Calorie Prediction using AI

This project leverages Multivariate Predictive AI Model to accurately predict calorie burn during physical activity, addressing the urgent global challenge of obesity and overweight. By enabling personalised insights, it empowers individuals, especially in high-risk regions to make informed health and fitness decisions.

Date : Jun 21, 2025

USD Project Team : 8

Members : Pankaj Shukla, Ashis Das and Bashir Ali Qule

Table of Contents

Introduction

The alarming global rise of obesity and overweight populations serves as the fundamental motivation and backdrop for the research and development of accurate calorie estimation methods during physical activity.

The **growing number of obese and overweight individuals** is a significant concern. This problem is widespread and, as noted, **no single nation has yet been able to fully resolve it**. Malaysia is specifically identified as having **the highest rate of obesity and overweight among Asian countries**, with 64% of men and 65% of women categorised as fat or overweight, and cases are **rising at an alarming rate** nationwide.

This increase in obesity and overweight is attributed to several modern lifestyle factors:

- **Sedentary lifestyles, long workdays, and a general lack of physical activity** are directly linked to increased rates of obesity, heart disease, and other chronic health issues.
- While genetics play a part, **environmental and lifestyle factors, such as physical activity and eating habits, are considered vitally important**.
- A key underlying cause is an **energy imbalance between calories consumed and calories expended**. People, often due to a lack of time, tend to **intake more junk food than healthy options**, which directly increases their total calorie rate and contributes to obesity. Furthermore, less active individuals may experience **muscle deterioration and a slowed metabolism**, making it more difficult to maintain a healthy Body Mass Index (BMI).

The consequences of this rising prevalence are severe:

- These conditions have a **substantial negative effect on general health and quality of life**.
- They also **raise the cost of healthcare**. The escalating health concerns stemming from inactive lifestyles underscore the urgent need for effective interventions.

In this larger context of growing obesity and overweight, there is **crucial need for accurate calorie estimation methods for**:

- **Promoting Healthy Lifestyles:** Regular exercise is key to managing obesity, balancing calorie intake, and sustaining a healthy lifestyle.
- **Empowering Informed Decisions:** Knowing your calorie burn helps you make smarter fitness and nutrition choices for better health outcomes.
- **Addressing Barriers to Exercise:** Personalized calorie burn tools can motivate and guide obese individuals through their fitness journey.

Current Estimation Methods for Calorie Burn

Method	Overview	Strengths	Limitations
Metabolic Equations	Estimate oxygen uptake (VO_2) and energy expenditure using standard equations (e.g., 1 MET = 3.5 ml·kg^{-1}·min^{-1}).	Standardized, used in clinical settings, endorsed by ACSM.	Inaccurate at high intensities, assumes "one size fits all," based on fixed rates.
Heart Rate (HR) Based	Uses HR to estimate VO_2 and calorie burn.	Easy to measure, reflects intensity changes.	Requires calibration, less accurate at low intensity, affected by stress.
Conventional HR	Estimates based on steady HR levels.	Useful in steady-state exercise.	15–35% error, assumes constant effort.
Beat-by-Beat HR (e.g., Firstbeat)	Analyzes detailed HR data with metabolic	High accuracy (7–10%), no lab needed,	Advanced method, may require specific

	modeling.	works under all conditions.	devices.
Wearables & Motion Sensors	Devices use motion (accelerometry/pedometers) to estimate energy use.	Low cost, user-friendly, tracks daily movement.	High error (20–60%), poor for non-vertical activities, not metabolism-based.
Calorimetry & DLW	Gold standard methods: Direct/Indirect calorimetry and Doubly Labeled Water (DLW).	Most accurate (≤5% error), reliable for both short and long term.	Expensive, not practical for daily use.
Activity Diaries & Questionnaires	Self-reported physical activity logs.	Inexpensive, promotes self-awareness.	High error (20–60%), prone to bias, not real-time.
Machine Learning Models	Predict calories using algorithms trained on multiple features.	High precision, personalized, scalable.	Requires quality data, still developing in real-time applications.

Project Overview

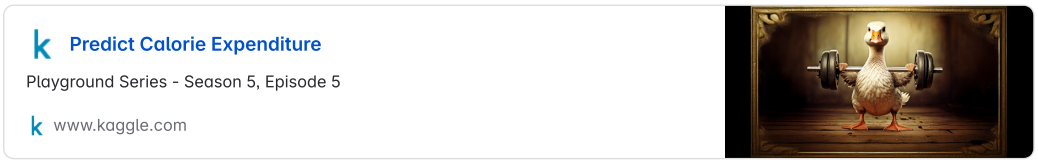
This project aims to develop an accurate and personalised model for estimating caloric expenditure during physical activity by integrating a range of physiological and demographic variables. Traditional methods often rely on generalised estimations that overlook individual variability. By adopting a data-driven approach grounded in physiological science, this project seeks to model energy expenditure with higher precision, contributing to advancements in personalised health monitoring, fitness tracking, and metabolic research.

