Implementing the Nearest Neighbor Algorithm on the Iris Dataset

Sivaram Senthilkumar

Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst

622: Predictive Analytics and Statistical Learning

Professor Michael Prokle

Feb 16, 2025

Abstract

This study employs the K-Nearest Neighbors (KNN) algorithm to classify different species of Iris flowers using Scikit-Learn's Iris dataset. The objective is to explore the dataset, visualize key features, and evaluate the classification approach and model performance. Instead of utilizing sklearn.neighbors.KNeighborsClassifier, this study manually implements the k-NN algorithm and assesses its effectiveness. The model's performance is evaluated based on its classification accuracy, providing insights into its predictive capabilities. The results highlight the potential of k-NN for species classification while emphasizing the importance of parameter selection and dataset characteristics in achieving optimal performance.

Keywords: Iris Classification, K Nearest Neighbors(KNN), Model Accuracy, Train-Test, Feature Attributes, Manual Implementation.

Implementing the Nearest Neighbor Algorithm on the Iris Dataset

This analysis focuses on evaluating the overall accuracy of the K-Nearest Neighbors (KNN) model, assessing its ability to classify each Iris species, and determining its effectiveness as a classification approach.

Methodology

The dataset is obtained from Scikit-Learn's package, sklearn.datasets. Classification is performed using multiple attributes of Iris flowers, with the k-NN algorithm assigning species based on their nearest neighbors. The four attributes considered are

1. Sepal length (cm)
2. Sepal width (cm)
3. Petal length (cm)
4. Petal width (cm)

Dataset Overview

Number of samples considered: 150

Target names : Setosa, Versicolor, Virginica

Attributes considered: sepal length, sepal width, petal length, petal width

Feature Analysis Insights

The scatter plots generated for feature analysis reveal the following key insights (see Appendix Feature Analysis Insights Fig 1 for the scatter plot key insights). Petal length and petal width provide a clear distinction between the three species, making them the most effective features for classification. Sepal width alone does not offer strong separation among the species, indicating it may be less relevant in isolation. Setosa is the most easily separable class across

different feature combinations, while Versicolor and Virginica exhibit some degree of overlap.

Model Implementation

The K Nearest Neighbor classifier is initialized with K=1. The dataset is split into training and testing sets with a 68-32 split ratio. The model is trained on the training set and predictions were made on the test set. The classification is based on the Euclidean distance, which measures the similarity between data points. The Overall accuracy was calculated without using built-in functions.

Discussion

The overall accuracy of the KNN model is computed by comparing the model's predictions with actual labels. The accuracy for each type of the Iris species is calculated individually: 100% for Setosa, 94% for Versicolor, and 95% for Virginica, resulting in an overall accuracy of 96% (see Appendix Discussion Fig 2 for the accuracy bar chart). This accuracy suggests that the model performs well in classifying the test instances. I would classify the model as good, given its high accuracy on the test set. However, further evaluation on more diverse datasets and comparison with built-in classifiers would provide a more comprehensive assessment of its robustness.

References

Brownlee, J. (2022, January 30). Train-Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery.https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning- algorithms/

Gongde Guo, Hui Wang, David A. Bell, Yaxin Bi. (2004, August). KNN Model-Based Approach in Classification. International Journal of Machine Learning, 25(3), 123-145. https://www.researchgate.net/publication/2948052_KNN_Model-Based_Approach_in

