**Introduction**

Cancer is a disease characterized by rapid growth of cells in the body, in form of tumor. Tumors can either be malignant (cancerous) or benign (non-cancerous). Malignant tumors destroy healthy body tissues. The breast cancer is a form malignant tumor that developed from cells in the breast.

Breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among women. It is estimated by world health organization’s International Agency for Research on cancer (IARC) that more than 400,000 women expire each year due to breast cancer.

A key factor in this trend is the early detection and accurate diagnosis of this disease. This is possible by performing various tests like MRI, mammogram, ultrasound and biopsy but these tests lacks high diagnostic accuracy in identifying malignant breast cancer.

The objective of this project is to develop a predictive model with the help of a Classification algorithm, k-Nearest Neighbour algorithm (kNN) that can classify cancerous tumor as benign or malignant with high accuracy and better performance. This will help pathologists to take accurate and timely decisions regarding this disease.

**Literature Review**

The research on medical diagnosis of breast cancer has been conducted several times in past and majority of them reported high classification accuracies.

In 1996, Quinlan reached 94.74% classification accuracy using 10-fold cross-validation with C4.5 decision tree method.

In 2007, Polat and Gunes, an accuracy of 98.53% was obtained through least square SVM.

In 2000, Setiono reported accuracy of 98.10 based on a feed forward neural network rule extraction algorithm.

**Dataset**

In this project, we will be downloading and preprocessing the Wisconsin Breast Cancer Dataset which was collected from University of Wisconsin Hospitals and made available at UCI Machine Learning repository. It comprises 699 patients with a total of 11 different variables. These variables will be used to create model that predict diagnosis as benign or malignant of a particular patient. Each record in the dataset represents one breast cancer tissue sample. Below is a summary of the attributes of data.

1. Sample code number: ID number

2. Clump Thickness

3. Uniformity of Cell Size

4. Uniformity of Cell Shape

5. Marginal Adhesion

6. Single Epithelial Cell Size

7. Bare Nuclei

8. Bland Chromatin

9. Normal Nucleoli

10. Mitoses

11. Class: (2 for benign, 4 for malignant).

The dataset consists of nine features (from #2 to #10 above), each of which represented as an integer between 1 and 10. The 458 samples of the dataset belong to benign class (shown as 2) and other 241 are of malignant class (shown as 4). Please note that these features are computed from biopsy results of breast masses that describes the characteristics of the cell nuclei in the image.

*URL to download data:*

<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data>

**Approach**

After downloading dataset, headers were added with the help of R coding in RStudio. Then ID attribute was removed as it was redundant for our experiment. The 16 records that have missing data were also removed. Then all nine features have been converted to numeric data type. The upcoming work is it to visualize data to understand features interaction and class distribution before creating our model.

R Codes are available at below link:

[**https://github.com/farazahma/CKME136-Capstone/tree/master/Literature-Review-and-Data-Description**](https://github.com/farazahma/CKME136-Capstone/tree/master/Literature-Review-and-Data-Description)

****