FACULTY OF MANAGEMENT

TDS2101

DATA SCIENCE FUNDAMENTALS

Project Report

T1 2023/2024

Body Fat Extended Dataset

Lecturer - Nur Syuhaidah Binti Nor Azni

Name	**ABDELKERIM ALI HASSAN**
ID	**1211302792**

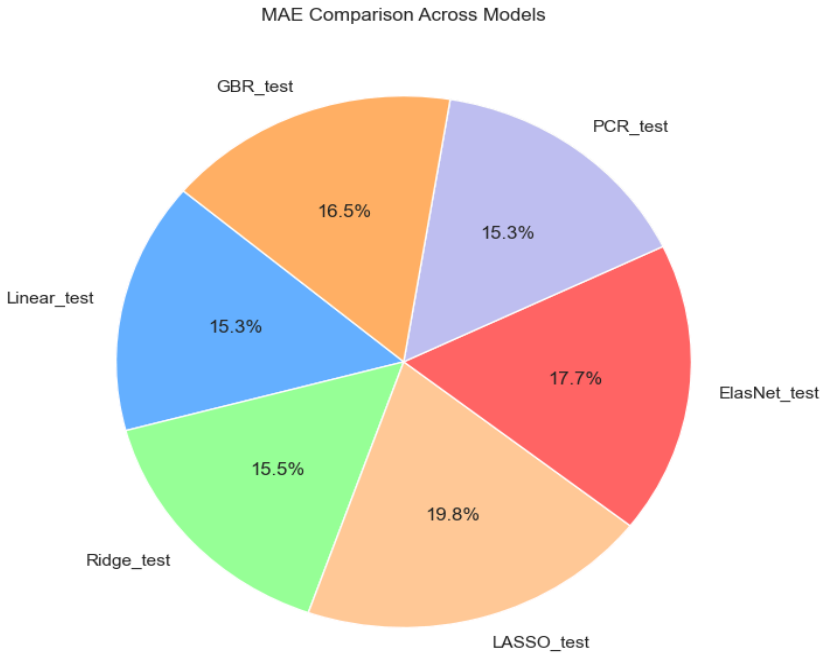
TABLE OF CONTENTS

1. Introduction:

The "Body Fat Extended Dataset" project has set out to revolutionize body fat percentage estimation by utilizing machine learning algorithms. This section gives an in-depth introduction, outlining the objectives, background, and significance of the project in the health and fitness domain. The introduction serves as a roadmap for readers to understand the overarching goals and motivations behind the project.

1.1 Problem Statement:

Accurately predicting body fat percentage through non-invasive measurements is crucial for health assessment and fitness planning. However, traditional methods are often costly and intrusive, highlighting the need for a more accessible and reliable alternative. This section explores the challenges associated with current methods and emphasizes the project's goal of developing a convenient and trustworthy means of estimating body fat percentage.



1.2 Executive Summary:

The executive summary is a brief summary of the project designed for stakeholders and decision-makers. It summarizes the project's objectives, approach, and potential impact, providing a quick overview of its significance. In this project, we have developed a practical and precise method for estimating body fat percentage.

2. Data Description:

The data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use it for non-commercial purposes.
The size of the data is 33.8 KB
with 252 males measures and 184 females

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 436 entries, 0 to 435
Data columns (total 16 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   BodyFat   436 non-null    float64
 1   Original  436 non-null    object
 2   Sex       436 non-null    object
 3   Age       436 non-null    int64
 4   Weight    436 non-null    float64
 5   Height    436 non-null    float64
 6   Neck      436 non-null    float64
 7   Chest     436 non-null    float64
 8   Abdomen   436 non-null    float64
 9   Hip       436 non-null    float64
 10  Thigh     436 non-null    float64
 11  Knee      436 non-null    float64
 12  Ankle     436 non-null    float64
 13  Biceps    436 non-null    float64
 14  Forearm   436 non-null    float64
 15  Wrist     436 non-null    float64
dtypes: float64(13), int64(1), object(2)
memory usage: 54.6+ KB
```

3. Data Cleaning:

Data cleaning plays a crucial role in preparing a dataset for accurate analysis. In this section, we will discuss the methods and techniques used to handle issues such as imbalances, outliers, and missing values. This phase ensures that the dataset is clean and reliable, which sets the foundation for accurate modeling in later stages. I want to emphasize that there is no missing data in this dataset.

