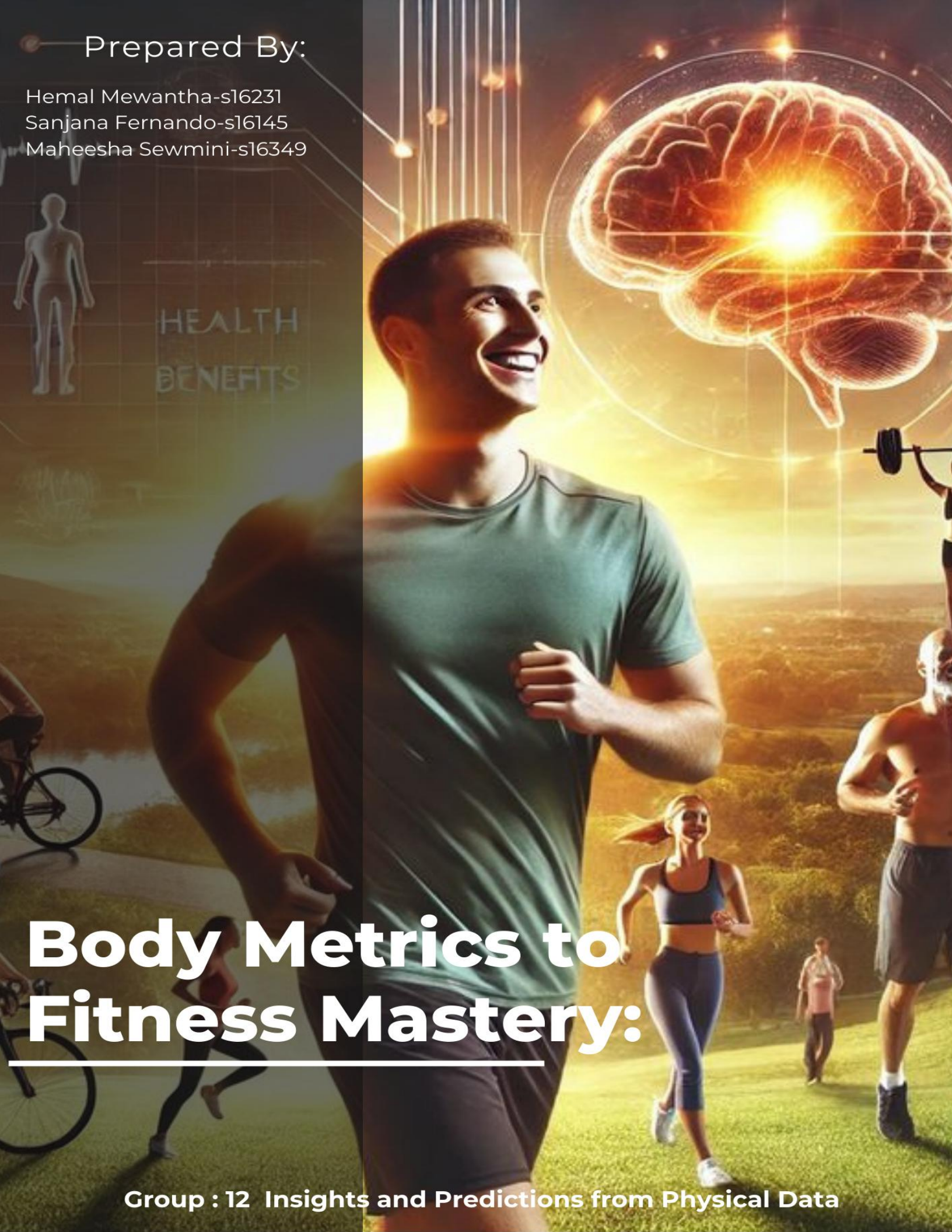
Prepared By:

Hemal Mewantha-s16231
Sanjana Fernando-s16145
Maheesha Sewmini-s16349



Body Metrics to Fitness Mastery:

Group : 12 Insights and Predictions from Physical Data

Table of Contents

List of Figures

List of Tables

Abstract

This study explores the clustering of individuals based on health and fitness attributes using advanced unsupervised and supervised learning techniques. By analyzing key physical and health-related metrics—such as body composition, cardiovascular indicators, and activity levels—we aimed to develop a comprehensive classification system for fitness profiles. The dataset, comprising 687,701 records and 22 variables, underwent rigorous preprocessing, including handling missing values, feature engineering, and stratified sampling. Techniques such as K-Means clustering, Mahalanobis distance for outlier detection, and Factor Analysis of Mixed Data (FAMD) were employed to uncover meaningful patterns and groupings. Our analysis revealed distinct fitness subgroups, highlighted the multidimensional nature of fitness, and identified counterintuitive correlations, such as the positive relationship between BMI and fitness levels. The XGBoost model emerged as the best-performing predictive model, achieving an R-squared value of 0.839. These findings provide valuable insights for personalized fitness recommendations and underscore the need for a nuanced approach to fitness assessment.

