**Optimizing Body Fat Prediction: A Comparative Analysis of Regression Models**

**Abstract**

This report investigates the application of machine learning (ML) regression models to predict body fat percentage, offering a cost-effective alternative to traditional, expensive methods of body fat assessment. Traditional techniques, while accurate, are not accessible to the wider population due to their high costs, underscoring the need for innovative solutions. The advent of ML technologies, particularly regression models, provides promising approaches by analysing readily available body measurements, such as abdomen, chest, hip, thigh sizes, and weight. This study not only represents a technological advancement but also marks a crucial evolution in health analytics. By enabling accurate, non-invasive, and affordable estimation of body fat percentage, ML models can facilitate early preventative measures against obesity-related comorbidities, aligning with public health objectives to enhance preventive healthcare access and mitigate high body fat-associated diseases. This research explores the effectiveness of five prominent regression algorithms, evaluating their performance based on metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). The goal is to identify the most precise model for practical use, thereby contributing to the broader efforts in combating obesity and associated health risks with innovative, accessible technologies.

**Introduction**

The global obesity pandemic presents one of the most daunting public health challenges of the 21st century, affecting individuals across all age groups and demographics. Recent statistics reveal a troubling surge in obesity rates worldwide, with over 650 million adults classified as obese, which accounts for approximately 13% of the global population. This figure is alarming, not only because of its magnitude but also due to the myriad health complications associated with obesity, which significantly diminish the quality of life and lifespan of affected individuals. Obesity is a complex condition with multifaceted effects on human health, acting as a catalyst for various comorbidities, including type 2 diabetes, cardiovascular diseases, certain types of cancer, and respiratory disorders. These health issues contribute to an increased mortality rate among obese individuals, with studies indicating that obesity can reduce life expectancy by up to 10 years compared to those with a healthy body weight. The reduction in lifespan is directly correlated with the severity of obesity and the onset of related health conditions, which often result in premature death. Moreover, the economic impact of the obesity pandemic cannot be overstated, with healthcare systems worldwide grappling with the escalating costs of treating obesity-related conditions. This economic burden underscores the urgent need for effective preventive measures and treatment strategies that can mitigate the impact of obesity on individuals and society at large.

In light of these challenges, the exploration of machine learning regression models to predict and manage body fat percentage offers a beacon of hope. By providing an accessible and cost-effective tool for early detection and intervention, machine learning technologies have the potential to revolutionize the approach to obesity management. This advancement could lead to significant improvements in public health outcomes, reducing the prevalence of obesity and its associated health risks. Ultimately, the successful integration of these technologies in the fight against the obesity pandemic could pave the way for a healthier future for populations worldwide.

**Methods**

**Data acquirement**

The acquisition of data is a critical initial step in the analytical process, serving as the foundation upon which subsequent analyses are built. For the project at hand, data was sourced from a dataset titled "Body Fat Prediction," which is hosted on the Kaggle platform. This dataset was published by Roger W. Johnson from the Department of Mathematics & Computer Science at the South Dakota School of Mines & Technology. It comprises a collection of key features and parameters that are instrumental in estimating body fat percentage, a vital metric in assessing human health and fitness levels.

The dataset is notable for its high quality, as it has undergone thorough pre-processing by the contributor. This pre-processing involved the removal of null values, the correction of any missing data, and the elimination of duplicated rows. Such meticulous data cleaning is advantageous as it simplifies the initial stages of the machine learning workflow, allowing for a more straightforward and efficient analysis. The absence of data anomalies such as null or duplicated rows ensures the integrity and reliability of the dataset, thereby facilitating a more accurate and robust predictive modelling process.

**Exploratory Data Analysis**

Following the acquisition of the dataset, an Exploratory Data Analysis (EDA) was conducted to scrutinize the data for integrity and to gain insights into its characteristics and distribution. The dataset was imported into pandas data frame. This step was crucial for ensuring that the dataset was free from discrepancies that could potentially skew the analysis.

