Prediction of Living Status for Breast Cancer Patients Using Machine Learning

Yaqian Han *

Department of Arts&Science, University of Toronto (St.George), Toronto, Canada, M5S1A1

* Corresponding Author Email: yvonnehan777@163.com

Abstract. Due to reasons such as diet, living habits, and genes, etc., breast cancer has gradually become a common high mortality disease and is not limited to women. Many patients suffer from terrible illness or even death due to untimely treatment or incorrect prediction of tumors. This study has 500 observations in total through utilizing the combination of oversampling and undersampling. What's more, by using advanced machine learning, the method of recursive feature elimination and random forest, this research constructs a model and select four types of protein content in the body and age as the most relevant features in predicting patients' living status. The model has a relative high accuracy 98.052% and can provide physicians relative accurate information to give targeted treatment. It also can be utilized as a warning to remind patients of their current physical condition. Early treatment and accurate diagnosis can significantly improve patient survival rates.

Keywords: Breast cancer, Machine learning, Random Forest, Recursive feature elimination.

1. Introduction

According to the statistics, breast cancer is the second most common cause of cancer-related deaths among human and one of the serious diseases that spreads the fastest [1]. Nearly one eighth of women will be diagnosed in their lifetime [2]. As the cancerous malignant lumps start to grow in breast cells, breast cancer begins to develop [1]. Therefore, early symptoms are difficult for physicians to detect and rely more on their experience. Sometimes misdiagnosis of tumors may cause patients to miss the best opportunity for treatment. Different breast cancer treatment has become necessary for these patients and by collecting basic testing resulting, physicians could implement corresponding therapies to treat as much as possible [2]. There is another research aiming at doing breast cancer diagnosis and prognosis by using machine learning. It specifies at the general use of decision tree, k-nearest neighbor, and support vector machines, etc. [3]. In the past few decades, many breast cancer patients were suffering from the pain and physicians could only depend on their experience to make diagnosis. Therefore, the accuracy cannot be guaranteed. However, the machine learning makes great contribution in the healthcare systems to enhance the accuracy of prognosis depending on the huge electronic database, such as the European Friedreichs Ataxia Consortium for Translational Studies (EFACTS) patient database [3]. What's more, the improved Gini index in random forest method is also used as more accurate measure algorithm to select features and make diagnosis [4]. Such intelligent techniques in healthcare system helps physicians understand the patient's physical conditions and detect latent diseases within the patient to some extent.

The data in this research resourced from Kaggle.com which contains the living status of patients and other information of corresponding patients [5]. The goal of this study is to predict the living status of these women who suffer from breast cancer base on the collected information of age, different parameters of four kinds of proteins, histology, ER status, PR status, HER2 status, etc. The predicted model could better provide patients prognosis based on their own physical conditions and timely do adjustment of treatment plan to some extent.

2. Data reprocessing

2.1. Data description

The dataset contains 334 observations and 15 variables including five numerical variables and ten categorical variables. The response variable in this study is patient status and our goal is to find how other predictor variables affect it. This is the description of all variables in the dataset as shown in Table 1.

Table 1. Description of all variables in dataset.

Name of Variables	Descriptions
Age	Numerical variable, recording the patients' age
Gender	Categorical variable, recording the patients' gender
Portein1	Numerical variable, parameters for certain protein 1
Portein2	Numerical variable, parameters for certain protein 2
Portein3	Numerical variable, parameters for certain protein 3
Portein4	Numerical variable, parameters for certain protein 4
Tumor_Stage	Categorical variable, representing patients' tumor stage they are located
Histology	Categorical variable, categorizing histology into two different carcinomas
ER status	Categorical variable, testing whether patients suffer from breast cancer with Estrogen receptor or not
PR status	Categorical variable, testing whether patients suffer from breast cancer with Progesterone receptor or not
HER2 status	Categorical variable, testing whether patients' breast tissue have extra HER2 receptors or not
Surgery_type	Categorical variable, giving four different kinds of surgery types as treatment
Date of Surgery	Categorical variable, recording the date of patient's surgery
Date_of_Last_Visit	Categorical variable, recording the date of patients' visiting
Patient_Status	Categorical variable, showing the living status of patients

The first step we need to do is that for some missing data in the response variable, we directly drop corresponding rows. After doing this step, our observations have been deleted into 321. Next, for these categorical variables: gender, HER2 status, Patient status, histology, tumor stage and surgery type, it's necessary to categorize and transform the label values into numeric values by using one-hot encoding since some methods in machine learning algorithms cannot operate label values. These categorical variables don't have the ordinal relationship between each other, and one-hot encoding used here could eliminate the effect of natural ordering in the model and give a better performance in the prediction [6]. As shown in Table 2. The form indicates how each variable in different types mapping to 0 and 1.

