Catalog

BREAST CANCER PREDICTION PROJECT REPORT
BY OLADELE AJAYI

1.0 Introduction

Breast cancer is one of the most common cancers among women worldwide. Early detection and accurate diagnosis are crucial for effective treatment and improved survival rates. This project aims to develop and evaluate machine learning models to predict breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset available at http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29.

2.0 Data Description

The WDBC dataset contains 569 instances of various features extracted from digitized images of breast mass. The dataset comprises 32 columns, including the ID number, diagnosis (M for malignant and B for benign), and 30 real-valued features computed for each cell nucleus.

Data Features:

- Mean, standard error, and worst (mean of the three largest values) of:
    - Radius
    - Texture
    - Perimeter
    - Area
    - Smoothness
    - Compactness
    - Concavity
    - Concave points
    - Symmetry
    - Fractal dimension

3.0 Data Cleaning

- The dataset was loaded and checked for null values. No missing values were found.
- The ID number column was dropped as it is not useful for prediction.

4.0 Exploratory Data Analysis (EDA)

- The distribution of the diagnosis was visualized using a count plot, showing a higher prevalence of benign cases compared to malignant ones.
- Histograms were plotted for each feature to understand their distributions.

5.0 Feature Engineering

The diagnosis column was encoded to numerical values (M = 1, B = 0) using LabelEncoder. Features were standardized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1.

6.0 Feature Selection

To enhance model performance, feature selection was performed by analyzing the correlation of each feature with the dependent variable (diagnosis). Features with low correlation were dropped to reduce noise and improve model accuracy. The correlation matrix was computed, and features with a correlation coefficient below a certain threshold were excluded from the final dataset.

7.0 Model Development

Several machine learning models were developed and evaluated for breast cancer prediction. These models include:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Support Vector Classifier (SVC)
- K-Nearest Neighbours (KNN) Classifier

8.0 Results and Discussion

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.964912	0.953488	0.953488	0.953488	0.96266
Decision Tree	0.938596	0.95	0.883721	0.915663	0.927776
Random Forest	0.95614	0.952381	0.930233	0.941176	0.951032
Gradient Boosting	0.95614	0.952381	0.930233	0.941176	0.951032
Support Vector Machine	0.95614	0.975	0.906977	0.939759	0.946446
K-Nearest Neighbour	0.964912	0.97561	0.930233	0.952381	0.958074

As presented in the table above, the performance of the various machine learning models for breast cancer prediction was evaluated using several metrics: accuracy, precision, recall, F1 score, and ROC AUC. The results reveal distinct strengths and weaknesses for each estimator.

8.1 Logistic Regression

Logistic Regression demonstrated the highest accuracy (0.964912) and performed well across all other metrics, with precision, recall, and F1 score all at 0.953488, and an ROC AUC score of 0.962660. This indicates that the model is highly reliable in distinguishing between benign and malignant cases, maintaining a good balance between correctly identifying positive instances and minimizing false positives.

8.2 Decision Tree Classifier

The Decision Tree Classifier had the lowest accuracy (0.938596) among the models, and although it achieved a high precision (0.950000), its recall was comparatively lower (0.883721), resulting in an F1 score of 0.915663. The ROC AUC score (0.927776) also lagged behind the other models, suggesting that while it can accurately identify malignant cases, it may miss some instances.

8.3 Random Forest Classifier

The Random Forest Classifier showed a strong performance with an accuracy of 0.956140, precision of 0.952381, recall of 0.930233, and an F1 score of 0.941176. Its ROC AUC score

(0.951032) indicates robust discriminative power. This model effectively balances precision and recall, making it a reliable choice for this prediction task.

8.4 Gradient Boosting Classifier

Gradient Boosting also achieved an accuracy of 0.956140, with precision and recall both at 0.952381 and 0.930233, respectively. Its F1 score matched that of Random Forest at 0.941176, and its ROC AUC score (0.951032) was identical. This model demonstrates that boosting techniques can enhance performance by reducing errors from previous models.

8.5 Support Vector Classifier (SVC)

SVC performed on par with Random Forest and Gradient Boosting in terms of accuracy (0.956140). It had a high precision (0.975000) but a slightly lower recall (0.906977), resulting in an F1 score of 0.939759. Its ROC AUC score (0.946446) was slightly lower, indicating that while SVC is excellent at correctly identifying malignant cases, it may produce more false negatives compared to some other models.

