**PROJECT REPORT**

**ON**

**CALORIE BURN PREDICTION**

**(CSE III SEMESTER MINI PROJECT)**



**GUIDED BY-**

**PROF.AMIT GUPTA**

**SUBMITTED BY-**

**ADITYA PHARSWAN**

**ROLL NO- 2318218**

**CSE-K-III-SEM**

**SESSION-2024-25**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA HILL UNVERSITY, DEHRADUN**

**CERTIFICATE**

### Certified that Mr. Aditya Pharswan (Roll No.- 2318218) has developed mini project on “Calorie burn prediction” for the CS III Semester Mini Project in Graphic Era Hill University, Dehradun. The project carried out by Students is their own work as best of my knowledge.

Date : 11/01/2025

**Prof. Akash Chauhan**

Class Co-ordinator

**CSE-K-III-Sem**

GEHU Dehradun

**Project Guide**

**Prof. Amit Gupta**

(CSE Department)

GEHU Dehradun

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to everyone who contributed to the successful completion of this mini project.

First and foremost, I extend my heartfelt thanks to my mentor,

**Prof. Amit Gupta**, for their invaluable guidance, support, and encouragement throughout this project. Their expertise and insights have been instrumental in shaping the outcome of this work.

I am also grateful to my institution, **Graphic Era Hill University**, and the faculty members of the **CSE Department** for providing the resources and learning environment necessary to undertake this project.

Lastly, I would like to thank my family, friends, and peers for their constant motivation and encouragement during this journey.

Thank you all for your support and belief in my efforts.

**Mr. Aditya Pharswan**

**Roll No. - 2318218**

**CSE-K-III-Sem**

**Session: 2024-2025**

**GEHU, Dehradun**

**TABLE OF CONTENTS**

**CHAPTER 1: INTRODUCTION**

* 1. Objective of Project
  2. Motivation
  3. Scope of Project

**CHAPTER 2: PROJECT DETAILS**

* 1. Technical Stack
  2. Data Collection
  3. Data Pre-processing
  4. Exploratory Data Analysis
  5. Feature Engineering
  6. Model Development
  7. Model Evaluation

**CHAPTER 3: RESULT AND SNAPSHOTS**

* 1. Visualization and Results
  2. Model accuracy

**CHAPTER 4: CONCLUSION**

* 1. Summary of Achievements
  2. Challenges Faced
  3. Future Work and Learning Outcomes

**CHAPTER 1**

**INTRODUCTION**

* 1. **Objective of Project**

The objective of this project is to develop an efficient and user-friendly calorie prediction system using machine learning techniques. This system aims to accurately predict the number of calories burned based on user inputs such as age, gender, height, weight, exercise duration, heart rate, and body temperature.

The primary goals include:

1. **Promoting Health Awareness:**  
   To help individuals understand the energy expenditure associated with their physical activities and maintain a healthy lifestyle.
2. **Advancing Machine Learning Applications:**  
   Leveraging advanced machine learning models, specifically the XGBoost Regressor, to enhance the accuracy and efficiency of predictions.
3. **Data-Driven Decision Making:**  
   Utilizing exploratory data analysis and visualization to uncover trends and correlations in the dataset, thereby improving the understanding of factors influencing calorie burn.

**1.2 Motivation**

The motivation behind this project arises from the increasing importance of maintaining a healthy lifestyle and the role of technology in enabling personalized health tracking

By offering personalized calorie predictions using data such as age, gender, weight, heart rate, and body temperature, the system provides tailored insights for fitness routines.

Leveraging advancements in machine learning, specifically XGBoost, the project creates accurate, efficient, and scalable predictive models for fitness tracking.

Additionally, the project fosters educational and professional growth, enhancing coding skills and knowledge in the intersection of technology and healthcare.

**1.3 Scope of Project**

The scope of this project encompasses the development and application of a machine learning-based calorie prediction system, transforming raw data into valuable insights about energy expenditure.

Below are the key dimensions of its scope:

#### 1. **Data Analysis and Insights :** conducting exploratory data analysis (EDA) to uncover trends and correlations in calorie consumption across various factors, including gender, age, weight, and exercise duration. It also provides enhanced data interpretability through visual insights such as heatmaps, statistical summaries, and other data visualizations.

2. **Predictive System Development *:*** Building an efficient predictive model using the XGBoost Regressor to estimate calories burned based on input parameters such as heart rate, body temperature, and exercise details. Model accuracy is ensured by evaluating performance metrics like Mean Absolute Error (MAE).

#### 3. **Applications in Health and Fitness :** Assisting fitness trainers and health professionals in designing personalized workout plans based on precise calorie predictions.

