Section A-Research paper
ISSN 2063-5346

A MACHINE LEARNING ALGORITHM'S LOGISTIC REGRESSION FOR BREAST CANCER DIAGNOSIS

Dr.K.Gomathi[1] Dr.S.Hemalatha[2] Dr.G.Manivasagam[3]

[1]Assistant Professor, Department of ICT & Cognitive Systems, Sri Krishna Arts and Science College, Coimbatore

[2]Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore

[3]Associate Professor, Department of CSA, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Vijayawada

Abstract

Breast cancer is the leading cause of death among women. It has become a common ailment with a rapid increase in occurrence. Earlier detection is the most efficient way to control breast cancer consequences. The healthcare industry benefits from machine learning because it can help make sense of the massive amounts of healthcare data generated daily within electronic health records. The process of breast cancer early prediction can be greatly aided by machine learning techniques, which are now a popular area of study due to their effectiveness. The risk prediction of breast tumours using algorithms for machine learning has been the subject of numerous studies. The purpose of this study is to use logistic regression to identify a woman's risk for breast cancer.

Keywords: Machine Learning, Logistic Function, Accuracy

1. Introduction

Breast cancer is a common and dangerous disease for women. Women will be diagnosed with more than 2.26 million new cases of breast cancer in 2020 [1]. Breast cancer treatment is very effective, 90% or higher survival rate achieved when the disease is detected early. Reduced mortality rates are only possible with early detection of breast cancer [2]. The ability to find breast cancer in its earlier stages is necessary for earlier treatment, though. Early detection requires an accurate diagnosis process that enables doctors to differentiate between benign and malignant breast tumours.

In recent years, machine learning methods have become increasingly popular in prediction, particularly in medical diagnosis. One of the most difficult problems in medical applications is a medical diagnosis. Breast cancer data classification can help predict the outcome of some ailments or discover the hereditary behaviour of tumours. As a result, classifier systems are increasingly being used in medical diagnosis. So, the study's goal is to use logistic regression to determine the most important breast cancer risk factors and to estimate the overall risk.

2. Background study

The logistic analysis dataset is available on the Kaggle website (https://www.kaggle.com). This study aims

to categorize tumours into benign (non-cancerous) or malignant (cancerous) tumours and to analyze the classification of these tumours using machine learning. The Breast Cancer dataset consists of 569 records and 32 attributes [3]. Jupyter Lab, a more powerful and flexible data science application software, is used for the analysis of data.

3. Machine Learning

Many different industries, including the healthcare sector, use machine learning extensively. The technology includes artificial intelligence (AI) software to enable systems, automatically learn from experience, and get better without explicit programming [4].

In addition, machine learning refers to the use of algorithms to analyze data, learn from it, and then predict or determine something regarding the outside world [5].

Machine learning is frequently used to solve two main types of issues: regression and classification. Regression algorithms are typically used with numerical data, and binary and multi-category classification problems are included. Furthermore, machine learning algorithms are classified into two types: supervised learning and unsupervised learning [6]. Unsupervised learning is used to infer natural structures within a dataset, whereas supervised learning uses predefined labels in output values.

4. Literature Review

Shakkeera L, Rahul Raj Pandey et al[7] proposed the classification and model of prediction with accuracy. This paper shows XGBoost has higher accuracy compared with other machine learning algorithms to detect the early stage of the tumour.

Gomathi K et al[8] Weka tool to access data mining algorithms efficiently and more easily to predict breast cancer. This shows that Naïve Bayes Classifier has better accuracy compared to other algorithms.

Nalini C, Meera D, et al[9] worked on the classification techniques Naive Bayes and J48 used to analyze the execution time and performance of accuracy, concluding that Naïve Bayes had higher accuracy with a minimum execution time compared to J48.

Ravi Kumar G, Ramachandra G A, Nagamani K, et al. focused on different data mining techniques to predict breast cancer using Naïve Bayes, Decision Tree, J48, Logistic regression, KNN, and SVM, and they compared predictive accuracy. SVM had the highest accuracy compared to the other algorithm [10].

Ganjar Alfian, Muhammad Syafrudin, et al[11] proposed the Extra tree model with SVM and also with other methods. An extra tree model which used for feature selection, finally the result shows that the Machine learning algorithm has improved the performance metrics of accuracy

Elham Bahmani, Mojtaba Jamshidi, Abdusalam Abdulla Shaltooki et al[12] MATLAB used and constructed the proposed model that Naïve Bayes combined with K-means clustering and RBF. It is used to detect the tumour based on performance metrics.

Shahan Yamin Siddiqui, Iftikhar Naseer et al[13] It has been suggested that the CF-BCP model has a great deal of potential for diagnosing various forms of breast cancer. The CF-BCP model achieves 97.41% accuracy in multimodal medical imaging

Section A-Research paper
ISSN 2063-5346

fusion in detecting breast cancer phases, and 97.97% accuracy in identifying different types of breast cancer following decision-based fusion enabled by fuzzy logic.

