Logistic Regression Model for Predicting the Malignancy of Breast Cancer

Muneeba Ahmed[1]

[1]*Systems Engineer, Infosys Limited, India*

Abstract - *In today's modern environment, recognising breast cancer is critical. Breast cancer is one of the most serious tumours that can affect women, and it can be fatal. Breast cancer is classified into two types: benign (non-cancerous) and malignant (cancerous). Machine learning is the process through which a machine learns increasingly on its own. The ML model is a mathematical technique used in artificial intelligence. A computer that thinks for itself and mimics human intelligence is referred to as artificial intelligence. Just like a human, the computer improves at its work as it gets "experience." There are several Machine Learning approaches available for analysing breast cancer data. This paper describes a Machine Learning model for diagnosing breast cancer. Logistic Regression model is used for detecting breast cancer. This algorithm falls under the category of supervised machine learning.*

Key Words: **Breast Cancer, Artificial Intelligence, Machine Learning, Logistic Regression**

1. INTRODUCTION

Breast cancer refers to the uncontrolled cell development in the breast. Both men and women can get breast cancer, but women are more likely to have it. Breast cancer has been one of the main causes of female mortality when compared to other malignancies. Breast cancer symptoms include changes in the breast's size and form, the thickness of the tissue around the breast, as well as crust, scales, and redness of the skin. Changes in environmental variables, hormones, and lifestyle lead to breast cancer, which raises the risk factor. The lymphatic vessels allow lymphatic fluid from the breast to pass through. If the breast contains cancerous cells, they go into the lymphatic vessels and start to multiply in the lymph nodes. Although many breast cancer patients have no symptoms at all, breast cancer is typically discovered after the beginning of symptoms. To prevent mortality, early detection of breast cancer is crucial. For the ability to detect breast cancer in its early stages, earlier therapy is required. A reliable and efficient diagnostic method that enables clinicians to differentiate between benign and malignant breast tumours is required for early identification. For the current medical issue, the automated identification of breast cancer is significant. It is crucial to create an efficient and reliable diagnostic strategy. Clinical applications face a major problem with clinical diagnosis. Breast cancer data classifications can be used to predict the outcomes of specific diseases and to determine the genetic activity of tumours [1]. Numerous methods for estimating breast cancer have been identified in the last year. During biopsy screening, the breast tissues are used for the biopsies. Although the testing yields more trustworthy results, the method for collecting breast biopsies is incredibly painful and pitiable [2]. The majority of patients are not interested in this testing as a result. Since mammography produces 2D projection images of the breast, it is the most widely used method for estimating breast cancer. The two most frequently utilised mammogram techniques are digital mammography and screen-film mammography [3]. Screenfilm mammography is used on female breasts that are asymptomatic. It takes roughly 20 minutes to do a traditional mammogram. Benign cancer cannot be found with this method. Digital mammography offers a solution to the screening mammography problem. It is connected to a computing equipment since a computer is where digital mammography data is saved. Digital mammography uses image processing techniques to enhance the quality of the images that are recorded. Digital mammography performs better for incorrectly diagnosed samples. Magnetic resonance imaging, another common technique, is primarily used to find breast cancer [4]. The MRI is a challenging procedure. Additionally, certain malignancies that mammography would have detected could be missed. In women who have been given a breast cancer diagnosis, MRI is used to measure the breast's actual size and spot numerous disorders in the breast.

In the past year, machine learning techniques have been used more and more in prediction, especially in the field of medicine [5]. It gives systems the ability to learn from the past in order to extrapolate intricate insights from massive data sets. In a variety of clinical settings, these methods are most frequently employed to identify and classify malignancies. In order to diagnose and cure breast cancer, machine learning has been used first and foremost [6].

2. RELATED WORK

This section discusses some of the related research on machine learning-based breast cancer diagnosis that has been conducted in the past.

