ML Project

Predicting Calories Burned through Exercise with XGBoost

C Naganna Gari Chowde Gowd

M.Tech Aerospace Engineering

IIT Bombay

Contents

1. Introduction

The objective of this project is to create a predictive model for estimating the number of calories burned during exercise based on various individual attributes and exercise parameters. The dataset used comprises exercise records and associated attributes of individuals. The ultimate goal is to develop a reliable machine learning model capable of accurately predicting the calories burned during exercise.

2. Data Collection and Processing:

The project commenced with the loading of two datasets: "calories.csv" and "exercise.csv". These datasets were imported into pandas dataframes for further analysis and processing. By merging the datasets using the common "User_ID" identifier, we were able to combine the exercise data with individual attributes.

3. Data Analysis:

Initial data analysis was conducted to understand the structure of the dataset and the distribution of variables. Statistical measures, such as the mean, standard deviation, and quartiles, were calculated to provide insights into the central tendency and dispersion of the data. Additionally, visualizations were generated to gain a visual understanding of the distributions, using techniques like histograms and density plots.

4. Data Preprocessing:

To prepare the data for modeling, we converted categorical data, such as "Gender," into numerical values using mapping. Moreover, irrelevant features such as "User_ID" were removed to ensure that only pertinent attributes were utilized for training the model.

5. Feature Selection:

The dataset was divided into feature matrices (X) and target variables (Y). The feature matrix consisted of relevant attributes related to exercise and individual characteristics, while the target variable contained the "Calories" burned during the exercise.

6. Data Splitting:

To assess the model's performance, the data was divided into training and testing sets using the train_test_split function. This separation allowed us to train the model on one subset and evaluate its performance on unseen data.

7. Model Training:

The model chosen for this project was the XGBoost Regressor. This decision was based on the algorithm's ability to handle regression tasks effectively. The model was loaded, trained using the training data, and its predictive capacity was evaluated.

8. Evaluation:

The model's performance was evaluated using the Mean Absolute Error (MAE) metric. The MAE quantifies the average absolute difference between the predicted and actual calorie values. Lower MAE values indicate better predictive accuracy.

9. Results:

The project successfully developed a predictive model using the XGBoost Regressor algorithm. The model was able to predict the calories burned during exercise based on the provided attributes. The

evaluation, conducted using the MAE metric, demonstrated the model's effectiveness in estimating calorie burn accurately.

10. Conclusion:

In conclusion, this project showcased the application of machine learning techniques to predict calories burned during exercise. The utilization of the XGBoost Regressor model, coupled with appropriate preprocessing and evaluation, yielded valuable insights into the relationship between individual attributes and exercise parameters in relation to calorie burn. With further refinement, this model could contribute significantly to health and fitness analytics.

11. References:

The project was conducted following established machine learning methodologies and practices. While no external references were directly mentioned, the project's approach aligns with standard data preprocessing, model training, and evaluation practices.

12. Limitations and Future Work:

The model's performance could potentially be enhanced through hyperparameter tuning and cross-validation techniques.

Incorporating additional relevant features and attributes, such as heart rate data or metabolic rates, might improve the model's accuracy.

Expanding the dataset with a broader range of exercise types and conditions could lead to more comprehensive and accurate predictions.