Exploratory Data Analysis of Fitness Tracker Data and building a Calorie Burnt and Workout Prediction Model using Machine Learning.

Pratyasha Das, Tapaja DasRoy,
B.Sc. Physics Hons.
Seth Anandram Jaipuria College

Mentor
Sabyasachi Ghosh

Period of Internship: 19th May2025 - 15th July 2025

Report submitted to: IDEAS - Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

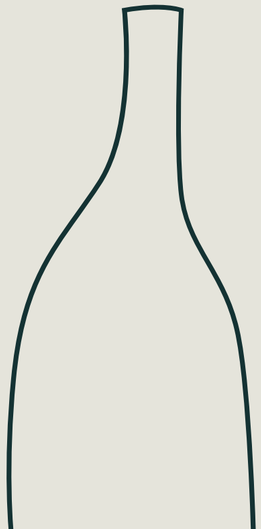
Table of Contents

Project Relevance

- Explores whether heart rate is directly correlated with calories burned
- Evaluates the effectiveness of a custom Intensity Index in predicting calorie expenditure
- Traditional calorie calculators often ignore real-time effort indicators like heart rate
- Aims to build more personalized and accurate predictions using physiological and workout data
- Supports development of self-use fitness tools and informed health decisions
- Helps bridge the gap between data science and practical health applications

Background Survey & Literature Review

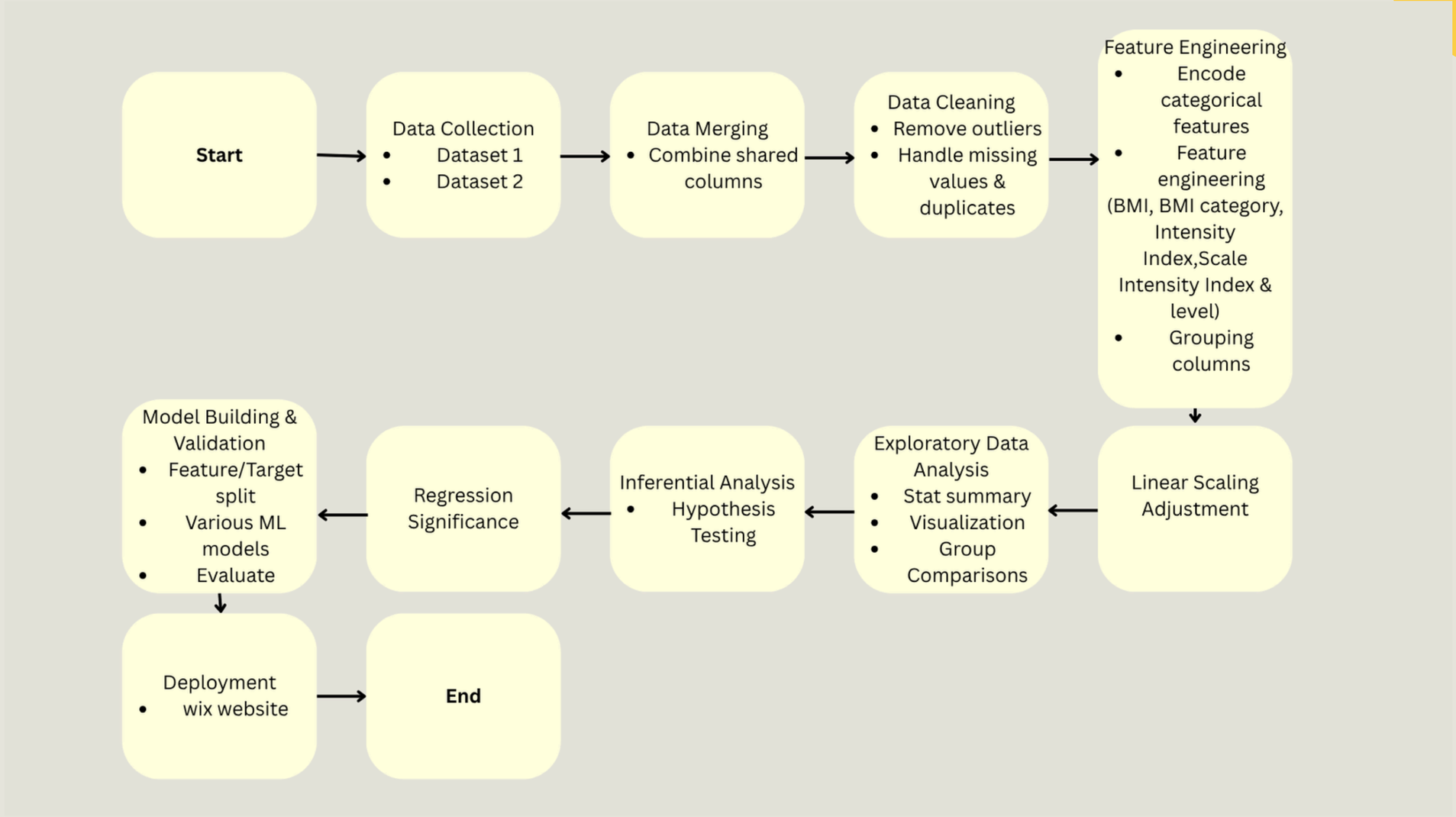
- Metabolism & BMR – Energy used at rest for vital functions
- Muscle Mass – More muscle = higher calorie burn
- Height & Weight – Larger bodies burn more calories
- Age – Metabolism slows with age due to muscle loss
- Sex Differences – Men usually burn more than women due to body composition difference.
- Exercise Intensity – Higher effort = more calories burned
- Heart Rate – Strong indicator of calorie burn during activity
- BMI Limitations – Not a reliable calorie burn predictor by itself



Project Objective

- **Calorie Prediction Model**
  - Inputs: Gender, Age, Height, Weight, Duration, Heart Rate
  - Goal: Predict calories burned using ML
- **To determine whether Heart Rate is a significant and direct predictor of Calories Burnt during exercise.**
- **Exploratory Data Analysis (EDA)**
  - Find trends, key factors, and outliers
  - Visualize insights to guide model development
- **Intensity Index Feature**
  - Create a custom metric using heart rate, weight, and duration
  - Ensure correlation with calories burned
- **Wix Website Deployment**
  - User inputs data → get predicted calories
  - Includes BMI & Ideal Weight calculators
  - Goal: Real-world application & demo-ready tool

Methodology



Final Adjustments Overview:

The final dataset for EDA and Model training was a merged dataset with the following features:
Merging:

- Dataset 1 - 9000
- Dataset 2 - 6608 (others were not considered)

Merged dataset count(before outlier removal): 15608
Merged dataset count (After outlier removal): 13782

Intensity Index

$$Intensity\ Index = \frac{Heart\ rate * Duration * \sqrt{(Weight)}}{1000}$$

Purpose:

- Measures how intense a workout is
- Gives one score to compare effort across users

Why these variables?