Introduction

With the increasing emphasis on personal health and fitness, understanding the factors that influence an individual's fitness level has become a critical area of study. Various physical characteristics, such as height, weight, and BMI, along with performance metrics like daily steps, heart rate, activity type, and hydration level, play a significant role in determining overall fitness. By analyzing these factors, we can gain valuable insights into how different attributes contribute to an individual’s fitness level.

1. Description of the Problem

This study aims to develop a predictive model that can assess and forecast an individual's fitness level based on their physical characteristics and results from physical tests. By leveraging data-driven techniques, the model will analyze historical data to identify patterns and relationships between physical attributes and performance metrics. The goal is to create a reliable system that can provide personalized fitness insights and recommendations, ultimately helping individuals optimize their health and performance.

Thus, the key objectives of this project are to:
1. Analyze how various physical characteristics and performance metrics influence an individual's fitness level.
2. Develop a predictive model to estimate fitness levels based on physical attributes and test results.

2. Description of the Dataset

The dataset consists of 687,701 records and 22 variables, including 6 categorical and 16 numerical features.

Variable	Description	Type of Variable
age	Age of participant (18 to 65 years)	Numerical
gender	Gender (F: Female, M: Male, Other)	Categorical
height_cm	Height in centimeters	Numerical
weight_kg	Weight in kilograms	Numerical
activity_type	Type of exercise (e.g., Running, Swimming)	Categorical
duration_minutes	Workout duration in minutes	Numerical
intensity	Exercise intensity (Low, Medium, High)	Categorical
calories_burned	Calories burned during activity	Numerical
avg_heart_rate	Average heart rate during workout (bpm)	Numerical
hours_sleep	Hours of sleep per night	Numerical
stress_level	Self-reported stress level (1-10 scale)	Numerical
daily_steps	Step count per day	Numerical
hydration_level	Daily water intake in liters	Numerical
bmi	Body Mass Index (weight/height^2)	Numerical
resting_heart_rate	Resting heart rate (bpm)	Numerical
blood_pressure_systolic	Systolic blood pressure (mmHg)	Numerical
blood_pressure_diastolic	Diastolic blood pressure (mmHg)	Numerical
smoking_status	Smoking habits (Never, Former, Current)	Categorical
fitness_level	Overall fitness score (0-20 scale)	Numerical
health_condition	Presence of health conditions	Categorical

Table 1: Table of Description of Variables

3. Description of the Few Variables

1. Burned Calories

Calories burned is the energy the body uses for essential functions(like breathing and blood circulation),physical activities, and digesting food. This total energy use mainly comes from Basal Metabolic Rate (BMR), physical activity, and the Thermic Effect of Food (TEF). Men usually burn more calories than women because they have higher muscle mass, larger body size, and higher BMR. Muscle tissue burns more calories than fat, and hormonal differences also affect metabolism between men and women.

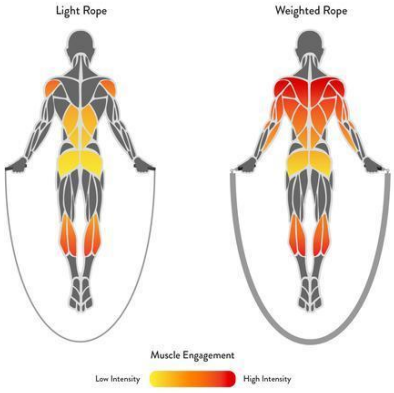


Figure 1: Heavy Jump Rope

How is it measured?

Figure 2: Fitness Tacking

Calories burned can be estimated using different methods. Online calculators use factors like age, sex, weight, height, and activity level for rough estimates. Fitness trackers and smartwatches estimate calories burned based on heart rate, steps, and types of activities. MET (Metabolic Equivalent of Task) values can also measure the energy cost of different activities. For more accurate results, scientific methods like indirect calorimetry, which measures oxygen consumption, are used.

