Predicting Calories Burnt Using Machine Learning

Anonymous CVPR submission

Paper ID Group-62

Abstract

This project focuses on developing a machine-learning model to predict calories burned during physical activities. Traditional methods for estimating calorie expenditure often fall short due to generalized assumptions and limited personalization. Our model will leverage key individual factors, such as age, weight, heart rate, and exercise type, to provide more accurate and tailored predictions. Multiple models including Linear Regression, Random Forest, XGBoost, and Support Vector Machines (SVMs) have been deployed to predict calories burned. This report discusses the dataset, preprocessing steps, model development, results, and analysis.

1. Introduction

Technology integration into personal fitness and health monitoring has gained significant traction recently. One critical aspect of health monitoring is accurately tracking the calories burned during various physical activities, which helps individuals optimize their exercise routines, manage weight, and improve overall well-being.

Traditional methods of calorie estimation are often limited by simplistic assumptions, such as linear relationships between exercise duration and calories burned, which may lead to inaccurate estimates. These inaccuracies stem from factors like heart rate, body temmperature, and individual attributes such as age, weight, and gender, which play a critical role in determining calorie expenditure.

By utilizing data-driven approaches, machine learning models can account for the intricate interactions between physiological factors, improving the precision of calorie estimates across a wide range of activities and individuals. This project explores different machine learning algorithms, including linear regression, ensemble learning methods like XGBoost and Random Forest, SVMs to develop a model for predicting calories burned based on individual attributes.

The goal is to evaluate these models' effectiveness, compare their performance, and determine the most accurate approach for personalized calorie prediction. By enhancing calorie tracking accuracy, this project aims to contribute to more effective fitness tracking and healthier lifestyle choices.

2. Literature Survey

A number of studies have applied machine learning to predict calories burned during physical activities. Kadam et al. [1] used Random Forest, finding it to be the best-performing model with an MAE of 5.33, an MSE of 68.92, and an RMSE of 8.3. Likhon et al. [2] applied XGBoost and achieved the best prediction results, with an MAE of 1.48. Panwar et al. [3] highlighted the importance of hyperparameter tuning in machine learning models, with XGBoost showing the lowest RMSE and a perfect R^2 score of 1.0. These studies emphasize the effectiveness of ensemble methods in improving the accuracy of calorie predictions and suggest the importance of feature selection and tuning.

3. Dataset

The dataset used for this project consists of 15,000 data points, each representing a unique exercise session. Each data point includes the following features:

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

These features were selected based on their known impact on calorie expenditure. The dataset provides a broad representation of various exercise sessions, making it ideal

for training machine learning models to predict calorie burn based on different individual characteristics.

4. Data Preprocessing

Before training the models, we performed several preprocessing steps:

- Filling missing values: Some features had missing values, which were handled by either imputing the missing data using mean imputation or removing rows with excessive missing data.

- Scaling: Features like age, weight, and heart rate were normalized to ensure that they were on the same scale, preventing features with larger values from dominating the model training.

- Label Encoding: Categorical variables, such as gender, were encoded into numerical labels for model compatibility.

- Outlier Detection: We identified and handled outliers in the dataset that could distort model performance, especially for continuous variables like heart rate and exercise duration.

5. Methodology and Model Details

5.1. Exploratory Data Analysis (EDA)

EDA revealed that certain features, such as height and weight, displayed a positive correlation, while others, such as heart rate and exercise duration, showed nonlinear relationships. Scatter plots revealed that males and females had distinct patterns in several features, indicating possible gender-based differences in calorie expenditure.

Many scatter plots highlight distinct separations between males and females across various features. For instance, certain features might have a more pronounced clustering of one gender, suggesting possible correlations between those features and gender. Features like height and weight typically show a positive correlation, where an increase in one corresponds to an increase in the other.

Certain features display nonlinear relationships, which could indicate the need for polynomial regression or other nonlinear modeling techniques. Some pairs, such as those involving height and weight, exhibit strong linear relationships. This multicollinearity could impact regression analyses, leading to inflated standard errors for the coefficients.

