**Disease prediction**

**First Update – Explanation Document**

**Software Requirements:**

* Python 3
* Anaconda (Jupyter Notebook)

**Processing Steps:**

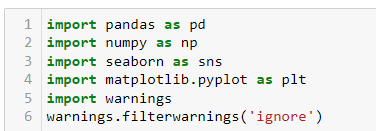
**Data Collection:**

* Data collection is the process of gathering and measuring information on variable of interest.
* Machines initially learn from the data what we give to them.
* Collecting the data is the most important process, so that our machine learning model can find the correct patterns.
* The quality of data that we feed to the machine will also determine how accurate our model to predict the outcomes.

Dataset Link: https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction

**Data Preprocessing:**

* Data preprocessing is a process of transforming the raw data into an understandable format.
* To make the data suitable, Data Preprocessing is the first and crucial step.
* Since the dataset contains huge number of data points, it is necessary to filter out the uncertainties such as missing value, Null values and irrelevant data from the dataset.
* This is because, the uncertainties have a negative impact on the accuracy and hence they must be eliminated from the dataset.
* First, Importing the necessary libraries

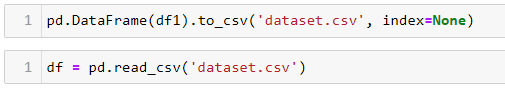


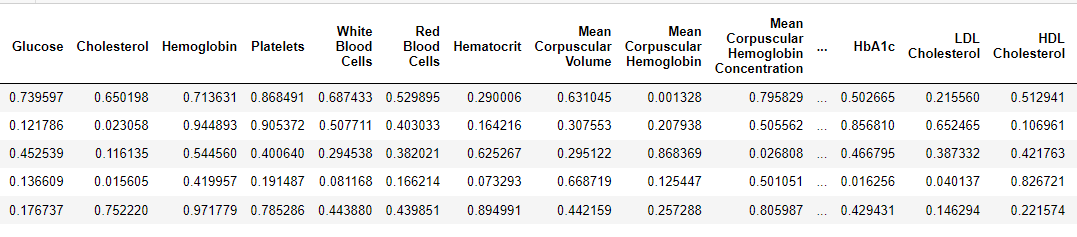
**Concat()** - The concat() function concatenates an arbitrary amount of Series or Data frame objects along an axis while performing optional set logic (union or intersection) of the indexes on the other axes.

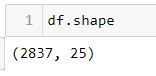


**Loading of Dataset:**

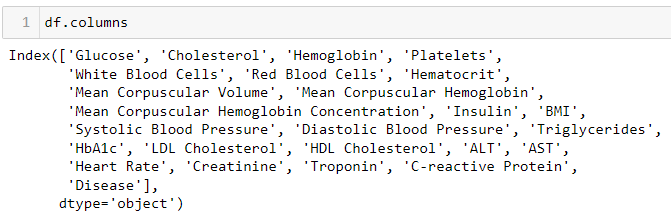
Dataset from the public repository are seperated into two files. For further implentation purpose, the dataset are concat and stored into single file. The below table shows the sample values of data.





