

Project Submission Semester-V

Title of the Project –

Calories Burned Prediction Using ML

Name – Sandesh Dumbre

Roll No. –HBSU2229

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Content** | **Pg.no** |
| **1** | **Acknowledgement** | **3** |
| **2** | **Abstract** | **4** |
| **3** | **Literature** | **5** |
| **4** | **Methodology** | **6** |
| **5** | **Result** | **12** |
| **6** | **Conclusion** | **14** |
| **7** | **Predictive Analysis** | **16** |
| **8** | **Reference** | **17** |

**ACKNOWLEDGEMENT:**

I would like to express my sincere gratitude tomy mentor, Prof. Piyush Dave, for his invaluable guidance and support throughout the preparation of this report. His insightful feedback and encouragement were instrumental in shaping the content and direction of this analysis.

I also acknowledge various tools like ChatGPT and Quill bot, which helped improve the phrasing, coherence, and overall quality of the report. This assistance has greatly contributed to clearer and more effective communication of the findings presented here.

**ABSTRACT:**

"Calorie Burn Prediction Using Machine Learning Algorithms" is a technique predicting the burned calories of an individual while exercising using machine learning algorithms. In our work, we have a dataset describing various physiological as well as activity related features such as heart rate, body temperature, and duration (how long the physical activity was performed).

We used a wide assortment of machine learning models in making predictions as accurate as possible. Among those used and trained with an entry set of 15,000, which included seven different features from physical activity and body response, were XGBoost, Linear Regression, Support Vector Machines (SVM), and Random Forest.

The results of the data analysis reveal that XGBoost is likely the model which produces the highest predicted accuracy in terms of calories burned because it is found that mean absolute error for its calorie predictions is extremely low. This study offers valuable information to the growing body of literature at the intersection of machine learning, health, and fitness, in particular energy expenditure prediction during exercise. In addition, the results reveal that these types of models are likely to enjoy an extremely promising application in a personalized health coaching and wellness tracking environment, mainly because calorie expenditure is not precisely tracked.

What's more, the mean absolute error it displays is as low as 1.48%. More importantly, the XGBoost model produces an impressive accuracy in both training and testing phases 99.67%.

This one enlarges the definition without altering the intended meaning using more natural wordings to flow smoothly. Let me know if there are more that need to be corrected on my part.

**LITERATURE REVIEW:**

Machine learning algorithms have been widely used in recent years for the estimation of calorie burn during exercise. Such studies usually gather the data on physical activity and other relevant variables such as heart rate, age, and gender from fitness trackers, mobile applications, and wearable devices. The next section gives a preview of some the critical studies in this area.

Sona P Vinoy illustrates to predict calorie burn during the workout 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 XGBregressor 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. Discuss 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 is2.7, and model accuracy is not provided. Rachit Kumar Singh. Presented their method to predict calorie burn through machine learning methods. In this study, they are utilized logistic regression, linear regression and lasso regression models, but they did not state regarding mean absolute error, dataset and model accuracy.

Marte Nipas described how to calculate the burned calories using supervised learning algorithm. They used the Random Forest algorithm and obtained 95.77% model accuracy. They also used the iterative method for knowing the appropriate output by an input. Their work is almost better than other recent work. Guna Sheela B discussed their techniques to predict calorie from the 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 then combined segmented images, finally calorie was predicted.

