CS 254 Machine Learning
County Level COVID-19 Prediction

Zahar, Matt
Strawbridge, Parker
Padberg, Isaac

1. Introduction

Our country's response to COVID-19 was varied and not organized. Each state had their own idea of what response was needed, and as a result, some states fared better than others. Because of the different responses and ideologies in each state and county, it is hard for communities in these states and counties to gauge the potential for a COVID-19 outbreak where they live. The problem we are addressing is predicting the number of COVID-19 cases in each county. We plan to inspect New York Times COVID-19 data, mask-use data based on specific counties, county education level data, and census data to predict the number of COVID-19 cases in each county. The results from this model will be helpful for counties to predict the number of positive cases they have and for hospitals, communities, and local governments to make plans of action. It will also inform the general public of the predicted trend in their county. Another step that will be taken is to examine all attributes in the data we have to determine how each of our factors directly affect the number of positive cases and deaths in each county. If prediction models had been used on foreign COVID-19 data prior to the pandemic reaching The United States our national, or local, governments could have been able to predict areas with increased risk and create stronger restrictions earlier leading to fewer deaths. Access to pandemic prediction models, when in the right hands, could save hundreds of thousands of deaths.

2. Problem Definition and Algorithm

2.1 Task Definition

Our inputs will be: percentages of varying levels of mask use, percentages of highest levels of education, county population, county size.
Our output will be the predicted number of confirmed cases of COVID-19 for each county.

With the US's current state being that of divisiveness and hostility, it is very important to understand the full picture of what is going on in each region of the US. The prediction of the number of COVID-19 cases in each county is important because it will help hospitals, city officials, and the general public understand the need for social distancing and mask usage in their respective counties. Additionally, a finely tuned model could help us to predict at such an accurate level that it would be able to perfectly inform counties with the most risk to take greater precautions and ways to reduce their risk. The possibility of an accurate prediction would allow for less testing and provide governments the resources needed to reduce the number of cases in each county. With that being said, there are many other features being ignored, including them could create a better generalizing, more diversified model. This could include: political affiliations, race statistics, average age of population, geographic location, underlying diseases, availability of medical support, and awareness about COVID-19 using web scraping techniques or finding other unique datasets online.

2.2 Algorithm Definition

We looked closely at three different machine learning algorithms in order to diversify our approach to the problem at hand. Each project member focused on a different algorithm with a deep look at Support Vector Regression (SVR), Random Forest Regressor (RFR), and a Multi-layer Perceptron (MLP). The results were examined and compared to each other to determine the regressor with the least error when predicting the number of confirmed COVID-19 cases per county.

After an initial effort to use linear regression as our baseline model, we switched over to a Support Vector Regression. The linear regression model was not able to accurately capture all of the features because the data used was non-linear. The tuning of the SVR was originally done by hand, but then later done using grid search to fully optimize it. We found the optimal C value was 31 and the optimal epsilon value was .3. Using these hyperparameters, the model was trained on 14 features and got an MSE of .57 compared to the linear regression models MSE of 1.8.

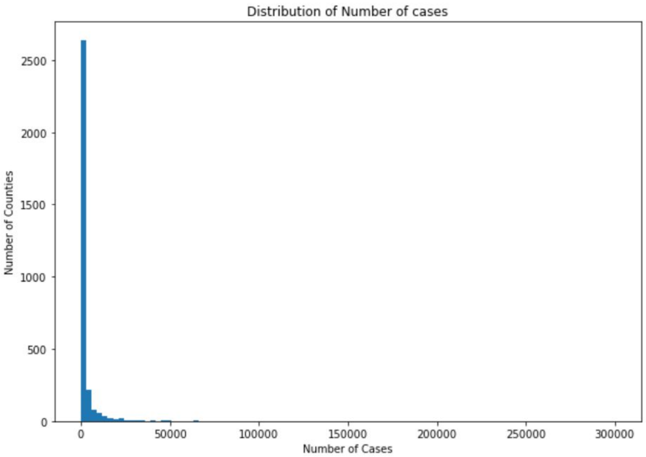
The Random Forest Regressor began by struggling to fit the data due to very noisy data with lots of seemingly useless features. After examining the importance of each feature selected when branching it was found that the population of any given county was the most important feature when predicting the number of confirmed COVID-19 cases. It was very obvious after initial testing that bootstrapping would outperform training on the entire set of data. A final bootstrap sampling size of 75% was used. A total of 120 n_estimators, or trees, were used in the forest. When deciding which features to randomly split on the limit of four random features was found to be the best. The ideal max depth of the tree was found to be 11 nodes deep. With all these hyperparameters tuned using GridSearchCV the final model using the hyperparameters found would average an MSE of about 0.355.

