Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/17/20240652

Breast cancer survival data prediction using machine learning model

Yucheng Zhao

Department of Biostatistics, School of Global Public Health, New York University, 50 West 4th Street, New York, NY, 10012, USA

yz9439@nyu.edu

Abstract. Breast cancer is the most common form of cancer affecting women, exerting a significant impact on individuals, families, and societies globally. With its multifaceted nature, breast cancer research and awareness efforts have gained substantial momentum, leading to transformative breakthroughs in understanding its causes, diagnosis, treatment, and prevention. Survival analysis is a pivotal statistical tool in understanding the dynamic and often complex trajectory of breast cancer. As a disease that evolves, breast cancer research benefits immensely from survival analysis, which provides insights into patient outcomes, treatment efficacy, and the influence of various factors on survival. In this paper, Haberman's Survival Dataset is used to analyze the data on breast cancer. The primary objective of this study is to establish the correlation between input and output variables, along with identifying significant features. The overarching aim of this research is to assess and compare the efficacy of various machine learning models in order to ascertain the optimal one.

Keywords: Breast Cancer, Prediction, Survival Analysis, Machine Learning Model.

1. Introduction

Breast cancer is one of the most widespread types of cancer that affects women on a global scale. This kind of cancer has the second highest mortality rate, just lower than lung cancer. According to the Cancer Statistics, 12% of newly diagnosed cancer cases are attributed to breast cancer. [1]. For women, this proportion goes up to a quarter. In developed countries, one in every nine women has a risk of breast cancer [2].

To measure the effect of certain treatments, it is very effective to calculate the survival rate. Survival analysis is a statistical approach that focuses on the time until an event of interest occurs, such as disease recurrence or death. In the context of breast cancer, this approach enables researchers and clinicians to explore not only whether these events occur but also when they occur. This temporal perspective is particularly relevant given the variable and often extended nature of breast cancer survival times [3].

Breast cancer survival analysis considers various factors that influence patient outcomes. These elements may include clinical factors (like tumor size, stage, and lymph node participation), molecular attributes (such as hormone receptor status and HER2 expression), treatment strategies (surgery, chemotherapy, radiation, targeted therapies), and individual patient factors (age, genetics, lifestyle). [4]. By incorporating these variables into survival models, researchers can identify patterns, prognostic indicators, and treatment effects.

DOI: 10.54254/2753-8818/17/20240652

Recently, the combination of machine learning methods with survival data related to breast cancer has emerged as a potent method for advancing our comprehension of the disease's behavior, forecasting patient prognoses, and refining treatment approaches. Machine learning models leverage the complexity, and volume of survival data to uncover patterns, relationships, and prognostic factors that traditional statistical methods might overlook. Machine learning encompasses a range of algorithms that can analyze vast amounts of data, identify intricate associations, and make predictions based on patterns. Applied to breast cancer survival data, machine learning models can incorporate diverse features, such as patient characteristics, genomic data, treatment information, and so on, to predict survival outcomes [5].

For the data, Haberman's Survival Dataset was used in the research. This dataset is widely recognized in the realms of survival analysis and biomedical research. Marvin Zelen and his associates introduced this in 1976, mainly focused on studying the survival rates of patients who underwent breast cancer surgery at the University of Chicago's Billings Hospital between 1958 and 1970 [6]. This dataset has been widely used in statistical and medical research to explore survival analysis techniques and gain insights into factors affecting patient outcomes after surgery for breast cancer.

In this paper, the basic situation of the dataset will be described and analyzed. Afterward, various machine learning models will be employed to assess their effectiveness in predicting outcomes. Early detection greatly improves the survival rate and accuracy of prediction. Therefore, an effective machine learning approach has the potential to enhance the overall standard of healthcare by forecasting illnesses, lowering treatment expenses, and preserving lives.

This paper combined the idea of survival analysis and machine learning. The paper introduced a technique for assessing the precision of diverse machine learning models by dividing the dataset into training and validation sets. Each model was first trained with the data from their subsets, and then the performance of each model was calculated by comparing the prediction and validating sets.

