วิชาเหมืองข้อมูล (Data Mining)

บทที่ 2

ความรู้พื้นฐานของเหมืองข้อมูล

Asst. Prof. Dr. Nattapong Songneam

Program in Computer Science , Faculty of Science and Technology,

Phranakhon Rajabhat University

ล่าสุด 12.11.2568

บทที่ 2 ความรู้พื้นฐานของเหมืองข้อมูล

- 2.1 เหมืองข้อมูลคืออะไร (What is Data Mining)
- 2.2 วัตถุประสงค์เหมืองข้อมูล (Objectives of Data Mining)
- 2.3 ส่วนประกอบของเหมืองข้อมูล (Components of Data Mining)
- 2.4 เทคนิคที่ใช้ในเหมืองข้อมูล (Techniques Used in Data Mining)
- 2.5 ประโยชน์ของเหมืองข้อมูล
- 2.6 ประเภทของการทำเหมืองข้อมูล

บทที่ 2 พื้นฐานของเหมืองข้อมูล (ต่อ)

- 2.7 การประยุกต์ใช้งานเหมืองข้อมูล
- 2.8 ขั้นตอนในการทำเหมืองข้อมูล
- 2.9 รู้จักข้อมูล
- 2.10 การเตรียมข้อมูลและการนำเข้าข้อมูล
- 2.11 การเตรียมข้อมูลด้วยภาษาไพธอน
- 2.12 แบบฝึกหัด

2.1 เหมืองข้อมูลคืออะไร

เหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม

การทำเหมืองข้อมูล (Data Mining) เปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล หรือจะแยกๆ เป็นข้อๆ ได้ดังนี้

- **กระบวนการหรือการเรียงลำดับของการค้นข้อมูลจำนวนมากและเก็บข้อมูลที่เกี่ยวข้อง**
- **การนำมาใช้โดยหน่วยงานทางธุรกิจและนักวิเคราะห์ทางการเงินหรือการนำมาใช้งานในด้านวิทยาศาสตร์เพื่อเอาข้อมูลขนาดใหญ่ที่สร้างโดยวิธีการทดลองและการสังเกตการณ์ที่ทันสมัย**
- **การสกัดหรือแยกข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่หรือฐานข้อมูล**
- **การวางแผนทรัพยากรขององค์กรโดยสามารถวิเคราะห์ทางสถิติและตรรกะของข้อมูลขนาดใหญ่เป็นการมองหารูปแบบที่สามารถช่วยการตัดสินใจได้**

2.1 เหมืองข้อมูลคืออะไร

ความหมายของเหมืองข้อมูล

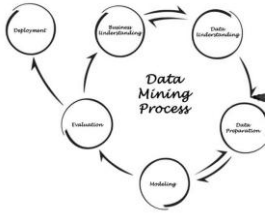
เหมืองข้อมูล (Data Mining) หมายถึง การวิเคราะห์ข้อมูลจากข้อมูลจำนวนมาก (big data) เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกประเภท รูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กัน และหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ ที่สามารถนำไปใช้ประกอบการตัดสินใจในด้านต่าง ๆ เช่น ตลาดหลักทรัพย์,ทางธุรกิจ, ทางด้านการแพทย์, ยุทธศาสตร์ทหาร เป็นต้น

2.2 องค์ประกอบของเหมืองข้อมูล

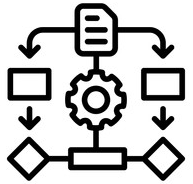
1 แหล่งข้อมูล (Data Sources)

2 การเตรียมข้อมูล (Data Preparation)

3 การวิเคราะห์ข้อมูล (Data Mining Techniques)



4 อัลกอริทึม (Algorithms) ใช้ในการสร้างโมเดลข้อมูล



5 เครื่องมือการเหมืองข้อมูล (Data Mining Tools)

6 การประเมินผลและแสดงผล (Evaluation and Visualization)

7 ผู้ใช้และการประยุกต์ใช้งาน (Applications)

2.2 องค์ประกอบของเหมืองข้อมูล

องค์ประกอบของเหมืองข้อมูล (Data Mining) ประกอบด้วยส่วนสำคัญดังนี้

1. แหล่งข้อมูล (Data Sources) คือชุดข้อมูลที่นำมาใช้ในเหมืองข้อมูลมีทั้งแบบโครงสร้างและไม่มีโครงสร้าง เช่น:

- ฐานข้อมูล (Databases)
- คลังข้อมูล (Data Warehouses)
- ข้อมูลบนเว็บ (Web Data)
- ข้อมูลจากเซ็นเซอร์ (Sensor Data)
- ข้อมูลข้อความหรือรูปภาพ

2. การเตรียมข้อมูล (Data Preparation) คือ ขั้นตอนนี้เป็นการจัดการข้อมูลให้เหมาะสมกับการวิเคราะห์ เช่น:

- การทำความสะอาดข้อมูล (Data Cleaning)
- การรวมข้อมูล (Data Integration)
- การเลือกข้อมูล (Data Selection)
- การลดมิติข้อมูล (Dimensionality Reduction)

2.2 องค์ประกอบของเหมืองข้อมูล

องค์ประกอบของเหมืองข้อมูล (Data Mining) ประกอบด้วยส่วนสำคัญดังนี้ (ต่อ)

3. การวิเคราะห์ข้อมูล (Data Mining Techniques) คือ เทคนิคต่าง ๆ ที่ใช้ค้นหารูปแบบในข้อมูล เช่น:

- การจำแนกประเภท (Classification)
- การจัดกลุ่ม (Clustering)
- การวิเคราะห์ความสัมพันธ์ (Association Rule Mining)
- การคาดการณ์ (Prediction)

4. อัลกอริทึม (Algorithms) ใช้ในการสร้างโมเดลข้อมูล เช่น:

- Decision Trees
- Neural Networks
- Support Vector Machines (SVM)
- K-Means Clustering

2.2 องค์ประกอบของเหมืองข้อมูล

องค์ประกอบของเหมืองข้อมูล (Data Mining) ประกอบด้วยส่วนสำคัญดังนี้ (ต่อ)

5. เครื่องมือการเหมืองข้อมูล (Data Mining Tools) คือ โปรแกรมและซอฟต์แวร์ที่ใช้ เช่น:

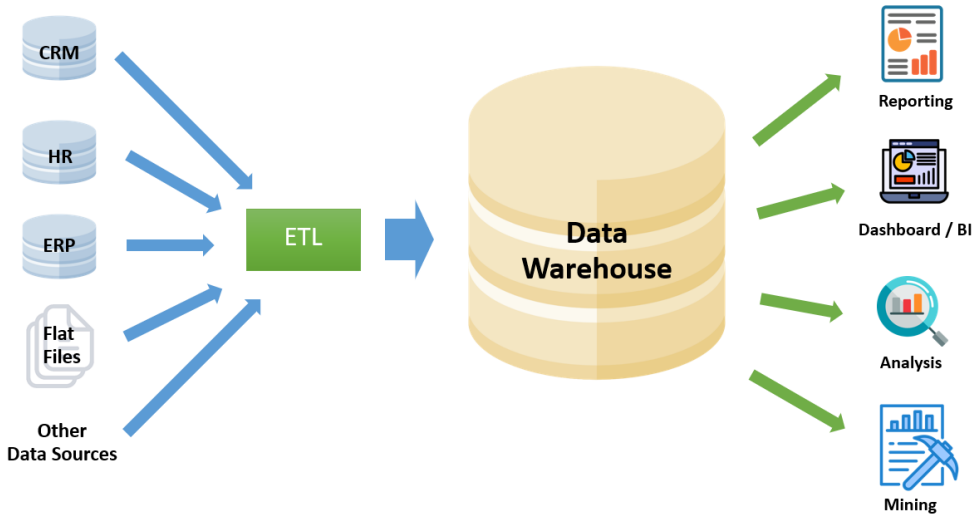
- **Weka**
- **RapidMiner**
- **Python (ไลบรารีเช่น scikit-learn, pandas)**
- **R**

6. การประเมินผลและแสดงผล (Evaluation and Visualization) ประเมินคุณภาพของโมเดลและแสดงผลข้อมูลในรูปแบบที่เข้าใจง่าย เช่น กราฟหรือแผนภูมิ

7. ผู้ใช้และการประยุกต์ใช้งาน (Applications)

