FITBURN

MACHINE LEARNING PROJECT

Prepared By :
Herath H.M.H.K

Index No:
258777A

Contents

Introduction

In recent years, Sri Lanka has experienced a significant shift in lifestyle patterns due to rapid urbanization, increased use of technology, and changes in occupational structures. A large portion of the working population is now engaged in sedentary jobs, while physical activity levels among both adults and young people have gradually declined. At the same time, dietary habits have shifted towards higher calorie consumption, influenced by the growing availability of fast food, processed meals, and irregular eating schedules. These lifestyle changes have contributed to a noticeable rise in non-communicable diseases such as obesity, diabetes, cardiovascular diseases, and metabolic disorders across the country. According to national health statistics, Sri Lanka faces an increasing burden of lifestyle-related illnesses, making personal health monitoring and preventive care a critical national concern.

One of the key challenges in maintaining a healthy lifestyle is understanding how daily physical activities impact calorie expenditure. Calories burned during physical exercise vary significantly from person to person, depending on factors such as age, gender, body composition, heart rate, body temperature, and exercise duration. In the Sri Lankan context, many individuals engage in physical activities such as walking, jogging, cycling, gym workouts, and home-based exercises without having a clear understanding of how effective these activities are in terms of calorie burning. This lack of awareness often leads to unrealistic fitness expectations, ineffective workout routines, and poor health decisions, especially among beginners who are trying to lose weight or manage medical conditions.

Traditionally, calorie estimation methods rely on generalized formulas or fitness charts that assume average human characteristics. However, these approaches fail to capture individual physiological differences and may produce inaccurate results. For example, two individuals of the same weight performing the same activity may burn different amounts of calories due to variations in heart rate, metabolic efficiency, and body temperature. In Sri Lanka, where access to advanced fitness tracking devices is still limited for many people, relying solely on wearable technologies is not always feasible. Therefore, there is a strong need for a more accessible, data-driven, and personalized solution that can accurately predict calorie burn using commonly available personal and physiological data.

Machine learning provides an effective solution to this problem by learning complex patterns from historical data and generating accurate predictions without relying on rigid mathematical formulas. By analyzing multiple features simultaneously, machine learning models can capture non-linear relationships between physiological factors and calorie expenditure. In this project, calorie burn prediction is treated as a regression problem, where the goal is to estimate the number of calories burned based on user-specific inputs such as age, gender, height, weight, heart rate, body temperature, and workout duration. This approach allows for personalized predictions that are more realistic and reliable compared to traditional estimation techniques.

The importance of such a system is particularly relevant in the Sri Lankan healthcare and fitness ecosystem. Many individuals begin exercise routines based on advice from social media, trainers, or peers, without objective feedback on their actual energy expenditure. A calorie prediction system can act as a decision-support tool, helping users evaluate whether their workout intensity and duration align with their fitness goals. Furthermore, for individuals managing chronic conditions such as diabetes or obesity, understanding calorie burn is essential for maintaining proper energy balance and preventing health complications. A predictive system can therefore contribute to both preventive healthcare and lifestyle management.

Another major limitation of many existing machine learning solutions is the lack of user-friendly interfaces. Even when accurate models are developed, they often remain confined to notebooks or research environments, making them inaccessible to non-technical users. In Sri Lanka, where digital literacy levels vary widely, it is essential that predictive systems are presented through simple and intuitive interfaces. Integrating the trained calorie prediction model into a web-based application ensures that users can easily input their details and receive instant predictions without requiring technical knowledge. This aligns with the growing adoption of web applications and mobile-friendly platforms across the country.

This project addresses the identified problem by developing a complete machine learning based calorie prediction system integrated into a web application. The system allows users to enter personal and workout-related information and instantly receive an estimate of calories burned. By combining data preprocessing, feature scaling, regression modeling, and front-end integration, the project demonstrates how machine learning can be applied to solve a real-world health problem in a Sri Lankan context. The application emphasizes accessibility, personalization, and practical usefulness, making it suitable for fitness enthusiasts, students, office workers, and individuals seeking to improve their overall health.

In summary, the calorie prediction problem represents a meaningful and practical application of machine learning that addresses a growing health concern in Sri Lanka. By moving beyond generalized formulas and adopting a personalized, data-driven approach, this project contributes to smarter fitness decision-making and increased health awareness. The integration of the predictive model into an interactive web application further enhances its real-world relevance, bridging the gap between machine learning theory and practical deployment. This makes the project not only technically valuable but also socially impactful within the Sri Lankan context.

Methodology

Dataset Collection

The dataset used in this project was synthetically generated to represent realistic fitness and physiological data relevant to the Sri Lankan population. Due to the absence of publicly available, large-scale, locally contextualized datasets for calorie burn prediction in Sri Lanka, a synthetic dataset was created based on domain knowledge, existing physiological research, and commonly observed fitness patterns among Sri Lankan individuals. The dataset generation process was designed to ensure realism while maintaining ethical compliance by avoiding the use of sensitive or personally identifiable real-world data.

The synthetic dataset reflects common demographic and physiological characteristics found within Sri Lanka, such as age distributions typical of working adults and youth, gender proportions, body measurements aligned with South Asian body compositions, and exercise intensities commonly associated with activities such as walking, jogging, gym workouts, and home-based fitness routines. By incorporating these contextual factors, the dataset aims to closely simulate real-world usage scenarios while remaining suitable for academic experimentation and machine learning model development.

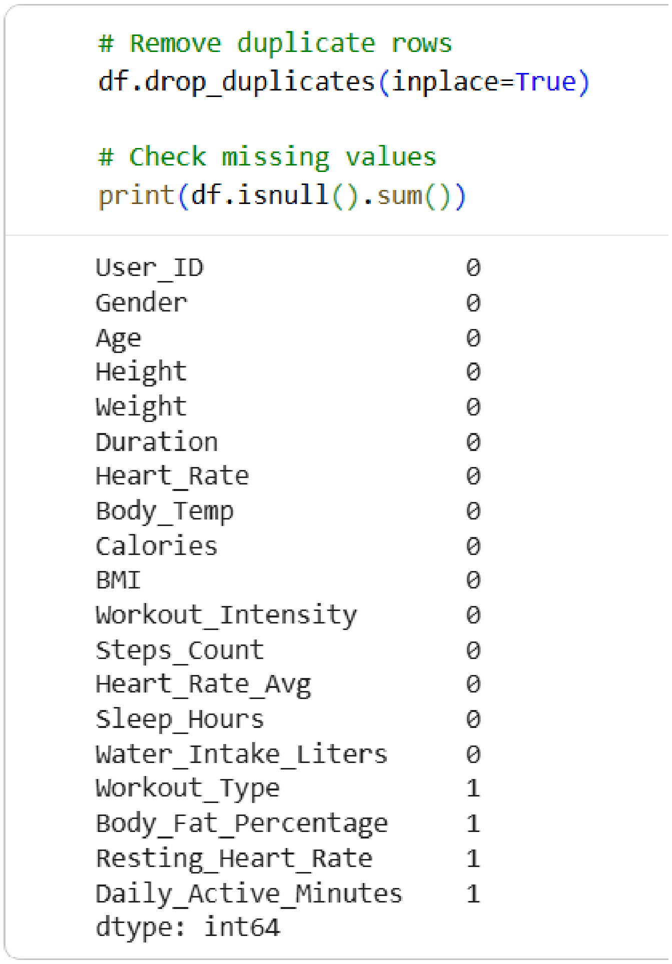
The dataset consists of approximately several hundred records, each representing an individual exercise session. The feature set includes both demographic and physiological attributes. Demographic features include gender and age, which influence metabolic rate and energy expenditure. Physical attributes such as height and weight are included to represent body composition, which plays a significant role in calorie burn. Physiological measurements such as heart rate and body temperature are incorporated to capture exercise intensity and metabolic response during physical activity. Additionally, workout duration is included to represent the time component of energy expenditure.

