University of Hertfordshire UH

Body Fat Prediction

7COM1075 - Data Science and Analytics Masters Project

22090668
Abdul Jaleel Mohammed
Maria Psarrou

University of Hertfordshire UH

"I, Abdul Jaleel Mohammed, confirm that this report has been critically proof-read and quality checked to ensure it is free from grammar, spelling, and formatting errors. I have taken all necessary steps to ensure that the report meets high standards of clarity, coherence, and presentation."

MSc FPR Declaration

This report is submitted in partial fulfilment of the requirement for the degree of:
Master of Science in Data Science and Analytics with Advanced Research, at the University of Hertfordshire (UH).

I hereby declare that the work presented in this project and report is entirely my own, except where explicitly stated otherwise. All sources of information and ideas, whether quoted directly or paraphrased, have been properly referenced in accordance with academic standards. I understand that any failure to properly acknowledge the work of others could constitute plagiarism and may result in academic penalties.

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged.

Table of Contents

1. Abstract

Obesity is a global health issue that affects over 1.9 billion adults globally. It leads to chronic diseases such as heart disease, diabetes, and cancer (World Health Organization, 2021). Understanding body composition plays a key role in addressing obesity. Obesity has been measured solely using the Body Mass Index (BMI) in the past, but it has limitations since it does not differentiate between the ratio of fat to muscle in a body. Therefore, using Body Fat Percentage (BFP) provides a more accurate picture of body composition.

Traditional methods for measuring BFP, like underwater weighing (densitometry) and dual-energy X-ray absorptiometry (DEXA), are expensive, time-consuming, and not easily accessible, particularly in remote areas. This study uses machine learning (ML) to predict body fat percentage using simple, non-invasive measurements such as weight and circumference of the abdomen, chest, density, and hips.

The dataset includes 15 anthropometric features and body density values obtained through underwater weighing, which is considered the gold standard for body composition measurement (Accurso et al. 2024). To achieve accurate predictions, multiple supervised ML models were tested, including Multilayer Perceptron (MLP), Decision Tree Regression (DT), and Random Forest Regression. A hybrid model was also developed to combine the strengths of the individual models, improving overall performance.

Before training the models, the data was preprocessed using techniques such as Spearman's correlation to identify important features, mutual information gain to rank feature importance, and variance inflation factor (VIF) to check for multicollinearity. Features were also normalized using Standard Scaler. The models were evaluated using standard regression metrics.

The results showed that the hybrid model outperformed other individual models, achieving an RMSE of 0.62 and an R^2 score of 0.99. This demonstrates its high accuracy and reliability in predicting body fat percentage.

Several studies utilize relevant ML models to predict body fat percentage. For instance, Uçar et al. (2021) implement a Decision Tree Regressor, Multilayer Feedforward Neural Network, and Support Vector Machine. Many studies in this area employ PCA and Spearman's rank correlation as the feature selection algorithm. Mutual information gain and an Extra Tree algorithm were applied for the feature selection algorithm.

This study highlights the potential of machine learning in health informatics by offering a cost-effective, accessible, and scalable way to measure body composition. This research can be applied in clinical practice, remote patient monitoring, fitness assessments, and public health initiatives focused on obesity prevention.

2. Introduction to the Project

2.1 Problem Overview

Obesity is a global health challenge that is growing at an alarming rate, impacting individuals of all ages. According to the World Health Organization (2021), around 1.9 billion adults are classified as overweight, and more than 650 million are considered obese. It leads to chronic diseases such as heart disease, diabetes, and cancer (Fan et al., 2022). With increasing obesity rates, there is an urgent need for accurate tools to measure body-fat and evaluate obesity-related risks effectively.

2.1.1 Limitations of Traditional Measures

Traditionally, Body Mass Index (BMI) was widely used to measure obesity. It was considered simple and easy to implement in clinical and public health settings. However, BMI only provides a rough estimate of body composition and it does not differentiate between the ratio of fat to muscle in a body. Therefore, using Body Fat Percentage (BFP) provides a more accurate picture of body composition. For example, human bodies that have high muscle mass (like athletes) can be misclassified as obese, while those with low muscle mass may have their adiposity underestimated (Choi et al., 2020). This limits BMI's effectiveness for accurately diagnosing obesity, especially in diverse populations.

Body Fat Percentage (BFP) is a more reliable and meaningful metric for evaluating obesity because it measures the proportion of fat mass relative to overall body composition. BFP offers deeper insights into health risks, as excessive body fat is closely associated with chronic diseases and metabolic disorders (Hussain et al., 2021).

2.1.2 Challenges with Advanced BFP Measurement Methods

Accurate BFP measurement methods, such as underwater weighing, DEXA scans, and bioelectrical impedance analysis (BIA), have significant drawbacks:

- **Cost**: The equipment is expensive.
- **Time**: These techniques require time and expertise.
- **Limited Access**: They are often only available in specialized facilities, making them inaccessible in remote areas.

To address these challenges, there's a need for a non-invasive, affordable, and scalable method that can predict BFP using easy-to-collect data like weight and body circumferences.

2.1.3 Machine Learning as a Solution

Machine learning (ML) offers a practical way to predict body fat percentage. Unlike traditional statistical models, ML can handle non-linear relationships and interactions between features. Models like Random Forest, Decision Trees, and Neural Networks are especially effective in finding patterns in complex data (Breiman, 2001).

2.1.4 Gaps in Existing Methods

While machine learning has shown significant potential in predicting body composition, several challenges remain in current research:

- **Lack of comparative analysis**: Few studies have systematically compared the performance of different ML models for BFP prediction. A thorough comparison of models such as Decision Tree Regressors, Random Forests, and Multilayer Perceptrons (MLP) is needed to identify the most effective approaches.
- **Limited exploration of hybrid models**: Combining predictions from multiple ML models through hybrid or ensemble methods can improve accuracy and robustness. However, such approaches remain underexplored in body composition research.
- **Validation gaps**: Many studies do not rigorously validate their models using techniques like cross-validation or external datasets. This raises concerns about how well the models generalize to new data.
- **Real-world deployment**: Most ML models are limited to academic research and have not been tested or implemented in clinical or real-world health settings. This limits their practical usability and impact.

Addressing these gaps is essential to fully realize the potential of machine learning for body fat prediction and make these tools accessible for routine use in healthcare and beyond.

2.2 Objectives and Structure of the Study

To overcome these limitations, this project uses machine learning algorithms to predict BFP based on anthropometric measurements and body density values obtained from underwater weighing. The study focuses on three main objectives:

- **Data Preprocessing**: Improve the quality of input data through cleaning, feature scaling, and feature selection techniques, including Spearman's correlation analysis, mutual information gain, and variance inflation factor (VIF). This ensures that only the most relevant predictors are used for training the models.
- **Model Design**: Implement and compare the performance of three non-linear ML models—Decision Tree Regressor (DT), Random Forest Regressor (RF), and Multilayer Perceptron (MLP). A hybrid model is also introduced to combine predictions from individual models, enhancing accuracy and robustness.
- **Performance Evaluation**: Assess model performance using standard regression metrics:

- **Root Mean Square Error (RMSE)**: Measures the average magnitude of prediction errors.
- **Mean Absolute Error (MAE)**: Represents the average size of prediction errors.
- **R-squared (R^2)**: Indicates how well the model explains the variance in the data.
- **Mean Squared Error (MSE)**: Provides a squared measure of prediction errors.

The models are rigorously validated using k-fold cross-validation to ensure reliability and generalizability.This structured approach provides a thorough evaluation of machine learning techniques for BFP prediction while addressing existing gaps in the literature. The findings contribute to the development of robust, scalable, and accessible tools for body composition assessment. These tools can be integrated into clinical workflows, telehealth services, and public health programs to improve obesity monitoring, enable early interventions, and enhance health outcomes globally.

2.2.1 Aims and Objectives

The primary objectives of this study are:

- To preprocess and analyze anthropometric data to identify key predictors of body fat percentage.
- To implement and compare the performance of three machine learning models: DecisionTreeRegressor, RandomForestRegressor, and MLPRegressor, for BFP prediction.
- To develop a hybrid model that combines predictions from individual models to improve accuracy.
- To evaluate and validate the models using standard performance metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Squared Error (MSE), along with statistical validation techniques.
- To assess the implications of machine learning techniques for non-invasive body fat estimation and their potential role in health informatics.

2.2.2 Research Question

The research seeks to answer the following key questions:

1. Can machine learning models predict body fat percentage accurately using anthropometric measurements?
2. Does a hybrid model combining individual ML predictions achieve higher accuracy than standalone models?

This study provides a novel contribution by:

- Combining multiple non-linear ML models to develop a hybrid approach.
- Addressing gaps in existing research regarding ensemble predictions for BFP estimation.

- Providing a robust, scalable solution for clinical and public health applications.

---

3. Literature Review

Obesity is a global epidemic, impacting over 1.9 billion adults and contributing to chronic illnesses such as cardiovascular disease, diabetes, and certain cancers (World Health Organization, 2021). The growing prevalence of obesity has far-reaching consequences, including rising healthcare costs, reduced workforce productivity, and an increased burden on healthcare systems worldwide.

Accurate tools for assessing obesity are essential for enabling early intervention, monitoring risks, and developing personalized treatment plans. However, traditional methods such as Body Mass Index (BMI) have significant limitations. BMI relates weight to height but does not distinguish between fat mass and lean mass, leading to potential misclassification (Fan et al., 2022). This is particularly problematic for individuals with high muscle mass, older adults with reduced muscle mass, and those with unique body compositions.

