MA324 Project Report

Spectral Clustering

Bachelor of Technology
in
Mathematics and Computing

Submitted by

Roll No	Names of Students
210123010	Ayush Kumar
210123011	Ayush Kumar Jaiswal
210123012	Ayush Verma
210123014	Chandrashekar A. Giridharan

Under the guidance of
Dr. Arabin Kumar Dey

Department of Mathematics
Indian Institute of Technology Guwahati
Guwahati, Assam – 781039

Winter Semester 2024

Objective:

This project aims to optimize data clustering effectiveness by employing a systematic approach. By meticulously cleaning the dataset, implementing classic clustering methodologies such as K-Means and Gaussian Mixture Models (GMM), and leveraging dimensionality reduction with Principal Component Analysis (PCA) followed by spectral clustering, the project seeks to achieve a refined clustering structure. Furthermore, the project aims to rigorously compare the performance of these varied clustering techniques using silhouette score, providing valuable insights into their efficacy and applicability in real-world scenarios.

K – Means Clustering

K-means iteratively partitions data into K clusters by updating centroids based on assigned points' means, minimizing within-cluster variance. Given data points x_i, centroids μ_j, and cluster assignments r_{ij}, the objective function minimizes total within-cluster variance:

$$J=\sum_{j=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|x_i-\mu_j\right\|$$

Where $r_{ij} = 1$, if x_i is assigned to cluster j, and 0 otherwise.

Elbow Method: aids in determining the optimal number of clusters in a dataset by plotting the within-cluster sum of squares against the number of clusters and identifying the "elbow" point where the rate of decrease sharply changes.

Gaussian Mixture Model

Gaussian Mixture Model (GMM) represents data as a mixture of K Gaussian distributions. Given data points x_i, mixture component means μ_j, covariances Σ_j, and mixing coefficients π_j, the probability density function of GMM is:

$$p(x_i) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)$$

Where $N(x_i | \mu_j, \Sigma_j)$ represents the Gaussian probability density function. The EM algorithm iterates through two steps:

1. **Expectation Step:** Compute the responsibility of each component for each data point using Bayes' theorem

$$\gamma(z_{ij}) = \frac{\pi_j \mathcal{N}(x_i|\mu_j,\Sigma_j)}{\sum_{k-1}^{K} \pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k)}$$

2. **Maximization Step:** Update parameters:
   - Mean:
     $$\mu_j = \frac{\sum_{i-1}^{n} \gamma(z_{ij}) x_i}{\sum_{i-1}^{n} \gamma(z_{ij})}$$
   - Covariance:
     $$\Sigma_j = \frac{\sum_{i-1}^{n} \gamma(z_{ij})(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i-1}^{n} \gamma(z_{ij})}$$
   - Mixing coefficients:
     $$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ij})$$

Bayesian Information Criterion (BIC) Method: assesses the goodness of fit of a statistical model by balancing model complexity and goodness of fit, aiming to select the model that minimizes the BIC score.

Principle Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the most important features in a dataset. It finds a new set of orthogonal axes, called principal components, such that the variance of the data is maximized along these axes. Here's a concise mathematical overview:

1. **Data Centering**: $\text{Centered Data} = \text{Data} - \text{Mean}(\text{Data})$
2. **Covariance Matrix**: $\Sigma = \frac{1}{n} \times \text{Centered Data}^T \times \text{Centered Data}$
3. **Eigenvalues and Eigenvectors**: $\Sigma v = \lambda v$
4. **Principal Components**: Eigenvectors corresponding to highest eigenvalues.
5. **Dimensionality Reduction**: Project data onto selected principal components.

Spectral Clustering

Spectral clustering partitions data by eigen-decomposing a similarity matrix. Nearest neighbour and Radial Basis Function (RBF) kernel are common approaches.

1. **Nearest Neighbor**:
   - Construct a graph W with nodes for data points.
   - Define W_{ij} as 1 if i is among the nearest neighbors of j, else 0.
   - Compute the Laplacian $L = D - W$, where D is the degree matrix.
   - Solve the eigenproblem $L\mathbf{v} = \lambda\mathbf{v}$ for eigenvectors $\mathbf{v}$.
2. **RBF Kernel**:
   - Compute similarity matrix W using RBF kernel: $W_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}$.
   - Compute Laplacian $L = D - W$.
   - Proceed with eigen-decomposition.

Spectral clustering captures intricate data structures effectively, robust to nonlinearities, making it versatile for various data types.

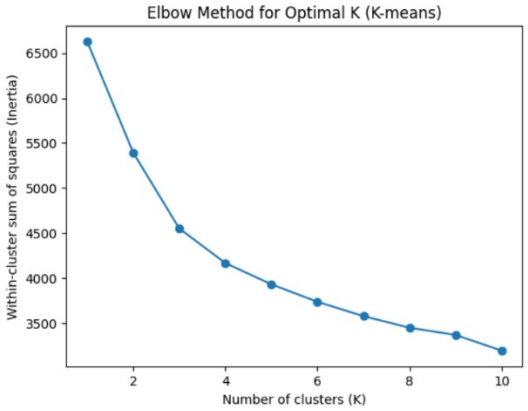
Eigengap Heuristic method: aids in determining the ideal number of clusters in spectral clustering by analysing the difference between consecutive eigenvalues of the Laplacian matrix. It identifies the point where the gap between eigenvalues is maximized, suggesting the optimal clustering structure.

Silhouette score: measures the quality of clustering by assessing how well-separated clusters are and how cohesive the data points within each cluster are, with values ranging from -1 to 1. Higher scores indicate better-defined clusters, with values close to 1 suggesting dense, well-separated clusters.

OBSERVATIONS:

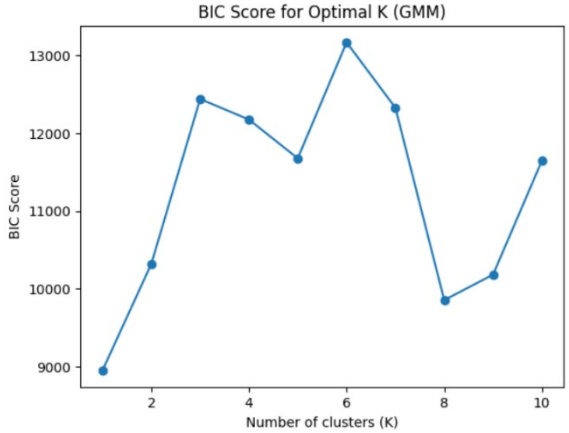
- For Breast Cancer Dataset

The elbow method using WCSS(within cluster sum of squares) was employed to find optimal k for K-means clustering and the following graph was observed.



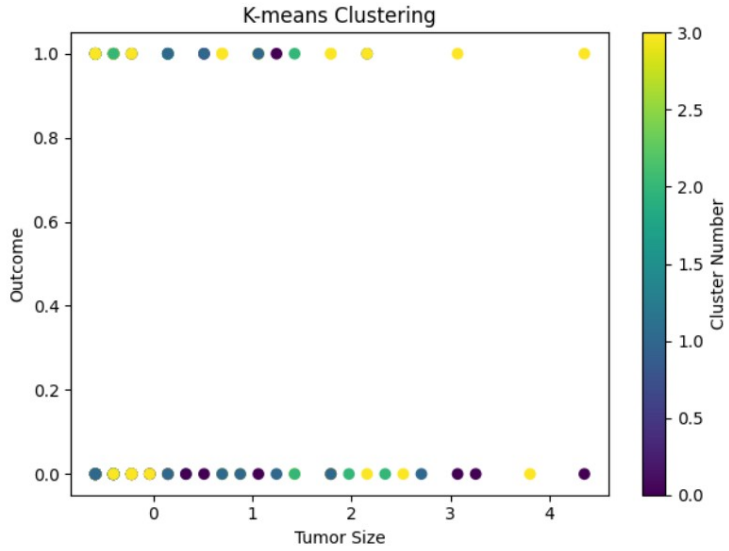
So the value of optimal k is 4.

The BIC scores for various values of K were plotted to find an optimal k for the GMM algorithm.



So optimal value of k is 2.

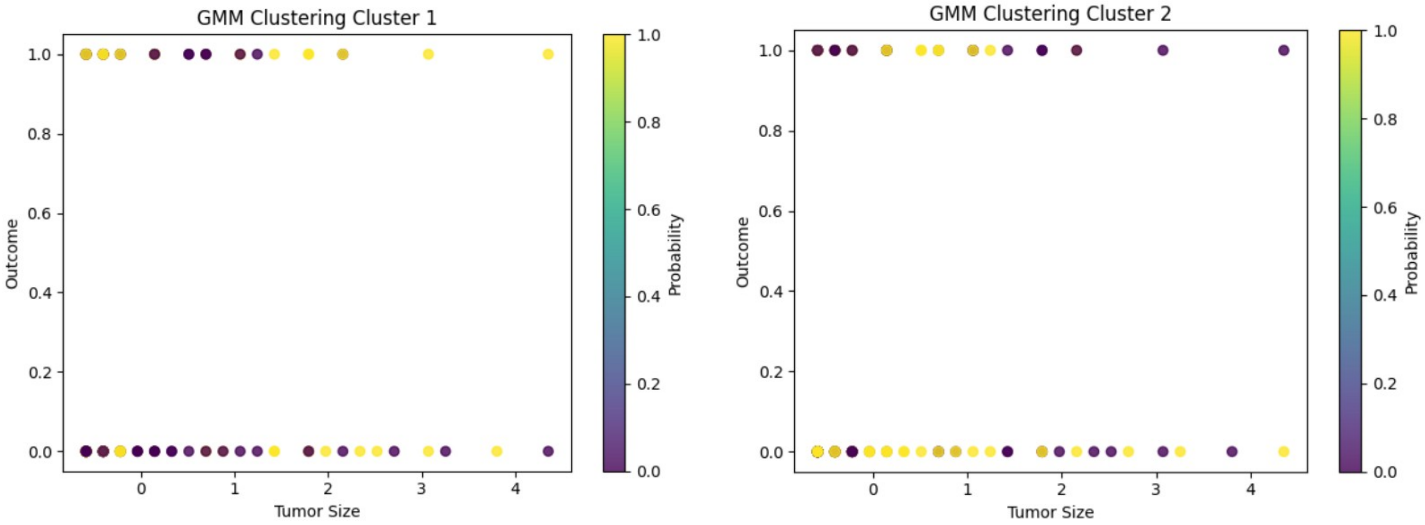
The clusters obtained after K-Means Clustering was applied:



Clusters obtained when plotting Tumour size vs outcome.

