Title

A

ADM Course Project Report

in partial fulfilment of the degree

Bachelor of Technology

in

Computer Science & Engineering

By

D. Sathwik Reddy (230351989)

P. Surendra Reddy (2303A51981)

CH. Venkata Sai Sri Ram (230351973)

N. Sai Charen (2303A510H3)

D. Vishnu Teja (2303A510F4)

Under the guidance of

Bediga Sharan

Assistant Professor

Submitted to

School of Computer Science and Artificial Intelligence

DEPARTMENT OFCOMPUTER SCIENCE& ENGINEERING.

CERTIFICATE

This is to certify that the **Application Of Data Mining – Course Project** Report entitled

"D. Sathwik Reddy, P. Surendra Reddy, CH. Venkata Sai Sri Ram, N. Sai Charen, D. Vishnu Teja"is a record of bonafide work carried out by the student(s) bearing 2303A51989, 2303A1981, 2303A51973, 2303A510H3, 2303A510F4 Hallticket No(s) during the academic year 2024-25 in partial.

Fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science &**

Engineering by the SR University, Warangal.

Supervisor	**Head of the Department**
(Mr. Bediga Sharan)	(Dr. M. Sheshikala)
Assistant Professor	Professor

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our project guide **Mr. Bediga Sharan**, Assoc.Prof.CS and AI as well as Head of the CSE Department Dr.M.Sheshikala, Associate Professor for guiding us from the beginning through the end of the Capstone Phase–II project with the irintellectual advices and insightful suggestions. We truly value their consistent feedback on our progress ,which was always constructive and encouraging and ultimately drove us to the right direction.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support

ABSTRACT

Accurate estimation of calories burned during physical activity plays a vital role in fitness monitoring and personal health management. Conventional methods often use fixed formulas that fail to account for individual differences and real-time physiological factors, resulting in less reliable predictions. This project presents a machine learning-based approach to predict the number of calories burned by analyzing personal attributes such as age, gender, height, and weight, along with exercise-related parameters including duration, heart rate, and body temperature.

Using a dataset containing detailed activity records, various regression algorithms were trained and evaluated to identify the most accurate model for calorie prediction. The proposed system aims to deliver personalized and precise estimations, potentially improving the effectiveness of fitness plans and aiding in better lifestyle choices. The final model can be integrated into fitness applications or wearable technologies to enhance health tracking capabilities.

KEYWORDS

numpy, pandas, matplotlib, plt, seaborn, gridspec, sklearn, Random Forest Classifier.

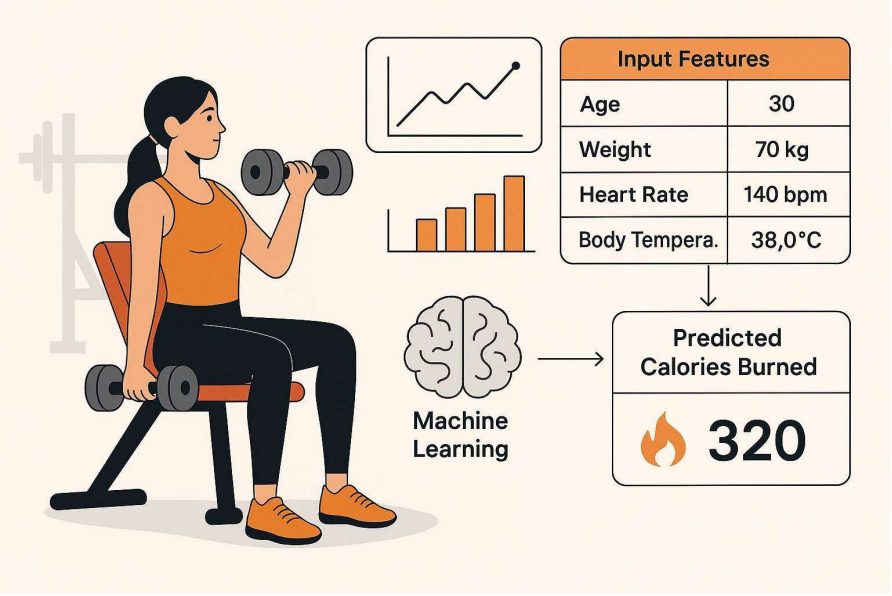
INTRODUCTION

In recent years, there has been a growing awareness and emphasis on personal health, fitness, and lifestyle management. One of the key elements in achieving fitness goals is understanding how many calories are burned during physical activity. Calorie tracking not only helps individuals monitor their energy expenditure but also aids in designing effective workout and diet plans tailored to specific health objectives such as weight loss, maintenance, or muscle gain.

