Project Documentation

Iris Species Classification

Wonkwon Lee

Introduction

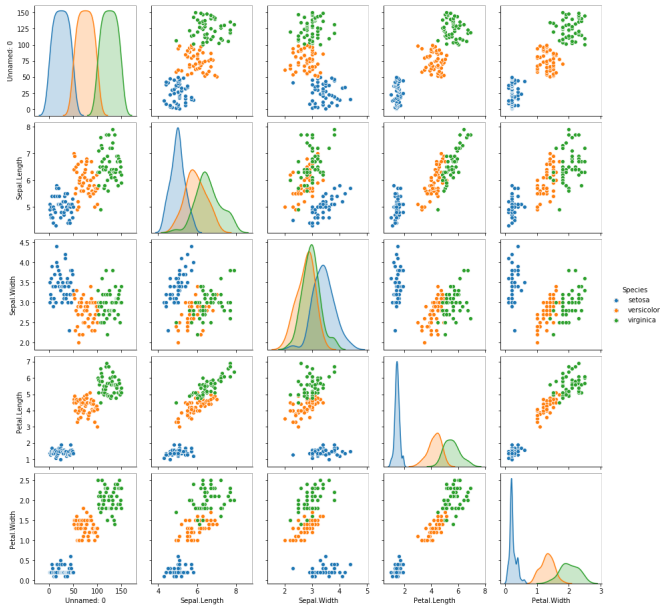
The first Computer Vision project I conducted in 2024 was Iris Classification, a classic machine learning project in Kaggle or Dacon. As a Volunteer Software Engineer, I have implemented image classification models that classify different types of iris flowers. This project explores classifying Iris flower species (Iris setosa, Iris versicolor, and Iris virginica) using machine learning algorithms. For this project, I employ the classic Iris dataset and investigate the efficacy of different classification models, including supervised and unsupervised learning. The analysis includes data exploration, model training, evaluation, and final selection of the best-performing model.

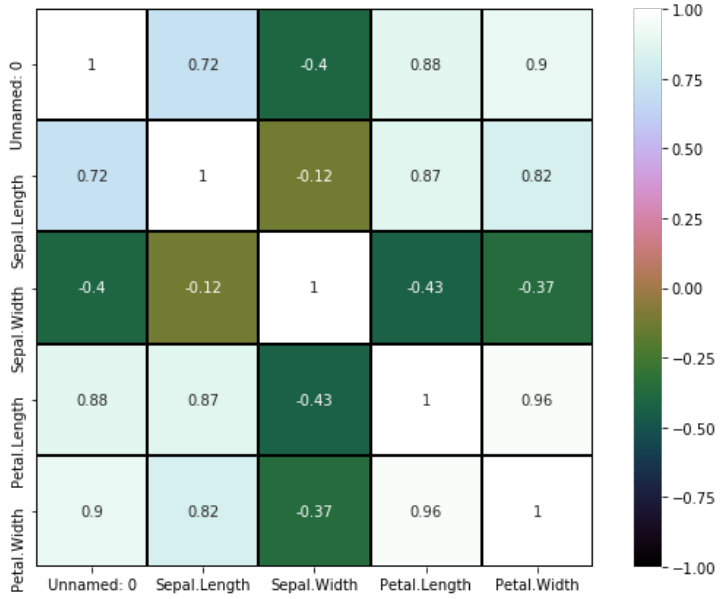
Motivation

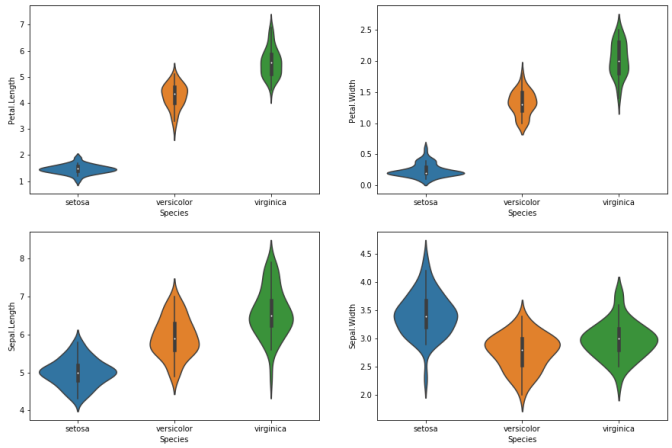
- **Problem Statement**: The objective is to develop a machine learning model capable of accurately classifying Iris flower species based on Sepal and Petal measurements.
- **Motivation**: This project demonstrates the application of machine learning in biological classification tasks. It highlights the process of dataset analysis, model selection, and performance evaluation, which form the foundation of many real-world machine learning use cases.

