ARTIFICIAL INTELLIGENCE
&
MACHINE LEARNING

Project Report

Semester-IV (Batch-2022)

Calories Burnt Prediction

Supervised By:

Mr. Talib

Submitted By:

Simran - 2210992382

Simrit - 2210992384

Srishti - 2210992404

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

Abstract

With the increasing prevalence of sedentary lifestyles and associated health issues, there is a growing interest in accurately predicting and monitoring physical activity and its effects on caloric expenditure. In this project, we propose a novel approach to predict the calories burned during physical activities using Artificial Intelligence and Machine Learning techniques.

Our dataset consists of various physical activity features such as heart rate, duration of activity, age, weight, and gender, collected from a diverse group of participants. We employ machine learning algorithms including Linear Regression, Decision Trees, Random Forest and Support Vector Machines to develop predictive models.

The models are trained and validated using a portion of the dataset and tested on the remaining data to assess their accuracy. Our results demonstrate that the proposed models can accurately predict the calories burned during physical activities with a high degree of accuracy.

Accurate prediction of calories burned during physical activities can help individuals make informed decisions about their exercise routines, leading to improved health and fitness outcomes.

Table of Contents

Introduction

In recent years, there has been a growing concern about sedentary lifestyles and their impact on health. Physical inactivity is associated with a range of health problems, including obesity, cardiovascular diseases, and diabetes. To combat these issues, there is a need for accurate methods to monitor and predict the caloric expenditure during physical activities.

The advancement of Artificial Intelligence (AI) and Machine Learning (ML) techniques offers promising solutions to this challenge. By leveraging AI and ML algorithms, it is possible to develop predictive models that estimate the calories burned during various types of physical activities. These models can take into account a range of factors such as heart rate, duration of activity, age, weight, and gender to provide personalized predictions.

In this project, we aim to develop an accurate and reliable model for predicting calories burned during physical activities using AI and ML techniques. We will use a dataset containing information about different physical activities and the corresponding calories burned, along with other relevant features.

By accurately predicting caloric expenditure during physical activities, individuals can make more informed decisions about their exercise routines, leading to better health outcomes and an improved quality of life.

Features

1. Data Collection and Preprocessing:

- Gather a diverse dataset containing information about physical activities, such as heart rate, duration, type of activity, age, weight, and gender.
- Clean and preprocess the dataset to handle missing values, outliers, and inconsistencies.

2. Feature Engineering:

- Select relevant features that have a significant impact on caloric expenditure during physical activities.
- Explore feature interactions and transformations to enhance model performance.

3. Model Selection:

- Experiment with various AI and ML algorithms such as Linera Regression, Decision Trees, Random Forest, Support Vector Machines.
- Evaluate the performance of each model using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

4. Model Training and Validation:

- Split the dataset into training and validation sets.
- Train the selected models on the training set and fine-tune hyperparameters using cross-validation techniques.

5. Model Evaluation:

- Evaluate the performance of the trained models on the validation set.
- Compare the performance of different models and select the one with the best performance.

6. Model Testing:

- Test the selected model on a separate test dataset to assess its generalization performance.
- Evaluate the model's ability to accurately predict caloric expenditure during physical activities.

Problem Statement

"Investigate and analyze the relationship between individual health metrics, including calorie intake, gender, age, height, weight, activity duration, heart rate, and body temperature. Identify patterns, correlations, and factors influencing overall health and wellness. Develop insights to inform personalized health and fitness strategies tailored to individual needs and goals." Explore how factors such as gender, age, and body composition influence dietary habits and physical activity levels, and how these, in turn, impact physiological indicators like heart rate and body temperature. Identify potential outliers or discrepancies within the dataset and discern whether they represent unique health profiles or data anomalies.

The purpose of this capstone project is to address the above problem by doing a comprehensive and effective analysis of Calories burnt prediction. This project aims to provide a detailed understanding of the patterns of health and fitness strategies.

Objectives

1. Develop Accurate Prediction Models:

Utilize AI and ML techniques to develop predictive models that accurately estimate the calories burned during various physical activities.

2. Personalized Prediction:

Develop models that can provide personalized predictions by taking into account individual factors such as heart rate, duration of activity, type of activity, age, weight, and gender.

