IRIS FLOWER CLASSIFICATION

By AKPOTU EMMA

INTRODUCTION

The Iris flower dataset comprises three distinct species: setosa, versicolor, and virginica, each characterized by specific measurements. The objective of this task is to develop a machine learning model that can accurately classify these species based on their respective measurements. By analyzing features such as sepal length, sepal width, petal length, and petal width, the model will learn to differentiate between the species, enabling it to make predictions on new, unseen data. This approach not only highlights the potential of machine learning in botanical classification but also demonstrates the effectiveness of data-driven decision-making in understanding and categorizing biological entities.

REQUIREMENT GATHERING: The provided dataset is the Iris flower dataset, consisting of 150 rows and 6 columns, designed to evaluate classification performance across different Iris flower species. This dataset facilitates predictions and aids in the categorization of biological entities, showcasing the power of data analysis in understanding and differentiating among various species of Iris flowers.

DATA CLEANING: The dataset underwent a quality assessment, which included checks for duplicates and missing values using Python. During this assessment, we found that the data types for the features were recorded as float for numerical measurements and Object for categorical variables. This information is crucial for conducting accurate statistical analysis.

DATA ANALYSIS: Data visualization was conducted using Seaborn (sns) to illustrate the relationships between different features in the dataset. This method was employed to facilitate a systematic analysis of the data and enhance the understanding of how various measurements interact with one another.

KEY METRICS

Sepal Length and Sepal Width: The Iris flower dataset presents distinct variations in both sepal length and sepal width among its three species: Iris-setosa, Iris-versicolor, and Iris-virginica.
In terms of sepal length, Iris-setosa has relatively small sepals, averaging around 5.0 cm. Iris-versicolor exhibits a moderate sepal length at approximately 5.9 cm, while Iris-virginica displays the longest sepals, averaging about 6.6 cm. This clear differentiation in sepal lengths provides a distinct pattern that can significantly enhance the classification process. The pronounced differences underscore the biological diversity within the Iris genus and indicate that sepal length is a robust feature for distinguishing among these species. Therefore, utilizing sepal length could improve the accuracy of machine learning models tasked with classifying Iris flowers.
When examining sepal width, notable differences also emerge but to a lesser degree. Iris-setosa has the widest sepals, with an average width of about 3.4 cm, setting it apart from the other species. In comparison, Iris-

versicolor and Iris-virginica present average sepal widths of around 2.8 cm and 2.9 cm, respectively, indicating narrower sepals. Although there are variations in sepal width, the differences are less pronounced than those in sepal length, suggesting that sepal width may offer less distinguishing power for species classification.
Overall, while both sepal length and width contribute to the classification of Iris species, the more substantial variation in sepal length makes it a particularly critical feature. Understanding these differences is essential for developing effective classification models, as they collectively inform the differentiation process of these Iris species.

Petal Length and Petal Width: The analysis of petal lengths and widths across the three Iris species—Iris-setosa, Iris-versicolor, and Iris-virginica—reveals significant morphological distinctions that are pivotal for species classification.
In terms of petal length, Iris-setosa has the shortest petals, averaging approximately 1.5 cm, which clearly differentiates it from the other species. Iris-versicolor, with a mean petal length of around 4.3 cm, occupies a middle ground, while Iris-virginica stands out with the longest petals, averaging about 5.5 cm. This gradient in petal lengths not only defines their respective identities but also provides a robust criterion for distinguishing the species in classification tasks. Similarly, petal width exhibits pronounced differences among these species. Iris-setosa has a notably small average petal width of about 0.2 cm, which serves as a strong identifying trait. On the other hand, Iris-versicolor presents an intermediate petal width of roughly 1.3 cm, whereas Iris-virginica features the largest average width at about 2.0 cm.
These variations in both petal length and width highlight their significance as critical features in the classification of the Iris species. The clear distinctions in these morphological traits not only enhance our understanding of each species' unique characteristics but also emphasize their importance as predictive features in machine learning models. By integrating petal length and width with other morphological measurements, classifiers can more effectively differentiate among Iris-setosa, Iris-versicolor, and Iris-virginica, thereby improving the accuracy of botanical identification and study.

Correlation Between Features: A robust positive correlation exists between petal length and petal width across all species of the Iris dataset. This relationship suggests that, as the length of the petals increases, there is a corresponding tendency for the petal width to expand as well. Such a correlation indicates a consistent pattern in petal morphology, implying that these two measurements are likely influenced by similar biological factors.
Additionally, a strong correlation is observed between sepal length and petal length. This finding further reinforces the interconnectedness of these morphological traits within the Iris species. The relationship between sepal length and petal length suggests that changes in one measurement may be associated with changes in the other, reflecting underlying biological or environmental influences that shape the growth patterns of these flowers. Together, these correlations highlight the utility of petal length, petal width, and sepal length as significant features for understanding the relationships among the Iris species and for developing effective classification models.

