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

1. Abstract

This project aims to build a calorie prediction model using demo workout and physiological data. The datasets include key features such as age, gender, height, weight-, workout duration, heart rate, BMI, Intensity Index etc. Extensive exploratory data analysis (EDA), feature engineering and multivariate analysis were conducted to uncover patterns and refine features. The data was pre-processed, cleaned, and engineered to enhance model performance. Multiple regression algorithms were tested to identify the best-performing model for predicting calories burned. The final model helps users estimate calorie expenditure based on individual and workout-specific parameters. A user-friendly web application is planned for deployment to make the tool easily accessible. This project demonstrates the practical use of data science in detailed exploration of fitness data and personalized fitness tracking. Insights from the EDA also highlight demographic and physiological factors influencing calorie burn. Overall, this work bridges fitness tracking and data-driven prediction for better health management and understanding.

2. Introduction

Project's Relevance

This project aims to investigate whether heart rate has a direct correlation with calories burned, and how a custom-built Intensity Index can improve calorie prediction accuracy. Traditional calorie estimation methods often overlook dynamic workout parameters like heart rate and workout intensity, leading to generalized results. Our approach uses individual physiological data; such as age, gender, BMI, along with real-time workout data, to assess how strongly heart rate reflects actual energy expenditure.

By introducing and validating a scaled Intensity Index, we aim to quantify workout effort in a more precise and standardized way. This research not only contributes to building smarter, self-serving fitness tools, but also demonstrates how data science can enhance health tracking and decision-making. The findings may benefit individuals aiming to personalize their fitness plans, as well as health professionals and app developers focused on delivering data-driven wellness solutions.

Link to the files: https://github.com/Pratyasha-Tapaja/calorie_prediction_project

Technologies Involved

- **Python (Colab)** - main environment for analysis
- **Pandas, NumPy** - data cleaning and manipulation
- **Matplotlib, Sea born** - visualization and EDA
- **Scikit-learn** - machine learning model building
- **Power BI -** advanced EDA visualization
- **Render –** Flask API
- **Wix –** Model deployment

Background Material Survey

A background material survey was conducted to understand the existing methods for estimating calories burned, including the use of physiological data and exercise parameters in popular fitness apps and related studies.

How and why do we burn calories?

Metabolism is the process by which the body changes food and drink into energy. During this process, calories in food and drinks mix with oxygen to make the energy the body needs.

Even at rest, a body needs energy for all it does. This includes breathing, sending blood through the body, keeping hormone levels even, and growing and repairing cells. The number of calories a body at rest uses to do these things is known as basal metabolic rate, also called **basal metabolism**.

Muscle mass is the main factor in basal metabolic rate. People who are larger or have more muscle burn more calories, even at rest. Men usually have less body fat and more muscle than do women of the same age and weight. That means men burn more calories. With aging, people tend to lose muscle. More of the body's weight is from fat, which slows calorie burning.

Besides the basal metabolic rate, two other things decide how many calories a body burns each day:

How the body uses food. Digesting, absorbing, moving and storing food burn calories. About 10% of calories eaten are used for digesting food and taking in nutrients. This can't be changed much.

How much a body moves. Any movement, such as playing tennis, walking to a store or chasing the dog, makes up the rest of the calories a body burns each day. This can be changed a lot, both by doing more exercise and just moving more during the day.

A closer look at physical activity and metabolism

We can't easily control the speed of our basal metabolic rate, but we can control how many calories we burn through physical activity. The more active we are, the more calories we burn. In fact, some people who seem to have a fast metabolism are probably just more active. To burn more calories, the Physical Activity Guidelines for Americans recommends the following:

- **Aerobic activity**. As a general goal, aim for at least 30 minutes of moderate physical activity every day. If you want to lose weight, maintain weight loss or meet specific fitness goals, you may need to exercise more.
- Moderate aerobic exercise includes activities such as brisk walking, biking, swimming and mowing the lawn. Vigorous aerobic exercise includes activities such as running, heavy yard work and aerobic dancing.

- **Strength training.** Do strength training exercises for all major muscle groups at least two times a week. Strength training can include use of weight machines, your own body weight, heavy bags, resistance tubing or resistance paddles in the water, or activities such as rock climbing.

Men burn more calories than women:

Men and women burn calories at different rates. This is why sex is included as a variable in the equation, along with age and weight, which also affect the number of calories a person burns.

Men generally have less body fat than women. They also typically have more muscle mass, which means the body burns more calories at rest.

Men generally burn more calories than women overall. But a person's body composition and hormone levels plays an important role as well.

How height affects calories burnt:

Taller individuals tend to have a higher basal metabolic rate (BMR), which is the number of calories burned at rest, because they have more body mass to maintain.

Thus often shorter people struggle losing weight than taller people because their daily dietary calorie intake is already significantly lower than that of tall people thus they cannot calorie deficit equal amount as tall people and will have to rely more on calorie burning.

How weight affects calories burnt:

Calories are just a measure of energy, so the more we weigh, the more energy it takes to move our body." Put differently, of two people with different weights, the one who weighs more will burn more calories, because they have a greater energy expenditure when moving.

People with larger bodies also tend to have larger internal organs (such as the heart, liver, kidneys, and lungs), which is a significant factor in how many calories are burned during exercise and at rest, because these organs and their processes require energy. One study [1] found that up to 43 percent of the variation in total calorie burn between people could be explained by differences in the size of their internal organs.

How age affects calories burnt:

After age 30, we begin to lose as much as 3 to 5 percent of our muscle mass per decade. The reasons for this aren't perfectly understood, but one review [2] explains that it's likely because your body becomes more resistant to hormones that promote the protein synthesis that's key to muscle maintenance. This loss of muscle mass lowers our metabolic rate — the speed at which we burn calories — at rest and during exercise.

A study published in 2021[3] made headlines for its findings that metabolic rate may not decline throughout adulthood, but rather that it plateaus between the ages of 20 and 60 and then begins its decline. But this doesn't necessarily mean that everyone's calorie burn stays constant through adulthood.

How fitness level affects calories burnt:

The more we do a certain type of workout, the easier it seems. That's not in our head — our body actually does adapt to do things more easily over time. It means that we can run faster or for longer with practice, and our muscles will be able to lift heavier weights with proper training.

But it also affects our calorie burn. As our body adapts to training, we will burn less calories with the same workouts. That's why a newbie might burn significantly more calories than someone who's been doing the same workout for years. It's also why changing our workout routine (such as switching the time of day we work out or the type or order of exercises) can increase pur fitness level and potentially enhance calorie burn.

How intensity affects calories burnt:

It’s also possible that two people doing the same workout are burning a different number of calories because they’re not actually doing the same workout. someone exercising at a high intensity, meaning they’re breathing heavily and can’t carry on a conversation, can burn twice as many calories in the same amount of time as someone exercising at a low intensity. And just because they’re covering the same distance as someone else, or going through the same motions, doesn’t mean that the two of them are working out at the same intensity.

Heart rate and calories burnt:

Yes, we generally burn more calories when our heart rate is higher. When we exercise, our heart beats faster to send more oxygen-rich blood to our muscles. The harder we exercise, the more our heart has to work and the more oxygen our muscles need. Because of this, our body uses more calories to keep up with the extra energy needed.

During the intense parts of a workout (HIIT) , our heart rate goes up a lot, which helps us burn a lot of calories both while exercising and afterward. This is called excess post-exercise oxygen consumption (EPOC) or the afterburn effect, meaning our body keeps burning calories even after we finish our workout.

BMI (Body Mass Index) is not a reliable measure for estimating calorie burn by itself. BMI primarily reflects weight relative to height, not body composition or metabolic activity, which are key factors in calorie expenditure. Heart rate is generally a more accurate indicator of calorie burn because it reflects the body's physiological response to physical activity.

Topics taught at IDEAS ISI, Kolkata

1st Week (Common Theme)
1. Introduction - Welcome Note - What to expect from this internship
2. Data Science Introduction
3. How to do a Research Project
4. Data Visualization - Power BI
5. Career & Life-Design: Shaping Your Future

2nd Week (Track Specialized): DATA SCIENCE GROUP 1 (Students with no / low Python coding Skills)
1. Python Programming for Data Science
2. Survey & Questionnaire Design 5
3. Sentiment Analysis, Text Analytics (no code tools)
4. Foundations of Generative AI and LLM
5. Gen AI support for BI

3. Project Objective

(i) Build a Calorie Prediction Model

Using user attributes like:

- Gender
- Age
- Height
- Weight
- Duration of exercise
- Heart rate

➤ Goal: Predict calories burned using machine learning.

