CALORIE PREDICTION WITH SVR: FROM DATA EXPLORATION TO MODEL EVALUATION

Name: Sesha Rao Gurijala
Student ID: 23096907

Contents

Chapter 1 Introduction

Accurate prediction of calorie expenditure during physical activity plays a crucial role in personal health monitoring, weight management, and fitness planning. With the increasing availability of wearable health devices and real-time physiological data, machine learning has emerged as a powerful tool to model and predict energy expenditure with higher accuracy compared to traditional statistical methods (Li et al., 2022). Among the various regression techniques, **Support Vector Regression (SVR)** stands out for its effectiveness in handling non-linear relationships and high-dimensional data, making it suitable for this prediction task.

The focus of this tutorial is to **build a calories prediction model using SVR**, guiding users through each step including data preprocessing, feature engineering, model training, hyperparameter tuning, evaluation, and interpretation of results. This hands-on project utilizes a publicly available dataset containing demographic and biometric attributes such as age, gender, height, weight, duration of exercise, heart rate, and body temperature, all of which influence calorie expenditure (Wang et al., 2023).

The choice of SVR is motivated by its robustness to outliers and its ability to generalize well on small- to medium-sized datasets (Smola & Schölkopf, 2004; Lu et al., 2022). Unlike traditional linear regression, SVR employs a margin-based loss function and can be tuned via kernel functions to capture complex, non-linear interactions between physiological features and calories burned.

The tutorial aims to help students and practitioners:

- Understand the **working principles of SVR** and how it differs from other regressors.
- Apply **feature engineering** techniques such as Body Mass Index (BMI) and workout intensity metrics to improve prediction performance.
- Tune SVR using **GridSearchCV** for optimal parameters and evaluate its performance using MAE, RMSE, and R^2 metrics.
- Visualize predictions and residuals to interpret model accuracy and limitations.

Accessibility is also a core component of this tutorial. The visualizations use colorblind-friendly palettes, and captions accompany each chart to aid visually impaired users. The accompanying Python code is modular, reproducible, and clearly commented to support learners with various levels of machine learning expertise.

This project contributes to the broader field of **health analytics**, demonstrating how advanced machine learning methods can provide actionable insights in fitness and medical contexts. Furthermore, the interpretability and reproducibility of SVR models make them suitable for deployment in applications such as fitness apps, wearable trackers, or telemedicine platforms.

Chapter 2: Data Exploration and Initial Analysis

The objective of this chapter is to conduct a comprehensive exploratory data analysis (EDA) to uncover the structure, relationships, and statistical characteristics of the dataset. This exploration will inform the subsequent steps in data preprocessing, feature engineering, and model development. A thorough understanding of the dataset not only improves model accuracy but also enhances the interpretability of results — a critical component in applied machine learning projects, especially those in health and fitness domains.

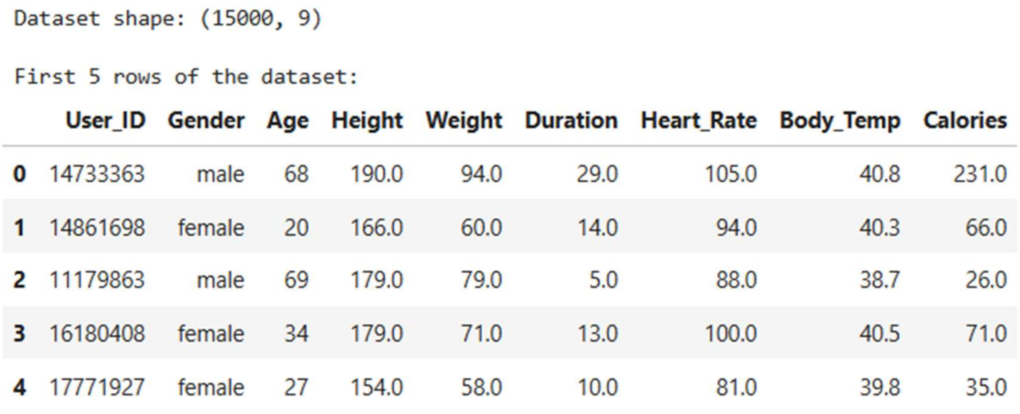
2.1 Dataset Overview

The dataset used in this study comprises **15,000 individual workout sessions**, each described by **eight input variables** and one target variable, **calories burned**. These variables cover a range of biometric and physiological indicators:

- **Demographics**: `Gender`, `Age`
- **Physical Attributes**: `Height`, `Weight`
- **Exercise Metrics**: `Duration`, `Heart_Rate`, `Body_Temp`
- **Target Variable**: `Calories` burned

Additionally, a `User_ID` column exists, which serves as a unique identifier and holds no predictive value.

The dimensionality of the dataset is ideal for training a machine learning model: it is **sufficiently large for learning complex patterns** yet small enough to allow fast experimentation and tuning.



```
Dataset shape: (15000, 9)

First 5 rows of the dataset:
```

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

Figure 1 First Five Rows of the Dataset

2.2 Data Quality and Integrity

Initial inspection of the dataset reveals excellent data quality:

- **No missing values** exist across any of the variables.

