Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Physical Activities Recommender System Based on Sequential Data Use K-Mean Clustering

Rizky Haffiyan Roseno[1)], Z. K. A Baizal[2)*], Ramanti Dharayani[3)]
[1,2,3)] School of Computing, Telkom University Bandung, Indonesia
[1)]rihano@student.telkomuniversity.ac.id, [2)]baizal@telkomuniversity.ac.id,
[3)]dharayani@telkomuniversity.ac.id

Submitted : Jan 9, 2024 | **Accepted** : Jan 12, 2024 | **Published** : Jan 17, 2024

Abstract: Physical activities such as Exercise are essential in maintaining health and fitness, especially for those who adopt a healthy lifestyle. Irregularity in doing Exercise can hurt the body and health, especially if it is not done according to one's physical capacity. In the framework of this research, we developed a Recommender System that aims to provide exercise suggestions according to the user's preferences, especially in the categories of cycling, running, walking, and horse riding. The primary considerations of the variables include heart rate (Average Heart Rate) and pace (Speed Rate). This research approach uses the FitRec Dataset and applies the K-Mean Clustering algorithm, with the support of Apache Spark, for large-scale data processing, given the large data size in the FitRec dataset. Grouping is done using the FitRec dataset and K-Mean. Users are grouped according to heart rate and pace information; this provides appropriate Exercise for users. The test results show that the proposed system performs well, as indicated by the Silhouette Score = 0.596, Calinski-Harabasz Score = 2133.09, and Davies-Bouldin Score = 0.480. These test metrics reflect the system's ability to cluster. Indirectly, the accuracy performance of the system is assessed through these metrics, showing good accuracy test results.
Keywords: Apache Spark; K-Mean Clustering; Physical Activities; Recommender System;

INTRODUCTION

Physical activity such as sports is an important activity to maintain physical health. By doing sports activities, the body becomes healthier (Lamusu, 2018). In this modern era, sports activities can be monitored using smartwatches and smartphones to find out what pace to keep and what the heart rate is when exercising. We can understand our performance by knowing the heart rate and pace measurements during exercise (Ni et al., 2019). Evaluating post-exercise effects is vital to understanding the impact on the body, given that exercising beyond one's limits can lead to disease risk. Therefore, it is essential to set limits appropriate for the body's capabilities when exercising so that the activity can provide health benefits without risking disease (Malm et al., 2019).

A recommender system is a platform that can offer product suggestions or options that match a user's preferences (Jannach et al., 2021). The existence of a recommender system has significant value in supporting users' daily activities. Currently, not only limited to e-commerce platforms but even in social media platforms such as Telegram, recommendation systems are applied to suggest sports activities that match user preferences (Muhammad et al., 2023). To understand the user's preferences, information such as exercise records can be an indicator. Performance analytics such as pace, heart rate, and type of exercise performed are key (Ni et al., 2019).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13374 p-ISSN : 2541-044X

Big data holds comprehensive and valuable information, holding great potential as an object of research, especially in the context of research on recommendation systems (Wang et al., 2020). In general, the concept of big data consists of three main dimensions, namely Volume (large data size), Velocity (speed of data growth or processing), and Variety (variety of data types) (Patgiri & Ahmed, 2016). The data contained in big data can be processed to be implemented into a recommendation system to identify groups based on specific criteria.

K-Mean clustering is an unsupervised machine learning; K-Mean clustering is closely related to data processing and analysis (Sinaga & Yang, 2020). By utilizing sequential data such as heart rate and pace, this research uses the ability and power of K-Mean clustering to provide personalized and practical exercise recommendations using sequential data.

Research on recommender systems using k-mean clustering, such as movie recommender systems with k-mean clustering, has been done before (Ahuja et al., 2019). A similar approach was applied to an effective running route recommender system (Loepp & Ziegler, 2018). In line with these developments, this research directs attention to applying K-Mean clustering in the context of personalized exercise recommendation, hoping to contribute to developing a more adaptive system that matches each user's unique characteristics.

We propose developing a sequential data-based physical activity recommendation system using the K-Mean clustering method. The main objective of this research is to assess the ability of K-Mean clustering to form a recommendation system. The evaluation is done by considering the quality of clustering results generated, focusing on the formation of clusters that can provide physical activity recommendations by a user's heart rate and pace parameters.

LITERATURE REVIEW

Currently, recommendation systems provide various benefits to users, especially in physical health, with their ability to monitor heart rate and pace to evaluate performance during sports activities. In 2020, (Ali et al.) researched to develop a system that can predict and monitor heart disease based on body sensor data, especially heart rate. The approach in this research uses base ensemble deep learning and future fusion technology.

The utilization of sequential data in datasets makes it easier for researchers to develop more sophisticated recommendation systems. (Abdulaziz et al., 2021) They researched to create a personalized recommendation system by utilizing sequential data. The system aims to determine user categories in sports based on heart rate, pace, and altitude parameters during physical activities. Datasets that present complete information in sequential form have a close relationship with the concept of Big Data. In this context, (Ismail Ebada et al., 2022) performed big data processing using Apache Spark to develop a health prediction system based on information from big data.

