# Name: Paras Kumar

# Batch: 06-DSAI

# Project Report:

# Breast Cancer Prediction Using Machine Learning

## 1. Introduction

Breast cancer is one of the most common cancers among women worldwide. Early detection through medical screening and diagnosis can help in reducing mortality rates. This project focuses on predicting breast cancer using machine learning techniques, leveraging clinical data. The project involves building a classification model that predicts whether a breast tumor is benign or malignant.

## 2. Objective

The main objective of this project is to create a machine learning model that can accurately predict whether a patient has benign or malignant breast cancer based on features extracted from their medical records.

## 3. Dataset Description

The dataset used in this project likely contains clinical and histopathological features of breast tumors. These features are used as input variables to predict the target variable (benign or malignant). Typically, the dataset for breast cancer prediction includes features like:

- Mean radius
- Mean texture
- Mean perimeter
- Mean area
- Mean smoothness

The target variable is:

- 0 for benign
- 1 for malignant

## 4. Libraries Used

The following Python libraries were used in the project:

- Pandas: For loading and managing the dataset.
- NumPy: For numerical computations.
- Matplotlib & Seaborn: For data visualization.
- Scikit-learn: For data preprocessing, model training, and evaluation.
- Warnings: To ignore minor, non-critical warnings.

## 5. Workflow

- The project follows a typical machine learning workflow, which can be broken down into several steps:

### 5.1 Data Preprocessing

- **Data Loading**: The dataset is loaded using pandas.
- **Exploratory Data Analysis (EDA)**: This step includes generating summary statistics, visualizing distributions of features, and identifying missing or erroneous data.
- **Data Cleaning**: Handling missing values, encoding categorical variables, and feature scaling are applied to prepare the data for model training.

### 5.2 Feature Engineering

- **StandardScaler**: Used to standardize the features by removing the mean and scaling them to unit variance.
- **LabelEncoder**: Encodes the target labels (benign or malignant) into numerical values.

### 5.3 Data Splitting

The dataset is split into:

- **Training set**: 70-80% of the data, used to train the model.
- **Test set**: 20-30% of the data, used to evaluate the model's performance on unseen data.

### 5.4 Model Building

Two machine learning models were explored:

- **Logistic Regression**: A linear model for binary classification.
- **Linear Regression**: Although this algorithm is generally used for regression tasks, it may have been explored for comparison.

### 5.5 Model Evaluation

The model's performance is evaluated using the following metrics:

- **Accuracy**: The percentage of correct predictions.
- **Classification Report**: Includes precision, recall, F1-score, and support for each class (benign/malignant).

### 6. Results and Discussion

After building the model and evaluating it on the test data, the performance of the model is assessed in terms of its ability to classify benign and malignant tumors. Common evaluation metrics include:

- **Accuracy**: Represents the proportion of correct predictions made by the model.
- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall (Sensitivity)**: The ratio of correctly predicted positive observations to all observations in the actual class.
- **F1 Score**: The weighted average of precision and recall, balancing both concerns.

### 7. Conclusion

- The project successfully developed a machine learning model to predict breast cancer. Based on the dataset and applied techniques, the logistic regression model likely demonstrated decent predictive accuracy. Early detection models like this can assist in medical diagnosis, but it's essential to validate these models with large, diverse datasets to ensure their robustness and generalization.