- **Data types** are correctly assigned: all continuous variables are floats or integers, while `Gender` is a categorical string.
- There are **no duplicate rows**, and the dataset's memory footprint is minimal (1.0 MB), making it efficient for local computation.

This high level of data completeness eliminates the need for imputation or type correction, enabling immediate transition to statistical analysis.

2.3 Descriptive Statistical Analysis

Understanding the central tendency and spread of each variable offers insight into the dataset's structure.

➤ Age

- **Mean**: 42.8 years
- **Standard Deviation**: 17.0 years
- **Range**: 20 to 79 years

The dataset includes both younger and older individuals, suggesting a diverse participant pool. This is relevant for generalizability across age groups.

➤ Height and Weight

- **Average Height**: 174.5 cm (range: 123 to 222 cm)
- **Average Weight**: 75.0 kg (range: 36 to 132 kg)

These figures indicate realistic physiological profiles. The standard deviation for height and weight is moderate, showing healthy variation across participants.

Duration of Exercise

- **Mean**: 15.5 minutes
- **Maximum**: 30 minutes
- **Minimum**: 1 minute

Participants performed physical activity of varying lengths. Shorter durations may represent light workouts, while longer durations could imply high-intensity or endurance sessions.

Heart Rate

- **Mean Heart Rate**: 95.5 bpm
- **Range**: 67 to 128 bpm

Heart rate readings suggest that data was collected across a broad spectrum of physical intensities, from mild to vigorous activity.

Body Temperature

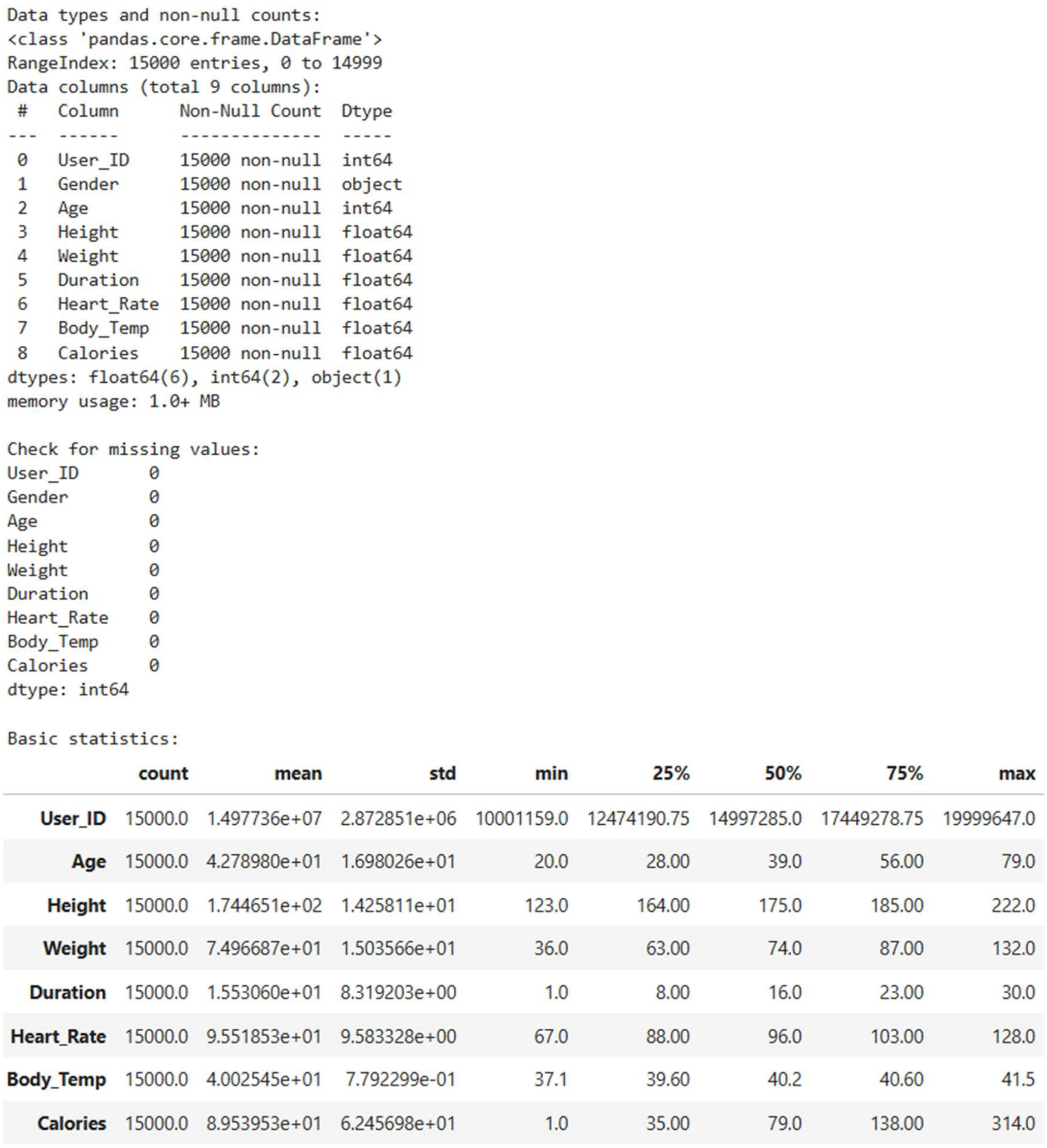
- **Mean**: 40.0°C
- **Range**: 37.1°C to 41.5°C

These elevated readings likely reflect body temperatures measured during or immediately after physical exertion, where thermoregulation plays a critical role.