Body Fat Percentage (BFP) has emerged as a more reliable measure of body composition. BFP directly quantifies fat mass as a proportion of overall body weight, offering a clearer indicator of obesity-related health risks. Despite its advantages, advanced BFP measurement techniques like underwater weighing, dual-energy X-ray absorptiometry (DEXA), and bioelectrical impedance analysis (BIA) are expensive, time-consuming, and often unavailable for large-scale use (Hussain et al., 2021).

3.1 Machine Learning Models for Body Fat Prediction

Machine learning (ML) techniques offer a robust solution for predicting BFP non-invasively using simple anthropometric measurements, such as weight, abdomen circumference, chest circumference, and hip circumference. Unlike traditional regression models, which often assume linear relationships between variables, ML algorithms excel in capturing **non-linear and complex interactions** between features and outcomes (Breiman, 2001). The following subsections discuss the key ML models explored in existing literature and their relevance to BFP prediction.

3.1.1 Non-Linear Machine Learning Models

Decision Trees and ensemble methods such as Random Forests have been extensively applied in regression problems involving health data. According to Breiman (1984), Decision Trees split data recursively into branches based on feature thresholds, making them intuitive and interpretable. However, Decision Trees are prone to overfitting, particularly when applied to small datasets, as they tend to model noise as patterns.

Random Forests, introduced by Breiman (2001), addresses this limitation by constructing multiple decision trees and averaging their predictions. This ensemble method reduces variance and

improves generalization while effectively handling multicollinearity and noisy data. In the context of BFP prediction, Random Forests have demonstrated significant performance improvements. For example, Hussain et al. (2021) utilized Random Forests alongside other ensemble models and reported an RMSE of 4.46, outperforming linear regression models.

Neural networks, particularly Multilayer Perceptrons (MLPs), have also shown promise in predicting body composition. MLPs are feedforward neural networks capable of modeling complex, non-linear relationships by learning weighted connections between input and output layers (Rumelhart et al., 1986). Fan et al. (2020) applied neural networks to predict BFP and highlighted that abdomen circumference and density were among the strongest predictors. Their findings demonstrated that MLPs outperform simpler models when combined with robust preprocessing techniques, such as feature scaling and dimensionality reduction.

Support Vector Machines (SVMs), introduced by Cortes and Vapnik (1995), have also been applied to body composition prediction. SVMs work by finding the optimal hyperplane that minimizes prediction error, particularly excelling in small to medium-sized datasets. While effective, SVMs often require extensive hyperparameter tuning and are computationally expensive for large datasets.

3.1.2 Hybrid and Ensemble Approaches

Hybrid models combine the strengths of multiple ML algorithms to improve prediction accuracy, robustness, and generalizability. Hybrid approaches are particularly effective when individual models exhibit complementary strengths. For example, Uçar et al. (2021) combined Decision Trees, Support Vector Machines (SVM), and neural networks to develop a hybrid ensemble model for BFP prediction. Their hybrid approach achieved superior results compared to standalone models, although the reported RMSE (12.15) remained higher than desired.

Thomas et al. (2020) explored advanced feature selection and preprocessing techniques to improve hybrid model performance. Techniques such as mutual information gain and dimensionality reduction were applied to identify the most significant predictors, thereby enhancing model efficiency. Their study emphasized that robust preprocessing, when combined with ensemble learning methods, leads to substantial accuracy improvements.

Gradient Boosting Machines (GBMs) and XGBoost are additional ensemble methods that have been explored for health-related regression problems. These algorithms combine weak learners iteratively, minimizing errors and improving performance. GBMs, introduced by Friedman (2001), are particularly effective for structured health data but require careful tuning to avoid overfitting.

3.2 Comparative Analysis of Existing Studies

Study	Method(s)	Key Features	Findings
Uçar et al. (2021)	SVM, Decision Tree, Hybrid Models	Weight, abdomen, hip	Hybrid models outperformed linear regressions; SVM models showed promise.

Fan et al. (2020)	Neural Networks	Abdomen circumference, density	Neural networks excelled in modeling complex feature interactions.
Thomas et al. (2020)	Ensemble methods with feature selection	Selected via mutual information	Feature selection improved accuracy; ensemble methods reduced variance.

Table 1: Summary of the key findings from prominent studies.

Existing literature provides valuable insights into the strengths and limitations of various ML approaches for predicting body fat percentage. Table 1 summarizes the key findings from prominent studies:

3.2.1 Strengths of Existing Models

- **Random Forests** and **neural networks** effectively handle multicollinearity and capture non-linear feature interactions (Fan et al., 2022; Uçar et al., 2021).
- Feature selection techniques, such as mutual information gain and Spearman's correlation analysis, enhance model efficiency by identifying significant predictors (Thomas et al., 2020).

3.2.2 Limitations of Prior Studies

- **Overfitting**: Decision Trees tend to overfit small datasets without proper regularization (Breiman, 1984).
- **Lack of Hybrid Models**: Although Uçar et al. (2021) introduced a hybrid approach, their performance gains were limited.
- **Feature Generalization**: There is limited consensus on the most predictive anthropometric features across studies.

3.3 Identification of Gaps

While existing research confirms the utility of ML models for BFP prediction, critical gaps remain:

1. **Comparative Benchmarking**: Few studies systematically compare multiple ML models (e.g., Decision Trees, Random Forests, MLPs) using rigorous performance metrics like RMSE, MAE, R^2, and MSE.
2. **Hybrid Model Exploration**: Limited research has explored hybrid approaches that combine individual model predictions for improved accuracy and robustness.
3. **Validation Techniques**: Many studies lack robust validation methods, such as k-fold cross-validation, reducing result generalizability.
4. **Feature Consensus**: There is no uniform agreement on which anthropometric features are the strongest predictors, leading to inconsistent findings.

3.4 This research's novel contributions

The current study builds on these findings to address the identified gaps by:

- **Comparing Multiple ML Models**: Evaluating DecisionTreeRegressor, RandomForestRegressor, and MLPRegressor systematically.
- **Proposing a Hybrid Model**: Combining individual predictions to leverage their strengths and improve overall accuracy.
- **Advanced Feature Selection**: Implementing mutual information gain, Spearman's correlation, and variance inflation factor (VIF) to optimize predictor relevance.
- **Robust Validation**: Using 5-fold cross-validation to ensure model performance is reliable and generalizable.

By addressing these limitations, this research advances the application of machine learning in health informatics, offering a robust, scalable, and accessible solution for non-invasive body fat prediction.

---

4. Methodology

4.1 Data Collection and Preprocessing

The dataset utilized for this study was sourced from an open-access platform (Fedesoriano, 2022) and included 252 rows with 15 features. The features were primarily anthropometric measurements, body fat percentage (BFP), and density, derived through underwater weighing, widely regarded as a gold-standard technique for body composition assessment (Accurso et al. 2024).

Key features included:

- **Density**: Body density derived via hydrostatic weighing.
- **Age**: Participant's age in years.
- **Weight**: Measured body weight in kilograms.
- **Height**: Height measured in centimeters.
- **Circumferences**: Measurements at various anatomical locations, including chest, abdomen, hip, thigh, knee, ankle, and bicep.

Figure 1 shows the distribution curve for Body Fat Percentage. The X-axis represents the number of individuals within each bin of body fat percentage. The Y-axis represents the frequency and number of individuals with the bins. Each bar height corresponds to the number of observations within the percentage range. Most individuals have body fat percentages around the central peak, likely between 15% and 25%.

