Analyzing COVID-19 Search Trends and Hospitalization

Helen Ren （260846901） Shuntian Xiao,(260823600) lanhui Shi,(260860344)

Group: 29

Abstract:

The aim of this project is using machine learning techniques to analyze Covid-19-related datasets and predict hospitalization cases given the Google search trend for health symptoms in some regions in the United States in 2020. The study is performed as a component of Project 1 of COMP551 at McGill University.

Two datasets being explored aggregate trends in Google search patterns for health symptoms and public COVID-19 data across sixteen US regions during 2020. We preprocess the raw datasets with standard techniques including data cleaning, data integration, and data reduction, and we leverage PCA as a dimensionality reduction technique and then cluster and visualize the merged data. Afterwards, we investigate the performance of two supervised learning frameworks, namely k-nearest neighbours and decision trees. In addition, K-cross validation is used to test the effectiveness of our machine learning models. We used two different strategies to split our data into test sets and train sets, namely region based split and time based split. After an extensive analysis of two regression models with careful experimentation, we conclude that the k-nearest neighbour regression approach achieves better accuracy than decision trees and is significantly faster to train on predicting hospitalization cases given related search trend data.

1. Introduction:

1.1 Task description:

Coronavirus diseases (COVID-19) is an infectious disease that has spread to a number of countries around the world. Most people infected will experience mild to moderate health symptoms, signs and conditions, while older people are more likely to develop serious illness. As of October 2020, the United States has reported the most Covid-19 cases than any other country. During the outbreak of Covid-19, machine learning techniques can be deployed to visualize big datasets and to support the diagnosis and prognosis of Covid-19.

In this project, we are going to explore how COVID-19 hospitalization cases are reflected by Google search trends of some health symptoms in the United States, then we are able to make predictions of the hospitalization cases given the search trends data. To tackle this task, regression is used to learn the relationship between the input-output pairs. We can use two important supervised learning frameworks related to regression: K-nearest neighbours (KNN) and decision trees. K-nearest neighbour(KNN) performs regression by finding similar instances in the training set, whereas a decision tree divides the input space into regions using a tree structure and assigns a prediction to each label. In this project, we train on a prepressed merged dataset which aggregates weekly Google search trends of each symptom and weekly hospitalized cases across sixteen regions in the United states from March 09 2020 to September 28 2020. The possible patterns in the search trend dataset that can be explored by deploying dimensionality reduction techniques, such as the Principal Component Analysis (PCA) and K-means clustering method. And we discover that a well preprocessed dataset can contribute to the overall models accuracy. Furthermore, we use different train-validation strategies for training, and we compare the overall regression performance of KNNS and decision trees by their efficiency and accuracy.

1.2 Important findings

(Adding normalization dramatically improved performance on the validation sets.)

We found the search trend of symptoms 'Viral pneumonia ' and 'shallow breathing' are directly correlated to a higher accuracy.

KNN has better regression performance for region-based train-validation split and decision tree have better performance for time-based train-validation split.

2. Datasets

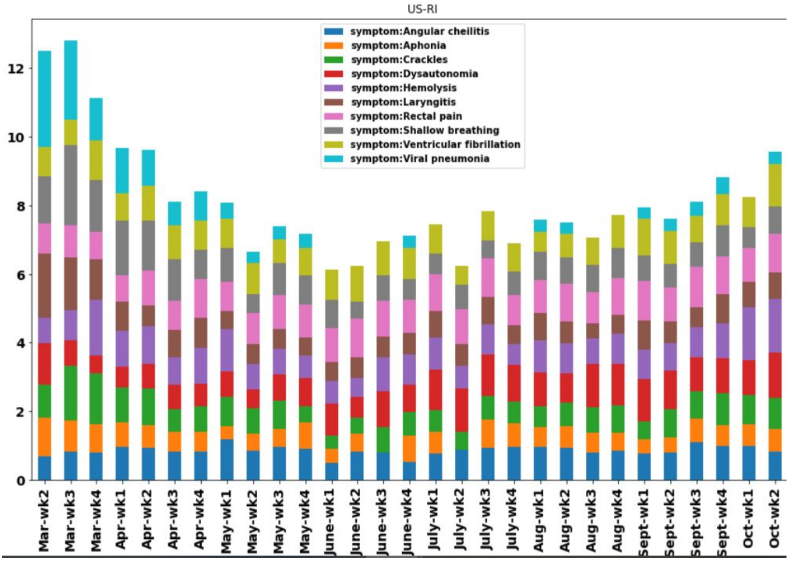
The provided dataset of symptoms search trends from Google searches contains weekly resolution for more than 400 symptoms over sixteen states in the United States. The hospitalization dataset shows information related to patients of COVID-19 and hospitals in a daily resolution.

