



Assignment

เรื่อง การใช้งาน RapidMiner Studio

จัดทำโดย

นายวงศ์รัชต์ มณีพันธ์ 5735512049

นายณัฐวัตร ทองอร่าม 5735512125

เสนอ

อาจารย์ อัมรินทร์ ดีมะการ

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา 242-425 Data Mining

มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตภูเก็ต

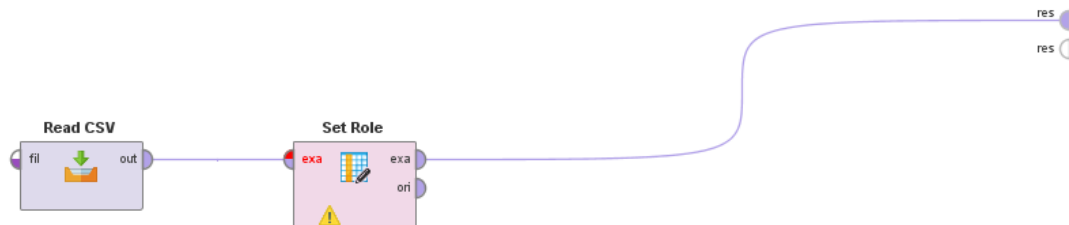
ภาคเรียนที่ 1 ปีการศึกษา 2560

1. Data Exploration

Operator ที่ใช้

Read CSV อ่านไฟล์ CSV จากภายนอก

อ่านไฟล์ Data01_customer-churn.csv



Set Role เชื่อมหน้าที่ของ Attribute

เช็ต Customer_id เป็น id

เช็ต Churn? เป็น Label

Parameters ✕

Set Role

attribute name Customer_id ▼

target role id ▼

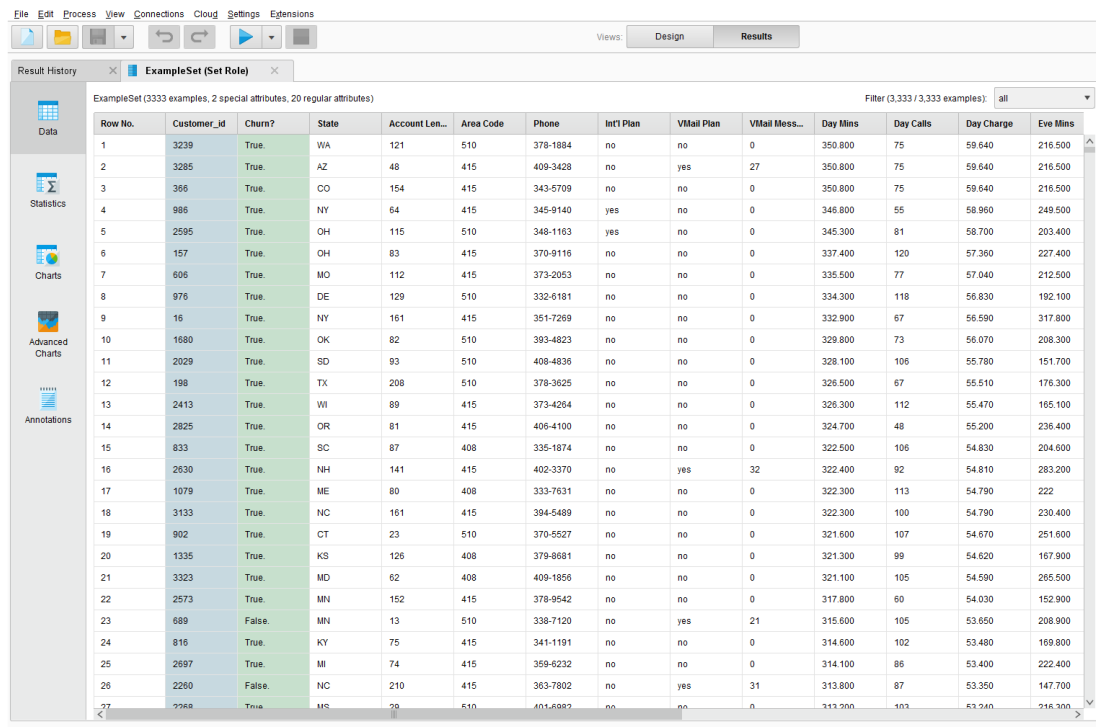
set additional roles Edit List (1)...

Edit Parameter List: set additional roles ✕

Edit Parameter List: **set additional roles**
This parameter defines additional attribute role combinations.

attribute name	target role
Churn? ▼	label ▼

Result



Row No.	Customer_id	Churn?	State	Account Len...	Area Code	Phone	Int'l Plan	VMail Plan	VMail Mess...	Day Mins	Day Calls	Day Charge	Eve Mins
1	3239	True	WA	121	510	378-1884	no	no	0	350.800	75	59.640	216.500
2	3285	True	AZ	48	415	409-3428	no	yes	27	350.800	75	59.640	216.500
3	366	True	CO	154	415	343-6709	no	no	0	350.800	75	59.640	216.500
4	986	True	NY	64	415	345-9140	yes	no	0	346.800	55	58.960	249.500
5	2595	True	OH	115	510	348-1163	yes	no	0	345.300	81	58.700	203.400
6	157	True	OH	83	415	370-9116	no	no	0	337.400	120	57.360	227.400
7	606	True	MO	112	415	373-2053	no	no	0	335.500	77	57.040	212.500
8	976	True	DE	129	510	332-6181	no	no	0	334.300	118	56.830	192.100
9	16	True	NY	161	415	351-7269	no	no	0	332.900	67	56.590	317.800
10	1680	True	OK	82	510	393-4823	no	no	0	329.800	73	56.070	208.300
11	2029	True	SD	93	510	408-4836	no	no	0	328.100	106	55.780	151.700
12	198	True	TX	208	510	378-3625	no	no	0	326.500	67	55.510	176.300
13	2413	True	WI	89	415	373-4264	no	no	0	326.300	112	55.470	165.100
14	2825	True	OR	81	415	406-4100	no	no	0	324.700	48	55.200	236.400
15	833	True	SC	87	408	335-1874	no	no	0	322.500	106	54.830	204.600
16	2630	True	NH	141	415	402-3370	no	yes	32	322.400	92	54.610	283.200
17	1079	True	ME	80	408	333-7631	no	no	0	322.300	113	54.790	222
18	3133	True	NC	161	415	394-6489	no	no	0	322.300	100	54.790	230.400
19	902	True	CT	23	510	370-5527	no	no	0	321.600	107	54.670	251.600
20	1335	True	KS	126	408	379-8681	no	no	0	321.300	99	54.620	167.900
21	3323	True	MD	62	408	409-1856	no	no	0	321.100	105	54.590	265.500
22	2573	True	MN	152	415	378-9542	no	no	0	317.800	60	54.030	152.900
23	689	False	MN	13	510	338-7120	no	yes	21	315.600	105	53.650	208.900
24	816	True	KY	75	415	341-1191	no	no	0	314.600	102	53.480	169.800
25	2697	True	MI	74	415	359-6232	no	no	0	314.100	86	53.400	222.400
26	2260	False	NC	210	415	363-7802	no	yes	31	313.800	87	53.350	147.700
27	2268	True	MS	28	510	401-6982	no	no	0	313.300	103	53.240	216.300