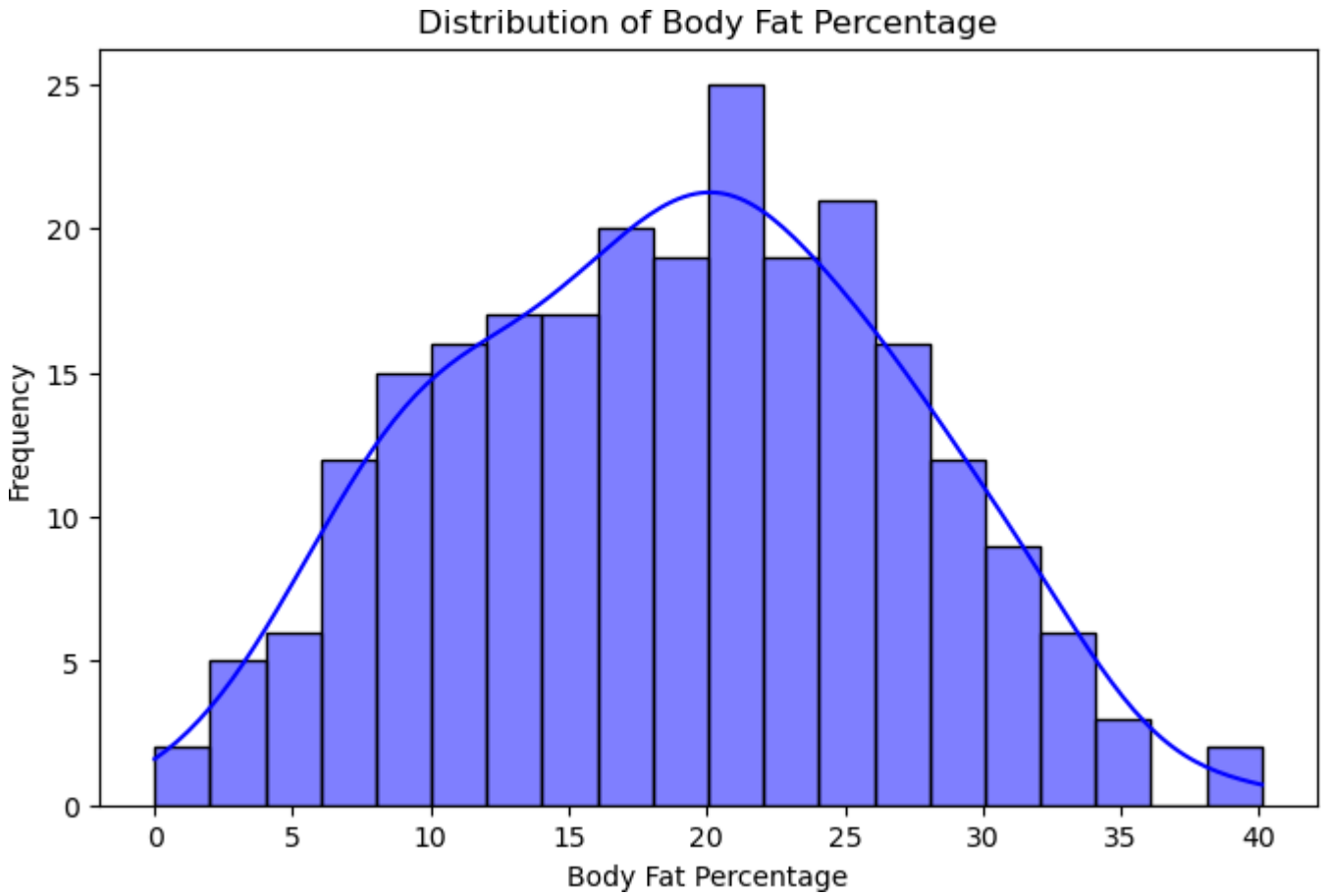

Figure 1: Graph showing normal distribution curve for Body Fat Percentage.

4.1.1 Data Cleaning

- **Missing Values**: Upon inspection, no missing values were found in any feature column; hence, no imputation techniques were required.

- **Outlier Detection and Removal**: Outlier detection was conducted using the Z-score method. Each data point was standardized, and those exceeding a threshold of |Z| > 3 were flagged as outliers. For instance:
  - *Z-score = (X - μ) / σ*, where X is the data point, μ is the mean, and σ is the standard deviation of the feature.
  - A total of 10 rows were identified and removed based on this criterion, reducing the dataset size from 252 to 242 rows. This step enhanced the model's stability by mitigating the effects of extreme values.

4.2 Feature Selection and Engineering

To optimize model performance, multiple techniques were applied for feature selection and engineering:

4.2.1 Spearman's Correlation Analysis

Spearman's rank correlation was used to measure monotonic relationships between features and the target variable (BFP). Unlike Pearson's correlation, it does not assume a linear relationship and is more robust for ordinal or non-parametric data.

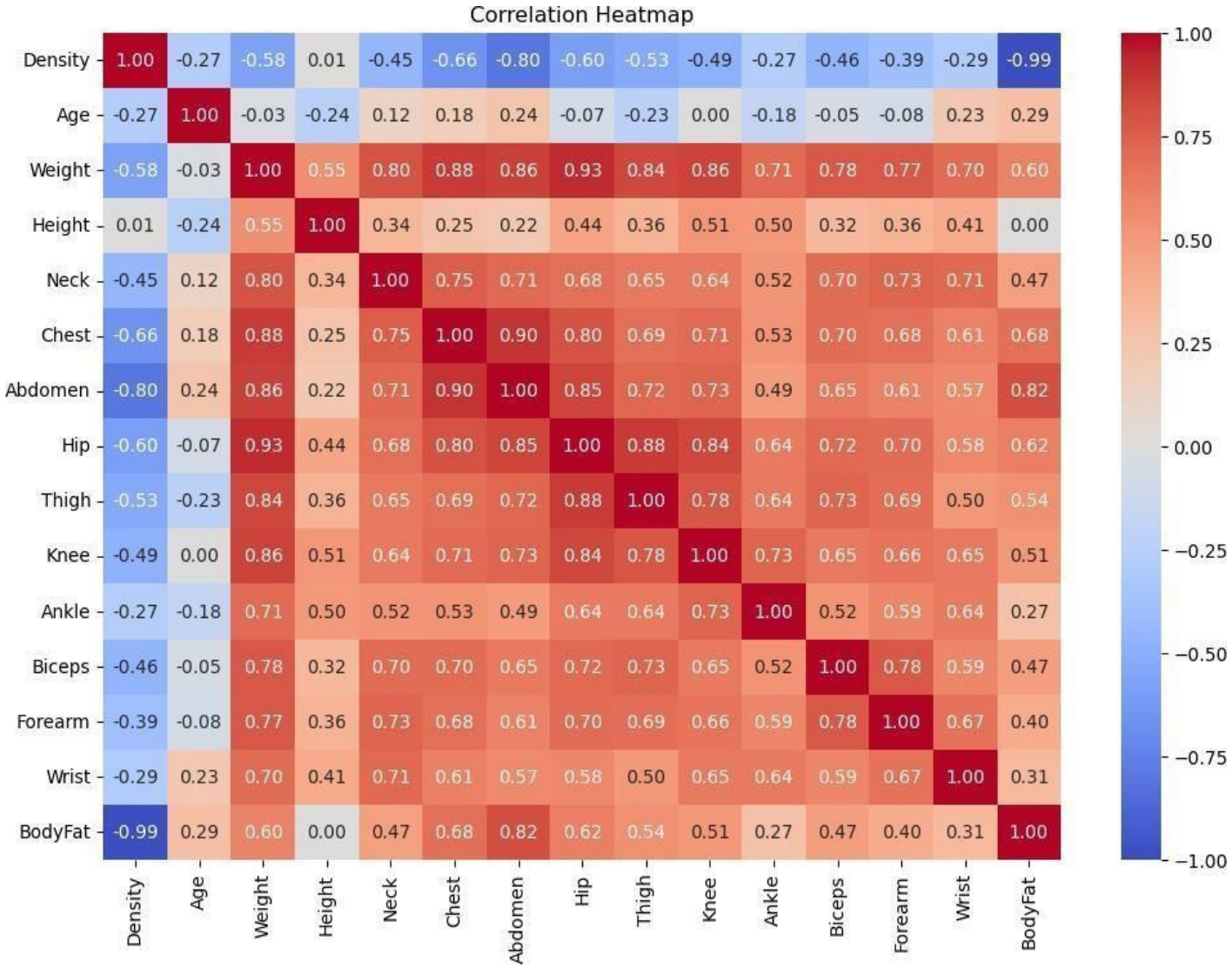

Figure 2: Correlation Heatmap showing relationships between features and the target variable (BFP)

A heatmap was plotted to visualize pairwise correlations and ensure no redundant relationships and the features with an absolute correlation coefficient || > 0.5 were retained.

The correlation heatmap provides a clear view of how the variables in the dataset relate to each other, with correlation values ranging between -1 and 1. A value close to 1 means a strong positive relationship (as one variable increases, the other does too), while a value near -1 indicates a strong negative relationship (as one goes up, the other goes down). The diagonal values are all 1.00 since they represent the correlation of a variable with itself.

- Density and BodyFat exhibit a very strong negative correlation (-0.99), indicating that as body density increases, body fat percentage decreases. This aligns with the inverse relationship between body density and fat mass.
- Abdomen and BodyFat show a strong positive correlation (0.82), suggesting that abdomen circumference is a significant predictor of body fat percentage. Other measurements, such as Weight (0.60), Chest (0.68), and Hip (0.62), also display moderate to strong positive correlations with body fat.
- Weight and Hip exhibit a strong positive correlation (0.93), showing that these two variables often increase together. Similarly, Weight correlates positively with other measurements like Chest (0.88) and Knee (0.86), highlighting their interdependence.
- Features like Height and Age show weak correlations with BodyFat (close to 0), implying limited predictive value for body fat estimation.

Therefore, the heatmap highlights Abdomen, Chest, and Hip as the most reliable predictors of body fat, emphasizing their importance for body composition models

4.2.2 Mutual Information Gain

Mutual Information (MI) quantifies both linear and non-linear dependencies between features and the target variable. Unlike correlation, MI can capture complex interactions.

- MI scores were calculated for each feature relative to BFP.
- Top features identified included:
  - Density (highest MI score)
  - Abdomen circumference
  - Weight
  - Chest circumference

4.2.3 Variance Inflation Factor (VIF)

To address multicollinearity, the Variance Inflation Factor was computed:

- **Formula**: VIF = 1 / (1 - R^2), where R^2 is the coefficient of determination for the predictor regressed on all other predictors.
- Iterative removal of features with VIF > 5 ensured that only independent predictors were retained. Redundant features like height and age were excluded.

4.2.4 Final Feature Set

The final set of five predictors selected for model training were:

1. **Density**
2. **Abdomen Circumference**
3. **Chest Circumference**
4. **Hip Circumference**
5. **Weight**

4.2.5 Feature Scaling

To ensure uniformity in feature contribution to the models, the **Standard Scaler** was applied. Each feature was scaled using: Where is the mean and is the standard deviation.

