BREAST CANCER DIAGNOSIS (WISCONSIN) USING ML

Sharath Pai

1. Introduction

The Breast Cancer Wisconsin Dataset is a publicly available dataset containing information about 569 breast cancer patients. The dataset is used to train and evaluate machine learning models for classifying breast cancer patients as benign or malignant. The dataset contains data on 30 attributes for each patient, which can be broadly categorized into the following groups:

- Patient information (age, menopause status, tumor size)
- Cell characteristics (cell radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension)
- Tumor characteristics (tumor node status, tumor node caps, malignancy degree)
- Diagnostic information (patient's diagnosis, either benign or malignant)

2. Approach

This report aims to analyze the Breast Cancer Wisconsin Dataset using classification algorithms used in supervised machine learning techniques to classify breast cancer patients as benign or malignant. The following steps were taken:

- Data Preprocessing: The dataset was preprocessed to handle missing values, outliers, and create dummy variables for categorical variables.
- Exploratory Data Analysis (EDA): EDA was performed to understand the distribution of variables, relationships between variables, and identify any patterns or anomalies in the data.

- Model Training: Various machine learning algorithms were trained using cross-validation to identify the best performing model for classification.
- Model Evaluation: The best model was selected on the basis of accuracy of each model.

3. Findings

- EDA: The EDA revealed that some variables were useless since their values existed between 0 and 1 which won't have much impact on the model.
- Model training: We took a split ratio of 70:30 for train and test data respectively and used classification algorithms to find out the accuracy of the best model.
- Model Selection: Gradient boost and Adaptive boost emerged as the best performing model, achieving an accuracy of 97.83% on the testing dataset.

4. Rationale behind Model Choices

Gradient boost and Adaptive boost were chosen as the best performing models due to its robustness to outliers, its ability to handle high-dimensional data, and its strong performance in classification tasks. The model's ability to generalize well to the testing dataset further supports its suitability for this problem.

5. Code Implementation

https://github.com/Sharath1036/task-intern-career/tree/main/Breast%20Cancer%20Diagnosis

6. Conclusion

The analysis of the Breast Cancer Wisconsin Dataset using machine learning techniques demonstrated the effectiveness of boosting in classifying breast cancer patients accurately. The findings from this study can be used to develop improved diagnostic tools and treatment plans for breast cancer patients.