e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

ACCURATE CALORIE BURN PREDICTION WITH MACHINE LEARNING: XGBOOST IN FOCUS

Gurrappagaru Sanjana Reddy*1, Singareddy Velankini Suhas*2

*1Department Of Information Science And Engineering, JSS Academy Of Technical Education,
Bengaluru, Karnataka, India.

*2Department Of Computer Science And Engineering, Vellore Institute Of Technology,
Bhopal, Madhya Pradesh, India.

DOI : https://www.doi.org/10.56726/IRJMETS45191

ABSTRACT

In an age where health and fitness are of increasing importance, accurately estimating calorie expenditure plays a crucial role in enhancing personal well-being. This paper conducts a comprehensive investigation into the use of machine learning, particularly the XGBoost regression algorithm, to achieve precise calorie burn predictions. Our research capitalizes on datasets containing diverse physical activity and biometric data, facilitating the creation of robust predictive models. By conducting a rigorous comparative analysis against alternative regression methods, our study underscores the superior performance of XGBoost in the realm of calorie burn prediction, employing two datasets comprising over 15,000 data points. These predictions are grounded in the metabolic equivalent of task (MET) chart and associated formulas.

Keywords: Comparative Analysis, XGBoost Regression, Regression Performance, Accuracy Assessment, Predictive Modeling, Mean Absolute Error.

I. INTRODUCTION

In the realm of human physiology, calories underpin the measurement of energy expended during specific tasks, forming the fundamental basis for assessing dietary intake, with each food item harboring its distinct calorie content.

Engaging in physical activity induces physiological changes, including elevated body temperature and heart rate, driven by the metabolic breakdown of carbohydrates into glucose, subsequently converted into energy through oxygen utilization. The accurate prediction of energy expenditure necessitates the consideration of a multitude of variables, encompassing exercise duration, average heart rate per minute, temperature, height, weight, gender, and age.

In pursuit of precise calorie burn estimation during physical activity, this research leverages the XGBoost machine learning regression algorithm. This approach integrates exercise duration, temperature, height, weight, and age as input parameters, culminating in the development of a comprehensive model that delivers accurate predictions of calorie expenditure.

II. METHODOLOGY

In the pursuit of accurately estimating an individual's calorie burn, this study embarked on a meticulous process that commenced with the meticulous curation of an appropriate dataset. This dataset served as the cornerstone for training our machine learning models, ensuring that we had the right foundation to make reliable predictions.

The journey towards precise calorie prediction necessitated data preprocessing, a crucial step aimed at enhancing the quality and suitability of the dataset. This involved various operations such as data cleaning and feature engineering to ensure that the data was in an optimal state for analysis. These pre-processing efforts were essential for eliminating noise and inconsistencies within the data, ensuring that our machine learning models could operate effectively.

Once the data had been meticulously prepared, the next phase involved transforming it into meaningful insights. Through the utilization of visualization techniques, the data is organized into comprehensible diagrams and graphs. These visualizations provided a clear representation of trends and relationships within the dataset, facilitating a deeper understanding of the factors influencing calorie burn. Within this context, we employed the XGBoost regressor as our machine learning model. It served as the cornerstone of our analysis, enabling us to not only compare various models but also evaluate their performance rigorously. The combination of a well-curated dataset, diligent data preprocessing, and powerful machine learning techniques paved the way for a robust and accurate estimation of calorie expenditure.

