Final Project

Group Members (UCI email ID):

Jose Maldonado(jemaldo2), Joshua Si (jjsi), Nomin Amgalan (namgalan)

Problem Statement:

Our group decided that due to current circumstances, we will project the rate of infection for coronavirus based on the weather/climate of the area currently infected. For example, if the weather report says the next 2 days will be rainy, then the next 2 days it's cloudy, then the next 3 days it's sunny, how would the rate of infection change on a day-by-day basis? We will use clustering methods on multiple COVID-19 data sources to identify weather features correlated with higher risk and susceptibility of infection.

For specifically coronavirus and weather, there has not been a lot of research on this. However, there have been several theories that we *may* see certain patterns emerge as weather changes from spring to summer to fall, as seen in both the Harvard News and VOA news, both predicting a possible rise in cases once the weather gets warmer. Their logic being that, if it is generally hotter, then more people will be willing to break quarantine and go outside to simply walk about, greatly increasing the possibility of getting infected. On the machine learning side of the spectrum, there have been several instances of machine learners being used to predict coronavirus spread throughout several countries. While not a very reliable method of prediction as there are many factors that go into the spread of a virus, there have been efforts to try and use different machine learning methods to predict where the virus might go next. These articles we found particularly useful as it supported the idea that weather can be used in conjunction with machine learning to predict the rate of transmission.

Papers/References Relevant to 1(b):

Harvard News warm weather:

https://news.harvard.edu/gazette/story/2020/04/covid-19-may-not-go-away-in-warmer-weather-as-do-colds/

Voa News warm weather:

https://www.voanews.com/covid-19-pandemic/will-warmer-weather-slow-covid-19-spread

Towards Data Science, machine learning and coronavirus spread:

https://towardsdatascience.com/machine-learning-the-coronavirus-9cb8352e1b36

Decomposition:

First Milestone: Gather and compile the necessary data.

The weather reports were found on the National Centers for Environmental Information website (https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/2020/). The data with daily cases in USA counties was collected from a github repository (https://github.com/nytimes/covid-19-data).

Second Milestone: Clean up and further analyse the data (complete with graphs) and package the data into csvs for future use.

Third milestone: Try out different Machine Learning methods on the dataset prepared and determine which features may affect the growth of cases.

Fourth milestone: Interpret errors and biases on the model and draw conclusions.

Coding the Project:

My (Jose Maldonado) experience with coding my chunk of the project was extracting the raw data and essentially forcing a merge between the coronavirus data and the weather data based on both the county and date. Since one dataset was in a text file and the other was in a csv,

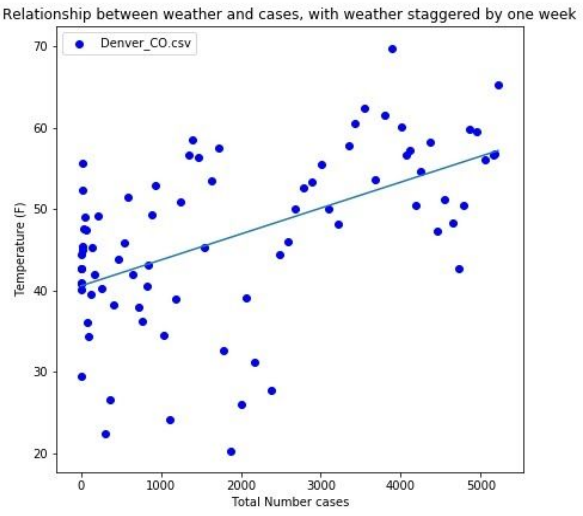
I also worked on making the text file human-readable (a csv). Once I compiled the merged csv data, I drew the graphs in python using matplotlib. Any data that I merged and/or worked with was turned into a csv in case we needed it for future use.

I (Nomin Amgalan) helped with the weather report collection and some parts of the dataset preparation. My main part was working with the data. While making comparisons between the features and which ones were relevant, I used 3 different methods. As for the models, I used Scikit-Learn's libraries and documentation as well as the mltools library provided in CS 178 with Professor Mandt, which I took before. I built my models on various combinations of the features to test whether the correlations were right or wrong.

