**ENDG 319 - Fall 23**

**CURE Project – Deliverable 3 – Groupwork**

**Maximum Points: 30**

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

**Define a classification problem and analyze data graphically**

**Instructions:**

1. This deliverable is worth 6% of your overall grade and should be completed as a group and submitted by a member of the group. **The mark you get is based on the quality of the work**.
2. 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 the file name: ENDG 319\_Group No\_CURE Deliverable 3. One report needs to be submitted on behalf of each group. Please make sure to put the names of all the group members in the report.
3. Submit the pdf file to Assessments> Dropbox > ‘CURE Project Deliverable 3’ by Oct 30, 2023, 11:59 pm (MT).
4. 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

(vi) Working in a group effectively.

**Relevant course content:** Descriptive Statistics, Machine learning, Python

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

**Task 1 [ 6 marks]**

Discuss your response and the data that you individually generated in ‘Task 4 of CURE deliverable 2’.

*Guidelines for discussion: Each of you can take two minutes to share his/her data with the group. Tell why you chose the dataset (maybe it is an important issue in general or you are personally interested in that topic, etc.). Also give any constructive feedback to others’ dataset (interesting, good problem for classification etc.*

Then report the following completed discussion summary as a summary/result of your discussion (read Task 2 before you complete this part):

Discussion summary:

We summarized the four classification problems proposed by each of the group member in ‘Task 4 of CURE deliverable 2’ as shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Member | Topic | No of classes | No of attributes | No of instances |
|  | Classification of … |  |  |  |
|  | Classification of … |  |  |  |
|  | Classification of … |  |  |  |
|  | Classification of … |  |  |  |

After discussing each problem, we chose to expand the …… dataset on ….

Or

The discussion led us to choose a new dataset on …..….

**Task 2**

**In your CURE project, you will work with a larger dataset as a group. This dataset should have all the following criteria: (i) at least two categories, (ii) at least three attributes, and (iii) a total of at least 100 instances, and (iv) at least 30 instances in each class.**

*Guidelines: You can choose to select one of the group member’s dataset from ‘CURE deliverable 2’ if all of you find it interesting for the project and expand it as per the criteria discussed above. Or you can work with a new dataset (e.g., distinguishing between different types of things/living beings/weather/personality/vehicle/industry/sports/types of houses/professionals/places/time periods/levels of performance/levels of satisfaction/peoples’ attitude etc.). The classification problem should be a new one, but the dataset may contain data that are either new (generated by you by any rough measurements) or collected from public domain. Please cite the references/sources from which you collect the data.*

(a) Define the classification problem by a single statement: ‘Classification of ...’ **[ 2 marks]**

(b) Describe the dataset (*Guidelines: give a short one/two-line description of the dataset: what it is all about and why it is important*) **[ 2 mark]**. Then provide the following statistics:

**[ 1 mark]**. Number of instances:

**[ 1 mark]**. Number of attributes:

**[ 1 mark]**. Number of classes:

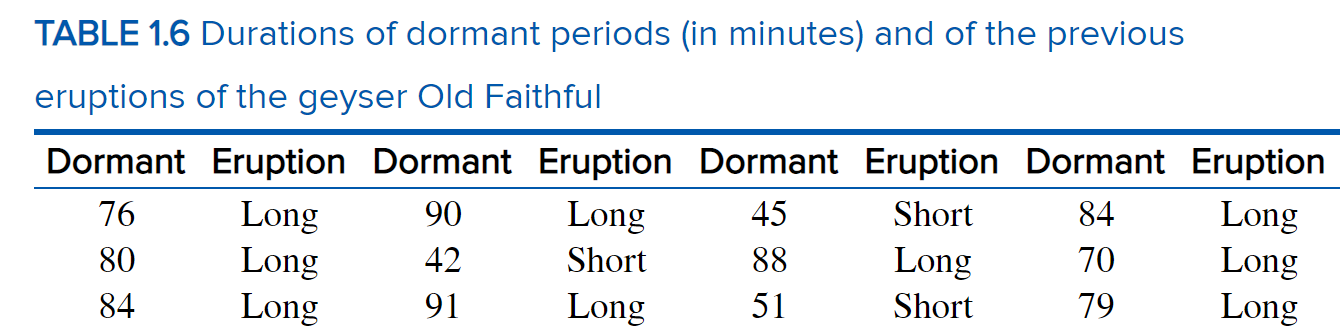
**[ 1 mark]**. Creator:

Now provide a table with all the data **[ 5 marks]** . Give a title of the table (read the titles of different datasets in the textbook). Mention the source of the data.