The target variable of the dataset is the total number of calories burned during the exercise session. This value is generated using realistic physiological relationships between the input features, ensuring that higher intensity workouts, longer durations, and elevated heart rates correspond to higher calorie expenditure. By defining calorie burn as a continuous numerical variable, the problem is formulated as a regression task. Overall, the dataset structure enables the machine learning model to learn complex interactions between multiple input features and predict calorie burn in a manner that is both realistic and applicable to the Sri Lankan fitness context.

Data Preprocessing

Before training the machine learning model, several data preprocessing techniques were applied to improve data quality, ensure consistency, and enhance model performance. Preprocessing is a critical step in any machine learning pipeline, as raw data often contains inconsistencies, scale variations, and potential noise that can negatively impact learning.

Initially, the dataset was examined for missing values and inconsistencies. Since the data was synthetically generated, missing values were minimal; however, validation checks were performed to ensure that all features contained valid numerical or categorical values within realistic ranges. Any out-of-range values, such as abnormal heart rates or body temperatures beyond human physiological limits, were identified and corrected or removed to maintain data integrity.



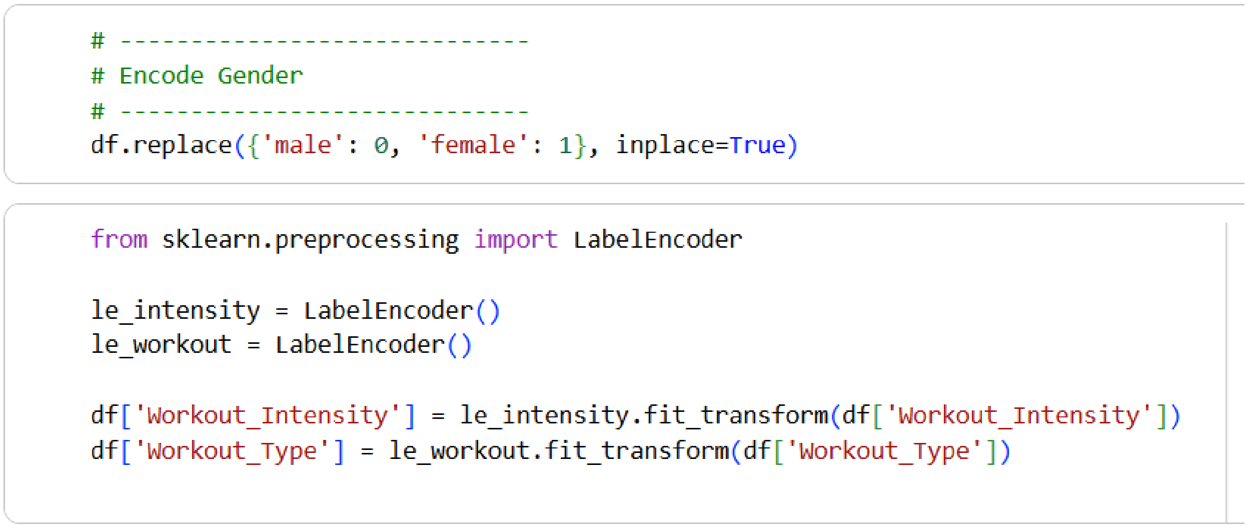
```python
# Remove duplicate rows
df.drop_duplicates(inplace=True)

# Check missing values
print(df.isnull().sum())
```

```
User_ID                 0
Gender                  0
Age                     0
Height                  0
Weight                  0
Duration                0
Heart_Rate              0
Body_Temp               0
Calories                0
BMI                     0
Workout_Intensity       0
Steps_Count             0
Heart_Rate_Avg          0
Sleep_Hours             0
Water_Intake_Liters     0
Workout_Type            1
Body_Fat_Percentage     1
Resting_Heart_Rate      1
Daily_Active_Minutes    1
dtype: int64
```

Figure 1 - Data Preprocessing

Categorical variables, such as gender, were converted into numerical representations using encoding techniques to ensure compatibility with the machine learning algorithm. Numerical features including age, height, weight, heart rate, body temperature, and duration were analyzed for scale differences. Because these features exist in different units and ranges, feature scaling was applied using normalization or standardization techniques. This step ensures that no single feature dominates the learning process purely due to its numerical magnitude.



```python
# ----------------------------
# Encode Gender
# ----------------------------
df.replace({'male': 0, 'female': 1}, inplace=True)
```

```python
from sklearn.preprocessing import LabelEncoder

le_intensity = LabelEncoder()
le_workout = LabelEncoder()

df['Workout_Intensity'] = le_intensity.fit_transform(df['Workout_Intensity'])
df['Workout_Type'] = le_workout.fit_transform(df['Workout_Type'])
```

Figure 2 - Encoding

Outlier analysis was also conducted to identify extreme values that could disproportionately influence model training. Although some variability is expected in fitness data, excessively extreme observations were treated carefully to balance realism with model stability. The dataset was then split into training and testing subsets to evaluate model generalization. By applying these preprocessing steps, the dataset was transformed into a clean, structured, and model-ready format suitable for regression analysis.

Selecting the machine learning algorithm

Selecting an appropriate machine learning algorithm is a crucial step in the development of a reliable predictive system, particularly when the problem domain involves human health and physiological behavior. In this project, the objective is to predict the number of calories burned during physical activity using a set of demographic and physiological input variables. Since calorie burn is a continuous numerical value rather than a categorical label, the task is formulated as a supervised regression problem. The choice of regression algorithm directly influences the accuracy, robustness, interpretability, and real-world applicability of the developed system. After evaluating the nature of the problem and the characteristics of the dataset, the XGBoost Regressor was selected as the most suitable machine learning algorithm for this application.

The calorie burn prediction problem exhibits several challenging characteristics that make simple regression models insufficient. Human energy expenditure is influenced by multiple interdependent factors such as age, gender, body weight, height, heart rate, body temperature, and exercise duration. These variables interact in complex and non-linear ways. For example, an increase in workout duration does not result in a uniform increase in calorie burn across all individuals; the effect varies depending on exercise intensity, cardiovascular response, and body composition. In the Sri Lankan context, these variations are further influenced by lifestyle patterns, fitness levels, and environmental factors such as climate. Therefore, the selected machine learning algorithm must be capable of learning non-linear relationships and capturing interactions between multiple features without relying on oversimplified assumptions.

