K-means optimization with bat algorithm for predicting diabetes and hypertension risk in athletes' comparison with machine learning

A'yunin Sofro [a,1,*], Danang Ariyanto [a,2], Junaidi Budi Prihanto [b,3], Dimas Avian Maulana [a,4], Riska Wahyu Romadhonia [a,5], Asri Maharani [c,6], Affi Oktaviarina [a,7], Ibnu Febry Kurniawan [d,8], Khusnia Nurul Khikmah [e,9], Muhammad Mahdy Al Akbar [f,10]

[a] Actuarial Science Department, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
[b] Sport Education Department, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
[c] School of Health Sciences, Department of Nursing, Manchester Metropolitan University, Bonsall St, Manchester, M15 6GX, United Kingdom
[d] Data Sciences Department, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
[e] Mathematics Department, Universitas Palangka Raya, Palangka Raya, Central Kalimantan, 74874 Indonesia
[f] Mathematics Department, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
[1] ayuninsofro@unesa.ac.id; [2] danangariyanto@unesa.ac.id; [3] junaidibudi@unesa.ac.id; [4] dimasmaulana@unesa.ac.id;
[5] riskaromadhonia@unesa.ac.id; [6] asri.maharani@manchester.ac.uk; [7] affiatioktaviarina@unesa.ac.id; [8] ibnufebry@unesa.ac.id;
[9] khusnia.nurulkhikmah@mipa.upr.ac.id; [10] muhammad.21061@mhs.unesa.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 24, 2024
Revised August 11, 2025
Accepted September 13, 2025
Available online November 30, 2025

Keywords
Bat Optimization
Extremely Randomized Trees
Machine learning Classification
Support Vector Classification

ABSTRACT

This research aims to develop an analytical approach to classification statistics. The proposed approach combines machine learning with optimization. Considering the urgency of research related to exploring the best methods to apply to sports data. This study proposes a novel framework that combines the k-means clustering results with the bat algorithm to optimize performance prediction for athletes in Indonesia. The proposed method aims to explore the data by comparing the classification performance of random forests, extremely randomized trees, and support vector machines. We conducted a case study using primary data from 200 respondents at Surabaya State University and the East Java National Sports Committee. The accuracy results in this study indicate that, based on the performance evaluation metric, the best approach is random forest clustering using k-means with bat algorithm optimization, achieving 81.25% accuracy, compared with other machine learning approaches. This research contributes to the field of classification statistics by introducing a novel hybrid framework that integrates machine learning, clustering, and optimization techniques to improve predictive accuracy, particularly in sports analytics. Beyond sports science, the proposed approach can be adapted to other domains that require robust performance prediction and decision support, such as health analytics, educational assessment, and human resource selection.

1. Introduction

Classification is a fundamental branch of statistics whose main task is to place data points into predefined categories and is a key component of its use in daily life, assisting decision-making and diagnostics [1]. Therefore, the development is characterized by the emergence of the latest classification approaches in data analysis. Traditional and newer classification approaches require particular study to

determine the merits of these methods. In addition, quality data and the right type for analysis are fundamentals in the methodological approach. One type of data is categorical. Categorical data is a type of data used to represent different categories so that it does not have a numerical order. This data also has several types, some of which are nominal and ordinal data. Nominal data is without any order or ranking among the categories, while ordinal data is the opposite [2]. The latest development in this data analysis approach is the k-means clustering method in machine learning. The k-means method aims to group data and then classify it based on the similarity of features, using the resulting grouping to inform the classification, making this method simple, efficient, and easy to understand [3]. However, it is an initialization-sensitive method, which can lead to variability in the final results. Therefore, this research proposes a clustering optimization approach to minimize the clustering error arising from the shortcomings of the k-means method, thereby enabling further classification [4].

The latest development in this optimization approach is the bat algorithm method [5]. The bat algorithm is a key component of the proposed optimization approach. It is an optimization method inspired by the echolocation behavior of bats in dark caves to detect the presence of prey. This optimization method identifies the optimal solution because it has relatively few parameters that must be set efficiently [6]. Previous research indicates that the bat algorithm achieves the highest routing accuracy for handling industrial internet problems, with lower computational cost than the bat algorithm and particle swarm optimization methods. The findings of this study state that the bat algorithm method shows better accuracy results than the particle swarm optimization method, with an accuracy of 85.50% [7].

The latest classification method approach is machine learning. Machine learning is a branch of artificial intelligence that allows learning from complex data quickly [8]. This research is significant as it provides new insights for researchers in the field of data analysis and machine learning. Various machine learning classification algorithms provide researchers with new insights to determine the best approach to achieve the goal. Three machine learning algorithms are proposed in this study, namely random forest [9], extremely randomized trees [10], and support vector classification [11] to be compared. Previous research found that classification results with random forest algorithms were better than those of decision trees in highway-rail class assessment. Other research examines the extremely randomized trees approach, where this algorithm is an extension of random, and the accuracy results obtained for classification are very good [10]. Another recent research related to classification is support vector classification, where research in [1] found that this approach has good classification accuracy. Therefore, this research aims to compare the performance of random forests, extremely randomized trees, support vector classification, and random forest methods obtained by k-means clustering with BA optimization.