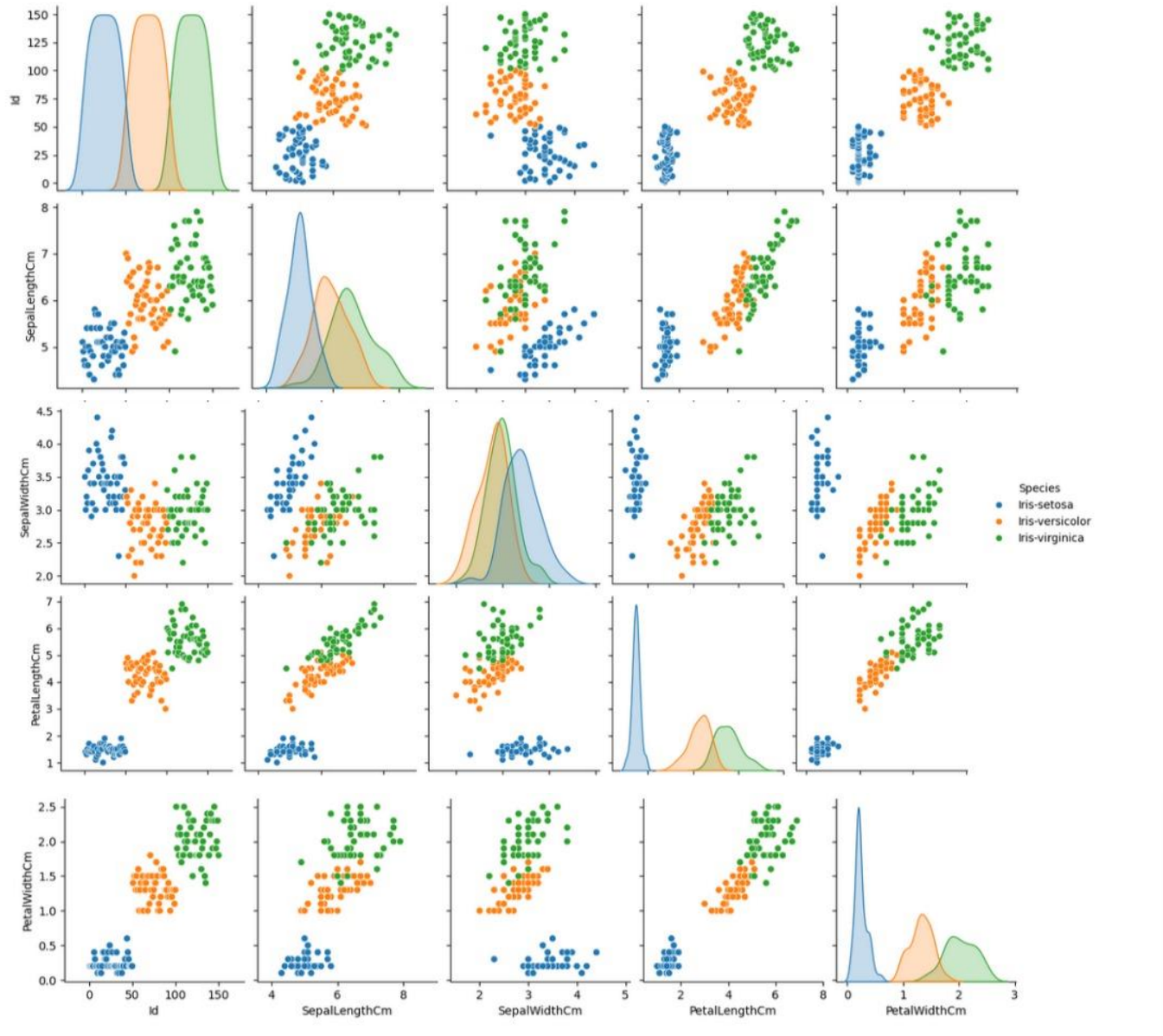
Visualization Insights

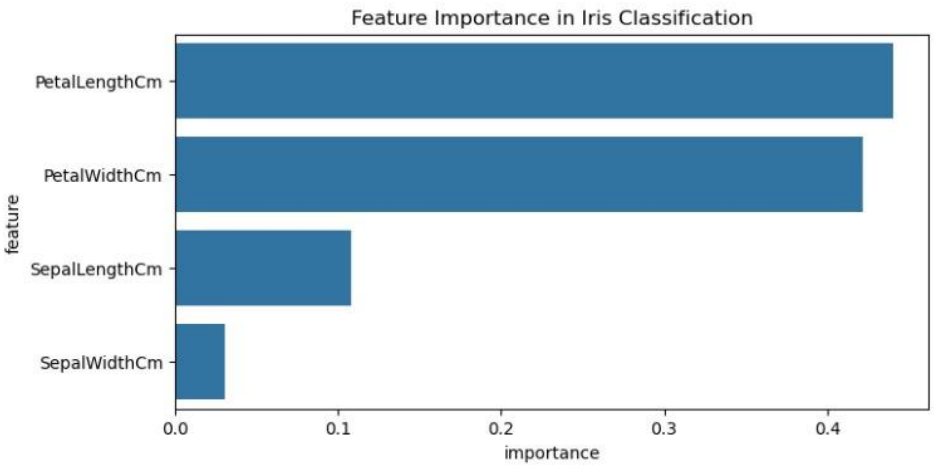
Visualizing the data using scatter plots or pair plots offers valuable insights into the separability of Iris species based on their features. A pair plot, in particular, effectively examines relationships and distributions across multiple dimensions within the dataset. Focusing on petal length and petal width, the pair plot showcases how distinctly different species—Iris-setosa, Iris-versicolor, and Iris-virginica—can be differentiated. Each species is typically represented by a unique color, making it easy to identify clusters or overlaps. The plot reveals a

significant separation among the species, with Iris-setosa distinctly clustered in the lower region due to its smaller petal dimensions. In contrast, Iris-versicolor and Iris-virginica display a wider range of petal lengths and widths, with Iris-virginica generally exhibiting larger measurements.

This visualization not only highlights the differences in petal dimensions among the species but also emphasizes the utility of petal length and petal width as effective features for classification tasks. By providing a clear graphical representation of species separability, the pair plot enhances initial exploratory data analysis and lays the groundwork for more advanced modeling and classification efforts.

```
# Create a pair plot to visualize relationships between features
sns.pairplot(flower, hue='Species')
plt.show()
```




Feature Importance in Iris Classification
PetalLengthCm
PetalWidthCm
SepalLengthCm
SepalWidthCm
feature
0.0
0.1
0.2
0.3
0.4
importance