Table 2. One-hot encoding on categorical variables.

Name of Variables	0	1
Gender	Female	Male
HER2.status	Negative	Positive
Tumor_Stage	Not "I"	"I"
Tumor_Stage	Not "II"	"II"
Histology	Not "Infiltrating Ductal Carcinoma"	"Infiltrating Ductal Carcinoma"
Histology	Not "Infiltrating Lobular Carcinoma"	"Infiltrating Lobular Carcinoma"
Surgery_type	Not "Lumpectomy"	"Lumpectomy"
Surgery_type	Not "Radical Mastectomy"	"Radical Mastectomy"
Surgery_type	Not "Simple Mastectomy"	"Simple Mastectomy"
Patient_Status	Dead	Alive

However, it's not necessary to select all variables in the dataset but select some important variables and implement the feature selection in the next step. Variables ER status and PR status have been deleted since they both show positive in all results which means that these two features are the same among all patients and are useless for the following filter feature selection. Also, we'll not focus on variables date of surgery and date of last visit even though we can fill it with some special values for missing data and keep all observations with complete data. However, each patient is independent with each other and normal time-series plot for different patients is meaningless for studying the relationship between relevant factors and patients' living status. Therefore, these two variables will also be deleted.

2.2. Variable normalization

Then, it's mandatory to standardize some numerical variables into [0, 1] so that it can eliminate the impact of dimensionality and avoid the problem of letting some variables with large-value dominate other relatively small-value variables [7]. It's hard for the prediction model to capture information in some small-value variables and cause contribution power to the model is relatively low. Here, we use min-max normalization method on variable age to make it have the same scale with other variables.

$$x^* = \frac{x-min}{max-min} \tag{1}$$

Where *min* is the minimum of sample data, max is the maximum of sample data.

Need to note that this method is sensitive to outliers since both nominator and denominator is determined by the maximum and minimum. But the distribution of age in our dataset is in reasonable range and we can use this method to standardize age. Finally, merge the data organized in the above steps.

2.3. Resampling:

After doing the data reprocessing, we can observe that the dead patients are 66, but alive patients are 255 in total. Such imbalanced data will cause less information to be extracted from dead patients and have more tendency in extracting from alive patients. Therefore, we combine oversampling and undersampling together for imbalanced classification. Such method reweights data by decreasing the bias in majority class and increasing the bias in minority class. By doing resampling, we get 500 balanced observations finally and there is not much difference between these two categories, with 244 alive patients and 256 dead patients.

3. Feature Selection:

Feature selection is a technique that can select the most relevant features among the subset of all possible features and give a more accurate prediction in machine learning algorithm. In this step, we'll conduct recursive feature selection which could fit the model and remove the features that have less relationship for each loop until it keeps the specific number of features. The rank is depended on the model's coefficients and the attribution of feature importance. But recursive feature selection doesn't always know the optimal number of features that it should keep [8]. Therefore, we combine the use of recursive feature elimination algorithm and 10-fold cross validation. Cross validation refers to a technique that can calculate and select the best collection of features based on accuracy among all different feature subsets. By using recursive feature selection, it could eliminate dependencies and collinearity that could potentially exist in the model [8].

After doing the feature selection, we find the most relevant variables: 'Protein4', 'Protein2', 'Protein1', 'Protein3', 'Age', 'I', 'II' (Tumor_Stage), 'Lumpectomy', 'Simple' (Surgery_type), 'HER2.status', 'Ductal', and 'Lobular' (Histology).

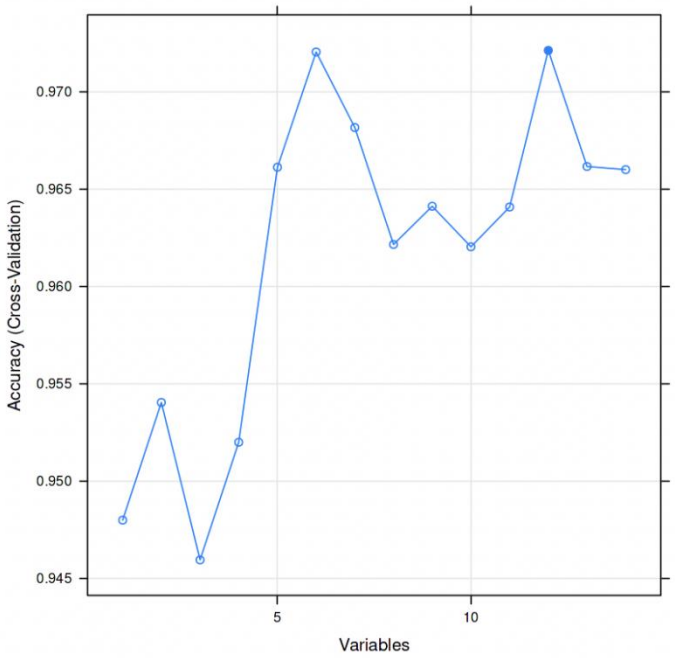


Figure 1. Accuracy of Cross-validation vs. number of variables.

