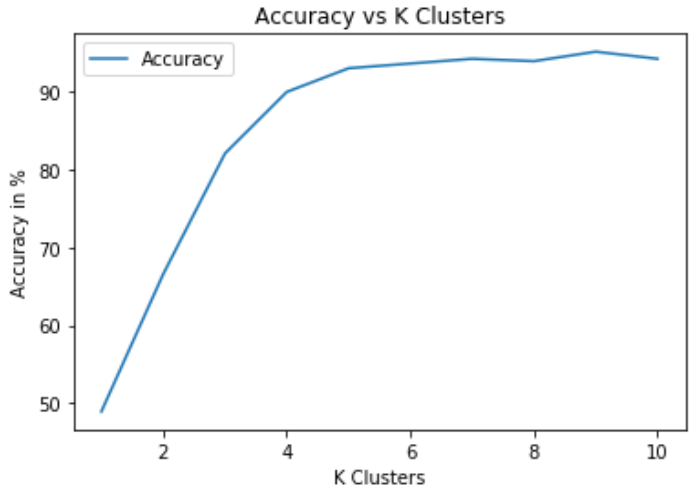
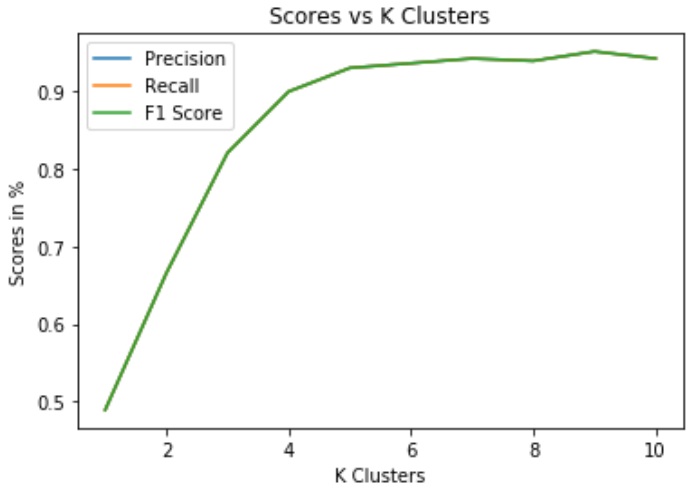
Data Mining

K-Mean Clustering

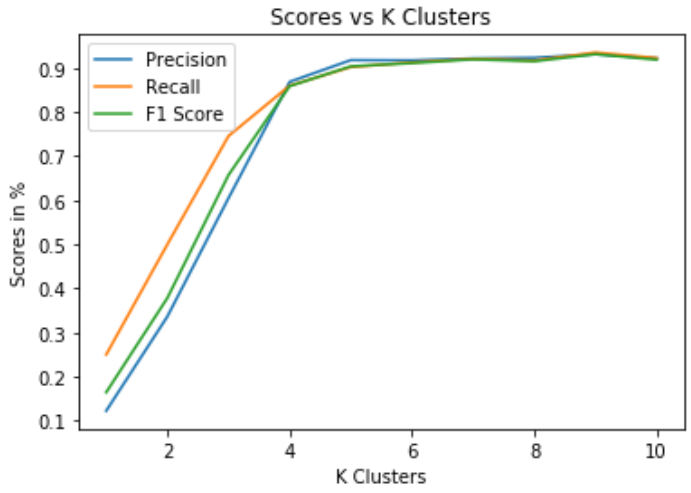
Name Kaustav Vats
Roll No 2016048

Q2- **K Mean Clustering** using **Euclidean Distance** over **Unnormalized Feature** values