III. MODELING AND ANALYSIS

A. Data Collection:

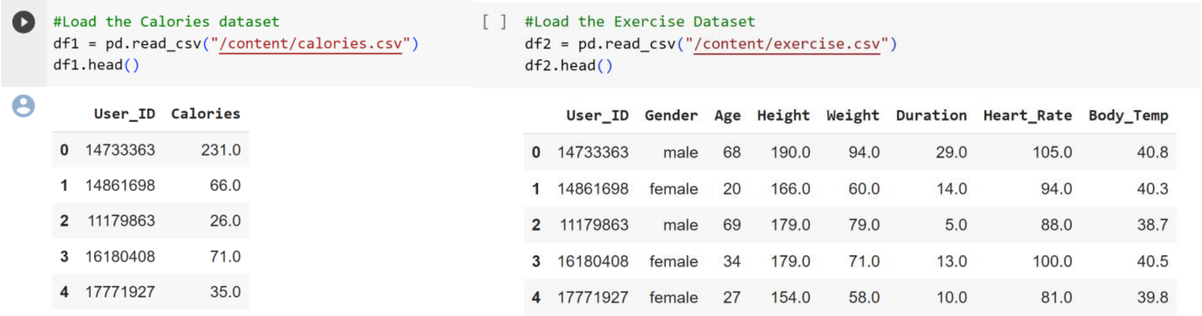
Our research journey commences with the pivotal step of data collection. To assemble the requisite dataset, we turn to Kaggle, a prominent and trusted data repository renowned for its diverse and comprehensive datasets. Specifically, we utilize Kaggle as the primary data source for our study. This rich dataset is subsequently loaded into the Colaboratory (Colab) program, a robust and versatile platform for data analysis. Notably, the data we have retrieved encompasses a wide range of information, comprising both categorical and numerical attributes.

B. Data Preprocessing:

Within this dataset, we discover a treasure trove of valuable information, comprising a substantial collection of 15,000 instances. These instances span across two distinct CSV files, aptly named "exercise.csv" and "calorie.csv." Each entry in this comprehensive dataset corresponds to an individual, encapsulating a wealth of attributes, including but not limited to height, weight, gender, age, exercise duration, heart rate, and body temperature. This meticulously collected dataset forms the bedrock of our analysis, allowing us to delve into the intricate relationships between these attributes and calorie burn.

C. Data Analysis:

In the pursuit of insightful data analysis, we rely on the capabilities of the Colab platform, chosen for its prowess in processing and visualizing complex datasets. To initiate our analysis, we upload the aforementioned dataset files, "exercise.csv" and "calorie.csv," into the platform. During the course of our analysis, we uncover intriguing observations, such as the notable presence of an average body temperature of approximately 40 degrees Celsius. Additionally, it becomes apparent that individuals engaged in physical activity tend to exhibit higher body temperatures, shedding light on the relationship between exercise and physiological responses. Notably, heart rate and core body temperature emerge as pivotal factors during this analytical phase. To enhance the comprehensibility of our findings, we employ various visualization techniques, ranging from tables to charts. Furthermore, we delve into the exploration of correlation types, including positive and negative correlations, to gain deeper insights into the interplay between different data records. Subsequently, we introduce the XGBoost Regressor model, a powerful machine learning tool, and rigorously evaluate its predictive capabilities. This evaluation involves testing the model with a specific dataset and comparing its predicted calorie burn values with the actual values, enabling a thorough assessment of its performance.



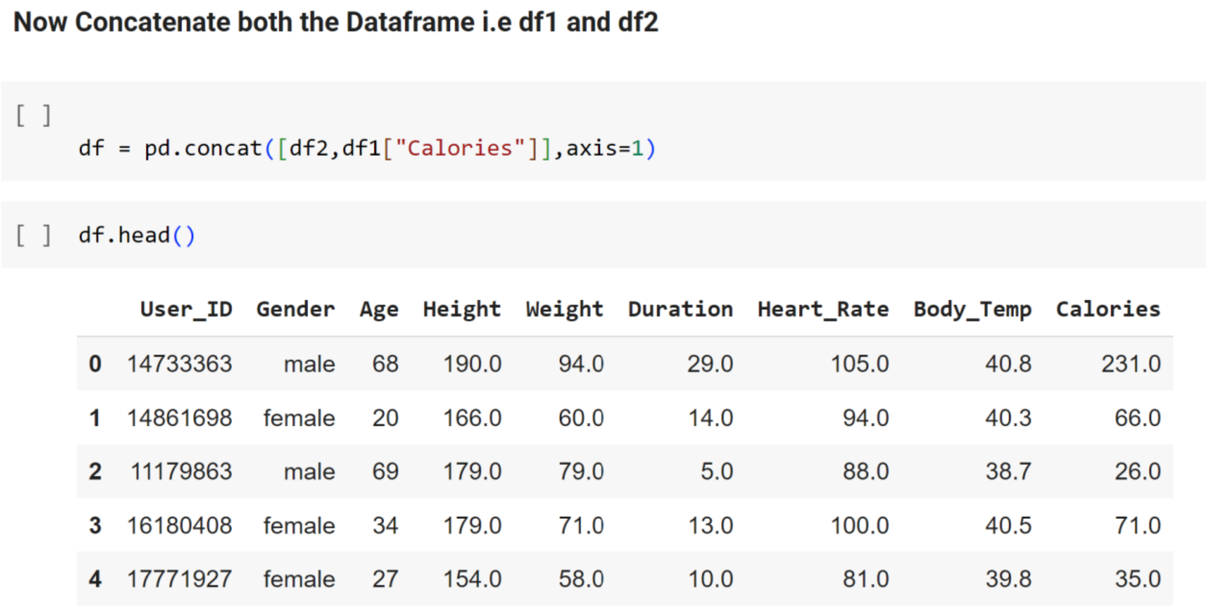
```
#Load the Calories dataset
df1 = pd.read_csv("/content/calories.csv")
df1.head()
```

