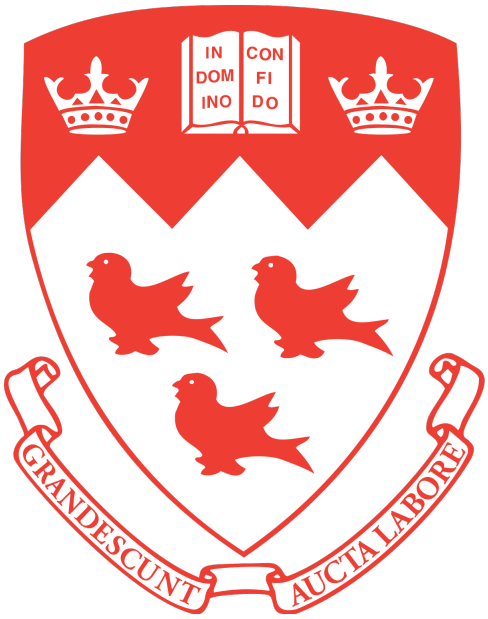


Comp 551: Applied Machine Learning
Mini-Project 1: Analyzing COVID-19 Search Trends and Hospitalization
Group 53

Alexandre Martini Santos - 260798352
Ameer Ibrahim Osman - 260682723
Iyatan Atchoro - 26065932

October 21, 2020

1 Abstract

The objective of the project was to explore different supervised and unsupervised learning frameworks for two COVID-19 related datasets, to tackle real-world datasets. The first dataset consists of aggregated, anonymous Google search trends for the United States. This dataset contains various symptom counts respective of each state. The second dataset consists of aggregate public COVID-19 hospitalizations, cases, deaths, and other attributes. The unsupervised learning models explored in this report are Principal Component Analysis and K-Means clustering to visualize and further understand the data. The supervised learning frameworks explored are K-nearest-neighbours regression and Decision Trees regression. The goal is to train these models to predict hospitalization cases given the search trends data. In our results we have KNN and Decision Tree models that achieve 5-fold-cross-validation errors of 0.0057 & 0.0133 respectively.

1.1 Abbreviation and Notations

PCA: Principal Component Analysis
MSE: Mean-Squared-Error
WSS: Within-Cluster Sum-Squared-Error
Z-Normalization: Equivalent to Standard Normalization
K-NN: K-Nearest-Neighbours
MAE: Mean-Absolute-Error
RMSE: Root-Mean-Squared-Error

2 Introduction

In this mini-project, the team is assigned to train supervised learning frameworks KNN and Decision Trees regression. The goal is to train these regression models such that they can predict hospitalization cases given the search trends data. We implement these models with the assistance of the following notable python libraries numpy, pandas, sklearn, and matplotlib. We utilize Jupyter notebook to provide the results in this report. We would also like to note that results vary between Jupyter notebook, Google Colab, and Pycharm. The dataset version used is dated to October 21st, 2020.

The two datasets we used initially was the weekly search trends and the daily hospitalization dataset of sizes (640,430) and (102912, 62) respectively. Upon removing the various missing entries from the weekly search trends dataset at a 69% threshold, we noticed that the number of symptoms dropped from 422 symptoms to 2 symptoms. Thus, we reverted to using the daily search trends dataset of size (14790, 430), indicating more datapoints for each symptom.

Our most important findings for PCA is that it is extremely difficult to interpret the complex mixture of feature data projected onto the 2 axes. We also found that the optimal clustering preformed with K-means is at a value of hyper-parameter K = 5. For the KNN regression, we found that a hyper-parameter K value of 12 provides the smallest 5-fold-cross-validation error. Furthermore, the KNN date strategy revealed that a K value of 14 results in the lowest MSE. For the Decision Tree regression, we found that the 5-fold-cross-validation error is 0.0014 at a maximum feature consideration of 20 features.

3 Datasets

3.1 Dataset Source, Pre-processing & Size

The daily search trends dataset consists of aggregate Google searches for a broad set of symptoms. This dataset can be found on the Google-research repository. A broad set of symptoms is mapped to a limited set of symptoms, from which we work with. This datasets reflects relative popularity of symptom searches within each geographical region of the United States. The pre-processing preformed includes normalizing the symptom searches for each region by the total number of search users for that region, resulting in the normalized popularity of a symptom. Then the normalized popularity of a symptom is scaled by the maximum value of the normalized popularity for each symptom and region separately. The result is a dataset of size or shape of (14790, 430) containing 422 symptoms, time-stamps, and region information. Hereinafter we will refer to this dataset as dataset 1.