XGBoost, which stands for Extreme Gradient Boosting, is an advanced ensemble learning algorithm based on the gradient boosting framework. Ensemble learning refers to the technique of combining multiple weak learners to form a strong predictive model. In XGBoost, the weak learners are decision trees, and they are built sequentially in a manner that allows each new tree to correct the errors made by the previous ensemble. This boosting mechanism enables the model to progressively improve its predictions by focusing on difficult-to-predict data points. Unlike traditional decision tree models, which may suffer from instability and overfitting, XGBoost constructs a robust ensemble that generalizes well to unseen data.

One of the primary reasons for choosing XGBoost Regressor in this project is its ability to handle non-linear relationships inherently. Traditional linear regression models assume a linear relationship between input variables and the target variable. While linear regression is simple and interpretable, it fails to model the true complexity of physiological processes such as calorie burning. Accurately modeling such processes would require extensive manual feature engineering, including polynomial features and interaction terms, which increases model complexity and reduces interpretability. XGBoost eliminates the need for such manual transformations by learning non-linear decision boundaries directly from the data through its tree-based structure.

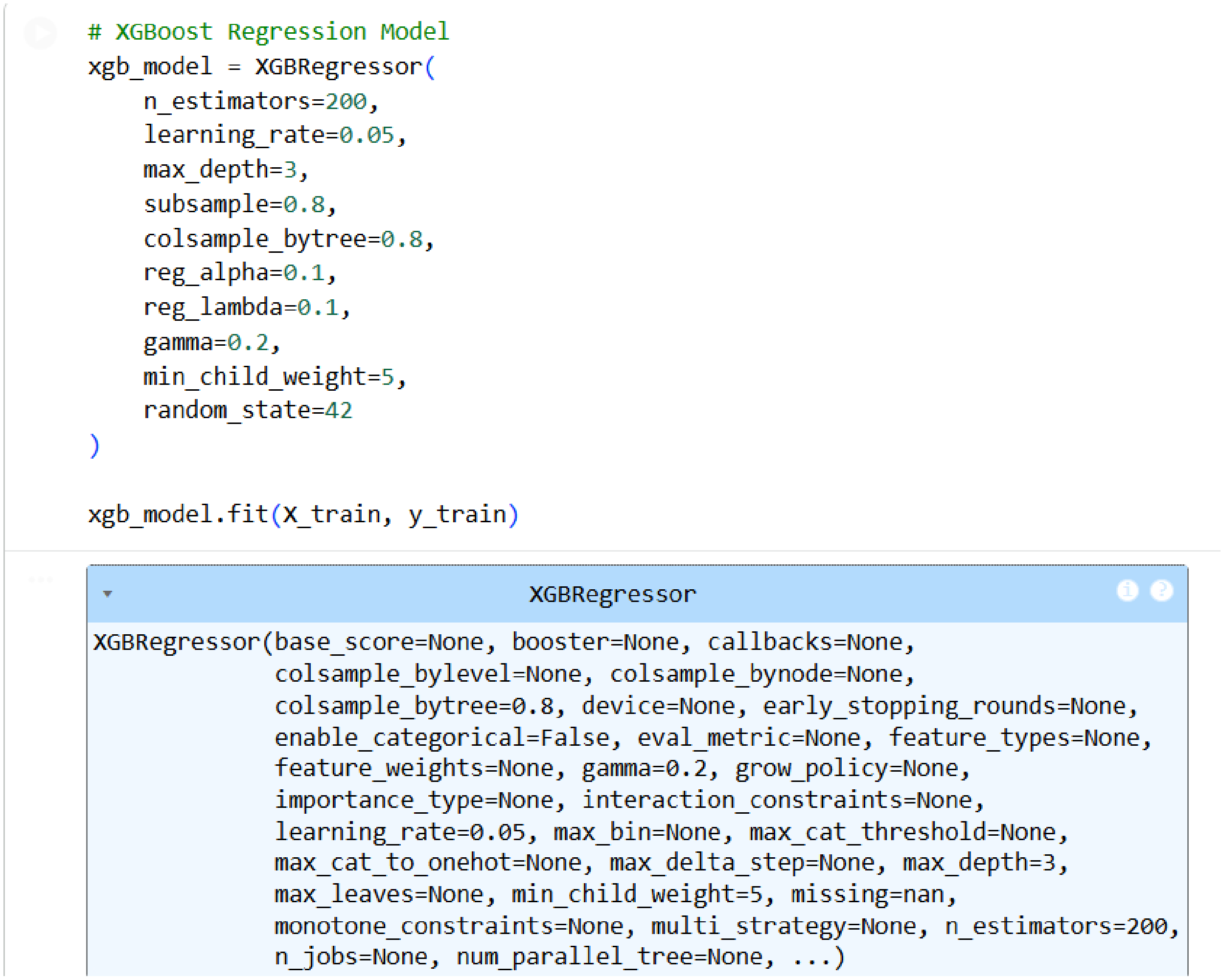
Another significant advantage of XGBoost is its robustness against overfitting, which is particularly important when working with synthetic or moderately sized datasets. Overfitting occurs when a model learns noise or specific patterns in the training data that do not generalize to new data. XGBoost addresses this issue through built-in regularization techniques, including L1 and L2 regularization on leaf weights. These regularization mechanisms penalize overly complex models and encourage simpler, more generalizable structures. Additionally, hyperparameters such as maximum tree depth, learning rate, subsampling ratio, and column sampling allow fine control over model complexity. This makes XGBoost a reliable choice for achieving balanced performance without excessive variance.

When compared to standard decision tree regression, XGBoost offers substantial improvements in stability and predictive accuracy. A single decision tree can easily overfit the training data, especially when the tree becomes deep. Furthermore, small changes in the dataset can lead to entirely different tree structures, resulting in unstable predictions. XGBoost mitigates these issues by combining multiple trees into an ensemble and by training them sequentially to minimize prediction error. This ensemble-based learning process produces smoother and more consistent predictions, which is essential for a health-related application where reliability is important.

Random Forest regression is another popular ensemble learning technique that could be considered for this problem. Random Forests build multiple decision trees independently using random subsets of data and features, and the final prediction is obtained by averaging the outputs of all trees. While Random Forests improve robustness compared to single decision trees, they lack the error-correcting mechanism present in boosting algorithms. XGBoost differs fundamentally in that each new tree is trained with explicit awareness of the errors made by previous trees. This targeted learning strategy allows XGBoost to achieve higher accuracy with fewer trees and better capture subtle patterns in the data, making it more effective for regression tasks involving complex relationships.

Compared to more complex machine learning approaches such as deep neural networks, XGBoost offers a favorable balance between performance and interpretability. Neural networks are powerful function approximators capable of modeling highly non-linear relationships; however, they often require large amounts of data, extensive hyperparameter tuning, and significant computational resources. Moreover, neural networks are typically considered black-box models, making it difficult to explain individual predictions. In contrast, XGBoost provides feature importance measures and integrates seamlessly with explainability techniques such as SHAP, enabling transparent interpretation of model behavior. This is particularly important in a calorie prediction application, where users and evaluators must be able to understand how different factors influence the predicted outcome.