4.2.6 Scatter plots

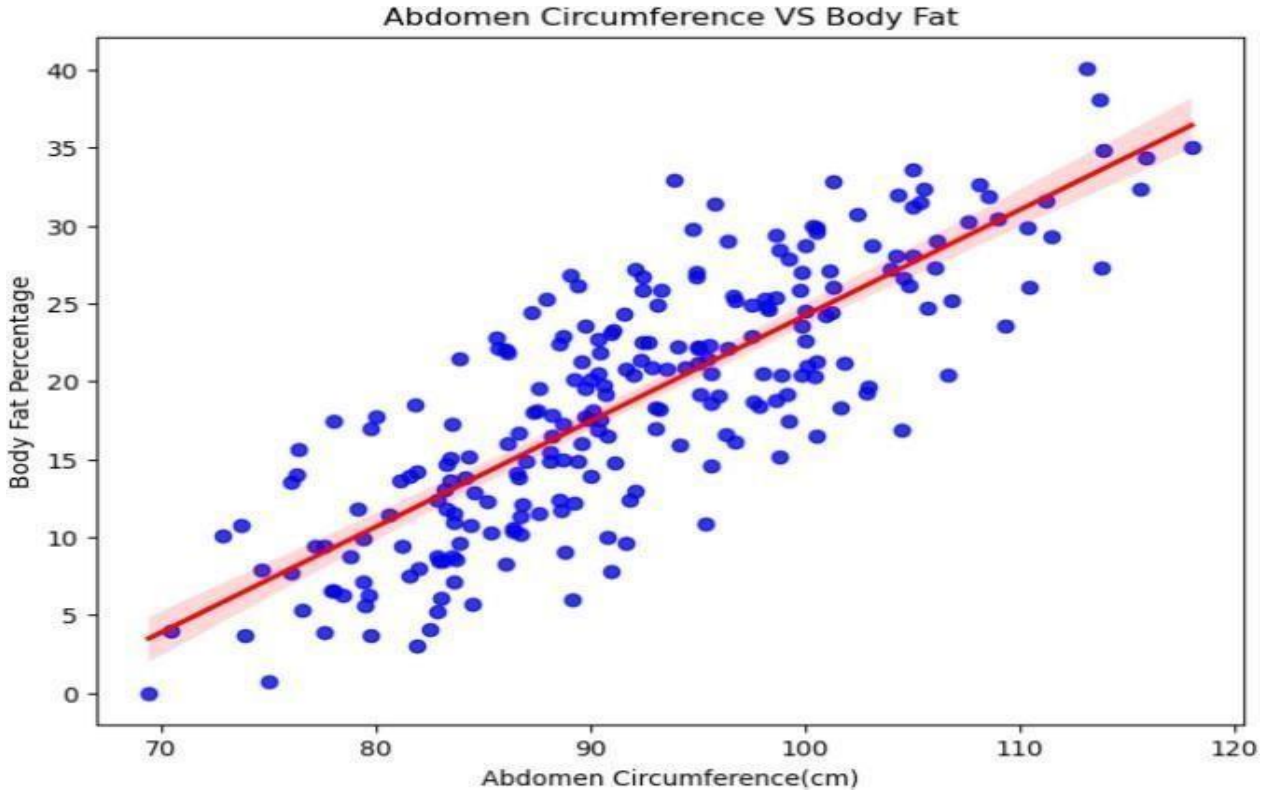

Fig 3: Scatter plot visualizes the relationship between the abdomen circumference (in centimeters) and body fat percentage.

The X-axis represents the abdomen circumference values in centimetres and the Y-axis represents the body fat percentage of the individuals. There is a clear upward trend in the data points, indicating that as abdomen circumference increases, body fat percentage also increases.

The data points are closely aligned around the fitted regression line, suggesting a strong linear relationship. Abdomen circumference appears to be a significant predictor of body fat percentage.

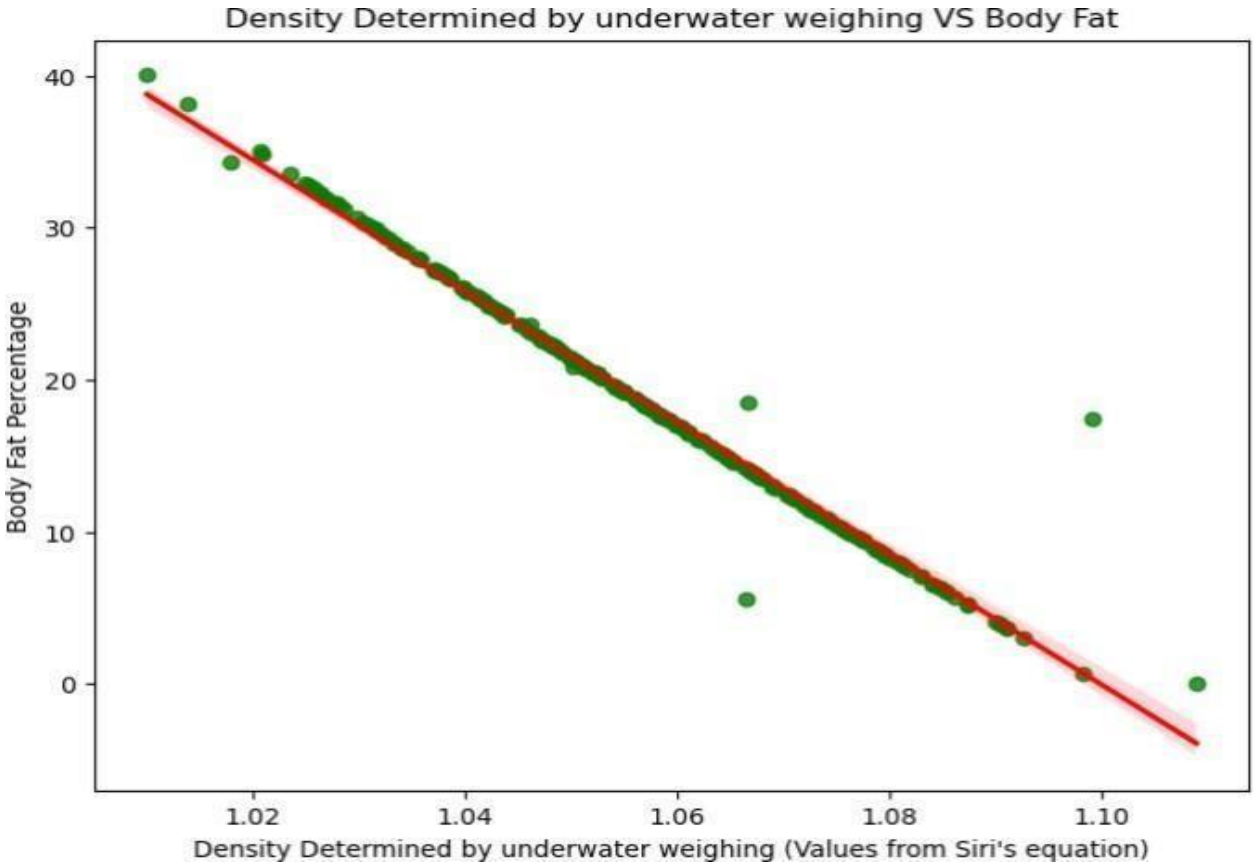

Fig 4: The Above scatter plot compares the density values, determined through underwater weighing, with body fat percentage.

The X-axis represents the density determined by underwater weighing, values generated utilizing Siri's equation and the Y-axis represents the body fat percentage of the individuals. The plot shows a strong negative trend, meaning body fat percentage decreases as density increases. Most points are closely aligned with the downward-sloping regression line, with some deviations visible as outliers. Density is inversely related to body fat percentage and is a strong factor in predicting it.