The daily problem of categorical data is data in the sports industry. The data in this sports case were selected for this study based on the importance of selecting athletes for various necessary needs [12]. Competitive competition in various sports makes athletes the primary concern. Optimal athletic performance is determined not only by talent and intensive training but also by excellent physical and health conditions [13]. This physical condition reflects information related to the influence of athlete performance so that the risk of injury can be minimized and improved performance, thereby extending the athlete's career [14]. This problem is of particular concern to related parties across various matters related to national and international competitions in Indonesia. The selection of athletes is crucial in helping Indonesian fighters improve their national and international sports achievements. This selection requires fair conditions for each individual, requiring various related information about the individual. The individual's primary health condition can be optimized to improve their chances. Therefore, the individual's health history and socio-demographics are essential considerations.

Previous research examined the relationship between athletes and hypertension factors and found that hypertension was the most risky factor in becoming an athlete [15]. Research [16] states that risk factors for hypertension are critical because hypertension is one of the leading causes of heart disease, which can cause death. In addition, the high-risk factor of hypertension in athletes, as detailed in [17],

is that it causes damage to small blood vessels in the muscles, which increases the risk of cramps and muscle tears. Other research suggests that health history with risk factors for diabetes also significantly affects athlete performance [18]. The effect of diabetes studied by [19] on athletes states that athletes with a history of diabetes have lower performance than those without diabetes, which increases the risk of muscle, bone, and nerve damage or diabetic neuropathy. These two health factors are special considerations in the selection of athletes. These considerations can be obtained through statistical analysis, a valuable tool for decision-making in the sports industry. Therefore, the proposed research examines the performance of k-means clustering with bat optimization, followed by prediction using random forest, which was compared with predictions using random forest, support vector classification, and extremely random trees, intending to become one of the tools that contribute to the advancement of sports analytics and maximize the potential of achievements that can be achieved.

2. Method

2.1. K-Means Algorithm

The most popular clustering algorithm in machine learning is k-means. This approach generally works by identifying group centers and then clustering the data based on their closest group. Specifically, this approach divides n objects into k predefined clusters into one cluster [20] where data in one cluster has a very high level of similarity between its members, while with other clusters, the level of similarity is low.

The k-means approach is popularly used in analyzing because of its algorithm simplicity and efficiency against complex data. However, it becomes a problem when the parameters are improperly initialized. The general stages of clustering with the k-means algorithm are as follows [21].

1. Initialization stage
   a. Determine the number n of clusters to be formed.
   b. Randomly select the center point (centroid), k.
2. Calculate the distance of each object to the center point. To calculate the distance, use Euclidean Distance, where x is the object and y is the center point. Where $(x_i - y_i)$ is the difference in distance between the object (x_i) and the center point (y_i).

$$d_{Euclidean}(x.y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

3. Put each object into the cluster with the closest distance.
4. Update the centroid by calculating the average of all objects in each cluster. Where the new centroid is calculated, if the kth cluster is symbolized by C_k and N is the number of objects in a cluster, the equation follows.

$$C_k = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (2)$$

5. Repeat step 2 until there is no change in the centroid.

2.2. Bat Algorithm

The Bat Algorithm, a ground-breaking concept first proposed by Xin-She Yang in 2010, is a simulation of the echolocation behavior of microbats. This innovative algorithm simulates a population of bats in search of food, each representing a potential solution. When a bat finds food, its pulse-sending speed increases, shortening the echolocation time and improving the location accuracy [22]. The movement of bats in generating an algorithm, where each individual (i) has a previous position $x_i(t-1)$ and a previous velocity $v_i(t-1)$ in the search space, it will be updated as the number of iterations increases. The new position $x_i(t)$ and velocity $v_i(t)$ at time t with values $f_{\min}$ is 0 and f_{max} is 1, can be defined as follows [5].

$$f_i = f_{min} + (f_{max} - f_{min}) \cdot U$$
$$v_i(t) = v_i(t-1) + (x_i(t-1) - X^*) \cdot f_i \quad (3)$$
$$x_i(t) = x_i(t-1) + v_i(t)$$

whre U is a random vector where $U \sim Uniform[0,1]$ and X^* is the current global optimal solution obtained by comparing with all solutions among n bats.

The ability of the bat algorithm in global or local search depends on its parameters, so it needs the best solution to balance between the two. This solution for each individual i can be generated using a random walk.

$$x(t) = x(t-1) + \varepsilon A^t \quad (4)$$

where ε is a random value $\varepsilon \in [-1,1]$ and A^t is the average population loudness. Furthermore, the loudness is updated when the bat finds prey and can be controlled (up and down) at each level by performing a global search. If the controlled loudness is symbolized by $A_i(t+1)$ and the pulse is symbolized by $r_i(t+1)$, then the global search can be defined as follows.

$$r_i(t+1) = r_i(0)\left[1 - e^{(-\gamma t)}\right] A_i(t+1) = \alpha A_i(t) \quad (5)$$

where α and γ are constants where $\alpha > 0$ and $\gamma > 0$. $r_i(0)$ is the initial value of the pulse [23].

