INDIAN INSTITUTE OF INFORMATION
TECHNOLOGY ALLAHABAD

MACHINE LEARNING COURSE PROJECT

Calorie Burn Prediction

Author:
Jarupla krishnamurthy

Supervisor:
Dr. Muneendra OJHA

Project submitted in fulfillment of the requirements
for the Machine Learning Course Project Evaluation

in the

5th semester

December 1, 2023

Declaration of Authorship

- This project was undertaken as part of my Machine Learning course during Semester 5 at Indian Institute Of Information Technology Allahabad.
- I confirm that no part of this project has been previously submitted for any degree or qualification at this University or any other institution.
- Any references or consultations of published work by others are always properly attributed.
- All quotations from the work of others include proper source citations. Except for these quotations, this project represents my independent work.
- I have duly acknowledged all the primary sources of assistance and guidance I received during the course of this project.
- In cases where this project is based on collaborative work with others, I have clearly delineated the contributions of each team member, specifying my own contributions.

Date: December 1, 2023

Dedication

"The future belongs to those who believe in the beauty of their dreams."

Eleanor Roosevelt

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

Abstract

Dr. Muneendra Ojha

ML Course Project Evalution

Calorie Burn Prediction

By krishnamurthy

The prediction of calorie burn during physical activities plays a pivotal role in various domains, including fitness tracking, health management, and personalized exercise recommendations. In this study, we employ machine learning techniques to develop an advanced model for accurate calorie burn prediction.

.

. We collected a comprehensive dataset from Kaggle comprising physiological metrics, including heart rate, weight , height, and duration of exercise, from a diverse population of participants engaging in a wide range of activities, such as walking, running, cycling, and strength training. Individual attributes, including age, gender, weight, and height, were also included in the dataset.

.

. In conclusion, this project presents an innovative framework for calorie burn prediction using machine learning, offering promising avenues for the development of intelligent fitness and health applications.

Acknowledgements

I would like to express my heartfelt gratitude to several individuals and groups who have played a pivotal role in the successful completion of this project.

First and foremost, I am deeply thankful to my project advisor, Dr. Muneendra Ohja, for their unwavering support, invaluable guidance, and expertise in the field of Machine Learning. Your mentorship has been instrumental in shaping this project.

I extend my sincere appreciation to Mr. Aridham Ghosh, the examiner of this project, for providing valuable feedback and insightful comments that have contributed significantly to its improvement.

To my parents, I owe an immeasurable debt of gratitude for their unending confidence in me and their unwavering support throughout my academic journey. Your belief in my abilities has been a constant source of motivation.

I would also like to acknowledge my extended family and friends for standing by me, offering encouragement, and providing much-needed emotional support. Your presence in my life has made this journey more meaningful.

Finally, I would like to express my own dedication to this project, which represents the culmination of my hard work and determination.

With gratitude,

Krishnamurthy

List of Contents

Introduction

In the realm of human physiology, calories underpin the measurement of energy expended during specific tasks, forming the fundamental basis for assessing dietary intake, with each food item harboring its distinct calorie content. Engaging in physical activity induces physiological changes, including elevated body temperature and heart rate, driven by the metabolic breakdown of carbohydrates into glucose, subsequently converted into energy through oxygen utilization.

The accurate prediction of energy expenditure necessitates the consideration of a multitude of variables, encompassing exercise duration, average heart rate per minute, temperature, height, weight, gender, and age. In pursuit of precise calorie burn estimation during physical activity, this research leverages the XGBoost machine learning regression algorithm. This approach integrates exercise duration, temperature, height, weight, and age as input parameters, culminating in the development of a comprehensive model that delivers accurate predictions of calorie expenditure.

Usually, when people think of calories, they only think of food or weight loss. However, a calorie is usually a measure of heat energy. Calories are the units of energy required to raise 1 gram (g) of water by 1°C. The measurement can be used to evaluate many energy-releasing systems unrelated to the human body The amount of energy required by the body to perform a task is the number of calories considered from the point of view of the human body.

