ความแม่นยำการพยากรณ์อาชีพภายหลังการสำเร็จการศึกษา โดยการใช้เทคนิคเหมืองข้อมูลกรณีศึกษา หลักสูตรธุรกิจดิจิทัล

Accuracy of Career Prediction After Graduation Using Data Mining Techniques.

ทอแสง ใจงาม
นภสร ใจทับทิม

ภาคนิพนธ์เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบริหารธุรกิจบัณฑิต
รายวิชาโครงงานด้านเทคโนโลยีสารสนเทศ
ภาควิชาธุรกิจดิจิทัล คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยสยาม
พ.ศ. 2567

หัวข้อภาคนิพนธ์	ความแม่นยำการพยากรณ์อาชีพภายหลังการสำเร็จการศึกษา โดยการใช้เทคนิคเหมืองข้อมูลกรณีศึกษา หลักสูตรธุรกิจดิจิทัล Accuracy of Career Prediction After Graduation Using Data Mining Techniques		
หน่วยกิตของภาคนิพนธ์	3 หน่วยกิต		
คณะผู้จัดทำ	นางสาวทอแสง	ใจงาม	6105000007
	นางสาวนภสร	ใจทับทิม	6105000008
อาจารย์ที่ปรึกษา	อาจารย์ศรัญธร มั่งมี		
ระดับการศึกษา	บริหารธุรกิจบัณฑิต		
สาขาวิชา	ธุรกิจดิจิทัล		
ปีการศึกษา	2566		

อนุมัติให้ภาคนิพนธ์นี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบริหารธุรกิจบัณฑิต สาขาวิชาธุรกิจดิจิทัล

คณะกรรมการสอบภาคนิพนธ์

.....ประธานกรรมการ

(ดร. กันทิมา คงสถิตสุวรรณ)

..... กรรมการสอบ

(ผศ.ดร. พิชญากร เลค)

..... อาจารย์ที่ปรึกษา

(อาจารย์ศรัญธร มั่งมี)

หัวข้อภาคนิพนธ์	ความแม่นยำการพยากรณ์อาชีพภายหลังการสำเร็จการศึกษา โดยการใช้เทคนิคเหมืองข้อมูลกรณีศึกษา หลักสูตรธุรกิจดิจิทัล		
หน่วยกิตของภาคนิพนธ์	3 หน่วยกิต		
คณะผู้จัดทำ	นางสาวทอแสง	ใจงาม	6105000007
	นางสาวนภสร	ใจทับทิม	6105000008
อาจารย์ที่ปรึกษา	อาจารย์ศรัญธร มั่งมี		
ระดับการศึกษา	บริหารธุรกิจบัณฑิต		
สาขาวิชา	ธุรกิจดิจิทัล		
ปีการศึกษา	2566		

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ 1) เพื่อศึกษา วิเคราะห์ และเปรียบเทียบการคัดเลือกคุณลักษณะที่มีความเหมาะสมด้วย เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้ง (Bagging) และ เทคนิคการจำแนกข้อมูลเนอีฟเบย์ (Naive Bayes) และ 2) เพื่อประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพสำหรับ นักศึกษาระดับปริญญาตรี หลักสูตรด้านคอมพิวเตอร์ของมหาวิทยาลัยสยาม ด้วย รูปแบบที่วิเคราะห์ได้จากวัตถุประสงค์ข้อที่ 1 และประเมินประสิทธิภาพการใช้งาน

ผลจากการศึกษาส่วนการศึกษาแบบจำลองการทำเหมืองข้อมูลเพื่อแนะนำอาชีพด้านไอทีด้วยเทคนิคการศึกษา ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้ง และเทคนิคเนอีฟเบย์ พบว่าเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 98.86% ค่าความคลาดเคลื่อน 0.02% และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้ง มีค่าความถูกต้อง 98.86% ค่าความคลาดเคลื่อน 0.02% แสดงให้เห็นว่าเทคนิคต้นไม้ตัดสินใจและเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้ง นั้นเป็นแบบจำลองตามอัลกอริทึมดังกล่าว ซึ่งมีประสิทธิภาพมากที่สุดเหมาะจะนำไปพัฒนาระบบการแนะแนวอาชีพให้นักศึกษาระดับปริญญาตรี โดยประยุกต์ใช้แบบจำลองตามอัลกอริทึมข้างต้นพัฒนา

คำสำคัญ: การพยากรณ์อาชีพ, การจำแนกข้อมูล, เทคนิคเหมืองข้อมูล

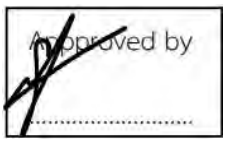
Project Title	Accuracy of Career Prediction After Graduation Using Data Mining Techniques		
Credits	3 Credits		
By	Miss Tosang	Jaingam	6105000007
	Miss Napasorn	Jaithubthim	6105000008
Advisor	Ms. Saranthon	Maungmee	
Degree	Bachelor of Business Administration		
Major	Digital Business		
Faculty	Information Technology		
Academic year	2023		

Abstract

This research aims: 1) to study, analyze, and compare the selection of appropriate characteristics by Decision Tree, Bagging, and Naive Bayes techniques; and 2) to apply data mining techniques to recommend careers for undergraduate students in the Computer Science program of Siam University using the analyzed model from Objective 1 and evaluate the efficiency of use.

The results of the data mining model to recommend IT careers, found that Decision Tree technique hads 98.86% accuracy with 0.02% error, and Bagging technique hads 98.86% accuracy with 0.02% error, indicating that Decision Tree technique and Bagging technique are the most efficient models according to the algorithm thegare suitable for developing a career guidance system for undergraduate students by applying the model according to the above algorithm.

Keywords: career forecasting, data classification, data mining techniques



กิตติกรรมประกาศ

ภาคนิพนธ์นี้จัดทำขึ้นเพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบริหารธุรกิจบัณฑิต สาขาธุรกิจดิจิทัล คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม การนำข้อมูลภาวะการมีงานทำกับข้อมูลระเบียนประวัติของผู้สำเร็จการศึกษาคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม ตั้งแต่ปี พ.ศ. 2553 – 2561 จำนวน 1,055 ระเบียน และข้อมูลเกรดเฉลี่ยของแต่ละรายวิชาที่นักศึกษาได้นำมาใช้ประโยชน์กับการทำงาน

ผู้วิจัยขอขอบคุณข้อมูลจากสำนักทะเบียนและวัดผล มหาวิทยาลัยสยามที่ให้ความร่วมมือในการให้ข้อมูลที่เป็นประโยชน์ต่องานวิจัย จนงานวิจัยสามารถสำเร็จลุล่วงไปได้ด้วยดี รวมทั้งผู้ที่มีส่วนเกี่ยวข้องที่ไม่สามารถระบุนามได้ ที่ให้ข้อมูลเพิ่มเติมและข้อเสนอแนะที่ทำให้งานวิจัยมีความสมบูรณ์ครบถ้วนตามวัตถุประสงค์ที่กำหนดไว้ ทางคณะผู้วิจัยขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

คณะผู้จัดทำ

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูปภาพ

ภาพที่ หน้า

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการศึกษาของประเทศไทยมีการพัฒนาและก้าวหน้ามากขึ้น มีเทคโนโลยีต่าง ๆ เข้ามาช่วยในการศึกษาให้มีประสิทธิภาพที่ดี จึงทำให้มีอาชีพมากมายในตลาดแรงงาน ส่งผลให้สถานศึกษาต้องเปิดหลักสูตรใหม่ ๆ ให้เป็นที่รับรองต่อตลาดแรงงานให้ได้มากที่สุด ดังนั้น จึงทำให้อนาคตในด้านการศึกษาของนักเรียนและนักศึกษาเป็นสิ่งสำคัญ การเลือกเรียนในด้านต่าง ๆ จึงเป็นตัวชี้วัดเส้นทางการทำงาน หรือสายอาชีพการทำงานในอนาคต เพื่อให้เพิ่มโอกาสในการเลือกอาชีพที่เหมาะสมกับความต้องการสอดคล้องกับความสามารถ สถาบันอุดมศึกษาจึงต้องผลิตบัณฑิตให้มีความรู้ความสามารถครบทุกด้าน พร้อมกับการส่งเสริมแนะแนวทางอาชีพให้กับนักศึกษา โดยเน้นให้คำปรึกษาทางด้านการประกอบอาชีพ อาทิเช่น การเลือกอาชีพ วิธีสมัครงาน โดยนักศึกษาจะมาขอคำปรึกษาและความช่วยเหลือจากอาจารย์หรือเจ้าหน้าที่ให้คำปรึกษาประจำศูนย์ให้คำปรึกษาการแนะแนวอาชีพในสถาบันอุดมศึกษาก็ยังไม่ได้ผลเต็มที่ นักศึกษาจำนวนมากประสบปัญหาในการตัดสินใจเลือกอาชีพให้ตรงกับตัวเอง มีเพียงจำนวนน้อยคนที่ตัดสินใจเลือกอาชีพที่ตรงกับตัวเอง การเลือกอาชีพผิดอาจจะทำให้เกิดการดำรงชีวิตที่ยากขึ้นต่อสภาพเศรษฐกิจในปัจจุบัน ซึ่งอาจจะส่งผลต่อความไม่พอใจในการทำงานปฏิบัติหน้าที่ได้ไม่เต็มประสิทธิภาพ เกิดความวิตกกังวลเกิดความเครียดและไม่ใส่ใจต่อสังคม

ผู้วิจัยจึงสนใจที่จะหาวิธีแก้ปัญหาดังกล่าว โดยการนำข้อมูลภาวะการมีงานทำกับข้อมูลระเบียนประวัติของผู้สำเร็จการศึกษาคณะเทคโนโลยีสารสนเทศ ตั้งแต่ปี พ.ศ. 2553 – 2561 จำนวน 1,055 ระเบียน และข้อมูลเกรดเฉลี่ยของแต่ละรายวิชาที่นักศึกษาได้นำมาใช้ประโยชน์กับการทำงานหรือความสอดคล้องต่อการมีงานทำอย่างไร มาศึกษาเทคนิคการพยากรณ์อาชีพโดยใช้เทคนิคเหมืองข้อมูล (Data Mining) ในการจำแนกข้อมูล (Classification) เพื่อเป็นการเพิ่มความแม่นยำในการจำแนกข้อมูลให้มีความ

ถูกต้องมากที่สุด จึงจำเป็นจะต้องเลือกใช้เทคนิควิธีจำแนกข้อมูลที่เหมาะสมกับข้อมูลและได้ค่าความแม่นยำที่อยู่ในระดับที่ยอมรับได้ด้วยการใช้คุณลักษณะทางด้านผลการเรียนของผู้สำเร็จการศึกษาที่ได้รับจากหลักสูตรด้านอาชีพของนักศึกษา ด้านเพศ ด้านตำแหน่งงาน ด้านสาขาวิชา จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ซึ่งข้อมูลที่ถูกเก็บไว้ในฐานข้อมูลหากเก็บไว้เฉย ๆ จึงไม่เกิดประโยชน์ ดังนั้น ต้องมีการสกัดเพื่อให้ได้สารสนเทศที่มีประโยชน์ การสกัดสารสนเทศ หมายถึง การคัดเลือกข้อมูลออกมาใช้งานในส่วนที่ต้องการในปัจจุบันการวิเคราะห์ข้อมูลจากฐานข้อมูลเดียวอาจไม่ให้ความรู้ที่เพียงพอและลึกซึ้งสำหรับการดำเนินงานวิจัยภายใต้ภาวะที่มีการแข่งขันสูงและมีการเปลี่ยนแปลงที่รวดเร็วจึงจำเป็นที่จะต้องรวบรวมข้อมูลหลาย ๆ ฐานข้อมูลเข้าด้วยกัน เรียกว่าคลังข้อมูล (Data Warehouse) เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่เดียวกัน

งานวิจัยนี้ได้เสนอเทคนิคการจำแนกข้อมูล 3 เทคนิคได้แก่ ต้นไม้ตัดสินใจ (Decision Tree – J48) แบ็กกิ้ง (Bagging) และนาอีฟเบย์ (Naïve Bays -Naïve Bayes) สำหรับวิเคราะห์ค่าความแม่นยำในการพยากรณ์ และนำเทคนิคที่ให้ค่าความแม่นยำมากที่สุดไปใช้เป็นตัวแบบในการพัฒนาระบบพยากรณ์เพื่อการแนะแนวอาชีพต่อไป

<u>วัตถุประสงค์การวิจัย</u>

1. เพื่อศึกษาการจัดทำเหมืองข้อมูลและสร้างโมเดล (Model) การพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรี คณะเทคโนโลยีสารสนเทศ สาขาธุรกิจดิจิทัลโดยใช้เทคนิคเหมืองข้อมูลที่เหมาะสม

2. เพื่อเปรียบเทียบความแม่นยำ (Accuracy) ของการวิเคราะห์เพื่อพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรี คณะเทคโนโลยีสารสนเทศ สาขาธุรกิจดิจิทัลโดยใช้เทคนิคเหมืองข้อมูล

3. เพื่อเป็นเครื่องมือในการตัดสินใจเกี่ยวกับงานภายหลังสำเร็จการศึกษา

ขอบเขตของการวิจัย

1. ขอบเขตของการศึกษา

ใช้เทคนิคการทำเหมืองข้อมูล (Data mining) เพื่อวิเคราะห์ค่าความแม่นยำในการพยากรณ์อาชีพ โดยนำข้อมูลภาวะการมีงานทำกับข้อมูลระเบียนประวัติของผู้สำเร็จการศึกษา คณะเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยสยาม ตั้งแต่ปี พ.ศ. 2553 – 2561 จำนวน 1,055 ระเบียน และข้อมูลเกรดเฉลี่ยของแต่ละรายวิชาข้อมูลประกอบไปด้วย อาชีพของนักศึกษา เพศ ความสอดคล้องสาขา และสาขาวิชา วิชาที่นำมาวิเคราะห์เป็นวิชาเอกบังคับดังตาราง

ตารางที่ 1.1 แสดงตัวอย่างข้อมูลที่ใช้ในการวิจัย

ลำดับที่	ชื่อแอตทริบิวต์	ค่าตัวแปร	คำอธิบาย
1	PREFIX_T	อักษร	เพศ
2	137-302	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์2
3	137-407	ระดับ	สัมมนาคอมพิวเตอร์ธุรกิจ
4	114-202	ระดับ	ภาษาอังกฤษธุรกิจ
5	130-202	ระดับ	การวิเคราะห์เชิงสถิติทางธุรกิจ
6	137-301	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์ 1
7	GPA	ระดับ	ผลสัมฤทธิ์การศึกษา
8	Target	อักษร	ความสอดคล้องทางอาชีพ

2. ขอบเขตด้านฮาร์ดแวร์ (Hardware) มีรายละเอียดดังนี้

2.1 CPU Intel Core i3-4030U ความเร็ว (1.90 GHz)

2.2 Hard Disk 1 TB.

2.3 RAM 4 GB

2.4 DVD-RW Driver

3. ขอบเขตด้านซอฟต์แวร์ (Software) มีรายละเอียดดังนี้

3.1 Windows 10

3.2 Microsoft Word Version 2016

3.3 Microsoft Excel Version 2016

3.4 Weka Version 3.8.6

4. อัลกอริทึมที่ใช้ในการวิเคราะห์

4.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

4.2 เทคนิคเบ็กกิ้ง (Bagging)

4.3 เทคนิคนาอีฟเบย์ (Naive Bayes)

ขั้นตอนการดำเนินงานวิจัย

1. รู้และเข้าใจปัญหา (Business understanding)

นักศึกษาคณะเทคโนโลยีสารสนเทศ สาขาธุรกิจดิจิทัล มหาวิทยาลัยสยาม หลังสำเร็จการศึกษามีการประกอบอาชีพตรงกับความสามารถของตนตามหลักสูตรหรือไม่ ผู้วิจัยจึงทำการเก็บรวบรวมข้อมูลอาชีพปัจจุบันของนักศึกษาที่สำเร็จการศึกษาแล้ว มาทำการพยากรณ์อาชีพจากเทคนิคที่เลือกทั้ง 3 เทคนิค เพื่อทำการวิเคราะห์ผลการศึกษาเพื่อหาค่าความถูกต้อง และทำการใช้เครื่องมือมาช่วยในการวิเคราะห์เพื่อให้ได้ผลลัพธ์การเลือกอาชีพที่เหมาะสมแก่นักศึกษา

2. สร้างฐานข้อมูลให้ครบ (Create a database)

ตารางที่ 1.2 แสดงข้อมูลนักศึกษาสาขาธุรกิจดิจิทัล ระหว่างปี พ.ศ. 2553-2561จำนวน 1,055 ระเบียน

ชนิดข้อมูล	จำนวนข้อมูลปี 2553 - 2561								
	2553	2554	2555	2556	2557	2558	2559	2560	2561
ข้อมูลนักศึกษา	156	140	124	154	145	117	69	61	89

3. เตรียมข้อมูลให้พร้อมใช้ (Data preparation)

3.1 ทำการสอบถามข้อมูลอาชีพปัจจุบันของศิษย์เก่า

3.2 คัดเลือกเกรดแต่ละรายวิชาตามหลักสูตรที่กำหนดเพื่อนำมาวิเคราะห์

3.3 แปลงข้อมูลให้เหมาะสมกับการวิเคราะห์

4. จัดทำและเลือกโมเดลที่ใช้ (Modeling) ข้อมูลแบ่งเป็น 2 ส่วนคือ

4.1 โมเดลแบ่งแยกตามภาควิชาต่าง ๆ ของหลักสูตร เช่น สาขาธุรกิจดิจิทัล

4.2 สร้างโมเดลด้วยเทคนิค Decision Tree ซึ่งจะได้โมเดลที่เข้าใจง่าย คำตอบของความสอดคล้องทางวิชา (Target) จะแบ่งเป็น 2 ประเภท คือ

4.2.1 ถ้าผลสัมฤทธิ์มากกว่าหรือเท่ากับ 3 จะสามารถวิเคราะห์ผลลัพธ์ของอาชีพ

4.2.2 ถ้าผลสัมฤทธิ์น้อยกว่า 3 โปรแกรมจะทำการวิเคราะห์ต่อไปจนกว่าจะไม่ตรงกับผลลัพธ์ใดๆ

5. ประเมินประสิทธิภาพของโมเดล (Evaluation)

5.1 ทดสอบข้อมูลด้วยเครื่องมือที่นำมาทำการวิเคราะห์

5.2 คำนวณหาค่าความถูกต้อง

6. การนำโมเดลที่ได้ไปใช้งาน (Deployment)

6.1 ทำการพิจารณาจากเกรดตามโมเดลที่สร้าง

6.2 นำเสนอผลลัพธ์ที่ได้ เพื่อนำไปใช้ประโยชน์ในการวางแผน กำหนดกลยุทธ์ และดำเนินการต่าง ๆ เพื่อผลประโยชน์ต่อไป

<u>ประโยชน์ที่คาดว่าจะได้รับ</u>

1. เพื่อเข้าใจเทคนิคการจัดทำเหมืองข้อมูล และการสร้างโมเดลเพื่อวิเคราะห์ผลที่เป็นไปได้อย่างแม่นยำ

2. ทำให้การพยากรณ์มีแม่นยำน่าเชื่อถือ และได้ข้อมูลที่มีประสิทธิภาพมากที่สุด

3. สามารถนำผลลัพธ์ของโมเดลที่ได้ทำการวิเคราะห์ไปใช้ในการตัดสินใจภายหลังสำเร็จการศึกษา

บทที่ 2
แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

จากการศึกษาและค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้องกับการศึกษาเรื่องการใช้เหมืองข้อมูลช่วยในการพยากรณ์อาชีพของนักศึกษาคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม ทำให้สามารถจำแนกและประมวลความรู้ที่ได้จากการทบทวนเอกสาร ออกเป็น 2 ส่วนได้แก่ แนวคิดทฤษฎี และงานวิจัยที่เกี่ยวข้อง ซึ่งมีรายละเอียด ดังนี้

แนวคิดและทฤษฎีที่เกี่ยวข้อง

เหมืองข้อมูล (Data Mining)

ในปัจจุบันด้วยความก้าวหน้าทางเทคโนโลยีสารสนเทศทำให้เราสามารถจัดเก็บข้อมูลจำนวนมากเพื่อช่วยในการตัดสินใจและมีผลต่อการดำเนินการต่าง ๆ ดังนั้น การทำเหมืองข้อมูลจึงเป็นเทคนิคหนึ่งที่สำคัญที่จะเข้ามาช่วยจัดการข้อมูลที่มีขนาดใหญ่และมีจำนวนมากโดยมีรายละเอียดดังนี้

ความหมายของเหมืองข้อมูล (Data Mining)

การค้นหารูปแบบความสัมพันธ์ของความรู้จากฐานข้อมูลขนาดใหญ่ที่ซับซ้อน สะดวกมีประสิทธิภาพ ลดระยะเวลาและค่าใช้จ่าย จากเหตุดังกล่าวจึงจำเป็นจะต้องมีการทำเหมืองข้อมูล เนื่องจากเป็นวิธีการที่ช่วยตอบสนองความต้องการนี้ได้เป็นอย่างดีอาจกล่าวได้ว่าการทำเหมืองข้อมูล เป็นเครื่องมือที่เสริมสร้างสารสนเทศและข้อความรู้เพื่อการตัดสินใจที่สำคัญในกรณีที่มีข้อมูลขนาดใหญ่ ซึ่งวิธีการสอบถามข้อมูลและวิธีการวิเคราะห์เชิงสถิติโดยทั่วไปอาจไม่สามารถตอบสนองได้ในลักษณะเดียวกัน โดยความหมายของการทำเหมืองข้อมูลนั้นได้มีผู้ให้ความหมายไว้หลากหลาย ดังนี้

สุรพงศ์ เอื้อวัฒนามงคล (2557) ได้ให้ความหมายการทำเหมืองข้อมูลว่าเป็นขบวนการวิเคราะห์ข้อมูลที่มีขั้นตอนเพื่อให้ได้มาซึ่งตัวแบบ (Pattern) ซึ่งแสดงความสัมพันธ์ระหว่างข้อมูลโดยผลลัพธ์ความรู้เกี่ยวกับข้อมูลที่ถูกต้องสามารถนำไปใช้ในการตัดสินใจและดำเนินงานได้โดยไม่ผิดพลาดหรือสร้างความเสียหายจากการนำไปใช้งาน

สายชล สินสมบูรณ์ทอง (2558) ได้ให้ความหมายการทำเหมืองข้อมูลว่า เป็นกระบวนการทำงานที่สกัดข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่เพื่อให้ได้สารสนเทศที่มีประโยชน์ที่เรายังไม่ทราบโดยเป็นสารสนเทศที่มีเหตุผลและสามารถนำไปใช้ได้ซึ่งเป็นสิ่งสำคัญที่จะช่วยการตัดสินใจในการดำเนินงานต่าง ๆ โดยการทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (KDD) ซึ่งการทำเหมืองข้อมูลจะสามารถนำมาคาดการณ์การพัฒนาการวิวัฒนาการของอนาคตได้ซึ่งการทำเหมืองข้อมูลนับเป็นหนึ่งใน 10 เทคโนโลยีที่เกิดขึ้นใหม่ที่จะทำให้เกิดการเปลี่ยนแปลงเนื่องจากองค์กรต่าง ๆ ได้มีการเก็บข้อมูลไว้ในคลังข้อมูลจำนวนเพิ่มมากขึ้น สารสนเทศที่จะนำมาวิเคราะห์เพื่อให้ประสบผลสำเร็จตามกลยุทธ์และเป้าหมายนั้นจะต้องพิจารณาจากข้อมูลที่มีอยู่ว่าสามารถนำมาทำอะไรได้บ้าง

สุชาดา กีระนันทน์ (2545) ได้ให้ความหมายของการทำเหมืองข้อมูลว่า เป็นกระบวนการค้นหาสารสนเทศหรือข้อความรู้ที่อยู่ในฐานข้อมูลขนาดใหญ่ที่ซับซ้อน เพื่อนำความรู้ที่ได้ไปใช้ประโยชน์ในการตัดสินใจสารสนเทศที่ได้อาจนำมาสร้างการพยากรณ์หรือสร้างตัวแบบสำหรับการจำแนกหน่วยหรือกลุ่มหรือแสดงความสัมพันธ์ระหว่างหน่วยต่าง ๆ หรือให้ข้อสรุปของสาระในฐานข้อมูลการทำเหมืองข้อมูลประกอบขึ้นด้วยการนำกระบวนการทางสถิติและการเรียนรู้ผ่านระบบคอมพิวเตอร์ เพื่อสร้างตัวแบบกฎเกณฑ์รูปแบบการพยากรณ์และข้อความรู้จากฐานข้อมูลขนาดใหญ่ โดยการทำเหมืองข้อมูลมีขั้นตอนการดำเนินงานหลายขั้นตอนซึ่งต้องอาศัยเทคนิคหรือวิธีการต่าง ๆ เช่น วิธีการจัดกลุ่มการค้นหาความสัมพันธ์การพยากรณ์ เป็นต้น ดังนั้น ถ้ามีฐานข้อมูลขนาดใหญ่ที่มีข้อมูลคุณภาพดีเทคโนโลยีการทำเหมืองข้อมูลจะช่วยในการค้นหรือแสวงหาโอกาสทางธุรกิจใหม่

โอม ศรนิล (2556) ได้ให้ความหมายของการทำเหมืองข้อมูลว่าเป็นการค้นหารูปแบบและความสัมพันธ์ในชุดข้อมูลขนาดใหญ่โดยทำงานในลักษณะกึ่งอัตโนมัติอาศัยความสามารถในการคำนวณของ

คอมพิวเตอร์และความรู้เกี่ยวกับธุรกิจของผู้ใช้ซอฟต์แวร์จะช่วยค้นหารูปแบบที่เป็นไปได้จากข้อมูลขน าดใหญ่

David Hand Heikki Mannila and Padhraic Smyth (2001) ได้ให้ความหมายของการทำเหมืองข้อมูลว่า เป็นการวิเคราะห์เซตข้อมูลเชิงสังเกต (ขนาดใหญ่) เพื่อหาความสัมพันธ์ที่ไม่ได้มีการคาดการณ์ไว้ล่วงหน้าและเพื่อสรุปข้อมูลในวิธีที่เข้าใจได้และเป็นประโยชน์ต่อเจ้าของข้อมูลความสัมพันธ์ และส่วนสรุปต่าง ๆ ผ่านการทดลองทำเหมืองข้อมูลที่ใช้อ้างถึงในฐานะต้นแบบหรือโครงสร้าง ตัวอย่างเช่น สมการเชิงเส้น กฎ คลัสเตอร์ กราฟ แขนงโครงสร้าง และโครงสร้างที่วนซ้ำในช่วงเวลา