The daily hospitalization dataset, also provided by Google, provides aggregate data for COVID-19 cases, deaths, tests, hospitalizations, discharges, and other attributes. This dataset can also be found on the Google-research repository. This data is for a broader geographical region, not restricted to the United States. The shape of this dataset is (102912,62) containing the region, dates, and the various attributes. Hereinafter we will refer to this dataset as dataset 2.

3.2 Dataset Processing

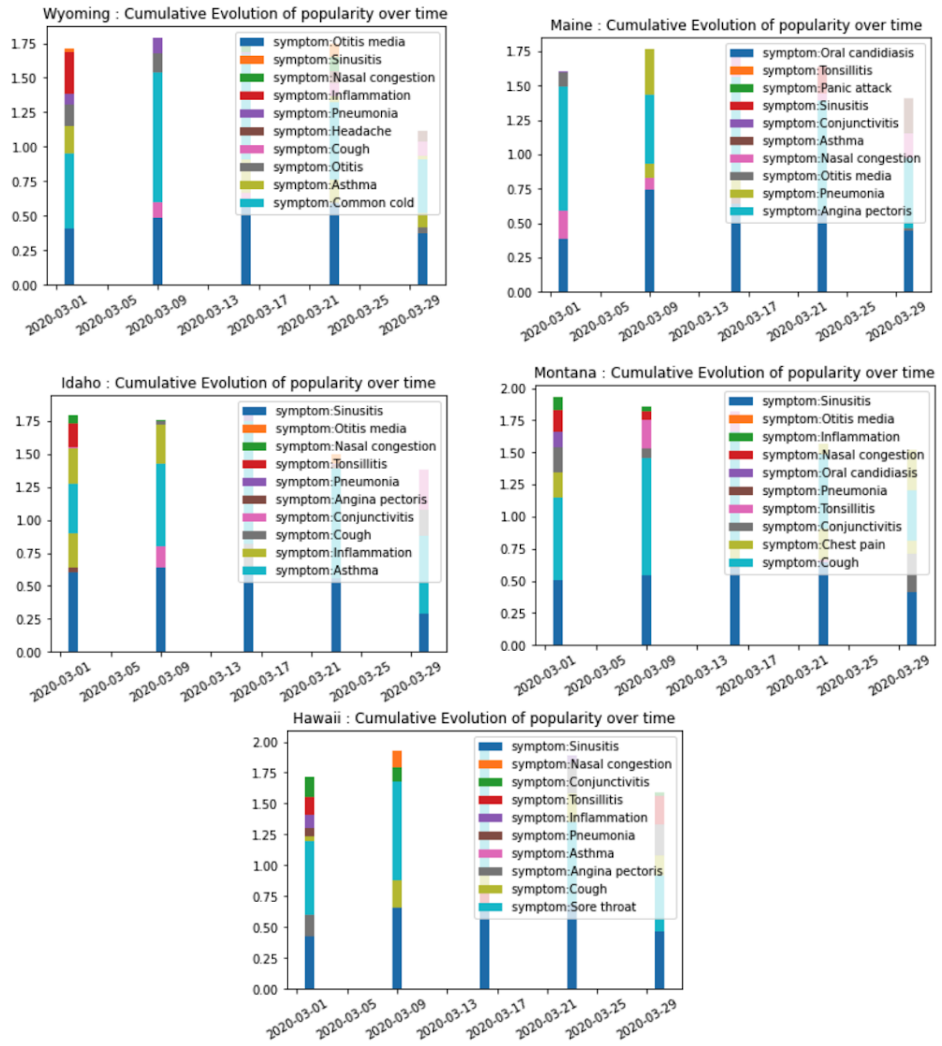
In our project, we had noticed that this step is the most vital to obtaining accurate results, thus we had spent the majority of our time in this step. The processing we preformed is as follows. Both datasets were loaded as Pandas dataframes using the URL's by utilizing the CSV library. For dataset 2, we only keep the regions which have at least 1 non-zero value for 'hospitalized new' column, as we have selected that as our primary label. We then remove the missing data 'NaN' at a 95% threshold for both datasets 1 & 2. We then Z-normalize dataset dataset 2 and collapse it to a weekly resolution. Then, dataset 1 is Z-normalized by region for each symptom, in order to allow across-region comparisons. Then, dataset 1 is collapsed to weekly resolution, and datasets 1 & 2 are merged. Moreover, we eliminate rows before 2020/03/02 and rows after 2020/09/08 due to date incompatibility in both the datasets (If this wasn't done, new NaN values would arise after merging). Finally, we normalize the merged dataset using Min-Max normalization in order for the final data to be within the range of 0 to 1 for each symptom. We also fill the remaining NaN values with the mean of the merged dataset for each symptom to deal with the remaining missing values.

