Calorie Burn Prediction: A Machine Learning Approach using Physiological and Environmental Factors

Author Info
Niharikareddy Meenigea
Data Analyst
Virginial International University
USA

Abstract

Today, people have very tight schedules due to changes in their lifestyle and work commitments. But it requires regular physical activity to stay fit and healthy. People do not pay attention to their eating habits leading to obesity. Obesity is becoming a big and widespread problem in today's lifestyle. This makes people choose diets and do a lot of exercise to stay fit and healthy. The main thing here is that everyone should have complete knowledge of calories consumed and burned, it is easy to keep track of their calories as it is available on product labels or on the internet. Tracking calories burned is the tricky part because there are so few devices for this. Calories burned by an individual based on the MET chart and formula. The main program of this study is to predict calories burned using a boost XG regression model like an ML (machine learning) algorithm to display accurate results. The model is given more than 15,000 data and its mean absolute error is 2.7. This error will improve over time by providing additional data for the enhanced XG regression model.

Keywords: Machine Learning, Environment, Data Source, Data Analysis and XGB Regressor.

INTRODUCTION

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.

METHODOLOGY

To determine how many calories an individual will burn, this study involved gathering the right data set to train our machine learning models. Pre-processing of records is necessary before performing the operation that provides statistics. After that, the data processing is complete and the data is organized into diagrams and graphs using a number of visualization techniques. Here, we use the XG Boost regressor as the ML (machine learning) model to compare and then evaluate these models.

Work Flow:

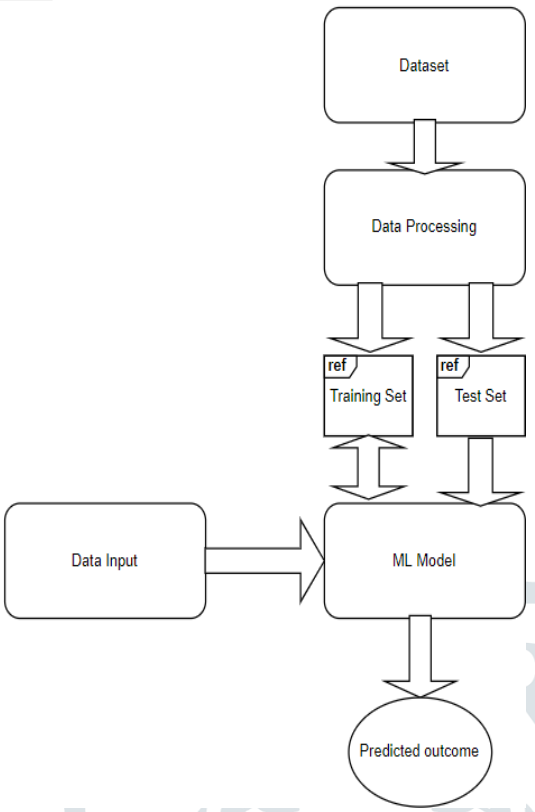

Figure 1 - Work Flow

Data sourse:

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.

Table 1: Attributes and their values

Attribute	**Function**
Gender_individual	Gender (female : 1, male : 0)
Age_Individual	Age in years
Height_Individual	Height of a person
Weight_Individual	Weight of a person
Heart_rate_Individual	Average heart rate of an individual during exercise (normal heart rate 75 beats/min)
Body_temp_individual	Average body temperature recorded over the entire course workout (above 37 degrees Celsius)
Duration_individual	Training time in minutes.
Calories_individual	The total amount of calories burned while workout.

There are 2 datasets in CSV file format that need to be uploaded together, mainly used for online data processing. The data frame is used for processing and analysis purposes. This provides some statistical measure of the data.

```
calories_data.head()
```

	User_ID	Gender	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
0	14733363	0	68	190.0	94.0	29.0	105.0	40.8	231.0
1	14861698	1	20	166.0	60.0	14.0	94.0	40.3	66.0
2	11179863	0	69	179.0	79.0	5.0	88.0	38.7	26.0
3	16180408	1	34	179.0	71.0	13.0	100.0	40.5	71.0
4	17771927	1	27	154.0	58.0	10.0	81.0	39.8	35.0

Figure 3 - data frame

The description of the above dataset can be seen in figure 4.

```
# get some statistical measures about the data
calories_data.describe()
```

