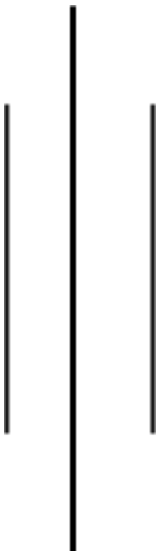
A
Report On

PROJECT - III

For
Machine Learning Course
Fuse.ai Microdegree in AI



Submitted By:
Anish Bhusal
bhusal.anish12@gmail.com

Date of Submission: 20th April,2020

1. DATA EXPLORATION

This project analyzes datasets mainly two datasets:
a) Literacy Rate b) Life Expectancy and Per Capita Income for Human Development marks.

Following information has been explored from the datasets:

I. Ramechhap has highest life expectancy of 72.9 years
II. Dolpa has lowest life expectancy among other districts i.e. 61.2 years
III. Manang has highest per capita income of 3166 USD whereas Bajhang has lowest per capita income of 487 USD.

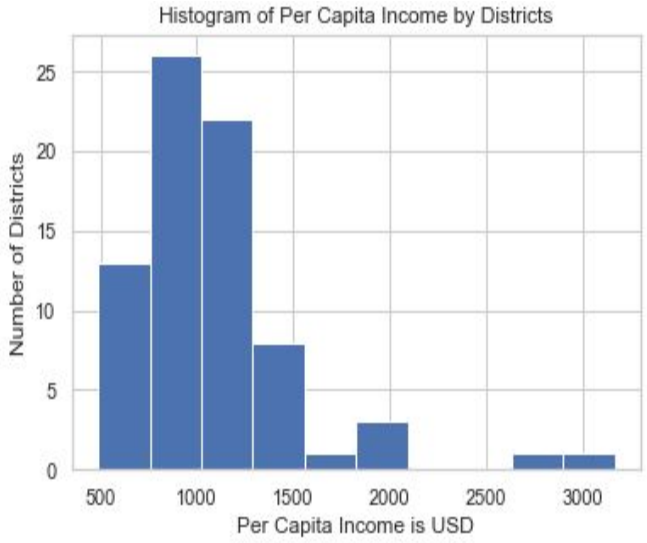


Fig.: Distribution of Per Capita Income by Districts

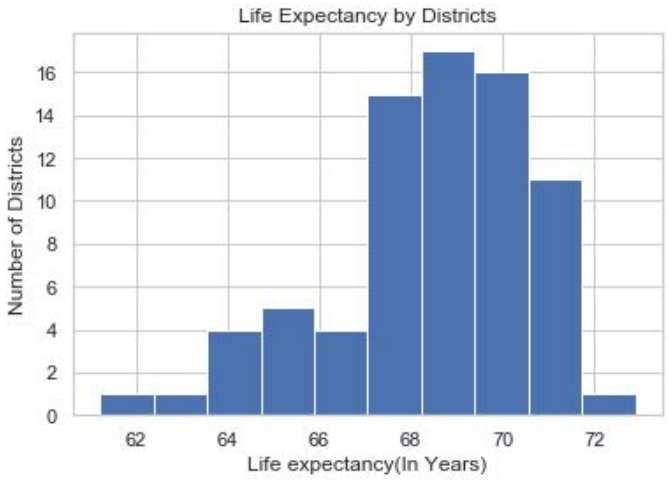


