WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predicting Breast Cancer Using Logistic Regression: A Machine Learning Approach

Abhay Singh Rawat, Vidit

Students, Dept. of Computer Science, Maharaja Surajmal Institute, Janakpuri, New Delhi
viditkumar474@gmail.com, abhayrawat26jun@gmail.com

ABSTRACT:-

This research paper explores the use of logistic regression, a machine learning algorithm, in predicting breast cancer. The study utilizes a publicly available breast cancer dataset to train and evaluate the logistic regression model, highlighting its effectiveness and accuracy. The results demonstrate the potential of logistic regression in clinical settings for early detection and improved patient outcomes.

Keywords: Breast Cancer, Machine Learning, Logistic Regression, Prediction Model, Early Detection

1. Introduction

Background: Breast cancer is a significant health concern worldwide, and early detection is crucial for improving survival rates. Traditional diagnostic methods are often time-consuming and expensive.

Objective: This study aims to develop and evaluate a logistic regression model for predicting breast cancer, providing a cost-effective and accurate diagnostic tool.

Thesis Statement: Logistic regression, a widely-used machine learning algorithm, can effectively predict breast cancer using clinical and demographic data.

There are different types of breast cancer which occurs when affected cells and tissues spread throughout the body:

1. **Ductal Carcinoma in Situ (DCIS)** is type of the breast cancer that occurs when abnormal cells spread outside the breast it is also known as the non-invasive cancer
2. **Ductal Carcinoma (IDC)** and it is also known as infiltrative ductal carcinoma. This type of the cancer occurs when the abnormal cells of breast spread over all the breast tissues and IDC cancer is usually found in men
3. **Mixed Tumors Breast Cancer (MTBC)** is the third type of breast cancer and it is also known as invasive mammary breast cancer. Abnormal duct cell and lobular cell causes such kind of cancer
4. **Lobular Breast Cancer (LBC)** which occurs inside the lobule. It increases the chances of other invasive cancers.
5. **Mucinous Breast Cancer (MBC)** is the fifth type that occurs because of invasive ductal cells, it is also known as colloid breast cancer. It occurs when the abnormal tissues spread around the duct.
6. **Inflammatory Breast Cancer (IBC)** is last type that causes swelling and reddening of breast. It is a fast growing breast cancer, when the lymph vessels block in break cell, this type of cancer starts to appear.

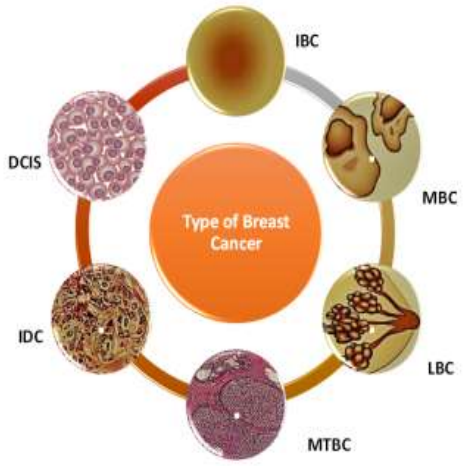


Fig: Demonstration of major types of Breast Cancer

2. Literature Review

Current Methods:

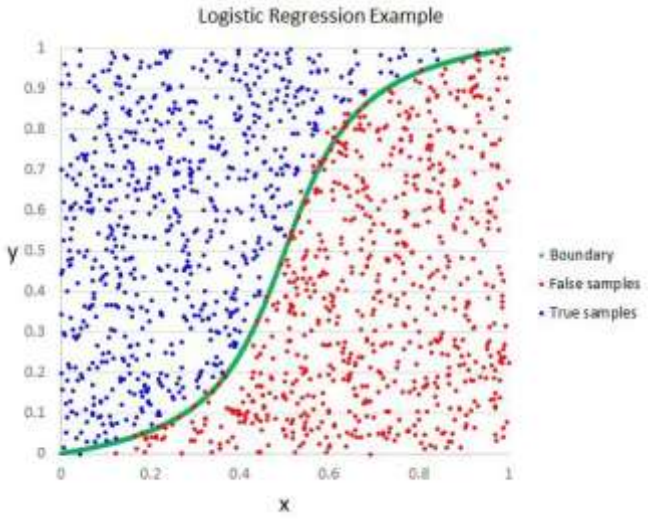
Review of current breast cancer prediction methods, including mammography, genetic testing, and other machine learning approaches.