	User_ID	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
count	1.500000e+04	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000
mean	1.497736e+07	42.789800	174.465133	74.966867	15.530600	95.518533	40.025453	89.539533
std	2.872851e+06	16.980264	14.258114	15.035657	8.319203	9.583328	0.779230	62.456978
min	1.000116e+07	20.000000	123.000000	36.000000	1.000000	67.000000	37.100000	1.000000
25%	1.247419e+07	28.000000	164.000000	63.000000	8.000000	88.000000	39.600000	35.000000
50%	1.499728e+07	39.000000	175.000000	74.000000	16.000000	96.000000	40.200000	79.000000
75%	1.744928e+07	56.000000	185.000000	87.000000	23.000000	103.000000	40.600000	138.000000
max	1.999965e+07	79.000000	222.000000	132.000000	30.000000	128.000000	41.500000	314.000000

Figure 4 - Data Description

The count of the gender is equally distributed in the dataset which can be seen in the below figure (figure 3).

```
# plotting the gender column in count plot
sns.countplot(calories_data['Gender'],palette="Set1")
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorator
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f4303134110>
```

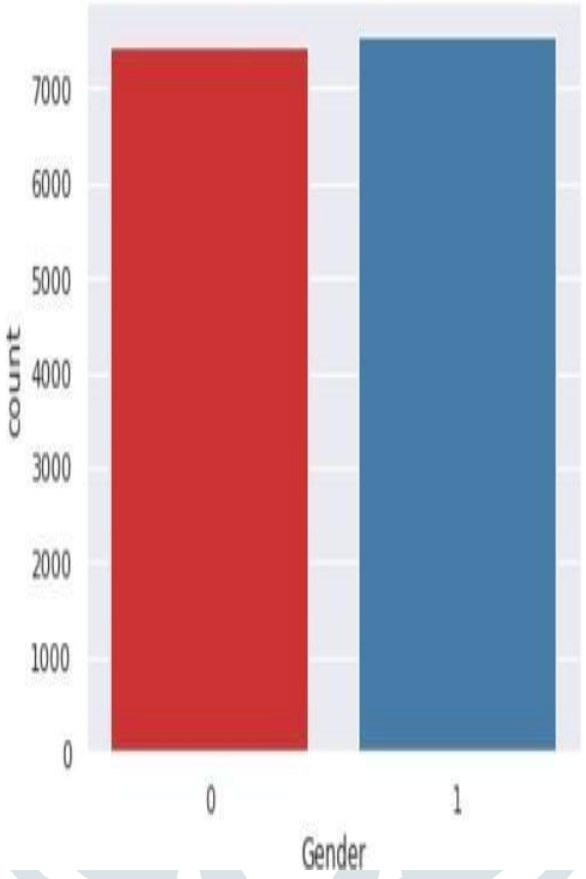

Figure 5 - Gender Distribution

In addition, we have mean values for age, height and weight shown in the figure below.

```
# finding the distribution of "Age" column
sns.distplot(calories_data['Age'], color = "R")
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `di
  warnings.warn(msg, FutureWarning)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2681: MatplotlibDeprecat
  color=hist_color, **hist_kws)
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:177: MatplotlibDeprecati
  kws["color"] = to_rgba(color, alpha)
<matplotlib.axes._subplots.AxesSubplot at 0x7f43031cbe10>
```

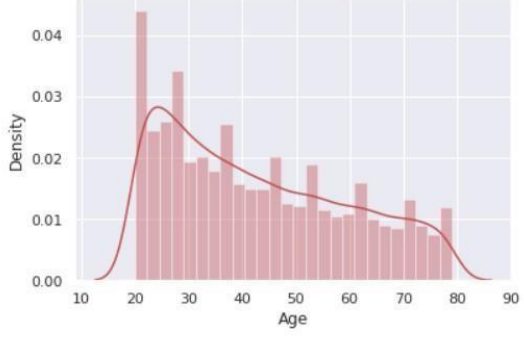

Figure 6 - Mean Age

For age, more values can be observed from 20 to 30 years old. Get an in-depth look at the average curve we created using 15,000 instances. A decreasing curve can be seen as people tend not to exercise at an older age

Figure 7 - Average Height and Weight

Then the relationship between the different records will be examined. There are two types of correlation: positive correlation and negative correlation. The amount of calories burned will increase as the workout time increases. The values are therefore equivalent, that is, have the same meaning and are indisputably linked.

