REVOLUTIONIZING WORKOUT ANALYTICS: MACHINE LEARNING MODELS FOR CALORIE BURN ESTIMATION

M.S. Bhuiyan[1], M.N.H. Likhon[1] , A.K.M.A. Habib[1], M.A. Fahim[1] and A.Z. Apurba[1,*]

[1]North South University, Dhaka, 1229, Bangladesh

[*]Corresponding Author's Email: afifa.zain@northsouth.edu

Article History: Received September 16, 2024; Revised October 10, 2024; Accepted October 15, 2024

ABSTRACT: Caloric estimation during exercise plays an integral role in customizing one's fitness regime. Individual differences are a major hindrance to the accurate determination of energy expenditure. This study aims to use machine learning tools to enhance the prediction of calories burned. Method: Five models (KNN, DT, AB, SVM, XGB) were trained with default values and optimized hyperparameters. This study emphasized hyperparameter tuning to achieve optimal results using any given model. Regarding fitness analytics, XGBoost has shown promising RMSE values of 2.13 and R^2 values of 1.00, signifying the efficacy of a machine-learning approach in fitness analysis. This research shows that it is possible to apply machine learning to forecast calorie loss from individual data, hence improving fitness and health programs.

KEYWORDS: *AdaBoost; XGBoost; hyperparameter optimization; feature importance.*

1.0 INTRODUCTION

Heat energy is often defined in terms of calories, which is needed to raise the temperature of 1g of water by 1 degree. Food and drinks provide an essential source of calories for our body, which are then metabolized together with oxygen to usher out energy. Physical activity plays a significant role in keeping us fit in our everyday chores. While exercising, there is a need for more oxygen, leading to increased heart rate and blood circulation [1]. So as to get viable muscle energy, oxygen is sent by blood pumped through the heart and arteries into them, getting rid of carbon dioxide produced after use.

Consequently, bodily temperature increases alongside sweating due to heat loss. Some other things that can affect calorie expenditure during training are age and the duration of their exercises. Additional investigations show that only 20 per cent of grown-ups do sufficient exercise routinely[2]. The human body is adversely affected by excess calories since they can lead to several diseases like coronary artery disease or diabetes[3]. If people eat balanced meals and participate in physical activities, they might be fit and live healthy lives. One way to gauge how many calories you burn is by relying on smart device estimates despite it being impossible to measure this percentage accurately. However, depending on the unique product, it is possible to get variations in accuracy. We intend to use previously crafted answers and a machine learning model to enhance an individual's skill in determining the number of calories expended during different exercises.

Aldraimli and his companions [4] used discretized measures of visceral fat to predict the risk of diseases related to it. The dataset for this analysis, obtained from UK Biobank access, consisted of 8,453 individual data points with 4,327 females and 4,126 males. The dataset contained individuals aged 40 to 70 years, but it needed to be more representative since it was a real-life data set. To overcome the problem of unbalanced data, they applied regression methods on SMOTE. Later on, six machine-learning methods were used for disease prediction. Among all tested models, the RF model stands out for having the greatest "True Positive Rate" with 79.3% men and 85% of women suspected of visceral fat-related diseases.

Using machine learning, Kaur and his team [5] were the first to guess about obesity risk and recommend meals for reducing the obese. They gathered two datasets: a) an obesity-predicting one from the UCI ML respiratory consisting of 2111 instances and (b) an open-source website with 93 meal plans for nutrition. They used different supervised algorithms to predict obesity and an unsupervised nearest neighbour method to predict nutrient meals. The gradient boosting classifier model recorded the highest accuracy of around 0.9811 based on the data ratio (90:10).

In addition, Manjunathan and his colleagues [6] implemented a machine-learning model that predicts calories burnt during workouts through different feature selection techniques. This paper used the exercise dataset by the UCI machine learning storage unit, which comprises 15,000 people's details and has eight independent and one

dependent feature ("Calories"). They applied various machine learning models to which decision trees and gradient boosting performed best, with almost equal accuracy at about 0.99 before and after scaling features.

