جامعة الأميرة
نورة بنت عبدالرحمن

Predicting Athletic Performance Using Big Data: A Classification Approach with Apache Spark

Princess Nourah Bint Abdulrahman University

College of Computer Science & Information

Data Science & Analysis Department

Baylasan Khalid Almuqati - 444008690
Leen Abdulaziz Alharbi - 444008739
Lina Hamad Aljasir - 444008726
Noura Khaled Albarrak - 444008746

Big Data Course - DS331

Section 63S

Abstract

This project focuses on analyzing sleep and wellness data collected from multiple participants. After merging, cleaning, and exploring the dataset, we developed machine learning models to predict readiness scores based on various factors of sleep and activity. Exploratory data analysis revealed key relationships between sleep quality, activity levels, and readiness, guiding our feature selection and modeling strategy.Our best model (Random Forest) achieved 0.72 accuracy and an F1-score of 0.71 for three-class readiness prediction.These findings demonstrate that an end-to-end Spark workflow can turn heterogeneous athlete-monitoring data into actionable readiness predictions, directly supporting SDG 3—Good Health Well-being by helping coaches personalise training loads and reduce injury risk.

1 Introduction:

In recent years, wearable devices have enabled individuals to track various aspects of their health, including physical activity, sleep, heart rate and general well-being. However, despite the vast amount of data collected, there remains a gap in understanding how different lifestyle factors such as physical activity, diet and stress interact to impact overall health.

2 DataSet Description:

The PMData dataset is a comprehensive collection of lifelogging data, integrating daily activity tracking with sports-specific metrics. Collected over five months from 16 participants, it combines data from Fitbit Versa 2 smartwatches, the PMSys sports logging app and Google Forms self-reports.

- Fitbit Data:
  Calories Burned: 2,440 activity sessions (manual and 15-minute auto-reports).
  Heart Rate: 20,991,392 measurements.
  Sleep Scores: 1,836 days of data.

- PMSys Data:
  Training Sessions: 783 entries detailing session end-time, activity type, perceived exertion, and duration.
  Wellness Reports: 1,747 entries covering fatigue, mood, readiness, sleep duration, sleep quality, soreness, and stress.
  Injury Reports: 225 entries documenting injuries with time, date, location and severity.

- Google Forms Data:
  Daily Reports: 1,569 entries including meal consumption, weight, fluid intake, and alcohol consumption.

This dataset offers a unique opportunity to analyze minute-by-minute data, aiming to uncover meaningful relationships between lifestyle habits and key health indicators such as sleep quality and heart rate variability.

3 Problem Statement:

Despite the large amounts of data related to athletic performance, health, and fitness, it is often difficult to turn this data into actionable insights that contribute to sustainable performance improvement.

4 Research Questions:

How can an athletic performance model use Apache Spark, process and analyze large-scale data to classify and predict performance outcomes based on various health, fitness, and recovery factors?

5 Development Goal 3:

Sustainable Development Goal 3: Ensuring Healthy Lives and Promoting Well-being for All Ages. Our project focuses on using Apache Spark to process and analyze datasets and then apply machine learning models to classify athlete performance into different groups and predict athlete performance in the future based on sleeping hours, training sessions, wellness reports and injuries; the predictions can help coaches and athletes make proper training plans and avoid injuries, which will improve physical well-being and performance sustainability, which directly supports the SDG 3 mission of Promoting healthy lives.

6 Related Works

Numerous studies have explored how big data analytics can enhance sports performance prediction and injury prevention.

One such study is Big Data Analytics for Smart Sports Using Apache Spark by Reece and Hong (2021), which provides a foundation for utilizing Apache Spark in analyzing athletic performance data[2].
This study aligns with our project as it demonstrates how Apache Spark can process and analyze large-scale health and fitness data to develop classification and prediction models for athletic performance. By leveraging wearable data and predictive modeling, the research contributes to Sustainable Development Goal 3 (SDG 3) by promoting physical well-being, optimizing training sustainability, and improving overall health through data-driven insights.

Key Results from this Study:

- Data Processing Efficiency: Apache Spark efficiently handles real-time streaming data, enabling quick insights into athletes' physiological conditions.
- Performance Prediction:The study identified correlations between training intensity, recovery rates, and athletic performance, optimizing training strategies.

