CS559: Machine Learning: Fund & Apps
Spring 2020 - Semester Project

\ Introduction:

The World Health Organization (WHO) was alerted to several cases of pneumonia in Wuhan City, Hubei Providence of China. The virus did not match with any other known virus. The new virus called COV19 raised concern as it resulted in nearly 704,095 positive tests, a total of 33,509 death, and only 148,824 people were reported recovery. The details of confirms, deaths, and recovers can be seen from the visual dashboard made from John's Hopkins University.

As of today (3/29/2020), the United States reported the highest number of confirms and the number of deaths raised exponentially. The area of New York City and North New Jersey said the most number of positives and deaths throughout the nation.

Objective:

As an expert data scientist, you are going to analyze the given COV19 Data from Kaggle to answer the following objectives:

1. Predict the spread of coronavirus - The outbreak of COV19 is developing into a major international crisis, and it's starting to influence important aspects of daily life. Develop a model that predicts the progression of the virus throughout March 2020 and how the virus could spread across different countries and regions may be able to help mitigation efforts.
2. Predict how the epidemic will end - You are going to develop a model that predicts how people are going to recover based on old recovery records.
3. The correlation between growth rate and types of mitigation across countries - The task is to evaluate the effectiveness of mitigation by trying to see if a correlation can be discovered between the different types of mitigation and the growth rate of confirmed cases.
4. The confirmation of correlation between weather conditions and coronavirus spread - the key question is whether the weather conditions and COV19 are correlated or not. From this investigation, we can see how our summer will be.
5. Develop a sophisticated question of coronavirus that might shed light on the understanding/expectation of the coronavirus.

Data:
Required Data: The data source has been recently updated as the date of 3/28/2020. You are welcome to use the updated data.

The required dataset is the dataset from COV19 Data from Kaggle. The datasets you can download are good enough for objectives #1 to #4. However, the use of any open datasets from any organizations and/or research institutes for your study is welcome.

For objective #5, this requires creativity more than any other objective. You are welcome to have any sophisticated questions you want to investigate/answer your own questions. You can do this only using the given data sets. However, this project is not limited to use only the given datasets. You are welcome to use any kind of open-source datasets from any organization and/or research institutes for your study.

CS559: Machine Learning: Fund & Apps
Spring 2020 - Semester Project

Submission:
The project is due 4/30th by 11:59 PM.

You are going to submit a research type report, which includes the abstract, introduction, data description, methodology on data preprocessing and modeling, results, discussion and conclusion, and reference. The methodology and results must be explained for each objective.

In the report, you are welcome to present any kinds of figures, graphs, and tables. Here we have the details of result submissions.
1. The studies of the U.S. is must, study details on at least three countries from each continent.
2. Objective #1: Must contain a table that summarizes the data for your selected countries and a table of accuracy.
3. Objective #2: Must have a table that presents the numerical value of prediction. It can be probabilities, the mean square errors, etc.…
4. Objective #3 & #4: Present the numerical value of correlations.
5. Objective #5: You have all your freedom of presentation.
6. As creativity is a crucial part of storytelling, and we tell stories with the data, we all have the freedom to include many plots, tables, etc.…
7. When you explain the model, make sure to include the choice of techniques and the reasons. For example, if you are going to use SVM, you should explain why you decided to use SVM, advantage, and disadvantage of SVM, etc.…

Video Presentation:
You are going to submit a 10 to 15 minutes long presentation. You can simply record it as you have done throughout the semester. If the size of the video is significant, you can submit the link.

Code:
It must be your own.

Limitations of outside resource:
ML Techniques - you are welcome to use any including techniques we have not covered. Although our datasets are labeled, you can use unsupervised learning if you would like to.
Datasets - You are welcome to use any as long as the database has credentials.
Kaggle Kernels & Discussions - The participants of these projects in Kaggle communicate with each other. You are welcome to do so. However, you must not steal their ideas, comments, codes.