Building a Personalized Fitness Recommendation Application based on Sequential Information

Manal Abdulaziz[1], Bodor Al-motairy[2], Mona Al-ghamdi[3], Norah Al-qahtani[4]
Department of Information Systems,
Faculty of Computing & Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—Now-a-days sports plays a very important role in the life of the human being and it allows to keep him healthy and make him always active. Sport is essential for people to have a healthy mind. However, the practice of a sport can have negative effects on the body and human health if it is practiced incorrectly or if it is not adapted to the body or the human health. This is why, in this paper, we have proposed a recommendations system that allows the selection of the right person to practice the right sport according to several factors such as heart rate, speed and size. The implementation was applied to the FitRec dataset with the help of SPARK tool, and the results show that the proposed method is capable of generating the appropriate training for different groups according to their information, where each group gets the appropriate training. The grouping of this data was done by the k-means method.**

***Keywords*—Big data; big data processing; recommendation system; sport analysis; K-means**

I. Introduction

Sport is an important healthy behavior for improving public health, superior utilization of energies, and it constitutes excellent protection for the human being. In the 21st century, Activities of Sport are modern phenomena to make everybody keep healthy [1]. It's referred to a normal physical effort or skill practiced under agreed rules with the aim of entertainment, competition, pleasure, distinction, developing skills or strengthening self-confidence. No two people disagree on the importance of sport to both physical and psychological health. However, despite this, it has been proven that some sports can have negative results at times parallel to their positive effects on those who practice them. It is difficult for people who have not previously practiced sports to know the correct way to practice some exercise, which makes the possibility of injuries greater. Therefore, it is best to use a sports coach to design a special program and see the appropriate sports for the person in proportion to the individual's capabilities and health so that the latter does not suffer any possible injuries in the future or negatively affect the body health. And because we saw that every person who practices sports needs to know what sport is appropriate for his body and health, we have proposed in this paper a recommendation approach, which recommends the appropriate sport for the person based on several parameters such as heart rate, speed, and height. These parameters are recorded by smart wearable devices (ex. Apple Watch) as shown in Fig. 1. Improving wearable technology enables people to measure and control their behaviors via mobile devices such as Fitbit and Apple watches [2].

The remainder of the paper is ordered as follow, Second Section provide a background of the big data analytic and processing, the third section related work to a spotlight on relevant work that has been conducted in this field. Fourth section concentrates on description of data set. Discussion the methodology and tools that used to build recommendation approach will be in the fifth section, and then the rest sections will be for implementation and results and conclude this paper with conclusion and references.

II. Background

A. Big Data Analytics Definition and Importance

The big data analytics (BDA) term was defined for the first time in 2012 by Gantz and Reinsel [3]. They stated that BDA includes three main dimensions, they are the huge data, the technologies used to analyze the data, and the valuable insights delivered from the data to create business values. Some researchers used the term BDA to refer to applying advanced analytic technologies against huge, heterogeneous, and diverse data that coming from different sources to extract values [4]. In Fact, big Data analytic is a revolution in the data science area. The organizations' need to big data analytics are increased, since their data are becoming bigger and heterogeneous. However, this huge amount of data does not provide any value in its raw nature [5]. Organizations have to employ advanced technologies to gain new values, enable smarter decisions, solve economic and social issues. Big data analytics assist process to be more efficient and facilitate in growing the organizations' profitability. As stated in the literature, there are also substantial advantages to performing big data analytic, they are [6][7].

- Enhancing decision makers: big data analytics enhance decision makers since it facilitates a understanding holistic data, creates predictive model, detects data trends and associations, and discovers data patterns.
- Predicting results: big data analytics tools able to model scenarios and then predicts results for each scenario, allowing decision makers to take a right action at the right time.
- Detecting Fraud: big data processing helps to quickly detect and correct deviations and outliers in data.
- Increasing data transparency: big data analytics expand the volume of exchange data between users while

preserving privacy. This therefore allows to access more valuable information.

- Improving organization productivity: big data analytics are considered an essential productivity driver since it provides opportunities for managing change and encouraging workers to be more conscious of their job patterns and practices.
- Improving customer service: big data analysis offers a rich knowledge of both customer issues and sentiment. It helps organizations to incorporate approaches to mitigate customer problems efficiently and proactively.
- Improving market intelligence: big data analysis provides insights into both the current and future state of the market and environment.
- Saving cost: big data tools are cost-effective since a huge amount of data can be processed and stored in commodity servers quickly.
- Developing new products: big data analytics allow organizations to develop new products according to consumers' needs based on analytics results of consumers' pattern and loyalty.

B. Type of Big Data Analytics

There are four major forms of BDA:

- Descriptive analysis: It is also known as data mining. It is the simplest and most common form of data analysis. It translates the immense volume of data into a statistical number that describe the data. Usually, findings of descriptive analysis are visualized in dashboards [8].
- Diagnostic analysis: it is used to detect the data patterns. Actually, diagnostic analysis assist to recognize the problems from their roots and understand reasons behind them [8].
- Predictive analysis: The organizations use predictive analysis to expect what may occur in the future based on both historical and current data [8].
- Prescriptive analysis: it is the most sophisticated form of data analysis as it makes a recommendation and recommends what actions can be taken. [8].

C. Big Data Processing

With the quick expansion of emerging applications such as, sensor networks, semantic web, LBS and social network applications, a variety of data to be processed continues to expand rapidly. Data processing is a collection of mechanics or models of programming to arrival large-scale data to elicit beneficial information for propping and providing decisions active handling of large-scale data (big data) poses an interesting but crucial challenge.

The ability to handle and process continuous data streams is becoming an essential part of building a data-driven organization. Data streams are sequences of unbounded tuples generated continuously in time. Unlike traditional batch processing which involves processing of static data [9], below sections illustrate the difference between them.

1) Batch processing: Batch processing has a long background in the field of big data. This processing includes operating over a huge, constant dataset and generating the result when the computation is complement later in time [10]. The sets of data in batch processing are:

1. Confined: datasets of batch are a limited data collection [10].

2. Continual: data is almost ever support by some kind of constant storage [10].

3. Huge: Often, batch operations are the only way to handle large data sets [10].

2) Stream processing: Streaming includes processing continuous or dynamic data [11]. The capability to manage and process continued data streams is becoming a major portion of constructing a data-driven organization [9]. Instead of determining operations to using to a whole set of data, stream processors realize operations that will be utilized to each singular data element as it push through the system [10]. The data set here considered "unbounded", this has a few major consequences:

1. Only the amount of data that has entered the system so far is known as the total dataset.

2. Perhaps the working dataset is more important and is restricted at one time to a single object.

Processing is event-based and unless expressly stopped, does not "end". Results are available instantly and will be revised constantly as new data arrives.

