Advance Big Data Analysis and Modeling on Workout FitRec Data

Nihar Patel
Schaefer School of Engineering & Science
Stevens Institute of Technology
Hoboken, NJ-07087
Email: npatel17@stevens.edu

Abstract—**In today's fast-paced world, maintaining good health is crucial for improving the quality of life and enhancing productivity. Among various strategies, daily workouts have emerged as an effective approach to achieving fitness and well-being. The FitRec project delves into the analysis of real workout data collected from individuals to uncover valuable insights into fitness trends, workout efficiency, and personalized recommendations. Leveraging big data technologies such as PySpark, this paper presents a comprehensive exploration of workout patterns, intensity levels, and their impact on health outcomes. By analyzing large-scale datasets, this research aims to bridge the gap between raw fitness data and actionable insights for health-conscious individuals and fitness professionals.**

1. Introduction

In recent years, the emphasis on personal fitness has grown exponentially, driven by the increasing awareness of health and well-being. Regular workouts, combined with proper dietary habits, are widely regarded as effective means to maintain physical fitness and prevent chronic diseases. With the advent of wearable fitness trackers and mobile applications, a significant volume of workout data is being generated daily. This vast amount of data presents an opportunity to analyze and extract meaningful insights that can guide users toward better fitness practices.

The FitRec project leverages the capabilities of big data technologies, specifically PySpark, to process and analyze workout datasets efficiently. This project aims to study the behavioral patterns, workout preferences, and health metrics of individuals using fitness data collected from real-world sources. By conducting a deep analysis, the project seeks to offer actionable insights, such as identifying effective workout regimens, predicting fitness goals, and enhancing personalized fitness recommendations.

This paper outlines the methodologies employed in the FitRec project, focusing on the integration of big data frameworks to handle the scale and complexity of the datasets. The findings aim to provide valuable contributions to the fitness industry, researchers, and individuals striving to achieve their health goals.

2. Background

In recent years, the intersection of health and technology has seen rapid growth, with wearable devices and fitness applications becoming integral to daily life. These tools generate a wealth of data, offering opportunities to analyze and understand workout patterns, health behaviors, and fitness outcomes on a large scale. However, the challenge lies in effectively managing, processing, and extracting meaningful insights from this vast and complex data.

The FitRec project was initiated to address these challenges by providing a comprehensive dataset that represents real-world workout activities. Collected from diverse users over extended periods, the dataset includes details such as activity types, duration, intensity, calories burned, and heart rate. By combining demographic information with workout data, the FitRec project creates a unique resource for studying individual and group fitness behaviors.

The FitRec dataset is particularly valuable because it mirrors real-world scenarios, where data is often messy, incomplete, or noisy. This presents an opportunity for researchers to explore advanced data processing techniques and build models that are robust and scalable. Furthermore, the dataset allows for the application of big data technologies like PySpark to process and analyze large volumes of data efficiently, facilitating insights that were previously inaccessible due to technological constraints.

By enabling detailed studies of fitness trends and behaviors, the FitRec project contributes to the broader goals of improving public health, designing personalized fitness recommendations, and advancing the capabilities of fitness tracking systems. The project not only empowers individuals to achieve their fitness goals but also serves as a benchmark dataset for researchers developing innovative solutions in sports science, health analytics, and machine learning.

2.1. Existing Methods

Predicting heart rate and analyzing workouts require sophisticated methods that leverage physiological data and advanced computational techniques. Over the years, several approaches have been developed, ranging from traditional statistical methods to advanced machine learning and deep learning models.

Statistical and regression-based models rely on predefined relationships between physiological parameters such as age, weight, exercise intensity, and heart rate. The Karvonen Formula is a widely used approach to estimate target heart rate during workouts based on resting and maximum heart rates[4]. Similarly, linear regression establishes a direct relationship between variables like workout duration and heart rate changes, while generalized linear models (GLMs) allow more flexibility in modeling non-linear relationships between predictors. However, these models often assume simplistic relationships that may not fully capture the complexity of physiological responses during varied workout types.

