

Part 1: Clustering

Introduction to the dataset and its meta data.

ข้อมูลที่ใช้คือ cluster data set

att1	att2	att3	att4
1.268	0.715	-1.899	-4.136
-4.266	-3.566	-4.802	1.462
-4.316	-2.033	-5.393	2.586
-4.881	-4.101	-4.635	2.503
-3.678	-3.7	1.088	-4.221
-0.033	4.187	-3.736	2.747
-5.565	2.975	-0.759	2.181
1.676	-2.232	4.935	2.372
-4.464	-5.364	-3.049	2.397
-0.531	3.533	4.652	1.07
-0.029	1.252	0.152	0.336
-3.622	2.003	-0.31	-2.328

รูปที่1 ตัวอย่างdata set

ในการทำการ clustering ได้ทำการใช้ operation k-mean ในการทำclustering

และทำการดูว่าoperator ไหนเหมาะกับการทำ subjective และ objective มากที่สุด

Clustering algorithms used

1. Clustering with K-mean

Cluster Model

```
Cluster 0: 89 items
Cluster 1: 180 items
Cluster 2: 79 items
Cluster 3: 85 items
Cluster 4: 67 items
Total number of items: 500
```

รูปภาพแสดงถึงจำนวน **Cluster** ที่ **K-mean** ทำการจำแนกกลุ่มออกมาได้มี

ผลลัพธ์ค่อนข้างดีและง่ายต่อการเข้าใจ

นำมาทำการหารหาจุด **Centroid** เพื่อทำการ **objective** ด้วย เครื่องมือ

Objective measures of clusters found

นำมาทำการหารหาจุด **Centroid** เพื่อทำการ **objective** ด้วย เครื่องมือ **operator**
cluster distance performance

