Jurnal Mandiri IT
Vol. 15 No. 1, July (2026), pp. 27-36

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Machine learning-based sports preference classification using demographic and behavioral factors

Hendra Apriawan[1], Suhendro Yusuf Irianto[2]
[1,2]Magister of Informatics, Institut Bisnis dan Informatika Darmajaya, Indonesia

A R T I C L E I N F O

Article history:

Received May 17, 2026
Revised May 25, 2026
Accepted Jun 6, 2026

Keywords:

Behavioral Factors;
Decision Tree;
Demographic Factors;
Machine Learning;
Naive Bayes;
Sports Preference Classification.

ABSTRACT

Sports preference is influenced by demographic and behavioral factors, making data-driven classification important for supporting personalized physical activity programs and public health decision-making. However, previous studies have mostly relied on descriptive analysis and have rarely integrated demographic and behavioral variables into a predictive machine learning framework. This study aims to classify community sports preferences using the Naive Bayes algorithm and compare its performance with the Decision Tree model. A quantitative data mining approach was applied using questionnaire data collected from 286 respondents selected from a population of 1,000 individuals using the Slovin formula with a 5% margin of error. The dataset included demographic attributes, such as age, gender, location, occupation, and socioeconomic status, as well as behavioral attributes, including exercise frequency, exercise place, motivation, barriers, and activity preference. Data preprocessing involved data cleaning, attribute transformation, binary label construction, and data leakage removal. Model evaluation was conducted using 10-fold stratified cross-validation with accuracy, precision, recall, F1-score, and Cohen's kappa. The results show that Naive Bayes outperformed Decision Tree, achieving 63.63% accuracy, 57.84% precision, 56.14% recall, 56.97% F1-score, and 0.132 kappa. These findings indicate that Naive Bayes provides moderate but better predictive performance and can serve as an initial baseline for data-driven sports recommendation systems, although further model development is needed to improve reliability.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Hendra Apriawan,
Magister of Informatics,
Institut Bisnis dan Informatika Darmajaya,
Pagar Alam St, No. 93, Gedong Meneng, Bandar Lampung City, 35145, Indonesia
Email: hendralpmpbkl@gmail.com

1. INTRODUCTION

Physical activity and sports participation are important parts of public health because they support physical fitness, mental well-being, and the prevention of non-communicable diseases (Martín-Rodríguez et al., 2024). However, insufficient physical activity remains a global concern because it increases the risk of cardiovascular disease, stroke, cancer, and diabetes (Katzmarzyk et al., 2022). In 2022, around 31% of adults aged 18 years and older were reported to be insufficiently active, showing that physical inactivity is still a major challenge for health promotion and disease prevention (Strain et al., 2024). Therefore, understanding how people choose and maintain sports activities is necessary to support targeted physical activity programs, personalized recommendations, and public health decision-making.

Sports preference is shaped by many interconnected factors, including demographic, behavioral, social, and environmental characteristics (Avsar & Kizilaslan, 2025). Demographic

factors such as age, gender, occupation, location, and socioeconomic status may influence access to sports facilities, available time, and the type of physical activity chosen by individuals (Tsartsapakis et al., 2026). At the same time, behavioral factors such as exercise frequency, motivation, preferred exercise place, and perceived barriers may affect whether a person prefers indoor or outdoor sports . Previous studies have also shown that barriers and facilitators of physical activity differ across populations and activity domains, including leisure, transportation, work, education, and household activities (Van Uffelen et al., 2017) For this reason, sports preference needs to be examined through a multidimensional framework that combines demographic and behavioral attributes rather than through a single-factor perspective (Li et al., 2017).

Along with the growth of data-driven research, data mining and machine learning have increasingly been used in behavioral and health-related studies because they can reveal complex patterns from multidimensional datasets. In physical activity research, machine learning has been applied for activity recognition, behavioral profiling, correlate analysis, and prediction tasks. Classification algorithms are relevant in this context because they can group individuals into predefined categories based on observed characteristics (Farrahi et al., 2020). Naive Bayes and Decision Tree are often used as baseline classification models because they are simple, interpretable, and suitable for structured data analysis. Naive Bayes uses a probabilistic classification approach, while Decision Tree produces a rule-based structure that is easier to interpret by researchers and practitioners.