As the Figure 1. shown above, the accuracy of the cross validation reaches the peak when the number of variables is 12 even though we can observe that it's not much different from the accuracy when the number of predictors is 6. With 12 predictors, it shows a high accuracy which is over 0.975.

4. Modeling

After using the combination of oversampling and undersampling, we have 500 observations in total. The response variable living status has two values, 0 and 1, which means that it's a binary classification. We'll conduct a common technique in machine learning for constructing prediction model called random forest. The goal of doing random forest is to reduce the number of predictors so that it can decrease the burden of collecting data and improve efficiency through selecting subsets containing important features [9]. A decision tree is the unit of constructing random forest which could split the dataset using decision nodes recursively until nodes have been left with pure leaf nodes [10]. The optimal split is to maximize the entropy gain which is a measure of unpredictability. However, single decision tree is highly sensitive to the training data, and it requires multiple trees to make less sensitive and reduce variance [10]. By selecting a subset of features randomly to reduce correlation for each tree, fit a new data point with each tree and choose the majority result of prediction (0 or 1) in classification. The process is called bagging and it shows how random forest works. In our study, first, the data has been split into train data and test data with the split ratio equal to 0.7. Next, we fit the random forest into the training dataset and use the model to predict the test set.

5. Results

The out-of-bag error could measure the prediction error of the random forest which is 8.09% in our study. We also obtain the confusion matrix for the test set which gives the accuracy and prediction power of this model as shown in Table 3.

Table 3. Confusion matrix for the test set

		Actual values	
		1	0
Predicted values	1	TP=72	FP=0
	0	FN=3	TN=79

TP refers to true positive that is we predicted positive and it's also true. TN refers to true negative that is we predicted negative and it's also true. FP refers to false positive that is we predicted positive but it's false. FN refers to false negative that is we predicted negative but its false [7]. The diagonal of the confusion matrix represents the correct positive prediction and correct negative prediction which are both equal to 1 and 0 for predicted value and true value, respectively. In other words, the diagonal represents the number of living status results that the prediction model could accurately predict. And other entries imply the number of living status results that have been wrongly predict. As we can observe, correct predictions occupy the majority, with 151 (72+79) in total.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} * 100\% \quad (2)$$

After calculation, the accuracy of this prediction model is around 98.052% which implies a relatively high prediction power.

The confusion matrix could be used to determine the efficiency of model based on the number of correct predicted positive results against the number of all positive cases in total [4], defined as sensitivity shown below:

$$Sensitivity = \frac{TP}{TP+FN} * 100\% \quad (3)$$

The sensitivity of this model is 96% after calculation.

Also, the confusion matrix could be used to determine the efficiency of model based on the number of correct predicted negative results against the number of all negative cases in total [4], defined as sensitivity shown below:

$$Specificity = \frac{TN}{TN+FP} * 100\% \quad (4)$$

The sensitivity of this model is 100% after calculation.

This visualization of random forest shown in Figure 2. Includes three different lines. For colored lines, they represent errors for different classes as trees increase and the black line represents the error of out-of-bag samples as trees increase. This plot shows that over 300 trees, errors are becoming generally stable.

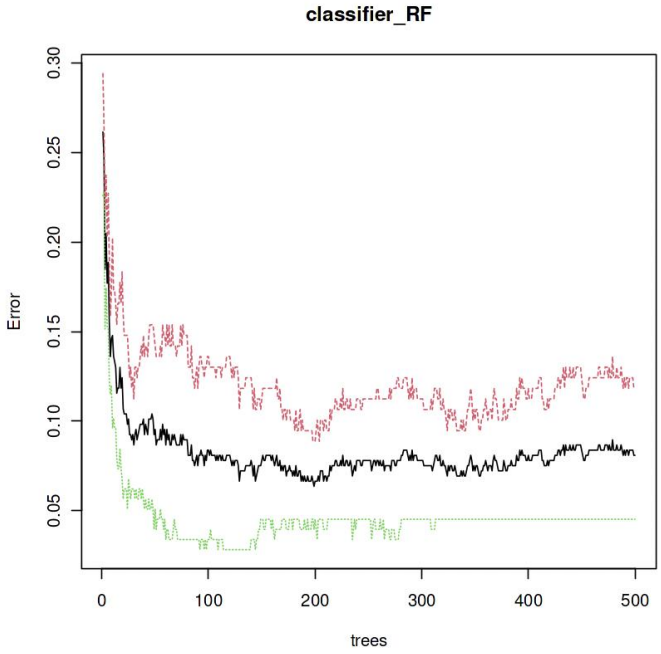


Figure 2. Error analysis of Random Forest.

