การใช้เทคนิคดาต้าไมน์นิงเพื่อการศึกษา
The Using Data Mining Techniques for Education

นคร ละลอกน้ำ*

บทคัดย่อ

ดาต้าไมน์นิง (Data mining) หรือการทำเหมืองข้อมูล เป็นการวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่ เพื่อให้ทราบความสัมพันธ์ รูปแบบของข้อมูลนั้นๆ ให้สามารถนำไปใช้ประโยชน์ในการตัดสินใจ วางแผน ทำนายแนวโน้มสิ่งที่จะเกิดขึ้นในอนาคต หรือแปรเปลี่ยนข้อมูลไปสู่ความรู้ใหม่ๆ การทำดาต้าไมน์นิงสามารถแบ่งประเภทข้อมูลได้ดังนี้ (1) การจัดกลุ่ม (Clustering) (2) การจัดความสัมพันธ์ (Association Rule) (3) การจัดจำแนก (Classification) และ (4) สมการถดถอย (Regression) กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านดาต้าไมน์นิง เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อว่า “CRISP-DM” มี 6 ขั้นตอน คือ (1) Business Understanding (2) Data Understanding (3) Data Preparation (4) Modeling (5) Evaluation และ (6) Deployment และทุกครั้งที่มีการสร้างโมเดลขึ้นมาจะต้องมีการทดสอบประสิทธิภาพของโมเดลที่สร้างได้ ซึ่งจะแยกเป็น 2 ส่วน คือ (1) วิธีการแบ่งข้อมูลเพื่อทำการทดสอบโมเดล (2) ตัวที่ใช้วัดประสิทธิภาพโมเดล เราสามารถประยุกต์ดาต้าไมน์นิ่งเพื่อการศึกษา 3 ด้าน หลัก คือ (1) งานด้านบริหารการศึกษา (2) งานด้านการเรียนการสอน และ (3) งานด้านบริการการศึกษา

คำสำคัญ : การทำเหมืองข้อมูล / การศึกษา

Abstract

Data mining or data mining. Analyze data from large databases. To know the relationship the format of the data. It can be used to make decisions, plan to predict trend in the future or change the information to new knowledge. Data mining can be classified as follows: (1) Clustering (2) Association Rule (3) Classification and (4) Regression The standardized process for data mining is called the "Cross-Industry Standard Process for Data Mining" or "CRISP-DM". There are 6 steps: (1) Business Understanding (2) Data Understanding (3) Data Preparation (4) Modeling (5) Evaluation and (6) Deployment. The performance of the model was built will be test for effective of model. It is divided into two parts: (1) the method of dividing the data to test the model (2) the model used to measure the model We can apply data mining for three main areas: (1) educational administration (2) instruction and 3) educational services

Keywords : Data Mining / Education

*อาจารย์ ดร. ภาควิชานวัตกรรมและเทคโนโลยีการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

บทนำ

จากพัฒนาการของเทคโนโลยียุคปัจจุบันที่มีการเปลี่ยนแปลงอย่างรวดเร็ว เทคโนโลยีเหล่านี้ไม่ว่าจะเป็นเทคโนโลยีคอมพิวเตอร์ เทคโนโลยีการสื่อสาร เทคโนโลยีสารสนเทศ ล้วนเข้ามาเป็นส่วนหนึ่งในการดำรงชีวิตของมนุษย์อย่างปฏิเสธไม่ได้ ทั้งในกิจวัตรประจำวัน การทำงาน การผักผ่อน เช่น การใช้โทรศัพท์หรือเครื่องมือสื่อสารในการเช็คข่าวสารที่เกิดขึ้นในแต่ละวัน การอัพเดทสถานะตนเองในแต่ละวันผ่านสื่อออนไลน์ การใช้ระบบเครือข่ายคอมพิวเตอร์สืบค้น บันทึกข้อมูลต่างๆ การจับจ่ายซื้อของอุปโภค บริโภค การช้อปปิ้งออนไลน์ที่มีบริการมากมาย เช่น Lazada, eBay, Shopee ฯลฯ ซึ่งข้อมูลจากพฤติกรรมการใช้เทคโนโลยีดังกล่าวจะถูกเก็บไว้ในรูปแบบดิจิทัล ทำให้สามารถเข้าถึงข้อมูลได้ง่าย เมื่อพิจารณาเทคโนโลยีที่ก้าวไกลในโลกออนไลน์ และพฤติกรรมการดำเนินชีวิตประจำวันทำให้เกิดข้อมูลปริมาณมากมาย มหาศาล เพราะมีข้อมูลเพิ่มขึ้นตลอดเวลาจากการใช้สื่อออนไลน์ของคนทั้งโลก จนมีผู้เชี่ยวชาญด้านข้อมูลให้นิยามศัพท์ข้อมูลมหาศาลอย่างนี้ว่าบิ๊กดาต้า (Big Data) ซึ่งมีลักษณะ 3 ประการ คือ (1) ปริมาณมาก (Volume) (2) ความหลากหลายของข้อมูล (Variety) และ (3) ความเร็วในการเปลี่ยนแปลงข้อมูล (Velocity) การเกิดข้อมูลจำนวนมหาศาลอย่างนี้ ได้มีการคิดหาวิธีการนำข้อมูลเหล่านี้มาใช้ให้ตรงกับความต้องการและเกิดประโยชน์สูงสุด วิธีที่นิยมใช้กันอีกวิธี คือ“การขุดเหมืองข้อมูล หรือ เทคนิคการใช้ดาต้าไมน์นิง (Data Mining)” ซึ่งปัจจุบันมีการนำวิธีการนี้ไปช่วยตัดสินใจในการดำเนินงาน ด้านต่างๆ มากมาย เพราะดาต้าไมน์นิง เป็นเทคโนโลยีสารสนเทศที่สามารถกลั่นกรอง วิเคราะห์ข้อมูลที่มีปริมาณมหาศาลเพื่อให้ได้ข้อมูลที่มีประโยชน์หรือได้ข้อมูลที่ซ่อนเร้นอยู่ในข้อมูลที่มีปริมาณมหาศาล เพื่อช่วยในการตัดสินใจ โดยเฉพาะด้านธุรกิจ เช่น การทำนายผลการตอบสนองกับการเปิดตัวสินค้าใหม่ การทำนายยอดขาย เมื่อมีการลดราคาสินค้า การทำนายกลุ่มลูกค้าที่น่าจะใช้สินค้าของเรา หรืองานการเงินการธนาคาร เช่น การคาดการณ์ถึงโอกาสในการชำระหนี้ของลูกค้าว่าสูงเท่าไหร่? การค้นหาลูกค้าขาดคุณภาพ เพื่อหลีกเลี่ยงความเสี่ยงในการปล่อยกู้ การค้นหาลูกค้าชั้นดีเพื่อเสนอการปล่อยกู้ การทำนายแนวโน้มของพฤติกรรมการใช้บัตรเครดิต จากความจำเป็นที่ต้องใช้ดาต้าไมน์นิงมาวิเคราะห์ข้อมูล ผู้เขียนจะขอเสนอบทความนี้ในประเด็น แนวคิดเกี่ยวกับดาต้าไมน์นิ่ง ขั้นตอนการทำดาต้าไมน์นิ่ง และในปัจจุบันควรนำเทคนิคการใช้ดาต้าไมน์นิงมาใช้เพื่อการศึกษาได้อย่างไรบ้าง เช่น ใช้ในการวิเคราะห์พฤติกรรมผู้เรียน การวางแผนการเรียน การพัฒนาหลักสูตร การแนะแนวการเรียนที่ตรงกับความสามารถของผู้เรียน การแนะนำการเรียนแบบเรียลไทม์ (real time) ผ่านบทเรียนแบบออนไลน์

ความหมายของดาต้าไมน์นิง

ดาต้าไมน์นิง หรือ การทำเหมืองข้อมูล ได้มีผู้เชี่ยวชาญให้ความหมายไว้อย่างหลากหลาย ดังนี้

ดาต้าไมน์นิง หมายถึง กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น เพื่อช่วยในการตัดสินใจ บอกแนวโน้มสิ่งที่จะเกิดขึ้นในอนาคต (ชนวัฒน์ ศรีสอ้าน, 2551)

ดาต้าไมน์นิง หมายถึง เครื่องมือในการค้นหาความรู้ เพื่อเพิ่มประสิทธิภาพในการทำงานด้านต่างๆ เช่น จะขายสินค้าให้ใครดี หรือจะลดการสูญเสียในการผลิตได้อย่างไร จะเทรดหุ้นตัวไหนในช่วงเวลานี้ ยาตัวนี้ได้ผลหรือไม่

ดาต้าไมน์นิง หมายถึง กระบวนการของการกลั่นกรองสารสนเทศที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อทำนายแนวโน้มและพฤติกรรม โดยอาศัยข้อมูลในอดีต และเพื่อใช้สารสนเทศเหล่านี้ในการสนับสนุนการตัดสินใจในการทำงานด้านต่างๆ (สายชล สินสมบูรณ์ทอง, 2558)