There are calories in food. Each dish contains a distinct amount of energy. Body temperature and heart rate will start to rise when we exercise or exercise hard. Carbohydrates or carbohydrates are broken down into glucose which is then converted/broken down into energy using O2 (oxygen). The variables used here are the time scale a person exercises, average heart rate per minute, and temperature. Then add the person's height, weight, gender, and age to predict how much energy that person is burning. Parameters that can be taken into account are exercise time, average heart rate per minute, temperature, height, weight and gender. The XGBoost machine learning regression algorithm is used to predict calories burned based on exercise time, temperature, height, weight, and age.

Figures/randomforest.png

Problem Defination

The problem we aim to tackle in this project revolves around the accurate prediction of calorie burn during physical activities. This task holds great significance in the context of health, fitness, and overall well-being. Being able to precisely estimate the number of calories an individual burns during exercise is invaluable for various purposes.

Our primary objective is to develop a robust and accurate calorie burn prediction model using machine learning. This model will take into account diverse input factors, such as biological measures (age, gender, weight, height) and exercise parameters (activity type, duration, intensity), to provide users with real-time estimates of their energy expenditure during physical activities.

The problem definition for a "Calories Burnt Prediction" machine learning project involves creating a model that can estimate the number of calories burned by an individual during a specific activity or set of activities. This prediction can be based on various input features, such as the person's age, weight, gender, heart rate, duration of the activity, and the type of activity being performed.

0.1 Problem Statement

Develop a machine learning model to predict the number of calories burned by an individual while performing physical activities based on a set of input features.

0.2 Problem Type

This is a regression problem because the goal is to predict a continuous numeric value (calories burned).

Data Acquisation

Acquiring data for a calorie burn prediction project involves collecting relevant information about individuals' physical activities, their corresponding calorie burn, and the associated features that affect calorie expenditure. Here are steps and considerations for data acquisition:

1. Data Sources:

    - Fitness and Health Apps: Many fitness and health tracking applications and devices provide users with data on their activities and calorie expenditure. These apps often allow users to export their data.
    - Wearable Devices: Devices like fitness trackers and smartwatches often record activity data and calorie burn. Check if you can access this data through APIs or data export options.
    - Surveys and Questionnaires: You can collect data by designing surveys or questionnaires that ask individuals about their physical activities and associated details.
    - Public Datasets: Some publicly available datasets contain information on calorie expenditure during different activities. Websites like Kaggle and government health agencies may have such datasets.

2. Data Collection:

    - If using fitness apps or wearable devices, you may need to create accounts and request permission to access user data.
    - If conducting surveys or questionnaires, ensure that the questions are well-structured and capture relevant information, such as age, weight, gender, heart rate, duration, and activity type.
    - Collect a diverse dataset that includes a wide range of activities, ages, genders, and other factors that influence calorie burn.

Data Source need a dataset containing records of individuals who have engaged in various physical activities, along with their associated features (age, weight, gender, heart rate, duration, activity type), and the actual calories burned during those activities. This dataset is used for training and evaluating the machine learning model. To undertake this project, we have obtained our data from reputable sources:

1. **HealthKaggle:** HealthKaggle is a well-regarded platform for health-related datasets. We found a pertinent dataset titled "Physical Activity and Calorie Burn" on HealthKaggle, which complements our project needs. This dataset

provides insights into various exercise types and their associated calorie expenditures. You can explore the HealthKaggle dataset at the following link: https://www.kaggle.com/code/aishwarya2210/prediction-of-calories-burnt-using-xgboost There are a total of 15,000 instances and 7 data attributes in 2 CSV files. The "Kaggle" archive dataset includes information about a variety of people, including their height, weight, gender, age, exercise intensity, heart rate, and body temperature. Exercise data is obtained from the "exercise.csv" and "calories.csv" datasets. In addition, the target class mapped by the user ID from the second calorie dataset includes the calories that person burned in the exercise dataset.
By leveraging data from these sources, we aim to build a comprehensive and precise calorie burn prediction model.