A descriptive analysis was performed to provide a comprehensive overview of the dataset's attributes, including measures of central tendency and variability. This analysis facilitated an understanding of the general behaviour of the data and its distribution patterns. It was observed that the data exhibited a distribution that closely approximated normality, as evidenced by Kernel Density Estimate (KDE) histograms. These histograms were instrumental in visualizing the distribution of each feature within the dataset.

To identify potential outliers, box plots were generated for each feature. Outliers are extreme values that deviate significantly from the majority of the data, and their presence can influence the results of subsequent analyses. Identifying these outliers is crucial for deciding whether they should be retained, adjusted, or removed to improve model performance.

Additionally, to ascertain the factors contributing to body fat estimation, an extensive literature review was conducted alongside a correlation analysis of the features. The correlation analysis revealed the interrelationships among the features, which were visualized through a heatmap. This heatmap not only highlighted the degree of correlation between each pair of features but also shed light on multicollinearity within the dataset—a condition where predictors are highly correlated with each other, which can affect the predictive model's performance.

**Feature Engineering**

Upon completing the Exploratory Data Analysis (EDA), the next phase involved strategic Feature Engineering to optimize the dataset for predictive modelling. A crucial step in this process was the establishment of a stringent threshold for the correlation coefficient. Features that did not meet this predetermined correlation threshold (Corr < 0.5) were considered less relevant for the prediction of body fat and, consequently, were excluded from the dataset. This approach ensured that only features with significant predictive power were retained, thereby enhancing the model's accuracy.

Further refining the dataset, an Interquartile Range (IQR) method was employed to identify and remove outliers. Outliers can disproportionately affect the model's performance, and their removal is essential for improving the predictability of the outcomes. Additionally, features exhibiting skewness were evaluated. Skewness in data distribution can lead to biases in model prediction. Thus, features that were both skewed and exhibited low correlation with the target variable were dropped from the analysis. This meticulous process of Feature Engineering aimed at streamlining the dataset, ensuring that it was primed for the development of an accurate and robust predictive model.

**Model training and Evaluation**

The model training and evaluation phase constituted a comprehensive approach to identify the most effective predictive model for estimating body fat percentage. A diverse set of regression models, including Linear Regressor, Ridge Regressor, Support Vector Regressor, Random Forest Regressor, and Decision Tree Regressor, were chosen for this purpose. The selection encompassed models of varying complexity to facilitate a thorough comparison, thereby enabling the identification of the model that best balances prediction accuracy with computational efficiency.

The dataset was divided into training and testing subsets using a 70:30 ratio, ensuring a substantial amount of data for both model training and evaluation phases. To mitigate the risk of data leakage and prevent the models from memorizing specific data points, the data was processed through a pipeline. This structured approach to data handling is critical for maintaining the integrity of the evaluation process.

Model performance was rigorously assessed using K-fold cross-validation, a technique that divides the data into 'K' subsets (K=5, commonly used) to evaluate the models multiple times, each time with a different subset as the testing set. This method enhances the reliability of the performance assessment. Evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R²) value, were employed to quantify the models' predictive capabilities. These metrics provide insights into the accuracy, consistency, and explanatory power of the models.

The model demonstrating superior performance across these metrics was selected as the optimal solution. For the sake of transparency and reproducibility, the code and model weights are included alongside this report, ensuring that the results can be independently verified and applied in future research endeavours.

**Observations and Results**

The analytical process undertaken in this study illuminated several key insights regarding the dataset and the predictive models applied. Initially, descriptive analysis and exploratory data analysis (EDA) unveiled that the dataset closely resembled a normal distribution (Fig1). However, it was noted that certain features, including weight, height, knee size, ankle size, and biceps size, exhibited a degree of skewness.