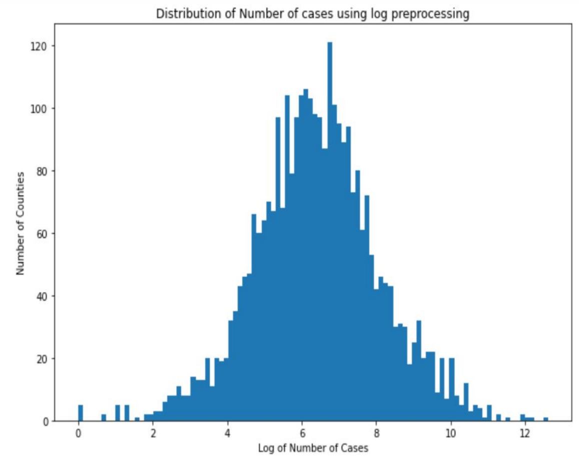
For the Multi-layer Perceptron (MLP) even from the base model where there was only a single hidden layer it immediately showed that it was going to out perform the SVR and Random Forest. The neural network allowed for it to handle the bulk of noisy data that we were feeding our models and allow it to still give a small MSE of about 0.3. My neural network was built out of 8 dense layers each using ReLu activation since that fit our model the best. When building the model the two optimizers that were close in performance were Adam and SGD. Both gave relatively similar performance until the education data was implemented into the model. Once the education data was added the SGD MSE rose to about 0.5 while the MSE for Adam stayed about the same around a 0.23. Which is why in the end the neural network optimizer used Adam. After adding in early stopping and adding more layers to the model the MSE had gotten as low as 0.081 but that was with an R2 value of .68. So by attempting to compromise and get both values to a really strong point I added another layer and increased my early stopping to allow for more data which did raise my MSE to the final value of 0.2 and a final R2 value of .91.

3. Experimental Evaluation

3.1 Methodology

The main evaluator of our model is Mean Squared Error (MSE). Our models are trying to test our expectation that the number of COVID-19 cases will be higher where there is little mask use and high population density while also taking into account population of the county along with levels of education. This is the right way to measure performance because we can directly compare what we are trying to predict with what we have predicted. However, since the data we are working with does not account for people who have not been tested but have COVID-19, the actual number of people who have COVID-19 will be higher than our predictions by an unknown amount. This is true because we are unable to predict based off of data that has not been collected. The training/test data split was chosen by randomly selecting counties to be trained/tested on. This data is realistic because there is a large variation in the population density and mask usage throughout the US. This random split should get enough variation in the training data to mimic the country as a whole so that there is little to no underfitting or overfitting. The performance data we collected is the difference of confirmed cases from predicted cases. This number will show us how far off the predicted number of cases was from the current number of cases, which is a direct measurement of accuracy. The biggest dataset that had a major impact on our model was obviously the dataset with the number of cases.





The main issue as you can see from the graph above is that the case data is severely skewed to the left with outliers up at around 290k which originally gave us an MSE of about 58 million. By taking the log of cases it allowed for our case data to appear on a normal distribution allowing for our MSE to fall below 1 which is a huge increase in precision. Our best MSE that we recorded was 0.198. By comparing the MSE from before and after the log was introduced to the cases you can see exactly why that change needed to be made. It made our data usable.