จะเห็นได้ว่า Customer_id จะเป็นสีฟ้า ซึ่งเป็น ID และ Churn? จะเป็นสีเขียว ซึ่งเป็น label

2. Data Bending

Operator ที่ใช้

Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

ดึงข้อมูล Retrieve Data02_customer-info และ Data03_customer-churn

Join ใช้สำหรับเชื่อมโยงข้อมูลจาก 2 ตารางเข้าด้วยกัน มีลักษณะเหมือนคำสั่ง join ใน SQL

เชื่อมโยงข้อมูล Retrieve Data02_customer-info และ Data03_customer-churn

Rename ใช้สำหรับเปลี่ยนชื่อ Attributes ต่าง ๆ

เปลี่ยนชื่อ Attributes Place_ofBirth เป็น Hometown

Generate Attributes ใช้สำหรับสร้าง Attributes ใหม่ขึ้นมา

เปลี่ยน Attributes Age เป็นไปตามโค้ด `ate_get(date_now(),DATE_UNIT_YEAR) - date_get(DOB,DATE_UNIT_YEAR)`

Select Attributes เลือก Attributes

เลือก Attributes Area Code, DOB, Hometown, Phone, State, email, first_name, last_name

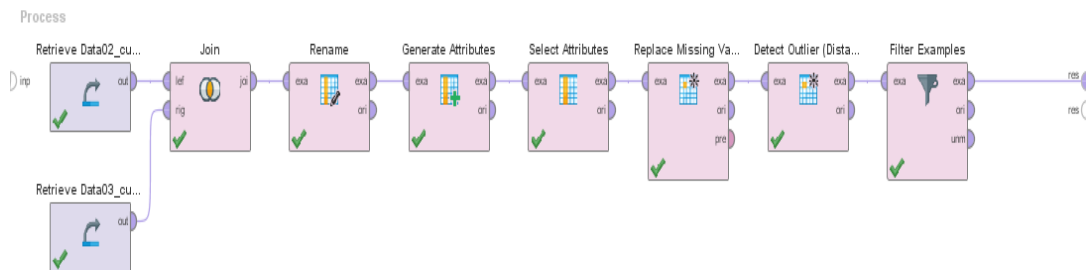
Replace Missing Values เป็นการแทนที่ค่าที่หายไป

แทนที่ค่า Attribute Age ด้วยค่าคงที่ 30 และ แทนที่ Attributes gender ด้วยค่าเฉลี่ย

Detect Outlier (Distances) ตรวจจับ Outlier

Filter Examples ใช้สำหรับเลือกข้อมูลที่น่าสนใจออกมาแสดงผล

เลือกข้อมูลที่ Outlier เท่ากับ False



Result

Row No.	id	outlier	gender	Account Len...	Int'l Plan	VMail Plan	VMail Mess...	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge
1	1	false	Female	128	no	yes	25	265.100	110	45.070	197.400	99	16.780
2	3	false	Woman	137	no	no	0	243.400	114	41.380	121.200	110	10.300
3	5	false	Female	75	yes	no	0	166.700	113	28.340	148.300	122	12.610
4	9	false	Female	117	no	no	0	184.500	97	31.370	351.600	80	29.890
5	12	false	Female	74	no	no	0	187.700	127	31.910	163.400	148	13.890
6	17	false	Female	85	no	yes	27	196.400	139	33.390	280.900	90	23.880
7	18	false	Female	93	no	no	0	190.700	114	32.420	218.200	111	18.550
8	24	false	Female	111	no	no	0	110.400	103	18.770	137.300	102	11.670
9	25	false	Female	132	no	no	0	81.100	96	13.790	245.200	72	20.840
10	27	false	Female	57	no	yes	39	213	115	36.210	191.100	112	16.240
11	29	false	Female	20	no	no	0	190	109	32.300	258.200	84	21.950
12	31	false	Female	142	no	no	0	84.800	95	14.420	136.700	63	11.620
13	32	false	Female	75	no	no	0	226.100	105	38.440	201.500	107	17.130
14	33	false	Female	172	no	no	0	212	121	36.040	31.200	115	2.650
15	35	false	Female	57	no	yes	25	176.800	94	30.060	195	75	16.580
16	37	false	Female	36	no	yes	30	146.300	128	24.870	162.500	80	13.810
17	39	false	Female	136	yes	yes	33	203.900	106	34.660	187.600	99	15.950
18	42	false	Female	135	yes	yes	41	173.100	85	29.430	203.900	107	17.330
19	46	false	Female	59	no	yes	28	120.900	97	20.550	213	92	18.110
20	51	false	Woman	52	no	no	0	191.900	108	32.620	269.800	96	22.930
21	53	false	Woman	10	no	no	0	186.100	112	31.640	190.200	66	16.170
22	54	false	Woman	96	no	no	0	160.200	117	27.230	267.500	87	22.740
23	56	false	Woman	81	no	no	0	175.500	67	29.840	249.300	85	21.190
24	57	false	Woman	141	no	no	0	126.900	98	21.570	180	62	15.300
25	58	false	Woman	121	no	yes	30	198.400	129	33.730	75.300	77	6.400
26	60	false	Woman	125	no	no	0	229.300	103	38.980	177.400	126	15.080
27	61	false	Woman	174	no	no	0	182.100	97	33.660	166.900	94	14.440

เชื่อมโยงข้อมูล Retrieve Data02_customer-info และ Data03_customer-churn เข้าด้วยกัน

