver operating characteristic (ROC) curve. This shows its high discrimination and performance level classification skills.

Keywords: classification tasks; student performance; machine learning; ML; Gaussian process classifier; GPC; population-based vortex search algorithm; PVSA; smell agent optimisation; SAO.

Reference to this paper should be made as follows: Huang, K. and Wang, C. (2025) 'Utilising a Gaussian process classifier integrating with meta-heuristic optimisers to predict and classify performance systems', *Int. J. Internet Manufacturing and Services*, Vol. 11, No. 5, pp.1–30.

Biographical notes: Kaifeng Huang graduated from Zhengzhou University in 2003 with a Bachelor's degree in Computer Science and Technology. He obtained a Master's degree in Computer Applications from Henan University of Science and Technology in 2012. Currently, he serves as an Associate Professor in the School of E-commerce at Luoyang Normal University. His main research interests include computer applications, e-commerce, and related fields.

Chun Wang graduated from Changchun University with a Bachelor's degree in Computer Science and Technology. At present, he serves as the Deputy Director of Changchun Cigarette Factory, Jilin Tobacco Industry Co., Ltd. In December 2010, he obtained the Professional and Technical qualification of intermediate engineer.

1 Introduction

At the core of every academic institution lies the fundamental objective of fostering a robust learning environment, and the hallmark of effective education lies in the accurate assessment of students. Traditionally, evaluation methods encompassed class tests, assignments, practical and laboratory work, and comprehensive semester examinations (Hernandez and Nunez, 2023). This holistic evaluation process is designed to elucidate the academic standing of each student across various courses within a class. Undoubtedly, student performance emerges as a pivotal aspect within the realm of educational institutions, as emphasised by Solomon et al. (2018). Recognising and accurately gauging the academic achievements of students not only serves as a testament to the efficacy of the educational system but also plays a crucial role in shaping the trajectory of individual learners (Fauzy et al., 2024). The commitment to thorough and comprehensive evaluation remains an integral part of the educational ethos, ensuring a robust and meaningful learning experience for every student (Ameen et al., 2019). A substantial number of students enrolled at the state university encounter various challenges, particularly during their inaugural year (Gomathy and Venkatasbramanian, 2023). Consequently, academic performance during this initial year has been identified as a crucial indicator predicting the likelihood of timely graduation (Alkasi et al., 2024). The inaugural year serves as a pivotal juncture, and successfully navigating its challenges is integral to ensuring a favourable trajectory toward graduation (Joshi et al., 2023).

In the pursuit of enhancing student retention, educators and researchers have delved into a comprehensive exploration of factors influencing retention rates (Krishnan et al., 2024). It is widely posited that the early identification of students facing a heightened risk of academic struggles provides a valuable opportunity for timely intervention (Legista et al., 2024). By promptly implementing necessary measures and interventions, educators can significantly bolster the prospects of these students, consequently elevating the overall graduation rate (Patacsil, 2020). As articulated by Mallincrodt and Sedlacek (1987), the retention rate emerges as a focal point warranting extensive scrutiny in the context of student retention within the university. The acknowledgment of this factor highlights its significance in the broader landscape of educational research, underscoring the imperative of devising strategies for comprehensive student support and success (Basañes et al., 2023).

In research by Quiroz (2000), dropouts were questioned openly about why they had left school. Poor academic performance and family problems were the primary reasons for dropout rates (39% of the time), followed by employment obligations (29% of the time) and teacher-related issues (24% of the time). Additionally, these students were asked to list the aspects of school they liked and didn't. These children loved socialising, professors, athletics, and counsellors, among other aspects of school (Mompel and Lombrio, 2024). However, among their dislikes were overcrowded classrooms, violent gangs, and teachers who were bored and indifferent (Ashifa and Büyük, 2024). It's interesting to note that among the dropouts surveyed, instructors were rated as both the finest and worst aspects of education (Basañes and Alentajan, 2024). This study indicates that the school's environment and effectiveness are also significant factors in the dropout rate. Dalton et al. (2009) also offered a variety of justifications for why students discontinue their education. The most prevalent causes, according to the study, were missing too many school days, receiving poor grades, falling behind on assignments, experiencing issues with instructors and other students, not feeling like one belongs, getting punished, becoming pregnant, obtaining employment, and providing for the family (Oqaidi et al., 2022).

One popular strategy to address the issue of school dropout is the use of machine learning (ML). The development of student prediction systems has been the subject of several studies carried out in industrialised nations (Solis et al., 2018). The rapid advancement of AI (Boden, 1996), ML (El Naqa and Murphy, 2015), and data-driven methodologies (Cerquitelli et al., 2021), coupled with the substantial accumulation of educational data by universities, has ushered in the potential to address the persistent challenge of early school leaving. In response to this burgeoning intersection of technology and education, a novel scientific discipline has emerged: educational data mining (EDM) (Romero and Ventura, 2007). The primary goal of EDM is to proactively tackle issues associated with student attrition by harnessing the wealth of available educational data (Kiss et al., 2019). This interdisciplinary field focuses on deploying advanced analytics to uncover patterns and insights within educational datasets (Padmavathy et al., 2024). A prevalent and consequential research objective within EDM involves the prediction of course grades and grade point average (GPA) (Del Río and Insuasti, 2016; Hellas et al., 2018; Vijayarani et al., 2023). By leveraging the power of predictive modelling, EDM endeavours to identify early warning signs and risk factors contributing to students disengaging from their academic journeys (Moyo and Nithyanantham, 2024). The synthesis of AI and educational data holds significant promise in facilitating targeted interventions and fostering student success, ultimately mitigating the challenges associated with early school leaving (Adekola and Aribisala, 2023).

This study aligns with recent research trends that highlight a growing interest in predicting students at risk of dropout. Del Río and Insuasti (2016) conducted a comprehensive study employing classification approaches on a large and diverse dataset, which includes over 32,500 student records from the University of Washington, to estimate the occurrences of dropouts (Aulck et al., 2016) reliably. Similarly, Yukselturk et al. (2014) examined data mining methods for predicting online program dropout (Yukselturk et al., 2014). Thammasiri et al. (2014) contributed to this conversation by utilising ML techniques to forecast whether first-year students will enroll in a second semester. Their study is notable for giving a thorough summary of relevant papers and comparing different categorisation techniques. Lin et al. (2009) explored the efficacy of

neural networks on a dataset encompassing 1,508 engineering students, incorporating both cognitive and non-cognitive attributes. In-depth research on sociodemographic factors was conducted by Kovacic (2010), who illuminated their possible impact on student dropout rates (Bastareche, 2024).

Furthermore, the research described in Golding and Donaldson 2006) concludes that a student's estimation of their theoretical achievement after earning a degree is greatly influenced by their performance in first-year computer science courses. This study examines the data collected from 85 students in UTECH's School of Informatics and Computer Science during their degree program (Padmanabhan et al., 2023). They discovered that the BSCIT program's total GPA could be accurately predicted by taking first-year basic courses, such as C programming, exposure to computer systems, computer digital and logical design, and three additional courses. By employing statistical methods, specifically regression analysis, they demonstrate a strong relationship that does not require further data mining or classification between the academic success of students in first-grade computer classes and their overall achievement in the BSCIT program. At 0.499, this correlation accounts for 70.6% of the variance in students' academic success (Pulivarthy, 2024). The authors also found no evidence of a connection between academic achievement and student demographics (Nithyanantham, 2023).

The research presented in Zimmermann et al., (2011) utilise RF data mining technology, which is essentially a collection of decision trees (DT). The objective is to predict, based on the achievements of their Bachelor of Science (B.Sc.), the performance of learners at the advanced standing level (M.Sc. or Master of Science). The authors collected information from 176 students majoring in computer science at ETH Zurich who were pursuing university degrees. A total of 135 variables are used in their analysis, which include variables like age, gender, and individual course performance (first and final examination attempts) (Masangu et al., 2021), as well as different study lengths and GPAs (e.g., first-year, second-year, third-year, necessary program, third-year, choice program, etc.). The component they are attempting to anticipate is the GPA for the M.Sc. (Roy and Garg, 2017) course. They discover that a small set of 14 characteristics, the majority of which are scores, account for 55% of the variance in undergraduate output (Venkatasubramanian et al., 2023). Pradeep et al. (2015) examined characteristics influencing student performance using data mining approaches for education, providing insightful information for the predictive modelling of dropout rates. In a different strategy, Burgos et al. (2018) employed logistic regression models to predict dropout from an online learning environment, along with a carefully designed tutoring action plan aimed at reducing the dropout rate. The domain has also seen widespread use of DT classifiers (Dekker et al., 2009; Shaleena and Paul, 2015), with a comprehensive review of educational dropout prediction encapsulated in Kumar et al. (2017).

While prior research has made substantial progress in predicting student dropout using a range of ML and statistical techniques, several critical gaps hinder the advancement of more robust, interpretable, and generalisable models (Putri et al., 2024). First, integrating advanced meta-heuristic optimisation algorithms with traditional classifiers is still limited. Most existing studies rely on standard classifiers – such as DT, logistic regression, and neural networks – with minimal hyperparameter tuning, thereby neglecting the potential performance gains achievable through hybrid frameworks that combine base models with optimisation strategies (Rajest et al., 2023). Second, many existing approaches focus predominantly on binary classification tasks, such as

distinguishing between students who drop out and those who do not (Rizal et al., 2024). This oversimplification overlooks the more nuanced continuum of academic outcomes, particularly the intermediate states, such as students who remain enrolled but have not graduated (Sebastian et al., 2024). Addressing this limitation through a multi-class classification approach allows a more comprehensive understanding of student trajectories (Sharma et al., 2024). Third, while some large-scale datasets have been utilised, a significant portion of the literature is based on relatively small or medium-sized datasets, often with fewer than 2,000 records (Shruthi and Aravind, 2023). This restriction may limit the statistical power and generalisability of the resulting models, particularly in diverse and evolving educational contexts (Tripathi and Al-Zubaidi, 2023). Fourth, many existing studies have noted a notable lack of interpretability and methodological transparency. Although predictive performance is often prioritised, limited efforts are made to explain how input features influence outcomes, making it difficult for educators and policymakers to draw actionable insights (Varmann et al., 2023). Furthermore, many studies provide insufficient detail regarding model construction, parameter settings, and optimisation processes, reducing the reproducibility and applicability of their findings.

