Project 1

A company intends to construct a diagnostic tool to detect individuals who were infected by Plasmodium vivax (Pv) in the last 6 months. The diagnostic tool will be based on antibody measurements ONLY. With the purpose, a study was conducted in Brazil (data_project_1_brazil.csv). The data dictionary is the following:

- Age (in years)
- sex (male, female)
- T_last_Pv (time to last Pv infection in days) – in this variable, consider NA as individuals who had a Pv infection a long time ago, 0 means that an individual was currently infected at the time of sampling.
- W1, …., W60 are antibody measurements of 60 different antibodies

Create a new variable indicating whether an individual had or not an infection in the last 6 months.

Before conducting any formal analysis, present and describe the data with appropriate statistics and plots.

Construct a diagnostic model based on the antibody data to predict the new variable. Use any feature selection that you find appropriate. Then, evaluate the overall performance of your diagnostic model by calculating the ROC curve and AUC. Decide the optimal sensitivity and specificity for your diagnostic model. Justify your decision. Is your model fair in terms of sex of the individuals? Is your model (diagnostic tool) fair in terms of age groups (children 0-5 years old, adolescents - 6-18 years old, adults - >18 years old)? To answer these two last questions, use a statistical test if you find appropriate.

The company conducted a follow-up study in Thailand to validate your results (data_project_1_thailand.csv), this file has the same data dictionary as data_project_1_brazil.csv. Apply your diagnostic model to this new data set. What is the respective performance? Do you have statistical evidence (apply a statistical test) that the diagnostic model has the performance in both data sets?

What are your overall conclusions?

Important:

Prepare a 15-minute presentation with your main findings. There will be a penalty of 0.5 points in your project grade if you exceed the time for your presentation. Upload your R script/R Markdown for code verification. Also upload your presentation as a pdf file. Failure

to upload these files before classroom evaluation leads to a penalty of 0.5 points in your project grade.

Reference (for a similar context only):

Longley RJ, White MT, Takashima E, et al. Development and validation of serological markers for detecting recent Plasmodium vivax infection. Nat Med. 2020;26(5):741-749. doi:10.1038/s41591-020-0841-4