IT402: Soft Computing

Lab Assignment 1 and 2

K-means clustering and Fuzzy C-means clustering

Name: Chinmayi C. Ramakrishna **Date of Submission:** 28th January, 2021

Roll No.: 181IT113

K-means Clustering

It is an unsupervised machine learning concept. It's a partitioning method to form k clusters based on similarity of data points.

Algorithm

1. Arbitrarily choose k data points from the dataset as initial centroids.
2. Repeats :
    - Assign each data point in the dataset to the most similar cluster evaluated using the mean value of data points in the cluster.
    - Update cluster means.
    - Repeat until no change in data points and centroids.

Dataset 1: IRIS dataset

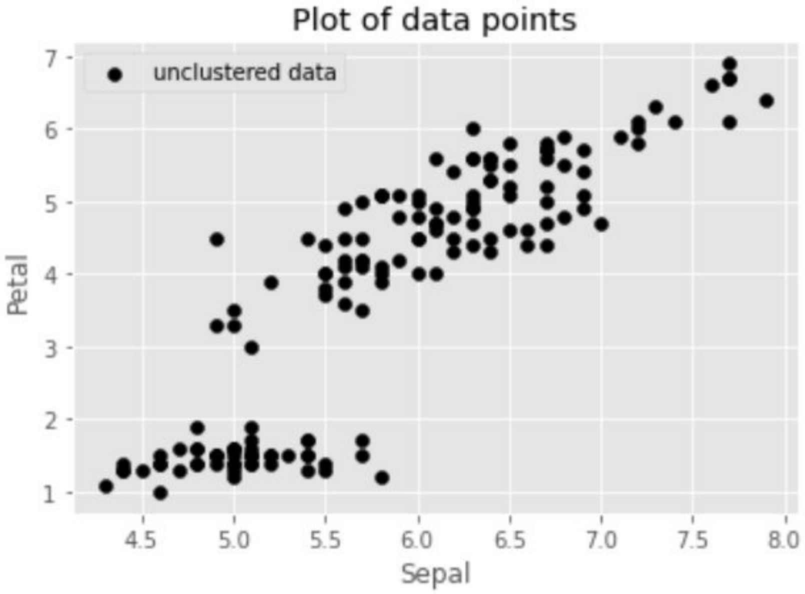


Fig 1. Petal length vs Sepal length unclustered data

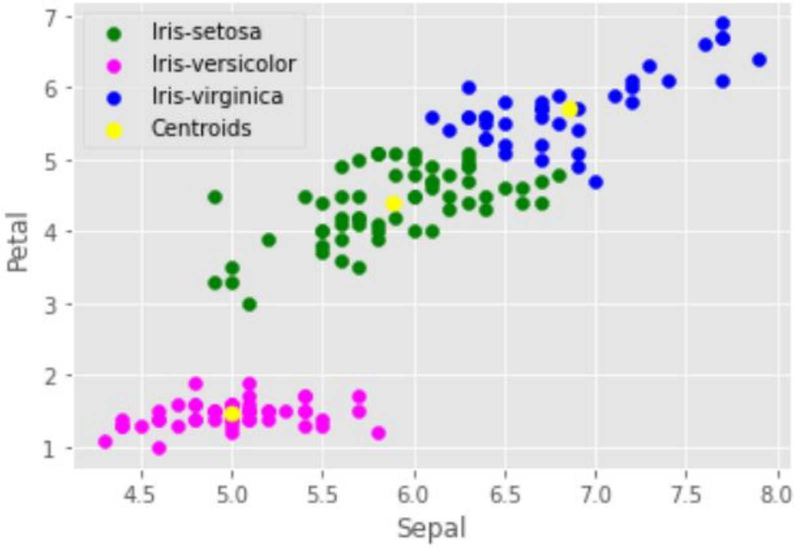
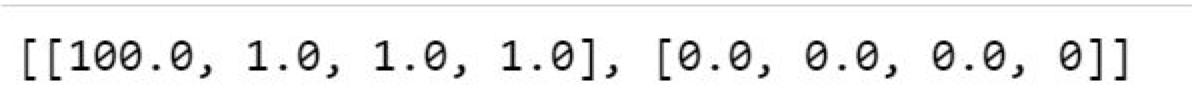