Fig.: Life Expectancy Distribution by Districts

IV. Kathmandu has highest literacy rate of 86.3% and Rautahat has least literacy rate of only 41.7 %

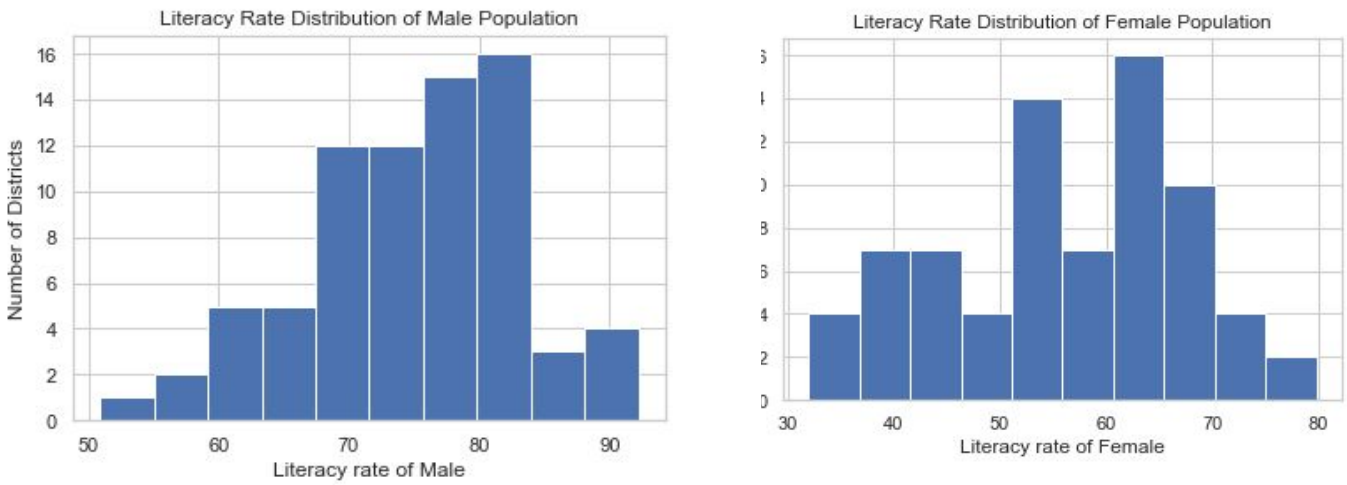


Fig.: Literacy Rate Distribution of Male and Female Population by Districts

V. Literacy Rate was computed for Province level.

	Provinces	Total Literacy	Male Literacy	Female Literacy
0	Province 1	77.9	84.2	61.5
1	Province 2	54.5	67.0	36.6
2	Province 3	86.3	92.2	60.6
3	Gandaki	82.4	90.1	69.6
4	Province 5	76.2	84.9	57.9
5	Karnali	73.1	82.0	48.7
6	Province 7	70.7	81.3	52.9

From the table, it is clear that Province 3 has the highest literacy rate and also the highest male literacy rate. Gandaki province has the highest female literacy rate. Similarly, Province 2 has lowest literacy rate for both male and female.

