**ARTIFICIAL INTELLIGENCE**

**PROJECT:**

**Regression Analysis**

**Estimating the Body Fat Percentage based on a number of Features**

**Student:                                                                Vesa Ibrahimi**

Contents

[**1. Introduction** 3](#_Toc180349950)

[**2. Data Description** 4](#_Toc180349951)

[**3. Identifying the Most Important Variables (Choosing Indicators)** 4](#_Toc180349952)

[**4. Predicting body fat percentage using Multiple Linear Regression and Random Forest method** 6](#_Toc180349953)

[**4.1 Multiple Linear Regression method** 7](#_Toc180349954)

[**4.2 Random forest method** 9](#_Toc180349955)

[**4.3 Comparison of the models** 11](#_Toc180349956)

[**5. Identifying the most important independent variables (X) in my regression model** 11](#_Toc180349957)

# **1. Introduction**

In this project, several indicators, including age, height, neck, abdomen, hip, forearm, and wrist measurements, were analyzed to estimate Body Fat Percentage (Y). Initially, multiple regression was applied to identify relationships between these variables. However, a Random Forest Regressor provided a better fit, as evidenced by a higher R-squared value (0.66) and lower mean squared error (15.63). Abdomen was identified as the most important predictor, contributing over 70% to the model’s performance, while height, wrist, and neck had moderate importance. Random Forest proved more effective than multiple regression in capturing the complex interactions between the variables, making it the preferred model for estimating Body Fat Percentage.

# **2. Data Description**

In this project, I worked with a dataset containing various body measurements aimed at estimating body fat percentage. The target variable, PercentBodyFat, represents the percentage of fat in an individual's body. To predict this, a set of independent variables were considered, including age, weight, height, and several body circumference measurements.

The independent variables comprised several key body measurements. The dataset included age, weight (in pounds), and height (in inches) as standard physical characteristics. In addition to these, there were 10 circumference measurements, which provided more detailed insight into different parts of the body. These measurements included neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist circumferences, all measured in centimeters. By analyzing these indicators, I aimed to uncover which ones were most useful in predicting body fat percentage.

I checked for any missing values in the dataset. Handling missing data is a critical step, as gaps in the dataset can lead to biased or inaccurate results. Fortunately, the dataset did not contain any missing values, so I was able to proceed without needing to employ techniques such as imputation or the removal of incomplete rows. This ensured that I could work with the full dataset, capturing the variability present in each indicator.

My primary focus was on understanding the relationships between these variables and their impact on body fat percentage. I aimed to identify the key predictors of body fat percentage and develop a model to estimate it accurately.

# **3. Identifying the Most Important Variables (Choosing Indicators)**

In response to the question, "Which indicators (independent variables) provide a reliable estimation of the Body Fat Percentage?", I conducted a multiple regression analysis using OLS (Ordinary Least Squares) to identify which variables have a statistically significant impact on predicting body fat percentage.

import pandas as pd

import statsmodels.api as sm

# Loaded the dataset

file\_path = '/content/drive/MyDrive/PercentBodyFat.xlsx'

df = pd.read\_excel(file\_path)

# Defined the independent variables (X) and the dependent variable (y)

X = df[['Age', 'Weight', 'Height', 'Neck', 'Chest', 'Abdomen', 'Hip', 'Thigh', 'Knee', 'Ankle', 'Biceps', 'Forearm', 'Wrist']]

y = df['PercentBodyFat']

