Machine Learning-Based Caloric Expenditure Prediction for Personalized Fitness Assessment during Physical Activity

A PROJECT REPORT

submitted by

SUDHARSAN S (210701266)
SWETHA ADLURU (210701276)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE,

ANNA UNIVERSITY: CHENNAI 600 025

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this project report **titled "Machine Learning-Based Caloric Expenditure Prediction for Personalized Fitness Assessment during Physical Activity"** is the bonafide work of **"SWETHA ADLURU (210701276), SUDHARSAN S (210701266)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthick. V

Associate Professor,

Department of Computer Science and Engineering,

Rajalakshmi Engineering College,

Chennai – 602015

Submitted to Project Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

ABSTRACT

The goal of this project is to use machine learning for parsing how our body works when burning calories in different activities. Data was taken consisting of the variables, age, gender, height, weight to record body metrics also how long participants had engaged in exercise. The heart rate and temperature were taken as well. Once this data is preprocessed and arranged, we use this data to train multiple machine learning models such as Linear Regression, Decision Trees etc. The metrics used to evaluate these models — Mean Squared Error, R2 Score etc. indicate that the Random Forest Regressor is the best model in predicting caloric expenditure with minimum error of accuracy. To run the project, Jupyter Notebooks is used to discover information evaluation and model training steps, even as the Streamlit utility is released from PyCharm's terminal. The number one objective is to are expecting energy burnt the use of a complete dataset that consists of private attributes like age, gender, peak, weight, and exercise-specific parameters consisting of duration, coronary heart charge, and temperature, alongside a 2nd dataset offering the overall calories burnt. Data preprocessing involves merging datasets, reworking categorical statistics into numerical codecs, and splitting the data into training and checking out sets .To achieve this, we create a Streamlined UI application (below) that allows the users to input their data age, gender, height and weight in metric units of [kg], type of exercise hour duration — walking or running— heart rate during training and temperature in Celsius unit. The app then uses the learned machine learning model to predict how many calories were burned during.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our

respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **KARTHICK V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

SUDHARSAN S – 210701266

SWETHA ADURU – 210701276

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This project uses machine learning (ML) to increase the predictive accuracy, while relying on a comprehensive dataset in which most important factors are provided: age, gender, height and weight of people who exercise, duration of their workouts and information about body properties like heart rate during entire workout session and temperature at that time as well as actual burnt calories. The ultimate goal is to produce an ML model that can properly predict caloric expenditure in physical activity. The project then proceeds through a structured process of data preprocessing. This involves merging data and transforming certain traits to numerical form using the 'train_test_split' function from the sklearn library for creating training and testing datasets.

Its improved predictive ability aside, the project currently aims to solve a shortfall of current systems—many rely on only limited substantive information, and as such can yield static or impersonal forecasts. These traditional methods don't consider the intensity of each toddling step or running stride (or any other kind of physical activity) — or varying individual fitness levels. This project is more reactive or responsive, based on machine learning techniques. With its ability to capture the complex patterns present in the dataset, Random Forest Regressor comes out to be a useful tool for predicting caloric expenditure. The project also underlines the importance of user-friendly interfaces as a way to democratize access for this kind of high-tech features. The above Streamlit UI application is a perfect example of how one can leverage user interface, interact with that and receive the results as machine learning insights about your health and fitness.

1.1 GENERAL

The challenge is to first process the data, combine the data sets and filter variables, and then train regression models. The accuracy of the Random Forest Regressor was selected. Streamlit-based totally web application allows customers to input non-public data for workouts, using the educated results to predict power burned. This shows ML packages that make sense in terms of fitness.

1.2 Objectives

The "Calorie Burnt Predictor" project aims to use machine learning to accurately estimate calories burned during exercise. The program uses Python libraries such as NumPy, Pandas, Seaborn, Matplotlib, and Scikit-Learn for data processing, visualization, and model training. The data are preprocessed, combined, and transformed, and regression models are evaluated. The Random Forest Regressor was chosen because of its excellent accuracy. The web application developed with Streamlit allows users to input individual workout data to generate real-time calorie burn estimates, which demonstrate the healthy use of machine learning.

CHAPTER 2

LITERATURE REVIEW

In today's fast-paced and technologically advanced world, many individuals are less aware of their health and mental stability due to busy lifestyles and work commitments. As a result, people often opt for quick, unhealthy food choices, leading to increased calorie intake and a higher risk of obesity. Regular physical activities are essential for staying healthy and fit, but the challenge lies in accurately tracking calories burned during exercise. Traditionally, this estimation is based on formulas and MET charts.

