

Cairo University

Faculty of Computers and Artificial Intelligence

Pattern Recognition

Research Project

|  |  |
| --- | --- |
| Name**:** | ID**:** |
| **1. Ahmed Khaled Mohamed** | **20210016** |
| **2. Habiba Ahmed Mohamed** | **20210120** |
| **3. Mina Albert** |  |

Iris Flower Dataset

Introduction to Iris Dataset:

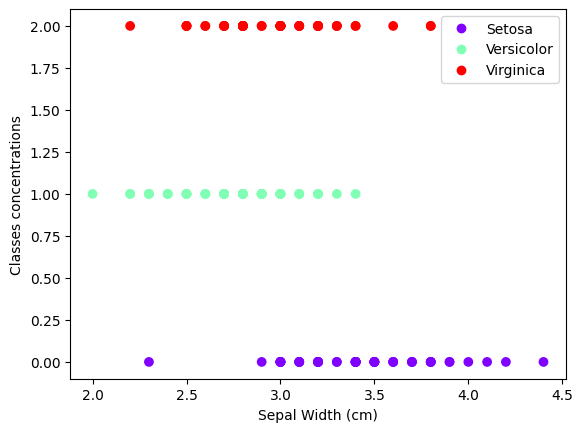
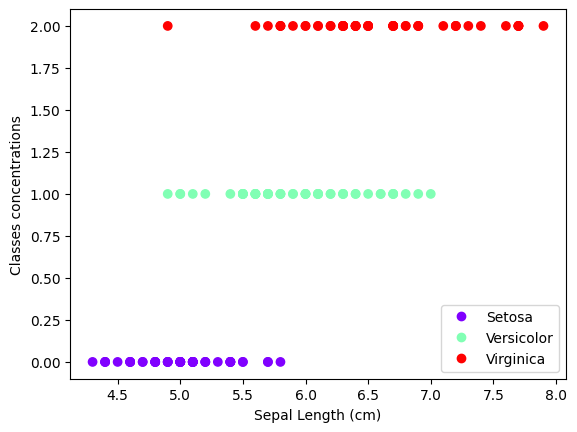
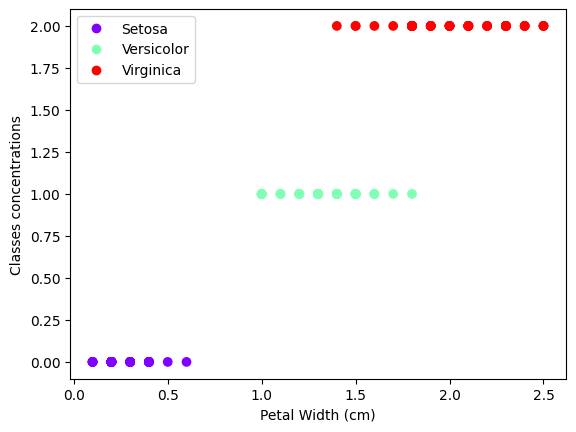
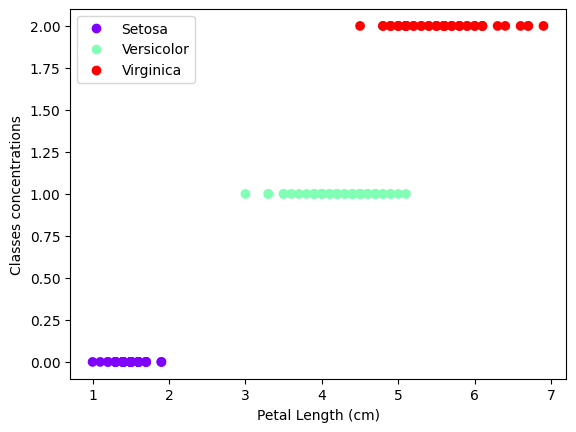
The Iris flower dataset is a classic and widely used dataset in machine learning, it features measurements of Iris flowers from three different species: Iris-setosa, Iris-versicolor, and Iris-virginica. This dataset is a gold standard for showcasing fundamental machine learning concepts in a concise and interpretable way.

The dataset contains four features for each of the 150 Iris flowers. These features are: Sepal length (cm), Sepal width (cm), Petal length (cm) and Petal width (cm).

Target: The target variable is the iris species, classified as Iris-setosa, Iris-versicolor, and Iris-virginica.

The dataset's simplicity and clear separation between some of the classes make it a popular choice for introducing machine learning concepts like classification, visualization, and dimensionality reduction techniques.

Our Methodology:



From just looking at these simple graphical representations of the features, we can already tell that the petal lengths and widths are likely better suited as potential features two separate between the three flower classes.

