**Group 03**

**Final project:**

**16242- Chathuni Amasha**

**16157 - Sandunika**

**16226 – Ashani**

## Abstract

The increasing use of wearable devices and fitness trackers has generated large volumes of data related to individuals’ daily activities, biometric measures, and lifestyle factors. This project aims to utilize these data to develop a predictive model that classifies a person’s health status into categories such as *healthy*, *high blood pressure*, or *diabetes*. Using machine learning algorithms, the model will analyze health-related features such as BMI, activity type, heart rate, blood pressure, stress level, and sleep hours. The ultimate goal is to create an application that provides personalized health insights and recommendations based on user input, helping individuals better understand and manage their health.

## Table of Contents

## Abstract........................................................................................................................................ 1

## List of Figures ..............................................................................................................................1

## List of Tables................................................................................................................................1

## Introduction ..................................................................................................................................2 Description of the Question.......................................................................................................... 2 Description of Dataset................................................................................................................... 2

## Data Preprocessing ........................................................................................................................3

## Results of Descriptive Analysis.....................................................................................................3

## Results of Advanced Analysis....................................................................................................... 9

## Conclusion.....................................................................................................................................10

## Appendix ......................................................................................................................................10

## References.....................................................................................................................................10

## List of Figures

Figure 1 : pie chart of health\_condition

Figure 2 : Boxplot of weight vs health\_condition

Figure 3: Boxplot of height vs health\_condition

Figure 4 : bar chart of gender vs health\_condition

Figure 5 : CORRELATION HEATMAP

Figure 6 : CRAMER’S HEATMAP

**Figure 7 : EXPLAINED VARIANCE BY FAMD COMPONENTS**

Figure 8 : average silhouette score plot

Figure 9 : Scatter plot of multivariate outliers

## List of Tables

**TABLE 1: DESCRIPTION OF DATASET ATTRIBUTES**

**TABLE 2: PERFOMANCE COMPARISON OF SELECTED MODELS**

## Introduction

In the modern era, wearable technology plays a vital role in monitoring physical activity and health-related metrics. Devices such as smartwatches and fitness bands continuously collect data that reflect an individual's lifestyle and physiological conditions. This data, when properly analyzed, can offer valuable insights into potential health risks and overall well-being.  
This project focuses on building a predictive health classification system that leverages health and fitness data to categorize users according to their health conditions. By analyzing patterns in features such as body mass index (BMI), heart rate, activity intensity, and stress level, the system aims to identify critical indicators that contribute to health deterioration.  
The study not only provides predictive insights but also promotes preventive healthcare by allowing users to monitor their risk levels in real-time through a user-friendly dashboard or application.

## Description of the Question

Wearable devices and fitness trackers collect large amounts of data related to daily activities, health indicators, and lifestyle factors. These data can be used to identify health risks and support personalized recommendations. Our project focuses on building a predictive model to classify health status (health\_condition) into categories such as healthy, high blood pressure, and diabetes, based on features like BMI, activity levels, heart rate, sleep hours, stress level, and other lifestyle indicators.

## Description of the Dataset**.**

The health\_fitness\_dataset contains real health and fitness tracking data from 3,000 participants. This dataset captures daily activities, important health indicators, and lifestyle factors, and is of great value for health analysis and predictive modeling.