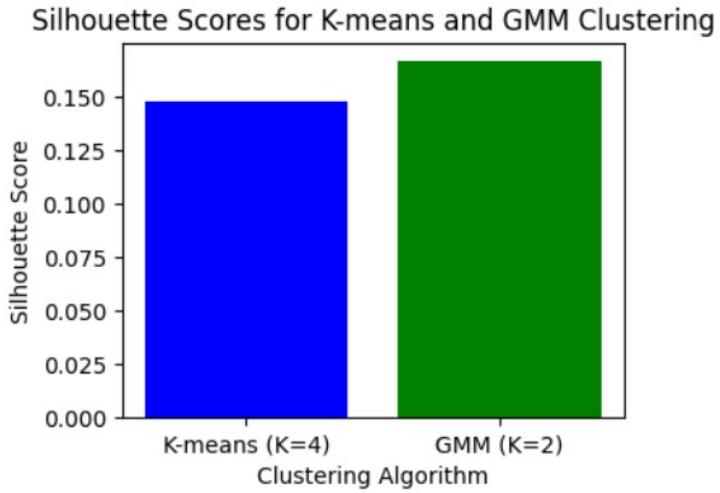
Clusters obtained for GMM

- The following clusters were obtained when plotting tumour size vs outcome



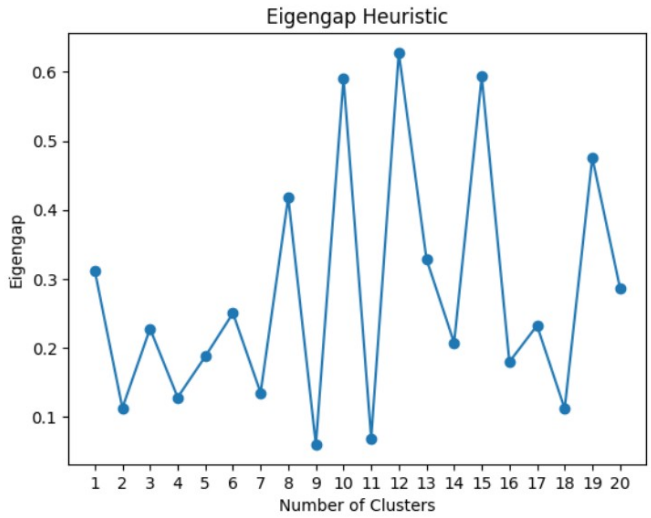
The quality of clusters was compared using Silhouette Scores and the results were:

	K-Means(K = 4)	GMM(K =2)
Silhouette Score	0.14733952349983268	0.1665229656854858



Then dimensionality reduction using PCA was done to the dataset so that it now contains only two columns, and a normalized dataset was generated for testing purposes.

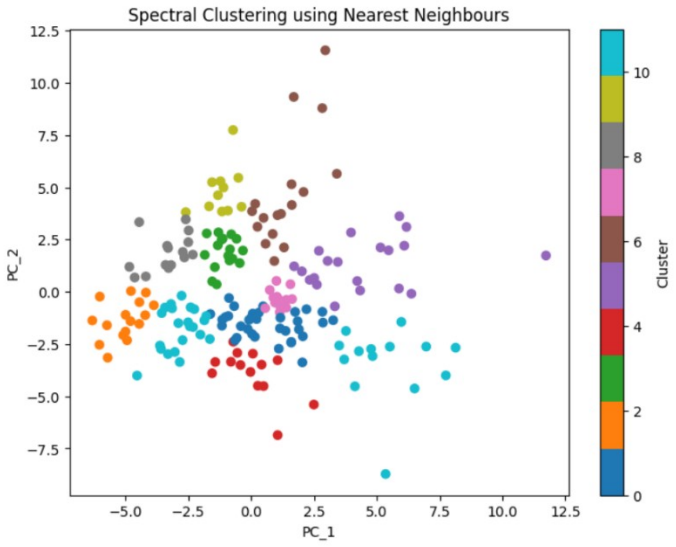
Spectral Clustering on the reduced dataset(non-normalised)



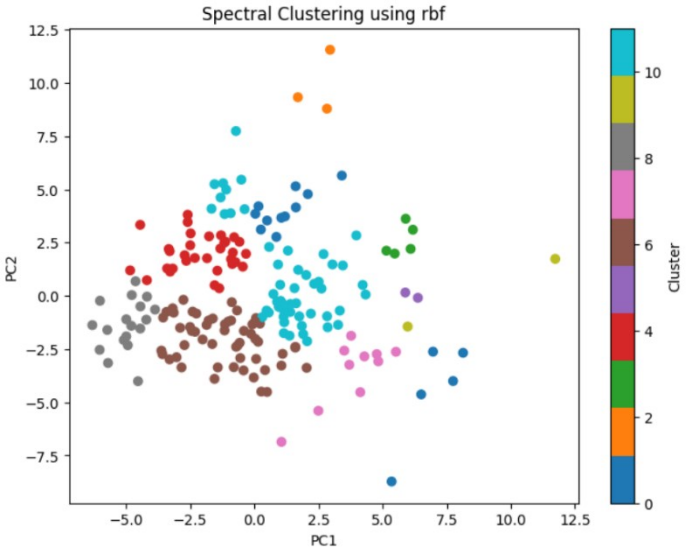
As evident, optimal value of k is

Argmax(eigengaps) = 12

Using affinity as *'nearest-neighbours'*

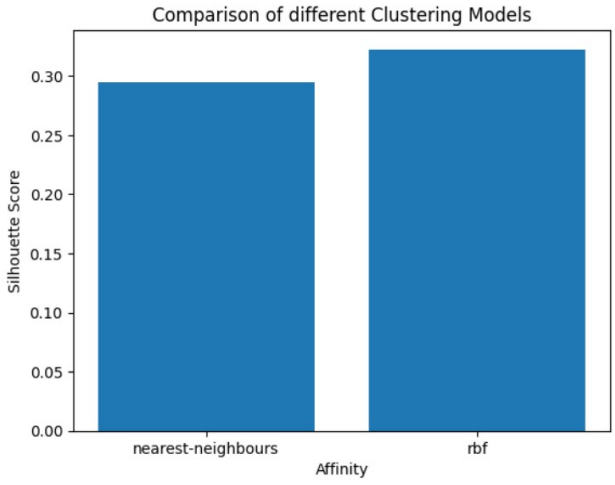


Using affinity as *'rbf'*

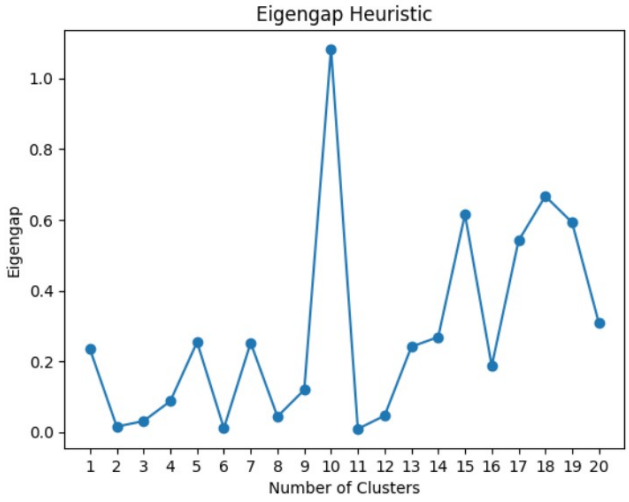


When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.2944166285420264	0.3225949196906876



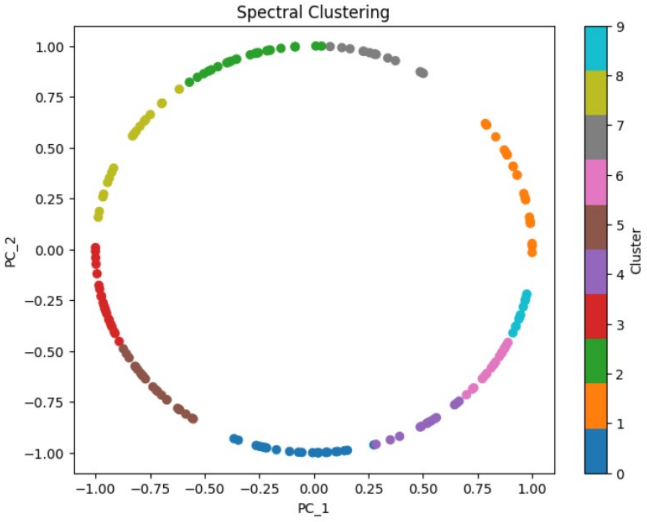
Spectral Clustering on the normalized dataset