ดาต้าไมน์นิง หมายถึง ขบวนการทำงานที่เรียกว่า process ที่สกัดข้อมูลจากฐานข้อมูลขนาดใหญ่เพื่อให้ได้สารสนเทศที่เรายังไม่รู้โดยเป็นสารสนเทศที่มีเหตุผล และสามารถนำไปใช้ได้ ซึ่งเป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจ (ปณิธิ แก้วสวัสดิ์, 2553)

ดาต้าไมน์นิง หมายถึง การค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

สรุป ดาต้าไมน์นิง หรือ การทำเหมืองข้อมูล หมายถึง การวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่ เพื่อให้ทราบความสัมพันธ์ รูปแบบของข้อมูลนั้นๆ ให้สามารถนำไปใช้ประโยชน์ได้ เช่น การตัดสินใจ การวางแผน การทำนายแนวโน้มสิ่งที่จะเกิดขึ้นในอนาคต หรือเป็นการแปรเปลี่ยนข้อมูลไปสู่ความรู้ใหม่ๆ

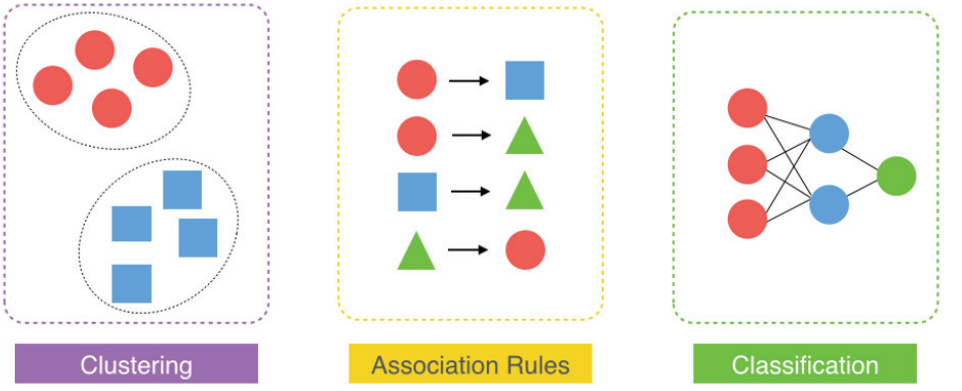
ความสำคัญของดาต้าไมน์นิง

ความสำคัญของดาต้าไมน์นิงสามารถนำมาช่วยในการวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่ ดังนี้

1. สกัด คัดเลือกข้อมูลมาใช้ในส่วนที่ตรงกับความต้องการจริงๆ
2. คาดการณ์สิ่งที่จะเกิดขึ้นล่วงหน้า เพื่อใช้ในการวางแผนพัฒนา
3. ดึงข้อมูลจากฐานข้อมูลขนาดใหญ่หลายๆ ฐานข้อมูลนำมาใช้ประโยชน์สูงสุด เพราะได้คำตอบจากข้อมูลที่ซับซ้อน ซ่อนเร้นที่อยู่ในข้อมูลขนาดใหญ่หลายๆ ฐาน
4. ลดความเสี่ยงในการทำงานขององค์กรต่างๆ เพราะมีฐานความรู้ที่ได้จากการทำดาต้าไมน์นิงมาใช้ในการสร้างกรอบการทำงาน วางแผนกลยุทธ์ทันกับการเปลี่ยนแปลงตรงกับพฤติกรรมผู้รับบริการ
5. วิเคราะห์ข้อมูลแบบเรียลไทม์ (real time) ให้ผลย้อนกลับทันที ปรับปรุงแก้ไขปัญหาได้ทันเหตุการณ์

เราใช้ดาต้าไมน์นิงทำอะไรได้บ้าง

ในทางปฏิบัติที่มีการนำดาต้าไมน์นิงมาใช้อย่างแพร่หลาย เป็นรูปธรรม จะเป็นการนำไปใช้แก้ไขปัญหาที่เกี่ยวข้องกับธุรกิจ ไม่ว่าจะเป็นการวิเคราะห์กลุ่มลูกค้า การเสนอหรือการให้คำแนะนำเพิ่มเติมแก่ลูกค้า การค้นหาความผิดปกติของชุดข้อมูลโดยเฉพาะที่เกี่ยวกับการเงิน หรือนำมาใช้แก้ปัญหาด้านการศึกษา ไม่ว่าจะเป็นการวิเคราะห์พฤติกรรมการเรียน การเสนอหรือการให้คำแนะนำเพิ่มเติมแก่ผู้เรียนเป็นรายบุคคล การเลือกเรียนในสาขาที่เหมาะสมกับความสามารถของตนเอง การพยากรณ์หรือคาดคะเนแนวโน้มที่จะเกิดขึ้น เช่น จะสอบผ่าน/ไม่ผ่านในรายวิชาที่ลงทะเบียนเรียน ในการทำดาต้าไมน์นิ่งต้องมีการจัดกลุ่มประเภทลักษณะของข้อมูลให้ชัดเจน ซึ่งสามารถแบ่งประเภทข้อมูลได้หลายแบบ ดังภาพประกอบที่ 1



ภาพประกอบที่ 1 การจำแนกแบ่งประเภทข้อมูลในการทำดาต้าไมน์นิง
ที่มา : https://dataminingtrend.com

การจัดกลุ่มประเภทลักษณะของข้อมูลหลักๆ มีรายละเอียดสรุปได้ ดังนี้

1. การจัดกลุ่ม (Clustering)

เป็นการจัดข้อมูลออกเป็นกลุ่มย่อยตามลักษณะความคล้ายคลึงของตัวข้อมูลเอง โดยต้องการให้ข้อมูลที่อยู่ในกลุ่มเดียวกันมีความคล้ายคลึงกันมากที่สุด หรือข้อมูลที่อยู่ต่างกลุ่มกันมีความต่างกันมากที่สุด หรือ “การจัดกลุ่มในพวกเดียวกัน คล้ายคลึงกัน” เช่น การแบ่งสินค้าตามรูปร่าง แบ่งสินค้าตามราคา หรือแบ่งสินค้าตามชนิดการใช้งาน สิ่งที่เห็นจากการใช้ข้อมูลประเภทนี้ในทางการตลาดที่แพร่หลาย เช่น การแบ่งกลุ่มลูกค้าที่มีลักษณะความต้องการที่คล้ายคลึงกัน แล้วจัดโปรโมชั่นให้ตรงกับความต้องการของกลุ่มนั้นๆ ได้ตรงใจกลุ่มเป้าหมายแต่ละกลุ่ม ข้อมูลประเภทนี้เป็นการวิเคราะห์ดาต้าไมน์นิงโดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) ซึ่งเป็นเทคนิคที่เน้นพิจารณาข้อมูลเป็นหลัก เพื่อหาความสัมพันธ์ของข้อมูลหรือแบ่งกลุ่มข้อมูลนั้นๆ โดยใช้อัลกอริทึม (เป็นกระบวนการแก้ปัญหาที่สามารถอธิบายออกมาเป็นขั้นตอนที่ชัดเจน โดยวิธีการในการอธิบาย Algorithm ได้แก่ 1. อธิบายแบบใช้ภาษาที่เราสื่อสารกันทั่วไป (Natural Language) 2. อธิบายด้วยรหัสจำลองหรือรหัสเทียม (Pseudocode) และ 3. อธิบายด้วยแผนผัง (Flowchart)) ที่มีหลายรูปแบบมาช่วยในการการวิเคราะห์ เช่น K-means และ Hierarchical ; Agglomerative Clustering ; Density base ตัวอย่างที่ได้จากการวิเคราะห์ประเภทนี้ คือ (1) การจัดกลุ่มลูกค้า (Segmentation) (2) การจัดกลุ่มดอกไม้ (3) การจัดกลุ่มของพื้นที่จากข้อมูลภาพถ่ายดาวเทียม ฯลฯ

2. การจัดความสัมพันธ์ (Association Rule)

เป็นการเก็บข้อมูลพฤติกรรมการรับบริการด้านต่างๆ อาจเป็นพฤติกรรมการเรียน พฤติกรรมการซื้อสินค้า หรือพฤติกรรมการท่องเที่ยว ฯลฯ เพื่อนำมาวิเคราะห์หากฎความสัมพันธ์ (Association Rule) หรือ “เป็นการหาความสัมพันธ์ของข้อมูลที่เกิดร่วมกัน” เช่น ในทางธุรกิจจะใช้ค้นหาสินค้าที่มีการซื้อร่วมกันบ่อย ๆ แล้วนำเสนอให้ลูกค้ารายใหม่ เช่น เมื่อลูกค้าซื้อน้ำพริกจะซื้อผักด้วย หรือ เมื่อลูกค้าซื้อของแต่งบ้านจะซื้อเบียร์ด้วย ข้อมูลประเภทนี้เป็นการวิเคราะห์ดาต้าไมน์นิงโดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน โดยอัลกอริทึมที่นิยมใช้ในการวิเคราะห์ คือ Apriori algorithm และ Frequent Pattern Growth (FP) ตัวอย่างที่ได้จากการวิเคราะห์ประเภทนี้ คือ การค้นหาพฤติกรรมการซื้อสินค้าข้ามสายผลิตภัณฑ์จากข้อมูลการซื้อในอดีตเพื่อแนะนำให้ลูกค้าใหม่ได้ทราบเป็นทางเลือก

