# DATA MINING

## Concepts and Applications

## แนวคิดการทำเหมืองข้อมูล และการประยุกต์ใช้

นริศรา เอี่ยมคณิตชาติ
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่

# DATA MINING

## Concepts and Applications

## แนวคิดการทำเหมืองข้อมูล และการประยุกต์ใช้

นริศรา เอี่ยมคณิตชาติ
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่

# คำนำ

หนังสือ “แนวคิดการทำเหมืองข้อมูลและการประยุกต์ใช้ (Data Mining: Concepts and Applications)” เล่มนี้ ถูกเรียบเรียงขึ้นเพื่อเป็นหนังสือทางวิชาการที่ครอบคลุมศาสตร์การทำเหมืองข้อมูล (Data Mining) อย่างครบถ้วน ตั้งแต่รากฐานทางทฤษฎีไปจนถึงการประยุกต์ใช้จริงในยุคที่ข้อมูลมีปริมาณมหาศาล ความสามารถในการสกัดองค์ความรู้และข้อมูลเชิงลึกที่ซ่อนอยู่จึงถือเป็นปัจจัยสำคัญในการสร้างความได้เปรียบเชิงกลยุทธ์ทั้งในภาคธุรกิจและวิชาการ เนื้อหาในหนังสือเล่มนี้มุ่งเน้นการสร้างความเข้าใจที่สมบูรณ์ให้แก่ผู้อ่าน

เนื้อหาของหนังสือครอบคลุมตั้งแต่บทนำและพื้นฐานของการทำเหมืองข้อมูล (Data Mining) รวมถึงความสัมพันธ์กับกระบวนการค้นหาความรู้จากข้อมูล (Knowledge Discovery in Databases: KDD) และขั้นตอนสำคัญของการประมวล (ผล) ข้อมูลก่อน (Data Preprocessing) ซึ่งเป็นพื้นฐานจำเป็นก่อนเข้าสู่การวิเคราะห์จริง ต่อมาเป็นเทคนิคหลักของการทำเหมืองข้อมูล เช่น การจำแนกประเภท (Classification) การพยากรณ์ (Regression) การจัดกลุ่ม (Clustering) การค้นหากฎความสัมพันธ์ (Association Rule Mining) การตรวจจับข้อมูลผิดปกติ (Anomaly Detection) การทำเหมืองข้อความ (Text Mining) และการทำเหมืองเว็บ (Web Mining)

ในส่วนหัวข้อขั้นสูงและการประยุกต์ใช้ หนังสือจะกล่าวถึงการประเมินแบบจำลอง (Model Evaluation) แนวโน้มและเทคนิคขั้นสูงในอนาคต เช่น การทำเหมืองข้อมูลขนาดใหญ่ (Big Data Mining) รวมถึงการอภิปรายด้านจริยธรรมและความเป็นส่วนตัวในการทำเหมืองข้อมูล (Ethics and Privacy) เพื่อให้ผู้อ่านตระหนักถึงความสำคัญของการใช้ข้อมูลอย่างรับผิดชอบ

ทุกบทของหนังสือประกอบด้วยการอธิบายแนวคิด สมการที่เกี่ยวข้อง ภาพประกอบที่ผู้เขียนวาดขึ้นเอง แบบฝึกหัดแสดงการคำนวณ ตัวอย่างการประยุกต์ใช้จริงด้วยโค้ดภาษา Python รวมทั้งคำถามทบทวนและคำถามชวนคิดท้ายบท หนังสือเล่มนี้จึงเหมาะสำหรับนักศึกษาระดับปริญญาตรีและปริญญาโท สาขาวิศวกรรมคอมพิวเตอร์และสาขาที่เกี่ยวข้อง รวมถึงผู้สนใจงานด้านการวิเคราะห์ข้อมูล ผู้เขียนหวังว่าหนังสือนี้จะช่วยให้ผู้อ่านได้รับความรู้ที่สมบูรณ์ สามารถเลือกและประยุกต์ใช้อัลกอริทึมการทำเหมืองข้อมูลได้อย่างมีประสิทธิผล และตระหนักถึงการใช้งานข้อมูลอย่างมีจริยธรรมในโลกแห่งความเป็นจริง

