Predict Calorie Expenditure

yixu liu

May 2025

1 Dataset Description

The competition data are provided in three files, all synthetically generated from a deep learning model trained on the original Calories Burnt Prediction dataset. Feature distributions closely mirror the originals but include slight shifts to encourage exploration of distributional differences.

1.1 Files and Dimensions

- **train.csv:**
  - $N_{\text{train}} = 750{,}000$ samples, 9 columns.
  - Memory usage: ≈ 51.5 MB.
  - Data types:
    - `int64`: `id`, `Age`
    - `float64`: `Height`, `Weight`, `Duration`, `Heart_Rate`, `Body_Temp`, `Calories`
    - `object`: `Sex`
- **test.csv:**
  - N_{test} samples, same columns as `train.csv` except for `Calories`.
- **sample_submission.csv:** A template with two columns: `id` and `Calories`.

Column	Description
`id`	Unique sample identifier
`Sex`	Participant sex (male/female)
`Age`	Participant age (years)
`Height`	Participant height (cm)
`Weight`	Participant weight (kg)
`Duration`	Workout duration (minutes)
`Heart_Rate`	Avg. heart rate during exercise (bpm)
`Body_Temp`	Avg. body temperature during exercise (°C)
`Calories`	Target: calories burned (continuous)

Table 1: Columns in `train.csv`.

	count	mean	std	min	25%	50%	75%	max
Age	750000	41.42	15.18	20	28	40	52	79
Height	750000	174.70	12.82	126	164	174	185	222
Weight	750000	75.15	13.98	36	63	74	87	132
Duration	750000	15.42	8.35	1	8	15	23	30
Heart_Rate	750000	95.48	9.45	67	88	95	103	128
Body_Temp	750000	40.04	0.78	37.10	39.60	40.30	40.70	41.50
Calories	750000	88.28	62.40	1	34	77	136	314

Table 2: Descriptive statistics for numeric features in `train.csv`.

Sex	Count
female	375 721
male	374 279

Table 3: Counts by sex in `train.csv`.

id	Sex	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
0	male	36	189.0	82.0	26.0	101.0	41.0	150.0
1	female	64	163.0	60.0	8.0	85.0	39.7	34.0
2	female	51	161.0	64.0	7.0	84.0	39.8	29.0
3	male	20	192.0	90.0	25.0	105.0	40.7	140.0
4	female	38	166.0	61.0	25.0	102.0	40.6	146.0

Table 4: First five entries from `train.csv`.

1.2 Competition Overview

- **Goal:** Predict calories burned during a workout.
- **Timeline:** Launched 25 days ago; closes in 6 days.
- **Evaluation:** Root Mean Squared Logarithmic Error (RMSLE):

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(\ln(\hat{y}_i + 1) - \ln(y_i + 1)\big)^2},$$

where n is the number of test observations, $\hat{y}_i$ the prediction, and y_i the true value.

2 Baseline Model

As a first-pass benchmark, I implemented a closed-form physiological equation—often referred to as the Keytel formula—to estimate calories burned per minute, then multiplied by workout duration to get total calories. This formula is widely used in exercise physiology and provides a reasonable "off-the-shelf" estimate without any learned parameters.

$$\text{kcal/min} = \begin{cases} \dfrac{-55.0969 + 0.6309\,\text{HR} + 0.1988\,W + 0.2017\,A}{4.184}, & \text{if Sex} = \text{male}, \\ \dfrac{-20.4022 + 0.4472\,\text{HR} - 0.1263\,W + 0.074\,A}{4.184}, & \text{if Sex} = \text{female}, \end{cases}$$

where

- HR is average heart rate (beats per minute),
- W is weight in kg,
- A is age in years.

a public RMSLE score of

$$\text{RMSLE}_{\text{public}} = 0.32672.$$

This establishes our baseline against which any more sophisticated ML models must improve.

3 Exploratory Data Analysis and Feature Engineering

3.1 Univariate Exploration

To gain an initial understanding of each feature, I performed three simple steps:

1. **Histograms.** I plotted each variable's distribution with 30 bins. The histograms revealed:
   - *Age* is right-skewed, with most people in their 20s–40s and fewer in older age brackets.
   - *Height* and *Weight* are approximately bell-shaped, suggesting a roughly normal spread around typical adult values.
   - *Duration* is nearly uniform over the 1–30 minute range, reflecting that exercise sessions are evenly distributed.
   - *Heart_Rate* clusters around 90–100 bpm, with a slight left tail (resting values) and right tail (high exertion).

