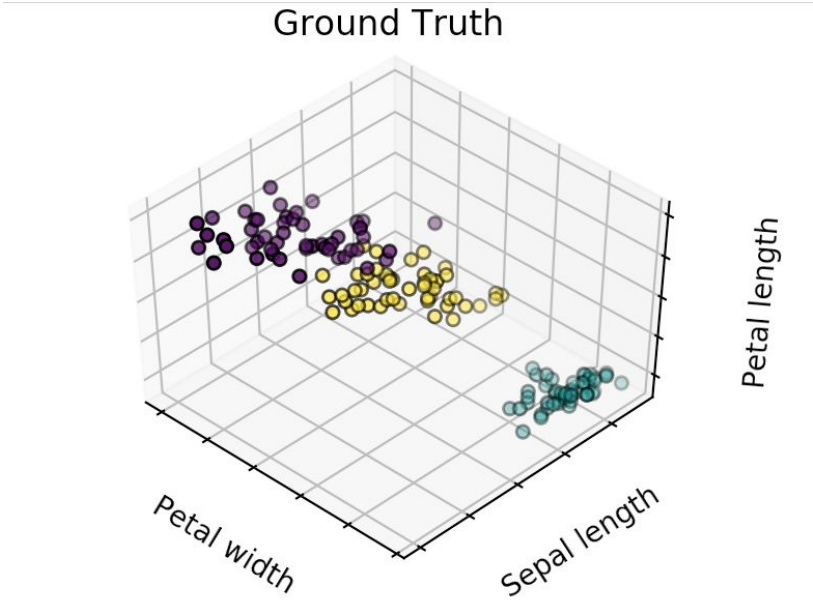
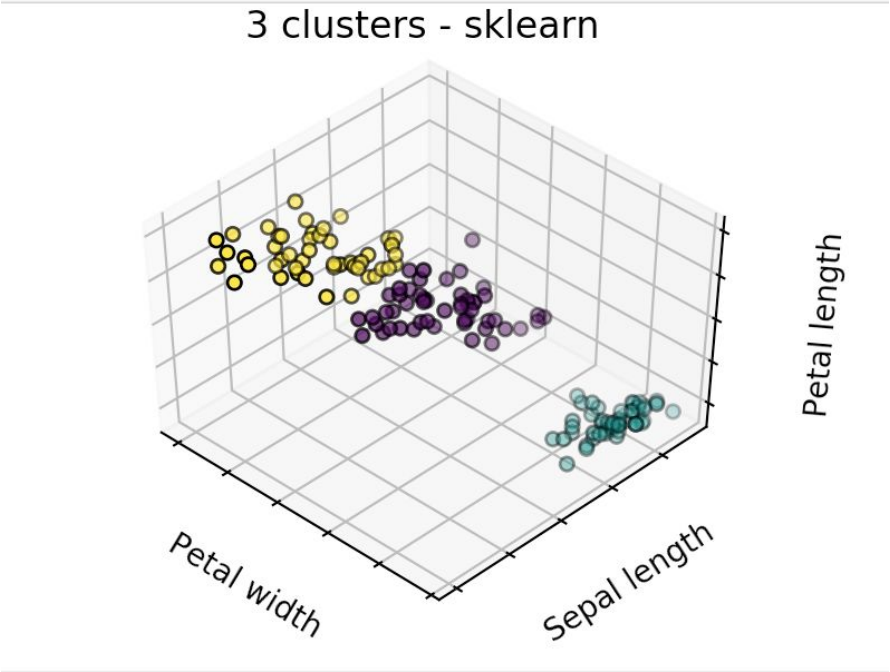
Raj Pokhrel
CMPSC 497
Project 3

