**Case study 1 - Breast Cancer Classification**

**Business challenge**

1. Most common cancer among women accounting 25% of all cancer
2. 2.1 Million people in 2015
3. Early diagnosis of cancer can relatably help in survival of life
4. Key challenge is underdiagnosis of breast cancer screening
5. Also key challenge is to classify tumors into malignant or benign
6. Research indicates that most experienced physicians can diagnose cancers with 79 percent accuracy.
7. 91 percent correct diagnosis is achieved using machine learning techniques in this case study.

-What is meaning of cancer

a very serious disease in which cells in one part of the body start growing and form lumps in a way that is not normal

The average size of a human cell is about 100 μm in diameter meaning 100 micrometer

In contrast to normal cells, cancer cells often exhibit much more variability in cell size—some are larger than normal and some are smaller than normal.

-What is malignant or benign

Tumors can be benign (noncancerous) or malignant (cancerous). Benign tumors tend to grow slowly and do not spread. Malignant tumors can grow rapidly, invade and destroy nearby normal tissues, and spread throughout the body.

-How doctors recognize breast cancer?

A mammogram is an X-ray of the breast. For many women, mammograms are the best way to find breast cancer early, when it is easier to treat and before it is big enough to feel or cause symptoms. Having regular mammograms can lower the risk of dying from breast cancer.

-How Machine learning will detect Braest cancer?

Transfer learning is effective in detecting breast cancer by categorizing mammogram images of the breast with general accuracy, sensitivity, specificity, precision, F-score, and accuracy

**Task**

-Our task is to classify tumors into malignant or benign tumors using features of patients from several cell images.

-1st step is the FANNI process, which simply means to extract the cell images from the tumor and at the extraction stage we do not know whether it is malignant or benign.

-Later on once we have the images of the cells, looking at the features of how the cells are like radius, texture, perimeter, area, smoothness, etc. we will identify based on the classification that whether the cell of the tumor is malignant or benign.

- Input we are going to give is 30 features that we got from the extraction of cells images .

-In this dataset we are going to have 569 instances

-Out of this 569 dataset, 212 are malignant and 357 are benign

-So we are going to have to have 2 main classes here, malignant and benign to describe our output data, whether the given image is malignant or benign.

- we have one class of malignant and other benign, we want to draw such line which separates both the classes andhelp us figure out the classes.

-So the objective of training our model using the support vector machine is to find out the best line that separate the two classes

- the line which separates two classes at maximum margin is called hyperplane

-in order to get this line of hyperplane, we use the maximum margin distance and that is the objective of support vector machine classifier.

- support vector machine is basically where we assume points of boundaries and we do not know what is the answer whether malignant or benign.

**Visualization of Data**

1. Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used **to visualize random distributions**.
2. Matplotlib and Seaborn **act as the backbone of data visualization through Python**. Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python.
3. Matplotlib-It is utilized for making basic graphs. Datasets are visualised with the help of bargraphs, histograms, piecharts, scatter plots, lines and so on.
4. Seaborn-Seaborn contains a number of patterns and plots for data visualization. It uses fascinating themes. It helps in compiling whole data into a single plot. It also provides distribution of data.

**Model Training**

1. Drop the target column so that we can train the features data.
2. Define variable y to the target column
3. Split the data into training and testing

**Building the Model**

1. From sklearn.svm which is support vector machine import support vector classifier(svc).
2. Also import classification report, confusion matrix for the purpose of evaluation of matrix
3. Fit the training data into the model

**Evaluating the Model**

1. **Using model.predict , predict the xtest which is the training data.**
2. **With the help of classification report understand the output numbers of ytest and ypredict**

**Improving the Model**