The distribution plots reveal some potential outliers, especially in features with wide ranges. These outliers may need to be examined further, as they can significantly affect modeling outcomes. In some scatter plots, the overlapping points suggest high density, indicating that many individuals share similar values for those features. This clustering could be explored further to understand the underlying factors influencing these values.

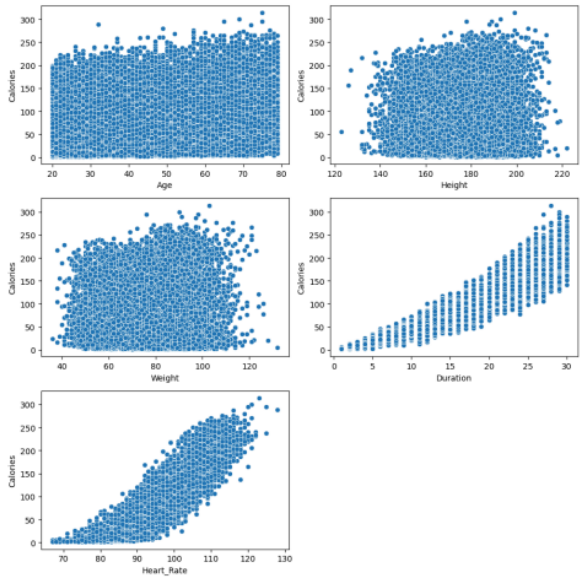

Figure 1. Exploratory Data Analysis

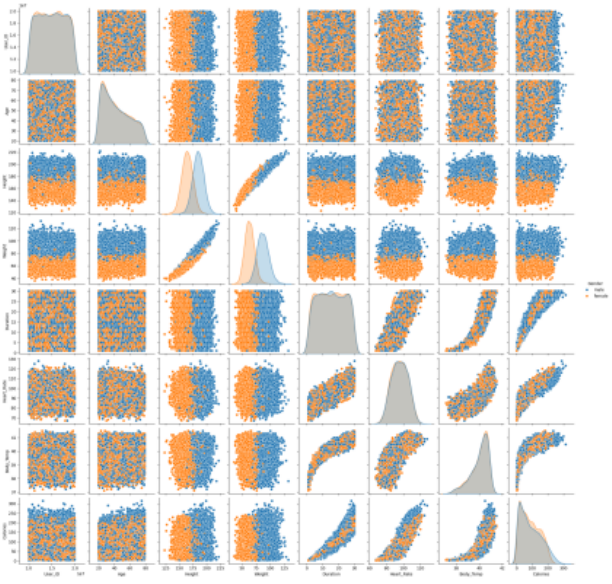

Figure 2. Exploratory Data Analysis

5.2. Feature Generation

To enhance the model's performance, we generated new features that were ratios or squares of existing continuous

features, such as weight-to-heart rate ratios and duration squared. Feature selection was based on the correlation of each feature with the target variable (calories burned), ensuring that only relevant features were included in the models.

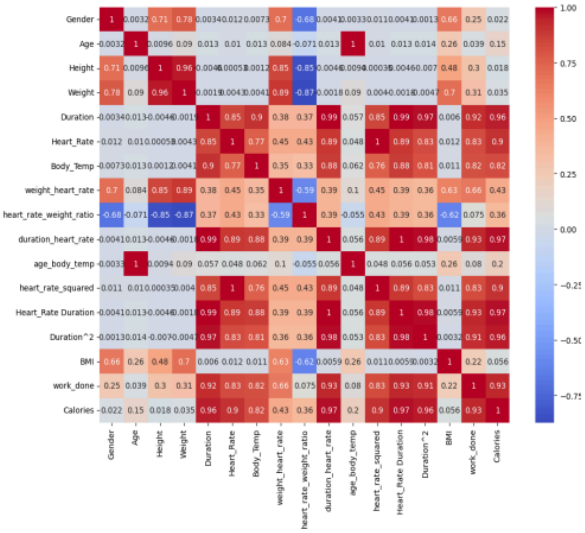

Figure 3. Correlation matrix

6. Methodology and Model Details

6.1. Model Selection

For this project, we implemented and evaluated several machine learning models to predict calories burned based on different features, such as age, weight, height, heart rate, and exercise duration. These models include linear regression, ensemble models (Random Forest and XGBoost), and Support Vector Machines (SVMs). We trained each model on two sets of features: 1. Original Features: Features, including Age, Height, Weight, Duration, Heart Rate, and Body Temperature. 2. All Features: All the features available in the dataset and the features obtained via feature engineering.

