## Part A: Data Preprocessing

- Label chosen to be month
- Table containing : City, Temperature, Gender, Age Group , Month , Resolved, Fatal
- Need to first one-hot encode the categorical columns
- Categorical columns found out to be : City, Gender, Age Group
    - Using pd.get_dummies command in pandas was able to encode these columns
- Next need to normalize other numerical columns, in this case “Tempreture”
    - Normalized temperature values to be between 0 and 1
- Remaining columns are Month ( which is to be used as the label) ,Resolved, and Fatal
    - Converted Resolved, Fatal to numerical values (i.e. True = 1)
- Found there was missing data in the temperature column
    - Filled missing values with 0

## Part B: Classification

## B.2

| | Accuracy | Precision | Recall | Time to construct of model(seconds) |
|---|---|---|---|---|
| Decision Tree | 0.912372234935164 | 0.9149541214458579 | 0.912372234935164 | 0.14775780297350138 |
| Gradient Boosting | 0.8363081617086193 | 0.8385215982173284 | 0.8363081617086193 | 21.951980715035546 |
| Random Forest | 0.9018459191456903 | 0.9038207360037677 | 0.9018459191456903 | 0.4420440249959938 |

We would rank the quality of the models produced by the three algorithms by comparing the predictive accuracy and speed of construct the model. Though measuring the accuracy, we can see there are high accuracy, high precision, and high recall in these three algorithms. Especially, the decision tree algorithm has around 91% of accuracy, precision, and recall. Also, the f-measure is high in each algorithm model which means most of the instances been classified correctly and it does not miss a significant number of instances. By comparing the time to construct of model, from the table it indicates that the Decision Tree algorithms use the least number of seconds among to others. Random forest is faster than gradient boosting. The main difference between Random Forest and Gradient Boosting is random forests build each tree independently while gradient boosting builds one tree at a time. For robustness, there are some missing values in the datasets, therefore, random forest result in better performance than random forests in handling noise and missing values. Overall, the Decision Tree algorithms rank first among these three algorithms, the second one is random forest, and the third one is Gradient boosting.

## B.3

From this data we can conclude that predicting the month that a case was from given the selected features easy very feasible. We find out that temperature was the most prominent feature also the locations was also used a bit. The month and the temperature are high correlation. In model, the city Toronto have more feature than other cities. This may due to the population density in city Toronto is higher than other locations. We indicate that from September to December there are more and more positive cases. Therefore, in the quarter 4, the temperature getting lower, the Covid-19 virus spreads faster.