Traditional methods of estimating calorie burn often rely on fixed formulas or approximations based on limited variables like activity type and duration. However, these methods overlook important personal and physiological factors such as age, gender, height, weight, heart rate, and body temperature, leading to generalized and often inaccurate predictions.

With the advancement of data science and machine learning, it is now possible to create more accurate and personalized models for calorie prediction. By analyzing large datasets that include both physical attributes and exercise metrics, machine learning algorithms can learn complex patterns and relationships to provide precise calorie burn estimations.

This project aims to develop a machine learning-based predictive model that estimates the number of calories burned using a combination of personal and exercise-related features. The goal is to create a tool that enhances the accuracy of calorie tracking, ultimately supporting better decision-making for users in their fitness journey.



OBJECTIVE

The objective of this project is to develop a machine learning model that accurately predicts the number of calories burned during physical activity based on individual physiological and workout-related factors. The project aims to analyze the influence of features such as age, gender, weight, height, heart rate, body temperature, and exercise duration on calorie expenditure. It involves preprocessing the dataset for model training, evaluating and comparing multiple regression algorithms to determine the most effective one, and building a reliable predictive model that generalizes well to new data. Additionally, the project explores integrating the model into a user-friendly interface for practical use in fitness tracking and health monitoring.

APPROACH

The Calories Burned Prediction project begins with clearly defining the objective: to predict the number of calories burned based on biometric and activity-related data. The next step involves collecting a relevant dataset that includes features such as age, gender, height, weight, duration of activity, heart rate, and the target variable—calories burned. After data collection, preprocessing is essential and includes handling missing values, encoding categorical variables like gender, and scaling numerical features for consistency. Exploratory Data Analysis (EDA) follows, where patterns, distributions, and correlations between features are visualized and analyzed to understand the data better. Feature selection or engineering may be applied to enhance model performance. For the modeling phase, various regression algorithms such as Linear Regression, Random Forest, and XGBoost will be tested. Finally, the chosen models will be trained using a split of the dataset into training and testing sets, and their performance will be evaluated using appropriate regression metrics such as MAE, MSE, or R^2.

PROBLEM STATEMENT

Calories Burned Prediction

In today's health-conscious society, accurately tracking calories burned during physical activity is crucial for effective fitness planning, weight management, and overall well-being. Traditional fitness trackers often rely on generalized formulas that may not account for individual differences such as age, gender, height, and weight. This can lead to inaccurate calorie estimations, which may hinder progress toward personal health goals.

The objective of this project is to develop a machine learning model that can accurately predict the number of calories burned based on key input features such as:

Age

Gender

Height

Weight

Duration of physical activity

Heart rate

Body temperature

By leveraging historical activity data, the model aims to provide more personalized and accurate calorie burn estimations. This can be used in health apps, wearable devices, or fitness platforms to enhance user experience and promote better health outcomes.

IMPLEMENTATION

CODE :-

```
# import required libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from xgboost import XGBRegressor
from sklearn.linear_model import LinearRegression
#from sklearn.linear_model import Ridge,Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
from statsmodels.stats.outliers_influence import variance_inflation_factor
import pickle

import warnings
from warnings import filterwarnings
filterwarnings("ignore")

sns.set()
```