To address these limitations, this study proposes a novel hybrid framework that integrates the k-nearest neighbour (KNN) classification algorithm with two recently developed nature-inspired optimisation algorithms (Velmonte, 2023): the northern goshawk optimiser (NGO) and the Tasmanian devil optimisation (TDO). The framework aims to improve the classification accuracy and robustness of student academic performance prediction across three meaningful categories – enrolled, graduated, and dropout – using a real-world dataset comprising 4424 student records. The key contributions of this study are as follows:

- Development of two hybrid classifiers: this research introduces two optimized models, KNN-NGO and KNN-TDO, in which the parameters of the base KNN classifier are fine-tuned using the NGO and TDO algorithms, respectively. These hybrid models demonstrate significant improvements over baseline models in classification performance.
- Comprehensive performance evaluation: the models are rigorously evaluated using a suite of classification metrics – accuracy, precision, recall, and F1-score – thereby offering a well-rounded assessment of their predictive capabilities.
- Interpretability and feature analysis: beyond quantitative assessment, the study examines the relationships between input features and classification outcomes, thereby contributing to a deeper understanding of the factors influencing student academic status.
- Methodological transparency and reproducibility: detailed justifications for selecting algorithms, descriptions of hyperparameter tuning procedures, and clear explanations of evaluation strategies are provided to ensure clarity, replicability, and potential for practical implementation.

2 Materials and methodology

2.1 Data processing

This study aims to develop a robust methodology for accurately evaluating students' academic performance, taking into account contextual factors that influence their performance. The key step involves pre-processing the dataset and converting textual data into numerical values. This foundational process is crucial for effectively employing ML and advanced statistical techniques in data analysis (Kannan et al., 2023). The dataset encompasses diverse variables categorized as follows:

- Student demographics: student's marital status, student's nationality, displaced candidate, student's gender, student's age at enrolment, student's nationality status.
- Parental information: mother's education qualification, father's education qualification, job of student’s mother, job of student’s father.
- Financial and support information: job of student’s mother, job of student’s father, special education requirements, debtor, academic fee situation, scholarship holder.
- Economic indicators: academic unemployment rate, educational inflation rate, GDP associated with learning.
- Enrolment information: mode of application to enroll the course, students' application order, specialised field of study.
- Academic performance: attendance regime values, past educational credentials, curricular units (credited), curricular units (enrolled), curricular units (evaluations), curricular units (approved), curricular units (grade), curricular units (without evaluation).

This study aims to forecast and categorise students into three distinct groups: enrolment, graduation, and dropout. Table 1 indicates the statistical properties of the variables. Data contains 4,424 samples, of which 1,421 belonged to dropout, 794 enrolled, and 2,207 graduate classes. The dataset underwent key pre-processing steps to enhance generalisability, including the imputation of missing values (using the mean or median for numerical values, and the mode for categorical values), label encoding, min-max normalisation, and outlier handling. Stratified data splitting and K-fold cross-validation were applied to ensure a balanced and robust model evaluation. Figure 1 illustrates the interplay between various parameters and their impact on the target, encompassing all three categories, employing colour coding for enhanced visual comprehension. The main diagonal of the square signifies the self-effect of each parameter, prominently denoted in bold red. Notably, an examination of the final row in the square delineates the substantial positive influences on student performance, primarily associated with factors such as tuition fees up to date, scholarship recipient, curricular units (evaluations), curricular units (approved), and curricular Units (grade). On the other hand, characteristics like age at enrolment, gender, and debt have negative impacts that range from –0.2 to –0.4. This visual representation provides a comprehensive insight into the intricate relationships between parameters and their collective impact on students' academic outcomes.

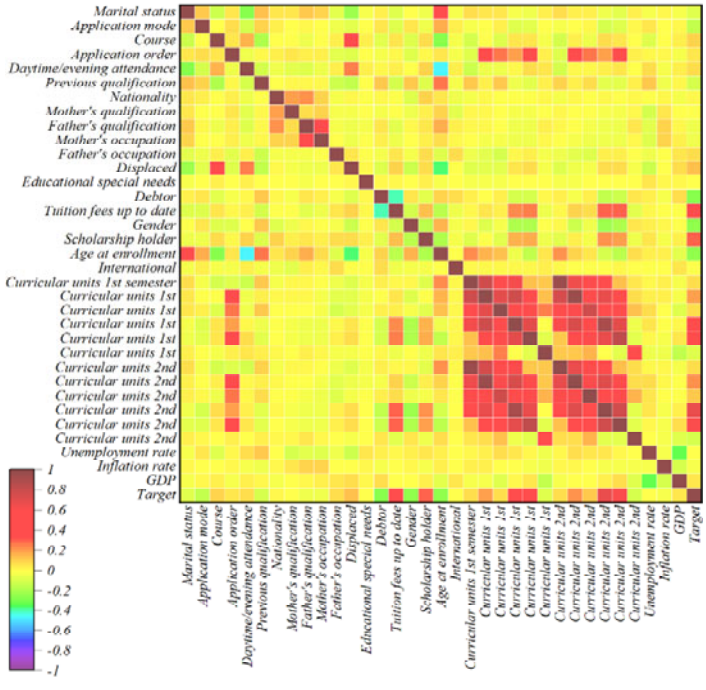
Table 1 Statistical properties of the variables

Variables	*Indicators*				
	Min	*Max*	*Avg.*	*Std*	*Median*
Marital status	1	6	1.179	0.606	1
Application mode	1	18	8.005	6.458	7
Course	1	17	1.735	1.353	1
Application order	33	9991	8857	2063	9238
Daytime/evening attendance	0	1	0.891	0.312	1
Previous qualification	1	17	2.209	3.158	1
Nationality	1	21	8.953	6.618	10
Mother's qualification	1	34	10.355	7.424	10
Father's qualification	1	34	6.429	4.548	5
Mother's occupation	1	46	6.906	5.537	7
Father's occupation	1	46	7.721	3.315	7.5
Displaced	0	1	0.548	0.498	1
Educational special needs	0	1	0.012	0.107	0
Debtor	0	1	0.114	0.317	0
Tuition fees are up to date	0	1	0.881	0.324	1
Gender	0	1	0.352	0.478	0
Scholarship holder	0	1	0.248	0.432	0
Age at enrollment	17	70	23.265	7.587	20
International	0	1	0.025	0.156	0

Table 1 Statistical properties of the variables (continued)

Variables	*Indicators*				
	Min	*Max*	*Avg.*	*Std*	*Median*
Curricular units 1st semester (credited)	0	20	0.710	2.360	0
Curricular units 1st semester (enrolled)	0	26	6.271	2.480	6
Curricular units1st semester (evaluations)	0	45	8.299	4.179	8
Curricular units 1st semester (approved)	0	26	4.707	3.094	5
Curricular units 1st semester (grade)	0	18.875	10.641	4.843	12.286
Curricular units 1st semester (without evaluations)	0	12	0.138	0.691	0
Curricular units 2nd semester (credited)	0	19	0.542	1.918	0
Curricular units 2nd semester (enrolled)	0	23	6.232	2.196	6
Curricular units 2nd semester (evaluations)	0	33	8.063	3.948	8
Curricular units 2nd semester (approved)	0	20	4.436	3.014	5
Curricular units 2nd semester (grade)	0	18.571	10.230	5.210	12.2
Curricular units 2nd semester (without evaluations)	0	12	0.150	0.754	0
Unemployment rate	7.6	16.2	11.566	2.664	11.1
Inflation rate	–0.8	3.7	1.228	1.383	1.4
GDP	–4.06	3.51	0.002	2.270	0.32

Figure 1 For the input and output variables, a correlation matrix (see online version for colours)



2.2 Assessment procedures

In assessing classification issues, accuracy is often used as a metric to evaluate a model's overall performance. This metric is dependent on four key components: false positives (FP), which indicate false positive predictions; false negatives (FN), which indicate false negative predictions; and true positives (TP), which indicate accurate positive forecasts. However, when the data is not balanced, accuracy loses its utility because it favours the dominant class, which restricts the interpretation that can be made of it. To address this shortcoming, three additional assessment metrics are commonly employed: recall, precision, and F1-Score. From an alternative standpoint, these metrics provide a sophisticated understanding of a model's effectiveness, especially when dealing with imbalanced class distributions, as expressed by equations (1) to (4). Taken together, these metrics contribute to a comprehensive and refined evaluation, providing an extensive assessment of a classification model's efficacy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\_score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

2.3 Machine learning algorithms

2.3.1 Gaussian process classifier

Based on Bayesian principles (Ojha et al., 2017), the Gaussian process (GP) classifier operates by establishing a firm foundation through the creation of a Gaussian prior distribution over the estimated function, denoted as $p(x) = w^T x + b.$ This initial distribution serves as the cornerstone for the subsequent probabilistic estimation process. To construct a probabilistic estimator, the classifier incorporates a sigmoid function, as detailed in equation (5).

$$p(y = 1 \mid x) = sigmoid(f(x)) \tag{5}$$

The determination of the distribution of the latent variable y for a test sample follows a 2-step process. Initially, it is calculated by employing the posterior over the latent variables, indicated as $p(f|X, y)$. Subsequently, the second step involves computing a posterior using equation (6), which leverages the outcomes from the initial step. This approach enables the probabilistic estimation of the latent variable y, offering valuable insights and predictions.

$$p(y = 1 \mid X, y, x) \tag{6}$$

Within this framework, the training samples are denoted by X and y, while the test sample is represented by x. The task of performing inference in this context is typically intricate, but there exist approximations that approach an optimal solution with increasing data size. Notably, the kernel versions of this process present a more straightforward approach (Kuss et al., 2005).

