Digital Skill Fair 35.0

CLASSIFICATION OF BREAST CANCER DATASET

Created By : Cindy Indriyani

INTRODUCTION

Background

- In 2022, breast cancer was **diagnosed in 2.3 million women** worldwide, **leading to 670,000 deaths** (World Health Organization, 2024)
- Breast cancer may not show symptoms in many women, making regular screening crucial for early detection (American Cancer Society, 2025)

Objective

- Determine the best classification model for detecting breast cancer
- Identify important features.

METHODOLOGY

Data and Data Sources

569 rows of data and **32 variables**

Responsive Variables:

Diagnosis:

- **M: Malignant**
- **B: Benign**

Predictor Variables:

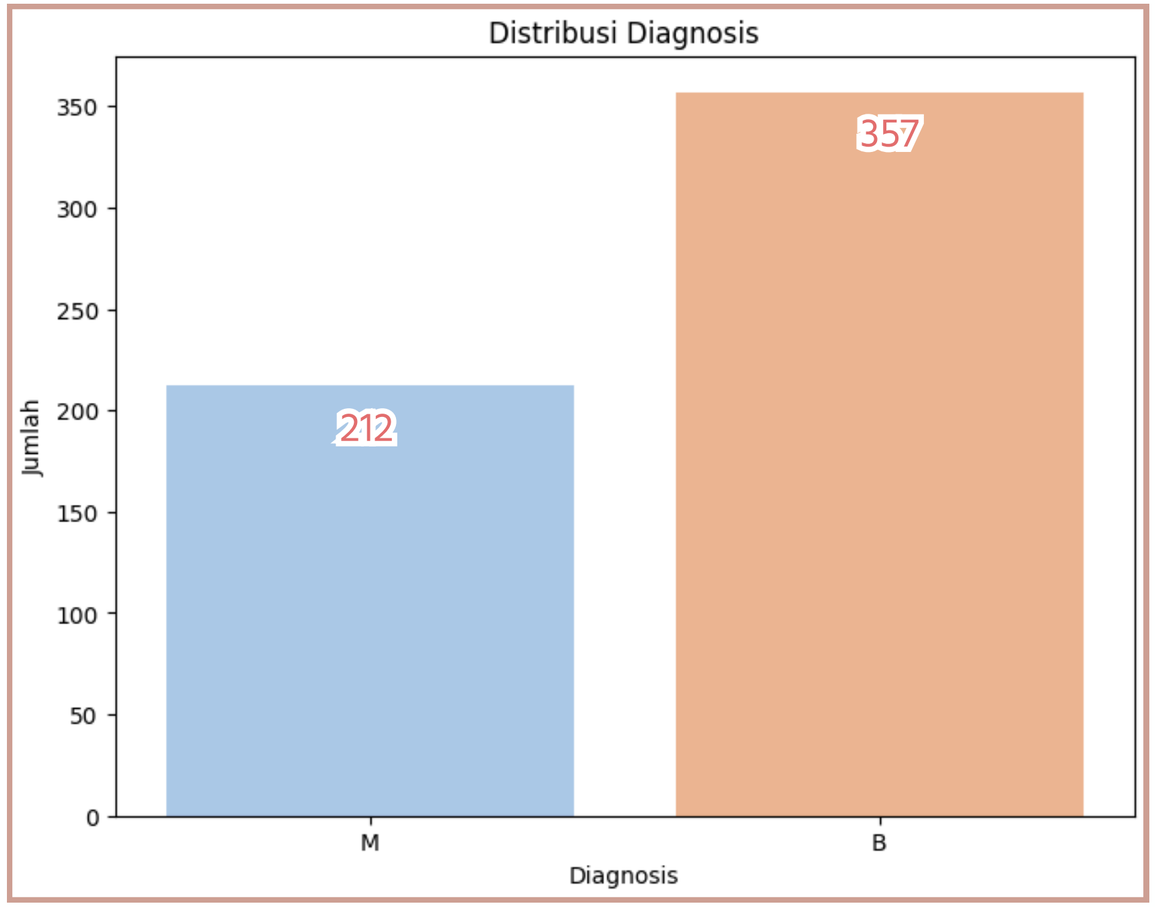
- **Radius**
- **Texture**
- **Perimeter**
- **Area**
- **Cmpactness**
- **Concavity**
- **Concave Points**
- **Symmetry**
- **Fractal Dimension**
- **Smoothness**