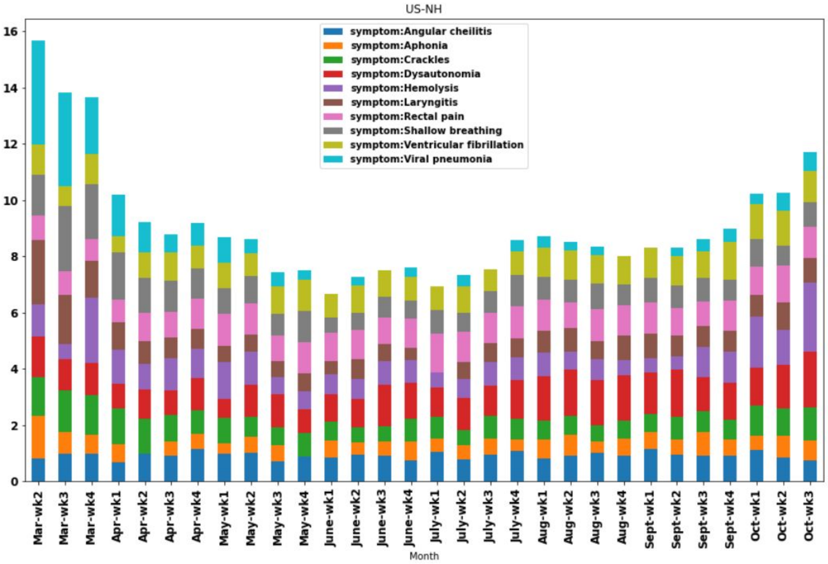
To convert the raw datasets provided into a well-readable format to be used by a machine learning model, all symptoms and regions that have no data or not enough data available are removed. Since the search trend dataset has a region-specific normalization factor, divisive normalization is applied to make the dataset comparable. Finally, two datasets are merged into one by bring both the datasets at the weekly resolution

Three types of plot are drawn for visualizing the data. Our observation is that in March, the time COVID-19 just started in the US, there was an extremely large amount of search on viral pneumonia, shallow breathing, and Laryngitis, rank from low to high. Then the search of the three symptoms gradually decreases over time. All the other symptoms have a relatively stable search amount from March to October. A very similar pattern is observed for all the other regions, one plot is listed as an example.

Left: Figure: Stacked Bar Graph of the Search Trend Dataset for all symptoms for region US-RI over time.

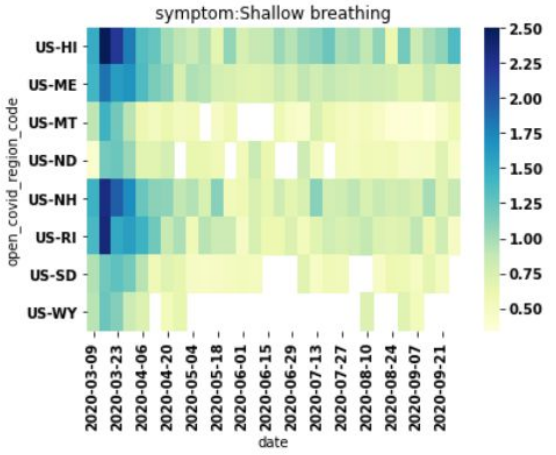
Right: Figure: Stacked Bar Graph of the Search Trend Dataset for all symptoms for region US-NH over time.





Left: Figure: HeatMap of the Search Trend Dataset for symptom shallow breathing across regions over time.

