

Anagha Sankar

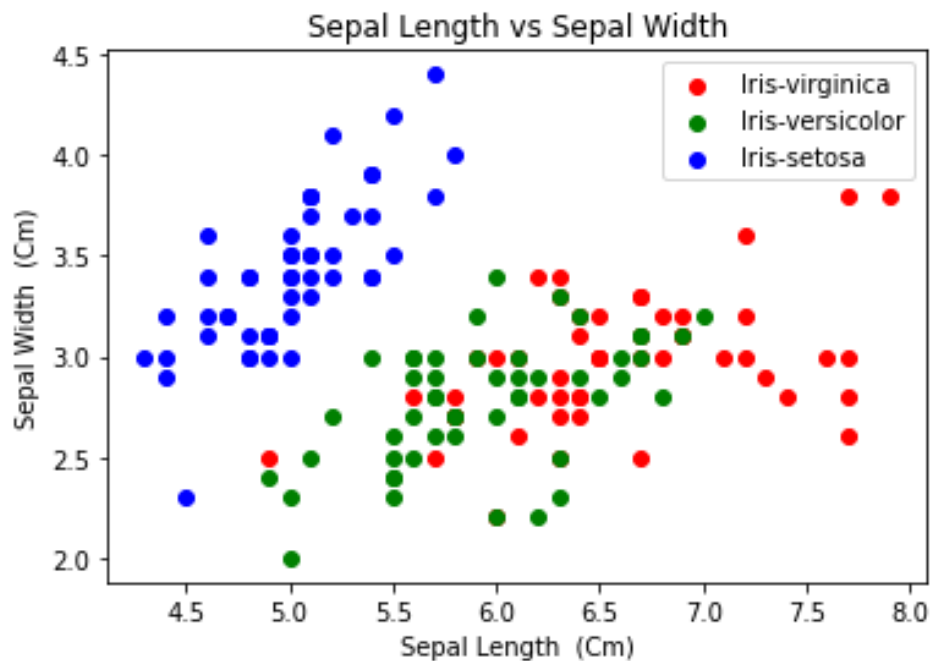
221CS2294

Mtech Information Security

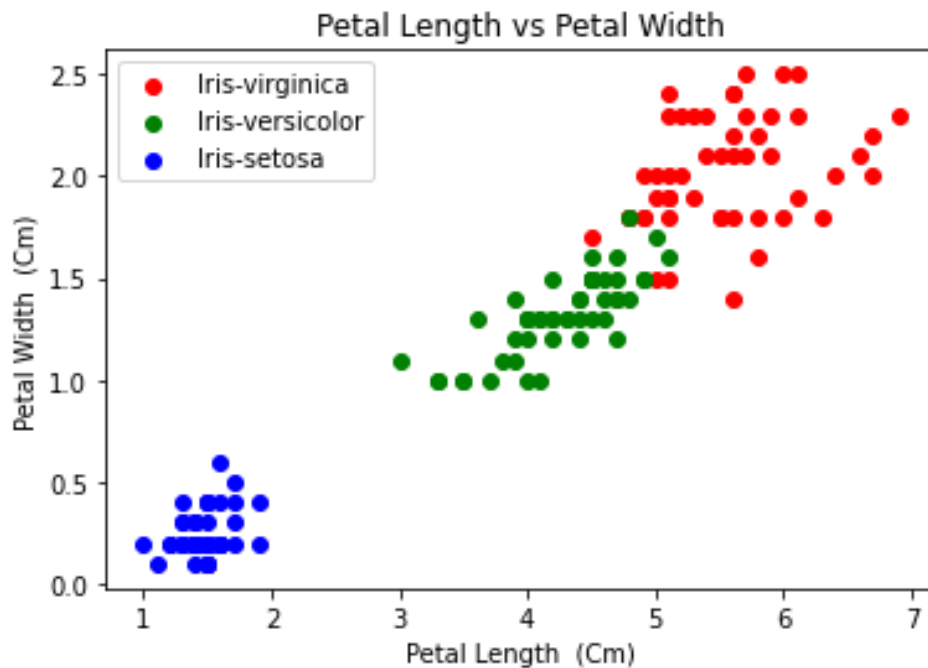
National Institute of Technology, Rourkela