โดยใช้ Primary key เป็น ID

3. Decision Tree

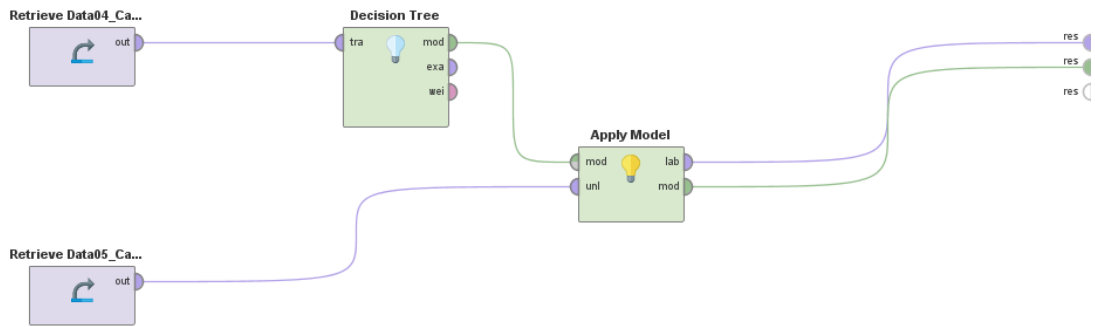
Operator ที่ใช้

Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

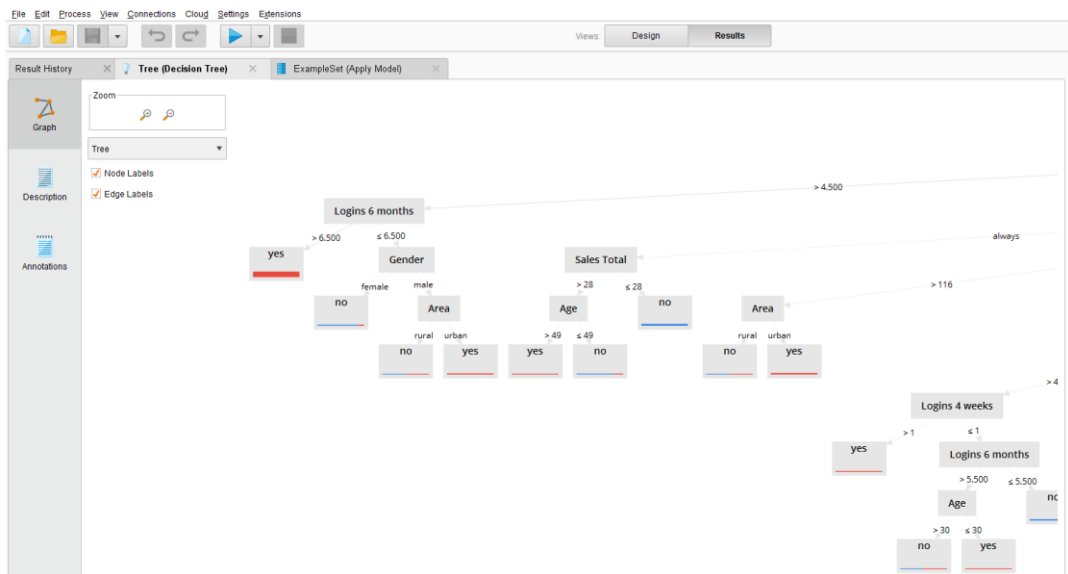
ดึงข้อมูล Data04_Campaign และ Data05_Campaign-new

Decision Tree เป็นอัลกอริทึมในการจำแนกข้อมูล

Apply Model เป็นการนำ training set และ testing set เพื่อทำนายผลลัพธ์



Result



แผนผังต้นไม้ โดยใช้แบบ information_gain ในการทำนายผลลัพธ์

4. Decision Tree, Naive Bayes and K-NN

Operator ที่ใช้

Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

ดึงข้อมูล Data04_Campaign และ Data05_Campaign-new

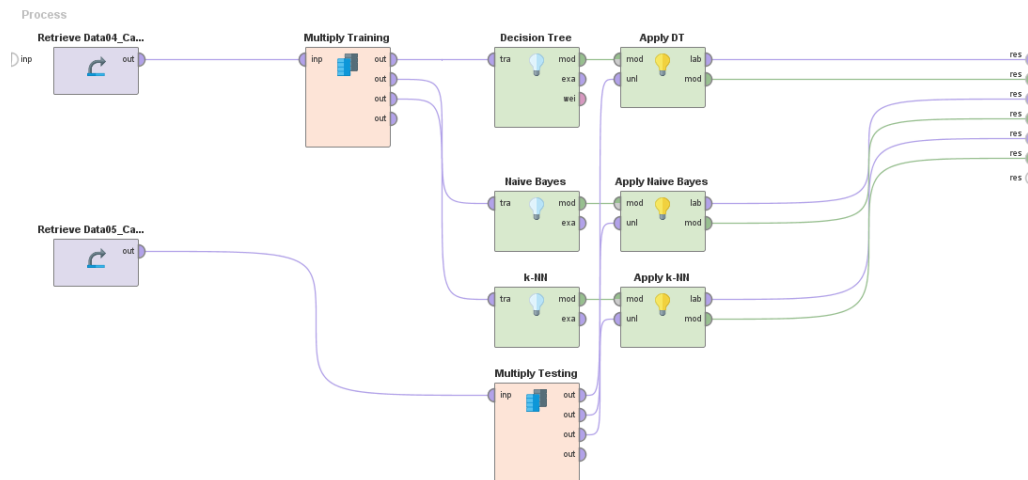
Multiply คือการนำ input เดียวให้สามารถนำไปใช้ได้หลาย ๆ อัน