3. Improve Fitness Tracking:

Create a tool that allows individuals to track their caloric expenditure during physical activities more accurately, helping them to achieve their fitness goals more effectively.

4. Enhance Health Monitoring:

Provide a means for individuals to monitor their physical activity levels and caloric expenditure, which can help in preventing health issues such as obesity, cardiovascular diseases, and diabetes.

5. Optimize Exercise Routines:

Enable individuals to make informed decisions about their exercise routines by providing them with insights into the calories burned during different types of physical activities.

Requirements

Software Requirements:

1. Libraries

NumPy: Fundamental package for scientific computing with Python.

Pandas: Data manipulation and analysis library.

Scikit-learn: Machine learning library for Python.

Matplotlib: Plotting library for Python.

Seaborn: Statistical data visualization based on Matplotlib.

Linear regression, Decision Tree Regression, Random Forest Regressor, Support Vector machine.

2. Google Collab

Hardware Requirement:

For the Calories Burned Prediction Project using AI and ML, the hardware requirements depend on the size of your dataset, complexity of your machine learning models, and the scale of your project. Here are the general hardware requirements:

Processor (CPU): A multi-core processor is recommended, preferably with a clock speed of 2 GHz or higher. For larger datasets and more complex models, a faster CPU with more cores will speed up training times.

Random Access Memory (RAM): At least 8 GB of RAM is recommended for most machine learning tasks. For larger datasets and more complex models, 16 GB or more may be necessary.

Development Devices: Development devices such as desktop computers, laptops, or workstations with adequate processing power and memory to support software development and testing activities.

Data Backup: Ensure you have a reliable data backup system in place to prevent data loss.

Proposed Solution

By following this proposed solution, we aim to develop an accurate and reliable model for predicting caloric expenditure during physical activities using AI and ML techniques, with the ultimate goal of improving personalized fitness tracking and health monitoring.

Linear Regression:

Linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the independent variables and the dependent variable.

The model equation for simple linear regression is: $y = mx + b$, where y is the dependent variable, x is the independent variable, m is the slope, and b is the intercept.

Decision Trees:

Decision Trees are non-parametric supervised learning models used for classification and regression tasks.

They work by recursively splitting the dataset into subsets based on the most significant attribute.

Each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value.

Decision Trees are easy to understand and interpret, making them useful for decision-making.

Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to create a more robust and accurate model. It builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Each tree in the Random Forest is trained on a random subset of the training data and a random subset of the features. Random Forest is widely used for classification and regression tasks and is known for its high accuracy and robustness.

Support Vector Machine (SVM):

Support Vector Machine is a supervised learning model used for classification and regression tasks.

SVM tries to find the hyperplane that best separates the classes in the feature space.

It works by finding the hyperplane that maximizes the margin between the two classes.

SVM can handle both linear and non-linear data using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid.

Results

Chart 1

- A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions.
- This Pie chart aims to show the distribution of the gender.

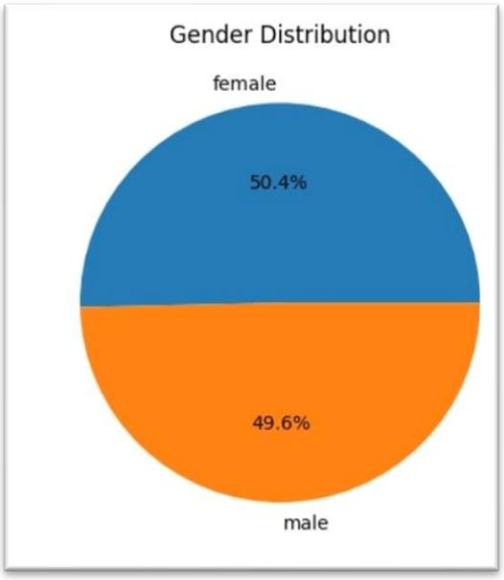


Chart 2

- Bar plots are excellent for comparing the counts or frequencies of different categories.
- In this graph we are comparing the mean values of different variables for males and females separately.