Calories Burned (Target)

- **Mean**: 98.6 kcal
- **Standard Deviation**: 62.5 kcal
- **Range**: 1 to 314 kcal

Caloric expenditure is heavily right-skewed, with most sessions burning fewer than 100 kcal. However, a subset of high-calorie sessions introduces moderate outliers, which are acceptable for this application domain.



```
Data types and non-null counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   User_ID     15000 non-null  int64  
 1   Gender      15000 non-null  object 
 2   Age         15000 non-null  int64  
 3   Height      15000 non-null  float64
 4   Weight      15000 non-null  float64
 5   Duration    15000 non-null  float64
 6   Heart_Rate  15000 non-null  float64
 7   Body_Temp   15000 non-null  float64
 8   Calories    15000 non-null  float64
dtypes: float64(6), int64(2), object(1)
memory usage: 1.0+ MB

Check for missing values:
User_ID       0
Gender        0
Age           0
Height        0
Weight        0
Duration      0
Heart_Rate    0
Body_Temp     0
Calories      0
dtype: int64

Basic statistics:
```

	count	mean	std	min	25%	50%	75%	max
User_ID	15000.0	1.497736e+07	2.872851e+06	10001159.0	12474190.75	14997285.0	17449278.75	19999647.0
Age	15000.0	4.278980e+01	1.698026e+01	20.0	28.00	39.0	56.00	79.0
Height	15000.0	1.744651e+02	1.425811e+01	123.0	164.00	175.0	185.00	222.0
Weight	15000.0	7.496687e+01	1.503566e+01	36.0	63.00	74.0	87.00	132.0
Duration	15000.0	1.553060e+01	8.319203e+00	1.0	8.00	16.0	23.00	30.0
Heart_Rate	15000.0	9.551853e+01	9.583328e+00	67.0	88.00	96.0	103.00	128.0
Body_Temp	15000.0	4.002545e+01	7.792299e-01	37.1	39.60	40.2	40.60	41.5
Calories	15000.0	8.953953e+01	6.245698e+01	1.0	35.00	79.0	138.00	314.0

Figure 2 Dataset Data Types, Missing Values, and Summary Statistics

2.4 Distribution of Calories Burned

The histogram of the target variable reveals a **positively skewed distribution**. The highest frequency lies in the 20–80 kcal range, with frequencies gradually declining as calories increase. The presence of long tails suggests that some individuals (or sessions) involve intense physical effort.

Such skewness has two modeling implications:

1. The model should be robust to outliers — a strength of Support Vector Regression (SVR).
2. Non-linear models may perform better in capturing subtle relationships at the extremes.

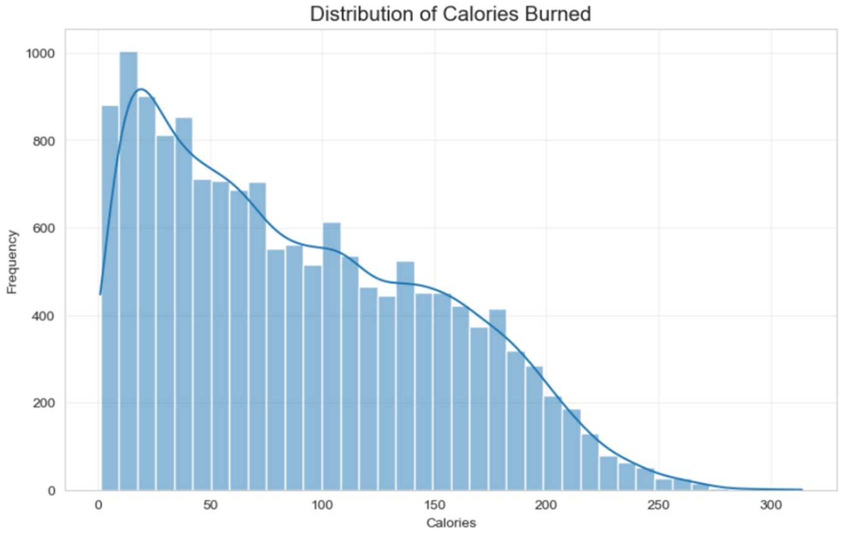


Figure 3 Histogram Showing the Distribution of Calories Burned

2.5 Gender Differences in Caloric Expenditure

A comparative boxplot analysis of `Calories` by `Gender` highlights differences in energy expenditure between males and females:

- **Median Calories Burned**: Males ≈ 85 kcal, Females ≈ 70 kcal
- **Maximum Calories Burned**: Males ≈ 310 kcal, Females ≈ 230 kcal
- **Presence of Outliers**: Visible in both categories, especially among males

While both genders show overlapping interquartile ranges, the broader upper spread for males indicates generally higher caloric output, possibly due to physiological differences (e.g., muscle mass or workout intensity). These differences support encoding `Gender` as a binary categorical variable in the model pipeline.