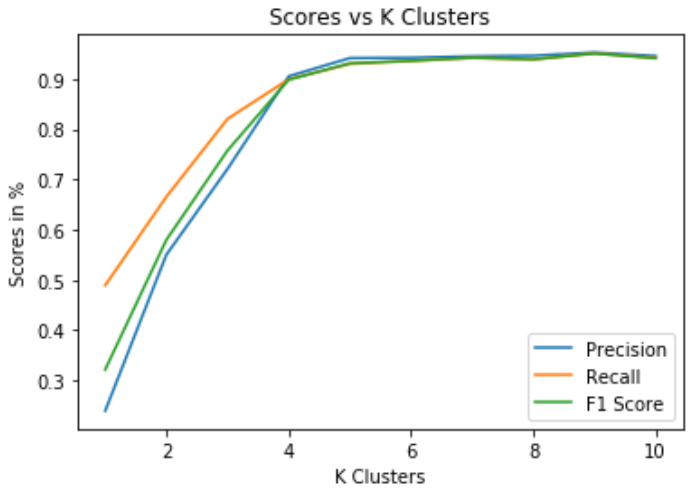




Average = micro



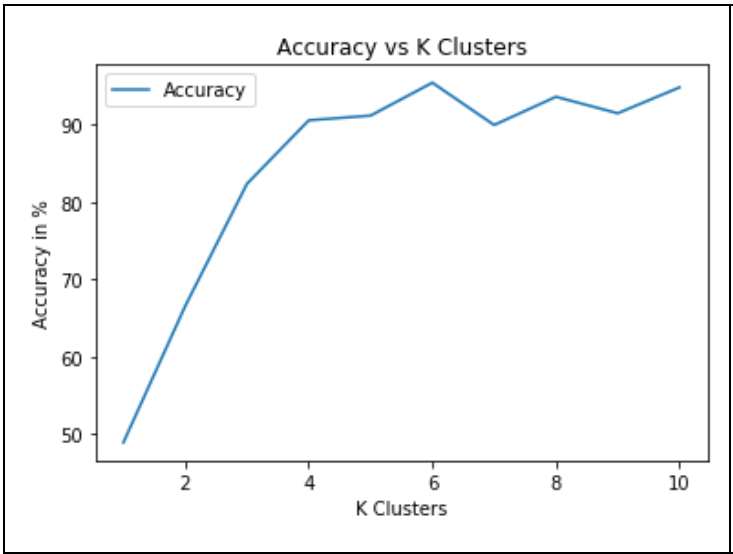
Average = macro

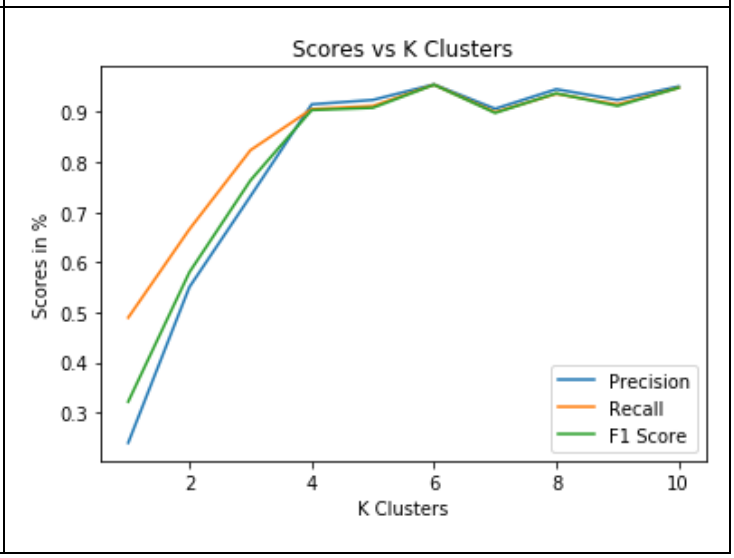


Average = weighted

Best K=9
Considering precision , recall and F1 score. All are high for k9 as compared to other k values.