#### 4. **Scalability and Integration :** Future development could include real-time calorie tracking and integration with other health metrics, such as sleep and nutrition data.

5. **Future Enhancements *:*** Expanding the dataset by incorporating additional features such as exercise type, fitness goals, and environmental conditions for more detailed predictions and also enhancing model performance through hyperparameter tuning and the integration of more sophisticated algorithms.

**CHAPTER 2**

**PROJECT DETAILS**

* 1. **Technical Stack**

programming tools, libraries, and platforms; these components are carefully chosen to ensure efficient data handling, accurate predictions, and seamless project execution.

**Programming Language:** The project is implemented in Python, a versatile and widely-used programming language known for its simplicity and extensive ecosystem of libraries that cater to data science and machine learning tasks.

**Data Manipulation and Analysis:**

* **Pandas:** Used for loading, cleaning, and pre-processing the dataset, ensuring the data is structured and ready for analysis.
* **NumPy:** Utilized for numerical computations and array manipulations to support data transformations and feature engineering.

**Visualization Tools:**

* **Matplotlib and Seaborn:** Employed to create insightful visualizations, such as heatmaps and statistical plots, to analyse data trends and correlations effectively.

**Machine Learning Frameworks:**

* **Scikit-learn:** Leveraged for splitting the dataset into training and testing sets, as well as evaluating model performance using metrics like Mean Absolute Error (MAE).
* **XGBoost:** A powerful and efficient machine learning library used for training the regression model, chosen for its ability to handle large datasets and deliver high prediction accuracy.
  1. **Data Collection**

The data collection process is critical to the development of this calorie prediction system. This project utilizes two key datasets: the **Calories Dataset** and the **Exercise Dataset**, both of which provide the essential information required for analysis and model development.

The **Calories Dataset** contains information about the number of calories burned during various physical activities and **Exercise Dataset** includes demographic and physiological parameters such as gender, age, height, weight, exercise duration, heart rate, and body temperature.

The two datasets were merged into a single dataframe by aligning records based on common identifiers, the resulting dataframe included all input features and the target variable (calories burned).

The combined dataset was thoroughly inspected for its dimensions, including the number of rows and columns. Each column's data type was reviewed to ensure it was compatible with the subsequent processing steps. Additionally, missing values were identified and addressed to maintain the consistency and integrity of the data.

To gather reliable and comprehensive data that forms the foundation for exploratory analysis, visualization, and the development of machine learning models. Ensuring that the dataset accurately represents the factors influencing calorie burn is crucial for enabling the system to make precise predictions.

* 1. **Data Pre-processing**

Data pre-processing is a crucial step in the development of this calorie prediction system, ensuring the dataset is clean, structured, and ready for analysis and model training.

The steps involved in data pre-processing are as follows:

1. **Combining Datasets:** The **Calories Dataset** and **Exercise Dataset** were merged into a single DataFrame to consolidate all relevant features and the target variable.
2. **Handling Missing Values:** The combined dataset was checked for missing values using functions like isnull().sum(). Any missing or null values were handled appropriately to maintain data integrity and prevent biases in the model.
3. **Data Cleaning:** Unnecessary columns, such as User\_ID, were dropped as they do not contribute to the prediction process, any anomalies or inconsistencies in the data were addressed during this step.
4. **Encoding Categorical Data:** The "Gender" column, which contained text values ("male" and "female"), was converted into numerical format (0 for male and 1 for female) using the where() function, ensuring compatibility with machine learning algorithms.
5. **Feature and Target Separation:** The features (input variables) were separated from the target variable (calories burned).Features included physiological and exercise parameters such as gender, age, height, weight, duration, heart rate, and body temperature and the target variable was identified as the "Calories" column.
6. **Data Splitting:** The dataset was split into training and testing sets using an 80-20 split ratio. This was achieved using the train\_test\_split function from Scikit-learn to ensure the model is trained and evaluated on distinct subsets of the data.
7. **Ensuring Data Readiness:** The data was reshaped where necessary to match the input format expected by the machine learning model. This included converting the input data into NumPy arrays and reshaping it into a format compatible with the prediction system.

**1.4 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a critical step in understanding the dataset and identifying patterns, trends, and relationships among variables.

Descriptive statistics were first calculated using the describe() function, providing essential metrics like mean, median, standard deviation, and the range (minimum/maximum) for all numerical columns, which helped summarize the data distribution.

Correlations between variables were examined using the corr() function, highlighting relationships between features such as exercise duration, heart rate, and calories burned. A heatmap was then created to visually represent these correlations, making it easier to spot strong positive and negative relationships; notably, exercise duration and heart rate had a strong positive correlation with calories burned.

Boxplots were employed to analyze the distribution of calories and identify any potential outliers.

