Lab Journal

Rienk Heins

9/11/2020

Lab Journal

08/09/2020

Decision which project to work on. Discussing the faults in the test EDA provided by the project.

10/09/2020

Unpacking the data in Rstudio, reading the supporting literature delivered with the data. Inspecting the data

11/09/2020

Inspecting the data further, starting on EDA, formulation of research question.

can the "strength" of the allergic reaction be predicted through machine learning with information about the composition of macro nutrients in the persons diet?

Machine Learning

For the algorithm there is a difference in how bad a wrong prediction is, as the possible predictions are no response, partly response and full response. This is because if a full response is predicted for a partly response that is still better than predicting no response. As with a prediction of no response the diet wouldn't be used of course. So when a wrong prediction is made it is also important to look at if the prediction is right in that there will be a response, as a partly response also is preferable for the patient.

Further the algorithm should be most accurate in guessing the diets with response right. As a working diet guessed wrong will simply not be used which is unfortunate but doesn't harm. But a diet that doesn't work guessed as responding will be used and a patient will then follow this diet with no effects. Not that this will harm the person but it still is a waste of time. Because of this the sensitivity or as it is called in the weka outcome recall will be used as the quality metric.

```
library(knitr)
classifier <- c("ZeroR", "OneR", "Naïve Bayes", "Simple Logistics", "SVM", "Nearest Neighbor(IBK)", "J4
accuracy <- c("41", "29", "29", "35", "35", "47", "32", "41")
TP_rate <- c("0,412", "0,294", "0,294", "0,353", "0,353", "0,471", "0,324", "0,412")
FP_rate <- c("0,412", "0,394", "0,369", "0,386", "0,370", "0,262", "0,357", "0,303")
recall <- c("0,412", "0,294", "0,294", "0,353", "0,353", "0,471", "0,324", "0,412")
weka_table <- data.frame("classifier" = classifier, "accuracy" = accuracy, "TP rate" = TP_rate, "FP rat
kable(weka_table)
```

classifier	accuracy	TP.rate	FP.rate	recall
ZeroR	41	0,412	0,412	0,412
OneR	29	0,294	0,394	0,294
Naïve Bayes	29	0,294	0,369	0,294

classifier	accuracy	TP.rate	FP.rate	recall
Simple Logistics	35	0,353	0,386	0,353
SVM	35	0,353	0,370	0,353
Nearest Neighbor(IBK)	47	0,471	0,262	0,471
J48	32	0,324	0,357	0,324
Random Forest	41	0,412	0,303	0,412

One of the algorithms used for optimization will be IBK as it not only is the most accurate, it's mostly accurate on no response and partly response and puts most wrongly guessed full response instances as partly response, which is a better mistake than putting it as no response as stated in the second part of the machine learning log.

Further