Q3 - **K Mean Clustering** using **Euclidean Distance** over **Normalized Feature** values

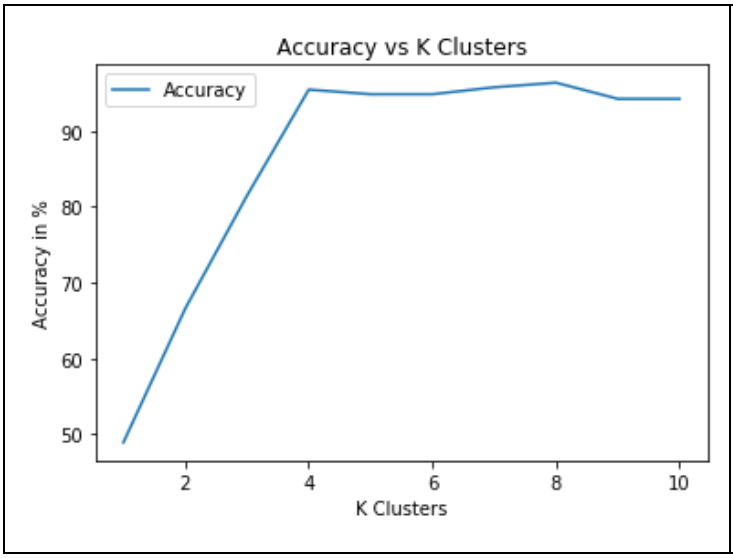




Best K=6

Considering precision , recall and F1 score. All are high for k6 as compared to other k values.

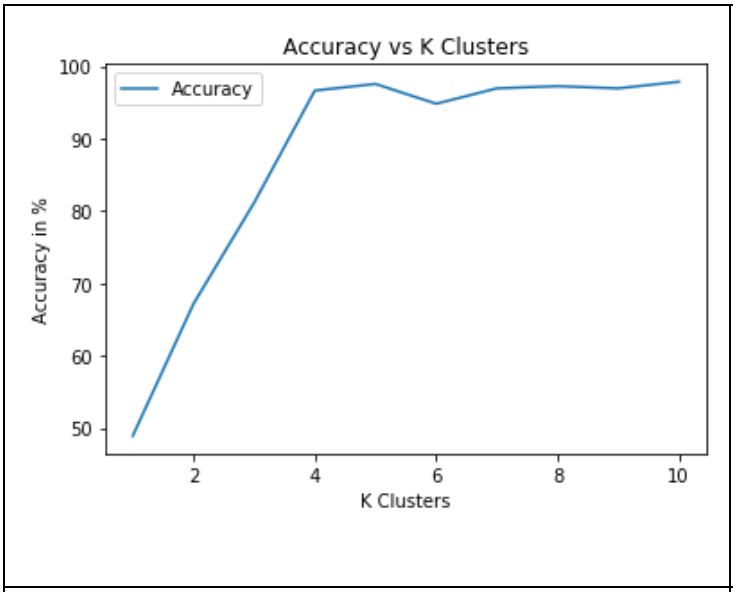
Q4 - **K Mean Clustering** using **Manhattan Distance** over **Unnormalized Feature** values

