# Breast Cancer Dataset Analysis and Model Evaluation

## Introduction

The Breast Cancer dataset from Scikit-learn is used for binary classification to diagnose malignant or benign tumors. The dataset consists of 30 numerical features computed from digitized images of breast mass samples. This study explores the dataset, performs preprocessing, and evaluates two supervised learning models: Decision Tree and Logistic Regression.

## Exploratory Data Analysis (EDA)

The dataset contains 569 samples with 30 features and a binary target variable (0: malignant, 1: benign). Basic statistical analysis reveals that features such as mean radius, mean texture, and mean perimeter contribute significantly to classification. Visualizations such as pair plots and correlation matrices help identify relationships between features. The dataset exhibits class imbalance, with more benign cases than malignant ones.

### Key Insights:

- The mean radius and mean perimeter show a strong correlation.

- Most malignant cases exhibit higher values in texture and radius measurements.

- No missing values were found in the dataset.

### Data Preprocessing

Before model training, preprocessing steps were applied:

1.Feature Scaling: Standardization using “StandardScaler()” ensures that all features have zero mean and unit variance.

2. Train-Test Split: The dataset was split into 80% training and 20% testing to evaluate model performance fairly.

3.Class Balance Consideration: The distribution of target labels was examined to ensure no severe imbalance.

## Model Selection

Two supervised learning models were selected for classification:

Decision Tree Classifier: A tree-based model that learns simple rules to classify data points.

Logistic Regression: A linear model that estimates probabilities for binary classification.

## Model Training and Evaluation

Both models were trained using the preprocessed data and evaluated using accuracy, classification reports, and confusion matrices.

| Model | Accuracy | Precision | F1-Score |
| --- | --- | --- | --- |
| Logistic Regression | 97% | 97% | 97% |
| Decision Tree | 94% | 95% | 95% |

### Key Findings:

- Logistic Regression outperforms the Decision Tree classifier in terms of accuracy (97% vs. 94%).

- The confusion matrix indicates that Logistic Regression makes fewer misclassifications.

- Decision Trees are prone to overfitting, while Logistic Regression generalizes better.

### Visualizations

1. Feature Correlation Matrix: Identifies highly correlated features.

2. Confusion Matrix Heatmaps: Show model predictions and errors.

### Conclusion

The Breast Cancer dataset analysis demonstrates that Logistic Regression performs better than the Decision Tree classifier, achieving higher accuracy and generalization capability. Preprocessing steps such as feature scaling and train-test splitting were essential in ensuring reliable model performance. Future improvements could involve hyperparameter tuning and ensemble methods to enhance classification results.