Assignment1: Implementing a Naive Bayes Classifier on the Iris Dataset

Objective

Implement a Naive Bayes classifier from scratch i.e. without using any built-in API/function/method from any library to classify Iris flower species based on the widely used Iris dataset. The assignment involves data loading, preprocessing, splitting the dataset into training and testing sets, training the Gaussian Naive Bayes model, performing classification on the test set, and evaluating accuracy.

Problem Description

The Iris dataset consists of 150 samples from three species of Iris flowers*: Iris-setosa, Iris-versicolor, and Iris-virginica*. Each sample has four continuous features:

- Sepal length (cm)
- Sepal width (cm)
- Petal length (cm)
- Petal width (cm)

You are asked to:

- Load the dataset from a CSV file
- Randomly split it into training and testing sets according to a user-defined ratio
- Train a Gaussian Naive Bayes classifier by calculating the mean and standard deviation of each feature per class
- Predict the class of each test sample and compute class probabilities
- Output predictions along with probabilities and correctness for each test sample
- Calculate and display overall accuracy on the test set

Input Details

- Dataset file: `iris.csv`
    - CSV format without header (or header row can be ignored)
    - Columns (in order): sepal length, sepal width, petal length, petal width, class label
    - Class labels: Iris-setosa, Iris-versicolor, Iris-virginica

- User input:
    - A float number defining the training set ratio (e.g., 0.7 means 70% training data, 30% testing)

Output Details

For each test instance, output:

- Feature values
- Predicted class label
- Probability assigned to each class (normalized to sum to 1)
- A correctness indicator ("Correct" if prediction matches the true label, otherwise "Wrong")

At the end, output:

- Overall accuracy on the test data as a percentage

Submission deadline

August 14, 10AM, 2025