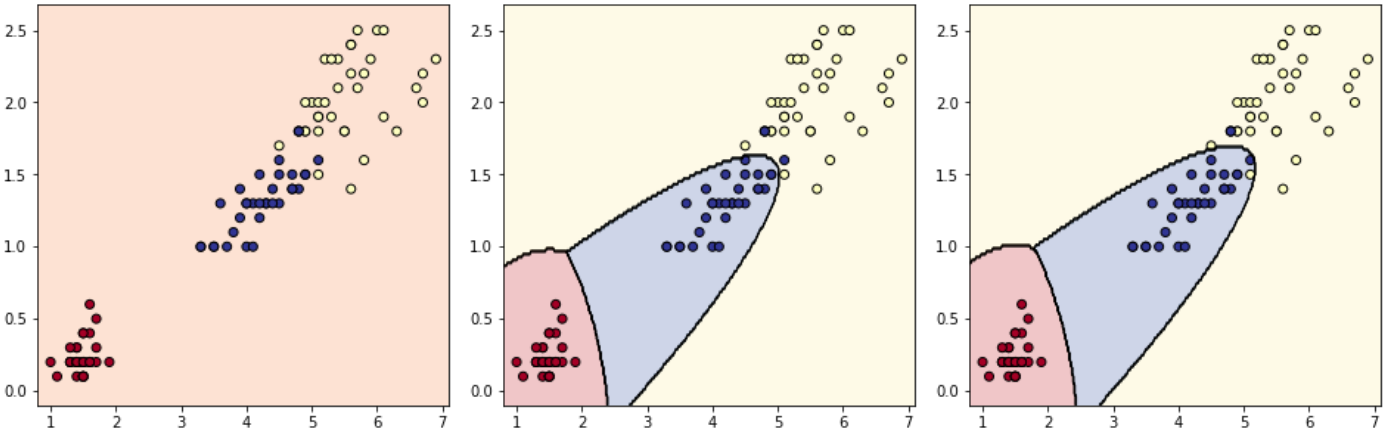
LAB ASSIGNMENT 4
Lab Report

Shubh Goyal (B21CS073)

QUESTION 1-

Task 1 and 2: The class Guassian_Bayes_Classifier was implemented with the required methods (train, predict, test and plot decision boundary).

Task 3: The following decision boundaries are obtained when two features are used to train the model (feature 1 and feature 2 i.e. petal length and petal width respectively) in the order of given cases respectively:



In the first case, the decision boundary cannot be seen maybe because it is not in the limit of the points or because the model was not able to form one and started recognizing every point as a single class. The decision boundaries in case 2 and 3 are quadratic and almost similar.
In case of training on two features (specified above) instead of all, the accuracies obtained were 34.28%,98.09% and 98.09% respectively.
(Using different columns as features for decision boundary plots created different shapes and gave different accuracies. HIghest accuracy was obtained in case of the specified features itself.)

The accuracy on the test data of the models which were trained on all the four features were 31.11%, 100% and 100% respectively which is quite good. In case 1,

the model classified every point as a single class and is thus not working well and is not viable to be used here. But 100% accuracy in case2 and case3 along with the decision boundaries plotted indicates that the models are working well here.

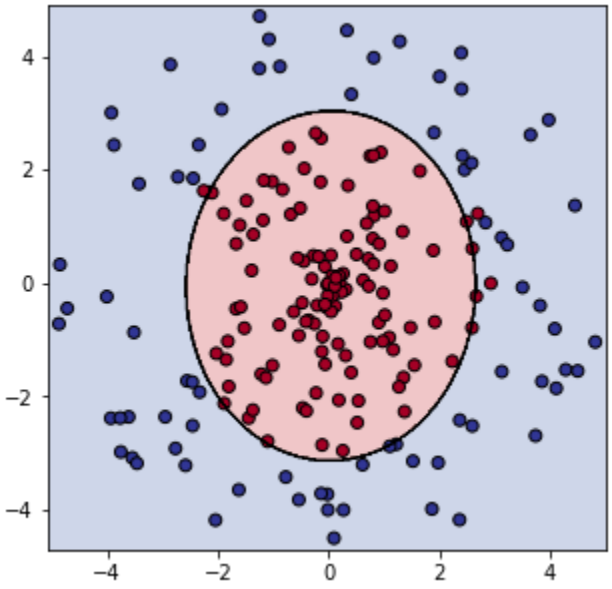
The models in case 2 and 3 are generalizable to the data quite considerably but the model in case 1 is not at all generalizable, rather can be said to be in a condition of complete bias towards a single class.

Task 4: A function for cross validation was written. The data was splitted into 5 parts using pre pre-defined numpy function.