- Heart Rate: Shows physical effort
- Duration: Longer workouts = more effort
- Weight: Heavier people burn more (√Weight avoids overweight bias)

Scaling:

- Divided by 1000 to keep values between 0–10

Exclusions (BMI and Gender):

- BMI: Already reflected via weight
- Gender: Heart rate reflects effort across all genders.

Correlation with Calories Burned: +0.52

Use Cases:

- Classify workouts (Very Low to Very High)

Scaled Index	Intensity	Workout Examples
0.0 – 2.0	Very Low	Stretching, casual walking, seated yoga
2.0 – 4.0	Low	Brisk walk, beginner yoga, slow cycling
4.0 – 6.0	Moderate	Jogging, Zumba, moderate cycling, circuit training
6.0 – 8.0	High	Running, swimming laps, HIIT, heavy lifting
8.0 – 10.0+	Very High	Sprints, Tabata, competitive sports, stair sprints

Key Statistical Value Summary

stats summary	age	bmi	calories	duration	heart_rate	height	weight	Intensity Index	scaled_intensity_index
count	13782	13782	13782	13782	13782	13782	13782	13782	13782
mean	41	25.49	199.2	32.4	104	174.6	78.0	33.05	2.50
median	39	24.67	124.0	21.0	99	175.0	77.0	18.93	1.85
mode	20	23.34	12.0	17.0	94	171.0	63.0	12.78	1.56
std	16	4.53	222.1	30.5	20	13.9	17.0	36.14	1.66
min	18	15.03	1.0	1.0	68	132.0	38.0	0.49	1.00
25%	28	23.10	56.0	12.0	90	164.0	64.0	9.50	1.41
50%	39	24.67	124.0	21.0	99	175.0	77.0	18.93	1.85
75%	53	26.40	224.9	40.0	111	186.0	90.0	41.57	2.89
max	79	42.98	999.0	119.0	165	200.0	128.0	196.42	10.00

Key insights from the summary

Demographics & Physical Metrics:

Sample size: 13,782 individuals

Average Age: 41.3 years

Average BMI: 25.5
→ Indicates borderline overweight

Average Height: 174.6 cm

Average Weight: 78.0 kg

Activity & Health Metrics:

Average Calories Burned: 199 kcal
→ Median: 124 kcal (Right-skewed distribution)

Average Duration: 32.4 minutes
→ Median: 21 minutes

Average Heart Rate: 104 bpm

Intensity & Efficiency:

Mean Intensity Index: 33.1

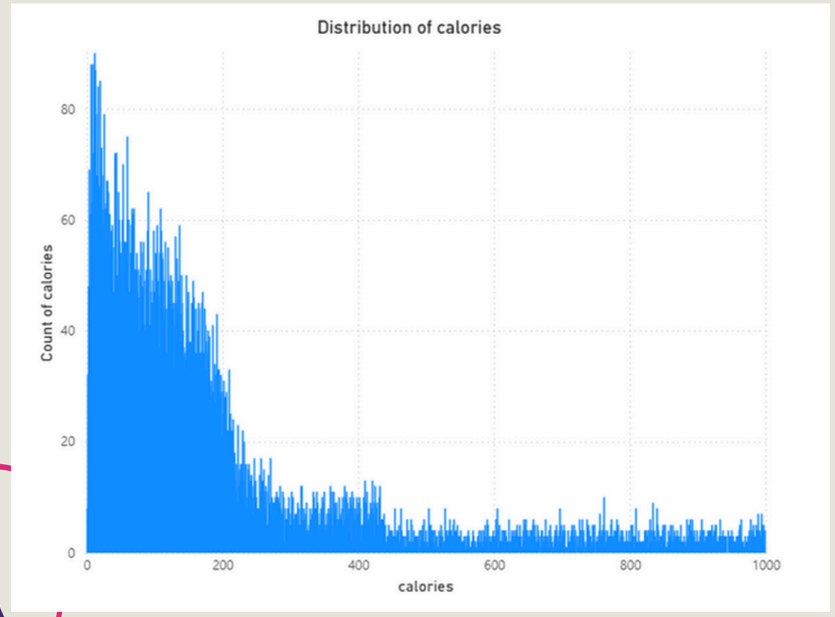
Scaled Intensity Index: 2.50

Calories per kg of Body Weight: 2.63
→ Standardized energy burn rate

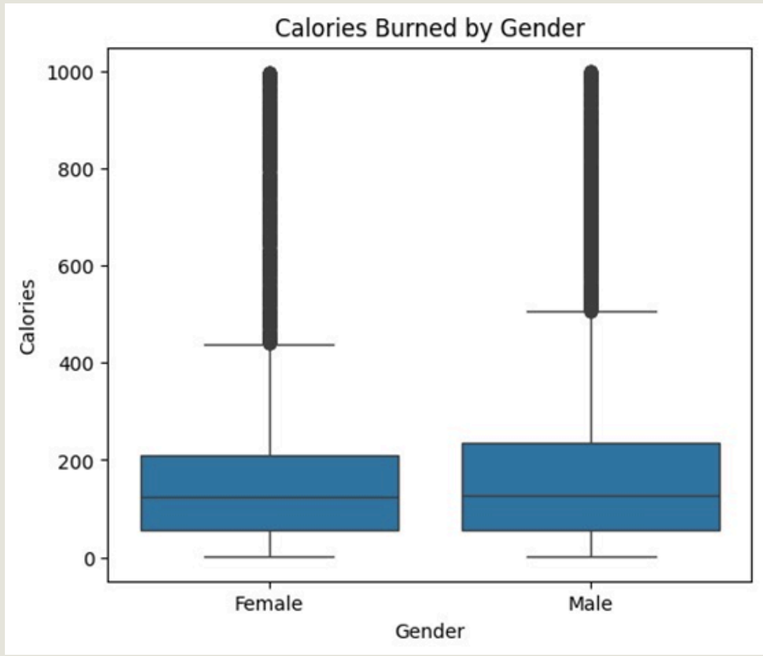
Data Analysis I

Descriptive Analysis

- ***Calories Burned Distribution***
  - Slight right-skew: Most burns are low to moderate

