# INTRODUCTION:

At first, I had planned to examine the diabetes dataset for my project. However, due to insufficient data, I have opted to focus on breast cancer prediction for the final project. Breast cancer is a type of cancer that affects the cells in the breasts. It is a significant concern for women and can also occur in men. Fortunately, thanks to scientific advancements, the mortality rates associated with breast cancer are decreasing. Detecting breast cancer at an early stage can greatly improve the chances of survival.

# DATASET SOURCE:

I found the breast cancer dataset on Kaggle (Pmotta. (2021, June 6). *Breast cancer prediction*. Kaggle) to be a helpful starting point, as it provides valuable information required for the analysis.

# MODELS:

I plan on building the following models with this dataset.

1. Decision Tree Implementation:

A decision tree is a useful tool for breaking down complex data into more manageable parts. From what I understand, decision trees are often used for prediction analysis, data classification, and regression.

1. Random Forest Classification:

Random forest classification helps with any overfitting issues with decision trees. Random Forest Classification also maintains its accuracy in case of missing data.

1. KNN Classification:

Since accuracy is also an important factor in prediction, I would like to test the KNN Classification model. Based on the outcomes from the models, I would also like to explore other models.

How do you plan to evaluate your results:

To evaluate a model's results effectively, it's important to go beyond the accuracy score alone. As such, I plan to construct a confusion matrix by utilizing the train-test split (80/20), cross\_val\_score, and accuracy score for a comprehensive evaluation of the model.

## What do you hope to learn:

Using this analysis, I am hoping to understand the factors and symptoms that cause Breast Cancer. Based on the outcome the objective is to help early detection of breast cancer.

## Assess any risks and ethical implications with your proposal:

The Kaggle dataset is sourced from samples that arrive periodically as Dr. Wolberg reports his clinical cases. The dataset has not been updated in years making it unclear if other factors could potentially cause breast cancer.

## Identify a contingency plan if your original project plan does not work out:

At this point, I am hoping we have enough information to begin the modeling process. However, based on my previous experience, I encountered an issue where I began to doubt the accuracy of my models and suspected that the dataset lacked essential features to make accurate predictions. In preparation for this scenario, I started working on a similar dataset within the same field. Similarly, for my current project, I will have a contingency plan based on the same grounds. I will primarily look for similar and more recent datasets in the same domain. As an alternative option, if the dataset doesn’t work, I’ll switch to the airline delay prediction dataset.

## Include anything else you believe is important:

Feature selection is key for model building. I plan on spending some time understanding the dataset to ensure the right features are selected for model building. I would also like to explore other models such as XGBClassifier and/or GaussianNB.

REFERENCES:

Pmotta. (2021, June 6). *Breast cancer prediction*. Kaggle - <https://www.kaggle.com/code/pmotta/breast-cancer-prediction/input>