IML PROJECT REPORT

FLOWER CLASSIFICATION MODEL

Prof. Avinash Sharma

GROUP TA NAME :JYOTISHMAN DAS

GROUP MEMBERS NAME :
Lagna Priyadarshini (B23CI1022)
Anjali Mehra (B23BB1008)
Esha Mandal (B23PH1008)
Jeetu Gurjar (B23ME1025)

1. INTRODUCTION

In this project, we aim to build a machine learning classification model to predict the species of flowers based on their physical characteristics. The model will use four key features of the flower: sepal length, sepal width, petal length, and petal width. The species to be predicted include Setosa, Versicolor, and Virginica. This classification task is based on the widely recognized Iris dataset, which is ideal for experimenting with supervised classification techniques like KNN classifier,logistic regression and SVM.

2. MOTIVATION

By focusing on flower classification, this project provides an opportunity to apply and understand core machine learning principles such as data preprocessing, feature selection, model training, and testing. Additionally, this project demonstrates how machine learning models can process large datasets and generalize to new data, laying the groundwork for more complex predictive tasks in various domains beyond botany, including healthcare, environmental monitoring, and industrial applications.

3. PROBLEM STATEMENT

Given four input features, we aim to classify each observation into one of the three flower species. This problem serves as an introduction to multi-class classification using simple yet effective machine learning techniques.

4. BACKGROUND STUDY

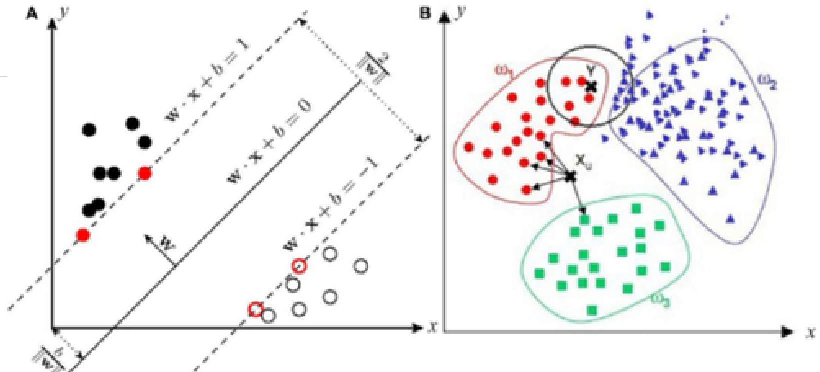


Fig: Description of knn and svm

k-Nearest Neighbors (k-NN)

Concept: A non-parametric algorithm that classifies data points based on the majority label of their 'k' nearest neighbors.

Theory: Relies on distance metrics (e.g., Euclidean) to determine proximity, with the choice of 'k' influencing sensitivity to noise and generalization.

Framework: Does not require a training phase but stores the entire dataset, making it suitable for smaller datasets; evaluated using cross-validation and confusion matrices.

Support Vector Machines (SVM)

Concept: A classification technique that finds the optimal hyperplane to separate different classes in high-dimensional space, maximizing the margin from support vectors.

Theory: Handles both linear and non-linear classification through kernel functions, allowing for complex decision boundaries.

Framework: Robust against overfitting, especially in high dimensions, employing optimization to determine the best hyperplane, evaluated using accuracy, F1-score, and ROC-AUC.

5. RELATED WORK

kNN:
https://drive.google.com/file/d/1cCVnt2eAC9VqtOLpMqpiKFS2LFvLQJhS/view?usp=sharing

SVM:
https://drive.google.com/file/d/1qVGVMzXZrLJx6dJwao3zQ_sLw_lEamzk/view?usp=sharing

6. ABOUT THE DATASET

The Iris dataset consists of 131 samples, Each sample includes:

Sepal Length,Sepal Width, Petal Length, Petal Width

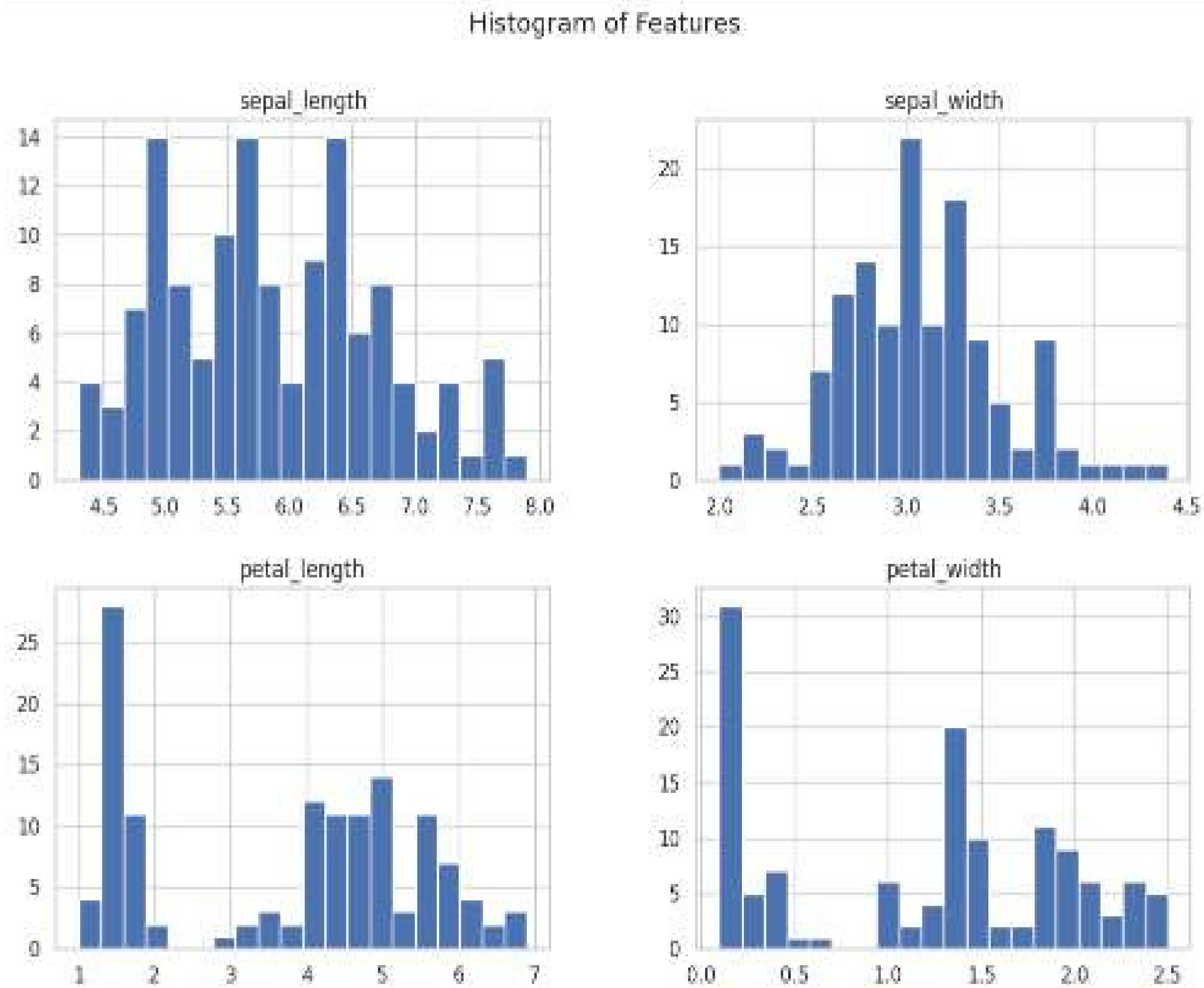
```
RangeIndex: 131 entries, 0 to 130
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal_length  131 non-null    float64
 1   sepal_width   131 non-null    float64
 2   petal_length  131 non-null    float64
 3   petal_width   131 non-null    float64
 4   species       131 non-null    object
dtypes: float64(4), object(1)
memory usage: 5.2+ KB
None
       sepal_length  sepal_width  petal_length  petal_width
count    131.000000   131.000000    131.000000   131.000000
mean       5.853435     3.068702      3.745038     1.183206
std        0.846558     0.434136      1.792729     0.764416
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.300000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
species
Iris-setosa        45
Iris-virginica     44
Iris-versicolor    42
Name: count, dtype: int64
```

7. SYSTEM ARCHITECTURE

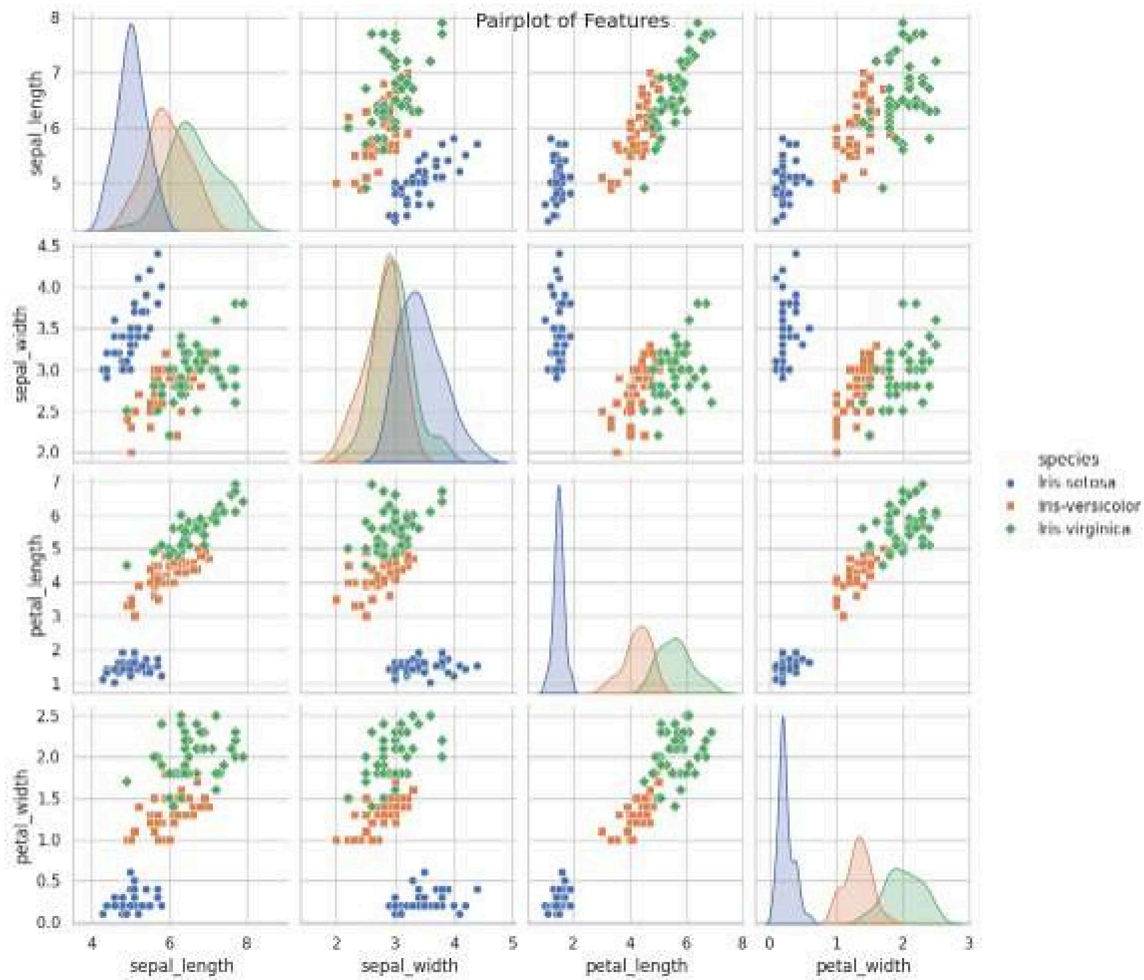
The machine learning pipeline for this project follows these steps:

Data Preprocessing: Clean and prepare the data for modeling, including normalization and splitting the data into training and testing sets.

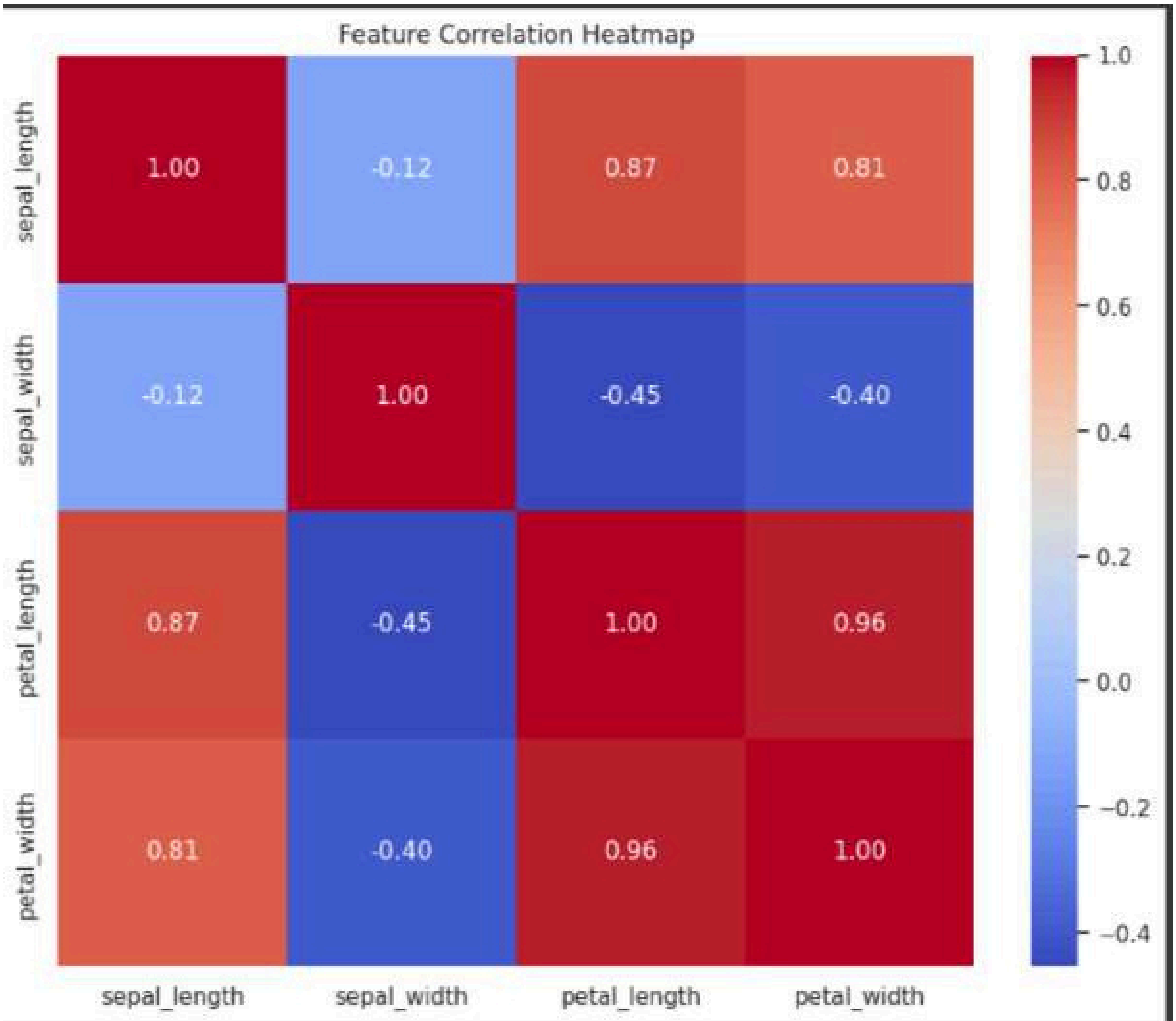
Feature Analysis: Histograms of Flower Characteristics



Feature Visualization: Pairplot of Flower Characteristics



Feature Correlation Analysis: Heatmap



Model Selection: Experiment with multiple models, including Logistic Regression, k-NN, and SVM.