- *Body_ Temp* shows a narrow peak near 40 °C but is right-skewed, indicating a few unusually high temperature readings.
- *Calories* burned is strongly right-skewed: many low-calorie sessions and a long tail of high-calorie outliers.

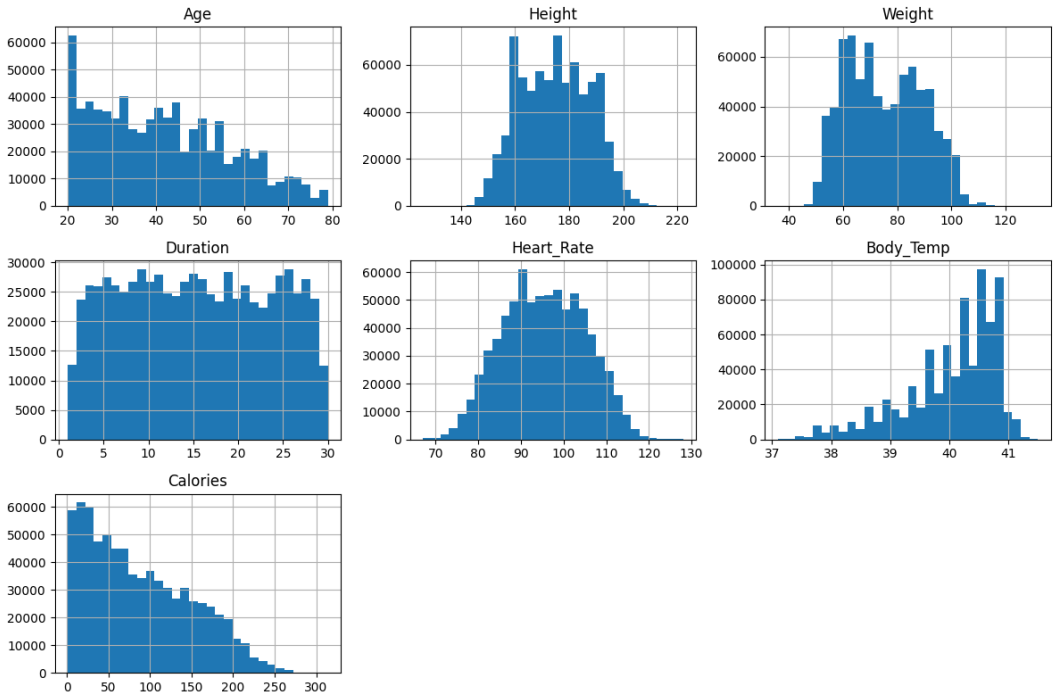


Figure 1: Histograms of all numerical features (30 bins each).

2. **Boxplots.** To spotlight potential outliers, I displayed boxplots for each variable. I found:
   - A handful of extreme *Calories* values above 250 kcal, likely very long or intense sessions.
   - Some high *Heart_ Rate* readings exceeding 120 bpm, which I will review for possible capping.
   - Occasional *Body_ Temp* points above 41 °C that may represent measurement errors.

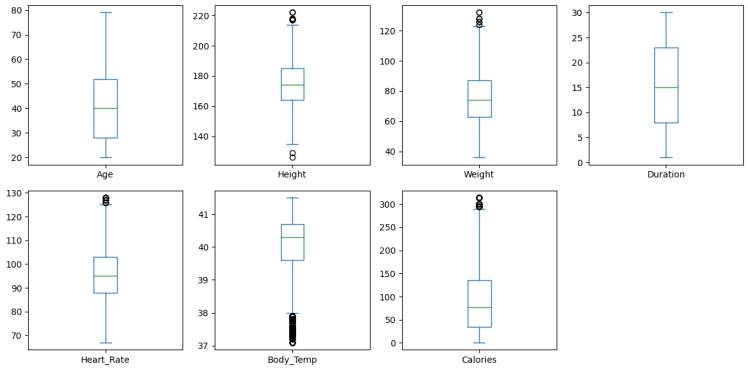


Figure 2: Boxplots of all numerical features, highlighting potential outliers.