	User_ID	Calories
0	14733363	231.0
1	14861698	66.0
2	11179863	26.0
3	16180408	71.0
4	17771927	35.0

```
[ ] #Load the Exercise Dataset
df2 = pd.read_csv("/content/exercise.csv")
df2.head()
```

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

Figure 1: df1 and df2, holding two separate datasets.



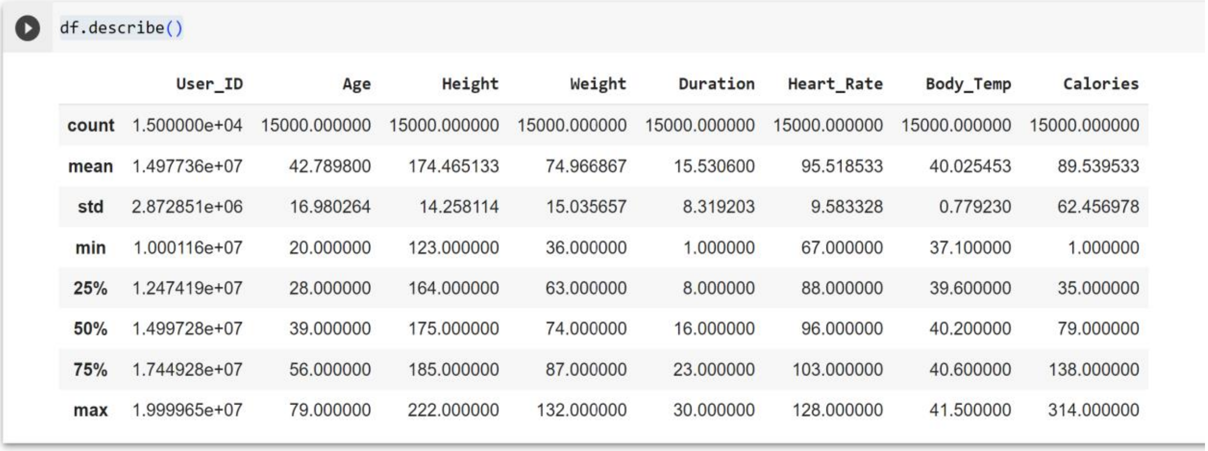
Now Concatenate both the Dataframe i.e df1 and df2

```
[ ]
    df = pd.concat([df2,df1["Calories"]],axis=1)
```

```
[ ]  df.head()
```

	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 2: Concatenating both the dataframes.



```
df.describe()
```

	User_ID	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
count	1.500000e+04	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000
mean	1.497736e+07	42.789800	174.465133	74.966867	15.530600	95.518533	40.025453	89.539533
std	2.872851e+06	16.980264	14.258114	15.035657	8.319203	9.583328	0.779230	62.456978
min	1.000116e+07	20.000000	123.000000	36.000000	1.000000	67.000000	37.100000	1.000000
25%	1.247419e+07	28.000000	164.000000	63.000000	8.000000	88.000000	39.600000	35.000000
50%	1.499728e+07	39.000000	175.000000	74.000000	16.000000	96.000000	40.200000	79.000000
75%	1.744928e+07	56.000000	185.000000	87.000000	23.000000	103.000000	40.600000	138.000000
max	1.999965e+07	79.000000	222.000000	132.000000	30.000000	128.000000	41.500000	314.000000

Figure 3: Descriptive statistics of the numerical columns.

