IC 272: Lab1: Data visualization and statistics from data

Deadline for submission: Aug 22, 2021, 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).

Write a report that should include the at least the following:

- Answers to these questions (including figures/plots). Note: Clearly label the x and y axis of figures/plots.
- The inference from the scatter plots in 2a and 2b.
- The inference from the correlation coefficient values in 3a and 3b.
- Observations from scatter plots and their relation with corresponding correlation coefficients
- Observations on the plots in 4, 5 and 6.
- Observations from the outcomes of question 1 and their comparison with the plots from question 6.
- Any other observations and inferences.

Instructions:

- **Your python program(s) should be well commented. Comment section at the beginning of the program(s) should include your name, registration number and mobile number.**
- **The python program(s) should be in the file extension .py**
- **Report should be strictly in PDF form. Write the report in word or latex form and then convert to PDF form.**
- **First page of your report must include your name, registration number and mobile number.**
- **Upload your program(s) and report in a single zip file. Give the name as <roll_number>_Assignment1.zip. Example: *b20001*_Assignment1.zip**
- **Upload the zip file in the link corresponding to your group only.**

In case the program found to be copied from others, both the person who copied and who help for copying will get zero as a penalty.