# Added a constant (intercept) to the model

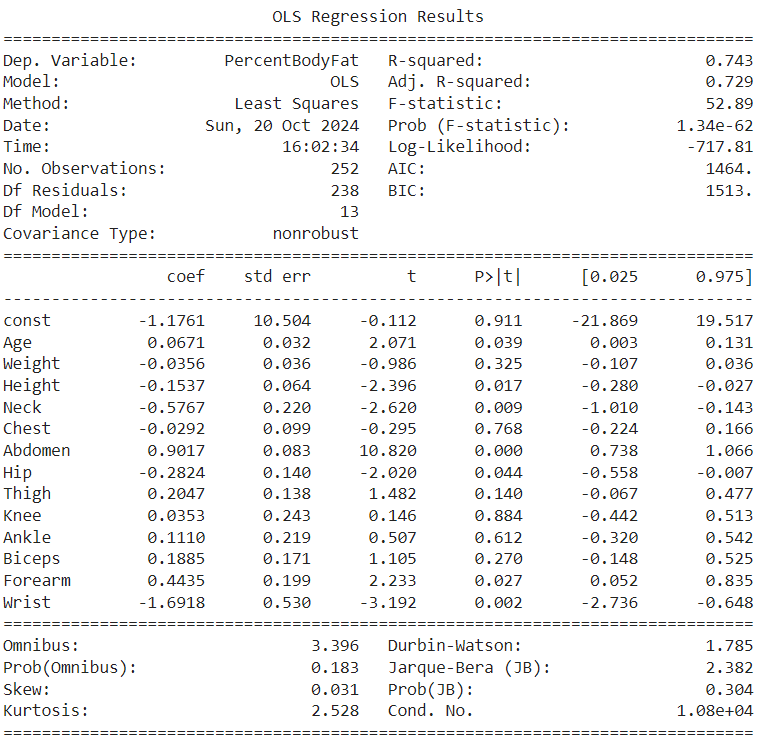
X = sm.add\_constant(X)

# Fited the OLS model

model = sm.OLS(y, X).fit()

# Printed the summary of the regression results

print(model.summary())



Based on the results, several key indicators emerged as reliable predictors of body fat percentage, determined by their p-values and coefficients. Variables with p-values less than 0.05 are considered statistically significant, meaning they have a strong relationship with body fat percentage.

**Significant Variables (p-value < 0.05):**

1. **Age** (Coefficient = 0.0671, p-value = 0.039):  
   Age shows a positive relationship with body fat percentage. As age increases, body fat percentage tends to increase slightly. This makes age a significant and reliable predictor, though its impact is modest.
2. **Height** (Coefficient = -0.1537, p-value = 0.017):  
   Height is negatively correlated with body fat percentage, suggesting that taller individuals generally have lower body fat percentages. This makes height a reliable indicator, with a statistically significant inverse relationship.
3. **Neck Circumference** (Coefficient = -0.5767, p-value = 0.009):  
   Neck circumference shows a negative association with body fat percentage, meaning larger neck circumferences are associated with lower body fat. This variable is statistically significant and provides a reliable estimation.
4. **Abdomen Circumference** (Coefficient = 0.9017, p-value < 0.001):  
   Abdomen circumference is the strongest predictor in the model. A larger abdomen circumference is strongly linked to higher body fat percentage, making it the most reliable and impactful indicator among all the variables.
5. **Hip Circumference** (Coefficient = -0.2824, p-value = 0.044):  
   Hip circumference has a negative relationship with body fat percentage. Although its effect is less pronounced than the abdomen's, it is still statistically significant and contributes to a reliable estimation of body fat.
6. **Forearm Circumference** (Coefficient = 0.4435, p-value = 0.027):  
   The forearm circumference has a positive association with body fat percentage. It shows a moderate relationship, making it another statistically significant and reliable predictor.
7. **Wrist Circumference** (Coefficient = -1.6918, p-value = 0.002):  
   Wrist circumference exhibits a strong negative correlation with body fat percentage. The smaller the wrist circumference, the higher the body fat percentage, making it a highly reliable predictor in the model.

Several variables, including **Weight**, **Chest**, **Thigh**, **Knee**, **Ankle**, and **Biceps**, did not show statistically significant relationships with body fat percentage in this model, as their p-values were greater than 0.05. While they may have some influence on body fat percentage, they are not reliable predictors based on this analysis.

# **4. Predicting body fat percentage using Multiple Linear Regression and Random Forest method**

The aim of this analysis of mine is to develop a model that provides the best possible fit between **Body Fat Percentage (Y)** and a set of significant predictors (X), which include **f**rom my previous analysis, the variables that have been identified as significant predictors (based on p-values and R-squared changes): **Age, Height, Neck, Abdomen (most important), Hip, Forearm, Wrist**.

The primary goal is to assess which model, **Multiple Linear Regression (MLR)** or **Random Forest Regression (RF)**, performs better in accurately predicting body fat percentage based on these variables.

Given this task, I approached the problem by first applying the classical method of **Multiple Linear Regression**. This method provides an easily interpretable model where each predictor has an associated coefficient, reflecting how changes in the predictor affect the dependent variable, in this case, body fat percentage. Following that, I implemented **Random Forest Regression**, which is a non-linear model capable of capturing more complex relationships between the predictors and the outcome. After fitting both models, I compared their performances based on key metrics like **Mean Squared Error (MSE)** and **R-squared (R²)** values to determine which approach was more suitable for this prediction problem.

