The Application of Supervised Learning Techniques in the Quantitative Estimation of Caloric Burn and Optimal Exercise Duration

Rafid Ahnaf Id: 19201143

Abstract—This report explores the application of a few popular supervised learning strategies for the quantitative estimation of caloric burn and most optimal exercising duration to keep body temperature under control. Leveraging a dataset that incorporates more than one capabilities like age, weight, height, heart rate, body temperature and exercising duration, the observe applies machine learning knowledge of models including Decision Trees, Linear Regression, Random Forests, Gradient Boosting, and Support Vector Regression. Through rigorous data preprocessing steps like missing value treatment and feature selection, the report lays the groundwork for model training. While the paper does now not delve into the overall performance metrics in element, it emphasizes the potential metrics that would be evaluated, inclusive of Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and the R^2 Score. The report concludes by highlighting the potential of supervised models in providing more accurate and personalised estimates, therefore contributing to advancements in the fields of fitness and health.

Index Terms—Supervised Learning, Caloric Burn Estimation, Optimal Exercise Duration, Data Preprocessing, Health and Fitness.

1 Introduction

THE quantification of caloric burn and optimal exercise duration holds significant importance in the realms of health and fitness. Accurate caloric measurement during physical activities not only aids in personalized fitness planning but also informs dietary adjustments and overall health monitoring. This report delves into the intricate application of multiple supervised learning models to estimate both these critical metrics. The machine learning models utilize a well-balanced dataset that includes multiple variables like age, weight, height, gender, average heart rate, and body temperature. Each of these features plays a very important role in determining the approximate amount of calories burned during exercise. In addition to that, health research shows that the maximum safe body temperature for humans during exercise is 40 degrees Celsius and so this report also aims to predict an optimal exercise duration in order to ensure that individuals do not overheat during their workouts, adding a layer of safety to their fitness routines. Measuring the optimal exercise time can help avoid various health concerns that can occur because of excess muscle damage or body temperature. By combining these various data points, the supervised learning algorithms employed—ranging from Linear Regression to Random Forest Regressor—seek to provide a comprehensive, accurate, and personalized approach to estimating caloric burn and determining the safest and most effective length of physical activity.

2 Methodology

2.1 Analyzing of Dataset

The dataset used in this study provides a rich and numerous set of features that shape the basis for constructive modeling. It encompasses key variables related to individual body structure and exercising styles, making it properly-appropriate for the dual goals of caloric burn estimation and optimal exercise duration prediction. One of the dataset's salient features is the inclusion of Body Temperature, a critical metric given that the maximum safe temperature all through workout have to no longer exceed forty degrees Celsius. The types and number of data of the dataset are given below:

TABLE 1: Dataset Features and Their Types

Feature	Type	Non-Null Count
User_ID	Integer	14,020
Calories	Float	13,740
Gender	Object	13,844
Age	Float	13,865
Height	Float	13,867
Weight	Float	13,876
Duration	Float	13,861
Heart_Rate	Float	13,857
Body_Temp	Float	13,859

The presence of a widespread range of null values throughout diverse capabilities requires meticulous preprocessing steps to ensure the integrity and reliability of the subsequent analyses. Overall, the dataset gives a robust basis for developing models that goal to provide customized, correct, and safe tips for caloric expenditure and workout length.

2.2 Preprocessing the dataset

2.2.1 Handling Null Values

Data cleaning is critical in data analysis, and one of its essential features is dealing with missing or null values. These missing numbers may bring inconsistencies into the model, and ignoring them can lead to computational mistakes during analysis. Unlike many datasets that may be

missing values-free, our dataset had a large number of null entries across various characteristics. Such insufficient data presents difficulties, particularly in a dataset as diverse as ours, which strives to offer a thorough basis for estimating calorie burn and appropriate exercise duration. Rather of deleting the partial information, we used imputation algorithms to address this problem.Techniques such as filling missing values with the mean or median of the respective feature were applied to preserve the integrity of the dataset and to maintain the robustness of our subsequent analyses. However this was not possible for categorical non-numerical values.