S.Vasundhara, B.V. Kiranmayee, and Chalumuru Suresh [7] proposed employing several machine learning methods to classify mammography pictures as benign, malignant, or normal. A comparison of Support Vector Machines, Convolutional Neural Networks, and Random Forest is performed. According to the simulation results, CNN is the best classifier since it produces instinctive classification of digital mammograms utilising filtering and morphological processes.

Arpita Joshi and Dr.Ashish Mehta [8] compared the classification results using KNN, SVM, Random Forest, and Decision Tree (Recursive Partitioning and Conditional Inference Tree). The dataset utilised was the Wisconsin Breast Cancer dataset from the UCI repository. According to simulation findings, KNN was the top classifier, followed by SVM, Random Forest, and Decision Tree.

Muhammet Fatih Ak [9] used Dr. William H. Walberg's dataset from the University of Wisconsin Hospital. This dataset was subjected to data visualisation and machine learning techniques such as logistic regression, k-nearest neighbours, support vector machine, naive Bayes, decision tree, random forest, and rotation forest. R, Minitab, and Python were chosen to be used for machine learning and visualisation. All of the procedures were subjected to a comparative study. The logistic regression model with all features included produced the greatest classification accuracy (98.1%), and the proposed approach exhibited an improvement in accuracy results.

Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel [10] compared the performance of four machine learning algorithms on the Wisconsin Breast Cancer (original) dataset: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbors (k-NN). According to the experimental data, SVM has the best accuracy (97.13%) and the lowest error rate. All experiments are carried out in a simulation environment using the WEKA data mining tool.

3. METHODOLOGY

3.1 Data Collection

The dataset is taken from Kaggle. It consists of 569 rows and 33 columns, the first of which is the ID number and the second of which is the diagnosis outcome (0-benign and 1-malignant). The other columns describe the shape and size of the nucleus of the target cancer cell. In a biopsy test, a sample of cells is obtained from the breast using the Fine Needle Aspiration (FNA) process. These characteristics are determined for each cell nucleus by examining it under a microscope in a pathology laboratory.

Table -1: Description of features of the dataset

Feature Name	Feature Description
Radius	Average of distance from center to circumference points
Texture	Standard deviation of gray scale vlaue
Perimeter	Gross distance between the snake points
Area	Total number of pixels on the inside of the snake along with one half of the pixels in the circumference
Smoothness	Local variance in length of radius, quantified by calculating the length difference
Compactness	Perimeter ^2/ Area
Concavity	Intensity of the contour concave points
Concave points	The number of contour concavities
Symmetry	The difference in the length between lines perpendicular to the major axis in both directions to cell boundary
Fractal Dimension	Coastline estimation. A higher value leads to a less normal contour representing a higher risk of malignancy.

3.2 Data Pre-processing

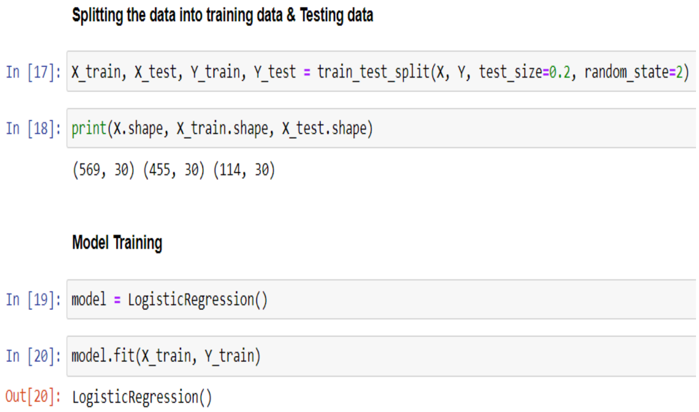
Data pre-processing is the initial step that starts the machine learning process while developing a model. Real-world data frequently lacks particular attribute values or trends and is frequently inconsistent, erroneous (contains mistakes and outliers), incomplete, and inconsistent. This is where data pre-processing enters the picture; it aids in calming, organising, and formatting the raw data so that machine learning models can use it immediately

3.3 Logistic Regression

The generated linear regression hyperplane cannot be utilised to predict the dependent variable in linear regression using the independent variable alone. Logistic regression is therefore employed when there are categorical data. Instead of forecasting anything continuous, logistic regression forecasts whether something is true or untrue. It serves as a classification tool. The dependent variable's independent variable is transformed using the sigmoid function into a probability expression with a range of 0 to 1. It is a well-liked Machine Learning technique due to its capacity to offer probabilities and categorize fresh samples using continuous and discrete data.

