Comprehensive Analysis and Visualization of Iris Flower Classification Using Support Vector Machine

Leela Krishna Sai Dangeti(101145083)

Madan Bandi(101142384)

Kotesh Ravula(101145486)

1. Data exploration and preparation:

We begin our research by exploring the Iris dataset, which is a comprehensive collection of measurements from many different kinds of iris blossoms. This dataset includes important parameters including sepal length, sepal breadth, petal length, as well as petal width. Our initial investigation of the dataset provides significant insights into its structure and all the data it contains. During this exploration, we learn important details like the total amount of samples in the collection of data and the number of characteristics connected with each sample. Before beginning model training, the dataset must be thoroughly prepared. This preparation requires dividing the dataset into a pair of independent parts: one to be used for training the algorithm and another to judge its performance. Furthermore, to maintain uniformity and comparability among all aspects.

2. Building the Model:

With an extensive knowledge of the data set at hand, that we begin building our model. Our preferred method is supported by SVM, an effective machine learning algorithm that excels at classification jobs like those provided by the Iris dataset. Using the SVM framework, that we select a linear kernel, which is also one of numerous options that determine how the SVM distinguishes between the different types of iris flowers. This strategic decision is consistent with our goal of effectively classifying iris flowers on the basis of relevant traits, paving the path for precise forecasts and insightful analysis.

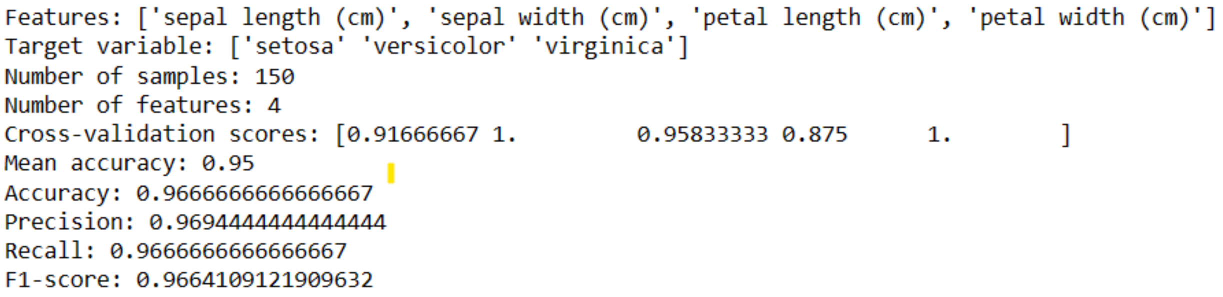
3. Assessing Model Performance:

To thoroughly evaluate the effectiveness of our model, we use an algorithm that's called K-fold cross-validation. This method divides the training data into many subsets, or "folds," trains the algorithm on each subset while validating the remainder, and then averages the performance over all folds. This strategy allows us to gain a more trustworthy prediction for the model's performance while reducing the likelihood of overfitting. Overfitting happens when an algorithm learns to perform very well on training data but is unable to generalize successfully to new, unknown data. We intend to use K-fold cross-validation to make sure that our model is robust and effective across a variety of data subsets, hence increasing its reliability and value in real-world scenarios.

4. Evaluating the Results:

After training and running our model on the validation data set, we thoroughly analyze its performance. To assess how efficient it is, we use a variety of metrics such as precision, recall, precision, and F1-score. Accuracy is the percentage of correctly classified examples to the total number of occurrences investigated. Precision examines the model's accuracy in making positive predictions, whereas recall examines its ability to properly identify positive examples in the dataset. The F1-score balances precision and recall, providing a comprehensive evaluation of the algorithm's performance.

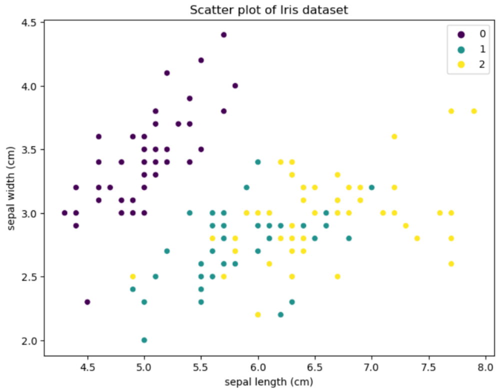
In addition to these essential measurements, we create a thorough classification report that combines these performance parameters for each type of iris flower. This report provides us with a sophisticated insight of how well the model performs across various flower varieties, allowing us to identify any potential gaps or strengths in its categorization capabilities. Using these measurements and reports, we receive crucial insights into the algorithm's overall performance and ability to generalize well across varied datasets.

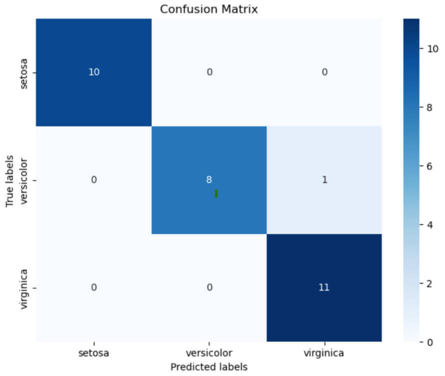


```
Features: ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
Target variable: ['setosa' 'versicolor' 'virginica']
Number of samples: 150
Number of features: 4
Cross-validation scores: [0.91666667 1.         0.95833333 0.875      1.        ]
Mean accuracy: 0.95
Accuracy: 0.9666666666666667
Precision: 0.9694444444444444
Recall: 0.9666666666666667
F1-score: 0.9664109121909632
```

5. Visualizing the Results:

We end our analysis by showing the distribution of the iris dataset using a scatter plot, with different colors representing different flower varieties. Additionally, we create a confusion matrix to visually examine the model's classification performance, noting any areas for improvement or misclassifications. The above resulted visualization methods offer a varied insights into the given dataset and as well as model's performance.





By following above approach, we built, evaluated, and understood our SVM model for the Iris dataset in a simple manner.