3. การจัดจำแนก (Classification)

เป็นการนำข้อมูลที่มีในอดีตมาสอนระบบเพื่อให้เรียนรู้รูปแบบที่เกิดขึ้นในข้อมูล จากนั้นนำมาสร้างเป็นสมการหรือโมเดลขึ้นมา เพื่อหาคำตอบให้สำหรับข้อมูลใหม่ หรือ “เป็นโมเดลที่ใช้สำหรับนำข้อมูลที่มีอยู่มาทำนายอนาคต” เช่น การจำแนกอีเมล์ออกเป็นแบบสแปมหรือแบบปกติซึ่งต้องมีการนำข้อมูลของอีเมล์แบบสแปมกับแบบปกติมาให้คอมพิวเตอร์ทำการเรียนรู้เสียก่อน หลังจากนั้นจึงสร้างโมเดลการจำแนกประเภทของอีเมล์และใช้การจำแนกอีเมลที่เข้ามาใหม่ว่าเป็นแบบสแปมหรือแบบปกติ (คำตอบจากการวิเคราะห์จะเรียกว่า “คลาส” หรือ “ลาเบล”) ถือได้ว่าข้อมูลประเภทนี้เป็นการวิเคราะห์ดาต้าไมน์นิ่งโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning) ซึ่งเน้นการเรียนรู้จากข้อมูลที่เก็บมาจากอดีตเพื่อนำมาสร้างโมเดลสำหรับทำนายหรือคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคตคาดว่าจะเกิดขึ้น โดยอัลกอริทึมที่นิยมใช้ในการวิเคราะห์ เช่น decision trees, naïve Bayesian, artificial neural networks, และ support vector machines. ตัวอย่างที่ได้จากการวิเคราะห์ประเภทนี้ เช่น (1) ให้ผู้เลือกตั้ง Vote เลือก ส.ส. แล้วสร้างโมเดลทำนาย (2) ให้ลูกค้าตอบแบบสอบถาม แล้วสร้างโมเดลทำนายประเภทลูกค้า (3) การจัดหมวดหมู่ของผู้ยื่นขอเครดิต (Credits) เป็นระดับต่ำ ระดับกลาง และระดับสูง ของความเสี่ยงที่จะได้รับจากการอนุมัติบัตรเครดิต ฯลฯ

4. สมการถดถอย (Regression) หรือ การประมาณค่าข้อมูล

จะมีลักษณะคล้ายกับการจัดจำแนก (Classification) แต่มีข้อแตกต่างกันที่คำตอบที่ต้องการทำนาย ซึ่ง Classification จะทำนายข้อมูลที่เป็นลักษณะ มาตรวัดนามบัญญัติ (Nominal Scale) คือ เป็นแค่การกำหนด

สัญลักษณ์หรือตัวเลขขึ้นมาเพื่อจำแนกประเภทสิ่งของหรือคุณลักษณะต่างๆ ออกเป็นกลุ่ม แต่จะไม่ได้แสดงถึงปริมาณ (มากหรือน้อย) หรือความสูง-ต่ำ ซึ่งไม่สามารถจัดลำดับก่อน-หลังได้ เช่น เพศ ก็แยกได้แค่เพศชายกับ เพศหญิงเท่านั้น ส่วนการประมาณค่าข้อมูล หรือ Regression จะใช้กับคำตอบที่เป็นเชิงปริมาณ หรือ จำนวนตัวเลขเป็นหลัก เช่น คาดการณ์ว่าพรุ่งนี้จะมีอุณหภูมิเท่าไหร่ ก็ต้องไปเก็บข้อมูลก่อนหน้านี้ ซึ่งจะสนใจเก็บข้อมูล ที่เป็นตัวเลข ข้อมูลประเภทนี้เป็นการวิเคราะห์ดาต้าไมน์นิงโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน โดยอัลกอริทึมที่นิยมใช้ในการวิเคราะห์ คือ สมการถดถอยเชิงเส้น (Multiple Linear Regression) และการถดถอยแบบโลจิสติกส์ (Logistic Regression) ตัวอย่างที่ได้จากการวิเคราะห์ประเภทนี้ เช่น ทำนายยอดขายสำหรับปีต่อๆ ปี โดยใช้จำนวนลูกค้าและงบการตลาดเป็นตัวแปรอิสระ หรือ ทำนายรายได้รวมต่อครอบครัว

ตัวอย่างการนำดาต้าไมน์นิงประยุกต์ใช้ในงานด้านต่างๆ

- ธุรกิจค้าปลีกใช้ในการพิจารณาหากลยุทธ์ให้เป็นที่สนใจกับผู้บริโภคในรูปแบบต่าง ๆ เช่น ที่ว่างในชั้นวางของจะจัดการอย่างไรถึงจะเพิ่มยอดขายได้ เช่นที่ Midas ซึ่งเป็นผู้แทนจำหน่ายอะไหล่สำหรับอุตสาหกรรมรถยนต์ งานที่ต้องทำคือการจัดการกับข้อมูลที่ได้รับจากสาขาทั้งหมด ซึ่งจะต้องทำการรวบรวมและวิเคราะห์อย่างทันท่วงที

- กิจการโทรคมนาคม เช่นที่ Bouygues Telecom ได้นำมาใช้ตรวจสอบการโกงโดยวิเคราะห์รูปแบบการใช้งานของสมาชิกลูกข่ายในการใช้งานโทรศัพท์ เช่น คาบเวลาที่ใช้จุดหมายปลายทาง ความถี่ที่ใช้ ฯลฯ หรือคาดการณ์ข้อบกพร่องที่เป็นไปได้ในการชำระเงิน

- การวิเคราะห์ผลิตภัณฑ์ เก็บรวบรวมลักษณะและราคาของผลิตภัณฑ์ทั้งหมดสร้างโมเดล และใช้โมเดลที่ได้ในการทำนายราคาผลิตภัณฑ์ตัวอื่น ๆ

- การวิเคราะห์การตัดสินใจในการที่จะให้เครดิตการ์ดกับลูกค้าหรือไม่ หรือ แบ่งประเภทของลูกค้าว่ามีความเสี่ยงในเรื่องเครดิต ต่ำ ปานกลาง หรือสูง หรือป้องกันปัญหาเรื่องการทุจริตบัตรเครดิต

- การวิเคราะห์ลูกค้า เช่น ช่วยแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะผลิตและเสนอสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม ทำนายว่าลูกค้าคนใดจะเลิกใช้บริการจากบริษัทภายในอนาคต

- การวิเคราะห์การขาย เช่น ช่วยในการโฆษณาสินค้าได้อย่างเหมาะสมและตรงตามเป้าหมาย ช่วยในการจัดวางสินค้าได้อย่างเหมาะสม

- การวิเคราะห์พฤติกรรมของลูกค้าในการใช้เว็บไซต์ใดตามลำดับก่อนหลัง เพื่อนำข้อมูลพิจารณาว่าส่วนใดของเว็บไซต์ที่ควรปรับปรุงหรือควรเรียงลำดับการเชื่อมโยงในแต่ละหน้าอย่างไรเพื่อให้สะดวกกับผู้เข้าเยี่ยมชมมากที่สุด

- ด้านการแพทย์ : นำมาช่วยวิเคราะห์อาการของคนไข้, วิเคราะห์การจ่ายยา , พยากรณ์แนวโน้มการเกิดโรคระบาด

- ด้านเกษตรกรรม : นำมาช่วยวิเคราะห์และพยากรณ์ราคาสินค้า, ทำนายมูลค่าการส่งออกสินค้า ฯลฯ

- ด้านแรงงาน : นำมาวิเคราะห์ความต้องการตลาดแรงงานในแต่ละพื้นที่ แต่ละอุตสาหกรรม เพื่อวางแผนพัฒนาฝีมือแรงงานหรือจัดหาให้ตรงกับความต้องการและทันกับการเปลี่ยนแปลง

อัลกอริทึมที่นิยมใช้ในการวิเคราะห์ข้อมูลเพื่อการศึกษา

อัลกอริทึมที่นำมาช่วยในการวิเคราะห์ข้อมูลทางการศึกษามีหลายเทคนิคสรุปไว้ ดังนี้ (จิราภา เลาหะวรนันท์และคณะ, 2558)