2. Methods

2.1. Dataset source

This dataset includes patients who underwent breast cancer surgery and their conditions after receiving treatment. This dataset originates from the UCI Machine Learning Repository and comprises documentation of patients who underwent treatment for breast cancer at the University of Chicago's Billings Hospital between 1958 and 1970. It encompasses a total of 306 instances, each characterized by three features: age, year of operation, and the count of positive axillary nodes. The outcome is whether patients survive for more than 5 years after surgery. For this dataset, there are no missing values and all the columns are of the integer data type. Among the patient population, 225 individuals survived beyond a five-year span, while 81 patients passed away within the initial five years, resulting in an imbalanced dataset.

The Haberman Survival Dataset consists of four features. The "age" parameter denotes the patient's age at the time of the surgery. "Year of operation" signifies the year when the surgical procedure took place. "Number of positive axillary nodes" indicates the count of identified positive axillary lymph nodes in the patient. Lymph node involvement is often indicative of disease severity. Survival status is a binary feature that indicates whether the patient survived for more than 5 years (denoted as 1) or less than 5 years (denoted as 2) after surgery. All variables are integers, and Table 1 lists their basic information.

Table 1. Basic Elements of Haberman Survival Dataset.

Elements	Range	Mean
Age	[30,83]	52.458
Year of Operation	[58,69]	62.853
Number of Positive Axillary Nodes	[0,52]	4.026
Survival Status	{1,2}	1.265

DOI: 10.54254/2753-8818/17/20240652

2.2. Methodology introduction

The research utilized Haberman's Survival Dataset sourced from the 'Centre for Machine Learning and Intelligent Systems, University of California, Irvine'. Initially, the paper scrutinized the breast cancer dataset to discern the input and output variables. In order to achieve the most accurate machine learning model, it is crucial to partition the complete dataset into training and validation sets. This enables us to ultimately assess and determine the optimal model.

Univariate analysis primarily serves to provide a summary and description of the attributes of a solitary variable. To obtain a more detailed understanding of the relationship between each input variable and output variable, studying the probability density function (PDF) is a good way. PDF refers to the likelihood of the variable assuming a specific value x. which can also be regarded as the smoothed version of the histogram. Here, three images reflecting the relationship are shown below. The vertical extent of the bar indicates the proportion of data points within the respective category.

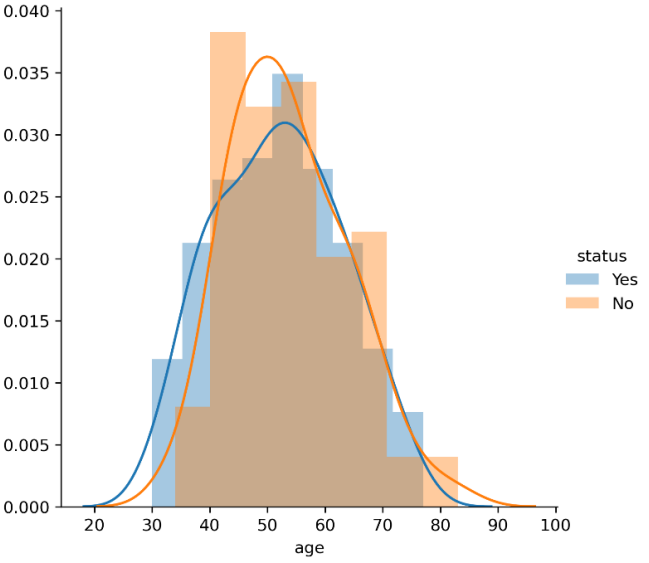

Figure 1. PDF of age.

