# Beyond the Cure: Modeling the Risk of Breast Cancer Recurrence

29 April 2025

Brandon Thomas

bthomas@bellarmine.edu

## Executive Summary

This project aimed to develop predictive models to forecast breast cancer recurrence by leveraging clinical and demographic patient data. Using machine learning techniques, including Random Forest, Support Vector Machines (SVM), and XGBoost, a breast cancer dataset containing attributes such as tumor size, lymph node status, and age was analyzed. Through rigorous exploratory data analysis, preprocessing, and model development, several important patterns emerged. Notably, higher recurrence rates were associated with factors such as increased numbers of invasive lymph nodes and specific tumor localization within the breast.

Despite challenges related to missing data and class imbalance, the predictive models demonstrated strong potential for supporting personalized clinical decision-making. By identifying patients at higher risk for recurrence early, healthcare providers could proactively adjust treatment strategies, such as recommending more aggressive adjuvant therapies like chemotherapy, implementing closer post-treatment surveillance protocols, or tailoring the frequency and type of follow-up imaging studies. For instance, a patient flagged by the model as high-risk based on lymph node involvement and tumor characteristics could be offered more intensive monitoring to detect recurrence at an earlier, more treatable stage. The interpretability of models like Random Forest and XGBoost, allows healthcare providers to not only trust the model's outputs but also to understand the key factors driving each patient's risk assessment, thus facilitating shared decision-making with patients.

Furthermore, the project results suggest opportunities for stratifying patients into risk tiers, enabling a more resource-efficient allocation of healthcare services. This could lead to the development of clinical decision support tools that dynamically update patient risk profiles during survivorship care of first-time cancer patients. Future work will involve refining model performance through advanced hyperparameter tuning, integrating genomic data such as BRCA1 and BRCA2 mutation status to capture genetic predisposition to recurrence, and validating model predictions against real-world patient outcomes. Incorporating genomic features would further personalize risk predictions and potentially uncover novel interactions between genetic and clinical factors. Ultimately, this project not only enhanced my technical skills in data preprocessing, feature engineering, and model evaluation, but also established a strong foundation for translating predictive modeling into actionable tools that could meaningfully improve breast cancer management and patient outcomes.

## Introduction

Breast cancer remains one of the most prevalent and impactful malignancies worldwide, representing a major cause of morbidity and mortality among women. Despite advances in early detection and treatment, recurrence—whether local, regional, or distant—continues to pose a significant threat to long-term patient survival and quality of life. Recurrence not only affects individual outcomes but also places a considerable burden on healthcare systems, necessitating prolonged monitoring, additional treatments, and complex clinical decision-making. Accurately predicting a patient's risk of recurrence is therefore a critical component of effective oncology care. Reliable risk assessments allow for the development of personalized treatment strategies, ensuring that patients receive interventions that are appropriately aggressive while minimizing unnecessary treatments that may carry harmful side effects. Additionally, accurate recurrence predictions enable the creation of individualized care plans, optimizing the timing and intensity of follow-up screenings and monitoring.

The aim of this project was to develop a consistent, robust, and interpretable predictive model capable of estimating breast cancer recurrence risk using clinical and demographic variables readily available through routine healthcare examinations. These features, including tumor size, lymph node status, menopausal status, and patient age, represent standard data collected during the diagnostic and treatment phases, making the model highly applicable to real-world clinical workflows. By integrating advanced machine learning techniques such as Random Forests, Support Vector Machines, and XGBoost with traditional clinical indicators, the project sought to bridge the gap between complex data-driven insights and everyday clinical practice. The ultimate goal was to contribute to the development of more accurate, patient-specific, and explainable predictive systems that clinicians could adopt to enhance decision-making processes, support early interventions, and improve overall patient outcomes in breast cancer care.

## Project Summary

The project proposal focused on developing a predictive model using a publicly available dataset from the UCI Machine Learning Repository, containing 286 patient cases with nine clinical attributes. These attributes included factors such as tumor size, number of invasive nodes, age, menopausal status, and treatment history. The target variable distinguished between patients who experienced recurrence and those who did not.