D. Recommendation Systems

Recommendation systems are tools that support the interaction with information spaces that are large and complex, by providing a personalized view of the information space through the prioritization of items that the user may find interesting [12]. The recommender systems' field was first identified in 1995 but became popular in the last decade [13]. According to [14], recommender systems are tools that provide the user with the appropriate recommendations. The definition provided during the inception of the field has been evolved [12]. In its early years, it was defined as a system in which people will provide the recommendations in the form of inputs, which will then be aggregated by the system and directed to the proper recipient. According to this definition, some cases depended primarily on aggregation for transformation while in different instances, the system's value depended on its ability to match recommenders and those in need of recommendations efficiently. This definition was based on collaboration among users. A recommendation system is therefore a system that will produce an output with an individualized recommendation. It can also be a system that can provide a user with personalized guidance to objects that are interesting or useful found in a large space with many possible options. Based on the evolved definition, which can also be formulated into a formal definition, two principles are identified about a recommender system. One is that it is personalized, meaning it caters to the needs of one user and not a group. Secondly, it offers assistance

to the user where discrete options are available, and the user has to make a selection [12]. In developing recommender systems, search needs were the main address point in addition to the selection of relevant products within Big Data from the internet, which is considered to be the biggest marketplace [15]. Recommendation systems are important because they provide customers with a direct connection to the products which they desire minimizing the browsing time used by the user. With the highest level of accuracy, a recommender system will predict a user's interests and provide a recommendation of the product that is perfectly matched to the user's interests [13]. With the assistance of recommendation systems, it is possible to recognize the tastes of a person and automatically find new content that would be desirable to them [16]. This means that the identification of a person's patterns, regardless of their variation in tastes. Recommendation systems are also important to the business using it since it can enhance its sales. This is made possible through the presentation of a variety of items that can match the user's interests since a recommended outcome is dependent on the commercial interests in addition to how ambiguous the customer is in formulating a request or query [15]. If a user or customer makes purchases of more items than they were looking for, through their search request, then the business makes a profit. They eliminate information overload by providing personalized recommendations [17].

1) Types of recommendation systems: The two major types of recommendation systems are content-based recommender systems and collaborative recommender systems. Also, there are hybrid recommender systems. A content-based recommender system uses a content-based filtering technique. A system model built using this approach is based on the representation of the product (item description) and the user preference (the customer's preference profile) [15][16]. In this approach, the products have features that are similar to products previously examined by the user or are currently being searched. The products or items are usually identified based on the user properties and the item properties. Other evaluations from the customers are not considered for this approach. The items recommended in this case are usually those that best match the items that have previously been rated. This approach has its drawbacks such as the system's is not able to provide suggestions for a variety of products. Another drawback is where the customer is required to rank several products beforehand so that they can get useful recommendations [16]. The collaborative recommender system uses collaborative filtering. This approach uses the knowledge of the relationship between the user and the item, product or service. It uses past behaviors (like previous purchases) or feedbacks (like ratings) by a customer in addition to decisions that are similar to this made by other customers [15][16]. This information and the relationship between the customer and the product help in predicting the recommended product. This approach has its drawbacks as well which include the need for extensive information on the customer's evaluation in order to achieve computations whose correlations and predictions are precise [15]. Another drawback is the fact that newly added customers or products are not included in the calculations. The hybrid recommender system applies the hybrid filtering technique [16]. This approach is a combination of the content-based and collaborative approaches while allows for the optimization of features from both approaches while minimizing the drawbacks associated with the two [15],[16]. In this research, the collaborative filtering approach has been adopted since it promotes building the model based on a user's past behavior and similar decisions made for other users to anticipate the most appropriate sport that the user needs.

Fig. 1. The Amount of Data that Collected from Smart Watch.

2) Applications of recommendation systems: The online market is one example where a third-party seller can trade their product on online marketplaces. The online market operator processes the transactions happening on the online marketplaces and the customers have a service that allows them to search for products using description or other properties in their knowledge [15]. Travel Industry is another application where customers are able to book hotels, purchase flight tickets directly as well as acquire holiday packages just from the mobile applications or from web pages without any additional costs that would otherwise be presented if the transaction was carried out physically. Recommender systems are being used in different places and for different purposes. It is being used in e-commerce companies such as Amazon to recommend items that a user can buy. Also, it is used for e-learning and e-library where users can get books and research documents in entertainment for movie suggestions, for instance, Netflix and YouTube. E-government is also using recommender systems [15].

III. Related Work

This research focuses sport recommendation and prediction (Fig. 1). We present the associated research from each related field as follows:

A. Context-aware Modeling

In several fields with ample contextual knowledge, context-aware models have been successfully adopted such as in recommender systems[18]. Naturally, like the other context-aware models, fitness and workout information has a heterogeneous input structure. Recommendation systems for tourism play a crucial role in supplying visitors with helpful trip planning. Recently researchers have been working to develop recommendation systems for routes. Users have been widely accepting such kind of systems [18]. Mehmood et al. (2019) [18] proposed a recommendation system to lead tourists in in South Korea.

The proposed system is based on the statistical analysis of user preference. Plus, the popularity of the sites, distance, traffic, weather and time are considered to recommend the routs. The system used Naïve Bayes classifier to calculate the probability of tourists to visit sites, and Haversine formula to measure the distance between the locations. The model evaluated using dataset that contains the patterns of tourist activity, obtained in the years 2016-2017 from tourists' smartphones. This real-time data was gathered from Wi-Fi routers distributed in 149 sites. The results show that, the tourists become able to visit more famous locations conveniently.

B. Sensor Data Mining

Ubiquitous computing combine technologies of processing data that collected from wearable devices, embedded devices and mobile applications [19]. Recently, there is a growing trend to model a recommendation system using these devices. Chowdhury et al. (2018) [20] proposed AdaBoost-based classifier to predict suitable exercise types in real-world contexts based on a limited data for training. First, the system extracts some features from the data such as the distance and heart rate. Then, the system mixes the extracted features with other features that related to each exercise like exercise duration. Finally, the system applies a 5-fold cross validation approach to classify the users. The experiment applied on a dataset contains 22 persons, who performed diverse exercises. The total of the sessions was 40.

Ni et al. (2019) [2] proposed a sequential context-aware model to extract the temporal patterns and personalized fitness data. The model gathers the specific activity data using wearable sensors. It also uses data from user's activity history. The model measures the average heart rate for each exercise, then it directs the user to a suitable exercise based on this biased. The model was evaluated on 250.000 records of workout and millions of measurements collected from the wearable devices.

C. Personalized Recommendation

Nowadays, academia and industry researchers have developed systems to recommend users based on their personalized behavior [21] [22]. Zhang et . (2016) [21] proposed personalized travel recommendations system that aim to recommend tailored routes based on users' needs and available time. The travel time volatility and point of interest opening hours are also considered to recommend the optimal rout. The model was evaluated on a real-time dataset. The results show that, the SE-SR algorithm helps in increasing performance but lesser quality than the optimal solutions, while PDFS algorithm is the most effective one. Loepp at el (2017) [22] proposed a system to provide runners to the optimal route based on their environment, goals and preferences. Of course, each route must be independently recommended, considering several distinct factors that decide whether a recommendation would eventually serve the runner. The system was evaluated on total of 11 runners from different ages and both genders.

