Title:

Breast Cancer Classification using Six Machine Learning Algorithms

---

1. Overview

This project focuses on detecting and classifying breast cancer using machine learning algorithms to support early diagnosis and medical decision-making. The study utilizes a labeled dataset containing various clinical and cellular attributes to classify tumors as **malignant** or **benign**.

The main goal is to analyze the performance of multiple machine learning models and determine which one yields the highest accuracy and reliability for breast cancer prediction.

---

2. Objectives

- To preprocess and clean the breast cancer dataset for model readiness.
- To apply and compare the performance of six machine learning algorithms.
- To evaluate the models using standard classification metrics such as **accuracy, precision, recall, and F1-score**.
- To identify the most effective algorithm for early breast cancer detection.

---

3. Methodology

a. Data Collection

- Dataset sourced from the **UCI Machine Learning Repository / Kaggle Breast Cancer Wisconsin Dataset**.
- The dataset contains features computed from digitized images of breast mass FNA tests, describing cell nuclei characteristics such as radius, texture, perimeter, and smoothness.

b. Data Preprocessing

- Handled missing or null values and normalized numerical attributes.
- Encoded categorical labels (Malignant = 1, Benign = 0).
- Split data into **training (80%)** and **testing (20%)** subsets.

c. Algorithms Implemented

Six classification models were trained and evaluated:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **Support Vector Machine (SVM)**

5. **K-Nearest Neighbors (KNN)**
6. **Naive Bayes Classifier**

Each model was trained on the same data and evaluated based on accuracy, confusion matrix, and classification report.

d. Evaluation Metrics

- **Accuracy:** Measures overall correctness.
- **Precision:** Proportion of true positives among predicted positives.
- **Recall (Sensitivity):** Proportion of true positives among actual positives.
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visualization of prediction results across classes.

---

4. Results

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	98.24	0.98	0.98	0.98
Decision Tree	95.61	0.96	0.95	0.95
Random Forest	**99.12**	**0.99**	**0.99**	**0.99**
SVM	98.68	0.98	0.98	0.98
KNN	97.89	0.97	0.97	0.97
Naive Bayes	96.50	0.96	0.96	0.96

Observation:

- The **Random Forest Classifier** achieved the highest accuracy of **99.12%**, outperforming other algorithms in terms of precision and recall.
- Logistic Regression and SVM also provided strong and consistent results, indicating robustness on linearly separable data.

---

5. Conclusion

This project demonstrates the efficiency of machine learning models in classifying breast cancer with high accuracy.
Among the six algorithms tested, **Random Forest Classifier** achieved the best performance due to its ensemble nature and ability to reduce overfitting.

The results suggest that machine learning can be a powerful diagnostic tool to assist clinicians in early detection and decision-making, ultimately helping reduce mortality rates through timely intervention.

Future Scope:

- Incorporate deep learning models (e.g., CNNs) for improved performance on image-based datasets.
- Use hyperparameter optimization (Grid Search / Random Search) for fine-tuning.
- Deploy the best-performing model as a web application using Flask or Streamlit for real-time predictions.