```
# Checking missing values.
print(df.isna().sum())
```

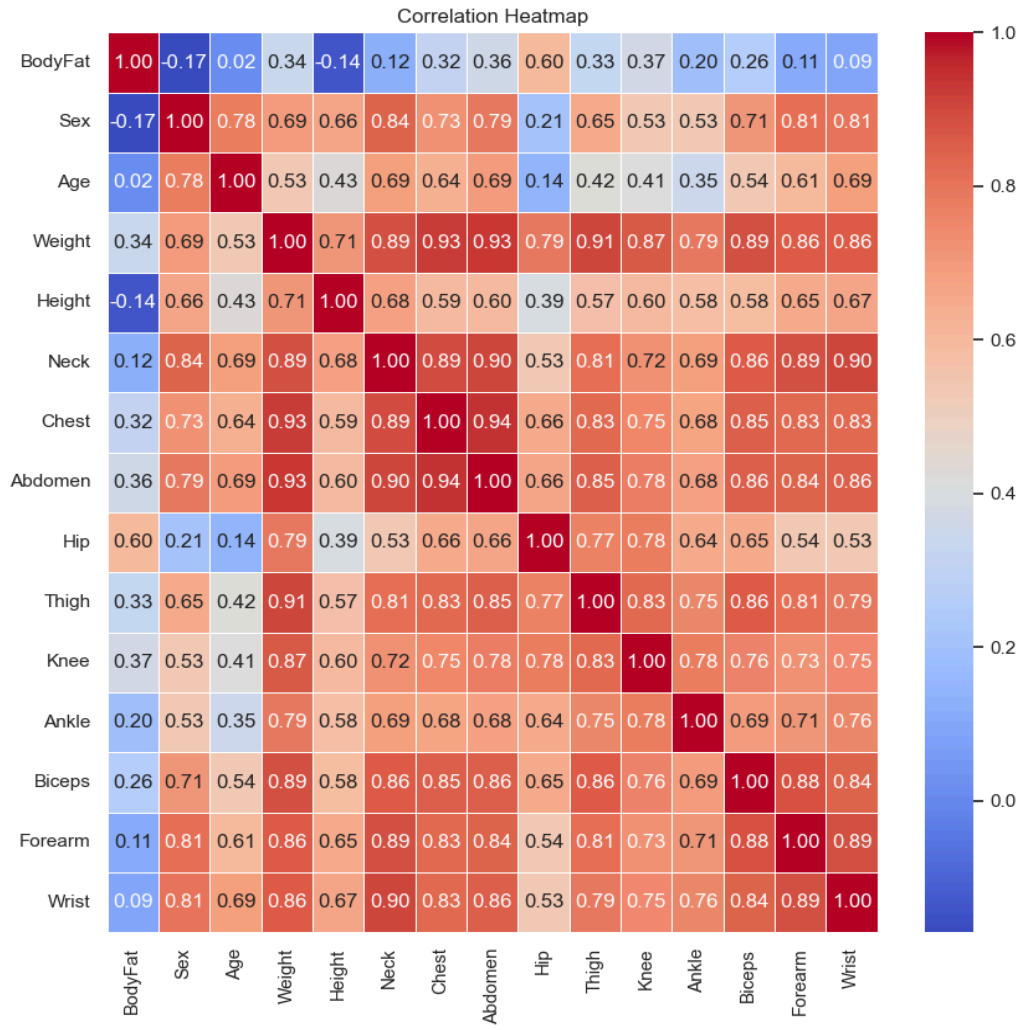
```
BodyFat     0
Original    0
Sex         0
Age         0
Weight      0
Height      0
Neck        0
Chest       0
Abdomen     0
Hip         0
Thigh       0
Knee        0
Ankle       0
Biceps      0
Forearm     0
Wrist       0
dtype: int64
```

4. Data Transformation:

To ensure the dataset is in a suitable format, it needs to undergo some transformations before moving on to model development. This process involves various techniques and procedures that will make the data more suitable for further analysis and model training. Proper data transformation is critical in achieving accurate and meaningful results.

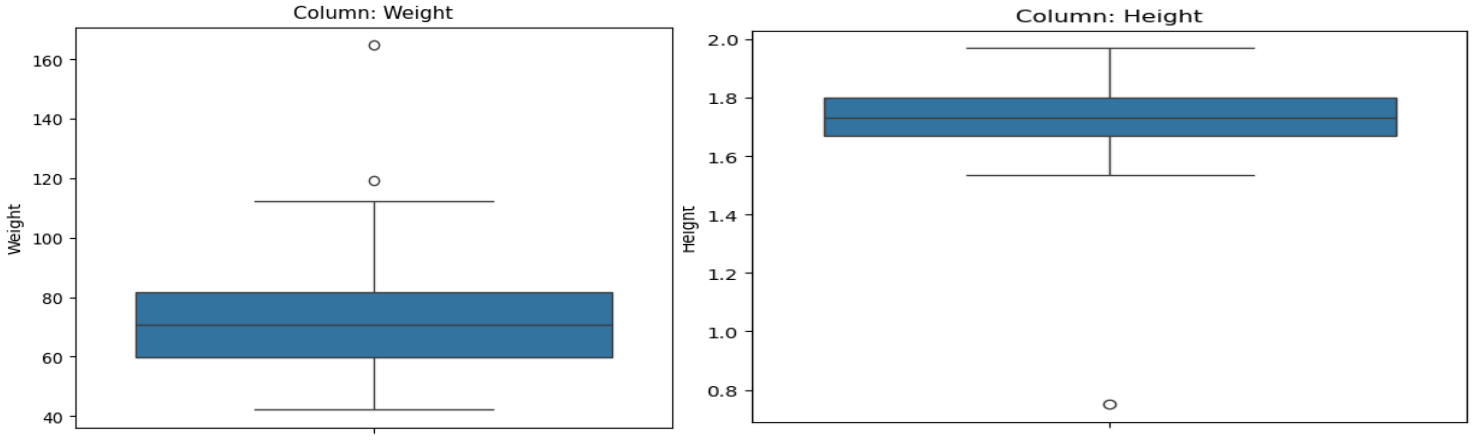
5. Exploratory Data Analysis:

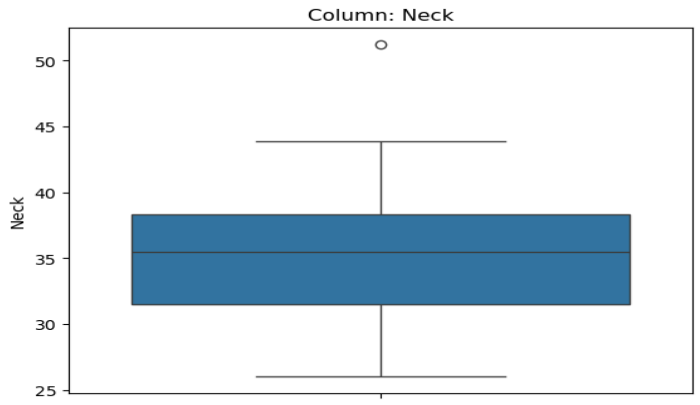
Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics of a dataset. In this phase, visualization tools are used to comprehend the distribution of body fat percentage and the relationships between different variables. This step lays the groundwork for subsequent modeling by