Breast Cancer Prediction

Saurabh Biswas & Gourav Verma

DSC-630 Predictive Analytics

Bellevue University, NE

Fall-2020

# Abstract

Breast cancer or breast carcinoma is uncontrolled growth of epithelial cells in the breast. The uncontrollable division of one cell results in visible mass named tumor. Tumor can be benign or malignant. By Johns Hopkins Pathology, benign tumors are non-malignant/non-cancerous tumor and malignant tumors are cancerous growths. A cancer is another word for a malignant tumor. Most benign tumors respond well to treatment. But, malignant tumors are often resistant to treatment, may spread to other parts of the body and they sometimes recur after they were removed. Under a study by University of Wisconsin, 569 patients (212 with cancer and 357 with benign masses) provided the data for diagnostic algorithm. Diagnostic features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34]. Breast cancer is the most common malignancy in women, with over 200,000 being diagnosed in the US every year. 40,000 women will die from it each year.

*Keywords:* K-Mean, AFINN, elbow method, clustering, scikit-surprise, recommendation model, BaselineOnly.

Table of Contents

[Abstract 2](#_Toc52620728)

[Data 5](#_Toc52620729)

[Methods 6](#_Toc52620730)

[Results 6](#_Toc52620731)

[Discussion 6](#_Toc52620732)

[Acknowledgments 6](#_Toc52620733)

[References 6](#_Toc52620734)

# Problem Statement

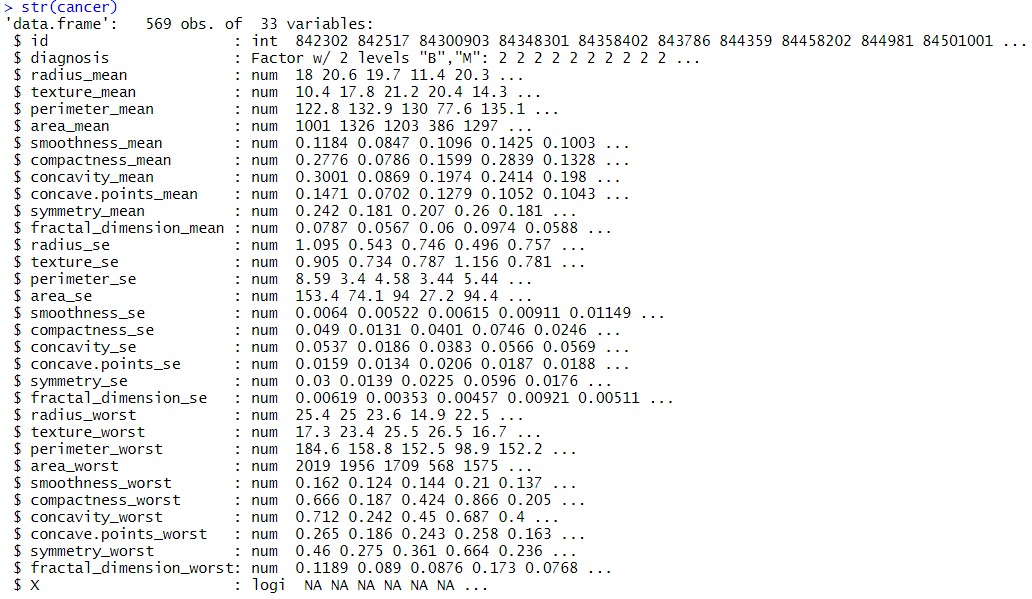
It is 2nd most common cancer in women. But, on rare scenario can also happen to men. It is 2nd leading cause of deaths of women after lung cancer. A fine needle biopsy is an effective tool in evaluating and diagnosing suspect lumps or masses. With early diagnosis of breast cancer, patients can be isolated for early treatment for a better chance of survival. Common biopsies for breast cancer diagnosis includes fine-needle aspiration (FNA), core needle biopsy, and MRI-guided biopsy. In this analysis, we will be using ten features of tumor cell nuclei extracted from the digital image processing of an FNA of a breast mass to predict breast cancer. The data is collected from UCI Machine learning repository.

# Data

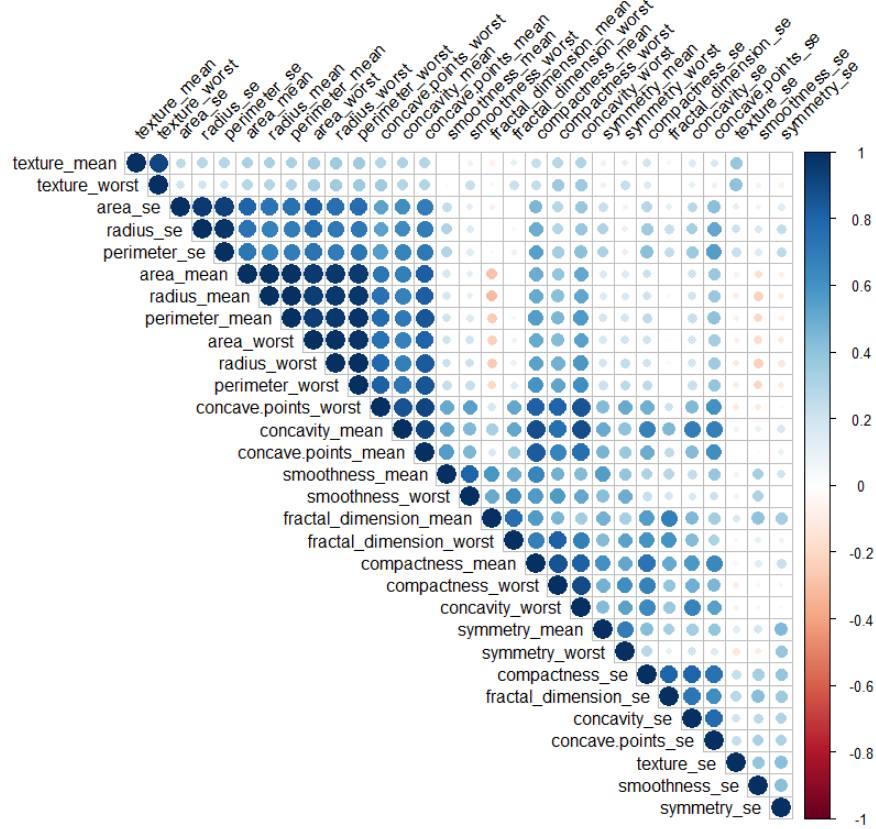
We are using the Wisconsin Breast Cancer [dataset](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29) was obtained from the University of Wisconsin Hospitals. As described the website following 10 features were computed for each cell nucleus and will be used as input for our analysis. All feature values are recoded with four significant digits.

1. radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. 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. There are no missing values in the dataset. The total numbers of data samples with different labels are not balanced. In fact, the total number (357) of data samples labeled as B (benign) is almost doubled the total number (212) of data samples labeled as M (malignant). Below is the statistical summary of the dataset.



Correlation between different variables.



# Methods

# Results

# Discussion

# Acknowledgments

# References