Nipas et al. [7] performed machine learning to predict burnt calories, where heart rate was the most prominent feature. The authors selected a pure dataset comprising eight numeric variables and a single categorical feature. This dataset was downloaded from Kaggle. They then preprocessed the data to check for similar or null values. This study utilized three machine learning models with varying accuracy; the Random Forest regressor had an optimal accuracy of 95%.

This paper presents a calorie burnt prediction system using a Machine-Learning Model. The datasets collected from Kaggle include personal physical activities and regular activity information which includes Kaggle. Handling missing data and replacing it with appropriate values is done during the preprocessing stage of the dataset. Once this phase is done, five machine-learning algorithms might be used to predict the output. All models are initially trained with their default hyperparameters before being optimized. Each model is evaluated for its predictions.

The structure of this document is as follows: The second section expounds on the proposed system along with various kinds of tables and dataset details. Also discussed in this section are methods and processes used in data preprocessing. The last predicted outcome and all discourse on findings are done in Section III. Section IV presents conclusions with citations to previous works.

In summary, this research adopts a gradual approach by meticulously working with the suggested system. Thus, the presented outcomes are demonstrated, and subsequent discussion includes recommendations for further studies.

2.0 PROPOSED SYSTEM

In this study, calorie consumption predictions are made using various machine learning methods that follow an explicit step-wise process.

Data Preparation: The beginning of data preparation involves handling the dataset's missing values. These steps could involve

methods like imputing or deleting missing data points. Moreover, preprocessing is done on the dataset so that the input features and target variables are in the right format for analysis.

Feature Selection: For instance, the system may influence input attributes that affect calorie burning. This step could be achieved by using different approaches, such as correlation analysis, feature importance, or locality of the area, to identify essential characteristics.

Model Training and Evaluation: The system uses a number of machine learning methods to train prediction models based on the given dataset. Some include K-Nearest Neighbors (KNN), SVM, and Decision Tree, AdaBoost and XGBoost. Each model's performance is compared against both default and improved hyperparameters.

Performance Evaluation: Within this context, measures such as mean square error (MSE), mean absolute error (MAE), or R-squared can be utilized to assess the performance of these trained models in their capacity to predict calorie loss. It provides a measure of the precision or accuracy of any given model when it comes to forecasting.

Hyperparameter Tuning: The system stresses the importance of tweaking hyperparameters to improve model accuracy. Changing hyperparameters, such as adjusting how many neighbours in KNN or changing the number of trees in Random Forest, can give the models a better accuracy rate.

This paper shows that these proposed machine learning algorithms that use individual characteristics and training data could accurately predict an individual's calorie expenditure. In addition, correctly estimating the calorie intake enables personalized health and fitness programs to be developed, allowing people to modify their workout and diet for optimum results.

2.1 DATASET

The data presented in this study, as shown in Table 1, suggested a solution obtained from the Kaggle website [8]. It contains 15,000 records with nine variables. Of these, eight are numerical, while one is categorical. It is important to note that the unprocessed data obtained from the website had no repeated rows or absent values [9].

Table 1: Summary of parameters and their value in the dataset used

Parameter	Value
Numeric variable	8
Categorical variable	1
Number of Observation	15000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

