Breast Cancer Staging Prediction Based on Logistic Regression Model

Tong Wu *

University of Washington - Seattle, Seattle, 98195, United State

* Corresponding Author Email: twu29@uw.edu

Abstract. Breast cancer is one of the most common diseases in women. It is important to diagnose whether a patient's breast cancer is benign or malignant as early as possible because there are different treatments for different stages. If the spread of cancer and tumors can be controlled early, patient's suffering can be reduced and survival rates improved. In order to determine if a patient has benign or malignant breast cancer, this article will develop a breast cancer staging prediction model using a popular machine learning approach called logistic regression. The Wisconsin Breast Cancer Diagnosis (WBCD) dataset, provided by the University of Irvine Machine Learning, is the basis for the logistic regression model used in this study. A confusion matrix is utilized to assess the model's accuracy as well as Type I and Type II errors. The Type I and Type II errors' percentages are very small, the accuracy of the logistic regression model is 94.958%, which is not very high thus it may not be recommended for people to use and rely on this model to predict breast cancer, people can consider other prediction models. Furthermore, there may be several factors which may lead to the low logistic regression model's accuracy, such as the size of the database, selection of samples, or the selection of variables.

Keywords: Breast cancer; logistic regression; confusion matrix; prediction model.

1. Introduction

The three primary components of a breast are connective tissue, which surrounds and holds everything together, ducts, and lobules [1]. The disease known as breast cancer is brought on by aberrant development of breast cells. Breast cancer can take many different forms, and the type that develops depends on which area of the breast develops cancer. The most common types of breast cancer are invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). IDC's cancer cells start in the ducts, and the ILC's cancer cells start in the lobules. If the growth and spread of breast cancer cells are not detected and controlled early, they may spread throughout the body and eventually lead to death [1, 2]. Breast cancer often has a variety of symptoms, such as changes in the shape, or appearance of the breast, abnormal or bloody nipple fluid, changes in the appearance of the nipple or the skin around it and thickening or lumps in the breast [2]. The World Health Organization (WHO) reports that 2.3 million women globally received a breast cancer diagnosis in 2020, and 685,000 of them lost their lives to the disease. Breast cancer is one of the most common diseases in the world now. According to WHO statistics, as of the end of 2020, a total of 7.8 million women had been diagnosed with breast cancer in the past five years [3]. Breast cancer can also occur in men, but it is much less common in men than in women [4].

According to the current study, five risk factors may lead to breast cancer: age, family history, reproductive factors, estrogen, and lifestyle [5]. Based on the Centers for Disease Control and Prevention (CDC), breast magnetic resonance imaging (MRI), diagnostic mammography, breast ultrasonography, and breast biopsy are the four major methods used to detect or diagnose breast cancer [6]. Among these methods, the most representative is biopsy. In this test, tissue, fluid, or cell samples are taken from the breast viewed under a microscope, and analyzed with multiple tests. Microscopic images of cells and tissues can be used to quantify numerical properties including area, perimeter, roughness, and radius. The obtained data is then analyzed and finally combined with various imaging data to estimate the chance of people suffering from malignant tumors [7, 8]. The

earlier breast cancer is diagnosed, the higher the patient's recovery rate and survival rate from treatment.

At the same time, it is important to diagnose whether the breast cancer is benign or malignant as early as possible because early detection and confirmation of whether the breast cancer is benign or malignant can take corresponding clinical treatment as soon as possible to control the spread of cancer cells and the progression of the tumor [9]. There are numerous models and techniques available now to determine if a patient has benign or malignant breast cancer [10]. With the rapid development of machine learning technology, the accuracy of many breast cancer prediction models established through machine learning continues improving, which is good for breast cancer patients. This article will use one of the datasets from University of Irvine Machine Learning Called Breast Cancer Wisconsin Diagnostic to create a model for predicting whether breast cancer is benign or malignant and compare the accuracy between models. The methods that will be used in this article include logistic regression and confusion matrix. The data in the Breast Cancer Wisconsin Diagnostic dataset is calculated from digitized images obtained by fine needle aspiration (FNA) of breast masses, one of the biopsy methods.