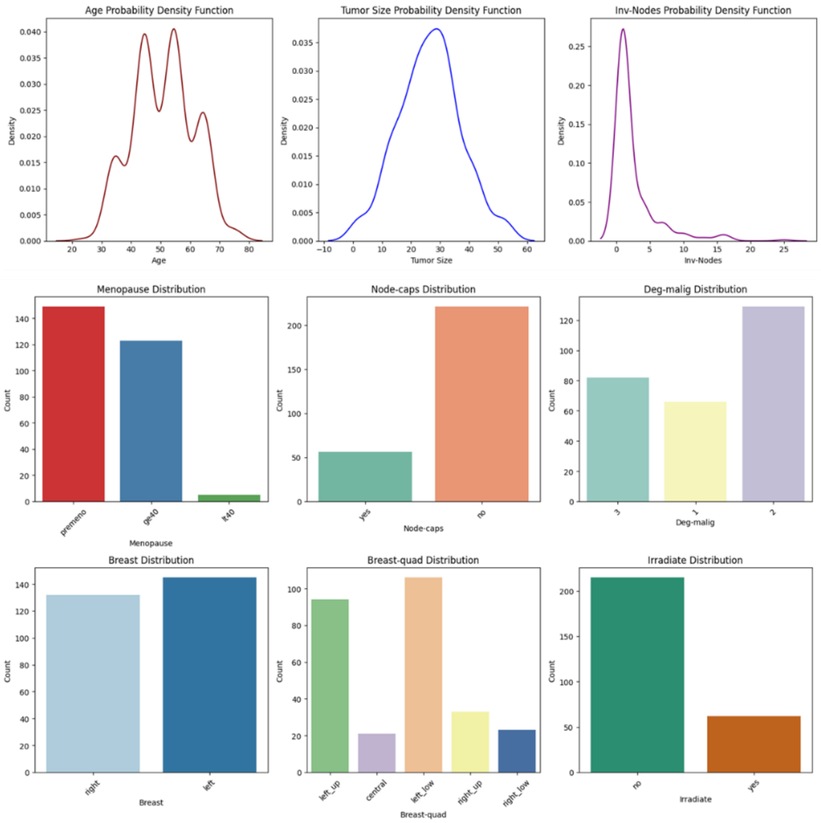
The exploratory data analysis began with a comprehensive examination of the dataset to ensure data integrity and to understand the underlying structure of the variables. Early in the analysis, missing data was identified, particularly within the ‘Node-Caps’ and ‘Breast-Quad’ variables. In the healthcare context, where decision-making can directly impact patient outcomes, the introduction of artificially imputed values could distort important relationships within the data. Therefore, to preserve the authenticity and reliability of the dataset, nine records containing missing values were excluded rather than subjected to imputation techniques such as mean substitution or regression-based filling to maintain accuracy.

Following data cleaning, each variable was carefully categorized based on its data type. Nominal variables such as 'Breast' and 'Menopause' were distinguished from ordinal variables like 'Tumor Size' and 'Invasive Nodes.' Appropriate preprocessing steps were taken to encode categorical features, enabling the data to be compatible with machine learning algorithms that require numerical inputs. Ordinal variables maintained their natural ordering to preserve inherent relationships, while nominal variables were one-hot encoded to avoid introducing artificial ordinal assumptions.

Descriptive statistical analysis and a variety of visualizations were used to uncover patterns and trends within the dataset. Crosstab analyses provided initial insights into associations between predictor variables and the recurrence outcome. One significant observation was that patients who experienced menopause before the age of 40 (‘lt40’ category) exhibited no recorded instances of cancer recurrence, suggesting a potentially protective effect of early menopause against recurrence. Furthermore, recurrence rates were found to be notably higher among patients with tumors located in the right-upper quadrant of the breast, compared to other quadrants, and among patients with left breast involvement overall. These findings indicated that tumor localization may have clinical relevance in predicting recurrence risk. The distribution of variables from which these preliminary conclusions were drawn is summarized in Figure 1.