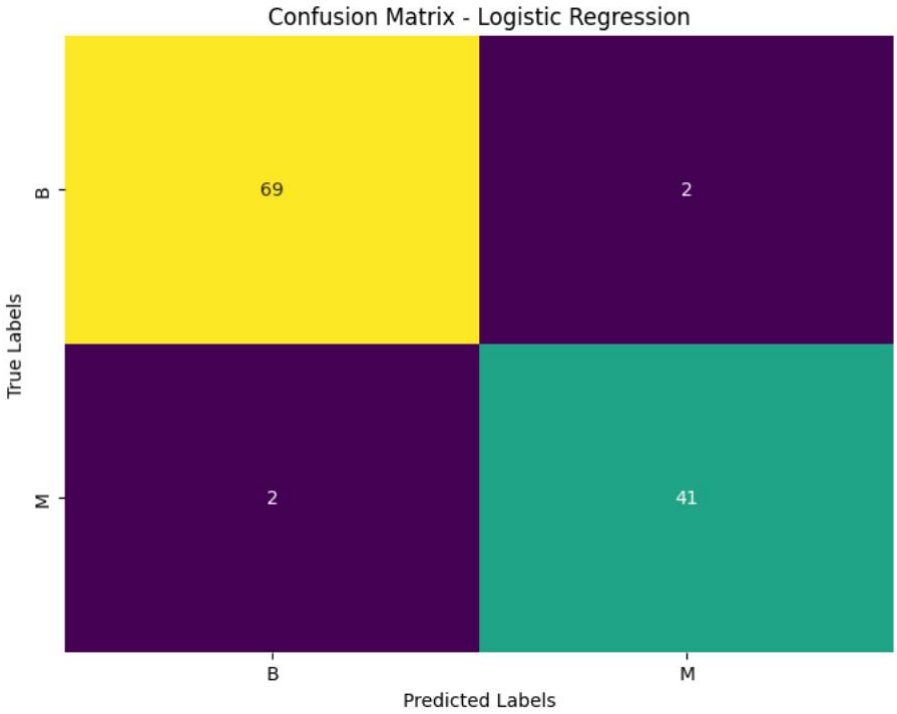
8.6 K-Nearest Neighbours (KNN) Classifier

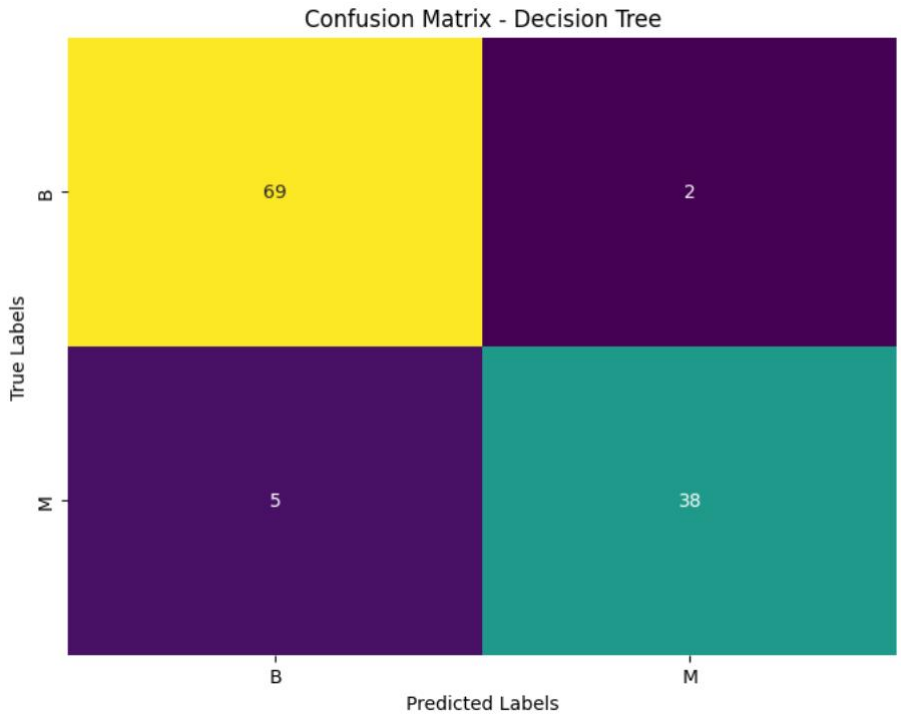
KNN showed high accuracy (0.964912), matching Logistic Regression. It also achieved the highest precision (0.975610) and a solid recall (0.930233), leading to an F1 score of 0.952381. The ROC AUC score (0.958074) was slightly lower than that of Logistic Regression, indicating strong overall performance but slightly less effectiveness in distinguishing between classes compared to Logistic Regression.

