TU Dublin Tallaght

Analysing Fitness Patterns and Predicting Calorie Expenditure for Gym Members

A Data Analysis and Predictive Modelling Report

By Alan Walker X00204054 & Roko Skugor X00203402

Computing with IT Management (TA\_KACTM\_B)

Abdelsalam Busalim

08.12.2024

Contents

[Executive Summary 3](#_Toc184482822)

[Objective 3](#_Toc184482823)

[Introduction 4](#_Toc184482824)

[Methodology 5](#_Toc184482825)

# Executive Summary

## Objective

The aim of this report is to investigate, analyse and present conclusions based on data extracted from gym members. This Is

# Introduction

# Methodology

## Data Description

The data that has been used for this report was a dataset found on Kaggle, the link of which can be found in the references. This CSV dataset that has been used contains 973 records, with each having 18 variables. These variables cover key areas that are important to analyse such as demographics (age, gender), physiological metrics (BMI, average heart rate) and workout details.

## Tools Used

We used a Jupyter notebook to record all our data input, cleaning, processing, output and analysing. This allowed us to annotate our code step by step as it was produced. The main libraries used for this project were:

* Pandas: This python library provides a useful way to interpret and clean data. Data is separated into rows and columns in a Dataframe, which can be easily manipulated and cleaned.
* NumPy: Useful for working with math, provides methods such as a square root method which streamline working with numbers.
* Matplotlib: Essential library used for data visualization.
* Seaborn: Like matplotlib, used for data visualization but has some differences & is a more intuitive tool.
* Scikit-learn: Very useful machine learning package which allows for the training of models such as linear regression and provides the tools for prediction.

## Data Cleaning

Once the data had been imported and passed into a Dataframe, the next step was to clean it. The process of data cleaning encompasses every step done to ensure quality, veracity and validity of data. This includes removing duplicate values, handling missing values, determining and dealing with outliers and formatting categorical values. The first step was to clean out missing values. Luckily, the dataset being used was whole and had no empty fields, so after confirming these facts, there was nothing else to be done. The next step was to identify outliers and handle them, which was done by removing the outliers using the IQR (Interquartile Range) Method.

After this, standardization of the categorical (Non-numerical) values was needed. For example, for the different workout types, the .strip() and .capitalize() methods would be used, which would remove extra whitespace & Capitalize the first letter respectively. To ensure consistency, all time units would be standardized to minutes and the column renamed.

Some important contextual data was also added. A categorical column called “BMI” was created based on the numerical BMI values provided by the dataset, to streamline the data processing and reduce the amount of calculations that needed to be done. The “Experience\_Level” column held numbers, which indicated it was Numerical. To make it easier to read this data, categorical values were mapped onto these numerical values.

The final few steps when cleaning the data were some minor touch-ups, such as checking and removing duplicates and verifying the data types of each column. If duplicates were not removed these could skew the data and make the data inaccurate, and if the data types were not appropriate (for example, weight being a String) these would mean the data would be incredibly harder to format. The final step was to ensure that the data was consistent in the sense of not having negative values where they shouldn’t be (like a negative amount of calories burned) and once these were all completed the cleaned Dataframe was saved and prepared to be used.

# Initial Data Exploration

## Descriptive Statistics

Once the data had been cleaned and prepared, some initial logic was applied to the dataset to summarize and describe it, giving an idea of where to start to find trends and dependent factors. Basic data manipulation techniques such as finding the mode, quartiles, mean and standard deviation provides an overview of what to expect from the data.

From the descriptive statistics, some important details can be inferred: from our data set, the average BMI is 24.91 which is bordering overweight. This data is above the average BMI of the general population. Another fact that can be determined from the data is that the amount of calories burned varies considerably from one entry to another. There seems to be a large amount of optimisation that can be done when it comes to burning calories in workouts.

A correlation matrix is an indispensable tool that can be used to visualize correlation between different variables. In this case, the following correlation matrix was produced. The strongest correlations that can be identified are that the amount of calories burned has a strong positive correlation with a workout session, meaning that longer sessions will strongly increase the amount of calories burned. There are strong negative correlations between fat percentage and both calories burned and session duration, but considering the previous fact that session duration and calories burned have such a strong correlation, the meaningful interpretation of these would be that an individual with a higher body fat percentage would have a tendency to work out for a short period of time. Finally, the slightly positive correlation is between calories burned and heart rate. A higher heart rate is indicative of a more intense workout, so it makes sense that this would correlate with more calories burned.

A correlation matrix mapping correlation between variables.
