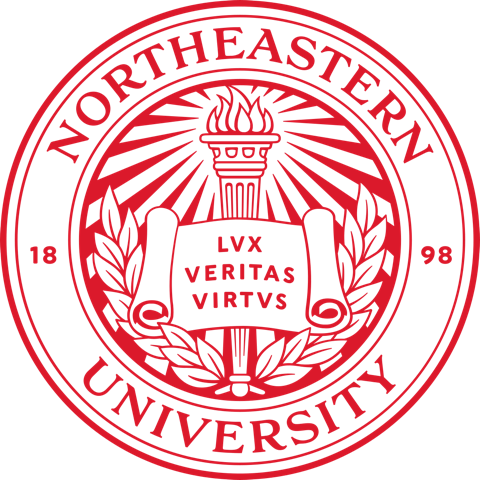
**Module 1 Midweek Project**

**Iris Classification**

Sunil Raj Thota



College of Professional Studies, Northeastern University ALY 6020 – Predictive Analytics

Prof. Justin Grosz

November 02, 2021

**Introduction:**

The Iris flower data set, sometimes known as Fisher's Iris data set, is a multivariate data set first published in 1936 by British statistician and biologist Ronald Fisher as an example of linear discriminant analysis in his paper. The use of numerous measurements in taxonomic problems. Because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species, it is frequently referred to as Anderson's Iris data set. We utilize a variety of flower characteristics to determine which sort of Iris they are. Two of the three species were collected in the Gaspé Peninsula, all from the same pasture, on the same day, and measured with the same equipment by the same person.

**Analysis:**

The goal of this report is to use the iris dataset to test among the most common supervised machine learning models, the k-NN/ k-Nearest Neighbor approach. The dataset's property Species is a parameter that will be forecasted using a k-NN algorithm. It is a classification task, in which we must determine if the plant belongs to the Versicolor, Setosa, or Virginica class based on some inputs to our classifier. The data for my analyses came from the sklearn package.

**Data Modelling:**

k-nearest neighbor is a non-parametric method used in classification. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In kNN regression, the output is the property value for the object. The characteristics are highly linearly associated, but their distributions vary between classifications, as we discovered throughout our data analysis. The K-NN method calculates commonalities among each occurrence, in this example each flower and its metrics, to produce forecasts.

1. **What was the overall accuracy of the model?**

The value of K is evaluated to be 3 in our case. On the foundation of vote and label matching, the model's accuracy value is computed. They match 11 out of 12 times, resulting in a 92 % accuracy, which is a great score.

1. **What was the accuracy of each type of iris?**

Let’s test 3 different values of K. The accuracy results with relation to the k-value are shown in the table below.

|  |  |
| --- | --- |
| k-value | Accuracy (%) |
| 2 | 92 |
| 3 | 92 |
| 4 | 92 |

1. **Would you classify the model as a good model or not?**

I would rate the algorithm as good because we were able to get a 92 percent accuracy, which is a great outcome. Accuracy rate stays unchanged regardless of the k-value. As a result, I believe it to be a well-balanced dataset.

**Conclusion:**

This demonstrates that the k-Nearest Neighbor Model, also known as the Lazy Learner Model, is among the easiest supervised machine learning techniques because it uses all the data to train but the model's output is excellent enough to be trusted. We calculated the Euclidean distance (feature names) and the values for target names (Classes of iris dataset). Regardless of the k-values, the accuracy rate appeared to be steady at 92 percent. As a result, I would consider this model to be well-balanced.

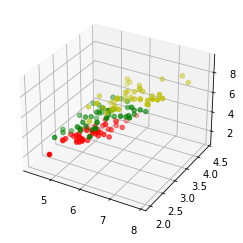
**References:**

Zhang Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational*

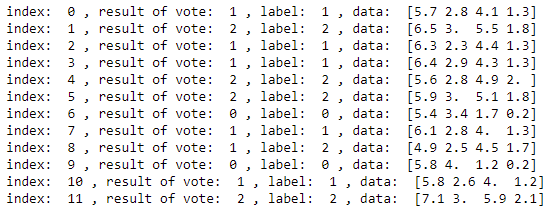
*medicine, 4*(11), 218. https://doi.org/10.21037/atm.2016.03.37

**Appendix:**

**Figure 1: Species Classification**

****

**Figure 2: k-NN Results**

****