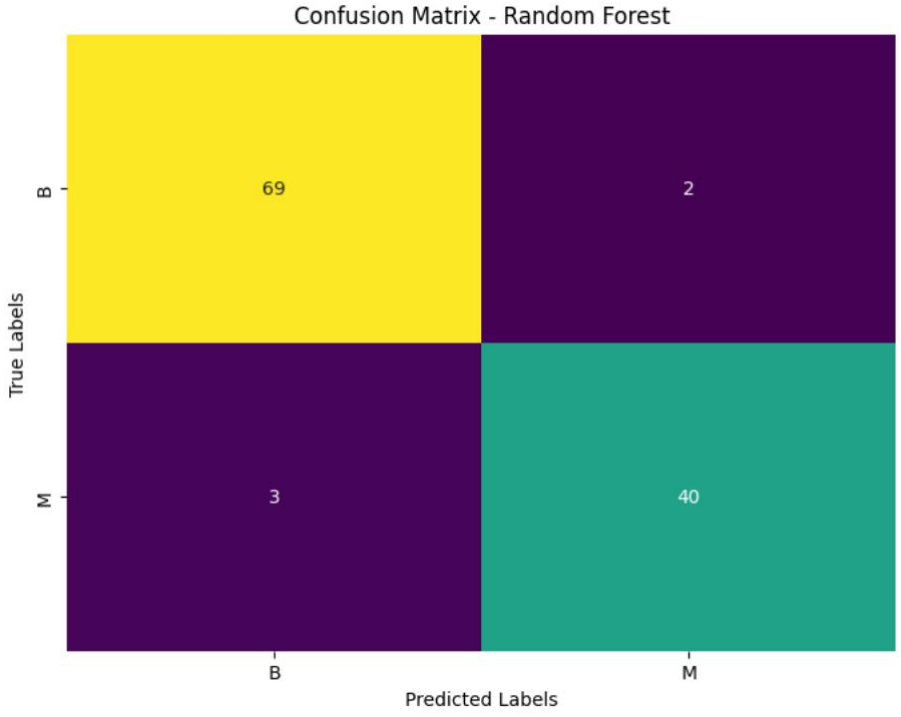
In summary, Logistic Regression emerged as the best overall model due to its superior balance across all evaluation metrics, particularly excelling in accuracy and ROC AUC. The model's simplicity and interpretability also contribute to its suitability for this task. While other models like Random Forest, Gradient Boosting, and SVC also showed strong performance, they did not surpass Logistic Regression in terms of overall effectiveness and efficiency.

