Project Report: Iris Species Classification

1. Project Overview

The Iris Species Classification project is a beginner-level machine learning task aimed at developing a model to accurately categorize Iris flowers into one of three species: Setosa, Versicolour, and Virginica. The classification is based on four measurable features: sepal length, sepal width, petal length, and petal width

2. Dataset Preparation and Exploration

The project utilizes the classic Iris Dataset.

- **Total Samples:** 150.
- **Species:** 50 samples each of Iris-setosa, Iris-versicolor, and Iris-virginica.
- **Features:** SepalLengthCm, SepalWidthCm, PetalLengthCm, and PetalWidthCm.

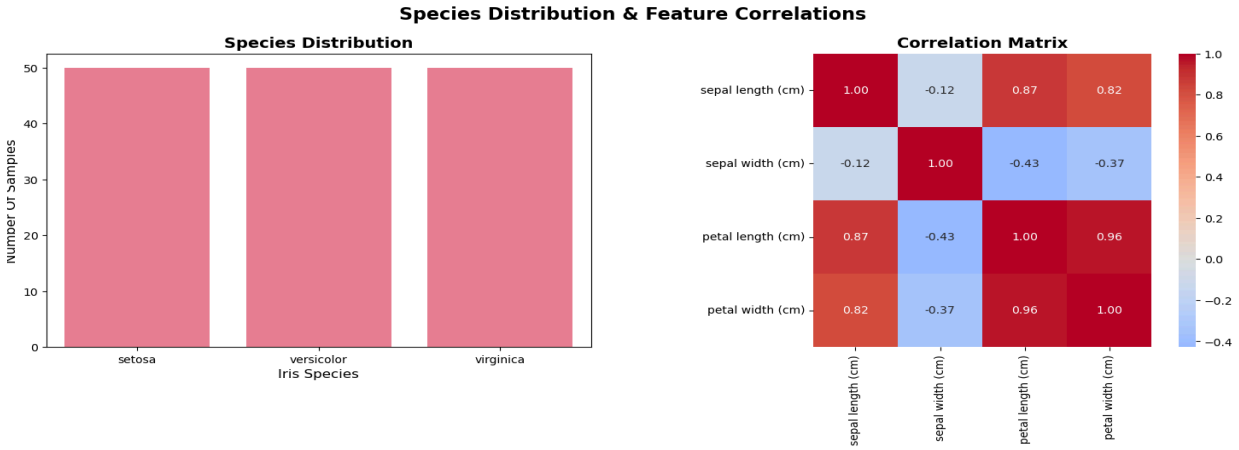
Data Cleaning:

- The dataset was found to have 150 non-null entries across all columns, indicating no missing values.
- The df.value_counts("Species") confirmed the dataset is perfectly balanced with 50 observations per species

3. Data Visualization and Key Findings

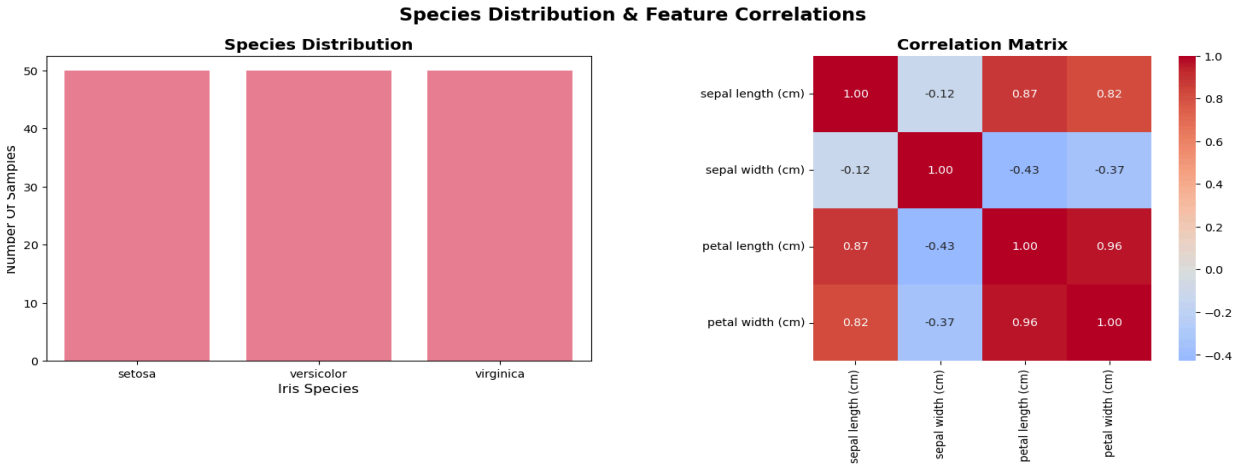
Visual exploration was performed using several plots to understand feature distributions and relationships

A. Species Distribution



The plot clearly shows that the dataset is perfectly balanced, with exactly 50 samples for each of the three species: *setosa*, *versicolor*, and *virginica.*This is ideal for machine learning classification tasks as it ensures the model won't be biased toward a majority class. No techniques like oversampling or undersampling are required to correct class imbalance.