Signal processing techniques, on the other hand, analyze raw physiological data such as ECG signals or photoplethysmograms (PPG) obtained from wearable devices. Methods like Fourier Transform are used to analyze periodic components of heart rate data, while wavelet transforms are applied to assess heart rate variability (HRV) during workouts, capturing changes across frequency bands[3]. Time-series analysis techniques, such as autocorrelation and spectral analysis, help identify trends and patterns in heart rate data over time. Despite their usefulness, these methods often require high-quality, noise-free data, which can be challenging to acquire in real-world scenarios.

Machine learning approaches have emerged as powerful tools for modeling the non-linear relationships inherent in heart rate and workout data. Algorithms like Support Vector Machines (SVMs) classify workout types based on heart rate patterns, while ensemble methods such as random forests and gradient boosting combine multiple features to predict heart rate and analyze workout efficiency. Simpler methods like k-Nearest Neighbors (kNN) predict heart rate trends by finding similarities in historical data. These models adapt well to varying datasets and can incorporate multiple input features, including demographics, activity intensity, and environmental conditions, making them robust for diverse applications[2]arxivXGBoostScalable.

Deep learning has further transformed the field by enabling the analysis of large-scale data and uncovering intricate patterns. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are designed to handle sequential data, making them ideal for predicting heart rate by capturing temporal dependencies. Convolutional Neural Networks (CNNs), often applied to raw signal data, excel in extracting spatial and temporal features, aiding in the classification of workout types and intensity levels. Autoencoders, on the other hand, are employed for dimensionality reduction and anomaly detection. While these methods excel in handling noisy, unstructured data, they often demand significant computational resources and large training datasets.

Hybrid approaches have shown promise in achieving higher accuracy and robustness by combining multiple methods. For instance, preprocessing noisy heart rate data using signal processing techniques, such as wavelet transforms, before feeding it into machine learning or deep learning models, enhances prediction accuracy. Multi-modal models, which integrate data from diverse sources like GPS, accelerometer, and heart rate monitors, further improve performance by providing a holistic view of workout and physiological data.

Despite these advancements, several challenges persist in the field. Data quality remains a major concern, as wearable data often suffers from noise, missing values, and inconsistencies. Personalization is another limitation, with many models failing to account for individual physiological differences, which reduces their effectiveness. Additionally, real-time analysis for low-latency predictions during workouts poses significant computational challenges[1].

The increasing availability of large-scale datasets, such as FitRec, and advancements in big data technologies like PySpark provide new opportunities to overcome these challenges. These innovations hold the potential to improve the scalability, personalization, and accuracy of methods for heart rate prediction and workout analysis, paving the way for robust and efficient systems in health and fitness.

3. About Data

The dataset used in this study is derived from the FitRec project, a large-scale dataset designed to support research on workout activity and fitness tracking. The data was collected from fitness tracking devices and applications used by a diverse population over an extended period. The dataset includes workout logs, physiological measurements, and metadata recorded during various physical activities such as running, cycling, walking, and other exercises. Data was anonymized to protect user privacy, ensuring compliance with ethical standards for data collection.

The FitRec dataset is a comprehensive collection encompassing 6.11 GB of data. It includes records from 1,104 users, covering 253,020 workout sessions and over 11 million individual data points, making it one of the largest publicly available fitness datasets.

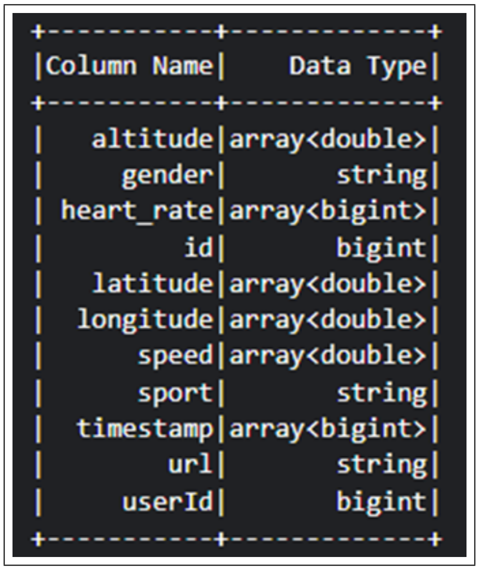
Each workout session in the dataset is represented as a record with multiple fields capturing a comprehensive view of the activity. Key fields in the dataset include:

- **User ID**: An anonymized identifier for each user, ensuring privacy while allowing for longitudinal analysis of individual workout patterns.
- **Activity Type**: A categorical variable specifying the type of physical activity performed, such as running, cycling, or walking.
- **Start Time and End Time**: Timestamps indicating the duration of the activity, enabling the analysis of temporal trends and workout intensity.
- **Heart Rate Data**: A time-series field capturing heart rate measurements during the workout, recorded at regular intervals.
- **Calories Burned**: An estimate of the total energy expenditure during the session, calculated based on activity type and user-specific factors.
- **Distance Covered**: The total distance traveled during the workout, relevant for activities like running and cycling.

