Premier University, Chittagong

Department of Computer Science and Engineering
Chittagong, Bangladesh

Project Final Report

Calories Burnt Prediction

Group Name: BurnIt

Submitted by:
Name MD. Emran Hossen
ID: 1803510201694
Name Inquyad Bin Mahbub
ID: 2103910202110
Name Mohammad Ashikur Rahman
ID: 2103910202111

Submitted to:
Avisheak Das
Lecturer
Dept. of Computer Science and Engineering
Premier University, Chittagong

Abstract

In today's world, people are having very tight schedules due to the changes in their lifestyles and work commitments. But it requires regular physical activity to stay fit and healthy. People do not concentrate on their food habits, leading to obesity. Obesity is becoming a major and common problem in today's lifestyle. This leads people to choose their diet and do an equal amount of exercise to stay fit and healthy. The main part here is people should have enough knowledge about their calorie consumption and burn, keeping track of their calorie intake is easy as it's available on the product label or the internet. Keeping track of calories burnt is a difficult part as there are very few devices for that. Calories burned by an individual are based on MET charts and formulas. The main agenda of this study is a prediction of the burnt calories with the help of an XGboost regression model as the ML (machine learning) algorithm to show accurate results. The model is fed with more than 15,000 data and its mean absolute errors will become better over time by feeding the XGboost regression model with more data.

Contents

List of Tables

List of Figures

1 Introduction

The human body needs calories to survive, without energy he would die, people absorb this energy from food and drink. From some studies [10], we learned that if people consumed only the number of calories needed every day, they would probably have healthy lives. Calorie consumption that is too low or too high will eventually lead to health problems. And that's why the human body needs to burn calories.

In our present time most often, when individuals think of calories, they only think of food or weight reduction. But that is not the real case in measuring calories. The amount of energy in an item of food or drink is measured in calories. When we eat and drink more calories than we use up. Our bodies store the excess as body fat, if this continues, over time we may put on weight. A calorie, however, is often a measure of heat energy. Calories are the units of energy needed to elevate 1 gram(g) of water by 1°C. The measurement may be used to assess a variety of energy-releasing systems unrelated to the human body. The amount of energy needed by the body to carry out a task is how many calories are considered from the perspective of the human body. From some research [13] we can say that a calorie is basically the amount of energy that is needed to raise 1 gram (g) of water by 1°C. This measurement can be applied to lots of different energy-releasing mechanisms outside of the human body, for the human body calories are a measure of how much energy the body needs to function. There are calories in food. Each and every item has a distinct quantity of energy included in it since various foods have varying calorie counts. As a guide [8], an average man needs around 2.500kcal (10.500kj) a day to maintain a healthy body weight. For the average woman, that figure is around 2.000kcal (8.400kj) a day. The temperature of the body and the heartbeat will start rising up when we perform exercise or some heavy workout. The carbohydrates or carbs are broken down into glucose which is further converted/broken down into energy using O2(oxygen). From research [9], we got the idea to take the variables used here are the timescale the person is training, the average heart rate per minute, and the temperature. Then get more height, weight, gender, and age of the person to predict the tonnage of energy that the person burns. Parameters that can be considered for input are the duration of exercise, average heart rate per minute, temperature, height, weight, and gender. A machine learning XGBoost regressor algorithm is used to predict calories burned depending on exercise time, temperature, height, weight, and

age.
Exercising is a good way to lose more calories if you need to burn a heavy amount of calories. From [6], we can describe that exercise requires energy to be done which is known in kilos of calories. This energy source comes from fat or glycogen. It is really often found in public society that if someone does exercise continuously it will make them get the ideal body weight and body shape. This is to ensure a healthier lifestyle and to reduce health problems. This paradigm has been very united in people who are obese. The rise of obesity as a global epidemic makes it immensely important to monitor the food habits of a modern-day person. Because obese people are apprehensive about losing weight, they regularly need to check the weight lost during exercise. From a bulletin [15], there author says that One of the easiest sports and easy to do to get the ideal weight is jogging. Jogging is a sport that is very effective in burning calories in the body. Jogging is a type of casual running that is meant to be less stressful than running. Usually, this type of exercise is performed remotely and for a period of time. Calories burned while running cause 90% more weight loss than just walking alone. Jogging is a very effective sport so that an athlete must monitor the calorie burning in his body.

1.1 Problem Statement

The problem we are trying to solve is the inaccuracy of traditional methods of calculating calorie burn during exercise. These methods are often based on general formulas and do not take into account individual differences such as age, gender, and weight. This inaccurate calculation can lead to individuals overestimating or underestimating their calorie burn, which can negatively impact their fitness progress and overall health outcomes.
The goal is to create a user-friendly and accurate model by using machine learning that can estimate the number of calories an individual burns based on various input parameters such as activity type, duration of exercise, intensity, and user-specific factors (individual characteristics) like Age, Weight, and Gender.
The system should basically use machine learning techniques to continuously improve its reductions over time, adapting to individual variations and providing personalized insights to help users manage their fitness goals effectively. The primary challenge is to gather and integrate diverse data sources, ensure model accuracy across a range of activities, and create a seamless user

experience for tracking and analyzing calorie expenditure. This application can be helpful for individuals to achieve their fitness goals more efficiently and effectively, which will lead them to improved health outcomes and a better quality of life.

1.2 Motivation

This Calories Burnt Prediction project aims to empower individuals to achieve their fitness goals more effectively and efficiently by providing real-time, personalized insights into their calorie spending during various activities. By using advanced machine learning algorithms [9], the system encourages a healthier lifestyle by helping users make informed decisions about their exercise routines and dietary choices. As it provides a more accurate calorie burn calculation, it can also help individuals better understand their fitness progress. This can ultimately lead to improved health outcomes and a better quality of life.

