- Dataset ที่เลือกใช้

Dataset ที่เลือกใช้เป็นชุดข้อมูลที่เกี่ยวกับการตัดสินใจเลือกซื้ออุปกรณ์อิเล็กทรอนิกส์ระหว่างLaptop(Notebook) กับ Tablet(iPad , Galaxy Tabs) ขอบเขตคือเฉพาะนักศึกษาในมหาลัยลาดกระบัง เท่านั้น มี Attribute ทั้งหมด 9 อย่าง คือ grade(ชั้นการศึกษา),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 นี้เข้าไปยังโมเดล ขั้นตอนแรกจะทำการ Encode data ให้กลายเป็นตัวเลขเพื่อให้เอาไปสามารถคำนวณในตัวโมเดลต่างๆได้ หลังจาก Encode เสดจะทำให้ตัดตัวอย่างออกให้ Dataset เกิดความ Balance กัน เป็นดังรูปต่อไปนี้

ไฟล์Datasetทั้งหมด

	A	B	C	D	E	F	G	H	I	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)	สื่อบันเทิง (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	สื่อบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	สื่อบันเทิง (สตรีมมิ่งเพลงและวิดีโอต่าง ๆ)	ทุกวัน	ชาย	Tablet (Ipad , Galaxy Tabs)
9	นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	25,001 - 30,000 บาท	การศึกษา	ดัดต่อวิดีโอ	ดัดต่อวิดีโอ	ไม่พกพาเลย	หญิง	Laptop (Notebook)
10	นักศึกษาชั้นปีที่ 3	คณะวิทยาศาสตร์	วิทยาการคอมพิวเตอร์	25,001 - 30,000 บาท	การศึกษา	เล่นเกม	เล่นเกม	3-5วัน	ชาย	Laptop (Notebook)

ตัวอย่างไฟล์Dataset ตอนแรก

	A	B	C	D	E	F	G	H	I
1	faculty	price	purpose1	purpose2	purpose3	freaquency	gender	class label	
2	2	6	5	0	0	1	1	Laptop (Notebook)	
3	2	6	6	0	0	0	1	Laptop (Notebook)	
4	2	6	6	0	0	3	1	Laptop (Notebook)	
5	2	2	5	7	7	1	1	Laptop (Notebook)	
6	2	4	0	4	4	2	2	Tablet (Ipad , Galaxy Tabs)	
7	2	2	5	0	0	0	1	Laptop (Notebook)	
8	2	1	0	4	4	2	1	Tablet (Ipad , Galaxy Tabs)	
9	2	3	0	1	1	3	2	Laptop (Notebook)	
10	2	3	0	7	7	1	1	Laptop (Notebook)	

ตัวอย่างไฟล์ Dataset หลังจากทำการ encode และ balance

หลังจากทำการ Encode และ Balance ตัว Dataset แล้วจะเหลือจำนวน Dataset ทั้งหมด 300 ตัว

- อธิบายผลการทดลอง

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

Direct Classification -เลือกใช้ K-Nearest Neighbors Algorithm

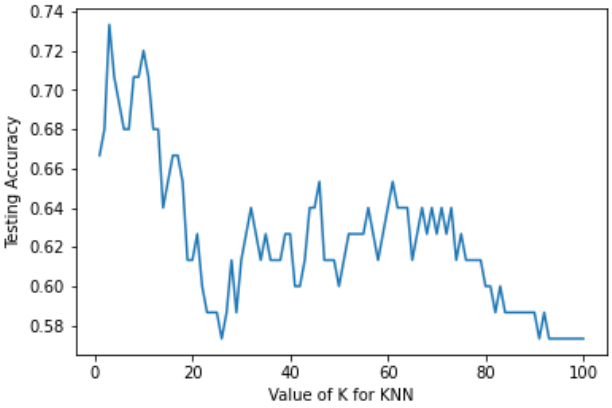
Parameter ที่ใช้ มี 2 อย่าง คือ weights ใช้ distance และ n_neighbors จำนวนเพื่อนบ้าน ทดลองโดยรันสุ่มค่า k ไปเรื่อยๆออกมาเป็นกราฟเพื่อหาดูว่า ค่า k ใดที่ให้ค่า Accuracy สูงสุด กับ Dataset มากที่สุด ทำช่วงไว้ที่ 1-100 และดู Confusion Matrix

Traditional classification -เลือกใช้ Decision Tree Parameter ที่ใช้มี 2อย่าง คือ random_state=0 คือจะไม่มีการสุ่มต้นไม้มั่วๆ max_depth กำหนดความลึกของต้นไม้ โดยจะ กำหนดไว้ที่ 2 กับไม่กำหนดขั้นต่ำเพื่อดูผลลัพธ์ของ ต้นไม้ทดลองโดยการรันและดู Confusion Matrix และ ดูค่า Accuracy