**Shape ()** – This returns the no. of rows and columns in the dataset.

**Columns () -** This returns the column labels of datasets.



**The feature used in this dataset and it's description or the (Blood parameters):**

1. Cholesterol: This is the level of cholesterol in the blood, measured in milligrams per deciliter (mg/dL).

2. Hemoglobin: This is the protein in red blood cells that carries oxygen from the lungs to the rest of the body

3. Platelets: Platelets are blood cells that help with clotting

4. White Blood Cells (WBC): These are cells of the immune system that help fight infections

5. Red Blood Cells (RBC): These are the cells that carry oxygen from the lungs to the rest of the body

6. Hematocrit: This is the percentage of blood volume that is occupied by red blood cells

7. Mean Corpuscular Volume (MCV): This is the average volume of red blood cells

8. Mean Corpuscular Hemoglobin (MCH): This is the average amount of hemoglobin in a red blood cell

9. Mean Corpuscular Hemoglobin Concentration (MCHC): This is the average concentration of hemoglobin in a red blood cell

10. Insulin: This is a hormone that helps regulate blood sugar levels

11. BMI (Body Mass Index): This is a measure of body fat based on height and weight

12. Systolic Blood Pressure (SBP): This is the pressure in the arteries when the heart beats

13. Diastolic Blood Pressure (DBP): This is the pressure in the arteries when the heart is at rest between beats

14. Triglycerides: These are a type of fat found in the blood, measured in milligrams per deciliter (mg/dL)

15. HbA1c (Glycated Hemoglobin): This is a measure of average blood sugar levels over the past two to three months

16. LDL (Low-Density Lipoprotein) Cholesterol: This is the "bad" cholesterol that can build up in the arteries

17. HDL (High-Density Lipoprotein) Cholesterol: This is the "good" cholesterol that helps remove LDL cholesterol from the arteries

18. ALT (Alanine Aminotransferase): This is an enzyme found primarily in the liver

19. AST (Aspartate Aminotransferase): This is an enzyme found in various tissues including the liver and heart

20. Heart Rate: This is the number of heartbeats per minute (bpm)

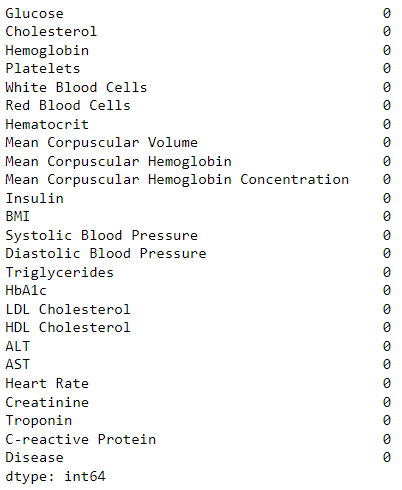
21. Creatinine: This is a waste product produced by muscles and filtered out of the blood by the kidneys

22. Troponin: This is a protein released into the bloodstream when there is damage to the heart muscle

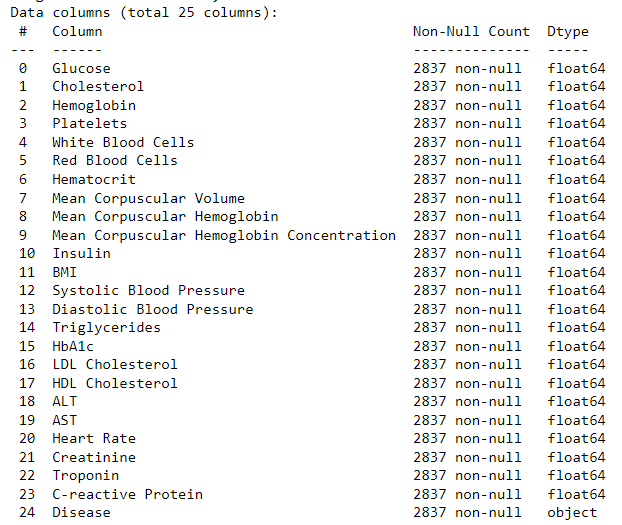
23. C-reactive Protein (CRP): This is a marker of inflammation in the body

24. Disease: This indicates whether he has a specific disease or not.

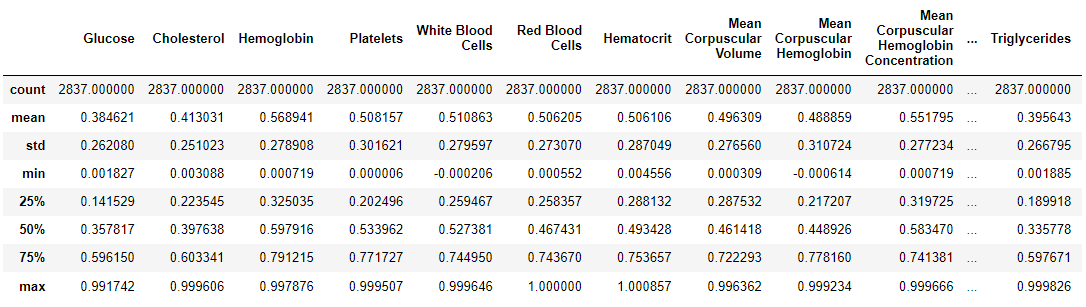
**isnull().sum ( ) -** This returns a Series about the null values in dataset



