JURNAL INFORMATIKA
PROXIES

Vol.10, No.1, July, 2026

Data-Driven Athlete Performance Grouping A Comparative Study of K-Means and K-Medoids With MKNN Accuracy Assessment

Adonis Max Balinda[1], Shinta Estri Wahyuningrum[2]

[1] Teknik Informatika Fakultas Ilmu Komputer, Soegijapranata Catholic University, adonismaxb@gmail.com
[2] Teknik Informatika Fakultas Ilmu Komputer, Soegijapranata Catholic University, shinta@unika.ac.id

Corresponding Author Email: shinta@unika.ac.id

https://doi.org/10.18280/proxies.v10i1.15427

Received: 2026-04-29
Revised: 2026-07-20
Accepted: 2026-07-20
Available online: 2026-07-23

Keywords:
Athlete Fitness 1, Classification 2, Clustering 3, K-Means 4, K-Medoids 5, Machine Learning 6, MKNN 7, Pseudo-Labeling 8

ABSTRACT

The aim of this study is to introduce a hybrid machine learning solution to classify athletes' fitness levels using unlabelled physical activities using clustering and classification techniques. This work focuses on overcoming the challenge posed by datasets of fitness levels which lack defined fitness levels labels required for direct classification processes. The selected dataset in this paper is the Workout & Fitness Tracker Dataset which consists of about 10,000 samples with selected features such as Steps Taken, Calories Burned, Distance, Workout Duration, and Daily Calories Intake. In the first step, K-Means and K-Medoids clustering techniques were implemented to form clusters in a certain number of cluster formations (K=3, K=4, K=5, K=6), then generating the pseudo-labels. Clustering effectiveness was analysed by evaluating the Silhouette Scores. Finally, in the second step, the generated pseudo-labels were fed into the Modified K-Nearest Neighbour (MKNN) classification algorithm. The experiments have demonstrated that the K-Means clustering technique provided the maximum Silhouette Score of 0.1476 with the formation of K=6. On the other hand, K-Medoids clustering provided K=3 and resulted in 94.80% accuracy and 94.79% F1-Score. It should be mentioned that clustering efficiency may not guarantee the highest classification accuracy.

1. INTRODUCTION

1.1. Background

The rapid development of wearable devices and fitness tracking applications has enabled the collection of large amounts of physical activity data, such as steps taken, calories burned, distance travelled, and workout duration. These data provide important information for evaluating athlete fitness and physical performance. However, one major challenge in fitness data analysis is the absence of predefined fitness labels, which limits the direct application of supervised learning methods for classification tasks.

Classification methods such as K-Nearest Neighbour (KNN) and Modified K-Nearest Neighbour (MKNN) have been widely used in various prediction and classification tasks because of their simplicity and effectiveness in handling numerical data. In sports-related analysis, classification methods can help identify athlete conditions, performance patterns, and fitness levels based on physical activity data.

However, supervised learning methods require labelled datasets, while many real-world fitness datasets are unlabelled. In this condition, unsupervised learning methods can be used to discover hidden patterns and generate data groups automatically. Clustering methods such as K-Means and K-Medoids are commonly used to group similar data objects based on feature similarity. The resulting cluster labels can then be used as pseudo-labels for the classification process.

Based on this problem, this study proposes a hybrid machine learning approach by combining clustering and classification methods to classify athlete fitness levels from unlabelled physical activity data. The clustering stage is used to generate pseudo-labels, while the classification stage is used to predict athlete fitness levels based on the generated labels.

1.2. Scope

This study focuses on the analysis of athlete fitness data using the Workout & Fitness Tracker Dataset obtained from Kaggle. The dataset consists of approximately 10,000 physical activity records. The selected numerical features used in this study include Steps Taken, Calories Burned, Distance (km), Workout Duration (mins), and Daily Calories Intake.

This study applies two clustering algorithms, namely K-Means and K-Medoids, with cluster configurations of K = 3, K = 4, K = 5, and K = 6. The clustering results are evaluated using the Silhouette Score and then used as pseudo-labels in the classification stage. The classification process uses the Modified K-Nearest Neighbour (MKNN) algorithm and is evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

1.3. Objective

The objectives of this study are to analyse the clustering performance of K-Means and K-Medoids in grouping athlete fitness data, to evaluate the effectiveness of pseudo-label generation for fitness classification, and to measure the performance of the Modified K-Nearest Neighbour (MKNN) algorithm in classifying athlete fitness levels based on clustering-generated labels. In addition, this study aims to analyse the effect of different cluster configurations on clustering quality and classification performance.