## **4.1 Multiple Linear Regression method**

The regression equation can be written as:

*Y=β0+β1⋅Age+β2⋅Height+β3⋅Neck+β4⋅Abdomen+β5⋅Hip+β6⋅Forearm+β7⋅Wrist*

I computed the **Mean Squared Error (MSE)** and **R-squared** values to evaluate the model's performance.

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from sklearn import linear\_model

from sklearn.metrics import mean\_squared\_error, r2\_score

from sklearn.model\_selection import train\_test\_split

# Loadede the dataset (replace with your dataset path)

data = pd.read\_excel('/content/drive/MyDrive/PercentBodyFat.xlsx')

# Dependent variable: Body Fat Percentage (Y)

Y = data['PercentBodyFat']

# Independent variables (X): Age, Height, Neck, Abdomen, Hip, Forearm, Wrist

X = data[['Age', 'Height', 'Neck', 'Abdomen', 'Hip', 'Forearm', 'Wrist']]

# Splited the data into training/testing sets

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=42)

# Initialized the linear regression model

regr = linear\_model.LinearRegression()

# Trained the model using the training sets

regr.fit(X\_train, Y\_train)

# Make predictions using the testing set

Y\_pred = regr.predict(X\_test)

# The coefficients

print("Coefficients: \n", regr.coef\_)

# The mean squared error

print("Mean squared error: %.2f" % mean\_squared\_error(Y\_test, Y\_pred))

# The coefficient of determination (R-squared)

print("R-squared: %.2f" % r2\_score(Y\_test, Y\_pred))

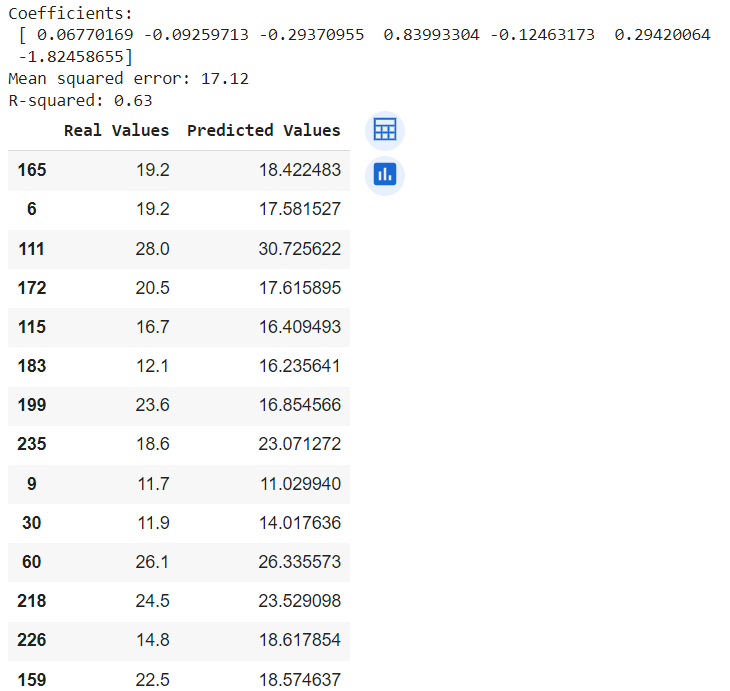
# Save the real and predicted values to a CSV

df = pd.DataFrame({'Real Values': Y\_test, 'Predicted Values': Y\_pred})

df.to\_csv("C:\\path\_to\_save\\BodyFatPredictionOutput.csv", encoding='utf-8', index=False)

# Show the results

df



The output of this model provided coefficients that reflected the relationship between each predictor and body fat percentage. For instance, **Abdomen** had the highest positive coefficient, indicating that as the abdominal measurement increases, body fat percentage increases significantly.

