DATA ANALYTIC

Method- Classification

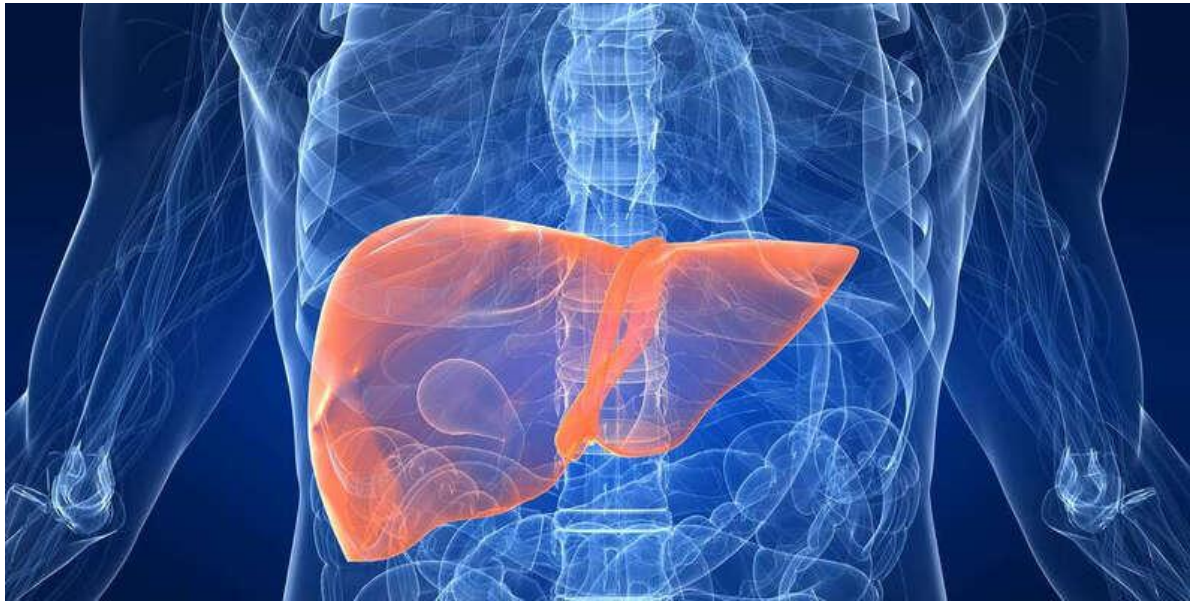
PROJECT - LIVER CIRRHOSIS STAGE CLASSIFICATION

University of cassino and southern Lazio

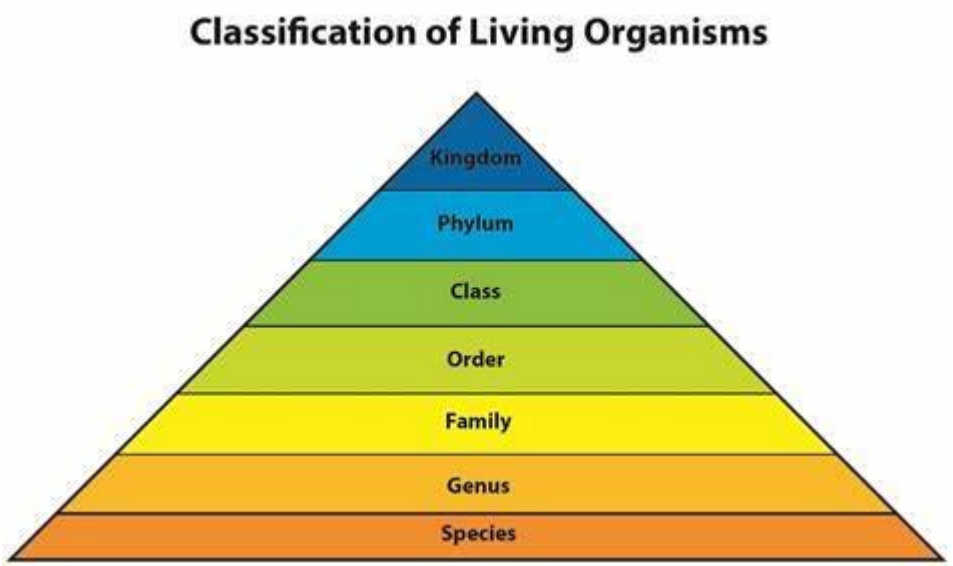
- Name- Dhrukumar Rajeshbhai valand
- Student id – 062957
- Subject- data analytics (L-33)
- DATA – 25-05-2024
- E-MAIL – dhrukumarrajeshbhai.valand@studentmail.unicas.it