Analysis Procedure:

Data Input → **Data Preprocessing** → **Data Exploration** → **Data Splitting** → **Handling Data Imbalance: SMOTE** → **Model Training** → **Model Evaluation with Test Data** → **Interpretation**

RESULTS

Data Exploration

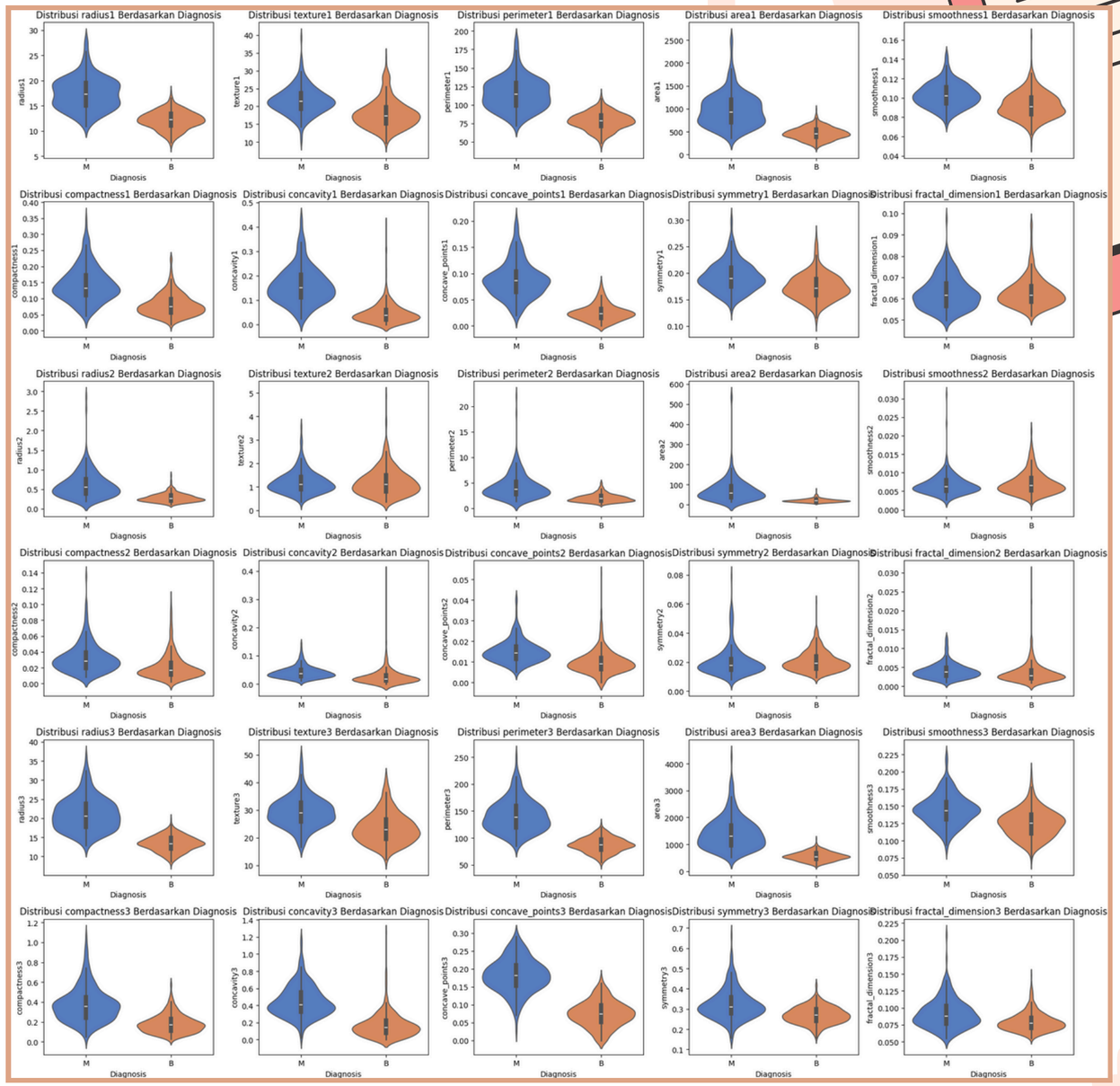


The **benign breast cancer** diagnosis has the **highest number of cases,** with **357 instances.**

RESULTS

Data Exploration

A **wider violin for malignant** cases might indicate that malignant tumors vary more widely in several characteristic features than benign ones, like **perimeter1**.