Ref: https://www.kaggle.com/datasets/evan65549/health-and-fitness-dataset

|  |  |  |
| --- | --- | --- |
| Variable Name | Type |  |
| Participant\_id | Integer | Sequential identification number of each individual |
| Age | Quantitative | Age of the individual |
| Gender | Categorical | Gender of the individual (e.g., Male, Female). |
| date | Quantitative | Date of record |
| Height\_cm | Quantitative | Height in centimeters |
| Weight\_kg | Quantitative | Weight in kilograms |
| BMI | Quantitative | Body Mass Index |
| Activity\_type | Categorical | Exercise type (e.g. running, swimming) |
| Duration\_minutes | Quantitative | Duration of activity |
| intensity | Categorical | Exercise intensity (low/medium/high) |
| Calories\_burned | Quantitative | Estimated calories burned |
| Daily\_steps | Quantitative | Total steps in a day |
| Avg\_heart\_rate | Quantitative | Average heart rate during activity |
| Resting\_heart\_rate | Quantitative | Resulting heart rate |
| Blood\_pressure\_sys | Categorical | Systolic blood pressure |
| Blood\_pressure\_dia | Quantitative | Diastolic blood pressure |
| Endurance\_level | Quantitative | Endurance level (1-10) |
| Sleep\_hours | Quantitative | Hours of sleep per day |
| Stress\_level | Quantitative | Stress level (1-10) |
| Hydration\_level | Quantitative | Water intake (liters/day) |
| Smoke\_status | Categorical | Smoking history (never/before/now) |
| Healthy\_condition | Categorical | Health status (target variable : healthy, high blood pressure ,diabetes ,etc) |
| Fitness\_level | Quantitative | Fitness level |

**TABLE 1: DESCRIPTION OF DATASET ATTRIBUTES**

## Data Preprocessing

* **Checked for duplicate**s there were no duplicates in the dataset.
* Missing values in the target variable **‘health\_condition’** were replaced with the category **“healthy”**
* Object-type columns (text-based variables such as gender, activity\_type, or smoke\_status) were identified and converted to **categorical** data type.
* **Split the dataset** into **80% training and 20% testing sets** for model development and evaluation.
* BMI was calculated using participants’ height and weight measurements, and a new column **bmi** was added to the dataset.
* A check for missing values across all variables .The results indicated that there are **no missing values**

## Results Of Descriptive Analysis

***Proportion of Categories in the Response Variable***

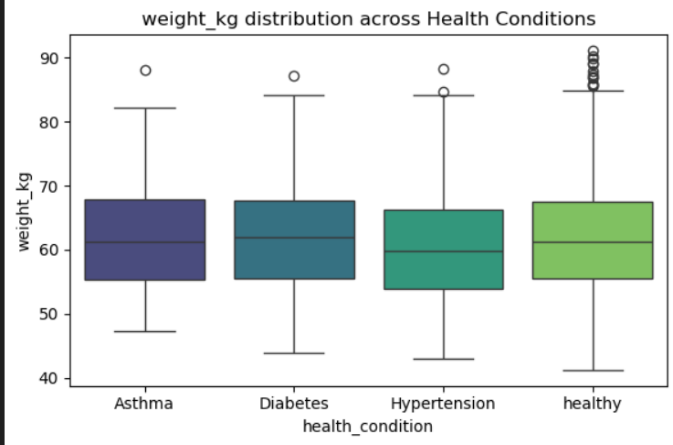
***A graph showing a number of different colored squares

AI-generated content may be incorrect.***

The bar chart illustrates the distribution of individuals according to their health conditions. It shows that the majority of the participants are **healthy**, with a count exceeding 1600, making it the most common category in the dataset. In contrast, the numbers of individuals with **Hypertension** and **Diabetes** are considerably lower, each ranging between approximately 300 and 500. **Asthma** has the least representation, with fewer than 200 individuals.

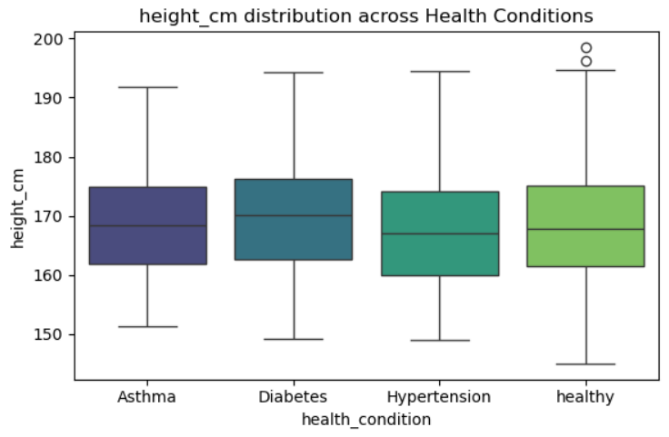
Figure 1 : pie chart of health\_condition

This indicates that the dataset is **highly imbalanced**, with the healthy group dominating the sample, while the other health conditions are relatively underrepresented. Such an imbalance should be considered during any data analysis or model development to ensure accurate and unbiased results.