Model Training: Train each model on the training data.

Evaluation: Assess model performance using metrics such as accuracy, precision, recall, and F1-score on the test set.

Prediction and Validation: Make predictions on unseen data and validate the model's accuracy.

8. <u>OBJECTIVES</u>

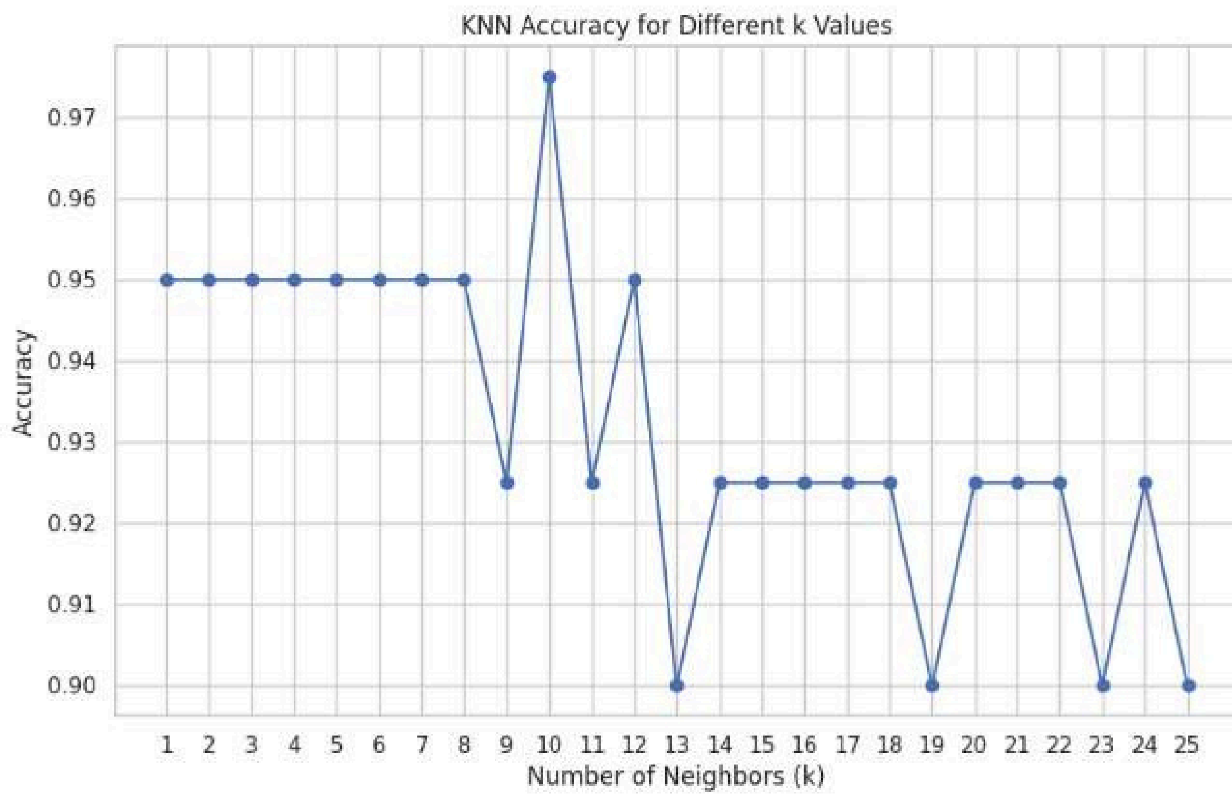
The main objectives of this project are:

1)To build an accurate model for classifying flowers based on their measurements.

2)To compare different classification algorithms and determine the most effective one for this dataset.

