Data Mining and Machine Learning

Project Proposal

Team Members :

BDSF22M023 Iqra Ishaq

BDSF22M040 Aisha Munir

Project Title: Predicting Calorie Expenditure from Workout Data with Streamlit Interface

Problem Statement:

Accurate estimation of calorie expenditure during physical activity is crucial for individuals managing their weight, optimizing fitness routines, and monitoring overall health. While various wearable devices and apps provide calorie estimates, these often rely on generalized formulas and may not be accurate for diverse populations or specific workout conditions. This project aims to develop a machine learning model that can more accurately predict calorie expenditure based on a dataset of workout-related features and provide an interactive interface for users to input their workout data and receive personalized calorie estimates. This is important for providing users with more personalized and reliable information to support their health and fitness goals. The problem is challenging due to the complex relationship between various activity metrics and individual physiological factors that influence calorie burn.

Objectives:

1. Develop a predictive model using machine learning techniques to accurately estimate calorie expenditure based on the provided workout dataset.
2. Evaluate the performance of different regression models (e.g., Linear Regression, Random Forest Regressor, Gradient Boosting Regressor) in predicting calorie expenditure.
3. Design and implement a user-friendly web interface using Streamlit to allow users to input workout parameters and visualize predicted calorie expenditure.
4. Identify the most significant features in the dataset that contribute to calorie expenditure and analyze their relationships.

Proposed Methodology:

- **Machine Learning Models/Techniques:**
  - We will explore several regression models, including:
    - **Linear Regression:** A fundamental algorithm for understanding linear relationships between features and the target variable.
    - **Random Forest Regressor:** An ensemble method that combines multiple decision trees to improve prediction accuracy and handle non-linear relationships.

        - **Gradient Boosting Regressor (e.g., XGBoost, LightGBM):** Powerful ensemble techniques that sequentially build trees, correcting errors from previous trees, often achieving state-of-the-art performance in regression tasks.
    - We will train and evaluate these models using the training dataset. The best-performing model will be selected for integration into the Streamlit application.
- **Frontend Development:**
    - We will use Streamlit to create an interactive web application. Streamlit allows for rapid development of data-driven web applications using Python.
    - The application will include input widgets for users to enter workout parameters (features from the dataset).
    - The application will display the predicted calorie expenditure based on the user's input, using the trained machine learning model.
    - We will also incorporate data visualization elements (e.g., charts) within the Streamlit app to illustrate the relationship between input features and predicted calories, and to show model performance.
- **Feature Engineering:** We will analyze the existing features to identify potential transformations or combinations that may improve model performance. This may include scaling, normalization, or creating interaction terms.
- **Evaluation Metrics:** The performance of the models will be evaluated using the Root Mean Squared Logarithmic Error (RMSLE) during training and testing. We will also evaluate the user experience and usability of the Streamlit application.

Dataset Description:

The dataset is obtained from the Kaggle Playground Series - Season 5, Episode 5 competition. It contains synthetic data generated from a deep learning model trained on a calorie expenditure dataset. The dataset includes various features related to workout activities (the specific features are not detailed in the prompt, but will be analyzed in the project). The goal is to predict the Calories variable, representing the amount of calories burned. The training data has approximately 750,000 rows, and the test data has approximately 250,000 rows. The data is in CSV format. We will perform exploratory data analysis (EDA) to understand the distribution of features, identify any missing values, and check for outliers.

Expected Outcomes:

This project aims to deliver a robust and user-friendly application for predicting calorie expenditure. We expect to:

- Achieve a competitive RMSLE score on the test dataset.
- Develop a functional and intuitive Streamlit web application for users to interact with the calorie expenditure prediction model.
- Provide insights into the key factors that influence calorie expenditure during workouts.
- Develop a reproducible and well-documented machine learning pipeline and web application.
- Potentially identify limitations of the synthetic dataset and suggest areas for improvement in future data generation.

Brief Timeline of Activities:

- **Day 1-2:** Data Exploration and Preprocessing (EDA, data cleaning, feature engineering)
- **Day 3-4:** Model Development and Training (model selection, hyperparameter tuning)
- **Day 5-6:** Streamlit Application Development and Integration (UI design, model integration, testing)
- **Day 7:** Model Evaluation, Reporting, and Deployment (Performance analysis, results visualization, report writing)