DISTANCE BASED CLASSIFIER

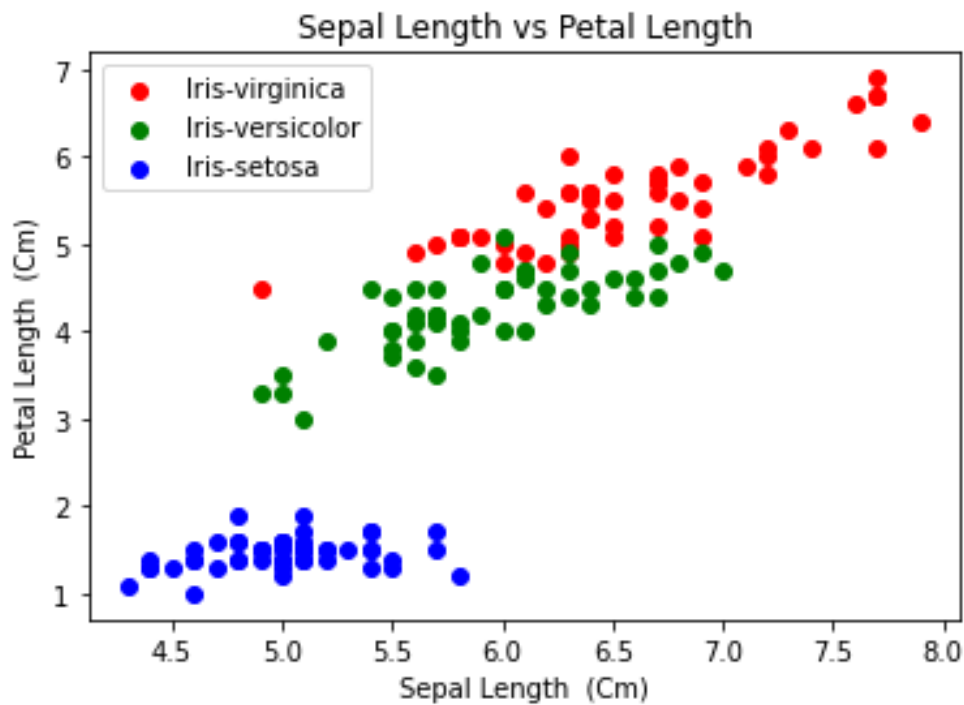
DATA ANALYSIS



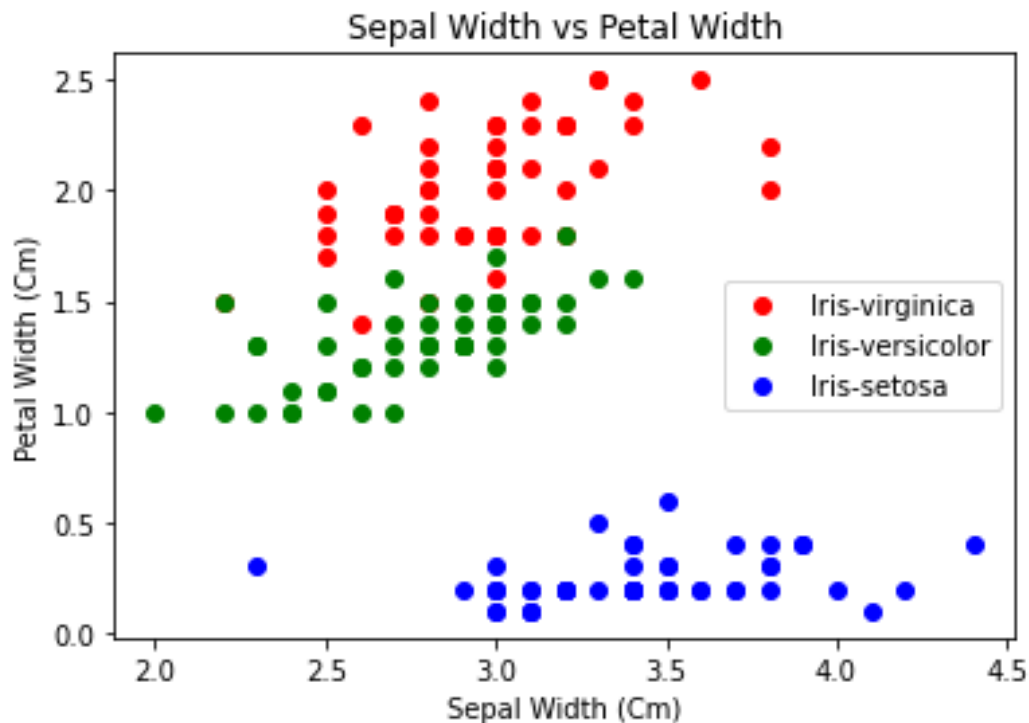
Comment: The classes Iris-virginica and Iris-Versicolor are almost inseparable.