RESULTS

Data Splitting

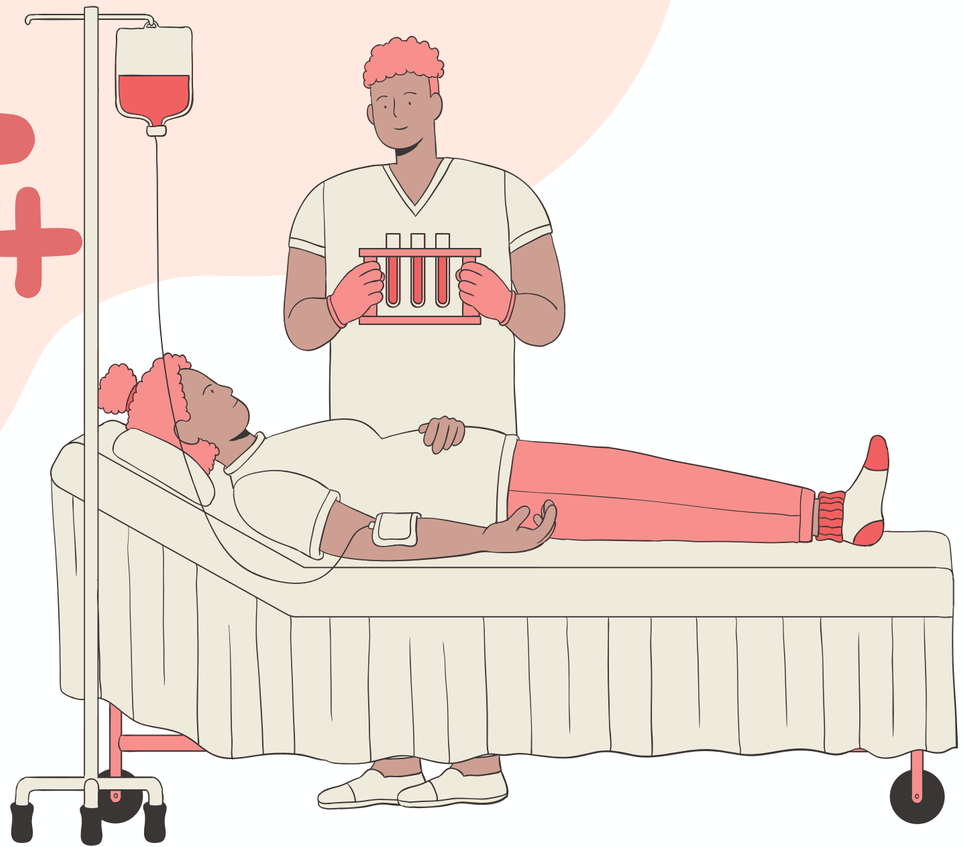
- **Training data**: 398 rows of data
- **Test data**: 171 rows of data