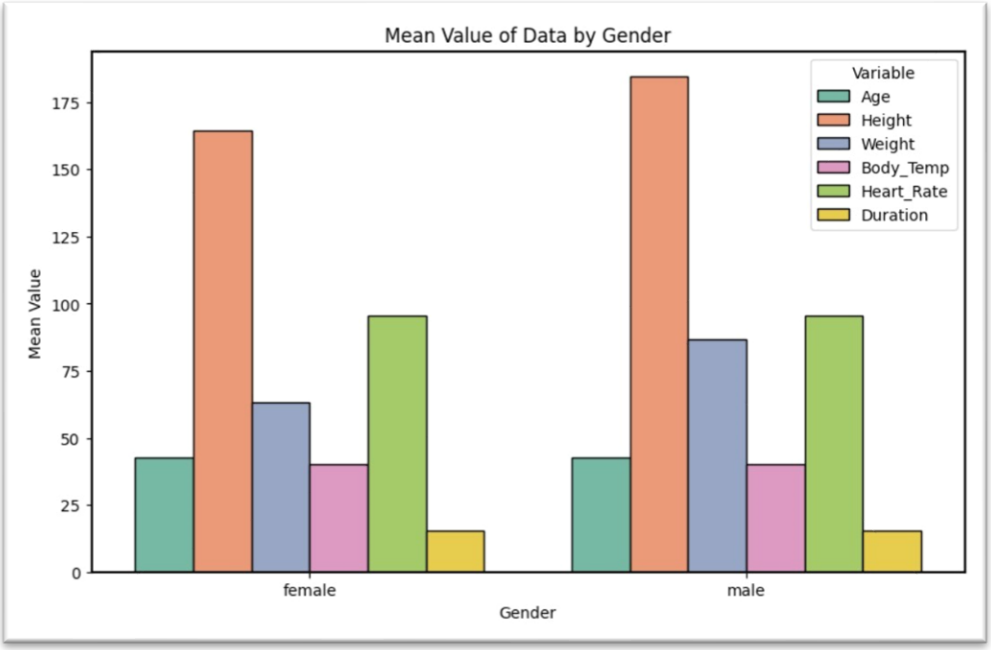


Chart 3

- A bar plot, also known as a bar chart, is a graphical representation of categorical data with rectangular bars.
- This aims to show the distribution of the gender in terms of mean calories.

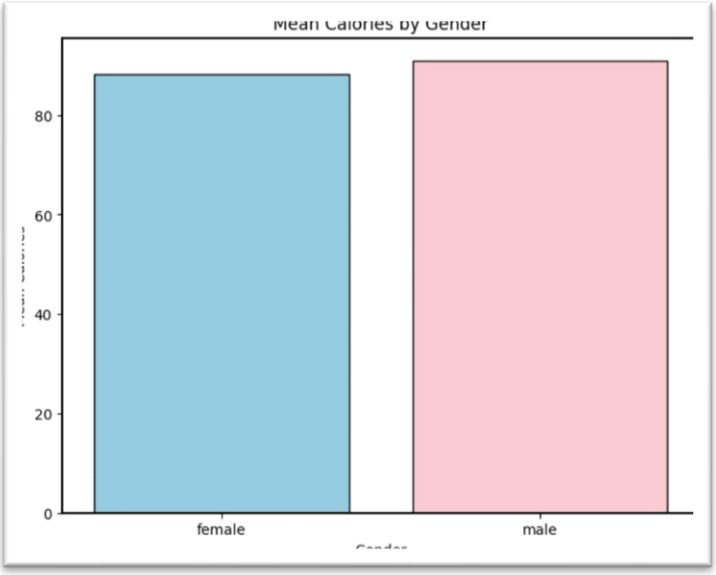


Chart 4

- A violin plot is a method of plotting numeric data and is a combination of a box plot and a kernel density plot.
- The plot is between Body Temperature Vs calories.