2. LITERATURE STUDY

Previous studies have shown that machine learning methods, especially classification algorithms, have been widely applied in sports performance analysis and prediction. A study on athlete performance prediction using machine learning showed that predictive models can effectively analyse athlete performance patterns based on multiple features, proving the potential of machine learning in sports analytics [8].

The K-Nearest Neighbour (KNN) algorithm has also been widely used in sports-related classification tasks. Previous research on badminton athlete potential classification showed that KNN can classify athlete potential based on performance indicators using distance-based similarity calculations [2]. Similarly, athlete anxiety classification before competition also showed that KNN can achieve high classification performance in sports datasets [3].

Modified K-Nearest Neighbour (MKNN) is a development of the traditional KNN algorithm by adding a weighting mechanism to improve classification accuracy. Previous studies showed that MKNN can improve prediction accuracy in classification problems because it considers both distance and data validity in the classification process [1][9].

In the context of unlabelled data, clustering methods are commonly used to identify hidden patterns and group similar data objects. K-Means is widely used because of its efficiency in grouping data into clusters based on centroid distance, while K-Medoids provides better robustness against outliers by using actual data objects as cluster centres [12][13]. These clustering methods are useful for generating pseudo-labels in datasets that do not have predefined class labels.

Based on previous studies, most classification-based research relies on supervised learning methods that require labelled datasets, while clustering-based studies are generally limited to grouping purposes only. In athlete fitness analysis, there is still limited research that combines clustering methods and MKNN classification in a pseudo-labelling framework. Therefore, this study addresses this research gap by implementing K-Means and K-Medoids clustering to generate pseudo-labels, followed by MKNN classification, while analysing the effect of different cluster configurations (K = 3, K = 4, K = 5, and K = 6) on clustering quality and classification performance.

3. RESEARCH METODOLOGY

This study uses a hybrid machine learning approach consisting of two main stages: clustering and classification. This approach is applied because the dataset used in this study does not contain predefined fitness level labels, making direct supervised classification difficult to implement. Therefore, clustering is first used to generate pseudo-labels, which are then used as labels in the classification stage.

The overall research workflow is illustrated in Figure 1. The workflow describes the sequence of processes starting from dataset collection, preprocessing, clustering, pseudo-label generation, classification, and performance evaluation.

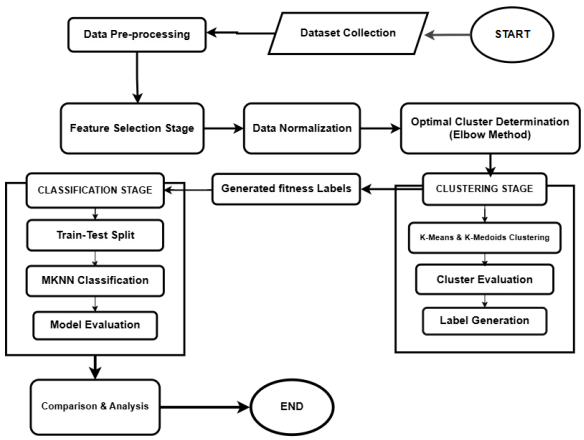

Figure 1 Workflow of the Proposed Fitness Classification System

Based on Figure 1, the research process begins with dataset collection followed by data preprocessing, feature selection, and normalization. The clustering stage is then performed using K-Means and K-Medoids to generate pseudo-labels based on data similarity patterns. After that, the generated labels are used in the classification stage using the Modified K-Nearest Neighbour

(MKNN) algorithm. Finally, the model performance is evaluated using Silhouette Score, accuracy, precision, recall, and F1-score to measure the effectiveness of the proposed hybrid approach.

Clustering Stage

The clustering stage aims to group athlete physical activity data based on similarities in selected features. Since the dataset is unlabelled, clustering is used as an unsupervised learning approach to form data groups automatically. The output of this stage is cluster assignment results, which are then used as pseudo-labels for classification.