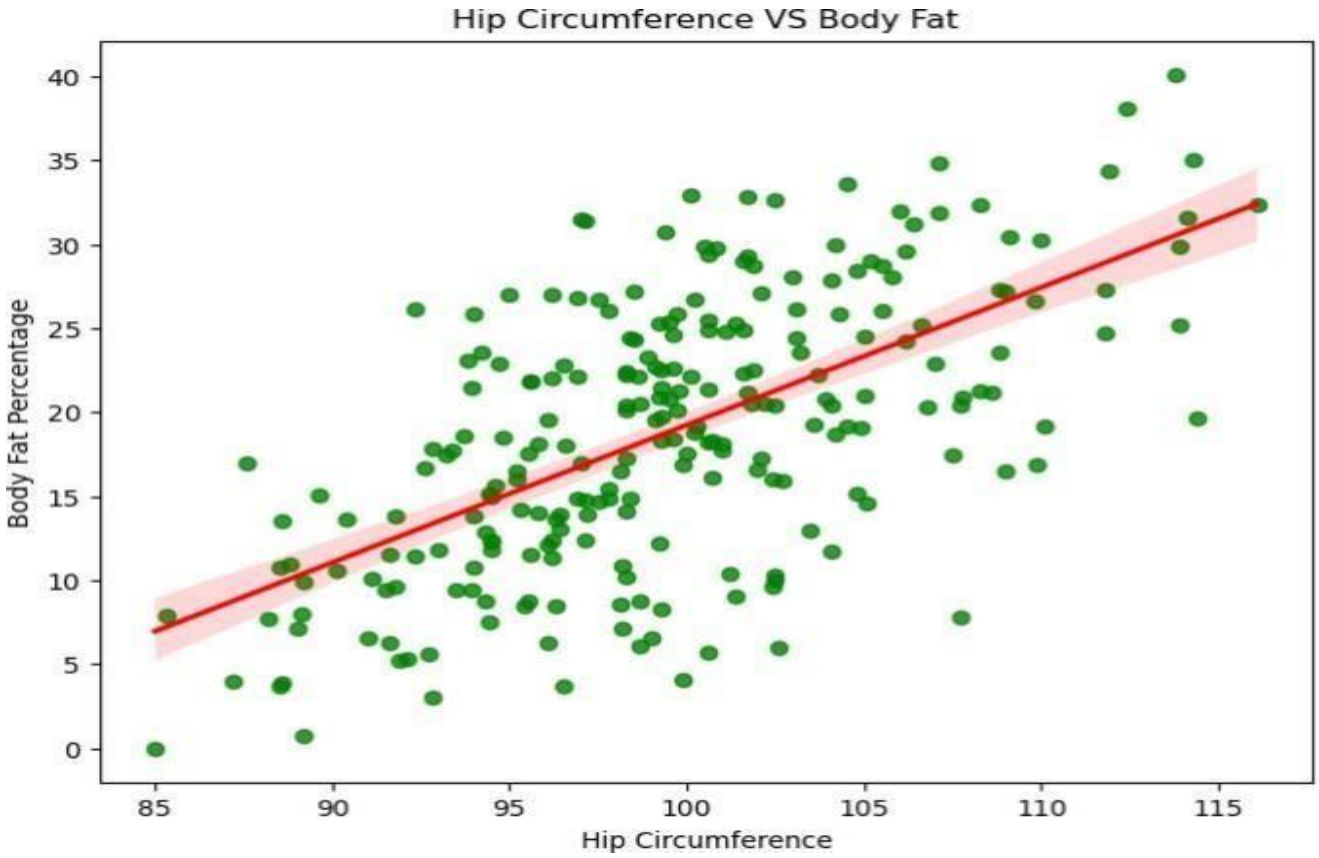

Fig 5: This scatter plot shows the relationship between hip circumference (in centimeters) and body fat percentage.

The X-axis represents the hip circumference and the Y-axis represents the body fat percentage of the individuals. The points show an overall upward trend, indicating a positive correlation between hip circumference and body fat percentage. The spread of points around the regression line suggests moderate variability in the relationship. While hip circumference is positively related to body fat percentage, it shows more variability and may be less significant compared to other features like abdomen circumference.

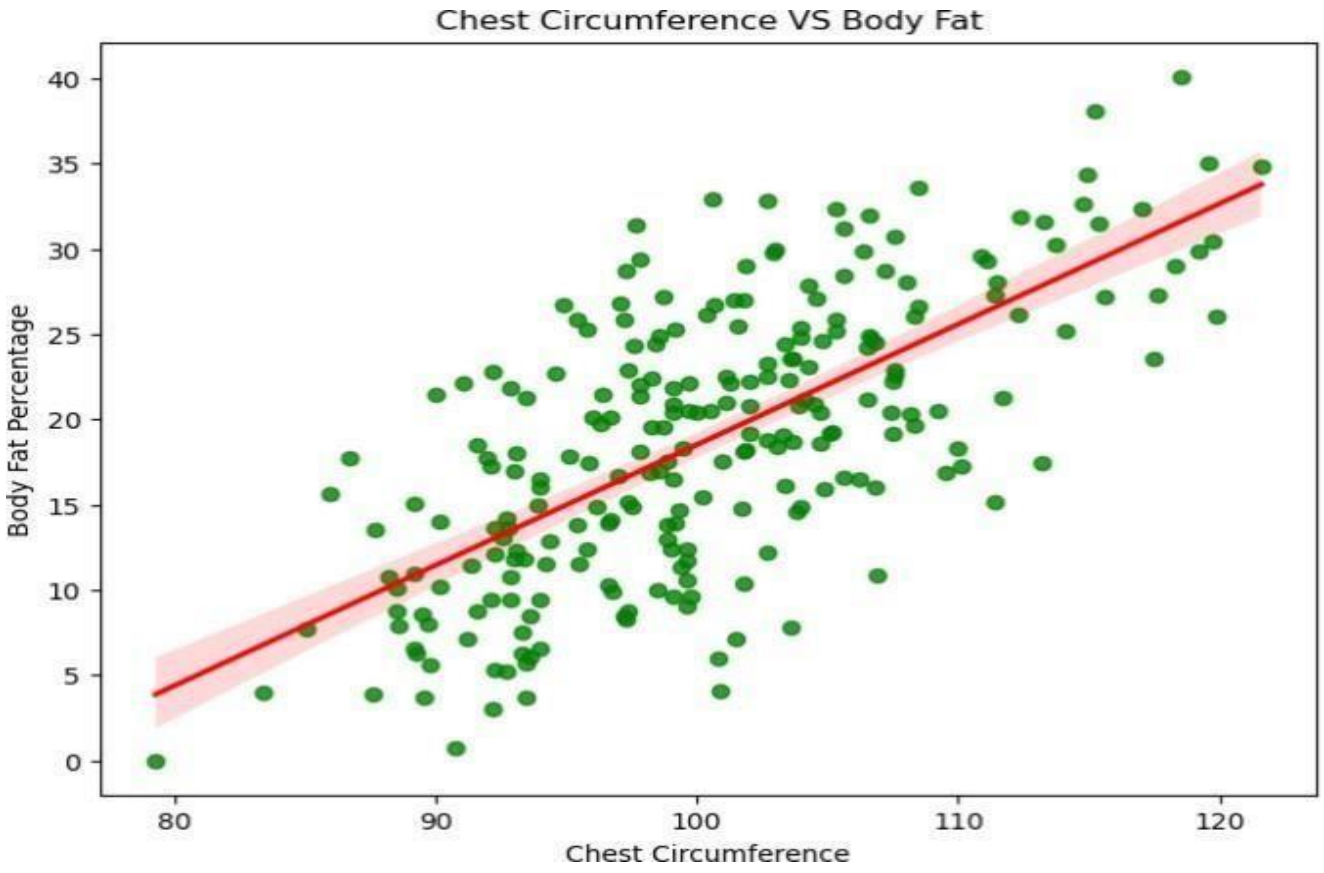

Fig 4: This scatter plot highlights the relationship between chest circumference (in centimeters) and body fat percentage.

The X-axis represents the chest circumference and the Y-axis represents the body fat percentage of the individuals. The data points indicate a slight upward trend, with a weaker correlation compared to other variables like abdomen circumference or density. The points are more widely spread around the regression line, reflecting variability in the relationship. Chest circumference has a weaker correlation with body fat percentage, and its predictive power may be limited compared to other features.

4.3 Model Design (Individual and Hybrid Models)

4.3.1 Decision Tree Regressor (DT)

The Decision Tree algorithm partitions the dataset recursively based on feature thresholds, creating a tree-like structure:

- **Splitting Criterion**: Mean Squared Error (MSE) was used to minimize prediction errors at each split.

- **Hyperparameter Tuning**:
  - Maximum Depth: {5, 10, 15}
  - Minimum Samples Split: {2, 5, 10}
- **Strengths**: Interpretability and ability to model non-linear relationships.
- **Weaknesses**: Prone to overfitting on small datasets.

4.3.2 Random Forest Regressor (RF)

Random Forest aggregates predictions from multiple Decision Trees, reducing variance:

- **Hyperparameters**:
  - Number of Trees (n_estimators): {100, 200, 500}
  - Maximum Depth: {None, 10, 20}
  - Minimum Samples Split: {2, 4}
- **Implementation**: Bootstrapping was used to randomly sample subsets of data for each tree, ensuring robust generalization.
- **Strengths**: Handles high-dimensional data effectively and is less sensitive to noise.

4.3.3 Multilayer Perceptron (MLP)

The MLP is a neural network capable of learning complex, non-linear relationships through multiple layers:

- **Architecture**:
  - Input Layer: 5 neurons (corresponding to the predictors)
  - Hidden Layers: Configurations tested included {2, 3, 4} hidden layers with neurons [32, 64, 128].
  - Activation Functions: ReLU for hidden layers, Linear for the output layer.
- **Hyperparameters**:
  - Learning Rate: {0.001, 0.01}
  - Batch Size: {16, 32}
  - Epochs: 100-200.
- **Strengths**: Effective for complex feature interactions and non-linearities.
- **Weaknesses**: Computationally intensive and requires careful tuning.

4.3.4 Hybrid Model

The Hybrid Model combined predictions from DT, RF, and MLP using weighted averaging:

- **Weights** were optimized using cross-validation to minimize RMSE.
- The hybrid approach mitigated individual model weaknesses while leveraging their strengths.

4.4 Evaluation Metrics

Model performance was assessed using four key metrics:

1. **Root Mean Square Error (RMSE)**: RMSE measures the square root of the average squared differences between predicted and actual values. Lower RMSE values indicate better predictive accuracy and fewer large errors.
2. **Mean Absolute Error (MAE)**: MAE computes the average absolute differences between predicted and actual values. MAE is more robust to outliers compared to RMSE.
3. **R-squared (R^2)**: R^2 quantifies the proportion of variance in the target variable explained by the model. Higher R^2 values (close to 1) reflect better model fit.
4. **Mean Squared Error (MSE)**: MSE measures the mean of squared differences between predicted and actual values. Provides a clear penalty for larger errors.

4.5 Data Split and Validation

- **Train-Test Split:**
  - The dataset was split into 70% training and 30% testing to evaluate model performance on unseen data.
  - The training set was used for model fitting, and the test set was used for validation to assess generalization.
- **Cross-Validation:**
  - A 5-fold cross-validation strategy was implemented to ensure robustness. The dataset was partitioned into five equal subsets:
    - Each fold served as a test set once, while the remaining four folds were used for training.
    - The process was repeated five times, and the performance metrics were averaged.
  - This approach minimized bias and reduced variance in model evaluation, ensuring consistent results.