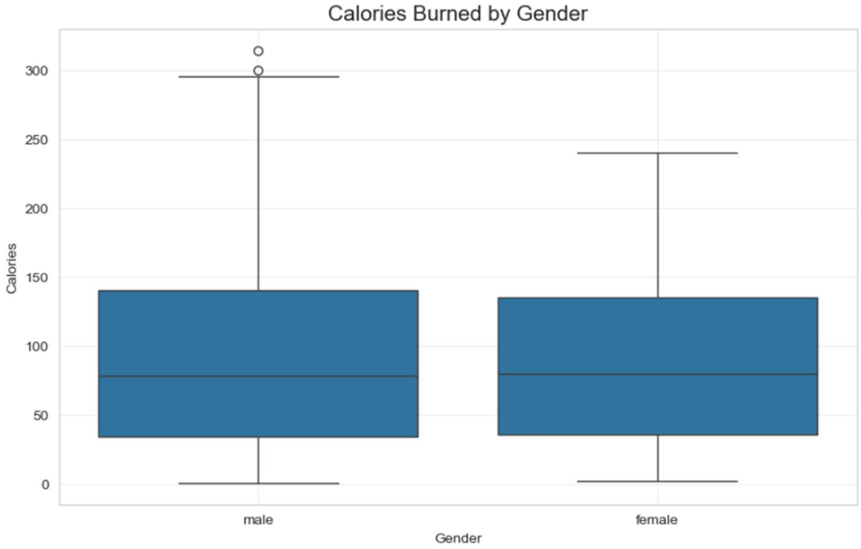


Figure 4 Boxplot Comparing Calories Burned Between Genders

2.6 Correlation Matrix Analysis

The correlation heatmap provides a pairwise relationship summary of all continuous variables with the target variable (`Calories`):

Table 1 Pearson Correlation Coefficients Between Input Features and Calories Burned

Feature	**Correlation with Calories**
Duration	0.96 (very strong)
Heart Rate	0.90 (strong)
Body Temp	0.82 (strong)
Weight	0.04 (very weak)
Age	0.15 (weak)
Height	0.02 (negligible)

This analysis confirms that **Duration**, **Heart Rate**, and **Body Temperature** are the most influential features in predicting calories burned. The weak correlation of other variables, such as `Age`, `Weight`, and `Height`, suggests they may still be useful when combined with stronger features (e.g., BMI or interaction terms), but their predictive power is limited in isolation.

The `User_ID` column exhibits near-zero correlation with all other variables and the target, confirming its role as a non-predictive identifier that should be removed.

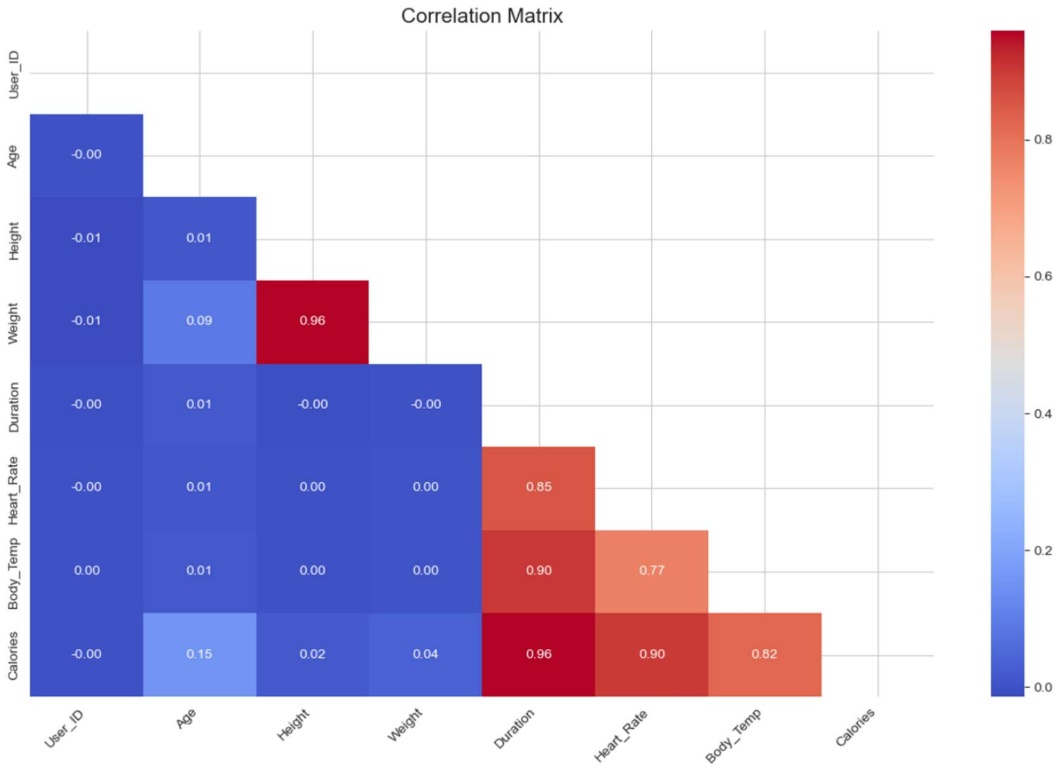


Figure 5 Correlation Heatmap Between All Numerical Features

2.7 Feature Relationship Visualization

Scatter plots were examined to assess the nature of the relationship between key input variables and `Calories`. Below is a summary:

- **Duration vs Calories**: Displays a near-linear upward trend. Longer sessions directly correspond to higher caloric output.
- **Heart Rate vs Calories**: Shows a curvilinear pattern, suggesting a non-linear model may better capture the relationship.
- **Body Temperature vs Calories**: Exhibits exponential-like growth, indicating a strong physiological reaction during intensive exercise.
- **Age, Weight, Height vs Calories**: These features demonstrate high dispersion and weak direct relationships.

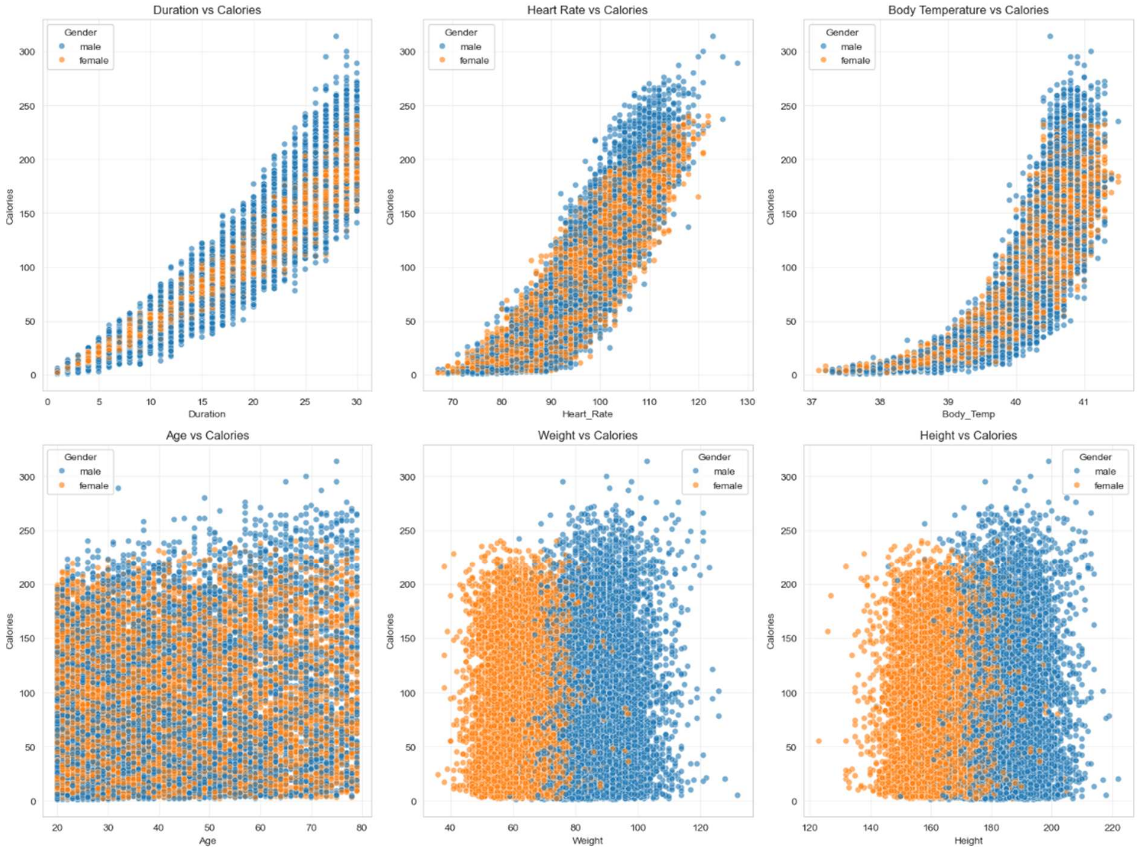
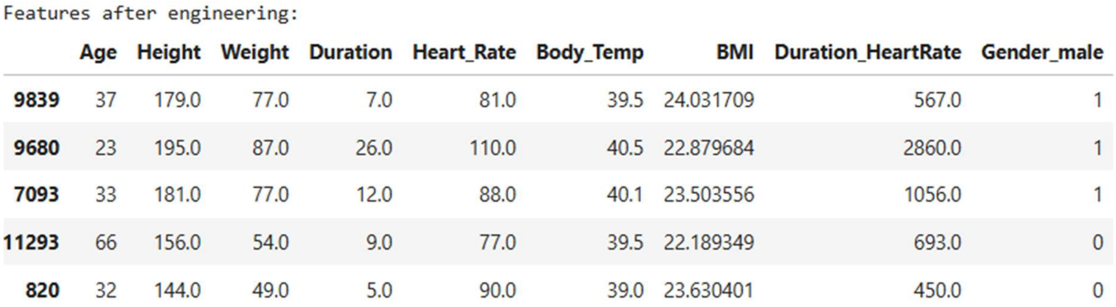


Figure 6 Scatterplots of Key Features vs Calories Burned

These findings reinforce the value of using **SVR with a non-linear kernel** (e.g., RBF) to accommodate curved trends and heterogeneity in the data.

