Breast Cancer Diagnostic Machine Learning Report

0. Project Workflow (Flowchart Style)

Data → EDA → Preprocessing → Model Training → Evaluation → Future Improvements

1. Project Overview

This project predicts malignant versus benign breast tumors using the Wisconsin Breast Cancer Diagnostic Dataset (Kaggle). It demonstrates beginner-to-intermediate machine learning skills including EDA, preprocessing, model training, evaluation, and future planning. It represents complete workflow understanding and readiness for ML engineering roles in healthcare.

2. Dataset Summary

• Samples: 569 rows, 32 columns.

• Target: diagnosis (M = malignant, B = benign).

• Missing values: No missing values found in this dataset.

• Features: Numeric measurements describing cell nuclei (radius, texture, symmetry, etc.).

• Cleaning: The ID column was removed because it does not carry predictive information.

3. Exploratory Data Analysis (EDA)

• I used distribution plots for several numerical features to understand how the data is spread, check skewness, and quickly notice any outliers.
• I used univariate comparison plots (boxplots) to take a closer look at two highly correlated features and see how they behave with the diagnosis labels.
• I created a correlation heatmap to clearly see how each numerical feature relates to the target diagnosis and to identify features contributing strongly to the prediction.

4. Preprocessing Steps

• Label encoding (M → 1, B → 0).

• Standard scaling applied to numerical features.

• 80/20 train-test split.

• Separating X (features) and y (labels).

5. Models Used

• Logistic Regression – simple and interpretable baseline.

• Random Forest – ensemble of decision trees for non-linear patterns.

• XGBoost – gradient boosting model that often performs well on tabular data.

6. Evaluation Approach

• Use train-test split so that evaluation is done on unseen data.

• Compute accuracy for overall correctness.

• Use precision and recall for malignant class evaluation.

• Check F1-score for balanced performance.

• Recall is important for reducing false negatives in medical ML.

7. Main Results (Test Set)

• Logistic Regression – Accuracy 0.9737, Precision 0.9756, Recall 0.9524, F1 0.9639.

• Random Forest – Accuracy 0.9649, Precision 1.0000, Recall 0.9048, F1 0.9500.

• XGBoost – Accuracy 0.9737, Precision 1.0000, Recall 0.9286, F1 0.9630.

All three models performed strongly. Logistic Regression and XGBoost were treated as primary reference models.

8. Future Improvements

• Use GridSearchCV to tune hyperparameters for Random Forest and XGBoost.

• Try SVC and KNN on the same dataset for comparison.

• Add cross-validation to improve robustness.

9. What This Project Demonstrates

This project shows beginner to intermediate skill using a real medical dataset. It demonstrates that I can structure an end-to-end ML workflow from raw data to evaluation, use multiple models and compare them fairly, maintain strong workflow understanding with clear logic behind algorithm selection, and focus on clinically meaningful metrics such as recall for malignant cases. I plan model improvements and next steps while keeping explanations readable for people who are not ML experts. My background in teaching helps me explain technical ideas in simple language, which is useful when working with cross-functional teams in healthcare. Overall, this project demonstrates readiness for technical ML roles.