Based on a web page [11] we can say that this innovative technology works to enhance overall well-being and lifestyle and contributes to a more proactive approach to fitness management. The project's significance lies in its potential to provide individuals and fitness professionals with a more accurate tool for estimating calories burnt during various activities. This personalized approach can contribute to more effective fitness planning and healthier lifestyle choices.

The potential significance of this project lies in its ability to offer individuals a personalized and accurate tool for estimating calorie expenditure during various physical activities.

1.3 Contribution

In this calories burnt prediction project, our primary contribution is to select and prepare the dataset carefully and ensure its suitability for robust model training and evaluation.

The dataset was sourced from a diverse range of individuals engaging in various physical activities, combining/merging data from wearable fitness trackers, physiological sensors, and self-reported inputs. As we did the project while sitting together, my role involved cleaning the data to address missing values, ensuring a complete representation of demographics, and searching for various sources that could help us while working.

To tackle missing values, a careful analysis was conducted to understand the nature and distribution of gaps within the dataset. For numerical features, strategic suggestion techniques were employed, using mean, median, and advanced methods like the k-Nearest Neighbors algorithm. Categorical variables were handled by creating a dedicated category for missing values. This accurate process is aimed at maintaining the properties of the dataset while fixing the impact of missing information on model performance.

Moreover, a crucial aspect of those contributions was standardizing and normalizing numerical features. This step was pivotal in preventing certain features from dominating the model due to disparate scales.

By addressing missing values and standardizing numerical features, the dataset was transformed into a refined and homogeneous input for the calories burnt prediction model. This foundational work lays the groundwork for subsequent phases of the project, enhancing the model's ability to learn and generalize across diverse activities and demographics.

1.4 Outline of This Report

In the Section 1 for Introduction, we are giving some basic knowledge about calories, how calories are measured. The Problem statement for this project, some motivation and contributions. At last, we are showing this Outline of the project subsection that shows the overall Sections definition. Section 2 is about the Literature Review where all the previous work done on this is described. Section 3 is about Dataset. Here all the information about our datasets is given and its source is also given from where we have taken the dataset. Section 4 Represents the main structures and the working procedures of the project. It is mainly working on analyzing and processing the datasets. Section 5 carries all the results that we get after and before processing the datasets. Section 6 is the Conclusion for this project. After the conclusion, all the References used here are given at the end.

2 Literature Review

According to a research source [2], The variety of burned energy in day-today life is directly related to weight maintenance, weight gain, or weight loss. People need to burn more calories than they consume, causing a calorie deficiency. But they want to know how many calories they burn every day.

Most people think that calories are most effectively associated with food and weight loss. As Considered in a study [12], Calories are variously defined units of energy or heat. For men or women trying to gain, lose, or maintain weight, it is essential to know how many calories they are consuming each day. In a study [3], we observed that the global obesity crisis has been continuously increasing, and thus far no nation has been able to turn it around. The World Health Organization identifies an energy imbalance between calories ingested and calories expended as the root cause of obesity. But mounting data indicates that the idea of calorie imbalance might not be enough to control and stop the obesity pandemic.

To examine the calorie imbalance idea and its components as a weight-management tool as well as any potential drawbacks, with the goal of highlighting the need for an updated theory about the origins of obesity. This revision could better direct public health initiatives to control obesity by avoiding weight increase or encouraging weight reduction.

From an article and research [14], Understanding the factors that affect calorie burning might help someone modify their diet or exercise routine to achieve the desired results. Various studies in the literature [1], used machine learning and data mining to diagnose these problems. When compared to today's study, some articles published two to three years ago have a lower accuracy for predicting calorie burn problems. Prior research has explored machine learning applications in health and fitness, but a comprehensive model considering various parameters for calorie burn prediction is less explored. This project builds upon existing studies to create a more holistic and accurate prediction model. A research [11] has shown that machine learning algorithms, particularly regression models and neural networks, have been successfully applied to predict calories burnt based on various input features such as activity type, intensity, duration, and user characteristics. Additionally, [7] study has investigated the integration of physiological parameters like oxygen consumption and metabolic rate to enhance prediction accuracy. The importance of data pre-processing techniques and feature selection is highlighted in the literature, emphasizing the significance of obtaining high-quality input data for reliable predictions. Some studies also discuss the challenges associated with personalized prediction models, considering the variability in individual metabolic rates and responses to different activities. Furthermore, recent advancements in the field of human activity recognition using deep learning techniques, such as convolution neural networks (CNNs) and recurrent neural networks (RNNs), are explored for their potential to improve accuracy in predicting calories burnt during complex and dynamic physical activities. Now here are some further ideas and features for calories burnt prediction machine learning on the basis of previous works and studies

1. **Sensor Integration and Fusion:** Many studies have focused on integrating data from multiple sensors, such as accelerometers and heart rate monitors, to capture a holistic view of physical activities. The fusion of these sensor inputs enhances the model's ability to account for different types and intensities of exercises.

2. **User-Specific Adaptation:** Considering the uniqueness of individuals, certain works have explored methods to adapt prediction models to specific user characteristics. Penalization factors may include age, weight, fitness level, and physiological parameters, allowing for more individualized and accurate predictions.

3. **Incorporation of Environmental Factors:** Some studies have investigated the impact of environmental conditions, such as temperature and humidity, on calorie expenditure. Integrating environmental factors into prediction models acknowledges their influence on the energy cost of physical activities.

