Breast Cancer Classification Project Report

1.Introduction

This project aims to develop a predictive model to classify breast cancer tumors as malignant (M) or benign (B) based on various features. The dataset used in this analysis is the Breast Cancer Wisconsin (Diagnostic) dataset.

2.Data Exploration and Cleaning

Explored the dataset's shape and found it to be (569, 33).

Checked for duplicate rows and found none.

Explored summary statistics for each column.

Checked for missing values and found none.

Dropped the 'id' and 'Unnamed: 32' columns as they are not informative.

Encoded the target variable 'diagnosis' using Label Encoder.

3.Data Visualization

Utilized various visualization techniques to understand the data distribution and relationships.

Plotted count plots, bar plots, box plots, histograms, and a pie chart.

Created a correlation heatmap to visualize feature correlations.

4.Building Models

Trained and evaluated the performance of different classification models:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- Naive Bayes

5.Model Evaluation

Utilized metrics such as accuracy, precision, recall, and F1-score for model evaluation.

Plotted confusion matrices for each model to visualize true positive, true negative, false positive, and false negative predictions.

6.Model Pipelines

Created pipelines for each model, including data preprocessing steps such as scaling.

7.Hyperparameter Tuning

Performed hyperparameter tuning using GridSearchCV for the following models:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- Gradient Boosting
- Naive Bayes

8.Model Comparison

Evaluated and compared the accuracy of each model after hyperparameter tuning.

9.Conclusion

Achieved the best accuracy with the Gradient Boosting model after hyperparameter tuning.

The Gradient Boosting model had the highest accuracy among the models considered.

The project provides a comprehensive overview of data preprocessing, model building, and evaluation techniques in the context of breast cancer classification.