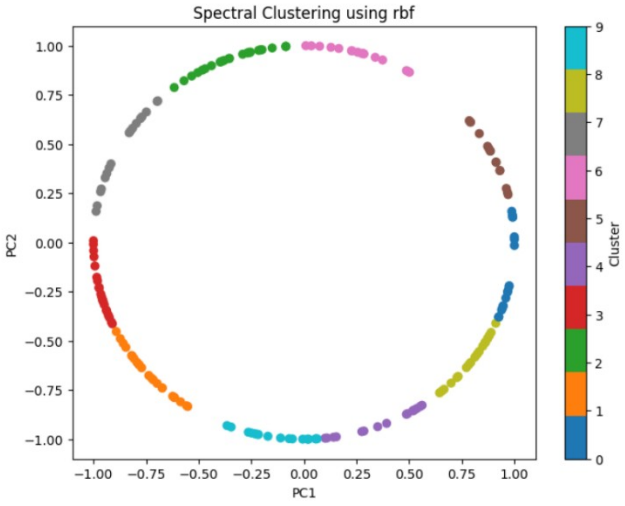
As evident, optimal value of k is

Argmax(eigengaps) = 12

Using affinity as *'nearest-neighbours'*

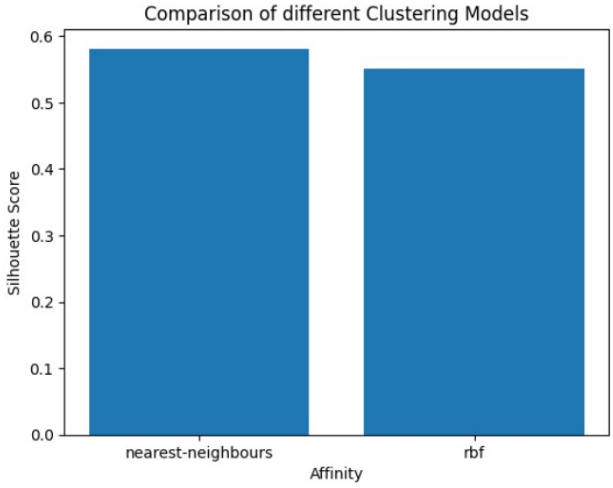


Using affinity as '*rbf*'



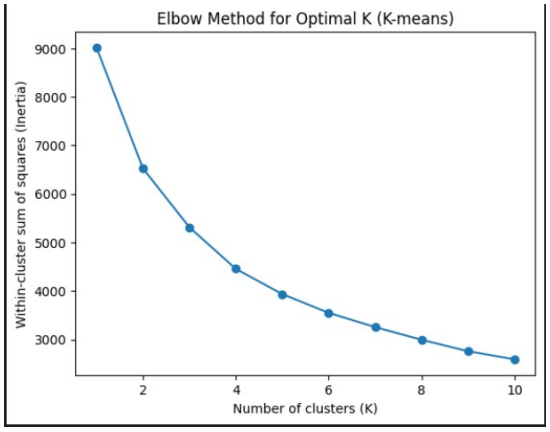
When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.5810503049046665	0.5505503126362913



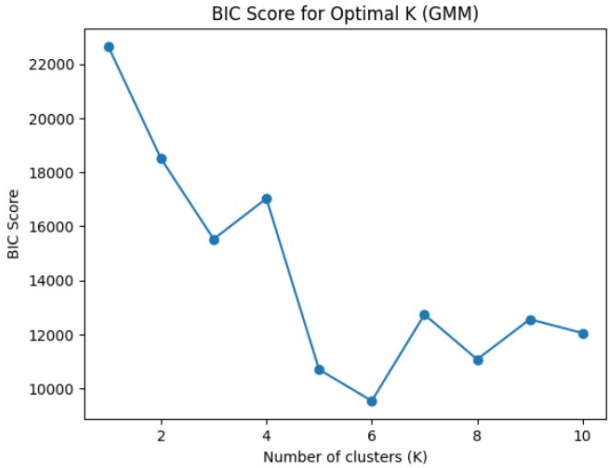
- For Airfoil Self Noise Dataset

The elbow method using WCSS(within cluster sum of squares) was employed to find optimal k for K-means clustering and the following graph was observed.



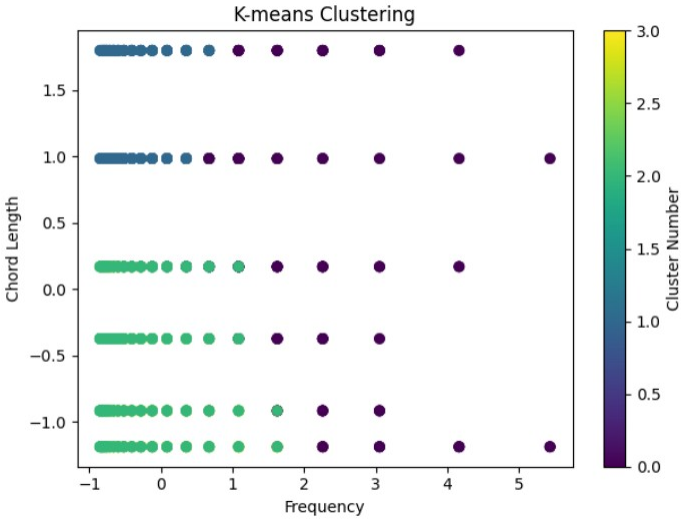
So the value of optimal k is 4.

The BIC scores for various values of K were plotted to find an optimal k for the GMM algorithm.



So optimal value of k is 2.

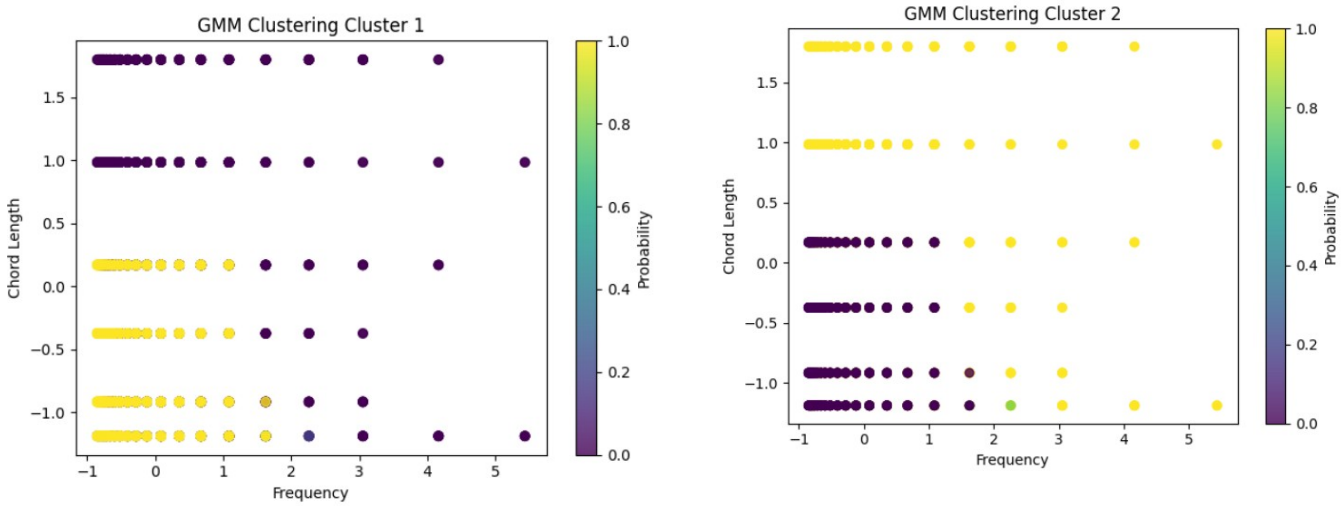
The clusters obtained after K-Means Clustering was applied:



Clusters obtained when plotting Frequency vs Chord Length.

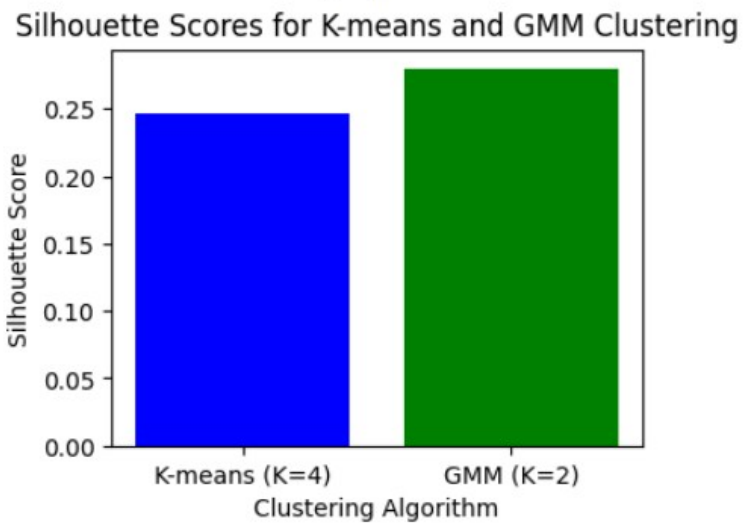
Clusters obtained for GMM

- The following clusters were obtained when plotting frequency vs chord length



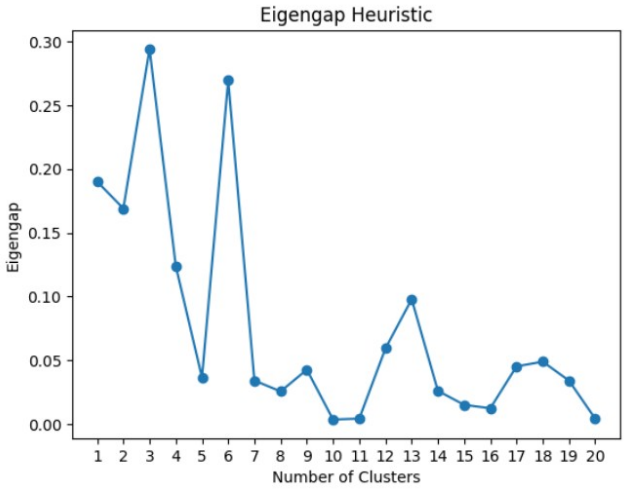
The quality of clusters was compared using Silhouette Scores and the results were:

	K-Means(K = 4)	GMM(K =2)
Silhouette Score	0.24655010919074216	0.2799819296327214



Then dimensionality reduction using PCA was done to the dataset so that it now contains only two columns, and a normalized dataset was generated for testing purposes.