Model Objectives

1. To design and implement a predictive model that leverages features such as age, gender, height, weight, heart rate, body temperature, and duration of activity to estimate individual calorie burn with high accuracy
2. To analyse and quantify the relative impact of each input variable on energy expenditure, identifying key predictors and exploring their interdependencies through statistical and machine learning techniques.

Data

The data was sourced from Kaggle Competition. The dataset for this competition (both train and test) was generated from a deep learning model trained on the Calories Burnt Prediction dataset



Dataset Details:

No. of Observation - 7,50,000, No. of Fields - 8
Target Feature (1) - Calories
Explanatory Features (7) - Sex, Age, Height, Weight, Duration, Heart_Rate, Body_Temp

Data collection method is unknown, but it seems to have come from a healthcare clinic collected at least 7 years ago. It is fair to assume that this would have been the clinic must have employed the most accurate way to measure calorie scientifically as per the technology available.

Exploratory Analysis

a) Variables

The 8 features or variables can be categorised as follows:

Numerical Variable	Categorical Variable
Continuous - Height, Weight, Duration, Heart_Rate, Body_Temp, Calories Discrete Variable - Age	Sex

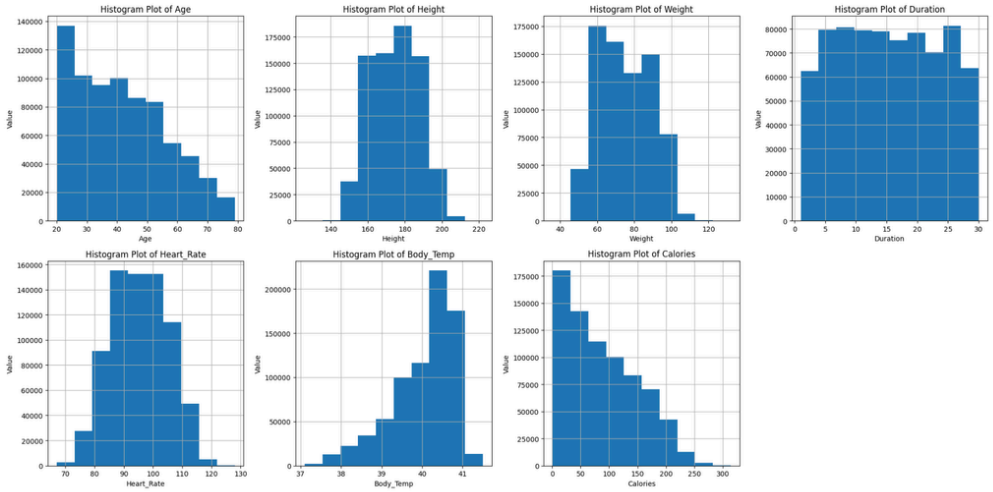
There are no null data in this dataset. Hence, we can go ahead with exploring these variables at length with an objective to find out key insights along the way.

b) Distribution Analysis

Data Description for Continuous Variables [using Pandas library]

	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
count	750000.000000	750000.000000	750000.000000	750000.000000	750000.000000	750000.000000	750000.000000
mean	41.420404	174.697685	75.145668	15.421015	95.483995	40.036253	88.282781
std	15.175049	12.824496	13.982704	8.354095	9.449845	0.779875	62.395349
min	20.000000	126.000000	36.000000	1.000000	67.000000	37.100000	1.000000
25%	28.000000	164.000000	63.000000	8.000000	88.000000	39.600000	34.000000
50%	40.000000	174.000000	74.000000	15.000000	95.000000	40.300000	77.000000
75%	52.000000	185.000000	87.000000	23.000000	103.000000	40.700000	136.000000
max	79.000000	222.000000	132.000000	30.000000	128.000000	41.500000	314.000000

Histogram for Continuous Variables [Using matplotlib library]



c) Skewness Analysis

Feature	Skewness	Comment
Age	0.43	Right Skewed
Height	0.05	No/Minimal Skewness
Weight	0.21	Slight Right Skewed
Duration	0.02	No/Minimal Skewness
Heart Rate	-0.005	No/Minimal Skewness
Body Temperature	-1.02	Left Skewed
Calories	0.53	Right Skewed

Skewness Interpretation:

Skewness ≈ 0: The data is approximately symmetric. A skewness close to 0 indicates a normal (Gaussian) distribution. In our dataset, we see this for **Height, Duration and Heart Rate.**

Skewness > 0: The distribution is positively skewed, meaning the right tail is longer than the left tail. The data tends to have relatively few larger values. In our dataset, we see this for **Age, Weight and Calories.**

Skewness < 0: The distribution is negatively skewed, meaning the left tail is longer than the right tail. The data tends to have relatively few smaller values. In our dataset, we see this for **Body Temperature.**

d) Kurtosis Analysis

Feature	Kurtosis	Comment
Age	-0.74	Platykurtic
Height	-0.83	Platykurtic
Weight	-0.99	Platykurtic
Duration	-1.19	Platykurtic
Heart Rate	-0.67	Platykurtic
Body Temperature	0.51	**Leptokurtic - Heavy Tails**
Calories	-0.68	Platykurtic

Kurtosis Interpretation:

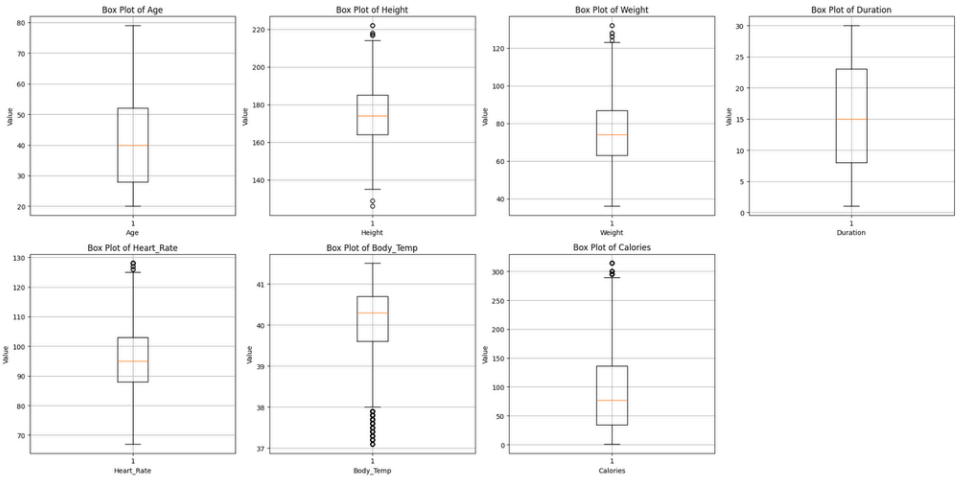
Excess Kurtosis < 0 → **Platykurtic** (light tails)

Excess Kurtosis ≈ 0 → **Mesokurtic** (normal-like)

Excess Kurtosis > 0 → **Leptokurtic** (heavy tails)

Insight 1 : Body Temperature Variable has high kurtosis, suggesting high number of outliers from the statistical test.

e) Outlier Analysis



	Column	IQR	Lower Bound	Upper Bound	Below Lower Bound	Above Upper Bound	z_scores
0	Age	24.0	-8.00	88.00	0	0	0
1	Height	21.0	132.50	216.50	2	12	38
2	Weight	24.0	27.00	123.00	0	9	104
3	Duration	15.0	-14.50	45.50	0	0	0
4	Heart_Rate	15.0	65.50	125.50	0	36	175
5	Body_Temp	1.1	37.95	42.35	14919	0	3368
6	Calories	102.0	-119.00	289.00	0	139	312