After SMOTE

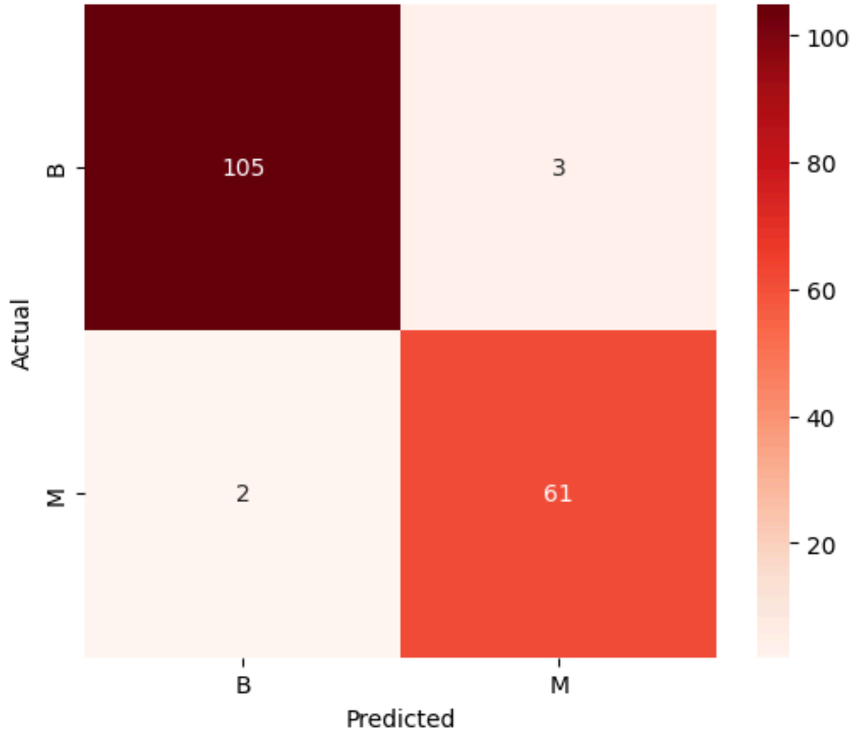
Training data: 498 rows of data

RESULTS

Confussion Matrix



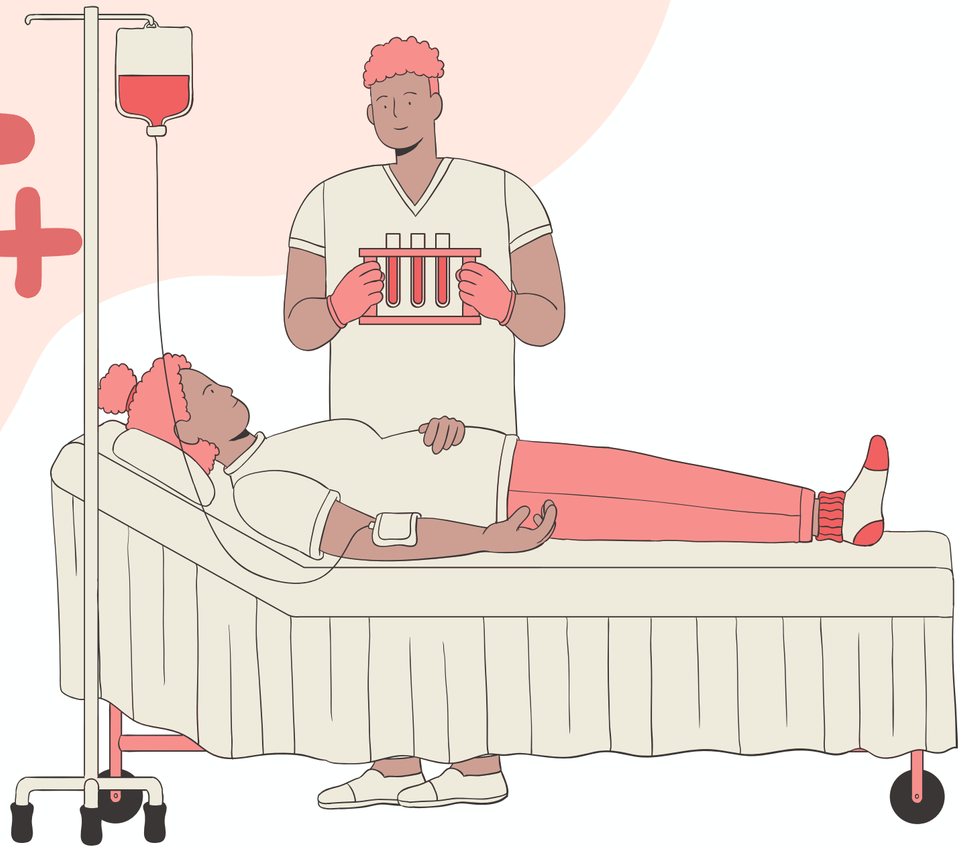