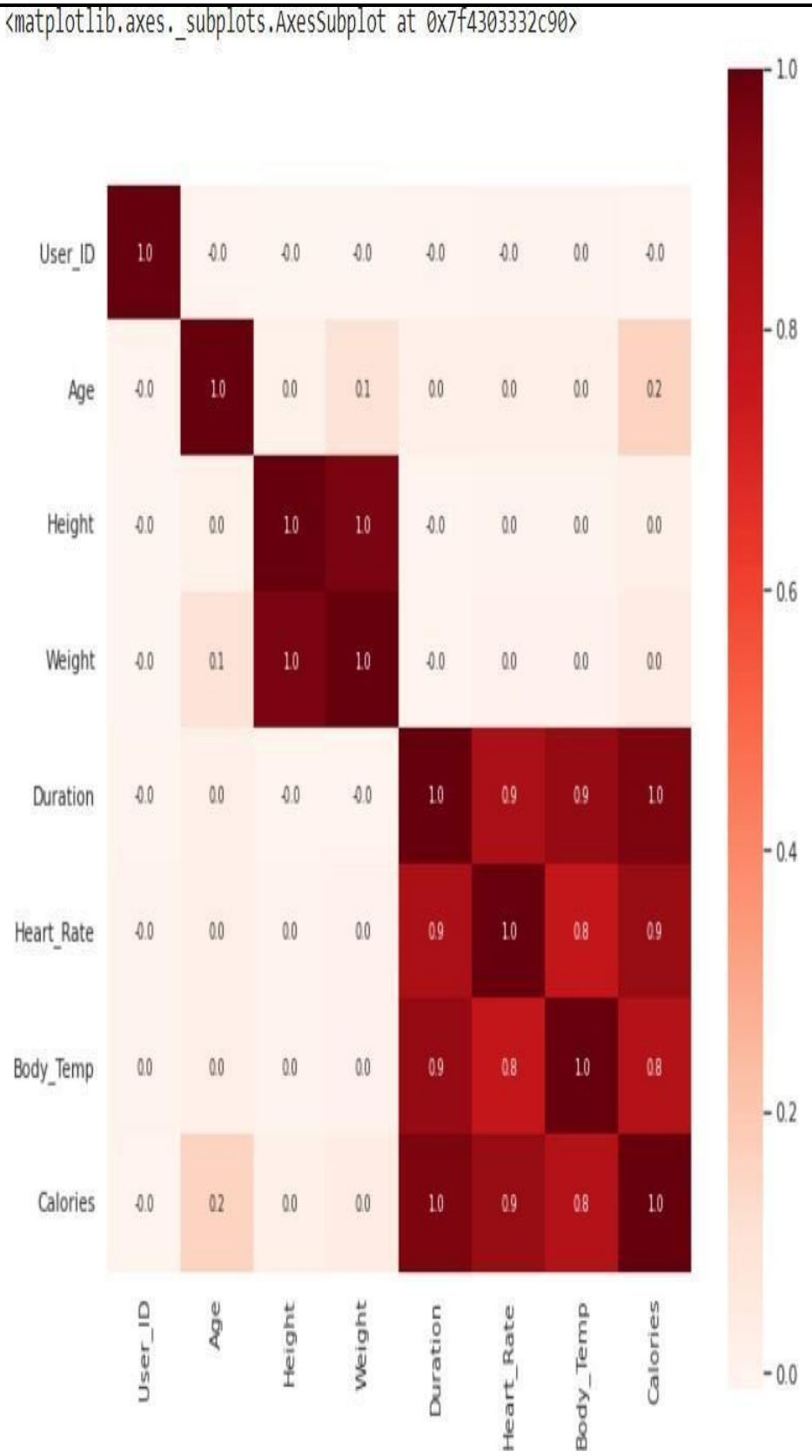

Figure 8 - Correlation of attribute

A. Data collection

Data retrieval is the first step. Kaggle is the data store we use. It is loaded into the Collab program. The information collected is both categorical and numerical.

B. Data Preprocessing

15,000 instances and 7 data attributes are contained in two csv files ("exercise.csv" and "calorie.csv."

Each person's attributes are included in the Kaggle data collection.

Including their size , weight, gender, age, exercise duration, heart rate, and body temperature.

Data preprocessing is an important step in the machine learning process because the quality of the data and the insights that can be extracted from that data directly affect the trainability of the model ours. It is important that we preprocess our data before providing it to our model as output.

C. Data Analysis

Colab, the platform used for the processing, requires the upload of two dataset csv files ("exercise.csv" and "calorie.csv"). The average body temperature is 40. People who exercise will have a higher body temperature. Heart rate and coronary temperature were the most important results of this analysis. The data is then visualized using a few tables and charts. Two types of correlation, positive and negative, were then studied between different records. Then load the XGB Regressor model and evaluate the prediction using the test data. This test data and the calories burned for the X test are run in the model. Similarly, compare our model's expected values with the original values.