9. EXPERIMENTS AND RESULTS

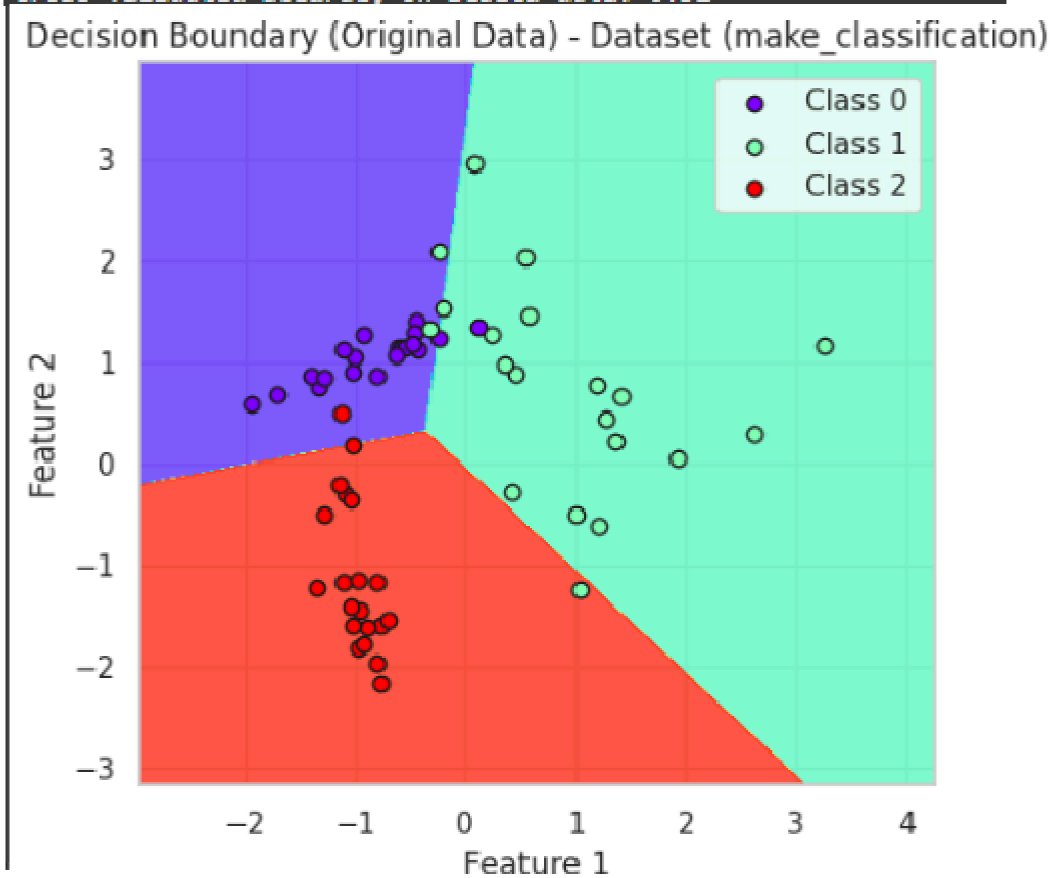
1)KNN: We implemented kNN for different values 1 to 25 and we got different accuracy values for each k value .
It is shown in the graph below:



We can observe the best accuracy at k=10 which is 97.5%.

SVM

```
Model: Dataset (make_classification)
Accuracy on original data: 0.88
Accuracy on scaled data: 0.88
Cross-validated accuracy on original data: 0.91
Cross-validated accuracy on scaled data: 0.91
```



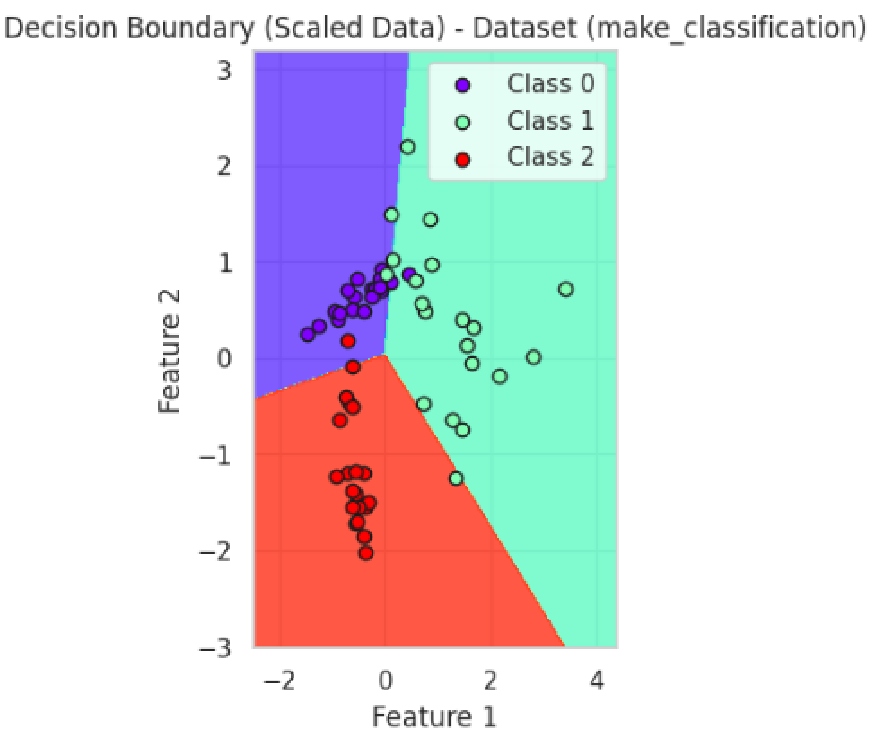
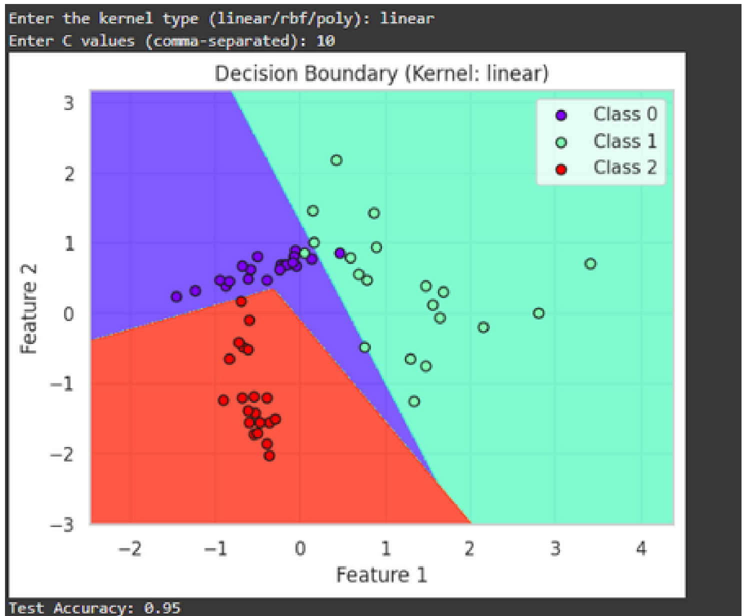