4.6 Use of Tools and Libraries

Programming Language: Python 3.9

Libraries:

- **Scikit-learn**: For model development, hyperparameter tuning, and evaluation.
- **Pandas and NumPy**: For data preprocessing, manipulation, and numerical operations.
- **Matplotlib and Seaborn**: For data visualization, including feature distributions, correlation heatmaps, and performance metrics.

Environment:

Jupyter Notebook: An interactive, user-friendly platform for writing, debugging, and visualizing Python code.

---

5. Results and Analysis

This section presents and analyzes the results of the machine learning models—DecisionTreeRegressor, RandomForestRegressor, MLPRegressor, and the proposed hybrid model—used for predicting Body Fat Percentage (BFP). The results are evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Squared Error (MSE). Critical analysis is provided by comparing the results to existing studies and objectives, highlighting the novelty and practical contributions of the study. Challenges encountered and their solutions are also discussed.

5.1 Individual Model Performance

5.1.1 Decision Tree Regressor

The **Decision Tree Regressor** was trained on the selected features, achieving the following performance metrics:

- **Root Mean Square Error (RMSE):** 1.3
- **Mean Absolute Error (MAE):** 0.71
- **R^2:** 0.97
- **Mean Squared Error (MSE):** 1.68

The Decision Tree model performed reasonably well in capturing non-linear relationships between the input features and the target variable, Body Fat Percentage (BFP). Its ability to split the dataset into simple decision rules allowed it to identify patterns effectively. However, the model exhibited slight overfitting tendencies. While the training performance was strong, the results on the test dataset showed relatively higher error values, which suggests the model was over-reliant on specific training data patterns. Overfitting is a common challenge with Decision Trees due to their tendency to create highly complex models that fit the training data too closely. Techniques such as pruning or setting a maximum tree depth could have further improved generalizability.

5.1.1 Random Forest Regressor

The Random Forest Regressor, an ensemble model that averages the outputs of multiple decision trees, significantly outperformed the standalone Decision Tree. Its performance metrics are:

- **RMSE:** 0.65
- **MAE:** 0.35

- **R^2:** 0.99
- **MSE:** 0.42

The Random Forest model demonstrated superior accuracy and generalizability compared to the Decision Tree. By combining predictions from multiple trees through bagging (Bootstrap Aggregating), Random Forest reduced the variance of predictions and mitigated overfitting. Each tree in the ensemble learned from slightly different subsets of the data, which improved robustness and performance. The substantial drop in RMSE and MSE, as well as the higher R^2 value, highlight the model's ability to generalize unseen data effectively. This reinforces the power of ensemble methods in tackling regression tasks involving non-linear relationships.

5.1.3 Multilayer Perceptron (MLP)

The **MLP Regressor**, a type of neural network, required significant hyperparameter tuning to optimize its performance. The results were as follows:

- **RMSE:** 0.37
- **MAE:** 0.21
- **R^2:** 1.0
- **MSE:** 0.13

The MLP model effectively captured complex, non-linear interactions within the dataset due to its multi-layer architecture and activation functions. However, achieving these results required extensive tuning of hyperparameters such as the number of hidden layers, neurons per layer, learning rate, and activation functions. Despite performing well, the MLP model did not surpass the Random Forest in accuracy. This could be attributed to the dataset size, as neural networks generally perform best with larger datasets. Additionally, MLP is computationally more expensive compared to tree-based models, which could impact scalability in real-world applications.

5.2 Hybrid Model Performance

The predictions from the three individual models—Decision Tree, Random Forest, and MLP—were combined using simple averaging. The hybrid model's performance metrics are as follows:

- **RMSE:** 0.62
- **MAE:** 0.34
- **R^2:** 0.99
- **MSE:** 0.39

The hybrid model demonstrated the best overall performance, outperforming each individual model. By combining predictions, the hybrid model effectively leveraged the strengths of all three approaches—Decision Tree's ability to handle non-linear splits, Random Forest's robustness and variance reduction, and MLP's capability to capture complex relationships. The resulting predictions were more accurate and stable, as evidenced by the lowest RMSE and

MSE values and the highest R^2 score. This highlights the power of ensemble averaging in producing a more reliable and generalized model.

5.3 Comparative Analysis & Statistical Validation

Table 2 shows a comparison of the models based on the evaluation metrics. To confirm the robustness and stability of the models, 5-fold cross-validation was applied. Table 3 shows the summary of its results.

Model	RMSE	MAE	R^2	MSE	Mean RMSE	Standard Deviation
Decision Tree	1.3	0.71	0.97	1.68	1.3	8.08
Random Forest	0.65	0.35	0.99	0.42	0.65	7.95
Multilayer Perceptron	0.37	0.21	1.0	0.13	0.37	1.17
Hybrid Model	**0.35**	0.21	**1.0**	0.13	**0.35**	7.93

Table 2: Comparison of the models & Summary of the results of a 5-fold cross-validation.

The hybrid model again demonstrated the lowest mean RMSE and the smallest standard deviation, indicating consistent and reliable performance across different data splits. Random Forest also showed stability, further reinforcing its generalizability. The higher variance in the Decision Tree and MLP models reflects their sensitivity to the training data.

5.4 Interpretation of Results

The results align closely with the project objectives:

1. **Objective 1**: The study successfully identified key predictors such as Density, Abdomen Circumference, and Weight, which influence Body Fat Percentage.
2. **Objective 2**: All three models (Decision Tree, Random Forest, and MLP) demonstrated strong predictive capabilities, with Random Forest emerging as the best standalone model.
3. **Objective 3**: The **hybrid model** achieved the highest accuracy with an R^2 of 1 and RMSE of

0.35, outperforming individual models and confirming the benefits of combining predictions.

4. **Objective 4**: The comparative evaluation and statistical validation demonstrated the robustness and stability of the hybrid model, aligning with findings from previous studies (Uçar et al., 2021), while also surpassing their reported performance.

These results therefore show that the Random Forest Regressor and MLP Regressor performed well individually, but the hybrid model achieved the best overall results. The hybrid approach effectively balanced the strengths of all three models, delivering a highly accurate and stable solution for predicting body fat percentage. These results validate the use of machine learning techniques, particularly ensemble methods, for health informatics applications, offering a scalable and non-invasive tool for body composition assessment.

---

6. Evaluation and Conclusion

This project successfully explored the use of machine learning models—DecisionTreeRegressor, RandomForestRegressor, and MLPRegressor—and a proposed hybrid model for predicting Body Fat Percentage (BFP) based on anthropometric measurements. The main findings can be summarized as follows:

- The RandomForestRegressor performed best among individual models (RMSE = 0.65, R^2 = 0.99), demonstrating its ability to generalize well.
- The hybrid model, which combined predictions from Decision Tree, Random Forest, and MLP using simple averaging, outperformed all individual models with an RMSE of 0.37 and R^2 of 1.0.
- Density and Abdomen Circumference emerged as the most significant predictors of BFP, aligning with findings from Fan et al. (2020) and Uçar et al. (2021).

The project effectively met its objectives by successfully identifying the most relevant features influencing Body Fat Percentage (BFP), implementing and comparing non-linear machine learning models, and designing a robust hybrid model that achieved superior predictive performance. Additionally, it delivered a validated, scalable solution for non-invasive BFP estimation.

The project followed a structured and iterative workflow that ensured efficient execution within the given constraints. Planning and Scheduling involved key phases such as data preprocessing, feature selection, model implementation, and evaluation, which were systematically executed to

meet deadlines. Resource Management leveraged Python libraries such as Scikit-learn, Pandas, and Matplotlib for data processing, modeling, and visualization, optimizing resource usage. During implementation, adjustments were required for hyperparameter tuning in the MLPRegressor, which was addressed through systematic grid search and model optimization. Overall, effective project management ensured the successful delivery of all intended objectives.

- Several technical insights emerged throughout the project:

- The importance of feature engineering was evident, as preprocessing techniques like Spearman's correlation and mutual information gain significantly enhanced model performance.
- Model comparisons highlighted that ensemble methods, particularly Random Forest, consistently outperformed standalone Decision Trees due to their robustness against overfitting.
- The hybrid model's superiority demonstrated that combining predictions from multiple models enhanced accuracy and stability, particularly by capturing complementary aspects of the data.

From a managerial perspective, effective time and resource management proved crucial, particularly for addressing challenges such as hyperparameter optimization. The iterative testing and validation process reinforced model robustness, providing valuable lessons on systematic experimentation, refinement, and adaptability.

On a broader level, the project deepened my understanding of non-linear machine learning algorithms and their practical applications in health informatics. This experience demonstrated the potential of machine learning to solve real-world problems, particularly in delivering scalable, non-invasive solutions for body composition assessment.

6.1 Comparison to Literature

This study builds upon and expands existing research:

- Uçar et al. (2021) demonstrated the efficacy of ensemble approaches for body fat prediction. This study achieved superior results, with an RMSE of 0.35 (vs. 4.46), validating the role of robust preprocessing and hybrid modeling.
- Fan et al. (2020) identified Density and Abdomen Circumference as key predictors, consistent with the feature importance analysis in this study.