Data Collection

The dataset used in this study is the Workout & Fitness Tracker Dataset consisting of approximately 10,000 physical activity records. The dataset contains several attributes related to workout performance and physiological activity. In this study, five numerical features were selected for analysis: Steps Taken, Calories Burned, Distance (km), Workout Duration (mins), and Daily Calories Intake.

Data Cleaning

The data cleaning stage was performed to ensure data quality before clustering. This process included checking for missing values, invalid data, and data type consistency. Outlier analysis was also conducted using the Z-score method to observe the impact of extreme values on cluster formation.

Feature Selection

Feature selection was conducted to identify the most relevant attributes contributing to athlete fitness patterns. This process aimed to reduce irrelevant features and improve clustering quality. The selected features were used consistently in both clustering and classification stages.

Normalization

Normalization was performed using the Min-Max Scaling method to transform all feature values into the range of 0 to 1. This process ensures that all features contribute equally to distance calculations in clustering and classification.

Classification Stage

After the clustering process, the resulting cluster labels were used as pseudo-labels for classification. The dataset was divided into training data (80%) and testing data (20%). The classification process was performed using the Modified K-Nearest Neighbour (MKNN) algorithm to predict athlete fitness levels based on the pseudo-labelled data.

Performance Evaluation

The clustering results were evaluated using the Silhouette Score to measure cluster compactness and separation. The classification results were evaluated using confusion matrix, accuracy, precision, recall, and F1-score. These evaluation metrics were used to measure the effectiveness of the proposed hybrid model in classifying athlete fitness levels.

4. RESULT AND DISCUSSION

4.1 Clustering Results

The clustering process was performed using K-Means and K-Medoids algorithms with cluster configurations of K = 3, K = 4, K = 5, and K = 6. The purpose of this stage was to generate pseudo-labels from unlabelled athlete fitness data based on similarity patterns in physical activity attributes.

To determine additional cluster configurations beyond the initial K = 3 and K = 4, the Elbow Method was applied by analysing the inertia trend across different cluster numbers. Based on the elbow curve pattern, K = 5 and K = 6 were selected as additional cluster candidates for further evaluation.

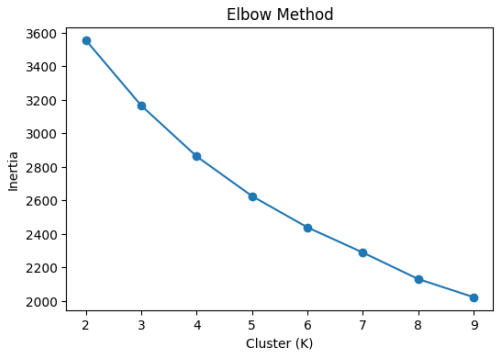

Figure 2 Elbow Method Graph for Determining the Optimal Number of Clusters

Based on Figure 1, the inertia values decreased as the number of clusters increased, indicating that the data points became more compact within each cluster. However, the decrease became less significant after K = 4, forming an elbow pattern. This pattern indicates that K = 5 and K = 6 could be considered additional cluster configurations for further analysis.

The clustering quality was then evaluated using the Silhouette Score. The results showed that K-Means consistently produced slightly higher Silhouette Scores than K-Medoids across all cluster configurations. The highest Silhouette Score was obtained by K-Means with K = 6 (0.1476), while the lowest score was obtained by K-Medoids with K = 3 (0.1304). Although the Silhouette Score values were relatively low, the results still

4.2 Classification Results

After the clustering process was completed, the generated cluster labels were used as pseudo-labels for the classification stage. The Modified K-Nearest Neighbour (MKNN) algorithm was then applied to classify athlete fitness levels using the pseudo-labelled dataset.

The classification performance was evaluated using accuracy, precision, recall, and F1-score. These evaluation metrics were used to measure the predictive consistency of the MKNN model across all clustering scenarios.

Table 1 Comparison of Clustering Quality and Classification Performance

Method	K	Silhouette score	Accuracy	Precision	Recall	F1-Score
K-Means	3	0.1336	0.9455	0.9455	0.9455	0.9455
K-Means	4	0.1454	0.934	0.9342	0.934	0.9339
K-Means	5	0.1446	0.9185	0.9189	0.9185	0.9185
K-Means	6	**0.1476**	0.9155	0.9158	0.9155	0.9154
K-Medoids	3	0.1304	**0.9480**	**0.9481**	**0.9480**	**0.9479**
K-Medoids	4	0.1420	0.9275	0.9276	0.9275	0.9275
K-Medoids	5	0.1411	0.9185	0.9187	0.9185	0.9184
K-Medoids	6	0.1409	0.9105	0.9106	0.9105	0.9104

Based on Table 1, the classification results showed high performance across all experimental scenarios, with all evaluation metrics consistently above 0.91. This indicates that the pseudo-labels generated from the clustering stage were sufficiently representative for supervised classification.

