JimjimValkema-EDA-conclusie-discussie

Jim Jim Valkema

9/8/2020

conclusion

This data set has a great amount of attributes however it does suffer from a great amount of missing values and only has 39 usable instances (fig 1). The data has 155 attributes that have more than 60% missing data excluding completely empty attributes. However there are 544 attributes left when the empty attributes and those with more than 60% missing are removed. But it might be challenging to effectively train a machine learning algorithm on a data set with only 39 instances.
The ranges of different attributes are quite significant (fig 2) so the data needs to be normalized. It is found that scaling by the standard deviation would be the most effective method for normalization (fig 3) for a future proof machine learning model. This is because this type of normalization can handle data that falls outside of the range of the training data set. Unlike min max normalization which fails in this regard. It is important that the end product can handle these ranges because the training data set has only a few instances and therefore might not contain the entire range of values for each attribute. The data also benefits from a log2 transformation because it seems to be more normal distributed this way (fig 4 & 5). Most attributes that were mentioned by the previous study are not correlated with each other (fig 6 & 7). Which could be an indication that they contain a good amount of information for the machine algorithm to train on.

discussion

There only a few of the many attributes that are looked at in this study. These attributes were picked based on the findings of the previous study. This is because of time limitations however it could be interesting to look at the rest of the data with more time efficient techniques and tools. However the final product can't have too many attributes as its input because that would be too cumbersome to use and most machine learning algorithms would probably benefit much from this many attributes. There are also a fair amount of missing values removed which might still be valuable because a missing value on its own can also convey information in some cases.

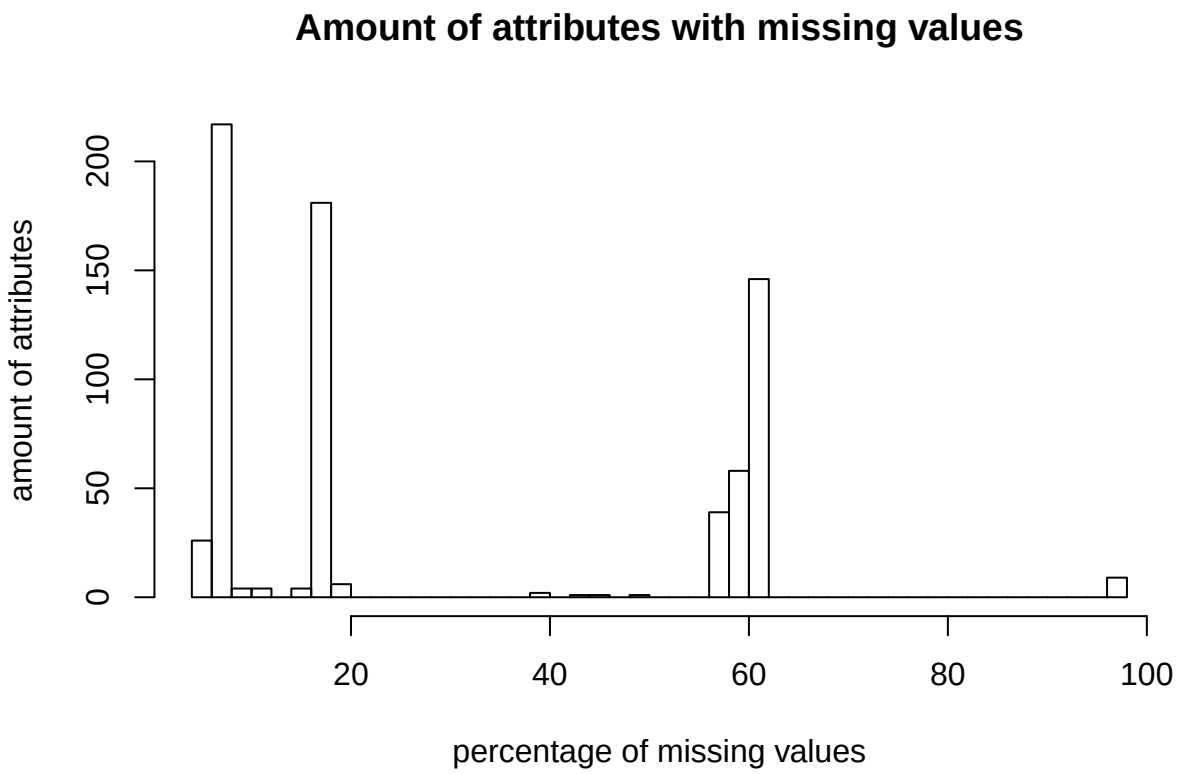


Fig 1: The amount of attributes on the y axis and the percentage of missing values these attributes have. This figure shows that a lot of attributes has below 20% missing but most notably are those with around 60% missing.

```
FALSE No id variables; using all as measure variables
```