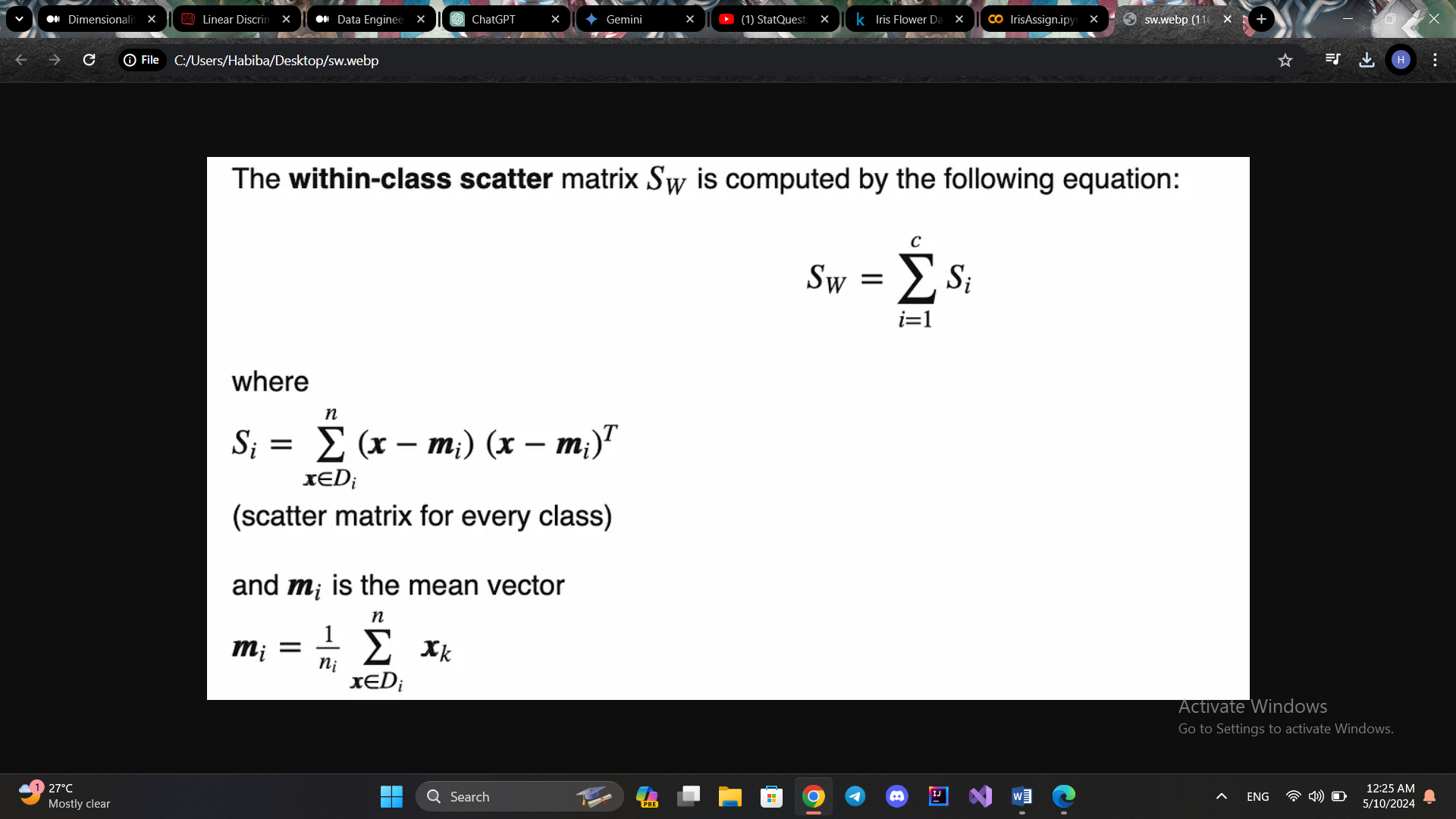
So we want to find the best way to separate the classes and can also potentially improve the accuracy of classification. So the well-suited technique for the Iris dataset is Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA):