Why I did this. Univariate exploration helps me detect unusual values, understand each feature's scale and distributional shape, and form hypotheses about transformations (e.g. log-scaling Calories) or outlier handling before moving on to modeling.
Key takeaways.

- Most features (Height, Weight, Heart_Rate) lie in a fairly symmetric range and can be used directly.
- Right-skewed variables (Age, Calories, Body_Temp) may benefit from transformation or outlier treatment.
- Uniform Duration suggests no obvious bias in session lengths.

3.2 Bivariate Exploration

Building on our univariate insights, I next examined pairwise relationships to see which features jointly explain Calories burned.

3.2.1 Correlation Heatmap

Figure 3 shows the Pearson correlation coefficients between all numeric features.

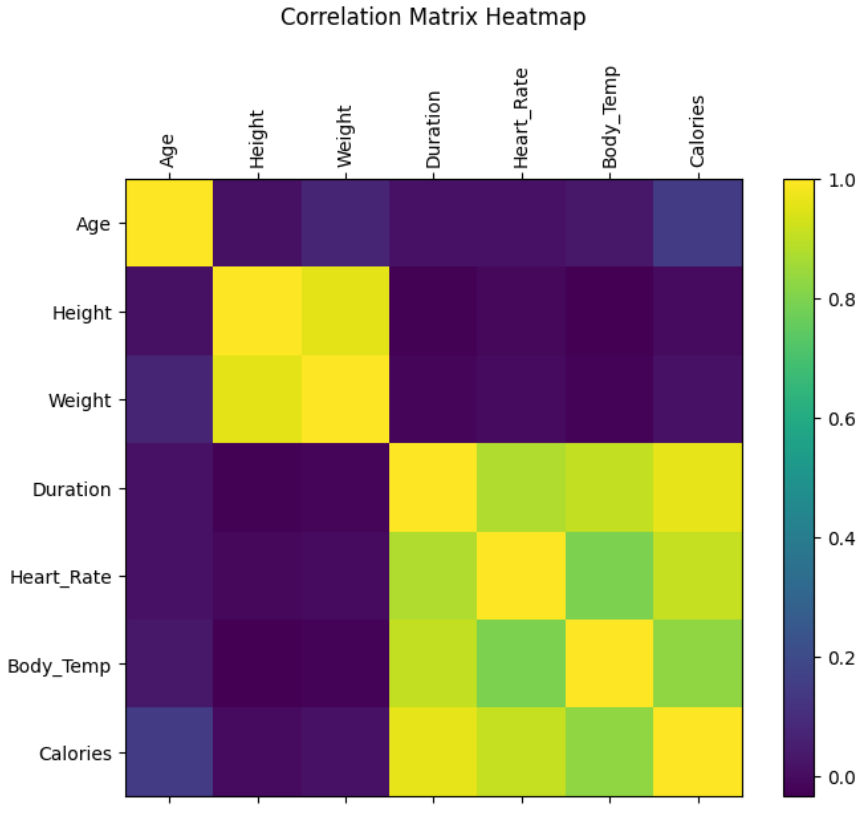


Figure 3: Correlation matrix heatmap of features.

Table 5 summarizes the most important correlations and their interpretations.

Pair	Corr. coeff.	Interpretation
Duration ↔ Calories	≈ 0.98	Nearly perfect linear relationship—longer workouts burn more.
Heart_Rate ↔ Calories	≈ 0.82	Higher average heart rates go hand-in-hand with more calories.
Body_Temp ↔ Calories	≈ 0.88	Warmer body temps correlate strongly with exertion.
Duration ↔ Heart_Rate	≈ 0.92	Longer sessions tend to raise average heart rate.
Duration ↔ Body_Temp	≈ 0.94	Longer workouts also heat you up more.
Weight ↔ Height	≈ 0.90	Heavier individuals in this dataset are generally taller.
Age with any other feature	< 0.20	Almost no linear effect of age on burn or vitals here.

Table 5: Key pairwise correlations and interpretations.

Take-away. Duration, Heart_Rate, and Body_Temp are all tightly linked to each other and to Calories. Weight and Height correlate strongly with each other but have little direct effect on Calories. Age shows negligible linear relationships.