D. Machine Learning Model:

In the realm of machine learning, this stage is pivotal, as it marks the application of our chosen algorithms to estimate the mean absolute error—a critical metric for gauging prediction accuracy. In this instance, we leverage multiple algorithms, including the XGBoost regressor, Linear Regression, Decision Tree Regressor, and Random Forest Regressor, to scrutinize and assess their respective performance levels. Utilizing key performance indicators, we gauge the models' ability to produce accurate predictions, shedding light on the precision of each algorithm's calorie burn estimations. Importantly, the XGBoost regression algorithm has been selected for its demonstrated effectiveness and efficiency in predicting calorie expenditure.

E. Evaluation:

Our analysis of this dataset extends to making predictions regarding calorie expenditure based on exercise duration and a myriad of additional factors, encompassing age, gender, body temperature, and heart rate at various time points during physical activity. The overarching objective is to identify a machine learning model capable of delivering lower mean absolute error, indicative of more accurate and reliable calorie burn predictions. Through the rigorous application of various machine learning methodologies, we aim to enhance our understanding of the intricate relationship between physiological variables and calorie expenditure during exercise, ultimately striving to provide more precise estimations of calorie burn for informed health and fitness decisions.

For a variety of reasons, it becomes evident that the XGBoost Regressor model stands out as the optimal choice for calorie burn prediction:-

1. High Predictive Accuracy:

The XGBoost Regressor exhibits a very high R-squared (R2) score, which indicates that it can explain a significant portion of the variance in the data. This high R2 score suggests that the model's predictions are highly accurate and closely aligned with the actual calorie burn values.

2. Lowest Error Metrics:

The XGBoost Regressor has the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) among the compared models. These metrics quantify the model's prediction errors, and lower values indicate better accuracy. The XGBoost model's consistently lower error metrics imply that it provides more precise predictions of calorie burn.

3. Robust Generalization:

Unlike the Decision Tree Regressor, which achieved a perfect R2 score but may be overfitting the data, the XGBoost Regressor strikes a balance between predictive accuracy and generalization. It performs exceptionally well without showing signs of overfitting, which is crucial for the model's ability to make accurate predictions on new, unseen data.

4. Ensemble Learning:

XGBoost is an ensemble learning technique that combines the predictions of multiple decision trees. This ensemble approach often results in superior predictive performance by mitigating the weaknesses of individual trees. It leverages the collective wisdom of multiple trees to make more accurate predictions.

5. Robustness to Noise:

XGBoost is known for its robustness to noisy data and outliers. It can handle complex relationships in the data while being less susceptible to the influence of individual data points that deviate significantly from the overall pattern. This is particularly valuable in real-world datasets where data quality can vary.

6. Model Tuning:

XGBoost offers extensive hyperparameter tuning capabilities, allowing you to fine-tune the model for optimal performance. This ability to adjust the model's parameters can lead to further improvements in prediction accuracy.

7. Wide Applicability:

XGBoost has been widely adopted in various machine learning competitions and real-world applications due to its exceptional predictive power. Its versatility makes it a suitable choice for complex regression tasks like calorie burn prediction.

8. Scalability:

XGBoost is efficient and can handle large datasets, making it suitable for scenarios where you have a significant amount of data to work with.

IV. RESULTS AND DISCUSSION

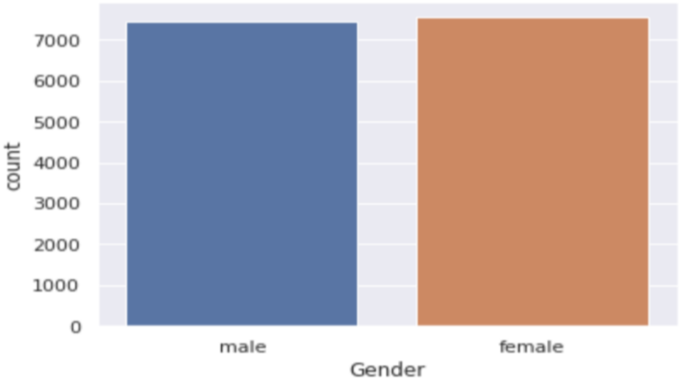


Figure 4: count plot of the "Gender" column in the DataFrame.