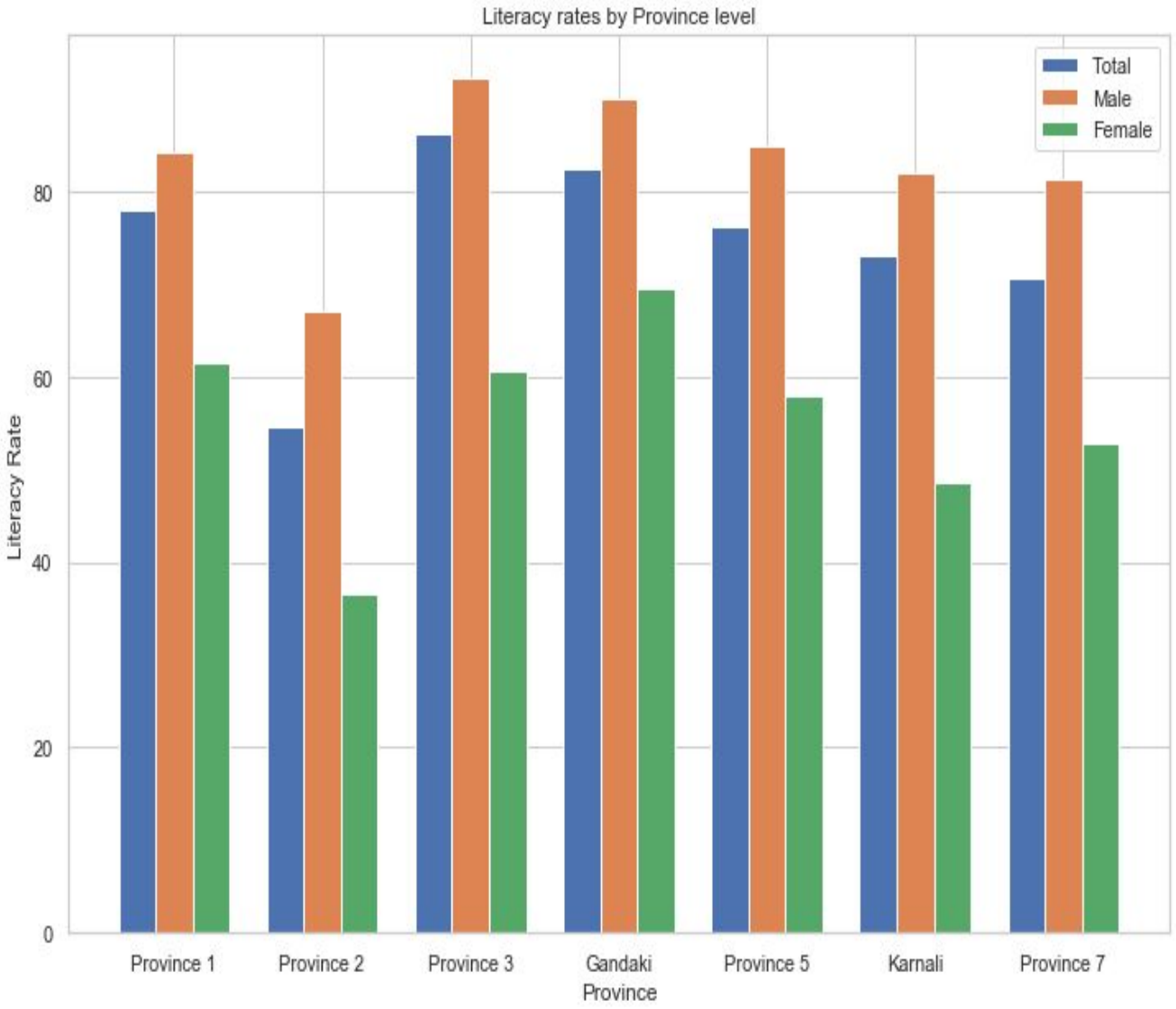


Fig.: Literacy Rates by Province Level

2. DATA PRE-PROCESSING AND FEATURE SELECTION

Two datasets : literacy rate and life expectancy were combined. The “Total Literacy Rate” column was only used from literacy rate and rest were dropped while combining. The final dataset looked like this:

	District	Life expectancy(In Years)	Per Capita Income(In USD)	Total
0	Ramechhap	72.90	951	86.3
1	Gorkha	71.70	1039	82.5
2	Saptari	71.34	801	82.4
3	Siraha	71.29	689	81.7
4	Rautahat	70.99	757	77.9

Then the data was normalized using *sklearn-StandardScaler.*

3. KMeans - APPROACH AND FINDINGS

Before building the model, the elbow method was used to find the optimal number of clusters. A graph of inertia vs number of cluster was plotted:

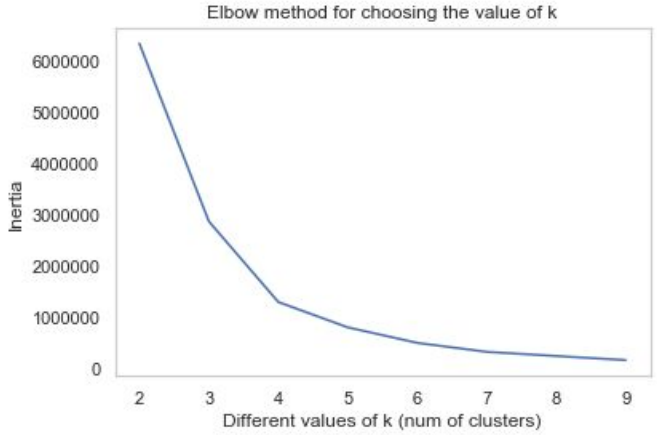


Fig.: Elbow method graph

From the above graph, the number of clusters=3 were chosen.

The final model was built using following params:

n_clusters=3	*random_state=1*

To visualize the final clusters, PCA was used with three principal components.

