COMP551 Mini Project 1

Grace Hu 260776936
Xingyu Chen 260786048
Jiahui Peng 260782511

October 21, 2020

1 Abstract

In this project, we explored 2 COVID-19-related datasets, and used Google symptoms search trends data to predict weekly COVID-19 new hospitalization cases in different regions across the world. We investigated the performance of 2 regression models, K-nearest neighbours (KNN) and decision trees, and found that KNN achieved better performance than decision trees.

After cleaning the data and removing features (symptoms in search trends) with too many NaN values, we employed techniques such as recursive feature elimination, PCA decomposition, N-fold cross-validation, and feature selection to improve the accuracy of our models' predictions. The data was split into training and testing sets in two ways: by region (e.g. delegating 80% of regions to training set, and 20% to validation set) and by time points.

With our **KNN**, we achieved an mean square error of around 5000 when splitting the data by regions, and we achieved an mean square error around 2000 when splitting the data by time points. We achieved a mean square error of 1510.64 with the K=18 when splitting the data by regions, and a MSE of 5605.54 with K=28 when splitting the data by time points.

With our **decision tree model**, we achieved an mean square error around 15000 when splitting the data by regions, and we achieved an mean square error around 6000 when splitting the data by time points.

2 Introduction

There have been previous research done with Google search trends data, such as Walker, Hopkins, and Surda's 2020 paper which investigated loss-of-smell related searches to identify the symptom as a potential effect of COVID-19 [1]. The 2020 Sousa-Pinto et al. paper also examined Google search trends for 4 keywords coronavirus, cough, anosmia, and ageusia over 5 years and in 2020 to establish a correlation with media coverage of COVID-19 and these keywords [2].

In our project, we're predicting the prevalence of hospitalizations related to COVID-19 from the search trends data. Our research question is as follows: Is there a connection between symptom-related Google searches and COVID-19 cases? We're hypothesizing that an increase in searches for certain symptoms is related to an increase in new COVID-19 hospitalization cases.

To inform ourselves about how to develop an accurate machine learning model, we explored the importance of feature selecting for huge data set by reading the paper written by Dash and Liu [3]. Later on, we utilised this idea to improve the performance for both regression models, KNN and decision tree.

Instead of using the percentage accuracy to display the performance of the models, we used mean square error which more pertinent for regression models.

3 Datasets

Both datasets originated from Google, and we combined the October 4th versions of the two datasets (daily aggregated international COVID-19 hospitalization data [4], weekly US Google search trends data [5]) to obtain our training and validation datasets. Since the resolution of the time series was different for the two datasets, we aggregated the daily hospitalization data into weeks by summing up the number of "hospitalized_new" cases.

We also made sure the two datasets shared the same time points. For example, if the hospitalization dataset provided daily data for 2020-01-01 to 2020-02-30 and the search trends dataset provided weekly data for weeks 2020-01-06, 2020-01-13 and so forth, we discarded the daily data from 2020-01-01 to 2020-01-05 and began aggregating the days starting from 2020-01-06 in the hospitalization dataset.

We also cleaned the data by eliminating NaNs: for the hospitalization dataset, we removed all regions that contained 100% NaNs in the "hospitalized_new" column and thus were irrelevant. For the search trends dataset we removed all symptoms whose columns contained more than 40% NaNs, and the remaining symptoms became features for the machine learning models. A total of 407/422 symptoms were removed, with 15 symptoms remaining.

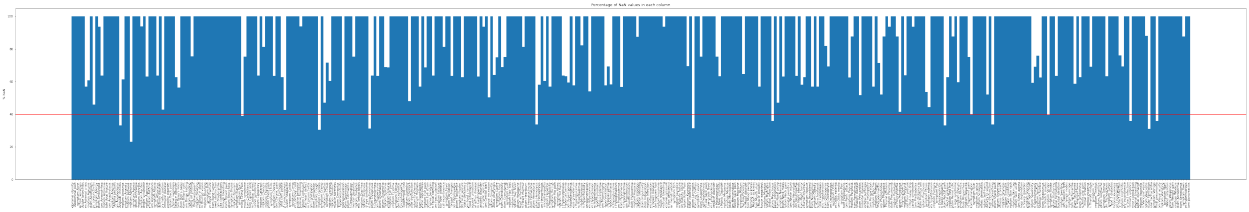


Figure 1: Bar plot of NaN % of symptoms, red line is 40% NaN

Since the values in the search trends dataset were previously normalized by the total number of searches in a region and linearly scaled to a value between 0-100 [6], we reversed this normalization so we would be able to compare values across regions. We did this by calculating the median for each symptom in each region, then taking the medians of these symptom medians and dividing all values in that given region by that final median.

