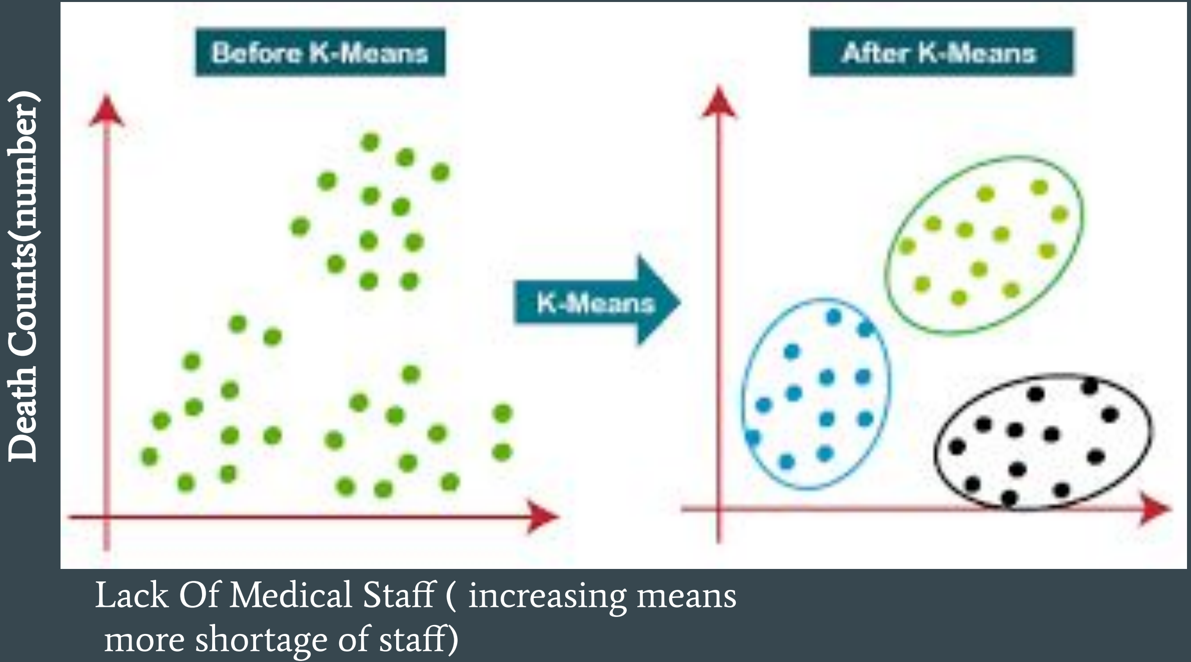
Community Clustering

Data used

1. Merged Data Csv (updated)
2. Quantile_normalized merged_data

...

** For each clusters I have also stored the labels in a column class. This tells what rows are grouped together. It has been kept so someone else could utilize it. Please find it here : link



K means is used to group similar data points together in groups also called clusters

****Note** Here the labels are taken for representation purpose only, they don't define true relationship between the two labels

What K means did and what it told us ?

1. The x and y axis are any two features that are to be tested. E.g. Death Count vs Lack of medical Staff.

2. The K means algorithm splits the data into three parts that are : **(Right figure)**

BLUE=> Less shortage of medical staff and low death rates (More staff can effectively control and increase chances of survival)

GREEN=>More deaths are caused if there is large shortage of doctor and the death count rises with increasing degree of shortage (Complimentary to previous one)

BLACK=> The third group is ideal one . Here the shortage of doctors is not causing death toll to rise . (No logics but maybe some other factor is contributing to it like people following good practices to prevent the spread itself.)

Here we can see that Black cluster is an interesting group. Analysis it further can give better insights to dealing with the problem.

Data Preprocessing Done

1. **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.
2. **Correlation test:** is used to drop highly correlated values or features. Highly correlated features usually provide similar informations 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 25 columns(after normalization on data).

Note :I have also stored the name of the features dropped for various correlation and variance values.