4 Discussion & Results

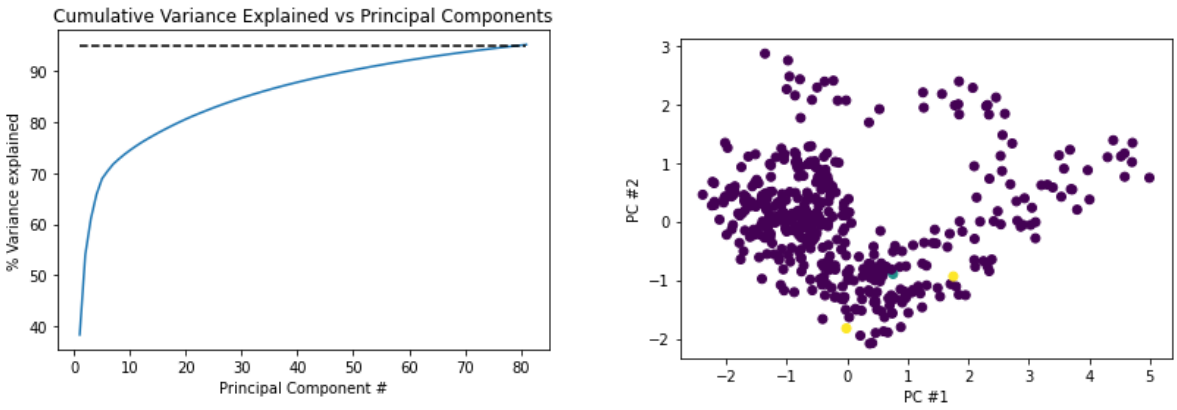
4.1 Data Visualization Discussion & Results

To visualize the evolution of popularity of various symptoms across different regions over the time,we utilized stacked bar plot from matplotlib.We first choose the top ten most frequent symptoms from our cleaned dataset and then use it on

different regions. This is shown for a few regions in the figures below.