5. Logistic Regression

Logistic regression is a machine learning algorithm that predicts the likelihood of classes based on a set of dependent variables. This model computes the logistic of the result by adding the input features [14]. The technique's name, logistic regression, was inspired by the logistic function, the method's main element. Statistics experts created the logistic function known as the sigmoid function to describe the ecological characteristic of population growth, which rises quickly and peaks at the environment's carrying capacity. This S-shaped curve can be used to convert any real-valued number into a value between 0 and 1, but never precisely between values [15].

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

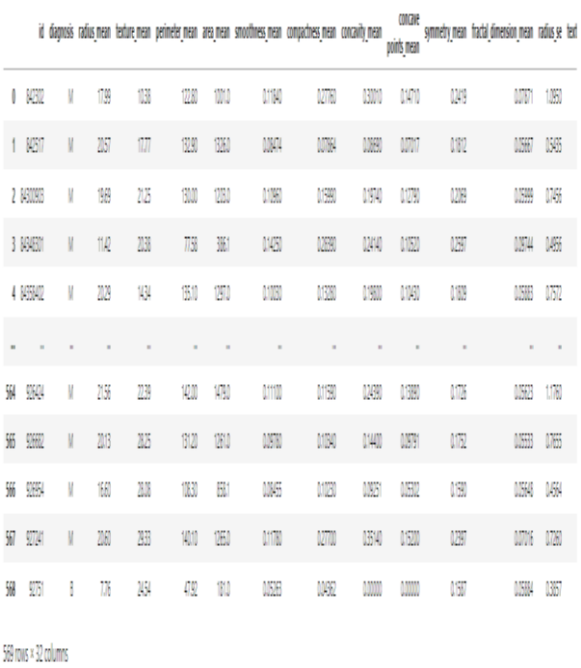
Sigmoid Function

In the formula, z is the function's input, e is the base of the natural log, and () is the output between 0 and 1 (probability estimate)[9].

6. Methodologies

The procedures that were used to create the machine learning logistic regression model.

1. The Kaggle website can be used to collect data.

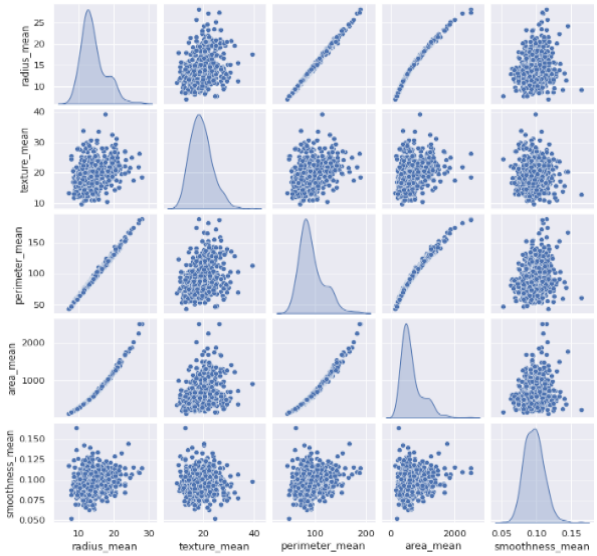


id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	text
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	1.0950
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	0.5435
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	0.7456
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	0.4956
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	0.7572
...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	1.1760
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	0.7655
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	0.4564
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	0.7260
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	0.3857

568 rows × 32 columns

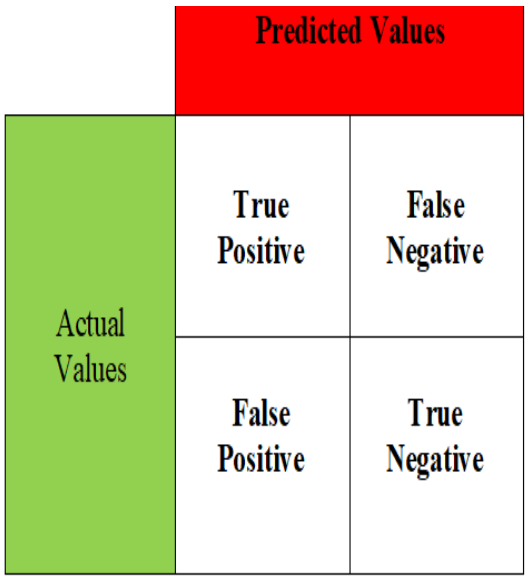
2. Pre-processing of data is necessary to construct a more precise ML model. The cleaning process of data is called data pre-processing. In this, missing, noisy, and inconsistent data are identified [16].