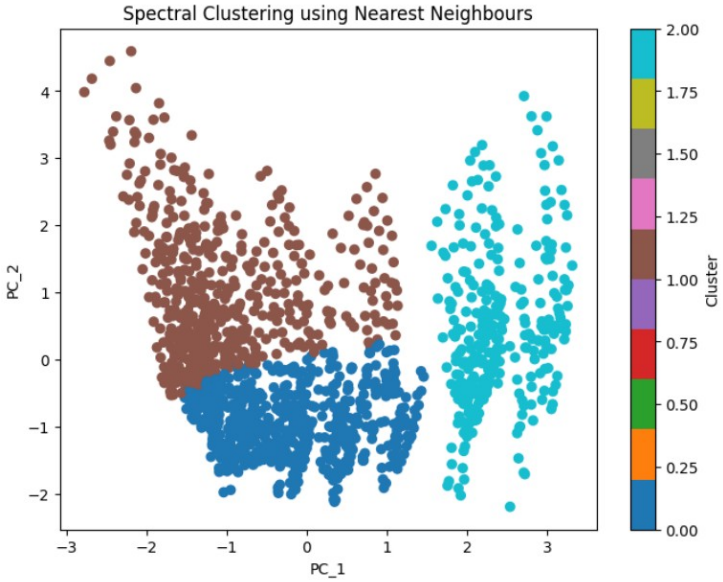
Spectral Clustering on the reduced dataset(non-normalised)



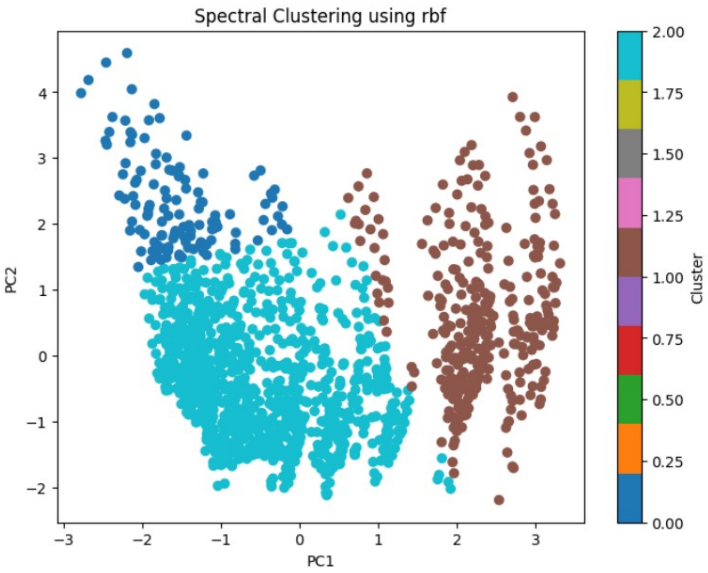
As evident, optimal value of k is

Argmax(eigengaps) = 3

Using affinity as *'nearest-neighbours'*

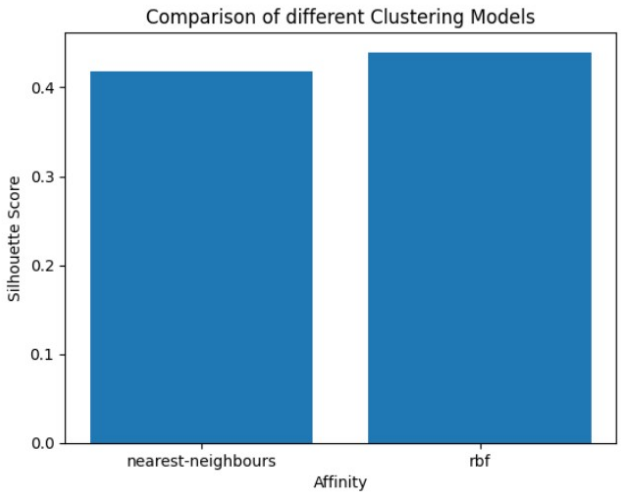


Using affinity as *'rbf'*

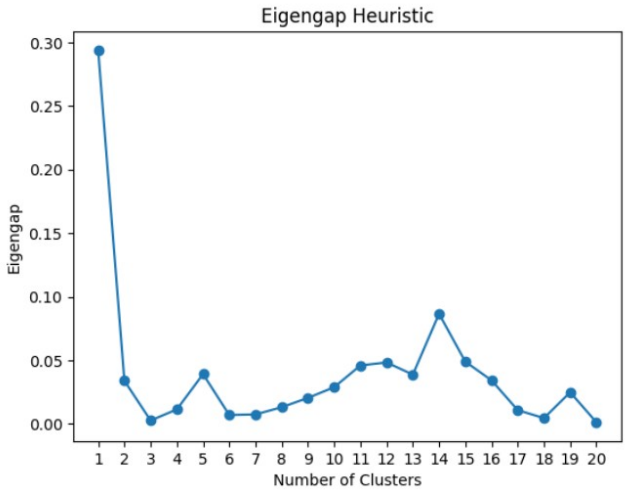


When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.41744824901865635	0.43962899867753225



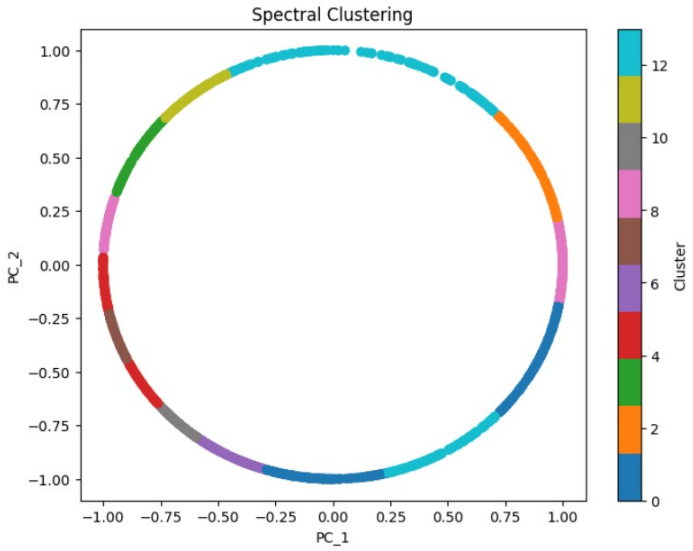
Spectral Clustering on the normalized dataset



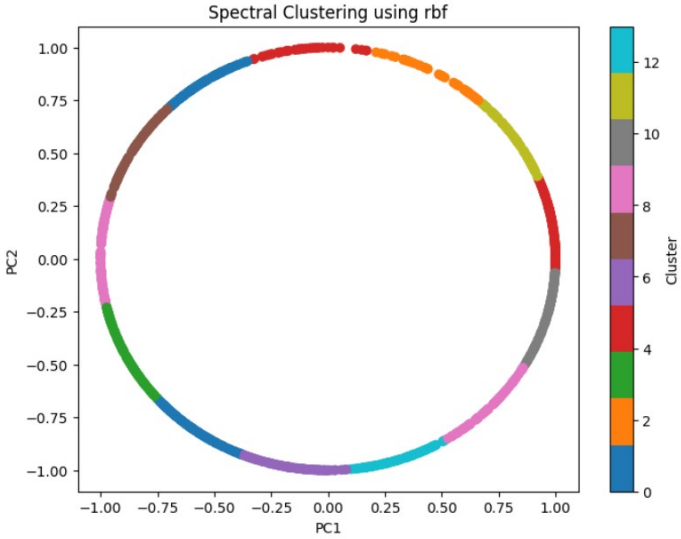
As evident, optimal value of k is

Argmax(eigengaps) = 14

Using affinity as *'nearest-neighbours'*

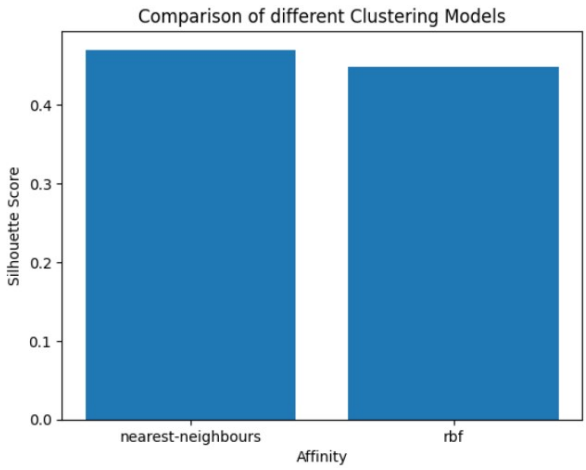


Using affinity as *'rbf'*



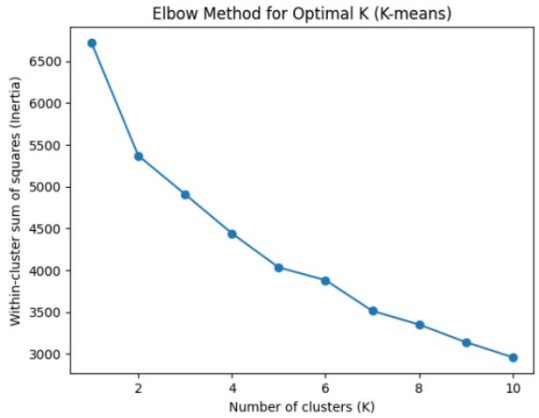
When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.46977160929634826	0.44834754527164733



