DATA PRE-PROCESSING

SPLITTING THE DATASET INTO DEPENDENT AND INDEPENDENT VARIABLES

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

- In machine learning, the concept of dependent variable (y) and independent variables(x) is important to understand. Here, Dependent variable is nothing but output in dataset and independent variable is all inputs in the dataset.
- With this in mind, we need to split our dataset into the matrix of independent variables and the vector or dependent variable. Mathematically, Vector is defined as a matrix that has just one column.

To read the columns, we will use iloc of pandas (used to fix the indexes for selection) which takes two parameters — [row selection, column selection].

Let’s split our dataset into independent and dependent variables.

1. The independent variable in the dataset would be considered as 'x'.
2. The dependent variable in the dataset would be considered as 'y'.

Now we will split the data of independent variables.

```
# The input variables/features are all the inputs except Dataset. The prediction or label is 'Dataset' that determines whether the patient has liver disease or not.
X = df.drop(['Dataset'], axis=1)
y = df['Dataset']
```

In the above code we are creating array or list of the independent variable x with our selected columns and for dependent variable y we are only taking the dependent or output or target column.