Projet Python for Data Analysis

Maxime de la Tour

The Dataset

The dataset is "*Estimation of obesity levels based on eating habits and physical condition*" (Palechor, F. M., & de la Hoz Manotas, A. (2019)).

The goal is to predict obesity levels.

It is the result of data collection from a survey conducted in southern american countries and extrapolation from said data to obtain balanced obesity levels. 77% of the data is synthetic. The resulting dataset has 2111 observations for 17 attributes.

Column name	Description
Gender	Self explanatory
Age	Self explanatory
Height	Self explanatory
Weight	Self explanatory
family_history_with_overweight	Self explanatory
FAVC	Eats caloric food frequently
FCVC	Frequency eating vegetables
NCP	number of meals
CAEC	eats btween meals
SMOKE	Self explanatory
CH2O	frequency drinking water
SCC	frequency monitors calories
FAF	phys. activity frequency
TUE	time using tech. devices
CALC	frequency drinking alcohol
MTRANS	Usual transportation
NObeyesdad	Obesity level

The Dataset

The dataset came in two file formats, ARFF and CSV.

The goal was initially to load the ARFF file into a pandas DataFrame using scipy's arff library. However, this resulted in glitches with the strings (see image).

Loading the the CSV using panda's read_csv function yields perfectly normal results.

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	
0	b'Female'	21.0	1.62	64.0	b'yes'	b'no'	2.0	3.0	b
1	b'Female'	21.0	1.52	56.0	b'yes'	b'no'	3.0	3.0	b
2	b'Male'	23.0	1.80	77.0	b'yes'	b'no'	2.0	3.0	b
3	b'Male'	27.0	1.80	87.0	b'no'	b'no'	3.0	3.0	b

Improperly loaded data

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC
0	Female	21.0	1.62	64.0	yes	no	2.0
1	Female	21.0	1.52	56.0	yes	no	3.0
2	Male	23.0	1.80	77.0	yes	no	2.0
3	Male	27.0	1.80	87.0	no	no	3.0

Properly loaded data

Data Exploration

According to the paper, the obesity level is simply obtained from the mass body index (MBI) given by :

$MBI = Weight/Height^2$

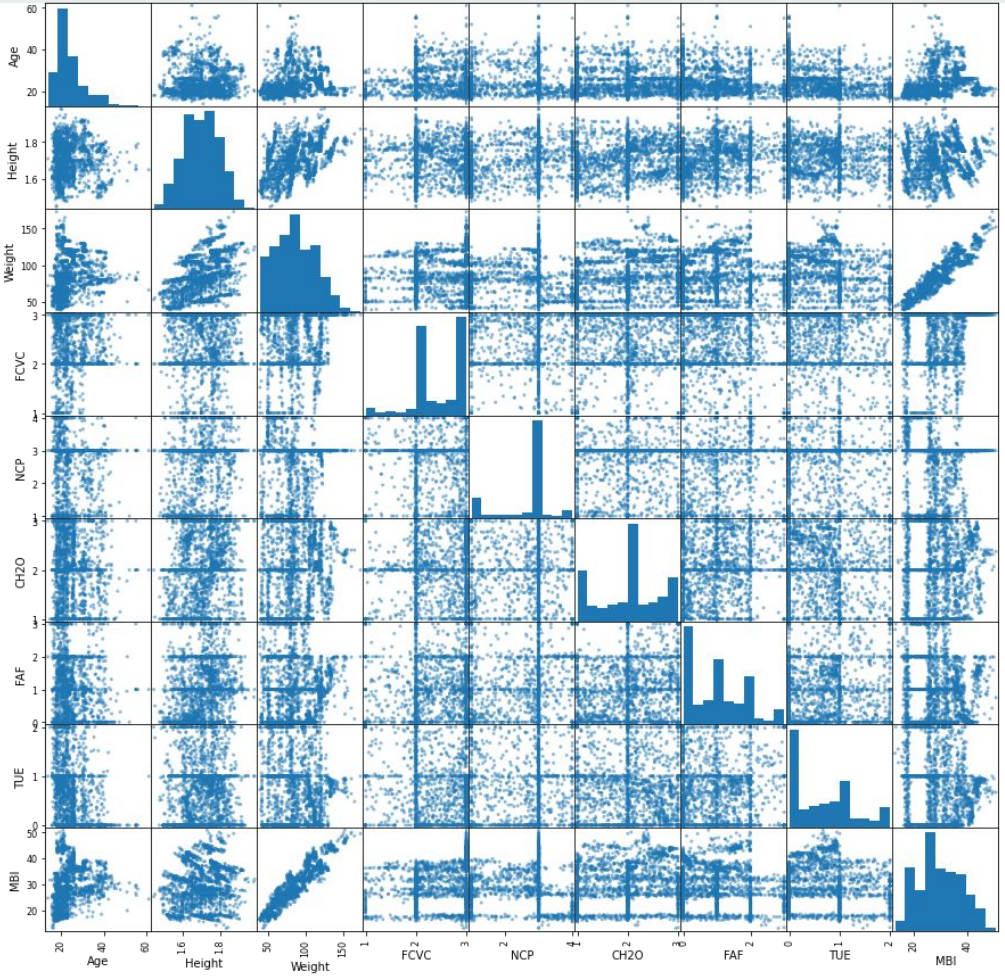
- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 40

This allows us to create a custom continuous variable to measure correlations.

Data Exploration

Once the custom column created, we can take a look at the raw data .