About the dataset

- This dataset created for Cirrhosis results from prolonged liver damage, leading to extensive scarring, often due to conditions like hepatitis or chronic alcohol consumption. The data provided is sourced from a Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984.
- in these dataset we have 2500 observations and 19 variables
- The target variable is “STAGE”. Our task is to figure out if the patient is in stage 1, 2 or 3



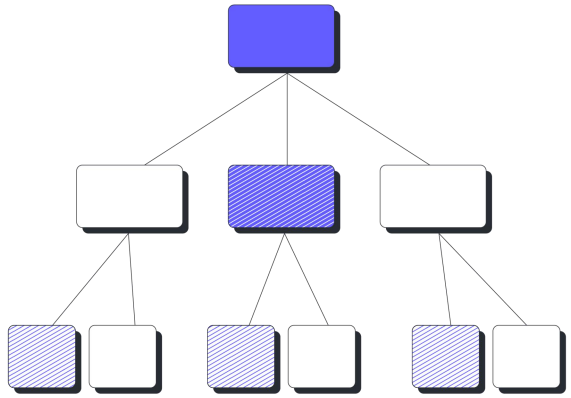
Classification



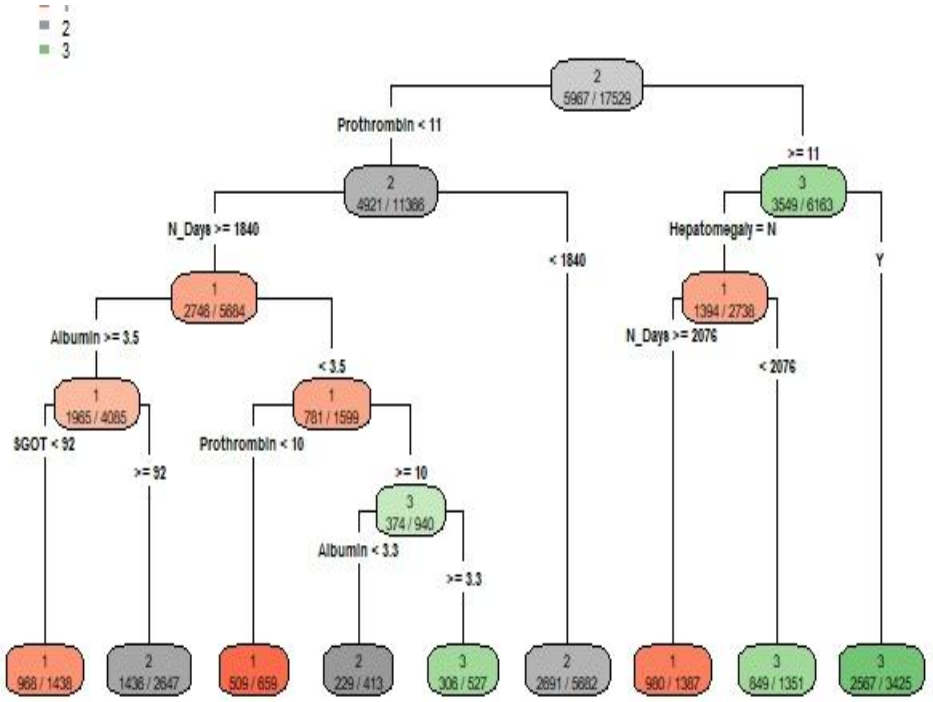
- Classification is a process in machine learning and statistics where items are assigned to predefined categories or classes based on their attributes. This involves building a model using a dataset with known labels (training data), which the model uses to learn patterns and relationships. Once trained, the model can predict the category of new, unseen instances based on their attributes.
- **Examples of Classification Algorithms**
- 1. Decision Trees
- 2. Random Forests

Decision tree

- Decision tree is a decision support model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
- Decision tree was build on rpart function, common way for building prediction model is like dividing the dataset into two
- 1) training
- 2) Testing
- Assigning some variables (characters) as factor(since its our Classification analysis) its better to assign them as factor.



