CSE 375/381 Project: COVID-19 Outcome Prediction

Project Description

The data used in this project will help to identify whether a person is going to recover from coronavirus symptoms or not based on some pre-defined standard symptoms. These symptoms are based on guidelines given by the World Health Organization (WHO).

This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.

imp

The data is available from 22 Jan, 2020. Data is in "data.csv".

The dataset contains 14 major variables that will be having an impact on whether someone has recovered or not, the description of each variable are as follows,

1. *Country*: where the person resides
2. *Location*: which part in the *Country*
3. *Age*: Classification of the age group for each person, based on WHO Age Group Standard
4. *Gender*: Male or Female
5. *Visited_Wuhan*: whether the person has visited Wuhan, China or not
6. *From_Wuhan*: whether the person is from Wuhan, China or not
7. *Symptoms*: there are six families of symptoms that are coded in six fields.
13. *Time_before_symptoms_appear:*
14. *Result*: death (1) or recovered (0)

It is required to design different classifiers to the predict the outcome (death/recovered) when a new person is admitted to the hospital. The data is already cleaned and preprocessed.

You will have to divide the data into three partitions: training, validation, and testing. You need to design the following classifiers:

1. K-Nearest Neighbors
2. Logistic Regression
3. Naïve Bayes (due end of week 11)
4. Decision Trees
5. Support Vector Machines (due end of week 14)

For each classifier, try to find the optimal hyperparameters.

You also need to compare the performance of all classifiers using different metrics such as the precision, recall, F1-score, and ROC/AUC curves.