(ii) Perform EDA (Exploratory Data Analysis)

Analyze the dataset(s) to:

- Understand trends in calorie burn
- Identify the most influential factors
- Spot outliers, patterns, or missing values

➤ Goal: Summarize insights & EDA outcomes visually, Shape the feature selection and model building.

(iii) To determine whether Heart Rate is a significant and direct predictor of Calories Burnt during exercise.

(iv) Feature engineer an Intensity Index which correlates to calories burnt

- An attempt to correlate Heart rate, duration and weight directly to calories burnt using a simple formula..

(v) Build a website on wix

- Let users input their data (age, weight, duration, etc.)
- Predict the number of calories they'll burn
- BMI calculator
- Ideal weight calculator

➤ Goal: Real-world utility and demo-readiness.

4. Methodology

1. **Data Collection**

1.1. Sources of data

Two datasets were used

1.1.1 First Dataset

- Source: Public fitness tracker dataset (`calories.csv`) obtained from Kaggle.
- Contains demographic and activity-related data for 5000 individuals

1.1.2 Second Dataset

- Source: Synthetic dataset (`workout_fitness_tracker_data.csv`) to augment the original records from Kaggle
- Includes similar features to ensure consistency

No direct survey was conducted for this project.

1.2 Steps of Data Collection

1.2.1 Data Sourcing: Downloaded raw datasets.

1.2.2 Merging: Combined data for a better distributed dataset.

2. Tools Used

- **Python (Colab)** - main environment for analysis
- **Pandas, NumPy** - data cleaning and manipulation
- **Matplotlib, Seaborn** - visualization and EDA
- **Scikit-learn** - machine learning model building
- **Power BI -** advanced EDA visualization
- **Render –** Flask API
- **Wix –** Model deployment
- **MS PowerPoint** - presentation design

3. Data Cleaning & Pre-processing

The following steps were taken:

1.**Inspection:** Checked for missing values, duplicates, outliers.
2.**Handling Missing Values:** Imputed or dropped rows/columns as necessary.
3□.**Outlier Detection:** Visualized numeric features (boxplots) and removed extreme outliers.
4□.**Feature Engineering:**

- Created ***BMI*** using Height & Weight.
- Computed ***Intensity Index*** using Heart Rate, Duration and Weight.

5. **Linear Scaling:** Applied standard scaling for model consistency.

   Justification:

   During our data validation phase, we observed that the Underweight and Obese BMI categories exhibited unusually high average calorie expenditure values compared to the Normal and Overweight groups. While it is physiologically plausible for individuals with obesity to burn more calories due to higher body mass, the magnitude of the difference in our dataset appeared unrealistic when cross-checked against typical energy expenditure patterns.

   For the Underweight category, the consistently high calorie burn contradicted established physiological expectations — as individuals with very low BMI typically have less muscle mass and energy reserves, and are less likely to perform very high-intensity or long-duration activities.

   To address these inconsistencies without discarding valuable records, we applied a proportional (linear) scaling correction to the calorie values for these two BMI groups. This method preserved the original distribution and relative differences within each group but shifted the average calorie expenditure to align with realistic physiological trends:
   - **Underweight:** Adjusted to ensure the average calorie expenditure is lower than that of the Normal BMI group.
   - **Obese:** Adjusted to ensure the average calorie expenditure remains slightly higher than the Overweight group but avoids unrealistically large values. (10% higher than Overweight)

   This approach allowed us to retain all valid data points, minimize the risk of biasing our machine learning model with outlier-driven patterns, and maintain the overall integrity of the dataset. Linear scaling is transparent, easy to reproduce, and does not distort the relationships between features, which is critical for valid modeling and interpretability.

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

6. **.Encoding Categorical Variables:** Converted *Gender* to numeric (e.g., Male = 0, Female = 1).

4. Exploratory Data Analysis (EDA)

- Univariate analysis: Histograms, summary statistics.
- Bivariate analysis: Correlation heat map, scatterplots.
- Group analysis: Calories burned by gender, age groups, duration buckets.
- Insights identified relationships, e.g., Heart Rate strongly influences calorie expenditure.

5..Feature Engineering

5.1 Intensity Index:

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

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

Justification:

1. *Purpose:*

The goal of the Intensity Index is to quantify how physically demanding a workout is based on observable metrics available in the dataset. It is used to estimate workout effort in a single comparable value across users.

2. *Variables Used:*

- Heart Rate: Reflects real-time physical effort and cardiovascular load.
- Duration: Longer workouts typically indicate greater energy expenditure.
- Weight: Heavier individuals tend to burn more calories for the same exercise. We apply the square root to weight to reflect diminishing returns and avoid overweight bias.

3. *Scaling:*

The division by 1000 is to bring the index to a readable scale (~0 to 10). It doesn’t affect correlation or relative comparisons, only readability.

4. *Correlation with Calories:*

This formula was selected because it has a strong positive correlation (+0.52) with calories burned, validating it as a good proxy for workout intensity.

5. *Exclusions (BMI and Gender):*

- BMI: Already reflected via weight, and doesn't add significant value beyond it.
- Gender: Excluded for simplicity. Heart rate reflects effort across all genders.

6. *Use Cases:*

- Categorizing workouts into effort bands (Very Low to Very High).
- Visualizing effort in Power BI dashboards.
- Adjusting calorie estimates or fitness scoring systems.

5.2 BMI:

$$\boldsymbol{BMI = \frac{weight(kg)}{Height^2(m)}}$$

5.3 BMI Category:

BMI	Category
<18.5	Underweight
18.5-25	Normal
25-30	Overweight
30+	Obese

6. Model Building & Validation

6.1 Model Selection

- Tested multiple algorithms:
    - Linear Regression
    - Random Forest Regressor
    - Gradient Boosting Regressor
    - Cat Boost
    - LightGBM
- Compared using cross-validation scores.

6.2 Training and Testing Split

- Data split: **80% training**, **20% testing** using train_test_split from Scikit-learn.
- Random state fixed for reproducibility.

6.3 Model Evaluation

- Evaluated using:
    - Mean Absolute Error (MAE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score
- Visual check: Plot of *Predicted vs Actual* calories burned.

6.4 Final Model.

- LightGBM performed the best with an R^2 of 0.6526
- Hyper parameter tuning (GridSearchCV) improved performance further.

7. Deployment

- Developed a basic **WIX website** for users to input features and get calorie predictions
- Code structured in app.py and tested locally.

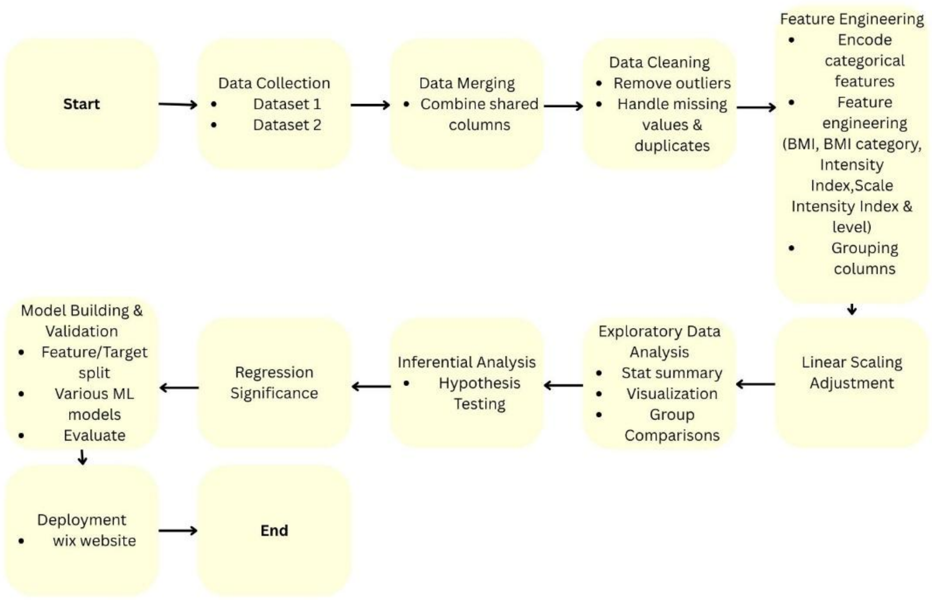


Figure 4.0: Flow chart of the Procedure

5.Data Analysis and Results

We have used two datasets for our project:

1) calories.csv, **2)** workout_fitness_tracker_data.csv.