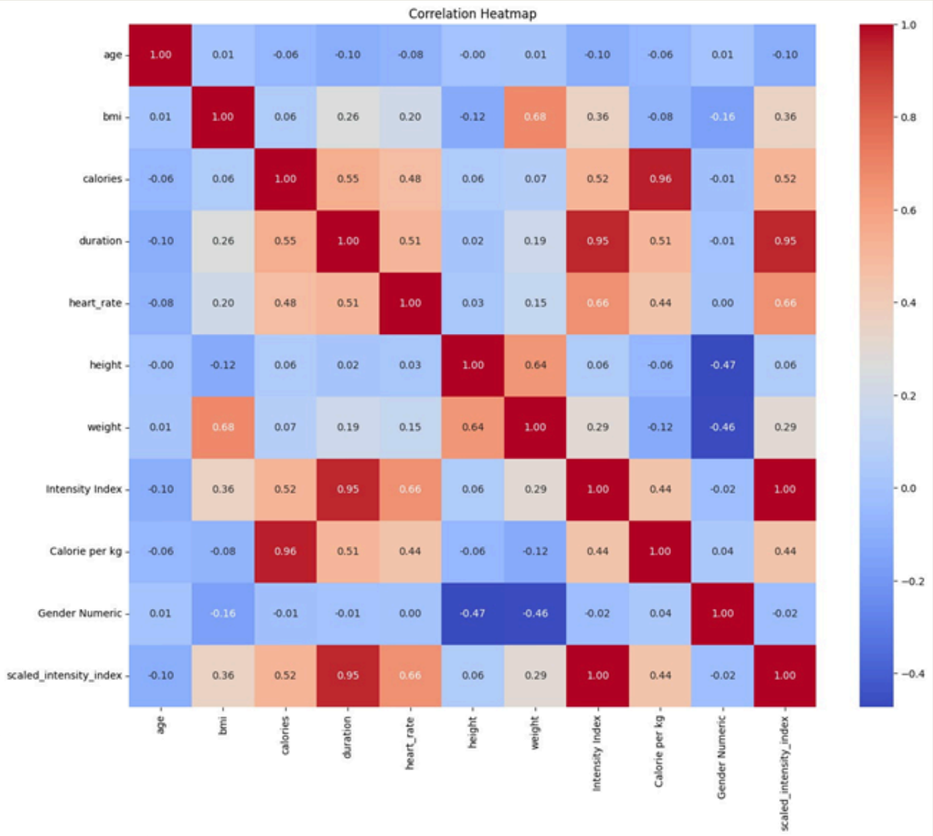


- ***Calories by Gender (Boxplot)***
  - Males burn slightly more calories on average than female participants.
  - Mean Calories Female: 197.06
  - Mean Calories Male: 201.33



Data Analysis II

Correlation Heatmap

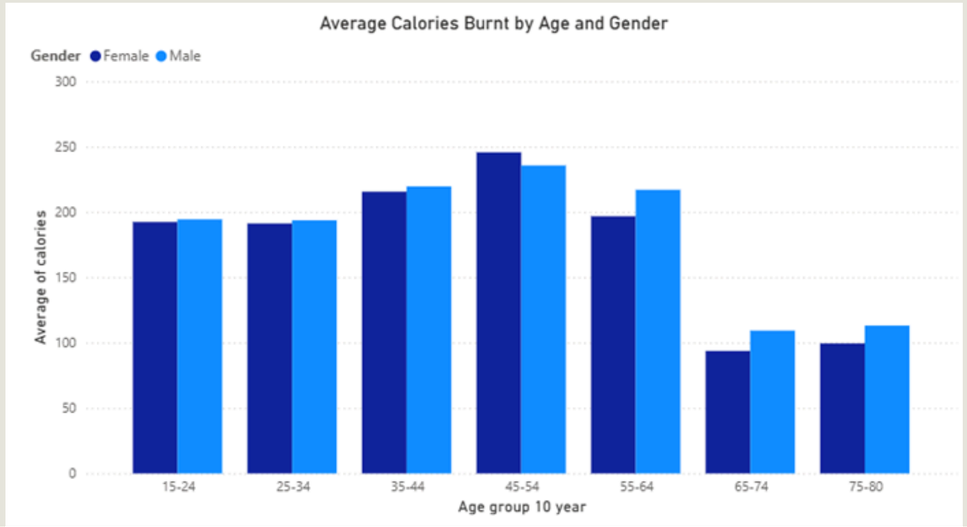


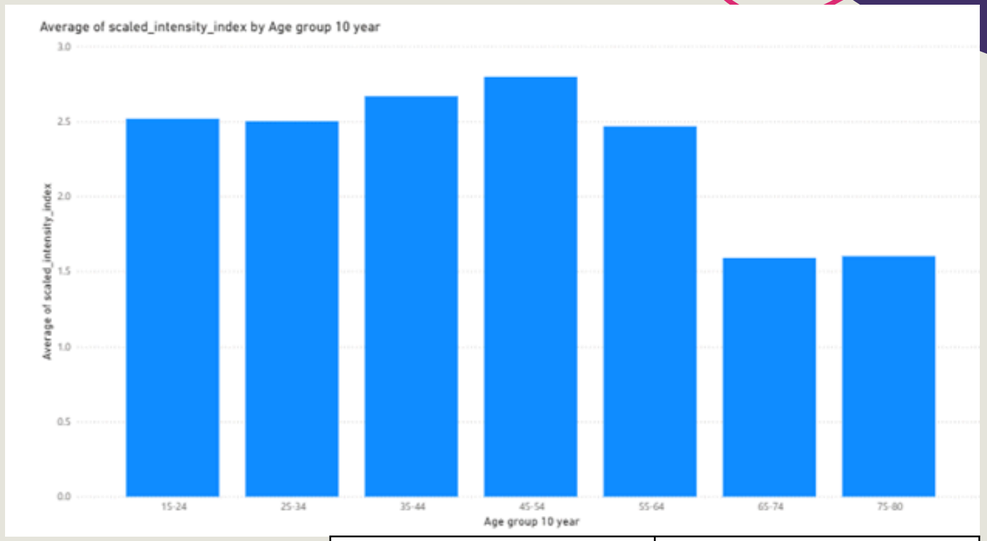
Variable Pairs	Correlation Coefficient
Heart Rate & Calories	0.48
Duration & Calories	0.55
BMI & Calories	0.06
Intensity Index & Calories	0.52

The heat map shows that Heart Rate, Duration & the feature engineered Intensity Index are the strongest predictors of calories burned.

Data analysis III

Calories burnt by Age,Gender & Does Intensity depend on Age?





