Calorie Burnt Prediction using Machine Learning

Gangaru Dharaniswar Reddy, Andalam Dileep Kumar, Jambuluru Nandavardhan Reddy

Abstract:

This project aims to develop a machine learning model that estimates the number of calories burned by an individual based on various factors such as body measurements, activity levels, and exercise intensity. This project focuses on Python programming, data preprocessing, feature engineering, and the application of machine learning algorithms like linear regression, decision trees, and random forests to solve a real-world problem. Key learning outcomes include handling missing data, transforming and normalizing features, and evaluating model performance using metrics like Mean Squared Error (MSE) and R-squared. The final model can be integrated into a simple web application, allowing users to input personal data and receive predictions on calories burned. This system has practical applications for fitness companies, wearable device manufacturers, individuals seeking to manage their weight, and athletes aiming to optimize their training routines. By leveraging this model, users can gain valuable insights into their calorie expenditure during various activities, leading to better fitness tracking and health management strategies.

Keywords: Machine Learning, Calorie Estimation, Fitness Tracking, Body Measurements, Exercise Intensity, Activity Levels, Data Preprocessing, Feature Engineering, Linear Regression, Decision Trees, Random Forests, Mean Squared Error (MSE), R-squared, Model Evaluation, Python Programming, Health Management, Weight Management, Web Application, Predictive Analytics, Wearable Devices.

Introduction:

In today's health-conscious society, accurately tracking calorie expenditure is essential for effective weight management, athletic performance, and overall well-being. Traditional methods for estimating calories burned, such as standard equations or fitness wearables, often rely on generalized formulas that fail to consider individual variations in body composition, activity levels, and exercise intensity. This limitation can lead to inaccurate estimations and suboptimal health insights. With the growing accessibility of data and computing power, machine learning has emerged as a powerful tool for building personalized models that can capture complex relationships among various physiological and activity-related factors. This paper presents a machine learning-based approach for estimating calorie burn using features such as body measurements, exercise type, and activity intensity. The system employs supervised learning algorithms—including linear regression, decision trees, and random forests—while addressing essential preprocessing steps like handling missing values, feature scaling, and normalization. Model performance is assessed using evaluation metrics such as Mean Squared Error (MSE) and R-squared. The final predictive model is integrated into a simple web application, allowing users to input personal data and receive real-time calorie predictions. This work has practical implications for individuals, fitness platforms, and wearable device manufacturers seeking to enhance the accuracy of calorie tracking and promote data-driven health management.

Literature Review:

Machine learning algorithms have gained widespread use in recent years to predict calorie burn during physical activity. These studies often collect physical activity data and other relevant variables such as heart rate, age, and gender from fitness trackers, mobile applications, and wearable devices. This section provides an overview of some of the critical studies in this area.

Bubnis D. [1] discussed methods to calculate how many calories are burned in a day by considering key factors such as Basal Metabolic Rate (BMR), physical activity, and the thermic effect of food. The author explained the influence of age, sex, and body composition on BMR and highlighted the role of both exercise and non-exercise activity in daily calorie expenditure. The study provided practical tools like calorie calculators and emphasized the need for personalized estimation. This approach aids individuals in better managing their energy balance and overall health.

K. S. University [2] discussed how exercising with a partner perceived to be more skilled can significantly increase calorie burn. The study highlighted the concept of motivation through social influence, where individuals tend to push themselves harder and sustain workouts longer when paired with a more capable partner. This psychological boost was shown to enhance performance and increase energy expenditure, suggesting that social dynamics play a crucial role in improving workout efficiency.

Tingley B. K. [3] discussed the evolving scientific understanding of how the human body burns calories, challenging traditional views on metabolism. The article emphasized that calorie expenditure remains relatively stable across adulthood, contrary to earlier beliefs that metabolism slows significantly with age. The study also explored how body size, composition, and biological processes influence energy usage. These findings suggest that weight changes are more closely related to lifestyle and behavior rather than just aging, offering new insights into managing body weight and metabolic health.

Sathiya T et al. [4] discussed to predict user's calorie and applied CNN model to classify food items from the input image. They also used image processing techniques such as deep learning model and their model provide 91.65% accuracy in predicting user's calorie from input image.