Despite these developments, several gaps remain in previous studies. First, research on sports preference and physical activity participation has often relied on descriptive or statistical approaches, while predictive classification models have received less attention. Second, demographic and behavioral factors are frequently analyzed separately, even though sports preference is likely influenced by the combined effect of both factor groups. Third, the use of Naive Bayes for classifying indoor and outdoor sports preferences is still limited, especially when compared with a rule-based model such as Decision Tree. The main novelty of this research is therefore not merely the use of a classification algorithm, but the integration of demographic and behavioral variables into one predictive framework that specifically classifies sports preference into indoor and outdoor categories. This position distinguishes the present study from previous sports behavior studies that mainly describe participation patterns without testing how far these factors can be transformed into predictive knowledge (Sun et al., 2025).

To address these gaps, this study proposes a machine learning-based classification model to predict community sports preferences using demographic and behavioral factors. The novelty of this study lies in integrating demographic and behavioral attributes into a single classification framework for predicting indoor and outdoor sports preferences. Unlike previous studies that focused mainly on descriptive analysis, this study emphasizes predictive modeling and comparative algorithm evaluation using 10-fold stratified cross-validation and multiple evaluation metrics, including accuracy, precision, recall, F1-score, and Cohen's kappa. The contribution of predictive modeling in this study is that it converts questionnaire-based sports behavior data into class predictions that can become an early analytical basis for data-driven sports recommendation systems. In practical terms, the model can help identify whether individuals are more likely to prefer indoor or outdoor sports, so that recommendations, facility planning, and intervention programs can be better aligned with behavioral barriers, exercise frequency, and demographic characteristics. The objective of this study is to classify community sports preferences using Naive Bayes and compare its performance with Decision Tree. Theoretically, this study contributes to the application of machine learning in sports behavior analytics, while practically, the findings may serve as a baseline for data-driven sports recommendation systems and targeted physical activity programs.

2. RESEARCH METHOD

Research Design

This study employed a quantitative research design using a data mining approach to classify community sports preferences based on demographic and behavioral factors. A supervised machine learning classification framework was applied because the target variable had been predefined into two classes, namely indoor and outdoor sports preference. The classification process was conducted by comparing the performance of Naive Bayes as the primary model and Decision Tree as the benchmark model. The use of classification algorithms was considered

appropriate because the objective of this study was to assign each respondent into a specific sports preference category based on observed attributes (Chen et al., 2021; Zhao et al., 2021).

Data Source and Respondents

The dataset used in this study was obtained from primary data collected through a structured questionnaire. The population consisted of 1,000 individuals representing community members with diverse demographic and behavioral characteristics. A probability-based sampling approach was applied to obtain a representative sample from the population. The sample size was determined using the Slovin formula, as shown in Equation (1). The use of a 5% margin of error indicates that the sample was designed to tolerate a relatively small sampling deviation from the population estimate. With 286 respondents, the dataset can be considered adequate for representing the target population in an exploratory classification study. However, this margin of error also implies that the findings should be interpreted as population estimates within a defined tolerance range rather than as absolute representations of all community sports preferences. Therefore, the model results are suitable for identifying general predictive tendencies, while broader generalization requires larger and more diverse samples.

$$n = \frac{N}{1+N(e)^2} \quad (1)$$

Given that the population size was N = 1000 and the margin of error was e = 0.05, the sample size was calculated as follows:

$$n = 285.71 \approx 286 \quad (2)$$

Research Variables

The variables used in this study consisted of demographic factors, behavioral factors, and the target variable. Demographic factors were used to represent the general characteristics of respondents, while behavioral factors were used to describe exercise-related habits, motivations, and barriers. The target variable was constructed by grouping sports preferences into two categories: indoor and outdoor. Sports preference was treated as a binary categorical label because the main analytical objective was to distinguish two broad environments of physical activity participation rather than to predict specific sport types. This simplification was also methodologically useful because it reduced excessive label fragmentation, made the supervised classification task more stable for a dataset of 286 respondents, and allowed Naive Bayes and Decision Tree to be compared using consistent binary classification metrics.