Age	Calories Burnt Male	Calories Burnt Female
15-24	194.42	192.42
25-34	193.7	191.28
35-44	219.6	215.57
45-54	235.65	245.69
55-64	217.02	196.8
65-74	109.27	93.69
75-80	113.17	99.53

<u>Average Intensity: 2.31</u>

We can observe that although there is a slight decrease in the intensity index for ages 55+, intensity itself does not depend completely on age but on the type of workout, heart rate and duration performed.

Age	Average Intensity (Scaled)
15-24	2.52
25-34	2.5
35-44	2.67
45-54	2.79
55-64	2.47
65-74	1.59
75-80	1.69

Data Analysis IV

Relation of Heart Rate and Duration

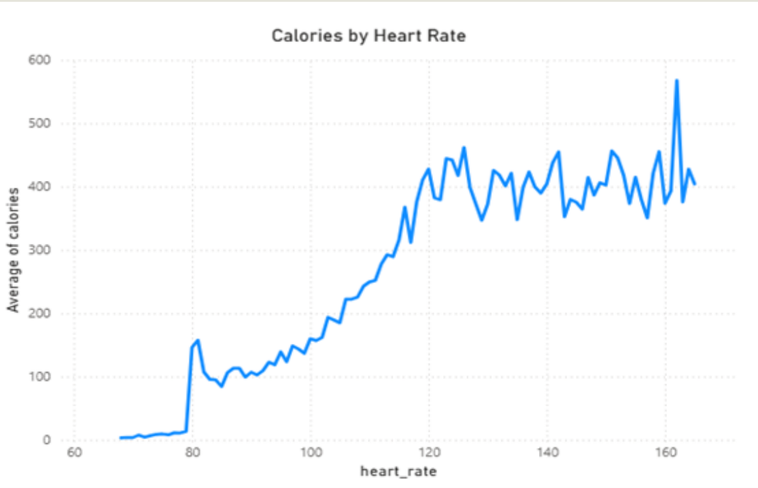


Fig: Calories Burnt by Heart Rate

It can be concluded that increase in heart rate is directly correlated with increase in duration.
Thus, an increase in calories burnt will also be observed.

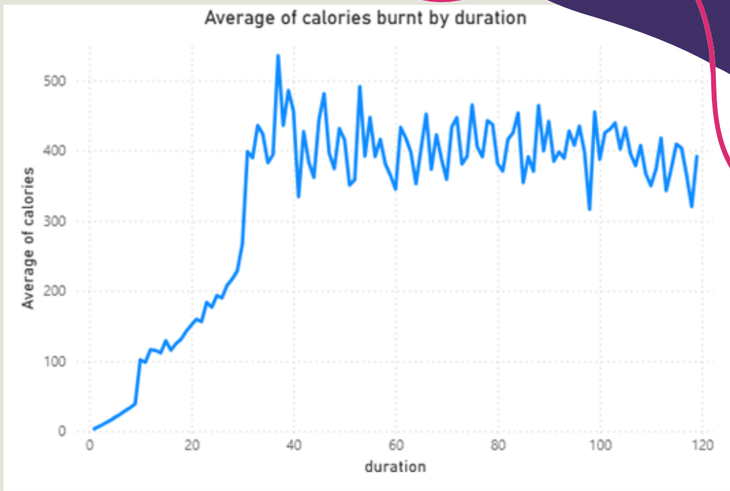


Fig: Calories Burnt by Duration

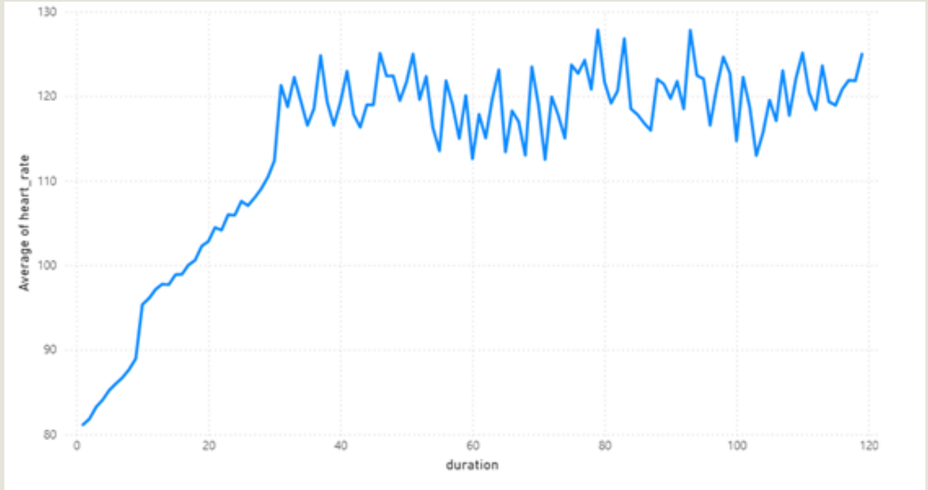


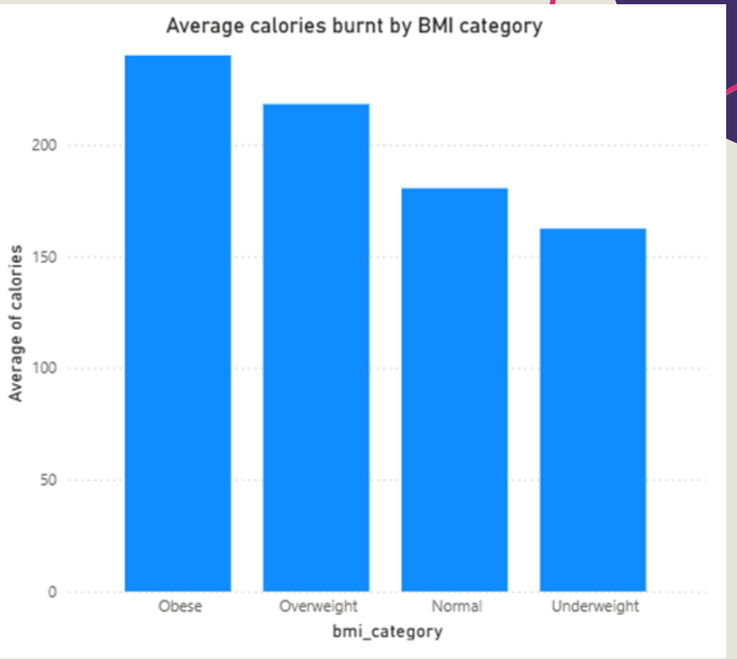
Fig: Heart Rate and Duration Relation

Heart Rate Range	Average Duration (Mins)	Average Calories (Kcal)
Low(<=79 bpm)	3.87	10.91
Moderate (80-99 bpm)	21.21	118.66
High (100-119 bpm)	34.26	221.42
Very High (120+ bpm)	63.41	404.01