Mean decrease Gini represents the importance of features or how these features contribute to leaves and nodes in random forest. The higher value of Mean decrease Gini indicates the higher

importance of corresponding features in the prediction model. As shown in Table 4. And Figure 3. Age and four types of protein have significantly higher mean decrease Gini than that of other predictor variables, with four kinds of protein approaching 30 and age 26.30. Other predictor variables show their mean decrease Gini scores are below 5. The scores also mean that their contribution to this prediction model is relatively large. In the future prediction of breast cancer, we should mainly focus on these five important variables to improve the accuracy of predicting patients' living status.

Table 4. Mean decrease Gini score.

Variable Names	Mean Decrease Gini
Age	26.301248
Protein1	29.052891
Protein2	29.932734
Protein3	27.046964
Protein4	30.188256
HER2.status	3.140749
I	4.400058
II	3.421577
Ductal	2.984370
Lobular	2.569223
Lumpectomy	4.565382
Simple	2.906093

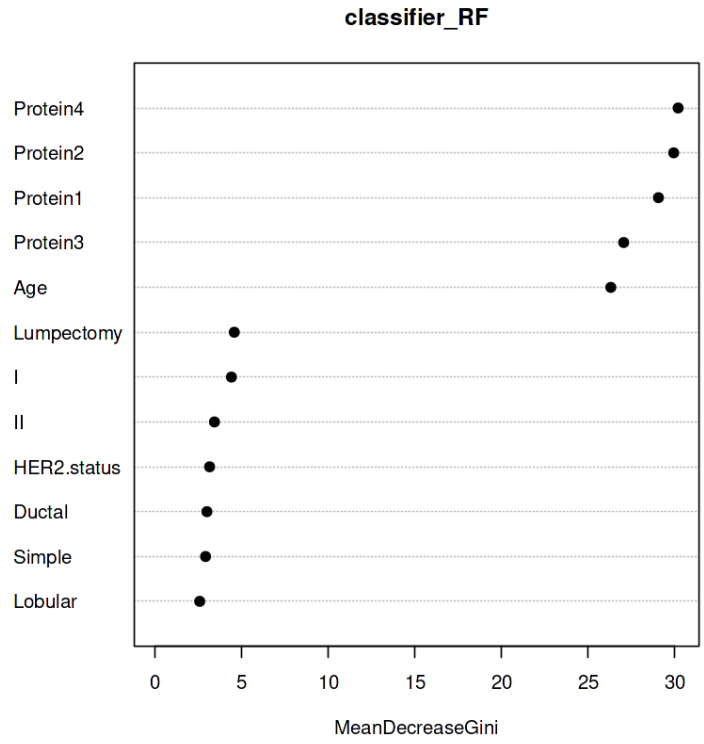


Figure 3. Mean decrease Gini score and its rank.

6. Conclusion

Overall, this research starts with doing data reprocessing, using one-hot encoding and variable normalization so that machine learning algorithm could process. Then, combine oversampling and undersampling to get 500 observations in total for the further model construction. By using recursive feature selection, variables have been reduced into only twelve as the most important relevant features in predicting patients' living status. Finally, implement random forest and obtain five important

variables with high mean decrease Gini score, four types of protein and age. The prediction model has 98.052% accuracy based on the results in confusion matrix.

This model also has certain biological and clinical application value, for example, prediction can help patients to carry out targeted early prevention of breast cancer. Meanwhile, relevant important features should be paid more attention to at the early stage of breast cancer to prevent and treat tumor deterioration. Based on model data, it can diagnose diseases with high accuracy and help doctors better provide information for treatment.

Reference:

[1] Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019, June). Machine learning classification techniques for breast cancer diagnosis. In IOP Conference Series: Materials Science and Engineering (Vol. 495, p. 012033). IOP Publishing.

[2] Maughan, K. L., Lutterbie, M. A., & Ham, P. S. (2010). Treatment of breast cancer. American family physician, 81(11), 1339–1346.

[3] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[4] Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., & Alzahrani, A. K. (2022). Fuzzy neural network expert system with an improved Gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia. Big Data and Cognitive Computing, 6(1), 13.

[5] Rajani, K. (2023, March 24). Breast cancer survival dataset. Kaggle. https://www.kaggle.com/datasets/kreeshrajani/breast-cancer-survival-dataset

[6] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[7] Singh, B. K., Verma, K., & Thoke, A. S. (2015). Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. International Journal of Computer Applications, 116(19).

[8] Sharma, N. V., & Yadav, N. S. (2021). An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers. Microprocessors and Microsystems, 85, 104293.

[9] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101.

[10] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.