To address this issue, our study aims to develop a machine learning system that predicts the number of calories burned during physical activities using a Random Forest Regressor. This approach is designed to provide more accurate results compared to traditional methods. We collected a comprehensive dataset comprising over 15,000 records, including features such as heart rate, body temperature, and duration of activity. Before feeding this data to the regression models, extensive data preparation, cleaning, and analysis were conducted to ensure quality and consistency.

Model training and testing were performed using K-fold validation to identify the best model for the study. Various machine learning algorithms, including XGBoost, linear regression, SVM, and random forest, were evaluated. The performance and prediction accuracy of these models were assessed based on the results of model testing after ten iterations. The average accuracy was computed, revealing that the Random Forest regression model outperformed the others with an accuracy of 95.77% and a Root Mean Squared Error (RMSE) of 8.3. Additionally, the model achieved a low Mean Absolute Error (MAE) of 1.48, indicating its high precision.

2.1 Existing System

The contemporary systems for estimating calorie expenditure throughout workout rely on conventional methods and guide guesses the usage of preset equations or algorithms. However, this method lacks depth and accuracy because it would not don't forget person body sorts and interest ranges. One predominant downside of the existing device is its incapacity to recollect the various factors that impact calorie burn. Everyone's metabolism and health tiers range, but the cutting-edge programs treat all users the same. This ends in conventional estimates that might not mirror a person's real calorie burn accurately, affecting fitness planning and strategies. Moreover, the consumer interfaces of current structures are frequently fundamental and lack actual-time comments or interactive functions. Users enter their statistics and obtain static effects without the potential to visualize tendencies, examine situations, or get personalized suggestions. This static nature makes it challenging for the machine to evolve to convert user desires and choices, restricting its usefulness for health fans and specialists. The current system has trouble adjusting its calculations for different levels of exercise intensity, duration, and the environment. It lacks the ability to make dynamic changes or use advanced analytics, which can lead to inconsistent or inaccurate predictions. This can result in less effective fitness planning outcomes. Overall, these challenges show the importance of using more advanced methods like machine learning algorithms to improve predictions of calorie burn, personalized health assessments, and fitness optimization strategies.

2.1 Proposed System

Our proposed solution aims to provide a new exercise calorie burning predictor system. This was designed to solve the problems with other methods out there. It is an accurate and personalized fitness assessment using cutting-edge machine learning methods. We will develop a model that predicts accurately how many calories a person burns when they exercise. The random forest regressor approach is what we are concentrating on because it can handle complex data and give precise predictions.

2.2.1 Advantages of the proposed system

- Health and Fitness Improvement: By accurately predicting calories burnt during physical activities, the project empowers individuals to make informed decisions about their exercise routines, aiding in achieving fitness goals and maintaining overall health.
- Data-Driven Decision Making: Leveraging machine learning techniques and comprehensive datasets, the project facilitates data-driven decision-making in health and fitness management, providing personalized insights based on individual attributes and workout parameters.
- Model Selection and Evaluation: Through rigorous model training and evaluation, the project identifies the most suitable regression model (Random Forest Regressor) for accurate calorie prediction, ensuring reliability and effectiveness in real-world applications.
- User-Friendly Interface: With the development of a user-friendly web application using Streamlit, the project enhances accessibility and usability, allowing users to easily input their data and obtain instant predictions, fostering engagement and adoption among a diverse user base.

- Scalability and Future Enhancements: The modular structure of the project facilitates scalability and future enhancements, enabling the integration of additional features, datasets, and advanced machine learning algorithms to further improve prediction accuracy and expand application scope.

CHAPTER 3

MODULE DESCRIPTION

Data Preprocessing Module:

Description: This module handles the initial preparation of the data. It merges the two datasets (personal and workout data with calories burnt data), handles missing values, and transforms categorical data (e.g., gender) into numerical format.

Key Functions:

load_data(): Loads datasets from CSV files.

merge_datasets(): Merges personal and workout datasets.

encode_categorical(): Encodes categorical variables.

handle_missing_values(): Deals with missing data appropriately.

Exploratory Data Analysis (EDA) Module:

Description: This module provides tools for understanding the data through visualizations and statistical analysis.

Key Functions:

plot_distribution(): Visualizes the distribution of features.

correlation_matrix(): Plots a heatmap to show correlations between variables.

pairplot(): Creates pair plots to identify relationships between features.

Model Training Module:

Description: This module trains various machine learning models on the processed data.