We performed an inner join between the hospitalized_new column and the cleaned search trends dataset on the keys open_covid_region_code and date to obtain the final dataset for our model to use.

4 Results

4.1 Principle Component Analysis and Clustering

In order to determine the optimal number of clusters for the KMeans clustering algorithm, we conducted a simple hyperparameter search by iterating over possible values of k. We can see in Figure 2 that 5 was the optimal number of clusters or the optimal value of k.

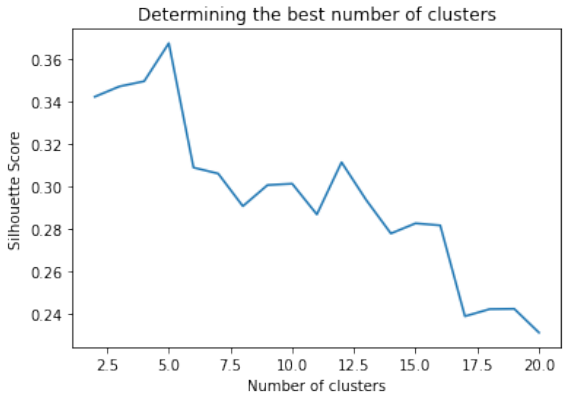
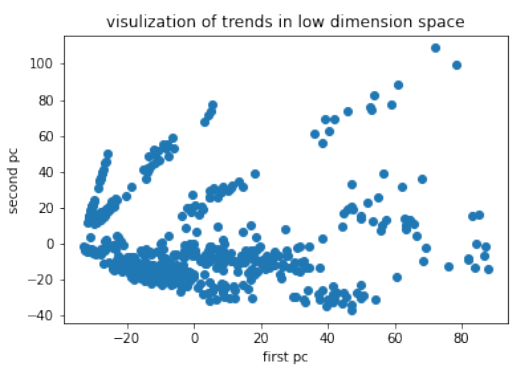
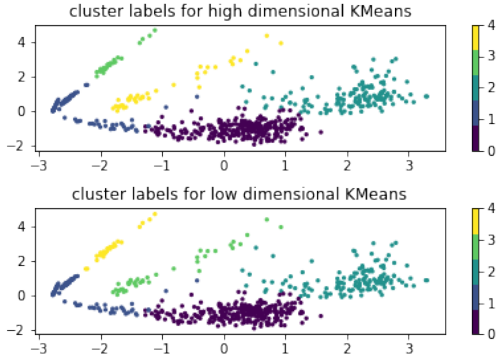


Figure 2: Hyperparameter Tuning for KMeans

We present in figure 3(a) the visualization of search trend data using the first 2 principle components. Figure 3(b) shows the clustering results of the original and PCA-reduced search trend data. From the colored label distribution we can infer that the divided groups remain consistent for original and PCA-reduced data.



(a) PCA visualization



(b) K-means clustering of high and low dimensional data

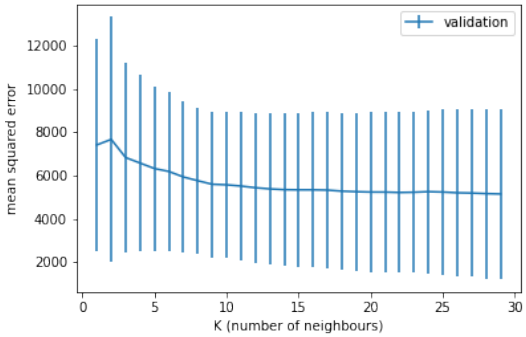
Figure 3: PCA and K-means

4.2 Comparison between KNN and Decision Tree

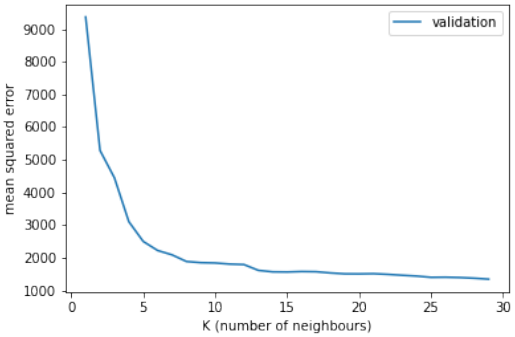
In this section, we split the data based on region and time respectively. 2 different supervised learning models are used: KNN and decision tree. Here we show the performance of each model on 2 differently split train-valid sets and compare the 2 models. The performance is shown by Mean Square Error (MSE) of the validation set. We also try to improve the model performance by removing some features in the merged dataset. (Creativity)

4.2.1 K-Nearest-Neighbour

For time-split data, figure 6(a) shows the validation error of different k values. k=18 gives the best validation error. Therefore, we chose k=18 for the KNN model on time-split data. At k=18, the MSE is 1510.64. For region-split data(fig. 6(b)), we used 5-fold cross validation and k=28 because this k value gives us the best as well as the simplest model within one standard deviation of the lowest validation error. The MSE at k=28 is 5605.54. Therefore KNN model performed better in time-split training set.