3.2 Results

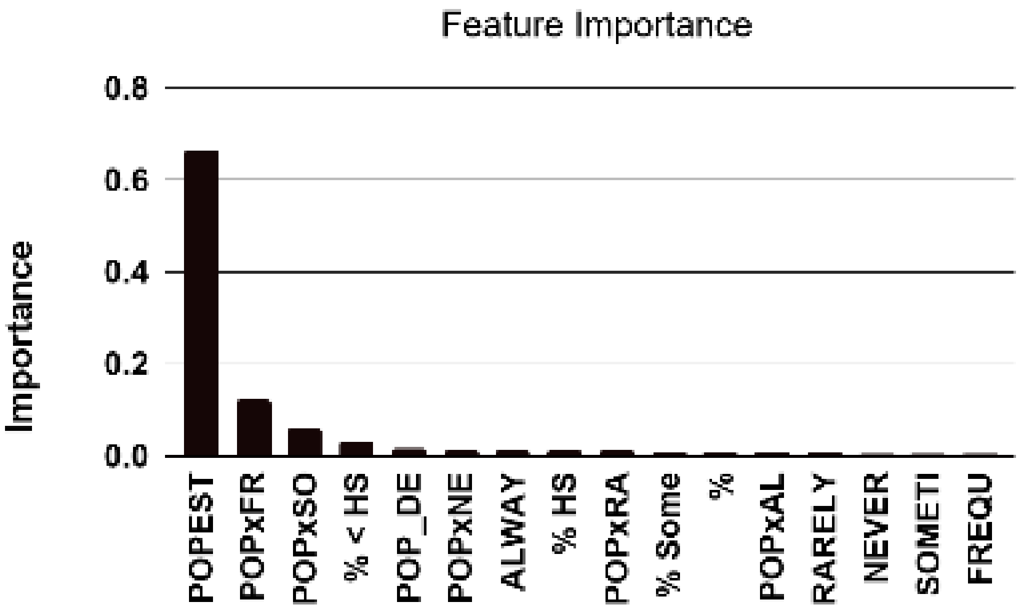


Figure I.
The above figure summarizes the importance of each feature used in our model. The importance is a percentage of how often each of the features are chosen when training the model, most importantly in the random forest. This shows that there is a high importance, heavy weight, on the population feature. The remaining features have little importance and describe the noise of the data. Random forests perform poorly on very noisy data which lead to a higher MSE than the neural network.

