การวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์

BEHAVIORS ANALYSIS FOR PREDICTING TYPE OF CARS USAGE

ณัฏฐินี อินทรชิต
วนิดา ลำใย
วรรณวิสา มานพวงษ์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2560

BEHAVIORS ANALYSIS FOR PREDICTING TYPE OF CARS USAGE

NUTTHINEE INTORRACHIT
WANIDA LUMYAI
WANWISA MANOPWONG

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF SCIENCE IN APPLIED MATHEMATICS
DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2017

หัวข้อปัญหาพิเศษ	การวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ Behaviors Analysis for Predicting Type of Cars Usage
ชื่อนักศึกษา	นางสาวณัฏฐินี อินทรชิต รหัสนักศึกษา 57050045 นางสาววนิดา ลำใย รหัสนักศึกษา 57050120 นางสาววรรณวิสา มานพวงษ์ รหัสนักศึกษา 57050125
ปริญญา	วิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์)
ภาควิชา	คณิตศาสตร์
ปีการศึกษา	2560
อาจารย์ที่ปรึกษา	ผศ.ดร. กาญจนา คำนึงกิจ

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์) ประจำปีการศึกษา 2560

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ดร.อาทิตย์ แข็งธัญการ ประธานกรรมการ	
ดร. กัมปนาท นามงาม กรรมการ	
ผศ.ดร. กาญจนา คำนึงกิจ กรรมการและอาจารย์ที่ปรึกษา	

หัวข้อปัญหาพิเศษ	การวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์
ชื่อนักศึกษา	นางสาวณัฏฐินี อินทรชิต รหัสนักศึกษา 57050045
	นางสาววนิดา ลำใย รหัสนักศึกษา 57050120
	นางสาววรรณวิสา มานพวงษ์ รหัสนักศึกษา 57050125
ปริญญา	วิทยาศาสตรบัณฑิต คณิตศาสตร์ประยุกต์
ภาควิชา	คณิตศาสตร์
คณะ	วิทยาศาสตร์
มหาวิทยาลัย	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)
ปีการศึกษา	2560
อาจารย์ที่ปรึกษาปัญหาพิเศษ	ผศ.ดร. กาญจนา คำนึงกิจ

บทคัดย่อ

ปัญหาพิเศษนี้ได้ทำการศึกษาเรื่องการหาความสัมพันธ์ระหว่างปัจจัยในการเลือกซื้อรถยนต์กับรถยนต์ที่กลุ่มตัวอย่างใช้ แล้วนำมาวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ที่ประชากรใช้งานว่ามีความถูกต้องมากน้อยเพียงใด โดยใช้วิธีการทำเหมืองข้อมูล (Data Mining) ซึ่งการศึกษาเริ่มตั้งแต่การรวบรวมข้อมูลพฤติกรรมการใช้รถยนต์ประเภทต่าง ๆ ของกลุ่มตัวอย่างบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา จำนวน 450 ตัวอย่าง มาทำการเตรียมข้อมูลและคัดเลือกข้อมูลให้พร้อมเข้าสู่กระบวนการการทำเหมืองข้อมูล (Data Mining) โดยกำหนดประเภทของรถยนต์ที่ใช้พิจารณาเป็น 4 ประเภท คือ รถเก๋ง รถอเนกประสงค์ รถกระบะและไม่ใช้รถยนต์ ซึ่งได้ทดลองผ่านวิธี 2 วิธี คือ วิธี Naïve Bayesian และวิธี Decision Tree โดยใช้โปรแกรม RapidMiner Studio ซึ่งให้ผลลัพธ์เป็นโมเดลและการทำนายค่าใหม่ของกลุ่มตัวอย่าง จากนั้นจึงนำโมเดลที่ได้มาหาค่าความถูกต้อง (Accuracy) และทำนายตัวอย่างที่ไม่รู้คลาสจำนวน 45 ตัวอย่าง ผลการทดลองพบว่าการจำแนกกลุ่มด้วยวิธี Decision Tree และ Naïve Bayesian มีค่าความถูกต้องในการทำนายค่าใหม่เป็น 72.73% และ 59.09% ตามลำดับ

Title	Behaviors Analysis for Predicting Type of Cars Usage		
Student Name	Miss Nutthinee	Intorrachit	Student ID 57050045
	Miss Wanida	Lumyai	Student ID 57050120
	Miss Wanwisa	Manopwong	Student ID 57050125
Degree	Degree of bachelor of Science Applied Mathematics		
Department	Mathematics		
Faculty	Science		
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)		
Academic Year	2017		
Advisor	Asst. Prof. Kanchana Kumnungkit		

Abstract

We have studied the relationship among the factors of car purchased that the sample using to forecast the type of car used by the population through "Data Mining". The study used the samples of King Mongkut's Institute of Technology Ladkrabang and Chachoengsao, 450 samples. The types of car considered are sedan, pick-up truck, sport-utility vehicle, and non-car group in this study. We used the Naïve Bayesian method and the Decidion Tree method using the Rapid Miner Studio program. We found that the accuracy values from the methods were 59.09% and 72.73% respectively.

กิตติกรรมประกาศ

ปัญหาพิเศษฉบับนี้สำเร็จลุล่วงได้ด้วยดี โดยได้รับความอนุเคราะห์อย่างดีจาก ผศ.ดร. กาญจนา คำนึงกิจ ที่กรุณาเสียสละเวลาอันมีค่าและให้คำแนะนำ ชี้แนะแนวทางการค้นคว้าข้อมูลเพิ่มเติมอันเป็นประโยชน์ทำให้ปัญหาพิเศษฉบับนี้สมบูรณ์ คณะผู้จัดทำขอขอบพระคุณและจารึกพระคุณนี้ไว้ในความทรงจำมิรู้ลืมเลือน

ขอขอบพระคุณ ผู้ให้ความอนุเคราะห์และเสียสละเวลาอันมีค่าในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา ในการให้ข้อมูลพฤติกรรมการใช้รถยนต์อันเป็นข้อมูลที่สำคัญยิ่งในการทำปัญหาพิเศษฉบับนี้

คุณค่าและประโยชน์ใดๆ ที่อาจมีจากงานปัญหาพิเศษฉบับนี้ คณะผู้จัดทำขอมอบเป็นเครื่องบูชาพระคุณของบิดามารดาที่ให้กำเนิดและเลี้ยงดูให้การศึกษา ตลอดจนครูบาอาจารย์และผู้ที่มีพระคุณทุกท่านที่มีส่วนในการวางรากฐานการศึกษาให้แก่คณะผู้จัดทำ

คณะผู้จัดทำหวังเป็นอย่างยิ่งว่า ปัญหาพิเศษฉบับนี้จะเป็นประโยชน์แก่ผู้ที่ต้องการศึกษาการใช้เหมืองข้อมูลที่ช่วยในการศึกษาพฤติกรรมการใช้รถยนต์ และหากมีข้อผิดพลาดประการใดในปัญหาพิเศษฉบับนี้ คณะผู้จัดทำขอกราบขออภัยอย่างสูงมา ณ ที่นี้ด้วย

นางสาวณัฏฐินี อินทรชิต
นางสาววนิดา ลำใย
นางสาววรรณวิสา มานพวงษ์

คำย่อ/สัญลักษณ์

คำย่อ/สัญลักษณ์	คำอธิบาย
$P(C_i \mid X)$	ความน่าจะเป็นที่กลุ่มของข้อมูล X จะอยู่ในกลุ่มข้อมูล C กลุ่มที่ i
$P(A \cap B)$	ความน่าจะเป็นที่เหตุการณ์ A และ เหตุการณ์ B เกิดขึ้นร่วมกัน
$P(C_i)$	class prior probability ของกลุ่มของข้อมูลกลุ่มที่ i
$X = (x_1, x_2, x_3, \dots, x_n)$	กลุ่มของข้อมูล X ที่ประกอบด้วยค่าของค่าของแอททริบิวต์ x ทั้งหมด n ค่า
S	จำนวนข้อมูลทั้งหมดในกลุ่มของข้อมูล
S_i	จำนวนข้อมูลทั้งหมดในกลุ่มของข้อมูลที่ i
$Gain(A)$	ค่าเกนความรู้สำหรับการพิจารณา แอททริบิวต์ A จะสามารถคำนวณได้จากค่าความแตกต่างระหว่างปริมาณข้อมูลที่ต้องการในการระบุถึงหมวดหมู่ของข้อมูลสำหรับเรคคอร์ดหนึ่ง ๆ กับจำนวนข้อมูลที่คาดว่าจะใช้สำหรับทำการแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย โดยทำการพิจารณาแอททริบิวต์ A
$Gini(D)$	เป็นตัวชี้วัดที่จะทำการพิจารณาความไม่บริสุทธิ์ของชุดข้อมูล D ที่ซึ่งจะมีเซตของเรคคอร์ดที่มีหมวดหมู่ของข้อมูลไม่เหมือนกันอยู่
$\mathrm{C_i(for\ i = 1, \dots, m)}$	แอททริบิวต์ที่เป็นหมวดหมู่ของข้อมูล มีค่าที่เป็นไปได้ทั้งสิ้น m หมวดหมู่
$C_{i,D}$	เซตของเรคคอร์ดที่อยู่ในหมวดหมู่ C_i
$Info(D)$	ค่าเอนโทรปี (entropy of D) หรือค่าเฉลี่ยของปริมาณข้อมูลที่ต้องการในการระบุถึงหมวดหมู่ของข้อมูลเรคคอร์ดหนึ่ง ๆ ในชุดข้อมูลที่จะขึ้นกับอัตราส่วนของจำนวนเรคคอร์ดที่สอดคล้องกับแต่ละหมวดหมู่
$Info_A(D)$	จำนวนข้อมูลที่คาดว่าจะใช้สำหรับทำการแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย โดยทำการพิจารณาแอททริบิวต์ A

สารบัญ

สารบัญ (ต่อ)

เรื่อง **หน้า**

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1

บทนำ

ในบทนี้เป็นการนำเสนอความเป็นมา วัตถุประสงค์ของงานวิจัย ขอบเขตงานวิจัย ประโยชน์ที่ได้รับและระยะเวลาดำเนินงานของงานวิจัยนี้ เพื่อเป็นการกำหนดแนวทางในการดำเนินงานต่อไป

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันรถยนต์เป็นยานพาหนะในการเดินทางที่ประชากรในประเทศไทยมีผู้ใช้งานเป็นจำนวนมากในทุกสาขาอาชีพหรือทุกธุรกิจ ซึ่งมีการใช้งานรถยนต์หลากหลายประเภทตามความต้องการของประชากร เช่น การใช้งานรถยนต์รถเก๋ง 4 ประตูใน อาชีพพนักงานบริษัท การใช้รถกระบะในอาชีพค้าขาย การใช้รถตู้ในครอบครัวใหญ่ หรือไม่ใช้รถยนต์ในกลุ่มนักศึกษา เป็นต้น ดังนั้น จะเห็นว่าประชากรมีการเลือกซื้อรถยนต์ที่เหมาะสมกับการใช้งานที่ทำให้เกิดประโยชน์สูงสุด ด้วยเหตุนี้ทางคณะผู้จัดทำจึงเลือกทำปัญหาพิเศษที่เกี่ยวกับการเลือกใช้งานรถยนต์ ที่เหมาะสมกับการใช้งานของประชากร โดยทำการเก็บข้อมูลพฤติกรรมการใช้งานรถยนต์จำกกลุ่มตัวอย่างในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร-ลาดกระบังและจังหวัดฉะเชิงเทรา แล้วนำมาวิเคราะห์เพื่อหาพฤติกรรมโดยรวมในการใช้งานรถยนต์ผ่านกระบวนการการทำเหมืองข้อมูล (Data mining) เพื่อหาพฤติกรรมการใช้งานรถยนต์ของประชากรและสามารถทำนายประเภทของรถยนต์ของประชากรจากข้อมูลส่วนตัวของประชากรได้

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อศึกษาพฤติกรรมการใช้งานรถยนต์และทำนายข้อมูล จากกลุ่มตัวอย่างในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา

2) เพื่อศึกษาความสัมพันธ์ของปัจจัยในการเลือกซื้อรถยนต์กับรถยนต์ที่กลุ่มตัวอย่างใช้ในชีวิตประจำวัน

3) เพื่อเรียนรู้กระบวนการการสร้างเหมืองข้อมูล (Data mining)

1.3 ขอบเขตของงานวิจัย

เริ่มจากการเก็บข้อมูลพฤติกรรมการใช้งานรถยนต์ เช่น การใช้งานรถยนต์เพื่อไว้สำหรับการทำงาน ตัวอย่างเช่น ทำงานขายสินค้าควรใช้รถยนต์ประเภทใดที่เหมาะสมกับการใช้งาน โดยทำการเก็บข้อมูลพฤติกรรมการใช้งานรถยนต์จำกกลุ่มตัวอย่างจากบริเวณสถาบันเทคโนโลยีพระจอมเกล้า

เจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทราเป็นจำนวน 450 ตัวอย่าง แล้วนำข้อมูลมาวิเคราะห์เพื่อหาพฤติกรรมการใช้งานรถยนต์โดยรวมในการใช้งานรถยนต์ผ่านกระบวนการการทำเหมืองข้อมูล (Data mining) โดยเมื่อทำการวิเคราะห์เรียบร้อยแล้ว จะได้ผลการทดสอบประสิทธิภาพของกระบวนการการทำเหมืองข้อมูล (Data mining) ว่ามีความถูกต้องในการทำนายมากน้อยเพียงใด จากนั้นจะนำผลที่ได้ ไปทำนายตัวอย่างที่ไม่รู้คลาส

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถวิเคราะห์พฤติกรรมและทำนายการซื้อรถยนต์ของประชากรได้

1.5 ระยะเวลาการดำเนินงาน

10 เดือน

ระยะเวลาการดำเนินงานตามที่แสดงไว้ในตารางที่ 1.1

ตารางที่ 1.1 แสดงระยะเวลาการดำเนินงานตามแผนงาน

การดำเนินงาน	ระยะเวลา									
	ปี 2560					ปี 2561				
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.
1) รวบรวมและศึกษาข้อมูลทั่วไปเกี่ยวกับData Mining	←	→								
2)รวบรวมและศึกษาการใช้โปรแกรม RapidMiner Studio 7		←	→							
3)ออกแบบ แบบสำรวจที่จะเก็บพฤติกรรมการใช้รถยนต์จาก		←	→							
4)เก็บข้อมูลพฤติกรรมการใช้รถยนต์จากกลุ่มตัวอย่าง			←	→						
5)วิเคราะห์เพื่อหาพฤติกรรมการใช้งานรถยนต์					←	→				
6)ประเมินและสรุปผล						←	→	→		
7)จัดทำเล่มปัญหาพิเศษ พร้อมทั้งนำเสนอ		←	→	→				←	→	→

จากวัตถุประสงค์ของการทำปัญหาพิเศษเล่มนี้ คือ เพื่อศึกษาพฤติกรรมการใช้งานรถยนต์และทำนายข้อมูล จากกลุ่มตัวอย่างในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา จำเป็นต้องมีความรู้พื้นฐานทางด้านการทำเหมืองข้อมูลและข้อมูลที่ต้องทำการสำรวจ เช่น ความหมายของคำว่าพฤติกรรม รถยนต์ประเภทต่างๆ เทคนิค Decision Tree และ Naïve Bayesian ที่ใช้ในการทำเหมืองข้อมูล และโปรแกรม RapidMiner Studio ซึ่งจะกล่าวถึงในบทที่ 2 การกำหนดประชากรและเลือกกลุ่มตัวอย่าง การสร้างแบบสอบถาม การเตรียมข้อมูล การทดสอบข้อมูล การทดสอบข้อมูลที่ไม่รู้คลาส ซึ่งจะกล่าวถึงในบทที่ 3 ผลการวิเคราะห์ผลการทดสอบของวิธี Decision Tree และวิธี Naïve Bayesian ซึ่งจะกล่าวถึงในบทที่ 4 สรุปผลการวิจัยและข้อเสนอแนะซึ่งจะกล่าวถึงในบทที่ 5

บทที่ 2

ความรู้พื้นฐานที่เกี่ยวข้อง

บทนี้เป็นการนำเสนอความรู้พื้นฐานที่ใช้เป็นกรอบในการทำปัญหาพิเศษเล่มนี้ทั้งด้านการทำเหมืองข้อมูลและการใช้โปรแกรมคอมพิวเตอร์ในการจำแนก ศึกษา และทำนายข้อมูลที่ได้เก็บรวบรวมมา

2.1 พฤติกรรม (Behavior)

พฤติกรรม หมายถึง การแสดงและกิริยาท่าทางของสิ่งมีชีวิตที่เกิดร่วมกันกับสิ่งแวดล้อม เป็นการตอบสนองของระบบหรือสิ่งมีชีวิตต่อสิ่งเร้าหรือการรับเข้าทั้งหลาย ไม่ว่าจะเป็นภายในหรือภายนอก มีสติหรือไม่มีสติระลึก ชัดเจนหรือแอบแฝง และโดยตั้งใจหรือไม่ได้ตั้งใจ นอกจากนั้นยังมีนักวิชาการได้ให้ความหมายไว้หลากหลาย เช่น

สมโภชน์ เอี่ยมสุภาษิต ผู้เชี่ยวชาญด้านการปรับพฤติกรรมได้ให้ความหมายของคำว่า พฤติกรรม หมายถึง สิ่งที่บุคคลกระทำ แสดงออกมา ตอบสนอง หรือโต้ตอบต่อสิ่งใดสิ่งหนึ่ง สภาพการณ์ใด สภาพการณ์หนึ่งโดยที่ผู้อื่นสามารถสังเกตได้

พฤติกรรม หมายถึง กริยาอาการหรือปฏิกิริยาที่แสดงออก หรือเกิดขึ้นเมื่อเผชิญกับสิ่งเร้า ซึ่งจะมาจากภายในร่างกายหรือภายนอกร่างกายก็ได และปฏิกิริยาที่แสดงออกนี้มิได้เป็นพฤติกรรมทางกายเท่านั้น แต่รวมถึงพฤติกรรมที่เกี่ยวกับจิตใจด้วย คำว่า Behavior ใช้แทนกันได้กับคำว่า Action นักจิตวิทยาถือว่าการเคลื่อนไหวของอินทรีย์ทุกชนิดที่ปรากฏออกมาเป็นพฤติกรรม หรือกล่าวอีกนัยหนึ่งพฤติกรรมจะเกิดขึ้นได้ตองมีมูลเหตุอย่างใดอย่างหนึ่ง (อุทัย หิรัญโต, 2526: 14)

พฤติกรรม คือ อาการ บทบาท ลีลา ท่าที การประพฤติ การปฏิบัติ การกระทำที่แสดงออกให้ปรากฏสัมผัสด้วยประสาทสัมผัสทางใดทางหนึ่ง คือ โสตสัมผัส ชิวหาสัมผัส และทางผิวหนัง หรือมิฉะนั้นก็สามารถวัดได้โดยเครื่องมือ (กันยา สุวรรณแสง, 2538:92)

พฤติกรรม หมายถึง การแสดงออกในลักษณะต่าง ๆ ของสิ่งมีชีวิตซึ่งอาจจะเกิดขึ้นได้ทั้งมนุษย์และสัตว์ พืช จุลินทรีย์ ซึ่งเป็นการตอบสนองสิ่งเร้าที่เกิดขึ้นภายในร่างกายหรือภายนอกร่างกาย พฤติกรรมนี้สามารถสังเกตได้โดยตรงหรือใช้เครื่องมือวัดได หรืออาจสังเกตได้ในทางออม เช่น การพูด การเคลื่อนไหว การทำงานของระบบต่าง ๆ ภายในร่างกาย การจำ การคิด ตลอดจนความรูสึก ทัศนคติ (เฉลิมพล ตันสกุล,2541: 2)

พฤติกรรมในมนุษย์ หมายถึง อาการกระทำ หรือกิริยาที่แสดงออกมาทางร่างกาย กล้ามเนื้อ สมองในทางอารมณ์ ความคิด และความรูสึก พฤติกรรมเป็นผลจากการตอบสนองต่อสิ่งเร้า เมื่อมีสิ่งกระตุ้นมาจะมีการตอบสนองทันที (ลักขณา สรวัฒน,2544: 17)

2.1.2 องค์ประกอบของพฤติกรรม

พฤติกรรมของมนุษย์มีองค์ประกอบที่สำคัญ ดังนี้

1. การรับรู้ เป็นการแปลความหมายจากการสัมผัส โดยเริ่มตั้งแต่การมีสิ่งเร้ามากระทบกับอวัยวะรับสัมผัสทั้งห้า และส่งกระแสประสาทไปยังสมองเพื่อการแปลความ

2. การเรียนรู้ เป็นการเปลี่ยนแปลงพฤติกรรมของบุคคลค่อนข้างถาวร อันเป็นผลมาจากประสบการณ์หรือการฝึกฝน มิใช่ผลจากการตอบสนองของสัญชาติญาณ อุบัติเหตุ หรือความบังเอิญ

3. การคิด เป็นกระบวนการของสมองในการสร้างสัญลักษณ์หรือภาพให้ปรากฏในสมอง เพื่อเป็นตัวแทนของวัตถุ สิ่งของ เหตุการณ์ หรือสถานการณ์ต่างๆ

2.2 รถยนต์ (Cars)

รถยนต์ หมายถึง ยานพาหนะทางบกที่ขับเคลื่อนที่ด้วยพลังงานอย่างใดอย่างหนึ่งและถ่ายทอดลงสู่ล้อเพื่อพาผู้ขับ ผู้โดยสาร หรือสิ่งของไปยังจุดหมายปลายทาง ปัจจุบันรถยนต์โดยส่วนมากได้รับการออกแบบอย่างซับซ้อนในทางวิศวกรรม และหลากหลายประเภทตามความเหมาะสมของการใช้งานหรือใช้สำหรับงานเฉพาะกิจ โดยในปัญหาพิเศษนี้ ทางคณะผู้จัดทำขอแบ่งประเภทของรถยนต์เป็น 3 ประเภท คือ รถเก๋ง รถ SUV และรถกระบะ

2.2.1 ประเภทของรถยนต์

ในปัญหาพิเศษนี้ ทางคณะผู้จัดทำขอแบ่งประเภทของรถยนต์เป็น 3 ประเภท ดังนี้

2.2.1.1 รถเก๋ง

รถเก๋ง (sedan หรือ saloon) เป็นรถยนต์ชนิดหนึ่งสำหรับนั่งส่วนบุคคล มีเครื่องยนต์อยู่หน้ารถ หลังรถมีกระโปรงเก็บของ มีที่นั่งสำหรับผู้โดยสาร 4 ที่นั่งหรือมากกว่า โดยอาจมีประตู 4 หรือ 2 แห่ง หลังคารถเป็นส่วนหนึ่งของตัวรถ ไม่สามารถถอดออกหรือเปิดประทุนรถได้

2.2.1.2 SUV

รถอเนกประสงค์ประเภท SUV นั้น ย่อมาจากคำว่า Sport Utility Vehicle นั่นก็คือ รถที่มีอรรถประโยชน์มากกว่ารถยนต์ทั่วไป แต่ยังคงความสวยงาม หรือพื้นฐานในแบบรถยนต์นั่งทั่วไป โดยอรรถประโยชน์ที่กล่าวมานั้น อาจจะเป็นความสมบุกสมบันที่สามารถลุยได้มากกว่ารถยนต์ทั่วไป หรือสามารถบรรทุกสัมภาระในการเดินทางได้มากขึ้น รวมไปถึงเทคโนโลยีสิ่งอำนวยความสะดวกที่ครบครัน ซึ่งปัจจุบันกระแสของรถประเภท SUV กำลังได้รับความนิยมเป็นอย่างมาก โดยหากจะพูดถึง SUV ก็คงเทียบได้กับรถ Honda CR-V, MG GS, Mazda CX-5, Nissan X-Trail ฯลฯ

2.2.1.3 รถกระบะ

รถกระบะตามความหมายของพจนานุกรมราชบัณฑิตยสถานได้ให้ความหมายไว้ว่า รถยนต์ชนิดหนึ่ง มีห้องคนขับอยู่ส่วนหน้ารถ ด้านหลังทำเป็นกระบะบรรทุกของ

2.3 การทำนาย (Forecasting)

การทำนายหรือการพยากรณ์ หมายถึง การคาดคะเน หรือการทำนายการเกิดเหตุการณ์หรือสภาพการณ์เกิดเหตุการณ์ต่าง ๆ ที่อาจเกิดขึ้นในอนาคต เช่น การพยากรณ์ยอดขายเพื่อประมาณความต้องการวัตถุดิบหรือเพื่อวางแผนการส่งเสริมทางการตลาด การพยากรณ์อัตราดอกเบี้ยเพื่อบริหารเงินสดหมุนเวียนในองค์กร เป็นต้น โดยการพยากรณ์จะทำการศึกษาแนวโน้มและรูปแบบการเกิดเหตุการณ์จากข้อมูลในอดีตและ/หรือใช้ความรู้ ความสามารถ ประสบการณ์และดุลยพินิจของผู้พยากรณ์ (นิภา นิรุตติกุล, 2549)

เทคนิคการพยากรณ์ (อัจฉรา จันทร์ฉาย, 2544) สามารถแบ่งออกเป็น 2 ประเภทใหญ่ ๆ คือ

2.3.1.เทคนิคการพยากรณ์เชิงปริมาณ (Quantitative Forecasting Methods)

เทคนิคการพยากรณ์เชิงปริมาณ (Quantitative Forecasting Methods) เป็นการพยากรณ์ที่ใช้ข้อมูลในอดีตและใช้ตัวแบบทางคณิตศาสตร์หรือวิธีการทางสถิติมาใช้ในการพยากรณ์ ซึ่งสามารถแบ่งเป็น 2 กลุ่มหลัก ๆ คือ

2.3.1.1 เทคนิคอนุกรมเวลา (Time Series Techniques)

เทคนิคอนุกรมเวลา (Time Series Techniques) เป็นเทคนิคที่ใช้ข้อมูลในอดีตเพื่อพยากรณ์ในอนาคต โดยข้อมูลในอดีตจะเก็บรวบรวมเป็นวัน รายสัปดาห์ รายเดือน หรือรายปีอย่างต่อเนื่อง สามารถแบ่งเป็น 2 กลุ่มดังนี้

ก. กลุ่มค่าเฉลี่ยเคลื่อนที่ (Moving Average) เทคนิคนี้เหมาะกับข้อมูลที่มีลักษณะคงที่ไม่เปลี่ยนแปลงมากในแต่ละงวด เช่น ยอดขายของสินค้าหรือบริการที่ไม่มีอิทธิพลของฤดูกาล (seasonal) เข้ามาเกี่ยวข้อง ในบางกรณีอาจใช้เทคนิคค่าถัวเฉลี่ยแบบถ่วงน้ำหนัก ซึ่งผลรวมของน้ำหนักที่ได้ต้องมีค่าเท่ากับ 1

ข. กลุ่มเทคนิคปรับเรียบเส้นโค้ง (Smoothing Technique) เป็นเทคนิคที่เหมาะสมกับข้อมูลที่ค่อนข้างเปลี่ยนแปลง ใช้หลักการเดียวกับค่าถัวเฉลี่ยแบบเคลื่อนที่อย่างง่าย คือ ใช้ข้อมูลในอดีตมาถ่วงน้ำหนัก แต่น้ำหนักที่ถ่วงข้อมูลกับข้อมูลในอดีตไม่เท่ากัน

2.3.1.2 เทคนิคความสัมพันธ์ของข้อมูล (Causal Models)

เทคนิคความสัมพันธ์ของข้อมูล (Causal Models) เป็นเทคนิคที่เน้นความสัมพันธ์ของตัวแปรในการพยากรณ์ เช่น การวิเคราะห์การถดถอยแบบง่ายหรือการพยากรณ์เชิงเดี่ยว (Simple Regression) เช่น การหาความสัมพันธ์ของยอดขายกับค่าโฆษณา เป็นต้น และการวิเคราะห์การถดถอยแบบพหุ (Multiple Regression) เช่น การหาความสัมพันธ์ของยอดขาย งบโฆษณา และจำนวนพนักงาน เป็นต้น

2.3.2. เทคนิคการพยากรณ์เชิงคุณภาพ (Qualitative Forecasting Methods)

เทคนิคการพยากรณ์เชิงคุณภาพ (Qualitative Forecasting Methods) เป็นการพยากรณ์ที่ใช้ข้อมูลในอดีตและไม่ใช้ตัวแบบทางคณิตศาสตร์หรือสถิติในการพยากรณ์ ผู้พยากรณ์ต้องมีความรู้ ความสามารถ ประสบการณ์ และดุลยพินิจในเรื่องที่จะพยากรณ์ เช่น การสำรวจ การวิจัยตลาด เป็นต้น สามารถแบ่งออกได้เป็น 3 กลุ่มใหญ่ ๆ ดังนี้

2.3.2.1 เทคนิคที่ใช้วิจารณญาณ (Subjective Assessment Methods)

เทคนิคที่ใช้วิจารณญาณ (Subjective Assessment Methods) ใช้วิจารณญาณ ประสบการณ์ของผู้พยากรณ์ ในการพยากรณ์

2.3.2.2 วิธีการค้นหา (Exploratory)

วิธีการค้นหา (Exploratory) เริ่มจากการศึกษาสภาพแวดล้อมในปัจจุบันและพยากรณ์ว่าอะไรจะเกิดขึ้นในอนาคตและเกิดขึ้นเมื่อใด ตัวอย่างเทคนิคได้แก่

ก. Scenario Analysis เป็นเทคนิคที่มีการพัฒนาจินตนาการเกี่ยวกับอนาคตด้านต่าง ๆ โดยการกำหนดสมมติฐานและพัฒนาทางเลือกการสร้าง Scenario เป็นประโยชน์ในการวางแผนกลยุทธ์

ข. Delphi เป็นเทคนิคที่ให้ผู้บริหารแต่ละคนออกความคิดเห็นเป็นอิสระ โดยไม่ได้พบปะปรึกษาหารือแบบเผชิญหน้า แต่ใช้การออกแบบสอบถามและส่งคืนภายหลัง จากนั้นจะรวบรวมและส่งคืนให้สมาชิกประเมินคำตอบใหม่ โดยชี้ให้เห็นว่าคนส่วนมากมีความคิดเห็นอย่างไร ทำเช่นนี้จนกว่าการคาดคะเนเกิดจากความเห็นพ้องกัน

2.3.2.3 เทคนิคด้าน Normative

เทคนิคด้าน Normative เริ่มจากค้นหาจุดมุ่งหมายและทิศทางแล้วจึงค้นหาวิธีการพัฒนาและเทคโนโลยีเพื่อทำให้บรรลุจุดมุ่งหมาย ตัวอย่างเทคนิคได้แก่

ก. Relevance Trees คล้ายกับ Decision Trees เทคนิคนี้เป็นการระบุความต้องการในอนาคตและค้นหาว่าอะไรที่ต้องทำเพื่อให้บรรลุเป้าหมาย

ข. System Dynamic เป็นการวิเคราะห์ระบบโดยมีเป้าหมายในการพิจารณาความสัมพันธ์ซึ่งกันและกันของส่วนต่าง ๆ ในระบบหรือสภาวะแวดล้อมมากกว่าที่จะปรึกษาแต่ละส่วนแยกกัน

นอกจากวิธีการพยากรณ์ที่กล่าวมาข้างต้นแล้ว ยังมีวิธีการพยากรณ์อีกอย่างหนึ่ง นั่นก็คือ การพยากรณ์ด้วยเทคนิคการทำเหมืองข้อมูล (Data mining)

2.4 ความสำคัญของการทำเหมืองข้อมูล (Data mining)

ในปัจจุบันนี้ Data Mining ได้รับความสนใจเป็นอย่างมากทั้งในอุตสาหกรรมต่าง ๆ และด้านสังคม เนื่องจากว่าข้อมูลที่เกี่ยวข้องด้วยนั้นมีจำนวนมากและมีความจำเป็นที่จะต้องใช้ประโยชน์จากข้อมูลหรือองค์ความรู้เหล่านั้น โดยอาจจะนำข้อมูลมากมายเหล่านี้ไปใช้เพื่อการวิเคราะห์ข้อมูลทางการตลาด การตรวจสอบการทุจริต และการเก็บรักษาข้อมูลลูกค้าไว้เพื่อใช้ในการควบคุมการผลิตและสำรวจความต้องการของลูกค้า เป็นต้น

การทำ Data Mining เป็นผลมาจากวิวัฒนาการของเทคโนโลยีสารสนเทศ โดยมีจุดเริ่มมาจากอุตสาหกรรมระบบฐานข้อมูล โดยเริ่มจาก

1. การเก็บรวบรวมข้อมูลและการสร้างฐานข้อมูล (Data collection and database creation)

2. การจัดการข้อมูล (Data management) ซึ่งประกอบด้วย การเก็บและค้นหาข้อมูล (Data storage and retrieval) และกระบวนการประมวลผลรายการของฐานข้อมูล (Database transaction processing)

3. การวิเคราะห์ข้อมูลขั้นสูง (Advanced data analysis) จะเกี่ยวข้องกับ Data warehouse และ Data mining

ในการพัฒนาเริ่มต้นของกลไกการเก็บรวบรวมข้อมูลและการสร้างฐานข้อมูล จะทำงานเท่าที่จำเป็นก่อน ต่อมามีการพัฒนากลไกที่มีประสิทธิภาพเพื่อการเก็บ การค้นหาข้อมูล และการประมวลผลรายการ และด้วยฐานข้อมูลมีขนาดใหญ่นี้จึงมีการคิวรีหรือประมวลผลรายการเพื่อให้ได้ข้อมูลที่ต้องการ จากการดำเนินการดังกล่าวนี้จึงกลายเป็นจุดเริ่มต้นของการวิเคราะห์ข้อมูลขั้นสูงในเวลาต่อมา

ปัจจุบันนี้การพัฒนาแอพพลิเคชันทางด้านระบบฐานข้อมูล เช่น การเก็บสื่อมัลติมิเดียต่าง ๆ มีการพัฒนากันอย่างรวดเร็ว และระบบสารสนเทศบนอินเตอร์เน็ตที่เกิดขึ้นอย่างมากเช่นเดียวกัน ข้อมูลเหล่านี้มีบทบาทสำคัญกับอุตสาหกรรมข้อมูล และความก้าวหน้าทางด้านเทคโนโลยีฮาร์แวร์ก็

มีสูง ทำให้อุปกรณ์การจัดเก็บข้อมูลมีประสิทธิภาพสูงด้วย ส่งผลให้เกิดระบบฐานข้อมูลที่แตกต่างกัน ทำให้มีการจัดเก็บข้อมูลได้ในหลายรูปแบบแตกต่างกัน ทั้งระบบปฏิบัติการหรือการจัดเก็บฐานข้อมูล ซึ่งการนำข้อมูลทั้งหมดมารวมและจัดเก็บไว้ในรูปแบบเดียวกัน เรียกว่า Data warehouse เพื่อความสะดวกในการจัดการต่อไป

ดังนั้นโดยสรุปแล้วการมี Data Mining เนื่องจากสาเหตุ คือ มีข้อมูลจำนวนมากที่ถูกเก็บไว้ในฐานข้อมูลหากเก็บไว้ไม่ได้ใช้ก็จะไม่เกิดประโยชน์ ดังนั้น จึงต้องมีการดึงข้อมูลหรือเลือกข้อมูลมาใช้ประโยชน์ให้ตรงตามความต้องการ และในอดีตการค้นหาข้อมูลจากฐานข้อมูลจะขึ้นอยู่กับผู้พัฒนาระบบฐานข้อมูลเป็นหลัก โดยผู้พัฒนาจะทำการสร้างเงื่อนไขขึ้นมาตามภูมิปัญญาของผู้พัฒนา การมี Data Mining จึงเป็นเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลและหาความเป็นไปได้ของข้อมูลทั้งหมดที่เป็นประโยชน์ออกมาอย่างเป็นระบบมากยิ่งขึ้น

2.5 ความหมายของเหมืองข้อมูล (Data mining)

เหมืองข้อมูล หรือ Data mining จะเกี่ยวข้องกับการดึงหรือการขุดเจาะองค์ความรู้จากจำนวนข้อมูลที่มีมากมาย ดังนั้น การทำเหมืองข้อมูล(Data mining) คือ การค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูลแต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก โดยการทำเหมืองข้อมูลจะเหมาะกับการแก้ปัญหาบางชนิดเท่านั้น มีเทคนิคต่าง ๆ ที่ใช้ในการแก้ปัญหาอยู่หลายเทคนิค ซึ่งไม่มีเทคนิคใดสามารถแก้ปัญหาได้ทุกปัญหา ดังนั้น ความหลากหลายของเทคนิคเป็นสิ่งที่จำเป็นที่จะนำไปสู่วิธีการแก้ปัญหาที่ดีที่สุดของการทำเหมืองข้อมูล ในกระบวนการขุดเจาะหรือสกัดให้ได้องค์ความรู้มานี้จะเรียกว่า Knowledge Discovery from Data หรือ KDD ซึ่งเป็นกระบวนการสืบค้นองค์ความรู้

กระบวนการสืบค้นองค์ความรู้ ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ใหม่ โดยเป็นกระบวนการสร้างแบบจำลองหรือกฎจากข้อมูลที่มีขึ้นมา เพื่อให้เกิดความเข้าใจในความสัมพันธ์ของข้อมูลและสามารถแยกประเภททำนายข้อมูลที่เป็นข้อมูลที่มีประโยชน์ต่อความต้องการออกมาได้ ซึ่งแต่ละขั้นตอนสามารถวนกลับไปทำงานยังขั้นตอนที่ผ่านมาได้โดยมีส่วนประกอบด้วยลำดับขั้นตอนดังนี้

1. การกลั่นกรองข้อมูล (Data cleaning) เป็นการกำจัดข้อมูลขยะ ข้อมูลผิดปกติ ข้อมูลที่ไม่สมบูรณ์ (Missing value / Noising data) ออกไปก่อน เพื่อเป็นการเพิ่มคุณภาพให้ข้อมูลก่อนทำการค้นหาองค์ความรู้ เพราะว่าข้อมูลทั้งหมดที่ได้มาอาจจะยังมีข้อผิดพลาดอยู่

2. การรวบรวมข้อมูล (Data integration) เป็นการรวบรวมข้อมูลที่เป็นรูปแบบเดียวกันหรือใกล้เคียงกันจากหลาย ๆ แหล่งมารวมกันเป็นข้อมูลชุดเดียวกัน โดยจะเก็บไว้ใน Data warehouse

3. การคัดเลือกข้อมูล (Data selection) เป็นการระบุลักษณะข้อมูลที่ต้องการแล้วคัดเลือกข้อมูลที่มีความสัมพันธ์กับการวิเคราะห์ ซึ่งหลักเกณฑ์ในการเลือกจะแตกต่างกันออกไปตามวัตถุประสงค์ของการทำเหมืองข้อมูล

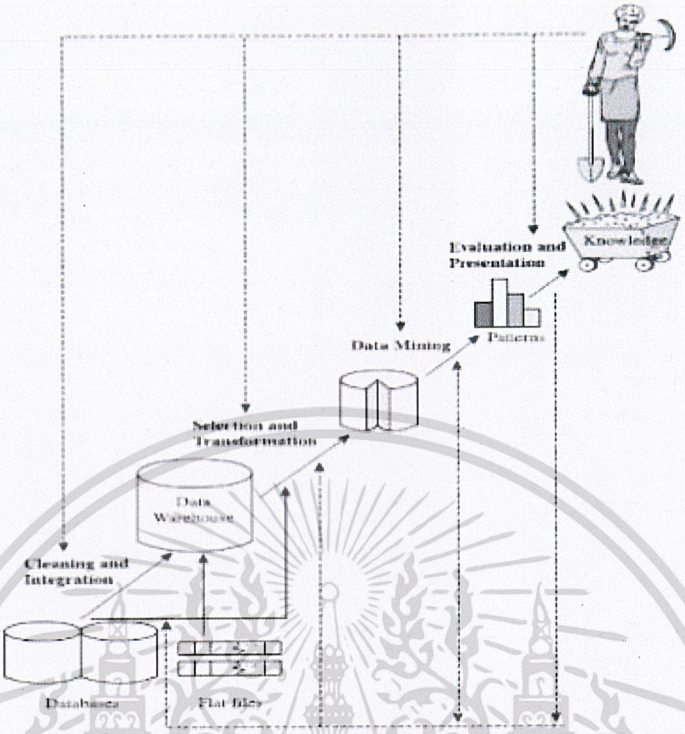
4. การแปลงรูปข้อมูล (Data transformation) เป็นการแปลงข้อมูลให้อยู่ในรูปที่พร้อมจะนำไปวิเคราะห์ตามหลักของ data mining เช่น ข้อมูลอายุ จะเป็นข้อมูลเป็นตัวเดี่ยว เช่น 17, 36, 28 อาจจะจัดเป็นกลุ่มช่วงอายุเพื่อสะดวกในการใช้งานและเกิดความเข้าใจมากขึ้น หรือเทคนิคการแปลงกลุ่มประเภทให้เป็นตัวเลขเพื่อความสะดวก เช่น การใช้รหัสแทนชื่อของสิ่งของใดๆ โดยขั้นที่ 3 และ 4 สามารถสลับลำดับกันได้ตามสมควร

ใน 4 ขั้นตอนข้างต้นนี้ อาจเรียกรวมกันได้เป็นการเตรียมข้อมูล (Data Preparation) ซึ่งเป็นขั้นตอนที่สำคัญมากและใช้เวลานานที่สุด เพราะว่าหากเกิดข้อผิดพลาด อาจทำให้ผลที่ได้ออกมาคลาดเคลื่อนหรือผิดจากจุดประสงค์ที่วางไว้ไปมาก ดังนั้น อาจถือว่าการเตรียมข้อมูล (Data Preparation) เป็นหัวใจสำคัญของงาน

5. การทำเหมืองข้อมูล (Data mining) เป็นการประมวลผลความสัมพันธ์ตามอัลกอริทึมที่ได้วางแผนเอาไว้ โดยจะนำเอาข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลมาแล้วมาใช้ในการประมวลผล โดยเมื่อทำขั้นตอนนี้แล้ว อาจมีการย้อนกลับไปทำขั้นตอนการเตรียมข้อมูลใหม่อีกครั้ง ซึ่งในขั้นตอนนี้มีรูปแบบของอัลกอริทึมหลายแบบ

6. การประเมินรูปแบบหรือกฎที่ใช้ได้ (Pattern evaluation) เป็นขั้นตอนการวิเคราะห์และประเมินผลข้อมูลที่ได้จากการทำ mining ว่าได้ตรงตามความต้องการที่ตั้งเอาไว้หรือไม่ ข้อมูลที่ได้จะน่าสนใจแค่ไหน เพื่อนำไปใช้หรือแก้ไขต่อไป

7. การนำเสนอองค์ความรู้ (Knowledge representation) การนำข้อมูลที่ผ่านกระบวนการขั้นต้นแล้ว ไปนำเสนอในรูปแบบต่างๆ ที่เข้าใจง่ายและนำเอาไปประยุกต์ใช้งานจริง หากว่าได้ผลลัพธ์เป็นที่น่าพึงพอใจตามจุดประสงค์ที่ตั้งไว้



รูปที่ 2.1 Data mining as a step in the process of knowledge discovery

ที่มา: slideplayer.com/slide/1372480

จะเห็นว่า Data mining เป็นหนึ่งในกระบวนการสืบค้นองค์ความรู้ อย่างไรก็ตาม ในอุตสาหกรรมและในงานวิจัยทางด้านฐานข้อมูล เทอมของ Data mining กำลังกลายเป็นที่น่าสนใจกว่าการค้นพบความรู้ (Knowledge discovery) จากข้อมูล เนื่องจากมีการนำเอาหน้าที่การทำงานของ data mining มาใช้กันในมุมมองที่กว้างขึ้น

Data mining คือ กระบวนการของการค้นพบองค์ความรู้ที่น่าสนใจจากจำนวนข้อมูลที่มีมากมายที่ถูกเก็บไว้ในฐานข้อมูล data warehouses หรือที่เก็บข้อมูลอื่น ๆ ดังนั้น โดยสรุปแล้ว กระบวนการ data mining นี้ เปรียบเสมือนการทำเหมืองแร่ที่ใช้เครื่องจักรคัดแยกแร่ที่เป็นที่ต้องการออกจากกองหิน กรวด ดินที่ปะปนมากับสายแร่ เพียงแต่ในกระบวนการ data mining สิ่งที่ได้จากกองข้อมูลมหาศาล คือ ความรู้ (knowledge) ที่ซ่อนอยู่ในกองข้อมูล ความรู้นี้จะช่วยให้เข้าใจลักษณะของข้อมูลและเข้าใจปัจจัยที่ทำให้เกิดลักษณะบางอย่างขึ้นในข้อมูลบางกลุ่ม ซึ่งจะช่วยให้สามารถทำนายแนวโน้มของข้อมูลใหม่ที่จะเกิดขึ้นในอนาคตได้ รวมถึงเข้าใจความสัมพันธ์ที่เชื่อมโยงข้อมูลแต่ละกลุ่มย่อยเข้าด้วยกัน ซึ่งจากกระบวนการสืบค้นความรู้ สามารถสรุปเป็นกระบวนการทำ data mining ได้ 4 ขั้นตอนดังนี้

ขั้นตอนที่ 1 เตรียมข้อมูล (data preparation) : ถ้าข้อมูลไม่อยู่ในรูปแบบที่ถูกต้องหรือเหมาะสม จะต้องมีการปรับข้อมูลให้อยู่ในรูปแบบที่โปรแกรม data mining จะเรียก

ใช้งานได้ เช่น การแทนค่าของข้อมูลที่ขาดหายไปด้วยค่าของข้อมูลที่ปรากฏบ่อยที่สุด หรือทำการแทนด้วยค่าของข้อมูลที่มีค่าเชิงสถิติสูงที่สุด เป็นต้น

ขั้นตอนที่ 2 ลดขนาดของข้อมูล (data reduction) : การจะหาโมเดลหรือ pattern ที่ข้อมูลส่วนใหญ่แสดงลักษณะเหล่านั้นออกมาเหมือนกัน จำเป็นต้องใช้ข้อมูลตัวอย่างจำนวนมาก ถ้าข้อมูลน้อยเกินไปอาจจะหาลักษณะร่วมเหล่านั้นไม่พบ แต่ในทางตรงกันข้าม ถ้าข้อมูลมีปริมาณมากเกินไป การค้นหาโมเดลหรือแพทเทิร์นจากกลุ่มข้อมูลขนาดใหญ่ต้องใช้เวลามาก ซึ่งถ้าลดจำนวนข้อมูลลงด้วยสัดส่วนที่ถูกต้อง โมเดลที่ได้ยังคงเป็นเช่นเดิมในขณะที่โปรแกรมใช้เวลาในการค้นหาโมเดลสั้นลง

การลดขนาดของข้อมูลทำได้ในสองลักษณะ คือ ลดจำนวนเรคคอร์ดและลดจำนวนแอททริบิวต์ของแต่ละเรคคอร์ด ข้อมูลที่ผ่านการลดขนาดแล้วจะถูกแบ่งออกเป็นสองส่วน คือ ส่วนแรกใช้ในกระบวนการค้นหาแพทเทิร์นหรือความสัมพันธ์จากข้อมูล เรียกข้อมูลส่วนนี้ว่า training set ส่วนที่สองใช้ตรวจสอบความถูกต้องของแพทเทิร์น เรียกข้อมูลส่วนนี้ว่า test set

ขั้นตอนที่ 3 ค้นหาโมเดลหรือความสัมพันธ์จากข้อมูล (data modeling/discovery) : กระบวนการค้นหาโมเดลหรือความสัมพันธ์จะเริ่มจากข้อมูลเริ่มต้นจำนวนไม่มากนัก จากนั้นนำผลที่ได้จากกระบวนการค้นหา (learning process/method) ไปยืนยันกับข้อมูลทดสอบ ถ้าผลที่ได้ยังไม่น่าพอใจอาจจะต้องปรับค่าพารามิเตอร์บางตัวของ learning method และเริ่มกระบวนการค้นหาใหม่กับข้อมูลจำนวนมากขึ้น จนกว่าผลที่ได้มีความถูกต้องอยู่ในระดับที่ยอมรับได้จึงจะจบกระบวนการค้นหา

ขั้นตอนที่ 4 ตรวจสอบและวิเคราะห์ผล (solution analyses) : โมเดลหรือความสัมพันธ์ที่หามาได้ในขั้นตอนที่ 3 จะต้องถูกนำมาทดสอบอัตราความผิดพลาดและวิเคราะห์ความซับซ้อนของรูปแบบโมเดล ถ้าอัตราความผิดพลาดยังสูงเกินไป อาจจะต้องย้อนกลับไป ที่ขั้นตอนที่ 3 อีกครั้ง เพื่อปรับปรุงโมเดลให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกัน ถ้าโมเดลที่หามาได้มีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจจะต้องย้อนกระบวนการกลับไปที่ขั้นตอนที่ 3 เพื่อให้หาโมเดลใหม่ที่มีความถูกต้องเท่าเดิม แต่มีรูปแบบที่ซับซ้อนน้อยลง

สำหรับสถาปัตยกรรมของระบบ Data mining โดยทั่วไปอาจจะมีส่วนประกอบหลักซึ่งประกอบด้วย

Database, Data warehouse, World Wide Web หรือที่เก็บข้อมูลอื่น ๆ เป็นแหล่งข้อมูลสำหรับการทำเหมืองข้อมูลหนึ่งที่มีกระบวนการกลั่นกรองข้อมูล (Data cleaning) และการรวบรวมข้อมูล (Data integration)

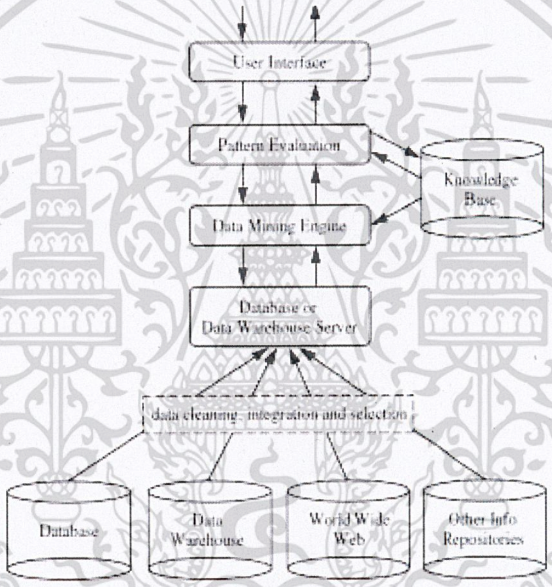
Database หรือ data warehouse server ทำหน้าที่ในการนำข้อมูลเข้าตามความต้องการของผู้ใช้งาน

