College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 1 Midweek Project

Submitted on:

Jan 18, 2024

Submitted to:

Prof: BEHZAD AHMADI

Submitted by:

NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The iris dataset consists of 150 observations of iris flowers from three different species: Setosa, Versicolor, and Virginica.

This dataset is publicly available and included in many data analysis libraries and tools, making it one of the most accessible datasets for beginners in data science.

With an equal number of observations for each class (species), the Iris dataset is a balanced dataset, which is ideal for learning classification algorithms since it avoids bias toward a particular class.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. The dataset's features are as follows:

1. Sepal length: the length of the sepal (the part of the plant that encases the budding flower)
2. Sepal width: the width of the sepal
3. Petal length: the length of the petal (the colorful part of the plant that most people consider the flower)
4. Petal width: the width of the petal

These features are continuous variables, and the dataset does not contain any missing values.

EDA

- Petal Measurements: The heatmap highlights a very strong positive correlation (0.96) between petal length and petal width. This indicates that these two attributes tend to increase together and may carry similar information.

- The sepal length also correlates strongly with both petal length (0.87) and petal width (0.82), indicating that larger sepals are generally associated with larger petals.

The results of this heatmap (**Figure 1**) can inform feature selection for machine learning models.

Iris setosa can be easily distinguished from the other species by its shorter sepal length and wider sepal width. In contrast, Iris versicolor and Iris virginica show some overlap in sepal size but can be more effectively differentiated by their petal measurements.

Petal length and width clearly separate the three species, with setosa having the smallest petals, versicolor intermediate, and virginica the largest. These differences in petal size are significant and suggest that petal features are particularly useful for species classification within the Iris dataset.

From the scatterplots (**Figure 3 and 4**), we can infer that setosa is the most distinguishable species based on sepal and petal dimensions, as its data points form a distinct cluster separate from the other two species. Versicolor and virginica have some points in overlap.

KNN MODEL

Initially, the dataset was divided into two segments: 70% dedicated to training the model, and the remaining 30% for testing purposes. This helps us build and check a model to predict income based on factors like age, job type, and education

We then applied the MinMaxScaler function for standardization to both the training and testing datasets to adjust all feature values to a common scale, enhancing the accuracy and efficiency of our model.

Then k-nearest neighbors classification is used with varying k values (from 1 to 19) on a training dataset. For each value of k, the model's accuracy on the test dataset is computed and this can be used to determine the optimal k value for the dataset.

The optimal value of k is 3 here in **Figure 5** because this is where the accuracy begins to stabilize without further significant increases.

With an accuracy of **97.78%** on the Iris dataset, the KNN model demonstrates a high level of performance in classifying the species accurately. Model achieved **100%** accuracy for Setosa and Virginica, and approximately **94%** accuracy for Versicolor.

From the confusion matrix (**Figure 6**), we can conclude:

- The model has a perfect classification for Setosa (16/16) and Virginica (11/11).
- The model has slightly lower accuracy for Versicolor, with 17 correctly classified and 1 misclassified as Virginica.

CONCLUSION

In summary, the K-Nearest Neighbors (KNN) algorithm is well-suited for the Iris dataset due to its simplicity and the relevance of the features for classification tasks. These characteristics contribute to KNN's ability to achieve high accuracy in classifying the three species of Iris in the dataset, making it a great choice for this particular machine learning task.

REFERENCES

GeeksforGeeks. (2023, November 9). *K Nearest neighbor KNN algorithm*. https://www.geeksforgeeks.org/k-nearest-neighbours/