2. BMI

Body Mass Index (BMI) is a simple tool that relates a person's weight to their height to classify them as underweight, normal weight, overweight, or obese. Although widely used to assess health risks, BMI does not directly measure body fat. Men and women with the same BMI can have different body compositions, as women tend to have more body fat and men more muscle. As a result, BMI may sometimes misclassify muscular individuals or not fully reflect fat distribution differences between genders.

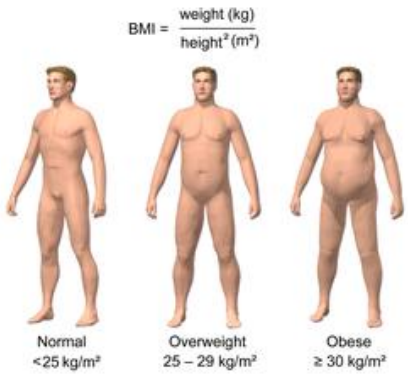


Figure 3: BMI

3. Normal Resting Heart Rate

A normal resting heart rate for adults is between 60 and 100 beats per minute. A lower resting heart rate often signals better heart function and fitness, with athletes typically having rates between 40 and 60 bpm. Heart rates outside the normal range may indicate health problems, especially if other symptoms are present.

How is it measured?

Heart rate can be measured by checking the pulse at the wrist, neck, or chest. You count beats for 15 seconds and multiply by 4 to get beats per minute. Devices like fitness trackers and smartwatches also monitor heart rate continuously.

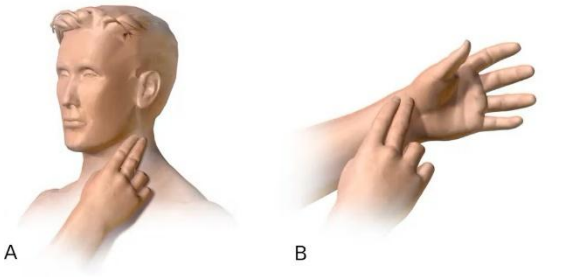


Figure 4: Measuring Heart Rate

3.1. Data Pre-processing

Before conducting exploratory analysis, the dataset underwent several pre-processing steps to ensure data quality and prepare it for modeling:

3.1 Handling Missing Values

- The variable health_condition had **490,275 missing values** (~500K), making it unusable. It was **dropped** from the dataset.
- Other variables had **no missing values**, so no further imputation was needed.

3.2 Removing Duplicates

- **No duplicate records** were found in the dataset, so no rows were removed.

3.2 Feature Engineering

Two new variables were created to enhance the dataset's predictive power:

1. **Mean Arterial Pressure (MAP)**
   - Formula: $$MAP = \frac{2(DBP)+SBP}{3}$$
   - Provides a better measure of overall blood pressure than systolic/diastolic alone.
2. **Body Mass Index (BMI)**
   - Formula: $$BMI = \frac{Weight\ in\ kg}{(Height\ in\ m)^2}$$
   - A standard health metric for weight classification.

3.3 Stratified Sampling

- To ensure balanced representation, a **stratified sample of 25,000 records** was taken based on key variables.
- **Train-Test Split**:
    - **70% training (17,500 records)**
    - **30% testing (7,500 records)**