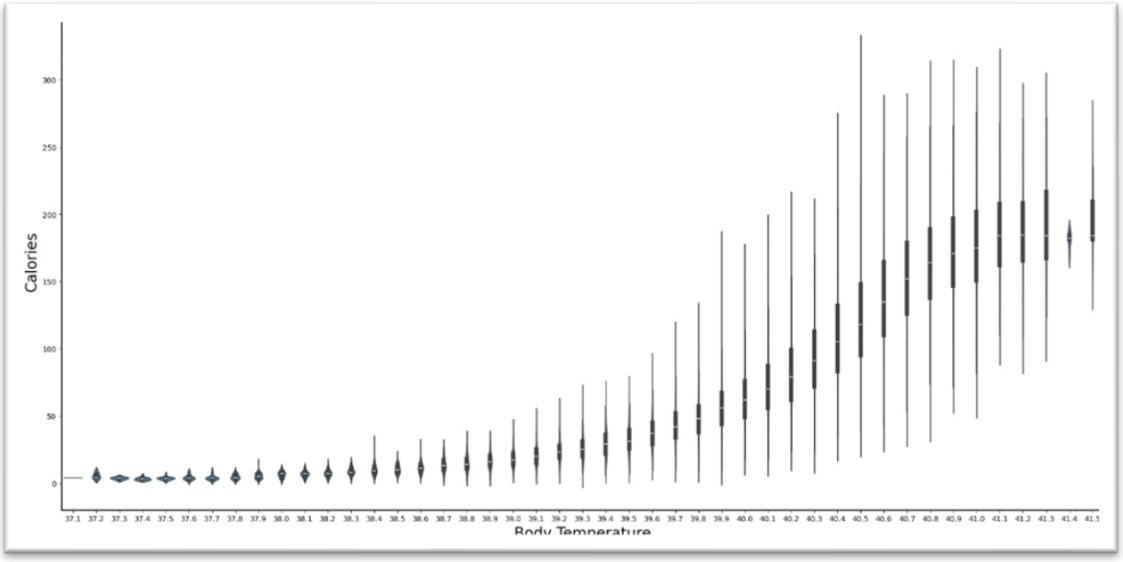


Chart 5

- Line plots are particularly useful for displaying trends and patterns over continuous variables(in this case, the age from 20 to 80
- Males has sharp increase in Calories Burnt in age between 70 to 80. Similarly females has sharp increase in Calories burnt in age 55 to 60.

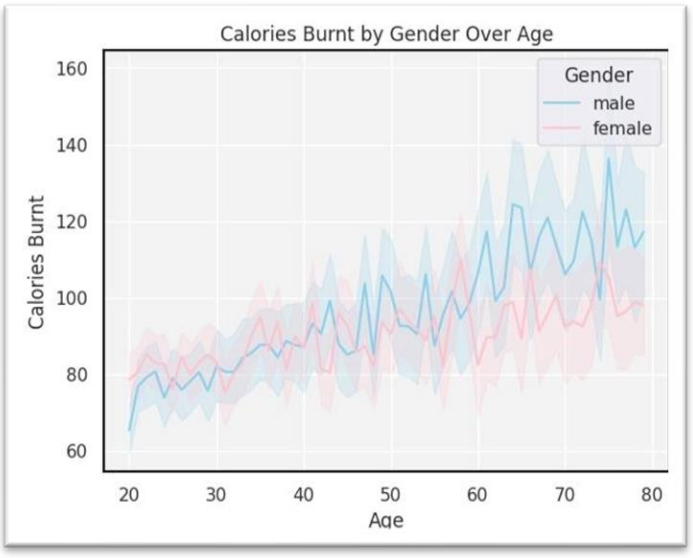


Chart 6

- The chosen chart is a count plot, which is suitable for visualizing the distribution of categorical variables, in this case, 'Gender' and 'Age Group'. The use of hue separates the bars by age group within each gender category.