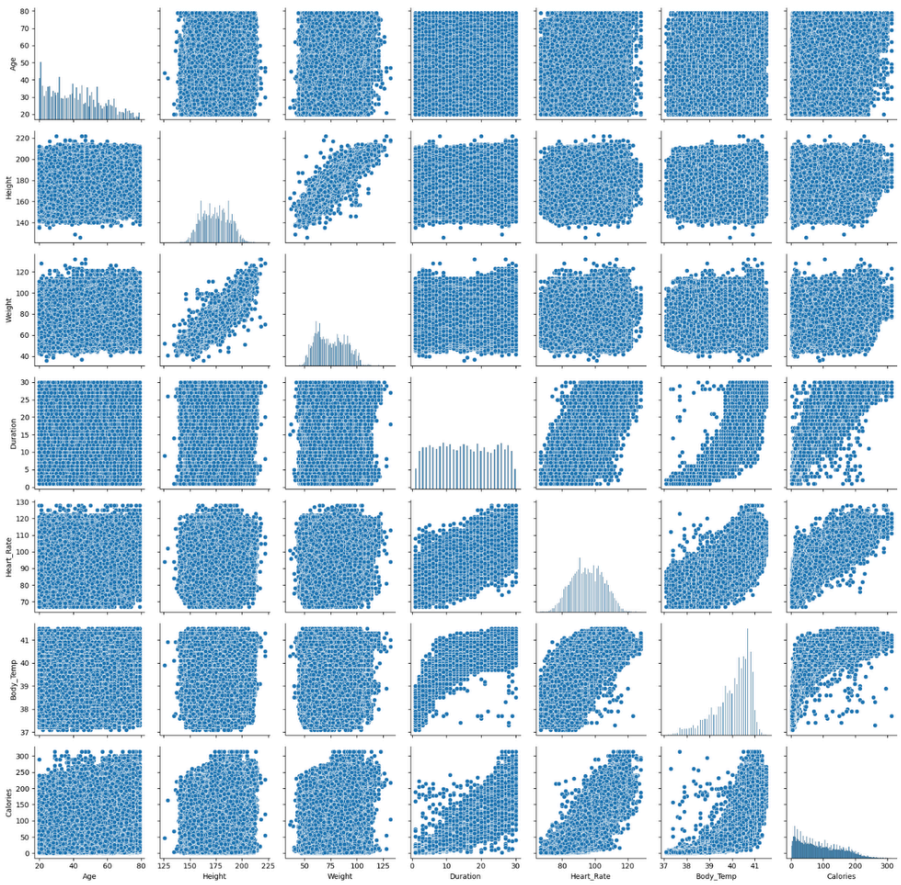
Both from visual analysis from box plots and from IQR stats, we see ~14,919 values in Body_Temp are lower than the lower bound of 1.5*IQR (Inter Quartile Range). But they all are greater than 35 degrees.

Could it be just a false positive?

On further investigation, we see that the body_temp distribution is right skewed and has a low standard deviation is **0.78**. Also, from common knowledge we know that the temperature between 35-37.95 degrees is actually not an outlier.

Hence, we are not removing outliers from the body temperature variable.

f) Pair-plot Analysis

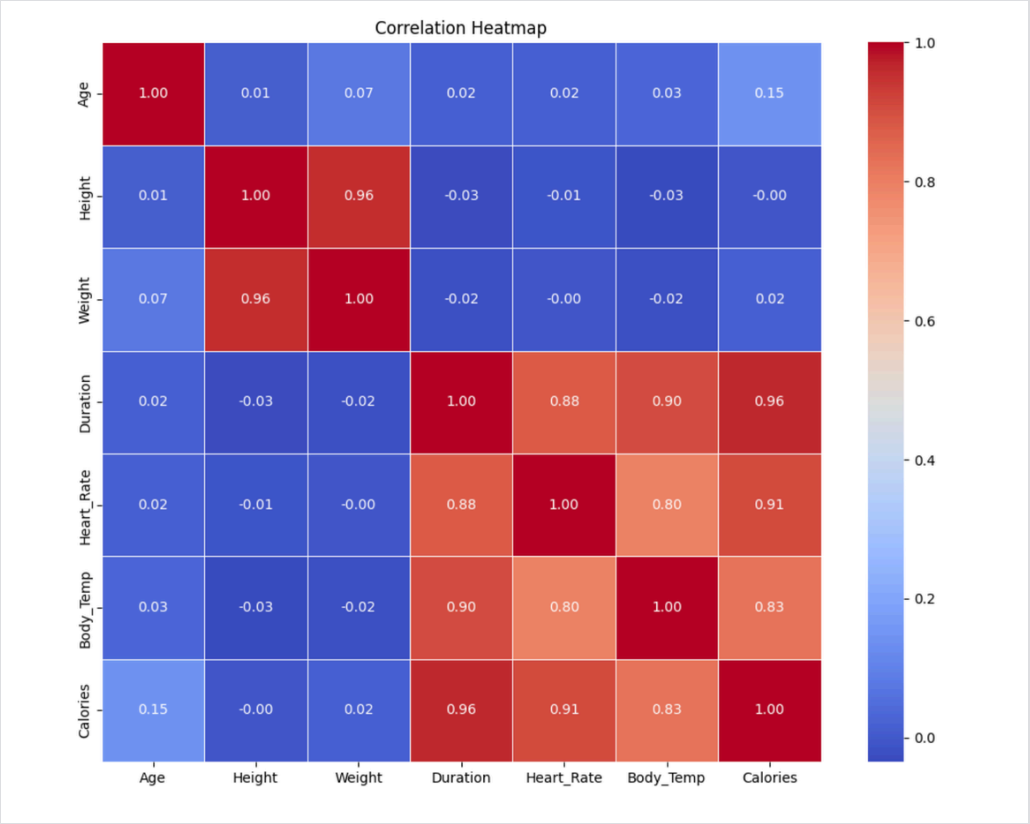


The pair-plot analysis of target (Calories) and explanatory variables show the below observations: Calorie has **positive linear relationship with duration, heart rate and body_temperature.** Calorie doesn't have seem to have any sort of relationship with Age, Height and Weight variables.