The models implemented are detailed below:

6.2. Linear Regression

We implemented three variations of the Linear Regression model to predict calories burned:

- Model 1: Linear regression using only the original set of features.
- Model 2: Linear regression using all available features, including the ones generated.
- Model 3: Linear regression using a subset of the most correlated features.

The model was trained using a training set with 70% of the data and tested on the remaining 30% of the data. The performance of the models was evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score.

The models were trained using both Original Features and all features, and their performance was evaluated based on various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. The results for each model are presented below.

6.3. Linear Regression (Model 1: Original Features)

The first linear regression model, which uses only the features present in the dataset (Gender, Age, Height, Weight, Duration, Heart Rate, and Body Temperature), returned the following performance:

Model	MSE	MAE	RMSE	R^2
Linear Regression (Model 1)	11.529	8.451	3.394	0.966

Table 1. Performance of Linear Regression (Model 1) on Original Features

6.4. Linear Regression (Model 2: All Features)

The second linear regression model, which uses all generated features as well as the original features, returned improved results as follows:

Model	MSE	MAE	RMSE	R^2
Linear Regression (Model 2)	5.866	4.293	2.425	0.991

Table 2. Performance of Linear Regression (Model 2) on All Features

6.5. Linear Regression (Model 3: Original Features)

The third linear regression model, which uses a subset of features specifically selected based on their relevance (including Gender, Age, Height, Weight, Duration, Heart Rate, BMI, and several generated features like weight-to-heart rate ratio and age-body temp), returned the following results:

Model	MSE	MAE	RMSE	R^2
Linear Regression (Model 3)	5.857	4.280	2.421	0.9913

Table 3. Performance of Linear Regression (Model 3) on Original Features

6.6. Comparison of Linear Regression Models

The following table summarizes the performance of all three linear regression models evaluated:

Model	MSE	MAE	RMSE	R^2
Model 1	11.529	8.451	3.394	0.966
Model 2	5.866	4.293	2.425	0.991
Model 3	5.857	4.280	2.421	0.9913

Table 4. Summary of Linear Regression Model Performance

6.7. Random Forest

The Random Forest Regressor is an ensemble method known for its ability to handle overfitting and its high accuracy in prediction tasks. We trained the Random Forest model with both original and all features. The model was tuned with 100 estimators, and the performance was evaluated based on the same metrics as linear regression.

The results for the **Random Forest** models (on Original Features and all features) are presented in the results section.

6.8. XGBoost

We also implemented the XGBoost Regressor, which is an advanced ensemble method that is known to handle overfitting more effectively than traditional decision trees. The model was trained using the same sets of features as the Random Forest model. XGBoost was used to evaluate the importance of each feature and to determine how well the model performed with both original and all features.

6.9. Support Vector Machines (SVM)

The Support Vector Machine (SVM) model was implemented using the SVR (Support Vector Regression) approach. SVM is particularly effective when dealing with non-linear data and can provide a good generalization in regression tasks. The model was tested with both original and all features to evaluate its performance.

6.10. Decision Tree Regressor

Finally, a Decision Tree Regressor was implemented. Decision trees are widely used due to their simplicity and interpretability. However, they are prone to overfitting if not properly tuned. We tested the model on both original and all features and evaluated its performance on the test data.

6.11. Model Training and Evaluation

For all models, the training set consisted of 70% of the dataset, and the test set consisted of the remaining 30%. We evaluated the models based on the following metrics:

- Mean Squared Error (MSE): A measure of the average squared differences between predicted and actual values. Lower values indicate better performance.