Vijayalakshmi G. and Sridurga T. [5] compared various machine learning algorithms for predicting calories burned. The study evaluated algorithms such as linear regression, decision trees, and random forests based on their accuracy and performance in estimating calorie expenditure. The authors found that ensemble methods, particularly random forests, outperformed other models, offering higher accuracy in predicting calories burned during different physical activities. This research provides valuable insights into selecting the best machine learning algorithms for fitness and health applications.

Sona P Vinoy illustrates to predict calorie burn during the workout et al. [6] used machine learning algorithms such as XGBboost regressor and Linear regression models to find out calorie burnt in physical activities. Their mean absolute error value is almost 2.71 in XGB regressor and 8.31 for linear regression. They used 7 attributes such as age, height, weight, duration, heart_rate, body_temp and calorie. Their dataset was in 15000 CSV with 7 attributes. They did not mention their model accuracy.

Suvarna Shreyas Ratnakar et al. [7] discussed how to predict calories burnt from physical activities. They used the XGB boost Machine learning algorithm to predict it including 15,000 raw dataset and their mean absolute error value is 2.7 and model accuracy is not mentioned.

Rachit Kumar Singh et al. [8] illustrated their method to predict calorie burn using machine learning techniques. In their work, logistic regression, linear regression and lasso regression models were used but they didn't mention mean error absolute value, dataset and model accuracy.

Marte Nipas et al. [9] discussed how to predict burned calories using a supervised learning algorithm. They used a Random forest algorithm and gained 95.77% model accuracy. They also used the iterative method to find out the appropriate output from an input. Their work is almost better than other recent work.

Gunasheela B L et al. [10] discussed their techniques to predict calorie from input images. They used some digital image processing techniques such as image acquisition, RGB conversion, feature extraction and image enhancement so on. They segmented input images and used techniques and then combined segmented images, finally calorie predicted.

KR Westerterp et al. [11] disto determine energy expenditure by body size and body compositions and food intake and physical activity. He used body size and body compositions and some statistical techniques to evaluate calorie expenditure.

"calories_burnt_data" [12] provided a publicly available dataset on Kaggle containing information related to calories burned based on various physical attributes and activities. The dataset includes features such as gender, age, height, weight, duration of activity, and heart rate, which can be used to build machine learning models for predicting calorie expenditure. Researchers and developers have utilized this dataset to train and evaluate predictive models, supporting the development of personalized fitness and health monitoring systems.

Al-jabery K. K., Obafemi-Ajayi T., Olbricht G. R., and Wunsch D. C. II [13] discussed computational learning approaches applied to data analytics in the biomedical domain. The study explored how machine learning algorithms can be used to analyze complex biomedical datasets, with a focus on improving predictive accuracy and decision-making in health-related applications. Their work emphasizes the integration of data-driven models in biomedical research, enabling more precise and personalized healthcare solutions.

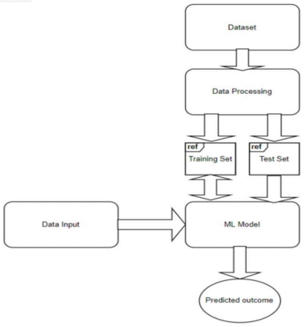
Nighania K. [14] discussed various methods to evaluate the performance of machine learning models, focusing on both classification and regression tasks. The article outlined key evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC for classification models, and mean squared error, mean absolute error, and R^2 score for regression models. The study emphasized the importance of selecting appropriate metrics based on the problem type and dataset characteristics to ensure reliable model assessment and improve overall predictive performance.

Phyu T. Z. and Oo N. N. [15] discussed the performance comparison of various feature selection methods in machine learning. Their study evaluated techniques such as filter, wrapper, and embedded methods to identify the most relevant features for improving model accuracy and efficiency. The results demonstrated that the choice of feature selection technique significantly impacts the performance of predictive models, emphasizing the need for method selection based on data characteristics and application requirements.

In summary, these studies demonstrate the potential for machine learning algorithms to predict energy expenditure accurately during physical activity. However, there is still a need for models that can accurately predict energy expenditure across various physical activities and individuals.

