ML Project

Body Fat Prediction

This task aims for predicting body fat percentage from a group of input features. Measuring fat percentage requires many body measurements. This is an attempt to build a model based on the input data to estimate the Body Fat with the sufficient number of features to do that, which could make fat percentage estimation more convenient.

The dataset comprises a set of features of 252 men (i.e. samples) in order to estimate Body Fat Percentage. The number of features is 13; 12 as input features and one (BodyFat) as target variable.

The project in steps:

A. The kind of task of this project represents a regression kind of Machine Learning models since the target variable is continuous variable.

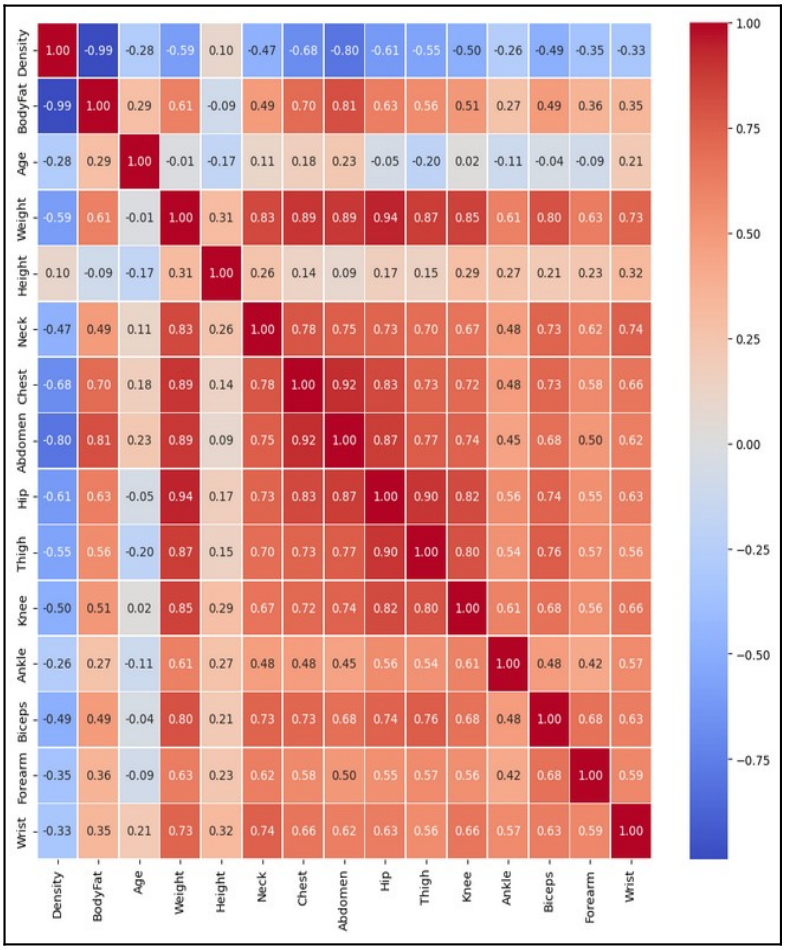
B. Exploratory Data Analysis(EDA):

1. The dataset has neither duplicated values nor null values.

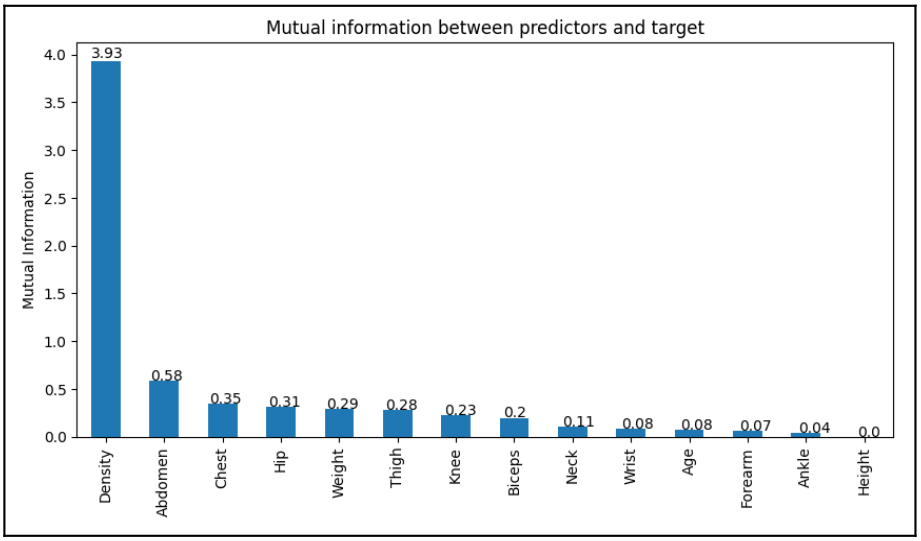
2. Plotting pair plots to project the relationships between each pair of the variables. The plots showed that none of the variables has skewed distribution. Data distributions follows normal distribution which is worth considering in linear models. As well, the pair plot shows linear relationship between Density and Body Fat variables.

3. It was noticed both Height and Weight own different unit scale comparing to other features' unit scale. So, the got converted into same unit scale (SI unit).

4. Heatmap based on pearson's correlation was created. It showed that Body Fat has strong negative correlation with Density (-0.99). Also, Body Fat has high correlation with abdomen and moderate correlation with some other features.



5. Mutual Information score was implemented which showed that Density can tell much information about Body Fat. No any other input feature was close to that.



C. Data Cleaning:

1. Outliers exists in most of the variable, most noticing in Weight variable. There were two options to remove outliers: using standard deviation (Z-score), or clipping the clipping the outlier values to min/max values based interquartile range (IQR). For the small number of samples, option two was implemented to include all samples for training and testing.

2. Two samples got dropped since it had unrealistic body fat percentage less than 2%.

3. At the end of data cleaning, the mean value of Weight variable has changed significantly from 178.92 (before data cleaning) to 81.32 (after data cleaning)

D. ML Models:

1. Model 1 based on Density only got 0.999 R2 score on testing data.

2. Model 2 based on all features on all features expect Density got 0.669 R2 score on testing data.

It can be concluded that Density is crucial for predicting body fat percentage. Model 1 can achieve high results on one feature only. Though, it is risky if Density is missing in new samples.