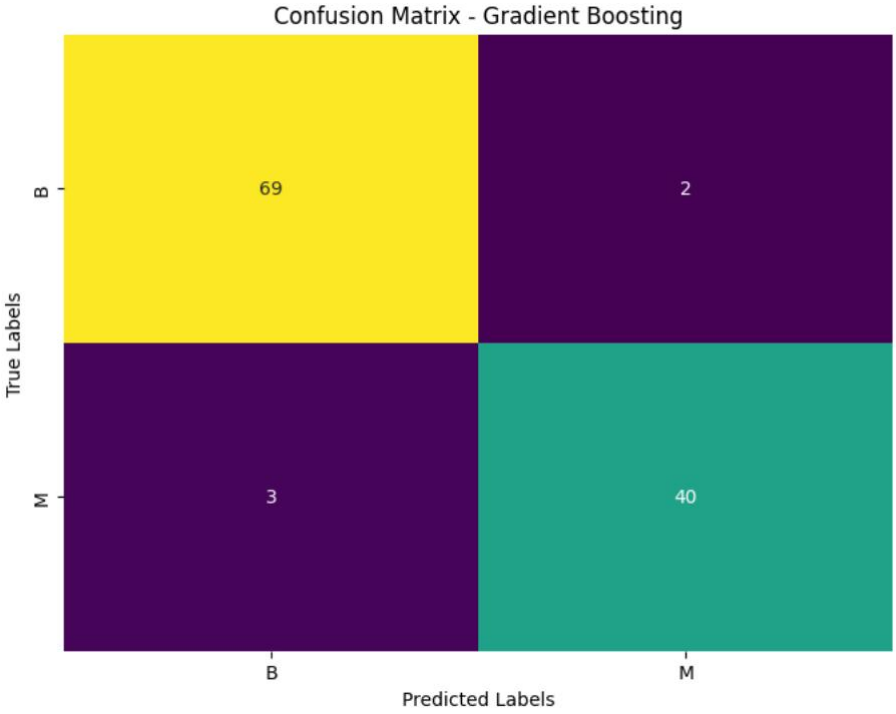
9.0 Confusion Matrix

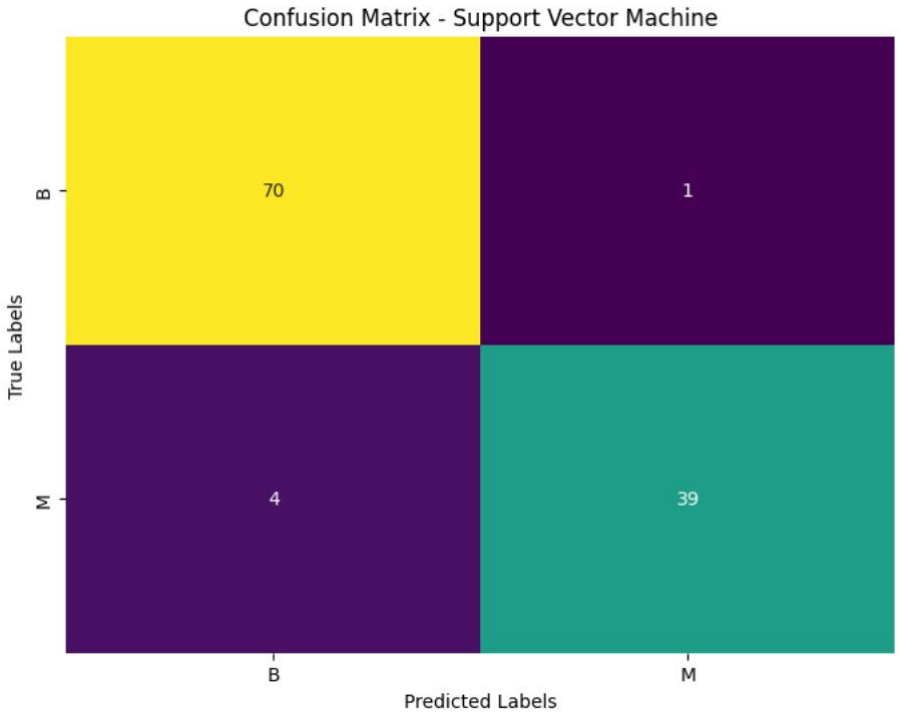
The following confusion matrix displays the instances of correct and incorrect predictions made by each model. In each matrix, the yellow cells represent the cases where individuals have benign conditions and were correctly predicted as benign by the model. The green cells on the diagonal represent the cases where individuals have malignant conditions (mammogram) and were correctly predicted as malignant by the model.

