Group 4: AFF HACKERS

Introduction

- Breast cancer is one of the most common cancers worldwide.
- Early and accurate detection is crucial for effective treatment.
- This project uses machine learning to classify tumors as benign or malignant.

Dataset Overview

This is the **Wisconsin Breast Cancer (Diagnostic)** dataset widely available on Kaggle and the UCI Machine Learning Repository. It's composed of diagnostic data derived from digitized images of fine needle aspirate (FNA) samples of breast masses

It contains **569 instances** with **32 attributes** (including an ID and diagnosis). The remaining 30 features represent the **mean**, **standard error**, and **"worst" (mean of three largest)** values for each of ten nuclear characteristics (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension)

The target variable is **diagnosis**: 'M' for malignant, 'B' for benign. The dataset includes **357 benign** and **212 malignant** samples, and it is known for having **no missing values**

It’s a clean, well-structured dataset—ideal for practicing classification tasks. Researchers and learners often use it to implement, test, and compare algorithms like logistic regression, decision trees, and ensemble methods

Exploratory data analysis helps identify which nuclear features distinguish malignant from benign cases. For example, the “area worst” feature showed statistically significant differences across classes, providing insight into early indicators used in screening

Code Steps;

1. Import libraries
2. Load dataset
3. Exploratory data analysis
4. Data processing
    1. Missing Values
    2. ML Based Imputation
    3. Dealing with Outliers
    4. Encoding
5. Machine learning
    1. Import models
    2. Define models
    3. Apply Gridsearch CV
    4. Evaluation
    5. Retrain best model
    6. Save model

Summary

	Model	Best Params	Cross-Validated Accuracy	Accuracy	Precision	Recall	F1
0	LogisticRegression	{}	0.628571	0.576087	0.576087	0.576087	0.576087
1	RandomForest	{'model__max_depth': 20, 'model__n_estimators'...	0.678912	0.652174	0.652174	0.652174	0.652174
2	XGBoost	{'model__learning_rate': 0.1, 'model__max_dept...	0.678912	0.663043	0.663043	0.663043	0.663043
3	SVC	{'model__C': 1, 'model__gamma': 'scale', 'mode...	0.634014	0.592391	0.592391	0.592391	0.592391
4	KNeighbors	{'model__n_neighbors': 7, 'model__weights': 'd...	0.600000	0.538043	0.538043	0.538043	0.538043
5	GradientBoosting	{'model__learning_rate': 0.05, 'model__max_dep...	0.672109	0.652174	0.652174	0.652174	0.652174
6	AdaBoost	{'model__learning_rate': 0.5, 'model__n_estima...	0.648980	0.641304	0.641304	0.641304	0.641304
7	HistGradientBoosting	{'model__learning_rate': 0.05, 'model__max_dep...	0.669388	0.646739	0.646739	0.646739	0.646739