**info()** - This method returns information about a datasets including the index dtype and columns, non-null values and memory usage

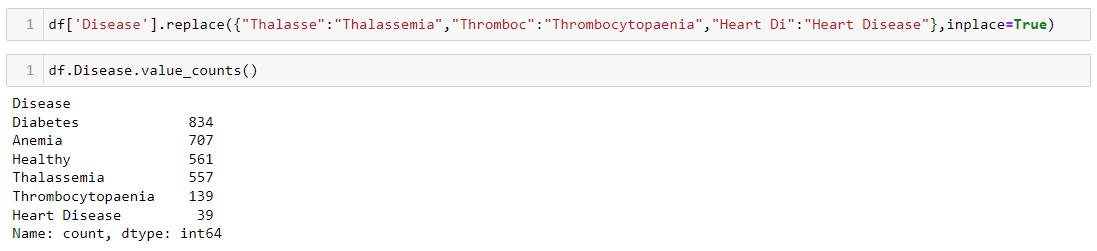


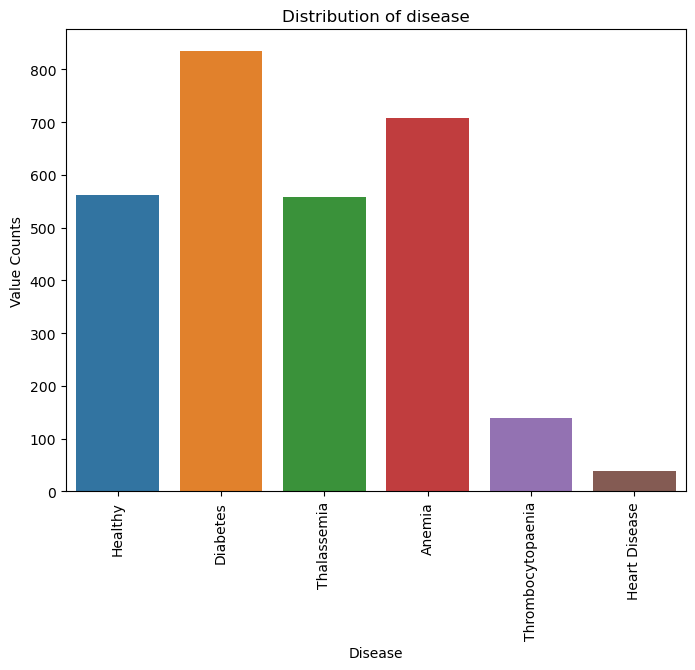
**describe() -** This returns description of the data in the Dataset.



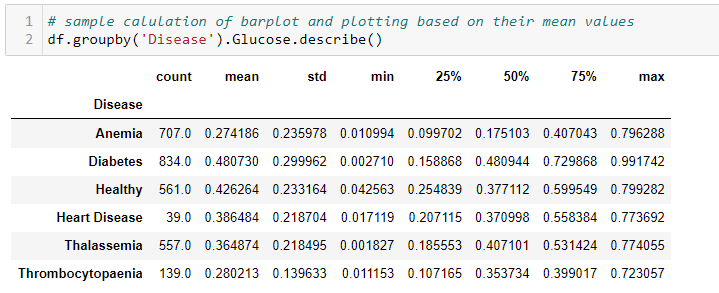
**Replace() :** replace the naming based on comfortable level.

