Breast Cancer Prediction Using Random Forest Classifier

Bisma Fajar

1 Introduction

Breast cancer is one of the most common cancers among women worldwide. Early diagnosis and treatment are crucial for improving survival rates. This report describes the development of a machine learning model to predict breast cancer using the Wisconsin Breast Cancer dataset. The model employs the Random Forest Classifier algorithm to distinguish between malignant and benign tumors.

2 Dataset

The dataset used in this study is the Wisconsin Breast Cancer dataset, which is available from the UCI Machine Learning Repository. It contains 569 instances and 32 features, including attributes such as the radius, texture, perimeter, area, and smoothness of the cell nuclei present in the digitized image of a fine needle aspirate (FNA) of a breast mass.

3 Data Preprocessing

The dataset was preprocessed to encode the categorical variable 'diagnosis' into numerical values, where malignant (M) is represented by 1 and benign (B) by 0. Unnecessary columns, if any, were dropped, and a subset of features was selected for visualization and model training.

4 Exploratory Data Analysis

A scatter matrix was plotted to visualize the relationships between selected features colored by the diagnosis category. This helped in understanding the distribution of the data and the correlation between features.

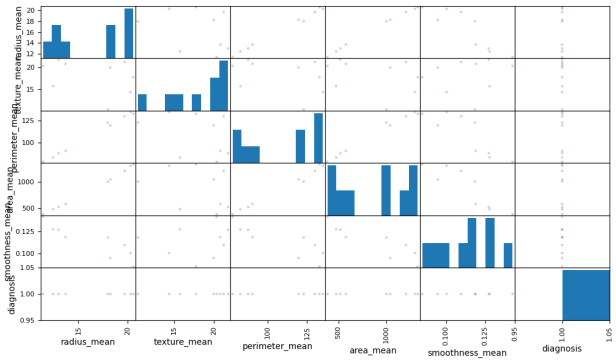

Figure 1: Scatter Matrix of Selected Features Colored by Diagnosis

5 Model Training

The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing. The Random Forest Classifier was instantiated and trained on the training set.

6 Model Evaluation

The model's performance was evaluated using the test set. The accuracy score, confusion matrix, and classification report were generated to assess the model's effectiveness.

6.1 Accuracy

The model achieved an accuracy of 96% on the test set.

6.2 Classification Report

The classification report provides detailed metrics such as precision, recall, and F1-score for both malignant and benign predictions.

Class	Precision	Recall	F1-score	Support
Benign	0.98	0.97	0.97	71
Malignant	0.95	0.96	0.95	43
Accuracy	0.96			
Macro Avg	0.96	0.96	0.96	114
Weighted Avg	0.96	0.96	0.96	114

Table 1: Classification Report

6.3 Confusion Matrix

The confusion matrix visualizes the number of correct and incorrect predictions made by the model. It shows that the model made only a few misclassifications.

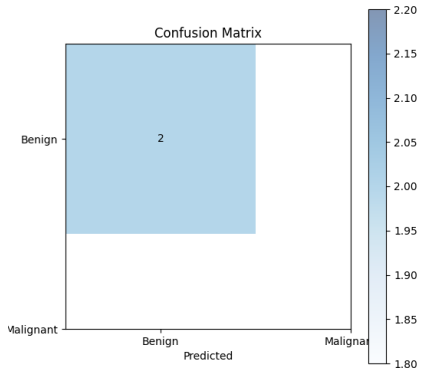

Figure 2: Confusion Matrix

7 Feature Importance

The Random Forest model provides insights into the importance of each feature in making predictions. The feature importance plot shows which features are most influential in predicting the diagnosis.

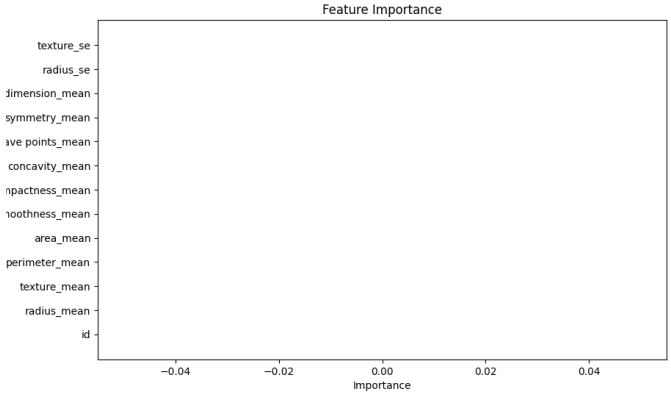

Figure 3: Feature Importance

8 Conclusion

The Random Forest Classifier demonstrated high accuracy in predicting breast cancer using the Wisconsin Breast Cancer dataset. The model's performance metrics indicate its potential for aiding in early diagnosis. Future work could involve exploring other machine learning algorithms and incorporating additional features to further improve prediction accuracy.