3.4 Training and Testing the Model

The issue of overfitting usually occurs during model training. When a model performs incredibly well on the data we used to train it but struggles to generalize successfully to new, unexpected data points, a problem has arisen. The model performs poorly even when tested on the training set of data, which is under fitting. Creating several data samples for the model's training and testing phases is the most popular method for locating these kinds of problems. After the analysis, we will use 80% of the data to train the machine. Using data to guide the machine is referred to as "training" in this context. The remaining 20% of the data points are used to test the machine's performance after the first 80% of the data were used to train it. To put it another way, we can measure the amount of specialized process knowledge the machine picked up.



Splitting the data into training data & Testing data

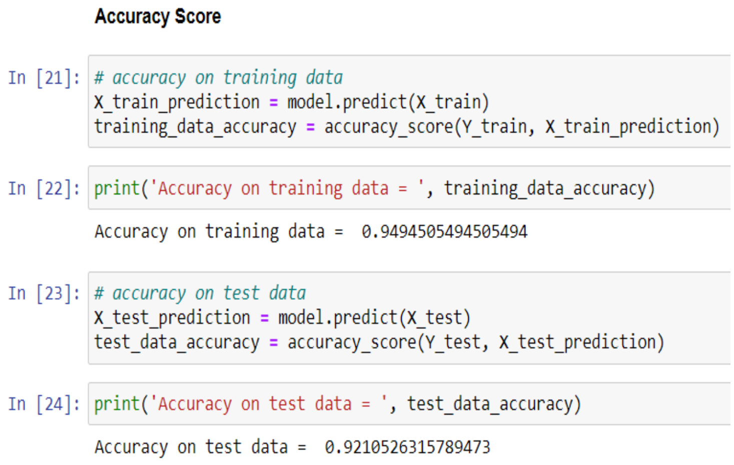
```
In [17]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

In [18]: print(X.shape, X_train.shape, X_test.shape)

         (569, 30) (455, 30) (114, 30)
```

Model Training

```
In [19]: model = LogisticRegression()

In [20]: model.fit(X_train, Y_train)

Out[20]: LogisticRegression()
```

Fig -1: Training and Testing the model

3.5 Model Evaluation



Accuracy Score

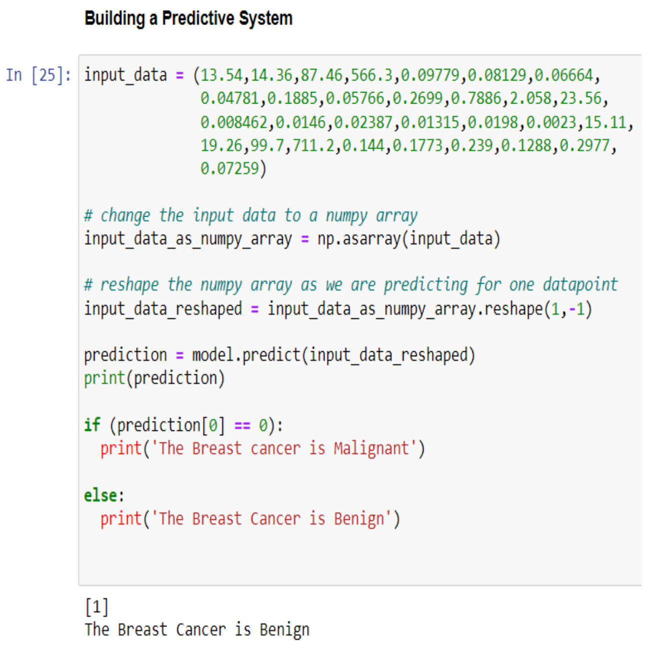
```
In [21]: # accuracy on training data
         X_train_prediction = model.predict(X_train)
         training_data_accuracy = accuracy_score(Y_train, X_train_prediction)

In [22]: print('Accuracy on training data = ', training_data_accuracy)

         Accuracy on training data =  0.9494505494505494

In [23]: # accuracy on test data
         X_test_prediction = model.predict(X_test)
         test_data_accuracy = accuracy_score(Y_test, X_test_prediction)

In [24]: print('Accuracy on test data = ', test_data_accuracy)

         Accuracy on test data =  0.9210526315789473
```