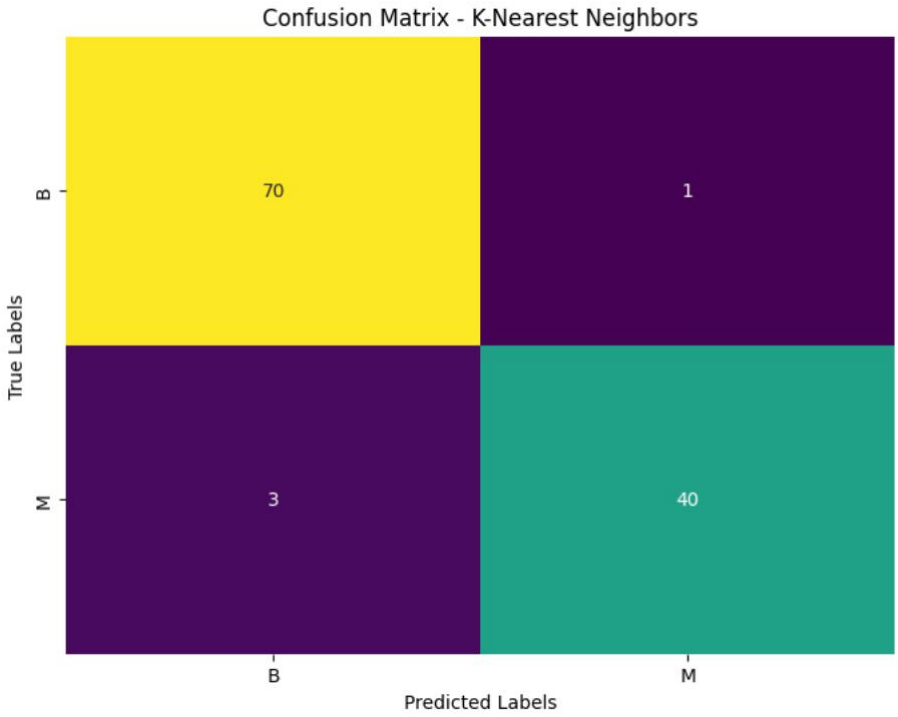



Confusion Matrix - Decision Tree
69
2
5
38
B
M
True Labels
B
M
Predicted Labels


Confusion Matrix - Random Forest
69
2
3
40
B
M
True Labels
B
M
Predicted Labels

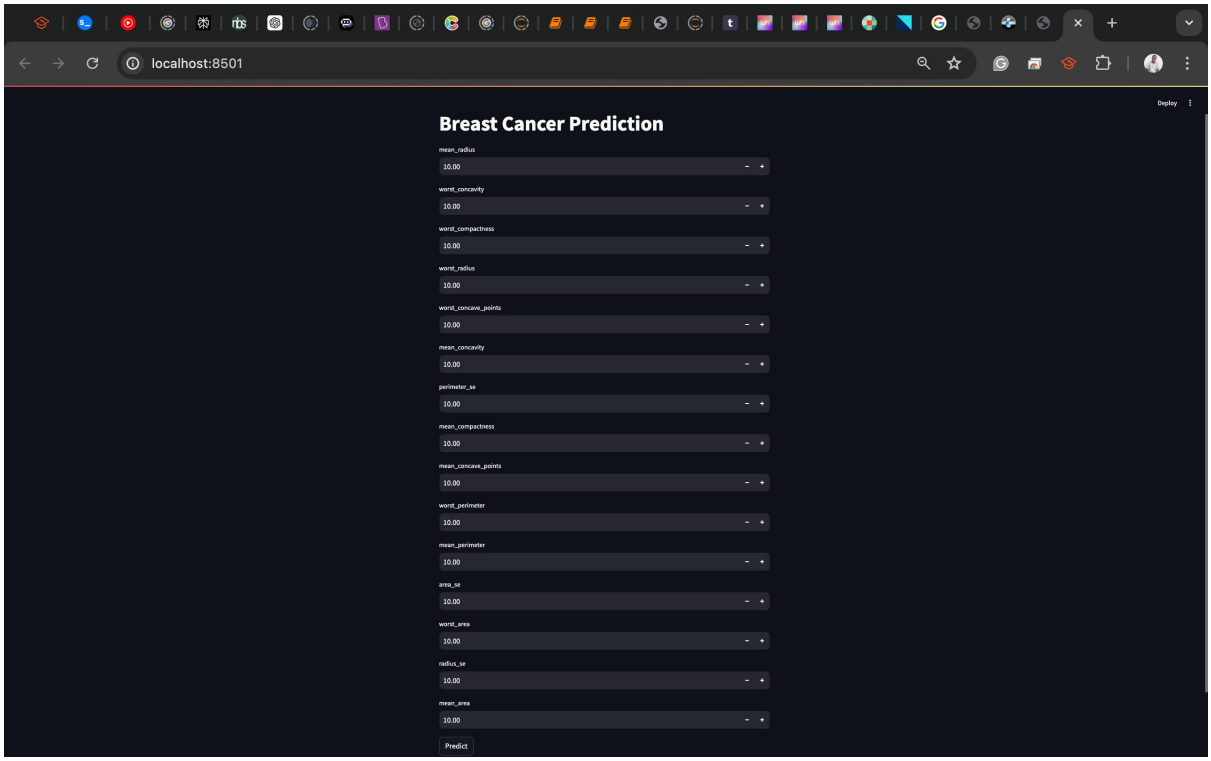

Confusion Matrix - Gradient Boosting
True Labels
B
M
69
2
3
40
B
M
Predicted Labels


Confusion Matrix - Support Vector Machine
True Labels
B
M
70
1
4
39
B
M
Predicted Labels



10.0 Model Deployment

The Logistic Regression model, which demonstrated the highest performance among the evaluated models, was successfully deployed using Streamlit. Streamlit is an open-source app framework that allows for the creation of interactive web applications. This deployment enables users to input relevant features and obtain real-time predictions on whether a breast cancer diagnosis is benign or malignant.



Model deployment