Decision Tree เป็นอัลกอริทึมในการจำแนกข้อมูล

Naive Bayes เป็นอัลกอริทึมในการจำแนกข้อมูล

k-NN เป็นอัลกอริทึมในการจำแนกข้อมูล

Apply Model เป็นการนำ training set และ testing set เพื่อทำนายผลลัพธ์



Operator ที่ตรงกับที่เรียน

Decision Tree เป็นอัลกอริทึมในการจำแนกข้อมูล

Naïve Bayes เป็นอัลกอริทึมในการจำแนกข้อมูล

k-NN เป็นอัลกอริทึมในการจำแนกข้อมูล

Result

Decision Tree

ExampleSet (1000 examples, 3 special attributes, 11 regular attributes)													
Row No.	prediction(response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 wee...	
1	no	1	0	CARVER	54	female	rural	free	always	0	7	0	
2	no	1	0	GONZALES	88	male	urban	free	never	3	3	0	
3	no	1	0	MICHEL	31	female	urban	free	always	0	7	0	
4	no	0.996	0.004	TATE	54	female	rural	free	never	0	0	0	
5	no	0.996	0.004	CARR	25	male	urban	free	never	0	0	0	
6	no	0.996	0.004	ARNOLD	68	female	urban	free	never	0	0	0	
7	no	1	0	PICKETT	27	female	urban	free	yes	0	0	0	
8	yes	0	1	MCCLELLAN	34	male	urban	free	yes	0	0	0	
9	no	0.875	0.125	CONLEY	29	female	urban	free	never	5	5	0	
10	no	0.996	0.004	BUSH	50	female	rural	free	never	0	0	0	
11	no	0.996	0.004	MCCRAY	64	male	rural	free	never	0	0	0	
12	yes	0.011	0.989	O'BRIEN	61	female	urban	free	never	9	12	0	
13	no	1	0	OLSEN	19	male	rural	free	yes	0	0	0	
14	yes	0.011	0.989	HOLDER	58	female	urban	free	yes	8	8	0	
15	yes	0	1	LYNCH	55	male	urban	free	yes	4	4	0	
16	no	0.996	0.004	SHEPHERD	30	female	rural	free	never	2	2	0	
17	no	0.792	0.208	BARO	59	female	urban	free	yes	0	7	0	
18	yes	0	1	CHLDWELL	66	male	urban	premium	yes	0	0	0	
19	yes	0	1	MCMAHON	37	female	urban	premium	never	0	2	0	
20	no	1	0	BRUCE	57	male	rural	free	yes	0	7	0	
21	yes	0	1	SIMON	42	male	urban	free	never	5	5	0	
22	no	1	0	YANG	20	male	urban	free	never	0	5	0	
23	no	0.792	0.208	TERRY	21	female	urban	free	yes	2	2	0	
24	yes	0.011	0.989	MILES	32	male	rural	free	yes	14	14	0	
25	no	0.996	0.004	AYERS	73	male	rural	free	never	0	0	0	
26	yes	0.014	0.986	RIVERS	38	male	rural	premium	never	0	0	0	

Naïve Bayes

Row No.	prediction(response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 we...
1	no	1.000	0.000	CARRIER	54	female	rural	free	always	0	7	0
2	no	1.000	0.000	GONZALES	88	male	urban	free	never	3	3	0
3	no	1.000	0.000	MCNEIL	31	female	urban	free	always	0	7	0
4	no	1.000	0.000	TATE	54	female	rural	free	never	0	0	0
5	no	1.000	0.000	CARR	25	male	urban	free	never	0	0	0
6	no	1.000	0.000	ARNOLD	68	female	urban	free	never	0	0	0
7	no	1.000	0.000	PICKETT	27	female	urban	free	yes	0	0	0
8	no	1.000	0.000	MCCLAIN	34	male	urban	free	yes	0	0	0
9	no	0.983	0.017	CONLEY	29	female	urban	free	never	5	5	0
10	no	1.000	0.000	BUSH	50	female	rural	free	never	0	0	0
11	no	1.000	0.000	MCCRAY	64	male	rural	free	never	0	0	0
12	yes	0	1	O'BRIEN	61	female	urban	free	never	9	12	0
13	no	1.000	0.000	OLSEN	19	male	rural	free	yes	0	0	0
14	yes	0.000	1.000	HOLDER	58	female	urban	free	yes	8	8	0
15	no	0.508	0.492	LYNCH	55	male	urban	free	yes	4	4	0
16	no	1.000	0.000	SHEPHERD	30	female	rural	free	never	2	2	0
17	no	1.000	0.000	BARO	59	female	urban	free	yes	0	7	0
18	no	1.000	0.000	CALDWELL	66	male	urban	free	yes	0	0	0
19	no	1.000	0.000	MCMAHON	37	female	urban	premium	never	0	2	0
20	no	0.998	0.002	BRUCE	57	male	rural	free	yes	0	7	0
21	yes	0.046	0.954	SIMON	42	male	urban	free	never	5	5	0
22	no	1.000	0.000	YANG	20	male	urban	free	never	0	5	0
23	no	0.999	0.001	TERRY	21	female	urban	free	yes	2	2	0
24	yes	0	1	MILES	32	male	rural	free	yes	14	14	0
25	no	1.000	0.000	AYERS	73	male	rural	free	never	0	0	0
26	no	1.000	0.000	RIVERS	18	male	rural	premium	never	0	0	0

k-NN

Row No.	prediction(response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 we...
1	no	1	0	CARRIER	54	female	rural	free	always	0	7	0
2	no	1	0	GONZALES	88	male	urban	free	never	3	3	0
3	no	1	0	MCNEIL	31	female	urban	free	always	0	7	0
4	no	1	0	TATE	54	female	rural	free	never	0	0	0
5	no	1	0	CARR	25	male	urban	free	never	0	0	0
6	no	1	0	ARNOLD	68	female	urban	free	never	0	0	0
7	yes	0	1	PICKETT	27	female	urban	free	yes	0	0	0
8	yes	0	1	MCCLAIN	34	male	urban	free	yes	0	0	0
9	yes	0	1	CONLEY	29	female	urban	free	never	5	5	0
10	no	1	0	BUSH	50	female	rural	free	never	0	0	0
11	no	1	0	MCCRAY	64	male	rural	free	never	0	0	0
12	no	1	0	O'BRIEN	61	female	urban	free	never	9	12	0
13	no	1	0	OLSEN	19	male	rural	free	yes	0	0	0
14	yes	0	1	HOLDER	58	female	urban	free	yes	8	8	0
15	yes	0	1	LYNCH	55	male	urban	free	yes	4	4	0
16	no	1	0	SHEPHERD	30	female	rural	free	never	2	2	0
17	yes	0	1	BARO	59	female	urban	free	yes	0	7	0
18	no	1	0	CALDWELL	66	male	urban	free	yes	0	0	0
19	no	1	0	MCMAHON	37	female	urban	premium	never	0	2	0
20	yes	0	1	BRUCE	57	male	rural	free	yes	0	7	0
21	no	1	0	SIMON	42	male	urban	free	never	5	5	0
22	yes	0	1	YANG	20	male	urban	free	never	0	5	0
23	no	1	0	TERRY	21	female	urban	free	yes	2	2	0
24	yes	0	1	MILES	32	male	rural	free	yes	14	14	0
25	no	1	0	AYERS	73	male	rural	free	never	0	0	0
26	no	1	0	RIVERS	18	male	rural	premium	never	0	0	0

