Eosinofiele Oesofagitis dataset EDA

Jim Jim Valkema

9/8/2020

Dataset

TODO neocate en ffed Nutritional intake was assessed in 40 Dutch adult EoE patients participating in the Supplemental Elemental Trial (SET) using 3-day food diaries. In this randomized controlled trial, diagnosed patients received either a four-food elimination diet alone (FFED) or FFED with addition of an amino acid-based formula (Neocate) for 6 weeks. Disease severity was assessed by peak eosinophil count/high power field (PEC) in esophageal biopsy specimens. Multiple linear regression analyses were performed to assess associations between the intake of nutrients and foods per 1000 kCal and PEC, both at baseline and after the 6 weeks diet, while controlling for baseline variables. What do vars mean datatypes, units, values What is the dependent class variable

Introduction

Research question

Can the peak eosinophil baseline in individuals with eosinophilic oesophagitis be predicted with classical machine learning using the measured nutrient intake data?

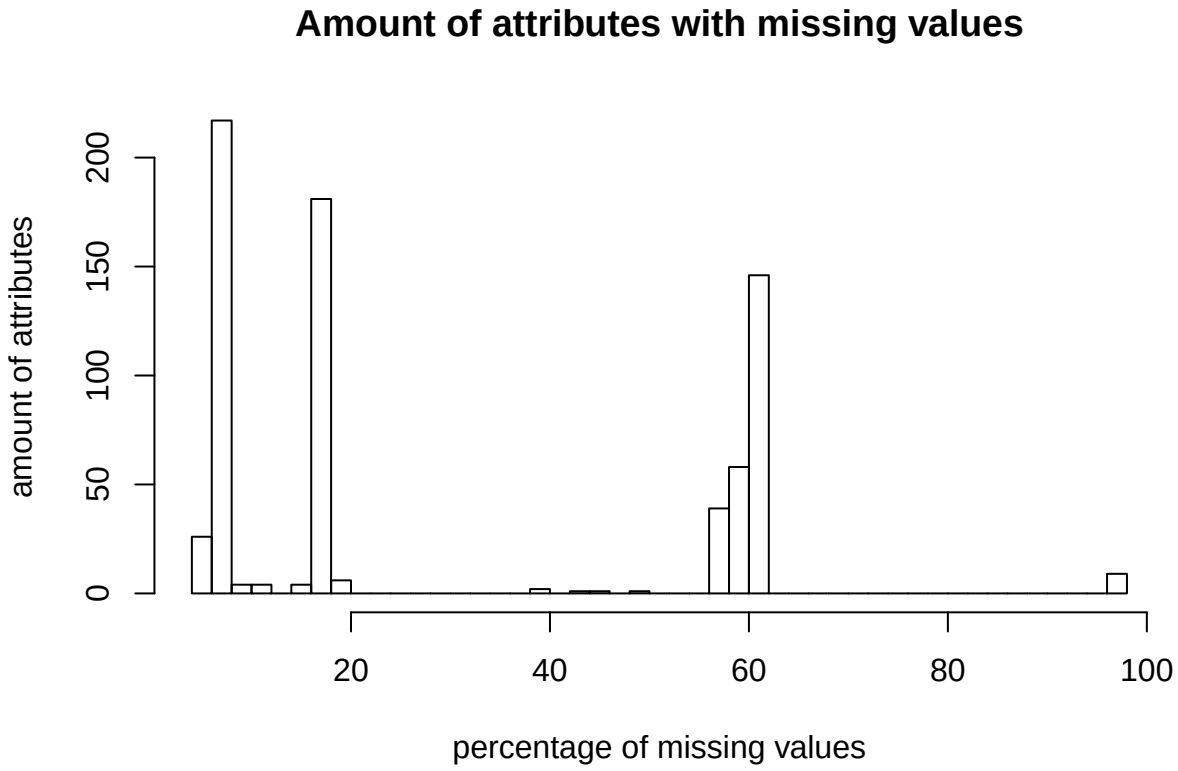
Missing data

Missing columns TODO There are quite a bit of missing attributes and missing values in this data set. Luckily there are plenty of attributes that do have data from where the machine learning algorithm can make its prediction.