Data Analysis V

Calories burnt by BMI category:

The calories burnt by BMI category follows the trend of Obese>Overweight>Normal>Underweight



BMI Classification

BMI	Category
Lower than 18.5	Underweight
18.5 up to 25	Optimal
25 up to 30	Overweight
30 upwards	Obese

BMI Category	Calories burnt
Underweight	162.41
Normal	180.46
Overweight	218.13
Obese	239.95

Inferential Analysis

Objective:

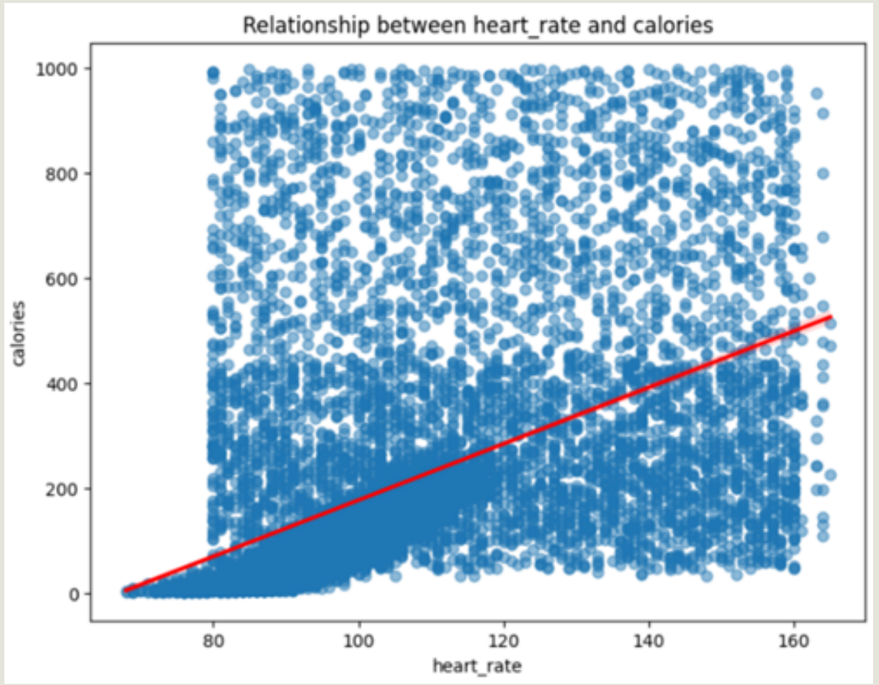
Test if Heart Rate significantly affects Calories Burned using statistical inference.

Hypothesis Test

- Null Hypothesis (H_0):
  Heart rate has no effect on calories burned.
- Alternative Hypothesis (H_1):
  Heart rate does affect calories burned.

Conclusion

- p-value < 0.05 → Reject H_0
- Heart rate has a statistically significant positive effect on calories burned
- As heart rate increases, calories burned also increase linearly



Metric	5.37
Slope	5.37
p-value	0
R-squared	0.232
Significance Level α	0.05

Regression Significance I

simple linear regression of Calories on duration show:

Feature: duration

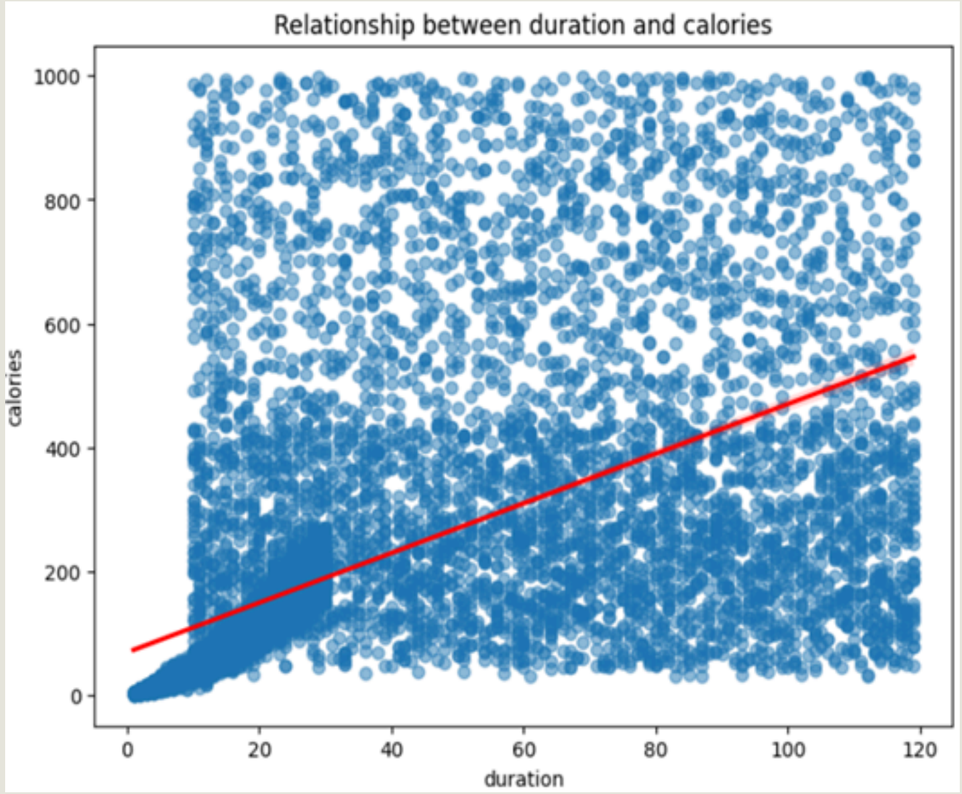
$y = 4.01x + 69.33$

$R^2 = 0.303$

Parameter	Coefficient	t-statistic	p-value
Duration	4	77.41	<0.001

Given that $p<0.001$ it can be concluded that duration is an important parameter for calories burnt.

For every minute increase in duration additional 4kcal is burnt.