Knowledge base เป็นความรู้เฉพาะด้านในงานที่ทำ ซึ่งจะเป็นประโยชน์ต่อการ ค้นหาหรือประเมินความน่าสนใจของรูปแบบผลลัพธ์ที่ได้

Data mining engine เป็นสิ่งที่จำเป็นต่อระบบ data mining ประกอบไปด้วยโมดูลการทำงาน เช่น การอธิบายลักษณะ การวิเคราะห์หาความสัมพันธ์ การจำแนก การทำนาย การวิเคราะห์การจัดกลุ่มและการวิเคราะห์การประเมิน

Pattern evaluation module เป็นส่วนประกอบที่ใช้โดยทั่วไปในการวัดสิ่งที่เกิดขึ้นโดยรวมและโต้ตอบกับ Data mining engine ซึ่งจะเน้นไปที่การค้นหา pattern ที่น่าสนใจ อาจจะใช้เกณฑ์วัดเพื่อกรองให้ได้ pattern ที่ค้นหา

User interface เป็นส่วนติดต่อประสานระหว่างผู้ใช้งานกับระบบการทำเหมืองข้อมูล



รูปที่ 2.2 Architecture of a typical data mining system

ที่มา: www.ques10.com/p/142/give-the-architecture-of-typical-data-mining-sys-1

2.6 ประเภทของข้อมูลที่สามารถใช้ในการทำเหมืองข้อมูล (Data mining)

โดยหลักแล้ว data mining สามารถที่จะประยุกต์ใช้กับที่เก็บข้อมูลประเภทใด ๆ ก็ได้ ดังนั้นในหัวข้อนี้จะยกตัวอย่างแหล่งที่เก็บข้อมูลคือ Relational database, data warehouses, Transaction databases และนอกจากนี้ยังมีที่เก็บข้อมูลจากแหล่งอื่น ๆ อีก เช่น Advanced database systems, Flat files, data streams, และ World Wide Web

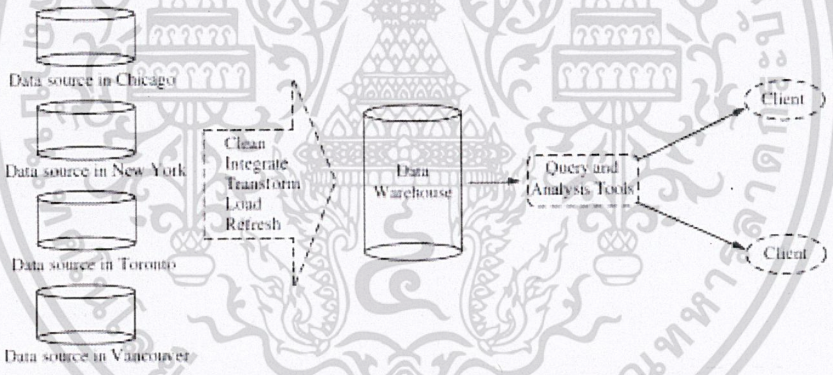
2.6.1 Relational Databases

ฐานข้อมูล (Database) เป็นกลุ่มของข้อมูลที่สัมพันธ์กัน โดยมีระบบการจัดการฐานข้อมูล (Database Management System : DBMS) ซึ่งเป็นซอฟต์แวร์ที่ทำหน้าที่ในการจัดการและควบคุมการเข้าถึงข้อมูลในฐานข้อมูล

ฐานข้อมูลแบบเชิงสัมพันธ์ (Relational database) เป็นฐานข้อมูลที่ออกแบบมาให้มีความยืดหยุ่นในการจัดการหรือใช้งานมากขึ้น โดยที่ข้อมูลจะถูกเก็บในรูปแบบของตารางที่ประกอบด้วยแถว (Row) และคอลัมน์ (Column) และแต่ละตารางก็จะสัมพันธ์เชื่อมโยงกัน เรียกความสัมพันธ์นี้ว่า รีเลชั่น (Relation)

2.6.2 Data Warehouses

Data warehouse คือ ที่เก็บข้อมูลจากหลายแหล่ง ซึ่งถูกเก็บภายใต้โครงสร้างเดียวกัน และปกติจะอยู่ในไซต์เดียว Data warehouse จะถูกสร้างด้วยกระบวนการ Data cleaning, Data integration, Data transformation, Data loading, และช่วงระยะเวลาในการ refresh ข้อมูล (การทำข้อมูลให้เป็นปัจจุบัน)



รูปที่ 2.3 โครงสร้างและการใช้งานของ Data warehouse

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

สิ่งสำคัญที่จะต้องทำในการทำ Data Mining ก็คือ การกำหนดข้อมูลที่เหมาะสม ในการ mining ดังนั้น Data mining จึงต้องการแหล่งข้อมูลที่มีการจัดเก็บและรวบรวมข้อมูลไว้อย่างดีและมีความมั่นคง เหตุผลที่ต้องมี Data warehouse ที่มีการจัดเก็บข้อมูลที่ดีสำหรับเตรียมข้อมูลเพื่อทำการ mining ก็คือ

Data warehouse จะเตรียมการจัดเก็บข้อมูลที่มีความมั่นคงและพร้อมใช้งานซึ่งเป็นสิ่งที่จำเป็นสำหรับการ mining ที่ต้องการความแน่ใจในความถูกต้องแม่นยำของข้อมูลเพื่อใช้สำหรับสร้างโมเดล และ Data warehouse ยังเป็นประโยชน์สำหรับการ mining ข้อมูลจากแหล่งข้อมูลหลาย ๆ แหล่งที่ค้นพบมากมายเท่าที่จะเป็นไปได้ ซึ่ง

Data warehouse จะบรรจุข้อมูลจากแหล่งข้อมูลเหล่านั้น ในการเลือกส่วนย่อย ๆ ของ record และ fields ที่ตรงประเด็นเพื่อการทำ Data mining และยังจำเป็นต้องใช้การ query ข้อมูลจาก Data warehouse การศึกษาผลที่ได้จากการทำ Data mining จะเป็นประโยชน์อย่างมาก โดยเฉพาะสำหรับการทำงานในอนาคตซึ่งมี Data warehouse เป็นแหล่งจัดเก็บข้อมูลที่ผ่านมาหรือข้อมูลในอดีต

ดังนั้น Data mining และ Data warehouse จะเป็นสิ่งคู่กัน โดยทั่วไปแล้วผู้ขายจำนวนมากจะหาวิธีที่จะนำเทคโนโลยี Data mining และ Data warehouse มารวมกัน

2.6.3 Transaction Database

Transaction Database เป็นฐานข้อมูลที่เก็บข้อมูลที่แทนเหตุการณ์ในขณะใดขณะหนึ่ง เช่น เก็บข้อมูลใบเสร็จรับเงิน เป็นต้น โดยทั่วไปแล้ว Transaction database จะประกอบด้วยไฟล์ โดยที่แต่ละเรคคอร์ดจะแทนการประมวลผลรายการหนึ่งๆ (Transaction) โดย Transaction หนึ่ง ๆ จะประกอบด้วยหมายเลข Transaction ที่เป็นหนึ่งเดียว (Tran_ID) และลิสต์ของ item ที่สร้าง Transaction ขึ้นมา เช่น item ของการสั่งซื้อใน store ตัวอย่าง**รูปที่ 2.4** ส่วนของ Transaction database สำหรับการขายของบริษัท AllElectronics

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

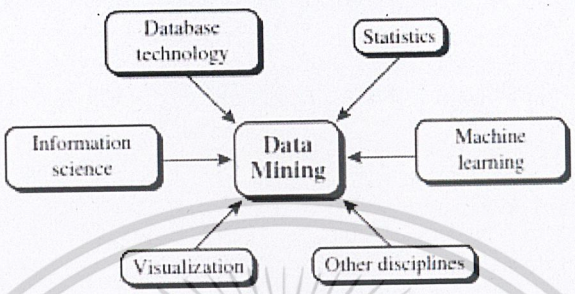
รูปที่ 2.4 ส่วนของ Transaction database สำหรับการขายของบริษัท AllElectronics

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

2.7 การจำแนกประเภทของระบบ Data mining (Classification of Data Mining Systems)

Data mining เป็นศาสตร์ที่สามารถประยุกต์ใช้ความรู้ในสาขาต่าง ๆ เข้าไว้ด้วยกัน ซึ่งประกอบด้วย Database systems, Statistics, Machine learning (การเรียนรู้ของเครื่อง) Visualization (การสร้างแบบจำลองให้เห็นภาพ) และ Information science นอกจากนี้ยังขึ้นอยู่กับวิธีในการทำ data mining ที่จะนำเทคนิคจากสาขาต่าง ๆ มาใช้ประยุกต์ เช่น นิวรอลเน็ตเวิร์ค (Neural networks) ฟัซซี่หรือราฟเซต (Fuzzy and/or rough set theory) การแทนองค์ความรู้ (Knowledge representation) เป็นต้น ทั้งนี้ขึ้นอยู่กับชนิดข้อมูลที่ต้องการจะทำ mining ซึ่งอาจจะนำไปรวมเทคนิคอื่น ๆ ได้อีก เช่น เทคนิคทางด้านการสืบค้นข้อมูล (Information Retrieval),

Pattern recognition, Image analysis, Signal processing, Computer graphics, Web technology, Economics (เศรษฐศาสตร์), Business, Bioinformatic (ชีวสารสนเทศ) หรือ psychology (จิตวิทยา)



รูปที่ 2.5 Data mining เป็นจุดรวมของสาขาวิชาต่างๆ

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

เนื่องจากสาขาวิชาต่าง ๆ เอื้อต่อการทำ data mining ทำให้งานวิจัยทางด้านนี้สร้างระบบ data mining ที่มีความหลากหลายจำนวนมากขึ้น ดังนั้น จึงมีการจำแนกระบบ data mining ซึ่งจะช่วยให้ผู้ใช้สามารถใช้งานได้ตรงตามความต้องการ

2.7.1 เกณฑ์ในการแบ่งชนิดของ Data mining

2.7.1.1 การแบ่งตามชนิดของฐานข้อมูล (kinds of databases)

การแบ่งตามชนิดของฐานข้อมูล (kinds of databases) เป็นการแบ่งตามชนิดของฐานข้อมูล (kinds of databases) เป็นการแบ่งตามระบบฐานข้อมูลที่จัดเก็บ เช่น แบ่งด้วย data model หรือประเภทของข้อมูลหรือแอพลิเคชั่นที่เกี่ยวข้อง ตัวอย่างเช่น ถ้าแบ่งตาม data model อาจจะมี Relational, Transactional, Object-relational, หรือ Data warehouse mining system เป็นต้น

2.7.1.2 การแบ่งตามชนิดขององค์ความรู้

การแบ่งตามชนิดขององค์ความรู้ เป็นการแบ่งโดยพิจารณาลักษณะงานหรือหน้าที่ที่ได้จากการ mining เช่น งานที่เกี่ยวข้องกับการอธิบายลักษณะงานที่เกี่ยวข้องกับการวิเคราะห์ความสัมพันธ์ งานที่เกี่ยวข้องกับการจำแนก การทำนาย การจัดกลุ่ม เป็นต้น

2.7.1.3 การแบ่งตามชนิดของเทคนิคที่นำมาใช้

การแบ่งตามชนิดของเทคนิคที่นำมาใช้ เป็นการแบ่งโดยพิจารณาจากเทคนิคที่นำมาใช้ หรือวิธีในการนำมาวิเคราะห์ข้อมูล เช่น Machine learning,

Statistics, Visualization, Pattern recognition, Neural networks เป็นต้น นอกจากนี้การแบ่งประเภทนี้ก็จะรวมวิธีที่มีการนำเทคนิคหรือวิธีการในการ mining มารวมกัน

2.7.1.4. การแบ่งตามแอพลิเคชั่นที่สร้างขึ้น

การแบ่งตามแอพลิเคชั่นที่สร้างขึ้น เป็นการแบ่งตามแอพลิเคชั่นที่ใช้สร้าง Data mining

2.8 เทคนิคต่าง ๆ ของเหมืองข้อมูล (Data mining)

การแก้ปัญหาของงานชนิดต่าง ๆ โดยใช้วิธี Data mining ในแต่ละงานก็จะมีเทคนิคที่จะนำมาใช้ได้อย่างเหมาะสม โดยเทคนิคที่นำมาใช้ในปัจจุบันนั้นมีมากมาย ส่วนใหญ่มาจากศาสตร์ทาง AI (Artificial Intelligence) หรือศาสตร์อื่นๆ ซึ่งจะขอยกตัวอย่างของเทคนิคที่ถูกใช้กันค่อนข้างแพร่หลาย ตัวอย่างเช่น

1. Association rule Discovery
2. Classification & Prediction
3. Database clustering หรือ Segmentation
4. Deviation Detection
5. จีเนติก อัลกอริทึม (Genetic Algorithms : GA)

2.8.1 Association rule Discovery

Association rule Discovery เป็นเทคนิคหนึ่งของ Data Mining ที่สำคัญและสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่าง ๆ หลักการทำงานของวิธีนี้ คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่าง ๆ หรือมากจากการวิเคราะห์การซื้อสินค้าของลูกค้า เรียกว่า “Market Basket Analysis” ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

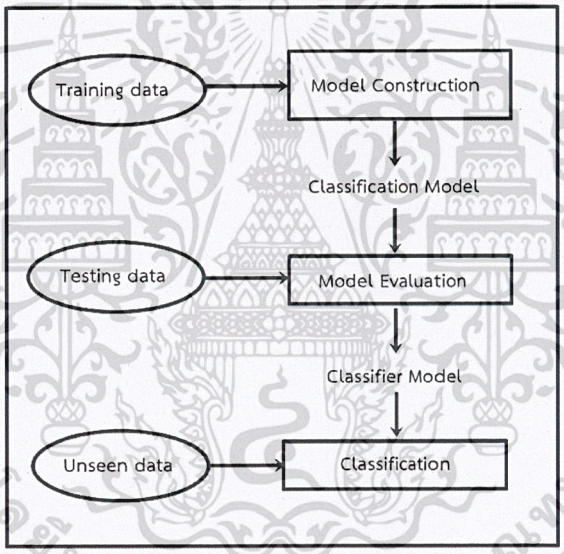
ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่ง ๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน ตัวอย่างเช่น

buys(x , database) -> buys(x , data mining) [80%,60%]

หมายความว่า เมื่อซื้อหนังสือ database แล้วมีโอกาสที่จะซื้อหนังสือ data mining ด้วย 60% และมีการซื้อทั้งหนังสือ database และหนังสือ data mining พร้อมๆกัน 80 %

2.8.2 Classification & Prediction

Classification & Prediction เป็นกระบวนการสร้าง model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ classification นี้แบ่งออกเป็น 3 ขั้นตอน



รูปที่ 2.6 กระบวนการ Classification

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

Model Construction (Learning) เป็นขั้นการสร้างโมเดลจากการเรียนรู้ข้อมูลที่มีการกำหนดคลาสไว้เรียบร้อยแล้ว (Training data) ซึ่งโมเดลที่ได้อาจแสดงในรูปของแบบต้นไม้ (Decision Tree) หรือแบบนิวรอลเน็ต (Neural Net) หรืออื่น ๆ ขึ้นกับว่าจะนำเทคนิคใดมาใช้

Model Evaluation (Accuracy) เป็นขั้นการประเมินความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (Testing data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จากการทำนายจาก model เพื่อทดสอบความถูกต้อง

Model Usage (Classification) เป็นการนำโมเดลที่ได้ไปใช้เพื่อแก้ปัญหาในการจำแนกข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen data) โดยจะทำการทำนายกลุ่มให้กับข้อมูลที่ไม่เคยเห็นมาก่อนนี้

2.8.3 Database clustering หรือ Segmentation

Database clustering หรือ Segmentation เป็นกระบวนการสร้าง model เพื่อแบ่งข้อมูลเป็นแบบกลุ่ม โดยที่ไม่รู้ล่วงหน้าว่าจะมีทั้งหมดกี่กลุ่ม โดยการจัดกลุ่มข้อมูลดังกล่าวได้จากการพิจารณาคุณสมบัติในหลาย ๆ มิติของข้อมูล ถ้ารายการในข้อมูลมีลักษณะคล้ายคลึงกันจัดเป็นกลุ่มเดียวกันได้ เช่น เมื่อเราพิจารณาจากการกระจายข้อมูล เราจะเห็นได้ว่า หากพิจารณาการกระจายข้อมูลใน 2 มิติ ข้อมูลนั้นควรแบ่งได้เป็น 3 กลุ่ม ความรู้ใหม่ที่ได้คือในสเปซของข้อมูลทั้งหมด จะมีกลุ่มที่ต่างกันเพียง 3 กลุ่มเท่านั้น

ยกตัวอย่างการนำไปใช้ บริษัทผลิตรถยนต์แห่งหนึ่งมีข้อมูลที่ผ่านมาเกี่ยวกับข้อมูลเงินเดือนลูกค้า ซึ่งมีตัวเลขเงินเดือนที่แตกต่างกันออกไป เราไม่สามารถกำหนดได้ว่ากลุ่มลูกค้ามีรายได้ในระดับใด มีระดับต่ำหรือระดับสูง จะใช้ค่าใดเป็นตัวแยกระดับ

รูปที่ 2.7 ตัวอย่างการ cluster ข้อมูล

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

2.8.4 การตรวจหาค่าความเบี่ยงเบน (Deviation Detection)

การตรวจหาค่าความเบี่ยงเบน (Deviation Detection) เป็นกรรมวิธีในการหาค่าของข้อมูลที่แตกต่างไปจากมาตรฐานหรือค่าที่คาดคิดไว้ว่ามีความแตกต่างมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติหรือการแสดงให้เห็นภาพ (Visualization) สำหรับตัวอย่างที่นำวิธีการดังกล่าวนี้ไปใช้งานการตรวจสอบลายเซ็นปลอมหรือบัตรเครดิตปลอม รวมทั้งหารตรวจหาจุดบกพร่องในชิ้นงานในโรงงานอุตสาหกรรม

2.8.5 จีเนติกอัลกอริทึม (Genetic Algorithms : GA)

จีเนติกอัลกอริทึม เป็นทฤษฎีที่จำลองกระบวนการวิวัฒนาการทางธรรมชาติ คือ การคัดเลือกทางธรรมชาติและอาศัยพื้นฐานความคิดทางพันธุกรรมในการถ่ายทอดลักษณะ

ต่าง ๆ ไปยังรุ่นถัดไป ที่สามารถนำมาพัฒนาใช้ในการหาคำตอบที่เหมาะสมที่สุดของแต่ละปัญหา จีเนติกอัลกอริทึมเป็นวิธีการหาคำตอบโดยการพิจารณา และดำเนินการจากกลุ่มของคำตอบของปัญหาที่ถูกสร้างขึ้นมาโดยการเข้ารหัส คือ การแปลงค่าตัวแปรหรือพารามิเตอร์ของปัญหาให้อยู่ในรูปโครงสร้างของโครโมโซม (Cromosomes) ที่กำหนด เพื่อคัดเลือกโครโมโซมคำตอบที่เหมาะสมสำหรับสร้างวิวัฒนาการของคำตอบให้ดีขึ้นตามกระบวนการทางพันธุศาสตร์ โดยการแลกเปลี่ยนค่าพารามิเตอร์ต่าง ๆ ระหว่างโครโมโซมที่ถูกคัดเลือกอันจะทำให้คำตอบของปัญหาถูกปรับปรุงให้ดีขึ้น

2.9 การเปรียบเทียบประสิทธิภาพวิธีในการจำแนกและทำนายข้อมูล

ในการเปรียบเทียบประสิทธิภาพของวิธีในการจำแนกและทำนายข้อมูลเราจะทำการประยุกต์ใช้เกณฑ์ ดังต่อไปนี้

1. ความถูกต้อง (Accuracy) : จะเกี่ยวข้องกับความสามารถของตัวจำแนกข้อมูลที่ถูกสร้างขึ้น ที่จะสามารถจำแนกข้อมูลที่ไม่เคยพบเจอมาก่อนได้อย่างถูกต้อง โดยในการวัดความถูกต้องอาจประเมินได้จากการใช้ชุดข้อมูลหนึ่ง ๆ (หรือมากกว่าหนึ่งชุดก็ได้) ที่แยกจากชุดข้อมูลเรียนรู้ (training dataset)

2. ความเร็ว (Speed) : จะเกี่ยวข้องกับเวลาที่ใช้ในการคำนวณทั้งในส่วนของการสร้างตัวจำแนกข้อมูลและการจำแนก/ทำนายข้อมูล

3. ความทนทาน (Robustness) : จะเกี่ยวข้องกับความสามารถของตัวจำแนกหรือตัวทำนายข้อมูลที่จะทำการทำนายได้อย่างถูกต้องจากข้อมูลตั้งต้นที่มีสิ่งรบกวนหรือมีการขาดหายไปของข้อมูล

4. ความยืดหยุ่นต่อปริมาณข้อมูล (Scalability) : จะเกี่ยวข้องกับความสามารถในการสร้างตัวจำแนกข้อมูลหรือตัวทำนายข้อมูลได้อย่างมีประสิทธิภาพ เมื่อมีข้อมูลที่ต้องพิจารณาเป็นปริมาณมาก

5. ความสามารถในการเข้าใจ (Interpretability) : เกี่ยวเนื่องกับระดับความสามารถที่จะถูกเข้าใจในตัวจำแนกหรือทำนายข้อมูลจากผู้ใช้งาน

2.10 ข้อดีข้อเสียของเหมืองข้อมูล (Data mining)

2.10.1 ข้อดีของเหมืองข้อมูล (Data mining)

1. ค้นหาข้อมูลที่สำคัญและเป็นที่ต้องการที่แอบซ่อนในข้อมูลขนาดใหญ่ได้ อาจกล่าวได้ว่า สามารถงมเข็มในมหาสมุทรได้

2. สามารถตอบสนองต่อการปรับเปลี่ยนกลยุทธ์ได้อย่างรวดเร็ว

3. ค่าใช้จ่ายน้อย เก็บข้อมูลได้มาก

4. มีการใช้ Data mining อย่างแพร่หลายทั้งภาครัฐและเอกชน

2.10.2 ข้อเสียของเหมืองข้อมูล (Data mining)

Data Mining เป็นเพียงเครื่องมือที่ใช้ในการวิเคราะห์เท่านั้น ไม่สามารถเข้าใจธุรกิจหรือเข้าใจข้อมูลได้ดีเท่าคน ดังนั้นผู้ใช้ Data Mining จึงจำเป็นต้องมีความรู้ความเข้าใจในข้อมูลธุรกิจ เครื่องมือ และอัลกอลิทึมได้เป็นอย่างดี อย่างไรก็ตาม Data Mining จะช่วยหารูปแบบและความสัมพันธ์ของข้อมูล แต่ไม่ได้ระบุว่า ค่าของข้อมูลจริง หรือค่าที่แสดงความสัมพันธ์เป็นจริงเสมอไป เป็นเพียงแค่การทำนายเท่านั้น ผู้ใช้ต้องทำการตัดสินใจอีกครั้ง

เป็นความเข้าใจผิดที่ว่า Data Mining จะช่วยค้นหาคำตอบโดยที่ไม่ต้องถามคำถามใด ๆ อันที่จริงแล้ว Data Mining ยังต้องการให้ผู้ใช้บอกรูปแบบของการค้นหาคำตอบด้วย อนึ่ง Data Mining ไม่ได้เข้ามาแทนที่ความชำนาญของนักวิเคราะห์ แต่จะเป็นเครื่องมือที่จะช่วยให้นักวิเคราะห์หรือนักบริหารในการต่อสู้กับคู่แข่งได้เป็นอย่างดี นอกจากนี้ ผู้บริโภคมักถูกหลอกในแง่มุมต่าง ๆ เช่น การใช้จ่ายที่ฟุ่มเฟือย ประชาชนถูกหลอกด้วยเทคนิคการตลาดในทางการเมืองได้ เช่น ในการจัดวางสินค้าหลังจากที่วิเคราะห์ข้อมูลโดยใช้ Data mining ทำให้ประชาชนมีการใช้จ่ายที่ฟุ่มเฟือยแทนที่จะไปซื้อของแค่หนึ่งอย่างกลับได้อย่างอื่นตามมาด้วย

2.11 ทฤษฎีของเบย์ (Bayes' Theorem)

ให้ X เป็น data sample ที่ไม่รู้ class และ H เป็นสมมติฐานว่า data sample X อยู่ในคลาส C ในปัญหา classification เราต้องการกำหนด $P(H|X)$ คือความน่าจะเป็นที่สมมติฐาน H มี data sample X โดย $P(H|X)$ เป็น posterior probability , posteriori ของ H ที่มีเงื่อนไขบน X Posterior probability มีสมการดังนี้

$$P(H|X) = \frac{P(X \cap H)}{P(X)}$$

โดยP $(H|X)$ คือ ค่า conditional probability หรือค่าความน่าจะเป็นที่เกิดเหตุการณ์ X ขึ้นก่อนและจะมีเหตุการณ์ H ตามมา

$P(X \cap H)$ คือ ค่า joint probability หรือค่าความน่าจะเป็นที่เหตุการณ์ X และ เหตุการณ์ H เกิดขึ้นร่วมกัน

$P(X)$ คือ ค่าความน่าจะเป็นที่เหตุการณ์ X เกิดขึ้น

ตัวอย่างเช่น ให้ data sample คือ fruits ที่อธิบายถึง color และ shape

ให้ X = red , round และ H คือสมมติฐานว่า X เป็น apple

เพราะฉะนั้น $P(H|X) = X$ เป็น apple ที่มีความ red และ round

ในทางกลับกัน $P(H)$ คือ prior probability , priori probability ของ H

ตัวอย่าง ให้ data sample ใดๆ เป็น apple โดยไม่สนใจรูปร่างของ data sample. $P(H|X)$ จะขึ้นอยู่กับข้อมูลมากกว่า $P(H)$ (prior probability) ที่จะขึ้นกับ X

โดย $P(A|B)$ คือ ความน่าจะเป็นของเหตุการณ์ A โดยมีเงื่อนไข B เช่นเดียวกันกับ $P(X|H)$: X มีเงื่อนไขบน H

X : red and round จะได้ว่า X เป็น apple จาก H คือ สมมติฐานว่า X เป็น apple

ดังนั้น $P(X|H)$: red and round มีเงื่อนไขว่าต้องเป็น apple

$P(X), P(H), P(X|H)$ สามารถประมาณด้วยข้อมูลที่ได้มา หรืออาจใช้ Bays Theorem คำนวณโดย

$$P(X|H) = \frac{P(X|H)P(H)}{P(X)}$$

2.12 Naive Bayesian Classification

การเรียนรู้แบบเบย์เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์ (Bayes's Theorem) เพื่อหาว่าสมมติฐานใดน่าจะถูกต้องที่สุดโดยใช้ความรู้ก่อนหน้า (Prior Knowledge) ได้แก่ ความน่าจะเป็นก่อนหน้าสำหรับสมมติฐานหนึ่งๆ ร่วมกับข้อมูล เช่นความน่าจะเป็นที่สังเกตได้สำหรับสมมติฐานหนึ่งๆ เพื่อหาสมมุติฐานที่ดีที่สุด การเรียนรู้แบบเบย์อาศัยหลักการของการคำนวณความน่าจะเป็นของแต่ละสมมติฐาน (ในที่นี้คือคลาสเป้าหมายหรือผลลัพธ์การทำนาย) โดยการเรียนรู้แบบเบย์เป็นการเรียนรู้เพิ่มได้เนื่องจากตัวอย่างใหม่ที่ได้มาถูกนำมาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นทำให้มีการเรียนรู้ที่เปลี่ยนไป วิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้ โดยผนวกกับความรู้เดิมที่มีซึ่งการทำนายค่า ถ้าคลาสเป้าหมายของตัวอย่างใช้ความน่าจะเป็นมากที่สุดของสมมติฐาน

Bayesian Classification เป็นการแบ่งคลาสโดยใช้หลักสถิติสามารถทำนายคลาสความน่าจะเป็นของสมาชิกได้เหมือนกับที่ความน่าจะเป็นให้กลุ่มตัวอย่างแก่แต่ละคลาส จะใช้ในองค์ความรู้เกี่ยวกับ Bayes Theorem (ทฤษฎีของเบย์)

Naive Bayesian Classifier : ตัวอย่างข้อมูลได้จาก Classification algorithm ใช้เปรียบเทียบกับ decision tree และ neural network classifier โดย Bayesian Classifier จะใช้ได้ดีใน database ขนาดใหญ่และสามารถทำงานได้เร็ว โดยมีสมมุติฐานว่า ผลกระทบที่มีต่อแอททริบิวต์ของคลาสที่ให้มาขึ้นอยู่กับค่าของแอททริบิวต์อื่น เรียกสมมติฐานนี้ว่า class conditional independence มีประโยชน์คือลดความซ้ำซ้อนในการคำนวณ

Naive Bayesian Classifier หรือ simple Bayesian classifier มีการทำงานดังนี้

1. แต่ละ data sample จะแสดงเป็นมิติ n มิติ $X = (x_1, x_2, \ldots, x_n)$ ทำให้ว่า n สร้างตัวอย่าง จาก n-attributes $[A_1, A_2, \ldots, A_n]$

2. มี m คลาส $C_1, C_2, C_3, \ldots, C_m$ ที่ไม่รู้ data sample X (ไม่รู้คลาส) Classifier (การจำแนกคลาสจะทำนายว่า X จะอยู่ใน คลาสที่มี posterior probability สูงสุดที่มีเงื่อนไขบน X)

Naive Bayesian classifier จะกำหนดให้ X (ข้อมูลที่ไม่รู้คลาส) อยู่ในคลาส C_i ก็ต่อเมื่อ

$$P(C_i|X) > P(C_j|X)\,; 1 \leq j \leq m \text{ และ } i \neq j$$

$P(C_i|X)$ ที่มีค่าสูงสุดจะเรียกว่า maximum posterion hypothesis โดยนำมาประยุกต์ใช้กับ Bay's Theorem ได้ดังนี้

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

3. $P(X)$ เป็นค่าคงที่สำหรับทุกคลาส ถ้าเราไม่รู้ class prior probability เราจะสมมติได้ว่าทุกคลาสมีค่า class prior probabilit เท่ากัน คือ $P(C_1) = P(C_2) = \ldots = P(C_m)$ โดยแต่ละ class prior probability หาได้จาก

$$P(C_i) = \frac{S_i}{S}$$

โดย S_i คือ จำนวน Training sample ในคลาส

S คือ จำนวนข้อมูลตัวอ่านทั้งหมด

4. ถ้าให้ชุดข้อมูลกับแต่ละ attribute จะใช้การคำนวณเยอะ เราจะลดการคำนวณโดยใช้ class conditional independence (ความเป็นอิสระต่อกัน) โดยค่า condition independence จะเป็นอิสระต่อกัน จะมีความสัมพันธ์ต่อกันดังนี้

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

ที่มาจากเรื่องความเป็นอิสระต่อกัน $P(A \cap B) = P(A)P(B)$

เช่น การโยนเหรียญ 2 ครั้ง ถ้าออกหัวจะไม่มีผลต่อการโยนเหรียญครั้งต่อไปโดย ความน่าจะเป็นของ $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ สามารถมาจาก training samples โดย

a. ถ้า A_kไม่มีเงื่อนไข แล้ว

$$P(x_k|C_i) = \frac{S_{ik}}{S_i}$$

โดย S_{ik} = จำนวนข้อมูลในคลาส C_i ที่มีค่า x_k สำหรับ A_k

S_i = จำนวนข้อมูลในคลาส C_i

b. ถ้า A_k เป็นค่าที่มีความต่อเนื่อง ดังนั้นแอททริบิวต์นี้จึงสามารถสมมติโดยใช้การกระจายแบบเกาส์ (Gaussian distribution) ดังนี้

$$P(x_k|C_i) = g(x_k, M_{ci}, \sigma_{ci}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x_k - \mu_{C_i})^2}{2\sigma^2 C_i}}$$

โดย $g(x_k, M_{ci}, \sigma_{ci})$ คือ ความหนาแน่นของฟังก์ชั่นเกาเซียนแบบปกติ (Gaussian (normal) density function) สำหรับแอททริบิวต์ $\mathbf{A_k}$

$\mathbf{M_{ci}}$ คือ ค่าเฉลี่ย

$\boldsymbol{\sigma_{ci}}$ คือ ส่วนเบี่ยงเบนมาตรฐาน

โดยสูตรการกระจายแบบเกาเซียนนี้จะทำให้ได้ค่าของแอททริบิวต์ $\mathbf{A_k}$ สำหรับตัวอย่างข้อมูล ในคลาส C_i

5. ในการจัดกลุ่มของคลาส (classify) ของ X ที่เป็นข้อมูลที่ไม่รู้คลาส และ $P(x|C_i)P(C_i)$ มีค่าให้ กับคลาส C_i โดย X จะอยู่ในคลาส C_i ก็ต่อเมื่อ

$$P(C_i|X) > P(C_j|X)\,; 1 \leq j \leq m \text{ และ } i \neq j$$

หรือจะพูดอีกอย่างหนึ่งว่า X จะอยู่ในคลาส C_i ที่มีค่า $P(x|C_i)P(C_i)$ มากที่สุด

2.13 ประสิทธิภาพของการจำแนกคลาสแบบเบย์

ในทางทฤษฎี Bayesian Classifier มีอัตราความผิดพลาดน้อยถ้าเทียบกับการจำแนกคลาสอื่น อย่างไรก็ตามในทางทฤษฎีนั้นมีความแตกต่างกันกับการปฏิบัติงานจริง ทั้งในทางเงื่อนไข

ความน่าจะเป็นของข้อมูล Bayesian Classifiers มีประโยชน์ในการให้เหตุผลทางทฤษฎีชัดเจนมากกว่าการจำแนกคลาสแบบอื่น ตัวอย่างเช่น ภายใต้เงื่อนไขใดๆเราสามารถแสดงผลในรูปแบบของ neural network และ curre-fitting algonrithms ที่ได้ output เป็น maximun posteriori hypothesis ที่ได้มาจาก naive Bayrsian classifier

2.14 ตัวอย่างการทำนายคลาสโดยใช้วิธี Bayesian Classifiers

ตารางที่ 2.1 ข้อมูลตัวอย่างการทำนายคลาสโดยใช้วิธี Bayesian Classifiers

ลำดับ	อายุ (Age)	Income	Student	Credit_rating	Class : buy_computer
1	≤30	hight	No	Fair	No
2	≤30	Hight	No	Excellent	No
3	31-40	Hight	No	Fair	Yes
4	≥40	Medium	No	Fair	Yes
5	≥40	Low	Yes	Fair	Yes
6	≥40	Low	Yes	Excellent	No
7	31-40	Low	Yes	Excellent	Yes
8	≤30	Medium	No	Fair	No
9	≤30	Low	Yes	Fair	Yes
10	≥40	Medium	Yes	Fair	Yes
11	≤30	Medium	Yes	Excellent	Yes
12	31-40	Medium	No	Excellent	Yes
13	31-40	Hight	Yes	Fair	Yes
14	≥40	Medium	no	Excellent	No

ที่มา: Data Mining Concepts and Techniques, 2012

การทำนายคลาสโดย Bayesian classifier เราต้องการทำนายคลาสของตัวอย่างข้อมูลที่ไม่รู้คลาส โดยใช้ข้อมูลในตารางด้านบน data sample จะอธิบายแอททริบิวต์เกี่ยวกับ age, income, student และ credit_rating มีแอททริบิวต์ class label คือ buys_computer ที่มีค่า (namely , {yes,no})

ให้ C_1 = คลาส buys_computer = “yes”

C_2 = คลาส buys_computer = “no”

โดย unknow samples มีดังนี้

X = (age = “≤30”, income = “medium”, student = “yes”, credit_rating = “fair”)

คำนวณค่า $P(x|C_i)P(C_i)$ โดยที่ i = 1,2 $P(C_i)$ โดย prior probability ของแต่ละคลาส สามารถคำนวณใน training sample ได้โดย

P(buys_computer = “yes”) = 9/14 = 0.643

P(buys_computer = “no”) = 5/14 = 0.357

เมื่อคำนวณ $P(x|C_i)$, โดยที่ i = 1,2 เราสามารถคำนวณตามเงื่อนไขได้ดังนี้

P(age =“ 30 ” | buys_computer = “yes”) = 2/9 = 0.222

P(age =“ 30 ” | buys_computer = “no”) = 3/5 = 0.600

P(income =“medium” | buys_computer = “yes”) = 4/9 = 0.444

P(income =“medium” | buys_computer = “no”) = 2/5 = 0.400

P(student =“yes” | buys_computer = “yes”) = 6/9 = 0.667

P(student =“yes” | buys_computer = “no”) = 1/5 = 0.200

P(credit_rating =“fair” | buys_computer = “yes”) = 6/9 = 0.667

P(credit_rating =“fair” | buys_computer = “no”) = 2/5 = 0.400

จาก

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)\ ;\ i = 1, 2$$

จะได้

P (X | buys_computer = “yes”) = 0.222 × 0.444 ×0.667 ×0.667 = 0.044

P (X | buys_computer = “no”) = 0.600 ×0.400 ×0.200 ×0.400 = 0.019

จาก

$$P(x|C_i)P(C_i)$$

จะได้

P(x | buy_computer = “yes”) P(x | buy_computer = “yes”) = 0.044 ×0.643 = 0.028

P(x | buy_computer = “no”) P(x | buy_computer = “no”) = 0.019 ×0.537 = 0.007

ดังนั้นการทำนายคลาส buys_computer โดยวิธี naive bayesian classifier ได้ output เป็น yes สำหรับตัวอย่าง X

2.15 ต้นไม้การตัดสินใจ (Decision Tree)

การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจจะเป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลแนบอยู่ด้วย ต้นไม้ตัดสินใจจะประกอบไปด้วยโหนดต่าง ๆ (ที่ไม่ใช่โหนดใบ-nonleaf node) ที่ซึ่งถูกใช้ในการแสดงถึงเงื่อนไขหรือแอททริบิวต์หนึ่ง ๆ ของข้อมูล โดยที่แต่ละกิ่งก้านของโหนดหนึ่ง ๆ จะหมายถึงค่าที่เป็นไปได้จากการทดสอบกับแอททริบิวต์นั้นๆ และจะประกอบไปด้วยโหนดใบ (leaf node) ที่ซึ่ง จะมีหมวดหมู่ข้อมูลจัดเก็บอยู่ โดยตัวอย่างต้นไม้ตัดสินใจที่จะแสดงการทำนาย คุณลักษณะของลูกค้าที่จะทำการซื้อคอมพิวเตอร์จากร้านขายอุปกรณ์ไฟฟ้า โดยโหนดต่างๆที่ไม่ใช่โหนดใบจะถูกแทนด้วยสี่เหลี่ยมและโหนดใบจะถูกแทนด้วยวงรีตามลำดับ จะเห็นว่าโหนดใบจะเป็นโหนดที่บ่ง บอกถึงข้อมูลหมวดหมู่ของคำตอบที่เราต้องการ อาทิเช่น “yes” หมายถึง ลูกค้าจะซื้อคอมพิวเตอร์และ “no” หมายถึงลูกค้าจะไม่ซื้อคอมพิวเตอร์โดยต้นไม้ที่ถูกสร้างขึ้นอาจเป็นต้นไม้ที่มีลักษณะเป็นไบนารี่หรืออาจจะไม่เป็นไบนารี่ก็ได้

หลังจากทำการสร้างต้นไม้ตัดสินใจแล้ว จะสามารถใช้ต้นไม้ตัดสินใจในการจำแนกข้อมูลได้ โดยจะทำการจำแนกหมวดหมู่ของข้อมูลเรคคอร์ดหนึ่ง ๆ (ที่ประกอบไปด้วยแอททริบิวต์ต่าง ๆ แต่จะไม่ทราบหมวดหมู่ข้อมูลในเรคคอร์ดนั้น ๆ) ด้วยการเปรียบเทียบแอททริบิวต์ที่อยู่ในโหนดรากกับค่าของแอททริบิวต์ในเรคคอร์ดที่พิจารณา โดยจะทำการเปรียบเทียบจากโหนดรากไปจนถึงโหนดใบ เมื่อทราบถึงโหนดใบจะทำให้ทราบถึงหมวดหมู่ข้อมูลของเรคคอร์ดที่ทำการพิจารณา

ต้นไม้ตัดสินใจเป็นแนวทางในการแสดงลำดับของกฎที่มีผลลัพธ์หรือมีค่าและเป็นการพิจารณาเทคนิคการทำเหมืองข้อมูลที่เป็นที่นิยมมากที่สุด เป็นแผนภูมิคล้ายโครงสร้างต้นไม้ประกอบด้วย โหนดภายใน โหนดใบ และกิ่ง

โหนดภายใน (internal node) แทนการตัดสินใจหรือการทดสอบกับแอททริบิวต์

กิ่ง (branch) แทนผลลัพธ์ (outcome) ของการทดสอบ คือ ค่าที่เป็นไปได้ทั้งหมดของแอททริบิวต์

โหนดใบ (leaf nodes) แทนถึง Class หรือกลุ่มของข้อมูล และโหนดที่อยู่บนสุดของต้นไม้เรียกว่า โหนดราก (root node)

การจำแนกหมวดหมู่ข้อมูลด้วยต้นไม้ตัดสินใจได้รับความนิยมเป็นอย่างมาก และถูกประยุกต์ใช้ในหลายๆงาน อาทิเช่น การผลิตและการใช้ยา(medicine) การผลิตสินค้าต่างๆ (manufacturing and production) การวิเคราะห์ทางการเงิน(Financial analysis) ดาราศาสตร์ (astronomy) อณูชีววิทยา(molecular biology) เป็นต้น สาเหตุที่ต้นไม้ตัดสินใจได้รับความนิยมอันเนื่องมากจากเหตุผลหลายประการด้วยกัน เช่น

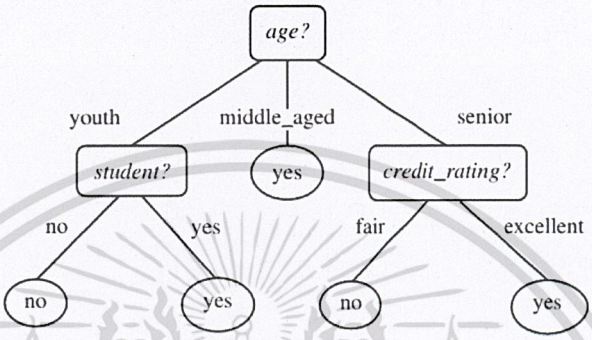
1. ไม่ต้องการองค์ความรู้ใดๆหรือการกำหนดค่าพารามิเตอร์ใดๆเพื่อที่จะทำการสร้างต้นไม้ตัดสินใจ

2. สามารถจัดการกับข้อมูลที่มีหลายๆมิติหรือข้อมูลที่มีหลายๆแอททริบิวต์ได้

3. ผลลัพธ์ที่ได้จะเป็นต้นไม้ตัดสินใจที่อยู่ในรูปแบบที่เข้าใจง่าย

4. ขั้นตอนการสร้างต้นไม้ตัดสินใจค่อนข้างง่ายและสามารถทำงานได้อย่างรวดเร็ว

5. มักจะให้ผลการจำแนกข้อมูลที่มีความถูกต้องค่อนข้างสูง ที่ซึ่งอาจขึ้นอยู่กับคุณลักษณะของข้อมูลที่เราทำการพิจารณาด้วยเช่นกัน



รูปที่ 2.8 ตัวอย่างต้นไม้ตัดสินใจสำหรับการจำแนกคุณลักษณะของลูกค้าที่ทำการซื้อคอมพิวเตอร์

ที่มา: https://z-p3-lookaside.fbs.com/file/10.WeekX

2.16 การสร้างต้นไม้ตัดสินใจ

ในช่วงปลายของยุค 1970 ได้มีนักวิจัยทางด้านการเรียนรู้ของเครื่อง (machine learning) คือ J. Ross Quinlan ได้คิดค้นอัลกอริทึมสำหรับสร้างต้นไม้ตัดสินใจที่มีชื่อว่า ID3 (Iterative Dichotomiser) ต่อมาเขาได้พัฒนาต่อยอด ID3 ไปเป็น C4.5 ซึ่งได้กลายมาเป็นอัลกอริทึมพื้นฐานที่ใช้สำหรับเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมต่างๆ ทางด้านการเรียนรู้แบบมีผู้สอน (Supervised Learning) ID3 และ C4.5 ได้ทำการประยุกต์ใช้วิธีการเชิงละโมภ (greedy approach) ในการสร้างต้นไม้ภายใต้วิธีการแบบ “top-down recursive divide-and-conquer” โดยทำการพิจารณาชุดข้อมูลสำหรับเรียนรู้ (training data, เซตของเรคคอร์ดของข้อมูลที่แต่ละเรคคอร์ดจะประกอบไปด้วยเซตของแอททริบิวต์ต่างๆและแอททริบิวต์ที่บ่งบอกถึงหมวดหมู่ของข้อมูลเรคคอร์ดนั้นๆ) ด้วยการแบ่งข้อมูลออกเป็นส่วนย่อยๆในระหว่างกระบวนการสร้างต้นไม้

ตัววัดการเลือกแอททริบิวต์ที่นิยม คือ Information Gain และ Gini Index

อัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจ (Generate_decision_tree)

อินพุตของต้นไม้ตัดสินใจ คือ ชุดข้อมูลสำหรับเรียนรู้ (D) ที่ประกอบไปด้วยเรคคอร์ดต่างๆ พร้อมกับหมวดหมู่ข้อมูล

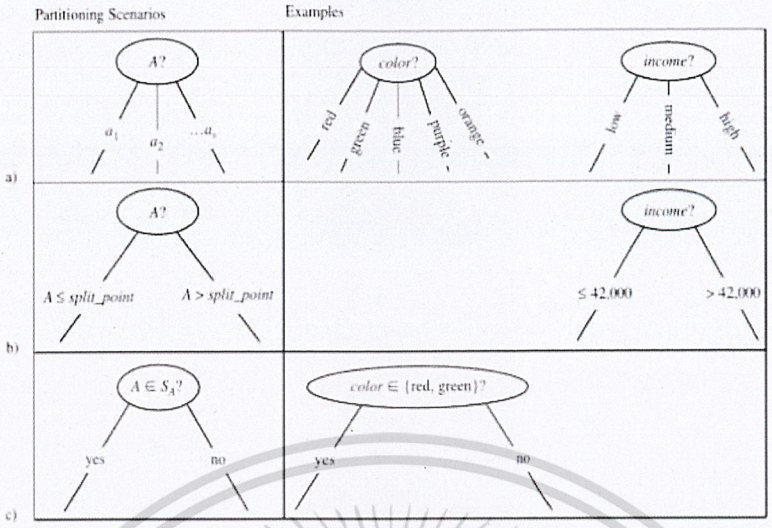
ลิสต์ของแอททริบิวต์ (attribute_list) คือ สิ่งที่อธิบายถึงคุณลักษณะของข้อมูล

วิธีการเลือกแอททริบิวต์(attribute_selection_method) เป็นวิธีการในการตัดสินว่าจะเลือกแอททริบิวต์ใดที่ดีที่สุดในการแบ่งข้อมูลออกตามหมวดหมู่ของข้อมูล

อัลกอริทึม การสร้างต้นไม้ตัดสินใจ (Generate_decision_tree)
อินพุต - ชุดข้อมูลสำหรับเรียนรู้ (D) ที่ประกอบไปด้วยเรคคอร์ดต่างๆพร้อมกับหมวดหมู่ข้อมูล
- ลิสต์ของแอทริบิว (attribute_list) ที่อธิบายถึงคุณลักษณะของข้อมูล
- วิธีการเลือกแอทริบิว (attribute_selection_method) เป็นวิธีการในการตัดสินว่าจะเลือกแอทริบิวใด (ที่ดีที่สุด) ในการแบ่งข้อมูลออกตามหมวดหมู่ของข้อมูล
เอาท์พุต ต้นไม้ตัดสินใจ
(1) สร้างโหนด N
(2) if ทุกๆเรคคอร์ดในชุดข้อมูล *D* มีหมวดหมู่ *C* เหมือนกันทั้งหมด then
(3) return N ในฐานะที่เป็นโหนดใบที่มีหมวดหมู่ C แนบอยู่
(4) if attribute_list เป็นเซตว่าง then
(5) return N ในฐานะที่เป็นโหนดใบที่มีหมวดหมู่ C (โดย C มาจากการใช้ majority vote กับหมวดหมู่ของข้อมูลที่ทำการพิจารณาทั้งหมด)
(6) ประยุกต์ใช้ Attribute_selection_method (D, attribute_list) เพื่อที่จะหาแอทริบิวที่ดีที่สุดที่จะใช้เป็นจุดแบ่งข้อมูล (find the "best" splitting_criterion)
(7) ทำการกำหนดชื่อของโหนด N ตามชื่อแอทริบิวที่ได้มาจากขั้นตอนที่ 6 (ตาม splitting_criterion)
(8) if แอทริบิวที่ได้จากขั้นตอนที่ 6 (splitting_attribute) มีค่าที่เป็นไปได้ในแอทริบิวแบบไม่ต่อเนื่อง (discrete-valued) และ attribute_selection_method ยอมให้ทำการแยกข้อมูลออกเป็นหลายๆส่วน (multi-way splits)
(9) ลบรายชื่อแอทริบิวที่ได้จากขั้นตอนที่ 6 ออกจากลิสต์ของแอทริบิว (attribute_list = attribute_list – splitting_attribute)
(10) for แต่ละค่าที่เป็นไปได้ j ของแอทริบิวจากขั้นตอนที่ 6
(11) กำหนดให้ D_j คือ เซตของเรคคอร์ดของข้อมูลใน D ที่มีค่าในแอทริบิวตรงกับค่า j (คือการแบ่งข้อมูลออกเป็นส่วนๆตามค่าที่เป็นไปได้ในแอทริบิวที่ทำการพิจารณา)
(12) if D_j เป็นเซตว่าง (ไม่มีเรคคอร์ดใดเลยใน D ที่มีค่าในแอทริบิวตรงกับค่า j)
(13) ทำการสร้างโหนดลูกให้กับโหนด N โดยโหนดที่สร้างขึ้นจะเป็นโหนดใบที่มี หมวดหมู่ C (โดย C มาจากการใช้ majority vote)
(14) else ทำการสร้างโหนดลูกให้กับโหนด N ด้วย Generate_decision_tree(Dj, attribute_list)
(15) return N

