Body Fat Prediction

Introduction

Background

The body fat percentage (BFP) of a human or other living being is the total mass of fat divided by total body mass, multiplied by 100; body fat includes essential body fat and storage body fat. Essential is necessary to maintain life and reproductive functions. The body fat percentage is a measure of fitness level, since it is the only body measurement which directly calculates a person's relative body composition without regard to height or weight. The widely used body mass index (BMI) provides a measure that allows the comparison of the adiposity of individuals of different heights and weights. While BMI largely increases as adiposity increases, due to differences in body composition, other indicators of body fat give more accurate results; for example, individuals with greater muscle mass or larger bones will have higher BMIs. As such, BMI is a useful indicator of overall fitness for a large group of people, but a poor tool for determining the health of an individual.

Problem Statement

The terms “overweight” and “obesity” refer to body weight that is greater than what is considered normal or healthy for a certain height. Overweight is generally due to extra body fat. However, overweight may also be due to extra muscle, bone, or water. This would impact the health and there are many other diseases caused by extra body fat. We would be creating a model to predict the body fat based on some parameters. We would like to determine whether the person is obese or not in this project.

Technical Approach

The Cross-Industry Standard Process for Data Mining (CRISP-DM). This process model has 6 phases that naturally describe the data science life cycle for this project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

During every phase of this project lifecycle, we might discover new aspects/finding which we will incorporate them in ways to improve the efficiency of our model.

Data Sources

The dataset that is selected has all the details that are needed for a model to identify the body fat

- Density determined from underwater weighing
- Percent body fat from Siri's (1956) equation
- Age (years)
- Weight (lbs)
- Height (inches)
- Neck circumference (cm)
- Chest circumference (cm)
- Abdomen 2 circumference (cm)

- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Biceps (extended) circumference (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

Analysis

We are planning to perform feature reduction and dimensionality reduction to select the most relevant variables for our model. Many predictive algorithms assume the model variables follow a normal distribution. There are inherent advantages to using normally distributed variables, so our approach will focus on columns that closely follow this distribution. We will determine which variables are normally distributed by conducting the following analyses:

- Summary statistics on all variables: concentrating on mean and SD values.
- Identify the best model.

Requirement Development

For this project, we are planning to use python. To complete this project in python the following are required for our development:

IDE: Jupyter Notebook

Libraries:

1. Numpy – extensively used for data analysis to handle multidimensional arrays.
2. Pandas – high-performance data structures and analysis tools for the labeled data.

3. Matplotlib –powerful visualizations which create several stories with the data visualized.
4. SciPy – used for high-level technical computations.
5. Seaborn – interface for drawing attractive and informative statistical graphics.

Model Deployment

We will use feature selection techniques to finalize our feature list for models. Using the results from this step, we will build a couple of classification models and evaluate their performance. Below are some example models.

1.) Random Forest

2.) Linear Regression

Testing and Evaluation

We will be splitting the data into 70% training and 30% test dataset. Using the test dataset, we will test the model. Cross-validation will also be used to decide the best model.

Expected Results

Using this model to determine whether the person is obese or not.

Ethical Considerations

Ethical considerations in research are a set of principles that guide our research designs and practices. We must always adhere to a certain code of conduct when collecting data from people. The goals of human research often include understanding real-life phenomena, studying effective treatments, investigating behaviors, and improving lives in other ways. What you decide to research and how you conduct that research involve key ethical considerations.