Kushal Bafna, Durga Toshniwal [23] proposed a dynamic framework for a feature-based overview of the views of customers on online products, which operates according to the product domain. When they elicit online feedback on a periodic basis for a product, they conduct the following work any time after extraction; firstly, it is done to define the characteristics of a product from the opinions of customers. Next, their corresponding views are extracted for each feature and their orientation or polarity (positive / negative) is identified. It calculates the final polarity of the feature-opinion pairs. Feature-based summaries of the reviews are then created by extracting the related excerpts for each pair of feature opinions and placing them in their respective feature-based cluster. Results show that in carrying out their tasks, the proposed methods are highly productive and successful. Now it has become much simpler for users to digest the data found in large numbers of products review corpus by making use.

Klavdiya Hammond and Aparna S. Varde in [24] represent study the predictive model implementation on cloud utilizing Hadoop and MapReduce programming notion. They suggest predictive analytics prototypes for text classification, recommendation framework and decision support using open-source cloud-based solutions, using open source software packages, specifically Apache Hadoop, Hive and Mahout, all of which are designed to be scalable to big data.

Jai Prakash Verma et al. [25] proposed system that offers review or summary of text data obtainable on web for an educational foundation. The suggested method generates the group of choose reviews as a summary of all feedback of big data set. In this system, Sequential Pattern Mining Framework (SPMF) has been used and Elki tool for clustering analysis. In comparison with manually selected reviews and feedback, the results produced by these tools are shown in the various graphs that introduced in paper and considered satisfactory.

Li Chen and Feng Wang [17] have identified that there was a problem with the identification of preference similarity among reviewers. They recognized that for an accurate recommendation to be generated for the buyer, there was a need to identify the essential preference similarities between the buyer and the product reviewers. To try and resolve this issue, Chen and Wang proposed a novel clustering method. The method is founded on the LCRM (Latent Class Regression model). This model makes it possible for reviewers' preference similarity to be identified through the consideration of the overall ratings as well as the feature-level opinion values as they are presented in the textual reviews. Using the model, they derived the cluster-level preferences and reviewer-level preferences, which is compared against active buyer's recommendations. They then used an experiment, applying two data sets from the real world to test the proposed recommender algorithm. The laptop dataset and digital camera dataset were used. The outcome from the experiments was the superior performance of the LCRM based on accuracy which used specifically included; the reviewer-level preferences that were derived were very stable, reviewers clustering were performed effectively and the recommendations generated for the buyer were more accurate regardless of how complete the buyer's preferences were [17].

Hongyan Liu and others [26] identified the low accuracy problem that traditional recommendation methods experienced due to the sparseness of data. They proposed a novel recommendation algorithm, called PORE. The proposed method identified the user's preferences by analyzing the difference between the user's ratings and opinions. It considers implicit opinion and explicit ratings, which helps in addressing the issue with data sparseness. The adverb-based opinion-feature

extraction method is used for the extraction of opinions and the features from the user reviews online. This method would enhance the accuracy of extraction. To evaluate the algorithm's performance, the researchers conducted an empirical study, which is based on the extraction methods. The study would be carried out on a real dataset of an online restaurant review and it would be in Chinese. The purpose of the study is to create a recommendation system for the restaurant as well as demonstrate the proposed method's effectiveness. The results from the experiments carried out show that in comparison to the already existing methods, the extraction method used here has a better performance and it can extract the most features and opinions. The results also show the capabilities of the recommender algorithm in dealing with data sparseness as well as its better accuracy and efficiency in making predictions [26].

Sandra Esparza and others [27] identified that Real-time web (RTW) services usually provide its users with a chance to put across their interests and opinions. RTW data is unstructured and is therefore not recognized in recommender systems. However, they recognized that RTW data can contain consumer reviews that are very useful with regards to the reviewed services, products and brands. Therefore, they proposed an approach that could utilize RTW data (Twitter-like short-form messages) for a product recommendation, where the RTW data is the source for retrieval information and indexing. The approach proposed making recommendations through the use of micro-blogging information. The researchers used four datasets of different products retrieved from the Blippr service to evaluate the micro-blog reviews. The results indicate that despite their use of language being inconsistent and messages being short form, the micro-blogging messages are capable of providing useful recommendations. The approach also proved, from the evaluation and based on accuracy and coverage, that the approach outperforms the traditional collaborative filtering approach [27].

Our contribution: previous studies have considered activity and fitness models from different aspects, but still there is a gap in term of modeling fitness applications based on sequential information such as heart rate. This inspires us to propose a fitness recommendation model based on as heart rate sequential information.

IV. Data Description

A. FitRec Datasets Description

Datasets of FitRec [28] include sports records of user from Endomondo. Data contains several sources of sequential sensor data like rate of heart, GPS, speed, as well as type of sport, weather condition (such as humidity and temperature) and gender of user. These sets of data are collected from wearable devices (e.g. Apple Watch, Fitbit, etc.) and such data are heterogeneous, noisy, diverse in scale and resolution, and have complex interdependencies. To clean these data, heuristics are utilized by filtering out those unnatural workout samples like too large magnitude, timestamps which are mismatching, sudden changes in coordinates of GPS. We also derive several variables such as distance and speed from the measurements.

Type of Data:

- Measurements data like timestamps, distance, speed, and heart rate,
- Contextual data like altitude, latitude, longitude, sport, user identity, and gender

Table I describe the datasets features, it contains 5 columns. The "variable" column represent the names of the features, while the type of each feature is represented in the "data type" column and the measurement unit is represented in "Unit" column. The Definition column describe the features. As shown in the table, the features is categorized into Measurement and Contextual.

B. Exploratory Data Analysis (EDA)

Table II shows the statistics of the dataset in respect of the total number of sports, workouts, genders, speed, and heart rate. As shown in the table, the majority of participants is males, while the minority is females. There are also **1185** unknown gender participants. The speed mean $\pm$SD = (20.962 $\pm$ 8.483 MPH) and the speed range is (min =0.0, max=74.859 MPH). The speed standard deviation is small, meaning that speed values of the participants are centralized around the mean. The min value of speed rate illustrates that there are sports that do not require fast movement. The heart rate mean $\pm$SD = (138.7 $\pm$ 18.961 BPM) and the heart rate range is (min = 0.0, max = 239.0 BPM). The large value of the standard deviation of the heart rate indicates that the heart rates values are scattered, meaning there are a notable difference between the users' heart rates. This could be due to their gender, weight, diseases, etc. Table III clarifies the differences in heart rate between male and females for each sport. These differences confirm that, the heart rates do not follow a specific pattern and they are varying from participant to another according to his/her personalized health. Therefore, the heart rate issues should be considered during the workouts for each participant alone, this is the value of our proposed recommendation system. There are a markable differences between the heart rates and speed rates between the females and males as shown in Table 3, Fig. 2, and 3. The females heart rates are usually higher than the males specially when they skate or play gymnastics. At the same time, the standards show that males are much faster than the females.

Fig. 4 and 5 shows that, generally, males' and females' heart rates are positively correlated with the altitude; this is natural phenomena since climbing heights takes great efforts. However, their heart rates are negatively correlated with the speed due to athletic heart syndrome [3]. It is a phenomenon that explains the natural changes that take place in the hearts of people participating in vigorous athletic training. Eventually, our recommendation model will address these two issues, it recommends participants the suitable sports. Plus, the participant's heartbeat will be observed during the workout to notify him if his heartbeat reach dangers threshold to take suitable action, such as slowing done or changing his path if there are altitudes.