นริศรา เอี่ยมคณิตชาติ
มิถุนายน 2569

# สารบัญ

หน้า

# ส่วนที่ 1
# บทนำและพื้นฐาน
# (Introduction and Fundamentals)

ในปัจจุบันเนื่องจากปริมาณของข้อมูลที่เพิ่มขึ้นอย่างมหาศาล ความสามารถในการสกัดองค์ความรู้และข้อมูลเชิงลึกที่ซ่อนอยู่มาใช้ประโยชน์นับเป็นปัจจัยสำคัญที่ก่อให้เกิดความได้เปรียบในการแข่งขัน ทั้งในภาคธุรกิจ การวิจัย และวงการวิทยาศาสตร์ ดังนั้นการทำเหมืองข้อมูล (Data Mining) จึงเป็นศาสตร์ที่ตอบสนองต่อความต้องการดังกล่าว

เนื้อหาในส่วนที่ 1 นี้เป็นการวางรากฐานที่สำคัญสำหรับการศึกษาด้านการทำเหมืองข้อมูลโดยจะครอบคลุมพื้นฐานตั้งแต่แนวคิดหลักไปจนถึงขั้นตอนการปฏิบัติที่จำเป็นก่อนเริ่มกระบวนการวิเคราะห์ข้อมูลจริง

➤ **บทที่ 1** กล่าวถึงนิยามความสำคัญของการทำเหมืองข้อมูลในปัจจุบัน และอธิบายความสัมพันธ์กับกระบวนการค้นพบองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Databases: KDD) พร้อมทั้งนำเสนอตัวอย่างการประยุกต์ใช้ในโลกธุรกิจและวิทยาศาสตร์

➤ **บทที่ 2** มุ่งเน้นศึกษาถึงหัวข้อ “ข้อมูล” โดยอธิบายธรรมชาติและประเภทของข้อมูล ทั้งข้อมูลเชิงปริมาณและเชิงคุณภาพ นำเสนอแนวทางการสำรวจข้อมูลเบื้องต้น (Exploratory Data Analysis) และการสรุปลักษณะสำคัญของข้อมูลด้วยสถิติพื้นฐาน

➤ **บทที่ 3** นำเสนอขั้นตอนการประมวล (ผล) ข้อมูลก่อน (Data Preprocessing) ซึ่งเป็นกระบวนการที่มีความสำคัญอย่างยิ่งและมักใช้เวลามากที่สุด เนื้อหาจะครอบคลุมเทคนิคที่จำเป็นในการปรับปรุงคุณภาพข้อมูลดิบให้พร้อมสำหรับการวิเคราะห์ เช่น การจัดการข้อมูลที่ขาดหายไป (Missing Values) การจัดการข้อมูลรบกวน (Noisy Data) การแปลงข้อมูล (Data Transformation) และการลดข้อมูล (Data Reduction) เพื่อให้แบบจำลองการวิเคราะห์สามารถทำงานได้อย่างมีประสิทธิภาพสูงสุด

เมื่อศึกษาเนื้อหาในส่วนนี้จบแล้วผู้อ่านจะมีความเข้าใจในหลักการและกระบวนการพื้นฐานทั้งหมด ซึ่งเป็นสิ่งจำเป็นสำหรับการศึกษาเทคนิคการทำเหมืองข้อมูลในระดับที่สูงขึ้นต่อไป

# บทที่ 1
# ความรู้เบื้องต้นเกี่ยวกับเหมืองข้อมูล (Introduction to Data Mining)

## 1.1 วัตถุประสงค์ (Objectives)

หลังจากศึกษาเนื้อหาในบทนี้จบแล้ว ผู้อ่านจะเข้าใจรายละเอียดในหัวข้อดังต่อไปนี้:

► ความสำคัญของการทำเหมืองข้อมูล ในบริบทของการปฏิวัติอุตสาหกรรมครั้งที่ 4 (Fourth Industrial Revolution) และยุคแห่งข้อมูลมหาศาล (Big Data Era) ได้

► นิยามของการทำเหมืองข้อมูล และแยกแยะความแตกต่างระหว่างการทำเหมืองข้อมูลกับการสืบค้นฐานข้อมูล (Database Query) และการวิเคราะห์ทางสถิติเบื้องต้น

► ขั้นตอน ในกระบวนการค้นพบองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Databases: KDD) และระบุได้ว่าการทำเหมืองข้อมูลเป็นขั้นตอนใดในกระบวนการดังกล่าว

► ระบุและอธิบายประเภทของงานหลัก ๆ ในการทำเหมืองข้อมูลได้ เช่น การจำแนกประเภท การพยากรณ์ การจัดกลุ่ม และการค้นหากฎความสัมพันธ์