รูปที่ 2.9 อัลกอริทึมสำหรับการสร้างต้นไม้ตัดสินใจ

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction



รูปที่ 2.10 เอาต์พุทของต้นไม้ตัดสินใจ

ที่มา: https://z-p3-lookaside.fbs.com/file/1.%Introduction

2.17 ตัวชี้วัดเลือกแอททริบิวต์ของต้นไม้ตัดสินใจ

ตัวชี้วัดการเลือกแอททริบิวต์ (attribute selection measure) หรือที่รู้จักกันในนาม splitting rules คือ วิธีการ/เกณฑ์ในการเลือกแอททริบิวต์ (splitting criterion) ที่ดีที่สุดที่จะถูกใช้เพื่อแบ่งส่วนชุดข้อมูลที่จะถูกใช้เพื่อแบ่งส่วนชุดข้อมูล D ออกเป็นชุดข้อมูลย่อยๆ การแบ่งชุดข้อมูลจะทำการแบ่งตามค่าที่เกิดขึ้นในแอททริบิวต์หนึ่ง ๆ ที่ถูกเลือกมา ชุดข้อมูลย่อยหนึ่ง ๆ จะประกอบไปด้วยเซตของเรคคอร์ดที่มีหมวดหมู่เหมือนกันมากที่สุด นอกเหนือจากการแบ่งชุดข้อมูลแล้ว ตัวชี้วัดการเลือกแอททริบิวต์ยังสามารถทำการจัดลำดับของแอททริบิวต์ที่ใช้ในการอธิบายชุดข้อมูล โดยหลังจากทำการจัดลำดับแอททริบิวต์ที่มีค่าคะแนนสูงสุดจะถูกเลือกไปเป็นแอททริบิวต์ที่ใช้แบ่งข้อมูล (splitting attribute) ในกรณีที่ splitting attribute มีค่าที่เกิดขึ้นในแอททริบิวต์แบบต่อเนื่อง หรือเราถูกจำกัดให้สร้างต้นไม้ตัดสินใจแบบไบนารี่ เราจะต้องทราบถึง split point หรือ splitting subset เพื่อที่จะสามารถประมวลผลได้ในการที่จะแบ่งชุดข้อมูล เราจะใช้นิยามดังต่อไปนี้

กำหนดให้ D คือ ชุดข้อมูลสอนที่ประกอบไปด้วยเซตของเรคคอร์ด ที่ซึ่งแต่ละเรคคอร์ดจะประกอบไปด้วยค่าในแอททริบิวต์ต่างๆที่สามารถอธิบายถึงคุณลักษณะของข้อมูลเรคคอร์ดนั้นๆ รวมถึงค่าในแอททริบิวต์ที่บ่งบอกถึงหมวดหมู่ของข้อมูลในเรคคอร์ดนั้นๆ สมมติให้แอททริบิวต์ที่เป็นหมวดหมู่ของข้อมูลมีค่าที่เป็นไปได้ทั้งสิ้น m หมวดหมู่ ดังนั้น ในเรคคอร์ดหนึ่งจะมีหมวดข้อมูลเป็น $C_i \left(\text{for i} = 1, \ldots, \text{m}\right)$ กำหนดให้ $C_{i,D}$ คือ เซตของเรคคอร์ดที่อยู่ในหมวดหมู่ C_i กำหนดให้ $|C_i|$ และ $C_{i,D}$ คือ จำนวนเรคคอร์ดที่อยู่ในชุดข้อมูล D และ $C_{i,D}$ ตามลำดับ

2.18 ค่าเกนความรู้ (Information gain) และดัชนีจีนี (Gini Index)

ค่าเกนความรู้จะเป็นตัวชี้วัดการแบ่งข้อมูลออกเป็นชุดข้อมูลย่อยที่ได้รับความนิยมอย่างแพร่หลาย โดยค่านี้จะถูกประยุกต์ใช้ในอัลกอริทึม ID3 ที่ซึ่งจะทำการเลือกแอททริบิวต์สำหรับแบ่งข้อมูลจากแอททริบิวต์ที่มีค่าเกนความรู้สูงที่สุด ที่ซึ่งจะเป็นการเลือกแอททริบิวต์ที่ต้องการข้อมูลที่น้อยที่สุดในการระบุ/แบ่งข้อมูลออกเป็นชุดข้อมูลย่อย ในการดำเนินงานเริ่มต้นจะต้องเริ่มจากการคำนวณหาค่า $Info(D)$ หรือที่เรียกอีกอย่างหนึ่งว่า ค่าเอนโทรปี (entropy of D) ที่จะหมายถึงค่าเฉลี่ยของปริมาณข้อมูลที่ต้องการในการระบุถึงหมวดหมู่ของข้อมูลเรคคอร์ดหนึ่งๆในชุดข้อมูลที่จะขึ้นกับอัตราส่วนของจำนวนเรคคอร์ดที่สอดคล้องกับแต่ละหมวดหมู่ ที่ซึ่งสามารถคำนวณได้ดังนี้

$$Info(D) = \sum_{i=1}^{m} p_i log_2 p_i$$

เมื่อ p_i คือ ค่าความน่าจะเป็นที่เรคคอร์ดหนึ่งๆจะมีหมวดหมู่ของข้อมูลเป็นหมวดหมู่ C_i ซึ่งสามารถคำนวณได้จาก

$$\frac{|C_{i,D}|}{|C_i|}$$

สมมติว่าต้องการที่จะแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อยโดยใช้แอททริบิวต์ A ที่มีค่าที่เกิดขึ้นในชุดข้อมูล v ค่าที่แตกต่างกัน ($\{a_1, a_2, \dots, a_v\}$) ถ้าแอททริบิวต์ A มีค่าที่เกิดขึ้นเป็นแบบไม่ต่อเนื่อง เราจะพิจารณาการแบ่งชุดข้อมูลออกเป็น v ชุดข้อมูลย่อย โดยสามารถแบ่งได้เป็น ($\{D_1, D_2, \dots, D_v\}$) โดยที่ D_j ใดๆจะบรรจุไปด้วยเซตของเรคคอร์ดที่แอททริบิวต์ A มีค่าเป็น a_j เป็นต้น โดยแต่ละชุดข้อมูลย่อยจะสอดคล้องกับกิ่งย่อยที่แตกออกจากโหนด N ที่กำลังทำการพิจารณาอยู่ ดังนั้นในการแบ่งชุดข้อมูลย่อย จะต้องพยายามทำให้แต่ละชุดข้อมูลย่อยประกอบไปด้วยเรคคอร์ดที่มีหมวดหมู่เหมือนกันทั้งหมด แต่อย่างไรก็ตามมักจะเป็นการยากที่จะมีเรคคอร์ดที่มีหมวดหมู่เหมือนกันมีค่าในแอททริบิวต์เหมือนกัน เนื่องจากในชุดข้อมูลย่อยที่ถูกแบ่งมักจะมีเรคคอร์ดที่มีหมวดหมู่ที่แตกต่างกัน ดังนั้นจะต้องทำการพิจารณาว่าเรคคอร์ดที่มีค่าในแอททริบิวต์เหมือนกันนั้นมีความเหมือนกันของหมวดหมู่เป็นเช่นไร โดยเราจะต้องทำการคำนวณค่า $Info_A(D)$ ที่ซึ่ง หมายถึงจำนวนข้อมูลที่คาดว่าจะใช้สำหรับทำการแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย โดยทำการพิจารณาแอททริบิวต์ A สามารถคำนวณได้ดังนี้

$$Info_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

เมื่อ $\frac{|D_j|}{|D|}$ หมายถึง จำนวนเรคคอร์ดในชุดข้อมูล D ที่มีค่าในแอททริบิวต์ A เป็น a_j หารด้วยจำนวนเรคคอร์ดทั้งหมดในชุดข้อมูล D

หลังจากคำนวณค่า $Info\ (D)$และค่า $Info_A\ (D)$ ค่าเกนความรู้สำหรับการพิจารณา แอททริบิวต์ A จะสามารถคำนวณได้จากค่าความแตกต่างระหว่างปริมาณข้อมูลที่ต้องการในการระบุถึงหมวดหมู่ของข้อมูลสำหรับเรคคอร์ดหนึ่ง ๆ กับจำนวนข้อมูลที่คาดว่าจะใช้สำหรับทำการแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย โดยทำการพิจารณาแอททริบิวต์ A ที่ซึ่งสามารถแสดงการคำนวณได้ดังนี้

$$Gain(A) = Info(D) - Info_A\ (D)$$

ค่า $Gain(A)$ จะบ่งบอกปริมาณเกนความรู้ที่เราจะได้รับด้วยการพิจารณาแอททริบิวต์ A และการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยตามการพิจารณาแอททริบิวต์ A ถ้าค่า $Gain(A)$ เป็นค่าเกนความรู้ที่มีค่ามากที่สุดในบรรดาค่าเกนความรู้ของแอททริบิวต์ทั้งหมด แอททริบิวต์ A จะเป็น แอททริบิวต์ที่ดีที่สุดสำหรับการจำแนก/แบ่งข้อมูลเป็นชุดข้อมูลย่อย เนื่องจากแอททริบิวต์ A ต้องการปริมาณข้อมูลที่น้อยที่สุดที่ใช้ในการจำแนกข้อมูล $(minimalInfo_A\ (D))$

ดัชนีจีนี่ (Gini index) จะเป็นตัวชี้วัดที่จะทำการพิจารณาความไม่บริสุทธิ์ของชุดข้อมูล D ที่ซึ่งจะมีเซตของเรคคอร์ดที่มีหมวดหมู่ของข้อมูลไม่เหมือนกันอยู่ โดยเริ่มจากการคำนวนหาค่าความไม่บริสุทธิ์ของชุดข้อมูล D ดังนี้

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

เมื่อ p_i คือ ค่าความน่าจะเป็นที่เรคคอร์ดหนึ่งๆจะมีหมวดหมู่ของข้อมูลเป็นหมวดหมู่ C_i ที่ซึ่งสามารถคำนวณได้จาก $\frac{|C_{i,D}|}{|C_i|}$

ในการแบ่งชุดข้อมูลโดยใช้ดัชนีจีนี่จะสามารถแบ่งชุดข้อมูลออกเป็น 2 ชุดข้อมูลย่อยเท่านั้น เมื่อทำการพิจารณาแอททริบิวต์ A ใดๆที่มีค่าที่เกิดขึ้นเป็นแบบไม่ต่อเนื่องและมีค่าที่แตกต่างกันทั้งสิ้น v ค่า $(\{a_1, a_2, \dots, a_v\})$ เราจะต้องทำการค้นหาการแบ่งชุดข้อมูลที่ดีที่สุดภายใต้การพิจารณาแอททริบิวต์ A ด้วยการพิจารณาทุกสับเซตที่เป็นไปได้ของค่าที่เกิดขึ้นในแอททริบิวต์ A ที่จะถูกจัดอยู่ในเซตย่อย 2 เซตด้วยกัน ตัวอย่างเช่น แอททริบิวต์รายได้ที่มีค่าที่เกิดขึ้นในแอททริบิวต์ 3 ค่าด้วยกันคือ $\{low, medium, high\}$ ซึ่งจากทั้ง 3 ค่าที่เกิดขึ้น จะสามารถแบ่งค่าทั้ง 3 ออกเป็นเซตย่อย 2 เซตได้ดังต่อไปนี้

$$\{(\{low, medium\}, \{high\}), (\{low, high\}, \{medium\}), (\{medium, high\}, \{low\}),$$

$$\{(\{low\},\{medium,high\}),(\{medium\},\{low,high\},),(\{high\},\{low,medium\})\}$$

เมื่อทำการแบ่งค่าที่เกิดขึ้นออกเป็น 2 เซตย่อย เราจะนำ 2 เซตใดๆที่แบ่งไว้แล้ว มาทำการคำนวณค่าน้ำหนักรวมของค่าความไม่บริสุทธิ์ โดยในการพิจารณาแอททริบิวต์ A ใดๆ จะทำการแบ่งชุดข้อมูล D ออกเป็น 2 ชุดข้อมูลย่อย คือ D_1 และ D_2 ดังนั้น เราจะสามารถหาค่าดัชนีจีนี่เมื่อทำการพิจารณาแอททริบิวต์ A ภายใต้การพิจารณา D_1 และ D_2 ได้เป็น

$$Gini_A(D) = \frac{|D_2|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

เมื่อทำการคำนวณค่าดัชนีจีนี่สำหรับ 2 เซตย่อยใดๆภายใต้แอททริบิวต์ A 2 เซต D_1 และ D_2 ใดที่ให้ค่า $Gini_A(D)$ น้อยที่สุดจะถูกเลือกเป็นตัวแทนของแอททริบิวต์ A

ในกรณีที่แอททริบิวต์ที่ทำการพิจารณามีค่าแบบต่อเนื่อง จะดำเนินการด้วยวิธีการเดียวกันกับการคำนวณหาค่าเกนความรู้ ด้วยการหาจุดแบ่งที่ดีที่สุด โดยทำการเรียงลำดับค่าทั้งหมดจากน้อยไปมาก แล้วทำการหาค่ากลางระหว่าง 2 ค่าใดๆ จากนั้นนำแต่ละค่ากลางมาพิจารณาเป็นจุดแบ่ง โดยค่าใดๆก็ตามที่มีค่าน้อยกว่าหรือเท่ากับค่ากลางจะถูกแบ่งไว้ในชุดข้อมูลย่อย D_1 และค่าใดก็ตามที่มีค่ามากกว่าค่ากลางจะถูกแบ่งไว้ในชุดข้อมูลย่อย D_2 ตามลำดับ จากนั้นค่ากลางใดที่ให้ค่า $Gini_A(D)$ น้อยที่สุดจะถูกเลือกเป็นตัวแทนของ แอททริบิวต์ A

ขั้นตอนสุดท้ายของการแบ่งชุดข้อมูลด้วยการใช้ค่าดัชนีจีนี่ เราจะต้องทำการคำนวณส่วนหักลบของความไม่บริสุทธ์ ซึ่งก็คือ $\Delta Gini(A)$ โดยเมื่อทำการคำนวณค่า $\Delta Gini(A)$ สำหรับทุก ๆ แอททริบิวต์แล้ว แอททริบิวต์ใดก็ตามที่มีค่า $\Delta Gini(A)$ มากที่สุด (หรือมีค่า $Gini_A(D)$ น้อยที่สุด) แอททริบิวต์นั้นจะถูกเลือกเพื่อเป็นแอททริบิวต์สำหรับแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อย

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

ตัวอย่างการสร้างต้นไม้ตัดสินใจด้วยค่าเกนความรู้กำหนดให้ชุดข้อมูลสอน D จากร้านขายอุปกรณ์ไฟฟ้า ประกอบไปด้วยเซตของเรคคอร์ด *14* เรคคอร์ด ที่ซึ่งเป็นข้อมูลคุณลักษณะของลูกค้าที่จะทำการซื้อ/ไม่ซื้อคอมพิวเตอร์ (Class : Buys_Computer) ที่ประกอบไปด้วยข้อมูลหลัก 4 แอททริบิวต์ที่มีค่าแบบไม่ต่อเนื่อง คือ Age (อายุ), Income (รายได้), Student (อาชีพนักศึกษาหรือไม่) และ Credit_Rating (ข้อมูลประวัติเครดิตทางการเงิน) และ 1 แอททริบิวต์ หมวดหมู่ของลูกค้าที่ประกอบไปด้วย *2* หมวดหมู่ คือ yes (หมวดหมู่ของลูกค้าที่ซื้อคอมพิวเตอร์จากร้านค้า) และ no (หมวดหมู่ของลูกค้าที่ไม่ได้ซื้อคอมพิวเตอร์จากร้านค้า) เมื่อทำการสังเกตชุดข้อมูล จะเห็นว่ามี *9* เรคคอร์ดที่เป็นข้อมูลลูกค้าในหมวดหมู่ของ yes (การซื้อคอมพิวเตอร์) และมี 5 เรคคอร์ดที่เป็น no (ลูกค้าที่ไม่ซื้อคอมพิวเตอร์) ตามลำดับ

ตารางที่ 2.2 ตัวอย่างข้อมูลสอนจากร้านขายอุปกรณ์ไฟฟ้า

RID	Age	Income	Student	Credit_Rating	Class : Buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

ที่มา: Data Mining Concepts and Techniques, 2012

ขั้นตอนแรกของการสร้างต้นไม้ตัดสินใจคือ การสร้างโหนด N สำหรับข้อมูลทั้งหมดในชุดข้อมูล D จากนั้นทำการค้นหาเกณฑ์/แอททริบิวต์ที่ใช้ในการแยกข้อมูลออกเป็นชุดย่อยๆด้วยการคำนวณหาค่าเกนความรู้ของแต่ละแอททริบิวต์ในตารางที่ 2 (ทำการหาค่าเกนความรู้ของแอททริบิวต์ Age, Income, Student และ Credit_Rating แต่ก่อนที่จะทำการหาค่าเกนความรู้ของแต่ละแอททริบิวต์ จะต้องทำการคำนวณค่าเอนโทรปีของชุดข้อมูล D เสียก่อน ซึ่งสามารถคำนวณได้ดังนี้

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

ขั้นตอนต่อไปจะทำการคำนวณค่าเกนความรู้ของแต่ละแอททริบิวต์ โดยเริ่มทำการพิจารณาแอททริบิวต์อายุเป็นลำดับแรก จากนั้นเราจะพิจารณาการกระจายของข้อมูลหมวดหมู่ต่าง ๆ ตามแต่ละค่าที่เป็นไปได้ของแอททริบิวต์อายุที่มีค่าที่เกิดขึ้น 3 ค่าด้วยกันคือ ‘youth’, ‘middle_aged’ และ ‘senior’ ตามลำดับ เมื่อพิจารณาช่วงอายุที่เป็น ‘youth’ จะเห็นว่ามี 2 เรคคอร์ดที่มีค่าอายุเป็น ‘youth’ และเป็นลูกค้าที่อยู่ในหมวดหมู่ที่ซื้อคอมพิวเตอร์ และมี 3 เรคคอร์ดที่มีค่าอายุเป็น ‘youth’

และอยู่ในหมวดหมู่ที่ไม่ซื้อคอมพิวเตอร์ ในส่วนของช่วงอายุที่เป็น ‘middle_aged’ จะมี 4 เรคคอร์ด ที่มีค่าอายุเป็น ‘middle_aged’ และเป็นลูกค้าที่อยู่ในหมวดหมู่ที่ซื้อคอมพิวเตอร์ แต่จะไม่มีลูกค้ามีค่าอายุเป็น ‘middle_aged’ และอยู่ในหมวดหมู่ที่ไม่ซื้อคอมพิวเตอร์เลย และในส่วนของช่วงอายุที่เป็น ‘senior’ จะมี 3 เรคคอร์ดที่มีค่าอายุเป็น ‘senior’ และเป็นลูกค้าที่อยู่ในหมวดหมู่ที่ซื้อคอมพิวเตอร์และ 2 เรคคอร์ดมีค่าอายุเป็น ‘senior’ และอยู่ในหมวดหมู่ที่ไม่ซื้อคอมพิวเตอร์ ตามลำดับ ดังนั้นเราสามารถคำนวณการกระจายของข้อมูลหมวดหมู่ต่างๆ ตามแต่ละค่าที่เป็นไปได้ของแอททริบิวต์อายุได้เป็น

$$\begin{aligned} Info_{Age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \right) \\ &\quad + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} \right) \\ &\quad + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \right) \\ &= 0.694 \end{aligned}$$

จากค่าการกระจายของข้อมูลหมวดหมู่ต่างๆตามแต่ละค่าที่เป็นไปได้ของแอททริบิวต์ Age สามารถคำนวณค่าเกนความรู้ของการแบ่งชุดข้อมูลด้วยแอททริบิวต์ Age ได้เป็น

$$Gain(Age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$$

หลังจากคำนวณค่าเกนความรู้ของแอททริบิวต์อายุแล้วจะต้องทำการคำนวณค่าเกนความรู้ของแอททริบิวต์อื่นๆ ทั้งหมดด้วยวิธีการเดียวกับข้างต้น ซึ่งสามารถคำนวณได้เป็น

$Gain(income)$ = 0.029

$Gain(student)$ = 0.151

$Gain(credit\_rating)$ = 0.048

เมื่อทำการเปรียบเทียบค่าเกนความรู้ของทุกแอททริบิวต์ จะสังเกตได้ว่าค่าเกนความรู้ของแอททริบิวต์ Age จะมีค่ามากที่สุด ดังนั้นจะเลือกแอททริบิวต์ Age ให้เป็นแอททริบิวต์สำหรับแบ่งชุดข้อมูลออกเป็นออกเป็นชุดข้อมูลย่อย ด้วยการแนบชื่อของโหนด N ด้วยชื่อของแอททริบิวต์ Age และทำการแตกกิ่งจากโหนด N ตามค่าทั้ง 3 ที่เกิดขึ้นในแอททริบิวต์ N (‘youth’, ‘middle_aged’, และ ‘senior’) จากนั้นทำการแบ่งข้อมูลในชุดข้อมูล D จะสังเกตได้ว่าช่วงอายุที่เป็น ‘middle_aged’ จะมีเฉพาะข้อมูลลูกค้าที่อยู่ในหมวดหมู่ที่ซื้อคอมพิวเตอร์ ดังนั้นจะทำการสร้างโหนดใบในกิ่งของช่วงอายุที่เป็น ‘middle_aged’ และกำหนดชื่อโหนดให้เป็น ‘หมวดหมู่ที่ซื้อคอมพิวเตอร์’ ตามลำดับ

จากนั้นจะดำเนินการแบบเวียนเกิดกับแต่ละกิ่งของต้นไม้ที่ไม่ใช่โหนดใบ เมื่อดำเนินการสร้างต้นไม้จนเสร็จสิ้นจะได้ต้นไม้ตัดสินใจ

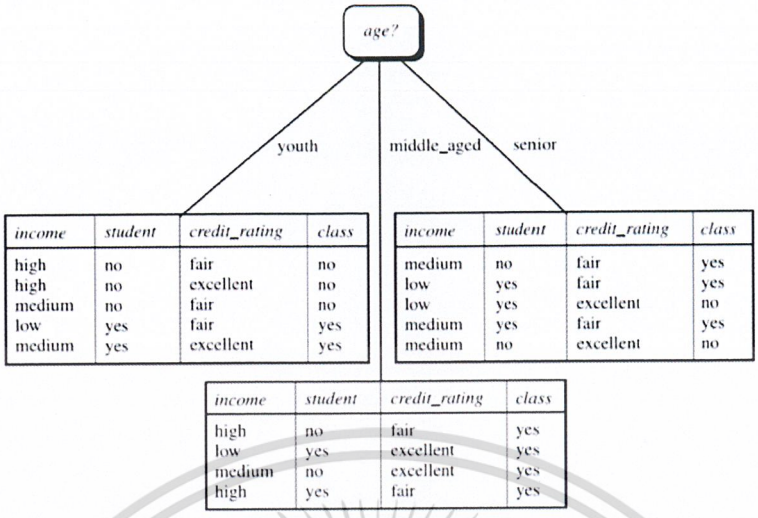
จากตัวอย่างข้างต้น จะสังเกตได้ว่าค่าที่เกิดขึ้นในทุกๆแอททริบิวต์จะมีค่าเป็นแบบไม่ต่อเนื่อง (หมายถึงเป็นค่าที่บ่งบอกถึงหมวดหมู่ที่แน่นอน ที่ซึ่งไม่ใช่ค่าที่แสดงถึงข้อมูลเชิงปริมาณ) แต่สำหรับแอททริบิวต์ที่มีค่าแบบต่อเนื่อง จะต้องคำนวณค่าเกนความรู้ในลักษณะที่แตกต่างออกไป เพื่อให้มีความเข้าใจมากขึ้นจะสมมติว่ามีแอททริบิวต์ A หนึ่งๆที่มีค่าเป็นแบบต่อเนื่อง อาทิเช่น แอททริบิวต์ Age ที่ซึ่งบ่งบอกถึงตัวเลขอายุของลูกค้าแต่ละราย จากค่าที่มีลักษณะแบบต่อเนื่องจะต้องทำการค้นหาจุดแบ่งที่ดีที่สุด (“best” splitpoint) สำหรับแอททริบิวต์ A ที่ซึ่งจุดแบ่งจะทำหน้าที่เปรียบเสมือนค่าขีดแบ่งในการแบ่งย่อยข้อมูลโดยใช้แอททริบิวต์ A ดังนั้น เพื่อที่จะทำการหาจุดแบ่งที่ดีที่สุด จะเริ่มจากการเรียงลำดับค่าที่เกิดขึ้นในแอททริบิวต์ A จากทุกๆเรคคอร์ดในชุดข้อมูลที่ทำการพิจารณาโดยเรียงลำดับจากน้อยไปมาก จากนั้นทำการหาค่ากลาง(midpoint) ระหว่าง 2 ค่าใดที่มีค่าติดกันโดยค่า a_i และ ค่า a_{i+1} ใดๆจะถูกพิจารณาเพื่อคำนวณค่ากลางได้ ดังนี้

$$\frac{a_i + a_{i+1}}{2}$$

ค่ากลางที่คำนวณได้จะถูกนำมาพิจารณาเพื่อเป็นจุดแบ่งชุดข้อมูล ดังนั้น ถ้าแอททริบิวต์ A มีค่าที่เกิดขึ้นในเรคคอร์ดทั้งหมด v ค่าจะสามารถหาค่ากลางได้ทั้งหมด $V-1$ ค่าด้วยกัน เมื่อคำนวณค่ากลางทั้งหมดแล้วจะต้องทำการพิจารณาว่าค่ากลางใดเป็นค่าที่เหมาะจะเป็นจุดแบ่งด้วยการหาค่า $Info_A(D)$ เมื่อจำนวนกลุ่มของข้อมูลที่ต้องการแบ่งออกเป็นชุดข้อมูลย่อย ๆ จะมีค่าเท่ากับ 2 ซึ่งก็คือ

1. กลุ่มของเรคคอร์ดที่มีค่าในแอททริบิวต์ A น้อยกว่าหรือเท่ากับค่ากลางที่ทำการพิจารณา
2. กลุ่มของเรคคอร์ดที่มีค่าในแอททริบิวต์ A มากกว่าค่ากลางที่ทำการพิจารณา

เมื่อทำการหาค่า $Info_A(D)$ ของทุกๆค่ากลางแล้ว จะเลือกค่ากลางที่มีค่า $Info_A(D)$ น้อยที่สุดเป็นจุดแบ่งสำหรับแอททริบิวต์ A ตามลำดับ



รูปที่ 2.11 การแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยด้วยแอททริบิวต์ Age

ที่มา: https://z-p3-lookaside.fbs.com/file/10.WeekX

2.19 ข้อมูล (Data)

ข้อมูล คือ ข้อเท็จจริง หรือค่าที่วัดได้จากลักษณะที่ต้องการศึกษาของหน่วยให้ค่าสังเกตเพียงหน่วยเดียว ค่าที่ได้นี้อาจเป็นตัวเลขเชิงปริมาณทางคณิตศาสตร์ หรือสัญลักษณ์ เช่น การเก็บข้อมูลของครัวเรือนเกษตรกรหนึ่งในจังหวัดชัยนาท พบว่า ครัวเรือนนี้มีสมาชิก 5 คน ประกอบด้วยชาย 3 คน และหญิง 2 คน พืชที่ปลูกส่วนใหญ่เป็นอ้อย กับข้าวโพด ค่าสังเกตของครอบครัวนี้คือ สมาชิก 5 คน ชาย 3 คน หญิง 2 คน อ้อย และข้าวโพด เป็นต้น

ข้อมูลสถิติ (Statistical Data) หมายถึง ค่าสังเกตของลักษณะใดลักษณะหนึ่งหรือหลายลักษณะที่ได้จากหน่วยให้ค่าสังเกตหลาย ๆ หน่วย ข้อมูลสามารถแบ่งได้หลายลักษณะขึ้นอยู่กับว่าจะใช้เกณฑ์ใดในการแบ่ง เช่นแบ่งตามการวัด สามารถแบ่งออกเป็น 4 ระดับ

1. ระดับนามบัญญัติ (Nominal Scale) เป็นการวัดในระดับต่ำสุด โดยกำหนดสัญลักษณ์หรือตัวเลขให้กับลักษณะที่ต้องการวัดของหน่วยให้ค่าสังเกต เพื่อใช้จำแนกกลุ่ม เช่น กำหนดเลข 1 แทนผู้ชาย เลข 2 แทนผู้หญิง ตัวอย่างนี้เลข 1 และเลข 2 ไม่ใช่ตัวเลขทางคณิตศาสตร์ เป็นเพียงสัญลักษณ์ที่ใช้ในการแบ่งผู้ชายออกจากผู้หญิง ตัวเลขเหล่านี้นำมาบวก ลบ ไม่ได้

2. ระดับอันดับบัญญัติ (Ordinal Scale) เป็นการวัดที่ให้รายละเอียดของค่าสังเกตมากกว่าระดับนามบัญญัติ คือ นอกจากแบ่งหน่วยให้ค่าสังเกตออกเป็นกลุ่มแล้ว ยังสามารถบอกอันดับได้ เช่น การประเมินความชอบของนักศึกษาที่มีต่อการเรียนสถิติเบื้องต้น โดยมีการประเมินดังนี้ เลข 1 น้อยที่สุด, เลข 2 น้อย, เลข 3 ปานกลาง, เลข 4 มาก, เลข 5 มากที่สุด จะเห็นว่าตัวเลขดังกล่าวนอกจากใช้แบ่งนักศึกษาออกเป็น 5 กลุ่ม ตามความชอบแล้ว ยังบอกได้ว่านักศึกษาคนใดชอบเรียนสถิติมากกว่ากัน ตัวเลขในระดับการวัดนี้เป็นเพียงสัญลักษณ์ที่บอกขนาดความชอบ ไม่ใช่ปริมาณ

ความชอบ ดังนั้น จึงไม่สามารถบอกปริมาณความแตกต่างระหว่างตัวเลขได้เช่นเดียวกับตัวเลขทางคณิตศาสตร์ การนำตัวเลขมาบวก ลบ หรือแสดงในรูปของอัตราส่วนไม่ก่อให้เกิดความหมายใด ๆ

3. ระดับช่วงบัญญัติ (Interval Scale) เป็นการวัดที่ครอบคลุมระดับอันดับบัญญัติ แต่ตัวเลขที่ให้สามารถบอกปริมาณของลักษณะที่ต้องการวัดได้ และระยะห่างระหว่างตัวเลขมีค่าเท่ากัน ดังนั้น จึงบอกปริมาณความแตกต่างได้ เช่น ผลต่างของคะแนนสอบวิชาสถิติเบื้องต้นของนักศึกษาที่สอบได้ 60 คะแนน กับ 40 คะแนน เท่ากับ 20 คะแนน เป็นต้น การวัดในระดับนี้ไม่มีจุดเริ่มต้นตามธรรมชาติในความหมายมี่ไม่มีค่า หรือศูนย์ที่แท้จริง เลข 0 จึงเป็นเพียง 0 สมมติ เช่น นักศึกษาคนหนึ่งสอบวิชาสถิติเบื้องต้นได้ 0 คะแนน ไม่ได้หมายความว่านักศึกษาคนนี้ไม่มีความรู้ในวิชาสถิติ แต่ที่สอบได้ 0 คะแนนอาจเป็นเพราะว่าข้อสอบยาก หรือเขาอ่านหนังสือเฉพาะบางเรื่องซึ่งไม่ตรงกับข้อสอบที่ออก เป็นต้น เพราะ 0 คะแนนไม่ได้บ่งบอกว่าไม่มีความรู้ เขาอาจมีความรู้บ้างซึ่งไม่จำเป็นต้องเท่ากัน ดังนั้น การวัดในระดับนี้จึงไม่สามารถแสดงในรูปอัตราส่วนได้

4. ระดับอัตราส่วนบัญญัติ (Ratio Scale) เป็นการวัดที่ครอบคลุมระดับช่วงบัญญัติ ระดับนี้มีจุดเริ่มต้นตามธรรมชาติ หรือศูนย์ที่แท้จริง เช่น น้ำหนัก 0 กิโลกรัม หมายความว่าไม่มีน้ำหนัก เป็นต้น ดังนั้นตัวเลขในระดับนี้จึงสามารถแสดงในรูปของอัตราส่วนได้ เช่น ส้ม 10 กิโลกรัม หนักเป็น 2 เท่าของส้ม 5 กิโลกรัม เป็นต้น แบ่งตามแหล่งกำเนิด สามารถแบ่งออกเป็น 2 ประเภท คือ ข้อมูลปฐมภูมิ (Primary Data) หมายถึง ข้อมูลที่ผู้ใช้เป็นผู้เก็บรวบรวมข้อมูลขึ้นเอง เช่น การเก็บแบบสอบถาม การทดลองในห้องทดลองและข้อมูลทุติยภูมิ (Discrete Data) หมายถึง ข้อมูลที่ผู้ใช้นำข้อมูลมาจากหน่วยงานอื่นหรือผู้อื่นที่ได้ทำการเก็บรวบรวมมาแล้วในอดีต เช่น รายงานประจำปีของหน่วยงานต่าง ๆ ซึ่งแบ่งตามลักษณะของข้อมูล สามารถแบ่งออกเป็น 2 ประเภท คือ ข้อมูลเชิงคุณภาพ (Qualitative Data) หมายถึง ข้อมูลที่ไม่สามารถบอกได้ว่ามีค่ามากหรือน้อย แต่สามารถบอกได้ว่าดีหรือไม่ดี หรือบอกลักษณะความเป็นกลุ่มของข้อมูล เช่น เพศ ศาสนา สีผม คุณภาพสินค้า ความพึงพอใจ ฯลฯ และข้อมูลเชิงปริมาณ (Quantitative Data) หมายถึง ข้อมูลที่สามารถวัดค่าได้ว่ามีค่ามากหรือน้อย ซึ่งสามารถวัดค่าออกมาเป็นตัวเลขได้ เช่น คะแนนสอบ อุณหภูมิ ส่วนสูง น้ำหนัก ปริมาณ ฯลฯ

ตัวแปร (Variable) หมายถึง ลักษณะ หรือคุณลักษณะของหน่วยให้ค่าสังเกตที่แปรค่าได้จากหน่วยหนึ่งไปยังอีกหน่วยหนึ่ง ตัวแปรแบ่งออกเป็น 2 ชนิด คือ ตัวแปรเชิงปริมาณ (Quantitative Variable) เป็นตัวแปรที่มีค่าแทนด้วยเลขจำนวนในทางคณิตศาสตร์ เช่น ส่วนสูง น้ำหนัก อายุ จำนวนบุตร จำนวนอุบัติเหตุ เป็นต้น และตัวแปรเชิงคุณภาพ (Qualitative Variable) เป็นตัวแปรที่มีค่าแสดงถึงสถานภาพ หรือคุณสมบัติ เช่น เพศ ภาควิชา ชั้นปี อาชีพ ศาสนา สถานสมรส สีตา เป็นต้น นอกจากนั้น ตัวแปรเชิงปริมาณยังแบ่งออกเป็น 2 ชนิด คือ ตัวแปรต่อเนื่อง (Continuous Variable) เป็นตัวแปรที่มีค่าต่อเนื่องกันตลอด ไม่สามารถบอกได้ว่ามีจำนวนเท่าใด ค่าเหล่านี้แทนได้ด้วยจุดบนเส้นจำนวนจริง หรือส่วนหนึ่งของเส้นจำนวนจริง กับตัวแปรไม่ต่อเนื่อง (Discrete Variable) เป็นตัว

แปรที่มีค่าแทนด้วยเลขจำนวนนับในทางคณิตศาสตร์หรือตัวแปรที่มีค่าแต่ละค่าแยกออกจากกันด้วยช่องว่าง เช่น จำนวนบุตร จำนวนอุบัติเหตุ เป็นต้น

ตารางแจกแจงความถี่ (Frequency Table) เป็นตารางที่แสดงจำนวนข้อมูลของแต่ละค่าสังเกตหรือช่วงของค่าสังเกตที่ไม่ซ้ำซ้อนกัน ตารางแจกแจงความถี่ในลักษณะแรกเรียกว่าตารางแจกแจงความถี่แบบไม่มีช่วง หรือไม่มีอันตรภาคชั้น (ungrouped frequency table) และลักษณะที่สองเรียกว่าตารางแจกแจงความถี่แบบมีช่วง หรือมีอันตรภาคชั้น (grouped frequency table) จำนวนข้อมูลในตารางสามารถแสดงได้ทั้งในรูปของ ความถี่ ความถี่สัมพัทธ์ หรือร้อยละ

การสร้างตารางแจกแจงความถี่แบบไม่มีอันตรภาคชั้น (ungrouped frequency table) มีขั้นตอนดังนี้

1. เขียนค่าสังเกตที่เป็นไปได้ทั้งหมดของตัวแปร เริ่มต้นจากค่าต่ำสุด จนถึงค่าสูงสุดของข้อมูลชุดที่กำลังพิจารณา
2. เขียนรอยขีด สำหรับทุกค่าสังเกตที่สอดคล้องกับค่าสังเกตที่เขียนไว้ในขั้นตอนที่ 1
3. หาความถี่ โดยนับจากรอยขีดในแต่ละค่าสังเกต
4. สร้างตารางแจกแจงความถี่ โดยเขียนเฉพาะค่าสังเกตที่มีความถี่เท่านั้น

การสร้างตารางแจกแจงความถี่แบบมีอันตรภาคชั้น (grouped frequency table) มีขั้นตอนดังนี้

1. หาค่าสูงสุด (maximum) และค่าต่ำสุด (minimum) ของข้อมูล
2. หาค่าพิสัย (range) เท่ากับ ค่าสูงสุด – ค่าต่ำสุด
3. กำหนดจำนวนชั้น (number of classes) แทนด้วย c การกำหนดจำนวนชั้นไม่มีหลักเกณฑ์ที่แน่นอน ขึ้นกับจำนวนข้อมูล แต่ไม่ควรมีน้อยหรือมากเกินไป ตามปกติจะอยู่ระหว่าง 5 ถึง 15 ชั้น

2.20 การกำหนดประชากรและกลุ่มตัวอย่าง

ประชากร หมายถึง หน่วยของข้อมูลทั้งหมดทุกหน่วยที่อยู่ในขอบข่ายที่ต้องการศึกษา ซึ่งเป็นได้ทั้งคน สัตว์ สิ่งของ และพืช ประชากรการวิจัยมี 2 ชนิด คือ ประชากรที่มีจำนวนนับได้แน่นอน (Finite population) และประชากรที่มีจำนวนนับได้ไม่แน่นอน (Infinite population) การวิจัยส่วนใหญ่ไม่สามารถศึกษาประชากรทั้งหมดได้จึงต้องเลือกประชากรบางส่วนมาศึกษาเรียกว่าตัวอย่าง

การเลือกกลุ่มตัวอย่าง ช่วยในการประหยัดเวลาและงบประมาณ ลดปัญหาด้านการบริหาร สามารถนำมาอ้างอิงถึงประชากร ในการเลือกผู้วิจัยจะต้องมีความรู้ความเข้าใจเกี่ยวกับขั้นตอนการเลือกกลุ่มตัวอย่าง ขนาดของตัวอย่าง และวิธีการเลือกตัวอย่างให้ถูกต้อง

2.20.1ขั้นตอนในการเลือกกลุ่มตัวอย่าง

ข้อแนะนำในการเลือกกลุ่มตัวอย่าง เพื่อให้ปราศจากอคติ

1. วิเคราะห์วัตถุประสงค์ของงานวิจัยให้เข้าใจอย่างแจ่มชัดว่าต้องการศึกษาอะไร ประชากรที่จะศึกษาคืออะไร กลุ่มตัวอย่างมีลักษณะเป็นเช่นไร จะทำการวัดอย่างไร และนำผลที่ได้จากการวัดไปใช้ทำอะไร

2. ควรทราบประชากรเป้าหมายว่ามีลักษณะอย่างไร ต้องให้คำจำกัดความของประชากรว่าหมายถึงใคร มีลักษณะเช่นไร และมีขอบเขตเพียงใด

3. กำหนดกรอบตัวอย่าง (Sampling) คือ การแสดงรายชื่อของทุก ๆ หน่วยที่เป็นประชากรที่ศึกษา ซึ่งจะใช้ในการเลือกกลุ่มตัวอย่าง กรอบตัวอย่างที่ดีจะต้องมีข้อมูลที่เป็นปัจจุบัน หรือเป็นกรอบที่ดีที่สุดเท่าที่จะทำได้ ทั้งนี้ขึ้นอยู่กับเวลาและงบประมาณที่มีอยู่

4. ควรทราบขนาดที่เหมาะสมกับการวิจัย และเทคนิคหรือวิธีการเลือกกลุ่มตัวอย่างที่ถูกต้อง นำมาวางแผนในการเลือกกลุ่มตัวอย่าง เพื่อให้ได้ตัวแทนที่ดีของประชากรเป้าหมาย และสามารถนำผลการวิจัยนั้น ไปสรุปอ้างอิงแทนประชากรเป้าหมายได้ ในขั้นนี้ผู้วิจัยต้องตัดสินใจว่าจะเลือกกลุ่มตัวอย่างโดยวิธีการใดที่สอดคล้องกับลักษณะของประชากร เวลา และงบประมาณ

2.20.2 ลักษณะกลุ่มตัวอย่างที่ดี

กลุ่มตัวอย่างที่ดีจะต้องสามารถเป็นตัวแทนที่ดีของประชากรเป้าหมาย ซึ่งควรมีลักษณะดังนี้ คือ

1. เป็นตัวแทนที่ดีของกลุ่มประชากรที่ต้องการศึกษา โดยมีลักษณะของกลุ่มประชากรครบถ้วน

2. มีขนาดพอเหมาะ คือ จำนวนกลุ่มตัวอย่างไม่น้อยเกินไป ควรเพียงพอที่จะอนุมานถึงประชากรได้ และมีจำนวนไม่มากเกินไปจนเป็นปัญหาในด้านเวลาและงบประมาณในการวิจัย

3. กลุ่มตัวอย่างต้องมีความเชื่อถือได้ คือ ทุกหน่วยของตัวอย่างควรได้รับการเก็บข้อมูลอย่างถูกต้องทางเทคนิค และถูกเลือกโดยไม่มีความลำเอียง

2.20.3 การกำหนดขนาดของกลุ่มตัวอย่าง

การกำหนดขนาดของกลุ่มตัวอย่าง มีวิธีการกำหนด 3 วิธี ดังต่อไปนี้

1. การใช้เกณฑ์ หรือการประมาณจากจำนวนประชากร ในการกำหนดขนาดของกลุ่มตัวอย่างโดยวิธีนี้ ผู้วิจัยต้องทราบจำนวนประชากรที่แน่นอนแล้วจึงนำมาคำนวณหากลุ่มตัวอย่างจากเกณฑ์ดังนี้ (บุญชม ศรีสะอาด, 2535: 38)

- จำนวนประชากรทั้งหมดเป็นหลักร้อย ขนาดของกลุ่มตัวอย่าง 15-30 %
- จำนวนประชากรทั้งหมดเป็นหลักพัน ขนาดของกลุ่มตัวอย่าง 10-15 %
- จำนวนประชากรทั้งหมดเป็นหลักหมื่น ขนาดของกลุ่มตัวอย่าง 5-10 %

2. การใช้สูตรคำนวณ สามารถใช้กำหนดขนาดของกลุ่มตัวอย่างได้ทั้งประชากรที่มีจำนวนแน่นอน และมีจำนวนไม่แน่นอน ดังนี้

ก. กรณีที่ประชากรมีจำนวนไม่แน่นอน (Infinite population) ซึ่งผู้วิจัยไม่สามารถทราบจำนวนประชากร ทราบเพียงว่ามีจำนวนมากใช้สูตรดังนี้ (Roscoe, 1969: 156-157)

$$N = (z_c\sigma/e_M)^2$$

เมื่อ N = จำนวนตัวอย่างประชากร

$\mathbf{z_c}$ = คะแนน z ตามระดับความมีนัยสำคัญที่ผู้วิจัยกำหนดให้ α

Z = 1.96 ที่ระดับความมั่นใจ 95% (α = .05)

Z = 2.58 ที่ระดับความมั่นใจ 99% (α = .01)

e_M = ค่าความคาดเคลื่อนมากที่สุดที่ยอมรับได้

σ = ส่วนเบี่ยงเบนมาตรฐานของประชากร

ในการวิจัยทางพฤติกรรมศาสตร์นั้นส่วนใหญ่จากประชากรที่มีขนาดไม่แน่นอน จะใช้กลุ่มตัวอย่างตั้งแต่ 30 หน่วย ถึง 500 หน่วย (ประคอง กรรณสูตร, 2538: 10) ซึ่ง รอสโก (Rosco, 1969: 157) กล่าวว่าขนาดตัวอย่าง 500 หน่วย จากประชากรที่ไม่แน่นอน ค่าความคาดเคลื่อนเนื่องจากการสุ่มตัวอย่างจะไม่เกิน σ/10 ดังนั้น หากจะใช้ประชากรตั้งแต่ 30 ถึง 500 ก็ใช้ e_M=σ/10

เป็นที่น่าสังเกตว่าขนาดของกลุ่มตัวอย่างที่คำนวณจากสูตรที่ 1 นั้น ไม่ว่าขนาดของประชากร มีมากเพียงใด หากกำหนดให้ความคาดเคลื่อนมากที่สุดที่ยอมรับได้เท่ากับ 1/10 ของส่วนเบี่ยงเบนมาตรฐาน ($e_M = \sigma/10$) และระดับมีนัยสำคัญทางสถิติที่ระดับ .05 แล้วเมื่อแทนค่าในสูตรจะต้องใช้ขนาดของกลุ่มตัวอย่าง 384 หน่วยเสมอไป

นอกจากสูตรที่กล่าวข้างต้น ในกรณีที่ประชากรมีจำนวนไม่แน่นอน ยังสามารถใช้สูตรอื่น ดังนี้ (บุญชม ศรีสะอาด, 2535: 38)

$$\mathbf{n = \frac{P(1-P)z^2}{e^2}}$$

เมื่อ **n** = จำนวนกลุ่มตัวอย่าง

P = สัดส่วนของประชากรที่ผู้วิจัยกำหนดสุ่ม

z = ระดับความมั่นใจ ที่ผู้วิจัยกำหนด

z = 1.96 ที่ระดับความมั่นใจ 95% (α = 0.5)

z = 2.58 ที่ระดับความมั่นใจ 99% (α = .01)

e = สัดส่วนของความคาดเคลื่อนที่ยอมรับให้เกิดขึ้นได้

ข. กรณีที่ประชากรมีจำนวนแน่นอน (Finite population) ใช้สูตรดังต่อไปนี้ เมื่อระดับความมีนัยสำคัญทางสถิติเท่ากับ .05 (Yamane, 1970: 580-581)

$$n = \frac{N}{1 + Ne^2}$$

เมื่อ n = จำนวนกลุ่มตัวอย่าง

e= ความคาดเคลื่อนของการสุ่มตัวอย่าง

N = ขนาดของประชากร

หรือใช้สูตรดังนี้ (บุญธรรม กิจปรีดาบริสุทธิ์, 2535: 68)

$$n = \frac{400N}{399 + N}$$

นอกจาก 2 สูตรที่กล่าวแล้วข้างต้น กรณีที่ประชากรมีจำนวนแน่นอนนั้น สามารถใช้สูตรดังนี้ (บุญชม ศรีสะอาด, 2535: 39)

เมื่อ n = จำนวนกลุ่มตัวอย่าง

N = จำนวนประชากร

P = สัดส่วนของประชากรที่ผู้วิจัยกำหนดสุ่ม

e = สัดส่วนของความคาดเคลื่อนที่ยอมให้เกิดขึ้นได้

3. การใช้ตารางสำเร็จรูป มีผู้เสนอตารางสำเร็จรูปสำหรับขนาดของตัวอย่างประชากรในกรณีที่ทราบกรอบประชากร มีดังนี้

ก. ตารางสำเร็จรูปของเครซี่ และมอร์แกร (Krejcie & Morgan, 1970: 607-610) R.V.Krejcie และ D.W. Morgan ได้เสนอตารางจำนวนของกลุ่มตัวอย่าง

ตั้งแต่ประชากร 10 คน ถึง 100,000 คน และในกรณีที่ประชากรมีจำนวนไม่ตรงกับที่ปรากฏในตารางให้ใช้หลักของบัญญัติไตรยางค์คำนวนกลุ่มตัวอย่าง

ข. ตารางสำเร็จรูปที่กำหนดระดับความเชื่อมั่น 95% (Arkin & Colton, 1963: 151-152 อ้างถึงในเพชรน้อย สิงห์ช้างชัย, 2539: 143)

2.20.4 วิธีเลือกกลุ่มตัวอย่าง

วิธีเลือกกลุ่มตัวอย่าง เป็นวิธีการเพื่อให้ได้ตัวอย่างที่มีคุณสมบัติเป็นตัวแทนของประชากรเป้าหมาย สามารถนำผลการวิจัยที่ได้ไปสรุปผลอ้างอิงไปยังประชากรทั้งหมด ซึ่งวิธีการเลือกกลุ่มตัวอย่างจะแบ่งออกเป็น 2 ประเภทใหญ่ ๆ คือ

1.การเลือกกลุ่มตัวอย่างโดยไม่ใช้หลักความน่าจะเป็น (Non-probability sampling)

