COMP 551 - Mini Project 1
Analyzing COVID-19 Search Trends and Hospitalization

David Castonguay (260804528), Marco Guida (260803123), Sean Smith (260787775), ECSE and Computer Science McGill University

***Abstract*—The essence of this project lies in the analysis of two datasets related to the current COVID-19 situation. That is, a dataset provided by search engines which aggregates search popularity index to various health symptoms, another which provides data for hospitalization. Using both datasets, the performance of two regression models, k-nearest neighbours and decision trees, was investigated. Before applying these models, K-means clustering and PCA reduction were used to understand and visualize the aggregate data. As for supervised learning, the two regression approaches work in fundamentally different ways, yet both approaches achieve similar accuracy. Though, by evaluating them on time performance, the KNNs model has a clear advantage over its counterpart.**

I. Introduction

The task: Examine if Google symptom search trends are related to COVID-19. More precisely, can search trends be used to predict hospitalizations. If so, this could allow public health experts to use search trends of a symptom within a community to detect outbreaks in communities earlier. This deduction can be made since a similar study was performed in India, where Google Search trends of various symptoms showed a "strong correlation" with chikungunya, dengue fever, Malaria and enteric fever reportings in Chandigarh and Haryana (M.Verma et. al). In order to accomplish this task, supervised learning techniques can be applied on a merged dataframe consisting of two datasets described in the following paragraph.

The first dataset, "COVID-19 Search Trends symptoms dataset", provided by Google is an aggregate dataset containing the relative search popularity of symptoms for states over given time. Each data entry has a date, state, and relative search popularity for certain health symptoms. For each state, a scaling factor was used to normalized the features, thus disallowing the direct comparison of symptom search popularity across different regions or time. The second dataset, "Open COVID-19 Data" provided by Google is an open source aggregation of public COVID-19 data. This dataset was used for the hospitalizations data, given our task was to use Google Search trends to predict hospitalizations and the first dataset did not contain data on hospitalizations.

An important finding in task 2 was that the PCA reduced data was clustered into 3 groups. Within these groups were the data of distinct states. One of these clusters contained a higher rate of hospitalities due to COVID, which could suggest that COVID search terms and outbreaks are region-correlated. Task 3 identifies the difference between two regression methods. KNNs are much faster in training than regression trees, though regression trees end up having a lower mean squared error than that of the KNNs. Through analysis, we can determine which symptoms are closely related to the hospitalisation cases.

II. Dataset

Two datasets compiled by Google were used - one that included weekly data reflecting the volume of Google searches for health symptoms across multiple US states, and the other that included data for daily and cumulative COVID-19 hospitalizations for each US state.

The symptoms dataset was processed in order to remove individual symptoms and US states that did not contain enough data. This was done by measuring the data fill rate for symptoms (that is, the percentage of the data that was non-empty for a given symptom), and discarding those symptoms whose data fill rate was below 50% (i.e. more empty than complete). US states were processed similarly, discarding those US states whose data fill rate average across all weeks was below 40%. Though certain weeks for certain states still lacked data for some symptoms after processing, threshold rates of 50% and 40% were selected such that enough US states remained for tasks 2 and 3 with enough symptoms to analyze.

The hospitalizations dataset was processed to remove US states that did not record any data. Further, using the daily COVID-19 hospitalization data, the dataset was brought to the weekly resolution, so that it could be later merged with the symptoms dataset.

The merging of the two datasets involved keeping data for US states contained in both the filtered symptoms dataset and the filtered hospitalizations dataset. After the cleaning was finished, the merged dataset included 7 US states with search trend data for 24 heath symptoms and data for COVID-19 hospitalizations.

Finally, the search trend data was normalized so that different US states could be directly compared in later tasks. The scheme used to normalize the data was to divide each datapoint by its region-symptom median, that is, the median value for a US state for a specific symptom. This scheme was selected since states that had larger search trend values were divided by a larger median, whereas smaller search trend values were divided by a smaller median. Consequently, the normalized search trend values across all US states ranged between 0 and 3, which made the data more meaningful when comparing different US states. However, a caveat of this normalization scheme is that the resulting search trend data is less comparable between symptoms for the same US state due to dissimilar normalization medians being used for the same US state, depending on the symptom that was normalized.

