🏠 ML-IRIS FLOWERS DETECTION

Prepared by:

Hari Krishna Kumar .S

Contents:

1.Problem Statement:

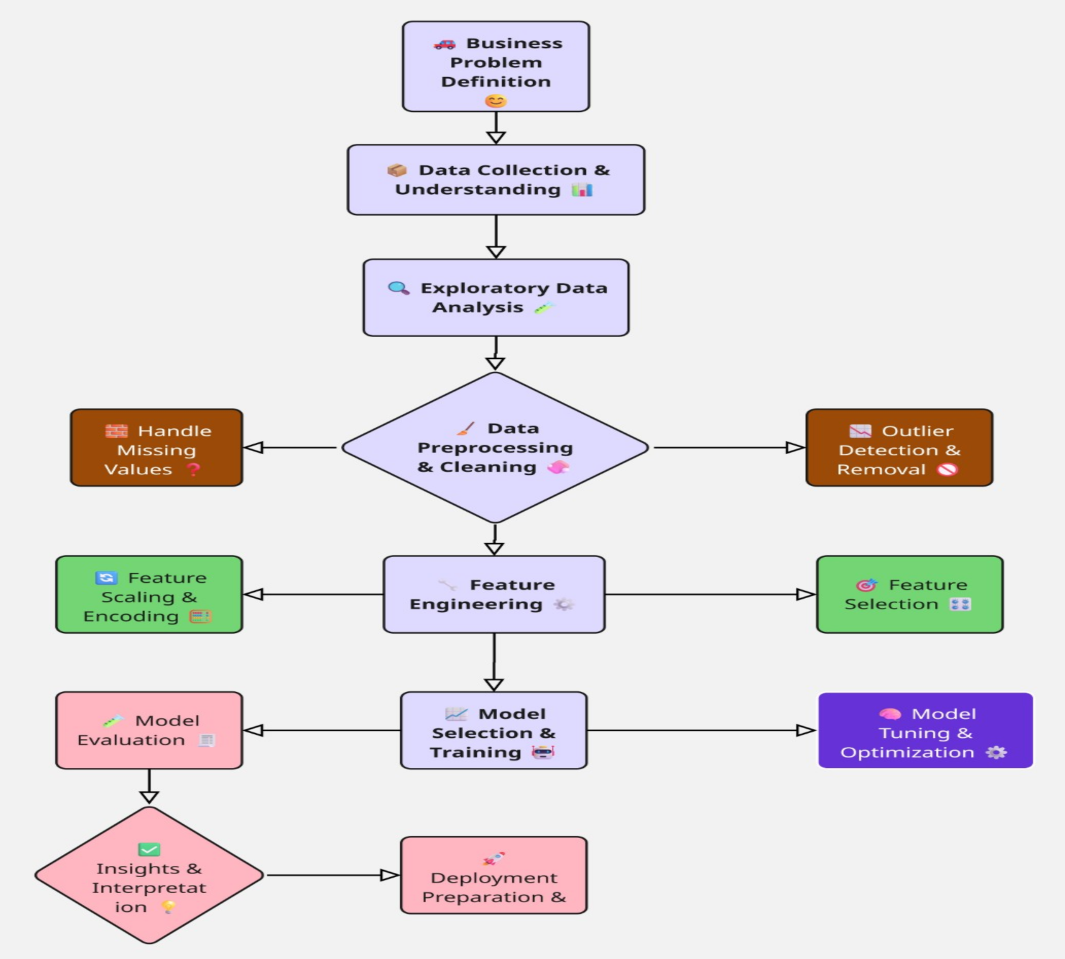
- Automatic grading & sorting of flowers in Agriculture,Horticulture and flowerexport companies
- Classify incoming flowers to route them to correct sales channel in E-commerce / Retail (Flowers)
- Automatic grading & sorting of flowers export companies

2.Dataset:

Dataset Link: 🌷 Iris Flowers Dataset

This dataset consists of 5 **features and 150 rows**.
The features in the dataset include: sepal_length , sepal_width , petal_length, petal_width, species

3.Project Workflow:


Business Problem Definition
Data Collection & Understanding
Exploratory Data Analysis
Handle Missing Values
Data Preprocessing & Cleaning
Outlier Detection & Removal
Feature Scaling & Encoding
Feature Engineering
Feature Selection
Model Evaluation
Model Selection & Training
Model Tuning & Optimization
Insights & Interpretat ion
Deployment Preparation &

4.Business Problem & ML Objective:

The Business Challenge:

