

# Project Report

Machine Learning Project

## Body Fat Percentage Predictor

UML501

Group 3CS3



THAPAR INSTITUTE  
OF ENGINEERING & TECHNOLOGY  
(Deemed to be University)

**Submitted By**

Nimish Duggal

102217073

**Submitted To**

Dr Sumit Kumar

## Introduction

Body fat percentage (BFP) is a critical metric for assessing overall health, with implications for obesity, cardiovascular disease, and mortality risk. While measures like Body Mass Index (BMI), waist circumference, and waist-to-hip ratios are widely used, they do not fully account for body composition differences due to age, gender, and other factors. Advanced techniques such as dual-energy X-ray absorptiometry (DXA) and skinfold measurements provide more precise insights but require specialized equipment and expertise.

This project aims to develop an accessible and accurate machine learning model to predict BFP using easily obtainable fitness metrics, such as BMI, workout patterns, heart rate, other anthropometric features and workout metrics. By integrating these diverse predictors, the model seeks to improve upon traditional methods, offering a robust, scalable solution for health monitoring and obesity risk assessment in a general population.

It aims to provide insights for fitness professionals and enthusiasts to monitor health trends. The project uses machine learning techniques, incorporating real-world fitness data, to address the increasing demand for precise, data-driven health assessments.

## Methodology

### About the dataset

This dataset has been taken from Kaggle, given at <https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset/data>.

This dataset comprises detailed fitness and health data for 973 gym members, capturing their physical attributes, exercise routines, and demographic information. It is designed to provide a comprehensive understanding of fitness patterns and health trends.

### Key Features:

- Age: Age of the gym member.
- Gender: Gender of the gym member (Male or Female).
- Weight (kg): Member's weight in kilograms.
- Height (m): Member's height in meters.
- Max\_BPM: Maximum heart rate (beats per minute) during workout sessions.
- Avg\_BPM: Average heart rate during workout sessions.
- Resting\_BPM: Heart rate at rest before workout.
- Session\_Duration (hours): Duration of each workout session in hours.
- Calories\_Burned: Total calories burned during each session.
- Workout\_Type: Type of workout performed (e.g., Cardio, Strength, Yoga, HIIT).
- Water\_Intake (liters): Daily water intake during workouts.
- Workout\_Frequency (days/week): Number of workout sessions per week.
- Experience\_Level: Level of experience, from beginner (1) to expert (3).

- BMI: Body Mass Index, calculated from height and weight.

This dataset is divided into test and training sets with test size being 20% of the entire dataset.

## Data Pre-Processing

### Data Cleaning

- **Missing Values** – The dataset was already clean in terms of missing values. There were no missing values found in the dataset
- **Handling Outliers** – To detect and remove outliers, we used Z-score method, with a threshold of 3.5, to ensure robust model performance.
  - **Z-Score Method**  
The Z-score method is a statistical technique used for standardizing data by expressing the distance of a value from the mean in terms of standard deviations. It is widely utilized for detecting and handling outliers in datasets. The method assigns a Z-score to each data point, calculated as:  

$$Z = (X - \mu) / \sigma,$$
 where  $\sigma$  represents standard deviation and  $\mu$  represents mean of the data. Values with absolute Z-scores beyond a certain threshold (commonly 3 or 3.5) are considered outliers. This approach assumes data follows a normal distribution, ensuring outlier detection is based on statistical significance. It is effective in preprocessing tasks for machine learning to improve model robustness and reduce noise.

### Data Transformation

- Categorical Values into Numerical Features
  - **Gender Feature** – This Feature is actually a binary feature with two inputs, Male and Female. Here we can represent Male with 1 and Female with 0 or vice versa.
  - **Workout\_Type Feature** – We will be using **One Hot Encoding** to convert this categorical feature into numerical feature. This feature has unique values – Strength, Cardio, Yoga and HIIT. Each category is represented as a unique binary vector, where one element is set to 1 (indicating presence) and all others are set to 0.
- Feature Scaling
  - All features were scaled to a range of 0 to 1 using **Min-Max Scaling** to ensure uniform feature influence in the model. Min-Max Scaling uses the formula:  

$$X' = (X - X_{min}) / (X_{max} - X_{min}),$$
 where X is the original value, Xmin and Xmax are the minimum and maximum values of the feature. This method preserves the relationships between values and is particularly useful when features have varying ranges, ensuring all inputs contribute proportionally to machine learning models.

