Practical Machine Learning

Lab Exercises: Classification

Q. 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 patients (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: (Label Column) 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:

1) Analyze the data

- Find out if there are any attributes with correlation more than 0.80
- Visualize the attributes using – Kernel density estimation (KDE/Histogram plots) – Write down your observations

1. **2) Curate the data (if required)**

- Identify the missing values and fill them with an appropriate method, if there are any missing values

3) Build a disease classifier using:

i. Decision Tree

ii. Naive Bayes

iii. Random Forest

iv. Gradient Boost

v. XGBoost

vi. SVM

vii. Logistic Regression

- Perform 5-fold cross validation

Evaluate all the models using – accuracy, precision, recall and F1-Score, FP

Compare the results in a table as shown below and write down your observations:

Table 1: Comparison of model evaluation metrics

Algorithm	Accuracy	Precision	Recall	F1-Score	False Positives
Decision Tree					
Naive Bayes					
Random Forest					
Gradient Boost					
XGBoost					
SVM					
Logistic Regression					

- **Tune all models using Hyperparameter Optimization and note improved results in following table**

Table 2: Model Evaluation metrics after Hyperparameter Optimization

Algorithm	Accuracy	Precision	Recall	F1-Score	False Positives
Decision Tree					
Naive Bayes					
Random Forest					
Gradient Boost					
XGBoost					
SVM					
Logistic Regression					