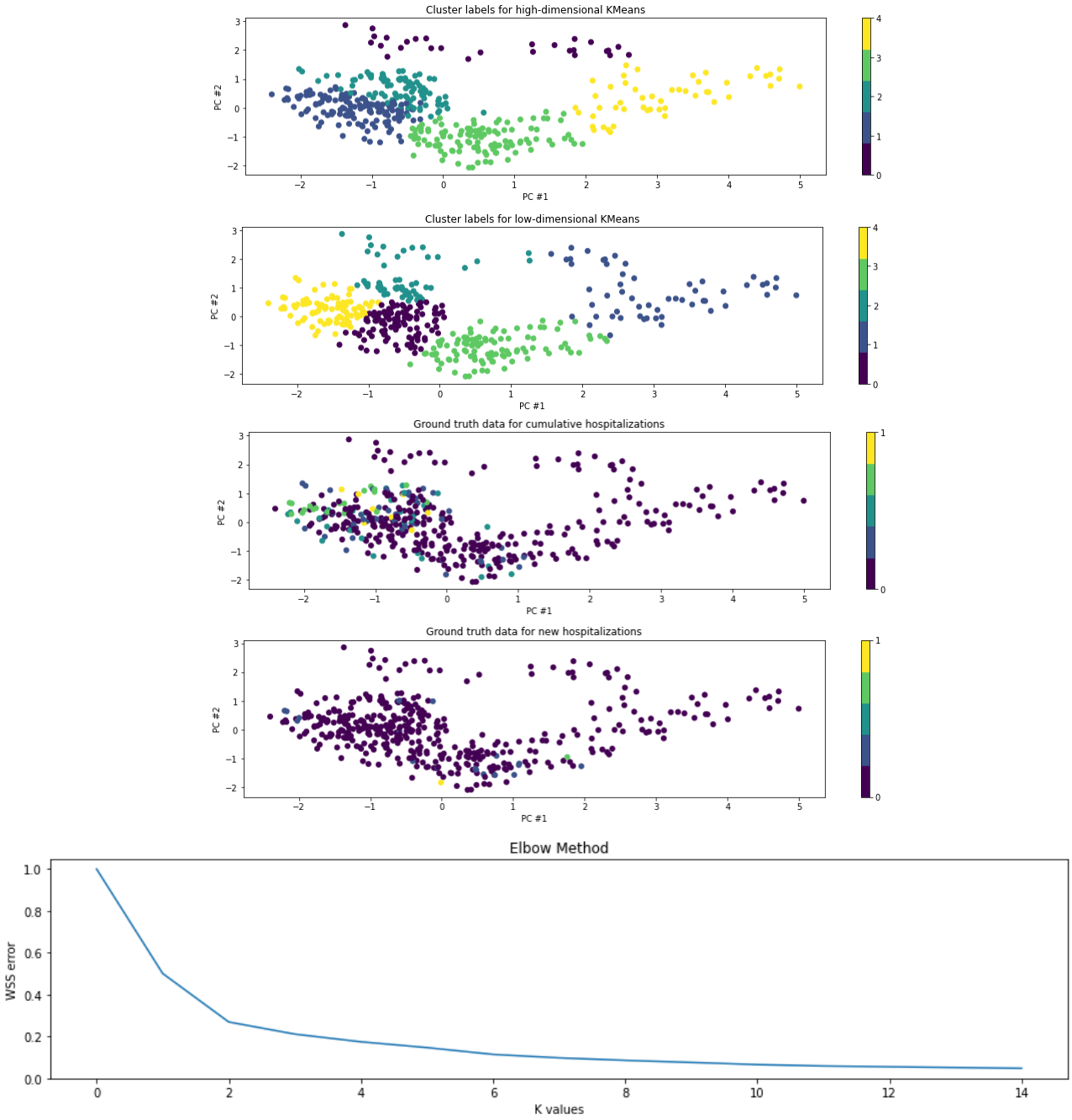


To further analyze the data, we must reduce the higher dimensional data into a lower dimensional space. This is done with PCA, applied to the features with no information from the classes. This indicates that PCA is an unsupervised learning method. We first scale the data using the min-max scaler. Then we determine the number of components or features to keep. To do this, we analyze the Cumulative variance against the number of principal components, and find the point of 95% variance. The 95% variance happens to occur at 80 principal components, as shown in the figure below. This means that all the components that cover 95% of the information within the higher-dimensional data are selected. The plot below also showcases the reduced components plotted. We found it difficult to extract information from this plot, as it is a complex mixture of the original features. The shape of the reduced feature dataframe is thus (434, 81).



The Variance for the first few components is: [0.38408167 0.15651706 0.07017481 0.04731291 0.03086474 0.01513891]

Finally, the last unsupervised learning method explored is K-means. Although the labels are utilized here, we do not explicitly label the data, and is hence still an unsupervised learning method. K-means is used to cluster the PCA reduced data. We see that higher dimensional data provides inaccurate clustering in lower dimensions, which hints at the need of the dimensional reduction. The plots below clearly show that the clustering is exponentially more accurate in lower dimensions than in higher dimensions. Furthermore, in order to determine the optimal value of hyper-parameter K, we use the Elbow method. In this method, we plot values of K against a measure of distance for datapoints in clusters. The measure of distance we used is Euclidean distance shown as the WSS. Our plots show that the scaled WSS error significantly drops at around K = 5 to 14.5%. This optimal value of K = 5 is used for all the cluster plots below. Moreover, we provide the ground truth data plots for cumulative hospitalizations & new hospitalizations.