37 empty instances 180 empty attributes Lots of columns with more then 70 % missing values

```
FALSE
FALSE Attaching package: 'janitor'

FALSE The following objects are masked from 'package:stats':
FALSE
FALSE     chisq.test, fisher.test

FALSE Loading required package: lattice

FALSE Loading required package: MASS

FALSE
FALSE Attaching package: 'memisc'

FALSE The following objects are masked from 'package:stats':
FALSE
FALSE     contr.sum, contr.treatment, contrasts

FALSE The following object is masked from 'package:base':
FALSE
FALSE     as.array
```

```
FALSE
FALSE Attaching package: 'ggplot2'
FALSE The following object is masked from 'package:memisc':
FALSE
FALSE     syms
FALSE Removing 37 empty rows of 79 rows total (46.8%).
FALSE Removing 180 empty columns of 879 columns total (Removed: Date, Eoscutoff, Lenght, Weight, BMI, Ag
```

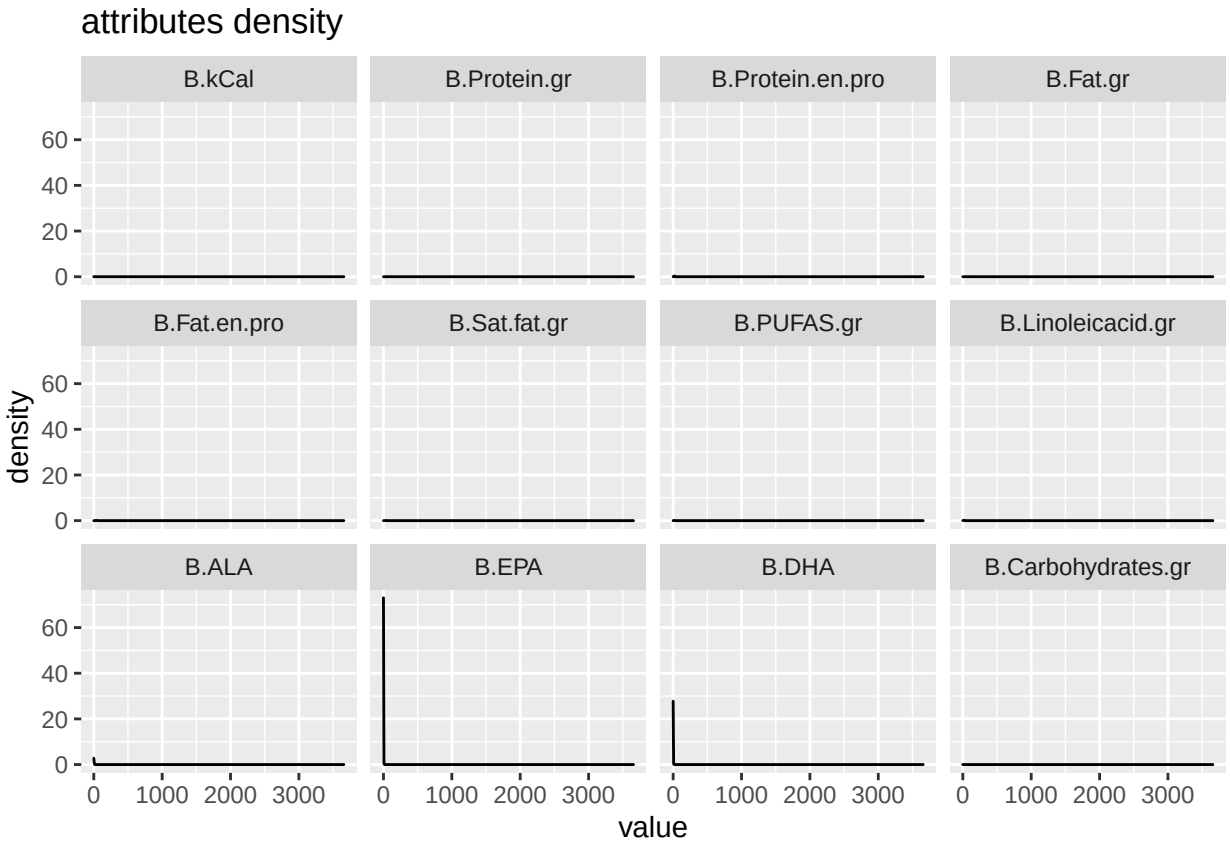



```
FALSE [1] "columns removed because too many missing values"
FALSE [1] 155
FALSE [1] "columns romved in total"
FALSE [1] 335
FALSE [1] 544
```