Comment: The classes Iris-virginica and Iris-Versicolor are almost separable, whereas the class Iris-setosa is totally distinct from the other two classes.



Comment: The classes Iris-virginica and Iris-Versicolor has very little overlapping, whereas the class Iris-setosa is totally distinct from the other two classes.



Comment: The classes Iris-virginica and Iris-Versicolor are almost separable, whereas the class Iris-setosa is totally distinct from the other two classes.

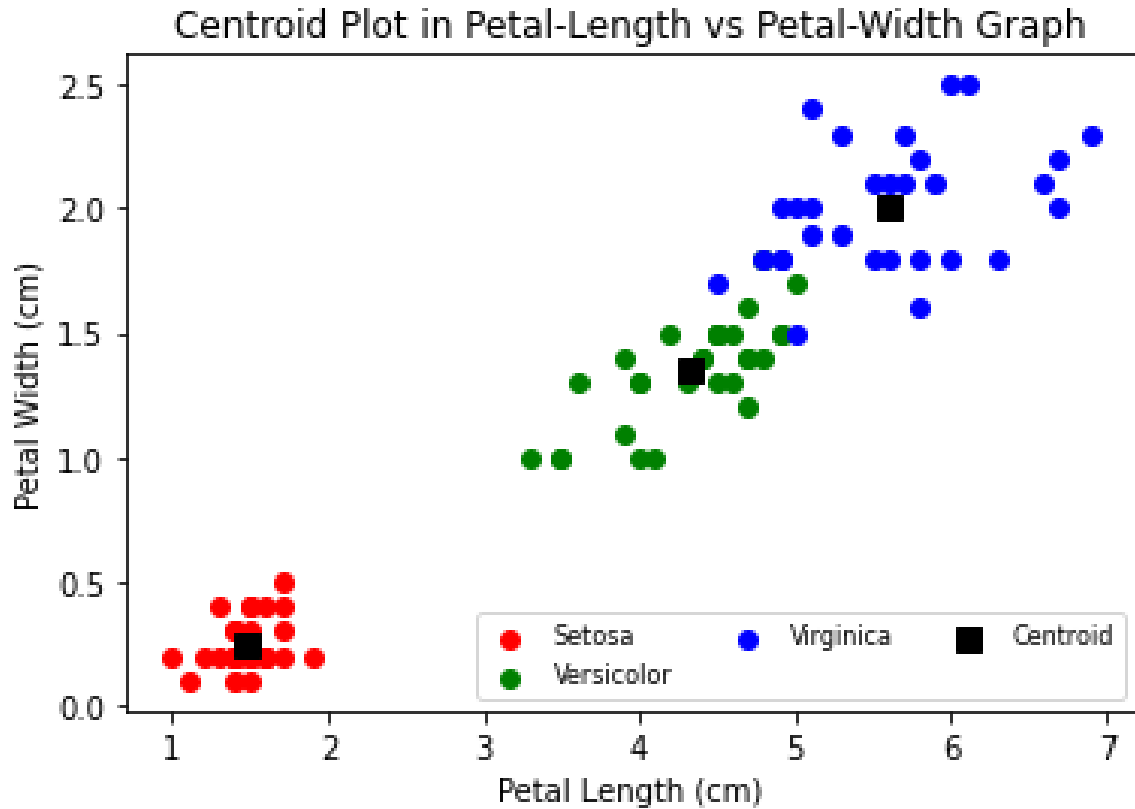
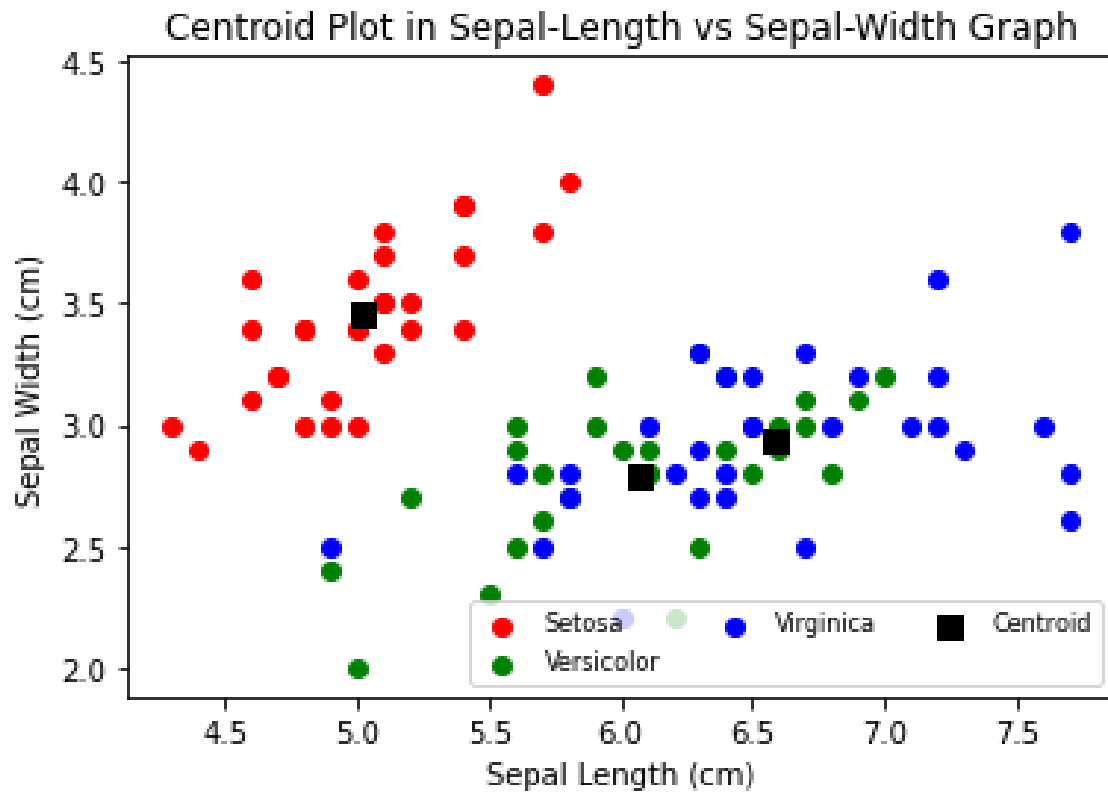
NOTE: The classes are clearly separable (with very less overlapping) using just two features. There is no need of using all the features to classify.