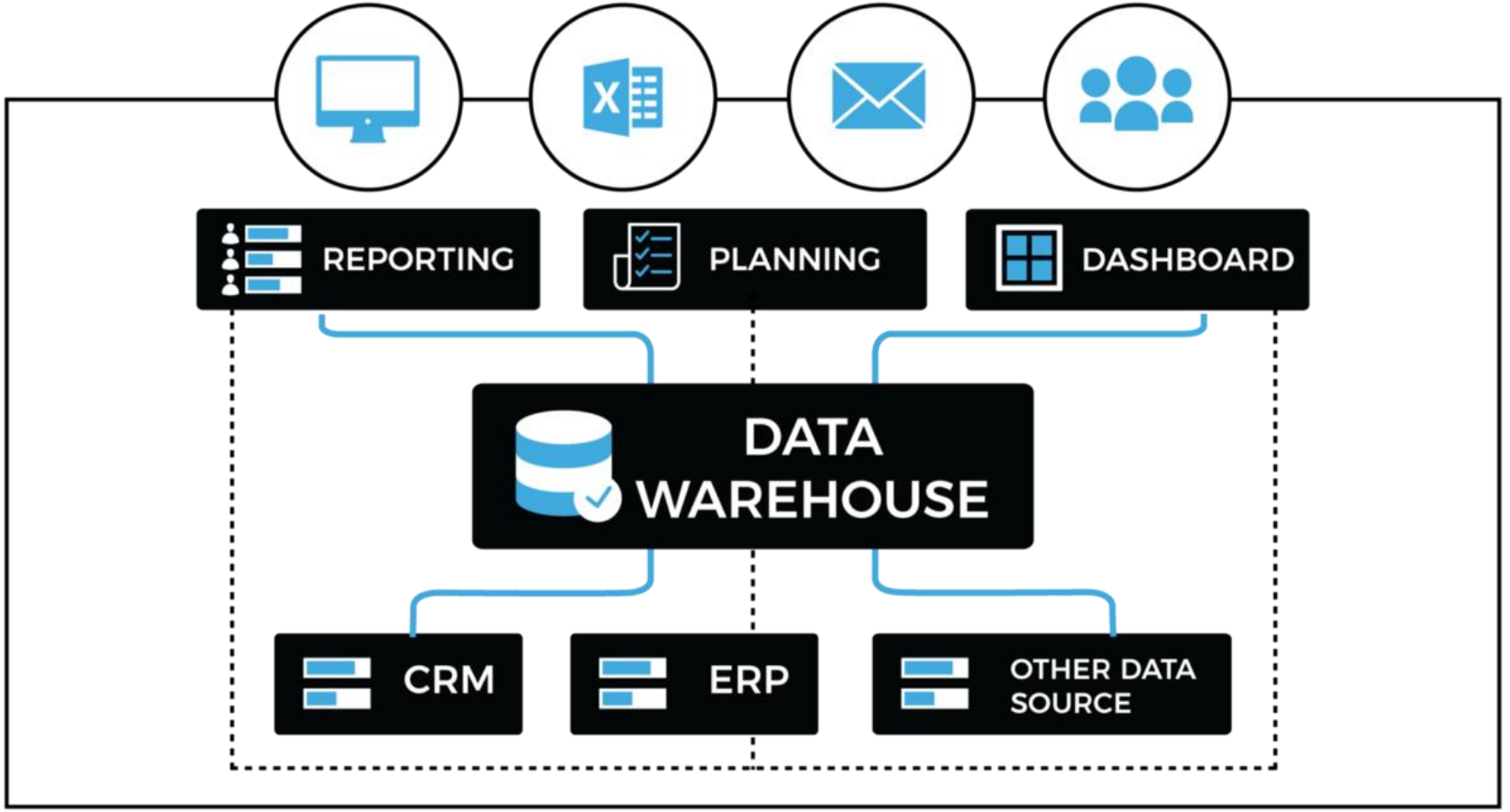
2.3 ระบบคลังข้อมูล (Data Ware-house)

- **ระบบคลังข้อมูล หรือ Data Ware-house คือ ระบบการจัดเก็บ รวบรวมข้อมูล ที่มีอยู่ในระบบปฏิบัติการต่าง ๆ ขององค์กร โดยข้อมูลเหล่านั้นมักเป็นข้อมูลกระจัดกระจาย ให้มารวมไว้เป็นศูนย์กลางข้อมูล ขององค์กร และสามารถเก็บข้อมูลย้อนหลังได้หลายๆ ปี เพื่อใช้เป็นข้อมูลช่วยสนับสนุนการตัดสินใจ (Decision Support System) หรือใช้ในการวิเคราะห์ข้อมูล ที่ถูกต้อง และมีประสิทธิภาพ โดยการวิเคราะห์ต้องทำได้แบบหลายมิติ (Multidimensional Analysis) ตลอดจนการวิเคราะห์ทางธุรกิจ เช่น การพยากรณ์ (Forecasting), What-If Analysis, Data Mining เป็นต้น**



หัวข้อ	Database (ฐานข้อมูล)	Data Warehouse (คลังข้อมูล)
วัตถุประสงค์หลัก	ใช้สำหรับการประมวลผลธุรกรรม (OLTP) เช่น CRUD (Create, Read, Update, Delete)	ใช้สำหรับการวิเคราะห์ข้อมูล (OLAP) และสร้างรายงาน
ประเภทข้อมูล	ข้อมูลปัจจุบันและกำลังดำเนินการ	ข้อมูลในอดีตที่รวมจากหลายแหล่ง
โครงสร้างข้อมูล	ปกติเป็นแบบ normalized (จัดการซ้ำซ้อนของข้อมูล)	มักใช้ denormalized เพื่อง่ายต่อการวิเคราะห์
ความเร็ว	เร็วสำหรับการประมวลผลธุรกรรมจำนวนมาก	เร็วในการอ่านข้อมูลปริมาณมากเพื่อการวิเคราะห์
การอัปเดตข้อมูล	มีการอัปเดตข้อมูลบ่อย ๆ	ข้อมูลมักจะโหลดเป็นรอบ (เช่น รายวัน รายสัปดาห์)
ปริมาณข้อมูล	ข้อมูลขนาดเล็กถึงกลาง	ข้อมูลขนาดใหญ่มาก (หลาย TB หรือมากกว่านั้น)
ผู้ใช้งานหลัก	พนักงานปฏิบัติการ ฝ่ายบริการลูกค้า ฯลฯ	นักวิเคราะห์ข้อมูล ผู้บริหาร
ตัวอย่างระบบ	MySQL, PostgreSQL, SQL Server, Oracle Database	Amazon Redshift, Google BigQuery, Snowflake, SAP BW