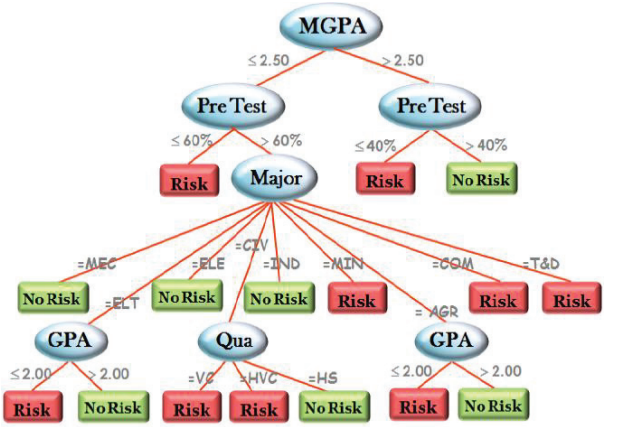
1. ต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคที่นิยมใช้ในการจัดหมวดหมู่ข้อมูลมักใช้ในการตรวจสอบข้อมูลเพื่อพยากรณ์ เทคนิคนี้จะมีลักษณะคล้ายโครงสร้างต้นไม้ โดยการ แตกแขนงไปตามเงื่อนไขหรือเส้นทางของกิ่งไม้และข้อมูลที่คาดคะเน ไว้ว่าจะเกิดขึ้น ซึ่งจะใช้กฎในรูปแบบ “ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์)” (If-then Rule) มา

ประกอบการสร้างโครงสร้างต้นไม้ตัดสินใจ สำหรับโครงสร้างต้นไม้ตัดสินใจประกอบด้วย ดังนี้

• โหนดภายใน (Internal Node) คือโหนดที่แสดงถึงคุณลักษณะ (Feature) ที่นำมาใช้ในการแบ่งกลุ่มของข้อมูลซึ่งมีโหนดราก (Root Node) อยู่บนสุดของโครงสร้าง ซึ่งเป็นโหนดที่มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุด

• กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนดที่ใช้เป็นเงื่อนไขหรือทางเลือกของการกระทำ ซึ่งมาจากผลลัพธ์แต่ละตัวของทุกตัวทำนาย (Predictor) หรือคุณสมบัติ (Feature)

• โหนดใบ (Leaf Node) เป็นโหนดที่แสดงผลลัพธ์ของเงื่อนไข หรือการกระทำตามเงื่อนไขที่เกิดขึ้น



ภาพประกอบที่ 2 การใช้ต้นไม้ตัดสินใจในการจำแนกนักศึกษาออกเป็นกลุ่มเสี่ยง/ไม่เสี่ยง
ที่มา : สุวิมล สิทธิชาติ (2560)

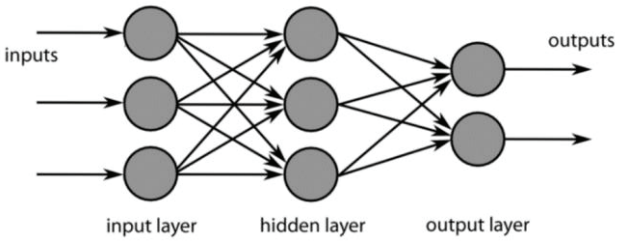
สำหรับการสร้าง Decision Tree เริ่มต้นที่การคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้ (Root Node) หลังจากนั้นจะทำการแตกกิ่งแอตทริบิวต์ออกไปเรื่อยๆ จนสามารถแบ่งข้อมูลออกเป็นคลาสได้ชัดเจน

2. นาอีฟเบย์ (Naïve Bayes) เป็นเทคนิคการทำเหมืองข้อมูลอย่างง่าย โดยนำโมเดลมาใช้ในการคิด แยกประเภทข้อมูลผ่านหลักความน่าจะเป็นที่อยู่บนพื้นฐานของทฤษฎี Bayes และสมมติฐานของการเกิดของเหตุการณ์เป็นอิสระต่อกัน เทคนิคนี้จะไม่มีการหมุนวนที่ซับซ้อนส่งผลให้สามารถทำงานได้ดีและมีประโยชน์กับชุดข้อมูลขนาดใหญ่อย่างมาก ทำให้เทคนิคนี้ถูกใช้อย่างแพร่หลาย เทคนิคนี้จะคำนวณจากทฤษฎีโดยสันนิษฐานว่าผลลัพธ์หรือค่าที่เกิดจากตัวที่ใช้ทำนาย (predictor) เป็นอิสระต่อกันโดยเขียนเป็นสมการดังนี้

$$P(c \mid x) = P(x \mid c)\, P(c) / P(x)$$

โดย $P(c \mid x)$ คือค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c; $P(x \mid c)$ คือ ค่าความน่าจะเป็นที่ข้อมูลในชุดข้อมูลสอนที่มีคลาส c และมีแอตทริบิวต์ x โดยที่ $x = x_1 \cap x_2 \ldots \cap x_M$ โดยที่ M คือ จำนวนแอตทริบิวต์; P(c) คือ ค่าความน่าจะเป็นของคลาส C; และ P(x) คือ ค่าความน่าจะเป็นของแอตทริบิวต์ x

3. โครงข่ายประสาท (Neural Network) เป็นการใช้โมเดลทางคณิตศาสตร์มาประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) โดยได้แนวคิดมาจากการจำลองการทำงานของเซลล์สมองมนุษย์ที่แต่ละเซลล์ประสาทจะประกอบไปด้วยเดนไดรท์ (Dendrite) หรือปลายในรับกระแสประสาท ซึ่งเป็นตัว input ของ เซลล์ และแอกซอน (Axon) เป็นเสมือน output ของเซลล์ที่จะส่งกระแสประสาทไปยังเซลล์ตัวอื่น ถ้าสมองได้รับกระแสไฟฟ้ามากพอ จะทำให้เซลล์ส่งต่อกระแสประสาทไปเรื่อยๆ

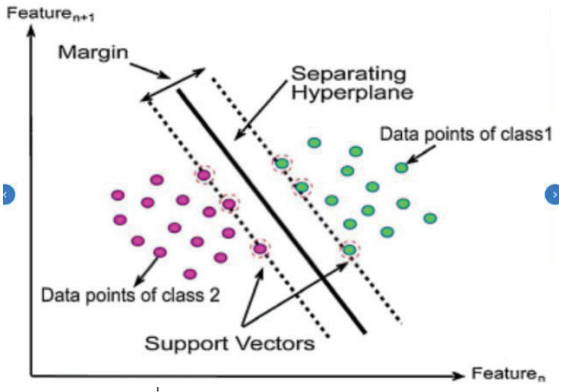


ภาพประกอบที่ 3 โครงสร้าง Layer ของโครงข่ายประสาทเทียมแบบหลายชั้น

ที่มา :https://commons.wikimedia.org/wiki/File:MultiLayerNeuralNetworkBigger_english.png

สำหรับโครงสร้างของประสาทเทียมจะประกอบด้วย input และ output เช่นกัน โดยแบ่งเป็นชั้นหรือ layer ซึ่งจะมีชั้นคั่นตรงกลางคือ hidden layer โดยโครงสร้างประสาทเทียมจะมีหน่วยย่อย เรียกว่า perceptron ซึ่งเทียบเท่าได้กับเซลล์สมองของมนุษย์หนึ่งเซลล์

4. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) เป็นอัลกอริทึมที่นำมาใช้ในการวิเคราะห์และจำแนกข้อมูลอย่างกว้างขวางโดยอาศัยโมเดลทางคณิตศาสตร์ในเรื่องของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นเลือกเส้นที่แบ่งข้อมูลได้ดีที่สุด โดยไม่จำเป็นต้องเป็นเส้นตรงเสมอไป

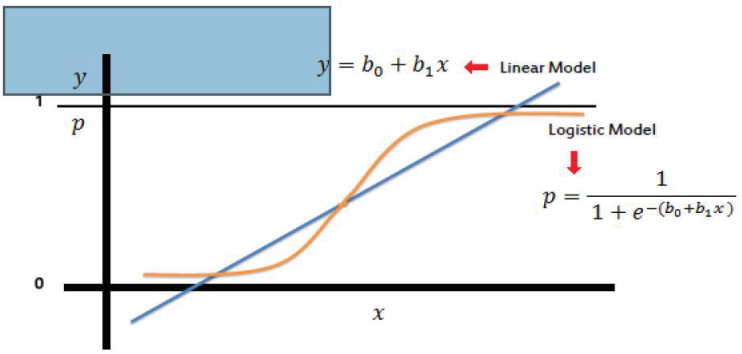


ภาพประกอบที่ 4 โครงสร้างซัพพอร์ทเวกเตอร์แมชชีน

ที่มา : https://www.researchgate.net/figure/Linear-support-vector-machine-example-modified-from-Burges-1998_fig3_303095519

หลักของเทคนิคนี้คือ การนำข้อมูลที่ใช้สอนมากระจายเป็นเวกเตอร์ในระนาบ หรือ สเปซ N มิติ (Feature Space) จากนั้นคำนวณหาเส้นไฮเปอร์เพลน (Hyperplane) ที่จะแยกกลุ่มของเวกเตอร์อินพุตออกเป็นประเภทต่างๆ โดยเส้นไฮเปอร์เพลนต้องเป็นเส้นที่มีค่า margin หรือระยะห่างระหว่างจุดของข้อมูลที่อยู่ใกล้ กับไฮเปอร์เพลนทั้งสองด้าน คือ d+ และ d- มากที่สุด เพราะถ้าค่า margin น้อย แสดงว่าอาจมีจุดหนึ่งของไฮเปอร์เพลนที่อยู่ใกล้ข้อมูลของแต่ละคลาสมากเกินไป ทำให้ข้อมูลใหม่ที่อยู่ห่างออกไปเล็กน้อยเกิดการทำนายผิดพลาดได้ โดยข้อมูลที่อยู่บนขอบของ margin เรียกว่า support vector