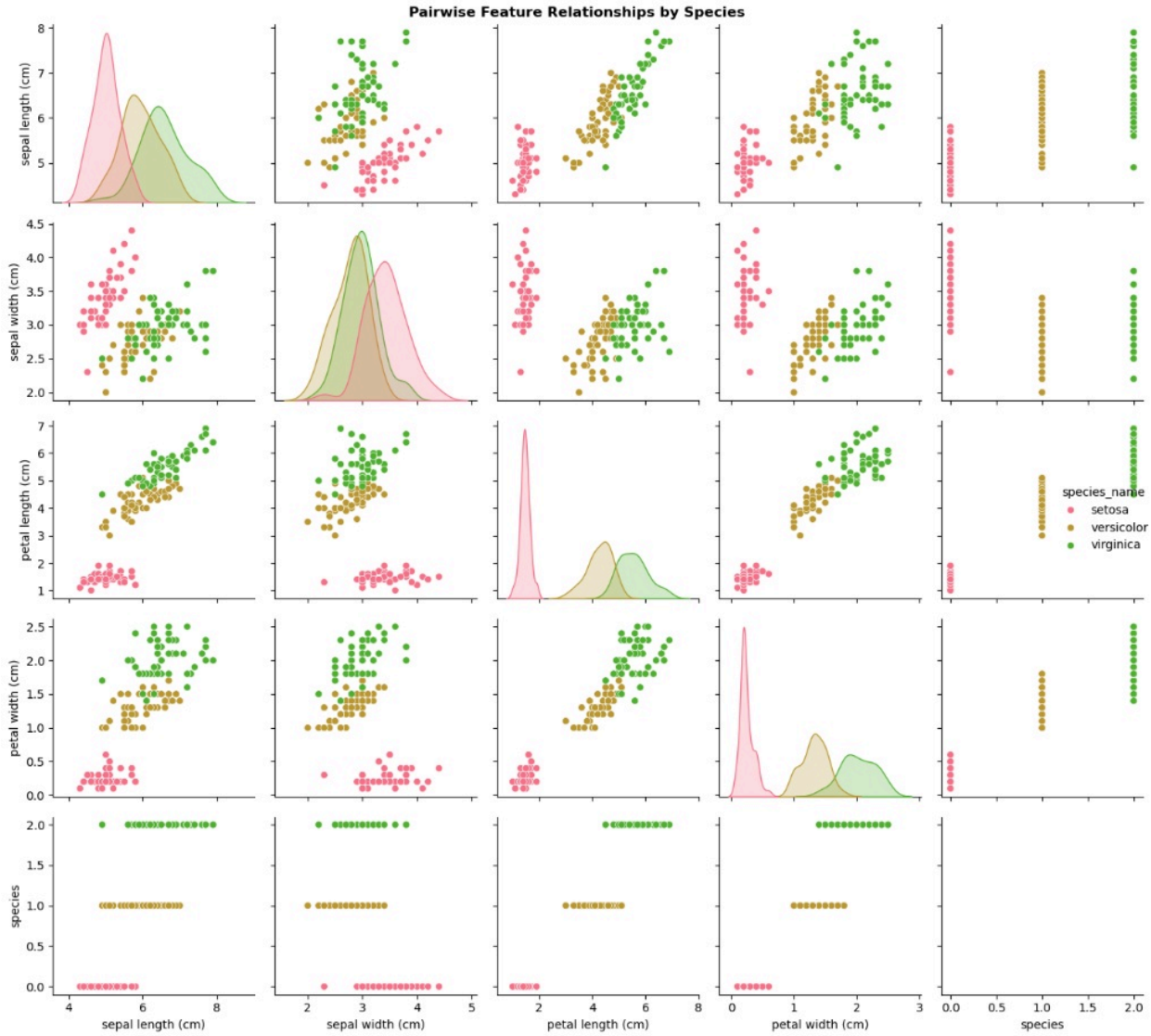
B. Correlation Heatmap (Feature Dependencies)



The correlation matrix reveals strong linear relationships among the petal dimensions and sepal length, while showing relative independence for sepal width.