2.3 ระบบคลังข้อมูล (Data Ware-house)



ภาพที่ 1. ระบบคลังข้อมูล

ที่มา : https://medium.com/@patipolchiammunchit/intro-to-data-warehouse-data-warehouse-คืออะไร-แตกต่าง กับ-database-ยังไง-ff88ec04cb6c

2.4 ปัญหาของ Database

- **ปัญหาของ Database**

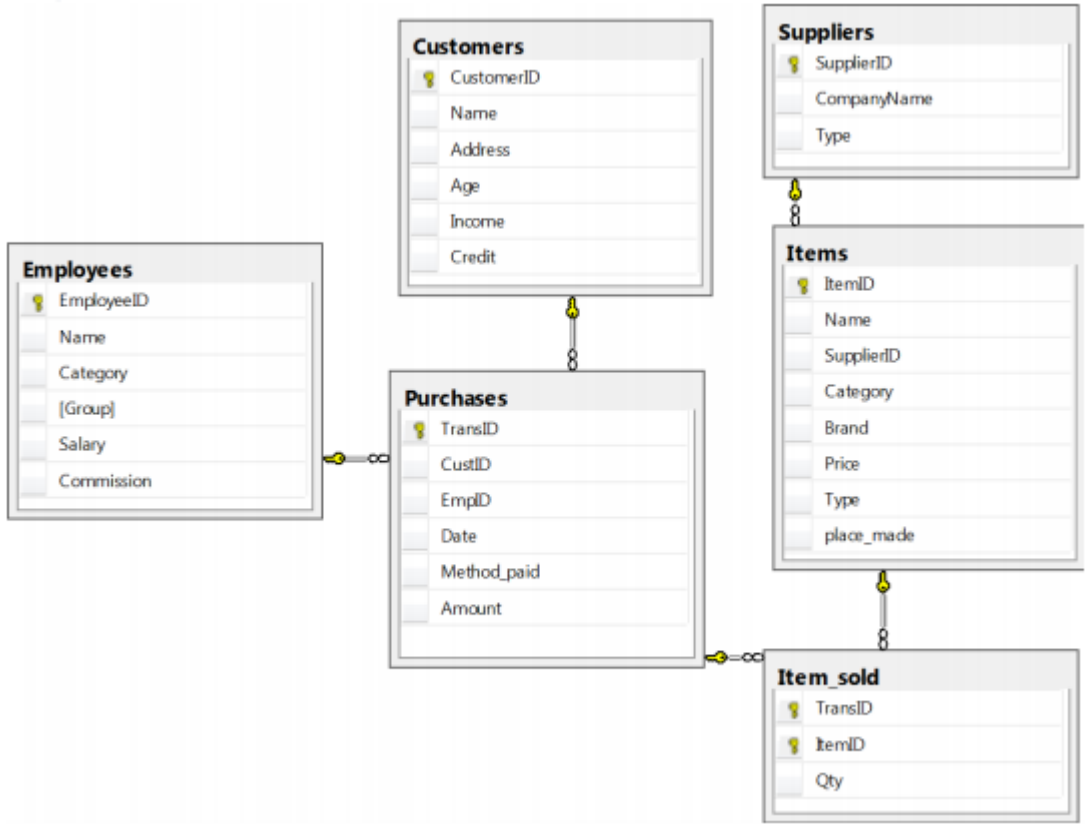
ลองนึกภาพว่าคุณทำงานอยู่แผนก IT ในบริษัทแห่งหนึ่ง โดยมีหน้าที่คอย Query SQL เพื่อคอยตอบคำถามให้กับผู้บริหาร คำถามที่ผู้บริหารส่วนใหญ่ถามจะเป็นประมาณนี้

- **ยอดขายของสินค้าที่มาจากบริษัท ABC ในปี 2015**
- **สรุปยอดขายรายปีของ 5 ปีล่าสุด**
- **แล้วคุณมี Relational Database ซึ่งออกแบบด้วยวิธี Normalization ดังนี้**

https://medium.com/@patipolchiammunchit/intro-to-data-warehouse-data-warehouse-คืออะไร-แตกต่าง-กับ-database-ยังไง-ff88ec04cb6c