The count plot as shown in Fig.4. is a useful way to visualize the distribution of categorical data, such as the number of occurrences of each gender category in the dataset. It provides insights into the balance or imbalance of gender categories within the dataset.

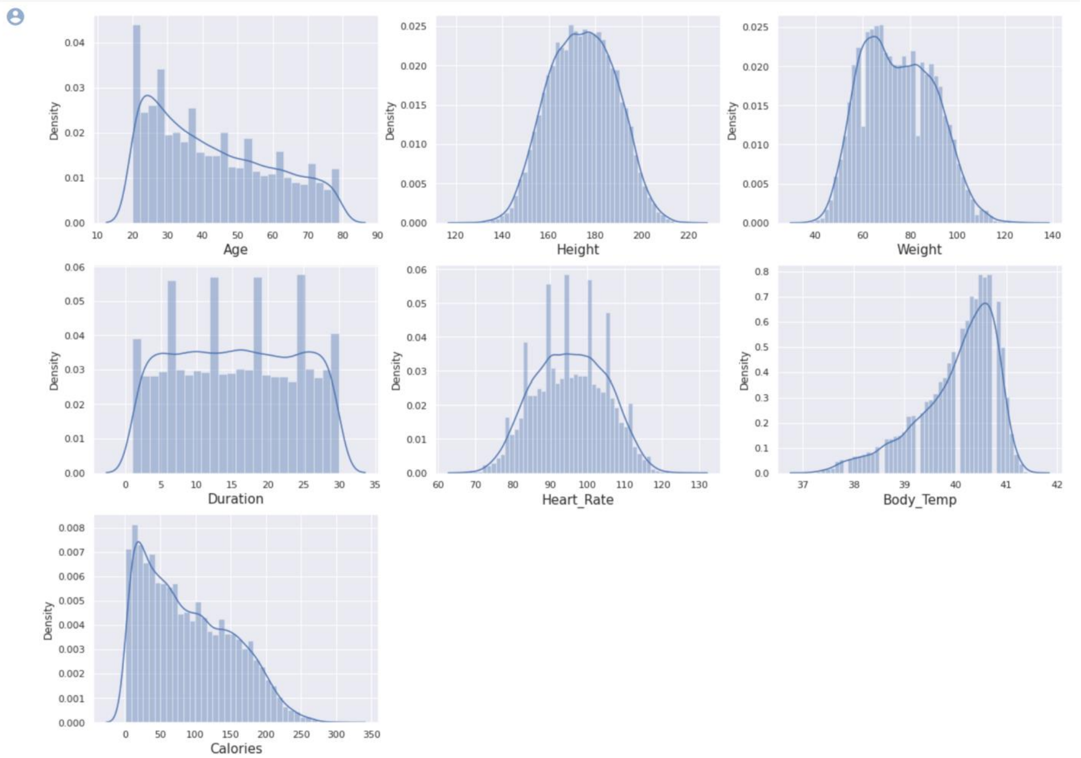
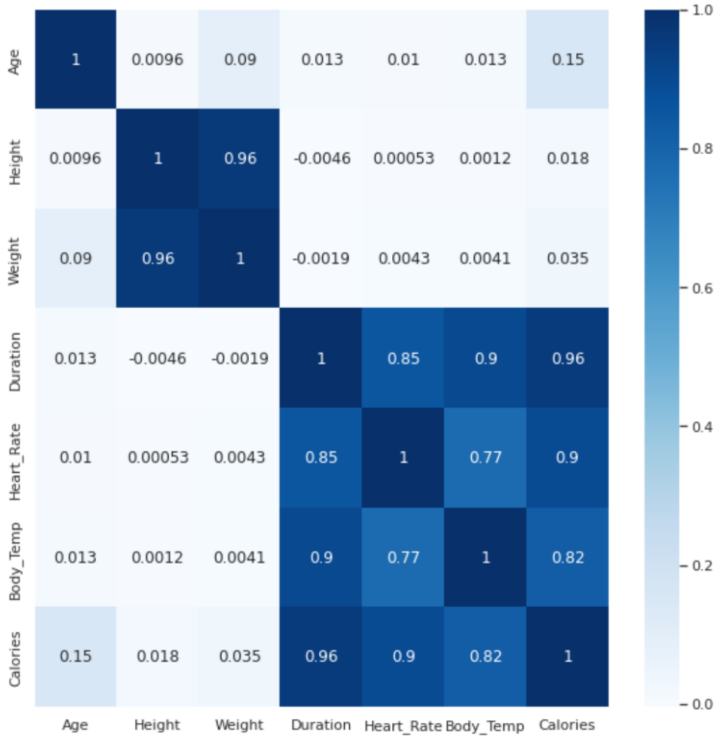