To help us better understand our data and which symptoms we can expect to potentially be more important than others as we perform future tasks, the data for US states was visualized via plots of certain symptoms against the hospitalization cases for that US state. Some of the visually strongest correlations are demonstrated in Figure 1.

III. Results

The data attained from merging the two datasets contained a total of 24 features, which were the relative symptom search popularities for each state at a given date. The target of each datapoint was the number of COVID-19 hospitalizations recorded that week. An initial analysis of symptom search popularity for every state yielded a total of 24 graphs, 1 for each feature. Each graph shows the popularity increase or decrease of a search symptom, where each line is a States' respective popularity change as shown in the

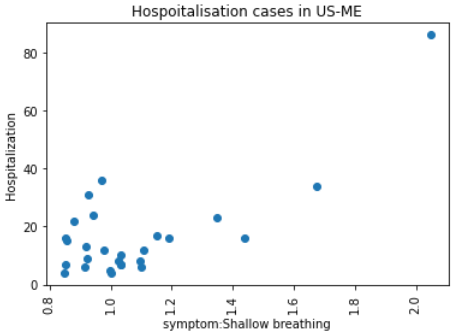

Fig. 1. Visually strong correlation between shallow breathing and hospitalisation cases in US-ME.

graph legend. Since many symptom search records were omitted on a given date for some symptoms, there were several blank entries. In order to deal with these blank entries, they were replaced with the value 0. Additionally, symptoms which had at least half as many blanks as entries were not entered in the graph. This allowed for more consistent data visualization. Figure 2 is an example of one of these graphs.

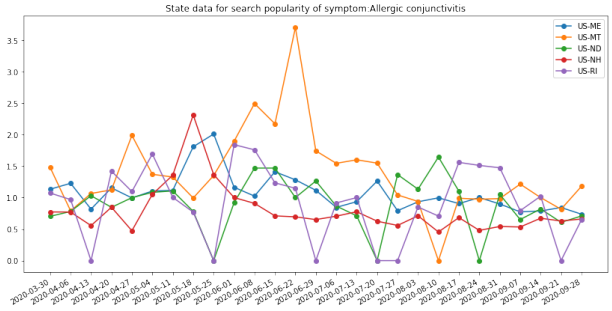

Fig. 2. While there are noticeable trend lines for the popularity of searching the symptom "Allergic conjunctivitis" in each state, no conclusions can be drawn from this figure alone, nor from comparison with other similar graphs. This applies to all 23 other graphs produced in this section.

In another attempt to make sense of the data in the merged dataframe, instead of plotting 24 graphs for each symptom, one might assume it is more useful to plot 7 graphs for each state, namely the symptom search trends for all symptoms per state. Although this results in less subplots, each plot is convoluted with symptom data, as shown in Figure 3.

As observed at this point is that it is highly impractical to attempt to visualize data and draw conclusions in high-dimensions due to "the Curse of Dimensionality". Principal component analysis (PCA) was used to reduce the 24-dimensional data to 2-dimensions with maximum fidelity, after which the data can be clustered using

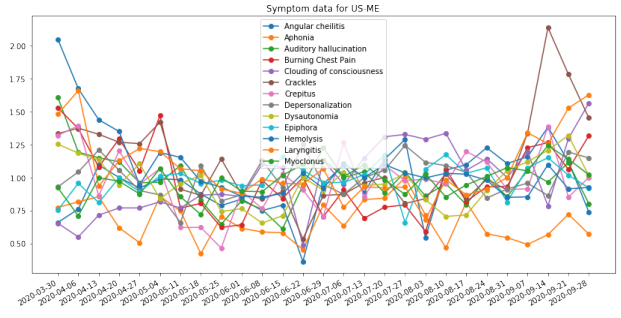

Fig. 3. Symptom search trends for all symptoms for Maine (state code US-ME). As with Fig.4, no conclusions can be drawn from this graph, or the 6 other graphs for each state.

k-means clustering.

