Homework#2.1

6410504055 น.ส.ชาลิสา สกุลวิจิตรัสันธุ

Q0. สมมติว่าคุณเป็นนักวิเคราะห์ข้อมูลที่ทำงานให้กับบริษัทอีคอมเมิร์ซ โปรดระบุแหล่งข้อมูลหลักอย่างน้อยสามแหล่งและแหล่งข้อมูลรองอย่างน้อยสามแหล่งที่คุณจะใช้ในการวิเคราะห์พฤติกรรมการซื้อของลูกค้า อธิบายว่าแต่ละแหล่งข้อมูลมีประโยชน์อย่างไร

แหล่งข้อมูลหลัก

1) ข้อมูลการซื้อขาย (Transaction Data) เป็นข้อมูลที่ช่วยในการวิเคราะห์พฤติกรรมการซื้อของลูกค้าได้โดยตรง เนื่องจากสามารถวิเคราะห์ได้หลากหลายมุม เช่น สามารถวิเคราะห์ว่าสินค้าหรือผลิตภัณฑ์ไหนที่ลูกค้ามักเลือกซื้อ , ช่องทางการชำระเงินไหนที่ลูกค้าส่วนใหญ่เลือก

2) ข้อมูลของการเข้าใช้งานแพลตฟอร์ม เช่น การเลือกเข้าชม , การยกเลิกสินค้าภายในตะกร้า สามารถช่วยวิเคราะห์พฤติกรรมการซื้อได้ เช่น ลูกค้าคนนี้มักเลือกเข้าชมสินค้าประเภทนี้ซึ่งนำไปสู่การแนะนำสินค้าเพื่อพัฒนายอดขาย

3) ข้อมูลส่วนตัว เช่น ข้อมูลเพศ , อายุ , สถานะภาพ , รายได้ ซึ่งได้จากตอนลงทะเบียนหรือการทำแบบสอบถาม โดย สามารถแบ่งกลุ่มลูกค้าเพื่อวิเคราะห์พฤติกรรมการซื้อได้ เช่น วัยรุ่นมีพฤติกรรมการซื้อเสื้อผ้ามากกว่าวัยเด็ก

แหล่งข้อมูลรอง

1) ข้อมูลประชากรศาสตร์จากรัฐบาล เพื่อวิเคราะห์โครงสร้างหลักของประชากร นำไปสู่การปรับปรุงให้ตอบสนองต่อความต้องการและพฤติกรรมของลูกค้า

2) ข้อมูลเศรษฐกิจในปัจจุบันจากนักวิเคราะห์ เพื่อวิเคราะห์กำลังซื้อและความต้องการซึ่งมีผลต่อพฤติกรรมการซื้อของลูกค้า

3) ข้อมูลจากโซเชียลมีเดีย เช่น ข้อมูลการแสดงความคิดเห็น , การรีวิว , การแชร์ มีผลต่อการตัดสินใจเลือกซื้อของลูกค้า

Q1. จงอธิบายและเปรียบเทียบข้อมูลสองประเภทต่อไปนี้ พร้อมยกตัวอย่างข้อมูลแต่ละประเภท

a) Categorical Data และ Numerical Data
b) Nominal Data และ Ordinal Data
c) Interval Data และ Ratio Data

a) Categorical Data หรือ ข้อมูลเชิงคุณภาพ คือ ข้อมูลเชิงหมวดหมู่ที่ใช้แยกประเภทหรือกลุ่ม โดยไม่มีการเปรียบเทียบกันในเชิงปริมาณ สามารถแบ่งได้เป็น Nominal Data , Ordinal Data และ Binary Data คือข้อมูลที่มี 2 ประเภท เช่น true/false , yes/no ยกตัวอย่างข้อมูลเชิงคุณภาพ ได้แก่ เพศ , สี , ประเทศ , เกรด , สถานะ
Numerical Data หรือข้อมูลเชิงปริมาณ คือ ข้อมูลในเชิงตัวเลข ที่สามารถวัดและนำมาคำนวณได้ ยกตัวอย่างข้อมูลเชิงปริมาณ ได้แก่ อายุ , จำนวนนักเรียน , ส่วนสูง , น้ำหนัก

b) Nominal Data คือข้อมูลเชิงคุณภาพที่ไม่มีการจัดลำดับ สามารถเปรียบเทียบได้จากความเหมือนไม่เหมือน ส่วนใหญ่อยู่ในลักษณะของประเภท , สถานะ ยกตัวอย่าง Nominal Data ได้แก่ สีผม , เลขรหัสไปรษณีย์ , รหัสประจำตัวประชาชน
Ordinal Data คือข้อมูลเชิงคุณภาพที่มีการจัดลำดับ แต่ระยะห่างระหว่างลำดับไม่มีความหมายในเชิงปริมาณ ยกตัวอย่าง Ordinal Data ได้แก่ เกรด (A, B, C, D) , ขนาด (เล็ก, กลาง, ใหญ่)

