Prediction Analysis Procedure For Priority Groups

We prepared two datasets, one for the states - **state_data.csv** and the other having districts of each state - **district_data.csv**

Dataset Features

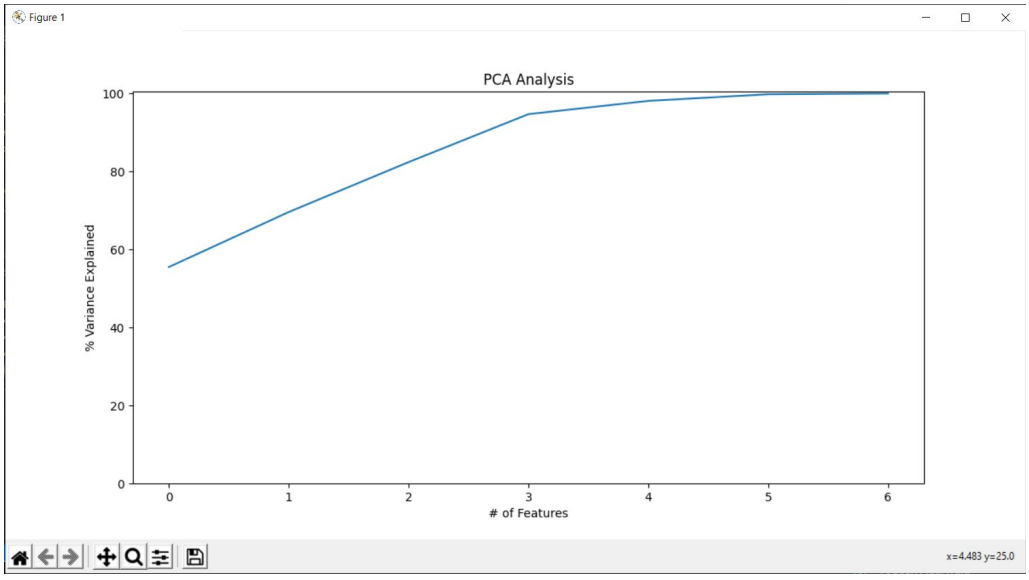
We went through various census report estimates, WHO reports, and other state websites to fetch and combine all these data features and finally made our datasets. We then add and update the data for the **number of covid cases** in each state daily and carry out our algorithms.

And hence we ended up with the following features for the Batch Predictions:
State Data:

1. **Active Cases**
2. **Population_2020(estimated: since the census is done every 10 years(2021 next))**
3. **Heath Workers Present in the State**
4. **Senior Citizens(60+)**
5. **Children(0-14yrs)**
6. **Accessibility of each state**
7. **Allotted Hospital Beds for Covid Patients in each state**

Feature Reduction

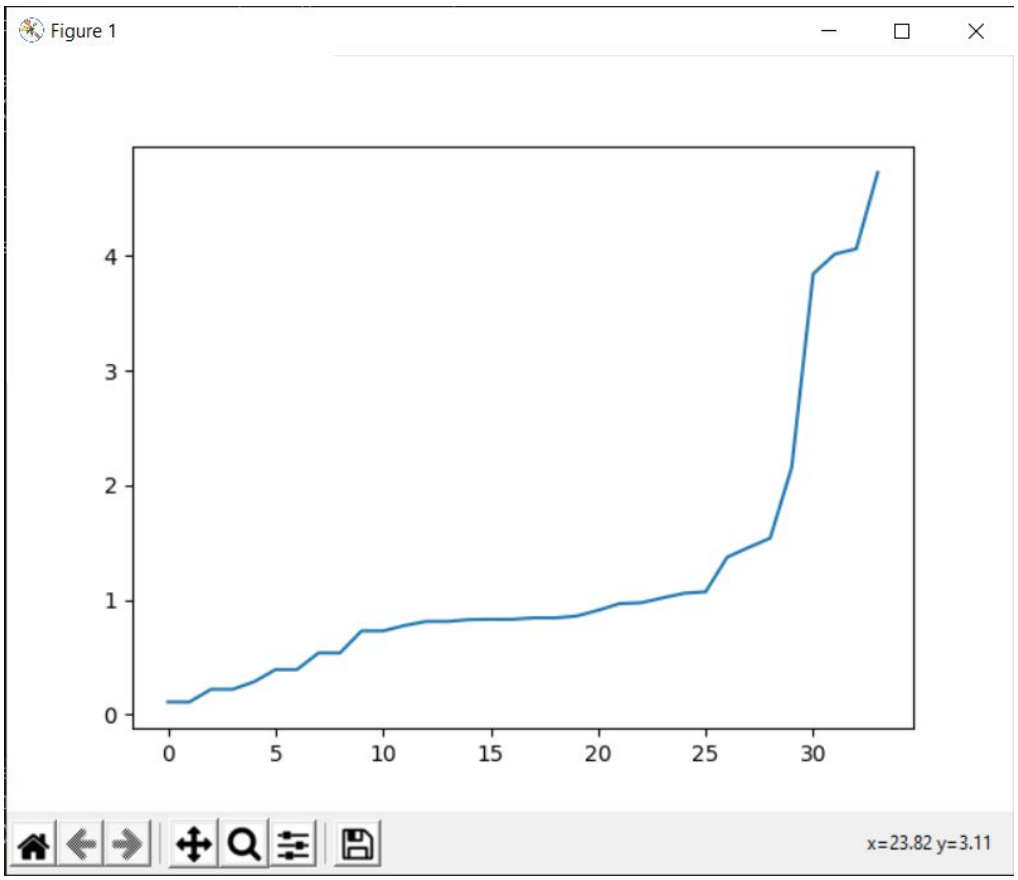
We then preprocessed our data features and then carried out **Principal Component Analysis** on the feature set to determine the optimum no. of features for the clustering algorithm.



