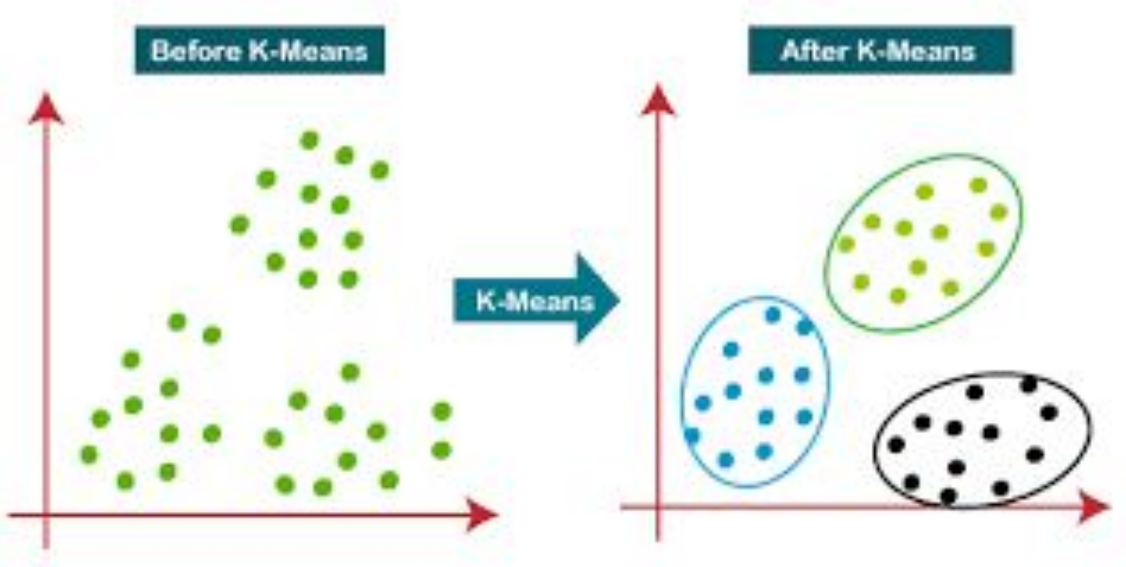
Community Clustering

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

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 35 columns.

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

** See code for more details

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

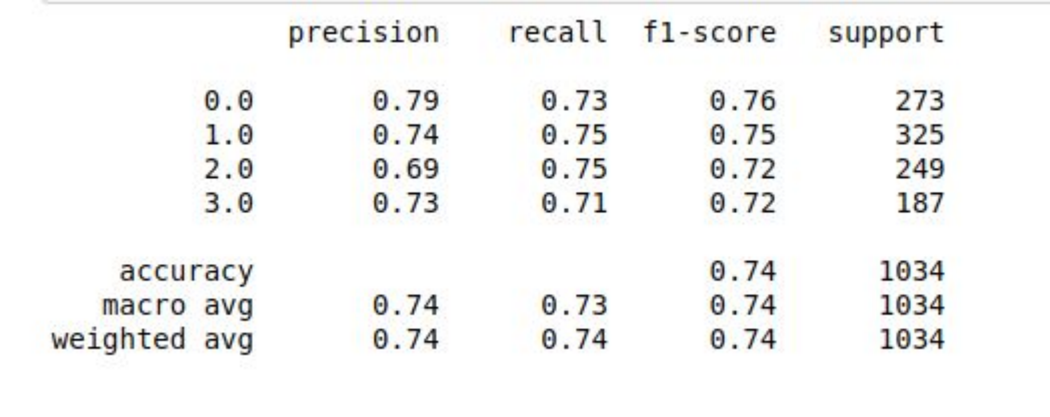
The predicted labels are used as the class labels for each clusters and this class is added to original data.