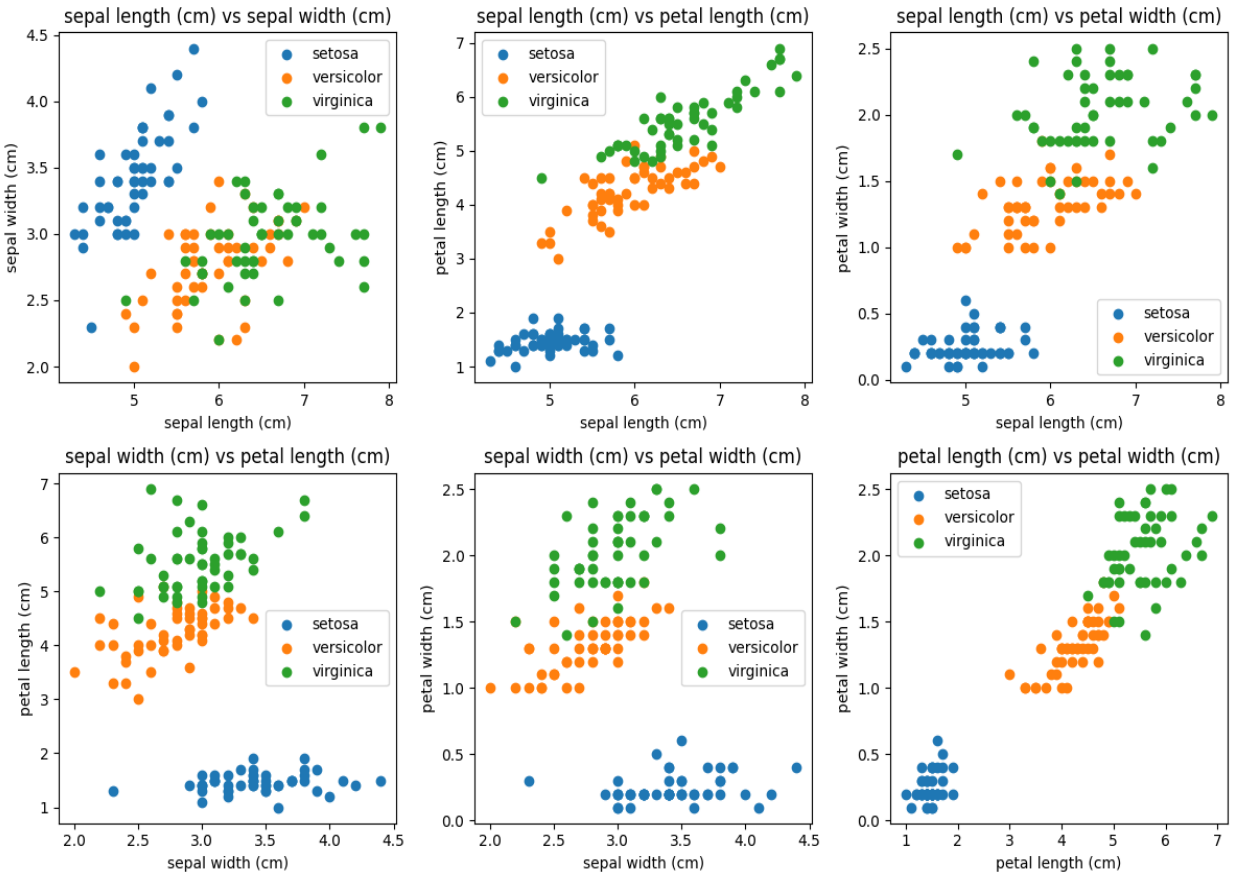
Classification

Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, Chih-Fong Tsai (2016). The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. https://springerplus.springeropen.com/articles/10.1186/s40064-016-2941-7

Appendix

Feature Analysis Insights

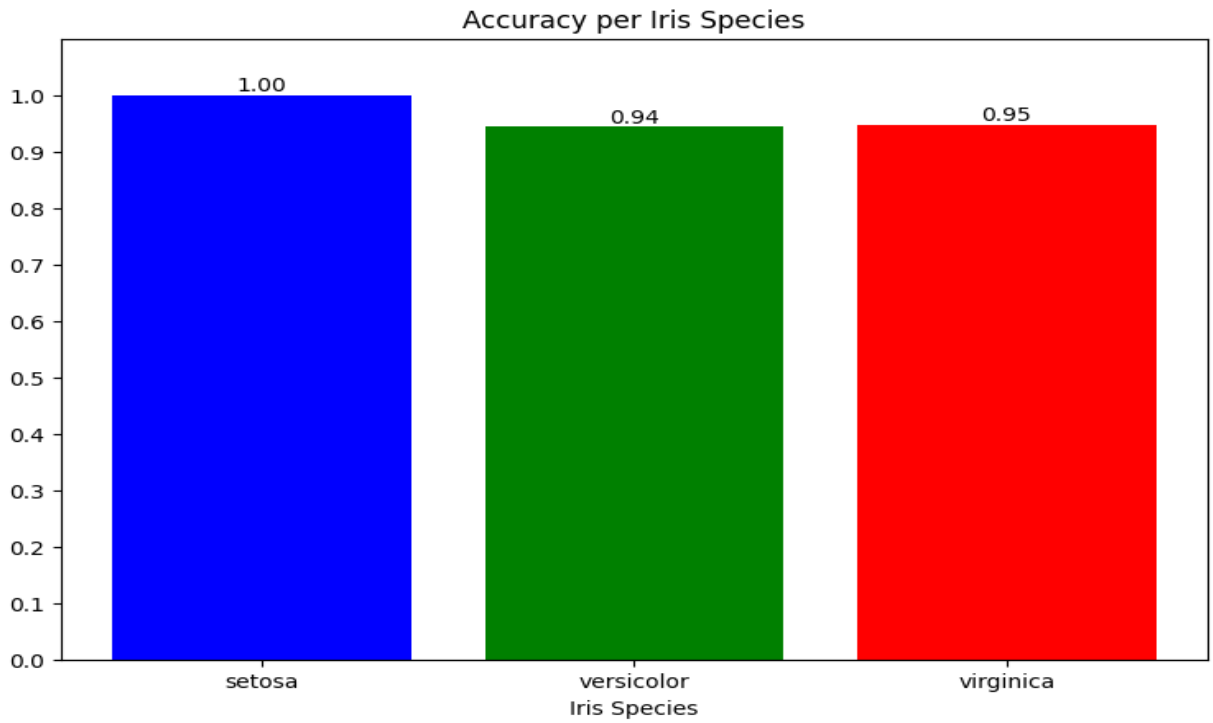
Fig 1



This Figure shows all the combination feature projection in the feature space.

Discussion

Fig 2



The figure illustrates the accuracy of the predicted iris species. Lower accuracy for Versicolor and Virginica is a result of their overlapping features.