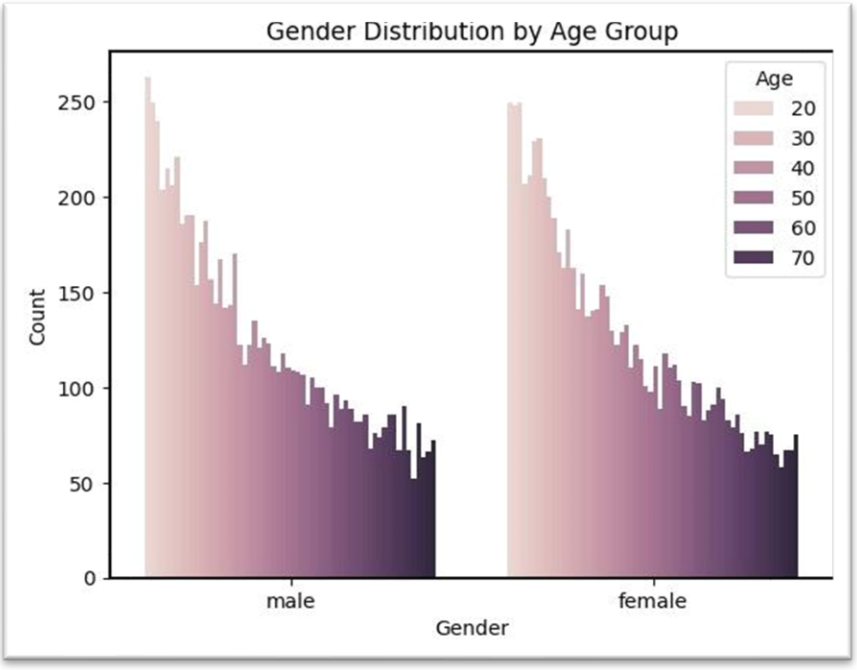


Chart 7

- The selected chart is a scatter plot, with points colored by gender, which is suitable for visualizing the relationship between two continuous variables: 'Heart Rate' and 'Calories', while also considering the gender differences.

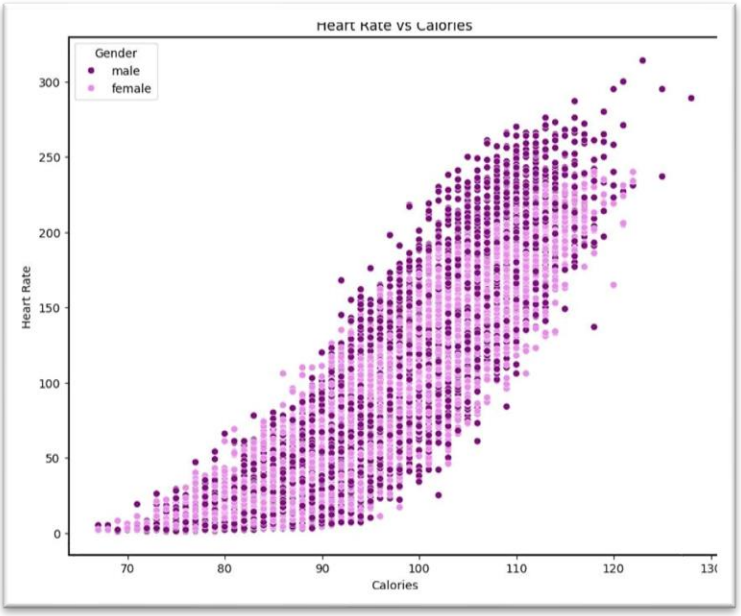


Chart 8

- A scatter plot is a type of mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.
- This plot shows between Age Vs Calories with Weight Variation

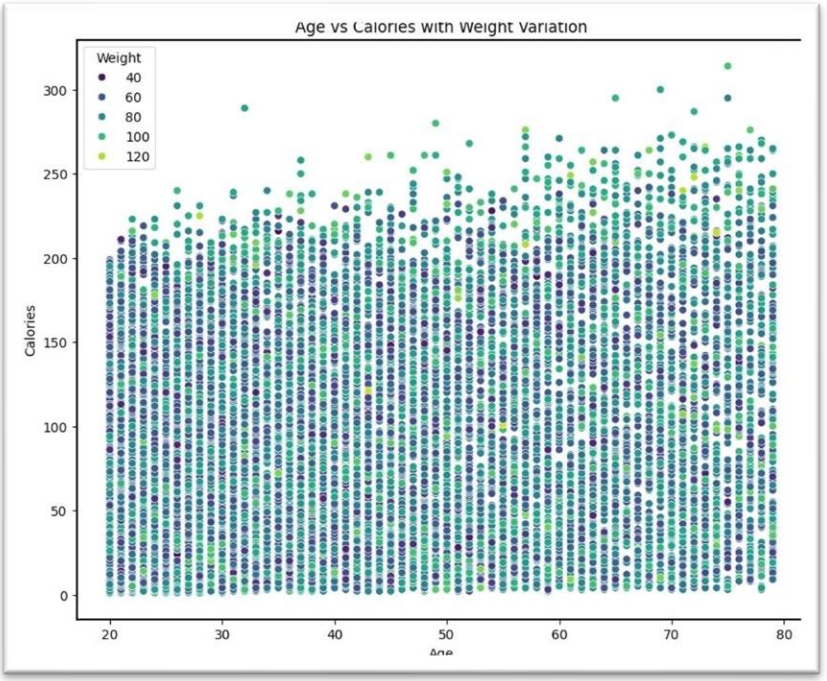


Chart 9

- The specific chart, which consists of three scatter plots showing the relationships between 'Age', 'Weight', and 'Calories', was chosen because it allows for a visual examination of how these variables interact with each other.

