

Iris classification



Sudip Dey





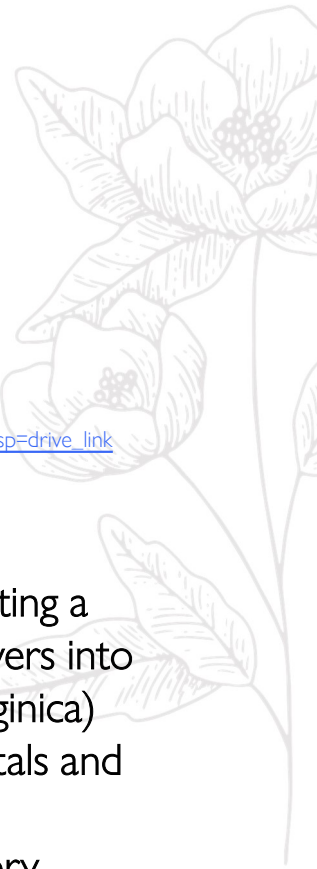
Project Overview



Dataset is available in the given link:

https://drive.google.com/file/d/19_BzURCJBh4xKrrix60E7SkHmHBXO9R1B/view?usp=drive_link

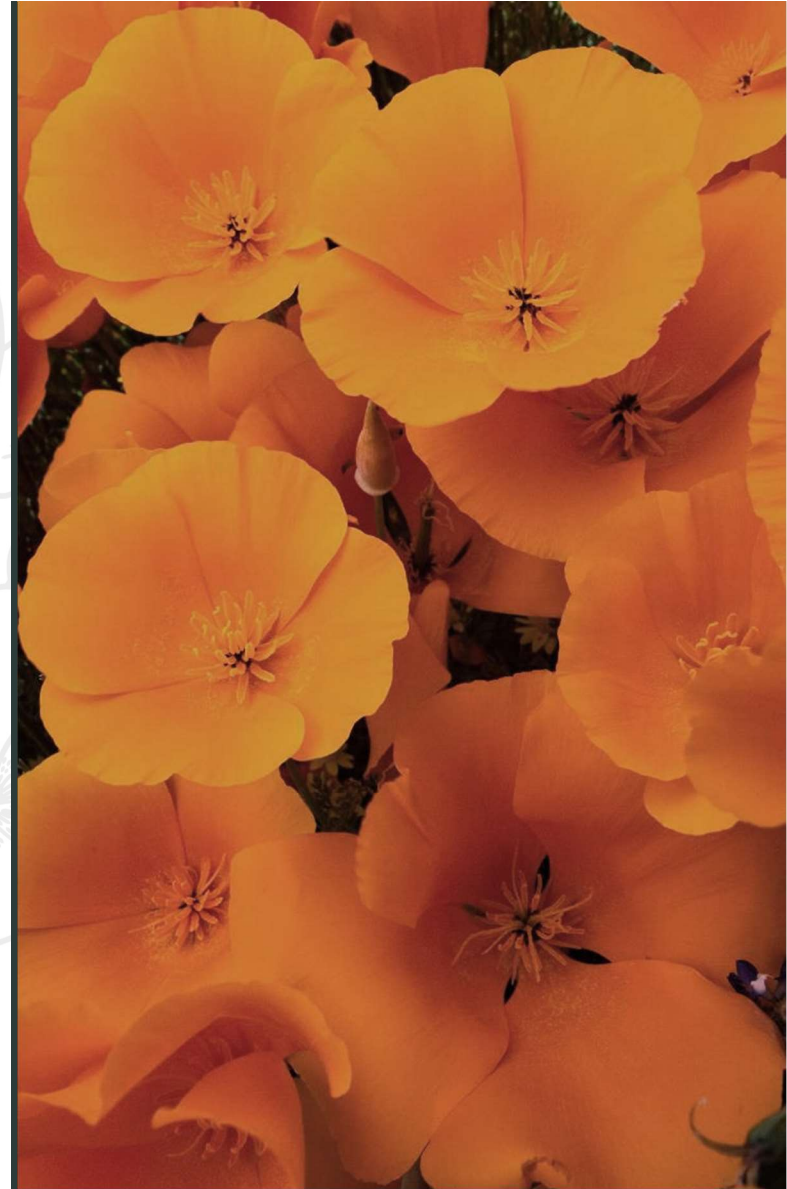
The Iris Classification project involves creating a machine learning model to classify iris flowers into three species (Setosa, Versicolour, and Virginica) based on the length and width of their petals and sepals. This is a classic problem in machine learning and is often used as an introductory example for classification algorithms.



