COMP40370 Practical 1

DATA EXPLORATION AND PREPROCESSING (Part B)

Prof. Tahar KECHADI

Academic year 2023-2024

The aim of this practical is to get familiar with some basic tools of data preprocessing and data exploration, and also use some of the concepts discussed in the lectures so far. Python is the programming language to use to complete this practical. For this practical download *diabetes.csv again, and work on the original version (not the version that you processed in part A)*

Assignment Files

- ./Practical-01-B.pdf	assignment questions (this file).
- ./*diabetes.csv*	data file for the questions.
- ./ages.csv	data file for the questions

Expected output files

- ./Prcatical-01-B.ipynb	Python notebook programs.
- ./Prcatical-01-B.html	Notebook in HTML showing the outputs.

Packages

Feel free to use the following packages:

- Python 3.8+, pandas 1.3+, numpy 1.20+, sklearn 0.24+.
- seaborn 0.11+, matplotlib 3.5+, scipy 1.9+.

Note

It is not recommended to loop/iterate a dataframe under any circumstances. Answers with dataframe iterations will not get full marks.

Question 1: Outliers removal and transformation

a. What are the kurtosis and skewness values of the Insulin attribute?
b. Filter the dataframe and select only the rows where Insulin is not equal to 0, and use the new version (filtered dataframe) for the rest of the practical
c. Draw the histogram of Insulin column along with kernel density estimation (*KDE*) curb.
d. Identify outliers of Insulin using Inter Quartile Range (IQR) approach and impute them with lower band and upper band values appropriately (any outliers above the upper band, replace it with the value of upper band, and any outliers below the lower band, replace it with the value of the lower band). Then draw the histogram again to observe the difference
e. Transform Insulin column using $\log_e$ (x+1) formula to make the Insulin values follow the normal distribution.
f. Find the kurtosis and skewness of Insulin after the transformation
g. Draw two QQ-plots to compare before and after the $\log_e$ (x+1) transformation for Insulin column.

h. Similarly detect and correct outliers (step d) in the 'Glucose','BloodPressure' and 'SkinThickness' columns.

i. Display the correlation matrix using the seaborn heatmap function between continuous variables; Pregnancies, Glocose, BloodPressure, SkinThickness, Insulin, BMI, Age.

Question 2: Data processing

a. Group the patients by number of pregnancies, along with their average BloodPressure.

b. Group the patients by number of pregnancies, along with the average age, sum of all ages in that group, and the count of patients in that group

c. Add a new column named ‘BMI/Age’ with the value of BMI over the Age, without using any loops.

d. Without looping the rows, add a new column named ‘risk’ with the following rules:

    a. High: if BMI > 45 and BloodPressure > 100

    b. Medium: if 30 < BMI < 45 and BloodPressure < 100

    c. Low: if BMI < 30

    d. Unknown: otherwise

e. The file named ages.csv provides the age group names. Merge the two dataframes by the age, the resulted dataframe must contain all the columns of diabetes.csv, in addition to their age group in the column named ‘AgeGroup’.

f. one-hot encode the categorical variable ‘AgeGroup’ and add the resulting columns to the dataframe, and remove AgeGroup.

The final deadline for the submission of Practical 01 (Part A and B) is Wednesday, 4th of October at 23:00. Submissions should be in a single file with FirstName_LastName-P1.zip (or tar.gz) format. All submissions must be done in Brightspace.