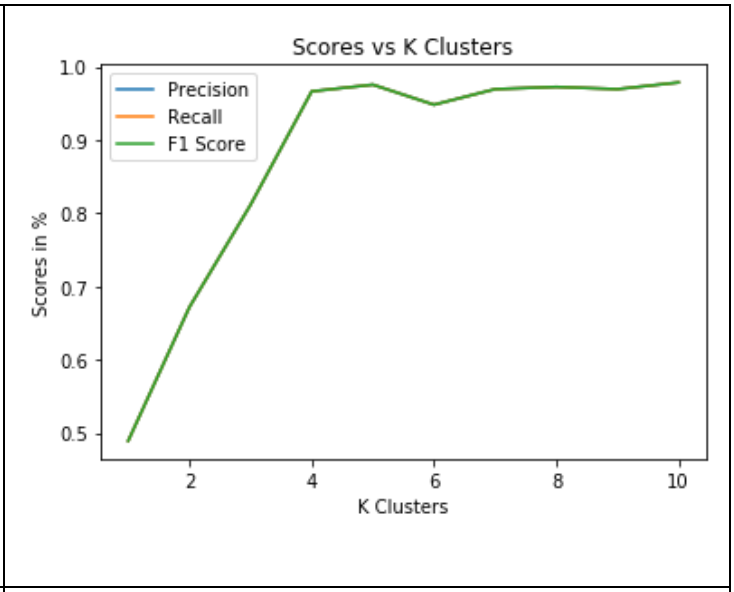


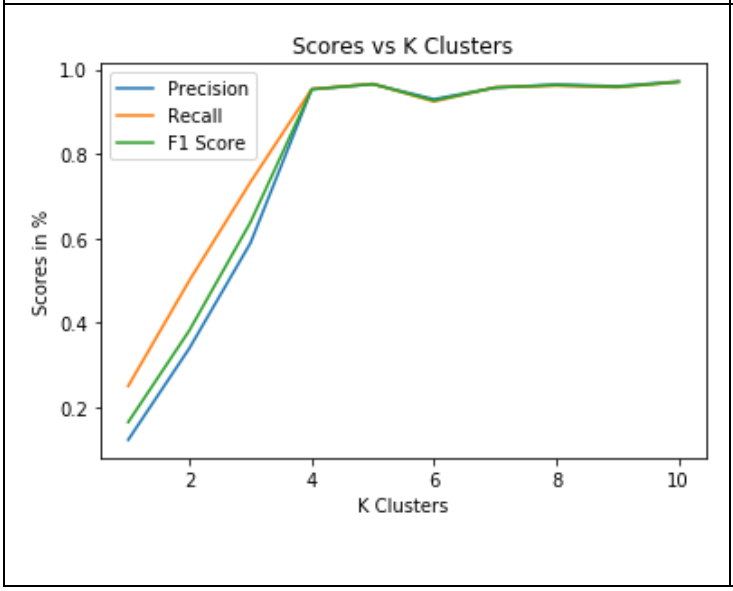
Best K=4 or k=8

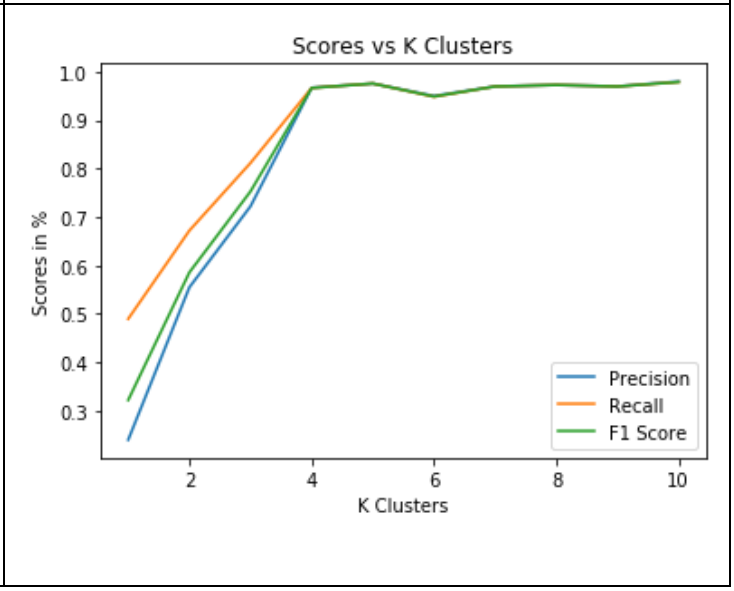
Considering precision , recall and F1 score. All are high for k4 and k8 as compared to other k values.

Q5 - **K Mean Clustering** using **Cosine Similarity** over **Unnormalized Feature** values









Best K=5, Considering precision , recall and F1 score. All are high for k5 as compared to other k values.

Overall it seems like **Manhattan Distance/Cosine Similarity** is the best distance metric with optimal K=5 value over unnormalized data. For normalized features i observed lot of spikes in precision recall f1 score and also the results are not competitive as compared to other metrics.