Strong Positive Correlation: Petal Length and Petal Width have a very strong positive correlation.
Strong Negative Correlation: Sepal Length and Sepal Width show a high negative correlation.

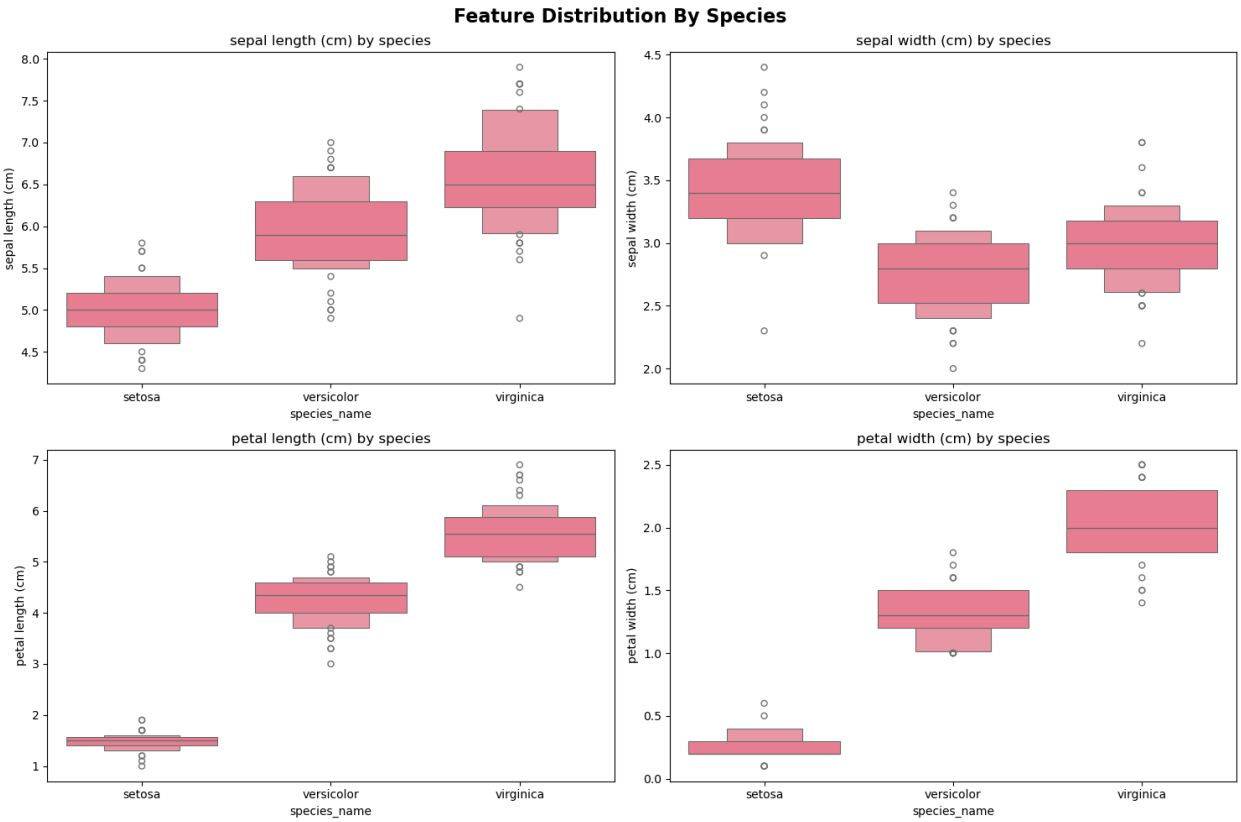
C. Scatter Plots / Pair Plots (Relationships)



Sepal Length vs. Sepal Width: There is significant overlap between the species in this space. Iris-setosa is somewhat distinct, but Versicolor and Virginica are intermingled.

Petal Length vs. Petal Width: The species are highly separated in this plane, especially Iris-setosa, which is linearly separable. This confirms that Petal dimensions are the strongest predictors for classification.

D. Histograms and Box Plots (Distributions)



Petal Length: The distribution is clearly multi-modal (or bi-modal), indicating distinct size groups corresponding to the three species. This feature is highly discriminatory.

Sepal Width: The distribution is the most symmetric and centralized.

