Community Clustering

Introduction

The main aim of the Community Clustering task is to group the Covid (Country-level) data collected from different sources into similar groups. What this would enable is the ability to take similar policy decisions and other needed actions to ensure better well being of people. It shows the area that show similar behaviour on certain factors such as :

- No of deaths
- Hospitalization Rate
- Economic harm caused (loss of employment, organizational losses(revenue))
- .... and many more

It is because the regions belonging to any specific cluster will have similarities with other members of the groups and thus the actions taken for any specific member of the cluster might also be helpful for many others too.

Research Questions to answer?

1. Which clustering algorithms to use and decide the best number of clusters to be made?
2. What are the most important features contributing towards the formation of the clusters?
3. What was the quality of clustering done?
4. What are some insights about the clusters formed?

ML Approached used:

DATA PREPROCESSING:

Variance Thresholding is used to drop low variance data. It means that all those columns are dropped which don't have much feature variation. Since all the points are very close so they don't provide much information for the model to make predictions. So we keep data that have a certain minimum threshold value.

Correlation test is used to drop highly correlated values or features. Highly correlated features usually provide similar information so having just one of them is good enough. After dropping highly correlated data(threshold=0.5) . The number of features were dropped from 93 to 32 columns

The features left after applying correlation test are:

Deaths	Average Daily PM2.5	Average Reading Performance
Years of Potential Life Lost Rate	% Severe Housing Problems	Black/White Segregation Index
% With Access to Exercise Opportunities	% Drive Alone to Work	Suicide Rate (Age-Adjusted)
% Driving Deaths with Alcohol Involvement	% Long Commute - Drives Alone	Juvenile Arrest Rate
Chlamydia Rate	Life Expectancy	% less than 18 years of age
Primary Care Physicians Rate	Child Mortality Rate	% American Indian & Alaska Native
% With Annual Mammogram	HIV Prevalence Rate	% Asian
High School Graduation Rate	Drug Overdose Mortality Rate	% Native Hawaiian/Other Pacific Islander
Social Association Rate	% Uninsured	% Hispanic
Violent Crime Rate	% Disconnected Youth	% Female
internet_ratio	Hypertension Death Rate	

Methodology:

Steps Followed:

1. K means clustering has been used so far as it is fast .
2. The best number of clusters to be kept was found to be 4. It means that the data should ideally be divided in 4 parts only.
3. This performance has been verified using some clustering performance metrics such as : Inertia, Silhouette Score, Davis_bouldin_score,etc.
4. These ensure that the clusters formed have :

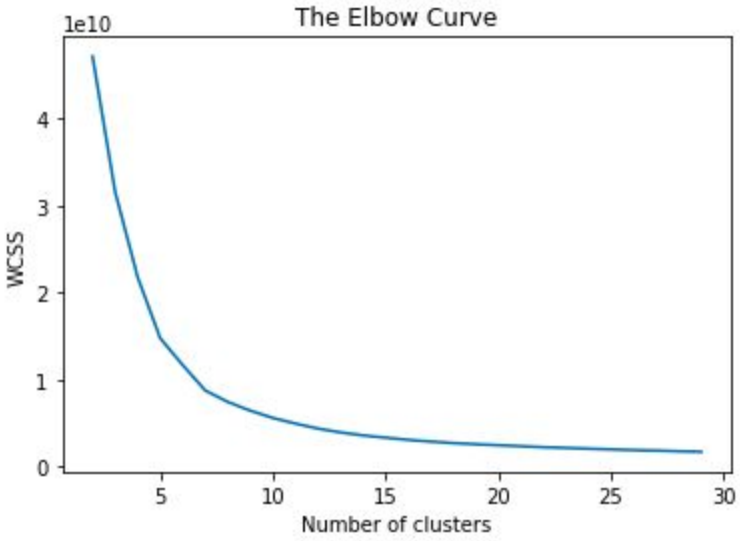
   high density , significant difference from each other (clusters are far) and many others.

K-means clustering has been taken as the choice for the partitive algorithm with kmeans++ initializer. The k-means algorithm is implemented for a number of clusters in range 2-10. The optimal number of clusters to be formed is decided based on multiple criterias starting with the elbow method.

Elbow Method:

The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.



The elbow is formed at 4 and 6.