Before continuing, it is important to note that before plotting the data, datapoints whose z-score was greater than 2.9 were removed. This equates to 3 datapoints whose hospitalization values were much higher than all others. If these datapoints had been included, it would be impossible to visualize the relative hospitalizations since the colormap scale used for the subplot would be distorted by the outliers. It is also important to note that all blanks (NaNs) in the dataframe were replaced with zeros. This was deemed the safer alternative to removing all rows which contained NaNs, since doing this would result in an unreasonably large reduction of datapoints. A hybrid implementation of replacing some NaNs with zeros and removing some rows with too many NaNs also resulted in a large reduction of datapoints and would result in inconsistent data. The last important note is that values in the dataframe were normalized to a unit-scale using the StandardScaler library from sklearn. This is important since feature scaling has an impact on the results of PCA. The standard sklearn PCA library might wrongly deduce the axis of maximum variance due to features not being scaled properly (scikit-learn.org).

In reducing the dimensionality of the symptom data using PCA, we also lose information. PCA reduction of a dataset can be summarized as a change of coordinates for the dataset $\mathbf{X}$. Given D dimensions, this transformation results to using $D' < D$ components in the new coordinate system for dataset $\mathbf{X}$. In this project, the first 2 components in the new coordinates were used, where the first component captures the highest variance, the second component captures the second highest variance. By using the "explained_variance_ratio_" attribute from sklearn's PCA library, allows us to see the difference in variance across components, as well as information lost by reducing dimensions. There is 33.56 percent of the information in the first component (q1), 28.21 percent of the information is in (q2). Rest of the components contain the rest of the information. In using $D' = 2$, we lost 44.23 percent of the variance from the original dataset. This equates to a lot of information lost. Increasing to $D' = 3$ only gained 6 percent variance of the total dataset variance. Unfortunately, since visualization is no more than 3 dimensions, no more data fidelity could be encapsulated. Fig.4 is the plot for the 2D and PCA reduced dataset. The 3D PCA dataset is in Task2.2 of the code.

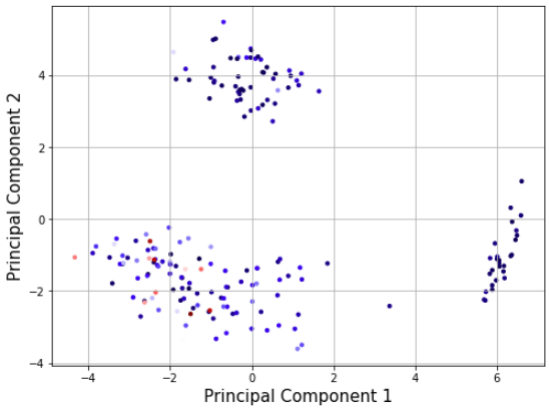

Fig. 4. 2 Component PCA for State Search Trends vs. Hospitalizations. The dark blue spots represent least hospitalizations. Strong red represents most hospitalizations. Faded colours are values in between.

K-medoids was used to evaluate possible groups in the PCA reduced dataset. K-medoids was chosen in order to choose more meaningful points as means, instead of arbitrary means in the plot. From the 2-dimensional PCA reduced dataset, 3 clusters are clearly visible by observation, therefore K=3 was selected as a hyper-parameter for K-Means clustering, although better methods exist for choosing k, such as the elbow method. Another improvement which could be made is using k-means++ to find a cluster which is withing $O(log(k)) * OPT$, given OPT is the optimal solution. The clusters are represented in Figure 5.

Upon further inspection of this data, it was found that each cluster contained a cluster of States, which was surprising. For example, the orange cluster contained all points for US-MT, ND, RI and SD (with the exception of 1 point), the blue cluster contained all points for US-ME and US-NH and the green cluster contained all the datapoints for US-WY (with the exception of 1 point). Unfortunately, this is only 7 states out of the 50, so no concrete conclusions can be made about this observation, however some hypotheses can be made. The orange cluster contains all of the points which contain high number of hospitalizations (as seen in fig.4,

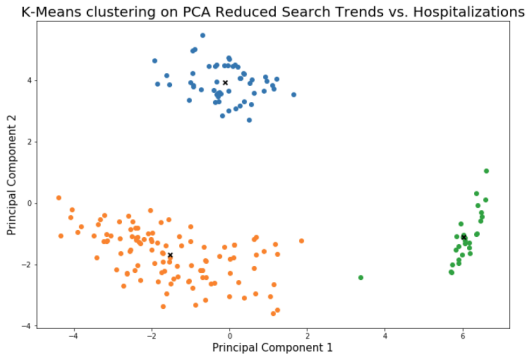

Fig. 5. K-medoids with k=3. Between The orange and blue clusters, there is a singular outlier centroid. This is a datapoint which belongs to the state ND. It is the only green datapoint which is not WY.