Figure 5: distributions of numerical data in the dataset.

Grid of subplots as shown in Fig.5. is used to display the distribution of numerical columns, which is beneficial for quickly visualizing the distributions of numerical data in the dataset, helping to understand the data's central tendencies and spread.



	Age	Height	Weight	Duration	Heart_Rate	Body_Temp	Calories
Age	1	0.0096	0.09	0.013	0.01	0.013	0.15
Height	0.0096	1	0.96	-0.0046	0.00053	0.0012	0.018
Weight	0.09	0.96	1	-0.0019	0.0043	0.0041	0.035
Duration	0.013	-0.0046	-0.0019	1	0.85	0.9	0.96
Heart_Rate	0.01	0.00053	0.0043	0.85	1	0.77	0.9
Body_Temp	0.013	0.0012	0.0041	0.9	0.77	1	0.82
Calories	0.15	0.018	0.035	0.96	0.9	0.82	1

Figure 6: Heatmap using Seaborn to visualize the correlation matrix of the numerical columns.

The resulting heatmap as shown in Fig.6, visually represents the strength and direction of correlations between numerical features. Positive correlations are shown in lighter shades of blue, while negative correlations are shown in darker shades. This visualization is helpful for identifying relationships between variables in the dataset, which are valuable for feature selection and understanding the data's underlying patterns.

