# BRAC University Advanced AI

# CSE 710

# MID Term 2018

**Total : 100**  **Time: 48 Hours** (Take Home Open Book Exam. Students are not permitted to discuss among each other.)

**Classify (use Python/R or weka) the following dataset using Naïve Bayes Classifier ( You need to divide the dataset into training, validataion and testing sets.)**

The Iris Dataset is a multivariate dataset. It has 5 attributes, the first one is sepal length (Numeric), second is sepal width (Numeric) third one is petal length (Numeric), the fourth one is petal width (Numeric) and the last one is the class itself. We have 150 rows that are equivalent to 150 flowers collected those flowers are divided into the different category. They are similar flowers that are Iris but the different category like Iris-setosa, Iris-versicolor, and Iris-virginica. It is important to know about these patterns because in future if you see similar data we can say that this data belong to the certain pattern. Based on these data, we can predict which kind of the Iris flower does new flower belongs. It is supervised data since we have the class (Nominal). **[ More Description of the dataset is given]**

1. Why do you need to divide the dataset into training, validation and testing sets? **[15 points]**
2. Visualize the data : ( Show the different classes of Iris) **[15 points]**
3. Complete the following tasks on the training data :[**15 points]**
   1. Separate Data by class
   2. Calculate Mean
   3. Calculate Standard Deviation
   4. Summarize Dataset
   5. Summarize attribute by class

What insights do you get from doing the above ?

1. Now you can do the prediction. Do the followings : **[40 points]**
   1. Calculate Gaussian Probability
   2. Calculate Class Probabilities
   3. Make a Prediction
   4. Estimate accuracy

The task is to assign a New data to one or more classes or categories is classification or categorization.

The following results are required:

Correctly Classified Instances

Incorrectly Classified Instances

Kappa Statistic

Mean Absolute Error

Root Mean absolute Error

Root Relative Squared Error

Total Number of Instances

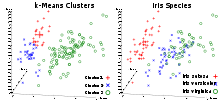
1. Build a Confusion matrix. With the help of it, how many of the data are rightly classified as Iris-setosa, Iris-versicolor and iris-virginica. How many of them were wrongly classified. Can you explain why the wrong classifications may have happened? **[ 15 points]**

**Description of the Dataset**

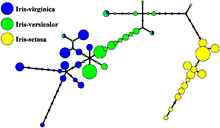
The ***Iris* flower data set** or **Fisher's *Iris* data set** is a [multivariate](https://en.wikipedia.org/wiki/Multivariate_statistics) [data set](https://en.wikipedia.org/wiki/Data_set) introduced by the British [statistician](https://en.wikipedia.org/wiki/Statistician) and [biologist](https://en.wikipedia.org/wiki/Biologist) [Ronald Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher) in his 1936 paper *The use of multiple measurements in taxonomic problems* as an example of [linear discriminant analysis](https://en.wikipedia.org/wiki/Linear_discriminant_analysis).[[1]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-fisher36-1) It is sometimes called **Anderson's *Iris* data set** because [Edgar Anderson](https://en.wikipedia.org/wiki/Edgar_Anderson) collected the data to quantify the [morphologic](https://en.wikipedia.org/wiki/Morphology_(biology)) variation of [*Iris*](https://en.wikipedia.org/wiki/Iris_(plant)) flowers of three related species.[[2]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-anderson36-2) Two of the three species were collected in the [Gaspé Peninsula](https://en.wikipedia.org/wiki/Gasp%C3%A9_Peninsula) "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".[[3]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-anderson35-3)

The data set consists of 50 samples from each of three species of *Iris* ([*Iris setosa*](https://en.wikipedia.org/wiki/Iris_setosa), [*Iris virginica*](https://en.wikipedia.org/wiki/Iris_virginica) and [*Iris versicolor*](https://en.wikipedia.org/wiki/Iris_versicolor)). Four [features](https://en.wikipedia.org/wiki/Features_(pattern_recognition)) were measured from each sample: the length and the width of the [sepals](https://en.wikipedia.org/wiki/Sepal) and [petals](https://en.wikipedia.org/wiki/Petal), in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