2. Methods

2.1. Data Description

The WBCD database is available on the Kaggle platform website, and it was provided by the trusted organization UCI Machine Learning Repository. The dataset contains 569 observations and 32 variables in the .CVS format and does not have any missing values.

2.2. Sample Selection

The analysis in this paper uses only 400 randomly selected observations and 11 variables. Table 1 shows the names of the 11 variables and the meaning of each variable. A new variable called "diagnosis_numeric" was added to the dataset, which contains just two numbers: 1 and 0, with 1 representing malignant and 0 representing benign. Since the original variable named "diagnosis" is not a numeric variable, it is difficult to use R for model analysis. Therefore, this paper created a new variable in the data set to replace this variable, turning the categorical variable into a numeric variable.

Table 1. Variable Explanation

Name of Variables	Definition of the Variables
diagnosis_numeric	diagnosis of breast tissues, 1 represents malignant, 0 represents benign
radius_mean	average of the distances between the center and the outermost points
texture_mean	standard deviation of values in grayscale
perimeter_mean	the core tumor’s average size
area_mean	average area of the tumor
smoothness_mean	average of local variation in radius lengths
compactness_mean	average of perimeter^2 / area - 1.0
concavity_mean	average of severity of concave portions of the contour
concave_point_mean	average for the contour's quantity of concave sections
symmetry_mean	average symmetry of the tumor
fractal_dimension_mean	mean for "coastline approximation" - 1

The Wisconsin Breast Cancer (Diagnosis) dataset this paper used for analysis contains a total of 400 patients, with 250 breast cancer patients diagnosed as benign breast cancer and 150 breast cancer patients diagnosed as malignant breast cancer (Figure 1). There are more patients with benign breast cancer in the dataset.

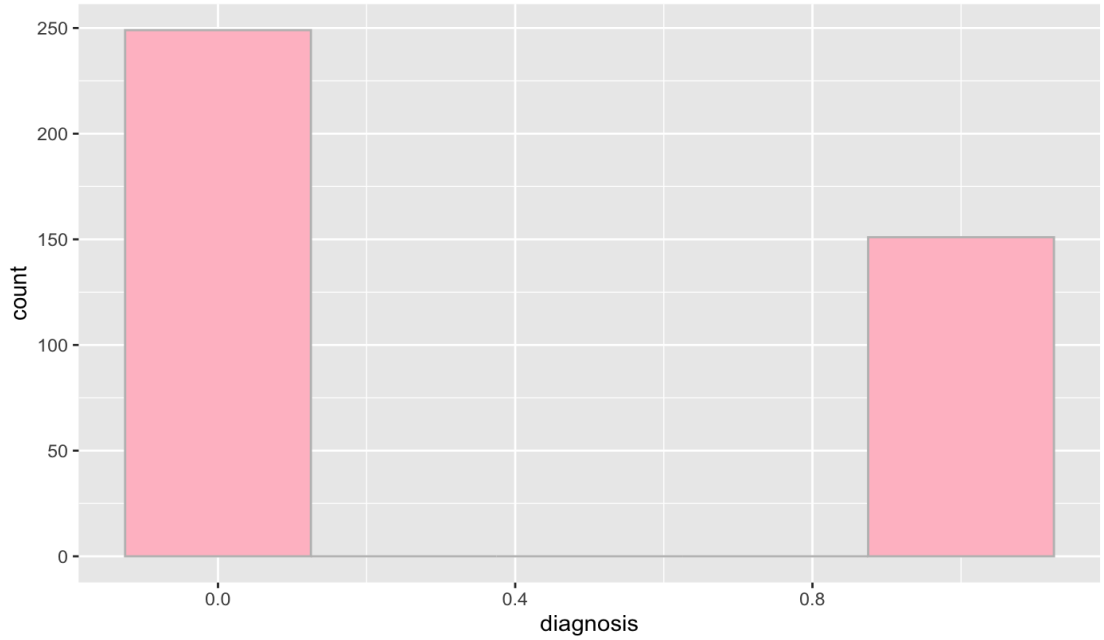

Fig. 1 Distribution of Diagnosis of Breast Tissues (0 represents malignant, 1 represents benign)