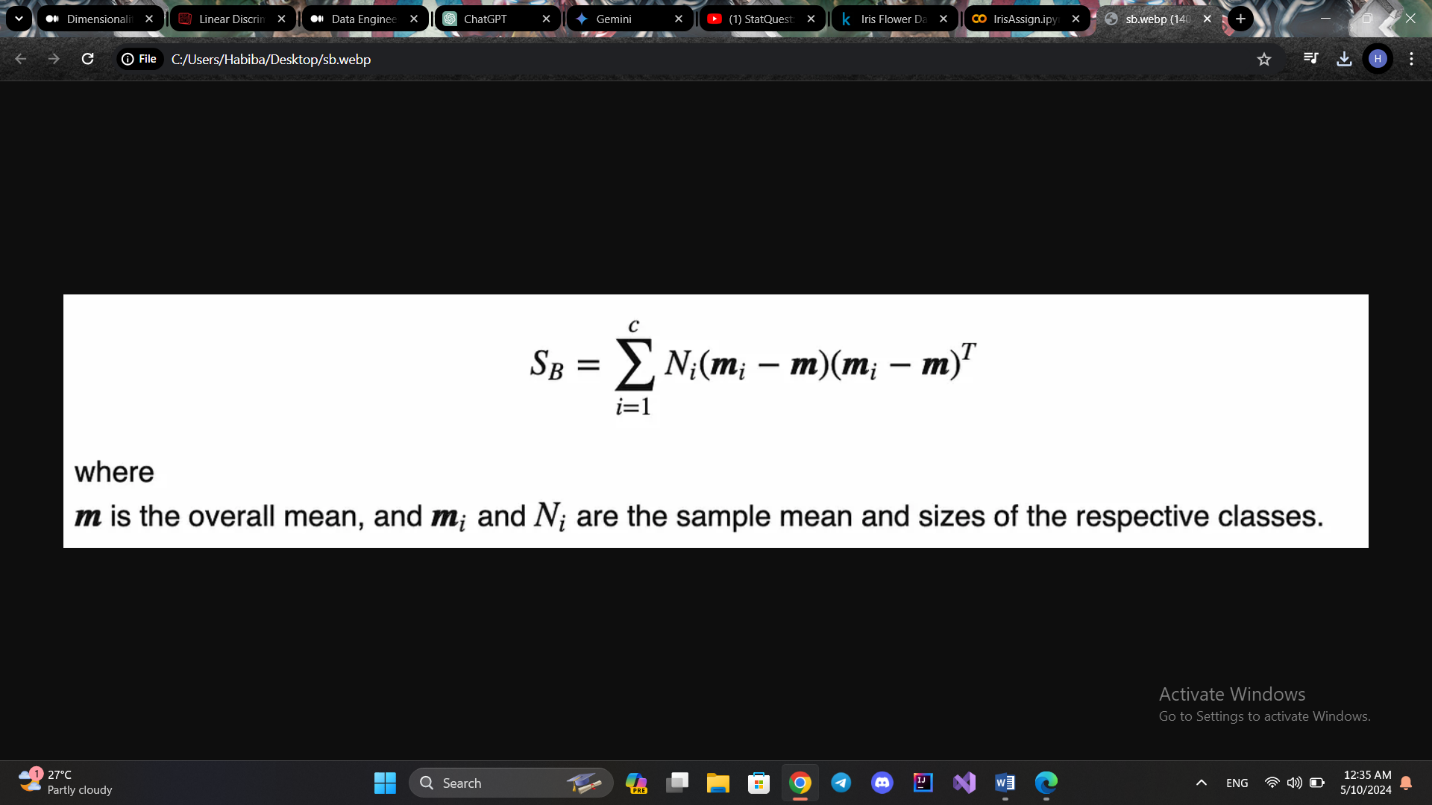
Linear Discriminant Analysis (LDA) is a method used for both reducing the dimensionality of data and for classification purposes. Its primary goal is to enhance the separation between different classes present in a dataset. LDA finds significant application in supervised learning settings, especially when the categories of data points are known in advance.

Steps in LDA:

1. Begins by calculating the mean vectors for each of the three distinct flower classes, denoted as i = 0, 1, 2. Each vector contains the mean of the 4 features in the dataset for the specific class.
2. Calculate the within-class scatter matrix (Sw), which illustrates how the data is distributed within each class.



1. Calculate the between-class scatter matrix (Sb), which captures the variability between different classes, using the following formula:



1. Compute the Eigenvalues and Eigenvectors of Sw-¹Sb (similar to PCA)
2. Arrange the eigenvectors in descending order of their corresponding eigenvalues and select the top k eigenvectors. Once the eigenpairs are sorted accordingly, we can construct our d times k dimensional eigenvector matrix (let’s denote it as W) based on the two most informative eigenpairs.
3. Use the matrix W to transform our samples onto the new subspace via the equation: Y = X\*W, where X is the original data frame in matrix format (150 X 4 matrix in our case) and Y is the transformed dataset (150 X 2 matrix).

PCA vs LDA:

* PCA reduces dimensions on features with the most variations, this is useful for plotting data with a lot of dimensions however in this case we are interested in maximizing the separability between groups so that we can take best decisions.
* LDA is like PCA but it focuses on maximizing the separability among known categories.

The lower-dimensional representation created by LDA can then be used by a machine learning classifier to potentially achieve better classification accuracy compared to using all the original features, especially if some features are redundant or noisy. And for that we will use Logistic Regression

Logistic Regression:

After this dimensionality reduction by LDA, the resulting lower-dimensional representation of the Iris flower data is fed into a Logistic Regression model. Logistic Regression excels at finding the decision boundaries between the classes in this new space. It analyzes the data points in the reduced space and learns a model to predict the most likely Iris species for a new, unseen flower based on its measurements.

The combination of LDA for feature reduction and Logistic Regression for classification is a powerful approach for the Iris dataset and similar classification problems with well-defined classes. However, it's essential to consider the specific characteristics of your dataset and task when choosing the most suitable techniques.