**Breast Cancer Wisconsin Classification Project**

**📄 Breast Cancer Classification: Wisconsin Dataset**

**🔍 Project Objective**

This project aims to develop a robust supervised machine learning model to classify tumors as **benign** or **malignant** using the **Wisconsin Breast Cancer dataset**. The goal is to achieve high accuracy, precision, and recall while ensuring the model generalizes well to unseen data.

**📚 Dataset Overview**

* **Source:** scikit-learn’s load\_breast\_cancer function.
* **Samples:** 569 instances of digitized images of fine needle aspirate (FNA) of breast masses.
* **Features:** 30 numeric features describing characteristics of the cell nuclei.
* **Target Classes:** 0 = malignant, 1 = benign.

**2️⃣ Exploratory Data Analysis (EDA)**

* Checked for correlations:
  + Heatmap revealed some strong correlations (e.g., mean radius & mean perimeter).
* Visualized class distribution and feature relationships using pair plots and box plots.

**3️⃣ Data Preprocessing**

* **Scaling:** Used StandardScaler to normalize features.
* **Train-Test Split:**  
  Typical split was 60% training, 20% for Holdout Validation Set to evaluate generalization, 20% testing for Final Evaluation.

**4️⃣ Model Building**

* **Algorithms Tested:**
  + Logistic Regression
  + K-Nearest Neighbors Classifier (KNN)
  + Support Vector Machine Classifier (SVC)
  + Random Forest Classifier
  + Ada Boost Classifier
  + Cat Boost Classifier
  + LightGBM Classifier
  + XG Boost Classifier
* **Hyperparameter Tuning:**  
  GridSearchCV used to find optimal parameters for Logistic Regression (C, Fit Intercept, Penalty, l1\_ratio=0, max\_iter, solver).

**5️⃣ Model Evaluation**

* **Metrics:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
* **Best Model:**
  + **Logistic Regression** performed the best with accuracy ~97%.
  + Confusion matrix showed very few false negatives, which is critical in cancer detection.

**✅ Key Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression (Best) | 98% | 98% | 98% | 98% |
| KNN | 96% | 96% | 96% | 96% |
| SVC | 95% | 95% | 95% | 95% |
| Random Forest | 97% | 97% | 97% | 97% |
| Decision Trees | 94% | 94% | 94% | 94% |
| XGBoost | 96% | 96% | 96% | 96% |
| AdaBoost | 96% | 96% | 96% | 96% |
| LightGBM | 96% | 96% | 96% | 96% |
| CatBoost | 97% | 97% | 97% | 97% |

**🔑 Insights & Recommendations**

* **Logistic Regression** is well-suited for this dataset due to its performance on smaller datasets.
* Feature importance analysis suggests mean radius and mean texture are highly predictive.
* For deployment, further validation on real-world samples is recommended.
* Imbalanced data handling could be improved with techniques like SMOTE if needed.

**🚀 Next Steps**

* Test the Final Model’s Robustness on Real-World Samples.
* Package the model with a simple user interface for medical staff.
* Test with real patient data to monitor performance drift.

**Prepared By:** *Pranjal Shinde*  
**Date:** *July 2025*