## Use of the data set

[](https://en.wikipedia.org/wiki/File:Iris_Flowers_Clustering_kMeans.svg)

Unsatisfactory [k-means clustering](https://en.wikipedia.org/wiki/K-means_clustering) result (the data set does not cluster into the known classes) and actual species visualized using [ELKI](https://en.wikipedia.org/wiki/Environment_for_DeveLoping_KDD-Applications_Supported_by_Index-Structures)

[](https://en.wikipedia.org/wiki/File:Principal_tree_for_Iris_data_set.png)

An example of the so-called "metro map" for the *Iris* data set.[[4]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-GorbanZinovyev2010-4) Only a small fraction of *Iris-virginica* is mixed with *Iris-versicolor*. All other samples of the different *Iris* species belong to the different nodes.

Based on Fisher's linear discriminant model, this data set became a typical test case for many [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) techniques in [machine learning](https://en.wikipedia.org/wiki/Machine_learning) such as [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machines)[[5]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-5).

The use of this data set in [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains *Iris setosa*, while the other cluster contains both *Iris virginica* and *Iris versicolor* and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in [data mining](https://en.wikipedia.org/wiki/Data_mining): Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.[[6]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-6)

Nevertheless, all three species of *Iris* are separable in the projection on the nonlinear branching principal component.[[7]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-7) The data set is approximated by the closest tree with some penalty for the excessive number of nodes, bending and stretching. Then the so-called "metro map" is constructed.[[4]](https://en.wikipedia.org/wiki/Iris_flower_data_set#cite_note-GorbanZinovyev2010-4) The data points are projected into the closest node. For each node the [pie diagram](https://en.wikipedia.org/wiki/Pie_chart) of the projected points is prepared. The area of the pie is proportional to the number of the projected points. It is clear from the diagram (left) that the absolute majority of the samples of the different *Iris* species belong to the different nodes. Only a small fraction of *Iris-virginica* is mixed with *Iris-versicolor* (the mixed blue-green nodes in the diagram). Therefore, the three species of Iris (*Iris setosa*, *Iris virginica* and *Iris versicolor*) are separable by the unsupervising procedures of nonlinear principal component analysis. To discriminate them, it is sufficient just to select the corresponding nodes on the principal tree.

Data Set

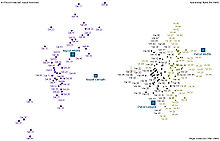
The dataset contains a set of 150 records under 5 attributes - Petal Length , Petal Width , Sepal Length , Sepal width and Class.

[](https://en.wikipedia.org/wiki/File:Iris_versicolor_3.jpg)

[*Iris versicolor*](https://en.wikipedia.org/wiki/Iris_versicolor)

[](https://en.wikipedia.org/wiki/File:Iris_virginica.jpg)

[*Iris virginica*](https://en.wikipedia.org/wiki/Iris_virginica)

[](https://en.wikipedia.org/wiki/File:Spectramap_Biplot_Iris_Flower_Data_Set_FULL.jpg)