5. Decision Tree, Naive Bayes and K-NN with subprocesses

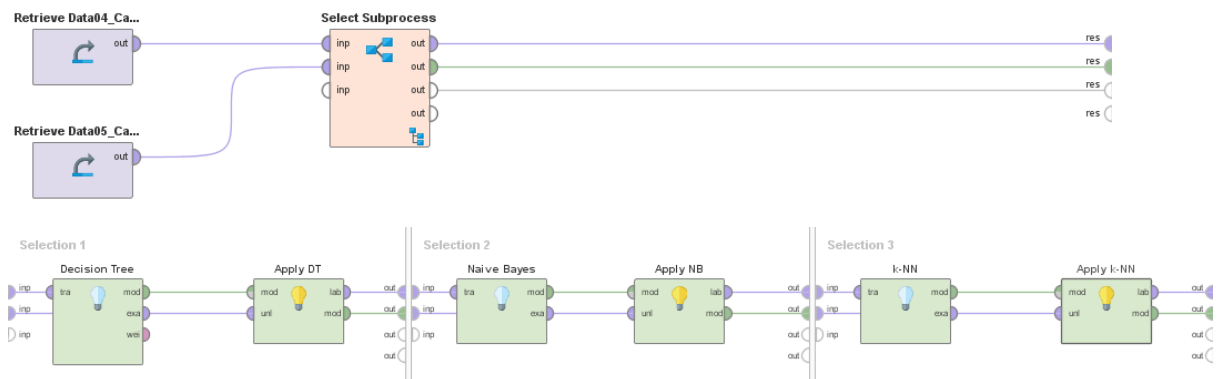
Operator ที่ใช้

Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

ดึงข้อมูล Data04_Campaign และ Data05_Campaign-new

Subprocess ใช้สำหรับรวบรวม process ที่สร้างขึ้นมาไว้ภายใน operator

ภายในได้รวบรวมโปรเซส Decision Tree, Naïve Bayes and k-NN



Result

Parameter 1

Subprocessor เลือกใช้งาน Decision Tree

ExampleSet (1000 examples, 3 special attributes, 11 regular attributes)													
Row No.	prediction(Response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 we...	
1	no	1	0	CARVER	54	female	rural	free	always	0	7	0	
2	no	0.997	0.003	GONZALES	88	male	urban	free	never	3	3	0	
3	no	1	0	MCNEIL	31	female	urban	free	always	0	7	0	
4	no	0.997	0.003	TATE	54	female	rural	free	never	0	0	0	
5	no	0.997	0.003	CARR	25	male	urban	free	never	0	0	0	
6	no	0.997	0.003	ARNOLD	68	female	urban	free	never	0	0	0	
7	no	0.922	0.078	PICKETT	27	female	urban	free	yes	0	0	0	
8	yes	0	1	MCCLAINE	34	male	urban	free	yes	0	0	0	
9	no	0.857	0.143	CONLEY	29	female	urban	free	never	5	5	0	
10	no	0.997	0.003	BUSH	50	female	rural	free	never	0	0	0	
11	no	0.997	0.003	MCCRAY	64	male	rural	free	never	0	0	0	
12	yes	0.009	0.991	OBRIEN	61	female	urban	free	never	9	12	0	
13	no	1	0	OLSEN	19	male	rural	free	yes	0	0	0	
14	yes	0.009	0.991	HOLDER	58	female	urban	free	yes	8	8	0	
15	yes	0	1	LYNCH	55	male	urban	free	yes	4	4	0	
16	no	0.997	0.003	SHEPHERD	30	female	rural	free	never	2	2	0	
17	no	0.922	0.078	BAIRD	59	female	urban	free	yes	0	7	0	
18	yes	0	1	CALDWELL	66	male	urban	free	yes	0	0	0	
19	yes	0.090	0.910	MCMAHON	37	female	urban	premium	never	0	2	0	
20	no	1	0	BRUCE	57	male	rural	free	yes	0	7	0	
21	yes	0.143	0.857	SIMON	42	male	urban	free	never	5	5	0	
22	no	0.997	0.003	YANG	20	male	urban	free	never	0	5	0	
23	no	0.922	0.078	TERRY	21	female	urban	free	yes	2	2	0	
24	yes	0.009	0.991	MILES	32	male	rural	free	yes	14	14	0	
25	no	0.997	0.003	AYERS	73	male	rural	free	never	0	0	0	
26	yes	0.090	0.910	RIVERS	38	male	rural	premium	never	0	0	0	
27	no	0.007	0.993	BEFET	22	male	urban	free	never	0	0	0	

Parameter 2

Subprocessor เลือกใช้งาน Naïve Bayes

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Result History X SimpleDistribution (Naive Bayes) X ExampleSet (Apply NB) X

ExampleSet (1000 examples, 3 special attributes, 11 regular attributes) Filter (1,000 / 1,000 examples): all

Row No.	prediction(Response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 wee...
1	no	1.000	0.000	CARVER	54	female	rural	free	always	0	7	0
2	no	1.000	0.000	GONZALES	88	male	urban	free	never	3	3	0
3	no	1.000	0.000	MCNEIL	31	female	urban	free	always	0	7	0
4	no	1.000	0.000	TATE	54	female	rural	free	never	0	0	0
5	no	1.000	0.000	CARR	25	male	urban	free	never	0	0	0
6	no	1.000	0.000	ARNOLD	68	female	urban	free	never	0	0	0
7	no	1.000	0.000	PICKETT	27	female	urban	free	yes	0	0	0
8	no	1.000	0.000	MCCLAIN	34	male	urban	free	yes	0	0	0
9	no	0.983	0.017	CONLEY	29	female	urban	free	never	5	5	0
10	no	1.000	0.000	BUSH	50	female	rural	free	never	0	0	0
11	no	1.000	0.000	MCCRAY	64	male	rural	free	never	0	0	0
12	yes	0	1	O'BRIEN	61	female	urban	free	never	9	12	0
13	no	1.000	0.000	OLSEN	19	male	rural	free	yes	0	0	0
14	yes	0.000	1.000	HOLDER	58	female	urban	free	yes	8	8	0
15	no	0.508	0.492	LYNCH	55	male	urban	free	yes	4	4	0
16	no	1.000	0.000	SHEPHERD	30	female	rural	free	never	2	2	0
17	no	1.000	0.000	BAIRD	59	female	urban	free	yes	0	7	0
18	no	1.000	0.000	CALDWELL	66	male	urban	free	yes	0	0	0
19	no	1.000	0.000	MCMAHON	37	female	urban	premium	never	0	2	0
20	no	0.998	0.002	BRUCE	57	male	rural	free	yes	0	7	0
21	yes	0.046	0.954	SIMON	42	male	urban	free	never	5	5	0
22	no	1.000	0.000	YANG	20	male	urban	free	never	0	5	0
23	no	0.999	0.001	TERRY	21	female	urban	free	yes	2	2	0
24	yes	0	1	MILES	32	male	rural	free	yes	14	14	0
25	no	1.000	0.000	AYERS	73	male	rural	free	never	0	0	0
26	no	1.000	0.000	RIVERS	38	male	rural	premium	never	0	0	0
27	no	1.000	0.000	PEREZ	22	male	urban	free	never	0	0	0

Parameter 3

Subprocessor เลือกใช้งาน k-NN

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Result History X KNNClassification (k-NN) X ExampleSet (Apply k-NN) X

ExampleSet (1000 examples, 3 special attributes, 11 regular attributes) Filter (1,000 / 1,000 examples): all