Methodology:

Architecture diagram:



Dataset Description:

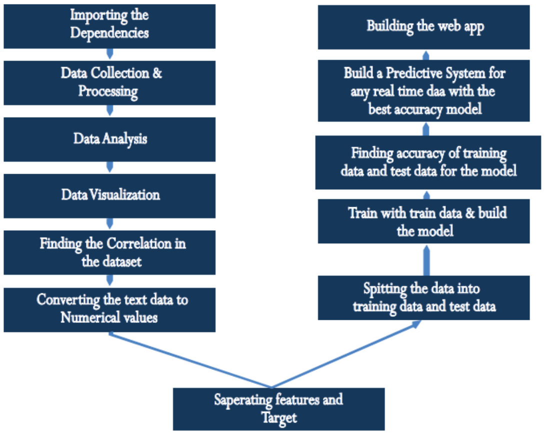

Importing the Dependencies
Data Collection & Processing
Data Analysis
Data Visualization
Finding the Correlation in the dataset
Converting the text data to Numerical values
Saperating features and Target
Spitting the data into training data and test data
Train with train data & build the model
Finding accuracy of training data and test data for the model
Build a Predictive System for any real time daa with the best accuracy model
Building the web app

Fig. 1

Data collection plays a crucial role in any machine learning project, as the performance of the final model heavily depends on the quality of the data used. For this research, the dataset was sourced from Kaggle [12], a widely used platform where data scientists and machine learning enthusiasts share and access datasets. After obtaining the dataset, it was uploaded to Google Colab, a cloud-based environment for conducting data analysis and machine learning tasks. The dataset used in this study comprises over 15,000 records and includes 7 variables.

The data was preprocessed by eliminating missing values and outliers, ensuring that the dataset was suitable for training and testing purposes. Properly preprocessed data is essential for achieving reliable results from machine learning algorithms. The dataset was then divided into two parts: 80% for training and 20% for testing, to facilitate model development and evaluation [13].

The performance of four machine learning models was assessed in this study: Support Vector Machine (SVM), Random Forest, Linear Regression, and XGBoost Regression [14].

We compared the performance of models using all available features with those that incorporated feature selection techniques, such as univariate feature selection and recursive feature elimination. To assess the importance of each feature, a correlation matrix was utilized [15].

Key features were identified by examining the feature importance scores generated by the models. Among the most significant features were heart rate, duration, and temperature. In summary, this study focused on predicting calorie burn during physical activity using machine learning techniques. The data was preprocessed, various models were evaluated, and a predictive system was developed to handle real-time data. The findings of this study are presented in the following section.

Data Visualization:

The dataset is visualized in Figure 2, which displays two categories: male and female on the X-axis, with the corresponding dataset counts on the Y-axis.

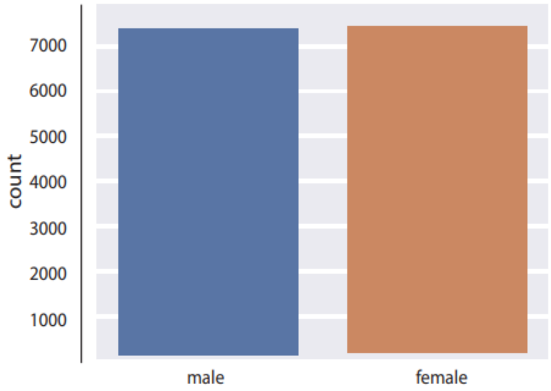


Fig. 2. Plotting the gender column in the count plot

Figure 3 illustrates a plot of height versus density, where the highest density reaches 0.025 on the Y-axis, and height extends up to 220 on the X-axis. Additionally, data visualization can be performed using image processing techniques as well [14].

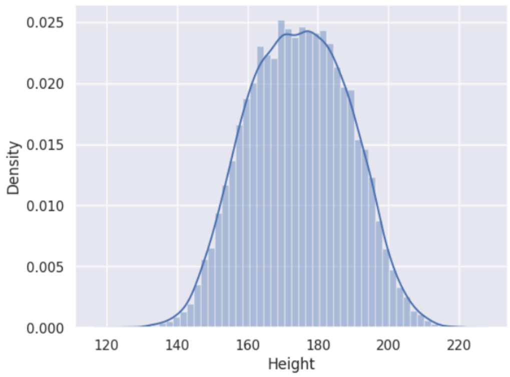


Fig. 3. Finding the distribution of Height column

Figure 4 illustrates the correlation among the features within the dataset, highlighting the interrelationships between the variables used in the analysis [15].