ญาใจ ลิ่มปิยกรณ์ (2553) ได้ให้ความหมายของการทำเหมืองข้อมูลว่าเป็นการค้นพบความรู้ในฐานข้อมูล (Knowledge Discovery in Databases-KDD) ซึ่งเป็นการค้นหาแบบรูปและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลขนาดใหญ่ ในปัจจุบันข้อมูลปริมาณมหาศาลถูกกักเก็บอยู่ในคอมพิวเตอร์แต่ข้อมูลเหล่านั้นไม่สามารถถูกนำมาใช้ประโยชน์อย่างเต็มที่ กระบวนการทำเหมืองข้อมูลเป็นวิธีการที่ใช้ในการสกัดความรู้ (Knowledge) ใหม่ ๆ ออกมาจากข้อมูลที่ถูกเก็บไว้ เพื่อนำความรู้นั้นไปใช้ประโยชน์ต่อไปหลายด้าน การทำเหมืองข้อมูลจึงเป็นที่นิยมแพร่หลายทั้งในแวดวงธุรกิจและในเชิงงานวิจัย

เนื่องด้วยปัจจุบันเป็นยุคที่ข้อมูลสารและสนเทศมีความสำคัญ การเผยแพร่และสื่อสาร ข้อมูลข่าวสาร ที่ตรงกับความต้องการของผู้ใช้จึงเป็นสิ่งจำเป็น การประยุกต์เทคโนโลยีสารสนเทศเพื่อช่วยในการสื่อสารข้อมูลจำนวนมากให้แก่ผู้ใช้ เช่น การให้บริการเว็บไซต์เพื่อเผยแพร่ข้อมูลข่าวสารและแลกเปลี่ยนความรู้ จึงเป็นเครื่องมือที่สำคัญในการสื่อสารข้อมูลถึงผู้ใช้จำนวนมาก ดังนั้นการศึกษาเกี่ยวกับพฤติกรรมของผู้ใช้บริการเว็บไซต์ จะช่วยให้องค์กรสามารถนำข้อมูลมาใช้ในการวางแผนพัฒนาเว็บไซต์ ให้ตรงกับความต้องการใช้งานหรือใช้ในการวางแผนกลยุทธ์ เพื่อสร้างความได้เปรียบทางการแข่งขัน

การทำเหมืองข้อมูลการทำเหมืองข้อมูล (Data Mining) คือ การค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูลแต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากโดยการทำเหมืองข้อมูลจะเหมาะกับการแก้ปัญหาบางชนิดเท่านั้น มีเทคนิคต่าง ๆ ที่ใช้ในการแก้ไขปัญหาอยู่หลายเทคนิค ซึ่งไม่มีเทคนิคใด

สามารถแก้ปัญหาได้ทุกปัญหา ดังนั้น ความหลากหลายของเทคนิคเป็นสิ่งที่จำเป็นที่จะนำไปสู่วิธีการแก้ปัญหาที่ดีที่สุดของการทำเหมืองข้อมูล

การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

จากความหมายของการทำเหมืองข้อมูลข้างต้น จะพบว่าการทำเหมืองข้อมูลเป็นสิ่งสำคัญที่มีการนำข้อมูลขนาดใหญ่มาค้นหารูปแบบความสัมพันธ์ของผลลัพธ์ที่ดีที่สุด ช่วยในการวิเคราะห์และตัดสินใจซึ่งอาศัยสภาพของข้อมูลที่มีอยู่นำมาเข้ากระบวนการวิเคราะห์ตามหลักคณิตศาสตร์สถิติและการเรียนรู้ของเครื่องได้ค้นพบความรู้ใหม่ที่ซ่อนอยู่ในข้อมูลซึ่งคุณประโยชน์นี้สามารถนำมาประยุกต์ใช้ได้กับบริษัทสถาบันองค์กรงานด้านต่าง ๆ ได้มากมายทำให้เกิดความเข้าใจจากผลสะท้อนข้อมูลในอดีต และนำผลลัพธ์การเรียนรู้ที่ได้มาปรับปรุงให้มีประสิทธิภาพที่ดีในอนาคตและวางแผนการดำเนินงานได้ต่อไป

สำหรับปัจจัยที่ทำให้การทำเหมืองข้อมูลได้รับความนิยม มีดังนี้

การผลิตข้อมูลที่มีขนาดใหญ่และข้อมูลมีการขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันข้อมูลขนาดใหญ่มีการขยายตัวอย่างรวดเร็วโดยผ่านทางอินเตอร์ดาวเทียมและแหล่งผลิตข้อมูลอื่น ๆ เช่น เครื่องอ่านบาร์โค้ดเครดิตการ์ดและอีคอมเมิร์ซเป็นต้น ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบสนับสนุนการตัดสินใจ เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจส่วนมากข้อมูลจะถูกจัดเก็บแยกมาจากระบบปฏิบัติการ โดยจัดอยู่ในรูปแบบของคลังสินค้า (data warehouse) หรือเหมืองข้อมูล (data mining) ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้ระบบคอมพิวเตอร์สมรรถนะสูงมีราคาถูกลง

ปัจจุบันเทคนิคการทำเหมืองข้อมูลที่มีอัลกอริทึมซับซ้อนมีความสามารถในการคำนวณสูงมาใช้งานได้พร้อมกับเริ่มมีเทคโนโลยีที่นำเครื่องคอมพิวเตอร์ขนาดเล็ก (microcomputer) จำนวนมากเชื่อมต่อกันโดยเครือข่ายความเร็วสูง (PC cluster) ทำให้ได้ระบบคอมพิวเตอร์สมรรถะสูงในราคาถูก การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้าเนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสาหกร รมและการค้ามี

การผลิตข้อมูลไว้จำนวนมากแต่ไม่ได้นำข้อมูลมาใช้ให้เกิดประโยชน์จึงมีความจำเป็นอย่างยิ่งที่ต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูลความรู้ที่ได้รับสามารถนำไปวิเคราะห์เพื่อการตัดสินใจในการจัดการในระบบต่าง ๆ

วิวัฒนาการของ Data Mining

ปี 1960 Data Collection คือ การนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือและป้องกันการสูญหายได้เป็นอย่างดี

ปี 1980 Data Access คือ การนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ต่อกันในข้อมูลเพื่อประโยชน์ในการนำไปวิเคราะห์และการตัดสินใจอย่างมีคุณภาพ

ปี 1990 Data Warehouse & Decision Support คือ การรวบรวมข้อมูลมาจัดเก็บลงไปในฐานข้อมูลขนาดใหญ่โดยครอบคลุมทุกแง่ทุกมุมขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ

ปี 2000 Data Mining คือ การนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล โดยการสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

ประเภทข้อมูลที่ใช้ทำเหมืองข้อมูล

1. Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตารางโดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย Entity

2. Relationship Model Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ที่เดียวกัน

3. Transactional Database ประกอบด้วยข้อมูลที่แต่ละทรานแซกชั่นแทนด้วย เหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูปชื่อลูกค้าและรายการสินค้าที่ ลูกค้ารายซื้อ

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่น ๆ เช่น ข้อมูลแบบ Object Oriented ข้อมูลที่เป็น Text File ข้อมูลมัลติมีเดีย ข้อมูลในรูปของ Web

รูปแบบข้อมูลของการทำเหมืองข้อมูล

เนื่องจากข้อมูลที่จะนำมาวิเคราะห์ด้วยวิธีการทำเหมืองข้อมูลมีได้หลายรูปแบบ การวิเคราะห์ข้อมูลจึงผิดแผกไปตามรูปแบบของข้อมูล ข้อมูลที่จะนำมาวิเคราะห์อาจแบ่งได้เป็น 2 รูปแบบ คือ

รูปแบบที่ 1 ข้อมูลแบบมีโครงสร้าง (Structured Data) เช่น ข้อมูลที่อยู่ในรูประเบียน (Record) ตาราง หรือในรูปแบบรายการข้อมูล (Transactional Data) เป็นต้น ข้อมูลแบบมีโครงสร้างโดยทั่วไปจะอยู่ในรูปแบบ ตารางซึ่งประกอบด้วยแถวและคอลัมน์ในการวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูลส่วนใหญ่จะเรียกข้อมูลแต่ละ “แถว” ว่า “ตัวอย่าง (example)” หรือ “อินสแตนซ์ (instance)” และข้อมูล แต่ละ “คอลัมน์” ว่า “แอตทริบิวต์ (attribute)” หรือ “ฟีเจอร์ (feature)”

รูปแบบที่ 2 ข้อมูลแบบไม่มีโครงสร้างแน่นอน (Unstructured Data) เช่น ข้อมูลในรูปข้อความ (Text) ข้อมูลในรูปเว็บซึ่งประกอบด้วยข้อความและลิงก์ที่ชี้ไปยังเว็บเพจอื่น ๆ ข้อมูลในรูปแบบกราฟ เป็นต้น นอกจากนี้ ข้อมูลส่วนใหญ่จะเป็นแบบข้อมูลแบบที่ไม่มีโครงสร้าง เช่น ข้อความหรือรูปภาพ ต่าง ๆ แต่ข้อมูลเหล่านี้ก็มีความสำคัญด้วยเช่นกัน

ส่วนใหญ่ข้อมูลที่นำมาทำเหมืองข้อมูลมักจะอยู่ในรูปแบบที่มีโครงสร้าง เช่น ระเบียนของข้อมูลหรือตารางข้อมูลเป็นต้น ในปัจจุบันการทำเหมืองข้อมูลกับข้อมูลที่ไม่มีโครงสร้างได้มีการดำเนินการมากขึ้น เช่น การทำเหมืองข้อมูลบนข้อมูลข้อความ (Text Mining) และการทำเหมืองข้อมูลกับข้อมูลที่เป็นเว็บเพจ ซึ่งเรียกว่า Web Mining เป็นต้น เนื่องจากข้อมูลส่วนใหญ่ที่นิยมนำมาวิเคราะห์ด้วยการทำเหมืองข้อมูลจะเป็นแบบมีโครงสร้างคือลักษณะข้อมูลที่มีโครงสร้างเป็นหลักข้อมูลแบบมีโครงสร้างมักประกอบด้วย Attributes หรือตัวแปรของข้อมูล ตัวอย่างเช่น ระเบียนข้อมูลของลูกค้าแต่ละราย ประกอบด้วยตัวแปร ได้แก่ หมายเลขบัตรประชาชน อายุ เพศ สถานะสมรส รายได้ต่อปี เป็นต้น โดยตัวแปรของข้อมูลอาจมีหลายชนิด ดังต่อไปนี้

ข้อมูลที่บอกคุณภาพ (Categorical Data) มีลักษณะเป็นข้อมูลที่มีค่าไม่ต่อเนื่อง (Discrete) สามารถแทนค่าด้วยสายอักษร ตัวอย่างเช่น เพศ สี เกรด เป็นต้น ข้อมูลประเภทนี้ยังแบ่งออกได้เป็น

2 ชนิดย่อย คือ ชนิดที่ 1 Nominal Data เป็นข้อมูลที่สามารถนำมาเปรียบเทียบกันว่าเท่ากันหรือไม่เท่ากัน ตัวอย่างเช่น เพศ สี เป็นต้น และชนิดที่ 2 Ordinal Data เป็นข้อมูลที่ค่าสามารถนำมาเปรียบเทียบว่าเท่ากันหรือไม่เท่ากัน ตัวอย่างเช่น เกรด เป็นต้น ข้อมูลที่บ่งบอกปริมาณ (Numerical Data) ซึ่งมีค่าต่อเนื่อง (Continuous) ดังนั้น นอกจากสามารถนำมาเปรียบเทียบได้เช่นเดียวกับ Categorical Data ยังสามารถนำมาคำนวณ เช่น การบวก ลบ คูณ หรือหารได้ ตัวอย่างเช่น น้ำหนัก ส่วนสูง อายุ เป็นต้น

สำหรับ Numerical Data ที่ สามารถนำมาบวกลบกันได้เท่านั้น เรียกว่า Interval Data เช่น วัน เวลา อุณหภูมิ เป็นต้น Numerical Data ที่สามารถนำมา บวก ลบ คูณ หรือหาร (หาค่าสัดส่วนระหว่างกันได้) เรียกว่า Ratio Data เช่น จำนวนนับ อายุ ความ สูง เป็นต้น

<u>ขั้นตอนการทำเหมืองข้อมูล</u>

ขั้นตอนการทำงานของการทำเหมืองข้อมูลจากขั้นตอนการทำงานของการทำเหมืองข้อมูลประกอบไปด้วย 4 ขั้นตอนหลัก ๆ ดังนี้ (Roiger, R., and Geatz, M., 2003)

1. การระบุปัญหาที่เกิดขึ้นกับธุรกิจเป็นการระบุขอบเขตของข้อมูลที่จะนำมาทำการวิเคราะห์เพื่อหาความได้เปรียบทางการตลาดหรือเพื่อนำมาแก้ไขปัญหา

2. ส่วนของการทำเหมืองข้อมูลเป็นการนำเทคนิคของการทำเหมืองข้อมูลไปใช้ถ่ายทอดหรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลจะนำไปใช้ได้จริงในทางธุรกิจ

3. การนำเอาข้อมูลที่เป็นผลลัพธ์ของส่วนการทำเหมืองข้อมูลมาลองปฏิบัติจริงกับธุรกิจ

4. การวัดประสิทธิภาพของเทคนิคการทำเหมืองข้อมูลที่จะนำมาใช้จากผลลัพธ์ เช่น วัดจากส่วนแบ่งของตลาด วัดจากปริมาณลูกค่าหรือวัดจากกำไรสุทธิเป็นต้น จากทั้ง 4 ขั้นตอนที่ DPU 6 กล่าวมาข้างต้นคือการนำเอา Data Mining ไปใช้กับระบบทางธุรกิจโดยแต่ละขั้นตอนจะพึ่งพาอาศัยกันผลลัพธ์จากขั้นตอนหนึ่งจะกลายมาเป็นการนำเข้า (Input) จากอีกขั้นตอนต่อไป ซึ่ง Data Mining (เหมืองข้อมูล) จะเปลี่ยนข้อมูลดิบให้เป็นข้อมูลประยุกต์ ดังนั้นการระบุแหล่งข้อมูลที่ถูกต้องจึงเป็นสิ่งที่สำคัญอย่างยิ่งต่อผลลัพธ์ที่ได้จากการวิเคราะห์

การทำงานของการทำเหมืองข้อมูลในทางปฏิบัติจริงการทำเหมืองข้อมูลจะประสบความสำเร็จกับงานบางกลุ่มเท่านั้น และต้องอยู่ภายใต้ภาวะที่จำกัดปัญหาเหมาะสมกับการใช้เทคนิคการทำเหมืองข้อมูล จะเป็นปัญหาที่ต้องใช้เหตุผลในการแก้เป็นปัญหาที่เกี่ยวข้องกับเศรษฐศาสตร์และการเงินซึ่งจะสามารถจัดรูป แบบของธุรกิจให้อยู่ในรูปแบบของงานทั้ง 6 งานได้ ดังนี้

1. การจำแนกข้อมูล (Classification)

การจำแนกข้อมูล (Classification) การจัดหมวดหมู่ถือว่าเป็นงานธรรมดาทั่วไปของการทำเหมืองข้อมูล เพราะการทำความเข้าใจและการติดต่อสื่อสารต่าง ๆ ก็เกี่ยวข้องกับการแบ่งเป็นหมวดหมู่การจัดแยกประเภท และการแบ่งแยกชนิดโดยการจัดหมวดหมู่ประกอบด้วยการสำรวจจุดเด่นของวัตถุที่ปรากฏออกมาและทำการกำหนดจุดเด่นนั้น ๆ เป็นตัวที่ใช้แบ่งหมวดหมู่งานในการแบ่งหมวดหมู่ คือ การบ่งบอกลักษณะโดยการอธิบายจุดเด่นที่เป็นที่รู้จักดีในหมวดหมู่นั้น และชุดข้อมูลเรียนรู้ (Training Set) ของตัวอย่างในแต่ละหมวดหมู่ ซึ่งมีภาระหน้าที่ในการสร้างโมเดลของบางชนิดที่ไม่สามารถจะจัดหมวดหมู่ของข้อมูลได้ให้สามารถจัดเป็นหมวดหมู่ได้ ตัวอย่างของการจัดหมวดหมู่ เช่น การจัดหมวดหมู่ของผู้ยื่นขอเครดิต (Credits) เป็นระดับต่างระดับกลางและระดับสูงของความเสี่ยงที่จะได้รับเป็นต้น

งานวิจัยเทคนิคการจำแนกประเภทข้อมูลที่ผ่านมาส่วนใหญ่เน้นกระทำกับฐานข้อมูลเชิงสัมพันธ์ ได้แก่ Ross เสนออัลกอริทึม C4.5 สำหรับการจำแนกประเภทข้อมูล (data classification) บนฐานข้อมูลเชิงสัมพันธ์, Mehta เสนออัลกอริทึม SLIQ ซึ่งเป็นอัลกอริทึมจำแนกประเภทข้อมูลที่สามารถจัดการกับข้อมูลขนาดใหญ่ได้โดยใช้เทคนิค pre-sorting เพื่อเพิ่มความเร็วโดยการลดการคำนวณแอทริบิวท์ที่เป็นตัวเลข Shafer เสนออัลกอริทึม SPRINT ซึ่งใช้ Pre-Sorting เช่นเดียวกับ SLIQ แต่ไม่มีข้อจำกัดเรื่องหน่วยความจำ สามารถขยายการแบ่งข้อมูลและนำไปทำการประมวลผลแบบขนานได้ นอกจากนี้มีงานวิจัยบางส่วนออกแบบมาสำหรับสืบค้นความรู้บนฐานข้อมูลเชิงสัมพันธ์ที่มีหลายลำดับชั้น ได้แก่ งานวิจัย Han และ Fu เสนออัลกอริทึมการหากฎความสัมพันธ์หลายลำดับชั้นจากฐานข้อมูลเชิงสัมพันธ์ งานวิจัย Han

เสนอแนวทางการทำดาต้าไมน์นิงบนฐานข้อมูลหลายลำดับชั้น และงานวิจัย Fortin เสนอเทคนิคการหาความสัมพันธ์ข้อมูลแบบลำดับชั้นมาประยุกต์กับฐานข้อมูลเชิงสัมพันธ์โดยใช้หลักการเชิงวัตถุ เป็นต้น

2. การประมาณค่า (Estimation)

การประมาณค่า (Estimation) การประมาณค่า เป็นวิธีการใช้ค่าสถิติที่ได้จากตัวอย่างไปประมาณค่าพารามิเตอร์ เป็นการหาข้อสรุปที่เกี่ยวกับพารามิเตอร์ในลักษณะของการประมาณ ซึ่งมักแสดงในรูปตัวเลข เช่น ประมาณค่าเฉลี่ยของประชากร ประมาณค่าสัดส่วนของประชากร เป็นต้น อาจกล่าวได้ว่ากระบวนการในการประมาณค่า เป็นการนำตัวเลขค่าสถิติที่ได้มาจากกลุ่มตัวอย่างไปประมาณหาค่าความจริงระดับประชากร ในเรื่องเดียวกันนั้น

การประมาณค่ามี 2 แบบ ดังนี้

2.1 การประมาณค่าแบบจุด (Point Estimation) เป็นการประมาณค่าพารามิเตอร์ของประชากรด้วยค่าเพียงค่าเดียว (Single valued Estimation หรือ Point Estimation) ซึ่งการประมาณค่าแบบนี้อาจจะมีค่าเท่ากับค่าพารามิเตอร์หรืออาจมีโอกาสที่จะได้ค่าที่คลาดเคลื่อนไปจากค่าพารามิเตอร์ได้มาก ทั้งนี้ขึ้นอยู่กับหน่วยตัวอย่างที่นำมาวิเคราะห์ (ถ้าหน่วยตัวอย่างนั้นได้มาจากการสุ่มตัวอย่างก็จะสามารถควบคุมความคลาดเคลื่อนได้ระดับหนึ่ง) หมายเหตุ การประมาณค่าพารามิเตอร์ซึ่งเป็นลักษณะของประชากรโดยใช้ข้อมูลตัวอย่าง หรือทำการประมาณค่าพารามิเตอร์ด้วยค่าสถิติ สัญลักษณ์ที่ใช้แทนค่าพารามิเตอร์และค่าสถิติ

2.2 การประมาณค่าแบบช่วง (Interval Estimation) ค่าที่ประมาณได้จากการประมาณค่าแบบช่วง จะได้ช่วงของตัวเลขที่ประมาณ เรียกช่วงการประมาณ เช่น ค่าเฉลี่ยค่าใช้จ่ายในการรักษาพยาบาลของผู้ป่วยโรคหอบที่มารับการรักษาในโรงพยาบาลสุโขทัยอยู่ระหว่าง 3,370 - 6,480 บาท ในการประมาณค่าแบบช่วงนิยมเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์ แทนค่าที่ทำการประมาณโดยครอบคลุมค่าต่ำสุด - สูงสุด เช่น หากเป็นการประมาณค่าเฉลี่ยของประชากร คือ μ จะเขียนเป็น $a < \mu < b$ เรียกค่า a และ b ว่า ค่าต่ำสุด และค่าสูงสุดของช่วงประมาณ μ ในการประมาณค่าแบบช่วง นอกจากขึ้นอยู่กับค่าที่ต้องการ

ประมาณ และการแจกแจงความน่าจะเป็นของค่าที่ต้องการประมาณแล้ว ยังขึ้นกับระดับความเชื่อมั่น (confidence level) อีกด้วย ระดับความเชื่อมั่นนี้จะเป็นค่าที่บอกเราว่า ช่วงประมาณที่สร้างขึ้นจะครอบคลุมค่าพารามิเตอร์ด้วยความน่าจะเป็นมากน้อยเพียงใด

ช่วงความเชื่อมั่น (confidence interval) หมายถึง ช่วงของค่าประมาณที่ประกอบไปด้วยค่าต่ำสุด (a) และค่าสูงสุด (b) ที่คำนวณขึ้นมา ช่วงดังกล่าวจะคลุมค่าของพารามิเตอร์ ด้วยความน่าจะเป็นตามที่กำหนด ตัวอย่างเช่น ช่วงความเชื่อมั่น 90% ของค่าใช้จ่ายโดยเฉลี่ยของผู้ป่วยโรคหอบที่มารับการรักษาในโรงพยาบาลสุโขทัย อยู่ระหว่าง 3,370 - 6,480 บาท หมายถึงว่า “มีความมั่นใจ 90% ที่ช่วงของการประมาณค่าใช้จ่ายโดยเฉลี่ยที่ได้ (3,370 - 6,480 บาท) จะครอบคลุมค่าใช้จ่ายที่เป็นค่าเฉลี่ยจริงของผู้ป่วย” ที่ระดับความเชื่อมั่นนี้จะเขียนแทนด้วยสัญลักษณ์ $(1-\alpha)100\%$ โดยที่ $0 < \alpha < 1$ และเรียกค่า $1-\alpha$ ว่า สัมประสิทธิ์ของความเชื่อมั่น (confidence coefficient)

3. การทำนายล่วงหน้า (Prediction)

การทำนายล่วงหน้า (Prediction) การทำนายล่วงหน้าก็เป็นงานที่มีลักษณะคล้ายกับการจัดหมวดหมู่หรือการประเมินค่ายกเว้นเพียงแต่จะใช้สถิติการบันทึกของการจัดหมวดหมู่ในการทำนายอนาคตของพฤติกรรมหรือการประเมินค่าที่จะเกิดขึ้นในอนาคต ตัวอย่างของงานการทำนายล่วงหน้า เช่น การทำนายการเปลี่ยนแปลงพฤติกรรมของตลาด หรือการทำนายจำนวนลูกค้าที่จะออกจากธุรกิจของเราใน 6 เดือนข้างหน้าเป็นต้น

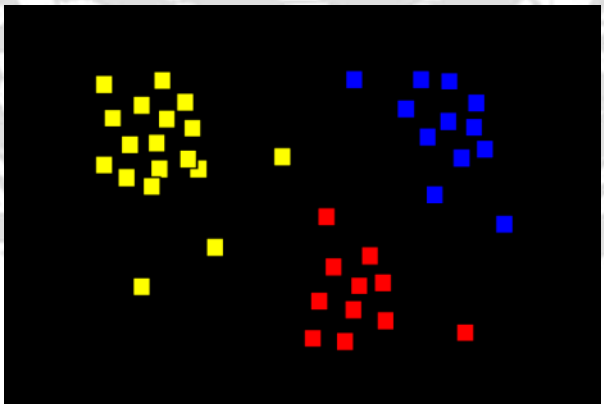
4. การจัดกลุ่มโดยอาศัยความใกล้ชิด (Affinity Group)

การจัดกลุ่มโดยอาศัยความใกล้ชิด (Affinity Group) หรือการวิเคราะห์ของตลาดงานในการจัดกลุ่ม หรือการวิเคราะห์ตลาด คือ การตัดสินใจรวมสิ่งที่สามารถไปด้วยกันเข้าไว้ในกลุ่มเดียวกัน ตัวอย่างของการจัดกลุ่มโดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ตลาด เช่น การตัดสินใจว่าลูกค้าคนใดจัดอยู่ในกลุ่มค้าประเภทใด

5. การรวมตัว (Clustering)

การรวมตัว (Clustering)[1] การรวมตัวคืองานที่ทำการรวมส่วนต่างๆในแต่ละส่วนที่ต่างชนิดกันให้อยู่ในรวมกันเป็นกลุ่มย่อย หรือคลัสเตอร์ (Clusters) โดยในแต่ละกลุ่มย่อย อาจจะประกอบด้วยส่วนต่าง ๆ ที่ต่างชนิดกัน ซึ่งความแตกต่างของการรวมตัวจากการจัดหมวดหมู่ คือ การรวมตัวจะไม่พึ่งพาอาศัยการกำหนดหมวดหมู่ล่วงหน้า และไม่ใช้ตัวอย่างข้อมูลจะรวมตัวกันบนพื้นฐานของความคล้ายในตัวเอง