3.2.2 Scatter-Matrix ("Pair-Plot")

I also generated a scatter-matrix (Figure 4) to visualize joint distributions.
Notable observations:

- *Duration vs. Calories:* a razor-thin upward band, confirming a near-perfect linear trend.
- *Heart_Rate vs. Calories:* clear positive slope, with more scatter than duration.
- *Body_Temp vs. Calories:* clean upward curve, again reinforcing a strong relationship.
- *Weight/Height vs. Calories:* a flat cloud, indicating little direct effect.
- *Age vs. Calories:* no visible trend at all.

3.2.3 Next Steps: 3D Scatterplots

Given the very high correlations among Duration, Heart_Rate, Body_Temp, and Calories, the most informative 3D plots to create are:

1. Duration (x) vs. Heart_Rate (y) vs. Calories (z)
2. Duration (x) vs. Body_Temp (y) vs. Calories (z)
3. Heart_Rate (x) vs. Body_Temp (y) vs. Calories (z)

These will allow me to inspect any nonlinear interaction effects before building our predictive models.

3.3 3D Scatterplots

To inspect potential interaction effects among the top predictors, I plotted three 3D scatterplots:
(a) Duration × Heart Rate → Calories (Figure 5a):

- *Rising "shelf":* Calories increase nearly linearly as both Duration and Heart Rate climb.

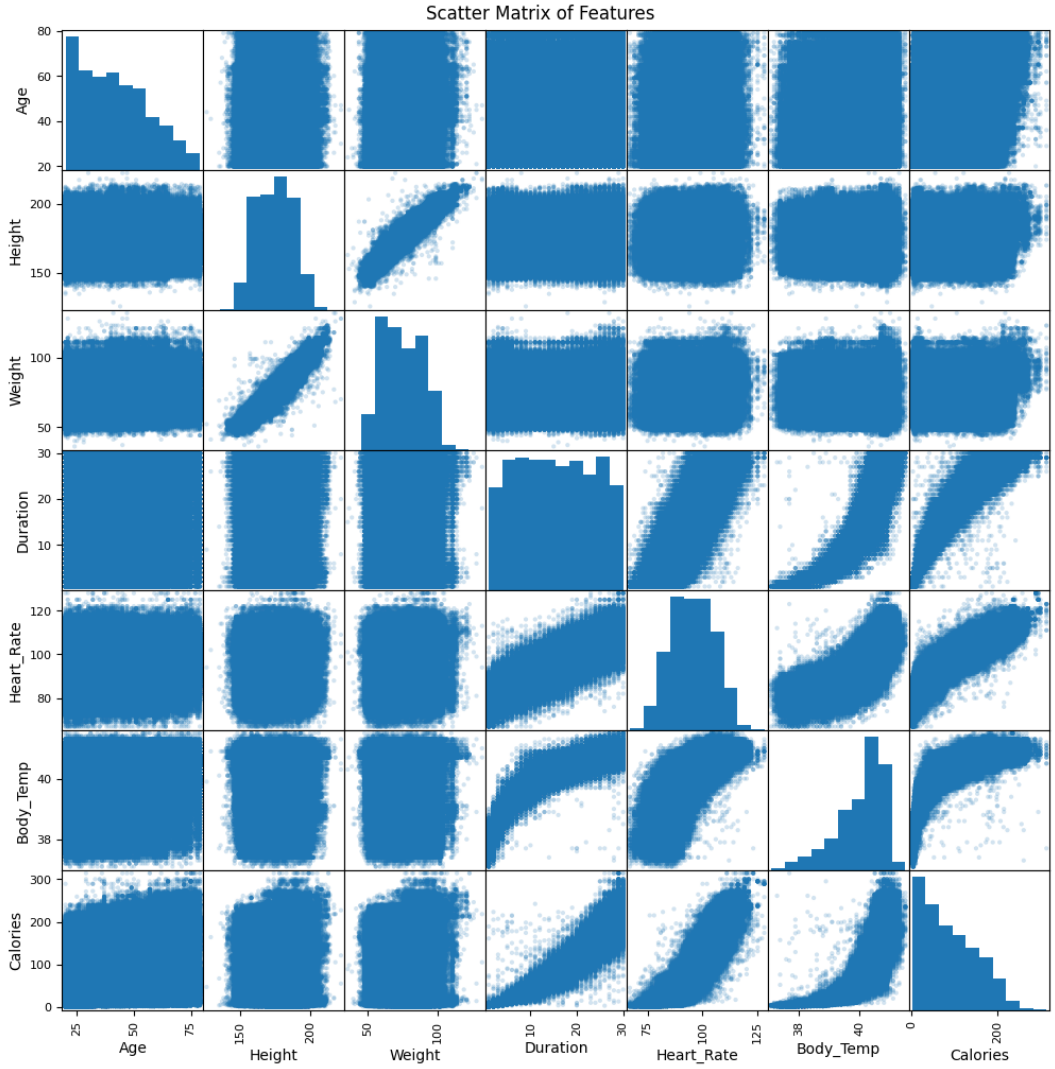
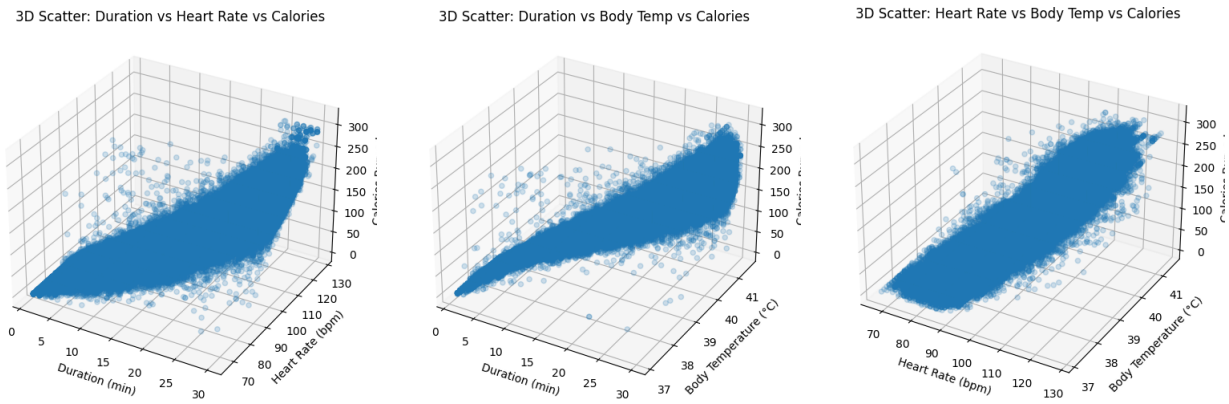


