Breast Cancer Detection Using K-Nearest Neighbor Algorithm

1-Introduction:

Breast cancer is one of the moat common types of cancer to occur in women in the world. This type cancer affects a significant percentage of women affecting and risking their life. Even after the huge development in the diagnosis field and the appearance of new technologies it's still so hard to accurately diagnose the cancer and even harder to determine whether it's a benign or a malignant cancer.

In the sack of improving the accuracy of the diagnosis of the breast cancer we have implemented a K-Nearest Neighbor Algorithm using many distance metrics such as Manhattan distance, Euclidean distance, Chebyshev distance and Cosine distance and the different normalization techniques include Min-Max, Z-Score and Decimal Scaling.

We tested the accuracy of each technique and the highest result came from the Manhattan normalization with accuracy of 98.24% using KNN algorithm using Manhattan distance metric, with K=14, along with Decimal scale normalization.

2-Experiment Environment:

Python is used for implementing this algorithm.

3-Breast Cancer dataset:

This paper is using a dataset named Wisconsin (Diagnostic) Breast Cancer dataset which is obtained from the University of Wisconsin.

It compromises of 569 instances out of which 212 cases are malignant and 357 cases are benign.

The Features of this dataset are computed from a digitized image of a fine needle aspirate of a breast mass which describes characteristic of the cell nuclei present in the image.

Ten real-valued features are figured for every cell core:

1. Radius

This is calculated by taking the mean of all the distances from the center of the cell nuclei to the points on its perimeter.

2. Texture

The texture of a cell nucleus (Street, Wolberg, Mangasarian & Goldgof, 1993) is calculated by taking the standard deviation of the grayscale values.

3. Perimeter (the sum of the distances between consecutive boundary points)

The total perimeter is calculated by summing up the distances between consecutive boundary points of the cell nuclei.

4. Area

The nuclear area (Street, Wolberg, Mangasarian & Goldgof, 1993) is calculated simply by counting the number of pixels on the interior of the cell and adding one-half of the pixels in the perimeter.

5. Smoothness

The smoothness of a nuclear contour (Street, Wolberg, Mangasarian & Goldgof, 1993) is quantified by measuring the distance between the length of the radial line and the mean length of the lines surrounding it.

6. Compactness

Compactness is used to calculate how compact a cell is. It is calculated by squaring the perimeter of a cell per its unit area.

7. Concavity

It is the estimation of the severity of concave portions of the contour.

8. Concave points

It is the measure of number of concave portions of the contour.

9. Symmetry

It is evaluated by taking the relative contrast between pairs of lines perpendicular to the axis of the cell.

10. Fractal dimension

The fractal dimension is approximated using the “Coastline Approximation” described by Mandelbrot (Mandelbrot, 1977).

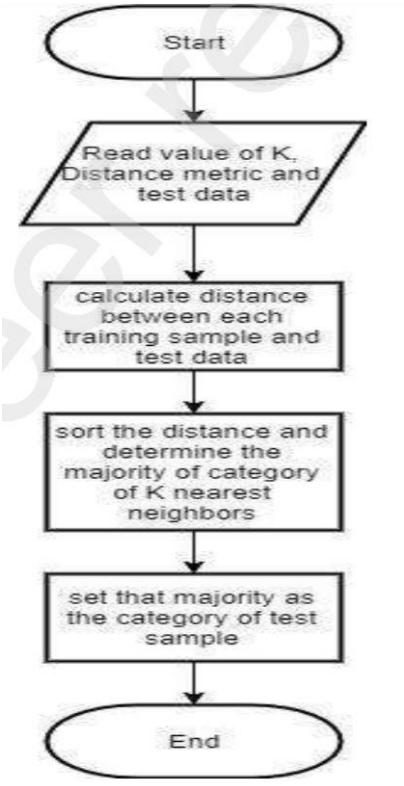
The mean value, worst case and standard error value are computed for each of the features, thus making it a total of 30 attributes. Apart from these, the other 2 attributes in the data set are id and diagnosis result.

4-K-Nearest Neighbor (K-NN):

In the K-Nearest Neighbor algorithm objects are classified by the majority of its neighbors.

The value of K determines the number of neighbors to be considered for classification.

The figure below describes the flow chart of the K-Nearest Neighbor Algorithm.



5-Normalization Techniques:

Normalization is considered as a preprocessing stage which is a scaling technique used to transform all the features of a dataset into an equal wight.

6-Distance Metrics:

Distance Metrics are used for majoring the distance between each point and its nearby points.

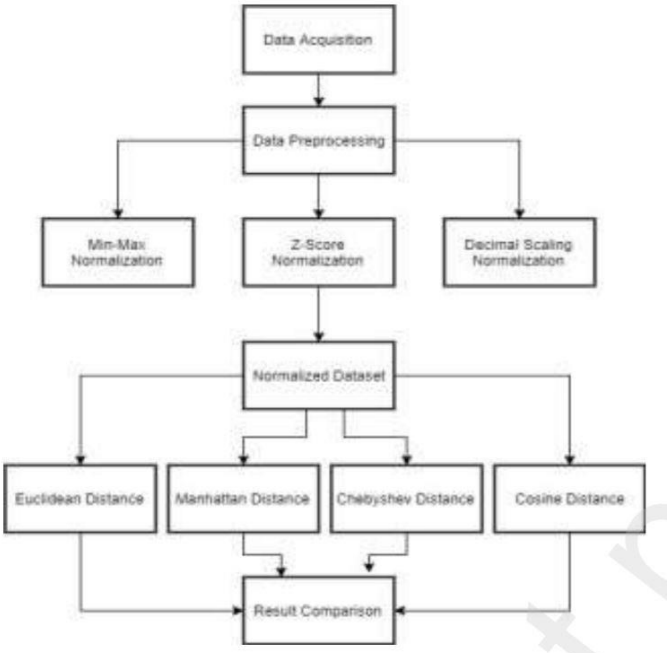
Selecting better distance function plays a crucial role in improving the performance of our algorithm.

7-Block Diagram:

We start performing the preprocessing of the data using the normalization technique which are Min-Max, Z score, and Decimal scaling normalization as shown in figure 2.

After that, we start calculating the distance between the test sample and the training instances using the different distance metrics which are Euclidean, Manhattan, Chebyshev and Cosine.

Then we start comparing the result of each sequence of these preprocessing techniques to find out which one is the most accurate one of them.



8-Collective Summarization of obtained results:

DISTANCE FUNCTION / NORMALIZATION TECHNIQUE	EUCLIDEAN DISTANCE	MANHATTAN DISTANCE	CHEBYSHEV DISTANCE	COSINE DISTANCE
MIN-MAX	95.6140, K=8	97.3684, K=6	92.1053, K=6	86.8421, K=14
Z-SCORE	95.6140, K=4	95.6140, K=6	94.7368, K=4	94.7368, K=4
DECIMAL SCALING	94.7368, K=4	98.2456, K=14	48.2456, K=4	68.4211, K=10

9-Experiamental results:

After performing the K-Nearest Neighbor algorithm using 3 different normalization techniques which are Min-Max, Z-score, and Decimal Scaling normalization techniques along with 4 different distance metrics which are Euclidean, Manhattan, Chebyshev and Cosine.

The accuracy percentage of each of the 12 combinations was noted in the previews table.

We can conclude that the most accurate method of implementing the K-Nearest Neighbor Algorithm for diagnosing Breast Cancer is by using Decimal Scaling normalization technique, Manhattan Distance Function, and K value equal to 14.

By using this combination we have reached an accuracy percentage of 98.2456%.