Breast Cancer Survival Analysis

Xinyu Li, Danwen Li, Huixin Yang

Introduction and Background

Breast cancer, a disease in which abnormal breast cells grow out of control and form tumors, was first noted by ancient Egyptians around 3,500 years ago. Since then, “it has been mentioned in almost every period of recorded history” (Mandal, 2019). Today, breast cancer ranks as the most common type of cancer among women in the United States, except for skin cancer. Despite continuing advances in medical science and healthcare treatments, breast cancer still “remains the second leading cause of cancer death among women overall” (CDC, 2023).

Objective and Goals

Our team is composed entirely of females, and within our families, some individuals have either experienced or are currently facing the challenges of breast nodules, mastitis, and other breast diseases. Due to our identities and our family backgrounds, we all share a profound curiosity and concern for breast cancer. We are acutely aware of the significance of breast cancer in women’s health and realize its enduring presence as a critical health concern, so our group’s topic of interest is centered around the survival analysis of breast cancer. We aim to uncover the relationship between breast cancer stages, various treatment methods, and the survival rates of patients at different follow-up intervals.

Dataset

To explore this topic, we sought a comprehensive dataset and identified one on Kaggle, generously contributed by Queen’s University Belfast Cancer Research. Their research aims to refine early detection methods and determine the best treatment plans for multiple subtypes of breast cancer (*Patrick G Johnston Centre for Cancer Research*, 2021). This dataset focuses on a group of breast cancer patients who have undergone tumor removal surgeries. It is structured in the CSV format, encompassing detailed clinical profiles of 334 patients within 16 columns. While each row represents the clinical information of a patient, each column signifies a specific variable. With the aim to model and predict the Patient_Status, which is a binary variable indicating "Alive" or "Dead", all other features like age, gender, protein markers, tumor stage, histology, hormone and receptor statuses, type and date of surgery, and last visit date, serve as independent variables for the analysis. This wealth of information paves the way for constructing predictive models that could significantly impact patient status prediction by forecasting potential outcomes based on the myriad of variables provided.

Project Plan

To better understand the dataset in depth, we will perform exploratory data analysis through 8 visualizations on different variables.

1. Heatmap: We will employ a heatmap to investigate the correlation between each pair of numerical variables. This visualization allows us to see if there's a multicollinearity issue in our dataset.

2. Histogram: We will use histogram to evaluate the age distribution of all patients in the dataset. Then, we can identify if ages are evenly distributed or skewed towards a certain age group. Meanwhile, we can check if there are outliers in the dataset (ensuring ages fall within a reasonable range).

3. Pie chart: We will use a pie chart to see the distribution of histology. This allows us to evaluate the percentage of each Carcinoma classification among all patients. In this way, we can quickly compare which type of Carcinoma is most common among the patients through a quantitative measure.

4. Scatter plot: Through a scatter plot, we can analyze the relationship between two variables, intuitively understand the distribution of data, examine if there is any clustering, and identify potential outliers. For example, we can use a scatter plot to analyze attributes like protein1-4 to investigate their correlations.

5. Box plot: A box plot can help us analyze the distribution of data, observe the maximum, minimum, median, and the 25th and 75th percentile values of the data. For example, we can use a box plot to analyze the differences in the age distribution of patients among different types of cancers (Histology category), allowing us to understand information such as the average age of onset for different types of breast cancer.

6. Bar chart: To understand the distribution of gender, we propose using a bar chart. This visualization tool will provide a clear and concise representation of the number of individuals or the percentage distribution for each gender category.

7. Table: To begin our exploration of the dataset, we are utilizing a table format to present descriptive statistics for variables. Tables provide a comprehensive snapshot, revealing measures of central tendency, spread, and the range for each variable. Specifically, the table lists count, mean, standard deviation (std), minimum (min), 25th percentile, median (50%), 75th percentile, and maximum (max) values.tage distribution for each gender category.