The first dataset (15000 rows) had almost all of the personalized columns needed, except Exercise Type. Hence, we started to look for a dataset that had a column mentioning the exercise type, with similar columns of physical attributes. The second dataset (10000) we found had almost all of these columns moreover,

The final dataset for EDA and Model training was a merged dataset with the following rows-

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

5.1 Descriptive Analysis

This section presents the descriptive statistical analysis of the dataset, summarizing the basic features and distributions of the data.

Table 4.1: Summary Statistics of Key Variables

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	41.2875	25.4876	199.203	32.3957	104.183	174.617	78.0346	33.051591	2.495514
median	39	24.6704	124	21	99	175	77	18.925993	1.846644
mode	20	23.3377	12	17	94	171	63	12.776348	1.564155
std	15.6537	4.52884	222.056	30.4933	19.8998	13.8653	16.9826	36.138661	1.660058
min	18	15.034	1	1	68	132	38	0.494975	1
25%	28	23.0954	56	12	90	164	64	9.496482	1.413491
50%	39	24.6704	124	21	99	175	77	18.925993	1.846644
75%	53	26.3958	224.923	40	111	186	90	41.569821	2.886806
max	79	42.9806	999	119	165	200	128	196.42058	10

Distribution Plots:

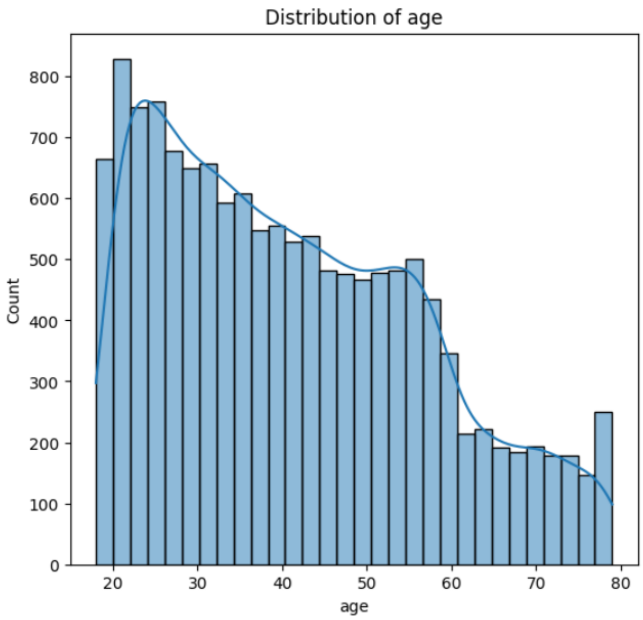


Fig 5.1: Distribution of Age

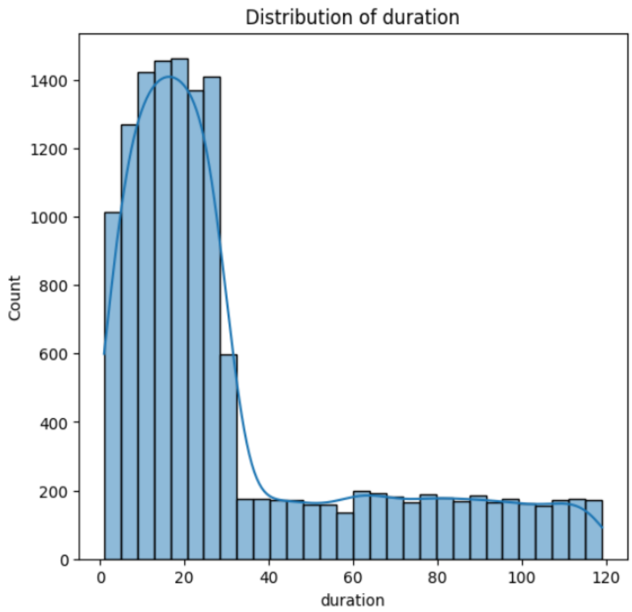


Fig 5.2: Distribution of Duration

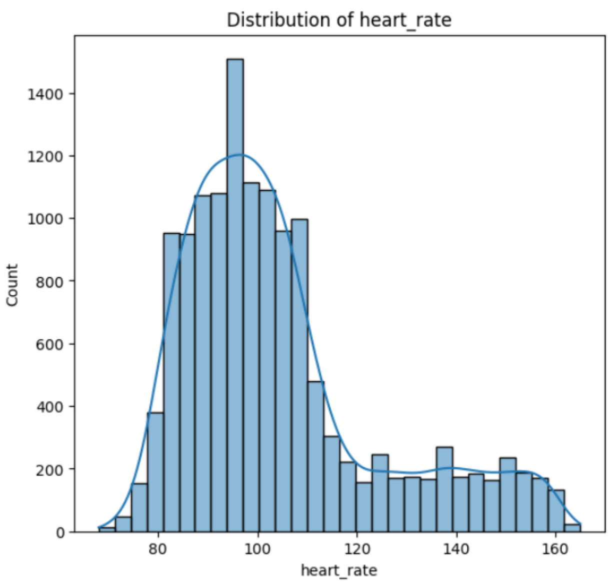


Fig 5.3: Distribution of Heart

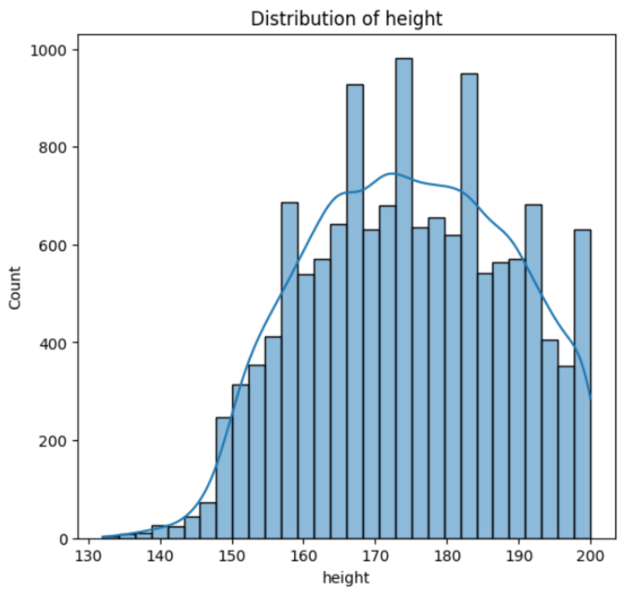


Fig 5.4: Distribution of Height

Distribution Plots:

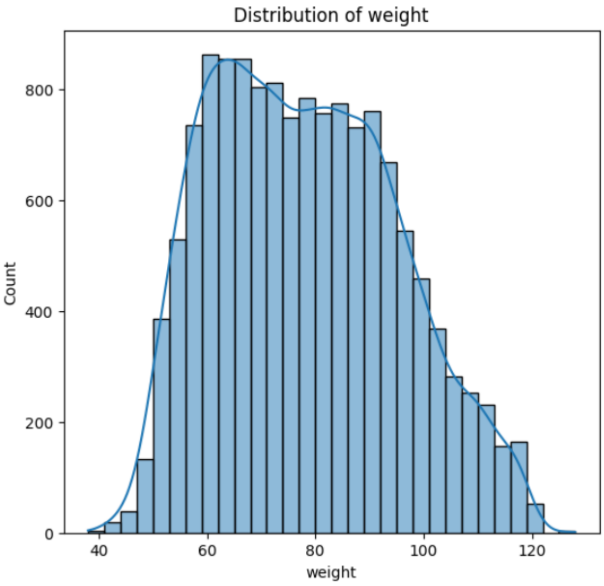


Fig 5.5: Distribution of Weight

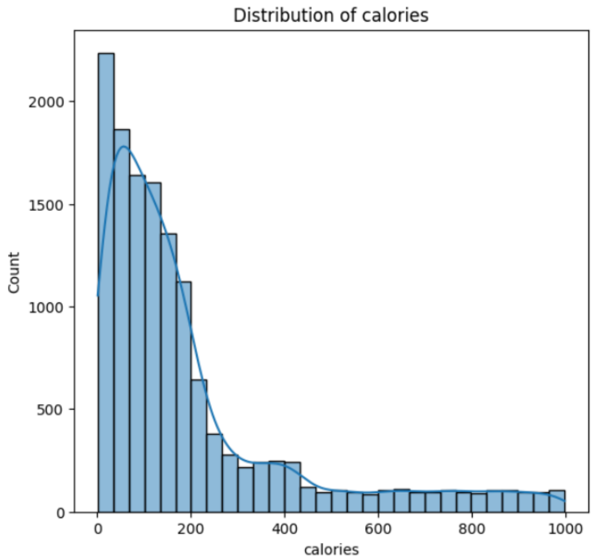


Fig 5.6: Distribution of Calories Burnt

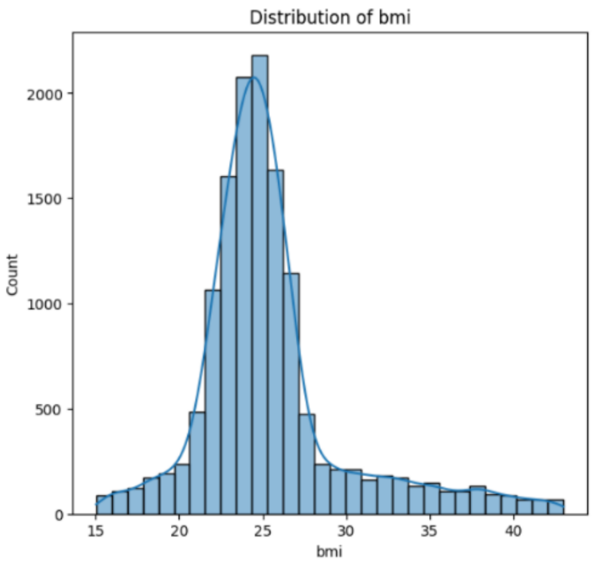


Fig 5.7: Distribution of BMI

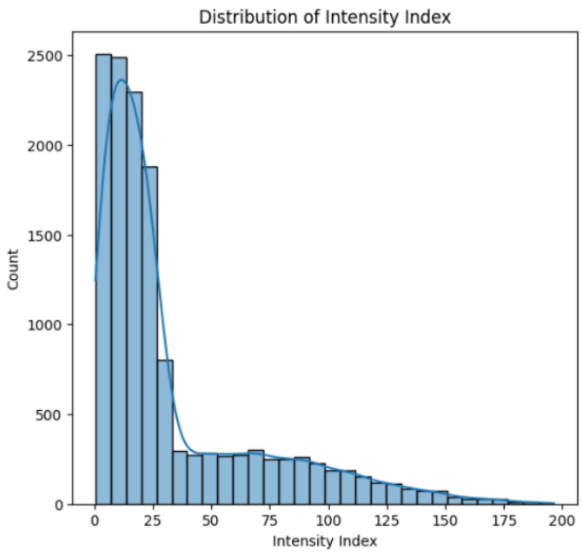
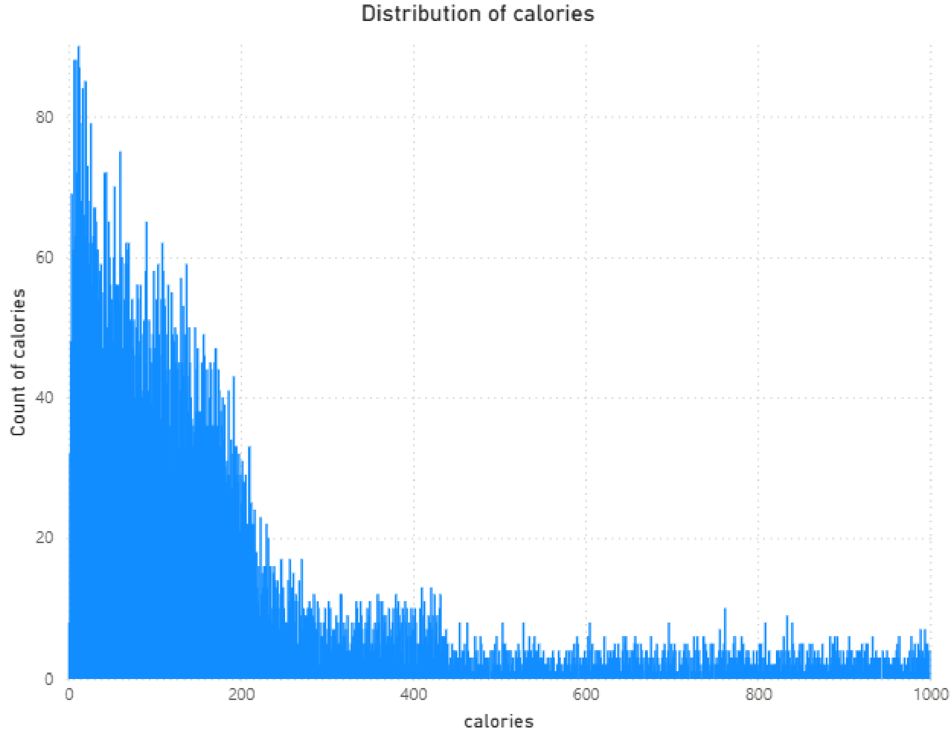


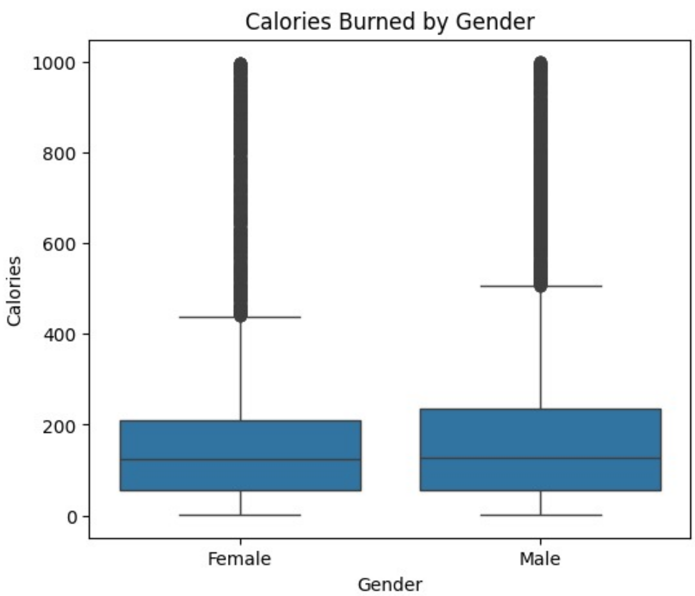
Fig 5.8: Distribution of Intensity

Figure 5.9: Histogram of Calories Burned



The histogram shows a slightly right-skewed distribution, indicating a higher frequency of low to moderate calorie burns, with few instances of very high values.

Figure 5.10: Boxplot of Calories Burned by Gender

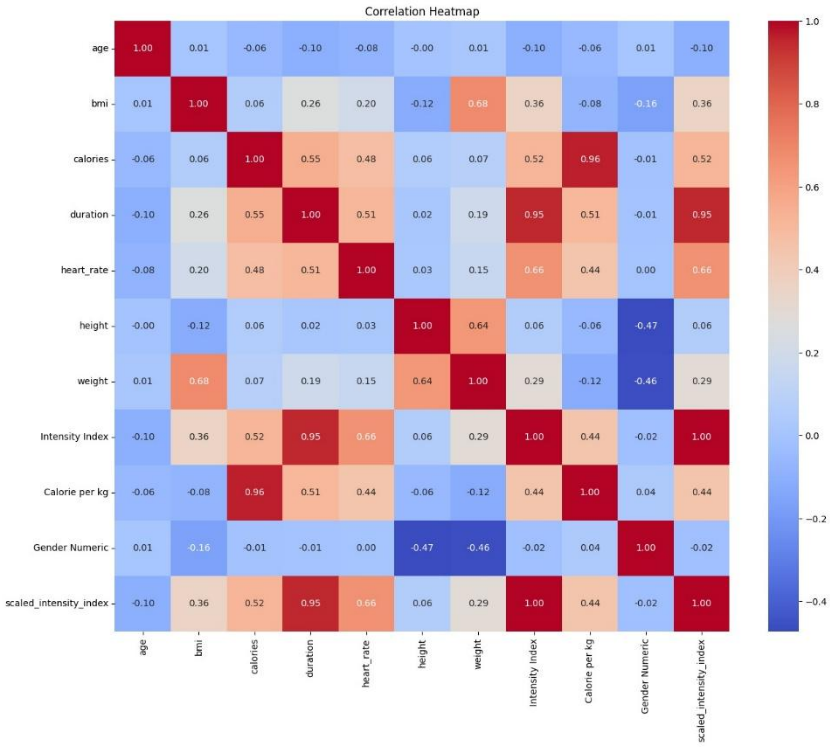


Mean Calories Female: 197.06

Mean Calories Male: 201.33

Observation: Male participants, on average, burn slightly more calories than female participants, potentially due to differences in body composition and duration.