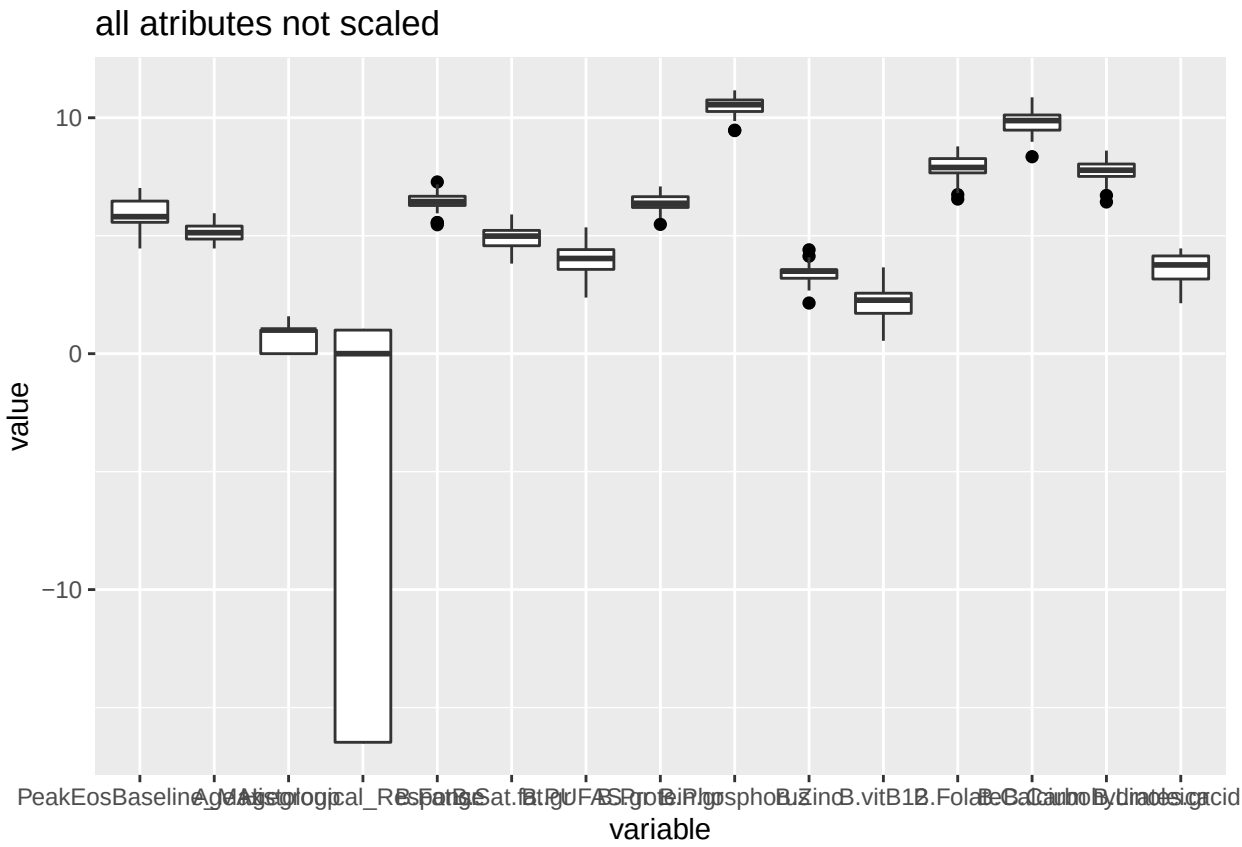


Figure 2: Shows the selected attributes distributions. Note that the range differ significantly which can interfere with most machine learning algorithms.

```
FALSE No id variables; using all as measure variables
```

