Survival Analysis of Breast Cancer Patients: Unveiling Prognostic Factors and Treatment Outcomes

Project Members : Ajinkya Phanse Sajeev Singh Heamesh Choudary

Abstract :

In this project, we delved into a dataset containing 2509 entries of breast cancer patients, aiming to uncover insights into their clinical and molecular characteristics for predicting survival outcomes. Through extensive exploratory data analysis, we meticulously examined key variables like age at diagnosis, overall survival, mutation count, and lymph node status. Visualizations were instrumental in showcasing age's distribution, revealing a prominent peak around the mid-50s, while the variability in overall survival hinted at diverse prognoses. Notably, the skewed distribution of mutation counts shed light on the spectrum of mutation burdens among patients. Additionally, we utilized boxplots to emphasize the variable presence of cancer in lymphatic tissue through positive lymph nodes. Addressing missing data was crucial, and we employed effective imputation techniques to ensure the robustness of model training. By encoding categorical variables and preprocessing the dataset, we laid the groundwork for predictive modeling using a Gradient Boosting Survival Analysis approach. Ultimately, the model generated survival probabilities for breast cancer patients, offering potential insights into personalized treatment strategies and prognosis assessment.

**Introduction :**

Breast cancer remains a significant public health concern worldwide, necessitating a comprehensive understanding of its various clinical and molecular aspects to improve patient outcomes. In this project, we delve into the analysis of a dataset obtained from the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) study, which contains information on 2509 patients diagnosed with breast cancer.

The dataset encompasses diverse variables, including demographic factors like age at diagnosis, clinical features such as tumor size and stage, treatment modalities like chemotherapy and hormone therapy, as well as molecular characteristics like hormone receptor status and genetic mutations. However, the dataset is not without its challenges, as it contains missing values that require careful handling to ensure the integrity of the analyses.

To address these challenges, we embark on a journey of data exploration, preprocessing, and analysis. We begin by examining the distribution of key variables, such as age at diagnosis, overall survival, and mutation count, through visualizations like histograms and box plots. These exploratory analyses provide insights into the demographic and clinical characteristics of the patient cohort and lay the groundwork for subsequent investigations.

Furthermore, we implement strategies to handle missing data, including imputation techniques tailored to the type of variable and its clinical relevance. By imputing missing values and standardizing features, we ensure the dataset's suitability for subsequent modeling and analysis.

Building upon this foundation, we employ survival analysis techniques to investigate key endpoints such as overall survival and relapse-free survival. Utilizing Kaplan-Meier estimators and advanced machine learning models like Gradient Boosting Survival Analysis, we aim to uncover prognostic factors and develop predictive models for patient outcomes.

Through this project, we aspire to contribute to the ongoing efforts in breast cancer research by leveraging data-driven approaches to gain insights into disease mechanisms, identify predictive biomarkers, and ultimately, improve clinical decision-making and patient care.

Materials and Methods :

Data Collection and Preprocessing:

* The dataset used for this project was sourced from the METABRIC study, consisting of 2509 entries with 34 features.
* Python's pandas library was utilized to load the dataset from a CSV file.
* Initial data exploration was conducted using methods like info() and describe() to gain insights into the dataset's structure, missing values, and statistical summary.
* Missing values were handled using appropriate techniques, such as imputation with mean/mode values or specific strategies based on column characteristics.

Exploratory Data Analysis (EDA):

* Matplotlib and Seaborn libraries were employed for data visualization.
* Histograms and boxplots were generated to visualize the distributions and key statistics of variables like age at diagnosis, overall survival, mutation count, and positive lymph nodes.
* Insights were drawn from the EDA, highlighting key observations and clinical relevance of the analyzed variables.

Data Encoding and Feature Transformation:

* Categorical variables were encoded using Label Encoder from scikit-learn to convert them into numerical format suitable for model training.
* Target variables, including 'Overall Survival Status' and 'Relapse Free Status', were transformed into the required format for survival analysis.

Right Censoring of Survival Data:

* Right censoring, a common occurrence in survival analysis, was illustrated using a graphical representation.
* Patients with different outcomes, such as those experiencing events within the study period, surviving beyond the study end, or withdrawing from the study, were depicted.