5. การถดถอยโลจิสติกส์ (Logistic Regression) เป็นการพยากรณ์ความน่าจะเป็นของผลลัพธ์ที่จะเกิดขึ้น ซึ่งสามารถเป็นได้เพียง 2 ค่า เช่น ใช่ หรือไม่ใช่ โดยการพยากรณ์จะขึ้นกับตัวแปรที่ส่งผลตอเหตการณ์นั้นๆ ซึ่งอาจจะมีเพียงหนึ่งตัวหรือมากกว่า



ภาพประกอบที่ 5 โครงสร้างการถดถอยโลจิสติกส์
ที่มา : จิราภา เลาหะวรนันท์ และคณะ, (2558)

การถดถอยโลจิสติกส์ (Logistic Regression) สามารถสร้าง curve จากการใช้ ลอการิทึม (logarithm) “odds” ในตัวแปรเป้าหมาย ซึ่งจำกัดค่า 0 ถึง 1

ขั้นตอนการทำดาต้าไมน์นิง

กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านดาต้าไมน์นิงได้พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือกันของ 3 บริษัท คือ DaimlerChrysler SPSS และNCR กระบวนการทำงานนี้เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อว่า “CRISP-DM” ดังภาพประกอบที่ 6

ดังภาพประกอบที่ 6 แสดงขั้นตอนการทำดาต้าไมน์นิง
ที่มา : https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

แต่ละขั้นตอนของ CRISP-DM มีรายละเอียด ดังนี้ (C. Shearer, 2000)

(1) Business Understanding เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางดาต้าไมน์นิง พร้อมทั้งวางแผนในการดำเนินการคร่าวๆ ตัวอย่างเช่น การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ หรือการใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนสารนเทศ การใช้เทคนิคดาต้าไมน์นิงเพื่อ

พัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์ หรือ การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพัฒนาโปรแกรมการเรียนรู้เฉพาะบุคคล สำหรับเด็กที่มีความต้องการพิเศษระดับปฐมวัย

(2) Data Understanding ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูลที่เกี่ยวข้องกับขั้นที่ 1 แล้วตรวจสอบความถูกต้องและความน่าเชื่อถือของข้อมูล และพิจารณาปริมาณข้อมูลที่เพียงพอในการนำไปวิเคราะห์

(3) Data Preparation เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมาอาจอยู่ในรูปแบบข้อความ ตัวอักษร ภาพ เสียง ฯลฯ ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ (แปลงให้ถูกต้อง เป็นระเบียบ จัดกลุ่ม หมวดหมู่ให้ง่ายในการวิเคราะห์ ใส่รหัสเพื่อให้โปรแกรมคอมพิวเตอร์มองออก) โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป

ขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM จึงมีการพัฒนาเครื่องมือหรือซอฟต์แวร์สำหรับงานนี้ขึ้นมาใช้โดยเฉพาะ เช่น RapidMiner, Weka, R, SPSS เป็นต้น

(4) Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมน์นิง ซึ่งต้องเลือกเทคนิคการวิเคราะห์ที่เหมาะสมกับข้อมูลที่ออกแบบไว้ในขั้นตอนที่ 3 โดยเทคนิคการวิเคราะห์ข้อมูลจะแบ่งหลักๆ คือ 1) การจัดกลุ่ม (Clustering) จะนิยมใช้ K-means, Agglomerative Clustering, Density base 2) การจัดความสัมพันธ์ (Association Rule) จะนิยมใช้ Apriori algorithm และ 3) การจัดจำแนก (Classification) จะนิยมใช้ decision trees, naïve Bayesian, neural networks, และ support vector machines.

(5) Evaluation เป็นการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรกหรือ มีความน่าเชื่อถือมากน้อยเพียงใดก่อนนำไปใช้จริง กรณีที่มีการสร้างโมเดลด้วยเทคนิค Classification การวัดประสิทธิภาพของโมเดลจำเป็นต้องแบ่งข้อมูลออกเป็น 2 ส่วน โดยส่วนที่ 1 ใช้เพื่อสร้างโมเดล หรือ เรียกว่า ส่วน Training data ส่วนที่ 2 ให้โมเดลทำนายคลาสออกมา หรือ เรียกว่าส่วน Testing data การแบ่งข้อมูลเพื่อทำการทดสอบนี้มี 3 วิธีการใหญ่ๆ คือ 1) วิธี Self-Consistency Test 2) วิธี Split Test 3) วิธี Cross-validation Test

(6) Deployment เป็นการนำองค์ความรู้จากการทำดาต้าไมน์นิงไปใช้จริงในองค์กรหรือบริษัท เช่น การสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่นที่ตรงกับกลุ่มเป้าหมาย หรือการทำโฆษณาใน google หรือ amazon, lazada, youtube ระบบเล่นอัตโนมัติ เป็นต้น

การทดสอบประสิทธิภาพของโมเดลจากการทำดาต้าไมน์นิ่ง

การสร้างโมเดลขึ้นทุกครั้งต้องมีการทดสอบประสิทธิภาพของโมเดลที่สร้างได้ ซึ่งจะแบ่งเป็น 2 ส่วน คือ 1) วิธีการแบ่งข้อมูลเพื่อทำการทดสอบโมเดล 2) ตัวที่ใช้วัดประสิทธิภาพโมเดล (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

1. วิธีการแบ่งข้อมูลเพื่อทำการทดสอบโมเดล สามารถแบ่งข้อมูลเพื่อทำการทดสอบนี้มี 3 วิธีการ ดังนี้

(1) วิธี Self Consistency Test หรือบางครั้งเรียกว่า Use Training Set เป็นวิธีการที่ง่ายที่สุด นั่นคือข้อมูลที่ใช้ในการสร้างโมเดล และข้อมูลที่ใช้ในการทดสอบโมเดลเป็นข้อมูลชุดเดียวกัน การวัดประสิทธิภาพด้วยวิธีนี้จะให้ผลการวัดประสิทธิภาพที่มีค่าสูงมาก (อาจจะเข้าใกล้ 100%) เนื่องจากเป็นข้อมูลชุดเดิมที่ระบบได้ทำการเรียนรู้มาแล้ว แต่ผลการวัดที่ได้ไม่เหมาะที่จะนำไปรายงานในงานวิจัยต่างๆ ซึ่งวิธีการนี้เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพเพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น

(2) วิธี Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% หรือ 80%) ใช้ในการสร้างโมเดลและข้อมูลส่วนที่สอง (30% หรือ 20%) ใช้ใน การทดสอบประสิทธิภาพของโมเดล ข้อดีของวิธีการนี้คือใช้เวลาในการสร้างโมเดลน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่

(3) วิธี Cross-validation Test เป็นวิธีที่นิยมในการทำงานวิจัย เพื่อใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ โดยวิธีนี้นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลัง จากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้

2. ตัวที่ใช้วัดประสิทธิภาพของโมเดล โดยทั่วไปแล้วจะมีตัววัดประสิทธิภาพที่นิยมใช้ คือ

- Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส
- Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส
- F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส
- Accuracy เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส

วิธีการคำนวณ Precision, Recall, F-measure, Accuracy มีรายละเอียด ดังนี้

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

$$\text{F–measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

$$\text{Accuracy} = \text{จำนวน True Positive ของทุกคลาสรวมกัน โดยสมการในการทดสอบ คือ}$$

$$= \frac{\text{True Positive + True Negative}}{\text{True Positive + True Negative + False Positive + False Negative}}$$

เพื่อให้เห็นภาพตัววัดประสิทธิภาพของโมเดลแต่ละตัว แสดงดังตัวอย่างต่อไปนี้ (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

ตารางที่ 1 ข้อมูลสภาพอากาศย้อนหลัง 10

No.	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	mild	normal	false	Yes
10	rainy	mild	normal	false	yes

วัน

จากตารางที่ 1 มีคลาสคำตอบอยู่ 2 ค่า คือ yes และ no นำมาสร้างตาราง confusion matrix ได้เป็นตารางขนาด 2x2 ดังในตารางที่ 2 โดยข้อมูลด้านคอลัมน์ คือ คลาสที่อยู่ในข้อมูลเทรนนิงดาต้า (actual) และข้อมูลในแนวแถว คือ คลาสที่โมเดลทำนายมาได้ (predicted)

ตารางที่ 2 แสดงตาราง confusion matrix ของข้อมูล weather ซึ่งมี 2 คลาส

คลาสที่โมเดลทำนายมาได้ (predicted)	ข้อมูลเทรนนิงดาต้า (actual)	
	yes	**no**
yes	TP	FP
no	FN	TN