## 1.2 บทนำและความสำคัญในยุคอุตสาหกรรม 4.0

### 1.2.1 การก้าวสู่ยุคแห่งข้อมูล

โลกในศตวรรษที่ 21 กำลังเผชิญกับการเปลี่ยนแปลงครั้งยิ่งใหญ่ที่เรียกว่า การปฏิวัติอุตสาหกรรมครั้งที่ 4 (Fourth Industrial Revolution หรือ Industry 4.0) ซึ่งมีลักษณะเด่นคือการบูรณาการของเทคโนโลยีดิจิทัล ปัญญาประดิษฐ์ (Artificial Intelligence) ระบบอัตโนมัติ (Automation) และอินเทอร์เน็ตของสรรพสิ่ง (Internet of Things: IoT) เข้าด้วยกัน การปฏิวัติครั้งนี้ได้เปลี่ยนโลกให้ก้าวเข้าสู่ยุคแห่งข้อมูล (Data Era) อย่างสมบูรณ์ ที่ซึ่งข้อมูลกลายเป็นทรัพยากรที่มีคุณค่าสูงสุดของยุค เทียบได้กับความสำคัญของน้ำมันในศตวรรษที่ผ่านมา [2], [5] โดยในปัจจุบัน ข้อมูลถูกสร้างขึ้นอย่างต่อเนื่องและรวดเร็วจากแหล่งที่มาที่หลากหลาย ไม่ว่าจะเป็น:

**ภาคธุรกิจและการค้า [2], [3], [5]**

► ระบบจุดขาย (Point of Sale: POS) บันทึกธุรกรรมการซื้อขายหลายล้านรายการต่อวัน

► แพลตฟอร์มอีคอมเมิร์ซติดตามพฤติกรรมการท่องเว็บและการซื้อสินค้าของผู้ใช้นับล้านคน

► ระบบบริหารความสัมพันธ์ลูกค้า (Customer Relationship Management: CRM) เก็บรวบรวมข้อมูลการติดต่อและประวัติของลูกค้า

**สื่อสังคมออนไลน์และสื่อดิจิทัล [43], [48]**

► เฟซบุ๊ก (Facebook) มีผู้ใช้งานมากกว่า 3,000 ล้านคนทั่วโลก ซึ่งสร้างเนื้อหาหลายล้านโพสต์ต่อชั่วโมง

► ยูทูบ (YouTube) มีการอัปโหลดวิดีโอมากกว่า 500 ชั่วโมงต่อนาที

► ติ๊กต็อก (TikTok) สร้างข้อมูลจากการดูวิดีโอ การกดถูกใจ และการแชร์นับพันล้านครั้งต่อวัน

**อุปกรณ์อินเทอร์เน็ตของสรรพสิ่ง (Internet of Things: IoT) และเซ็นเซอร์**

► สมาร์ตโฟนของผู้ใช้ทั่วโลกส่งข้อมูลตำแหน่ง การใช้งานแอปพลิเคชัน และข้อมูลเซ็นเซอร์ต่าง ๆ อย่างต่อเนื่อง

► รถยนต์สมัยใหม่มีเซ็นเซอร์หลายร้อยตัว สร้างข้อมูลหลายกิกะไบต์ต่อชั่วโมงการขับขี่

► เมืองอัจฉริยะ (Smart City) ใช้กล้องวงจรปิด เซ็นเซอร์ตรวจจราจร และเซ็นเซอร์คุณภาพอากาศที่สร้างข้อมูลมหาศาล

**ภาคการแพทย์และสาธารณสุข**

► เครื่องมือทางการแพทย์ เช่น เครื่อง เครื่องตรวจเอ็มอาร์ไอ (Magnetic Resonance Imaging: MRI) เครื่องเอกซเรย์คอมพิวเตอร์ (Computed Tomography: CT Scan) และอัลตราซาวด์ (Ultrasound) สร้างข้อมูลภาพจำนวนมหาศาล

► ระบบบันทึกสุขภาพอิเล็กทรอนิกส์ (Electronic Health Records: EHR) เก็บประวัติการรักษาของผู้ป่วยนับล้านราย

► อุปกรณ์สวมใส่ที่ตรวจสอบสุขภาพ (Wearable Devices) เช่น สมาร์ตวอทช์ ติดตามข้อมูลสุขภาพตลอด 24 ชั่วโมง