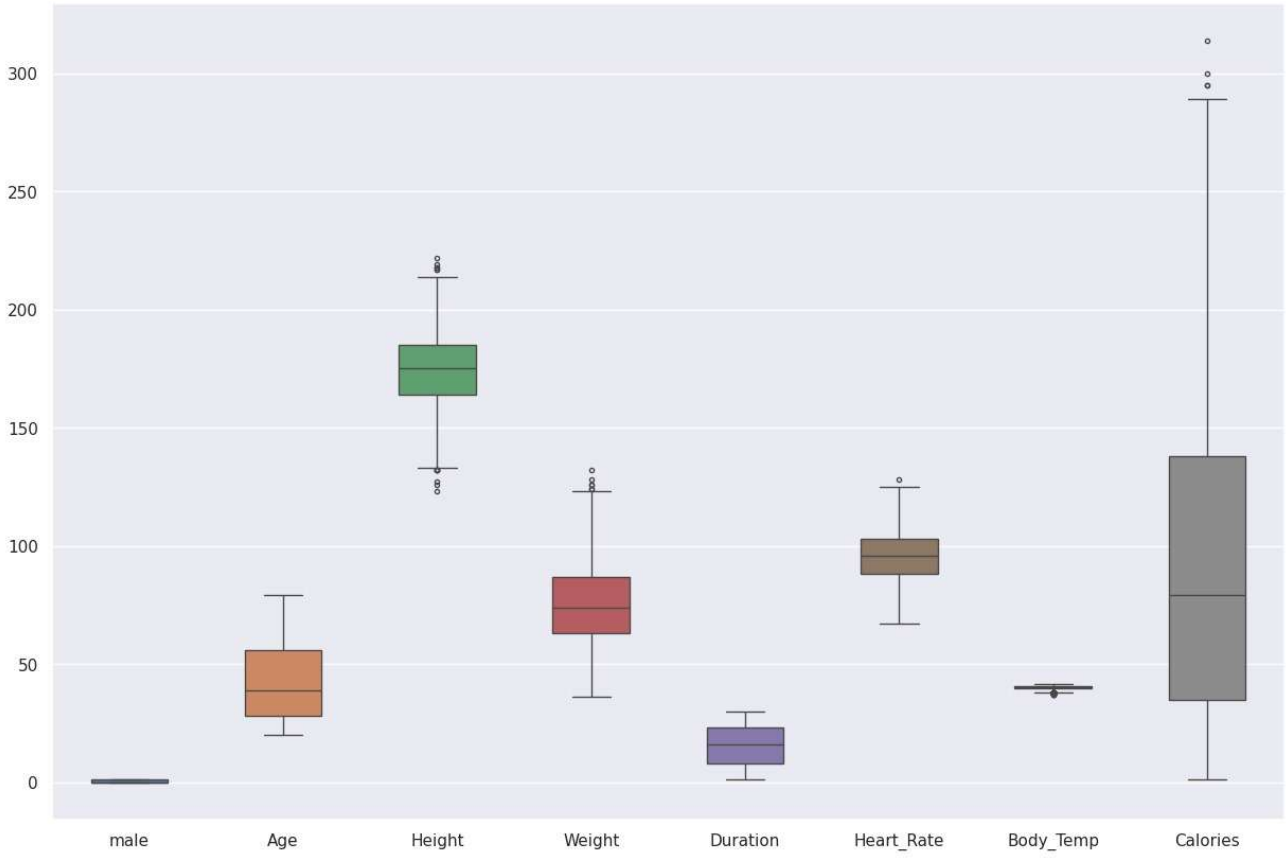
```
#from sklearn import metrics
def predict(ml_model):
    model=ml_model.fit(X_train,y_train)
    print('Score : {}'.format(model.score(X_train,y_train)))
    y_prediction=model.predict(X_test)
    print('predictions are: \n {}'.format(y_prediction))
    print('\n')

    r2_score=metrics.r2_score(y_test,y_prediction)
    print('r2 score: {}'.format(r2_score))

    print('MAE:',metrics.mean_absolute_error(y_test,y_prediction))
    print('MSE:',metrics.mean_squared_error(y_test,y_prediction))
    print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,y_prediction)))

    sns.distplot(y_test-y_prediction)
```

```
plt.figure(figsize=(20,15))
plotnumber = 1

for column in data:
  if plotnumber <= 8:
    ax = plt.subplot(3,3,plotnumber)
    sns.distplot(data[column])
    plt.xlabel(column,fontsize=15)
  plotnumber+=1
plt.show()
```

