National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

Machine Learning for Engineering Applications – CT1 **Date: 06.03.2021**

Duration: 1 Hr **Time:** 05:30 – 06:30 PM

Total Marks: 20

1. If a dataset has the target labels associated with all the samples and the machine learning algorithm also considers all these labels to take a decision, then which category does this learning algorithm belong to? **(1 M)**

   (a) Supervised **(b)** Unsupervised **(c)** Semi-supervised **(d)** Reinforcement

2. (i) Draw Box-and-Whisker Plot for the values: **4, 7, 9, 8, 12, 80, 15** **(4 + 2 = 6 M)**

   (ii) Discuss about the spread of data in the above plot using Q1, Q2, Q3 and IQR values.

3. Match the following: **(1 M)**

A. Feature Binning	(i) Creates separate features for each unique value that is present in categorical column
B. Feature Engineering	(ii) Converts the "n" unique values in categorical columns to values between 0 and n-1
C. Label Encoding	(iii) Should be utilized when one wants to replace an existing feature with more meaningful additional features
D. One Hot Encoding	(iv) Should be applied on columns that has large number of unique values

   (a) A -> (ii); B -> (iii); C -> (iv); D -> (i) **(b)** A -> (iii); B -> (i); C -> (iv); D -> (ii)

   (c) A -> (i); B -> (iv); C -> (ii); D -> (iii) **(d)** A -> (iv); B ->(iii); C -> (ii); D -> (i)

4. (i) Write the names of various binning methods. **(1.5 + 5 = 6.5 M)**

   (ii) Consider the following data and apply any two binning methods *[**Hint:** Bin Size = 3]*

   15, 21, 45, 6, 11, 17, 45, 19, 12, 4, 9, 5

5. Write the name of the function that has to be utilized in order to display the following rows of a data frame: (i) First 5 rows; (ii) Last 5 rows. **(1 M)**

Dr. R. Bala Krishnan, Assistant Professor, Department of CSE

6. (i) Write the various values that can be utilized to replace the NULL values. **(1 + 1 = 2 M)**
   (ii) What is the role of the parameter “inplace = True” in “fillna” method.

7. Write the name of the function that can be utilized to define the datatype of a particular column. **(1 M)**

8. Write the names of various visualization techniques. **(1.5 M)**

---------- END ----------

Satyam Singh
11211906
CT-1
CSOE18

Sol8.

1. Column Chart
2. Bar graph
3. Stacked Bar graph
4. Line Graph
5. Dual-Axis Chart
6. Mekko chart
7. Pie chart
8. Scatter plot Chart
9. waterfall chart
10. Funnel chart.

Sol5. (i) df.head()
(ii) df.tail()

Sol6. (i)
- Replace with default value
- Mean
- Median
- Mode

(ii) inplace=true, updates dataframe in which you are working on

Sol1. a) Soperviséd

Sol3 d) A - (iv) B - (iii)
C - (ii) D - (i)

Sol7. dtype = {"user-id" : int}
It is used to define the datatype of particular column while loading a database.
astype() is used to change the datatype of a particular column while working on a database

Sol4. (i) There are two types of binning:
- Unsupervised Binning :- • Eaual width binning
  • Eaual freaveney binning.
- Supervised Binning :- • Entropy based binning

112119066

Sol4.(ii)

Data :- 15, 21, 45, 6, 11, 17, 45, 19, 12, 4, 9, 5

Data in ascending order:-

4, 5, 6, 9, 11, 12, 15, 17, 19, 21, 45, 45

So,

By applying Bin mean (Binsize=3)

Bin1 4 5 6 ⇒ mean = 5

Bin2 9 11 12 ⇒ mean = 10.67

Bin3 15 17 19 ⇒ mean = 17

Bin4 21 45 45 ⇒ mean = 37

By Bean mean

Bin 1 5 5 5 ⇒ mean = 5

Bin 2 10.67 10.67 10.67 ⇒ mean = 10.67

Bin 3 17 17 17 ⇒ mean = 17

Bin 4 37 37 37 ⇒ mean = 17

By applying Bin Boundry method

Bin 1 4 4 6

Bin 2 9 12 12

Bin 3 15 15 19

Bin 4 21 45 45

112119066

Sol2.

(i) Values:- 4, 7, 9, 8, 12, 80, 15

Values in ascending order :- 4, 7, 8, 9, 12, 15, 80

Q_1 = 7, Q_2 = 9, Q_3 = 15

Inter - Quartile Range

$(IQR) = Q_3 - Q_1 = 15 - 7$

$= 8$

1st quartile = 7
2nd quartile = 9
3rd quartile = 15
Smallest = 4
Largest = 80

Lower whicker $= Q_1 - \frac{3}{2} \times (IQR)$

$= 7 - \frac{3}{2} \times 8$

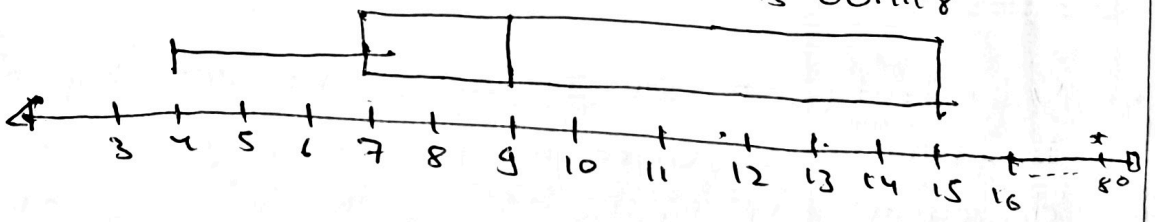
$= -5$

Upper whicker $= Q_3 + \frac{3}{2} \times (IQR)$

$= 15 + \frac{3}{2} \times 8$

$= 27$

$4 > -5$ and $80 > 27$ so 80 is outlier



(ii) The above represented data is positively skewed and the data between Q_1 and Q_2 is closely packed since $Q_2 - Q_1 < Q_3 - Q_2$ and data between Q_2 and Q_3 is loosly packed.

80 is considered as outlier.