NaiveBayesClassifierforBreastCancerDiagnosis

1 Introduction

The objective of this report is to implement a Naive Bayes classifier to predict whether a tumor is benign or malignant based on various features of cell nuclei present in the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset contains 569 instances with 30 numerical features and a target variable indicating the diagnosis: malignant (M) or benign (B).

2 Dataset Overview

The dataset, sourced from the UCI Machine Learning Repository, includes features like mean radius, mean texture, mean perimeter, mean area, etc., which describe the characteristics of the tumor. The target variable, *Diagnosis*, is a binary classification where 'M' represents malignant and 'B' represents benign.

3 Preprocessing Steps

Several preprocessing steps were necessary to prepare the data for the model:

- **Loading the Data**: The dataset was loaded using `pandas` and the column names were assigned meaningful labels.
- **Handling Missing Data**: No missing values were found in the dataset.
- **Target Variable Transformation**: The target variable was mapped to numerical values: 0 for malignant and 1 for benign.
- **Dropping Irrelevant Columns**: The column containing the patient ID was dropped as it does not provide useful information for diagnosis prediction.

4 Data Visualization

To understand the distribution of the data, various visualizations were employed:

- A histogram of the target variable showed the distribution of benign and malignant cases.
- A scatter plot was created to observe the relationship between two key features, *mean radius* and *mean texture*, for benign and malignant cases.

5 Model Implementation

The model was built using the *Gaussian Naive Bayes* classifier from `scikit-learn`. The following steps were taken:

- **Train-Test Split**: The dataset was split into training and testing sets using a 78%-22% ratio.
- **Model Training**: The Naive Bayes model was trained on the training set.
- **Prediction**: The model predicted the diagnosis on the test set.

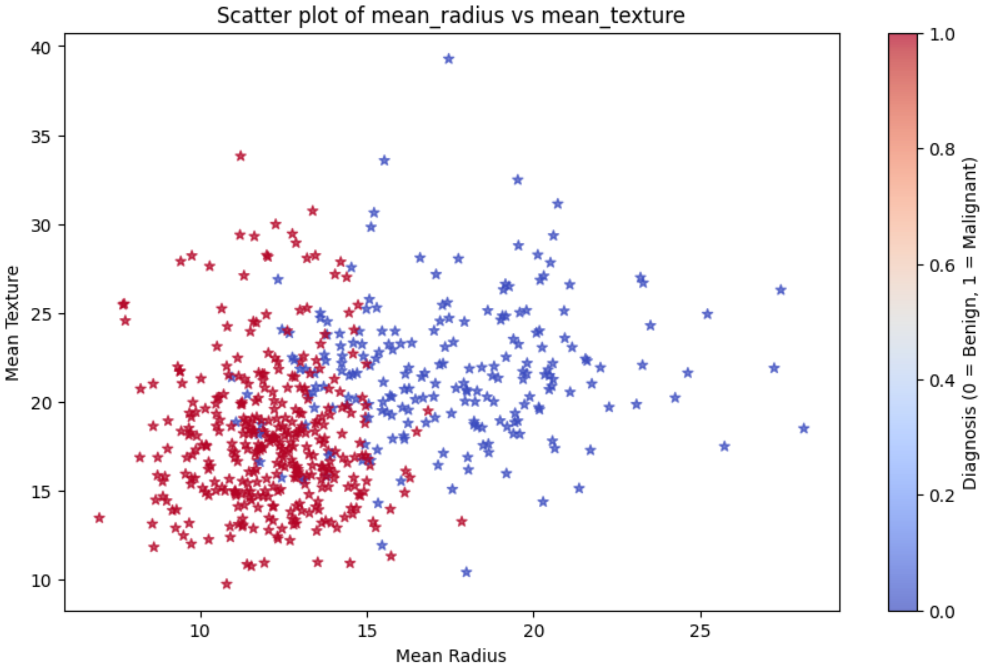

Figure 1: Scatter plot of mean radius vs mean texture.