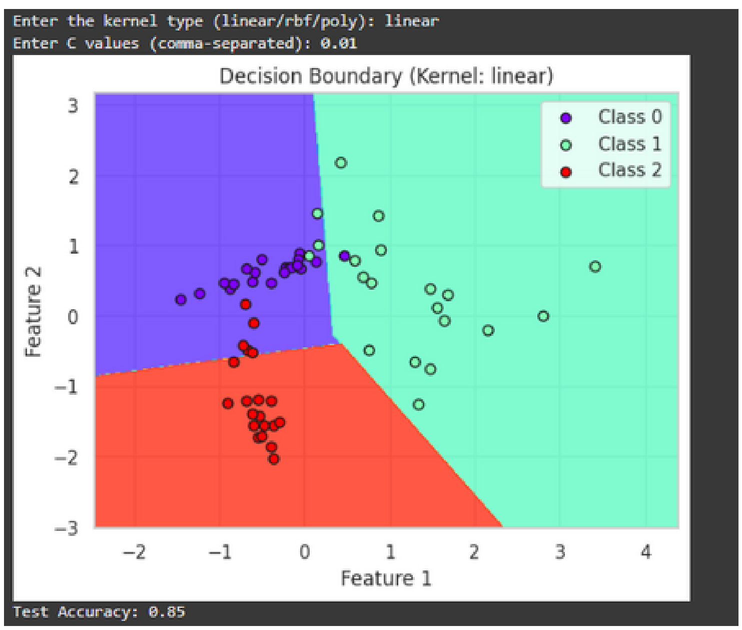
Fig: Decision boundaries by linear SVC on original and scaled data