Figure 4: Scatter-matrix of all feature pairs (diagonals show univariate histograms).

- *Low-duration plateau:* Sessions under about 5 min burn little even at high heart rates, showing duration is a gating factor.
- *Spread at long durations:* Beyond ~ 20 min, more scatter appears—long workouts vary more in exertion.
- *Take-away:* A simple interaction term Duration $\times$ Heart_Rate could capture most of this signal.

(b) Duration $\times$ Body Temp $\rightarrow$ Calories (Figure 5b):

- *Similar shape:* Longer sessions heat the body more, and hotter bodies burn more.
- *Tighter band:* Body temperature is less noisy than heart rate, producing a thinner "sheet."
- *Low-temp floor:* Below $\sim 37.5C$ almost no calories are burned, confirming temperature as an exertion proxy.



(a) Duration vs. Heart Rate vs. Calories

(b) Duration vs. Body Temp vs. Calories

(c) Heart Rate vs. Body Temp vs. Calories

Figure 5: Three-dimensional views of Calories burned against pairs of top predictors.

- *Take-away:* Try a feature like Duration $\times$ (Body_Temp $-$ 37) to encode length and intensity.

(c) Heart Rate $\times$ Body Temp $\rightarrow$ Calories (Figure 5c):

- *Diagonal ridge:* As both vitals rise together, calories increase along a narrow elbow.
- *No flat zone:* Burn stays near zero only when both HR and Temp remain low.
- *Take-away:* Heart rate and temperature carry largely redundant information. Consider using only one or combining them (e.g. an average z-score).