การเลือกกลุ่มตัวอย่างโดยไม่ใช้หลักความน่าจะเป็น (Non-probability sampling) หมายถึง การเลือกตัวอย่างโดยไม่คำนึงถึงโอกาสที่จะถูกเลือกของประชากรแต่ละหน่วย เพราะลักษณะบางอย่างของประชากรไม่อำนวยให้เลือกโดยวิธีใช้หลักความน่าจะเป็น การเลือกตัวอย่างประเภทนี้จะง่ายในการปฏิบัติเพราะไม่ต้องทำกรอบบัญชีรายชื่อ จะคำนึงถึงความสะดวกในการเก็บข้อมูล และผลที่ได้จะสรุปเฉพาะกลุ่มที่ศึกษา การเลือกกลุ่มตัวอย่างโดยไม่ใช้หลักความน่าจะเป็นสามารถจำแนกได้เป็น 4 วิธี ดังนี้

1.1 การเลือกตัวอย่างแบบบังเอิญ (Accidental sampling) หมายถึง การเลือกตัวอย่างโดยไม่มีกฎเกณฑ์ จะเลือกใครก็ได้เท่าที่หาได้จนครบตามจำนวนที่ต้องการ โดยคำนึงถึงความสะดวกสบายของผู้วิจัยเป็นหลักและผู้วิจัยจะเก็บบันทึกข้อมูลจากหน่วยนั้นซึ่งยินดีให้ความร่วมมือ และบังเอิญอยู่ในสถานที่ที่ผู้วิจัยกำลังเก็บข้อมูล การเลือกตัวอย่างแบบบังเอิญจะคำนึงถึงการได้ข้อมูลมาวิเคราะห์เป็นประเด็นสำคัญ ดังนั้น จึงควรระมัดระวังในการตีความหมายของข้อมูล

1.2 การเลือกกลุ่มตัวอย่างแบบเจาะจง (Purposive sampling) หมายถึง การเลือกตัวอย่างโดยการกำหนดคุณสมบัติของกลุ่มตัวอย่างไว้ ว่าจะเลือกกลุ่มตัวอย่างที่มีคุณสมบัติอย่างไรจึงจะเหมาะสมกับการวิจัย และเป็นตัวแทนของประชากรที่ศึกษา ถ้าพบตัวอย่างใดที่มีคุณสมบัติครบถ้วนตามกำหนดก็จะเลือกตัวอย่างมาศึกษาได้ ซึ่งการเลือกตัวอย่างโดยวิธีนี้พบมากในการวิจัยทางการพยาบาล

1.3 การเลือกกลุ่มตัวอย่างแบบกำหนดโควตา (Quota sampling) หมายถึง การเลือกตัวอย่างที่ผู้วิจัยกำหนดโควตาของสัดส่วนขนาดตัวอย่างย่อย

(Subgroup) ไว้ล่วงหน้า โดยใช้องค์ประกอบหรือคุณลักษณะของประชากร เช่น เพศ อายุ การศึกษา ฯลฯ เป็นเครื่องกำหนด ซึ่งการเลือกแบบนี้เหมาะสำหรับงานวิจัยที่ต้องการศึกษากลุ่มประชากรที่มีลักษณะย่อยต่างกัน

1.4 การเลือกกลุ่มตัวอย่างตามสะดวกหรือสมัครใจ (Coveninient or Voluteer sampling) หมายถึง การเลือกตัวอย่างตามความสะดวกและสมัครใจของผู้วิจัยและผู้ถูกวิจัย

2. การเลือกกลุ่มตัวอย่างโดยใช้หลักความน่าจะเป็น (Probability sampling)

การเลือกกลุ่มตัวอย่างโดยใช้หลักความน่าจะเป็น (Probability sampling) หมายถึง การสุ่มตัวอย่างโดยคำนึงถึงโอกาสที่ทุกหน่วยประชากรจะถูกเลือก และสามารถประมาณค่าความน่าจะเป็นได้ เป็นการสุ่มตัวอย่างที่มีระบบและกฎเกณฑ์ จะต้องจัดทำกรอบตัวอย่าง (Sample frame) ไว้ การสุ่มตัวอย่างโดยใช้หลักความน่าจะเป็นสามารถจำแนกได้เป็น 5 วิธี ดังนี้

1.1 การสุ่มตัวอย่างแบบง่าย (Simple random sampling) หมายถึง การสุ่มอย่างง่าย ๆ ที่ทุกหน่วยประชากรมีโอกาสถูกเลือกได้เท่ากัน มี 2 วิธีคือ การจับฉลาก และการใช้ตารางเลขสุ่ม

1.2 การสุ่มตัวอย่างแบบมีระบบ (Systematic sampling) หมายถึง การสุ่มตัวอย่างจากประชากร โดยยึดช่วงห่างของลำดับที่ประชากรเป็นเกณฑ์ในการเลือก การสุ่มวิธีนี้ใช้ได้เฉพาะกรณีที่มีกรอบตัวอย่างของประชากร

1.3 การสุ่มตัวอย่างแบบแบ่งชั้น (Stratified sampling) หมายถึง การสุ่มตัวอย่างจากประชากรที่มีหลายลักษณะรวมกัน โดยมีหลักว่าต้องแบ่งชั้นโดยให้ประชากรที่อยู่ในกลุ่มเดียวกันมีความคล้ายคลึงกันมากที่สุด และประชากรที่อยู่ต่างกลุ่มกันมีความแตกต่างกันมากที่สุด จากนั้นจึงสุ่มตัวอย่างออกมาจากแต่ละกลุ่มโดยทำการสุ่มตัวอย่างแบบง่ายหรือแบบมีระบบก็ได้ ให้ได้จำนวนประชากรตามสัดส่วนที่ต้องการ การแบ่งชั้นของประชากรอาจแบ่งตามเพศ อายุ อาชีพ ระดับการศึกษา ศาสนา รายได้ เป็นต้น

1.4 การสุ่มตัวอย่างแบบแบ่งกลุ่ม (Cluster sampling or Area sampling) หมายถึง วิธีการสุ่มตัวอย่างโดยการแบ่งกลุ่มประชากรออกเป็นกลุ่มตามพื้นที่ทางภูมิศาสตร์ สถาบัน หน่วยงานหรือสมาคม แล้วทำการเลือกมาเพียงบางส่วนด้วยวิธีการสุ่มแบบธรรมดาหรือแบบเป็นระบบก็ได้ โดยให้ประชากรในกลุ่มแตกต่างกันมากที่สุด และประชากรระหว่างกลุ่มคล้ายคลึงกันมากที่สุด การเลือกจึงสามารถนำมาศึกษาเพียงหนึ่งหรือสองกลุ่มก็ได้

1.5 การสุ่มตัวอย่างแบบหลายขั้นตอน (Multi-stage sampling) หมายถึง กระบวนการสุ่มตัวอย่างจากประชากร โดยดำเนินการสุ่มตั้งแต่ 3 ขั้นขึ้นไป เป็นวิธีที่เหมาะกับประชากรที่มีขอบข่ายกว้าง ไม่สามารถหากรอบบัญชีรายชื่อที่ประกอบด้วยทุกหน่วย

ประชากรได้โดยตรง ซึ่งอาจเนื่องมาจากความไม่สะดวก ความสิ้นเปลือง ดังนั้น จึงทำเพียงกรอบบัญชีรายชื่อเฉพาะกลุ่มที่เลือกได้เท่านั้น

2.21 ไฟล์ข้อมูล CSV

ประเภทข้อมูลที่สามารถ Import เข้ามาใน RapidMiner เพื่อ Process มีถึง 16 ฟอร์แมท แต่ที่นิยมนำมาใช้กันคือ ฟอร์เมทที่เป็น CSV

ไฟล์ CSV ย่อมาจาก Comma Separated Value เป็นไฟล์ประเภทหนึ่งที่ใช้สำหรับเก็บข้อมูลในรูปแบบตาราง โดยใช้จุลภาค (,) คั่นระหว่างค่า โดยปกติเราสามารถบันทึกไฟล์จาก Microsoft Excel ออกมาเป็น CSV ไฟล์ได้โดยตรง หรือ อาจได้ไฟล์ CSV จากการ export ไฟล์จากระบบฐานข้อมูลอื่น ๆ เมื่อข้อความและตัวเลขถูกบันทึกในไฟล์ CSV การย้ายจากโปรแกรมหนึ่งไปอีกโปรแกรมหนึ่งจึงเป็นเรื่องง่าย ไฟล์ CSV แตกต่างจากไฟล์สเปรดชีตชนิดอื่นตรงที่คุณไม่สามารถบันทึกการจัดรูปแบบและสูตรได้

2.22 RapidMiner Studio 7

RapidMiner เป็นแพลตฟอร์มซอฟต์แวร์ข้อมูลวิทยาศาสตร์ที่พัฒนาขึ้นโดยบริษัทที่มีชื่อเดียวกันซึ่งมีสภาพแวดล้อมแบบรวมสำหรับการจัดเตรียมข้อมูลการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ลึกการทำเหมืองข้อความและการวิเคราะห์เชิงคาดการณ์ ใช้สำหรับการใช้งานทางธุรกิจและเชิงพาณิชย์รวมทั้งการวิจัยการศึกษาการฝึกอบรมการสร้างต้นแบบอย่างรวดเร็วและการพัฒนาแอพพลิเคชันและสนับสนุนขั้นตอนทั้งหมดของกระบวนการเรียนรู้ของเครื่องรวมถึงการเตรียมข้อมูลการสร้างภาพผลการตรวจสอบรูปแบบและการเพิ่มประสิทธิภาพ RapidMiner ได้รับการพัฒนาบนโมเดลแกนเปิด RapidMiner Studio Free Edition ซึ่งมีข้อ จำกัด อยู่ที่ 1 ตัวประมวลผลเชิงตรรกะและแถวข้อมูล 10,000 ชุดสามารถใช้ได้ภายใต้ใบอนุญาต AGPL การกำหนดราคาทางการค้าเริ่มต้นที่ 2,500 เหรียญและสามารถหาได้จากผู้พัฒนา

2.22.1 การเริ่มต้นใช้งาน

เมื่อเริ่มต้นใช้งาน RapidMiner Studio 7 จะแสดงหน้าต่างเริ่มต้นซึ่งประกอบด้วย 4 เมนูหนักดังนี้

- GET STARTED แสดงวิธีเริ่มต้นการใช้งานในรูปแบบวีดีโอ คือ

- การนำข้อมูลเข้า (Import Data)

- การสร้างโปรเซส (Build Process)

- การประมวลผล (Run Process)

- ตรวจสอบและวิธีการดูผลการประมวลผลของโปรเซส (Inspect Result)

- การปรับตั้งค่าในโปรเซส (Improve Process)

- LEARN เป็นหน้าที่รวบรวมและแสดงวิธีการใช้งานของ RapidMiner Studio 7 ซึ่งทำ Link ไปยังหน้าเวปที่ แสดงการใช้งานในรูปแบบ document,VDO และมีบทเรียนฝึกหัดให้ทำตาม 3 บทคือ

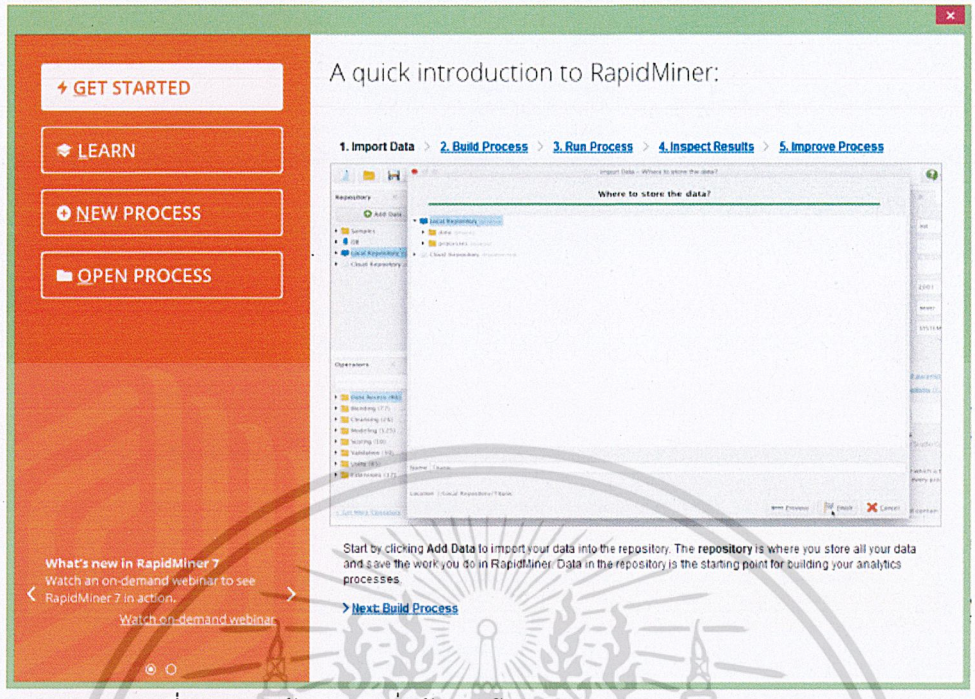
- Basic สอนพื้นฐานการใช้งาน

- Data Handling สอนการจัดการข้อมูล

- Modeling, Scoring and Validation โดยสอนเกี่ยวกับการสร้างโมเดลเพื่อทำ Prediction, การวิเคราะห์ผลและนำผลลัพธ์ข้อมูลเชิงลึกที่คาดการณ์ได้มาใช้จริง และสอนกระบวนการยืนยันความถูกต้องของโมเดล

- NEW PROCESS สร้างโปรเซสใหม่เพื่อเริ่มการใช้งาน RapidMiner ซึ่งทุกครั้งที่ต้องสร้างงานใหม่ที่แตกต่างจะต้องสร้างโปรเซสใหม่

- OPEN PROCESS เปิดโปรเซสเก่าที่เคยสร้างไว้เพื่อดูหรือแก้ไข โดยโปรเซสที่สร้างไว้แล้วสามารถ Reuse ได้ หรือ ส่งให้คนอื่นได้



รูปที่ 2.12 หน้าต่างค่าเริ่มต้นของโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

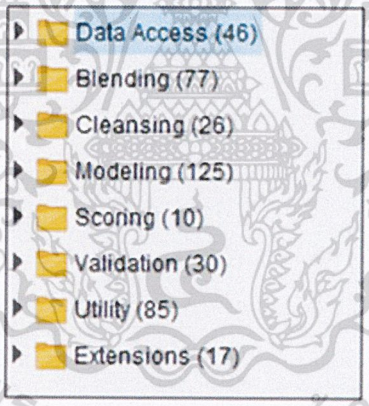
รูปที่ 2.13 องค์ประกอบของหน้าต่าง Design ใน RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

หน้า Design RapidMiner Studio 7 ประกอบด้วย 5 ส่วนหลัก ๆ คือ

1. Operators เป็นส่วนที่ใช้เก็บตัวโอเปอร์เรเตอร์ที่ใช้ในการทำงานทั้งหมด ซึ่งจัดเป็นกลุ่ม ๆ โดยกลุ่มที่ใช้งานคล้ายคลึงกันจะจะอยู่ในกลุ่มเดียวกัน มี 8 กลุ่ม โอเปอร์เรเตอร์คือ

   a. Data Access
   b. Blending
   c. Cleansing
   d. Modeling
   e. Scoring
   f. Validation
   g. Utility
   h. Extensions



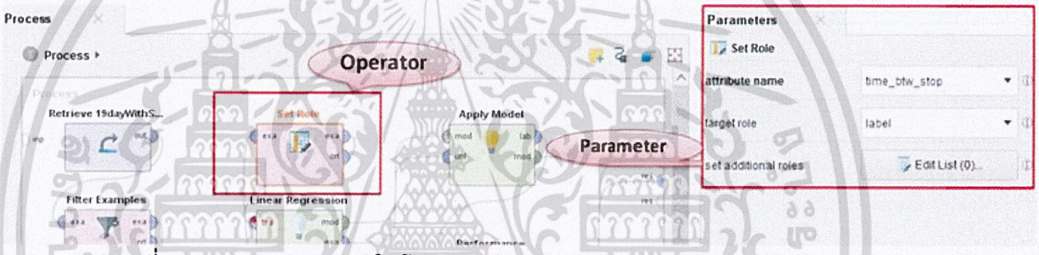
รูปที่ 2.14 Operator ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

2. Repository ส่วนนี้เป็นส่วนจัดการไฟล์ RapidMiner Studio 7 จะจัดการข้อมูลจาก 3 แหล่ง คือ DB (ดาต้าเบส), Local (ในเครื่องคอมพิวเตอร์ที่ใช้อยู่) และ Cloud Repository (ในคลาวด์) โดยเก็บไฟล์ Data Set และ Process ต่าง ๆ แยกเก็บคนละโฟลเดอร์กัน

3. Process เป็นหน้าหลักในการทำงานในการสร้างโปรเซสสำหรับทำ Machine Learning ของซอฟต์แวร์นี้ โดยจะนำโอเปรอเรเตอร์มาประกอบเพื่อสร้างโปรเซสขึ้นมาตามวัตถุประสงค์ของโจทย์ที่ตั้งไว้

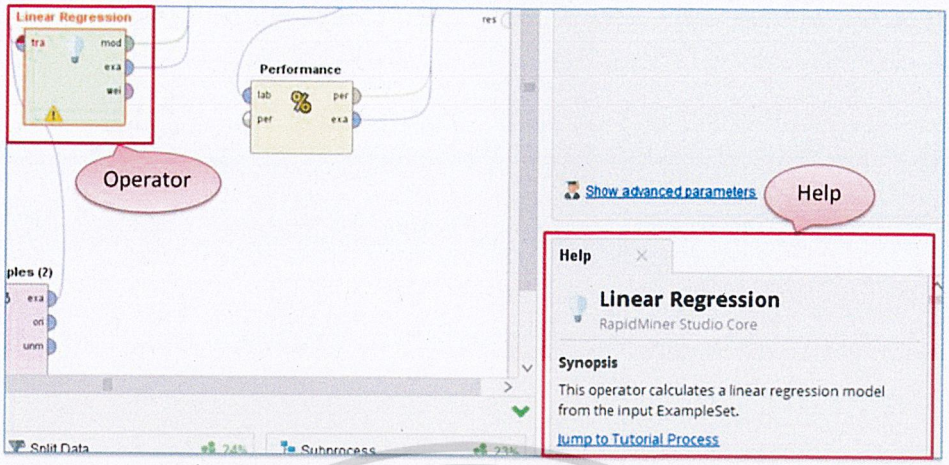
4. Parameters เป็นส่วนที่กำหนดพารามิเตอร์ที่เป็นรายละเอียดของโอเปอร์เรเตอร์ ที่เลือกใช้งาน เช่น โอเปอเรเตอร์ Set Role เป็น Operator ที่มีพารามิเตอร์ที่เกี่ยวข้องสองพารามิเตอร์ คือ แอททริบิวต์เนม ซึ่งจากตัวอย่าง เลือก time_btw_stop เป็นเอาท์พุทของโปรเซส และ พารามิเตอร์ target role เพื่อระบุว่าพารามิเตอร์เอาท์พุทเป็น label เท่านั้น



รูปที่ 2.15 Parameter ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

5. Help เป็นส่วนช่วยเหลือซึ่งจะแสดงรายละเอียดของตัวโอเปอเรเตอร์ที่เลือกใช้งานอยู่ ส่วนช่วยเหลือของ RapidMiner Studio 7 จะบอกเพียงหน้าที่และรายละเอียดคร่าวๆของโอเปอเรเตอร์หากต้องการดูรายละเอียดมากกว่านี้ต้องไปที่ Jum to Tutorai Process ซึ่งจะ Link ไปยังเวปไซต์ที่มีรายละเอียดของเกี่ยวกับ Operator ที่ใช้อยู่ เช่น โอเปอเรเตอร์ชื่อ Linear Regression ใน หน้า Help ก็จะบอกว่าเป็นโอเปอเรเตอร์ ที่ใช้คำนวณข้อมูลจากจาก data set โดยใช้วิธี Linear Regression



รูปที่ 2.16 คำสั่ง Help ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

2.22.2 การ Import Data

การ Import Data เป็นการนำข้อมูลเข้ามาใช้งานเพื่อวิเคราะห์หรือสร้าง model ใน RapidMiner โดยข้อมูลที่ Import เข้ามาจะถูกเก็บอยู่ใน “Repository” ซึ่งเป็นศูนย์กลางเก็บข้อมูลและโปรเซสใน RapidMiner เพื่ออำนวยความสะดวกที่จะไม่ต้องโหลดข้อมูลทุกครั้ง ก่อนที่จะนำข้อมูลเข้าจะต้องเตรียมข้อมูลก่อน

1. Arrival Time Database

ข้อมูลที่จะนำเข้าเป็นข้อมูล เวลาที่ใช้เดินทางระหว่างป้าย มีรายละเอียดดังตาราง

	A	B	C	D	E	F	G
1	dayofweek	time5mins	stop1	stop2	linkID	distance	timebtwstops
2	F	231	2860	2013	2860-2013	270.5925952	197.4666667
3	W	189	3614	3616	3614-3616	279.7474122	192.2
4	H	65	390	1522	390-1522	1102.743599	171.2
5	W	142	3989	3990	3989-3990	195.2409243	170.5166667
6	T	49	4003	4004	4003-4004	375.3132785	109.2333333
7	F	96	3596	3597	3596-3597	257.4390034	102.8666667
8	S	204	2860	2013	2860-2013	270.5925952	76.4
9	U	67	3991	3992	3991-3992	269.2960554	67.16666667
10	F	258	3990	3991	3990-3991	218.9085515	63.46666667
11	H	50	3571	3572	3571-3572	463.4337579	60.08333333
12	F	246	3988	3989	3988-3989	277.1124285	58.91666667

รูปที่ 2.17 Arrival Time Databaseในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

- A dayofweek เป็นวันอาทิตย์-วันเสาร์
- B time5mins เป็นช่วงเวลาใน 1 วัน
- C stop1 หมายเลย Link ที่เริ่มเดินทาง
- D stop2 หมายเลย Link ที่รถจอด
- E linkID ต้นและปลาย link หมายถึงช่วงถนนระหว่างป้ายรถ
- F distance หรือระยะทางระหว่างป้าย หน่วยเป็นเมตร
- G timebtstops คือ ระยะเวลาเดินทางระหว่างป้าย หน่วยเป็นนาที

2. คำนิยามเกี่ยวกับ DATA

ข้อมูล แถว(colum) A-G เรียกว่า แอททริบิวต์ (attribute)

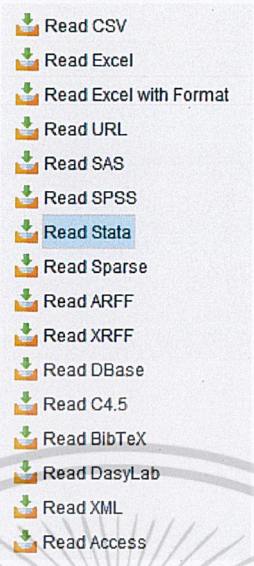
ข้อมูล แนว(row) 1-12 เรียกว่า example

3. ประเภทของข้อมูลแต่ละแอททริบิวต์

- Integer คือ ข้อมูลประเภทตัวเลขจำนวนเต็ม
- Real คือ ข้อมูลประเภทตัวเลขทศนิยม
- Date time คือ ข้อมูลประเภทวันที่และเวลา
- Polynominal คือ ข้อมูลประเภทที่เป็นตัวเลขและมีมากกว่า 2 ตัวเลือก เช่น dayofweek

4. ประเภทข้อมูลที่สามารถ Import เข้ามาใน RapidMiner เพื่อ Process

RapidMiner สามารถใช้ข้อมูลมาโปรเซสถึง 16 ฟอร์เมท ดังรูปด้านล่าง แต่หนึ่งในนั้นที่นิยมนำมาใช้กัน คือ ฟอร์เมทที่เป็น CSV และ Excel ซึ่งในตัวอย่างที่แสดงให้ดูจะใช้ CSV



รูปที่ 2.18 ประเภทข้อมูลที่สามารถ Import เข้ามาใน RapidMiner เพื่อ Process

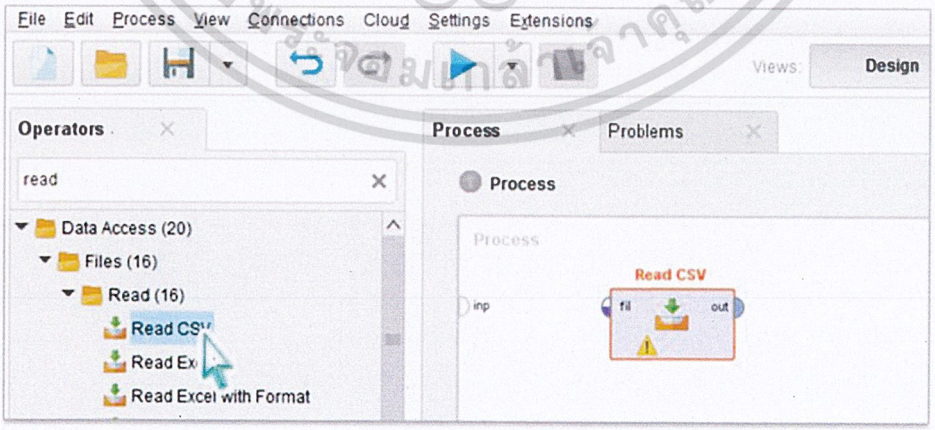
ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

5. Import data

- Import ไฟล์ CSV เข้ามาใน Repository

- ไปที่ค้นหาของเมนู Operator พิมพ์คำว่า read เพื่อเรียกใช้ operator สำหรับ Import ข้อมูล

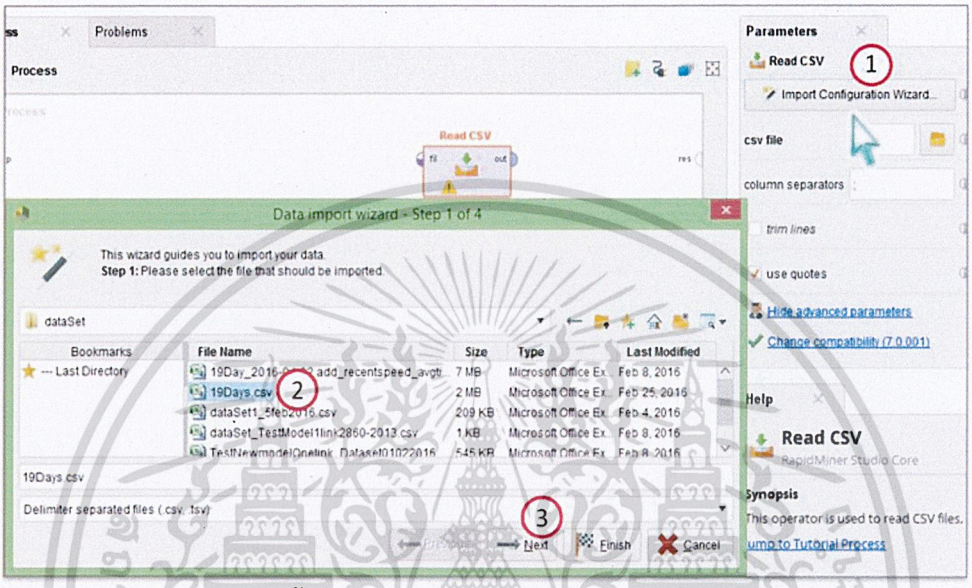
- เลือก Operator ชื่อ Read CSV ลากมาวางที่หน้าต่าง Main



รูปที่ 2.19 Import ไฟล์ CSV เข้ามาใน Repository ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

- คลิกที่ Operator แล้วไปที่หน้าต่าง Parameter เพื่อตั้งค่า Operator “Read CSV”



รูปที่ 2.20 การตั้งค่า Operator ในโปรแกรม RapidMiner Studio 7

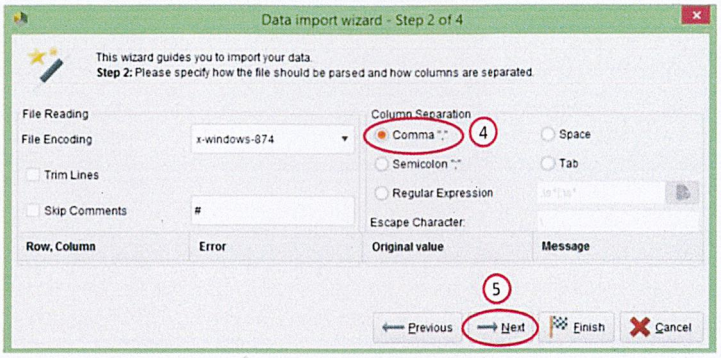
ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

- คลิกที่ 1 Import Configuration Wizard

- เลือกไฟล์ 2 ที่จะ Import เข้ามาใน RapidMiner ซึ่งในที่นี้เราเลือกไฟล์ CSV ชื่อ “19Days”

- คลิก 3 Next

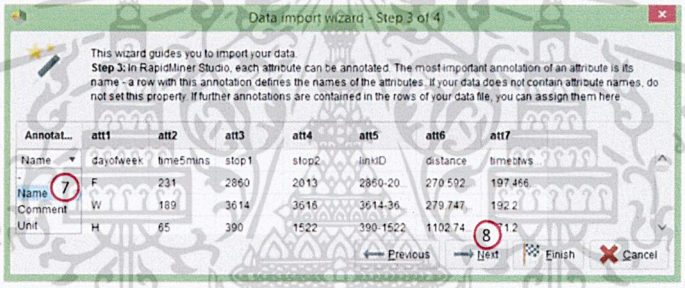
- เลือก 4 Comma และคลิก 5 Next [การแยกข้อมูลระหว่างคอลัมน์นั้นเลือก comma เนื่องจาก ฟอร์แมท CSV คั่นข้อมูลระหว่างคอลัมน์ด้วย comma]



รูปที่ 2.21 การแยกข้อมูลระหว่างคอลัมน์ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

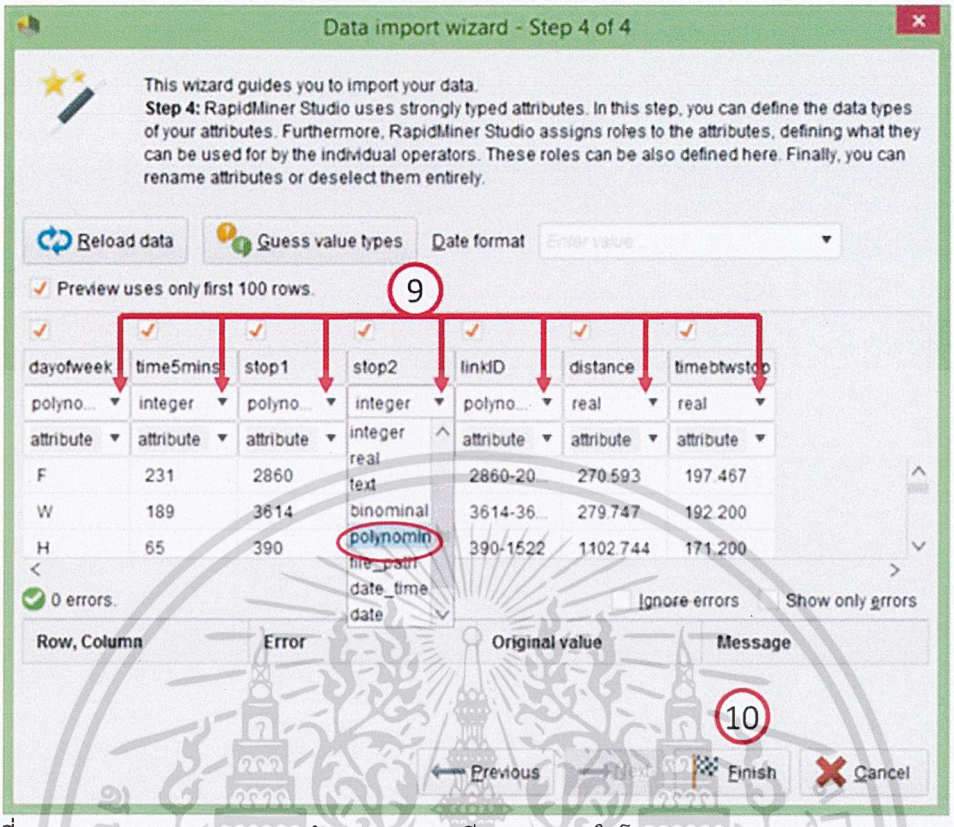
- กำหนดรายละเอียดให้กับแอตทริบิวต์ โดยกำหนดให้แถวแรกเป็นชื่อ 7 ของแอททริบิวต์ และคลิก 8 Next



รูปที่ 2.22 การกำหนดรายละเอียดแอททริบิวต์ ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

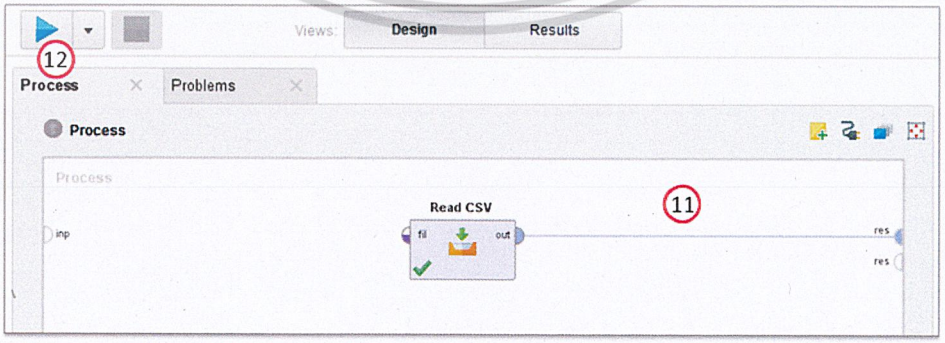
6. ความหมายของการกำหนดรายละเอียดของแถว
   - – คือ กำหนดให้แถวนั้นๆเป็นข้อมูล
   - Name คือ กำหนดแถวนั้นให้เป็นชื่อของแอททริบิวต์
   - Comment คือ กำหนดให้แถวนั้นเป็นการบรรยายรายละเอียดของแอททริบิวต์
   - Unit คือ กำหนดให้แถวนั้นหน่วยของแอททริบิวต์
   - ตรวจสอบ 9 แอททริบิวต์ของข้อมูลทั้งหมดว่าถูกต้องหรือไม่ซึ่งในที่นี้ต้องเปลี่ยนแอททริบิวต์ของ stop1, stop2 เป็นประเภท Polynominal คลิก 10 Finish



รูปที่ 2. 23 ความหมายของการกำหนดรายละเอียดของแถวในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

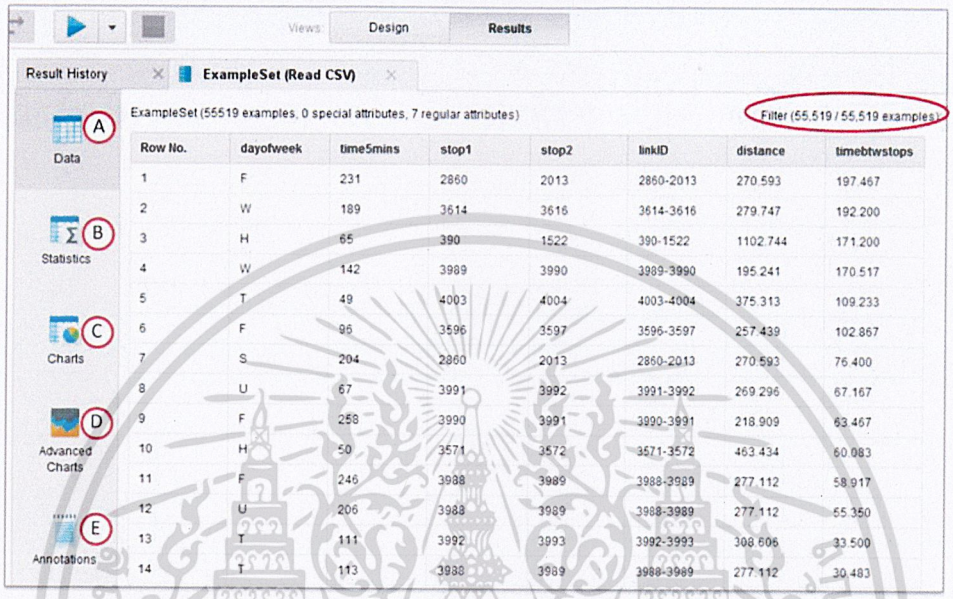
- 11 ลากต่อเส้นจากพอร์ค exa ของโอเปอเรเตอร์ Read CSV ไปยังพอร์ต res คลิก ปุ่ม เพื่อตรวจสอบว่า Import ข้อมูลเข้ามาสำเร็จหรือไม่



รูปที่ 2.24 การตรวจสอบ Import ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

- ข้อมูลที่ Import เข้ามาใน RapidMiner Studio 7 แสดงในหน้าต่าง Results ตรวจสอบได้จากรายละเอียดข้อมูลและจำนวนแถวข้อมูลที่เข้ามา



Row No.	dayofweek	time5mins	stop1	stop2	linkID	distance	timebtwstops
1	F	231	2860	2013	2860-2013	270.593	197.467
2	W	189	3614	3616	3614-3616	279.747	192.200
3	H	65	390	1522	390-1522	1102.744	171.200
4	W	142	3989	3990	3989-3990	195.241	170.517
5	T	49	4003	4004	4003-4004	375.313	109.233
6	F	96	3596	3597	3596-3597	257.439	102.867
7	S	204	2860	2013	2860-2013	270.593	76.400
8	U	67	3991	3992	3991-3992	269.296	67.167
9	F	258	3990	3991	3990-3991	218.909	63.467
10	H	50	3571	3572	3571-3572	463.434	60.083
11	F	246	3988	3989	3988-3989	277.112	58.917
12	U	206	3988	3989	3988-3989	277.112	55.350
13	T	111	3992	3993	3992-3993	308.606	33.500
14	T	113	3988	3989	3988-3989	277.112	30.483

รูปที่ 2.25 ข้อมูลที่ Import ในโปรแกรม RapidMiner Studio 7

ที่มา: https://z-p3-lookaside.fbs.com/file/SlideWeek8.pdf

ข้อมูลที่แสดงที่หน้าต่าง Result หรือข้อมูลที่ได้จากเอาท์พุทของโอเปอเรเตอร์แสดง 5 รูปแบบ คือ

A. แสดงข้อมูลในรูปแบบตาราง(Data) ซึ่งเป็นค่าเริ่มต้นในการแสดงหลังจาก Run เสร็จ หรือเป็นค่าเริ่มต้น

B. แสดงข้อมูลในรูปของค่าทางสถิติ(Statistic) ที่สรุปมาเป็นค่า Min, Max, Average, Least, Most

C. แสดงกราฟแบบต่างๆ สามารถเลือกชนิดกราฟได้

D. แสดงกราฟแบบต่างๆ สามารถตั้งค่าการพล๊อตกราฟเองได้ โดยสามารถปรับสีและ Font เองได้

E. แสดงรายละเอียดของข้อมูล เช่น ข้อมูลที่ Import เข้ามา ได้มาจากแหล่งข้อมูลแหล่งใด ยกตัวอย่างเช่น D:\suporn\bus checkin\algorithm_predictTravel time\dataSet\19Days.csv

จากความรู้พื้นฐานในบทนี้จะนำมาใช้ในการจำแนกข้อมูลที่เก็บรวบรวมมาและนำมาวิเคราะห์พฤติกรรมการซื้อรถยนต์ ซึ่งบทถัดไปจะอธิบายถึงวิธีการดำเนินการทำปัญหาพิเศษเล่มนี้

บทที่ 3

วิธีดำเนินการทำปัญหาพิเศษ

ปัญหาพิเศษนี้จะกล่าวถึงขั้นตอนการทำงานและเทคนิคการใช้เหมืองข้อมูลมาช่วยในการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ โดยมีขั้นตอนการดำเนินการทำปัญหาพิเศษดังนี้

3.1 การกำหนดประชากรและเลือกกลุ่มตัวอย่าง

3.1.1 ประชากรที่ใช้ในการวิจัย

ประชากรที่ใช้ในการวิจัย คือ ผู้ที่มีอายุ 18 ปีขึ้นไป เนื่องจากสามารถมีใบอนุญาติขับขี่รถยนต์ และอาศัยอยู่ในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง และจังหัวฉะเชิงเทรา

3.1.2 กลุ่มตัวอย่างที่ใช้ในการวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คือ ผู้ที่มีอายุ 18 ปีขึ้นไป ที่มีรถยนต์หรือไม่มีรถยนต์อยู่ในครอบครองในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง และจังหัวฉะเชิงเทรา เนื่องจากไม่ทราบจำนวนประชากรที่แท้จริง จึงกำหนดขนาดกลุ่มตัวอย่างจากสูตรการคำนวณแบบไม่ทราบประชากร (กัลยา วานิชบัญชา. 2549: 25-56) ที่ระดับความเชื่อมั่น 95% คือ ยอมให้เกิดความคลาดเคลื่อนในกลุ่มตัวอย่างไม่เกินร้อยละ 5 ได้ขนาดกลุ่มตัวอย่าง 385 ตัวอย่าง เพื่อให้การเก็บข้อมูลของแบบสอบถามสมบูรณ์ จึงเพิ่มจำนวนตัวอย่างอีก 65 ตัวอย่าง ดังนั้น ขนาดกลุ่มตัวอย่างในการทำวิจัยครั้งนี้ คือ 450 ตัวอย่าง โดยใช้สูตรการคำนวณ ดังนี้

$$n = \frac{Z^2}{4e^2}$$

เมื่อ n คือ ขนาดตัวอย่าง

Z คือ ระดับความเชื่อมั่น ที่ผู้วิจัยกำหนดไว้ คือ 95% ดังนั้น จะมีค่า $Z_{1-\frac{\alpha}{2}}$ หรือ $Z_{.975}$ = 1.96

e คือ ความคลาดเคลื่อนที่จะยอมให้เกิดขึ้นได้ โดยกำหนดให้ค่าความคลาดเคลื่อน 5% (e = 0.05)

เมื่อแทนค่า จะได้ $n = \frac{(1.96)^2}{4(0.05)^2}$ = 385 ตัวอย่าง

จากการคำนวณกลุ่มตัวอย่างเท่ากับ 385 ตัวอย่าง และเก็บจำนวนตัวอย่างเพิ่มอีก 65 ตัวอย่าง รวมเป็นจำนวนตัวอย่างทั้งหมด 450 ตัวอย่าง โดยใช้วิธีการสุ่มตัวอย่างแบบเจาะจง (Purposive Sampling) โดยเจาะจงผู้ที่มีอายุ 18 ปีขึ้นไป ที่มีรถยนต์หรือไม่มีรถยนต์อยู่ในครอบครองในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา และวิธีการสุ่มตัวอย่างแบบโควตา (Quota Sampling) และวิธีการสุ่มตัวอย่างตามความสะดวก (Convenience Sampling)

3.2 การสร้างแบบสอบถาม

จากการที่คณะผู้จัดทำได้เลือกหัวข้อเรื่องการใช้เหมืองข้อมูลมาช่วยในการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ จึงต้องสร้างแบบสอบถามเพื่อสำรวจข้อมูลส่วนตัวและพฤติกรรมการใช้รถยนต์จำนวน 450 ตัวอย่าง ในบริเวณสถาบันเทคโนโลยีพระจอมเกล้า-เจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา ประกอบด้วยข้อมูลส่วนตัวและพฤติกรรมการใช้รถยนต์ โดยแบ่งประเภทของรถยนต์เป็น 3 ประเภท คือ รถเก๋ง รถกระบะ รถเอนกประสงค์และไม่ใช้รถยนต์ โดยรถยนต์แต่ละประเภทนั้นจะบอกเพียงประเภทของรถยนต์ ไม่ได้บ่งบอกถึงมูลค่าของรถยนต์ประเภทนั้น ๆ จากนั้นทำการสร้างตัวแบบพยากรณ์ (Model) โดยมีการเก็บข้อมูลต่าง ๆ ดังนี้

ตารางที่ 3.1 คุณลักษณะที่ใช้ในการสร้างตัวแบบพยากรณ์

ประวัติส่วนตัว	พฤติกรรมการใช้รถยนต์
1. เพศ	9. ลักษณะการใช้งานรถ
2. อายุ	10. ความเร็วเฉลี่ยในการขับขี่
3. สถานภาพ	11. ระยะทางจากบ้านถึงสถานที่ทำงาน/
4. ระดับการศึกษา	สถานศึกษา
5. อาชีพ	12. ค่าใช้จ่ายในการจอดรถเฉลี่ย

ตารางที่ 3.1 คุณลักษณะที่ใช้ในการสร้างตัวแบบพยากรณ์ (ต่อ)

ประวัติส่วนตัว	พฤติกรรมการใช้รถยนต์
6. รายได้ต่อเดือน 7. จำนวนสมาชิกในครอบครัว 8. ใบอนุญาตในการขับขี่รถ	13.ค่าบำรุงรักษารถต่อปี

โดยคำถามแต่ละข้อ มีลักษณะดังนี้

1. เพศ ใช้ระดับการวัดข้อมูลแบบนามบัญญัติ (Nominal Scale)

2. อายุ ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale) ผู้ตอบแบบสอบถามมีอายุระหว่าง 18-75 ปี เนื่องจากผู้ที่สามารถมีใบอนุญาตขับขี่รถยนต์ต้องมีอายุตั้งแต่ 18 ปีบริบูรณ์ ซึ่งเป็นช่วงเริ่มต้นต่ำสุดของข้อมูล และช่วงสูงสุดของข้อมูล คือ 75 ปี พร้อมทั้งกำหนดช่วงปลายเปิด และกำหนดจำนวนชั้นของข้อมูล คือ 6 ชั้น ดังนั้น ความกว้างของอันตรภาคชั้น มีดังนี้

ความกว้างอันตรภาคชั้น = (ข้อมูลที่มีค่าสูงสุด – ข้อมูลที่มีค่าต่ำสุด)/จำนวนชั้น

= (75-18)/6

= 9 ปี

จากการคำนวณช่วงอายุข้างต้น สามารถแบ่งช่วงอายุของกลุ่มตัวอย่าง ดังนี้

1. 18-26 ปี
2. 27-35 ปี
3. 36-44 ปี
4. 45-53 ปี
5. มากกว่า 53 ปี

3. สถานภาพ ใช้ระดับการวัดข้อมูลแบบนามบัญญัติ (Nominal Scale)

4. ระดับการศึกษา ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale)

5. อาชีพ ใช้ระดับการวัดข้อมูลแบบนามบัญญัติ (Nominal Scale)

6. รายได้ต่อเดือนต่อคน ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale) โดยคณะผู้จัดทำสนใจผู้ตอบแบบสอบถามที่มีรายได้ 10,000 ถึง 70,000 บาท พร้อมทั้งกำหนดช่วงปลายเปิด และกำหนดจำนวนชั้นของข้อมูล คือ 6 ชั้น ดังนั้น ความกว้างของอันตรภาคชั้น มีดังนี้

ความกว้างอันตรภาคชั้น = (ข้อมูลที่มีค่าสูงสุด – ข้อมูลที่มีค่าต่ำสุด)/จำนวนชั้น

= (70,000 – 10,000)/6

= 10,000 บาท

จากการคำนวณช่วงรายได้ข้างต้น สามารถแบ่งช่วงรายได้ของกลุ่มตัวอย่าง ดังนี้

1. น้อยกว่าหรือเท่ากับ 10,000 บาท

2. 10,001 – 20,000 บาท

3. 20,001 – 30,000 บาท

4. 30,001 – 40,000 บาท

5. 40,001 – 50,000 บาท

6. มากกว่า 50,000 บาท

7. จำนวนสมาชิกในครอบครัวที่เป็นสายเลือดเดียวกันหรืออาศัยอยู่ในที่เดียวกัน ใช้ระดับการวัดข้อมูลแบบระดับอัตราส่วน (Ratio Scale)

8. ใบอนุญาติการขับขี่รถ ใช้ระดับการวัดข้อมูลแบบนามบัญญัติ (Nominal Scale)

9. ลักษณะการใช้งานรถ ใช้ระดับการวัดข้อมูลแบบนามบัญญัติ (Nominal Scale)

10. ความเร็วเฉลี่ยในการขับขี่ ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale) ผู้ตอบแบบสอบถามใช้ความเร็วในการขับขี่รถยนต์ระหว่าง 10-180 กิโลเมตรต่อชั่วโมง โดยคณะผู้จัดทำสนใจผู้ตอบแบบสอบถามที่ใช้ความเร็วในการขับขี่รถยนต์ระหว่าง 90-140 กิโลเมตรต่อชั่วโมง พร้อมทั้งกำหนดช่วงปลายเปิด และกำหนดจำนวนชั้นของข้อมูล คือ 5 ชั้น ดังนั้น ความกว้างของอันตรภาคชั้น มีดังนี้

ความกว้างอันตรภาคชั้น = (ข้อมูลที่มีค่าสูงสุด – ข้อมูลที่มีค่าต่ำสุด)/จำนวนชั้น

= (140 - 90)/5

= 10 กิโลเมตรต่อชั่วโมง

จากการคำนวณช่วงความเร็วข้างต้น สามารถแบ่งช่วงความเร็วในการขับขี่รถยนต์ของกลุ่มตัวอย่าง ดังนี้

1. น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง

2. 91-100 กิโลเมตรต่อชั่วโมง

3. 101-110 กิโลเมตรต่อชั่วโมง

4. 111-120 กิโลเมตรต่อชั่วโมง

5. มากกว่า 120 กิโลเมตรต่อชั่วโมง

11. ระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale) โดยคณะผู้จัดทำสนใจผู้ตอบแบบสอบถามที่ระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา ระหว่าง 10-40 กิโลเมตร พร้อมทั้งกำหนดช่วงปลายเปิดและกำหนดจำนวนชั้นของข้อมูล คือ 3 ชั้น ดังนั้น ความกว้างของอันตรภาคชั้น มีดังนี้

ความกว้างอันตรภาคชั้น = (ข้อมูลที่มีค่าสูงสุด – ข้อมูลที่มีค่าต่ำสุด)/จำนวนชั้น

= (40 - 10)/3

= 10 กิโลเมตร

จากการคำนวณช่วงระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษาข้างต้น สามารถแบ่งช่วงระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา ของกลุ่มตัวอย่าง ดังนี้

1. น้อยกว่าหรือเท่ากับ 10 กิโลเมตร

2. 11-20 กิโลเมตร

3. มากกว่า 20 กิโลเมตร

12. ค่าใช้จ่ายในการจอดรถเฉลี่ยต่อเดือน ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale)

13. ค่าบำรุงรักษาต่อปีต่อคัน ใช้ระดับการวัดข้อมูลประเภทเรียงลำดับ (Ordinal Scale)

ในการทำปัญหาพิเศษนี้ ทางคณะผู้จัดทำได้สร้างแบบสอบถามสำรวจข้อมูลส่วนตัวและพฤติกรรมการใช้รถยนต์ในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา จำนวน 450 ตัวอย่าง โดยข้อมูลที่จะใช้ประกอบการพิจารณาะมีทั้งหมด 18 แอททริบิวต์ ประกอบด้วยแอททริบิวต์ ดังนี้ คือ ประเภทของรถยนต์ เพศ อายุ สถานภาพ ระดับการศึกษา อาชีพรายได้ต่อเดือน จำนวนสมาชิกในครอบครัว ใบอนุญาตในการขับขี่รถ ลักษณะการใช้งานรถที่ประกอบด้วย การบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน ไม่จำเป็นต้องใช้และอื่น ๆ ความเร็วเฉลี่ยในการขับขี่ ระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา ค่าใช้จ่ายในการจอดรถเฉลี่ยและค่าบำรุงรักษารถต่อปี โดยทำการรวบรวมข้อมูลจากแบบสอบถามมาให้อยู่ในรูปของฐานข้อมูลแบบเดียวกัน ซึ่งแบบสอบถามมีลักษณะดังนี้

แบบสอบถาม

เรื่อง การวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์

คำชี้แจง

ข้อมูลในแบบสอบถามนี้ใช้เพื่อประกอบการทำปัญหาพิเศษเรื่องการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ ของสาขาวิชาคณิตศาสตร์ประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โปรดตอบคำถามและให้ข้อมูลของท่านตามความจริงเพื่อประโยชน์ในการทำการศึกษาปัญหาพิเศษ และข้อมูลของท่านจะถูกเก็บไว้เป็นความลับ

คำถามตรวจสอบข้อมูล

ปัจจุบันท่านใช้รถยนต์ประเภทใดในชีวิตประจำวัน

☐ รถเก๋ง ☐ รถกระบะ

☐ รถเอนกประสงค์ (SUV) ☐ ไม่ใช้

ส่วนที่ 1 ข้อมูลทั่วไป

1. เพศ

☐ ชาย ☐ หญิง

2. อายุ

☐ 18-26 ปี ☐ 27-35 ปี ☐ 36-44 ปี

☐ 45-53 ปี ☐ > 53 ปีขึ้นไป

3. สถานภาพ

☐ โสด ☐ สมรส/อยู่ด้วยกัน

☐ หย่าร้าง/แยกกันอยู่ ☐ หม้าย

4. ระดับการศึกษา

☐ มัธยมศึกษาตอนต้นหรือต่ำกว่า

☐ มัธยมศึกษาตอนปลายหรือ ปวช.