Spectramap biplot of Fisher's iris data set

| Fisher's *Iris* Data hide | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Dataset Order** | **Sepal length** | **Sepal width** | **Petal length** | **Petal width** | **Species** |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | *I. setosa* |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | *I. setosa* |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | *I. setosa* |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | *I. setosa* |
| 5 | 5.0 | 3.6 | 1.4 | 0.3 | *I. setosa* |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | *I. setosa* |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | *I. setosa* |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | *I. setosa* |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | *I. setosa* |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | *I. setosa* |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | *I. setosa* |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | *I. setosa* |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 | *I. setosa* |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 | *I. setosa* |
| 15 | 5.8 | 4.0 | 1.2 | 0.2 | *I. setosa* |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 | *I. setosa* |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 | *I. setosa* |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 | *I. setosa* |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 | *I. setosa* |
| 20 | 5.1 | 3.8 | 1.5 | 0.3 | *I. setosa* |
| 21 | 5.4 | 3.4 | 1.7 | 0.2 | *I. setosa* |
| 22 | 5.1 | 3.7 | 1.5 | 0.4 | *I. setosa* |
| 23 | 4.6 | 3.6 | 1.0 | 0.2 | *I. setosa* |
| 24 | 5.1 | 3.3 | 1.7 | 0.5 | *I. setosa* |
| 25 | 4.8 | 3.4 | 1.9 | 0.2 | *I. setosa* |
| 26 | 5.0 | 3.0 | 1.6 | 0.2 | *I. setosa* |
| 27 | 5.0 | 3.4 | 1.6 | 0.4 | *I. setosa* |
| 28 | 5.2 | 3.5 | 1.5 | 0.2 | *I. setosa* |
| 29 | 5.2 | 3.4 | 1.4 | 0.2 | *I. setosa* |
| 30 | 4.7 | 3.2 | 1.6 | 0.2 | *I. setosa* |
| 31 | 4.8 | 3.1 | 1.6 | 0.2 | *I. setosa* |
| 32 | 5.4 | 3.4 | 1.5 | 0.4 | *I. setosa* |
| 33 | 5.2 | 4.1 | 1.5 | 0.1 | *I. setosa* |
| 34 | 5.5 | 4.2 | 1.4 | 0.2 | *I. setosa* |
| 35 | 4.9 | 3.1 | 1.5 | 0.2 | *I. setosa* |
| 36 | 5.0 | 3.2 | 1.2 | 0.2 | *I. setosa* |
| 37 | 5.5 | 3.5 | 1.3 | 0.2 | *I. setosa* |
| 38 | 4.9 | 3.6 | 1.4 | 0.1 | *I. setosa* |
| 39 | 4.4 | 3.0 | 1.3 | 0.2 | *I. setosa* |
| 40 | 5.1 | 3.4 | 1.5 | 0.2 | *I. setosa* |
| 41 | 5.0 | 3.5 | 1.3 | 0.3 | *I. setosa* |
| 42 | 4.5 | 2.3 | 1.3 | 0.3 | *I. setosa* |
| 43 | 4.4 | 3.2 | 1.3 | 0.2 | *I. setosa* |
| 44 | 5.0 | 3.5 | 1.6 | 0.6 | *I. setosa* |
| 45 | 5.1 | 3.8 | 1.9 | 0.4 | *I. setosa* |
