Diabetes_Exploratory_Data_Analysis

20 November 2023 16:30

The dataset consist of few independent variables & one target variable

Independent variables include - no. of pregnancies, Age, insulin level, BMI, Skin thickness etc

Matplotlib - data visualization python library
Seaborn - advanced data visualization library

- data = pd.read_csv("/kaggle/input/pima-indians-diabetes-database/diabetes.csv") --- takes in the excel file.
- data.head() -- prints out the top data
- data.describe() -- gives out the mathematical description about the data
- data.info() -- gives out the null values
- data.isna().sum() -- also to check out the null values
- data.duplicated().sum() -- to check duplicate values

Outlayers describe the maximum and minimum of a feature
hue - describes in which column d you want to show the difference

Here we are using the **kth nearest neighbor algorithm** to figure out the patterns