การวิเคราะห์คลัสเตอร์ในตัวเองไม่ใช่อัลกอริทึมแต่เป็นการทำงานร่วมกันของอัลกอริทึมที่หลากหลายเพื่อแก้ปัญหาในการทำงาน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (similarity) หรือ ความใกล้ชิด (proximity) โดยจะแบ่งชุดข้อมูล (มักจะเป็นเวกเตอร์) ออกเป็นกลุ่ม (cluster) นำข้อมูลที่มีคุณลักษณะเหมือนกัน หรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน การคำนวณจากการวัดระยะระหว่างเวกเตอร์ของข้อมูลเข้า โดยใช้การวัดระยะแบบต่าง ๆ เช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบแมนฮัตตัน (Manhattan distance) การวัดระยะแบบเชบิเชฟ (Chebychev distance)



ภาพที่ 2.1 แสดงภาพผลการแบ่งกลุ่มจำนวน 3 กลุ่ม

[1] https://bigdata.go.th/big-data-101/4-types-of-clustering/

การวิเคราะห์คลัสเตอร์เริ่มมีการกล่าวถึงครั้งแรกในปี พ.ศ. 2475 โดย ไดรฟ์เวอร์ และโครเบอร์ และมีการนำมาใช้งานในด้านจิตวิทยาในปี พ.ศ. 2481

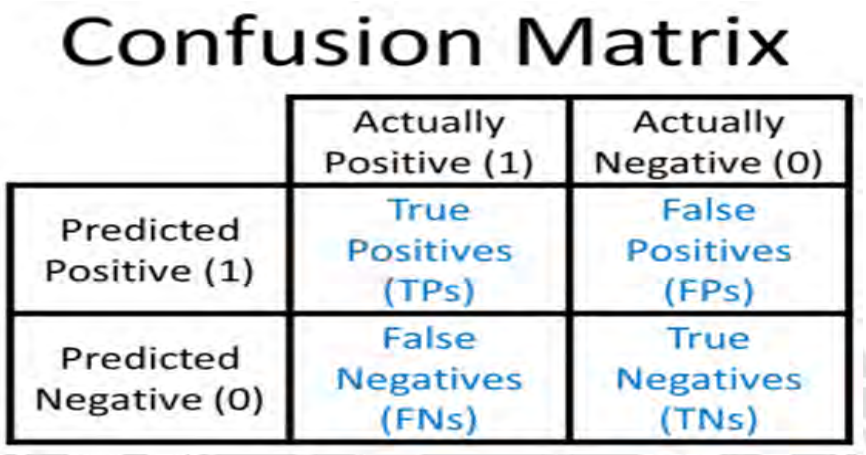
การแบ่งกลุ่มข้อมูลจะแตกต่างจากการแบ่งประเภทข้อมูล (classification) โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดประเภทของข้อมูลไว้ก่อน จึงกล่าวได้ว่าการแบ่งกลุ่มข้อมูลเป็นการเรียนรู้แบบไม่มีผู้สอน ขั้นตอนวิธีการแบ่งกลุ่ม ได้แก่ k-means clustering, hierarchical clustering, self-organizing map (som)

การแบ่งกลุ่มข้อมูลอาจใช้เป็นข้อตอนเบื้องต้นของการวิเคราะห์ข้อมูล เพื่อช่วยในการลดขนาดข้อมูล (แยกเป็นหลาย ๆ กลุ่มและคัดเฉพาะบางกลุ่มเพื่อทำการวิเคราะห์ต่อไป หรือแยกการวิเคราะห์ออกเป็นสำหรับแต่ละกลุ่ม) ก่อนที่จะนำไปวิเคราะห์ด้วยวิธีการอื่นต่อไป ขั้นตอนวิธีในการแบ่งกลุ่มข้อมูลโดยทั่วไปแบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ การแบ่งแบบเป็นลำดับขั้น (hierarchical) และการแบ่งแบบตัดเป็นส่วน (partitional) การแบ่งแบบเป็นลำดับขั้นนั้น จะมีทำการแบ่งกลุ่มจากกลุ่มย่อยที่ถูกแบ่งไว้ก่อนหน้านั้นซ้ำหลายครั้ง ส่วนการแบ่งแบบตัดเป็นส่วนนั้น การแบ่งจะทำเพียงครั้งเดียว การแบ่งแบบเป็นลำดับขั้น จะมี 2 ลักษณะคือ แบบล่างขึ้นบน (bottom-up) หรือเป็นการแบ่งแบบรวมกลุ่มจากกลุ่มย่อยให้ใหญ่ขึ้นไปเรื่อย ๆ โดยเริ่มจากกลุ่มเล็กสุดคือในแต่ละกลุ่มมีข้อมูลเพียงตัวเดียว และแบบบนลงล่าง (top-down) หรือเป็นการแบ่งแบบกลุ่มจากกลุ่มใหญ่ให้ย่อยไป เรื่อยๆ โดยเริ่มจากกลุ่มใหญ่ที่สุด คือ กลุ่มเดียวมีข้อมูลทุกตัวอยู่ในกลุ่ม

6. การบรรยาย (Description)

การบรรยาย (Description) ในบางครั้งวัตถุประสงค์ของการทำเหมืองข้อมูล คือ ต้องการอธิบายความสับสนของฐานข้อมูลในทางที่จะเพิ่มความเข้าใจในส่วนของประชากร ผลิตภัณฑ์ หรือขบวนการให้มากขึ้น เทคนิคการทำเหมืองข้อมูล ส่วนใหญ่ต้องการเทรนนิงข้อมูลจำนวนมากที่ประกอบด้วยหลายๆ ตัวอย่างเพื่อจะสร้างกฎที่ใช้ในการจัดหมวดหมู่ กฎของความสัมพันธ์ กลุ่มย่อยการทำนายล่วงหน้า ดังนั้นชุดของข้อมูลขนาดเล็กจะนำไปสู่ความไม่น่าไว้วางใจของผลสรุปที่ได้ไม่มีเทคนิคใดเลยที่จะสามารถแก้ปัญหา

ของการทำเหมืองข้อมูลได้ทุกปัญหา ดังนั้นความหลากหลายของ เทคนิคจึงเป็นสิ่งที่จำเป็นในการไปสู่วิธีการแก้ปัญหาของ Data Miningได้ดีที่สุด



ภาพที่ 2.2 แสดงตาราง Confusion Matrix
ที่มา : ประกรณ์ เกตุชาลี (2562)

ไม่มีเทคนิคหรือเครื่องมือเพียงชนิดเดียวของการทำเหมืองข้อมูลที่เหมาะสมกับงานทุกชนิด งานในแต่ละชนิดก็จะมีเทคนิคของการทำเหมืองข้อมูลที่แตกต่างกันไปขึ้นอยู่กับชนิดของงาน

<u>ส่วนประกอบของระบบการทำเหมืองข้อมูล</u>

1. Database, Data Warehouse, World Wide Web และ Other Info Repositories เป็นแหล่งข้อมูลสำหรับการทำเหมืองข้อมูล

2. Database หรือ Data Warehouse Server ทำหน้าที่นำเข้าข้อมูลตามคำขอของผู้ใช้

3. Knowledge Base ได้แก่ ความรู้เฉพาะด้านในงานที่ทำจะเป็นประโยชน์ต่อการสืบค้นหรือประเมินความน่าสนใจของรูปแบบผลลัพธ์ที่ได้

4. Data Mining Engine เป็นส่วนประกอบหลักประกอบด้วยโมดูลที่รับผิดชอบงานทำเหมืองข้อมูลประเภทต่างๆ ได้แก่ การหากฎความสัมพันธ์ การจำแนกประเภท การจัดกลุ่ม

5. Pattern Evaluation Module ทำงานร่วมกับ Data Mining Engine โดยใช้มาตรวัดความน่าสนใจในการกลั่นกรองรูปแบบผลลัพธ์ที่ได้ เพื่อให้การค้นหามุ่งเน้นเฉพาะรูปแบบที่น่าสนใจ

6. Graphic User Interface ส่วนติดต่อประสานระหว่างผู้ใช้กับระบบการทำเหมืองข้อมูล ช่วยให้ผู้ใช้สามารถระบุงานทำเหมืองข้อมูลที่ต้องการทำ ดูข้อมูลหรือโครงสร้างการจัดเก็บข้อมูล ประเมินผลลัพธ์ที่ได้

<u>การประยุกต์ใช้ Data Mining</u>

การประยุกต์ใช้ Data Mining จะมีได้หลากหลายแต่สามารถจัดกลุ่มกว้างๆ ได้เป็นสองกลุ่ม คือ กลุ่มที่ใช้เพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย (สุรพงศ์ เอื้อวัฒนามงคล, 2557)

การทำ Data Mining เพื่อการทำนาย เป็นการนำความรู้ที่เรียนรู้มาจากข้อมูลที่มีอยู่เพื่อประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต เช่น จากข้อมูลลูกค้าของแผนกสินเชื่อของธนาคารที่ได้มีการจัดลำดับชั้นของลูกค้าไว้แล้วว่าใครเป็นลูกค้าชั้นดี ใครเป็นลูกค้าในระดับปานกลาง และใครเป็นลูกค้าที่มักจะผิดนัดชำระหนี้ Data Mining สามารถเรียนรู้จากข้อมูลเหล่านี้และค้นหาโมเดลที่สามารถใช้อธิบายลักษณะของลูกค้าชั้นดี ลูกค้าระดับปานกลาง และลูกค้าที่ไม่เป็นที่ต้องการ จากโมเดลที่ได้นี้สามารถนำไปใช้ทำนายลูกค้าใหม่ที่มาขอสินเชื่อได้ว่าน่าจะเป็นลูกค้าประเภทใด

การทำ Data Mining เพื่อการอธิบาย เป็นการค้นหารูปแบบที่น่าสนใจจากกลุ่มข้อมูล รูปแบบนี้มักจะเป็นความสัมพันธ์หรือลักษณะที่เชื่อมโยงกันของข้อมูล การทำแบบนี้ต่างจากแบบแรกตรงที่ผู้ใช้ไม่ได้กำหนดล่วงหน้าว่าจะให้โปรแกรม Data Mining ค้นหารูปแบบหรือโมเดลของอะไร แต่ให้ค้นหาทุกรูปแบบที่น่าสนใจจากข้อมูล

การวัดประสิทธิภาพโมเดล (Confusion Matrix) ถือเป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย หรือ Prediction ที่ทำนายจาก Model ที่เราสร้างขึ้นใน Machine learning

โดยมีไอเดียจากการวัดว่า สิ่งที่เราคิด (Model ทำนาย) กับสิ่งที่เกิดขึ้นจริง มีสัดส่วนเป็นดังนี้ (ประกรณ์ เกตุชาลี, 2562)

True Positive (TP) = สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริงในกรณีทำนายว่าจริง และสิ่งที่เกิดขึ้น คือ จริง

True Negative (TN) = สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นในกรณีทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น คือ ไม่จริง

False Positive (FP) = สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือ ทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง

False Negative (FN) = สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือ ทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

โดย TP, TN, FP, FN ในตารางจะแทนด้วยค่าความถี่ สามารถใช้ Confusion Matrix มาคำนวณการประเมินประสิทธิภาพของการทำนายด้วย Model ของเราในรูปแบบค่าต่าง ๆ ได้หลายค่า ได้แก่ Accuracy (ความถูกต้องที่เราทายได้ตรงกับสิ่งที่เกิดขึ้นจริง)

Accuracy (ความถูกต้อง) = (TPs + TNs) / (TPs + TNs + FPs + FNs) หรือกล่าวได้ว่า Accuracy = ผลรวมของตัวเลขบนเส้นทแยงมุมในตาราง Confusion Matrix / จำนวน observations ทั้งหมด โดยความเป็นจริงแล้ว Confusion matrix ไม่จำเป็นต้องเป็นแบบ 2x2 หรือมีผลลัพธ์แค่ 2 แบบเสมอไป โดยอาจเป็น 3x3, 4x4, nxn ก็ได้ โดยวิธีการหา Accuracy ก็ใช้แบบเดิม คือ ผลรวมของตัวเลขบนเส้นทแยงมุมในตารางConfusion Matrix / จำนวน observations ทั้งหมด

Precision (ค่าความแม่นยำ) เป็นการเปรียบเทียบ การทำนายที่ถูกต้องว่า จริง และก็เกิดขึ้นจริง (TP) กับการทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง (FP)

$$\text{Precision} = TPs / (TPs + FPs)$$

Recall (ความถูกต้องของการทำนาย) เป็นการหาความถูกต้องของการทำนายว่าจะเป็น “จริง” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งทำนาย และเกิดขึ้น ว่า “เป็นจริง

$$\text{Recall} = TPs/(TPs+FNs)$$

F1 score เป็นค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall จุดประสงค์ของการสร้าง F1 ขึ้นมา คือ เพื่อเป็น single metric ที่วัดความสามารถของโมเดล

F1 = 2 x (Precision x Recall)/(Precision + Recall)

$$F1 = 2 * \left(\frac{precision * recall}{precision + recall}\right)$$

ภาพที่ 2.3 แสดงสูตรการหาค่า Recall

ตัวอย่าง แทนค่าในสมการ F1 = 2 * ((0.625 * 0.526) / (0.625 + 0.526)) = 57.1%

ตัวอย่างเสริมการอธิบาย เพื่อให้เข้าใจมากขึ้น

		Actual	
		Spam	Ham
Prediction	Spam	20	12
	Ham	18	50

ภาพที่ 2.4 แสดงตัวอย่างตารางสรุปผลอีเมล

ในกรณีที่ต้องการทราบว่า โมเดลทำนายแม่นขนาดไหน คือ ทายถูกต้องว่าเป็น Spam จากการพยายามทำนายทั้งหมด ต้องใช้ Precision ก็คือ

Precision ของทำนายSpam = 20 / 32 = 0.625

แต่ถ้าต้องการทราบว่า โมเดลที่เราสร้างขึ้น สามารถตรวจจับ Spamได้ถูกต้องขนาดไหนจากSpam Email ทั้งหมด ต้องใช้ Recall

Recall ของการตรวจจับSpam = 20 /38 = 0.526

แต่ในกรณีที่ต้องการหาประสิทธิภาพของโมเดลการทำนายนี้ ที่ต้องมีทั้งการทายถูกต้องว่าสิ่งที่เจออันนั้นคือ Spam จริง ๆ และในขณะเดียวกันก็ต้องตรวจจับ Spamได้ด้วย ต้องเลือกใช้ F1 score ก็คือ เป็นค่าเฉลี่ยของทั้ง Precision และ Recall

F1 ของ Model นี้ = 2 * ((0.625 * 0.526) / (0.625 + 0.526)) = 0.571

<u>เทคนิคการทำเหมืองข้อมูล</u>

การทำเหมืองข้อมูลมีวิธีการต่าง ๆ ที่ใช้ในการแก้ปัญหาอยู่หลายวิธีการแต่ไม่มีวิธีการใดวิธีการหนึ่งที่สามารถแก้ปัญหาของการทำเหมืองข้อมูลได้ทั้งหมดแต่ความหลากหลายของวิธีการก็เป็นสิ่งที่จะนำไปสู่การแก้ปัญหาที่ดีที่สุดของการทำเหมืองข้อมูล เพื่อให้ได้ความรู้ตามที่ต้องการจากการค้นหาความรู้ในฐานข้อมูลต่าง ๆ ดังนั้น การเลือกเทคนิควิธีการทำเหมืองข้อมูลจึงเป็นขั้นตอนที่มีความสำคัญ มักแบ่งลักษณะการทำงานของการทำเหมืองข้อมูลตามลักษณะเป้าหมายของการแก้ปัญหาเพื่อตอบโจทย์คำถามของการวิเคราะห์ข้อมูล โดยทั่วไปแล้วปัญหาของการทำเหมืองข้อมูลสามารถมองได้ 2 ลักษณะใหญ่ ๆ ตามเป้าหมายของการทำงาน เป้าหมายแรกคือ เพื่ออธิบายลักษณะและพฤติกรรมทั่วไปของข้อมูลและเป้าหมายที่สอง คือ เพื่อนำข้อมูลในฐานข้อมูลที่เรามีมาใช้คาดคะเนข้อมูลอื่น ๆ โดยทั่วไปแล้วนั้นมักแยกส่วนของการทำเหมืองข้อมูลได้ดังนี้คือ (วันวิสาข์ ชนะประเสริฐ, 2559)

1. การอธิบายลักษณะประเภท (Concept/Class Description) กล่าวคือ ข้อมูลโดยทั่วไป เราจะสามารถแบ่งได้โดยอาจระบุการแบ่งเป็นคลาส (Class) คือ ประเภท ซึ่งสามารถมองเป็นรูปธรรมได้ชัดเจน เช่น “ประเภทคอมพิวเตอร์” “ประเภทเครื่องพิมพ์” เป็นต้น หรือแบ่งแบบคอนเซปต์ (Concept) คือ แบ่งตามแง่คิดหรือมุมมอง เช่น “ประเภทฟุ่มเฟือย” “ประเภทประหยัด” เป็นต้น หลักการนี้จะมีเป้าหมายเพื่อพยายามสรุปหรือพยายามจัดกลุ่มข้อมูลว่าสามารถทำได้ในลักษณะใดบ้าง หลักการนี้สามารถนำไปใช้ประโยชน์ในกระบวนการเตรียมข้อมูลหรือการนำเสนอตามมุมมองได้

2. การวิเคราะห์ความสัมพันธ์ (Association Analysis) คือ กระบวนการที่พยายามจะค้นหาลักษณะความสัมพันธ์ของข้อมูลเพื่อระบุออกมาเป็นกฎ โดยจะดูความถี่ของกลุ่มที่เกิดขึ้น ร่วมกันใน

ขอบเขตของกฎดังกล่าวเป็นหลัก มักนำไปใช้กับการวิเคราะห์ข้อมูลที่มีลักษณะเป็น Transaction คือ มีการเกิดขึ้นเป็นประจำ

3. การจำแนกประเภทและการทำนายข้อมูล (Classification and Prediction) ในความเป็นจริงแล้ว ทั้งการจำแนกประเภทและการทำนายข้อมูลนั้นถูกจัดอยู่ในส่วนเดียวกัน เนื่องจากทั้งสองกระบวนการสามารถมองเป็นการวิเคราะห์เพื่อทำนายข้อมูลได้ทั้งคู่เพียงแต่การแยกกันนั้น พิจารณาจากหลักกระบวนการทำงาน ซึ่งมีบริบทที่ต่างกันคือ การจำแนกข้อมูลนั้นมีวัตถุประสงค์เพื่อทำนายเป็นชื่อกลุ่ม คือ ค้นหาที่มาที่ไปของการที่ข้อมูลใด ๆ ถูกจัดอยู่ในกลุ่มที่มีการระบุไว้ก่อนแล้ว ซึ่งเราเรียกมันว่า คลาสจากข้อมูลเก่า เพื่อใช้ในการระบุข้อมูลใหม่ซึ่งไม่เคยมีการระบุคลาสมาก่อนเพื่อระบุหรือทำนายคลาสให้กับมัน ส่วนการทำนายข้อมูลจะทำนายค่าตัวเลข คือ การนำค่าข้อมูลตัวเลขที่เกี่ยวข้องมาคำนวณเพื่อระบุเป็นสมการ เพื่อใช้ในการหาค่าตัวเลขที่ต้องการจะทราบ โดยสามารถนำไปประยุกต์กับการหาข้อมูลตัวเลขบางจุดที่ไม่มีการระบุไว้ได้ทั้งสองลักษณะนี้เรียกรวมกันว่า เป็นการพยากรณ์ข้อมูล เช่น ต้นไม้ตัดสินใจ (Decision Trees) แบบจำลองโครงข่ายประสาทเทียม (Neural Networks) การวิเคราะห์การถดถอย (Regression)

4. การวิเคราะห์การจัดกลุ่มข้อมูล (Cluster Analysis) คือ กลุ่มข้อมูลที่ถูกแบ่งกันไว้ตามลักษณะของมัน หลักการการจัดกลุ่มข้อมูลต่างจากหลักการจำแนกประเภท เพราะหลักการจำแนกประเภท สนใจวิเคราะห์จำแนกตามประเภทของข้อมูลหรือคลาสที่มีการระบุไว้แล้วในข้อมูลที่นำมาวิเคราะห์เป็นหลัก แต่หลักการการจัดกลุ่มข้อมูลสนใจวิเคราะห์ในระดับข้อมูลเป็นหลักโดยไม่สน ในการอ้างอิงของคลาสเข้ามาช่วยจำแนก หากข้อมูลที่นำมาใช้ทำการวิเคราะห์ไม่เคยมีการระบุคลาสมาก่อน กระบวนการวิเคราะห์การจัดกลุ่มข้อมูลสามารถช่วยจัดกลุ่มข้อมูลเพื่อช่วยในการสร้างคลาสเพื่ออ้างอิง ได้ตัวอย่างของการวิเคราะห์เพื่อจัดกลุ่ม ได้แก่ การหาค่าเฉลี่ย (K-mean Algorithm) การรวมและการแบ่งกลุ่มโดยการจัดลำดับชั้น (Agglomerative and Division Hierarchical Clustering) และการลำดับตำแหน่งเพื่อแสดงโครงสร้างการจัดกลุ่ม (Ordering Points to Identify the Clustering Structure) เป็นต้น

5. การวิเคราะห์ข้อมูลแปลกแยก (Outlier Analysis) การที่ข้อมูลในฐานข้อมูลที่มีปริมาณมากมาย อาจมีบางข้อมูลที่มีพฤติกรรมแปลกแยกไปจากข้อมูลส่วนใหญ่ กระบวนการนี้จะทำการตรวจสอบหาข้อมูลดังกล่าว โดยใช้พื้นฐานทางสถิติเป็นหลักในการตรวจสอบ เพื่อค้นหาลักษณะการกระจายและระยะห่างระหว่างข้อมูลเพื่อพิจารณาความเป็นไปได้ที่เกิดขึ้นของข้อมูลว่าเป็นข้อมูลที่จัดอยู่ในกลุ่มการจัดข้อมูลหรือนอกกลุ่มการจัดข้อมูล

6. การวิเคราะห์วิวัฒนาการ (Evolution Analysis) กระบวนการนี้เน้นหลักไปที่การวิเคราะห์ เพื่อหารูปแบบในการอธิบายพฤติกรรมของข้อมูลที่มีการเปลี่ยนแปลงไปบนพื้นฐานของช่วงเวลาที่เปลี่ยนแปลงไป (Time-Series Data Analysis) โดยอาจจำเป็นต้องใช้หลักการอื่นๆ ของการทำเหมืองข้อมูลที่เกี่ยวข้องเข้ามาช่วยในการวิเคราะห์ร่วมด้วย

<u>เทคนิคต้นไม้ตัดสินใจ (Decision Tree)</u>

ในช่วงปลายของยุค 1970 ได้มีนักวิจัยทางด้านการเรียนรู้ของเครื่อง (machine learning) คือ J. Ross Quinlan ได้คิดค้นอัลกอริทึมสำหรับสร้างต้นไม้ตัดสินใจที่มีชื่อว่า ID3 (Iterative Dichotomiser) ต่อมาเขาได้พัฒนาต่อยอด ID3 ไปเป็น C4.5 ซึ่งได้กลายมาเป็นอัลกอริทึมพื้นฐานที่ใช้สำหรับเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมต่าง ๆ ทางด้านการเรียนรู้แบบมีผู้สอน (Supervised Learning)

ID3 และ C4.5 ได้ทำการประยุกต์ใช้วิธีการเชิงละโมบ (greedy approach) ในการสร้างต้นไม้ภายใต้ วิธีการแบบ “top-down recursive divide-and-conquer” โดยทำการพิจารณาชุดข้อมูลสำหรับเรียนรู้ (training data, เซตของเรคคอร์ดของข้อมูลที่แต่ละเรคคอร์ดจะประกอบไปด้วยเซตของแอตทริบิวต่าง ๆ และแอตทริบิวที่บ่งบอกถึงหมวดหมู่ของข้อมูลเรคคอร์ดนั้น ๆ) ด้วยการแบ่งข้อมูลออกเป็นส่วนย่อย ๆ ในระหว่าง กระบวนการสร้างต้นไม้

ต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองเพื่อการทำนายเป็นแบบจำลองที่มีลักษณะคล้ายกับแบบจำลองที่มีลำดับขั้นของการตัดสินใจวิธีการที่ได้รับความนิยมเนื่องจากมีความซับซ้อนน้อยเมื่อเปรียบเทียบกับอื่น ๆ ซึ่งต้นไม้การตัดสินใจเป็นการนำข้อมูลทดสอบ (Training Data) มาสร้างแบบจำลอง

เพื่อพยากรณ์มีการทำงานแบบการเรียนรู้แบบมีผู้สอนคือสามารถสร้างแบบจำลองได้จากกลุ่มตัวอย่างของข้อมูลได้อัตโนมัติและสามารถสร้างพยากรณ์กลุ่มตัวอย่างของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่หรือข้อมูลทดสอบ (Testing Data) การแสดงรูปแบบของต้นไม้ตัดสินใจประกอบไปด้วยโหนด (Node) แรกสุดเรียกว่าโหนดรากและแตกออกเป็นโหลดย่อยจนหลอดสุดท้ายเรียกว่า โหนดปลาย การวิจัยครั้งนี้ผู้วิจัยเลือกใช้วิธีการสร้างต้นไม้ตัดสินใจด้วยขั้นตอนวิธีแบบ 4C.5 โดยกำหนดให้

$$Entropy(s) = -\sum_{i=1}^{n} p_i \log p_i$$

โดยที่ p_i จำนวนความถี่ของคลาส (Class) i ในโหนด (Node) s เพื่อใช้ h สำหรับคำนวณค่าความน่าจะเป็น ซึ่งจะเป็นหนึ่งคลาสโดยใช้ค่า Entropy จะมีค่าเป็นนั้นหมายถึงทุกคลาสค่าความหน้าจะเป็นที่เท่ากันซึ่งมีโอกาสเกิดขึ้น 1 และมีค่าเป็น 0ได้โดยนิยาม

$$p = (k_i \mid N)$$

ที่ N เท่ากับค่าทั้งหมดของรูปแบบกลุ่มของคลาส โดย K_i เท่ากับเหตุการณ์ที่เกิดขึ้นใน N การคำนวณหาค่า Information Gain ในการแบ่งกลุ่ม p ในกลุ่ม k ด้วยการวัดผล โดยนำค่า Gain ของ p ที่มีค่าน้อยที่สุดในการรวมกันจากกลุ่มย่อย k แล้วนำไปลบออกจากค่าของ Entropy(p) โดยค่าคุณสมบัติ (Attributes) จะใช้สำหรับการเลือกโหนด (Node) ในการแบ่งกลุ่ม โดยเลือกค่า Gain ที่มีค่ามากที่สุดของ k โดยมีนิยามดังนี้

$$GAIN_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i))$$

จากนั้นจึงทำการแจกแจงข้อมูลในแต่ละกลุ่ม p ตามโหนด ที่ได้ทำการแบ่งไว้แล้วข้างต้น ตัวจำแนกข้อมูล C4.5 ได้ทำการขยายส่วนของการจำแนกข้อมูลที่เป็นตัวเลข ด้วยการแบ่งข้อมูลให้เป็นช่วง เพื่อใช้ในการสร้างต้นไม้ตัดสินใจโดยการแบ่งค่าต่อเนื่องออกเป็นช่วง (Discretization)

