**ENDG 319 - Fall 23**

**CURE Project – Deliverable 2**

**>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>**

**Graphical data analysis to develop a machine-learning model**

**Instructions:**

* This deliverable is worth 3% of your overall grade and should be submitted individually.
* The assignment will require you to use Python **and you will have to paste screenshots of your work in a Word file to show your work that involves coding**. Please keep the original Word file. Once done, convert the Word file into pdf with file name: Last Name\_First Name\_UCID\_CURE Deliverable 2.
* Submit the pdf file to Assessments > Dropbox > ‘CURE Project Deliverable 2’ by Oct 15, 2023, 11:59 pm (MT).
* You must submit this deliverable on time to be able to submit the upcoming deliverables.

**>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>**

**Research skill:**

(i) Dealing with real-world data

(ii) Graphical analysis of data, interpreting figure

(iii) Critical thinking - drawing conclusions from data

(iv) Use of engineering tools – python

(v) Developing a research question

**Relevant course content:** Descriptive Statistics (Ch 1), Machine learning – classification

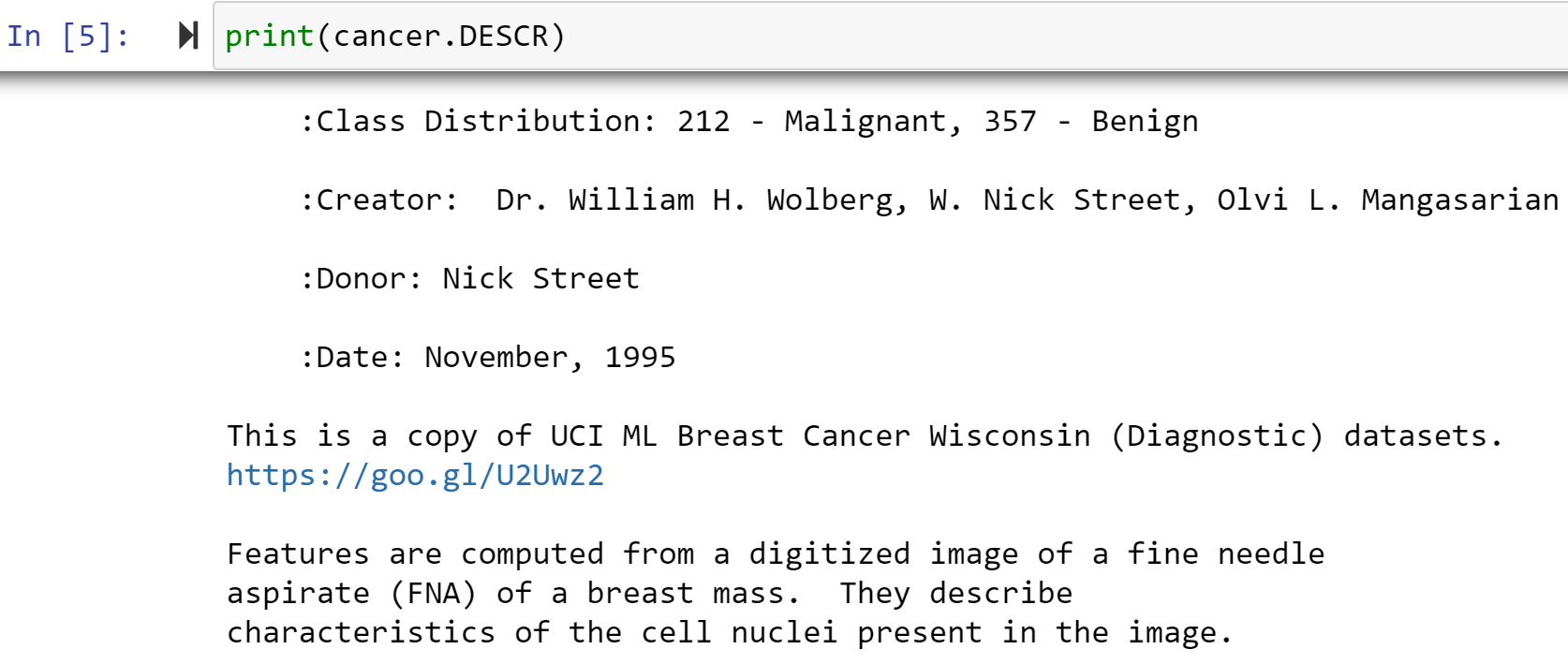
**>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>**

**Exploring the breast\_cancer dataset in sklearn**

In the breast\_cancer database there are 30 features and 2 classes, as shown below.