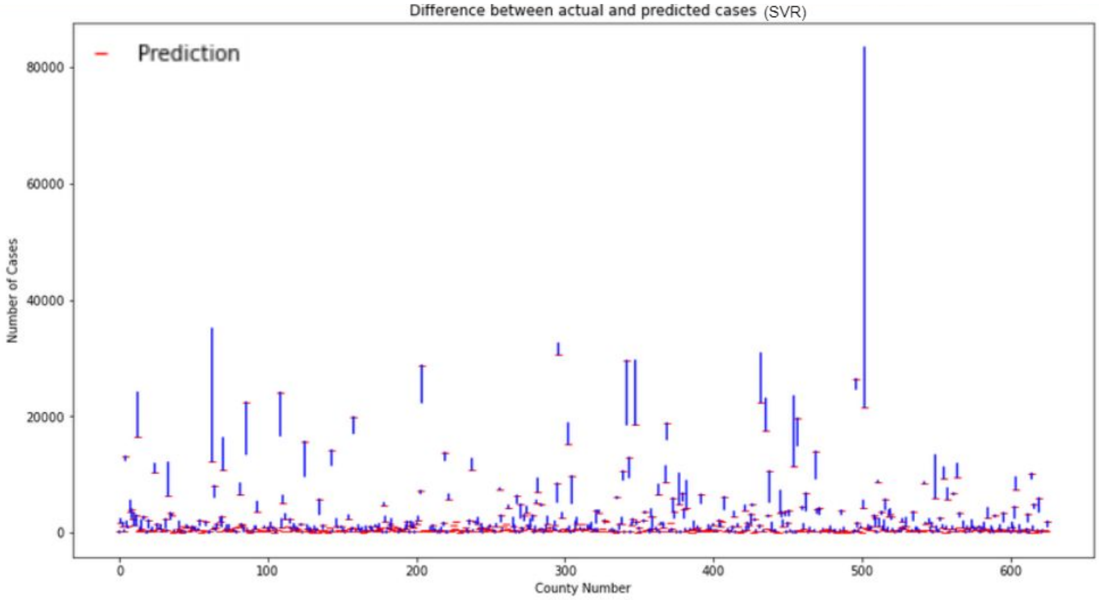


Figure II.

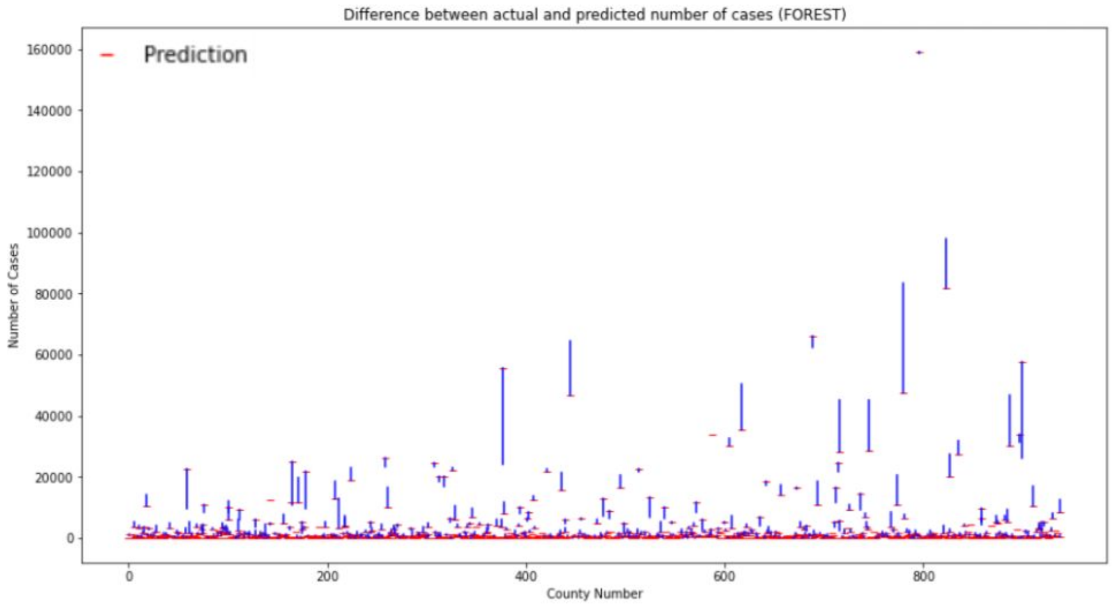


Figure III.