2.3.2 Population-based vortex search algorithm

The VSA has effective application characteristics that greatly enhance quick execution. It was designed as a metaheuristic with an emphasis on a single solution (Doğan and Ölmez, 2015). The VS algorithm generates potential solutions by employing a Gaussian distribution, assembling them in a circle around the centre. However, in certain situations, population-dependent algorithms perform better in the early exploration stage of a search region, especially when there are unknown sites that require research. Despite efforts to encourage variety within the search area, this strategy may lead to premature convergence. Such techniques generate fresh coordinates by leveraging information gathered for each position in the previous iteration (Sağ, 2022).

- Initialising: important control parameters are specified during the algorithm's startup phase. These parameters are the mutation probability (η_m), vortex size (*vsize*), population size (*psize*), and termination criteria. The parameter *psize* implies the total count of contender solutions produced in a single repetition, evenly divided into two parts to yield *vsize*, where *vsize* equals *psize*/2. In the initial phase, the creation of candidate solutions (CS) corresponds to the value of *vsize*. Following that, additional CSs, ranging from (*vsize* + 1) to *psize*, are generated in the subsequent stage. When

the predetermined limit of function evaluations (*maxFEs*) is reached, the algorithm breaks. The polynomial alteration process in the next stage relies on the prospect parameter η_m. Furthermore, equations (7) and (8) are applied in a specific order to calculate μ_0 and q_0.

$$\mu_0^i = \frac{upper_i + lower_i}{2} \tag{7}$$

$$q_0^i = \sigma_0^i = \frac{\left(max\left(upper_i\right) - min\left(lower_i\right)\right)}{2} \tag{8}$$

- First phase: during the first iteration of this phase, random processes are used to construct a whole population of size *psize*. Subsequent versions, referred to as ***X****size*, limit the distribution of random generation to half of the population. After this phase is finished, the best solution discovered is used to update the central point (μ). Following the instructions in equation (9) of the original VS method, half of the population is created throughout this updating step using a Gaussian distribution. Prioritising the optimal centre and applying it to only half of the population, the other half is revised using a population-oriented method that takes elements of selection pressure into account. When solutions are found to be outside of the prescribed range, they are recalculated following the rules given in equations (10).

$$s_i^t(x_i^t \mid \mu_t, v) = \left((2\pi)^d \mid v \mid\right)^{-\left(\frac{1}{2}\right)} e^{\left(-1/2\left(x_i^t - \mu_t\right)^T v^{-1}\left(x_i^t - \mu_t\right)\right)} \tag{9}$$

$$s_i\left(lower_i \vee s_i\right) upper_i \rightarrow s_i = rand \times \left(upper_i - lower_i\right) + lower_i \tag{10}$$

- In the context of the original VS algorithm, the initial central point μ_0, though not directly part of the initial population, actively contributes to its formation. Subsequently, the selection of the central point is exclusively determined by candidates within the existing population. Modifying the VS algorithm yields distinct iterations of the PVS algorithm. Whereas PVS_b leaves μ_0 out of the starting population, PVS_a includes it. μ_0 is the major contender solution POP (1) in the population in the first step of PVS_a, and the other *psize*–1 competitor solutions POP (2: *psize*) are created at random. On the other hand, the first population for PVS_b is created by selecting *psize* candidate solutions at random from the set POP (1: *psize*).
- Second phase: in contrast to algorithms that depend on individual solutions, population-based algorithms necessitate the interaction during the search process between potential solutions across several cycles to modify their placements. While the updating procedure in population-based algorithms may vary depending on the specific algorithm employed, the underlying approach revolves around representing the experiences of individual and collective contender solutions in vector format, enabling the transfer of data. Within the context of the PVS algorithm, a selection method grounded in proportionality is initiated. With this method, the ABC algorithm's observer bee phase candidate solution locations are updated, and specific modifications are made to handle reduction issues. Equation (11) may be utilised to calculate the selection probability vector (PB) for every potential response.

$$pb_i = csum_i \,/\, csum_{psize}$$

$$csum_i = \sum_{j=1}^{i} normp_i \ and$$

$$normp_i = p_i \Big/ \sum_{i=1}^{psize} p_i \ and$$

$$p_i = 0.9 \times \left(max\left\{ \vec{f} \right\} - f_i \right) + 0.1 \tag{11}$$

Within the given context, the i^{th} solution's health metric is represented by f, the population's maximum fitness value is marked by $max\left\{ \vec{f} \right\}$, and the adjusted fitness score of the i^{th} solution for reduction is indicated by p_i. By shifting the objective function values from a minimisation to a maximisation viewpoint, this change is achieved. *Normp* stands for the probability that arises from fitting p-values into the 0.5–1 range. These possibilities are limited to the range of 0.5 to 1. They are obtained by normalising the p-values. A random nearby solution is chosen from the solutions labeled as CS_i where i varies from $vsize + 1$ to $psize$, in the latter portion of the population. A randomly chosen dimension's value is changed using equation (12) to generate a new solution known as CS_{new}. After, the resulting measurement value is assessed to signify if it exceeds specific thresholds, as defined in equation (12). This process is guided by the prob vector.

$$CS_{new} = CS_{current} then CS^i_{new} = CS^i_{current} + \left(CS^i_{current} - CS^i_{neighbor} \right) \times (r - 0.5) \times 2 \tag{12}$$

$$CS_{new} = \begin{cases} lower_i, & CS^i_{new} < lower_i \\ CS^i_{new}, & lower_i \le CS^i_{new} \le upper_i \\ upper_i, & CS^i_{new} > upper_i \end{cases} \tag{13}$$

To assess the fitness of the newly created solution, CS_{new} a randomly selected number, r, between 0.5 and 1, is used. If the fitness of CS_{new} exceeds that of $CS_{current}$, the former replaces the latter. Subsequently, the fitness of the existing solution, $CS_{current}$ is compared to the newly calculated fitness. When CS_{new} is not able to surpass $CS_{current}$, a mutant solution is known as CS_{mutant} Is created via the polynomial mutation process, as stated in equation (14).

$$CS_{mutant} = CS_{current} + \delta_q \times (upper - lower)$$

$$\delta_q = \begin{cases} \left[2r + (1-2r) \left(1 - \delta_1\right)^{\eta_m + 1} \right]^{\frac{1}{\eta_m + 1}}, & if \ r \le 0.5 \\ 1 - \left[2(1-r) + 2(r-0.5) \left(1 - \delta_1\right)^{\eta_m + 1} \right]^{\frac{1}{\eta_m + 1}}, & otherwise \end{cases} \tag{14}$$

$$\delta_1 = \frac{CS_{current} - lower}{upper - lower}$$

$$\delta_2 = \frac{upper - CS_{current}}{upper - lower}$$

In such instances, a stochastic value, denoted as *rnd* is individually generated for each dimension, spanning from 0.5 to 1. If *rnd* is smaller than the calculated η_m value is determined as the reciprocal of the dimensionality of the problem under consideration. The process proceeds to the subsequent stages. Previous scholarly investigations have supported the polynomial mutation operator as an optimal technique for overcoming the challenge of avoiding localised optima and ensuring diversified exploration of the search space, a major roadblock encountered in metaheuristics. The polynomial mutation operator induces a disturbance effect by introducing disturbances into the solution through the polynomial probability distribution. After that, $CS_{current}$ and $CS_{mutantare}$ compared using a selection procedure that gives preference to the better solution. After this phase is finished, the best-identified solution is used to rejuvenate the central point (μ). After the current group is finished, the assessment of equation (15) leads to a decrease in the radius size for the next generation. The PVS algorithm continues to run until it reaches the maximum number of purpose assessments. During the first step, solutions inside the lowered radius, which add up to *vsize*, are repeated. In the second stage, responses defining the remaining fraction of the population are established using random data.

$$r_t = \sigma_0 \times \frac{1}{x} \times \Gamma\left(x, a_t\right) \tag{15}$$

where $a_t = \dfrac{(MaxFEs - Fes)}{MaxFEs}$

then if $(a_t \leq 0)\ a_t = 0.1$

2.3.3 Smell agent optimisation

The sense of smell is among the main senses that organisms use to navigate and understand their surroundings. Through this olfactory sense, many species can identify potentially hazardous compounds in their surroundings (Buck, 2004; Sakalli et al., 2020; Axel, 2005). Considering human olfaction in the development of SAO is not an unusual occurrence. Three modes guide the general framework of SAO, each drawing inspiration from the steps involved in the perception of smell. The agent first detects the scent molecules, determines where they are, and then decides whether to ignore them or search for their source actively. Based on the choice made in the previous phase, the agent then pursues the fragrance molecules as they search for the source. Finally, the last mode is designed to prevent agents from getting trapped in local minima by ensuring that the agent does not lose its trail (Axel, 2005; Chapman and Cowling, 1990; Abdechiri et al., 2013).

- Sniffing mode: as smell chemical compounds propagate towards the agent, the process commences by randomly defining the initial location of these molecules. The initialisation of small molecules can be achieved through the application of equation (16):

$$x_i^{(t)} = \begin{bmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,D)} \\ . & . & . \\ x_{(N,1)} & x_{(N,2)} & x_{(N.D)} \end{bmatrix} \tag{16}$$

In this case, N stands for all the scent molecules, and D for all the decision factors. Equation (17) may be used to produce the position vector in of equation (16), which gives the agent the ability to choose its ideal location inside the search space.

$$x_i^{(t)} = lb_i + r_0 \times \left(ub_i - lb_i\right) \tag{17}$$

where r_0 denotes a random number between 0 and 1, and ub and lb stand for the upper and lower limits, respectively, established by the decision variables. To commence, every odor molecule is assigned a starting speed and spreads out from the origin using equation (18).

$$v_i^{(t)} = \begin{bmatrix} v_{(1,1)} & v_{(1,2)} & v_{(1,D)} \\ . & . & . \\ v_{(N,1)} & v_{(N,2)} & v_{(N.D)} \end{bmatrix} \tag{18}$$

Each smell molecule serves as an indicator of a candidate solution, with the position of these candidates derived from the location vector. $x_i^{(t)} \in R^N$ as represented in equation (16), and their velocity, denoted as $v_i^{(t)} \in R^N$, provided in equation (18). The speed of the molecules is then efficient using equation (19).

$$x_i^{t+1} = x_i^{(t)} + v_i^{t+1} \times \Delta t \tag{19}$$

when $\Delta t = 1$, it means that as the optimisation is happening, the agent moves forward on the path at the same time. Using equation (20), the revised position of the scent molecules is determined:

$$x_i^{t+1} = x_i^{(t)} + v_i^{t+1} \tag{20}$$

Because every fragrance molecule has a unique diffusion velocity, its position within the search criteria may be updated and evaporated. Equation (21) is applied to calculate the updated velocity of scent molecules.

$$v_i^{t+1} = v_i^{(t)} + v \tag{21}$$

where v is the inform mutable of the velocity, determined by utilising equation (22).

$$v = r_1 \times \sqrt{\frac{3KT}{m}} \tag{22}$$

The smell constant, i, normalises the effects of temperature and mass on the kinetic energy of scent molecules. The scent molecules' mass and temperature are represented by the letters m and T, respectively. The modified positions of the scent molecule in equation (20) are tested for fitness. When this happens, the sniffing mode is finished, and the agent. $x_{agent}^{t}{}'s$ location may be found.

- Trailing mode: during the additional mode, the agent simulates seeking performances to locate the source of a smell. It explores new locations with a heightened emphasis on smell molecules. Utilising equation (23), the agent takes advantage of the chance to go in the direction of these new locations:

$$x_i^{t+1} = x_i^{(t)} + r_2 \times olf \times \left(x_{agent}^t - x_i^{(t)}\right) - r_3 \times olf \times \left(x_{worst}^t - x_i^{(t)}\right) \tag{23}$$

In this case, r_2 and r_3 are integers between 0 and 1. *Olf* influence on x_{agent}^t is penalised by r_2 while its influence on x_{agent}^t is penalised by r_3. The program logs the values of x_{agent}^t and x_{worst}^t, which are acquired via the sniffing mode. As seen in equation (24), the algorithm can more effectively balance exploration and exploitation by utilising this data.

- Random mode: the intensity of a scent may alter over time in situations when smell molecules are spread across great distances. This variance may confuse the agent, potentially leading to a loss of scent and making trailing difficult. Given the agent's incapacity to maintain a consistent trail, It might end up stuck in nearby minima. In these conditions, the agent transitions to the random mode, as demonstrated by equation (25):

$$x_i^{t+1} = x_i^{(t)} + r_4 \times SL \tag{24}$$

The step length, or SL, is a random number that stochastically penalises the quantity of SL. The step length, denoted by SL, is a random number that stochastically penalizes the SL magnitude in this case. SL denotes the step length. r_4 is used to punish the amount of SL stochastically. SL denotes the step length in this case and the random value r_4 stochastically penalises the SL magnitude.

Alphabet 1 comprises the SAO algorithm's pseudo-code.

Algorithm 1 SAO

Initialise parameters
Initialise smell molecule's initial position
Assess fitness
Prepare the location of the agent and the worst position of molecules
While (Itr < Itr_{max}) **do**:
for (i =1 to molecules) **do**:
for (j=1 to position) **do**:
update molecules' velocity and position (sniffing)
end for
Assess fitness
if (new fitness is better) **then**:
Update fitness
Update agent and worst molecules
end if
end for
for (i = 1 to molecules) **do**:
for (j = 1 to position) **do**:
update position (trailing)
end for

Assess fitness
end for
if (new fitness is better) **then:**
grant new fitness
update position
else
for (i = 1 to molecules) **do**:
for (j = 1 to position) **do**:
Implement random mode
end for
end for
end if
end while
return optimum solution

2.4 K-fold cross-validation

Table 2 presents the accuracy values obtained from five iterations of K-fold cross-validation, denoted as K1 through K5. The accuracy scores for each fold are as follows: 0.701 (K1), 0.740 (K2), 0.725 (K3), 0.739 (K4), and 0.763 (K5). These results highlight the stability and robustness of the proposed model across varying subsets of the dataset. While slight fluctuations are observed, the overall variation is minimal, suggesting that the model generalises well and is not overly dependent on any particular training-test split. The lowest accuracy of 0.701 in fold K1 may be attributed to a higher concentration of complex or imbalanced instances within that subset.

In contrast, the highest score of 0.763 in K5 demonstrates the model's potential under more favourable data conditions. The average accuracy across all folds is approximately 0.734, confirming that the model performs consistently across different data segments. This consistent performance across folds supports the reliability of the classification model and indicates that it is well-suited for practical application in real-world educational settings.

Table 2 Result of K-fold cross-validation

Index values	*Models*				
	K1	*K2*	*K3*	*K4*	*K5*
Accuracy	0.701	0.740	0.725	0.739	0.763

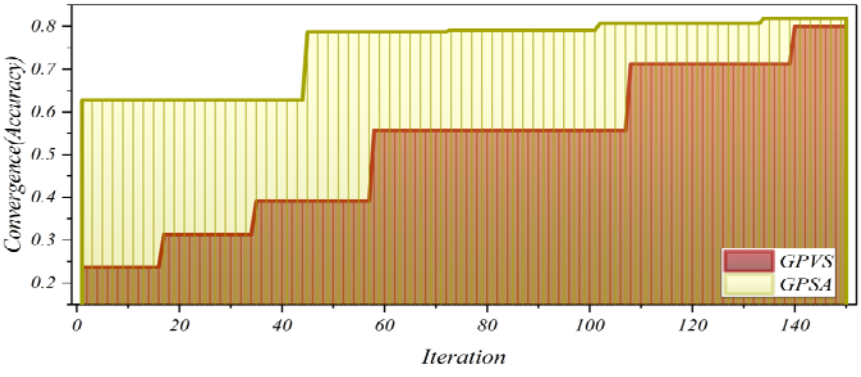
3 Results and discussion

3.1 Convergence and hyperparameter results

This study aims to fine-tune the hyperparameters of the Gaussian process classifier (GPC), tailoring its functionality to suit distinct datasets and problem domains. This optimisation project utilises smell agent optimisation (SAO) and the population-based

vortex search algorithm (PVSA), a significant step toward enhancing the predictive power of this basic ML algorithm. Optimisation performance evaluation entails thoroughly examining how the chosen algorithms influence the accuracy of the GPC over multiple iterations. Figure 2 depicts two convergence curves, GPVS and GPSA, wherein the GPSA model exhibits a noteworthy advantage over the GPVS model in the preliminary iterations, positioning itself at a higher level. As the iterations progress, the sustained superiority of the GPSA model persists.

Figure 2 Line plot for convergence of hybrid models (see online version for colours)



Conversely, the GPVS model demonstrates commendable advancement, taking substantial steps forward, albeit without surpassing the GPSA model. Despite the significant progress achieved by the GPVS model throughout the iterative process, the outcome underscores the enduring dominance of the GPSA model. The GPVS model, while showcasing resilience and substantial strides, falls short of dethroning the GPSA model, which maintains its position as the top-performing model. This nuanced dynamic highlights the intricate interplay of factors influencing both models' performance trajectories, emphasising the GPSA model's sustained excellence as the predominant model throughout the iterative continuum. Furthermore, Table 2 highlights the fine-tuning of hyperparameters for two models – GPVS and GPSA. GPVS employs more aggressive optimisation settings (n_restarts_optimiser = 5, max_iter_predict = 200, n_jobs = 22), indicating a more exhaustive search and a greater utilisation of computational resources for potentially higher accuracy. In contrast, GPSA uses lighter settings (n_restarts_optimiser = 1, max_iter_predict = 131, n_jobs = 3), favouring efficiency and faster computation. This contrast highlights a trade-off between performance robustness and resource efficiency, providing flexibility for various deployment environments.

Table 3 Result of fine-tuning values

Hyperparameter	*GPVS*	*GPSA*
n_restarts_optimiser	5	1
max_iter_predict	200	131
n_jobs	22	3

3.2 Evaluating outcomes from forecasting models

In predicting academic outcomes using ML, this study integrates diverse student data, focusing on their academic progression, including enrolment, graduation, and dropout rates. The dataset is pivotal for training and assessing 3 GPC models: GPC, GPVS, and GPSA. For every stage of the prediction process, the research methodically computes important performance measures such as F1-score, accuracy, precision, and recall. This detailed analysis aims to identify the most effective predictive model, providing insights to enhance students' academic success. All relevant metric values, covering training, testing, and overall model performance, are outlined in Table 3 and visualised in Figure 3. During the practice phase, a substantial portion of the data (approximately 70%) is utilised, leaving a smaller portion (around 30%) for testing. Consequently, the results from the test phase are expected to be diminished. However, during the evaluation, consideration should encompass the phase datasets. In comparing models during this comprehensive assessment, the GPSA model exhibited superior performance, with an average of 0.816, surpassing the GPVS models, which had an average of 0.794, and the GPC model, with an average of 0.75. This performance distinction is further illustrated in Figure 3, where the GPSA model outperforms others in the Pie chart, depicting the attainment of developed models based on evaluators.

Figure 3 Pie chart for the achievement of developed models based on evaluators (a) Accuracy (b) F1_score (c) Recall (d) Precision (see online version for colours)

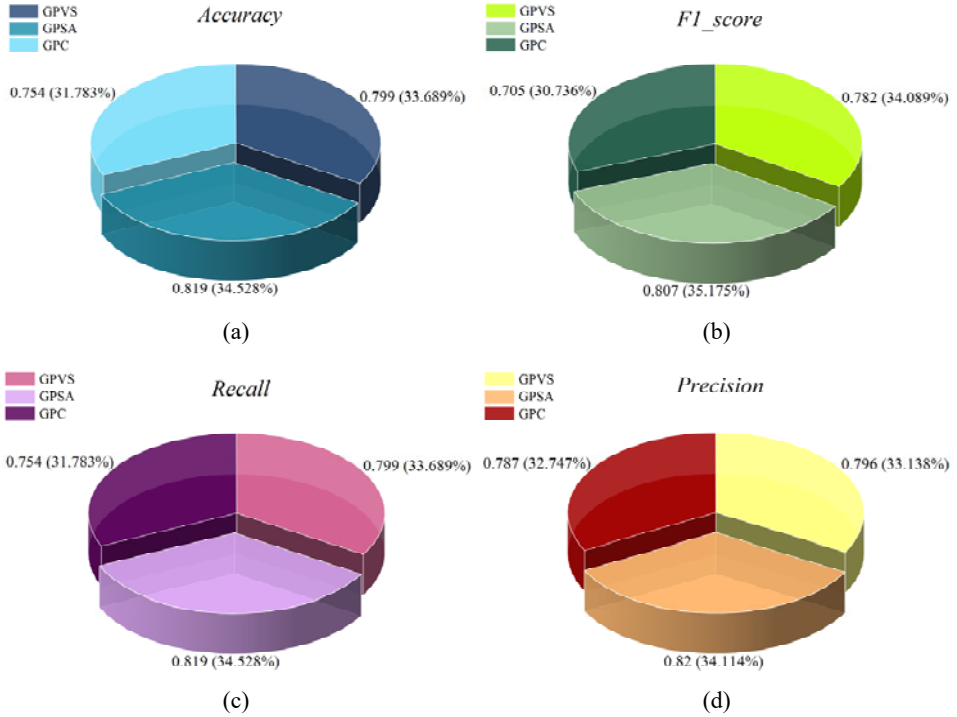


Table 4 Result of developed models

Phase	*Index values*	*Models*		
		GPC	*GPSA*	*GPVS*
Train	Accuracy	0.769	0.852	0.822
	Precision	0.815	0.856	0.825
	Recall	0.769	0.852	0.822
	F1-score	0.723	0.843	0.808
Test	Accuracy	0.719	0.742	0.745
	Precision	0.707	0.728	0.726
	Recall	0.719	0.742	0.745
	F1-score	0.663	0.720	0.720
All	Accuracy	0.754	0.819	0.799
	Precision	0.787	0.820	0.796
	Recall	0.754	0.819	0.799
	F1-score	0.705	0.807	0.782

3.3 Classification outcomes

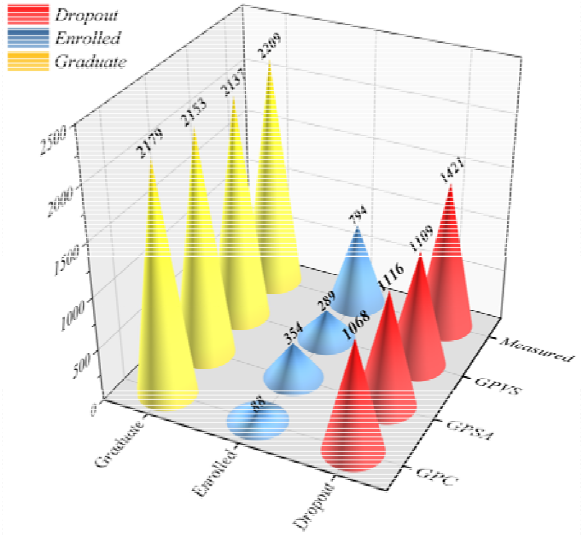
Table 4 provides a comprehensive overview of the recall, precision, and F1_score metrics related to student assignments based on academic status. This table's metrics provide useful information about how well the model predicts positive outcomes, given its ability to accurately identify real positives, as well as its effectiveness in classifying students based on their academic performance, including enrolment, dropout rates, and graduation rates.

- Dropout: in this subset, comprising approximately 32.12% of the dataset and involving 1,421 students who discontinued their education, the GPSA model demonstrated superior predictive performance. This finding highlights the model's effectiveness in accurately predicting instances of student dropout, underscoring its potential as a valuable tool for identifying and addressing factors contributing to educational discontinuation.
- Enrolled: with greater recall, precision, and F1-Score values, the GPSA model outperformed the other models in this cohort of 794 students in categorising academic achievement. The GPC model performed noticeably worse than the other models in the investigation, outperforming them in precision but falling short in Recall and F1-Score measures.
- Graduate: the GPSA model displayed exceptional applicability, consistently outperforming its counterparts with the highest average values across all metrics. Notably, it achieved an impressive average of 0.88, surpassing GPC (0.83) and GPVS (0.87). This highlights the GPSA model's superior performance and versatility in effectively addressing the task at hand (Table 5).

Table 5 Performance metrics for classification models based on three categories

Model	*Category*	*Index values*		
		Precision	*Recall*	*F1-score*
GPC	Dropout	0.88	0.75	0.81
	Enrolled	0.85	0.11	0.2
	Graduate	0.7	0.99	0.82
GPSA	Dropout	0.9	0.79	0.84
	Enrolled	0.74	0.45	0.56
	Graduate	0.79	0.97	0.88
GPVS	Dropout	0.88	0.78	0.83
	Enrolled	0.7	0.36	0.48
	Graduate	0.78	0.97	0.86

Figure 4 3D bars plot for the comparison between the measured and predicted values (see online version for colours)



In Figure 4, a classification of the number of students in the enrolled graduated and dropout categories is presented through the utilisation of 3 models: a singular model and two optimised models. The visual representation utilises cones to represent the number of students in each category, facilitating the visual identification of the model with the least deviation from the measured values. Through a meticulous comparison of the values generated by the models and the actual measured values, a numerical discrepancy is discerned. Specifically, the GPSA model demonstrates a variance of 801, the GPC model exhibits a difference of 1089, and the GPVS model shows a deviation of 889. Consequently, the GPSA model with the smallest disparity from the measured values is identified as the superior model. This analytical approach provides a quantitative basis for evaluating model performance, emphasising the practical significance of selecting a model that closely aligns with the actual observed values. This, in turn, contributes to the

accuracy and reliability of the predictive model in classifying students into specified categories.

Figure 5 Confusion matrix for models' accuracy (see online version for colours)

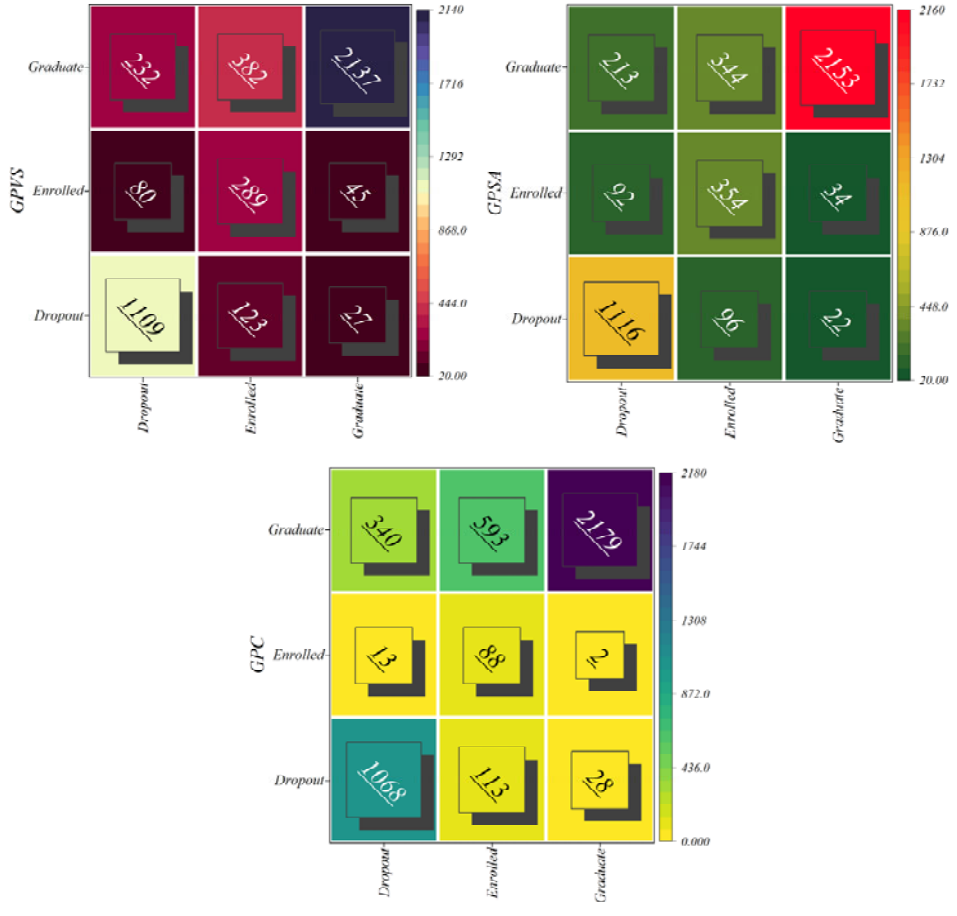
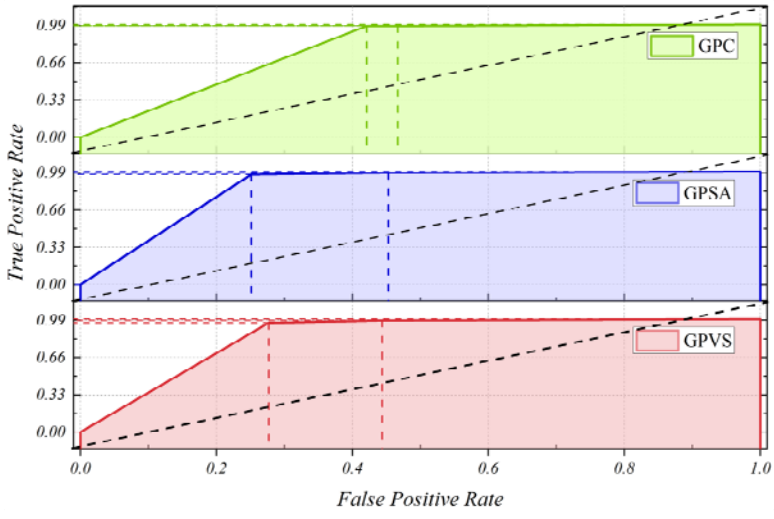


Figure 5 presents a comprehensive view through a confusion matrix detailing the precise categorisation of students and instances of misclassification within three models: GPC, GPSA, and GPVS. In the GPC model, 3,335 students were accurately classified, while 1089 were misclassified. The GPSA model exhibited accurate classification for 3,623 students, with 801 misclassifications, and the GPVS model accurately classified 3,535 students, with 889 misclassifications. The matrix presents the nuanced performance of each model, highlighting its advantages and disadvantages in accurately classifying students into the appropriate groups. This detailed breakdown aids in a more nuanced understanding of the predictive capabilities of each model, informing decisions on model selection based on the trade-off between correct classifications and misclassifications.

Figure 6 presents the receiver operating characteristic (ROC) curve, a crucial tool for evaluating model performance across varying thresholds of true positive and false positive rates. One statistic that is often used to assess the overall prediction performance of the GPC model and its two improved variants is the Area Under the ROC Curve

(AUC). The analysis of model performance highlights a notable weakness in the GPC model, evidenced by its significantly smaller AUC compared to the two optimised models. A closer examination of the AUC values for the optimised models reveals nuanced differences in performance. The GPSA model outperforms the GPVS model, albeit with a slight distinction, substantiated by its portrayal of the largest Area under the curve. The effectiveness of the GPSA model in differentiating between real positive and false positive rates is further demonstrated by this. The meticulous consideration of these metrics contributes to a thorough comprehension of model effectiveness, aiding in informed decisions on model selection for predictive analytics.

Figure 6 ROC curve of the best hybrid models (see online version for colours)



3.4 Limitation of study

One notable limitation of this study is its reliance on the quality and diversity of the educational datasets used for analysis. While incorporating various datasets provides a broad perspective, the findings may be influenced by the specific characteristics and biases inherent in these datasets. Additionally, the study's predictive models and optimisation algorithms, though advanced, may require extensive computational resources and expertise to implement effectively. This could limit the practicality and scalability of the proposed solutions in educational institutions with limited technical infrastructure or expertise. Furthermore, while the study aims to generalise its findings across different educational contexts, the variability in educational systems, curricula, and student populations may affect the applicability and generalisability of the results. Future research should explore the integration of diverse datasets and the development of more accessible and scalable modelling techniques to address these challenges.

3.5 Future suggestion

- Expansion of datasets: future research should incorporate a broader range of educational datasets from diverse regions and educational systems to enhance the

generalisability of the findings. This would help validate the model's effectiveness across diverse academic environments.

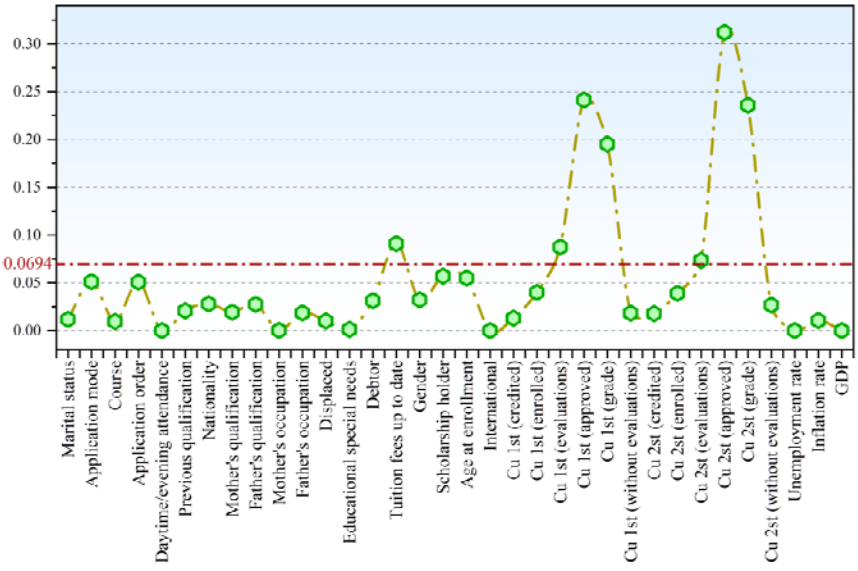
- Real-time implementation: developing real-time predictive systems based on the GPC model optimised with SAO could provide educators with timely insights, allowing immediate interventions to support students at risk of poor performance.
- Interdisciplinary approaches: combining insights from educational psychology, cognitive science, and data science could lead to more nuanced models that account for a broader range of factors influencing academic performance, including psychological and socio-economic variables.
- Scalability and accessibility: simplifying the implementation of the proposed models and optimisation algorithms can make them more accessible to educational institutions with limited technical resources. Future work could focus on creating user-friendly software tools and platforms that facilitate the adoption of these advanced techniques.
- Longitudinal studies: conducting longitudinal studies to track the impact of predictive modelling and optimisation algorithms over time would provide deeper insights into their long-term benefits and potential areas for improvement.
- Integration with educational policies: exploring how predictive models can be integrated into educational policies and practices to support strategic decision-making at the institutional and policy levels would further enhance their impact on improving educational outcomes.
- Ethical considerations: future studies should also address ethical considerations related to data privacy, bias, and fairness in predictive modelling to ensure that the implementation of these technologies promotes equity and inclusiveness in education.

By addressing these suggestions, future research can build on the current study's foundation, leading to more robust, practical, and impactful applications of ML in educational settings.

3.6 Features analysis

Figure 7 illustrates the line-symbol plot that showcases the influence of each variable on student academic performance based on SHAP sensitivity analyses. Mutual_info_classif was employed to achieve this objective. Mutual_info_classif is a feature selection method that measures the mutual information between a feature and a target variable. Mutual information measures how much information one variable provides about another. In feature selection, a high mutual information score indicates that a feature is informative about the target variable and is, therefore, likely helpful for classification. Mutual_info_classif is a non-parametric method, making no assumptions about the data distribution. This makes it a versatile method that can be used with various datasets. In addition to being non-parametric, mutual_info_classif is also very efficient. It can be calculated quickly, even for large datasets. As is evident from the figure, the threshold value for identifying the most impactful inputs is 0.0694. Only seven inputs surpass this threshold, while the remaining variables hover closely around it.

Figure 7 The line-symbol plot, used to explore the impact of input variables on student performance output, is based on SHAP sensitivity analyses (see online version for colours)



3.7 Practical deployment considerations

The successful real-world application of student performance forecasting models hinges on predictive accuracy and several key practical factors, including computational requirements, feasibility for real-time implementation, and ethical implications.

3.7.1 Computational requirements

The proposed hybrid models – KNN optimised with Northern Goshawk optimiser (KNN-NGO) and Tasmanian devil optimisation (KNN-TDO) – are computationally efficient and scalable. The KNN algorithm, being non-parametric, benefits significantly from optimisation techniques that fine-tune hyperparameters to reduce unnecessary computation during prediction. Both NGO and TDO are designed to converge quickly, minimising processing overhead during model training. On a standard machine with moderate processing capabilities (e.g., a quad-core CPU and 8GB of RAM), the training and evaluation processes are completed reasonably quickly, making these models suitable for integration into institutional data systems.

3.7.2 Real-time prediction feasibility

Once trained, the KNN-based models are capable of delivering near real-time predictions. Although KNN traditionally requires storing the full dataset for comparison during inference, optimised indexing structures and dimensionality reduction techniques can be applied to ensure fast query responses. This is especially critical in institutional dashboards where administrators or counsellors may need to assess a student's academic

risk profile immediately. The models can be embedded into learning management systems (LMS) or student information systems (SIS) for seamless access and interaction.

3.7.3 Ethical implications

The deployment of predictive models in educational settings raises significant ethical concerns, particularly related to data privacy, bias, and the potential for misclassification. Institutions must ensure that data collection adheres to privacy regulations, such as the GDPR or FERPA, and obtain informed consent for the use of student data. Moreover, caution should be exercised to avoid reinforcing systemic biases, such as overpredicting dropout risk for students from underrepresented backgrounds, by ensuring balanced training data and transparency in feature selection. It is also critical that these models are used as support tools rather than deterministic labels; predictions should inform interventions, not replace human judgment. Ethical review boards and institutional oversight should guide the deployment to safeguard against misuse and uphold fairness and transparency.

3.7.4 Integration of the optimisation

The implementation of optimisation algorithms plays a pivotal role in advancing the effectiveness of predictive models in education. In this study, two nature-inspired metaheuristic algorithms, the NGO and the TDO, are integrated with the KNN classifier to enhance the classification of students based on academic status. These hybrid models, KNN-NGO and KNN-TDO, are designed to optimise hyperparameters more efficiently than manual tuning or conventional grid-based methods, resulting in improved predictive accuracy and generalisation capability. These findings are supported by consistent performance gains across multiple evaluation metrics, including accuracy, precision, recall, and F1-score. As demonstrated in Tables 2 and 3 and visualised in Figures 3 and 4, the optimised models offer superior classification across the enrolled, graduated, and dropout categories. This performance consistency reinforces the suitability of using metaheuristic optimisation to strengthen the predictive capabilities of ML models in educational settings. Moreover, integrating optimisation strategies aligns with modern educational analytics workflows, where precision, scalability, and interpretability are increasingly prioritised. The structured modelling sequence, data pre-processing, algorithm selection, optimisation, validation, and interpretability analysis ensure the system is compatible with current data-driven academic decision-making practices. This approach enhances the technical robustness of the predictive models. It supports the feasibility of real-time deployment, making it suitable for integration into academic early warning systems and learning management platforms. From a practical standpoint, the hybrid KNN-NGO and KNN-TDO models empower institutions to make timely and informed interventions. Educators can offer targeted support and reduce dropout rates by accurately identifying at-risk students and understanding the factors contributing to their academic trajectories.

Additionally, the optimisation-enhanced models operate with computational efficiency, making them viable for deployment in real-time dashboards without incurring significant resource overhead. In conclusion, incorporating advanced optimisation algorithms into the ML framework enhances model performance and supports a comprehensive and modern approach to academic analytics. These strategies provide

educators with reliable, interpretable, and scalable tools to enhance student outcomes and institutional performance.

4 Conclusions

In conclusion, this study stands at the forefront of advancing academic performance prediction through the innovative integration of a powerful GPC model. To enhance its predictive power, the model utilises two state-of-the-art optimisation methods: SAO and PVSA. The primary goal is to achieve accurate predictions of students' comprehensive performance, which will significantly contribute to the broader mission of enhancing educational outcomes, particularly in higher education, where predicting student performance plays a pivotal role in strategic decision-making and reducing the dropout rate. The research analyses diverse educational datasets using ML techniques, with a focus on dimensionality reduction. Taking a proactive approach empowers educators with data-informed decision-making capabilities, providing timely guidance for academic improvement. In addition, the project aims to improve education in general by classifying people according to their innate abilities, with a focus on reducing dropout rates. Predictive modelling, especially when applied to ML, empowers the academic community to take proactive measures in addressing problems, which in turn creates a more encouraging learning environment and eventually improves student results. The outcomes revealed an error rate of 18.1% for the GPSA model, 20% for the GPVS model, and 24.6% for the GPC model. Consequently, the GPSA model demonstrated superior performance when contrasted with the other models, exhibiting the lowest error rate. This underscores its enhanced ability to distinguish performance levels, making it a crucial tool for accurate academic performance prediction. The findings suggest that incorporating these optimisation algorithms has the potential to enhance model performance in educational settings, leading to more accurate and reliable student classifications.

Declarations

- Competing interests: the authors declare no competing interests.
- Availability of data and materials: data can be shared upon request.
- Funding: this work was supported by 2025 General Project of Henan Provincial Education Science Planning in research on the Multi-Collaborative Training Path of ‘artificial intelligence + e-commerce’ Professional Talent with No.2025YB0169.
- Authors' contributions: the author contributed to the study's conception and design. Data collection, simulation and analysis were performed by ‘Yao Zhang’.

Acknowledgements

I would like to take this opportunity to acknowledge that there are no individuals or organisations that require acknowledgment for their contributions to this work.

- Ethical approval: the research paper has received ethical approval from the institutional review board, ensuring the protection of participants' rights and compliance with the relevant ethical guidelines.

References

Abdechiri, M., Meybodi, M.R. and Bahrami, H. (2013) 'Gases Brownian motion optimization: an algorithm for optimization (GBMO)', *Applied Soft Computing*, Vol. 13, No. 5, pp.2932–2946.

Adekola, R.A. and Aribisala, O.O. (2023) 'Impact of entrepreneurial skill on attaining sustainable development in business education', *FMDB Transactions on Sustainable Social Sciences Letters*, Vol. 1, No. 3, pp.180–188.

Alkasi, S., Surya, S., Tohir, S., Alpiah, S. and Oktapiana, S. (2024) 'Community service journal: promoting ethics and responsibility in social media use for high school students', *AVE Trends in Intelligent Social Letters*, Vol. 1, No. 4, pp.177–186.

Ameen, A.O., Alarape, M.A. and Adewole, K.S. (2019) 'Students' academic performance and dropout predictions: a review', *Malaysian Journal of Computing*, Vol. 4, No. 2, pp.278–303.

Ashifa, K.M. and Büyük, İ. (2024) 'Measuring the level of knowledge about addiction to substances among university students', *FMDB Transactions on Sustainable Techno Learning*, Vol. 2, No. 1, pp.1–10.

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016) *Predicting Student Dropout in Higher Education*, arXiv Available: https://arxiv.org/abs/1606.06364 (accessed 28 April 2017).

Axel, R. (2005) *Scents And Sensibility: A Molecular Logic of Olfactory Perception*, Nobel Lecture, Angewandte Chemie International Edition, Vol. 44, No. 38, pp.6110–6127.

Basañes, R., Odango Alentajan, J., Basañes-Erfe, G. and Erfe, J.E. (2023) 'Creating a student-centered learning climate for public elementary and secondary schools', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 4, pp.211–219.

Basañes, R.A. and Alentajan, J.O. (2024) 'Evaluating the programs and services offered by the student affairs and services division in a state university', *AVE Trends in Intelligent Social Letters*, Vol. 1, No. 2, pp.82–92.

Bastareche, R.D. (2024) 'Global educational collaboration: Its platform, premises, and promises', *FMDB Transactions on Sustainable Techno Learning*, Vol. 2, No. 2, pp.88–101.

Boden, M.A. (1996) *Artificial Intelligence*, Elsevier, Amsterdam, Netherland.

Buck, L.B. (2004) 'Unraveling the sense of smell', *Les Prix Nobel. Nobel Prize*, Vol. 44, No. 38, pp.6128–6140.

Burgos, C., Campanario, M.L., de la Peña, D., Lara, J.A., Lizcano, D. and Martínez, M.A. (2018) 'Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout', *Computers and Electrical Engineering*, Vol. 66, No. 2, pp.541–556.

Cerquitelli, T., Pagliari, D.J., Calimera, A., Bottaccioli, L., Patti, E., Acquaviva, A. and Poncino, M. (2021) 'Manufacturing as a data-driven practice: methodologies, technologies, and tools', *Proceedings of the IEEE*, Vol. 109, No. 4, pp.399–422.

Chapman, S. and Cowling, T.G. (1990) *The Mathematical Theory of Non-Uniform Gases: An Account of The Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases*, Cambridge University Press, Cambridge, England United Kingdom.

Dalton, B., Glennie, E. and Ingels, S.J. (2009) *Late High School Dropouts: Characteristics, Experiences, and Changes Across Cohorts*, Descriptive Analysis Report (NCES 2009–307), National Center for Education Statistics, Washington, United States of America.

Dekker, G.W., Pechenizkiy, M. and Vleeshouwers, J.M. (2009) 'Predicting students drop out: a case study', in *Proceedings of the 2nd International Conference on Educational Data Mining* (*EDM 2009*), Córdoba, Spain, pp.41–50.

Del Río, C. and Insuasti, J.P. (2016) *Predicting Academic Performance in Traditional Environments at Higher-Education Institutions Using Data Mining: A Review*, La Academia Técnica del Norte, Ecos, Vol. 2, No. 4, pp.185–201.

Doğan, B. and Ölmez, T. (2015) 'A new metaheuristic for numerical function optimization: vortex search algorithm', *Information Sciences*, Vol. 293, No. 2, pp.125–145.

El Naqa, I. and Murphy, M.J. (2015) *What is Machine Learning*? Springer, Cham, Switzerland.

Fauzy, I., Padilah, D., Alzena, F.C., Azzahra, F. and Kamodin, F. (2024) 'Optimizing learning by enhancing students' motivation', *FMDB Transactions on Sustainable Humanities and Society*, Vol. 1, No. 2, pp.91–99.

Golding, P. and Donaldson, O. (2006) 'Predicting academic performance', in *Proceedings. Frontiers in Education. 36th Annual Conference*, San Diego, CA, United States of America, pp.21–26.

Gomathy, V. and Venkatasbramanian, S. (2023) 'Impact of teacher expectations on student academic achievement', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 2, pp.78–91.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C. and Nam Liao, S. (2018) 'Predicting academic performance: A systematic literature review', in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, Larnaca, Cyprus, pp.175–199.

Hernandez, J.J. and Nunez, N. (2023) 'Technopreneurs' intention of IT students: an application of Ajzen's theory of planned behavior', *FMDB Transactions on Sustainable Social Sciences Letters*, Vol. 1, No. 4, pp.198–218.

Joshi, M., Shen, Z. and Kausar, S. (2023) 'Enhancing inclusive education on leveraging artificial intelligence technologies for personalized support and accessibility in special education for students rth diverse learning needs', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 3, pp.25–142.

Kannan, R., Abarna, K.T.M. and Vairachilai, S. (2023) *Student Academic Performance Prognosticative Using Optimized Hybrid Machine Learning Algorithms*, Research Square Platform LLC, Durham, United States of America.

Kiss, B., Nagy, M., Molontay, R. and Csabay, B. (2019) 'Predicting dropout using high school and first-semester academic achievement measures', in *2019 17th International Conference on Emerging eLearning Technologies and Applications* (*ICETA*), Starý Smokovec, Slovakia, pp.383–389.

Kovacic, Z. (2010) 'Early prediction of student success: mining students' enrolment data', *Proceedings of Informing Science and IT Education Conference* (*InSITE*) 2010, Cassino, Italy.

Krishnan, A., Mol, K.P.S., Saiju, L., Babu, J. and Ashifa, K.M. (2024) 'Tribal education in Kerala and achieving sustainable development goal (SDG)-4: prospects and challenges', *FMDB Transactions on Sustainable Techno Learning*, Vol. 2, No. 1, pp.41–49.

Kumar, M., Singh, A.J. and Handa, D. (2017) 'Literature survey on educational dropout prediction', *International Journal of Education and Management Engineering*, Vol. 7, No. 2, p.8.

Kuss, M., Rasmussen, C.E. and Herbrich, R. (2005) 'Assessing approximate inference for binary Gaussian process classification', *Journal of Machine Learning Research*, Vol. 6, No. 5, p.10.

Legista, M., Nurlaili, L. and Masriah, I. (2024) 'Development of an adaptive learning system utilizing deep learning to support independent learning processes for students', *AVE Trends in Intelligent Techno Learning*, Vol. 1, No 2, pp.107–120.

Lin, J.J., Imbrie, P.K. and Reid, K.J. (2009) 'Student retention modelling: an evaluation of different methods and their impact on prediction results', *Proceedings of Research in Engineering Education Symposium*, Palm Cove, Australia.

Mallinckrodt, B. and Sedlacek, W.E. (1987) 'Student retention and the use of campus facilities by race', *NASPA Journal*, Vol. 24, No. 3, pp.28–32.

Masangu, L., Jadhav, A. and Ajoodha, R. (2021) 'Predicting student academic performance using data mining techniques', *Advances in Science, Technology and Engineering Systems Journal*, Vol. 6, No. 1, pp.153–163.

Mompel, J.T. and Lombrio, C. (2024) 'Impact of physical education as a form of recreational activities: academic performance of the students', *FMDB Transactions on Sustainable Humanities and Society*, Vol. 1, No. 3, pp.113–123.

Moyo, S.B. and Nithyanantham, V. (2024) 'Evaluating the influence of parental involvement in early childhood education programs', *AVE Trends in Intelligent Techno Learning*, Vol. 1, No. 1, pp.47–60.

Nithyanantham, V. (2023) 'Study examines the connection between students' various intelligence and their levels of mathematical success in school', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 1, pp.32–59.

Ojha, T., Heileman, G.L., Martinez-Ramon, M. and Slim, A. (2017) 'Prediction of graduation delay based on student performance', in *2017 International Joint Conference on Neural Networks* (*IJCNN*), Alaska, United States of America, pp.3454–3460.

Oqaidi, K., Aouhassi, S. and Mansouri, K. (2022) 'Towards a students' dropout prediction model in higher education institutions using machine learning algorithms', *International Journal of Emerging Technologies in Learning*, Vol. 17, No. 18, p.103.

Padmanabhan, J., Rajest, S.S. and Veronica, J.J. (2023) 'A study on the orthography and grammatical errors of tertiary-level students', in *Handbook of Research on Learning in Language Classrooms Through ICT-Based Digital Technology*, pp.41–53, IGI Global, USA.

Padmavathy, V., Venkateswaran, P.S., Begum, S. and Sheela, K. (2024) 'A review of the faculty engagement and performance in higher educational institutions', *AVE Trends in Intelligent Techno Learning*, Vol. 1, No. 2, pp.76–87.

Patacsil, F.F. (2020) 'Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models', *Universal Journal of Educational Research*, Vol. 8, No. 9, pp.4036–4047.

Pradeep, A., Das, S. and Kizhekkethottam, J.J. (2015) 'Students dropout factor prediction using EDM techniques', in *2015 International Conference on Soft-Computing and Networks Security* (*ICSNS*), Coimbatore, India, pp.1–7.

Pulivarthy, P. (2024) 'Research on oracle database performance optimization in IT-based university educational management system', *FMDB Transactions on Sustainable Computing Systems*, Vol. 2, No. 2, pp.84–95.

Putri, M.F.J.L., Sunarso, S. and Samsuri, S. (2024) 'Citizenship education as an effort to prevent radicalism and extremism', *FMDB Transactions on Sustainable Humanities and Society*, Vol. 1, No. 2, pp.82–90.

Quiroz, P. (2000) 'A comparison of the organizational and cultural contexts of extracurricular participation and sponsorship in two high schools', *Educational Studies*, Vol. 31, No. 3, pp.249–275.

Rajest, S.S., Moccia, S., Chinnusamy, K., Singh, B. and Regin, R. (Eds.) (2023) 'Handbook of research on learning in language classrooms through ICT-based digital technology', *Advances in Educational Technologies and Instructional Design*, DOI:10.4018/978-1-6684-6682-7, USA.

Rizal, V., Riani, V., Muhtadi, T., Lestari, Y. and Sulistyorini, Y.F. (2024) 'Enhancing critical thinking skills among Islamic high school student leaders through interactive training programs: a case study', *FMDB Transactions on Sustainable Management Letters*, Vol. 2, No. 3, pp.121–136.

Romero, C. and Ventura, S. (2007) 'Educational data mining: a survey from 1995 to 2005', *Expert Systems with Applications*, Vol. 33, No. 1, pp.135–146.

Roy, S., and Garg, A. (2017) 'Predicting academic performance of student using classification techniques', in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics* (*UPCON*), Mathura, India, pp.568–572.

Sağ, T. (2022) 'PVS: a new population-based vortex search algorithm with boosted exploration capability using polynomial mutation', *Neural Computing and Applications*, Vol. 34, No. 20, pp.18211–18287.

Sakalli, E., Temirbekov, D., Bayri, E., Alis, E.E., Erdurak, S.C. and Bayraktaroglu, M. (2020) 'Ear nose throat-related symptoms with a focus on loss of smell and/or taste in COVID-19 patients', *American Journal of Otolaryngology*, Vol. 41, No. 6, p.102622.

Sebastian, J.E., Babu, J. and Ashifa, K.M. (2024) 'Family resilience during COVID-19: A study of adaptation and well-being among social work students', *AVE Trends in Intelligent Social Letters*, Vol. 1, No. 2, pp.93–103.

Shaleena, K.P. and Paul, S. (2015) 'Data mining techniques for predicting student performance', in *2015 IEEE International Conference on Engineering and Technology* (*ICETECH*), Coimbatore, India, pp.1–3.

Sharma, V.Y., Vashist, S., Deeksha, R. and Khusba, T.A. (2024) 'Educational intervention on menstrual hygiene knowledge and attitude among adolescent girls in Dehradun: a quasi-experimental study', *FMDB Transactions on Sustainable Health Science Letters*, Vol. 2, No. 2, pp.94–109.

Shruthi, S. and Aravind, B.R. (2023) 'Engaging ESL learning on mastering present tense with nearpod and learningapps.org for engineering students', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 1, pp.21–31.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T. and Hernandez, M. (2018) 'Perspectives to predict dropout in university students with machine learning', in *2018 IEEE International Work Conference on Bioinspired Intelligence* (*IWOBI*), Alajuela Province, Costa Rica, pp.1–6.

Solomon, D., Patil, S. and Agrawal, P. (2018) 'Predicting performance and potential difficulties of university student using classification: survey paper', *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 18, pp.2703–2707.

Thammasiri, D., Delen, D., Meesad, P. and Kasap, N. (2014) 'A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition', *Expert Systems with Applications*, Vol. 41, No. 2, pp.321–330.

Tripathi, S. and Al-Zubaidi, A. (2023) 'A study within Salalah's higher education institutions on online learning motivation and engagement challenges during COVID-19', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 1, pp.1–10.

Varmann, S.S., Hariprasath, G. and Kadirova, I. (2023) 'Optimizing educational outcomes: H2O gradient boosting algorithm in student performance prediction', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 3, pp.165–178.

Velmonte, G.L. (2023) 'Preferred college degree programs among senior high school students: a policy recommendation', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 3, pp.143–155.

Venkatasubramanian, S., Gomathy, V. and Saleem, M. (2023) 'Investigating the relationship between student motivation and academic performance', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 2, pp.111–124.

Vijayarani, K., Nithyanantham, V., Angelene Christabel, and Marupaka, D. (2023) 'A study on relationship between self-regulated learning habit and achievement among high school students', *FMDB Transactions on Sustainable Techno Learning*, Vol. 1, No. 2, pp.92–110.

Yukselturk, E., Ozekes, S. and Türel, Y.K. (2014) 'Predicting dropout student: an application of data mining methods in an online education program', *European Journal of Open, Distance and E-Learning*, Vol. 17, No. 1, pp.118–133.

Zimmermann, J., Brodersen, K.H., Pellet, J-P., August, E. and Buhmann, J.M. (2011) 'Predicting graduate-level performance from undergraduate achievements', in *EDM*, Citeseer, New Jersey, United States of America, pp.357–358.