MACHINE LEARNING APPROACH FOR FLOW CYTOMETRY DATA ANALYSIS

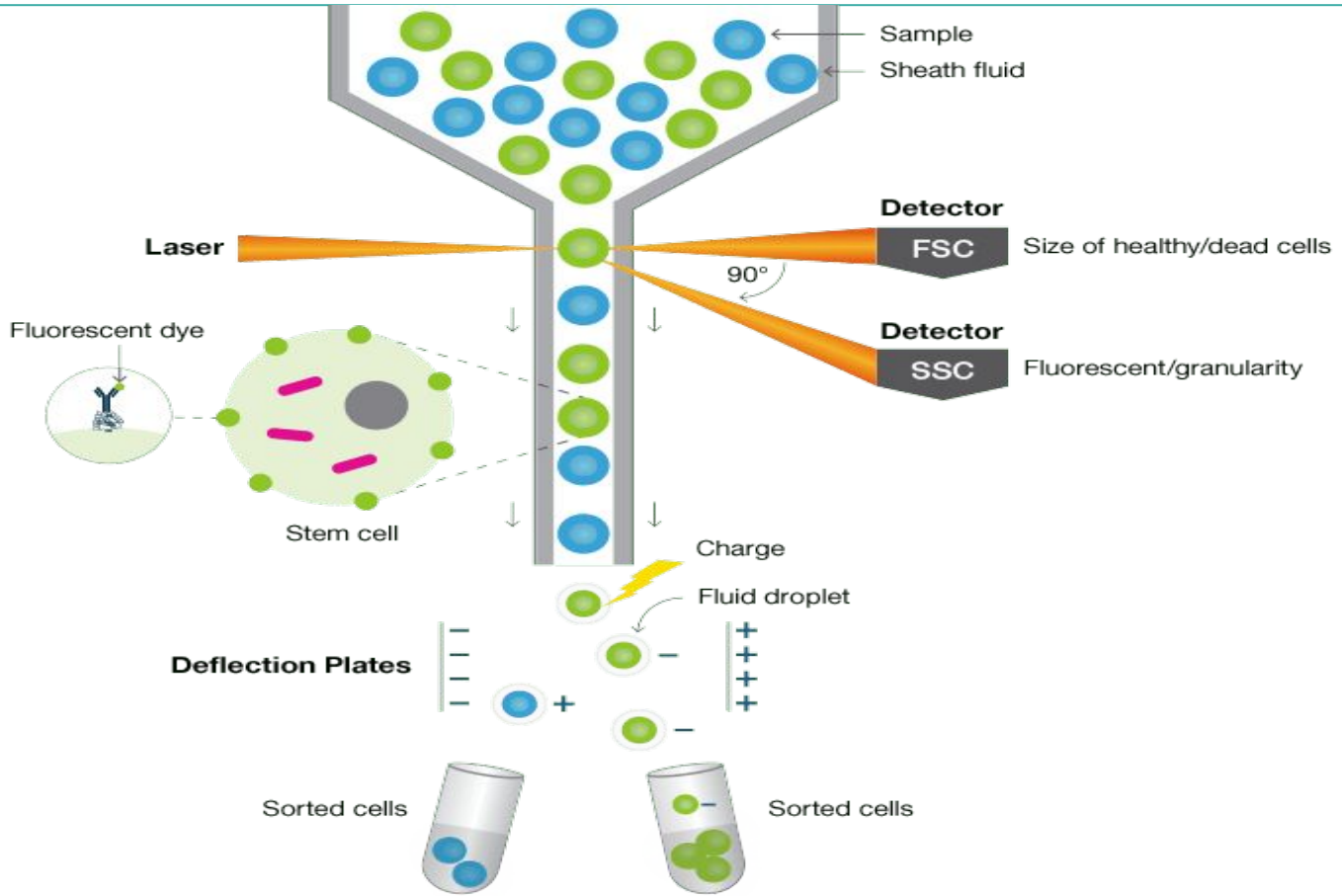
Group no.-16

Group Members :

Anuradha Kumari 2019149
Kunal 2019430
Himani varolia PhD21115

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Flow Cytometry



Problem Statement

- **Flow cytometry** is a powerful tool to study and to measure different parameters within the samples, which includes the quantification of various cell populations and the specific markers present on the cell's surface.the single cell
  - Applications**:** Immunology, molecular biology, bacteriology, virology, cancer biology and infectious disease monitoring.
- **Problems**:
  - High dimensionality of data
  - No automation
  - Big data to visualize and explore clusters in 2D graphs through a manual tedious process.