Table 1. Research variables used in the classification model

Category	Attribute	Description	Data Type	Role
Demographic factor	Age	Respondent age group	Categorical	Predictor
Demographic factor	Gender	Respondent gender	Categorical	Predictor
Demographic factor	Location	Respondent residential area	Categorical	Predictor
Demographic factor	Occupation	Respondent occupational category	Categorical	Predictor
Demographic factor	Socioeconomic status	Respondent economic status	Categorical	Predictor
Behavioral factor	Exercise frequency	Frequency of exercise activity	Categorical	Predictor
Behavioral factor	Exercise place	Preferred place for exercise	Categorical	Predictor
Behavioral factor	Motivation	Main motivation for exercising	Categorical	Predictor
Behavioral factor	Barrier	Main barrier to exercise participation	Categorical	Predictor
Target variable	Sports preference	Indoor or outdoor sports preference	Binary categorical	Label

The classification label was defined as follows:

$$Y = \begin{cases} Indoor, \text{ if the preferred sport was conducted in an indoor setting} \\ Outdoor, \text{ if the preferred sport was conducted in an outdoor setting} \end{cases} \quad (3)$$

Data Preprocessing

Data preprocessing was performed to improve data quality and ensure that the dataset could be processed appropriately by the classification algorithms. This stage was considered important because the quality of input data can directly affect the performance and reliability of machine learning models (Cho et al., 2022; Mohammed et al., 2025).

The preprocessing stage consisted of four main procedures: data cleaning, attribute transformation, label construction, and data leakage removal. Data cleaning was conducted by checking incomplete, duplicate, and inconsistent records. Attribute transformation was performed

to ensure that categorical variables were represented in a format that could be processed by the classification models. Label construction was carried out by converting various types of sports activities into two target classes, namely indoor and outdoor. Data leakage removal was applied by excluding attributes that could directly reveal the target class, so that model evaluation could reflect a more realistic prediction process.

Table 2. Data Preprocessing procedures

Preprocessing Stage	Description	Output
Data cleaning	Incomplete, duplicate, and inconsistent records were checked and removed when necessary.	Cleaned dataset
Attribute transformation	Categorical attributes were transformed into a model-compatible format.	Structured categorical dataset
Label construction	Sports activities were grouped into indoor and outdoor categories.	Binary target label
Data leakage removal	Attributes that could directly reveal the target class were removed.	Valid predictor set
Role assignment	The sports preference attribute was assigned as the label, while other variables were assigned as predictors.	Final dataset for classification

Classification Algorithms

Two classification algorithms were used in this study: Naive Bayes and Decision Tree. Naive Bayes was used as the primary model, while Decision Tree was used as the comparison model.

a. Naive Bayes, is a probabilistic classification algorithm based on Bayes' theorem. It assumes that each predictor contributes independently to the probability of a class, although this assumption may not always be fully satisfied in real-world data. Despite this limitation, Naive Bayes has been widely used because of its computational efficiency and effectiveness in handling structured and categorical datasets (Jadhav & Channe, 2013; Zhang, 2016). Bayes' theorem is expressed in Equation (4).

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (4)$$

For a set of predictors $X=(x_1,x_2,...,x_n)$, the Naive Bayes classification rule is written as follows:

$$\hat{C} = \arg\max_{C} P(C) \prod_{i=1}^{n} P(x_i \mid C) \quad (5)$$

The class with the highest posterior probability was selected as the predicted sports preference category.

b. Decision Tree, is a rule-based classification algorithm that represents decision rules in a tree structure. The model divides the dataset into branches based on attribute values and produces a final class prediction at the leaf nodes. This algorithm was used as a benchmark model because it is interpretable and commonly applied in classification tasks involving structured data (Mienye & Jere, 2024).

In a decision tree, attribute selection is generally performed using information-based criteria such as entropy and information gain. Entropy is calculated using Equation (6).

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (6)$$

Information gain is calculated using Equation (7).

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (7)$$

The attribute with the highest information gain was selected to split the data at each node.

Experimental Setup

The classification experiment was conducted using RapidMiner. The experimental workflow consisted of data retrieval, attribute generation, attribute selection, role assignment, model training, model testing, and performance evaluation. A 10-fold stratified cross-validation strategy was used to evaluate the models. Stratified cross-validation was applied to preserve the proportion of indoor and outdoor classes in each fold, thereby producing a more stable and representative evaluation result.