K-Means Programming Assignment

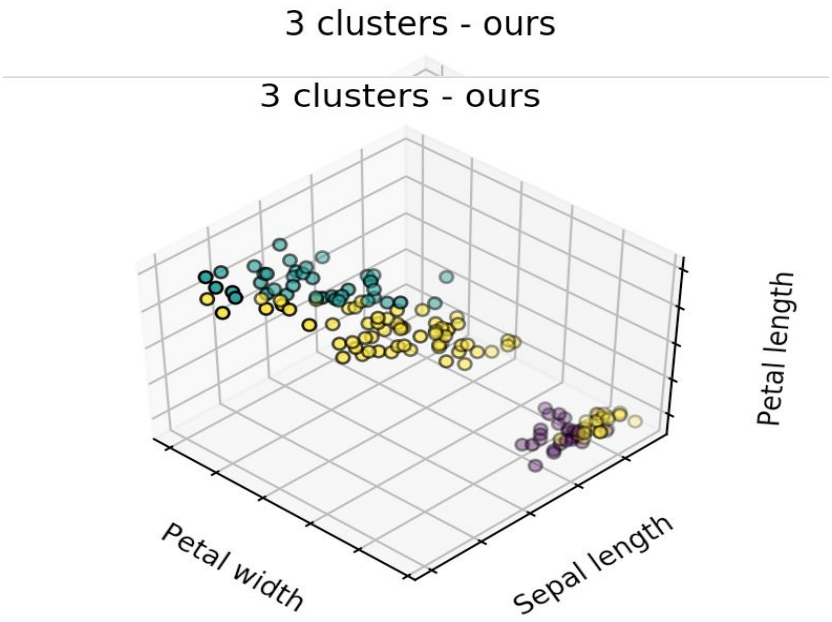
Plot of Ground Truth and Sklearn



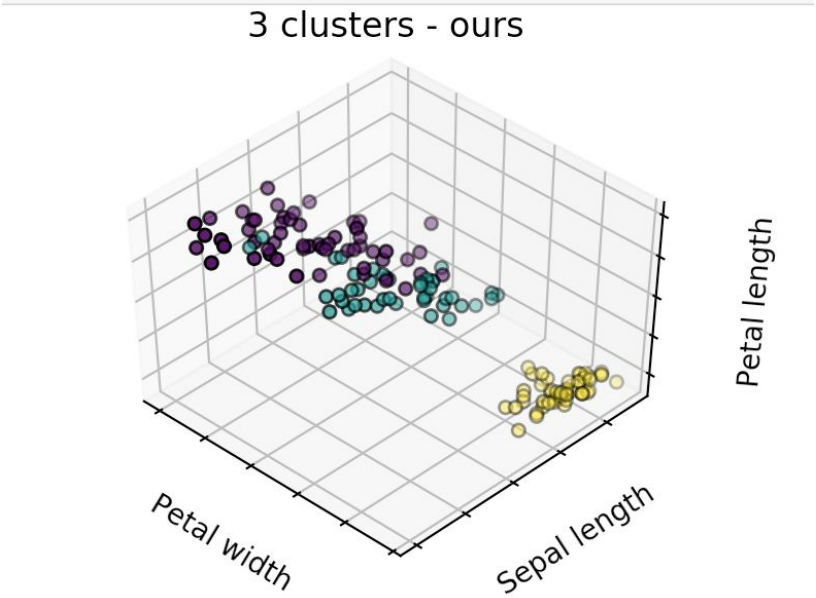


Task 1

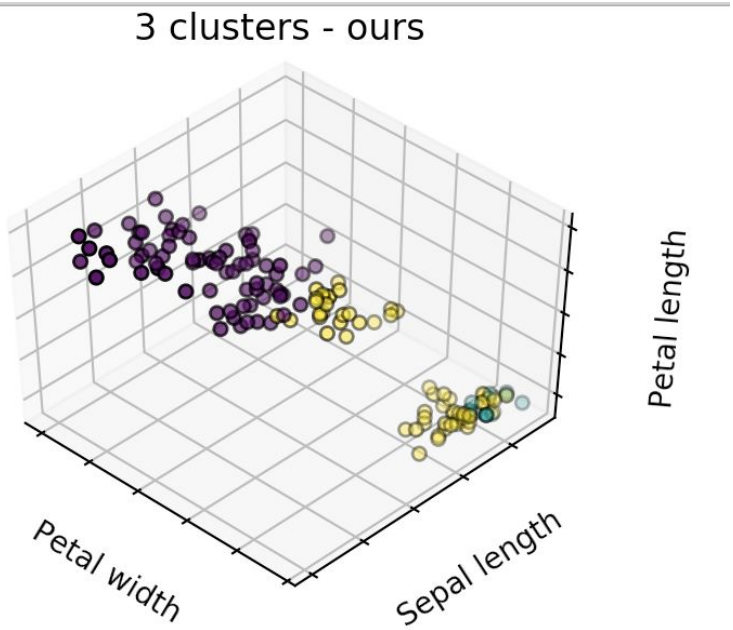
- The plots below represent the best three clusters that I was able to obtain using the K-means



Gini of Majority Vote Label 0 is 0.88
Gini of Label 1 is 0.88
Gini of Label 2 is 0.88



Gini of Majority Vote Label 0 is 0.86
Gini of Label 1 is 0.866
Gini of Label 2 is 0.86



Gini of Majority Vote Label 0 is 0.80
Gini of Label 1 is 0.80
Gini of Label 2 is 0.80

- The skit-learn clusters had more precise cluster than my clusters because the different label was clustered properly as you can see on the graph above.

