**Project Name –** Breast Cancer Classification

**Domain Knowledge –**

* Breast cancer is an uncontrolled growth of breast cells. Cancer occurs as a result of abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy.
* The genes are in each cell’s nucleus, which acts as the “control room” of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new.
* Cells take over as old ones die out. But over time, mutations can change cells by “turning on” certain genes and “turning off” others in a cell. Changed cell gains the ability to keep dividing without control, producing more cells just like it and forming a tumor
* A tumor can be benign (not dangerous to health) or malignant (dangerous to health).
* Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body.
* Malignant tumors are cancerous, where the cancer cells can invade and damage tissues and organs near the tumor.

I found three datasets for Breast Cancer Classification.

**Following is the first one:**

1. **Basic Dataset Details:**

Data description

1. Sample code number ID number

2. Clump Thickness 1 - 10

3. Uniformity of Cell Size 1 - 10

4. Uniformity of Cell Shape 1 - 10

5. Marginal Adhesion 1 - 10

6. Single Epithelial Cell Size 1 - 10

7. Bare Nuclei 1 - 10

8. Bland Chromatin 1 - 10

9. Normal Nucleoli 1 - 10

10. Mitoses 1 - 10

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

**Attributes Description**

**Clump thickness:** Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer.

**Uniformity of cell size/shape:** Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.

**Marginal adhesion:** Normal cells tend to stick together. Cancer cells tend to loose this ability. So loss of adhesion is a sign of malignancy.

**Single epithelial cell size:** Is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.

**Bare nuclei:** This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.

**Bland Chromatin:** Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.

**Normal nucleoli:** Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.

**Following is the second dataset:**

**Ten real-valued features are computed for each cell nucleus:**

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

These features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image

**For each characteristic three measures are given:**

* Mean
* Standard error
* Largest/ “worst”

**I am using this second type of dataset which has the following points:**

1. Mean, Standard error and “worst” calculated for each of the real – valued features. Hence these become total 30 features.
2. 31st feature is the class/diagnosis defined for each record (M = Malignant and B = Benign)
3. 32nd feature is the “ID” of the sample.

**Example –**

Sample Code Number – 1017122

Clump thickness – 8

Uniformity of Cell Size - 10

Uniformity of cell shape- 10

Marginal adhesion – 8

Single epithelial cell size – 7

Number of bare nuclei – 10

Bland chromatin – 9

Number of normal nuclei – 7

Mitosis – 1

Classified Class (Malignant) – 4

**Description of the above sample record -**

1. More Clump thickness specifies that the cell might be of malignant type.
2. Malignant cells vary in size. We can see that Uniformity of Cell Size and shape are maximum , hence classifying the cell to be of malignant type
3. Epithelial cell size and Marginal Adhesion is more indicating that the cell might be of type = malignant.
4. More the value of Bland chromatin specifies that the cells are more rough or harsh in texture. Hence classifying the record/cell into a malignant tumor type.