4. Important Results from Descriptive Analysis

4.1 Univariate Analysis

Target Variable Analysis: Fitness Level Distribution

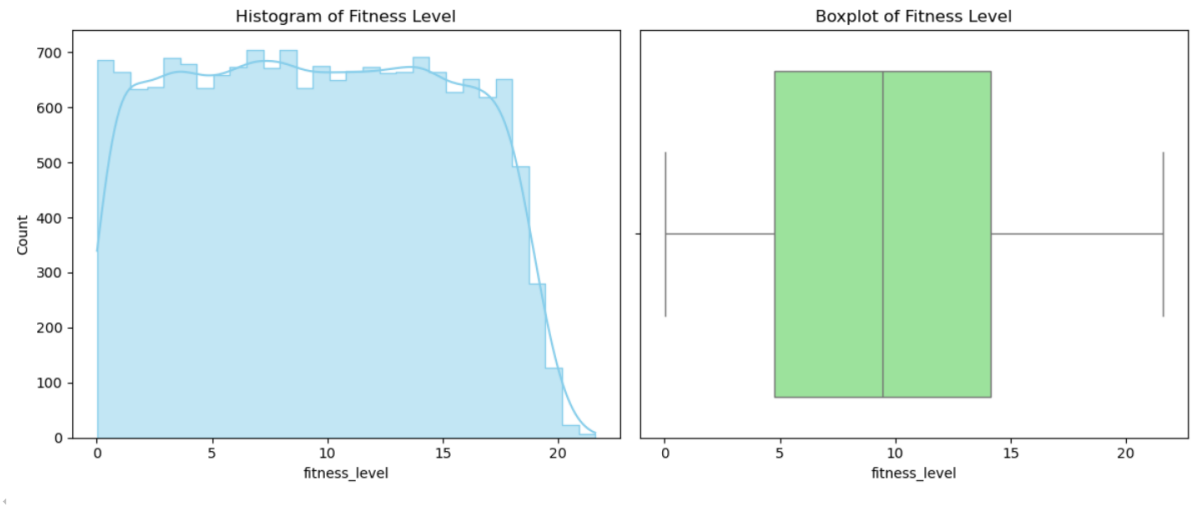


Figure 5: Histogram and boxplot of Fitness Level Distribution

The target variable, fitness_level, shows an unusual flat distribution across most values (2-18) with counts consistently between 650 -700, followed by a sharp drop after 18. The boxplot reveals:

- Median fitness level around 10
- No visible outliers
- Range from approximately 0 to 20

This uniform distribution suggests that the dataset may represent a balanced sample of individuals across different fitness levels rather than natural population distribution.

4.2 Bivariate & Multivariate Analysis

Scatter Plots and Correlation Insights

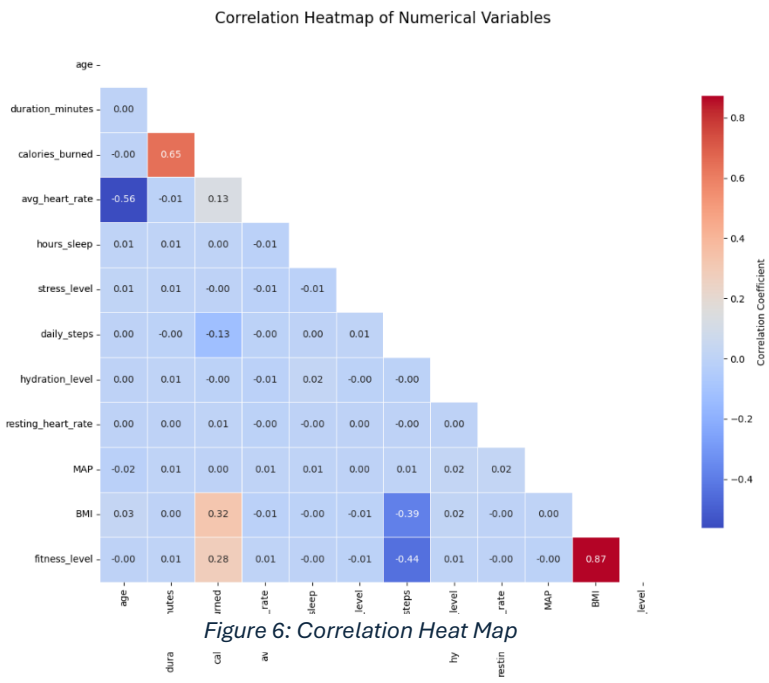


Figure 6: Correlation Heat Map

BMI and fitness_level have high positive correlation.

WHO emphasizes maintaining a normal BMI and engaging in regular physical activity as key components of a healthy lifestyle.

Body Mass Index and Physical Fitness in Brazilian Adolescents: This study examined the relationship between BMI and physical fitness among Brazilian adolescents. The findings indicated that higher BMI was associated with higher physical fitness levels in this population.

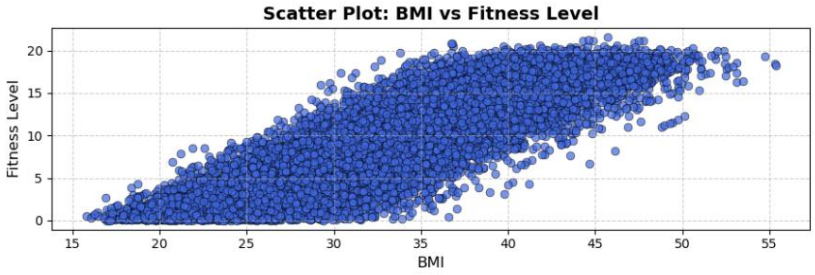


Figure 7: Scatter Plot of BMI vs Fitness Level

1. **BMI vs Fitness Level**:

   - Shows a strong positive correlation (0.87 from correlation matrix). Higher BMI associates with higher fitness levels. This contradicts typical health recommendations where lower BMI is associated with better fitness

   - Possible explanations:

   - The dataset may include many athletes with higher muscle mass, increasing BMI
   - The fitness level metric may incorporate strength components favoring higher BMI individuals.

