Group 11: An EDA on Malignant Breast Cancer Cells

Problem Statement

Breast cancer is one of the most commonly diagnosed cancers among women worldwide and the major contributor to the global cancer morbidity. According to the WHO data, in 2022, breast cancer ranked second in overall incidence and remained the most common cause of cancer death in women. In the United States specifically, breast cancer is the second leading cause of cancer death among women, only after lung cancer (CDC, 2025).

This project applies end-to-end Python data science to a public deidentified breast cancer dataset containing cell/tumor morphology features - radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension - with patients' attributes in the dataset such as age.

Our goals are to:

a) Quantify how malignancy rates vary with age and tumor characteristics
b) Identify which features most strongly distinguish malignant from benign tumors
c) Visualize trends with distribution plots and partial dependence views

Datasets

We will be using the following datasets in our project:

Breast Cancer Wisconsin (Original) Data Set

https://www.kaggle.com/datasets/mariolisboa/breast-cancer-wisconsin-original-data-set

Dataset obtained from the University of Wisconsin Hospitals, including:
Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses.
All quantitative.

Breast Cancer Coimbra

https://www.kaggle.com/datasets/atom1991/breast-cancer-coimbra

Dataset originated from UC Irvine Breast Cancer Coimbra dataset, including:
Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1.
All quantitative.

Breast Cancer Wisconsin (Diagnostic) Data Set

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Dataset with features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Including:
Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, Fractal Dimension.

Current Knowledge and Our Analysis

Our goal is to build an analytical framework for understanding how tumor morphology and patient factors relate to malignancy risk and severity. Although there is prior research on breast cancer diagnostics, we see areas to expand on. Many factors, such as how shape irregularities (concavity, compactness) combined with size- and texture-based features are still underexplored. We aim to focus on the

intersection of these signals, and move beyond simple conclusions. Visualizations we are looking to utilize are:

- Box/violin plots to compare key morphometrics across benign vs. malignant cases
- Regression/likelihood plots to examine how features like concavity relate to malignancy probability
- Correlation heatmaps (and clustering) to rank features and expose redundancy
- Stacked bar plots to explore malignancy prevalence across age brackets and tumor size/texture bins
- Feature-importance and SHAP charts to identify the strongest global and local predictors

Real-world applications include: informing triage by prioritizing high-risk cases for biopsy, improving model transparency for clinicians and patients through feature-level explanations, flagging subgroups where calibration drifts, and guiding data-driven quality checks for screening workflows.

Project Steps

Steps	Description	Timeline	Group Members
Data Collection & Cleaning	Gather and preprocess datasets, handle missing values, normalize fields.	Finished by Monday, November 24th	Zhenhao and Jiaqi
Exploratory Data Analysis	Visualize distributions, correlations, and variable relationships.	Finished by Monday, December 1st	Zhijian, Jiaqi, and Juliana
Visualization & Insights	Create visuals to showcase our findings.	Finished by Sunday,	Shijie and Juliana