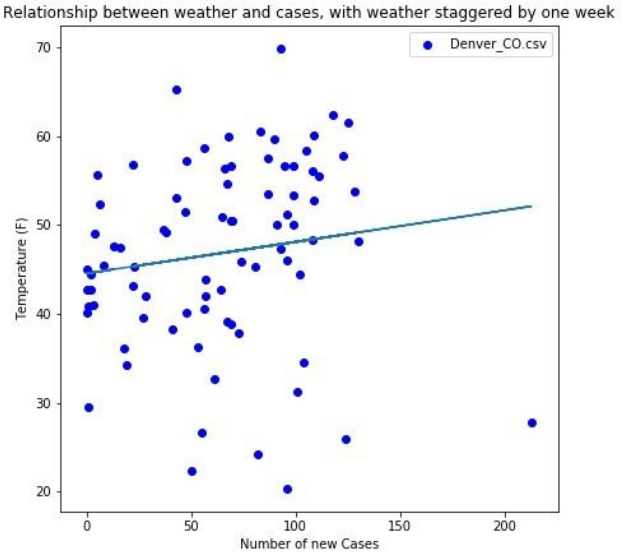
I (Joshua Si) helped interpret the different features in the data and found. I used Jupyter notebook to run KNN with the datasets. I used Nomin and Jose's processed and sorted data and ran clustering models using Scikit-Learn's sklearn.neighbors, evaluating the correlations among features, including longitude, latitude, elevation, temperature, precipitation, and wind speed, and seeing if they have impact on COVID-19 cases. After we created the models, I helped with documenting what we tried and creating the demo code and README for the project.

Experience and Results:

First, we wanted to see if there was any sort of correlation between the weather and the amount of total/new cases. This was done to check and see if our machine learner could be at least decently correct, and to have a measurable metric when reviewing the models' "correctness".0 Thus, I (Jose Maldonado) created one graph for each county as shown to the right. Note that temperature data is staggered a



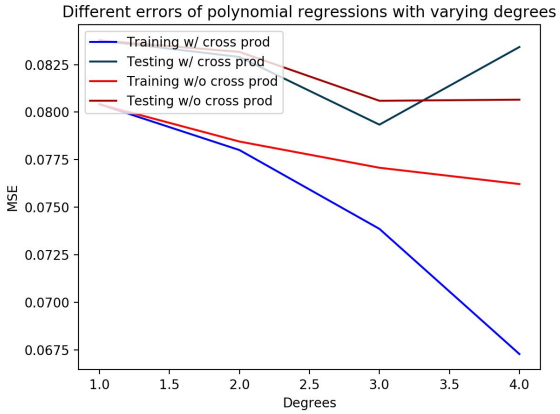
week behind the case date. This is because the incubation period for the virus is around a week, thus we need temperature data a week before the new cases are confirmed. Since we tested a total of 50 counties, we won't be adding all our graphs (all other graphs can be found in our .zip file, under the folder "graphs"). However nearly every graph looked identical to this one; a positive correlation between the total number of cases and the daily temperature. I also created a similar graph for the number of new cases per day, as shown to the right. This particular data shows us two major pieces of information. One, it shows the general average of how many new cases there are. In this particular example, the number of new cases rarely exceeds 140. Two, it shows the temperature range at which the bulk of the new cases are confirmed . In this particular example, most of the new cases happen between temperatures 40 and 60. Once again, like my previous graphs, there are too many to add to this document (the rest are in our .zip file, under the folder "graphs"), however for the most part they all tell the same story: hotter temperatures usually means more new cases. Thus, with all this information gathered, we can safely conclude that, in these 50 particular counties, there is at the very least a small positive correlation between temperature and number of new cases (per county).