Silhouette Coefficient

Silhouette Coefficient score is used in cases when True labels are not present the clusters are to be evaluated using the model only. A higher coefficient score refers to better clusters formed. The Silhouette Coefficient is defined for each sample and is composed of two scores:

1. The mean distance between a sample and all other points in the same class.
2. The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample. The scores vary in range of -1 to 1 with following meanings:

Score near	**Interpretation**
-1	Incorrect clustering
0	Overlapping clusters
1	Highly dense and separated clusters

The silhouette score for 4 clusters was ~0.48. It signifies clusters are separated with a little overlap.

Supervised Analysis:

To check for the efficiency of clusters made, the predicted labels using k means are used as a target column and the classification report is generated. ExtraTreeClassifier has been used to check for the feature importance score and classification model performance on the dataset. This classifier implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

Fig: Classification report when using ExtraTreeClassifier.

	precision	recall	f1-score	support
0	0.96	0.92	0.94	354
1	0.94	0.97	0.95	553
2	0.89	0.70	0.78	23
3	1.00	0.96	0.98	93
accuracy			0.95	1023
macro avg	0.95	0.89	0.91	1023
weighted avg	0.95	0.95	0.95	1023

Above reports show a good model performance in predicting new samples obtained and assigning them correct labels.

Fig: Top 10 features contributing the most to model performance

Feature_name	Feature_importance_score
Years of Potential Life Lost Rate	0.113612
Deaths	0.034183
Child Mortality Rate	0.028482
internet_ratio	0.027169
% With Access to Exercise Opportunities	0.02209
% Uninsured	0.019218
% With Annual Mammogram	0.018246
Suicide Rate (Age-Adjusted)	0.016503
Average Daily PM2.5	0.01439
Hypertension Death Rate	0.014173

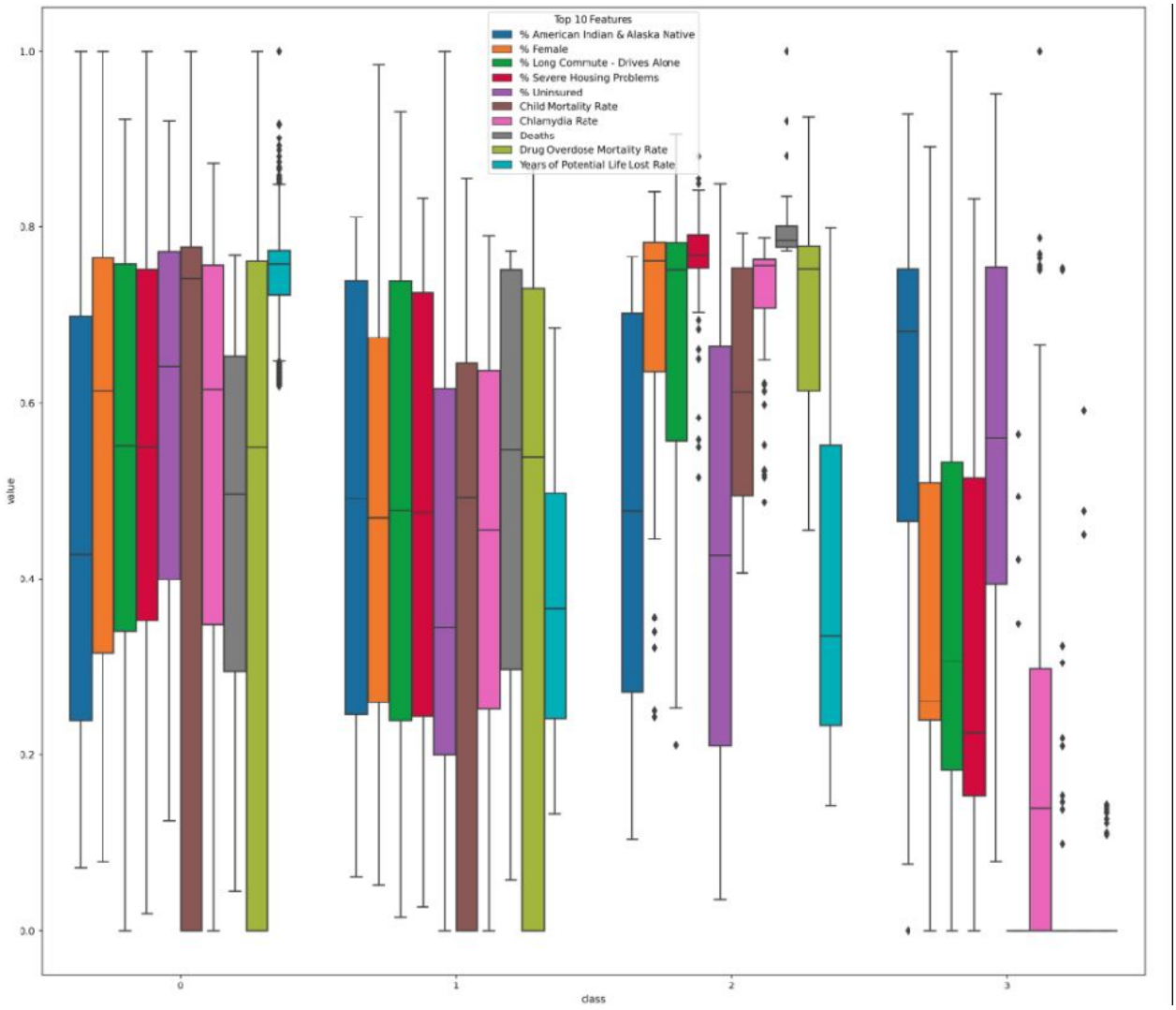
Then individual clusters were analysed to find out which features were responsible for them. It was done as follows:

1. For any class label , follow steps 2-3
2. Convert that class label to 1 and all others to 0, making it a binary classification problem. E.g. points in the first cluster vs points not in the first cluster (second,third and fourth).
3. Randomly sample points from both the classes and divide them into training and testing data. Using Recursive Feature Elimination (RFE), find the top ten most important features and store them.
4. Repeat the steps for all other initial class labels.

All the features are stored in a table. It can be seen from the table that individual clusters have only some different features than the other clusters and not all the same. It ensures that clusters formed draw out different meanings and are not just based on the same features.The following features were found to be most important for the individual clusters:

First Cluster	Second Cluster	Third Cluster	Fourth Cluster
Deaths	Deaths	Deaths	Hypertension Death Rate
Years of Potential Life Lost Rate	% Asian	Years of Potential Life Lost Rate	% Native Hawaiian/Other Pacific Islander
% Asian	Drug Overdose Mortality Rate	% Hispanic	% Asian
% Uninsured	Child Mortality Rate	Black/White Segregation Index	% American Indian & Alaska Native
Drug Overdose Mortality Rate	% Long Commute - Drives Alone	% With Annual Mammogram	% less than 18 years of age
High School Graduation Rate	Violent Crime Rate	High School Graduation Rate	internet_ratio
Child Mortality Rate	HIV Prevalence Rate	internet_ratio	% Hispanic
Violent Crime Rate	Years of Potential Life Lost Rate	Violent Crime Rate	Social Association Rate
Average Daily PM2.5	% Driving Deaths with Alcohol Involvement	Average Daily PM2.5	% Female
% Long Commute - Drives Alone	Chlamydia Rate	Child Mortality Rate	Years of Potential Life Lost Rate

Boxplots are also constructed clusterwise for the top 10 most important features. It is shown in the diagram below:



Observations:

Some of the observations about each cluster that could be made are:

Cluster 1:

1. People in this group have lost a high amount of their days from what their previous life expectancy could be.
2. There are more womens in this group who don't have very long commutes and are also not affected much by housing problems.
3. The death rates are moderate with less children dying than people with drug-overdose.

Cluster 2:

1. This cluster has an almost equal number of males and females who take shorter commutes alone.
2. People face a little housing issues and the group has the most percentage of people having insurance.
3. Death rates are higher than cluster 1 but people have only lost a small amount of their previous life expectancy.

Cluster 3:

1. Females make up most of this cluster with severe housing problems and having to commute over longer distances.
2. This group has a very high death rate (highest among all) . It has a high Chlamydia Rate with a large section of death resulting from drug overdose.
3. Strangely enough, the high death rates have not caused a significant decrease in the life expectancy of people.

Cluster 4:

1. This group consists more of males who are American Indian & Alaskan Natives.
2. It has low housing problems with no major disease outbreak like Chlamydia. It has a lot of outliers for multiple features.
3. It has a low death rate with people requiring shorter commutes.