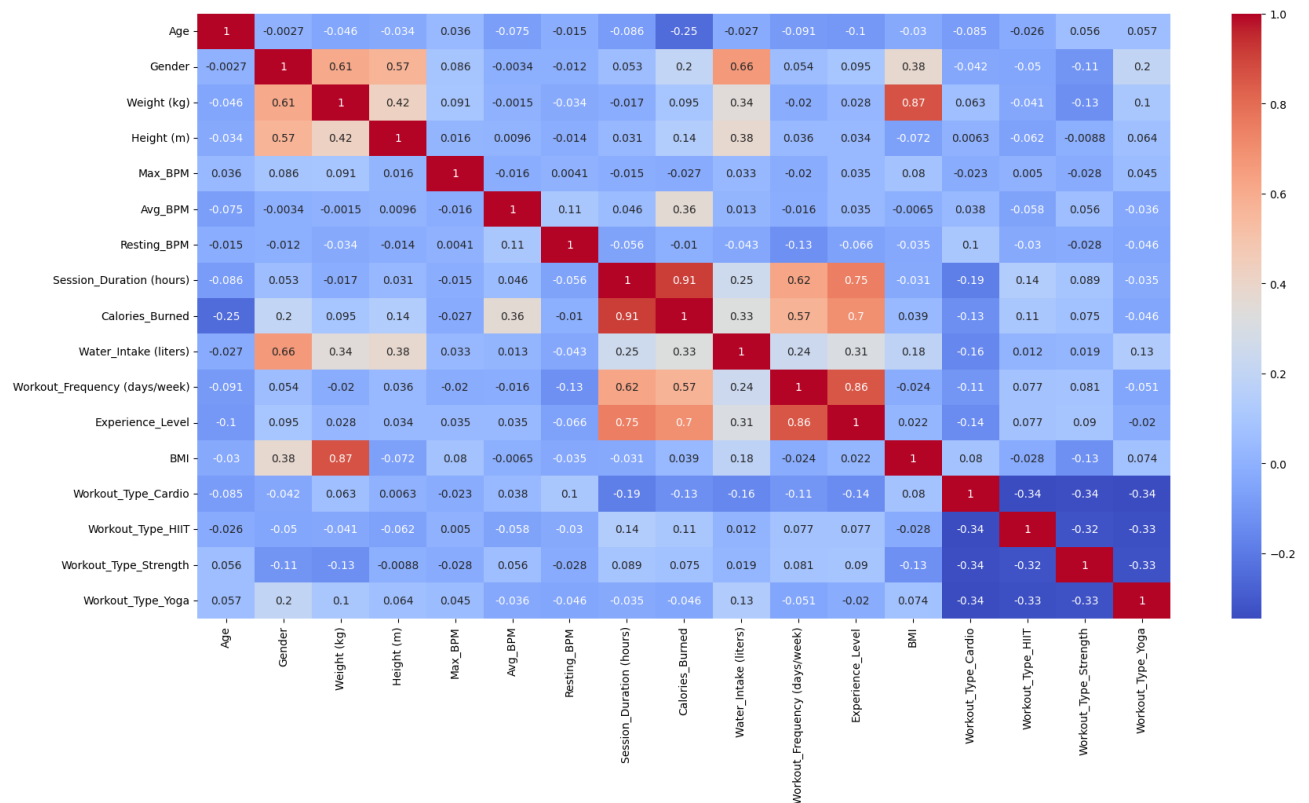
### Feature Selection

#### Correlation Analysis

Correlation analysis examines the relationship between variables to identify multicollinearity, where features are highly interrelated. A heatmap was used to visualize Pearson correlation coefficients:

- Values close to +1 or -1 indicated strong relationships.
- Features with high inter-correlation, such as *Session Duration* and *Workout Frequency* (correlation > 0.8), were removed to enhance model interpretability and avoid redundancy.