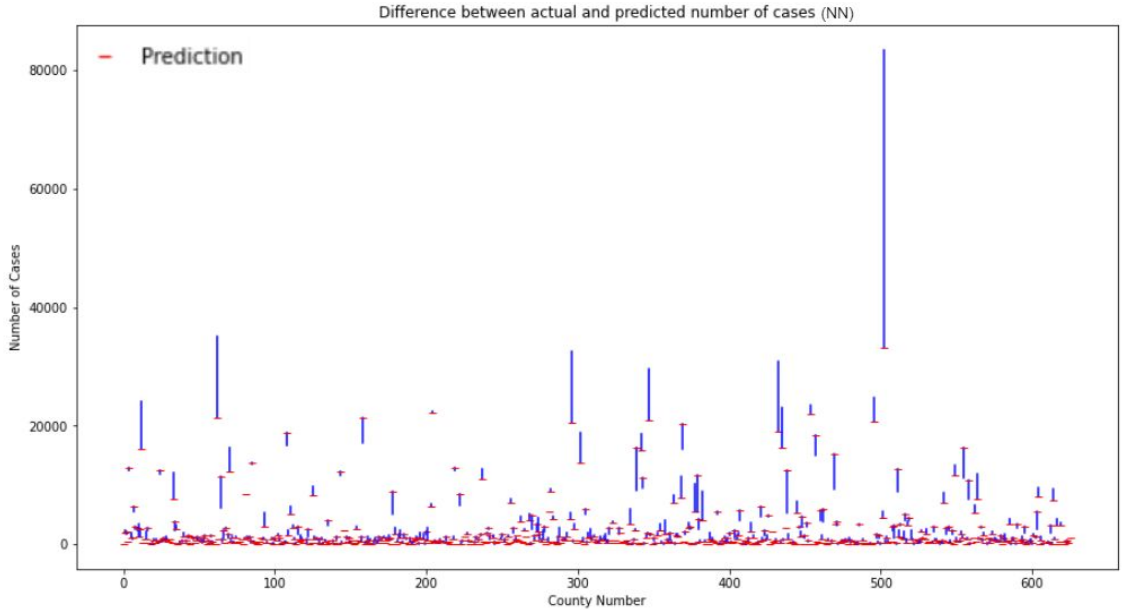


Figure IV.

Figures II,III and IV:
These figures show the difference in predicted cases vs the actual number of cases in each county. Each figure is representative of one of the machine learning algorithms described above.

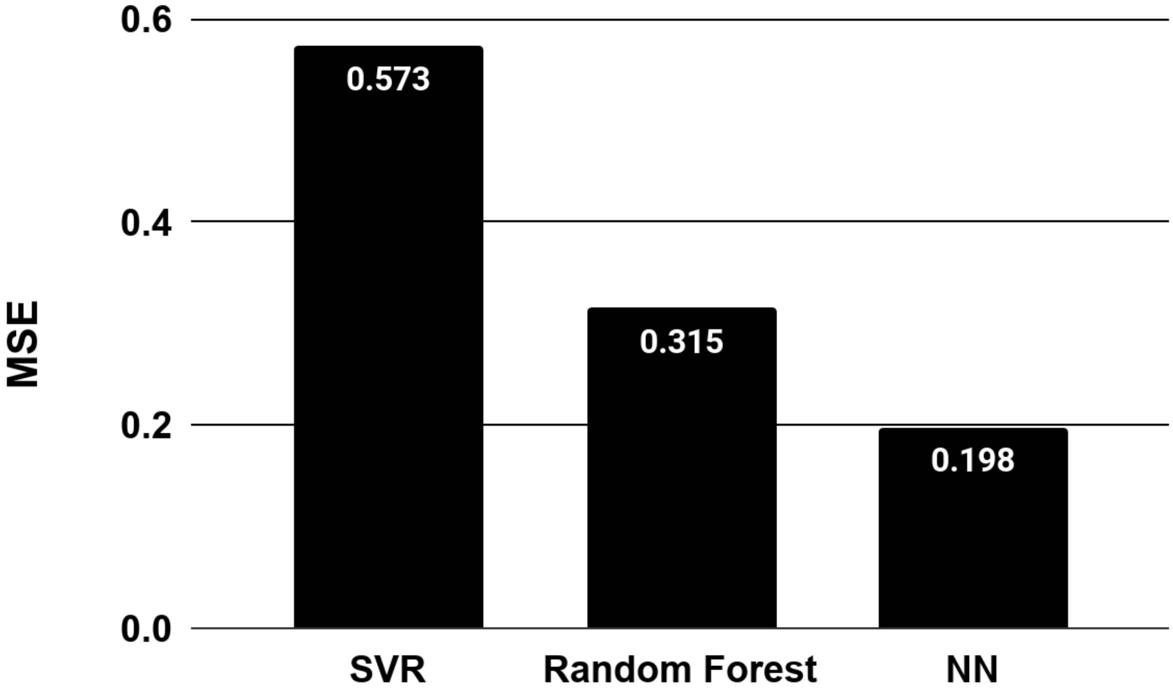


Figure V.