Right: Figure: HeatMap of the Search Trend Dataset for symptom rectal pain across regions over time



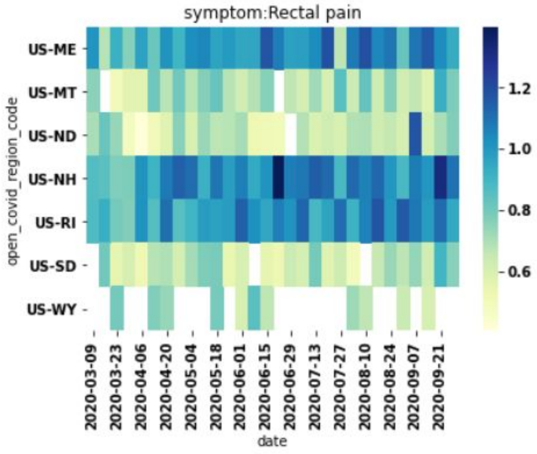
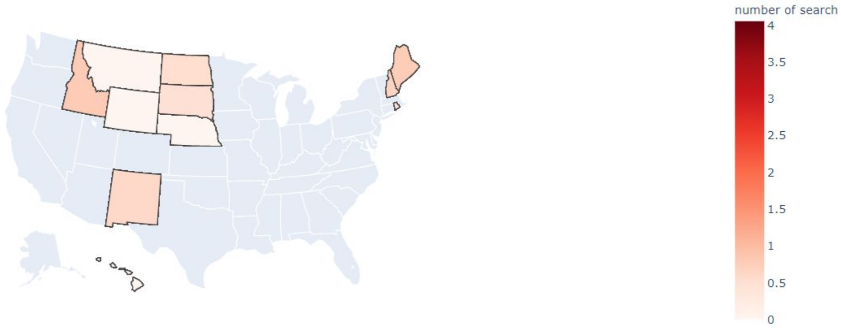


Figure below: Choropleth Map of the Search Trend Dataset that compares the total search of symptom in different region



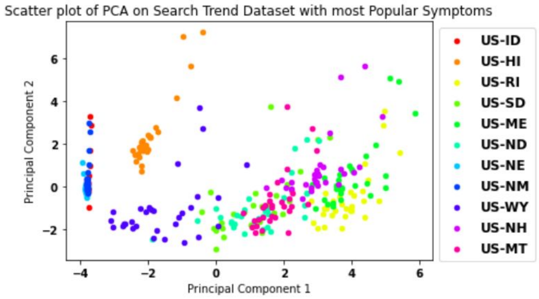
In addition, we have a new dataset about the crude oil price in the United State. From other people's research, as a result of the COVID-19 pandemic, global industrial production is reduced causing lower oil consumption that decreases the barrel price. Therefore, we thought it would be interesting to test whether this new feature can give a better performance on the validation test.

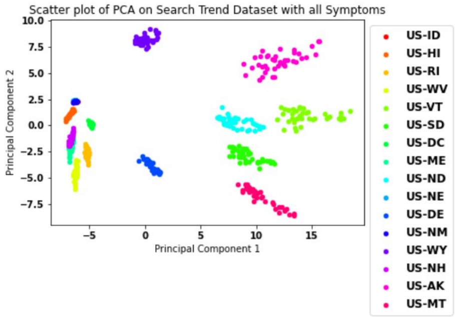
3. Results

3.1 Principal Component Analysis (PCA):

In order to have a better understanding of the Google search trends dataset, we visualize how the distribution of search frequency of each symptom aggregated across different regions changes over time. For the search trends dataset, we treat each weekly time point for a single region in the United States as an independent data point. A single point consists of multiple symptoms, so it is high-dimensional. To visualize the dataset in 2D, we need to faithfully represent the search trends dataset in low dimension, specifically, we use principal component analysis (PCA).

Following two scatter plots are the result of PCA.





The scatter plot on the left shows the first two principal components of the cleaned search trend data which contains the weekly search trend data of most popular symptoms in the ten most popular regions. Whereas the scatter plot on the right shows the first two principal components of the search trend data which contains the weekly search trend data of all the symptoms in all sixteen regions.