Deep Learning -เลือกใช้ Multilayer Perceptron Classifier Parameter ที่ใช้มี 4อย่าง คือ alpha คือ ค่า regularization โดยใช้ L2 penalty ข้อดีของมันคือทำให้เราควบคุม 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 วิธีการทดลองปรับparameter ไปเรื่อยๆ โดยการปรับ parameter แต่ละครั้ง จะทำการวัด Confusion Matrix และ Accuracy ทั้งหมด 10 รอบ เนื่องจากตัว MLPClassifier เมื่อเริ่มต้น model ค่า weight จะถูกสุ่มเลยทำให้การวัดค่า Accuracy ในแต่ละครั้งไม่เท่ากันเลยตัดสินใจรัน 10 ครั้ง เพื่อหาค่าเฉลี่ย ในการทดลองนี้ทำการดูเวลาที่ใช้ในการคำนวณคร่าวๆด้วย

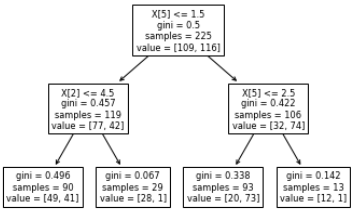
- ผลการทดลอง

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

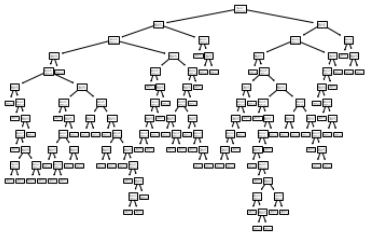


จากกราฟนี้พบว่า ค่า k ที่ทำให้เกิดค่า Accuracy สูงที่สุด คือ k = 3 ได้ 73.33% แต่การที่จะหาค่า K ที่เหมาะสมนั้น ต้องดูจากภาพรวมของ Data Set ทั้งหมดมีสูตรการหาค่า k คือ k=sqrt(n) โดยที่ n คือ จำนวน Data Set