| 46 | 4.8 | 3.0 | 1.4 | 0.3 | *I. setosa* |
| 47 | 5.1 | 3.8 | 1.6 | 0.2 | *I. setosa* |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 | *I. setosa* |
| 49 | 5.3 | 3.7 | 1.5 | 0.2 | *I. setosa* |
| 50 | 5.0 | 3.3 | 1.4 | 0.2 | *I. setosa* |
| 51 | 7.0 | 3.2 | 4.7 | 1.4 | *I. versicolor* |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 | *I. versicolor* |
| 53 | 6.9 | 3.1 | 4.9 | 1.5 | *I. versicolor* |
| 54 | 5.5 | 2.3 | 4.0 | 1.3 | *I. versicolor* |
| 55 | 6.5 | 2.8 | 4.6 | 1.5 | *I. versicolor* |
| 56 | 5.7 | 2.8 | 4.5 | 1.3 | *I. versicolor* |
| 57 | 6.3 | 3.3 | 4.7 | 1.6 | *I. versicolor* |
| 58 | 4.9 | 2.4 | 3.3 | 1.0 | *I. versicolor* |
| 59 | 6.6 | 2.9 | 4.6 | 1.3 | *I. versicolor* |
| 60 | 5.2 | 2.7 | 3.9 | 1.4 | *I. versicolor* |
| 61 | 5.0 | 2.0 | 3.5 | 1.0 | *I. versicolor* |
| 62 | 5.9 | 3.0 | 4.2 | 1.5 | *I. versicolor* |
| 63 | 6.0 | 2.2 | 4.0 | 1.0 | *I. versicolor* |
| 64 | 6.1 | 2.9 | 4.7 | 1.4 | *I. versicolor* |
| 65 | 5.6 | 2.9 | 3.6 | 1.3 | *I. versicolor* |
| 66 | 6.7 | 3.1 | 4.4 | 1.4 | *I. versicolor* |
| 67 | 5.6 | 3.0 | 4.5 | 1.5 | *I. versicolor* |
| 68 | 5.8 | 2.7 | 4.1 | 1.0 | *I. versicolor* |
| 69 | 6.2 | 2.2 | 4.5 | 1.5 | *I. versicolor* |
| 70 | 5.6 | 2.5 | 3.9 | 1.1 | *I. versicolor* |
| 71 | 5.9 | 3.2 | 4.8 | 1.8 | *I. versicolor* |
| 72 | 6.1 | 2.8 | 4.0 | 1.3 | *I. versicolor* |
| 73 | 6.3 | 2.5 | 4.9 | 1.5 | *I. versicolor* |
| 74 | 6.1 | 2.8 | 4.7 | 1.2 | *I. versicolor* |
| 75 | 6.4 | 2.9 | 4.3 | 1.3 | *I. versicolor* |
| 76 | 6.6 | 3.0 | 4.4 | 1.4 | *I. versicolor* |
| 77 | 6.8 | 2.8 | 4.8 | 1.4 | *I. versicolor* |
| 78 | 6.7 | 3.0 | 5.0 | 1.7 | *I. versicolor* |
| 79 | 6.0 | 2.9 | 4.5 | 1.5 | *I. versicolor* |
| 80 | 5.7 | 2.6 | 3.5 | 1.0 | *I. versicolor* |
| 81 | 5.5 | 2.4 | 3.8 | 1.1 | *I. versicolor* |
| 82 | 5.5 | 2.4 | 3.7 | 1.0 | *I. versicolor* |
| 83 | 5.8 | 2.7 | 3.9 | 1.2 | *I. versicolor* |
| 84 | 6.0 | 2.7 | 5.1 | 1.6 | *I. versicolor* |
| 85 | 5.4 | 3.0 | 4.5 | 1.5 | *I. versicolor* |
| 86 | 6.0 | 3.4 | 4.5 | 1.6 | *I. versicolor* |
| 87 | 6.7 | 3.1 | 4.7 | 1.5 | *I. versicolor* |
| 88 | 6.3 | 2.3 | 4.4 | 1.3 | *I. versicolor* |
| 89 | 5.6 | 3.0 | 4.1 | 1.3 | *I. versicolor* |
| 90 | 5.5 | 2.5 | 4.0 | 1.3 | *I. versicolor* |
| 91 | 5.5 | 2.6 | 4.4 | 1.2 | *I. versicolor* |
| 92 | 6.1 | 3.0 | 4.6 | 1.4 | *I. versicolor* |
| 93 | 5.8 | 2.6 | 4.0 | 1.2 | *I. versicolor* |
| 94 | 5.0 | 2.3 | 3.3 | 1.0 | *I. versicolor* |
| 95 | 5.6 | 2.7 | 4.2 | 1.3 | *I. versicolor* |
| 96 | 5.7 | 3.0 | 4.2 | 1.2 | *I. versicolor* |
| 97 | 5.7 | 2.9 | 4.2 | 1.3 | *I. versicolor* |
| 98 | 6.2 | 2.9 | 4.3 | 1.3 | *I. versicolor* |
| 99 | 5.1 | 2.5 | 3.0 | 1.1 | *I. versicolor* |