Traditional Diagnostic Methods

Traditional diagnostic methods for breast cancer include mammography, ultrasound, magnetic resonance imaging (MRI), and biopsy. Mammography, the most common screening tool, has limitations such as false positives and negatives, radiation exposure, and reduced sensitivity in dense breast tissue. Ultrasound and MRI provide additional imaging options but come with higher costs and limited accessibility. Biopsy, considered the gold standard, is invasive and not suitable for routine screening.

Logistic Regression in Medical Predictions

Logistic regression is a statistical method commonly used for binary classification problems. It models the probability of a binary outcome (such as benign or malignant) based on one or more predictor variables. Logistic regression is valued for its simplicity, interpretability, and effectiveness in many applications, making it a popular choice in medical research.



Data Sources for Breast Cancer Prediction

Various datasets have been utilized for developing breast cancer prediction models, including:

- **Wisconsin Breast Cancer Dataset**: One of the most widely used datasets, containing features computed from digitized images of fine needle aspirate (FNA) of breast masses.

- **Breast Cancer Surveillance Consortium (BCSC) Dataset**: Contains detailed mammography data along with patient demographics and clinical histories.
- **National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Database**: Provides comprehensive cancer statistics, including incidence and survival data.

Challenges and Limitations in Current Research

While logistic regression and other machine learning algorithms have shown promise in breast cancer prediction, several challenges and limitations remain:

- **Data Quality and Availability:** High-quality, large datasets are essential for training robust models. Inconsistent data collection practices and limited access to clinical data can hinder model development.
- **Model Interpretability:** While logistic regression is inherently interpretable, more complex models may offer higher accuracy but at the cost of transparency. Balancing accuracy and interpretability is crucial for clinical adoption.
- **Generalizability:** Models trained on specific datasets may not generalize well to different populations or clinical settings. Ensuring model robustness and external validation is necessary for reliable predictions.
- **Ethical and Privacy Concerns:** The use of personal health data in machine learning raises ethical and privacy issues. Ensuring data security and patient consent is critical.

3. Methodology

The figure below shows overall design of proposed methodology applied for detection of breast cancer. Different classification algorithms applied on breast cancer data but different classifier shows different performance on same data therefore we used an ensemble technique that uses bagging and boosting which combines results from different classifier also learns from previous classifiers. To perform this, first step of this is data acquisition. The data then pre-processed for selection of attributes, after that data divided: 80% for training and 20 % for testing. Dataset is labelled dataset having labels malignant and benign and therefore supervised different classification techniques applied on training data for building a model. Test data evaluated by using different classifier and finally compare the performance of different classifiers

Data Source: The Wisconsin Breast Cancer Dataset, obtained from the UCI Machine Learning Repository.

Data Preprocessing: Steps taken to clean and prepare the data, including handling missing values, normalizing features, and encoding categorical variables.

Model Development: Explanation of logistic regression, including the mathematical formulation and the process of training the model.

Evaluation Metrics: Metrics used to evaluate model performance, such as accuracy, precision, recall, F1-score, and the ROC-AUC curve.