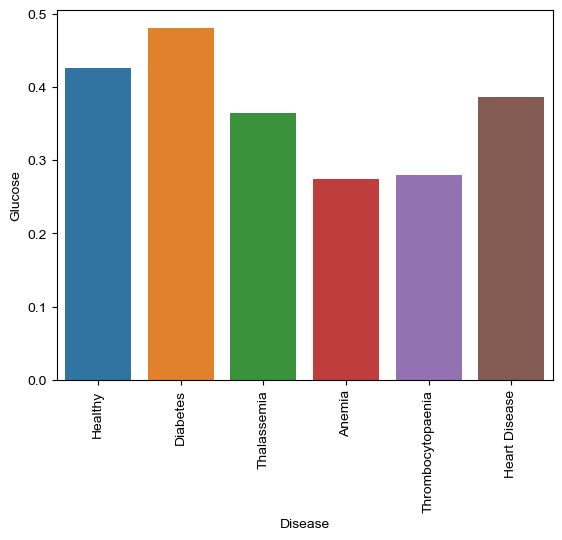
**Value\_counts()** : calculate the total number of counts in particular diseases.





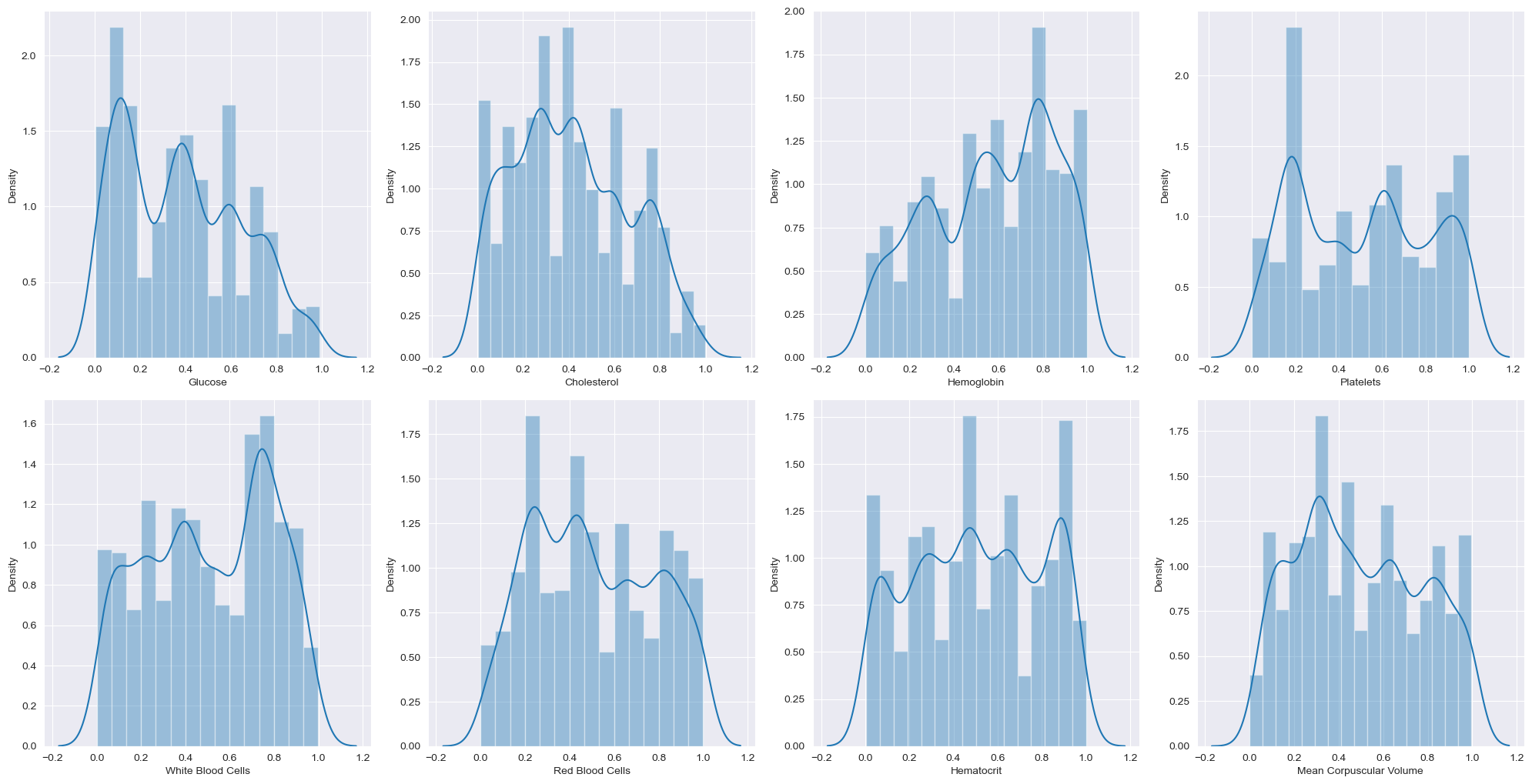
**Barplot** - grouping of categorical values into numerical values, it calculates mean,std, min, max and visualize in the form of graphs. Here, grouping of disease based on their gulcose level and their mean values are plotted in the graphs below. Similary, all independent variables and their graph are shown in notebook file. Kindly refer..,