4. **Real-Time Feedback Systems:** Research has been conducted on developing real-time feedback systems that provide users with immediate insights into their calorie expenditure during ongoing physical activities. This encourages more informed decision-making and adjustments to achieve fitness and health goals.

5. **Mobile Health Technologies:** The proliferation of mobile health technologies, including smartphone applications and wearables, has inspired research on leveraging these devices for accurate calorie prediction. Incorporating data from everyday devices enhances accessibility and user engagement.

6. **Biometric Data Integration:** Review works that integrate biometric data, such as metabolic rate, body composition, and genetics, to enhance the accuracy of calorie burn predictions. Understand how these factors contribute to a more personalized approach.

3 Dataset Description

This dataset is collected from the platform Kaggle[1]. The name of this dataset is Calories Burnt Prediction as given as Kaggle.

1. **Overview:** The dataset is designed for building a machine learning model to predict the number of calories burnt during physical activities. It includes a diverse set of individuals engaging in various exercises and activities, with corresponding measurements and bio-metric data.

2. **Data Source:** In this project work, we used ”kaggle” to download our dataset, there are two ”.csv” files one named as exercise.csv and the other one is ”calories.csv”, which represents the data, and contains a total of 15000 entries and 10 attributes, the “kaggle” repository's contains attributes information about a variety of people including their age, height, weight, gender, body temperature, heart rate, workout duration. The ”exercise.csv” dataset is taken as the training data. This dataset carries 8 columns containing individual information. The second ”calories.csv” dataset comprises the calories burned by the corresponding person. This dataset carries 2 columns of burnt calorie data and the individual's User_ID.

3. **Features:** These two datasets have various attributes. Each attribute carries a meaning, and how those attributes are being used here, and their units.

Attribute	Function
Gender Individual	Gender (male: 1, female: 0)
Age Individual	Age is mentioned in years
Height Individual	Height of a person, mentioned in centimeter
Weight Individual	Weight of a person, mentioned in kilogram
Heart rate Individual	Average Heart Rate of an Individual during work out (Normal heart rate beat/min)
Body temp Individual	Average Body temperature captured in the course of the entire workout (greater than 37 degrees Celsius)
Duration Individual	Duration of exercising in minutes
Calories Individual	The total amount of calories burned during workout

Table 1: Dataset Attribute and Functions

[1] https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos

4. **Target Variable:** Calories Burnt: The number of calories burnt during a specific activity for any individual.

5. **Data Size:** The dataset comprises a sufficient number of records to ensure diversity and represent different demographics and activities. Ideally, the dataset includes a mix of short and long-duration activities to capture variations in calorie burn patterns.

6. **Data Pre-Processing:**

   (a) Address any missing values in the dataset.

   (b) Standardize or normalize numerical features.

   (c) Encode categorical variables appropriately.

   (d) Handle outliers and anomalies.

7. **Data Split:** Split the dataset into training and testing sets to evaluate the model's performance. By splitting the data, we can see, what number of Data is going for test and train. Our project focused on predicting calories burnt, a crucial step involves the careful splitting of the dataset into training, validation, and testing sets. This process is essential to ensure the robustness and generalization ability of our predictive model. We employed a standard split ratio, allocating a significant portion, typically 70-80%, to the training set. This large proportion allows the model to learn intricate patterns and relationships within the data. The validation set, comprising 10-15% of the dataset, serves as a means to fine-tune hyperparameters and prevent overfitting during the training phase. Finally, the remaining 10-20% is reserved for the testing set, which remains untouched during model development and is only used for evaluating the model's performance on unseen data. This meticulous data split methodology ensures the reliability and effectiveness of our calories burnt prediction model, providing a solid foundation for accurate real-world application.

8. **Potential Challenges:** Working with large datasets can be challenging. It can provide valuable insights and help to uncover hidden patterns in the data, but it can also present a number of challenges. Here are a few potential challenges that may arise when working with large datasets:

(a) **Data cleaning:**
Large datasets often contain a lot of missing, duplicate, or incorrect data. Cleaning and pre-processing the data can be time-consuming and may require a significant amount of computational power. Without cleaning process the prediction system may not work accurately that will not bring the best outcomes we expect.

(b) **Data storage:**
Storing large datasets can be challenging, and it may require a significant amount of storage space and specialized software to manage the data.

(c) **Data analysis:**
Analyzing large datasets can be computationally intensive and may require specialized tools and techniques. It can also be difficult to make sense of the data and identify meaningful insights.

(d) **Data visualization:**
Visualizing large datasets can also be challenging, as it may require specialized software or programming skills. It can also be difficult to present the data in an easy-to-understand and interpretable way.

(e) **Data privacy and security:**
Large datasets often contain sensitive information, and it is important to ensure that the data is protected and used ethically. Companies must ensure that they are compliant with regulations and laws regarding data privacy and protection.

(f) **Scalability:**
As the data grows, the model or algorithm you are using may not be able to handle it, and it might be required to scale it to handle the new size and complexity of the data.

While working with our dataset we did not face many challenges. As this dataset was created with such accurate values. We normally did not have to make many changes to this. Basically while using datasets, there people may face various difficulties. In the data source from where we have taken our dataset. This was a most used dataset and reliable also. In further explanation, we can see the null or duplicate data check result. That will show the accuracy of this dataset.

4 Methodology

This study gathers the right data set and trains the ML model to determine how many calories are burnt by an individual. Firstly, pre-processing of the dataset is carried out to make data free from null values and keep the important attributes in the dataset. After this data is plotted into different graphs to understand the relationship among attributes using different visualization Techniques. We used different regressive and linear machine learning algorithms to compare and find optimal solutions. This study shows the variance among results of different approaches using graphs.

