- Dataset ที่เลือกใช้

Dataset ที่เลือกใช้เป็นชุดข้อมูลที่เกี่ยวกับการตัดสินใจเลือกซื้ออุปกรณ์
อิเล็กทรอนิกส์ระหว่างLaptop(Notebook) กับ Tablet(iPad, Galaxy Tabs) ขอบเขตคือ
เฉพาะนักศึกษาในมหาลัยลาดกระบัง เท่านั้น มี Attribute ทั้งหมด 9 อย่าง คือ grade(ชั้น
การศึกษา),faculty(คณะ),department(สาขา),price(งบในการซื้อเป็นช่วงตัวเลข),
purpose1(จุดประสงค์ในการซื้อลำดับที่1), purpose2(จุดประสงค์ในการซื้อลำดับที่2)
,purpose3(จุดประสงค์ในการซื้อลำดับที่3),frequency(ความถี่ในการพกพาต่ออาทิตย์)
,gender(เพศ) มีจำนวนข้อมูลทั้งหมด ทั้งสิ้น 303 ตัวอย่าง จำนวนคนที่ตอบ
Tablet มีจำนวน 153 ตัวอย่าง และ laptop 150 ตัวอย่าง และประเภทของข้อมูลทุก Attribute
เป็นข้อมูลประเภท Nominal

ก่อนที่จะนำ Dataset นี้เข้าไปยังโมเคล ขั้นตอนแรกจะทำการ Encode data ให้ กลายเป็นตัวเลขเพื่อให้เอาไปสามารถคำนวณในตัวโมเคลต่างๆได้ หลังจาก Encode เสคจะ ทำให้ตัดตัวอย่างออกให้ Dataset เกิดความ Balance กัน เป็นดังรูปต่อไปนี้

dataset
dataset_encoded_balance

11/10/2020 10:37 AM Microsoft Excel C... 12/18/2020 9:50 PM Microsoft Excel C...

114 KB

12 KB

ไฟล์Datasetทั้งหมด

A	В	C	D	E	F	G	Н	1	J
1 grade	faculty	department	price	purpose1	purpose2	purpose3	freaquency	gender	class label
2 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	มากกว่า 30,000 บาท	เขียนโปรแกรม	การศึกษา	การศึกษา	3-5วัน	ชาย	Laptop (Notebook)
3 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	มากกว่า 30,000 บาท	เล่นเกม	การศึกษา	การศึกษา	1-2วัน	ชาย	Laptop (Notebook)
4 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	มากกว่า 30,000 บาท	เล่นเกม	การศึกษา	การศึกษา	ไม่พกพาเลย	ชาย	Laptop (Notebook)
5 นักศึกษาชั้นปีที่ 2	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	20,001 - 25,000 บาท	เขียนโปรแกรม	เล่นเกม	เล่นเกม	3-5วัน	ชาย	Laptop (Notebook)
6 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	5,001 - 10,000 บาท	การศึกษา	สื่อบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	สือบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	ทุกวัน	หญิง	Tablet (Ipad , Galaxy Tabs)
7 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	20,001 - 25,000 บาท	เขียนโปรแกรม	การศึกษา	การศึกษา	1-2วัน	ชาย	Laptop (Notebook)
8 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	15,001 - 20,000 บาท	การศึกษา	สือบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	สือบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	ทุกวัน	ชาย	Tablet (Ipad , Galaxy Tabs)
9 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	25,001 - 30,000 บาท	การศึกษา	ดัดต่อวิดิโอ	ตัดต่อวิดีโอ	ไม่พกพาเลย	หญิง	Laptop (Notebook)
0 นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	25,001 - 30,000 บาท	การศึกษา	เล่นเกม	เล่นเกม	3-5วัน	ชาย	Laptop (Notebook)

ตัวอย่างไฟล์Dataset ตอนแรก

	Α	В	С	D	E	F	G	Н	1
1	faculty	price	purpose1	purpose2	purpose3	freaquency	gender	class label	
2	2	6	5	0	0	1	1	Laptop (N	otebook)
3	2	6	6	0	0	0	1	Laptop (N	otebook)
4	2	6	6	0	0	3	1	Laptop (N	otebook)
5	2	2	5	7	7	1	1	Laptop (N	otebook)
6	2	4	0	4	4	2	2	Tablet (Ipa	ad , Galaxy Tabs)
7	2	2	5	0	0	0	1	Laptop (N	otebook)
8	2	1	0	4	4	2	1	Tablet (Ipa	ad , Galaxy Tabs)
9	2	3	0	1	1	3	2	Laptop (N	otebook)
10	2	3	0	7	7	1	1	Laptop (N	otebook)