| 100 | 5.7 | 2.8 | 4.1 | 1.3 | *I. versicolor* |
| 101 | 6.3 | 3.3 | 6.0 | 2.5 | *I. virginica* |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 | *I. virginica* |
| 103 | 7.1 | 3.0 | 5.9 | 2.1 | *I. virginica* |
| 104 | 6.3 | 2.9 | 5.6 | 1.8 | *I. virginica* |
| 105 | 6.5 | 3.0 | 5.8 | 2.2 | *I. virginica* |
| 106 | 7.6 | 3.0 | 6.6 | 2.1 | *I. virginica* |
| 107 | 4.9 | 2.5 | 4.5 | 1.7 | *I. virginica* |
| 108 | 7.3 | 2.9 | 6.3 | 1.8 | *I. virginica* |
| 109 | 6.7 | 2.5 | 5.8 | 1.8 | *I. virginica* |
| 110 | 7.2 | 3.6 | 6.1 | 2.5 | *I. virginica* |
| 111 | 6.5 | 3.2 | 5.1 | 2.0 | *I. virginica* |
| 112 | 6.4 | 2.7 | 5.3 | 1.9 | *I. virginica* |
| 113 | 6.8 | 3.0 | 5.5 | 2.1 | *I. virginica* |
| 114 | 5.7 | 2.5 | 5.0 | 2.0 | *I. virginica* |
| 115 | 5.8 | 2.8 | 5.1 | 2.4 | *I. virginica* |
| 116 | 6.4 | 3.2 | 5.3 | 2.3 | *I. virginica* |
| 117 | 6.5 | 3.0 | 5.5 | 1.8 | *I. virginica* |
| 118 | 7.7 | 3.8 | 6.7 | 2.2 | *I. virginica* |
| 119 | 7.7 | 2.6 | 6.9 | 2.3 | *I. virginica* |
| 120 | 6.0 | 2.2 | 5.0 | 1.5 | *I. virginica* |
| 121 | 6.9 | 3.2 | 5.7 | 2.3 | *I. virginica* |
| 122 | 5.6 | 2.8 | 4.9 | 2.0 | *I. virginica* |
| 123 | 7.7 | 2.8 | 6.7 | 2.0 | *I. virginica* |
| 124 | 6.3 | 2.7 | 4.9 | 1.8 | *I. virginica* |
| 125 | 6.7 | 3.3 | 5.7 | 2.1 | *I. virginica* |
| 126 | 7.2 | 3.2 | 6.0 | 1.8 | *I. virginica* |
| 127 | 6.2 | 2.8 | 4.8 | 1.8 | *I. virginica* |
| 128 | 6.1 | 3.0 | 4.9 | 1.8 | *I. virginica* |
| 129 | 6.4 | 2.8 | 5.6 | 2.1 | *I. virginica* |
| 130 | 7.2 | 3.0 | 5.8 | 1.6 | *I. virginica* |
| 131 | 7.4 | 2.8 | 6.1 | 1.9 | *I. virginica* |
| 132 | 7.9 | 3.8 | 6.4 | 2.0 | *I. virginica* |
| 133 | 6.4 | 2.8 | 5.6 | 2.2 | *I. virginica* |
| 134 | 6.3 | 2.8 | 5.1 | 1.5 | *I. virginica* |
| 135 | 6.1 | 2.6 | 5.6 | 1.4 | *I. virginica* |
| 136 | 7.7 | 3.0 | 6.1 | 2.3 | *I. virginica* |
| 137 | 6.3 | 3.4 | 5.6 | 2.4 | *I. virginica* |
| 138 | 6.4 | 3.1 | 5.5 | 1.8 | *I. virginica* |
| 139 | 6.0 | 3.0 | 4.8 | 1.8 | *I. virginica* |
| 140 | 6.9 | 3.1 | 5.4 | 2.1 | *I. virginica* |
| 141 | 6.7 | 3.1 | 5.6 | 2.4 | *I. virginica* |
| 142 | 6.9 | 3.1 | 5.1 | 2.3 | *I. virginica* |
| 143 | 5.8 | 2.7 | 5.1 | 1.9 | *I. virginica* |
| 144 | 6.8 | 3.2 | 5.9 | 2.3 | *I. virginica* |
| 145 | 6.7 | 3.3 | 5.7 | 2.5 | *I. virginica* |
| 146 | 6.7 | 3.0 | 5.2 | 2.3 | *I. virginica* |
| 147 | 6.3 | 2.5 | 5.0 | 1.9 | *I. virginica* |
| 148 | 6.5 | 3.0 | 5.2 | 2.0 | *I. virginica* |
| 149 | 6.2 | 3.4 | 5.4 | 2.3 | *I. virginica* |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 | *I. virginica* |
|  |  |  |  |  |  |

Source : https://en.wikipedia.org/wiki/Iris\_flower\_data\_set