V. The Proposed Framework

First when person open the application at first time, there is some information must record it in app, like gender, and ID (id may be generated by app), and app then will join with watch. Participants wear smart watch and start to do workout. During workout, some important data will be generated like speed,

TABLE I. FITREC DATA DESCRIPTION

	Variable	Data type	Unit	Definition
Measurement	Heart Rate	Array¡bigint¿	Beat per Minute (BPM)	The speed of heartbeats for a specific user.
	Timestamp	Array¡bigint¿	Unix Timestamp	Tracking the time of the specific event occurred in form of seconds
	Distance	Array¡double¿	Mile	Measured by calculating the haversine formula which considers longitudes and latitudes to determine the distance among certain points.
	Speed	Array¡double¿	Mile per Hour(MPH)	Mesures the movement of an object which is calculated by dividing the distance over time.
Contextual	ID	Bigint	-	The ID for a specific workout.
	UserID	Bigint	-	The user's national ID
	Sport	String	-	The user makes physical exertion and practicing different sports such as bike riding, walking, running, skating, and so on.
	Gender	String	Male,Female	The user can be male or female.
	Altitude	Array¡double¿	Meter	The perpendicular distance from the specific object to the earth's surface.
	Longitude	Array¡double¿	Degree	The geographic coordinations which determine the exact location on the earth's surface.
	Latitude	Array¡double¿	Degree	
	URL	String	-	The link of the user's account

TABLE II. EDA

Feature	Measures	
Sport	43	
Workout	167,783	
Gender	Male	156717
	Female	9881
	Unknown	1185
Speed	Max	74.859
	Min	0.0
	Mean	20.962
	Standard Deviation	8.483
Heart Rate & Standard Deviation & 18.961	Max	239.0
	Min	0.0
	Mean	138.7

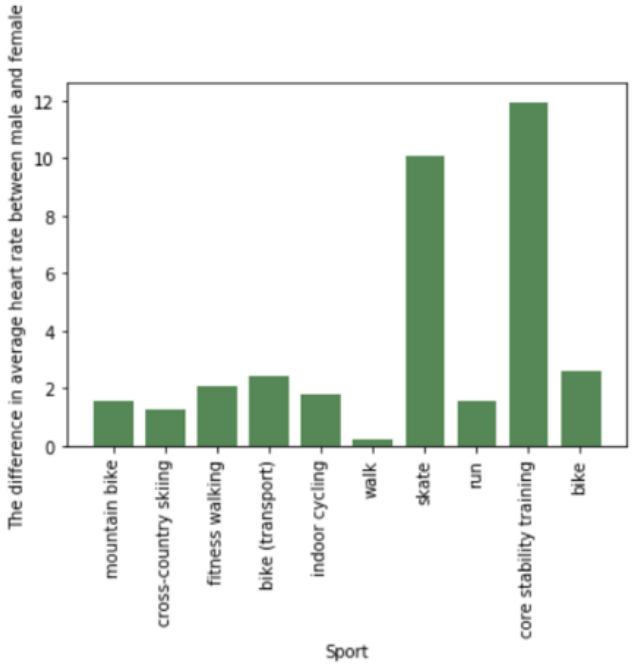

Fig. 2. The Differences in Average Heart Rate between Males and Females.

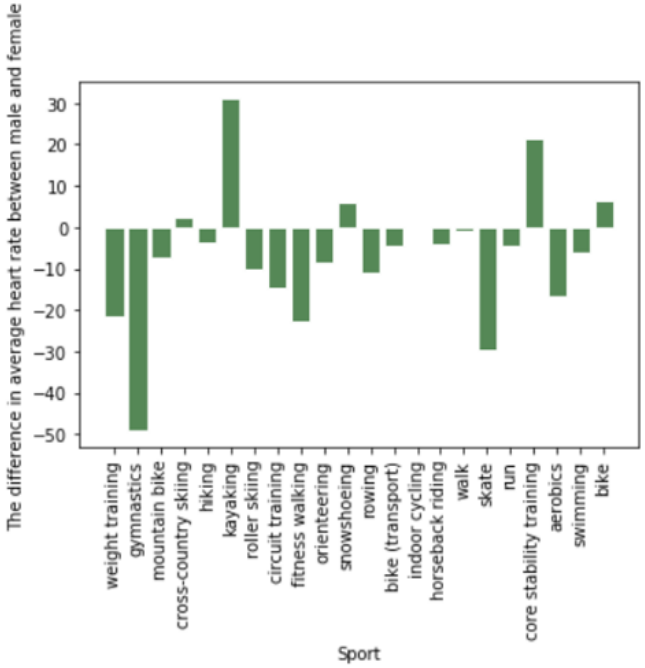

Fig. 3. The Differences in Average Speed Rate between Males and Females.

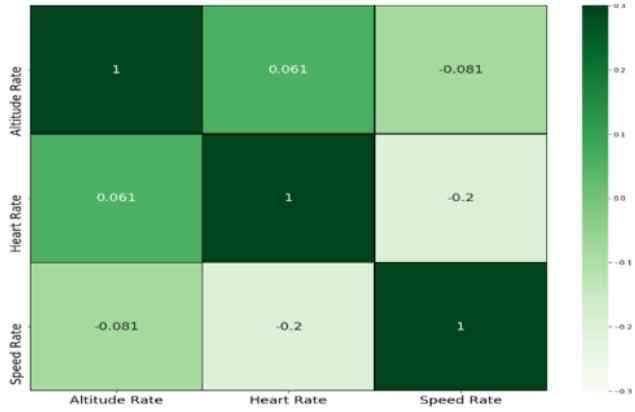

Fig. 4. Male Pearson Correlation of Features.

heart rate and altitude depend on place of person (longitude and latitude). Based on some processes in sequential steps (explained in framework) the application will recommend the suitable workout for person. Fig. 6 shows the system overview. Framework is divided into three layers:

A. Application Layer

1) Data gathering: The data was collected by first: registering in the application, second through workout where data is measured by wearable device (smart watch). These data will

