Breast Cancer Prediction using Machine Learning (Random Forest)

MAMTA

02/09/2025

Problem Statement

Breast cancer is one of the leading causes of cancer deaths in women.

Early detection can save lives.

Goal → Build an ML model to classify **Malignant (cancerous)** vs **Benign (non-cancerous)** tumors.

Dataset Overview

Dataset: *Breast Cancer Wisconsin Diagnostic Dataset*

Size: ~569 samples, 30 features + target (diagnosis)

Target variable:

- **M → Malignant (1)**
- **B → Benign (0)**

Dropped: `id`, `Unnamed: 32`

Data Preprocessing

Converted diagnosis labels (`M` → 1, `B` → 0).

Checked for missing values.

Standardized features using **StandardScaler**.

Split data: 80% training, 20% testing (stratified).

Model Selection

Algorithm used: **Random Forest Classifier**

Why Random Forest?

- Handles high-dimensional data.
- Reduces overfitting.
- Provides **feature importance** ranking.

Parameters: `n_estimators = 100`, `random_state=42`

Model Performance

Accuracy: **~97.3684%**

Feature Importance

Bar chart of **top features responsible for prediction**.

Example: *radius_mean, concave points_mean, area_worst...*

Insight: Certain tumor shape & size metrics strongly indicate malignancy.

Prediction Example

Input new tumor data (30 features).

Model scales & predicts.

Example Output:

"Prediction: Malignant Tumor (Cancerous)"

Conclusion

Achieved **high accuracy with Random Forest**.

Model can assist doctors in early detection.

Next steps:

- Try other ML models (Logistic Regression).