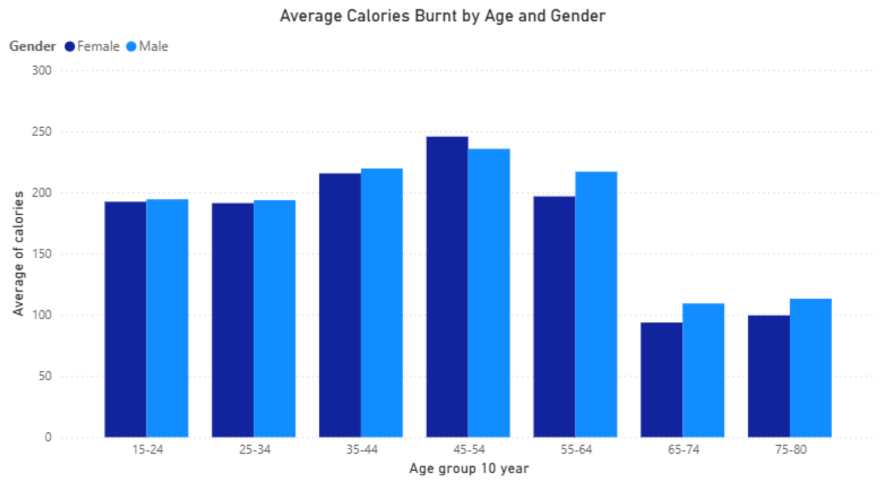
Figure 5.11: Correlation Heatmap



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

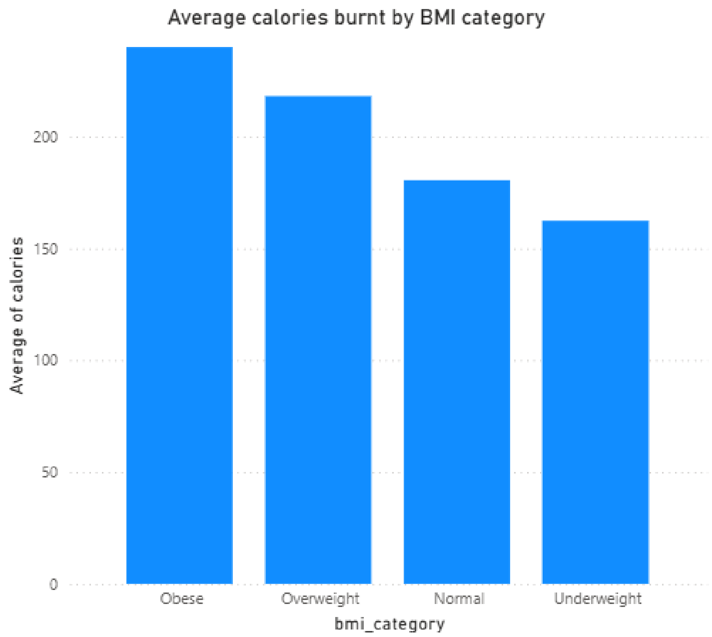
Variable Pairs	Correlation Coefficient
Heart Rate & Calories	+0.48
Duration & Calories	+0.55
BMI & Calories	+0.06
Intensity Index & Calories	+0.52

Figure 5.12: Average calories burnt (Age and Gender)



Age	Calories Burnt Male	Calories Burnt Female
15-24	194.42	192.42
25-34	193.70	191.28
35-44	219.60	215.57
45-54	235.65	245.69
55-64	217.02	196.80
65-74	109.27	93.69
75-80	113.17	99.53

Figure 5.13: Average calories burnt by BMI category



BMI Category	Calories burnt
Underweight	162.41
Normal	180.46
Overweight	218.13
Obese	239.95
It can be observed that the calories burnt on average increases with increase in BMI	

Relation between Heart Rate and Duration

Figure 5.14: Calories Burnt by Heart rate

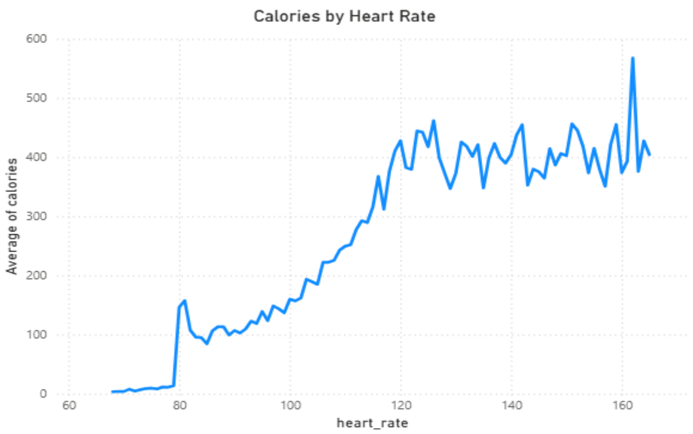


Figure 5.15: Calories burnt by Duration

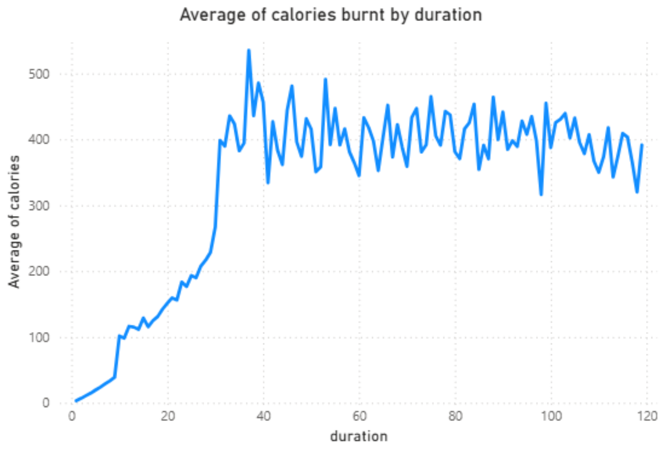
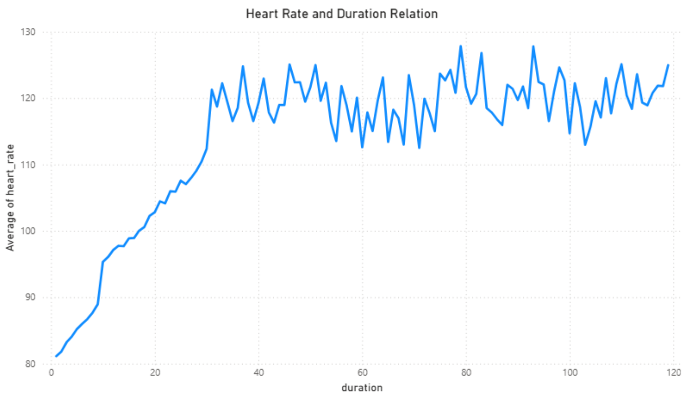


Figure 5.16: Heart Rate and Duration Relation

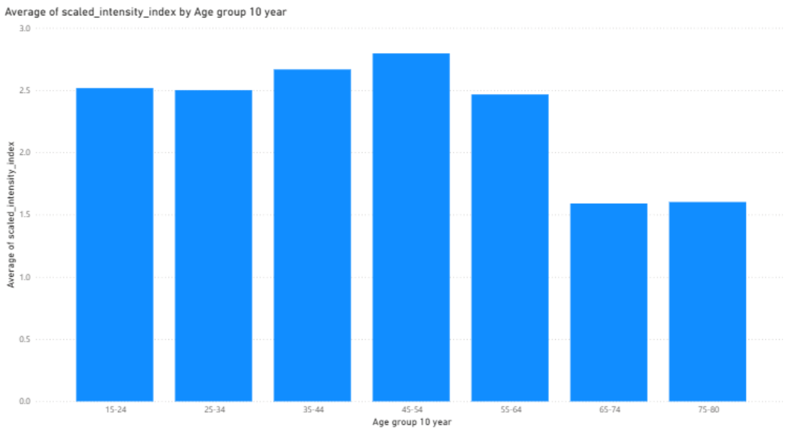


It can be concluded that increase in heart rate is directly correlated with increase in duration.