From a practical implementation perspective, XGBoost is computationally efficient and well-suited for academic and real-world environments. The algorithm is optimized for speed and memory efficiency and supports parallel processing, which allows faster training even on standard computing resources such as Google Colab or personal laptops. This efficiency enabled iterative experimentation with different hyperparameter settings during model development, facilitating performance optimization without excessive computational cost.



```python
# XGBoost Regression Model
xgb_model = XGBRegressor(
    n_estimators=200,
    learning_rate=0.05,
    max_depth=3,
    subsample=0.8,
    colsample_bytree=0.8,
    reg_alpha=0.1,
    reg_lambda=0.1,
    gamma=0.2,
    min_child_weight=5,
    random_state=42
)

xgb_model.fit(X_train, y_train)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.8, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             feature_weights=None, gamma=0.2, grow_policy=None,
             importance_type=None, interaction_constraints=None,
             learning_rate=0.05, max_bin=None, max_cat_threshold=None,
             max_cat_to_onehot=None, max_delta_step=None, max_depth=3,
             max_leaves=None, min_child_weight=5, missing=nan,
             monotone_constraints=None, multi_strategy=None, n_estimators=200,
             n_jobs=None, num_parallel_tree=None, ...)
```

Figure 3- Hyperparameters of XGBoost

The training process of the XGBoost Regressor involved dividing the preprocessed dataset into training and testing subsets to evaluate generalization performance. During training, the algorithm iteratively added decision trees to minimize a loss function that measures the difference between predicted and actual calorie values. At each iteration, gradients of the loss function were computed, and a new tree was fitted to correct the residual errors. This process continued until the model converged or reached a predefined number of trees. Evaluation metrics such as Mean Absolute Error, Root Mean Squared Error, and R² score were used to assess the effectiveness of the trained model and ensure that it met the performance expectations of the project.

In the Sri Lankan context, the choice of XGBoost is particularly appropriate due to the variability in fitness behavior and physiological characteristics among individuals. The model's ability to adapt to diverse patterns and learn from multiple interacting features makes it suitable for representing realistic calorie burn scenarios. Furthermore, its compatibility with explainability frameworks and front-end integration tools enhances its suitability for responsible deployment in a user-facing application.

In conclusion, the selection of XGBoost Regressor for this project is justified by its strong predictive performance, ability to model complex non-linear relationships, robustness against overfitting, and superior interpretability compared to many standard machine learning models. When compared with linear regression, single decision trees, Random Forests, and neural networks, XGBoost offers a balanced and practical solution tailored to the calorie burn prediction problem. These characteristics make it an ideal algorithm for developing a reliable, interpretable, and context-aware calorie prediction system suitable for the Sri Lankan fitness domain.

SHAP Analysis

To enhance the transparency and interpretability of the calorie prediction model, SHAP (SHapley Additive exPlanations) was employed as the primary explainability technique. While the XGBoost regressor demonstrates strong predictive performance, understanding how individual input features contribute to its predictions is essential, particularly in health and fitness applications. SHAP provides a principled approach to model interpretation by quantifying the contribution of each feature to a given prediction, based on concepts derived from cooperative game theory. This enables both global and local explanations, allowing insight into overall model behavior as well as individual prediction outcomes.

The SHAP summary analysis reveals that workout duration is the most influential feature in determining calorie burn predictions. This finding aligns well with physiological understanding, as longer exercise durations generally result in higher energy expenditure. Heart rate emerges as another highly significant contributor, reflecting the intensity of physical activity and the cardiovascular response of the individual. Elevated heart rates typically indicate higher metabolic activity, leading to increased calorie burn. Body temperature also shows a noticeable influence, as it serves as an indirect indicator of exertion level and metabolic response during exercise.