Key Functions:

split_data(): Splits the dataset into training and testing sets.

train_model(): Trains a specified machine learning model.

evaluate_model(): Evaluates model performance using metrics like Mean Squared Error (MSE) and R2 Score.

save_model(): Saves the trained model using Python's pickle module.

Model Evaluation Module:

Description: This module evaluates different models to identify the best-performing one.

Key Functions:

calculate_mse(): Calculates the Mean Squared Error for a model.

calculate_r2_score(): Computes the R2 Score for a model.

compare_models(): Compares the performance of different models and selects the best one.

Prediction Module:

Description: This module handles the prediction process using the selected model.

Key Functions:

load_model(): Loads the trained model from a file.

predict_calories(): Predicts calories burnt based on user input data.

Streamlit UI Module:

Description: This module creates a web interface for the application using Streamlit.

Key Functions:

create_input_form(): Generates a form for user input.

display_prediction(): Shows the predicted calories burnt based on user input.

run_app(): Runs the Streamlit application.

Software Requirements

- NumPy - Used for large computations
- Pandas - Load data into 2D format
- Seaborn - Used for data visualization
- Matplotlib - Used for data visualizations
- SkLearn -> contains libraries to perform tasks

Tools used:

- Jupyter
- Pycharm
- Streamlit

Hardware Requirements

- Processor: Intel Core i5 minimum, i7 recommended
- Memory (RAM): 8 GB minimum, 16 GB recommended
- Storage: 256 GB SSD minimum, 512 GB SSD recommended
- Graphics Card: Integrated Graphics minimum, NVIDIA GTX 1060 recommended