D. Machine Learning model

This is the step where we apply our chosen algorithm (in this case, the XGBoost regressor) to determine the mean absolute error. The XGB regression procedure was used and the results obtained. For this, we use indicators that indicate the level of errors of the version

The XGBoost regression algorithm has been proven to be an effective and efficient method for predicting calories burned.

E. Evaluation

This dataset was analyzed to make predictions about how many calories were burned based on exercise duration as well as factors such as age, gender, body temperature and heart rate at different time points. different points during exercise. We are looking for a machine learning model with lower mean absolute error that produces more accurate results using these machine learning methods.

RESULT

A. First five rows of the dataset:

Tabular view of the first 5 records of the dataset:

```
calories_data.head()
```

	User_ID	Gender	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
0	14733363	male	68	190.0	94.0	29.0	105.0	40.8	231.0
1	14861698	female	20	166.0	60.0	14.0	94.0	40.3	66.0
2	11179863	male	69	179.0	79.0	5.0	88.0	38.7	26.0
3	16180408	female	34	179.0	71.0	13.0	100.0	40.5	71.0
4	17771927	female	27	154.0	58.0	10.0	81.0	39.8	35.0

Figure 9 - data frame

B. Conversion of text data to numerical:

```
[21] calories_data.replace({"Gender":{'male':0,'female':1}}, inplace=True)

[22] calories_data.head()
```

	User_ID	Gender	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
0	14733363	0	68	190.0	94.0	29.0	105.0	40.8	231.0
1	14861698	1	20	166.0	60.0	14.0	94.0	40.3	66.0
2	11179863	0	69	179.0	79.0	5.0	88.0	38.7	26.0
3	16180408	1	34	179.0	71.0	13.0	100.0	40.5	71.0
4	17771927	1	27	154.0	58.0	10.0	81.0	39.8	35.0

Figure 10 - Gender data conversion

C. Slicing of Data:

Splitting the data into training data and Test data

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

[ ] print(X.shape, X_train.shape, X_test.shape)

    (15000, 7) (12000, 7) (3000, 7)
```

Figure 11 - Slicing of data

D. Training data in XG Boost Regressor

```
[ ] # loading the model
    model = XGBRegressor()

[ ] # training the model with X_train
    model.fit(X_train, Y_train)

    [16:36:32] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
    XGBRegressor()
```

Figure 12 - Training of data

E. Mean absolute error:

```
[ ] test_data_prediction = model.predict(X_test)

[ ] print(test_data_prediction)

    [129.06204  223.79721   39.181965 ... 145.59767   22.53474   92.29064 ]
```

Mean Absolute Error

```
[ ] mae = metrics.mean_absolute_error(Y_test, test_data_prediction)

[ ] print("Mean Absolute Error = ", mae)

    Mean Absolute Error =  2.7159012502233186
```

CONCLUSION

XGB Regressor produces results more accurate results. The mean absolute error indicates that the absolute error should be as small as possible. It is nothing more than the difference between the observed values and the values predicted by the models. 2.71 is a good value for the mean absolute value that XGBRegressor gives us. The error rate is quite low. Therefore, we can say that XG Boost Regressor is the best model to predict calories consumed. The flexibility of the proposed technique can also be improved with variations.
In this study, we focused on seven main factors that influence how many calories our bodies burn, but there are other factors that play a role as well. It is also important to understand how many calories we consume if we want to stay healthy and fit. Alternatively, ML can be used to build this (machine learning). A user interface is also required for the user to enter their values and get results showing how many calories they have burned. Moreover, we can create a fully functional app with all these features and our recommended diet and exercise program.

REFERENCES

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. *Cambridge University*.
[2] MacKay, D.J, & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*.Cambridge University press.
[3] Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*
[4] https://www.kaggle.com/fmendes/fmendesdat263xdemos
[5] https://machinelearningmastery.com/xgboost-for- regression/
[6] https://www.medicalnewstoday.com/articles/319731#factors-influencing-daily-calorie-burn-and-weight-loss
[7] https://zenodo.org/record/6365018
[8] https://devhadvani.github.io/calorie.htm
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5496172
[10] https://www.jetpac.com/
[11] World Health Organization. (2011, October) Obesity Study. [Online]http://www.who.int/mediacentre/factsheets /fs311/en/index.html