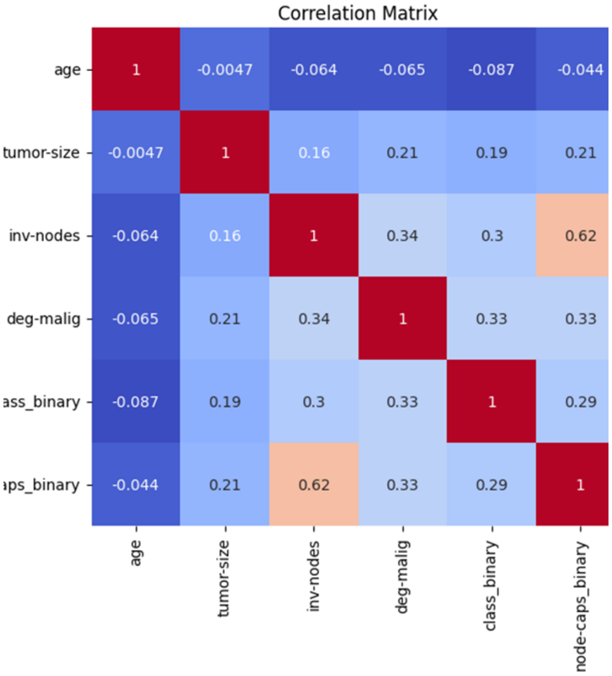
Further, the probability density functions (PDFs) for ‘age’ and ‘tumor size’ revealed relatively normal, unimodal distributions, indicating that these variables were fairly evenly represented across the patient cohort without extreme skewness or clustering (Figure 1). In contrast, the distribution of the 'invasive nodes' variable was heavily right skewed, suggesting that while most patients had few or no invasive nodes, a smaller subset exhibited much higher values. This skewness is clinically expected, as extensive lymph node involvement is less common but strongly prognostic when present.

**Figure 1 – Distribution of Variables**



Correlation matrices were constructed to quantify the strength and direction of relationships between continuous or ordinal variables (Figure 2). Although most correlations were weak, a modest positive correlation was detected between tumor size, number of invasive lymph nodes, and the likelihood of recurrence. Specifically, as tumor size increased, the number of positive invasive nodes also tended to rise, and both were associated with higher recurrence risk.

**Figure 2 – Correlation of Numerical Variables**



Overall, the EDA process not only clarified the structure of the dataset but also uncovered several biologically and clinically meaningful patterns that informed the selection of modeling strategies. Recognizing the importance of variables like tumor size, lymph node involvement, menopausal status, and tumor localization early in the analysis allowed for more targeted feature engineering and improved the interpretability of subsequent machine learning models.

Following data preparation, a range of machine learning models were implemented to predict breast cancer recurrence, with a focus on balancing predictive performance and interpretability. The models selected included Random Forests, Support Vector Machines (SVM), and XGBoost. Each model was chosen for its distinct strengths: Random Forests for their robustness against overfitting and ability to rank feature importance, SVM for their strong performance in high-dimensional spaces and handling of non-linear decision boundaries, and XGBoost for its efficiency and superior handling of complex, non-linear interactions through gradient boosting techniques.

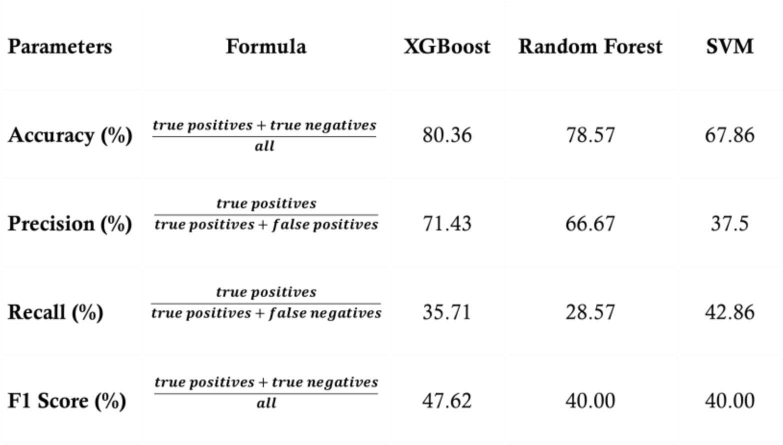
Given the imbalance in the dataset—with a higher proportion of non-recurrence cases compared to recurrence cases—class balancing techniques were critical. The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to address this imbalance. SMOTE generates synthetic examples of the minority class (recurrence cases) by interpolating existing minority instances. This approach helped prevent models from becoming biased toward predicting the majority class and ensured that performance metrics like recall, which measures the model's ability to correctly identify true positives, remained meaningful.

Model evaluation was conducted using multiple metrics to capture different aspects of performance. Accuracy was calculated to measure the overall proportion of correct predictions, while precision and recall provided deeper insights into the model's behavior concerning false positives and false negatives, respectively. The F1-score, the harmonic mean of precision and recall, was particularly emphasized due to its balanced view of model effectiveness in the presence of class imbalance. In addition to these metrics, confusion matrices were generated to visually assess the distribution of true positives, true negatives, false positives, and false negatives across models.

