TITLE: BREAST CANCER PREDICTION USING MACHINE LEARNING

NAME: KELLIE NJOKI NDARU

DATE: 30/08/2024

INTRODUCTION

Project Objective: To develop a model to predict whether a breast tumor is benign or malignant.

> Dataset Used: Breast Cancer Wisconsin (Diagnostic) Dataset

> Stakeholder: Healthcare providers, patients

PROBLEM STATEMENT

> **Problem**: Early and accurate diagnosis of breast cancer is crucial for effective treatment and patient survival.

➤ **Goal**: To build a reliable model that can assist in early detection of malignant tumors.

DATA OVERVIEW

- Dataset description:
- o 569 samples, 31 features

> Target Variable:

Diagnosis (B=Benign, M=Malignant)

DATA PREPROCESSING

Steps taken:

- Dropped irrelevant columns (ID column)
- Handled missing columns (dropped unamed32 column)
- Feature scaling
- > Split the data into training and testing sets

MODEL SELECTION

> Models Evaluated: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting.

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score.

MODEL PERFORMANCE

> Logistic Regression:

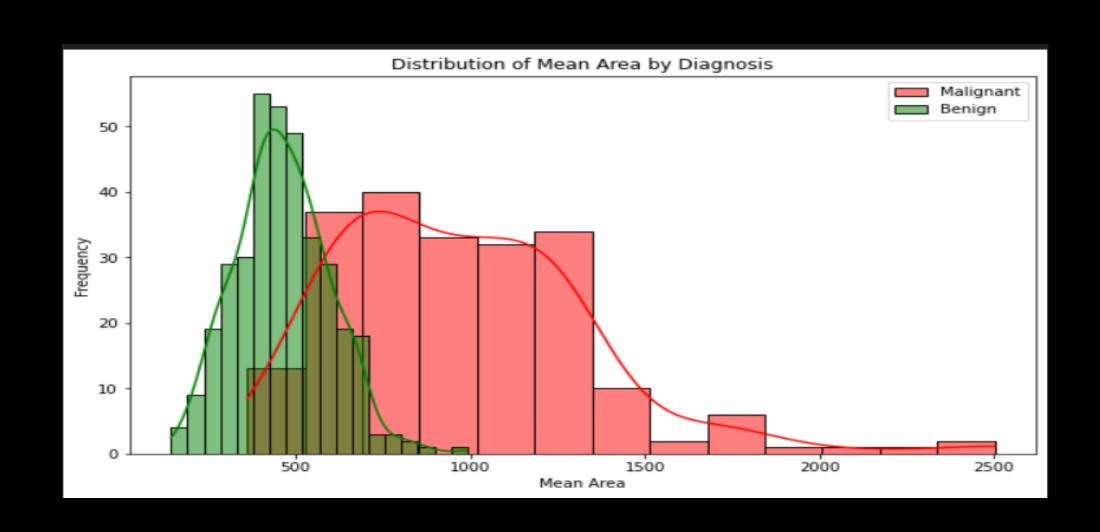
Accuracy: 98%

Precision and Recall: High performance in both classes.

F1-Scores: This indicates a better balance between precision and recall.

VISUALIZATIONS

DISTRIBUTION OF FEATURE VALUES FOR EACH DIAGNOSIS



CORRELATION HEATMAP

		Correlation Heatmap of Features													
texture_worst -	1	0.4	0.4	0.4	0.3	0.3	0.4	0.2	0.1	0.9	-0.08	0.3	0.2	0.3	0.2
concavity_worst -	0.4	1	0.6	0.9		0.5	0.6	0.5	0.6	0.3	0.04	0.9	0.4	0.5	0.4
radius_worst -	0.4	0.6	1	0.8	0.8	1	1	0.2	0.2	0.4	-0.1	0.7	0.7	1	0.8
concave points_worst -	0.4	0.9	0.8	1	0.9		0.8	0.5	0.5	0.3	-0.03	0.9	0.5	0.7	0.5
concave points_mean -	0.3	0.8	0.8	0.9	1	0.8	0.9	0.5	0.5	0.3	0.1	0.9	0.7	0.8	0.7
area_worst -	0.3	0.5	1	0.7	0.8	1	1	0.2	0.2	0.3	-0.1	0.7	0.8	0.9	0.8
perimeter_worst -	0.4	0.6	1	0.8	0.9	1	1	0.2	0.3	0.4	-0.1	0.7	0.7	1	0.8
smoothness_worst -	0.2	0.5	0.2	0.5	0.5	0.2	0.2	1	0.2	0.08	-0.01	0.4	0.1	0.1	0.1
compactness_se -	0.1	0.6	0.2	0.5	0.5	0.2	0.3	0.2	1	0.2	0.4	0.7	0.4	0.2	0.3
texture_mean -	0.9	0.3	0.4	0.3	0.3	0.3	0.4	0.08	0.2	1	0.009	0.3	0.3	0.3	0.3
symmetry_se -	-0.08	0.04	-0.1	-0.03	0.1	-0.1	-0.1	-0.01	0.4	0.009	1	0.2	0.2	-0.1	0.1
concavity_mean -	0.3	0.9	0.7	0.9	0.9	0.7	0.7	0.4	0.7	0.3	0.2	1	0.6	0.7	0.6
radius_se -	0.2	0.4	0.7	0.5	0.7	0.8	0.7	0.1	0.4	0.3	0.2	0.6	1	0.7	1
radius_mean -	0.3	0.5	1	0.7	0.8	0.9	1	0.1	0.2	0.3	-0.1	0.7	0.7	1	0.7
area_se -	0.2	0.4	0.8	0.5	0.7	0.8	0.8	0.1	0.3	0.3	0.1	0.6	1	0.7	1
	orst -	orst -	orst -	iorst -	nean -	orst -	orst -	orst -	- s	ean -	, se -	ean -	- as s	ean -	a se -

