Assignment 3

1. Load Lab 2 Dataset.csv from https://github.com/sanadv/MLCourse
2. Use Lab_2.ipynb as an example
3. Predict the costumer response using

- Logistic Regression
- GaussianNB
- generalized_linear_model
- LogisticRegression
- Deep Learning
- KNeighborsClassifier
- LinearDiscriminantAnalysis
- QuadraticDiscriminantAnalysis
- GaussianProcessClassifier
- AdaBoostClassifier
- DecisionTreeClassifier
- XGBClassifier

4. Print the Loss, Accuracy, Precision, Recall, F1 Score
5. Plot the models' performance.
6. Perform Preprocessing and Data Cleaning operations then run the models and compare the performance before and after cleaning.

Interview Questions

- **Scenario:** You are developing a machine learning model to classify news articles into multiple categories such as politics, sports, technology, and entertainment. The dataset contains thousands of articles, each labeled with one of these categories. The text data is highly dimensional due to the vast vocabulary and the dataset includes both short and long articles.
- **Question:** Discuss how you would approach building this classifier. What steps would you take to handle the high dimensionality and variability in article length? Explain your choice of classification algorithm considering the nature of the data and the need for model interpretability.

Scenario: You have developed a model to predict whether a customer will default on a loan. The initial model, a logistic regression, performs adequately, but you believe performance can be improved.
Question: Describe the steps you would take to optimize this model. What alternative models might you consider and why? How would you handle the deployment of this model in a production environment to ensure it remains accurate over time? Discuss how you would set up a feedback loop for continuous model improvement.

#*** Read and summarize Lesson 4