Here, I'll be using the features Petal-length and petal-width.

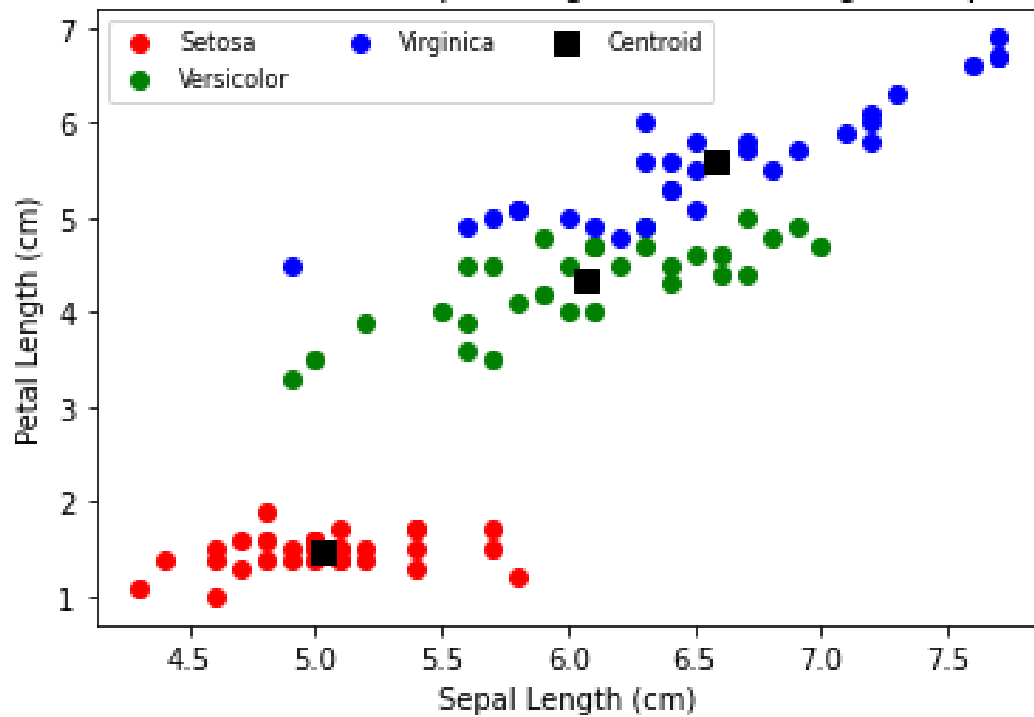
TRAINING

Training phase involves finding the centroid of each class.

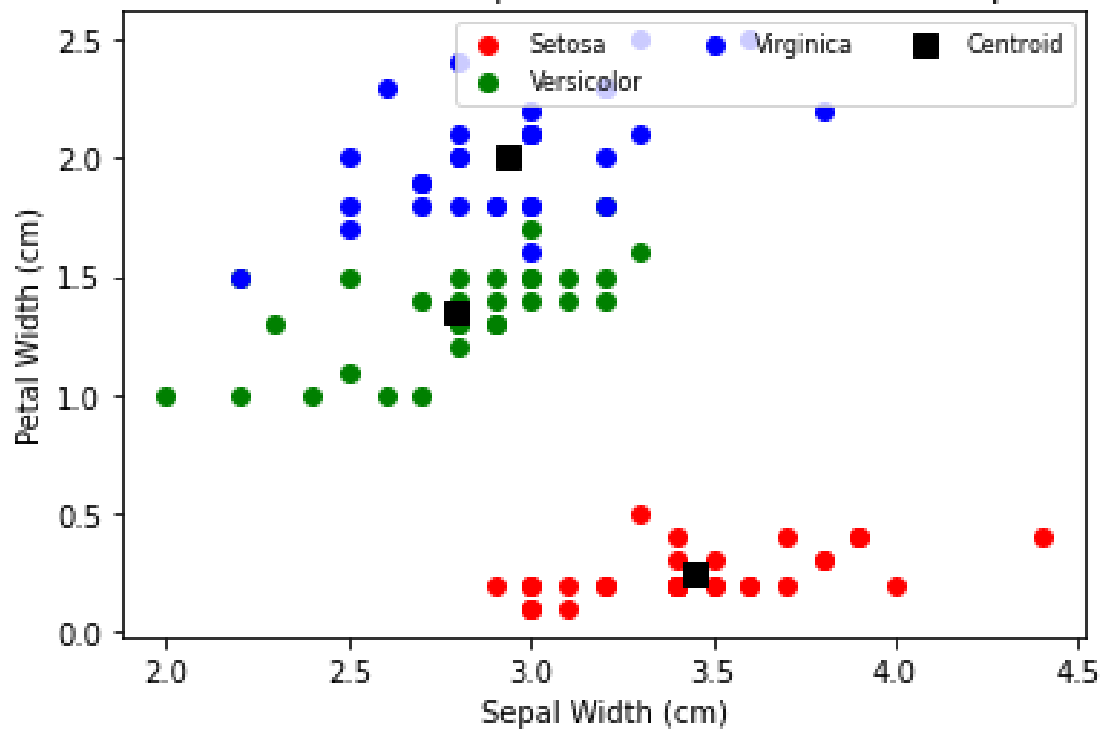
```
[[5.02666667 3.45      1.47333333 0.24666667]
 [6.07      2.79      4.33333333 1.35333333]
 [6.58333333 2.93333333 5.60333333 2.00666667]]
```



