## DATA PRE-PROCESSING EXPLORATORY DATA ANALYSIS

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

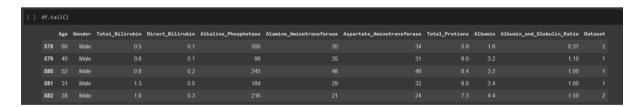
Exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods and used for determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

**head()**: To check the first five rows of the dataset, we have a function called **head()**.



Head() method is used to return top n (5 by default) rows of a DataFrame or series.

Tail(): To check the last five rows of the dataset, we have a function called tail().



## **Understanding Data Type and Summary of features**

How the information is stored in a DataFrame or Python object affects what we can do with it and the outputs of calculations as well. There are two main types of data those are numeric and text data types.

Numeric data types include integers and floats.

- Text data type is known as Strings in Python, or Objects in Pandas. Strings can contain numbers and / or characters.
- Or example, a string might be a word, a sentence, or several sentences.

Will see how our dataset is, by using the info() method.

```
[ ] df.info()
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 583 entries, 0 to 582
    Data columns (total 11 columns):
                                    Non-Null Count Dtype
         Column
     0
                                    583 non-null
                                                   int64
        Age
     1
        Gender
                                    583 non-null
                                                  object
        Total Bilirubin
                                   583 non-null
                                                   float64
     2
                                   583 non-null
        Direct Bilirubin
     3
                                                  float64
                                 583 non-null
        Alkaline_Phosphotase
     4
                                                  int64
     5
        Alamine Aminotransferase
                                  583 non-null
                                                  int64
        Aspartate_Aminotransferase 583 non-null
     6
                                                  int64
     7
        Total Protiens
                                    583 non-null
                                                  float64
     8
        Albumin
                                    583 non-null
                                                  float64
     9
         Albumin and Globulin Ratio 579 non-null
                                                   float64
     10 Dataset
                                    583 non-null
                                                   int64
    dtypes: float64(5), int64(5), object(1)
    memory usage: 50.2+ KB
```

We notice that dataset contains both categorical and numerical columns. But it is not necessary that all the continuous data which we are seeing has to be continuous in nature. There may be a case that some categorical data is in the form of numbers but when we perform info() operation we will get numerical output. So, we need to take care of those type of data also.

**describe():** functions are used to compute values like count, mean, standard deviation and IQR(Inter Quantile Ranges) and give a summary of numeric type data.