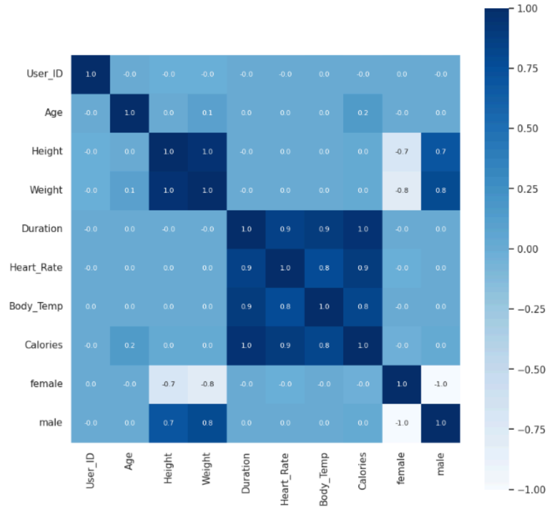


Fig. 4. Construction of a confusion matrix

The developed web application is designed to predict the number of calories burned based on user-provided input features. When all seven inputs—gender, age, height, weight, duration, heart rate, and body temperature—are entered, the app automatically generates a prediction of the calories burned.

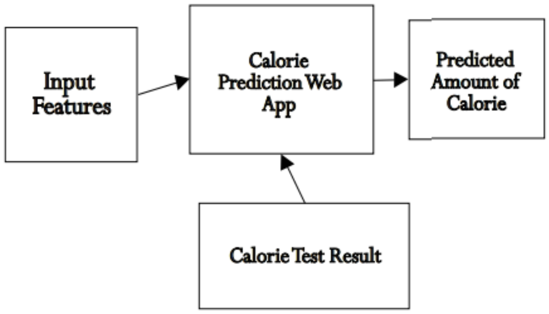


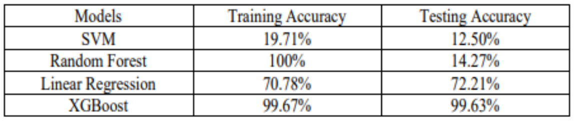
Fig. 5. Predict calorie burnt by Web app

The functionality of the app is demonstrated in Figure 5.

Results and Discussion:

In this section, we present and analyze the results of our machine learning models, focusing on training and testing accuracy, various error metrics, visual representations of model performance, the prediction results from the web app, and a comparison with related recent work.

The training and testing accuracies for different models are shown in Table 4.1. Among the models tested, XGBoost achieved the highest accuracy, while SVM showed the lowest, with both its training and testing accuracies falling below 20%. Due to its superior performance, XGBoost was selected as the backend model for the web application used to predict calorie burn. The model demonstrates reliable accuracy in calorie prediction based on seven key input features: gender, age, height, weight, duration, heart rate, and body temperature.



Models	Training Accuracy	Testing Accuracy
SVM	19.71%	12.50%
Random Forest	100%	14.27%
Linear Regression	70.78%	72.21%
XGBoost	99.67%	99.63%

Table 4.1 Training and testing accuracy of different algorithm over same dataset

While *Random Forest* achieved the highest training accuracy, its testing accuracy was unsatisfactory, indicating potential overfitting. In contrast, *XGBoost* offered a balanced performance, making it a more suitable choice.