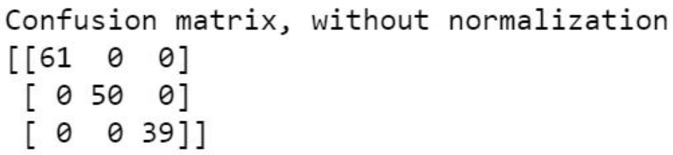
Fig 2. Petal length vs Sepal length clustered data



```
[[100.0, 1.0, 1.0, 1.0], [0.0, 0.0, 0.0, 0]]
```

Fig 3. Accuracy Results: [Accuracy in %, Precision, Recall, F1 Score]

The results show a 100% accuracy in clustering data. All the four data points have been used for clustering the data. Since the data points are non overlapping and distinct, K-means clustering gives the desired results.



```
Confusion matrix, without normalization
[[61  0  0]
 [ 0 50  0]
 [ 0  0 39]]
```

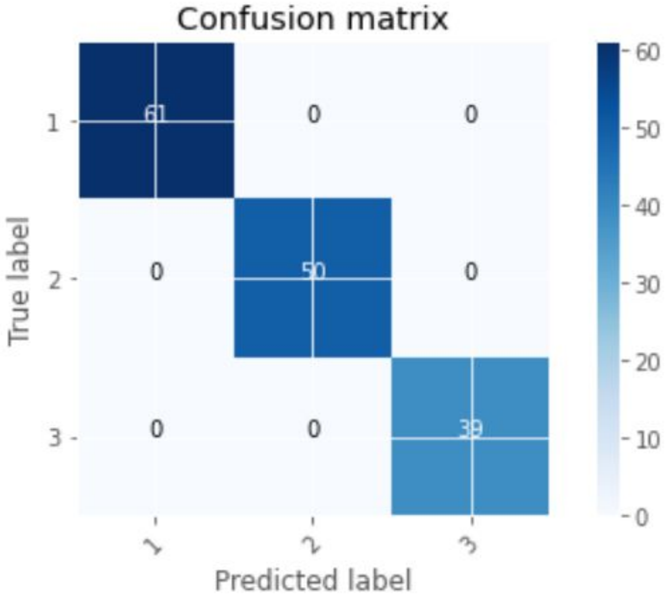


Fig 4. Confusion Matrix for K-means clustering

There are no false positives and false negatives as can be observed from the confusion matrix.