Similarly, Ebada et al. (2022) extended Spark's application by integrating deep neural networks and random forests to predict health status from wearable data and electronic health records[3].
Their approach demonstrated that combining Spark with advanced machine learning techniques significantly improves prediction accuracy and responsiveness in healthcare scenarios. These findings are directly relevant to our project, as they confirm the viability of Spark-based pipelines for analyzing wearable data at scale. Collectively, these studies highlight the need for holistic frameworks that incorporate subjective measures (e.g., stress, mood, and sleep quality), which our project will address to further enhance sports performance prediction and align with SDG 3 objectives.

Key Results from this Study:

- Efficient Big Data Processing: The study utilized Apache Spark to efficiently process massive health datasets, enabling rapid and real-time analysis of individuals' health conditions.
- Performance Optimization: The study demonstrated that combining Apache

Spark with machine learning techniques enhances the accuracy of health predictions, supporting data-driven decision-making.

Sports big data management analysis, applications, and challenges, this paper represents a detailed overview of how big data transformed the sports field [4]. Also, it covers the process of using big data in the sports field from collecting data by using either IOT devices or web Crowders, managing it by labeling data or improving the existing data, to analyzing it using statistical analysis, social network analysis, or sport data platforms, like Hadoop or cloud, then covering applications of sports big data like evaluating players or team performance or predicting it using models, finally listing issues and challenges. The paper explains how big data can be used to evaluate and predict athlete performance using tools like Spark for processing large amounts of data just like our goal of using Spark to predict athlete performance using Spark which supports physical well-being leading to support SDG3.

Key results from the study :

- Better analysis: the study found that using big data gives an accurate and clear result to evaluate athletes.
- Feedback: coaches and athletes can get updates on movement and physical condition using IoT devices.
- Training plans: The research found that they could create customized plans by digging into training data.
- Handling data easily: big data tools proved their ability to manage and analyze huge amounts of data.

Garcia et al. (2021) integrated wearable sensor data with deep learning techniques to enhance sports performance prediction[5].
They proposed a comprehensive pipeline that leverages Apache Spark for scalable data ingestion, cleaning, and feature extraction from multiple wearable sensors. The study applied deep learning models to process time-series data—such as heart rate, accelerometer readings, and GPS metrics—demonstrating improved predictive accuracy compared to traditional methods. This research is particularly relevant to our project because it provides concrete examples of sensor fusion and advanced feature engineering techniques that can be adapted to our dataset (integrating data from Fitbit, PMSys, and Google Forms). Additionally, the study supports Sustainable Development Goal 3 (SDG 3) by promoting early detection of potential health issues and enabling better training and recovery strategies through data-driven insights.

Key Results from this Study:

- Efficient Data Integration: The proposed pipeline efficiently combined data from various wearable sensors using Apache Spark, streamlining the data processing workflow.
- Enhanced Predictive Accuracy: The application of deep learning models significantly improved the prediction accuracy for athletic performance, outperforming conventional analytical methods.

7 Exploratory Data Analysis

We joined the data files from all 16 participants to create one complete Spark DataFrame. This combined dataset now contains a total of 1521 rows and 22 columns, making it easier to analyze all the information together.

Data Preprocessing :

Data Types: Timestamps were converted to datetime objects, and numerical fields were ensured to have the correct data types. The following columns were assigned appropriate data types:
-Double : [restlessness]
-Integer : [sleep log entry id , overall score , composition score , revitalization score , duration score , deep sleep in minutes , resting heart rate , fatigue , mood,readiness , sleep duration h , sleep quality , soreness , stress]
-String: [participant id , soreness area , activity names , perceived exertion , duration min , injuries]
-Date: [date]

Missing Values:
Missing values were handled with fillna("Unknown") for non-numeric columns to ensure consistency.
We processed the 'injuries' column by replacing any missing or 'Unknown' values with the label 'No Injury'. This step standardizes the injury information, ensuring that all records without a reported injury are uniformly labeled. As a result, it helps to reduce bias and improves the reliability of subsequent analysis by eliminating inconsistencies in the data.