We found out that for **3 # of features is sufficient to explain variance to 90%** of our previous seven featured datasets.

Clustering Algorithms and Evaluation Metrics

After that, we ran different Clustering Models like KMeans Clustering, DBSCAN Clustering, and Hierarchical Clustering models and evaluated the elbow curve and Silhouette Score for DBSCAN KMeans Clustering Algorithms.



We decided upon the Kmeans Clustering algorithm based on the metrics since it gave us better Silhouette scores **(given below with max score given at the bottom of the image)** than the DBSCAN algorithm.

```
[0.03136797698350208, 0.04023821894648165, 0.04690767522914846, 0.08486178214996729, 0.11292302997508355, 0.114405625597
9578, 0.1562603307148037, 0.19457789205206227, 0.2165507215402576, 0.23639574061598914, 0.2367823773229422, 0.2564392766
3661115, 0.27411876446581085, 0.2866342356383575, 0.2919654139718154, 0.3510742221530088, 0.3539586151722027, 0.36788956
69025565, 0.37391411962180865, 0.37429426959522377, 0.37506598322937545, 0.39276820772641147, 0.40838594673836565, 0.408
6472366212832, 0.42341992800824046, 0.4444625186030132, 0.4491621792955222, 0.46467271526397547, 0.5126071574172856, 0.5
157408153439585, 0.5217288148291129, 0.5295399699982226]
0.5295399699982226
```

Hence we run the KMeans clustering algorithm through all the number of clusters possible and note the silhouette scores for each one of them and then cluster the features into the optimum Cluster Number based on the silhouette scores for each number(of the clusters)

Finally, we ran through the clustered dataset, calculated the mean observation for each Cluster, and ran those mean observations representing each cluster into our ranking algorithm.

Ranking Algorithm

We chose the best ranking algorithm (Rank Exponential Weighting System) for ranking our features. Based on various metrics and common practical knowledge and facts, we came to the following ranking criteria:

Rank No. 1 (Most imp Criteria): No. of Active Cases (The more => More weightage)
Rank No. 2: Ratio of Hospital Beds/No. of active cases (The less => More weightage)
Rank No. 3: No. of Health Workers (The less => More weightage, since fewer doctors to handle patients)
Rank No. 4: No. of Senior Citizens
Rank No. 5: No. of Children
Rank No. 6 (Least Imp Criteria): Accessibility of the State

The rank is calculated using the following formula:

$$S_i = \sum_{j=1}^{n} c_{ij} w_j$$

$\mathbf{S}_i$ is the score for each observation; $\mathbf{w}_j$ is the weight of each criterion (feature) calculated by the RES method. $\mathbf{C}_{ij}$ is the normalized features of an observation-**i** in the dataset, respectively.

Where the formula calculates the weight:

$$w_j(RE) = \frac{(n - r_j + 1)^p}{\sum_{k=1}^{n} (n - r_k + 1)^p},$$

Where **n** is the number of criteria and $\mathbf{r}_j$ for each criterion, and **p** is the variable used to vary the weights for each ranking criterion.

The behavior of the generated numerical weights depending on the parameter *p* of the rank component method.

Rank	0	0.5	1	2	3		10
1	0.20	0.27	0.33	0.45	0.56		0.90
2	0.20	0.24	0.27	0.29	0.29	.	0.10
3	0.20	0.21	0.20	0.16	0.12	.	0.00
4	0.20	0.17	0.13	0.08	0.03	.	0.00
5	0.20	0.11	0.07	0.02	0.00	.	0.00
Sum	**1**	**1**	**1**	**1**	**1**		**1**

We calculated the rank weights by varying **p - (1,1.2,1.3,1.5, 2,2.5, and 3)** and ranked our features using them.

Finally, we decided upon using **p = 1.2** since it gave the most optimum rank predictions in line with the general practical expectations.

To check our Ranking system, we used the well-known subjective method for determining weight is the **Analytic Hierarchy Process** method (**AHP**) proposed by Saaty [1980]. In this method, The preferences of the decision criteria are compared in a pairwise manner with regard to the criterion preceding them in the hierarchy.

The References to the **Ranking Methods**, **weightage systems**, and **AHP** has been given below**:** Attached here.

Project: Priority Covid-19 Vaccine Delivery
Team TechHD
IIT Madras

Note:
A similar process has been carried for the districts (**district_data.csv)** so that each state can be further divided into different batches based on features: The number of Active Covid cases and each district's population.