2.3. Random Forest (RF)

In 2001, Breiman introduced an ensemble learning algorithm, a method widely used in classification and regression tasks. The random forest, a popular and robust approach, is a unique combination of an ensemble technique and bagging [24]. Its simplicity and effectiveness in classification are attributed to its random selection of n samples from the training data, a process known as bootstrapping, and the subsequent generation of new data samples [25]. The random forest method is robust to overfitting, owing to its randomized mechanism that mitigates this risk. Moreover, it is robust to data noise, enabling it to perform well across diverse data scenarios. This adaptability makes it a preferred classification strategy, as it only requires the optimization of two parameters: the number of trees ($ntree$) and the number of features considered in splitting the decision tree nodes randomly ($mtry$) [26].

Previous research found that the best number of trees for cancer cases was 100 [27]. In contrast, the number of $mtry$ Used in the random forests does not have definite rules. However, the general guideline is that for data with many features, a small mtry value, namely by using the square root of the number of features, gives the best results.

Generally, the stages of classification analysis using the random forest method are as follows.

1. Determine the number of trees to be generated (k).
2. Each tree is randomly sampled using n returns to the training data.
3. Select a random subset of m explanatory variables for each tree with $m < p$.
4. Repeat steps 2 and 3 for every k trees.
5. The random forest prediction result is determined by the most votes based on the classification results of k trees.

2.4. Extremely Randomized Trees (Extra Trees)

In the mid-2000s, the extremely randomized trees method was developed. This method, a variation of the random forest, is known for its robustness to noise. Unlike the random forest, where feature selection and threshold value selection are based on impurity reduction, the extremely randomized trees method randomly selects both, leading to a high tree diversity [28]. This algorithm, used for regression

and classification tasks, is renowned for its accuracy and robustness, preventing overfitting and often outperforming the random forest [10]. The extremely randomized trees method employs a unique approach to tree generation. It creates a highly randomized tree, an ensemble-based classification strategy. This strategy combines a single tree with severe randomization. The feature selection and point assignment for node splitting are based on the randomization technique. As a result, each tree constructed from the training data is highly randomized to reduce drift.

The extremely randomized trees method, using training data in general, has the following classification stages.

1. The stage of selecting the best split:
   a. Randomly select m features
   b. Randomly select k-cut points
   c. Determine the best splitting criteria
   d. Repeat steps a to c until reaching the stopping criterion so that the prediction result of one tree is obtained.
2. Repeating step a until m trees are formed
3. Combining the estimation results obtained from each classification tree using the most votes.

2.5. Support Vector Classification (SVC)

The support vector classification method was introduced in the early 1990s and was initially used for classification, regression, and outlier detection. This method can classify high-dimensional data while being robust to overfitting and effective on nonlinear data [1]. However, with its development, this method is best known for its classification capabilities. The initial concept of this approach was for binary classification ($k = 2$). Problems that occur around not only require binary classification but also multi-class classification. Therefore, research on the development of the support vector classification approach to this multi-class problem was conducted. Previous research introduced the issue of multi-class classification ($k > 2$) with support vector classification in schizophrenia cases, where the analysis conducted showed good accuracy results. This research has practical implications across a wide range of fields, including data science, machine learning, and agricultural research, and informs future studies and applications [29].

In general, in the support vector classification method of non-linear training data, the kernel technique replaces the dot product between two vectors in the input space with a function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with an implicit mapping from the input space to the high-dimensional feature space is $k(x, x') = \langle \phi(x), \phi(x') \rangle \mathcal{F}$ for each vector x where $x' \in \mathcal{X}$ and $\phi: \mathcal{X} \rightarrow \mathcal{F}$. If the Lagrange multiplier is symbolized by α_i and the absolute value is represented by y_i, where $y_i = 1$ is if x_i is on the positive side of h and path w and -1 otherwise and b is the term of bias. For the $sign(\ldots)$ is sign function and represents a real number as its input and output. Then, the support vector classification function prediction $f(x)$ is defined as follows [30].

$$f(x) = sign(\sum_i \alpha_i y_i k(x_i, x) + b) \tag{6}$$

2.6. Data

This study uses primary data obtained from 200 athlete selection participants at Surabaya State University and the East Java Indonesian National Sports Committee, with the flowchart of the analysis provided in Fig. 1. The analysis began with data preprocessing, splitting data (training data and testing data), followed by clustering using the K-means method optimized by the Bat Algorithm (BA) to find hidden patterns by following integration process on previous study [31], minimizing variance within clusters or maximizing the silhouette score. The resulting cluster labels and cluster center distances were added to the original dataset as new features see Fig 1. Then, the clustering results (new features) are used for classification using the random forest method, which is compared with the classification of the

original data using the random forest method, support vector classification, and extremely randomized trees. Next, the data is split, where the training data is used for modelling and the test data is used for evaluation. Modelling is then performed using the selected model as shown in Fig 2. The model is evaluated using accuracy, sensitivity, specificity, and F1 score metrics. Finally, the best model is determined based on the prediction results.