Overall Insights & Next Steps

- **Duration is the master switch:** Without sufficient time, even high intensity yields little burn.
- **Body temperature edges out heart rate:** It is a cleaner, less noisy proxy for exertion.
- **Almost all variation in Calories** lies on the 3D "sheet" formed by (Duration, Intensity) $\rightarrow$ Burn.
- **Actionable features to engineer:**
  - Interaction: Duration $\times$ Body_Temp
  - Normalized intensity: $\frac{\text{Body\_Temp}-37}{\text{Duration}}$
  - Threshold flags: e.g. $\mathbf{1}\{\text{Temp} > 38\}$ or $\mathbf{1}\{\text{HR} > 100\}$

3.4 Derived-Feature Ideas

Based on our 3D exploration, I engineered several candidate features to better capture workout intensity and its effect on calorie burn.

3.4.1 Normalized Intensity

I define a resting-temperature baseline and compute a per-minute intensity score:

$$\text{baseline} = 37.0, \quad \text{norm\_intensity} = \frac{\text{Body\_Temp} - \text{baseline}}{\text{Duration}}$$

with zero-duration sessions set to zero. This yields:

$$\text{corr}(\text{norm\_intensity}, \text{Calories}) \approx 0.88.$$

3.4.2 Flipped Ratio

I also tried the inverse ratio—

$$\text{dur\_over\_delta} = \frac{\text{Duration}}{\text{Body\_Temp} - \text{baseline}},$$

again filling infinities/NaNs with zero. Its correlation with Calories is:

$$\text{corr}(\text{dur\_over\_delta}, \text{Calories}) \approx 0.75.$$

3.4.3 Raw Temperature Rise

Finally, the simple temperature rise above baseline:

$$\delta_{\text{temp}} = \text{Body\_Temp} - \text{baseline},$$

has

$$\text{corr}(\delta_{\text{temp}}, \text{Calories}) \approx 0.88.$$

3.4.4 Threshold Flags

I examined binary splits to flag "high-intensity" sessions:

- **Temp_High:** $\mathbf{1}\{\text{Body\_Temp} > 38°\text{C}\}$
  - Mean Calories when 0: 4.7 kcal (2.5% of sessions)
  - Mean Calories when 1: 90.4 kcal (97.5% of sessions)
  - *Take-away:* Extremely imbalanced; mostly flags "no exercise" vs. "exercise."
- **HR_High:** $\mathbf{1}\{\text{Heart\_Rate} > 100\,\text{bpm}\}$
  - Mean Calories when 0: 53.5 kcal (67% of sessions)
  - Mean Calories when 1: 158.6 kcal (33% of sessions)
  - *Take-away:* A balanced split with a ~ 105 kcal lift—useful as a binary indicator.

3.4.5 K-Means Heart-Rate Zones

To obtain a more nuanced HR categorization, I applied K-Means (K=4) to Heart_Rate, yielding centroids at approximately [82, 91, 99, 108] bpm. Mean calorie burns by zone are summarized in Table 6.

Take-away: Treat `hr_zone_km` as a categorical feature (one-hot encode) so the model can learn separate intercepts per zone.

Zone	Centroid (bpm)	# Sessions	Mean Calories
0	82.0	173 k	23.0 kcal
1	91.0	178 k	32.7 kcal
2	99.5	163 k	43.3 kcal
3	108.2	133 k	53.9 kcal

Table 6: Heart-rate zones derived via K-Means (K=4) and their corresponding mean calorie burns.

3.4.6 Age-Based Duration Zones

Clustering `Duration` into five K-Means bins yields the following ranges, session counts, and mean calorie burns:

Zone	Duration Range (min)	# Sessions	Mean Calories
0	20–27	172,080	23.04 kcal
1	28–38	177,636	32.73 kcal
2	39–48	162,560	43.32 kcal
3	49–60	133,333	53.98 kcal
4	61–79	104,391	67.52 kcal

Table 7: Duration zones derived via K-Means (K=5) and their corresponding session counts and mean calorie burns.

Encode this as a one-hot categorical feature (`duration_km_bin`) in subsequent models.

3.5 Key Findings Overview

`Feature`	How It's Defined	Corr / Impact
`dur_temp`	Duration × Body_Temp	**0.9609**
`dur_over_delta`	Duration / (Body_Temp − 37)	**0.9358**
`delta_temp`	Body_Temp − 37	**0.8287**
`temp_high`	Body_Temp > 38°C (binary)	+85 kcal; 97.5%
`hr_high`	Heart_Rate > 100 bpm (binary)	+105 kcal; 33%
`hr_zone_km`	4-cluster K-Means on Heart_Rate	[23, 62, 127, 187] kcal
`BMI`	Weight / (Height/100)^2	0.05
`age_decade`	Age binned by decade (20s, 30s, ...)	78 → 112 kcal
`age_km_bin`	5-cluster K-Means on Age (20–27, ..., 62–79)	–

Table 8: Key engineered features, their definitions, and correlation/impact on Calories burned.