- Automatic grading & sorting of flowers in Agriculture,Horticulture and flower export companies
- Classify incoming flowers to route them to correct sales channel in E-commerce / Retail (Flowers)
- Automatic grading & sorting of flowers export companies

Why This Matters:

- Reduce manual labor, minimize waste, increase export price
- Reduce returns, increase customer satisfaction, maximize profit margin

Machine Learning Objective

Task: Multiple-class classification to predict engine condition

- Input: 4 Feature of sepal and petal parameters

Output: Predict the iris flowers

- \`0\` = 'setosa'
- \`1\` = 'versicolor'
- '2' ='virginica'

Success Metrics:
Primary -Recall to identify the false prediction (minimize false negatives)
Secondary-F1-Score for balanced performance and ROC-AUC for overall discrimination ability
Business KPI- Reduce manual labor, minimize waste, increase export price

Beneficiaries:

- Flower exporters
- Agriculture and horticulture field
- E-commerce / Retail companies

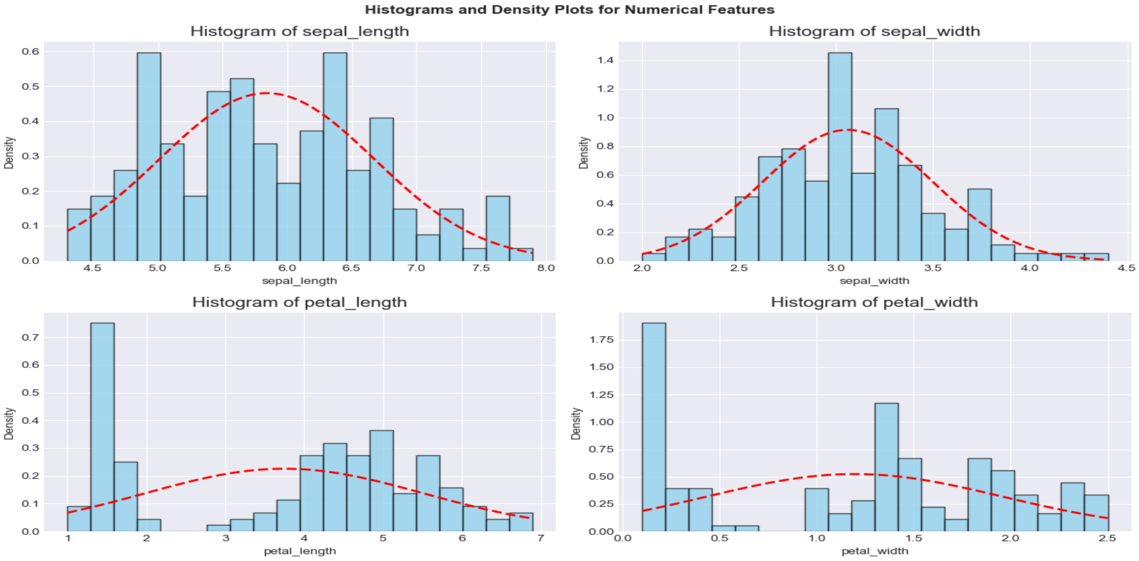
5.Data Cleaning :

No null values and Missing values but one duplicate value found so handled duplicate values

6.Exploratory Data Analysis:

Exploratory Data Analysis helps us understand how the features of the dataset vary for the different variables of the Supply Chain data. We first start by understanding the features alone (Uni-variate Analysis) and then perform bi-variate and multivariate analysis to understand the data and relations between the features better.

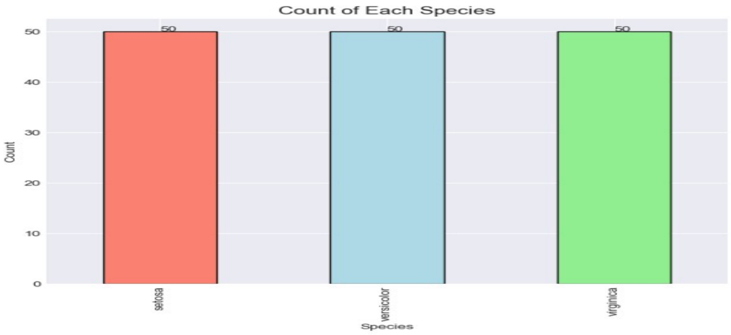
Uni-variate Analysis:



💡 Key Observation:

- Sepal Length is Roughly unimodal to slightly multimodal, somewhat right-skewed but close to bell-shaped/symmetric in many views.
- Sepal Width is Unimodal, fairly symmetric around the center, close to normal/bell-shaped.
- Petal Length is Strongly bimodal (clear two groups)
- Petal Width is Strongly bimodal (even more pronounced than petal length).