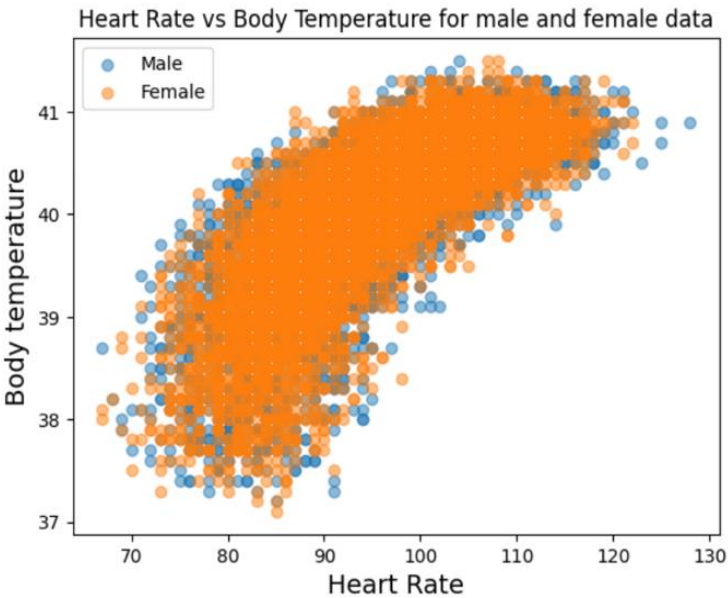

Figure 1: Heart rate vs. Body temperature for male and female data.

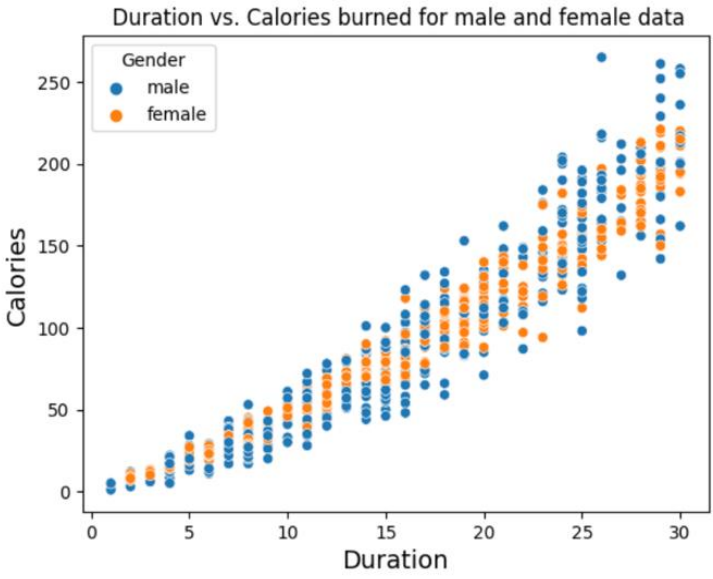

Figure 2: Duration vs. Calories Burned for male and female data

Figure 1 illustrates Heart rate vs. Body temperature for male and female data, whereas Figure 2 illustrates the calories males and females burn regarding their workout duration. Figure 1 and Figure 2 show the relationship between heart rate, body temperature, and exercise duration in the two figures for both males and females. They also show the differences in calorie expenditure concerning gender and workout duration. This situation highlights the need for specific fitness prescriptions for each individual.

2.2 DATASET PREPROCESSING

This solution is based on a Dataset of 15000 observations, including nine attributes [8]. The dataset was created using a data frame on Google Colab IDE, and various preprocessing techniques were employed afterwards. First, we have made some checks for duplicates in the dataset, but luckily, no duplicate values were found. Secondly, null checks were carried out; again, no null values were available in any of the features. The dataset contained a categorical variable called 'Gender' that was changed to a numeric variable to improve its efficiency during the processing. The next step was to separate the dataset into features and target variables. We used the necessary libraries from sci-kit-learn to split the final dataset into 'test_train_split'.

Different EDA methods were applied to detect outliers from the original data sets. At first, basic boxplots were drawn using visualizer libraries such as 'Seaborn' or 'pyplot' as per the features. In addition, this analysis pointed out outliers in attributes like 'Height', 'Weight', 'Duration', 'Body-temperature', 'Heart rate' and 'Calories'. Furthermore, we plotted KDE graphs for Ages, Density, Gender and Calories vis-a-vis their corresponding freaks.

Also, we have utilized histograms, swarm plots, and strip diagrams to represent the anomalies and outliers in the dataset. Applying these techniques helped us better comprehend the dispersion of the data points.