ALGO 2 with Important variables Through “information gain”



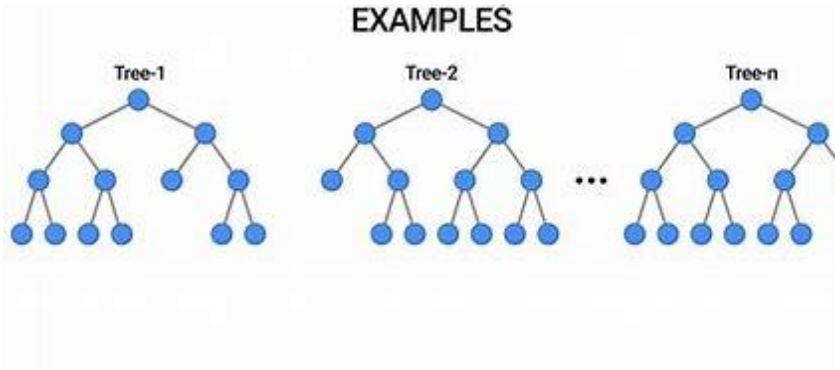
Algo 2 represents the algorithm of DT

it represents the model with optimum number of variables, obtained through information again. which eliminates the overfitting problem, aslo specified the minimum observations limit for considering a variable as important through “M insplit” function defined as “150” and finally split is based on Information again

some important attributes

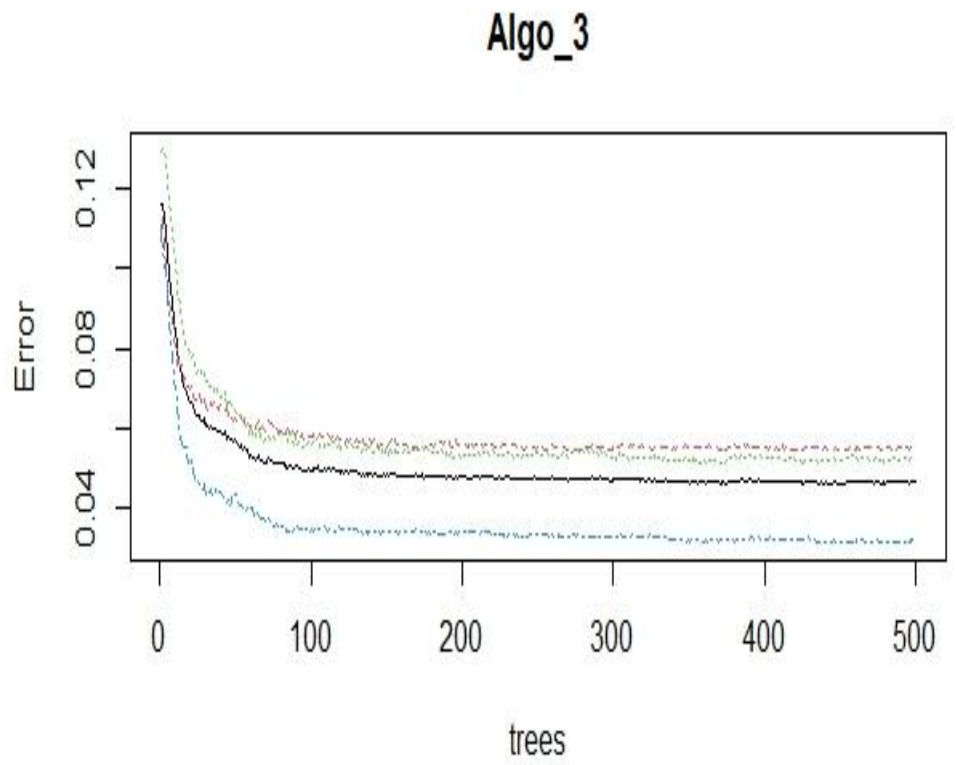
*** prothrombin**
*** N_days**
*** Albumin**
*** SGOT**

Random forest



- Random forest is a Supervised Machine Learning Algorithm. It builds decision trees on different samples and takes their majority vote for classification. It is easy to build and Its widespread popularity stems from its user-friendly nature and adaptability. Random Forest offers a combination of high accuracy, robustness, feature importance analysis, and scalability.
- Random forest provides high prediction accuracy especially when dealing with complex datasets and provides a measure of feature importance, which helps identify the most influential variables in the prediction task.
- In random forest 500 observation are there for the best result.