6 Model Performance

The performance of the classifier was evaluated using the following metrics:

- **Accuracy**: The proportion of correctly classified instances among all instances.
- **Precision**: The proportion of correct positive predictions out of all predicted positives.
- **Recall**: The proportion of actual positives correctly classified.
- **F1 Score**: The harmonic mean of precision and recall, giving a balanced evaluation.
- **Confusion Matrix**: A table representing the true positives, true negatives, false positives, and false negatives.

The results were:

- **Accuracy**: 96.03%
- **Precision**: 95.65%
- **Recall**: 93.62%
- **F1 Score**: 94.62%
- **Confusion Matrix**:

$$\begin{bmatrix} 77 & 2 \\ 3 & 44 \end{bmatrix}$$

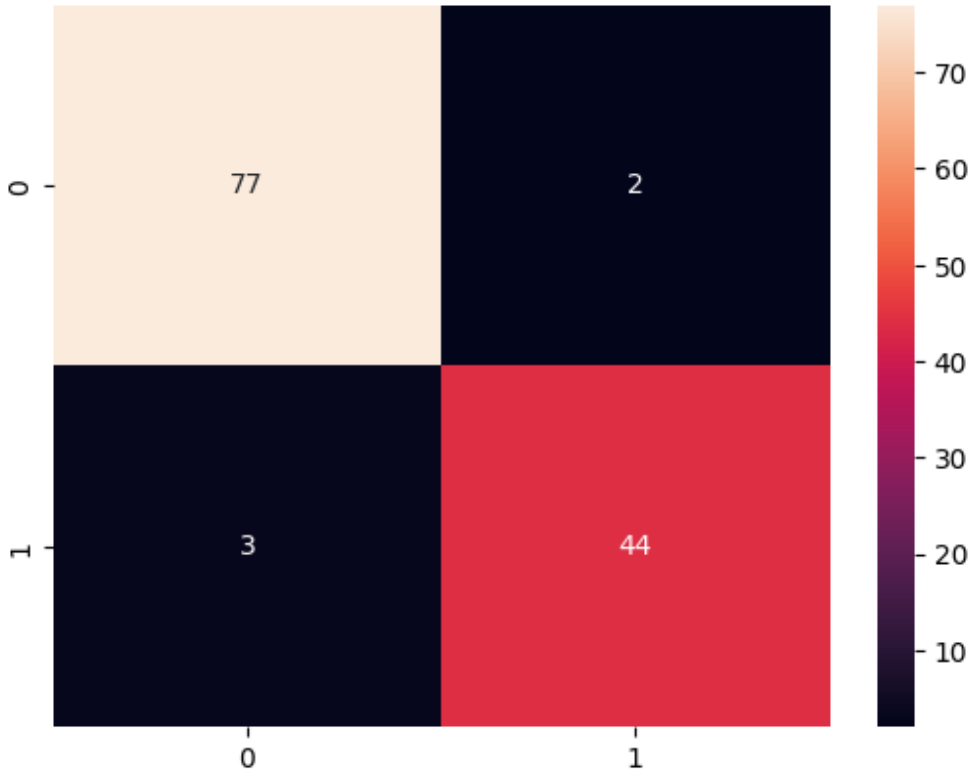

Figure 2: Confusion Matrix.

7 Insights Gained

The model demonstrated high accuracy, precision, and recall, indicating that the Naive Bayes classifier is effective in diagnosing breast cancer. Even though the dataset is slightly imbalanced (more benign cases than malignant), the model successfully identified the majority of malignant cases, achieving a recall of 93.62%. This means that the classifier is particularly strong in detecting actual cancer cases.

Additionally, feature exploration showed that characteristics like *mean radius* and *mean texture* provide good separability between benign and malignant cases.

8 Conclusion

In this report, we built and evaluated a Naive Bayes classifier to predict breast cancer diagnosis using the Breast Cancer Wisconsin (Diagnostic) dataset. The model achieved high accuracy and recall, proving to be an effective tool for medical diagnosis. Further work could explore alternative classifiers and feature selection methods to improve performance even more.