Assignment 10: WCE Smart Forest

Objective: Build a variant of random forest!

Download the following dataset (same as assignment 9) :

https://www.kaggle.com/datasets/erdemtaha/cancer-data/data

1. Drop Id column
2. Use the Diagnosis column as the target with Classes B and M
3. Perform a test train split. 80% into train and 20% in test
4. Following manipulation is performed to increase the skew in the data (only for this assignment. This is not to be done in practice!)
    a. From train data
        i. Consider all the rows that has diagnosis label =M ,
        ii. Of these rows , remove random 120 rows with label M and append these rows into test data
5. Build 10 decision trees using feature bagging and sample bagging(if size of train data is N, choose N samples with replacement).
    a. Feature bagging does not mean restricting the number of input features to trees. Each tree is trained using full set of feature. Just at the time of node split , it does not check all features , but uses a random subset. Use '**max_features'** ***parameter of sklearn decision tree***
6. Combine feature importance of all the features from each tree either using simple avg or weighted avg with accuracy of the tree as a weight
    a. You can either use '**feature importances** **'** attribute of decision tree or compute permutation importance using ***sklearn.inspection.permutation_importance***
7. Shortlist the features to use and drop other features from train data
8. Train 10 tress again using shortlisted features
9. Build following two models with input to them as shortlisted features + outcome of 10 trees trained in prev step
    a. Logistic regression model
    b. New Master decision tree.
10. On test data :
    a. Just retain the data of shortlisted features
    b. Make predictions using those 10 decision tress
    c. Use those predictions plus shortlisted features as input and
        i. Predict the label using logistic regression model
        ii. Predict the label using the master decision tree
    d. Observe which of these two approaches has highest accuracy? And how much improvement they offer over the 10 decision treess

11 Note : We have amplified the skew in the training data, by reducing the size of minority class. Therefore, use class weights while training. While measuring accuracy keep an eye on recall of minority class.