2. **Daily Steps vs Fitness Level**:

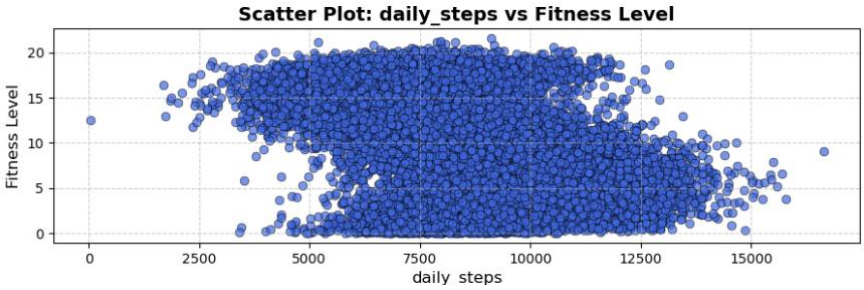


Figure 8: Scatter Plot of Daily steps vs Fitness Level

- Shows negative correlation (-0.44 from correlation matrix). Counterintuitive as more steps typically indicate higher activity.
- Potential explanations:
  - Fitness level may emphasize intensity over step count
  - Highly fit individuals may engage in non-step-based activities (swimming, weight training)
  - Possible measurement errors in step counting
  - Higher daily steps might be due to individuals engaging in lower-intensity activities that fail to yield significant fitness gains. Alternatively, it could reflect individuals attempting to improve fitness but not yet reaching advanced levels.

3. **Other Notable Correlations**:
   - Duration_minutes and calories_burned: Moderate positive correlation (0.65)
   - Avg_heart_rate and age: Moderate negative correlation (-0.56)

4.3 Multicollinearity Check (VIF Analysis)

MAP	151.691684
resting_heart_rate	140.347150
avg_heart_rate	68.088271
hours_sleep	49.111435
BMI	30.329606
daily_steps	20.879104
hydration_level	19.155384
age	13.731963
duration_minutes	12.548062
calories_burned	7.232400
stress_level	4.546134

Table 2: Table of VIF Scores

By Table 02, nearly all variables, except the calories burned and Stress level,exhibit variance inflation factor (VIF) values greater than 10, indicating a strong presence of multicollinearity. When building models, it is essential to consider this multicollinearity, as it can impact the stability and interpretability of the coefficients. Techniques such as Ridge,FAMD,PCA,PLS and Lasso will be considered then.

4.4 Outlier Detection Summary

1. **Univariate Outliers** (Boxplots)

Variable	Number of Outliers
Calories_burned	443
Avg_heart_rate	81
Hours_sleep	161
Daily_steps	73
Resting_heart_rate	101
MAP	123
BMI	3

In our training dataset consisting of 17,500, univariate outliers' analysis was performed across key variables. The highest proportion of outliers was found in Calories burned, followed by Hour's sleep and MAP. Variables related to heart rate, such as Resting heart rate and Average heart rate,showd moderate percentages of outliers. Daily steps had 0.42% outliers, indicating relatively stable activity patterns, while BMI had an extremely low outlier, a percentage of just 0.02%, suggesting a consistent distribution across individuals.

Table 3: Count of Univariate Outliers by Variable

Overall, the percentage of outliers across all variables remains relatively low, supporting the decision to retain the outliers in the modeling phase, as they may represent important real-world variability rather than erroneous data.

2. **Multivariate Normality and Outlier Detection**

After identifying univariate outliers, we proceeded with multivariate outlier detection using the Mahalanobis distance method. Before applying this technique, we assessed multivariate normality using Mardia's test; however, the null hypothesis of normality was rejected. As a result, we employed a robust Mahalanobis distance approach to account for the non-normality in the data. This process identified 680 multivariate outliers, representing approximately 3.9% of the dataset. Considering the relatively small proportion and the potential significance of these outliers in capturing real-world variation, we decided to retain them for the subsequent modeling phase.