Algo 3 with all the variables and OOB error rate



Algo 3 represents the model with all the attributes make in a relation with target variables of "stage" and not specifiying the Number of trees

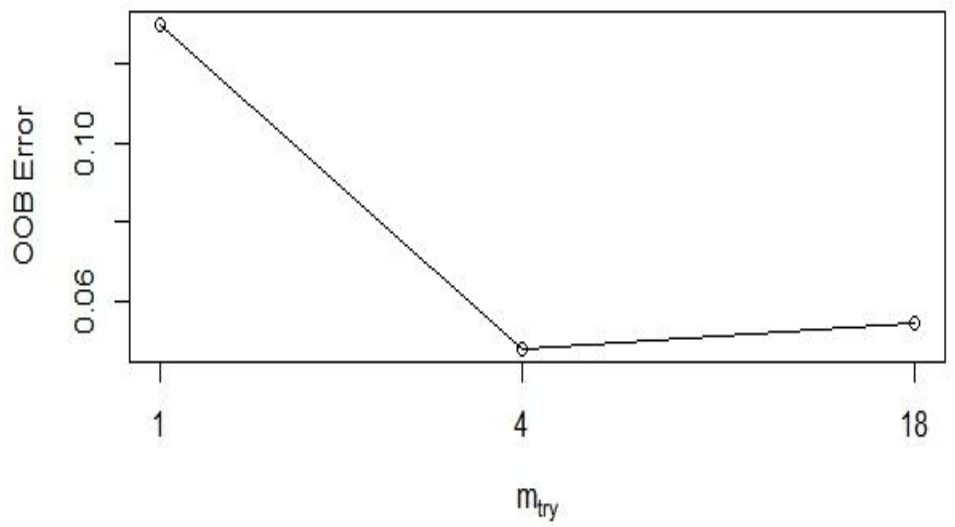
r**esults of the Algo_4 :-**

obb error rate = 4.65 %

the accuracy will be 93.35%

and trees and variables used by the model 500 and 4 variables at each split

Tunerf method for optimizing the Error rate

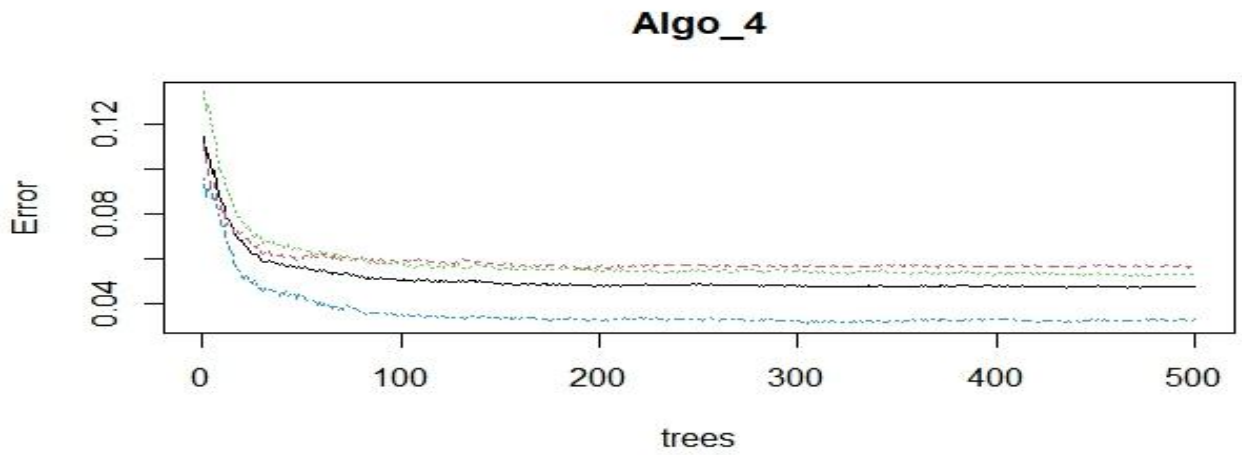


TUNERF Function :-

Tune Rf function is inbuild function of random forest . it assists the model to find the optimum number of variables to used for the model in order to reduce and control the stability of the model
*the graph represents OOB error on the X asix and
* on the y axis, it represents Mtry" number of variables tried at each split"

so the variables to try at split will be 4

Algo 4 with optimum of trees and variables



Algo 4 model OOb error rate:-

*the error line slopes downwards as the number of trees increases
*it's generally a good for the model . It indicates the model is improving its ability to predict the target variable accurately.
* the error is being stable after 200 trees

Conclusion :-

after comparing both models, the results of the model are pretty good in terms of performance and accuracy

DECISION TREE = 60.45 %

RANDOM FOREST = 95.26 %

Each algorithm works on different method.
Random forest have shown significant accuracy over Decision tree