SEAN XANDER B. AQUINO BSCS 2A **PROBLEM SET 3**

The analysis utilizes the Breast Cancer Dataset which exists within the UCI Machine Learning Repository. The database holds clinical and pathological data about breast cancer patients to forecast recurrence events. The breast cancer data includes both categorical along with numerical features which require encoding for categorical values and numeric value scaling before analysis.

The data contains the listed characteristics:

- **Age** presented as categorical groups which include age divisions matching 30-39, 40-49 etc.
- **Menopause Status** (Categorical: premeno, ge40, lt40)
- **Tumor Size** (Categorical: ranges from 0-4 mm to 50-54 mm)
- **Lymph Node** presents information in the form of lymph node involvement numbers.
- **Node Caps** (Binary: Yes/No)
- **Breast** (Binary: left/right)
- **Breast Quadrant** consists of left-up, left-low, right-up categories and similar groups.
- **Irradiation Treatment** (Binary: Yes/No)
- **Class Label** (Target Variable): Recurrence-events (1) or No-recurrence-events (0)

The categorical variables in our dataset received Label Encoding treatment for converting each category into numerical values. Standardization through *StandardScaler* standardized numerical values to obtain better results when training the model.

Logistic Regression Performance

The first logistic regression model received its training data after preprocessing. The test set accuracy of the initial model measured 70.69% which indicates a comparatively poor predictive effectiveness. Model generalization alongside overfitting prevention was achieved by implementing L1 (Lasso) and L2 (Ridge) regularization methods against each other for performance evaluation.

The analysis of regularization required training two separate logistic regression models.

L1 Regularization (Lasso) works through coefficient penalization that uses absolute coefficient value magnitude to choose features by setting selected coefficients at zero.

L2 Regularization (Ridge) applies penalty to squared coefficient magnitudes to produce smaller non-zero values which produce better generalization outcomes.

Model Accuracy

- Original Model - 68.96%
- L1 Regularization - 67.24%
- L2 Regularization - 68.96%