Assignment 2 Report

From the datasets available, we chose to work with the Breast Cancer Dataset[[1]](#footnote-1). While loading the data we made the following determinations:

1. Number of Features
   1. data-numpy array of shape (569, 31) = **30 Features**
2. Names of the Features
   1. Feature Names: ['radius1', 'texture1', 'perimeter1', 'area1', 'smoothness1', 'compactness1', 'concavity', 'concave\_points', 'symmetry1', 'fractal\_dimension1', 'radius2', 'texture2', 'perimeter2', 'area2', 'smoothness2', 'concavity2', 'concave\_points2', 'symmetry2', 'fractal\_dimension2', 'radius3', 'texture3', 'perimeter3', 'area3', 'smoothness3', 'compactness3', 'concavity3', 'concave\_points3', 'symmetry3', 'fractal\_dimension3']
3. Name of Target
   1. Target Name:  **Diagnosis**
4. Number of Samples
   1. data-numpy array of shape (569, 31) = **569 Samples**
5. First Five Rows of the Data
   1. Row 0 : [ 17.99 10.38 122.8 1001. 0.118 0.278 0.3 0.147

0.242 0.079 1.095 0.905 8.589 153.4 0.006 0.049

0.054 0.016 0.03 0.006 25.38 17.33 184.6 2019.

0.162 0.666 0.712 0.265 0.46 0.119]

* 1. Row 1 : [ 20.57 17.77 132.9 1326. 0.085 0.079 0.087 0.07

0.181 0.057 0.543 0.734 3.398 74.08 0.005 0.013

0.019 0.013 0.014 0.004 24.99 23.41 158.8 1956.

0.124 0.187 0.242 0.186 0.275 0.089]

* 1. Row 2 : [ 19.69 21.25 130. 1203. 0.11 0.16 0.197 0.128

0.207 0.06 0.746 0.787 4.585 94.03 0.006 0.04

0.038 0.021 0.022 0.005 23.57 25.53 152.5 1709.

0.144 0.424 0.45 0.243 0.361 0.088]

* 1. Row 3 : [ 11.42 20.38 77.58 386.1 0.142 0.284 0.241 0.105 0.26

0.097 0.496 1.156 3.445 27.23 0.009 0.075 0.057 0.019

0.06 0.009 14.91 26.5 98.87 567.7 0.21 0.866 0.687

0.258 0.664 0.173]

* 1. Row 4 : [ 20.29 14.34 135.1 1297. 0.1 0.133 0.198 0.104

0.181 0.059 0.757 0.781 5.438 94.44 0.011 0.025

0.057 0.019 0.018 0.005 22.54 16.67 152.2 1575.

0.137 0.205 0.4 0.163 0.236 0.077]

1. Histogram of Radius1

A blue graph with white text

Description automatically generated

1. Histogram of Texture1

A graph of a blue column

Description automatically generated with medium confidence

1. Scatter Plot of Perimeter vs. Symmetry Based on Target

A diagram of red and blue dots

Description automatically generated

Knowing that our problem is a binary classification between Malignant and Benign for Breast Cancer we followed the k-NN algorithm. Before we started training the model, we had to determine the correct number of neighbors to be used in the model. To determine this, we compared test accuracy verse training accuracy with differing numbers of neighbors with the lowest number of neighbors being 1 and the highest being where *n* is the number of samples in our dataset. From this we computed the created the following chart.

A graph of a graph with blue and orange lines

Description automatically generated

Using the previous chart, we decided to use 5 as our number of neighbors. To check how well our choice is we ran this number of neighbors, with the same training and test data as before, with Cross-validation and StratifiedKFold using 5 folds. The following chart shows the resulting scores.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Mean |
| Training Accuracy | 0.953 | 0.871 | 0.976 | 0.859 | 0.918 | 0.915 |
| Test Accuracy | 1.0 | 0.966 | 0.862 | 1.0 | 0.893 | 0.950 |

The lower training accuracy when compared to the test accuracy was expected as the Accuracy vs. N\_Neighbors chart labeled the accuracy to be higher at this number of neighbors.

1. <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic> [↑](#footnote-ref-1)