Project Proposal
Breast Cancer Prediction

Project title:

Prediction of Breast Cancer Using Data Collected from Fine Needle Aspiration (FNA) method.

Purpose and Outcome:

- *Purpose:* This study assesses the correlation between the collected features and the benign or malignant nature of breast tumors. Based on this analysis, a predictive model is developed to determine tumor malignancy.
- *Outcome:* Identify features that are strongly associated with breast cancer and to use them to accurately predict the malignancy of tumors.

Dataset:

- *Description:*
    - The Breast Cancer Wisconsin (Diagnostic) Dataset is a medical dataset widely used to predict whether a breast tumor is malignant (M) or benign (B). The data is derived from digitized images of fine needle aspirates (FNA) of breast masses, and the features describe cell nuclei characteristics.
    - Total Samples: 569
    - Features: 30
    - Target variable: **Diagnosis** (M= Malignant, B = Benign)

- *Source:* Wisconsin Diagnostic Breast Cancer (WDBC) dataset https://www.kaggle.com/datasets/khansaafreen/breastdataset?select=data.cs

- *Structure:*
    - Features Explained: The dataset contains **mean, standard error (SE), and worst values** for each measurement of the tumor's shape and texture
        - **Radius**: Radius of the tumor
        - **Texture**: Variation in grey-scale intensity
        - **Perimeter**: Tumor's boundary length
        - **Area**: Size of the tumor
        - **Smoothness**: How smooth the tumor surface is
        - **Compactness**: How compact or irregular the tumor is
        - **Concavity**: How deep the indentations are
        - **Concave Points**: Number of concave portions
        - **Symmetry**: Tumor symmetry
        - **Fractal Dimension**: "Roughness" of the tumor boundary
    - ID and Target column:
        - **ID**
        - **Diagnosis**: M= Malignant, B = Benign

Initial Analysis Plan

- **Data cleaning:** Handle missing values, correct data types, and remove duplicates.
- **EDA:** Generate summary statistics, remove multicolinearity features, find correlation between target column and features.
- **Analysis:** Build and evaluate Logistic regression model.
- **Visualization:** Create boxplots to check outliers of the features, create kernel density estimate (kde) to find the difference between 2 target variables, create heatmaps, scatter plots to check correlation between features.
- **Data Storytelling:** Presenting the predictive model and offering recommendations for subsequent diagnostic evaluations and treatment planning.