TABLE III. INFORMATION FROM THE DATASET ABOUT SPEED AND HEART RATE FOR EACH SPORT

Sport	Heart Rate			Speed		
	Male	Female	Diff_Heart	Male	Female	Diff_Speed
Mountain bike	135.334108	142.596029	-7.261921	20.403874	18.785404	1.618470
Cross-country skiing	138.137305	135.610060	2.527246	14.566997	13.248498	1.318498
Citness walking	104.853535	127.391624	-22.538089	9.420005	7.262868	2.157137
Bike (transport)	126.564361	131.209013	-4.644652	23.519709	21.048684	2.471026
Indoor cycling	133.235132	133.297185	-0.062054	27.456905	25.587658	1.869247
Walk	102.351934	103.368996	-1.017062	6.477019	6.214970	0.262049
Skate	118.626881	148.212842	-29.585961	29.880860	19.772502	10.108359
Run	146.752812	151.428408	-4.675596	11.681297	10.079193	1.602104
Core stability training	131.464555	109.941488	21.523067	17.818819	5.822604	11.996215
Bike	133.692914	127.367465	6.325449	27.192178	24.523801	2.668377
Weight training	107.399738	129.021000	-21.621262	—	—	—
Gymnastics	104.550400	153.598000	-49.047600	—	—	—
Hiking	110.213909	114.032450	-3.818541	—	—	—
Kayaking	123.132933	92.156000	30.976933	—	—	—
Roller skiing	129.653974	139.967600	-10.313626	—	—	—
Circuit training	118.425579	133.212308	-14.786729	—	—	—
Orienteering	146.885183	155.238818	-8.353635	—	—	—
Snowshoeing	127.474727	121.362000	6.112727	—	—	—
Rowing	131.076264	142.158111	-11.081847	—	—	—
Horseback riding	138.429000	142.704000	-4.275000	—	—	—
Aerobics	141.228800	157.691000	-16.462200	—	—	—
Swimming	118.511500	124.460667	-5.949167	—	—	—

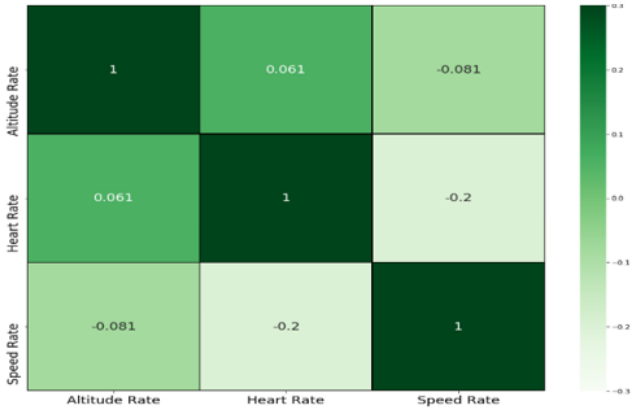

Fig. 5. Female Pearson Correlation of Features.

store in in Repository Layer.

B. Processing Layer (Pre-processing, comparison, prediction, clustering and recommendation)

1) Recommendation: depending on the characteristics of person, app will recommend a suitable workout for him to reach the goal of workout without any bad effects on participant's heart.

2) Clustering: The system will proceed with K-means clustering which is the most straightforward and well-known unsupervised machine learning algorithm. The purpose of applying K-means is to aggregate similar data points into a set of clusters. There are several advantages behind using this technique which are simplicity of its implementation as it has only two steps. Also, scaling to big data sets as well as the algorithm works in high speed. Furthermore, the technique ensures convergence and the ease to adjusts to new examples [29]. In essence, the similar participants are clustered based on their recorded measures of heart rate speed and altitude of all their previous workouts per sport.

3) Comparison (stream processing): The system check the participants' heart rate and altitude. It sends an alarm when detecting unusual heart rate and recommend the participant with the suitable action.

4) Feature selection: The only workouts that have speed, altitude and heart rate are selected. So, from the overall type of workout, 15 workouts are selected. Average speed, average heart rate, and average altitude are selected to develop the cluster model.

5) Pre-processing:

Encoding

- Timestamp encoded into 4 periods, here we will divide 24 hours into 4 period [0 from 12:00 a.m. to 5:00 a.m.], [1 from 6:00 a.m. to 11; 00 a.m.] , [2 from 12:00 p.m. to 9:00 p.m.] and last period [3 from 6:00 p.m. to 11:00 p.m.] . We chose to divide the day into four periods that came from our knowledge of some of the current clubs with high competence in dividing the sport periods into 4 periods.
- Encoding is the process of converting categorical variables into numerical. Binary encoding technique was used to encode the verified and official features, '0' represents the account which is not verified or official the account. Gender data will convert to 0 and 1 (Male =0, female=1).
- The sport feature was encoded using one-hot encoding; it is a type of encoding method [30]. It points to dividing the column which includes numerical categorical data to multiple columns basing on the categories number existing in that column, each column contains "0" or "1" matching to which column it has been placed. The sports which recorded with (speed, altitude and heartrate) only considered. So, the overall total of workout is 15.

Aggregation

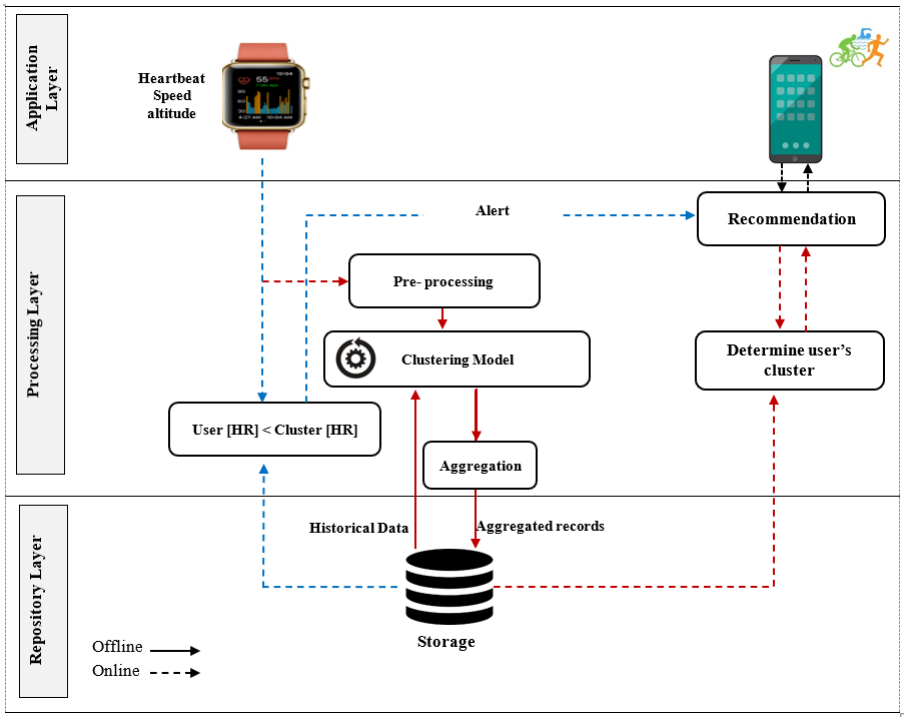

Fig. 6. System Overview.

To implement this step, there is a need first to understand the nature of inputted data for different attributes like heart rate, speed and altitude which each of them is taken as an array of successive values. First, the system will calculate the average heart rate, speed and altitude for each user per workout. Then, it will aggregate all these averages and calculate the overall average for each participant's recored per sport.

Missing value

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values. The missing values will be replaced by the average values of the its feature, depending on the gender.

Vectorization

It is the way to represent the data set into a set of vectors. Many different techniques are available, the simplest one is count vectorizer. This process aims to discover some patterns or relationships within the targeted data [31]. The system will set up co-ordinate vectors for each user prior implementing the K-mean. in detail, these co-ordinate vectors will be assembled dependent on the marks which expresses an hour period that is divided into 4 marks with values of (0,1,2,3). Each of these marks will have a count for each user Id and it will be updated constantly but the mark that has not existed will be assigned with zero. Also, gender is taking 0 for a male and 1 for a female. Furthermore, calculating the average of different attributes such as speed, heart rate and altitude grouping by workouts in order to constitute vectors of m sports.