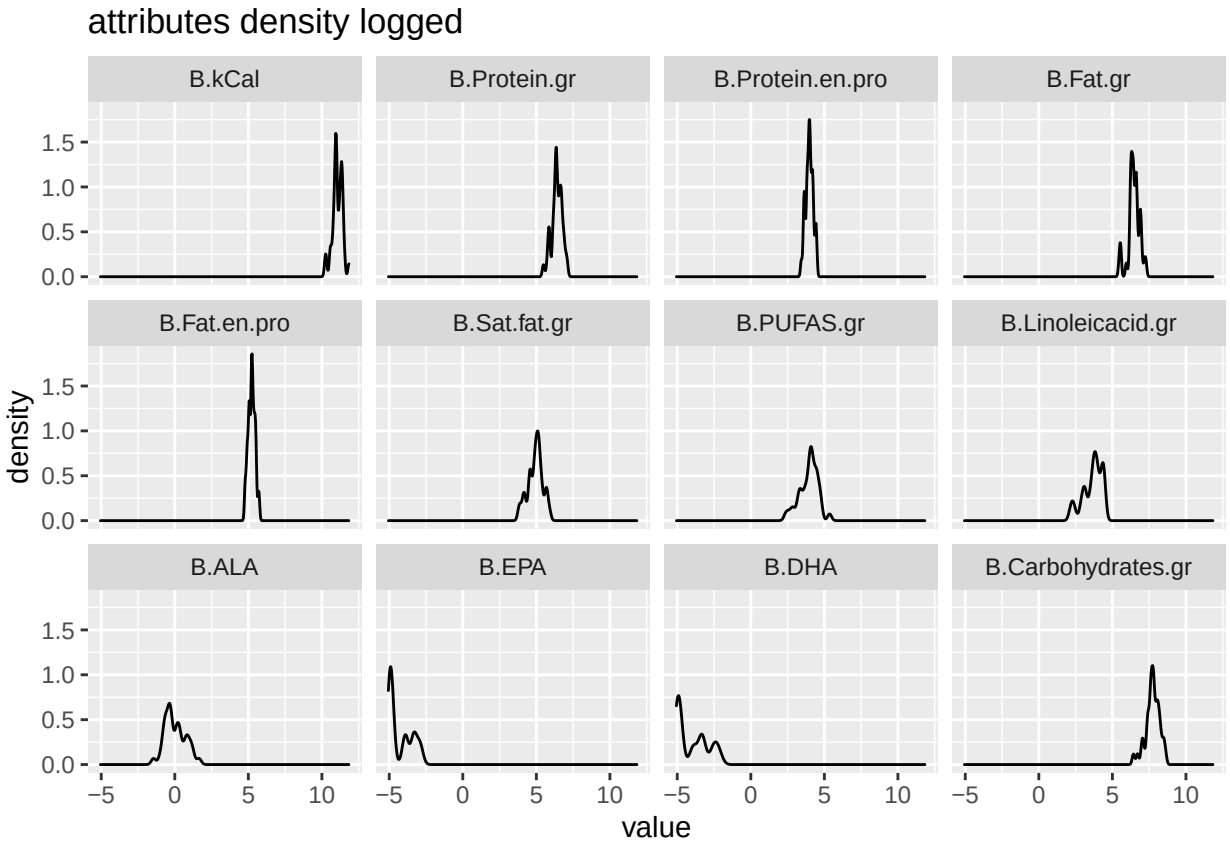
distrubution

Most attributes seem normally distributed how ever their range differ quite significantly so a normalization step might be required other wise the range can bias