**Effect on predictors to the response**

The box plot illustrates the **distribution of weight (in kilograms)** across different health conditions — Asthma, Diabetes, Hypertension, and Healthy. Overall, the median weight for all groups lies around **60–65 kg**, indicating a relatively similar central tendency among the categories. The **interquartile ranges (IQRs)** are also comparable, showing that weight variation within each group is fairly consistent. A few **outliers** are present in all categories, especially among the healthy individuals, where several weights exceed 85 kg, indicating individuals with higher body weights than the general range.

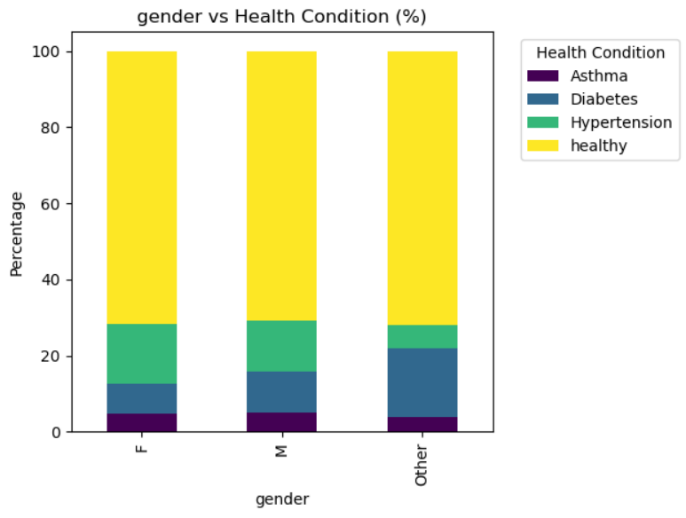
Figure 2 : Boxplot of weight vs health\_condition

Furthermore, an **ANOVA test** was conducted to statistically assess whether there are significant differences in mean weight across the groups. The resulting **p-value of 0.0298** indicates that there is a **statistically significant difference** among at least one of the health condition groups (since p < 0.05). This suggests that while the visual differences appear subtle, the weight distributions across health conditions are **not identical**, implying that weight may have some influence on health status in this dataset.

The box plot illustrates the **distribution of height (in centimeters)** across different health conditions. The median height for all groups appears to be around **170 cm**, showing that the central tendency is quite similar among them. The **interquartile range (IQR)** is also consistent, indicating that most individuals across these health conditions fall within a similar height range. A few **outliers** are observed in the healthy group, representing individuals with heights above 195 cm.

Figure 3: Boxplot of height vs health\_condition

To statistically verify whether height differences among these groups are significant, an **ANOVA test** was performed. The test yielded a **p-value of 0.0449**, which is below the conventional 0.05 threshold. This indicates a **statistically significant difference** in average height among at least one of the health condition groups. Although the visual differences between the box plots appear subtle, the ANOVA result suggests that **height may vary slightly but meaningfully across different health conditions** in this dataset.



The stacked bar chart illustrates the **percentage distribution of health conditions across different genders** , Female (F), Male (M), and Other. In all gender categories, the **Healthy** group (yellow) represents the majority, occupying the largest proportion, while the remaining health conditions (**Hypertension**, **Diabetes**, and **Asthma** ) make up smaller portions of the total. The proportions appear relatively similar across genders, suggesting that most individuals, regardless of gender, are classified as healthy, with only a smaller fraction experiencing specific health conditions.

Figure 4 : bar chart of gender vs health\_condition

However, a **Chi-square test** was conducted to statistically evaluate whether there is an association between gender and health condition. The resulting **p-value of 0.0363** (which is less than 0.05) indicates a **significant relationship** between gender and health condition. This means that the distribution of health conditions is **not independent of gender**, implying that certain health conditions may occur more frequently in specific gender groups within the dataset.

***Further Analysis***

Correlation Heatmap

A diagram of heat map

AI-generated content may be incorrect. A screenshot of a graph

AI-generated content may be incorrect.