- **GPS Data**: Geospatial information recorded during the activity, enabling the analysis of routes and environmental influences on workout performance.
- **Step Count**: The number of steps taken during the session, commonly recorded for walking and running activities.
- **Elevation Gain**: The total elevation change during the workout, providing insights into workout intensity for outdoor activities.
- **User Attributes**: Demographic and physical attributes such as age, gender, weight, and height, used to personalize analysis and predictions.

The dataset is particularly valuable due to its size and diversity, encompassing millions of records across users with varying fitness levels, demographics, and workout preferences. This richness enables researchers to develop and validate models for heart rate prediction, workout classification, and fitness analytics with high accuracy and generalizability.

Data preprocessing steps are required to handle missing values, noise, and inconsistencies commonly present in wearable data. Additionally, normalization and transformation techniques are applied to prepare the data for machine learning and deep learning algorithms. The comprehensive nature of the dataset and the breadth of its features make it an ideal choice for exploring advanced analytics in the domain of health and fitness.



```
+-----------+------------+
|Column Name|   Data Type|
+-----------+------------+
|   altitude|array<double>|
|     gender|      string|
| heart_rate|array<bigint>|
|         id|      bigint|
|   latitude|array<double>|
|  longitude|array<double>|
|      speed|array<double>|
|      sport|      string|
|  timestamp|array<bigint>|
|        url|      string|
|     userId|      bigint|
+-----------+------------+
```

Figure 1. Fields and its data types

4. Exploratory Data Analysis

In the exploratory data analysis (EDA), we compared various variables such as users and sports, gender and sport, and user and heart rate. Additionally, we analyzed single-variable distributions, such as location, to gain insights into workout trends and user behavior.

4.1. Users vs. Sports

Analyzing the distribution of users across different sports revealed significant insights into their preferences and activity levels. Among the sports, **running**, **biking**, and **mountain biking** emerged as the most popular, contributing substantially both in terms of user participation and total workouts. On a per-user basis, the contributions of these sports are as follows: running accounts for **25.1%**, biking for **23.1%**, and mountain biking for **9.8%**.

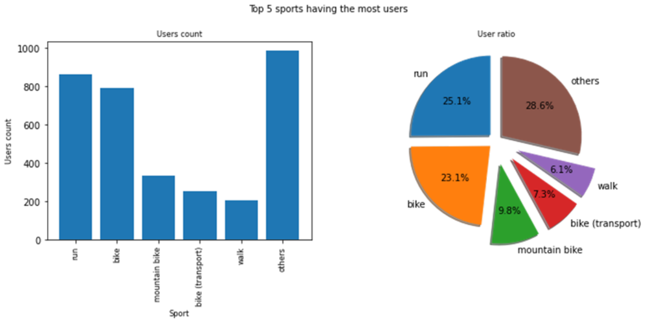


Figure 2. Sports based on Users

However, when examining the total number of workouts recorded in the dataset, running accounts for a much higher proportion at **46.6%**, followed by biking at **38.7%**, and mountain biking at **5.3%**.