Decision boundaries by kernel method

1. Linear Kernel

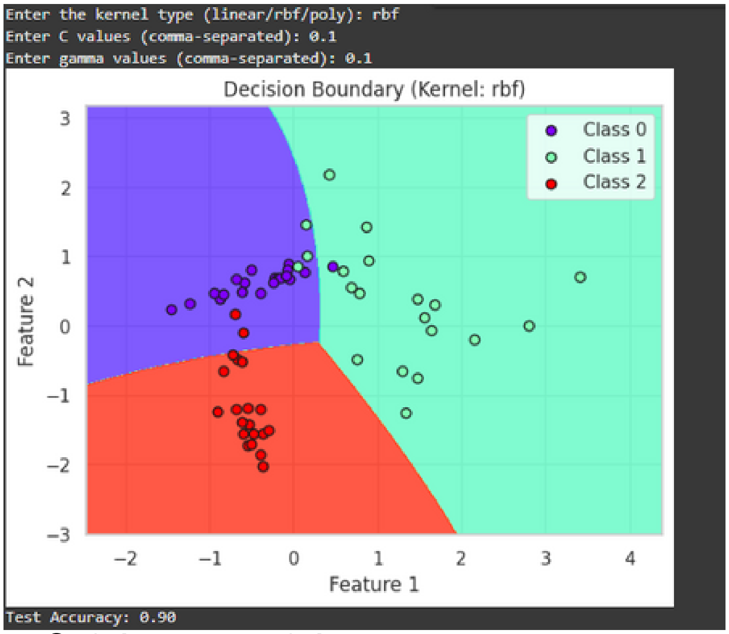


For C=10

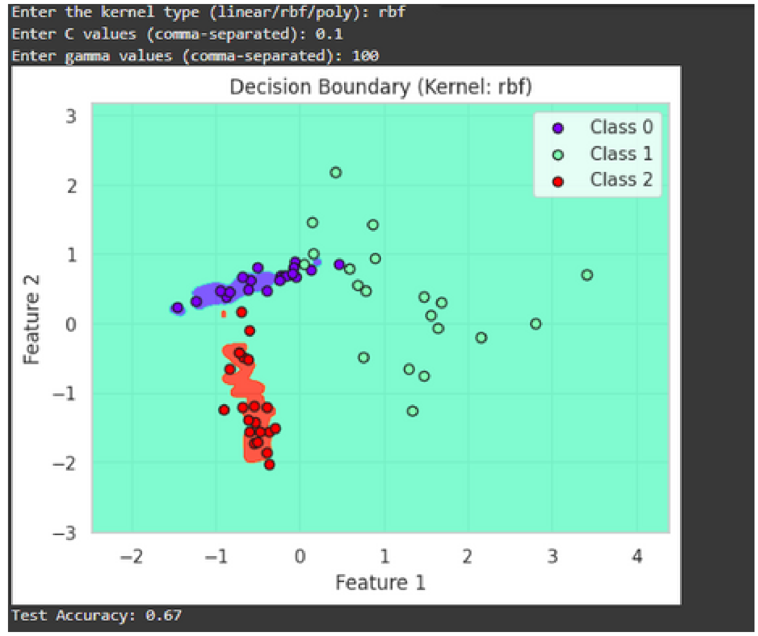


For C=0.01

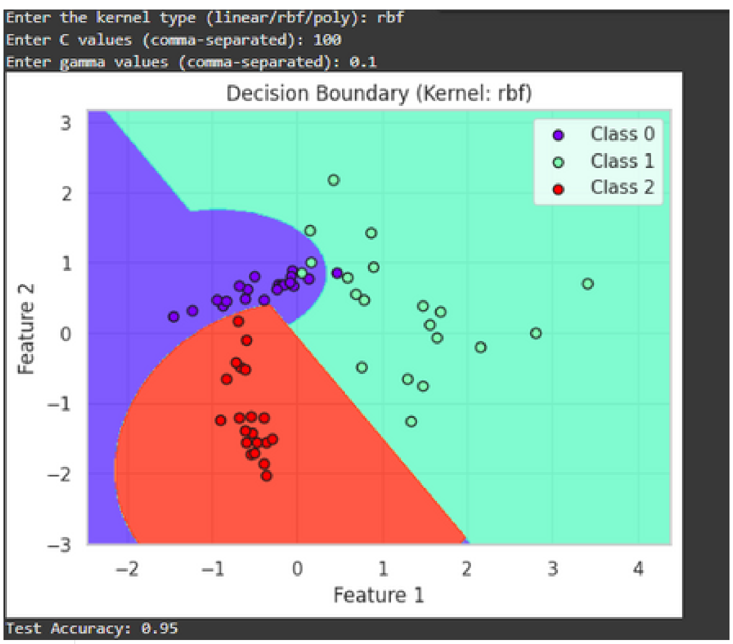
2. RBF (Radial Basis Function)



