BDAT 1008 – Assignment 2

The COVID19 pandemic has been devastating for hospitals as limited resources can be stretched. One area of work that is being investigated is the use of simulators that can determine the number of active cases likely to happen at a hospital – for example http://covid19simulator.ca/

This simulator uses machine learning algorithms to predict number of patients that may possibly enter into a hospital in the process helping hospitals predict their resource needs. In this assignment, you will try to run a machine learning algorithm in Spark that predicts fatalities.

The following is a dataset from the city of Toronto on COVID19 cases

https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Using the above dataset, write a Spark machine learning algorithm in order to predict the fatality rates in the Toronto area.

Note that since majority of COVID-19 cases result in recovery, this ***dataset is not balanced***. For example, if you have an algorithm that simply makes all cases as "resolved", then it would be 99% accurate (since 99% of the cases are "resolved") even though it did not predict a single fatality correctly! As a result, you cannot use the dataset as is and must balance the dataset. We did not go over the concept of balancing in the Spark Machine Learning lessons but you should have been exposed to this concept in other courses. You will therefore need to investigate how to balance a dataset in Spark. Here is an example tutorial with random forest.

https://www.linkedin.com/pulse/multi-class-classification-imbalanced-data-using-random-burak-ozen/

Once you have a balance dataset, you can run your algorithm on the balanced dataset and report your accuracy.

Deliverables

1. Your machine learning algorithm code in Spark (as a simple text file)

2. Formal report on your findings