Beyond the Cure: Modeling the Risk of Breast Cancer Recurrence

Brandon Thomas

bthomas@bellarmine.edu

06 February 2024

Exploratory Data Analysis Report

**Introduction**

The goal of this project is to develop a predictive model for breast cancer recurrence, leveraging advanced machine learning techniques to improve clinical decision-making and patient outcomes. Breast cancer recurrence, whether local or metastatic, remains a significant challenge in cancer care, and accurate prediction of recurrence risks is crucial for personalized treatment planning and long-term monitoring. The project will utilize a comprehensive dataset, which includes clinical and demographic features like tumor size, lymph node status, and age among other variables that was obtained from the UCI Machine Learning Repository.[[1]](#footnote-1) The following report includes a data set description, summary statistics, graphical exploration, concluding with a summary of initial findings.

**Data Set Description**

The Breast Cancer dataset from the University Medical Centre, Institute of Oncology in Ljubljana, Yugoslavia, is a widely recognized resource in the machine learning community. It is frequently used in research for classification tasks related to breast cancer recurrence. The dataset comprises 286 patient cases, each described by nine attributes that provide essential information about the patients and their clinical conditions. With its relevance to healthcare and machine learning, this dataset has proven to be instrumental in developing predictive models that can aid in understanding the factors influencing breast cancer recurrence. The primary target variable in the dataset is the "Class" variable which indicates whether a patient experienced a recurrence of breast cancer. The possible outcomes are "no-recurrence-events" and "recurrence-events," with 201 instances belonging to the former and 85 to the latter. This imbalance in class distribution reflects the real-world scenario, where most patients do not experience recurrence. The remaining attributes capture demographic, pathological, and treatment-related details, making the dataset rich with predictive potential.

One key attribute is "Age," which describes the patient’s age at the time of diagnosis, categorized into decades such as "30-39," "40-49," and so on. Another significant attribute is "Menopause," which identifies the menopausal status of the patient, classified into "lt40" (less than 40), "ge40" (greater than or equal to 40), and "premeno" (premenopausal). These attributes are critical for understanding how age and hormonal factors contribute to breast cancer recurrence. Pathological attributes in the dataset include "Tumor Size," which measures the size of the tumor in millimeters and is grouped into intervals such as "0-4" and "5-9." Similarly, "Inv-Nodes" describes the number of positive axillary lymph nodes detected, that is, lymph nodes to which cancer has metastasized, with ranges like "0-2," "3-5," etc. These attributes are important markers of disease progression. Additionally, "Node-Caps" indicates whether cancerous nodes exhibit capsular invasion, recorded as either "yes" or "no," while "Deg-Malig" reflects the degree of malignancy on a scale of 1 (low), 2 (medium), and 3 (high). Capsular invasion in cancer refers to the spread of cancer cells beyond the capsule—a fibrous, protective layer or membrane that often surrounds organs or tumors. Normally, this capsule serves as a barrier, containing the tumor and preventing it from invading nearby tissues. However, when cancer cells break through this capsule and invade the surrounding tissues, it is referred to as capsular invasion. Further details about the tumor’s location and affected breast are captured through the "Breast" and "Breast-Quad" attributes. "Breast" specifies which breast was affected—either "left" or "right"—and "Breast-Quad" identifies the quadrant of the breast where the tumor is located, such as "left-up," "right-low," or "central." Finally, the dataset includes "Irradiat," which documents whether the patient underwent radiation therapy, with "yes" or "no" as possible values. The Breast Cancer dataset is not only comprehensive in its representation of patient and clinical data but also includes challenges such as missing values in certain attributes like "Node-Caps" (Table 1). This adds an additional layer of complexity for researchers, requiring careful preprocessing and imputation techniques before applying machine learning models.

**Table 1 – Data Description**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Data Type** | **Possible Values** | **Percent Missing Data (%)** |
| Class | Nominal | No-recurrence-events, recurrence-events | 0 |
| Age | Ordinal | 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 | 0 |
| Menopause | Nominal | lt40, ge40, premeno | 0 |
| Tumor Size | Ordinal | 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 | 0 |
| Inv-Nodes | Ordinal | 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 | 0 |
| Node-Caps | Nominal | Yes, no | 2.7972 |
| Deg-Malig | Ordinal | 1, 2, 3 | 0 |
| Breast | Nominal | Left, Right | 0 |
| Breast-Quad | Nominal | left-up, left-low, right-up, right-low, central | 0.3497 |
| Irradiat | Nominal | Yes, No | 0 |

**Data Set Summary Statistics and Graphical Exploration**

1. Numerical Summarization/ Correlation

A screenshot of a black and white calculator

AI-generated content may be incorrect.

Figure 1 – Brief Numerical Description of Dataset

As depicted in Figure 1, most of the variables included in the dataset are categorical. The figure illustrates total data points, unique data points, and most frequent data points for each variable.

|  |
| --- |
| A screenshot of a cell phone  Description automatically generated |

Figure 2 – Crosstabs of Each Variable Compared to Binary Class Variable

The crosstabs or contingency tables (Figure 2) display the relationships between ‘class’ and each other variable to examine the frequency distributions. Notable differences between recurrence classes include: there were no records of cancer recurrence in patients who experienced menopause at less than 40 years old; the proportion of cancer recurrence is greater in patients with left breast cancer; the proportion of recurrence is greater in patients with right upper tumor localization; and the proportion of recurrence is greater in patients who did receive radiation as part of their cancer treatment.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 3 – Correlation Matrix of Quantitative Variables

The correlation matrix (Figure 3) shows the relationships between different variables, with correlation values ranging from -1 to 1. Age has a very weak negative correlation with ‘tumor size’ (-0.0047), ‘invasive nodes’ (-0.064), and ‘class\_binary’ (-0.087), indicating that age has little to no influence on these factors. ‘Tumor size’ exhibits a weak positive correlation with ‘invasive nodes’ (0.16) and ‘class\_binary’ (0.19), suggesting that larger tumors may be slightly associated with a higher likelihood of positive classification and more invasive nodes. Additionally, ‘invasive nodes’ have a moderate positive correlation with ‘class\_binary’ (0.3), implying that a greater number of invasive nodes is somewhat linked to a positive classification. Overall, the correlations are relatively low, indicating weak linear relationships between these variables.