The k-means clustering method has become common in data analysis and statistics research. In 2023 (Ikotun et al.) researched the k-means clustering algorithm in the era of big data. The study highlighted the vulnerability of the clustering process, which is highly dependent on data quality as it can affect the final result. K-means clustering aims to minimize internal variability within clusters by maximizing variability between clusters. The K-means method works by grouping data into predefined groups of clusters, showing that k-means clustering is effective for numerical data and can produce good clusters, especially when the group structure in the data is quite clear (Ahmed et al., 2020).

In 2021 (Puspasari et al.) conducted research that implemented the k-means clustering algorithm in a job recommendation system. This research was conducted to understand how the k-means clustering algorithm can be used in constructing a job recommendation system. (Ashari et al., 2023) researched classifying flood-affected areas in Jakarta using the k-means clustering algorithm. The clustering process is evaluated using the Silhouette, Davies-Bouldin, and Calinski-Harabasz metrics to get the conclusion of the clustering process.

In 2017, (Zhang et al.) researched the development of a tourist and travel route recommendation system. The purpose of this research is to recommend routes based on the needs and available time of the user, taking into account the total travel time and opening hours of tourist attractions as determining factors in making recommendations.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The presence of a recommendation system for physical health can have a positive impact on user survival. Previously, users often lacked a preference for exercise. With technology and recommendation systems, maintaining health and exercising can improve mental health and reduce the risk of depression (Pieh et al., 2020).

METHOD

This study aims to implement the K-Mean clustering algorithm in developing a Recommender System. The final result of this study is a cluster model derived from the FitRec Project dataset. The cluster model is evaluated using relevant evaluation metrics. Furthermore, the cluster results were clustered to identify the various user characteristics in the data that had gone through the clustering process. The steps involved converting the JSON format dataset into a CSV format dataset, data preprocessing, implementing the K-Mean clustering algorithm, and model evaluation, as illustrated in Figure 1 and Figure 3. Subsequently, comprehensive data preprocessing is conducted, refining the dataset for practical application in the subsequent clustering process. The central methodological component comprises implementing the K-Mean clustering algorithm, a pivotal step to derive clusters indicative of distinct patterns within the fitness data.

This study particularly emphasizes the application of evaluation metrics such as the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score to ensure the robustness and reliability of the derived clusters. The process, illustrated in Figure 1, embodies a systematic approach to producing an effective Recommender System based on user characteristics identified through the K-Mean clustering algorithm applied to the FitRec dataset. The decision to use metric evaluation is helpful to assist in modeling cluster grouping. By integrating the metric evaluation, this research takes a systematic approach to building an effective Recommendation System. In particular, the model is designed to reference user preferences accurately, considering individual characteristics such as heart rate and pace. The application of metric evaluation increases confidence in the model and provides a solid empirical basis for further interpretation and discussion of the results.

