**DISCRETIZATION**

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: ANANDAKRISHNAN K V**

**Batch Id: 19042021**

**Topic: Preliminaries for Data Analysis**

**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 |

**Ans:-**

**Python code:-**

######################## Discretization ####################

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

from sklearn.impute import SimpleImputer as si

### import data set "iris" as dataframe ###

a\_main = pd.DataFrame(iriscsv)

d = a\_main.copy(deep=True)

## Treat species as output column. so not gonna do discretization technique up on that perticular column

## sepal length

d.iloc[:,1].describe()

bins=[4,5,6,7,8] ### [4,5],(5,6],(6,7],(7,8]

group\_names= ['[4,5]','(5,6]','(6,7]','(7,8]'] ## (a,b] => a not included; but b included

d.iloc[:,1]= pd.cut(d.iloc[:,1],bins, labels = group\_names)

## sepal width

d.iloc[:,2].describe()

bins=[2,3,4,5] ## [2,3],(3,4],(4,5]

group\_names= ['[2,3]','(3,4]','(4,5]'] ## (a,b] => a not included; but b included

d.iloc[:,2]= pd.cut(d.iloc[:,2],bins, labels = group\_names)

## petal length

d.iloc[:,3].describe()

bins=[1,3,5,7] ## [1,3],(3,5],(5,7]

group\_names= ['[1,3]','(3,5]','(5,7]'] ## (a,b] => a not included; but b included

d.iloc[:,3]= pd.cut(d.iloc[:,3],bins, labels = group\_names)

## petal width

d.iloc[:,4].describe()

bins=[0,1,2,3] ## [0,1],(1,2],(2,3]

group\_names= ['[0,1]','(1,2]','(2,3]'] ## (a,b] => a not included; but b included

d.iloc[:,4]= pd.cut(d.iloc[:,4],bins, labels = group\_names)

## conclusion ##

### discretized data set

print("discretized data set is:",d)

**Hints:**

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