Figure 5 : correlation heatmap Figure 6 : cramer’s heatmap

The **first heatmap** shows correlations among numerical features. Strong positive relationships are seen between **height and weight (r = 0.76)**, **weight and BMI (r = 0.61)**, and **age and endurance level (r = 0.72)**, indicating natural links among these variables. Most other correlations are weak, suggesting that factors like **daily steps, calories burned, and duration minutes** are largely independent. No strong negative correlations appear, meaning the numerical features do not inversely affect each other.

The **second heatmap** illustrates Cramér’s V associations among categorical variables. Most features are nearly independent, with very low association values except for slight links between **gender and health condition (0.04)** and **intensity and health condition (0.032)**. This implies that while these may have minor relationships, overall, the categorical variables contribute distinct information without major overla

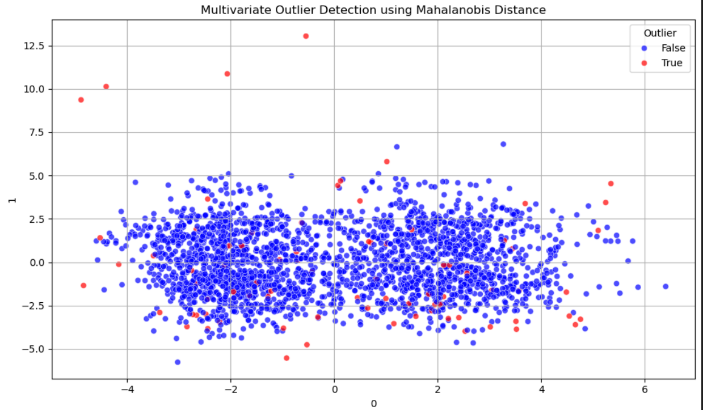
FAMD plot

The **explained variance plot** shows that the **first two FAMD components capture only a small percentage of the total data variance**, indicating that they do not adequately represent the dataset’s overall structure. As a result, performing cluster analysis using only these two components would not yield meaningful groupings. When considering all **nine FAMD components**, the **average silhouette scores** across different numbers of clusters remain very low, suggesting that the data does not form any well-defined or distinct clusters. Therefore, the results indicate that **clustering is not suitable for this dataset**, as the variables do not naturally separate into clear groups based on the available features.

A graph with a line

AI-generated content may be incorrect.A graph with a line going up

AI-generated content may be incorrect.Figure 7 : explained variance by famd components Figure 8 : average silhouette score plot



The analysis identified **82 multivariate outliers** when considering all **10 FAMD components**. These outliers were detected using the **Mahalanobis distance**, which measures how far each observation lies from the center of the multivariate distribution, taking into account correlations between variables. The scatter plot provides a **two-dimensional illustration** of these results using only the **first two FAMD components** for visualization. In the plot, the **red points represent the detected outliers**, while the **blue points indicate normal observations**. Although the visualization is limited to two components, it helps illustrate the presence of observations that deviate significantly from the overall data pattern in the multivariate space.

Figure 9 : Scatter plot of multivariate outliers

## Results of Advanced Analysis

Advanced analysis involves using more sophisticated techniques such as machine learning models to uncover deeper insights and improve predictive performance. The following table shows the results of some selected models used in this analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | Accuracy | | Sensitivity | | F1 score | |
|  | Training | Test | Training | Test | Training | Test |
| Logistic Regression | 0.713 | 0.712 | 0.713 | 0.712 | 0.593 | 0.591 |
| Ridge classifier | 0.715 | 0.705 | 0.715 | 0.705 | 0.602 | 0.589 |
| LDA | 0.714 | 0.695 | 0.714 | 0.695 | 0.616 | 0.586 |
| XG Boost | 0.828 | 0.712 | 0.828 | 0.712 | 0.8 | 0.595 |
| SVM | 0.933 | 0.51 | 0.933 | 0.252 | 0.932 | 0.251 |
| Random Forest | 1 | 0.7 | 1 | 0.7 | 1 | 0.595 |

Table 2:perfomance comparison of Selected models

## Conclusion

**Best Model**

## 12. Appendix: Python Code

## 13. References