Dataset 2: SPECT_F dataset

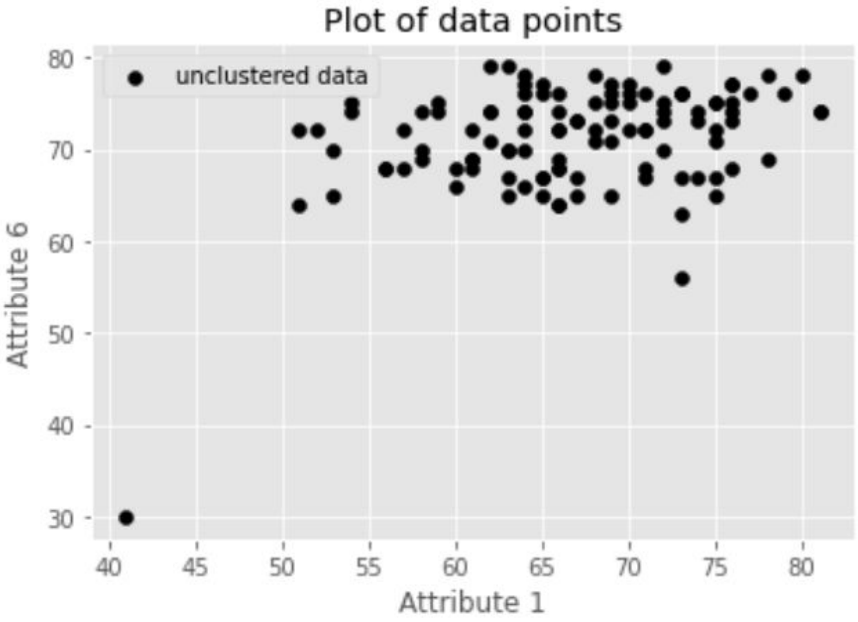


Fig 5. Attribute 6 vs Attribute 1 unclustered data

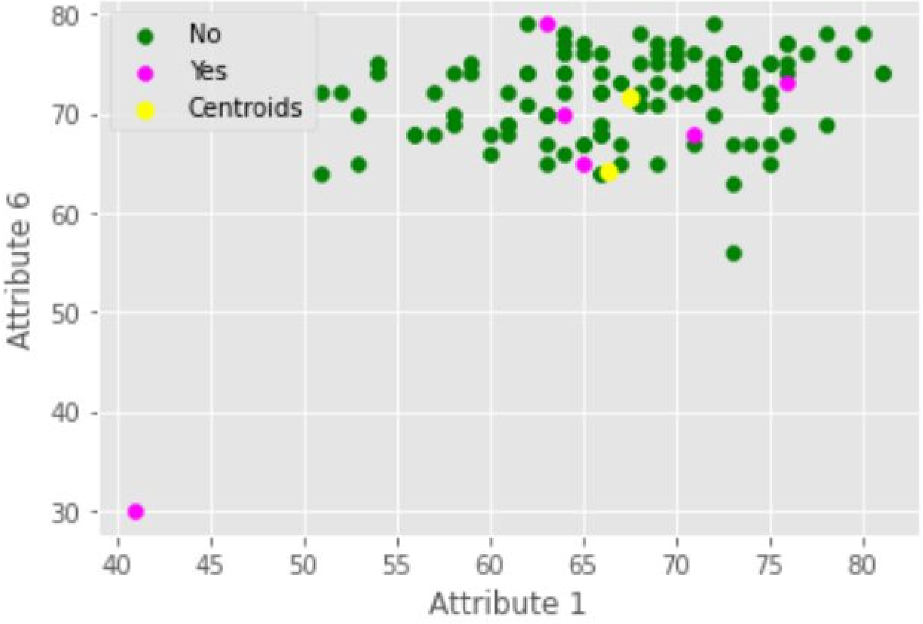
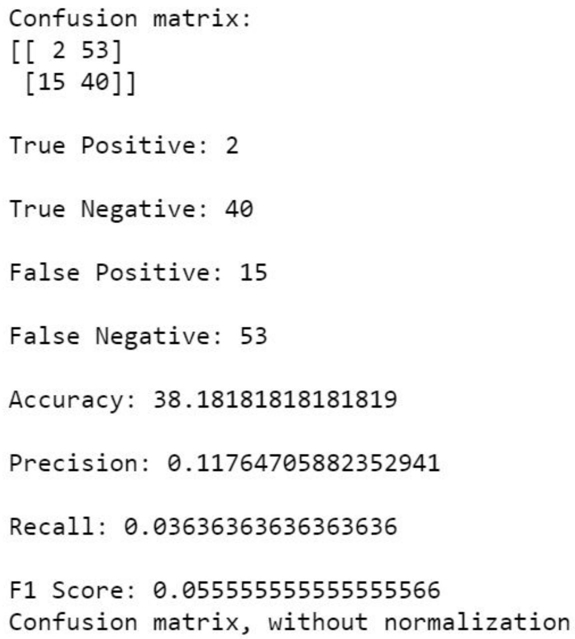