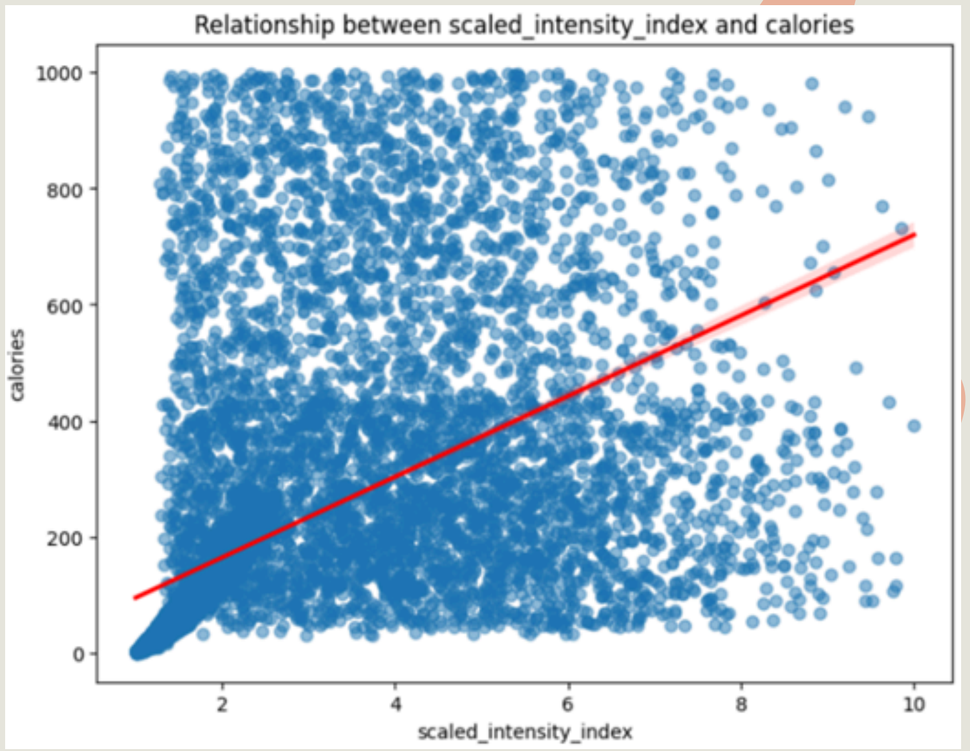


Regression Significance II

simple linear regression of Calories on Scaled Intensity Index shows:

Feature: Scaled Intensity Index
y = 69.40 * x + 26.01
R-squared = 0.269

Parameter	Coefficient	t-statistic	p-value
Scaled Intensity Index	69.4	71.24	<0.001



Given that p<0.001 it can be concluded that scaled intensity index is an important parameter to predict calories burnt.

For every unit value increase in Scaled Intensity Index additional 69.4kcal is burnt.

Machine Learning Model Performance

To predict calories burned, multiple models were trained and compared.
Age, Gender Encoded(Male=0 & Female=1), Weight, Height, BMI, Duration, Heart Rate, Scaled Intensity Index were the features used.

LightGBM achieved the best performance with an R^2 of 0.6526, indicating high predictive accuracy.

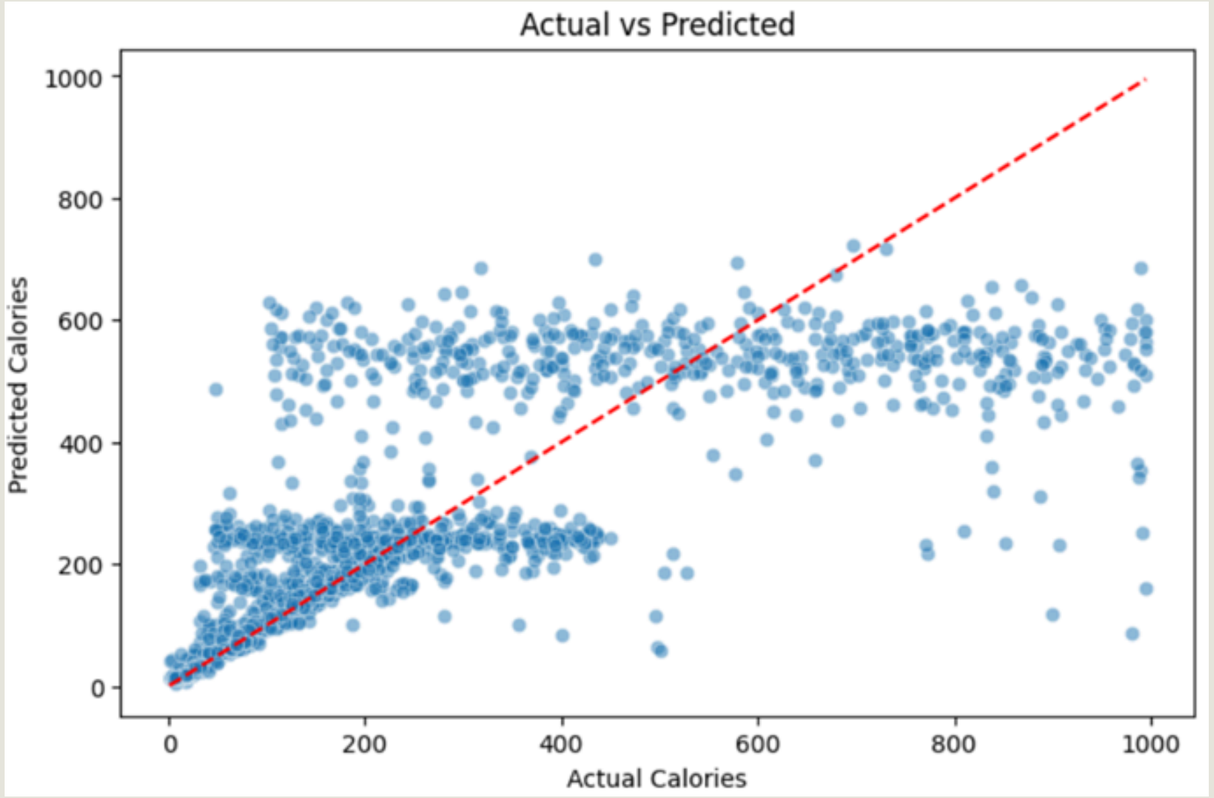
Model	Hyperparameter Type	Hyperparameter Values
LGBMRegressor	**Tuned Hyperparameters:**	n_estimators=1000, learning_rate=0.05, max_depth=7, random_state=42, eval_metric='l1', callbacks= early_stopping(stopping_rounds=50), log_evaluation(period=50)