Survival Analysis - Kaplan-Meier Estimators:

* Survival analysis was performed using Kaplan-Meier estimators to analyze the probability of survival over time.
* Separate survival functions were plotted for 'Overall Survival Status' and 'Relapse Free Status' to visualize survival probabilities over the study period.
* Insights were drawn from the survival functions, highlighting differences in survival probabilities based on the analyzed variables.

Predictive Modeling - Gradient Boosting Survival Analysis:

* Gradient Boosting Survival Analysis, implemented using the scikit-survival library, was employed for predictive modeling.
* A pipeline was constructed to preprocess the data, including imputation and standardization of features.
* The preprocessed data was split into training and testing sets for model training and evaluation.
* The trained model was used to predict survival functions for new data, and the results were visualized to demonstrate the predicted survival probabilities over time.

Conclusion:

* The materials and methods presented in this project encompassed data collection, preprocessing, exploratory analysis, survival analysis, and predictive modeling.
* Through these methodologies, insights were gained into breast cancer patient data, survival probabilities, and predictive modeling for patient outcomes.

Results :

In our analysis of the breast cancer dataset, we observed several significant findings. Firstly, upon preprocessing the data, we found that most patients were diagnosed with breast cancer around their mid-50s, highlighting the prevalence of the disease in middle-aged individuals. We also noted a wide distribution of overall survival months, indicating considerable variability in patient outcomes post-diagnosis, with some patients surviving for extended periods. Furthermore, the distribution of mutation counts was right-skewed, suggesting that while most patients had a low number of mutations, a few exhibited a high mutation burden. Notably, the count of living patients outnumbered deceased patients, underscoring the success of treatments and interventions in improving survival rates. In survival analysis, Kaplan-Meier estimators illustrated survival functions for overall survival and relapse-free status, offering insights into the probability of survival over time for individual patients. These findings collectively suggest the heterogeneity of breast cancer outcomes and underscore the importance of personalized treatment strategies tailored to each patient's unique characteristics and disease profile.

DISCUSSION :

The findings from our analysis provide valuable insights into the characteristics and outcomes of breast cancer patients, shedding light on the complexities of the disease and its management.

Firstly, the preprocessing stage revealed important demographic and clinical trends among the patient cohort. The fact that the majority of patients were diagnosed in their mid-50s suggests a common age range for breast cancer onset in this population. Additionally, the wide distribution of overall survival months highlights the variability in patient outcomes, indicating the need for personalized treatment approaches tailored to individual needs.

Examining patient outcomes, we observed that the count of living patients exceeded that of deceased patients. This is an encouraging indication of the effectiveness of current treatments and interventions in prolonging survival and improving quality of life for breast cancer patients.

Survival analysis using Kaplan-Meier estimators allowed us to visualize survival functions for overall survival and relapse-free status. These analyses underscored the heterogeneity of breast cancer outcomes, with distinct survival trajectories observed among patients. Such variability underscores the importance of considering multiple factors, including tumor characteristics, treatment modalities, and patient demographics, in predicting prognosis and guiding treatment decisions.

Overall, these results emphasize the need for personalized and multidisciplinary approaches to breast cancer management. Tailoring treatment strategies to individual patient characteristics and disease profiles is crucial for optimizing outcomes and enhancing the quality of life for breast cancer patients. Furthermore, ongoing research efforts aimed at unraveling the molecular mechanisms underlying breast cancer heterogeneity hold promise for the development of more targeted and effective therapeutic interventions in the future.

LITERATURE CITATIONS AND LEARNING :

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394-424.
2. DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., ... & Jemal, A. (2019). Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(6), 438-451.
3. Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., ... & Cronin, K. A. (Eds.). (2020). SEER Cancer Statistics Review, 1975-2017. National Cancer Institute. Retrieved from https://seer.cancer.gov/csr/1975\_2017/
4. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481.
5. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., ... & Smith, P. G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer*, 35(1), 1-39.
6. Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175-2197.
7. Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer.
8. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.