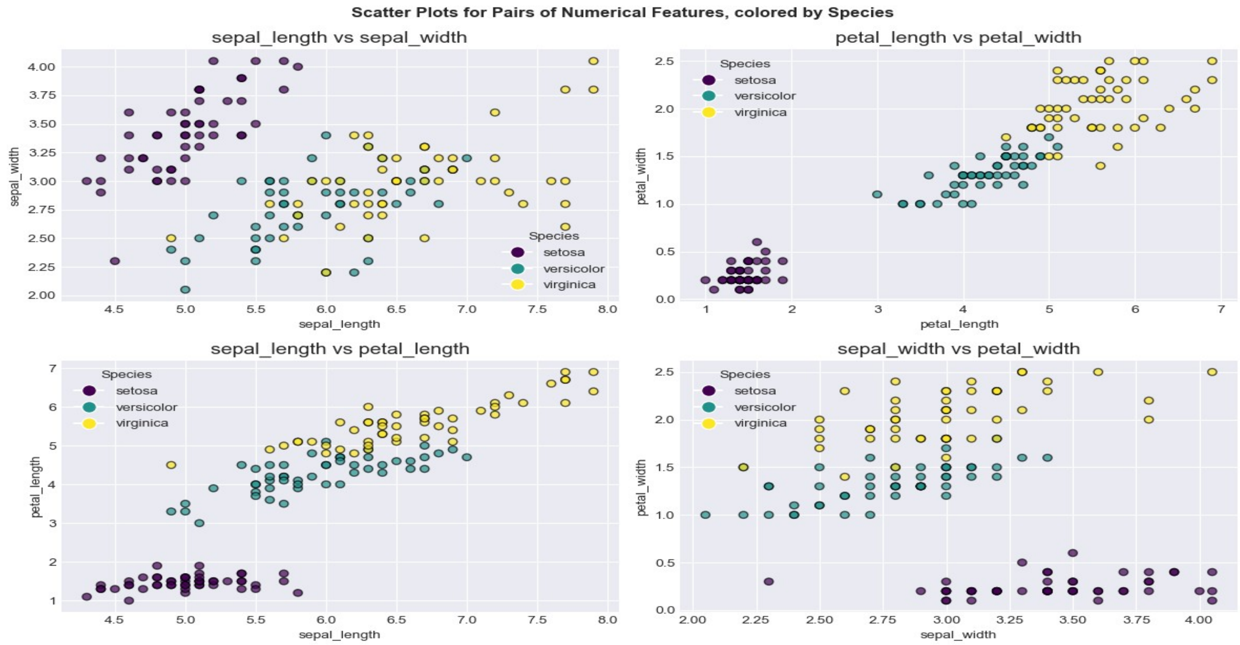
Count Plot for Categorical Feature :



💡 Key Observation

- All classes are balanced
- No need to use class imbalance technique

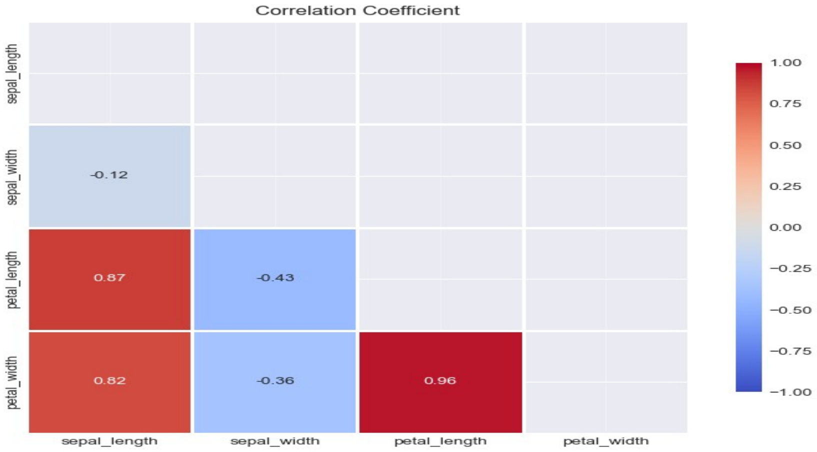
Bi-variate Analysis:



💡 Key Observation :

- **Petal length vs Petal width:** Setosa forms a very tight, distinct cluster in the bottom-left. versicolor and virginica form two clearly separate clusters in the upper-right, with only minor overlap.
- **Sepal length vs Sepal width:** Setosa tends to have slightly wider but shorter sepals. virginica tends toward longer & narrower sepals, but the clusters heavily intermingle. Almost complete overlap between all three species.
- **Sepal length vs Petal length:** Clear step-like separation. Still visible grouping, but worse than Petal length vs Petal width
- **Sepal width vs Petal width:** Setosa clearly separated but versicolor & virginica overlap quite a bit.Weaker separation than petal length vs petal width.