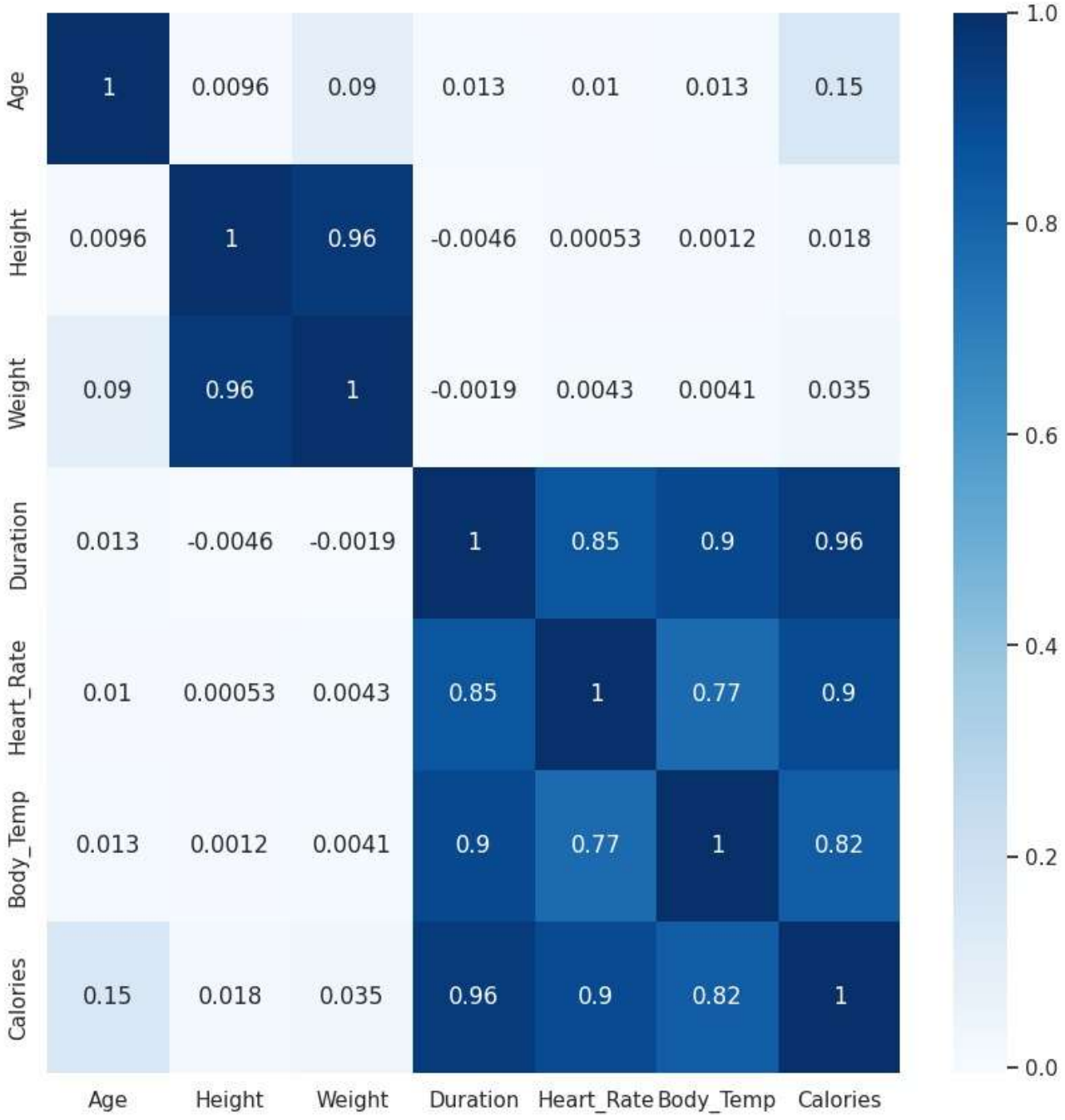
RESULT SCREENS :-

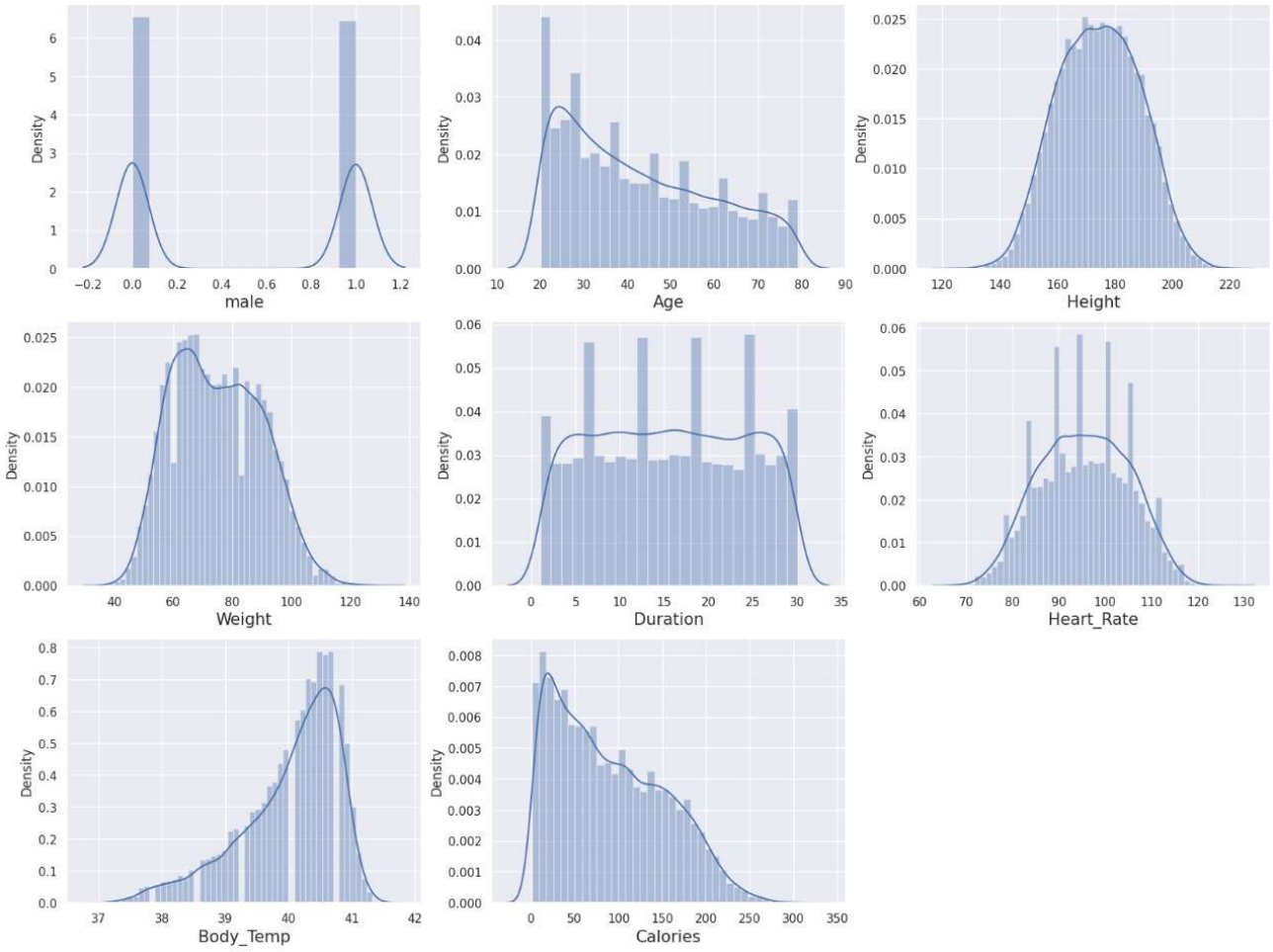
Categorical and

Numerical.



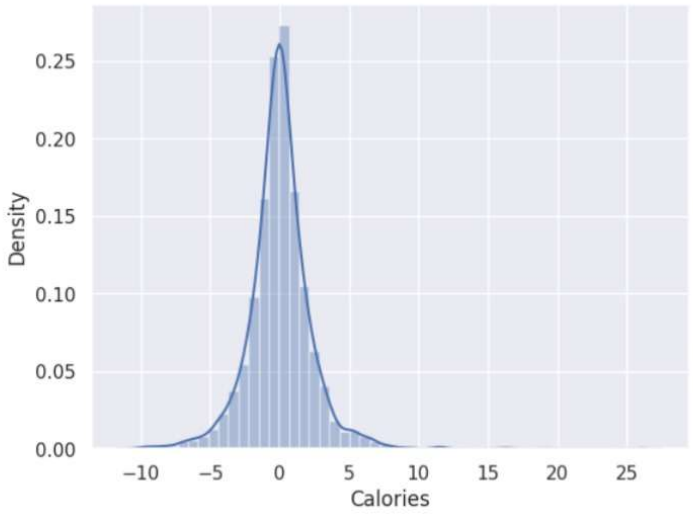
Heatmap for correlation.




Density
male
Age
Height
Weight
Duration
Heart_Rate
Body_Temp
Calories