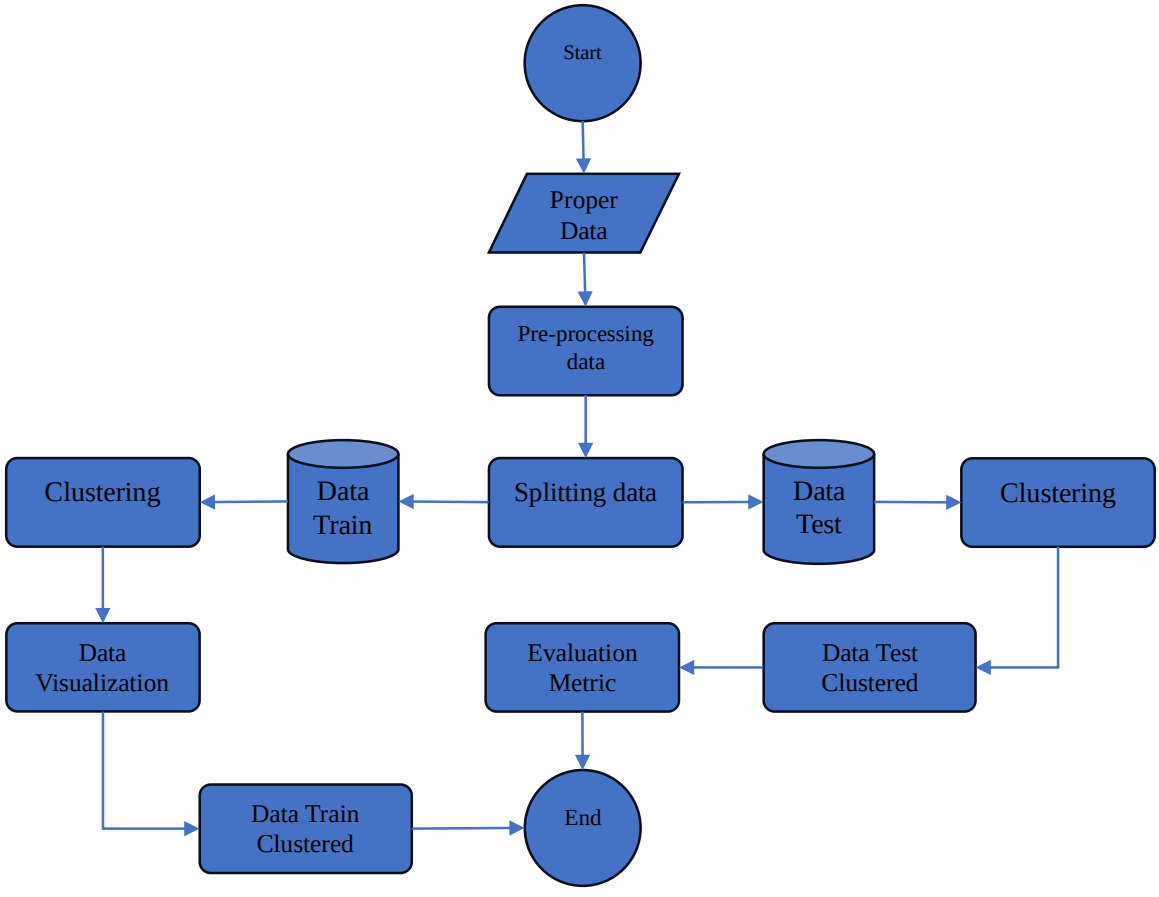

Figure 1. Flowchart of the constructed system

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dataset

The data used in this study came from the FitRec Project (Ni et al., 2019). This dataset comes from users' exercise activity records on the endomondoHR app and includes 167,783 exercise records from 956 unique users. The dataset is formatted in a JSON file with a size of 5 GB and consists of columns such as user-id, gender, sport, id, heart rate, speed, longitude, altitude, latitude, and timestamp. For more information on the data type of each column, please refer to Table 1.

Table 1. FitRec Project dataset description and overview

column_name	data_type	unit	description
altitude	Array <double>	Meter	Altitude recordings in exercise
gender	String	Male, Female, Unknown	The gender of each user
heart_rate	Array <bigint>	Beat per Minute (BPM)	Heart rate recordings during exercise
id	Bigint	-	Unique ID of each user's exercise record
latitude	Array <double>	Degree	Geographic recording of users when exercising
longitude	Array <double>	Degree	
speed	Array <double>	Mile per Hour (MPH)	Record of user’s pace during exercise
sport	String	-	Sports that users do
timestamp	Array <bigint>	Unix Timestamp	Time recording of users' exercise
URL	String	-	The URL link of each user's sports recording
User-id	Bigint	-	ID for each user

Explanatory Data Analysis (EDA)

To handle a large-scale dataset, performing an EDA to understand the dataset’s content more detailed and compactly is crucial. This analysis can be found in Table 2, which includes essential information such as the number of workouts, users, gender, heart rate, and speed. The table indicates that the dataset consists of 43 workouts, with male users dominating the number at 156,717, followed by female users at 9,881, and unknown gender at 1,185. This fact illustrates that the number of male users is more dominant than the number of female users and the unknown gender in this dataset.

Table 2. Explanatory Data Analysis Result

Descriptive Data Analysis (EDA)		
Workout		167,783
User		956
Sport		43
Gender	Male	156,717
	Female	9,881
	Unknown	1,185
Heart_rate	Min	0.0
	Max	239.0
	Mean	138.7
	Standard Deviation	18.96
Speed	Min	0.0
	Max	74.85
	Mean	20.96
	Standard Deviation	8.48

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

From 167,783 exercise records with 956 users, the average heart rate in this dataset is 138.7 BPM, with a standard deviation of 18.96 BPM. The heart rate ranges from 0.0 BPM to 239.0 BPM. The large standard deviation of the heart rate indicates a significant variation in values, which can be influenced by factors such as gender, weight, health conditions, and other variables. The mean speed was 29.96 MPH, with a standard deviation of 8.48 MPH and a speed range from 0.0 MPH to 74.85 MPH. The minimum speed value indicates that some sports do not require fast movements. The relatively high standard deviation of the speed shows that the user's pace data is widely dispersed, and the majority is centered around the mean value.

Apache Spark

Because the initial dataset is in JSON format and has a large size, it needs to be processed using big data preprocessing. The dataset is processed using the Pyspark library, a specialized Python library designed to manage data through the Apache Spark framework (Verma et al., 2015). Apache Spark, as a technical framework for big data processing (Ismail Ebada et al., 2022), provides advantages in processing big data through Pyspark. It should be noted that performance and conceptual differences between Pyspark and pandas can be observed, as illustrated in Figure 2. The illustration can be seen in Table 1, where the average column has an array data type, and data preprocessing using Pyspark is required.