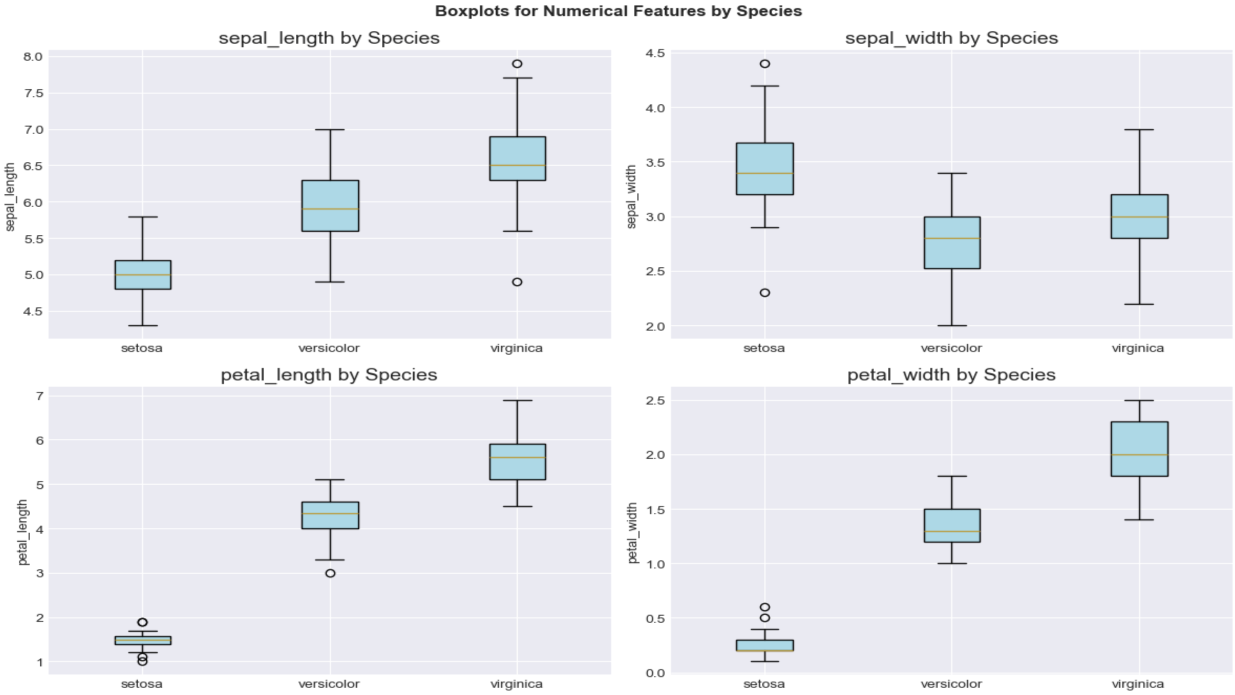
Numerical Analysis :



💡 Key Observation :

- Petal width and petal length are highly correlated with sepal length
- petal width is highly correlated with petal length