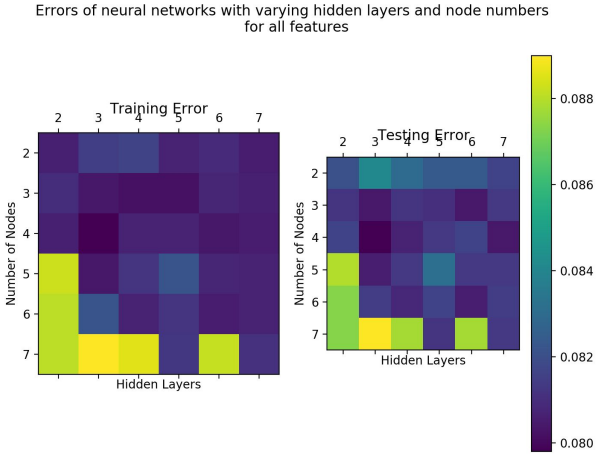
To further check if temperature really affects the number of cases, we have made 3 rankings of the features, source code of which can be found in the *rankingFeatures.py*. First ranking was made by the LassoCV library in Sklearn. Second was made by the Pandas' correlation function. Third was checking each features' variance.

Conclusions:

I (Nomin Amgalan) have tried Linear and Neural Network Regressions. For the Linear Regression, I made various polynomial models with varying degrees as well as both with cross product and without. Models where the cross product was considered yielded better results. As shown below.



For the Neural Network Regression, I tried out different parameters such as number of hidden layers and number of nodes at each layer. I evaluated all models by their Mean Square Error, where both Linear and Neural Network models performed similarly. The image below shows the training and testing MSE for varying parameters on all features



For both regression models, we calculated the MSE of various combinations of features, mainly top 3 features (Temperature, Rain and Latitude) versus bottom 3 features (Wind Speed, Elevation and Longitude). Surely, the MSE of the model trained on only the bottom 3 features performed worse than the other one. All the graphs made can be found in the folder RegressionGraphs.

Thus, with this specific model with these particular parameters, we trained our model and created a small application that, when given longitude, latitude, elevation, temperature, wind speed and

rain, it predicts a percentage of how the number of cases will have grown in a week. This application is called “predictingCases.py” and can be found in our included zip folder. This would help everyone in predicting how the future weather might affect the COVID-19 situation and help be better prepared for the upcoming growth if there is any.

While we weren’t able to conclusively prove whether weather has a significant enough impact on growing COVID-19 case numbers, we did discover a significant correlation between temperature and COVID-19 cases. This correlation may be attributed to people being less likely to go outside or be in contact with others when it is cold, and more likely to want to go outside when it is hot. Thus, perhaps due to this correlation, our model is able to fairly accurately predict how COVID-19 may grow in the coming days when given certain weather parameters.

Works Cited

Almukhtar, S et al. (2020, March 03). Coronavirus in the U.S.: Latest Map and Case Count. Retrieved June 06, 2020, from https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

Amit, Tomer. "Machine Learning & The Coronavirus." Medium, Towards Data Science, 21 Mar. 2020, towardsdatascience.com/machine-learning-the-coronavirus-9cb8352e1b36.

Baragona, Steve. "Will Warmer Weather Slow COVID-19 Spread?" Voice of America, Voa News, 12 May 2020, www.voanews.com/covid-19-pandemic/will-warmer-weather-slow-covid-19-spread.

"COVID19 Global Forecasting (Week 1)." Kaggle, Kaggle, 19 Mar. 2020, www.kaggle.com/c/covid19-global-forecasting-week-1/data.

"COVID19 Global Forecasting (Week 2)." Kaggle, Kaggle, 26 Mar. 2020, www.kaggle.com/c/covid19-global-forecasting-week-2

"COVID19 Global Forecasting (Week 3)." Kaggle, Kaggle, 2 April 2020, www.kaggle.com/c/covid19-global-forecasting-week-3

"COVID19 Global Forecasting (Week 4)." Kaggle, Kaggle, 9 April 2020, www.kaggle.com/c/covid19-global-forecasting-week-4

davidbnn92. "Weather Data." Kaggle, Kaggle, 12 Apr. 2020, www.kaggle.com/davidbnn92/weather-data.

designer, JHCHS website. "Situation Reports on the Novel Coronavirus Identified in China." Johns Hopkins Center for Health Security, 3 June 2020, www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-SituationReports.html.

Powell, Alvin. “Warm Weather May Have No Impact on COVID-19.” Harvard Gazette, 14 Apr. 2020, news.harvard.edu/gazette/story/2020/04/covid-19-may-not-go-away-in-warmer-weather-as-do-colds.

Smith et al. “Nytimes/Covid-19-Data.” *GitHub*, 6 June 2020, github.com/nytimes/covid-19-data.