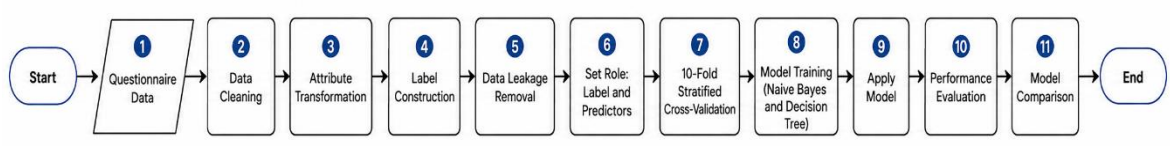

Figure 1. Proposed research workflow

Table 3. Experimental configuration

Component	Configuration
Software	RapidMiner
Dataset type	Primary questionnaire data
Number of respondents	286
Learning task	Binary classification
Target classes	Indoor and Outdoor
Main algorithm	Naive Bayes
Benchmark algorithm	Decision Tree
Validation method	10-fold stratified cross-validation
Evaluation metrics	Accuracy, precision, recall, F1-score, and Cohen's kappa

Model Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and Cohen's kappa. These metrics were selected to provide a more comprehensive evaluation of classification performance, especially because accuracy alone may not sufficiently represent model quality when class distribution is imbalanced. The confusion matrix used in this study is shown in Table 4.

Table 4. Confusion matrix structure

Actual Class / Predicted Class	Predicted Indoor	Predicted Outdoor
Actual Indoor	True Positive (TP)	False Negative (FN)
Actual Outdoor	False Positive (FP)	True Negative (TN)

Accuracy measures the proportion of correctly classified instances among all instances and is calculated using Equation (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive and is calculated using Equation (9).

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall measures the proportion of correctly predicted positive instances among all actual positive instances and is calculated using Equation (10).

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1-score represents the harmonic mean of precision and recall and is calculated using Equation (11).

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

Cohen's kappa was used to measure the agreement between predicted and actual classes while considering agreement occurring by chance. It is calculated using Equation (12).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (12)$$

The overall methodological flow of this study began with questionnaire-based data collection and continued with sample size determination, data preprocessing, label construction, and classification modeling. The final dataset was evaluated using 10-fold stratified cross-validation to compare the predictive performance of Naive Bayes and Decision Tree. The model with the best performance was determined based on the comparison of accuracy, precision, recall, F1-score, and Cohen's kappa.

3. RESULTS AND DISCUSSIONS

Model Performance Evaluation

The performance of the classification models was evaluated using 10-fold stratified cross-validation to obtain a more stable and representative estimation of model performance. Two

classification algorithms were compared in this study, namely Naive Bayes as the primary classifier and Decision Tree as the benchmark classifier. The evaluation was conducted using five performance metrics: accuracy, precision, recall, F1-score, and Cohen's kappa. These metrics were selected because classification performance should not be evaluated only from accuracy, especially when the distribution of classes may affect the predictive results.

Table 5. Performance Comparison of Naive Bayes and Decision Tree Models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Kappa
Naive Bayes	63.63	57.84	56.14	56.97	0.132
Decision Tree	55.21	53.35	53.59	53.47	0.067

As shown in Table 5, the Naive Bayes model achieved the highest performance across all evaluation metrics. The accuracy of Naive Bayes reached 63.63%, which was higher than the Decision Tree accuracy of 55.21%. This result indicates that Naive Bayes provided better overall classification performance in predicting sports preferences based on demographic and behavioral factors. However, the accuracy value also suggests that the model performance remained moderate rather than high, indicating that sports preference classification is a relatively challenging task. From a model reliability perspective, an accuracy of 63.63% means that the model can correctly classify more than half of the respondents and is useful as an initial baseline, but it should not yet be considered sufficiently reliable for autonomous decision-making. In practical implementation, the model should be used as a decision-support tool that assists recommendation design rather than as the only basis for determining individual sports programs.

The precision value of Naive Bayes was 57.84%, while Decision Tree achieved 53.35%. This result indicates that Naive Bayes was more reliable in producing correct positive predictions than Decision Tree. In terms of recall, Naive Bayes obtained 56.14%, which was slightly higher than the Decision Tree recall of 53.59%. This finding indicates that Naive Bayes was also more capable of identifying actual class instances. The F1-score of Naive Bayes reached 56.97%, whereas Decision Tree obtained 53.47%, showing that Naive Bayes produced a better balance between precision and recall.

