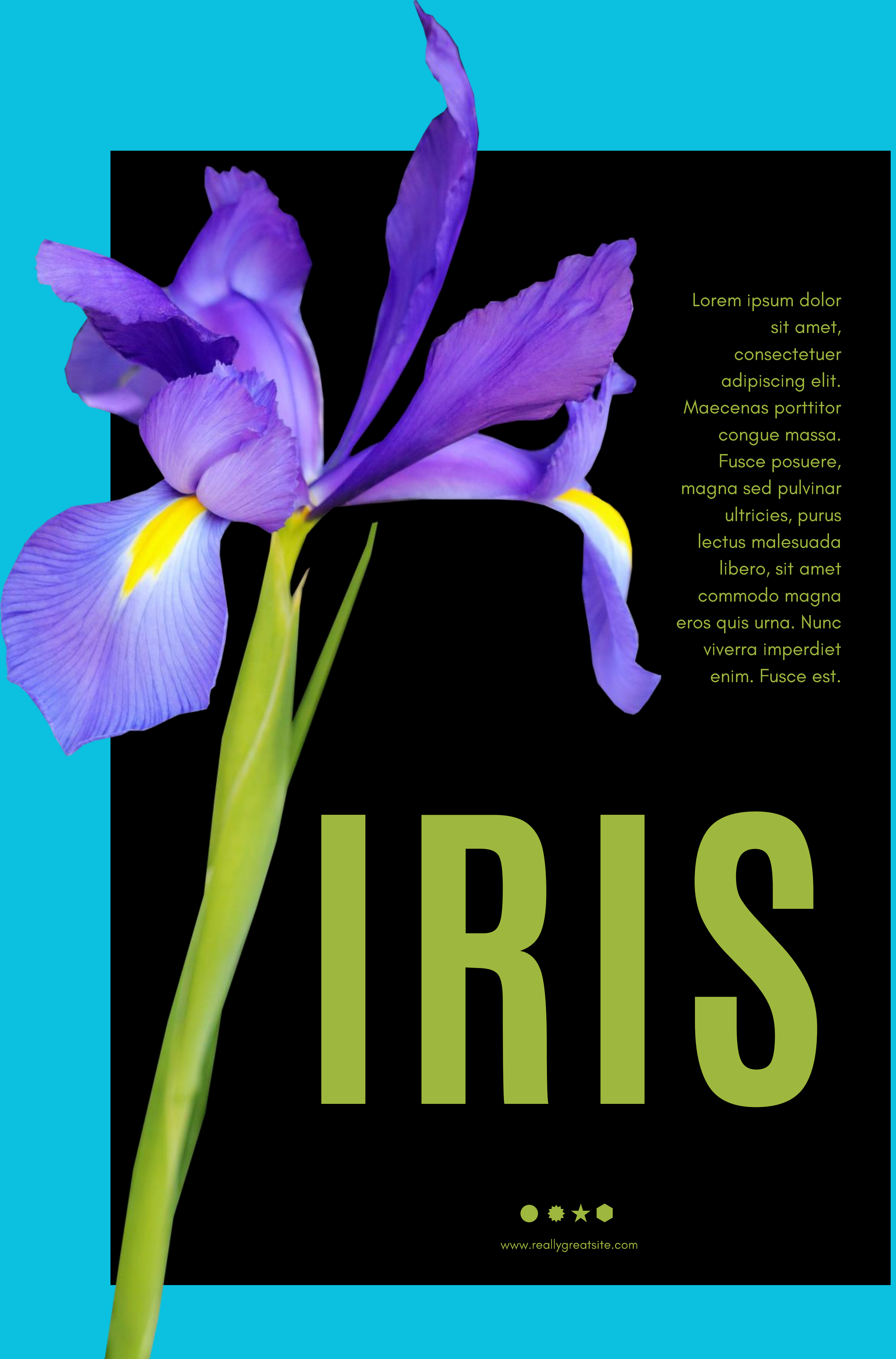