3.6 Multicollinearity Assessment

Variance Inflation Factor (VIF) quantifies how much a given feature X_j is linearly explained by the other features. If it regress X_j on all other predictors and obtain R_j^2, then

$$\mathrm{VIF}_j \;=\; \frac{1}{1 - R_j^2}\,.$$

- VIF = 1: no linear correlation with other features.
- $1 <$ VIF < 5: acceptable collinearity.
- VIF > 5: increasing multicollinearity concern.

Table 9: Variance Inflation Factors for Predictor Variables

Feature	**VIF**	**Interpretation**
BMI	1.04	negligible collinearity
age_1	1.56	low collinearity
age_2	1.55	low collinearity
age_3	1.49	low collinearity
age_4	1.43	low collinearity
hr_1	2.52	moderate collinearity
hr_2	3.97	moderate collinearity
hr_3	5.49	approaching concern
delta_temp	8.24	borderline high
dur_over_delta	31.61	very high
dur_temp	54.82	very high
const	344.33	intercept (ignore)

The predictor matrix X was assembled by combining both engineered continuous features and categorical cluster indicators, as follows:

- **dur_temp**: product of exercise `Duration` and `Body_Temp`, capturing total thermal load.
- $\boldsymbol{\delta}_{\text{temp}}$: deviation of `Body_Temp` from normal (37 °C), isolating temperature effects.
- **dur_over_delta**: ratio $\frac{\text{Duration}}{\delta_{\text{temp}}}$, reflecting rate of temperature change.
- **hr_zone_km** dummies: heart–rate zones (4 clusters via K-Means), one-hot encoded (drop first).
- **age_km_bin** dummies: age bins (5 clusters via K-Means), one-hot encoded (drop first).

Key Takeaways

- High VIFs for `dur_temp` and `dur_over_delta` reflect that both combine `Duration` and `Body_Temp` in nearly linear ways.
- δ_temp (VIF$\approx$ 8.2) is borderline but still under the usual cutoff of 10.
- Age and BMI features show minimal collinearity.
- HR-zone dummies have moderate collinearity, as expected with one-hot encoding.

4 Random Forest Training

4.1 Feature Engineering

Based on our exploratory analysis, I constructed the following features:

$$\begin{aligned} \text{dur\_temp} &= \text{Duration} \times \text{Body\_Temp}, \\ \delta_{\text{temp}} &= \text{Body\_Temp} - 37, \\ \text{dur\_over\_delta} &= \frac{\text{Duration}}{\delta_{\text{temp}}} \quad \text{(with zeros replaced by NaN)}. \end{aligned}$$

I also applied K-Means clustering to capture non-linear effects:

- **Heart-Rate zones:** 4 clusters on `Heart_Rate`, reordered by centroid.
- **Age bins:** 5 clusters on `Age`, reordered by centroid.

4.2 Preprocessing Pipeline

I encapsulated preprocessing and modeling in a single `Pipeline` to avoid data leakage:

1. **One-hot encode** categorical cluster features (`hr_zone_km`, `age_km_bin`), dropping the first level to keep full rank.
2. **Pass through** the continuous features $\{\text{dur\_temp}, \text{dur\_over\_delta}, \delta_{\text{temp}}\}$.

4.3 Random Forest Configuration

I chose a `RandomForestRegressor` with:

- `n_estimators` $= 200$: enough trees to stabilize predictions.
- `max_depth` $= 10$: limits overfitting on noise.
- `random_state` $= 42$: ensures reproducibility.
- `n_jobs` $= -1$: uses all CPU cores for training.

4.4 Evaluation

I ran 5-fold cross-validation using RMSLE:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(\ln(y_i+1) - \ln(\hat{y}_i+1)\big)^2}\,.$$

CV result:

$$\text{RF RMSLE} = 0.1224 \pm 0.0008\,.$$