uncovering patterns and trends within the dataset.Hence correlation heatmap is useful for identifying the relationships between variables and understanding potential multicollinearity

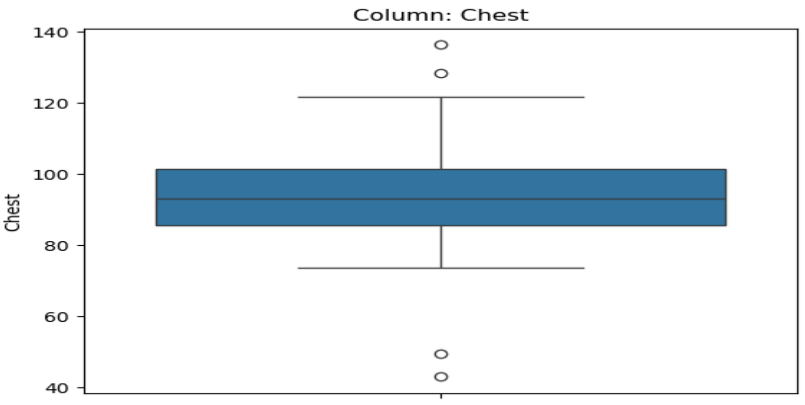


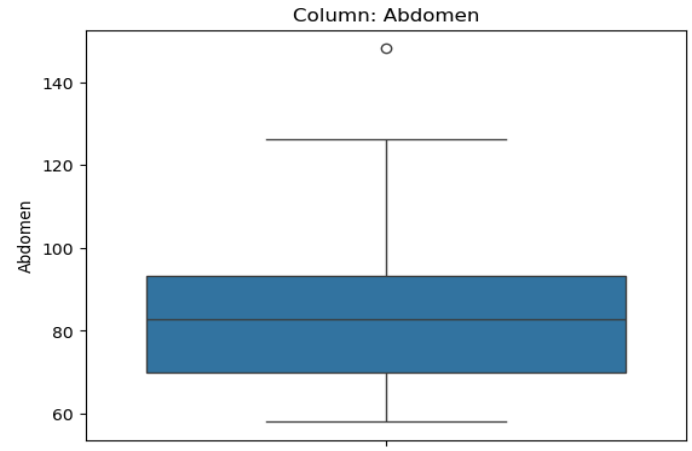
6. Data Mining:

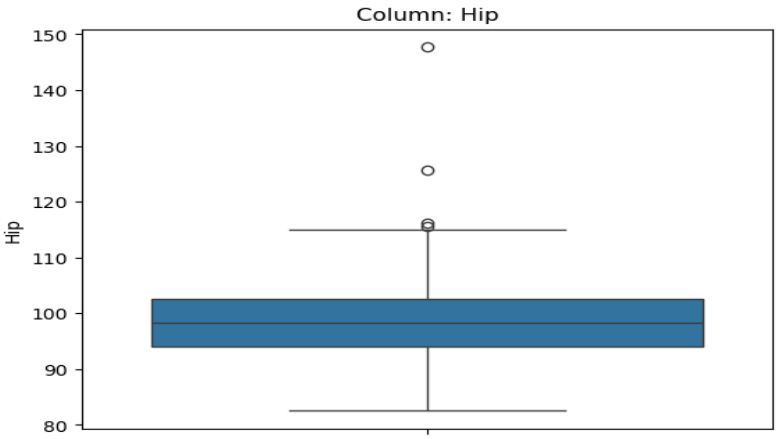
The data mining phase involves exploring techniques for feature selection in order to identify the most relevant variables that significantly contribute to predicting body fat percentage. This step is crucial in optimizing model performance and enhancing the accuracy of predictions.

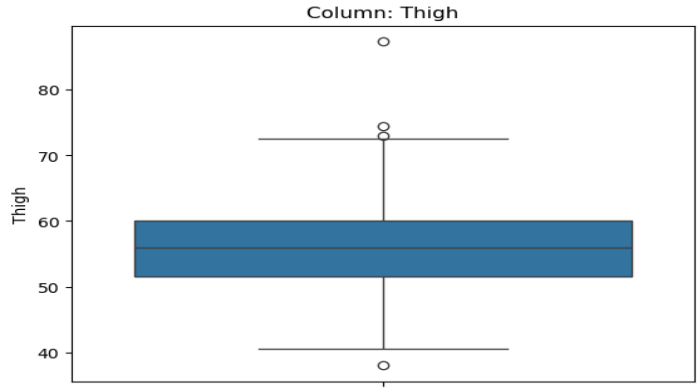


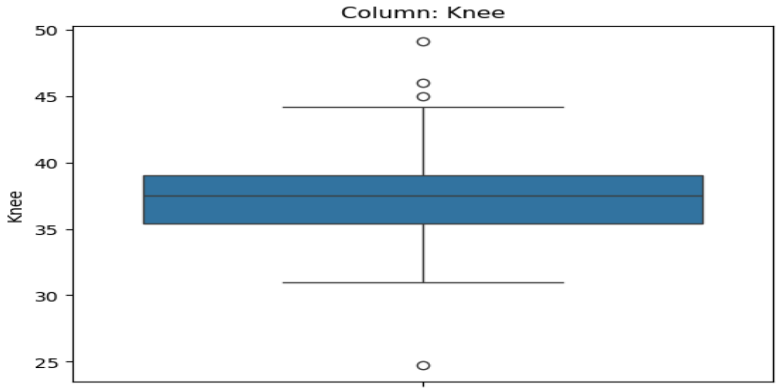

Column: Neck
Neck
50
45
40
35
30
25

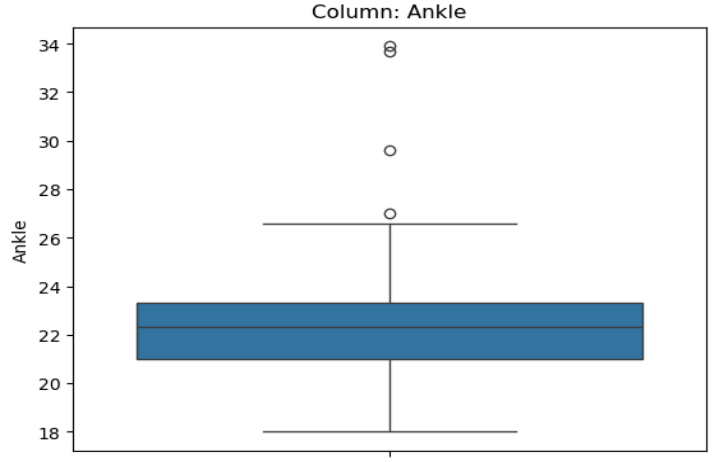

Column: Chest
Chest
140
120
100
80
60
40

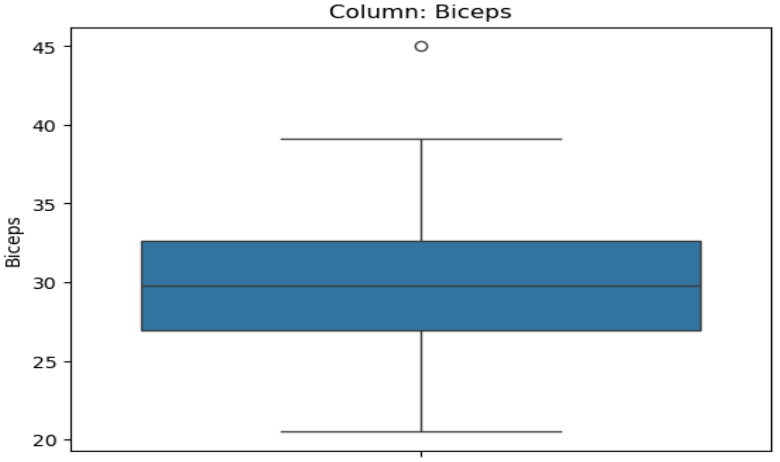

Column: Abdomen
Abdomen
140
120
100
80
60

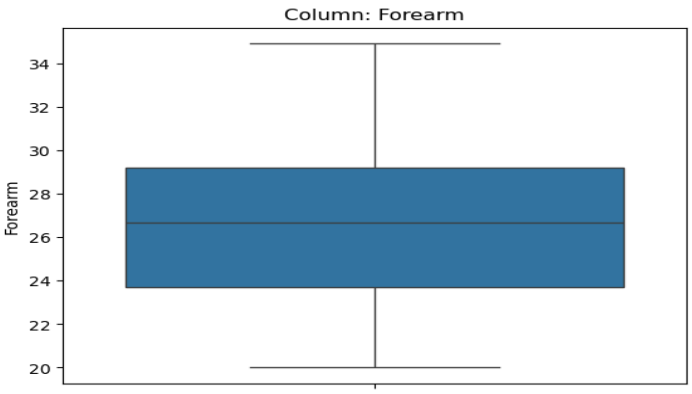

Column: Hip
Hip
150
140
130
120
110
100
90
80

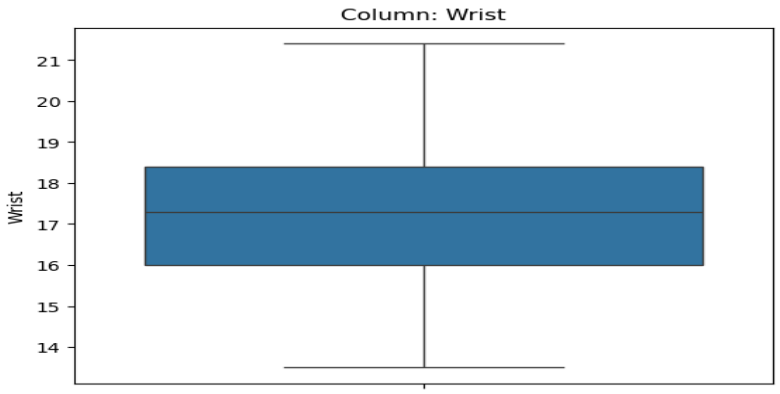

Column: Thigh
Thigh
80
70
60
50
40


Column: Knee
Knee
50
45
40
35
30
25