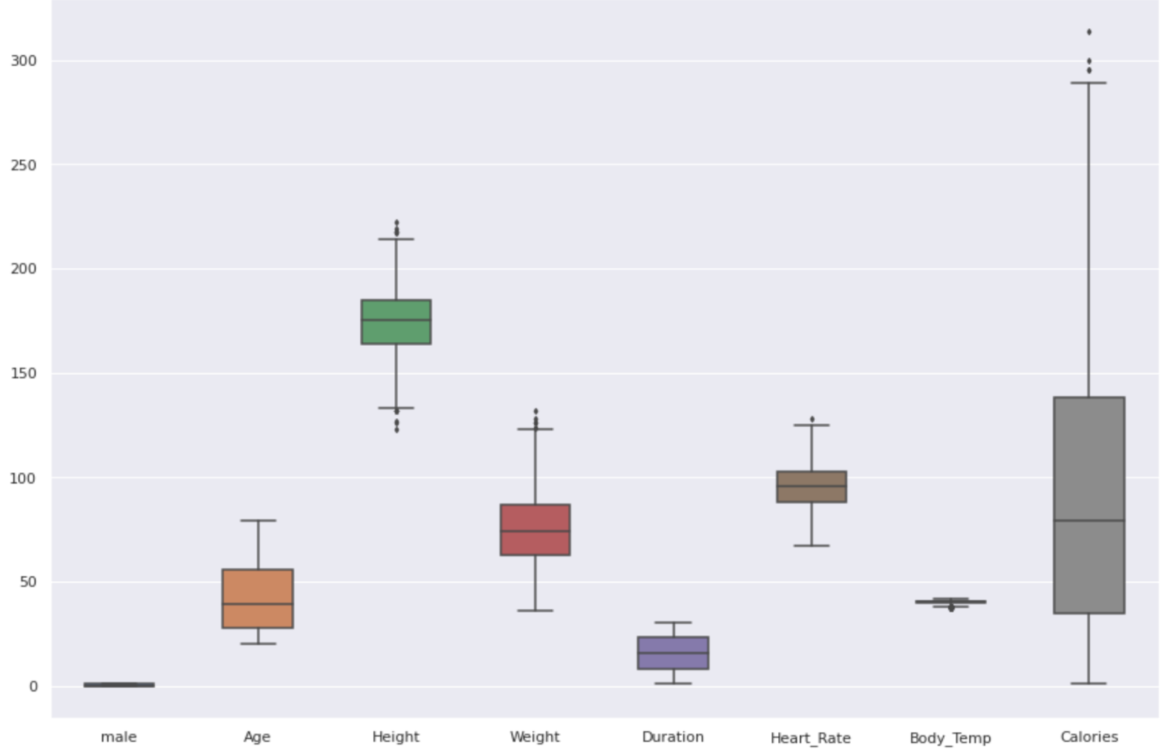


Figure 7: Creating a box plot using Seaborn to visualize the distribution and statistics of numerical data.

The resulting box plot as shown in Fig.7, displays the distribution of numerical data in the “data” dataframe, showing key statistics such as the median, quartiles, and potential outliers for each numerical variable. It is a useful visualization for understanding the spread and central tendency of the data and for identifying any data points.

XGB Regressor

Evaluation of the model's performance using various metrics:-

r2 score(R-squared (R2)) score: 0.9962383848898505

MAE (Mean Absolute Error): 2.748654327124357

MSE (Mean Squared Error): 15.102769892432217

RMSE (Root Mean Squared Error): 3.8862282347325174

Linear Regression

Evaluation of the model's performance using various metrics:-

r2 score(R-squared (R2)) score: 0.9655977245826504

MAE (Mean Absolute Error): 8.479071745987955

MSE (Mean Squared Error): 138.12408611460899

RMSE (Root Mean Squared Error): 11.752620393538157

Decision Tree Regression

Evaluation of the model's performance using various metrics:-

r2 score(R-squared (R2)) score: 0.9926221108057348

MAE (Mean Absolute Error): 3.4726666666666666

MSE (Mean Squared Error): 29.622

RMSE (Root Mean Squared Error): 5.442609668164712

Random Forest Regression

Evaluation of the model's performance using various metrics:-

r2 score(R-squared (R2)) score: 0.99766556152047

MAE (Mean Absolute Error): 1.80648

MSE (Mean Squared Error): 9.372699266666668

RMSE (Root Mean Squared Error): 3.061486447245303

In summary, the XGBoost Regressor excels in terms of both predictive accuracy and generalization ability, as evidenced by its high R2 score, low MAE, MSE, and RMSE, and its avoidance of over fitting. This makes it the best-performing model among the four considered for this regression task.

V. CONCLUSION

In conclusion, the application of machine learning models, including the XGBoost regressor, Linear Regression, Decision Tree Regressor, and Random Forest Regressor, allowed us to evaluate prediction accuracy. Our results demonstrated that the XGBoost regression model consistently outperformed other algorithms, offering the most precise and reliable calorie burn estimations. Furthermore, our research delved into exploring correlations between variables and visualizing data trends, enhancing our understanding of the complex relationships involved. We also identified the key factors contributing to accurate calorie burn prediction.

In the pursuit of optimizing health and fitness monitoring, our study presents a valuable contribution. By accurately estimating calorie burn during physical activity, individuals can make informed decisions regarding exercise and dietary choices, ultimately leading to improved well-being. Our findings emphasize the potential for machine learning models, particularly the XGBoost regressor, in enhancing personalized health management and fitness tracking.

In a world increasingly focused on health and fitness, our research equips individuals with a robust tool for better understanding and managing their calorie expenditure, promoting a healthier and more informed lifestyle. This study highlights the potential for data-driven approaches to revolutionize health and fitness monitoring, providing individuals with the knowledge they need to achieve their wellness goals.

VI. REFERENCES

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University.

[2] MacKay, D.J, & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge University press.

[3] Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM

[4] https://www.kaggle.com/fmendes/fmendesdat263xdemos

[5] https://machinelearningmastery.com/xgboost-for- regression/

[6] https://www.medicalnewstoday.com/articles/319731#factors-influencing-daily-calorie-burn-and-weight-loss

[7] https://zenodo.org/record/6365018

[8] https://devhadvani.github.io/calorie.htm

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5496172

[10] https://www.jetpac.com/

[11] World Health Organization. (2011, October) Obesity Study. [Online]
http://www.who.int/mediacentre/factsheets/fs311/en/index.html