the machine learning algorithm to specific attributes.

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

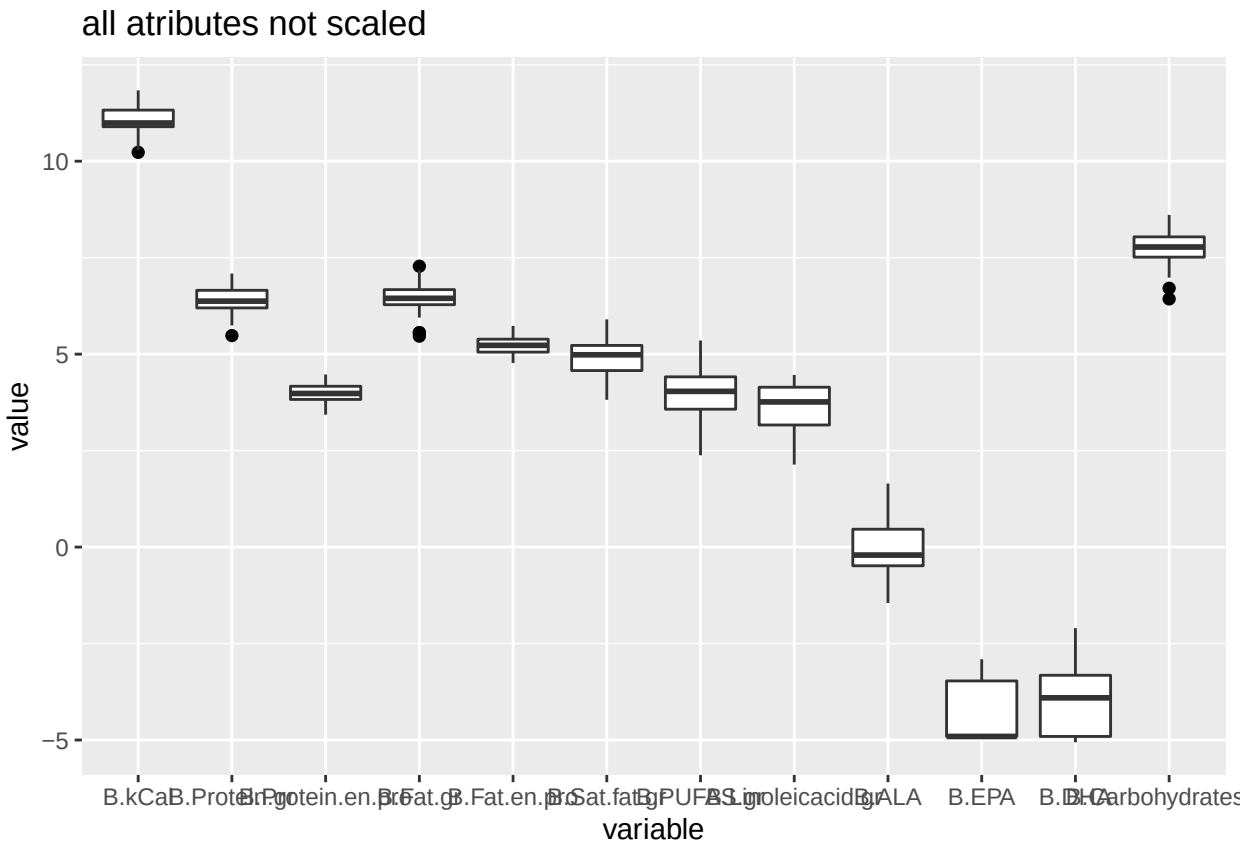



```
FALSE No id variables; using all as measure variables
```