Thus, an increase in calories burnt will also be observed

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

Figure 5.17: Average Intensity of workout by age



Age	Average Intensity (Scaled)
15-24	2.52
25-34	2.50
35-44	2.67
45-54	2.79
55-64	2.47
65-74	1.59
75-80	1.69

Average Intensity: 2.31

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.

5.2 Inferential Analysis

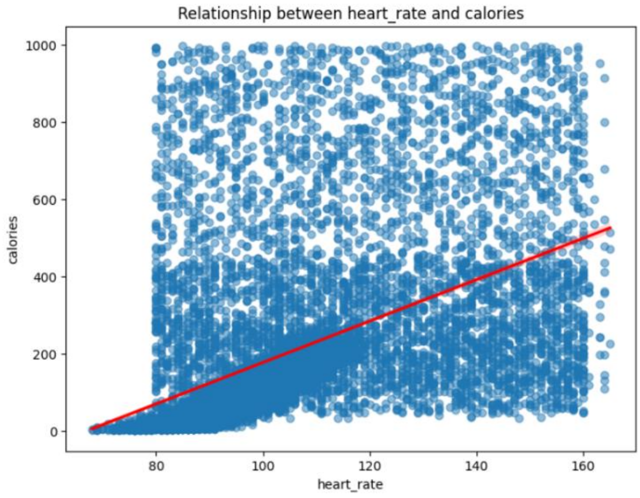
This section presents hypothesis testing and inferential statistics to validate patterns in the data.

5.2.1 Hypothesis Test: Does Heart rate affect calories burnt?

To determine whether **heart rate** has a statistically significant effect on **calories burned** during exercise using simple linear regression analysis.

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

Fig 5.18: Relationship between Heart Rate and Calories

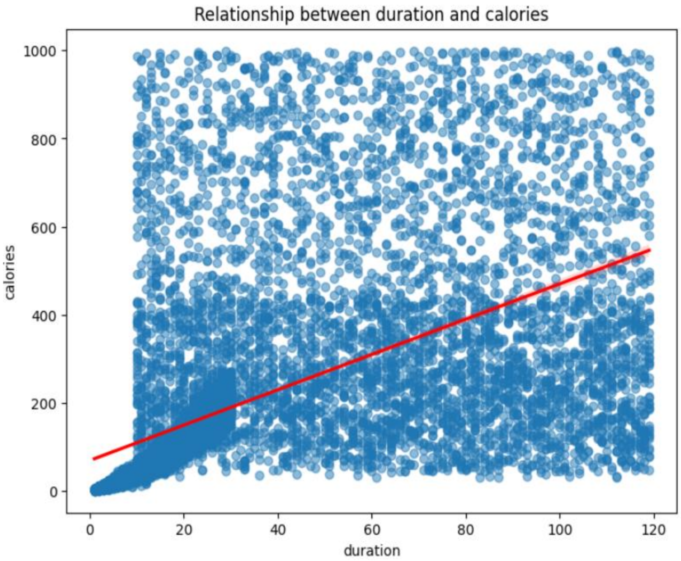


Metric	Value
Slope	5.37
p-value	0.00
R-squared	0.232
Significance Level α	0.05

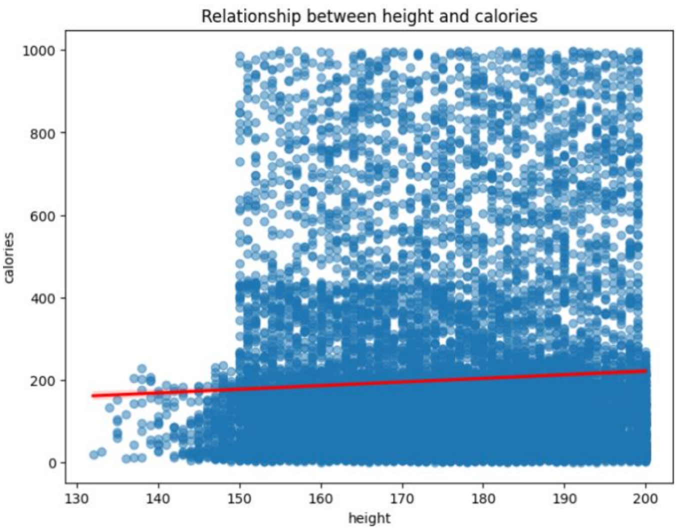
- Since the **p-value < 0.05**, we **reject the null hypothesis**.
- There is statistically significant evidence that **heart rate affects calories burned**.
- The relationship is **positive**, suggesting that higher heart rates are associated with greater calories burned.

5.2.2 Regression Significance

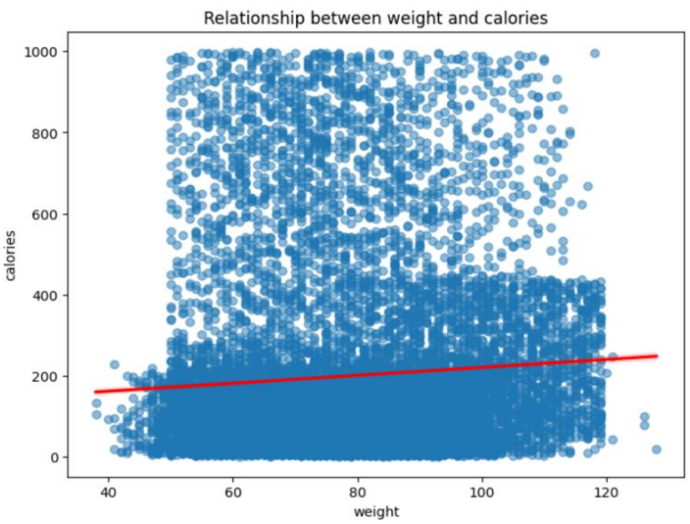
A simple linear regression of Calories on each parameters show:



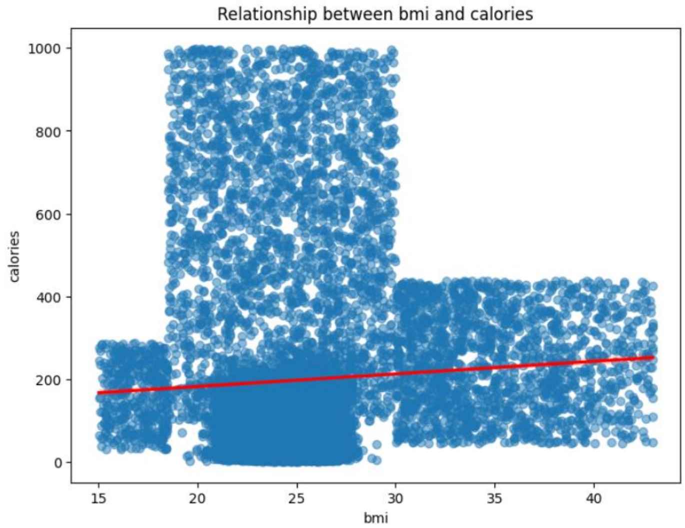
Feature: duration
y = 4.01 * x + 69.33
R-squared = 0.303



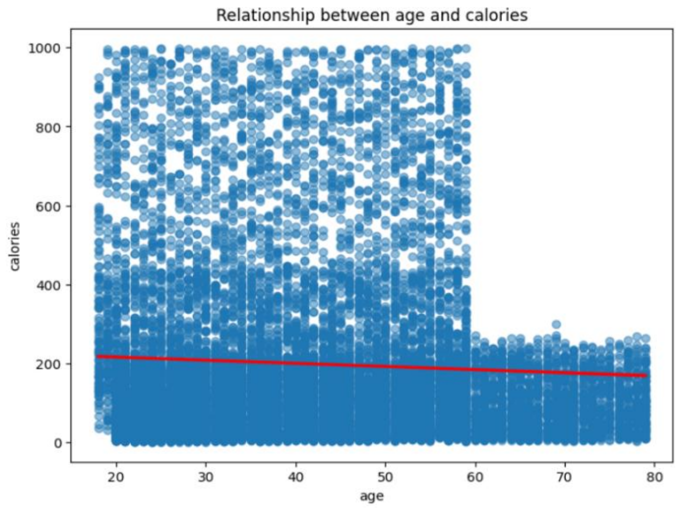
Feature: height
y = 0.88 * x + 45.31
R-squared = 0.003