☐ ปริญญาตรี

☐ สูงกว่าปริญญาตรี

5. อาชีพ

☐ ข้าราชการ/เจ้าหน้าที่ของรัฐ ☐ พนักงานบริษัทเอกชน

☐ พนักงานรัฐวิสาหกิจ ☐ เจ้าของธุรกิจ/ประกอบอาชีพอิสระ

☐ กำลังศึกษาอยู่ ☐ อื่นๆ (โปรดระบุ).....

6. รายได้ต่อเดือน

☐ ≤10,000 บาท ☐ 10,001-20,000 บาท

☐ 20,001-30,000 บาท ☐ 30,001-40,000 บาท

☐ 40,001-50,000 บาท ☐ > 50,000 บาทขึ้นไป

7. จำนวนสมาชิกในครอบครัว(รวมตัวท่านด้วย).....

8. ใบอนุญาตในการขับขี่รถ ☐ มีและพกพา ☐ มีแต่ไม่ได้พกพา

☐ ไม่มี

9. ลักษณะการใช้งานรถ (เลือกได้มากกว่า 1 ประเภท)

☐ บรรทุกของ ☐ ท่องเที่ยวผจญภัยต่างจังหวัด

☐ เดินทางไปทำงาน ☐ ไม่จำเป็นต้องใช้

☐ อื่นๆ (โปรดระบุ).....

10.ความเร็วเฉลี่ยในการขับขี่

☐ ≤ 90 กม./ชม. ☐ 91-100 กม./ชม.

☐ 101-110 กม./ชม. ☐111-120 กม./ชม.

☐ > 120 กม./ชม.

11. ระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา

☐≤ 10 กิโลเมตร ☐11-20 กิโลเมตร

☐ >20 กิโลเมตร

12. ค่าใช้จ่ายในการจอดรถเฉลี่ย

☐ ไม่มี ☐ มี (โปรดระบุ).....บาทต่อเดือน

13.ค่าบำรุงรักษารถต่อปี

☐ ไม่ได้เป็นผู้เสียค่าบำรุงรักษาด้วยตัวเอง

☐ มี (โปรดระบุ).....บาท

ทางคณะผู้จัดทำขอขอบพระคุณอย่างยิ่งที่กรุณาสละเวลาอันมีค่ามาทำแบบสอบถาม

3.3 การเตรียมข้อมูล

หลังจากการสร้างแบบสอบถามและเก็บข้อมูล ต้องมีการเตรียมข้อมูลที่จะนำไปใช้ โดยการเตรียมข้อมูลให้เป็นมาตรฐานเดียวกันและลดความหลากหลายของข้อมูล เพื่อให้โมเดลมีประสิทธิภาพสูงสุด ได้แก่ การโอนย้ายข้อมูล (Data Transfer) และการทำความสะอาดข้อมูล (Data Cleaning) โดยทางคณะผู้จัดทำเลือกในโปรแกรม Microsoft Excel เป็นแหล่งจัดเก็บข้อมูล โดยวิธีการทำความสะอาดข้อมูล คือ หากข้อมูลขาดหายเป็นประเภทตัวเลขจะใช้การตัดทิ้งข้อมูลนั้นๆ หากข้อมูลขาดหายเป็นประเภทข้อความ จะให้การเติมข้อมูลสูญหายด้วยสัญลักษณ์ “N/A” หรือ “?” ลงไปในข้อมูลที่ขาดหาย (นิตยา เกิดประสพ; และคณะ. 2546: 3)

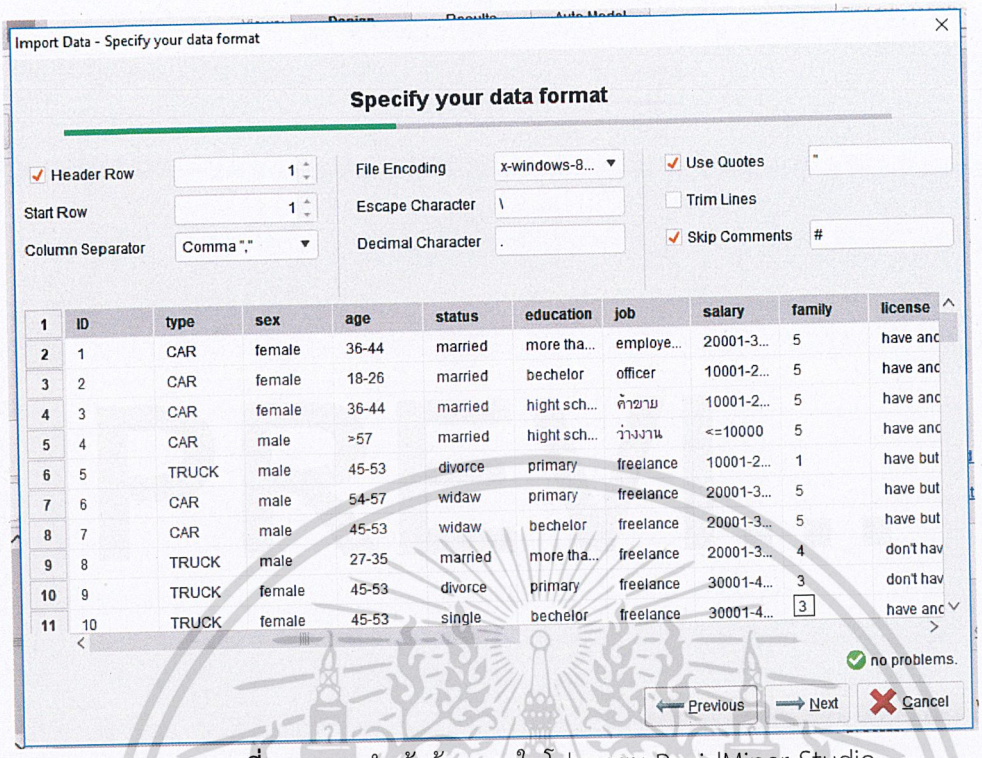
การคัดเลือกข้อมูลที่นำมาทำเหมืองข้อมูล เพื่อให้ได้ประสิทธิภาพสูงสุดจะต้องทำการเตรียมข้อมูล 2 ขั้นตอน คือ การโอนย้ายข้อมูล (Data Transfer) และการทำความสะอาดข้อมูล (Data Cleaning)

3.3.1 การโอนย้ายข้อมูล (Data Transfer)

การโอนย้ายข้อมูล เป็นการย้ายข้อมูลเดิมที่สร้างและจัดเก็บอยู่ในรูปแบบไฟล์ Excel ที่ได้จากการกรอกข้อมูลจากแบบสอบถามที่ผู้จัดทำแต่ละคนได้ทำการสำรวจให้มารวมกัน และบันทึกข้อมูลให้อยู่ในรูปแบบไฟล์ .CSV เพื่อให้สามารถนำไปใช้งานได้และสะดวกในการนำเข้าข้อมูลในโปรแกรม RapidMiner Studio

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	ID	type	sex	age	status	education	job	salary	family	license	carry	travel	work	dispenable	other	speed	distance	parking	maintain
2	1	CAR	female	36-44	married	more than bechelor	employees	20k-30k	5	have and can	no	yes	yes	no	no	121-130	<=10	don't have	don't have
3	2	CAR	female	18-26	married	bechelor	officer	10k-20k	5	have and can	no	yes	yes	no	no	<=90	10-20	don't have	9000
4	3	CAR	female	36-44	married	hight school	ค้าขาย	10k-20k	5	have and can	no	yes	yes	no	no	<=90	<=10	don't have	don't have
5	4	CAR	male	>57	married	hight school	ว่างงาน	<=10k	5	have and can	no	no	no	yes	จ่ายตล	91-100	<=10	don't have	don't have
6	5	TRUC	male	45-53	divorce	primary	freelance	10k-20k	1	have but not	no	yes	yes	no	no	101-120	10-20	don't have	19000
7	6	CAR	male	54-57	widaw	primary	freelance	20k-30k	5	have but not	no	no	yes	no	no	91-100	10-20	don't have	10000
8	7	CAR	male	45-53	widaw	bechelor	freelance	20k-30k	5	have but not	no	yes	no	no	no	101-120	>30	don't have	15000
9	8	TRUC	male	27-35	married	more than bechelor	freelance	20k-30k	4	don't have	yes	yes	yes	no	no	101-120	10-20	don't have	12000
10	9	TRUC	female	45-53	divorce	primary	freelance	30k-40k	3	don't have	yes	no	yes	no	no	91-100	>30	don't have	17000
11	10	TRUC	female	45-53	single	bechelor	freelance	30k-40k	6	have and can	yes	no	yes	no	no	91-100	<=10	don't have	19000
12	11	TRUC	male	27-35	single	bechelor	officer	10k-20k	6	have and can	no	yes	yes	no	no	<=90	>30	don't have	15000
13	12	TRUC	male	45-53	divorce	bechelor	officer	30k-40k	6	don't have	no	yes	yes	no	no	101-120	<=10	don't have	17000
14	13	TRUC	male	45-53	single	more than bechelor	officer	20k-30k	6	don't have	yes	yes	yes	no	no	91-100	>30	don't have	9000
15	14	TRUC	male	27-35	married	more than bechelor	officer	30k-40k	4	have and can	yes	yes	yes	no	no	101-120	>30	don't have	don't have
16	15	CAR	female	18-26	single	more than bechelor	study	<=10k	3	have but not	no	yes	yes	no	no	101-120	>30	500	don't have
17	16	TRUC	male	18-26	single	more than bechelor	private employee	10k-20k	6	have but not	yes	yes	yes	no	no	<=90	>30	don't have	15000
18	17	TRUC	male	18-26	married	more than bechelor	employees	20k-30k	2	have and can	no	yes	yes	no	no	<=90	>30	don't have	11000
19	18	TRUC	female	18-26	single	bechelor	freelance	10k-20k	6	have but not	yes	yes	yes	no	no	121-130	>30	don't have	don't have
20	19	TRUC	female	18-26	single	bechelor	freelance	10k-20k	4	don't have	yes	yes	yes	no	no	121-130	>30	don't have	19000
21	20	TRUC	female	36-44	married	primary	freelance	10k-20k	3	don't have	yes	yes	yes	no	no	>30	<=90	don't have	don't have
22	21	SUV	male	27-35	married	primary	ค้าขาย	30k-40k	5	don't have	no	yes	yes	no	no	121-130	>30	don't have	don't have

รูปที่ 3.1 ตัวอย่างการรวมรวมข้อมูลให้มาอยู่ในฐานข้อมูลเดียวกัน



1	ID	type	sex	age	status	education	job	salary	family	license
2	1	CAR	female	36-44	married	more tha...	employe...	20001-3...	5	have anc
3	2	CAR	female	18-26	married	bechelor	officer	10001-2...	5	have anc
4	3	CAR	female	36-44	married	hight sch...	ค้าขาย	10001-2...	5	have anc
5	4	CAR	male	>57	married	hight sch...	ว่างงาน	<=10000	5	have anc
6	5	TRUCK	male	45-53	divorce	primary	freelance	10001-2...	1	have but
7	6	CAR	male	54-57	widaw	primary	freelance	20001-3...	5	have but
8	7	CAR	male	45-53	widaw	bechelor	freelance	20001-3...	5	have but
9	8	TRUCK	male	27-35	married	more tha...	freelance	20001-3...	4	don't hav
10	9	TRUCK	female	45-53	divorce	primary	freelance	30001-4...	3	don't hav
11	10	TRUCK	female	45-53	single	bechelor	freelance	30001-4...	3	have anc

รูปที่ 3.2 การนำเข้าข้อมูลลงในโปรแกรม RapidMiner Studio

3.3.2 การทำความสะอาดข้อมูล (Data Cleaning)

การแปลงค่าให้มีความหมายเดียวกันและมาตรฐานเหมือนกัน มีวิธีการทำดังนี้

ตารางที่ 3.2 การแปลงค่าของข้อมูล

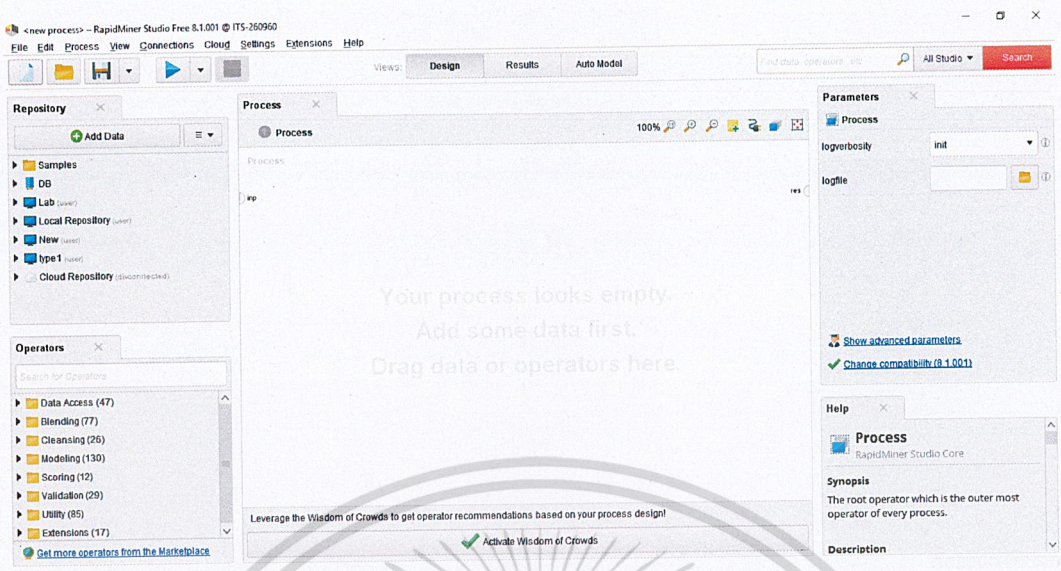
ข้อมูลเดิม	ข้อมูลใหม่
ประเภทของรถยนต์	type
เพศ	Sex
อายุ	Age
สถานภาพ	status
ระดับการศึกษา	education
อาชีพ	Job
รายได้ต่อเดือน	salary
จำนวนสมาชิกในครอบครัว	family
ใบอนุญาตในการขับขี่รถ	license

ตารางที่ 3.2 การแปลงค่าของข้อมูล (ต่อ)

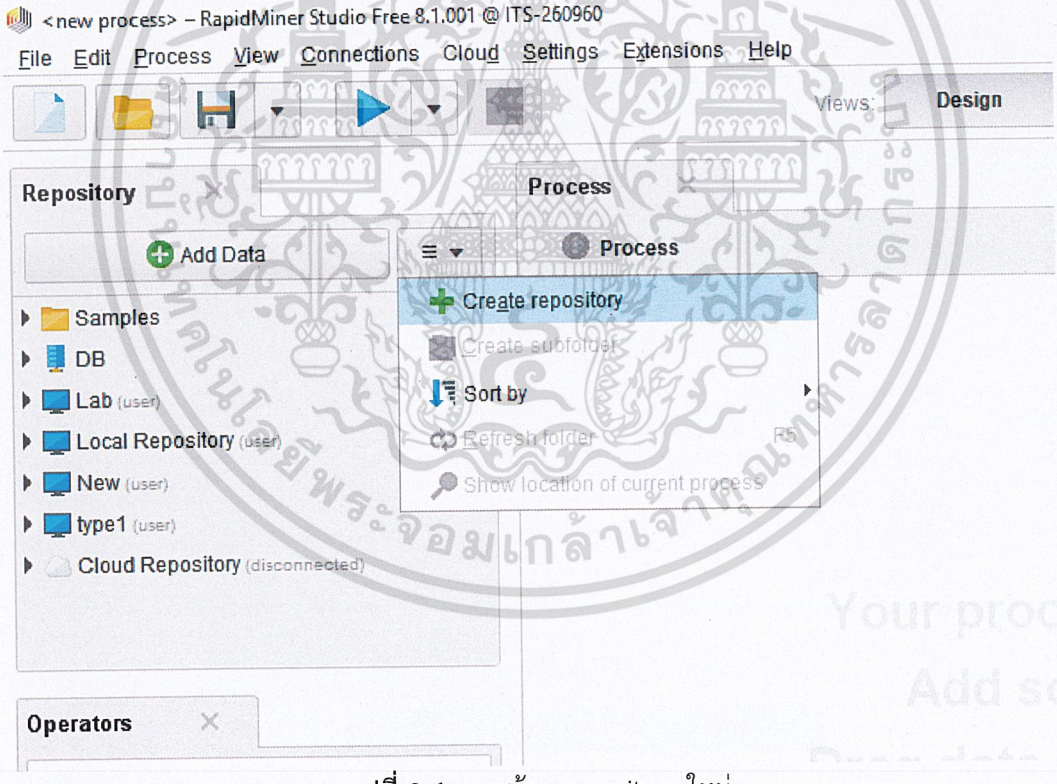
ข้อมูลเดิม	ข้อมูลใหม่
ลักษณะการใช้งานรถ - การบรรทุกของ - ท่องเที่ยวผจญภัยต่างจังหวัด - เดินทางไปทำงาน - ไม่จำเป็นต้องใช้ - อื่น ๆ	 carry travel work dispensable other
ความเร็วเฉลี่ยในการขับขี่	speed
ระยะทางจากบ้านถึงสถานที่ทำงาน/สถานศึกษา	distance
ค่าใช้จ่ายในการจอดรถเฉลี่ย	parking
ค่าบำรุงรักษารถต่อปี	maintain

นอกจากการแปลงค่าข้อมูล ทางคณะผู้จัดทำได้มีการกำจัดข้อมูลที่มีค่าไม่ตรงกันผ่านโปรแกรม RapidMiner Studio โดยใช้โอเปอร์เรเตอร์ Replace, FilterExample และ Replace Missing Value โดยมีขั้นตอน ดังนี้

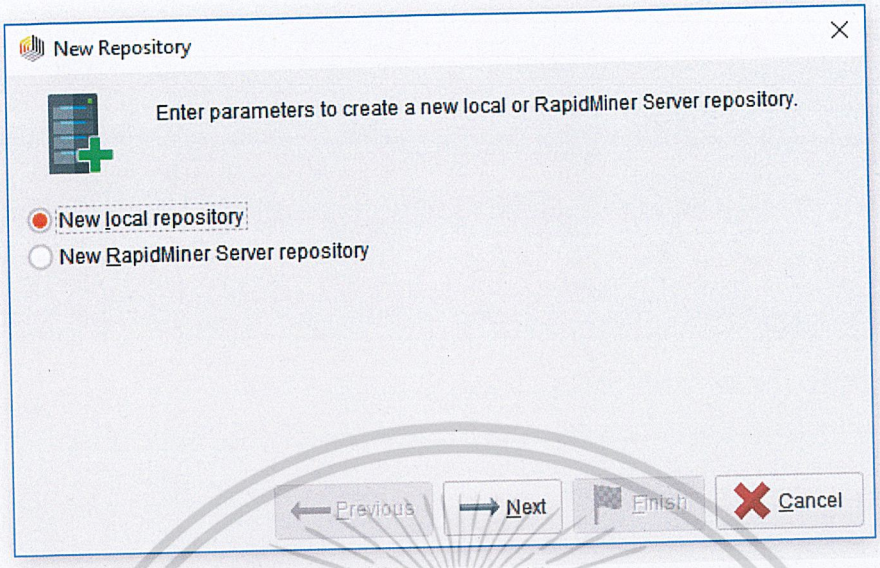
ขั้นตอนที่ 1 เริ่มจากเปิดโปรแกรม RapidMiner Studio จะปรากฏหน้าต่างดัง**รูปที่ 3.3** หน้าโปรแกรม RapidMiner Studio จากนั้นทำการสร้าง repository ใหม่โดยการคลิกที่ปุ่มที่อยู่ข้างปุ่ม Add data แล้วเลือก Create repository ดัง**รูปที่ 3.4** การสร้าง repository ใหม่ เมื่อกดแล้วจะปรากฏหน้าต่าง New Repository เลือก New local repository แล้วกด Next ดัง**รูปที่ 3.5** การเลือกตำแหน่งที่เก็บ repository



รูปที่ 3.3 หน้าโปรแกรม RapidMier Studio

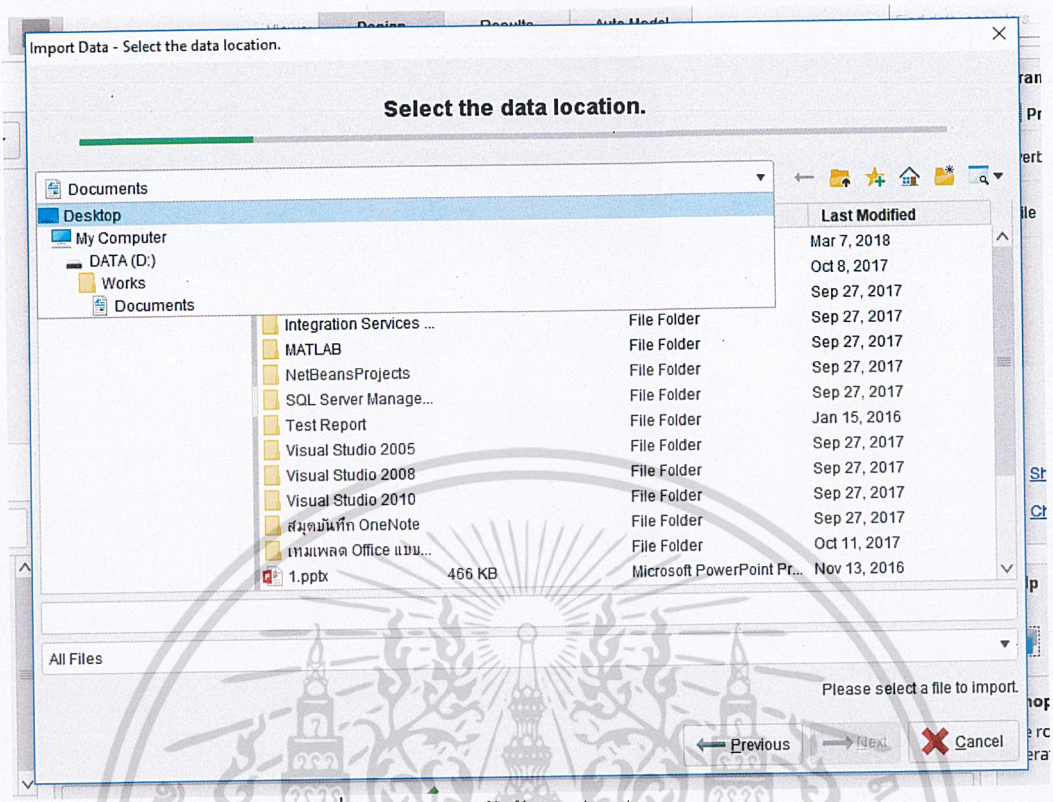


รูปที่ 3.4 การสร้าง repository ใหม่

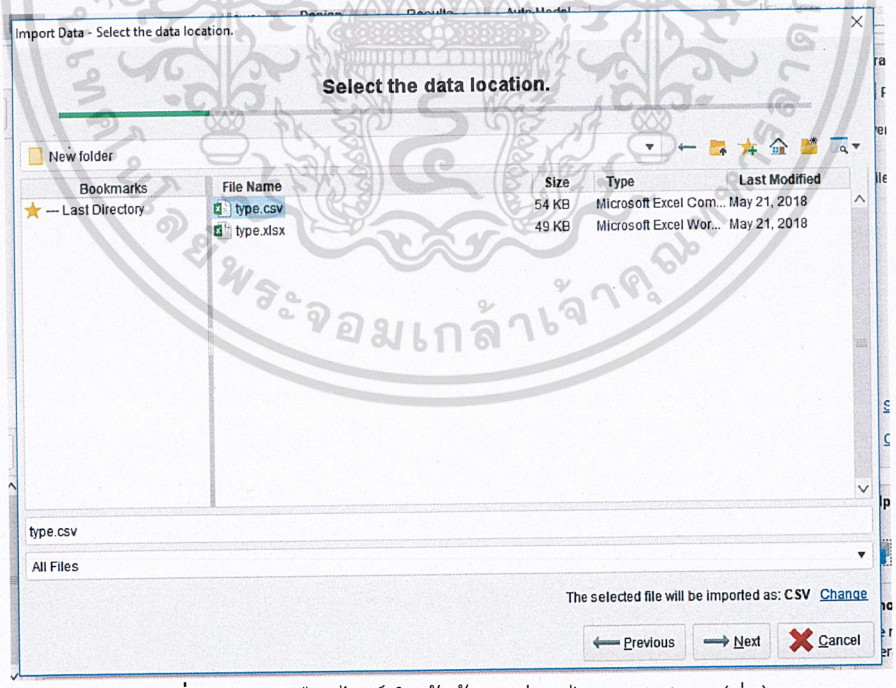


รูปที่ 3.5 การเลือกตำแหน่งที่เก็บ repository

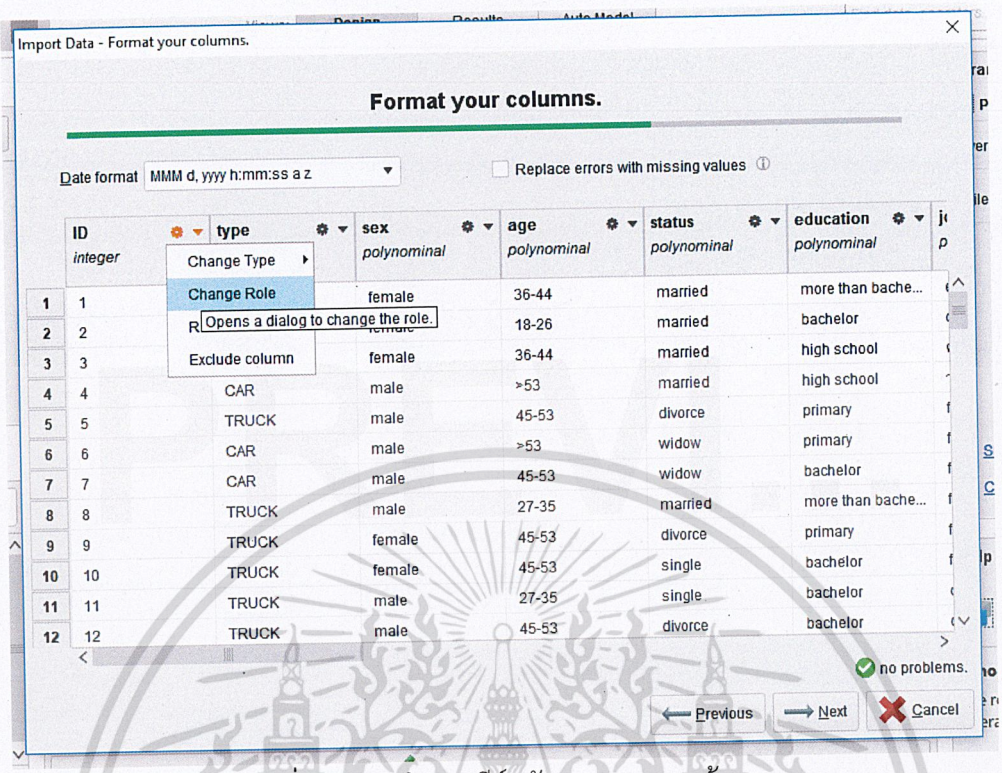
ขั้นตอนที่ 2 ทำการนำเข้าข้อมูลเข้าสู่โปรแกรมโดยคลิกที่ปุ่ม Add Data จากนั้นจะปรากฏหน้าต่าง Import Data – Select the data location เพื่อเลือกข้อมูลที่ต้องการนำเข้า ดัง**รูปที่ 3.6** การนำเข้าข้อมูลผ่านปุ่ม Add data ต่อมาจะเป็นการกำหนดคลาสของข้อมูล โดยการคลิกที่รูปฟันเฟืองที่อยู่แถวแรกของข้อมูล แล้วเลือก Change Role จากนั้นจะปรากฏหน้าต่าง Change Role พิมพ์ id เมื่อต้องการให้แอททริบิวต์นั้นเป็นคีย์หลัก และพิมพ์ label เมื่อต้องการให้แอททริบิวต์นั้นเป็นคลาสของข้อมูล ดัง**รูปที่ 3.7** การกำหนดคีย์หลักและคลาสของข้อมูล และ**รูปที่ 3.9** การกำหนดคลาสให้กับข้อมูล หลังจากนั้นทำการเลือกแหล่งที่ต้องการเก็บข้อมูล โดยในปัญหาพิเศษนี้จะเก็บข้อมูลไว้ที่ typrcar ดัง**รูปที่ 3.10** การเลือกที่เก็บข้อมูลนำเข้า



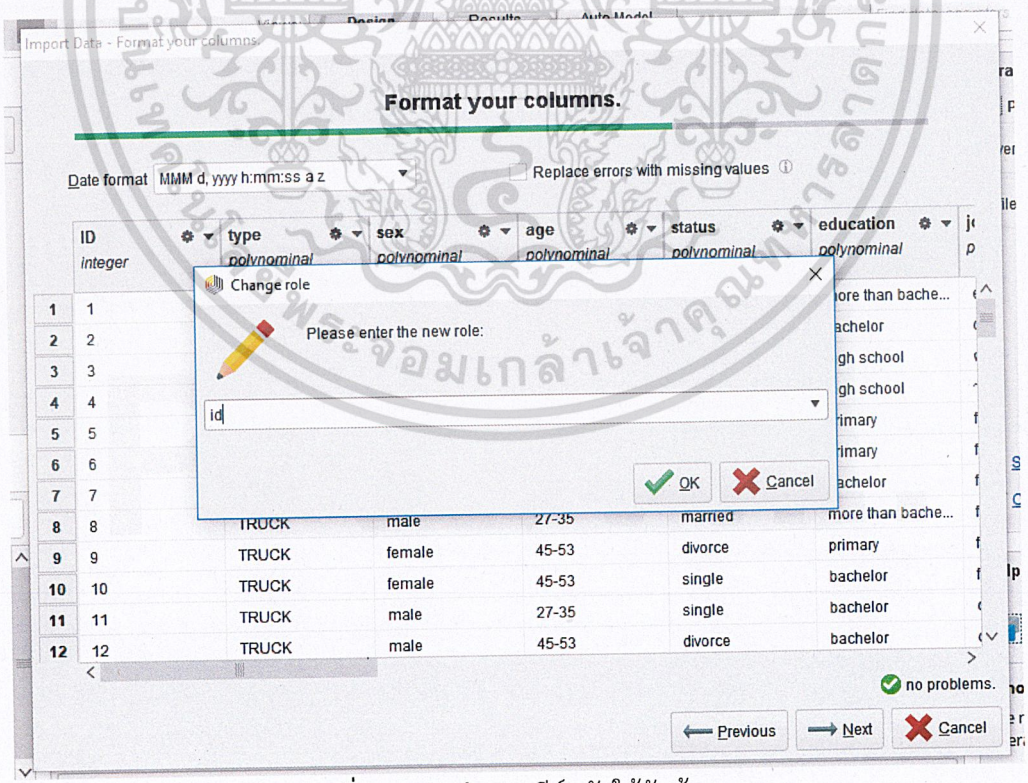
รูปที่ 3.6 การนำเข้าข้อมูลผ่านปุ่ม Add data



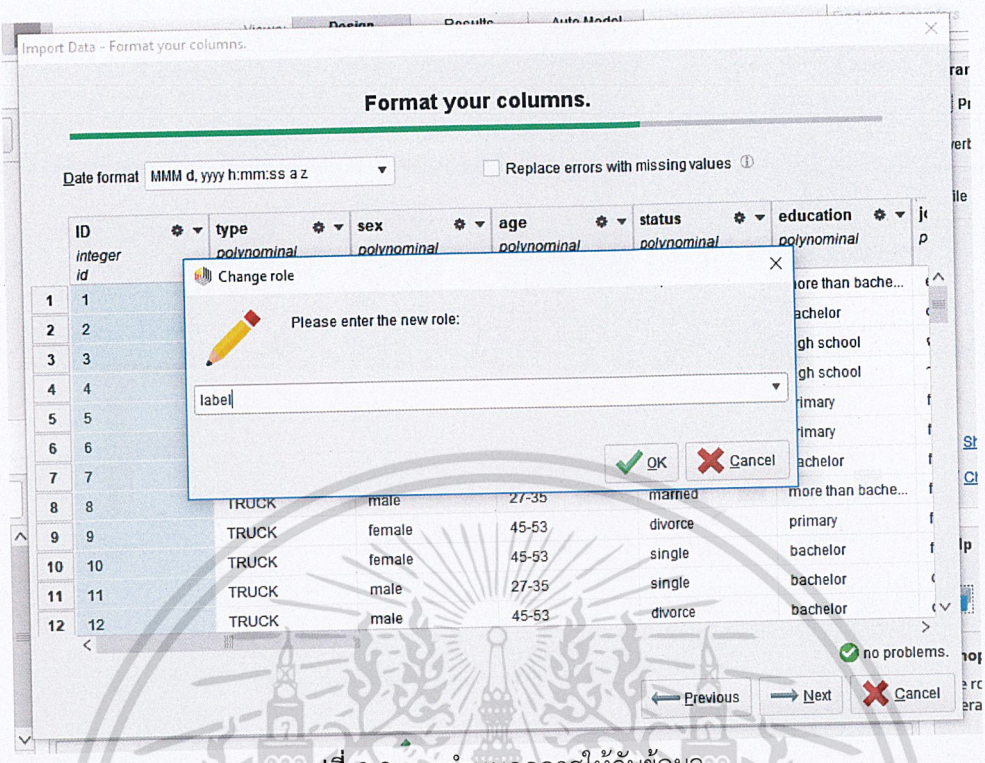
รูปที่ 3.6 การเลือกไฟล์นำเข้าข้อมูลผ่านปุ่ม Add data (ต่อ)



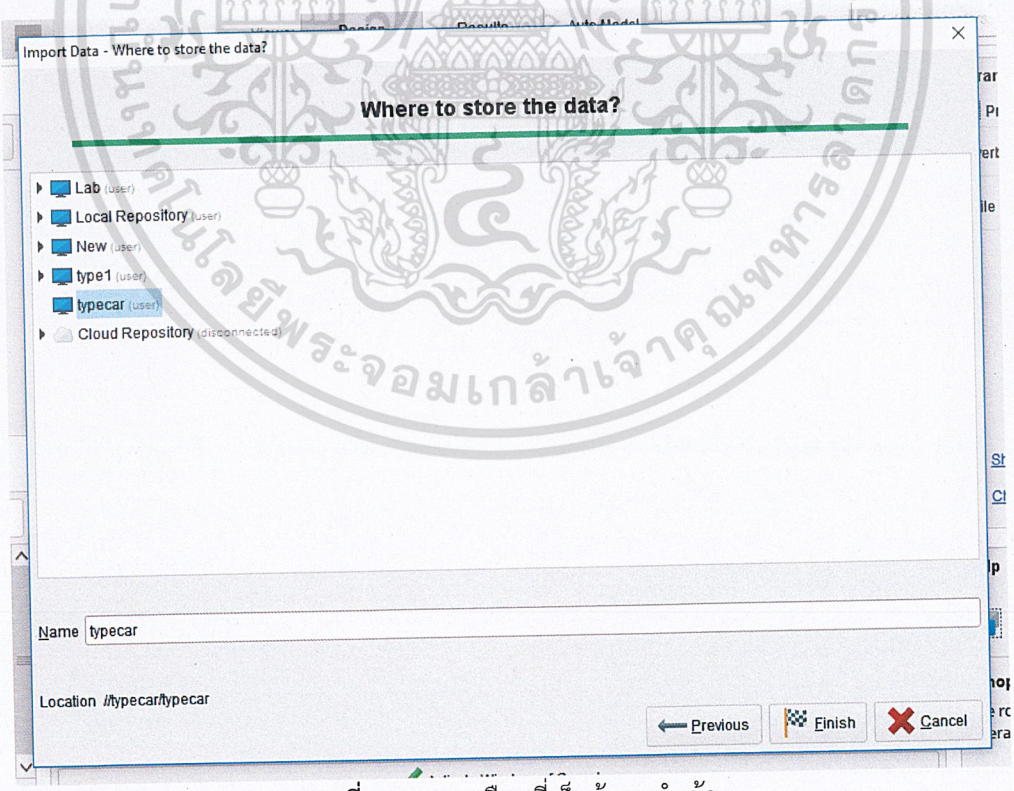
รูปที่ 3.7 การกำหนดคีย์หลักและคลาสของข้อมูล



รูปที่ 3.8 การกำหนดคีย์หลักให้กับข้อมูล



รูปที่ 3.9 การกำหนดคลาสให้กับข้อมูล



รูปที่ 3.10 การเลือกที่เก็บข้อมูลนำเข้า

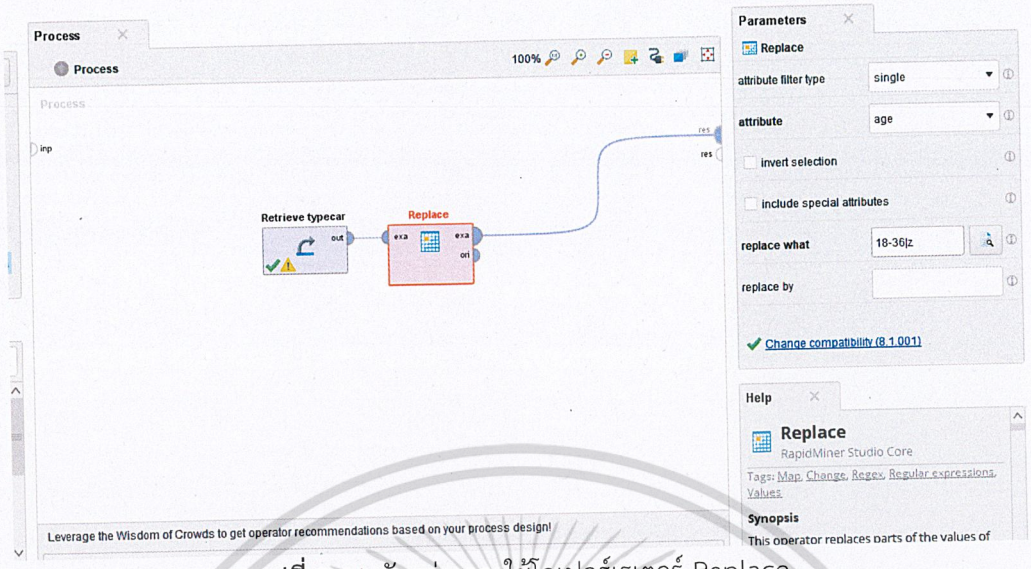
ขั้นตอนที่ 3 ทำความสะอาดข้อมูล (Cleaning Data) นำข้อมูลที่นำเข้ามาลากไว้ที่หน้าต่าง Process แล้วทำการทำความสะอาดข้อมูลโดยใช้โอเปอร์เรเตอร์ Replace จากนั้นไปยังหน้าต่าง Parameter เพื่อเลือกแอททริบิวต์ที่ต้องการทำความสะอาด โดยช่อง attribute คือ ช่องที่ให้เลือกแอททริบิวต์ที่ต้องการทำความสะอาดและช่อง replace what คือ ช่องที่มีไว้เพื่อกรอกตัวอย่างข้อมูลที่ผิดพลาด และช่อง replace by คือ ค่าที่ต้องการไปแทนค่าที่ผิดพลาด

กรณีที่ต้องการกำจัดข้อมูลออก จะใช้โอเปอร์เรเตอร์ Filter Example เชื่อมกับข้อมูลนำเข้า จากนั้นจะพบกับหน้าต่าง Create Filters: filters ในช่องแรกให้เลือกแอททริบิวต์ที่มีค่าที่ต้องการกำจัด ในช่องที่สอง คือ การกำหนดให้ข้อมูลที่ต้องการไม่มีค่าของข้อมูลที่กำจัดออกไป และช่องที่ 3 คือ ตัวอย่างข้อมูลที่ต้องการกำจัดออก ดัง**รูปที่ 3.13** การใช้โอเปอร์เรเตอร์ Filter Examples

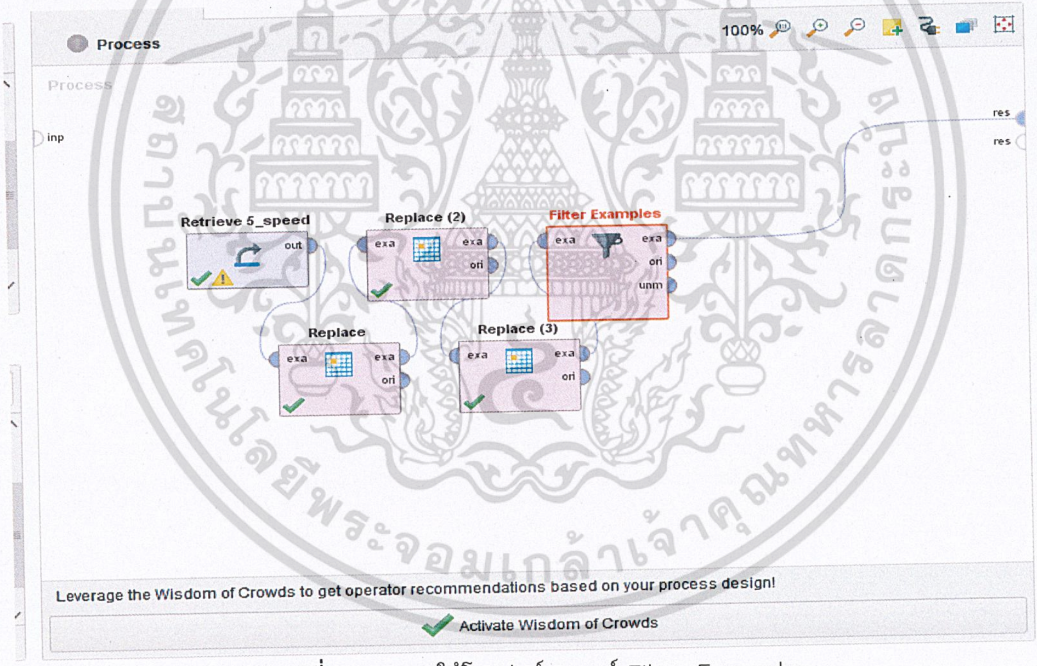
กรณีที่ต้องการแทนค่าข้อมูลขาดหาย จะใช้โอเปอร์เรเตอร์ Replace Missing Values โดยในช่อง Paremeter ในช่วงของ defult คือ การเลือกค่าที่ต้องการแทนในข้อมูลขาดหาย โดยในปัญหาพิเศษนี้จะเลือกการแทนค่าโดยค่า “N/A” ดัง**รูปที่ 3.14** การใช้โอเปอร์เรเตอร์ Replace Missing Value

กรณีที่ต้องการจัดกลุ่มข้อมูล จะใช้โอเปอร์เรเตอร์ Discretize โดยในช่อง Paremeter ให้กดปุ่ม Edit List จะปรากฏหน้าต่าง Edit Parameter List Classes: ในช่อง Class name คือ การกำหนดชื่อกลุ่มที่ต้องการแบ่ง และช่อง Upper limit คือ ค่าสูงสุดของการแบ่งแต่ละกลุ่ม ดัง**รูปที่ 3.15** ตัวอย่างการใช้โอเปอร์เรเตอร์ Discretize

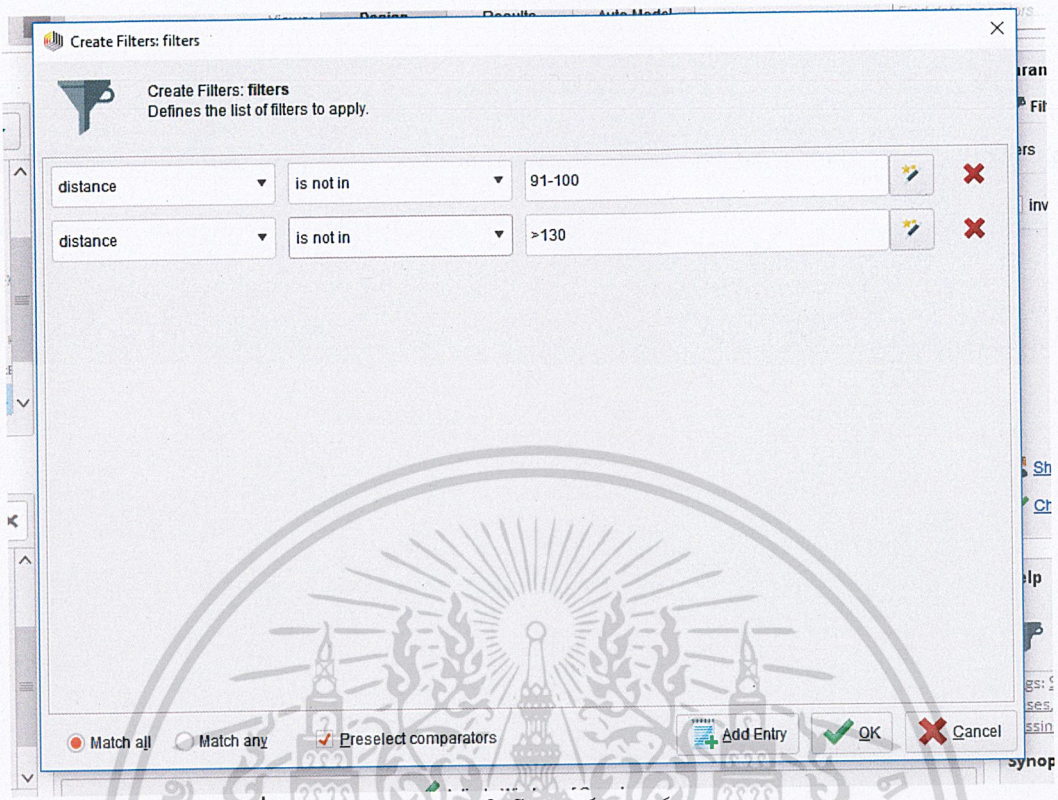
สุดท้ายเมื่อทำความสะอาดข้อมูลเรียบร้อยแล้ว ทำการบันทึกข้อมูลโดยการไปที่หน้า Result คลิกขวาที่แถบแสดงผลแล้วเลือก Store ExampleSet in Repository



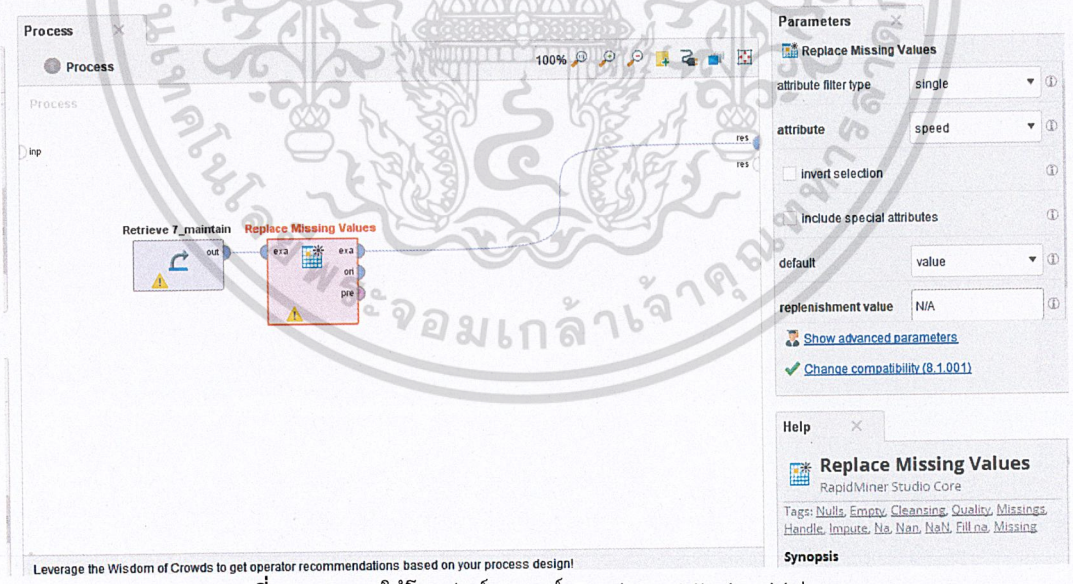
รูปที่ 3.11 ตัวอย่างการใช้โอเปอร์เรเตอร์ Replace



