DATA PRE-PROCESSING

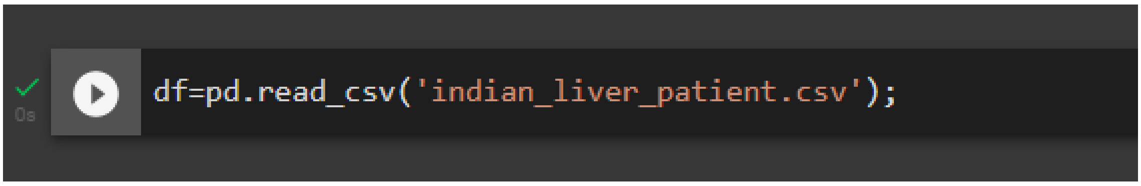
READING THE DATASET

Date	19 November 2022
Team ID	PNT2022TMID04221
Project Name	Project – Statistical Machine Learning Approaches to Liver Disease Prediction

- You might have your data in .csv files, .excel files or .**tsv** files or something else. But the goal is the same in all cases. If you want to analyse that data using pandas, the first step will be to read it into a data structure that's compatible with pandas.

- Let's load a .csv data file into pandas. There is a function for it, called **read_csv().**We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program).

- names on Windows tend to have backslashes in them. But we want them to mean actual backslashes, not special characters.



```
df=pd.read_csv('indian_liver_patient.csv');
```

If the dataset is in the same directory of your program, you can directly read it, without giving raw as r.

This 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".

Our Data consists of Following Columns:

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin

- Alkaline Phosphotase
- Alamine Aminotransferas
- 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)