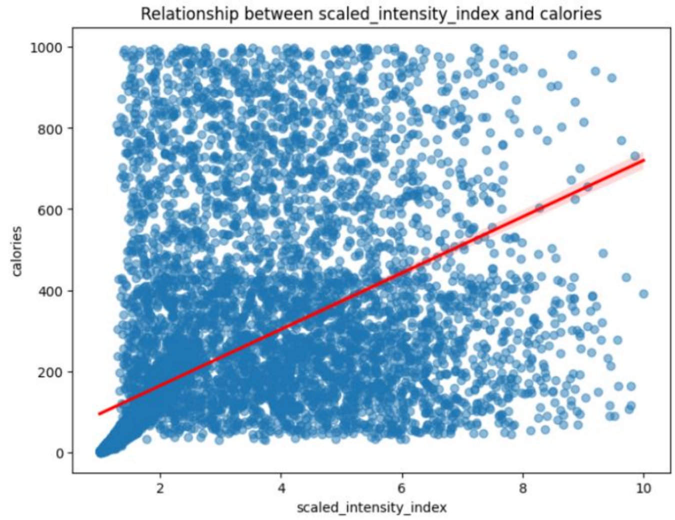
Feature: weight
y = 0.98 * x + 122.93
R-squared = 0.006



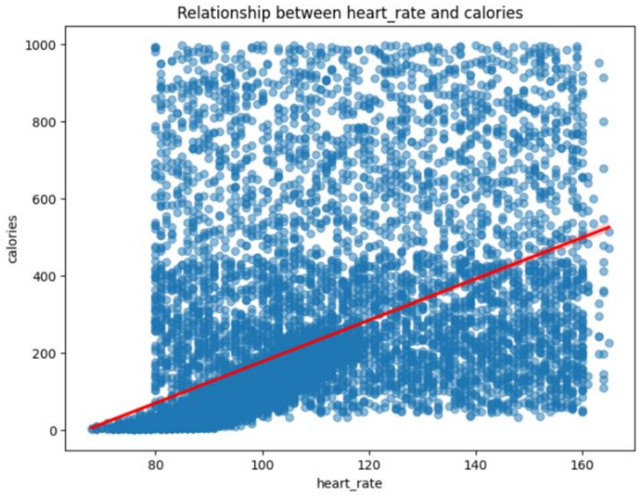
Feature: BMI
y = 3.04 * x + 121.81
R-squared = 0.004



Feature: age
y = -0.80 * x + 232.05
R-squared = 0.003



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



Feature: Heart Rate

y = 5.37 * x + -360.35
R-squared = 0.232

Parameter	Coefficient	t-statistic	p-value
Duration	4.00	77.41	<0.001
Heart Rate	5.37	71.24	<0.001
Weight	0.97	8.79	1.53
BMI	3.03	7.28	3.42
Height	0.88	6.49	1.01
Age	-0.79	-6.59	4.44
Scaled Intensity Index	69.40	71.24	<0.001

The positive coefficient indicates that:

- For every minute increase in duration it is associated with an average increase in 4kcal calories burnt.
- For every unit increase in Scaled Intensity it is associated with an average increase of 69.40kcal burnt.
- For every bpm rise it is associated with an increase of 5kcal burnt.

5.3 Machine Learning Model Performance

To predict calories burned, multiple models were trained and compared.

Table 5.2: Comparative Performance of Models

Models	**MAE**	**RMSE**	**R²**
Linear Regression *(simple)*	99.22	165.20	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	0.6180
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

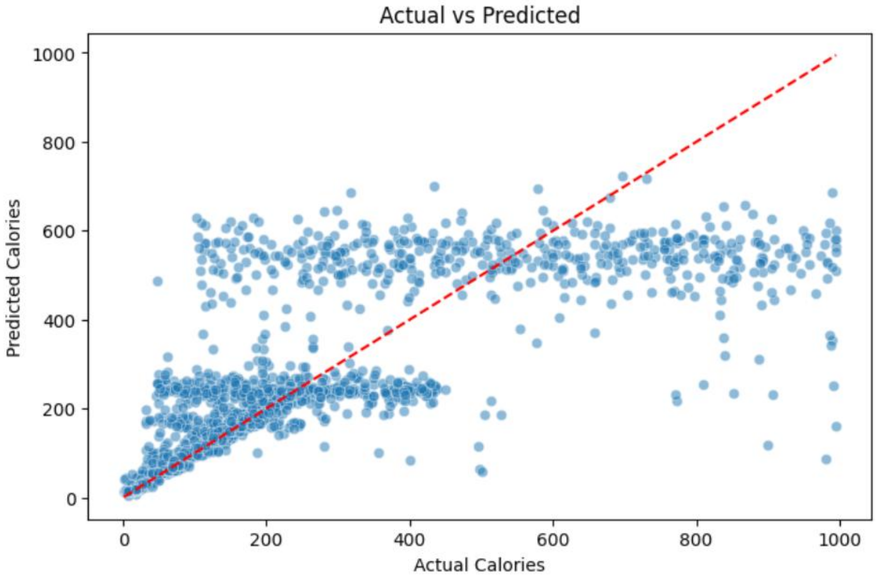
Table 5.3: Comparison of hyperparameters of models

Models	**Hyperparameters Type**	**Hyperparameters Values**
Random Forest	Default Hyperparameters	n_estimators=100, max_depth=5, random_state=42

Gradient Boosting	Default Hyperparameters	n_estimators=100, max_depth=5, random_state=42
	Tuned Hyperparameters	n_estimators=2000, learning_rate=0.01, max_depth=7, subsample=0.8, random_state=42
XGBoost	Default Hyperparameters	n_estimators=100, max_depth=5, random_state=42
	Tuned Hyperparameters	n_estimators=2000, learning_rate=0.05, max_depth=7, early_stopping_rounds=50, eval_metric='mae', random_state=42
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)
CatBoostRegressor	Tuned Hyperparameters:	iterations=2000, learning_rate=0.05, depth=7, early_stopping_rounds=50, verbose=100, random_state=42

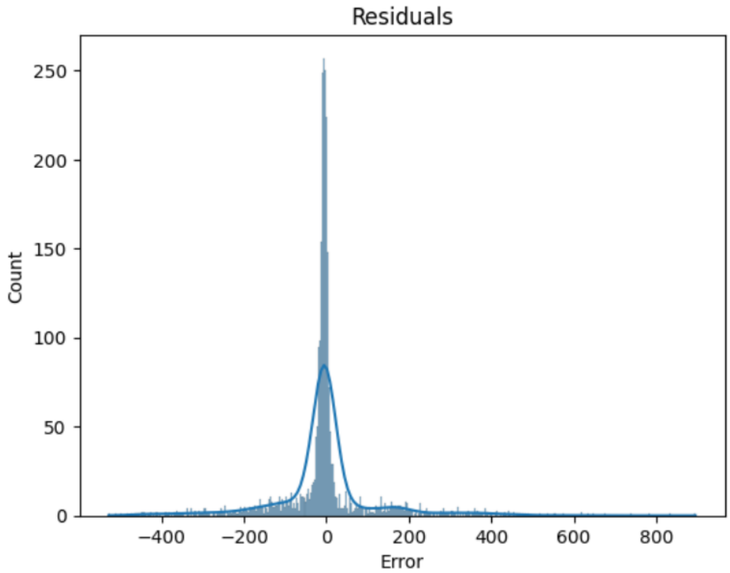
LightGBM achieved the best performance with an **R^2 of 0.6526**, indicating high predictive accuracy.

