Calories Burnt Prediction

Adarsh Singh
DSAI
211020405
IIIT Naya Raipur
adarsh21102@iiitnr.edu.in

Sanchit Namdeo
DSAI
211020443
IIIT Naya Raipur
sanchit21102@iiitnr.edu.in

***Abstract*— The prediction of calories burned during physical activity is an important area of research, as it can help individuals better understand and plan their exercise routines. The project aims to predict the number of calories burnt after a person has done the workout based on several factors such as gender, duration of workout, heart rate, body temperature,etc.**

The analysis includes data exploration, feature selection, and model creation as part of the machine learning technique employed. In general, the study has the potential to offer crucial insights into the course of calories burnt and contribute to the creation of methods to accurately predict it.

***Keywords*— Calories, Linear Regression, R2 Score, MSE, MAE, Supervised Learning**

I. INTRODUCTION

Regular physical exercise is an essential component of a healthy lifestyle. It has numerous benefits, including weight management, improved cardiovascular health, and enhanced mental well-being. One important aspect of exercise is the number of calories burnt during a workout session, which can provide useful information for individuals seeking to achieve specific fitness goals.

Predicting the number of calories burnt during exercise can be challenging, as it depends on various factors such as age, weight, gender, and the intensity of the workout. Machine learning techniques such as linear regression and clustering can be used to analyze these factors and predict the number of calories burnt during a workout session. In the study, we explore the use of machine learning techniques to predict the number of calories burnt during exercise. Specifically, we use linear regression and clustering algorithms to analyze data collected from individuals performing different types of exercises at various intensities.

The results of this study could have significant implications for fitness enthusiasts, personal trainers, and healthcare professionals seeking to help individuals achieve their fitness goals. The ability to accurately predict the number of calories burnt during exercise could enable individuals to better monitor their progress and adjust their workout routine. Overall, the project seeks to provide insights into the use of machine learning techniques for predicting the number of calories burnt during exercise. By doing so, we hope to contribute to the growing body of research on the application of machine learning in the field of fitness and health.

II. Dataset

The dataset is obtained from the ‘fmendes-DAT263x-demos’, platform on Kaggle. The dataset contains 15000 entries of the different people who participated in the experiment. The dataset contains information about the participants such as gender, age, height, weight, duration of workout, heart rate just after the workout, and body temperature. Overall, it contains 9 features.

III. Methodology

Following Steps were performed: -

Data Cleaning

Identifying and correcting or removing inaccurate, incomplete, or irrelevant data in a dataset. It involves detecting and correcting errors, filling in missing values, and removing duplicates or outliers to improve the quality and usefulness of the data.

Feature Selection

We have used correlation analysis, feature importance analysis, and domain knowledge to select the most relevant features for our model so that we can get the most accurate model.

Conversion of Data

We have also got text data (gender), which needs to be converted into numerical data, so to be useful in our project, for which we have used Label Encoding.

Model Selection

Choosing the best model from a set of candidate models for a given dataset. It involves evaluating the performance of different models using metrics such as accuracy or mean squared error, and selecting the model that performs best on the validation set or through cross-validation.

Model Training

Since the data is quantitative, we have used regression and clustering techniques to train the model.

In linear regression, the model is trained using a set of input features and a continuous target variable. The goal is to find a linear relationship between the features and the target variable.

In polynomial regression, the model is trained using a set of input features and a continuous target variable. The goal is to find a polynomial relationship between the features and the target variable. We had taken 2 degree polynomial, as it suited best according to the visualisation of data.

In clustering, the model is trained using only a set of input features but no output features. The goal of the model is used to a certain pattern in data, such that it suits the model best.

Model Testing

Both models are then tested using a separate test dataset to evaluate their performance.

Performance Evaluation

Performance evaluation is an essential step in machine learning to assess the quality of the trained models. Multiple performance metrics are used to evaluate the model's accuracy, such as MAE, MSE, WCSS (Squared distance between the point and the centroid of the cluster), RMSE, and R2 score. Evaluating the models on the test data helps in selecting the best-performing model for deployment.

Visualization

Visualization is an important tool in exploratory data analysis and helps to understand patterns and relationships in the data. Different types of visualizations, such as scatter plots, histograms, and heat maps, are used to visualize the data and relationships between variables. Visualization can provide insights into the data and guide the selection of features and models for machine learning tasks.

IV. Evaluation Metrics

Evaluation metric refers to a measure that we use to evaluate different models. We have used the following evaluation metrics:

MSE

MSE (Mean Squared Error) is a commonly used metric to evaluate the performance of regression models. It calculates the average squared difference between the predicted and actual values, providing a measure of how well the model fits the data.