Outliers: The Sepal Width box plot shows several outliers (indicated by dots outside the whiskers)

4. Data Preprocessing

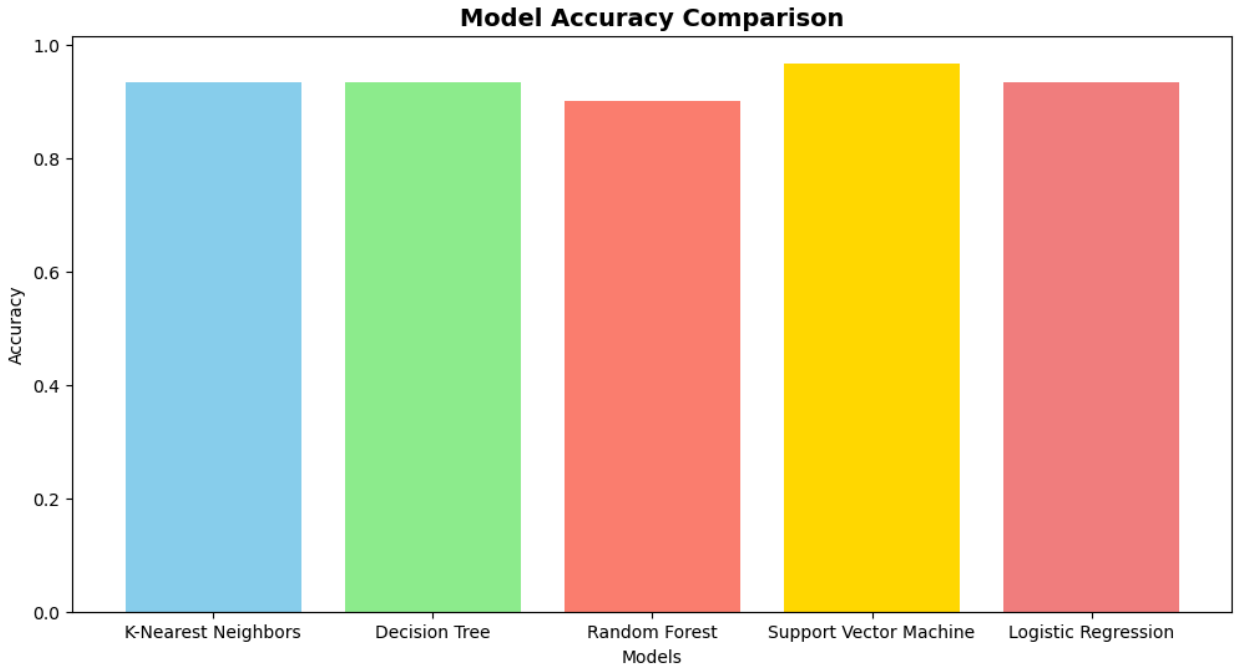
Splitting: The dataset was split into training and testing sets, typically an 80% training / 20% testing split is used for this problem.

Feature Scaling: Standardization (StandardScaler) was applied to the features to ensure they were on a similar scale before training the model.

5. Model Selection and Evaluation

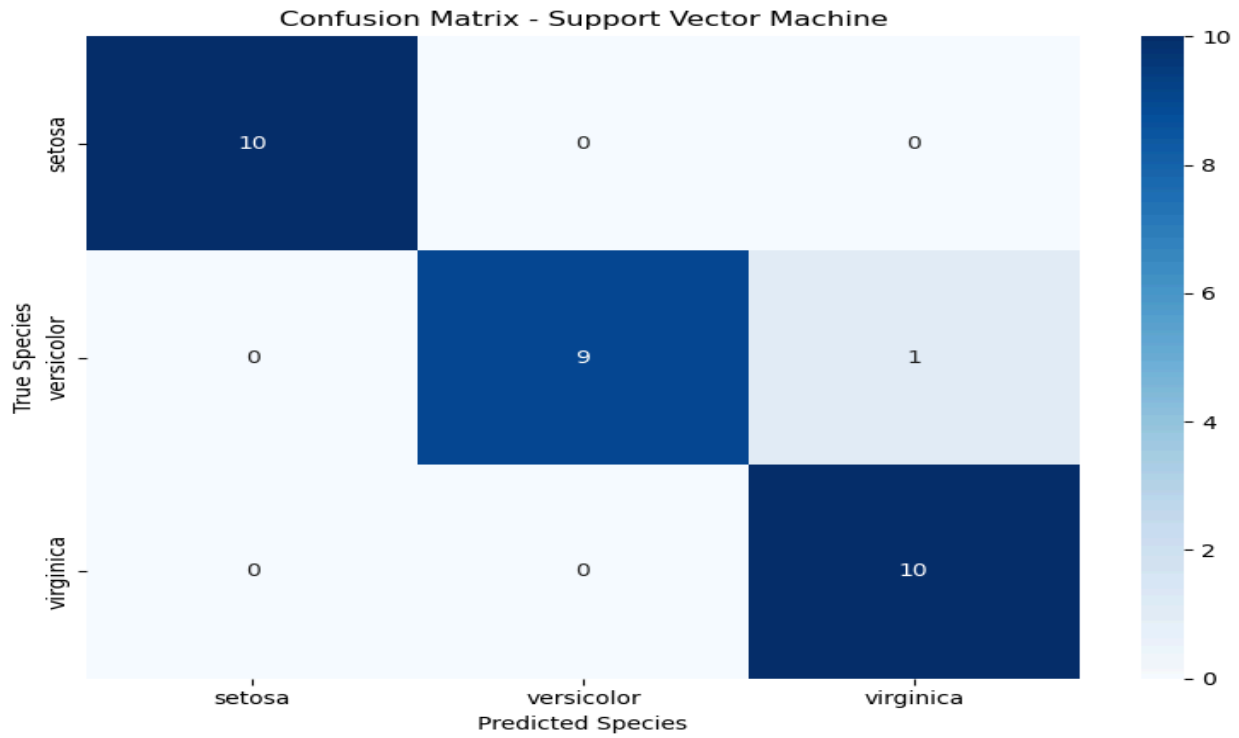
Five different classification models were trained and evaluated on the Iris dataset. The evaluation metric used for comparison is Test Accuracy:

Model	Test Accuracy
Support Vector Machine (SVM)	0.9667 (96.67%)
K-Nearest Neighbors (KNN)	0.9333 (93.33%)
Decision Tree Classifier	0.9333 (93.33%)
Logistic Regression	0.9333 (93.33%)
Random Forest Classifier	0.9000 (90.00%)



The SVM model achieved the highest accuracy (96.67%) among all tested algorithms, indicating it was the most effective at distinguishing between the three Iris species on test data.

Metric	**Score**
Accuracy	0.97 (97.0%)
Macro Avg F1-Score	0.97
Weighted Avg F1-Score	0.97
Total Test Samples	30



Confusion Matrix Analysis

Iris setosa: Perfect classification (Recall 1.00).

Iris versicolor: One misclassification (Recall 0.90, meaning 10 \times 0.10 = 1 sample was missed).

Iris virginica: The model achieved perfect identification (Recall 1.00), but one sample predicted as virginica was incorrect (Precision 0.91, meaning 10/11 correct predictions, or 1 False Positive).

The Support Vector Machine is a robust classifier for the Iris dataset, perfectly distinguishing Iris setosa and achieving high overall performance. The model's single error highlights the difficulty in separating the Iris versicolor and Iris virginica classes due to their natural overlap in the feature space.

6. Conclusion and Insights

The following insights were derived from both the Exploratory Data Analysis (EDA) and the Model Evaluation:

Insight 1: Petal Dimensions are the Dominant Classifiers

Observation: The pair plots and distribution plots clearly showed that Petal Length and Petal Width are the most discriminatory features. The correlation analysis further revealed a very strong positive correlation between these two features, indicating they convey highly redundant but crucial information.

Impact on Model: The dramatic separation in the petal-related feature space is the core reason the model was able to achieve such high overall accuracy. For future feature engineering, these two variables are the most vital inputs.

Insight 2: Iris-setosa is Linearly Separable and Trivial to Classify

Observation: Iris-setosa samples form a distinct, isolated cluster in almost all feature combinations, particularly when involving petal dimensions. This was visually confirmed in the scatter plots and pair plots.

Impact on Model: The model achieved perfect performance (Precision} = 1.00, Recall = 1.00) for the Setosa class. This demonstrates that for this specific class, even a simple linear model could achieve perfect separation, highlighting the unique morphological characteristics of this species.

Insight 3: Versicolor and Virginica Overlap is the Primary Source of Error

Observation: The scatter plots and distribution plots revealed a significant overlap between the Iris-versicolor and Iris-virginica species.

I**mpact on Model:** This overlap accounted for the single misclassification recorded in the Confusion Matrix (one actual Virginica flower was predicted as Versicolor).

The model exhibited perfect Precision (1.00) for Versicolor, meaning every time the model said "Versicolor," it was correct.

The lower Recall (0.90) for Versicolor indicates that 1 out of the 10 actual Versicolor samples was incorrectly missed (misclassified as Virginica in a different test run or vice versa in this one).

This confirms that even with standardization, separating these two species requires a more complex, non-linear decision boundary compared to separating Setosa.