c) Interval Data หรือข้อมูลช่วง เป็นข้อมูลประเภทข้อมูลเชิงปริมาณ คือ ข้อมูลที่สามารถนำมาบวก, ลบได้ และเป็นข้อมูลที่ไม่มี 0 จริง ยกตัวอย่าง Interval Data ได้แก่ อุณหภูมิ เช่น 0°C , 50°C คือ 0°C ไม่ได้แปลว่าไม่มีอุณหภูมิ , ข้อมูลปี
Ratio Data หรือข้อมูลอัตส่วน เป็นข้อมูลประเภทข้อมูลเชิงปริมาณ คือ ข้อมูลที่สามารถนำมาบวก, ลบ, คูณและหารได้ เป็นข้อมูลที่มี 0 จริง ยกตัวอย่าง Ratio Data ได้แก่ ส่วนสูง เช่น 100 ซม , 0 ซม. คือ 0 ซม.แปลว่าไม่มีส่วนสูงจริงๆ , อายุ , น้ำหนัก , รายได้

6410504055 น.ส.ชาลิสา สกุลวิจิตร์สันธุ

Q2. กำหนดให้ข้อมูลต่อไปนี้

X = [1, 1, 0, 1, 0] และ Y = [0, 1, 1, 1, 0)

จงคำนวณหาค่าความเหมือน (similarity) ระหว่าง X และ Y ด้วยมาตรวัดต่อไปนี้

a) Simple Matching Coefficient
b) Jaccard Coefficient
c) Cosine Similarity

a) Simple Matching Coefficient

$$SMC = \frac{q+d}{a+b+c+d} = \frac{2+1}{2+1+1+1} = \frac{3}{5} = 0.6$$

โดย q = จำนวนของคุณลักษณะที่ทั้ง 2 ตัวอย่างมีค่าเหมือนกันและเป็น 1
d = จำนวนของคุณลักษณะที่ทั้ง 2 ตัวอย่างมีค่าเหมือนกันและเป็น 0
b = จำนวนของคุณลักษณะที่ตัวอย่างแรกมีค่าเป็น 1 และตัวอย่าง 2 มีค่าเป็น 0
c = จำนวนของคุณลักษณะที่ตัวอย่างแรกมีค่าเป็น 0 และตัวอย่าง 2 มีค่าเป็น 1

b) Jaccard Coefficient

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

$$\frac{q}{q+r+s} = \frac{2}{2+1+1} = \frac{2}{4} = 0.5$$

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

c) Cosine Similarity

$$\cos(x,y) = \frac{x \cdot y}{|x||y|} = \frac{(1\times0)+(1\times1)+(0\times1)+(1\times1)+(0\times0)}{(\sqrt{1^2+1^2+0^2+1^2+0^2})(\sqrt{0^2+1^2+1^2+1^2+0^2})} = \frac{2}{3} = 0.667$$

Q3. ให้จุดสองจุดในพื้นที่สามมิติ

A = [2, 4, 6] และ B = [1, 3, 5] จงคำนวณหาค่า Euclidean Distance ระหว่าง A และ B แสดงวิธีการคำนวณทีละขั้นตอน

Euclidean Distance

$$d_{ij} = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$$

$$d_{ij} = \sqrt{(2-1)^2+(4-3)^2+(6-5)^2}$$
$$d_{ij} = \sqrt{3}$$
$$d_{ij} = 1.732$$

6410504055 น.ส.ชาลิสา สกุลวิจิตรัสินธุ

Q4. Consider the following dataset where X1, X2 are Nominal and X3, X4 are Numerical.

X1	X2	X3	X4
A	A	3.6	3.5
B	B	3.2	3.3
C	C	2.8	2.9
A	A	3.8	3.9
B	D	3.3	3.4
B	D	3.4	3.5
A	C	3.7	3.8
A	A	3.9	4.0

Answer the following questions. Show your calculation step.

a) Calculate Similarity between X1 and X2.
b) Calculate Dissimilarity between X1 and X2.
c) Calculate Dissimilarity between X3 and X4.

a) $SMC = \frac{S}{S+3} = \frac{5}{8} = 0.625$

b) $Dissimilarity = 1 - SMC = 1 - \frac{5}{8} = 0.375$

c) $Euclidean\ Distance = \sqrt{(3.6-3.5)^2+(3.2-3.3)^2+(2.8-2.9)^2+(3.8-3.9)^2+(3.3-3.4)^2+(3.4-3.5)^2+(3.7-3.8)^2+(3.9-4.0)^2}$

$= \sqrt{0.08}$

$= 0.2828$