IC 272: Lab1: Data visualization and statistics from data

Deadline for submission: Aug 30, 2022, 10:00 PM

You are given the **Pima Indians Diabetes Database** as a csv file. This data-set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females with at least 21 years old of Pima Indian heritage. It contains following 9 attributes.

pregs: Number of times pregnant
plas: Plasma glucose concentration 2 hours in an oral glucose tolerance test
pres: Diastolic blood pressure (mm Hg)
skin: Triceps skin fold thickness (mm)
test: 2-Hour serum insulin (mu U/mL)
BMI: Body mass index (weight in kg/(height in m)^2)
pedi: Diabetes pedigree function
Age: Age (years)
class: Class variable (0 or 1)

Write a python program (with pandas) to read the given data and display following:

1. Mean, median, mode, minimum, maximum and standard deviation for all the attributes excluding the attribute 'class'.
2. Obtain the scatter plot between
    a. 'Age' and each of the other attributes, excluding 'class'
    b. 'BMI' and each of the other attributes, excluding 'class' (You can use `matplotlib` library).
3. Find the value of correlation coefficient in the following cases:
    a. 'Age' with all other attributes (excluding 'class').
    b. 'BMI' with all other attributes (excluding 'class').
4. Plot the histogram for the attributes 'preg' and 'skin' (You may use "`hist`" function from pandas)
5. Plot the histogram of attribute 'preg' for each of the 2 classes individually (Use "`groupby`" function to group the tuples according to their "class")
6. Obtain the boxplot for all the attribute excluding 'class' (Use "`boxplot`" function).