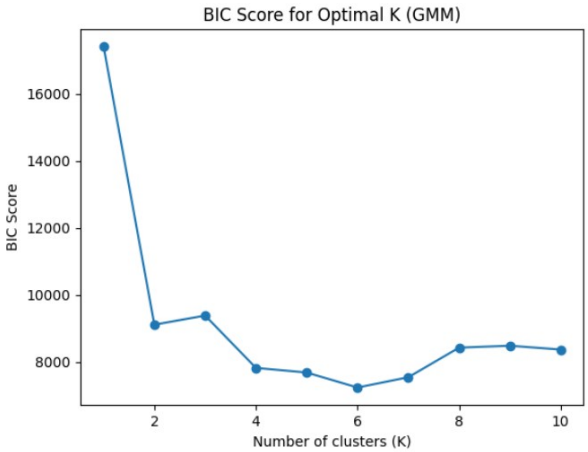
- For Forest Fire Dataset

The elbow method using WCSS(within cluster sum of squares) was employed to find optimal k for K-means clustering and the following graph was observed.



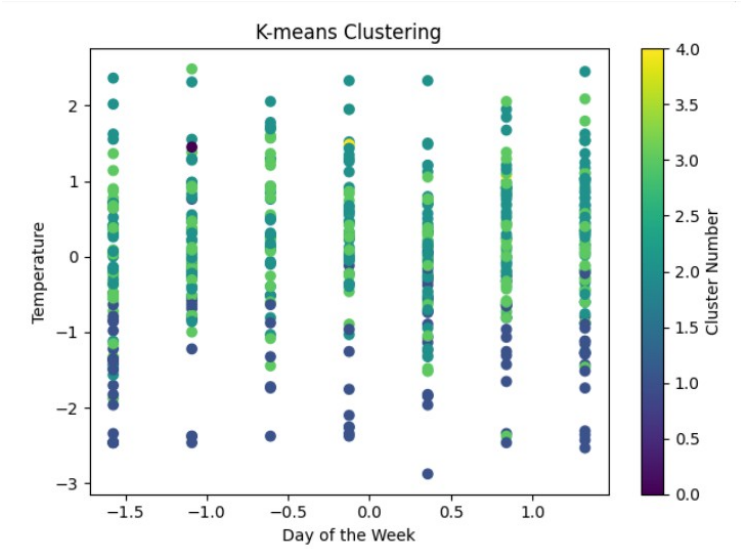
So the value of optimal k is 5.

The BIC scores for various values of K were plotted to find an optimal k for the GMM algorithm.



So optimal value of k is 6.

The clusters obtained after K-Means Clustering was applied:

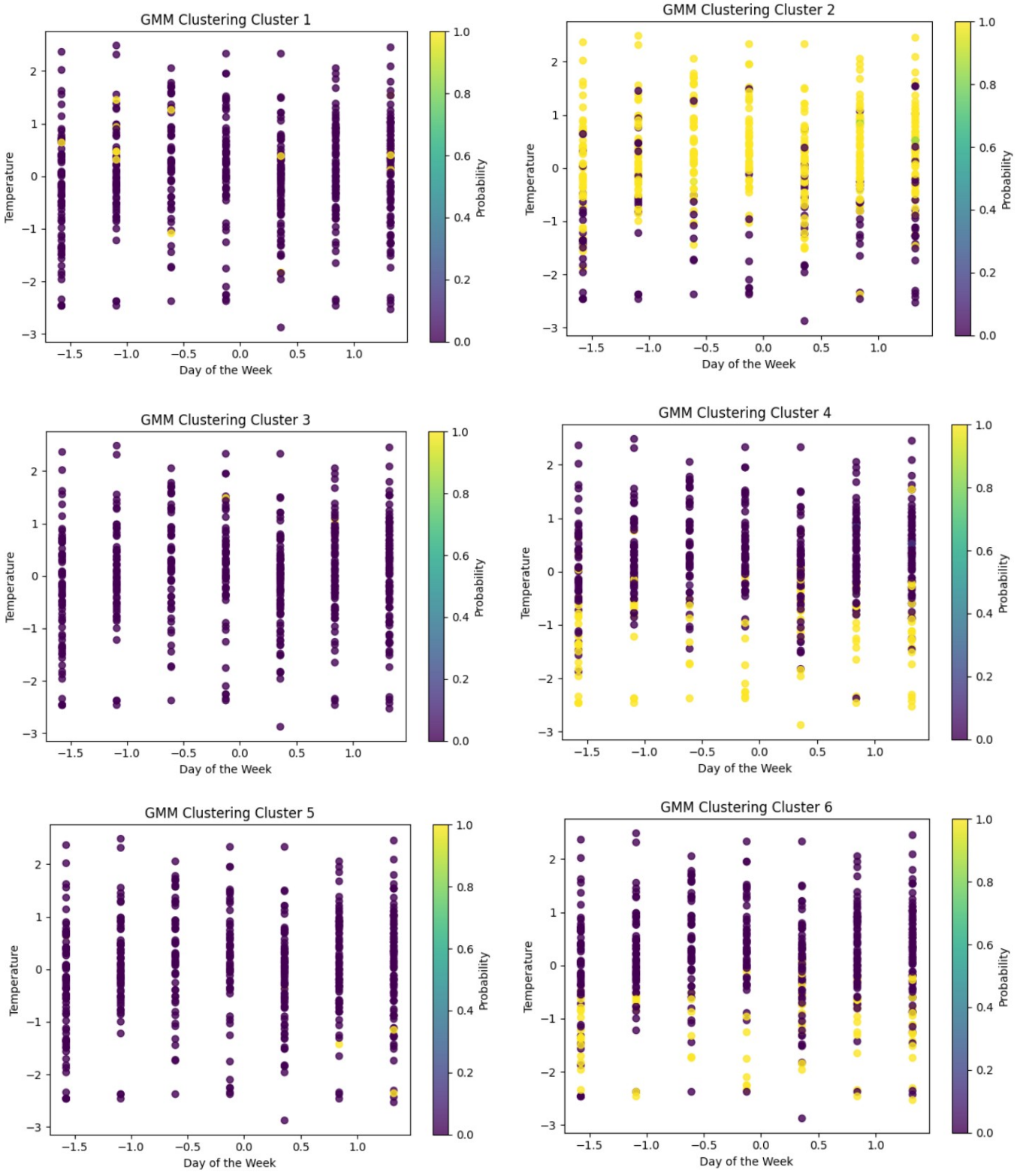


Clusters obtained when plotting

Day of Week vs Temperature

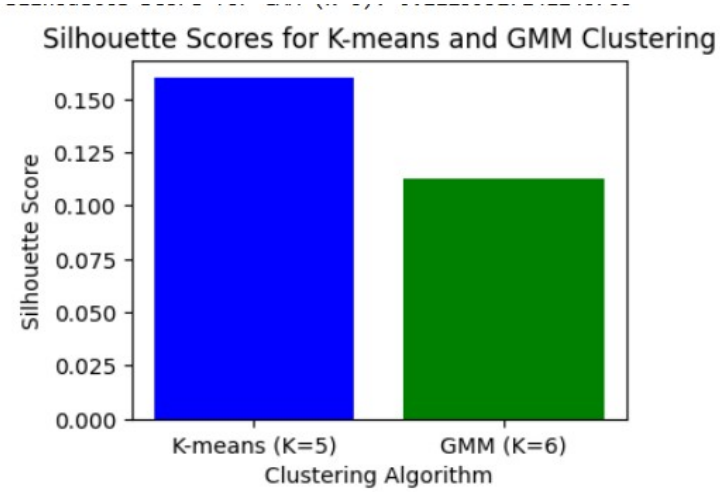
Clusters obtained for GMM

- The following clusters were obtained when plotting day of the week vs temperature



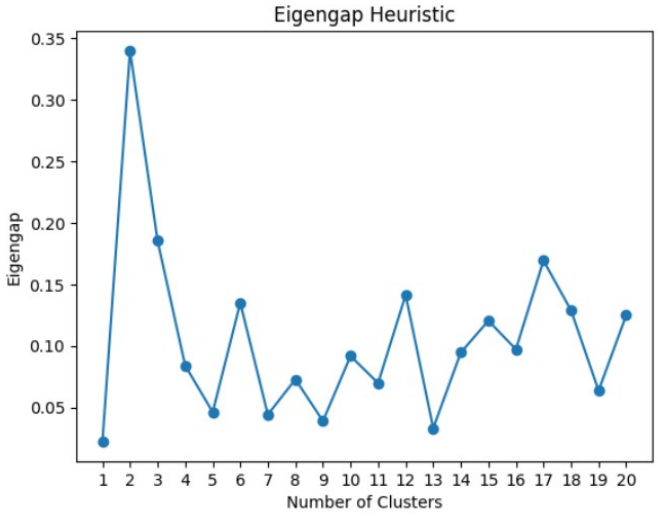
The quality of clusters was compared using Silhouette Scores and the results were:

	K-Means(K = 5)	GMM(K = 6)
Silhouette Score	0.16005850853660933	0.11259527141248768



Then dimensionality reduction using PCA was done to the dataset so that it now contains only two columns, and a normalized dataset was generated for testing purposes.