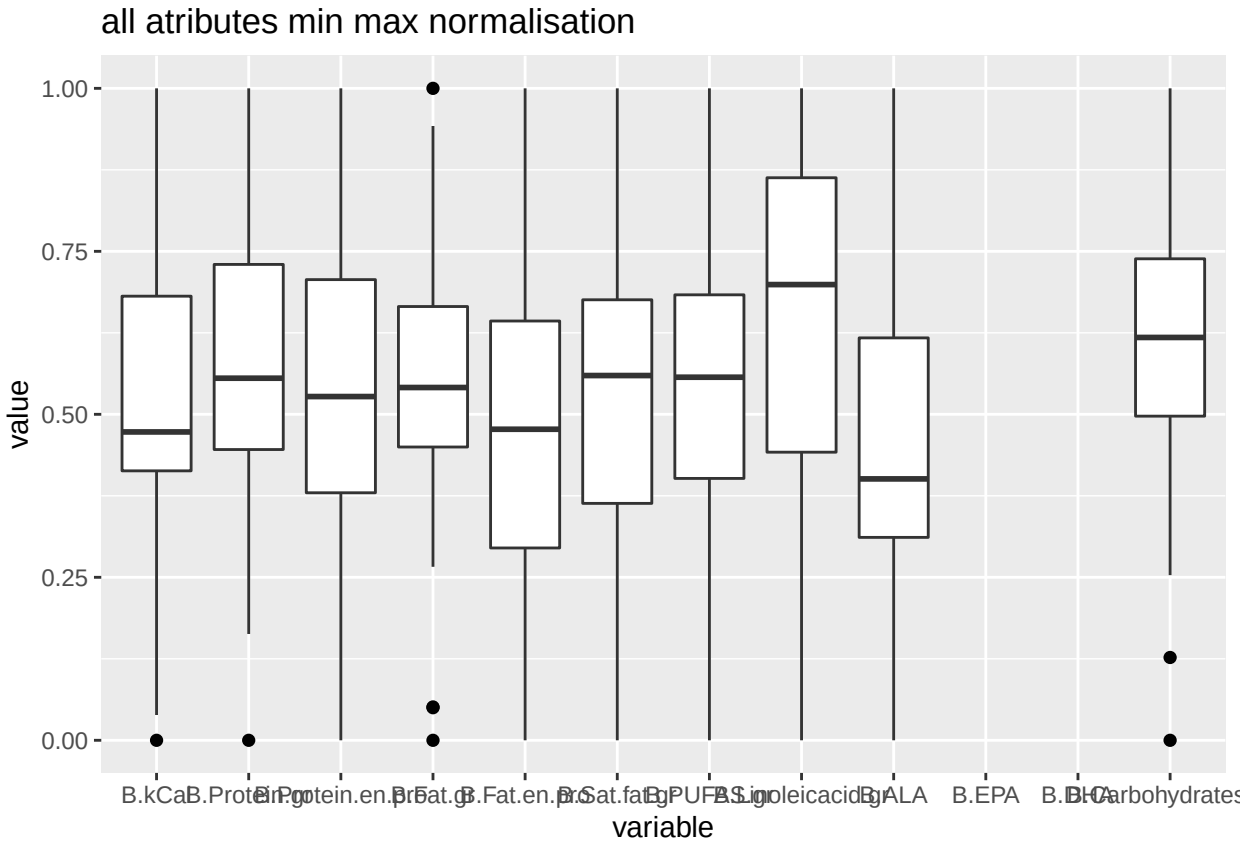


Normalisation This is all attributes compared with no normalization, min-max and scaling normalization.