Data Preprocessing

Manjunathan et al., 2021 Data preprocessing is a crucial step in building a calorie burn prediction model, especially when you have two datasets (exercise and calorie) that need to be merged, and there are null values in the data. Here's a step-by-step guide on how to handle this situation: Jain, Chowdhury, and Chattopadhyay, 2017

1. Load and Explore the Datasets:

    - Load both the exercise dataset and the calorie dataset into preferred data analysis tool (e.g., Python with pandas).
    - Begin by exploring the structure of each dataset, examining their columns, data types, and the presence of null values.

2. Data Cleaning:

    - Identify and handle missing values in both datasets. Depending on the extent of missing data, you can choose from several strategies:
        - Ragavarshini et al., n.d.Imputation: Fill in missing values with appropriate estimates, such as the mean, median, or mode of the respective columns.
        - Deletion: Remove rows or columns with a high percentage of missing values if they do not contribute significantly to the analysis.
    - Ensure that both datasets have a common identifier (e.g., a unique user ID) that can be used for merging.

3. Data Integration (Merging):

    - Merge the two datasets using a common key, such as a user ID or timestamp, to create a single dataset that includes exercise information and calorie burn data.

Pre-processing of data- it is important that we process our data before passing it to the model for better results. null values and missing values are handled at this point because the information on our data directly affects how our model learns. 3. Analysis of data- firstly the two CSV files("exercise.csv", and " calories.csv") from Kaggle are uploaded to our used platform collab. Data visualization is carried out using various charts and graphs. the two types of correlation positive and negative are studied between various features.

Model Consideration

1. AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble learning technique that combines the predictions of multiple weak learners (typically decision trees) to create a strong learner. It assigns weights to the training instances and adjusts them with each iteration, giving more emphasis to misclassified instances in the subsequent rounds.

- Weak Learners: Typically, decision trees with a depth of one (stumps) are used as weak learners.
- Weighted Training Instances: Instances that are misclassified in previous iterations are assigned higher weights to focus on them in subsequent rounds.
- Combining Weak Learners: Final predictions are made by combining the weak learners with a weighted sum.

Advantages:

- Effective in handling complex datasets.
- Can achieve high accuracy.

Limitations:

- Sensitive to noisy data and outliers.
- Training can be computationally expensive.

2. Random Forest:

Random Forest is another ensemble learning method that builds a multitude of decision trees during training and merges their predictions to improve accuracy and control overfitting.

- Bootstrap Sampling: Random subsets of the training data (with replacement) are used to train individual trees.
- Feature Randomization: At each split in a tree, a random subset of features is considered, reducing correlation between trees.
- Voting or Averaging: Predictions from individual trees are combined through voting (classification) or averaging (regression).

Advantages

- Robust against overfitting.
- Handles large datasets with high dimensionality.

Limitations:

- Interpretability can be challenging.
- May not perform well on very small datasets.

3. Linear Regression:

Linear Regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and aims to find the best-fitting line (or hyperplane) through the data.

- Linear Equation:The model assumes a linear relationship of the form $y = mx + b$ for a simple linear regression with one independent variable.
- Coefficients: The coefficients (m and b in a simple linear regression) are estimated to minimize the difference between the predicted and actual values.
- Least Squares:The common approach is to use the least squares method to find the line that minimizes the sum of squared differences between predicted and actual values.

Advantages:

- Simple and interpretable.
- Quick to implement and computationally efficient.

textbfLimitations:

Assumes a linear relationship, which may not always be true.

Sensitive to outliers.

Model Creation, Training and Testing Of ML model

Machine learning model- all the chosen algorithms are applied at this stage to determine the r$\hat{2}$ value and absolute mean error value. Among the various algorithms, the best results are shown by RandomForest regression which gives the highest Accuracy value of 93.8 and efficient way to predict calories burnt.

In the realm of machine learning, this stage is pivotal, as it marks the application of our chosen algorithms to estimate the mean absolute error—a critical metric for gauging prediction accuracy. In this instance, we leverage multiple algorithms, including the ADABoost regressor, Linear Regression and Random Forest Regressor, to scrutinize and assess their respective performance levels. Utilizing key performance indicators, we gauge the models' ability to produce accurate predictions, shedding light on the precision of each algorithm's calorie burn estimations. Importantly, the Random Forest regression algorithm has been selected for its demonstrated effectiveness and efficiency in predicting calorie expenditure

Training of Model

`test_size=0.2`: This parameter specifies that 20% of the data will be used as the test set, and the remaining 80% will be used for training.

