PREDICTING BREAST CANCER RECURRENCE

A DATA-DRIVEN APPROACH

CECI NGUYEN, DCN558
FEB. 25TH, 2025
STUBBS, COE379L

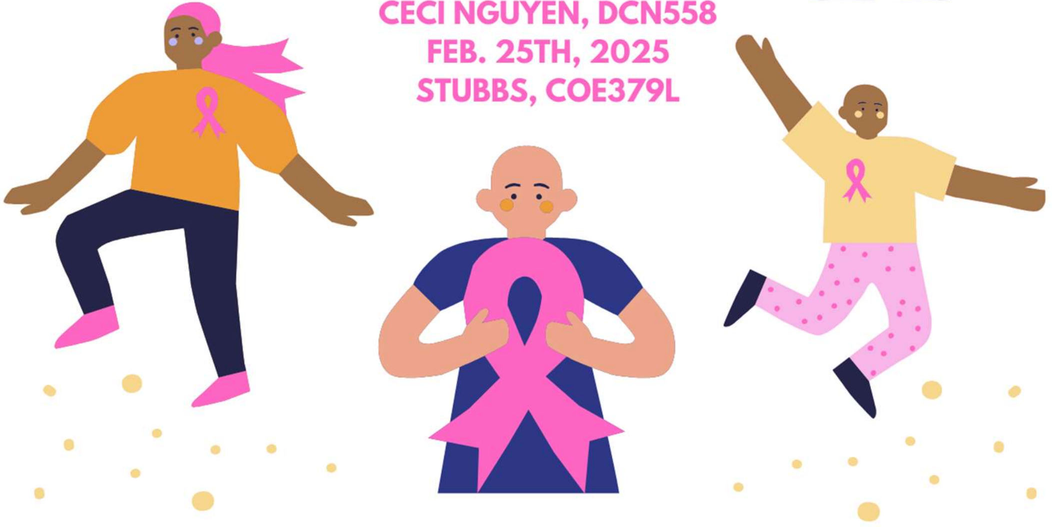

Table of Contents

Introduction

Breast cancer recurrence prediction is crucial for improving patient outcomes and guiding treatment decisions. In this study, I analyzed a dataset containing breast cancer patient records to develop a classification model that predicts recurrence. The project involved data preparation, exploratory data analysis (EDA), and the implementation of machine learning models to assess their predictive capabilities.

Data Preparation

To prepare the dataset, I performed several key steps to ensure that the data was clean, structures, and suitable for analysis. First, I checked for any missing values and confirmed that all columns contained appropriate data types. For example, I found by just calling ***bc_data.info()*** that there was two null values in the data set.

I then checked for and dropped duplicates, remedied the null values in the ***tumor-size*** and ***inv-nodes***, and treated any invalid data values. After all the data was clean, I converted all the column's data types to be categorical except for ***deg-malig*** since it was already in a proper data type. First, I converted the ***age, tumor-size, and inv-nodes*** columns into ordinal categories by mapping their ranges. Next, I utilized the ***groupby*** function for the ***menopause*** column since the column is dependent on the age column. Then, I converted the ***class, node-caps, breast, and irradiat*** columns by utilizing binary one-hot encoding. Lastly, I utilized one-hot encoding for the ***breast-quad*** columns for the five valid locations.

Before any model training occurred, I also used summary statistics to understand the distributions of the data frame. I created histograms to visualize the distribution of key variables like age and tumor size and generated heatmaps to analyze correlations between variables.

Insights from Data Preparation

Through EDA, I found that breast cancer cases were most frequent in the 40-49 and 50-59 age groups, with a significant number of tumors located in the left lower quadrant of the breast. The malignancy levels were skewed towards higher values, indicating a prevalence of severe cases. Additionally, the dataset showed class imbalance, with more non-recurrence cases than recurrence cases, which could affect model performance.

Model Training Procedure

To build a classification model for predicting recurrence, I followed these steps:

1. The dataset was divided into training (80%) and testing (20%) subsets while maintaining class distribution using stratified sampling.
2. I experimented with two classification models:
    a. K-Nearest Neighbors (KNN)
    b. Logistic Regression
3. For KNN, I used GridSearchCV to determine the optimal number of neighbors.
4. Both models were trained using the training dataset and then evaluated on the test set.

Model Performance

The model results indicated that the default KNN model (K=5) had an accuracy of 59%, while the optimized KNN model (K=16) and Logistic Regression both achieved 65% accuracy. However, accuracy alone did not reflect the model's effectiveness in predicting recurrence cases. The KNN model with the best K value performed well for non-recurrence cases but completely failed to detect recurrence cases. Logistic Regression provided a slightly better balance, correctly identifying 21% of recurrence cases, though still struggling with overall recall. These results highlight the challenge of class imbalance and the need for further model improvements.

Model Confidence

Given the observed class imbalance and the performance results, I have moderate confidence in the model's ability to predict non-recurrence cases but low confidence in its ability to detect recurrence cases accurately. The poor recall for recurrence cases suggests that additional techniques, such as resampling methods (oversampling minority class or undersampling the majority class), feature engineering, or trying different model architectures (e.g., decision trees or ensemble methods), could improve performance.

Overall, while the current model provides a baseline understanding, further refinement is necessary to make it a reliable predictor of breast cancer recurrence.