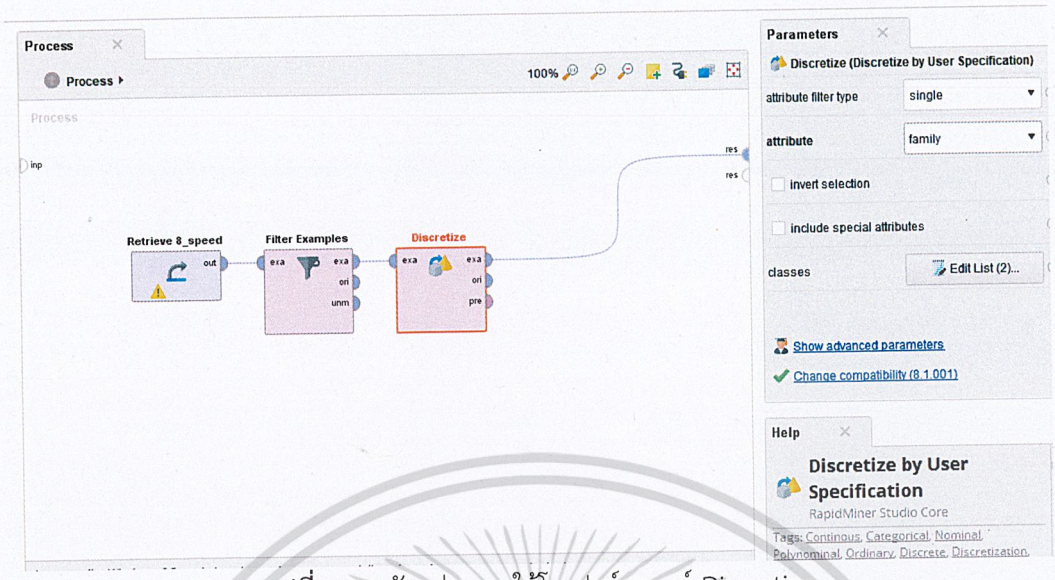
รูปที่ 3.12 การใช้โอเปอร์เรเตอร์ Filter Examples



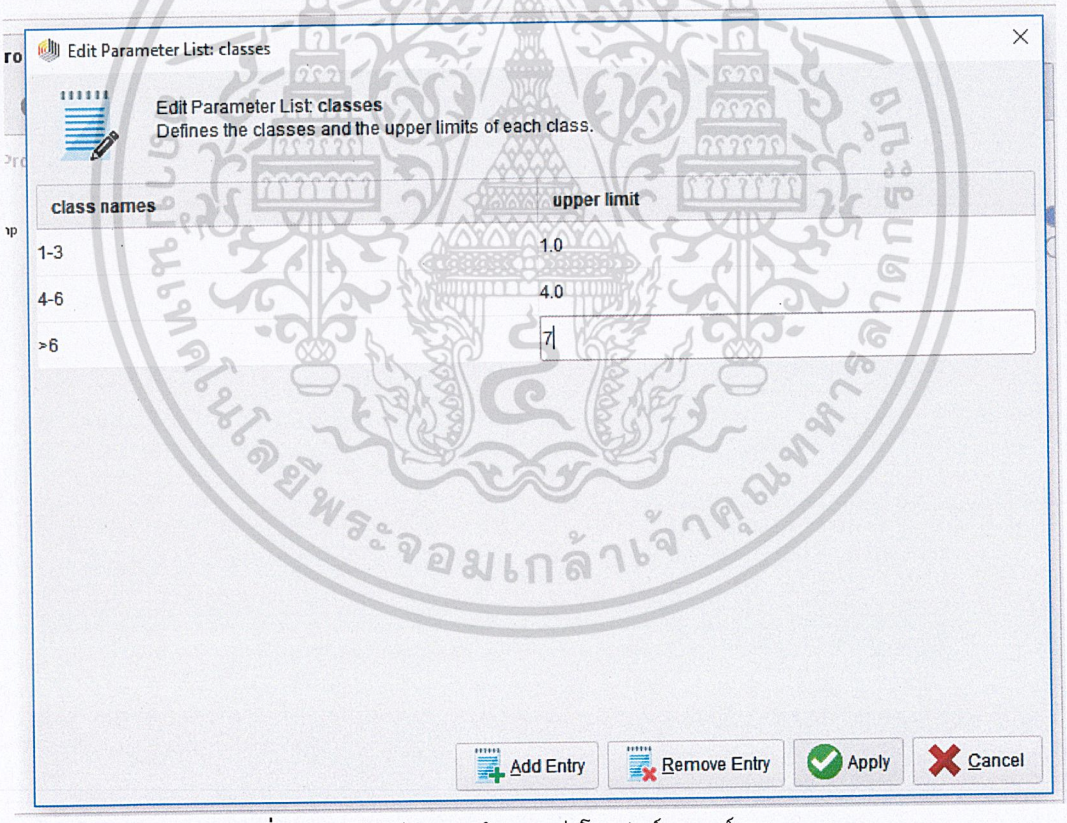
รูปที่ 3.13 การกำหนดค่าในโอเปอร์เรเตอร์ Filter Examples



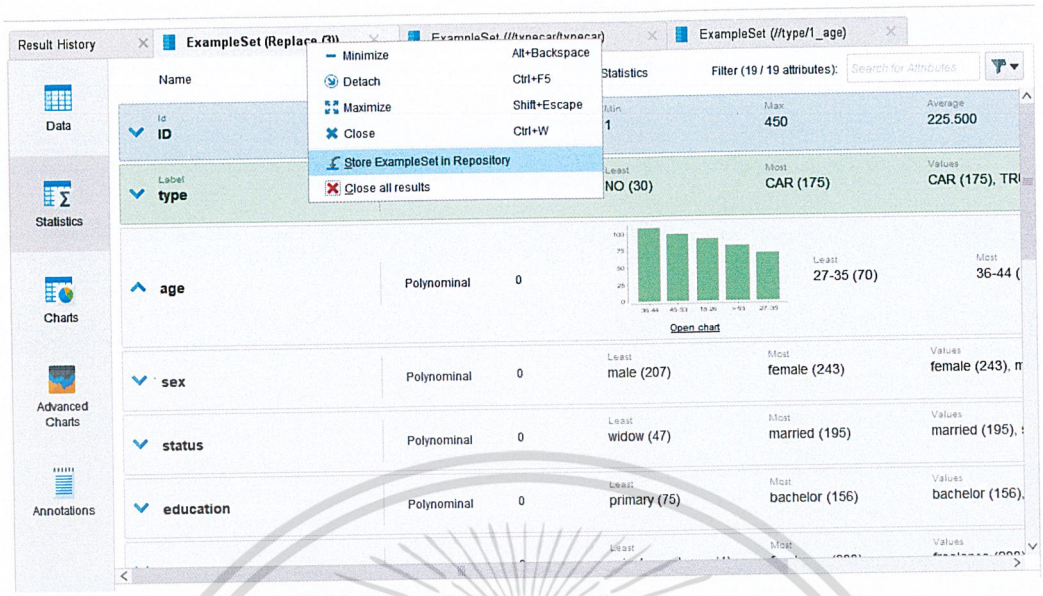
รูปที่ 3.14 การใช้โอเปอร์เรเตอร์ Replace Missing Values



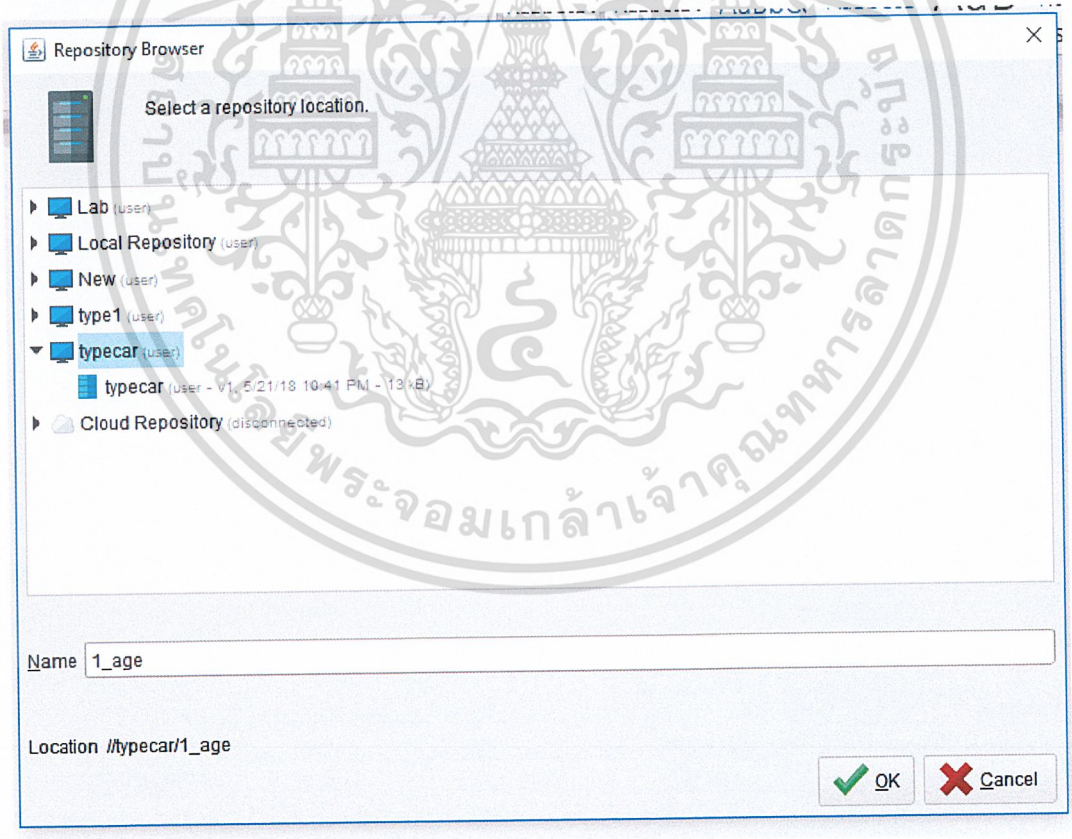
รูปที่ 3.15 ตัวอย่างการใช้โอเปอร์เรเตอร์ Discretize



รูปที่ 3.16 ตัวอย่างการกำหนดค่าโอเปอร์เรเตอร์ Discretize



รูปที่ 3.17 การบันทึกข้อมูลที่ทำความสะอาดแล้ว



รูปที่ 3.18 การตั้งชื่อและบันทึกข้อมูลที่ทำความสะอาดแล้ว

3.4 การเลือกเทคนิคที่เหมาะสม

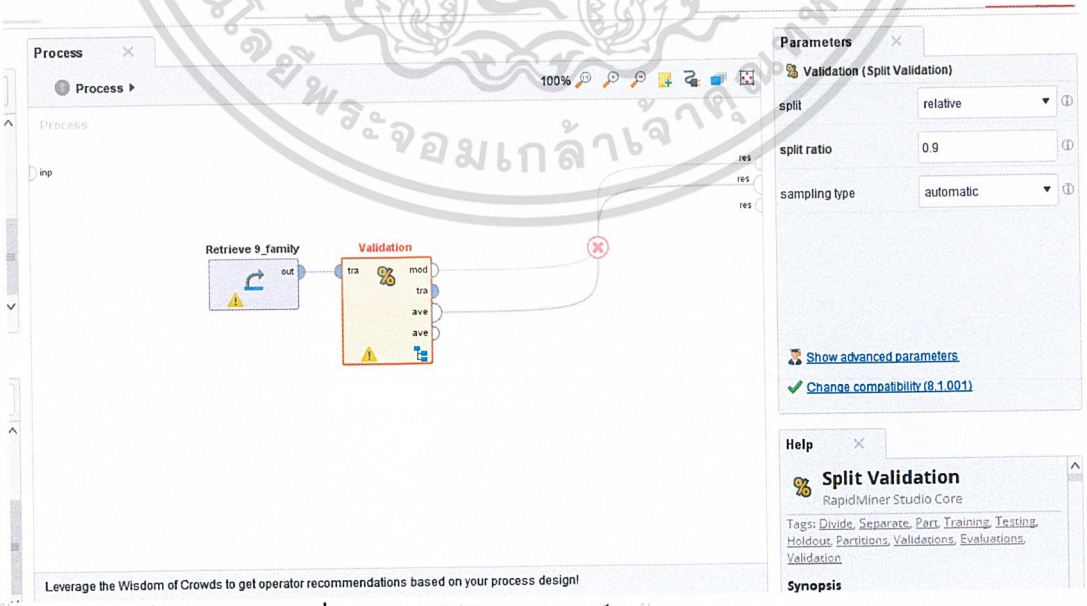
หลักจากการโอนย้ายข้อมูลและทำความสะอาดข้อมูลแล้ว ได้ทำการนำข้อมูลดังกล่าวมาทดสอบในโปรแกรม RapidMiner Studio ซึ่งได้เลือกเทคนิค Decision Tree และ Naïve Bayesian เป็นตัวทดสอบ โดยแต่ละเทคนิคจะให้ผลลัพธ์ที่แตกต่างกัน และทำการหาค่าความถูกต้อง (Accuracy) ที่แตกต่างกัน

3.5 การทดสอบข้อมูล

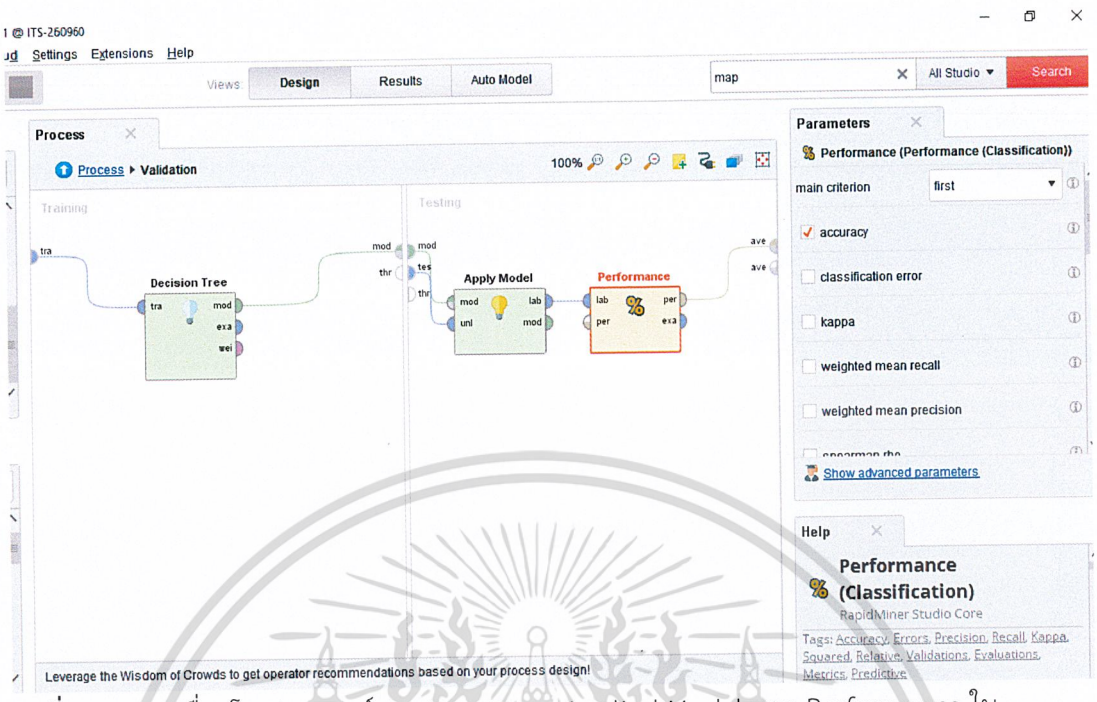
หลังจากทำการทำความสะอาดข้อมูลเรียบร้อยแล้ว จะต้องทำการทดสอบข้อมูลโดยใช้เทคนิค Decision Tree และ Naïve Bayesian โดยมีขั้นตอนดังนี้

3.5.1 การทดสอบข้อมูลโดยใช้วิธี Decision Tree

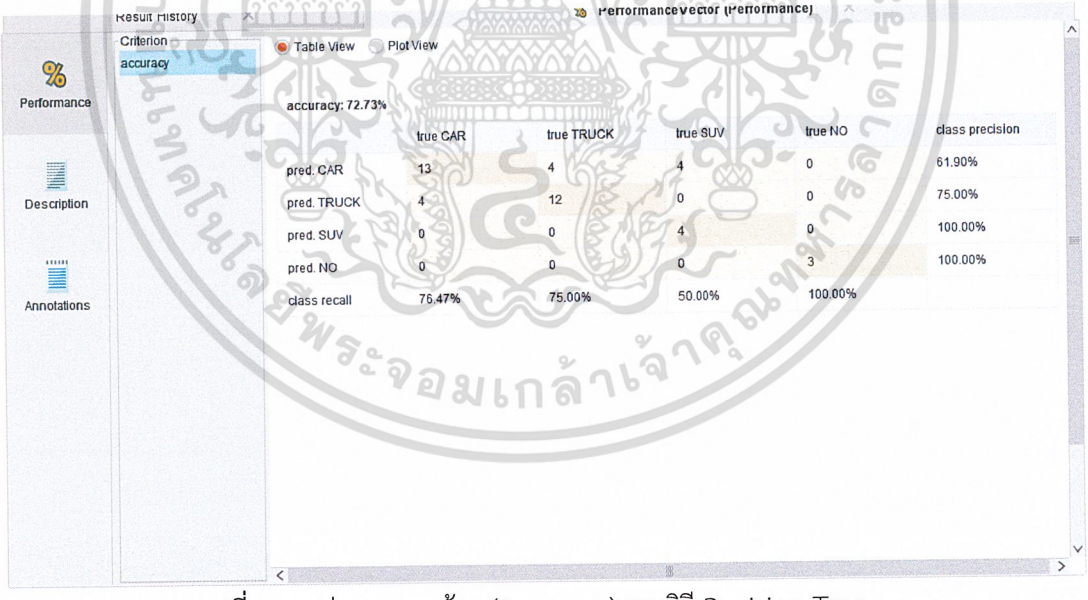
ขั้นตอนการทดสอบข้อมูลและสร้างโมเดลด้วยวิธี Decision Tree มีขั้นตอนการทำ คือ ทำการเชื่อมข้อมูลกับโอเปอร์เรเตอร์ Split Validation จากนั้นดับเบิ้ลคลิกที่โอเปอร์เรเตอร์ Split Validation ในส่วนของ Traning ให้ลากโอเปอเรเตอร์ Decision Tree ลงไป ส่วนส่วนของ Testing ให้ลากโอเปอเรเตอร์ Apply Model และ Performance ลงไป แล้วทำการ run จากนั้นจะได้ค่า Accuracy และแผนภาพต้นไม้ สุดท้ายคือการทำ Rule Model โดยการใช้โอเปอร์เรเตอร์ Tree to Rules เชื่อมต่อกับข้อมูล จากนั้นดับเบิ้ลคลิกที่โอเปอร์เรเตอร์ Tree to Rules แล้วลากโอเปอร์เรเตอร์ Dicision Tree ลงใน Process แล้วทำการ run จึงจะได้ Rule Model ของวิธี Dicision Tree ในรูปแบบของ if else ดัง**รูปที่ 3.25** ตัวอย่างของ Rule Model



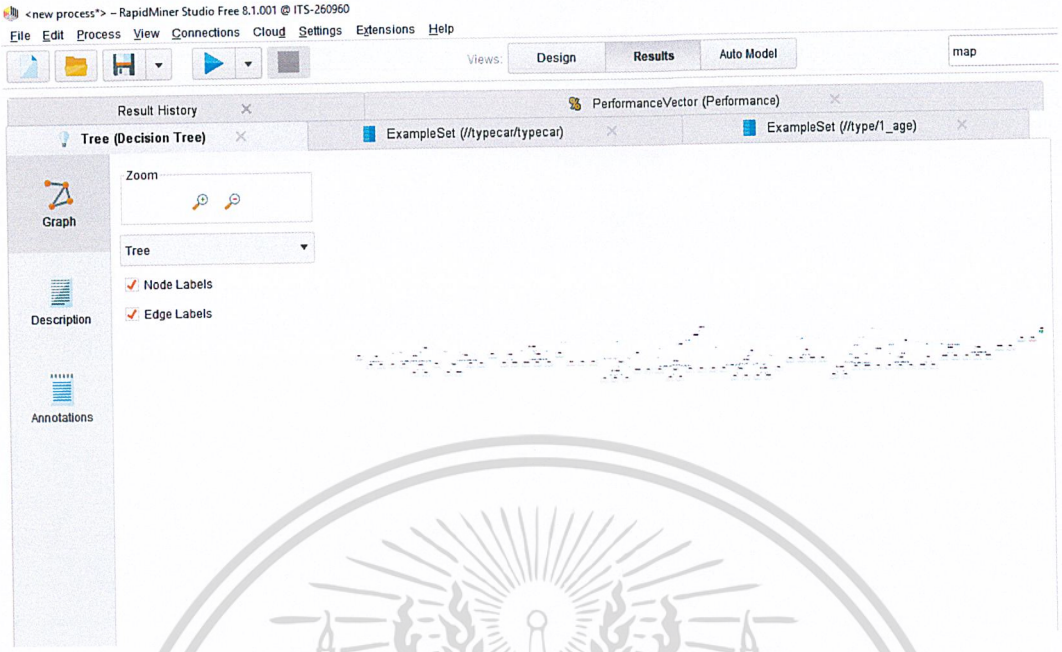
รูปที่ 3.19 การใช้โอเปอเรเตอร์ Split Validation



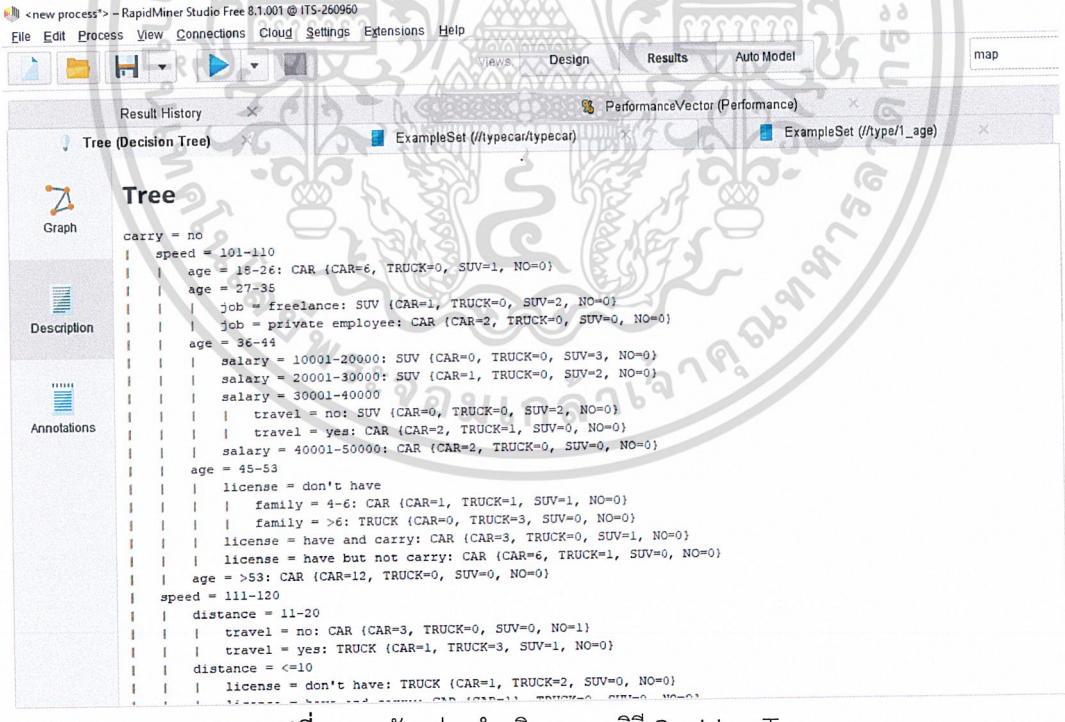
รูปที่ 3.20 การเชื่อมโอเปอเรเตอร์ Decision Tree, Applied Model และ Performance ใน โอเปอร์เรเตอร์ Validation



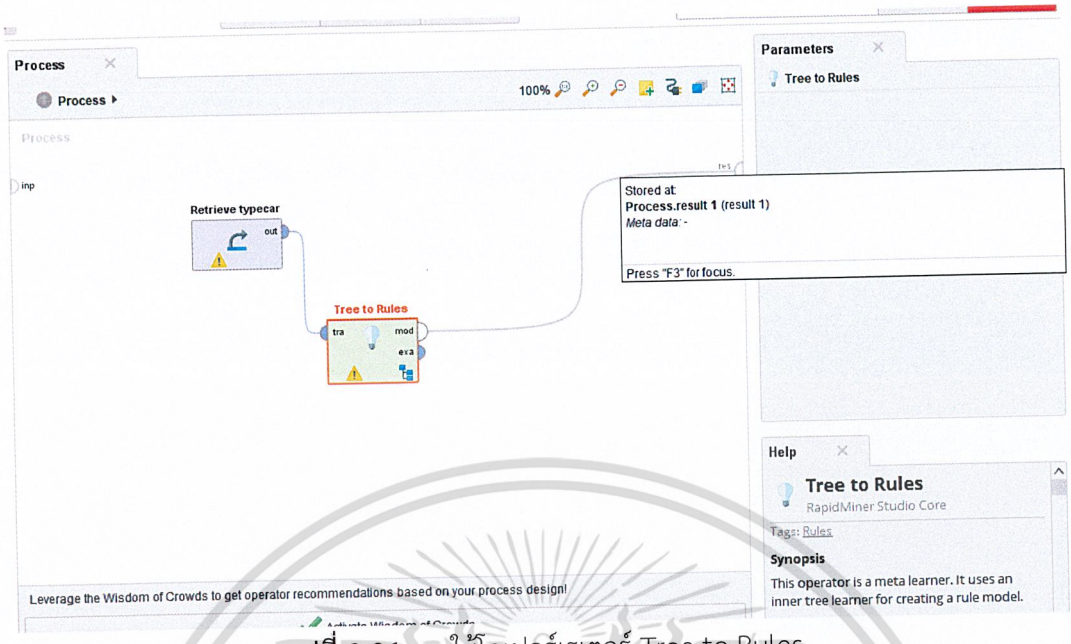
รูปที่ 3.21 ค่าความถูกต้อง (Accuracy) ของวิธี Decision Tree



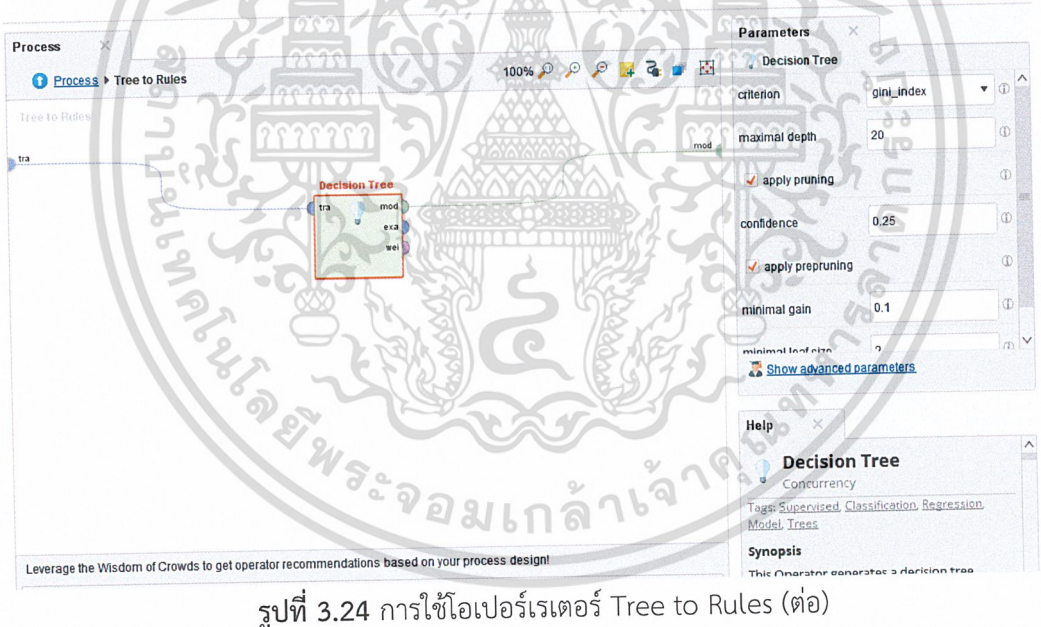
รูปที่ 3.22 ตัวอย่าง Decision Tree



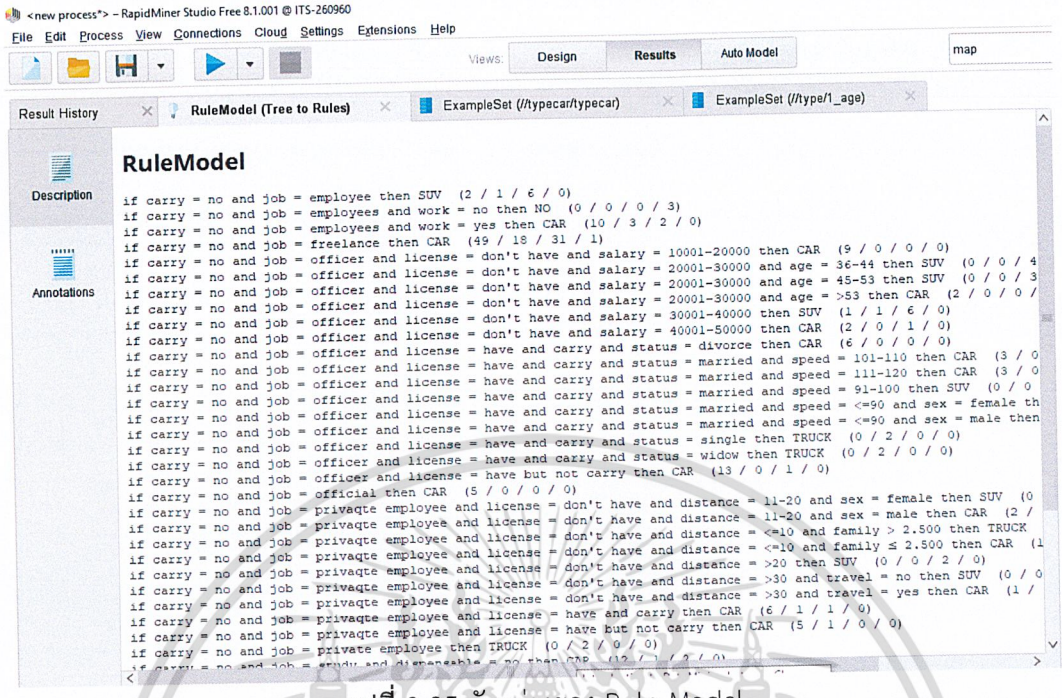
รูปที่ 3.23 ตัวอย่างคำอธิบายของวิธี Decision Tree



รูปที่ 3.24 การใช้โอเปอร์เรเตอร์ Tree to Rules



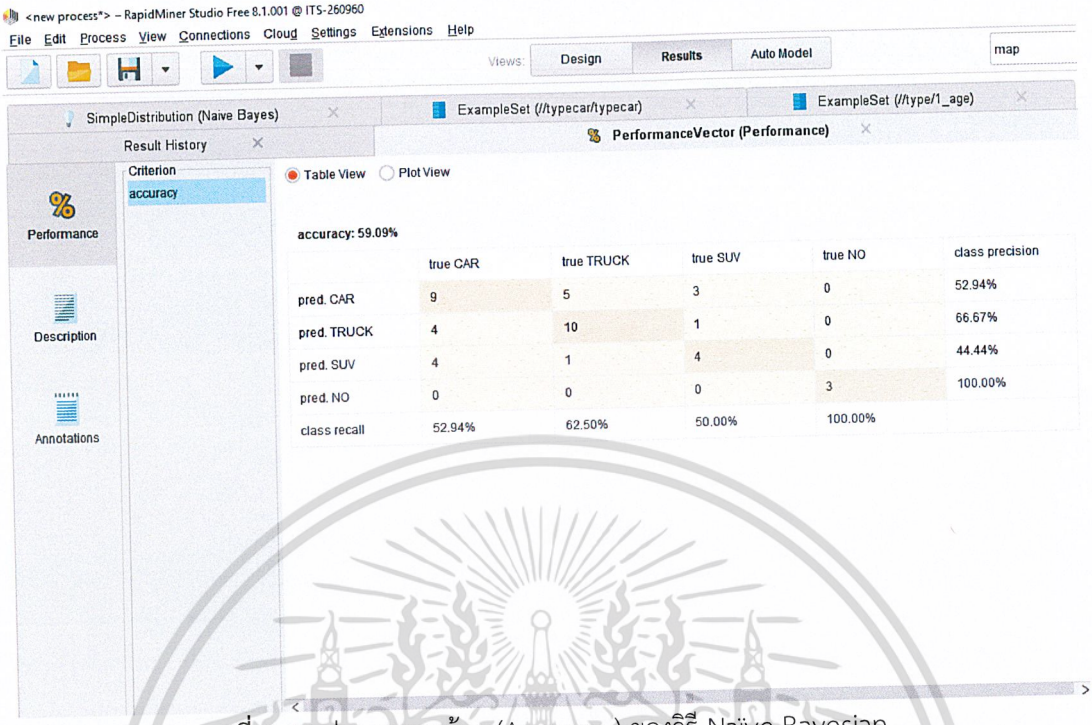
รูปที่ 3.24 การใช้โอเปอร์เรเตอร์ Tree to Rules (ต่อ)



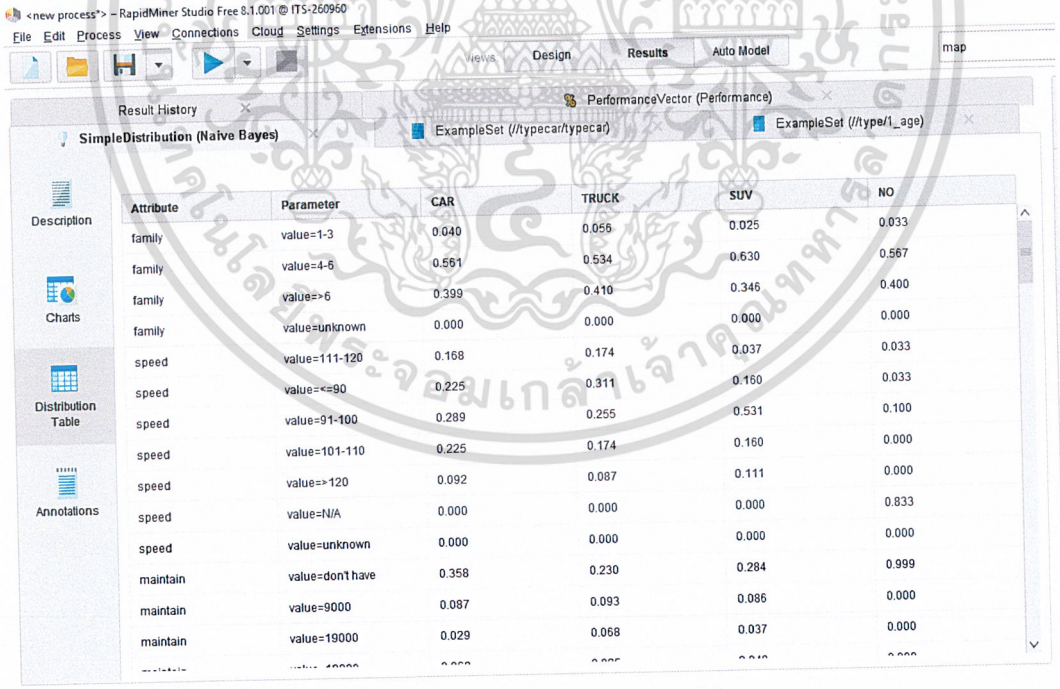
รูปที่ 3.25 ตัวอย่างของ Rule Model

3.5.2 การทดสอบข้อมูลโดยใช้วิธี Naïve Bayesian

ขั้นตอนการทดสอบข้อมูลและสร้างโมเดลด้วยวิธี Naïve Bayesian มีขั้นตอนการทำคล้ายกับการทดสอบข้อมูลและสร้างโมเดลด้วยวิธี Decision Tree ซึ่งมีขั้นตอนการทำ คือ ทำการเชื่อมข้อมูลกับโอเปอร์เรเตอร์ Split Validation ดัง**รูปที่ 3.19** การใช้โอเปอเรเตอร์ Split Validation จากนั้นดับเบิ้ลคลิกที่โอเปอร์เรเตอร์ Split Validation ในส่วนของ Traning ให้ลากโอเปอเรเตอร์ Naïve Bayesian ลงไป ส่วนของ Testing ให้ลากโอเปอเรเตอร์ Apply Model และ Performance ลงไป ดัง**รูปที่ 3.20** การเชื่อมโอเปอเรเตอร์ Decision Tree, Applied Model และ Performance ในโอเปอเรเตอร์ Validation แล้วทำการ run จากนั้นจะได้ค่า Accuracy และตารางค่าความน่าจะเป็นของแต่ละคลาสของวิธี Naïve Bayesian



รูปที่ 3.26 ค่าความถูกต้อง (Accuracy) ของวิธี Naïve Bayesian



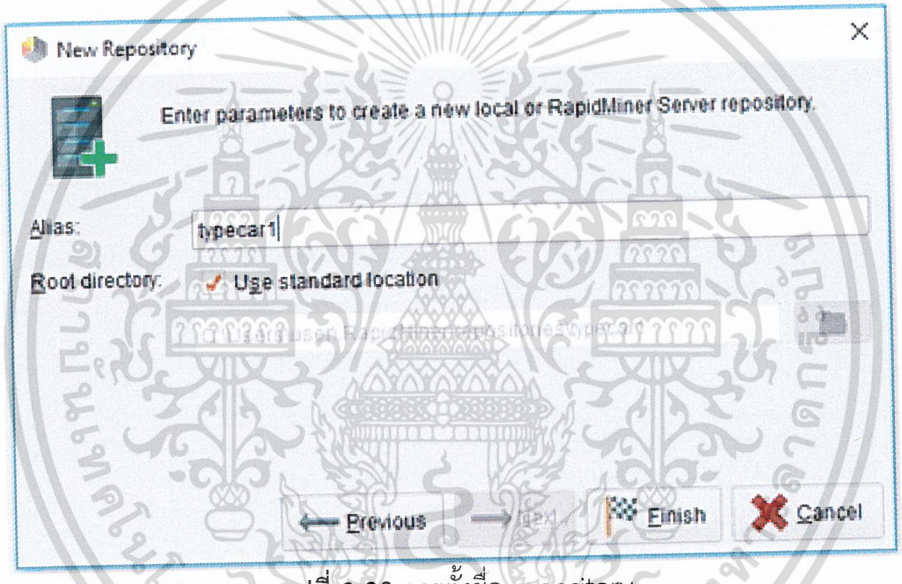
รูปที่ 3.27 ค่าความน่าจะเป็นของแต่ละคลาสของวิธี Naïve Bayesian

3.6 การทดสอบข้อมูลที่ไม่รู้คลาส

3.6.1 การทดสอบข้อมูลด้วยวิธี Decision Tree

การทดสอบข้อมูลด้วยวิธี Dicision Tree มีขั้นตอนการดำเนินการ ดังนี้

ขั้นตอนที่ 1 เริ่มจากเปิดโปรแกรม RapidMiner Studio จะปรากฏหน้าต่างดัง**รูปที่ 3.2** หน้าโปรแกรม RapidMiner Studio จากนั้นทำการสร้าง repository ใหม่โดยการคลิกที่ปุ่มที่อยู่ข้างปุ่ม Add data แล้วเลือก Create repository **รูปที่ 3.4** การสร้าง repository ใหม่ เมื่อกดแล้วจะปรากฏหน้าต่าง New Repository เลือก New local repository แล้วกด Next การเลือกตำแหน่งที่เก็บ repository จากทำการตั้งชื่อ repository โดยในปัญหาพิเศษนี้ตั้งชื่อว่า typecar1

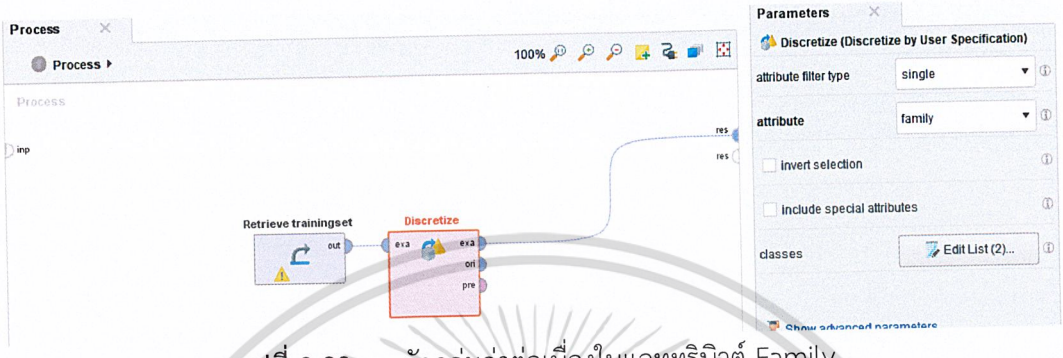


รูปที่ 3.28 การตั้งชื่อ repository

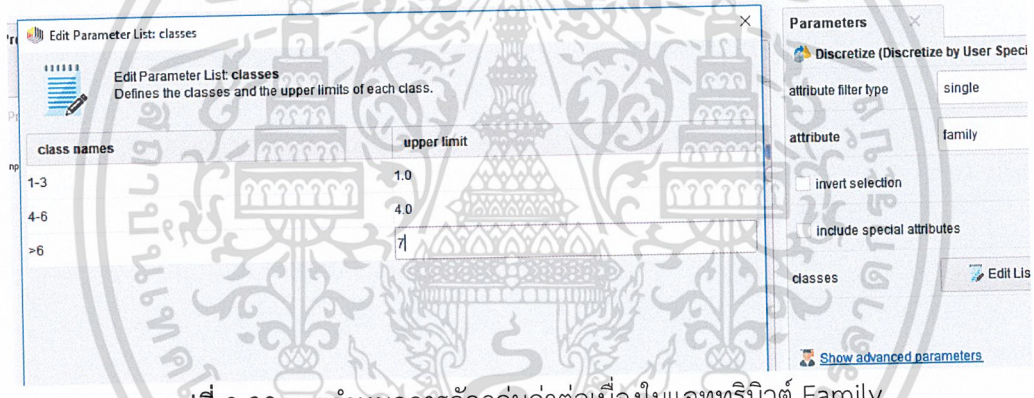
ขั้นตอนที่ 2 ทำการนำเข้าข้อมูลเข้าสู่โปรแกรมโดยคลิกที่ปุ่ม Add Data จากนั้นจะปรากฏหน้าต่าง Import Data – Select the data location เพื่อเลือกข้อมูลที่ต้องการนำเข้า การนำเข้าข้อมูลผ่านปุ่ม Add data ต่อมาจะเป็นการกำหนดคลาสของข้อมูล โดยการคลิกที่รูปฟันเฟืองที่อยู่แถวแรกของข้อมูล แล้วเลือก Change Role จากนั้นจะปรากฏหน้าต่าง Change Role พิมพ์ id เมื่อต้องการให้แอททริบิวต์นั้นเป็นคีย์หลัก ดัง**รูปที่ 3.8** การกำหนดคีย์หลักให้กับข้อมูล หลังจากนั้นทำการเลือกแหล่งที่ต้องการเก็บข้อมูล โดยในปัญหาพิเศษนี้จะเก็บข้อมูลไว้ที่ typecar1

ขั้นตอนที่ 3 เลือกข้อมูลที่นำเข้าไว้ข้างต้น โดยในปัญหาพิเศษนี้ใช้ข้อมูลชื่อ Retrieve trainingset จากนั้นทำการจัดกลุ่มค่าต่อเนื่อง โดยใช้โอเปอเรเตอร์ “Discretize” โดยกำหนด Parameters คือ attribute filter type เลือก single และเลือกแอททริบิวต์

ที่ต้องการ ในปัญหาพิเศษคือ family ดัง**รูปที่ 3.29** การจัดกลุ่มค่าต่อเนื่องในแอททริบิวต์ Family กำหนด upper limit คือ 1 สำหรับกลุ่ม 1-3, 4 สำหรับกลุ่ม 4-6 และ 7 สำหรับกลุ่ม >6 ตามลำดับ ทำการบันทึกโดยใช้ชื่อ testset

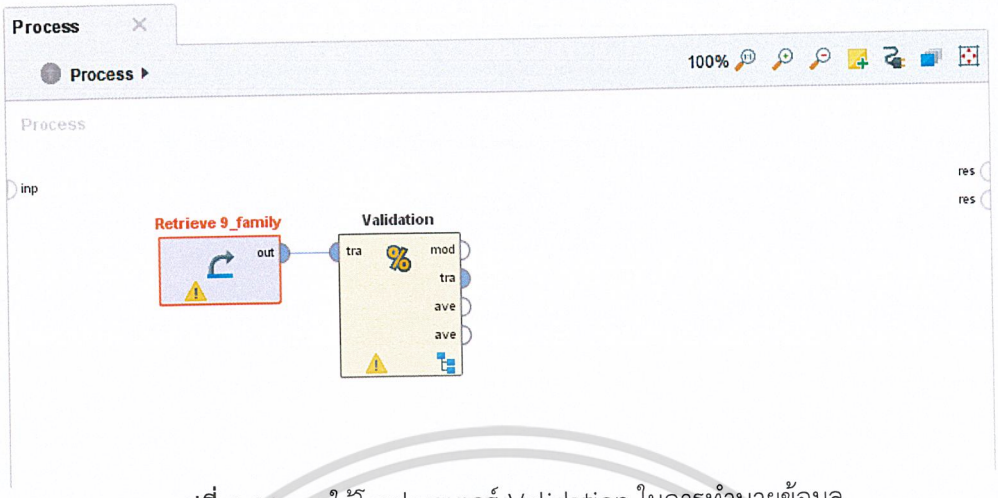


รูปที่ 3.29 การจัดกลุ่มค่าต่อเนื่องในแอททริบิวต์ Family



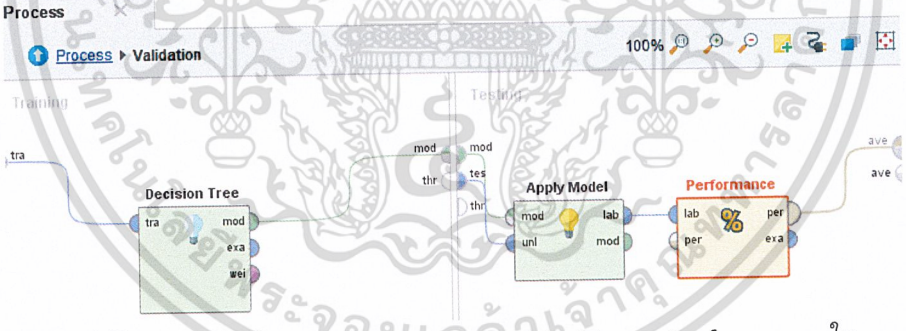
รูปที่ 3.30 การกำหนดการจัดกลุ่มค่าต่อเนื่องในแอททริบิวต์ Family

ขั้นตอนที่ 4 เลือกข้อมูลเก่าที่มีชื่อว่า 9_family มาเชื่อมกับโอเปอเรเตอร์ “Validation” ดัง**รูปที่ 3.31** การใช้โอเปอเรเตอร์ Validation ในการทำนายข้อมูล



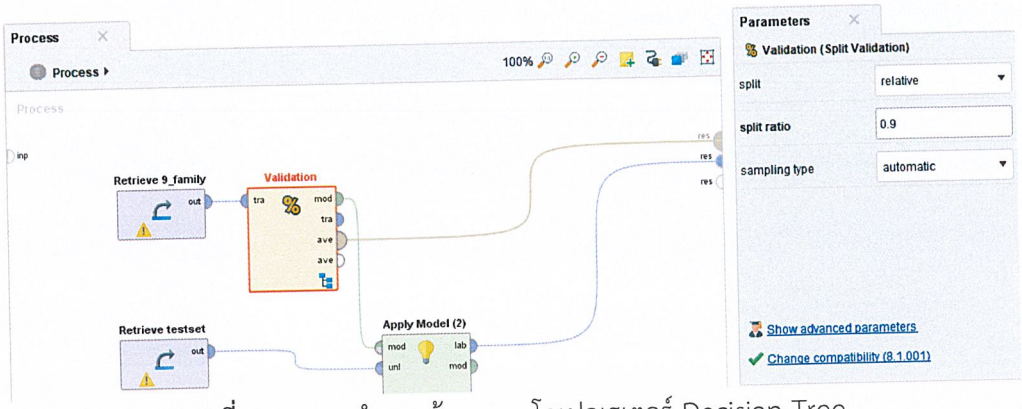
รูปที่ 3.31 การใช้โอเปอเรเตอร์ Validation ในการทำนายข้อมูล

ขั้นตอนที่ 5 ดับเบิ้ลคลิกที่ตัวโอเปอเรเตอร์ “Validation” จะได้หน้าต่างใหม่ จากนั้นเลือกโอเปอเรเตอร์ “Decision Tree”, “Apply Model”, “Performance” และทำการเชื่อมโอเปอเรเตอร์ดัง**รูปที่ 3.32** การใช้โอเปอเรเตอร์ Decision Tree, Apply Model และ Performance ในการทำนายข้อมูล

