Milestone 2 Report

2.1) Data Splitting

Before splitting the dataset, we decided to encode categorical data to make sure every attribute in the dataset is numerical for calculation purposes. To encode data, we used 'LabelEncoder' and 'OneHotEncoder' functions from sklearn library. Label encoder was used to encode date_confirmation data, and one hot encoder was used to encode sex, country, and province attributes because we did not want to have any correlation in our data which is why we used one hot encoder. Then for future testing purposes, we used the 'train_test_split' function from sklearn library to split our data into 80% training data, and 20% testing data.

2.2) Build Models

AdaBoost Model: We decided to go with the AdaBoost model because it was one of the main options regarding boosting trees. However, it turned out that the test and training accuracies were pretty low even in a 10-fold cross validation run on the training data, it resulted in an average of ~47.38% which is very low compared to our other models. If we had the choice to go back, we would try another model instead.

Random Forests

Random forest usually beats other tree classifiers when it comes to large datasets, and it performs pretty well regarding overfitting of the data. This classifier was our best model. It resulted in 10-fold cross validation of 81% on the training data.

Out of the two models above, RandomForest classifier resulted in better accuracy scores for all the outcome labels such as decreased, hospitalized, etc.

2.3) Evaluation

```
>> Ada Boost Model:
Train Accuracy: 0.48634954277695214
Test Accuracy: 0.48509736388925684
10-fold CV for the model:
[0.48718798 0.48048185 0.48532516 0.44239765 0.44372232 0.49182432
 0.53760815 0.48439311 0.47967379 0.40627587]
CV Accuracy for the model: 47.38890190949704
--- Classification report for train data:
                precision    recall  f1-score   support

      deceased       0.00      0.00      0.00      3084
   hospitalized       0.58      1.00      0.73     96837
nonhospitalized       0.99      0.28      0.43     74274
     recovered       0.39      0.00      0.01     67372

      accuracy                           0.49    241567
     macro avg       0.49      0.32      0.29    241567
  weighted avg       0.64      0.49      0.43    241567

--- Classification report for test data:
                precision    recall  f1-score   support

      deceased       0.00      0.00      0.00       725
   hospitalized       0.58      1.00      0.73     24240
nonhospitalized       0.99      0.27      0.43     18512
     recovered       0.45      0.00      0.01     16915

      accuracy                           0.49     60392
     macro avg       0.51      0.32      0.29     60392
  weighted avg       0.66      0.49      0.43     60392
```

Fig 1. Evaluation of AdaBoost Classifier

```
ADA Boost
    learning rate: 1, train: 0.705, test: 0.860
    learning rate: 3, train: 0.647, test: 0.610
    learning rate: 5, train: 0.565, test: 0.660
    learning rate: 7, train: 0.367, test: 0.720
    learning rate: 9, train: 0.477, test: 0.790
    learning rate: 11, train: 0.703, test: 0.710
    learning rate: 13, train: 0.365, test: 0.710
    learning rate: 15, train: 0.703, test: 0.710
    learning rate: 17, train: 0.703, test: 0.710
    learning rate: 19, train: 0.680, test: 0.710
```

Fig 2. Scores for different hyperparameters (2.4)

```
>> Random Forst Model:
Train Accuracy: 0.9807837991116336
Test Accuracy: 0.8121605510663664
10-fold CV for the model:
[0.80949621 0.80870969 0.80933063 0.81388417 0.81127623 0.8086269
 0.80705386 0.80803941 0.81068886 0.81706408]
CV Accuracy for the model: 81.04170045632723
--- Classification report for train data:
                precision    recall  f1-score   support

      deceased       0.95      0.93      0.94      3084
   hospitalized       0.97      0.99      0.98     96837
nonhospitalized       1.00      1.00      1.00     74274
     recovered       0.98      0.95      0.97     67372

      accuracy                           0.98    241567
     macro avg       0.97      0.97      0.97    241567
  weighted avg       0.98      0.98      0.98    241567

--- Classification report for test data:
                precision    recall  f1-score   support

      deceased       0.09      0.08      0.09       725
   hospitalized       0.76      0.80      0.78     24240
nonhospitalized       1.00      1.00      1.00     18512
     recovered       0.70      0.66      0.68     16915

      accuracy                           0.81     60392
     macro avg       0.64      0.63      0.64     60392
  weighted avg       0.81      0.81      0.81     60392
```

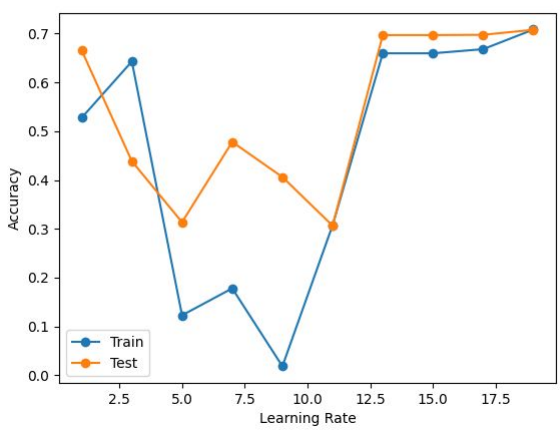
```
Random Forest
    trees: 1, train: 0.924, test: 0.921
    trees: 6, train: 0.969, test: 0.969
    trees: 11, train: 0.985, test: 0.987
    trees: 16, train: 0.990, test: 0.991
    trees: 21, train: 0.994, test: 0.994
    trees: 26, train: 0.995, test: 0.996
    trees: 31, train: 0.997, test: 0.997
    trees: 36, train: 0.997, test: 0.998
    trees: 41, train: 0.998, test: 0.999
    trees: 46, train: 0.999, test: 0.999
```

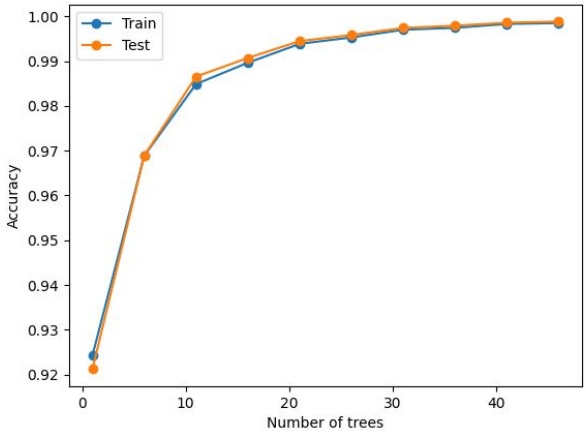
Fig 3. Evaluation of RandomForest Classifier Fig 4. Scores for different hyperparameters (2.4)

To understand the metrics more, please see learning rate and number of trees (i.e., depth) in the overfitting section.

2.4) Overfitting

To check for overfitting, we compared accuracy of the trained models on the training dataset and the test dataset .The models were trained and tested on the same datasets, but with tweaked parameters. For random forests, the number of trees was changed and for ada boost the learning rate was changed. Below are the plotted results. Left figure is the ada boost classifier, right figure is the random forest classifier.





From looking at the graphs, there seems to be sparse signs of overfitting if any. Accuracy from testing and training data sets are close to each other, leading us to believe that no overfitting may be occurring in the models.