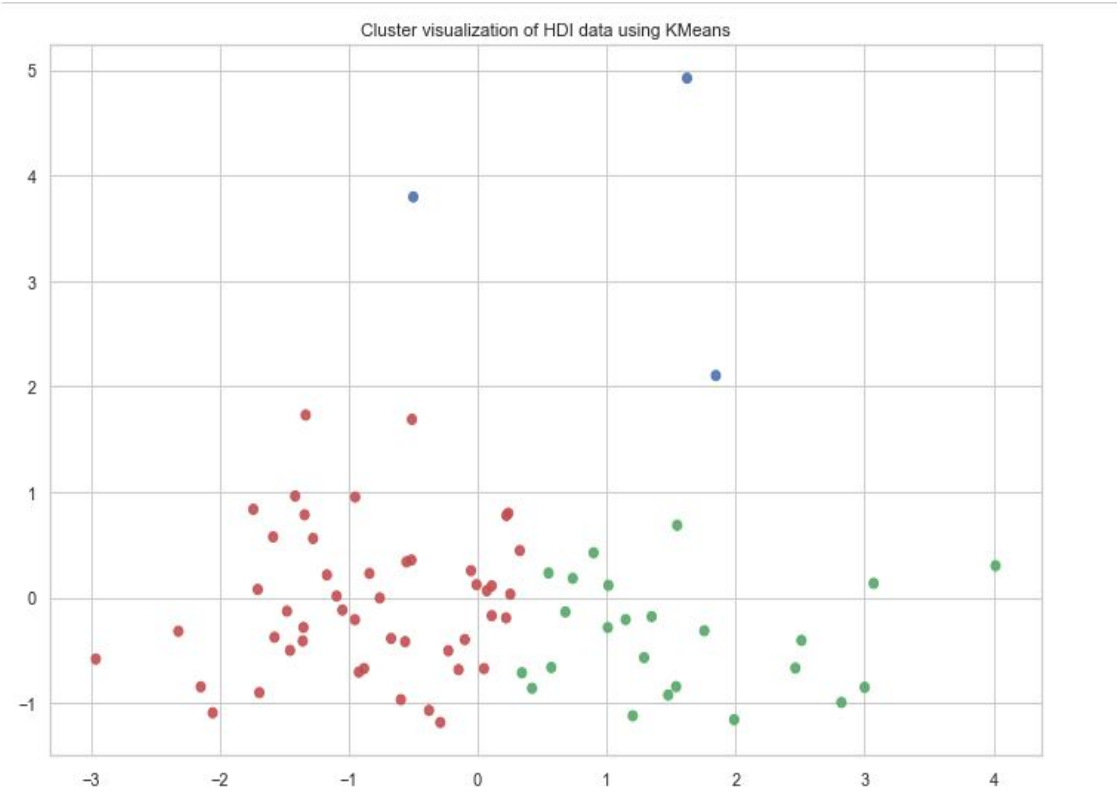


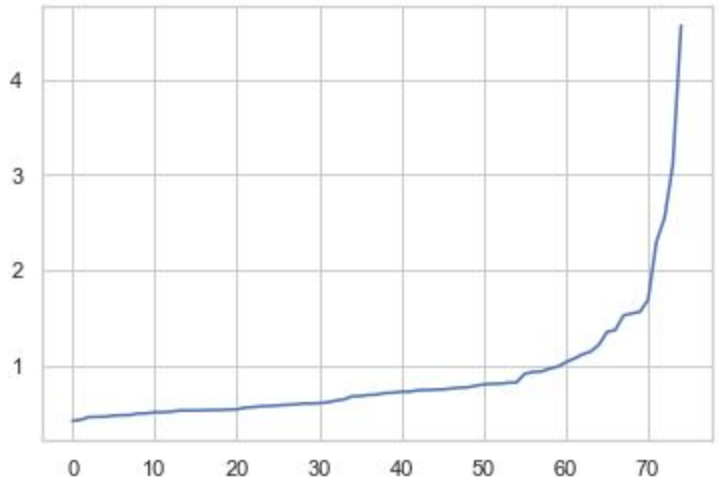
Fig.: Cluster visualization for KMeans

The evaluation of KMeans clusters provided following scores:

Silhouette Score	0.4405

4. DBSCAN

To choose the value of epsilon, Nearest Neighbours algorithm was used and following graph was plotted:



From the graph, the value of epsilon=2 was chosen.

The model was built using following params:

eps=2	*min_samples=3*	*metric='euclidean'*

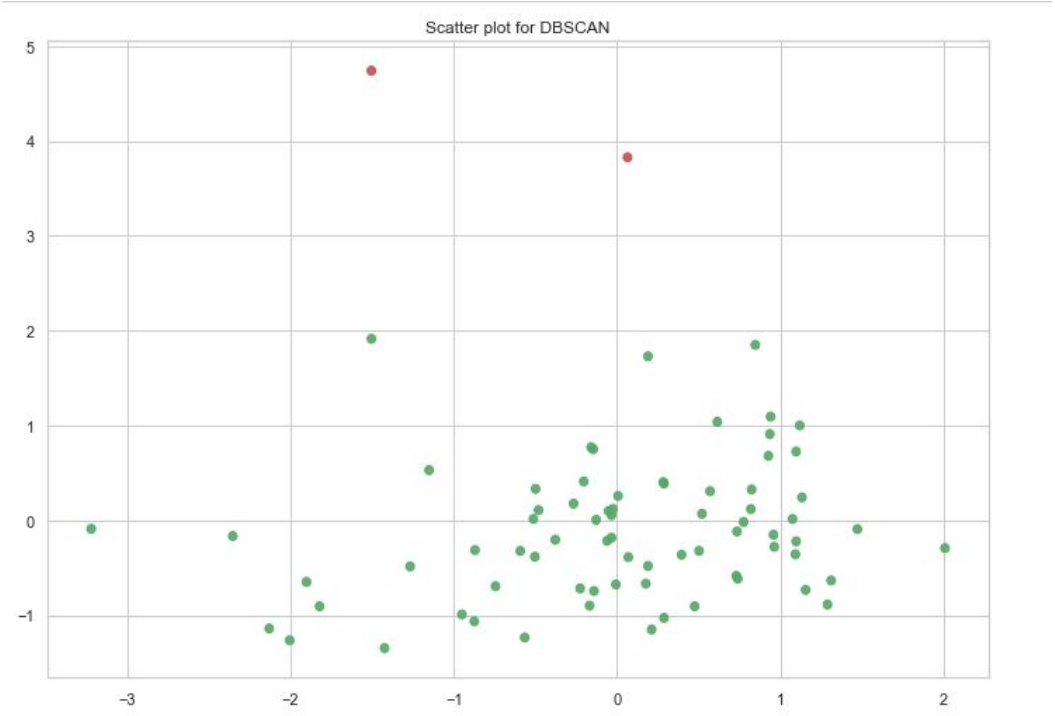


Fig.: Scatter Plot Visualization for DBSCAN
Only single cluster was formed by DBSCAN

with Silhouette Score of 0.58

5. GMM

Following params were used to compute Gaussian Mixture Model:

n_components=3

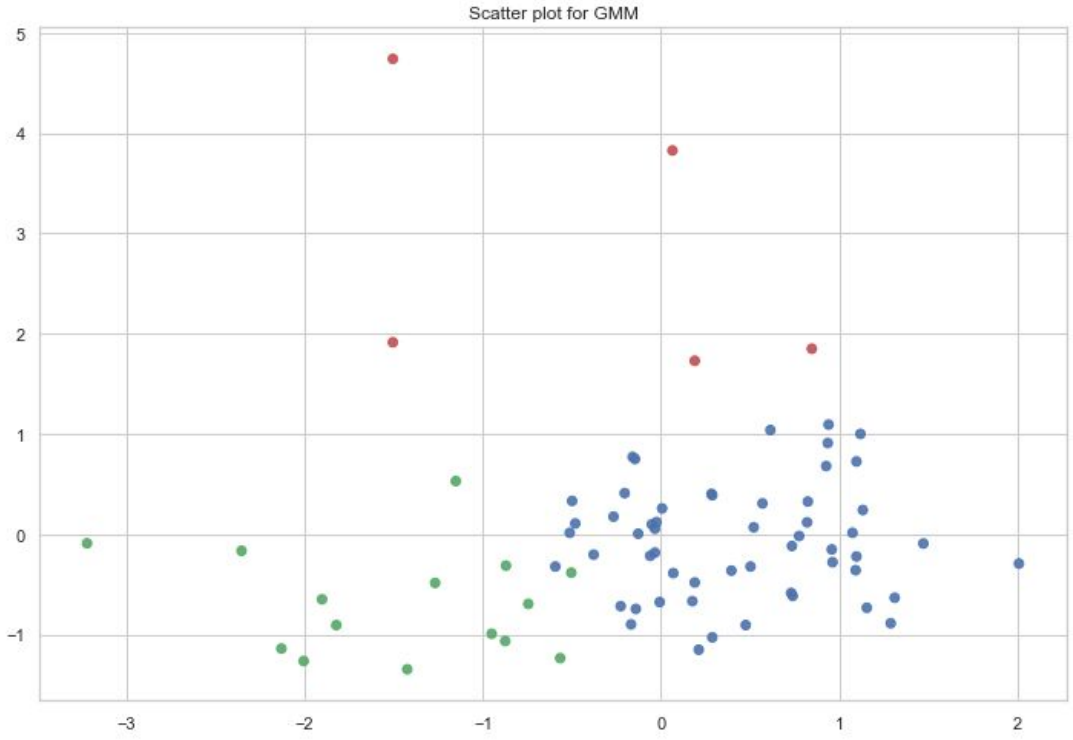


Fig.: Scatter Plot Visualization for GMM

Silhouette score for final three clusters formed:

Score	0.4366

6. CONCLUSION

Three models were used for clustering HDI dataset on parameters: Life Expectancy, Per Capita Income and Literacy Rate. On evaluating the performance of all models, it can be concluded that KMeans with number of components=3 performed well than other model achieving Silhouette score of 0.4405