Breast Cancer Diagnosis: Data Analysis and Model Report

Objective

The primary goal of this analysis is to predict the diagnosis of breast cancer (benign or malignant) based on various tumor characteristics. The workflow involves exploratory data analysis (EDA), data pre-processing, model training, evaluation, and feature importance assessment.

Dataset Overview

- **Source:** Provided CSV file (assumed to contain medical tumor data).
- **Features:**
  - Mean characteristics of tumors: radius, texture, perimeter, area, smoothness.
- **Target Variable:** `diagnosis` (categorical: benign or malignant).

Initial Data Insights

- **Head:** The first five rows were inspected to understand the dataset structure.
- **Info:** Data types and non-null values were verified.
- **Describe:** Summary statistics were reviewed for numerical columns.

Data Preprocessing

1. **Missing Values:**
   - No missing values were found.
2. **Feature Scaling:**
   - Standardized numerical features (`mean_radius`, `mean_texture`, `mean_perimeter`, `mean_area`, `mean_smoothness`) using `StandardScaler`.

Model Development

Model: Random Forest Classifier

- Random Forest is a robust, ensemble-based algorithm known for handling non-linear relationships and feature importance evaluation.

Data Splitting:

- **Training Set:** 80% of the data.
- **Testing Set:** 20% of the data.

Model Training:

- The model was trained on the scaled training set.

Model Evaluation

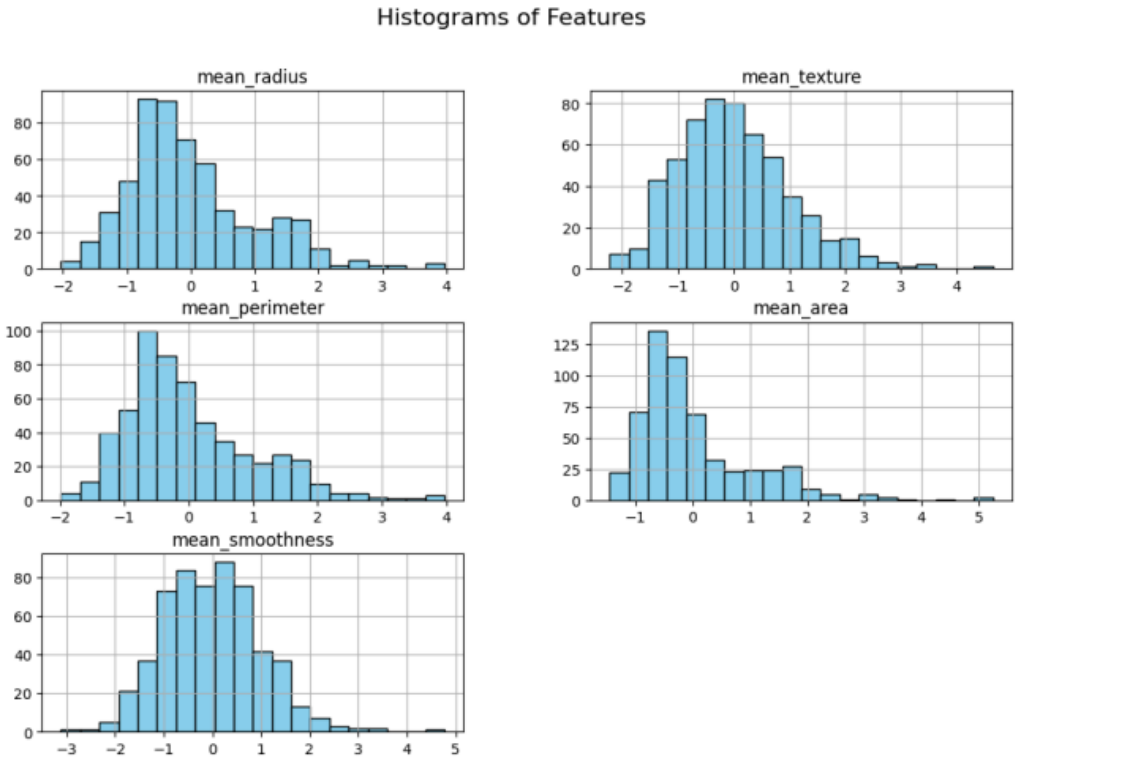
- **Accuracy:** Achieved a high accuracy score of **95%**
- **Classification Report:**
  - Provided detailed metrics:
    - **Precision:** Percentage of correct positive predictions.
    - **Recall:** Sensitivity of the model to positive cases.
    - **F1-Score:** Balance between precision and recall.
- Key Insight: The model performs well, with minimal misclassification.