3.3 Discussion

Our three different models all supported our hypothesis that the areas with higher population would generally have a higher concentration of cases. As shown in Figure II and IV , the SVR and the NN were not good at predicting the high concentration of COVID-19 cases in counties that had a large population. Compared to the SVR and NN, the random forest algorithm performed much better when it came to these counties. That is to say, the random forest algorithm handled outliers the best, as shown in Figure III. In Figure V. the MSE of each model used in this project are compared. The lower the MSE, the better. As one can see, the NN performed the best on our dataset with an MSE of .198, the neural network was able to generalize better than the other two models. By doing so it allowed for the MSE of the NN to get as low as 0.198 proving that it was the best model to represent our data. Followed by the Random Forest at .315 which was heavily impacted by the noise of the data which resulted in a higher MSE due to the extreme importance of the population feature. Our baseline SVR performed the worst with an MSE of .573. These observations reinforced our initial thoughts when approaching this problem. For future improvements the introduction of new features could prove extremely valuable and the use of temporal data could allow model translation to other days and gain insight on the future trend of positive cases.

4. Related Work

The problem is to predict the amount of cases in individual counties across the country. Our approach to the problem of predicting COVID-19 cases is different because we have broken it down to smaller regions. This will help us with the overall accuracy of our model and hopefully result in a better outcome. One way in which we could improve our model would be to add data about the number of tests done in each county. This would allow us to better understand the state of each county and predict their covid numbers with better accuracy.

A more complete but alternate approach which we thought about during the inception of this project, was to use past data to predict the COVID-19 cases trend in all of the United States counties. This would take an enormous amount of work as data from the past months would have to be collected and sorted with time stamps so that the model we designed picked up patterns that were real and not fabricated by poor timestamps on the data. A model of this type was created for Iran based off of the data collected in China. They only used three parameters: time-dependent transmission rate, time-dependent recovery rate, and time dependent death rate. This model only needed to be injected with the current number of cases in Iran and then it could predict the number of cases for the next month. While this model could be improved by adding features such as population and mask use, its basic nature allows it to be very general and thus work for different areas around the world. However, extremely general models should be taken with a grain of salt because there is a high likelihood that the model is too general and severely underfit.

5. Code and Dataset

https://github.com/CS254-ML-FinalProject/Final-Project

6. Conclusion

Our results have shown us that there is certainly a relationship between mask use, population, population density, and COVID-19 cases. Although the relationships between these features are different than previously expected, they are still interesting and can perhaps give us insight on COVID-19 and its ability to spread across the United States. One of our findings which stands out is that population density is not the most important feature when calculating COVID-19 cases. Instead, the population of a county is a significantly better feature to calculate the number of cases on.. With the extra importance being placed on the population feature, there is inherently less importance to all the other features which we have discussed.

