This experiment is based on the IRIS data set. The data set contains a total of 150 samples, which has three classes: setosa, versicolor and virginica, and four features: sepal length, sepal width, petal length, petal width. Combine the 4 features in the sample in pairs, we can construct 12 combinations, as shown in the figure:

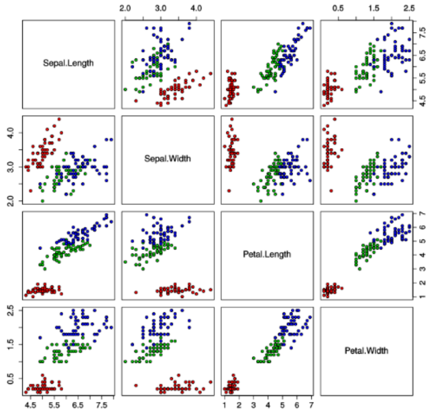

Figure. Iris Data Set

Table. Experimental Design

	Method Description	
Normalization	z-score	Min-max
Distance measure	Euclidean distance	Manhattan distance
Voting scheme	k weighted	k majority
K value	1-30	

According to the ratio of 8:2, 120 samples are randomly divided into training sets from 150 samples, and the remaining 30 are test sets. In this assignment, a KNN classifier will be implemented based on the above experimental design. Next, I will analyze their impact on the performance of the classifier from the four perspectives: normalization, distance measure, voting scheme, and k value. Because of the small amount of data in this experiment, the results obtained may not be representative, so the following part will perform ten runs for each comparison to get the average of the final results.

A) Normalization:

1. Based on Euclidean distance, weighted vote, and k=10, the output accuracy when using z-score normalization is 1. The average accuracy of 10 runs is 0.9715.
2. Based on Euclidean distance, weighted vote, k=10, the output accuracy when using min-max normalization is 0.9333. The average accuracy of 10 runs is 0.9244.

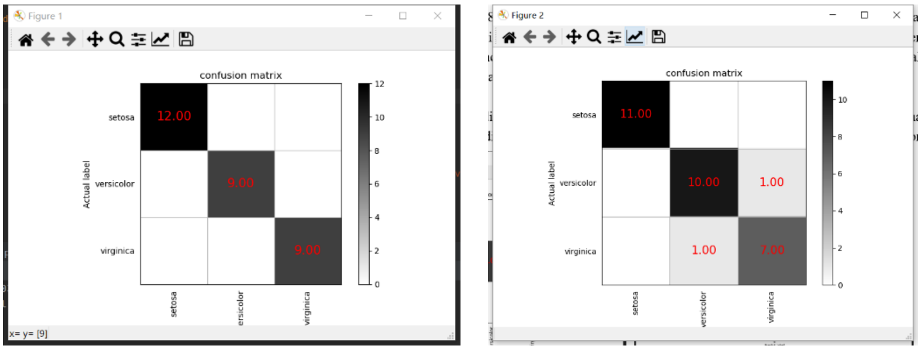

Figure. Output confusion matrix (z-score, min-max)

B) Distance Measure:

1. Based on weighted vote, k = 10, z-score normalization, the output accuracy when using Euclidean distance is 0.9333. The average accuracy of 10 runs is 0.9191
2. Based on weighted vote, k = 10, z-score normalization, the output accuracy when using Manhattan distance is 0.4333.The average accuracy of 10 runs is 0.5615

