Assignment 2 - Generative vs. Discriminative Models

Narahara Chari Dingari, Ph.D.

DS552 - Generative AI

Assignment Instructions:

Use the Penguins dataset open source dataset, focusing on two species only (e.g., Adelie and Gentoo), to compare the performance of Naive Bayes (Generative Model) and Logistic Regression (Discriminative Model). Please ensure to upload your Jupyter Lab Notebook along with the corresponding code and markdown explanations on GitHub. Once complete, share the repository link for submission.

1. Accuracy Comparison:

- Evaluate and report the accuracy of both the **Naive Bayes** and **Logistic Regression** models on the **training** and **test** datasets.
- Compare the performance of these models in terms of accuracy to determine which model better distinguishes between the two penguin species. Provide a brief explanation of your findings.

2. AUC (Area Under the ROC Curve) Comparison:

- Calculate the AUC for both Naive Bayes and Logistic Regression on the training and test datasets.
- Interpret the AUC values for both models to assess how effectively each model discriminates between the two penguin species.
- Provide insights into which model is more effective based on the AUC metric.

3. Lift and Gain Charts:

- Generate **Lift** and **Gain** charts for both Naive Bayes and Logistic Regression using **10 deciles**.
- Use a **dual y-axis** plot with **deciles on the x-axis**, and **Lift and Gain** on the y-axis.
- Evaluate the Lift and Gain charts to understand how well each model ranks the predicted probabilities and how effective each model is in prioritizing the classification of the two species.

4. Model Performance Comparison:

- Based on the results from accuracy, AUC, and Lift/Gain charts, compare the overall performance of Naive Bayes and Logistic Regression models.
- Discuss which model performs better in classifying the two penguin species and provide reasons for your conclusion.

5. Performance on a Complex Dataset:

- Extend your analysis by applying both Naive Bayes and Logistic Regression to a more complex dataset, such as **MNIST** (handwritten digits).
- Compare how the performance of generative models (e.g., Naive Bayes) and discriminative models (e.g., Logistic Regression) differs when dealing with **image data** (MNIST) compared to the two-species penguin dataset.
- Discuss the differences in performance and behavior across these datasets.