Regression is the technique for investigating the relationship between variables and outcomes. It is used in predictive modeling in which algorithms are used to predict continuous outcomes. [5]

Linear regression

This model consists of a predictor variable and a dependent variable related linearly to each other. It is used to find the dependency between two variables. It calculates the amount of increase in temperature with the amount of exercise done.

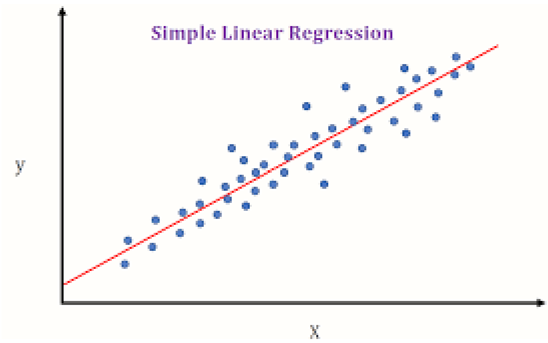


Figure 1: Linear Regression

How does LR-Model Works: Linear Regression is a statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data as Figure 1. The model assumes that the relationship between the variables is linear, meaning that a change in the independent variable(s) corresponds to a proportional change in the dependent variable. The linear regression model aims to find the best-fitting line through the data points, minimizing the sum of squared differences between the observed and predicted values. The coefficients of the linear equation represent the slope and intercept of the line, indicating the strength and direction of the relationship. Once the model is trained, it can be used to make predictions on new data by applying the learned coefficients to the independent variables. Linear Regression is widely employed in various fields for tasks such as predicting outcomes, understanding relationships, and identifying patterns in data.

Benefits: Benefits we can have by using LR-model:

1. It provides a simple yet effective way to understand and quantify relationships between variables.

2. Linear Regression is computationally efficient.

3. Its ease of implementation and straightforward assumptions make it accessible for both beginners and experienced practitioners.

Limitations: There are some limitations for which we did not use this model in our project. Those are:

1. If the relationship is inherently non-linear, a linear regression model might provide inaccurate predictions.

2. Linear regression is sensitive to outliers in the data, which can affect the performance.

3. The variance of the errors is constant across all levels of the independent variable.

Random forest regression

It is a supervised learning algorithm this model combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It is powerful and accurate.

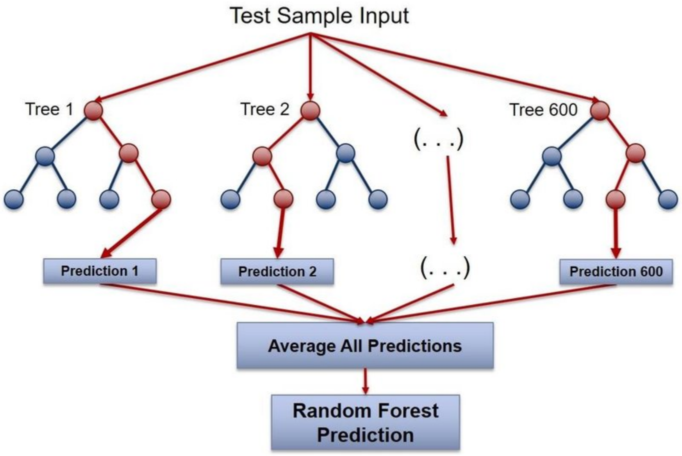


Figure 2: Random Forest Model

How does Random Forest model works: Random Forest Regression is an ensemble learning method that combines the predictive power of multiple decision trees to improve the accuracy and robustness of regression predictions. In this model, a collection of decision trees is constructed during training as Figure 2, each using a subset of the input features and a random subset of the training data. During the prediction phase, the individual tree predictions are averaged to generate the final output. This ensemble approach helps mitigate overfitting and captures complex relationships within the data. The randomness introduced in feature selection and data sampling enhances the model's generalization capabilities. **Benefits:** By using this model, our beneficial points could be:

1. As it uses multiple decision trees, It improves overall prediction accuracy and reduces overfitting.

2. It is resistant to outliers and noise, making it suitable for datasets with diverse and potentially noisy features.

3. This model is particularly effective for handling large datasets with numerous features.

Limitations: The reason or the limitations for which we are not using Random Forest Model, are:

1. It lacks explanation or clarification.

2. The model's complexity can be a disadvantage in scenarios where a transparent and straightforward explanation of predictions is crucial.

3. The potential for overfitting, particularly when the Random Forest is configured with a large number of trees.

4. It might not perform optimally in situations where relationships between variables are non-linear or highly complex.

XGBoost regression

It stands for extreme gradient boosting. This ensemble learning model produces strong predictions by combining the prediction of multiple weak models. It handles large datasets and provides efficient handling of missing values. **How does XGBoost Regression Works:** XGBoost Regression Model operates by sequentially building an ensemble of decision trees, aiming to minimize the prediction error. It employs a gradient boosting framework, where each subsequent tree corrects the errors made by the previous ones. Initially, the model assigns equal weights to all data points and fits a shallow tree to predict the target variable. The subsequent trees focus on the residuals of the previous ones, adjusting their weights based on the errors, enhancing the model's predictive accuracy. XGBoost employs regularization techniques to prevent overfitting, and it incorporates a learning rate to control the contribution of each tree. The model combines the predictions from all trees to produce a final output, creating a robust and accurate regression model known for its efficiency and effectiveness in handling complex datasets.

Benefits: There can be several benefits to using XGBR Model. Those are:

1. It has high predictive performance, as XGBoost is an ensemble learning algorithm that combines the predictions of multiple weak learners (typically decision trees) to produce a strong, accurate model.