จากในตารางที่ 2 ค่าที่แสดงในช่องต่างๆ ของตาราง มีความหมาย คือ

- True Positive (TP) คือ ทำนายคำตอบถูกว่าเป็นคำตอบถูก
- True Negative (TN) คือ ทำนายคำตอบถูกว่าเป็นคำตอบผิด
- False Positive (FP) คือ ทำนายคำตอบผิดว่าเป็นคำตอบถูก
- False Negative (FN) คือ ทำนายคำตอบผิดว่าเป็นคำตอบผิด

ตารางที่ 3 แสดงข้อมูลแอตทริบิวต์ Play จากเทรนนิงดาต้า (actual) 10 ตัวแรก และค่าที่ทำนายได้ (predicted)

No.	Actual	Predicted
1	no	no
2	no	no
3	yes	no
4	yes	yes
5	yes	no
6	no	yes
7	yes	yes
8	no	no
9	yes	no
10	yes	yes

จากตารางที่ 3 โดยที่กำลังพิจารณาคลาส Play = yes จะสามารถสรุปได้ว่า

- True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = yes
  - มีจำนวน 3 ตัว (แถวที่ 4, 7 และ 10)
- True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = no
  - มีจำนวน 3 ตัว (แถวที่ 1, 2 และ 8)
- False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = yes
  - มีจำนวน 1 ตัว (แถวที่ 6)
- False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = no
  - มีจำนวน 3 ตัว (แถวที่ 3, 5 และ 9)

ตารางที่ 4 แสดงตาราง confusion matrix ของข้อมูล weather ซึ่งมี 2 คลาส

คลาสที่โมเดลทำนายมาได้ (predicted)	ข้อมูลเทรนนิงดาต้า (actual)	
	yes	no
yes	3	1
no	3	3

จากตารางที่ 4 นำไปคำนวณค่า Precision, Recall, F-measure และ Accuracy ตามวิธีการคำนวณข้างต้น

แนวทางการใช้ดาต้าไมน์นิ่งเพื่อการศึกษา

การใช้ระบบสารสนเทศและเครือข่ายคอมพิวเตอร์เพื่อการศึกษามีการใช้อย่างแพร่หลาย สามารถแบ่งได้ 3 ลักษณะใหญ่ๆ ดังนี้

1. คอมพิวเตอร์เพื่อการบริหาร การใช้คอมพิวเตอร์ในการบริหารจะทำให้ผู้บริหารสามารถกำหนดนโยบาย วางแผนงานด้านต่างๆ ได้อย่างสะดวก รวดเร็ว และน่าเชื่อถือ ช่วยให้บริหารเป็นไปอย่างมีประสิทธิภาพเพราะมีความพร้อมของข้อมูลครบถ้วน รอบด้าน เช่น ระบบบริหารจัดการงบประมาณ ระบบฐานข้อมูลนักเรียน/นักศึกษา ระบบบริหารทรัพยากรมนุษย์ ระบบจัดตารางเรียน ระบบงานทะเบียนและวัดผลการศึกษา ระบบบริหารจัดการครุภัณฑ์ ฯลฯ

2. คอมพิวเตอร์เพื่อการเรียนการสอน เป็นการนำคอมพิวเตอร์ที่ผู้สอนมาใช้ลักษณะถ่ายทอดเนื้อหาและอำนวยความสะดวกในการเรียนของผู้เรียน เช่น คอมพิวเตอร์กับการจัดการเรียนการสอน (Computer - Managed Instruction : CMI) โดยการจัดโปรแกรมการเรียนให้สอดคล้องกับความต้องการของผู้เรียน และเปิดโอกาสให้ผู้เรียนเรียนรู้ตามความสามารถและความถนัดของตน คอมพิวเตอร์ช่วยสอน (Computer Assisted Instruction : CAI) โดยเสนอบทเรียนเป็นเกม แบบฝึกหัด แบบทดสอบ สถานการณ์จำลอง หรือการสอนโดยใช้เว็บเป็นฐาน (Web base Instruction : WBI) โดยสอนผ่านระบบเครือข่ายอินเทอร์เน็ตผู้เรียนและผู้สอนสามารถติดต่อสื่อสารถึงกันได้ และอาจารย์สามารถติดตามพฤติกรรมการเรียน ตลอดจนผลการเรียนของผู้เรียนได้

3. คอมพิวเตอร์เพื่อบริการการศึกษา เป็นลักษณะการใช้คอมพิวเตอร์ในลักษณะสนับสนุนการเรียน เช่น ฐานข้อมูลห้องสมุด ฐานข้อมูลสืบค้น บริการสื่อการศึกษา พิพิธภัณฑ์และอุทยานการศึกษา ปฏิสัมพันธ์ทางเสียง/ภาพ เป็นต้น

จากลักษณะการใช้ระบบสารสนเทศและเครือข่ายคอมพิวเตอร์เพื่อการศึกษาใน 3 ด้าน หลักๆที่ผ่านมาเป็นข้อมูลที่เก็บไว้มากมาย มหาศาล และสามารถที่จะเข้าถึงได้ง่ายเพราะส่วนใหญ่เก็บในรูปแบบดิจิทัล แต่ข้อมูลที่มีเหล่านี้จะอยู่แบบแยกส่วนกัน ไม่มีการเชื่อมโยงกัน ไม่สามารถใช้ประโยชน์จากข้อมูลที่มีอยู่เท่าที่ควร จากเทคโนโลยีการวิเคราะห์ด้วยเทคนิคดาต้าไมน์นิง เราสามารถนำข้อมูลเหล่านี้ หรือการเก็บเพิ่มเติมตามปัญหาที่เราสนใจ นำไปจัดประเภทข้อมูลด้วยเทคนิคการจัดกลุ่ม (Clustering) การจัดความสัมพันธ์ (Association Rule) การจัดจำแนก (Classification) และ สมการถดถอย (Regression) ด้วยเทคนิคการวิเคราะห์ที่เหมาะสม เช่น K-means, decision trees, naïve Bayesian, artificial neural networks, support vector machines, Logistic Regression เพื่อให้ทราบความสัมพันธ์ รูปแบบของข้อมูล ให้สามารถนำไปใช้ประโยชน์ทางการศึกษา ได้อย่างมีประสิทธิภาพ ดังนี้

1. งานด้านบริหารการศึกษา สามารถรวบรวมข้อมูลจากระบบฐานข้อมูลจากแหล่งสารสนเทศต่างๆ มาใช้ประโยชน์ในการตัดสินใจ กำหนดนโยบาย วางแผน จัดการ เช่น

- การนำข้อมูลบุคลากรในหน่วยงานมาวิเคราะห์โดยใช้เทคนิคการจัดกลุ่ม เพื่อแบ่งกลุ่มบุคลากรที่มีลักษณะความต้องการที่คล้ายคลึงกัน แล้วจัดหา จัดทำ บริการ พัฒนาให้ตรงกับความต้องการของกลุ่มนั้นๆ ได้ตรงใจ

กลุ่มเป้าหมายแต่ละกลุ่ม จะทำให้ผู้บริหารตัดสินใจในการวางแผนบริหาร จัดการและพัฒนาบุคลากรในสังกัดในด้านต่างๆ ให้ตรงกับความต้องการจริงๆ ซึ่งวิธีนี้เป็นการวิเคราะห์ดาต้าไมน์นิ่งโดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน เพื่อรวมรวบข้อมูลที่มีอยู่แล้วมาจัดกลุ่มลักษณะงาน/ความถนัด/ความสนใจ ฯลฯ ให้สามารถบริหารจัดการบุคลากรให้เหมาะสมกับงานสามารถทำงานตามเป้าหมายที่กำหนดไว้

- การนำข้อมูลพัสดุ ครุภัณฑ์ มาวิเคราะห์โดยใช้เทคนิคสมการถดถอย หรือ การประมาณค่าข้อมูล มาพยากรณ์หรือทำนายการใช้งานพัสดุ ครุภัณฑ์ จะเพียงพอหรือไม่ในอนาคต มีการใช้หมุนเวียนคุ้มค่าเพียงใด วิธีนี้เป็นการวิเคราะห์ดาต้าไมน์นิ่งโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน โดยการนำข้อมูลการใช้พัสดุ ครุภัณฑ์ที่ผ่านมาในอดีตมาสอนระบบเพื่อให้เรียนรู้รูปแบบที่เกิดขึ้นในข้อมูล จากนั้นนำมาสร้างเป็นสมการหรือโมเดลขึ้นมา เพื่อหาคำตอบให้สำหรับข้อมูลใหม่