2.4 ปัญหาของ Database

แล้วคุณมี Relational Database ซึ่งออกแบบด้วยวิธี Normalization ดังนี้



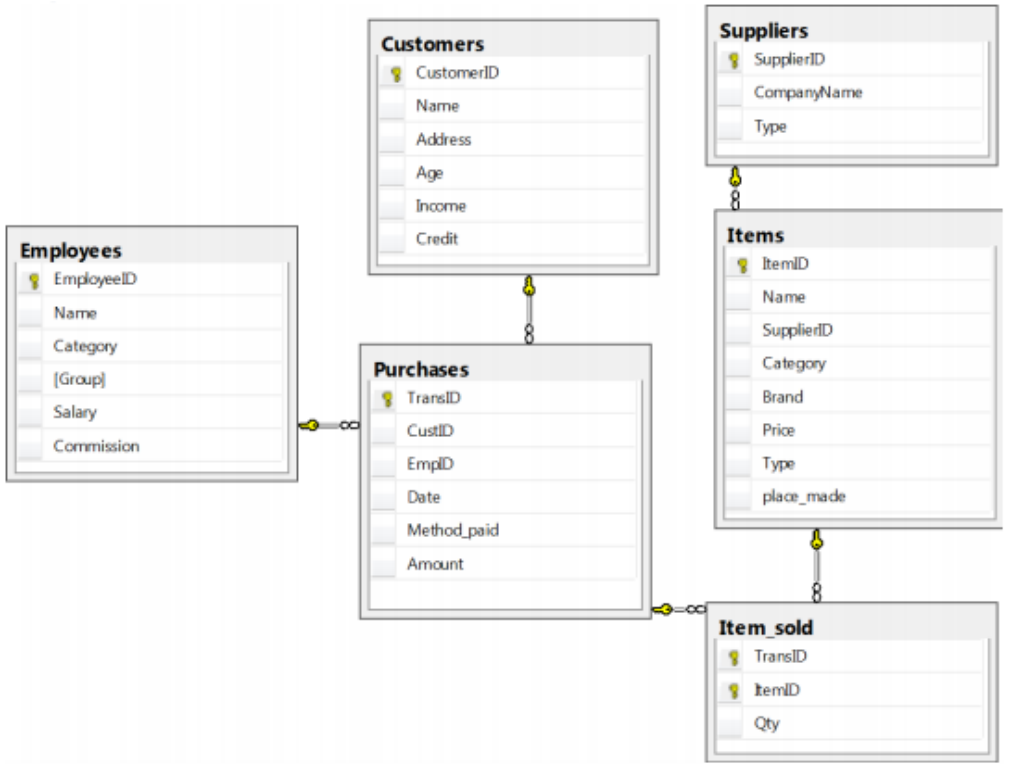
2.4 ปัญหาของ Database

แล้วคุณมี Relational Database ซึ่งออกแบบด้วยวิธี Normalization ดังนี้

- **ลองคิด SQL Command ที่ใช้ตอบคำถามเมื่อกี้นี้ดูนะครับ**
  ยอดขายของสินค้าที่มาจากบริษัท ABC ในปี 2015:

  SELECT Items.Name as ItemName, Items.Price * Item_sold.Qty as TotalSale FROM Purchases INNER JOIN Item_sold ON Purchases.TransID = Item_sold.TransID INNER JOIN Items ON Item_sold.ItemID = Items.ItemID INNER JOIN Suppliers ON Items.SupplierID = Suppliers.SupplierID WHERE CompanyName = 'ABC' AND DATEPART(yyyy, Purchases.Date) = 2015

  สรุปยอดขายรายปีของ 5 ปีล่าสุด:
  SELECT DATEPART(yyyy, Purchases.Date) AS Year, SUM(Purchases.Amount) AS Total FROM Purchases GROUP BY DATEPART(yyyy, Purchases.Date) ORDER BY DATEPART(yyyy, Purchases.Date) DESC



2.4 ปัญหาของ Database

แล้วคุณมี Relational Database ซึ่งออกแบบด้วยวิธี Normalization ดังนี้

- **SQL Command ที่เขียนไปข้างต้นอาจจะไม่ถูกต้อง 100% นะครับ แต่เราจะสังเกตได้ว่า**