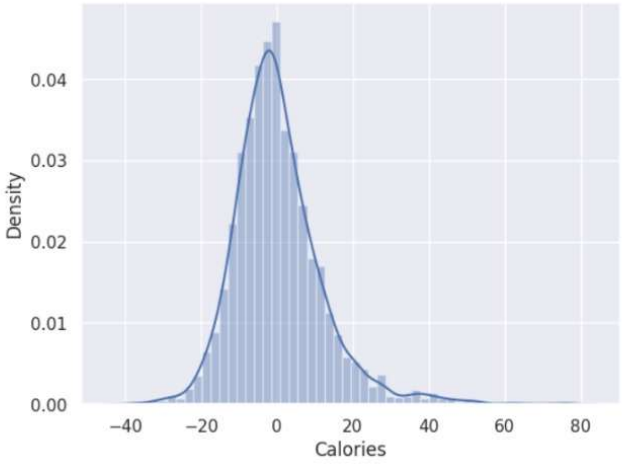
XGB Regressor.

```
Score : 0.9995380557081355
predictions are:
 [197.06581   70.867226 196.99498  ...  29.043041 104.09284   14.61472 ]


r2 score: 0.9986863132331905
MAE: 1.5521575984954834
MSE: 5.2744122853837005
RMSE: 2.2966088664340956
```