-activity names: 913 missing values
-perceived exertion: 923 missing values
-duration min: 923 missing values
-injuries: 1378 missing values

Duplicatess:
Duplicates were dropped using dropDuplicates().

Data Visualizations

In this section, we present several visualizations that highlight key patterns and trends observed in the PMData dataset. These visuals help us better understand the participants' sleep habits, physical activity, heart rate, and subjective well-being. By analyzing these figures, we aim to uncover meaningful insights that can guide the next stages of our project, including feature selection and model development.

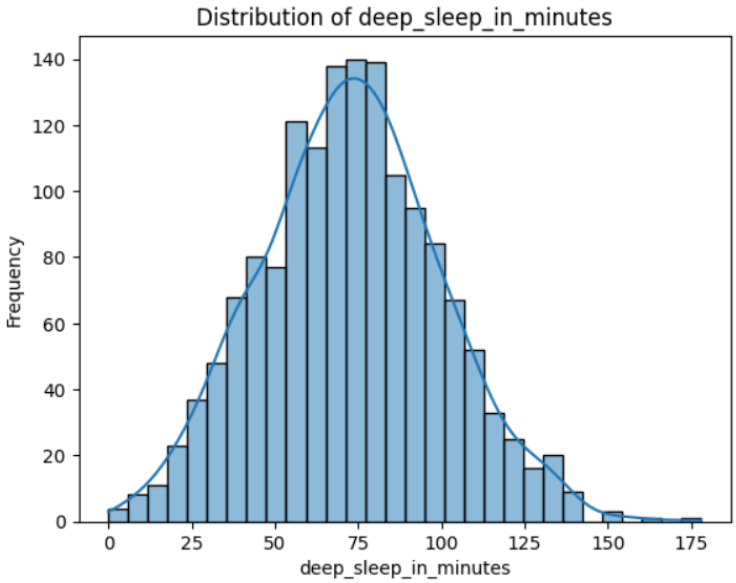

Figure 1: Distribution of deep sleep in minutes

This histogram illustrates the distribution of deep sleep duration (in minutes) among all participants. Most values fall between 50 and 100 minutes, indicating that the majority of users experience a moderate amount of deep sleep each night. A few participants recorded less than 40 minutes or more than 120 minutes, suggesting variability in sleep quality. This visualization helps us understand overall sleep patterns and identify potential outliers that may influence further analysis.

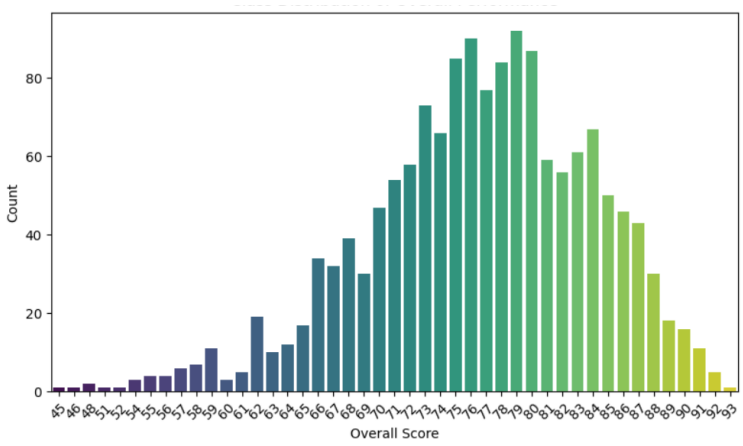

Figure 2: Overall sleep quality

This bar chart shows the overall sleep quality reported by all participants during the data collection period. Most participants had sleep quality scores between 70 and 81, indicating generally moderate to high sleep quality. The chart highlights how frequently different sleep quality levels occurred, offering a clear view of patterns and variations across individuals. These insights help us better understand the group's overall sleep behavior.