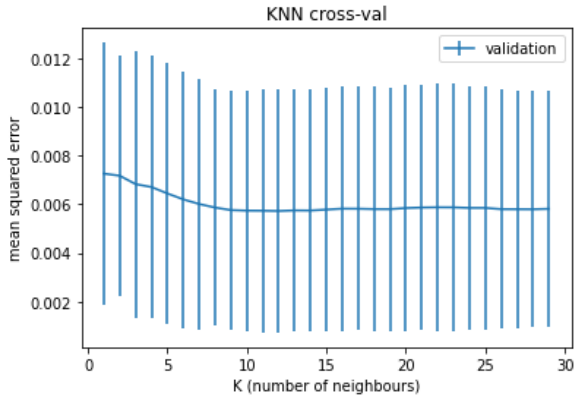


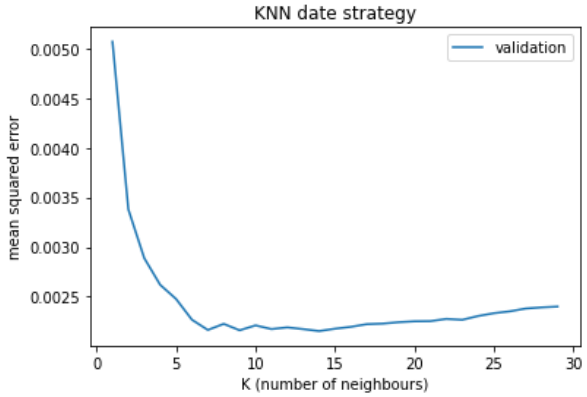
Note that the WSS axis is normalized, where 1.0 corresponds to 100% error.

It is noteworthy that our PCA and K-means algorithm sometimes doesn't converge due to Singular value decomposition not converging. This is a weird bug we encountered, which seemed to be fixed by either re-running the algorithm or running K-means in a separate cell.

4.2 Supervised Learning Discussion & Results

In this section we provide the results for our supervised learning models K-NN and Decision Trees regression(s). We split the data into a training and validation set using two strategies, based on regions and based on time. We perform an 80% - 20% split for the training set and validation set respectively. We selected the 'new hospitalization' as our label. Furthermore, we use 5-fold-cross validation, MSE, and other model-specific metrics to assess the performance of these models. We will refer to 5-fold-cross validation error as cross validation error.
In the KNN regression model, we find that the K value of 12 achieves the lowest cross-validation error of 0.00572, within one standard deviation of error. The date strategy reveals that at a value of K = 14 the lowest MSE of 0.00215. We selected the model with K = 12 that has the lowest cross validation error. The figures for the KNN cross-validation error and date strategy MSE are shown in the figures below.





In the Decision Trees regression model, our optimal results are found by setting the max features to 20, which indicates the maximum number of features to consider when looking for best splits. Moreover, the maximum length of the tree was left unbounded. We find the mean corss-validation error is 0.0133. Furthermore, our date strategy for this model reveals that the MSE is 0.0026, MAE is 0.0367, and RMSE is 0.0515.

5 Conclusion

In this project, we compared the performance of two methods for regression, namely K-NN and Decision Trees using 5-fold-cross-validation and other error metrics. We started by cleaning and normalizing the datasets using region specific Z-normalization and Min-max normalization. We then visualized the data over time, used PCA to reduce the dimensionality of the data and plotted it, and deployed K-Means to cluster the data. We've determined that our models for KNN and Decision Trees have significantly low cross-validation errors of 0.0057 & 0.0133 respectively.

We have come to realize that cleaning and pre-processing the data is the most vital aspect of any data analysis project. Hence, we spent most of our time trying to achieve clean datasets. We all gained experience working with powerful python libraries, Pandas for big dataframes, and Scikit-learn for data analysis. Moreover, we all gained experience in data visualization techniques, which significantly improves our understanding of data science in general.

In future implementations we would spend more time on tasks 1 & 2, emphasizing visualizing the merged dataset by deploying other techniques. PCA does not uncover much information aside from the variance of certain components. K-Means provide a good way to cluster the data, but since it's an unsupervised learning algorithm, without labelled data, it's difficult to interpret these clusters. We would also explore cross-validation for folds other than 5. Moreover, we would tweak the model hyper-parameters to gain deeper understanding and potentially select better models for KNN and Decision Trees.

6 Statement of Contributions

The work on this project was split among all three team members. We all worked on Task 1 equally. Iyatan worked on Task 2.1, Alexandre on Task 2.2, & Ameer on Task 2.3. Task 3 was entirely done by Alexandre. The report write up was done by all of us equally.

7 References

- Google LLC "Google COVID-19 Search Trends symptoms dataset". http://goo.gle/covid19symptomdataset, Accessed: 2020/10/21.