

COMP40370 Practical 1

DATA EXPLORATION AND PREPROCESSING (Part A)

Prof. Tahar KECHADI

Academic year 2023-2024

This practical aims to familiarise with some basic data pre-processing and exploration tools and use some concepts discussed in the lectures. Python is the programming language to use to complete this practical. The datasets needed to complete the practical are described below.

Assignment Files

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

Expected output files

- `./Prcatical-01.ipynb` Python notebook programs.
- `./Prcatical-01.html` Notebook in HTML showing the outputs.
- `./diabetes.csv` Original data file for the questions.

Requirements

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

Question 1: Data Exploration

The comma-separated file "*diabetes.csv*" consists of several medical variables and one target variable, `Outcome`. The variables include the number of pregnancies the patient has had, their BMI, insulin level, age, etc. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes based on specific diagnostic measurements included in the dataset. The dataset has been modified for the purpose of this assignment.

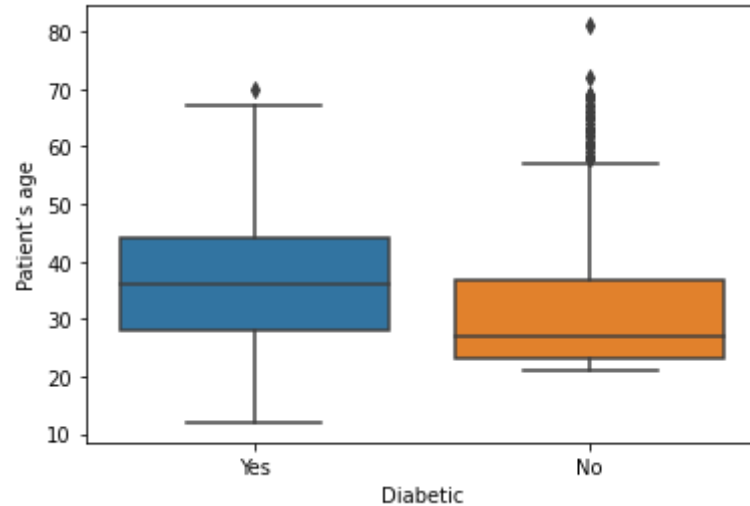
Write a Python program to answer the following:

1) Descriptive statistics

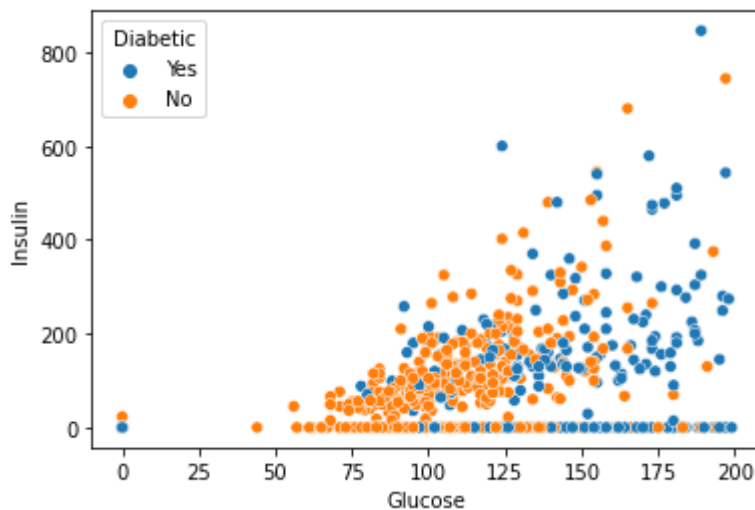
- 1- Read the data file into a pandas data frame and print the first 5 rows
- 2- Print the number of rows and columns
- 3- Calculate the min, max, mean, and std of the 'age' column using pandas.
- 4- What is the mode of the 'age' column? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
- 5- Use pandas to calculate the first quartile (Q1) and the third quartile (Q3) of the 'age' column.
- 6- What is the Interquartile Range of the 'age' column?
- 7- Print the five-number summary of the 'age' column.

II) Data visualisation

- 1-Show a boxplot of the 'age' column.
- 2-Show a boxplot of the 'age' column of the diabetic and no-diabetic groups side by side. The x-axis's label should be (*diabetic*) and 'yes' under positive patients and 'no' under negative patients, and the y-axis's label should be (*patient's age*), as follows:



- 3-Based on the boxplot of question 2, analyse the relationship between the patient's age and being diabetic.
- 4-Show a scatter plot with the patient's Blood pressure on the x-axis and the patient's BMI on the y-axis.
- 5-Show a scatter plot with the patient's Blood pressure on the x-axis and the patient's BMI on the y-axis, highlighting diabetic patients with different colours. The colouring label should be 'diabetic': yes and no.



- 6-Based on the scatterplot of question 5, analyse the relationship between BMI/Blood pressure and diabetes.

III) Data filtering

- 1-Select all patients with Insulin more than 400. How many patients are diabetic/no-diabetic among those selected?
- 2- Select all patients with Insulin greater than 400 and Glucose greater than 175. How many patients are diabetic/no-diabetic among those selected?

3- What is the average Glucose level of a patient with more than 5 pregnancies and older than 45?

4- Count the distinct values in the 'pregnancies' column.

5- List the distinct values of the 'pregnancies' column along with the percentage of diabetic/no-diabetic of each value. Example:

Pregnancies	Outcome	
0	0	0.648148
	1	0.351852
1	0	0.778626
	1	0.221374
2	0	0.818182
	1	0.181818
3	0	0.640000
	1	0.360000
4	0	0.640625
	1	0.359375
5	0	0.611111
	1	0.388889

Question 2: Data Cleaning

I) Duplicated removal

1. Identify any duplicated records by printing "True" if the row is duplicated and "False" otherwise.
2. For all duplicated records, keep one record and remove its duplicates.
3. What is the dimension of the data frame after removing the duplicates?
4. How many duplicated rows were there (before removing the duplicates)?

II) Missing values

- 1- How many missing values are in the "blood pressure" column?
- 2- Remove the missing records in the "blood pressure" column.
- 3- Copy the following columns into a separate data frame: 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'
- 4- On the newly copied data from (3). Use sklearn's train_test_split function to split the data into 90% training and 10% test.
- 5- On the test set, set the Glucose to null for those records. And keep a separate copy for evaluation in the following questions.
- 6- Fill in the missing values of the test set based on the mean of the Glucose of the training set (90%). Calculate the RMSEs for the imputed values of the test set (compared to the copy you have saved from (5)).
- 7- Fill in the missing values of the test set based on the median of the Glucose of the training set (90%). Calculate the RMSEs for the imputed values of the test set (compared to the copy you have saved from (5)).
- 8- Use scikit-learn SimpleImputer with the 'most_frequent' strategy, and calculate RMSE.
- 9- Use scikit-learn KNNImputer (for neighbours = 3), and calculate RMSE.
- 10- Which Imputer is better?

Please make sure that you have completed this practical. Next week, you will get the second part of the practical.