WHAT WE LEARNT :

1. Global Cancer Statistics: Bray et al. (2018) provided estimates of breast cancer incidence and mortality worldwide, highlighting the global burden of the disease. This information is crucial for understanding the broader context of breast cancer survival rates and patterns.
2. Breast Cancer Statistics in the US: DeSantis et al. (2019) offered detailed statistics on breast cancer in the United States, including incidence, mortality, and survival rates. This data provides valuable insights into the specific trends and disparities within the US population.
3. SEER Cancer Statistics Review: The SEER Cancer Statistics Review (Howlader et al., 2020) provides comprehensive data on cancer incidence, survival, and prevalence based on the Surveillance, Epidemiology, and End Results (SEER) program. This dataset is a fundamental resource for studying cancer outcomes and understanding long-term survival trends.
4. Kaplan-Meier Estimation: Kaplan and Meier (1958) introduced the nonparametric Kaplan-Meier estimator, which is widely used for estimating survival probabilities from incomplete observational data. This method is essential for analyzing censored survival data, such as time-to-event outcomes in breast cancer studies.
5. Design and Analysis of Clinical Trials: Peto et al. (1977) discussed the design and analysis of randomized clinical trials, emphasizing the importance of prolonged observation and appropriate statistical methods for evaluating treatment effects in cancer research. Understanding trial design principles is crucial for interpreting the effectiveness of interventions aimed at improving breast cancer survival.
6. Flexible Parametric Survival Models: Royston and Parmar (2002) introduced flexible parametric survival models, which offer advantages over traditional Cox proportional hazards models by allowing for more flexible modeling of the baseline hazard function and covariate effects. These models provide a powerful framework for prognostic modeling and estimating treatment effects in breast cancer survival analysis.
7. Modeling Survival Data: Therneau and Grambsch (2000) provided a comprehensive guide to modeling survival data, including the extension of the Cox proportional hazards model and techniques for assessing model assumptions and fit. Understanding these modeling techniques is essential for developing robust predictive models and analyzing survival data in breast cancer research.
8. Multiple Imputation for Missing Data: van Buuren and Groothuis-Oudshoorn (2011) introduced the concept of multiple imputation by chained equations (MICE) for handling missing data in survival analysis. This approach allows for more robust and efficient analysis by imputing missing values based on observed data patterns. Handling missing data appropriately is critical for ensuring the validity and reliability of survival analyses in breast cancer research.

APPENDICES :

Sajeev :

1. Data Loading and Initial Analysis:
   * Load the breast cancer dataset from the CSV file.
   * Perform an initial analysis of the dataset including checking the data types, missing values, and basic statistics.
2. Data Visualization:
   * Create visualizations to understand the distribution of variables such as age at diagnosis, overall survival, mutation count, and positive lymph nodes.
3. Handling Missing Values:
   * Implement a function to handle missing values by imputing them appropriately for different columns based on their types and relationships with other variables.

Ajinkya:

1. Label Encoding and Data Preparation:
   * Encode categorical variables using Label Encoder.
   * Prepare the dataset for model training by splitting it into features (X) and target variables (y).
   * Handle categorical variables and missing values using appropriate preprocessing techniques.
2. Survival Analysis Modeling:
   * Utilize the lifelines library to perform survival analysis.
   * Implement the Kaplan-Meier estimator to visualize survival functions for overall survival and relapse-free status.
3. Survival Analysis Prediction:
   * Train a survival analysis model (e.g., Gradient Boosting Survival Analysis) using the prepared dataset.
   * Predict the survival function for a sample patient and visualize it.

Heamesh:

1. Data Preprocessing Pipeline:
   * Implement a preprocessing pipeline using scikit-learn's Pipeline and Column Transformer to handle missing values and scale/encode features.
2. Preprocessed Data Inspection:
   * Verify the preprocessing steps by examining the preprocessed dataset.
   * Ensure that missing values are imputed, and features are scaled or encoded correctly.
3. Final Integration and Presentation:
   * Integrate the work from others into the final script.
   * Review the entire codebase for consistency and completeness.
   * Prepare the final presentation or report summarizing the work done, key insights, and visualizations.