(1) create a node *N*;
(2) if *samples* are all of the same class *C* then
(3) return *N* as a leaf node labeled with the class *C*;
(4) if *attribute-list* is empty then
(5)return *N* as a leaf node labeled with the most common class in *samples*; // majority voting
(6) select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
(7) label node *N* with *test-attribute*;
(8) for each known value a_i of *test-attribute*
(9) grow a branch from node *N* for the condition *test-attribute*= a_i;
(10) let s_i be the set of samples in *samples* for which *test-attribute*= a_i;
(11) if s_i is empty then
(12) attach a leaf labeled with the most common class in *samples*;
(13) else attach the node returned by Generate decision tree(s_i, *attribute-list, test-attribute*);

ภาพที่ 2.5 แสดงอัลกอริทึมพื้นฐานในการหาต้นไม้ตัดสินใจ

จากภาพที่ 2.5 จะเป็นการอธิบายรายละเอียดต่าง ๆ ของขั้นตอนการสร้างต้นไม้ตัดสินใจที่ซึ่งจะสามารถอธิบายได้ดังนี้

ขั้นตอนที่ (1) อัลกอริทึมการสร้างต้นไม้ตัดสินใจ จะต้องการอินพุต 3 อินพุตด้วยกันคือ

(1.1) D หมายถึง ชุดข้อมูลที่จะทำการพิจารณา โดยในตอนเริ่มต้นของการทำงาน ชุดข้อมูล D จะประกอบไปด้วยข้อมูลทุกเรค คอร์ด

(1.2) ลิสต์ของแอตทริบิว ที่ซึ่งหมายถึง เซตของแอตทริบิวที่ใช้ในการอธิบายคุณลักษณะของข้อมูล แต่ละเรคคอร์ดในชุดข้อมูล

(1.3) วิธีการในการเลือกแอตทริบิวที่ซึ่งเป็นวิธีการในการเลือกแอทริบิว ที่ดีที่สุดที่สามารถแยกความแตกต่างระหว่างเรคคอร์ดต่าง ๆ ในชุดข้อมูลตามหมวดหมู่ของข้อมูล

วิธีการเลือกแอตทริบิวนี้จะใช้ตัวชี้วัดการเลือกแอตทริบิว เช่น ค่าเกนความรู้(information gain) หรือค่าดัชนีจีนี่ (gini index) เป็นต้น โดยในการเลือกตัวชี้วัดการเลือกแอทริบิว เราอาจต้องพิจารณาข้อบังคับและคุณลักษณะของต้นไม้ด้วย เช่น ค่าดัชนีจีนี่จะเป็นตัวชี้วัดที่จะสามารถทำงานได้กับต้นไม้แบบไบนารี แต่สำหรับค่าเกนความรู้จะสามารถทำงานได้กับต้นไม้ที่มีหลายกิ่งและยอมให้ทำการแตกกิ่งออกเป็นหลายทิศทาง (multi-way split)

ขั้นตอนที่ (2) การสร้างต้นไม้จะเริ่มจากโหนด N เพียงแต่หนึ่งโหนดเท่านั้นที่ซึ่งแสดงถึงข้อมูลทั้งหมดในชุดข้อมูล D (ขั้นตอนที่ (1))

ขั้นตอนที่ (3) ในกรณีที่ข้อมูลทุกเรคคอร์ดในชุดข้อมูล D มีหมวดหมู่ของข้อมูลเดียวกัน คือ C— โหนด N จะกลายเป็นโหนดใบ และจะมีหมวดหมู่ C แนบอยู่ (ขั้นตอนที่ (2) และ (3)) - หมายเหตุ ขั้นตอนที่ (4) และ (5) คือ เงื่อนไขการจบการทำงาน ในขั้นตอนที่ (6) จะพบการเรียกใช้วิธีการเลือก แอตทริบิวที่จะตัดสินใจเลือกเกณฑ์ในการแบ่งข้อมูล โดยผลลัพธ์ที่ได้จะเป็นแอตทริบิวหนึ่งๆ ที่ดีที่สุดที่สามารถแบ่งข้อมูลเรคคอร์ดต่าง ๆ พร้อมกับหมวดหมู่ของข้อมูลออกจากชุดข้อมูล การเลือกแอทริบิวยังสามารถบอกได้ว่ากิ่งใดจะเติบโตจากโหนด N บ้าง ในการแบ่งข้อมูลออกเป็นส่วนๆ ถ้าส่วนใดก็ตามที่ถูกแบ่งมีเรคคอร์ดทั้งหมดที่มีหมวดหมู่เดียวกับส่วนของข้อมูลนั้น ๆ จะถูกเรียกว่า “pure partition” แต่อย่างไรก็ตามในการแบ่งข้อมูลอาจมีการแบ่งข้อมูลที่ไม่ได้มีหมวดหมู่เดียวกันอยู่ในส่วนเดียวกันก็เป็นได้—โดยในการแบ่งข้อมูลเราจะต้องพยายามให้ส่วนที่ถูกแบ่งนั้นมีเรคคอร์ดของข้อมูลที่มีหมวดหมู่เหมือนกันมากที่สุดเท่าที่จะมากได้

ขั้นตอนที่ (4) โหนด N จะถูกตั้งชื่อด้วยชื่อแอตทริบิวที่ได้จากขั้นตอนที่ (6) และแต่ละกิ่งที่เติบโตจากโหนด N จะหมายถึงการเติบโตจากค่าที่เป็นไปได้ในแอทริบิวนั้นๆ โดยในการแบ่งข้อมูลและการเพิ่มกิ่งก้านให้แก่โหนด N (ในขั้นตอนที่ (10) และ (11)) จะเกิดขึ้นภายใต้เงื่อนไข 3 เงื่อนไข โดยกำหนดให้ A

เป็นแอตทริบิวที่ได้จากการเลือกหรือกล่าวอีกนัยหนึ่งว่า A เป็น splitting_attribute โดยที่ A มีค่าที่เป็นไปได้ทั้งหมดที่เกิดขึ้นในชุดข้อมูล เป็น $\{a_1, a_2, \dots, a_v\}$

กรณีที่ A มีค่าแบบไม่ต่อเนื่อง จะทำการแตกกิ่งจากโหนด N ไปตามค่าที่ปรากฏใน แอตทริบิว A โดยกิ่งหนึ่งจะถูกแทนด้วยค่า a_j หนึ่ง ๆ ในแอตทริบิว) ที่ซึ่งแอตทริบิว A คือแอตทริบิวที่บ่งบอกถึงข้อมูลสีของเรคคอร์ด โดยมีสีที่เป็นไปได้ 5 สีด้วยกัน คือ สีแดง เขียน น้ำเงิน ม่วง และส้ม ในกรณีนี้—เราจะทำการแตกกิ่งจากโหนด N ออกตามค่าที่เป็นไปได้ โดยในการแตกกิ่งจากโหนด N เราจะต้องทำการแบ่งข้อมูลจากชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย $\{D_1, D_2, \dots, D_v\}$ โดยชุดข้อมูลย่อย D_j ใดๆจะสอดคล้องกับค่า a_j ที่เกิดขึ้นในแอตทริบิว A ที่ซึ่งทุก ๆ เรคคอร์ดในชุดข้อมูลย่อย D_j จะมีค่า a_j เกิดขึ้นในแอทริบิว A ดังนั้น เมื่อทำการแตกกิ่งสำหรับโหนด N (แอทริบิว A) แล้ว แอทริบิว A จะไม่ถูกพิจารณาอีกต่อไป สามารถลบแอทริบิว A ออกจากลิสต์ของแอตทริบิวที่ทำการพิจารณาได้ (ขั้นตอนที่ (8) และ (9))

กรณีที่ A มีค่าแบบไม่ต่อเนื่อง จะทำการตรวจสอบโหนด N ภายใต้ 2 เงื่อนไข โดยเราจะต้องทำการหาจุดแบ่งเงื่อนไข (split_point) ถ้าค่าในเรคคอร์ดใด ๆ ก็ตามที่มีค่าน้อยกว่า หรือเท่ากับจุดแบ่งเงื่อนไขจะกำหนดให้ค่านั้น ๆ อยู่กลุ่มของชุดข้อมูลย่อยทางฝั่งซ้าย แต่ถ้าค่าในเรคคอร์ดมีค่ามากกว่าจุดแบ่งเงื่อนไขจะกำหนดให้อยู่ในกลุ่มของชุดข้อมูลย่อยทางฝั่งขวา จุดแบ่งเงื่อนไขจะได้มาจากการเรียกใช้ attribute selection method (โดยทั่ว ๆไป จุดแบ่งเงื่อนไขมักเป็นค่ากลางของค่าทั้งหมดที่เกิดขึ้นในแอทริบิว เช่น ในแอตทริบิวมีค่าเกิด ขึ้นอยู่ในช่วงระหว่าง 1 - 9 ดังนั้น เราจะสามารถนำค่า 5 ซึ่งเป็นค่ากลางของข้อมูลมาใช้ เป็นจุดแบ่งเงื่อนไขได้) ข้อมูลรายได้ของคนจะถูกแบ่งตามจุดแบ่ง เงื่อนไขซึ่งมีค่าเท่ากับ 42,000 ถ้าข้อมูลเรคคอร์ดใด ๆ ก็ตามมีค่าเงินรายได้น้อยกว่าหรือเท่ากับ 42,000 เรคคอร์ดนั้น ๆ จะถูกเก็บไว้ในชุดข้อมูลย่อย D_1 ในส่วนกรณีอื่น ๆ จะถูกเก็บ อยู่ในชุดข้อมูลย่อย D_2 ตามลำดับ

กรณีที่ A มีค่าแบบไม่ต่อเนื่อง และเราต้องการสร้างต้นไม้ให้อยู่ในรูปแบบของต้นไม้ ไบนารี (อาจจะเกิดจากข้อจำกัดของตัวชี้วัดการเลือกแอตทริบิวที่สามารถสร้างต้นไม้ได้ในลักษณะที่เป็น

ต้นไม้ไบนารีเท่านั้น) เราจะต้องมีเซตของค่าในแอตทริบิว A ที่ใช้ในการแบ่งชุดข้อมูล ออกเป็นสองส่วน (เซต S_A ซึ่งถูกเรียกว่า splitting subset) โดยเซต S_A จะได้มาจากการ เรียกใช้ attribute_selection_method เซต S_A จะประกอบไปด้วยค่าต่างๆที่เกิดขึ้นในแอทริบิว A ในเซต S_A ที่เป็น splitting subset จะประกอบไปด้วยสองแอตทริบิวด้วยกัน คือ สีแดงและสีเขียว ดังนั้น เมื่อทำการพิจารณาเรคคอร์ดหนึ่งๆแล้ว ถ้าค่าในแอทริบิว A ของเรคคอร์ดนั้น ๆ มีค่าอยู่ในเซต S_A เราจะทำการจัดเก็บเรคคอร์ดนั้นไว้ในชุดข้อมูลย่อย D_1 ส่วนในกรณีอื่นจะเก็บไว้ในชุดข้อมูลย่อย D_2 ตามลำดับ

ขั้นตอนที่ (4) อัลกอริทึมการสร้างต้นไม้ตัดสินใจจะใช้กระบวนการแบบเวียนเกิด (recursive) ในการสร้างหรือแตกกิ่งต้นไม้/แบ่งข้อมูลออกเป็นชุดข้อมูลย่อยต่อไปเรื่อย ๆ (ขั้นตอนที่ (14))

ขั้นตอนที่ (5) การเวียนเกิดจะเสร็จสิ้นก็ต่อเมื่อการทำงานผ่านเงื่อนไขใดเงื่อนไขหนึ่งดังต่อไปนี้

ทุก ๆ เรคคอร์ดในชุดข้อมูล D ที่กำลังพิจารณาอยู่ขณะที่โหนด N มีหมวดหมู่ของข้อมูลเดียวกัน (ขั้นตอนที่ (2) และ (3))

ไม่มีแอตทริบิวใดเหลืออยู่ในลิสต์ของแอตทริบิวที่ยังไม่ทำการพิจารณา (ขั้นตอนที่ (4)) ในกรณีนี้ เราจะทำการแตกกิ่งที่เป็นโหนดใบให้แก่โหนด N แล้วทำการแนบหมวดหมู่ของข้อมูล C ให้แก่โหนดใบ โดยที่ C ได้มาจากการใช้ majority vote (ขั้นตอนที่ (5))

ไม่มีเรคคอร์ดใดๆเลยสำหรับชุดข้อมูลย่อย D_j (ขั้นตอนที่ 12) ที่ได้หลังจากการแตกกิ่งในกระบวนการสร้างต้นไม้ ในกรณีนี้ - โหนดใบจะทำถูกสร้างด้วย majority vote ของหมวดหมู่ของข้อมูล

<u>เทคนิคการจำแนกข้อมูลด้วยวิธีนาอีฟเบย์ (Naive Bayes)</u>

อัลกอริทึมนาอีฟเบย์ หมายถึง เครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น ตามทฤษฎีของเบย์(Bayes Theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ใช้การคำนวณหาความน่าจะเป็นในการทำนายผลเป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้

จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมาก และคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันโดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

อัลกอริทึมเบย์อย่างง่าย (Naïve Bayes) เป็นรูปแบบการหาความสัมพันธ์ที่ไม่ซับซ้อนและได้ผลลัพธ์ดี ใช้วิเคราะห์หาความน่าจะเป็นของเหตุการณ์ที่ยังไม่เคยเกิดขึ้น โดยคาดเดาจากเหตุการณ์ที่เคยเกิดขึ้นมาก่อน รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย ได้ใช้อัลกอริทึมเบย์อย่างง่าย ในการทำนายลักษณะจำแนกผู้ป่วย โรคเบาหวาน โดยมีผลการทำนายค่าความถูกต้อง โดยเฉลี่ย 75.59% (รุ่งโรจน์ บุญมา, และนิเวศ จิระวิชิตชัย, 2563) อัลกอริทึมนี้ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem (Roiger & Geatz, 2003)

นาอีฟเบย์ (Naïve Bayes) เป็นขั้นตอนวิธีที่ได้รับความนิยมและถูกนำมาใช้อย่างแพร่หลายในงานจำแนกหมวดหมู่เอกสาร เนื่องจากความเรียบง่ายของขั้นตอนวิธีและให้ประสิทธิภาพการจำแนกที่ดี นาอีฟเบย์เป็นขั้นตอนวิธีที่มีพื้นฐานมาจากทฤษฎีเบย์ (Bayes' Theorem) ซึ่งอาศัยหลักความน่าจะเป็นในการทำนายผลลัพธ์ โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ สำหรับรูปแบบการคำนวณความน่าจะเป็นของนาอีฟเบย์สามารถคำนวณได้จากสมการที่ 3

$$p(\text{Class}_j|\text{Place}_i) = \frac{p(\text{Class}_j)\times p(\text{Place}_i|\text{Class}_j)}{p(\text{Place}_i)}$$

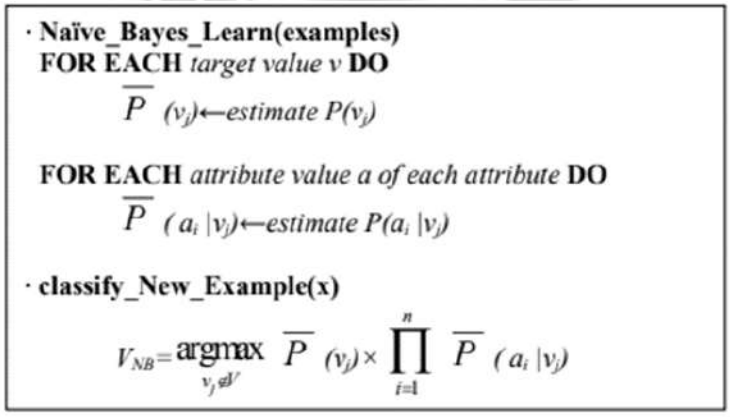
โดยที่ $p(\text{Class}_j|\text{Place}_i)$ ความน่าจะเป็นที่สถานที่ (Place) ที่ i จะอยู่ในหมวดหมู่ (Class) ที่ j เมื่อ $1 \le i \le n$, $1 \le j \le 5$ และ n คือ จำนวนสถานที่ทั้งหมด

$p(\text{Class}_j)$ ความน่าจะเป็นของหมวดหมู่ (Class) ที่ j

$p(\text{Place}_i)$ ความน่าจะเป็นของสถานที่ (Place) ที่ i

$p(Place_i|Class_j)$ ความน่าจะเป็นที่คุณลักษณะ (f_{1-n}) ของสถานที่ ($Place$) ที่ i ปรากฏในหมวดหมู่ ($Class$) ที่ j สามารถคำนวณได้จากสมการที่ 4

$$p(Place_i|Class_j) = p(f_1, f_2, \dots, f_n|Class_j) = \prod_{k=1}^{n}(f_k|Class_j)$$


· **Naïve_Bayes_Learn(examples)**
FOR EACH *target value v* **DO**

$\overline{P}$ $(v_j) \leftarrow$ *estimate* $P(v_j)$

FOR EACH *attribute value a of each attribute* **DO**

$\overline{P}$ $(a_i|v_j) \leftarrow$ *estimate* $P(a_i|v_j)$

· **classify_New_Example(x)**

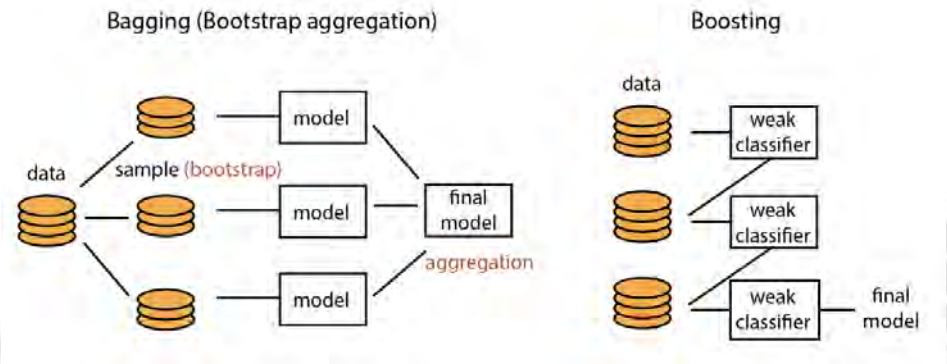
$$V_{NB} = \underset{v_j \in V}{\text{argmax}} \ \overline{P}\ (v_j) \times \prod_{i=1}^{n} \overline{P}\ (a_i|v_j)$$

ภาพที่ 2.6 แสดงอัลกอริทึมสำหรับการจำแนกประเภทข้อความโดยใช้การเรียนรู้นาอีฟเบย์

<u>เทคนิคเบ็คกิ้ง (Bagging)</u>

Bagging (ย่อมาจาก Bootstrap Aggregation) ซึ่งเป็นพื้นฐานของ Random Forest Classifier ใน scikit-learn library เป็นพื้นฐานของอัลกอริทึมที่คนใช้กันบ่อยมากๆได้แก่ Random Forest Classifier นั่นเอง คำว่า Bagging นั้นย่อมาจาก “bootstrap aggregation” ซึ่งถ้ามีใครเคยเรียนวิชาสถิติจะรู้ว่า bootstrap คือ การสุ่มข้อมูลมาจากข้อมูลประชากร เพื่อใช้คำนวณค่าทางสถิติของประชากรกลุ่มเล็ก ๆ ที่เราสุ่มออกมา และ aggregation ก็คือ การเอารวมกัน ดังนั้น Bagging หรือ bootstrap aggregation ก็คือ การสุ่มตัวอย่างข้อมูลออกมาแล้วสร้าง classifier ขึ้นมานี่เอง สำหรับวิธีการสุ่มข้อมูลออกมา เราใช้วิธีสุ่มแบบแทนที่ (random with replacement) ซึ่งหมายความว่าข้อมูลที่เรามียังอยู่เหมือนเดิม ไม่ได้ลดลงหลังจากการสุ่ม เราสามารถสุ่มข้อมูลหลาย ๆ รอบเพื่อให้ได้ classifier หลาย ๆ ตัว แล้วเวลาทำนายก็ใช้ classifiers ทุกตัวที่สร้างขึ้นมาเพื่อทำนายชุดข้อมูลใหม่ที่เจอ การทำนายก็มีได้หลายแบบ ได้แก่ การเฉลี่ย

หรือการโหวตก็ได้ แล้วแต่ว่าเราทำนายความน่าจะเป็นหรือทำนายประเภทข้อมูลจะใช้เทคนิคการจำว่า bagging คือ การสุ่มข้อมูลมาเป็นถุง ๆ แล้วสร้างโมเดลจากถุงข้อมูลที่หยิบออกมาก็ได้นอกจากสุ่มข้อมูลแล้วเรายังสามารถสุ่ม features ของข้อมูลได้อีกด้วย วิธีการสุ่มข้อมูลนี้ทำให้เราได้โมเดลหลาย ๆ ตัวมาช่วยกันทำนาย ทำให้เราสามารถลดแนวโน้มที่โมเดลจะ over fit ข้อมูลที่เรามีอีกทางหนึ่ง เช่น ข้อมูลมี 20 features (20 columns) เราก็สามารถสุ่มข้อมูลออกมาโดยใช้ features เพียงครึ่งเดียวได้



ภาพที่ 2.7 แสดงความแตกต่างระหว่าง Bagging และ Boosting (Machine Learning)

<u>คุณลักษณะของอาชีพ</u>

<u>1. โปรแกรมเมอร์ (Programmer)</u>

Programmer คือ ผู้นำข้อมูลไปออกแบบรายละเอียดการวางโครงสร้างระบบคอมพิวเตอร์ โดยเขียนโปรแกรมด้วยภาษาทางคอมพิวเตอร์ที่แตกต่างกัน เช่น ภาษาซี และจาวา พวกเขามีหน้าที่เขียนและทดสอบรหัสหรือโค้ดเพื่อทำให้คอมพิวเตอร์และซอฟต์แวร์ทำงานได้ และทำการตรวจสอบรหัสที่เกิดข้อผิดพลาดหรือซ่อมแซมแก้ไข ในกรณีที่เกิดปัญหาจากการทำงาน

Programmer หรือคอมพิวเตอร์โปรแกรมเมอร์ ทำหน้าที่รับรายละเอียดของความต้องการของผู้ใช้งานจากนักวิเคราะห์ระบบ (System Analyst) แล้วจัดทำแผนขั้นตอนการทำงาน (Flow Chart) ที่ละเอียดและถูกต้อง เพื่อประโยชน์ในการเขียนโปรแกรมสำหรับการประมวลผลด้วยคอมพิวเตอร์ โดยจะทำหน้าที่ร่วมกันกับนักพัฒนาซอฟต์แวร์

โดยโปรแกรมที่ถูกสร้างมีจุดประสงค์เพื่อจัดการรหัสคอมพิวเตอร์ที่มีความซับซ้อน ซึ่งโปรแกรมที่ใช้ในการเขียนแอปพลิเคชันในมือถือมีจะมีความซับซ้อนน้อยกว่าเครือข่ายอื่นๆ ซึ่งระยะเวลาการทำงานในโปรแกรมที่ง่ายอาจใช้เวลาอันสั้น แต่สำหรับโปรแกรมที่มีความซับซ้อน เช่น ระบบคอมพิวเตอร์ อาจใช้เวลานานเป็นปีหรือมากกว่านั้น

ขั้นตอนการทำงาน

1. ทำความเข้าใจวัตถุประสงค์ในการทำงานของโปรแกรม และวางแผนโดยเขียนแผนภาพ ขั้นตอนของโปรแกรมโดยละเอียด
2. เขียนโปรแกรมด้วยภาษาทางคอมพิวเตอร์ต่างๆ เช่น ภาษา C++ ภาษา Java
3. อัปเดตและพัฒนาขยายโปรแกรม
4. ทดสอบโปรแกรมและแก้ไขปัญหาหรือข้อบกพร่องที่เกิดขึ้น
5. สร้างและทดสอบรหัสในการพัฒนาระบบคอมพิวเตอร์

คุณสมบัติของนักโปรแกรมเมอร์

ลักษณะการทำงานของนักโปรแกรมเมอร์ จะทำหน้าที่นำข้อมูลการออกแบบรายละเอียดการวางโครงสร้างระบบคอมพิวเตอร์ จากนักวิเคราะห์ระบบงานมาเขียนเป็นโปรแกรมต่าง ๆ ภาษาที่ใช้ในการเขียนโปรแกรมจะแตกต่างกันไปตามลักษณะเครื่องของระบบฐานข้อมูล ทดสอบระบบ และส่งให้นักวิเคราะห์ระบบทำการตรวจสอบอีกครั้งเพื่อหาจุดบกพร่องและแก้ไขก่อนนำไปใช้จริง

โปรแกรมเมอร์ยังต้องทำหน้าที่ รับรายละเอียดของความต้องการของผู้ใช้ระบบ (User) จากนักวิเคราะห์ระบบ (System Analyst) จัดทำแผนภูมิ (Flowchart) ขั้นตอนการทำงานที่ละเอียด และถูกต้องตามหลักวิชา เพื่อประโยชน์ในการเขียนโปรแกรมสำหรับการประมวลผลด้วยคอมพิวเตอร์ วิเคราะห์แผนภูมิหรือแผนผังสายงาน แต่เพียงบางส่วนหรือทั้งหมด

คุณสมบัติที่จำเป็นในการเป็นโปรแกรมเมอร์

1. มีความรู้ทางด้านคอมพิวเตอร์ซึ่งสามารถเข้ารับศึกษาได้ในสถาบันการศึกษาที่เรียนทำการสอนหรือสำเร็จการศึกษาปริญญาตรีทางด้านคอมพิวเตอร์