2. It effectively handles complex relationships and non-linear patterns in data, making it suitable for a wide range of applications.

3. It has more efficiency and scalability than others, which makes it suitable for large datasets and parallel computing.

4. XGBoost can handle both regression and classification tasks, and its flexibility enables fine-tuning of hyperparameters to optimize performance.
5. This provides feature importance scores, allowing users to understand and interpret the significance of each variable in the model

While working with any application each team needs to follow a well-maintained workflow that is going to be followed by everyone. It is mandatory work to choose a better work plan for achieving a better application.

The workflow of the project looks like as shown in the following Figure 3 below-

Working Flow:

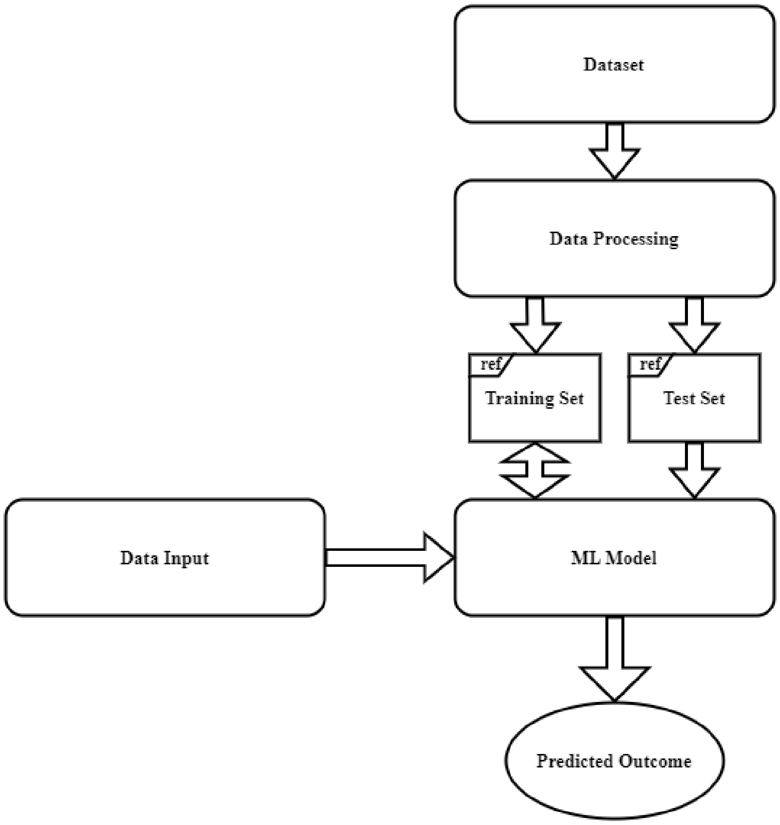


Figure 3: Working Flow for the project

In this innovative AI machine learning project, we support advanced Artificial Intelligence techniques to predict calorie burn accurately and effectively. To make this possible and work as the work flow we need to collect data set, analyse data, pre-process and other steps to follow. Now here we will explain the steps of our work-

4.1 Data Collection

Dataset collection is the primary step in almost every AI project. We also need to take a good and reliable Data set that may help us to build a better application. To get a solid and effective data set, we used Kaggle as the data repository.
We have found two datasets those were used previously for similar kind of work. After downloading both data sets, Data is then uploaded to the colab platform. The data used here is both categorical and numerical.

1. **Calories Dataset:** The first dataset we get from kaggle is calories.csv. Calories Dataset carries 2 columns (User_Id and Calories). Starting 5 rows from calories dataset show at Table 2:

User_ID	**Calories**
14733363	231.0
14861698	66.0
11179863	26.0
16180408	71.0
17771927	35.0

Table 2: Calories.csv Dataset

2. **Exercise Dataset:** The exercise dataset carries some individuals relate information as Age, Height. Weight etc. It has 8 columns and we are showing starting 5 rows from exercise dataset in Table 3

User_ID	**Gender**	**Age**	**Height**	**Weight**	**Duration**	**Heart_Rate**	**Body_Temp**
14733363	male	68	190.0	94.0	29.0	105.0	40.8
14861698	female	20	166.0	60.0	14.0	94.0	40.3
11179863	male	69	179.0	79.0	5.0	88.0	38.7
16180408	female	34	179.0	71.0	13.0	100.0	40.5
17771927	female	27	154.0	58.0	10.0	81.0	39.8

Table 3: Exercise.csv Dataset

4.2 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. The data is then split into test and training data. the used regression models are loaded. test data is used to assess the prediction. In the analysis sector we will also see if there is any null or duplicate data available on those datasets or not.

1. **Check Null and Duplicate:** From the merged dataset, we tried to figure out all the null and duplicate values. We found no null values and checking duplicates also resultant false. From Table 4, we can see that.

Null Values in the Dataset:	
User_ID	0
Calories	0
Gender	0
Age	0
Height	0
Weight	0
Duration	0
Heart_Rate	0
Body_Rate	0
dtype: int64	
check_duplicates(data): False	

Table 4: Checking Null and Duplicates

2. **Histograms:** This is the graphical representation and potted view for the attributes of both Marge Data sets. Here the plots are for the User Id, Calories, Age (M, F) in the Figure 4. Height, Weight and Duration shown in the Figure 5. the last Figure 6 shows the Heart Rate, Body Temperature and Gender of individual.