The average accuracy in case 1 is 34.28%.
The average accuracy in case 2 is 96.19%.
The average accuracy in case 3 is 97.14%.

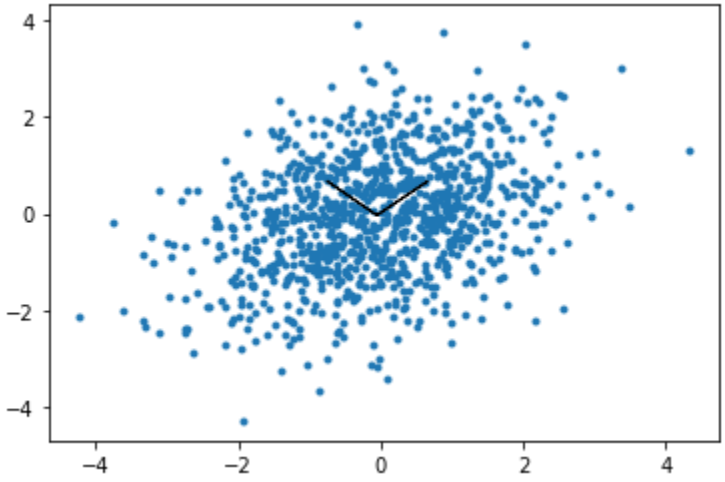
Model in case 1 is not generalizable. While models in case 2 and case 3 are performing well and generalized. The model in case 3 is best generalized here.

Task 5: A synthetic dataset with 400 values having 200 values from each class was generated using numpy. Classes were made on the basis of a points euclidean distance from the origin.
The Gaussian Bayes model (case 3) gave an accuracy of 96.5% in case of this synthetic dataset. The decision boundary created was also close to the actual boundary.



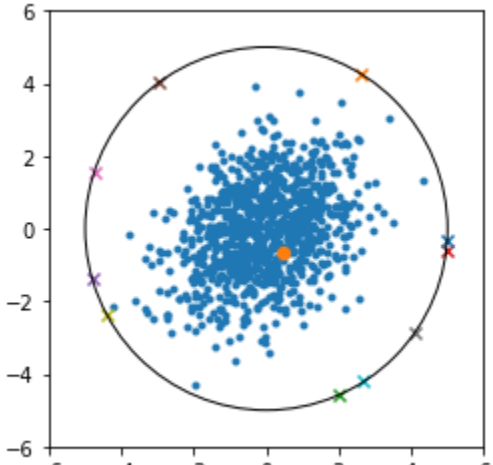
QUESTION 2-

Task 1: Sample X was created using numpy and a function was made for calculating covariance matrix. Eigenvalues and vectors were also calculated using numpy and the following plot was obtained. The blue points represent sample X and the black arrows/line segments represent the eigenvectors.

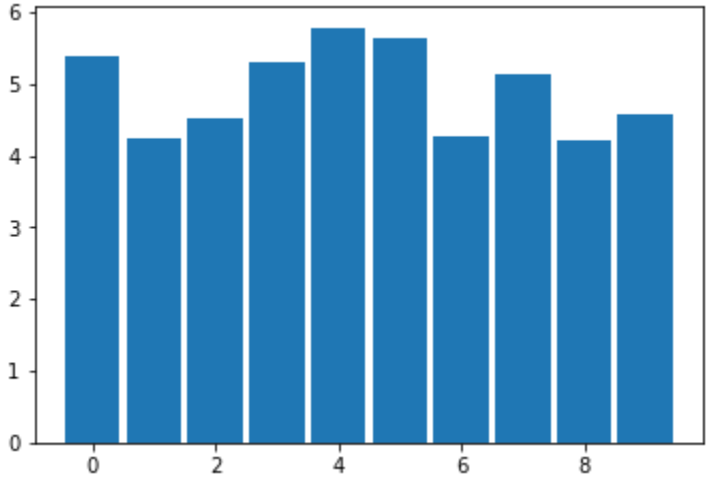


Task 2: After the transformation, the covariance matrix calculated was found to be a diagonal matrix which represents that the covariance between features is zero or infinitely small.
The purpose of the transformation was to create a sample with independent features.

Task 3: A random uniform sample (named P) of 10 data points was created on the curve x^2 + y^2 = 25.
The following plot was obtained when the sample P (cross marks) was plotted with sample X (blue points).



The following bar plot represents the euclidean distance of the points in sample P from their mean.



Task 4: The transformation was performed and the following plots were obtained. The bar plot shows the distance of new points from their mean. The other plot is plotted with Y (blue dots) and Q (cross marks).