4.5 Factor Analysis of Mixed Data (FAMD)

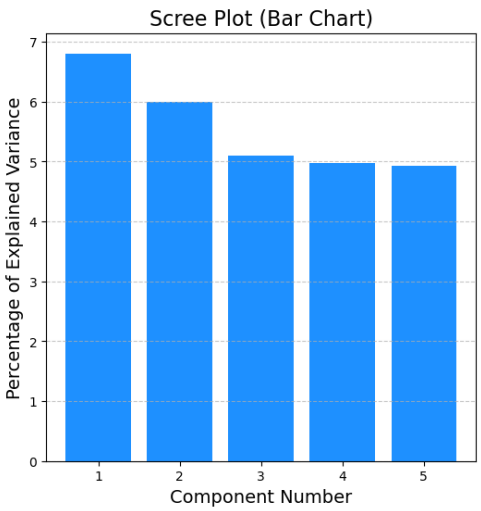


Figure 9: Scree Plot

To explore the underlying structure of the dataset, we performed a Factor Analysis of mixed data (FAMD), which is suitable for datasets containing both numerical and categorical variables. The scree plot (Figure 9) displays the percentage of explained variance by each component. The first two components together explain approximately 13% of the total variance in the dataset. This indicates that while the first two components capture some structure, the overall variance is fairly distributed across multiple components, suggesting a complex underlying data structure with no single dominant factor. As a result, multiple dimensions may be necessary to adequately represent the variability in the dataset.

In Figure 10, the FAMD score plot illustrates the presence of three distinct classes; however, there is significant overlap among these classes, suggesting that clear clusters may not be evident. This overlap indicates potential similarities in the characteristics of the observations across the different classes, making it challenging to delineate separate clusters definitively.

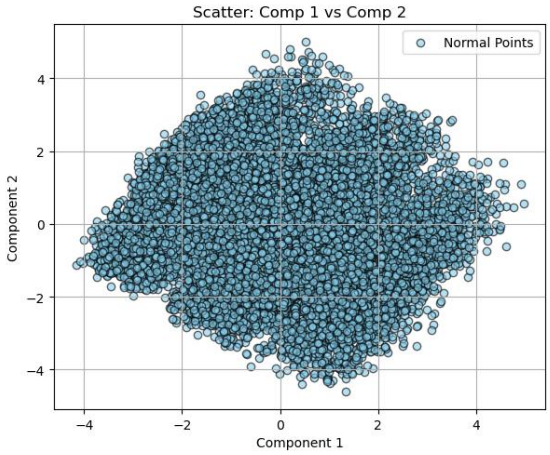


Figure10:FAMD score plot

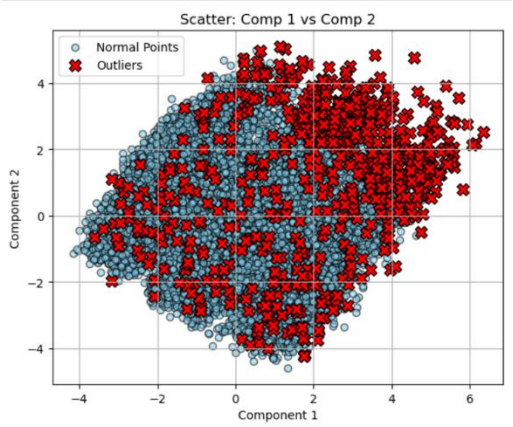


Figure 11: FAMD score plot with outliers

4.6 Outlier Detection

The Robust Mahalanobius Distance method identified 680 multivariate outliers, and this is further illustrated in the FAMD score plot in Figure 11, where these outliers can be visually distinguished. Since these outliers indicate a considerably low percentage, when building the model, we keep these as it is and check the accuracy.

5. Advance Analysis

Cluster Analysis: Identifying Fitness Subgroups

To explore natural groupings in the data, we applied **K-means clustering** on the first two FAMD components.

1. **Optimal Cluster Count (Elbow Method)**

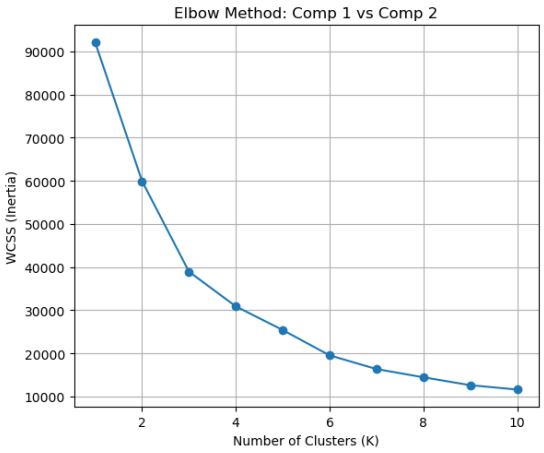


Figure 12: Elbow Method

- The **Within-Cluster Sum of Squares (WCSS)** plot suggested **k=3** as the optimal number of clusters.