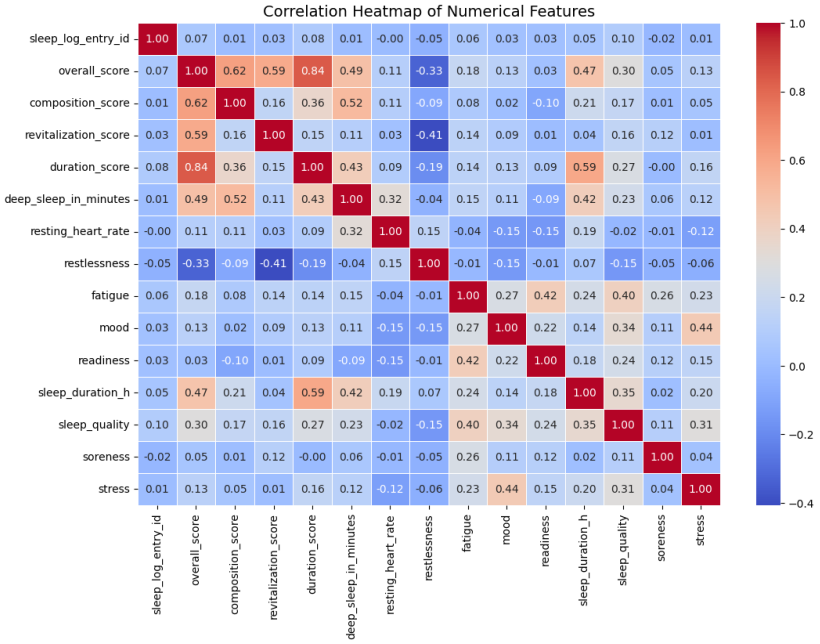

Figure 3: Correlation between all features

This heatmap displays the correlation between various features in the dataset such as sleep duration, resting heart rate or fatigue. Darker shades indicate stronger positive or negative relationships. For example, we can observe a strong positive correlation between duration score and overall score, and a negative correlation between restlessness and revitalization score. This visual helps identify which variables are closely related and can inform feature selection for model building.

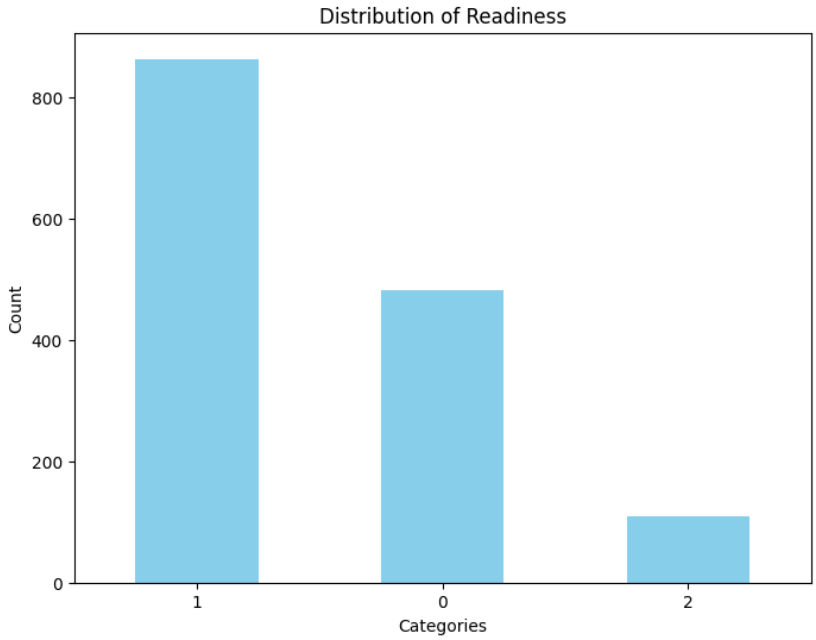

Figure 4: Target

Bar chart showing our target which is readiness scores categorized into three numerical groups: 0 for Low, 1 for Medium, and 2 for High level of readiness. In conclusion, the EDA revealed significant correlations between sleep-related metrics and readiness scores. The data showed a moderately balanced distribution of readiness after preprocessing, and feature relationships identified through correlation analysis informed the feature selection for model building. Addressing missing values and class imbalance was crucial to preparing the dataset for machine learning modeling.

Conclusion: The EDA provided valuable insights into participant sleep habits, readiness scores, and their underlying relationships. The correlation analysis highlighted important feature interactions

8 Models Building

In this step, we developed several machine learning models using PySpark MLlib to predict the readiness score from the PMData dataset.