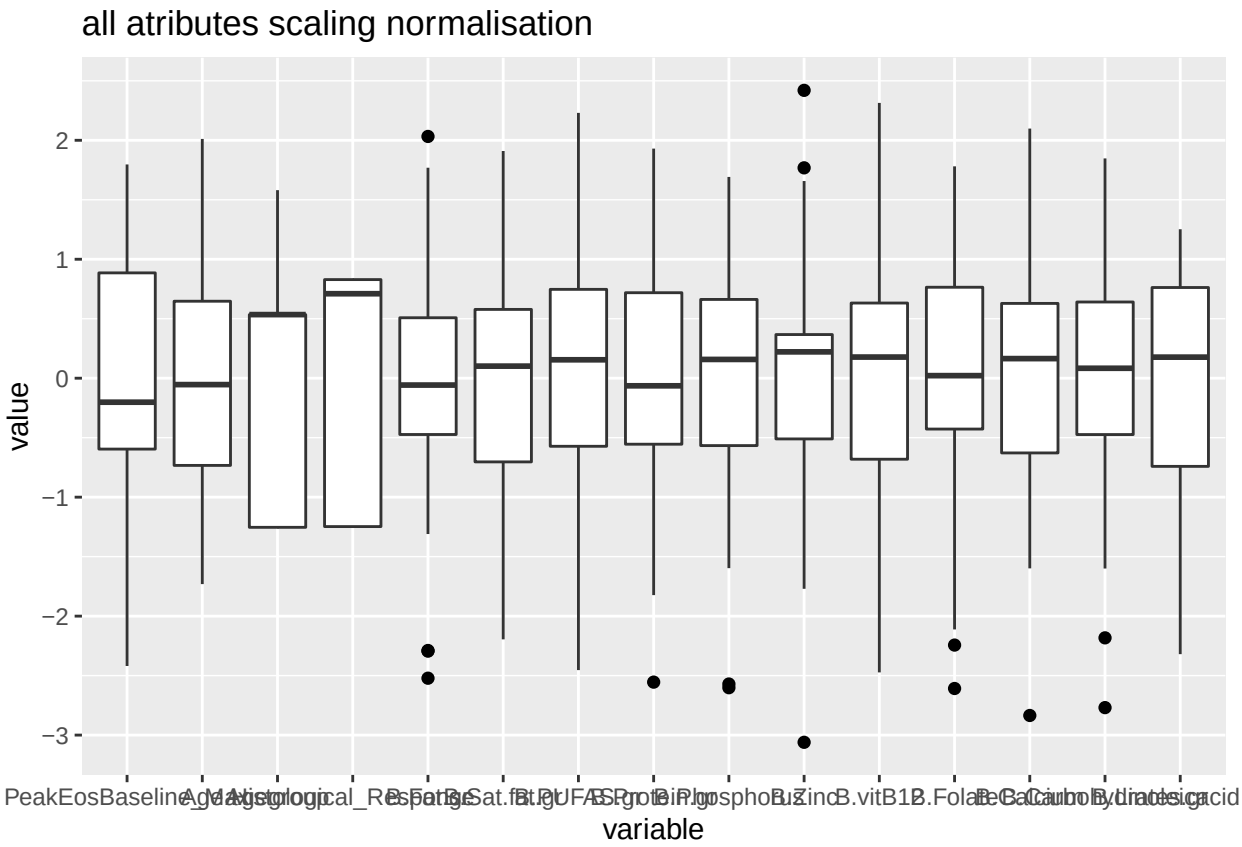


Figure 3: Shows the distributions of the same attributes as figure 2 but with scaling normalisation which makes for much more manageable ranges.

```
FALSE No id variables; using all as measure variables
```

