Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230796

An empirical comparison of machine learning methods for predicting breast cancer probability

Wenqiang Ge

Rensselaer Polytechnic Institute, School of Science, Troy, NY, 12180, U.S.

gew@rpi.edu

Abstract. Nowadays, there are many kinds of cancer that people cannot prevent early, so those diseases cannot be controlled effectively when people detect them. To solve the problem, more and more scientists started to study the features and possible reasons for those diseases. This study concentrates on the relationship between some basic physical features, breast cancer diagnosis, and patients' possibilities of getting breast cancer. The main methods in this research are machine learning. It includes logistic regression, decision tree, random forest, k-nearest neighbor, and gradient boosting model. These methods use their main ideas to construct and fit the training data, then they predict the testing data and compare them with the testing set. Each model has its prediction results and performance values. After that, they will also herald the patients' probability of getting breast cancer with their physical conditions. According to the evaluation values, the random forest model performed best. Regarding probability prediction, the models have almost similar results. This study is a simple attempt to study the possible reasons for breast cancer. People could concentrate on increasing the number of independent variables in future studies. For example, they can add potential triggers of breast cancer. This suggestion will improve the efficiency of disease prevention and explore some helpful treatments.

Keywords: machine learning, breast cancer, classification.

1. Introduction

With the continuous improvement of living standards, people have started paying more attention to their daily health. In the study of probability prediction, people are working hard to get more accurate probability results for those widespread diseases. Cancer is one of the most common and severe diseases among those prevalent diseases. Because some specific kinds of cancer are not solved completely, people cannot detect or treat them on time before the illness aggravates. Since the fatality rate is very high for most kinds of cancer and the condition worsens quickly and severely, people should discover and prevent them as soon as possible.

To check the cancer condition, scientists begin to research some physical features of patients and judge the diagnosis through these features. In addition, scientists also want to detect and prevent diseases on time, so they are trying to model and predict the probability of patients getting cancer. Scientists have already studied the possibility and effective treatment for those common types of cancer disease. After people find a method to calculate the probability of getting cancer, people will reserve more time to prevent and even avoid cancer.

DOI: 10.54254/2755-2721/14/20230796

However, some other types of cancer are not very common, so people need to research and try to prevent them in the future. One specific kind of cancer, breast cancer, is to be taken out for more detailed research.

A science-based machine learning method is used to learn and make a prediction. The main idea for machine learning is first to select some data from the whole data set as the training set. Then it will use these data to construct and adjust a model to make classification, prediction, or calculation [1]. Since machine learning already has many helpful classification and feature selection algorithms, it is also appropriate to use these algorithms to predict the probability of breast cancer.

Before getting the probability prediction, the machine learning method will separate the whole data set into two groups. One is called training set, and the other one is testing data. The machine learning algorithms usually use the training set to construct the model for later judgment. There are several different algorithms chosen to make a comparison of model performance. To be more specific about the model performance, the prediction accuracy will be calculated, and the ROC curve and AUC score will be quoted. After these values are estimated, they will be visualized in the plots for later comparison and analysis.

Finally, about the probability analysis, the same as the previous performance comparison, the predicted probabilities from each algorithm will also be visualized and analyzed together.

2. Method

2.1. Dataset

The dataset used in this research project is mainly about some physical features of each possible patient and their corresponding diagnosis. The dataset describes the primary condition of each breast cancer patient, including the breast's size, area, texture, and smoothness. This is the target variable for the diagnosis, and it only has two results, 0 or 1. The diagnosis result is a binary variable. 0 means that the patient does not get breast cancer recently, and one indicates that this patient was confirmed to have breast cancer. The dataset has 570 patients, and to fit the models, it is separated into two sets. One of the sets is the training data, accounting for 80% of the whole dataset randomly. The training data used to fit the model can help the model learn the tendency and features of this data set. The other one is called testing data. The testing data means that it is used to test the model, and then it will give feedback to the model for later improvement. Since the research will utilize several different models to fit and predict the diagnosis of patients, it is indispensable to distribute the same patient information into training and testing groups. Therefore, before training the models, the data set is first marked by the seed number, randomly separating the whole dataset into groups. After that, the same training and testing data groups will be utilized in the model fitting and prediction. The advantage of this setting is that this will provide a fairer result for the comparison and conclusion.

