Description:

- Implement the ***naive Bayes classifier (NBC)*** and ***logistic regression with gradient descent*** algorithms in Python or MATLAB to solve a ***binary classification problem*** for a dataset of your choosing with continuous input and binary output.
- You may use all built-in functions. The built-in NBC and logistic regression models in Python and MATLAB that can be used for this assignment are:

o Python - Gaussian naïve Bayes (for continuous input): https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
o Python – logistic regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html with the "loss" parameter set to 'log_loss'
o MATLAB - Gaussian naïve Bayes (for continuous input): https://www.mathworks.com/help/stats/classificationnaivebayes.html (default is Gaussian/normal)
o MATLAB – logistic regression: https://www.mathworks.com/help/stats/incrementalclassificationlinear.html with the "learner" parameter set to 'logistic'

Train and test your models with a dataset of your choosing that meets the following criteria:

o Number of input features: 3+
o Input characteristics: Continuous real-valued
o Output characteristics: Binary

Note: If you find a dataset that has 3 or more output classes, you can use it by removing the data examples from other classes

- Use 80% of the dataset for training and 20% for testing your model.
- Use one of the following sources to find a dataset:

o University of California, Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php
o Kagglehttps://www.kaggle.com/
o Awesome Public Datasets https://github.com/awesomedata/awesome-public- datasets
o GoogleDatasetSearchEnginehttps://datasetsearch.research.google.com/ o Microsoft Research Open Data https://msropendata.com/
o U.S. Government's Open Data https://www.data.gov/
o Registry of Research Data Repositories https://www.re3data.org/
o CMU Libraries https://guides.library.cmu.edu/machine-learning/datasets

Summarize your approach and results in a report that includes at least the following:
o The dataset you used, its source and characteristics.
o The data preprocessing steps you took (if any).
o Thesolution$\boldsymbol{w}$(parametervector)forlogisticregression.Thisvectorshould include the intercept (bias term).

- Relevant evaluation metrics for NBC (accuracy, sensitivity, specificity, f1 score, log loss) for BOTH the training and test datasets.
- Relevant evaluation metrics for logistic regression (accuracy, sensitivity, specificity, f1 score, log loss) for BOTH the training and test datasets.