Insight 2:

Calorie has **positive linear relationship with duration, heart rate and body_temperature.** Height and Weight have a strong linear relationship which is expected from common knowledge.

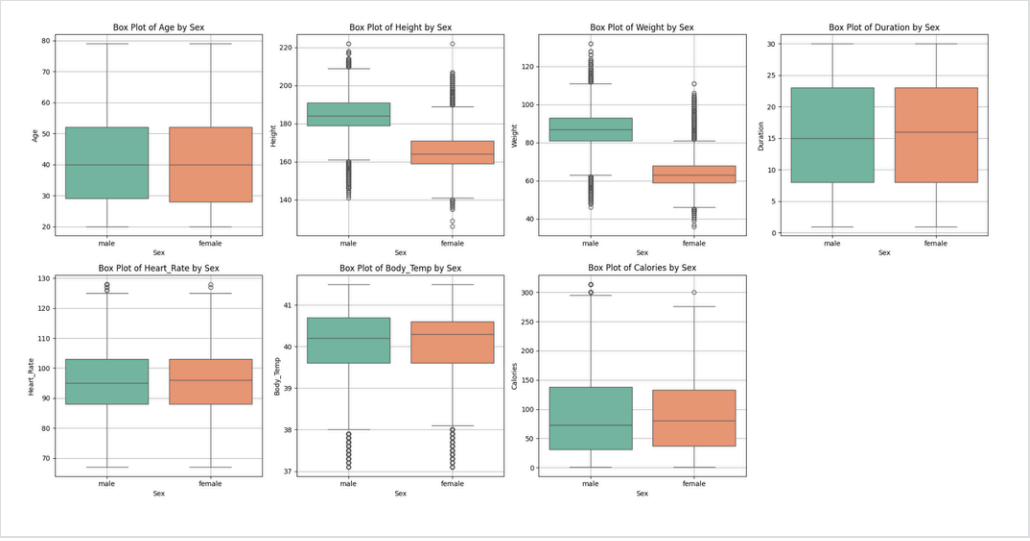
g) Correlation Analysis



Insight 3:
High correlation observed between Height and Weight.

Insight 4:
High correlation observed between Duration, Heart_Rate and Body Temperature. In feature selection for model building, need to be mindful of selecting all of these three features together. As doing so means distributing the impact of one overall between the three, and none of them may turn out significant then.

h) Outlier Analysis segmented by Sex Information



From the above box plots, we don't see significant change in any of the features when segmented by gender apart from **height and weight variables**, where men's height and weight is higher than that of females. This is also expected from common knowledge.

Insight 5:
Homogenous or balance distribution of Sex feature in our dataset.

Both Male and Female Gender have **similar mean and dispersion** on Age, Heart_Rate, Body_Temp and Calories Burnt

Inferential Analysis

a) Shapiro-Wilk Normality Test

	Variable	Test Statistic	p-value	Normal Distribution?
0	Age	0.9538	0.0	No (p≤0.05)
1	Height	0.9843	0.0	No (p≤0.05)
2	Weight	0.9678	0.0	No (p≤0.05)
3	Duration	0.9544	0.0	No (p≤0.05)
4	Heart_Rate	0.9907	0.0	No (p≤0.05)
5	Body_Temp	0.9137	0.0	No (p≤0.05)
6	Calories	0.9432	0.0	No (p≤0.05)