Problem Statement

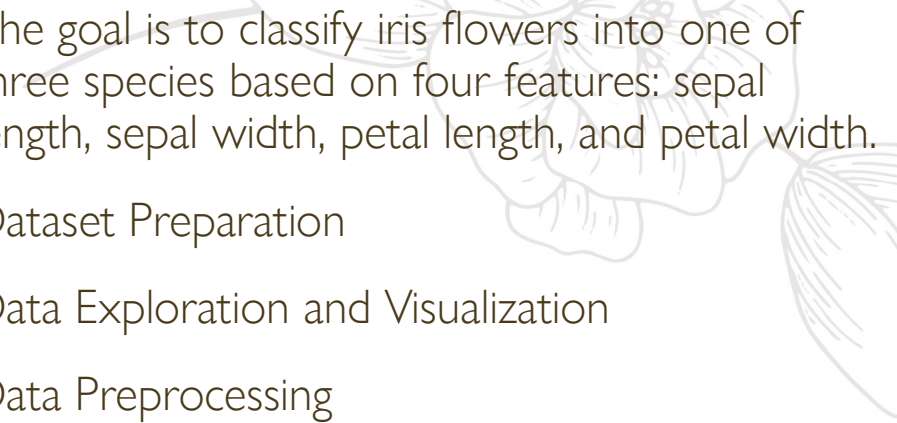


- The model should achieve a high level of accuracy in classifying iris species.
- The model's predictions should be consistent and reliable, as measured by cross-validation.
- The final report should provide clear and comprehensive documentation of the project, including all code, visualizations, and findings

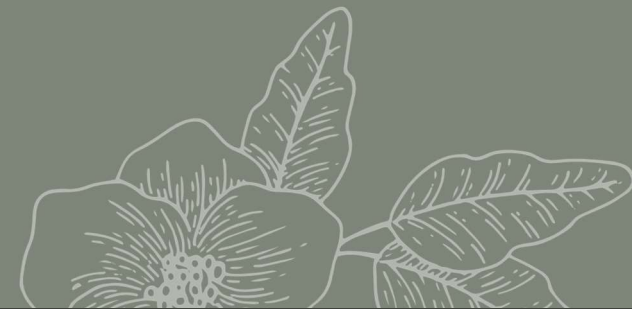


Project Goal



- The goal is to classify iris flowers into one of three species based on four features: sepal length, sepal width, petal length, and petal width.
 - Dataset Preparation
 - Data Exploration and Visualization
 - Data Preprocessing
 - Model Selection and Training
 - Model Evaluation
 - Hyperparameter Tuning
 - Model Interpretation and Insights
- 

About Dataset



It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

Number of Instances: 150

Number of Features: 5

- **Sepal Length** (in cm, numeric)
- **Sepal Width** (in cm, numeric)
- **Petal Length** (in cm, numeric)
- **Petal Width** (in cm, numeric)
- **Species** (categorical: Setosa, Versicolor, Virginica)

Target Variable:

- **Species** (3 classes: Setosa, Versicolor, Virginica)

Feature Types:

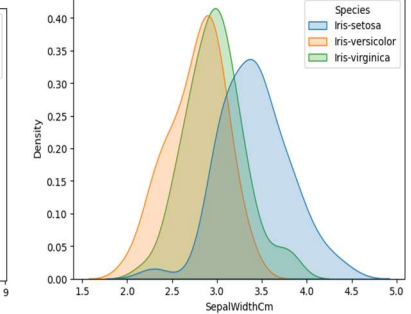
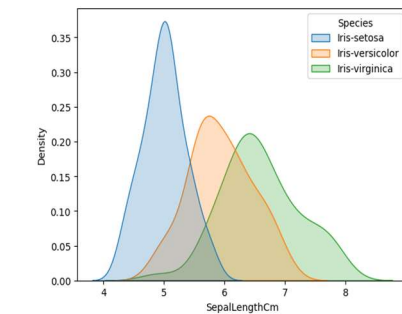
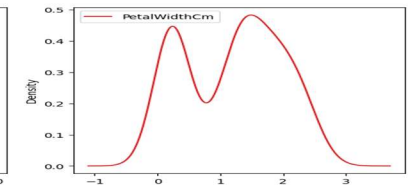
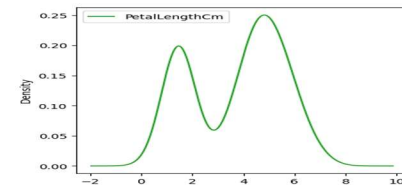
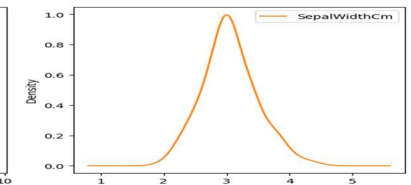
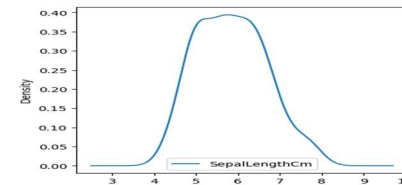
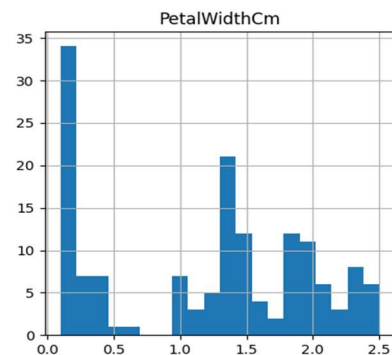
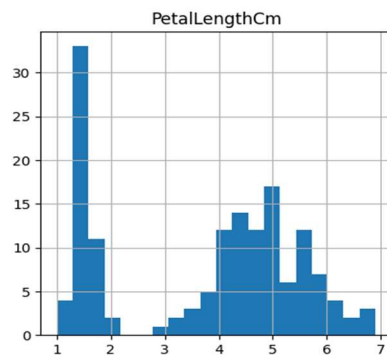
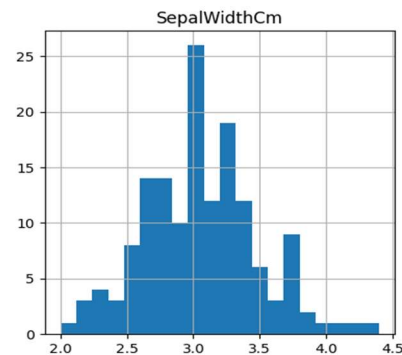
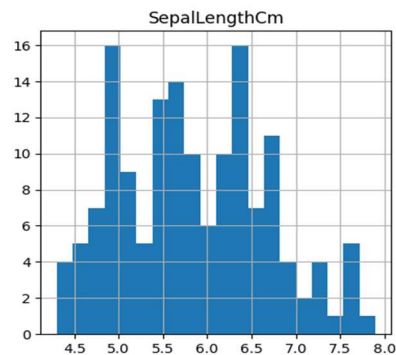
- 4 Continuous Numeric Features
- 1 Categorical Feature (Target)