2.2. Models

2.2.1. Logistic regression. For the first model, it is very convenient to choose the logistic regression. Logistic regression, a machine learning algorithm, connects and predicts the relationship between given features and target variables [2]. The reason for choosing Logistic regression is that this model can fit the relationship between the continuous features and the discrete variable. Logistic regression provides more precise results on numerical data values [2]. To be more specific, the predicted results from the logistic regression not only contain the continuous value but also can describe the discrete results [2]. From the perspective of machine learning, it is a supervised model. Given that it makes use of the categorical answer variable, it is also a classification method. Finding a correlation between certain traits and the likelihood of an outcome for a dependent variable is the goal of logistic regression.

2.2.2. Decision tree. When using the decision tree as a model to fit the data, it can be understand to use many feature questions to separate the whole dataset into several classified groups. For example,

DOI: 10.54254/2755-2721/14/20230796

considering the patients in the dataset have physical features that describe their physical condition, each of their physical conditions can be regarded as healthy or unhealthy. Therefore, just following this kind of procedure, the diagnosis of patients will be known in the last step of classification. Each feature question is like a tree branch, so the method is named decision tree. Each component in the decision tree has an index called Gini which describes the purity of each time separation [3]. The smaller Gini is, the higher the data purity in each separation. Unlike logistic regression, before getting final results from the decision tree, there should be a threshold to stop the model; otherwise, the decision tree will overfit the dataset, making the predictions inaccurate [4].

2.2.3. Random forest. There are many options for feature questions in each branch, and each feature question has a different emphasis, so the classification results will also be different. To minimize the effect of feature question selection, there is another method called random forest. The random forest method is a combination of many decision trees. After trying many decision trees, each tree will have different classification results. At that time, the random forest model will choose the combination of one specific tree to make the accuracy as greater as possible.

2.2.4. K-nearest neighbor. The k-nearest Neighbor method, also known as the KNN method, is supervised and unsupervised. As a supervised method, just like the condition in this research, the diagnosis of each patient will be the final result of the prediction. As an unsupervised method, the classification groups will be the results of this method. The general idea of the supervised method of KNN in this background is to determine the category of k samples. Specifically, if there are k samples in the same group, this group's category is determined by the most number of categories [5].

2.2.5. Gradient boosting. Generally, gradient boosting is to fit many times for each data group and minimize the difference between the actual value and predicted result. In particular, given a judgment object made up of several weak learners, the cumulative model loss after including the weak learner must match the negative gradient of the loss function in the direction of the negative gradient [6].

2.3. Evaluation

After completing the whole process of fitting and predicting from each model, several essential evaluation scores are used to judge the performance of the model's prediction. These evaluation scores include the accuracy, error, precision, recall, and F-1 score, which describe different aspects of the model's prediction. The quotient between the number of correct predictions and the total number of predictions is the formula of accuracy. The more accurate score is, the more precise the model's prediction is.

The accuracy score in machine learning is an evaluation value that measures how many the correct predictions a model makes concerning the total number of predictions made. The error of a set of prediction could be achieved by the quotient between the number of correct predictions and the number of predictions [7]. The model's error is just the result of one minus the model's accuracy. It indicates the proportion of incorrect predictions made by the model. Precision in machine learning is calculated by dividing the number of correct-positive predictions (the positive number means the diagnosis of 1) by the total number of positive results. The precision of machine learning indicates the quality of how well the positive predictions made by the model. The recall value can use truly positive cases divide the sum of true positive and wrongly negative (negative here means the diagnosis of 0). About the recall score, it evaluates the rate of actual positive results predicted by the model [8]. Finally, the F1-score is calculated by dividing the product of recall and precision by the sum of recall and precision and then times the result by two. The F1-score is always gotten by operating the values of recall and precision, and it is another way to get the 'average' score from recall and precision values.