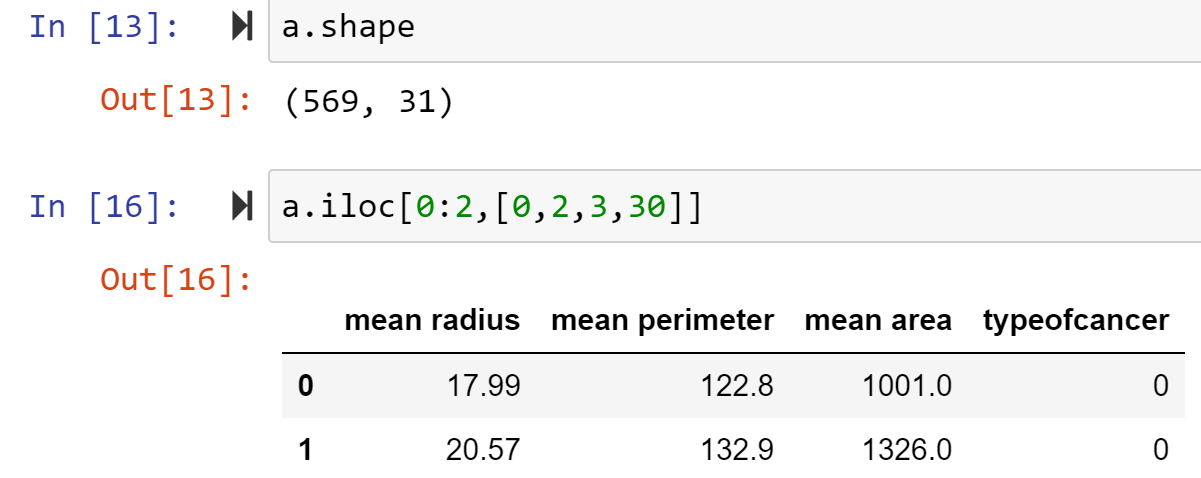


More information is available in the description. Read the following snippet from the description.