- ระบบงานทะเบียนและวัดผลการศึกษา สามารถใช้ดาต้าไมน์นิ่งวิเคราะห์ผลการเรียนของผู้เรียนโดยใช้เทคนิคสมการถดถอย ซึ่งเมื่อนักศึกษาได้รับการติดตามการเรียนอย่างใกล้ชิดและได้รับคำแนะนำที่ถูกจุดก็จะสามารถช่วยลดอัตราการถอนรายวิชาตรงนี้ได้ การใช้การวิเคราะห์เข้ามาช่วยจะช่วยให้เราสามารถมองเห็นข้อมูลเชิงลึกของผลการเรียนของผู้เรียนในอนาคตได้ โดยการทำนายล่วงหน้านี้จะสามารถช่วยให้ทางสถาบันการศึกษาปรับเปลี่ยนโปรแกรมการสอนได้ทันหากผลการคาดการณ์ออกมาในแง่ลบ เป็นการช่วยให้สถาบันการศึกษาพัฒนาหลักสูตรให้ดีขึ้น โดยคนเรียนเก่งก็ได้จะรับการสนับสนุนทักษะที่ชำนาญให้ยิ่งเกิดความเชี่ยวชาญ ในขณะที่คนที่มีผลการเรียนด้อยก็จะได้รับการขัดเกลาที่ถูกจุดพัฒนาให้เต็มศักยภาพที่พึงมีของแต่ละคน

- การนำข้อมูลการเข้าศึกษาในมหาวิทยาลัยมาวิเคราะห์โดยใช้เทคนิคการจัดจำแนก มาทำนายจำนวนนักศึกษาใหม่ เพื่อให้ผู้บริหารนำผลการวิเคราะห์มาวางแผนการรับนิสิต การจัดสรรงบประมาณ และจัดการเรียนการสอนที่เหมาะสมต่อไป

2. งานด้านการเรียนการสอน สามารถประยุกต์ใช้ดาต้าไมน์นิง

- ช่วยพัฒนาผลการเรียน และศักยภาพของผู้เรียน โดยการนำเอาข้อมูลของนักเรียนรายบุคคลมาวิเคราะห์ เช่น พฤติกรรมการสืบค้นข้อมูล พฤติกรรมการส่งงาน พฤติกรรมการทำกิจกรรม ดูความสัมพันธ์การทำข้อสอบหรือแบบฝึกหัดผิด/ถูก ในแต่ละข้อ แต่ละเนื้อหาหรือแต่ละเรื่อง สามารถใช้เทคนิคการจัดความสัมพันธ์ หรือ “เป็นการหาความสัมพันธ์ของข้อมูลที่เกิดร่วมกัน” เพื่อให้สามารถจัดโปรแกรมการเรียนที่เหมาะสมกับพัฒนาการเรียนรู้ และความสามารถของผู้เรียนแต่ละคน นอกจากนี้ยังสามารถนำข้อมูลพฤติกรรมการเรียนจากบทเรียนที่พัฒนาขึ้นมาวิเคราะห์โดยใช้เทคนิคการจัดกลุ่ม เพื่อช่วยในการแบ่งกลุ่มผู้เรียนในการทำรายงานหรือแบ่งกลุ่มเรียน โดยจะแบ่งตามลักษณะของผู้เรียนที่มีพฤติกรรมและรูปแบบการเรียนที่ใกล้เคียงกันหรือมีความชอบคล้ายคลึงกันมาอยู่กลุ่มเดียวกัน เพื่อให้ผู้สอนสามารถจัดรูปแบบการสอน โปรแกรมการเรียนได้ตรงกับความต้องการและความสามารถผู้เรียนแต่ละกลุ่มมากที่สุด

- ช่วยให้ผู้เรียนเลือกสาขาวิชา/วิชาเอก ที่ตรงกับความสามารถของตนเอง หรือแนวโน้มการสอบผ่าน/ไม่ผ่านในรายวิชาที่ตนเองลงทะเบียนเรียน หรือ สาเหตุที่ทำให้ผลการเรียนตกต่ำ ซึ่งสามารถใช้เทคนิคการจัดจำแนก มาพยากรณ์หรือบอกแนวโน้มที่จะเกิดในอนาคตได้ ซึ่งจะช่วยนิสิตในการตัดสินใจและวางแผนการเรียนของตนเองได้ สามารถลงทะเบียนในรายวิชาที่เหมาะสมกับตนเองมากที่สุด สามารถเลือกเรียนในสาขาที่ตนเองถนัดและตรงกับความสามารถจริงๆ ผู้สอนสามารถวางแผนการสอน ปรับเปลี่ยนวิธีการสอนที่เหมาะสมกับผู้เรียนมากยิ่งขึ้น

- ช่วยในการวัดประเมินและให้ข้อมูลย้อนกลับแบบเรียลไทม์ เมื่อผู้เรียนลงทะเบียนผ่านระบบ เช่น บทเรียนออนไลน์ บทเรียนคอมพิวเตอร์ช่วยสอน ระบบบริหารจัดการเรียนการสอน เราสามารถใช้ดาต้าไมน์นิ่ง นำข้อมูลการเรียนแต่ละคนจากระบบเหล่านี้มาวิเคราะห์โดยใช้เทคนิคการจัดความสัมพันธ์ เพื่อให้อาจารย์ผู้สอนสามารถช่วยเหลือได้

อย่างเจาะจงและถูกจุดสำหรับผู้เรียนแต่ละคน ซึ่งจะนำไปสู่การทำความเข้าใจในบทเรียนที่รวดเร็วขึ้นและผลการเรียนที่ดียิ่งขึ้น อีกทั้งตัวระบบยังเปิดโอกาสให้อาจารย์ผู้สอนได้ติดตามผู้เรียนทุกคนแบบเรียลไทม์ ทำให้สามารถนำข้อมูลเหล่านั้นไปปรับปรุง พัฒนาบทเรียนเรียนและจัดโปรแกรมการเรียนหรือคอร์สเรียนได้อย่างรวดเร็วตรงกับความสามารถผู้เรียนแต่ละคนจริงๆ และระบบวิเคราะห์ยังสามารถช่วยบอกได้อีกว่าจุดไหนหรือหัวข้อไหนที่ยากต่อการทำความเข้าใจของนักเรียน โดยวิเคราะห์จากความถี่ในการอ่านหนังสือแต่ละบท ระยะเวลาที่ใช้งาน/ศึกษาในแต่ครั้ง และจำนวนคำถามที่เกี่ยวกับบทเรียนนั้น ซึ่งดาต้าไมน์นิ่งสามารถให้ข้อมูลเชิงลึกของผู้เรียนแต่ละคนว่าเรียนเป็นอย่างไรบ้างในแต่ละระดับ ซึ่งผู้เรียนแต่ละคนก็จะมีรูปแบบการเรียนที่แตกต่างกันออกไป และวิธีการเรียนที่แตกต่างกันนั้นเองที่มีผลต่อผลการเรียนในรายวิชานั้นๆ

3. งานด้านบริการการศึกษา เป็นลักษณะการใช้คอมพิวเตอร์ในลักษณะสนับสนุนการเรียน เช่น ฐานข้อมูล ห้องสมุด ฐานข้อมูลสืบค้น บริการสื่อการศึกษา พิพิธภัณฑ์และอุทยานการศึกษา ปฏิสัมพันธ์ทางเสียง/ภาพ ซึ่งจากพฤติกรรมการเข้าใช้งานระบบเหล่านี้เราสามารถใช้เทคนิคดาต้าไมน์นิงมาใช้ประโยชน์ได้อย่างหลากหลาย เช่น

- จากพฤติกรรมการใช้ฐานข้อมูลห้องสมุด ฐานข้อมูลสืบค้น สามารถรวบรวมข้อมูลจากสถิติการใช้นำมาวิเคราะห์จัดหมวดหมู่โดยใช้เทคนิคการจัดกลุ่ม และใช้เทคนิคการจัดความสัมพันธ์ เพื่อนำผลการวิเคราะห์นำมาออกแบบและพัฒนาระบบการจัดเก็บและสืบค้นได้อย่างรวดเร็วตรงกับความต้องการของผู้ใช้ เทคนิคนี้จะพบบ่อยๆในหลายเว็บ เช่น เมื่อเราสืบค้นบทความเกี่ยวกับ Big Data เมื่อเราคลิกเข้าไปอ่านบทความนั้นก็จะมีบทความที่เกี่ยวข้องกับ Big Data แสดงไว้ตอนท้ายบทความมากมาย หรือ การแนะนำสินค้า/ข้อมูล ที่เราสนใจขึ้นมาให้เราเลือกแบบอัตโนมัติมากมาย

- นำข้อมูลการใช้บริการสื่อการศึกษา พิพิธภัณฑ์และอุทยานการศึกษา ซึ่งเป็นส่วนสำคัญในการสนับสนุนการเรียนการสอน ถ้านำข้อมูลพฤติกรรมการใช้ในลักษณะต่างๆ มาวิเคราะห์ ก็จะทำให้สามารถจัดหา และพัฒนาได้ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น บริหารจัดการได้อย่างมีประสิทธิภาพ