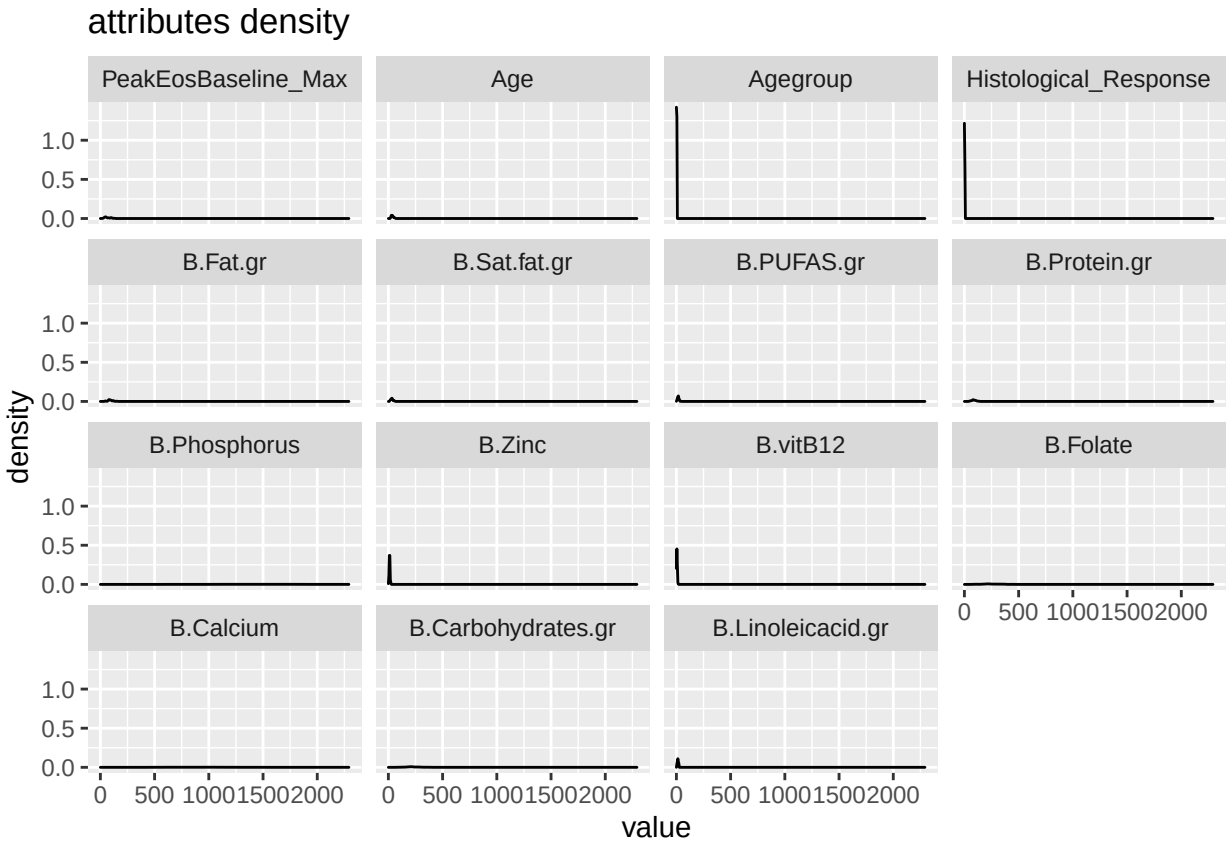


Figure 4: show the distribution of the attributes on a density graph. A lot of data is completely shifted toward the x axes which show that it might benefit from a log2 transform to achieve a more normal distribution.

```
FALSE No id variables; using all as measure variables
```