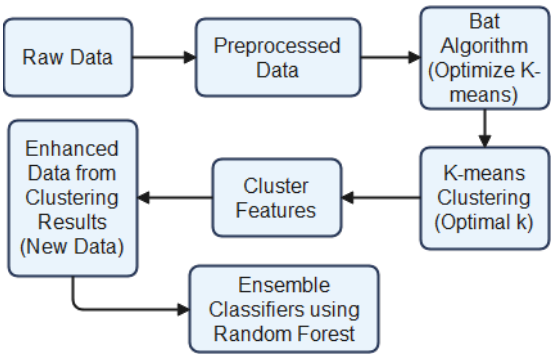

Fig. 1. Flowchart of detailed hybridization

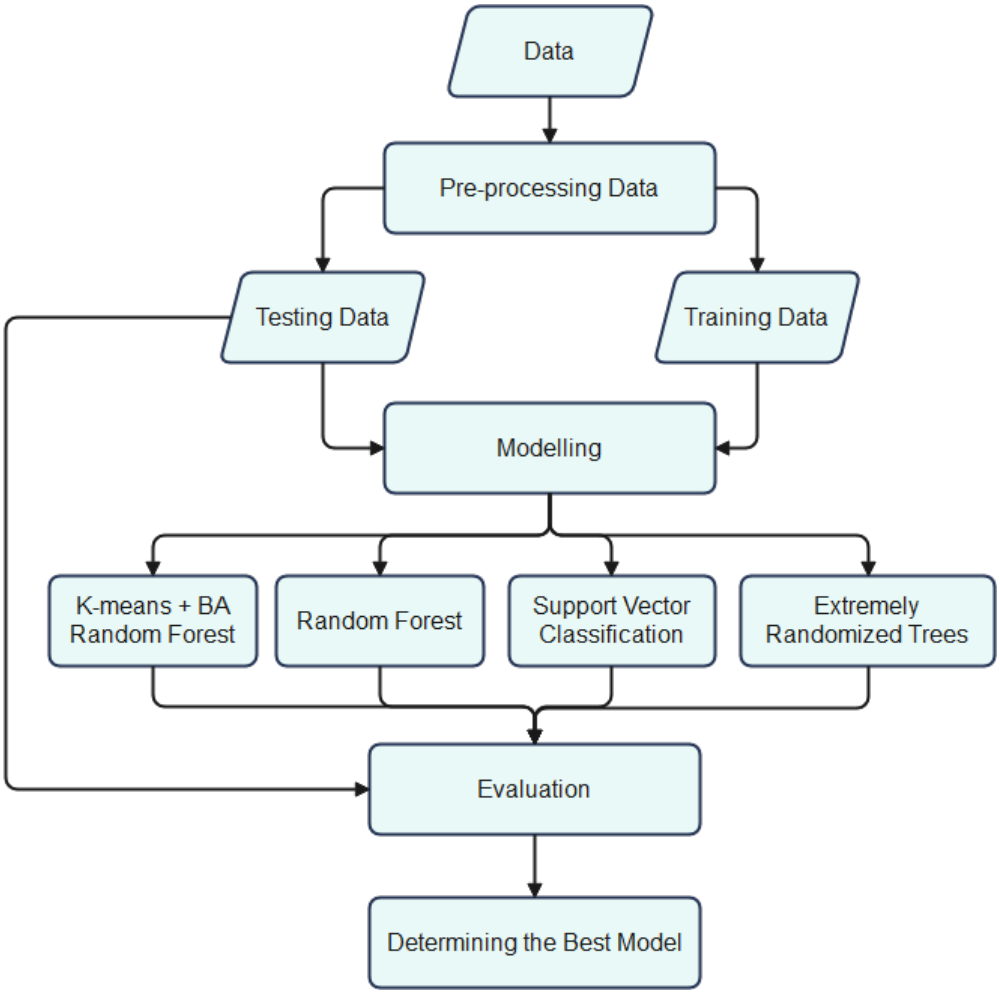

Fig. 2. Flowchart of analysis

The data have been used in previous analyses [32], which have a different purpose from this study, whichc focuses on exploring the relationship between socio-demographic and anthropometric factors. The response variable (Y) used in this study comprises 119 respondents who did not qualify and 41, who

qualified as athletes. Thus, Fig. 1 shows that the distribution of the response variable categories used in this study is unbalanced. This value is presented in Fig. 3.

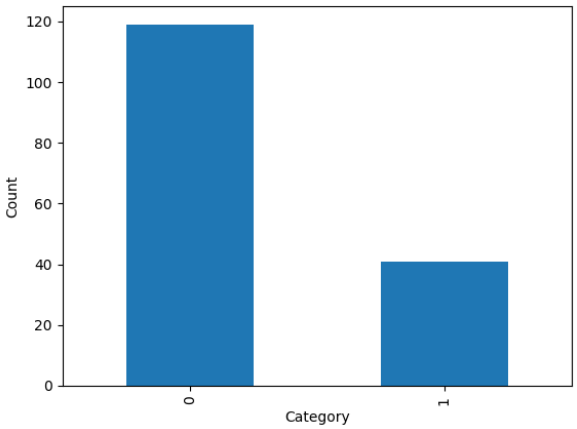

Fig. 3. Distribution of response variable category

3. Results and Discussion

3.1. K-Means with Bat Optimization

In accordance with the research objectives, this study examines the performance of k-means clustering with bat optimization, followed by prediction using a random forest, which is compared with predictions from support vector classification and extremely randomized trees. This study then analyzed the general characteristics of the data used, as clearly described in previous studies [32]. The comparison of the performance of the approaches proposed in this study begins by dividing the data into two groups, namely 80% training data and 20% Test data. This division is significant because it allows training on a large portion of the data, while the test data serves as an independent dataset for evaluating model performance. Meanwhile, the three classification approaches proposed in the study are random forest, extremely randomized trees, and support vector classification.