2. มีทักษะในการเขียนภาษาคอมพิวเตอร์

3. มีความคิดสร้างสรรค์สามารถประยุกต์และดัดแปลงความรู้ความสามารถทางด้านโปรแกรมคอมพิวเตอร์ได้เป็นอย่างดี

ผู้ที่ต้องทำงานด้วย

1. โปรแกรมเมอร์ร่วมทีมในการสร้างโปรแกรมแต่ละครั้งอาจมีขนาดงานที่ใหญ่เกินกว่าโปรแกรมเมอร์คนหนึ่งจะแบกรับไว้ได้ ผู้ร่วมทีมจะช่วยประสานและแบ่งงานกันให้ภารกิจเสร็จลุล่วงได้อย่างถูกต้อง ทันเวลา

2. Business Analyst ทำหน้าที่ประสานงานและรับโจทย์จากลูกค้าหรือผู้บริหารที่ต้องการโปรแกรมคอมพิวเตอร์ แล้วนำมาถ่ายทอดส่งต่อให้ทีมโปรแกรมเมอร์ดำเนินการ

3. System Analyst ช่วยทำหน้าที่จัดสรรและกระจายงานต่าง ๆ ไปให้โปรแกรมเมอร์ในทีมสร้างสรรค์โปรแกรมตามความถนัดและตามโจทย์ที่ได้รับจาก Business Analyst โดย System Analyst ต้องมีความรู้และเข้าใจระบบการทำงานของโปรแกรมเมอร์ เพื่อจะสามารถบริหารงานได้อย่างราบรื่น

4. Graphic Designer งานโปรแกรมที่เราเห็นสวยงามได้ ไม่ได้เกิดขึ้นจากการเขียน Code จากโปรแกรมเมอร์เพียงอย่างเดียว แต่ต้องรวมองค์ประกอบจากเนื้อหาและงานภาพที่สวยงามจากการออกแบบของ Graphic Designer ด้วย

ความก้าวหน้าในการประกอบอาชีพ

นักเขียนโปรแกรมคอมพิวเตอร์สามารถเลื่อนตำแหน่งให้สูงขึ้นได้ หากมีความสามารถในการวิเคราะห์ระบบและมีทักษะในการสื่อสารและถ่ายทอดความรู้ที่ดี สามารถก้าวไปยังตำแหน่ง นักวิเคราะห์ระบบงานหรือตำแหน่งที่สูงขึ้นไปอีกก็ได้ หรืออาจจะหาอาชีพเสริมได้ด้วยการรับสอนภาษาคอมพิวเตอร์

และรับเขียนโปรแกรมและวางระบบคอมพิวเตอร์ในหน่วยงานหรือองค์กรต่าง ๆ รับเขียนโปรแกรมสำเร็จรูป หรือจัดตั้งบริษัทที่ปรึกษาทางด้านคอมพิวเตอร์ก็ได้

ความต้องการของตลาดแรงงานการพัฒนาศักยภาพทางคอมพิวเตอร์ยังสามารถขยายตัวไปได้อีกมาก จำนวนโปรแกรมเมอร์ที่มีอยู่ในปัจจุบันจึงยังไม่เพียงพอกับความต้องการขยายตลาดวงการไอที อาชีพนี้จึงยังมีแนวโน้มความต้องการในตลาดแรงงานค่อนข้างสูงและให้ผลตอบแทนสูง สำหรับผู้ที่มีประสบการณ์และความสามารถมาก ดังนั้นโปรแกรมเมอร์จึงควรที่จะศึกษาหาความรู้เพิ่มขึ้นเพื่อนำมาปรับปรุงใช้ในงานและพัฒนาฝีมือให้เป็นที่ยอมรับมากขึ้น

2. นักออกแบบเว็บไซต์ (Website Designer)

Web Design คือ ผู้ที่ทำหน้าที่ออกแบบเว็บไซต์ ซึ่งสำคัญมาก ในหลายบริษัทที่ทำธุรกิจทางด้านสร้างและออกแบบเว็บไซต์โดยตรง จะต้องมีพนักงานที่ทำหน้าที่ Web Design ซึ่งมีหน้าที่ในการออกแบบจัดทำเรื่องราวที่เกี่ยวกับกราฟิก สีสัน เลย์เอาต์ ของหน้าเว็บเพจทั้งหมด ตามที่ Web master ได้ทำการกำหนดทิศทางรูปแบบของเว็บไว้แล้ว หรืออาจจะนำเสนอสิ่งสร้างสรรค์ให้ Web master พิจารณา ผู้ที่จะทำหน้าที่ Web Design ควรมีความคิดในเชิงสร้างสรรค์มีจินตนาการ สามารถใช้งานโปรแกรมประเภทกราฟิก ดีไซน์ ได้อย่างคล่องตัว เช่น โปรแกรม Photo Shop, Flash เป็นต้น และเป็น Creator ที่ดี ไม่ลอกผลงานผู้อื่น หรือหากมีการนำกราฟิกจากที่อื่นมาใช้ควรให้เครดิตผู้สร้างสรรค์ แต่ในบริษัทบางแห่งอาจรวมคนเขียน Code HTML ให้ทำ Graphic ไปด้วย โปรแกรมที่นิยมใช้ในการสร้างและออกแบบเว็บไซต์ เช่น Frontpage, Dreamweaver เป็นต้น

คุณสมบัติของนักออกแบบเว็บไซต์

Website-Designer ได้แก่ ผู้ออกแบบ และเขียนโปรแกรมคอมพิวเตอร์ สำหรับการนำเสนอในเว็บไซต์ เพื่อโฆษณาสินค้าและบริการ โครงการรณรงค์ต่าง ๆ รวมทั้งการเผยแพร่ข้อมูลของสถาบันหรือ

หน่วยงานของสถานประกอบการที่มอบหมายให้จัดทำ หรือสิ่งอื่น ๆ เพื่อเสนอต่อสาธารณชนทางระบบเครือข่ายทางอินเทอร์เน็ต

ผู้ประกอบนักออกแบบเว็บไซต์ (Website-Designer) ต้องมีคุณสมบัติ ดังนี้

1. สำเร็จการศึกษาขั้นต่ำตามระเบียบบังคับของกระทรวงศึกษาธิการ มีปฏิภาณไหวพริบดี

2. มีความสามารถในการเขียนโปรแกรมการสร้างเว็บไซต์ และออกแบบได้ เช่น Java HTML

3. มีความคิดสร้างสรรค์ ชอบในงานศิลป์ และสนใจในการใช้ระบบงานคอมพิวเตอร์

4. มีความสามารถในการเรียนรู้สิ่งใหม่ ๆ และพร้อมที่จะพัฒนาตนเองอยู่เสมอ

5. เป็นคนที่มีมุมมองไม่เหมือนคนอื่น และมีแง่มุมหลายมุมมอง

6. เป็นคนทันสมัย มีความรู้รอบตัว มีความคิดกว้างไกล และมีจินตนาการ

7. มีทัศนคติที่ดี ตรงต่อเวลา มีความรับผิดชอบสูงทั้งต่อลูกค้าและสังคม

8. มีความซื่อสัตย์ในอาชีพ ไม่ใช้ความรู้ ความสามารถในการดัดแปลงข้อมูลเพื่อประโยชน์ส่วนตัว มีความรับผิดชอบในงานที่ได้รับมอบหมาย ควรจะมีมนุษยสัมพันธ์ที่ดีเนื่องจากหน่วยงานที่ว่าจ้าง และผู้เข้าชมเว็บไซต์อาจจะต้องการความช่วยเหลือและคำแนะนำในด้านการใช้งานจึงต้องมีความสามารถชี้แจง ให้ข้อเสนอแนะในการปฏิบัติงานให้แก่ผู้ใช้ระบบงาน รวมทั้งต้องรับฟังความคิดเห็นหรือข้อเสนอแนะจากผู้อื่น

ผู้ที่จะประกอบนักออกแบบเว็บไซต์ (Website-Designer) ควรเตรียมความพร้อม คือ เมื่อสำเร็จการศึกษาตามกฎ ข้อบังคับของกระทรวงศึกษาธิการ หรือสำเร็จการศึกษาในระดับประกาศนียบัตรวิชาชีพและสนใจศึกษาในการเขียนโปรแกรมเพื่อสร้างเว็บไซต์ รวมทั้งมีความสามารถในเชิงศิลป์ หากมีประสบการณ์และประสบความสำเร็จในการสร้างเว็บไซต์จะยิ่งเป็นที่สนใจในการว่าจ้าง โดยอาจจะโฆษณารับเขียนเว็บไซต์ทางระบบอินเทอร์เน็ตหรือสำเร็จการศึกษาระดับปริญญาตรีทางด้านคอมพิวเตอร์ มีความรู้ในการเขียนภาษาคอมพิวเตอร์มีทักษะทางด้านคณิตศาสตร์และภาษาอังกฤษ

สำหรับผู้ที่สนใจประกอบนักออกแบบเว็บไซต์ (Website-Designer) แต่ไม่ได้ศึกษาทางด้านคอมพิวเตอร์มาในระดับปริญญาตรีแต่มีความสนใจในการเขียนโปรแกรมสร้างเว็บไซต์ อาจเข้ารับการอบรม

ตามสถาบันสอนคอมพิวเตอร์ทั่วไป และหากมีความสามารถในการออกแบบเว็บไซต์และมีความเชี่ยวชาญมากพอก็สามารถประกอบนักออกแบบเว็บไซต์ (Website-Designer) ได้เช่นกัน

ลักษณะของงานที่ทำ

1. รับรายละเอียดความต้องการของผู้มอบหมายงานในการจัดทำเว็บไซต์ ศึกษาข้อมูลสิ่งที่ต้องการนำเสนอ เช่น ผลิตภัณฑ์ที่ต้องการโฆษณา โครงการ หรือสถาบันต่าง ๆ ว่ามีจุดกำเนิดอย่างไร มีจุดยืนอย่างไร ต้องการเชิญชวนกลุ่มเป้าหมายใดให้มาสนใจ ด้วยถ้อยคำอย่างไร

2. วิเคราะห์ข้อมูลที่ได้รับ มาใช้สร้างหรือกำหนดลำดับขั้นตอนของการนำเสนอ รวมทั้งกำหนดประเภทและแบบของ การเขียนโปรแกรมในการนำเสนอในเว็บไซต์

3. ออกแบบ การจัดวางเนื้อหาและการเชื่อมสู่รายละเอียดในแต่ละรายการที่ต้องการนำเสนอ (Sitemap) และโครงร่าง (Outline) ของเว็บไซต์

4. ปรึกษาหารือกับผู้ควบคุมงาน และผู้แทนของหน่วยงานต่าง ๆ ที่เกี่ยวข้อง เพื่อพิจารณาแก้ไขปัญหาที่สำคัญ ในการนำเสนอ การนำข้อมูลเข้าระบบ ขอบเขตของการแสดงข้อมูล

5. ออกแบบ การจัดวางภาพ และข้อความ (layout) ในแต่ละเว็บเพจ ซึ่งอาจจะมีผู้ออกแบบกราฟฟิก (Graphic Designers) เป็นผู้ช่วยทำให้การนำเสนองานมีความสมบูรณ์ ก่อนจะส่งให้ผู้ว่าจ้างพิจารณา

6. เปลี่ยนข้อมูลและภาพให้เป็นข้อมูลและภาพที่สามารถนำเสนอในเว็บไซต์ได้

7. ทดสอบความถูกต้องของโปรแกรมและข้อมูลที่นำเสนอ และแก้ไขความคลาดเคลื่อนของโปรแกรมใหม่ให้ถูกต้อง

8. จัดเตรียมข้อสั่งหรือคู่มือการใช้งานระบบนั้นๆ และชี้แจงให้เจ้าหน้าที่ผู้ใช้เครื่อง ได้ใช้เป็นแนวทางในการทำงาน

ผู้ปฏิบัติงานนักออกแบบเว็บไซต์-Website-Designer จะต้องใช้เครื่องคอมพิวเตอร์ในการเขียนและทดสอบ ดังนั้นสถานที่ทำงานจะเป็นสำนักงานที่มีอุปกรณ์ สิ่งอำนวยความสะดวกเช่นสำนักงานทั่วไป มี

การออกไปติดต่อผู้ใช้งานระบบ เพื่อขอข้อมูลเพิ่มเติมเป็นครั้งคราว สำหรับนักออกแบบเว็บไซต์อิสระสามารถทำงานที่บ้านของตนเองได้ รวมทั้งการประสานงานบางอย่าง อาจใช้ระบบการสื่อสารทางอินเทอร์เน็ตช่วยโดยไม่ต้องเดินทางไปสถานประกอบการก็ได้

งานออกแบบเว็บไซต์เป็นงานที่ต้องนั่งอยู่หน้าจอคอมพิวเตอร์เป็นเวลานาน วันหนึ่งประมาณ 6 - 7 ชั่วโมง หรือมากกว่านั้น ต้องใช้ประสาทสัมผัสของสายตาและมือ บางครั้งอาจมีปัญหากับสายตาได้เนื่องจากอยู่กับจอคอมพิวเตอร์เป็นเวลานาน

ประโยชน์ ของ Web Design

1. เพื่อใช้ในการประชาสัมพันธ์สินค้าหรือบริการขององค์กร หรือบริษัท รวมถึงการสร้างภาพลักษณ์ที่ดีน่าเชื่อถือให้เกิดขึ้นกับกลุ่มเป้าหมาย

2. เพื่อใช้ในลักษณะการให้บริการแก่ลูกค้าหรือสมาชิกขององค์กร หรือบริษัท โดยการนำเว็บไซต์มาเป็นเครื่องมือ ช่วยในการอำนวยความสะดวกแก่ลูกค้าหรือสมาชิกเป็นหลัก

3. ผู้ดูแลระบบเครือข่าย (System Administrator)

System Administratorมีหน้าที่บริหารและจัดการเครือข่ายคอมพิวเตอร์ขององค์กร ที่คอยดูแลและจัดการเครือข่ายคอมพิวเตอร์ที่มีความหลากหลายขึ้นอยู่กับหน่วยงานหรือโครงการ แต่โดยทั่วไปแล้วจะมีหน้าที่ติดตั้ง ตอบคำถาม ดูแลเซิร์ฟเวอร์หรือระบบคอมพิวเตอร์ รวมถึงการวางแผนงาน การดูแลควบคุมโครงการที่เกี่ยวข้อง

นอกจากนี้ อาจต้องทำหน้าที่โปรแกรมเมอร์ เช่น การเขียนโปรแกรม รวมถึงการสอนการใช้งานต่อผู้ใช้ทั่วไป ดังนั้นงานตำแหน่งนี้จึงต้องทำทุกอย่างเกี่ยวกับ ระบบ IT ในองค์กร รวมตั้งแต่ server AD, DNS, DHCP, MAIL, File, Printer, Router, internet link, switch, vlan, wan และอื่น ๆ ทุกอย่าง

จากขอบเขตความรับผิดชอบของงานทำให้งานตำแหน่งนี้ต้องมีความรู้ด้าน Operating System หรือ OS ซอฟต์แวร์ระบบปฏิบัติการคอมพิวเตอร์ ขณะเดียวกันยังต้องอัพเดตความรู้ใหม่ ๆ อยู่เสมอ

ดังนั้น ผู้ที่จะทำงานตำแหน่งนี้ได้ จึงต้องมีคุณสมบัติหรือมีความรู้ความสามารถในการดูแลรักษาระบบให้มีความปลอดภัย สามารถทำงานได้ตามปกติ และต้องทำ report performance ของ server ทำแบ็กอัป network monitoring ตรวจสอบความผิดปกติที่จะเกิดขึ้นกับระบบได้ ต้องติดตั้ง ฮาร์ดแวร์ ซอฟต์แวร์ application ต่าง ๆ รวมทั้ง patch และปรับแต่งระบบตามความต้องการขององค์กร แก้ไขปัญหาต่าง ๆ ที่เกิดขึ้นกับระบบและ support user

<u>คุณสมบัติของผู้ดูแลระบบเครือข่าย</u>

ผู้ดูแลระบบเครือข่ายจะช่วยสร้างบำรุงรักษาดูแลและแก้ไขปัญหาคอมพิวเตอร์ข้อมูลและเครือข่ายการสื่อสารที่มีประสิทธิภาพสำหรับองค์กรตามชุดการว่าจ้างผู้ดูแลระบบเครือข่ายของ TechRepublic ในส่วนคำอธิบายงานของชุดจ้างผู้ดูแลระบบเครือข่ายความรับผิดชอบเฉพาะบางรายการรวมถึงการออกแบบการกำหนดค่าเครือข่ายการวิเคราะห์และแก้ไขปัญหาเครือข่ายการปรับปรุงการทำงานของเครือข่ายและประสิทธิภาพการตรวจสอบความปลอดภัยเครือข่ายขององค์กร

ความต้องการผู้ดูแลระบบเครือข่ายคาดว่าจะเพิ่มขึ้น 6% จากปี 2559 เป็นปี 2569 เนื่องจากองค์กรต่าง ๆ ยังคงลงทุนในเทคโนโลยีล่าสุดตามรายงานของสำนักสถิติแรงงาน (BLS)

โดยเฉพาะผู้ดูแลเครือข่ายสามารถคาดหวังอัตราการเติบโตที่สำคัญในการออกแบบระบบคอมพิวเตอร์และอุตสาหกรรมบริการที่เกี่ยวข้องซึ่งคาดว่าจะเติบโต 20% ภายในปี 2569 ในขณะที่ธุรกิจขนาดเล็กถึงขนาดกลาง (SMBs) นำบริการคลาวด์มาใช้มากขึ้น และผู้ดูแลระบบคอมพิวเตอร์จะเพิ่มขึ้นเนื่องจาก SMB มักไม่ได้ติดตั้งแผนกไอทีของตนเองเพื่อการเชื่อมต่อเครือข่ายรายงาน BLS เสริม

เริ่มต้นจากบทบาทผู้ดูแลระบบเครือข่ายผู้เชี่ยวชาญด้านไอทีมีเส้นทางอาชีพที่หลากหลายให้เลือกตามการทำงานของ Zippia ในฐานะโปรไฟล์ผู้ดูแลระบบเครือข่าย ผู้ดูแลระบบเครือข่ายสามารถได้รับการเลื่อนขั้นเป็นผู้จัดการศูนย์ข้อมูลผู้ดูแลระบบอาวุโสผู้อำนวยการฝ่ายเทคโนโลยีสารสนเทศผู้จัดการระบบสารสนเทศและอื่น ๆ ฐานความรู้ที่จำเป็นในการเป็นผู้ดูแลระบบเครือข่ายสามารถนำไปใช้กับตำแหน่งไอทีอื่น ๆ ได้

ทักษะที่ดีที่สุดในการเรียนรู้ที่จะเป็นผู้ดูแลระบบเครือข่าย

ผู้ดูแลเครือข่ายมักจะจบปริญญาตรีสาขาวิทยาศาสตร์คอมพิวเตอร์วิศวกรรมสาขาอื่น ๆ ที่เกี่ยวข้องกับคอมพิวเตอร์หรือการจัดการธุรกิจตามคำอธิบายงานของผู้ดูแลระบบเครือข่ายของอันที่จริง ผู้สมัครอันดับต้น ๆ คาดว่าจะมีการแก้ไขปัญหาเครือข่ายหรือประสบการณ์ด้านเทคนิคอย่างน้อย 2 ปี บางองค์กรต้องการผู้สมัครที่มีวุฒิปริญญาโทด้วยความเชี่ยวชาญด้านเทคโนโลยีสารสนเทศ ผู้ดูแลระบบเครือข่ายควรจะมีความรู้ใน 5 เรื่อง ดังนี้

1. เชี่ยวชาญระบบปฏิบัติการ Linux Serverเนื่องจากในบริษัทหรือองค์กรต่าง ๆ ไม่ได้ใช้ Windows Server อย่างเดียวแต่อาจจะมี Linux Server ที่เป็นระดับ Enterprise อย่าง RedHat Enterprise Linux Server ที่ในปัจจุบันก็ใช้กันมากมายและหลากหลาย

2. ภาษาอังกฤษอยู่ในระดับที่ดีเพราะภาษาที่ใช้ในระบบคอมพิวเตอร์นั้น เกือบทั้งหมดเป็นภาษาอังกฤษ

3. มีความรู้ Technology Virtualization เพราะในปัจจุบันกว่า 95 % ของบริษัทขนาดใหญ่ใช้งาน Virtualization กันเกือบทั้งหมด เนื่องจาก Virtualization มีข้อดีในการช่วยประหยัดทรัพยากรของ Server

4. เข้าใจการทำงานระบบ Network เพราะในหลายๆบริษัท System Administrator อาจต้องทำหน้าที่ Network Administrator ร่วมไปด้วย

5. ต้องรู้ระบบ Docker System ซึ่งจะช่วยให้ ซึ่งจะเข้ามาช่วยพัฒนาโปรแกรมหรือภาษาต่าง ๆ ได้เร็วขึ้น หากคนที่ทำงานตำแหน่งนี้มีความรู้และทักษะในส่วนนี้จะสามารถดูแลรับผิดชอบทั้ง System และ Programing ได้ ซึ่งจะช่วยให้ได้รับการพิจารณา ให้ทำหน้าที่เป็นตำแหน่งนี้อีกด้วย

การจำลองเสมือนเซิร์ฟเวอร์ วิธีปฏิบัติที่ดีที่สุด

ผู้ดูแลระบบเครือข่ายควรทราบวิธีกำหนดค่าเครือข่ายที่ซับซ้อนพร้อมความสามารถในการจัดการควบคุมและตรวจสอบโครงสร้างพื้นฐานของเซิร์ฟเวอร์ตามชุดการว่าจ้างผู้ดูแลระบบเครือข่ายของ

TechRepublic Premium เครือข่ายไร้สายประเภทพื้นฐานที่ผู้เชี่ยวชาญเหล่านี้ควรทราบ ได้แก่ Local Area Network (LAN), Wide Area Network (WAN) และ Virtual Private Network (VPN)

องค์กรจะต้องปรับปรุงโครงสร้างพื้นฐานเครือข่ายให้ทันสมัยอยู่เสมอเพื่อให้สามารถโฮสต์เทคโนโลยีล่าสุดรวมถึงอัปเดตโปรโตคอลความปลอดภัยได้ตามต้องการ อุตสาหกรรมที่เปลี่ยนแปลงตลอดเวลานี้ต้องการผู้ดูแลระบบเครือข่ายที่ยืดหยุ่นซึ่งสามารถปรับเปลี่ยนให้เข้ากับการเปลี่ยนแปลงได้

บริษัทต่าง ๆ ให้ความสำคัญกับทักษะที่อ่อนนุ่มในผู้เชี่ยวชาญด้านไอทีและทักษะเหล่านี้จะนำไปใช้กับผู้ดูแลระบบเครือข่าย โดยเฉพาะอย่างยิ่งผู้ดูแลระบบเครือข่ายควรจะร่วมมือกับทีมทำหน้าที่เป็นทั้งผู้เล่นในทีมและผู้นำโต้ตอบอย่างมีประสิทธิภาพกับหลายระดับขององค์กรทำงานอย่างอิสระโดยไม่ต้องมีการควบคุมดูแลและสื่อสารกับเพื่อนร่วมงานได้ดี

<u>ความก้าวหน้าทางสายอาชีพ</u>

สำหรับผู้ทำงานตำแหน่งนี้ส่วนใหญ่จะจบการศึกษาระดับปริญญาตรีในสาขาวิทยาศาสตร์คอมพิวเตอร์, เทคโนโลยีสารสนเทศ, วิศวกรรมอิเล็กทรอนิกส์, หรือ วิศวกรรมคอมพิวเตอร์ สำหรับอัตราค่าจ้าง หรือรายได้ต่อเดือนของตำแหน่งนี้จะมีความแตกต่างตามขนาดขององค์กร โดยองค์กรขนาดกลางและเล็กจะมีอัตราเงินเดือนเริ่มต้น 15,000 -25,000 บาท ขณะที่องค์กรหรือบริษัทขนาดใหญ่จะมีอัตราค่าจ้างเริ่มต้น 22,000-30,000 บาทต่อเดือน

เนื่องจากเป็นกลุ่มบุคลากรที่มีความรู้ความสามารถและชำนาญใน ระบบ IT ทำให้ผู้ที่มีประสบการณ์ทำงานในตำแหน่งนี้ สามารถผันตัวไปทำงานในตำแหน่ง Network admin หรือ Network Engineer หรือ ผู้ดูแลเน็ตเวิร์ก และยังสามารถต่อยอดไปสู่ระดับ Manager/Senior Managerได้ในอนาคต

<u>งานวิจัยที่เกี่ยวข้อง</u>

สมฤทัย กลัดแก้ว (2558) วิจัยเรื่องการพยากรณ์อาชีพของนักศึกษาเทคโนโลยีสารสนเทศ การศึกษามีวัตถุประสงค์ 1) เพื่อศึกษาปัจจัยที่มีผลต่อการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถ

ของบัณฑิตด้วยการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่มและการจำแนกประเภทข้อมูล ด้วยวิธีต้นไม้ตัดสินใจ 2) เพื่อเปรียบเทียบความถูกต้องของการพยากรณ์ระหว่างการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม (Logistic Regression) และการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจ 3) เพื่อพัฒนาตัวแบบการตัดสินใจการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิต กลุ่มตัวอย่างที่ใช้ในการศึกษา ได้แก่ ข้อมูลภาวะบัณฑิตที่มีงานทำที่เข้ารับพระราชทานปริญญาบัตรในปีการศึกษา 2555-2557 จำนวน 1,933 คน เครื่องมือที่ ใช้ในการศึกษา ได้แก่ แบบสำรวจข้อมูลภาวะบัณฑิตมีงานทำของศูนย์คอมพิวเตอร์มหาวิทยาลัยราชภัฏหมู่บ้านจอมบึง สถิติที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่ การวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม และการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจ พบว่า เทคนิคการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจมีค่าความถูกต้องมากกว่าการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่มเล็กน้อย โดยค่าความถูกต้องของเทคนิคการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจเท่ากับ 57.37% และการวิเคราะห์การถดถอยโลจิสติคพหุกลุ่มมีค่าความถูกต้อง 56.3%

