Comparative Analysis Report

- **GDA on Microchip Raw Data**
- **GDA on data set having Gaussian distribution**

A) GDA on Microchip Raw Data

Dataset:

```
dataset=pd.read_csv('microchip.csv')
dataset = dataset.sample(frac=1).reset_index(drop=True)
dataset
```

	result1	result2	accepted
0	-0.697580	0.684940	0
1	0.961410	0.085526	0
2	-0.611180	-0.067982	0
3	0.483290	-0.184940	1
4	0.149190	-0.579680	0
...	...	...	...
113	0.016705	-0.404240	1
114	-0.467170	-0.133770	0
115	-0.392280	-0.258040	0
116	-0.375000	0.502190	1
117	0.863480	-0.082602	0

118 rows × 3 columns

Fig 1. Microchip Dataset

This raw dataset does not follow gaussian distribution among the two classes (0 and 1) which can be seen by analysing the dataset through its histogram representation.

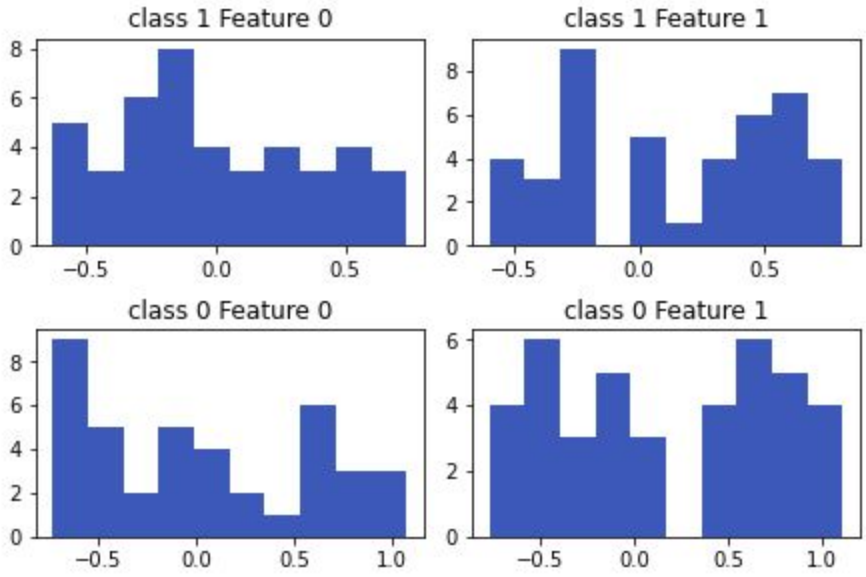

Fig 2. Features of Raw Microchip Data

The above graph shows the raw microchip data does not follow any gaussian distribution therefore if we apply GDA algo on this dataset the classify into 1 or 0 , we might not get the expected result and that is what is happening . In this dataset , on applying GDA we get an accuracy of 54 percent.

```
[12] test(trainX,trainY,testX,testY)

     Accuracy on raw data= 54.285714285714285
```

Fig 3. Accuracy for GDA on Raw Data

B) GDA on data set having Gaussian distribution :

Here we produce the two features ourselves by using **Bux- Muller transformation** and hence for both classes we get data following gaussian distribution.

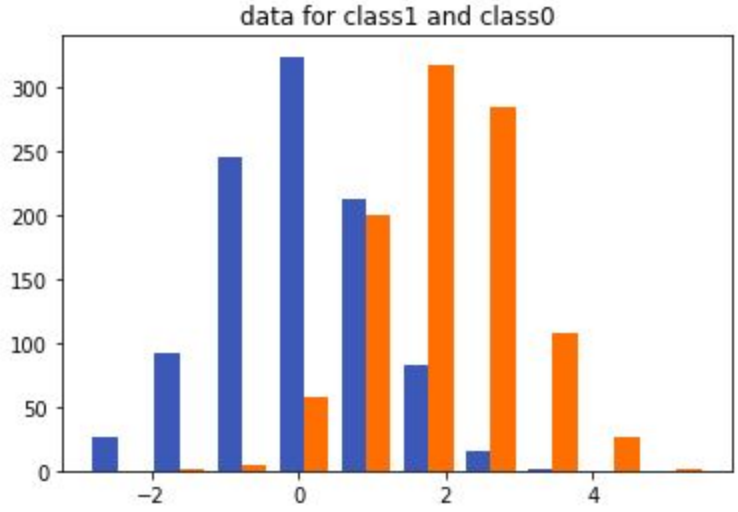

Fig 4. Data produced by BM algorithm

The above figure shows two gaussian distributions which are then used to create data for the two classes of the two features given below which is too approximately a gaussian distribution shown below.

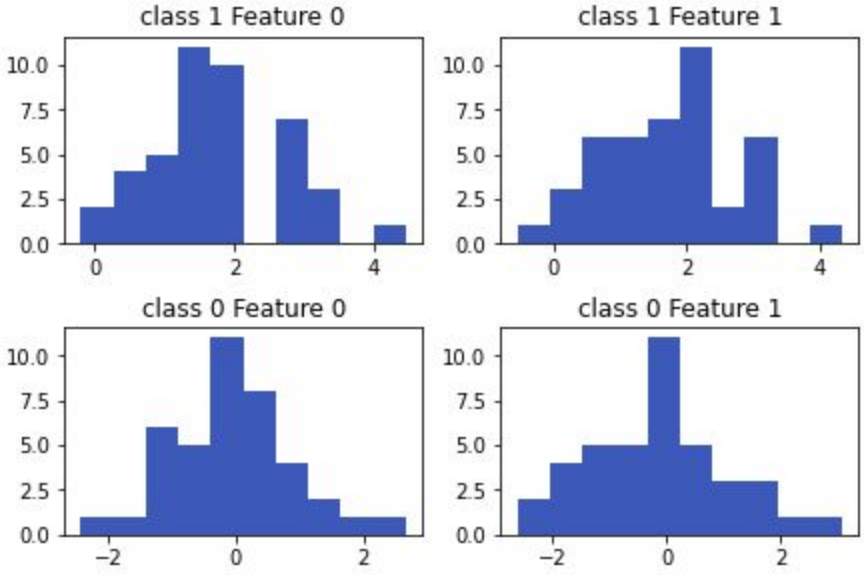

Fig 5. Features of self produced Gaussian Data

Since GDA works for a dataset following gaussian distribution, therefore the **accuracy of Part B is much more than that of part A.**

```
test(trainX,trainY,testX,testY)
```

```
Accuracy using Bux Muller = 94.28571428571428
```

Fig 6. Accuracy for GDA on Gaussian Data