Classification accuracy in this research proposes a new view of data classification obtained based on clustering results with optimization, compared to the original data of machine learning classification without clustering and optimization. Clustering aimed at using the k-means method and bat algorithm as optimization, with n-cluster parameters used where $n = 2$ Fig 4. The selection of $n = 2$ based on Fig 3, which is indicated by the cluster point at $n = 2$, is a turning point, meaning that adding more clusters with $n > 2$ will be less significant.

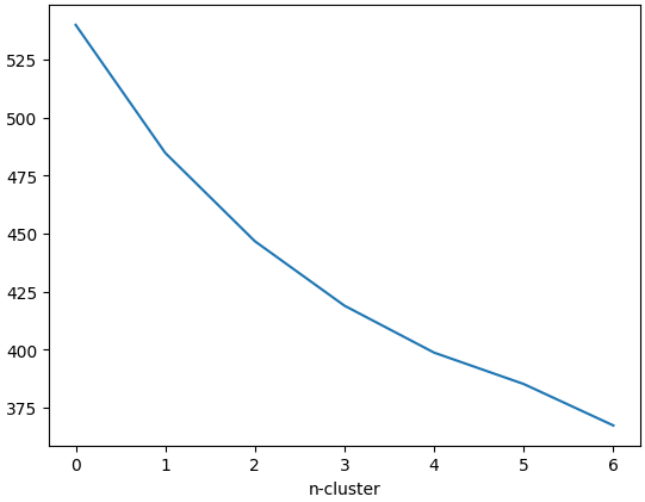

Fig. 4. The optimal cluster plot number of k-means

The bat algorithm optimization parameters used are detailed in Table 1. The bat algorithm plays a crucial role in optimization, helping to fine-tune the model's parameters to improve classification accuracy.

Table 1. Bat Algorithm Model's Parameter for Optimization

Methods	Parameter	Value
Bat Algorithm	Bats	20
	Iteration	1000
	Bat's pulse (α)	0.97
	Pulse emission rate (γ)	0.1
	f minimum	0
	f maximum	2
	Loudness	0.5
	Pulse rate	0.5
	Maximum iteration	10
Random Forest	n-estimators	100
	Max-depth	3
Extremely Randomized Trees	n-estimators	100
	Max-depth	3
Support Vector Classification	Kernel: Radial Basis Function (γ)	$\frac{1}{n\,features * X.var}$

The bat algorithm parameters are used with the proposed goodness measure to evaluate the quality of each optimization algorithm, and the algorithm quality measure to achieve the optimal value of the objective function is the maximum objective function. This optimization also proposes initializing the parameters α and γ at the outset, using the best values from prior research.

This algorithm has two stopping iterations where the iteration with a value of 1000 is the maximum iteration for convergent conditions while the maximum iteration of 10 is the iteration with conditions where the global test does not improve. The clustering results with $n = 2$ clusters are then classified using the random forest method and compared with data without clustering optimization. The random forest parameters used in both are proposed with the exact specification: the decision tree ($n - estimator$) is 100, this value chosen based on a previous study on [24], and the other model complexity control parameter used is the maximum depth (max_{depth}), which is 3, aligning with a previous study on [33]. The proposed classification evaluation is accuracy, sensitivity, specificity, precision, and F1-score. This value estimates the model's goodness in the form of percent. This evaluation value is obtained based on the error of the tested approach on the test data compared to its predicted result. This accuracy is shown in Table 2.

Table 2 shows the performance of the model where the best machine learning classification proposed are random forest and random forest clustering results of the k-means method with bat algorithm optimization. The difference in accuracy produced between the two is significant, or around 8.25%. According to previous research on [34], the best model accuracy category obtained is the good category. These model accuracy results are significant as they demonstrate the effectiveness of the proposed approach in improving data classification accuracy.

3.2. Machine Learning Classification

The analysis results with other approaches proposed in this study by adjusting the latest approaches to classification with machine learning are extremely randomized trees and support vector classification. These two approaches are proposed with data specifications without clustering to compare the novelty of the approach with the data optimization results. The initialization parameters of the proposed

extremely randomized trees approach generally have the exact specifications as random forest, namely $n-estimators$ 100 and max_{depth} 3. The parameter specification used in the support vector classification approach is a radial basis function kernel to map data to a higher feature space according to the characteristics of the data. The complexity of the model in the radial basis function kernel is symbolized by γ, where $\gamma = \frac{1}{n\,features * X.var}$. The performance measure used to evaluate these two approaches is the same as before: accuracy, computed from the prediction error relative to the original classification results on the test data. The accuracy of these two approaches is shown in Table 2. The model performance was evaluated and compared in a meticulous manner. The best classification approach obtained was support vector classification. However, when compared overall across the four proposed approaches, the random forest method obtained by k-means clustering with bat algorithm optimization was the best, based on the highest accuracy in Table 2.