Centroid Plot in Sepal-Length vs Petal-Length Graph



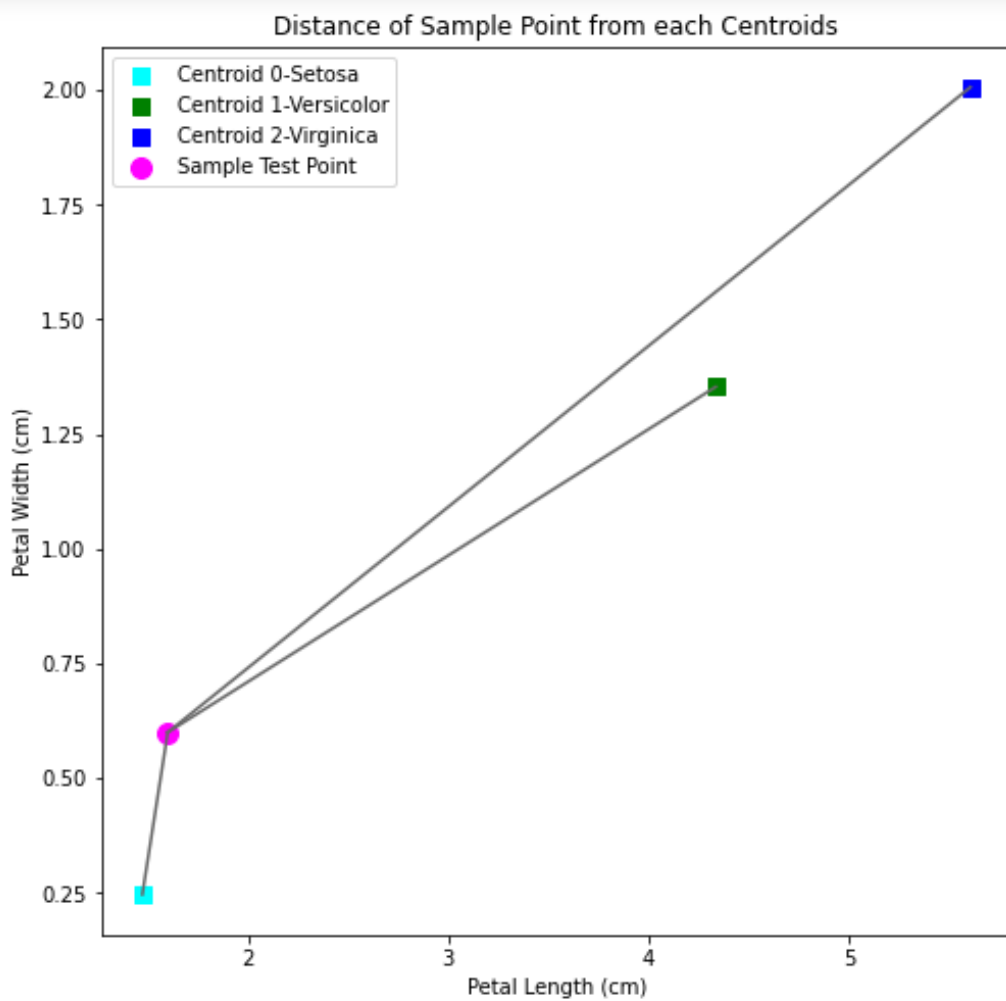
Centroid Plot in Sepal-Width vs Petal-Width Graph



ANALYSIS OF SAMPLE POINT USING EUCLIDEAN DISTANCE

Random Sample Point 1:

```
Sample Test Point Id 13  
True Class of Sample Test Point -> 0  
Predicted Class of Sample Test Point -> 0
```