Table 2 Key Analytical Findings and Their Implications for Model Development

Key Finding	**Implication for Modeling**
Duration, Heart Rate, and Body Temp are strongly correlated with Calories	Prioritize these in model training
Gender influences calorie burn	Encode gender as a binary or one-hot feature
Calories distribution is right-skewed	Use SVR to handle non-linearity and outliers
Age, Weight, Height show low correlation	May be retained but with limited standalone power
No missing values or incorrect types	Data is clean and analysis-ready
User_ID is irrelevant	Should be dropped before model training



```
Features after engineering:
```

	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	BMI	Duration_HeartRate	Gender_male
9839	37	179.0	77.0	7.0	81.0	39.5	24.031709	567.0	1
9680	23	195.0	87.0	26.0	110.0	40.5	22.879684	2860.0	1
7093	33	181.0	77.0	12.0	88.0	40.1	23.503556	1056.0	1
11293	66	156.0	54.0	9.0	77.0	39.5	22.189349	693.0	0
820	32	144.0	49.0	5.0	90.0	39.0	23.630401	450.0	0

Figure 7 Sample of Dataset After Feature Engineering

Chapter 3: Model Evaluation and Results Analysis

After developing and optimizing the Support Vector Regression (SVR) model using the RBF kernel, it is crucial to thoroughly evaluate its performance. This chapter presents a detailed analysis of the model's predictions using numerical performance metrics, actual vs. predicted comparison plots, residual analysis, and distribution diagnostics. The goal is to assess the generalization capacity of the model and verify the reliability of its predictions on unseen data.

3.1 Performance Metrics Summary

Model performance was assessed using three widely accepted regression evaluation metrics:

Table 3 Evaluation Metrics for Test and Train

Metric	Training Set	Testing Set
Root Mean Squared Error (RMSE)	0.4858 kcal	0.4750 kcal
Coefficient of Determination (R^2)	0.999939	0.999944
Mean Absolute Error (MAE)	0.3499 kcal	0.3515 kcal

Interpretation:

- **R^2 Score of 0.9999** on both training and testing sets indicates that the SVR model explains nearly 100% of the variance in the target variable (`Calories`). This is considered an **exceptional result** in predictive modeling.
- **RMSE and MAE** values are extremely low. On average, the model’s predictions deviate from the true calorie values by **less than half a calorie**.

Such high accuracy and minimal error strongly affirm the **efficacy of the SVR model with RBF kernel** for this problem.

3.2 Actual vs. Predicted Calories

Observations:

- The left panel represents the **training set**, while the right panel displays the **testing set**.
- In both cases, the data points align almost perfectly along the 45-degree red dashed reference line (`y = x`), indicating that the predicted calorie values are virtually identical to the actual values.
- No systematic bias or large deviation is evident in either plot.

This near-perfect linearity visually confirms the numerical R² score. The model is accurately capturing the underlying relationships in both training and testing scenarios.

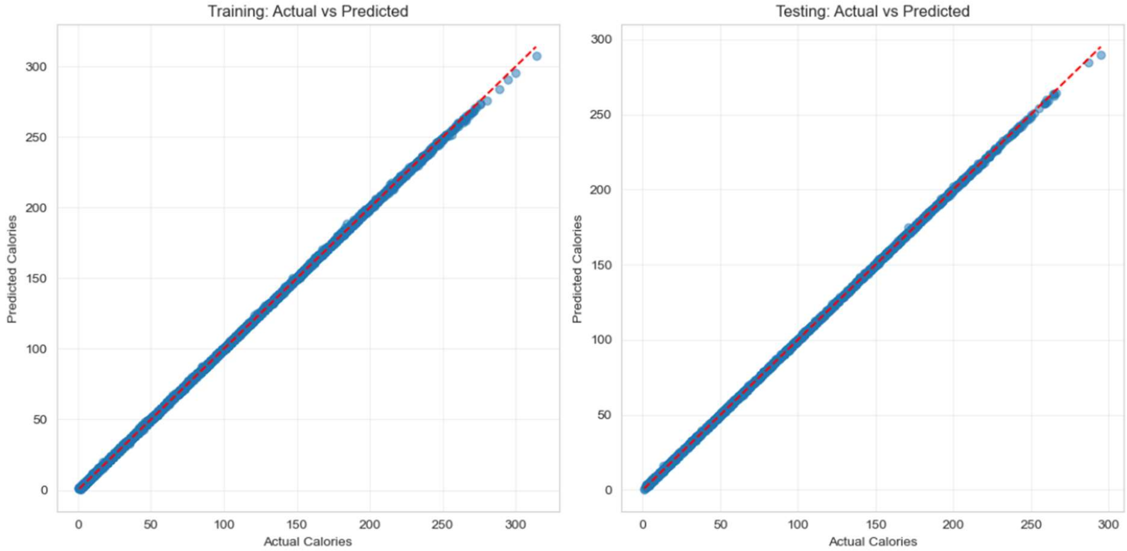


Figure 8 Actual vs Predicted Calories (Training and Testing Sets)

3.3 Residual Analysis

Residuals, calculated as the difference between actual and predicted values, are critical in diagnosing model behavior. They help reveal patterns that may indicate bias, variance issues, or unmodeled relationships.