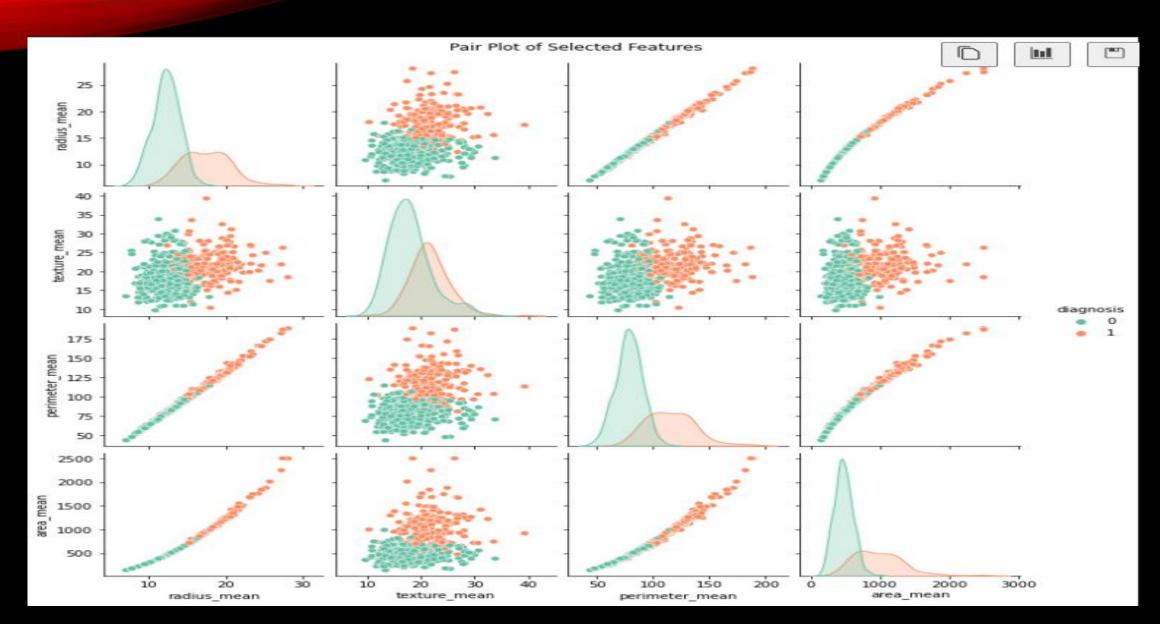
- 0.8

- 0.6

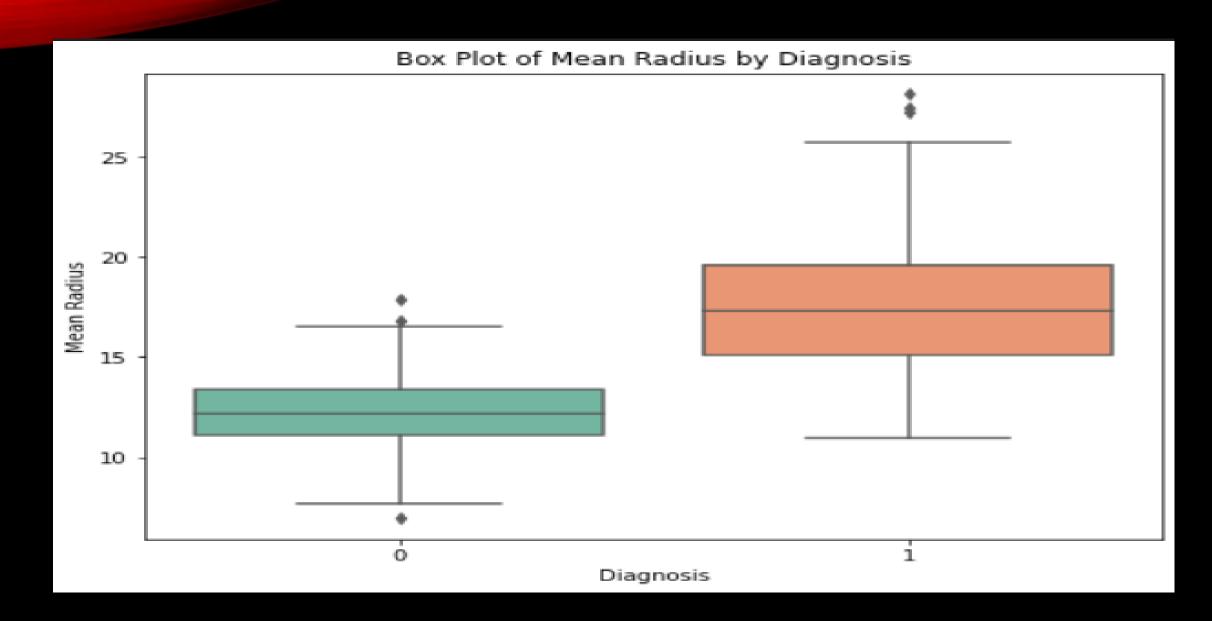
- 0.4

- 0.2

PAIR PLOT OF SELECTED FEATURES



BOX PLOT OF FEATURE VALUES BY DIAGNOSIS



RECOMMENDATIONS

> Deploy the Logistic Regression Model for real world use.

THANKYOU