Missing Values: None

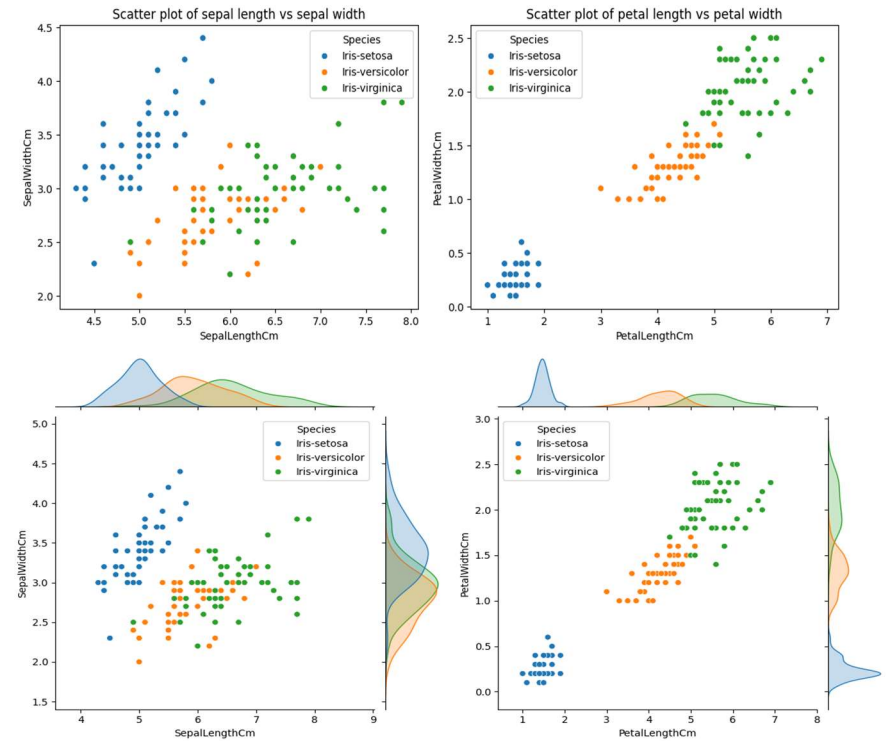
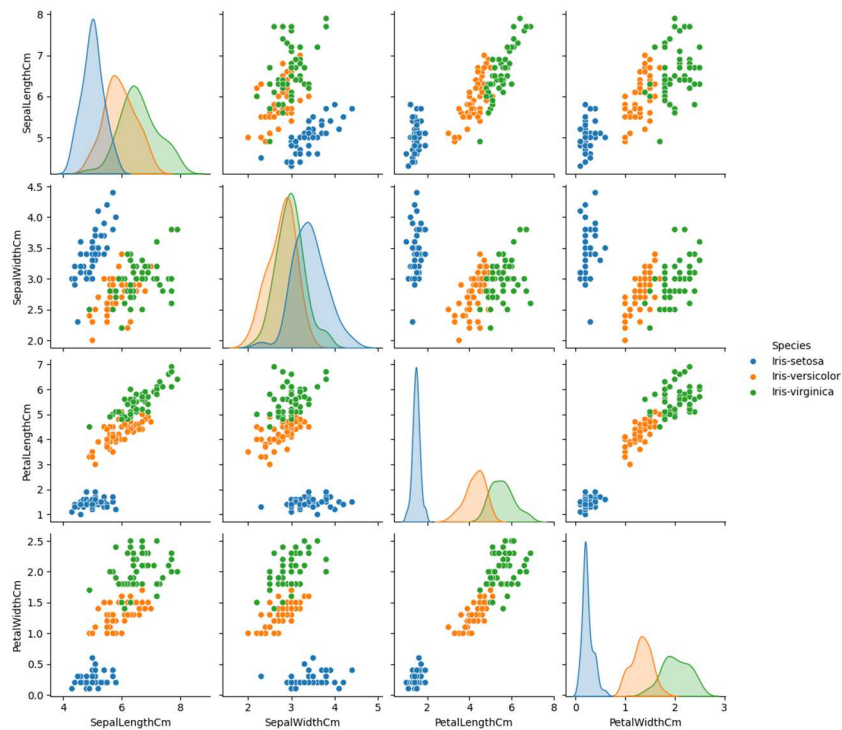
Class Distribution:

- **Setosa:** 50 instances
- **Versicolor:** 50 instances
- **Virginica:** 50 instances