- **Aim**: To study and design learning (ML) based efficient Tool for clustering FCM data.
  - Each row of the tabular data will act as an input point in the multidimensional space where each column is a dimension.
  - The output will be the clusters and detailed information of the possible significant clusters of the input cells.
  - This will discard the manual labor that the expert oncologists have to perform monotonously in order to identify these clusters before inferring its medical consequences.

Dataset

- Data1 used is Open Source **SDY50** - Analysis of the activation kinetics of signaling pathways involved with dendritic cell maturation.
- Data 2 used is **SDY820** - Human Immune Signature of Mycobacterium Tuberculosis infection.
- Both the datasets and supporting publication are available at ImmPort (https://www.immport.org, under study accession **SDY50 and SDY820**.
- Data 1 shape : (51812, 8)
- Data 2 shape : (300000, 17)

FSC-A	SSC-A	ALEXA 488-A	ALEXA 350-A	PE-A	ALEXA 750-A	ALEXA 647-A	Time
21471.119141	4921.560059	12.960001	212.400009	0.000000	-18.799999	-1.940000	0.0
43539.757812	73998.359375	41.040001	4717.440430	178.480011	77.080002	62.080002	0.1
20720.480469	28576.800781	16.200001	83.520004	496.640015	25.379999	36.860001	0.1
82162.960938	64777.324219	73.440002	211.680008	221.160004	212.440002	103.790001	0.1
143799.046875	81612.359375	35.640003	556.559998	6798.729980	106.220001	454.930023	0.1

Snippet of the dataset AA1168_Mix1_C1.32035.fcs

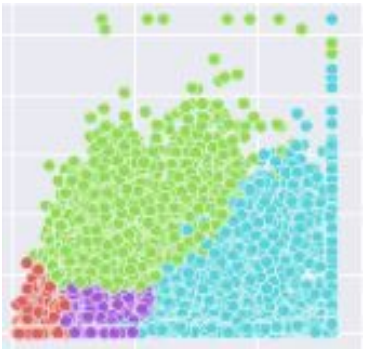
FSC-A	FSC-H	FSC-W	SSC-A	SSC-H	SSC-W	APC-A	Alexa Fluor 700-A	APC-eFluor 780-A	V500-A	FITC-A	PE-A	PerCP-Cy5-5-A	PE-Cy7-A	eFluor 450-A	BV650-A	Time
81213.859375	60287.0	88284.890625	17399.201172	12063.0	94526.570312	404.040009	1339.599976	3243.600098	2237.020020	8366.440430	359.720001	4229.680176	14791.140625	288.600006	851.400024	10.4
62029.242188	48986.0	82985.921875	16130.660156	13934.0	75867.593750	251.160004	10231.280273	13005.000000	2180.040039	10448.799805	1440.239990	6324.500000	1172.180054	562.400024	42.139999	10.5
5025.930176	5417.0	60804.750000	4668.300293	6333.0	48309.136719	33.670002	-108.120003	-38.080002	47.360001	-37.740002	18.360001	-15.470000	4.300000	34.779999	33.540001	10.5
73410.609375	56388.0	85320.242188	21208.460938	16592.0	83770.351562	594.230042	22195.880859	28566.121094	3142.780029	13877.219727	3143.640137	11279.450195	2371.020020	925.739990	17.200001	10.5
79933.492188	60990.0	85891.476562	23055.759766	20442.0	73915.578125	283.920013	17793.560547	23762.589609	2373.919922	9299.580078	2861.439941	9297.470703	1985.739990	520.960022	110.080002	10.5