1. SQL Command ค่อนข้าวยาวและซับซ้อน มีการเรียกใช้ Function ที่ไม่คุ้นตา เช่น DATEPART เป็นต้น

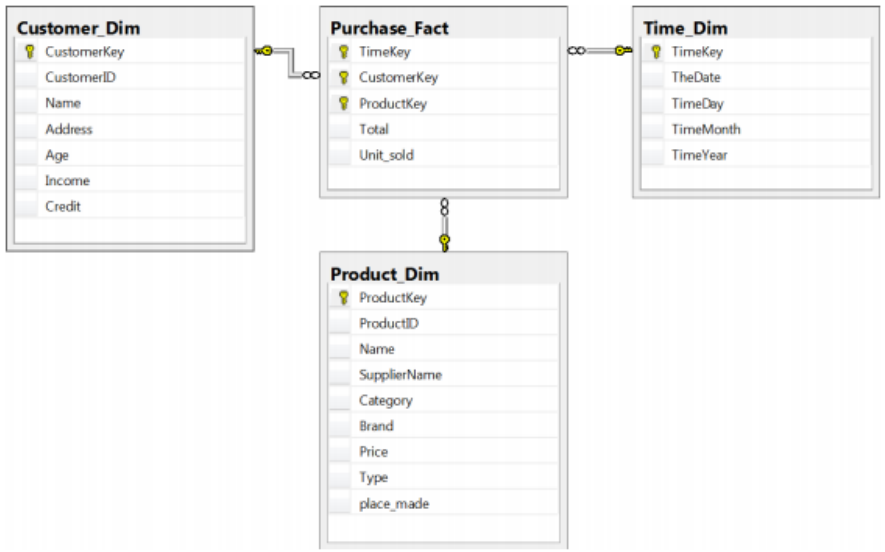
2. ใช้คำสั่ง JOIN ค่อนข้างเยอะ ซึ่งคำสั่ง JOIN เป็นคำสั่งที่ทำงานช้ามาก

3. เมื่อมีการเปลี่ยนแปลงคำถาม SQL Command จะเปลี่ยนรูปแบบการเขียนแตกต่างกันอย่างสิ้นเชิง

- **ซึ่งปัญหาที่กล่าวมานี้ ทำให้การออกแบบ Relational Database แบบ Normalization ไม่เหมาะกับการ Query เพื่อตอบคำถามผู้บริหาร**

2.4 ปัญหาของ Database

- **Data Warehouse...พระเอกของเรา**
- **Data Warehouse เป็น Relational Database ออกแบบด้วยวิธี Dimensional Modeling ซึ่งลักษณะของ Schema จะเหมือนเป็นรูปดาว หรือเราเรียกว่า Star Schema นั่นเอง**



ตัวอย่าง Data Warehouse (Relational Database ที่ออกแบบด้วยวิธี Normalization)

2.4 ปัญหาของ Database

ระบบคลังข้อมูล (Data Warehouse)

- ให้คุณลองใช้ SQL Command Query ข้อมูลจาก Data Warehouse นี้เพื่อตอบคำถามแบบเดียวกับที่ทำไปเมื่อซักครู่นะครับ
- ยอดขายของสินค้าที่มาจากบริษัท ABC ในปี 2015:
  SELECT Product_Dim.Name AS ItemName, Purchases_Fact.Total AS TotalSale FROM Purchase_Fact INNER JOIN Product_Dim ON Purchase_Fact.ProductKey = Product_Dim.ProductKey INNER JOIN Time_Dim ON Purchase_Fact.TimeKey = Time_Dim.TimeKey WHERE Product_Dim.SupplierName = 'ABC' AND Time_Dim.TimeYear = 2015
- สรุปยอดขายรายปีของ 5 ปีล่าสุด:
  SELECT Time_Dim.TimeYear AS Year, Total FROM Purchases_Fact INNER JOIN Time_Dim ON Purchase_Fact.TimeKey = Time_Dim.TimeKey ORDER BY Time_Dim.TimeYear

2.4 ปัญหาของ Database

ระบบคลังข้อมูล (Data Warehouse) : สังเกตได้ว่าจากตัวอย่างนี้

1. SQL Command ยาวและซับซ้อนน้อยลง ไม่มีการเรียกใช้ Function แปลก ๆ แล้ว (อาจจะเห็นว่า SQL มันยาวพอ ๆ กัน แต่ชื่อ Table ใน Data Warehouse มันยาวกว่านะ 55555)
2. จำนวนการ Join ลดลงจาก Database ต้องใช้ Join ถึง 3 ครั้ง เหลือแค่ 2 ครั้ง ยิ่งไปกว่านั้น ต่อให้มีคำถามที่ต้อง Join ทุก Table เพื่อตอบคำถาม ใน Data Warehouse ข้างต้นจะ Join กันเพียงแค่ 3 Table เท่านั้น ผิดกับ Database ที่ต้อง Join กันถึง 5 ครั้งเลยทีเดียว
3. วิธีการ Query ค่อนข้างจะคล้ายคลึงกัน

- เหตุผลที่ Database ที่ออกแบบด้วยวิธี Normalization ไม่เหมาะสำหรับการนำมาวิเคราะห์ เพราะ Normalization ใช้เพื่อออกแบบ Database ที่ใช้ในการจัดเก็บข้อมูลที่รวดเร็ว ส่วน Data Warehouse ถูกออกแบบขึ้นมาเพื่อ Query ข้อมูลไปวิเคราะห์ได้ง่ายขึ้น คนที่ใช้ SQL ไม่คล่องก็สามารถ Query ข้อมูลไปวิเคราะห์เองได้

