**Case Study 1: Clustering and Classification of Covid Data**

*Ben Appelbaum, Leo Kaiser*

*ESE 105*

*9/30/2023*

***Abstract -* In case study 1, we were given a set of COVID case data by county. It was our goal to develop a method of classifying this data using the k-means algorithm so that we could determine what region of the United States a given COVID vector was most likely from. In order to do this, we split up our training data based on geographic division. We then generated three centroids for each geographic division of the United States and compiled them all together as our starting centroids for the main k-means algorithm on the full training data set. This allowed us to start k-means with some sense of what we wanted our final centroids to correspond to. Using this method gave us an improved accuracy compared with random starting centroids. We were able to achieve an 93% accuracy when classifying our testing data.**

1. **Introduction**

Accurately clustering data is a valuable skill in scientific research. Algorithms such as k-means can unearth important insights about the relationships within a dataset. But, if you are not clustering data well, you can miss patterns or come up with connections that do not exist. The main challenge of this case study is that it was not enough to run the k-means algorithm with randomly set centroids, meaning we had to develop a method for choosing starting centroids that improved the accuracy of clustering. This case study gives us experience with developing and testing data clustering methods.

In addition, this project has important applications when it comes to modeling the spread of infectious diseases, COVID in this case. Being able to uncover information about the different ways diseases spread in different parts of the country can tell you a lot about the efficacy of different quarantine policies.

1. **Methods**

In order to achieve a high accuracy, we needed an algorithm to produce starting centroids that would lead to clustering of COVID data based on geography. Since there is no way to control how the starting centroids change when running the k-means algorithm, it is important to use starting centroids that have some sense of the way you want to cluster the data. Therefore, we wanted our starting centroids to be spaced out within each geographic division. In order to achieve this, we split up the main dataset into nine datasets for each geographic division. We then ran the k-means algorithm on each of these nine datasets with three centroids. This gave us three group representatives for each division. We then ran the main k-means algorithm on the entire dataset with all twenty-seven centroids as our starting centroids. We also implemented a method to track which centroids are supposed to correspond to each division since using more than one centroid for each division means it can be difficult to keep track of everything.

A group of different types of normalized blood pressure

Description automatically generated with medium confidence

* *Graph of all centroids for each geographic division.*

Looking at the above graphs. You can see the relationships between the centroids within each division. For example, the small spike right before the largest spike in cases is much smaller in the Pacific division compared to the South Atlantic division.

1. **Results and Discussion**

We achieved an 93% accuracy on classifying COVID data by division using our testing data set. This is a much more accurate result compared with chance, strongly suggesting that our method has been successful in creating accurate centroids for each division. Additionally, looking at our centroids, we have three centroids that correspond to a division for most of the divisions. So, our method of finding three centroids for each division has also been successful.  Some limitations with our method are that it does not deal with noisy data well. Later in the pandemic, certain counties decreased the frequency at which they reported numbers. This led to inconsistent and noisy data. The inclusion of this data in our training set meant that one of our centroids corresponded to this messy data and decreased the overall accuracy.

A graph showing a number of data

Description automatically generated

* *Inconsistent data reporting in the South Atlantic*

A graph showing the number of points

Description automatically generated

* *Inconsistent reporting leading to incorrect classification.*

1. **Conclusion**

In order to achieve the goal of this case study, there must be a meaningful relationship between the COVID data within regions. If there was no relationship, our clustering method would not yield accurate results. Achieving a 93% accuracy when clustering COVID test data based on region indicates that there is a relationship between COVID data in and the geographical division that data comes from.

Additionally, looking at the centroids for each region gives information about …