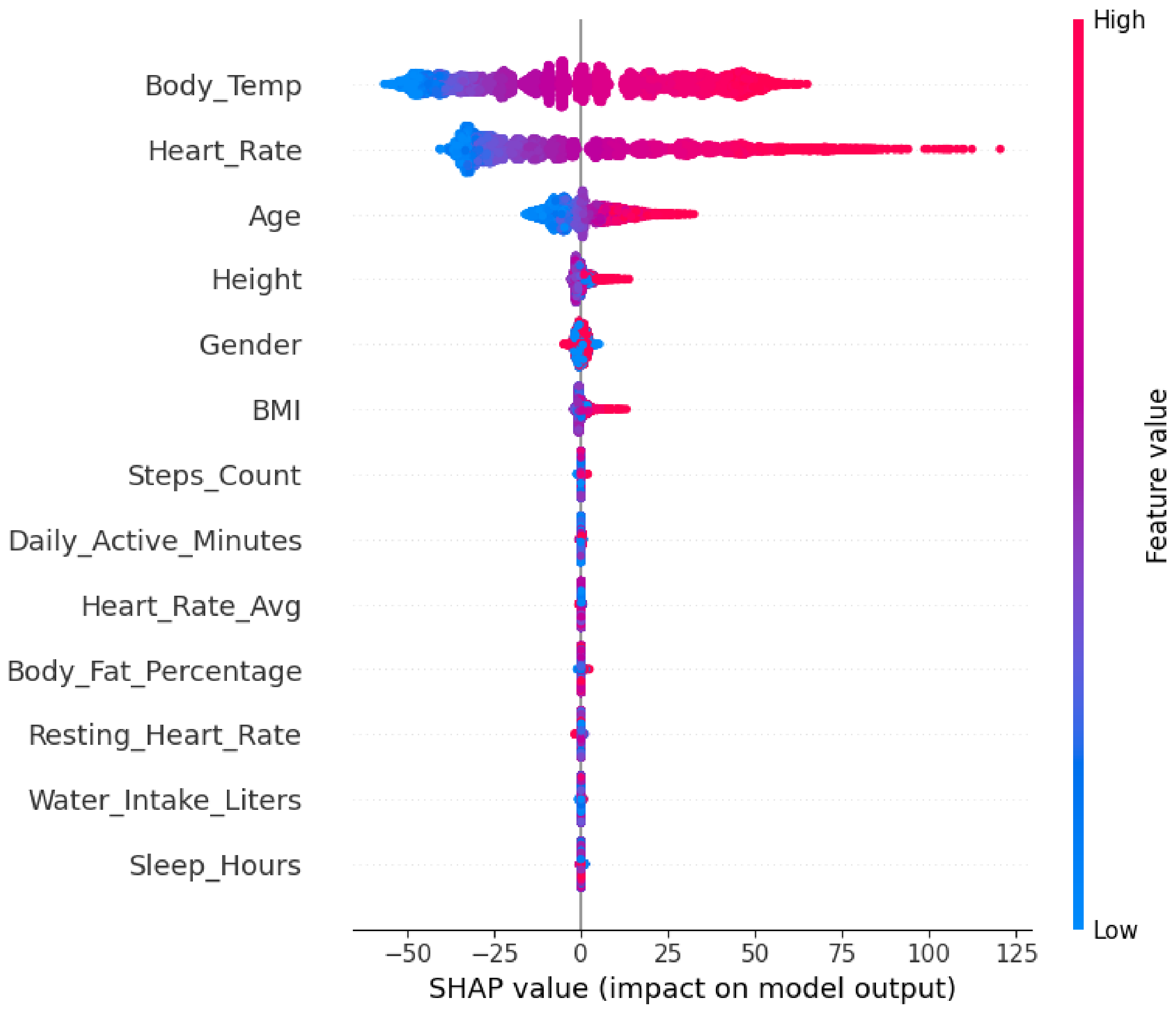


Figure 4 - SHAP summary plot

Features related to body composition, such as weight and height, contribute moderately to the model's predictions. Individuals with higher body mass tend to burn more calories for the same activity due to increased energy requirements. Age and gender exhibit comparatively lower, yet meaningful, contributions. These features capture differences in metabolic efficiency and physiological response, which vary across demographic groups. The presence of these effects in the SHAP results indicates that the model has learned realistic and biologically plausible relationships rather than relying on spurious correlations.

Local SHAP explanations further demonstrate how feature contributions differ across individuals. For example, in a high-calorie prediction scenario, longer workout duration and elevated heart rate contribute positively to the final output, while in lower-calorie predictions, shorter durations and lower intensity levels contribute negatively. This individualized interpretability is particularly valuable for user-facing applications, as it allows predictions to be explained in a personalized and intuitive manner.

Overall, the SHAP analysis confirms that the XGBoost model bases its predictions on meaningful physiological and demographic factors that are consistent with real-world fitness behavior. By providing transparent explanations for both global trends and individual predictions, SHAP enhances trust in the model and supports responsible use of the system. This interpretability is especially important in the Sri Lankan context, where users may rely on such applications for fitness guidance, and it ensures that the model's decisions remain understandable, justifiable, and aligned with domain knowledge.

Front End Integration

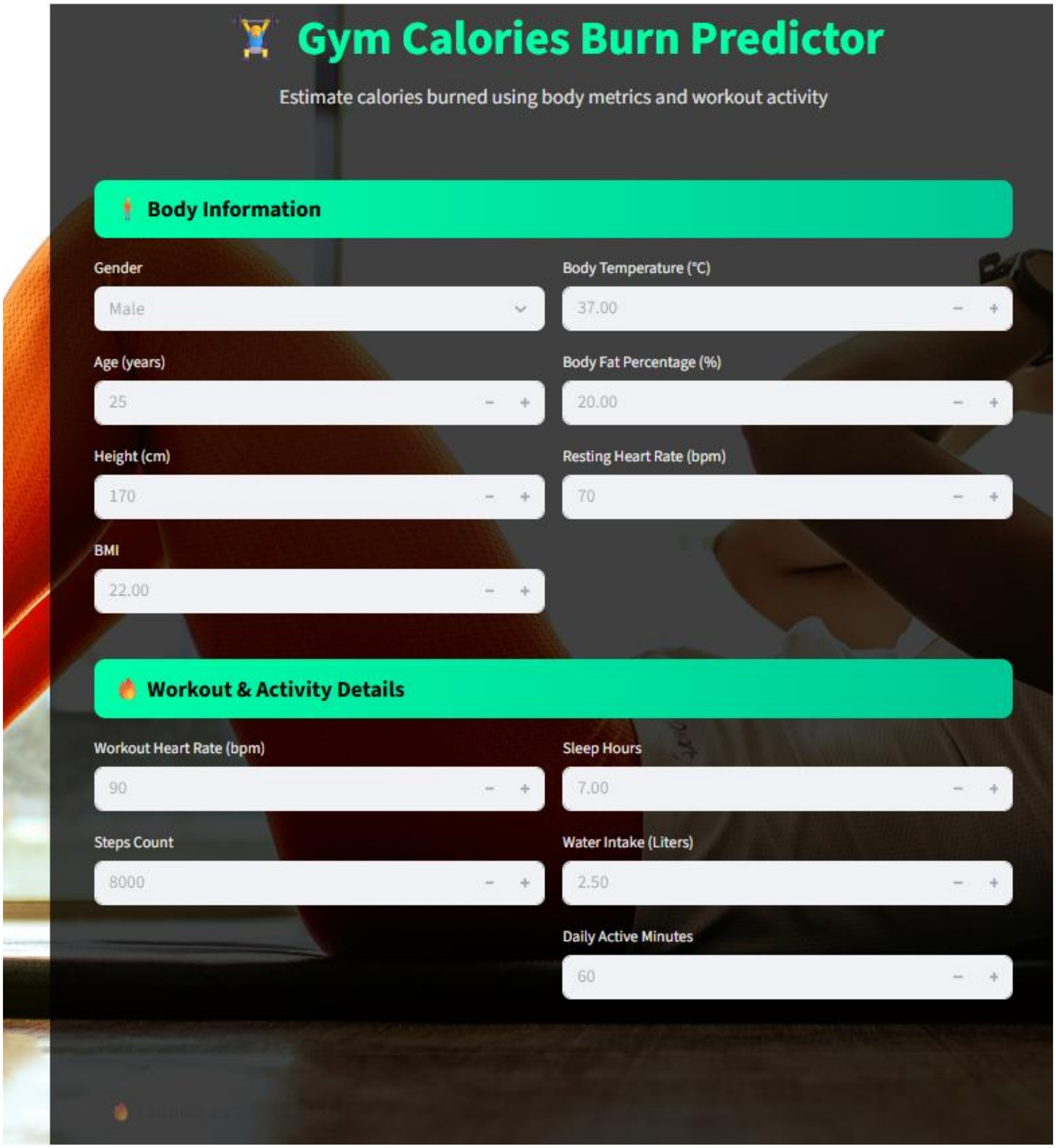
To ensure that the developed machine learning model is accessible and usable by non-technical users, a web-based front-end application was implemented using Streamlit. While machine learning models are typically developed and evaluated within programming environments such as Jupyter Notebook or Google Colab, their practical value is significantly enhanced when they are deployed through interactive user interfaces. In the context of this project, the front end serves as a bridge between the trained XGBoost regression model and end users who wish to estimate calorie expenditure during physical activity.

Streamlit was selected as the front-end framework due to its simplicity, rapid development capabilities, and seamless integration with Python-based machine learning workflows. Unlike traditional web frameworks that require extensive HTML, CSS, and JavaScript development, Streamlit enables the creation of interactive web applications directly using Python code. This makes it particularly suitable for academic projects and rapid prototyping, especially in environments with limited development resources.

The trained XGBoost model and associated preprocessing components, such as the feature scaler, were serialized and saved as files after model training. These files were then loaded into the Streamlit application at runtime to ensure consistency between the training and deployment phases. When a user enters input data through the web interface, the same preprocessing steps applied during model training are executed before generating predictions. This ensures that the input data format and scale match the conditions under which the model was trained, thereby maintaining prediction accuracy.