Scaling

The dataset consists of various independent variables. The value ranges for those different variables are almost certain to vary widely and to have their own scale. The value ranges of all variables should have a similar scale to promote each variable in contributing proportionately to the ultimate result [32]. The scaling feature is almost confined between 0 and 1 which are viewed as the minimum and maximum values. Accordingly, the min-max scaling methods will be used.

C. Repository Layer (Historical Data)

The dataset will be stored in Mongodb in JOSN format. MongoDB is an open source NoSQL database system built for storing semi-structured data. Plus, the data will be stored on the MongoDB after the clustering process. This structure provides flexibility and adaptability in a way that enables the programmer to create classes and objects instead of having a traditional row/column model [33].

VI. Experiment, Results and Evaluation

A. Clustering

The k-mean cluster was used to split the similar participants together according to their heart rate, speed rate, and altitude. One of the properties of the k-mean clustering algorithm is that the optimal number of clusters can be defined before the clustering process. The well-known elbow method was applied to determine the optimal clusters number [1]. The elbow method assumes that the optimal clusters number must produce small inertia, or total intra-cluster variation. To identify the optimal number, the k must be selected at the "elbow" point, after the inertia /distortion begins decreasing in a linear fashion. Fig. 7 shows the optimal number to cluster the data is k=5. Table IV shows the elbow points' value in k (2:10).

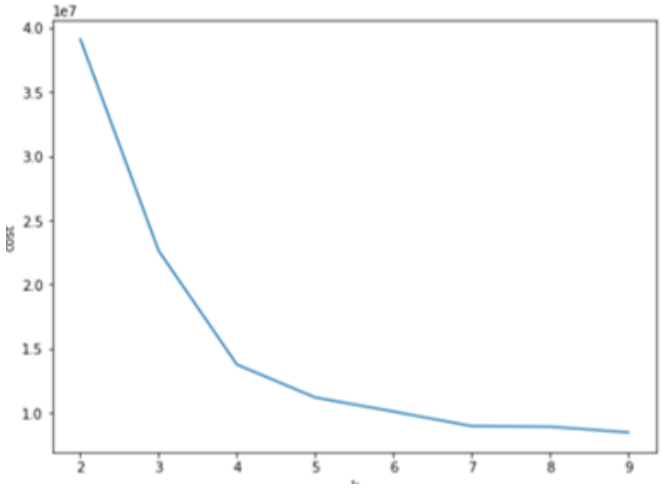

Fig. 7. The Changes on Elbow Values of k (2:10).

TABLE IV. THE ELBOW VALUES OF K (2:10)

K	elbow
2	0
3	0
4	22627014.820202056
5	39113248.973619014
6	13754284.605939962
7	11188781.682890285
8	10090998.661182716
9	8957156.883945007
10	8911367.440781211

The k-mean algorithm was trained on 70% of the data, then it was used to predict the remaining 30% sets of data. The algorithm performance is 0.73128 based on Silhouette Coefficient score. Silhouette Coefficient is a metric used to evaluate the clustering technique [2]. Its value varies between -1 and 1, where 1 indicates that clusters are distinct and clearly differentiated from each other, while -1 indicates the clusters are allocated in the wrong way, and 0 indicates that the clusters are indifferent [2]. Table V shows the properties of each cluster. As shown on the Table V, each cluster was given a name.The majority of the participants are belonged to the Moderate fit athlete cluster (the the account of the participant equal 16989), while the minority to the Unusual (the the account of the participant =176). In General, there is a big discrepancy between the number of participants in the Moderate fit athlete cluster and the other clusters. The Unusual cluster which has a very few participants may contain unusual observations. This appear when comparing its very high-altitude average which equal (1510.427) with other clusters altitude average. The participants who belongs to the Most fit athlete cluster achieved much higher altitude rate than the participants in the Moderate fit athlete cluster. At the same time, they recorded 130.553 which considered close to the normal heart rate. For adults, the average resting heart rate varies between 75 and 170 beats per minute during exercises [3]. A lower heart rate at rest usually means more successful heart function and greater cardiovascular health [3]. So, the participants included in Most fit athlete cluster are the most trained athlete. The participants who belong to the cluster Above-average fit and Adventurers recorded high altitude and heart rates, this metrics seems normal as most of them practiced hiking, mountain bike, and cross-country skiing as shown in Fig. 8. It seems that the participants in cluster Adventurers reached much higher peaks than those belonged to cluster Above-average fit, increasing their heart rate. Finally, there is no markable differences between the average speed of all the five clusters, all the users' speed rates are ranged in (17.35 and 19. 66).

B. Comparison

Based on the center of disease control and prevention [4], and The Heart Foundation charity organization [5], the intensity of exercises can be measured using heart rate measures. The heart rate should be 50% to 70% of the estimated maximum heart rate for exercising at a low to moderate intensity, but during robust exercises it's about 70-85% of maximum. Table VI shows a heart rate measures according to the different ages. For example, if the participant is 20 years old, the maximum estimated heart rate for exercising is 140 in case of intensity exercises and 170 in case of robust exercises. Because of the dataset is limited in term of the age, we will consider the average of all the heart rates as the heart rate threshold, resulting in 122.5 for intensity exercises and 149 for robust exercises.

According to the average altitude which achieved by participial for each cluster. The altitude of the Moderate fit athlete cluster (55) is consider as threshold of the altitude when the participant does not join to one of the robust sports since it represents the majority of the participants. On the other hand, the altitude of the Most fit athlete cluster (213) is consider a threshold of the altitude when the participant joins to one of the robust sports since it represents a high portion of the participants who exercise robust sports. The robust sports are mountain bike, cross-country skiing, hiking, orienteering, run, and bike. Fig. 9 represent the flowchart of the comparison phase. To implement the comparison phase of the system, the streaming data was simulated using pyspark. Fig. 9 shows the flowchart of the recommendation system.

VII. CONCLUSION

This research aimed at building a personalized fitness recommendation application based on sequential information such as heart rate. Specifically, the proposed application will recommend a proper workout type for the one who care to avoid the negative consequences that may affect his health. This target has been achieved through learning the historical workout sequences that is integrated in our proposed framework and applied K-means clustering algorithm to group similar users according to their heart rate, speed rate, and altitude. After the elbow method was applied to determine the optimal clusters number, the result shows that the optimal number to cluster the data is k=5. Furthermore, this clustering algorithm has been trained on 70% of the chosen dataset and then it was utilized to predict the remaining 30% sets of data. The performance of the applied algorithm is 0.73128 based on Silhouette Coefficient score which is indicated that that clusters are distinct and clearly differentiated from each other. This research aimed at building a personalized fitness recommendation application based on sequential information such as heart rate. Specifically, the proposed application will recommend a proper workout type for the one who care to avoid the negative