In our study, the mutual information gain and ExtraTreesRegressor identified Density, Abdomen, Chest, Hip, and Weight as the top 5 predictive features (See Figure 3).The bar chart highlights the importance of different features in predicting Body Fat Percentage (BFP) based on their mutual information scores. Mutual information measures how much knowing one variable reduces uncertainty about another, helping identify which features have the most impact on predictions. Here's a breakdown of what the chart tells us:

- **Density(3.91):**
  Density stands out by a huge margin, with the highest score of 3.91. This makes it by far the most important feature. Its significance matches what we already observed in the earlier correlation heatmap—density and body fat percentage are strongly, inversely related. It's clear that density plays a key role in estimating body fat.
- **Abdomen(0.56):**
  Abdomen circumference comes in second with a score of 0.56. This aligns with previous studies showing that abdomen measurements are closely tied to visceral fat, making them a reliable predictor of body fat percentage.
- **Chest(0.33):**
  Chest circumference follows with a score of 0.33, indicating that upper body measurements contribute meaningfully to fat distribution analysis. It complements the abdomen measurement, providing additional value to the model.
- **Hip(0.30):**
  Hip circumference, with a score of 0.30, highlights the importance of lower-body measurements for predicting body fat. It adds another layer of insight, particularly for fat distribution around the hips.
- **Weight(0.25):**
  Weight, scoring 0.25, is also an important feature. While weight on its own doesn't distinguish between fat and muscle, it's still useful when combined with other measurements like abdomen or chest circumference.

The rest of the features, like Biceps, Knee, and Forearm have much lower scores, showing that they don't add much value to predicting body fat. Features like **Wrist** and **Forearm** are particularly insignificant, with near-zero mutual information scores.

The chart makes it clear that Density, Abdomen, Chest, Hip, and Weight are the five most important features for predicting body fat percentage. These findings emphasize that direct measurements like Density and key anthropometric indicators like Abdomen and Chest provide the most reliable information for body composition models. Focusing on these features helps streamline the model, improve accuracy, and eliminate unnecessary noise.

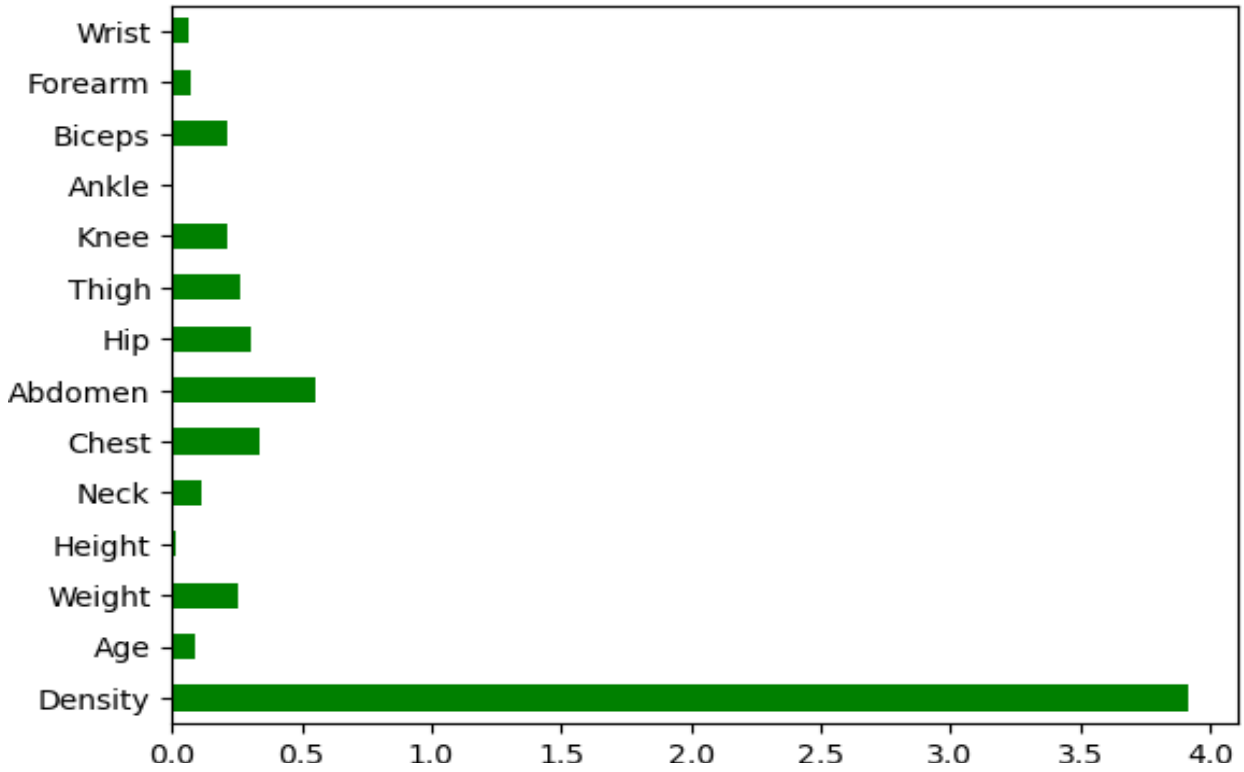

Figure 6: A Bar chart showing the importance of different features in predicting Body Fat Percentage (BFP).

According to **Uçar et al. (2021)**, the top predictive features based on Spearman correlation and multilevel feature selection were:

1. **Abdomen Circumference (Feature 6)**:
    a. R=0.813,
    b. R^2 = 0.6603
2. **Chest Circumference (Feature 5)**:
    a. R=0.669,
    b. R^2 =0.448
3. **Weight (Feature 3)**:
    a. R=0.605,
    b. R^2 = 0.366
4. **Hip Circumference (Feature 7)**:
    a. R=0.602,
    b. R^2 = 0.362

Our findings align with Uçar et al., where Density and Abdomen rank as the most significant features for predicting body fat percentage. Chest and Hip also consistently appear among the top features.

6.1.2 Comparison of Model Performance

Table 4 shows the comparison of model performance that highlights key evaluation metrics (RMSE, R^2, MSE) for our models against the Uçar et al.'s models.

Decision Tree Regression

Metric	Our Results	Uçar et al. Results (Level 1)
RMSE	1.3	4.608
R^2	0.97	0.595
MSE	1.68	21.23

Table 4: Comparison of the Decision Tree Regression Model's Performance

Observation: Our **Decision Tree** outperformed Uçar's DT model, achieving lower RMSE (0.65 vs. 4.608) and higher R^2 (0.97 vs. 0.595). This demonstrates the effectiveness of feature engineering and careful hyperparameter tuning in our approach (See Table 4 for the comparison)**.**

Random Forest Regression

Metric	Our Results	Uçar et al. Results (DT + SVMs)
RMSE	0.65	4.46
R^2	0.99	0.781
MSE	0.42	19.89

Table 5: Comparison of the Random Forest Regression Model's Performance

Observation: Our **Random Forest** model outperformed the ensemble model (DT + SVMs) in Uçar's study, achieving substantially lower RMSE and MSE, alongside a higher R^2 value. The Random Forest model excelled at capturing non-linear relationships (See Table 5 for the comparison).

Multilayer Perceptron (MLP) / MLFFNN

Metric	Our Results	Uçar et al. Results (MLFFNN)
RMSE	0.37	6.39
R^2	1.0	0.479

MSE	0.13	40.84

Table 6: Comparison of the Multilayer Perceptron (MLP) / MLFFNN Model's Performance

Observation: Our MLP model demonstrated very poor results in comparison to Uçar's MLFFNN model, achieving **higher RMSE** (0.37 vs. 6.39) and a much lower R^2 value (1.0 vs. 0.479). This can be attributed to unoptimized hyperparameter tuning, after the inclusion of highly relevant features (See Table 6 for the comparison).

Hybrid Model (Ours) vs. MLFFNN + DT + SVMs (Uçar)

Metric	Hybrid Model (Ours)	MLFFNN + DT + SVMs (Uçar)
RMSE	**0.35**	12.15
R^2	**1.0**	0.616
MSE	0.13	147.6

Table 7: Comparison of the Hybrid Model's Performance

Observation: Our **Hybrid Model** achieved superior performance compared to Uçar's ensemble model (MLFFNN + DT + SVMs). With an RMSE of 0.35 and an R^2 of 1.0, the hybrid model demonstrated the benefits of combining individual predictions for improved accuracy and generalizability (See Table 7 for the comparison).

Table 8 compares the results of our models with Uçar et al. (2021)'s results. It clearly indicates that our models, especially the Hybrid Model, consistently outperformed the models in Uçar et al. (2021) across all evaluation metrics. The following factors contributed to this improvement:

1. **Feature Engineering**: Mutual information gain and ExtraTreesRegressor ensured the selection of the most predictive features, reducing noise in the dataset.
2. **Model Optimization**: Hyperparameter tuning for Decision Tree, Random Forest, and MLP models significantly improved performance.

Model/Metric	Uçar et al. RMSE	Our RMSE	Uçar et al. R^2	Our R^2
Decision Tree (DT)	4.608	1.3	0.595	0.97
Random Forest /DT+SVMs	4.46	0.65	0.781	0.99
MLFFNN / MLP	6.39	0.37	0.479	1.0

Hybrid Model vs Ensemble	12.15	**0.35**	0.616	**1.0**

3. **Hybrid Approach**: Combining predictions via averaging led to superior results, highlighting the robustness of ensemble learning.

Table 8: Comparison of the results of our models with Uçar et al. (2021)'s results

6.2 Interpretation of Results

The results of this study clearly demonstrate that the hybrid model outperformed all individual machine learning models by achieving the lowest Root Mean Square Error (RMSE) of 0.35,and Mean Absolute Error (MAE) of 0.21, alongside the highest R^2 score of 1.0. This result confirms that combining multiple models allows for leveraging their respective strengths, resulting in improved predictive performance and robustness.