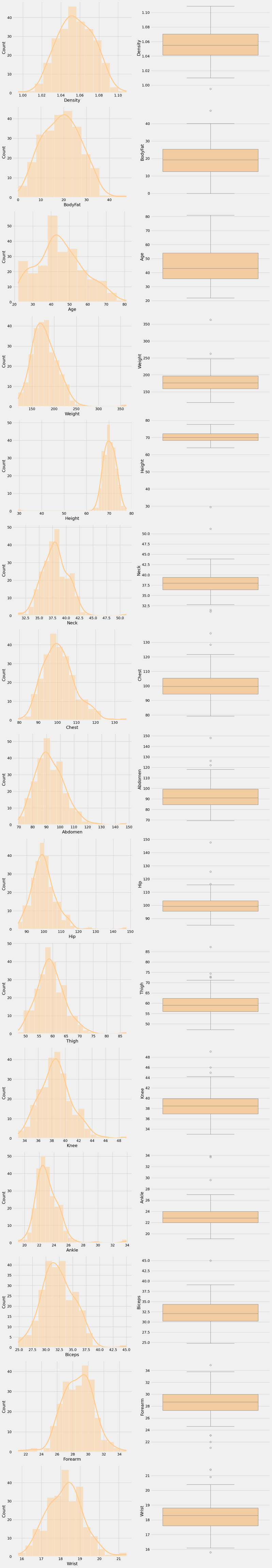


Fig1. Histogram and Boxplots representing the distribution of the partial dataset (complete figure in supplementary)

Additionally, the correlation heatmap (fig 2) indicated the presence of multicollinearity among the variables, leading to the implementation of a 0.5 coefficient correlation threshold (fig 3). This threshold facilitated the elimination of features that did not meet the specified criteria, streamlining the dataset for more effective analysis.