**1.2.2 การเติบโตของข้อมูลแบบทวีคูณ [2], [5]**

ปริมาณข้อมูลที่ถูกสร้างขึ้นในโลกมีลักษณะเป็น การเติบโตแบบเลขชี้กำลัง (Exponential Growth) ซึ่งสามารถอธิบายได้ด้วยแบบจำลองทางคณิตศาสตร์ดังนี้:

$$V(t) = V_0 e^{rt}$$ สมการ 1.1

โดยที่

$V(t)$: คือปริมาณข้อมูลในเวลา $t$

$V_0$: ปริมาณข้อมูลเริ่มต้น (ณ เวลาเริ่มต้น)

$e$: จำนวนออยเลอร์ (Euler's Number) มีค่าประมาณ 2.718

$r$: อัตราการเติบโต (Growth Rate)

$t$: เวลา

สมการนี้แสดงให้เห็นว่า ปริมาณข้อมูลไม่ได้เพิ่มขึ้นแบบเส้นตรง แต่เพิ่มขึ้นอย่างรวดเร็วในอัตราทวีคูณ ตัวอย่างเช่น ตามรายงานของ International Data Corporation (IDC) ปริมาณข้อมูลดิจิทัลทั่วโลกในปี 2020 มีประมาณ 64 เซตตะไบต์ (Zettabytes) และคาดการณ์ว่าจะเพิ่มขึ้นเป็น 175 เซตตะไบต์ในปี 2025 ซึ่งหมายถึงการเพิ่มขึ้นมากกว่า 2.7 เท่าในเวลาเพียง 5 ปี [43], [48]

**หมายเหตุ:** 1 เซตตะไบต์ (Zettabyte: ZB) = 1,000 เอกซะไบต์ (Exabytes) = 1,000,000 เพตะไบต์ (Petabytes) = 1,000,000,000 เทระไบต์ (Terabytes)

### 1.2.3 ความท้าทาย: จากภูเขาข้อมูลสู่ความรู้อันทรงคุณค่า

แม้ว่าข้อมูลจำนวนมหาศาลจะถูกสร้างขึ้นอย่างต่อเนื่อง แต่ข้อมูลเหล่านี้อยู่ในรูปแบบดิบ (Raw Data) นั้นไม่มีคุณค่ามากนักหากไม่สามารถแปลงให้กลายเป็นความรู้ที่นำไปใช้ได้จริง (Actionable Knowledge) [2], [5] องค์กรต่าง ๆ กำลังประสบกับปัญหาสำคัญหลายประการ ซึ่งภายใต้บริบทของความซับซ้อนของข้อมูล สามารถจำแนกได้ดังนี้:

► ข้อมูลมีความหลากหลายทั้งในเรื่องของโครงสร้าง ได้แก่ ข้อมูลแบบมีโครงสร้าง (Structured Data) ข้อมูลแบบกึ่งโครงสร้าง (Semi-Structured Data) และข้อมูลแบบไม่มีโครงสร้าง (Unstructured Data) [2], [3]

► ข้อมูลมาจากหลายแหล่งที่แตกต่างกัน มีรูปแบบไม่สอดคล้อง

► ข้อมูลมีมิติสูง (High-Dimensional) ซึ่งก่อให้เกิดความท้าทายต่อการวิเคราะห์ [3], [5]

**ข้อจำกัดของวิธีการวิเคราะห์แบบดั้งเดิม**

► การสืบค้นฐานข้อมูลแบบเดิม (Database Query) ต้องการให้ผู้ใช้ทราบล่วงหน้าว่าต้องการค้นหาอะไร

► การวิเคราะห์ทางสถิติเบื้องต้นต้องการสมมติฐานที่ชัดเจนและมักใช้กับข้อมูลขนาดจำกัด

► วิธีการแบบเดิมไม่สามารถจัดการกับข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ

**ความจำเป็นในการค้นพบรูปแบบที่ซ่อนอยู่**

► ความรู้ที่มีคุณค่ามักซ่อนอยู่ในรูปแบบ (Patterns) ที่ซับซ้อนและไม่ชัดเจน

► การค้นพบความสัมพันธ์ที่ไม่เคยทราบมาก่อนต้องใช้เทคนิคที่ทันสมัย

► องค์กรต้องการระบบที่สามารถ “เรียนรู้” จากข้อมูลได้เองโดยอัตโนมัติ

จากความท้าทายเหล่านี้ จึงเกิดความจำเป็นในการพัฒนาและใช้เทคโนโลยีการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นกระบวนการที่สามารถ “สกัดทองคำจากภูเขาข้อมูล” (Extract Gold from Data Mountains) [2], [5] หรือกล่าวอีกนัยหนึ่งคือการแปลงข้อมูลดิบจำนวนมหาศาลให้กลายเป็นความรู้เชิงลึกที่สามารถนำไปสร้างคุณค่าและความได้เปรียบทางการแข่งขันได้