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

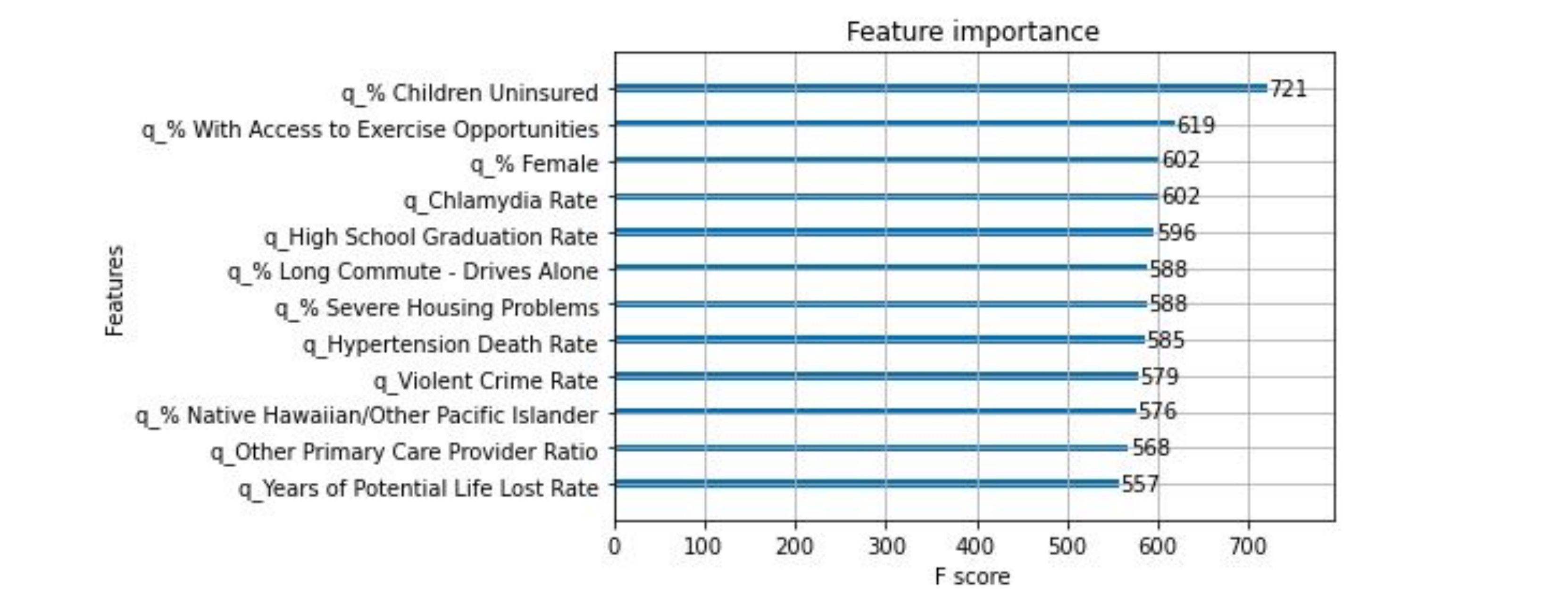


```
              precision    recall  f1-score   support

         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.

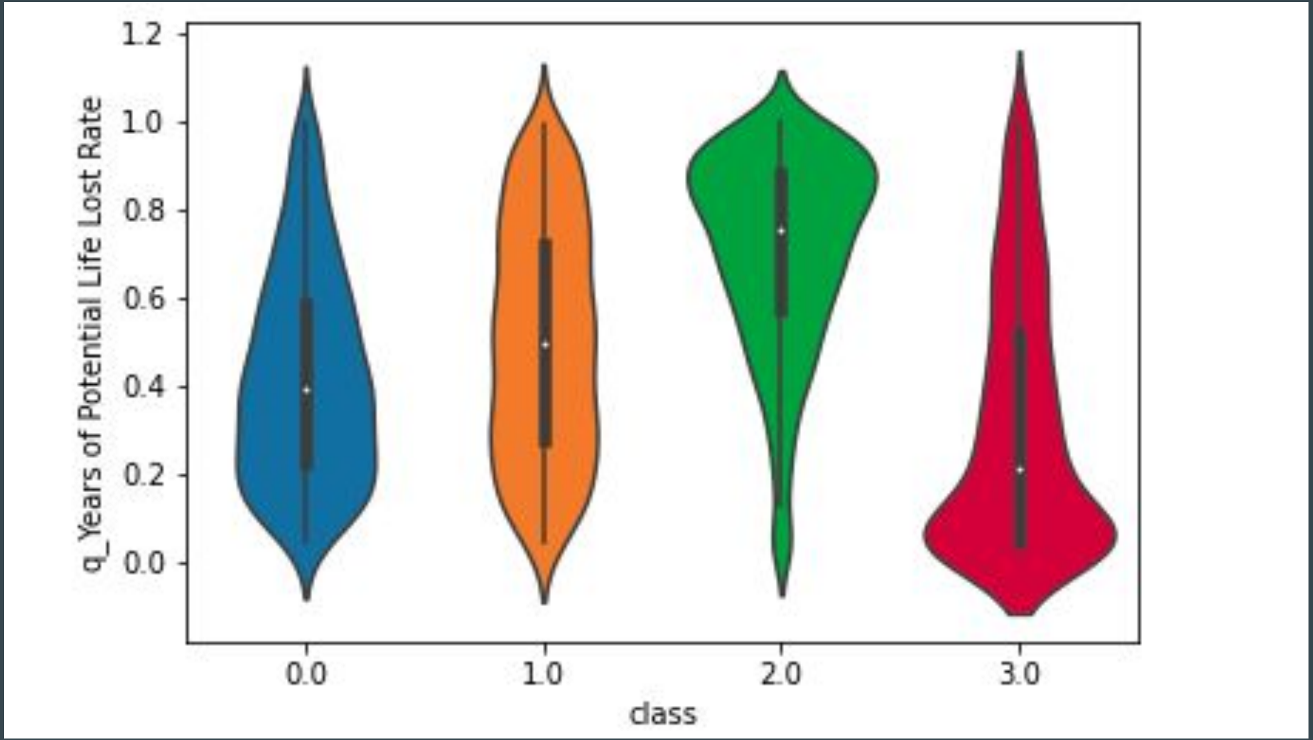
Important features

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_% Fair or Poor Health	q_% Fair or Poor Health	q_% Fair or Poor Health	q_% Fair or Poor Health
q_% With Access to Exercise Opportunities	q_Chlamydia Rate	q_Chlamydia Rate	q_% With Annual Mammogram	q_% With Annual Mammogram
q_Chlamydia Rate	q_% With Annual Mammogram	q_% With Annual Mammogram	q_Average Daily PM2.5	q_Average Daily PM2.5
q_Average Daily PM2.5	q_Average Daily PM2.5	q_Average Daily PM2.5	q_Population	q_Population
q_Life Expectancy	q_% Severe Housing Problems	q_% Severe Housing Problems	q_% Black	q_% Black
q_Child Mortality Rate	q_Population	q_Population	q_% American Indian & Alaska Native	q_% American Indian & Alaska Native
q_Population	q_% Rural	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_all	q_internet_all
q_internet_hhs	q_% Without Health Insurance	q_% Without Health Insurance	q_internet_hhs	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