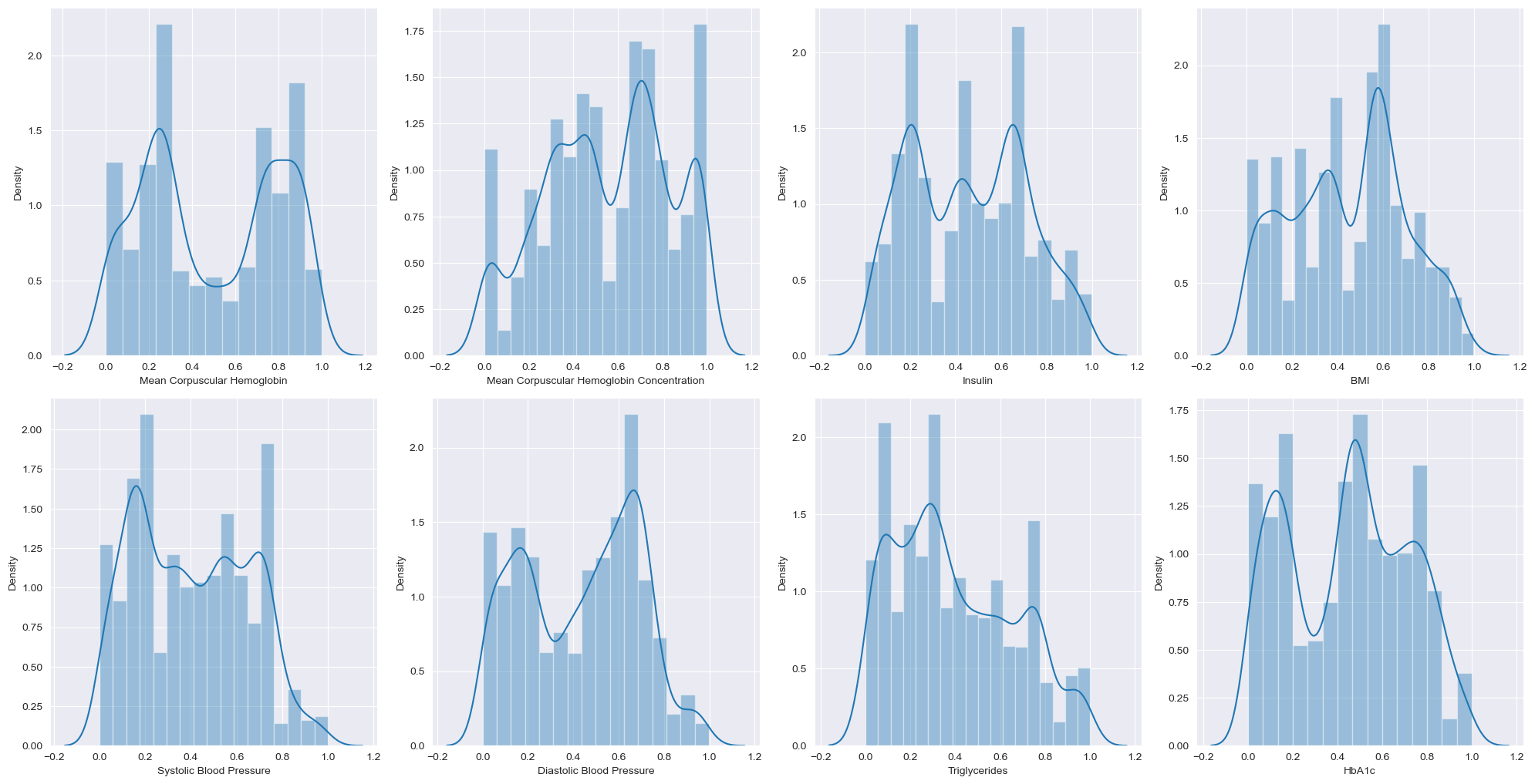


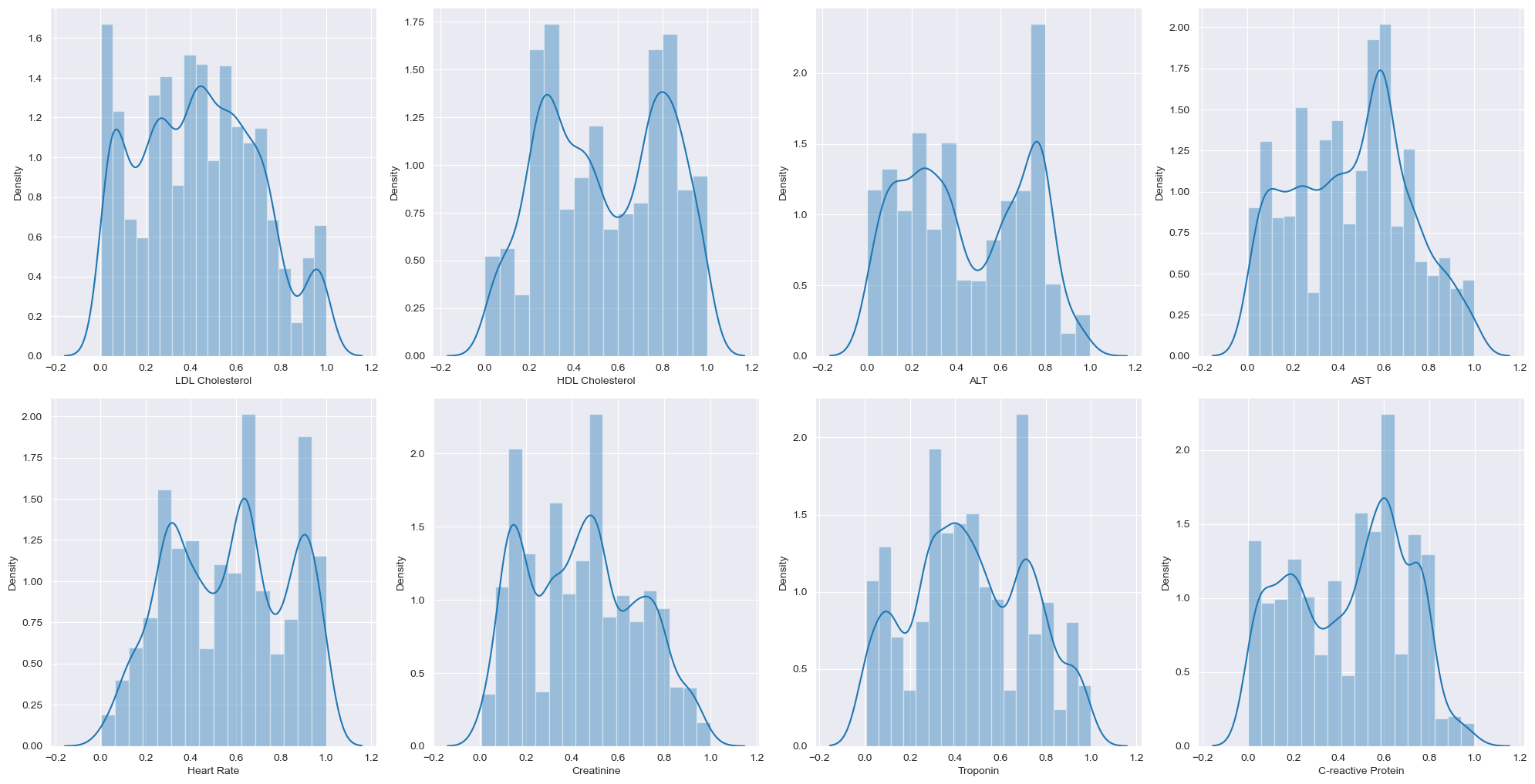


**Distplot:** A Distplot or distribution plot, depicts the variation in the data distribution.

Seaborn Distplot represents the overall distribution of continuous data variables.

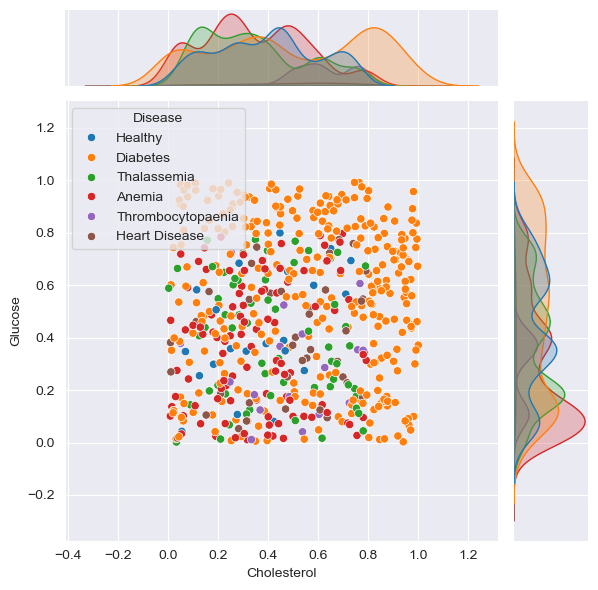






**Joint Plot** - A Jointplot is a figure that showcases the relationship between two variables, combining scatter plots, hexbin plots, regression plots, or 2D kernel density plots with histograms, KDE (Kernel Density Estimate) plots, or other representations of the univariate distribution of each variable on the margins.

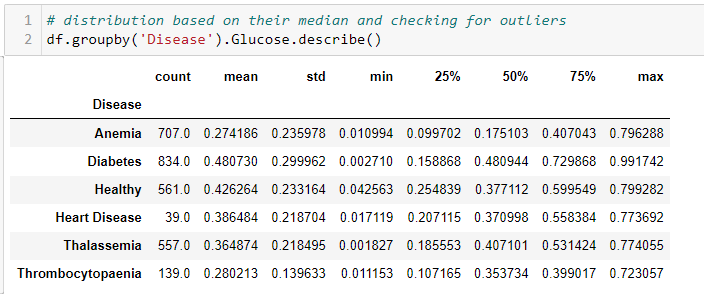
* From this, it is clearly shown that when cholestrol level greate than 0.8 and glucose level between 0-1 cause Diabetes.
* Similary, the relation between all the independent features are plotted in form of matrix pairplot are shown in notebook. Kindly refer..,

