# Module 1 Midweek Project | Iris Classification

Trang Tran

CPS Analytics, Northeastern University

ALY6020 | Predictive Analytics

Professor Prashant Mittal

Nov 2, 2023

**Introduction**

This project explores and classifies iris flowers based on their attributes in the “Iris dataset” by implementing a K-Nearest Neighbors (KNN) classification model. The dataset contains measurements of sepal length, sepal width, petal length, and petal width for three species: Setosa, Versicolor, and Virginica, with labels of 0,1, and 2.

**Data**

The Iris dataset contains 150 instances of iris flowers, each categorized into one of three species. Its balanced distribution (as follows) makes it an ideal benchmark for classification tasks.

```
Class 0: 50 occurrences (33.33%)
Class 1: 50 occurrences (33.33%)
Class 2: 50 occurrences (33.33%)
```

The 3D scatter plot in Figure 1, made using the 'Matplotlib' tool, shows data points grouped into three classes (0, 1, 2), each represented by a different color ("Red," "Green," and "Yellow"). What stands out is that the green and yellow points mix a lot, meaning that the things in class 1 and class 2 (green and yellow) are pretty similar. There's also some mixing between red and green points, but it's not as much as with the other two colors. These findings suggest that the similarities in values between Versicolor and Virginica species (classes 1 and 2) may pose challenges in accurately predicting and classifying the target data.

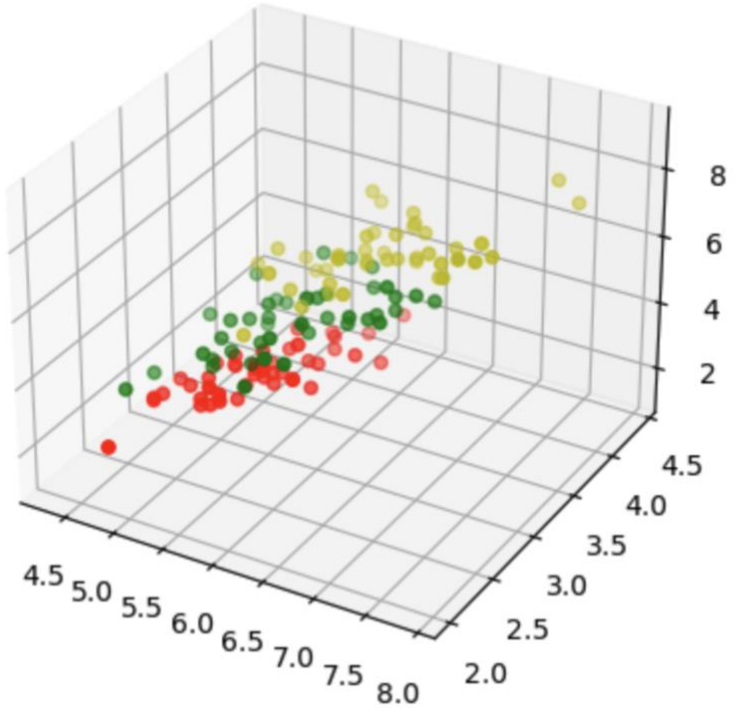


*Figure 1: 3D Plot of three different species*

**KNN Classification Model**

After loading the dataset, we separate the predictor variables and the target variable. Then we divide it into training and test sets and apply the K-NN model for classification. In this analysis, 14 test samples (<10% of the dataset) are evaluated after the training session with KNN models.

```
index:  0 , result of vote:  1 , label:  1 , data:  [6.1 2.8 4.7 1.2]
index:  1 , result of vote:  0 , label:  0 , data:  [4.6 3.2 1.4 0.2]
index:  2 , result of vote:  2 , label:  2 , data:  [6.2 3.4 5.4 2.3]
index:  3 , result of vote:  2 , label:  2 , data:  [5.7 2.5 5.  2. ]
index:  4 , result of vote:  1 , label:  1 , data:  [5.7 2.9 4.2 1.3]
index:  5 , result of vote:  1 , label:  1 , data:  [4.9 2.4 3.3 1. ]
index:  6 , result of vote:  2 , label:  2 , data:  [6.3 2.7 4.9 1.8]
index:  7 , result of vote:  1 , label:  2 , data:  [4.9 2.5 4.5 1.7]
index:  8 , result of vote:  2 , label:  1 , data:  [6.  2.7 5.1 1.6]
index:  9 , result of vote:  0 , label:  0 , data:  [5.1 3.5 1.4 0.3]
index:  10 , result of vote:  1 , label:  1 , data:  [5.1 2.5 3.  1.1]
index:  11 , result of vote:  1 , label:  1 , data:  [5.6 3.  4.5 1.5]
index:  12 , result of vote:  2 , label:  2 , data:  [6.2 2.8 4.8 1.8]
index:  13 , result of vote:  2 , label:  2 , data:  [7.2 3.6 6.1 2.5]
```

***The overall accuracy of the model***

Out of the 14 actual labels in the "label" column and the corresponding 14 predictions in the "results of vote" column, the model achieved an overall accuracy of 12 correct predictions out of 14, approximately 85.7%. It's worth noting that the two prediction failures occurred in classes 1 and 2 only.

***The accuracy of each type of iris***

- Setosa (Class 0): 2 correct predictions out of 2 actual labels, resulting in 100% accuracy.
- Versicolor (Class 1): 5 correct predictions out of 6 actual labels, roughly 83.3% accuracy.
- Virginica (Class 2): 5 correct predictions out of 6 actual labels, roughly 83.3% accuracy.

***Is the model a good model or not?***

In summary, the model demonstrates strong overall performance (85.7%), achieving high accuracy (100%) for Class 0 (Setosa) and reasonably good accuracy (83.3% ) for Classes 1 (Versicolor) and 2 (Virginica). The minor misclassifications in the latter two classes can be attributed to the high similarity between their values. However, it is advisable to further evaluate the model's performance by testing it with various train-test splits, especially considering the balanced distribution of the Iris dataset. The current test set exhibits an uneven distribution among the three classes, which may result in a biased assessment of the predictive accuracy for each iris type.

**References**

ALY6020 – Module 1 – Lesson 1-5 page.

https://northeastern.instructure.com/courses/160676/pages/lesson-1-5-running-the-algorithm?module_item_id=9500386