Diagnosis of Disease using Random Forest: Tree Ensemble

Objective:

- To Understand Data preprocessing steps for Tabular data
- To Understand effect of scaling and feature selection
- To Understand Implementation, evaluation metrics and loss function of Random Forest Tree Ensemble

Key Terms: Random Forest Tree Ensemble, Tree plots Hyperparameter tuning, Roc Curve, Confusion Matrix, Classification report.

Dataset:

The Dataset is available publicly on kaggle and is licensed to use for study purposes. though I don't have any information about who is the creator of the dataset. For ease I have uploaded the dataset in the “Datasets” folder in this repository.

Assignment 1. Data Preprocessing

- To load the dataset
- Class Balance check
- Clean the Dataset (Handling structural errors, Null values, duplicates, outliers,...e.t.c.)
- Convert Ordinal or Categorical Data into Numeric
- Transform the Data (Normalization, Standardization, Scalings techniques)
- Feature Engineering (Feature generation and selection)
- Split the data into train test and validation using *train test split* from *Sk-learn library.*

Assignment 2. Implement and Train

- Design the model
- Hyperparameter Tuning
- Train the model and make predictions using Trained model
- Evaluate the model on test data

Assignment 3. Predictions and Results

- Plot RoC curve to analyze the report
- Plot confusion matrix to understand the results better.
- Print classification report specifying precision and recall.
- Prepare a Detailed report for future reference