Predicting US COVID-19 Hospitalizations From Trends in Google Symptom Searches

Austin Zhang
260886286

Isaac Meadowcroft
260752377

Talise Wang
260829722

October 21, 2020

Abstract

In this project, we use a US Google symptom search trends dataset[1] and a US COVID-19 hospitalization cases dataset[2] in order to compare how well *K-Nearest Neighbour Regression* and *Decision Trees* perform when predicting hospitalization cases from related symptom searches. In doing this, we both compare the predictive power of two standard data science methods on real world data and provide valuable insight on the relationship between Google symptom searches and COVID-19 hospitalization. Due to the disparate nature of our two main datasets and the inconsistencies within the datasets themselves, heavy data preprocessing was essential and proved to have huge consequences on the performance of our models. Overall, we found that *Decision Trees* produced significantly better predictions than *K-Nearest Neighbour Regression*.

Introduction

COVID-19 is the largest global pandemic to occur during the age of widespread internet usage. This has led researches to investigate how internet data can be used to extract meaningful insights about the pandemic. Notably, Higgins and colleagues at Cedars Sinai Medical Center, Indiana University and Kentuckiana ENT published a study which found that "worldwide search terms for shortness of breath, anosmia, dysgeusia and ageusia, headache, chest pain, and sneezing had strong correlations ($r > 0.60$, $P < .001$) to both new daily confirmed cases and deaths from COVID-19."[3] In a similar vein, this project uses two standard machine learning methods, *K-Nearest Neighbour Regression* and *Decision Trees*, to predict US COVID-19 hospitalizations based on US Google symptom searches. To train our algorithms, we use a US Google symptom search trends dataset[1] and a US COVID-19 hospitalization dataset[2]. We first merge the datasets and preprocess the data. To better understand the data, we then visualize it and use *Principle Component Analysis* to reduce its dimensionality. We, moreover, cluster the data with *K-means*. Next, we use *K-Nearest Neighbour* and *Decision Trees* to predict COVID-19 hospitalizations based on Google search trends. Finally, we compare the performances of *K-Nearest Neighbour* and *Decision Trees*. Our principle finding is that *Decision Trees* produced significantly better predictions.

Datasets

Two open-source Google datasets were processed and merged in order to conduct our research experiments. Data regarding the prevalence of Google searches for various common symptoms was acquired from the Google COVID-19 Search Trends dataset[1]. This dataset measures the weekly frequency of Google searches for common medical symptoms across US states throughout the COVID-19 pandemic. Data measuring the number of COVID-19 hospitalizations was acquired from the Google Open COVID-19 dataset[2]. This dataset provides daily statistics on COVID-19 cases and hospitalizations across different countries and regions of the world. Due to the disparate nature of the two datasets and the inconsistencies within the datasets themselves, heavy data preprocessing was essential. First of all, since the hospitalization dataset was at a daily resolution and the search trends dataset was at a weekly resolution, we had to sum up the daily hospitalizations throughout every week in order to obtain weekly hospitalization counts. We then merged the two datasets and removed symptom search data points which were more than 35% of their values. Next, we replaced missing symptom search data points with the minimum symptom search value in the points' respective state. In order to consider internet population, we then scaled the symptom search points of each region by the region's internet population. This was done using a US Internet Usage dataset[4] which provides the number of active internet users in each US state in 2019. Finally, we removed certain states (e.g. Nebraska) from our dataset due to the extreme hospitalization spikes at certain points.

Results

To better understand the data, we first visualized the evolution of the popularity of two common symptoms, ventricular fibrillation[1] and crackles[2], across different US states throughout the pandemic. *Figure 1* and *Figure 2* below are histograms representing the prevalence of Google searches for ventricular fibrillation and crackles throughout the states considered in our preprocessed dataset. Note that having survived the preprocessing phase, these symptoms have a relatively large amount of data entries supporting them.

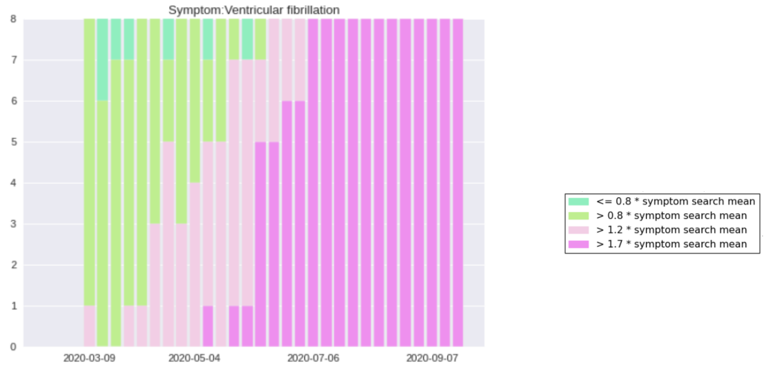


Figure 1: Ventricular Fibrillation Searches Throughout COVID-19

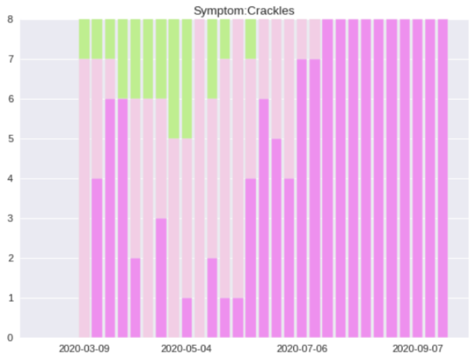


Figure 2: Crackles Searches Throughout COVID-19

The prevalence of Google searches for these symptoms can be further visualized by looking at the search trends specific to each of the states left in the dataset. *Figure 3* and *Figure 4* depict how the number of searches for ventricular fibrillation and crackles changes over time in each state. *Figure 5* and *Figure 6* show how the total number of searches for these two symptoms changes over time and what proportion of searches comes from each state.

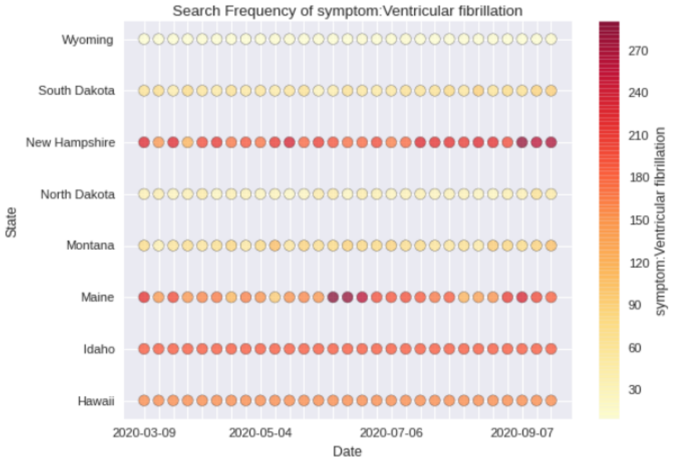


Figure 3: Statewide Ventricular Fibrillation Searches Throughout COVID-19

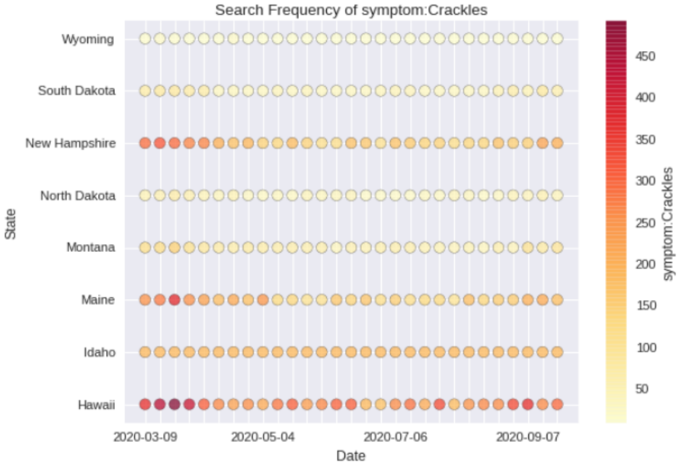


Figure 4: Statewide Crackles Searches Throughout COVID-19

[1] Very rapid uncoordinated fluttering contractions of the ventricles of the heart resulting in loss of synchronization between heartbeat and pulse beat.[5]

[2] A peculiar crackling sound audible with inspiration in pneumonia and other lung diseases.[6]

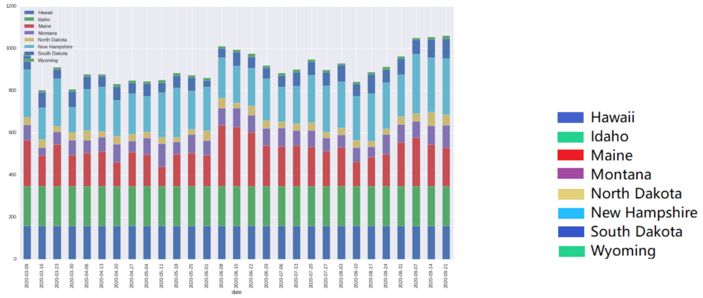


Figure 5: Ventricular Fibrillation Searches Throughout COVID-19

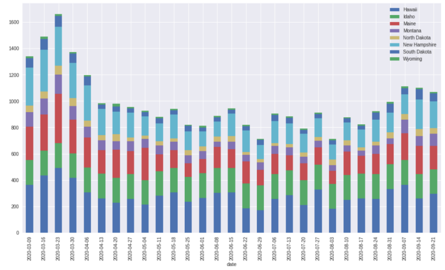


Figure 6: Crackles Searches Throughout COVID-19

To further understand the data, we used Principle Component Analysis (PCA) to faithfully represent the data in lower dimensions. *Figure 7* shows how the percentage of data variance explained changes with respect to the number of principle components (PCs) used. Note that in order to explain 95% of the data variance, at least four principal components are required. In *Figure 8*, we show the percent of data variance explained by each of those four principal components.

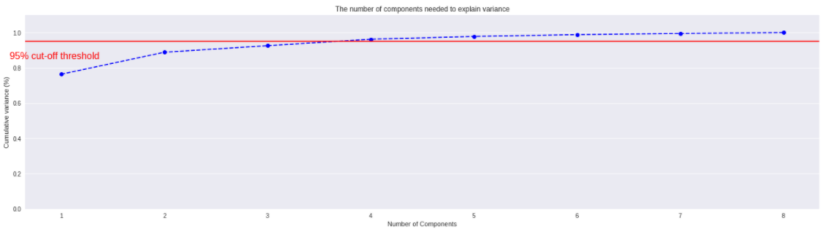


Figure 7: Number of PCs vs. % of Variance Explained

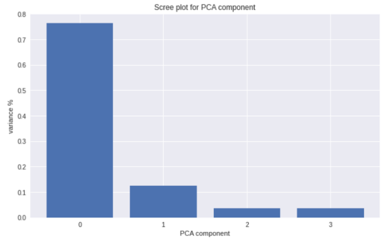


Figure 8: % of Variance Explained by First 4 PCs

Here we notice that over 85% of the variance in the data is accounted for by the first two principal components. We can therefore represent our data fairly faithfully in just two dimensions. In *Figure 9*, we plot the two dimensional, PCA-reduced data. *Figure 10* shows the same data in a standardized form.

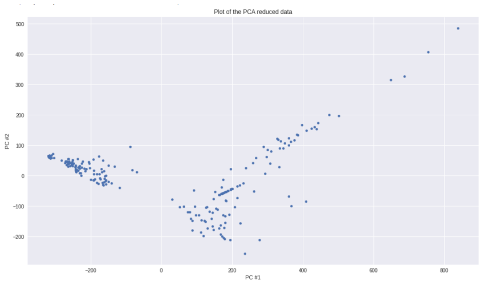


Figure 9: PCA-Reduced Data

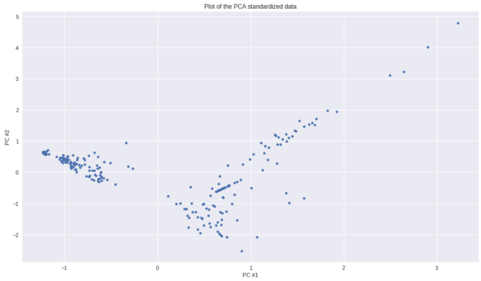


Figure 10: Standardized PCA-Reduced Data

To better understand how the data is grouped together, we used *K-means* clustering to cluster the data. In *Figure 11*, we show how the cost of the clustered data changes with the number of cluster centers. Note that the cost is calculated as the sum of squared distances between each point and its respective cluster center.

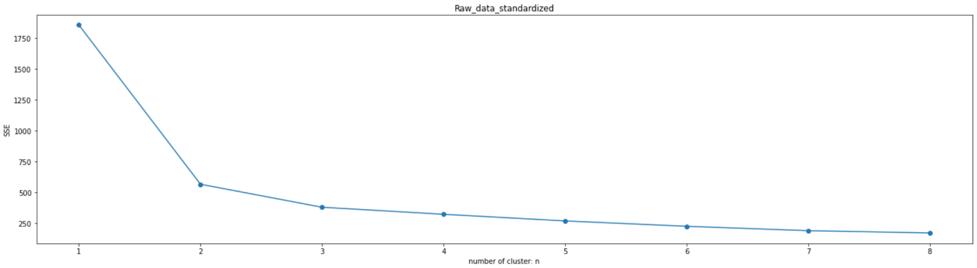


Figure 11: Number of Cluster Centers vs. SSE Between Each Point and Its Cluster Center

From *Figure 11*, we see that the elbow of the curve[3] is at three cluster centers, since the cost reduces considerably slower after this point. Therefore, we can achieve an adequate clustering with three cluster centers. We now cluster the non-PCA-reduced and the PCA-reduced data using *K-means* and three cluster centers ($k = 3$) as shown in *Figure 12* and *Figure 13* respectively.

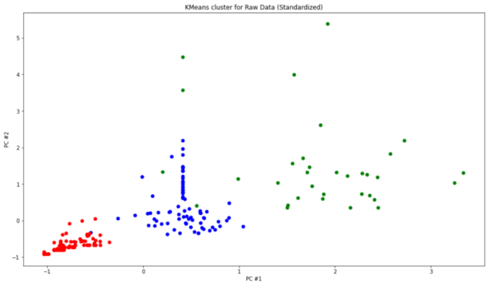


Figure 12: K-means Clustering with Non-PCA-Reduced Data

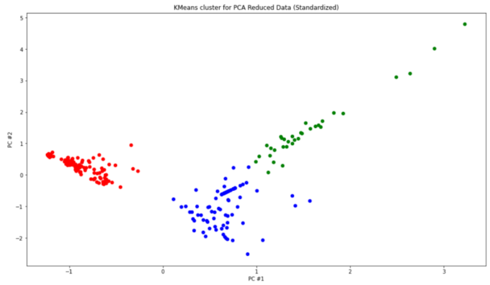


Figure 13: K-means Clustering with PCA-Reduced Data

Now that we have preprocessed and visualized the data, we move on to our main experiment: predicting the number of hospitalization cases from Google symptom search trends using *K-Nearest-Neighbors* and *Decision Tree*. In order to train and validate our preprocessed data, we split the data into train and validation sets using a region split and a date split strategy. We estimate the validation error for the region split data using cross-validation (i.e. calculating the average validation error by repeatedly keeping 80% of regions in training set and 20% in validation set). In contrast, we estimate the validation error of the date split data by keeping data after 2020-08-10 in the validation set and training on the rest. Inspired by the paper by Higgins and colleagues at Cedars Sinai Medical Center, Indiana University and Kentuckiana ENT which found "a high correlation with the Internet search data ($r > 0.7$) 8–10 days before new laboratory-confirmed cases"[3], we explored shifting the hospitalization data. Notably, we observed that both *Decision Trees* and *KNN* were best able to predict COVID-19 hospitalizations when symptom data was aligned with hospitalization data two weeks later, as shown by *Figure 16* and *Figure 17*. Intuitively, this observation makes sense since "the median time to ICU admission from the onset of illness or symptoms ranges from 10 to 12 days."[7] Although shifting improves the performance of the methods, the validation errors of our *K-Nearest-Neighbors* and *Decision Tree* methods are still very high. *Figure 14* and *Figure 15* show the validation error (measured using *mean squared error*) of the *KNN* and *Decision Tree* methods over different hyperparameters k. *Figure 16* and *Figure 17* compare the minimum MSE of *KNN* and *Decision Trees* using the region and date split data. Notably, *Decision Trees* performs significantly better than *KNN* regardless of how the data is split or whether the symptom search data is shifted by two weeks. Note that the minimum MSE of *K-Nearest-Neighbors* is more than 10 times bigger than the minimum MSE of *Decision Trees*.

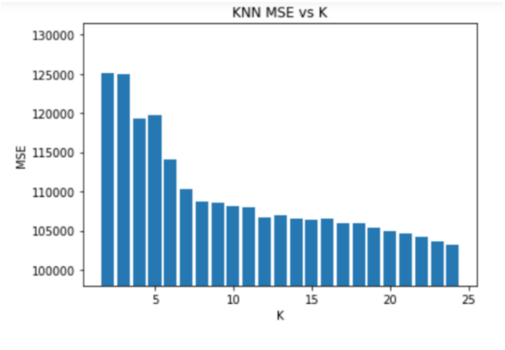


Figure 14: *KNN* MSE vs. k

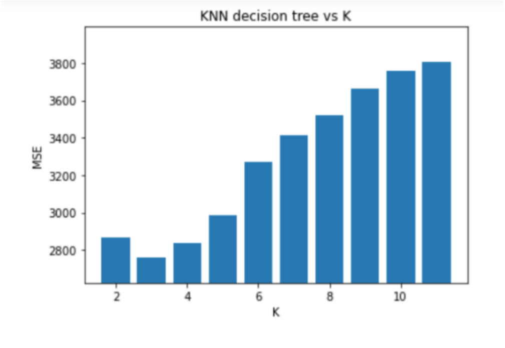
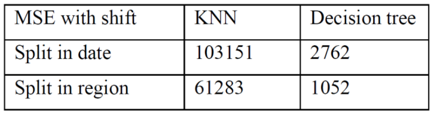
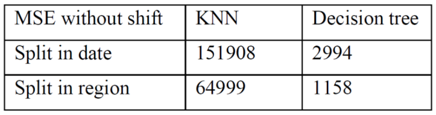


Figure 15: *Decision Tree* MSE vs. k



MSE with shift	KNN	Decision tree
Split in date	103151	2762
Split in region	61283	1052

Figure 16: Minimum *KNN* MSE vs. Min. *DT* MSE in Shifted Data



MSE without shift	KNN	Decision tree
Split in date	151908	2994
Split in region	64999	1158

Figure 17: Minimum *KNN* MSE vs. Min. *DT* MSE in Non-Shifted Data

[3] The point where diminishing returns are no longer worth the additional cost.

One problem with *K-Nearest-Neighbors* which is likely responsible for its poor performance here is that it is very sensitive to the scaling of features. Note that since we did not know the correct scaling of the features, the performance of our *KNN* algorithm likely suffered greatly. Looking at *Figure 14*, we observe that the MSE error of *KNN* continuously decreases despite the increasing of hyperparameter k. This behaviour is problematic since the power of *KNN* stems from its ability to make predictions based on nearby data points in a small neighborhood. With larger k, the *KNN* algorithm is averaging over more points and its predictions become less and less reliable. Clearly, on this data *KNN* fails since nearby points give insufficient information about a given data point. Note that since *Decision Trees* are invariant to scaling, our *Decision Tree* model did not suffer as our *KNN* model likely did. Looking at *Figure 16* we see that, despite the large error, the *Decision Tree* model acts as expected. That is, its MSE is minimized with small k and grows afterwards due to overfitting of the data. Looking at *Figure 15*, we see that the validation error using the region split data is about half of that when using the date split data for both *K-Nearest-Neighbors* and *Decision Trees.* This suggests that data before 2020-08-10 does not strongly predict data after 2020-08-10 or that it is not reliable to predict the future based on the past. Note that since the region-split data performs better, predicting COVID-19 hospitalizations using symptom search data of other states is reasonable. Nevertheless, even when using the best performing approach (i.e. *Decision Trees* algorithm with data split by region), the MSE is still quite large (MSE=1052). One reason for this is that there is not enough data. Notably, the merged and preprocessed dataset has only 8 features and 232 records.