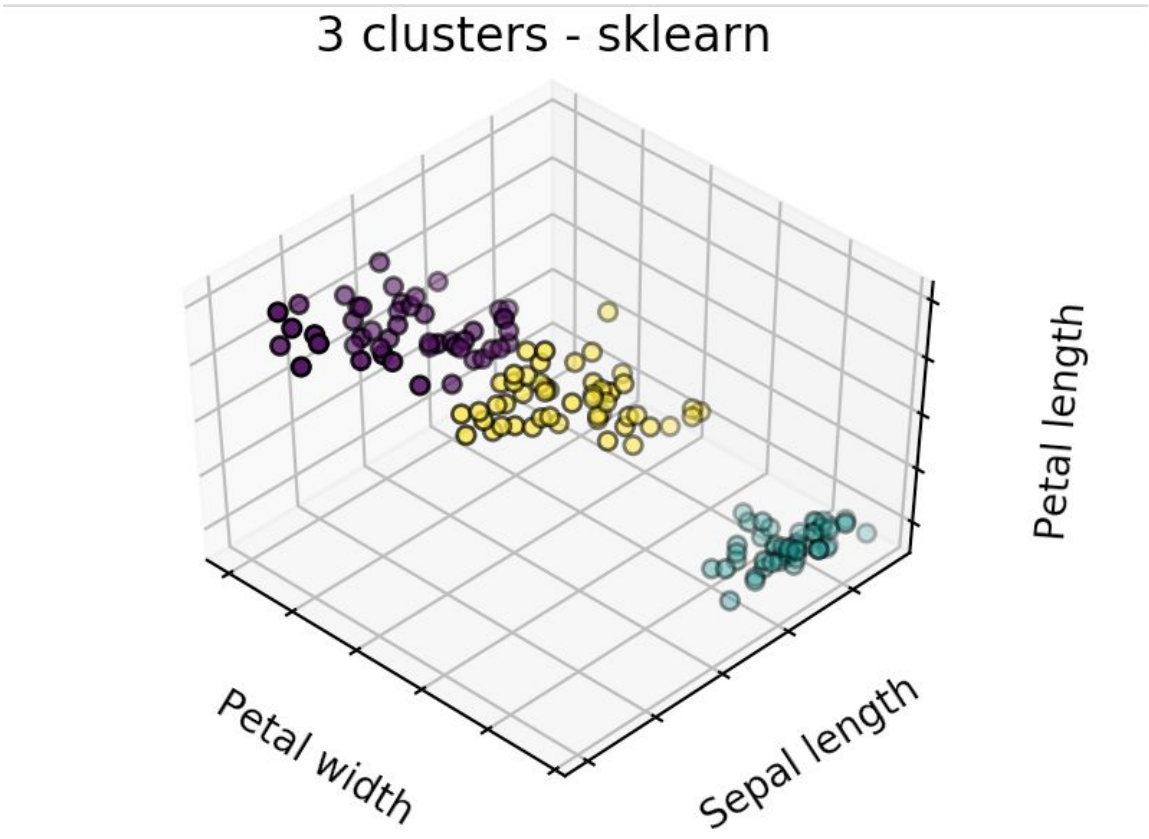
Task 2

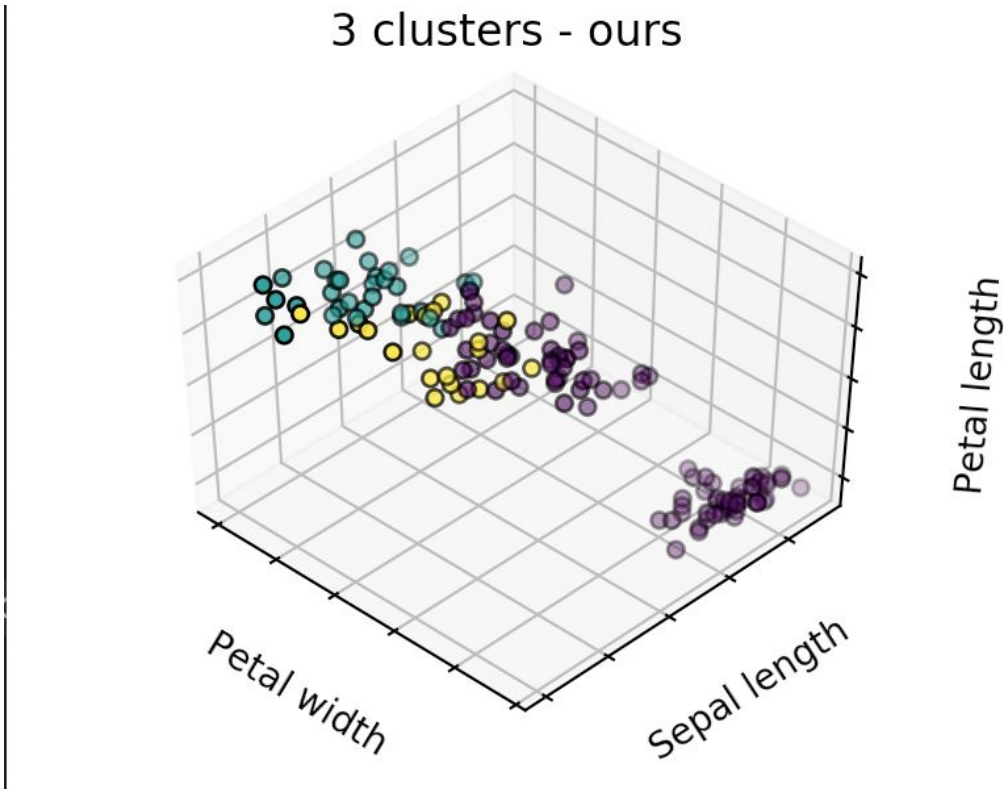
- It's very important to choose the proper initial centroids for K-means because the algorithm will find the suboptimal solution when the center is chosen incorrectly and different initial centroids can give you different clusters. Bad initialization may end up getting bad clusters. However, we can pick the initial centroids close to the final centroid or simply choose random point initially and run the K-means several times to pick the best cluster.