Table: Model performance comparison

Models	MAE	RMSE	R^2
Linear Regression *(simple)*	99.22	165.2	0.4054
Random Forest *(with default hyperparameters)*	68.43	128.77	0.6387
Gradient Boosting *(with default hyperparameters)*	65.37	127.55	0.6455
Gradient Boosting *(with tuned hyperparameters)*	65.77	131.08	0.6257
XGBoost *(with default hyperparameters)*	67.58	132.41	0.618
XGBoost *(with tuned hyperparameters)*	64.02	127.31	0.6469
LightGBM *(with tuned hyperparameters)*	**64.32**	**126.28**	**0.6526**
CatBoost *(with tuned hyperparameters)*	64.98	126.68	0.6504

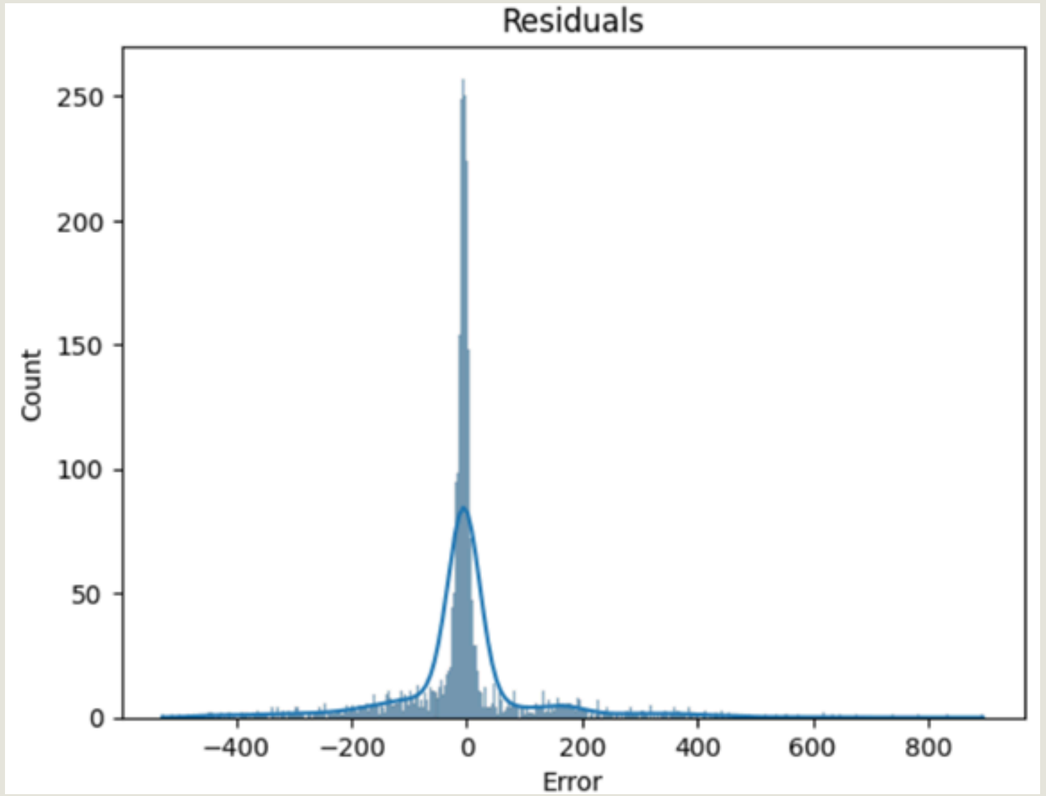
Actual vs Predicted Calories Plot

The plot shows that predictions align closely with actual values, with minimal spread, indicating low bias and variance.



Residuals Plot

The residuals are randomly scattered around zero, confirming no clear pattern and supporting model adequacy.



Linear Scaling Justification

Underweight & Obese groups showed unrealistically high calorie burn

- Underweight: High burn not expected physiologically
- Obese: Burn too high compared to Overweight

Fix:

- Applied linear scaling to calorie values
- Underweight < Normal
- Obese > Overweight (10% more than Overweight)

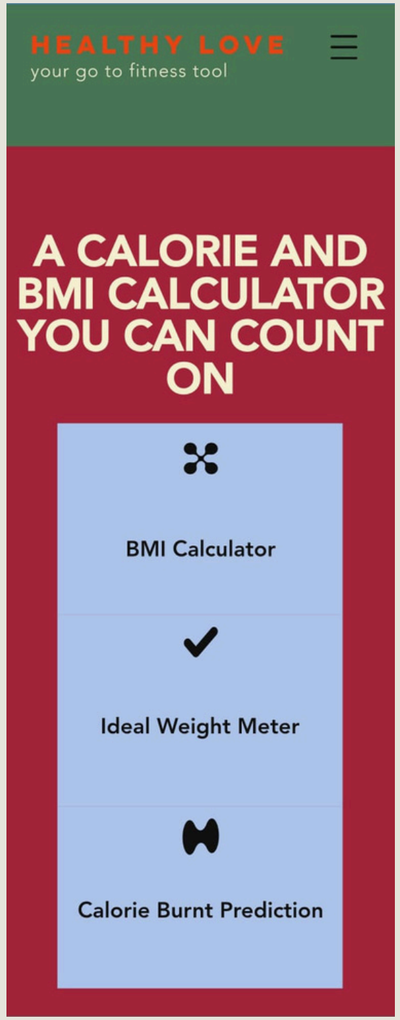
Why this method?

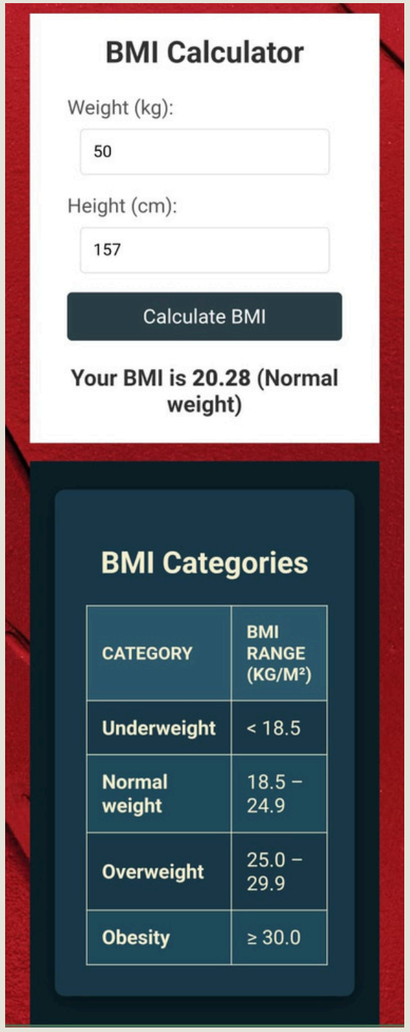
- Preserves data distribution
- Keeps all records
- Reduces model bias
- Easy to reproduce & interpret