(a) region-split

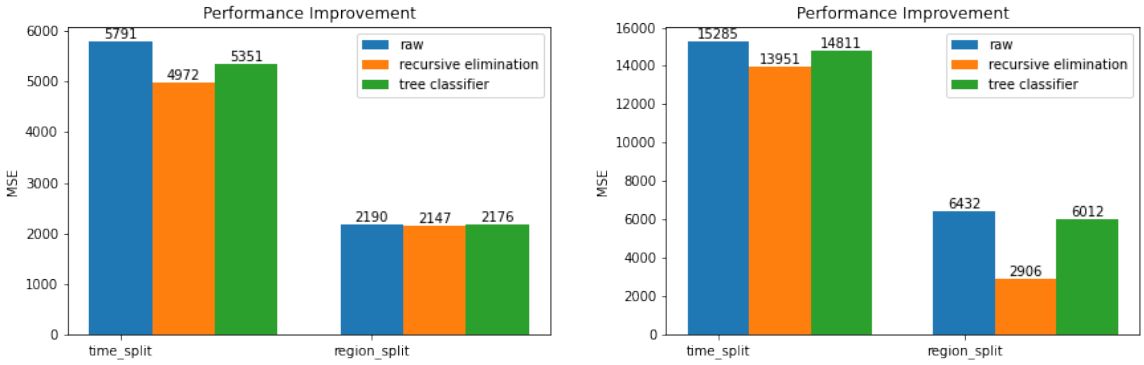


(b) time-split

Figure 4: KNN performance on region and time-split data

4.2.2 Further Exploration: Improving Performance on Selected Features

As a creative extension to our project, we used Extra Tree Classifier to extract the top 10 features for the dataset. Also, we used to Recursive Feature Elimination (RFE) to compare the performance which is recursively remove the unimportant features. Figure 5 presents the comparison of the performance of KNN and Decision Tree model on original features and selected features.



(a) KNN

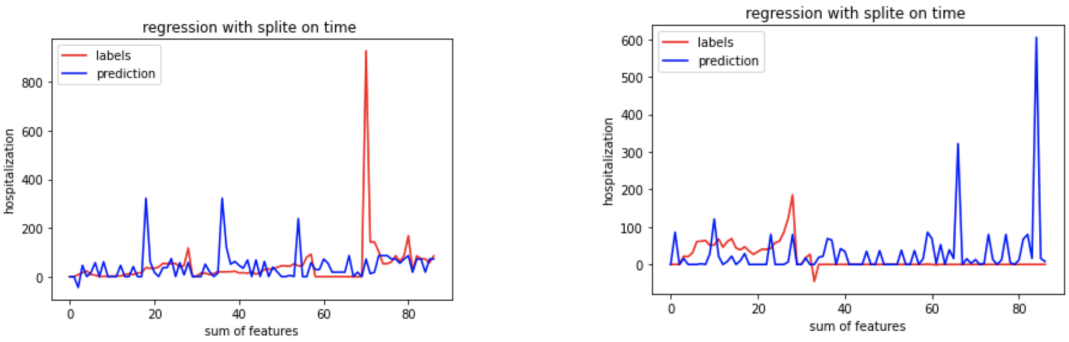
(b) Decision Tree

Figure 5: Performance of 2 models on original and selected features

In KNN model, we took the average MSE over 30 k values to assess the performance. In time-split training set, the MSE decreased by 14.16% and 7.60% by recursive elimination and extra tree classifier respectively. In region-split training set, the MSE decreased by 1.96% and 0.63% by recursive elimination and extra tree classifier. In decision tree model, after we removed features by recursive elimination, the MSE decreased by %8.72 and %54.82 in time-split and region-split training set. Via extra tree classifier, the MSE decreased by %3.10 and % 6.53.