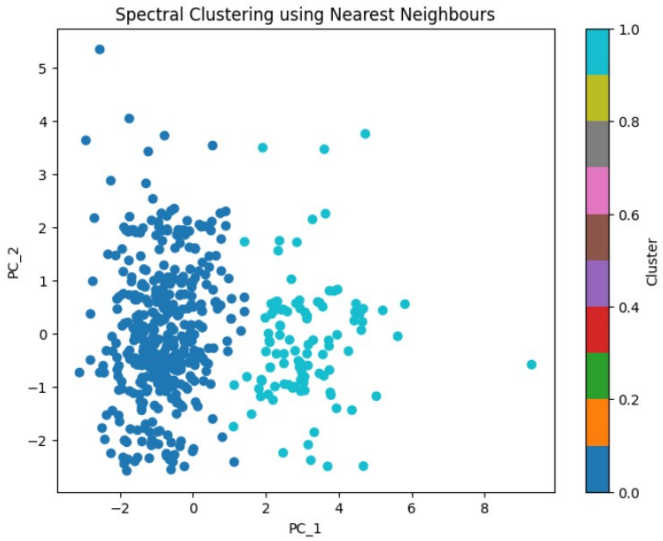
Spectral Clustering on the reduced dataset(non-normalised)



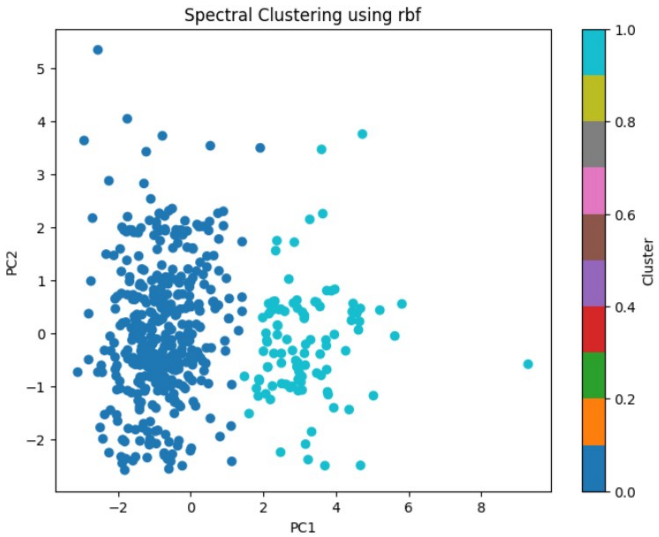
As evident, optimal value of k is

Argmax(eigengaps) = 2

Using affinity as *'nearest-neighbours'*

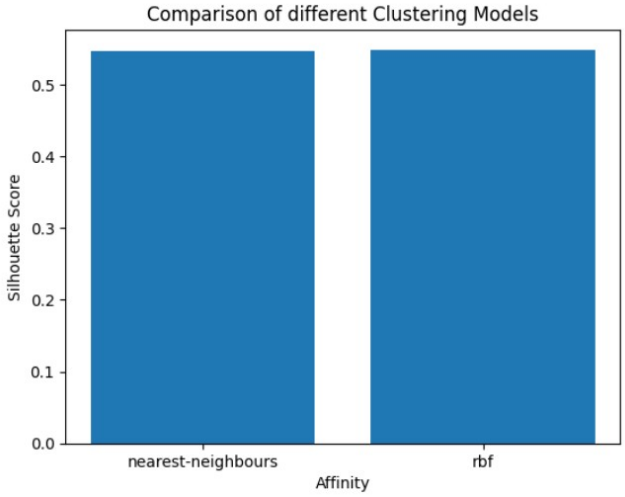


Using affinity as '*rbf'*

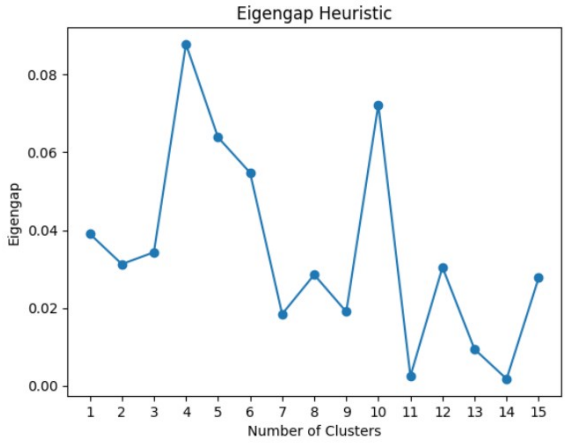


When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.546902634940505	0.5485962608731191



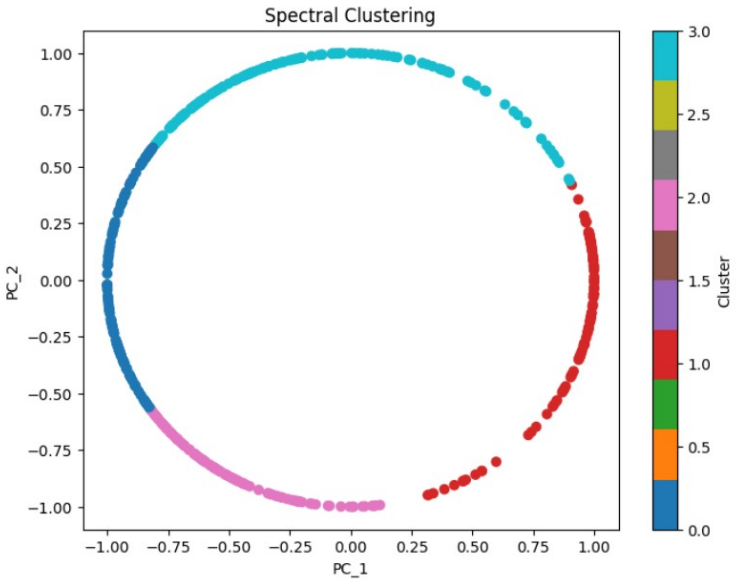
Spectral Clustering on the normalized dataset



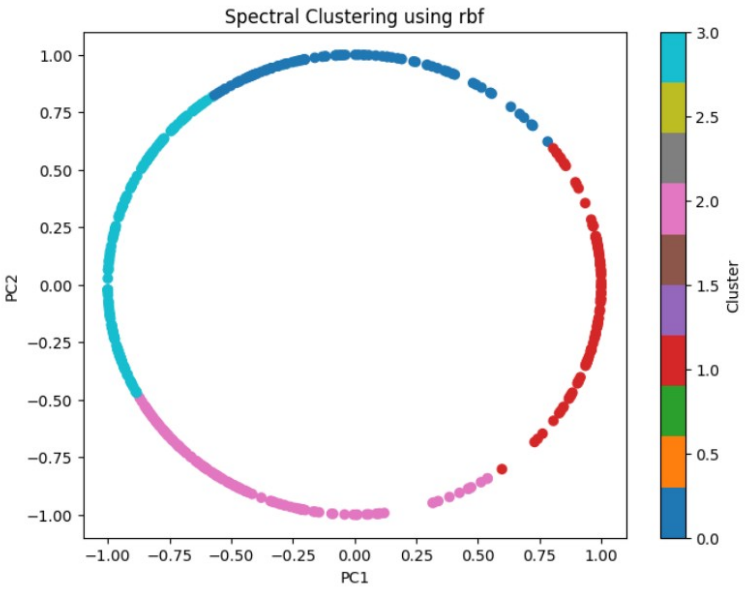
As evident, optimal value of k is

argmax(eigengaps) = 4

Using affinity as *'nearest-neighbours'*

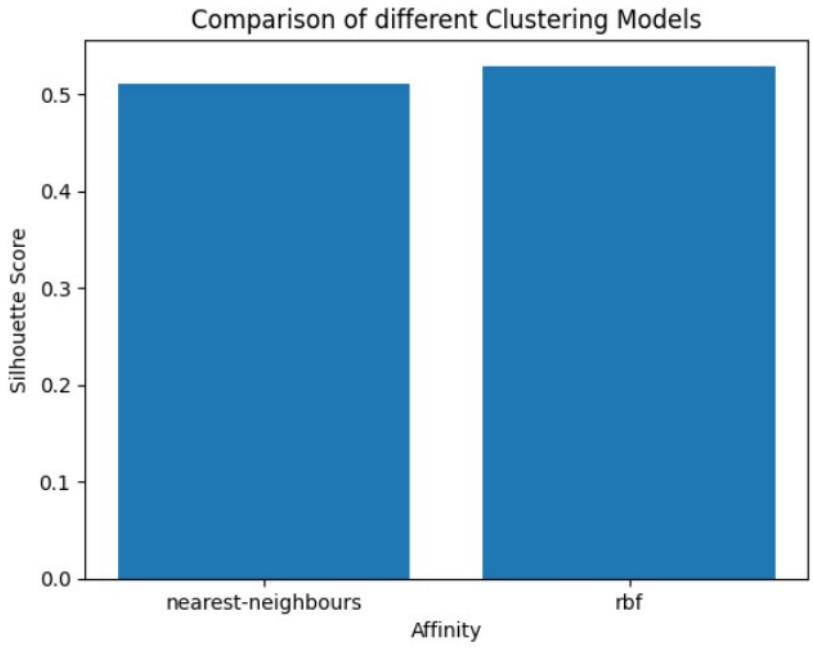


Using affinity as *'rbf'*



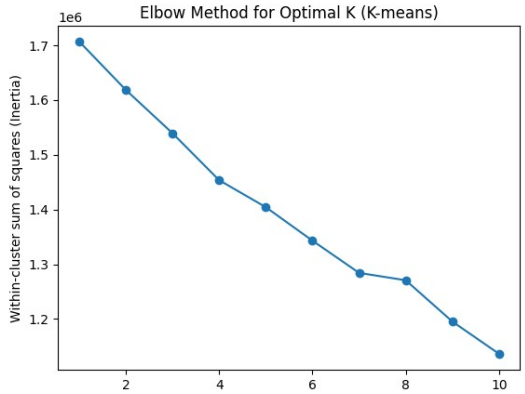
When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.5114582744003617	0.529438704732511

