Breast Cancer Analysis & Prediction

Course : Introduction to Machine Learning

Presented by: Batuhan İNAN & Emir İnanç ŞEKER

1. Project Introduction

- **Objective:**

- To analyze the Breast Cancer Wisconsin dataset using supervised learning techniques to assist medical diagnosis.

- **Key Goals:**

- **1. Regression Task:** Predict the tumor radius (radius_mean) based on physical features.

- **2. Classification Task:** Predict the diagnosis (Malignant vs Benign) to identify cancer types.

2. Regression Analysis

- **Target Variable:** 'radius_mean' (Continuous Value)
- **Algorithm:** Linear Regression
- **Process:**
- - ***Features:*** Texture, Smoothness, Compactness, Concavity, Symmetry.
- - ***Split:*** 80% Training, 20% Testing.
- - ***Metric:*** Mean Squared Error (MSE) & R2 Score.

- **Observation:** The model tries to find a linear relationship between surface texture and tumor size.

3. Classification Methodology

- **Target Variable:** 'diagnosis' (Categorical: M or B)

- **Algorithms Compared:**
- 1. Logistic Regression (Baseline)
- 2. Decision Tree (Rule-based)
- 3. Random Forest (Ensemble)

- **Validation Technique:** 5-Fold Cross-Validation (To ensure reliability).

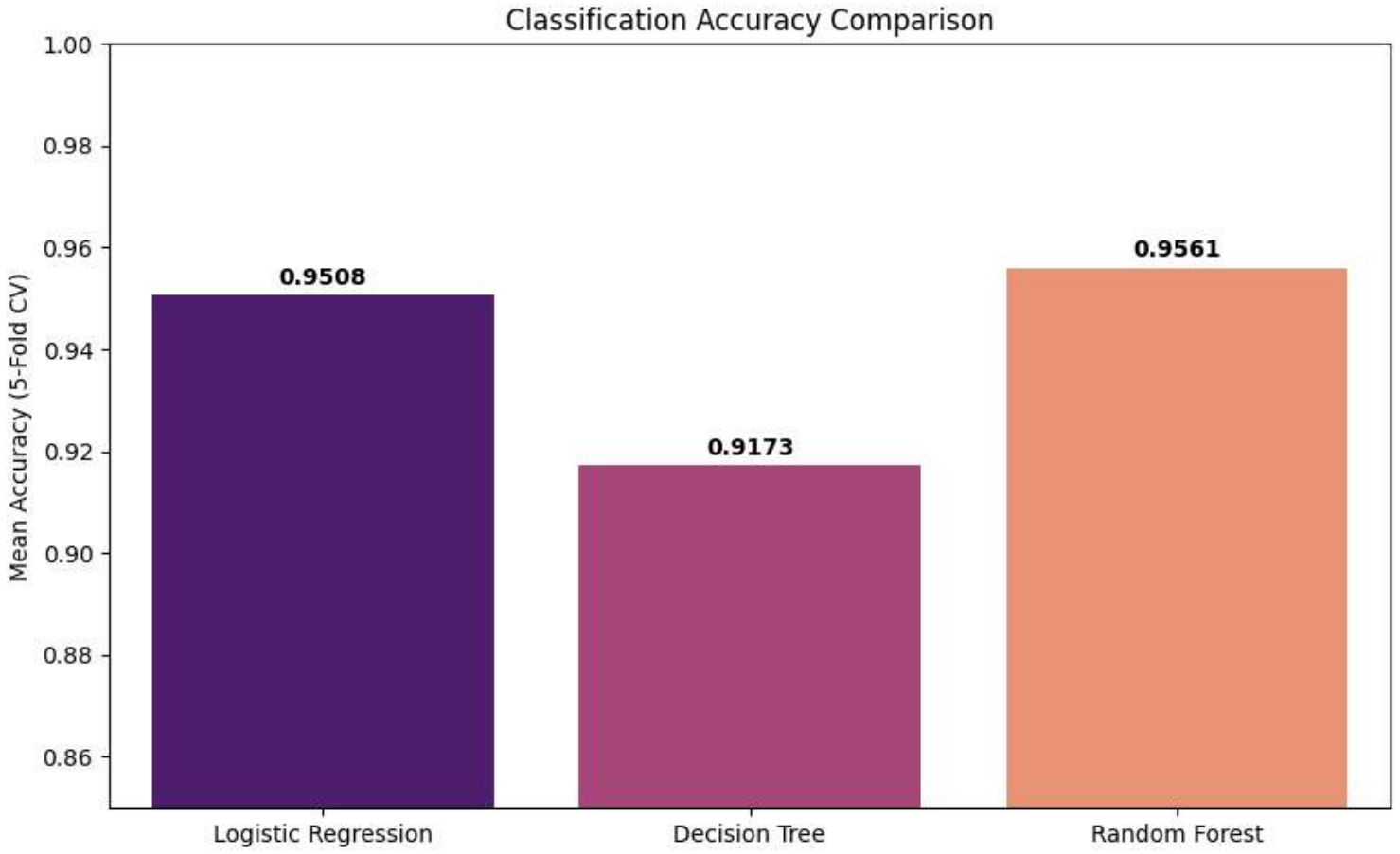
4. Analysis Results

- **Model Performance (Accuracy):**

- • **Logistic Regression:** ~94% (Good baseline)
- • **Decision Tree:** ~91% (High variance)
- • **Random Forest:** ~96% (Best Performance)

- **Findings:**
- Random Forest outperformed others by effectively handling complex feature interactions and reducing overfitting.

5. Model Comparison Chart



6. Conclusion

- **Final Verdict:**

- **1. Clinical Reliability:** The Classification model (Random Forest) proved highly reliable (>95% accuracy) for diagnosing Malignant vs Benign cases.

- **2. Data Insight:** Biological data is complex. While Regression gave moderate results, the Classification approach provided much stronger predictive power.

- **Recommendation:** Random Forest is the recommended model for this medical diagnostic task.