4.2.3 Decision tree

The mean-square error for splitting by regions is higher than splitting by time. The MSE for the former is around 15000 and for the latter is about 6000. One of the possible reasons to explain this phenomenon is that the data is more scattered across regions whereas some of the regions have much more hospitalization than the average hospitalization of all states. The figure 6 are obtained during the cross-validation. Both the labels and predicted values have spikes which is caused by the uneven distribution of hospitalization across regions.



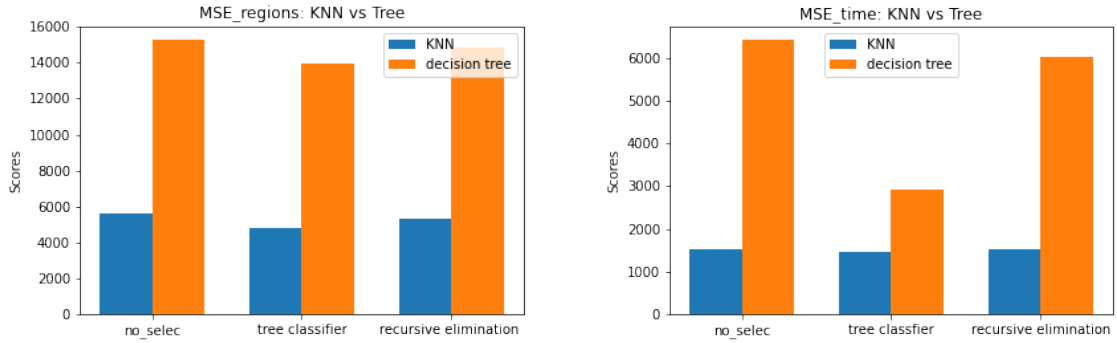
(a) region-split

(b) time-split

Figure 6: Decision Tree Performance on region and time-split data

4.2.4 Comparison between KNN and Decision Tree

Figure 7 shows the MFE values of KNN and Decision Tree model on 2 training set splitting scheme. On both original and selected features, the KNN has smaller MSE than Decesion Tree model and thus KNN has better performance.



(a) Region-split

(b) Time-split

Figure 7: KNN vs Decision Tree

5 Discussion and Conclusion

The purpose of this miniproject is to compare 2 models on the search trend symptom dataset by using 2 different splitting scheme to predict the weekly hospitalization.

From the project, we learned the importance of pruning data. As the model is fixed, besides tuning the parameter we can also improve performance by reducing irrelevant and rarely filled features. We removed the features with over 50% of NaN filled in and selected the most relevant feature. The observation shows a decrease in MSE. Moreover, the way of splitting the data also plays an important role. In our observation, splitting by time performs better than splitting by regions due to the characteristics of data where the symptoms are dependent on time.

The feature investigation could be normalizing across regions since the hospitalization in some regions is very far away from the mean. The normalization would bring them closer to the mean and therefore result a better performance when splitting by regions.

6 Statement of Contributions

Grace Hu was in charge of Task 1 and writing the report, Xingyu Chen was in charge of Task 2 and developing the KNN model in Task 3, while Jiahui Peng was in charge of developing the decision tree model in Task 3 and writing the report.

References

[1] Abigail Walker, Claire Hopkins, and Pavol Surda. "Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak". In: *International Forum of Allergy & Rhinology* 10.7 (2020), pp. 839–847. DOI: 10.1002/alr.22580. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/alr.22580.

[2] Bernardo Sousa-Pinto, Aram Anto, Wienia Czarlewski, et al. "Assessment of the Impact of Media Coverage on COVID-19–Related Google Trends Data: Infodemiology Study". In: *J Med Internet Res* 22.8 (Aug. 2020), e19611. ISSN: 1438-8871. DOI: 10.2196/19611. URL: http://www.ncbi.nlm.nih.gov/pubmed/32530816.

[3] M. Dash and H. Liu. "Feature Selection for Classification". In: *Intelligent Data Analysis* 1 (1997). 3, pp. 131–156. ISSN: 1571-4128. DOI: 10.3233/IDA-1997-1302. URL: https://doi.org/10.3233/IDA-1997-1302.

[4] Google Research. *Open COVID-19 Data*. URL: https://github.com/google-research/open-covid-19-data.

[5] Google LLC. *Google COVID-19 Search Trends symptoms dataset*. URL: http://goo.gle/covid19symptomdataset.

[6] Shailesh Bavadekar, Andrew Dai, John Davis, et al. *Google COVID-19 Search Trends Symptoms Dataset: Anonymization Process Description (version 1.0)*. 2020. arXiv: 2009.01265 [cs.CR].