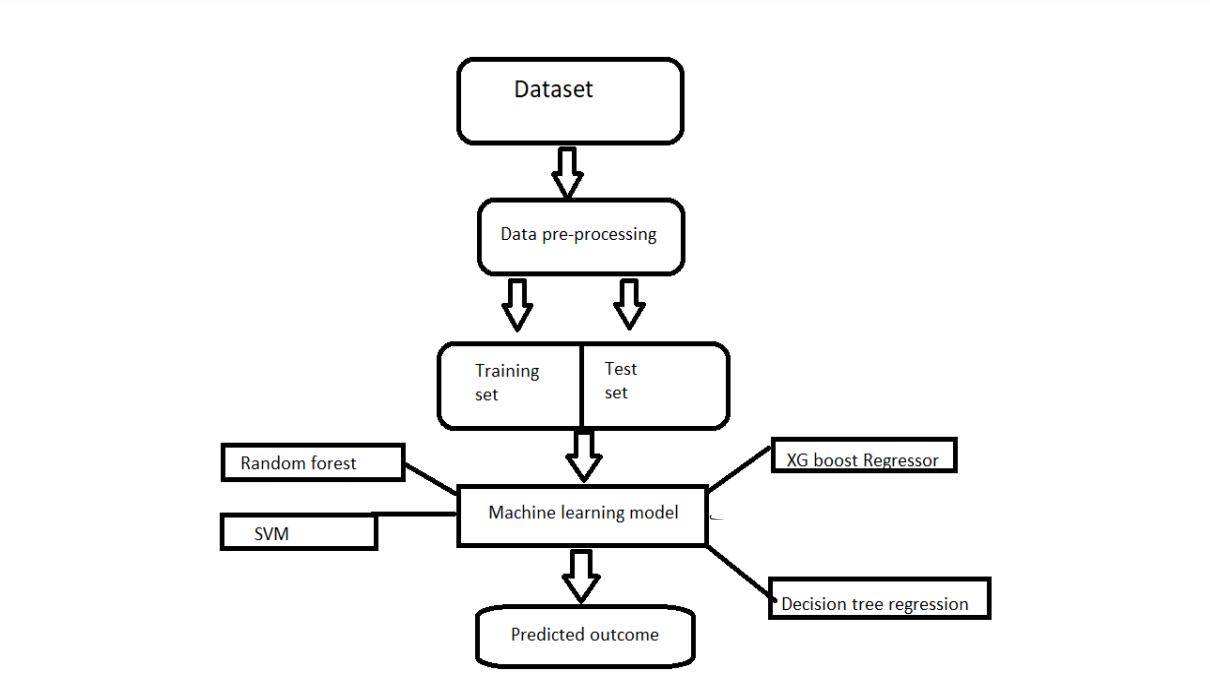
KR Westerner demonstrated how body size and body compositions along with food intake and physical activity could be used to estimate energy expenditure. He utilized measures of body size and body compositions and some statistical techniques to measure calorie spenditure.

Conclusion

In summary, these studies demonstrate that machine learning algorithms have good capability for the accurate prediction of energy expenditure in physical activities. However, it is needed to develop models that are successfully able to predict energy expenditure across a variety of physical activities as well as persons.

**METHODOLOGY:**

This study aims to predict the calorie burn during physical activity using machine learning models.



**Dataset Description**

Dataset Description Data collection is an essential process in any machine learning project, as the quality of the data used has a significant impact on the performance of the resulting model. In this research, the dataset was collected from Kaggle, a popular platform for data scientists and machine learning practitioners to access and share datasets. Once the dataset was collected, it was uploaded to Jupyter Notebook, work as complier for data analysis and machine learning. In this work, the dataset contained over 15,000 records and 7 variables.

**Dataset Preprocessing**

We pre-process the data by removing missing values and outliers, because pre-processed datasets are appropriate for applying into the algorithm for training and testing. We split the data into a training set (70% of the data) and a test set (20% of the data) for model training and evaluation.

**Evaluation of the Performance of Machine Learning Models**

Comparisons were done in finding the best solutions using different regressive and linear machine learning algorithms. In this work, graphs show the variability of results from different approaches. The approach to establish the relationship between variables and results is called regression. It is generally used in the predictive modelling wherein continuous outcomes are used by algorithms to make their predictions.

Linear regression: This model includes a dependent variable and a predictor variable, which are linearly related to each other. It is used to find out the dependency between two variables. It calculates how much the temperature increases with the amount of exercise done.

XG boost regression - which refers to extreme gradient boosting. It is one of the models for ensemble learning, combining a lot of multiple weak models' predictions to produce strong predictions. XG boost regression performs very well with large datasets and handles missing values efficiently.