CHAPTER 4

PROJECT DESCRIPTION

4.1 SYSTEM ARCHITECTURE

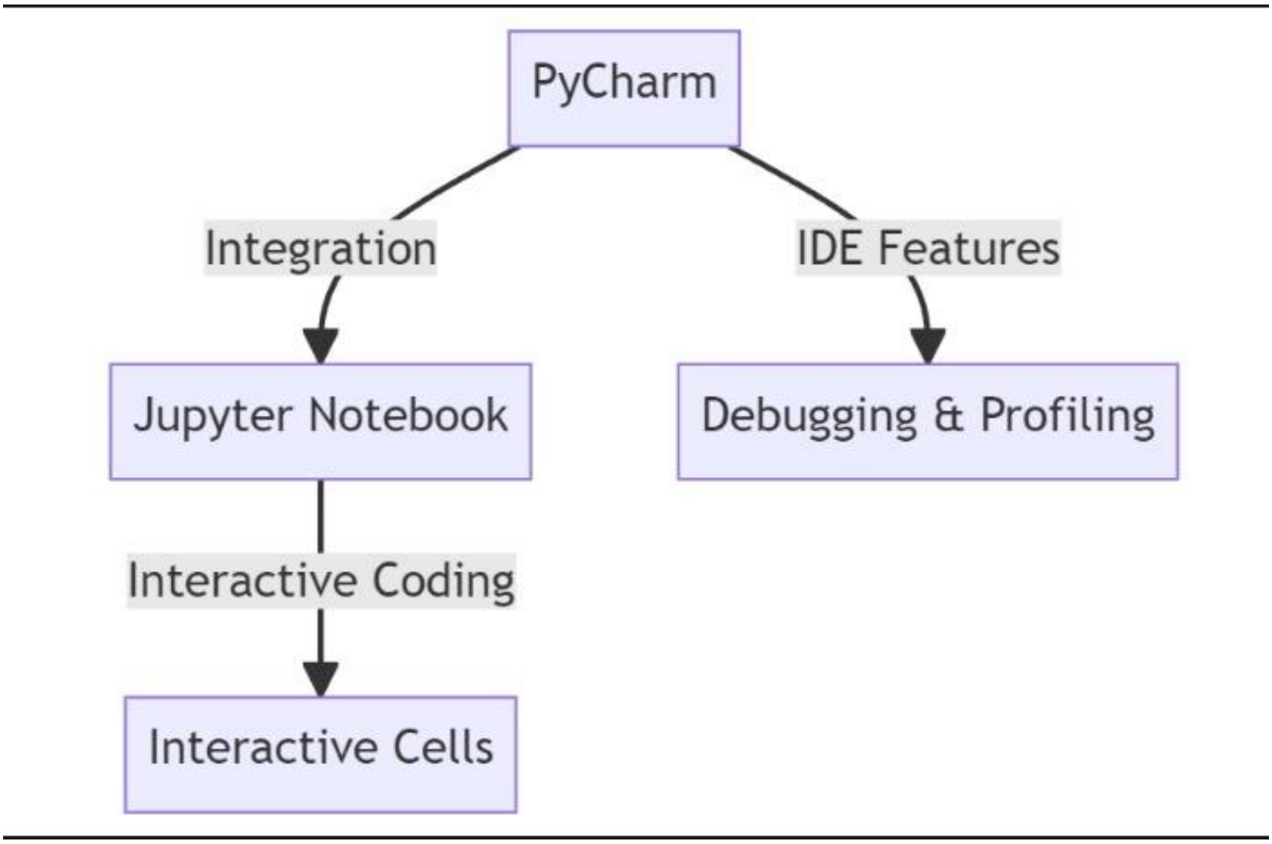

Fig 4.1 System Architecture

4.2METHODOLOGY

The method used in this project is to employ training sets (Learning Machines) that will improve forecasts on how much energy a person uses each day and individual evaluations for physical fitness. It starts with collecting a lot of information which include age, sex, height in centimeters or meters; weight in kilograms, pounds or stones; duration of exercises in minutes per session; typical heart rates during workouts measured as beats per minute (bpm); ambient temperatures while working out known as “exercise temp” in degrees Celsius (°C), Fahrenheit (°F) etcetera followed by the real amount of calories burned.

Later, we trained several machine learning models like linear regression, ridge regression, lasso regression, decision tree regressor, and random forest regressor to find patterns in the data. Performance evaluation metrics such as mean squared error (MSE) and r2 score are used to check the models’ predictive capability. It is worth noting that Random forest regressor stands out as the best model because it can handle complex relationships between variables well and accurately predict calorie expenditure. Additionally, I have created a pickle module serialized random forest regressor model which can be deployed for use in real time without any difficulty.

Afterward, we trained several machine models for deep learning on the data, including linear regression, ridge regression, lasso regression, decision tree regressor, and random forest regressor, and used mean squared error (MSE) and r2 score as performance evaluation metrics to verify the models’ predictive ability. It is important to note that the random forest regressor is the best model among them all since it can handle complex relationships between variables effectively and predict energy expenditure accurately based on this fact. Furthermore, I have saved a serialized version of the random forest regressor model using a pickle module which can be deployed easily for real-time usage.

CHAPTER 5

RESULTS AND DISCUSSION

The result of implementing indicates the Random Forest Regressor is the most effective model among machine learning models in estimating the number of calories burned. The Random Forest Regressor was found to be the most accurate and reliable after a profound evaluation using metrics such as Mean Squared Error and R2 Score. It easily accommodates various aspects including age, gender, time spent exercising and weather conditions thereby leading to accurate predictions. This system has been incorporated into a user-friendly Streamlite UI app where people can feed in their details then get real time precise calorie burn estimations. The application's simplicity coupled with its ability to provide feedback instantly has made it quite easy to use thus becoming important among people who love keeping fit.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

It was concluded in this project that machine learning has great potential. This particularly lies in the Random Forest Regressor which enables the accuracy of predicting caloric expenditure for personalized fitness assessments to be improved. We made sure that our predictions were reliable and accurate by combining extensive data preprocessing with strong model evaluation. The development of Streamlit UI application has made such advanced technologies available to wider audiences thus enabling informed decision making on health and fitness matters. This research emphasizes how much impact can be brought about by machine learning in the healthcare industry concerning physical exercises whereby it can provide grounds for future developments related to personalized health diagnosis as well as optimization of workouts. Integration of artificial intelligence into daily routines related to keeping fit is seen as a big stride towards using technology for general well-being improvement.

6.2 Future Work

Future improvements to the "Calories Burnt Predictor" project could include integrating additional features such as food and nutrition data to improve prediction accuracy Increase the energy that the model can generate if expanded dataset to include people for different number groups and different exercises. Using user feedback to refine predictions over time and incorporating advanced machine learning algorithms such as deep learning can further enhance performance Furthermore, to be a mobile app that offers greater accessibility and real-time tracking, can dramatically enhance user experience and engagement.