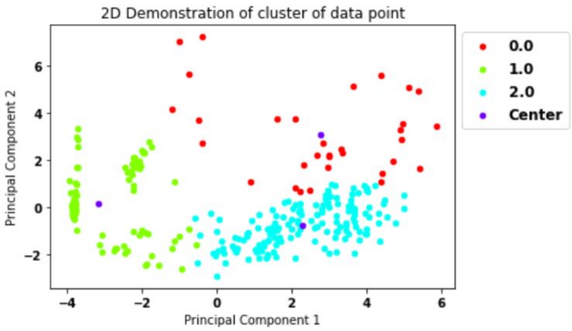
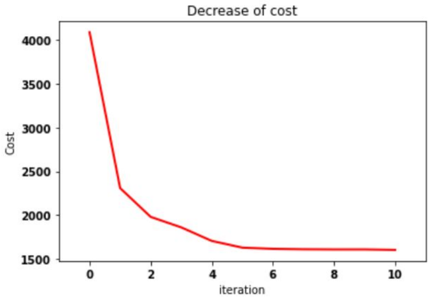
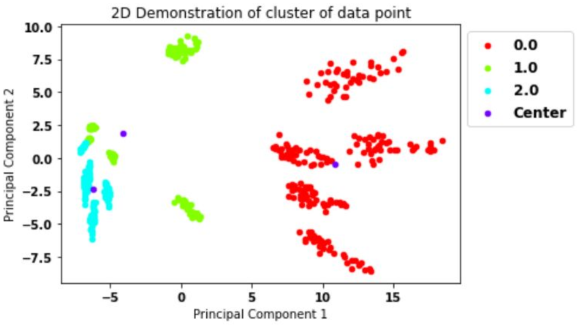
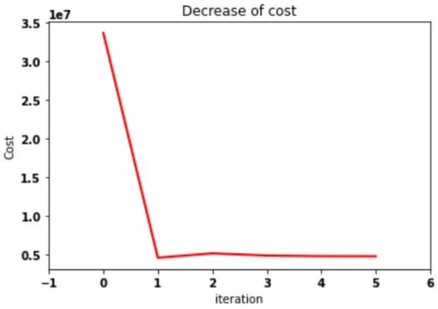
3.2 k-means:

We want to evaluate possible groups in the search trend dataset.

One approach is to cluster the points using regions, and the diagrams are above in section 3.1. Because from the Section 2-Dataset we find that the search frequency of symptoms doesn’t change much between different US regions, we think the better approach is to cluster the points using similarities between the search trend of symptoms.

Following are the 2D visualizations of clusters of preprocessed dataset and raw search trend dataset using the similarities between search frequency with deploying K-mean and PCA. The diagram in 3.2 compared with diagrams in 3.1 indeed show that using regions is not a good clustering method of data points, and preprocessing data have lower cost compared with raw search trend data.

The choice of K is 3, and this came by comparing the cost of different k.

	2D visualization of cluster of data points	Decrease of cost over iteration
Merged_data(preprocessed)		
Raw search trend data with nan filled with		

3.3 KNNS vs Decision trees:

We use two different split strategies for task 3, namely region based split and time based split, and for regression performance, we choose to use mean absolute error.

In region based split, we keep 20% of regions in the test set, and 80% of regions in the training set. And we use groupKfold as cross validation generator to make sure that the same group will not appear in two different folds. The

cross validation result (mean absolute error) for decision tree and KNN are displayed as follows.

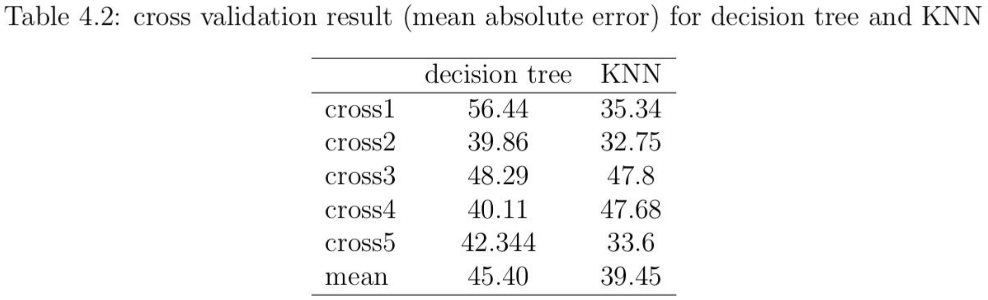


Table 4.2: cross validation result (mean absolute error) for decision tree and KNN

	decision tree	KNN
cross1	56.44	35.34
cross2	39.86	32.75
cross3	48.29	47.8
cross4	40.11	47.68
cross5	42.344	33.6
mean	45.40	39.45

In region based split, we keep the data before 2020-08-10 as train set and the data after it as test set. We then use both decision trees and knn to fit, and then calculate the mean absolute error, by using the result of prediction. The comparisons are displayed as follows.