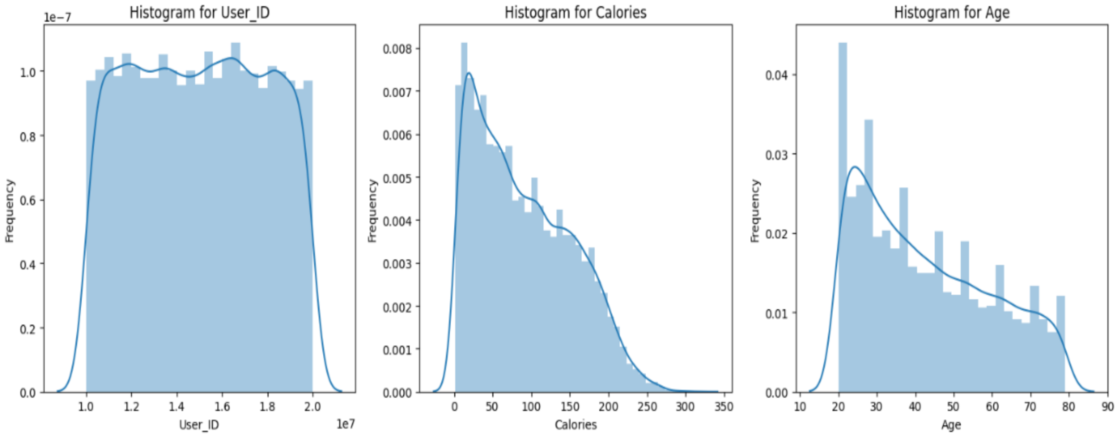


Figure 4: Histogram for User ID, Calories and Age (L to R)

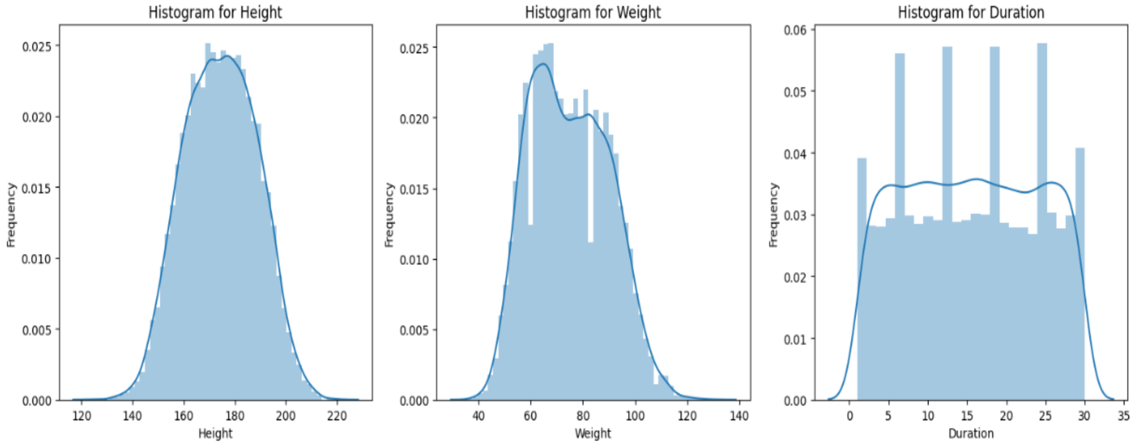


Figure 5: Histogram for Height, Weight and Duration (L to R)

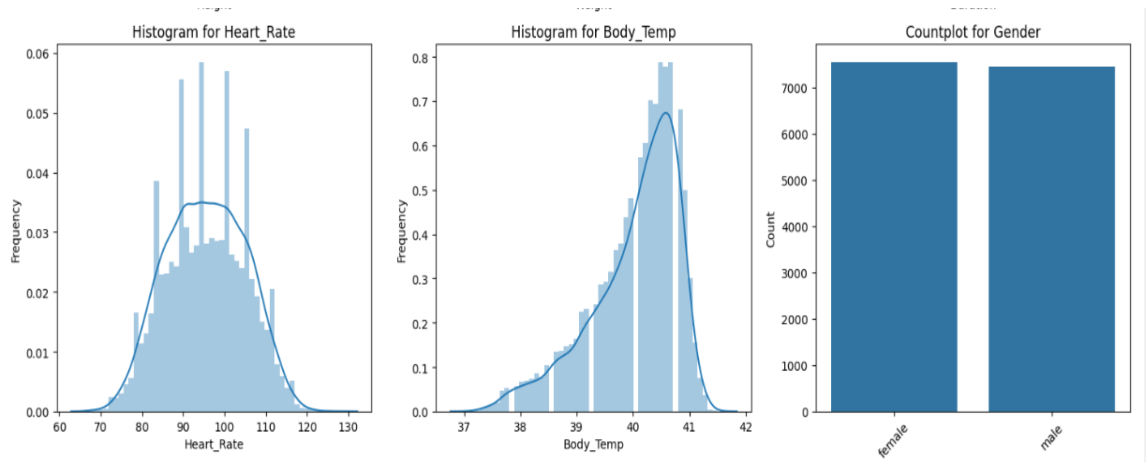


Figure 6: Histogram for Heart Rate, Body Temp and Gender (L to R)

Additionally, graphs can highlight any outliers or anomalies in the data, aiding in the identification and resolution of potential issues within the AI model. The visual depiction of the predicted versus actual calories burnt allows for a direct comparison, enabling a more comprehensive evaluation of the model's accuracy and performance. Moreover, graphical representations enhance the overall professionalism of the project report, making it visually appealing and engaging for readers. In conclusion, the use of graphs significantly improves the overall impact and effectiveness of conveying the findings of a calories burnt prediction AI project. [4]

4.3 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.