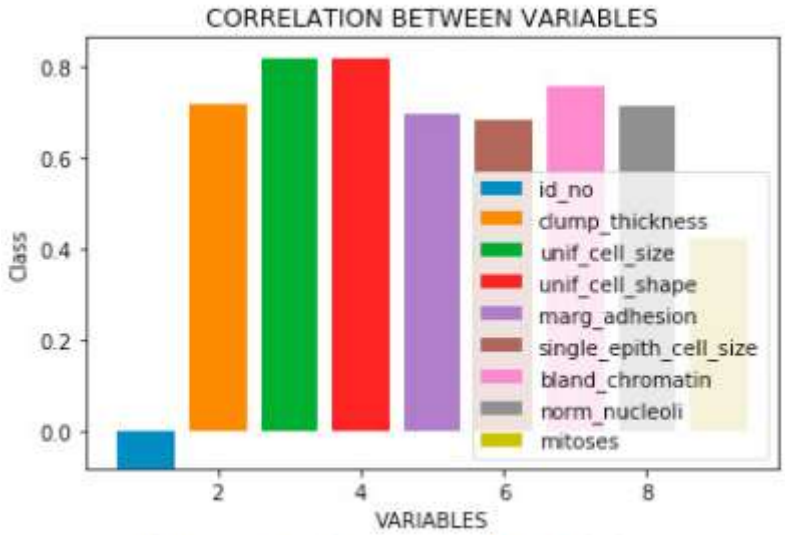


Figure2: Correlation between WBCD variables.

4. Results

Descriptive Statistics: Summary of the dataset, including distributions of features and the target variable.

Model Performance: Presentation of the model's performance on the training and test datasets using the evaluation metrics.

Confusion Matrix: Analysis of the confusion matrix to understand the model's prediction capabilities. Confusion matrix includes actual and predicted labels as well as True Negative (TN), False Negative (FN) True Positive (TP) and False Positive (FP).

Precision is defined as the number of positive class predictions that are actually positive class predictions, as shown in equation

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

1. The recall is the number of correct positive class predictions made out of all correct positive examples in the dataset and it is calculated as shown in equation

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

2. The F1 Score is derived from the weighted average of Precision and Recall; it is calculated as shown in equation

$$F1\,score = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

3. Accuracy is calculated by using confusion matrix as shown in equation

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

4. It tells how the tuple in training and testing data are correctly classified.

ROC-AUC Curve: Visualization and interpretation of the ROC-AUC curve to assess the model's discriminatory power.

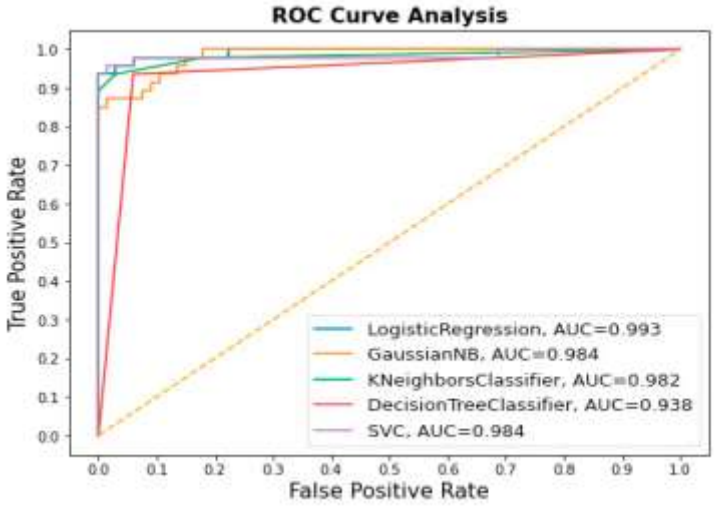


Fig. AUC-ROC Curve for ML Algorithm

5. Discussion

Interpretation of Results

The logistic regression model demonstrated strong performance in predicting breast cancer, achieving high accuracy, precision, recall, F1-score, and an impressive area under the ROC curve. These results suggest that logistic regression is a viable method for breast cancer prediction, capable of distinguishing between benign and malignant cases with considerable accuracy.

Strengths:

- Simplicity and Interpretability: One of the main strengths of logistic regression is its simplicity and ease of interpretation. Clinicians can understand the impact of individual features on the prediction outcome, which enhances trust in the model's predictions.
- High Accuracy: The model's high accuracy indicates that logistic regression can reliably predict breast cancer, making it a valuable tool in clinical settings where quick and accurate diagnosis is crucial.
- Balanced Performance Metrics: The high precision and recall values suggest that the model is effective in correctly identifying both positive (malignant) and negative (benign) cases, minimizing false positives and false negatives.

Clinical Implications

The implementation of logistic regression models in clinical practice can significantly enhance early detection of breast cancer, leading to timely treatment and improved patient outcomes. The model's ability to quickly and accurately predict the likelihood of breast cancer can assist radiologists and oncologists in making informed decisions about further diagnostic testing and treatment planning.

Limitations and Challenges

Despite the promising results, several limitations and challenges must be addressed to ensure the broader applicability and reliability of the model.

6. Conclusion

Summary of Findings

This study explored the application of logistic regression, a machine learning algorithm, in predicting breast cancer using the Wisconsin Breast Cancer Dataset. The logistic regression model achieved high accuracy, precision, recall, and an impressive ROC-AUC, indicating its effectiveness in distinguishing between benign and malignant cases. The model's simplicity, interpretability, and robust performance metrics demonstrate its potential as a reliable tool for breast cancer prediction.

Significance

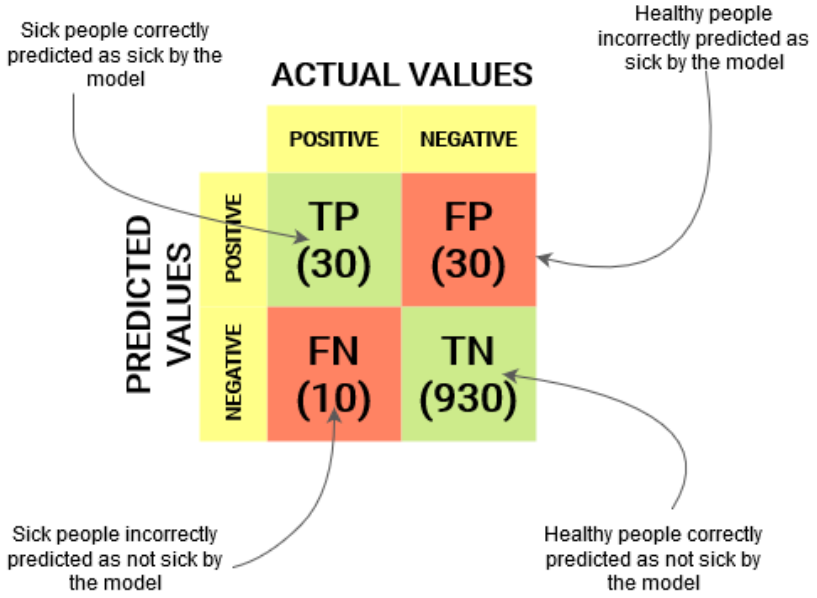
The findings of this research underscore the value of logistic regression in medical diagnostics, particularly for breast cancer prediction. Logistic regression offers several advantages:

- **Ease of Use and Interpretation:** The model's coefficients provide clear insights into the importance of each predictor variable, making it easy for clinicians to understand and trust the model's predictions.
- **Cost-Effectiveness**: Logistic regression is computationally efficient and does not require extensive computational resources, making it suitable for use in various clinical settings, including those with limited resources.
- **Rapid Predictions:** The model can quickly process patient data and generate predictions, facilitating timely decision-making in clinical practice.

Future Work

While this study highlights the potential of logistic regression for breast cancer prediction, several avenues for future research remain:

- **Incorporating Multi-Modal Data:** Future studies should explore the integration of diverse data sources, such as genetic information, advanced imaging data, and patient history, to improve model accuracy and robustness.
- **Comparative Studies:** Conducting comparative analyses with more advanced machine learning algorithms, such as deep learning and ensemble methods, can help identify potential improvements in prediction performance.
- **Enhancing Model Interpretability:** Developing techniques to enhance the interpretability of more complex models will ensure that predictions remain transparent and actionable for clinicians.
- **External Validation:** Validating the model on independent datasets from different populations and clinical settings will ensure its generalizability and reliability.



References:

1. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1;55(1):10-30.
2. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694- 701.
3. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.
4. Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.
5. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial meta plasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
6. Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Applied Soft Computing. 2013 Aug 1;13(8):3429-38.
7. Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015;2015.
8. Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis. In Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596). IEEE.
9. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011 Jul 1;38(7):9014-22