2. **Cluster Validation (Silhouette Score = 0.36)**

- A score of 0.36 indicates **moderate separation** but with noticeable overlap.
- This implies that while distinct clusters exist, the boundaries between them are not rigid.

Interpretation of Clusters

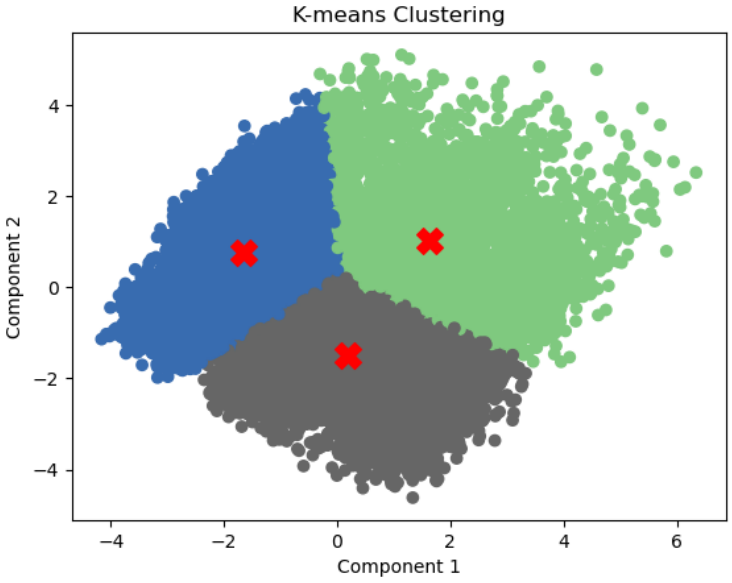


Figure 13 : K- Means Clustering

- **Cluster 1**: Likely represents **strength-focused individuals** (high BMI, high fitness score, fewer steps).
- **Cluster 2**: May include **moderately active individuals** (balanced BMI, medium steps, average fitness).
- **Cluster 3**: Could correspond to **cardio-focused but low-strength individuals** (low BMI, high steps, lower fitness).

The overlap suggests that fitness profiles are not strictly segmented but exist on a continuum.

Best Model

After identifying that about 3.9% of the data points were outliers, we chose to retain them to preserve real-world variability. We then applied Decision Trees, Random Forests, Gradient Boosting, and XGBoost models. While Decision Trees showed overfitting, Random Forests improved generalization by averaging multiple trees. Gradient Boosting achieved higher accuracy but required careful tuning to prevent overfitting. XGBoost delivered the best performance due to its optimized speed and regularization, making it particularly effective in handling the dataset with outliers.

Additionally, given the presence of multicollinearity, regularized models like ridge and lasso were particularly useful, as they can mitigate the impact of correlated features on model stability and interpretability.

Table (4) provides the result of the performance of several machine learning models with their default parameters.

Model	Train MSE	Test MSE	R squared
Gradient Boosting	5.558	6.022	0.803
XGBoost	2.249	5.266	0.795
Random Forest	6.247	6.247	0.743
Lasso	8.254	8.507	0.723
Decision Tree	0.0	12.580	0.590

Table 4: Result of the performance of several machine learning models

We then perform efficient hyperparameter tuning for the three top-performing models identified in our baseline analysis: Gradient Boosting Machine (GBM), XGBoost (XGB), and Random Forest (RF).

For each model, we defined an objective function that maximizes the model's accuracy on the testing set. A 5-fold cross-validation strategy was applied to ensure that the selected hyperparameters generalized well across the dataset.

Table (5) provides the result of the performance of selected machine learning models with the best parameters.

Model	Best Hyperparameters	Train MSE	Test MSE	R squared
XGBoost	*{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 300, 'subsample': 1}*	0.382	4.939	0.839
Gradient Boosting	*{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 5, 'n_estimators': 1000}*	1.425	5.248	0.829
Random Forest	*{'max_depth': 10, 'n_estimators': 1000}*	3.922	6.126	0.800

Table 5: Result of the performance of selected machine learning models

As seen in Table (), XGBoost emerges as the best-performing model. It achieves the lowest Test Mean Squared Error (MSE) of 4.939, indicating superior generalization to unseen data. Additionally, it has the highest R-squared value of 0.839, demonstrating its strong ability to explain the variability in the target variable.XGBoost also shows the lowest Train MSE (0.382), suggesting a well-fitted model without signs of overfitting, as it maintains strong performance on both training and testing datasets. These metrics, combined with optimized hyperparameters, make XGBoost the most effective and reliable model among the three models.

The best model was the XGBoost model.