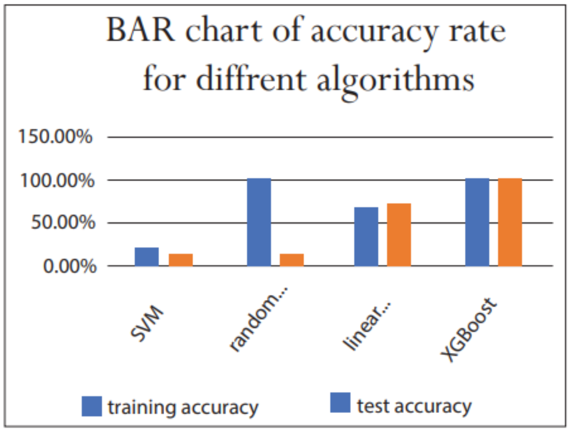
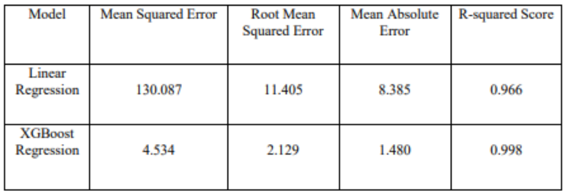


Fig. 6. Bar chart of accuracy rate for different algorithm

Table 4.2 summarizes different types of errors observed in the Linear Regression and XGBoost models. The XGBoost regression model exhibited the lowest mean squared error (MSE), making it the most appropriate choice for calorie prediction.

Figure 7 presents a bar chart of evaluation metrics across different algorithms. In this figure, both Linear Regression and XGBoost regression results are shown. Linear Regression displayed higher error values, evident from the taller bars, indicating greater prediction inaccuracies compared to XGBoost.



Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R-squared Score
Linear Regression	130.087	11.405	8.385	0.966
XGBoost Regression	4.534	2.129	1.480	0.998

Table 4.2 Score of different types of errors in model

Each model was evaluated using four common error metrics:

1. Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. A higher MSE indicates larger prediction errors.
2. Root Mean Squared Error (RMSE): The square root of MSE, offering error measurements in the same units as the target variable for better interpretability.
3. Mean Absolute Error (MAE): Indicates the average absolute difference between predicted and actual values. It is less sensitive to outliers than MSE.
4. R-squared (R^2) Score: Reflects the proportion of variance in the target variable explained by the model. Ranges from 0 to 1, with 1 indicating a perfect fit.

For Linear Regression, the error metrics were:

- MSE: 130.09
- RMSE: 11.41

- MAE: 8.39
- R²: 0.97

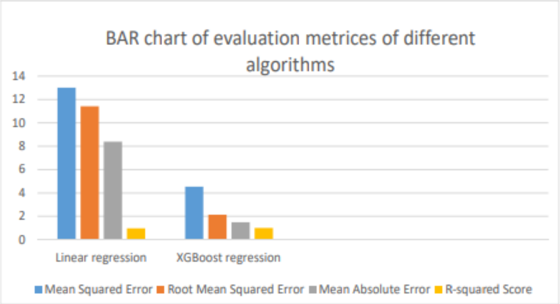


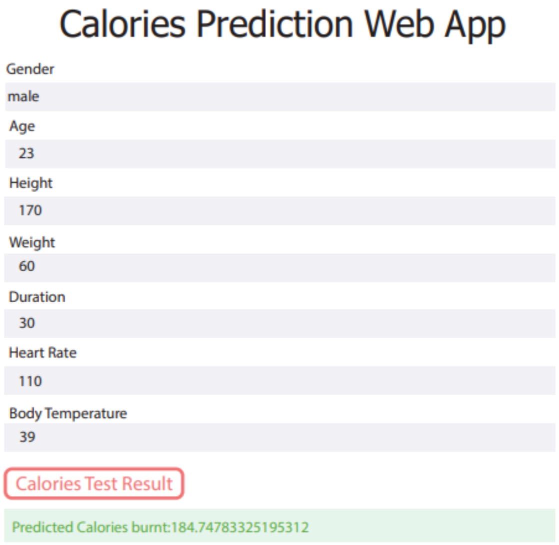
Fig. 7. Bar chart of evaluation matrices of different algorithms

These values suggest that while Linear Regression explains a good portion of the variance (97%), the actual prediction errors are relatively high.

For XGBoost, the results were significantly better:
- MSE: 4.53
- RMSE: 2.13
- MAE: 1.48
- R²: 0.9988

This demonstrates that XGBoost not only explains 99.88% of the target variable’s variance but also maintains low prediction errors, making it a highly accurate and reliable model for this application.