Snippet of the dataset Specimen_001_XT0141_005.605757.fcs

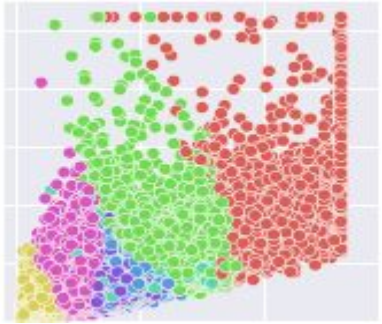
Progress: Until Mid-sem

We applied the baseline **K-Means** clustering which is an Unsupervised learning algorithm. It allows us to cluster the data into k different groups having similar properties. It is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

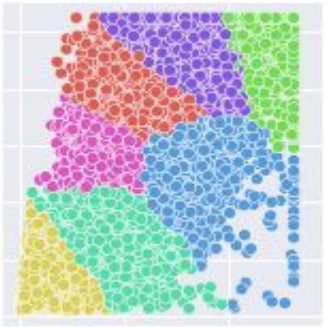
Analysis: Gating for identification of homogeneous cell populations that share a particular function, Interpretation i.e., finding correlations between some characteristics of the cell populations, and Visualization for any expert to interpret the results.



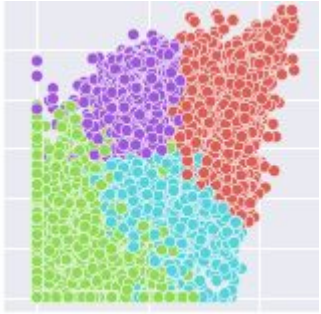
K-means Clustering Data 1, K =4



K-means Clustering Data 2, K =7



K-means Clustering Data 2, K =7

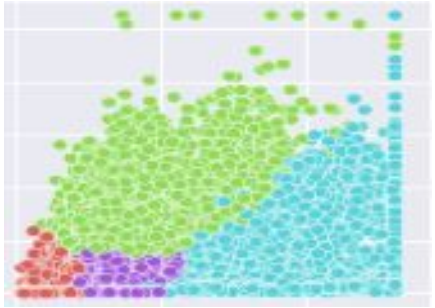


K-means Clustering Data 2, K =4

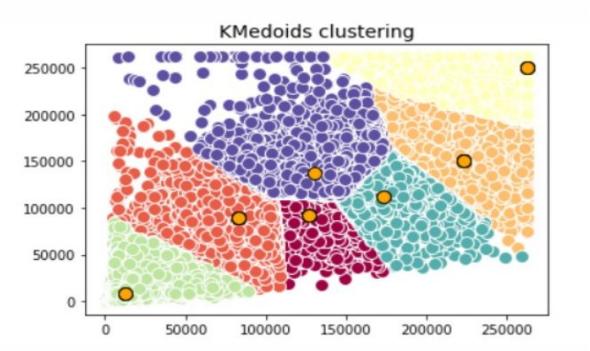
Approaches

- We have applied six popular clustering algorithms which are K-Means, K-medoids, Self Organizing Map, Gaussian Mixture, Agglomerative clustering, and DBSCAN.
- All the algorithms are applied with different configurations and different parameters like number of clusters, epsilon, min Points to cluster, etc to obtain satisfactory results on the two datasets.
- **Pre-processing and feature selection**
  - Sub-sampling, Standardization, MinMaxScaling and hyperbolic sine transformation
  - The total number of markers (columns) in this dataset1 are eight namely, ALEXA 647-A, ALEXA 488-A, SSC-A, ALEXA 350-A, ALEXA 750-A, PE-A, FSC-A are included and Time is discarded during processing. The total number of markers (columns) in this dataset2 are seventeen namely, in which FSC-A, SSC-A, APC-A, Alexa Fluor 700-A, APC-eFluor 780-A, V500-A, FITC-A, PE-A, PerCP-Cy5-5-A, PE-Cy7- A, eFluor 450-A, BV650-A are included and FSC-H, FSCW,SSC-H, SSC-W and Time are excluded.
- **Evaluation Metrics**
  - Silhouette Coefficient ,Calinski and Harabasz score, and Davies-Bouldin score
- **Visualization**
  - Scatter plot, Histogram , TNSE plot, PairPlot and PCA.