where the red points represent highest hospitalizations). Further-more, with the exception of US-RI, all the other states in the cluster are within close proximity to each other. One hypothesis could be that these clusters represent outbreaks for certain regions which correlate to search trends of symptoms. Perhaps regions within close proximity to each other are more likely to have similar number of hospitalizations and therefor similar search trends. To further verify search trends can predict outbreaks, supervised learning shall be explored.

A comparison of regression performance via mean squared error between KNNs and decision trees was done using two validation strategies - region-based cross-validation and time-based validation (where test set is all data after '2020-08-10').

KNNs and decision trees were performed with two symptoms/features (i.e. in low dimensions) to be able to both visualize the training and test sets, and, more importantly, to keep the validation error of our models in a relatively low order of magnitude so that they can be more readily compared. Though not visually represented with a figure in this report, our selections of k=3 for KNNs and depth=4 for decision trees were decided based on the lowest average validation error as k and depth increases for their respective models. Lastly, empty values that remained after cleaning the data were omitted in KNNs and decision trees, as they otherwise caused issues in computing euclidean distances for KNNs and created incorrect internal nodes for decision trees. However, the drawback of this omission was that the splitting of data between training and test sets would not perfectly represent a 4:1 split, thus making them less comparable, depending on the number of empty values omitted.

To compare the regression performance of KNNs and decision trees for the region-based and time-based validation strategies, the error for region-based cross-validation for KNNs and decision trees were visualized, as illustrated in Figure 6, and the error for time-based validation for KNNs and decision trees were visualized, as illustrated in Figure 7. Note that Figure 6 contains the results for only four of the cross-validation splits. This is due to the other splits having much higher validation error (∼15,000).

In observing the validation error for region-based cross-validation in Figure 6, it is noticeable that across all combinations of symptoms for KNNs and decision trees, the ten best (lowest) validation errors for each of the region splittings were, in general, lower for decision trees (green) than for KNNs (red) by 15-20%. Further, two outlier plots can be seen with lower than expected validation errors. These two outliers are represented by the same region splitting, that is, the splitting in which US-ND and US-NH form the test set. In this case, like the rest of the data, decision trees still had a lower validation error.

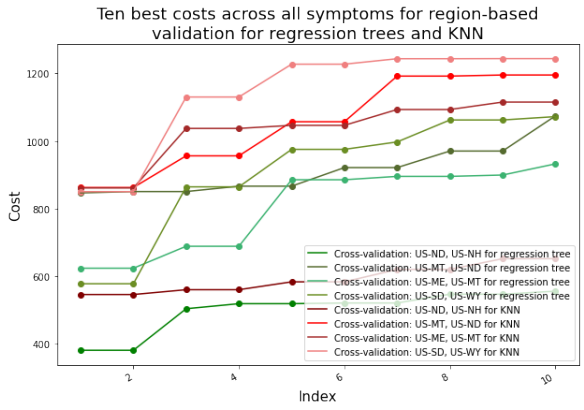

Fig. 6. Representation of the mean squared error for region-based cross-validation for the four best (with regards to cost) region splittings for both KNNs (red) and decision trees (green).

For time-based validation in Figure 7, the validation error is larger for both KNNs (red) and decision trees (green) than even the most costly splittings in region-based validation, illustrated in Figure 6. In comparing the results of KNNs and decision trees of only time-based validation, the results are opposite of those observed in region-based validation. That is, KNNs has a lower validation error than decision trees (without considering the two zero-error outliers in decision trees).

Though a certain algorithm may have a lower mean squared error (depending on the validation strategy),

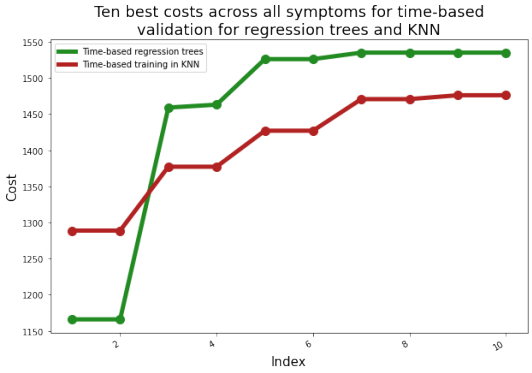

Fig. 7. Representation of the mean squared error for time-based validation for both KNNs (red) and decision trees (green).