Table 2. Model Evaluation

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score
K-Means Bat with Random Forest	81.25%	85.00%	80.00%	80.95%	82.93%
Random Forest	73.00%	75.00%	70.00%	71.43%	73.17%
Extremely Randomized Trees	65.00%	65.00%	65.00%	65.00%	65.00%
Support Vector Classification	70.00%	70.00%	70.00%	70.00%	70.00%

The significant difference in performance between the renewed extremely randomized trees approach and the proposed combination approach is likely due to the extremely randomized trees approach's sensitivity to unbalanced data and the support vector classification method's sensitivity to data scale. In comparison, this research has unbalanced data, with the features used as scaled data. However, comparing the accuracy of data classification results without optimization is in line with previous research on [35] in the case of food insecurity, which compares several machine learning methods. The results of that study, obtained by the random forest method, have the best classification accuracy compared to extremely randomized trees and rotation forests, with an accuracy of 65.8%. Research on [36] examines bat algorithm optimization in diabetes cases, where the accuracy is 72.4%. These results align with this study, which found that the Random Forest clustering results of the k-means method with bat algorithm optimization achieved the highest accuracy.

This research aims to enhance classification performance using the random forest approach by optimizing clustering results. The classification results also provide new insights into the comparison of classification performance with other machine learning approaches. Evaluation of the classification ability of random forest through clustering results with optimization proven to improve classification performance on sports data sets. The strength of the proposed approach is that it obtains the best accuracy in achieving the goal. The results of our analysis show that random forest classification using clustering and optimization achieves higher accuracy, with improvements of 8.25% over random forest, 11.25% over support vector classification, and 16.25% over extremely randomized trees. These improvements are significant in machine learning, indicating a substantial enhancement in classification performance. Our results support the consistency of findings with previous research on optimization performance using the bat algorithm. In addition, our results support a more dynamic view of the performance of machine learning approaches with optimization based on response criteria with futures.

According to previous research [37], the bat algorithm improves neural network classification performance over support vector classification by 5% on heart disease data in Switzerland. These findings also include multi-class cases where the bat algorithm optimization improves model performance. The neural network-based classification approach, which was targeted in [7] to assess performance reliability by optimizing with the bat algorithm, outperformed particle swarm optimization in data prediction.

The prediction results obtained in this study support the selection of athletes by Surabaya State University and the East Java Indonesian National Sports Committee. This study also shows that the factors of health history and demographics of athletes significantly affect athlete selection. The findings in this study support previous research, which shows that health history significantly affects a person's

ability to become an athlete [36]. The demographic influence of whether or not someone can become an athlete is also significantly influenced by socio-demographic factors, supporting previous research [37]. This result shows that athletes use their energy and body to support their activities. The athlete's prime condition is a significant factor that provides an excellent opportunity for athletes to improve their national sports achievements.

The results of this scientific and measurable approach can also minimize injuries that may occur in athletes due to their medical history. In addition, fairness in the national athlete selection process can provide a more ethical view of an athlete's chances of qualifying. The results of this study also have limitations, as shown in Table 1, particularly with respect to parameter initialization: the optimization parameters were set by the researcher at the outset, introducing potential bias. In addition, the data used in the study were unbalanced, and no unbalanced data handling was performed. Thus, the analysis results were based on the original data without preprocessing. Therefore, future research is expected to explore optimization methods combined on multi-class data with and without class imbalance to determine the proposed classification method's performance, reveal the approach's reliability, and determine the behaviour and interaction between the proposed response variable and its features with repetition experiments. In addition, parameter selection strategies across different datasets should be a particular focus of future research, with hyperparameter tuning aimed at identifying the optimal parameters for the proposed approach while minimizing running time. Future research should also address the weaknesses found in this study. In addition to selecting features for future research, it is expected to use an optimization that yields only features with a substantial influence.

4. Conclusion

The analysis classified the respondents' sports data for athlete selection using three machine learning approaches, with additional optimization via clustering, and yielded accurate results according to the accuracy evaluation matrix. The analysis indicates that the best machine learning method is random forest with clustering and bat optimization, with this conclusion based on an accuracy performance metric value of 81.25%. Additionally, the analysis compared machine learning approaches' performance and resilience to imbalanced data, with all methods achieving accuracy values $> 60.00\%$. This finding confirms that incorporating optimized clustering results as additional features can effectively capture hidden data structures and improve predictive capability beyond conventional classification approaches. Furthermore, the comparative analysis highlights that although all evaluated models achieved acceptable performance (accuracy above 60%), their robustness varied under class imbalance. Extremely randomized trees and support vector classification showed greater sensitivity to unbalanced data and data scaling, whereas the optimized random forest model exhibited superior stability and generalization. The observed accuracy improvement of up to 16.25% compared to non-optimized models underscores the practical benefit of hybridizing machine learning with metaheuristic optimization techniques in sports analytics. From an applied perspective, the proposed framework provides a reliable and objective decision-support tool for athlete selection, particularly by incorporating health history and socio-demographic factors that are critical to performance, injury prevention, and long-term athletic sustainability. Beyond sports science, the methodological contribution of this research lies in its hybrid analytical framework, which can be adapted to other domains that require robust classification under complex, imbalanced data conditions, such as health risk prediction, educational assessment, and human resource selection.

