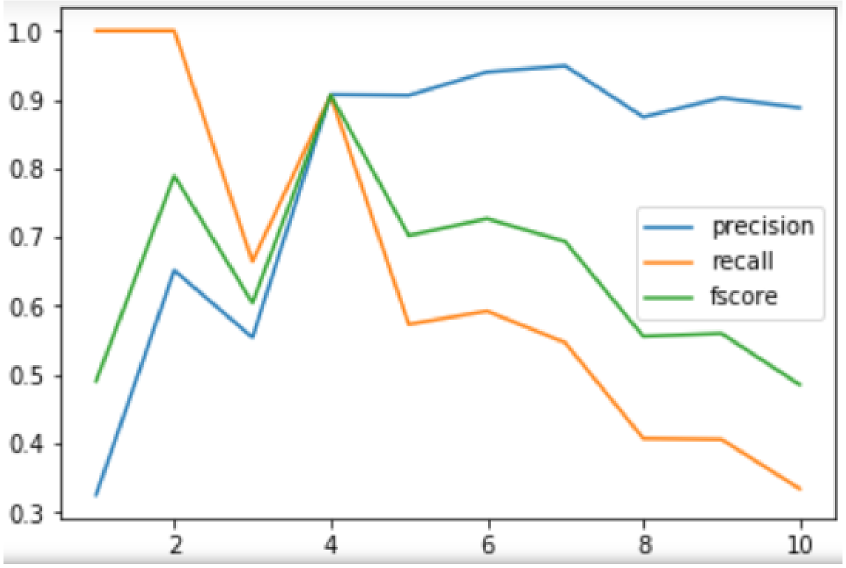
Assignment 4:

Data Set:

Here data is in 4 files animals, veggies, fruits and countries. We have to merge those 4 files into one file and then import data.

Preprocessing for normalized data: Make all the columns normalized except first column.

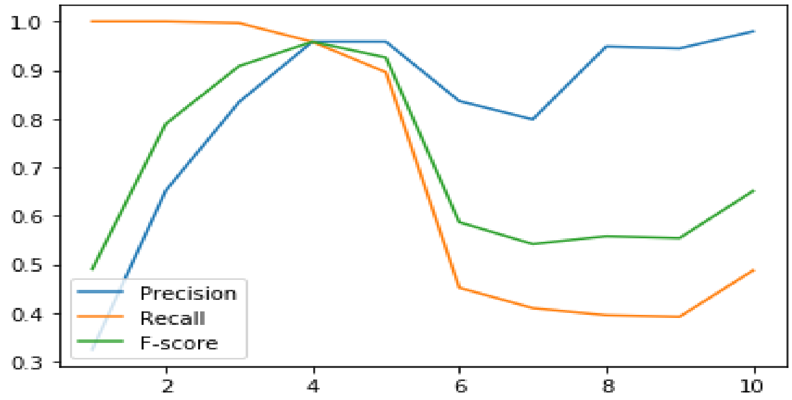
Q1:] K means using Eucledian distance:



Here in above diagram we can see that as number of cluster increases value of recall decreases, As number of cluster increases value of precision also increases. But we can see mix behavior for fscore.

Here we can also see that there is intersection of precision, recall and fscore at 4. Therefore this algorithm performs good when there are 4 clusters. At k=4 value of precision, recall and fscore is 0.923.

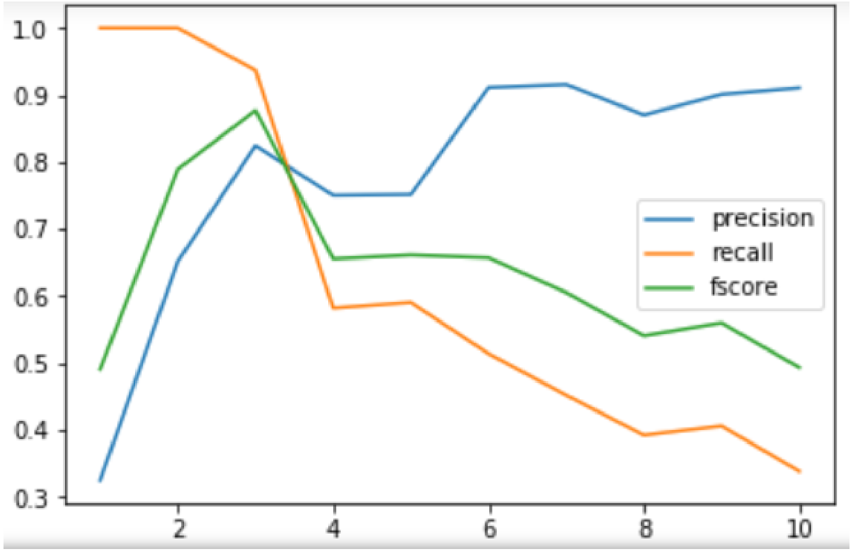
Q2:]K means on Normalized data:



Here in above diagram we can see that as number of cluster increases value of recall decreases, As number of cluster increases value of precision also increases but after k=4 we can see some fall in precision. But we can see mix behavior for fscore.

Here we can also see that there is intersection of precision, recall and fscore at 4. Therefore this algorithm performs good when there are 4 clusters. At k=4 value of precision, recall and fscore is 0.932.

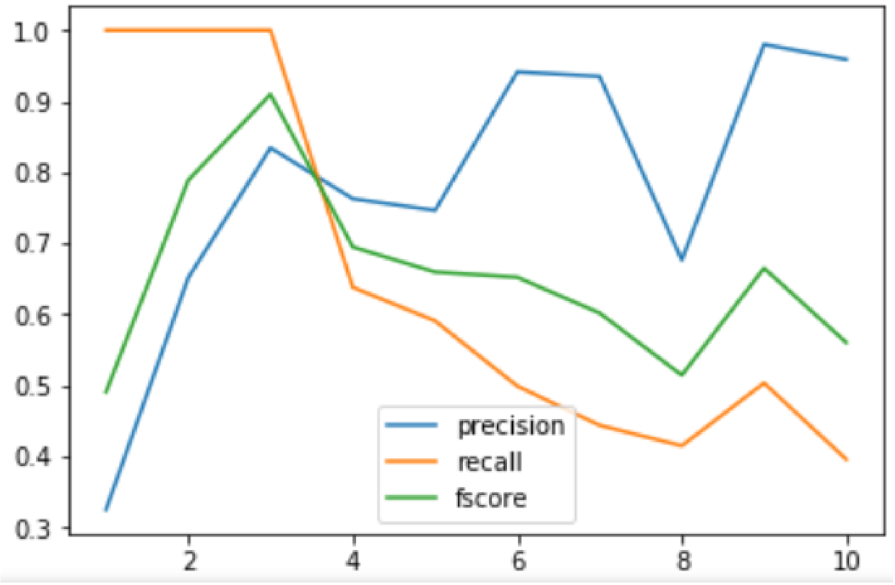
Q3:]K means using Manhattan distance



Here in above diagram we can see that as number of cluster increases value of recall decreases, As number of cluster increases value of precision also increases but after k=4 we can see some fall in precision. But we can see mix behavior for fscore.

Here we can also see that there is intersection of precision, recall and fscore at 4. Therefore this algorithm performs good when there are 4 clusters. At k=4 value of precision, recall and fscore is 0.802.

Q4:]K means by Cosine Similarity



Here in above diagram we can see that as number of cluster increases value of recall decreases, As number of cluster increases value of precision also increases but after k=4 we can see some fall in precision. But we can see mix behavior for fscore.

Here we can also see that there is intersection of precision, recall and fscore at 4. Therefore this algorithm performs good when there are 4 clusters. At k=4 value of precision, recall and fscore is 0.821.

Conclusion:

For clusters size = 4. In above all graphs we can see that algorithm works good on normalized data as precision recall and fscore intersect at 0.932 which is high than all other methods. After that it works good on k means by **eucledian** distance. Algorithm gives some what equal performance on cosine similarity and manhattan distance. But as we can see overall K means by Eucledian distance is good because after k = 4 other methods show slight unpredictable behavior for precision as there is rise and fall in precision value. But for k means with eucledian distance there is no rise and fall in precision value.

But in all methods precision recall and fscore intersect at k=4. Therefore best setting for k means clustering for this dataset is **4**.