In addition, another kind of evaluation indicator, the AUC score, is used in this model performance judgment. The AUC score is the total covered area under the ROC curve, according to its definition. The AUC score indicates the proportion of correct predictions under the probabilistic framework.

DOI: 10.54254/2755-2721/14/20230796

Compared with each model's accuracy, the AUC score can provide a more complex relationship and logic between independent and target variables [9]. The AUC calculation approach takes into account the classifier's capacity to identify positive and negative situations, and it can still evaluate the classifier in the presence of unbalanced samples. For instance, in a situation when there are only a few positive examples (assume 0.1%). If accuracy evaluation is used and all samples are predicted as negative examples, an accuracy rate of 99.9% can be obtained [10]. However, if using AUC and indicating all samples as negative examples, true positive rate and false positive rate are both 0 (without positive) and are connected, resulting in an AUC of only 0.5, successfully avoiding the problem caused by uneven samples [10]. Considering the probability of diagnosis, the AUC score of a model also has a great reference value compared with accuracy since it will reduce the impact of spam information or value.

3. Result

The model results are mainly those evaluation values and the testing patients' probabilities of getting breast cancer or not. Each model also includes a fuzzy matrix, which compares the condition of predicted and actual results.

3.1. Quantitative result

After getting each model's evaluation values, comparing each model's performance with different plots is convenient. Figure 1 visualizes each algorithm's evaluation values, with values in demonstrated in Table 1. Regarding accuracy, the logistic regression model has the greatest value. This shows that the logistic regression has correct predictions compared to the actual results. Random forest and logistic regression both have the highest value in terms of accuracy. The KNN method has the most recall value, and the random forest has the greatest F1 score.

Table 1. Performances of each machine learning algorithms.

Evaluation Matrix	Logistic Regression	Decision Tree	Random Forest	KNN	Gradient Boosting
Accuracy	0.9298	0.8772	0.9386	0.9123	0.9123
Precision	0.8958	0.8750	0.8958	0.8125	0.8750
Recall	0.9348	0.8400	0.9556	0.9750	0.9130
F1-Score	0.9149	0.8571	0.9247	0.8864	0.8936

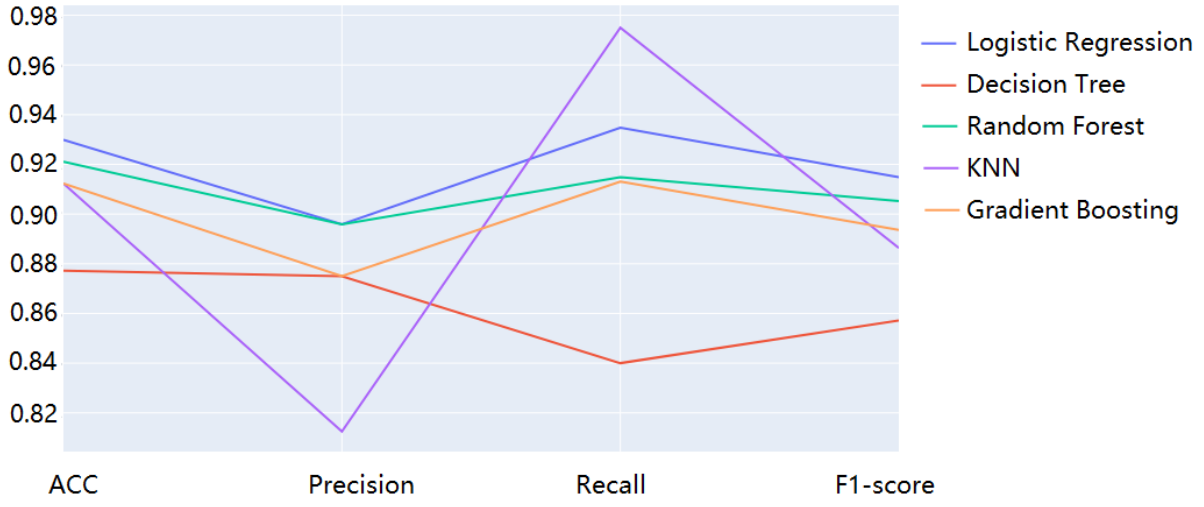

Figure 1. Line chart of Evaluation values.

