Team Meeting

16 SEPTEMBER 2022 / 1:30 PM / Zoom

Attendees

Ni, Stefan, Xavier, Ate, Zexi, Jiadong

Agenda

Latest results from 833 using the new approach - Ni

Spartan - Xavier, Stefan, Zexi

Report - Ate

Checklist

To Do

Latest results - Ni

Note:

- Missing values in both training and testing datasets are replaced by mean;

- Haven't removed incorrect sample in the Denmark dataset

Initial Approach - Using E-Risk as training data				
AUC	Stacking	Voting	PCA	Nature
E-Risk	0.8516	0.8357	0.8127	0.739
BSGS	0.8092	0.8031	0.7612	0.774
Denmark	0.6301	0.6587	0.629	0.563
E-MTAB	0.6356	0.6782		0.522
AMDTSS	0.7173	0.7139		0.648

Final Approach - 4 as training and 1 as testing				
AUC	rf	lr	training_rf	training_lr
E-MTAB	0.671	0.6775	0.7966	0.8073
AMDTSS	0.6373	0.6913	0.8117	0.8425
DENMARK	0.6392	0.6144	0.799	0.8538
BSGS	0.7869	0.8229	0.7977	0.8342
E-Risk	0.6823	0.7146	0.8049	0.8764

Spartan - Zexi, Xavier, Stefan

- All 5 datasets has been converted to the same format
- Name for the datasets: “xxx_ALL.csv”
- Row name: CPG, Zygosity
- Zygosity: “MZ”, “DZ”
- All samples with missing / incorrect labels are removed

Report - Ate

Checklist

1. I’ve filled in the missing values in training data by their column means, should I do the same for testing or should I drop them?
2. Should we still consider using a stacking / voting model? If so, should they be built on the selected variables (either by rf or lr)?
3. As 5 different training methods (4 training and 1 testing repeated for 5 times) would give 5 different sets of selected variables. Do we provide an intersection of those 5 as our final selected variables?

   If so, does this violate our rule of not using the testing data to select variables?

   If not, do we provide all 5 different sets of selected variables corresponding to each training method?

To Do

1.