Kristy Nguyen

NSHE: 5006243601

Thomas Bryant

NSHE: 2000193948

CS 422-1001

Assignment #3

Assignment 3 Summary Report

**Dataset, Source, and Characteristics**

We used the ‘Breast Cancer Wisconsin (Diagnostic) Data set for this project from data.world (https://data.world/health/breast-cancer-wisconsin). The characteristics of the dataset have 30 real-valued input features with the output characteristics being either malignant or benign for the diagnosis.

**Data Preprocessing Steps**

For data preprocessing, we first have to use the StandardScaler() object to help standardize the input features by scaling them to unit variance. Then, using the fit\_transform() to get the mean and standard deviation for each input feature from the training dataset, and then transforms the training dataset via standardization. Finally, we use transform() on the test dataset as it will use the parameters from the training set to transform the test dataset – ensuring consistency and reliability in scaling both datasets.

**Relevant Evaluation Metrics for NBC**

|  |  |  |
| --- | --- | --- |
| Training Dataset Metrics: | B | M |
| Accuracy | 0.96 | 0.90 |
| Sensitivity | 0.96 | 0.90 |
| Specificity | 0.90 | 0.96 |
| f1 score | 0.98 | 0.95 |

|  |  |  |
| --- | --- | --- |
| Test Dataset Metrics: | B | M |
| Accuracy | 0.96 | 0.89 |
| Sensitivity | 0.96 | 0.89 |
| Specificity | 0.89 | 0.96 |
| f1 score | 0.98 | 0.94 |
| log loss | 0.58 | 0.58 |