2.3 Machine Learning Models

The term "machine learning model" refers to algorithms enabling computers to identify patterns and make predictions or judgments without explicit coding. These models use past data to generate generalizations and forecasts for future, previously unseen data. The models:

1. XGBoost: XGBoost is often perceived as a superlative gradient-boosting algorithm known for its excellent outcomes. It slowly builds a set of weak predictors that are generally tree-based. To improve the effectiveness of models, XGBoost relies on regularization and gradient descent optimization techniques.
2. Decision tree: A Decision Tree typically employs a feature-based branching algorithm to generate a tree-like pattern. The algorithms first travel down this path from the root (the topmost node) to the leaves (the bottommost nodes) to arrive at a value judgment at each point along the way: i.e., it will use different numbers or classes until it reaches the outcome. Users understand it quickly since it functions using both qualitative and quantitative information.
3. SVM: Support vector machine (SVM) is like a jack of all trades in classification and regression tasks. It works by constructing a hyperplane that better divides two or more data sets representing different classes. Moreover, it uses kernel functions that can handle both classes of data that can be separated linearly and others that cannot, thus making it ideal for complex decision boundaries.

4. KNN: KNN is the best algorithm for classification and regression problems. It predicts output by majority consensus from K nearest data points or averages of those points. KNN does not use any parameter but relies heavily upon how similar the points are in the feature space.
5. AdaBoost: AdaBoost combines many weak learners to create a strong learner. Thus, later models can focus on improving predictions because they are assigned higher weights where there has been misclassification. To make this guide more durable and effective, AdaBoost modifies parameters iteratively.

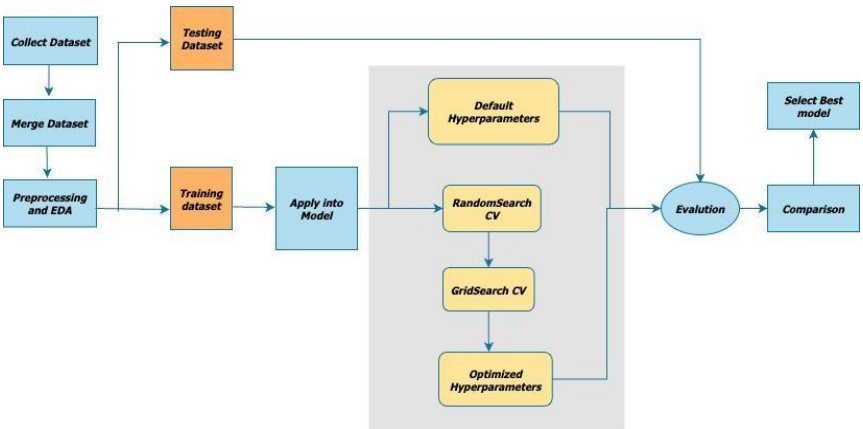

Figure 3: Working sequences of the proposed calories burnt prediction system.

The research procedure is stated in Figure 3 with the flowchart. In Figure 3, the simplified flowchart presents the stages of data collection, preprocessing, model training, and evaluation in clear chronological order. Data collection and merging are performed; after that, preprocessing and EDA are carried out. In the next step of our procedure, we partition the data set into training and testing datasets. A range of models is applied to this data, and their performance is evaluated using default and optimized hyperparameters obtained from random search or grid search cross-validation, respectively. Ultimately, the best model was chosen by comparing the results obtained from these models.

3.0 RESULT AND DISCUSSION

This paper utilized five unique machine learning (ML) models to forecast output. After evaluating each model, XGBoost showed the least MAE, MSE, and RMSE values, the greatest accuracy, and an R2 of 1.00, which suggests a perfect fit for the data. In comparison, the AdaBoost model suggests larger errors when it comes to prediction challenges; for instance, its R2 of 0.39 indicates that it may fail in accurately predicting fundamental issues confronting this study case.. KNN does better than that by committing relatively lower errors and having a high R2 figure(0.99) per the report matrix.