ตัวอย่างไฟล์ Dataset หลังจากทำการ encode และ balance

หลังจากทำการ Encode และ Balance ตัว Dataset แล้วจะเหลือจำนวน Dataset ทั้งหมด 300 ตัว

- อธิบายผลการทคลอง

Dataset -ทำการแบ่ง Train Set และ Test Set เป็นจำนวน 75% และ 25% โดยไม่มีการrandom สมาชิกในแต่ละแบบ

Direct Classification -เลือกใช้ K-Nearest Neighbors Algorithm

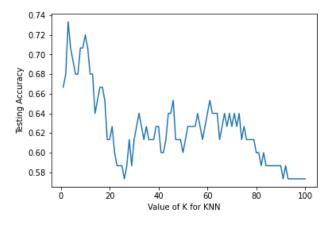
Parameter ที่ใช้ มี 2 อย่าง คือ weights ใช้ distance
และ n_neighbors จำนวนเพื่อนบ้าน ทคลอง โดยรันสุ่มค่า kไป
เรื่อยๆออกมาเป็นกราฟเพื่อหาดูว่า ค่า k ใดที่ให้ค่า Accuracy
สูงสุด กับ Dataset มากที่สุด ทำช่วงไว้ที่ 1-100 และดู
Confusion Matrix

Traditional classification -เลือกใช้ Decision Tree Parameter ที่ใช้มี 2อย่าง คือ
random_state=0 คือจะ ไม่มีการสุ่มต้น ไม้มั่วๆ max_depth
กำหนดความถึกของต้น ไม้ โดยจะ กำหนด ไว้ที่ 2 กับ ไม่
กำหนดขั้นต่ำเพื่อดูผลลัพธ์ของ ต้น ไม้ทดลอง โดยการรัน
และดู Confusion Matrix และ ดูค่า Accuracy

Deep Learning -เลือกใช้ Multilayer Perceptron Classifier Parameter ที่ใช้มี 4อย่าง คือ alpha คือ ค่า regularization โดยใช้ L2 penalty ข้อดีของมันคือ ทำให้เราควบคุม ค่า weight ได้ กำหนดไว้ที่ 0.01เป็นค่าเริ่มต้น hidden layer sizes คือ จำนวนโหนดในชั้นซ่อน กำหนดไว้ที่ 8 เป็นค่าเริ่มต้น มาจาก Log n ฐาน 2 โดยที่ n เป็นจำนวนDataset learning_rate คือ ค่าการเรียนรู้ยิ่งน้อยยิ่งดี กำหนดไว้เป็น adaptive คือสามารถปรับค่า learning rate ระหว่างคิดได้ max_iter คือ จำนวนรอบที่คำนวณ กำหนดเริ่มต้นที่ 1000 วิธีการทดลอง ปรับparameter ไปเรื่อยๆ โดยการปรับ parameter แต่ละครั้ง จะทำ การวัด Confusion Matrix และ Accuracy ทั้งหมด 10 รอบ เนื่องจาก ตัว MLPClassifier เมื่อเริ่มต้น model ค่า weight จะถูกสุ่มเลยทำให้ การวัคค่า Accuracy ในแต่ละครั้งไม่เท่ากันเลยตัดสินใจรัน 10 ครั้ง เพื่อหาค่าเฉลี่ย ในการทดลองนี้ทำการดูเวลาที่ใช้ในการคำนวณ คร่าวๆด้วย

ผลการทดลอง

K-Nearest Neighbors Algorithm -ทำการทดลองค่า k เพื่อหาค่า k ที่ทำให้เกิดค่า Accuracy สูงที่สุด เป็นดังกราฟต่อไปนี้

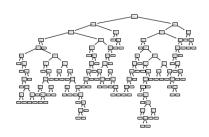


จากกราฟนี้พบว่า ค่า k ที่ทำให้เกิดค่า Accuracy สูงที่สุด คือ k=3 ได้ 73.33% แต่การที่จะหาค่า K ที่เหมาะสมนั้น ต้องดูจากภาพรวมของ Data Set ทั้งหมดมีสูตรการหาค่า k คือ k=sqrt(n) โดยที่ n คือ จำนวน Data Set