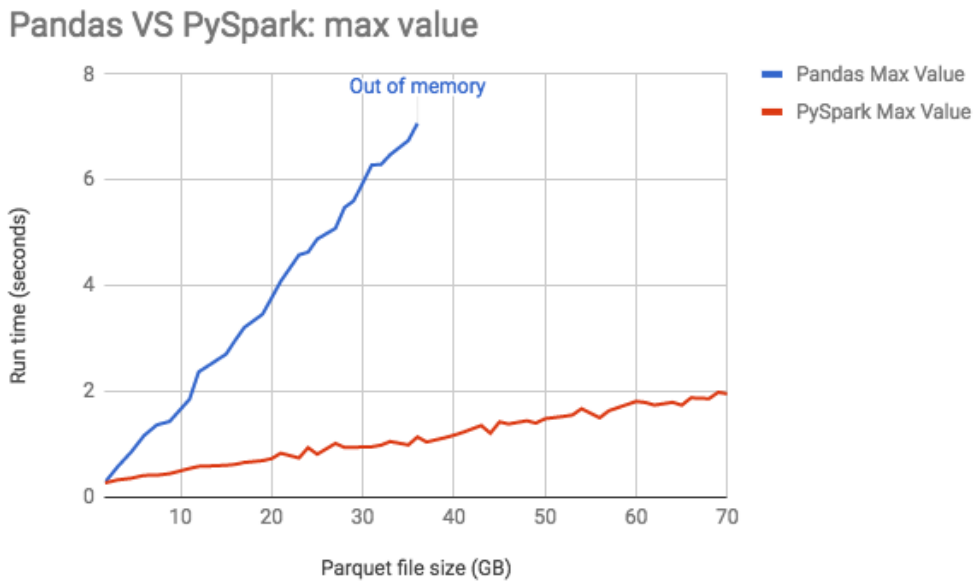

Figure 2. Differences in performance between pandas vs. pyspark

Figure 3 shows the process of obtaining proper data using Pyspark. The first step to obtaining appropriate data is aggregating the heart rate and speed columns to get the average value based on the workout ID. Both columns in the original data contain array data. Before aggregating, the process of deleting columns that are not needed, such as longitude, altitude, latitude, and URL columns. Then, a data cleaning step is carried out with the drop null value method to eliminate null values in the data row.

Furthermore, a drop null value is performed on the unknown gender to increase the validity of user data so that the gender available in the data only consists of males and females. After the column deletion, data cleaning, and gender unknown drop processes, the number of data rows decreased from 167,783 to 31,554. With these changes, some aspects of the data were also modified, such as the number of users from 956 to 737 and the number of sports from 43 to 15 in the dataset.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

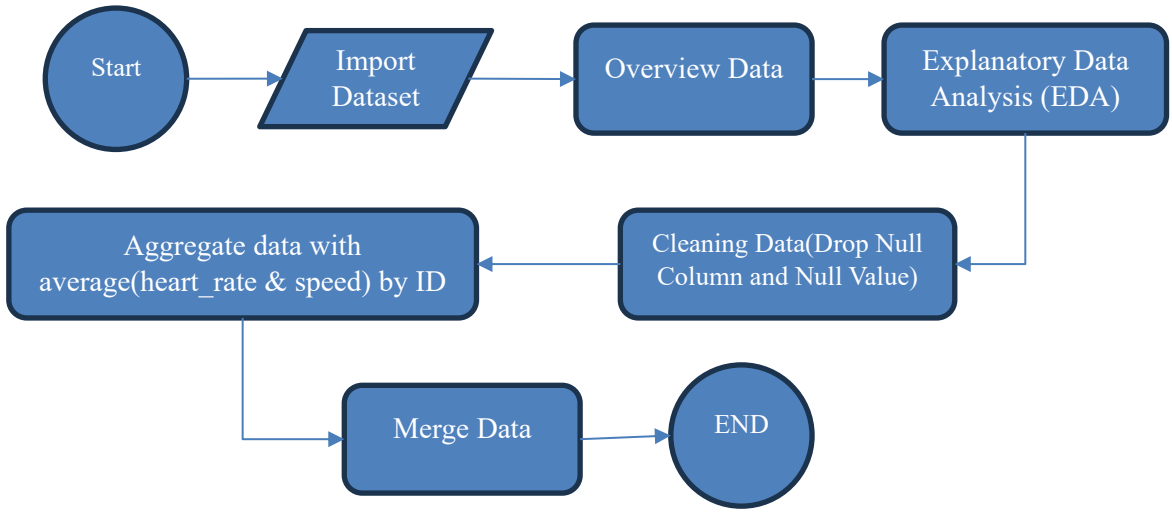

Figure 3. Flowchart of Apache Spark process to get proper data

Pre-processing Data

The following method applied is data pre-processing using data that has been adequately processed. Before the clustering stage, the data pre-processing step is vital to ensure that clustering results can be obtained quickly and prevent errors (Abdulaziz et al., 2021). The first step in data pre-processing is to remove the timestamp column, which allows the use of data in the clustering process, considering that the timestamp column is a DateTime data type.

After deleting the columns, continue with the data summary process to calculate the average in the heart rate and speed columns. This data summary process uses the group by function based on the user-id, gender, and sports columns. This approach was taken because the number of rows of data was still quite large, and there were device and time limitations for running with a large number of rows of data. Through the data summary process, the number of data rows is from 31,554 to 1,253; this happens because a user has many sports records in one type of sport before the data is summarized.

The next step is to label the data to convert it into numeric values. This data labeling process aims to correct data with a string data type into numeric values under the clustering with K-Mean, which requires a dataset with numeric values. Two columns are labeled in the data labeling process, namely the gender and sports columns. For gender labeling, the male is converted to the value "1" while the female is converted to "0". Furthermore, the sports column or sports activities are labeled in order. Although the number of rows of data became 1,253, the number of sports activities remained 15. Table 3 provides an example of the labeling results.