ชัชชฎา วันดี (2556) ศึกษาวิจัยเรื่อง เปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพของนิสิต ระดับปริญญาตรีหลังสำเร็จการศึกษา โดยในงานวิจัยนี้ได้ใช้ชุดข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูล ระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2550-2554 จำนวน 12 คุณลักษณะ และ 2,515 ระเบียน ซึ่งได้นำเทคนิคแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) เทคนิคโครงข่ายประสาทเทียม (Neural Network) และเทคนิคการเรียนรู้แบบเบย์ (Naïve Bayes) มาทำการเปรียบเทียบประสิทธิภาพ ผลจากการศึกษาพบว่าประสิทธิภาพในการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ มีประสิทธิภาพในการจำแนกสูงสุดด้วยค่าเฉลี่ย 88.62% และปัจจัยสำคัญที่ทำให้การเลือกอาชีพตรง หรือไม่ตรงกับสาขา มี 4 ปัจจัย คือ สาขาวิชาที่เรียน เกรดเฉลี่ยเฉพาะวิชาสาขา เพศ และเกรดเฉลี่ยรวม

พฤฒิพงศ์ เพ็งศิริ (2557) นำเสนอการพยากรณ์โดยการประยุกต์ใช้กระบวนการตัดสินใจด้วยเทคนิคต้นไม้ตัดสินใจ เป็นเทคนิคอย่างหนึ่งในการทำเหมืองข้อมูล ซึ่งอาศัยความสัมพันธ์ของปัจจัยข้อมูล

นักศึกษาเป็นการบ่งชี้ถึงระดับผลการเรียนของนักศึกษาผลการทดลองพบว่าปัจจัยข้อมูลของนักศึกษาที่เหมาะสมในการเรียน มีทั้งหมด 7 ตัวแปร จำนวน 4,591 ข้อมูล ข้อมูลนำเข้าทั้งหมด 12 ตัวแปร ทั้งนี้ 7 ตัวแปรมาจากต้นไม้ตัดสินใจที่ได้มาสามารถสรุปเลือกเฉพาะกิ่งที่มีผลมากที่สุดโดยวัดค่าความแม่นยำ (Accuracy) ได้ค่าสูงถึง 84.78% ซึ่งถือว่าอยู่ในเกณฑ์สูงดังนั้นตัวแปรทั้งหมดนี้น่าจะเป็นปัจจัยที่ส่งผลกระทบผลการเรียนที่จบการศึกษาสูงสุดคือความสม่ำเสมอการเข้าเรียนในการเรียน

ภัทร์พงศ์ พงศ์ภัทรกานต์ (2559) งานวิจัยนี้นำเสนอการทดสอบวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา โดยใช้เทคนิคการจำแนกแบบต้นไม้ใช้ข้อมูลการเข้าใช้บริการผ่านประตูอัตโนมัติในช่วงเดือนกุมภาพันธ์ถึงตุลาคม 2559 ที่มี 9 ปัจจัยพื้นฐาน คือ วันที่เข้าใช้บริการ ช่วงเวลา เพศ คณะ ชั้นปี จังหวัดที่เกิด หมู่เลือด จำนวนพี่น้อง และเกรดเฉลี่ยสะสม จำนวน 79,953 ชุดข้อมูล ทำการประมวลผลด้วยอัลกอริทึม C5.0, Neural Network และ CART เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการคัดแยกข้อมูล ผลการศึกษา พบว่า อัลกอริทึม C5.0 ให้ค่าความถูกต้อง 97.78 % และใช้ระยะเวลาในการประมวลผล น้อยกว่าอัลกอริทึมที่นำมาเปรียบเทียบ

ภัทธิรา สุวรรณโค, ดร.นิศาชล จำนงศรี และ ผศ. ดร.จิติมนต์ อั่งสกลุ (2558) งานวิจัยฉบับนี้ได้วิเคราะห์ถึงค่าแม่นยำและค่า AUC ของแบบจำลองพยากรณ์ความเสี่ยงในการเกิดอุบัติเหตุทางถนนในเทศกาลปีใหม่ ด้วยเหมืองข้อมูล ในการวิเคราะห์ได้ใช้ ขั้นตอนการทดลองในการทำเหมืองข้อมูล โดยใช้ข้อมูลจากศูนย์กลางข้อมูลภาครัฐ ระหว่างปี 2550 - 2558 จากข้อมูลมีจำนวน 214,951 รายการ ใช้หลักการแยกด้วยวิธีการ 10-fold cross validation เทคนิคที่นำมาใช้ในแบบจำลองคือ Naïve Bayes Multilayer Perceptron และ Meta Bagging ซึ่งมีผลดังนี้ การสร้างแบบจำลองเหมืองข้อมูลโดยใช้ Multilayer Perceptron และ Meta Bagging มีค่าความถ่วงดุลเท่ากันโดยคิดเป็นร้อย 97.5ซึ่งถือว่ามากที่สุด แต่เมื่อพิจารณาค่า AUC ซึ่งใช้ประกอบการประเมินประสิทธิภาพ พบว่า Meta Bagging ให้ค่าประสิทธิภาพมากที่สุดโดยคิดเป็นร้อยละ 77.3% ซึ่งสามารถสรุปได้ว่าเทคนิค Meta Bagging นี้มีประสิทธิภาพในการพยากรณ์มากที่สุดและมีความเหมาะสมในการนำไปใช้การพยากรณ์ต่อไป

วีรศักดิ์ ฟองเงิน, วรปภา อารีราษฎร์ และ เผด็จ พรหมสาขา ณ สกลนคร (2560) งานวิจัยนี้ได้นำข้อมูลที่เป็นปัจจัย ที่มีผลต่อการเปลี่ยนแปลงระดับน้ำ ประกอบด้วย ปริมาณน้ำไหลเข้าเขื่อน ปริมาณน้ำในเขื่อน ปริมาณการปล่อยน้ำและอัตราการระเหย โดยรวบรวมข้อมูลรายวัน ตั้งแต่ปี พ.ศ 2535 - 2559 จำนวน 9,300 รายการ โดยมีการแยกข้อมูลรายเดือน เพื่อนำมาพยากรณ์ ด้วยเทคนิคพยากรณ์ 4 เทคนิคการวิเคราะห์การถดถอย เทคนิคโครงข่ายประสาทเทียม เทคนิคจำลองต้นไม้เอ็มไฟว์พี และเทคนิคซัพพอร์ตเวกเตอร์แมชชีน โดยใช้โปรแกรม Weka มาใช้ทำแบบจำลอง พบว่า ค่าสัมบูรณ์ของความคลาดเคลื่อนโดยวิธีแบบจำลองต้นไม้เอ็มไฟว์พี มีค่าสัมบูรณ์ของความคลาดเคลื่อนต่ำสุด ที่ 10.56 และเป็นวิธีที่เหมาะสมที่สุดสำหรับไปพัฒนาระบบพยากรณ์น้ำในเขื่อน

สำราญ วานนท์, ธรัช อารีราษฎร์ และจรัญ แสนราช (2561) ได้วิจัยเรื่องการพยากรณ์อาชีพ งานวิจัยนี้ได้นำข้อมูลของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษาย้อนหลัง 5 ปี คือ ปี พ.ศ 2555 - 2559 จำนวน 65,335 ระเบียน ในสาขาวิชาทางด้านคอมพิวเตอร์ ทดลองวัดความแม่นยำด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ เทคนิคการทำนายข้อมูลด้วยวิธีแรนดอมฟอร์เรส และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิง ผู้วิจัยได้ใช้โปรแกรม Weka ในการพยากรณ์ ผลการวิจัยพบว่า ความแม่นยำในการจำแนกประเภทข้อมูลจาก 3 เทคนิค พบว่าเทคนิคแรนดอมฟอร์เรสให้ความถูกต้องในการจำแนก ประเภทข้อมูลได้สูงที่สุดถึง 84.29% สรุปได้ว่า เทคนิคการจำแนกข้อมูลแบบแรนดอมฟอร์เรส เป็นตัวแบบที่เหมาะสมที่จะนำไปพัฒนาเป็นระบบการแนะแนวอาชีพให้กับนักศึกษาระดับปริญญาตรี สาขาคอมพิวเตอร์ต่อไป

ทัศนีย์ เพียรทำดี (2558) ได้วิจัยเรื่องการพยากรณ์คะแนนสอบมาตรฐานวิชาชีพได้ใช้การทดสอบประสิทธิภาพเปรียบเทียบจำแนกตามอัลกอริทึม นาอีฟเบย์, ต้นไม้ตัดสินใจ, ซีโร่อาร์ และเคเนียเรสเนเบอร์ ในการวิจัยนี้ได้ข้อมูลจริงจากฐานข้อมูลนักเรียน โดยนำข้อมูลของนักเรียนระดับปวช. 3 แผนกคอมพิวเตอร์ธุรกิจ ปีที่ใช้ พ.ศ 2555-2557 จำนวน 410 เรคคอร์ด การพยากรณ์นี้ใช้โปรแกรม Weka ผลการวิจัยพบว่า เทคนิคต้นไม้ตัดสินใจ ใช้สร้างตัวแบบการพยากรณ์คะแนนสอบมาตรฐาน วิชาชีพของ

นักเรียน ระดับปวช. 3 แผนกคอมพิวเตอร์ พยากรณ์ได้ถูกต้องถึง 82.68% ซึ่งมีความแม่นยำอยู่ในระดับค่อนข้างสูง การพยากรณ์คะแนนสอบมาตรฐานวิชาชีพของนักเรียน ระดับชั้นปวช. 3 แผนกคอมพิวเตอร์ ได้ถูกต้องว่าจัดอยู่ในกลุ่มใด เพื่อเป็นการกระตุ้นให้นักเรียน มีผลการเรียนที่ดีขึ้นซึ่งจะช่วยให้นักเรียนสอบผ่าน มาตรฐานวิชาชีพ และยังช่วยลดปัญหาการเรียนที่ไม่มีคุณภาพ

วันวิสาข์ ชนะประเสริฐ (2559) ผู้วิจัยได้วิเคราะห์เทคนิคเมืองข้อมูลเพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการทำนายผลสำหรับแนะแนวทางประกอบอาชีพนักศึกษาปริญญาตรี โดยมีข้อมูลประชากรคือ บัณฑิตผู้สำเร็จการศึกษาปริญญาตรี คณะโบราณคดี มหาวิทยาลัยศิลปากร ปีการศึกษา 2556-2558 จำนวน 400 คน มีการวัด ประสิทธิภาพแบบจำลองด้วยวิธี 10 - Fold Cross Validation และวัดค่าความถูกต้องแม่นยำด้วยการวัดค่าความถูกต้องของการจำแนกข้อมูล ใช้เทคนิคแรนดอมฟอร์เรส เทคนิค Neural Network และเทคนิคนาอีฟเบย์ โดยใช้โปรแกรม Weka ในการทดลอง ผลการวิจัยพบว่า เทคนิค Neural Network มีค่าความถูกต้องของถึง 91.35 มีการจำแนกข้อมูลในเกณฑ์ดี และมีประสิทธิภาพมากกว่าทุกเทคนิคที่นำมาใช้

ชณิดาภา บุญประสม และ จรัญ แสนราช (2561) ได้วิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธี Decision Tree, K-Nearest Neighbors และ Naive Bayes โดยใช้ข้อมูลจากฐาน ข้อมูลงานทะเบียนของมหาวิทยาลัยราชภัอุบลราชธานี ของนักศึกษาปริญญาตรี ระหว่างปีการศึกษา 2558-2560 มีจำนวน 13,729 ชุด ข้อมูลเมื่อนำมาวิเคราะห์ค่าน้ำหนักของแอตทริบิวต์ด้วยวิธีการ Information Theory พบว่า มีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาจำนวน 8 ปัจจัย นำปัจจัยที่ได้มาทำการสร้างเป็นโมเดลทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วย ค่า Accuracy เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด ผลการวิจัย พบว่า การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลพบว่าโมเดลที่สร้างด้วยเทคนิควิธี Naive Bayes มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องสูงถึง 93.58%

บทสรุปงานวิจัยที่เกี่ยวข้อง

จากการที่ศึกษางานวิจัยที่เกี่ยวข้องของนักวิจัยหลาย ๆ ท่านสามารถนำมาสรุปได้ดังนี้

ตารางที่ 2.1 แสดงงานวิจัยที่เกี่ยวข้อง

ชื่องานวิจัย (ผู้แต่ง,ปี)	ข้อมูลที่ใช้	เทคนิคที่ใช้	ผลการวิจัย
1) ระบบสนับสนุนการตัดสินใจการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิต สมฤทัย กลัดแก้ว (2558)	ข้อมูลภาวะบัณฑิตที่มีงานทำ ปี 2555-2557 จำนวน 1,933 ระเบียน	Decision Tree	เทคนิคต้นไม้ตัดสินใจ ค่าความถูกต้อง 57.37%
2) เปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพของนิสิต ระดับปริญญาตรีหลังสำเร็จการศึกษา ชัชชฎา วันดี (2556)	ข้อมูลภาวะการณ์มีงานทำ ปี พ.ศ. 2550-2554 จำนวน 2,515 ระเบียน	Decision Tree Neural Network Naïve Bayes	เทคนิคต้นไม้ตัดสินใจ ค่าความถูกต้อง 88.62%
3) การพยากรณ์ความสัมพันธ์ของปัจจัยข้อมูลนักศึกษา พฤฒิพงศ์ เพ็งศิริ (2557)	ข้อมูลนักศึกษา ปี พ.ศ. 2557 จำนวน 4,59 ระเบียน	Decision Tree	เทคนิคต้นไม้ตัดสินใจ ค่าความถูกต้อง 84.78%

ตารางที่ 2.1 แสดงงานวิจัยที่เกี่ยวข้อง (ต่อ)

ชื่องานวิจัย (ผู้แต่ง,ปี)	ข้อมูลที่ใช้	เทคนิคที่ใช้	ผลการวิจัย
4) การใช้เทคนิคเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา ภัทร์พงศ์ พงศ์ภัทรกานต์ (2559)	ข้อมูลการเข้าใช้บริการผ่านประตูอัตโนมัติ ปีพ.ศ. 2559 จำนวน 79,953 ชุด	Decision Tree	เทคนิค Decision Tree ค่าความถูกต้อง 97.78 %
5) แบบจำลองการพยากรณ์ความเสี่ยงในการเกิดอุบัติเหตุทางถนน ภัทธิรา สุวรรณโค ดร. นิศาชล (2558)	ข้อมูลจากศูนย์กลางข้อมูลภาครัฐ ปี 2550 – 2558 จำนวน 214,951 รายการ	Multilayer Perceptron Naïve Bayes Meta Bagging	เทคนิค Meta Bagging ค่าความถูกต้อง 77.3
6) การพยากรณ์ปริมาณน้ำในเขื่อนกิ่วลมโดยใช้เทคนิคเหมืองข้อมูล วีรศักดิ์ ฟองเงิน วรปภา อารีราษฎร์ และ เผด็จ พรหมสาขา ณ สกลนคร (2560)	ข้อมูลที่เป็นปัจจัย ที่มีผลต่อการเปลี่ยนแปลงระดับน้ำ ปี พ.ศ 2535 – 2559 จำนวน 9,300 รายการ	Decision Tree Support Vector Machine	เทคนิค Decision Tree มีค่าคลาดเคลื่อน 10.56%

ตารางที่ 2.1 แสดงงานวิจัยที่เกี่ยวข้อง (ต่อ)

ชื่องานวิจัย (ผู้แต่ง,ปี)	ข้อมูลที่ใช้	เทคนิคที่ใช้	ผลการวิจัย
7) การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาปริญญาตรีสาขาคอมพิวเตอร์ โดยใช้เทคนิคเหมืองข้อมูล สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช (2561)	ข้อมูลของบัณฑิต และข้อมูลระเบียนประวัติ ปี พ.ศ 2555 - 2559 จำนวน 65,335 ระเบียน	Decision Tree Random Forest Bagging	เทคนิคแรนดอมฟอร์เรส ค่าความถูกต้อง 84.29%
8) การพยากรณ์คะแนนสอบมาตรฐานวิชาชีพคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล นางทัศนีย์ เพียรทำดี (2558)	ข้อมูลของนักเรียนระดับ ปวช.3 แผนกคอมพิวเตอร์ธุรกิจ ปี พ.ศ 2555-2557 จำนวน 410 เรคคอร์ด	Decision Tree Naïve Bayes	เทคนิคต้นไม้ตัดสินใจ ค่าความถูกต้อง 82.68%

ตารางที่ 2.1 แสดงงานวิจัยที่เกี่ยวข้อง (ต่อ)

ชื่องานวิจัย (ผู้แต่ง,ปี)	ข้อมูลที่ใช้	เทคนิคที่ใช้	ผลการวิจัย
9) การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล ชณิดาภา บุญประสม และ จรัญ แสนราช (2561)	ข้อมูลงานทะเบียนนักศึกษา ปี 2558-2560 จำนวน 13,729 ชุด	K-Nearest Neighbors Naïve Bayes	เทคนิค Naive Bayes ค่าความถูกต้อง 93.58%
10) การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพคณะโบราณคดี มหาวิทยาลัยศิลปากร นางสาววันวิสาข์ ชนะประเสริฐ (2559)	ข้อมูลบัณฑิตผู้สำเร็จการศึกษาปริญญาตรี ปี 2556-2558 จำนวน 400 คน	Neural Network Random Forest Naïve Bayes	เทคนิค Neural Network ค่าความถูกต้อง 91.35%

บทที่ 3
วิธีการดำเนินงานวิจัย

ศึกษาปัญหาและวิเคราะห์ข้อมูล

งานวิจัยชิ้นนี้ผู้วิจัยเลือกเก็บข้อมูลจากการนำข้อมูลภาวะการมีงานทำกับข้อมูลระเบียนประวัติของผู้สำเร็จการศึกษาคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม ตั้งแต่ปี พ.ศ. 2553 – 2561 จำนวน 1,055 ระเบียน และข้อมูลเกรดเฉลี่ยของแต่ละรายวิชาที่นักศึกษาได้นำมาใช้ประโยชน์กับการทำงานหรือความสอดคล้องต่อสถานที่ทำงาน จากการศึกษาพบว่า คำถามสำคัญที่เกิดขึ้นในระหว่างการเรียนการสอนในแต่ละภาคการศึกษานั้น ปัจจัยใดเป็นปัจจัยสำคัญในการเลือกอาชีพที่เหมาะสมที่สุดเมื่อสิ้นสุดภาคการศึกษา ซึ่งคำถามดังกล่าวจะเกิดขึ้นตลอดระยะเวลาการศึกษา จากคำถามดังกล่าวนำมาสู่การแก้ปัญหาด้วยวิธีการปัจจัยที่ส่งผลต่อการพยากรณ์ผลสัมฤทธิ์ที่จะเกิดขึ้นกับผู้เรียนในอนาคต ผู้วิจัยจึงทำการพยากรณ์จำลองอาชีพของนักศึกษาเพื่อเป็นเครื่องมือช่วยในการตัดสินใจในอนาคตของนักศึกษา และได้ทำการรวบรวมข้อมูลที่สามารถนำมาใช้ในการวิจัยในอนาคต

พิจารณาข้อมูลได้จากกระบวนการศึกษาและอาชีพที่ได้จากการรวบรวมข้อมูล สามารถสรุปลักษณะข้อมูลที่ใช้ในการทำงานวิจัยได้ดังนี้ ข้อมูลที่ใช้สำหรับวิเคราะห์ ผู้วิจัยแบ่งข้อมูลออกเป็นดังนี้ 1. ข้อมูลจากการติดต่อสอบถามอาชีพในปัจจุบันผู้ที่สำเร็จการศึกษา คณะเทคโนโลยีสารสนเทศ สาขาธุรกิจดิจิทัล มหาวิทยาลัยสยาม ในปัจจุบันเพื่อทำการวิเคราะห์ปัจจัยการเลือกอาชีพและความสอดคล้องของสาขาที่ได้ศึกษา 2. ข้อมูลจากผลการเรียนของนักศึกษาเพื่อทำการนำข้อมูลมาพยากรณ์อาชีพ ที่สามารถนำไปวิเคราะห์อาชีพที่สอดคล้องกับวัตถุประสงค์และเป้าหมายของรายวิชาตามที่เรียนในหลักสูตร ชุดข้อมูลสำหรับการวิจัยสามารถแบ่งจำแนกคุณลักษณะข้อมูลได้ดังตาราง

ตารางที่ 3.1 แสดงรายละเอียดคุณลักษณะเพื่อนำมาสร้างตัวแบบ

ลำดับที่	ชื่อแอตทริบิวต์	ค่าตัวแปร	คำอธิบาย
1	PREFIX_T	อักษร	เพศ
2	137-302	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์2
3	137-407	ระดับ	สัมมนาคอมพิวเตอร์ธุรกิจ
4	114-202	ระดับ	ภาษาอังกฤษธุรกิจ
5	130-202	ระดับ	การวิเคราะห์เชิงสถิติทางธุรกิจ
6	137-301	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์ 1
7	GPA	ระดับ	ผลสัมฤทธิ์การศึกษา
8	Target	อักษร	ความสอดคล้องทางอาชีพ

การเตรียมข้อมูล

1. รู้และเข้าใจถึงปัญหา (Business understanding)

นักศึกษาคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม หลังสำเร็จการศึกษา มีปัจจัยในการเลือกอาชีพไม่ตรงกับความสามารถของตนตามหลักสูตร ผู้วิจัยจึงทำการหาเหตุและผลเพื่อทำการวิเคราะห์ปัจจัยเหล่านั้น และทำการใช้เครื่องมือมาช่วยในการวิเคราะห์เพื่อให้ได้ผลลัพธ์การเลือก อาชีพที่เหมาะสมแก่นักศึกษา ซึ่งเทคนิคที่ได้นำมาใช้เป็นเครื่องมือมี3 อัลกอริทึม ได้แก่

1.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

1.2 เทคนิคเบ็กกิ้ง (Bagging)

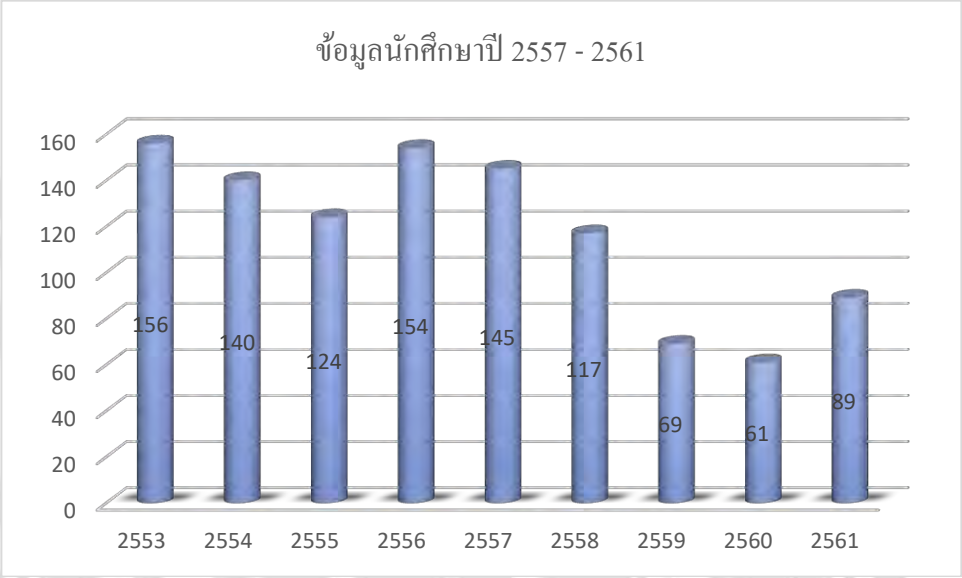
1.3 เทคนิคนาอีฟเบย์ (Naive Bayes)

2. สร้างฐานข้อมูลให้ครบ (Data understanding)

ข้อมูลนักศึกษาสาขาธุรกิจดิจิทัล ระหว่างปี พ.ศ. 2553-2561 จำนวน 1,055 ระเบียนดังตา

ตารางที่ 3.2 แสดงระเบียนข้อมูลนักศึกษาธุรกิจดิจิทัล ระหว่างปี พ.ศ. 2553-2561

ชนิดข้อมูล	จำนวนข้อมูลปี 2552 - 2560								
	2553	2554	2555	2556	2557	2558	2559	2560	2561
ข้อมูลนักศึกษา	156	140	124	154	145	117	69	61	89



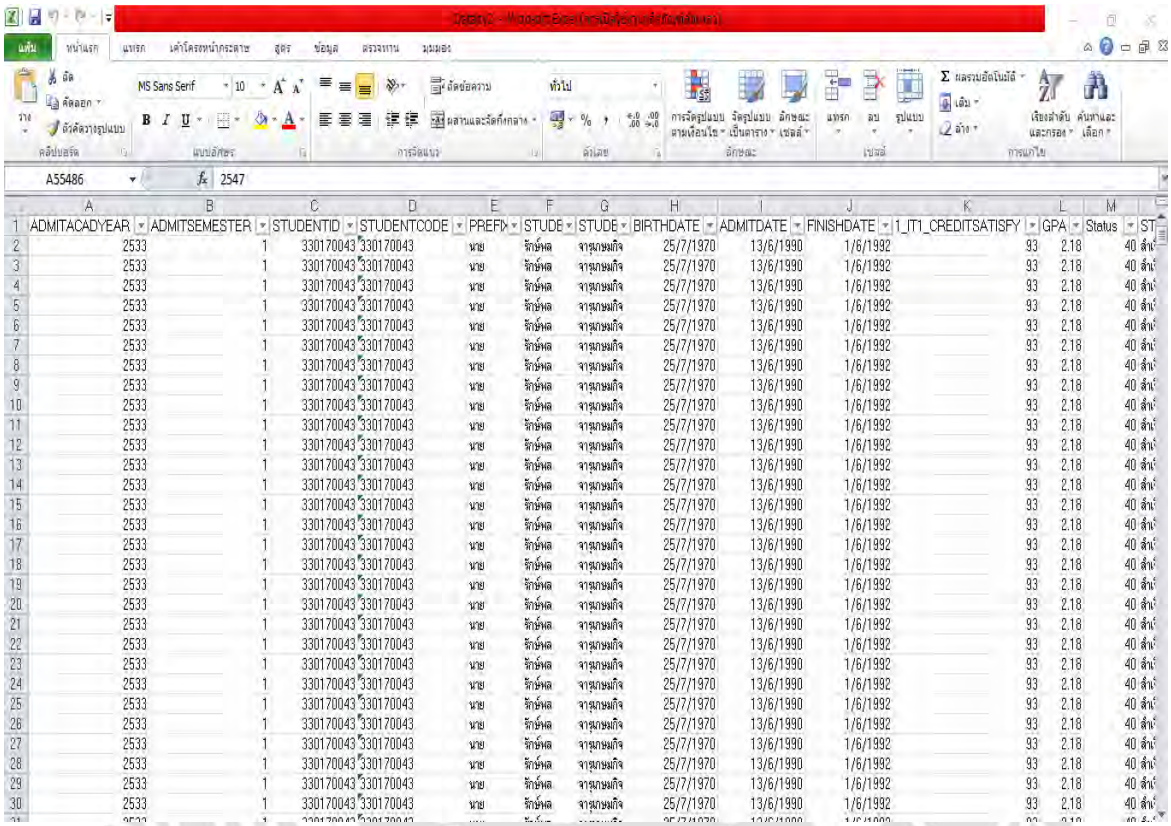
ภาพที่ 3.1 แผนภูมิแสดงจำนวนข้อมูลของนักศึกษา