The best classification performance was achieved by K-Medoids with K = 3, resulting in an accuracy of 0.9480 and an F1-score of 0.9479. In contrast, the highest Silhouette Score was obtained by K-Means with K = 6 (0.1476), but this configuration did not produce the best classification performance. This finding indicates that higher clustering quality does not always correspond to better classification performance.

4.3 Discussion

The experimental results demonstrate that clustering quality and classification performance do not always show a directly proportional relationship. Although K-Means with K = 6 produced the highest Silhouette Score (0.1476), the best classification performance was obtained by K-Medoids with K = 3, achieving an accuracy of 0.9480 and an F1-score of 0.9479. This finding indicates that internal clustering compactness is not the only factor influencing classification performance, especially when clustering results are used as pseudo-labels for supervised learning.

The use of pseudo-labelling in this study shows that the quality of generated labels plays a critical role in determining classification effectiveness. In the clustering-classification hybrid model, pseudo-label consistency becomes an important factor because the classification model learns directly from the generated labels. If cluster boundaries are too fragmented, the classifier may experience difficulties in learning stable decision boundaries. This explains why configurations with larger cluster numbers (K = 5 and K = 6) produced lower classification performance compared to smaller cluster configurations.

The superior performance of K-Medoids with K = 3 can be explained by its medoid-based cluster centre selection mechanism. Unlike K-Means, which uses centroids and is sensitive to extreme values, K-Medoids uses actual data points as cluster representatives, making it more robust to outliers and better at preserving cluster structure. This characteristic contributes to more representative pseudo-label generation and improved classification accuracy.

The classification results obtained using MKNN also show consistently high performance across all cluster configurations, with accuracy values above 0.91. This confirms that MKNN is effective in handling pseudo-labelled datasets and can maintain classification stability despite variations in cluster structures. The weighted neighbour mechanism in MKNN allows closer data points to contribute more significantly to classification decisions, improving prediction reliability.

These findings are consistent with previous studies that showed K-Nearest Neighbour-based methods can achieve strong classification performance when the class representation is sufficiently structured and meaningful [1], [4], [6]. In addition, previous clustering studies also reported that K-Medoids tends to provide better robustness in handling noisy data compared to centroid-based clustering methods such as K-Means [12].

Overall, the proposed hybrid approach successfully demonstrates that combining clustering and MKNN can be used as an effective solution for classification problems involving unlabelled athlete fitness datasets. The pseudo-labelling strategy provides an alternative approach for transforming unsupervised data into supervised learning problems, allowing machine learning models to be applied even when explicit labels are unavailable.

5. CONCLUSION

This study proposed a hybrid machine learning approach for athlete fitness classification by combining clustering and classification methods on unlabeled physical activity data. The clustering stage used K-Means and K-Medoids algorithms to generate pseudo-labels, while the classification stage applied the Modified K-Nearest Neighbor (MKNN) algorithm to predict fitness levels.

The experimental results showed that both clustering methods were able to form meaningful cluster structures, although the Silhouette Score values were relatively low, ranging from 0.1304 to 0.1476. K-Means produced slightly better internal clustering quality, while K-Medoids demonstrated better classification performance.

The classification results showed that the MKNN algorithm achieved high performance across all scenarios, with accuracy values above 91%. The best classification performance was obtained using K-Medoids with K = 3, achieving an accuracy of 0.9480 and an F1-score of 0.9479. This finding indicates that simpler cluster configurations can produce more stable and representative pseudo-labels for classification.

Overall, the proposed clustering-based pseudo-labeling approach proved effective for building a fitness classification model from unlabeled athlete data. This approach provides an alternative solution for classification tasks where labeled data are not available and can support data-driven fitness analysis for athletes.