** See code for more details

Remaining features

q_Years of Potential Life Lost Rate
q_% With Access to Exercise Opportunities
q_% Driving Deaths with Alcohol Involvement
q_Chlamydia Rate
q_% With Annual Mammogram
q_% Vaccinated
q_High School Graduation Rate
q_Social Association Rate
q_Violent Crime Rate
q_Average Daily PM2.5
q_Presence of Water Violation
q_% Severe Housing Problems
q_% Long Commute - Drives Alone

q_Average Reading Performance
q_Suicide Rate (Age-Adjusted)
q_Juvenile Arrest Rate
q_% less than 18 years of age
q_% Native Hawaiian/Other Pacific Islander
q_% Female
q_Hypertension Death Rate
q_% workers commuting by public transit
q_% Veterans in Civilian Adult Population
q_Child Mortality Rate
q_% Children Uninsured
q_Other Primary Care Provider Ratio

Algorithm Used

1. K means clustering has been used so far as it is fast .
2. The best number of clusters to be kept was found between 3-6 . It means that the data should ideally be divided in 3-6 parts only.
3. This performance has been verified using some clustering performance metrics such as : Inertia, Silhoutte Score, Davis_bouldin_score,etc.
4. These ensure that the clusters formed have :

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

Making Supervised cluster analysis

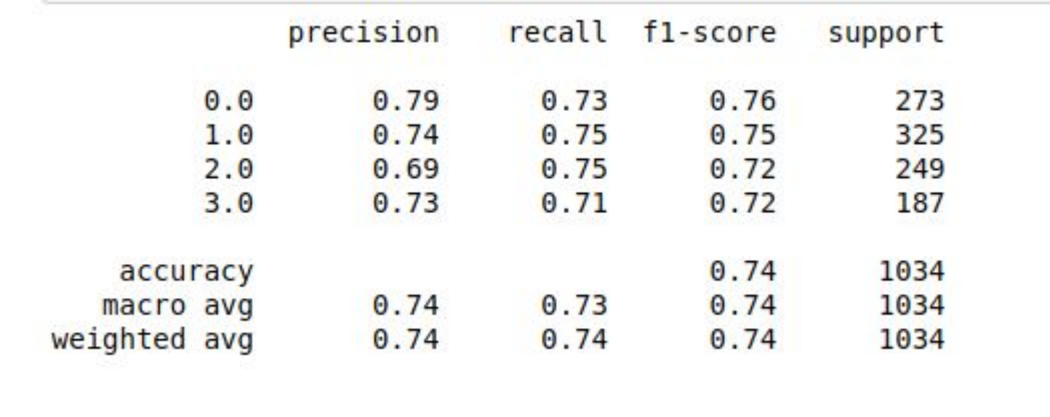
The predicted labels are used as the class labels for each clusters and this class is added to original data.

All new predictions are made on this data only. The following tests have been done:

1. What are the testing performance based on (accuracy,precision, etc)
2. Feature importance analysis for each cluster label.
3. Feature importance for every individual cluster for by considering it as a binary classification problem.

Cluster Performance

In case of 4 clusters for quantile normalized data. Class-wise cluster performance