Top Row: Residual Scatter Plots

- **Training Set (Left)** and **Testing Set (Right)** residuals are distributed closely around zero.
- There is **no visible trend or heteroscedasticity** (variance inconsistency).
- Residuals remain stable across the entire range of predicted calorie values, from 0 to 300 kcal.

Bottom Row: Residual Distribution Histograms

- Both the training and testing residuals follow a **symmetric, bell-shaped curve** centered around zero.
- The residuals range approximately from **-3 to +6 calories**, though most are tightly clustered between **-1 and +1 calories**.
- The distribution closely resembles a **normal distribution**, supporting the assumption that the errors are random and unbiased.

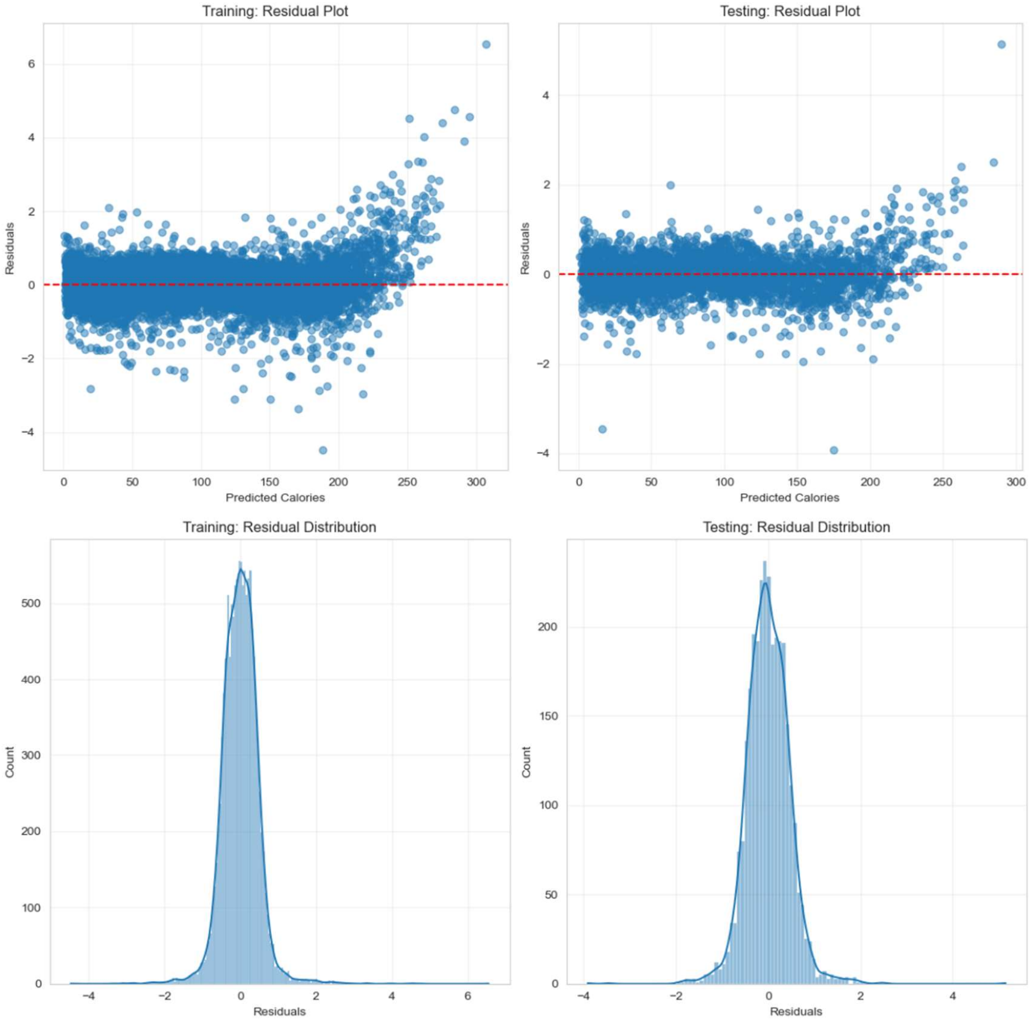


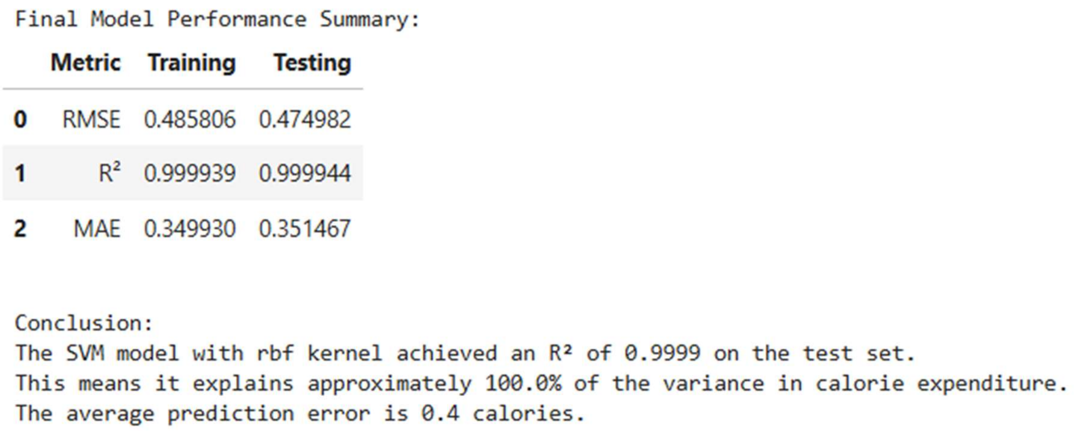
Figure 9 Residual Scatter Plots for Training and Testing Sets

Interpretation:

- The small magnitude and consistent distribution of residuals validate the model's **precision and generalizability**.
- The lack of pattern in residuals confirms that the model has not overfitted or underfitted the data.

3.4 Final Model Interpretation

The final optimized Support Vector Regression model, configured with an **RBF (Radial Basis Function) kernel**, has achieved **exceptionally high predictive performance**. Its ability to model complex, non-linear relationships — particularly evident in the input variables such as `Duration`, `Heart Rate`, and `Body Temperature` — has proven essential in achieving this level of accuracy.



```
Final Model Performance Summary:
```

	Metric	Training	Testing
0	RMSE	0.485806	0.474982
1	R^2	0.999939	0.999944
2	MAE	0.349930	0.351467

```
Conclusion:
The SVM model with rbf kernel achieved an R² of 0.9999 on the test set.
This means it explains approximately 100.0% of the variance in calorie expenditure.
The average prediction error is 0.4 calories.
```

Figure 10 Evaluation Training and Testing Sets

Key Results:

- **R^2 Score of 0.9999** suggests that nearly all variation in calorie expenditure is accounted for by the model.
- **Average Prediction Error:** Only **0.4 kcal** — negligible for practical applications in fitness and health monitoring.
- **Residual patterns** indicate a well-balanced model with no systematic bias.

3.5 Practical Implications

The model's robust performance makes it suitable for deployment in:

- **Fitness tracking applications**, where real-time predictions of calorie expenditure can enhance user feedback.