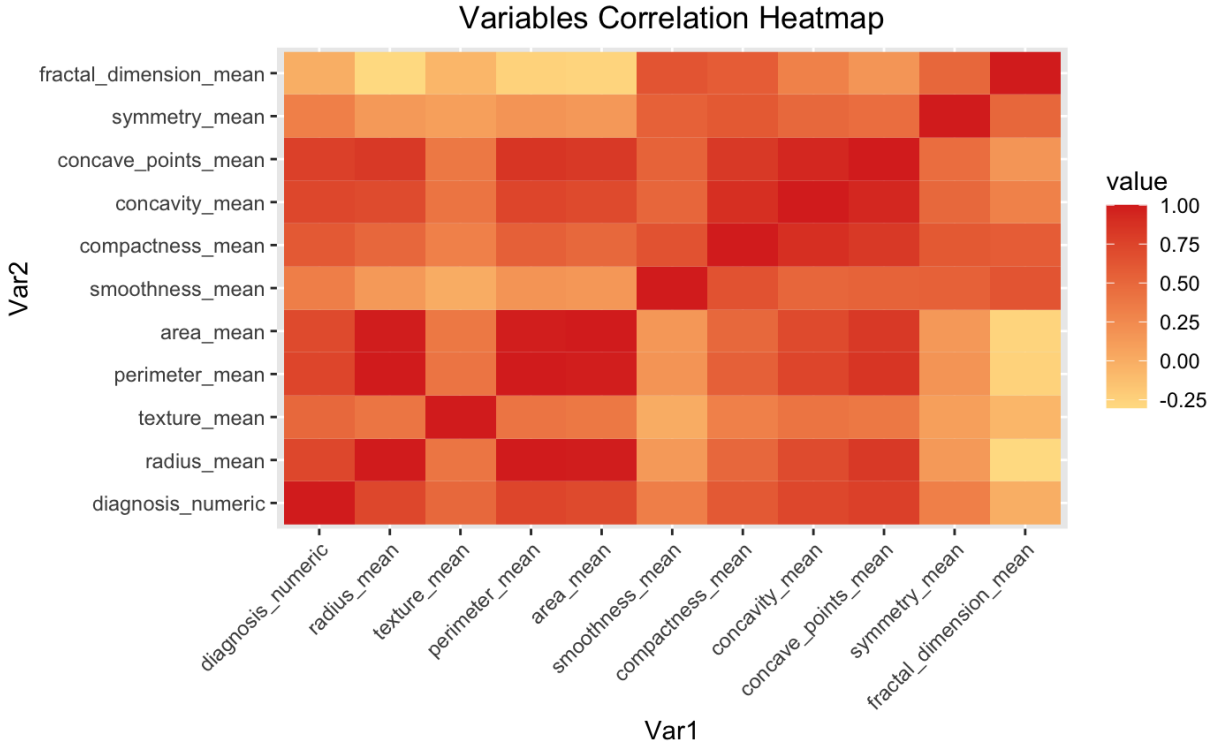

Fig. 2 Variables Correlation Heatmap

Figure 2 shows the correlation between the variables, and the results indicate the existence of multicollinearity among some variables. According to the variable correlation heat map, "radius_mean" is highly correlated with "perimeter_mean", "area_mean", and "concave_point_mean"; "perimeter_mean" is highly correlated with "area_mean" and "concave_point_mean"; "concavity_mean" is highly correlated with "concave_point_mean" and "compactness_mean".

Furthermore, the diagnosis of breast tissue has some correlation with "radius_mean", "perimeter_mean", "area_mean", "compactness_mean", "concavity_mean" and "concave_point_mean". Therefore, only these six factors will be used in this paper's logistic regression model, other variables do not seem to be sufficiently correlated with the diagnosis result, thus it is necessary to put other variables in the logistic regression model.

2.3. Method Introduction

This article is going to use the confusion matrix to analyze the logistic regression model which predicts whether the patients have malignant or benign breast cancer. This paper only selected the 6 variables obtained in Figure 2 that have a certain correlation with breast tissue diagnosis to build a logistic regression model. This paper separated the dataset into a training set and a testing set. The training set contains 70% of the breast cancer patients in the Wisconsin Breast Cancer (Diagnosis) dataset that this paper cleaned and the testing set contains 30% of the observations in the dataset. This paper used confusion matrix to see the false positive (type I error)'s percentage and false negative (type II errors)'s percentage and the accuracy of the logistic regression model.