Random Forest



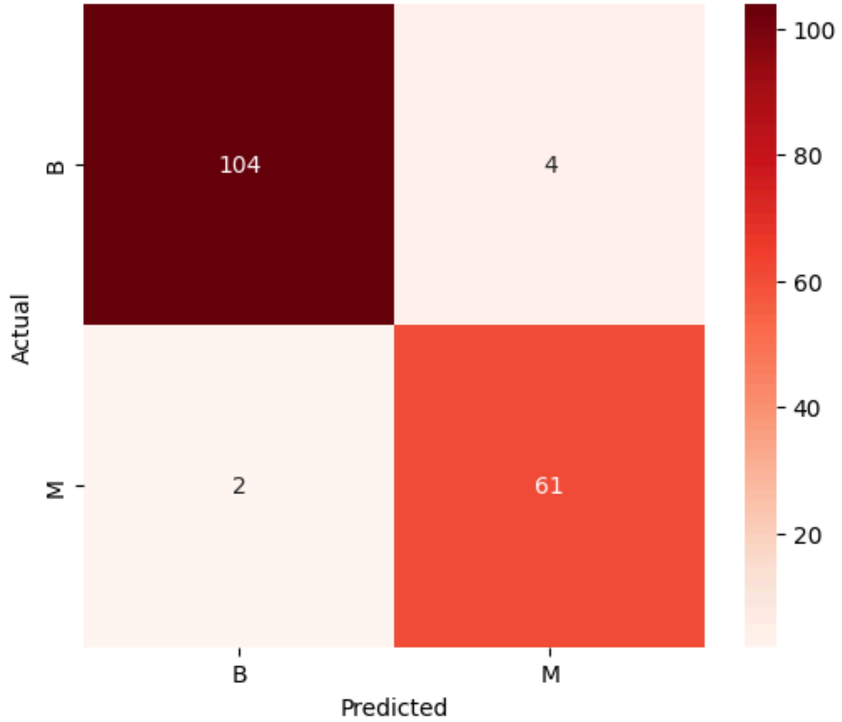
- **F1-Score**: 97%
- **Recall**: 97%
- **Precision**: 97%