This process ensured the final model used only meaningful, independent features, improving predictive performance.



## Model Training

Model training is a critical phase in the machine learning pipeline, where selected algorithms are used to learn patterns from the training dataset. In this project, three regression models—K-Nearest Neighbors (KNN), Linear Regression, and Random Forest Regressor—were trained and evaluated to predict body fat percentage using gym members' physiological and fitness data.

### Model Selection:

- Three diverse regression models were chosen to ensure a comprehensive comparison of performance.
  - **K-Nearest Neighbors (KNN):** A simple, non-parametric algorithm that predicts based on the average of the k-nearest neighbors.
  - **Linear Regression:** A parametric model assuming a linear relationship between predictors and the target variable.
  - **Random Forest Regressor:** An ensemble learning method using multiple decision trees for robust predictions.

### Training Process:

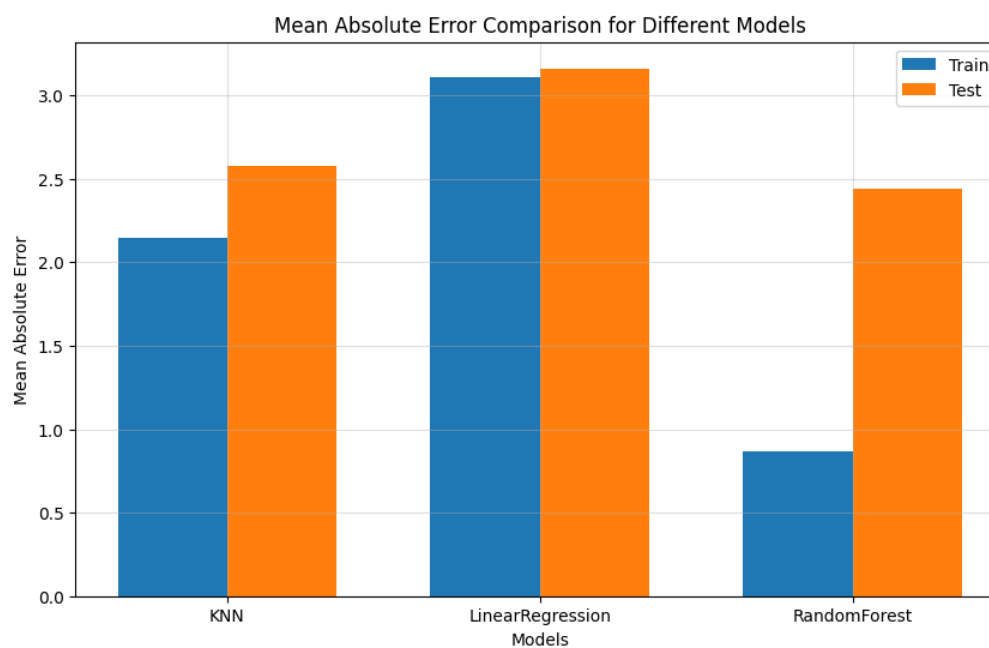
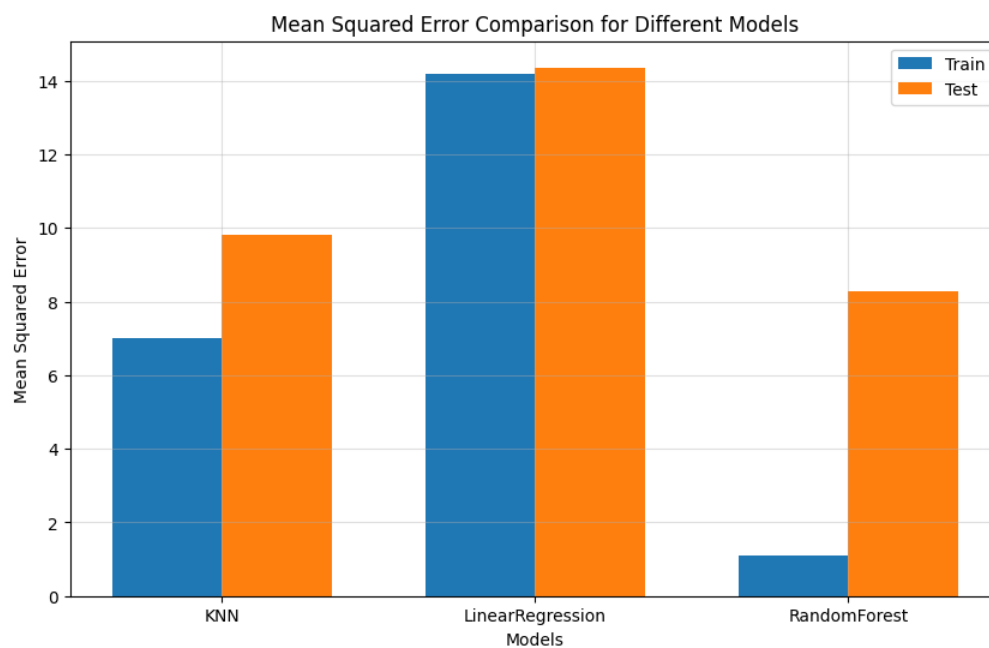
- Each model was trained using the fit method, where it learned patterns from the scaled training data.