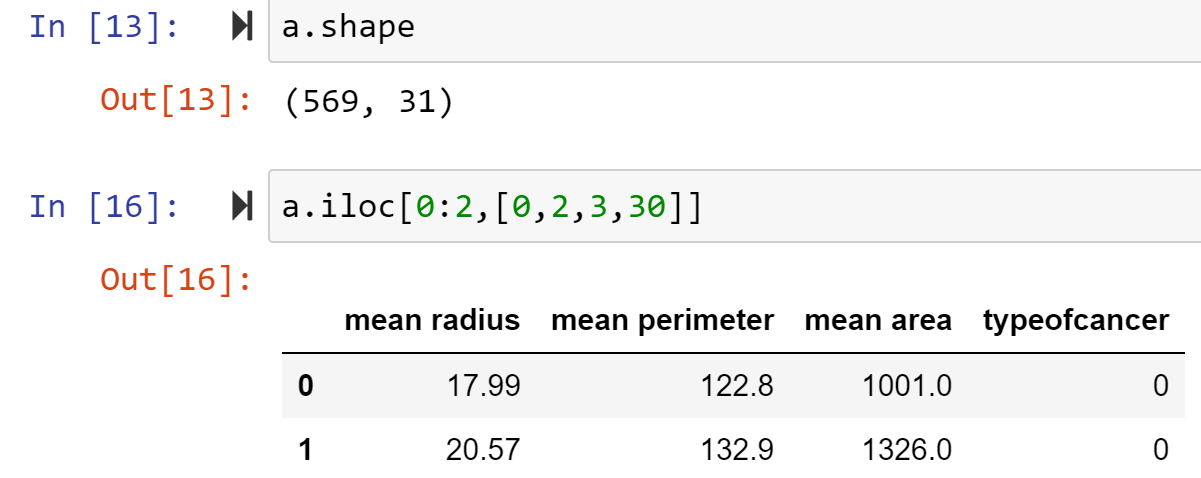


**Task 1**

Create a dataframe variable ‘a’ with this dataset. This dataframe should have all the 569 instances, 30 features and the class of 569 instances as 0 (Malignant) or 1 (Benign). The column that contains the classes should be labeled as ‘typeofcancer’. Show the output of the following input:



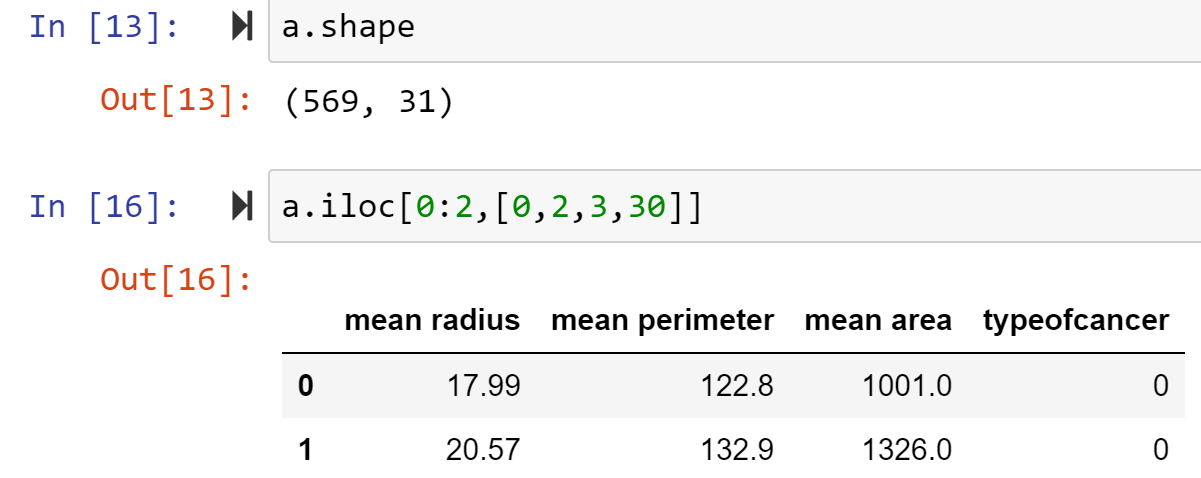
[Hints: the outputs should be same as below.



(b) Now create a dataframe variable ‘df’ by slicing dataframe ‘a’. The new datafraeme ‘df’ should have all the instances, their labels but with the following three features: mean radius, mean perimeter and mean area. [Hints: use .iloc method to extract necessary columns from ‘a’]

(i) Show the first two rows.

[Hint: The output should be same as below.



(ii) Show the rows with indexes 17, 18, 19, 20, 21.

**Task 2**

Suppose we want to explore the possibility of developing a machine learning model that can diagnose a new patient’s cancer condition as benign or malignant from the features in df.

(i) As a first step, you want to do some graphical analysis. Write the code to generate the following figure (Figure 1). Show screenshot of the code (input) and the figure (output) from your work. You are free to choose your favorite data marker and color in your figure.