### 1.2.4 บทบาทของเหมืองข้อมูลในยุคปัจจุบัน

การทำเหมืองข้อมูลได้กลายเป็นเครื่องมือสำคัญที่ช่วยให้องค์กรต่าง ๆ สามารถ [2], [5], [30]:

► เข้าใจพฤติกรรมและความต้องการของลูกค้า เพื่อสร้างกลยุทธ์ทางการตลาดที่มีประสิทธิภาพ

► คาดการณ์แนวโน้มในอนาคต เพื่อการวางแผนและตัดสินใจที่ดีขึ้น

► ค้นพบความผิดปกติและความเสี่ยง เพื่อป้องกันปัญหาที่อาจเกิดขึ้น

► เพิ่มประสิทธิภาพการดำเนินงาน ด้วยการหาจุดที่ต้องปรับปรุง

► สร้างนวัตกรรมใหม่ ๆ จากการค้นพบความรู้ที่ไม่เคยมีมาก่อน

บทนี้จะนำผู้อ่านเข้าสู่โลกของการทำเหมืองข้อมูลตั้งแต่แนวคิดพื้นฐาน กระบวนการทำงาน ประเภทของงานต่าง ๆ ไปจนถึงตัวอย่างการประยุกต์ใช้ในชีวิตจริง เพื่อสร้างความเข้าใจที่ชัดเจนและเตรียมความพร้อมสำหรับการศึกษาในเชิงลึกในบทต่อไป

## 1.3 นิยามและแนวคิดพื้นฐานของเหมืองข้อมูล

### 1.3.1 นิยามของเหมืองข้อมูล

การทำเหมืองข้อมูลคือกระบวนการวิเคราะห์ข้อมูลขนาดใหญ่และมีความซับซ้อน โดยใช้เทคนิคทางสถิติ คณิตศาสตร์ การเรียนรู้ของเครื่อง (Machine Learning) และปัญญาประดิษฐ์ เพื่อค้นพบ รูปแบบ ความสัมพันธ์ (Relationships) แนวโน้ม (Trends) หรือ องค์ความรู้ใหม่ ที่ซ่อนอยู่ภายในชุดข้อมูล ซึ่งเป็นความรู้ที่ ไม่เคยทราบมาก่อน (Previously Unknown) และมีศักยภาพในการนำไปใช้ประโยชน์ (Potentially Useful) [1], [2]

แนวคิดหลักของเหมืองข้อมูลสามารถสรุปได้ด้วยความสัมพันธ์เชิงฟังก์ชันที่เรียบง่ายแต่มีประสิทธิภาพสูง

$$K = f(D)$$ สมการ 1.2

โดยที่

$D$ : คือชุดข้อมูลนำเข้า (Input Dataset) ซึ่งเป็นวัตถุดิบตั้งต้น โดยทั่วไปมักเป็นข้อมูลขนาดใหญ่ (Large-Scale Data) ที่มีความซับซ้อนและอาจยังไม่ผ่านการประมวลผลเพื่อหาความหมายในเชิงลึก

$f$ : แทนฟังก์ชันการทำเหมืองข้อมูล (Mining Function) ซึ่งก็คือ อัลกอริทึมหรือแบบจำลองทางคณิตศาสตร์และสถิติที่ถูกประยุกต์ใช้กับชุดข้อมูล $D$ โดยมีหลากหลายประเภทตามวัตถุประสงค์ของการวิเคราะห์ เช่น การจำแนกประเภทข้อมูล (Classification) การจัดกลุ่ม (Clustering) หรือการค้นหากฎความสัมพันธ์ (Association Rule Mining)

$K$ : คือ องค์ความรู้หรือรูปแบบที่ค้นพบ (Discovered Knowledge and Patterns) ซึ่งเป็นผลลัพธ์ที่มีคุณค่าที่สกัดได้จากข้อมูล องค์ความรู้นี้คือรูปแบบ แนวโน้ม หรือความสัมพันธ์ที่มีนัยสำคัญซึ่งซ่อนอยู่ภายในชุดข้อมูล และไม่เคยถูกค้นพบหรือรับรู้มาก่อน (Previously Unknown and Potentially Useful)