Fig -2: Checking the accuracy score of the model

The model's accuracy was found to be 94.94% on training data and 92.10% on testing data.

3.6 Prediction

After Machine learning model is fit, the model can predict whether the patient has Malignant type of tumour that is patient is suffering from cancer or Benign type of tumour that is patient does not have cancer.



Building a Predictive System

```
In [25]: input_data = (13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,
                       0.04781,0.1885,0.05766,0.2699,0.7886,2.058,23.56,
                       0.008462,0.0146,0.02387,0.01315,0.0198,0.0023,15.11,
                       19.26,99.7,711.2,0.144,0.1773,0.239,0.1288,0.2977,
                       0.07259)

         # change the input data to a numpy array
         input_data_as_numpy_array = np.asarray(input_data)

         # reshape the numpy array as we are predicting for one datapoint
         input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

         prediction = model.predict(input_data_reshaped)
         print(prediction)

         if (prediction[0] == 0):
           print('The Breast cancer is Malignant')

         else:
           print('The Breast Cancer is Benign')

         [1]
         The Breast Cancer is Benign
```

Fig -3: Building a predictive system

Hence, a Logistic Regression Model was implemented to predict the malignancy of breast cancer.

4. CONCLUSION

Breast cancer is one of the most deadly diseases affecting women today. It is the most common cause of death in women. As a result, early identification of this cancer could save many lives. The effort in this study is to create a classification with the purpose of detecting breast cancer in its early stages. We examined Logistic Regression model for breast cancer detection.

REFERENCES

[1] Joshi, R. Doshi, and J. Patel, "Diagnosis and prognosis breast cancer using classification rules," International Journal of Engineering Research and General Science, vol. 2, no. 6, pp. 315–323, 2014.

[2] A. M. Ahmad, G. M. Khan, S. A. Mahmud, and J. F. Miller, "Breast cancer detection using Cartesian genetic programming evolved artificial neural networks," in Proceedings of the 14th annual conference on Genetic and evolutionary computation, pp. 1031–1038, Philadelphia Pennsylvania, USA, 2012.

[3] A. T. Azar and S. A. El-Said, "Probabilistic neural network for breast cancer classification," Neural Computing and Applications, vol. 23, no. 6, pp. 1737–1751, 2013.

[4] E. Warner, H. Messersmith, P. Causer, A. Eisen, R. Shumak, and D. Plewes, "Systematic review: using magnetic resonance imaging to screen women at high

risk for breast cancer," Annals of Internal Medicine, vol. 148, no. 9, pp. 671–679, 2008.

[5] G. R. Kumar, G. Ramachandra, and K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques," International Journal of Innovations in Engineering and Technology (IJIET), vol. 2, no. 4, p. 139, 2013.

[6] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, Sage Journals, vol. 2, 2006.

[7] S.Vasundhara , B.V. Kiranmayee and Chalumuru Suresh "Machine Learning Approach for Breast Cancer Prediction", International Journal of Recent Technology and Engineering, 2019.

[8] Arpita Joshi and Dr. Ashish Mehta "Comparative Analysis of Various Machine Leaning Techniques for Diagnosis of Breast Cancer," International Journal on Emerging Technologies, 2017.

[9] Muhammed Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications", Healthcare, MDPI, 2020.

[10] Hiba Asria, Hajar Mousannifb, Hasan Al Moatassime, Thomas Noeld "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science, Elsevier, 2016.