Predicting COVID-19 Test Result from Symptoms and Comorbidities	
Contact	Name: Runsheng Wang, Jinqi Lu Email:runsheng@bu.edu, jinqilu@bu.edu Cell Phone: N/A,
Organization	N/A
Organization Description	N/A
Project Type	Data Science
Project Description	We seek to predict COVID-19 test result from symptoms and comorbidities. We gather covid-19 test data along with their symptoms from the internet and choose which to use. We believe there is a relationship between specific symptom(s) and the possibility of the test result. We will try to unveil the symptoms that contribute most to the test result and build model to predict the result based on selected symptoms.
Data Sets	● Ccarbonhealth_and_braidhealth dataset from ORGANIZATION OR SOURCE ● Israeli dataset from ORGANIZATION OR SOURCE
Suggested Steps	We will first download preprocess dataset, dropping invalid values columns or rows. We will then perform feature selection, under sampling, and oversampling. We will feed the processed data into following models and use the best one: ● KNN ● Logistic ● Complement Naïve Bayes ● Decision Tree ● Random Forest
Questions to be answered in Analysis	# Very specific questions that the clients wants answered ● **What symptoms or comorbidities has a relationship with COVID-19?** ○ Preprocess data, handle N/A case properly. ○ Feature Selection Algorithm like chi2 ● **Will the patient test positive or negative based on specific symptoms?** ○ Various model like KNN, Logistic, Decision Tree, Complement Naïve Bayes, and Random Forest. ● **Which dataset can be used?**

Additional Information	

Limitations:

Multiple datasets contain different symptoms, merging multiple datasets will lead to losses of feature.
Patient data and test result are private information, and it is not always available to the public.
Some people may have different standard for symptoms like muscle sore or fatigue, and some feature may not be objective.
The test results are from different method like RT-PCR or qPCR, their accuracy is not always consistent.
The number of negative results is way too more than the positive result (100:1), this will create problem when applying some model.