RESULTS

Confussion Matrix



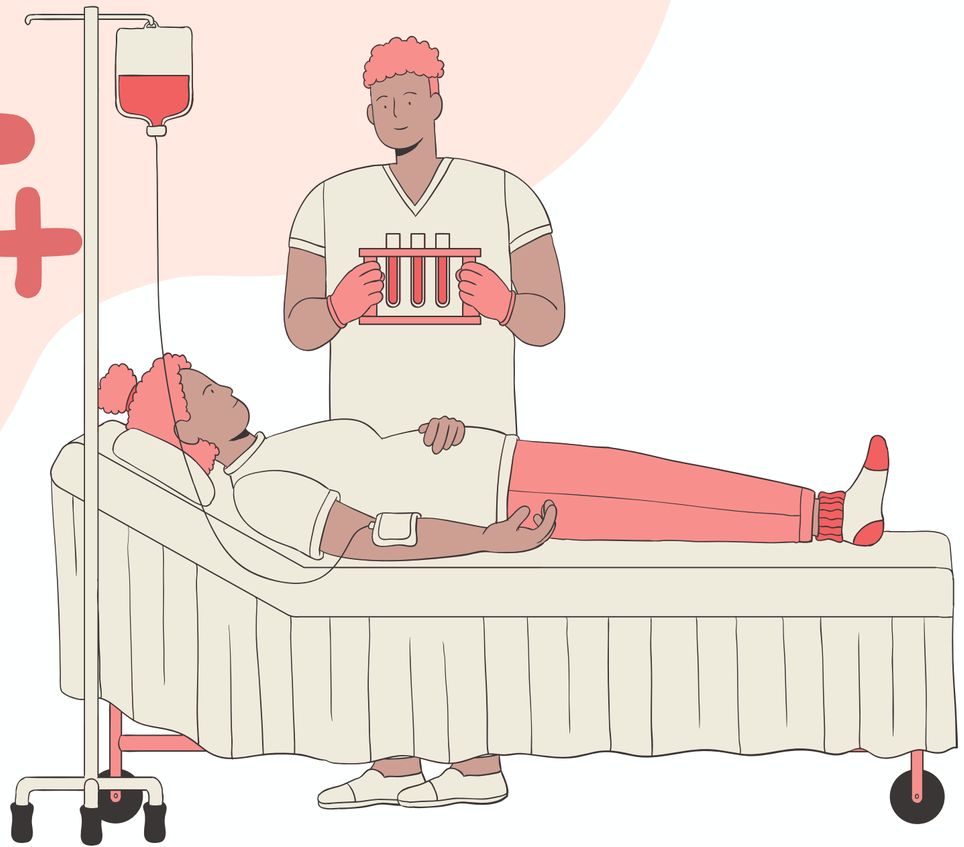
Gradient Boosting



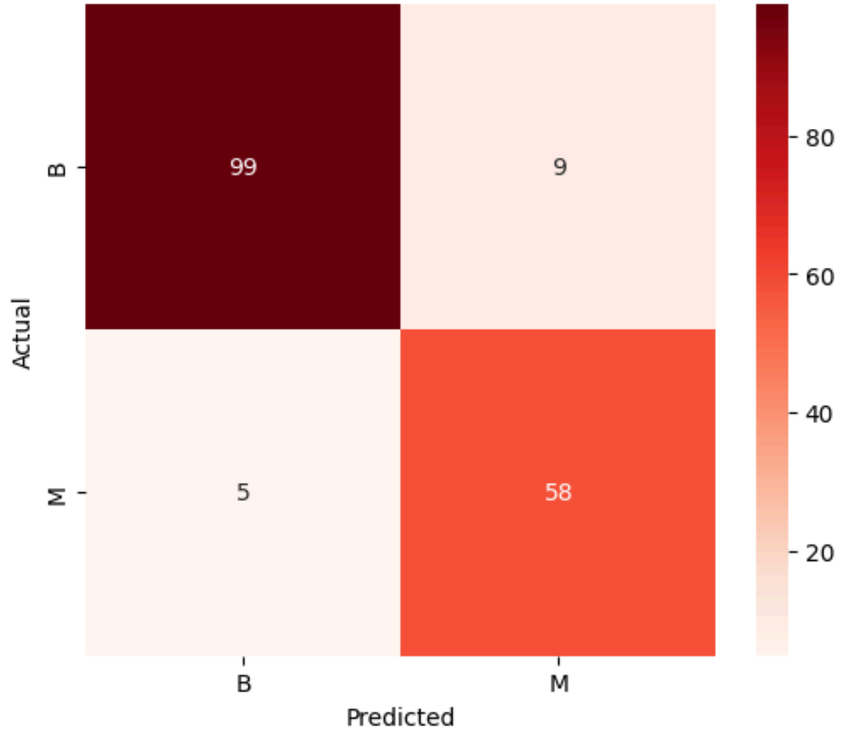
- **F1-Score**: 96%
- **Recall**: 97%
- **Precision**: 96%