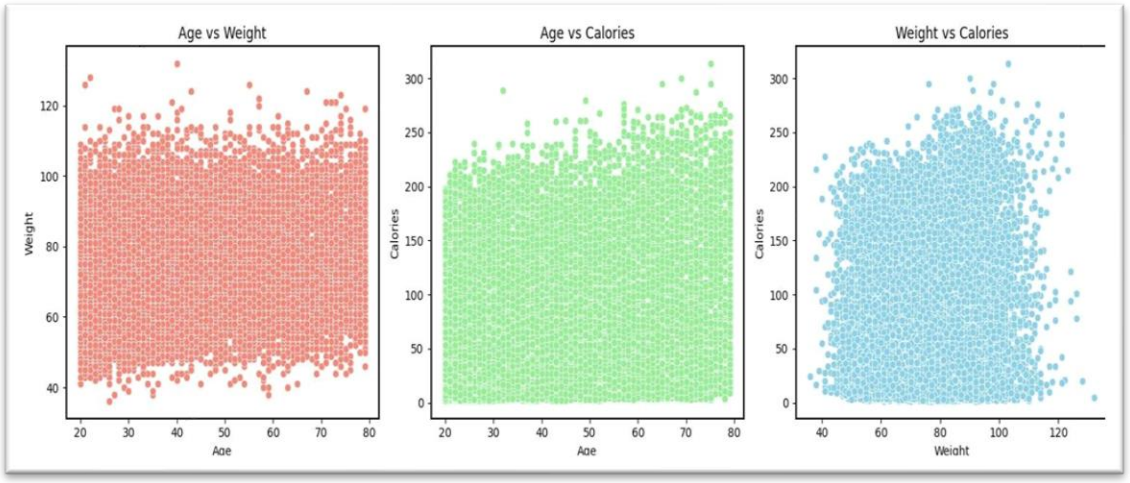


Chart 10

- Heatmaps are effective for displaying the correlation matrix, as they provide a clear visual representation of the strength and direction of relationships between variables.
- Height and Duration, Weight and Duration have a negative correlation.

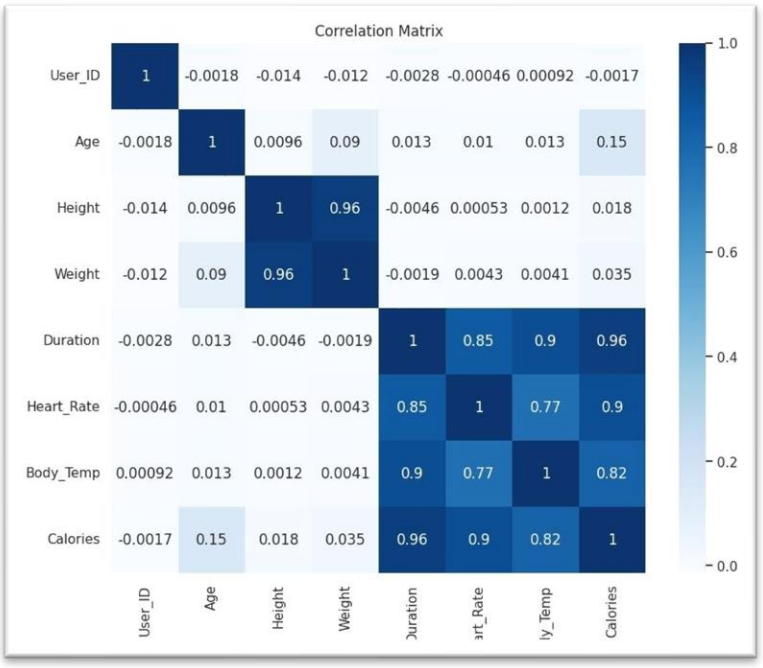


Chart 11

- A joint plot is a type of data visualization that combines multiple plots to display the relationship between two variables, usually a bivariate relationship.
- This plot is between Weight Vs Calories.