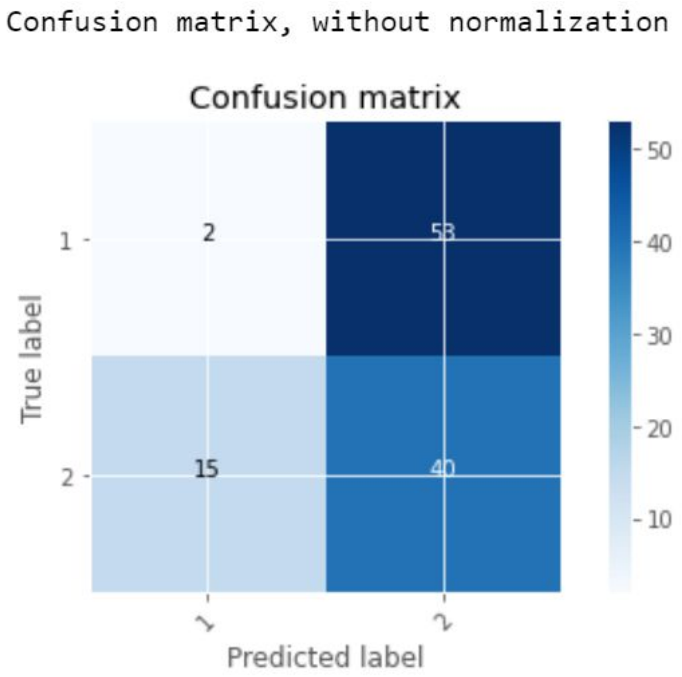
Fig 6. Attribute 6 vs Attribute 1 clustered data



```
Confusion matrix:
[[ 2 53]
 [15 40]]

True Positive: 2

True Negative: 40

False Positive: 15

False Negative: 53

Accuracy: 38.18181818181819

Precision: 0.11764705882352941

Recall: 0.03636363636363636

F1 Score: 0.05555555555555566
Confusion matrix, without normalization
```

Fig 7. Accuracy Results: [Accuracy in %, Precision, Recall, F1 Score]

The results show an accuracy of 38.18%. All the 44 attributes have been used to cluster the data points. Since the data points are overlapping, K-means clustering doesn’t provide the best performance matrix.



```
Confusion matrix, without normalization
```

Fig 8. Confusion matrix for Fuzzy C-means clustering

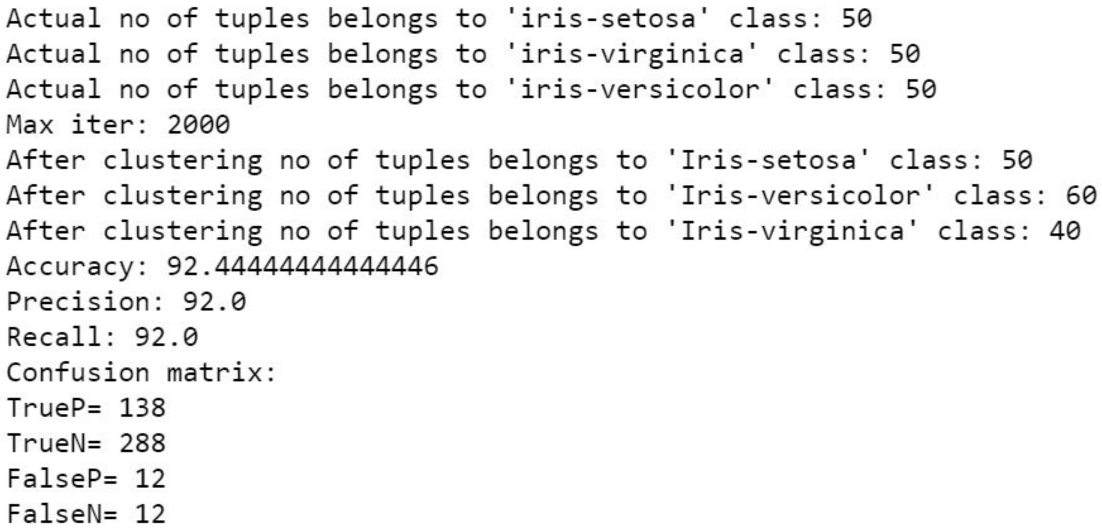
Fuzzy C-means Clustering

Fuzzy C-means is a technique that can be used on inseparable data points. Data points can belong to more than one cluster in this method.

