Breast Cancer Prediction

Milestone 3 - Preliminary Analysis

School: Bellevue University

Date: 07/12/2023

Class: DSC630-T301 Predictive Analytics (2237-1)

## Data Expectations

The selected breast cancer dataset contains key features that reflect the underlying health factors of breast cancer patients, which should be adequate for prediction purposes. However, before modeling, the dataset needs to be prepared and data balance checks must be conducted.

If I am unable to find the answers with the selected dataset, as an alternative, I have identified another dataset for breast cancer prediction (Architkuiya. (2022, November 28). Breast\_cancer\_detection).

## Visualizations

For this analysis, I would like to use the following visualizations:

* + 1. A Bar/Pie chart with the number of benign vs malignant tumors to identify data balance. This is useful to identify any imbalance in the dataset before modeling to avoid bias.
    2. Different Bar charts for features such as (Cell size, Clump Thickness, and Cell shape) vs Class (Benign/Malignant) to see the effect of these features on cancerous cells.
    3. A correlation matrix (heatmap) to observe the correlations between features.

I would also like to build line and/or scatter plots based on features with a strong correlation with the class variable (Benign/Malignant), based on the outcome from the correlation heatmap. However, since we already would have built a bar chart for different features vs Class, I doubt if the scatter/line charts would add any additional value.

## Data Preparation

As a first step in data preparation, I would perform some checks such as

* + 1. Duplicate checks – Remove duplicate rows from the dataset
    2. NA/Null checks – Replace missing values with the median value of the column
    3. Drop empty rows and columns, if any.

Additionally, if from visualization (1), the dataset turns out to be imbalanced, I would have to implement techniques such as SMOTE (Synthetic Minority Oversampling Technique) to balance the data before model building.

## Additional Model Selection

My main choices for modeling breast cancer prediction include Decision Tree, Random Forest, and KNN. If time allows, I also aim to investigate the implementation of an SVM model. This is because SVMs possess the capability to capture complex relationships among features, which is beneficial considering the non-linear nature of the data. SVMs are popular with health datasets, as they are recognized for their proficiency in performing both linear and non-linear classifications. The ultimate objective is to assess the accuracy of each model and determine which one yields the most precise predictions.

## Expectations

I believe my primary goal of predicting breast cancer is still achievable with the available dataset features. As a contingency plan, I have an alternate dataset for breast cancer and also an airline delay prediction dataset. I am confident that I can build decent models with the available breast cancer dataset.

## REFERENCES:

1. Architkuiya. (2022, November 28). *Breast\_cancer\_detection*. Kaggle. <https://www.kaggle.com/code/architkuiya/breast-cancer-detection/data>

**Corrections to Milestone 2 based on the feedback are highlighted below.**

# **MILESTONE - 2**

# 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 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 to build 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.

*7/12/2023:* I choose to evaluate the KNN Classification model as it is a non-parametric algorithm, which means it does not make assumptions about the underlying data distribution. This flexibility allows it to capture complex relationships between features, making it potentially suitable for breast cancer prediction. I do not know for sure if KNN would give a better accuracy over Random Forest Classification, but I assume the accuracy would be better as KNN is known for its ability to handle numerical data and can be effective in certain scenarios.

How do you plan to evaluate your results:

To evaluate a model's results effectively, it is 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.

*7/12/2023:* As such I plan to evaluate models using train-test-split, by dividing the dataset into a training and a test set. Training the model on the training set and evaluate its performance on the test set, which represents the distribution of real-world data.

I would also like to construct a confusion matrix which gives a comprehensive view of the model's performance for different classes and helps identify any specific issues, such as imbalanced class predictions.

## 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 does not work, I will switch to the airline delay prediction dataset.

## Include anything else you believe is important:

Feature selection is key for model building. I plan to spend 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>