Date: 27/01/23

Assignment#02

Name Sufyan Ahmad
RollNo 20198013
Subject Data Mining
Section A

Q#02

(a)

The entities identification problem in data integration refers to the process of identifying and resolving inconsistencies in the naming and representation of entities across different data sources for example:- In a data integration scenario involving a company's customer information stored multiple databases the entity "Customer" may be represented differently in each database (e.g "Client" in one database and "patron" in another). Identifying and resolving these inconsistencies.

Date: 27/01/23 Name: Sufyan Ahmad RollNo: 2019BCO13

(b)

For Nominal data redundancy analysis can be implemented by creating a contingency table where the frequencies of each combination of variable values are computed. **For example:** In data set with "Gender" and "Marital Status" variables, a contingency table can be created to determine the number of males and females in each marital status. If there is a high correlation between the two variables it may be redundant to include both variable in the data set.

For Numeric Data Correlation Analysis can be implemented by computing the correlation coefficient between two variables. A correlation cofficient of 1 indicates a perfect positive correlation a ~~negative~~ cofficient of -1 indicates a perfect negative correalation and a cofficient of

Date: 27/01/23 Name: Sufyan Ahmad RollNo: 20198013

0 indicates no correlation. For example: if there is a high correlation between a customer's income and their likelihood of making a purchase, it may be redundant to include both variables in the data set.

#01

Let $(42, 5, 22, 15)$ and $(35, 1, 15, 13)$
$x_1\ y_1\ z_1\ m_1$ $x_2\ y_2\ z_2\ m_2$

(a) Euclidean distance

Formula $= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2+(m_2-m_1)^2}$

$= \sqrt{(35-42)^2+(1-5)^2+(15-22)^2+(13-15)^2}$

$= \sqrt{(-7)^2+(-4)^2+(-7)^2+(-2)^2}$

$= \sqrt{49+16+49+4}$

$= \sqrt{118}$

$= 10.8627$

Date: 27/01/23 Name: Sufyan Ahmad Roll No: 2019B0013

(b)

Manhattan distance Formula

$$= \{|x_2 - x_1| + |y_2 - y_1| + |z_2 - z_1| + |m_2 - m_1|\}$$

$$= \{|35 - 42| + |1 - 5| + |15 - 22| + |13 - 15|\}$$

$$= \{|-7| + |-4| + |-7| + |-2|\}$$

$$= \{7 + 4 + 7 + 2\}$$

$$= 20$$

(C)

Supremum distance Formula

$$= \max\{|x_2 - x_1|, |y_2 - y_1|, |z_2 - z_1|, |m_2 - m_1|\}$$

$$= \max\{|35 - 42|, |1 - 5|, |15 - 22|, |13 - 15|\}$$

$$= \max\{|-7|, |-4|, |-7|, |-2|\}$$

$$= \max\{7, +4, 7, 2\}$$

$$= 7$$

Date: 27/01/23 Name: Sufyan Ahmad RollNo: 20198013

Q#03

Min

33, 35, 36, 36, 39, 40, 40, 41, 42, 42,
45, 45, 45, 45, 50, 53, 53, 55, 55, 55,
55, 56, 60, 65, 66, 72, 90

Q_2

max

(a)

Bin 1: 33, 35, 36, 36, 39, 40, 40, 41, 42

Bin 2: 42, 45, 45, 45, 45, 50, 53, 53, 55

Bin 3: 55, 55, 55, 56, 60, 65, 66, 72, 90

$$\text{Bin1} := \frac{33+35+36+36+39+40+40+41+42}{9}$$

$$= 342/9 \;, = 38$$

$$\text{Bin2} := \frac{42+45+45+45+45+50+53+53+55}{9}$$

$$= 433/9 = 48.11 \approx 48$$

$$\text{Bin3} := \frac{55+55+55+56+60+65+66+72+90}{9}$$

$$= \frac{574}{9} \;, = 63.77 \approx 64$$

Date: 27/01/23 Name: Sufyan Ahmed RollNo: 20198013

Bin1: 38, 38, 38, 38, 38, 38, 38, 38, 38

Bin 2: 48, 48, 48, 48, 48, 48, 48, 48, 48

Bin 3: 64, 64, 64, 64, 64, 64, 64, 64, 64

1b)

min = 33

max = 90

Q_2 = 45

Q_1 = 40+41/2 = 40.5

Q_3 = 55+55/2 = 55

IQR = $Q_3 - Q_1$ = 55 − 40.5 = 14.5

upper-limit = $Q_3 + 1.5(IQR)$

= 55 + 1.5 (14.5)

= 55 + 21.75

= 76.75

lower-limit = $Q_1 - 1.5(14.5)$

= 40.5 − 1.5(14.5)

= 40.5 − 21.75

= 18.75

outliner = 90

- It is upper-limit is greater. 90 and lower-limit has no outliners.

Date: 27/01/22 Name: Sufyan Ahmad RollNo: 20198003

(c)

min-max normalization

min = 33, max = 90, V = 55

$$V' = \frac{V - min_A}{max_A - min_A}(new\_max - new\_min_A) + new\_min_A$$

$$= \frac{55 - 33}{90 - 33}(1.0 - 0.0) + 0.0$$

$$= \frac{22}{57}(1.0) + 0.0$$

$$= \frac{22}{57}(1.0) + 0.0$$

$$= 0.3859(1.0)$$

min-max = 0.3859

(d)

V = 35

$$V' = \frac{V - \mu_A}{\sigma_A}$$

$$\mu = \frac{33+35+36+36+39+40+40+41+42+42+45+45+45+45+50+53+53+55+55+55+55+56+60+65+66+72+90}{27}$$

Date: 27/01/23 Name: Sufyan Ahmad Roll No: 20198013

$$\mu = \frac{1349}{27} = 49.96$$

$\sigma = 12.94$ (using calculator)

$$= \frac{35 - 49.96}{12.94}$$

$$= \frac{-14.96}{12.94} = -1.156$$