APPENDIX

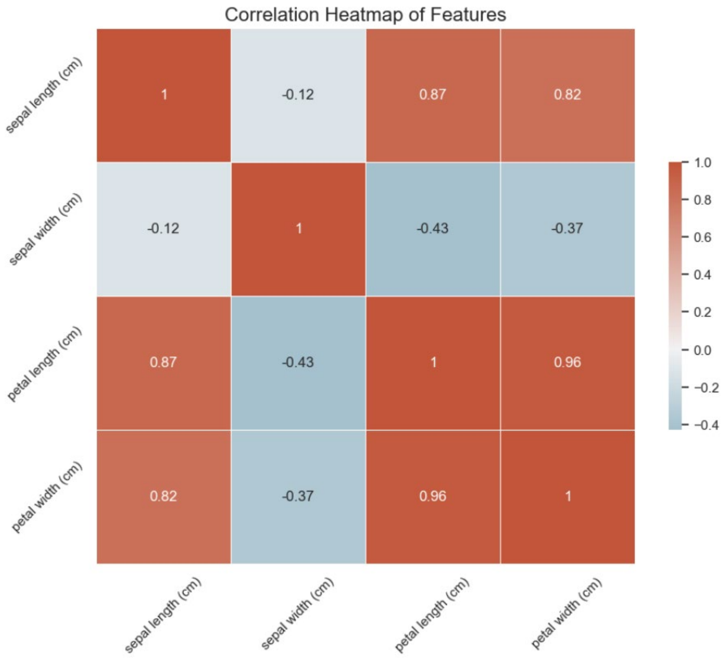

FIgure 1 – Correlation matrix

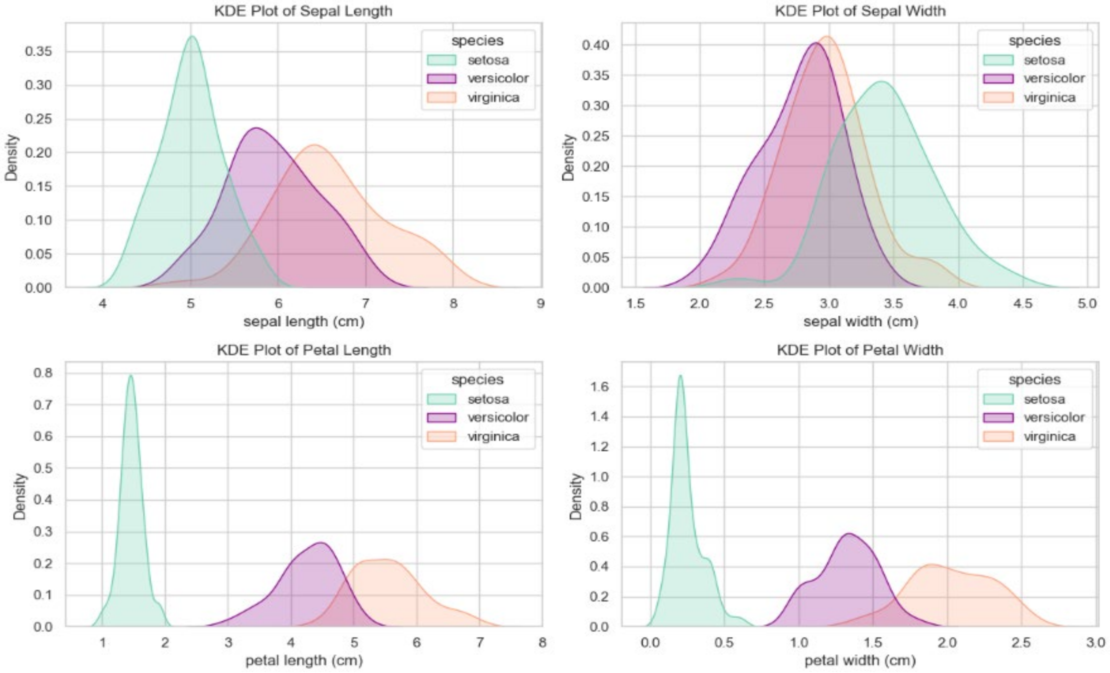

Figure 2 – Multiple density plot

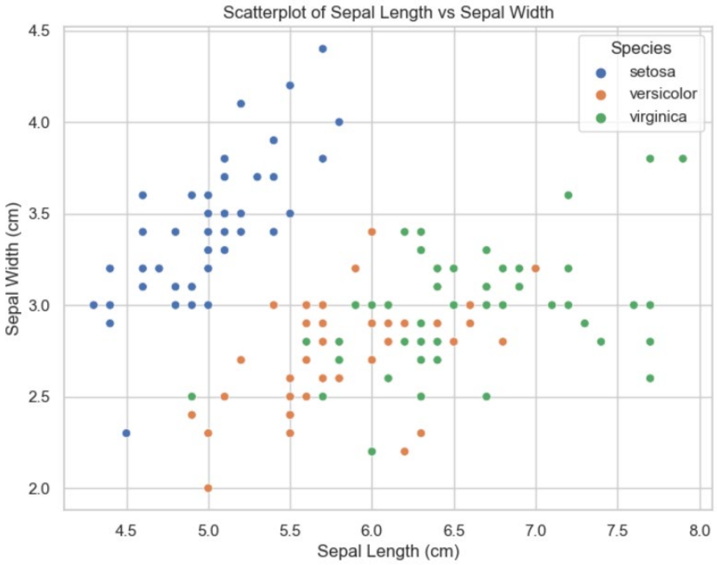

Figure 3 – Scatterplot A

