Data Science Project On

Liver Patient Prediction

PROJECT-ID = PRCP-1007-LiverPatientPred

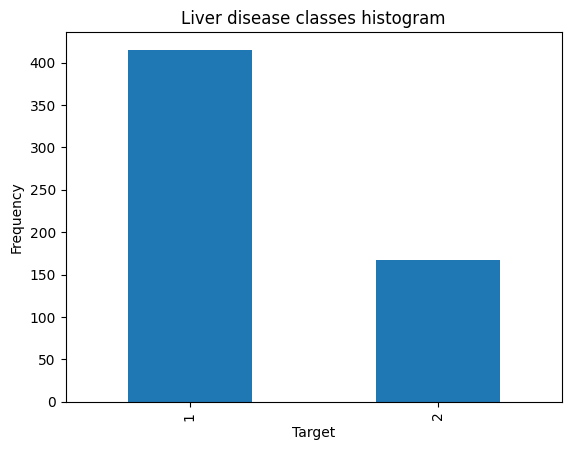
Business case:- Based on the given dataset we used to evaluate prediction algorithms in an effort to reduce burden on doctors.

Introduction:

With the problem of identification of the patients who have liver disease patients and who are not, the data set contains 416 liver patient records and 167 non \_liver patient records collected from North East of Andhra Pradesh, India

Domain Analysis & EDA:

* As we have target column it’s a supervised machine learning. Here we have 1 and 2 in target column 1 means patients have liver disease and 2 means patients doesn’t have liver disease.
* At First we have to change the above into 1 and 0 terms by using map.



* From the above we can clearly see that the most of the patients have liver disease.
* For univariate analysis I used Sweetviz, which is an open-source Python library, this is used for automated exploratory data analysis (EDA), it helps data analysts/scientists quickly generate beautiful & highly detailed visualizations. The output, we get is a fully self-contained HTML application. The system built reports around quickly visualizing the target values & comparing datasets.
* For bivariate analysis I used hist plot to show the distribution.

Data Preprocessing:

**1.Missing value:**

* Dataset have 4 missing values in **Albumin and Globulin Ratio.** We can drop the null values by using dropna().

**2.Categorical data:**

Dataset contain only one categorical column that is gender of patients, Here I used pandas get dummies for changing the categorical column.

**3.Outlier Handling:**

Here we have lot of outliers in different column, by using mean we can get rid of outliers in our dataset.

**4.Scaling:**

Here we use standard scaling for scale our data. It converts the data in range -3 to +3. Totally change the dataset in above range.

Feature Selection:

**1.Drop unique & Constant column:**

The dataset contains only one constant column.

**2.Checking correlation:**

* In this data no high correlated feature is present

**3.Check duplicates:**

* There are no duplicates present in data

Model Creation and Selection:

* Logistic regression model gets 85% model accuracy
* Random forest and decision tree also works very well with f1\_score more than 83%.
* But the SVM not perform that much good it have only 72% f1\_score.
* From the above we prefer the logistic model for better result of our model.

**NOTE:**

Hyperparameter tunning takes to much time.