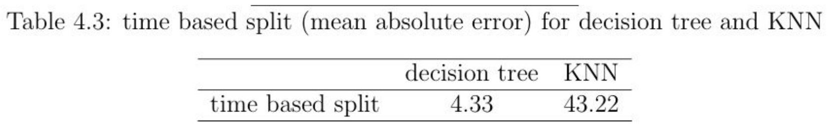
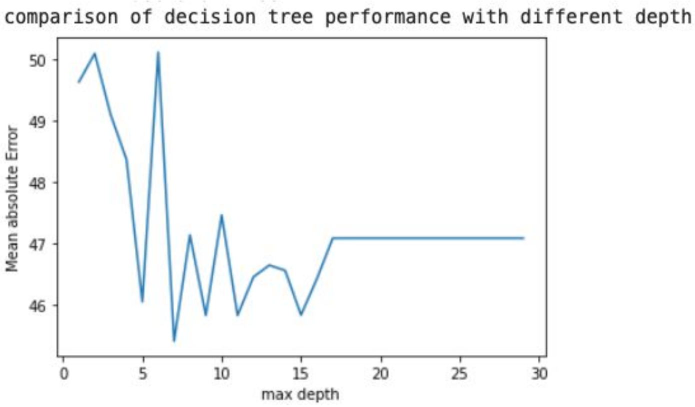


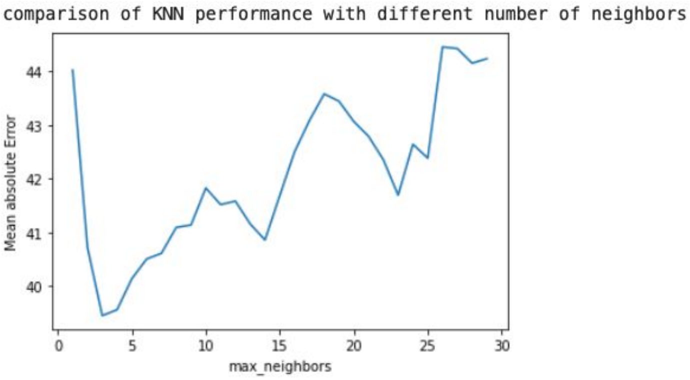
Table 4.3: time based split (mean absolute error) for decision tree and KNN

	decision tree	KNN
time based split	4.33	43.22

In conclusion, for region based split, KNN performed better, with a mean absolute error of 39.45. For time based split, the decision tree performed better, with a mean absolute error of 4.33.

Creativity point: 1. To make sure each model yields the best performance on the data, we use an automatic way to determine which parameter value would give us the best results. The comparison within a range of parameter value is listed below. From the graph, we can see that the best max depth in decision model is 7, and the best value of number of neighbors in KNN model is 3.





2. We also tried another regression model, multinomial naive bayes model, which yields a performance of 1.32 as mean absolute error, which is better than both decision tree and KNN. 3. Furthermore, we combined the crude oil price dataset integrated with the origin dataset to train our model, however a less satisfactory result is shown which means the symptom search trend has a stronger correlation with the number of hospitalizations.

4. Discussion and Conclusion

In this work, we explore different approaches to visualize the data. Using PCA allows us to reduce the dimensionality to 2 while retaining trends and patterns. Then, we observed that the decision tree has a better performance on time based split, whereas KNN performed better on region based split. Therefore, we conclude that certain models may appear to be dominant on specific datasets, but in order to obtain a more accurate prediction, combining multiple models would be a better option. For future work, we would like to explore different approaches on training our model on a larger dataset. Furthermore, we believe the accuracy on the validation set could have been improved by further extraction of new features.

Statement Of Contributions

All team members have made significant personal contributions towards this project:

Helen Ren: Data preprocessing, model training, linear regression, write up, implementation contribution (task1 and task3)

Shuntian Xiao: Data preprocessing, PCA , K-Means clustering , write up, implementation contribution

Lanhui Shi:Data preprocessing, data visualization, write up, paper finding, implementation contribution

References

[1] Coronavirus Search Trends. (n.d.). Retrieved October 22, 2020, from https://trends.google.ca/trends/story/US_cu_4Rjdh3ABAABMHM_en

[2]Pair-code.github.io. 2020. Explore COVID-19 Symptoms Search Trends. [online] Available at: <https://pair-code.github.io/covid19_symptom_dataset/?date=2019-06-17> [Accessed 22 October 2020]

[3]World Health Organization: WHO. (2020, January 10). Coronavirus. WHO. https://www.who.int/health-topics/coronavirus#tab=tab_1

[4]Donia Aloui, Stéphane Goutte, Khaled Guesmi, Rafla Hchaichi. COVID 19’s impact on crude oil and natural gas S&P GS Indexes. 2020. halshs-02613280f