These considerations work to

- Protect the rights of research participants

- Enhance research validity
- Maintain scientific integrity

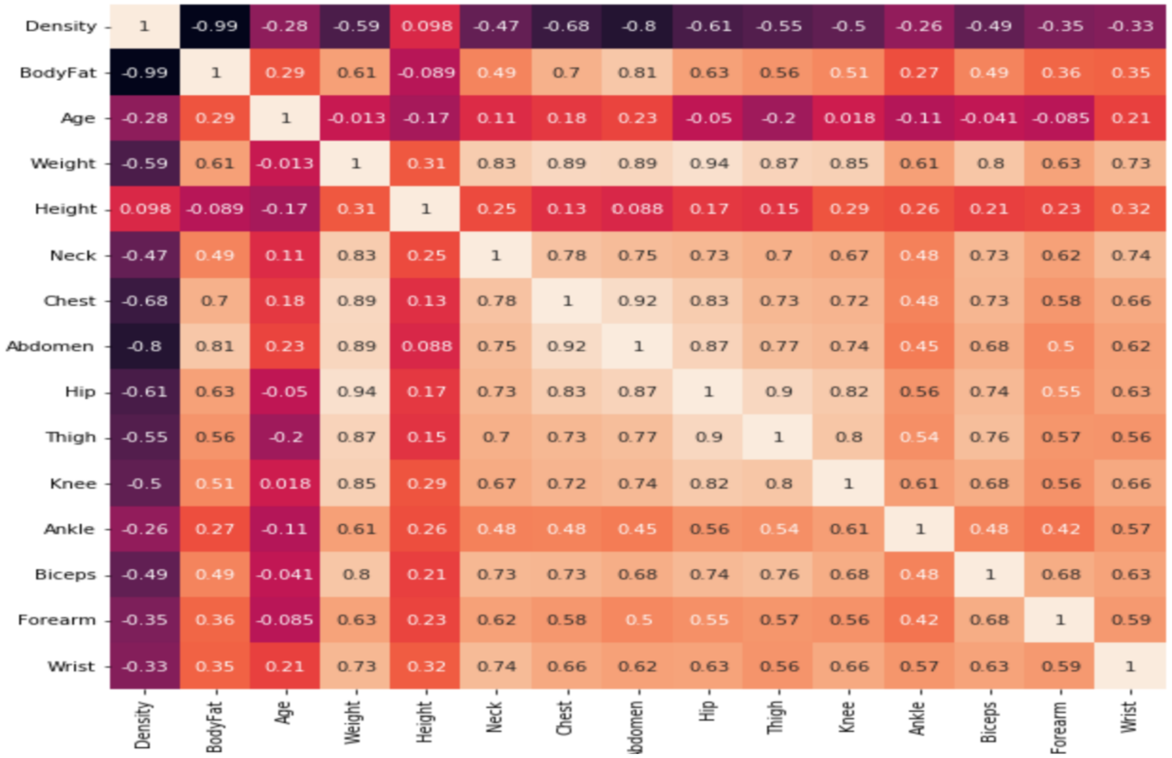
Challenges/Issues

The biggest challenge is conversion of the features into same units which would be required to improve the efficiency of the model.

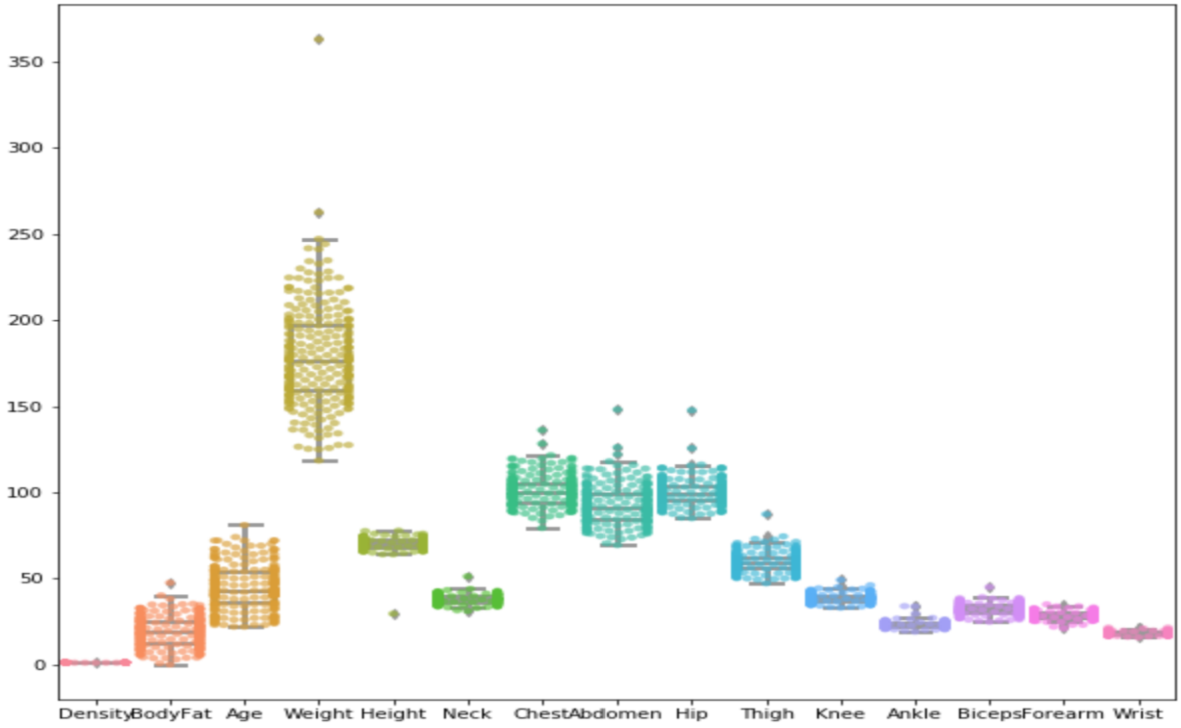
Reference:

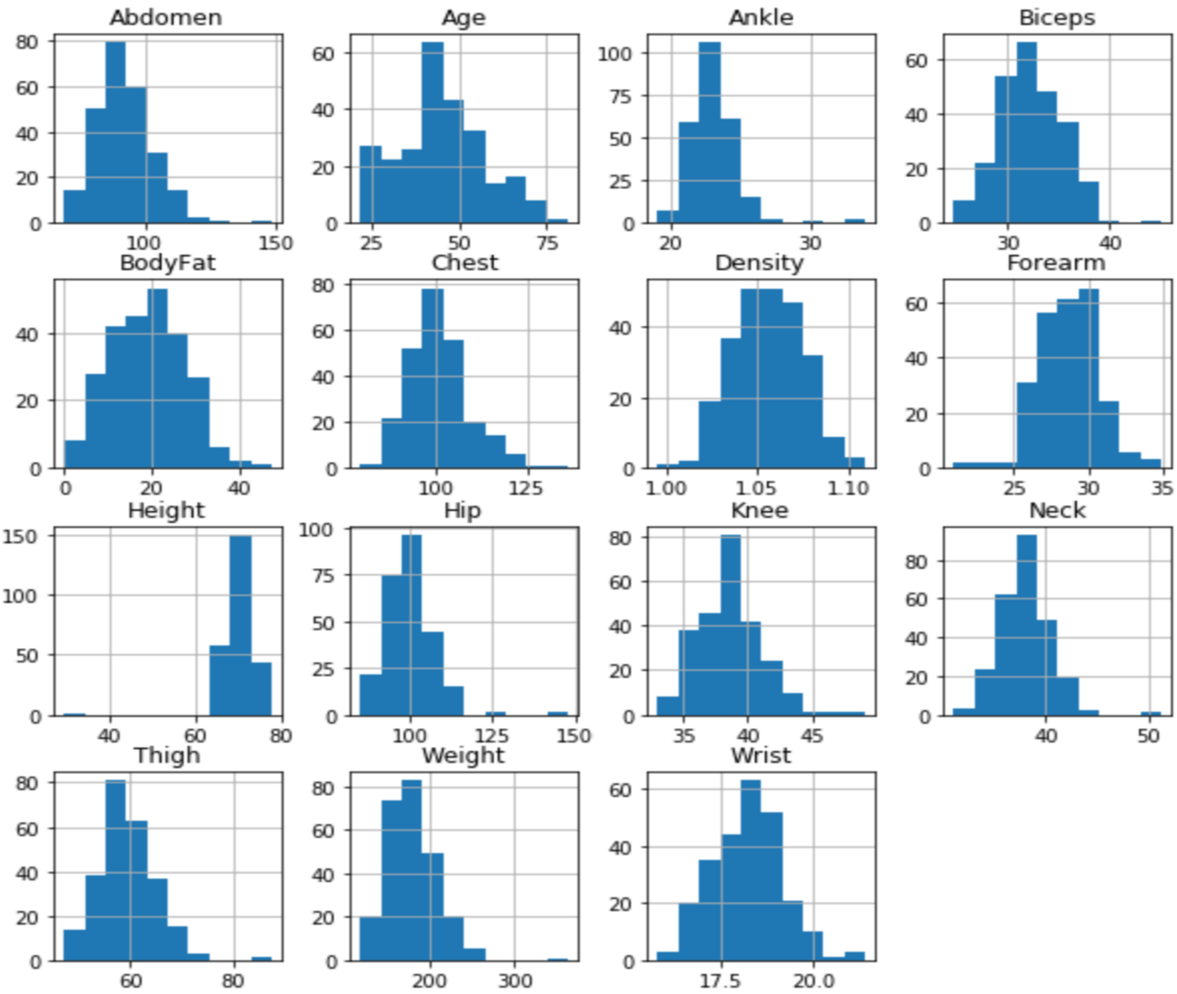
- Kaggle link for a dataset: https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset
- https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/definition-facts
- https://en.wikipedia.org/wiki/Body_fat_percentage#:~:text=The%20body%20fat%20percentage%20(BFP,maintain%20life%20and%20reproductive%20functions.
- https://www.scribbr.com/methodology/research-ethics/