A graph of a number of dots

AI-generated content may be incorrect.

Figure 4 – Scatterplot of Inv-Nodes vs. Tumor-Size

When comparing the relationship between ‘tumor-size’ and ‘invasive-nodes,’ there is a slight positive relationship, more so for recurrent cancer cases (Figure 4). Generally, as ‘tumor-size’ increases the number of ‘inv-nodes’ also increases.

1. Frequency Distributions and Proportions

A graph with a line

Description automatically generated

Figure 5 – Probability Density Functions of Quantitative Variables

The probability density functions for ‘age’ and ‘tumor-size’ exhibit relatively normal, unimodal distributions (Figure 5). ‘Age’ was transformed from age ranges to the mean of each age range which is why there are significant slope changes in the distribution. “Inv-nodes” has a right-skewed distribution which reflects the trend seen in healthcare as there are fewer patients with many invasive nodes.

A group of blue bars

Description automatically generated

Figure 6 – Distributions of Qualitative Variables

The distributions of qualitative variables (Figure 6) is not entirely balanced especially for ‘irradiate’ and ‘node-caps.’ However, there should still be sufficient data to implement machine learning models.

1. Data Anomalies

9 rows were dropped containing missing data to maintain data integrity and avoid potential biases. Since the dataset pertains to healthcare, any imputation technique, such as mean, median, or regression-based imputation, might introduce inaccuracies or distort the relationships within the data. Healthcare datasets involve sensitive and high-stakes information, where even minor misrepresentations can lead to misleading conclusions or compromised model performance. Thus, it is safer to exclude these rows to ensure that the subsequent machine learning model predictions remain reliable.A graph of different colored squares

AI-generated content may be incorrect.

Figure 7 – Proportions of Breast Quadrant vs. Recurrence

The heatmap depicted in Figure 7 illustrates the proportions of recurrence subdivided by breast quadrant. Right upper and left lower breast cancer localization have the highest proportion of breast cancer recurrence.

A green squares with white text

AI-generated content may be incorrect.

Figure 8 – Proportions of Degree of Malignancy vs. Recurrence

The heatmap depicted in Figure 8 illustrates the proportions of recurrence, subdivided by degree of malignancy. Patients with 3rd degree malignancy have the highest rate of breast cancer recurrence.

A graph of a comparison of menopause and recurrence proportions

AI-generated content may be incorrect.

Figure 9 – Proportions of Menopause vs. Recurrence

The heatmap depicted in Figure 9 illustrates the proportion of recurrence, subdivided by menopause category. Patients who have not yet experienced menopause, ‘premeno,’ have the highest rate of cancer recurrence.

**Summary of Findings**

The analysis explores the relationship between cancer recurrence and various factors using crosstabs, correlation matrices, and visualizations. Notably, no cases of cancer recurrence were recorded in patients who experienced menopause before the age of 40. Recurrence rates were higher in patients with left breast cancer, right upper tumor localization, and those who received radiation as part of their treatment. Correlation analysis revealed weak relationships between variables, with tumor size showing a slight positive correlation with invasive nodes and recurrence likelihood, while invasive nodes exhibited a moderate correlation with recurrence. Additionally, larger tumors were slightly associated with more invasive nodes, particularly in recurrent cases.

The distributions of age and tumor size appeared relatively normal, whereas invasive nodes followed a right-skewed distribution, reflecting the typical trend in healthcare where fewer patients have a high number of invasive nodes. Despite some imbalance in qualitative variables such as 'irradiate' and 'node-caps,' the dataset remains sufficient for implementing machine learning models. Heatmaps further highlighted key trends, showing the highest recurrence rates in patients with right upper and left lower breast quadrants, those with 3rd-degree malignancy, and premenopausal individuals.

The dataset presents several issues that could impact a prediction system. Formatting inconsistencies include extra apostrophes in categorical values and hyphenated ranges in columns like `age`, `tumor-size`, and `inv-nodes`, which required transformation, through comprehensive functions, into numerical values. Missing data exists in `node-caps` and `breast-quad`, which was handled by removing these rows as to not impact the accuracy of predicting health trends. All columns are currently read as strings, though some were converted to numerical format. Categorical variables needed proper encoding—binary features like ‘class’ were mapped to 0/1. Additionally, potential class imbalance in the `class` column could bias the model, necessitating techniques like synthetic minority oversampling techniques (SMOTE) or under sampling. Addressing these issues will improve data quality and enhance the predictive accuracy of machine learning models.

1. Lichman, M.: UCI machine learning repository, University of California, School of Information and Computer

   Science, Irvine, CA (2019). [http://archive.ics.uci.edu/ml/datasets/breast+cancer](http://archive.ics.uci.edu/ml/datasets/breast%2bcancer)  [↑](#footnote-ref-1)