Decision Tree -ทำการทดลองรันโดยเปลี่ยน จำนวนความลึกของต้นไม้เป็นรูปดังนี้

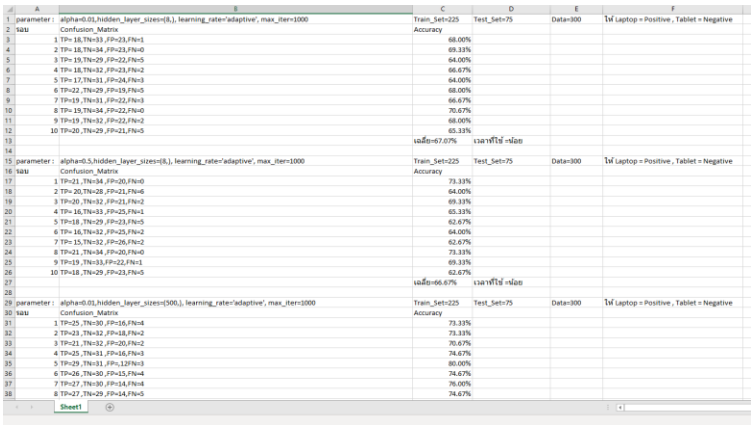


ต้นไม้ที่มีความลึกเท่ากับ 2 ค่า Accuracy = 78.67%



ต้นไม้ที่มีความลึกไม่จำกัด ค่า Accuracy = 69.33%

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



	A	B	C	D	E	F
1	parameter :	alpha=0.01,hidden_layer_sizes=(8,), learning_rate='adaptive', max_iter=1000	Train_Set=225	Test_Set=75	Data=300	ไฟ Laptop = Positive , Tablet = Negative
2	รอบ	Confusion_Matrix	Accuracy			
3	1	TP= 18,TN=33 ,FP=23,FN=1	68.00%			
4	2	TP= 18,TN=34 ,FP=23,FN=0	69.33%			
5	3	TP= 19,TN=29 ,FP=22,FN=5	64.00%			
6	4	TP= 18,TN=32 ,FP=23,FN=2	66.67%			
7	5	TP= 17,TN=31 ,FP=24,FN=3	64.00%			
8	6	TP=22 ,TN=29 ,FP=19,FN=5	68.00%			
9	7	TP=19 ,TN=31 ,FP=22,FN=3	66.67%			
10	8	TP= 19,TN=34 ,FP=22,FN=0	70.67%			
11	9	TP=19 ,TN=32 ,FP=22,FN=2	68.00%			
12	10	TP=20 ,TN=29 ,FP=23,FN=5	65.33%			
13			เฉลี่ย=67.07%	เวลาที่ใช้ -น้อย		
14						
15	parameter :	alpha=0.5,hidden_layer_sizes=(8,), learning_rate='adaptive', max_iter=1000	Train_Set=225	Test_Set=75	Data=300	ไฟ Laptop = Positive , Tablet = Negative
16	รอบ	Confusion_Matrix	Accuracy			
17	1	TP=21 ,TN=34 ,FP=20,FN=0	73.33%			
18	2	TP= 20,TN=28 ,FP=21,FN=6	64.00%			
19	3	TP=20 ,TN=32 ,FP=21,FN=2	69.33%			
20	4	TP= 16,TN=33 ,FP=25,FN=1	65.33%			
21	5	TP=18 ,TN=29 ,FP=23,FN=5	62.67%			
22	6	TP= 16,TN=32 ,FP=25,FN=2	64.00%			
23	7	TP= 15,TN=32 ,FP=26,FN=2	62.67%			
24	8	TP=21 ,TN=34 ,FP=20,FN=0	73.33%			
25	9	TP=19 ,TN=33,FP=22,FN=1	69.33%			
26	10	TP=18 ,TN=29 ,FP=23,FN=5	62.67%			
27			เฉลี่ย=66.67%	เวลาที่ใช้ -น้อย		
28						
29	parameter :	alpha=0.01,hidden_layer_sizes=(500,), learning_rate='adaptive', max_iter=1000	Train_Set=225	Test_Set=75	Data=300	ไฟ Laptop = Positive , Tablet = Negative
30	รอบ	Confusion_Matrix	Accuracy			
31	1	TP=25 ,TN=30 ,FP=16,FN=4	73.33%			
32	2	TP=23 ,TN=32 ,FP=18,FN=2	73.33%			
33	3	TP=21 ,TN=32 ,FP=20,FN=2	70.67%			
34	4	TP=25 ,TN=31 ,FP=16,FN=3	74.67%			
35	5	TP=29 ,TN=31 ,FP=,12FN=3	80.00%			
36	6	TP=26 ,TN=30 ,FP=15,FN=4	74.67%			
37	7	TP=27 ,TN=30 ,FP=14,FN=4	76.00%			
38	8	TP=27 ,TN=29 ,FP=14,FN=5	74.67%			

Sheet1

ตัวอย่างรูปข้อมูลใน Excel หลังจากการทดลองพบว่า ถ้าเราเพิ่มจำนวนโหนดเยอะๆ ค่า Alpha น้อยๆ และจำนวนรอบเยอะๆ พบว่าทำให้เกิดค่า Accuracy ได้สูงที่สุด คือ 73.87% แต่ก็จะแลกมากับเวลาที่มากขึ้น แต่ตัวอื่นๆมีค่า Accuracy ที่ใกล้เคียงกัน และใช้เวลาเท่าๆกัน

- สรุปผลการทดลอง

จากการทดลองทั้งหมดโมเดลที่ผมจะเลือกใช้ คือ Multilayer Perceptron Classifier เนื่องจากตัว KNN ไม่เหมาะกับข้อมูลที่เป็น nominal เลยตัดทิ้งเป็นอันดับแรก ส่วน Decision Tree ถ้าเราไม่กำหนดความลึกของต้นไม้พบว่า ตัว Accuracy ของต้นไม้ตัดสินใจยังได้น้อยกว่าตัว MLPClassifier แต่ข้อเสียของMLPClassifier ก็ยังมีอยู่คือ เวลาที่ใช้ในการคำนวณจะนานกว่าตัว Model อื่นๆ แต่จาก Data Set นี้ เวลาที่ใช้ในการคำนวณมันห่างกันไม่มาก(ประมาณ2-3วินาที) แต่ก็แลกมากับค่า Accuracy ที่สูง แต่ต้องเลือก จำนวนโหนด ,จำนวนรอบ,ค่า Alpha ให้ดีๆ ข้อเสียอีกอย่างของตัวMLPClassifier ก็คือมันไม่มีสามารถให้เหตุผลได้ว่าทำถึงเลือกคำตอบนี้ด้วย แต่มันเป็นข้อเสียที่ไม่ได้ส่งผลเสียมากมายอะไรเพราะเราไม่ได้อยากรู้ถึงเหตุผลในการเลือก เพราะฉนั้นตอนนี้เราสนแค่ว่า Model ใดเหมาะสมกับประเภทข้อมูลของเรา และ ค่า Accuracy ที่ออกมา แล้ว MLPClassifier ตอบโจทย์สำหรับผลการทดลองนี้มากที่สุด

Link Google Drive :

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