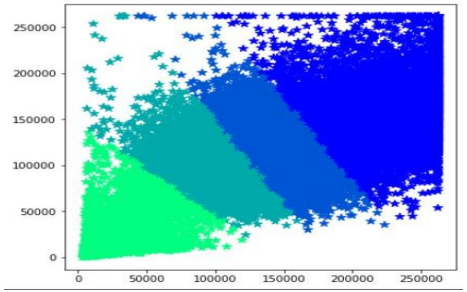
Results(Some Relevant Clustered Outputs)



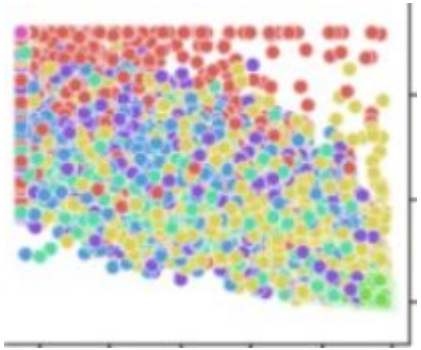
K-Means



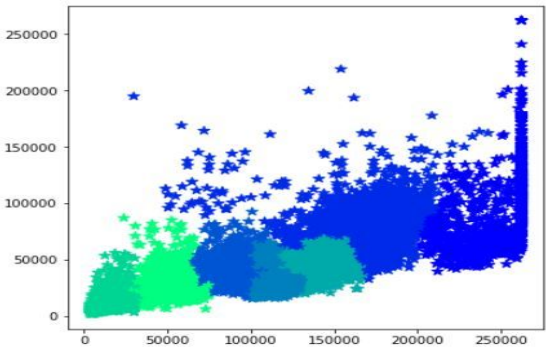
K-Medoids



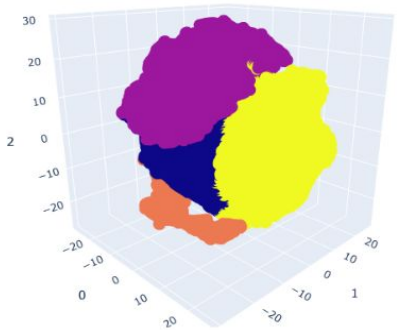
SOM



Gaussian Mixture



Agglomerative clustering



DBSCAN

Results and Analysis : Dataset 1

WITHOUT DATA TRANSFORMATION							
METHOD	EVALUATION METRICS	NUMBER OF CLUSTERS					
		4	5	6	7	8	9
K- MEANS	Silhouette_score :	0.46	0.42	0.42	0.40	0.40	0.39
	Calinski Score :	90489.59	89467.28	86645.27	84912.18	82862.89	80090.61
	Davies Bouldin Score :	0.73	0.81	0.83	0.84	0.86	0.84
K-Medoids	Silhouette_score :	0.44	0.39	0.37	0.38	0.37	0.36
	Calinski Score :	72030.46	69089.84	66821.39	67458.29	64321.48	61176.35
	Davies Bouldin Score :	0.75	0.87	0.90	0.85	0.88	0.91
Gaussian Mix	Silhouette_score :	-0.01	0.00	-0.03	-0.02	-0.02	-0.01
	Calinski Score :	7449.29	6482.83	6616.27	4772.82	6656.40	6590.22
	Davies Bouldin Score :	5.91	4.36	5.51	5.89	6.57	4.25
SOM	Silhouette_score :	0.37	0.36	0.36	0.35	0.32	0.32
	Calinski Score :	74562.47	77689.00	81448.55	82261.35	72558.91	73748.39
	Davies Bouldin Score :	0.86	0.92	0.97	1.01	1.21	1.61
Agglomerative Clustering	Silhouette_score :	0.63	0.39	0.39	0.38	0.37	0.37
	Calinski Score :	33085.94	30897.86	29077.75	28737.59	29378.12	29250.99
	Davies Bouldin Score :	0.62	0.76	0.73	0.76	0.75	0.85
		eps					
		<90	90.00	100.00	105.00	110.00	115.00
DBSCAN	Silhouette_score :	NA	0.20	0.22	0.22	0.21	
	Calinski Score :		20.46	24.24	24.06	27.26	
	Davies Bouldin Score :		0.25	0.48	0.51	0.68	