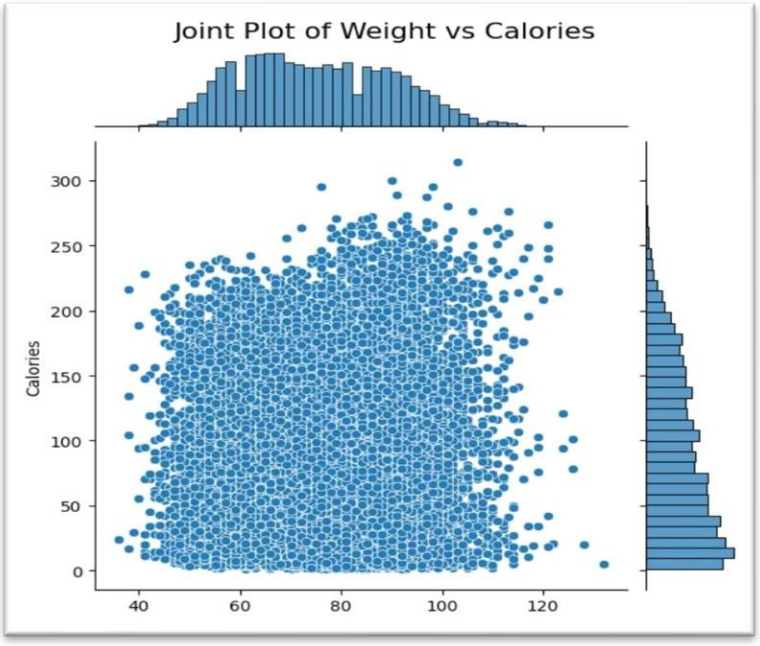
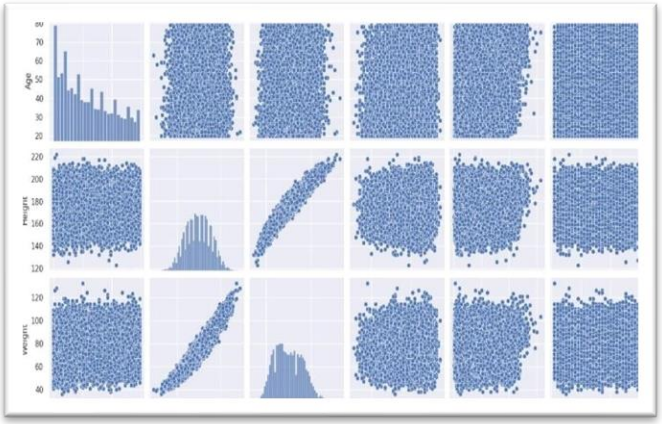
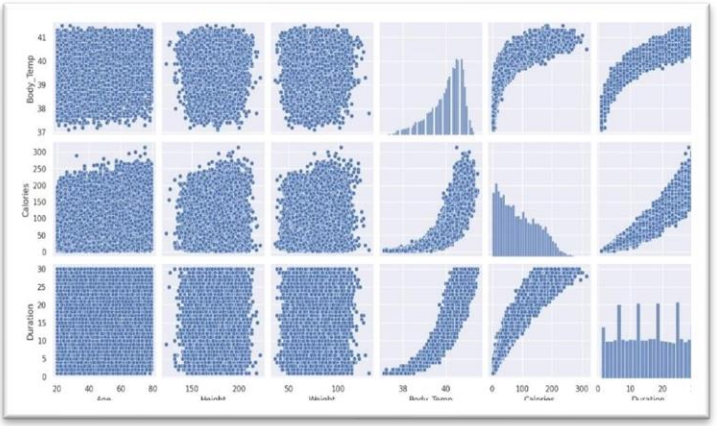


Chart 12

- Pair plots allow us to visualize relationships between every pair of variables in the dataset.
- The scatter plots in the pair plot can reveal patterns of correlation between pairs of variables





References

- We took the dataset from the Kaggle.
- Linear Regression link: https: //www.youtube.com/watch?v=zUQr6HAAKp4
- Decision Tree Regression link: https://www.youtube.com/watch?v=mvveVcbHynE&t=249s
- Random Forest Tree Regression link: https://www.youtube.com/watch?v=DXqxXe3rep0
- Support Vector Machine link: https://www.youtube.com/watch?v=xLkk6MUrvrw