2.4 ปัญหาของ Database

ระบบคลังข้อมูล (Data Warehouse) : SQL : Structure Query Language

- DDL
- DML
- DCL

2.5 ประเภทของระบบสารสนเทศ IS Information System

ประเภทของระบบสารสนเทศ IS Information System

- TPS Transaction Processing System
  ระบบประมวลผลรายการ
- MIS ระบบสารสนเทศเพื่อการจัดการ
- DSS ระบบสนับสนุนการตัดสินใจ
- EIS ระบบสารสนเทศเพื่อผู้บริหาร
- ES ระบบผู้เชี่ยวชาญ

2.6 วัตถุประสงค์ของการทำเหมืองข้อมูล (Objective of Data Mining)

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่ใช้เครื่องมือทางคณิตศาสตร์และการวิเคราะห์ข้อมูลเพื่อค้นหาความรู้ที่มีอยู่ภายในข้อมูลที่มีปริมาณมาก วัตถุประสงค์ของการเหมืองข้อมูลมีหลายด้านและสามารถนำไปใช้ในหลายสาขาด้วยกัน

โดยวัตถุประสงค์ของการทำเหมืองข้อมูล (Objective of Data Mining) มีหลายประการที่มุ่งเน้นการใช้ประโยชน์จากข้อมูลในเชิงลึกเพื่อนำไปใช้ในการตัดสินใจและแก้ปัญหาต่าง ๆ โดยวัตถุประสงค์หลักมีดังนี้

1. ค้นหารูปแบบที่ซ่อนอยู่ในข้อมูล (Discover Hidden Patterns) การทำเหมืองข้อมูลช่วยระบุรูปแบบหรือแนวโน้มที่ซ่อนอยู่ในข้อมูลจำนวนมากที่อาจไม่สามารถมองเห็นได้ง่ายด้วยการวิเคราะห์แบบทั่วไป

2. การคาดการณ์ (Prediction) ใช้ข้อมูลในอดีตเพื่อพัฒนาแบบจำลองที่สามารถคาดการณ์เหตุการณ์ในอนาคต เช่น การพยากรณ์ยอดขาย การคาดการณ์ความเสี่ยง หรือพฤติกรรมของลูกค้า

3. การจัดกลุ่ม (Clustering) แบ่งกลุ่มข้อมูลตามลักษณะหรือความคล้ายคลึงกัน เช่น การแบ่งลูกค้าตามพฤติกรรมการซื้อ

4. การจำแนกประเภท (Classification) สร้างโมเดลเพื่อจัดข้อมูลใหม่ให้อยู่ในประเภทที่เหมาะสม เช่น การพิจารณาว่าลูกค้าคนใดจะชำระหนี้หรือไม่

2.6 วัตถุประสงค์ของการทำเหมืองข้อมูล (Objective of Data Mining)

5. การเชื่อมโยงความสัมพันธ์ (Association Rule Discovery) ค้นหาความสัมพันธ์ระหว่างข้อมูล เช่น การวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) เพื่อดูว่าสินค้าใดมักจะถูกซื้อร่วมกัน

6. การตรวจจับความผิดปกติ (Anomaly Detection)
ระบุข้อมูลที่แตกต่างหรือไม่เป็นไปตามปกติ เช่น การตรวจจับการทุจริตทางการเงิน

7. การเพิ่มประสิทธิภาพการตัดสินใจ (Improved Decision-Making)
นำผลลัพธ์จากการวิเคราะห์ข้อมูลไปใช้ในการตัดสินใจเชิงกลยุทธ์หรือปรับปรุงกระบวนการทางธุรกิจ

8. การวิเคราะห์แนวโน้มและการเปลี่ยนแปลง (Trend and Change Analysis)
ตรวจสอบการเปลี่ยนแปลงในข้อมูลเพื่อระบุแนวโน้มที่สำคัญ เช่น การเปลี่ยนแปลงความนิยมของผลิตภัณฑ์

9. การเพิ่มมูลค่าให้กับข้อมูล (Data Monetization)
แปลงข้อมูลให้เป็นสินทรัพย์ที่มีมูลค่าโดยใช้ข้อมูลในการพัฒนาแผนธุรกิจและกลยุทธ์การตลาด

2.6 วัตถุประสงค์ของการทำเหมืองข้อมูล (Objective of Data Mining)

1. เพื่อการค้นพบองค์ความรู้ใหม่ในฐานข้อมูล (Knowledge discovery in databases)
2. เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่ (Knowledge extraction)
3. เพื่อจัดการกับข้อมูลในอดีต (Data archeology)
4. เพื่อสำรวจข้อมูล (Data exploration)
5. เพื่อค้นหา Pattern ของข้อมูลที่ซ่อนอยู่ (Data pattern processing)
6. เพื่อใช้ขุดเจาะข้อมูล (Data dredging)
7. เพื่อเก็บเกี่ยวผลประโยชน์ให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

2.7 ประโยชน์ของเหมืองข้อมูล

การเหมืองข้อมูลมีประโยชน์มากมายในหลายด้านของชีวิตประจำวันและธุรกิจ. นี่คือบางประโยชน์ที่สำคัญของการเหมืองข้อมูล:

1. **ค้นพบความรู้ที่ซ่อนอยู่:** การเหมืองข้อมูลช่วยในการค้นพบความรู้และแนวโน้มที่อาจซ่อนอยู่ภายในข้อมูลที่มีปริมาณมาก. นี้สามารถช่วยในการทำนายแนวโน้มอนาคต, ความสัมพันธ์ระหว่างตัวแปร, หรือปัจจัยที่ส่งผลต่อผลลัพธ์ทางธุรกิจ.