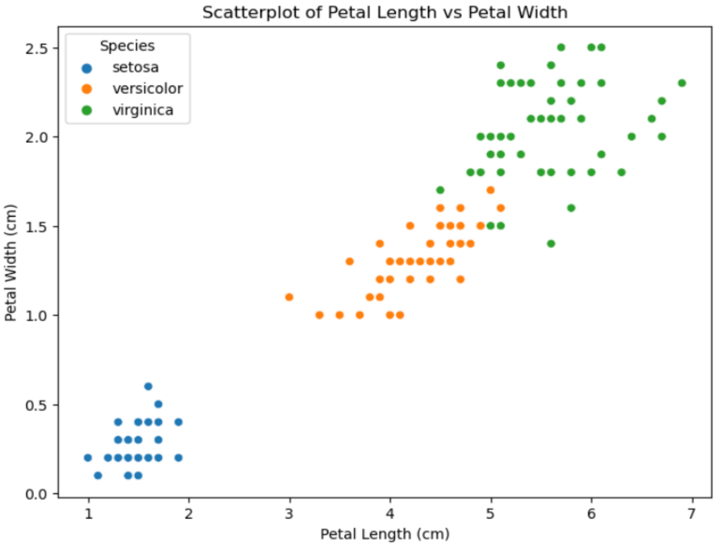

Figure 4 – Scatterplot B

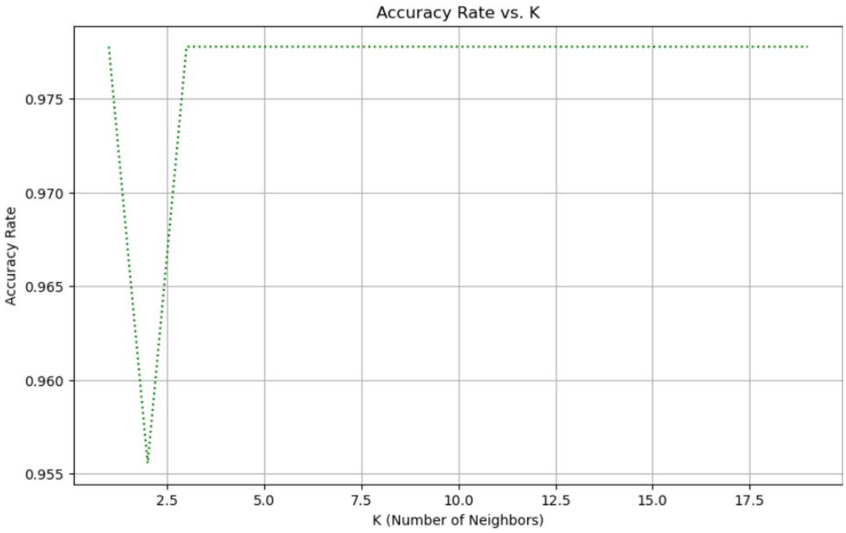

Figure 5 – Optimal k value selection

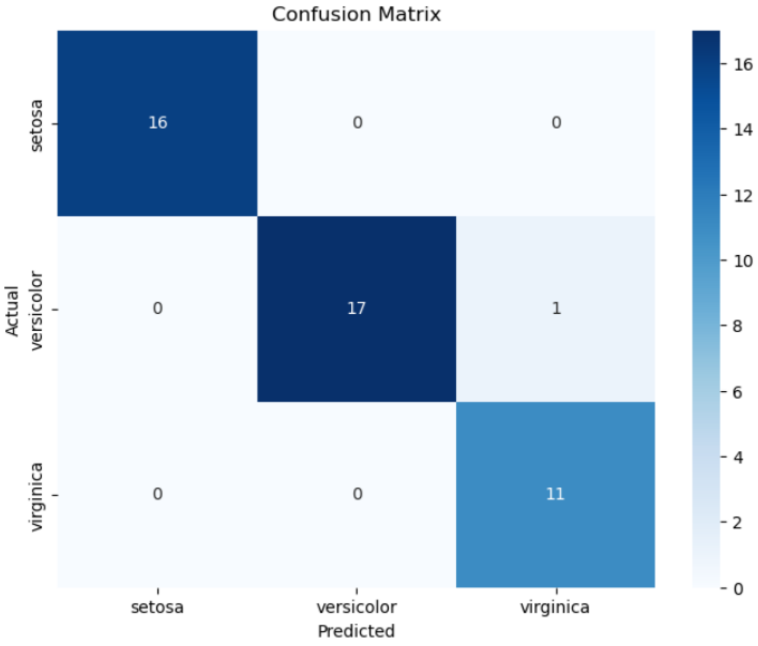

Figure 6 – Confusion matrix