# PERSONALIZED MEDICINE: Breast Cancer Data Analysis and Modeling

This project is about **“PERSONALIZED MEDICINE”** on topic Breast Cancer Early Detection can improve targeted therapies. Cancer is a worldwide epidemic that affects individuals of all ages and backgrounds. However, breast cancer is one of the most common cancers in women. Predicting and diagnosing cancer at an early stage is an area where machine-learning approaches may have a significant impact. **A model can be trained to predict whether a lump is benign or malignant.** This could lead to faster, more accurate diagnosis, enabling timely interventions and potentially saving lives.

**About Dataset:**

This dataset was extracted from Kaggle. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Features describe characteristics of the cell nuclei present in the image.This database is also available in UW CS ftp server:[**ftp.cs.wisc.edu**](ftp://ftp.cs.wisc.edu)

**Features Information:**

ID number

Diagnosis

Unnamed

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

a) radius (mean of distances from centre to points on the perimeter)  
b) texture (standard deviation of g ray-scale values)  
c) perimeter  
d) area  
e) smoothness (local variation in radius lengths)  
f) compactness (perimeter^2 / area - 1.0)  
g) concavity (severity of concave portions of the contour)  
h) concave points (number of concave portions of the contour)  
i) symmetry  
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius. All feature values are recoded with four significant digits.

## Overview

This project involves exploratory data analysis (EDA), pre-processing, and feature engineering, and modelling on a breast cancer dataset. The goal is to classify the diagnosis of breast cancer using various machine learning techniques.

## Libraries Used

* **EDA and Visualization**: Pandas, Numpy, Matplotlib, Seaborn
* **Pre-processing and Modelling:** sklearn.preprocessing, sklearn.model\_selection, sklearn.linear\_model, sklearn.metrics,

sklearn.decomposition, sklearn.cluster, sklearn.ensemble.

1. **Load the Dataset**

* df\_Breast\_cancer = pd.read\_csv ("data.csv")

1. **View the Dataset**

* head(), tail(), info(), describe()

1. **Check for Missing Values and Outliers**

* Missing values: isnull().sum()
* Outliers: IQR method and boxplots

1. **Exploratory Data Analysis (EDA):**

* Drop irrelevant columns
* Univariate, bivariate, and multivariate analysis using histograms, scatter plots, and correlation matrix

1. **Feature Engineering:**

* Create new features: arp\_mean(area,radius,perimeter=per,p), arp\_worst(area,radius=rad,),  rad\_per\_se
* Log transformation

1. **Data Pre-processing:**

* Value counts and visualization
* Scaling: RobustScaler, StandardScaler
* Label encoding

1. **Model Evaluation:**

* Principal Component Analysis (PCA)
* Random Forest Classifier
* Metrics: accuracy score, f1\_score

1. **Clustering:**

* K-Means clustering and silhouette score

1. **Hyperparameter Tuning:**

* Logistic Regression with GridSearchCV

**Note:**

(FNA: Fine Aspiration Needle is a medical procedure done to obtain sample cells from breast lump.)

Diagnosis: Benign = B, Malignant = M

SE = Standard Error

Link: <https://github.com/Niveditha-93/Data_Science> kindly give pull request!