Assignment 3: Breast Cancer Diagnosis Classification Report

The objective of this report is to classify breast cancer diagnosis as malignant or benign using machine learning classifiers. The dataset used for this task contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Data Loading and Preprocessing:

- The dataset "Assignment3-Breast-Cancer-Diagnose.csv" was loaded using pandas.
- Features and target variable were separated.
- Label encoding was applied to the target variable to encode 'M' (malignant) and 'B' (benign) as numerical values.

Classifier Implementation:

Three classifiers were implemented for the binary classification task:

1. **Decision Tree:** A single decision tree classifier.
2. **XGBoost:** An ensemble learning method known for its performance and scalability.
3. **Gradient Boosting (HistGradientBoostingClassifier):** A gradient boosting classifier that handles missing values.

Model Training and Evaluation:

- The dataset was split into training and testing sets using a 80-20 split.
- Each classifier was trained on the training data and evaluated on the testing data.
- Accuracy scores were computed for each classifier to evaluate their performance.

Encountered Errors:

Initially, the dataset contained missing values (NaNs), which caused errors when training the classifiers. Therefore, I replaced K-Nearest Neighbors (KNN) with Random Forest. However, it was later replaced with HistGradientBoostingClassifier to ensure compatibility with missing values and avoid errors.

Conclusion

In conclusion, the implemented classifiers achieved satisfactory accuracy in diagnosing breast cancer as malignant or benign. The use of HistGradientBoostingClassifier ensured compatibility with missing values in the dataset, resolving errors encountered during implementation. Further

optimizations and hyperparameter tuning could potentially improve the performance of the classifiers.