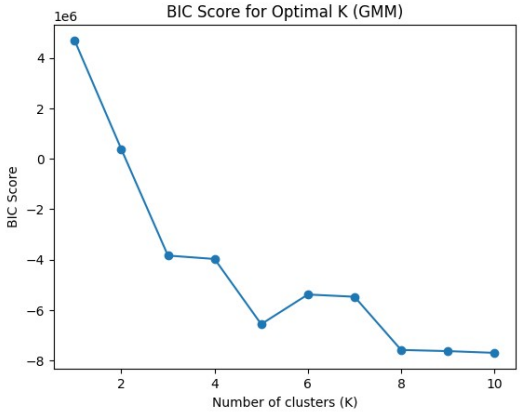


- For <u>Movies</u> Dataset

The elbow method using WCSS(within cluster sum of squares) was employed to find optimal k for K-means clustering and the following graph was observed.



So the value of optimal k is 7.

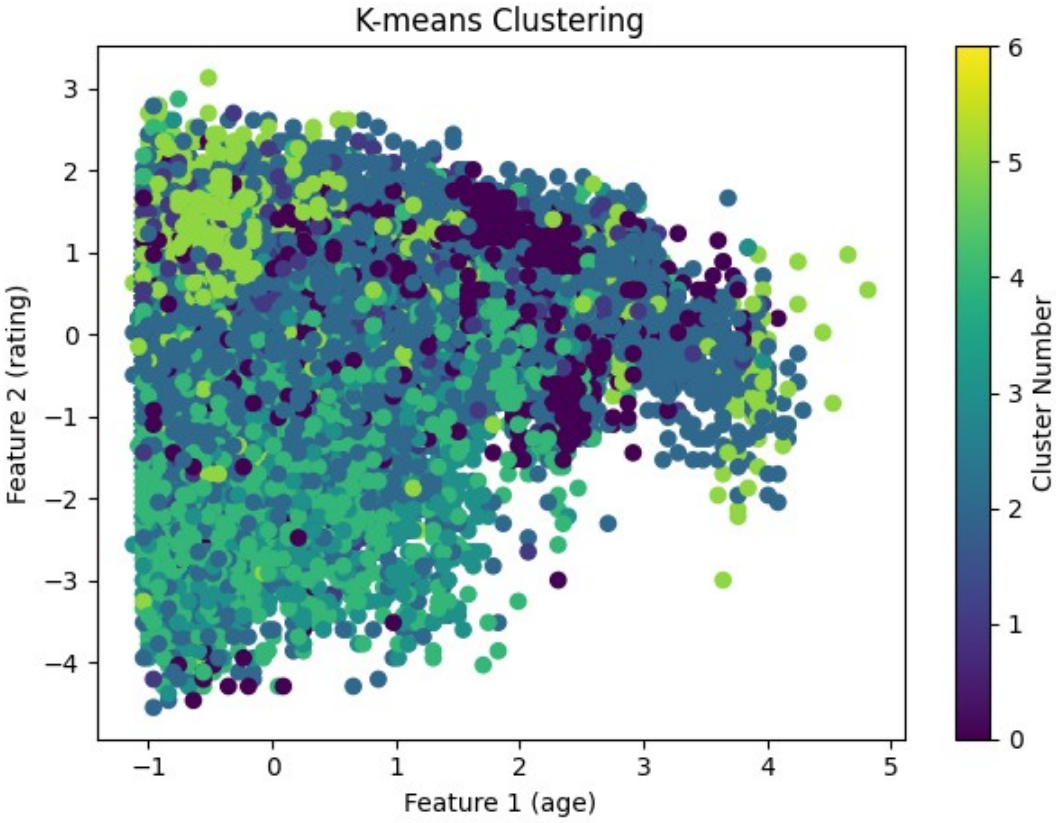


The BIC scores for various values of K were plotted to find an optimal k for the GMM algorithm.

So optimal value of k is 5.

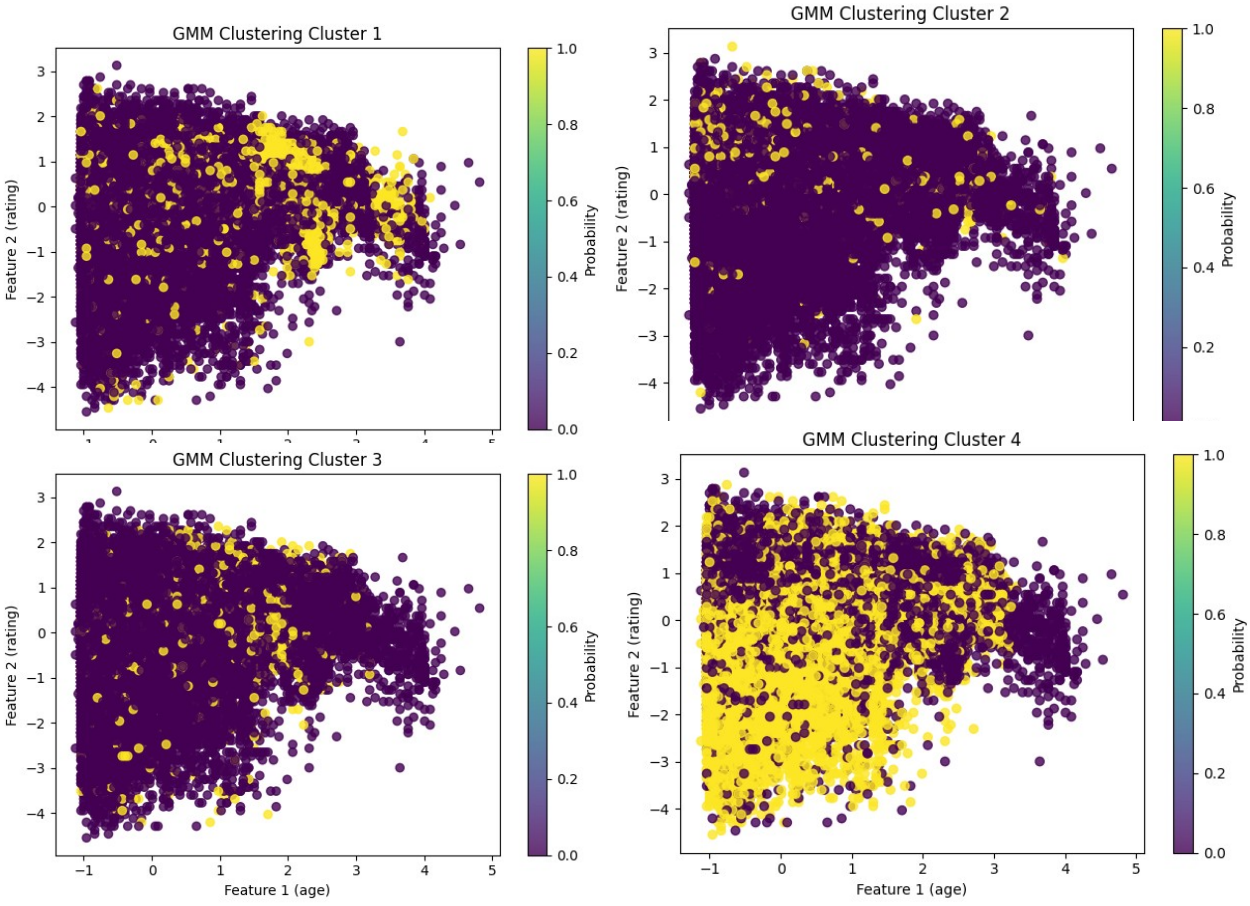
The clusters obtained after K-Means Clustering was applied:

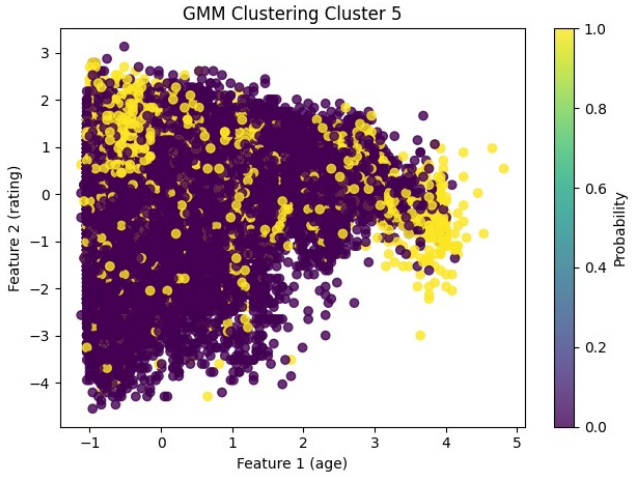
(plotting age vs rating)



Clusters obtained for GMM

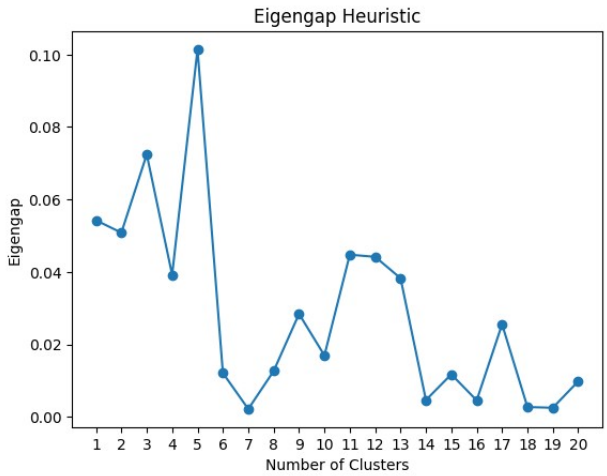
- The following clusters were obtained when plotting day of the week vs temperature





Then **dimensionality reduction using PCA** was done to the dataset so that it now contains only two columns, and a normalized dataset was generated for testing purposes.