Results and Analysis : Dataset 1

WITH DATA TRANSFORMATION							
METHOD	EVALUATION METRICS	NUMBER OF CLUSTERS					
		4	5	6	7	8	9
K- MEANS	Silhouette_score :	0.32	0.29	0.28	0.27	0.24	0.23
	Calinski Score :	35253.20	32081.24	29838.19	27395.94	25404.91	23681.85
	Davies Bouldin Score :	1.03	1.13	1.28	1.27	1.35	1.44
K-Medoids	Silhouette_score :	0.44	0.39	0.37	0.38	0.37	0.36
	Calinski Score :	72030.46	69089.84	66821.39	67458.29	64321.48	61176.35
	Davies Bouldin Score :	0.75	0.87	0.90	0.85	0.88	0.91
Gaussian Mix	Silhouette_score :	-0.02	-0.03	-0.02	-0.04	-0.04	-0.02
	Calinski Score :	4279.63	4368.68	4451.94	4375.00	4381.68	4396.75
	Davies Bouldin Score :	11.79	3.13	5.55	4.20	4.21	4.23
SOM	Silhouette_score :	0.54	0.56	0.57	0.58	0.58	0.58
	Calinski Score :	181919.45	247496.74	293880.38	397631.90	473434.87	521020.03
	Davies Bouldin Score :	0.51	0.49	0.48	0.47	0.47	0.48
Agglomerative	Silhouette_score :	0.63	0.39	0.39	0.38	0.37	0.37
	Calinski Score :	33085.91	30897.85	29077.72	28737.54	29378.05	29250.90
	Davies Bouldin Score :	0.62	0.76	0.73	0.76	0.75	0.85

Results and Analysis : Dataset 2

WITHOUT DATA TRANSFORMATION							
METHOD	EVALUATION METRICS	NUMBER OF CLUSTERS					
		4	5	6	7	8	9
K- MEANS	Silhouette_score :	0.51	0.52	0.46	0.44	0.40	0.37
	Calinski Score :	270170.43	259928.64	243650.94	223954.36	207636.05	196808.67
	Davies Bouldin Score :	0.71	0.80	1.00	1.00	1.12	1.18
K-Medoids	Silhouette_score :	0.49	0.33	0.33	0.32	0.32	0.28
	Calinski Score :	35340.99	29236.50	25145.05	22274.78	19621.94	17309.62
	Davies Bouldin Score :	0.74	1.16	1.26	1.18	1.21	1.29
Gaussian Mix	Silhouette_score :	0.34	0.36	0.34	0.33	0.29	0.33
	Calinski Score :	17729.41	15198.59	15518.28	14203.48	16204.33	13776.96
	Davies Bouldin Score :	1.08	1.85	2.08	2.27	2.00	2.41
DBSCAN	eps	4	125	130	131	132	133
	Silhouette_score :	NA	0.51106	0.32909	0.329748	0.32965	0.32947
	Calinski Score :		22.3056	26.06743	27.74058	30.14046	27.950214
	Davies Bouldin Score :		0.654098	0.867038	0.9410813	0.926204	1.056789
Agglomerative	Silhouette_score :	0.61	0.45	0.41	0.41	0.39	0.35
	Calinski Score :	37821.55	37679.44	36058.14	34472.50	33712.48	32456.68
	Davies Bouldin Score :	0.59	0.75	0.86	0.86	0.86	0.87
SOM	Silhouette_score :	0.43	0.45	0.41	0.41	0.36	0.36
	Calinski Score :	486206.10	450567.11	466593.49	469536.56	382692.29	427697.77
	Davies Bouldin Score :	0.75	0.75	0.86	0.87	0.93	0.95