The user interface was designed to be intuitive and user-friendly, allowing users to input personal and workout-related information such as age, gender, height, weight, heart rate, body temperature, and exercise duration. Streamlit widgets such as sliders, dropdown menus, and numeric input fields were used to reduce input errors and improve usability. Clear labels and headings were included to guide users, and domain-appropriate styling elements were incorporated to create a fitness-oriented appearance suitable for a calorie prediction application.

Once the user submits the input data, the application processes the values, feeds them into the trained XGBoost regression model, and displays the predicted calorie burn in real time. This immediate feedback enhances user engagement and demonstrates the practical applicability of the machine learning model. The application also supports easy modification and extension, allowing future integration of additional features such as visual explanations, progress tracking, or personalized recommendations.



Gym Calories Burn Predictor

Estimate calories burned using body metrics and workout activity

Body Information

Gender: Male

Body Temperature (°C): 37.00

Age (years): 25

Body Fat Percentage (%): 20.00

Height (cm): 170

Resting Heart Rate (bpm): 70

BMI: 22.00

Workout & Activity Details

Workout Heart Rate (bpm): 90

Sleep Hours: 7.00

Steps Count: 8000

Water Intake (Liters): 2.50

Daily Active Minutes: 60

Figure 5 - Front end integration

Overall, the integration of the machine learning model with a Streamlit-based front end transforms the calorie prediction system from a purely analytical model into a functional and interactive application. This deployment approach highlights the end-to-end nature of the project, demonstrating not only model development and evaluation but also real-world usability. In the Sri Lankan context, such lightweight web applications can play an important role in increasing accessibility to data-driven fitness tools and promoting health awareness among a broader audience.

Results and Discussion

The performance of the proposed calorie prediction model was evaluated using both quantitative metrics and visual diagnostic plots in order to comprehensively assess its predictive accuracy, reliability, and generalization capability. Since the objective of the study is to estimate calorie expenditure as a continuous variable, regression-based evaluation techniques were employed. The results demonstrate that the XGBoost regression model performs effectively in predicting calorie burn within the defined Sri Lankan fitness context.

One of the primary evaluation techniques used in this study is the Actual vs Predicted Calories Burn plot. This visualization provides an intuitive assessment of how closely the model's predictions align with the true observed values. In the generated plot, each data point corresponds to an individual test instance, with actual calorie values plotted along the horizontal axis and predicted calorie values plotted along the vertical axis. The diagonal reference line represents an ideal prediction scenario where predicted values perfectly match actual values.

The clustering of data points around the diagonal line indicates a strong agreement between actual and predicted calorie values. This suggests that the model has successfully learned the underlying relationships between input features such as age, height, weight, heart rate, body temperature, and workout duration, and the target variable representing calories burned. While minor deviations from the diagonal line are visible, these deviations are expected due to physiological variability among individuals and the inherent noise present in fitness-related data. Importantly, the absence of extreme outliers or large systematic deviations implies that the model does not exhibit severe prediction instability. Furthermore, the balanced dispersion of points above and below the diagonal indicates that the model does not consistently overestimate or underestimate calorie expenditure, supporting its robustness and fairness in prediction.

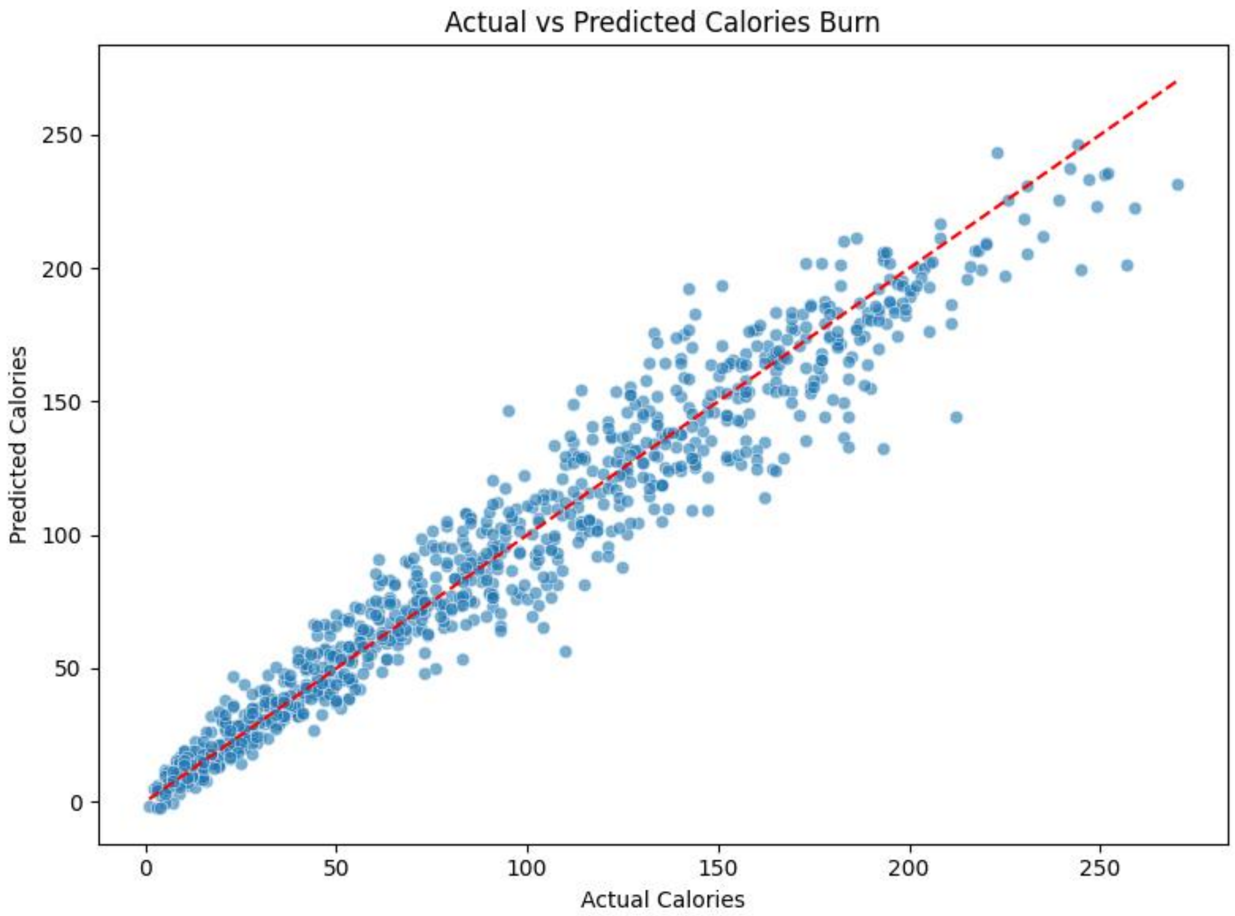


Figure 6 Evaluation Using Actual vs Predicted Values

To further analyze prediction errors, a residual error distribution plot was examined. Residuals, defined as the difference between actual and predicted calorie values, provide valuable insights into model bias and error behavior. Ideally, a well-performing regression model should produce residuals that are centered around zero and approximately normally distributed, indicating that errors are random rather than systematic.

