BREAST CANCER WISCONSIN DATASET ANALYSIS REPORT

INTRODUCTION AND DATA EXPLORATION

INTRODUCTION

This report details the approach and findings from the analysis of the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset, comprising features computed from breast mass images, is utilized to diagnose whether a breast mass is malignant or benign, presenting a binary classification problem.

DATA EXPLORATION

The initial phase involved loading the dataset and converting the 'diagnosis' column to a binary format, where 'M' (Malignant) was represented as 1 and 'B' (Benign) as 0. The dataset was then explored to understand its structure, features, and to check for missing data and outliers. This preliminary analysis was crucial for gaining insights into the dataset and guiding the subsequent preprocessing steps.

DATA PREPROCESSING, MODEL BUILDING, AND EVALUATION

DATA PREPROCESSING

During preprocessing, missing values were addressed, and feature standardization was performed to normalize the data. This step ensured that the models would not be biased towards variables with larger scales.

MODEL BUILDING AND SELECTION

For the classification task, three models were chosen: Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier. Each model was trained on the dataset, with their performance evaluated based on accuracy, precision, recall, and F1-score.

MODEL EVALUATION

The models were rigorously evaluated using a variety of metrics. The evaluation focused on not just accuracy but also on how well each model could balance precision and recall, especially important in medical diagnosis contexts.

MODEL TUNING AND CONCLUSION

MODEL TUNING

Hyperparameter tuning was conducted for each model using GridSearchCV. This process was crucial for optimizing model performance by finding the best combination of parameters. The best parameters for each model were identified and used to retrain the models, leading to improved performance.

CONCLUSION

The analysis of the Breast Cancer Wisconsin (Diagnostic) dataset demonstrated the effectiveness of machine learning in medical diagnosis. The chosen models, after careful evaluation and tuning, showed promising results in classifying breast masses as malignant or benign. This report highlights the importance of thorough data preprocessing, careful model selection, and rigorous evaluation in building effective machine learning models for critical applications like medical diagnosis.