Figure 5.19: Predicted vs Actual Calories Plot



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

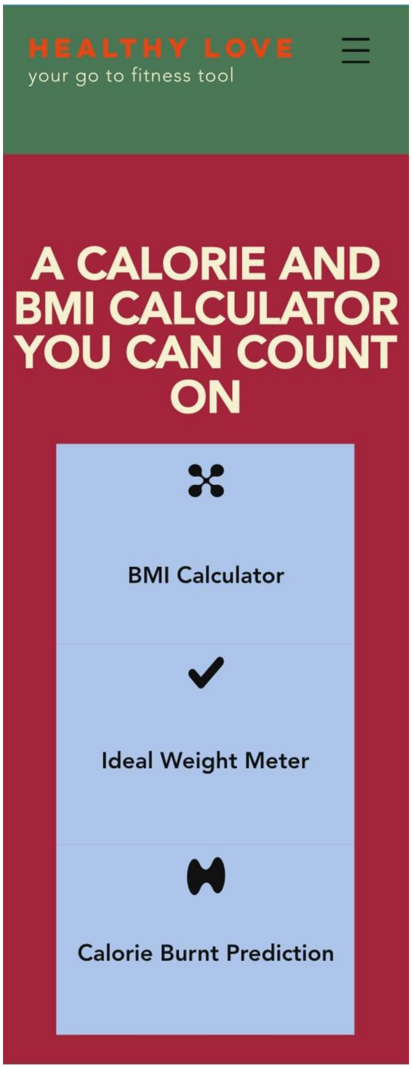
Figure 5.20: Residuals Plot

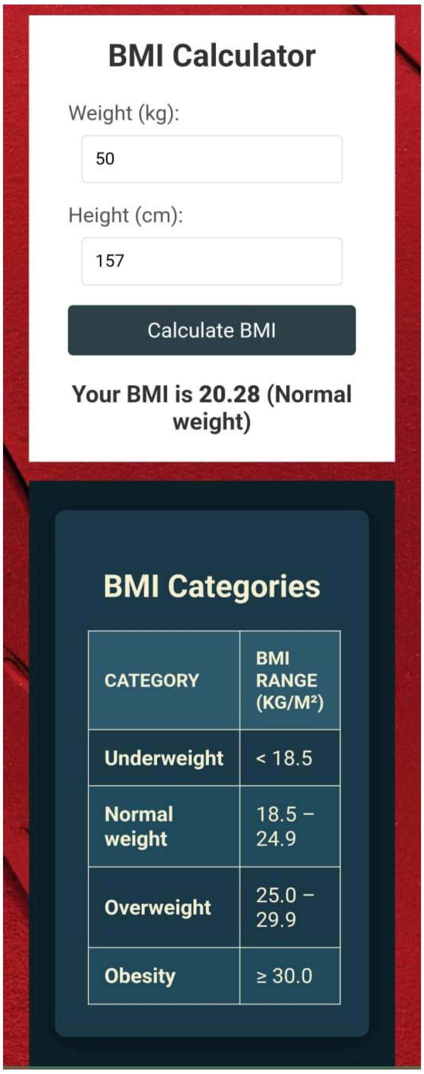


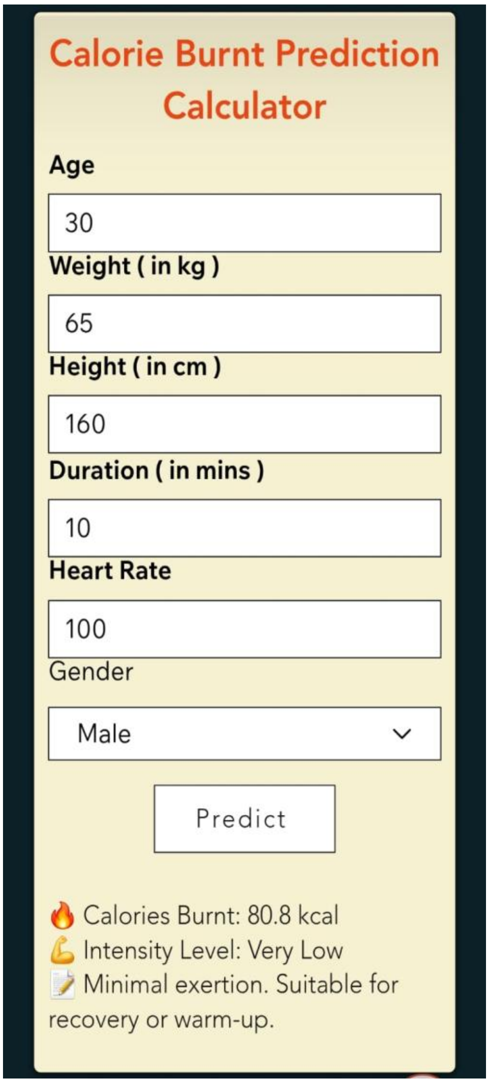
- The residuals are randomly scattered around zero, confirming no clear pattern and supporting model adequacy.

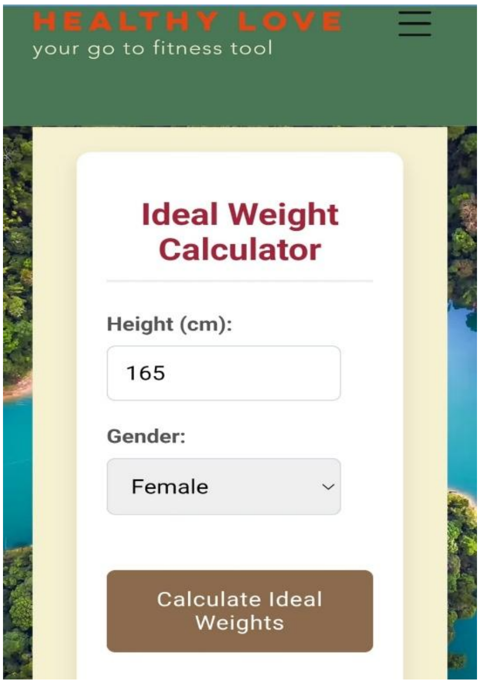
---

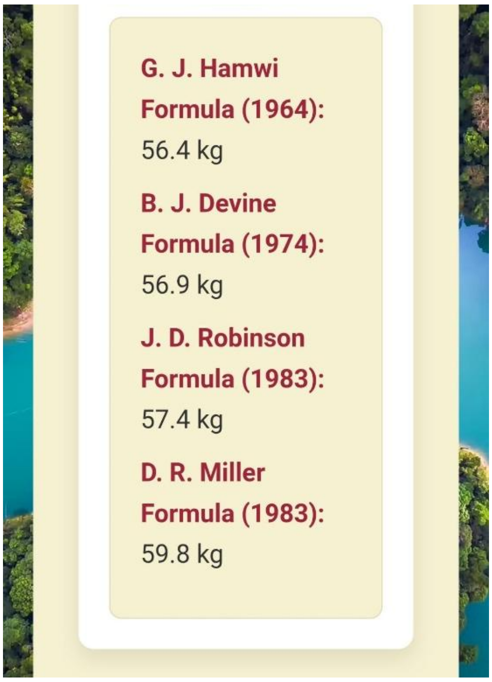
6. Website Overview: https://pdastdasroy.wixsite.com/healthylove











7.Conclusion

- Heart Rate and Duration are the strongest predictors of calories burned.
- Feature engineered Intensity Index had a strong correlation with calories burnt.
- Light Gradient Boost model outperform linear 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 exercises are more preferred than high intensity.
- Further model building can be done with consideration to training with activity type.

8.Project Limitations

As students who are new to data science and machine learning, these are some of the errors we found in our project.

- Both the datasets have no clear origin and can classify as synthetic.
- This model is not made for complete real world uses. It can be used to inspire future projects with real life proper sourced data.
- Linear scaling was only done on calories & not on heart rate which increases outlier potential.
- Model was trained using BMI too even with the presence of weight and height (Though BMI had a high feature importance in all our models for its output).
- The scaling of intensity index thus, the intensity levels were done solely based on our dataset and is subject to variation depending on the dataset used for training.
- The intensity index is an attempt to correlate calories to just one combined feature and can be subjective depending on the scenario (In our case O_2 values were not provided to calculate intensity based on MET).

9.Appendices

- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0022732
- https://www.sciencedirect.com/science/article/pii/S1568163716302719
- https://www.science.org/doi/10.1126/science.abe5017?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
- 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.urmc.rochester.edu/news/publications/health-matters/is-bmi-accurate
- https://www.everydayhealth.com/fitness/factors-that-can-affect-how-many-calories-you-burn
- 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**
- https://github.com/Pratyasha-Tapaja/calorie_prediction_project