COMPARITIVE ANALYSIS REPORT GDA

-RATNESH KUMAR RAI(MIT2019098)

Dataset:

We were given a dataset of Microchip containing two test features, and a result of pass(1) and fail(0).First, for applying box-muller we have to pre-process the data as the log function does not work for -ve values, so I used MinMaxScaler of SciKitLearn, which scale our data between 0 to 1. But again, log(0) will throw error, So instead of dropping the complete row, I replaced with **median value**, to avoid the error.

After that we applied box-muller Algo, to get a gaussion data. Then, I divided data into test and training, containing 30% of total data as test data.

Performance Analysis Of GDA:

Without BOX-MULLER	With BOX-MULLER
Testing data: 30% Accuracy: 0.472 Confusion Matrix: [6 17] [2 11]	Testing data: 30% Accuracy: 0.639 Confusion Matrix: [13 10] [3 10]
Testing data: 20% Accuracy: 0.416 Confusion Matrix : [4 13] [1 6]	Testing data: 20% Accuracy: 0.666 Confusion Matrix: [11 6] [2 5]
Testing data: 50% Accuracy : 0.491 Confusion Matrix: [5 29] [1 24]	Testing data: 50% Accuracy : 0.627 Confusion Matrix: [15 19] [3 22]

Conclusion:

As we can see the GDA works better for data which are Gaussian distributed or likely Gaussian in nature, as similar to we read in theory, if the data is not Gaussian, we should avoid using GDA, rather we can opt for Logistic Regression for classification. Although GDA has its own perks, computation cost is much less compared to Logistic Algorithms and It works faster for larger dataset.