TABLE V. THE OVERVIEW INFORMATION OF ALL THE FIVE CLUSTERS

Cluster	Cluster Name	Count	%	Avg(speed)	Avg(heart)	Avg (Altitude)	Sport
1	Moderate fit athlete	16989	67.4	19.31756492094167	136.92687567988688	55.29274236638334	mountain bike, cross-country skiing, hiking, kayaking, roller skiing, fitness walking, orienteering , bike (transport, indoor cycling, horseback riding, walk, skate, run, core stability, bike.
2	Unusual	176	0.0069	17.59108255126259	133.46338172392413	1510.426606091586	mountain bike, indoor cycling, walk, run, bike
3	Most fit athlete	5872	23.3	17.35909711437589	130.5529807309039	213.9624622595631	mountain bike, cross-country skiing, hiking, roller skiing, orienteering, bike (transport), indoor cycling, walk, skate, run, core stability, bike.
4	Above-average fit	1592	2.5	17.8432634544732	133.93142304348586	360.10695197968624	mountain bike, cross-country skiing—, hiking, roller skiing, fitness walking, orienteering, bike (transport), indoor cycling, horseback riding, walk, skate, run, core stability, bike.
5	Adventurers	568	2.2	19.66543325571946	139.49961339923698	822.0379459052019	mountain bike, cross-country skiing, roller skiing, indoor cycling, core stability, run, bike

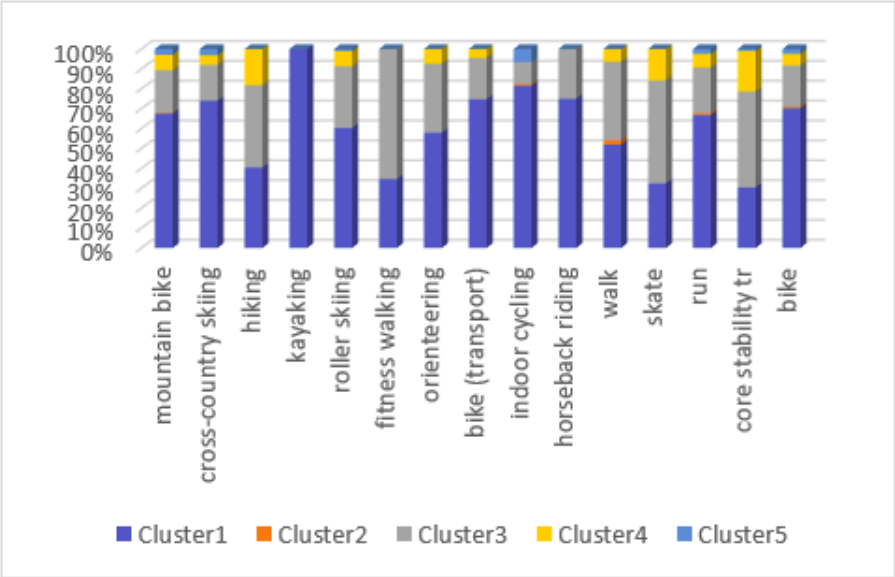

Fig. 8. The Distribution of Sports in each Cluster.

TABLE VI. ESTIMATED HEART RATES FOR EXERCISING

	Intensity Exercises	Robust Exercises
70	75 - 105 bpm	75-128 bpm
60	80 - 112 bpm	80-136 bpm
50	85 – 119 bpm	85-145 bpm
40	90-126 bpm	90-153 bpm
30	95-133 bpm	95-162 bpm
20	100-140 bpm	100-170 bpm

consequences that may affect his health. This target has been achieved through learning the historical workout sequences that is integrated in our proposed framework and applied K-means clustering algorithm to group similar users according to their heart rate, speed rate, and altitude. After the elbow method was applied to determine the optimal clusters number, the result shows that the optimal number to cluster the data is k=5. Furthermore, this clustering algorithm has been trained on 70% of the chosen dataset and then it was utilized to predict the remaining 30% sets of data. The performance of the applied algorithm is 0.73128 based on Silhouette Coefficient score which is indicated that that clusters are distinct and clearly differentiated from each other. Furthermore, this research is achieved a comparison for the recommendation system by simulating the streaming data using pyspark. In near future, some extra work can be done by integrating a third party that recommends the most appropriate path for the user instead of giving him a general message to change his route. Also, the research needs to employ more features or characteristics related to chronic diseases to have more logical reasons for grouping the clusters.

REFERENCES

[1] Y. G. Wibowo and B. Indrayana, "Sport: A review of healthy lifestyle in the world," *Indonesian Journal of Sport Science and Coaching*, vol. 1, no. 1, pp. 30–34, 2019.