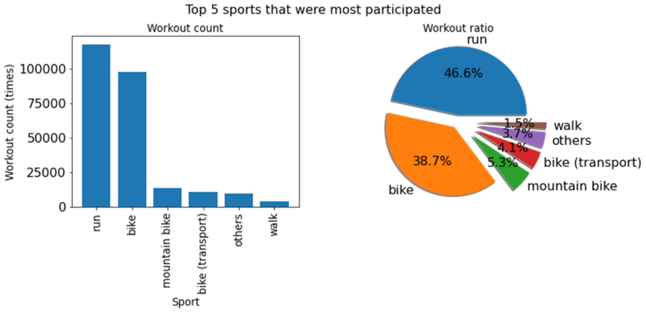


Figure 3. Sports based on Workouts

These findings highlight the dominant role of running and biking in the dataset, both in terms of user preference and workout volume. The disproportionate contribution of these sports to the total number of workouts suggests that users tend to engage in running and biking more frequently than other activities, underscoring their popularity and accessibility as fitness activities. This observation can serve as a key factor in model development, particularly in tasks like workout classification or activity prediction, where these sports form the majority of the dataset.

4.2. Gender and Sports

The analysis of gender-based participation in sports revealed notable patterns in user engagement. Out of the total **1,104 users**, **822 users** participated in more than one sport, highlighting the diversity of activities undertaken by the majority of individuals. On average, a user engaged in approximately **3 sports**. Interestingly, when disaggregating the data by gender, it was observed that **female users** demonstrated greater diversity in their participation, with an average involvement in **4 sports**, surpassing their male counterparts.

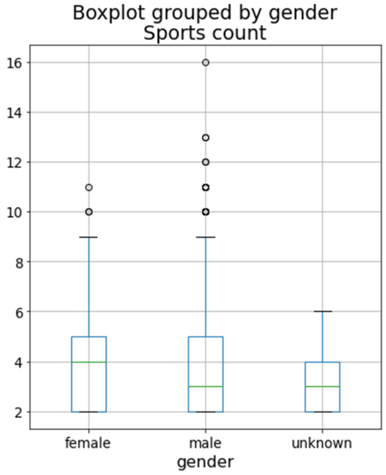


Figure 4. Distribution of sport on the basis of Gender

This finding underscores a higher propensity for multi-sport participation among female users, which could be indicative of broader fitness interests or varied workout routines. Such insights are crucial for tailoring recommendations or designing models that account for gender-specific workout preferences and behaviors.

4.3. Sports and Heart Rate

Heart rate analysis across different sports activities revealed significant variations depending on the type of activity and the gender of participants. For instance, in the case of **soccer**, male participants exhibited an average heart rate of approximately **150 beats per minute (bpm)**, indicating the high-intensity nature of the sport. Conversely, for **fitness walking**, a popular activity among female participants, the average heart rate was observed to be **130 bpm**, reflecting its moderate-intensity profile.

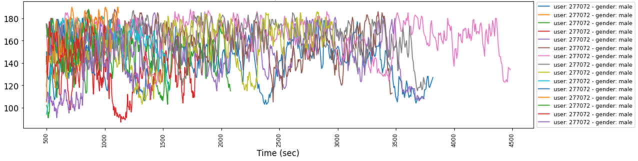


Figure 5. Heart rate of Male during Soccer

These variations emphasize the role of activity type and gender in influencing heart rate patterns. Such insights are valuable for personalized fitness recommendations, as they help in tailoring workout plans to align with the user's physiological responses and fitness goals.

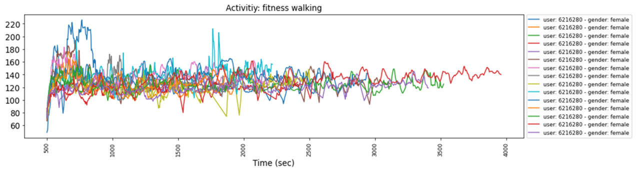


Figure 6. Heart rate of Female during fitness walking

4.4. Location Analysis

The analysis of location-based data revealed distinct patterns in activities depending on whether they are track-based or not. Activities such as **skiing** and **swimming** are inherently **track-based**, often confined to specific routes or facilities like ski trails or swimming pools. In contrast, activities such as **running**, **cycling**, and **mountain biking** are typically not track-based and can occur across diverse terrains and locations.

This distinction is crucial for understanding the spatial constraints of various sports and can inform models for route prediction or activity classification. Additionally, it highlights the need for different preprocessing and feature engineering strategies for track-based versus non-track-based activities, as the latter involves greater variability in environmental factors.