Decision Tree -ทำการทดลองรันโดยเปลี่ยน จำนวนความลึกของต้นใม้เป็นรูปดังนี้



ต้นไม้ที่มีความลึกเท่ากับ 2 ค่า Accuracy = 78.67%



ต้นไม้ที่มีความลึกไม่จำกัด ค่า Accuracy = 69.33%

Multilayer Perceptron Classifier -ทำการเปลี่ยน parameter เป็นเรื่อยๆ รันแต่ละครั้ง
เป็นจำนวน 10 รอบ เพื่อหาค่า เฉลี่ย Accuracy
ผมทำสรุปไว้ใน Excel ครับจะอยู่ในไดฟ์

2 sau Cora 3 1 TP- 4 2 TP- 5 3 TP- 6 4 TP- 7 5 TP- 8 6 TP- 9 7 TP- 0 8 TP- 1 9 7 TP- 2 10 TP- 2 10 TP- 3 3 Cora 4 5 parameter: alap 7 1 TP- 8 2 TP- 9 3 TP- 9 3 TP-	Julia - Gill. Juliani, Javar -	Train_Set-e25 Accuracy (8,000) (9,33) (6,001) (6,07) (6,007) (6,007) (7,007) (8,000) (6,77) (8,000) (6,77) (8,000) (7,33) (1,36) (1,37)	6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Data=300	To Luptop = Positive , Tablet = Negative To Luptop = Positive , Tablet = Negative To Luptop = Positive , Tablet = Negative
3 1 1 104 4 2 179 5 3 179-6 6 4 179-7 7 5 179-8 8 6 179-9 7 79-10 10 8 179-10 11 9 179-12 12 10 179-13 14 14 15 parameter: alp desired for aux Cord	** IL Thread, J. Perd, J. Perd	68,000 69,331 64,000 68,677 64,000 68,007 68,000 48,000 48,000 70,077 Train_Set=25 Accuracy 73,331	6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Data=300	bf Luptop = Positive , Tablet = Negative
4 2 TP- 5 3 TP- 6 4 TP- 7 5 TP- 8 6 TP- 9 7 TP- 10 8 TP- 11 9 TP- 12 10 TP- 13 14 15 parameter: alap 14 15 parameter: alap 17 1 TP- 18 3 TP- 19 3 TP-	** IL THO AS A PACE AND ** **	69.33' 66.00' 66.00' 66.00' 68.00' 68.00' 70.07' 68.00' 65.33' valide=67.07' Train_Set=225 Accuracy 73.33'	6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Deta=300	To Laptop = Positive , Tablet = Negative
5 3 TP: 6 4 TP: 7 5 TP: 8 6 6 TP: 9 7 TP: 11 9 TP: 12 10 TP: 13 14 14 15 parameter: alp 16 1au Co: 17 1 TP: 18 2 TP: 19 3 TP: 19 3 TP:	he 10, 10-20, 20	64.00' 66.77' 64.00' 68.00' 68.00' 70.07' 88.00' 88.00' 88.00' 88.00' 88.00' 10	ง 6 6 6 6 6 6 6 7 8 12คาที่ใช้ = slat	Data=300	Tvf Laptop = Positive , Tablet = Negative
6 4 TP/ 5 TP- 8 6 TP- 9 7 TP- 10 8 TP- 11 9 TP- 12 10 TP- 13 Dereter: alpho- 14 Sparameter: alpho- 16 sau Cor 17 1 TP- 19 3 TP- 19 3 TP-	** ILT MEST 2 PARL PINE* **PET TIMEST 1 PARL PI	66.67 64.00 68.00 70.67 70.67 68.00 63.33 salfer-67.07% Train_Set=22S Accuracy 73.33	6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Data=900	Tw Luptop = Positive , Tablet = Negative
7 5 TP: 8 6 TP: 9 7 TP: 10 8 TP: 11 9 TP: 12 10 TP: 13 14 14 15 parameter: alp 16 tau Cor 17 1 TP: 18 2 TP: 19 3 TP:	For T, The SLI, PAGA Media For SLI, The SLI,	64.00' 68.00' 70.67' 70.67' 68.00' 10.67' 10.00' 10.31' 10.48' 10.57' 17ain_Set=225 Accuracy 73.33'	6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Data=300	Tw Laptop = Positive , Tablet = Negative
8 6 TP1 9 7 TP1 10 8 TP1 11 9 TP1 12 10 TP1 13 14 15 parameter: alp 16 Satu Coi 17 1 TP1 18 2 TP1 19 9 3 TP1	Part J. The St. J. Part J. The St. J. Part J. The St. J. Part	68.00' 66.67' 70.676' 68.00' 63.33' vaffe-67.07% Train_Set=22S Accuracy 73.33'	ร์ ร์ ร์ ร์ เวลาที่ใช้ =น่อย Test_Set=75	Data=300	Tw Laptop = Positive , Tablet = Negative