The kappa value of Naive Bayes was 0.132, while Decision Tree achieved a kappa value of 0.067. Cohen's kappa is used to measure agreement between predicted and actual classes by considering agreement that may occur by chance. Although Naive Bayes outperformed Decision Tree, the kappa value indicates that the level of agreement was still relatively low. This means that the classification model was able to detect certain patterns in the dataset, but the predictive consistency between actual and predicted classes remained limited.

Visualization of Model Performance

The comparison of model performance is presented in Figure 2. The visualization shows that Naive Bayes consistently outperformed Decision Tree in all evaluation metrics, including accuracy, precision, recall, and F1-score.

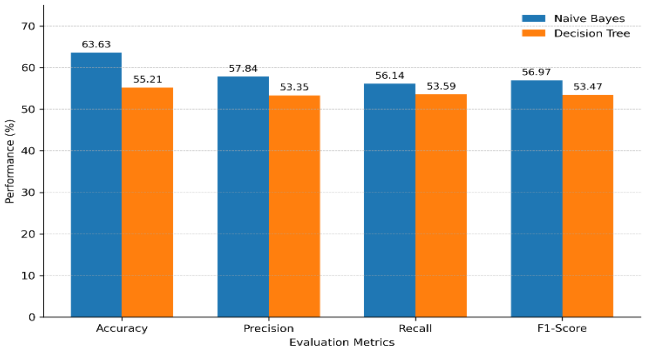

Figure 2. Performance comparison of naive bayes and decision tree models

The difference between Naive Bayes and Decision Tree was most visible in accuracy, where Naive Bayes achieved an improvement of 8.42 percentage points over Decision Tree. This improvement suggests that the probabilistic classification approach was more suitable for the dataset used in this study. Naive Bayes may perform effectively on structured datasets because it

estimates class probabilities based on the distribution of predictor attributes. In contrast, Decision Tree relies on hierarchical rule-based splitting, which may be less effective when the relationship between predictors and target classes is not clearly separable.

Confusion Matrix Analysis

The classification performance of the Naive Bayes model was further examined using a confusion matrix. The confusion matrix provides a more detailed view of correct and incorrect predictions for each class [10].

Table 6. Confusion Matrix Structure for the Naive Bayes Model

Actual Class / Predicted Class	Predicted Indoor	Predicted Outdoor
Actual Indoor	20	55
Actual Outdoor	18	193

Based on the confusion matrix analysis, the number of correct predictions for the outdoor class was higher than that for the indoor class. This result indicates that the Naive Bayes model was more effective in identifying outdoor sports preferences than indoor sports preferences. However, misclassification was still found in both classes, particularly in the indoor class. This condition may indicate that the predictor variables used in the model were not sufficient to fully distinguish indoor and outdoor sports preferences.

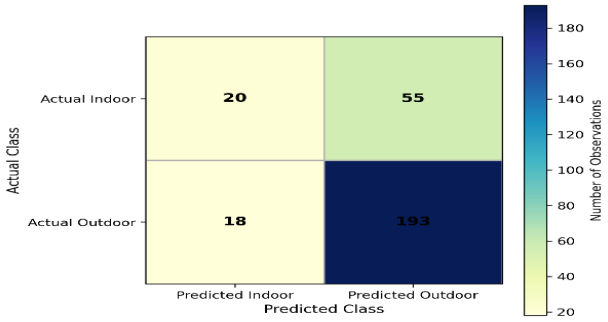

Figure 3. Confusion matrix of the naive bayes model

The higher prediction tendency toward the outdoor class may also be influenced by class distribution, respondent behavior, or overlapping characteristics between indoor and outdoor sports participants. For example, individuals with similar demographic profiles may still have different sports preferences depending on motivation, accessibility, time availability, and perceived barriers. Therefore, future model development should consider additional predictors, such as psychological factors, environmental accessibility, facility availability, and lifestyle patterns, to improve classification performance.

Feature Importance Analysis

Feature importance analysis was conducted using Information Gain to identify the contribution of each attribute to the classification process. The results showed that behavioral attributes had stronger contributions than several demographic attributes.