After both columns have been numerically labeled, the next step is to run a data normalization process. This process aims to optimize the performance of machine learning algorithms, as good data quality is a crucial factor in the smooth operation of the algorithm (Singh D & Singh B, 2020). Data normalization is performed using the MinMaxScaler function in the Python sci-kit-learn library. This function changes the value of a column to a range of 0 to 1. The columns that underwent normalization were the (avg)heart_rate and (avg)speed columns, as seen in Table 3. Data normalization aims to simplify the average value of the two columns, ensuring that the clustering process can occur optimally.

The last step in data pre-processing is data splitting, which is divided into two parts: training and testing data. The data division ratio is 70% for train data and 30% for test data. With a total of 1,253 rows of data, the train data has 877 rows, while the test data has 376 rows. This data division process is carried out to train the model using train data and evaluate the clustering quality through test data to obtain relevant evaluation metrics.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. Example of data after Pre-Processing data

userid	gender	(avg)heart_rate	(avg)speed	sport_encoded
2358	1	0.717181	0.448214	1
2358	1	0.695478	0.166723	13
3808	1	0.667869	0.334491	1
3808	1	0.641009	0.143253	13
3808	1	0.232345	0.073137	15

Clustering

The clustering process is a stage in which individuals in a dataset are grouped based on specific parameters or criteria, such as the similarity of features or attributes (Sinaga & Yang, 2020). The KMeans function of the Python sci-kit-learn library is used to implement this clustering process. The KMeans method helps identify patterns, relationships, or structures in data, providing straightforward interpretation and decision-making. The K-Means algorithm is a commonly used clustering method (Ashari et al., 2023).

Choosing the optimal number of clusters (K) is vital so that the clustering results provide a deep understanding of the data structure. The elbow method is one common approach to determining the optimal K value (Ashari et al., 2023). This method involves iterating data clustering by varying the number of clusters and measuring the intra-cluster variance for each cluster. As the number of clusters increases, the intra-cluster variance will generally decrease. However, the decrease slows down at a certain point, forming an "elbow" on the graph. When the variance decrease slows down, the K value can be considered optimal (Sinaga & Yang, 2020).

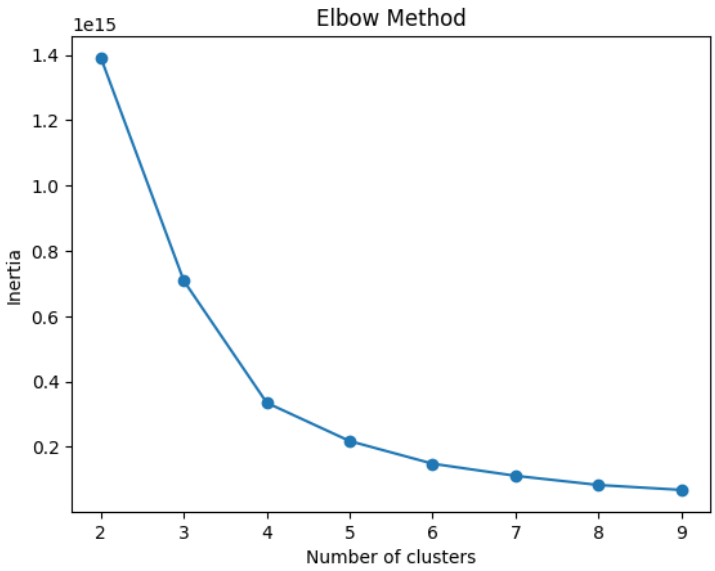

Figure 4. Elbow Method Result

In this elbow method, the number of K is set for a series of iterations with a range of K values between 2 and 10, using the data_train derived from the data splitting process. The results of finding the optimal K value through the elbow method are presented as a graphical plot, as shown in Figure 4. Implementing the elbow method yielded an optimal K value of K = 4. The assignment of this value is based on the observation that the graph shows a significant decrease at K = 4, and after that point, the line shows a smoother slope at subsequent K values.

After obtaining the optimal K value, the next step is to run the clustering process using the KMeans function with a value of K = 4. In Figure 5, it can be seen that there are four clusters with different colors. The clustering results use the avg_heart_rate, avg_speed, and user-id attributes for grouping

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13374
e-ISSN : 2541-2019
p-ISSN : 2541-044X

users using 3D plots. The visualization illustrates that each user has been grouped into a particular cluster according to their average heart rate and speed, indicating that the user has been attributed to a particular cluster based on their heart rate and speed capabilities.

After knowing the clustering results and displaying the 3D cluster visualization, the next step is to perform the cluster labeling process. This step aims to assign a label or category to each cluster formed. Cluster labeling is a step of interpreting the clustering results, which allows us to give meaning or understandable interpretation to the groups formed.

