Final Report

Feature Preprocessing and Clustering

Feature Preprocessing:

In feature preprocessing, we implemented,

1) Feature Selection
2) Kernels
3) Principal Component Analysis

1) Feature Selection:

In feature selection we successfully selected subset of features which either maintained the original accuracy or increased the accuracy.

In feature selection we implemented,

a) Sequential Forward Selection
b) Sequential Backward Selection

For both the algorithms we use Fisher's Discriminant Ratio as the criterion function.

2) Kernels

In kernels we successfully got the similarity in the higher dimensions which helped us in increasing the accuracy.

In kernels, we implemented,

a) Gaussian Kernel
b) Polynomial Kernel

Various degrees of polynomial kernels are tried and the accuracy increases by increasing its degree.

Gaussian kernel, the most famous kernel increased the accuracy by 3%.

3) Principal Component Analysis

In Principal Component Analysis, we successfully visualised Iris dataset and Internet Advertisement dataset by projecting them to 2 dimension.

We also projected the entire data into lower dimension which captures about 90% of the data.

Clustering :

In Clustering, we implemented,

1) K-Means Clustering
2) DBSCAN Clustering
3) Adaptive DBSCAN

1) K-Means Clustering:

In K-Means we successfully clustered the given data using the euclidean distance

Here, we also implemented cluster validity indices such as,

a) DB Index
b) Dunn Index

Both the cluster validity indices helped us in detecting the best value of k.

2) DBSCAN Clustering:

In DBSCAN we successfully clustered the given data based on the density. Here we successfully detected the outlier points and the data points.

Adaptive DBSCAN:

But the hard part in DBSCAN is getting the value of epsilon and minimum points.

We solved this by implementing the paper [1].

Here, we automatically find the epsilon value and the minimum points value using the algorithm given in [1] which successfully worked.

Feature Selection and Clustering:

Now, after applying both feature selection and clustering on the datasets. We combined both the methods which also worked similar to the original one or more accurate than the original one whose results are attached below.

Results :

a. K Means Algorithm.

Dataset	Method	Number of Features Selected	Minimum DB Index		Maximum Dunn Index	
			K	DB Value	K	Dunn Value
Arrhythmia	Feature Selection	100	16	1.029	2	0.939
Arrhythmia	Feature Selection	110	8	0.8164	2	0.9286
Arrhythmia	Feature Selection	120	6	0.9141	2	0.9746
Arrhythmia	Feature Selection	130	14	0.0947	4	0.7960
Arrhythmia	Feature Selection	140	16	0.9620	2	0.8902
Arrhythmia	Feature Selection	150	14	1.0082	2	0.8858
Arrhythmia	Feature Selection	160	16	0.9656	2	0.8409
Arrhythmia	Feature Selection	170	16	1.0014	2	0.8357
Arrhythmia	Feature Selection	180	11	1.0947	2	0.7453
Arrhythmia	Feature Selection	190	14	1.3120	2	0.7290
Arrhythmia	Feature Selection	200	14	1.296	2	0.7300
Arrhythmia	Polynomial Kernel	452	15	1.0580	2	0.7300
Arrhythmia	Gaussian Kernel	452	14	1.186	2	0.7302

Dataset	Method	Number of Features Selected	Minimum DB Index		Maximum Dunn Index	
			K	DB Value	K	Dunn Value
Internet Advertisement	**Feature Selection**	100	13	0.2542	2	4.1337
Internet Advertisement	**Feature Selection**	200	15	0.0845	2	4.1335
Internet Advertisement	**Feature Selection**	300	15	0.0129	2	4.1330
Internet Advertisement	**Feature Selection**	400	6	0.2745	2	4.0809
Internet Advertisement	**Feature Selection**	500	14	0.2285	2	4.0879
Internet Advertisement	**Feature Selection**	550	16	0.0061	2	4.0805
Internet Advertisement	**Feature Selection**	600	8	0	2	4.0803
Internet Advertisement	**Feature Selection**	650	9	0.2341	2	4.0801
Internet Advertisement	**Feature Selection**	700	8	0	2	4.0800
Internet Advertisement	**Feature Selection**	750	12	0.2785	2	4.0799
Internet Advertisement	**Feature Selection**	800	8	0.2643	2	4.0798
Internet Advertisement	**Feature Selection**	850	9	0.2294	2	4.0795
Internet Advertisement	**Feature Selection**	900	16	0.2296	2	4.0794

Internet Advertisement	**Feature Selection**	950	14	0.2591	2	4.0793
Internet Advertisement	**Feature Selection**	1000	7	0.3130	2	4.0792
Internet Advertisement	**PCA**	435	11	0.2370	2	4.0809
Internet Advertisement	**Gaussian Kernel**	100	8	0	2	4.0792
Internet Advertisement	**Gaussian Kernel**	200	6	0.2262	2	4.0792

b. Adaptive DBSCAN :

Dataset	Method	Number of Features Selected	Epsilon	Min - Points	Number of Clusters formed
Arrhythmia	-	-	379	2	2
Arrhythmia	-	-	418.77	2	2
Arrhythmia	-	-	413	19	1
Internet Advertisement	**Feature Selection**	400	35.99	57	2
			40.27	160	2
			44.24	78	2
Internet Advertisement	**Feature Selection**	450	48.14	49	2
			45.76	36	2

			44.76	5	2
Internet Advertisement	**Feature Selection**	500	35.9	57	2
			40.2	160	2
			44.2	78	2
Internet Advertisement	**Feature Selection**	550	48.14	49	2
			45.7	36	2
			44.4	5	2
Internet Advertisement	**Feature Selection**	600	35.99	56	2
			40.30	161	2
			44.26	78	2
Internet Advertisement	**Feature Selection**	650	35.99	56	2
			40.321	162	2
			44.27	78	2
Internet Advertisement	**Feature Selection**	700	35.99	56	2
			40.32	162	2
			44.27	78	2
Internet Advertisement	**Feature Selection**	750	33.59	56	2
			40.32	162	2
			44.2	78	2
Internet Advertisement	**Feature Selection**	800	35.99	56	2
			40	162	2

			44.2	78	2
Internet Advertisement	**Feature Selection**	850	36.02	55	2
			40.33	163	2
Internet Advertisement	**Feature Selection**	900	36	55	2
			49	163	2
			44.2	78	2
Internet Advertisement	**Feature Selection**	950	36.00	55	2
			40.343	163	2
			44.28	78	2
Internet Advertisement	**Feature Selection**	1000	36.01	55	2
			40.35	163	2
			44.28	78	2
Internet Advertisement	**PCA**	435	36.012	55	2
			40.35	163	2
			44.26	78	2
Internet Advertisement	**Gaussian Kernel**	100	36.0122	55	2
			40.357	163	2
			44.28	78	2
Internet Advertisement	**Gaussian Kernel**	200	1.6324	140	2
			1.93	108	2

	1.944	4	6