Row No.	prediction(Response)	confidence(no)	confidence(yes)	Name	Age	Gender	Area	Email	Mobile	Logins 4 we...	Logins 6 mo...	Sales 4 wee...
1	no	1	0	CARVER	54	female	rural	free	always	0	7	0
2	no	1	0	GONZALES	88	male	urban	free	never	3	3	0
3	no	1	0	MCNEIL	31	female	urban	free	always	0	7	0
4	no	1	0	TATE	54	female	rural	free	never	0	0	0
5	no	1	0	CARR	25	male	urban	free	never	0	0	0
6	no	1	0	ARNOLD	68	female	urban	free	never	0	0	0
7	yes	0	1	PICKETT	27	female	urban	free	yes	0	0	0
8	yes	0	1	MCCLAIN	34	male	urban	free	yes	0	0	0
9	yes	0	1	CONLEY	29	female	urban	free	never	5	5	0
10	no	1	0	BUSH	50	female	rural	free	never	0	0	0
11	no	1	0	MCCRAY	64	male	rural	free	never	0	0	0
12	no	1	0	O'BRIEN	61	female	urban	free	never	9	12	0
13	no	1	0	OLSEN	19	male	rural	free	yes	0	0	0
14	yes	0	1	HOLDER	58	female	urban	free	yes	8	8	0
15	yes	0	1	LYNCH	55	male	urban	free	yes	4	4	0
16	no	1	0	SHEPHERD	30	female	rural	free	never	2	2	0
17	yes	0	1	BAIRD	59	female	urban	free	yes	0	7	0
18	no	1	0	CALDWELL	66	male	urban	free	yes	0	0	0
19	no	1	0	MCMAHON	37	female	urban	premium	never	0	2	0
20	yes	0	1	BRUCE	57	male	rural	free	yes	0	7	0
21	no	1	0	SIMON	42	male	urban	free	never	5	5	0
22	yes	0	1	YANG	20	male	urban	free	never	0	5	0
23	no	1	0	TERRY	21	female	urban	free	yes	2	2	0
24	yes	0	1	MILES	32	male	rural	free	yes	14	14	0
25	no	1	0	AYERS	73	male	rural	free	never	0	0	0
26	no	1	0	RIVERS	38	male	rural	premium	never	0	0	0
27	no	1	0	PEREZ	22	male	urban	free	never	0	0	0

6. Cross Validation - Decision Tree

Operator ที่ใช้

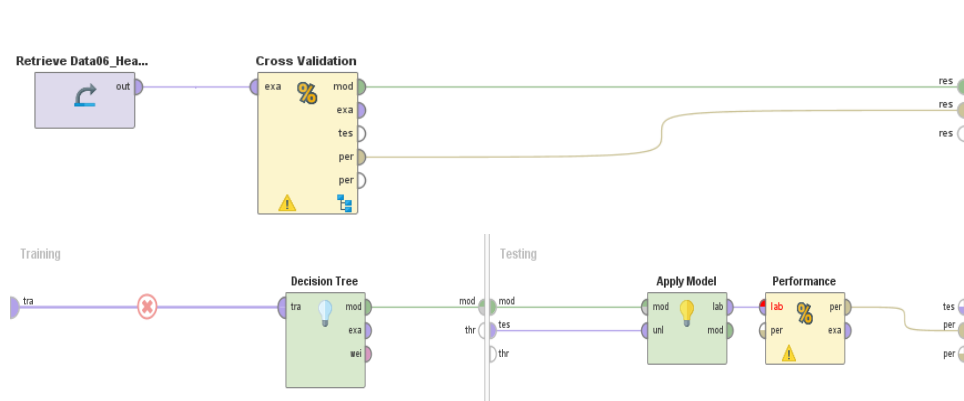
Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

ดึงข้อมูล Data06_Heart-Attack

Cross Validation ใช้ในการทำนายข้อมูล

Performant ใช้ในวัดประสิทธิภาพของข้อมูล

หลักการของ 10-fold cross-validation โดยพฤติกรรมของการทำ 10-fold จะทำการแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากจะนำข้อมูลจำนวน 9 ส่วนไปสร้าง Model (Train Model) เมื่อ Train Model เสร็จ ก็จะใช้ข้อมูล 1 ส่วน เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ นั่นหมายความว่ามีการ Train และ ทดสอบ 9 ครั้ง

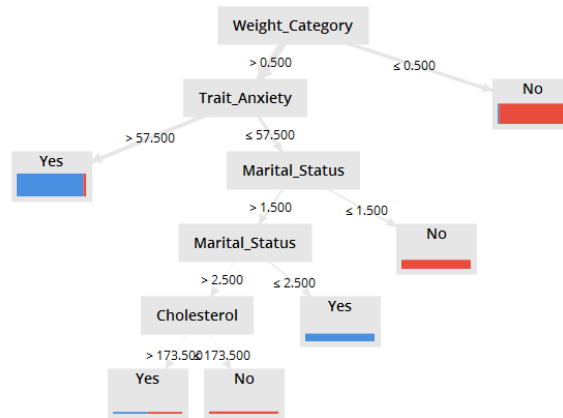


Confusion Matrix

accuracy: 94.95% +/- 5.60% (mikro: 94.93%)

	true Yes	true No	class precision
pred. Yes	67	6	91.78%
pred. No	1	64	98.46%
class recall	98.53%	91.43%	

Tree และกฎ ที่ได้ออกมา



Tree

```
Email = free
| Logins 4 weeks > 4.500
| | Logins 6 months > 6.500: yes {no=1, yes=93}
| | Logins 6 months ≤ 6.500
| | | Gender = female: no {no=7, yes=1}
| | | Gender = male
| | | | Area = rural: no {no=1, yes=1}
| | | | Area = urban: yes {no=0, yes=9}
| | Logins 4 weeks ≤ 4.500
| | | Mobile = always
| | | | Sales Total > 28
| | | | | Age > 49: yes {no=0, yes=3}
| | | | | Age ≤ 49: no {no=4, yes=1}
| | | | Sales Total ≤ 28: no {no=26, yes=0}
| | | | Mobile = never
| | | | | Sales 6 months > 116
| | | | | | Area = rural: no {no=1, yes=1}
| | | | | | Area = urban: yes {no=0, yes=16}
| | | | | Sales 6 months ≤ 116
| | | | | | Logins 6 months > 2.500
| | | | | | | Sales Total > 45
| | | | | | | | Logins 4 weeks > 1: yes {no=0, yes=4}
| | | | | | | | Logins 4 weeks ≤ 1
| | | | | | | | | Logins 6 months > 5.500
| | | | | | | | | | Age > 30: no {no=1, yes=1}
| | | | | | | | | | Age ≤ 30: yes {no=0, yes=2}
| | | | | | | | | Logins 6 months ≤ 5.500: no {no=14, yes=1}
| | | | | | | Sales Total ≤ 45
| | | | | | | | Logins 6 months > 10.500
| | | | | | | | | Gender = female: no {no=5, yes=1}
| | | | | | | | | Gender = male
| | | | | | | | | | Age > 66.500: no {no=2, yes=0}
| | | | | | | | | | Age ≤ 66.500: yes {no=0, yes=3}
| | | | | | | | Logins 6 months ≤ 10.500
| | | | | | | | | Logins 4 weeks > 3.500
| | | | | | | | | | Age > 39.500: no {no=7, yes=0}
| | | | | | | | | | Age ≤ 39.500: yes {no=1, yes=2}
| | | | | | | | | Logins 4 weeks ≤ 3.500: no {no=82, yes=0}
| | | | | | | | Logins 6 months ≤ 2.500: no {no=280, yes=1}
```