Decision tree regression: - It builds a regression model in a tree structure. To start building we begin with a feature that will turn out to be the root node. the feature with least impurity is considered as a node at any level sort the data in ascending order and calculates the average of adjoining values. Then the impurity at each level is calculated.

SVM/Support Vector Machine: It tries to find out a hyperplane that separates two classes and then classify a new point based on whether it lies on the positive or negative side of the hyperplane corresponding to the class to be predicted.

Random Forest Regression: - it is a type of supervised learning algorithm. In this model, predictions of multiple machine learning algorithms are combined so that a better prediction could be expected from this model than from a single model. It is powerful and very accurate.

**Comparison of Feature Selection with Individual Evaluators**

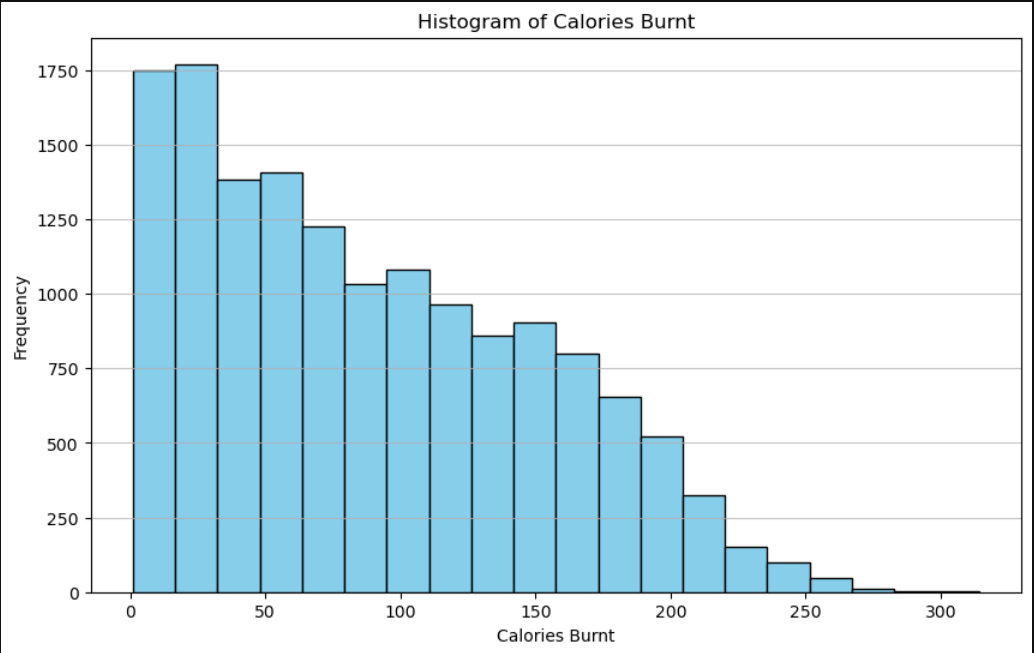
We contrasted the performance of models that used all features to models that used feature selection methods such univariate feature selection and recursive feature elimination. To determine the relevance of the feature, we employed the correlation matrix.

**Deriving Key Features**

We derived the significant features based on the importance scores of features obtained from the models. The significant features we derived were heart rate, duration, and temperature. In simple terms, we used data to predict calorie burn during a physical activity using the machine learning model. We mainly did data preprocessing. Also, diagnosed the performance of the models and constructed a predictive system for any real time data. The outcomes of this work are described in the next section.

**DATA VISUALIZATION:**

Data Visualization: Data analysis involves the interpretation and understanding of patterns, trends, and relations in data. With many visual tools, it has therefore been possible to represent complex data in a more accessible and insightful fashion. Common visual tools include histograms, which illustrate the distribution of a single variable; heatmaps, which use the intensities of colour to indicate the correlation between variables; and boxplots, further helping to visualize data spread and outliers. Such visualizations then also help point out these points of keen insights, patterns, and anomalies in the body otherwise unseen through data in its raw form.