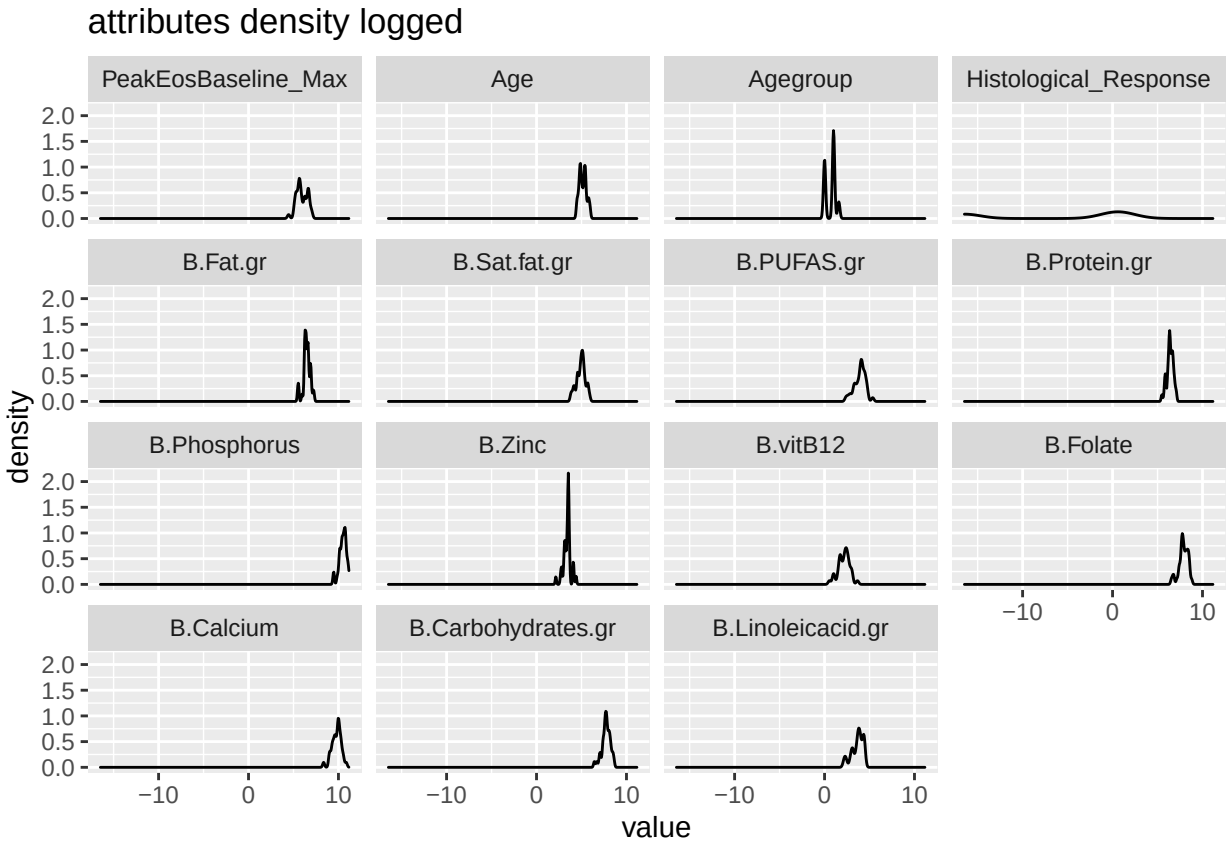
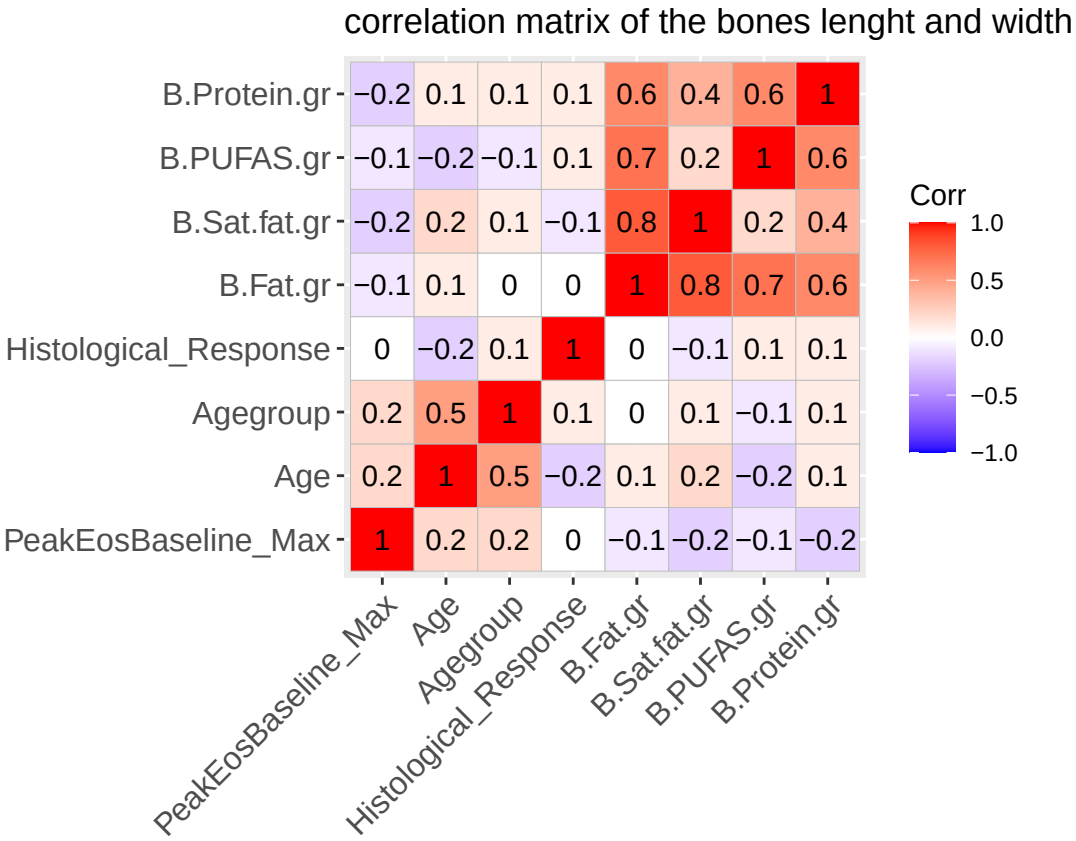


