**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:**

**Batch Id:**

**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 |

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| Sepal.Length | Length of Sepal | Quantitative,Continuous, Ratio | Relevant |
| Sepal.Width | Width of Sepal | Quantitative,Continuous, Ratio | Relevant |
| Petal.Length | Length of Petal | Quantitative,Continuous, Ratio | Relevant |
| Petal.Width | Width of Petal | Quantitative,Continuous, Ratio | Relevant |
| Species | Species Name | Quanlitative,Discrete, Nominal | Relevant |

**DATA UNDERSTANDING:**

1. There are no missing values in the dataset.
2. The data is continuous(float) in nature except Species(String)

**INSIGHTS FROM THE DATA:**

1. We have plotted a distribution plot with the continuous numerical value but didn’t have a standard visualization.
2. Then the numerical data is converted to discretized bins.
3. We plotted the discretized data distribution plot and we got a standard visualization plot.
4. We can also discretize the data by converting into bins by cut or qcut method.
5. We have used the pd.cut method and discretize the data into 3 bins.

Inferences: Binning the data can be a very useful strategy while dealing with numeric data to understand certain trends. Sometimes, we may need an age range, not the exact age, a profit margin not profit, a grade not a score. The Binning of data is very helpful to address those. Pandas library has two useful functions [cut](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.cut.html) and [qcut](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html) for data binding.

The discretization transform provides an automatic way to change a numeric input variable to have a different data distribution, which in turn can be used as input to a predictive model.

**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