

Breast Cancer Classification

Waleed Alfaifi

1. Introduction

This project focuses on building and evaluating machine learning models to classify breast cancer tumors as malignant (M) or benign (B) using the Breast Cancer Wisconsin dataset. The dataset contains 569 samples and 30 features describing cell nuclei characteristics.

2. Data Preprocessing

- Loaded dataset and explored shape (569 rows, 32 columns).
- Removed unnecessary columns (ID).
- Encoded target diagnosis (M=1, B=0).
- Verified missing values (none detected).
- Split data into training (75%) and testing (25%).

3. Model Training & Hyperparameter Tuning

We applied several machine learning algorithms and tuned them using GridSearchCV:

- Logistic Regression (Best: C=1, solver=liblinear)
- Support Vector Classifier (Best: C=10, kernel=rbf)
- Random Forest (Best: n_estimators=200)
- K-Nearest Neighbors (Best: n_neighbors=5)
- Gaussian Naive Bayes

4. Model Evaluation

All models achieved high accuracy (~94%-98%). The best-performing models were Random Forest and SelectKBest feature selection with k=15-20 achieving 98% accuracy.

5. Feature Selection

We applied SelectKBest with different k values (5, 10, 15, 20). Feature selection improved performance, with best results at k=15 and k=20.