Table 1. ……

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instance | Attribute 1 | Attribute 2 | … | Class |
| 0 | 2 | 3 |  | A |
| 1 | 4 | 5 |  | B |
| …. | … | …. |  | … |
| …. | …. | …. |  | … |
|  |  |  |  |  |

*[Guidelines and hints: As good examples of how to present your data in a table and write table titles (captions), see the titles of tables 1.1 to 1.6 from your textbook. Part of Table 1.6 is shown below.*

]

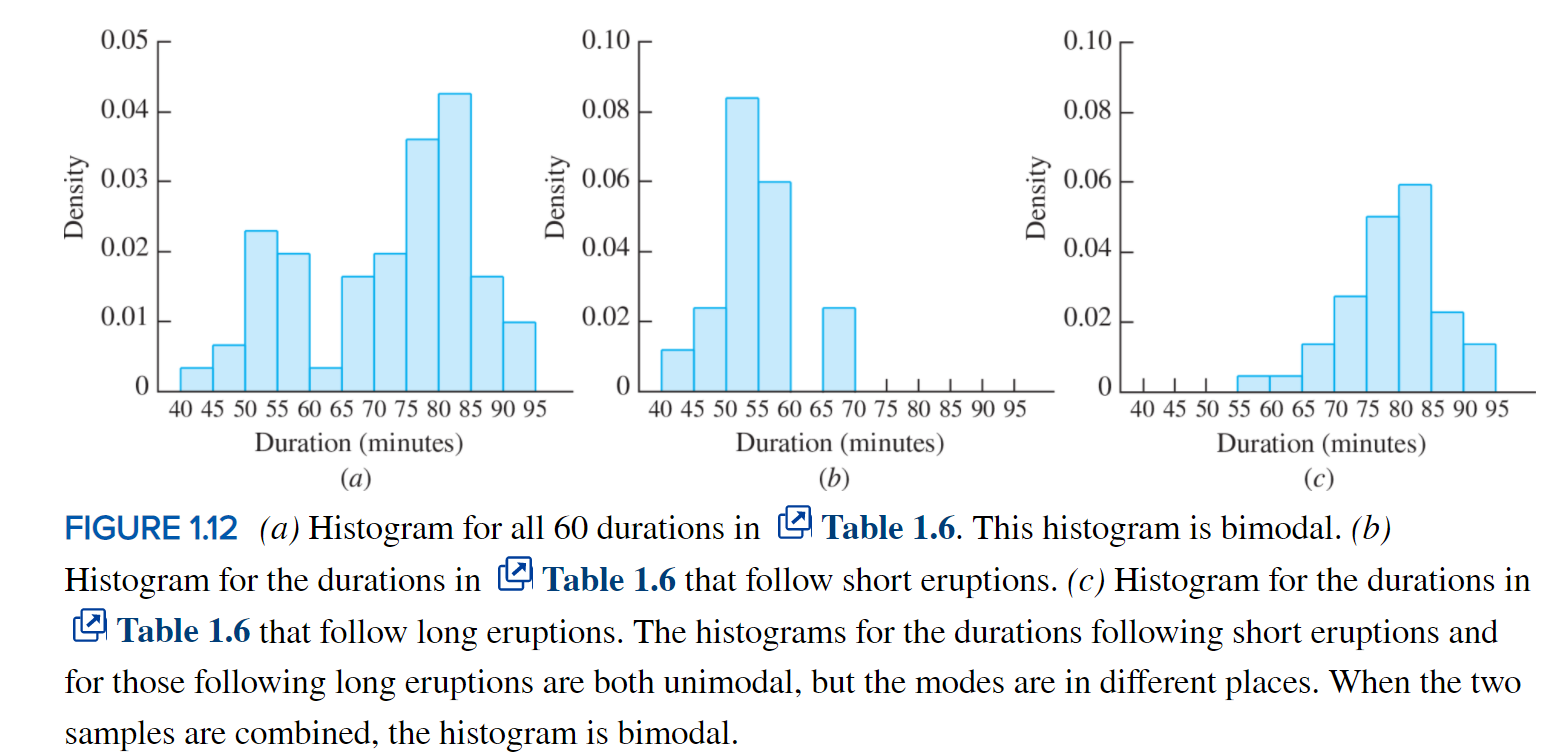
**Task 3**

Explore the possibility of developing a machine learning model through graphical analysis.

(a) Write the code to generate pair plots for your data (as you did in CURE deliverable 2). Show screenshot of the code (input). **[ 4 marks]**

(b) Show screenshot of the figure (pairplot). You are free to choose your favorite data marker and color in your figure. Write short but descriptive title (caption) for your figure(s) at the bottom of each figure. **[ 2 marks]**

*Guidelines and hints for figure caption: The caption should be concise but clear so that even if someone does not read details of your work, s/he can somewhat understand what you are trying to convey with the figure. Again, see some examples in the textbook (FIGURE 1.1 to 1.12), for example, read the caption of Figure 1.12 below.*



(c) Briefly describe what each of the subplots (except the histogram) in your figure(s) reveal about the data. **[ 4 marks]**

(d) Make a overall conclusion in one or two sentence whether a machine learning model will be able to successfully classify a new instance. **[ 1 mark]**