ดังนั้น สมการนี้จึงสรุปหลักของการทำเหมืองข้อมูลโดยประยุกต์จากแนวคิดของ [2], [5] ว่าเป็นกระบวนการแปลง ข้อมูลดิบ $(D)$ ให้กลายเป็น องค์ความรู้ใหม่ $(K)$ ผ่านการประยุกต์ใช้ อัลกอริทึมที่เหมาะสม $(f)$ อย่างเป็นระบบนั่นเอง

### 1.3.2 การเปลี่ยนจากข้อมูลสู่ความรู้

กระบวนการเหมืองข้อมูลสามารถมองได้เป็นการแปลงที่มีหลายระดับ [2], [3], [5]:

**ระดับที่ 1: ข้อมูลดิบ (Raw Data)** ซึ่งเป็นข้อมูลที่ยังไม่ผ่านการประมวลผล ไม่มีโครงสร้างที่ชัดเจนหรืออาจมีความผิดพลาด ตัวอย่างเช่น ไฟล์ log ของเว็บเซิร์ฟเวอร์ รายการธุรกรรมการซื้อขายที่ยังไม่จัดกลุ่ม

**ระดับที่ 2: ข้อมูลที่มีความหมาย (Meaningful Data)** ข้อมูลที่ผ่านการทำความสะอาดและจัดระเบียบแล้ว มีโครงสร้างที่ชัดเจน พร้อมใช้ในการวิเคราะห์ ตัวอย่างเช่น ตารางข้อมูลลูกค้าที่มีฟิลด์ครบถ้วน ไม่มีค่าที่หายไป

**ระดับที่ 3: ข้อมูลเชิงสารสนเทศ (Information)** ข้อมูลที่ถูกวิเคราะห์และสรุปในบริบทที่เฉพาะเจาะจง สามารถตอบคำถามพื้นฐานได้ เช่น “อะไร” “เมื่อไร” “ที่ไหน” ตัวอย่างเช่น รายงานยอดขายรายเดือน สถิติการเข้าชมเว็บไซต์

**ระดับที่ 4: องค์ความรู้ (Knowledge)** ความเข้าใจเชิงลึกที่ได้จากการค้นพบรูปแบบและความสัมพันธ์ สามารถตอบคำถามซับซ้อน เช่น “ทำไม” “อย่างไร” และ “จะเกิดอะไรขึ้นต่อไป” ตัวอย่างเช่น การค้นพบว่าลูกค้าที่ซื้อสินค้า A มักจะซื้อสินค้า B ภายใน 7 วัน หรือการทำนายว่าลูกค้ารายใดมีแนวโน้มจะเลิกใช้บริการ

### 1.3.3 การทำเหมืองข้อมูล: สิ่งที่เป็นและสิ่งที่ไม่ใช่

เพื่อเสริมสร้างความเข้าใจที่ชัดเจนเกี่ยวกับเหมืองข้อมูลจึงมีความจำเป็นต้องจำแนกความแตกต่างระหว่างการทำเหมืองข้อมูลและเทคนิคการจัดการข้อมูลอื่น ๆ ที่เกี่ยวข้อง โดยสามารถสรุปการเปรียบเทียบประเด็นหลักได้ดังตารางที่ 1.1

**ตารางที่ 1.1** การเปรียบเทียบเหมืองข้อมูลกับเทคนิคอื่น [2], [4]

| มิติเปรียบเทียบ | การสืบค้นฐานข้อมูล (Database Query) | การวิเคราะห์ทางสถิติ (Statistical Analysis) | การทำเหมืองข้อมูล (Data Mining) |
|---|---|---|---|
| วัตถุประสงค์ | ค้นหาข้อมูลตามเงื่อนไขที่กำหนด | ทดสอบสมมติฐานและหาความสัมพันธ์ | ค้นพบรูปแบบและความรู้ใหม่ที่ไม่เคยทราบ |
| สมมติฐาน | ต้องทราบว่าต้องการค้นหาอะไร | ต้องมีสมมติฐานล่วงหน้า | ไม่จำเป็นต้องมีสมมติฐานล่วงหน้า |
| ขนาดข้อมูล | ทำงานได้ดีกับข้อมูลขนาดเล็กถึงกลาง | เหมาะกับข้อมูลขนาดจำกัด | ออกแบบมาสำหรับข้อมูลขนาดใหญ่มาก |
| ประเภทข้อมูล | เน้นข้อมูลที่มีโครงสร้าง (Structured) | ส่วนใหญ่เป็นข้อมูลที่มีโครงสร้าง | จัดการได้ทั้งข้อมูลที่มีและไม่มีโครงสร้าง |
| ผลลัพธ์ | ชุดข้อมูลที่ตรงตามเงื่อนไข | การยืนยัน/ปฏิเสธสมมติฐาน | รูปแบบ กฎ หรือแบบจำลองใหม่ |
| ตัวอย่าง | “แสดงลูกค้าทั้งหมดที่อายุ > 30” | “รายได้มีความสัมพันธ์กับการใช้จ่ายหรือไม่?” | “ค้นหากลุ่มลูกค้าที่มีพฤติกรรมคล้ายกัน” |
| เครื่องมือ | SQL, NoSQL | SPSS, R (แบบดั้งเดิม), Excel | Python, R, Weka, RapidMiner |
| ความซับซ้อน | ต่ำ | ปานกลาง | สูง |