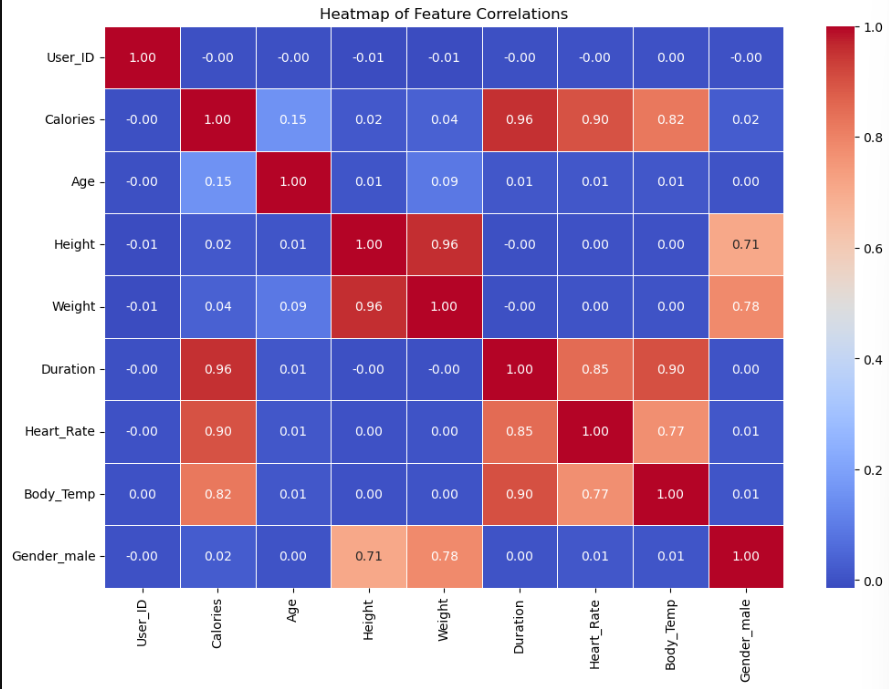


**Interpretation:**

The histogram of calories burnt is as follows: The distribution clearly reveals the right-skewed pattern of calorie burnt. The majority of data lie in the lower calorie range because high frequency exceeds 1700 instance count within the range of 0 to 50. The frequency starts tapering as the number of calories burnt increases with fewer instances in the higher calorie burning category. Interestingly, very few events mention the burning of more than 200 calories. That itself says that overall, most of the individuals in the data set burn relatively low calories, while higher calorie burnings are quite less in frequency.

