**Discretization Problem Statement**

**Name: \_\_\_\_\_RAJU BOTTA\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Batch Id: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_05102021\_\_\_\_\_\_\_\_**

**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:

**1)Objective:** Convert the continuous data into discrete classes on the **iris dataset**.

Prepare the dataset by performing the pre-processing techniques, to have the data which improve model performance.

|  |
| --- |
| **Name of Feature  |     Description                      |   Type                                        |  Relevance**  Sepal.Length         |     Length of sepal   |  Quantitative, Continuous          |  relevant  Sepal.Width           |     Width of Sepal      | Quantitative, Continuous           | relevant  Petal.Length          |      Length of Petal      | Quantitative, Continuous          | relevant  Petal.Width            |      Width of Petal       | Quantitative, Continuous          | relevant |
| Species                  |      Species                  | Quantitative, Continuous          | relevant |

import pandas as pd

data = pd.read\_csv("D:/DataSets/iris.csv")

data.head()

data.describe()

data.drop(data.columns[data.columns.str.contains('unnamed',case = False)],axis = 1, inplace = True)

**#rename columns**

data.rename(columns = {'Sepal.Length':'sl','Sepal.Width':'sw','Petal.Length':'pl','Petal.Width':'pw','Species':'species'}, inplace=True)

data.columns

**#Sepal Length**

data['sl\_new'] = pd.cut(data['sl'], bins=[min(data.sl) - 1, data.sl.mean(), max(data.sl)], labels=["Low","High"])

data.head()

data.sl\_new.value\_counts()

data.columns

**#Sepal Width**

data['sw\_new'] = pd.cut(data['sw'], bins=[min(data.sw) - 1, data.sw.mean(), max(data.sw)], labels=["Low","High"])

data.head()

data.sw\_new.value\_counts()

data.columns

**#Petal Length**

data['pl\_new'] = pd.cut(data['pl'], bins=[min(data.pl) - 1, data.pl.mean(), max(data.pl)], labels=["Low","High"])

data.head()

data.pl\_new.value\_counts()

data.columns

**#Petal Width**

data['pw\_new'] = pd.cut(data['pw'], bins=[min(data.pw) - 1, data.pw.mean(), max(data.pw)], labels=["Low","High"])

data.head()

data.pw\_new.value\_counts()

data.columns

**Conclusion:** Hence the columns are categorized as Low and High which called as discretisation.