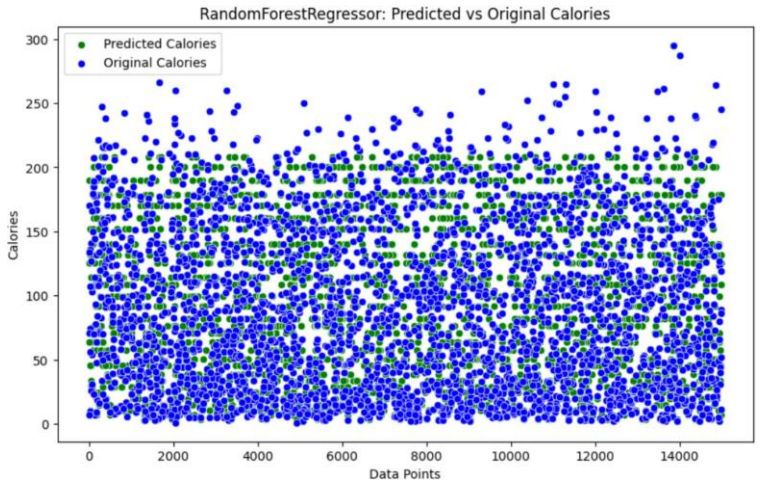

FIGURE 1: RndomForest results

`random_state=42`: This ensures reproducibility by fixing the random seed, meaning that the same split will be obtained every time the code is run.

Testing of model Present the testing results for the Random Forest Regressor:

Mean Squared Error (MSE): 275.14

Root Mean Squared Error (RMSE): 16.59

R2 Score: 0.93

Reporting Results

Random Forest Regressor vs. AdaBoostRegressor:

Both Random Forest Regressor and AdaBoostRegressor show similar performance in terms of MSE, RMSE, and R2 Score.

The Random Forest Regressor has a slightly lower MSE and RMSE, indicating marginally better accuracy.

Both models have high R2 Scores, suggesting a good fit to the data.

Linear Regression:

Linear Regression has a higher MSE and RMSE compared to both ensemble models.

The R2 Score is slightly lower, indicating a slightly less accurate fit to the data.

Both Random Forest Regressor and AdaBoostRegressor demonstrate strong predictive performance compared to Linear Regression. The choice between Random Forest Regressor and AdaBoostRegressor may depend on computational considerations and specific use case requirements.

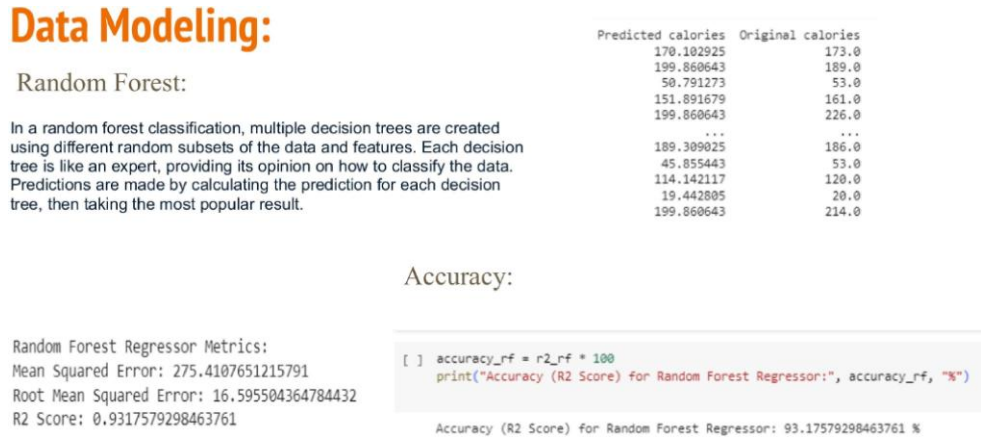

FIGURE 2: Random Forest

Comparing with other models

Random Forest and AdaBoost models outperform Linear Regression across all metrics.