Conclusion

ผลลัพธ์จากการใช้ **Opreator** cluster distance performance ในการทำ

Objective , **Operator** นี้ใช้วิธี **davied bouldin index** ในการคำนวณ

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -9.901
Avg. within centroid distance_cluster_0: -13.193
Avg. within centroid distance_cluster_1: -11.072
Avg. within centroid distance_cluster_2: -5.047
Avg. within centroid distance_cluster_3: -14.363
Avg. within centroid distance_cluster_4: -2.446
Davies Bouldin: -0.953
```

ค่าอยู่ที่ **-0.953** ในการใช้วิธี **davied bouldin index** ในการคำนวณยิ่งค่าผลลัพธ์

ออกมาน้อยแสดงว่า **Cluster** นั้นยังมีประสิทธิภาพ

Part 2: ANN

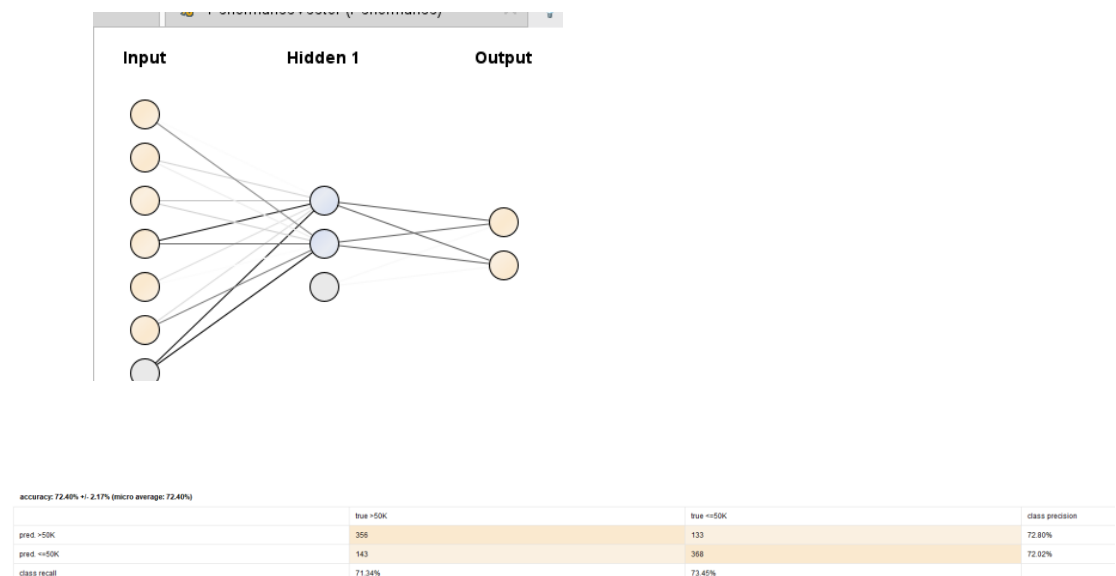
Introduction to the dataset and its meta data.

Data set ที่นำมาใช้ประมวลผล คือเงินเดือนของแต่ละคน โดยจะนำข้อมูลที่มาทำการทำนายว่าจะมีผลต่อเงินเดือนบ้าง

อธิบาย attribute แต่ละ attribute 1.age อายุ มีผลต่อรายได้ไหม 2.weight 3.capital รายรับ 4.capital loss รายจ่าย 5.education การศึกษา 6.hour per week ชั่วโมงการทำงาน 7.label เงินเดือน

Row No.	age	fnlwgt	education_n...	capital_gain	capital_loss	hours_per_...	label
1	0.698	0.024	-0.202	-0.200	-0.268	0.693	>50K
2	1.246	1.367	-0.574	-0.200	3.546	0.693	>50K
3	-0.008	0.172	0.541	-0.200	3.728	1.538	>50K
4	-1.262	0.843	-0.202	0.423	-0.268	0.524	>50K
5	-0.478	-0.431	1.284	0.423	-0.268	-0.574	>50K
6	-0.400	-0.902	-0.945	-0.200	-0.268	0.693	>50K
7	-0.165	-0.915	0.913	1.083	-0.268	1.538	>50K
8	-0.243	-1.555	0.169	-0.200	3.728	-0.151	>50K
9	0.670	1.502	0.012	0.200	0.268	0.151	<50K

Results of training using default parameters.



แสดงผลของการ run model เมื่อใช้ค่า default parameter และอธิบายเหตุผลว่าเพราะสาเหตุใดจึงได้ accuracy ดังกล่าว

ผลการทำนายโดยใช้ค่า default parameter ได้ค่าอยู่ที่ 72.40% โดยการที่ค่าดังกล่าวได้ผลเช่นนั้นเกิดจากการที่ parameter ยังไม่ได้ปรับให้ดีพอในการทำ classification


2.2 ทำการเปลี่ยนค่า parameter เพื่อดูความแตกต่างของผล

accuracy: 74.30% +/- 4.42% (micro average: 74.30%)		
	true >50K	true <=50K
pred. >50K	325	83
pred. <=50K	174	418
class recall	65.13%	83.43%

ทำการเปลี่ยนค่า parameter ในการทำ ANN มีผลดีขึ้น 2.30% โดยทำการปรับ parameter ดังนี้

1. ทำการปรับ hidden layer เพื่อช่วยให้โครงสร้างนี้สามารถเรียนรู้และความซับซ้อนของข้อมูลได้มากขึ้น, ซึ่งช่วยในการแก้ปัญหาที่ซับซ้อนและการจำแนกประเภทข้อมูลที่มีลักษณะที่ซับซ้อน
2. ทำการปรับ learning rate ไปที่ 0.7 ให้เรียนรู้ค่อนข้างเร็ว
3. ทำการปรับ momentum ที่ ส่งผลต่อการประมวลผล ไป ที่ 0.08 ส่งผลต่อการประมวลผลค่อนข้างต่ำ