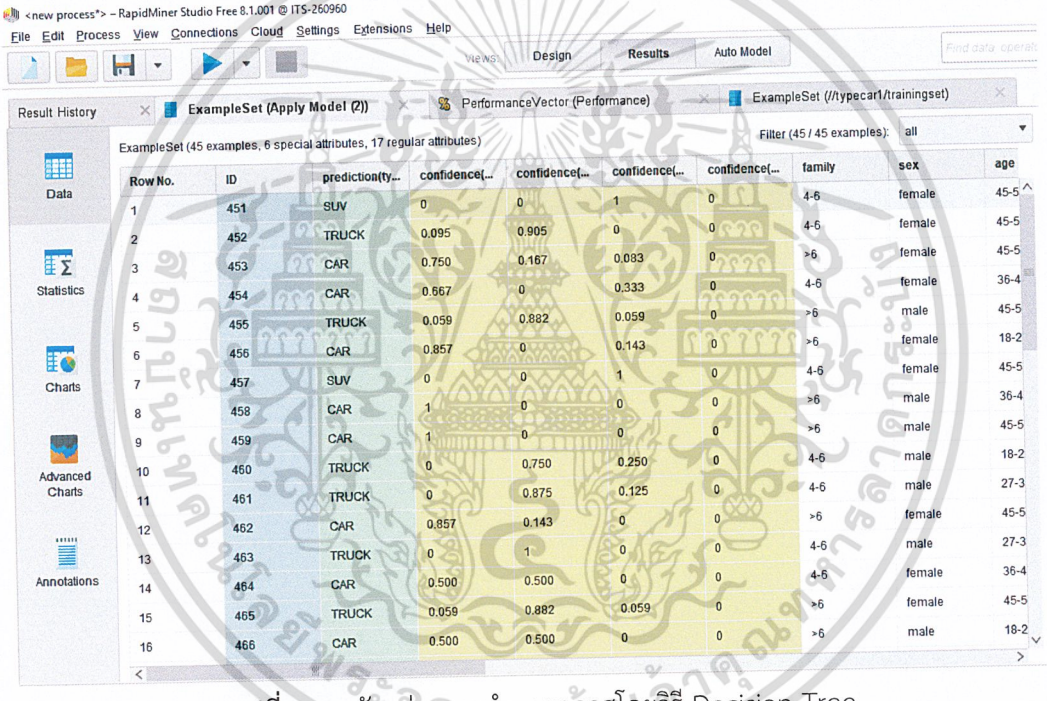


รูปที่ 3.32 การใช้โอเปอเรเตอร์ Decision Tree, Apply Model และ Performance ใน Validationในการทำนายข้อมูล

ขั้นตอนที่ 6 กลับไปที่หน้าต่าง Process เลือกข้อมูลที่ทำการบันทึกไว้ข้างต้น นั่นคือ ข้อมูลที่มีชื่อว่า testset และเลือกโอเปอเรเตอร์ “Apply Modeal” จากนั้นเชื่อมโอเปอเรเตอร์และข้อมูลต่าง ๆ โดยที่โอเปอเรเตอร์ “Validation” กำหนด Parameters คือ split เลือก relative, split ratio กำหนด 0.9 และsampling type ดัง**รูปที่ 3.33** ทำนายข้อมูลของโอเปอเรเตอร์ Decision Tree และจะได้ผลลัพธ์เป็นตารางการทำนายคลาสต่างๆ ดัง**รูปที่ 3.34** ตัวอย่างการทำนายคลาสโดยวิธี Decision Tree



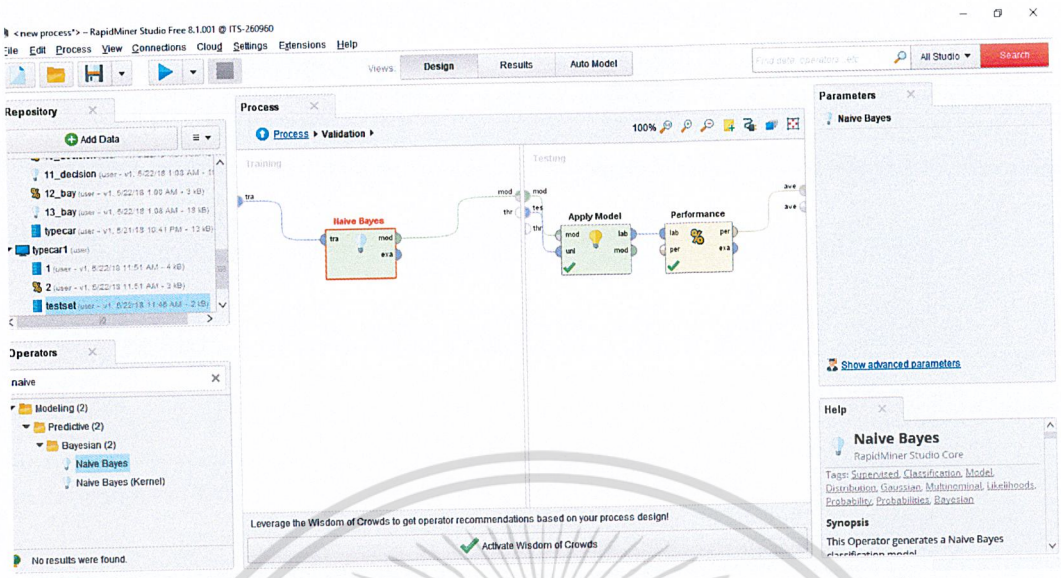
รูปที่ 3.33 การทำนายข้อมูลของโอเปอเรเตอร์ Decision Tree



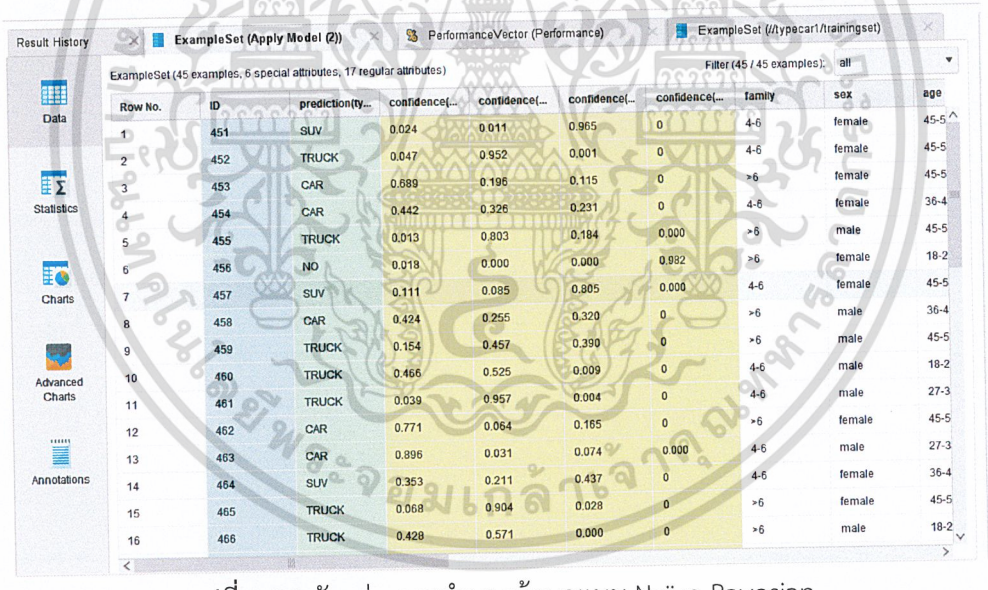
รูปที่ 3.34 ตัวอย่างการทำนายคลาสโดยวิธี Decision Tree

3.6.2 การทดสอบข้อมูลด้วยวิธี Naïve Bayesian

การทดสอบข้อมูลด้วยวิธี Naïve Bayesian มีขั้นตอนการดำเนินการ คือ เริ่มจากทำตามขั้นตอนที่ 1, 2, 3 และ 4 ของวิธี Dicision Tree จากนั้น ภายในโอเปอร์เรเตอร์ Validation ให้เปลี่ยนจาก Decision Tree เป็น Naïve bayes ดัง**รูปที่ 3.35** การใช้โอเปอเรเตอร์ Naïve Bayes, Apply Model และ Performance ใน Validationในการทำนายแล้วทำการ run ข้อมูล สุดท้ายจะได้ผลลัพธ์ดัง**รูปที่ 3.36** ตัวอย่างการทำนายข้อมูลแบบ Naïve Bayesian



รูปที่ 3.35 การใช้โอเปอเรเตอร์ Naïve Bayes, Apply Model และ Performance ใน Validation ในการทำนายข้อมูล



Row No.	ID	prediction(ty...	confidence(...	confidence(...	confidence(...	confidence(...	family	sex	age
1	451	SUV	0.024	0.011	0.965	0	4-6	female	45-5
2	452	TRUCK	0.047	0.952	0.001	0	4-6	female	45-5
3	453	CAR	0.689	0.196	0.115	0	>6	female	45-5
4	454	CAR	0.442	0.326	0.231	0	4-6	female	36-4
5	455	TRUCK	0.013	0.803	0.184	0.000	>6	male	45-5
6	456	NO	0.018	0.000	0.000	0.982	>6	female	18-2
7	457	SUV	0.111	0.085	0.805	0.000	4-6	female	45-5
8	458	CAR	0.424	0.255	0.320	0	>6	male	36-4
9	459	TRUCK	0.154	0.457	0.390	0	>6	male	45-5
10	460	TRUCK	0.466	0.525	0.009	0	4-6	male	18-2
11	461	TRUCK	0.039	0.957	0.004	0	4-6	male	27-3
12	462	CAR	0.771	0.064	0.165	0	>6	female	45-5
13	463	CAR	0.896	0.031	0.074	0.000	4-6	male	27-3
14	464	SUV	0.353	0.211	0.437	0	4-6	female	36-4
15	465	TRUCK	0.068	0.904	0.028	0	>6	female	45-5
16	466	TRUCK	0.428	0.571	0.000	0	>6	male	18-2

รูปที่ 3.36 ตัวอย่างการทำนายข้อมูลแบบ Naïve Bayesian

หลังจากการดำเนินการกำหนดประชากรและเลือกกลุ่มตัวอย่าง สร้างแบบสอบถาม เตรียมข้อมูล เลือกเทคนิคในการทดสอบข้อมูลและทดสอบข้อมูลเรียบร้อยแล้ว ในบทถัดไปจะกล่าวถึงผลของการทำปัญหาพิเศษและอภิปรายผล

บทที่ 4

ผลการวิจัย และอภิปรายผล

จากกระบวนการทำปัญหาพิเศษในบทที่ 3 เพื่อการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ ในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบังและจังหวัดฉะเชิงเทรา การวิเคราะห์ข้อมูลและการแปรความหมายของผลการวิเคราะห์ข้อมูล มีดังนี้

4.1 ผลการวิเคราะห์

ในการทดสอบและสร้างโมเดลนี้ ได้ทำการทดสอบในโปรแกรม RapidMiner Studio ข้อมูลที่ใช้ทดสอบมีจำนวน 450 ตัวอย่าง 18 แอททริบิวต์ ในการทดสอบจะแบ่งจำนวนข้อมูลศึกษา (Training set) และข้อมูลทดสอบ (Test set) เป็นระดับต่างๆ คือ ข้อมูลศึกษา 40% ของข้อมูลทั้งหมด ข้อมูลทดสอบ 60% ของข้อมูลทั้งหมด, ข้อมูลศึกษา 30% ของข้อมูลทั้งหมด ข้อมูลทดสอบ 70% ของข้อมูลทั้งหมด, ข้อมูลศึกษา 20% ของข้อมูลทั้งหมด ข้อมูลทดสอบ 80% ของข้อมูลทั้งหมด และข้อมูลศึกษา 10% ของข้อมูลทั้งหมด ข้อมูลทดสอบ 90% ของข้อมูลทั้งหมด และนำผลลัพธ์ที่ได้มาหาค่าเฉลี่ย โดยความหมายของแต่ละแอททริบิวต์มีความหมายดังตารางที่ 4 โดยแต่ละแอททริบิวต์มีค่าต่างๆ ดังตาราง

ตารางที่ 4.1 ค่าในแอททริบิวต์ type

TYPE	ความหมาย
CAR	รถเก๋ง
TRUCK	รถกระบะ
SUV	รถอเนกประสงค์
NO	ไม่มีรถยนต์

ตารางที่ 4.2 ค่าในแอททริบิวต์ sex

sex	ความหมาย
male	เพศชาย
female	เพศหญิง

ตารางที่ 4.3 ค่าในแอททริบิวต์ age

Age	ความหมาย
18-26	อายุอยู่ระหว่าง 18-26 ปี
27-35	อายุอยู่ระหว่าง 27-35 ปี
36-44	อายุอยู่ระหว่าง 36-44 ปี
45-53	อายุอยู่ระหว่าง 45-53 ปี
>53	อายุมากกว่า 53 ปี

ตารางที่ 4.4 ค่าในแอททริบิวต์ status

Status	ความหมาย
Single	สถานภาพโสด
Married	สถานภาพแต่งงานแล้ว
Divorce	สถานภาพหย่าร้างหรือแยกกันอยู่
Widow	สถานภาพหม้าย

ตารางที่ 4.5 ค่าในแอททริบิวต์ education

education	ความหมาย
primary	การศึกษาระดับการศึกษาขั้นพื้นฐาน
hight school	การศึกษาระดับมัธยมปลาย
bachelor	การศึกษาระดับปริญญาตรี
more than bachelor	การศึกษาระดับสูงกว่าปริญญาตรี

ตารางที่ 4.6 ค่าในแอททริบิวต์ job

job	ความหมาย
officer	อาชีพข้าราชการหรือเจ้าหน้าที่ของรัฐ
private employee	อาชีพพนักงานบริษัทเอกชน
employee	อาชีพพนักงานรัฐวิสาหกิจ
freelance	อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ
Study	กำลังศึกษา
Other	อาชีพอื่นๆนอกเหนือจากที่กล่าวมา

ตารางที่ 4.7 ค่าในแอททริบิวต์ salary

salary	ความหมาย
<=10000	รายได้น้อยกว่าเท่ากับ 10,000 บาทต่อเดือน
10001-20000	รายได้อยู่ระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน
20001-30000	รายได้อยู่ระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน
30001-40000	รายได้อยู่ระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน
40001-50000	รายได้อยู่ระหว่าง 40,001 ถึง 50,000 บาทต่อเดือน
>50000	รายได้มากกว่า 50,000 บาทต่อเดือน

ตารางที่ 4.8 ค่าในแอททริบิวต์ license

license	ความหมาย
have and carry	มีใบอนุญาติขับขี่และพกพา
have but not carry	มีใบอนุญาติขับขี่แต่ไม่พกพา
don't have	ไม่มีใบอนุญาติขับขี่

ตารางที่ 4.9 ค่าในแอททริบิวต์ carry

carry	ความหมาย
yes	ใช้รถยนต์ไว้เพื่อบรรทุกของ
no	ไม่ใช้รถยนต์ไว้เพื่อบรรทุกของ

ตารางที่ 4.10 ค่าในแอททริบิวต์ travel

travel	ความหมาย
yes	ใช้รถยนต์ไว้เพื่อท่องเที่ยว
no	ไม่ใช้รถยนต์ไว้เพื่อท่องเที่ยว

ตารางที่ 4.11 ค่าในแอททริบิวต์ work

work	ความหมาย
yes	ใช้รถยนต์ไว้เพื่อเดินทางไปทำงาน
no	ไม่ใช้รถยนต์ไว้เพื่อเดินทางไปทำงาน

ตารางที่ 4.12 ค่าในแอททริบิวต์ dispensable

dispensable	ความหมาย
yes	มีรถยนต์แต่ไม่จำเป็นต้องใช้
no	มีรถยนต์และจำเป็นต้องใช้

ตารางที่ 4.13 ค่าในแอททริบิวต์ other

other	ความหมาย
yes	มีรถยนต์ไว้เพื่อทำสิ่งต่างๆนอกเหนือจากที่กล่าวมา
no	ไม่ได้มีรถยนต์ไว้เพื่อทำสิ่งต่างๆนอกเหนือจากที่กล่าวมา

ตารางที่ 4.14 ค่าในแอททริบิวต์ speed

speed	ความหมาย
<=90	ความเร็วเฉลี่ยในการขับขี่รถยนต์น้อยกว่าเท่ากับ 90 กิโลเมตรต่อชั่วโมง
91-100	ความเร็วเฉลี่ยในการขับขี่รถยนต์อยู่ระหว่าง 91-100 กิโลเมตรต่อชั่วโมง
101-120	ความเร็วเฉลี่ยในการขับขี่รถยนต์อยู่ระหว่าง 101-120 กิโลเมตรต่อชั่วโมง
121-130	ความเร็วเฉลี่ยในการขับขี่รถยนต์อยู่ระหว่าง 121-130 กิโลเมตรต่อชั่วโมง
>130	ความเร็วเฉลี่ยในการขับขี่รถยนต์มากกว่า 130 กิโลเมตรต่อชั่วโมง

ตารางที่ 4.15 ค่าในแอททริบิวต์ distance

distance	ความหมาย
<=10	ระยะทางระหว่างบ้านถึงที่ทำงานหรือสถานศึกษาน้อยกว่าเท่ากับ 10 กิโลเมตร
11-20	ระยะทางระหว่างบ้านถึงที่ทำงานหรือสถานศึกษาอยู่ระหว่าง 11-20 กิโลเมตร
>20	ระยะทางระหว่างบ้านถึงที่ทำงานหรือสถานศึกษามากกว่า 20 กิโลเมตร

ตารางที่ 4.16 ค่าในแอททริบิวต์ parking

parking	ความหมาย
don't have	ไม่มีค่าใช้จ่ายในการจอดรถในแต่ละเดือน
ค่าที่เป็นตัวเลข	มีค่าใช้ใช้จ่ายในการจอดรถตามที่ระบุต่อเดือน

ตารางที่ 4.17 ค่าในแอททริบิวต์ maintain

parking	ความหมาย
don't have	ไม่มีค่าใช้จ่ายในการบำรุงรักษารถในแต่ละปี
ค่าที่เป็นตัวเลข	มีค่าใช้ใช้จ่ายในการบำรุงรักษารถตามที่ระบุต่อปี

ส่วนค่าในแอททริบิวต์ family ค่าที่ระบุจะเป็นจำนวนตัวเลข หมายถึง จำนวนสมาชิกที่อยู่ในครอบครัวตามที่ระบุไว้ โดยจากการทดสอบข้อมูล ได้ผลลัพธ์ ดังนี้

4.1.1 ผลการทดสอบของวิธี Decision Tree

การทดสอบข้อมูลด้วยวิธี Decision Tree โดยใช้ข้อมูลที่ไม่ทราบคลาสทั้งหมด 45 ตัวอย่าง ซึ่งมีลักษณะข้อมูลที่ทำนายได้ดังนี้

1. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิง และมีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส และมีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. มีอาชีพพนักงานบริษัทเอกชนที่มีเงินเดือน 10,001 ถึง 20,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือ เดินทางไปทำงาน โดยมีความเร็วเฉลี่ยในการขับขี่

ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร และมีค่าใช้จ่ายในการจอดรถเฉลี่ย 200 บาท มีค่าบำรุงรักษารถ 10,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 45-53 AND family = 4-6 AND distance = 11-20: SUV {CAR=0, TRUCK=0, SUV=3, NO=0}

2. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิง และมีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ และมีระดับการศึกษาปริญญาตรี อาชีพพนักงานรัฐวิสาหกิจ ที่มีเงินเดือน 40,001 ถึง 50,000 บาทต่อเดือน และมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 111-120 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 15,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

3. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิง ที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะหม้าย มีระดับการศึกษาสูงกว่าปริญญาตรี อาชีพข้าราชการ/เจ้าหน้าที่ของรัฐที่มีเงินเดือน 30,001 ถึง 40,000 บาทต่อเดือน โดยมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือ เดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 111 ถึง 120 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ มีค่าบำรุงรักษารถ 12,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 111-120 AND distance = >20: CAR {CAR=6, TRUCK=4, SUV=2, NO=0}

4. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี และมีสถานะสมรส มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 30,001 ถึง 40,000 บาทต่อเดือน และมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่มากกว่า 120 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 17,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = >120: CAR {CAR=15, TRUCK=6, SUV=6, NO=0}

5. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชายที่มีอายุระหว่าง 45 ถึง 53 ปี มีสถานะสมรส และมีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 20,001 ถึง 30,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 18,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

6. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด ระดับการศึกษาสูงกว่าปริญญาตรี อาชีพกำลังศึกษา และมีเงินเดือนน้อยกว่าหรือเท่ากับ 10,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือไม่จำเป็นต้องใช้ และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานศึกษาระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และไม่มีค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = 18-26: CAR {CAR=6, TRUCK=0, SUV=1, NO=0}

7. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45-53 ปี สถานะหม้าย มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 40,001 ถึง 50,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 45-53 AND family = 4-6 AND distance = >20: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}

8. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้าง/แยกกันอยู่ และมีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่เงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = yes AND sex = male: CAR {CAR=13, TRUCK=1, SUV=0, NO=0}

9. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชาย อายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส ระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระ และมีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 15,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 45-53 AND family = >6: TRUCK {CAR=1, TRUCK=2, SUV=1, NO=0}

10. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชาย และมีอายุระหว่าง18 ถึง 26 ปี สถานะสมรส ระดับการศึกษาสูงกว่าปริญญาตรี อาชีพพนักงานรัฐวิสาหกิจ มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะ การใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน ใช้มีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมงระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = employee AND sex = male: TRUCK {CAR=0, TRUCK=3, SUV=1, NO=0}

11. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชาย และมีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด ระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระ มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ อีกทั้งมีค่าบำรุงรักษารถ 16,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

12. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส/อยู่ด้วยกัน มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า และมีอาชีพเป็นพนักงานบริษัทเอกชนที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน และมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด

เดินทางไปทำงานและไม่จำเป็นต้องใช้ โดยใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร อีกทั้งมีค่าใช้จ่ายในการจอดรถเฉลี่ย 500 บาทต่อเดือน และมีค่าบำรุงรักษารถ 9,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = 45-53 AND license = have but not carry: CAR {CAR=6, TRUCK=1, SUV=0, NO=0}

13. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี และมีสถานะสมรส/อยู่ด้วยกัน ระดับการศึกษาปริญญาตรี อาชีพข้าราชการ/เจ้าหน้าที่ของรัฐ และมีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน มีการพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและไม่จำเป็นต้องใช้ มีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงาน ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ อีกทั้งมีค่าบำรุงรักษารถ 15,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = officer AND work = no: TRUCK {CAR=0, TRUCK=2, SUV=0, NO=0}

14. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี และมีสถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 7,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = yes AND sex = female AND salary = 30001-40000 AND family = 4-6: CAR {CAR= 1, TRUCK= 1, SUV=0, NO=0}

15. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส/อยู่ด้วยกัน มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า มีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ มีค่าบำรุงรักษารถ 8,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

16. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด มีระดับการศึกษาสูงกว่าปริญญาตรี และมีอาชีพกำลังศึกษาที่มีเงินเดือนน้อยกว่าหรือเท่ากับ 10,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถ คือ บรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน โดยใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานศึกษาระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

17. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะสมรส/อยู่ด้วยกัน ระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. อาชีพพนักงานบริษัทเอกชนที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 12,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

18. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ ระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 45-53 AND family = 4-6 AND distance = <=10: CAR {CAR=1, TRUCK=0, SUV=1, NO=0}

19. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะโสด ระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน โดยมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงาน น้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = 36-44 AND salary = 10001-20000: SUV {CAR=0, TRUCK=0, SUV=3, NO=0}

20. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า มีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของและเดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง โดยมีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 16,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

21.ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = yes AND sex = female AND salary = 10001-20000: CAR {CAR=6, TRUCK=0, SUV=2, NO=0}

22. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด มีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 17,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = 27-35 AND job = freelance: SUV {CAR=1, TRUCK=0, SUV=2, NO=0}

23. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหม้าย ระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนมากกว่า 50,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทาง

ไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = yes AND sex = female AND salary = >50000: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}

24. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 44 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด และเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 12,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

25. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี และมีสถานะสมรส/อยู่ด้วยกัน ระดับการศึกษาปริญญาตรี และมีอาชีพข้าราชการ/เจ้าหน้าที่ของรัฐที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน และมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 9,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = officer AND work = yes: CAR {CAR=10, TRUCK=2, SUV=0, NO=0}

26. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะแต่งงาน ระดับการศึกษาระดับปริญญาตรี อาชีพพนักงานบริษัทเอกชน มีรายได้ต่อเดือนระหว่าง 20,001 ถึง 30,000 บาท และมีใบอนุญาติขับขี่รถยนต์และพกพา ลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 10,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 111-120 AND distance = <=10 AND license = have and carry: CAR {CAR=11, TRUCK=0, SUV=0, NO=0}

27. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะแต่งงานแล้วและมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 40,001 ถึง 50,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ ลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 36-44: SUV {CAR=0, TRUCK=0, SUV=15, NO=0}

28. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ ลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111-120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

29. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้างและมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนมากกว่า 50,000 บาท มีและมีใบอนุญาตขับขี่รถยนต์และไม่พกพา ลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ มากกว่า 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 16,000 บาทต่อปีโมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

30. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะแต่งงานแล้วและมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 20,001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถโมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

31. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์และไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน และใช้ความเร็วในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 15,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

32. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน ประกอบอาชีพอื่น ๆ ที่นอกเหนือจากข้อมูลในแบบสอบถาม มีรายได้ต่อเดือนน้อยกว่าหรือเท่ากับ 10,000 บาท และมีใบอนุญาติขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยว โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน/สถานศึกษา คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = others AND license = have but not carry: TRUCK {CAR=1, TRUCK=2, SUV=0, NO=0}

33. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

34. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุมากกว่า 53 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาสูงกว่าระดับปริญญาตรี อาชีพพนักงานบริษัทเอกชน มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 101 ถึง 110 กิโลเมตร

ต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = >53: CAR {CAR=12, TRUCK=0, SUV=0, NO=0}

35. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด และมีระดับการศึกษาระดับปริญญาตรี อาชีพรับราชการ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์และพกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

36. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะหม้าย และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์และพกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและมีค่าบำรุงรักษารถอยู่ที่ 12,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

37. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง และมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 45-53 AND family = 4-6 distance = <=10: CAR {CAR=1, TRUCK=0, SUV=1, NO=0}

38. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะหย่าร้าง และมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไป

ทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

39. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี โมเดลทำนายว่าจะใช้รถอเนกประสงค์ (SUV)

กฎที่อ้างอิง : carry = no AND speed = 91-100 AND travel = no AND age = 27-35: SUV:{CAR=1, TRUCK=0, SUV=3, NO=0}

40. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุมากกว่า 53 ปี สถานะหม้าย และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 15,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

41. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีมากกว่า 53 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี ประกอบอาชีพอื่น ๆ ที่นอกเหนือจากข้อมูลในแบบสอบถาม มีรายได้ต่อเดือนน้อยกว่า 10,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยว โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน/สถานศึกษา คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และค่าบำรุงรักษารถ โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = others AND license = don't have: CAR {CAR=4, TRUCK=0, SUV=1, NO=0}

42. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน

อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนอยู่ที่ 40,0001 ถึง 50,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

43. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาอยู่ที่ปริญญาตรี ประกอบอาชีพพนักงานรัฐวิสาหกิจ มีรายได้ต่อเดือนอยู่ที่ 10,0001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 7,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = 101-110 AND age = 18-26: CAR {CAR=6, TRUCK=0, SUV=1, NO=0}

44. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่างร้าง และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนอยู่ที่ 20,0001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 19,000 บาทต่อปี โมเดลจะทำนายว่าจะใช้รถกระบะ (TRUCK)

กฎที่อ้างอิง : carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

45. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพรับราชการ มีรายได้ต่อเดือนอยู่ที่ 20,0001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและบำรุงรักษา โมเดลจะทำนายว่าจะใช้รถเก๋ง (CAR)

กฎที่อ้างอิง : carry = no AND speed = <=90 AND job = officer AND work = yes: CAR {CAR=10, TRUCK=2, SUV=0, NO=0}

4.1.2 ผลการทดสอบของวิธี Naïve Bayesian

การทดสอบข้อมูลด้วยวิธี Naïve Bayesian โดยใช้ข้อมูลที่ไม่ทราบคลาสทั้งหมด 45 ตัวอย่าง ซึ่งมีลักษณะข้อมูลที่ทำนายได้ดังนี้

1. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิง และมีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส และมีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. มีอาชีพพนักงานบริษัทเอกชนที่มีเงินเดือน 10,001 ถึง 20,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือ เดินทางไปทำงาน โดยมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร และมีค่าใช้จ่ายในการจอดรถเฉลี่ย 200 บาท มีค่าบำรุงรักษารถ 10,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.024, 0.011, 0.965 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

2. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิง และมีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ และมีระดับการศึกษาปริญญาตรี อาชีพพนักงานรัฐวิสาหกิจ ที่มีเงินเดือน 40,001 ถึง 50,000 บาทต่อเดือน และมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 111-120 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 15,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.047, 0.952, 0.001 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

3. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิง ที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะหม้าย มีระดับการศึกษาสูงกว่าปริญญาตรี อาชีพข้าราชการ/เจ้าหน้าที่ของรัฐที่มีเงินเดือน 30,001 ถึง 40,000 บาทต่อเดือน โดยมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือ เดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 111 ถึง 120 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ มีค่าบำรุงรักษารถ 12,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.689, 0.196, 0.115 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

4. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี และมีสถานะสมรส มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า และมีอาชีพเจ้าของ

ธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 30,001 ถึง 40,000 บาทต่อเดือน และมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่มากกว่า 120 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 17,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.442, 0.326, 0.231 และ 0.000 ตามลำดับ ดังนั้น คลาส คือ CAR

5. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชายที่มีอายุระหว่าง 45 ถึง 53 ปี มีสถานะสมรส และมีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 20,001 ถึง 30,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 18,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.013, 0.803, 0.184 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

6. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด ระดับการศึกษาสูงกว่าปริญญาตรี อาชีพกำลังศึกษา และมีเงินเดือนน้อยกว่าหรือเท่ากับ 10,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือไม่จำเป็นต้องใช้ และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานศึกษาระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และไม่มีค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.018, 0.000, 0.000 และ 0.982 ตามลำดับ ดังนั้นคลาส คือ NO

7. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45-53 ปี สถานะหม้าย มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือน 40,001 ถึง 50,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.111, 0.085, 0.805 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

8. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้าง/แยกกันอยู่ และมีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่เงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน และไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง

มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.424, 0.255, 0.320 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

9. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศชาย อายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส ระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระ และมีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 15,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.154, 0.457, 0.390 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

10. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชาย และมีอายุระหว่าง18 ถึง 26 ปี สถานะสมรส ระดับการศึกษาสูงกว่าปริญญาตรี อาชีพพนักงานรัฐวิสาหกิจ มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน ใช้มีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมงระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.466, 0.525, 0.009 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

11. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชาย และมีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด ระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระ มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ อีกทั้งมีค่าบำรุงรักษารถ 16,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.039, 0.957, 0.004 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

12. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส/อยู่ด้วยกัน มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า และมีอาชีพเป็นพนักงานบริษัทเอกชนที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน และมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัด เดินทางไปทำงานและไม่จำเป็นต้องใช้ โดยใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110

กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร อีกทั้งมีค่าใช้จ่ายในการจอดรถเฉลี่ย 500 บาทต่อเดือน และมีค่าบำรุงรักษารถ 9,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.771, 0.064, 0.165 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

13. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี และมีสถานะสมรส/อยู่ด้วยกัน ระดับการศึกษาปริญญาตรี อาชีพข้าราชการ/เจ้าหน้าที่ของรัฐ และมีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน มีการพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและไม่จำเป็นต้องใช้ มีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงาน ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ อีกทั้งมีค่าบำรุงรักษารถ 15,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.896, 0.031, 0.074 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

14. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี และมีสถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 7,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.353, 0.211, 0.437 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

15. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี และมีสถานะสมรส/อยู่ด้วยกัน มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า มีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 30,001 ถึง 40,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ มีค่าบำรุงรักษารถ 8,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.068, 0.904, 0.028 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

16. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คนเพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด มีระดับการศึกษาสูงกว่าปริญญาตรี และมีอาชีพกำลังศึกษาที่มีเงินเดือนน้อยกว่าหรือเท่ากับ 10,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถ คือ บรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน โดยใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึง

สถานศึกษาระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.428, 0.571, 0.000 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

17. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะสมรส/อยู่ด้วยกัน ระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. อาชีพพนักงานบริษัทเอกชนที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 12,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.063, 0.882, 0.054 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

18. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ ระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.434, 0.333, 0.233 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

19. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะโสด ระดับการศึกษาปริญญาตรี และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน โดยมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงาน น้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.712, 0.157, 0.131 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

20. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า มีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือบรรทุกของและเดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง โดยมีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่า

บำรุงรักษารถ 16,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.042, 0.815, 0.144 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

21.ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานน้อยกว่าหรือเท่ากับ 10 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 11,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.571, 0.170, 0.260 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

22. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด มีระดับการศึกษามัธยมศึกษาตอนปลายหรือ ปวช. และมีอาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน ไม่มีใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่ระหว่าง 101 ถึง 110 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 17,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.435, 0.307, 0.258 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

23. ผู้ที่มีครอบครัวจำนวนระหว่าง 4 ถึง 6 คน เพศหญิงที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหม้าย ระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนมากกว่า 50,000 บาทต่อเดือน มีและพกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน มีความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 14,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.137, 0.038, 0.825 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

24. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 44 ถึง 53 ปี สถานะหย่าร้าง/แยกกันอยู่ มีระดับการศึกษามัธยมศึกษาตอนต้นหรือต่ำกว่า อาชีพเจ้าของธุรกิจ/ประกอบอาชีพอิสระที่มีเงินเดือนระหว่าง 20,001 ถึง 30,000 บาทต่อเดือน โดยมีแต่ไม่พกพาใบอนุญาตในการขับขี่รถ และมีลักษณะการใช้งานรถคือบรรทุกของ ท่องเที่ยวผจญภัยต่างจังหวัด และเดินทางไปทำงาน ใช้ความเร็วเฉลี่ยในการขับขี่ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง และมีระยะทางจากบ้านถึงสถานที่ทำงานมากกว่า 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 12,000 บาทต่อปี ซึ่งคิดค่า confidence

ของ CAR, TRUCK, SUV และ NO ได้ 0.032, 0.967, 0.001 และ 0.000 ตามลำดับ ดังนั้น คลาส คือ TRUCK

25. ผู้ที่มีครอบครัวจำนวนมากกว่า 6 คน เพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี และมีสถานะสมรส/อยู่ด้วยกัน ระดับการศึกษาปริญญาตรี และมีอาชีพข้าราชการ/เจ้าหน้าที่ของรัฐที่มีเงินเดือนระหว่าง 10,001 ถึง 20,000 บาทต่อเดือน และมีและพกพาใบอนุญาตในการขับขี่รถ มีลักษณะการใช้งานรถคือท่องเที่ยวผจญภัยต่างจังหวัดและเดินทางไปทำงาน และมีความเร็วเฉลี่ยในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง มีระยะทางจากบ้านถึงสถานที่ทำงานระหว่าง 11 ถึง 20 กิโลเมตร ไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถ 9,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.816, 0.116, 0.068 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

26. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะแต่งงาน ระดับการศึกษาระดับปริญญาตรี อาชีพพนักงานบริษัทเอกชน มีรายได้ต่อเดือนระหว่าง 20,001 ถึง 30,000 บาท และมีใบอนุญาติขับขี่รถยนต์และพกพา ลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 10,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.908, 0.081, 0.011 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

27. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะแต่งงานแล้วและมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 40,001 ถึง 50,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ ลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.068, 0.036, 0.896 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

28. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ ลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111-120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือมากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ

ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.061, 0.906, 0.033 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

29. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะหย่าร้างและมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนมากกว่า 50,000 บาท มีและมีใบอนุญาตขับขี่รถยนต์และไม่พกพา ลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ มากกว่า 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 16,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.047, 0.912, 0.040 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

30. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะแต่งงานแล้วและมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 20,001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.138, 0.836, 0.026 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

31. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 36 ถึง 44 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์และไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน และใช้ความเร็วในการขับขี่น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 15,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.061, 0.928, 0.011 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

32. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน ประกอบอาชีพอื่น ๆ ที่นอกเหนือจากข้อมูลในแบบสอบถาม มีรายได้ต่อเดือนน้อยกว่าหรือเท่ากับ 10,000 บาท และมีใบอนุญาติขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยว โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่า

หรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน/สถานศึกษา คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.935, 0.060, 0.005 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

33. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยว และเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.122, 0.878, 0.001 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

34. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุมากกว่า 53 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาสูงกว่าระดับปริญญาตรี อาชีพพนักงานบริษัทเอกชน มีรายได้ต่อเดือนระหว่าง 10,001 ถึง 20,000 บาท มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 101 ถึง 110 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คืออยู่ระหว่าง 11 ถึง 20 กิโลเมตรและไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.589, 0.032, 0.379 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

35. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะโสด และมีระดับการศึกษาระดับปริญญาตรี อาชีพรับราชการ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์และพกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.211, 0.786, 0.003 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

36. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะหม้าย และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์และพกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ

และมีค่าบำรุงรักษารถอยู่ที่ 12,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.044, 0.953, 0.002 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

37. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่าร้าง และมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.280, 0.138, 0.582 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

38. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะหย่าร้าง และมีระดับการศึกษาระดับมัธยมปลาย อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.179, 0.817, 0.004 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

39. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.175, 0.213, 0.613 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ SUV

40. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศหญิงที่มีอายุมากกว่า 53 ปี สถานะหม้าย และมีระดับการศึกษาระดับปริญญาตรี อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนระหว่าง 30,001 ถึง 40,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมี

ค่าบำรุงรักษารถอยู่ที่ 15,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.256, 0.645, 0.098 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

41. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีมากกว่า 53 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาระดับปริญญาตรี ประกอบอาชีพอื่น ๆ ที่นอกเหนือจากข้อมูลในแบบสอบถาม มีรายได้ต่อเดือนน้อยกว่า 10,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยว โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน/สถานศึกษา คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและค่าบำรุงรักษารถ ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.920, 0.019, 0.061 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

42. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 27 ถึง 35 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนอยู่ที่ 40,0001 ถึง 50,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 91 ถึง 100 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 9,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.043, 0.940, 0.017 และ 0.000 ตามลำดับ ดังนั้น คลาส คือ TRUCK

43. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศชายที่มีอายุระหว่าง 18 ถึง 26 ปี สถานะแต่งงานแล้ว และมีระดับการศึกษาอยู่ที่ปริญญาตรี ประกอบอาชีพพนักงานรัฐวิสาหกิจ มีรายได้ต่อเดือนอยู่ที่ 10,0001 ถึง 20,000 บาท และมีใบอนุญาตขับขี่รถยนต์แต่ไม่พกพา มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางท่องเที่ยว และเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 111 ถึง 120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ อยู่ระหว่าง 11 ถึง 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 7,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.731, 0.247, 0.021 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

44. ผู้ที่มีจำนวนสมาชิกครอบครัวระหว่าง 4 ถึง 6 คน และเป็นเพศหญิงที่มีอายุระหว่าง 45 ถึง 53 ปี สถานะหย่างร้าง และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพเจ้าของธุรกิจหรือประกอบอาชีพอิสระ มีรายได้ต่อเดือนอยู่ที่ 20,0001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อบรรทุกของ เดินทางท่องเที่ยวและเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ ระหว่าง 111 ถึง

120 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ น้อยกว่าหรือเท่ากับ 10 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถ และมีค่าบำรุงรักษารถอยู่ที่ 19,000 บาทต่อปี ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.025, 0.974, 0.001 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ TRUCK

45. ผู้ที่มีจำนวนสมาชิกครอบครัวมากกว่า 6 คน และเป็นเพศชายที่มีอายุมากกว่า 53 ปี สถานะโสด และมีระดับการศึกษาอยู่ที่ระดับการศึกษาขั้นพื้นฐาน อาชีพรับราชการ มีรายได้ต่อเดือนอยู่ที่ 20,0001 ถึง 30,000 บาท และไม่มีใบอนุญาตขับขี่รถยนต์ มีลักษณะการใช้งานรถยนต์ คือ เพื่อเดินทางไปทำงาน โดยความเร็วเฉลี่ยในการขับรถ คือ น้อยกว่าหรือเท่ากับ 90 กิโลเมตรต่อชั่วโมง ระยะทางจากบ้านถึงที่ทำงาน คือ มากกว่า 20 กิโลเมตร และไม่มีค่าใช้จ่ายในการจอดรถและบำรุงรักษา ซึ่งคิดค่า confidence ของ CAR, TRUCK, SUV และ NO ได้ 0.763, 0.047, 0.190 และ 0.000 ตามลำดับ ดังนั้นคลาส คือ CAR

โดยตัวอย่างการทำนายตัวอย่างที่ไม่รู้คลาสทั้ง 45 ตัวอย่างนี้ เป็นส่วนหนึ่งของการทำนายค่าโดยโมเดลของวิธี Decision Tree และ Naïve Bayesian ซึ่งสามารถทำการทำนายตัวอย่างที่ไม่รู้คลาสได้อีกไม่จำกัด เพราะเป็นการใช้โมเดลในการทำนายที่มีอยู่แล้ว

ต่อมาจึงทำการสรุปผลการทดลองในวิธีต่างๆที่ได้ทำการทดลองและเสนอข้อเสนอแนะที่เป็นประโยชน์ต่อปัญหาพิเศษ ดังจะกล่าวในบทที่ 5 สรุปผลการวิจัยและเสนอแนะ

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

ปัญหาพิเศษนี้ได้ทำการศึกษาเกี่ยวกับการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์โดยการทำเหมืองข้อมูล ที่ใช้เทคนิค Decision Tree และ Naïve Bayesian เป็นตัวทดสอบได้ผลสรุป ดังนี้

5.1 ผลสรุปการวิจัย

ปัญหาพิเศษเรื่อง การวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์ มีวัตถุประสงค์เพื่อเพื่อศึกษาพฤติกรรมการใช้งานรถยนต์และทำนายข้อมูล จากกลุ่มตัวอย่างในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและจังหวัดฉะเชิงเทรา โดยมีขั้นตอนหลักๆ 5 ส่วน ได้แก่

ส่วนที่ 1 การกำหนดประชากรและเลือกกลุ่มตัวอย่างที่ต้องการทำการศึกษา โดยทางคณะผู้จัดทำได้เลือกกลุ่มตัวอย่างที่ต้องการศึกษา คือ ผู้ที่ใช้รถยนต์และไม่ใช้รถยนต์ที่อยู่ในบริเวณสถาบันพระจอมเกล้า เจ้าคุณทหาร ลาดกระบังและจังหวัดฉะเชิงเทรา และกำหนดจำนวนตัวอย่างที่ต้องเก็บจากสูตรการหาจำนวนตัวอย่างแบบไม่ทราบจำนวนประชากรที่แน่นอน จำนวน 450 ตัวอย่าง

ส่วนที่ 2 สร้างแบบสอบถามและเก็บข้อมูล จำนวน 450 ตัวอย่าง โดยทำการเก็บข้อมูลแบบวิธีการสุ่มตัวอย่างแบบเจาะจง (Purposive Sampling) โดยเจาะจงผู้ที่มีอายุ 18 ปีขึ้นไป ที่มีรถยนต์หรือไม่มีรถยนต์อยู่ในครอบครองในบริเวณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบังและจังหัวฉะเชิงเทรา และวิธีการสุ่มตัวอย่างแบบโควต้า (Quota Sampling) และวิธีการสุ่มตัวอย่างตามความสะดวก (Convenience Sampling)

ส่วนที่ 3 เก็บและรวบรวมข้อมูลจากแหล่งต่างๆที่เก็บมาให้มาอยู่ในแหล่งข้อมูลเดียวกันและเป็นมาตรฐานเดียวกันโดยใช้โปรแกรม Microsoft Excel เพื่อให้สามารถนำมาใช้กับโปรแกรม RapidMiner ได้โดยมี 2 ขั้นตอน คือ การโอนย้ายข้อมูล (Data Transfer) และการทำความสะอาดข้อมูล (Data Cleaning)

ส่วนที่ 4 เลือกเทคนิคที่ต้องการทำการทดสอบและทดสอบข้อมูลที่ไม่รู้คลาส คือ วิธี Decision Tree และ Naïve Bayes และทดสอบข้อมูล ซึ่งในการทำปัญหาพิเศษครั้งนี้ ได้ทดสอบเทคนิคการจำแนกกลุ่มจำนวน 2 วิธี คือ วิธี Decision Tree และ Naïve Bayes ในโปรแกรม

RapidMiner Studio จากการทดสอบข้างต้น พบว่า การจำแนกกลุ่มแบบ Decision Tree ให้ผลลัพธ์ดีที่สุด เนื่องจากมีความถูกต้องมากที่สุด สามารถแบ่งกลุ่มได้ตามเงื่อนไขที่ชัดเจนที่สุด โดยมีค่าความถูกต้องเป็น 72.73% และวิธี Naive Bayessian มีค่าความถูกต้องเป็น 59.09% จากนั้นทำการทดสอบข้อมูลที่ไม่รู้คลาสจำนวน 45 ข้อมูล

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
1. family = 4-6 AND sex = female AND age = 45-53 AND status = married AND education = hight school AND job = private employee AND salary = 10001-20000 AND licence = have and carry AND carry = no AND travel = no AND work = yes AND dispensable = yes AND other = no AND speed = 91-100 AND distance = 11-20 AND parking = 200 AND maintain = 10000	SUV	SUV	SUV
2. family = 4-6 AND sex = female AND age = 45-53 AND status =divorce AND education = bachelor AND job = employee AND salary = 40001-50000 AND licence = have and carry AND carry = yes AND travel = yes AND work = yes AND dispensable = yes AND other = no AND speed = 111-120 AND distance = >20 AND parking = don' have AND maintain = 15000	TRUCK	TRUCK	TRUCK
3. family = >6 AND sex = female AND age = 45-53 AND status = widaw AND education = bachelor AND job = officer AND salary = 30001-40000 AND licence = have and carry AND carry = no AND travel = no AND work = yes AND dispensable = yes AND other = no AND speed = 111-120 AND	CAR	CAR	TRUCK

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
distance = >20 AND parking = don' have AND maintain = 12000			
4. family = 4-6 AND sex = female AND age = 36-44 AND status = married AND education = primary AND job = freelance AND salary = 30001-40000 AND licence = have but carry AND carry = yes AND travel = yes AND work = yes AND dispensable = yes AND other = no AND speed = >120 AND distance = 11-20 AND parking = don' have AND maintain = 17000	CAR	CAR	CAR
5. family = >6 AND sex = male AND age = 45-53 AND status = married AND education = hight school AND job = freelance AND salary = 20001-30000 AND licence = don't have AND carry = yes AND travel = no AND work = yes AND dispensable = yes AND other = no AND speed = 91-100 AND distance = 11-20 AND parking = don' have AND maintain = 18000	TRUCK	TRUCK	TRUCK
6. family = >6 AND sex = female AND age = 18-26 AND status = single AND education = more than bachelor AND job = study AND salary = <10000 AND licence = don't have AND carry = no AND travel = no AND work = no AND dispensable = yes AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don' have AND maintain = don't have	CAR	NO	NO
7. family = 4-6 AND sex = female AND age = 45-53 AND status = widaw AND education = primary AND	SUV	SUV	SUV

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
job = freelance AND salary = 40001-50000 AND licence = don't have AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = >20 AND parking = don' have AND maintain = don't have			
8. family = >6 AND sex = male AND age = 36-44 AND status = divorce AND education = primary AND job = freelance AND salary = 30001-40000 AND licence = don't have AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = >20 AND parking = don' have AND maintain = 14000	CAR	CAR	CAR
9. family = >6 AND sex = male AND age = 45-53 AND status = married AND education = primary AND job = freelance AND salary = 30001-40000 AND licence = have and carry AND carry =no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = >20 AND parking = don' have AND maintain = 15000	TRUCK	TRUCK	CAR
10. family = 4-6 AND sex = male AND age = 18-26 AND status = married AND education = more than bechelor AND job = employee AND salary = 20001-30000 AND licence = have and carry AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = <=90 AND	TRUCK	TRUCK	TRUCK

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
distance = >20 AND parking = don' have AND maintain = 11000			
11. family = 4-6 AND sex = male AND age = 27-35 AND status = single AND education = hight school AND job = freelance AND salary = 10001-20000 AND licence = don't have AND carry =yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = >20 AND parking = don' have AND maintain = 16000	TRUCK	TRUCK	SUV
12. family = >6 AND sex = female AND age = 45-53 AND status = married AND education = primary AND job = private employee AND salary = 30001-40000 AND licence = have but not carry AND carry =no AND travel = yes AND work = yes AND dispensable = yes AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = 500 maintain = 9000	CAR	CAR	CAR
13. family = 4-6 AND sex = male AND age = 27-35 AND status = married AND education = bechelor AND job = officer AND salary = 10001-20000 AND licence = have and carry AND carry =no AND travel = yes AND work = no AND dispensable = yes AND other = no AND speed = <=90 AND distance = 11-20 AND parking = don't have maintain = 15000	TRUCK	CAR	TRUCK
14. family = 4-6 AND sex = female AND age = 36-44 AND status = divorce AND education = bechelor AND job = freelance AND salary = 30001-40000	CAR	SUV	TRUCK

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
AND licence = don't have AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = 11-20 AND parking = don't have maintain = 7000			
15. family = >6 AND sex = female AND age = 45-53 AND status = married AND education = primary AND job = freelance AND salary = 30001-40000 AND licence = don't have AND carry =yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = <=10 AND parking = don't have maintain = 8000	TRUCK	TRUCK	TRUCK
16. family = >6 AND sex = male AND age = 18-26 AND status = single AND education = bechelor AND job = study AND salary = <=10000 AND licence = don't have AND carry =yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don't have maintain = don't have	TRUCK	TRUCK	TRUCK
17. family = 4-6 AND sex = male AND age = 36-44 AND status = married AND education = hight school AND job = private employee AND salary = 20001-30000 AND licence = don't have AND carry =yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don't have maintain = 12000	TRUCK	TRUCK	TRUCK
18. family = 4-6 AND sex = female AND age = 45-53 AND status = divorce AND education = bechelor	CAR	CAR	CAR