The results from MLR were as follows:

* **Mean Squared Error**: 17.12
* **R-squared**: 0.63

## **4.2 Random forest method**

For this analysis, I followed similar steps to MLR:

import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_squared\_error, r2\_score

from sklearn.model\_selection import train\_test\_split

# Loaded the dataset (replace with your dataset path)

data = pd.read\_excel('/content/drive/MyDrive/PercentBodyFat.xlsx')

# Dependent variable: Body Fat Percentage (Y)

Y = data['PercentBodyFat']

# Independent variables (X): Age, Height, Neck, Abdomen, Hip, Forearm, Wrist

X = data[['Age', 'Height', 'Neck', 'Abdomen', 'Hip', 'Forearm', 'Wrist']]

# Splited the data into training/testing sets

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=42)

# Initialize the Random Forest Regressor

rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)

# Trained the model using the training sets

rf\_model.fit(X\_train, Y\_train)

# Make predictions using the testing set

Y\_pred = rf\_model.predict(X\_test)

# The mean squared error

print("Mean squared error: %.2f" % mean\_squared\_error(Y\_test, Y\_pred))

# The coefficient of determination (R-squared)

print("R-squared: %.2f" % r2\_score(Y\_test, Y\_pred))

# Feature importance (which predictors are most important)

feature\_importances = rf\_model.feature\_importances\_

print("Feature Importances: ", feature\_importances)

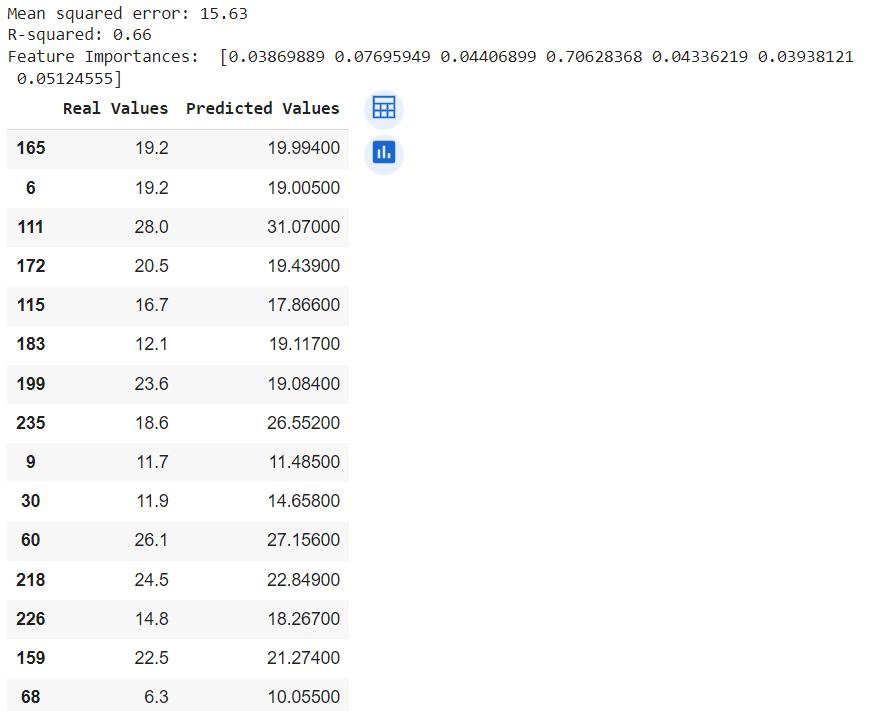
# Saved the real and predicted values to a CSV

df = pd.DataFrame({'Real Values': Y\_test, 'Predicted Values': Y\_pred})

df.to\_csv("C:\\path\_to\_save\\RandomForest\_BodyFatPrediction.csv", encoding='utf-8', index=False)

# Show the results

df



The Random Forest model yielded the following results:

* **Mean Squared Error**: 15.63
* **R-squared**: 0.66

In addition, the **feature importance** showed that **Abdomen** was by far the most important predictor, with a feature importance score of 0.71, confirming its strong influence on body fat percentage.

## **4.3 Comparison of the models**

When comparing both models, it is clear that **Random Forest** outperformed **Multiple Linear Regression** in this case:

* **Mean Squared Error**: Random Forest had a lower MSE (15.63) compared to MLR (17.12), indicating that Random Forest made more accurate predictions on average.
* **R-squared**: Random Forest explained a higher proportion of the variance in body fat percentage (R² = 0.66) compared to MLR (R² = 0.63).