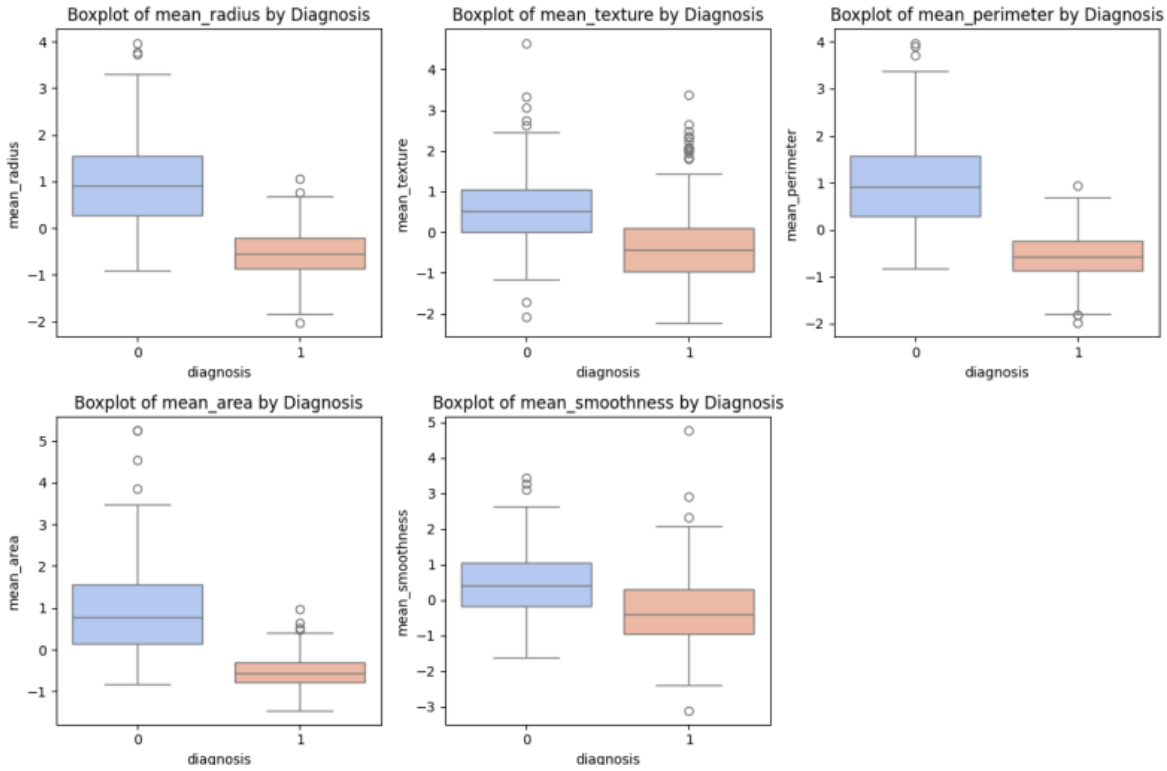
Feature Importance

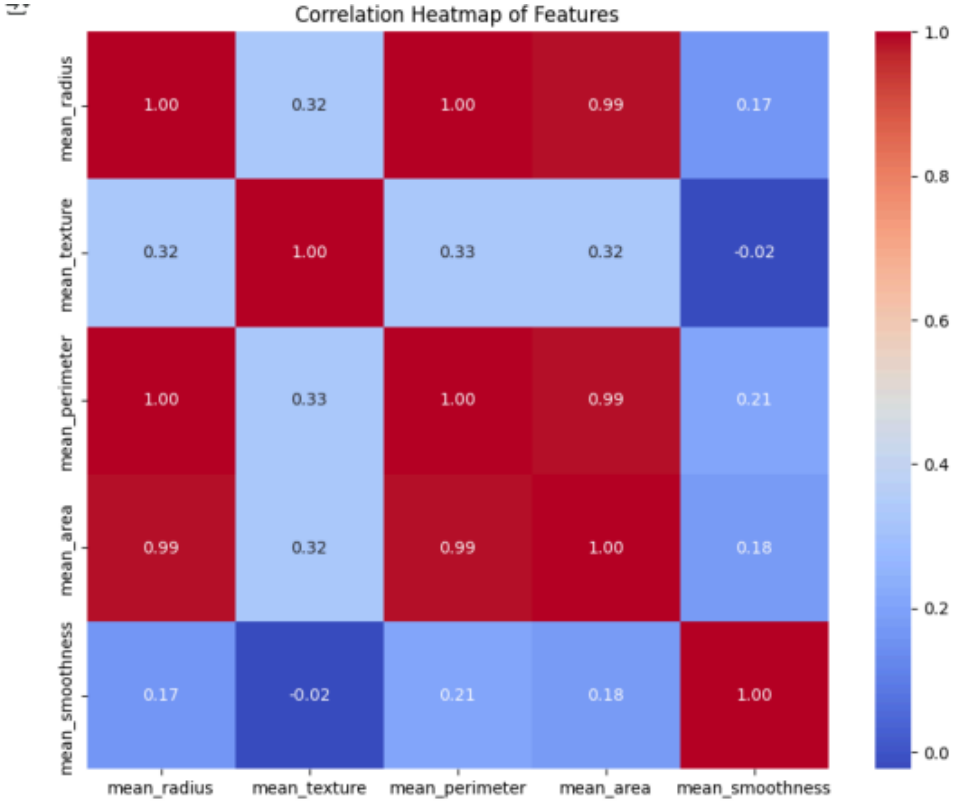
- The top features contributing to the model were visualized using a bar chart.
- Key Insight: Certain features, such as `mean_radius` and `mean_area`, were highly important for distinguishing between benign and malignant cases.