7. Neural Network with 10-Fold Cross Validation

Operator ที่ใช้

Retrieve ใช้สำหรับดึงข้อมูลที่เก็บไว้ใน Repository มาใช้งานใน Process

ดึงข้อมูล Data10_Student

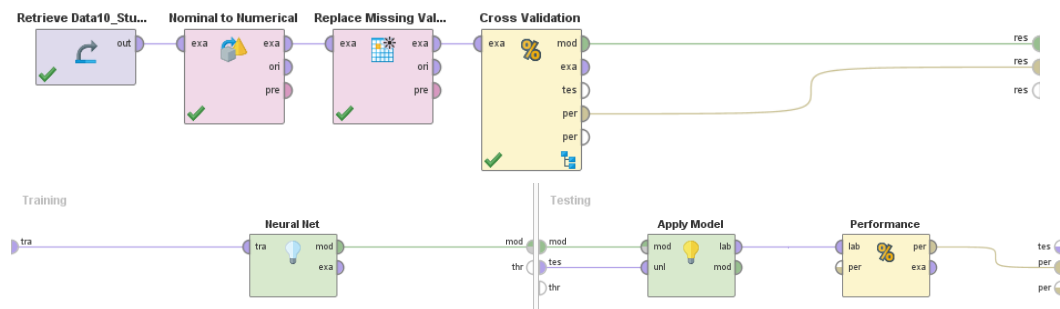
Cross Validation ใช้ในการทำนายข้อมูล

Performant ใช้ในวัดประสิทธิภาพของข้อมูล

Neural Net เป็นอัลกอริทึมในการจำแนกข้อมูล

Nominal to Numerical เป็นการแปลงข้อมูลจากข้อมูลที่แทนด้วยตัวหนังสือเป็นตัวเลข

Replace Missing Values เป็นการแทนที่ค่าที่หายไป

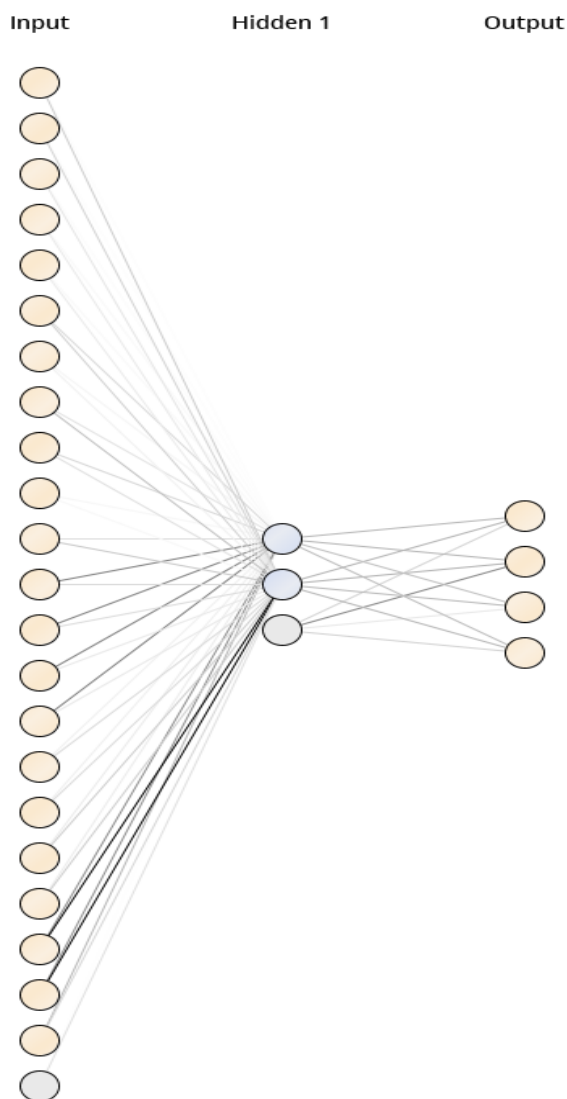


Confusion Matrix

accuracy: 83.90% +/- 3.73% (mikro: 83.90%)

	true computer	true marketing	true management	true hotel	class precision
pred. computer	315	15	26	15	84.91%
pred. marketing	5	100	7	10	81.97%
pred. management	10	4	102	10	80.95%
pred. hotel	20	18	21	322	84.51%
class recall	90.00%	72.99%	65.38%	90.20%	

Neural Network



8. บทความที่เกี่ยวข้องกับเนื้อหาในบทเรียน

1.1 บทความภาษาอังกฤษ

International Journal of Science and Research (IJSR)

ISSN (Online): 2319-7064

Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques

Sayali D. Jadhav¹, H. P. Channe²

^{1,2}Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, India

- a) ระบบทำอะไร เป็นการเปรียบเทียบเทคนิคในการจำแนกข้อมูลแบบ unsupervised learning
- b) ใช้เทคนิคอะไร ใช้เทคนิค k-nearest neighbors, Naïve Bayes และ Decision Tree
- c) ผลลัพธ์เป็นอย่างไร

Parameter	KNN	Naive Bayes	Decision Tree
Deterministic/Non-deterministic	Non-deterministic	Non-deterministic	Deterministic
Effectiveness on	Small data	Huge data	Large data
Speed	Slower for large data.	Faster than KNN.	Faster
Dataset	It can't deal with noisy data.	It can deal with noisy data.	It can deal with noisy data.
Accuracy	Provides high accuracy.	For obtaining good results it requires a very large number of records.	High accuracy

Table 2: Results of Accuracy of Classifiers

Dataset	Size of Dataset	KNN	Naïve Bayes	Decision Tree
Weather Nominal	Small (14 instances)	100%	92.857%	100%
Segment Challenge	Medium (1500 instances)	100%	81.667%	99%
Supermarket	Large (4627 instances)	89.842%	63.713%	63.713%