# **5. Identifying the most important independent variables (X) in my regression model**

import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_squared\_error, r2\_score

from sklearn.model\_selection import train\_test\_split

import numpy as np

# Loaded the dataset

file\_path = '/content/drive/MyDrive/PercentBodyFat.xlsx'

data = pd.read\_excel(file\_path)

# Dependent variable: Body Fat Percentage (Y)

Y = data['PercentBodyFat']

# Independent variables (X): Age, Height, Neck, Abdomen, Hip, Forearm, Wrist

X = data[['Age', 'Height', 'Neck', 'Abdomen', 'Hip', 'Forearm', 'Wrist']]

# Splited the data into training/testing sets

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=42)

# Initialize the Random Forest Regressor

rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)

# Trained the model using the training sets

rf\_model.fit(X\_train, Y\_train)

# Make predictions using the testing set

Y\_pred = rf\_model.predict(X\_test)

# The mean squared error

mse = mean\_squared\_error(Y\_test, Y\_pred)

print(f"Mean squared error: {mse:.2f}")

# The coefficient of determination (R-squared)

r\_squared = r2\_score(Y\_test, Y\_pred)

print(f"R-squared: {r\_squared:.2f}")

# Feature importance (which predictors are most important)

feature\_importances = rf\_model.feature\_importances\_

importance\_df = pd.DataFrame({

    'Feature': X.columns,

    'Importance': feature\_importances

}).sort\_values(by='Importance', ascending=False)

print(importance\_df)

# Visualize the feature importances

import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))

plt.barh(importance\_df['Feature'], importance\_df['Importance'], color='skyblue')

plt.xlabel('Importance')

plt.ylabel('Feature')

plt.title('Feature Importance in Random Forest Model')

plt.gca().invert\_yaxis()

plt.show()

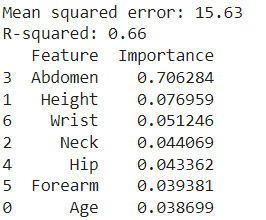
# Saved real and predicted values to CSV for further analysis

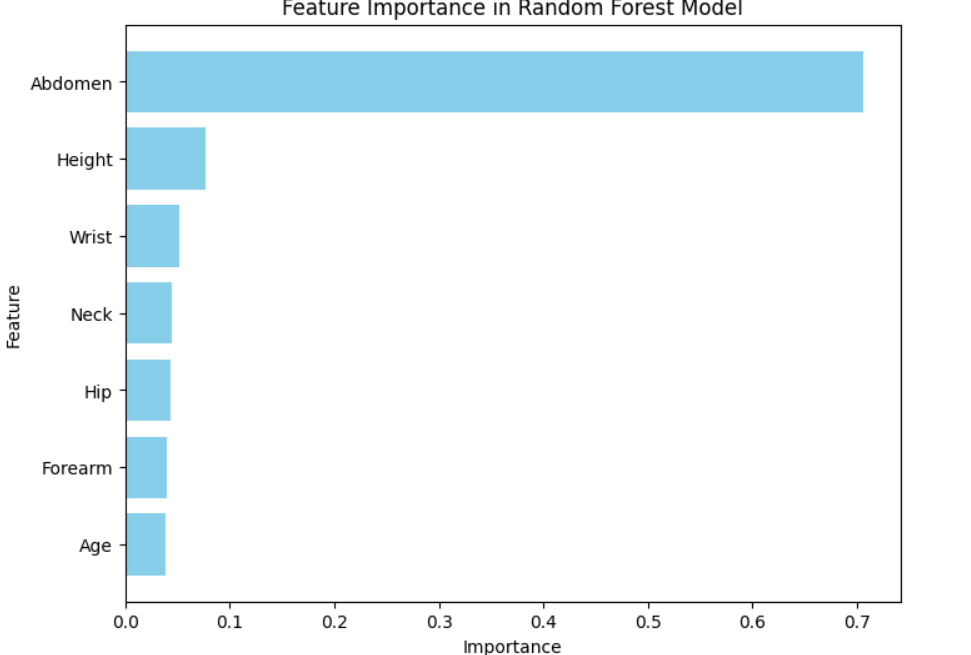
results\_df = pd.DataFrame({'Real Values': Y\_test, 'Predicted Values': Y\_pred})

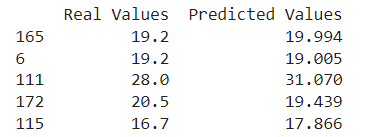
results\_df.to\_csv("C:\\path\_to\_save\\RandomForest\_BodyFatPrediction.csv", encoding='utf-8', index=False)

# Show the results

print(results\_df.head())







Conclusion: **Abdomen (0.7063):** This feature is by far the most important predictor of body fat percentage, contributing over 70% of the model's predictive power.