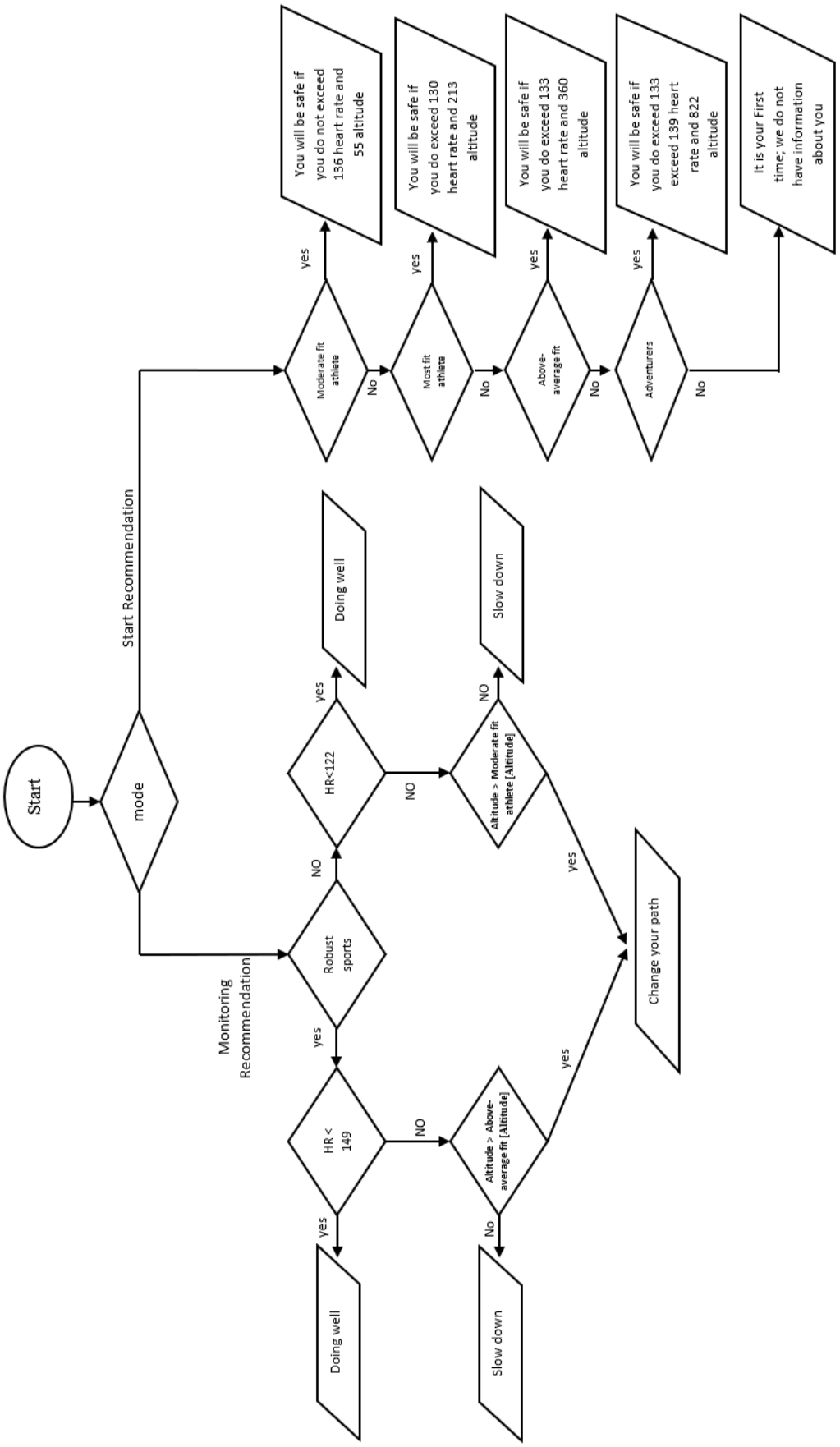

Fig. 9. Flowchart of the Recommendation Phase.

[2] J. Ni, L. Muhlstein, and J. McAuley, "Modeling heart rate and activity data for personalized fitness recommendation," in *The World Wide Web Conference*, 2019, pp. 1343–1353.

[3] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, no. 2012, pp. 1–16, 2012.

[4] G. George, E. C. Osinga, D. Lavie, and B. A. Scott, "Big data and data science methods for management research," 2016.

[5] P. Maroufkhani, M.-L. Tseng, M. Iranmanesh, W. K. W. Ismail, and H. Khalid, "Big data analytics adoption: Determinants and performances among small to medium-sized enterprises," *International Journal of Information Management*, vol. 54, p. 102190, 2020.

[6] S. Guha and S. Kumar, "Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap," *Production and Operations Management*, vol. 27, no. 9, pp. 1724–1735, 2018.

[7] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," *Information Systems and e-Business Management*, vol. 16, no. 3, pp. 547–578, 2018.

[8] B. M. Balachandran and S. Prasad, "Challenges and benefits of deploying big data analytics in the cloud for business intelligence," *Procedia Computer Science*, vol. 112, pp. 1112–1122, 2017.

[9] H. Isah, T. Abughofa, S. Mahfuz, D. Ajerla, F. Zulkernine, and S. Khan, "A survey of distributed data stream processing frameworks," *IEEE Access*, vol. 7, pp. 154 300–154 316, 2019.

[10] V. Gurusamy, S. Kannan, and K. Nandhini, "The real time big data processing framework: Advantages and limitations," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 12, pp. 305–312, 2017.

[11] S. Shahrivari, "Beyond batch processing: towards real-time and streaming big data," *Computers*, vol. 3, no. 4, pp. 117–129, 2014.

[12] R. Burke, A. Felfernig, and M. H. Göker, "Recommender systems: An overview," *Ai Magazine*, vol. 32, no. 3, pp. 13–18, 2011.

[13] S. Vinodhini, V. Rajalakshmi, and B. Govindarajalu, "Building personalised recommendation system with big data and hadoop mapreduce," *International Journal of Engineering Research and Technology*, vol. 3, no. 4, pp. 2310–2316, 2014.

[14] R. M and K. R. R. Raman, "Recommendation system: A big data application," *International Journal of Emerging Trends in Science and Technology 2348-9480*, vol. 3, pp. 39 –46, 09 2016.

[15] W. Serrano, "Intelligent recommender system for big data applications based on the random neural network," *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 15, 2019.

[16] J. P. Verma, B. Patel, and A. Patel, "Big data analysis: recommendation system with hadoop framework," in *2015 IEEE International Conference on Computational Intelligence & Communication Technology*. IEEE, 2015, pp. 92–97.

[17] L. Chen and F. Wang, "Preference-based clustering reviews for augmenting e-commerce recommendation," *Knowledge-Based Systems*, vol. 50, pp. 44–59, 2013.

[18] F. Mehmood, S. Ahmad, and D. Kim, "Design and development of a real-time optimal route recommendation system using big data for tourists in jeju island," *Electronics*, vol. 8, no. 5, p. 506, 2019.

[19] C. Anderson, I. Hübener, A.-K. Seipp, S. Ohly, K. David, and V. Pejovic, "A survey of attention management systems in ubiquitous computing environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–27, 2018.

[20] A. K. Chowdhury, A. Farseev, P. R. Chakraborty, D. Tjondronegoro, and V. Chandran, "Automatic classification of physical exercises from wearable sensors using small dataset from non-laboratory settings," in *2017 IEEE Life Sciences Conference (LSC)*. IEEE, 2017, pp. 111–114.

[21] C. Zhang, H. Liang, and K. Wang, "Trip recommendation meets real-world constraints: Poi availability, diversity, and traveling time uncertainty," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 1, pp. 1–28, 2016.

[22] B. Loepp and J. Ziegler, "Recommending running routes: framework and demonstrator," in *Workshop on Recommendation in Complex Scenarios*, 2018.

[23] K. Bafna and D. Toshniwal, "Feature based summarization of customers' reviews of online products," *Procedia Computer Science*, vol. 22, pp. 142–151, 2013.

[24] K. Hammond and A. S. Varde, "Cloud based predictive analytics: text classification, recommender systems and decision support," in *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013, pp. 607–612.

[25] J. P. Verma, B. Patel, and A. Patel, "Web mining: opinion and feedback analysis for educational institutions," *International Journal of Computer Applications*, vol. 84, no. 6, 2013.

[26] H. Liu, J. He, T. Wang, W. Song, and X. Du, "Combining user preferences and user opinions for accurate recommendation," *Electronic Commerce Research and Applications*, vol. 12, no. 1, pp. 14–23, 2013.

[27] S. G. Esparza, M. P. O'Mahony, and B. Smyth, "Mining the real-time web: a novel approach to product recommendation," *Knowledge-Based Systems*, vol. 29, pp. 3–11, 2012.

[28] [Online]. Available: https://sites.google.com/eng.ucsd.edu/fitrec-project/home

[29] A. automaticaddison, "Advantages of k-means clustering," Aug 2019. [Online]. Available: https://automaticaddison.com/advantages-of-k-means-clustering/

[30] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?" *Emerging Markets Finance and Trade*, pp. 1–11, 2020.

[31] G. Goel, "Why data is represented as a 'vector' in data science problems?" Jul 2020. [Online]. Available: https://towardsdatascience.com/why-data-is-represented-as-a-vector-in-data-science-problems-a195e0b17e99

[32] S. Yıldırım, "Data preprocessing with scikit-learn: Standardization and scaling," Jun 2020. [Online]. Available: https://towardsdatascience.com/data-preprocessing-with-scikit-learn-standardization-and-scaling-cfb695280412

[33] "What is mongodb? introduction, architecture, features & example." [Online]. Available: https://www.guru99.com/what-is-mongodb.html