Visualization played a key role throughout the analysis to ensure that findings were clearly communicated and interpretable. Visualizations were created using Matplotlib and Seaborn, allowing for the generation of heatmaps and correlation matrices that highlighted important trends in the data. The combination of careful preprocessing, balanced modeling strategies, evaluation metrics, and comprehensive visualizations ensured that the findings from this project were both statistically relevant and clinically interpretable.

Among the models developed, XGBoost achieved the best overall performance, balancing accuracy and recall. Feature importance analyses consistently highlight tumor size, degree of malignancy, and number of invasive nodes as the most influential predictors. Table 1 highlights how the four parameters of each model compare.

**Table 1 – Machine Learning Evaluator Comparison**



| Parameters | Formula | XGBoost | Random Forest | SVM |
|---|---|---|---|---|
| Accuracy (%) | $\frac{true\ positives + true\ negatives}{all}$ | 80.36 | 78.57 | 67.86 |
| Precision (%) | $\frac{true\ positives}{true\ positives + false\ positives}$ | 71.43 | 66.67 | 37.5 |
| Recall (%) | $\frac{true\ positives}{true\ positives + false\ negatives}$ | 35.71 | 28.57 | 42.86 |
| F1 Score (%) | $\frac{true\ positives + true\ negatives}{all}$ | 47.62 | 40.00 | 40.00 |

## Reflection

This project provided valuable insights into the complexities and challenges associated with predictive healthcare modeling. Several aspects of the project were particularly successful and laid a strong foundation for the overall outcomes. The data cleaning and exploratory data analysis phases were especially effective. Early in the process, careful examination of the dataset led to the identification and appropriate management of missing values and formatting inconsistencies. By addressing these issues thoughtfully—removing records with missing critical information rather than imputing values—the resulting dataset was structured, reliable, and better aligned for proper machine learning analysis. This decision prioritized data integrity, which is vital in clinical modeling where inaccuracies can mislead healthcare decision-making.

Preprocessing, including one-hot encoding of categorical variables and transforming ordinal variables while preserving their intrinsic order, enabled more meaningful analyses. It also ensured compatibility with various machine learning algorithms, ultimately supporting more accurate and interpretable model development. A major success of the project was the use of multiple models—Random Forest, Support Vector Machines (SVM), and XGBoost—to compare performance and extract common model insights. This comparative approach improved my understanding of how different algorithms are equipped to handle data characteristics such as non-linearity, feature interactions, and class imbalance. Notably, the XGBoost model achieved an accuracy score of approximately 80%. In a clinical context, a model capable of accurately predicting 8 out of 10 cases of breast cancer recurrence represents a significant step toward enhancing individualized patient care. It could enable oncologists to adjust treatment plans based on individualized risk profiles, potentially leading to earlier interventions and better long-term outcomes.

However, the project also encountered several notable challenges. The limited size of the dataset constrained model performance, reducing the volume of data available for both training and testing phases. Machine learning models, particularly those like SVMs and XGBoost, generally benefit from larger datasets to avoid overfitting and to generalize better across diverse patient populations. The small sample size not only limited the model's ability to detect subtle patterns but also amplified the risks associated with model overfitting and unstable predictions. Another key challenge was the inherent class imbalance between patients who experienced recurrence and those who did not. This imbalance caused some models to initially bias predictions toward the majority class (no recurrence), sacrificing sensitivity in detecting true recurrence cases. While class balancing techniques like SMOTE helped partially mitigate this bias, predicting the minority class with high sensitivity remained difficult. The ability to correctly identify true positive recurrence cases is crucial in clinical applications where false negatives (failing to predict recurrence) can have serious consequences for patient outcomes.

If another student or researcher were to embark on a similar project, several recommendations would be crucial to improving outcomes. First, it would be highly beneficial to incorporate a larger, more diverse dataset from the outset. Larger datasets enhance model learning, reduce the risk of overfitting, and improve the generalizability of predictions. Integrating additional features such as hormonal status or genetic would also significantly enrich the feature space, capturing complex biological mechanisms that influence recurrence risk but are not evident in standard clinical variables alone. Furthermore, consideration should be given to integrating imaging data from mammograms or MRIs, offering a multimodal approach that could uncover additional, non-obvious risk factors. In addition to expanding the dataset and feature set, investing more time into systematic hyperparameter tuning would yield better optimized models. This could involve the use of cross-validation and automated machine learning techniques to systematically search through parameters and identify the best model configurations. Finally, collaborating early with healthcare professionals would provide critical feedback on clinical relevance, interpretability needs, and deployment feasibility within real-world clinical workflows.

