Data Science Project Report on

Calories Burt Prediction Using Machine Learning

Submitted by:

Name: **Most. Maliha Hossain**
Roll: **09-006-09**

Date of submission:
16 february, 2025

CALORIES BURNT PREDICTION USING MACHINE LEARNING APPROACH

ABSTRACT

Calorie burnt prediction by machine learning algorithm" aim to predict the number of calories burnt by an individual during physical activity using machine learning techniques. We collected a dataset that includes features such as heart rate, body temperature, and duration of activity. We used various machine learning models, including XGBoost, linear regression, SVM and random forest, to predict calorie burn based on 15,000 records with nine features. The results indicate that the XGBboost model can accurately predict calorie burn with a minimum mean absolute error of calories. This work contributes to the growing body of research on using machine learning for health and fitness applications and has potential implications for personalized health coaching and wellness tracking. The highest accuracy of training and testing is gained by the XGBboost model with 94.86% with mean absolute error is almost 9.89%.

Purpose

A calories burnt prediction project estimates the number of calories a person burns during physical activities based on factors like activity type, duration, intensity, and personal characteristics (weight, age, gender). The goal is to help individuals track their fitness progress, plan workouts, and achieve specific health goals like weight loss or muscle gain. By using machine learning models trained on activity data, it provides personalized insights into calorie expenditure.

Introduction

Background

As a student of Statistics at the University of Barishal, I have always been fascinated by how data-driven insights can improve decision-making in real-world scenarios. During my studies, I became particularly interested in the application of statistical methods and machine learning to health and fitness. Observing the increasing popularity of fitness tracking and the demand for personalized health insights, I identified a gap in accurately predicting calories burnt during physical activities. This project aims to address that gap, using statistical models to analyze factors like activity type, intensity, and personal attributes. By combining my statistical knowledge with machine learning techniques, this project seeks to provide accurate, personalized predictions for users looking to optimize their fitness routines and health management.

Problem Statement

In today's world, where health and fitness have become a significant part of many people's daily lives, tracking physical activity and calorie expenditure is crucial. Accurately estimating the number of calories burned during physical activities enables individuals to make informed decisions about their exercise routines, diet, and overall health. With the increase in the use of fitness trackers and mobile apps, having an accurate, personalized method of predicting calorie expenditure is more important than ever.

Caloric burn during physical activity is influenced by various factors, including

the type of activity, its intensity, duration, and personal attributes like body weight, age, and gender. While fitness devices and applications already provide estimates for calories burnt, the predictions can often be imprecise.

This limitation calls for a more personalized approach, one that integrates personal data with detailed activity metrics to generate accurate and tailored predictions. This project aims to bridge this gap by developing a machine learning model that can predict the number of calories a person burns during different types of physical activities. By collecting data on activities, personal attributes, and relevant exercise details, the project seeks to create a robust model that provides real-time, personalized insights into caloric expenditure. Through this process, individuals will be able to monitor their energy expenditure more accurately, helping them optimize their workouts and achieve their health goals.

Mothodology

The methodology behind this project involves several key steps, starting with data collection and preprocessing, followed by the selection and training of machine learning models, and concluding with model evaluation and performance analysis.

Data Collection and Preprocessing

The first step in the project is data collection, which involves gathering information about physical activities and personal characteristics that influence calorie burn. The dataset used for this project includes records from a variety of physical activities, such as running, walking, cycling, and strength training. For each activity, the dataset contains the following features:

Age: Affects metabolism and energy expenditure. Younger individuals generally have a higher metabolic rate than older ones.
Gender: Males typically burn more calories than females due to differences in muscle mass and metabolic rate.

Height: Can influence calorie burn as it affects body mass and overall energy needs.
Weight: Heavier individuals burn more calories for the same activity due to increased energy requirements.
Duration: The length of physical activity; longer durations generally result in higher calorie expenditure.
Heart Rate: Higher heart rates indicate more intense physical activity, leading to increased calorie burn.
Body Temperature: Elevated body temperature can reflect higher metabolic activity, influencing calorie expenditure.

Additionally, heart rate data can be included if available, as it is directly correlated with the intensity of exercise and can help refine the calorie burn prediction.

Once the data is collected, the next step is preprocessing. This involves cleaning the data by handling missing values, removing duplicates, and normalizing numerical values. Some activities might have missing or inconsistent data, and it is important to address these gaps to ensure the quality of the model. Data normalization helps scale the values of numerical features (like weight or duration) so that they all fall within a similar range, preventing certain features from disproportionately influencing the model.

Feature Engineering

Feature engineering is a crucial step in the methodology, as it involves selecting the most relevant features and transforming the raw data into formats that can improve model performance. In this project, activity type is typically encoded as a categorical variable, which can be transformed into numerical format using techniques like one-hot encoding. Duration, weight, and intensity are kept as numerical features, with some additional transformations applied, such as calculating calories per minute or adjusting for intensity.

Model Selection & Training

Three machine learning models are implemented:

(i) Linear Regression
A simple regression model that assumes a linear relationship between input features and calories burnt. While computationally efficient, it struggles with complex, nonlinear relationships.

(ii) Random Forest
An ensemble learning method that combines multiple decision trees to reduce overfitting and improve accuracy. It captures complex patterns effectively and is more robust than Linear Regression.

(iii) XGBoost Regressor
An advanced gradient boosting algorithm that improves prediction accuracy using boosted decision trees. It is highly efficient and provides the best performance among the three models.

Model Evaluation

The trained models are evaluated using standard regression metrics:

Mean Absolute Error (MAE): Measures the average difference between predicted and actual calories.
Root Mean Squared Error (RMSE): Captures overall prediction error magnitude.
R-squared (R^2): Indicates how well the model explains variance in calorie expenditure.
6. Results and Insights
1) Linear Regression provides a baseline but lacks flexibility for complex interactions.

2)Random Forest performs significantly better due to its ability to capture nonlinear relationships.
3)XGBoost achieves the best results with lower error rates and higher predictive power.

Based on these findings, XGBoost is recommended for real-world deployment due to its accuracy and efficiency.

Conclusion

This project successfully demonstrates how machine learning models can improve calorie burn prediction compared to traditional estimation techniques. By leveraging Linear Regression, Random Forest, and XGBoost, we identified the best-performing model for this task.

Future improvements may include:

1) Real-time prediction integration with fitness trackers and smartwatches.
2) Incorporation of additional physiological factors, such as metabolism rate and oxygen consumption.
3) Deep learning approaches to enhance accuracy further.

The results highlight the potential of machine learning in personalized health tracking, making calorie estimation more precise and user-specific. This approach can be integrated into fitness applications to help users optimize workouts and achieve their health goals effectively.