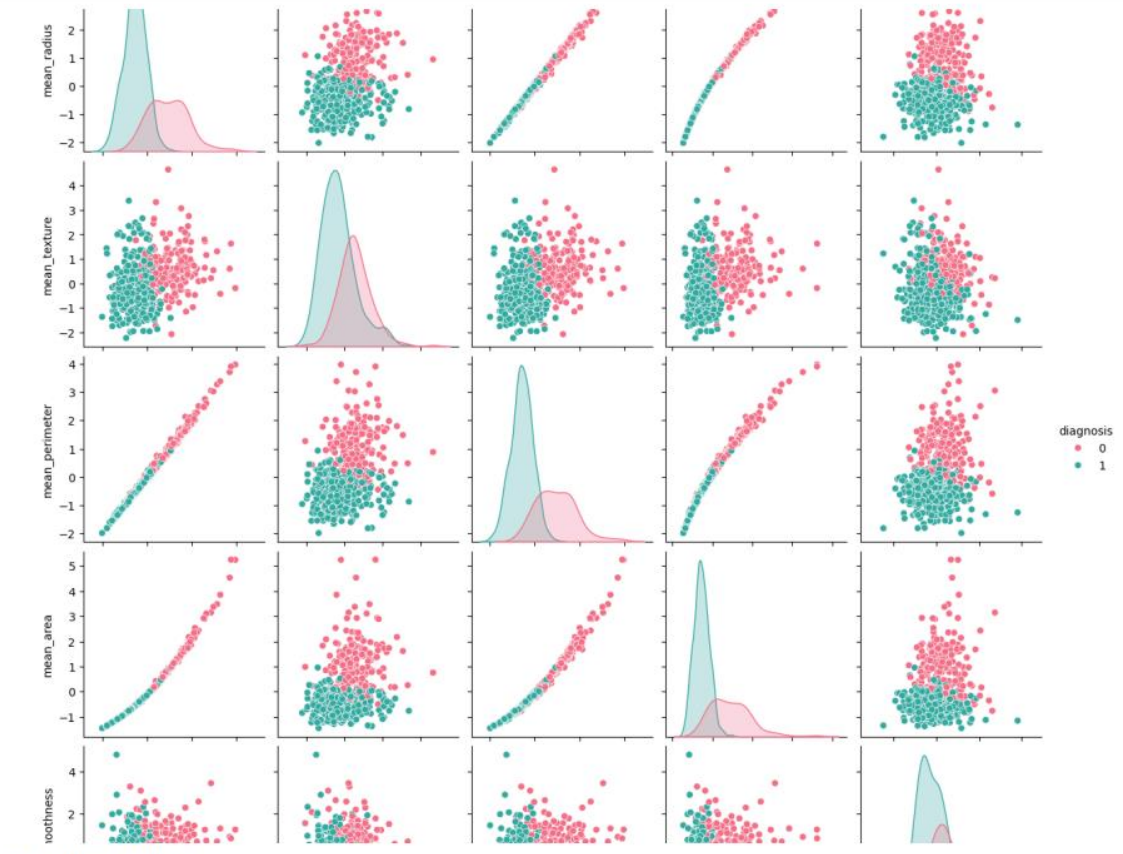
Key Insights

1. **EDA Results:**
   - Tumor characteristics show significant variation based on diagnosis.
   - High correlations between features suggest some redundancy.
2. **Model Performance:**
   - The Random Forest Classifier showed strong predictive ability, suitable for deployment.
3. **Feature Importance:**
   - Certain features dominate in importance, providing potential areas for further research or simplified modeling.