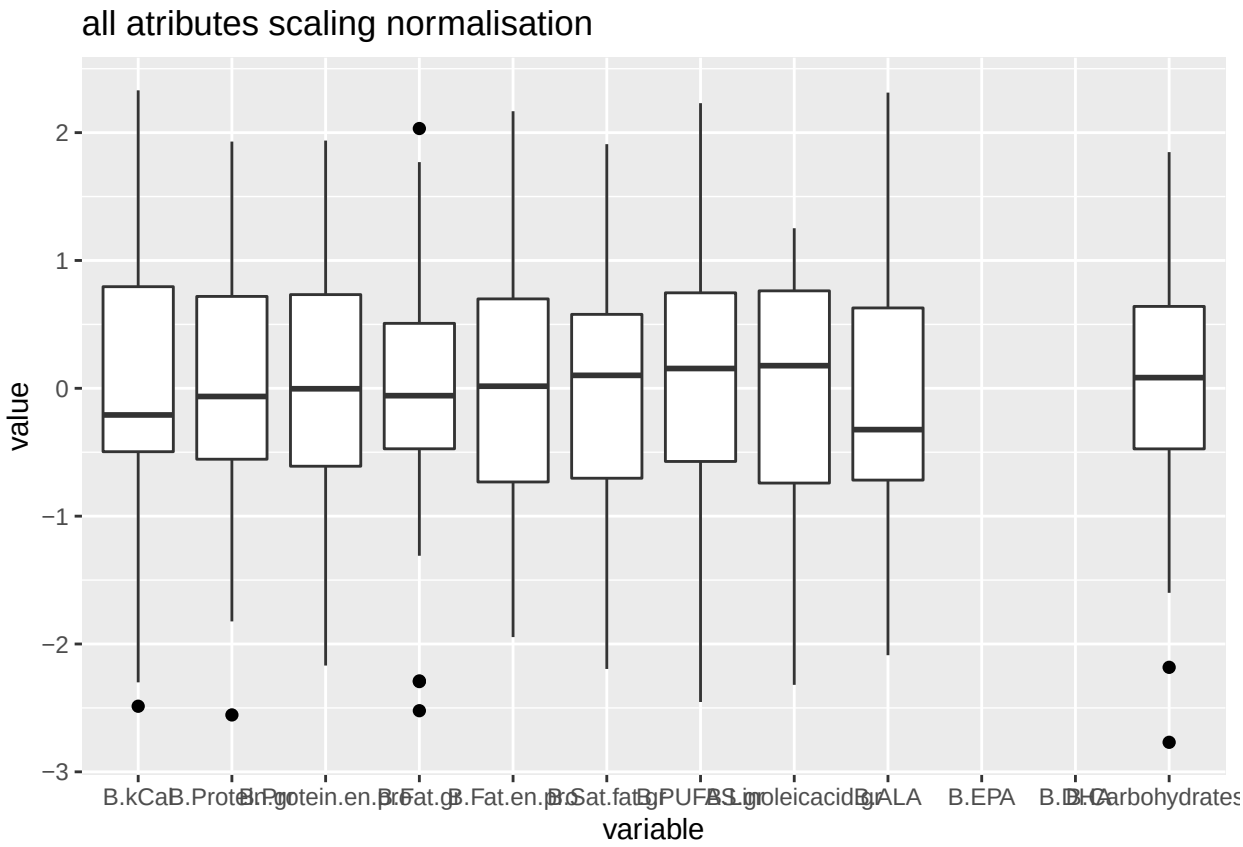
```
FALSE No id variables; using all as measure variables
```




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




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


all atributes scaling normalisation
value
variable
2
1
0
−1
−2
−3
B.kCal
B.Protein.g
B.Protein.en.perc
B.Fat.g
B.Fat.en.perc
B.Sat.fat.g
B.PUFA.g
B.Linoleicacid.g
B.ALA
B.EPA
B.DHA
B.Carbohydrates