9 7 TPs 10 8 TPs 11 9 TPs 12 10 TPs 13 14 15 parameter: alp 16 tats Cos 17 1 TPs 18 2 TPs 19 3 TPs	Host, Timos J., Price, Zimos Host, Timos J., Price, J., Reaming, Ender-Indiantive, max. Jen-1000 Inchiston, Music.	66.67 70.67 68.00 65.33 salle=67.07% Train_Set=225 Accuracy 73.33	% % % เวลาที่ใช้ =น้อย Test_Set=75	Data=300	Tw Laptop = Positive , Tablet = Negative
10 8 TP- 11 9 TP- 12 10 TP- 13 14 15 parameter: alp 16 tatu Cor 1 TP- 18 2 TP- 19 3 TP-	Po 13,714-64, FP-02,714-0 Po 13,714-62,714-0 Po 12,714-0 Po 12,714-0 Po 12,714-0 Po 12,714-0 Po 12,714-0 Po 13,714-0 Po 14,714-0 Po 14,714	70.675 68.007 65.331 valid=67.075 Train_Set=225 Accuracy 73.331	ร์ เรลาที่ใช้ =น่อย Test_Set=75	Data=300	lví Laptop = Positive , Tablet = Negativi
9 TP: 12 10 TP: 13 14 14 15 parameter: alpha 16 Satu Coi 17 1 TP: 18 2 TP: 19 3 TP:	Ped 3, Triko 3, Prio 2, Prio 2	68.001 65.331 เฉลีย=67.07% Train_Set=225 Accuracy 73.331	ร์ เวลาที่ใช้ =ฟอย Test_Set=75	Data=300	11/ Laptop = Positive , Tablet = Negative
12 10 TP- 13 14 15 parameter: alp 16 1au Coi 17 1 TP- 18 2 TP- 19 3 TP-	n-20,1%-29,f%-21,f%-5 ghan0.5,hidden_layer_sizes=(8), learning_rater*adaptive*, max_fler=1000 onfusion_Matrix =22,17%-34,f%-920,f%-00	65.33 salite=67.07% Train_Set=225 Accuracy 73.33	รลาที่ใช้ =พ่อย Test_Set=75	Data=300	Tw Laptop = Positive , Tablet = Negative
13 14 15 parameter : alp 16 sau	ighand-5,hidden_layer_sizes=(8,), learning_rate='adaptive', max_ite==1000 onfusion_Matrix =2-1,11%-14, Pe-20,Pk=0	เฉลีย=67.07% Train_Set=225 Accuracy 73.331	เวลาที่ใช้ =น้อย Test_Set=75	Data=300	Tw Laptop = Positive , Tablet = Negative
15 parameter : alp 16 sau Coi 17 1 TP: 18 2 TP: 19 3 TP:	onfusion_Matrix P=21,TN=34,FP=20,FN=0	Train_Set=225 Accuracy 73.331	Test_Set=75	Data=300	ให้ Laptop = Positive , Tablet = Negative
15 parameter : alp 16 sau Cor 17 1 TP: 18 2 TP: 19 3 TP:	onfusion_Matrix P=21,TN=34,FP=20,FN=0	Accuracy 73.33	_	Data=300	ให้ Laptop = Positive , Tablet = Negative
16 Sau Coi 17 1 TP1 18 2 TP1 19 3 TP1	onfusion_Matrix P=21,TN=34,FP=20,FN=0	Accuracy 73.33	_	Data=300	ให้ Laptop = Positive , Tablet = Negative
17 1 TP: 18 2 TP: 19 3 TP:	P=21 ,TN=34 ,FP=20,FN=0	73.33	6		
18 2 TP- 19 3 TP-		73.33	6		
19 3 TP:	P= 20,TN=28,FP=21,FN=6	64.00			
			6		
	P=20 ,TN=32 ,FP=21,FN=2	69.33	16		
20 4 TP4	P= 16,TN=33 ,FP=25,FN=1	65.33	6		
1 5 TP:	P=18 ,TN=29 ,FP=23,FN=5	62.67	4		
	P= 16,TN=32 ,FP=25,FN=2	64.00	6		
	P= 15.TN=02 ,FP=26.FN=2	62,67	4		
	P=21 ,TN=34 ,FP=20,FN=0	73.33	16		
	P=19 ,TN=33,FP=22,FN=1	69.33	6		
	P=18 .TN=29 .FP=23.FN=5	62.67	6		
27		เฉลีย=66.67%	เวลาที่ใช่ =พ่อย		
28					
parameter: alp	pha=0.01,hidden_layer_sizes=(500,), learning_rate='adaptive', max_iter=1000	Train_Set=225	Test_Set=75	Data=300	ให้ Laptop = Positive , Tablet = Negative
sau Cor	onfusion_Matrix	Accuracy			
	P=25 ,TN=30 ,FP=16,FN=4	73.33	4		
12 2 TP+	P=23 ,TN=32 ,FP=18,FN=2	73.33	16		
3 TP:	P=21 ,TN=32 ,FP=20,FN=2	70.67	6		
34 4 TP	P=25 ,TN=31 ,FP=16,FN=3	74.67	6		
5 TP:	P=29 ,TN=31 ,FP=,12FN=3	80.00	6		
6 TP4	P=26 ,TN=30 ,FP=15,FN=4	74.67	6		
7 TP:	P=27,TN=30,FP=14,FN=4	76.00	4		
38 8 TP-	P=27,TN=29,FP=14,FN=5	74.67	16		
She	heet1 +				