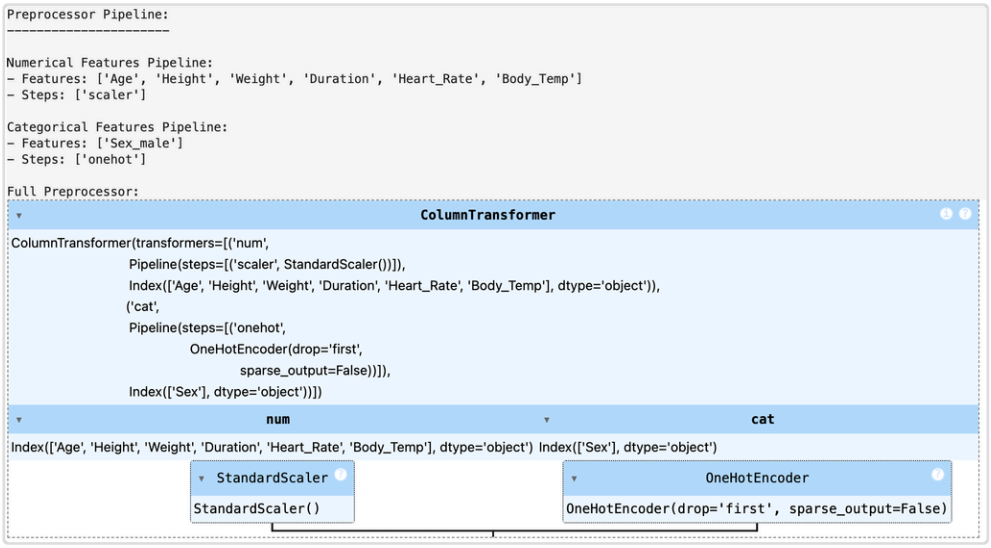
```
The Shapiro-Wilk test checks if data follows a normal distribution:
- Null hypothesis (H0): Data is normally distributed
- Alternative hypothesis (H1): Data is not normally distributed
- If p-value > 0.05: Data likely follows normal distribution
- If p-value ≤ 0.05: Data likely does not follow normal distribution
```

b) Mann-Whitney U test

	Variable	Male Median	Female Median	Difference	p-value	Significant Difference?
0	Age	40.0	40.0	0.0	0.0	Yes (p≤0.05)
1	Height	184.0	164.0	20.0	0.0	Yes (p≤0.05)
2	Weight	87.0	63.0	24.0	0.0	Yes (p≤0.05)
3	Duration	15.0	16.0	-1.0	0.0	Yes (p≤0.05)
4	Heart_Rate	95.0	96.0	-1.0	0.0	Yes (p≤0.05)
5	Body_Temp	40.2	40.3	-0.1	0.0	Yes (p≤0.05)
6	Calories	73.0	80.0	-7.0	0.0	Yes (p≤0.05)

```
Mann-Whitney U test compares distributions between male and female groups:
    - Null hypothesis (H0): No significant difference between groups
    - Alternative hypothesis (H1): Significant difference exists between groups
    - If p-value > 0.05: No significant difference
    - If p-value ≤ 0.05: Significant difference exists
```

Pre-Processing and Feature Engineering



- The data has been split into into features (X) and target variable (Calories)
- Creating preprocessing pipeline for numeric and categorical features:
- Numeric features: ['Age', 'Height', 'Weight', 'Duration', 'Heart_Rate', 'Body_Temp']
  Categorical features: ['Sex']
- Numeric features will be standardized using StandardScaler
- Categorical features will be one-hot encoded, dropping first category
- Top 5 most important features using f_regression

Best Features : Weight, Duration, Heart Rate and Sex

Feature	**Weight**	**Duration**	**Heart Rate**	**Sex**	Height	Body Temp	Age
Score	8.7e+06	3.54e+06	1.64e+06	16262	188	108	12
Rank	1	2	3	4	5	6	7

Model development

The Process involves configuring 5 machine learning models—Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost, while also setting up basic experiment settings such as test size (20%) with input features **X**, target **y**.

Multiple results were generated by experimenting with different combinations of features and model hyper parameters.

For each model, specific hyper-parameters are defined:

Linear Regression	Decision Tree	Random Forest	Gradient Boosting	XGBoost
fit intercept	max depth and minimum samples split	number of trees and max depth	number of trees and learning rate	number of trees, learning rate, and max depth

```
Model Building
---
Test Set Size: 0.2
Number of Batches: 10
Selected Batch: 1

Model Parameters:

Linear Regression Parameters:
Fit Intercept: True

Decision Tree Parameters (Two Models):
Max Depth: 3 and 5
Min Samples Split: 2

Random Forest Parameters (Two Models):
Number of Trees: 100
Max Depth: 3 and 5

Gradient Boosting Parameters (Two Models):
Number of Trees: 100
Learning Rate: 0.1 and 0.01

XGBoost Parameters (Two Models):
Number of Trees: 100
Learning Rate: 0.1 and 0.01
Max Depth: 5
```

Six set of Train Test data was created for Model Training with below features:

1. `Features0 (all): ['Sex', 'Age', 'Height', 'Weight', 'Duration', 'Heart_Rate', 'Body_Temp']`
2. `Features1 (subset): ['Sex', 'Age', 'Height', 'Duration', 'Heart_Rate', 'Body_Temp']`
3. `Features2 (subset): ['Sex', 'Age', 'Height', 'Weight', 'Duration', 'Body_Temp']`
4. `Features3 (subset): ['Sex', 'Age', 'Height', 'Duration', 'Body_Temp']`
5. `Best 5 Selected Features: ['Sex', 'Height', 'Weight', 'Duration', 'Heart_Rate']`
6. `Best 4 Selected Features: ['Sex', 'Weight', 'Duration', 'Heart_Rate']`

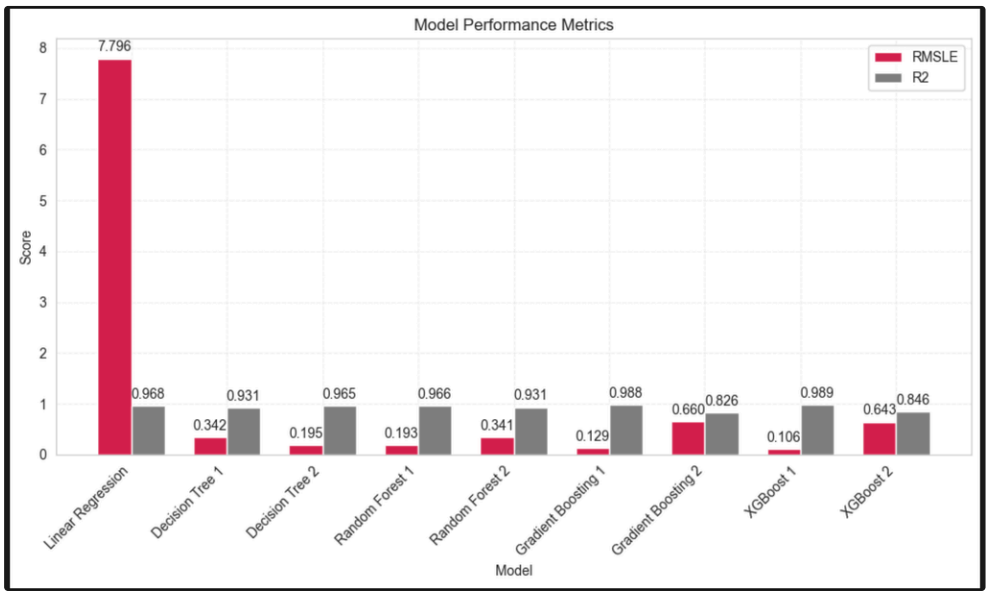
Evaluation

This evaluation method compares the performance of the given machine learning models—like Linear Regression, Decision Tree, Random Forest, and others—by training them on a selected portion of the dataset and testing them on 20% unseen data.

It uses two key measures: **RMSLE (Root Mean Squared Log Error)**, which checks how close the predicted values are to the actual ones while handling large value differences more gently, and **R^2 (R-squared)**, which tells how well the model explains the variability in the data.

Among all the models, the one that achieves the lowest RMSLE (indicating better prediction accuracy) and a higher R^2 is chosen as the best. The function prints performance scores, tracks configurations used, and saves the best-performing model for future use.

Prediction Results

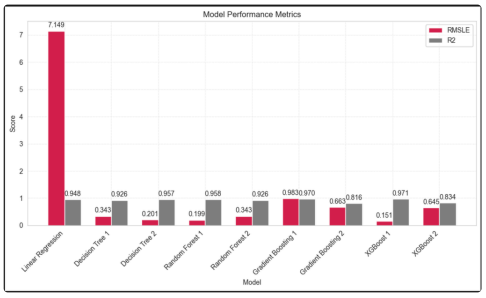


Best Performing Model:
Model: XG Boost
RMSLE Score: 0.106
R^2 Score: 0.989

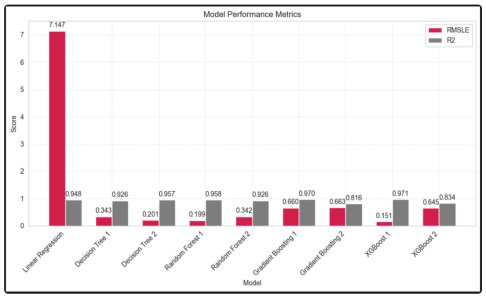
Configuration:
{'n_estimators': 100, 'max_depth': 5, 'LR' : 0.1}

Features:
'Sex', 'Weight', 'Duration', 'Heart_Rate'

Results from Iteration :



