COMP 551: Mini Project 1

Wisang Sugiarta[a]**, Viet Hoang**[a]**, and Eddie Cai**[a]

[a]Department of Computer Science, McGill University, Montreal, Quebec, Canada

This manuscript was compiled on October 21, 2020

The sources where information are coming from have never been more important as misinformation has been rampant in the age of this COVID-19 pandemic. People will often resort to Google Search to inform themselves of medical knowledge to see if they may have potential symptoms. In this paper, we extracted hospitalization and search trends data from Google Research's Open COVID-19 Data project to investigate the relationship between search trends and SARS-CoV-2 hospitalizations. We explore various types of visualizations for Google's search data pertaining to symptoms. We also created a machine learning model using K-Nearest Neighbours (K-NN), Random Forest and Decision Trees. These methods allowed to predict hospitalization rates in a given region. We determined that the K-NN method achieved greater performance compared to the Decision Tree and Random Forest regression models. We also demonstrate that using a 5-fold cross validation strategy based on region increased accuracy when compared to splitting the train and test sets of our models by date.

Machine Learning | Search Trends | MSE | K-NN | COVID-19

The 2019 novel coronavirus, SARS-CoV-2, has changed the landscape of everyday life for humans all around the globe. Since the beginning of the pandemic, many researchers have devoted countless hours to predict where and how outbreaks can occur based on a variety of factors. Corporations such as Google and Apple have helped this process by releasing important data sets about how populations have reacted to the pandemic. The data set of interest in this study is Google's symptom search trends. With the advances of machine learning techniques and data visualization, this study's objective is to create models that can predict hospitalization rates based on the symptom search trends in a region.

The study will focus on three supervised machine learning techniques. First, K-Nearest Neighbours (K-NN) is a technique that uses training examples represented as vectors in a multidimensional feature space, each with a class label. The learning phase in this case is called lazy learner. This phase consists of storing the feature vectors and class labels of the training samples. However, in the classification phase, the unlabeled input vector is classified by assigning it the label that is most frequent in k of the closest euclidean training samples, where k is a defined hyper-parameter(1).

Second, Decision Trees are models that use the form of a tree structure. The learning phase of this technique breaks down a training data set into smaller and smaller subsets (2^n), while keeping track of these divisions in a decision tree, which is incrementally developed. The final structure of the decision tree is composed of decision nodes and leaf nodes. The decision nodes represent the attribute being tested while the leaf node represent a numerical target which can be the output label for a given vector(1).

Lastly, Random Forest is another learning method for classification. They are implemented by constructing multiple decision trees during the training and outputting the class which represents the mode or average predictions of individual trees. In general, this implementation outperforms decision trees in terms of accuracy and does not over-fit as much as decision trees. Using these three methods and additional methods to visualize data, we will create models that predict a region's hospitalization rate based on Google symptom searches(1).

Datasets

A. How it was made. Two datasets were used in this study. The first dataset is Google's symptom search trends in a given week throughout all states in the US. This dataset provides a quantity that researchers at Google call the relative popularity of a symptom in a region. Essentially, it is the count of a given symptom search normalized by the probability of that symptom being searched on any given week and by a normalized population in the region. This dataset gives contains the relative popularity of 420 different symptoms. This data was obtained from this repository on 19/10/2020.

The second dataset is an open source dataset that aggregates daily public COVID-19 data, such as deaths and hospitalizations amongst many more, into one set. This dataset not only includes the states in US but all the regions in the world, which we must filter out. This data was obtained from this repository on 19/10/2020.

B. Preprocessing Set 1. This task proved to be difficult based on Google's limited data as well as its format. We imported the two aforementioned datasets and loaded it into two Pandas data frames. Since the second dataset contained many other regions outside of the US, we filtered all of them out. We decided to first eliminate all columns in either dataset that had more than 35% of rows with either no value of 0. We did the same for the rows next but with a more generous limit of 10%. For the second dataset, we saw took the summation of the values of daily data and for each region made them into weekly data.

An import note that Google made, is that the data regarding symptoms were not comparable region to region without having to be normalized in some way. The easiest way to be able to compare the data was to divide each point by the mean of the column. This enables us to have a normalized dataset in the goal of being able to compare regions for visualization. Next, we combined the two datasets based on their region code (state) and week. The then filtered out the the rows with more than 10% empty entries. We are the left with a dataset of 218 states and dates, 79 symptoms and the cumulative hospitalization rate.

C. Preprocessing Set 2. In addition to creating the first modified dataset, we made a second one, to incorporate more data points. Instead of using the the weekly symptom set from Google, we used the daily set as it contained more diverse information from additional regions. To do this, we summed the daily symptom data and added the regions/dates that didn't exist in the first set to a lager set. We then used the same procedure as set 1 to clean and ended up with a dataset of the shape (993, 427).

Results

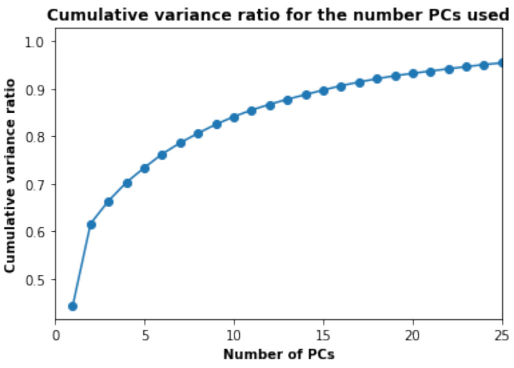

Fig. 1. The cumulative variance ratio is plotted as a function of the number of PCs. A higher cumulative variance ratio means a greater proportion of the total variance is explained. A ratio of 1.0 means 100% of the variance of the data is captured.

Principle Component Analysis. The dataset used in this study contains data about the search trends of numerous symptoms, making the task of visualizing it challenging. We used Principle Component Analysis (PCA) to reduce the dimensionality of the data so that it can be graphically displayed. We want the minimal number of PCs that is sufficient enough to describe the variance our data. The number of principle components (PC) to select is a hyper-parameter. Based off of Figure 1, we decided to select the first 15 PCs, which explains 90% of the variance. Figure 2 is the visualization of the transformed data using the first three PCs.

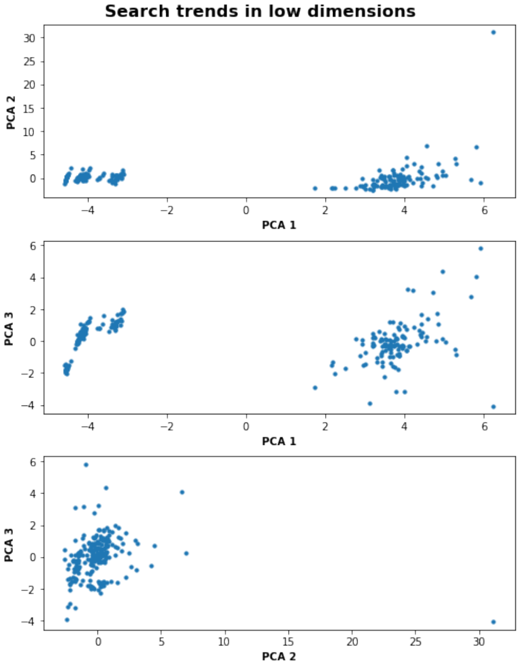

Fig. 2. The three plots represent various axes combinations in order to get an idea of what the data looks like in three dimensional space. PCA 1, 2, 3 are the first three PCs.

K-means Clustering. Just from the observation of the plot in Figure 2, there are clear groupings in our dataset. We used k-means clustering to get a better understanding of the groups present in our data. K-means clustering partitions the data into k-clusters and minimizes the sum of squared distance from each point in a cluster to the cluster center. Similar to PCA, the number of clusters (k) is a hyper-parameter. By using the 'Elbow method' on Figure 3, we decided to select $k = 3$ for the number of clusters.

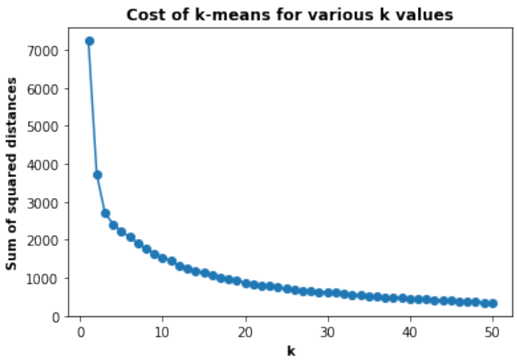

Fig. 3. The cost (sum of square distances of all the points from their cluster center) of k-means for various selections of k. Lower cost, to a certain extent, indicates better performance.

We fitted both the high dimensional and low dimensional data. Both high and low dimensional data are fitted in order to show that the clustering with PCA-reduced data is consistent with the clustering in higher dimensional data. From visual inspection of Figure 4, it appears that the clustering remains consistent. One of the clusters clusters consist of a single point, which could be an outlier. It does appear that two clusters is sufficient in this case.

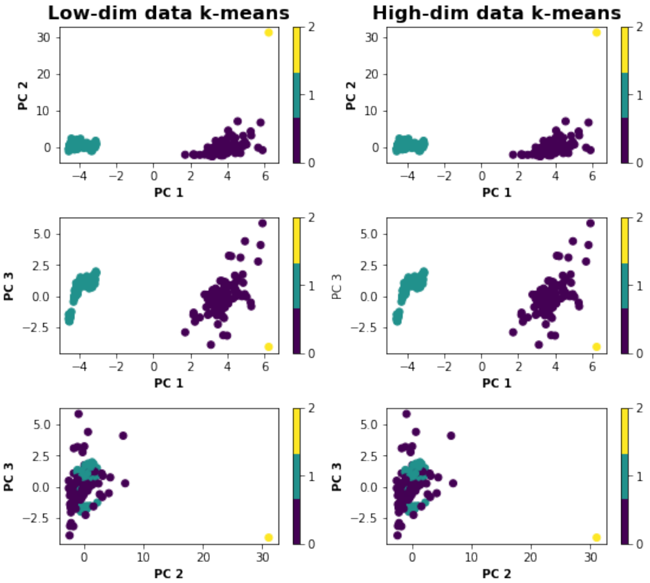

Fig. 4. Visualization of both high and low dimensional clusters using k-means. The three plots for both clusterings reperesent various axes combinations. PCA 1,2,3 are the first three PCs

Comparison of regression performance. The dataset was split into training and validation sets according to two strategies. The first was based on regions, keeping around 80% in the training set and the remainder in the validation set. We did this multiple times and took the average of the result for cross-validation. The second was based on time. Data of dates before 2020-08-10 was used for the training set, and the remainder as validation set. We have chosen to use Mean Absolute Error (MAE) and R2 as our metrics for evaluating these models, as they provide easy to interpret results. MAE describes the mean error between paired observations. R2 describes the amount of variance in the results that can be explained from the model. For the K-Nearest Neighbour model, we first performed hyper-parameter tuning on the value of K, and selected k=6 as the number of neighbours. For the Decision Tree model, hyper-parameter tuning suggested a depth of 6. For the Random Forest model, we chose to have 25 estimators based on training experiments.

As we can see in Figure 5, the K-Nearest Neighbour model with the cross-validation region split strategy performed the best of all models, with an MAE of 28 and an R2 score of -0.19. The MAE of 28 describes that on average, the prediction of the number of new hospitalizations is off by 28. The R2 score of -0.19 is a bit more troubling, as well as the fact that all the models have negative R2 scores. A negative R2 score suggests that the model performs worse than just a model that just predicts the mean value of the validation set's hospitalizations. Despite having a negative R2 score, the K-NN models and the Random Forest models have similar performances, while the Decision Tree models perform slightly worse.

There does appear to be a difference between the date split strategies, however it is difficult to interpret as the region split method has increased performance in the K-NN and Decision Tree models, but decreased performance in the Random Forest model, compared to the date split strategy.

Permutation Feature Importance. After our initial comparison, we decided to attempt to inspect the best K-NN model and determine the most important features in the search dataset for our model's predictions. We used a technique called permutation feature importance in which we run the model many times, while removing one feature at a time. We can then observe which feature, when removed, will increase the model's error the most. We determined that Orepitus, Depersonalization, Epiphora, Nasal Polyp and Rumination were the most important symptoms for our model's performance. Although

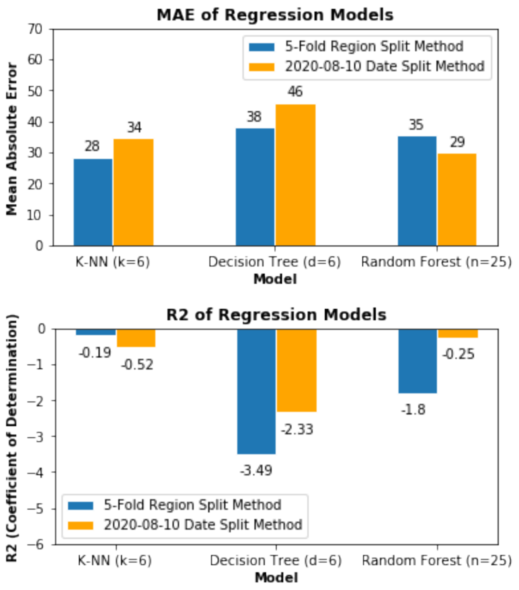

Fig. 5. Comparison of regression performance of tested models using Mean Absolute Error (MAE) and R2 scores. MAE measures error between paired observations. The lower the MAE, the better performance for the model. R2 is the coefficient of determination, which measures proportion of the variance of hospitalization cases that is predictable from the features. R2 of 1.0 means perfect predictor, while an R2 of 0.0 means no variance can be accounted for, and an R2 of negative means that the model does not follow the trend of the data.

these symptoms do not appear to have any direct relationship with having SARS-CoV-2, some of them do appear to be related to depression, which may correlate with peaks of hospitalization cases. To our surprise, further examination determines that common symptoms of SARS-CoV-2 such as Coughing and Fever do not appear in this dataset.

Visualization of search trends. We were interested in seeing how the popularity of search trends changed over time. Using the results from permutation feature importance, we plotted the the popularity of the top 3 most important symptoms. It is interesting to note that the top three symptoms maintain a relatively stable popularity over the course of this year and that they are consistently less popular than the state mean (which would have ratio of 1).

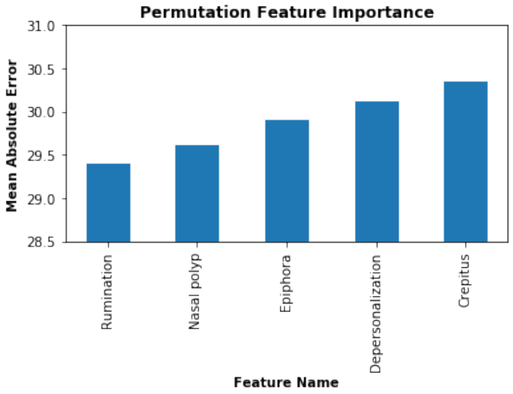

Fig. 6. Most important symptoms for the K-Nearest Neighbour model (K = 6), as extracted using the permutation feature importance model inspection technique. These are the feature that increases the model's error the most when removed from the dataset.

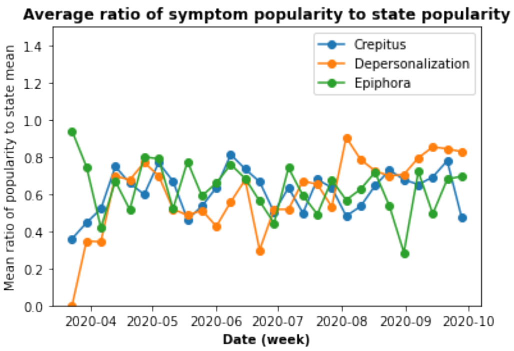

Fig. 7. The evolution of the popularity of the top 3 most important symptoms (based off of permutation feature importance) over time. Higher values indicate greater popularity with respect to the mean, which is centered at 1.

Discussion

Limitations of the data. The weekly dataset is missing data from many major states which hurts the generalizability of our results. Furthermore, as stated earlier, many major symptoms, such as coughing, fever, and nasal congestion were absent in the weekly dataset. This could explain why the most important symptoms (determined by permutation feature importance) are relatively unknown and unrelated to COVID-19. These symptoms are also consistently not very popular relative to the mean, as shown in Figure 7. It does make sense, however, that their popularity remains stable throughout the year as these symptoms typically do not arise for seasonal illnesses, such as the flu or common cold. Seasonal symptoms, such as coughing and fever, would normally see the most popularity during the fall and winter, which is seen in this interactive graph produced by Google. Unfortunately, the sparseness of the data has restricted our ability to draw conclusions regarding the popularity of COVID-19 symptoms and its relationship with hospitalization cases.

Using the second dataset. By using daily data to supplement the weekly dataset, we were able to generate more interpretable results. Figure 9 clearly demonstrates that the 5-fold region method is superior to the date split method. It produces a significantly lower MAE especially for decision trees and random forest. The 5-fold region produces also

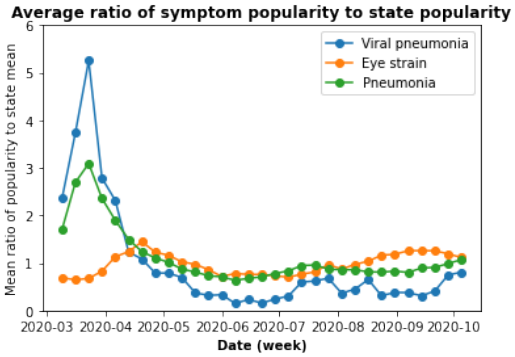

Fig. 8. The evolution of the popularity of the top 3 most important symptoms over time, determined using permutation feature importance on the K-NN model (k=6). This was generated from the new dataset that combined both weekly and daily symptom data.

consistently produces the higher R2 score, making it a better method for splitting the data. K-NN has the highest R2 score with an MAE in the 5-fold region splitting method comparable to other models, confirming that it is the best model for this dataset.

From the new dataset, the top three symptoms that were most important (according to permutation feature importance) were viral pneumonia, eye strain, pneumonia. From Figure 8, it is clear that these symptoms are extremely popular, with viral pneumonia reaching over 5 times the popularity of the mean in late March / early April. All three of these symptoms were more popular earlier in the year, which could be due to the arrival of COVID-19 in the U.S. Pneumonia is a condition where alveoli in the lungs are inflammed, producing pus, mucous and other respiratory problems. It can be caused by viral infections, such as COVID-19 (2). The rising concern of eye-strain is likely explained by the difficulties in adapting to the current working and living conditions. The establishment of quarantine forced the country's population to remain indoors and work from home. With all work being online and remote, it is unsurprising that many people would experience eye strain from sitting in front of a monitor for prolonged periods of time.

Conclusion

Despite the limitations in the initial dataset, the results looked very promising. Future studies using an aggregated dataset of daily and weekly search trends could yield interesting results. Especially since the presidential election will be occurring in the near future, it could be interesting to examine if the the search trends of COVID-19 symptoms is somehow related to whether a state is Republican or Democratic. Another potential study can look at the correlation between hospitalization cases and a state's strictness in quarantine and social distancing policies.

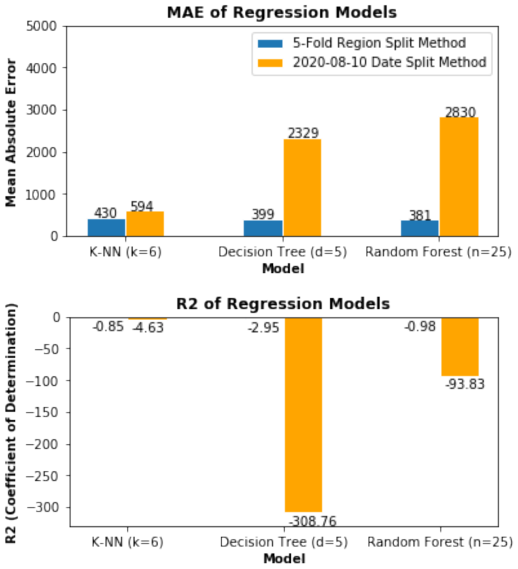

Fig. 9. Comparison of regression performance of tested models using MAE and R2. This was generated from the new dataset that combined both weekly and daily symptom data.

Statement of Contributions. All team-members contributed equally. W.Sugiarta was the lead contributor to data preprocessing,V.Hoang was the lead contributor to vizualizaing and E.Cai was the lead contributor to the supervised learning section.

References

1. K Murphy, Machine learning: A probabilistic perspective in *Texbook*. pp. CH1–5 (2012).
2. MBAZJP Rikinkumar S Patel, Neev Patel, Clinical perspective on 2019 novel coronavirus pneumonia: A systematic review of published case reports. *Cureus* **12** (2020).