4. กำหนดให้มีการสลับ shuffle data set ให้ได้เรียนรู้ข้อมูลหลายรูปแบบ
5. ทำการ normalize dataset ให้มีรูปแบบที่งานดีการประมวลผล data set
6. ทำการเพิ่ม cycle ในการ train data set ให้มีการทำที่มากขึ้น

Parameters ✕

 **Neural Net**

hidden layers

training cycles

800

learning rate

0.7

momentum

0.08

☐ decay

☒ shuffle

☒ normalize

Results when training using modified parameter settings.

อธิบายถึงผลกระทบต่อค่า parameter ของ algorithm

ผลจากการปรับ parameter ส่งผลต่อ accuracy ในทางที่ดีขึ้น จาก 72.40% เป็น 74.30 % โดยผลรับก่อนการปรับเป็นผลที่อยู่ในเกณฑ์ ที่รับ ได้อยู่แล้ว

Patterns in the data

อธิบายถึงผลและ pattern ใน dataset


Pattern ของ data set นี้เป็นการวิเคราะห์ จากข้อมูลที่ได้มาว่ามีสิ่งไหนส่งผลต่อ label บ้าง โดยจะดูจาก capital gain และ capital loss education และมาทำการคำนวณหา ผลลัพธ์ที่ถูกต้อง

Part3: Association Rules

Introduction to the dataset and its meta data.

Data set บอกถึงตัวแปรต่างๆของแต่ละคนและเงินเดือน ทำการวิเคราะห์โดยใช้ FP growth operator โดยให้ค่า support = 0.2 และให้สร้าง association rule ว่าอะไรมีผลต่อเงินเดือน

Frequent Itemset Discussion

 FP-Growth	
input format	items in separate columns
<input checked="" type="checkbox"/> trim item names	
min requirement	support
min support	0.2
min items per itemset	1
max items per itemset	0
max number of itemsets	1000000

ทำการใช้ FP growth โดยให้ค่า support = 0.2 และแสดงผลลัพธ์ดูว่าอะไรทำให้คนมีรายได้มากกว่า 50k

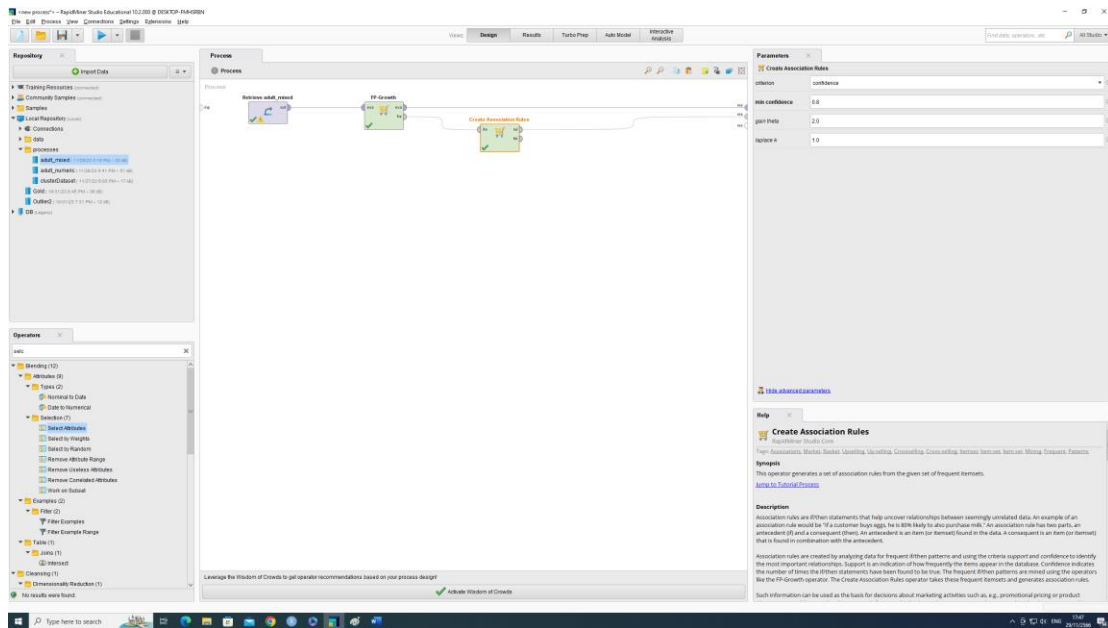
Size	Support ↑	Item 1	Item 2	Item 3	Item 4	Item 5
2	0.200	Private	9.0			
2	0.200	Private	HS-grad			
3	0.200	0.0	Private	9.0		
3	0.200	0.0	Private	HS-grad		
3	0.200	Private	9.0	HS-grad		
4	0.200	0.0	Private	9.0	HS-grad	
4	0.200	United-States	White	Husband	40.0	
5	0.200	0.0	United-States	White	Husband	40.0
5	0.200	United-States	White	Male	Husband	40.0
5	0.200	United-States	White	Married-cv-spouse	Husband	40.0
6	0.200	0.0	United-States	White	Male	Husband

Row No.	label	age	workclass	hthwt	education ↓	education_n...	marital_stat...	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_...	native_coun...
1	>50K	49.0	Self-emp-inc	191681.0	Some-college	10.0	Married-civ-s...	Exec-manag...	Husband	White	Male	0.0	0.0	50.0	United-States
4	>50K	24.0	Private	279472.0	Some-college	10.0	Married-civ-s...	Machine-op-L...	Wife	White	Female	7298.0	0.0	48.0	United-States