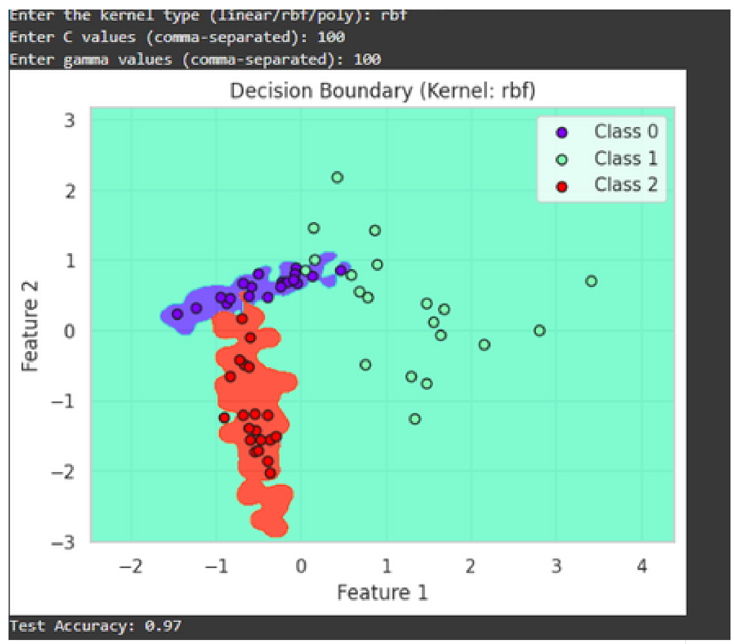
a. C=0.1, gamma=0.1



b. C=0.1, gamma=100

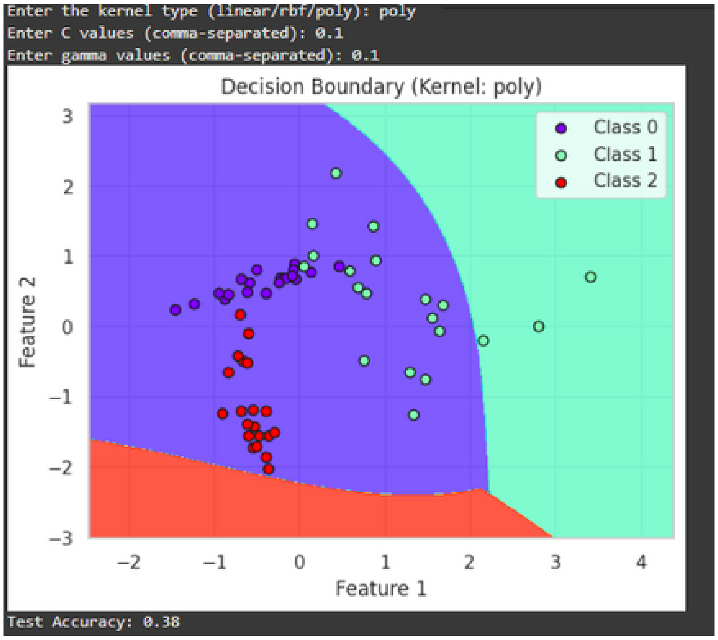


c. C=100, gamma=0.1

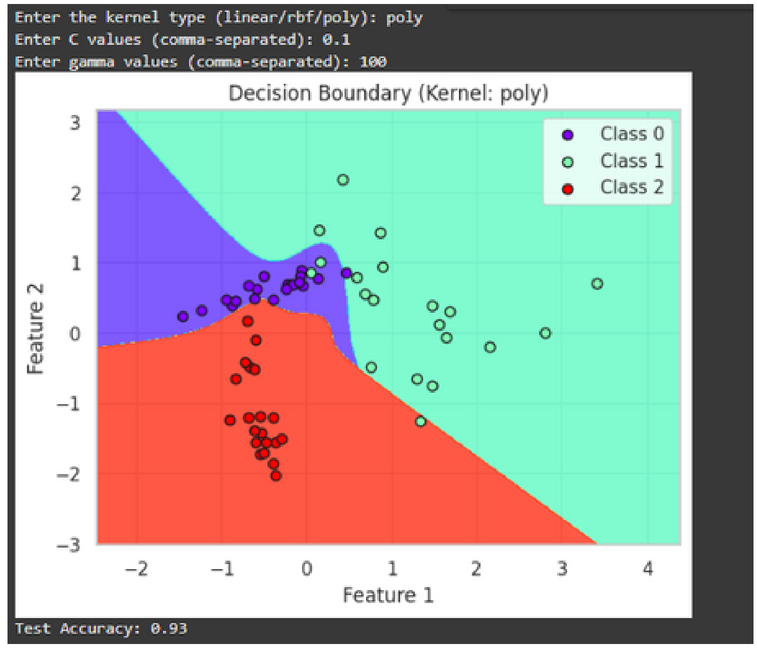


d. C=100, gamma=100

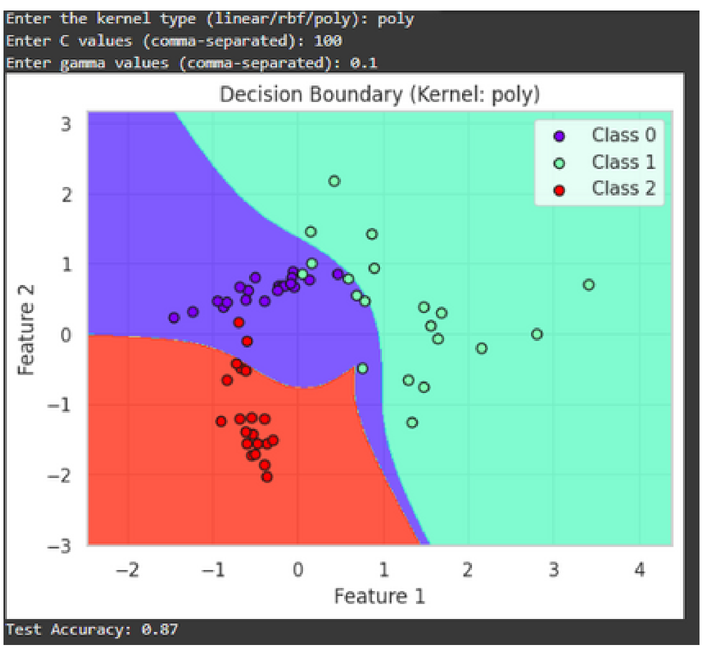
3. Polynomial function



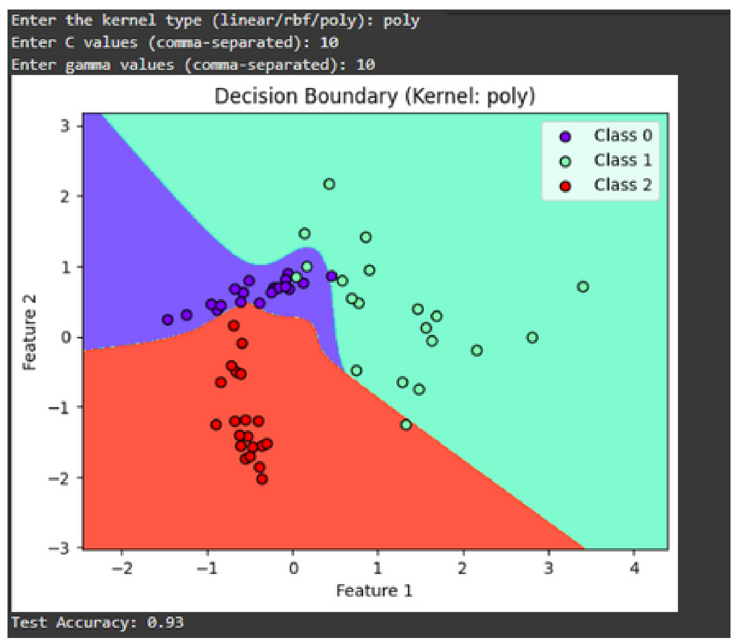
a. C=0.1, gamma=0.1



b. C=0.1, gamma=100



c. C= 100, gamma=0.1



d. C=10, gamma=10

10. CONCLUSION AND FUTURE SCOPE

This project demonstrated that machine learning algorithms could effectively classify flowers based on simple measurements. The best-performing model, in this case, was **kNN MODEL**, with an accuracy of 97.5%. Potential future work includes applying feature engineering, testing on larger and more complex datasets, or deploying the model in a real-time application.

11. REFERENCES:-

Dataset link:
https://drive.google.com/file/d/1vfQ3Ukr7Opep1jjXAXn-q8xiPly8_zaN/view?usp=sharing

- https://drive.google.com/file/d/16Q-aYsRrDIGnd72CKOWHhugaaraAFiiX/view?usp=share_link