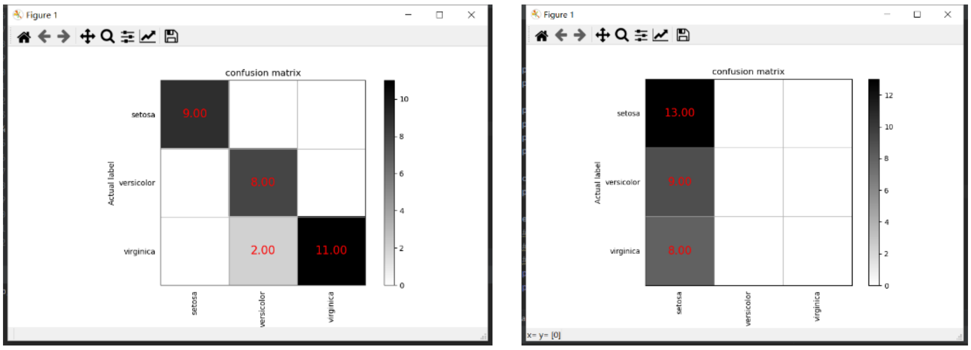

Figure. Output confusion matrix (Euclidean, Manhattan)

C) Voting Scheme:

1. Based on Euclidean distance, k = 10, and z-score normalization, the output accuracy when using weighted voting is 0.9667. The average accuracy of 10 runs is 0.9590.
2. Based on Euclidean distance, k = 10, and z-score normalization, the output accuracy when using majority voting is 0.9. The average accuracy of 10 runs is 0.9125.

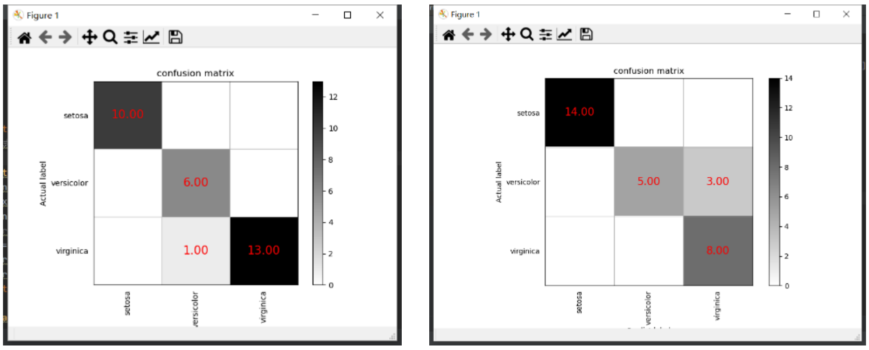

Figure. Output confusion matrix (weighted, majority)

D) K value:

In this part, other parameter settings will be fixed as Euclidean distance, z-score normalization, and weighted voting, the output accuracy when k is in range [1,30] is as follows.

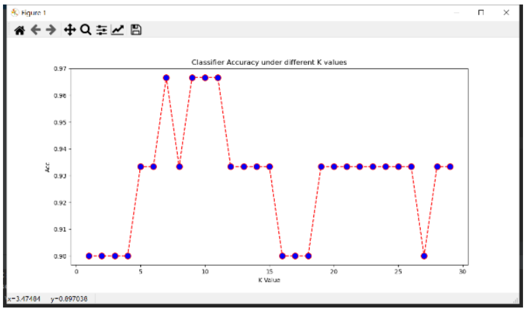

Figure. Output acc versus k value

Conclusion:

The min-max scales the eigenvalue range to (0,1), and the z-score scales the eigenvalue to near 0 to make it fit the Gaussian distribution better. After experimenting, I think that the performance of the two normalizations is basically the same, because the data set itself is showed as a good Gaussian distribution, so z-score performs better . Euclidean distance is better because, Manhattan distance calculation is not a straight line feature, so the results obtained often do not match the results predicted by the classifier, so the output is not representative. Since KNN only determines the category of the sample to be classified according to the categories of the nearest samples of the test points in the classification decision, in the KNN algorithm, the weighted voting effect is better.if k is too small and model is complex, the system will be easy to overfitting. if k is too high, the estimation error of learning will increase, and the prediction results are very sensitive to the instance points of the nearest neighbor. The greater K value can reduce the estimation error of learning, but the approximate error of learning will increase, and the training examples far away from the input examples will also play a role in the prediction. According to the experiment, k should take the smallest and most stable value, so it is probably best to take k around 10.