Assignment 2- ML – Classification – Naïve Bayes & Decision Tree

Instructions:

a) "Learning is the Goal"… "NOT grades".
b) Students are expected to have good knowledge in feature engineering and classification algorithms. Revise the concepts before solving the problem.
c) It is an individual assignment, not a group activity.
d) You must provide complete solution, with analysis, not just answer. Right approach with appropriate explanation/analysis will be appreciated even though final answer is wrong.
e) Model tuning and evaluation are mandatory.
f) Submit the solution as a softcopy with the file name as "RollNumber_Name_Assignment3_ML_NB_DT.ipynb". No other format is allowed.
g) The solutions will be evaluated automatically using scripts. Strictly follow the instructions.

Problem Statement

The data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.

Any patient whose age exceeded 89 is listed as being of age "90".

Attributes/Columns:

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase

- Aspartate Aminotransferase
- Total Protiens
- Albumin
- Albumin and Globulin Ratio
- Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

Use these patient records to determine which patients have liver disease and which ones do not. Perform the following operations.

- **Analyze the data**
    - Find out if there are any attributes with correlation more than 0.50
    - Visualize the attributes and find out if there are any outliers using box plot or any other related plot.
    - Plot distribution plots and explain the dispersion.
- **Curate the data (if required)**
    - Identify the missing values and fill them with an appropriate method, if there are any missing values
- **Normalize all the attributes using an appropriate method.**
- **Build a disease classifier using Decision Tree (NB) and Naïve Bayes (NB) algorithms.**
    - Build the model & Perform 5-fold cross validation
    - Perform hyperparameter optimization using Random Search, Grid Search and TPE. Which method gives best model? Explain.

Assignment 2- ML – Classification – Naïve Bayes & Decision Tree

- Evaluate the model using accuracy, confusion matrix, precision, recall and F1-Scores
- Compare the performance of both the models (NB and DT)

- **Store the results into a csv file.**
- **Store the model. Print and explain the content of the model.**
- **Design a simple interface and deploy the model, which can take the required attributes of a person and predict whether he/she has liver disease.**