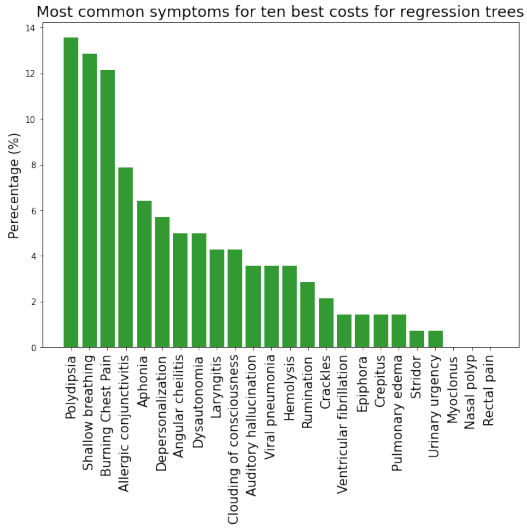

Fig. 8. Visualisation of the commonness of symptoms for the best (lowest) validation error results from the top 10 lowest error regression trees (through all region-based trees and the time-based tree)

time is also an important component in deciding whether KNNs or decision trees is more performant overall. It was observed that in changing the hyper-parameter K, the time complexity of KNNs did not suffer much when the value of K ranged within 1 to 50. This reflects the linearity of the time complexity $O(ND) + O(NK)$ (as oppose to more expensive higher-order complexities), and that in our case the value of K is irrelevant to the observable performance of KNNs at our scale. For decision trees, some components of the algorithm have a high computational complexity. For instance, the greedy test, which is used to calculate the node splitting value, has a time complexity of $O(ND)$. The problem arises when selecting a maximum depth d. At any depth d, the maximum number of splits is equal to $\sum_{n=1}^{d} 2^n$. This means the time complexity of the node splitting increases exponentially through the depth. Therefore, the real time complexity of the whole tree is $O((ND)^d)$. It is necessary to have a depth that is reasonably large e.g. higher than 4, in order to get pertinent expectations. Therefore, the regression tree will usually always take longer to compute than the KNNs algorithm.

Finally, we computed the commonness of symptoms using the best validation error results for KNNs and regression trees. Figure 8 represents the commonness of symptoms in the best results for decision trees. In Figure 8, the most common symptoms are polydipsia, shallow breathing, and burning chest pain. For KNNs (figure in Task 4.7 of our code), the most common symptoms are shallow breathing, allergic conjunctivitis, and ventricular fibrillation. As these symptoms produced the best costs for their respective learning models, their search trends are therefore the most strongly correlated to hospitalization cases from the dataset.

IV. Discussion and Conclusion

The PCA reduced data was clustered into three distinct groups. Within these groups were the data for certain states. Further, one of these clusters contained a higher rate of hospitalities due to COVID. This could suggest that COVID search terms and outbreaks could be related by region, yet this does directly address the task at hand. In order to address the task, supervised learning was used to conclude that symptom search trends can be used to predict COVID-19 hospitalizations. That is, many of the symptoms that one would expect[1] to be correlated to COVID-19 are those that produced the lowest validation error in our results and therefore most accurately predicted COVID-19 hospitalizations for the validation set.

V. Statement of Contributions

Task 1: David, Marco. Task 2: Sean. Task 3: Marco (KNNs), David (decision trees). Report: All.

References

[1] M. K. A. R. S. G. A. S. K. M. Verma, K. Kishore, "Google search trends, predicting disease outbreaks: An analysis from india," 2018.
[2] Scikit-learn.org, "Importance of feature scaling."
[3] M. of Health of the Province of Ontario, *COVID-19 Reference Document for Symptoms*, 2020. Available at http://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/2019_reference_doc_symptoms.pdf.

[1]See Ministry of Health COVID-19 Reference Document for Symptoms. For example, "shallow breathing" from our dataset corresponds to "shortness of breath" from the Reference document. Other symptoms follow this correspondence.