The residual distribution obtained in this study shows a clear concentration of residual values around zero, confirming that the majority of predictions are close to the actual calorie values. The symmetric shape of the distribution on both the positive and negative sides of the zero reference line indicates that the model does not favor over-prediction or under-prediction. Additionally, the gradual decline in frequency as the magnitude of residuals increases suggests that large prediction errors occur infrequently. This pattern is a desirable characteristic of a reliable regression model and demonstrates that the XGBoost regressor has achieved a balanced and unbiased learning outcome.

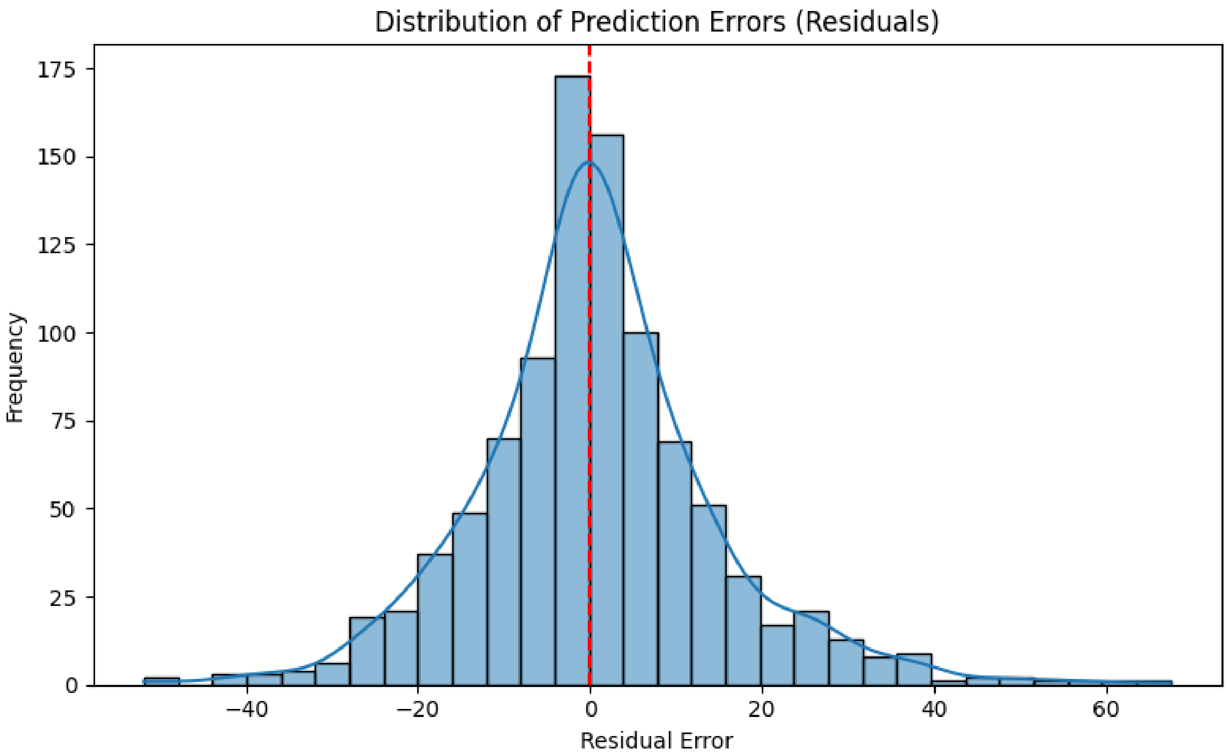
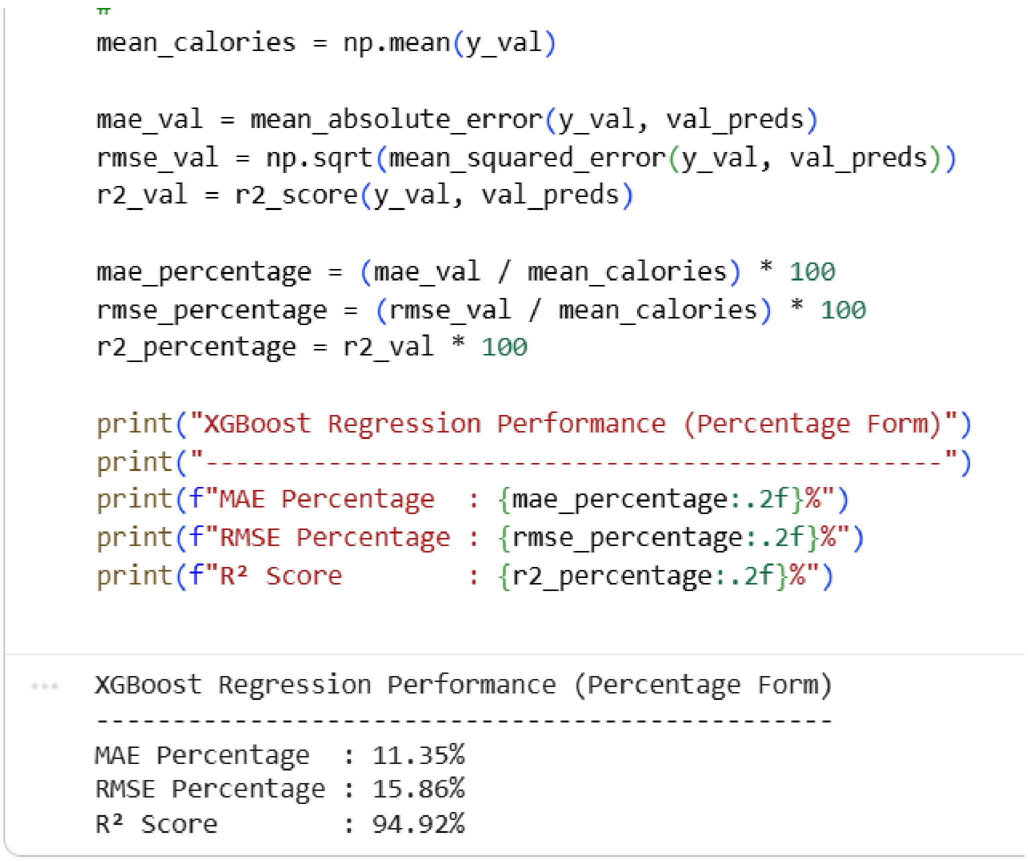


Figure 7 Residual Error Distribution Analysis

In addition to visual evaluation, quantitative performance metrics were computed and expressed in percentage form to provide an interpretable measure of model accuracy. The Mean Absolute Error (MAE) percentage of 11.35% indicates that, on average, the model's predictions deviate from the true calorie values by approximately eleven percent. Given the natural variability in human physiological responses to exercise, this level of error is considered low and acceptable for a fitness-oriented predictive application. The Root Mean Squared Error (RMSE) percentage of 15.86% reflects the model's sensitivity to larger errors and further confirms that significant prediction deviations are relatively uncommon.

