Logbook What urinary biomarkers in combination with PanRISC are most accurate at predicting pancreatic cancer

Nils Mooldijk

10/12/2021

```
#loading in the clean output data from the EDA.
data <- read.table("data/PDAC_cleaned_data.csv", header = TRUE, sep = ",")
```

Log

Weka log

There was some trouble loading the data into WEKA. This was because a 0.00 value was fed to log(2) which created a '-INF' value. This value isn't numerical which made the data incompatible with WEKA. To solve this, an extra ifelse() check has been added to the code section which log transforms certain values.

```
log_example <- data
#OLD
log_example[, 3] <- log(log_example[, 3], 2)
#NEW
log_example[, 3] <- ifelse(log_example[, 3] != 0, log(log_example[, 3], 2), 0)
```

In order to get the most accurate model, a lot of algorithms and settings have to be tried out:

J48, 10 cross-validation folds:

=== Stratified cross-validation === === Summary ===

Correctly Classified Instances 410 69.4915 % Incorrectly Classified Instances 180 30.5085 % Kappa statistic 0.5418 Mean absolute error 0.2232 Root mean squared error 0.4222 Relative absolute error 50.2892 % Root relative squared error 89.6255 % Total Number of Instances 590

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,710	0,123	0,722	0,710	0,716	0,590	0,819	0,630	1
0,668	0,162	0,692	0,668	0,680	0,510	0,787	0,651	2
0,709	0,174	0,675	0,709	0,691	0,528	0,799	0,602	3

Weighted Avg. 0,695 0,154 0,695 0,695 0,695 0,541 0,801 0,628

=== Confusion Matrix ===

a b c <– classified as 130 29 24 | a = 1 25 139 44 | b = 2 25 33 141 | c = 3

J48, 50 cross-validation folds:

=== Stratified cross-validation === === Summary ===

Correctly Classified Instances 415 70.339 % Incorrectly Classified Instances 175 29.661 % Kappa statistic 0.5544 Mean absolute error 0.2056 Root mean squared error 0.4093 Relative absolute error 46.3181 % Root relative squared error 86.881 % Total Number of Instances 590

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,710	0,113	0,739	0,710	0,724	0,604	0,838	0,686	1
0,663	0,165	0,687	0,663	0,675	0,503	0,806	0,660	2
0,739	0,169	0,690	0,739	0,714	0,561	0,833	0,668	3

Weighted Avg. 0,703 0,150 0,704 0,703 0,703 0,554 0,825 0,671

=== Confusion Matrix ===

a b c <– classified as 130 31 22 | a = 1 26 138 44 | b = 2 20 32 147 | c = 3