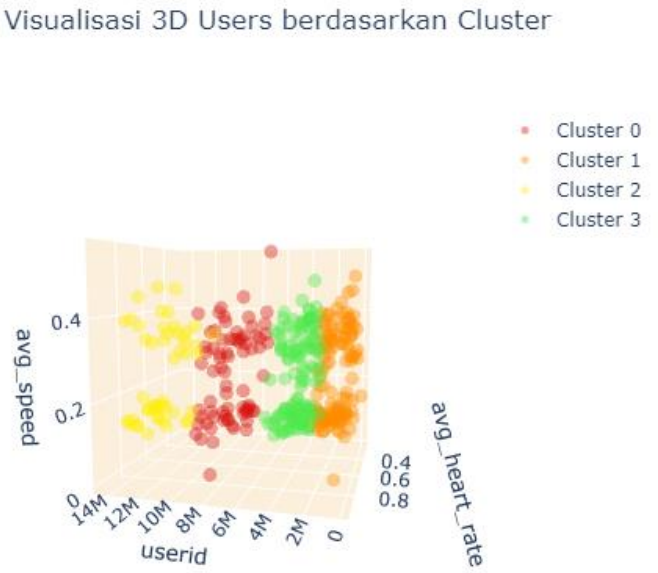

Figure 5. Visualization of clustering result with 3D Plot

Evaluation Metric

Metric evaluation after the clustering process is crucial to assess the quality of the clustering results. Metric evaluation provides insight into the extent to which the clusters formed to match the actual structure in the data. Some commonly used evaluation metrics in clustering include the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score (Ashari et al., 2023).

Silhouette Score (1) is an evaluation metric used to measure how well an object sits within a cluster compared to other clusters in cluster analysis. It indicates how well the cluster has been formed.

$$S(i) = \frac{\mathrm{b}(i) - \mathrm{a}(i)}{\max\{\mathrm{a}(i), \mathrm{b}(i)\}} \quad (1)$$

In this context, $S(i)$ represents the Silhouette Score for object i, $\mathrm{a}(i)$ is the average distance between object i with other objects in the same cluster (intra-cluster distance), and $\mathrm{b}(i)$ is the average distance between object i with objects from other clusters that are closest (inter-cluster distance). The function $\max\{\mathrm{a}(i), \mathrm{b}(i)\}$ is used to obtain the maximum value between $\mathrm{a}(i)$ and $\mathrm{b}(i)$. The Silhouette Score value ranges from -1 to 1. The higher the Silhouette Score value, the more optimal the placement of an object in the corresponding cluster and the better the overall cluster division. A positive value indicates that the object is in the appropriate cluster, while a negative value indicates that the object may be placed in the wrong cluster. Conversely, a value of zero indicates that the object is between two clusters that have the same level of similarity.

Calinski-Harabasz Score (2) is an evaluation metric used to measure the extent to which clustering results have been successfully formed. It focuses on measuring how well the inter-cluster variance

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13374 p-ISSN : 2541-044X

compares to the intra-cluster variance. If the Calinski-Harabasz score is high, it indicates that the clusters have a high degree of variance between each other, but low variance within the clusters.

$$\mathrm{CH} = \frac{B(k)}{W(k)} X \frac{N-k}{k-1} \quad (2)$$

As seen in equation (2), there are several variables to understand. The variable k indicates the number of clusters. $B(k)$ measures the variation between clusters, assessing how different the cluster means are from each other. $W(k)$ indicates the within-cluster variation, measuring the extent to which the data in the cluster is uniform. N notes the total amount of data in the dataset. The ratio $\frac{B(k)}{W(k)}$ indicates how large the variation between clusters is compared to the variation within clusters. If the variation between clusters is large enough compared to the variation within clusters, the CH value will be higher, indicating good clustering quality. The adjustment factor $\frac{N-k}{k-1}$ is used to correct for the effects of sample size and number of clusters. In a practical context, a higher CH value reflects an optimal number of clusters, creating a good and uniform data partition. Therefore, in cluster analysis, we look for the number of clusters that give the highest CH value as the optimal solution to the clustering problem.

Davies-Bouldin Score (DB Score) (3) is an evaluation metric used to assess the extent to which a cluster is optimally formed in the clustering process. It focuses on the concept of how close (similar) each cluster is to its closest cluster and how far (different) it is from other clusters. A lower Davies-Bouldin Score indicates better clustering results.

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq 1} \left(\frac{S_i - S_j}{d(c_i, c_j)} \right) \quad (3)$$

As seen in equation (3), it involves several variables such as the number of clusters k, the scatter within cluster S_i, and the distance between cluster centres $d(c_i, c_j)$. The DB Score reflects how effectively the clustering is able to separate groups of data. The lower the DB score, the more effective the separation between clusters, indicating the optimality of clustering. In the formula, S_i represents the scatter within cluster i, calculated as the average distance between each point in cluster i with the center of cluster i. The variable $d(c_i, c_j)$ is the distance between the centres of clusters i dan j, which measures how far cluster i is from cluster j in feature space. The key component in the DB formula is $max_{j \neq 1} \left(\frac{S_i - S_j}{d(c_i, c_j)} \right)$, which evaluates how effectively cluster i is separated from cluster j.

The overall DB Score is calculated as the average of these ratios for each cluster, reflecting the extent to which the clustering has achieved effective separation. The DB Score has a range of values from 0 to infinity. It provides a holistic picture of the quality of the clustering by considering both the dispersion within clusters and the distance between clusters. The use of DB Score, together with other evaluation metrics, can provide more comprehensive information about the extent to which the clustering is able to provide a good representation of the underlying data structure (Ashari et al., 2023).