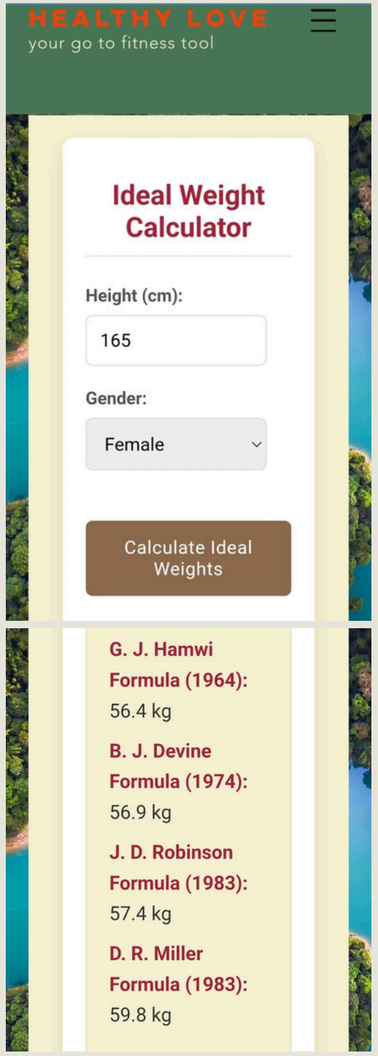
BMI Category	Mean Calories (Before Scaling)	Mean Calories (After Scaling)
Underweight	562.01	162.41
Normal	180.46	180.46
Overweight	218.13	218.13
Obese	546.32	239.94

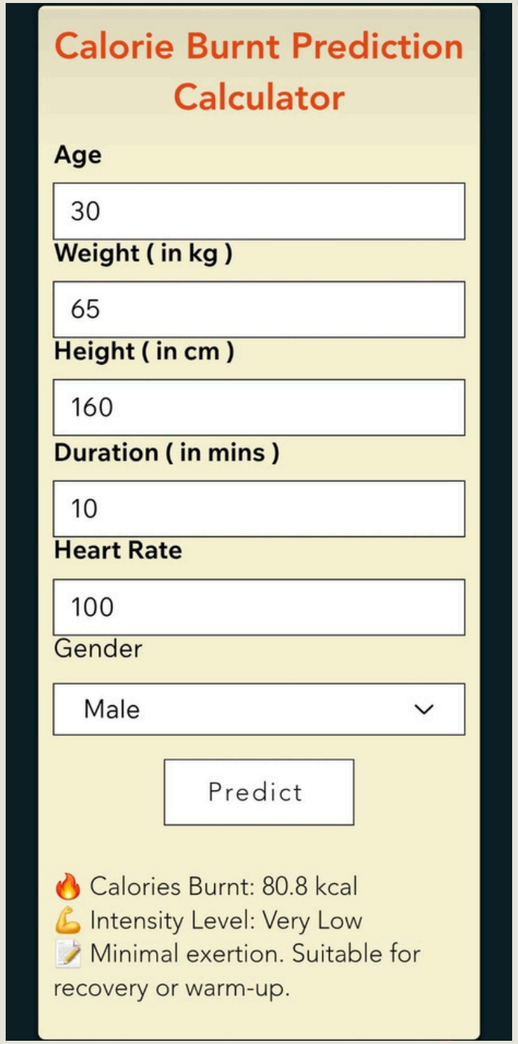
Website Overview

https://pdastdasroy.wixsite.com/healthylove

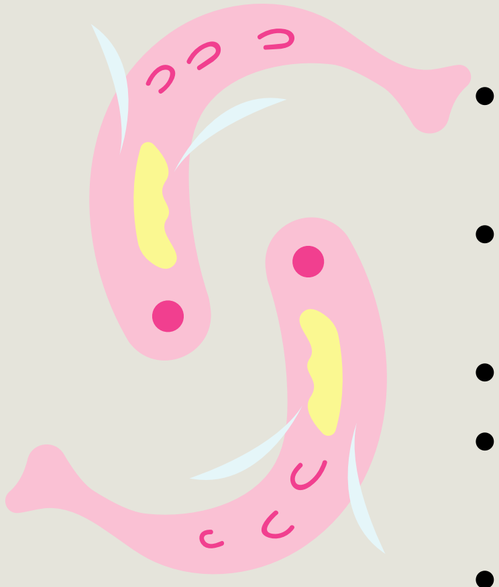








Conclusion



- Heart Rate and Duration are the strongest predictors of calories burned.
- Feature engineered Intensity Index had a strong correlation with calories burnt.
- LightGBM (with tuned hyperparameters) outperform other models for this dataset.
- Inferential analysis was proven true.
- Intensity is correlated to duration and heart rate more than age and gender.
- Low to Moderate intensity workouts are more preferred than high intensity workouts.
- Further model building can be done with consideration to training with activity type.

Project Limitations

- Both the datasets are potentially synthetic.
- Not designed for real-world use, only for academic/demo purposes.
- Linear scaling applied only to calories, not heart rate (outliers remain).
- BMI included in model training despite height & weight being present (BMI had a high feature importance for presenting calorie outputs).
- Intensity levels scaled based on our dataset — may not generalize.
- Intensity Index is custom & lacks MET-based accuracy (Due to no presence of O2 levels in the dataset).

Appendices

- https://www.mayoclinic.org/healthy-lifestyle/weight-loss/in-depth/metabolism/art-20046508
- https://www.healthline.com/health/fitness-exercise/how-many-calories-do-i-burn-a-day
- https://www.everydayhealth.com/fitness/factors-that-can-affect-how-many-calories-you-burn
- https://www.urmc.rochester.edu/news/publications/health-matters/is-bmi-accurate
- https://www.coospo.com/blogs/knowledge/calories-burned-by-heart-rate-understanding-the-connection
- https://www.mdanderson.org/publications/focused-on-health/How-to-determine-calorie-burn.h27Z1591413.html
- https://www.kaggle.com/datasets/ruchikakumbhar/calories-burnt-prediction
- https://www.kaggle.com/datasets/adilshamim8/workout-and-fitness-tracker-data
- https://pdastdasroy.wixsite.com/healthylove/calorie-burnt-prediction
- https://drive.google.com/drive/folders/1s5wugHS8bss--r9ka_63QApo86QHVih_?usp=drive_link
- https://github.com/Pratyasha-Tapaja/calorie_prediction_project

THANK YOU