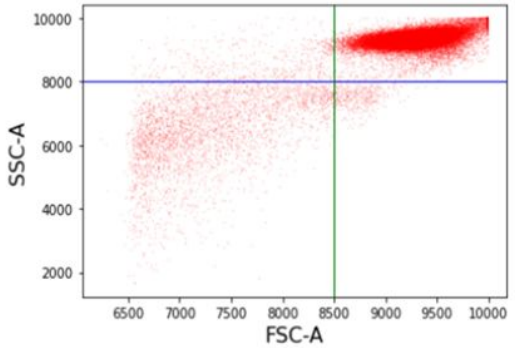
Results and Analysis : Dataset 2

WITH DATA TRANSFORMATION							
METHOD	EVALUATION METRICS	NUMBER OF CLUSTERS					
		4	5	6	7	8	9
K- MEANS	Silhouette_score :	0.35	0.38	0.40	0.37	0.37	0.37
	Calinski Score :	280139.67	256726.59	241655.03	234973.40	222261.22	209823.43
	Davies Bouldin Score :	1.08	1.01	0.93	1.10	1.18	1.21
K-Medoids	Silhouette_score :	0.50	0.43	0.40	0.37	0.31	0.30
	Calinski Score :	36570.27	31665.81	27291.48	26569.78	24059.50	21920.08
	Davies Bouldin Score :	0.72	0.98	1.00	1.12	1.22	1.24
Gaussian Mix	Silhouette_score :	0.34	0.35	0.34	0.35	0.34	0.32
	Calinski Score :	17729.93	14789.31	15518.28	18575.04	16438.19	11726.00
	Davies Bouldin Score :	1.08	2.50	2.08	1.56	1.74	2.63
Agglomerative	Silhouette_score :	0.64	0.54	0.54	0.53	0.54	0.55
	Calinski Score :	48339.18	54874.09	73017.18	80827.28	94571.96	102146.20
	Davies Bouldin Score :	0.40	0.49	0.53	0.50	0.51	0.50
SOM	Silhouette_score :	0.52	0.50	0.50	0.53	0.53	0.54
	Calinski Score :	39808.92	37599.91	36177.76	47778.00	47105.75	64813.08
	Davies Bouldin Score :	0.53	0.56	0.56	0.56	0.55	0.53

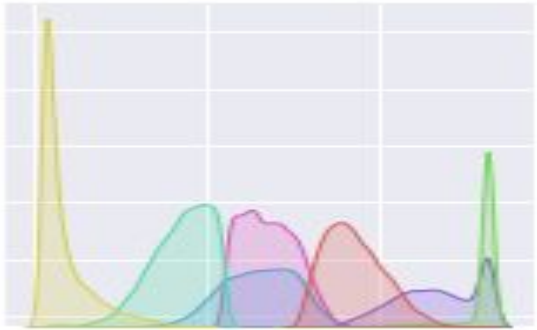
Results and Analysis

We have analyzed the pros and cons of the of baseline models using evaluation metrics (**Silhouette score, Callinkski score and Davies Bouldin score**). Based on the scores and the visualizations our interpretations are:

- K-means with number of clusters = 4 shows the best results.
- Best results: DBSCAN and SOM.
- Datasets have features with varying values therefore GMM shows poor results, cluster are not well separated.
- We are getting 1 large cluster in DBSCAN for Data2 since it has 1 or 2 overpowering features.
- No single model is best fit for all datasets.
- K-means is the best fit for Data1.
- DBSCAN and SOM is the best fit for Data2.
- Flow cytometry data is analyzed by four quadrant method and by applying gating on forward scatter, side scatter and marker expression scatter plots.
- Single parameter histograms are used to further investigate and to quantify specific populations of intere



Scatter plot



Histogram

Contribution

Himani Varolia performed literature survey, dataset finalization, applied two clustering methods : K-means and DBSCAN and analysis of results over the two open source dataset,evaluation metrics, visualizations and detailed analysis, and documentation.

Anuradha Kumari performed two clustering methods : kmediods and Gaussian mixture models,evaluation metrics, documentation, visualizations and analysis of result.

Kunal performed two clustering methods: Agglomerative clustering and SOM (Self Organizing Maps),evaluation metrics, documentation and visualization.

Thank you