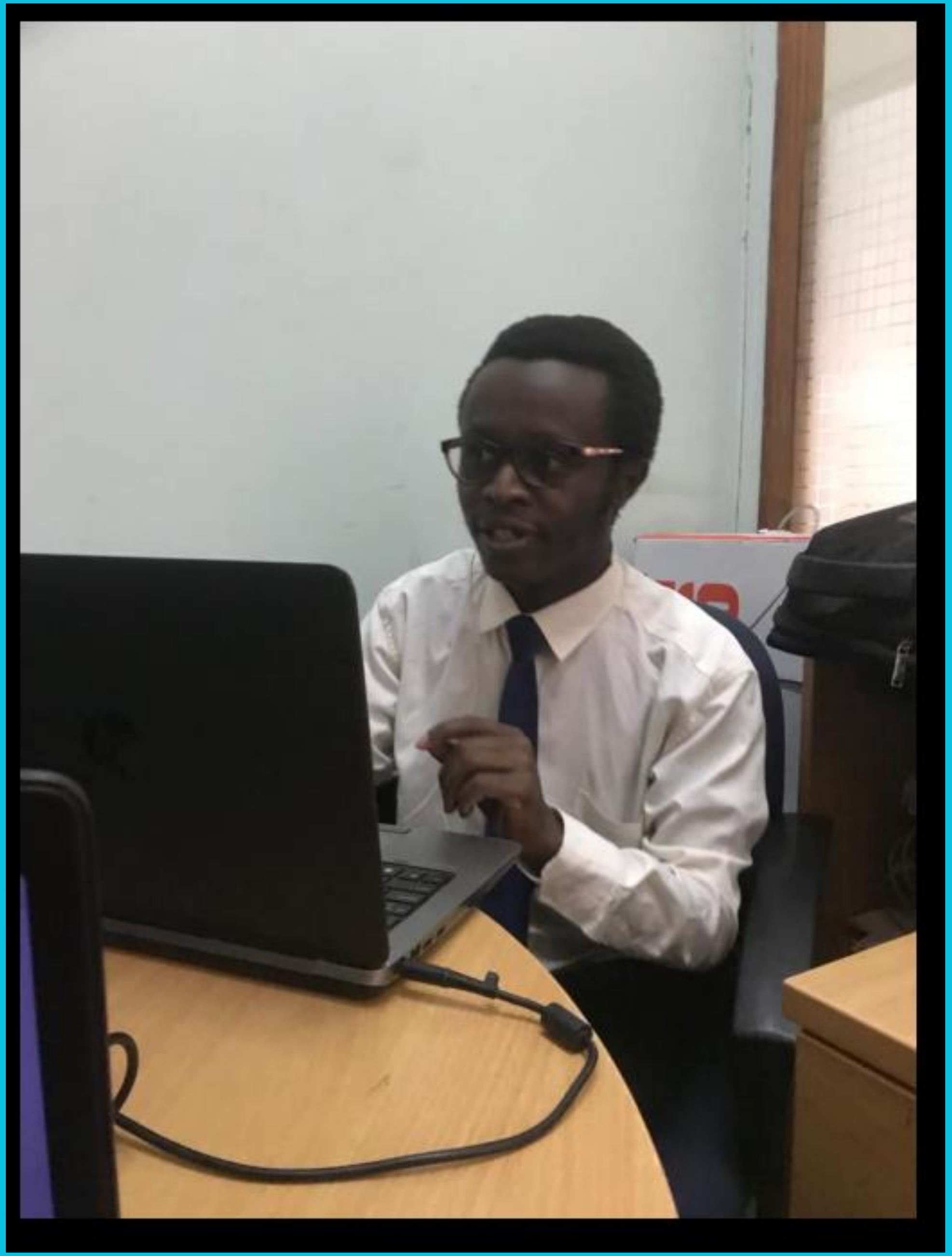


Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est.

Iris
Flower
Classification
Project Presentation

Data Scientist and Machine Learning Engineer



NICHOLAS MUCHIRI KINYUA

https://github.com/MuchiriKinyua

muchiri.kinyua6564@gmail.com

> “ Data is the new oil ~ Clive Humby

Content Outline

Topics for discussion

01 Business Understanding

02 Problem Statement

03 Data modelling

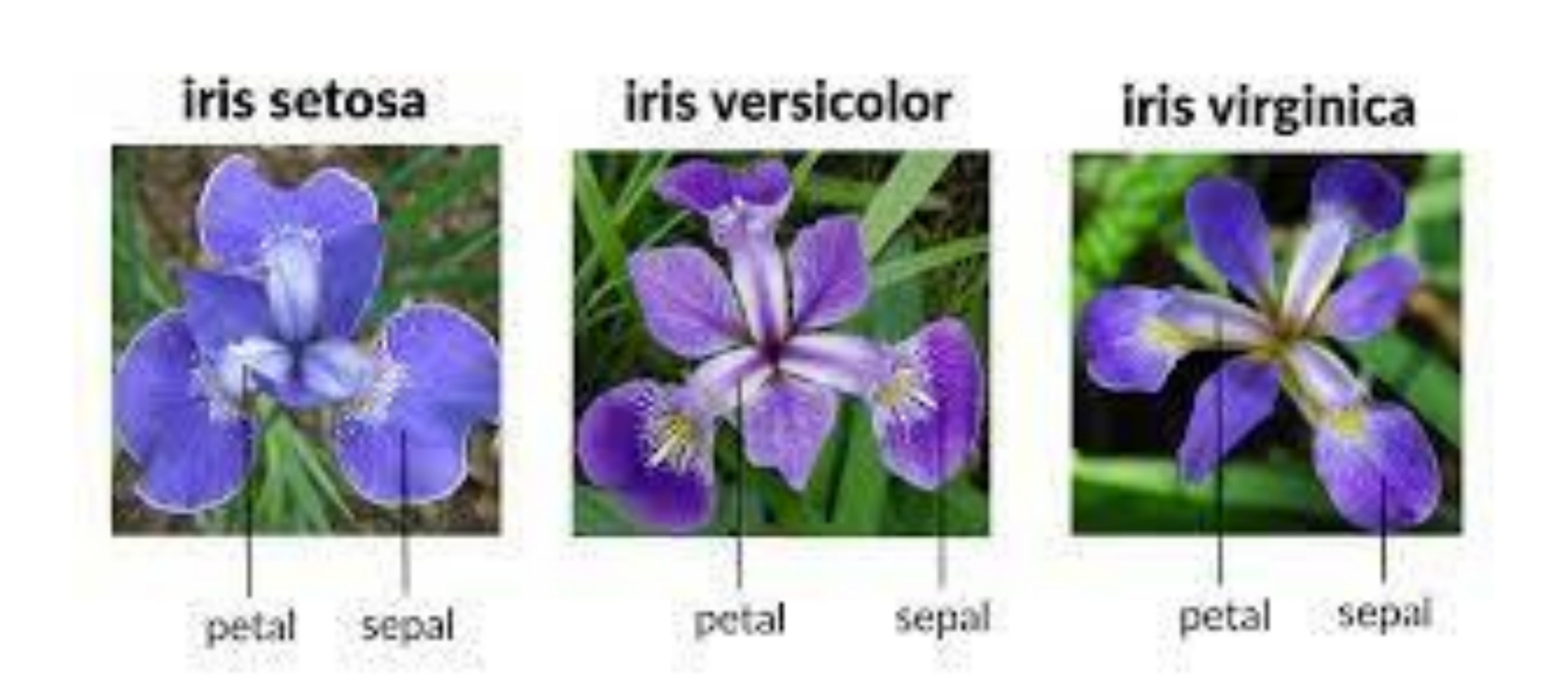
04 Model interpretation

05 Conclusions

06 Recommendations

Business Understanding

This project is aimed to classify iris flowers into three species (setosa, versicolor, virginica) based on sepal length, sepal width, petal length, and petal width.