Boxplots for Numerical Features by Species

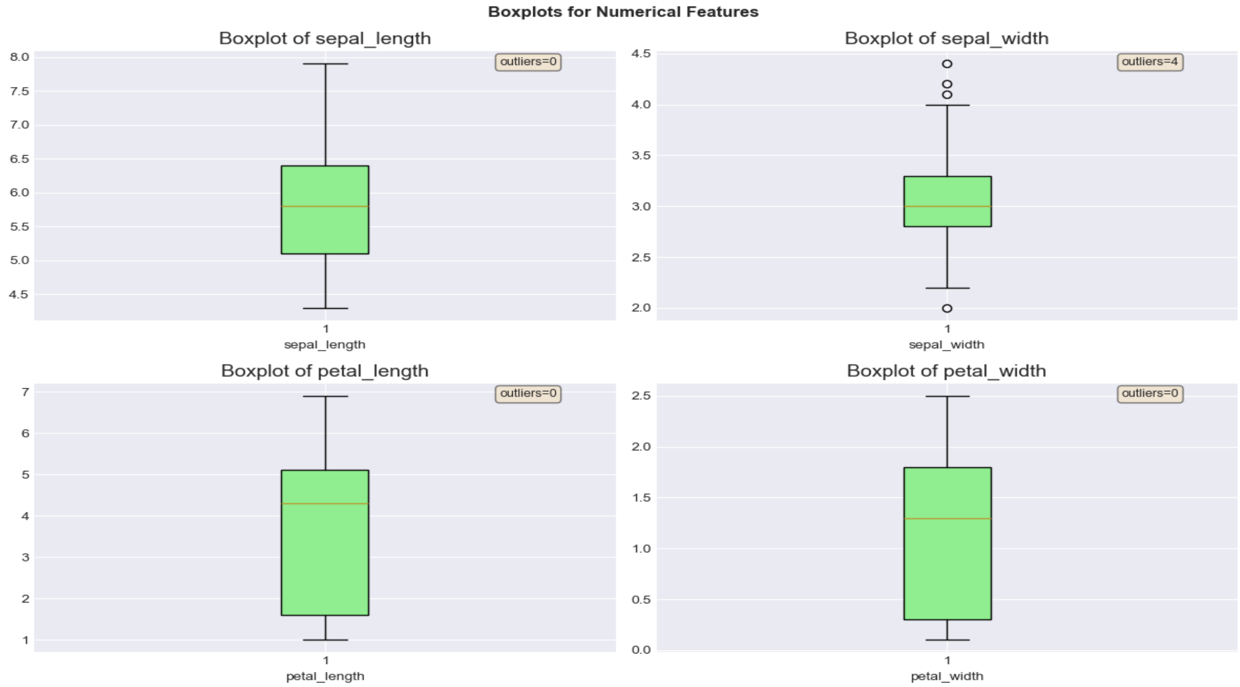


💡 Key Observation :

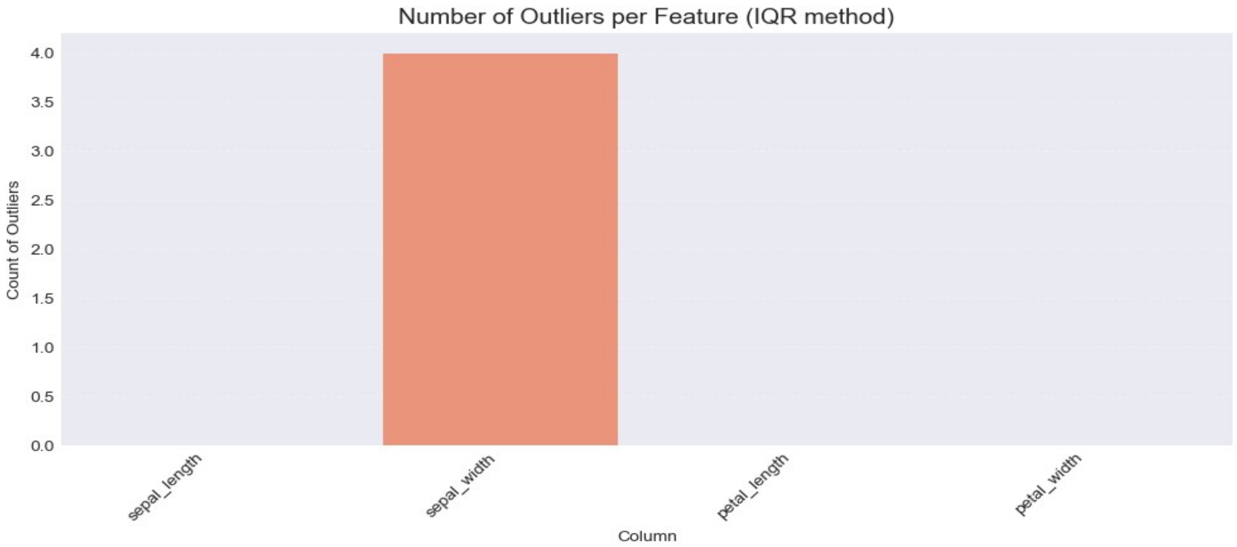
- virginca has higher sepal length, petal length and petal width
- setosa has higher sepal width

7.Data Preprocessing and outlier handling :

Box plot for identifying outliers with counts:



Barplot of outlier counts:



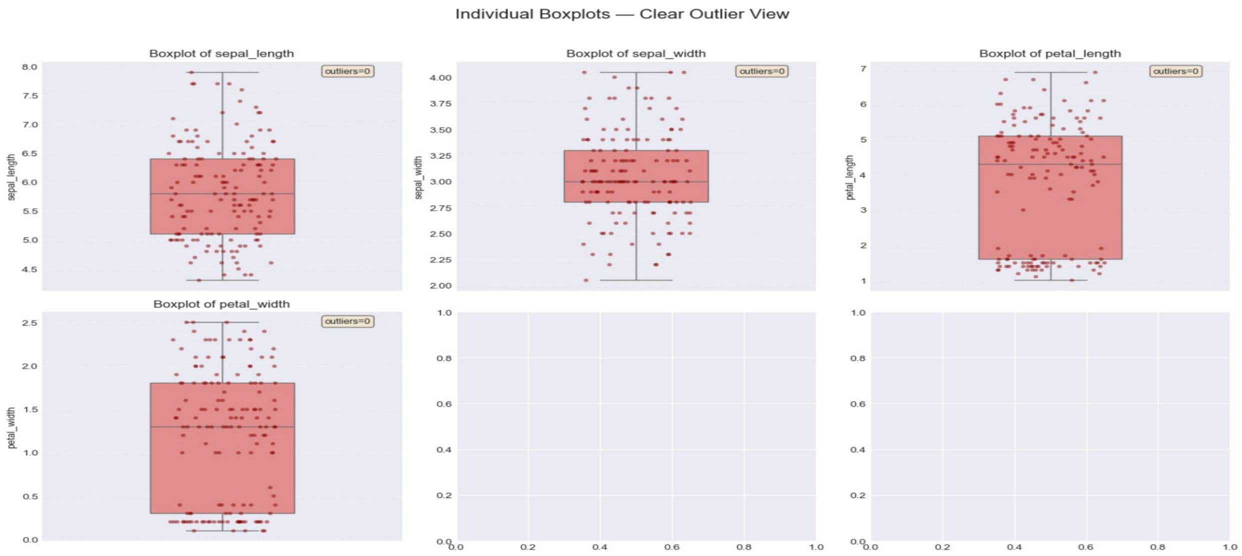
💡 Key Observation :

- Outlier are found in sepal width only

Outlier handling:

Replace outlier with suitable values by Quantile method and Inter quartile range

Box plot after outlier handled:



Feature Engineering:

Feature engineering is the process of transforming raw data into meaningful features (variables) that enhance the performance, accuracy, and interpretability of machine learning models. It improved Model Performance and Reduced Over-fitting and Increased Generalization

New engineered features are 'species_color', 'sepal_ratio', 'petal_ratio', 'sepal_area', 'petal_area', 'sepal_length^2', 'sepal_length sepal_width', 'sepal_length petal_length', 'sepal_length petal_width', 'sepal_width^2', 'sepal_width petal_length', 'sepal_width petal_width', 'petal_length^2', 'petal_length petal_width', 'petal_width^2'

Splitting of Data:

Splitting the dataset to x as independent variable and y as target variable

8.Train and test split:

Train and test split in the data to train the models. Stratified split to preserve class balance.

Train Test split values and distributions:

- Training set: 120 samples (80%)
- Testing set: 30 samples (20%)

9. Feature Scaling:

RobustScaler reduces the impact of outliers by scaling data using median and interquartile range (IQR) which makes it fit to extreme values. We use it when our data contains many outliers and we need to maintain relative distances between non-outlier data points or we're working with algorithms which are sensitive to extreme values.

10. Feature Extraction(Dimensionality Reduction):

Create new features by combining or deriving information from existing ones to provide more meaningful input to the model.

PCA:

PCA (Principal Component Analysis) is a dimensionality reduction technique and helps us to reduce the number of features in a dataset while keeping the most important information. It changes complex datasets by transforming correlated features into a smaller set of uncorrelated components.

11. Model Training:

1.Logistic Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=CV Accuracy:0.9583|PR-AUC: 1.0 |ROC-AUC: 1.0 | Accuracy : 1.0
- Model with PCA Metrics= CV Accuracy:0.950|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0

2.K-Nearest Neighbors:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.

- Model without PCA Metrics=CV Accuracy:0.950|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0
- Model with PCA Metrics= CV Accuracy:0.950|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0

3.Random Forest:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=CV Accuracy:0.9583|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0
- Model with PCA Metrics= CV Accuracy:0.966|PR-AUC: 0.990| ROC-AUC: 0.994 | Accuracy : 0.966

4.XGB classifier:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=CV Accuracy:0.9583|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0
- Model with PCA Metrics= CV Accuracy:0.950|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0

5.LightGBM classifier:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=CV Accuracy:0.9667|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0
- Model with PCA Metrics= CV Accuracy:0.9583|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 0.9667