1. **Dataset Before Encoding:** 5 rows are shown from the 2 merged data sets. Table 5 before conversion and Table 3 after conversion. In that Table initially, we will merge calories and exercise datasets by the User_ID and create a new data table. here by default it is showing 5 rows for 5 different users calories, gender, age, height, and other individual attributes for each person.

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

Table 5: Before Data Encoding

2. **Dataset After Encoding:** Table 6 is basically carries the same records as Table 5. The difference is, here we replaced the Data sets Gender attribute where we used ’1’ for the user whose Gender is ”male”, and ’0’ for a ”female” user. All the other attributes remains the same.

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

Table 6: After Data Encoding (on Gender column)

3. **Column Drop:** Deleting Unnecessary Columns. Here User ID has been dropped from the dataset (shown in Table 7) as it is no use in here. As we can see, this application is going to predict the calorie from some common or uncommon situation. So, User_ID has no interruption in this application.

	Gender	Age	Height	Weight	Duration	Heart_Rate	Body_Temp
0	1	68	190.0	94.0	29.0	105.0	40.8
1	0	20	166.0	60.0	14.0	94.0	40.3
2	1	69	179.0	79.0	5.0	88.0	38.7
3	0	34	179.0	71.0	13.0	100.0	40.5
4	0	27	154.0	58.0	10.0	81.0	39.8

Table 7: Dropping Column

4. **Pipelining:** Plotting the pipelining for XGBoostRegression model.

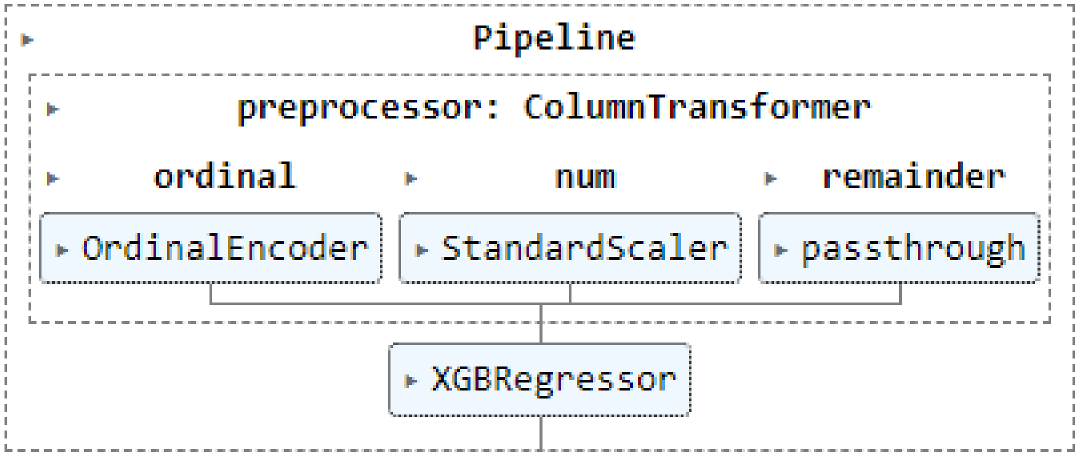


Figure 7: Pipelining for XGboost model

4.4 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 XGBoostRegression which gives the least absolute error value of 1.48 and efficient way to predict calories burnt.

4.5 Evaluation

The results of different algorithms are compared and the best among them is used to calculate the prediction of calories burnt during exercise along with various other factors such as age, gender height, weight, body temperature, and heart rate.

5 Results

1. **Model Accuracy:** Here, to understand which model is better, we used 3 different model and figured out their initial accuracy, error rate

and cross validation or cross check accuracy. and we found that XG boost model was the best one having the highest accuracy and lowest error rate. Here, we are showing the comparison before data pre-processing (Table 8) and after data pre-processing (Table 9).

Before Column Drop:

[[	Model_name,	Model_Accuracy,	Mean_absolute_error,	After_Cross_validation],
[	'LR',	0.9433768585675858,	8.4965348327468746,	0.9433763548523487],
[	'RF',	0.9576765324348346,	1.9578984768759485,	0.9576768393882934],
[	'XGBR',	0.9616823764827639,	1.55386453268753869,	0.9616823422309643]]

Table 8: Three different models outputs(Before pre-process)

After Column Drop:

[[	Model_name,	Model_Accuracy,	Mean_absolute_error,	After_Cross_validation],
[	'LR',	0.9672937151257295,	8.441513553849704,	0.9671402283675841],
[	'RF',	0.9982413634335349,	1.7105533333333334,	0.9978923367747573],
[	'XGBR',	0.9988678909361673,	1.4981198125282924,	0.9988510864545181]]

Table 9: Three different models outputs (After pre-process)

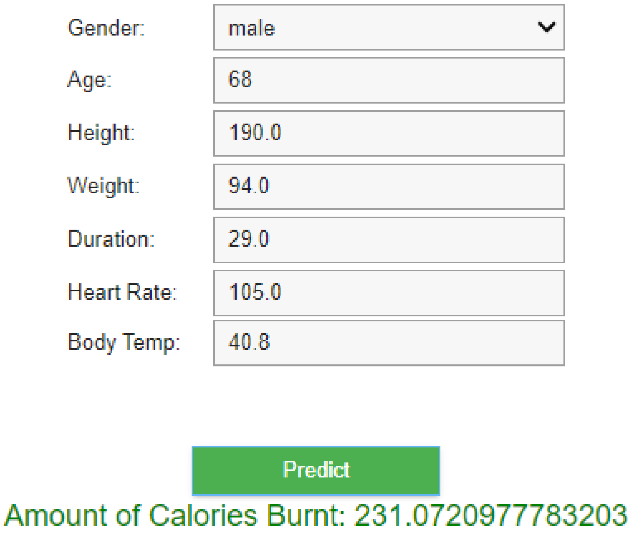
2. **Prediction and Sample:** Given sample Data to train the AI system. This represents the exact value of calorie need to burn. This is going to be a imaginary sample data outside from the data set. This is going to be work as a example for the AI system. Here, for some sample imaginary data, we have found a amount of calorie need to burn. This value is going to be compared with the amount of GUI. The we can finally compare whether the accuracy we gained is true or not.