Algorithm

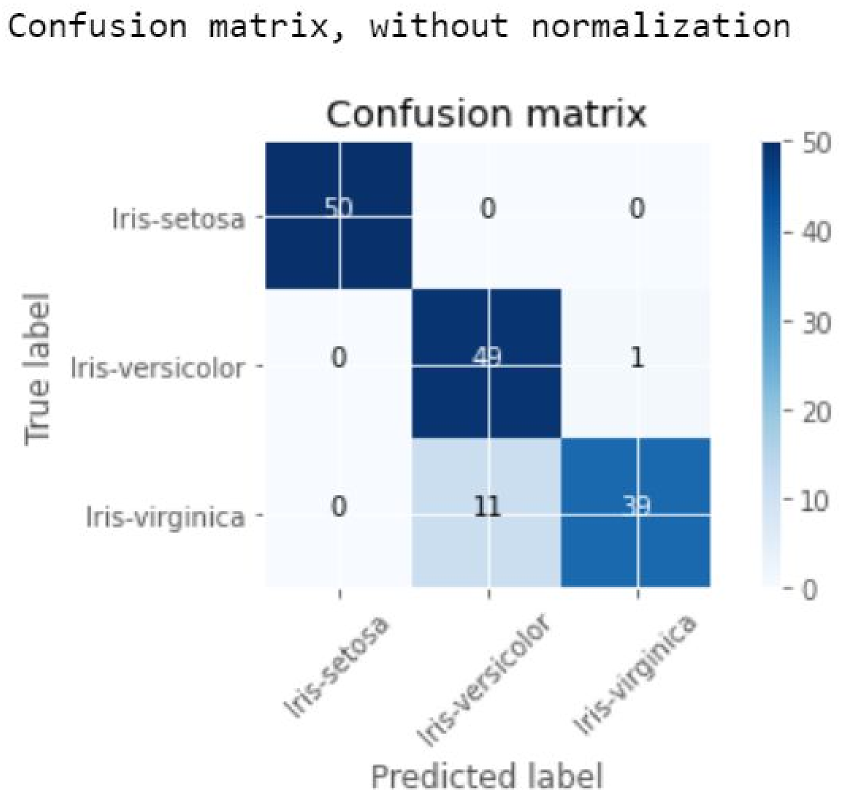
1. Randomly initialise membership matrix U with c clusters.
2. Repeat:
    - Calculate C_i.
    - Compute dissimilarity between centroids and data points.
    - Compute a new U.
3. Repeat until improvement over the previous iteration is below convergence threshold t.

Dataset 1: IRIS dataset



```
Actual no of tuples belongs to 'iris-setosa' class: 50
Actual no of tuples belongs to 'iris-virginica' class: 50
Actual no of tuples belongs to 'iris-versicolor' class: 50
Max iter: 2000
After clustering no of tuples belongs to 'Iris-setosa' class: 50
After clustering no of tuples belongs to 'Iris-versicolor' class: 60
After clustering no of tuples belongs to 'Iris-virginica' class: 40
Accuracy: 92.44444444444446
Precision: 92.0
Recall: 92.0
Confusion matrix:
TrueP= 138
TrueN= 288
FalseP= 12
FalseN= 12
```

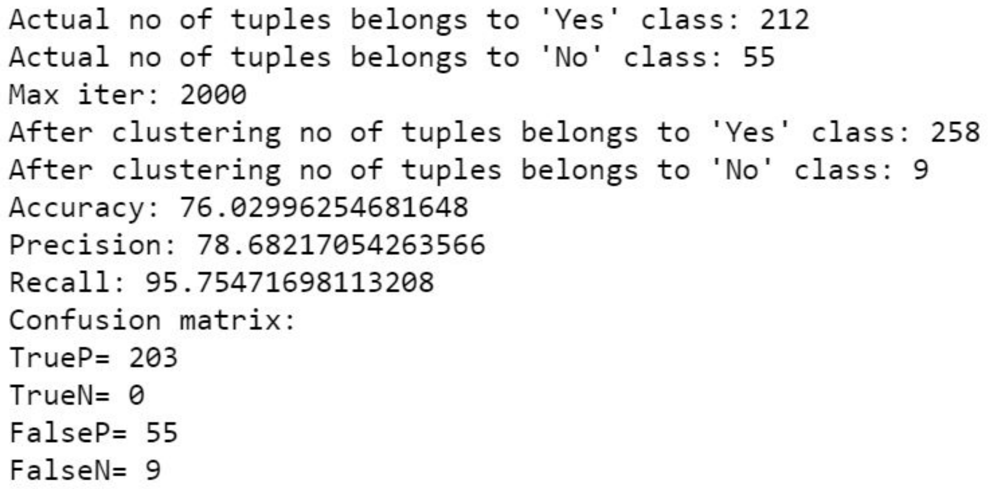
Fig 9. Results for Fuzzy C-means clustering



```
Confusion matrix, without normalization
```