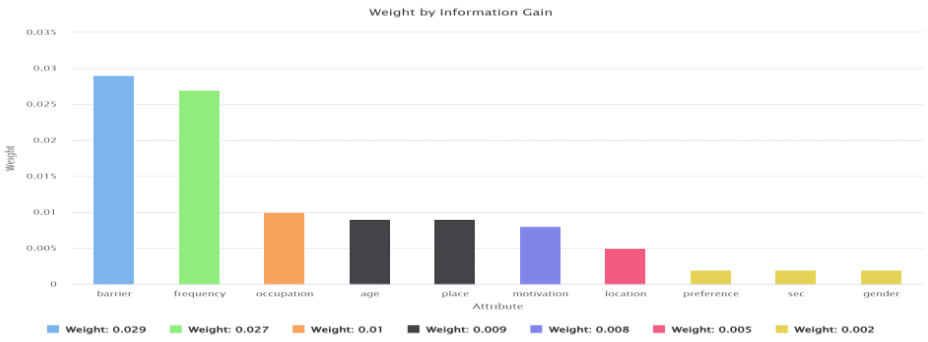

Figure 4. Feature importance based on information gain

The feature importance results indicate that barrier had the highest contribution to the classification process, with a weight of 0.029. Exercise frequency was the second most influential attribute, with a weight of 0.027. These results suggest that behavioral factors were more relevant in distinguishing sports preferences than several demographic factors. This finding is consistent with the view that physical activity participation is strongly influenced by behavioral barriers, motivation, and habitual activity patterns.

The relatively low contribution of demographic attributes, such as gender and socioeconomic status, suggests that these factors were not dominant predictors in the dataset used in this study. However, this result should not be interpreted as evidence that demographic factors are irrelevant in all contexts. The influence of demographic characteristics may vary depending on population structure, cultural background, accessibility to sports facilities, and social environment. Therefore, the lower contribution of demographic variables in this study may reflect the specific characteristics of the respondents rather than a universal pattern.

Discussion of Algorithm Performance

The experimental results showed that Naive Bayes performed better than Decision Tree in classifying sports preferences based on demographic and behavioral factors. This result indicates that the probabilistic learning mechanism of Naive Bayes was more suitable for the dataset characteristics, particularly because the predictors consisted mostly of categorical attributes such as age group, gender, occupation, exercise frequency, motivation, and exercise barriers. Several factors may explain why Naive Bayes outperformed Decision Tree. First, Naive Bayes estimates class membership through posterior probabilities, making it relatively robust for small and categorical datasets. Second, the model does not require a complex hierarchical splitting structure, so it is less vulnerable to unstable branches when the dataset contains overlapping respondent characteristics. Third, sports preference patterns in this dataset may be distributed across several weak predictors rather than concentrated in a few dominant variables, which makes a probabilistic aggregation approach more effective than a rule-based tree structure. Naive Bayes estimates the posterior probability of each class by considering the contribution of each predictor attribute, and the class with the highest probability is selected as the final prediction (Rizki et al., 2021). Although Naive Bayes assumes conditional independence among variables, this assumption does not necessarily reduce its practical effectiveness in real-world classification tasks, especially when the dataset is structured and relatively small (Wickramasinghe & Kalutarage, 2021)(Rizki et al., 2021)(Wickramasinghe & Kalutarage, 2021).

The superiority of Naive Bayes can also be explained by the behavioral nature of sports preference. Sports preference is not always formed through a simple linear rule, but may emerge from the interaction of demographic characteristics, exercise habits, motivation, perceived barriers, and access to facilities. In this context, a probabilistic model may be more flexible than a rule-based model because class membership is estimated through probability distributions rather than fixed hierarchical decision paths. In contrast, Decision Tree depends on attribute-splitting mechanisms using criteria such as entropy, information gain, or gain ratio (Yilmaz & Demirhan, 2023). This approach may become less effective when indoor and outdoor sports preferences contain overlapping patterns, where respondents from different classes share similar demographic profiles, exercise frequencies, motivations, or barriers.