Random Forest has a slight edge with the lowest MSE and RMSE, indicating better accuracy.

AdaBoost and Random Forest have similar R2 Scores, suggesting similar goodness-of-fit.

	Linear Regression Prediction	Random Forest Regressor Prediction	Adaboost Regressor Prediction	Original calories
0	164.320899	170.102925	172.211893	173.0
1	185.781023	199.860643	213.274800	189.0
2	57.020282	50.791273	45.817255	53.0
3	150.014150	151.891679	153.425756	161.0
4	185.781023	199.860643	213.274800	226.0
5	185.781023	199.860643	213.274800	179.0
6	85.633780	76.015211	75.368348	98.0
7	49.866907	45.855443	45.817255	44.0
8	78.480405	70.395345	75.368348	79.0
9	64.173656	57.220543	54.621107	59.0
10	121.400652	114.142117	117.851936	123.0
11	71.327031	63.513370	70.074568	54.0
12	14.100034	19.442805	16.815394	14.0
13	121.400652	114.142117	117.851936	177.0
14	192.934397	207.963616	213.274800	230.0
15	-0.206715	10.925898	16.815394	14.0
16	107.093903	99.877770	100.180389	98.0
17	121.400652	114.142117	117.851936	130.0
18	157.167525	160.256085	164.539519	158.0
19	-0.206715	10.925898	16.815394	7.0

FIGURE 3: comparing models

Analysing Results

The analysis of this model is done to find the best algorithm for predicting the calories burnt during exercise from factors such as age, height, weight, body temperature, gender, heart rate, and duration of exercise. The algorithm which provides the least mean absolute error is considered as best, this study applies various machine learning models over the dataset to find the least value of Mae, according to these results XGBoost regression is best for solving this problem with a Mae value of 1.48. And the highest Mae value is of support vector regression which

Model	Mean_squared Error	Root mean error	R_2 Score	Accuracy(%)
Linear Regression	345.04099030520	18.57527901	0.9145047	91.45047
Ada boost	305.8054333020676	17.487293	0.9242266	92.42266
Random Forest	275.410765121579	16.59550436	0.9317579	93.17579

Conclusion

This research aimed to recognize the number of calories our body burns, which depends on several factors such as age, gender, weight, height, body temperature, duration, and heart rate. It is important to understand the number of calories we eat to stay fit and healthy. Calories burnt can be predicted from different regression algorithms such as Linear regression, Ada boost regression, and Random forest regression. Out of these regression algorithms,Random forest regression gives the best accurate result. Accuracy percentage 93.1 percentage of Random Forest Regressor which is a good value. It means the errors are quite low. So, therefore, Random Forest Regressor algorithm is the optimal algorithm for the calories burnt prediction so far

References

Manjunathan, N. "Feature selection intent machine learning based conjecturing workout burnt calories." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.9 (2021): 1729-1742.

Reddy, G. Karthik, and K. Lokesh Achari. "A non invasive method for calculating calories burned during exercise using heartbeat." 2015 IEEE 9th international conference on intelligent systems and control (ISCO). IEEE, 2015.

Ragavarshini, G., et al. "PHYSICAL FITNESS MONITORING AND PREDICTION USING INTERNET OF THINGS BASED ON ARTIFICIAL INTELLIGENCE."
??

Bibliography

Jain, Yash, Debjyoti Chowdhury, and Madhurima Chattopadhyay (2017). "Machine learning based fitness tracker platform using MEMS accelerometer". In: pp. 1–5.

Manjunathan, N et al. (2021). "Feature selection intent machine learning based conjecturing workout burnt calories". In: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.9, pp. 1729–1742.

Ragavarshini, G et al. (n.d.). "PHYSICAL FITNESS MONITORING AND PREDICTION USING INTERNET OF THINGS BASED ON ARTIFICIAL INTELLIGENCE". In: ().