Key insights from the EDA showed that exercise duration and heart rate were the most influential factors in determining calories burned, with gender also playing a role, as different trends in calorie expenditure were observed between males and females.

**1.5 Feature Engineering**

Feature engineering is a crucial step in improving the performance of machine learning models by creating meaningful features from the raw data. In this calorie prediction system, feature engineering was performed to enhance the model’s predictive power and ensure that the dataset is well-prepared for training.

In the data preprocessing phase, categorical data, specifically the "Gender" column, was encoded into numerical values where "male" was replaced by 0 and "female" by 1 to make it compatible with machine learning algorithms.

For feature selection, irrelevant features like User\_ID were dropped, and relevant features such as age, height, weight, duration, heart rate, and body temperature were retained based on their correlation with the target variable (calories burned).

Missing values were identified and addressed by either filling them with the mean or median or removing rows with missing target values. Feature scaling was considered to ensure that features with larger numerical ranges (e.g., weight, height) did not dominate the learning process.

Interaction features, such as combinations of exercise duration and heart rate, were considered to potentially provide more detailed insights into calorie expenditure. Lastly, transformations of the target variable, calories burned, could have been applied to reduce skewness and improve model accuracy, though this step was optional.

**1.6 Model Development**

Model development is the core of the calorie prediction system, where machine learning algorithms are applied to learn from the data and make accurate predictions. The model development process for this project involved several key steps, including choosing the right algorithm, training the model, and evaluating its performance.

The dataset was split into training and testing sets using an 80-20 ratio, ensuring that the model was trained on 80% of the data and evaluated on the remaining 20%.

The XGBoost model was trained on the training data, learning the relationship between features and the target variable (calories burned). Hyperparameters were set to default values, though further fine-tuning could improve performance.

After training, the model was evaluated using the testing set, and Mean Absolute Error (MAE) was computed to assess predictive accuracy.

**1.7 Model Evaluation**

Model evaluation is a critical step to determine how well the trained model performs on unseen data. For this calorie prediction system, various metrics and techniques were used to assess the accuracy and effectiveness of the **XGBoost Regressor**.

**Prediction on Test Data:** After training the model on the training dataset, it was used to make predictions on the test data, The model’s predictions were generated using the predict() function, which takes the test features as input and returns the predicted calorie values.

**Performance Metrics:**

* + **Mean Absolute Error (MAE):**
    - The Mean Absolute Error (MAE) was the primary metric used to evaluate the model’s performance.
    - MAE measures the average absolute difference between the actual and predicted values.
    - The MAE provides a clear understanding of how far off the model's predictions are from the actual values, in terms of calories burned.

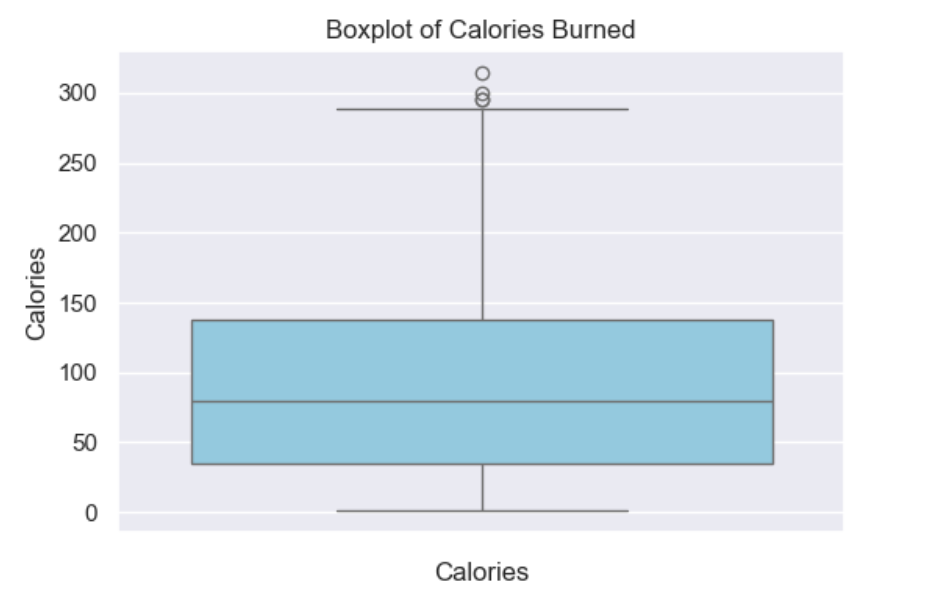
The model's predicted calorie values were compared with the actual values from the test dataset to assess its performance. An error analysis was conducted to investigate any patterns in the model's predictions, such as biases or systematic errors, by examining residuals (differences between predicted and actual values).