RESULTS

Confussion Matrix



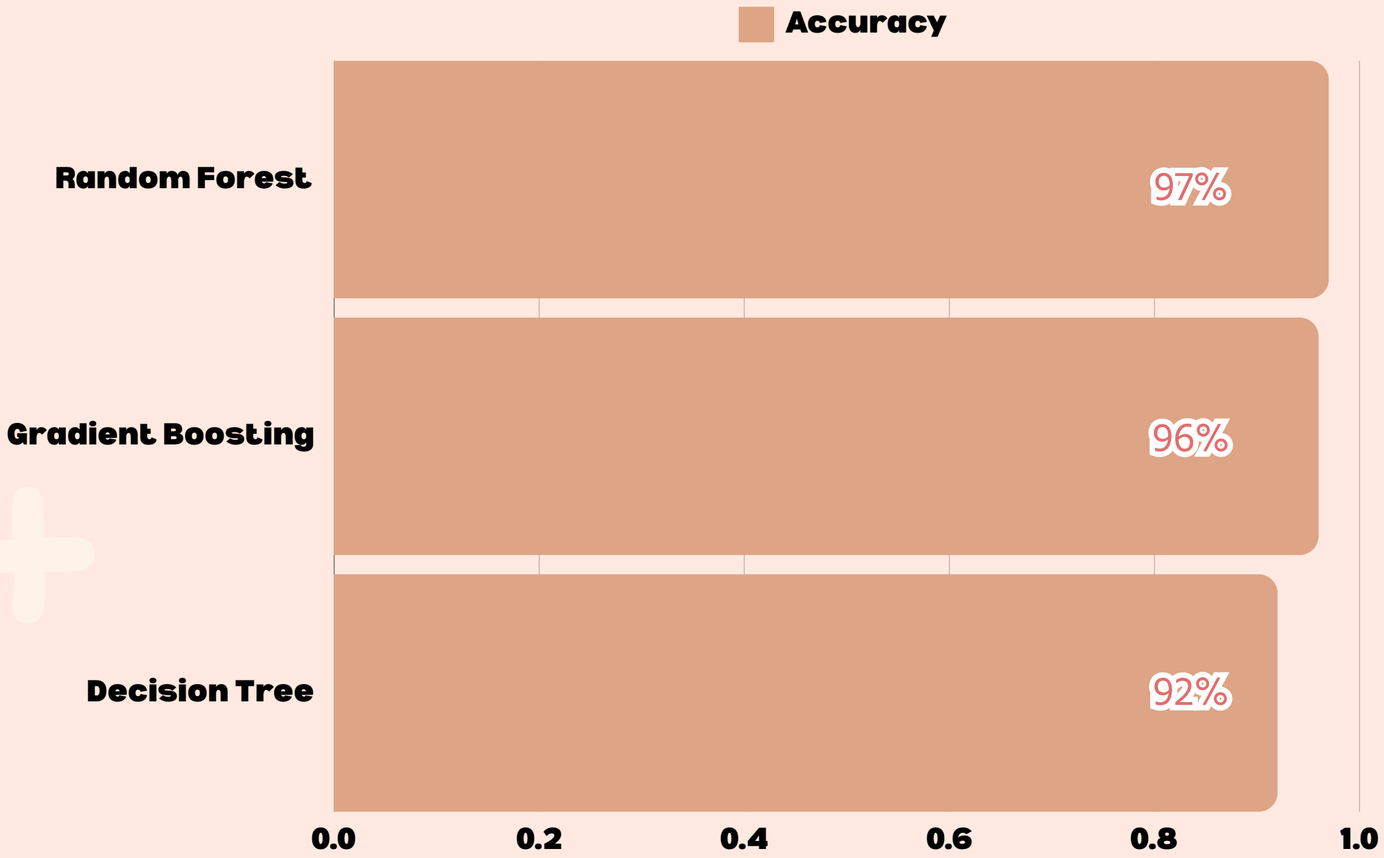
Decision Tree

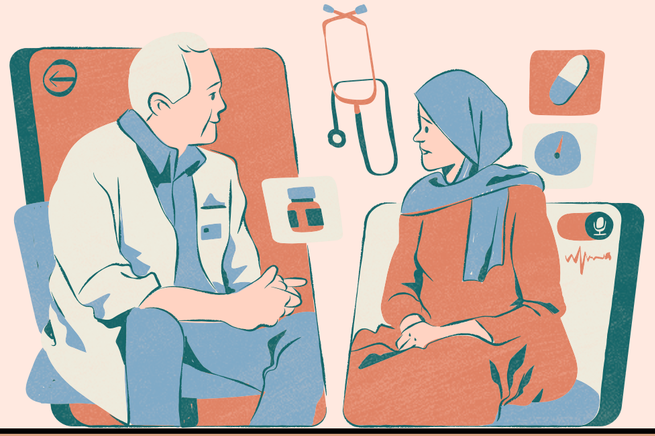


- **F1-Score**: 91%
- **Recall**: 92%
- **Precision**: 91%

RESULTS

Accuracy Comparison Across Models



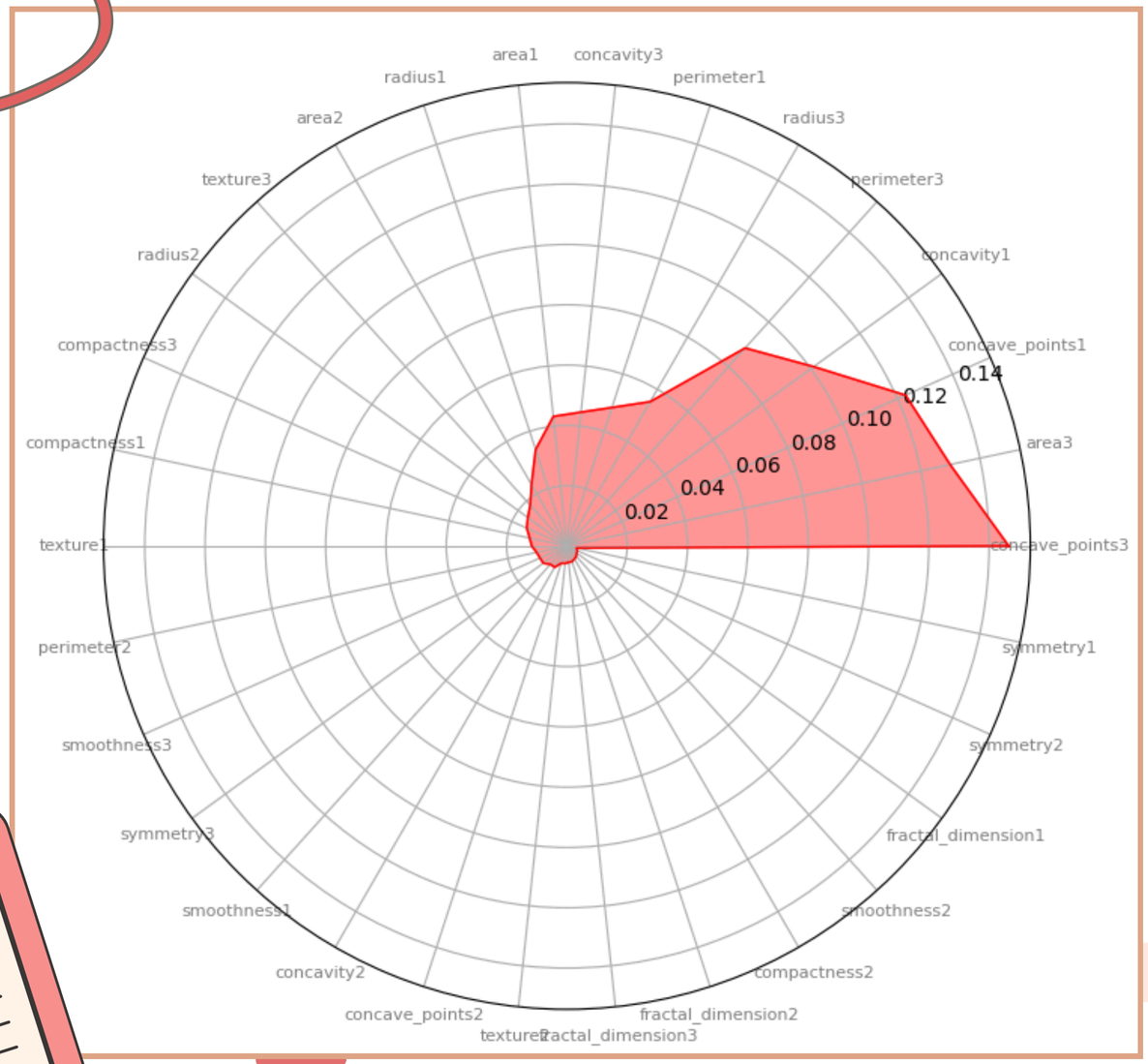


Random Forest has the **highest** accuracy, which is 97%.

RESULTS

Feature Importance

Concave_points3 is the feature (variable) with the **greatest contribution in the model** used (Random Forest), with an importance value **of 0.147**.



CONCLUSION

Random Forest is the chosen model for **breast cancer diagnosis classification** with an **accuracy of 97%**.

The **feature importance** value of 0.147 means that **Concave_points3** is one of the most important features used by the model to **classify the tumor as benign or malignant.**

THANK YOU FOR YOUR ATTENTION