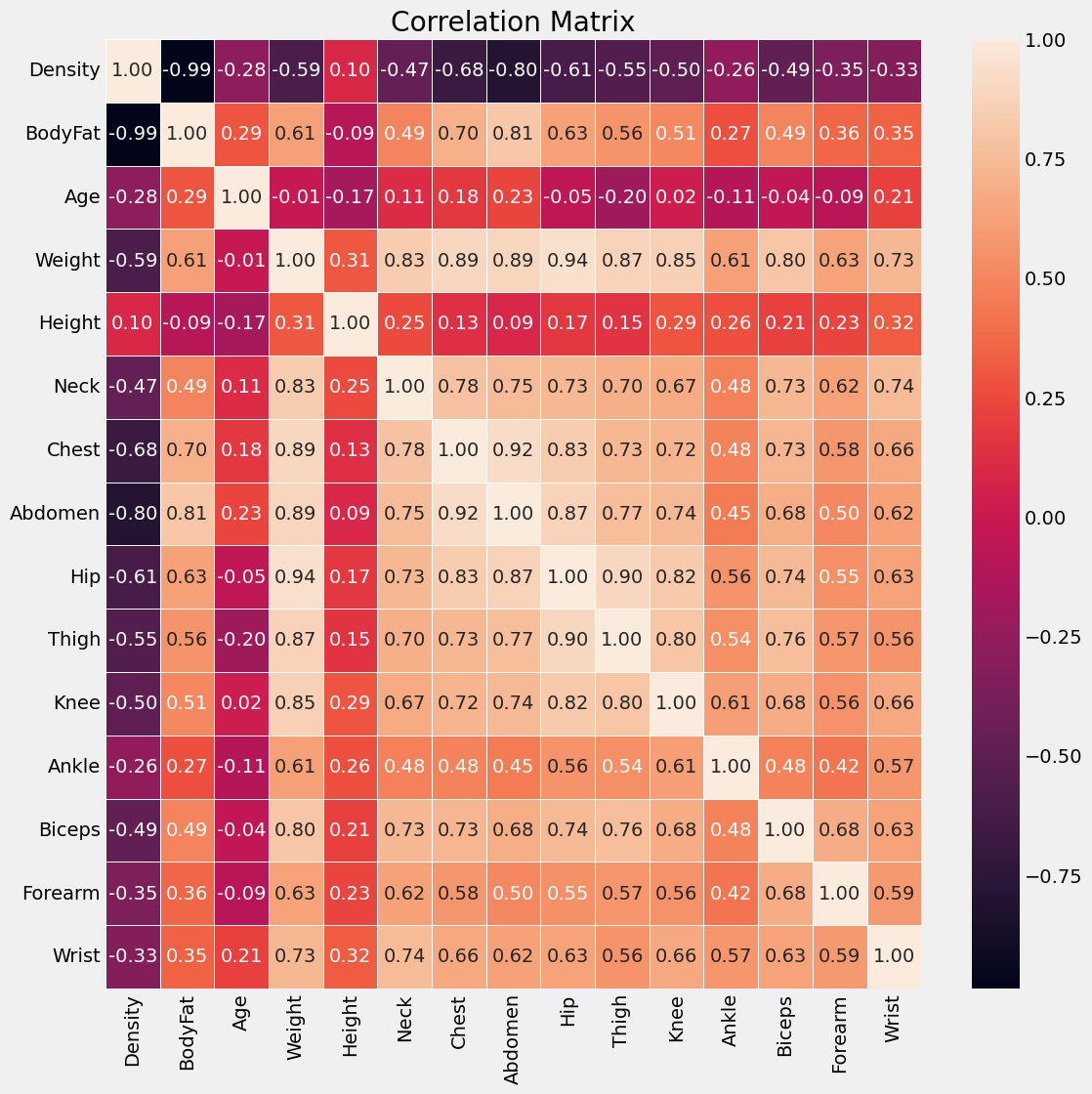


Fig 2. Correlation Matrix of each feature against each other

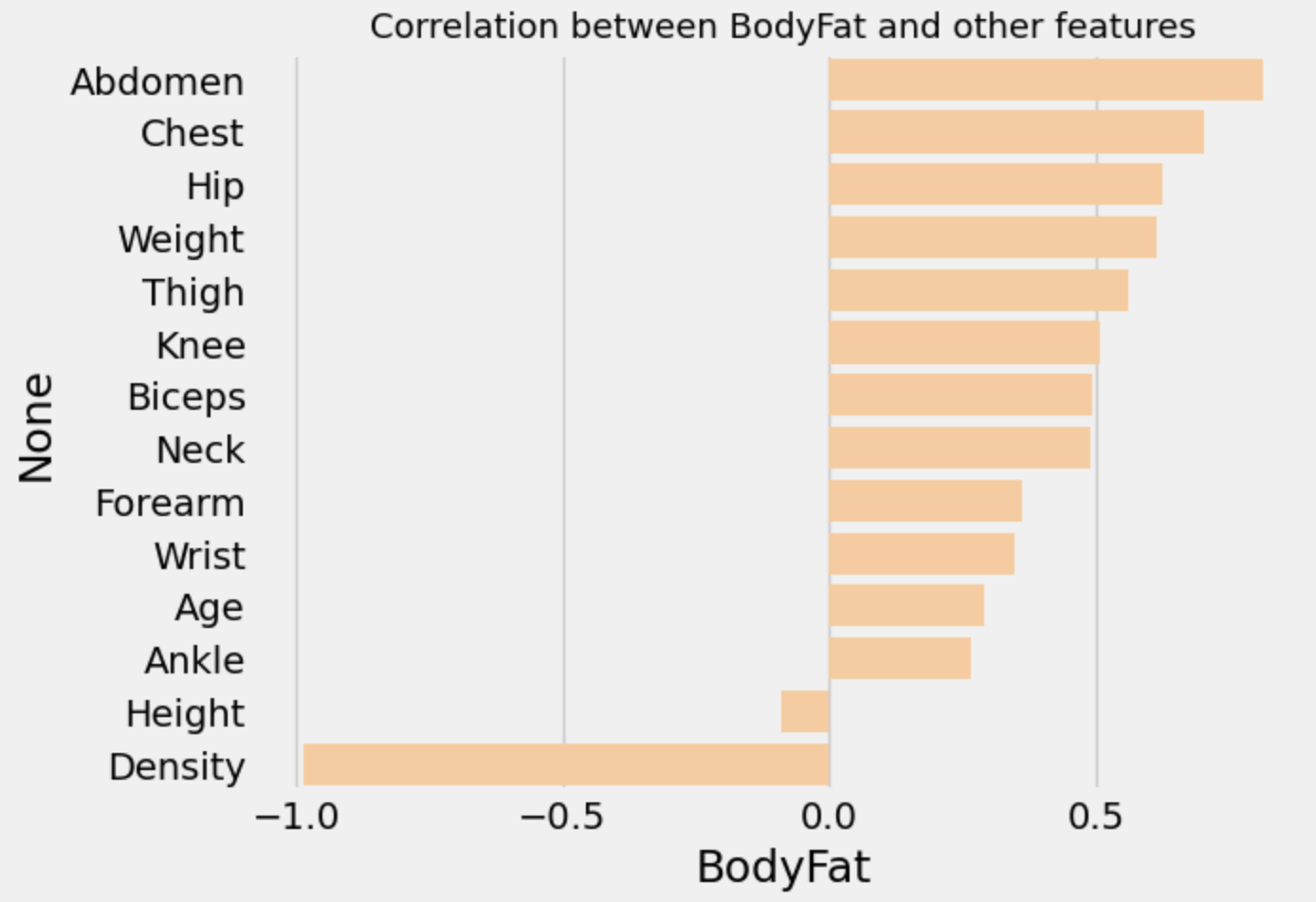
****

Fig 3. Correlation plot of Body Fat against other features.

Subsequent outlier detection, utilizing the Interquartile Range (IQR) method, further refined the dataset, preparing it for the model training phase. A selection of regression models was then evaluated, with the Support Vector Regression (SVR) model showing notably higher Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values, indicating a less favourable performance compared to other models (Table 1). Conversely, the Random Forest Regression model demonstrated the lowest MAE, MSE, and RMSE values, coupled with the highest R-squared (R²) value, signifying its superior ability to predict body fat percentage accurately.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | **MAE** | **MSE** | **RMSE** | **R2** | **Adjusted R2** |
| **1** | Linear Regression | 0.482500 | 1.863811 | 1.123463 | 0.968007 | 0.965142 |
| **2** | Support Vector Regression | 1.323009 | 6.313004 | 2.405135 | 0.907019 | 0.898692 |
| **3** | Ridge Regression | 0.517737 | 1.848089 | 1.137451 | 0.968387 | 0.965556 |
| **4** | Random Forest | 0.345478 | 1.582528 | 1.057788 | 0.973574 | 0.971208 |
| **5** | Decision Tree | 0.420449 | 2.399508 | 1.395142 | 0.957892 | 0.954121 |

Table 1. Comparison of metrics for each model trained

The comparative analysis of these metrics (fig.4) revealed that, although several models exhibited satisfactory performance, the Random Forest Regression model stood out due to its enhanced predictive accuracy and ability to account for a higher proportion of the variance within the dataset. This model's efficacy, particularly in medical-related predictions, underscores its applicability and robustness, making it the optimal choice for this specific dataset.