3. Results and Discussion

Based on the regression coefficient estimates, the following conclusions were drawn: for each additional unit of "radius_mean", the probability of having malignant breast cancer increases by 1.734 and the p-value is 0.9727; for each additional unit of "perimeter_mean", the probability of having malignant breast cancer decreases by 0.518 and the p-value is 0.6691; for each additional unit of "area_mean", the probability of have malignant breast cancer increases by 0.027 and the p-value is 0.0967; for each additional unit of "compactness_mean", the probability of have malignant breast cancer decreases by 4.330 and the p-value is 0.7655; for each additional unit of "concavity_mean", the probability of have malignant breast cancer increases by 25.907 and the p-value is 0.0257; lastly, for each additional unit of "concave_point_mean", the probability of have malignant breast cancer increases by 74.777 and the p-value is 0.0005 (Table 2). Based on Table 2, the mean for the number of concave portions of the contour has the most impact on malignant breast cancer since it has the largest coefficient and the smallest p-value; and the mean of compactness has the smallest coefficient and the largest p-value. The p-value for concave_points_mean indicates that it may be a highly significant predictor of the response variable in the logistic regression model because it is well below the significance level of 0.05 (Table 2).

Table 2. Logistic Regression Information

Term	Coefficient	P-value
radius_mean	1.734	0.6691
perimeter_mean	-0.518	0.3373
area_mean	0.027	0.0967
compactness_mean	-4.330	0.7655
concavity_mean	25.907	0.0257
concave_point_mean	74.777	0.0005

The confusion matrix method was used to analyze the logistic regression model. The accuracy of the logistic regression model in predicting patients with malignant or benign breast cancer was 94.9580%. Figure 3 shows that a major number of breast cancer patients who are predicted by the logistic model diagnosed as benign are benign breast cancer, and breast cancer patients who are

predicted by the logistic model diagnosed as malignant are malignant breast cancer. The percentages of false positive and false negative are low. Because there are 113 people receive correct conclusion (either true positive or true negative), and only 6 people receive false conclusion (3 people receive false positive and other 3 people receive false negative).