The performance gap between the two models was most evident in accuracy, where Naive Bayes achieved 63.63%, while Decision Tree achieved 55.21%, indicating an improvement of 8.42 percentage points. Naive Bayes also produced higher precision, recall, and F1-score, suggesting that it was more balanced in identifying class instances and reducing classification errors. The higher F1-score indicates a better trade-off between precision and recall, which is important when model performance cannot be evaluated by accuracy alone (Islam et al., 2022). However, the accuracy of 63.63% and F1-score of 56.97% still indicate moderate predictive performance. This suggests that the model was able to capture useful patterns, but its reliability remains limited for practical implementation. The Cohen's kappa values further support this interpretation, where Naive Bayes obtained 0.132 and Decision Tree obtained 0.067. Since Cohen's kappa measures agreement between predicted and actual classes while accounting for chance agreement, the low values indicate that both models still had limited classification agreement (Reddy & Chittineni, 2021).

The feature importance results showed that behavioral attributes, particularly exercise barriers and exercise frequency, contributed more strongly than several demographic attributes. This finding suggests that sports preference was more closely associated with behavioral conditions than demographic characteristics alone. In other words, how often individuals exercise and what barriers they experience may be more informative for predicting sports preference than general demographic variables such as gender or socioeconomic status. This result is consistent with studies emphasizing that physical activity participation is strongly influenced by perceived barriers, motivation, and behavioral patterns (Tangirala, 2020). Overall, Naive Bayes can be considered more appropriate than Decision Tree as a baseline model for this study, although future research should use larger datasets, improve class balance, add contextual predictors, and evaluate more advanced algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, and ensemble learning.

4. CONCLUSION

This study developed a machine learning-based classification model to predict community sports preferences by integrating demographic and behavioral factors into a binary classification framework of indoor and outdoor sports preferences. Using 10-fold stratified cross-validation, Naive Bayes demonstrated better performance than Decision Tree across all evaluation metrics, achieving 63.63% accuracy, 57.84% precision, 56.14% recall, 56.97% F1-score, and 0.132 Cohen's kappa, compared with Decision Tree, which achieved 55.21% accuracy, 53.35% precision, 53.59% recall, 53.47% F1-score, and 0.067 Cohen's kappa. These findings indicate that the probabilistic learning mechanism of Naive Bayes was more suitable for the dataset than the rule-based structure of Decision Tree, although the moderate accuracy and low kappa value suggest that the predictive agreement between actual and predicted classes remained limited. Theoretically, this study contributes to sports behavior analytics by demonstrating that demographic and behavioral attributes can be integrated into a unified machine learning classification framework. The feature importance results further indicate that behavioral factors, particularly exercise barriers and exercise frequency, contributed more strongly than several demographic factors, suggesting that sports preference is closely related to individual exercise habits and constraints. Practically, the findings have implications for the development of personalized physical activity programs because they show that recommendation design should consider behavioral barriers and exercise routines, not only demographic profiles. The model can serve as an early decision-support baseline for identifying whether a person is more likely to prefer indoor or outdoor activities, allowing program providers to design more relevant activity options, facility access strategies, and motivational interventions. However, this study is limited by the relatively small dataset of 286 respondents, the simplification of sports preference into two classes, and the exclusion of psychological, environmental, social, and accessibility-related variables. Future research has opportunities to extend this work by using larger and more diverse datasets, applying advanced algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, XGBoost, and ensemble learning, and comparing their ability to improve accuracy, recall, model stability, and generalization. Further studies may also develop multiclass or recommendation-based models that predict specific sport types and integrate contextual variables such as facility distance, weather, lifestyle patterns, and digital activity records.

REFERENCES

Avsar, F., & Kizilaslan, N. (2025). Life Satisfaction and Healthy Lifestyle Behaviors of Individuals According to Exercise Preferences of Outdoor and Indoor. *Public Health Nursing*, *42*(3), 1261–1271. https://doi.org/10.1111/PHN.13538

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing 2021 2021:1*, *2021*(1), 30-. https://doi.org/10.1186/S13634-021-00742-6

Cho, E., Chang, T. W., & Hwang, G. (2022). Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process. *Electronics 2022, Vol. 11, Page 477*, *11*(3), 477. https://doi.org/10.3390/ELECTRONICS11030477

Farrahi, V., Niemelä, M., Kärmeniemi, M., Puhakka, S., Kangas, M., Korpelainen, R., & Jämsä, T. (2020). Correlates of physical activity behavior in adults: A data mining approach. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1). https://doi.org/10.1186/S12966-020-00996-7