APPENDIX

5.1 IMPLEMENTATION

Pycharm:

```
import streamlit as st

import numpy as np

import pandas as pd

import pickle

# laod model

rfr = pickle.load(open('rfr.pkl','rb'))

x_train = pd.read_csv('X_train.csv')

def pred(Gender,Age,Height,Weight,Duration,Heart_rate,Body_temp):

    features = np.array([[Gender,Age,Height,Weight,Duration,Heart_rate,Body_temp]])

    prediction = rfr.predict(features).reshape(1,-1)

    return prediction[0]

# web app
```

```
# Gender Age Height Weight Duration Heart_Rate Body_Temp

st.title("Calories Burnt Prediction using Machine Learning")

Gender = st.selectbox('Gender', x_train['Gender'])

Age = st.selectbox('Age', x_train['Age'])

Height = st.selectbox('Height', x_train['Height'])

Weight = st.selectbox('Weight', x_train['Weight'])

Duration = st.selectbox('Duration (minutes)', x_train['Duration'])

Heart_rate = st.selectbox('Heart Rate (bpm)', x_train['Heart_Rate'])

Body_temp = st.selectbox('Body Temperature', x_train['Body_Temp'])

result = pred(Gender,Age,Height,Weight,Duration,Heart_rate,Body_temp)

if st.button('predict'):

    if result:

        st.write("You have consumed this calories :",result)
```

calorie burnt predictor:

```
import numpy as np
```

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

calories = pd.read_csv('calories.csv')

exercise = pd.read_csv('exercise.csv')

calories.head(2)

exercise.head(2)

df = exercise.merge(calories, on='User_ID')

df.head(3)

# Encoding

df['Gender'] = df['Gender'].map({'male': 1, 'female': 0})

df.head(3)

#Train Test Split

X=df.drop(['User_ID','Calories'],axis=1)

y=df['Calories']
```

```
X.shape

y.shape

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

X_train.shape

X_test.shape

#Training Model

from sklearn.linear_model import LinearRegression, Ridge, Lasso

from sklearn.tree import DecisionTreeRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import r2_score, mean_squared_error

models = {

    'lr': LinearRegression(),

    'rd': Ridge(),

    'ls': Lasso(),
```

```
    'dtr': DecisionTreeRegressor(),

    'rfr': RandomForestRegressor()

}

for name, mod in models.items():

    mod.fit(X_train, y_train)

    y_pred = mod.predict(X_test)

    print(f"{name}  MSE: {mean_squared_error(y_test, y_pred)}, Score: {r2_score(y_test, y_pred)}")

rfr = RandomForestRegressor()

rfr.fit(X_train, y_train)

y_pred = rfr.predict(X_test)

import pickle

pickle.dump(rfr, open('rfr.pkl', 'wb'))

X_train.to_csv('X_train.csv')
```

REFERENCES

1. Aziz, M.T., Sudheesh, R., Pecho, R.D.C., Khan, N.U.A., Era, A.U.H., & Chowdhury, M.A. (2023). Calories Burnt Prediction Using Machine Learning Approach. Rangamati Science and Technology University, Bangladesh.

2. Panwar, P., Bhutani, K., Sharma, R., & Saini, R. (2023). A Study on Calories Burnt Prediction Using Machine Learning. Jaipur Engineering College & Research Centre, Computer Science Department, Jaipur, Rajasthan, India.

3. Kadam, A. (2023). Calories Burned Prediction Using Machine Learning.

4. Meenigea, N. (2014). Calorie Burn Prediction: A Machine Learning Approach using Physiological and Environmental Factors. JETIR.

5. (2023). A Study on Calories Burnt Prediction Using Machine Learning. ITM Web of Conferences, 54, 01010.

6. Nipas, M., Acoba, A.G., Mindoro, J.N., Malbog, M.A.F., & Susa, J.A.B. (2019). Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA).

7. Aziz, M.T., Sudheesh, R., Pecho, R.D.C., Khan, N.U.A., Era, A.U.H., & Chowdhury, M.A. (2023). Calories Burnt Prediction Using Machine Learning Approach. Rangamati Science and Technology University, Bangladesh.

8. Panwar, P., Bhutani, K., Sharma, R., & Saini, R. (2023). A Study on Calories Burnt Prediction Using Machine Learning. Jaipur Engineering College & Research Centre, Computer Science Department, Jaipur, Rajasthan, India.

9. Kadam, A. (2023). Calories Burned Prediction Using Machine Learning.

10. Meenigea, N. (2014). Calorie Burn Prediction: A Machine Learning Approach using Physiological and Environmental Factors. JETIR.

11. (2023). A Study on Calories Burnt Prediction Using Machine Learning. ITM Web of Conferences, 54, 01010.

12. Nipas, M., Acoba, A.G., Mindoro, J.N., Malbog, M.A.F., & Susa, J.A.B. (2019). Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA).