Project Proposal for COVID19 prediction

1. **Questions**
2. **Hypothesis**
3. **Approach**

You will prepare a project proposal detailing the questions we are wanting to answer. The initial hypotheses about the data relationships and the approach you will take to get your answers.

- Proposal is just a plan.
- End goal is important

Section 1: Questions to Answer

What questions do you want to answer?

1. Why is your proposal important in today's world? How predicting a disease accurately can improve medical treatment?
2. How is it going to impact the medical field when it comes to effective screening and reducing health care burden.
3. If any, what is the gap in the knowledge or how your proposed method can be helpful if required in future for any other disease.

Section 2: Initial Hypothesis (or hypotheses)

1. Here you have to make some assumptions based on the questions you want to address based on the DA track or ML track.
    1. If DA track please aim to identify patterns in the data and important features that may impact a ML model.
    2. If ML track please perform part 'i' as well as multiple machine learning models, perform all required steps to check if there are any assumptions and justify your model. Why is your model better than any other possible model? Please justify it by relevant cost functions and if possible by any graph.

2. From step 1, you may see some relationship that you want to explore and will develop a belief about data.

Section 3: Data analysis approach

1. What approach are you going to take in order to prove or disprove your hypothesis?
2. What feature engineering techniques will be relevant to your project?
3. Please justify your data analysis approach.
4. Identify important patterns in your data using the EDA approach to justify your findings.

Section 4: Machine learning approach

1. What method will you use for machine learning based predictions of COVID19?
2. Please justify the most appropriate model.
3. Please perform necessary steps required to improve the accuracy of your model.
4. Please compare all models (at least 4 models).

Machine learning-based prediction of COVID-19 diagnosis based on symptoms

A speedy and accurate diagnosis of COVID-19 is made possible by effective SARS-CoV-2 screening, which can also lessen the burden on healthcare systems. There have been built prediction models that assess the likelihood of infection by combining a number of parameters. These are meant to help medical professionals all over the world treat patients, especially in light of the scarcity of healthcare resources. The current dataset has been downloaded from ‘ABC’ government website and contains around 2,78,848 individuals who have gone through the RT-PCR test. Data set contains 11 columns, including 8 features suspected to play an important role in the prediction of COVID19 outcome. Outcome variable is covid result test positive or negative. We have data from 11th March 2020 till 30th April 2020. Please consider 11th March till 15th April as a training and validation set. From 16th April till 30th April as a test set. Please further divide training and validation set at a ratio of 4:1.

- Please perform all appropriate feature engineering tasks.
- Perform important data visualization techniques to find the pattern in data.
- Report characteristics of important features, such as total number and percentage of each category in a table format after performing all relevant tasks.

- Perform multiple machine learning models relevant to your hypothesis, justify your model.
- Perform important cost functions to justify which model is better.

The following list describes each of the dataset’s features used by the model:

A. Basic information:

1. ID (Individual ID)

2. Sex (male/female).

3. Age ≥60 above years (true/false)

4. Test date (date when tested for COVID)

B. Symptoms:

5. Cough (true/false).

6. Fever (true/false).

7. Sore throat (true/false).

8. Shortness of breath (true/false).

9. Headache (true/false).

C. Other information:

10. Known contact with an individual confirmed to have COVID-19 (true/false).

D. Covid report

11. Corona positive or negative

Dataset: corona_tested_006