Sample_Input	**Result**
male, 68, 190.0, 94.0, 29.0, 105.0, 40.8	231.0721

Table 10: Sample and Prediction

3. **Graphical User Interface:** This graphical user interface takes inputs from a user and generates how much calories a person need to

burn by predicting itself. It shows the results from the datasets values from where it has been trained. From the below Figure 8 we can match the calories from sample prediction and justify its prediction.



Gender: male

Age: 68

Height: 190.0

Weight: 94.0

Duration: 29.0

Heart Rate: 105.0

Body Temp: 40.8

Amount of Calories Burnt: 231.0720977783203

Figure 8: Graphical User Interface

Now comparing the predicted result with the Expected result. To compare this we are taking a sample data as prediction. For XGBoostRegression model:

XGBoostRegression		
Input Data	Predicted Result	Expected Result
male, 68, 190.0, 94.0, 29.0, 105.0, 40.8	231.0720977783203	231.0721

Table 11: Comparing Predicted and Actual Result

Now we can say from the comparison shown in Table 11, the user interface is showing almost the exact and accurate calories value as we expected.

6 Conclusion:

We deduced from the analysis that the XGBRegressor produces more accurate findings. Mean absolute error suggests that absolute error should be as minimal as possible. It is nothing more than the discrepancy between values that were seen and those that were predicted by models. 1.49 is a good value for the mean absolute value that the XGBRegressor gives us. The mistake rates are quite low. Therefore, we can say that XGBoostRegressor is the best model for predicting calorie burn. The flexibility of the suggested technique can also be improved with variations. In this study, we have concentrated on the seven primary factors that influence how many calories our body burns, but there are other factors that also play a role. It's also crucial to understand how many calories we are consuming if we want to stay healthy and fit. Additionally, ML may be used to construct this (machine learning). A UI (user interface) is also required so that users may input their values and obtain results that show how many calories they have burned. Additionally, we are able to create a completely functional app with all of these features and our recommended diet and exercise regimen.

References

[1] Tom Baranowski, Karen W Cullen, Theresa Nicklas, Deborah Thompson, and Janice Baranowski. Are current health behavioral change models helpful in guiding prevention of weight gain efforts? *Obesity research*, 11(S10):23S–43S, 2003.

[2] Salvador Camacho and Andreas Ruppel. Is the calorie concept a real solution to the obesity epidemic?, 2017.

[3] Salvador Camacho and Andreas Ruppel. Is the calorie concept a real solution to the obesity epidemic? *Global health action*, 10(1):1289650, 2017.

[4] Viorica Rozina Chifu, Cristina Bianca Pop, Andrei Ciurianu, Emil St Chifu, and Marcel Antal. Machine learning-based approach for predicting health information using smartwatch data. In *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 391–397. IEEE, 2021.

[5] Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

[6] David W Hart, Steven E Wolf, David N Herndon, David L Chinkes, Sophia O Lal, Michael K Obeng, Robert B Beauford, and Ronald P Mlcak. Energy expenditure and caloric balance after burn: increased feeding leads to fat rather than lean mass accretion. *Annals of surgery*, 235(1):152–161, 2002.

[7] Amol Kadam, Anurag Shrivastava, Sonali K Pawar, Vinod H Patil, Jacob Michaelson, and Ashish Singh. Calories burned prediction using machine learning. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, volume 6, pages 1712–1717. IEEE, 2023.

[8] Peggy J Liu, James R Bettman, Arianna R Uhalde, and Peter A Ubel. 'how many calories are in my burrito?' improving consumers' understanding of energy (calorie) range information. *Public Health Nutrition*, 18(1):15–24, 2015.

[9] Ishaq Azhar Mohammed. Artificial intelligence for caregivers of persons with alzheimer's disease and related dementias: Systematic literature review. *International Journal of Emerging Technologies and Innovative Research (www. jetir. org— UGC and issn Approved), ISSN*, 2349:5162, 2019.

[10] Marion Nestle and Malden Nesheim. *Why calories count: from science to politics*, volume 33. Univ of California Press, 2012.

[11] Steven Ovadia. Researchgate and academia. edu: Academic social networks. *Behavioral & social sciences librarian*, 33(3):165–169, 2014.

[12] Punita Panwar, Kanika Bhutani, Rohit Saini, et al. A study on calories burnt prediction using machine learning. In *ITM Web of Conferences*, volume 54, page 01010. EDP Sciences, 2023.

[13] Gurrappagaru Sanjana Reddy and Singareddy Velankini Suhas. Accurate calorie burn prediction with machine learning: Xgboost in focus.

[14] N Shah, F Amirabdollahian, and R Costa. The dietary and physical activity habits of university students on health and non-health related courses. *Journal of Human Nutrition and Dietetics*, 24(3):303–304, 2011.

[15] Nur Zarna Elya Zakariya and Marshima Mohd Rosli. Physical activity prediction using fitness data: Challenges and issues. *Bulletin of Electrical Engineering and Informatics*, 10(1):419–426, 2021.