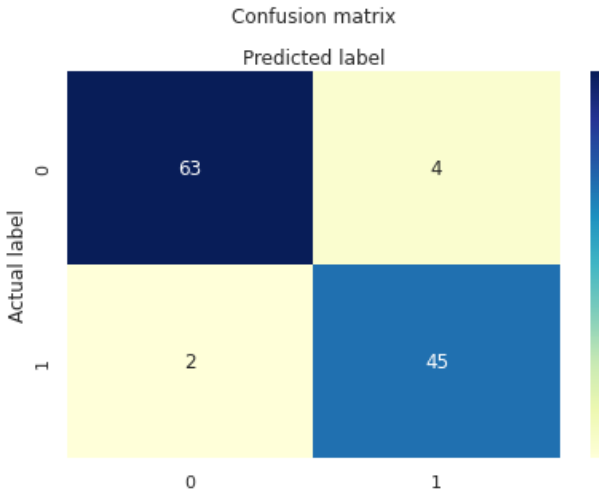
3. Logistic Regression Algorithm, implemented, is primarily used for prediction and determining success probability.



Confusion Matrix

A classification model's accuracy can be determined using a variety of metrics, but the confusion matrix is among the most straightforward. This is the most used method for evaluating logistic regression.





7. Result and Discussion

Accuracy, precision, recall, and F-score metrics can be calculated. The performance of machine learning models is critical because it allows us to understand the benefits and drawbacks of these models when making predictions in novel situations [17].

Accuracy

Accuracy is the ratio of true positives and true negatives to all positive and negative observations as a performance metric for machine learning classification models.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision

The calculation of the precision value is based on the feature classification of true positive and false positive prediction,

$$Precision = \frac{TP}{TP + FP}$$

Recall

The true positive and false negative feature classification is to classify recall value. It's stated,

$$Recall = \frac{TP}{TP + FN}$$

F-score

Precision and Recall value to be used to calculate F-score. It is expressed as:

$$F\ score\ = \frac{2 * Precision * Recall}{Precision + Recall}$$

Recall	94%
F1-score	95%

Accuracy	**97.81%**
Precision	97%

8. Conclusion

Breast cancer is the most serious infection that women face today. For women, it is the leading cause of death. So early prediction is very important to increase the survival rate of life. In recent days, Machine learning algorithms have mainly been involved in the healthcare industry for earlier prediction of diseases. The focus of this work is to classify benign and malignant, and Accuracy, Precision, Recall, and F1-score performance metrics can be calculated that are evaluated based on logistic regression.

References:

1. https://www.wcrf.org/cancer-trends/breast-cancer-statistics/
2. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
3. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
4. Das, S., Dey, A., Pal, A., & Roy, N, 2015. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. International Journal of Computer Applications, 115(9), 31-41
5. Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S.,2019. Applications of Artificial Intelligence in Transport: An Overview, Sustainability, 11(1), 189.
6. Sathya, R & Abraham, 2013.A Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.
7. Shakkeera L, Rahul Raj Pandey, Rahul Bhardwaj, Sidhya Virya Singh, Siddhartha S. Mukherjee, 2020. Analysis and Prediction of Breast Cancer using Machine Learning Techniques, International Journal of Engineering and Advanced Technology (IJEAT), Volume-10 Issue-2.
8. Gomathi K, 2012. An empirical study on breast cancer using data mining techniques, International Journal of Research in Computer Application and Management, Vol 2, Issue 7, ISSN 2231-1009.
9. Nalini C, Meera R, 2018. Breast cancer prediction system using Data mining methods, International Journal of Pure and Applied Mathematics, Volume 119 No. 12, 10901-10911.
10. Ravikumar G, Ramachandra G A, Nagamai, 2013.An Efficient Prediction of Breast Cancer Data using Data Mining Techniques, International Journal of Innovations in Engineering and Technology (IJIET), Vol: 2, Issue 4.
11. Ganjar Alfian, Muhammad Syafrudin, Imam Fahrurrozi, Norma Latif Fitriyani, Fransiskus Tatas Dwi Atmaji, Tri Widodo, Nurul Bahiyah, Filip Benes, Jongtae Rhee,

2022.Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method, https://doi.org/10.3390/computers11090136.

12. Elham Bahmani, Mojtaba Jamshidi, Abdusalam Abdulla Shaltooki, 2019. Breast Cancer Prediction Using a Hybrid Data Mining Model, International Journal on Informatics Visualization, VOL 3 NO 4.
13. Shahan Yamin Siddiqui, Iftikhar Naseer, Muhammad Adnan Khan, Muhammad Faheem Mushtaq, Rizwan Ali Naqvi, Dildar Hussain, Amir Haider, 2021. Intelligent Breast Cancer Prediction Empowered with Fusion and Deep Learning, DOI:10.32604/cmc.2021.013952
14. https://www.kdnuggets.com/2022/07/logistic-regression-work.html.
15. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
16. Thanuja Nishadi A S,2019. Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab, International Journal of Advanced Research and Publications, ISSN: 2456-9992.
17. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/