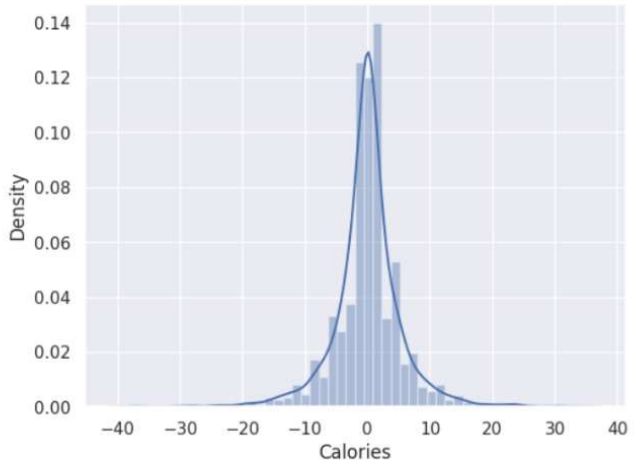
Linear Regression.

```
Score : 0.967592555473578
predictions are:
 [198.81182363  80.43555305 194.40940033 ...  22.14745631 118.63504926
 -11.98134672]


r2 score: 0.9655977245826504
MAE: 8.479071745987955
MSE: 138.12408611460899
RMSE: 11.752620393538157
```



DecisionTree Regression

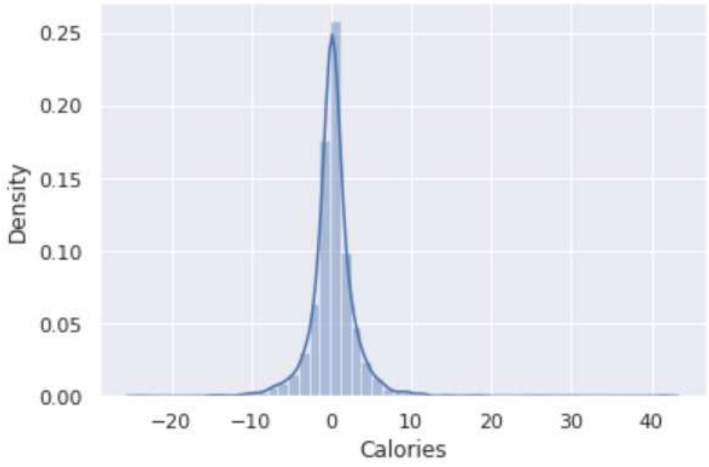
```
Score : 1.0
predictions are:
 [194.  75. 204. ...  30. 109.  13.]

r2 score: 0.9925279631413153
MAE: 3.508
MSE: 30.0
RMSE: 5.477225575051661
```



RandomForest

Regression

```
Score : 0.9996818696593301
predictions are:
 [197.36  67.8  195.67 ...  27.7  111.56  14.08]

r2 score: 0.99766556152047
MAE: 1.80648
MSE: 9.372699266666668
RMSE: 3.061486447245303
```



CONCLUSION

Dataset Overview:

Merged two datasets: calories.csv and exercise.csv, each containing 15,000 records.

Final dataset contains 8 features after dropping User_ID.

Missing Values:

No missing values were present in the dataset, ensuring data completeness.

Feature Engineering:

The categorical variable Gender was one-hot encoded to create a binary feature male.

All other features (Age, Height, Weight, etc.) are numerical and were kept as-is.

Exploratory Data Analysis (EDA):

Distribution plots revealed that most features (e.g., Calories, Duration, Heart Rate) are not perfectly normally distributed.

Outliers were detected in several variables (especially Calories, Body_Temp, Heart_Rate), visible through boxplots and distribution shapes.

Heatmap of correlation showed strong relationships between Duration, Heart_Rate, Body_Temp and Calories, suggesting they are important predictors.

Feature Preparation:

Features (X) and target (y) were separated successfully.

Dataset was split into training (80%) and testing (20%) sets for model evaluation.

GITHUB LINK :-

https://github.com/sriram453/ADM-PROJECT