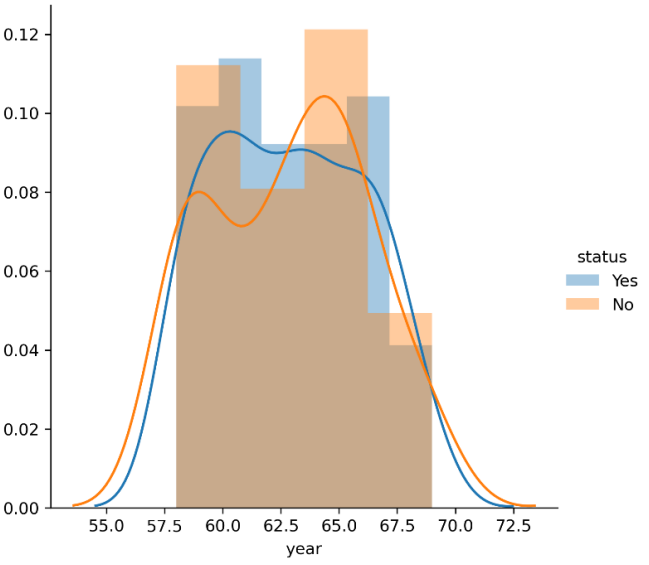

Figure 2. PDF of the year of operation.

DOI: 10.54254/2753-8818/17/20240652

For the PDF of age and year shown in Figure 1 and Figure 2, major overlapping is observed, which shows that a person's age has almost nothing to do with his survival chances. Thus, they cannot be used as a parameter to simply decide the patient's survival chances. In the PDF related to positive axillary nodes, patients with either no nodes or just one node are more inclined to survive. Conversely, the likelihood of survival drastically diminishes for cases involving 25 or more nodes. These conclusions are all based on a one-to-one relationship with a single input variable. However, comprehensive prediction requires a combination of multiple variables. Therefore, a machine-learning model is needed.

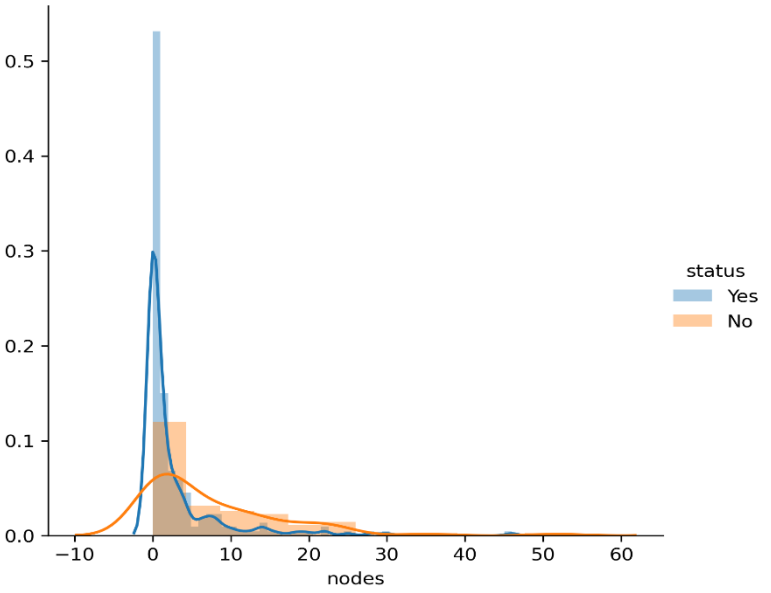

Figure 3. PDF of the detected positive axillary nodes.

3. Results and Discussion

3.1. Machine learning models

Several machine learning models can be used to predict this dataset, including XGBoost [7], linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbors, random forest, linear regression, logistic regression, decision tree, random forest, and neural network including multilayer perception [8] are all feasible models [9].

3.1.1. Linear Discriminant Analysis (LDA). This method is a supervised classification technique that identifies a linear combination of features that effectively distinguishes between various classes. It assumes that the features are normally distributed and have a common covariance matrix across classes. LDA operates by transforming the data into a lower-dimensional space through linear combinations, and it establishes decision boundaries by maximizing the variance between classes while minimizing the variance within classes. This makes LDA efficient for high-dimensional data and provides insights into class separability, although it may not perform optimally if the covariance assumption is not met. In this dataset, we can see from the graph above that the input features age and year are both nearly normally distributed. Therefore, the LDA model can perform well for the dataset.

3.1.2. Quadratic Discriminant Analysis (QDA). This is another supervised classification method similar to LDA, but it does not assume a common covariance matrix across classes. Instead, it estimates a unique covariance matrix for each class. As a result, QDA is more flexible and can handle cases in which classes have different covariance structures. However, it requires a larger dataset to accurately estimate these matrices and can be computationally more expensive than LDA because of the increased complexity of

DOI: 10.54254/2753-8818/17/20240652

decision boundaries. Since there are only 305 objects in the dataset, it is not necessary to use this model in the dataset to reduce the dimension costing much complexity. In other words, the performance of QDA should be similar to that of LDA but has more complex calculations.