XG Boost with LR=0.1
Feature : `'Sex', 'Age', 'Height', 'Duration', 'Body_Temp'`

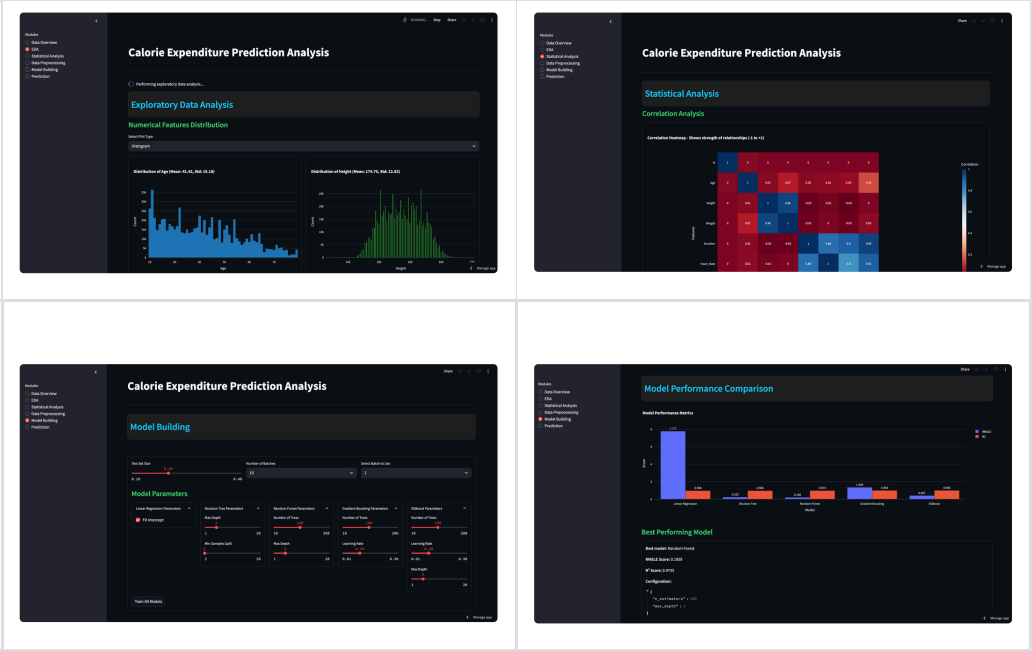


XG Boost with LR=0.1
Feature : `'Sex', 'Age', 'Height', 'Duration', 'Heart_Rate', 'Body_Temp'`

End to End Model Training Web Application

The solution was also transformed into a automated framework using streamlit based web application, which involves exploration of new data, building new calorie prediction models from a long-term solution point of view.

Access it from here : Calorie Expenditure Prediction



References

1. Hendelman, D., Miller, K., Baggett, C., Debold, E., & Freedson, P. (2000). *Validity of accelerometry for the assessment of moderate intensity physical activity in the field. Medicine & Science in Sports & Exercise, 32*(9 Suppl), S442–S449. NCBI - WWW Error Blocked Diagnostic
2. Lyden, K., Kozey, S. L., Staudenmayer, J. W., & Freedson, P. S. (2014). *A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. European Journal of Applied Physiology, 114*(2), 271–280. https://iopscience.iop.org/article/10.1088/0967-3334/35/2/253
3. Godfrey, A., Conway, R., Meagher, D., & ÓLaighin, G. (2008). *Direct measurement of human movement by accelerometry. Medical Engineering & Physics, 30*(10), 1364–1386. A Theoretical Method of Using Heart Rate to Estimate Energy Expenditure during Exercise - Robert W. Pettitt, Cherie D. Pettitt, Chad A. Cabrera, Steven R. Murray, 2007
4. Keytel, L. R., Goedecke, J. H., Noakes, T. D., Hiiloskorpi, H., Laukkanen, R., van der Merwe, L., & Lambert, E. V. (2005). *Prediction of energy expenditure from heart rate monitoring during submaximal exercise. Journal of Sports Sciences, 23*(3), 289–297. NCBI - WWW Error Blocked Diagnostic
5. Ceesay, S. M., Prentice, A. M., Day, K. C., Murgatroyd, P. R., Goldberg, G. R., Scott, W., & Spurr, G. B. (1989). *The use of heart rate monitoring in the estimation of energy expenditure: A validation study using indirect whole-body calorimetry. The American Journal of Clinical Nutrition, 50*(3), 591–600. NCBI - WWW Error Blocked Diagnostic
6. Data Collection : Discussion of Kaggle and video link shared on that.

   https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos/discussion/476313

   MSXDAT262017-V012200

   This dataset was used by Graeme Malcolm from Microsoft in a tutorial video

Github Link

GitHub - usd-aai-500-in1-group8/predict-calorie-expenditure: Multivariate Predictive Model for Estimating Calorie Burn Based on Physiological Signals