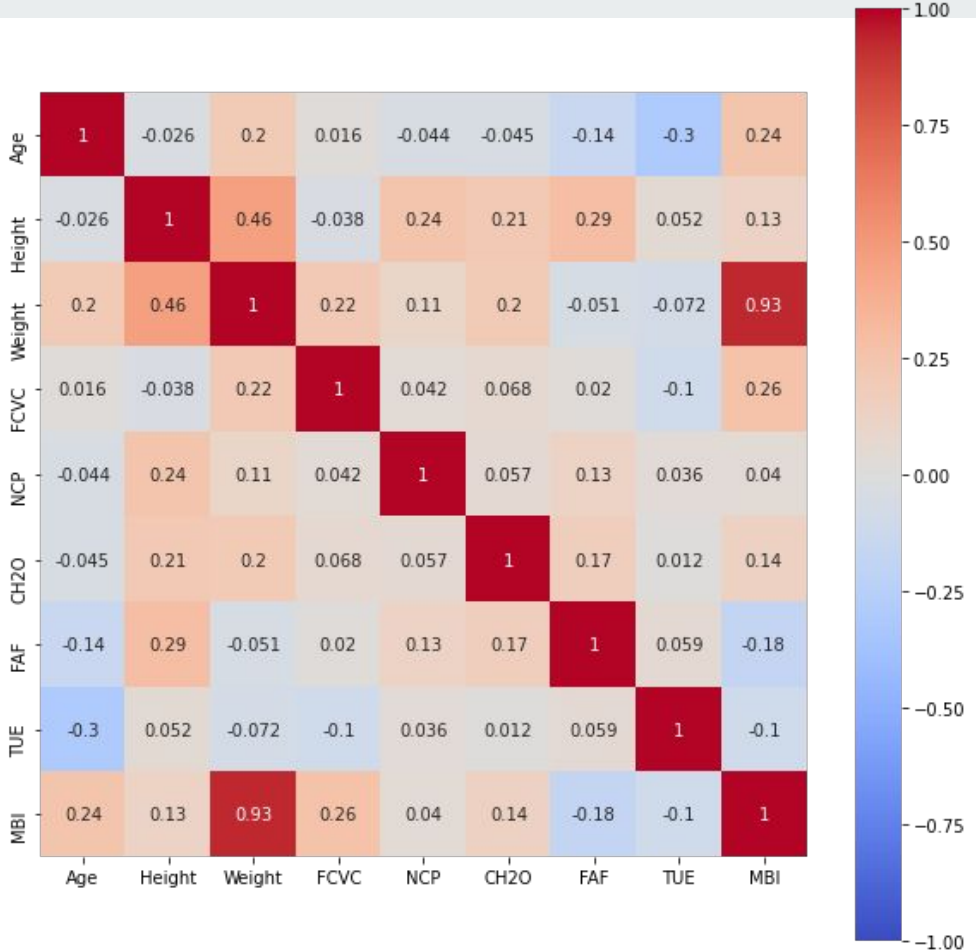
Weight jumps out as being a good predictor ob MBI. Of course, as it is used in the MBI calculation, we can't use it to make prediction, as for height.



Data Exploration

Now we can take a look at the correlation matrix. Looking at the MBI row, it appears that, besides height and weight (obviously), age and vegetable consumption are the most strongly correlated variables to the MBI.

We notice some other interesting correlations outside the scope of MBI prediction, like the inverse correlation between age and screen time, or height and physical activity.



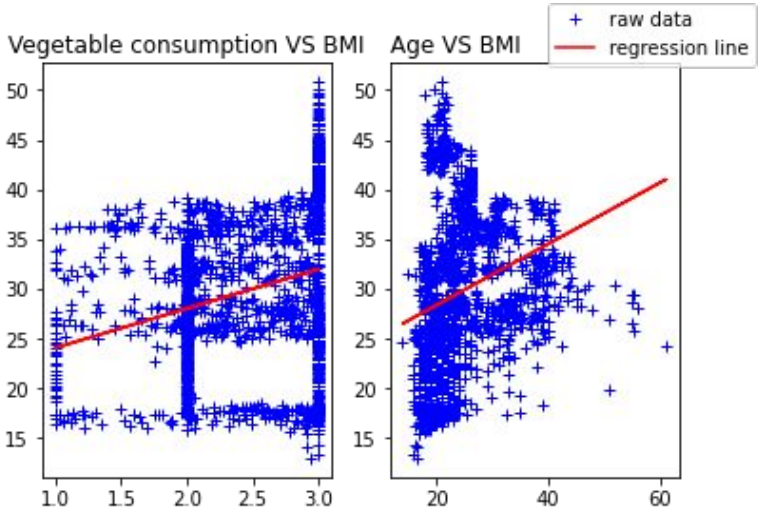
Visualisation & basic models

We use SKlearn to fit a linear regression model to each of those variables and Matplotlib to visualize the results :

Unsurprisingly, the r^2 scores are bad : 0.07 and 0.06.

However, fitting a multiple linear regression algorithm to both gives us a r^2 of 0.13.

Adding all features raises the r^2 to 0.18



Trying different models

We first split the dataset into a training set of the first 1500 rows, and the remaining 611 are used as a test set.

First, the multiple linear regression, which has a low r^2 score.

Then, we test with a multiple polynomial interpolation with ridge reggression, testing with different degrees :

- r^2 score for degree 2 : 0.29914294439952116
- r^2 score for degree 3 : 0.36620604700465076
- r^2 score for degree 4 : 0.45107032359982957
- r^2 score for degree 5 : 0.5816602568208353
- r^2 score for degree 6 : 0.699646460213473
- r^2 score for degree 7 : 0.3715665526502965

Trying different models

Beyond polynomials of degree 6 the r^2 score drops significantly. The best fit is achieved at degree 6, and the r^2 is close to 0.7, which is satisfactory.