3.1.3. K-Nearest Neighbours (KNN) Analysis. This is a non-parametric, instance-based algorithm employed for both classification and regression tasks. Unlike LDA and QDA, KNN doesn't undergo a distinct training phase and retains the entire dataset in memory. To classify a new data point, it identifies the 'k' closest neighbors in the feature space and determines the majority class among them. KNN is simple and intuitive and can handle complex decision boundaries. Nonetheless, it is crucial to select an appropriate value for 'k', and it can be computationally intensive, especially when dealing with sizable datasets, as it needs to compute distances between the new point and all existing data points. For this dataset, the input features are normally distributed. In addition, this survival dataset is not large, so the KNN method works well.

3.1.4. Linear Regression. This statistical technique is employed to characterize the connection between a continuous dependent variable and one or more independent variables. It presupposes a linear association between the variables, striving to identify the most suitable line (or hyperplane in multiple dimensions) that minimizes the sum of squared variances between the observed and predicted values. Linear regression is widely applied in fields such as economics, biology, and engineering for tasks such as predicting sales figures, estimating house prices, and modeling physical phenomena. It offers insights into how alterations in the independent variables correlate with changes in the dependent variable. This is a widely used model, but it is not appropriate for the dataset because all features are integers and the relationship between variables is not linear.

3.1.5. Logistic Regression. Despite its name, is not a regression technique but rather a classification algorithm. Logistic regression is employed when the dependent variable is binary, indicating two potential outcomes. This type of regression models the probability of the binary response variable based on one or more predictor variables. It utilizes a logistic function to confine predicted probabilities within the range of [0, 1]. This makes it well-suited for tasks such as predicting the likelihood of a disease occurrence, customer churn, or whether an email is spam or not. Logistic regression is widely employed in medicine, social sciences, and other fields where binary classification is necessary. For these reasons, this is a good choice for the Haberman Survival dataset.

3.1.6. Decision Tree. This machine-learning algorithm is adaptable and easily understood, suitable for both classification and regression tasks. It functions by iteratively dividing the feature space using defined criteria at each node, constructing a hierarchical arrangement of nodes and branches. This mechanism enables the tree to arrive at decisions by traversing a path from the root node to the leaf node. Decision trees are known for their interpretability, which makes them valuable for understanding the underlying patterns in data. However, they can be prone to overfitting if not properly pruned or if the tree grows too deep. Decision trees find applications across a broad spectrum of fields, spanning from finance and healthcare to natural language processing and recommendation systems. However, regarding the survival dataset, there can be significant differences between the patients even if they have similar conditions. The characteristics of different people lead to the inapplicability of the method.

3.1.7. Random Forest. This technique falls under ensemble learning in machine learning. It amalgamates multiple decision trees to forge a more resilient and precise predictive model. The process involves generating numerous decision trees in the training phase and producing the mode of the classes (for classification) or the mean prediction (for regression) derived from the individual trees. Random Forest introduces randomness through two main methods: firstly, by randomly picking a subset of data for training each tree (bootstrapping), and secondly, by utilizing only a random subset of features at each split in a tree. This diversity among the individual trees aids in mitigating overfitting and improves the

DOI: 10.54254/2753-8818/17/20240652

model's ability to generalize effectively [10]. As a result, Random Forest is widely used for tasks such as classification, regression, and feature importance ranking across various domains. Similar to the decision tree, it cannot achieve good performance in this dataset.

3.1.8. Multilayer Perceptron (MLP). This is a form of artificial neural network comprising numerous interconnected nodes, often referred to as "neurons," arranged in layers. It operates as a feedforward neural network, signifying that information progresses in a unidirectional manner—from input nodes through hidden nodes to output nodes. Within the hidden layers, each node computes a weighted sum of inputs, followed by the application of an activation function. This mechanism enables the network to discern intricate relationships within the data. MLPs are capable of learning and approximating non-linear functions, making them highly adaptable for various of tasks, including classification, regression, and pattern recognition. They are trained using techniques such as backpropagation, where errors are propagated backward through the network to adjust the weights and biases, enabling the model to improve its predictions over time.