<u>ตัวอย่างการนำเทคนิคดาต้าไมน์นิงหรือการทำเหมืองข้อมูลมาใช้ทางการศึกษา เช่น</u>

- สุวิมล สิทธิชาติ (2560) ได้ทำวิจัยเรื่อง การวิเคราะห์คุณลักษณะพื้นฐานทางการศึกษาด้วยเทคนิคเหมืองข้อมูล ได้ศึกษาคุณลักษณะของนักศึกษาคณะ วิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ที่มีผลต่อการเรียนแคลคูลัส 1 สำหรับวิศวกรด้วย เทคนิคเหมืองข้อมูลเพื่อจำแนกนักศึกษาออกเป็น 2 กลุ่ม คือ กลุ่มเสี่ยงที่จะไม่ผ่านรายวิชา และกลุ่มที่ไม่มีความเสี่ยง โดยการวิเคราะห์ข้อมูลได้เก็บรวบรวมข้อมูลคุณลักษณะทั้ง 3 ด้าน รวมถึงปัจจัยส่วนบุคคลอื่นๆ ของนักศึกษาจำนวน 453 คน จากผลการศึกษา พบว่า ความแม่นยำในการจำแนกข้อมูลจาก 50 คุณลักษณะนั้น วิธี ANN มีค่า 71.52% และ Decision Trees - J48 มีค่า 66.23% หลังจากทำการคัดเลือกคุณลักษณะแสดงให้เห็นว่าการจำแนกข้อมูลด้วยวิธี ANN จาก 5 คุณลักษณะแรกที่ได้จากการคัดเลือกโดยวิธี Filter Ranker Method ที่คำนวณค่าน้ำหนักด้วย Chi-Square ทำให้ได้ค่าความถูกต้องมีค่าสูงสุด คือ 80.13% และการจำแนกข้อมูลด้วย Decision Tree ก็ให้ผลไปในทางเดียวกัน โดยมีค่าความถูกต้องสูงสุดที่ 75.83% คุณลักษณะที่ผ่านการคัดเลือกนั้นแสดงถึงคุณลักษณะในด้านพื้นฐานความรู้ของนักศึกษาที่มีมาก่อน

- จิราภา เลาหะวรนันท์ และคณะ (2558) ได้ทำวิจัยเรื่อง การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือก แขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ โดยพัฒนา “ระบบแนะนำแขนงวิชา”เพื่อเป็นเครื่องมือช่วยตัดสินใจเลือกแขนงวิชา 4 แขนงวิชา (วิศวกรรมซอฟต์แวร์, เทคโนโลยีเครือข่ายและระบบ, การพัฒนาสื่อประสมและเกม, อัจฉริยะทางธุรกิจ) การวิจัยครั้งนี้ได้ใช้ข้อมูลผลการเรียนและผลการวัดความสามารถด้านต่างๆ ที่เกี่ยวข้องมาสร้างแบบจำลองพยากรณ์โดยเปรียบเทียบเทคนิคการทำเหมืองข้อมูล 5 เทคนิค และพยากรณ์ผ่านเทคนิค “Ensemble” ผลการวิจัยพบว่า การพยากรณ์มีความแม่นยำอยู่ที่ 72.92% โดยแขนงวิศวกรรมซอฟต์แวร์สามารถ

ทำนายได้แม่นยำถึง 86.67% ซึ่งประโยชน์จากการพัฒนาระบบนทำให้นักศึกษาทราบถึงแขนงวิชาที่เหมาะสมกับตนเองมากทสุดถึงน้อยที่สุดพร้อมระดับความน่าจะเป็น ซึ่งจะส่งผลให้นักศึกษาตัดสินใจเลือกสาขาที่ เหมาะกับตนเองได้และมีโอกาสประสบความสำเร็จในการเรียน

บทสรุป

การทำดาต้าไมน์นิง (Data mining) หรือการทำเหมืองข้อมูล เป็นขั้นตอนหลักที่สำคัญอย่างหนึ่งในกระบวนการหาความหมายที่แฝงอยู่ในกลุ่มข้อมูลจำนวนมากที่เก็บไว้ในฐานข้อมูล เพื่อให้ทราบความสัมพันธ์ รูปแบบของข้อมูลนั้นๆ ให้สามารถนำไปใช้ประโยชน์ได้ เช่น การตัดสินใจ การวางแผน การทำนายแนวโน้มสิ่งที่ จะเกิดขึ้นในอนาคต หรือเป็นการแปรเปลี่ยนข้อมูลไปสู่ความรู้ใหม่ๆ เห็นได้จากในปัจจุบันนี้เทคโนโลยีด้านต่างๆ เข้ามามีบทบาทกับพฤติกรรมการใช้ชีวิตของมนุษย์ทุกรูปแบบ เช่น การซื้อของ การทำงาน การท่องเที่ยว การพูดคุย การสื่อสาร ฯลฯ ล้วนใช้เทคโนโลยีทั้งสิ้น ซึ่งข้อมูลส่วนใหญ่ถูกเก็บในรูปแบบดิจิตอลไฟล์ทำให้เข้าถึง ถ่ายโอนได้ง่าย บริษัท หรือองค์กรสถาบันต่างๆ เริ่มให้ความสนใจกับข้อมูลที่ถูกจัดเก็บไว้เพิ่มมากขึ้น โดยมีวัตถุประสงค์เพื่อนำลักษณะ เฉพาะที่แฝงอยู่ภายในกลุ่มข้อมูล มาใช้สนับสนุนการตัดสินใจอย่างมีประสิทธิภาพ ในการดำเนินงานที่เป็นประโยชน์ต่อองค์กรของตนเองทั้งภาครัฐและเอกชน เช่น ทางธุรกิจช่วยแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะผลิตและเสนอสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม ทำนายว่าลูกค้าคนใดจะเลิกใช้บริการจากบริษัทภายใน 6 เดือนหน้า ทางการแพทย์ช่วยวิเคราะห์อาการของคนไข้, วิเคราะห์การจ่ายยา , พยากรณ์แนวโน้มการเกิดโรคระบาด ทางการศึกษา ทำนายผลการเรียน จัดกลุ่มผู้เรียนที่มีลักษณะคล้ายคลึงกันเพื่อจัดโปรแกรมการเรียนให้เหมาะสมกับกลุ่ม/พัฒนาการ ทั้งหมดนี้ล้วนแต่ใช้การทำดาต้าไมน์นิงวิเคราะห์ทั้งสิ้น

การทำดาต้าไมน์นิงสามารถแบ่งประเภทข้อมูลใหญ่ๆ คือ (1) การจัดกลุ่ม (Clustering) (2) การจัดความสัมพันธ์ (Association Rule) (3) การจัดจำแนก (Classification) และ (4) สมการถดถอย (Regression) หรือการประมาณค่าข้อมูล กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านดาต้าไมน์นิง เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อว่า “CRISP-DM” มี 6 ขั้นตอน คือ (1) Business Understanding (2) Data Understanding (3) Data Preparation (4) Modeling (5) Evaluation และ (6) Deployment และการสร้างโมเดลขึ้นมาจะต้องมีการทดสอบประสิทธิภาพของโมเดลที่สร้างได้ ซึ่งจะแยกเป็น 2 ส่วน คือ 1) วิธีการแบ่งข้อมูลเพื่อทำการทดสอบโมเดล 2) ตัวที่ใช้วัดประสิทธิภาพโมเดล และเราสามารถประยุกต์ดาต้าไมน์นิ่งเพื่อการศึกษา 3 ด้าน หลัก คือ (1) งานด้านบริหารการศึกษา (2) งานด้านการเรียนการสอน และ (3) งานด้านบริการการศึกษา

เอกสารอ้างอิง

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมานนท์. (2544). การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์. *NECTEC Technical Journal, 3*(11).

จิราภา เลาหะวรนันท์ และคณะ. (2558). การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. *วารสารเทคโนโลยีสารสนเทศลาดกระบัง, 4*(2).

ชนวัฒน์ ศรีสอ้าน. (2551). *ฐานข้อมูล คลังข้อมูล และเหมืองข้อมูล.* กรุงเทพฯ : สำนักพิมพ์มหาวิทยาลัยรังสิต

ปณิธิ แก้วสวัสดิ์. (2553). *เหมืองข้อมูล : การค้นหาความรู้และการขุดข้อมูล.* คอมพิวเตอร์และเทคโนโลยีขั้นสูง.

สายชล สินสมบูรณ์ทอง. (2558). *การทำเหมืองข้อมูล Data Mining.* กรุงเทพฯ : จามจุรีโปรดักท์.

สุวิมล สิทธิชาติ. (2560). การวิเคราะห์คุณลักษณะพื้นฐานทางการศึกษาด้วยเทคนิคเหมืองข้อมูล. *วารสารเทคโนโลยีสารสนเทศ, 13*(2).

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิงเบื้องต้น.* กรุงเทพฯ : เอเชีย ดิจิตอลการพิมพ์.

Berry, M. J. A., & Linoff, G. (1997). *Data Mining Techniques: for marketing, sales, and customer Support*. Wiley Computer Publishing,

Cabema, P. et al. (1998). *Discovering Data Mining: from concept to implementation*,.Prentice Hall Publishing,

C. Shearer. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.