3. เตรียมข้อมูลให้พร้อมใช้ (Data preparation)

3.1 ทำการสอบถามข้อมูลอาชีพปัจจุบันของนักศึกษา

3.2 คัดเลือกเกรดแต่ละรายวิชาตามหลักสูตรที่กำหนดเพื่อนำมาวิเคราะห์

3.3 แปลงข้อมูลให้เหมาะสมกับการวิเคราะห์



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ADMITACADYEAR	ADMITSEMESTER	STUDENTID	STUDENTCODE	PREFIX	STUDE	STUDE	BIRTHDATE	ADMITDATE	FINISHDATE	1_IT1_CREDITSATISFY	GPA	Status
2	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
3	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
4	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
5	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
6	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
7	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
8	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
9	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
10	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
11	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
12	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
13	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
14	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
15	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
16	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
17	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
18	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
19	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
20	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
21	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
22	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
23	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
24	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
25	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
26	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
27	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
28	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
29	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40
30	2533	1	330170043	330170043	นาย	รักษ์พล	จารุณพงกิจ	25/7/1970	13/6/1990	1/6/1992	93	2.18	40

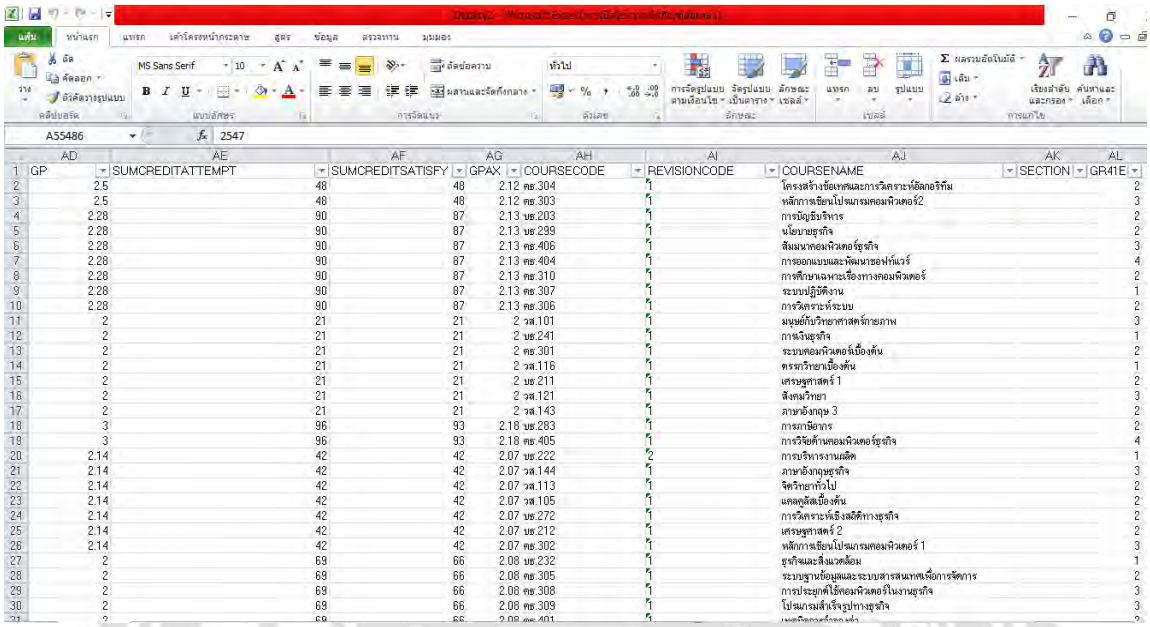
ภาพที่ 3.2 แสดงไฟล์ข้อมูลผลสัมฤทธิ์ แต่ละวิชา ในรูป Main Database (xlsx)

4. จัดทำและเลือกโมเดลที่ใช้ (Modeling)

การสร้างแบบจำลอง ได้แบ่งการพิจารณาจากการวิเคราะห์แต่ละวิธีจากอัลกอริทึม 3 แบบ ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) โครงข่ายประสาทเทียม (Neural Network) และเทคนิคนาอีฟเบย์ (Naive Bayes)

โดยข้อมูลแบ่งเป็น 2 ส่วน คือ

4.1โมเดลแบ่งแยกตามภาควิชาต่างๆของหลักสูตร เช่น สาขาเทคโนโลยีสารสนเทศกับสาขาคอมพิวเตอร์ธุรกิจ



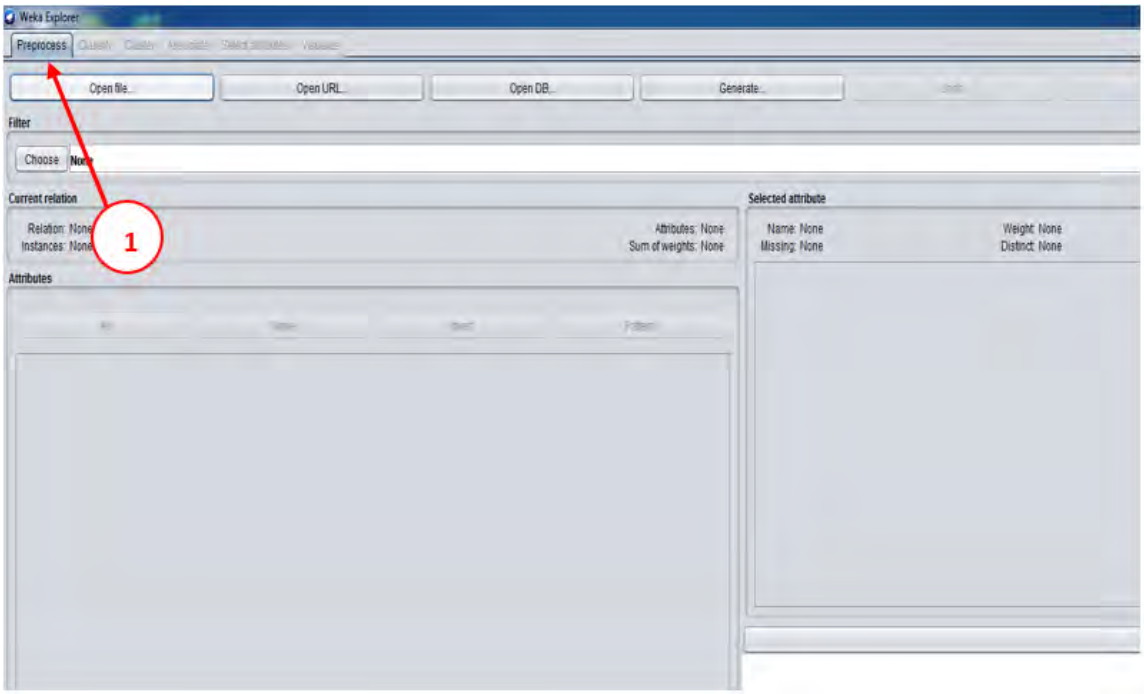
	GP	SUMCREDITATTEMPT	SUMCREDITSATISFY	GPAX	COURSECODE	REVISIONCODE	COURSENAME	SECTION	GR41E
2	2.5	48	48	2.12	ศธ.304	1	โครงสร้างข้อเทศและการวิเคราะห์อัลกอริทึม		2
3	2.5	48	48	2.12	ศธ.303	1	หลักการเขียนโปรแกรมคอมพิวเตอร์2		3
4	2.28	90	87	2.13	บธ.203	1	การบัญชีบริหาร		2
5	2.28	90	87	2.13	บธ.299	1	นโยบายธุรกิจ		2
6	2.28	90	87	2.13	ศธ.406	1	สัมมนาคอมพิวเตอร์ธุรกิจ		3
7	2.28	90	87	2.13	ศธ.404	1	การออกแบบและพัฒนาซอฟท์แวร์		4
8	2.28	90	87	2.13	ศธ.310	1	การศึกษาเฉพาะเรื่องทางคอมพิวเตอร์		2
9	2.28	90	87	2.13	ศธ.307	1	ระบบปฏิบัติงาน		1
10	2.28	90	87	2.13	ศธ.306	1	การวิเคราะห์ระบบ		2
11	2	21	21	2	วส.101	1	มนุษย์กับวิทยาศาสตร์กายภาพ		3
12	2	21	21	2	บธ.241	1	การเงินธุรกิจ		1
13	2	21	21	2	ศธ.301	1	ระบบคอมพิวเตอร์เบื้องต้น		2
14	2	21	21	2	วส.116	1	ดาราศาสตร์เบื้องต้น		1
15	2	21	21	2	บธ.211	1	เศรษฐศาสตร์ 1		2
16	2	21	21	2	วส.121	1	สังคมวิทยา		3
17	2	21	21	2	วส.143	1	ภาษาอังกฤษ 3		2
18	3	96	93	2.18	บธ.283	1	การภาษีอากร		2
19	3	96	93	2.18	ศธ.405	1	การวิจัยด้านคอมพิวเตอร์ธุรกิจ		4
20	2.14	42	42	2.07	บธ.222	2	การบริหารงานผลิต		1
21	2.14	42	42	2.07	วส.144	1	ภาษาอังกฤษธุรกิจ		3
22	2.14	42	42	2.07	วส.113	1	จิตวิทยาทั่วไป		2
23	2.14	42	42	2.07	วส.105	1	แคลคูลัสเบื้องต้น		2
24	2.14	42	42	2.07	บธ.272	1	การวิเคราะห์เชิงสถิติทางธุรกิจ		2
25	2.14	42	42	2.07	บธ.212	1	เศรษฐศาสตร์ 2		2
26	2.14	42	42	2.07	ศธ.302	1	หลักการเขียนโปรแกรมคอมพิวเตอร์ 1		3
27	2	69	66	2.08	บธ.232	1	ธุรกิจและสิ่งแวดล้อม		1
28	2	69	66	2.08	ศธ.305	1	ระบบฐานข้อมูลและระบบสารสนเทศเพื่อการจัดการ		2
29	2	69	66	2.08	ศธ.308	1	การประยุกต์ใช้คอมพิวเตอร์ในงานธุรกิจ		3
30	2	69	66	2.08	ศธ.309	1	โปรแกรมสำเร็จรูปทางธุรกิจ		3

ภาพที่ 3.3 แสดงไฟล์ข้อมูลที่พร้อมในการวิเคราะห์ในรูป Main Database (csv)

จากชุดข้อมูลเรียนรู้ที่ทราบสาเหตุ ผู้วิจัยจึงนำมาเทคนิคนี้มาใช้เพื่อหาอัลกอริทึมที่เหมาะสม โดยให้แบบจำลองที่มีประสิทธิภาพที่สุด โดยการนำชุดข้อมูลเรียนรู้มาทดสอบทีละชุดกับ อัลกอริทึมทั้ง 3 แบบ ใน โปรแกรม Weka

5. ประเมินผลก่อนตัดสินใจ (Evaluation)

ทดสอบข้อมูลด้วยเครื่องมือที่นำมาทำการวิเคราะห์ด้วยโปรแกรม Weka เพื่อคำนวณหาค่าความถูกต้อง การทดสอบจากขั้นตอนการหาแบบจำลองเมื่อได้แบบจำลองที่มีประสิทธิภาพแล้วขั้นตอนต่อไปจะนำแบบจำลองที่ได้มาทดสอบกับชุดข้อมูลทดสอบโดยใช้ชุดทดสอบมาทำการวิเคราะห์ด้วยโปรแกรมดังนี้



ภาพที่ 3.4 แสดงภาพการนำเข้าชุดข้อมูล

Preprocess เป็นส่วนที่ใช้ในการเลือกไฟล์ข้อมูลสำหรับเป็นอินพุต (input) เพื่อทำการวิเคราะห์ข้อมูล

Classify เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการจำแนกข้อมูล (classification) หรือทำนายข้อมูล (prediction) ซึ่งจะมีวิธีการต่าง ๆ ให้เลือกมากมาย

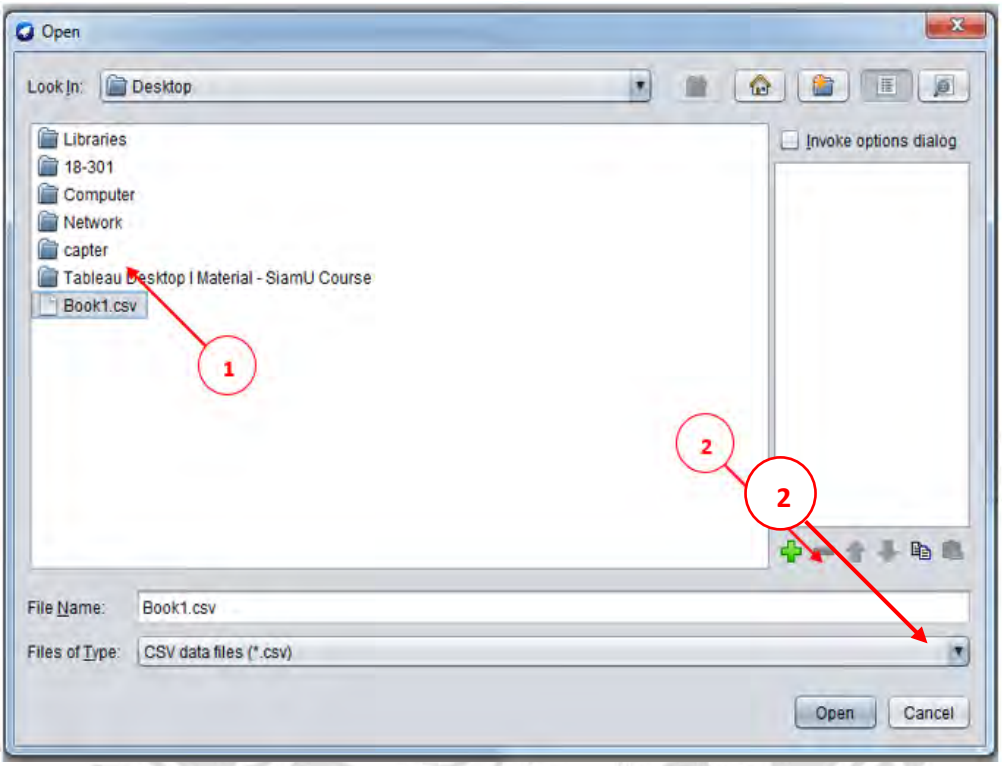
Cluster เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการจัดกลุ่มข้อมูล (clustering) โดยจะจัดกลุ่มข้อมูลที่มีลักษณะคล้าย ๆ กันหรือมีความสัมพันธ์กันเข้าไว้ด้วยกัน

Associate เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการหาความสัมพันธ์ของข้อมูล

Select attributes เป็นส่วนที่คล้าย ๆ กับส่วน Preprocess แต่จะเน้นที่การหาว่าตัวแปรไหนที่สำคัญและไม่สำคัญในชุดข้อมูลบ้าง ซึ่งตัวแปรที่ไม่สำคัญนี้จะถูกกำจัดทิ้งไป ก่อนที่จะวิเคราะห์ข้อมูลด้วยวิธีต่าง

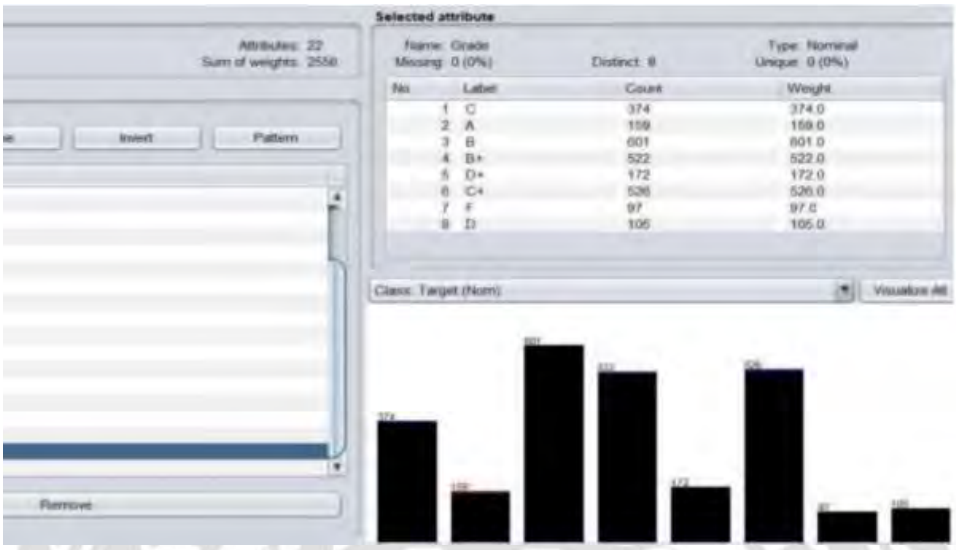
Visualize เป็นส่วนของการ plot จุดข้อมูลในรูปแบบ 2 มิติ

เมื่อทำการเข้ามาแล้ว ให้ทำการนำเข้าข้อมูลที่ได้เตรียมไว้เพื่อทำการวิเคราะห์ข้อมูล โดยคลิกที่ Open File เพื่อให้เกิดหน้าต่างดังรูปต่อไปนี้



ภาพที่ 3.5 แสดงไฟล์ข้อมูลที่ต้องการวิเคราะห์

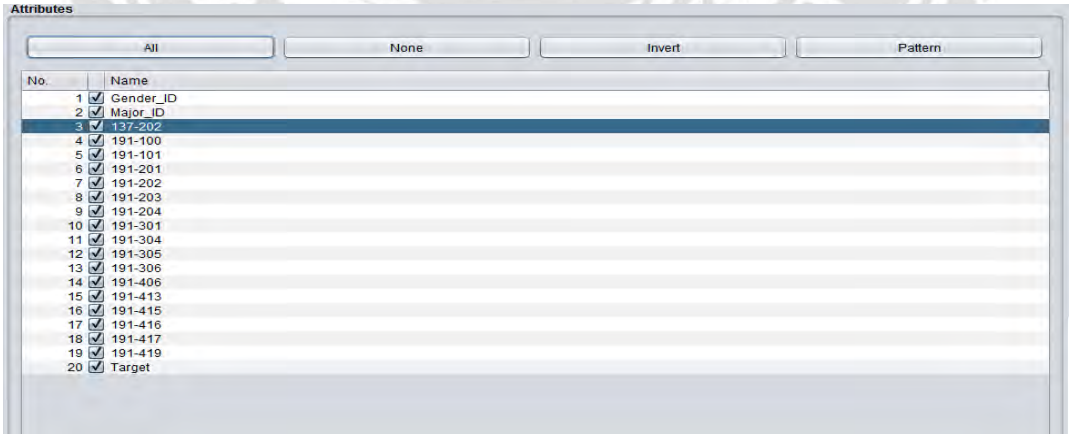
โดยการที่จะสามารถนำเข้าข้อมูลได้นั้น ข้อมูลจะต้องเป็นไฟล์ประเภท .csv เท่านั้น และเมื่อทำการนำเข้าข้อมูลแล้วเราก็จะทำการกด Open เพื่อทำการอินพุตไฟล์ เพื่อให้ได้ดังรูปต่อไป



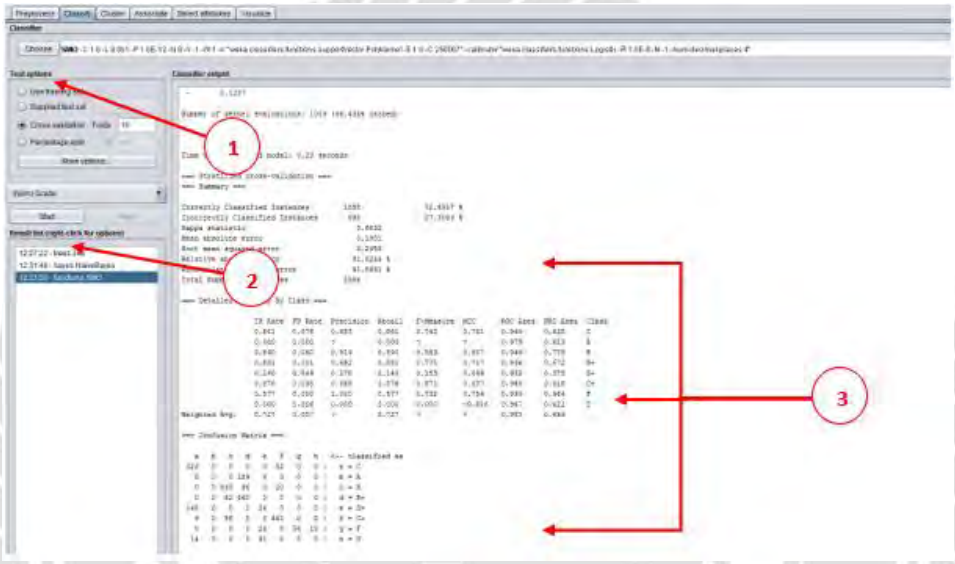
ภาพที่ 3.6 แสดงข้อมูลแอตทริบิวต์

ในส่วนนี้จะแสดงให้เห็นว่าชุดข้อมูลที่ต้องใช้ในการทำนายมีอะไรบ้าง จากภาพจะเห็นได้ว่าเมื่อทำการกดเลือก จะแสดงข้อมูลในแอตทริบิวต์ อย่างเช่นแอตทริบิวต์เกรด จะมีข้อมูลผลสัมฤทธิ์แต่ละรายวิชาเป็นต้น

ภาพที่ 3.7 แสดงการเลือกข้อมูลเพื่อทำการทำนาย



จากนั้นให้ทำการกดเลือกแอตทริบิวท์ที่จะทำการวิเคราะห์ข้อมูลโดยจะทำการกดเลือกข้อมูลทั้งหมดมาทำนาย และเมื่อทำการเลือกข้อมูลแล้ว ให้กดไปที่ Classify เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการจำแนกข้อมูล (classification) หรือทำนายข้อมูล (prediction) ซึ่งจะมีวิธีการต่าง ๆ ให้เลือก ดังรูปต่อไปนี้



ภาพที่ 3.8 แสดงการเลือกอัลกอริทึมเพื่อการทำนายข้อมูลและการแสดงผล

เมื่อทำการเลือกอัลกอริทึมที่กำหนดแล้วจะทำการกด Start เพื่อทำการทำนายข้อมูลเมื่อทำการทำนาย เสร็จแล้วนั้นจะปรากฏข้อมูลผลทำนายดังภาพหมายเลข 3 โปรแกรมจะทำการทำนายและแสดงผลลัพธ์ ในส่วน classifier output เราสามารถดูผลการทำนายได้ เช่น Correctly Classified Instances จะบอกความ ถูกต้องจากการทำนายผลสรุปของการพยากรณ์ทั้งหมดเพื่อให้เราได้ทำการหาผลสรุปถึงข้อมูลการใช้เทคนิค ทั้งหมดที่มีค่าความแม่นยำมากที่สุด

บทที่ 4

ผลการศึกษา

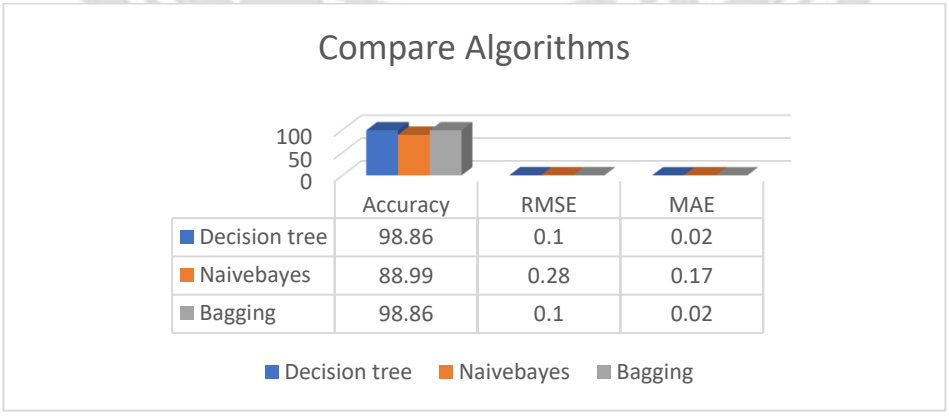
ผลการศึกษาเทคนิคการพยากรณ์จำแนกข้อมูลโดยใช้เทคนิคเหมืองข้อมูล

คณะผู้วิจัยได้ศึกษาเทคนิคเหมืองข้อมูลการจำแนกข้อมูล โดยใช้เทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) เทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้ง (Bagging) และเทคนิคการจำแนกข้อมูลนาอีฟเบย์ (Naive Bayes) วิเคราะห์เปรียบเทียบกับข้อมูลภาวะการมีงานทำใช้ค่าความถูกต้อง ความคลาดเคลื่อนเฉลี่ยกำลังสอง ความคลาดเคลื่อนสมบูรณ์ เพื่อวัดประสิทธิภาพของตัวแบบ ได้ผลการดำเนินงานดังตารางที่ 4.1 และภาพที่ 4.1

ตารางที่ 4.1 แสดงตารางการเปรียบเทียบผลการวิเคราะห์ค่าความถูกต้องของแต่ละเทคนิค

ที่	Algorithms	Accuracy (%)	RMSE	MAE
1	Decision tree	98.86	0.10	0.02
2	Naivebayes	88.99	0.28	0.17
3	Bagging	98.86	0.10	0.02