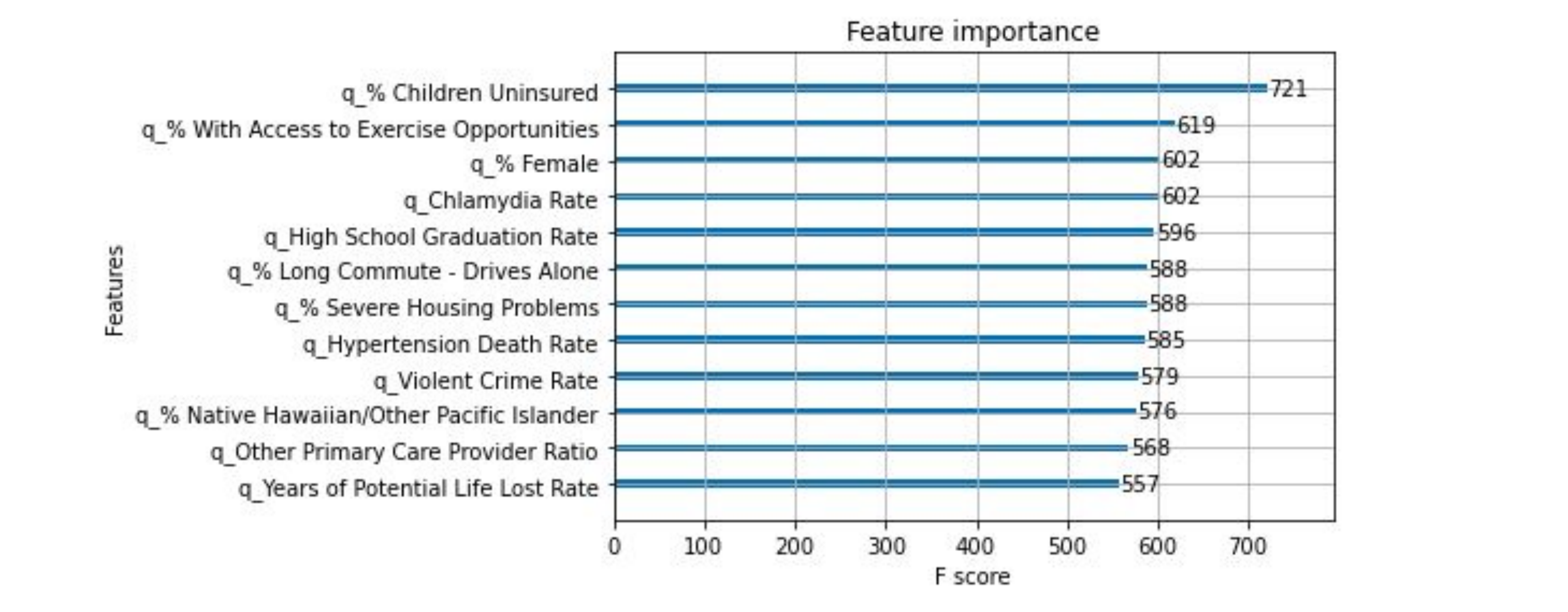


```
              precision    recall  f1-score   support

         0.0       0.79      0.73      0.76       273
         1.0       0.74      0.75      0.75       325
         2.0       0.69      0.75      0.72       249
         3.0       0.73      0.71      0.72       187

    accuracy                           0.74      1034
   macro avg       0.74      0.73      0.74      1034
weighted avg       0.74      0.74      0.74      1034
```