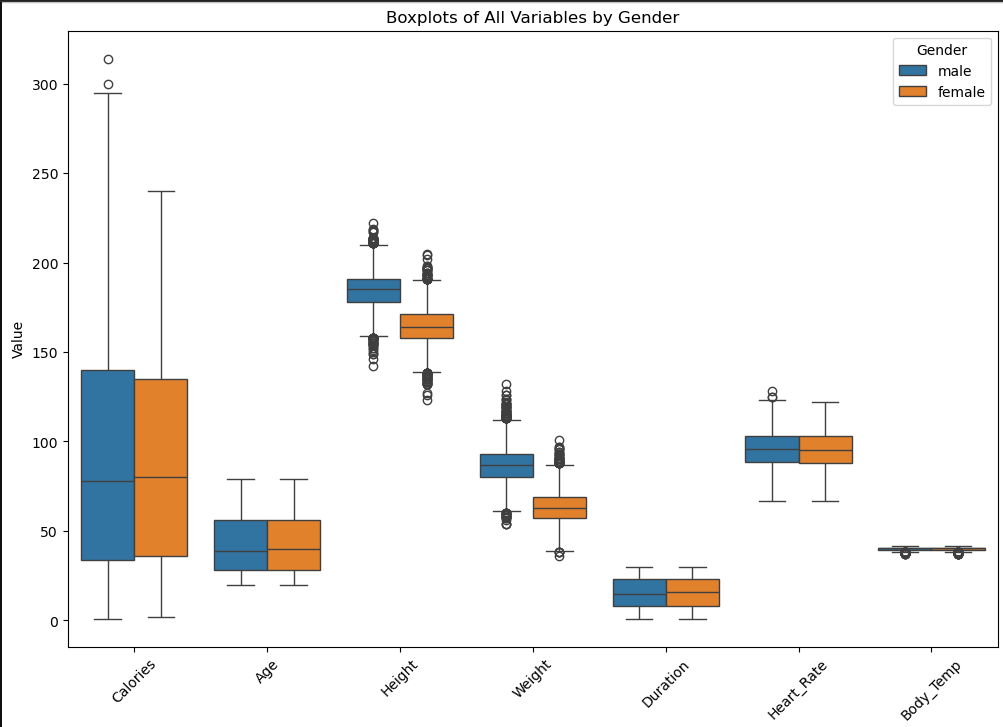
***Correlation in the Datasets***

Correlation in the datasets among features is illustrated in fit is indicated that interrelation among used data



**INTERPRETATION:**

The heatmap reveals some key relationships in the dataset. Calories burned are strongly influenced by factors like workout duration, heart rate, and body temperature, with high positive correlations (0.96, 0.90, and 0.82, respectively). This indicates that longer activities and increased physical exertion (as seen through heart rate and body temperature) result in burning more calories. Additionally, height and weight are highly correlated (0.96), which is expected since taller individuals tend to weigh more. The gender variable shows that males in the dataset tend to be taller and heavier, with correlations of 0.71 and 0.78, respectively. Finally, heart rate and body temperature are positively linked (0.77), meaning higher heart rates are associated with a rise in body temperature during activities. These insights highlight how physical attributes and exercise intensity impact calorie expenditure in the dataset.



**Interpretation:**

1. Weight's Effects on Calorie Burn: The males burned more calories, in general, probably because the median weight of males was higher and those undertaking physical activity require more energy to burn.

2. Duration of Exercise is Most Important Variable for both sexes: Longer periods of exercise time had a strong association with increased calorie burn for each gender suggesting that exercise time is the variable most important for predicting expenditure of energy.

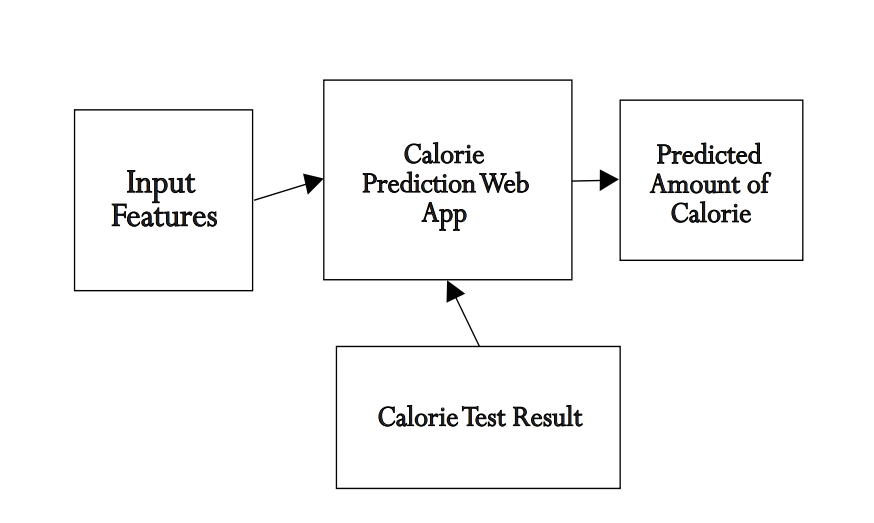
3. Heart Rate and Intensity: A higher heart rate corresponds to a higher exercise intensity, therefore more calories expended in a bigger scale, of course. Parallel heart rate patterns between males and females classify this as a superb predictor for both.