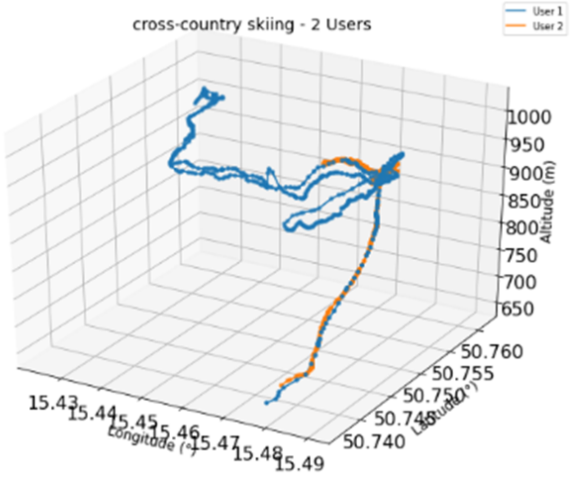


Figure 7. Skiing circuit

5. Model Development

The goal of the model was to predict the **speed** of a user based on features such as **gender** and **heart rate**. For this purpose, a **Random Forest Regressor** was utilized, leveraging its ability to handle non-linear relationships and interactions between variables effectively.

5.1. Model Configuration

The Random Forest Regressor was configured with the following hyperparameters:

- Number of Trees (`numTrees`): **50**
- Maximum Depth (`maxDepth`): **5**

These parameters were chosen to balance model complexity and computational efficiency.

5.2. Performance Evaluation

The model's performance was evaluated using the **Root Mean Squared Error (RMSE)**, which measures the average deviation between the predicted and actual speeds. The model achieved an **RMSE of 10.302636**, indicating reasonable accuracy given the limited feature set.

5.3. Observations

The model demonstrates the potential of using physiological data such as heart rate, along with demographic features like gender, to predict user speed during workouts. However, further improvement can be achieved by incorporating additional features such as activity type, location, or terrain to enhance prediction accuracy.

6. Future Work

The analysis and modeling presented in this study provide valuable insights into workout data, but there are several avenues for future exploration to enhance the understanding and utility of such data. One potential direction is to incorporate additional features, such as environmental factors (e.g., temperature, altitude) and user-specific attributes (e.g., age, fitness level), to improve the accuracy of predictive models. Another area of interest is the application of advanced machine learning algorithms, such as deep learning models, to analyze complex patterns in heart rate, speed, and other physiological data.

Furthermore, extending the analysis to include temporal aspects of the data, such as time-series trends or recovery patterns, could provide deeper insights into user performance and health outcomes. The integration of geospatial data for activities like running and cycling could also enable route-based performance analysis. Lastly, real-time prediction and personalized feedback mechanisms could be developed to provide actionable insights to users, fostering better workout experiences and health management.

7. Conclusion

This paper explored the FitRec dataset to understand user workout patterns, physiological responses, and their relationship to activities. Through exploratory data analysis, key trends were identified, such as the popularity of sports like running and biking and variations in heart rate across activities and genders. A Random Forest Regressor model was developed to predict user speed based on heart rate and gender, achieving a Root Mean Squared Error (RMSE) of 10.302636.

The findings underscore the potential of big data technologies in analyzing real-world workout data to derive meaningful insights. By leveraging tools like PySpark and machine learning algorithms, this study demonstrates how data-driven approaches can enhance our understanding of fitness behaviors and contribute to the development of more personalized and effective workout recommendations. Future work will build upon these foundations to address the limitations and explore new dimensions of this rich dataset.

References

[1] *A Wavelet-Based Approach for Estimating Time-Varying Connectivity in Resting-State Functional Magnetic Resonance Imaging — pmc.ncbi.nlm.nih.gov*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9271336/. [Accessed 13-12-2024].

[2] *Modeling personalized heart rate response to exercise and environmental factors with wearables data - npj Digital Medicine — nature.com*. https://www.nature.com/articles/s41746-023-00926-4. [Accessed 13-12-2024].

[3] *PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals - PubMed — pubmed.ncbi.nlm.nih.gov*. https://pubmed.ncbi.nlm.nih.gov/10851218/. [Accessed 13-12-2024].

[4] *The effects of training on heart rate; a longitudinal study - PubMed — pubmed.ncbi.nlm.nih.gov*. https://pubmed.ncbi.nlm.nih.gov/13470504/. [Accessed 13-12-2024].