2.2.2 Categorical Encoding

The format of the input data is critical when dealing with machine learning models. While some algorithms can handle categorical variables, the majority are designed to handle numerical inputs. Our dataset included a 'Gender' column that classified people as 'Male' or 'Female.' We used Label Encoding techniques to simplify this feature for computational analysis. In this transformation, we assigned the value 'Male' to '1' and the value 'Female' to '0,' turning the textual data into a numerical form that can be easily consumed by multiple algorithms. This critical step in data preparation not only made the dataset machine-friendly, but it also preserved the original information's essence.

2.2.3 Feature Selection

In predictive modeling, the quality of the input features can significantly influence the model's efficacy. While some variables contribute meaningfully to predictions, others may introduce noise or redundancy, diminishing the model's performance. To fine-tune our dataset, we utilized correlation analysis, which assigns a coefficient between -1 and 1 to pairs of features, revealing the nature and strength of their relationship.

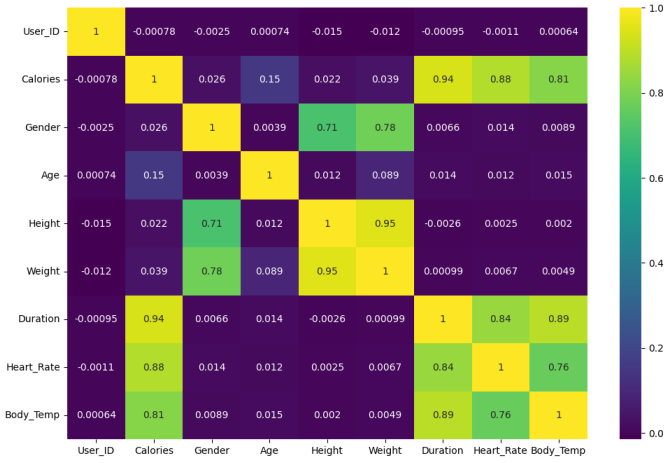


Fig. 1: Correlation Matrix

Our analysis led us to remove the User_ID variable, which showed little to no correlation with other attributes. Conversely, while 'Height' and 'Weight' as well as 'Exercise Duration', 'Heart Rate','Body Temperature' exhibited a high degree of correlation, we opted to preserve both in the dataset, recognizing their individual contributions to the model.

3 Supervised Models Used

3.1 Linear Regressor

Linear Regression is a supervised modeling technique that is used to analyze the relationship between a dependent variable Y and one or more independent variables X. The algorithm aims to find the best-fitting linear equation:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

Here, $\beta_0, \beta_1, \ldots, \beta_n$ are coefficients to be optimized, and ϵ is the error term. The optimization goal is to minimize the sum of squared differences between observed and predicted values, typically using the Least Squares method.

This model is popular for its simplicity and interpretability.

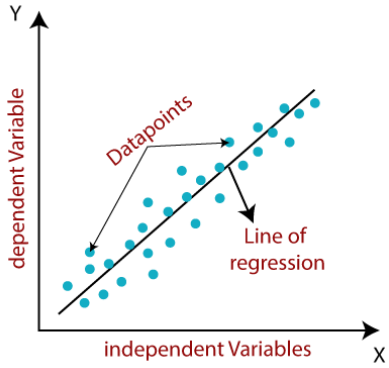


Fig. 2: Linear Regressor Model

3.2 SVM Regressor

Support Vector Machines (SVM) are a class of supervised learning models used for classification and regression tasks. The primary goal of SVM is to find a hyperplane that best separates the data into different classes. For a two-class problem, the mathematical representation of this hyperplane is given by:

$$\vec{w} \cdot \vec{x} - b = 0$$

Here, $\vec{w}$ is the weight vector and b is the bias term. The objective is to maximize the margin between the closest points (support vectors) of the different classes, subject to the following constraints:

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \quad \text{for all } i$$

SVM is renowned for its effectiveness in high-dimensional spaces and is often preferred for its robustness and accuracy.