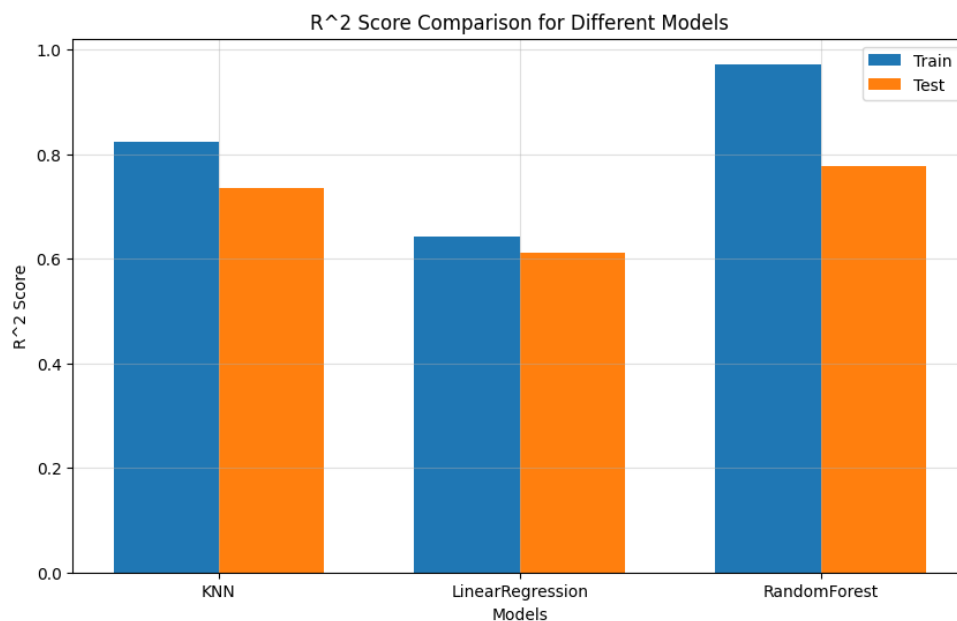
## Model Evaluation

Models were evaluated on both training and test data to ensure generalization.

The model was evaluated on the test set using metrics such as:

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.
- **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors.
- **R<sup>2</sup> Score:** Represents the proportion of variance in the target variable explained by the model.