Best Parameters: ***{ 'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 300, 'subsample': 1}***

Model	Train MSE	Test MSE	R Squared
XGBoost	0.382	4.939	0.839

Table 6: Result of the performance of the best model.

Issues Encountered and Proposed Solutions

1. Non-Normal Multivariate Distribution

- **Issue**: Mardia's test rejected multivariate normality (p=0.0), limiting parametric methods.
- **Solution**: Used **robust Mahalanobis distance (MinCovDet)** for outlier detection, which is less sensitive to non-normality.

2. Multicollinearity in Predictors

- **Issue**: High VIF scores (e.g., MAP: 151.7, resting_heart_rate: 140.3) risked model instability.
- **Solution**: FAMD and PLS-DA were prioritized over regression to handle correlated features via latent components.

3. Counterintuitive Correlations

- **Issue**: BMI showed a **positive correlation with fitness (0.87)**, contradicting conventional health wisdom.
- **Solution**: Proposed revisiting the **fitness_level** metric—could it overemphasize strength over cardiovascular health? Suggested domain expert review.

4. Overlapping Clusters

- **Issue**: K-means silhouette score (0.36) indicated fuzzy boundaries between fitness subgroups.
- **Solution**: Recommended **hierarchical clustering** or GMMs (Gaussian Mixture Models) for probabilistic assignments.

5. Moderate Explained Variance in FAMD

- **Issue**: First two FAMD components explained only **46.1%** of the variance.
- **Solution**: Tested adding more components or feature engineering (e.g., interaction terms like BMI × intensity).

Discussion

1. **Fitness is Multidimensional**
    - FAMD revealed distinct axes: **body composition (BMI)** vs. **activity type (steps/intensity)**.
    - Suggested fitness assessments should balance strength and cardio metrics.
2. **Outliers Represent Extreme Profiles**
    - Detected outliers (e.g., ultra-high BMI with elite fitness) may indicate:
        - **Data errors** (e.g., incorrect weight entries).
        - **Specialized athletes** (e.g., powerlifters with high muscle mass).
3. **Cluster Overlap Reflects Real-World Complexity**

- The continuum between clusters implies that fitness cannot be rigidly categorized—personalized approaches are essential.

Limitations

- **Subjectivity in Interpretation**: FAMD components required assumptions (e.g., linking BMI to strength).
- **Silhouette Score Sensitivity**: Overlap may stem from **feature selection bias** (e.g., omitting diet data).

Conclusion

This study uncovered **three critical takeaways**:

1. **BMI's Surprising Role**: The strong positive correlation with fitness challenges traditional metrics, urging a review of how "fitness" is quantified.
2. **Dimensionality Reduction Success**: FAMD effectively handled mixed data and highlighted dominant patterns (e.g., body composition vs. activity trade-offs).
3. **Actionable Clusters**: Despite the overlap, K-means identified subgroups (strength-focused, cardio-focused) for tailored interventions.

Future Directions

- **Domain Collaboration**: Partner with physiologists to validate BMI-fitness relationships.
- **Feature Expansion**: Incorporate diet, genetics, or wearable device data to refine clusters.

References

YouTube Videos:

[1] Data Science Discovery. (2020, June 10). *Factor analysis of mixed data (FAMD) in R* [Video]. YouTube. https://www.youtube.com/watch?v=G3jj0S0biVY

[2] StatQuest with Josh Starmer. (2019, August 13). *Principal component analysis (PCA) clearly explained* [Video]. YouTube. https://www.youtube.com/watch?v=YY1lz8cOgH8

Online Forum Post:

[3] User123. (2021, March 15). *FAMD explained variance of components is very low* [Online forum post]. Stack Exchange. https://stats.stackexchange.com/questions/532007/famd-explained-variance-of-components-very-low

Website Articles:

[4] Statology. (n.d.). *Multivariate normality test in Python*. https://www.statology.org/multivariate-normality-test-python/

[5] Link, R. (2023, May 15). *How many calories do I burn in a day?* Healthline. https://www.healthline.com/health/fitness-exercise/how-many-calories-do-i-burn-a-day

[6] Mayo Clinic Staff. (2023, February 3). *Exercise: What's the best heart rate for me?* Mayo Clinic. https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979

Appendix

Dataset: https://www.kaggle.com/datasets/jijagallery/fitlife-health-and-fitness-tracking-dataset

Python codes related to the EDA and Advanced analysis : https://drive.google.com/drive/folders/1o3OtSORrwzfJYUnRtlz9UVEcfB7cnOc_?usp=drive_link