Cody Kesselring and Owen Rooff

Breast Cancer Diagnostics

Dataset Description
Dataset : Wisconsin Breast Cancer Dataset
The dataset contains 569 instances of 30 continuous, real-valued features (e.g., mean radius, texture, perimeter) in a csv file. Using these values we are predicting whether an instance is Malignant or Benign.

Implementation/Technical Merit
The dataset is relatively clean with no missing values, the primary challenge is model performance due to high correlation among several features. In a decision tree, this collinearity can lead to redundant splits. To combat this we will implement the following:

- Stratified Splitting for the initial test set.
- Handling Bootstrapping (to create N unique training/validation sets).
- Random feature subset selection at each node split using Entropy for split evaluation.
- Selecting the best trees based on validation set accuracy.

Since the dataset has 30 attributes, we will artificially reduce the number of attributes during the forest classifier fit process, at every node. Instead of considering all 30 available attributes, the algorithm will randomly select X attributes as candidates for the optimal split.

Potential Impact of Results
The results of this project offer a comparison of multiple machine learning techniques on a single dataset. This comparative analysis identifies the best model for high-stakes diagnostics, which is very important in the medical field as people's lives are affected.

The stakeholders include medical researchers and scientists who are interested in methods and the fine tuning behind them to accurately assist medical professionals diagnose a patient.Medical Practitioners and patients benefit from the models accuracy and reliability as they can trust the diagnostics.

Citations
Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset
Libraries: NumPy, Pandas, Scikit-learn