**A picture containing shape, arrow

Description automatically generatedDISCRETIZATION**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shabezade Srikanth**

**Batch Id: DS 07012022**

**Topic: Data Pre-Processing**

**Problem Statement:**

Everything will revolve around the data in Analytics world. Proper data will help you to make useful predictions which improve your business. Sometimes the usage of original data as it is does not help to have accurate solutions. It is needed to convert the data from one form to another form to have better predictions. Explore on various techniques to transform the data for better model performance. you can go through this link:

<https://360digitmg.com/mindmap-data-science>

1. Convert the continuous data into discrete classes on iris dataset.

Prepare the dataset by performing the preprocessing techniques, to have the data which improve model performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

Answer:

import pandas as pd

import numpy as np

data1=pd.read\_csv('F:/Assignments/Assignment 4 data pre processing/DataSets/iris.csv')

data1.dtypes

data1.info()

data1.describe()

data1.columns

pd.cut(data1['Sepal.Length'],bins=3,labels=np.arange(3),right=False) #

pd.cut(data1['Sepal.Length'],bins=3,right=False) #used for binning

pd.cut(data1['Sepal.Width'],bins=3,labels=np.arange(3),right=False)

pd.cut(data1['Sepal.Width'],bins=3,right=False) #used for binning

pd.cut(data1['Petal.Length'],bins=4,labels=np.arange(4),right=False)

pd.cut(data1['Petal.Length'],bins=4,right=False) #used for binning

pd.cut(data1['Petal.Width'],bins=4,labels=np.arange(4),right=False)

pd.cut(data1['Petal.Width'],bins=4,right=False) #used for binning

data1

**A picture containing shape, arrow

Description automatically generatedHints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Hint: Refer to Iris.csv, which is a public dataset.
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function

**Grading Guidelines:**

**Note: 1. An Assignment submission is considered complete only when successful executable code(s), and documentation explaining the applied solution and results are provided. Failing to submit either of them will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline date will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline date
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of questions asked in assignments are completed. (or)
  + All assignments were submitted, however, after the given deadline
* **A picture containing shape, arrow

  Description automatically generatedGrade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of questions asked in assignments are completed. (or)
  + Less than 80% of questions asked in assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):** Assignments submitted after the Deadline and with 50% or less of questions
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of questions asked in the assignments are submitted after the deadline (OR)
  + Less than 30% of questions asked in the assignments are submitted before deadline
* **Grade F: (< 50):** Copied submission or No submission