This analysis helped identify any consistent overestimation or underestimation of calorie burn for specific user groups. The Mean Absolute Error (MAE) was used as a key metric to summarize the model's accuracy.

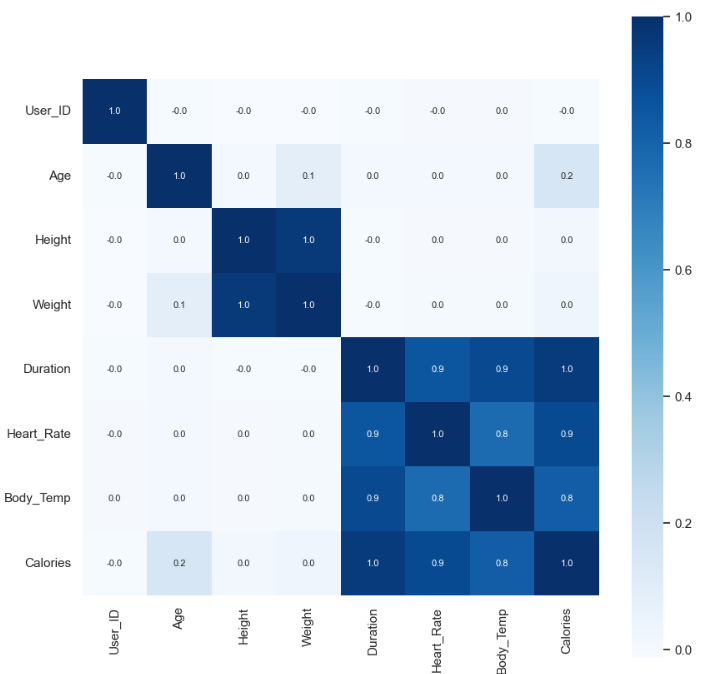
**CHAPTER 3**

**RESULTS AND SNAPSHOTS**

* 1. **Visualization and Results**
* Boxplot to analyze the distribution of calories



* heatmap to understand the correlation



* Scatter plot for visualization of Age vs. Calories



* 1. **Model Accuracy**
  + **Model Performance** :

Mean Absolute Error (MAE) : 1.48 (lower value is better)

* + **Sample Prediction** :
    - Input: Gender = 0, Age = 41, Height = 175 cm, Weight = 85 kg, Duration = 25 min, Heart Rate = 100 bpm, Body Temp = 40.7°C.
    - Predicted Calories Burned : 142.48 (actual value = 143 )

**CHAPTER 4**

**CONCLUSION**

* 1. **Summary of Achievements**

This project successfully developed a predictive system for calorie expenditure using machine learning. Key achievements include:

**Data Preprocessing**: Cleaned and encoded categorical data, handled missing values, and selected relevant features, ensuring the dataset was ready for modeling.

**Exploratory Data Analysis**: Conducted statistical analysis and visualized relationships between features, identifying key factors such as exercise duration and heart rate influencing calorie burn.

**Model Development**: Built and trained an XGBoost Regressor model, achieving accurate predictions by splitting the data into training and testing sets.

**Model Evaluation**: Assessed model performance using Mean Absolute Error (MAE) and conducted error analysis to identify potential biases.

**Actionable Insights**: Derived insights on key variables affecting calorie burn, with potential for further model improvement through hyperparameter tuning or feature engineering.

* 1. **Challenges faced**

**Data Cleaning**: Handling missing values and encoding categorical data required careful attention to avoid errors and ensure compatibility with the model.

**Feature Selection**: Identifying the most relevant features while avoiding overfitting was a challenge, requiring efficient feature selection techniques.

**Model Tuning**: Balancing model complexity to prevent overfitting or underfitting and fine-tuning hyperparameters to optimize performance proved time-consuming.

**Outliers**: Managing outliers in the data, particularly in the distribution of calories, required careful handling to maintain model accuracy.

**Evaluation Metrics**: Choosing the appropriate evaluation metric (e.g., MAE) and ensuring accurate performance interpretation was critical.

* 1. **Future and Learning Outcomes**

**Learning Outcomes** : Gained experience in data preprocessing, applying XGBRegressor for regression and evaluating model performance.

**Future Application** : Can be used in fitness apps for personalized calorie burn estimates and integrated into wearable devices for real-time tracking.

**REFERENCES**

1. **Pandas Documentation**: McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference (SciPy 2010).
2. **NumPy Documentation**: Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357-362
3. **Scikit-learn Documentation**: Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. **XGBoost Documentation**: 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
5. **Matplotlib Documentation**: Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
6. **IBM(machine learning using python)**