**Project 2: Breast Cancer Survival Prediction**

**Abstract**

This project endeavors to leverage a dataset comprising information about breast cancer patients from a prospective study, with the primary objective of constructing predictive models for estimating the survival months and survival status of these patients. In the course of our exploratory data analysis, noteworthy insights emerged: the distribution of predictive variables and the interrelationships among them suggest a departure from the assumptions inherent in linear models. Specifically, we observed a tendency towards interaction, as indicated by (). In response to these findings, we opted to develop two nonlinear models to enhance predictive accuracy—a survival analysis model for survival months and a logistic model for survival status.In summary, both models exhibited satisfactory predictive power, underscoring their efficacy in forecasting outcomes. Furthermore, the results provide valuable indications that () significantly influences the risk associated with breast cancer survival.

**Introduction**

In this breast cancer survival prediction project, our focus is on utilizing a dataset from a prospective study, encompassing variables such as age, tumor characteristics, and patient demographics. Our objective is to predict the risk of death based on these features and identify significant factors influencing survival. With an emphasis on model performance and fairness evaluation, we aim to address potential disparities between racial groups. The report will provide a comprehensive analysis, including data exploration, regression model considerations, and model diagnostics.

**Methods**

1. **Exploratory Data Analysis (EDA)**

For our EDA, we commenced by providing fundamental numeric descriptive statistics for the original data. Visual representations, such as density plots and histograms, were employed to illustrate variable distributions. The identification of outliers was conducted through boxplots. Additionally, the interrelationships between variables were elucidated via a correlation plot, offering a comprehensive overview of the dataset's key characteristics.

1. **Survival Analysis**
2. **Logistic Model**

***Predictors Examination***

We classified our predictors into categorical and continuous variables. For categorical ones, we assessed their distribution through histograms, conducted chi-square tests, and created correlation plots. Continuous variables were examined using density plots, rank-sum tests, and t-tests for median comparisons between status-stratified groups. Our findings led to strategic decisions: 1) exclusion of the 4th level of *grade* due to anaplastic mode, 2) removal of *differentiate*, *x6th\_stage*, and *estrogen\_status* to address multicollinearity, 3) elimination of *survival\_months*, and 4) retention of *regional\_node\_examined* for subsequent Walt test.

Furthermore, we explored interaction terms exhaustively, revealing two significant pairs: *a\_stage* with *regional\_node\_positive*, and *reginol\_node\_examined* with *regional\_node\_positive*. These interactions were incorporated into our models.

***Model Variables Selection and Diagnosis***

During this stage, we employed a comprehensive approach for variable selection in our logistic models. Utilizing the Stepwise Method, Random Forest, and LASSO, we gauged the importance of variables. The models generated through these methods underwent scrutiny via Nomogram, Calibration Curve, and Hosmer and Lemeshow Goodness-of-fit Test for comparison. The ultimate variables chosen for our logistic model were determined based on accuracy and kappa value, validated through a 5-fold cross-validation process.

**Result**

1. **EDA**
2. **Survival Analysis**
3. **Logistic Model**

In evaluating the performance of the models derived from different methods, the Stepwise Method demonstrated superior results with an accuracy of 85.7%, kappa of 18.7%, nomogram score of 7.6%, and H-L value of 69.4%. The model revealed insightful associations: 1) a one-year increase in age corresponds to a 0.02186205-fold rise in the risk of death; 2) for Black individuals, the probability of death is 0.58 times higher than that for White individuals; 3) in the interaction term, a unit increase in regional node positive, given distant and regional a\_stages, results in a 0.6762288-fold higher probability of death. Model efficacy was further assessed through Residuals vs. Fitted Plot, QQ plot, Scale-Location Plot, and Cook’s Distance.

For validation and real-world applicability, the model was executed with race-stratified data. Prediction accuracy was 86.3% for White individuals and 83.6% for other races, indicating a slight difference of 3%, deemed acceptable. To refine accuracy, a suggested approach involves separate modeling for each racial group, offering a tailored solution to enhance predictive performance.

**Conclusion**

**Contribution**