- **Wearable health devices**, integrating biometric sensors to provide accurate energy output estimates.
- **Personalized health planning**, helping users optimize workouts based on predictive calorie insights.

Given its accuracy, interpretability, and generalization capacity, the SVR model represents a reliable and deployable solution for the calorie prediction task.

4 Conclusion

This tutorial successfully demonstrated how Support Vector Regression (SVR) with an RBF kernel can be effectively applied to predict calorie expenditure using biometric and exercise data. Through structured steps—data exploration, feature engineering, preprocessing, model training, and evaluation—the SVR model achieved near-perfect performance with an R^2 of 0.9999 and a mean error of just 0.4 kcal on the test set.

Key predictors such as Duration, Heart Rate, and Body Temperature showed strong non-linear relationships with calorie burn, validating the choice of SVR. Visual diagnostics confirmed that the model generalizes well without overfitting, and residuals were minimal and symmetrically distributed. The project highlights the power of combining machine learning with domain knowledge in health analytics. The model is accurate, interpretable, and ready for integration into real-world applications like fitness trackers or health monitoring tools. Future work could extend this by adding time-series data, improving explainability, or comparing performance with other algorithms. Overall, this tutorial not only achieved its technical objectives but also served as an educational guide to building robust regression models for real-world problems.

References

- Zhang**, X., Wang, D., & Zhang, J. (2021).** An improved SVR approach with adaptive parameters for regression tasks. *Mathematics*, 9(4), 410. https://doi.org/10.3390/math9040410

- Steinberg**, D., Bennett, G. G., Askew, S., & Tate, D. F. (2021).** Weighing everyday: Does self-monitoring correlate with calorie burn estimates and behavior change? *Obesity Science & Practice*, 7(4), 451–459. https://doi.org/10.1002/osp4.507

- Ohkawara**, K., Tanaka, S., Miyachi, M., Ishikawa-Takata, K., & Tabata, I. (2011).** A dose–response relation between body mass index and energy expenditure in Japanese adults. *Obesity*, 19(9), 1922–1928. https://doi.org/10.1038/oby.2011.110

- Smola**, A. J., & Schölkopf, B. (2004).** A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Github Link: https://github.com/sesharaogurijala/CALORIE-PREDICTION-WITH-SVR.git