RESULT

In evaluating clustering results, several metrics are used to measure the quality and characteristics of the clusters formed. One commonly used metric is the Silhouette Score, which indicates the extent to which an object is in the cluster compared to its neighbours. In this study, the Silhouette Score reached a value of about 0.596. The positive score indicates the closeness of the objects in the cluster and the formation of an optimal cluster with a strong internal structure.

The Calinski-Harabasz Score metric, which reached a value of about 2133.09, indicates how well the separation between clusters is. The higher the score, the better the separation between clusters. The evaluation results show that the clusters are well-formed and have significant differences. This gives

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 1, January 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i1.13374 p-ISSN : 2541-044X

confidence that the clustering algorithm successfully separates the clusters according to the underlying data structure. The Davies-Bouldin Score (DB Score), with a value of about 0.481, provides information about the quality of cluster formation in cluster analysis. This score measures how well the clusters are formed and how far one cluster is from another. The DB Score evaluation shows that the clustering results are good quality, with low values indicating optimal results.

With 376 observations in the test data, the evaluation was performed on a dataset extensive enough to give a solid representation of the characteristics or patterns in the data. Overall, these metrics provide a positive picture of the quality of the clustering results. Clusters were formed with high internal similarity and significant differences between clusters. This evaluation supports using clustering algorithms in recommender systems based on sequential data. The clustering results revealed a close relationship between user characteristics based on heart rate and speed. The clusters reflect different exercise patterns and levels of physical activity intensity. These clusters can be interpreted as representative of a particular lifestyle or exercise preference.

Table 4. Evaluation Metric result

Metric	Value
Observation Number	376
Silhouette Score	0.5960659965645076
Calinski-Harabasz Score	2133.0926294152514
Davies-Bouldin Score	0.480788126271269

DISCUSSIONS

In this section, we present an analysis of the results of the study experiments. This discussion involves an evaluation of the study findings and provides a detailed view of the experimental results that have been generated. Some aspects commonly covered in the discussion section include comparisons between measured data and model results, comparisons between different modeling methods, and further interpretations of emerging new and significant findings.

As support for the discussion, Table 5 details the four clusters that resulted from this study. The clusters are distinguished by name, number of data, percentage of total data, average heart rate, average speed, and the predominant sport type within each cluster. This table provides a more detailed description of the characteristics of each cluster, facilitating further understanding of the patterns identified in the clustering experiment results. This detailed information enriches the knowledge of the clustering experiment patterns and provides valuable support for a broader discourse on the implications and applications of the research results. The systematic description of the attributes of each cluster increased the depth of discussion, allowing for a more in-depth exploration of the relationship between physical activity, physiological responses, and clustering outcomes.

Table 5. Overview information from four clusters for experiment result

cluster	cluster_name	count	%	avg_heart_rate	avg_speed	sports
0	Average-Fit	207	23.60	141.89	18.44	bike, run, fitness walking, bike (transport), walk, mountain bike, roller skiing, core stability training, indoor cycling, cross-country skiing, orienteering
1	Most-Fit	296	33.75	141.28	19.59	bike (transport), fitness walking, run, bike, core stability training, mountain bike, indoor cycling, walking, cross-country skiing, hiking, orienteering

*name of corresponding author

2	Moderate-Fit	121	13.79	141.58	18.96	run, bike, indoor cycling, mountain bike, orienteering, bike (transport), cross-country skiing, core stability training, walking, skating, kayaking, roller skiing
3	Above-average-Fit	253	28.84	140.86	19.44	bike, bike (transport), skate, mountain bike, run, indoor cycling, fitness walking, core stability training, walking, horseback riding, orienteering, cross-country skiing, roller skiing

In this section of the study experiment, it was observed that the data could be categorized into four groups or clusters, each exhibiting different physical fitness traits. The "Most-Fit" cluster dominates with the most significant percentage of the total data, featuring a relatively high average heart rate and speed. This group tends to be active in activities such as cycling (especially for transportation), fitness walking, and running. Meanwhile, the "Moderate-Fit" cluster has a lower percentage but shows significant variation in average heart rate and pace. The main activities in this group include running, cycling, and indoor sports such as indoor cycling.

Despite the relatively smaller percentage, the "Average-Fit" Cluster made a meaningful contribution with unique characteristics. Despite having a lower percentage, this group displays a stable average heart rate and moderate pace. Activities that dominate this group include cycling, running, and fitness walking. Meanwhile, the "Above-average-Fit" cluster features a relatively low average heart rate but a high pace. Sports such as cycling and running and indoor sports such as indoor cycling dominate the activities in this group.

Further discussion of these results may reveal possible patterns in exercise behavior based on physical characteristics. The existence of a "Most-Fit" group that tends to be active in cycling, fitness walking, and running may reflect specific preferences or habits in the community that were the focus of the study. These results may provide a basis for recommending fitness programs more tailored to specific physical characteristics. Comparisons between groups may also provide additional insights into the factors influencing the intensity and type of exercise individuals choose. Overall, the findings in this study contribute to our understanding of physical fitness patterns in the context of the population under investigation.