The R^2 score of 94.92% demonstrates that the model explains nearly ninety-five percent of the variance in calorie burn values within the dataset. This high coefficient of determination indicates an excellent fit between the model and the data, suggesting that the selected features collectively provide strong explanatory power for predicting calorie expenditure. Such a high R^2 value is particularly notable given that the dataset represents realistic, physiologically driven variability rather than purely deterministic relationships.



```
mean_calories = np.mean(y_val)

mae_val = mean_absolute_error(y_val, val_preds)
rmse_val = np.sqrt(mean_squared_error(y_val, val_preds))
r2_val = r2_score(y_val, val_preds)

mae_percentage = (mae_val / mean_calories) * 100
rmse_percentage = (rmse_val / mean_calories) * 100
r2_percentage = r2_val * 100

print("XGBoost Regression Performance (Percentage Form)")
print("------------------------------------------------")
print(f"MAE Percentage  : {mae_percentage:.2f}%")
print(f"RMSE Percentage : {rmse_percentage:.2f}%")
print(f"R² Score        : {r2_percentage:.2f}%")
```

```
XGBoost Regression Performance (Percentage Form)
------------------------------------------------
MAE Percentage  : 11.35%
RMSE Percentage : 15.86%
R² Score        : 94.92%
```

Figure 8 - Error values

When considered together, the evaluation plots and quantitative metrics provide consistent evidence of strong model performance. The alignment observed in the Actual vs Predicted plot, the unbiased residual distribution, and the high R² score collectively indicate that the XGBoost regression model generalizes well to unseen data and captures meaningful patterns rather than noise. These results validate the suitability of XGBoost for the calorie prediction task and support its deployment within the Streamlit-based front-end application.

In the context of Sri Lankan fitness and health applications, these results suggest that the developed model can serve as a reliable decision-support tool for estimating calorie expenditure during physical activity. While predictions should not be interpreted as medical advice, the model provides sufficiently accurate and consistent estimates to support fitness tracking, awareness, and lifestyle management. Overall, the evaluation confirms that the proposed system achieves its intended objectives with strong predictive performance and practical relevance.

Limitations

Despite the strong predictive performance achieved by the calorie prediction model, several limitations must be acknowledged to ensure responsible interpretation and use of the system. One of the primary limitations relates to data quality. The dataset used for training and evaluation was synthetically generated to reflect realistic Sri Lankan fitness and physiological patterns due to the lack of publicly available local datasets. While the synthetic data was carefully designed to capture plausible relationships between input features and calorie expenditure, it may not fully represent the diversity and complexity of real-world populations. Certain lifestyle factors, regional variations, and uncommon physiological conditions may be underrepresented, which could affect the model's generalization when applied to real users.

Another important limitation concerns the risk of bias and unfairness. Physiological responses to physical activity vary significantly based on individual health conditions, fitness levels, genetics, and metabolic differences. The model does not explicitly account for underlying medical conditions such as cardiovascular disease, diabetes, or respiratory disorders, which can significantly influence calorie burn. As a result, predictions may be less accurate for individuals outside the typical physiological range represented in the dataset. Additionally, demographic features such as age and gender are used as inputs, which, while physiologically relevant, may introduce subtle biases if not carefully interpreted. This highlights the importance of viewing model outputs as estimates rather than absolute measures.

The potential real-world impact of the system also presents limitations. Although the application provides useful calorie burn estimates for fitness awareness and lifestyle management, it is not designed to function as a medical or diagnostic tool. There is a risk that users may misinterpret predictions as precise or medically authoritative values, leading to inappropriate exercise decisions. Without proper guidance or disclaimers, such misinterpretation could result in overexertion or unrealistic fitness expectations. Therefore, responsible deployment requires clear communication of the system's intended purpose and limitations.

Ethical considerations further constrain the scope of the proposed system. If extended to real-world deployment with real user data, issues related to data privacy, consent, and security would need to be carefully addressed. Health-related data is sensitive, and improper handling could lead to privacy violations or misuse. Moreover, transparency in model behavior is essential to maintain user trust. While explainability techniques such as SHAP help mitigate this concern, continuous monitoring and validation would be required to ensure ethical and fair use.

In summary, the limitations of the proposed calorie prediction system stem from data representativeness, potential bias, real-world applicability, and ethical considerations. Recognizing these limitations is essential for responsible interpretation of the results and provides a foundation for future improvements, such as incorporating real-world datasets, personalized health indicators, and enhanced ethical safeguards.

Conclusion

This project successfully demonstrated the design, development, evaluation, and deployment of a machine learning based calorie prediction system tailored to a Sri Lankan context. The primary objective of the study was to estimate calorie expenditure during physical activity using demographic and physiological features, thereby supporting fitness awareness and informed lifestyle decisions. By framing the problem as a supervised regression task and employing a data-driven approach, the project addressed a real-world health-related challenge using modern machine learning techniques.

A synthetic dataset was carefully constructed to reflect realistic Sri Lankan fitness patterns in the absence of publicly available local datasets. Through systematic data preprocessing, feature scaling, and validation, the dataset was transformed into a reliable foundation for model training. The XGBoost regressor was selected as the core predictive algorithm due to its ability to capture complex non-linear relationships, robustness against overfitting, and strong generalization performance. Quantitative evaluation results, including a high R^2 score and low percentage-based error metrics, confirmed that the model accurately estimates calorie burn and performs consistently on unseen data.

To enhance transparency and trust, SHAP was integrated as an explainability framework. The SHAP analysis demonstrated that the model learned meaningful and physiologically plausible relationships, with workout duration, heart rate, and body temperature emerging as the most influential factors. This interpretability is particularly important in health and fitness applications, as it ensures that model predictions can be understood and justified rather than treated as black-box outputs.

The integration of the trained machine learning model into a Streamlit-based web application transformed the system from a purely analytical model into an interactive and user-friendly tool. The front-end interface allows users to input personal and workout-related details and receive real-time calorie predictions, thereby bridging the gap between machine learning theory and practical deployment. This end-to-end implementation highlights the feasibility of deploying data-driven fitness applications using lightweight web technologies.

Despite its effectiveness, the project also acknowledges limitations related to synthetic data usage, potential bias, and ethical considerations. These limitations emphasize the need for cautious interpretation of predictions and responsible deployment. Future work may involve incorporating real-world datasets, additional health indicators, and personalized recommendations to further enhance accuracy and applicability.

In conclusion, this project demonstrates the practical potential of machine learning in supporting health and fitness awareness within the Sri Lankan context. By combining robust predictive modeling, explainability, and accessible front-end deployment, the system provides a strong foundation for future enhancements and real-world adoption, while adhering to responsible and ethical AI principles.

References

[1] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, pp. 785–794, 2016.

[2] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[3] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, San Francisco, USA, 2012.

[5] J. A. Levine, “Measurement of Energy Expenditure,” *Public Health Nutrition*, vol. 8, no. 7A, pp. 1123–1132, 2005.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] A. Treuille, “Streamlit: Turning Data Scripts into Shareable Web Apps,” *Streamlit Documentation*, 2020.