Acknowledgment

The authors express their gratitude to the Directorate General of Higher Education, Ministry of Education and Culture, for the basic research grant that provided funding to support this research.

Declarations

Author contribution. A'yunin Sofro developed the research idea and designed the study. Danang Ariyanto, Dimas Avian Maulana, Riska Wahyu Romadhonia, and Affi Oktaviarina contributed to developing the research methods and data collection procedures. Danang Ariyanto, Dimas Avian

Maulana, Riska Wahyu Romadhonia, Affi Oktaviarina, and Muhammad Mahdy Al Akbar collected, organized, and managed the data used in the study. Khusnia Nurul Khikmah performed the data analysis and interpretation. A'yunin Sofro and Khusnia Nurul Khikmah drafted and revised the manuscript. Junaidi Budi Prihanto, Asri Maharani, and Ibnu Febry Kurniawan reviewed and edited the manuscript for clarity, coherence, and accuracy. A'yunin Sofro provided oversight and guidance to the research project and obtained funding for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The data used in this study are available to the corresponding author upon reasonable request. The data will be provided in CSV format and the analyses in this study were conducted using Python. The code used for the analyses is available upon reasonable request.

References

[1] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, no. September, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[2] G. Tutz, "Ordinal regression: A review and a taxonomy of models," *WIREs Comput. Stat.*, vol. 14, no. 2, pp. 1–28, Mar. 2022, doi: 10.1002/wics.1545.

[3] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renew. Sustain. Energy Rev.*, vol. 120, no. March, p. 109628, Mar. 2020, doi: 10.1016/j.rser.2019.109628.

[4] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions," *Appl. Sci.*, vol. 11, no. 23, p. 11246, Nov. 2021, doi: 10.3390/app112311246.

[5] T. Agarwal and V. Kumar, "A Systematic Review on Bat Algorithm: Theoretical Foundation, Variants, and Applications," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2707–2736, Aug. 2022, doi: 10.1007/s11831-021-09673-9.

[6] I. Karakonstantis and A. Vlachos, "Bat algorithm applied to continuous constrained optimization problems," *J. Inf. Optim. Sci.*, vol. 42, no. 1, pp. 57–75, Jan. 2021, doi: 10.1080/02522667.2019.1694740.

[7] A. Alharbi, W. Alosaimi, H. Alyami, H. T. Rauf, and R. Damaševičius, "Botnet Attack Detection Using Local Global Best Bat Algorithm for Industrial Internet of Things," *Electronics*, vol. 10, no. 11, p. 1341, Jun. 2021, doi: 10.3390/electronics10111341.

[8] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[9] R. Genuer and J.-M. Poggi, *Random Forests with R*, 1st ed. in Use R! Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-56485-8.

[10] U. Saeed, S. U. Jan, Y.-D. Lee, and I. Koo, "Fault diagnosis based on extremely randomized trees in wireless sensor networks," *Reliab. Eng. Syst. Saf.*, vol. 205, no. January, p. 107284, Jan. 2021, doi: 10.1016/j.ress.2020.107284.

[11] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*, Elsevier, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[12] D. Patel, D. Shah, and M. Shah, "The Intertwine of Brain and Body: A Quantitative Analysis on How Big Data Influences the System of Sports," *Ann. Data Sci.*, vol. 7, no. 1, pp. 1–16, Mar. 2020, doi: 10.1007/s40745-019-00239-y.

[13] K. Till and J. Baker, "Challenges and [Possible] Solutions to Optimizing Talent Identification and Development in Sport," *Front. Psychol.*, vol. 11, no. April, p. 525518, Apr. 2020, doi: 10.3389/fpsyg.2020.00664.

[14] M. Lopes Dos Santos *et al.*, "Stress in Academic and Athletic Performance in Collegiate Athletes: A Narrative Review of Sources and Monitoring Strategies," *Front. Sport. Act. Living*, vol. 2, p. 502979, May 2020, doi: 10.3389/fspor.2020.00042.

[15] V. Schweiger, D. Niederseer, C. Schmied, C. Attenhofer-Jost, and S. Caselli, "Athletes and Hypertension," *Curr. Cardiol. Rep.*, vol. 23, no. 12, p. 176, Dec. 2021, doi: 10.1007/s11886-021-01608-x.

[16] C. McHugh, K. Hind, J. Cunningham, D. Davey, and F. Wilson, "A career in sport does not eliminate risk of cardiovascular disease: A systematic review and meta-analysis of the cardiovascular health of field-based athletes," *J. Sci. Med. Sport*, vol. 23, no. 9, pp. 792–799, Sep. 2020, doi: 10.1016/j.jsams.2020.02.009.

[17] C. M. Baker-Smith and T. Tsuda, "Exercise Testing in Hypertension and Hypertension in Athletes," in *Pediatric Hypertension*, Cham: Springer International Publishing, 2023, pp. 827–842. doi: 10.1007/978-3-031-06231-5_12.