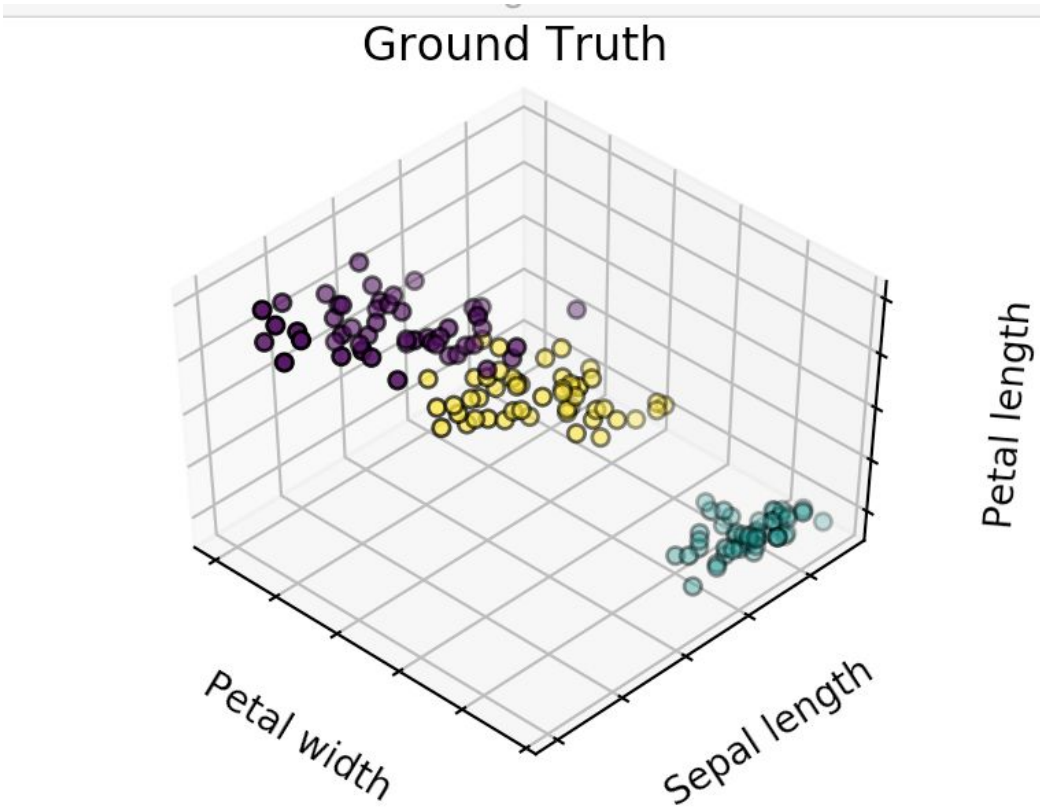
Task 3

- The scikit-learn cluster was slightly better than my own clusters in term of Gini index which the clusters as you can see above. The Gini index below is the best cluster of K-mean I was able to obtain.

  Gini of Majority Vote Label 0 is 0.88
  Gini of Label 1 is 0.88
  Gini of Label 2 is 0.88