Table 3: Results of Time taken for Classification

Dataset	Size of Dataset	Time	KNN	Naïve Bayes	Decision Tree
Weather Nominal	Small (14 instances)	To Build Model	0 sec	0 sec	0.02 sec
		To Test Model	0.02 sec	0 sec	0 sec
Segment Challenge	Medium (1500 instances)	To Build Model	0 sec	0.08 sec	0.16 sec
		To Test Model	0.42 sec	0.31 sec	0.06 sec
Super market	Large (4627 instances)	To Build Model	0.02 sec	0.06 sec	0.06 sec
		To Test Model	45.55 sec	0.28 sec	0.03 sec

ผลลัพธ์ที่ได้จากการทดลองจะเห็นได้ว่าแต่ละวิธีมีข้อดีแตกต่างกัน

- d) เนื้อหาที่ตรงกับที่เรียน k-nearest neighbors, Naïve Bayes และ Decision Tree

1.2 บทความภาษาไทย

NECTEC Technical Journal Vol. III, No. 11

การใช้เทคนิคดาต้าไมนิงเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์

กฤษณะ ไวยมัย, ชิคชนก ส่งศิริ และธนาวิทย์ รักธรรมานนท์

อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
และนิสิตปริญญาโทวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

- a) ระบบทำอะไร นำความรู้ทางด้านดาต้าไมนิงมาประยุกต์ใช้กับข้อมูลนิตินิต คณะวิศวกรรมศาสตร์ เพื่อเป็นแนวทางในการแก้ไขปัญหาต่าง ๆ อาทิเช่น ปัญหา การเลือกสาขาวิชาไม่ตรงกับความสามารถที่แท้จริง ปัญหาผลการเรียนของนิสิต ตกต่ำจนต้องออกจากสถาบันการศึกษา
- b) ใช้เทคนิคอะไร Association rule discovery, data classification และ data prediction
- c) ผลลัพธ์เป็นอย่างไร

โมเดลกลางการ จำแนกประเภทข้อมูล	โมเดลการจำแนก ประเภทข้อมูล ในแต่ละสาขาวิชา	โมเดลการพยากรณ์ ข้อมูล ในแต่ละสาขาวิชา
1. ความน่าเชื่อถือที่น้อย ประมาณ 50% เนื่องจาก กลุ่มเป้าหมายมาก	1. ผลการทดสอบที่ได้มี ความถูกต้องสูง 84.58 %	1. ผลการทดสอบที่ได้ มีความถูกต้องสูง 96.84 %
2. ข้อมูลนิตินิตในสาขา วิชาต่างๆ มีจำนวนแตกต่างกัน มาก ทำให้ โมเดลการทำนายโอน เชิงไปทางสาขาวิชาที่ มีนิตินิตมาก	2. ข้อมูลแต่ละสาขาวิชา ไม่ส่งผลกระทบต่อกัน เนื่องจากวิธีนี้ได้สร้าง โมเดลแยกกันในแต่ละ สาขาวิชา	2. ข้อมูลแต่ละสาขา วิชาไม่ส่งผลกระทบต่อกัน เนื่องจากวิธีนี้ได้ สร้างโมเดลแยกกันในแต่ละ สาขาวิชา
3. ต้องมีการจัดกลุ่มผล การเรียนในแต่ละราย วิชา (High, Medium, Low) เพื่อลดการ กระจายตัวของข้อมูล ทั้งนี้ ถ้าไม่มีการจัดกลุ่ม ข้อมูล จะทำให้โมเดลที่ ได้กระจายตัว ข้อมูลใน	3. ต้องมีการจัดกลุ่มผล การเรียนในแต่ละรายวิชา (High, Medium, Low) เพื่อลดการกระจายตัวของ ข้อมูล ทั้งนี้ ถ้าไม่มีการ จัดกลุ่มข้อมูล จะทำให้ โมเดลที่ได้กระจายตัว ข้อมูล ในแต่ละเส้นทางของ	3. ข้อมูลผลการเรียน ในแต่ละรายวิชาเป็น ข้อมูลผลการเรียนจริง (A, B+, B, C+, C, D+, D, F) ที่มีได้มีการจัด กลุ่ม ทำให้ข้อมูลที่น่า มาสร้างโมเดล Prediction นั้นมีความ

แต่ละเส้นทางของโมเดลมีจำนวนน้อย เป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	โมเดลมีจำนวนน้อยเป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	ละเอียดและแม่นยำมากกว่าการจัดกลุ่มดังเช่นโมเดลการจำแนกประเภทข้อมูล
4. โมเดลนำเสนอเพียงสาขาวิชาเดียวที่เหมาะสม ซึ่งส่งผลกระทบต่อตรงกับการตัดสินใจของนิสิต	4. โมเดลนำเสนอเฉพาะสาขาวิชาที่เหมาะสมให้กับนิสิตเท่านั้น สำหรับนิสิตบางส่วนที่มีผลการเรียนดี โมเดลจะเสนอทุกสาขาวิชาให้กับนิสิตเป็นสาขาวิชาที่เหมาะสม และสำหรับนิสิตบางส่วนที่มีผลการเรียนไม่ดี โมเดลจะไม่นำเสนอสาขาวิชาใดๆ ที่เหมาะสมให้กับนิสิตเลย ทำให้การตัดสินใจทั้งหมดไปตกอยู่กับนิสิต โดยที่โมเดลมิได้ช่วยนิสิตในกลุ่มเหล่านี้เลย	4. โมเดลนำเสนอแนวโน้มเกรดเฉลี่ยสะสมเมื่อจบการศึกษาของนิสิตในทุกสาขาวิชาทำให้นิสิตได้เห็นแนวโน้ม และเห็นความแตกต่างของผลการเรียนของตน เมื่อเข้าไปศึกษาในสาขาวิชาที่ต่างกันไป นอกจากนี้การนำเสนอได้เพิ่มในส่วนของ MEAN และ TOP30% ทั้งที่ได้เคยกล่าวไป ทำให้ช่วยให้นิสิตได้เห็นความแตกต่างในการเรียนในแต่ละสาขาวิชามากยิ่งขึ้น

ผลลัพธ์ที่ได้จากงานวิจัยนี้ค่อนข้างเป็นที่น่าพอใจ โดยมีเปอร์เซ็นต์ความถูกต้องค่อนข้างสูง แต่มีปัญหาบางประการ ได้แก่ จำนวนข้อมูลในบางสาขาวิชาที่มีปริมาณค่อนข้างน้อยทำให้โมเดลที่ได้ไม่แม่นยำเท่าที่ควร

- d) เนื้อหาที่ตรงกับที่เรียน Association rule discovery, data classification และ data prediction