Islam, S. S., Haque, M. S., Miah, M. S. U., Sarwar, T. Bin, & Nugraha, R. (2022). Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.898

Jadhav, S. D., & Channe, H. P. (2013). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR) ISSN*, *5*. www.ijsr.net

Katzmarzyk, P. T., Friedenreich, C., Shiroma, E. J., & Lee, I. M. (2022). Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries. *British Journal of Sports Medicine*, *56*(2), 101–106. https://doi.org/10.1136/BJSPORTS-2020-103640

Li, W., Procter-Gray, E., Churchill, L., Crouter, S. E., Kane, K., Tian, J., Franklin, P. D., Ockene, J. K., & Gurwitz, J. (2017). Gender and Age Differences in Levels, Types and Locations of Physical Activity among Older Adults Living in Car-Dependent Neighborhoods. *The Journal of Frailty & Aging*, *6*(3), 129–135. https://doi.org/10.14283/JFA.2017.15

Martín-Rodríguez, A., Gostian-Ropotin, L. A., Beltrán-Velasco, A. I., Belando-Pedreño, N., Simón, J. A., López-Mora, C., Navarro-Jiménez, E., Tornero-Aguilera, J. F., & Clemente-Suárez, V. J. (2024). Sporting Mind: The Interplay of Physical Activity and Psychological Health. *Sports*, *12*(1). https://doi.org/10.3390/SPORTS12010037

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, *132*, 102549. https://doi.org/10.1016/J.IS.2025.102549

Reddy, G. S., & Chittineni, S. (2021). Entropy based C4.5-SHO algorithm with information gain optimization in data mining. *PeerJ Computer Science*, *7*, 1–22. https://doi.org/10.7717/PEERJ-CS.424

Rizki, M., Arhami, M., & Huzeni, H. (2021). Perbaikan Algoritma Naive Bayes Classifier Menggunakan Teknik Laplacian Correction. *Jurnal Teknologi*, *21*(1), 39. https://doi.org/10.30811/TEKNOLOGI.V21I1.2209

Strain, T., Flaxman, S., Guthold, R., Semenova, E., Cowan, M., Riley, L. M., Bull, F. C., Stevens, G. A., Raheem, R. A., Agoudavi, K., Anderssen, S. A., Alkhatib, W., Aly, E. A. H., Anjana, R. M., Bauman, A., Bovet, P., Moniz, T. B., Bulotait, G., Caixeta, R., … Zoma, L. R. (2024). National, regional, and global trends in insufficient physical activity among adults from 2000 to 2022: a pooled analysis of 507 population-based surveys with 5·7 million participants. *The Lancet Global Health*, *12*(8), e1232–e1243. https://doi.org/10.1016/S2214-109X(24)00150-5

Sun, Y., Zhao, Y., & Yang, J. (2025). The impact of sports preferences on physical activity participation among college students: the mediating role of sports achievement emotions and exercise motivation. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/FPSYG.2025.1565998

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, (2), 612–619. https://doi.org/10.14569/IJACSA.2020.0110277

Tsartsapakis, I., Zafeiroudi, A., Trigonis, I., & Kouthouris, C. (2026). Exercise Participation Among Physically Active Adults: A Multidimensional Analysis of Demographic, Anthropometric, Personality, and Behavioral Factors. *International Journal of Environmental Research and Public Health*, *23*(2). https://doi.org/10.3390/IJERPH23020209

Van Uffelen, J. G. Z., Khan, A., & Burton, N. W. (2017). Gender differences in physical activity motivators and context preferences: A population-based study in people in their sixties. *BMC Public Health*, *17*(1). https://doi.org/10.1186/S12889-017-4540-0

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/S00500-020-05297-6

Yilmaz, A. E., & Demirhan, H. (2023). Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing*, *134*. https://doi.org/10.1016/J.ASOC.2023.110020

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, *4*(12), 241–241. https://doi.org/10.21037/ATM.2016.03.38

Zhao, L., Lee, S., & Jeong, S. P. (2021). Decision Tree Application to Classification Problems with Boosting Algorithm. *Electronics 2021, Vol. 10, Page 1903*, *10*(16), 1903. https://doi.org/10.3390/ELECTRONICS10161903