Deliverable 1

Predicting COVID-19 Test Result from Symptoms and Comorbidities

Group Members

Runsheng Wang

Jinqi Lu

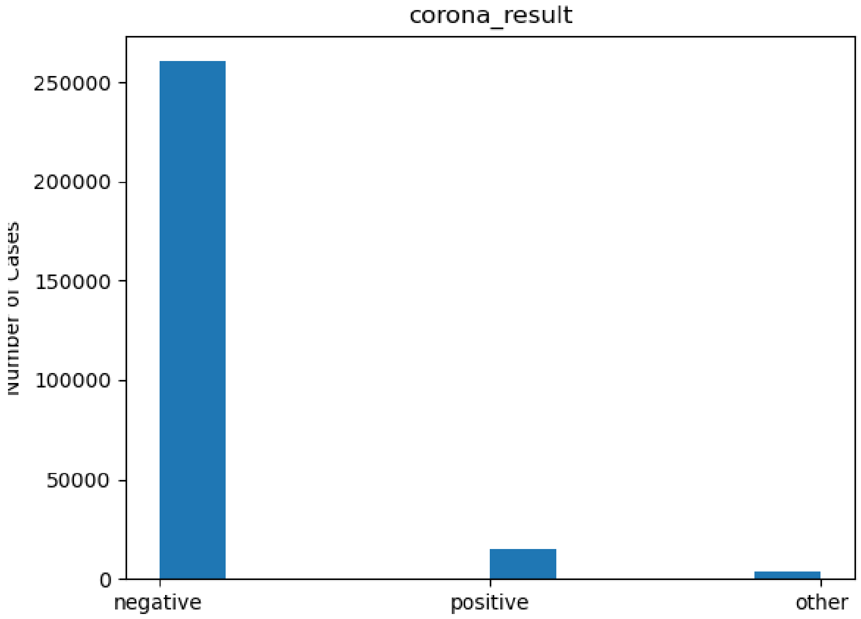
Data collection:

- corona_tested_individuals_ver_006.english.csv

Preliminary analysis:

The first impression is the dataset has about 278k cases which is a lot. The dataset is from Israeli, feature contains test date, cough, fever, sore throat, shortness of breath, headache, test result, age 60 above, and gender. Although it contains a large number of cases (278848), the feature is relatively limited, which brings us less flexibility.

Also, this is an imbalance dataset with about 19k positive and 260k negative. This will bring us new trouble when creating the model.



One Key Question:

Which dataset can be used to create the model?

The dataset is good, it contains less missing values with a significant number of cases. We can use this dataset, however, since the number of features is limited, we'd better find another source if possible.

New Limitations:

The number of negative results is way too larger than the positive result; this will create problem when applying some model.