3.3 Decision Tree Regressor

Decision Tree Regressor is a type of supervised learning algorithm that is primarily used for solving regression problems. The model works by partitioning the feature space into a set of rectangles, and then fitting a simple model (like a constant) in each one. The tree is constructed through a process known as binary recursive partitioning.

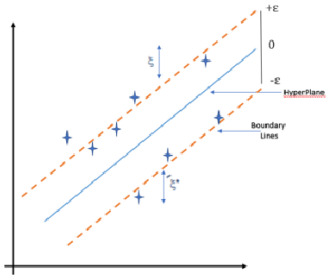


Fig. 3: Support Vector Machine Model

The objective is to minimize the sum of the squared differences between the observed and predicted values within each partition. Mathematically, the cost function $J(D)$ to minimize is:

$$J(D) = \frac{1}{|D|} \sum_{i \in D} (y_i - \hat{y})^2$$

Here, $|D|$ is the number of samples in the partition D, y_i are the actual values, and $\hat{y}$ is the average value of y in D.

Decision Trees are intuitive and easy to interpret but can easily overfit or underfit the data.

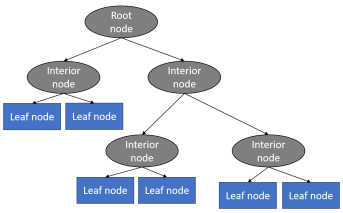


Fig. 4: Decision Tree Regressor Model

3.4 Random Forest Regressor

Random Forest Regressor is an ensemble learning method that combines multiple Decision Trees to improve the model's generalization and robustness. Each tree in the ensemble is built from a bootstrap sample of the data, and random subsets of features are used to split the nodes. The final prediction is obtained by averaging the predictions of all individual trees. Mathematically, the predicted value $\hat{Y}$ is given by:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

Here, $\hat{y}_i$ is the prediction from the i^{th} Decision Tree and n is the number of trees.

Random Forests offer high accuracy and the ability to handle large data sets with higher dimensionality.

3.5 Gradient Boosting Regressor

Gradient Boosting Regressor is another ensemble technique that builds multiple Decision Trees sequentially, where each tree aims to correct the errors of its predecessor. It can also ustilize linear regressors instead od decision trees if preferred. The method leverages the concept of boosting,

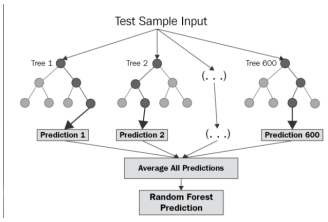


Fig. 5: Random Forest Regressor Model

focusing more on instances that were mispredicted by the previous models. The final prediction is a weighted sum of the individual tree predictions:

$$\hat{Y} = \sum_{i=1}^{n} w_i \hat{y}_i$$

Here, $\hat{y}_i$ is the prediction from the i^{th} Decision Tree and w_i is the weight assigned to this prediction.

Gradient Boosting is known for its high accuracy and effectiveness in predictive modeling but may require careful tuning of parameters.

Fig. 6: Gradient Boosting Regressor Model

4 Results

4.1 Estimated Caloric Burn

In this section, we present the performance outcomes of the various machine learning models employed in the study. A comprehensive summary of the evaluation metrics for each model is provided in the given table:

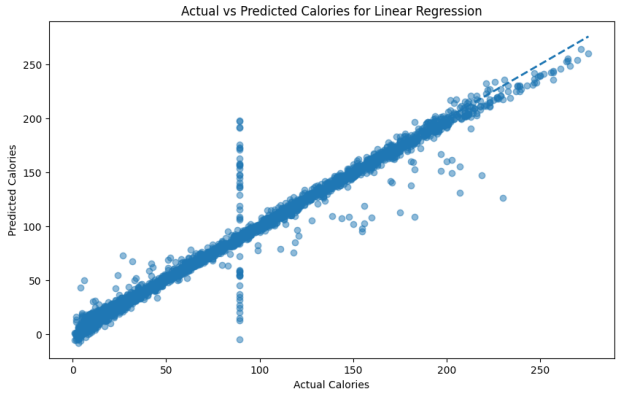
TABLE 2: Performance Metrics for Models

Model	MSE	MAE	RMSE	R^2
Linear Regressor	110.53	4.46	10.51	0.97
SVM	139.84	4.14	11.82	0.96
Decision Tree Regressor	195.08	5.73	13.96	0.95
Random Forest Regressor	96.27	3.99	9.81	0.97
Gradient Boosting Regressor	99.75	4.30	9.98	0.98

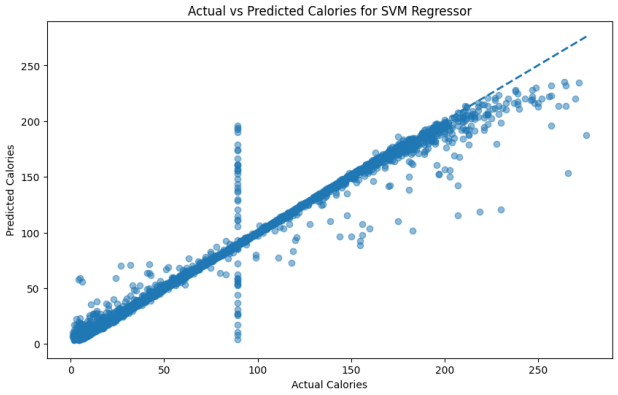
Upon evaluating the performance metrics, it is evident that the Gradient Boosting Regressor outperforms the other models across all key indicators. With the lowest MSE of 99.75, MAE of 4.30, and RMSE of 9.98, along with the highest R^2 value of 0.98, the Gradient Boosting Regressor proves to be the most efficient and accurate model for this study.

4.2 Optimal Exercise Duration

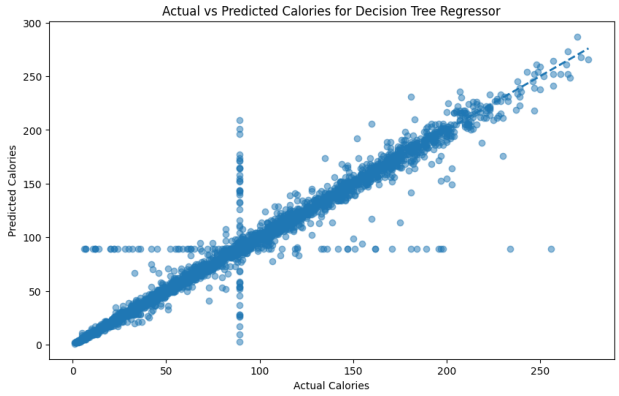
Further extending our analysis, the Gradient Boosting Regressor was employed to estimate the optimal exercise duration while keeping the body temperature below 40 degrees Celsius. For this specialized task, the model's performance metrics were also computed and are as follows:



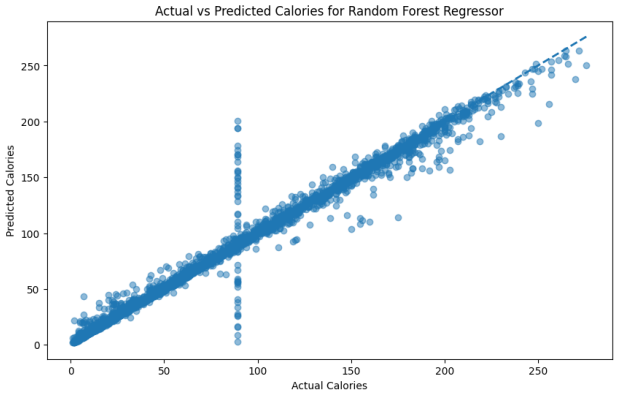
(a) Linear Regressor



(b) SVM



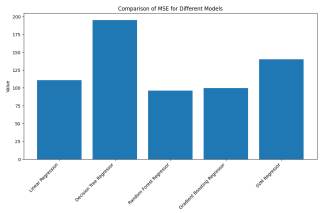
(c) Decision Tree Regressor



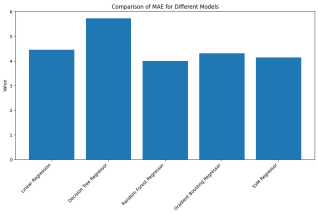
(d) Random Forest Regressor

(e) Gradient Boosting Regresso

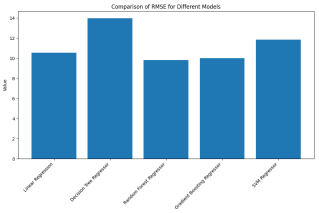
Fig. 7: Actual vs. Predicted values for different models



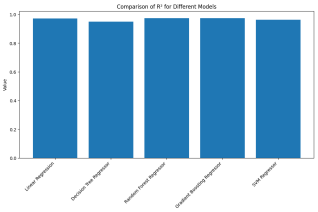
(a) MSE



(b) MAE



(c) RMSE



(d) R^2

Fig. 8: Comparison of Different Performance Metrics

MSE of 12.69, MAE of 2.85, RMSE of 3.56, and R^2 value of 0.81. These results underscore the model's efficacy not only in caloric burn estimation but also in determining a safe and optimal exercise duration. It successfully identifies the durations for which the body temperature remains below the critical 40-degree Celsius mark, thereby adding an extra

layer of safety to exercise routines.

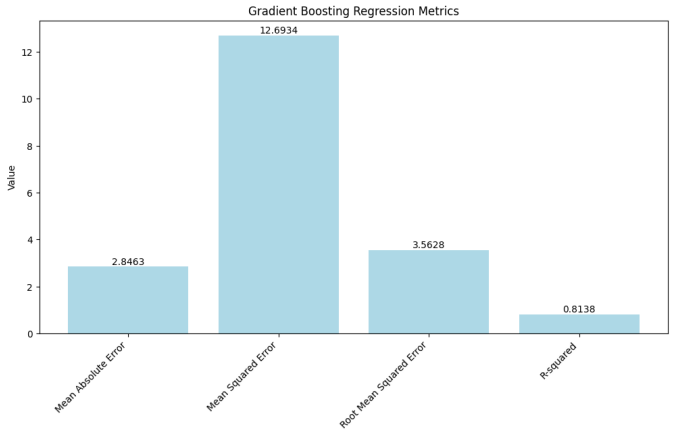


Fig. 9: Gradient Boosting Regressor for optimal time

5 Conclusion

This study aimed to apply various supervised learning models to quantitatively estimate caloric burn and optimal exercise duration. Among the models tested, the Gradient Boosting Regressor emerged as the most effective, achieving the highest performance metrics across all key indicators. Its efficacy was further demonstrated in a specialized task of estimating exercise durations that keep the body temperature below a critical 40-degree Celsius threshold.

The insights gained from this research hold significant practical implications. By accurately predicting caloric burn and safe exercise durations, this model can serve as a valuable tool for personalized fitness and health management. Moreover, the approach can be integrated into wearable devices or fitness apps to provide real-time, data-driven recommendations, thereby improving the overall quality and safety of physical exercise regimes.

Future work could explore more complex models or consider additional features, such as type of exercise or individual metabolic rates, to further improve the model's predictive power. Nonetheless, the current findings make a strong case for the adoption of machine learning techniques in the health and fitness domain.

The End