ตารางที่ 4.1 แสดงกราฟการเปรียบเทียบผลการวิเคราะห์ค่าความถูกต้องของแต่ละเทคนิค



จากตารางที่ 4.1 พบว่า แบบการพยากรณ์อาชีพด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุดที่ร้อยละ 98.86 และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้งให้ค่าความถูกต้องร้อยละ 98.86 และเทคนิคการจำแนกข้อมูลด้วยวิธีนาอีฟเบย์ให้ค่าความถูกต้องร้อยละ 88.99 มีความคลาดเคลื่อนเฉลี่ยกำลังสองเท่ากับ 0.1,0.1 และ 0.28 มีค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์เท่ากับ 0.02, 0.17 และ 0.02 ตามลำดับ เปรียบเทียบค่าความถูกต้องและนำผลไปใช้ในการดำเนินการในขั้นตอนถัดไป

ผลการเปรียบเทียบเทคนิคการพยากรณ์โดยใช้เทคนิคเหมืองข้อมูล

จากการเปรียบเทียบผลการวิเคราะห์ค่าความถูกต้องของแต่ละเทคนิคสร้างตัวแบบการพยากรณ์อาชีพจากข้อมูลภาวะการมีงานทำของบัณฑิต นำมาทดสอบค่าทางสถิติแบบ Paired T-Test พบว่าเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจและวิธีแบ็กกิ้งให้ค่าความถูกต้องสูงสุดจากเทคนิคทั้งหมด 3 เทคนิคอย่างมีนัยสำคัญทางสถิติที่ระดับค่าความเชื่อมั่น ดังตารางที่ 4.2

ที่	Algorithms	Accuracy (%)	Number of time
1	Decision tree	98.86	0/0/1
2	Naivebayes	88.99	0/0/1
3	Bagging	98.86	0/0/1

ตารางที่ 4.2 แสดงตารางการทดสอบค่าทางสถิติแบบ Paired T-Test

จากตารางที่ 4.2 พบว่าการทดสอบจากข้อมูลชุดเดียวกันด้วยเทคนิคทั้ง 3 โดยใช้เทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจและวิธีแบ็กกิ้งเป็นฐานจะได้ค่าทางสถิติที่ดีกว่าเทคนิคการจำแนกข้อมูลด้วยวิธีเทคนิคการจำแนกข้อมูลด้วยนาอีฟเบย์

ผลการทดสอบการพยากรณ์ด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ

การพยากรณ์ด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจและวิธีแบ็กกิ้ง โดยแบ่งชุดข้อมูลออกเป็น 10 ส่วน (10-fold) ผลลัพธ์ที่ได้ตามตาราง confusion matrix ที่ 4.3

ตารางที่ 4.3 แสดงตารางการพยากรณ์เทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ

		Prediction		Recall (%)
		Y	N	
Actual	Y	717	4	0.994
	N	8	325	0.976
Precision (%)		0.994	0.976	0.989

โดยที่

Y คือ นักศึกษาที่สำเร็จการศึกษาทำงานตรงตามสายอาชีพที่ได้ศึกษา

N คือ นักศึกษาที่สำเร็จการศึกษาทำงานไม่ตรงตามสายอาชีพที่ได้ศึกษา

จากตาราง Confusion Matrix ที่ตาราง 4.3 ผลการทำนายคลาสนักศึกษาที่สำเร็จการศึกษาทำงานตรงตามสายอาชีพที่ได้ศึกษา เป็น 717 และคลาสนักศึกษาที่สำเร็จการศึกษาทำงานไม่ตรงตามสายอาชีพที่ได้ศึกษาเป็น 325 รวมเป็น 1,051 จากจำนวนข้อมูลทั้งหมดให้ค่าความถูกต้องเป็น 98.86 %

บทที่ 5
สรุปผลและข้อเสนอแนะ

สรุปผลการวิจัย

งานวิจัยนี้เป็นการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อการเปรียบเทียบตัวแบบเทคนิคการพยากรณ์อาชีพของนักศึกษาที่สำเร็จระดับปริญญาตรีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยามซึ่งข้อมูลที่นำมาใช้ในงานวิจัยเป็นข้อมูลภาวะการมีงานทำจากสำนักทะเบียนมหาวิทยาลัยสยาม จากผลการวิจัย พบว่าค่าความแม่นยำในการจำแนกประเภทข้อมูลเฉลี่ยจาก 3 เทคนิค ตัวแบบการพยากรณ์อาชีพด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุดร้อยละ 98.86 และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิ้งให้ค่าความถูกต้องร้อยละ 98.86 และด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีนาอีฟเบย์ให้ค่าความถูกต้องร้อยละ 88.99 มีความคลาดเคลื่อนเฉลี่ยกำลังสองเท่ากับ 0.1, 0.1 และ 0.28 มีค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์เท่ากับ 0.02, 0.17 และ 0.02 จากผลการวิจัยสามารถสรุปได้ว่าตัวแบบเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจและวิธีเบ็กกิ้งเป็นตัวแบบที่เหมาะสมที่จะนำไปพัฒนาเป็นระบบการแนะแนวอาชีพให้กับนักศึกษาระดับปริญญาตรี คณะเทคโนโลยีสารสนเทศต่อไป

ข้อจำกัดของงานวิจัย

5.1.1 ข้อมูลที่นำมาทำการติดต่อเพื่อทำแบบสอบถามค่อนข้างเป็นข้อมูลเก่าเนื่องจากนักศึกษาบางท่านไม่สามารถติดต่อได้

5.1.2 สำหรับการทำแบบสอบถามนักศึกษาส่วนใหญ่ไม่สะดวกในการตอบแบบสอบถามเนื่องจากค่อนข้างไม่เข้าใจในกระบวนการ แต่สามารถตอบคำถามผ่านจากการโทรและพูดคุยเท่านั้น

ข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอตัวแบบเทคนิคการพยากรณ์เป็นเพียงการใช้เทคนิคส่วนหนึ่งของการทำเหมืองข้อมูลจากหลายเทคนิคที่มีในปัจจุบัน โดยนำข้อมูลภาวะการมีงานทำของคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม ซึ่งมีข้อมูลยังไม่มาก ในอนาคตหากมีการพัฒนางานวิจัยให้มีประสิทธิภาพ

มากขึ้นอาจจะต้องเพิ่มข้อมูลทางด้านสาขาวิชาอื่น ๆ เพื่อให้เกิดความหลากหลายของข้อมูลมากขึ้น และเพื่อให้ตัวแบบการพยากรณ์สามารถนำไปใช้ได้อย่างมีค่าความถูกต้องที่สุด

บรรณานุกรม

ชณิดาภา บุญประสม, และจรัญ แสนราช. (2561). การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล. *วารสารวิชาการครุศาสตร์อุตสาหกรรม,* 9(1), 142-151.

ชัชชฎา วันดี. (2556). *การศึกษาปัจจัยที่มีผลต่อการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษาโดยใช้เทคนิคเหมืองข้อมูล.* (ปริญญานิพนธ์บัณฑิต). มหาวิทยาลัยมหาสารคาม.

ญาใจ ลิ่มปิยกรณ์. (2553). การเพิ่มประสิทธิภาพความสามารถการเข้าถึงข้อมูลบนเว็บ สำหรับผู้พิการทางสายตา. *วารสารเทคโนโลยีสารสนเทศ, 10*(1), 31-42.

ทัศนีย์ เพียรทำดี. (2558). *การพยากรณ์คะแนนสอบมาตรฐานวิชาชีพ ของนักเรียนระดับประกาศนียบัตรวิชาชีพ ชั้นปีที่ 3 แผนกคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล.*

ประกรณ์ เกตุชาลี. (2562). *Confusion Matrix เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย ใน Machine learning.* https://medium.com/@pagongatchalee/confusion-matrix-machine-learning-fba6e3f9508c.

พฤฒิพงศ์ เพ็งศิริ. (2557). การพัฒนากระบวนความคิดกลุ่มเยาวชนผ่านนวัตกรรมการศึกษา ด้วยเกมคอมพิวเตอร์หมากรุกไทย. *วารสารรม่พฤกษ์, 32*(2), 126-134.

ภัทธิรา สุวรรณโค, นิศาชล จำนงศรี, จิติมนต์ อั่งสกุล. (2558). แบบจำลองการพยากรณ์ความเสี่ยงในการเกิดอุบัติเหตุทางถนนในเทศกาลปีใหม่ด้วยการทำเหมืองข้อมูล. *วารสารวิทยาการและเทคโนโลยีสารสนเทศ, 7*(2), 10-19.

ภัทร์พงศ์ พงศ์ภัทรกานต์. (2559). *การใช้เทคนิคเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา. PULINET Journal, 4*(2), 10-18.

วันวิสาข์ ชนะประเสริฐ. (2559). *การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพสำหรับนักศึกษาปริญญาตรี คณะโบราณคดี มหาวิทยาลัยศิลปากร.* มหาวิทยาลัยศิลปากร.

วีรศักดิ์ ฟองเงิน วรปภา อารีราษฎร์และเผด็จ พรหมสาขา ณ สกลนคร. (2017*).* การพยากรณ์ปริมาณน้ำในเขื่อนกิ่วลมโดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิทยาการจัดการสมัยใหม่ คณะวิทยาการจัดการมหาวิทยาลัยราชภัฏลำปาง, 10*(2), 121-131.

สมฤทัย กลัดแก้ว. (2557). *ระบบสนับสนุนการตัดสินใจการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิต.* (การศึกษาอิสระ ปริญญาวิทยาศาสมหาบัณฑิต). มหาวิทยาลัยรามคำแหง.

สายชล สินสมบูรณ์ทอง. (2558). การเปรียบเทียบประสิทธิภาพการแทนค่าข้อมูลสูญหายกับการจำแนกกลุ่ม 4 วิธี. *Thai Journal of Science and Technology, 9*(5), 586-588.

สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช. (2561). การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์ โดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิชาการการจัดการเทคโนโลยีสารสนเทศและนวัตกรรม, 5*(1), 164-171.

สุชาดา กีระนันทน์. (2545). *การอนุมานเชิงสถิติ : ทฤษฎีขั้นต้น*. มหาวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย.

สุรพงศ์ เอื้อวัฒนามงคล. (2557). *การทำเหมืองข้อมูล (Data Mining).* บางกอกบล๊อก.

โอม ศรนิล. (2556). การ*ออกแบบและพัฒนาคลังข้อมูล*. บางกอกบล๊อก.

David Hand Heikki Mannila and Padhraic Smyth. (2001). *Principles of Data Mining*. Retrieved from https://pzs.dstu.dp.ua/DataMining/bibl/MIT-PrinciplesofDataMining.pdf.

Roiger, R., and Geatz, M. (2003). *Data Mining*. Addison- Wesley, Boston.

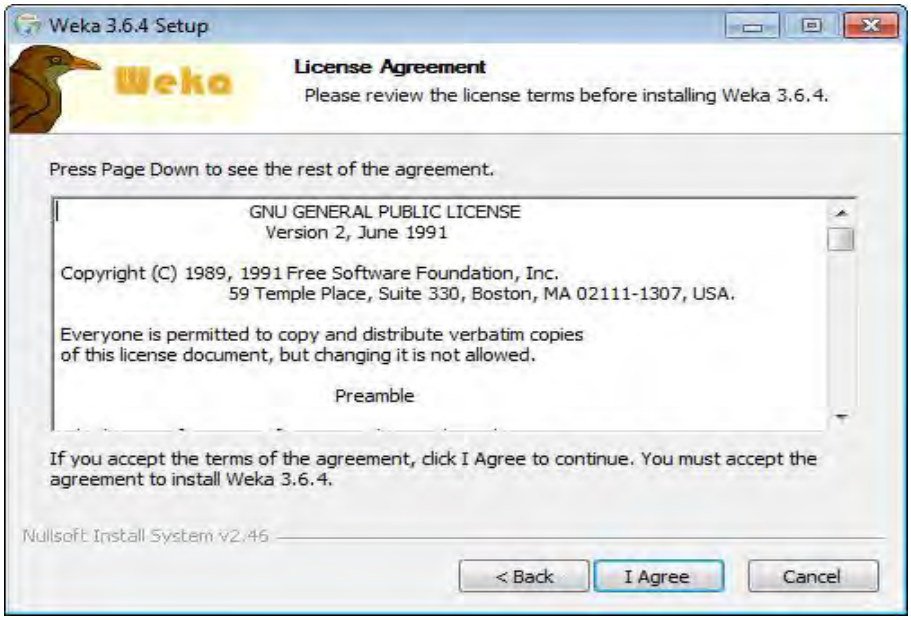
Saksit Srimarong. (2020). *4 ประเภทของการแบ่งกลุ่มข้อมูล (Clustering).* https://bdi.or.th/big-data-101/4-types-of-clustering/

ภาคผนวก ก

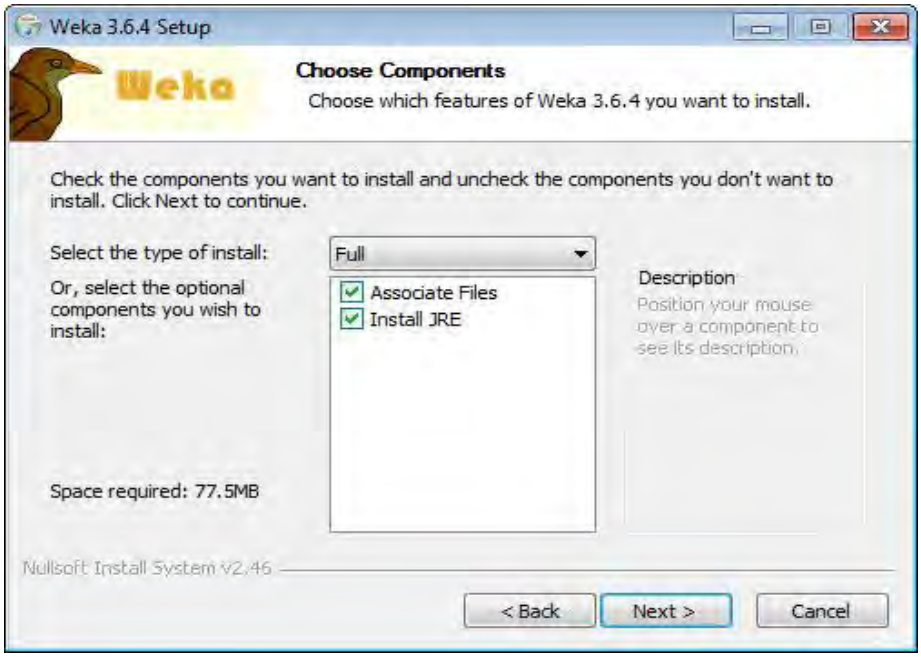
ขั้นตอนการลงโปรแกรม WEKA

1. ดับเบิลคลิก

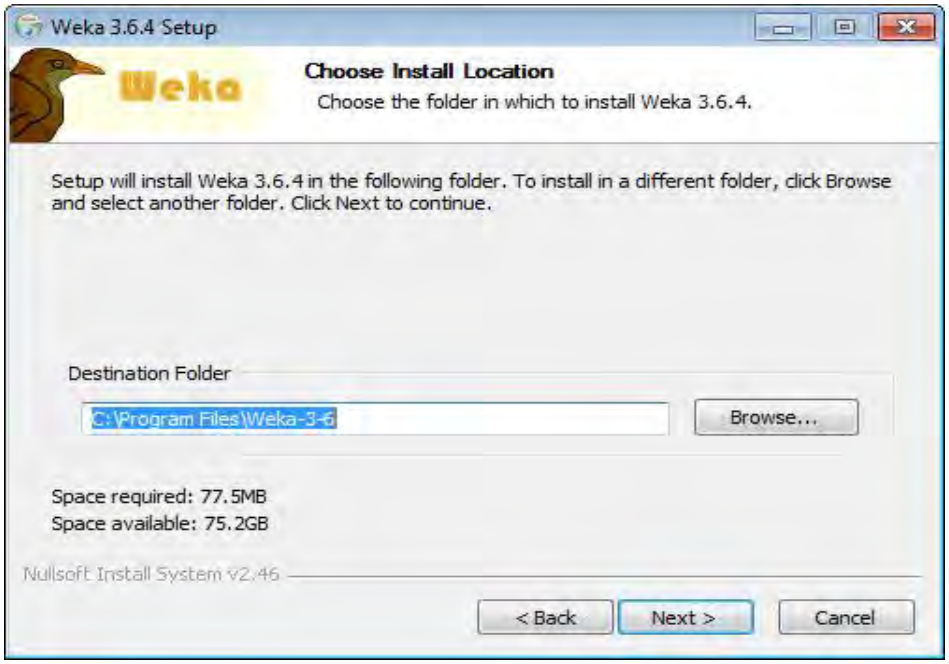
2. คลิก Next เพื่อทำการติดตั้ง



3. คลิก I Agree เพื่อทำการติดตั้ง

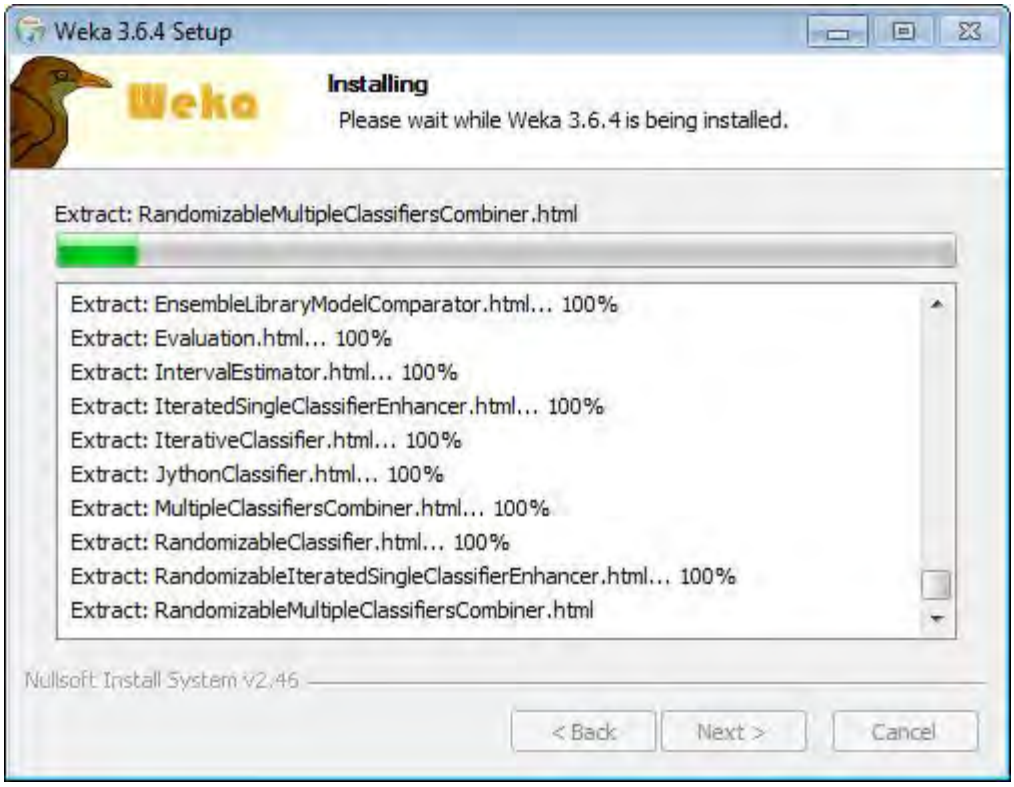


4. คลิก Next เพื่อทำการติดตั้ง

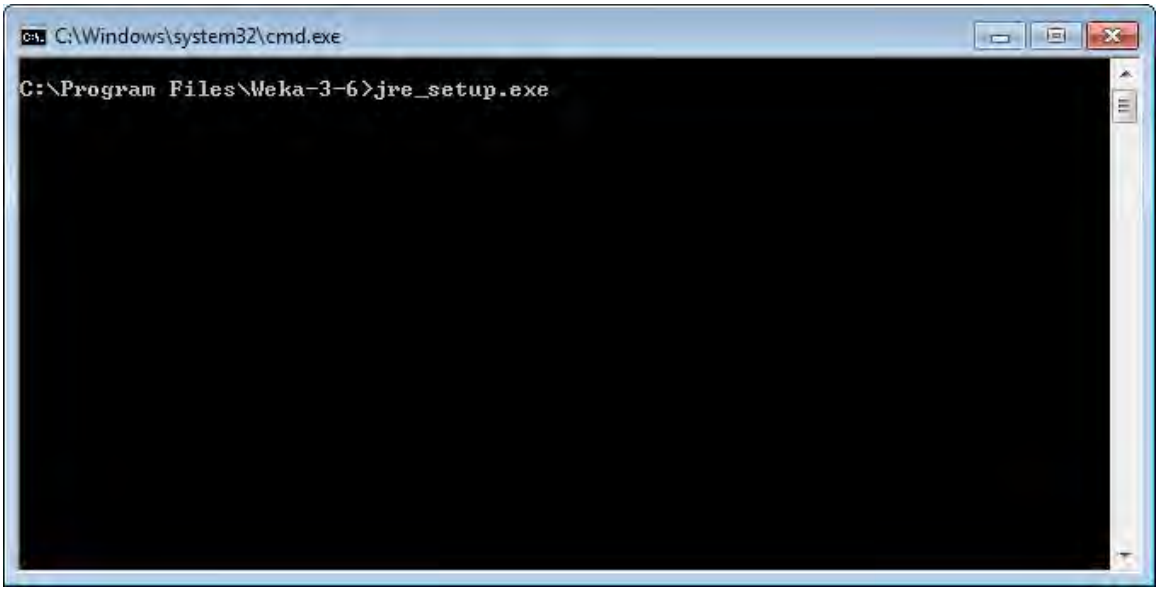


5. คลิก Next เพื่อทำการติดตั้ง

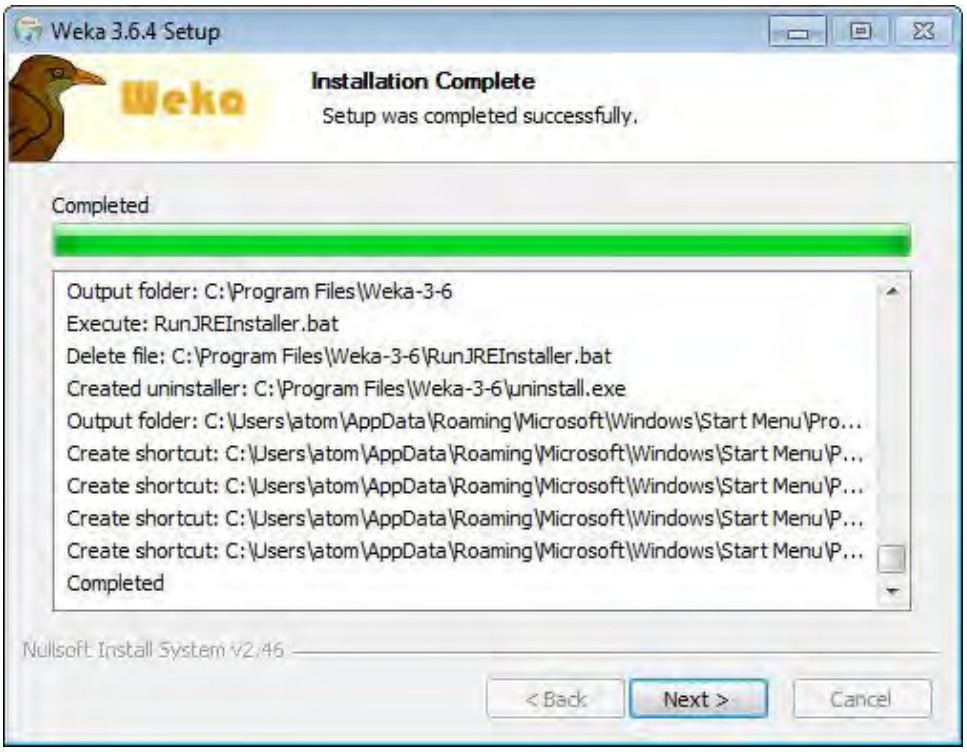
6. คลิก Install



ระหว่างประมวลผลจะขึ้นหน้าจอนี้ด้านล่างนี้



หน้าจอจะปิดไปเองเมื่อประมวลผลเสร็จ



7. คลิก Next


Weka 3.6.4 Setup
Completing the Weka 3.6.4 Setup Wizard
Weka 3.6.4 has been installed on your computer.
Click Finish to close this wizard.
Start Weka
< Back
Finish
Cancel

8. คลิก Finish

ภาคผนวก ข

ชุดข้อมูลรายการรายชื่อการทำวิจัย

ตารางที่ 3.1 รายละเอียดคุณลักษณะเพื่อนำมาสร้างตัวแบบ

ลำดับที่	ชื่อแอทริบิวต์	ค่าตัวแปร	คำอธิบาย
1	PREFIX_T	อักษร	เพศ
2	137-302	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์2
3	137-407	ระดับ	สัมมนาคอมพิวเตอร์ธุรกิจ
4	114-202	ระดับ	ภาษาอังกฤษธุรกิจ
5	130-202	ระดับ	การวิเคราะห์เชิงสถิติทางธุรกิจ
6	137-301	ระดับ	หลักการเขียนโปรแกรมคอมพิวเตอร์ 1
7	GPA	ระดับ	ผลสัมฤทธิ์การศึกษา
8	Target	อักษร	ความสอดคล้องทางอาชีพ

ตารางที่ 3.2 แสดงระเบียนข้อมูลนักศึกษาธุรกิจดิจิทัล ระหว่างปี พ.ศ. 2553-2561

ชนิดข้อมูล	จำนวนข้อมูลปี 2552 - 2560								
	2553	2554	2555	2556	2557	2558	2559	2560	2561
ข้อมูลนักศึกษา	156	140	124	154	145	117	69	61	89

ประวัติคณะผู้จัดทำ

รหัสนักศึกษา : 6105000007
ชื่อ-นามสกุล : นางสาว ทอแสง ใจงาม
คณะ : เทคโนโลยีสารสนเทศ
สาขาวิชา : ธุรกิจดิจิทัล
ที่อยู่ : 488/15 หมู่2 ต.อ้อมน้อย อ.กระทุ่มแบน จ.สมุทรสาคร 74130

รหัสนักศึกษา : 6105000008
ชื่อ-นามสกุล : นางสาว นภสร ใจทับทิม
คณะ : เทคโนโลยีสารสนเทศ
สาขาวิชา : ธุรกิจดิจิทัล
ที่อยู่ : 56/1 บ้านพักรถไฟธนบุรี ซ.6 แขวงศิริราช เขตบางกอกน้อย กรุงเทพมหานคร 10700