17	>50K	44.0	Self-emp-inc	320984.0	Some-college	10.0	Married-civ-s...	Sales	Husband	White	Male	5178.0	0.0	60.0	United-States
31	>50K	31.0	Private	157887.0	Some-college	10.0	Married-civ-s...	Exec-manag...	Husband	White	Male	0.0	0.0	65.0	United-States
42	>50K	50.0	Private	92079.0	Some-college	10.0	Married-civ-s...	Tech-support	Husband	White	Male	0.0	0.0	45.0	United-States
53	>50K	53.0	Private	211654.0	Some-college	10.0	Married-civ-s...	Craft-repair	Husband	White	Male	0.0	0.0	50.0	?
63	>50K	37.0	Private	205339.0	Some-college	10.0	Never-married	Tech-support	Not-in-family	White	Male	0.0	0.0	40.0	United-States
64	>50K	43.0	Local-gov	105862.0	Some-college	10.0	Married-civ-s...	Adm-clerical	Wife	White	Female	0.0	1902.0	40.0	United-States
65	>50K	43.0	Local-gov	96102.0	Some-college	10.0	Married-civ-s...	Protective-serv	Husband	White	Male	0.0	1887.0	40.0	United-States
68	>50K	49.0	Self-emp-inc	362954.0	Some-college	10.0	Married-civ-s...	Exec-manag...	Husband	White	Male	15024.0	0.0	50.0	United-States
69	>50K	54.0	Private	22743.0	Some-college	10.0	Married-civ-s...	Transport-mo...	Husband	White	Male	15024.0	0.0	60.0	United-States
72	>50K	27.0	Private	35032.0	Some-college	10.0	Married-civ-s...	Handlers-cle...	Husband	White	Male	0.0	0.0	40.0	United-States
76	>50K	48.0	Local-gov	31264.0	Some-college	10.0	Married-civ-s...	Exec-manag...	Wife	White	Female	5178.0	0.0	40.0	United-States
83	>50K	65.0	Self-emp-inc	208452.0	Some-college	10.0	Married-civ-s...	Sales	Husband	White	Male	0.0	0.0	35.0	United-States
89	>50K	31.0	Self-emp-inc	236415.0	Some-college	10.0	Married-civ-s...	Adm-clerical	Wife	White	Female	0.0	0.0	20.0	United-States
90	>50K	55.0	Private	157079.0	Some-college	10.0	Married-civ-s...	Protective-serv	Husband	Black	Male	0.0	0.0	40.0	?

จากรูปภาพแสดงให้เห็นถึงสาเหตุหลัก 4 อย่างด้วยกันที่ส่งผลต่อเงินเดือน นั่นคือ 1.education 2.marital satatus 3.race 4.education num

Rules Discussion

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	LIFT ↓	Conviction
445	Some-college	White, 10.0	0.201	0.893	0.980	-0.249	0.155	4.358	7.453
447	Some-college	0.0, White, 10.0	0.201	0.893	0.980	-0.249	0.155	4.358	7.453
448	0.0, Some-college	White, 10.0	0.201	0.893	0.980	-0.249	0.155	4.358	7.453
940	White, 10.0	Some-college	0.201	0.980	0.997	-0.209	0.155	4.358	39.719
941	White, 10.0	0.0, Some-college	0.201	0.980	0.997	-0.209	0.155	4.358	39.719
942	0.0, White, 10.0	Some-college	0.201	0.980	0.997	-0.209	0.155	4.358	39.719