Figure 5: Show the data after a log2 transform which makes for a dataset that appears to be normal distributed.


correlation matrix of the bones lenght and width
B.Protein.gr
B.PUFAS.gr
B.Sat.fat.gr
B.Fat.gr
Histological_Response
Agegroup
Age
PeakEosBaseline_Max
−0.2 0.1 0.1 0.1 0.6 0.4 0.6 1
−0.1 −0.2 −0.1 0.1 0.7 0.2 1 0.6
−0.2 0.2 0.1 −0.1 0.8 1 0.2 0.4
−0.1 0.1 0 0 1 0.8 0.7 0.6
0 −0.2 0.1 1 0 −0.1 0.1 0.1
0.2 0.5 1 0.1 0 0.1 −0.1 0.1
0.2 1 0.5 −0.2 0.1 0.2 −0.2 0.1
1 0.2 0.2 0 −0.1 −0.2 −0.1 −0.2
PeakEosBaseline_Max
Age
Agegroup
Histological_Response
B.Fat.gr
B.Sat.fat.gr
B.PUFAS.gr
B.Protein.gr
Corr
1.0
0.5
0.0
−0.5
−1.0

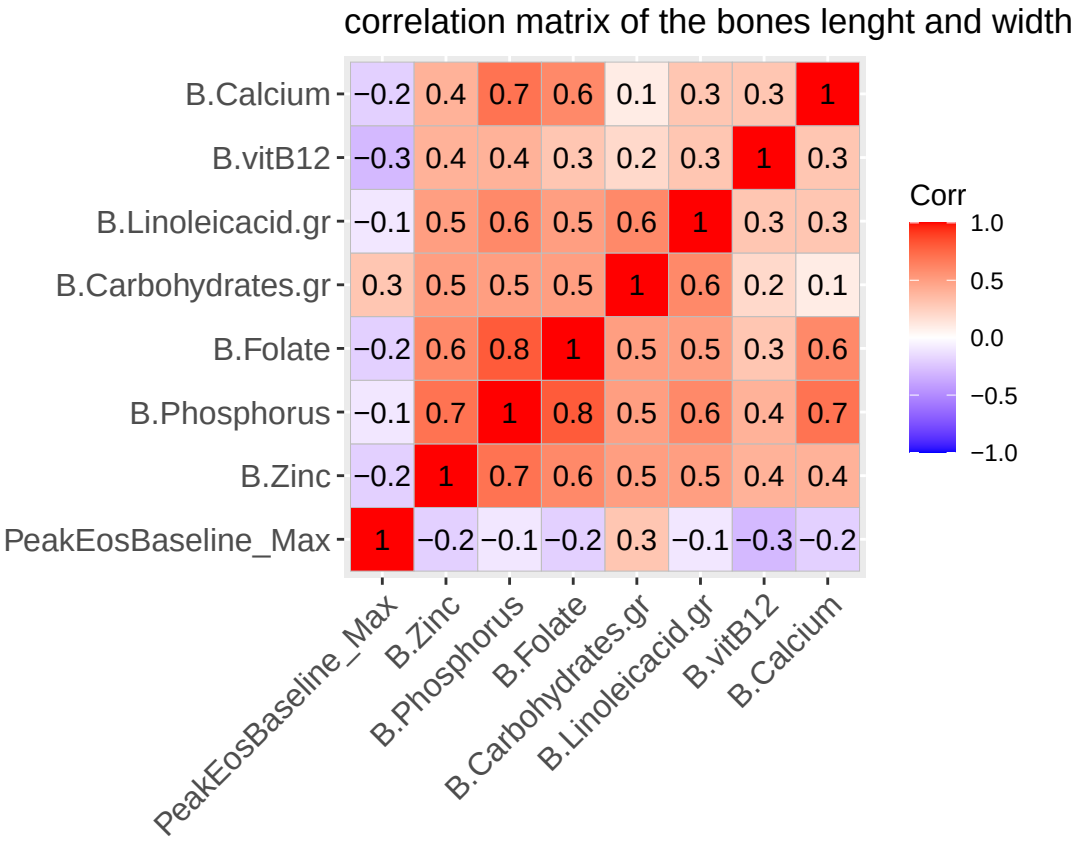


Figure 6 & 7: show the correlation between attributes. It seems that the majority isn't highly correlated which might be a good indication that there is enough information between attributes for the algorithm to train on.