ACKNOWLEDGMENT

I have been given a myriad of support, advice, and help throughout the process of writing this document. I would like to take this opportunity to show my deepest appreciation to my supervisor, Dr.Shinta Estri Wahyuningrum, S.Si., M.Cs., for her guidance, direction, and support in completing this thesis.

I would like to thank my family, especially my father, mother, and my sibling, for their endless love, prayers, and support throughout my stay at Soegijapranata Catholic University. I would like to thank my friends, in church and college, for their endless support and for being a venue for me to share my thoughts throughout the process of completing this thesis.

REFERENCES

[1] Abrianto P, Hidayat N, Dewi RK. Implementasi Metode Modified K - Nearest Neighbor (MK-NN) Untuk Klasifikasi Cedera Pada Pemain Futsal. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8917 (2021).

[2] Adi nur Rohkhim. ALGORITHM K-NN IN CLASSIFICING POTENTIAL AREAS OF MEN SINGLES BADMINTON PLAYERS IN INDONESIA. http://ejournal.unira.ac.id/index.php/insand_comtech/article/view/734 (2020).

[3] Munawaroh S, Rosyidah UA, Yanuarti R. Klasifikasi Tingkat Kecemasan Atlet Sebelum Bertanding Menggunakan Algoritma K–Nearest Neighbor (KNN) Berbasis Website. https://www.bios.sinergis.org/bios/article/view/120 (2024).

[4] Prasetyawan D, Gatra R. Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi. https://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/3129 (2022).

[5] Mestika D, Nazario R, Agung H. ALGORITMA K-NEAREST NEIGHBOUR DALAM MENGUKUR PERFORMANCE PLAYER GAME PUBGM. https://publikasi.mercubuana.ac.id/index.php/jitkom/article/view/10231 (2021).

[6] Ramadhani DH, Jumadi J, Sandi G. Implementasi Algoritma K-Nearest Neighbors (KNN) Untuk Prediksi Gizi Buruk. https://www.researchgate.net/publication/387395092_Implementasi_Algoritma_K-Nearest_Neighbors_KNN_Untuk_Prediksi_Gizi_Buruk (2024).

[7] Dhany HW. Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung. https://ejournal.pelitaindonesia.ac.id/ojs32/index.php/SENATIKA/article/view/1152 (2021).

[8] Szmygin A, Wojtowicz M, Świderska-Chadaj Ż, et al. Prediction of athletes' performance results using machine learning algorithms. In: 2023 24th International Conference on Computational Problems of Electrical Engineering (CPEE). https://www.researchgate.net/publication/374785260_Prediction_of_athletes'_performance_results_using_machine_learning_algorithms (2023).

[9] Ashraf Maghari. Prediction of Student's Performance Using Modified Knn Classifiers. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3704733 (2018).

[10] Riefky M, Pramesti W. Sentiment Analysis of Southeast Asian Games (SEA Games) in Philippines 2019 Based on Opinion of Internet User of Social Media Twitter with K-Nearest Neighbor and Support Vector Machine. https://journal.unhas.ac.id/index.php/jmsk/article/view/9947 (2020).

[11] Wahyuningrum SE, Sulastri A, Hendriks MPH, et al. THE INDONESIAN NEUROPSYCHOLOGICAL TEST BATTERY (INTB): PSYCHOMETRIC PROPERTIES, PRELIMINARY NORMATIVE SCORES, THE UNDERLYING COGNITIVE CONSTRUCTS, AND THE EFFECTS OF AGE AND EDUCATION. http://psjd.icm.edu.pl/psjd/element/bwmeta1.element.ojs-doi-10_5604_01_3001_0016_1339 (2022).

[12] Bahri S, Midyanti DM. Penerapan Metode K-Medoids untuk Pengelompokan Mahasiswa Berpotensi Drop Out. JTIIK 2023; 10: 165–172. https://jtiik.ub.ac.id/index.php/jtiik/article/view/6643 (2023).

[13] Dewi FP, Aryni PS, Umaidah Y. Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran. JISKa 2022; 7: 111–121. https://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/3348 (2022).

[14] Nurhalizah RS, Ardianto R, Purwono P. Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. Jur Ilm Komp & Infor 2024; 4: 61–72. https://jiki.jurnal-id.com/index.php/jiki/article/view/168 (2024).

[15] Workout Fitness dataset, https://www.kaggle.com/datasets/talhaanjum0/workout-fitness-dataset.