#### R2 Score for different Models:

- **KNN** - 0.734580444591002
- **Linear Regression** - 0.6117645100754843
- **Random Forest Regressor** - 0.776557273877883

### Results Analysis and Interpretation

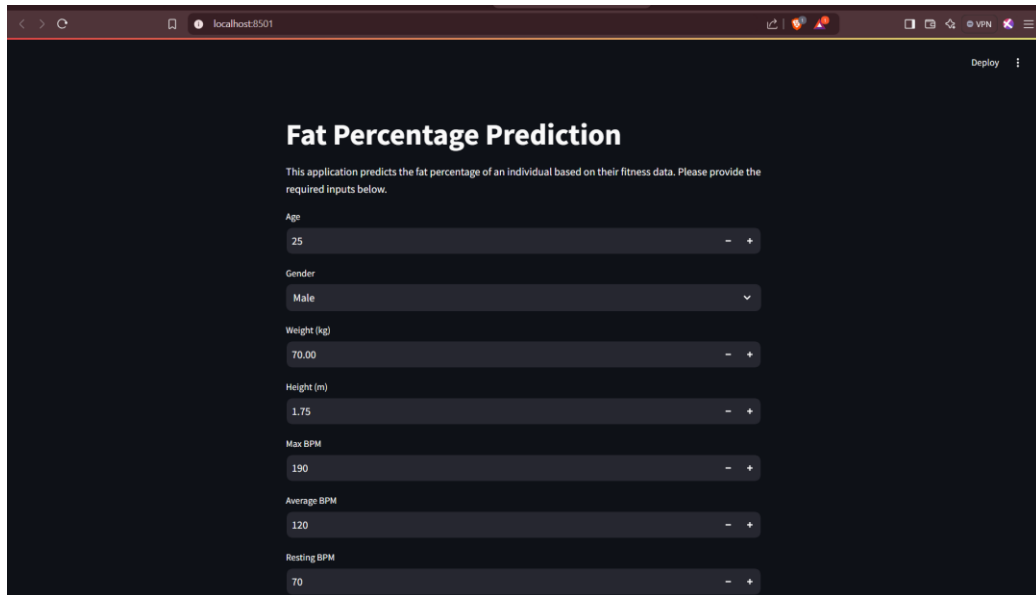
The model with the highest R<sup>2</sup> score on the test set was identified as the best-performing model. Random Forest achieved the highest test performance, suggesting its ability to capture complex patterns in the data. But while analysing the evaluation metrics, it was realised that Random Forest Regressor was not performing as well, as it was on testing data than it was performing on training data. This hinted that random forest regressor was overfitting on the data and was failing to generalize well on the data to some extent.

This is why the model, K-Nearest neighbours was used, which had the second best R<sup>2</sup> score and whose training and testing data metrics similar, which suggested that it was a good fit.

On the contrary, Linear Regression failed to recognize the underlying complex patterns in the data. This is because Linear Regression is comparatively a simpler model than the other models and it recognizes the linear patterns much well than the complex patterns.