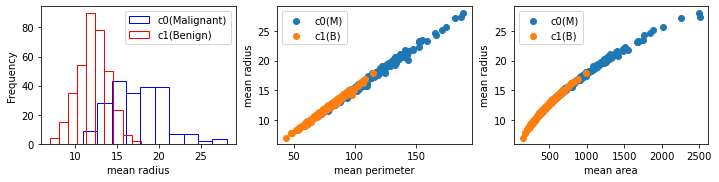
****

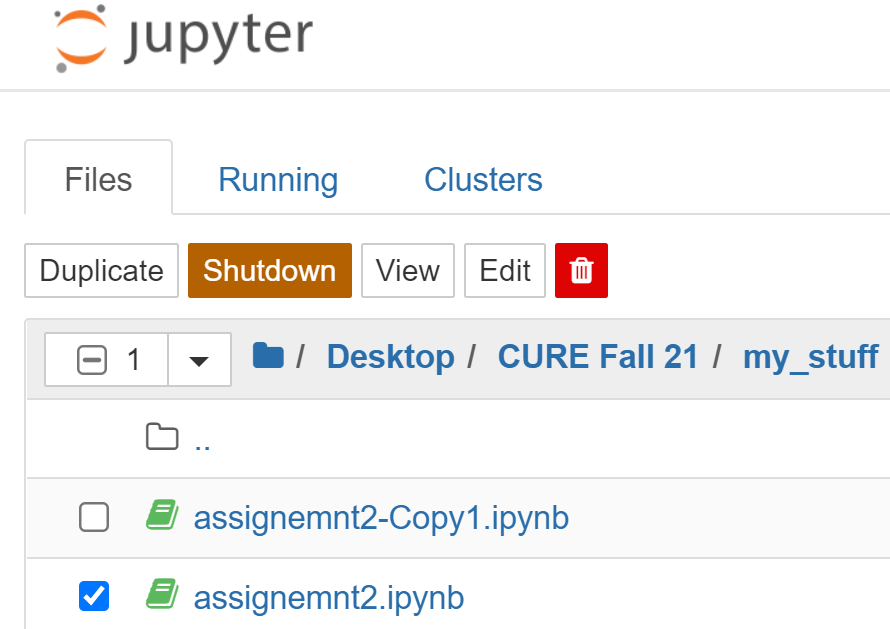
Figure 1: From left of right: histogram of ‘mean radius’ data for each class, scatter plot of ‘mean radius’ versus ‘mean perimeter’, scatter plot of ‘mean radius’ versus ‘mean area’.

(ii) Briefly describe what each of the subplots in Figure 1 reveal about the data.

**Task 3**

Save the file generated in Task 2. Copy it and modify the code to generate the following figure (Figure 2).

*Note: In juputer notebook you can use the ‘Duplicate’ button to create a copy of a file as shown below. The ‘Duplicate’ button will be available when you select the file you want to copy.*



(i) Generate the following figure (Figure 2). Show screenshot(s) of the code (input) and the figure (output) from your work. You are free to choose your favorite data marker and color in your figure.

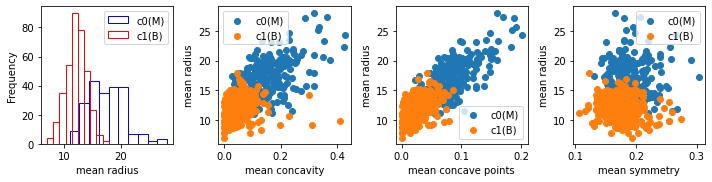
****

Figure 2: From left of right: histogram of ‘mean radius’ data for each class, scatter plot of ‘mean radius’ versus ‘mean concavity’, scatter plot of ‘mean radius’ versus ‘mean concave points’, and scatter plot of ‘mean radius’ versus ‘mean symmetry’.

(ii) Briefly describe what each of the subplots (except the histogram) in Figure 2 reveal about the data.

**Task 4**

Think of a classification problem from your own experience and interest (e.g., distinguishing between two types of things/living beings/weather/personality/vehicle/industry/sports/houses/professionals etc.). Collect some real-world data to solve that classification problem. Have at least two classes, two attributes and at least 5 instances per class. Create a dataframe with this data and paste a screenshot of it. Also give a short description of the data (number of instances, number of attributes and the list of attributes, list of classes, and creator: Your name). The attributes should be numerical.

The classification problem should be a new one, but the dataset may contain data that are new (generated by you by any rough measurements) or collected from public domain. Please cite the references/sources from which you collect the data.