Kristy Nguyen

NSHE: 5006243601

Thomas Bryant

NSHE: 2000193948

CS 422-1001

Assignment #4

Assignment 4 Summary Report

**Dataset, Source, and Characteristics**

As per the instruction given in the assignment, we used the ‘Breast Cancer Wisconsin (Diagnostic) Data set for this project from data.world that was used in the previous assignment (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**

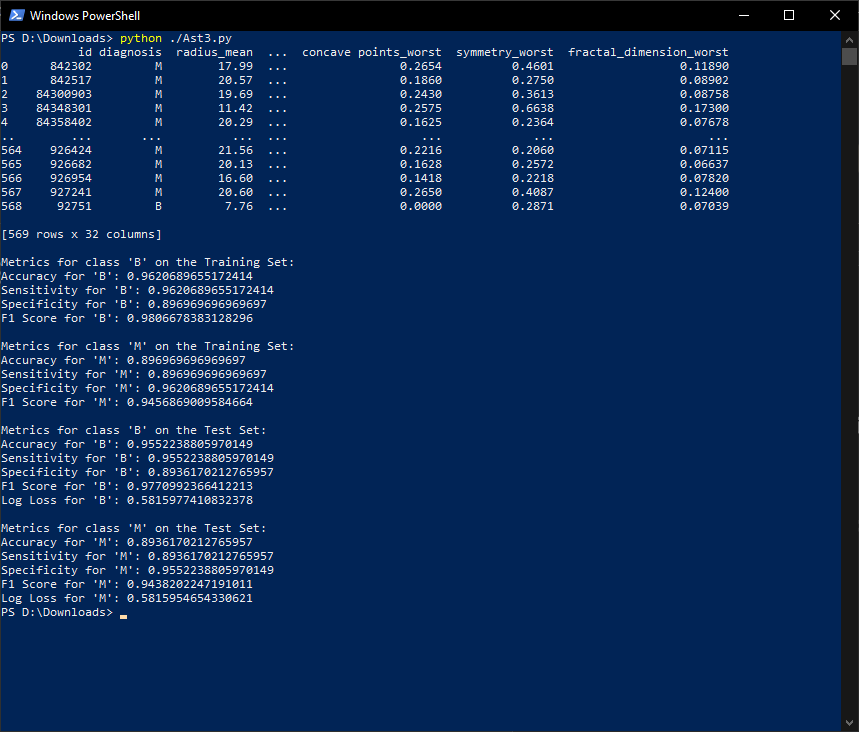
For data preprocessing, we first have to use **StandardScaler()** to help standardize the input features by scaling them to unit variance. Then using **fit\_transform()**, we obtain the mean and standard deviation for each input feature from the training dataset, and then we transform the training dataset 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 Naive-Bayes Classifier**

| Training Dataset Metrics: | B | M |
| --- | --- | --- |
| Accuracy | 0.9621 | 0.8970 |
| Sensitivity | 0.9621 | 0.8970 |
| Specificity | 0.8970 | 0.9621 |
| f1 score | 0.9807 | 0.9457 |

| Test Dataset Metrics: | B | M |
| --- | --- | --- |
| Accuracy | 0.9552 | 0.8936 |
| Sensitivity | 0.9552 | 0.8936 |
| Specificity | 0.8936 | 0.9552 |
| f1 score | 0.9771 | 0.9438 |
| log loss | 0.5816 | 0.5816 |

**Naive-Bayes Classifier:**

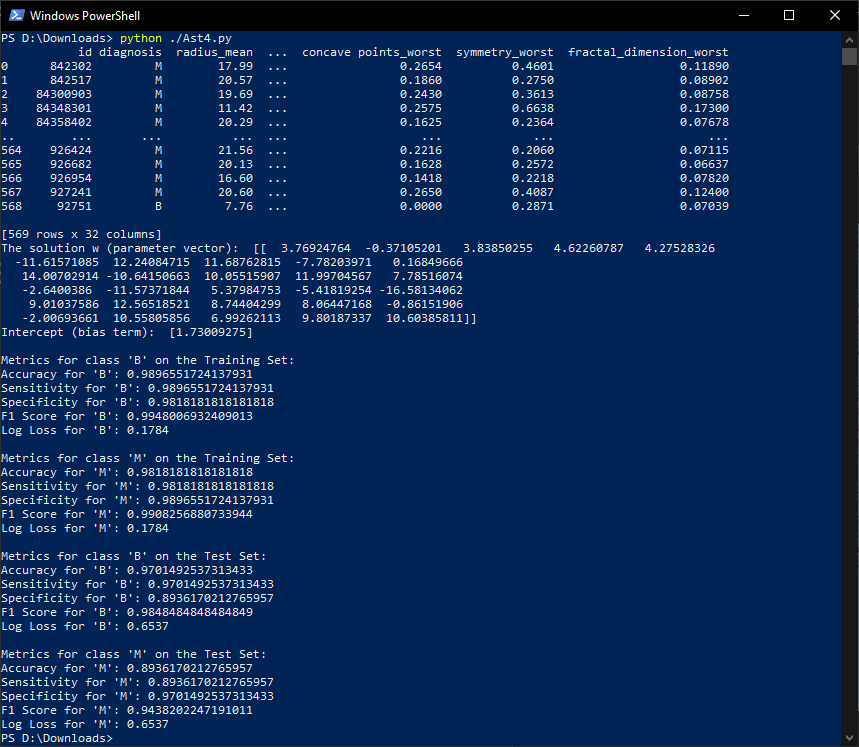


**Relevant Evaluation Metrics for Logistic Regression**

| Training Dataset Metrics: | B | M |
| --- | --- | --- |
| Accuracy | 0.9897 | 0.9818 |
| Sensitivity | 0.9897 | 0.9818 |
| Specificity | 0.9818 | 0.9896 |
| f1 score | 0.9948 | 0.9908 |
| Log Loss | 0.1784 | 0.1784 |

| Test Dataset Metrics: | B | M |
| --- | --- | --- |
| Accuracy | 0.9701 | 0.8936 |
| Sensitivity | 0.9701 | 0.8936 |
| Specificity | 0.8936 | 0.9701 |
| f1 score | 0.9848 | 0.9438 |
| log loss | 0.6537 | 0.6537 |

**Logistic Regression:**



**Model Comparison:**While the logistical regression model performed slightly better than the naive Bayes classification model, by only between 1% to 2% overall in performance improvement. It is also worth noting that while the logistic regression model performed well in identifying the malignant cells in the training data, the test data almost performed 10% worse in the test data - which given the heavier cost of missing a malignant tumor, would be worth considering in determining which model to use for predicting on this data set.