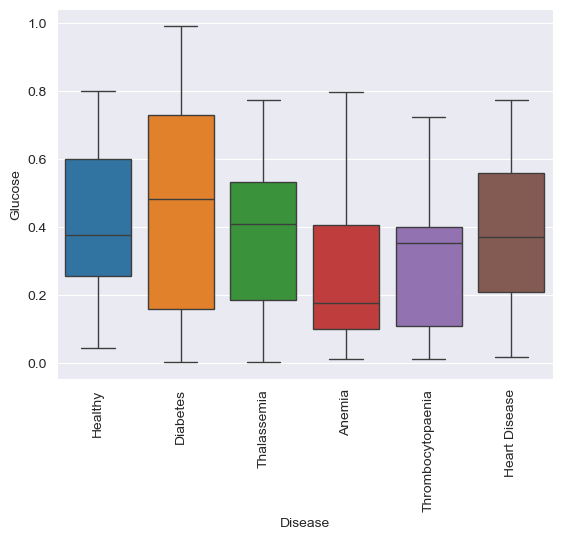


**Checking for outliers:**

* A box plot allows us to identify the uni-variate outliers, or outliers for one variable.
* Box plots are useful because they show minimum and maximum values, the median, and the interquartile range of the data.

The below graph shows the 25% as minimum value, 75% as maximum values and 50% shows the median range of Glucose by grouping up the class variable.





Healthy – 0.377112

Diabetes – 0.480944

Thalassemia – 0.407101

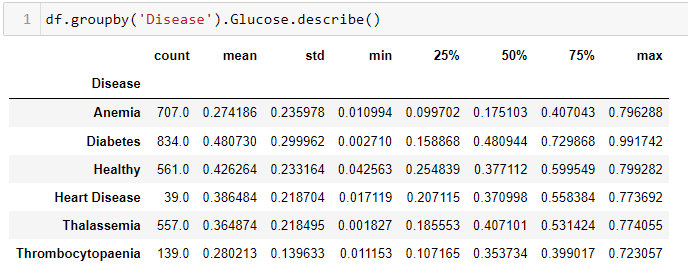
Anemia – 0.175103

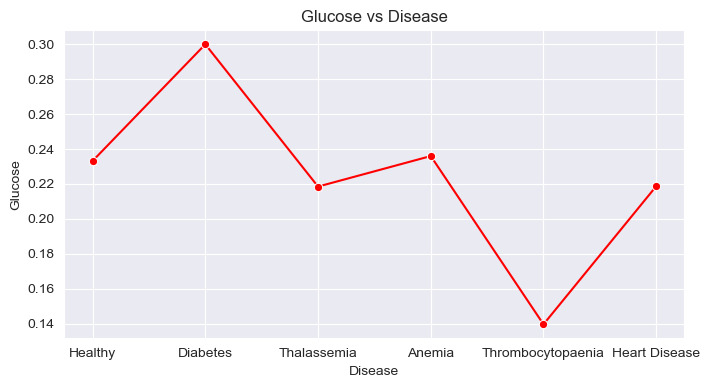
Thrombocytopaenia – 0.353734

Heart Disease – 0.370998

**Line plot:**

* A line chart is created using sample data points. Annotations displaying the x and y coordinates are added to each data point on the line chart for enhanced clarity.
* Below plot show the distribution of **standard deviation** regarding Glucose values and their respective disease.
* Similary, the relation between all the independent features are plotted in form of line plot are shown in notebook. Kindly refer..,





Healthy – 0.233164

Diabetes – 0.299962

Thalassemia – 0.218495

Anemia – 0.235978

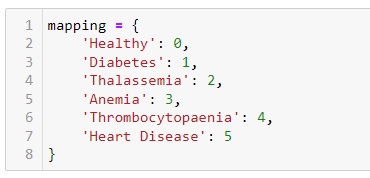
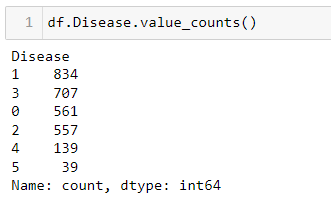
Thrombocytopaenia – 0.139633

Heart Disease - 218704

**Converting Categorical data into Numerical data:**

* Categorical data cannot typically be directly handled by machine learning algorithms, as most algorithms are primarily designed to operate with numerical data only.
* Therefore, before categorical features can be used as inputs, it must be converted into numerical data for better analysis.

**Mapping:**

****

**Correlation:**

* The Seaborn. Heat map() method is used to plot rectangular data in the form of a color-coded matrix.
* A heat map is one of the data visualization tools that shows the magnitude of a certain phenomenon in the form of colors.

****