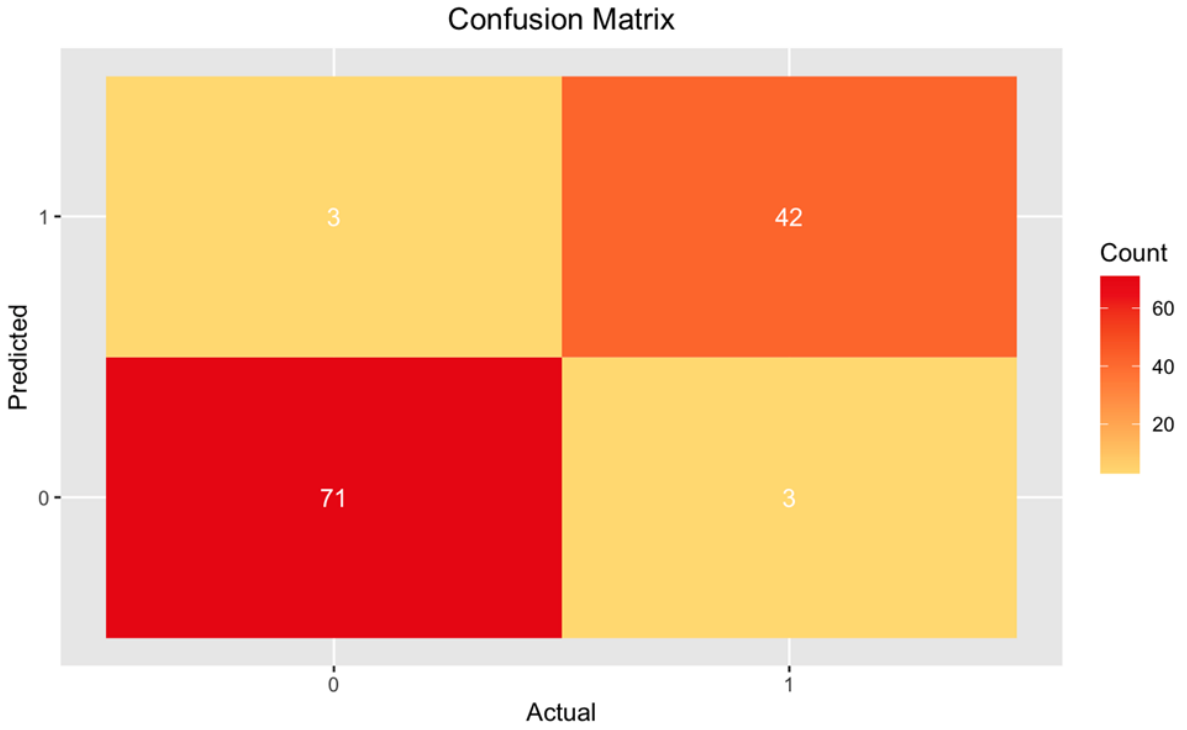

Fig. 3 Variables Correlation Heatmap

4. Conclusion

In conclusion, although the percentages of false positive and false negative are small, the accuracy percentage of the logistic regression model is not high enough, it is only around 94.9580%. Personally speaking, this paper shows the higher the accuracy of a medical disease prediction model, the better. The accuracy of the disease prediction model should be at least 98% or above before it can be truly used in the medical field. Moreover, the logistic regression model may not be the most useful and accurate model for predicting benign or malignant breast cancer. In the future, if people try to choose model to predict the stage of breast cancer, they can consider using other machine learning algorithms, such as random forests, k-nearest neighbors, or decision trees, etc., and then compare the accuracy of the models to find the best breast cancer staging prediction model with has the highest accuracy. In addition, the low accuracy of the logistic regression model may be due to insufficient sample size or variable selection factors. This paper only used 400 samples for testing. The sample size is slightly small, thus, there may be bias. Therefore, future research can consider increasing the dataset or sample size and adjusting the variable selection in the model to see whether the accuracy of the logistic regression model can be improved.

References

[1] Winters Stella, et al. Breast cancer epidemiology, prevention, and screening. Progress in molecular biology and translational science, 2017, 151: 1-32.

[2] Sun Yisheng et al. Risk Factors and Preventions of Breast Cancer. International journal of biological sciences, 2017, 13: 1387-1397.

[3] Ara Sharmin, Annesha Das, Ashim Dey. Malignant and benign breast cancer classification using machine learning algorithms. 2021 International Conference on Artificial Intelligence (ICAI). IEEE, 2021.

[4] Zheng Ying, Wu Chunxiao, Zhang Minlu. Epidemic status and disease characteristics of breast cancer in China. Chinese Journal of Cancer, 2013, 23(8): 561-569.

[5] Zheng Ying, Wu Chunxiao, Wu Fan. Current situation and development trend of breast cancer death in Chinese women. Chinese Journal of Preventive Medicine, 2011, 45(2): 5.

[6] Shen Yaqin, Sun Huimin, Wang Min, et al. Study on the status quo of social restrictions and influencing factors in patients with breast cancer undergoing chemotherapy after surgery. Journal of Nursing, 2023, 38(15): 30-34.

[7] Sun Qianqian, Ye Hongfang, Yang Li. Meta analysis of the intervention effect of acceptance and commitment therapy on breast cancer patients. Chinese Journal of Nursing, 2022, 57(9): 1070-1078.

[8] Wu Guofeng, Li Xinrui, Zhong Meimei, et al. Study on the effect of continuous nursing based on cloud platform on subthreshold depression of breast cancer patients. Chinese Journal of Nursing, 2024, 59(2): 142-148.

[9] Yang Ling, Li Liandi, Chen Yude, et al. Estimation and prediction of the incidence and death trend of breast cancer in China. Chinese Journal of Cancer, 2006, 28(6): 3.

[10] Zheng Ying, Li Delu, Xiang Yongmei, et al. Analysis of the epidemic situation and trend of breast cancer in Shanghai urban area. Surgical Theory and Practice, 2001, 6 (4): 3.