Deliverable 2

Predicting COVID-19 Test Result from Symptoms and Comorbidities

Group Members

Runsheng Wang

Jinqi Lu

Data collection:

- **corona_tested_individuals_ver_006.english.csv**

  data from Israeli, feature contains test date, cough, fever, sore throat, shortness of breath, headache, test result, age 60 above, and gender. 278848 cases included.

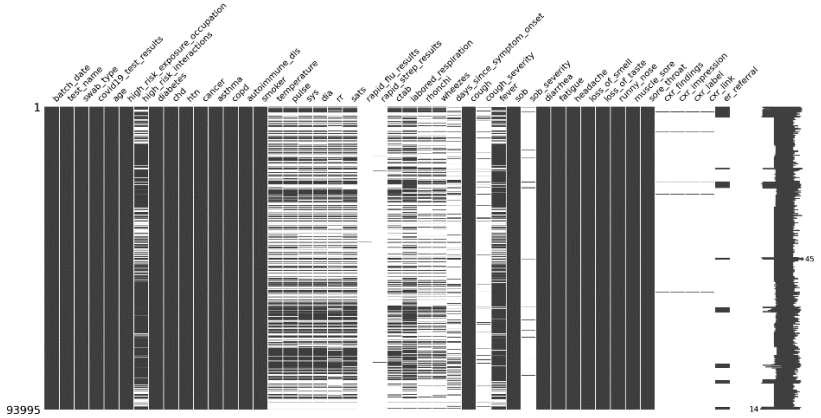
- **carbonhealth_and_braidhealth.csv**

  covid-19 test data with more features in a weekly basis. Features included: batch date, test method, swab type, test result, age, high risk exposure occupation, high risk interactions, diabetes, chd, htn, cancer, asthma, copd, autoimmune dis, temperature, pulse, sys, dia, rr, sats, rapid flu results, rapid strep results, ctab, labored_respiration, rhonchi,wheezes, days_since_symptom_onset, cough, cough_severity, fever, sob, sob_severity, diarrhea, fatigue, headache, loss_of_smell, loss_of_taste, runny_nose, muscle_sore, sore_throat. There are 29 files with 96493 cases included.

Preliminary analysis:

By comparing two datasets, the first impression is the Israeli dataset has significant higher number of cases (~278k) while the carbon dataset has only 96k.

The carbon dataset has more features, but some of them contains primarily NaN values.



The Israeli dataset, however, has no NaN values at all.

Since both datasets only has a limited number of common features, it is hard to merge them together, and we have to focus on only one.

One Key Question:

Which dataset can be used to create the model?

After carefully consideration, we decide to use the carbon dataset since it contains more features which brings us more flexibility. In order to utilize it, more preprocessing is required like over and under sampling or feature selection.

New Limitations:

There are lots of NaN values in some of the features in the carbon dataset. It is hard to say if we should get rid of those rows, filling them with some values, or simply remove the entire column from feature selection. If any not those step is not properly done, the final model will not be the most fit one.