** For Detailed result for all clusters visit code.

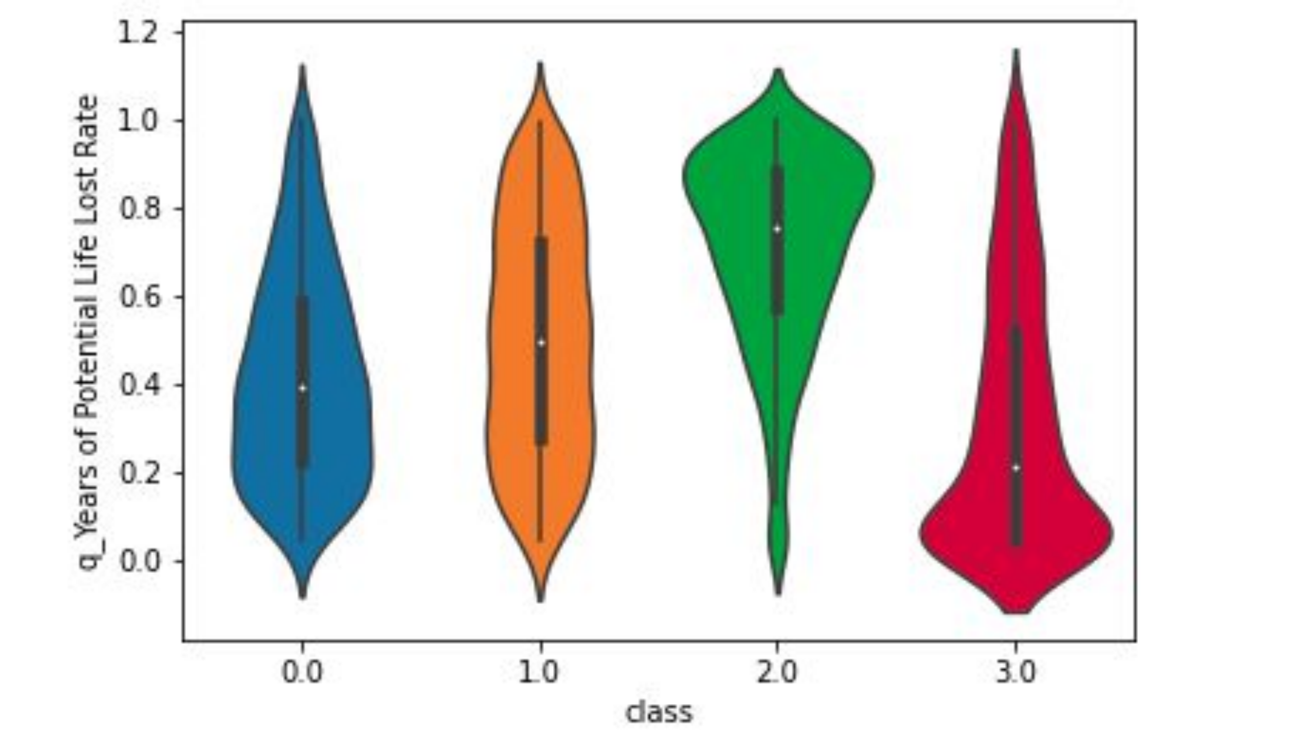
Important features

For 4 cluster case



How data is distributed across important features

For 4 clusters case



Number of clusters

Important Features

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters
q_% Fair or Poor Health	q_Years of Potential Life Lost Rate	q_% Fair or Poor Health	q_% Fair or Poor Health	q_% Fair or Poor Health
q_% With Access to Exercise Opportunities	q_% Fair or Poor Health	q_Chlamydia Rate	q_% With Annual Mammogram	q_% With Annual Mammogram
q_Chlamydia Rate	q_% With Access to Exercise Opportunities	q_% With Annual Mammogram	q_Average Daily PM2.5	q_Average Daily PM2.5
q_Average Daily PM2.5	q_Chlamydia Rate	q_Average Daily PM2.5	q_Suicide Rate (Age-Adjusted)	q_Population
q_Life Expectancy	q_Average Daily PM2.5	q_% Severe Housing Problems	q_Population	q_% Black
q_Child Mortality Rate	q_Child Mortality Rate	q_Population	q_% American Indian & Alaska Native	q_% American Indian & Alaska Native
q_Population	q_Black/White Segregation Index	q_% Rural	q_% Rural	q_% Rural
q_% Non-Hispanic White	q_countycode	q_countycode	q_countycode	q_countycode
q_countycode	q_internet_hhs	q_internet_hhs	q_internet_hhs	q_internet_all
q_internet_hhs	q_% workers commuting by public transit	q_% Without Health Insurance	q_% Without Health Insurance	q_internet_hhs

Results

- The data contains enough information to make between 2-6 different metrics as this same range of clusters were analysed to be good.
- Since k means is susceptible to outliers, quantile normalization is used.
- The prediction for any new data into one of the given groups is as high as 81% for 2 clusters. For others it's low due to lack of data points for individual clusters.
- The important features that are most helpful for designing each clusters are stored and are mostly the same ones.
- The clusters obtained are well balanced for cases of 2 to 6 clusters. It means that the data is almost equally divided for all groups. It ensures absence of selection bias

Further steps and analysis to be done:

1. Comparing the top important features with the already used ones for individual metrics. (It can help it addition or removal of some features to improve prediction)
2. Analyzing the data distribution on some important metrics like racial oppression vs severity. (It tells how are people belonging to different races being affected by Covid---> discrimination)
3. Creating synthetic data using various oversampling techniques such as SMOTE , Adasyn,etc. After confirming some choices for number of possible clusters (say 3,4,5)