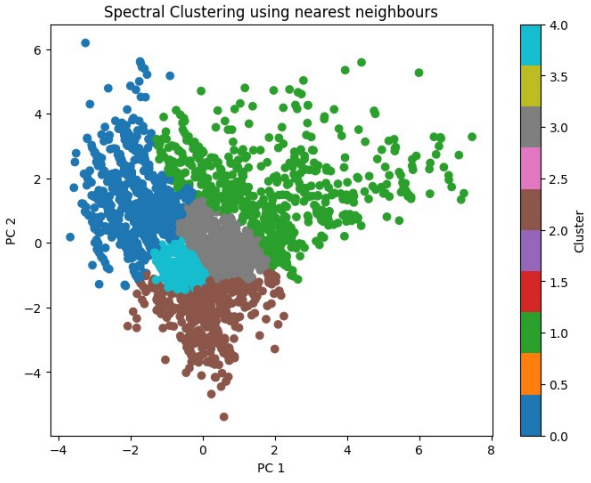
Spectral Clustering on the reduced dataset(non-normalised)



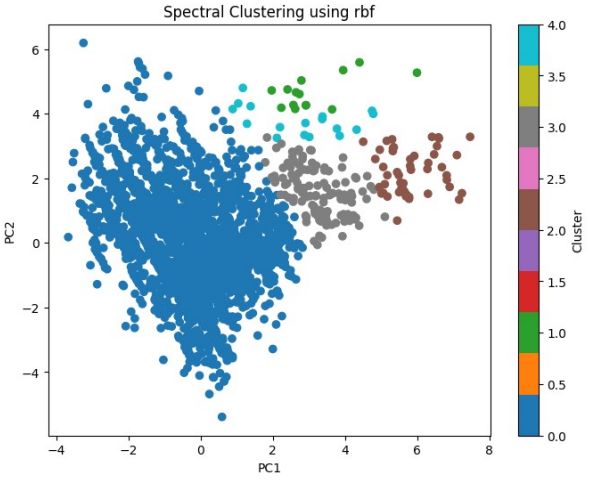
As evident, optimal value of k is

Argmax(eigengaps) = 5

Using affinity as *'nearest-neighbours'*

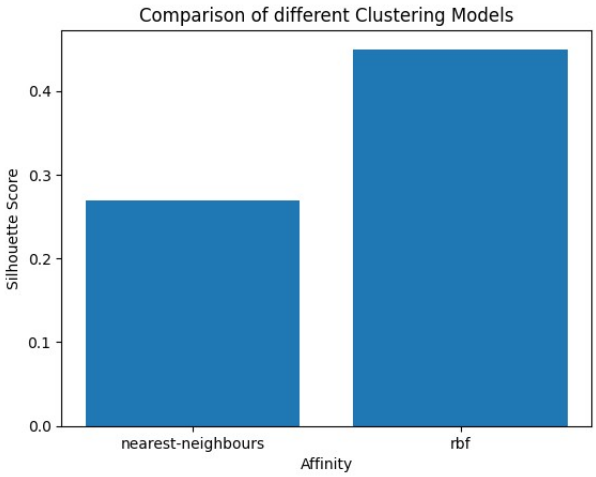


Using affinity as '*rbf*'

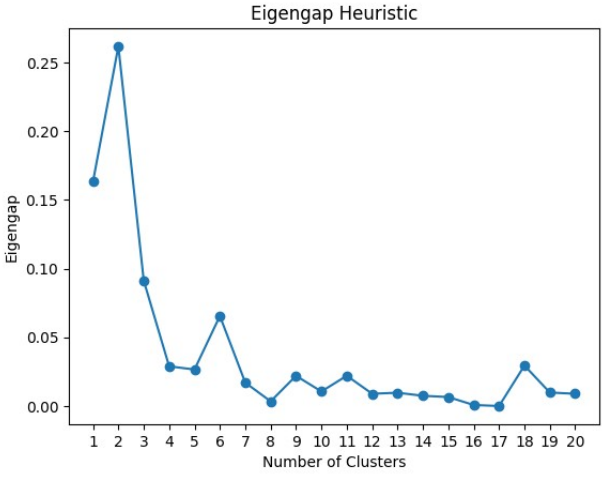


When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.26988	0.44986



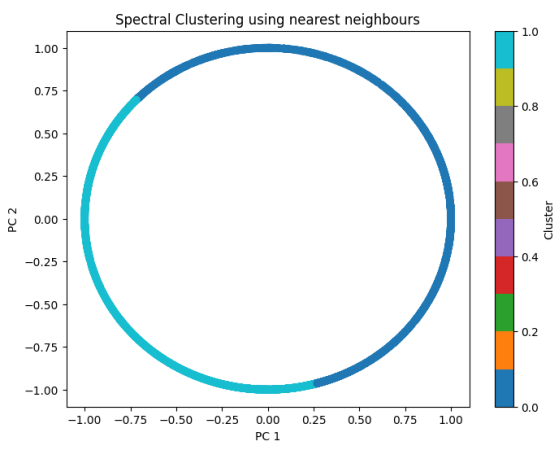
Spectral Clustering on the normalized dataset



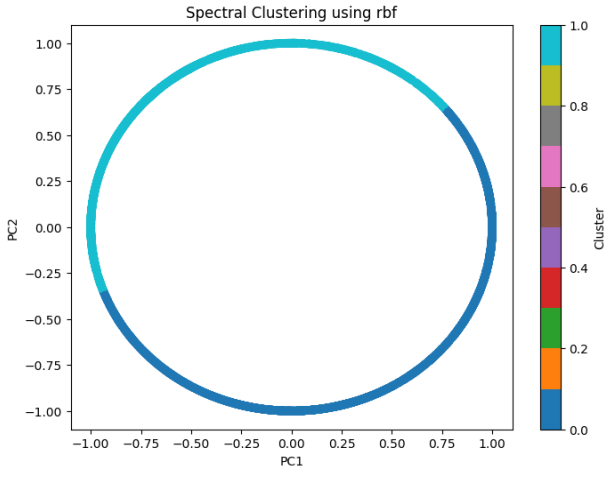
As evident, optimal value of k is

argmax(eigengaps) = 2

Using affinity as *'nearest-neighbours'*

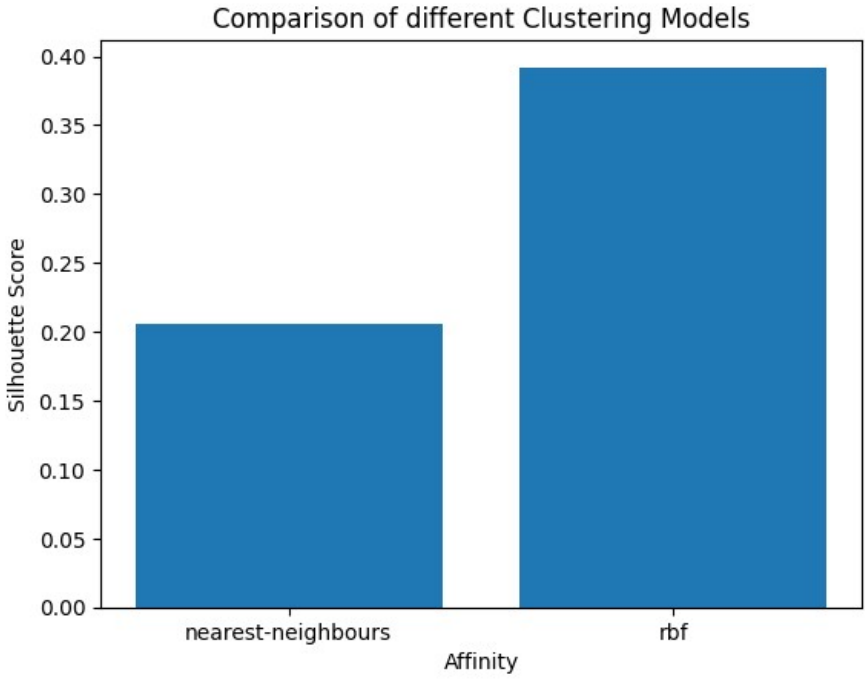


Using affinity as *'rbf'*



When the quality of the clusters in the two methods were compared:

	Nearest-neighbour	rbf
Silhouette Score	0.20545	0.39200



REFERENCES:

- Tayal, D. K., Ahuja, L., & Chhabra, S. (n.d.). Word Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering.
- Shlens, J. (2005). A Tutorial on Principal Component Analysis.
- Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. IEEE International Conference on Data Mining.
- https://www.omdbapi.com/
- https://grouplens.org/datasets/movielens/
- https://www.geeksforgeeks.org/ml-spectral-clustering/ by Alind Gupta.