PROBLEM STATEMENT

Classifying flowers can be challenging, especially with similarities in their morphological features. In botanical studies and horticulture, accurately identifying the species of a flower is crucial for various purposes, such as species conservation, breeding programs, and understanding ecological dynamics. However, manual classification based on morphological features can be time-consuming and subjective, leading to inconsistencies in identification.

To address this challenge, the project aims to develop a machine learning model that can accurately classify flowers based on their morphological characteristics. By leveraging a dataset containing measurements of several features of different flower species, such as sepal length, sepal width, petal length, and petal width, the project seeks to train a model capable of distinguishing between species with a high degree of accuracy.

MAIN OBJECTIVE

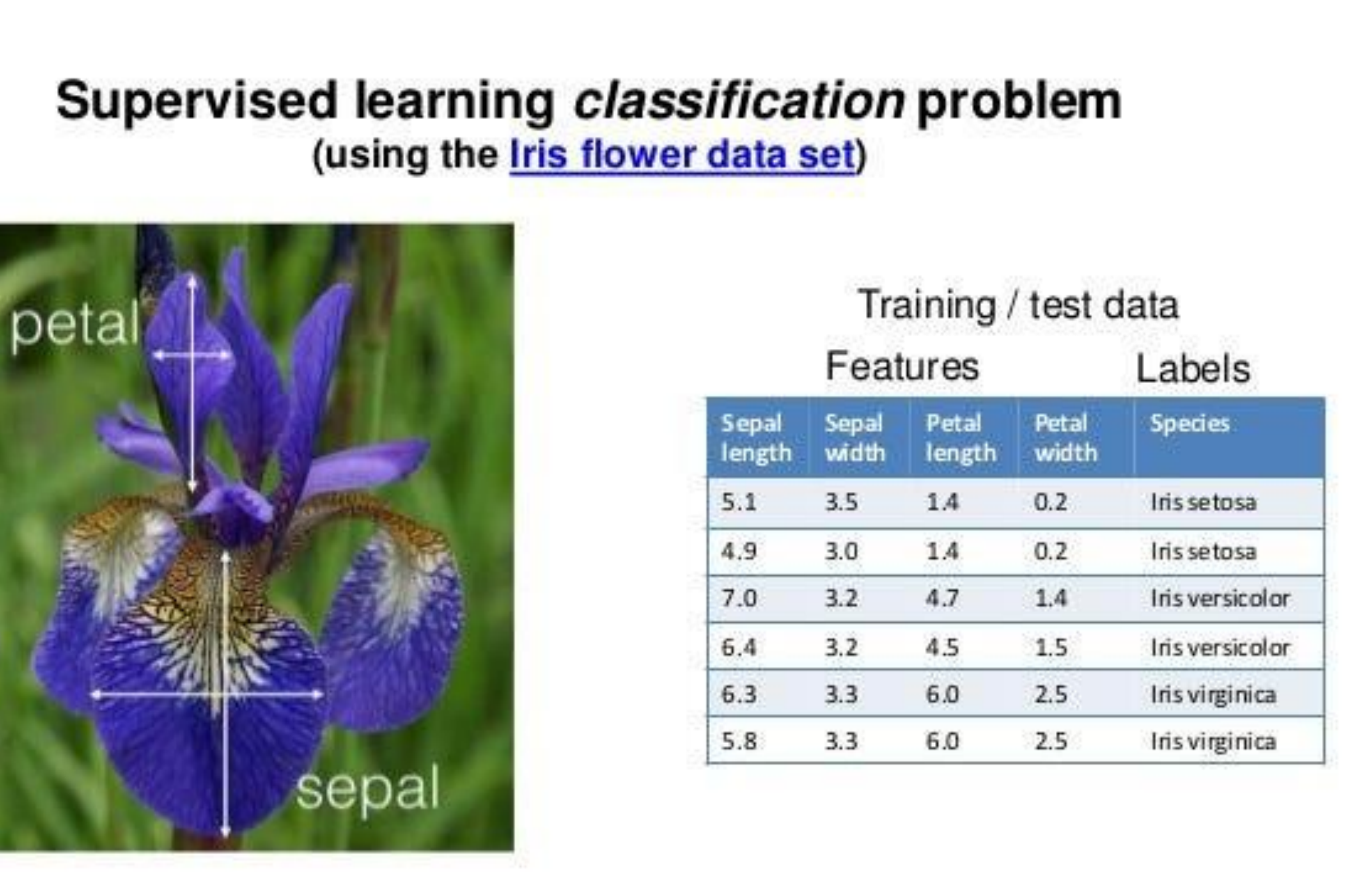
The objective is to develop accurate machine learning models capable of classifying new iris plants into their correct species categories.