Recommendations drawn from these clustering results include the potential development of more personalized exercise or physical activity plans. For example, the recommender system could offer advice on the type of exercise or activity that aligns with each cluster's heart rate and speed characteristics. This approach can help users optimize their health and fitness benefits based on exercise patterns that best suit their physical conditions.

CONCLUSION

This study implements the K-Mean clustering method for developing a recommender system based on sequential data, such as heart rate and pace. The goal is to achieve optimal clustering results in the system to facilitate exercise recommendations according to the user's heart rate and pace so the users can have sports preferences according to their capabilities. Determination of the optimal number of clusters, K=4, is done through the elbow method. System evaluation using metrics such as Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score is essential to measure system performance and the quality of processed data. The system shows a satisfactory evaluation using 30% of the data as test data, as seen from the Silhouette Score value of about 0.596, Calinski-Harabasz Score of about 2133.09, and Davies-Bouldin Score of about 0.481. This evaluation indicates that the clusters formed

*name of corresponding author

show high internal similarity and significant differences between clusters. The system successfully formed four groups with specific sports according to the users' average heart rate and average pace. Additional features, such as external factors, weather, or environmental conditions, should be considered for future research. Further exploration of the exercise behavior patterns within each cluster can provide deeper insights and support the development of a more personalized and relevant physical fitness recommendation system.

REFERENCES

Abdulaziz, M., Al-motairy, B., Al-ghamdi, M., & Al-qahtani, N. (2021). Building a Personalized Fitness Recommendation Application based on Sequential Information. *International Journal of Advanced Computer Science and Applications*, *12*(1). https://doi.org/10.14569/IJACSA.2021.0120173

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Ahuja, R., Solanki, A., & Nayyar, A. (2019). Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 263–268. https://doi.org/10.1109/CONFLUENCE.2019.8776969

Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K.-S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208–222. https://doi.org/10.1016/j.inffus.2020.06.008

Ashari, I. F., Dwi Nugroho, E., Baraku, R., Novri Yanda, I., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing*, *7*(1), 89–97. https://doi.org/10.30871/jaic.v7i1.4947

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Ismail Ebada, A., Elhenawy, I., Jeong, C.-W., Nam, Y., Elbakry, H., & Abdelrazek, S. (2022). Applying Apache Spark on Streaming Big Data for Health Status Prediction. *Computers, Materials & Continua*, *70*(2), 3511–3527. https://doi.org/10.32604/cmc.2022.019458

Jannach, D., Pu, P., Ricci, F., & Zanker, M. (2021). Recommender systems: Past, present, future. *AI Magazine*, *42*(3), 3–6. https://doi.org/10.1609/aimag.v42i3.18139

Lamusu, Z. (2018). Olahraga Dan Penyakit Zaman Modern. *Ideas: Jurnal Pendidikan, Sosial, Dan Budaya*, *4*(4), 537–552. https://jurnal.ideaspublishing.co.id/index.php/ideas/article/view/115

Loepp, B., & Ziegler, J. (2018). Recommending Running Routes: Framework and Demonstrator. *Workshop on Recommendation in Complex Scenarios*.

Malm, C., Jakobsson, J., & Isaksson, A. (2019). Physical Activity and Sports—Real Health Benefits: A Review with Insight into the Public Health of Sweden. *Sports*, *7*(5), 127. https://doi.org/10.3390/sports7050127

Muhammad, W. S. F., Baizal, Z. K. A., & Dharayani, R. (2023). Ontology-Based Recommender System for Personalized Physical Exercise in Obesity Management. *Sinkron*, *8*(3), 1699–1708. https://doi.org/10.33395/sinkron.v8i3.12689

Ni, J., Muhlstein, L., & McAuley, J. (2019). Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation. *The World Wide Web Conference*, 1343–1353. https://doi.org/10.1145/3308558.3313643

Patgiri, R., & Ahmed, A. (2016). Big Data: The V's of the Game Changer Paradigm. *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 17–24. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0014

*name of corresponding author

Pieh, C., Budimir, S., & Probst, T. (2020). The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (COVID-19) lockdown in Austria. *Journal of Psychosomatic Research*, *136*, 110186. https://doi.org/10.1016/j.jpsychores.2020.110186

Puspasari, B. D., Damayanti, L. L., Pramono, A., & Darmawan, A. K. (2021). Implementation K-Means Clustering Method in Job Recommendation System. *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 1–6. https://doi.org/10.1109/ICEEIE52663.2021.9616654

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*(Part B), 105524. https://doi.org/10.1016/j.asoc.2019.105524

Verma, J. P., Patel, B., & Patel, A. (2015). Big Data Analysis: Recommendation System with Hadoop Framework. *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, 92–97. https://doi.org/10.1109/CICT.2015.86

Wang, K., Zhang, T., Xue, T., Lu, Y., & Na, S.-G. (2020). E-commerce personalized recommendation analysis by deeply-learned clustering. *Journal of Visual Communication and Image Representation*, *71*, 102735. https://doi.org/10.1016/j.jvcir.2019.102735

Zhang, C., Liang, H., & Wang, K. (2017). Trip Recommendation Meets Real-World Constraints. *ACM Transactions on Information Systems*, *35*(1), 1–28. https://doi.org/10.1145/2948065

*name of corresponding author