Throughout the course of this project, several key lessons emerged that will inform potential future work in healthcare analytics. Perhaps one of the most important realizations was that clinical data demands a cautious and tailored approach to standard data science workflows. Traditional cleaning techniques, such as mean imputation for missing data, cannot always be applied without risking distortion of biologically meaningful patterns. Each step, from data cleaning to model interpretation, must be guided by a deep understanding of the clinical context and the potential real-world consequences of modeling decisions. Data quality and preprocessing are absolutely critical. Errors or careless imputations can propagate biases, distort feature relationships, and produce models that are inaccurate or even dangerous in a clinical setting.

In healthcare, predictive accuracy alone is insufficient; models must be explainable and transparent so that clinicians can trust and act on their outputs. Visualization played a pivotal role, not only for final presentation of the models but also during exploratory phases. Effective visualizations helped reveal subtle anomalies in the data, clarified variable distributions, and enhanced the understanding of feature relationships. In addition, maintaining clear documentation throughout the project proved to be essential. Consistent commentary on code and results allowed for reproducibility, made troubleshooting more manageable, and streamlined the transition between the various phases of the capstone project.

Ultimately, this project reinforced the importance of balancing technical performance with clinical relevance. Building models that are both high-performing and understandable is essential for transitioning machine learning advances from academic exercises into powerful tools that truly benefit patient care.

## Conclusion/ Future Work

In conclusion, this project successfully demonstrated the significant potential of machine learning techniques in predicting breast cancer recurrence using readily available clinical and demographic data. By applying advanced algorithms such as Random Forest, Support Vector Machines (SVM), and XGBoost, the project was able to uncover meaningful patterns within the data and develop models capable of differentiating between patients at higher and lower risk of recurrence. Despite challenges inherent to the dataset—including its relatively small sample size and the noticeable imbalance between recurrence and non-recurrence cases—the models produced promising results. Random Forest and XGBoost, in particular, identified critical predictive factors such as tumor size, degree of malignancy, and the number of invasive lymph nodes, aligning with established clinical knowledge and further validating the models' practical relevance.

Future work should focus on expanding the breadth and depth of the dataset to enhance model robustness and predictive power. Incorporating additional clinical variables, such as hormone receptor status, more detailed treatment histories, and patient lifestyle factors, would provide a more comprehensive view of the recurrence risk profile. Furthermore, the inclusion of genetic markers, such as BRCA1 and BRCA2 mutation status, could allow the models to capture inherited predispositions to cancer recurrence, significantly improving individual-level risk assessments. Integrating imaging data, such as mammogram or MRI-derived features, would further enrich the model's input space and enable multi-modal predictive modeling that could detect subtle anatomical indicators of recurrence risk missed by clinical variables alone.

In addition to expanding the dataset, more rigorous model validation will be essential to ensure generalizability beyond the initial cohort. This includes validating the models on external datasets collected from different hospitals, geographic regions, and patient demographics. Such external validation would not only strengthen the evidence for the models' utility within breast cancer populations but also open the door for potential extrapolation to predictive modeling in other types of cancer where recurrence is a critical concern.

Collaboration with healthcare professionals will be a crucial step in translating these models into clinical practice. By engaging oncologists, radiologists, and data scientists in a multidisciplinary approach, predictive models can be refined to fit naturally within clinical workflows. Integration into electronic health record (EHR) systems and development of clinician-friendly decision-support tools would enable providers to access real-time, personalized recurrence risk assessments during patient consultations. This would allow for increasingly data-informed treatment planning, improved patient counseling, monitoring strategies, and potentially leading to better patient outcomes.

Ultimately, the work completed in this project lays a strong foundation for advancing predictive modeling in oncology. It demonstrates how thoughtful application of machine learning can supplement clinical judgment with precise, data-driven insights, helping to optimize care planning and contributing to the broader movement toward precision medicine. By continuing to refine these models and bring them closer to clinical implementation, there is significant potential to positively impact the lives of breast cancer patients worldwide.