8. Stacked bar chart: For a comprehensive understanding of the biomarker distribution in our dataset, I propose using a stacked bar chart. This will effectively visualize the concurrent representation of the three breast cancer biomarkers: ER status, PR status, and HER2 status.

After performing the above visualizations, we would delve into predictive analysis using 3 methods to predict Patient_Status.

1.SMOTE: By observing the 'Patient_Status' categories in the dataset, it becomes apparent that this is an imbalanced dataset, with 79% of the data belonging to the 'Alive' class and only 21% belonging to the 'Dead' class. This can lead to a situation where classification models tend to predict 'Alive' more often.

To address this issue, we can employ the SMOTE (Synthetic Minority Over-sampling Technique) oversampling method to augment the minority class samples. This process balances the data by increasing the number of 'Dead' class samples, ensuring an equal representation of data between the 'Dead' and 'Alive' classes. This helps mitigate the bias in classification model predictions.

2. Logistic Regression: To forecast the 'Patient_Status', we aim to employ logistic regression, a statistical method designed for predicting binary outcomes. Logistic regression is suitable for this project because our target variable, 'Patient_Status', is binary. The model will calculate the odds of a patient being 'Alive' or Dead' based on the input variables. By analyzing the weights and significance of each variable, we can understand which factors are most influential in predicting patient outcomes. By utilizing logistic regression, we aim to provide clinicians with a tool to assess the risk and prognosis for patients based on several key factors, helping them make more informed decisions. Furthermore, understanding the contribution of each variable can lead to insights into potential areas of intervention or tailored patient care. This model, once validated, can be an asset in predicting patient outcomes and aiding in personalized treatment planning.

3. Random Forest: We can use more complex classification models than logistic regression to train and predict on the dataset. This allows us to compare the predictive performance of different classification models on the dataset and choose the one that performs better as the final model to use. We will use RandomForestClassifier from the sklearn package in python. We will split the dataset with 70% for training data and 30% for testing data. Then, we will use ROC to evaluate the accuracy of our model. Using random forest, we can rank what variables are more important in classifying Patient_Status as it will generate a score to each feature.

Data Dictionary:

Patient_ID: unique identifier id of a patient

Age: age at diagnosis (Years)

Gender: Male/Female

Protein1, Protein2, Protein3, Protein4: expression levels (undefined units)

Tumour_Stage: I, II, III

Histology: Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, Mucinous Carcinoma

ER status: Positive/Negative

PR status: Positive/Negative

HER2 status: Positive/Negative

Surgery_type: Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other

Date_of_Surgery: Date on which surgery was performed (in DD-MON-YY)

Date_of_Last_Visit: Date of last visit (in DD-MON-YY) [can be null, in case the patient didn't visited again after the surgery]

Patient_Status: Alive/Dead [can be null, in case the patient didn't visited again after the surgery and there is no information available whether the patient is alive or dead].

References

1. Am. (2021, August 5). *Real breast cancer data*. Kaggle. https://www.kaggle.com/datasets/amandam1/breastcancerdataset/data
2. Centers for Disease Control and Prevention. (2023, July 25). *Basic information about breast cancer*. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/index.htm
3. Mandal, Dr. A. (2019, February 26). *History of breast cancer*. News. https://www.news-medical.net/health/History-of-Breast-Cancer.aspx
4. *Patrick G Johnston Centre for Cancer Research*. Breast Cancer | The Patrick G Johnston Centre for Cancer Research | Queen's University Belfast. (2021, August 9). https://www.qub.ac.uk/research-centres/cancer-research/OurResearch/BreastCancer/
5. Guest_Blog. (2023, April 26). *10 Techniques to solve Imbalanced Classes in Machine Learning (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/
6. Keita, Z. (2022, September 21). *Classification in Machine Learning: An Introduction*. https://www.datacamp.com/blog/classification-machine-learning
7. R, S. E. (2023, October 14). *Understand random forest algorithms with examples (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/