**Homework 5**

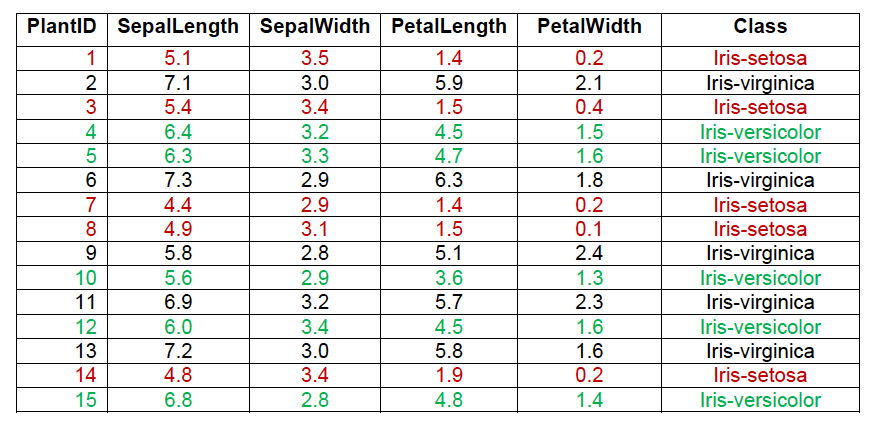
**Name: Megha Godwal**

**UID: U01695574**

**Nearest Neighbors**

We use a subset of the "Iris Plants Database" dataset (provided by WEKA, contained in the "iris.arff" file).Each plant record (i.e., example) is represented by the 5 attributes.

1. SepalLength: the sepal length in cm.
2. SepalWidth: the sepal width in cm.
3. PetalLength: the petal length in cm.
4. PetalWidth: the petal width in cm.
5. Class: the classification attribute, with the possible values {Iris-setosa, Iris-versicolor, Iris-virginica}.



We want to predict the class for each of the following plants:

* Plant #16. (SepalLength=4.6; SepalWidth=3.6; PetalLength=1.0; PetalWidth=0.2).
* Plant #17. (SepalLength=6.1; SepalWidth=2.8; PetalLength=4.0; PetalWidth=1.3).
* Plant #18. (SepalLength=7.7; SepalWidth=3.0; PetalLength=6.1; PetalWidth=2.3).

**Part 1 - Manual Computation**

Apply the Nearest Neighbour learning algorithm to classify the three to-be-predicted plants (i.e., Plants #16-18), to determine what kind of plant it is.

Try the three different values for the neighbourhood size; i.e., k=1; 3; and 5. Use one of the geometry distance functions (e.g., Manhattan or Euclidean distance function).

1. For k=1, convert the data of the set of Plants #16-18 (together with their predicted class) into the ARFF format, and save it in the "plants\_test1.arff" file.
2. For k=3, convert the data of the set of Plants #16-18 (together with their predicted class) into the ARFF format, and save it in the "plants\_test2.arff" file.
3. For k=5, convert the data of the set of Plants #16-18 (together with their predicted class) into the ARFF format, and save it in the "plants\_test3.arff" file.

**Answer: -**

* **Plant #16**

D= (|x1-y1|+|x2-y2|+|x3- y3|+---------+|xn-yn|), where x1, x2--- are the distances of instances 1-15, y1, y2… are the distances of instances 16, 17, 18

k=1

Applying nearest neighbor theorem: class is Iris-setosa, as the distance between the nearest neighbor is less than one

k=3

Applying nearest neighbor theorem: class is Iris–setosa as the distance in the range under 3 are classified as Iris-setosa

k=5

Applying nearest neighbor theorem: class is Iris–setosa as the distance in the range under 5 are Iris-setosa.

Therefore, **Plant #16 is Iris-setosa.**

* **Plant #17**

k=1

Applying nearest theorem, class in the range of k=1, is Iris-versicolor

k=3

Applying nearest theorem, class in the range of k=3, is Iris-versicolor

k=5

Applying nearest theorem, class in the range of k=5, is Iris-versicolor

Therefore, **Plant #17 is Iris-versicolor**

* **Plant #18**

k=1

Applying nearest theorem, class in the range of k=1, is Iris-virginica

k=3

Applying nearest theorem, class in the range of k=3, is Iris-virginica

k=5

Applying nearest theorem, class in the range of k=5, is Iris-virginica

so, **Plant #18 is class Iris-virginica**

**Part 2 - Analysis with WEKA**

* Convert the dataset containing 15 examples (i.e., Plants #1-15) into the ARFF format (supported by WEKA), and save it in the "plants\_train.arff" file.
* Launch the WEKA tool, and then activate the "Explorer" environment.
* Open the "plants\_train" dataset (i.e., saved in the "plants\_train.arff" file). - For each attribute and for each of its possible values, how many instances in each class have the feature value (i.e., the class distribution of the feature values)?
* Go to the "Classify" tab. Select the IBk classifier. In the "Test options" panel select the "Supplied test set" option. Activate the nearby "Set..." button and locate the "plants\_test1.arff" file. Run the classifier and observe the results shown in the "Classifier output" window.
* - How many instances used for the training? How many for the test?
* - How many instances are incorrectly classified?
* - What is the MAE (mean absolute error) made by the learned classifier?
* - What can you infer from the information shown in the Confusion Matrix?
* - Visualize the errors made by the learned classifier. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?
* - How can you save the learned classifier to a file?
* Now, click on the "IBk - K 1 - W 0" label (i.e., close to the "Choose" button). Set KNN equal to 3 (i.e., to use the neighbourhood size of 3), and then click the "OK" button to save the new setting. Activate the nearby "Set..." button and locate the "plants\_test2.arff" file. Run the classifier and observe the results shown in the "Classifier output" window.
* - How many instances are incorrectly classified?
* - What is the MAE (mean absolute error) made by the learned classifier?
* - What can you infer from the information shown in the Confusion Matrix?
* - Visualize the errors made by the learned classifier. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?
* Now, click on the "IBk - K 3 - W 0" label (i.e., close to the "Choose" button). Set KNN equal to 5 (i.e., to use the neighbourhood size of 5), and then click the "OK" button to save the new setting. Activate the nearby "Set..." button and locate the "plants\_test3.arff" file. Run the classifier and observe the results shown in the "Classifier output" window.
* - How many instances are incorrectly classified?
* - What is the MAE (mean absolute error) made by the learned classifier?
* - What can you infer from the information shown in the Confusion Matrix?

- Visualize the errors made by the learned classifier. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?

**Answer: -**

For k=1

Instances are 15 and MAE is 0.0741

From the confusion matrix, we can say that all the instances were classified correctly

Matrix

a b c 🡨 classified as

1 0 0 | a = Iris-setosa

0 1 0 | b = Iris-virginica

0 0 1 | c = Iris-versicolor

For k=3

No instances and MAE is 0.124

From the confusion matrix, we can say that all the instances were classified correctly.

Matrix

a b c 🡨 classified as

1 0 0 | a = Iris-setosa

0 1 0 | b = Iris-versicolor

0 0 1 | c = Iris-virginica

For k=5

No instances and MAE is 0.1566

From the confusion matrix, we can say that all the instances were classified correctly.

Matrix

a b c 🡨 classified as

1 0 0 | a = Iris-setosa

0 1 0 | b = Iris-versicolor

0 0 1 | c = Iris-virginica