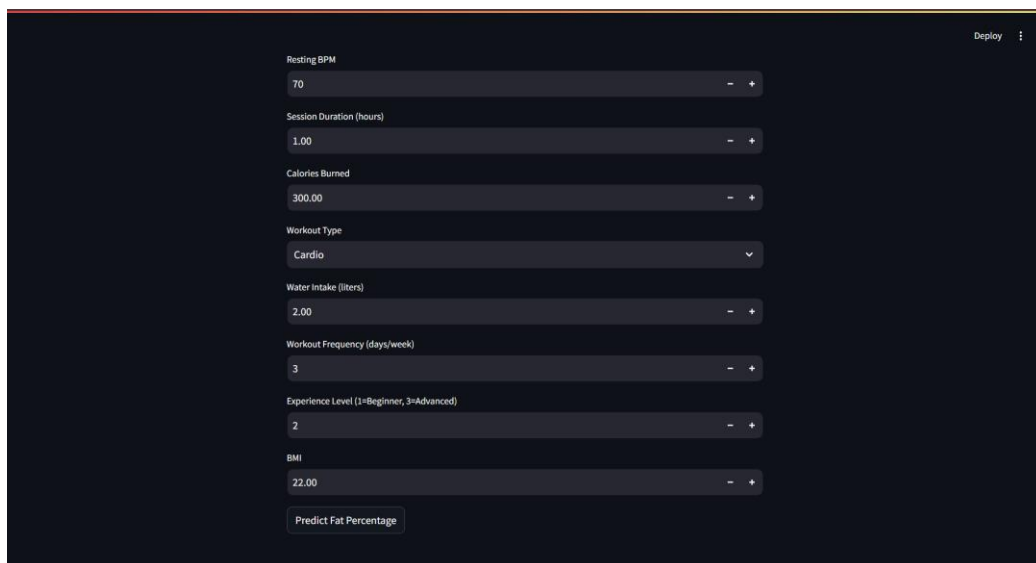
## Testing Interface

In addition to the predictive models, an intuitive user interface was developed using **Streamlit** to provide real-time predictions. This interface allows users to input key data such as age, gender, weight, height, workout type, and other fitness parameters. Upon submission, the app processes the input data, applies the trained machine learning model, and displays the predicted body fat percentage, offering a seamless experience for non-technical users. This easy-to-use interface ensures accessibility and usability of the model in real-world fitness applications.



The screenshot shows the top section of the 'Fat Percentage Prediction' app. The title 'Fat Percentage Prediction' is centered at the top. Below it, a subtitle reads: 'This application predicts the fat percentage of an individual based on their fitness data. Please provide the required inputs below.' The input fields are as follows:

- Age: 25
- Gender: Male
- Weight (kg): 70.00
- Height (m): 1.75
- Max BPM: 190
- Average BPM: 120
- Resting BPM: 70

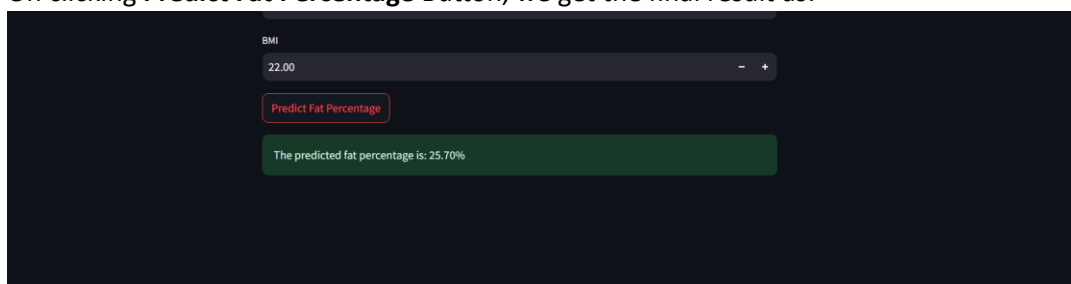


The screenshot shows the bottom section of the 'Fat Percentage Prediction' app. The input fields are as follows:

- Resting BPM: 70
- Session Duration (hours): 1.00
- Calories Burned: 300.00
- Workout Type: Cardio
- Water Intake (liters): 2.00
- Workout Frequency (days/week): 3
- Experience Level (1=Beginner, 3=Advanced): 2
- BMI: 22.00

At the bottom, there is a button labeled 'Predict Fat Percentage'.

On clicking **Predict Fat Percentage** Button, we get the final result as:



The screenshot shows the final result of the prediction. The BMI input field is still visible with the value 22.00. Below it, the 'Predict Fat Percentage' button is highlighted with a red border. The result is displayed in a green box: 'The predicted fat percentage is: 25.70%'.

## Conclusion

In this project, various machine learning models were applied to predict body fat percentage based on gym members' demographic and fitness data. By preprocessing the data, handling outliers, encoding categorical features, and applying appropriate scaling techniques, the dataset was transformed to be compatible with different regression models.

The Random Forest model, despite showing the highest  $R^2$  score, exhibited overfitting and struggled to generalize well on the test data. In contrast, K-Nearest Neighbors (KNN) performed more consistently across both training and testing data, making it the optimal choice for prediction. The Linear Regression model, being a simpler algorithm, failed to capture the complex relationships in the data, further reinforcing the importance of using more advanced models for accurate predictions in fitness and health applications.

The project successfully highlighted the potential of machine learning in fitness and health prediction, with future opportunities for improvement through hyperparameter tuning and more extensive datasets.