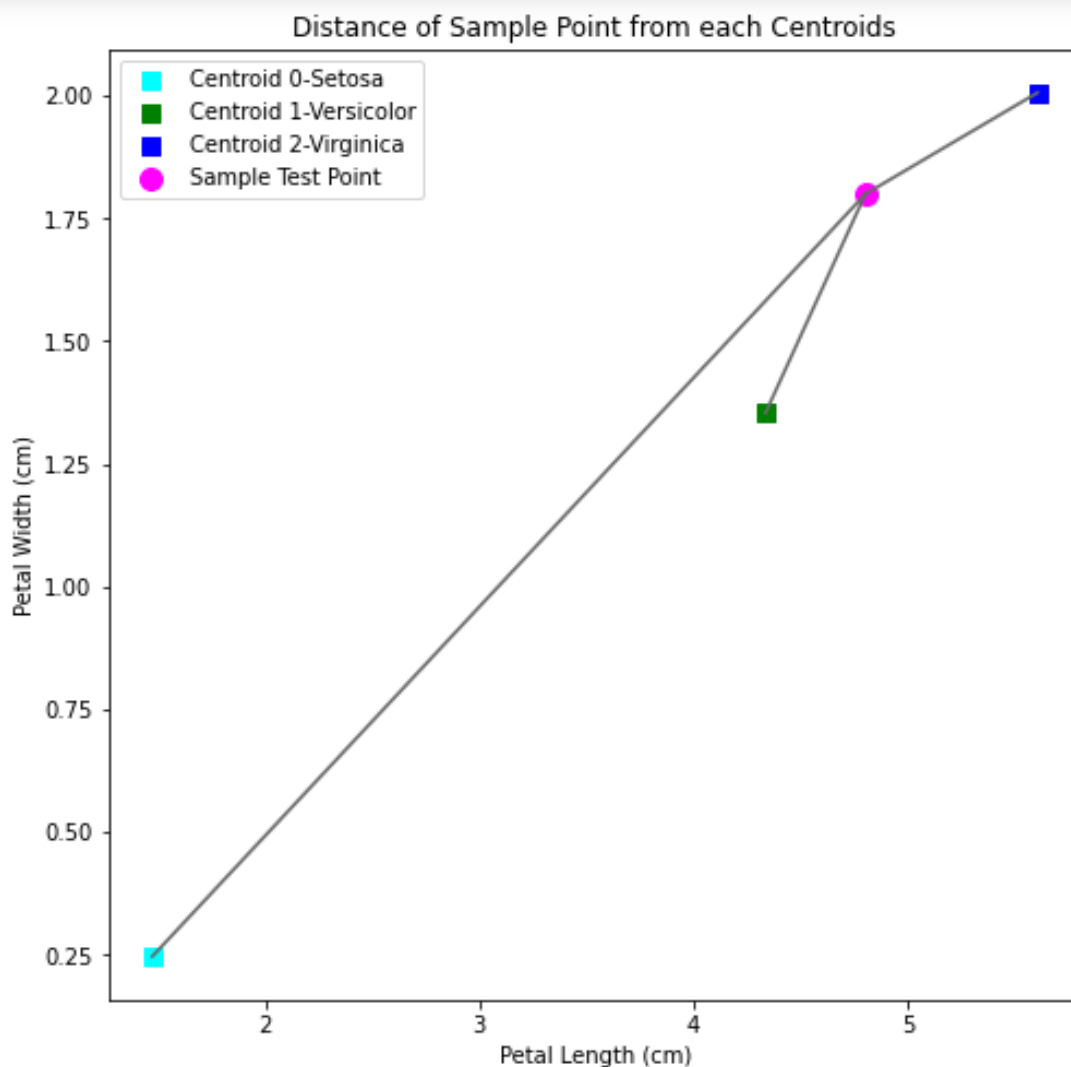
Comment: It's a case of correct classification. The distance of sample point to the centroid of class 0 is the least compared to its distance to the centroids of class 1 and 2.

Random Sample Point 2:

Sample Test Point Id 48

True Class of Sample Test Point -> 2

Predicted Class of Sample Test Point -> 1



Comment: It's a case of misclassification. The distance of sample point from both the centroids of class1 and class2 are equal. Thus, its classified to class 1 instead of class 2.