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
AND job = freelance AND salary = 20001-30000 AND licence = don't have AND carry =no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = <=10 AND parking = don't have maintain = 11000			
19. family = 4-6 AND sex = male AND age = 36-44 AND status = single AND education = bechelor AND job = freelance AND salary = 10001-20000 AND licence = have and carry AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = <=10 AND parking = don't have maintain = 14000	SUV	CAR	SUV
20. family = 4-6 AND sex = female AND age = 45-53 AND status = divorce AND education = primary AND job = freelance AND salary = 10001-20000 AND licence = don't have AND carry =yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = <=10 AND parking = don't have maintain = 16000	TRUCK	TRUCK	TRUCK
21. family = >6 AND sex = female AND age = 36-44 AND status = divorce AND education = hight school AND job = freelance AND salary = 10001-20000 AND licence = don't have AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND	CAR	CAR	SUV

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
distance = <=10 AND parking = don't have maintain = 11000			
22. family = 4-6 AND sex = female AND age = 27-35 AND status = single AND education = hight school AND job = freelance AND salary = 10001-20000 AND licence = don't have AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = >20 AND parking = don't have maintain = 17000	SUV	CAR	SUV
23. family = 4-6 AND sex = female AND age = 36-44 AND status = widaw AND education = primary AND job = freelance AND salary = >50000 AND licence = have and carry AND carry =no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = >20 AND parking = don't have maintain = 14000	SUV	SUV	SUV
24. family = >6 AND sex = female AND age = 45-53 AND status = widaw AND education = primary AND job = freelance AND salary = 20001-30000 AND licence = have but not carry AND carry =yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = >20 AND parking = don't have maintain = 12000	TRUCK	TRUCK	CAR
25. sex = female AND age = 18-26 AND status = married AND education = bachelor AND job = officer AND salary = 10001-20000 AND family = 5 AND	CAR	CAR	CAR

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
license = have and carry AND carry = no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed <= 90 AND distance = 11-20 AND parking = don't have AND maintain = 9000			
26.sex = male AND age > 53 AND status = married AND education = bachelor AND job = private employee AND salary = 20001-30000 AND family = 5 AND license = have and carry AND carry = no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 111-120 AND distance <= 10 AND parking = don't have AND maintain = 10000	CAR	CAR	CAR
27. sex = male AND age = 36-44 AND status = married AND education = high school AND job = freelance AND salary = 40001-50000 AND family = 4 AND license = don't have AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance = 11-20 AND parking = don't have AND maintain = don't have	SUV	SUV	SUV
28. sex = female AND age = 45-53 AND status = married AND education = primary AND job = freelance AND salary = 30001-40000 AND family = 5 AND license = have and carry AND carry = yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance > 20 AND parking = don't have AND maintain = don't have	TRUCK	TRUCK	TRUCK

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
29. sex = male AND age = 36-44 AND status = divorce AND education = bachelor AND job = freelance AND salary > 50000 AND family = 3 AND license = have but not carry AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed > 120 AND distance = 11-20 AND parking = don't have AND maintain = 16000	TRUCK	TRUCK	TRUCK
30. sex = female AND age = 27-35 AND status = married AND education = bachelor AND job = freelance AND salary = 20001-30000 AND family = 2 AND license = don't have AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don't have AND maintain = don't have	TRUCK	TRUCK	TRUCK
31. sex = male AND age = 36-44 AND status = married AND education = bachelor AND job = freelance AND salary = 10001-20000 AND family = 4 AND license = have and carry AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed <= 90 AND distance <= 10 AND parking = don't have AND maintain = 15000	TRUCK	TRUCK	TRUCK
32. sex = female AND age = 18-26 AND status = married AND education = primary AND job = others AND salary <= 10000 AND family = 2 AND license = have but not carry AND carry = no AND travel = yes AND work = no AND dispensable = no AND other =	TRUCK	CAR	CAR

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
no AND speed <= 90 AND distance <= 10 AND parking = don't have AND maintain = don't have			
33. sex = female AND age = 18-26 AND status = single AND education = bachelor AND job = freelance AND salary = 10001-20000 AND family = 6 AND license = have but not carry AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 111-120 AND distance > 20 AND parking = don't have AND maintain = don't have	TRUCK	TRUCK	TRUCK
34. sex = female AND age > 53 AND status = married AND education = more than bachelor AND job = private employee AND salary = 10001-20000 AND family = 5 AND license = have and carry AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don't have AND maintain = don't have	CAR	CAR	CAR
35. sex = male AND age = 18-26 AND status = single AND education = bachelor AND job = officer AND salary = 30001-40000 AND family = 5 AND license = have and carry AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed <= 90 AND distance > 20 AND parking = don't have AND maintain = don't have	TRUCK	TRUCK	TRUCK
36. sex = male AND age = 27-35 AND status = widow AND education = bachelor AND job = freelance AND salary = 30001-40000 AND family = 2 AND license =	TRUCK	TRUCK	TRUCK

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
= have and carry AND carry = yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed 111-120 AND distance > 20 AND parking = don't have AND maintain = 12000			
37. sex = female AND age = 45-53 AND status = divorce AND education = high school AND job = freelance AND salary = 30001-40000 AND family = 3 AND license = don't have AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance <= 10 AND parking = don't have AND maintain = don't have	CAR	SUV	SUV
38. sex = male AND age > 53 AND status = divorce AND education = high school AND job = freelance AND salary = 30001-40000 AND family = 6 AND license = have but not carry AND carry = yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed <= 90 AND distance <= 10 AND parking = don't have AND maintain = 9000	TRUCK	TRUCK	TRUCK
39. sex = male AND age = 27-35 AND status = married AND education = bachelor AND job = freelance AND salary = 30001-40000 AND family = 2 AND license = don't have AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance > 20 AND parking = don't have AND maintain = 9000	SUV	SUV	TRUCK
40. sex = female AND age > 53 AND status = widow AND education = bachelor AND job = freelance AND	TRUCK	TRUCK	CAR

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
salary = 30001-40000 AND family = 6 AND license = have but not carry AND carry = yes AND travel = no AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance > 20 AND parking = don't have AND maintain = 15000			
41. sex = female AND age > 53 AND status = married AND education = bachelor AND job = others AND salary <= 10000 AND family = 7 AND license = don't have AND carry = no AND travel = yes AND work = no AND dispensable = no AND other = no AND speed <= 90 AND distance <= 10 AND parking = don't have AND maintain = don't have	CAR	CAR	SUV
42. sex = male AND age = 27-35 AND status = single AND education = primary AND job = freelance AND salary = 40001-50000 AND family = 3 AND license = don't have AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 91-100 AND distance > 20 AND parking = don't have AND maintain = 9000	TRUCK	TRUCK	CAR
43. sex = male AND age = 18-26 AND status = married AND education = bachelor AND job = employees AND salary = 10001-20000 AND family = 2 AND license = have but not carry AND carry = no AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 101-110 AND distance = 11-20 AND parking = don't have AND maintain = 7000	CAR	CAR	CAR

ตารางที่ 5.1 ตารางเปรียบเทียบผลลัพธ์ระหว่างวิธี Decision Tree และ Naïve Bayesian ในแต่ละตัวอย่าง (ต่อ)

ข้อมูลที่ใช้ทำนาย	Decision Tree	Naïve Bayesian	ผลลัพธ์จริง
44. sex = female AND age = 45-53 AND status = divorce AND education = primary AND job = freelance AND salary = 20001-30000 AND family = 3 AND license = don't have AND carry = yes AND travel = yes AND work = yes AND dispensable = no AND other = no AND speed = 111-120 AND distance <= 10 AND parking = don't have AND maintain = 19000	TRUCK	TRUCK	TRUCK
45. sex = male AND age > 53 AND status = single AND education = primary AND job = officer AND salary = 20001-30000 AND family = 5 AND license = don't have AND carry = no AND travel = no AND work = yes AND dispensable = no AND other = no AND speed <= 90 AND distance > 20 AND parking = don't have AND maintain = don't have	CAR	CAR	CAR

5.2 ปัญหาและแนวทางการแก้ไข

1. การสร้างแบบสอบถามไม่ได้มีการประเมินคุณภาพและความเชื่อมั่นของแบบสอบถามด้วยวิธีหาค่าสัมประสิทธิ์แอลฟ่าโดยสูตรของครอนบัค (Cronbach's alpha) จึงทำให้ไม่สามารถพิจารณาได้ว่าแบบสอบถามที่สร้างนั้นมีค่าความเชื่อมั่นเป็นเท่าใด

2. การเก็บข้อมูลเข้าสู่ฐานข้อมูลมีการบันทึกข้อมูลไม่ครบถ้วน จึงทำให้เกิดค่าขาดหายที่จำเป็นต้องกำจัดข้อมูลนั้นออก หากมีการบันทึกข้อมูลให้ถูกต้อง สมบูรณ์ ก็จะทำให้ได้ผลการจำแนกกลุ่มที่สมบูรณ์และผลการวิจัยที่น่าเชื่อถือมากยิ่งขึ้น

5.3 ข้อเสนอแนะ

การทำเหมืองข้อมูลเรื่องการวิเคราะห์พฤติกรรมเพื่อทำนายประเภทการใช้รถยนต์สามารถเป็นต้นแบบในการจัดทำโปรแกรมที่ช่วยในการตัดสินใจหรือทำนายการใช้รถยนต์ได้

เอกสารอ้างอิง

กัลยา วานิชบัญชา. 1999. **การวิเคราะห์สถิติ : สถิติเพื่อการวิจัย.** พิมพ์ครั้งที่ 4. กรุงเทพฯ. โรงพิมพ์จุฬาลงกรณ์วิทยาลัย.

ดลชาติ ตันติวานิช. (2557). **สถิติเบื้องต้น.** พิมพ์ครั้งที่ 4. กรุงเทพฯ: โครงการตำรา คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

ชญานันท์ นวพนอนันต์. 2016. **Weak VIII RapidMiner Studio 7.4.** [Online]. Available : e:///C:/Users/student/Downloads/SlideWeek8.pdf.

ชญานันท์ นวพนอนันต์. 2016. **Weak XI Naïve Bayes Classification.** [Online]. Available : e:///C:/Users/student/Downloads/SlideWeek11.pdf.

ชญานันท์ นวพนอนันต์. 2016. **Weak XIII RapidMiner Studio 7.4** (Decision Tree). [Online]. Available : e:///C:/Users/student/Downloads/SlideWeek13.pdf.

ชญานันท์ นวพนอนันต์. 2016. **Weak XIII RapidMiner Studio 7.4** (Naïve Bayes). [Online]. Available : e:///C:/Users/student/Downloads/SlideWeek131.pdf.

ดร. โกเมศ อัมพวัน. 2005. **การจำแนกประเภทและการทำนายข้อมูล (Classification and Prediction.**[Online]. Available : https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-Classification.pdf.

ทิพย์ธิดา วงศ์พิพันธ์. 2013. **“การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ”.** วิทยานิพนธ์วิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์.

นิตยา เกิดประสพ. 2003. **“การศึกษาเปรียบเทียบเทคนิคการจัดการข้อมูลสูญหายในการทำ เหมืองข้อมูลประเภทงานจำแนก”**. หน้า 3. ใน 29th Congress on Science and Technology of Thailand. ขอนแก่น. : คลังปัญญา มหาวิทยาลัยเทคโนโลยีสุรนารี.

ภาณุวัฒน์ ชุ่มชื่น. 2012. **“พฤติกรรมการตัดสินในซื้อรถยนต์ ยี่ห้อโตโยต้า รุ่นพรีอุส ของผู้บริโภคในเขตกรุงเทพมหานคร”.** วิทยานิพนธ์บริหารธุรกิจมหาบัณฑิต สาขาวิชาการตลาด มหาวิทยาลัยศรีนครินทร์วิโรฒ.

Foster Provost and Tom Fawcett. 2013. **Data Science for Business What you need to know about data mining and data-analytics thinking.** O’Reilly Media.

Jaiwei Han, Micheline Kamber and Jian Pei. 2012. **Data Mining Concepts and Techniques.** 3th Edition. Massachusetts. Elsevier.

krutu2507. **ความหมายและประเภทของข้อมูล.** [เว็บบล็อก]. สืบค้นจาก

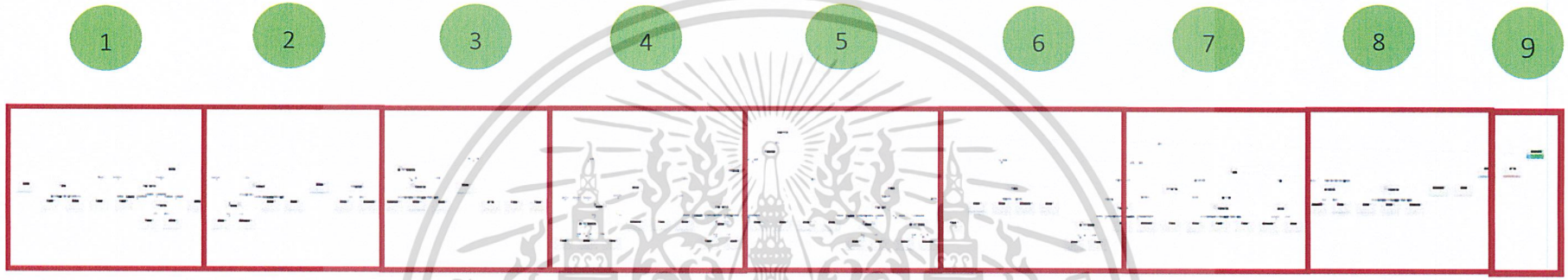
https://krutu2507.wordpress.com/%E0%B8%82%E0%B9%89%E0%B8%AD%E0%B8%A1%E0%B8%B9%E0%B8%A5/

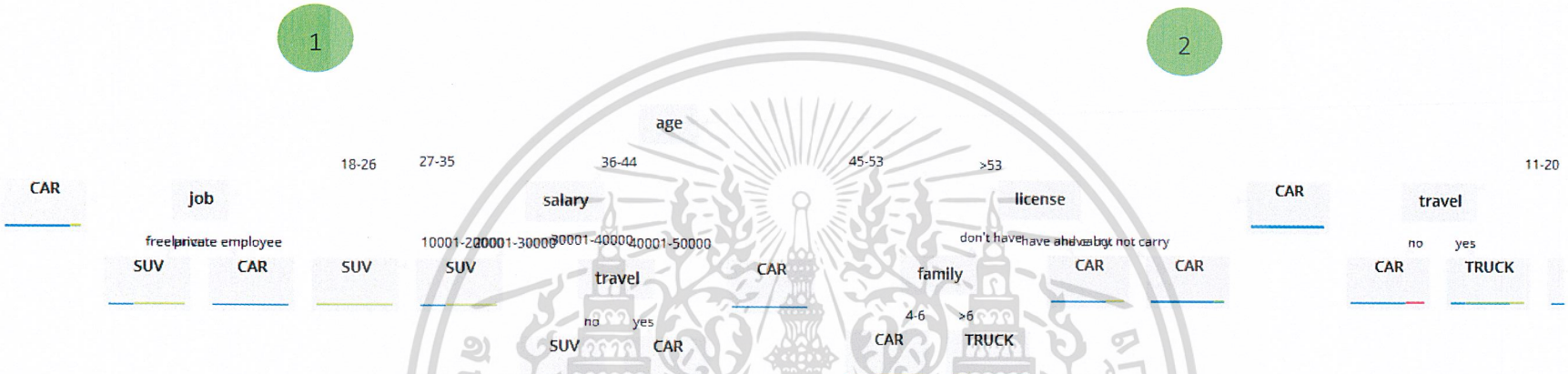
Rapid Miner. 2010. **Installing RapidMiner Studio.** [Online] . Available : https://docs.rapidminer.com/latest/studio/installation/.

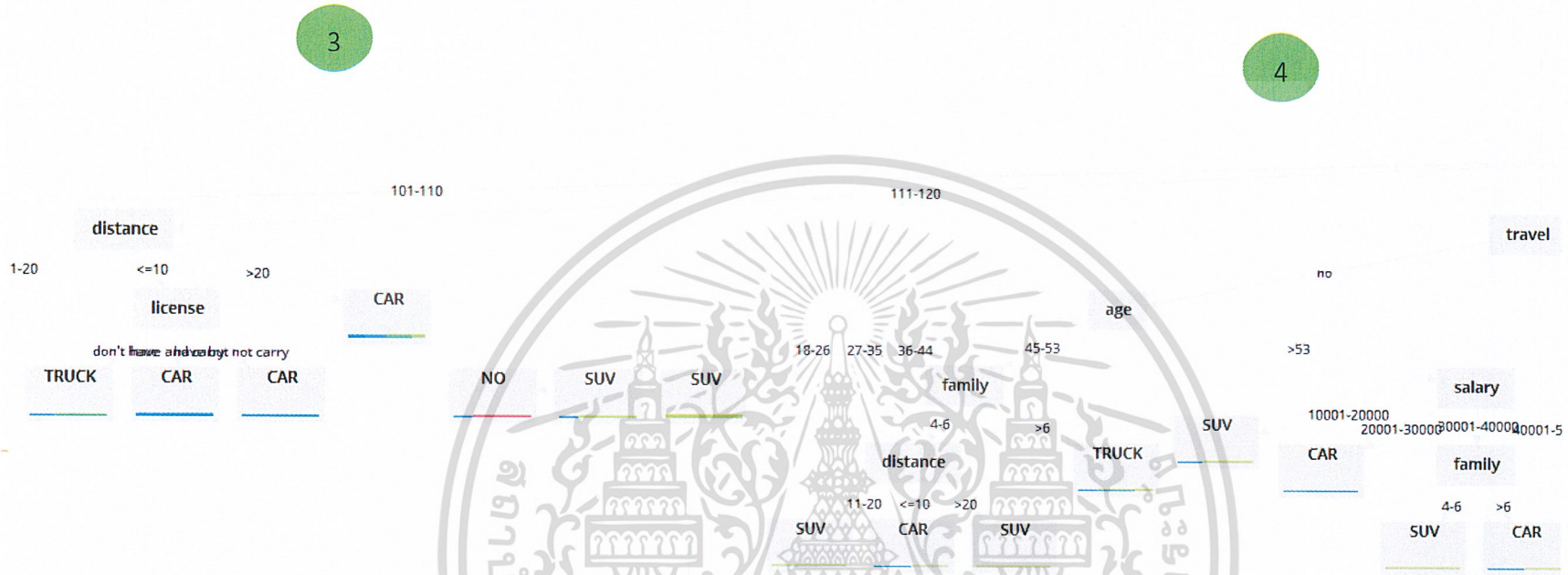
Rapid Miner. 2010. **Filtering rows and columns.** [Online] . Available : https://community.rapidminer.com/t5/RapidMiner-Studio-Forum/Filtering-rows-and-columns/td-p/6868.

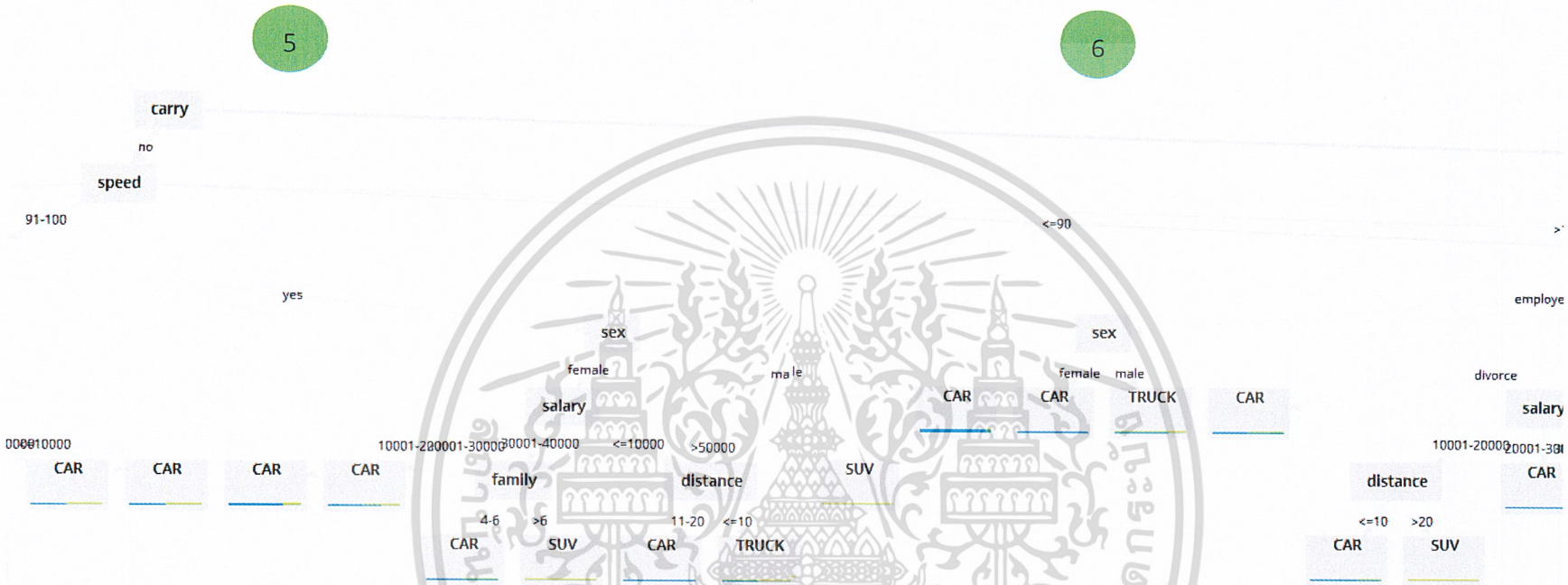
ภาคผนวก

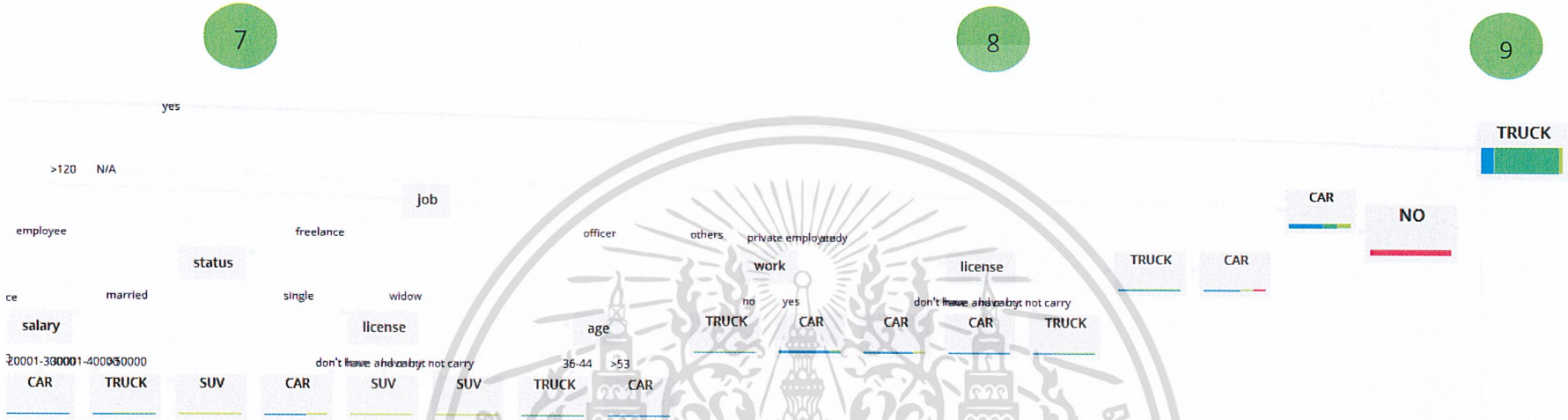
ต้นไม้ตัดสินใจแบบละเอียด








5
carry
no
speed
91-100
yes
sex
female
male
salary
family
distance
SUV
CAR
CAR
CAR
CAR
CAR
SUV
CAR
TRUCK
4-6
>6
11-20
<=10
10001-20000
30001-40000
<=10000
>50000
6
<=90
employe
sex
female
male
CAR
CAR
TRUCK
CAR
divorce
salary
distance
CAR
CAR
SUV
10001-20000
<=10
>20



ผลลัพธ์ของวิธี Decision Tree ในรูปแบบคำอธิบายแผนภาพต้นไม้

Tree

carry = no

| speed = 101-110

| | age = 18-26: CAR {CAR=6, TRUCK=0, SUV=1, NO=0}

| | age = 27-35

| | | job = freelance: SUV {CAR=1, TRUCK=0, SUV=2, NO=0}

| | | job = private employee: CAR {CAR=2, TRUCK=0, SUV=0, NO=0}

| | age = 36-44

| | | salary = 10001-20000: SUV {CAR=0, TRUCK=0, SUV=3, NO=0}

| | | salary = 20001-30000: SUV {CAR=1, TRUCK=0, SUV=2, NO=0}

| | | salary = 30001-40000

| | | | travel = no: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}

| | | | travel = yes: CAR {CAR=2, TRUCK=1, SUV=0, NO=0}

| | | salary = 40001-50000: CAR {CAR=2, TRUCK=0, SUV=0, NO=0}

| | age = 45-53

| | | license = don't have

| | | | family = 4-6: CAR {CAR=1, TRUCK=1, SUV=1, NO=0}

| | | | family = >6: TRUCK {CAR=0, TRUCK=3, SUV=0, NO=0}

| | | license = have and carry: CAR {CAR=3, TRUCK=0, SUV=1, NO=0}

| | | license = have but not carry: CAR {CAR=6, TRUCK=1, SUV=0, NO=0}

| | age = >53: CAR {CAR=12, TRUCK=0, SUV=0, NO=0}

| speed = 111-120

| | distance = 11-20

| | | travel = no: CAR {CAR=3, TRUCK=0, SUV=0, NO=1}

| | | travel = yes: TRUCK {CAR=1, TRUCK=3, SUV=1, NO=0}

| | distance = <=10

| | | license = don't have: TRUCK {CAR=1, TRUCK=2, SUV=0, NO=0}

| | | license = have and carry: CAR {CAR=11, TRUCK=0, SUV=0, NO=0}

| | | license = have but not carry: CAR {CAR=5, TRUCK=0, SUV=0, NO=0}

| | distance = >20: CAR {CAR=6, TRUCK=4, SUV=2, NO=0}

| speed = 91-100

| | travel = no

```
|  |  |  age = 18-26: NO {CAR=1, TRUCK=0, SUV=0, NO=3}
|  |  |  age = 27-35: SUV {CAR=1, TRUCK=0, SUV=3, NO=0}
|  |  |  age = 36-44: SUV {CAR=0, TRUCK=0, SUV=15, NO=0}
|  |  |  age = 45-53
|  |  |  |  family = 4-6
|  |  |  |  |  distance = 11-20: SUV {CAR=0, TRUCK=0, SUV=3, NO=0}
|  |  |  |  |  distance = <=10: CAR {CAR=1, TRUCK=0, SUV=1, NO=0}
|  |  |  |  |  distance = >20: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
|  |  |  |  family = >6: TRUCK {CAR=1, TRUCK=2, SUV=1, NO=0}
|  |  |  age = >53
|  |  |  |  salary = 10001-20000: SUV {CAR=1, TRUCK=0, SUV=2, NO=0}
|  |  |  |  salary = 20001-30000: CAR {CAR=3, TRUCK=0, SUV=0, NO=0}
|  |  |  |  salary = 30001-40000
|  |  |  |  |  family = 4-6: SUV {CAR=0, TRUCK=0, SUV=3, NO=0}
|  |  |  |  |  family = >6: CAR {CAR=1, TRUCK=0, SUV=1, NO=0}
|  |  |  |  salary = 40001-50000: CAR {CAR=1, TRUCK=0, SUV=1, NO=0}
|  |  |  |  salary = <=10000: CAR {CAR=2, TRUCK=0, SUV=2, NO=0}
|  |  travel = yes
|  |  |  sex = female
|  |  |  |  salary = 10001-20000: CAR {CAR=6, TRUCK=0, SUV=2, NO=0}
|  |  |  |  salary = 20001-30000: CAR {CAR=2, TRUCK=1, SUV=1, NO=0}
|  |  |  |  salary = 30001-40000
|  |  |  |  |  family = 4-6: CAR {CAR=1, TRUCK=1, SUV=0, NO=0}
|  |  |  |  |  family = >6: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
|  |  |  |  salary = <=10000
|  |  |  |  |  distance = 11-20: CAR {CAR=2, TRUCK=0, SUV=0, NO=0}
|  |  |  |  |  distance = <=10: TRUCK {CAR=0, TRUCK=1, SUV=1, NO=0}
|  |  |  |  salary = >50000: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
|  |  |  sex = male: CAR {CAR=13, TRUCK=1, SUV=0, NO=0}
|  speed = <=90
|  |  job = employee
|  |  |  sex = female: CAR {CAR=3, TRUCK=0, SUV=0, NO=0}
|  |  |  sex = male: TRUCK {CAR=0, TRUCK=3, SUV=1, NO=0}
```

| | job = freelance
| | | status = divorce: CAR {CAR=2, TRUCK=2, SUV=0, NO=0}
| | | status = married
| | | | salary = 10001-20000
| | | | | distance = <=10: CAR {CAR=2, TRUCK=1, SUV=0, NO=0}
| | | | | distance = >20: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
| | | | salary = 20001-30000: CAR {CAR=2, TRUCK=0, SUV=0, NO=0}
| | | | salary = 30001-40000: TRUCK {CAR=1, TRUCK=2, SUV=0, NO=0}
| | | | salary = >50000: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
| | | status = single
| | | | license = don't have: CAR {CAR=2, TRUCK=0, SUV=1, NO=0}
| | | | license = have and carry: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
| | | | license = have but not carry: SUV {CAR=0, TRUCK=0, SUV=2, NO=0}
| | | status = widow
| | | | age = 36-44: TRUCK {CAR=0, TRUCK=2, SUV=0, NO=0}
| | | | age = >53: CAR {CAR=3, TRUCK=0, SUV=0, NO=0}
| | job = officer
| | | work = no: TRUCK {CAR=0, TRUCK=2, SUV=0, NO=0}
| | | work = yes: CAR {CAR=10, TRUCK=2, SUV=0, NO=0}
| | job = others
| | | license = don't have: CAR {CAR=4, TRUCK=0, SUV=1, NO=0}
| | | license = have and carry: CAR {CAR=3, TRUCK=0, SUV=0, NO=0}
| | | license = have but not carry: TRUCK {CAR=1, TRUCK=2, SUV=0, NO=0}
| | job = private employee: TRUCK {CAR=1, TRUCK=5, SUV=0, NO=0}
| | job = study: CAR {CAR=3, TRUCK=0, SUV=1, NO=1}
| speed = >120: CAR {CAR=15, TRUCK=6, SUV=6, NO=0}
| speed = N/A: NO {CAR=0, TRUCK=0, SUV=0, NO=25}
carry = yes: TRUCK {CAR=22, TRUCK=113, SUV=6, NO=0}

RuleModel ของวิธี Dicision Tree

RuleModel

if carry = no and job = employee then SUV (2 / 1 / 6 / 0)

if carry = no and job = employees and work = no then NO (0 / 0 / 0 / 3)

if carry = no and job = employees and work = yes then CAR (10 / 3 / 2 / 0)

if carry = no and job = freelance then CAR (49 / 18 / 31 / 1)

if carry = no and job = officer and license = don't have and salary = 10001-20000 then CAR (9 / 0 / 0 / 0)

if carry = no and job = officer and license = don't have and salary = 20001-30000 and age = 36-44 then SUV (0 / 0 / 4 / 0)

if carry = no and job = officer and license = don't have and salary = 20001-30000 and age = 45-53 then SUV (0 / 0 / 3 / 0)

if carry = no and job = officer and license = don't have and salary = 20001-30000 and age = >53 then CAR (2 / 0 / 0 / 0)

if carry = no and job = officer and license = don't have and salary = 30001-40000 then SUV (1 / 1 / 6 / 0)

if carry = no and job = officer and license = don't have and salary = 40001-50000 then CAR (2 / 0 / 1 / 0)

if carry = no and job = officer and license = have and carry and status = divorce then CAR (6 / 0 / 0 / 0)

if carry = no and job = officer and license = have and carry and status = married and speed = 101-110 then CAR (3 / 0 / 1 / 0)

if carry = no and job = officer and license = have and carry and status = married and speed = 111-120 then CAR (3 / 0 / 0 / 0)

if carry = no and job = officer and license = have and carry and status = married and speed = 91-100 then SUV (0 / 0 / 2 / 0)

if carry = no and job = officer and license = have and carry and status = married and speed = <=90 and sex = female then CAR (2 / 0 / 0 / 0)

if carry = no and job = officer and license = have and carry and status = married and speed = <=90 and sex = male then TRUCK (0 / 2 / 0 / 0)

if carry = no and job = officer and license = have and carry and status = single then TRUCK (0 / 2 / 0 / 0)

if carry = no and job = officer and license = have and carry and status = widow then TRUCK (0 / 2 / 0 / 0)

if carry = no and job = officer and license = have but not carry then CAR (13 / 0 / 1 / 0)

if carry = no and job = official then CAR (5 / 0 / 0 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = 11-20 and sex = female then SUV (0 / 0 / 4 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = 11-20 and sex = male then CAR (2 / 0 / 0 / 1)

if carry = no and job = privaqte employee and license = don't have and distance = <=10 and family > 2.500 then TRUCK (0 / 4 / 0 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = <=10 and family ≤ 2.500 then CAR (1 / 0 / 1 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = >20 then SUV (0 / 0 / 2 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = >30 and travel = no then SUV (0 / 0 / 2 / 0)

if carry = no and job = privaqte employee and license = don't have and distance = >30 and travel = yes then CAR (1 / 1 / 0 / 0)

if carry = no and job = privaqte employee and license = have and carry then CAR (6 / 1 / 1 / 0)

if carry = no and job = privaqte employee and license = have but not carry then CAR (5 / 1 / 0 / 0)

if carry = no and job = private employee then TRUCK (0 / 2 / 0 / 0)

if carry = no and job = study and dispensable = no then CAR (12 / 1 / 2 / 0)

if carry = no and job = study and dispensable = yes then NO (0 / 0 / 0 / 22)

if carry = no and job = ค้าขาย and status = married and distance = <=10 then CAR (2 / 0 / 0 / 0)

if carry = no and job = ค้าขาย and status = married and distance = >20 then SUV (0 / 0 / 2 / 0)

if carry = no and job = ค้าขาย and status = married and distance = >30 then CAR (1 / 0 / 1 / 0)

if carry = no and job = ค้าขาย and status = widow then TRUCK (0 / 2 / 0 / 0)

if carry = no and job = ว่างงาน then CAR (13 / 5 / 3 / 3)

if carry = no and job = แม่บ้าน then CAR (3 / 2 / 3 / 0)

if carry = yes then TRUCK (22 / 113 / 6 / 0)

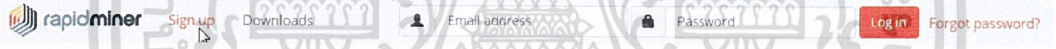
correct: 333 out of 450 training examples.

ขั้นตอนการดาวน์โหลดโปรแกรม RapidMiner Studio

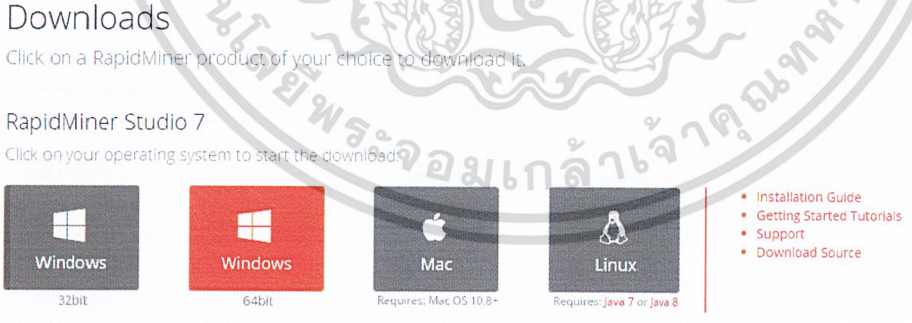
1. ไปที่ https://rapidminer.com/products/studio/ เพื่อทำการดาวน์โหลดโปรแกรม

2. คลิกปุ่ม Download ที่มุมบนขวา

3. หากต้องการสร้าง Account ให้กรอกข้อมูลต่างๆ ตามรูปด้านล่าง แต่ถ้าหากต้องการดสวน์โหลดโปรแกรมโดยไม่ต้องเข้าสู่ระบบ ให้คลิกปุ่ม Download



4. คลิกเลือกระบบปฏิบัติการที่ใช้ โดยโปรแกรมจะทำการไฮไลต์ระบบปฏิบัติการที่ใช้อยู่แล้ว



ขั้นตอนการติดตั้งโปรแกรม RapidMiner Studio ในระบบปฏิบัติการ Window

1. ดับเบิ้ลคลิกในไฟล์ที่ดาวน์โหลดมาแล้ว

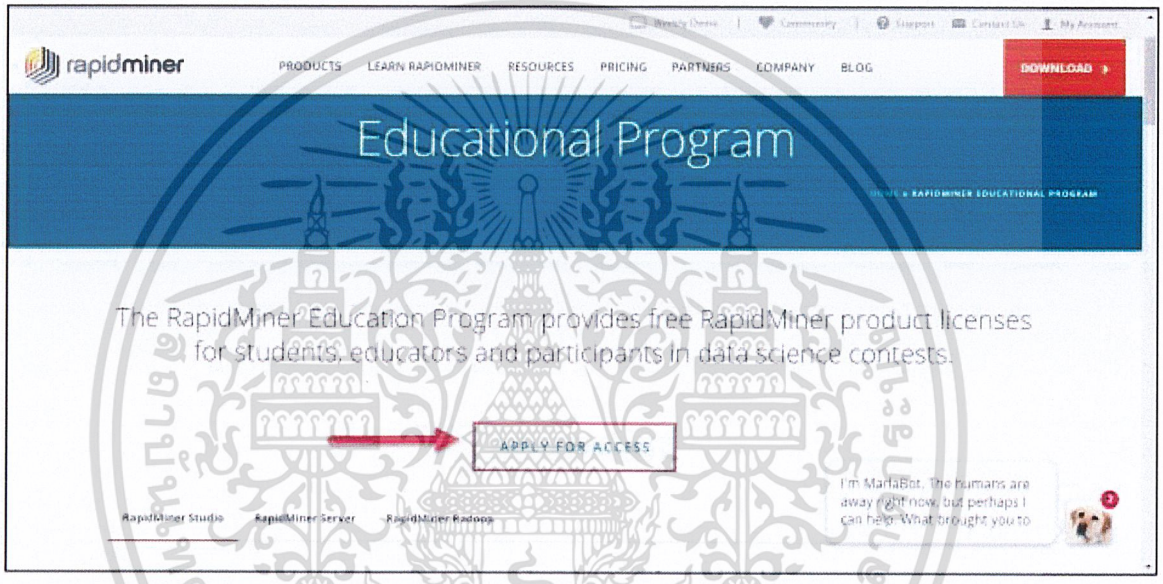
2. หากปรากฏหน้าต่างให้อนุญาติติดตั้งโปรแกรม ให้กดปุ่มอนุญาติ จากนั้นหน้าจอจะปรากฏหน้าต่างให้ติดตั้งโปรแกรม จากนั้นให้คลิก Next

3. เมื่อปรากฏหน้าต่างข้อตกลงและเงื่อนไขในการบริการให้คลิก Accepted เพื่อยอมรับเงื่อนไขการให้บริการ

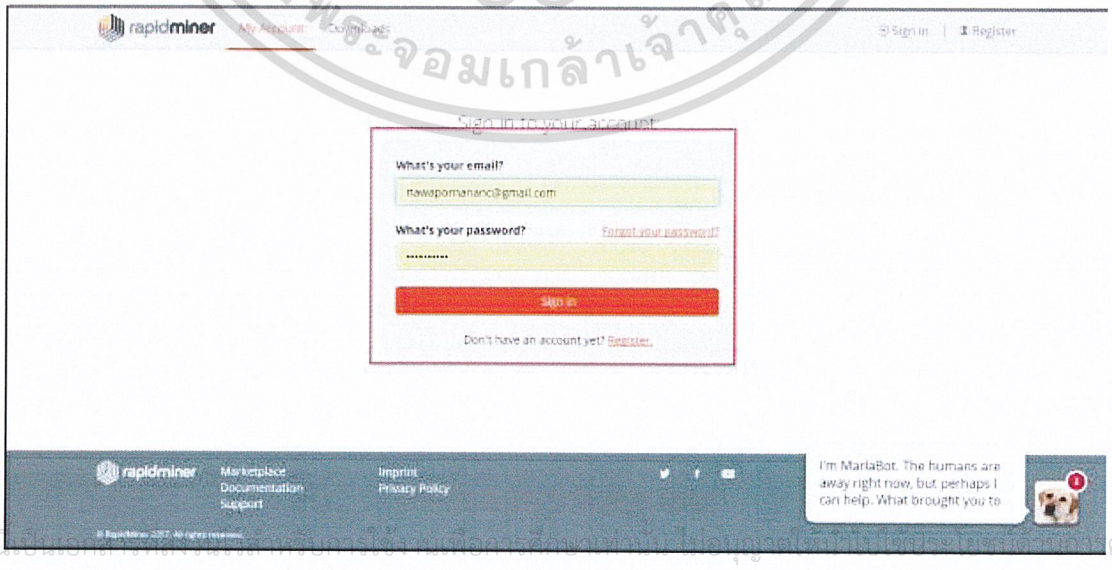
4. เลือกโฟล์เดอร์ทราต้องการจัดเก็บโปรแกรม จากนั้นคลิก Install เมื่อติดตั้งเสร็จเรียบร้อยแล้วให้คลิก Next และ Finish ตามลำดับ

ขั้นตอนการลงทะเบียนสำหรับ Educational Program

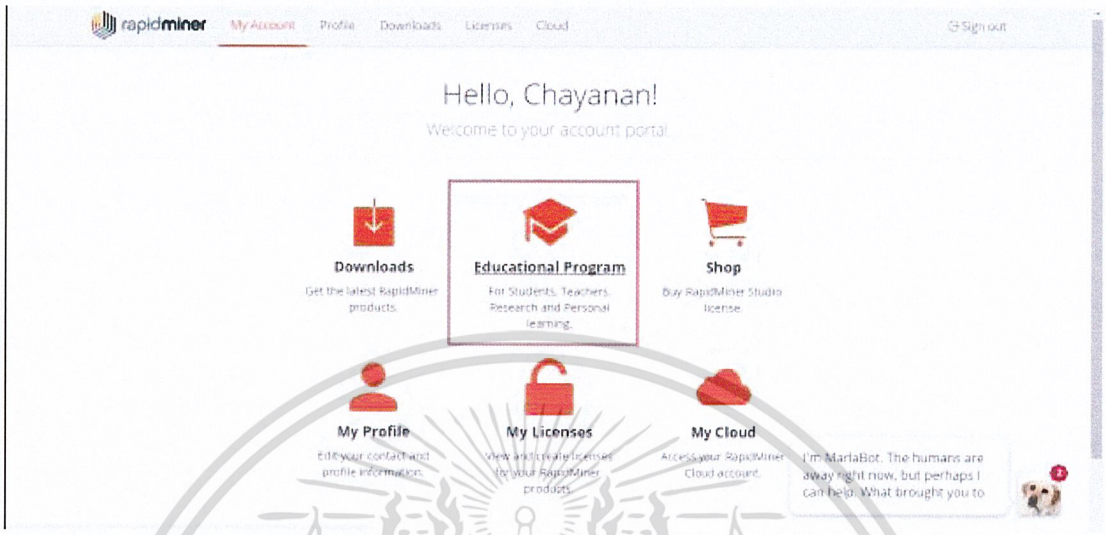
1. ยังไม่ต้องเปิดโปรแกรม Rapidminer
2. คลิกลิ้งค์ > https://rapidminer.com/educational-program/
3. ทำตามรูปด้านล่าง



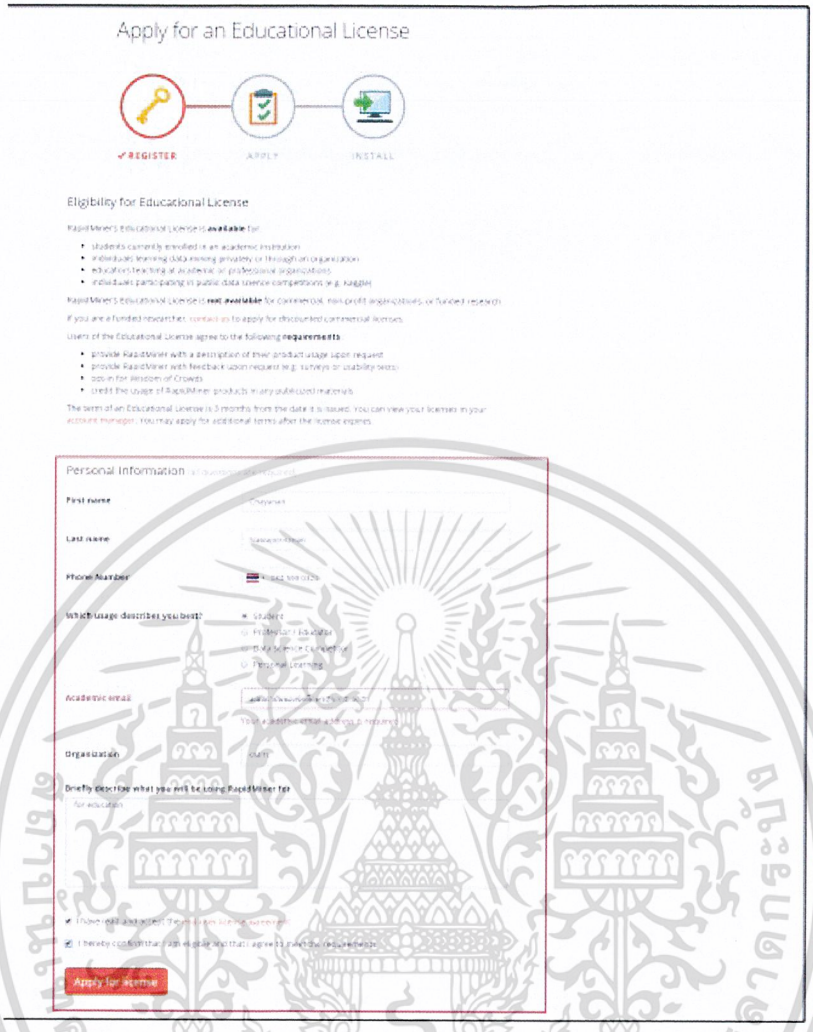
4. กรอก Email และพาสเวิร์ด ที่เคยลงทะเบียนกับระบบไว้ ตามรูปด้านล่าง



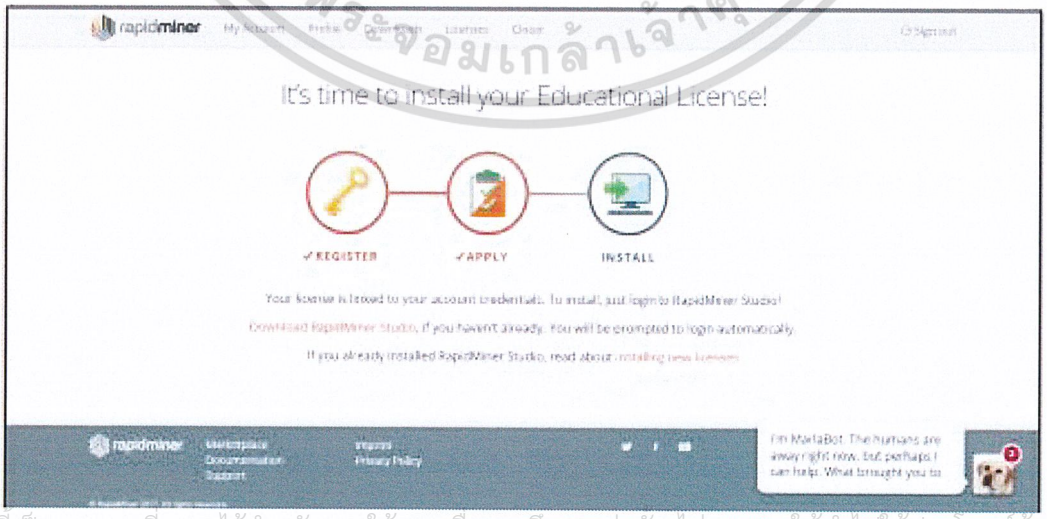
5. คลิกตามรูปด้านล่าง



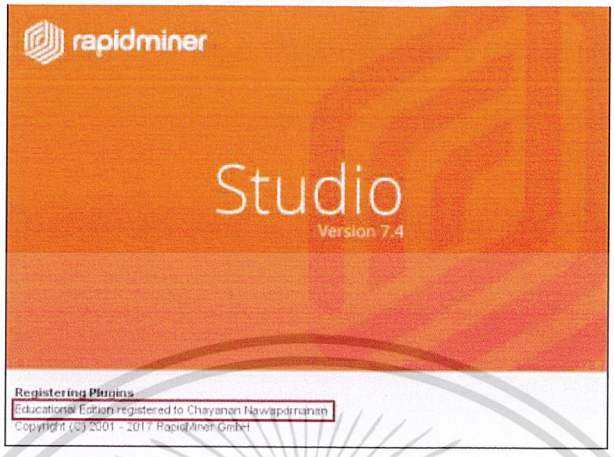
6. กรอกรายละเอียดของนักศึกษาตามรูปด้านล่าง



7. จากนั้นนักศึกษาจะได้อีเมล์ และหน้าจอจะแสดงดังรูปด้านล่าง



8. จากนั้นให้นักศึกษาเปิดโปรแกรม หน้าตาโปรแกรมจะเป็นแบบนี้



9. และพบหน้าหลักดังรูปด้านล่าง พร้อมทำงานต่อไป