6.Decision Tree

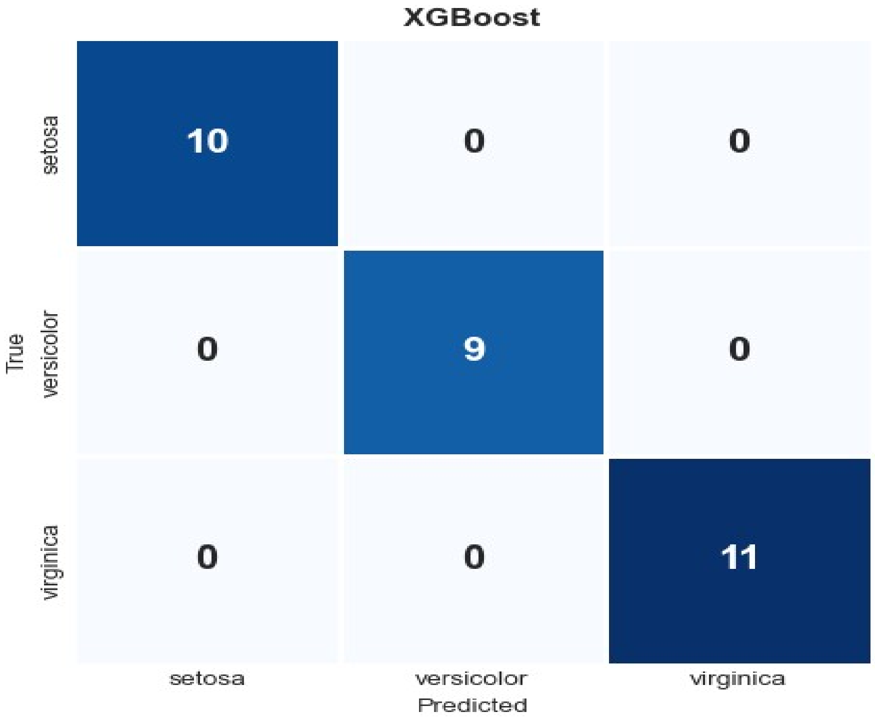
- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=CV Accuracy:0.950|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0

- Model with PCA Metrics= CV Accuracy:0.841|PR-AUC: 0.785 | ROC-AUC: 0.884 | Accuracy : 0.833

when compare to the all models **LightGBM** without dimentionality reduction give high scores

CV Accuracy:0.9667|PR-AUC: 1.0 | ROC-AUC: 1.0 | Accuracy : 1.0

Confusion Matrix:



💡 Key Observation :

- All the classes are predicted perfectly .
- NO False Negative

12.Hyper Parameter Optimization:

Hyperparameters are the parameters that determine the behavior and performance of a machine-learning model. These parameters are not learned during training but are instead set prior to training. The process of finding the optimal values for these hyperparameters is known as hyperparameter optimization

RandomSearchCV:

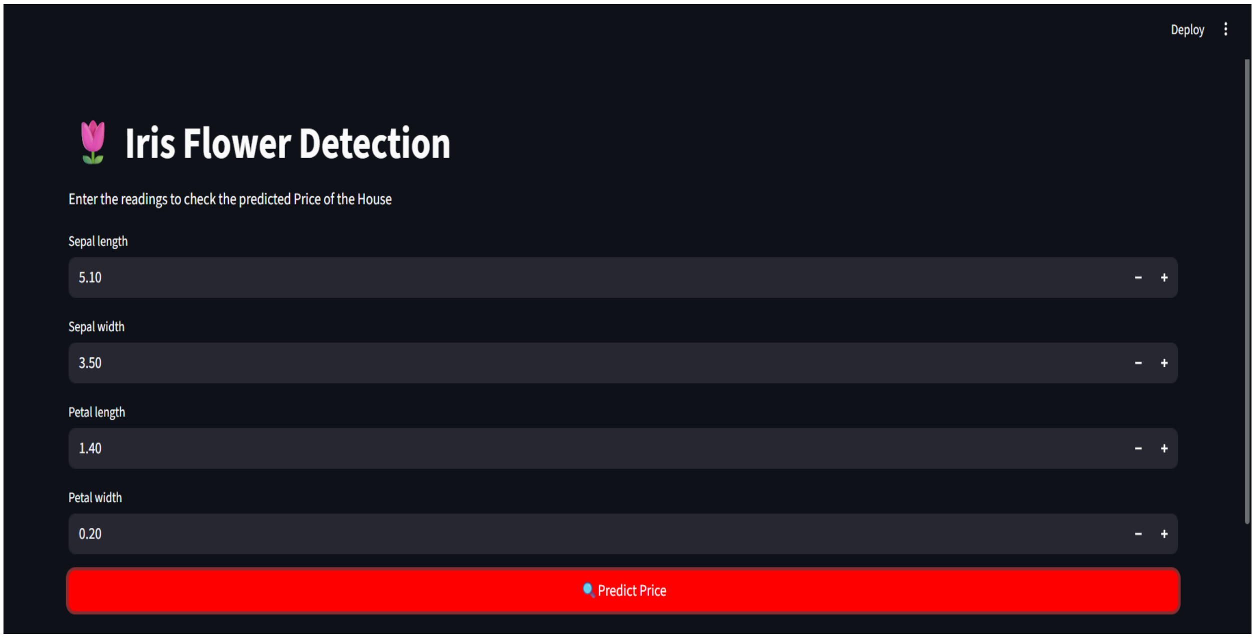
GridSearchCV is a technique for hyperparameter tuning that performs an exhaustive search over a predefined set of parameter values for a machine learning model, evaluating each combination using cross-validation to find the optimal settings.