Column: Ankle
Ankle
34
32
30
28
26
24
22
20
18


Column: Biceps
Biceps
45
40
35
30
25
20





7. Training the Model:

In this section, we will discuss the process of choosing a machine-learning algorithm suitable for regression tasks. We will also explain how the model's output is presented visually and how its performance is evaluated using appropriate metrics. Our main goal is to train a precise and reliable model for estimating body fat percentage based on non-invasive measurements.perform train test split in 80:20 ratio of 80% for training and 20% for testing
Training set: (348, 15)
Test set: (88, 15)

Random Forest Regression:
Train | Mean Squared Error: 6.101 | Mean Absolute Error: 1.961 | R-squared: 0.892
Test | Mean Squared Error: 17.109 | Mean Absolute Error: 3.363 | R-squared: 0.671

Best Hyperparameters: OrderedDict([('max_depth', 14), ('max_features', 0.6937904169224471), ('max_samples', 1.0), ('min_samples_split', 9), ('n_estimators', 108)])

8. Results:

The results section showcases the accuracy and reliability of the trained model in estimating body fat percentage. It presents the model's predictive capabilities, performance metrics, and noteworthy findings derived from the analysis.Several regression models were trained and evaluated, including linear regression, ridge regression, LASSO regression, elastic net regression, PCR regression, and gradient boosting regression. With the lowest errors and a high explanatory power, the Gradient Boosting Regressor (GBR) emerged as the best-performing model.

Multiple Linear Regression

	Linear_train	Linear_test
MAE	3.231037	3.288859
MSE	15.921647	15.919518
RMSE	3.990194	3.989927
MAPE	0.235390	0.189282
Adjusted R-Squared	0.720774	0.540687

Ridge Regression

	Ridge_train	Ridge_test
MAE	3.261007	3.327680
MSE	16.491550	16.397958
RMSE	4.060979	4.049439
MAPE	0.244362	0.194072
Adjusted R-Squared	0.710779	0.526883

LASSO Regression

	LASSO_train	LASSO_test
MAE	5.096955	4.263214
MSE	40.203390	30.076998
RMSE	6.340614	5.484250
MAPE	0.430328	0.256539
Adjusted R-Squared	0.294932	0.132212

Principal Component Regression

	PCR_train	PCR_test
MAE	3.231037	3.288859
MSE	15.921647	15.919518
RMSE	3.990194	3.989927
MAPE	0.235390	0.189282
Adjusted R-Squared	0.720774	0.540687

Elastic Net Regression

	ElasNet_train	ElasNet_test
MAE	4.417379	3.813896
MSE	30.262202	22.843439
RMSE	5.501109	4.779481
MAPE	0.362765	0.227764
Adjusted R-Squared	0.469276	0.340916

Gradient Boosting Regressors

	GBR_train	GBR_test
MAE	1.710644	3.546613
MSE	4.643947	18.320159
RMSE	2.154982	4.280206
MAPE	0.122903	0.203801
Adjusted R-Squared	0.918557	0.471423

9. Conclusion and Future Work:

The main point of our concluding section is that the project summarizes its findings, discusses their implications, and outlines potential avenues for future research and improvement. It serves as a reflection on the project's achievements while providing a roadmap for ongoing advancements in the field of non-invasive body fat estimation. The section aims to leave room for continuous exploration and enhancement beyond the project's initial scope.Such as developing the app that can calculate the body Fat and give an instruction on how to overcome your Obese.

10. References

- Body fat extended Dataset,Obtained from kaggle
  https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset/data