3.2. Accuracy score

In this study, the goal is to predict the outcome with only three input features. To achieve good performance, several machine learning models are used. The principles of these models were analyzed in a previous article. In addition, the matching degree of each model to the data set, and the analysis and discussion of the predicted results, have been discussed. However, all of these are based on the most basic machine learning models. K-fold cross-validation is an effective technique for assessing the performance of any model. It entails partitioning the dataset into k equally sized subsets, or "folds." The model undergoes training and evaluation k times, with each fold acting as the test set once, while the remaining k-1 folds are utilized for training. This iterative process enables a thorough evaluation of the model's performance across various data subsets. The outcomes from each fold are then averaged, yielding a more dependable estimate of the model's performance. This practice aids in mitigating overfitting risks and furnishes a more robust assessment of a model's ability to generalize.

According to each machine learning model, the average accuracy results are in Table 2:

Table 2. Accuracy score of each model.

	LDA	QDA	KNN	Linear	Logistic	Decision Tree	Random Forest	Multilayer Perceptron
Accuracy score	0.741	0.745	0.758	0.554	0.734	0.663	0.712	0.812

It's crucial to recognize that survival rates are approximations and may not precisely foretell the specific outcome for an individual because each person's case is unique. They serve as general guidelines for understanding the potential course of the disease and can help make informed decisions about treatment and care. From the table, Multilayer Perceptron is the best model for the survival dataset. In addition, it was found that lymph node does have a great impact on survival rates. Patients who have more than one detected lymph node are not anticipated to survive for more than five years.

4. Conclusion

From the results, it is a reliable way to forecast and classify the patient's survival time using the three input features. In addition, the input feature of the auxiliary nodes shows a strong correlation with the survival status which means that it can make the prediction more accurate. It is an effective way to diagnose breast cancer with the dataset by using machine learning models and data analysis methods.

In the above article, the measure was whether the patient survived for more than 5 years after surgery. The precision of the one-layer network is 81.2%. This paper outlines the process of examining the existing dataset and choosing a suitable machine learning model based on its specific attributes. To enhance performance, the study employs the k-fold cross-validation technique.

DOI: 10.54254/2753-8818/17/20240652

Normally, people think that the main factors affecting the disease are age and the year of surgery. Younger people indeed have a higher chance of survival. However, people of all ages have the probability of getting breast cancer. In this study, the positive axillary node was found to be an important feature. People with more nodes are less likely to survive. This can not only be a method to judge and predict survival but also be an important indicator of breast cancer. Those who were detected to have axillary nodes are very likely to develop the disease. In other words, there is a high sensitivity between these two income and outcome features. The dataset indicates that individuals with a low count of detected axillary nodes have a significantly elevated likelihood of survival, whereas a person with many detected axillary nodes has almost no chance of survival. Therefore, it can be a good way of prevention and early treatment that monitoring the number of axillary nodes regularly to avoid breast cancer and increase the chance of survival.

References

[1] Ferlay J, et al. 2019 Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, 144(8), 1941-1953.

[2] Lánczky A, et al. 2016 Survival Analysis Tips and Tricks Using R. *Briefings in Bioinformatics*, 17(2), 307–315.

[3] Harris L, et al. 2007 American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology*, 25(33), 5287-5312.

[4] Haberman S J 1976 Generalized residuals for log-rank tests. *Lifetime Data Analysis*, 2(2), 161-171.

[5] Cook R J and Lawless J F 2007 The statistical analysis of recurrent events. *New York: Springer*.

[6] Liu P, et al. 2021 Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer. *IEEE Transactions on Biomedical Engineering*, 68, 148-160.

[7] Lotfnezhad A H, et al. 2021 Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation. *Iran J Public Health*, 50(3), 598-605.

[8] Li J, et al. 2021 Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One*, 16(4).

[9] Montazeri M, et al. 2016 Machine learning models in breast cancer survival prediction. *Technol Health Care*, 24(1), 31-42.

[10] Mihaylov I, Nisheva M and Vassilev D 2019 Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. *Information*, 10(3), 93.