Table 2. Hyperparameter values' ranges for various ml model

Model	Hyperparameter Value Range	Optimized value
SVM	'C': np.logspace(-3, 3, 7), 'gamma': np.logspace(-3, 3, 7), 'kernel': ['linear', 'rbf']	C:10, gamma:1, Kernel: RBF
Decision Tree	min_samples_split = [2, 5, 10, 14] min_samples_leaf = [1, 2, 4, 6, 8] max_features = ['auto', 'sqrt', 'log2'] criterion': ['mse', 'friedman_mse', 'mae', 'poisson'] max_depth: linspace (10, 1000,10)	criterion='friedman_ mse', max_depth=560, max_features=auto, min_samples_leaf=2, min_samples_split=5
KNN	n_neighbors: [1, 3, 6, 8, 11, 13, 16, 19, 21, 24, 26, 29, 31, 34, 37, 39, 42, 44, 47, 50]}, 'weights': ['uniform', 'distance'], 'metric': ['Euclidean', 'Manhattan', 'Chebyshev', 'Euclidean', 'Minkowski'], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'Leaf_size': [20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109, 111, 113, 115, 117, 120]	n_neighbors:20, weights='distance', algorithm= 'ball_tree', Leaf_size=78, Metric= 'Manhattan'
XGBoost	max_depth: [10, 120, 230, 340, 450, 560, 670, 780, 890, 1000], n_estimators: (200, 2000, 100) learning_rate: [0.1, 0.01, 0.05]	max_depth:10, n_estimators:400 learning_rate:0.01
AdaBoost	'n_estimators': [100, 200, 300], 'Learning_rate': [0.1, 1, 10], 'loss': ['linear', 'exponential']	n_estimators: 7 'Learning_rate': [1],

		'loss': ['linear']

Similarly, SVM performance has been relatively better although it still suffers from high error rates compared to what was obtained before. Its R2 coefficient is only 0.94 to indicate how far departed the results were from actual observations. Conversely, even within the confines of decision trees, good performances were noted alongside minimal mistakes; hence, an appreciable range of 0.99 for its R2 coefficient could still show until now.

To conclude, this analysis asserts these characteristics of the models mentioned earlier while concentrating solely on the XGBoost model, whose predictions turned out best amongst them all irrespective of model performance choice since it recorded the least possible values for MSE MAE RMSE points calculation as well as perfect R2 measure observing upon it. Table 2 illustrates the ranges of hyperparameter values and the corresponding optimized hyperparameters for all the ML models. Performance metrics of various ML models with default hyperparameters have been illustrated in Table 3.

Table 3: Performance metrics of various ml models with default hyperparameters

Model	MAE	MSE	RMSE	R2 coefficient
XGBoost	1.48	4.53	2.13	1.00
SVM	10.62	243.29	15.6	0.94
KNN	5.05	51.46	7.17	0.99
AdaBoost	113.3	23966	154.8	0.39
Decision Tree	3.36	27.15	5.21	0.99

Table 4: Performance metrics of various ml models with optimized hyperparameters

Model	MAE	MSE	RMSE	R2 coefficient
XGBoost	1.38	4.85	2.20	1.00
SVM	0.04	0.03	0.06	1
KNN	4.40	41.65	6.45	0.99
AdaBoost	109	21367	146.17	0.45
Decision Tree	3.31	25.32	5.03	0.99

Performance metrics of various ML models with optimized hyperparameters have been illustrated in Table 4. In Table 4, we elaborate on the optimized hyperparameters that inform the figurative

representations used in connection with the models. For example, increasing the number of estimators in XGBoost and adjusting the kernel in SVM, among other changes, results in better model performance.

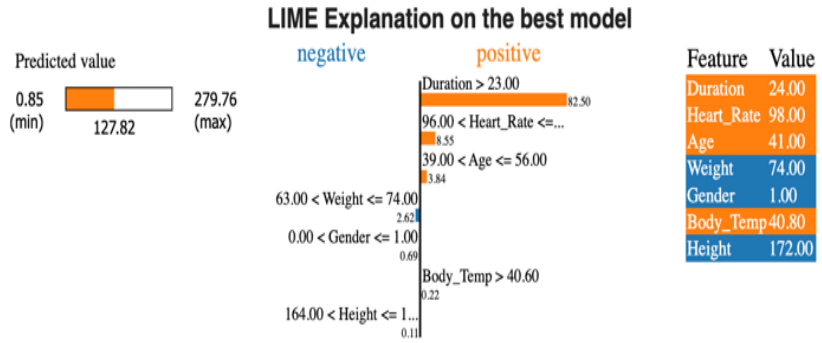

Figure 4: The LIME explainable AI library's machine learning model prediction.

