Milestone2

2.1 Splitting dataset

The data is split into train and test datasets with a ratio of 80:20 by applying train_test_split function in library sklearn.model_selection. Random_state is set to fix so that the accuracy would be consistent every time

2.2 Build models

Data modification:

- Before training the model, the preprocessing of data includes converting categorical data to numerical data and removing repeated attributes such as latitude and longitude.
- All three different models would aim to classify each instance into one of the four labels ('deceased': 0'hospitalized': 1, 'nonhospitalized':2, 'recovered':3)

Model selection:

1. AdaBoost 2. XGBoost 3. KNN

2.3 Evaluation

- For scoring, since this is a classification problem, we used the accuracy score function from the scikit-learn package for all three models. In addition, the function will be utilized twice to evaluate both the train and test set.
- For better visualization, we plot the output as a confusion matrix to see the strength and weakness of each model.

AdaBoost model (by Lizhou Ding):

AdaBoost model is used to train the dataset. The train and test accuracy are 0.777627 and 0.777653 without changing any hyperparameters. (default: n_estimators=50). The train and test accuracy are 0.787435 and 0.786898 when changing n_estimators=10.There is negligible difference between test and train accuracy. Also, the difference remains trivial when I tune hyperparameters. The output is shown below in confusion matrix 1

XGBoost model (By Siyuan Wu):

The model employed in this section is the XGBoost model. By default, the train set accuracy is 0.851135, and the test set accuracy is 0.850754. By tuning the max iteration to 100 from 50 and the learning rate to 0.1 from 1, the train set accuracy increases 0.335% and test set accuracy increases 0.272%. Therefore, we say the current iteration is optimum as more iterations would have less gain and can potentially raise the overfitting problem. The output is shown below in confusion matrix 2

KNN model (By ZiZe Zhang):

Last but not least, K nearest neighbor model is implemented in the last section. By default, the train set accuracy is 0.854494, and the test set accuracy is 0.853853. As the model is already having the sign of overfitting, no more effort is made into increasing learning iterations or adjusting the learning rate.

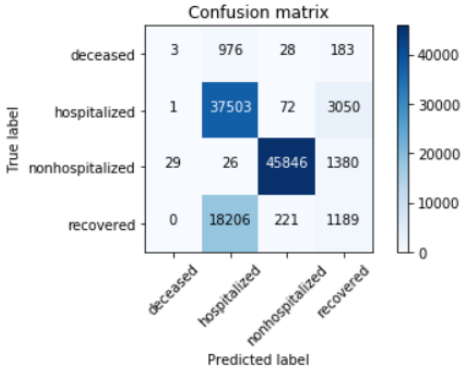

figure 1.

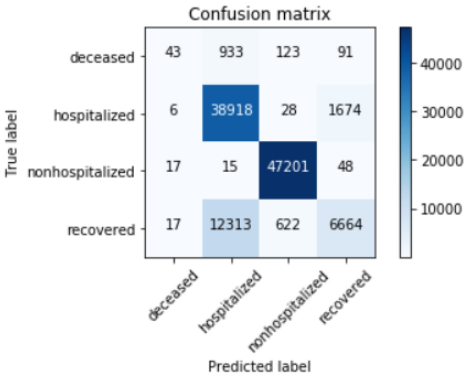

figure 2.

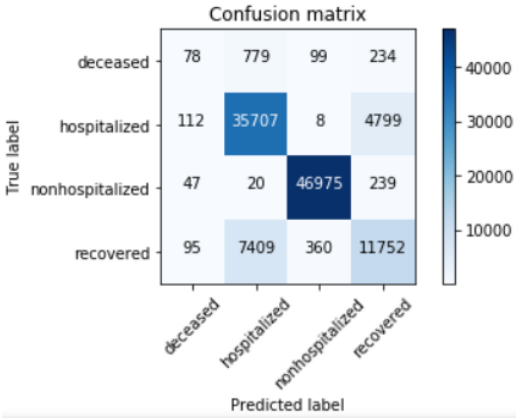

figure 3.

Conclusion: All three models delivered an accurate prediction in terms of hospitalized and non-hospitalized. Among these two outcomes, XGBoost returns the highest true value for hospitalized and non-hospitalized

2.4 Overfitting

In order to observe for overfitting, a threshold of 5% difference between train and test set accuracy is set. It is common for overfitting to occur when the returned train set accuracy is higher than test set accuracy. With all three models implemented, XGBoost and KNN models both have a higher train set accuracy. Since the differences between the training and testing set are 0.04% and 1.1%, which are below the threshold, all three models are considered non-overfitting. In addition, the differences between training and testing sets does not vary vastly with the tuning of hyperparameters including n_estimators and n_neighbors. Hence, overfitting does not apply for all three models.