Patients with liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. In an effort to reduce the burden on doctors, the government has hired you as a data scientist to build a predictive machine learning that would give an indication of whether a person would have a liver problem or not.

Now, as a data scientist, your goal is to build a logistic machine learning model that predicts whether a patient is healthy (non-liver patient) or ill (liver patient) based on some clinical and demographic features (or input variables) listed in the 'Data Description' section.

**Evaluation Criteria**

Submissions are evaluated using Accuracy Score. How do we do it?

Once you generate and submit the target variable predictions on evaluation dataset, your submissions will be compared with the true values of the target variable.

The True or Actual values of the target variable are hidden on the DPhi Practice platform so that we can evaluate your model's performance on evaluation data. Finally, an Accuracy score for your model will be generated and displayed

##### About the dataset

This data set contains liver patient records and non liver patient records collected from North East of Andhra Pradesh, India. The "Liver\_Problem" column is the target variable used to divide groups into liver patient ( Liver\_Problem == 1) or not ( Liver\_Problem == 2).

* Liver\_Problem == 1, implies the individual is a liver patient
* Liver\_Problem == 2, implies the individual is not a liver patient

To load the dataset in your jupyter notebook, use the below command:

import pandas as pd

liver\_data = pd.read\_csv('https://raw.githubusercontent.com/dphi-official/Datasets/master/liver\_patient\_data/indian\_liver\_patient\_dataset.csv')

**Data Description:**

* 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
* "Liver\_Problem" column is the **target variable** used to divide groups into liver patient (liver disease) or not (no disease).

Some of the data are slightly technical, you may refer online resources to learn more about them.

##### Evaluation Dataset

Load the evaluation dataset (name it as 'eval\_data'). You can load the data using the below command.

eval\_data = pd.read\_csv('https://raw.githubusercontent.com/dphi-official/Datasets/master/liver\_patient\_data/indian\_liver\_patient\_new\_testdataset.csv')

Here the Liver\_Problem column is deliberately not there as you need to predict it.

##### Reference:

This dataset is downloaded from the UCI Machine Learning Repository.