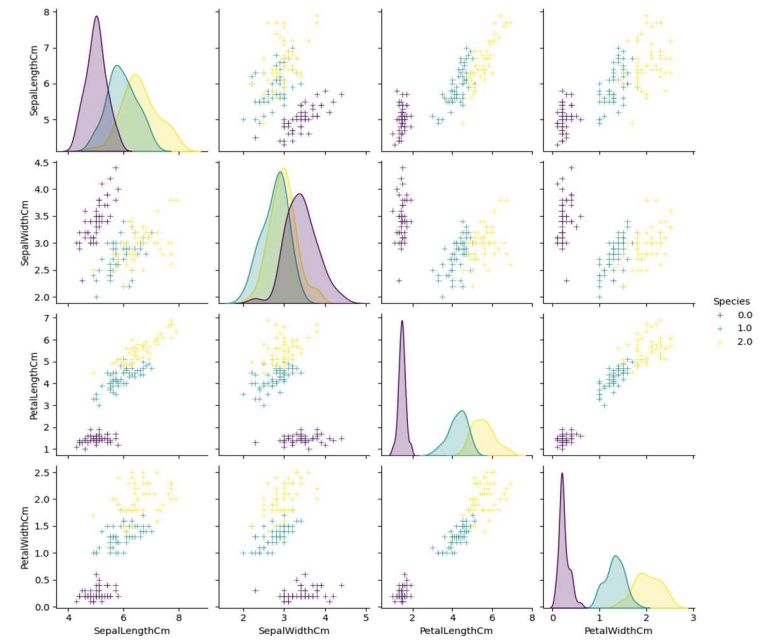
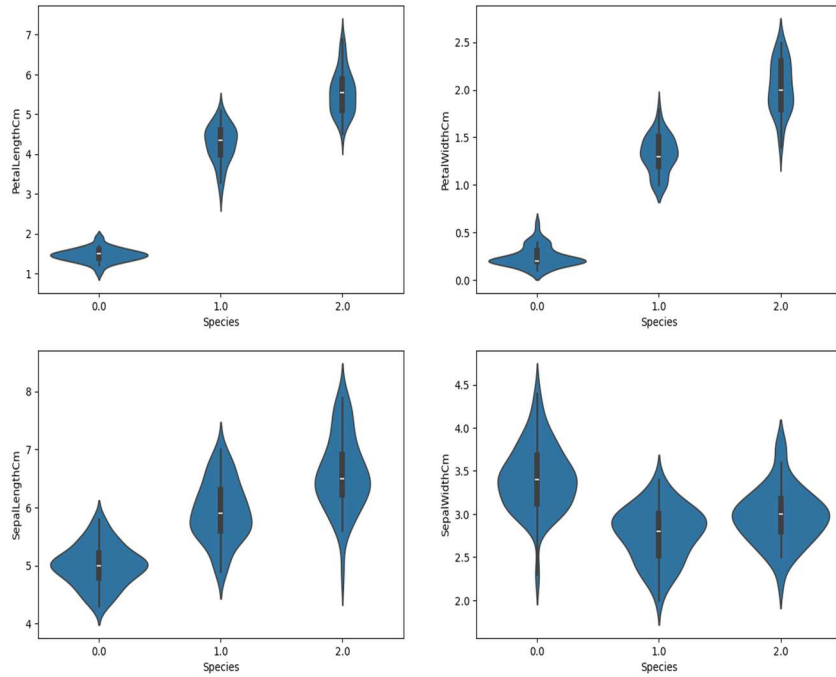
Univariate Analysis (Histograms) & Kernel density estimation (KDE)



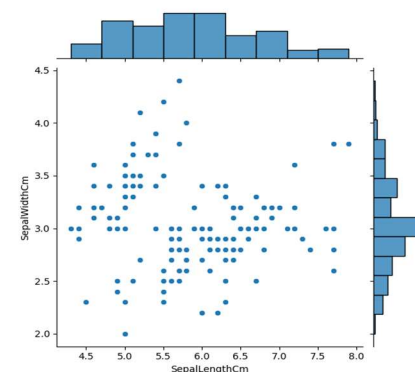
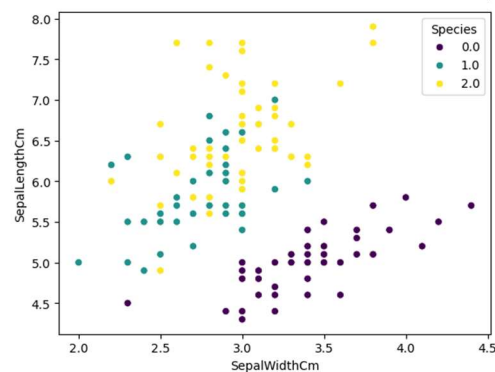
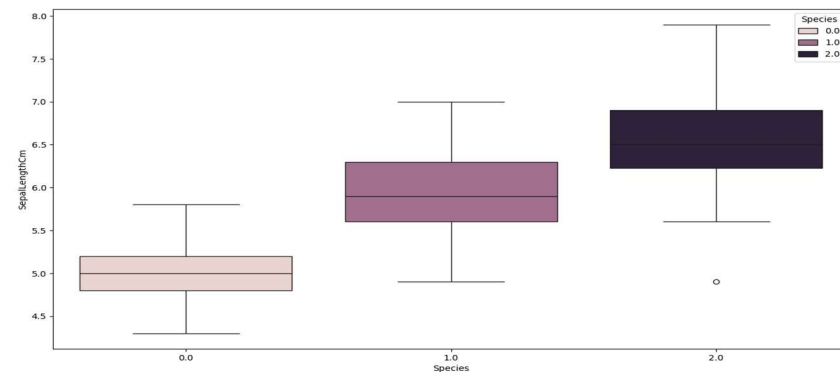
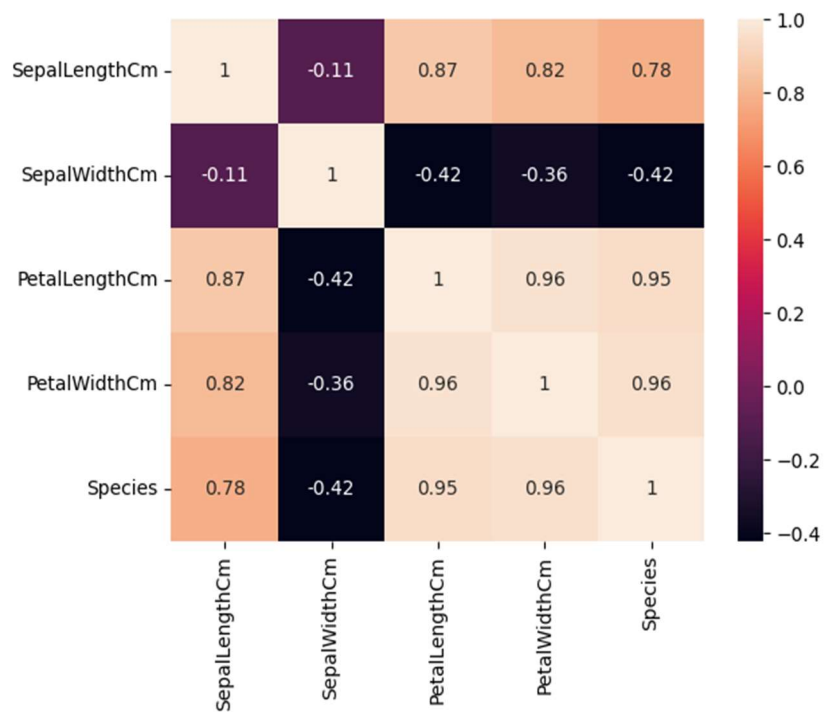
Pairwise plot of all features & Scatter plot and Joinplot