Comparison of True Class Labels and Predicted Class Labels

	True Labels	Predicted Labels
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	1	1
21	1	1
22	1	1
23	1	2
24	1	1
25	1	1
26	1	1
27	1	1
28	1	1
29	1	1
30	1	1
31	1	1
32	1	1
33	1	1
34	1	1
35	1	1
36	1	1
37	1	1
38	1	1
39	1	1
40	2	2
41	2	2
42	2	2
43	2	2
44	2	2
45	2	2
46	2	2
47	2	2
48	2	1
49	2	2
50	2	2
51	2	2
52	2	2
53	2	2
54	2	2
55	2	2
56	2	2
57	2	2
58	2	2
59	2	2

DISTANCE CLASSIFIERS COMPARISON

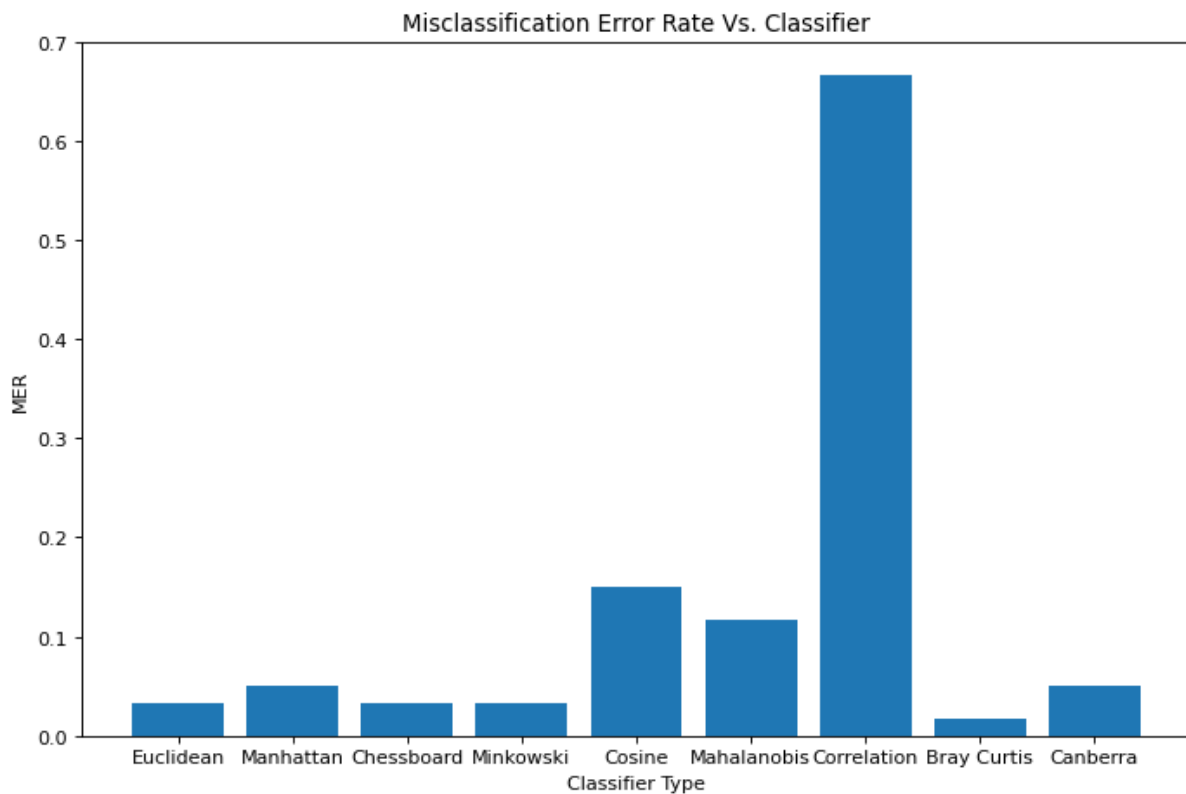
MISCLASSIFICATION ERROR RATE:

MER = No. of Incorrect prediction/No. of total predictions

ACCURACY OF MODEL:

Acc = 1 - MER

	MER	Accuracy
Euclidean	0.033333	96.666667
Manhattan	0.050000	95.000000
Chessboard	0.033333	96.666667
Minkowski	0.033333	96.666667
Cosine	0.150000	85.000000
Mahalanobis	0.116667	88.333333
Correlation	0.666667	33.333333
Bray Curtis	0.016667	98.333333
Canberra	0.050000	95.000000



Correlation distance model resulted in high misclassification compared to other distance models.