This section discusses the expected performance of the models with real datasets, and the pivotal role of feature selection and engineering is pointed out. Features like heart rate and duration were selected because of their substantial effect on calorie burn, as revealed by the LIME analysis, as shown in Figure 4. LIME AI reveals that the best model's main predictors are duration and heart rate. These two variables play an essential role in the outcome. Therefore, LIME AI insights indicate that duration and heart rate are crucial in the predictive model, giving more analysis and decision-making opportunities.

Table 5. Comparison of the proposed system with existing work

Ref.	Model	Accuracy/RMSE	R2 Coefficient
[7]	XGBoost Regressor	2.71	0.96
[2]	Linear Regression	8.38	0.89
This work	**XGBoost**	**2.13**	**1.00**

Table 5 compares the proposed calorie burn prediction results with our work. This section presents a performance comparison of the proposed system with other existing systems, showing that the accuracy of the proposed XGBoost model is the highest, and the error

rates are the lowest.

4.0 CONCLUSION

Results obtained from predicting calorie burn with various machine learning algorithms utilizing default and adjusted hyperparameters indicate that XGBoost outperforms other models according to its minimal MAE, RMSE values, and the more significant R2 coefficient. The Support Vector Machines (SVMs) also performed outstandingly, with few mistakes recorded.

Although they have more errors, KNN and Decision Tree have captured the main patterns of the dataset. As far as accuracy is concerned, AdaBoost is rated lower than other models. Different ensemble methods can be tried out in subsequent works for further improvements.

The scope of future research may also be extended to use fitness gear and wearables such as heart rate monitors and trackers to harness the potential of real-time physiological data for a more accurate and personalized prediction of calorie burn. This paper is concerned with integrating the above models with wearable fitness infrastructures. It highlights the importance of real-time physiological data in estimating the calories burnt and generating tailored recommendations for fitness regimes.

ACKNOWLEDGMENTS

We extend our heartfelt gratitude to our collaborators for their invaluable support throughout this research.

REFERENCES

[1] R. K. Shing, V. Gupta, "Calories burnt prediction using machine learning model," vol. 11, 2022.

[2] S. P. Vinoy, B. Joseph, "Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms," vol. 4, 2022.

[3] N. Mohanto, N.C. Nurunnabi, "Prevalence and risk factors of general and abdominal obesity and hypertension in rural and urban residents in Bangladesh," 2022.

[4] M. Aldraimli, D. Soria, J. Parkinson, E. L. Thomas, J. D. Bell, M. V.

Dwek, and T. J. Chaussalet, "Machine learning prediction of susceptibility to visceral fat associated diseases," Health and technology, vol. 10, pp. 925–944, 2020.

[5] R. Kaur, R. Kumar, and Gupta, "Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence," pp. 458–469, 2022

[6] N. Manjunathan, M. S. Devi, S. Sridevi, K. K. Bonala, A. Kavitha and K. Jayasree," Feature Selection Intent Machine Learning based Conjecturing Workout Burnt Calories," Turkish Journal of Computer and Mathematics Education, vol. 12, pp. 1729-1742, 2021.

[7] M. Nipas, A. G. Acoba, J. N. Mindoro, M. A. F. Malbog, J. A. B. Susa and J. S. Gulmatico, "Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm," pp. 1-4,2022

[8] https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos

[9] Titu, M. F. S., Emon, M. H. M., Aumi, S. A., Bhuiyan, M. S., Rahman, M. R., & Murshid, M. F. (2024). Kidney Cancerous Tumor Prediction using CNN System Architecture. Asian Journal Of Medical Technology, 4(1), 57-70.