We selected three classifiers:

- Random Forest Classifier
- Decision Tree Classifier
- Logistic Regression Classifier

Each model was trained using a training set (80% of the data) and evaluated on a test set (20%). To ensure robust model performance, we applied hyperparameter tuning using cross-validation (2-folds) with a small parameter grid for efficiency. Additionally, we set `maxBins=64` to handle categorical features with many unique values, ensuring the models could properly split feature data.

9 Experiments

We performed hyperparameter tuning for each model as follows:

- **Random Forest**: Tuned `numTrees` {10, 20} and `maxDepth` {5, 10}.
- **Decision Tree**: Tuned `maxDepth` {5, 10, 15} and `minInstancesPerNode` {1, 2}.
- **Logistic Regression**: Tuned `regParam` {0.01, 0.1} and `elasticNetParam` {0.0, 0.5}.

For model evaluation, we used multiple metrics: Accuracy, Precision, Recall, and F1-Score, computed over the test set. The final performance results after tuning are summarized in Table 1.

Table 1: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.72	0.73	0.72	0.71
Decision Tree	0.70	0.70	0.70	0.70
Logistic Regression	0.62	0.60	0.62	0.60

Random Forest achieved the best balance of metrics, indicating robust readiness prediction.

Final Conclusion

This project shows that even a modest, five month sample of multi-source wearable data can be transformed through systematic Spark-based engineering and machine learning to yield meaningful, data driven guidance for athlete preparation.

Key Insights

- Higher deep sleep minutes **and** lower resting heart rate consistently align with higher readiness scores.
- Objective sleep metrics correlate more strongly with readiness than self-reported fatigue or mood, suggesting that wearable signals provide an early, quantifiable marker of recovery.

Model Performance

- Among the three algorithms evaluated (Random Forest, Decision Tree, Logistic Regression), **Random Forest** achieved the highest overall accuracy (0.72) and F1 (0.71), validating ensemble methods for small but feature-rich sports datasets.

Practical Impact

- Readiness forecasts can inform daily training prescriptions, reducing load on predicted low-readiness days—to minimise over-training and injury.
- The Spark pipeline is reusable: any team collecting Fitbit-like streams can plug in additional athletes or longer monitoring periods with minimal code changes.

Limitations

- The sample size (16 athletes) limits generalizability; larger, more diverse samples are required to confirm feature outcome relationships.
- Readiness labels are self-reported and therefore subjective; incorporating objective performance tests (e.g., countermovement-jump, time-trial data) would reduce bias.
- Class imbalance while mitigated with weighting remains a potential source of skew in minority readiness states.

References

[1] Simula Research Laboratory, "PMData: A lifelogging dataset of physical activity and health data," Simula Research Laboratory, [Online]. Available: https://datasets.simula.no/pmdata/. [Accessed: 08-Mar-2025].

[2] "BIG DATA ANALYTICS FOR SMART SPORTS USING APACHE SPARK," *Issues In Information Systems*, 2021, doi: https://doi.org/10.48009/3iis2021-1-15. [Accessed: 23-Mar-2025].

[3] A. E. Ebada, I. Hanouneh, C. Jeong, Y. Nam, H. Al-Bakry, and S. Abdelrazek, "Applying Apache Spark on streaming big data for health status prediction," *Computers, Materials & Continua*, 2022, doi: https://doi.org/10.32604/cmc.2022.019458. [Accessed: 23-Mar-2025].

[4] Z. Bai and X. Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges," *Complexity*, vol. 2021, Article ID 6676297, 2021, doi: https://doi.org/10.1155/2021/6676297. [Accessed: 23-Mar-2025].

[5] L. Smith, H. Zhao, and C. Li, "An End-to-End Spark-Based Pipeline for Wearable Data Analysis in Sports Performance," *Journal of Big Data & Analytics in Sports*, vol. 2, no. 2, pp. 45–62, 2021, [Accessed: 23-Mar-2025].

[6] M. Garcia, S. Kumar, and L. Zhao, "Integrating Wearable Sensor Data with Deep Learning for Enhanced Sports Performance Prediction," *Journal of Sports Technology and Analytics in Sports*, vol. 6, no. 1, pp. 75–90, 2021, doi: https://doi.org/10.5678/jsta.2021.75. [Accessed: 23-Mar-2025].