Discussion

Despite the large MSE of our *K-Nearest-Neighbors* and *Decision Tree* algorithms, we gained many meaningful insights from our analysis. We further substantiated a core finding by Thomas S Higgins and colleagues. Namely, there is a connection between symptom search trends in a given region and the severity of COVID-19 in the region. We also showed that *Decision Trees* outperforms *KNN* on our data. This is likely due to the fact that *KNN* is very sensitive to the scaling of features while *Decision Trees* are invariant to feature scaling. Furthermore, we showed that both *KNN* and *Decision Trees* perform significantly better when the data is split based on region as opposed to date. Going forward, future investigation should be done with larger and more consistent datasets, as this will help improve the performance of *K-Nearest-Neighbors* and *Decision Tree* algorithms when trying to predict COVID-19 hospitalizations.

Conclusion

In this project we use a US Google symptom search trends dataset[1] and a US COVID-19 hospitalization cases dataset[2] in order to compare how well *K-Nearest Neighbour Regression* and *Decision Trees* perform when predicting hospitalization cases from related symptom search. Inspired by the findings of Higgins and colleagues at Cedars Sinai Medical Center, Indiana University and Kentuckiana ENT, which found a high correlation between internet search data and laboratory-confirmed cases 8 to 12 days later[3], we explored aligning US symptom search data with hospitalization data two weeks later. This significantly reduced the MSE of both our *KNN* and *Decision Tree* algorithms. Furthermore, we showed that *Decision Trees* outperforms *KNN* by a large margin on our data and has a MSE less than one tenth of that of *KNN.* This is likely due to fact that *KNN* is very sensitive to the scaling of features, while *Decision Trees* are invariant to feature scaling. We then showed that both *KNN* and *Decision Trees* perform significantly better when the data is split based on region as opposed to date. Despite the large MSE of our algorithms, we expect that with larger and more consistent datasets, the performance of *KNN* and *Decision Tree* algorithms when trying to predict COVID-19 hospitalizations can be significantly improved.

Statement of Contributions

Isaac Meadowcroft wrote the Latex writeup and did task one. Austin Zhang worked on task three and provided support for task one. Talise Wang worked on task two and provided many diagrams and plots.

References

[1] Google LLC “Google COVID-19 Search Trends symptoms dataset”.
`http://goo.gle/covid19symptomdataset`, Accessed: October 18, 2020.

[2] Google LLC “Google Open COVID-19 Data”.
`https://github.com/google-research/open-covid-19-data`, Accessed: October 18, 2020.

[3] Higgins TS, Wu AW, Sharma D, Illing EA, Rubel K, Ting JY, Snot Force Alliance Correlations of Online Search Engine Trends With Coronavirus Disease (COVID-19) Incidence: Infodemiology Study
JMIR Public Health Surveill 2020;6(2):e19702
URL: `https://publichealth.jmir.org/2020/2/e19702`
DOI: 10.2196/19702
PMID: 32401211
PMCID: 7244220

[4] “Internet Usage Penetration in the United States”
`https://www.statista.com/statistics/184691/internet-usage-in-the-us-by-state/`,
Accessed: October 18, 2020.

[5] “Ventricular fibrillation.” Merriam-Webster.com Medical Dictionary, Merriam-Webster,
`https://www.merriam-webster.com/medical/ventricular%20fibrillation`. Accessed 18 Oct. 2020.

[6] “Crepitant rale.” Merriam-Webster.com Medical Dictionary, Merriam-Webster,
`https://www.merriam-webster.com/medical/crepitant%20rale`. Accessed 18 Oct. 2020.

[7] “Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)”
`https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html`,
Accessed: October 18, 2020.