R2 Score

R2 score is a metric used to evaluate the performance of regression models. It measures the proportion of the variance in the dependent variable that can be explained by the independent variables, with a value of 1 indicating a perfect fit.

MAE

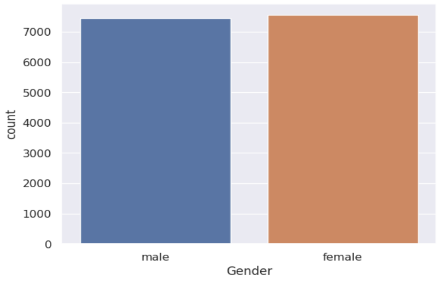
Mean Absolute Error (MAE) is a metric used to measure the average difference between predicted and actual values in a dataset, calculated by taking the absolute value of the differences and averaging them.

WCSS

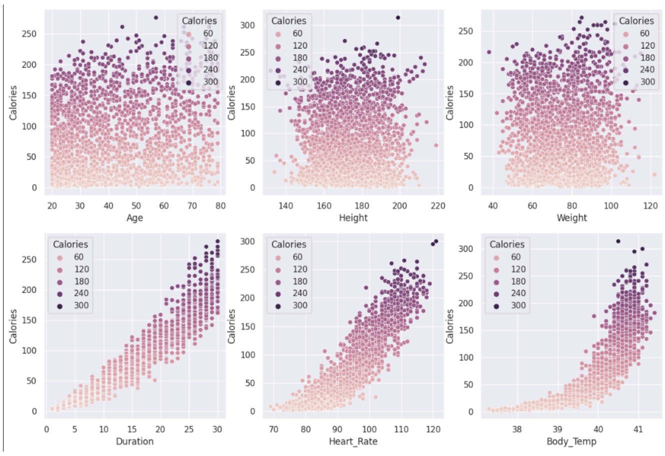
Within-Cluster Sum of Squares (WCSS) is a metric used to evaluate the quality of clustering in unsupervised machine learning. It measures the sum of squared distances between each data point and its centroid, indicating how tightly grouped the data points are within each cluster. Lower WCSS values suggest better clustering.

V. Observations / Result

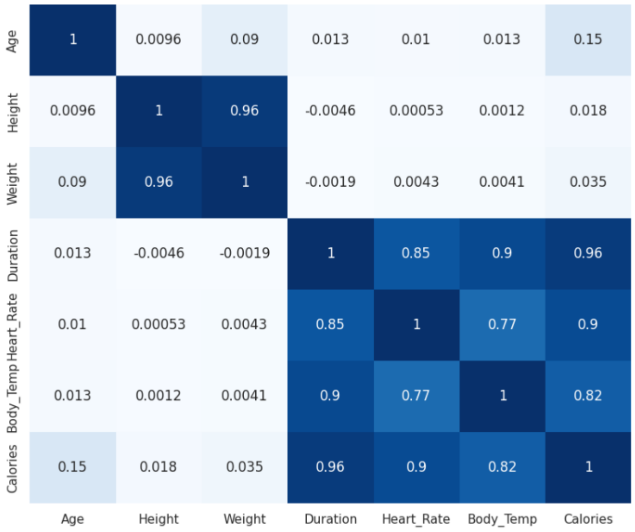
Gender Ratio of data:



Data Distribution:

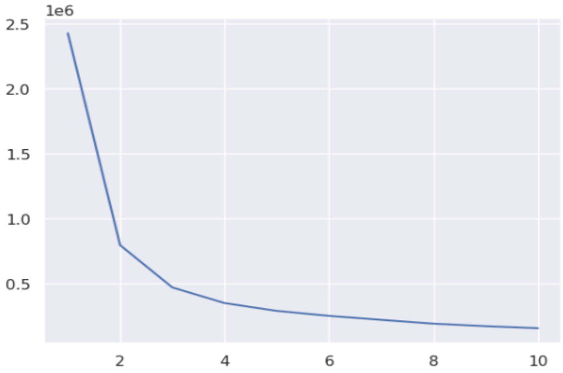


Correlation Matrix:



	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

WCSS:



Linear Regression (All Features)

Coefficient - [-1.3529688 0.49863216 -0.17806306 0.29718665 6.63137447 1.98567329 -16.92036832]

Intercept - [462.1763257347644]

R2 Score - 0.9663275127619021

MAE - 8.451945920538375

MSE - 132.91911580007886

Linear Regression (3 Correlated Feature)

Coefficient - [-16.67245864 2.00013293 6.61115255]

Intercept - [462.1763257347644]

R2 Score - 0.9450699827305398

MAE - 10.659186142185114

MSE - 216.8313042844934

Polynomial Regression with degree 2 (3 Correlated Feature)