Violinplot & Pairplot



Relationship between features



- ❖ Petal length and petal width are strong indicators for distinguishing species.
- ❖ Setosa is easily separable with smaller petal measurements.
- ❖ Versicolor and Virginica overlap but can be distinguished by petal measurements.
- ❖ Sepal length and sepal width show overlap and are less effective at distinguishing species.
- ❖ Petal length and width have non-overlapping regions for the species.
- ❖ Sepal width has a wide spread, making it less useful in classification.
- ❖ Boxplots show Setosa has a tighter distribution for sepal and petal length, while Versicolor and Virginica show variability.
- ❖ Petal length and width are highly positively correlated.
- ❖ Sepal width negatively correlates with petal measurements, limiting its usefulness.
- ❖ The dataset is balanced across the three species, with no class imbalance.
- ❖ Petal length and petal width show clear clusters for each species, making them useful for classification.
- ❖ Setosa has the smallest sepal length, Virginica the largest, with Versicolor and Virginica showing overlap.
- ❖ Setosa often has a higher sepal width, with significant overlap between Versicolor and Virginica.
- ❖ Sepal width and length show little correlation and overlap between species.
- ❖ Petal length and width show clear species separation, while sepal width has variability.
- ❖ Outliers are visible, especially in Virginica's sepal width, needing attention in preprocessing.
- ❖ Andrews curves show Setosa as distinct, while Versicolor and Virginica overlap.
- ❖ Categorical plots confirm Setosa's clear separation by petal length and width.
- ❖ Violin plots show tightly grouped features for Setosa, while Versicolor and Virginica have wider distributions.
- ❖ Petal measurements are key for classification, while sepal features show overlap between species.
- ❖ Outliers, particularly in Virginica's sepal width, may impact model training.

Insights from EDA on the Iris Dataset

Conclusion and Insights



- **Simple Dataset:** The Iris dataset is small and well-structured, with clear separations between species based on features.
- **Distinct Features:** Petal length and petal width are strong indicators that effectively differentiate the species.
- **Balanced Classes:** The dataset contains an equal number of samples for each species, preventing class imbalance issues.
- **Low Complexity:** The relationships among features are relatively simple, making it easier for models to capture patterns.
- **Effective Algorithms:** All tested algorithms (Logistic Regression, Decision Tree, SVM, k-NN) are capable of perfectly fitting the training data.
- **No Noise:** The dataset appears to have minimal noise or outliers, contributing to clearer decision boundaries for classification.
- **Training Size:** The small size of the dataset allows for memorization by the models, leading to perfect accuracy on the training data.

Google Colab(iris classification):https://colab.research.google.com/drive/1-7rBudadqj0SluY5k0eVm_-fg02pQDyy?usp=sharing

Thank you.