- Mean Absolute Error (MAE): A measure of the average absolute differences between predicted and actual values. Lower values indicate better performance.

- R^2 Score: A measure of how well the model explains the variance in the target variable. A higher R^2 score indicates a better fit.

The performance of each model was evaluated on both the Original Features and all features, and the results are compared in the following section.

7. Results and Analysis

The models were trained using both Original Features and all features, and their performance was evaluated based on various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. The results for each model are presented below.

7.1. Random Forest (Original Features)

For the Random Forest model, using Original Features (Age, Height, Weight, Duration, Heart Rate), the following results were observed:

Model	MSE	MAE	RMSE	R^2
Random Forest(original)	8.431	1.813	2.903	0.997

Table 5. Performance of Random Forest on Original Features

7.2. Random Forest(All Features)

When all features were included, the performance improved as follows:

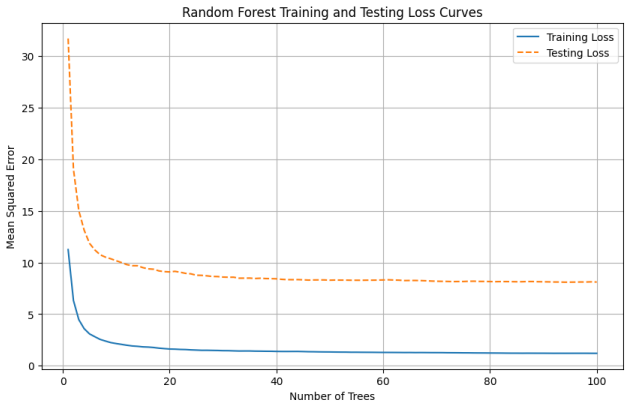

Figure 4. loss curve for Random Forest

Model	MSE	MAE	RMSE	R²
Random Forest (All Features)	8.110	1.739	2.847	0.997

Table 6. Performance of Random Forest on All Features

The feature importance for the Random Forest model on the Original Features was as follows:
Feature Importances: [0.005, 0.003, 0.000, 0.000, 0.000, 0.003, 0.000, 0.003, 0.001, 0.500, 0.023, 0.003, 0.455, 0.000, 0.000, 0.003]
These values indicate the relative importance of each feature in predicting the target variable, "Calories."

7.3. MLP Regressor (Original Features)

The MLP Regressor was also evaluated using Original Features:

Model	MSE	MAE	RMSE	R²
MLP (Original Features)	0.235	0.369	0.484	0.999

Table 7. Performance of MLP Regressor on Original Features

1

7.4. MLP (All Features)

For the MLP model with all features, the results were as follows:

Model	MSE	MAE	RMSE	R²
MLP (All Features)	0.202	0.352	0.449	0.999

Table 8. Performance of MLP Regressor on All Features

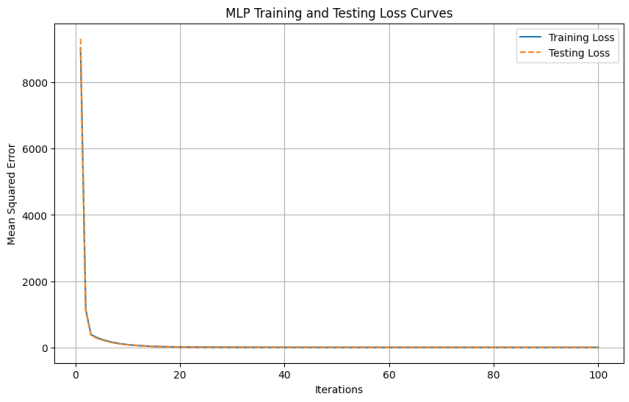

Figure 5. loss curve for MLP

7.5. Decision Tree (Original Features)

The Decision Tree Regressor results with Original Features are:

Model	MSE	MAE	RMSE	R²
DT(Original Features)	29.534	3.479	5.434	0.992

Table 9. Performance of Decision Tree Regressor on Original Features

7.6. Decision Tree (All Features)

The Decision Tree Regressor results with all features are:

Model	MSE	MAE	RMSE	R²
DT(All Features)	24.086	3.165	4.907	0.993

Table 10. Performance of Decision Tree Regressor on All Features

7.7. XGB-Boost

Model	MSE	MAE	RMSE	R²
XGB-Boost	2.835	1.178	1.683	0.999

Table 11. Performance of XGB-Boost Random Forest

7.8. GridSearch

Model	MSE	MAE	RMSE	R²
GridSearch	10.184	1.882	3.191	0.997

Table 12. Performance of GridSearch Random Forest

7.9. Comparison of Models

The table below summarizes the performance of all models evaluated using both Original Features and all features:

Model	MSE	MAE	RMSE	R²
Random Forest (Original)	8.731	1.813	2.903	0.997
Random Forest (All)	8.110	1.739	2.847	0.997
MLP Regressor (Original)	0.235	0.369	0.484	0.999
MLP Regressor (All)	0.202	0.352	0.449	0.999
Decision Tree (Original)	29.534	3.479	5.434	0.992
Decision Tree (All)	24.086	3.165	4.907	0.993
Grid Search	10.184	1.882	3.191	0.997
XGBBoost	2.835	1.178	1.683	0.999

Table 13. Model Performance Comparison (Original vs All Features)

7.10. Key Observations

- Random Forest (Original and All Features): Both versions of the Random Forest model performed well, with R² scores of 0.997. The version using all features

showed a slightly lower MSE of 8.110 compared to the Original Features version with an MSE of 8.731.

- MLP Regressor (Original and All Features): The MLP Regressor showed the best performance with very low MSE and MAE values. The version trained with all features performed slightly better, with an MSE of 0.202, MAE of 0.352, and RMSE of 0.449, along with a perfect R^2 score of 0.999.
- Decision Tree (Original and All Features): The Decision Tree models had higher error metrics than other models. The version trained with Original Features showed an MSE of 29.534 and MAE of 3.479, though it still achieved a high R^2 score of 0.992. The use of all features improved performance slightly, resulting in a reduced MSE of 24.086 and an MAE of 3.165, with a R^2 of 0.993.
- Grid Search: The Grid Search model achieved an MSE of 10.184, MAE of 1.882, and an R^2 score of 0.997, which is comparable to the performance of the Random Forest models.
- XGBoost: This model performed well with an MSE of 2.835, MAE of 1.178, and an RMSE of 1.683, reaching a very high R^2 score of 0.999.

8. Conclusion

The MLP Regressor demonstrated the best performance among all models, achieving the lowest error metrics and a perfect R^2 score of 0.999. The XGBoost and Random Forest models also showed strong performance, with R^2 scores of 0.999 and 0.997, respectively. The Decision Tree model, although effective, had a higher error rate compared to other models, but still produced good results with an R^2 of 0.993. The Grid Search model performed comparably to Random Forest.

Some of the major challenges we faced were feature generation and identifying overfitting in complex models. High model complexity can lead to overfitting, especially with small datasets. Determining which features influence the model's predictions can be non-trivial, especially with ensemble models like Random Forest or XGBoost. Identifying optimal hyperparameters can be time-consuming as well.

Contributions- Vinti Kumar Kushwah- Linear regression, XGBoost, GridSearch Saurabh Mehta- Data preprocessing, EDA, MLP, Decision Tree Aditya Gupta- Literature review, Feature Engineering, Random Forest

References

[1] A. Kadam, A. Shrivastava, S. K. Pawar, V. H. Patil, J. Michaelson, and A. Singh. Calories burned prediction using machine learning. In *Proceedings of the 6th International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1712–1717, 2023. 1

[2] Md. N. H. Likhon, F. Bhuiyan, Md. S. Bhuiyan, M. A. Fahim, and A. Hossain. Calories burnt prediction: A machine learning approach. In *TechRxiv*, 2022. 1

[3] P. Panwar, K. Bhutani, R. Sharma, and R. Saini. A study on calories burnt prediction using machine learning. In *ITM Web of Conferences*, 2022. 1