ตัวอย่างรูปข้อมูลใน Excel หลังจากการ
ทคลองพบว่า ถ้าเราเพิ่มจำนวนโหนคเยอะๆ ค่า
Alpha น้อยๆ และจำนวนรอบเยอะๆ พบว่าทำให้
เกิดค่า Accuracy ได้สูงที่สุด คือ 73.87% แต่ก็จะ
แลกมากับเวลาที่มากขึ้น แต่ตัวอื่นๆมีค่า Accuracy
ที่ใกล้เคียงกัน และใช้เวลาเท่าๆกัน

- สรุปผลการทคลอง

จากการทดลองทั้งหมดโมเคลที่ผมจะเลือกใช้ คือ Multilayer Perceptron Classifier เนื่องจากตัว KNN ไม่เหมาะกับข้อมูลที่เป็น nominal เลยตัดทิ้งเป็นอันดับแรก ส่วน

Decision Tree ถ้าเราไม่กำหนดความลึกของค้นไม้พบว่า ตัว Accuracy ของค้นไม้ตัดสินใจ ยังได้น้อยกว่าตัว MLPClassifier แต่ข้อเสียของMLPClassifier ก็ยังมีอยู่คือ เวลาที่ใช้ใน การคำนวณจะนานกว่าตัว Model อื่นๆ แต่จาก Data Set นี้ เวลาที่ใช้ในการคำนวณมันห่าง กันไม่มาก(ประมาณ2-3วินาที) แต่ก็แลกมากับค่า Accuracy ที่สูง แต่ต้องเลือก จำนวนโหนด ,จำนวนรอบ,ค่า Alpha ให้ดีๆ ข้อเสียอีกอย่างของตัวMLPClassifier ก็คือมันไม่มีสามารถให้ เหตุผลได้ว่าทำถึงเลือกคำตอบนี้ด้วย แต่มันเป็นข้อเสียที่ไม่ได้ส่งผลเสียมากมายอะไรเพราะ เราไม่ได้อยากรู้ถึงเหตุผลในการเลือก เพราะฉนั้นตอนนี้เราสนแค่ว่า Model ใดเหมาะสม กับประเภทข้อมูลของเรา และ ค่า Accuracy ที่ออกมา แล้ว MLPClassifier ตอบโจทย์ สำหรับผลการทดลองนี้มากที่สุด

Link Google Drive:

https://drive.google.com/drive/folders/1pFsuvVXShJHVbaDnSW9lGYgoHRTGAhI?usp=sharing