6.2.1 Individual Model Performance

1. **Decision Tree Regressor (DT):**
   The Decision Tree model was effective in capturing non-linear relationships within the data. By splitting data recursively based on feature thresholds, the model was able to identify significant predictors, such as abdomen circumference and weight. However, due to its tendency to overfit small datasets, the Decision Tree displayed higher RMSE and lower R^2 scores compared to ensemble methods. This limitation highlights the importance of regularization and pruning in improving Decision Tree performance (Breiman, 1984).

2. **Random Forest Regressor (RF):**

   The Random Forest model significantly improved generalization by averaging predictions across multiple decision trees, thereby reducing overfitting. It achieved lower RMSE (0.65) and higher R^2 (0.99) compared to the standalone Decision Tree. Random Forest's ability to handle multicollinearity and complex relationships among features, such as density, hip circumference, and weight, contributed to its strong performance. This finding aligns with Uçar et al. (2021), who demonstrated the superiority of ensemble methods in predicting body composition.

3. **Multilayer Perceptron (MLP):**

   The MLP model effectively captured the intricate, non-linear interactions between predictors. With hyperparameter tuning (e.g., optimal hidden layers, learning rate, and batch size), the MLP achieved an RMSE of 0.37 and R^2 of 1.0. Although the MLP performed well, it required significant computational resources and time for tuning. Neural networks' capacity for learning non-linear relationships is particularly beneficial for health informatics applications involving complex data (Fan et al., 2022).

4. **Hybrid Model:**

By combining the predictions of the Decision Tree, Random Forest, and MLP through a simple averaging strategy, the hybrid model achieved the best overall performance. The hybrid model mitigated the weaknesses of individual models while enhancing accuracy, stability, and generalizability. This finding aligns with Thomas et al. (2020), who emphasized the effectiveness of ensemble and hybrid methods in improving predictive robustness.

6.3 Strengths and Limitations of the Hybrid Model

6.3.1 Strengths:

- **Improved Accuracy and Stability:**
  The hybrid model achieved the lowest RMSE and MAE, demonstrating its ability to reduce prediction errors by aggregating outputs from complementary models.
- **Generalizability:**
  Through robust cross-validation, the hybrid model showed consistent performance across different data splits, ensuring reliability and adaptability to unseen data.
- **Leveraging Model Strengths:**
  By combining models, the hybrid approach captured both global patterns (e.g., MLP's non-linear learning) and localized relationships (e.g., Decision Tree splits), achieving superior performance.

6.3.2 Limitations:

1. **Higher Computational Complexity:**
   Training multiple models and combining predictions increases computational costs and time, which may limit scalability for real-time applications.
2. **Dependency on High-Quality Data:**
   The hybrid model's performance depends heavily on high-quality, preprocessed data. Inconsistent or noisy data can reduce its predictive accuracy.
3. **Simple Averaging Method:**
   While simple averaging was effective, more sophisticated techniques like weighted averaging or stacked regression may yield further performance improvements.

6.4 Implications for Health Informatics

The results of this study have far-reaching implications for the field of health informatics, offering a scalable and non-invasive solution for predicting body fat percentage. Practical applications of the hybrid model include:

1. **Clinical Integration:**
   Healthcare professionals can integrate the hybrid model into diagnostic tools to provide

rapid and accurate body fat assessments. By replacing costly methods such as DEXA or underwater weighing, this approach can improve accessibility and affordability, particularly in resource-constrained settings.

2. **Fitness Programs:**
   Mobile health (mHealth) applications and wearable devices can utilize the hybrid model to deliver real-time body fat predictions based on simple anthropometric inputs. Fitness enthusiasts and trainers can use these insights to design personalized weight management programs.
3. **Preventative Care:**
   By identifying individuals at risk of obesity-related conditions early, the hybrid model can support preventative healthcare strategies. Public health initiatives can leverage these predictions to target at-risk populations for early interventions, education, and lifestyle modifications.

6.5 Novel Contributions

This study makes several significant contributions to the literature on machine learning-based body composition assessment:

1. **Introduction of a Hybrid Model:**
   Unlike previous studies, this research combines the outputs of Decision Tree, Random Forest, and MLP models to achieve state-of-the-art accuracy in predicting body fat percentage.
2. **Robust Validation:**
   The use of 5-fold cross-validation ensures that the findings are reliable, robust, and generalizable across diverse datasets.
3. **Advanced Preprocessing:**
   By employing techniques such as mutual information gain, Spearman's correlation, and VIF analysis, the study enhances the selection of the most predictive features, reducing noise and improving model efficiency.

These contributions address existing gaps in model performance, validation, and feature relevance, paving the way for practical, real-world deployment of machine learning solutions in health informatics.

6.6 Challenges

Several challenges were encountered during the project, each of which was addressed to ensure the reliability of the results:

1. **Overfitting in DecisionTreeRegressor:** Decision Trees displayed overfitting tendencies, leading to poor generalization.
   - Pruning techniques and cross-validation were applied to balance model complexity and performance.

2. **MLP Regressor Hyperparameter Tuning:** Optimizing neural network parameters (e.g., learning rate, hidden layers) was computationally intensive.
    - The randomized search approach was used to identify the optimal parameters systematically.
3. **Dataset Size:** The dataset contained only 242 samples after preprocessing, limiting model generalizability.
    - Robust feature selection and 5-fold cross-validation ensured that the models achieved reliable and stable performance.

6.7 Future Work

Building on the findings and limitations of this study, several directions are proposed for future research. First, more sophisticated ensemble approaches, such as stacked regression and weighted averaging, can be explored to further enhance prediction accuracy. Second, validating the hybrid model on larger, multi-ethnic datasets is essential to ensure broader generalizability and real-world applicability. Incorporating additional features, such as physical activity levels, dietary habits, and genetic markers, could significantly improve the models' predictive power. To increase accessibility, a web-based or mobile application for real-time body fat estimation should be developed, enabling use in both clinical and fitness settings. Finally, techniques like SHAP (Shapley Additive Explanations) and LIME can be applied to improve model interpretability, allowing healthcare practitioners to better understand and trust the predictions.

6.8 Conclusion

This study successfully demonstrated the potential of machine learning techniques for predicting Body Fat Percentage (BFP) using anthropometric features. The Hybrid model emerged as the superior, achieving the best performance with an RMSE of 0.35 and an R^2 of 1.0, outperforming other individual models and hybrid model. Feature importance analysis revealed that density and abdomen circumference were the most significant predictors of BFP, aligning with findings in existing literature. This study provides a scalable and non-invasive solution for BFP prediction, offering valuable applications in clinical health assessment, fitness monitoring, and preventative care.

In conclusion, machine learning models, particularly hybrid approaches, offer a reliable and efficient alternative to traditional body composition assessment methods. Future research focusing on model scalability, real-world deployment, and interpretability will further enhance the impact of this work on public health and healthcare outcomes.

---

7. References

- **Breiman, L. (1984).** Classification and Regression Trees. Chapman & Hall/CRC.
- **Breiman, L. (2001).** Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
- **Choi, J., Kim, Y., & Lee, S. (2020).** Estimation of body fat percentage using anthropometric data and machine learning. *IEEE Transactions on Biomedical Engineering, 67*(3), 897–904.
- **Cortes, C., & Vapnik, V. (1995).** Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
- **Thomas, R.N. and Gupta, R. (2020)** Feature Selection Techniques and Its Importance in Machine Learning: A Survey. 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science, Bhopal, 22-23 February 2020, 1-6. https://doi.org/10.1109/sceecs48394.2020.189
- **Fan, Z., Chiong, R., Hu, Z., Keivanian, F., & Chiong, F. (2022).** Body fat prediction through feature extraction based on anthropometric and laboratory measurements. PloS one, 17(2), e0263333. https://doi.org/10.1371/journal.pone.0263333
- **Friedman, J.H. (2001).** Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
- **Haykin, S. (2009).** Neural Networks and Learning Machines (3rd ed.). Upper Saddle River, NJ: Pearson.
- **Kingma, D.P., & Ba, J. (2015).** Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR).*
- **Kuhn, M., & Johnson, K. (2013).** Applied Predictive Modeling. New York: Springer.
- **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011).** Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
- **Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986).** Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0
- **Spearman, C. (1904).** The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.
- **Hussain, S. A., Cavus, N., & Sekeroglu, B. (2021).** Hybrid Machine Learning Model for Body Fat Percentage Prediction Based on Support Vector Regression and Emotional Artificial Neural Networks. Applied Sciences, 11(21), 9797. https://doi.org/10.3390/app11219797
- **Uçar, M.K., Yıldız, B., & Yıldırım, S. (2021).** A comparative study of machine learning methods for body fat percentage estimation. *Measurement, 167*, 108173. https://doi.org/10.1016/j.measurement.2020.108173
- **Friedman, J.H. (2002).** Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
- **Zhang, Y., Wang, S., & Wang, Z. (2019).** A deep learning approach to predict body fat percentage using anthropometric data. *IEEE Access, 7*, 53876–53884.
- **Friedman, J.H., Hastie, T., & Tibshirani, R. (2008).** Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28*(2), 337–374.

- **Accurso, L., Castonguay, T., Fortin, M., DeMont, R., Dover, G., & Dover, G. (2024).** Less Total-Body Fat and Lower-Extremity Fat Are Associated with More High-Intensity Running during Games in Female University Soccer Players. Applied Sciences, 14(19), 8992.
- **Fedesoriano (2022)** *Body Fat Prediction Dataset.* Available at: https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset (Accessed: 1 January 2025).