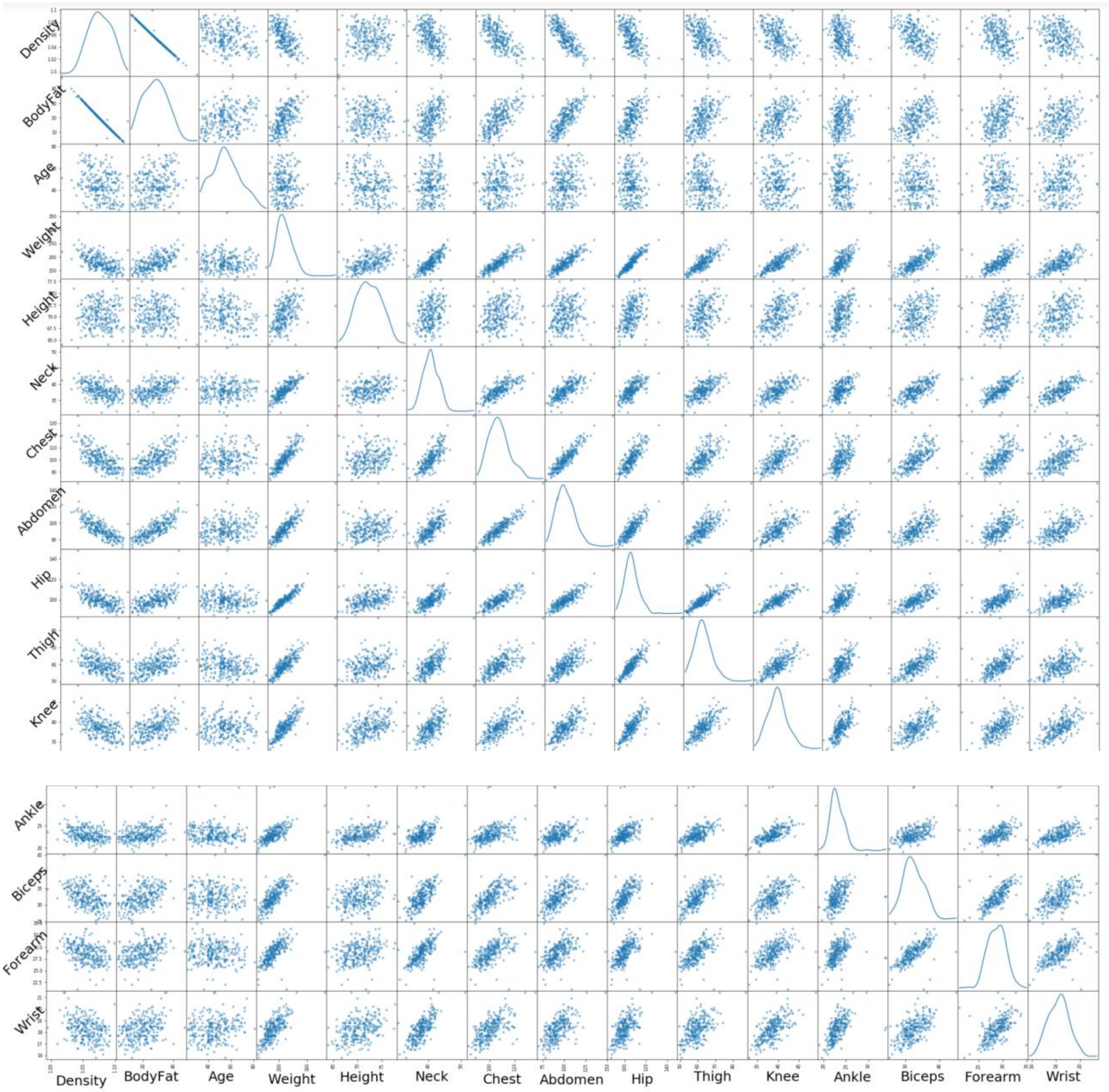
Illustrations:



	Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Density	1	-0.99	-0.28	-0.59	0.098	-0.47	-0.68	-0.8	-0.61	-0.55	-0.5	-0.26	-0.49	-0.35	-0.33
BodyFat	-0.99	1	0.29	0.61	-0.089	0.49	0.7	0.81	0.63	0.56	0.51	0.27	0.49	0.36	0.35
Age	-0.28	0.29	1	-0.013	-0.17	0.11	0.18	0.23	-0.05	-0.2	0.018	-0.11	-0.041	-0.085	0.21
Weight	-0.59	0.61	-0.013	1	0.31	0.83	0.89	0.89	0.94	0.87	0.85	0.61	0.8	0.63	0.73
Height	0.098	-0.089	-0.17	0.31	1	0.25	0.13	0.088	0.17	0.15	0.29	0.26	0.21	0.23	0.32
Neck	-0.47	0.49	0.11	0.83	0.25	1	0.78	0.75	0.73	0.7	0.67	0.48	0.73	0.62	0.74
Chest	-0.68	0.7	0.18	0.89	0.13	0.78	1	0.92	0.83	0.73	0.72	0.48	0.73	0.58	0.66
Abdomen	-0.8	0.81	0.23	0.89	0.088	0.75	0.92	1	0.87	0.77	0.74	0.45	0.68	0.5	0.62
Hip	-0.61	0.63	-0.05	0.94	0.17	0.73	0.83	0.87	1	0.9	0.82	0.56	0.74	0.55	0.63
Thigh	-0.55	0.56	-0.2	0.87	0.15	0.7	0.73	0.77	0.9	1	0.8	0.54	0.76	0.57	0.56
Knee	-0.5	0.51	0.018	0.85	0.29	0.67	0.72	0.74	0.82	0.8	1	0.61	0.68	0.56	0.66
Ankle	-0.26	0.27	-0.11	0.61	0.26	0.48	0.48	0.45	0.56	0.54	0.61	1	0.48	0.42	0.57
Biceps	-0.49	0.49	-0.041	0.8	0.21	0.73	0.73	0.68	0.74	0.76	0.68	0.48	1	0.68	0.63
Forearm	-0.35	0.36	-0.085	0.63	0.23	0.62	0.58	0.5	0.55	0.57	0.56	0.42	0.68	1	0.59
Wrist	-0.33	0.35	0.21	0.73	0.32	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59	1




Abdomen
0
20
40
60
80
100
150
Age
0
20
40
60
25
50
75
Ankle
0
25
50
75
100
20
30
Biceps
0
20
40
60
30
40
BodyFat
0
20
40
0
20
40
Chest
0
20
40
60
80
100
125
Density
0
20
40
1.00
1.05
1.10
Forearm
0
20
40
60
25
30
35
Height
0
50
100
150
40
60
80
Hip
0
25
50
75
100
100
125
150
Knee
0
20
40
60
80
35
40
45
Neck
0
20
40
60
80
40
50
Thigh
0
20
40
60
80
60
80
Weight
0
20
40
60
80
200
300
Wrist
0
20
40
60
17.5
20.0



Appendix:

- **Random forest:** It builds multiple decision trees and merges them to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

- **Logistic regression:** It is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.
- **Linear Regression:** Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.
- **Random Forest Regression:** Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.
- **Leave One Out:** The Leave-One-Out Cross-Validation, or LOOCV, procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- **Box Cox Transformation:** A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.
- **Decision Tree:** It belongs to the family of supervised learning algorithms. It can be used for solving regression and classification problems. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable

by learning simple decision rules inferred from training data. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. Based on the comparison, we follow the branch corresponding to that value and jump to the next node.

- Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns.
- A numpy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.