Breast Cancer ML Project Report

Project Title

Breast Cancer Diagnosis using Machine Learning

Technologies Used

1. Python: Main programming language.

2. Pandas & NumPy: Data manipulation and numerical operations.

3. Matplotlib & Seaborn: Data visualization.

4. missingno: Visualizing missing values.

5. Scikit-learn: ML algorithms (Logistic Regression, KNN, SVM, Decision Tree, Random Forest, Gradient Boosting).

6. XGBoost: Advanced boosting algorithm.

Project Overview

This machine learning project aims to classify breast cancer diagnoses as malignant (M) or benign (B) using a dataset from UCI.

Various ML algorithms are trained and tested to determine the most accurate model.

The dataset is preprocessed, features are selected based on correlation, and models are evaluated using accuracy and ROC.

Data Preprocessing

1. Loaded data using pandas.

2. Diagnoses converted: M = 1, B = 0.

3. Visualized missing values.

4. Dropped highly correlated features.

5. Applied StandardScaler for feature scaling.

Algorithms Used

1. Logistic Regression: Linear model for binary classification.

2. K-Nearest Neighbors (KNN): Classifies based on closest neighbors.

3. Support Vector Machine (SVM): Finds the best boundary to separate classes.

4. Decision Tree: Splits data into branches based on features.

5. Random Forest: Ensemble of decision trees.

6. Gradient Boosting: Sequential tree-based method optimizing loss function.

7. XGBoost: Efficient gradient boosting implementation with high performance.

Model Evaluation

Models are evaluated using:

- Accuracy

- Confusion Matrix

- Classification Report (Precision, Recall, F1 Score)

- ROC Curve and AUC Score

Bar charts compare Accuracy and ROC of all models.