

Breast Cancer Prediction

Ramakrishna Mission Vivekananda Educational & Research Institute

Belur Math, Howrah, West Bengal
Department of Computer Science
Machine Learning – Course Project Report

Student Name: Pritam Kayal

Student Id: B2530096

1 Problem Statement

Breast cancer is one of the most common and fatal diseases affecting women worldwide. Early diagnosis plays a critical role in improving patient survival rates and treatment outcomes. Traditional diagnosis methods, though effective, can be time-consuming and subject to human interpretation errors. Therefore, it becomes essential to use machine learning to assist in early and accurate prediction.

The primary objective of this project is to develop a predictive model that classifies breast tumors as **malignant (cancerous)** or **benign (non-cancerous)** based on cell nucleus features. The project employs **Logistic Regression** and **Random Forest** algorithms, evaluated through 10-Fold Cross Validation to ensure accuracy, robustness, and generalization of results.

2 Proposed Methodology

This project uses Logistic Regression, a supervised linear classification algorithm, to model the probability that a given tumor is malignant.

Logistic Regression is well-suited for binary classification problems as it outputs probabilities between 0 and 1, representing the likelihood of a positive (malignant) outcome.

Steps Followed:

1. Data Collection:

The dataset is collected in CSV format containing both malignant and benign cases, with cell characteristics as features.

2. Data Preprocessing:

- Load the dataset (Kaggle Breast Cancer Wisconsin Dataset).
- Handle missing or inconsistent values.
- Encode the target variable ($M \rightarrow 1$ for malignant, $B \rightarrow 0$ for benign).

3. Model Selection — Logistic Regression:

- Logistic Regression is a linear binary classification algorithm that predicts the probability of a sample belonging to the malignant class.
- The logistic function (sigmoid) maps input features to probabilities between 0 and 1:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

If $P(y = 1 \mid X) \geq 0.5$, then the tumor is classified as malignant; otherwise

4. K-Fold Validation:

To prevent overfitting and obtain a reliable performance estimate, 10-Fold Cross Validation is applied:

- The dataset is divided into 10 equal folds.
- The model is trained on 9 folds and tested on the remaining 1 fold.
- The process is repeated 10 times, and the average of all metrics is computed.

5. Random Forest Regression:

Random Forest is an ensemble method using multiple decision trees and averaging their predictions, reducing variance and capturing nonlinear relationships not accessible to simple linear models.

6. Model Evaluation:

Performance was evaluated on the test set using key metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**
- **ROC Curve and AUC Score**

3 Dataset Details

The dataset used for this project is the Breast Cancer Wisconsin (Diagnostic) Dataset, a widely used benchmark dataset for binary classification tasks in medical machine learning.

- **Source:** Kaggle
- **Format:** CSV file (dataset.csv)
- **Total Samples:** 569 patient records (rows)
- **Number of Features:** 31 input attributes describing cell nuclei characteristics measured from digitized images of fine needle aspirates of breast masses.
- **Target Variable:**
Diagnosis — indicates whether the tumor is
 - M (Malignant) → cancerous (encoded as 1)
 - B (Benign) → non-cancerous (encoded as 0)

4 Comparative Analysis & Results

This section presents the detailed evaluation of the Logistic Regression model trained for Breast Cancer Prediction.

1. Model Evaluation Approach:

To assess the effectiveness and robustness of the developed models — **Logistic Regression** and **Random Forest** — a systematic evaluation approach was followed. The goal was to ensure that the models not only perform well on the training data but also generalize effectively to unseen data. The following evaluation strategies and performance metrics were used.

2. Logistic Regression :

Logistic Regression is one of the most widely used supervised learning algorithms for binary classification problems.

Linear regression, which predicts continuous values, logistic regression predicts the probability that an input instance belongs to a specific class — in this case, whether a breast tumor is malignant (1) or benign (0).

Metric	Value
Accuracy	97.37%
Precision	97.62%
Recall	95.35%
F1-Score	96.47%
ROC Curve	0.997

- **Analysis:**

The Receiver Operating Characteristic (ROC) curve for Logistic Regression shows how the model balances between true positive rate (sensitivity) and false positive rate at different thresholds.

- The AUC Score (0.997) indicates almost perfect discrimination between malignant and benign tumors

3. Random Forest:

Random Forest is a powerful and widely used ensemble learning algorithm that builds multiple Decision Trees and combines their predictions to improve accuracy and reduce overfitting.

Metric	Test Result
Accuracy	96.49%
Precision	97.56%
Recall	93.02%
F1-Score	95.24%
ROC-AUC	0.995

- **Interpretation:**

The Receiver Operating Characteristic (ROC) curve for Random Forest demonstrates the trade-off between true positive rate and false positive rate at various thresholds.

- The AUC score (0.995) indicates that the model is almost perfect at distinguishing between malignant and benign cases.

4. Confusion Matrix:

The confusion matrix helps to understand how many samples were correctly or incorrectly classified.

Table 1: Confusion Matrices of Logistic Regression

Actual / Predicted	Benign (0)	Malignant (1)
Benign (0)	71	1
Malignant (1)	2	40

Table 2: Confusion Matrices of Random Forest

Actual / Predicted	Benign (0)	Malignant (1)
Benign (0)	70	1
Malignant (1)	3	40

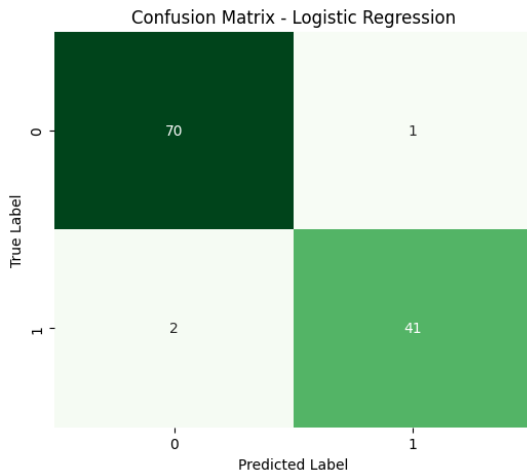


Figure 1: Logistic Regression

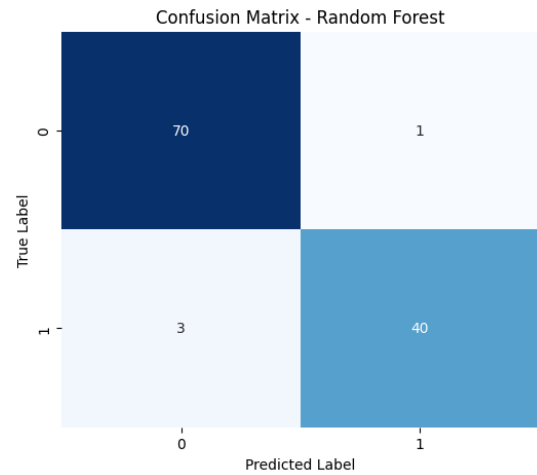


Figure 2: Random Forest

5. ROC Curve and AUC:

In this project, both **Logistic Regression** and **Random Forest** models achieved high AUC values, indicating strong classification capability and robust discrimination between malignant and benign tumors.

Interpretation :

- The Logistic Regression model achieved a slightly higher AUC (0.997) compared to Random Forest's AUC (0.995).
- This means Logistic Regression performs marginally better at distinguishing between malignant and benign cases across different probability thresholds.
- The ROC curve of Logistic Regression lies slightly above that of Random Forest, indicating better sensitivity for the same level of false positives.

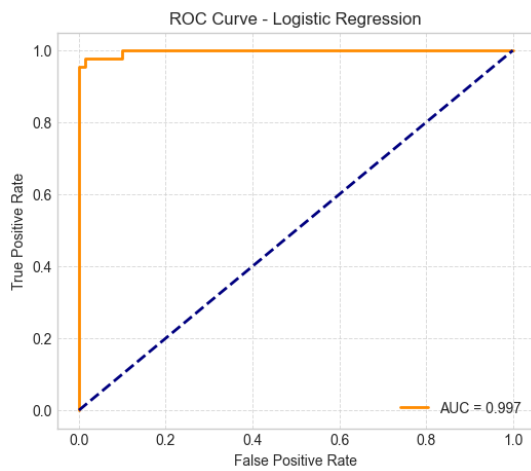


Figure 3: Logistic Regression

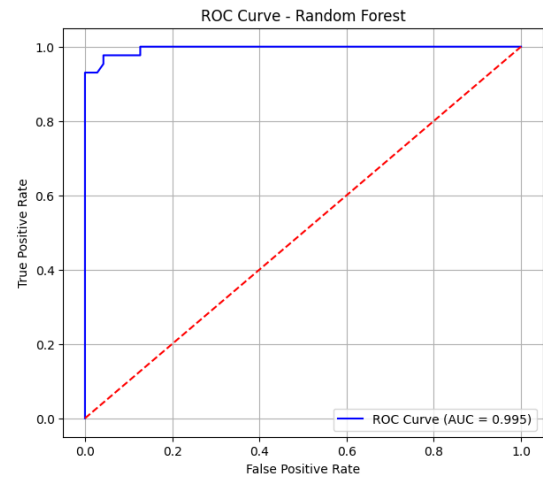


Figure 4: Random Forest

5 Conclusion

This project successfully demonstrates how Logistic Regression can be applied to predict breast cancer diagnosis with high accuracy using clinical diagnostic data. The model achieves strong performance in all key metrics, accuracy, recall, and ROC, and provides interpretable results that help to understand which characteristics contribute the most to tumor classification.

- Logistic Regression efficiently classifies tumors as benign or malignant.
- The most important predictive features include mean radius, texture, and concavity.

References

1. Kaggle, Breast Cancer Prediction_dataset.csv
2. scikit-learn, pandas and matplotlib documentation