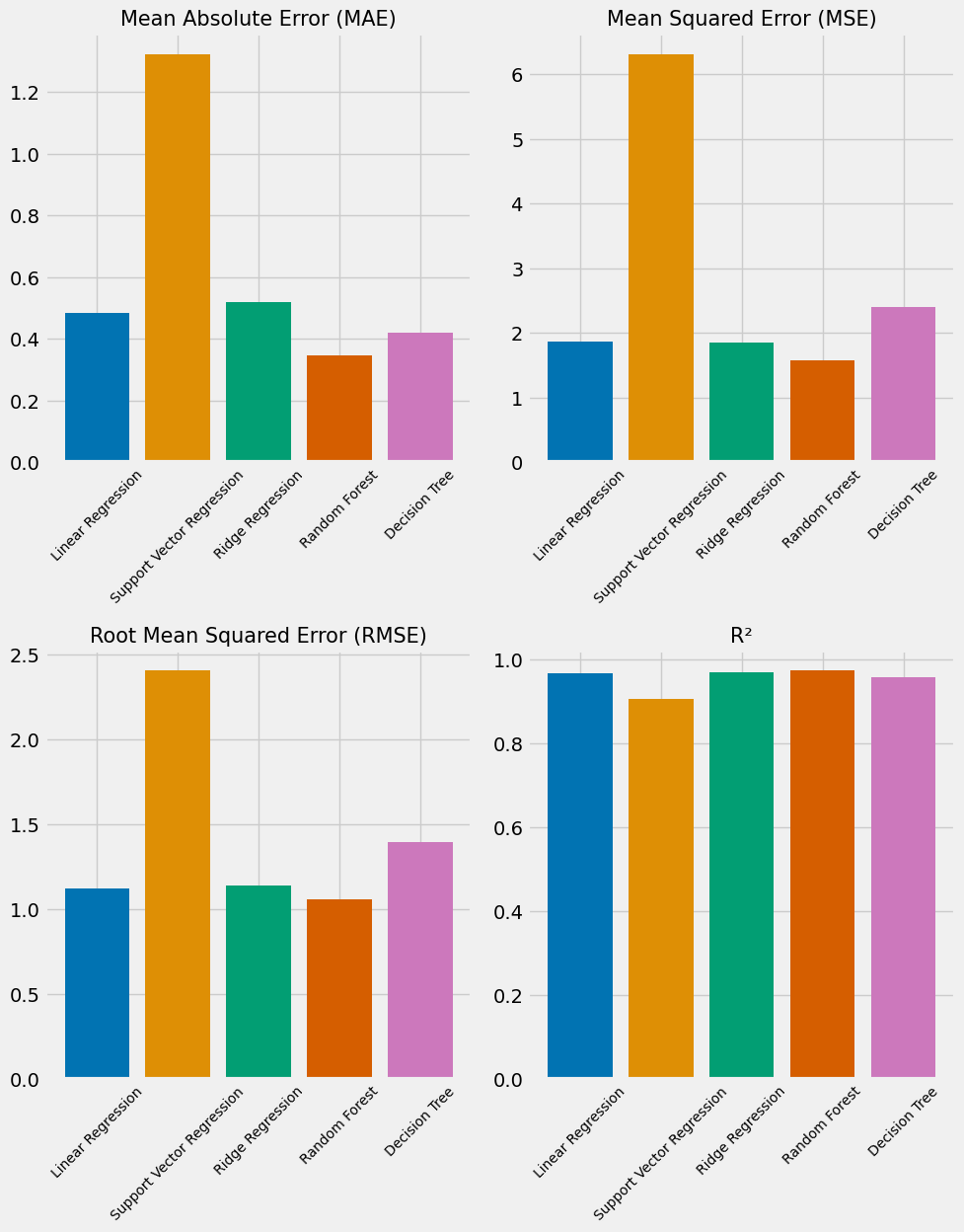


Fig 4. Comparison of MAE, MSE, RMSE and R2 for the 5 models trained.

**Discussions and Future Work**

The findings from this study offer significant insights into the predictive modelling of body fat percentage, highlighting the importance of rigorous data pre-processing and model selection. Initial exploratory data analysis (EDA) and descriptive statistics underscored the dataset's approximation to a normal distribution while identifying skewness in specific

features, necessitating careful feature engineering. The implementation of a 0.5 correlation coefficient threshold and the Interquartile Range (IQR) method for outlier removal were pivotal in refining the dataset, ensuring that only the most relevant and statistically sound features were retained for modelling.

Among the evaluated models, the Random Forest Regressor emerged as the most suitable for this dataset, outperforming others in terms of predictive accuracy as measured by MAE, MSE, RMSE, and R² metrics. This superiority can be attributed to the model's ability to handle multicollinearity and its robustness against overfitting, characteristics that are particularly valuable in the context of medical predictions. The study's methodology, from data acquisition through to model evaluation, adheres to stringent analytical standards, providing a replicable framework for future research.

Furthermore, the study’s approach to feature selection, based on correlation thresholds and the identification of outliers, highlights the critical role of data quality in developing effective predictive models. The success of the Random Forest Regressor in this context not only underscores its suitability for complex datasets but also its potential applicability in other domains requiring precise and reliable predictions.

**Supplementary figure** – describing the distribution of the remaining features.