Coefficient - [[6.68349523e+01, -1.65472755e+00, -1.45100540e+01, -9.39775398e-01, 6.19786363e-02, 1.79214643e-01, -4.78570314e-03, 1.35403929e-01, -1.36571020e-02]]

Intercept - [-1207.56179666]

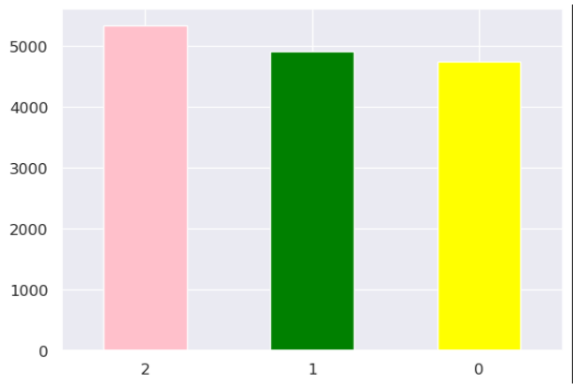
R2 Score - 0.961216

MAE - 8.296657

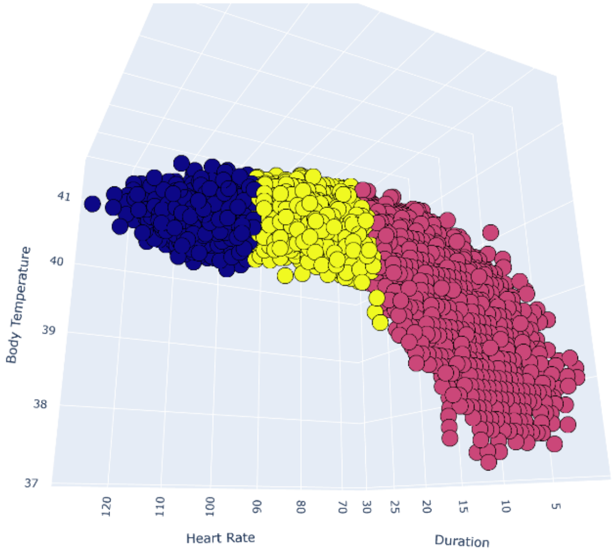
MSE - 153.098185

Clustering (3 Correlated Feature)

We obtained the optimal of 3 clusters with a count of:-



Clusters -



According to the elbow method, the clusters would have least error when the number of clusters were **Three(3)**.

VI. Conclusion

We have used various approaches to find the best optimum model. At starting we tried to find out the pattern hidden in the data using data distribution and scatter plots. After that, we tried linear regression which gave us a pretty good and accurate result. Additionally, we used 3 highly correlated features instead of all the features, which also gave approximately the same result which we obtained with all features. Thus, we can conclude that obtained 3 features are the best predictors of burnt calories.

We also tried polynomial regression with degree ‘2’ in the highly correlated features, which gave us a commendable result. In the end, we also tried to cluster the data based on that 3 features, which gave the best results when there are 3 clusters present, verified using the elbow method.

VII. Future Scope

- Developing more accurate prediction models: The development of more accurate and reliable prediction models is an area of active research.
- We could explore various machine learning algorithms and techniques, including deep learning and neural networks, to improve the accuracy of Calories Burnt prediction.

- Model optimization: Fine-tuning the regression models to improve their predictive performance, by experimenting with different model hyperparameters and regularization techniques.

- Feature engineering: Exploring additional features to include in the regression models, which could potentially enhance their accuracy and generalizability.

IX. ACKNOWLEDGEMENT

We want to express our sincere appreciation and thanks to Dr. Mallikarjuna Rao for his invaluable guidance and support throughout our report on Calories Burnt Prediction. His expertise in statistical data analysis has been instrumental in shaping our work and achieving the desired outcomes. We are grateful for his valuable feedback, insightful suggestions, and encouragement, which have helped us immensely in completing this report.

We would also like to thank the institute for offering this course on Statistical Data Analysis, which has helped us to develop a better understanding of the subject. We acknowledge the efforts of the faculty and staff in creating a conducive learning environment and providing us with the necessary resources to pursue our academic goals.

Once again, we express our heartfelt appreciation for Dr Mallikarjuna Rao's mentorship and support. His guidance has been crucial in our academic journey, and we are truly grateful for it. We also thank the institute for allowing us to learn and grow as students.

IX. REFERENCES

- Muskan Jha, Calories Burnt Prediction

  https://www.kaggle.com/code/muskanjha/calories-burnt-prediction

- Calories Burnt Prediction using Machine Learning - GeeksforGeeks

  https://www.geeksforgeeks.org/calories-burnt-prediction-using-machine-learning/