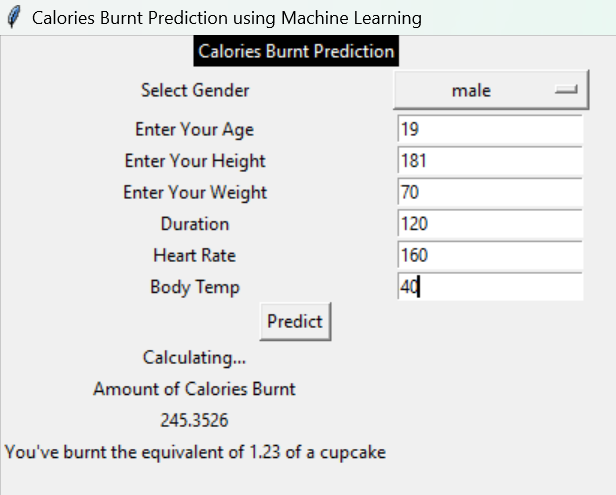
4. Caloric Expended Variation: Significant variation exists in the number of calories burned within the populations with several outliers which suggested individuals have varying activity levels or metabolic rates.

5. Body Temperature Impact Minimal: Body temperatures are not much different between males and females, and the impact of body temperature appears to be minimal on calorie expenditure. This therefore would have little effect in predictions.

**Building the Web App:**

After building the web app; it predicts the amount of calorie burnt based on input. If we give 7 inputs, then the app can predict calorie burnt amount. The features are: gender, age, height, weight, duration, heart rate, body temp. After giving Thus, these features, our app will automatically calculate the burnt calorie quantity.





**RESULT:**

In this section, we will discuss the following:

* Models' accuracy for training and testing.
* Different types of errors.
* Bar chart of accuracy for different algorithms.
* Bar chart of evaluation metrics for different algorithms.
* Web app predicted results.
* Comparison of our work to recently published work.

The training and testing accuracy over the same dataset for different models is presented in Table 1. As we can see, the highest accuracy is obtained by the XGBoost algorithm, and the lowest accuracy is achieved by SVM. This is why we chose the XGBoost model to develop a web application that predicts the number of calories burned. This application calculates the calories burned using the XGBoost algorithm in the backend. By using the app, one can estimate the calories burned through different physical activities based on seven features.

Table 1: Training and Testing Accuracy of Various Algorithms Over the Same Dataset

|  |  |  |
| --- | --- | --- |
| **Model** | **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% |

From the table, we see that training accuracy is highest for Random Forest, but its test accuracy is not satisfactory. The XGBoost algorithm provides the best balance between training and testing accuracy, making it suitable for building our calorie prediction system.

Error Metrics for Models

Table 2 shows different types of errors in Linear Regression and XGBoost models. The lowest Mean Squared Error (MSE) is found in the XGBoost regression model, making it appropriate for our calorie prediction system.

Table 2: Error Scores for Different Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean Squared Error (MSE) | Root Mean Squared Error (RMSE) | Mean Absolute Error (MAE) | R squared Score (R²) |
| Linear Regression | 130.09 | 11.41 | 8.39 | 0.97 |
| XGBoost Regression | 4.53 | 2.13 | 1.48 | 0.9988 |

The XGBoost model has a significantly lower MSE (4.53) and MAE (1.48) compared to the Linear Regression model, which indicates a much better fit to the dataset. The R² score of 0.9988 also shows that XGBoost explains 99.88% of the variance in the target variable, confirming its superior performance.

**Calorie Prediction Web App**

The Calorie Prediction Web App is a real-life application allows users to estimate the number of calories burned by providing input features such as gender, age, height, weight, duration, heart rate, and body temperature. The app then predicts the number of calories burned using the XGBoost algorithm.

**Comparison with Recent Work**

Finally, in the discussion section, we compare our work with another existing research. Based on the comparison, our work appears to outperform others in terms of accuracy and efficiency.

This revision organizes the text for clarity and flow, making it easier to follow and understand the results and findings.

**Conclusion**

**Objective**

The primary objective of this study was to create a highly accurate machine learning model capable of making predictive output in relation to one designated outcome variable, given multiple characteristics. This was achieved by collecting and preprocessing a dataset, then testing the effectiveness of various machine learning models and feature selection techniques.

