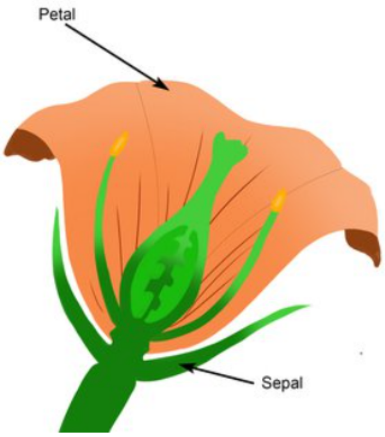
IC 272 - Data Science III

Assignment 3: Bayes Classifier

Deadline: October 18, 2023: 23.59 Hr.

Dataset Description:

You are given two CSV files “**iris_train.csv**” and “**iris_test.csv**” containing the measurements of three types of Iris flowers:



Independent variables/ Attributes/ Features:

i. “`SepalLengthCm`”: Sepal length in cm.

ii. ” `SepalWidthCm`”: Sepal width in cm.

iii. “`PetalLengthCm`”: Petal length in cm.

iv. “`PetalWidthCm`”: Petal width in cm.

Dependent variable/ Target Attribute/ Class:

i. “`species`”: Type of the flower corresponding to a set of measurements.

Problem Statements:

I. Consider the given files as the train and test datasets containing samples from three Gaussian distributions and do the following:

i) Reduce the **dimension** to **ONE** of both train and test samples using the PCA you implemented for Assignment-2.

ii) Build a Bayes classifier on the one-dimensional train set by estimating the parameters of univariate Gaussian distributions. Implement the **steps** taught in the class. Do NOT use **any built-in** classification function.

iii) Test the model by classifying each one-dimensional test sample. To do this you need to compute a test sample's likelihood for the classes. Implement the method taught in the class. Do NOT use **any built-in** classification function.

iv) Estimate the model's performance by computing a confusion matrix and the model's accuracy in percentage.

NOTE:

a) Do NOT use any built-in function to compute any statistical measure of the data.
b) Read the data file using Pandas.

II. Consider the original train and test datasets of four-dimensional samples drawn from three Gaussian distributions and do the following:

i) Build a Bayes classifier on the four-dimensional train set by estimating the parameters of multivariate Gaussian distributions. Implement the **steps** taught in the class. Do NOT use **any built-in** classification function.

iii) Test the model with each four-dimensional test sample by computing its likelihood for the classes. Use the built-in function defined in **scipy.stats** for computing the probability of a sample.

iv) Estimate the model's performance by computing a confusion matrix and the model's accuracy in percentage.

III. Compute and print the difference between the accuracies of the models built using the original and dimension-reduced data.

NOTE:

a) You may use built-in statistical functions, but using the functions, you already implemented for other assignments will bring **bonus marks**.