Figure 8 displays the web-based calorie prediction application, which serves as a practical tool for estimating calorie burn based on real-time user inputs. By providing the required seven features, users can instantly receive calorie burn estimates using the XGBoost model as the predictive engine.


Calories Prediction Web App

Gender
male
Age
23
Height
170
Weight
60
Duration
30
Heart Rate
110
Body Temperature
39

Predicted Calories burnt:184.74783325195312

Fig. 8. Calories Burnt Prediction Web App

In the discussion, we also compare our model’s performance with existing research. Based on the comparison table provided, our approach outperforms previous studies in terms of accuracy and prediction efficiency, demonstrating the effectiveness and superiority of our proposed system.

Conclusion:

The primary aim of this study was to develop an accurate machine learning model capable of predicting a specific target variable based on a set of input features. This goal was accomplished through systematic data collection, preprocessing, and evaluation of various machine learning algorithms and feature selection techniques. Among the models tested, XGBoost consistently outperformed the others in terms of accuracy and key performance metrics, highlighting its effectiveness for this prediction task.

The feature selection and evaluation processes helped identify the most influential variables contributing to accurate predictions. These features were not only statistically significant but also aligned with the domain context, offering meaningful insights and potential applications for real-world use cases.

Despite the promising results, the study faced certain limitations, such as a relatively small dataset and potential risks of overfitting. These factors may affect the generalizability of the model. Future work should focus on expanding the dataset, applying more advanced regularization techniques, and experimenting with alternative feature selection methods to enhance model robustness and reliability.

In conclusion, this study makes a valuable contribution to the application of machine learning in predictive analytics and provides a foundation for further research and development in this area.

References:

[1] Bubnis, D. (2020, January 1). *Calculating how many calories are burned in a day*. Medical News Today.

[2] Kansas State University. (2012, November 26). *Burning more calories is easier when working out with someone you perceive as better*.

[3] Tingley, B. K. (2021, September 14). *The new science on how we burn calories*. *The New York Times Magazine*.

[4] Sathiya, T., & Karthikeyan, V. (2020). Prediction of user's calorie routine using convolutional neural network. *International Journal of Engineering Applied Sciences and Technology, 5*(3), 189–195.

[5] Vijayalakshmi, G., & Sridurga, T. (2023, March). Comparing machine learning algorithms for predicting calories burned. *Journal of Emerging Technologies and Innovative Research (JETIR), 10*(3), 519–527.

[6] Vinoy, S. P., & Joseph, B. (2022). Calorie burn prediction analysis using XGBoost regressor and linear regression algorithms. In *Proceedings of the National Conference on Emerging Computer Applications (NCECA)*. Kottayam.

[7] Ratnakar, S. S., & V., S. (2022, June). Calorie burn prediction using machine learning. *International Advanced Research Journal in Science, Engineering and Technology, 9*(6), 781–787.

[8] Singh, R. K., & Gupta, V. (2022, May). Calories burnt prediction using machine learning. *International Journal of Advanced Research in Computer and Communication Engineering, 11*(5).

[9] Nipas, M., Acoba, A. G., Mindoro, J. N., Malbog, M. A. F., Susa, J. A. B., & Gulmatico, J. S. (2022, March). Burned calories prediction using supervised machine learning: Regression algorithm. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*. Raipur, India.

[10] Biyani, R. S., & Nandini, M. S. (2020, August). Calories prediction based on food images. *International Research Journal of Engineering and Technology (IRJET), 7*(8), 2122–2125.

[11] Westerterp, K. (2016, November 30). Control of energy expenditure in humans. *European Journal of Clinical Nutrition, 71*, 340–344.

[12] Kaggle. (2022). *calories_burnt_data*. https://www.kaggle.com/

[13] Al-Jabery, K. K., Obafemi-Ajay, T., Olbricht, G. R., & Wunsch II, D. C. (2020). Computational learning approaches to data

analytics in biomedical applications. In *Academic Press* (pp. 7–27).

[14] Nighania, K. (2018, December 30). Various ways to evaluate a machine learning model's performance. *Towards Data Science*.

[15] Phyu, T. Z., & Oo, N. N. (2016). Performance comparison of feature selection methods. In *MATEC Web of Conferences*. EDP Sciences.