**ตัวอย่างเปรียบเทียบความแตกต่าง**

สมมติว่ามีข้อมูลการซื้อสินค้าของลูกค้าในซูเปอร์มาร์เก็ต

► **การสืบค้นฐานข้อมูล** จะตอบคำถามที่ทราบล่วงหน้า เช่น “ลูกค้าที่ซื้อนมในวันนี้มีกี่คน” (ต้องทราบว่าต้องการหาอะไร)

► **การวิเคราะห์ทางสถิติ** จะใช้ทดสอบความสัมพันธ์ที่มีสมมติฐาน เช่น “รายได้ของลูกค้ามีความสัมพันธ์กับมูลค่าการซื้อหรือไม่” (มีสมมติฐานล่วงหน้า)

► **การทำเหมืองข้อมูล** จะค้นพบว่ารูปแบบความสัมพันธ์ เช่น “ลูกค้าที่ซื้อขนมปังและเนยมักจะซื้อแยมด้วยในธุรกรรมเดียวกันร้อยละ 68 ของกรณี” (ค้นพบความสัมพันธ์ที่ไม่เคยทราบมาก่อน)

### 1.3.4 คุณลักษณะสำคัญของเหมืองข้อมูล (Key Characteristics of Data Mining)

การทำเหมืองข้อมูลที่มีประสิทธิภาพสูงจำเป็นต้องอาศัยคุณลักษณะหลักหลายประการเพื่อบรรลุผลลัพธ์ที่เป็นประโยชน์และเชื่อถือได้ดังต่อไปนี้:

► **การค้นพบแบบอัตโนมัติ (Automatic Discovery)** ซึ่งหมายความว่าระบบมีความสามารถในการค้นหารูปแบบ (Patterns) ได้เองโดยมีการแทรกแซงจากมนุษย์น้อยที่สุด [1], [12] ซึ่งช่วยลดความจำเป็นในการระบุสมมติฐานล่วงหน้าก่อนการวิเคราะห์

► **ความสามารถในการขยายขนาด (Scalability)** เป็นคุณลักษณะที่สำคัญอย่างยิ่ง เนื่องจากระบบต้องสามารถจัดการกับชุดข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ โดยที่เวลาในการประมวลผลควรเพิ่มขึ้นในอัตราที่ยอมรับได้เมื่อปริมาณข้อมูลเพิ่มขึ้น [2], [5]

► **ความหลากหลายของประเภทข้อมูล (Data Type Versatility)** เป็นปัจจัยสำคัญที่ช่วยให้เหมืองข้อมูลมีความยืดหยุ่น โดยระบบต้องสามารถรองรับข้อมูลประเภทต่าง ๆ เช่น ข้อมูลตัวเลข ข้อความ รูปภาพ เสียง และวิดีโอ อีกทั้งยังต้องจัดการกับข้อมูลที่มีโครงสร้าง กึ่งโครงสร้าง และไม่มีโครงสร้างได้อย่างเหมาะสม

► **การค้นพบความรู้ที่น่าสนใจและมีประโยชน์ (Interesting and Useful Knowledge)** เป็นคุณลักษณะที่เน้นว่ารูปแบบที่ค้นพบจะต้องมีนัยสำคัญทางสถิติ และที่สำคัญที่สุดคือ สามารถนำไปใช้ในการตัดสินใจหรือแก้ปัญหาได้จริง