DOI: 10.54254/2755-2721/14/20230796

3.2. ROC curve and AUC score

The AUC scores and ROC curves are demonstrated in Table 2 and Figure 2 respectively. AUC is the area under the ROC curve, with the true positive rate on the vertical axis and the false positive rate on the horizontal axis of the ROC curve, respectively. Y=X occurs when the two are equivalent. The ROC curve and AUC score are not affected by the data proportion. Although the accuracy and precision of each model are not the same in the previous table, the AUC score of some models is the same. Since the training and testing data are set before, each model has used the same set of data to make a prediction. Therefore, these AUC scores are more persuasive in indicating the ability of the model's prediction.

Table 2. AUC of different algorithms.

Algorithm	AUC score
Decision Tree	0.979798
Gradient Boosting	0.974116
KNN	0.969381
Logistic Regression	0.969381
Random Forest	0.969381

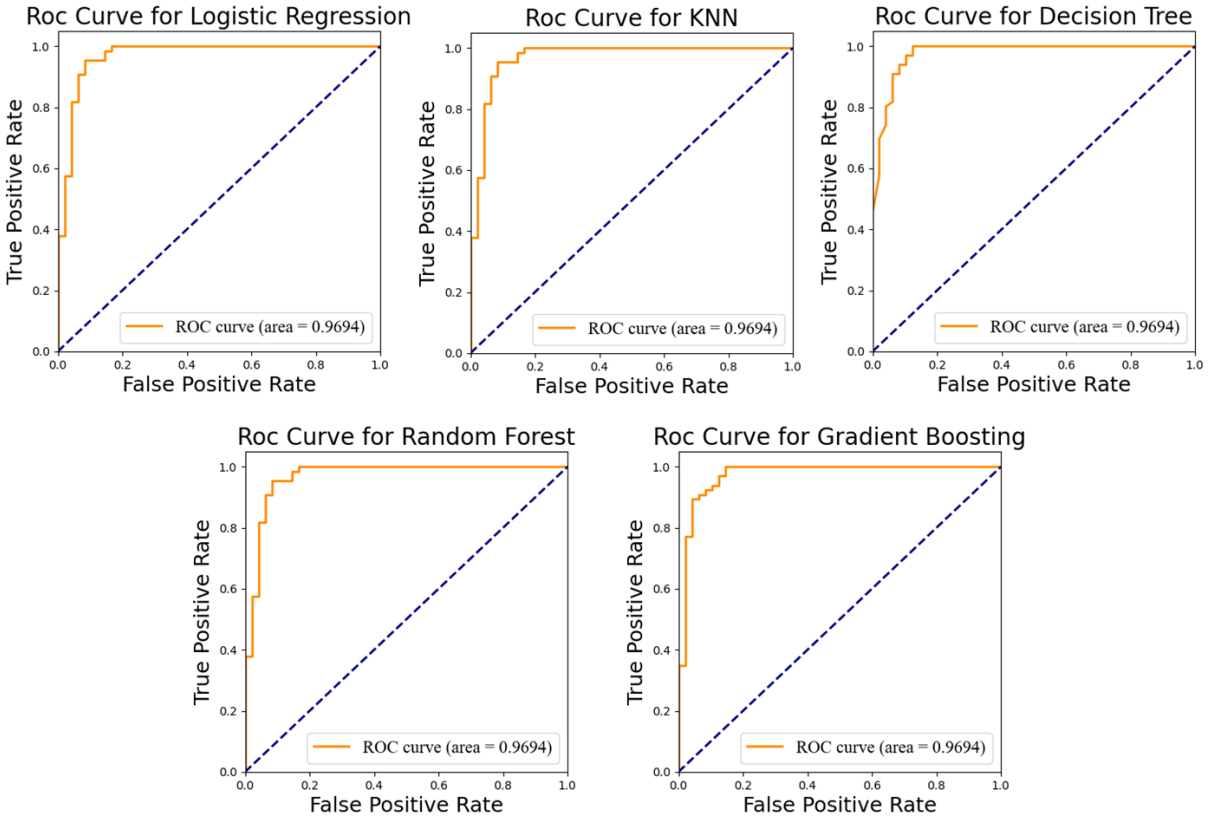

Figure 2. ROC curve of different models.

3.3. Probability of getting cancer

Figure 3 shows the predicted probabilities of patients getting cancer by each algorithm. Almost every patient they have a similar probability of getting cancer by each model. However, there is also a difference between some prediction results because of the model's calculation method.

DOI: 10.54254/2755-2721/14/20230796

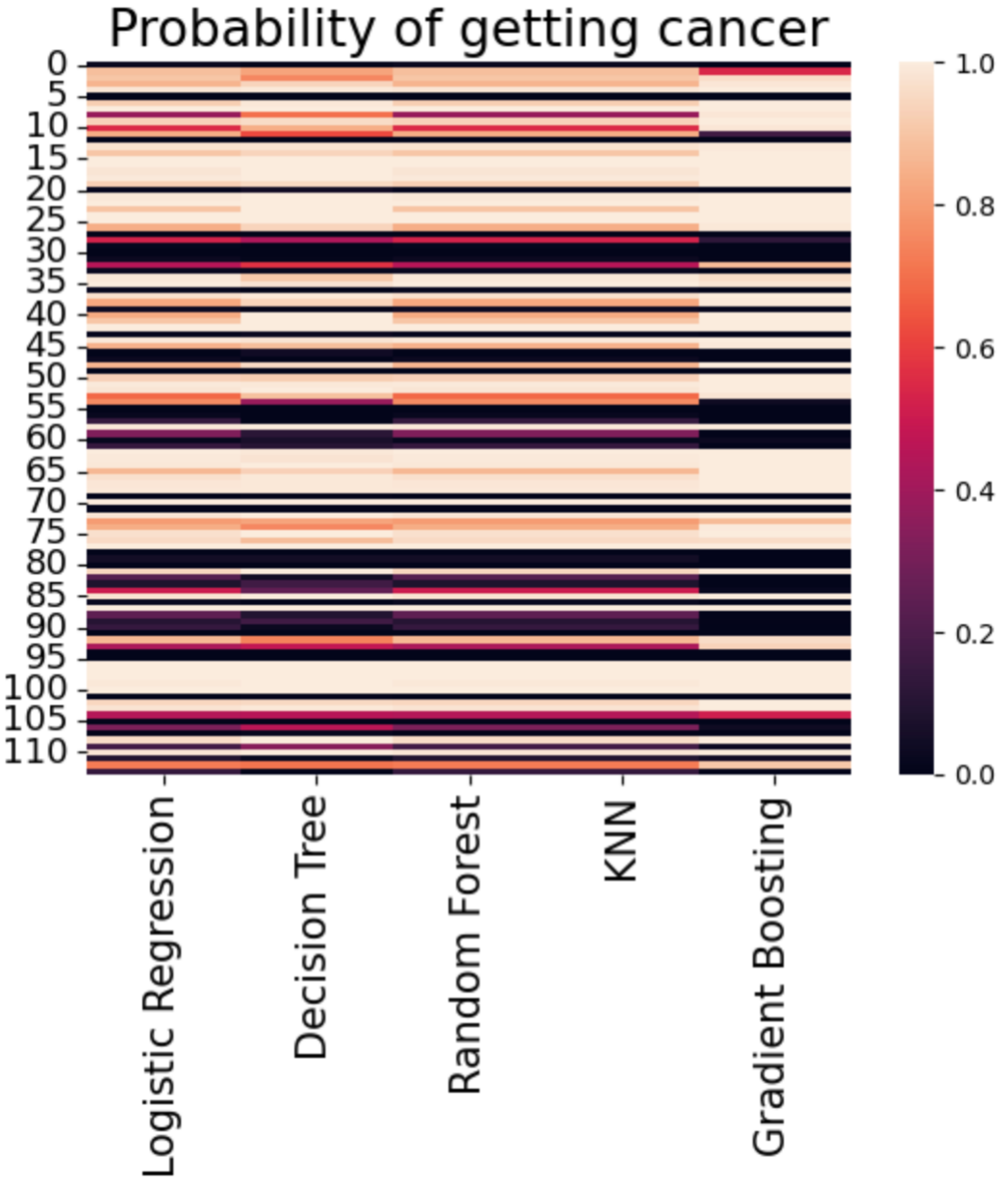

Figure 3. Probability of patients getting cancer.

4. Discussion

By the comparison of accuracy and AUC score, it is clear that the Logistic Regression has the greatest accuracy value but not the AUC, and Decision Tree has a relatively low accuracy but a high AUC. Therefore, maybe the thresholds of the Decision Tree model could be changed to increase its accuracy.

After getting the probabilities of breast cancer by different models, there are several cases of probability that are predicted very differently by the models. The reason is apparent that the different algorithms have different ideas to predict the diagnosis, so those physical features have various weights on prediction. For example, for the 28th patient in the heat map, the machine learning models have different results. Here are several ideas to improve the results to make them more similar. 1) adjust the parameters of each model (training size and testing size). 2) train and predict many times with different

DOI: 10.54254/2755-2721/14/20230796

random seeds. 3) take the median results to stand for the final probability. 4) compare and get the conclusion.

The limitation of machine learning is mainly the disadvantages of each model. The method to prevent this hold is to combine and train several models.

5. Conclusion

The machine learning methods have different kinds of ideas to train and fit the data, so the predicted results will also vary. According to the testing data comparison, the model with the greatest accuracy is the random forest. This means that the random forest can best fit this data set, so it can be used to predict the patients' diagnosis through those given physical features. Reversely, the decision tree has the lowest value of accuracy. By the train of thought of decision tree method, it didn't construct the model comprehensively, so its relatively low accuracy is understandable. Furthermore, many other excluded physical features may cause breast cancer, so this limitation will also affect the persuasiveness of prediction results. Considering the defect of the dataset, the ROC curve and AUC score are also quoted to judge the model performance. Regarding the AUC score, the decision tree has the most outstanding performance. To the advantage of the AUC score, this value will only be affected by the proportion of well-predicted results, so there are also some models with the same AUC score. The random forest performs best when evaluation values and AUC score results are combined. After model performance are judged by several key values, these model can start to find the probability of cancer. Although each model has different predicted probabilities of breast cancer, there can be a threshold to decide the final diagnosis for patients. This set of modeling and predicting can help scientists broaden the application of machine learning in medical research, especially on rare diseases. Additionally, this kind of search idea can be used to find effective treatments. Looking forward to detecting and controlling more and more diseases in the future.

References

[1] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[2] Gladence, L. M., Karthi, M., & Anu, V. M. (2015). A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN Journal of Engineering and Applied Sciences, 10(14), 5947-5953.

[3] Mathan, K., Kumar, P. M., Panchatcharam, P., Manogaran, G., & Varadharajan, R. (2018). A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design automation for embedded systems, 22, 225-242.

[4] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[5] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11), 1-7.

[6] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[7] Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 chi conference on human factors in computing systems, 1-12.

[8] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning, 233-240.

[9] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.

[10] Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?. Emergency Medicine Journal, 34(6), 357-359.