Gaurang Goel - GXG190015
Partner - Anurag Diwate - ASD190004

a. Training time: 27055 microseconds
   Coefficients: -8.08405 -2.9601
   Test accuracy: 1
   Test sensitivity: 1
   Test specificity: 1

   With Naive Bayes:
   Accuracy: 0.53252
   Sensitivity: 0
   Specificity: 1

b. Using the synthetic data, the logistic regression model achieved a test accuracy of 1, suggesting that it was able to properly categorize the test data. Nevertheless, the accuracy of the Titanic data was just 0.53252, demonstrating that the model did not perform as well on the real-world dataset. The sensitivity was zero, suggesting that the model was unable to accurately detect positive instances, which in this case would be persons who died. The specificity was one, suggesting that the model accurately identified the negative instances, in this case, the ones who lived. There might be various reasons why the model did not perform as well with Titanic data. One reason might be that the synthetic data was created with a clear distinction between the two groups, but real-world data is likely to be more complicated and lack such a distinction. Another cause might be that the model's features were insufficiently informative to discriminate between the two classes, and new features or more advanced feature engineering may be required.

c. Machine learning methods used for classification applications include generative classifiers and discriminative classifiers. Generative classifiers employ Bayes' rule to determine the conditional probability of the class label given the input features by modeling the joint probability distribution of the input features and the class labels. In contrast, discriminative classifiers explicitly predict the conditional probability of the class label given the input characteristics. Generative classifiers have the benefit of being able to represent the whole probability distribution of the input characteristics and class labels, which can be valuable for tasks such as missing data imputation or synthetic data synthesis. They also perform well when there are few training examples available since they can estimate the class label using the complete probability distribution. When the focus is on the classification job itself, discriminative classifiers perform better since they explicitly predict the likelihood of the class label given the input characteristics. Since they do not have to represent the complete probability distribution, discriminative classifiers can be more computationally efficient.

d. In machine learning, reproducible research refers to the technique of making research results and methodologies public and replicable so that other researchers may replicate the experiments and achieve the same results. This is significant for a number of reasons. For starters, it allows other researchers to validate the conclusions and methodology, which can boost trust in the findings. Second, it permits researchers to expand on earlier work in new areas. Finally, it fosters research transparency and accountability. There are numerous approaches to implementing repeatability in machine learning research. To monitor changes to the code and data over time, one method is to utilize version control systems such as Git. Another way is to utilize containerization technologies like Docker to build repeatable environments that can be readily shared and operated across several platforms. A third way is to leverage platforms like GitHub and Zenodo to freely share code and data, making it available for others to use and expand on.