Meeting with annie 13\_11\_2020

Pvalues:Chi2 test: have y scale as logarithmic instead of rescaling all values as log10

DO SPECIFY the threshold

Do backward feature selection recursive instead of comparing the 3 scores

RR: could still be important because you want to discriminate

Do HEAD to HEAD: bivariate vs multivariate! Tells you how much interaction there is (sort of) between variables!

Should drop MI test no one uses it

Draw conclusions on multi vs bivariate analysis!!!!!! Do keep bivariate to show why it’s important

Confounding OR interaction / compounding

What about FEATURE IMBALANCE IMPORTANCE!!!! Not enough nuance, a lot of incentive to MISCLASSIFY!!!

ALSO: binary variables:….. they’re not great….a “change” = going from 0 to 1 or the other war around.

LIME: drop the “negative” probability, you only care about positivity….

In the graph: make feature names bigger, remove the <= threshold; just make it more CLEAR!

USUALLY if its binary variable 1 = indicates presence!

Usually two seasons: dry, rainy; but in Tanzania: post rainy is floody, marshy… but rainy = too rainy for malaria to grow in. Mosquitoes are sensitive to elevation (they don’t like elevation) but they like a certain temperature (12 degrees and up) and definitely do like humidity/rain

1. Dry
2. Rainy
3. Post-rainy

Dataset clustering: what about features that are not relevant but used in the data point proximity algorithm (clustering? Then it would impact the groups…. And that’s not great). At the same time, is the dataset going to contain features that are totally irrelevant in the clustering of patients. Or maybe you should cluster manually based on features such as age and location before doing the surrogate model thing.

Do the LIME thing using SHAP

Confidence interval on the predictions themselves! So 0.64 chance of being malaria positive +/- whatever

Message Ridha

Message mid-term presentation