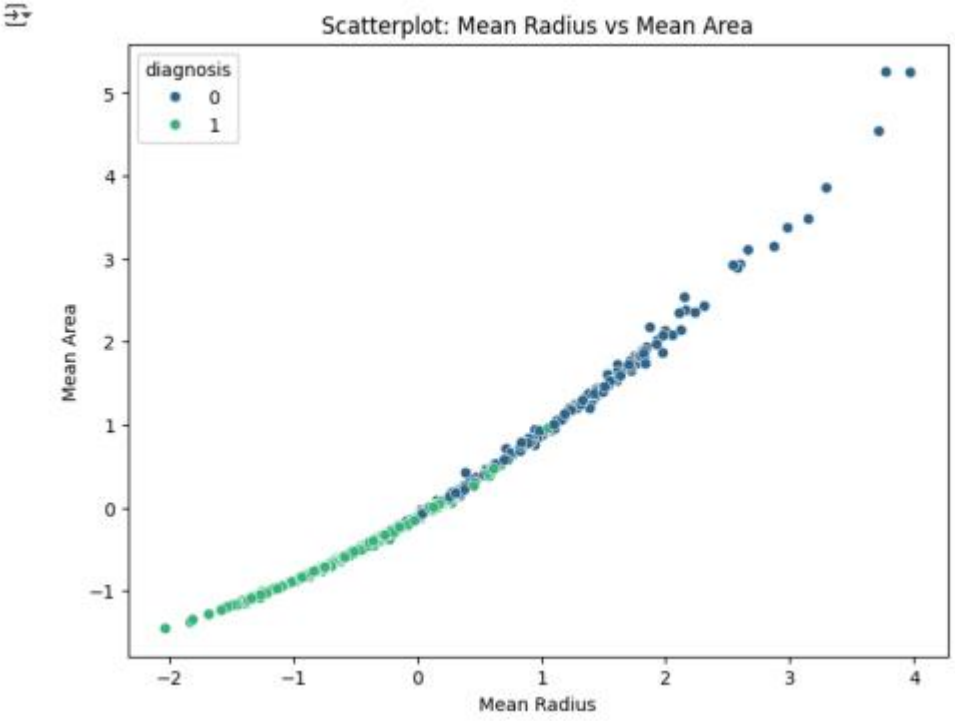

Histograms of Features
mean_radius
mean_texture
mean_perimeter
mean_area
mean_smoothness


Boxplot of mean_radius by Diagnosis
Boxplot of mean_texture by Diagnosis
Boxplot of mean_perimeter by Diagnosis
Boxplot of mean_area by Diagnosis
Boxplot of mean_smoothness by Diagnosis
diagnosis





Scatterplot for specific patterns


Scatterplot: Mean Radius vs Mean Area
diagnosis
0
1
Mean Area
Mean Radius
5
4
3
2
1
0
−1
−2
−1
0
1
2
3
4