Specific Objectives

To build a model that can accurately classify Iris flowers into their three known species (Iris Setosa, Iris Versicolor, Iris Virginica) based on the provided measurements.

To compare the classification accuracy achieved using PCA. This will help in understanding if dimensionality reduction through PCA benefits the model performance in this specific case.

To evaluate the performance of different models using appropriate metrics such as accuracy


petal
sepal

Supervised learning *classification* problem
(using the Iris flower data set)

Training / test data

Features | Labels

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

DATA UNDERSTANDING

The data 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. The dataset, Iris, has got 150 rows and 6 columns. columns ae as follows:

Id - The primary identifier

sepal length (cm) - length of the sepal in centimeters

sepal width (cm) - width of the sepal in centimeters

petal length (cm) - length of the petal in centimeters

petal width (cm) - width of the petal in centimeters

Target - Setosa, Versicolor, Virginica

DATA ANALYSIS

97% of the data collected was used.

DATA ANALYSIS ON IRIS DATA SET

MODELS USED

Model 1:
Logistic regression

The model appears to have an excellent performance, with an accuracy of 95%

Model 2:
Decision Trees

Decision Tree model achieved an accuracy of 94%.

Model 3:
Random Forest

Random Forest achieved an accuracy of 94%.

Model 4:
Neural Networks

Neural Networks chieved an accuracy of 94%.

Model 5:
KNN

KNN achieved an accuracy of 94%.

Model 6:
Nayes Bayes

Naive Bayes achieved an accuracy of 94%.

Evaluation

Evaluation

Ease of Interpretation: Accuracy is an intuitive and easily interpretable metric. It represents the percentage of correct predictions among all predictions made by the model. Users, including hotel staff, can readily understand and trust this metric.

Clear Benchmark: Accuracy provides a clear benchmark for evaluating model performance. It answers the basic question: "How often is the model correct?" This simplicity can be advantageous for communication and decision-making.

Balanced Classes: If the classes of interest (e.g., setosa, versicolor, virginica) are roughly balanced, accuracy can be an effective measure. It doesn't favor one class over another and provides a sense of overall correctness.

I tested and from that we choose the best 3 performing models which are :

Logistic regression = 95%
Random Forest = 94%
KNN = 94%

Conclusions and Recommendations

- It provides a good example for exploring various machine learning techniques, particularly classification algorithms.
- Petal length and sepal length are twice the size of petal width and sepal width.
- Feature importance analysis: It can provide insights into which features contribute the most to the classification task. This information can be valuable for understanding the underlying characteristics of the data and potentially simplifying the model by focusing on the most relevant features.
- Outlier detection and removal, as demonstrated in the analysis, can improve model performance by reducing the impact of noisy data points. However, it's essential to exercise caution and consider the domain knowledge when deciding how to handle outliers like I did. But mostly justify.
- It's essential to monitor the model's performance over time and periodically retrain it with new data to ensure its effectiveness.

Questions or comments?

Get in touch!

Let’s talk

Instructions: Feel free to ask questions at any time during or after the presentation. I'll address them shortly.

Contact Information: For any follow-up questions or discussions, please reach out to muchiri.kinyua6564@gmail.com.

You can also find the full project at: here

Thank you