[18] T. Trojian *et al.*, "American Medical Society for Sports Medicine Position Statement on the Care of the Athlete and Athletic Person With Diabetes," *Clin. J. Sport Med.*, vol. 32, no. 1, pp. 8–20, Jan. 2022, doi: 10.1097/JSM.0000000000000906.

[19] C. J. Holmes and M. K. Hastings, "The Application of Exercise Training for Diabetic Peripheral Neuropathy," *J. Clin. Med.*, vol. 10, no. 21, p. 5042, Oct. 2021, doi: 10.3390/jcm10215042.

[20] W. Xu, M. Tang, and Y. Li, "A new method for assessment of regional drought risk: information diffusion and interval mapping adjustment based on k-means cluster points," *J. Water Clim. Chang.*, vol. 13, no. 12, pp. 4302–4316, Dec. 2022, doi: 10.2166/wcc.2022.345.

[21] K. Zhou and S. Yang, "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering," *Pattern Anal. Appl.*, vol. 23, no. 1, pp. 455–466, Feb. 2020, doi: 10.1007/s10044-019-00783-6.

[22] B. Alsalibi, L. Abualigah, and A. T. Khader, "A novel bat algorithm with dynamic membrane structure for optimization problems," *Appl. Intell.*, vol. 51, no. 4, pp. 1992–2017, Apr. 2021, doi: 10.1007/s10489-020-01898-8.

[23] L. F. Zhu, J. S. Wang, H. Y. Wang, S. S. Guo, M. W. Guo, and W. Xie, "Data Clustering Method Based on Improved Bat Algorithm With Six Convergence Factors and Local Search Operators," *IEEE Access*, vol. 8, no. April, pp. 80536–80560, 2020, doi: 10.1109/ACCESS.2020.2991091.

[24] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[25] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, New York, NY: Springer New York, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7_5.

[26] S. Park, S.-Y. Hamm, and J. Kim, "Performance Evaluation of the GIS-Based Data-Mining Techniques Decision Tree, Random Forest, and Rotation Forest for Landslide Susceptibility Modeling," *Sustainability*, vol. 11, no. 20, p. 5659, Oct. 2019, doi: 10.3390/su11205659.

[27] A. Arfiani and Z. Rustam, "Ovarian cancer data classification using bagging and random forest," in *AIP Conference Proceedings*, American Institute of Physics Inc., Nov. 2019, p. 020046. doi: 10.1063/1.5132473.

[28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[29] T. V Rampisela and Z. Rustam, "Classification of Schizophrenia Data Using Support Vector Machine (SVM)," *J. Phys. Conf. Ser.*, vol. 1108, no. 1, p. 012044, Nov. 2018, doi: 10.1088/1742-6596/1108/1/012044.

[30] C. AVCI, M. BUDAK, N. YAĞMUR, and F. BALÇIK, "Comparison between random forest and support vector machine algorithms for LULC classification," *Int. J. Eng. Geosci.*, vol. 8, no. 1, pp. 1–10, Feb. 2023, doi: 10.26833/ijeg.987605.

[31] A. Sofro, D. Ariyanto, J. Budi Prihanto, D. A. Maulana, R. W. Romadhonia, and A. Maharani, "Integration of Bivariate Logistic Regression Models and Decision Trees to Explore the Relationship between Socio-Demographic and Anthropometric Factors with the Incidence of Hypertension and Diabetes in Prospective Athletes," *Sport Mont*, vol. 22, no. 1, pp. 71–78, Feb. 2024, doi: 10.26773/smj.240210.

[32] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1301, May 2019, doi: 10.1002/widm.1301.

[33] G. M. Foody, "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification," *Remote Sens. Environ.*, vol. 239, no. March, p. 111630, Mar. 2020, doi: 10.1016/j.rse.2019.111630.

[34] S. Md Mujeeb, R. Praveen Sam, and K. Madhavi, "Adaptive Exponential Bat algorithm and deep learning for big data classification," *Sādhanā*, vol. 46, no. 1, p. 15, Dec. 2021, doi: 10.1007/s12046-020-01521-z.

[35] A. K. Heather *et al.*, "Biological and Socio-Cultural Factors Have the Potential to Influence the Health and Performance of Elite Female Athletes: A Cross Sectional Survey of 219 Elite Female Athletes in Aotearoa New Zealand," *Front. Sport. Act. Living*, vol. 3, p. 601420, Feb. 2021, doi: 10.3389/fspor.2021.601420.

[36] N. Leite, J. Arede, X. Shang, J. Calleja-González, and A. Lorenzo, "The Influence of Contextual Aspects in Talent Development: Interaction Between Relative Age and Birthplace Effects in NBA-Drafted Players," *Front. Sport. Act. Living*, vol. 3, no. March, p. 642707, Mar. 2021, doi: 10.3389/fspor.2021.642707.

[37] A. Lisinskiene and M. Lochbaum, "The Coach–Athlete–Parent Relationship: The Importance of the Sex, Sport Type, and Family Composition," *Int. J. Environ. Res. Public Health*, vol. 19, no. 8, p. 4821, Apr. 2022, doi: 10.3390/ijerph19084821.