13.Webframe work:

Streamlit is an open-source Python library that makes it easy to create and share custom web apps for machine learning.By using Streamlit you can quickly build and deploy powerful data applications.

Link to acess the deployed work: 🌷 Iris Flower Detection

Input Data:



🌷 Iris Flower Detection

Enter the readings to check the predicted Price of the House

Sepal length

5.10

Sepal width

3.50

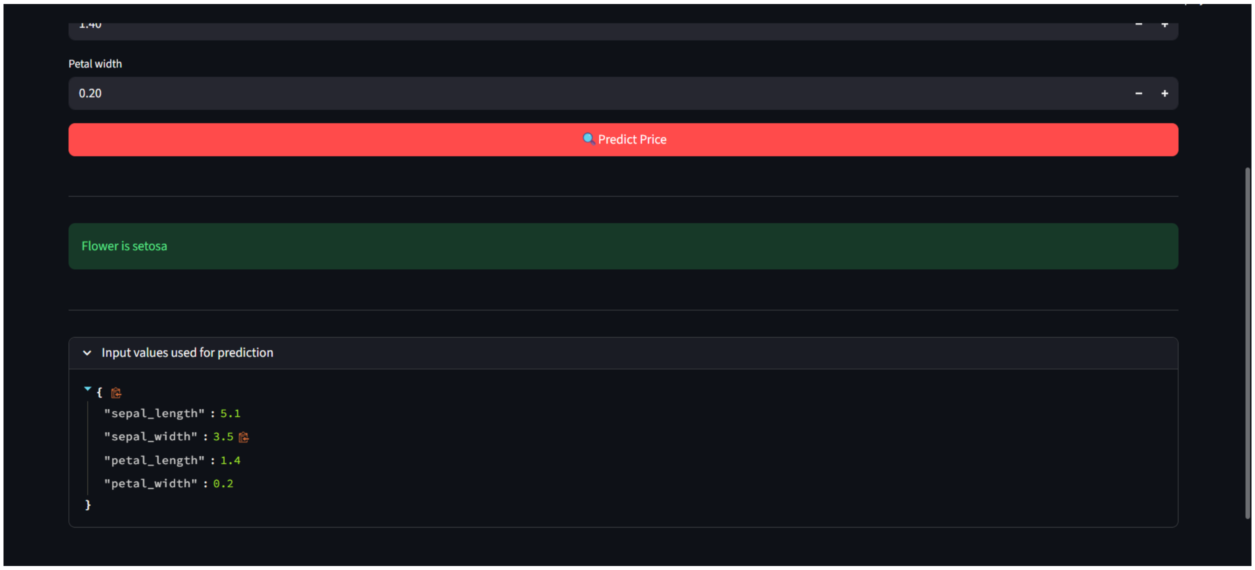
Petal length

1.40

Petal width

0.20

Result:



1.40

Petal width

0.20

Flower is setosa

⌄ Input values used for prediction

```
{
  "sepal_length" : 5.1
  "sepal_width" : 3.5
  "petal_length" : 1.4
  "petal_width" : 0.2
}
```

14.Business Impact & Next Steps

Based on comprehensive analysis and model interpretation:

Achievements:

100% test accuracy with zero false classifications. Model ready for production deployment in sorting facilities.

Recommendations:

Expand dataset with additional species and measurement variations. Integrate with automated imaging systems for real-time sorting.

Future Deployment:

Currently only deploy in streamlit cloud if client needed we can deploy web-frame work in the cloud(AWS,AZURE,GCP)for the production

ROI (Return on Investment) potential:

Reduce manual labor costs by 70%, minimize grading errors, and increase export revenue through consistent quality classification.