314	10.0	White, Some-college	0.201	0.874	0.976	-0.259	0.155	4.348	6.337
315	10.0	0.0, White, Some-college	0.201	0.874	0.976	-0.259	0.155	4.348	6.337
316	0.0, 10.0	White, Some-college	0.201	0.874	0.976	-0.259	0.155	4.348	6.337
643	10.0	United-States, Some-college	0.210	0.913	0.984	-0.250	0.162	4.348	9.085
644	10.0	0.0, United-States, Some-college	0.210	0.913	0.984	-0.250	0.162	4.348	9.085
645	0.0, 10.0	United-States, Some-college	0.210	0.913	0.984	-0.250	0.162	4.348	9.085



จากการทำ **rules** และตั้งค่า **parameter** ดังนี้ ให้ผลลัพธ์ว่า ส่วนที่
 ส่งผลต่อเงินเดือนทั้งมากกว่าและน้อยกว่ามากจาก **education** และ
capital gain และ **capital loss** กับ **education weight** เป็น
Rules ที่น่าสนใจ และมีค่า **lift** มากกว่า 1

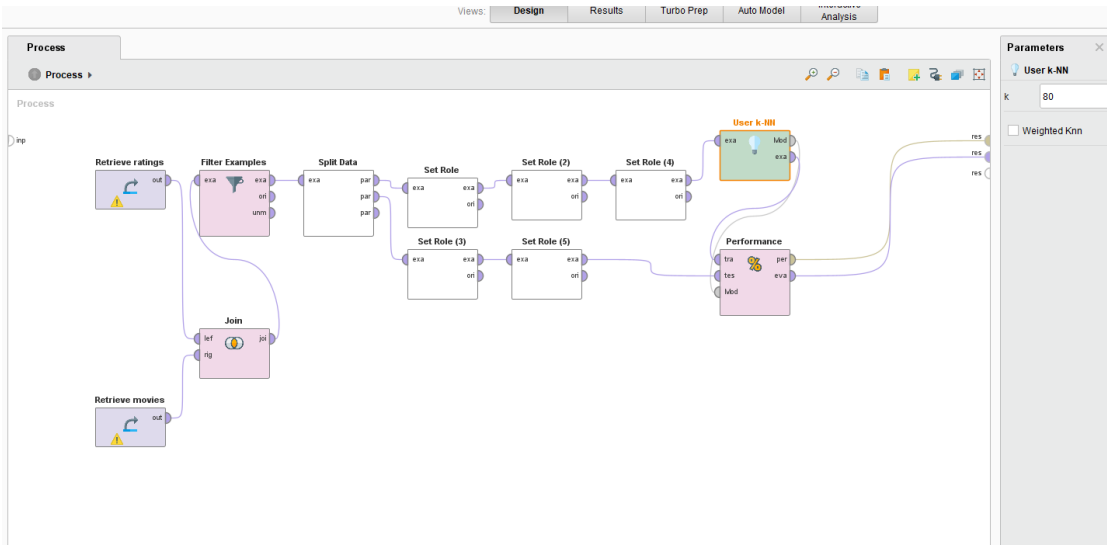
Part 4: Recommendation Engine

Introduction to the dataset and its meta data.

Data set ที่ใช้เป็น rating ของหนังที่ user ให้คะแนน และ movie ชื่อid ของหนังแต่ละเรื่อง โดยจะทำการเอาสอง Data set นี้มาทำการ join กัน และทำการสร้าง

Recommendation

Recommendation algorithms used



Row No.	userId	movieId	rating	title
81052	513	32	4	Twelve Monkeys (...)
81053	513	110	4	Braveheart (1995)
81054	513	150	4.500	Apollo 13 (1995)
81055	513	235	4.500	Ed Wood (1994)
81056	513	260	5	Star Wars: Episo...
81057	513	266	3.500	Legends of the F...
81058	513	296	4.500	Pulp Fiction (1994)
81059	513	318	5	Shawshank Red...
81060	513	457	4.500	Fugitive, The (19...
81061	513	593	4	Silence of the La...
81062	513	750	5	Dr. Strangelove a

Algorithm ที่ใช้ประกอบไปด้วย User knn และ performance item **Recommendation** โดย **user Knn** ใช้ในการหา **Recommendation** และใช้ performance ในการหาค่า **prec@k**

Results of training using default parameters.

Open in

 Turbo Prep


 Auto Model

Row No.	AUC	prec@5	prec@10	prec@15	NDCG	MAP
1	0.925	0.303	0.250	0.224	0.511	0.169

ผลลัพธ์ที่ได้ จาก **prec@10** อยู่ที่ **0.250**

Results when training using modified parameter settings.

Row No.	AUC	prec@5	prec@10	prec@15	NDCG	MAP
1	0.910	0.308	0.258	0.230	0.518	0.176

 **User k-NN**

k

☐ **Weighted Knn**

หลังจากได้ทำการลองปรับ **parameter** ดูแล้วพบว่าผลลัพธ์ได้เพิ่มขึ้นนิดนึง

Conclusion

ผลสรุป ได้ทำการ ใช้ **user Knn** และ **performance** ในการทำ **Recommendation** ของ **data set** และ **movie** ผลลัพธ์ที่ได้จากการใช้ **default parameter** ของ **prec@10** คือ **0.250** ทำการปรับค่า **parameter** ของ **user knn** ดูแล้ว ผลลัพธ์อยู่ที่ **0.258**