► **ความสามารถในการตีความและอธิบายผล (Interpretability)** เป็นสิ่งที่ไม่อาจมองข้ามได้ เพราะผลลัพธ์ที่ได้ควรจะสามารถเข้าใจและอธิบายได้ง่าย เพื่อให้ผู้ใช้ที่ไม่ใช่ผู้เชี่ยวชาญสามารถนำความรู้นั้นไปประยุกต์ใช้ในบริบทต่าง ๆ ได้อย่างมีประสิทธิภาพ สอดคล้องกับแนวทางการพัฒนาโมเดลที่เน้นการตีความได้ซึ่งผู้เขียนได้เคยนำเสนอไว้ [56]

## 1.4 กระบวนการค้นพบองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Databases: KDD)

### 1.4.1 ภาพรวมของกระบวนการ KDD

การทำเหมืองข้อมูล ไม่ใช่กระบวนการเดียวที่ทำงานอย่างโดดเดี่ยว หากแต่เป็นส่วนหนึ่งของกระบวนการที่ใหญ่กว่าที่เรียกว่า กระบวนการค้นพบองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Databases: KDD) ซึ่งถือเป็นโครงร่างเชิงระบบ (Systematic Framework) ที่ครอบคลุมตั้งแต่การเตรียมข้อมูลดิบจนถึงการนำเสนอความรู้ที่ค้นพบ [1], [2] ด้วยเหตุนี้ การทำเหมืองข้อมูลจึงเป็นเพียงหนึ่งในขั้นตอน แต่ถือเป็นส่วนสำคัญของกระบวนการ KDD ทั้งหมด คล้ายกับการที่เครื่องยนต์เป็นส่วนสำคัญของรถยนต์ แต่รถยนต์ที่สมบูรณ์ประกอบด้วยระบบย่อยอื่น ๆ อีกมากมาย การทำความเข้าใจกระบวนการ KDD ทั้งหมดจึงเป็นสิ่งสำคัญ เนื่องจากความสำเร็จของโครงการเหมืองข้อมูลไม่ได้ขึ้นอยู่กับอัลกอริทึม

ในโลกที่ข้อมูลเพิ่มขึ้นอย่างรวดเร็ว การมีข้อมูลจำนวนมากไม่ได้หมายถึงความได้เปรียบ หากขาดความสามารถในการวิเคราะห์และสกัด “องค์ความรู้” ที่มีความหมาย หนังสือเล่มนี้จึงมุ่งพัฒนาความเข้าใจด้านการทำเหมืองข้อมูล (Data Mining) อย่างเป็นระบบ เพื่อเปลี่ยนข้อมูลดิบให้เป็นเครื่องมือสำคัญในการตัดสินใจเชิงกลยุทธ์

หนังสือนำเสนอทั้งแนวคิดพื้นฐานและเทคนิคสำคัญของการวิเคราะห์ข้อมูล ครอบคลุมตั้งแต่การเตรียมข้อมูลไปจนถึงการสร้างและประเมินแบบจำลอง พร้อมเชื่อมโยงสู่บริบทการใช้งานจริงในยุคข้อมูลขนาดใหญ่ (Big Data) และระบบปัญญาประดิษฐ์สมัยใหม่

จุดเด่นของหนังสือเล่มนี้

- อธิบายแนวคิดอย่างเป็นลำดับ พร้อมภาพประกอบและสมการที่ช่วยสร้างความเข้าใจเชิงลึก
- มีตัวอย่างเชิงคำนวณและโค้ดภาษา Python สำหรับการประยุกต์ใช้งานจริง
- ครอบคลุมทั้งพื้นฐาน เทคนิคหลัก และแนวโน้มเทคโนโลยีสมัยใหม่ เช่น Explainable AI (XAI)
- ให้ความสำคัญกับจริยธรรม ความเป็นส่วนตัว และความรับผิดชอบในการใช้ข้อมูล

ด้วยการผสานระหว่างทฤษฎีและการประยุกต์ใช้ หนังสือเล่มนี้เหมาะสำหรับนักศึกษา นักวิจัย และผู้ปฏิบัติงานด้านข้อมูลที่ต้องการพัฒนาความเข้าใจเชิงลึก และสามารถนำเทคนิคการทำเหมืองข้อมูลไปใช้ได้อย่างมีประสิทธิภาพในสถานการณ์จริง

เพราะความได้เปรียบในยุคข้อมูล ไม่ได้อยู่ที่ “ปริมาณ” แต่อยู่ที่ความสามารถในการแปลงข้อมูลให้เป็น “ความรู้ที่ใช้ได้จริง”

CHIANG MAI
UNIVERSITY PRESS

$23.40
ISBN 978-0-9999-9999-8
52340
9 780999 999998