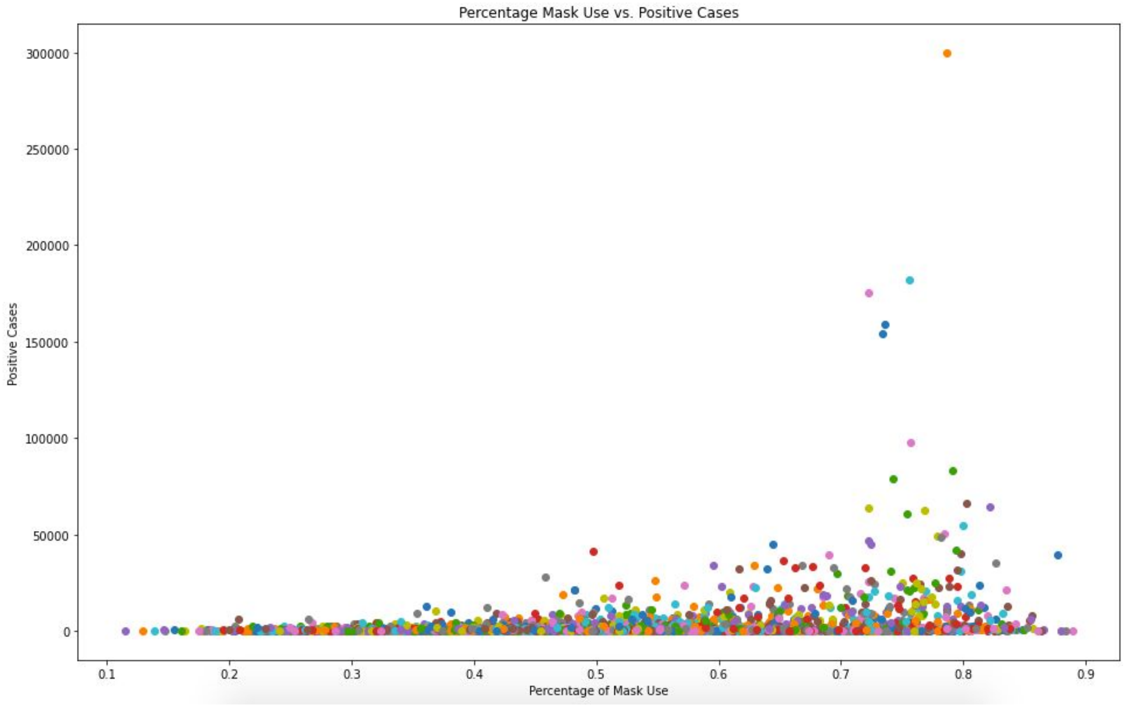


Figure VI.

One place where we see this is in Figure VI. We expected for the graph to have a high amount of cases where there was little mask use, and a low amount of cases where there was high mask use. It was exactly the opposite. This leads us to initially conclude that mask use in United States counties is a function of the number of cases in that county. As we have shown above, our original assumptions were on the right path, but not completely correct. One of the next steps to improve on our model is to collect temporal data so that we have a model that can predict cases/outbreaks in real time. Another improvement we could make is to include new features such as political affiliation, age groups, race and religion.With these improvements, our model could then have the possibility to evaluate and predict the total US cases and the trend over time. Our results will be useful for future projects/research on this topic because they show the importance of different features and how our three different algorithms behaved with this data.

Bibliography:

The New York Times. (2020, July 28). Nytimes/covid-19-data. Retrieved September 25, 2020, from https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv

The New York Times. (2020, March 03). Covid in the U.S.: Latest Map and Case Count. Retrieved September 25, 2020, from https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

The New York Times. (2020, September 25). Nytimes/covid-19-data. Retrieved September 25, 2020, from https://github.com/nytimes/covid-19-data/blob/master/live/us-counties.csv

United States Census. (2020, March 26). Retrieved September 25, 2020, from https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/

Bureau, U. (2020, August 06). USA Counties: 2011. Retrieved October 28, 2020, from https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html

Zareie B, Roshani A, Mansournia MA, Rasouli MA, Moradi G. A Model for COVID-19 Prediction in Iran Based on China Parameters. Arch Iran Med. 2020 Apr 1;23(4):244-248. doi: 10.34172/aim.2020.05. PMID: 32271597.
https://pubmed.ncbi.nlm.nih.gov/32271597

County level data sets - dataset by usda. (2016, August 18). Retrieved December 07, 2020, from https://data.world/usda/county-level-data-sets