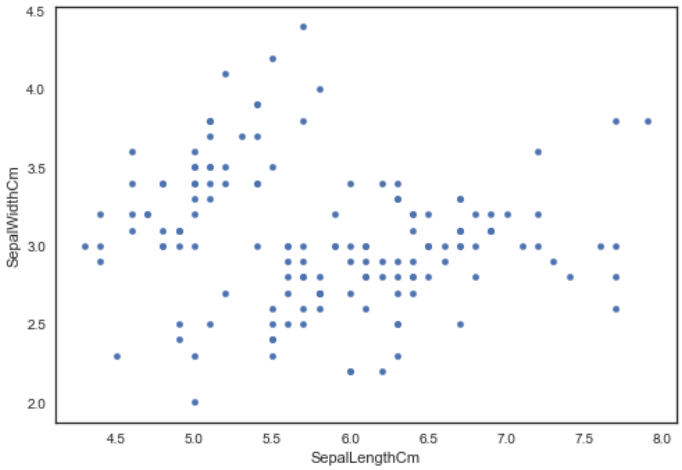
Dataset

One of the earliest known datasets used for evaluating classification methods, the Iris dataset is obtained from the UCI Machine Learning Repository. The dataset contains









150 samples with 4 features (Sepal Length, Sepal Width, Petal Length, Petal Width) and 3 target classes (Iris Setosa, Iris Versicolor, Iris Virginica) balanced with 50 samples each.

Exploratory Data Analysis (EDA)

In the above visualizations, pair plots and scatter plots reveal relationships and potential clusters among feature. Other plots illustrate the distribution of each feature and potential differences between species. The heatmap plot shows that Petals are expected to be more important than Sepal. While the Sepal Width and Length are not correlated, the Petal Width and Length are highly correlated.

Training and Evaluation

For this project, I have implemented several supervised and unsupervised machine learning models and evaluated the efficiency. The tested classification algorithms are Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree. The Iris dataset is split into training and testing sets with 0.7:0.3 ratio. Because Iris classification is well-known problem and is easy to solve, any further data splitting strategy is not needed for this specific task. For model implementation, Scikit-Learn Python framework is used for simple and efficient image classification task.

Metrics

- Accuracy: Overall percentage of correctly classified samples
- Precision, Recall, F1-Score: Provide a more granular class-wise performance analysis
- Confusion Matrix: Visualizes specific errors and highlights misclassifications

```
Confusion matrix:
 [[16  0  0]
 [ 0 14  1]
 [ 0  0  7]]
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        16
  versicolor       1.00      0.93      0.97        15
   virginica       0.88      1.00      0.93         7

    accuracy                           0.97        38
   macro avg       0.96      0.98      0.97        38
weighted avg       0.98      0.97      0.97        38
```

Model	Accuracy
Logistic Regression	0.96732
Support Vector Machine	0.96732
K-Nearest Neighbor	0.92892
Decision Tree	0.92892

Results

Among the four models, both Logistic Regression and SVM outperform the other models by achieving accuracy score of 0.96732. The KNN classifier and Decision Tree classifier both reports 0.92892 accuracy. In addition, the confusion matrix from Logistic Regression achieves high precision, recall, and F1-score, 0.96, 0.98, 0.97 respectively. Therefore, Logistic Regression performed very well in accuracy, precision, recall, and F1-score, demonstrating its efficiency in predictive analysis.

Discussion

The project demonstrates the importance of EDA and data visualizations. The EDA revealed clear distinctions between Iris species in the feature space. For instance, pair plots demonstrated that Iris setosa generally exhibits smaller sepals and petals compared to the other two species. This visual separation hinted that our features hold strong discriminatory power for classification.

Interestingly, while the Logistic Regression and SVM achieved an accuracy of 0.96732, simpler models like K-Nearest Neighbors also performed remarkably well. This reinforces two points: 1) More complex models don't always guarantee superior results on all datasets, and 2) Model selection should consider the balance between accuracy and interpretability, especially if understanding feature importance is crucial. As the dataset only contains four features – length and width of Petal and Sepal, making the problem easy to solve with simple machine learning techniques.

Though our results are promising, several avenues remain for further improvement. Introducing new features related to flower shape or color could enhance model performance. Additionally, feature scaling or dimensionality reduction could prove beneficial, especially if applying algorithms sensitive to feature scale or redundancy.

Key Takeaways

- The Iris dataset presents a well-defined classification problem, allowing us to demonstrate the power of EDA and the comparative analysis of machine learning algorithms
- Model choice involves careful consideration of performance metrics, computational costs, and the need for interpretability depending on the project's specific goals
- Even with a classic dataset like Iris, opportunities exist for refining classification performance through feature engineering and exploring more advanced techniques