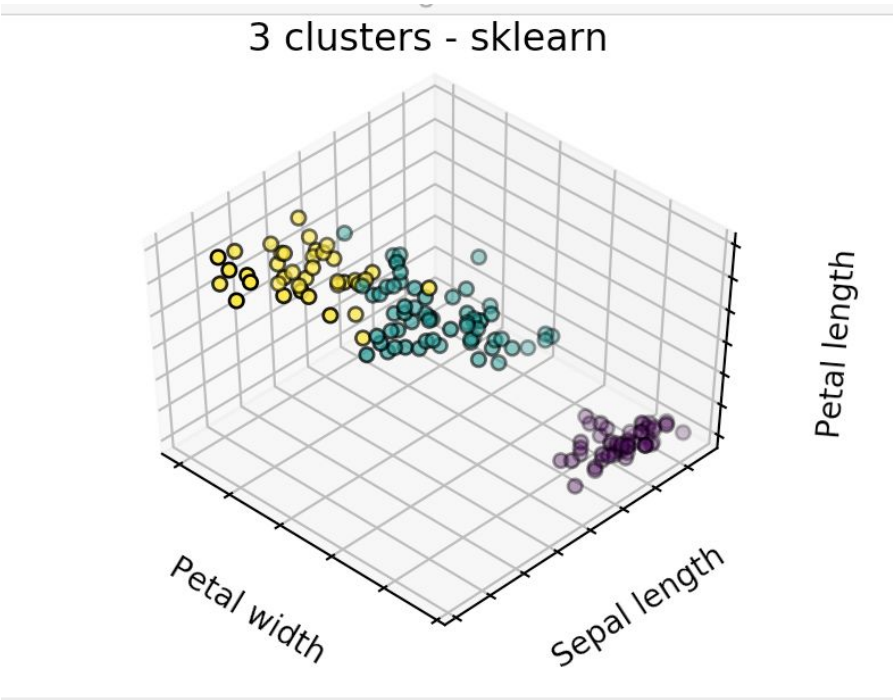



3 clusters - ours
Petal width
Sepal length
Petal length

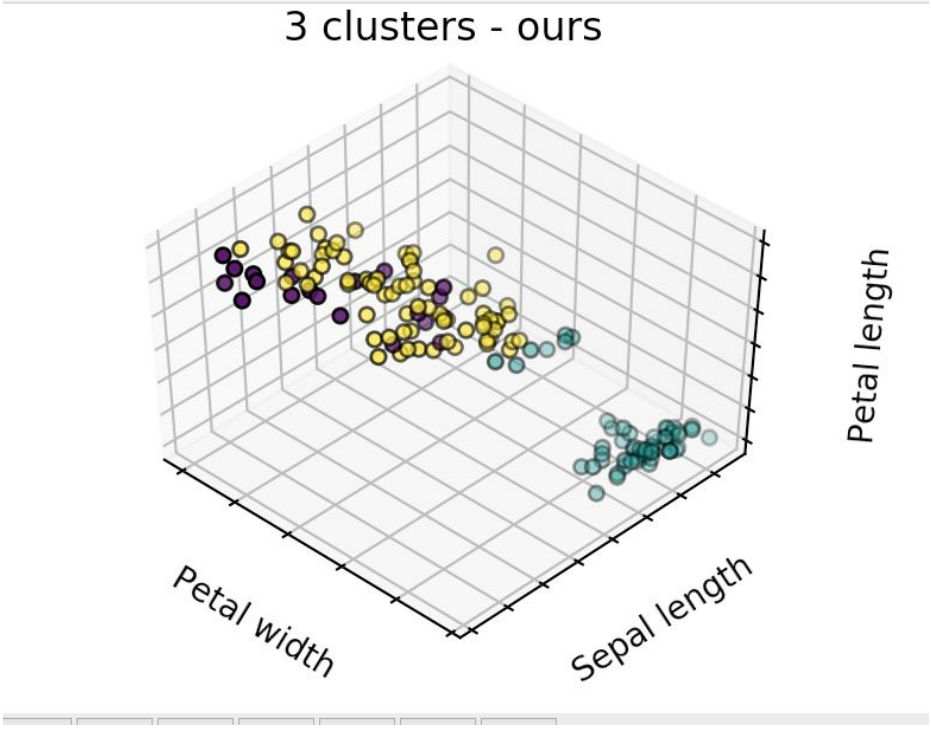

Ground Truth
Petal width
Sepal length
Petal length

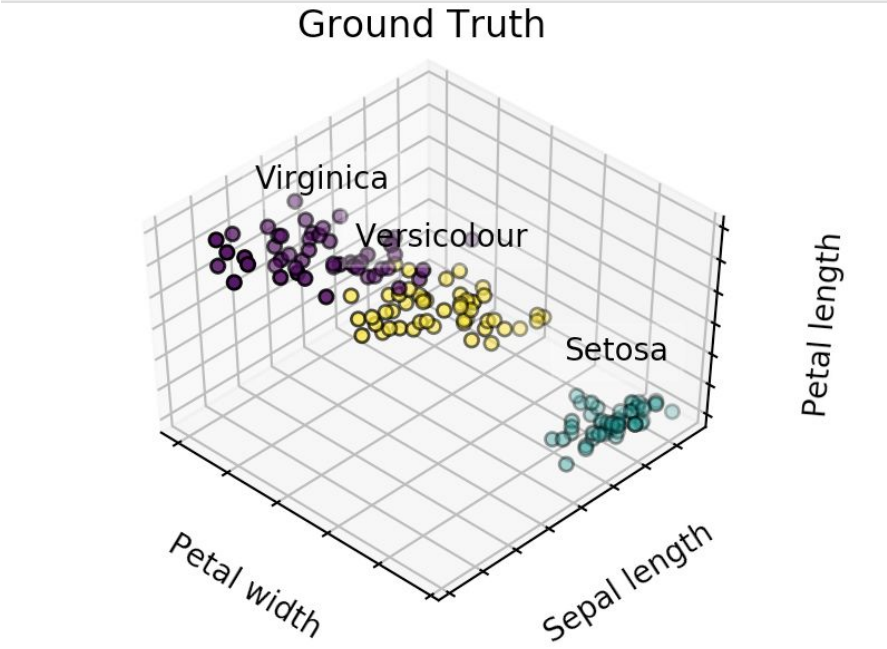
Task 4

- By normalizing we remove any bias in the datasets and the results will be very different compared to the raw data. Normalization gives the same importance to all the features in the datasets. In addition, normalized datasets give us better results as you can see in the plots above with normalize datasets and below without normalizing datasets. The Gini index for the normalized values was about 0. 88 and the unnormalized was about 0.69 which means normalizing the datasets will give us better results. The number of iterating to converge for me 1000 for both unnormalized and normalized datasets.



Gini of Majority Vote Label 0 is 0.88
Gini of Label 1 is 0.88
Gini of Label 2 is 0.88


3 clusters - ours
Petal width
Sepal length
Petal length


Ground Truth
Virginica
Versicolour
Setosa
Petal width
Sepal length
Petal length

Task 5

- The Impurity tells us about the percentage of data in a cluster that belongs to it's most frequent class. Therefore, when working with the large datasets impurity can be a very useful measure that might help us understand what the data better. From example, if the Gini index for one cluster is higher than the other one, we can same that the higher Gini index has better results.