**Work Integrated Learning Programmes Division**

**Machine Learning**

**S1-23_AIMLCZG565**

**First Semester, 2023 -24**

# Assignment 1 – Part 1

# Breast Cancer Detection Analysis

## Instructions for Assignment Evaluation

1. Please follow the naming convention as <Group no>_<Dataset name>.ipynb.

   Eg – for group 1 with a weather dataset your notebooks should be named as - Group1_WeatherDataset.ipynb.

2. Inside each jupyter notebook, you are required to mention your name, Group details and the Assignment dataset you will be working on.
3. Organize your code in separate sections for each task. Add comments to make the code readable.
4. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behavior.
5. Notebooks without output shall not be considered for evaluation.
6. Prepare a jupyter notebook (recommended - Google Colab) to build, train and evaluate a Machine Learning model on the given dataset. Please read the instructions carefully.
7. Each group consists of up to 3 members. All members of the group will work on the same problem statement.
8. Each group should upload in CANVAS in respective locations under ASSIGNMENT Tab. Assignment submitted via means other than through CANVAS will not be graded.
9. The executed ipynb file with clear subdivision of the codes and brief description of the purpose of respective code needs to be uploaded on Canvas. All the executed tables or graphs and results should be present in the ipynb file.
10. Only two files should be uploaded in canvas without zipping them. One is ipynb file and other one html output of the ipynb file. No other files should be uploaded.

# Problem Statement

Build a classifier to predict the outcome of a new patient with high accuracy. Also, remember that as a data-scientist working on healthcare problems, your intent should also be to minimize the number of false-negatives.

**Dataset: breastcancer dataset can be downloaded from drive**

https://drive.google.com/file/d/1EDaE8o8_ZgQKQihhrizqBO_T9nvsRbwP/view?usp=sharing

1. **Import Libraries/Dataset**

    a. Download the dataset
    b. Import the required libraries

2. **Data Visualization and Exploration [1 M]**

    **a.** Print at least 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
    **b.** Print the description and shape of the dataset.
    **c.** Provide appropriate visualization to get an insight about the dataset.
    **d.** Try exploring the data and see what insights can be drawn from the dataset.

3. **Data Pre-processing and cleaning [2 M]**

    **a.** Do the appropriate preprocessing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Apply appropriate feature engineering techniques for them.
    **b.** Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset.
    **c.** Do the correlational analysis on the dataset. Provide a visualization for the same.

4. **Data Preparation [2M]**

    a. Do the final feature selection and extract them into Column X and the class label into Column into Y.
    b. Split the dataset into training and test sets.

5. **Model Building [1 M]**

    **a.** Perform Model Development using logistic regression and decision tree, separately. Deep Learning Models are strictly not allowed.
    **b.** Train the model and print the training accuracy and loss values.

6. **Performance Evaluation [4 M]**

    a. Print the confusion matrix. Provide appropriate analysis for the same.
    b. Do the prediction for the test data and display the results for the inference.