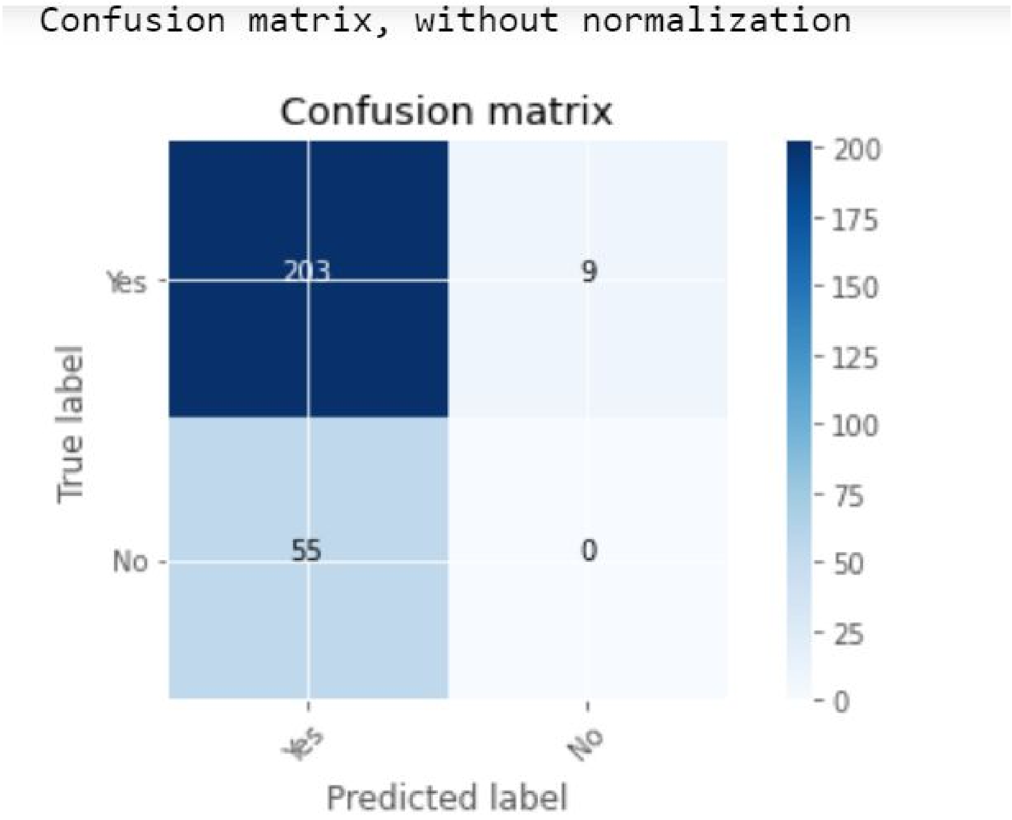
Fig 10. Confusion matrix for Fuzzy C-means clustering

Dataset 2: SPECT_F dataset



```
Actual no of tuples belongs to 'Yes' class: 212
Actual no of tuples belongs to 'No' class: 55
Max iter: 2000
After clustering no of tuples belongs to 'Yes' class: 258
After clustering no of tuples belongs to 'No' class: 9
Accuracy: 76.02996254681648
Precision: 78.68217054263566
Recall: 95.75471698113208
Confusion matrix:
TrueP= 203
TrueN= 0
FalseP= 55
FalseN= 9
```

Fig 11.Results



```
Confusion matrix, without normalization
```

	Yes	No
Yes	203	9
No	55	0

Fig .Confusion matrix for Fuzzy C-means clustering