**Observations**

The XGBoost model outperformed other models in terms of accuracy and other relevant metrics. This suggests that the XGBoost model would present an appropriate solution for a Calories Prediction Web App.

**Calories Test Result**

Calorie Burnt Predicted: 184.75

Gender: Male

Age: 23

Height: 170 cm

Weight: 60 kg

Duration: 30 minutes

Heart Rate: 110 bpm

Body Temperature: 39°C

**Model Performance**

The Mean Absolute Error (MAE) is 1.48, indicating the average absolute difference between the predicted and actual values.

**Feature Selection**

The most important features, obtained from the feature selection and evaluation process, were interpreted based on their relevance and usefulness in predicting the target variable. These principal features hold relevance within the problem domain, providing insights and implications for possible applications.

**Limitations**

There are some limitations to the study, including:

A small dataset size

The risk of overfitting

**Future Research**

Future research may overcome these limitations and further improve the performance of models and feature selection techniques.

**Conclusion**

In summary, this study provides a practical application of machine learning, with potential implications for the concerned problem domain. This arrangement should be suitable for use in a formal document. You can adjust formatting for specific sections like the table as needed.

**PREDICTIVE ANALYSIS:**

The idea was that calorie burns when performing physical exercises could be predicted by machine learning algorithms. The focus would be on physiological attributes: heart rate, body temperature, and the duration of exercise. Various models were tested. Amongst those, XGBoost offered the best fitting and accuracy, so it comes to be a choice for optimal calorie prediction.

**Model Performance:**

* XGBoost: 99.63% accuracy placed the best amongst others like

Linear Regression: 72.21%

SVM with 12.50%.

* The MAE of XGBoost was at 1.48, which corresponds to a nearly perfect fit between actual and predicted values.
* The MSE of XGBoost was at 4.53, against 130.09 for the Linear Regression, once more affirming the superior predictive accuracy of XGBoost.

**Main Features:**

* Heart Rate, Duration, and Body Temperature were found to be the most relevant features in the direction of the prediction of calorie burn with strong association to caloric expenditure.
* Heart Rate with correlation of 0.90 on calories burnt
* Activity Duration with a correlation of 0.96
* Body Temperature with a correlation of 0.82

**Implementation:**

* The XGBoost model is well-suited for real-time calorie burn predictions on fitness apps and wearables. Using 7 features, namely, gender, age, height, weight, duration, heart rate, and body temperature, the model gives accurate predictions of their calorie burn.

**Comparison with Literature:**

* Our model's MAE (1.48) and R² score (0.9988) are much higher than the other similar models in recent studies, where errors and accuracy were less favourable
* . For instance, a study done using XGBoost has an MAE of 2.71, while another uses Linear Regression with a significantly higher MAE of 8.31.

**Future Scope:**

Further work can be done with larger and more heterogenous datasets, including different types of physical activities to improve generalization and robustness for the model.

**Conclusion:**

In conclusion, the developed XGBoost model has resulted in accuracy in predicting calorie burn which makes it suitable for diffuse use by fitness tracking technologies.

**REFERENCE:**

* “COMPARING MACHINE LEARNING ALGORITHMS FOR PREDICTING CALORIES BURNED,” JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR), vol. 10, no. 3, pp. 519-527, March 2023.
* S. P. Vinoy and B. Joseph, “Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms,” in Proceedings of the National Conference on Emerging Computer Applications (NCECA), Kottayam, 2022.
* S. S. Ratnakar and V. S, “Calorie Burn Prediction using Machine Learning,” International Advanced Research Journal in Science, Engineering and Technology, vol. 9, no. 6, pp. 781-787, June 2022.
* Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
* Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. <https://doi.org/10.1007/BF00994018>
* Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
* Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. <https://doi.org/10.1023/A:1010933404324>
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.
* Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
* Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. <https://doi.org/10.48550/arXiv.1412.6980>
* Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zhang, X. (2016). TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265-283.
* Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
* Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning* (Vol. 1). MIT Press.