2. **การทำนายและการตัดสินใจทางธุรกิจ:** การเหมืองข้อมูลช่วยในการสร้างแบบจำลองทางสถิติหรือการเรียนรู้ของเครื่องเพื่อทำนายผลลัพธ์ ซึ่งสามารถนำไปใช้ในการตัดสินใจทางธุรกิจอย่างมีมูลค่า.

3. **การตรวจจับและป้องกันปัญหา:** การเหมืองข้อมูลช่วยในการตรวจจับแนวโน้มที่ไม่ปกติหรือข้อมูลที่ผิดปกติที่อาจชี้ว่ามีปัญหาเกิดขึ้น, ทั้งในด้านความปลอดภัย, การฉ้อโกง, หรือปัญหาอื่น ๆ.

4. **การปรับแต่งการตลาด:** การเหมืองข้อมูลช่วยในการทำนายและเข้าใจพฤติกรรมของลูกค้า, ซึ่งสามารถนำไปใช้ในการปรับแต่งแผนการตลาด เพื่อเพิ่มประสิทธิภาพและการตอบสนอง.

2.7 ประโยชน์ของเหมืองข้อมูล (ต่อ)

5. **การจัดกลุ่มลูกค้า (Customer Segmentation):** การเหมืองข้อมูลช่วยในการแบ่งกลุ่มลูกค้าตามลักษณะที่คล้ายกัน เพื่อสร้างกลยุทธ์การตลาดที่เป็นเอกลักษณ์สำหรับแต่ละกลุ่ม.
6. **การพัฒนาสินค้าและบริการ:** การเหมืองข้อมูลช่วยในการเข้าใจความต้องการและความพึงพอใจของลูกค้า, ทำให้ธุรกิจสามารถพัฒนาสินค้าและบริการได้ตรงกับความต้องการ.
7. **การปรับแต่งประสบการณ์ลูกค้า:** การเหมืองข้อมูลช่วยในการปรับแต่งประสบการณ์ลูกค้า, ทำให้ธุรกิจสามารถให้บริการและปรับปรุงประสบการณ์ลูกค้าได้ดียิ่งขึ้น.
8. **การวิเคราะห์การทำงานภายในองค์กร:** การเหมืองข้อมูลช่วยในการวิเคราะห์แนวโน้มและประสิทธิภาพของการทำงานภายในองค์กร, ทำให้สามารถทำการปรับปรุงและประสิทธิภาพได้.

การเหมืองข้อมูลมีผลทำให้ธุรกิจสามารถปรับตัวและตอบสนองต่อสถานการณ์ที่เปลี่ยนไปอย่างรวดเร็วขึ้น, ทำให้มีการตัดสินใจที่มีประสิทธิภาพและเหมาะสมกับสภาพแวดล้อมการทำธุรกิจในปัจจุบัน.

2.7 ประโยชน์ของ Data Mining

เหมืองข้อมูลเป็นกระบวนการที่มีประโยชน์อย่างมากในหลายด้าน โดยช่วยสนับสนุนการตัดสินใจและการคาดการณ์ผลลัพธ์จากการตัดสินใจได้อย่างมีประสิทธิภาพ นอกจากนี้ยังช่วยเพิ่มความเร็วในการวิเคราะห์ข้อมูลในฐานข้อมูลขนาดใหญ่ที่มีความซับซ้อน ทำให้สามารถประมวลผลและดึงข้อมูลที่มีค่าออกมาใช้งานได้อย่างรวดเร็วและแม่นยำ

กระบวนการทำเหมืองข้อมูลยังช่วยค้นหาส่วนประกอบที่ซ่อนอยู่ภายในเอกสารหรือข้อมูล รวมถึงความสัมพันธ์ระหว่างส่วนประกอบเหล่านั้น ทำให้องค์กรสามารถเข้าใจข้อมูลได้อย่างลึกซึ้งยิ่งขึ้น อีกทั้งยังส่งเสริมการเชื่อมโยงและบูรณาการระหว่างหน่วยงานต่าง ๆ ภายในองค์กร ทำให้การทำงานมีความสอดคล้องและมีประสิทธิภาพมากขึ้น

อีกหนึ่งประโยชน์สำคัญคือการจัดกลุ่มข้อมูล เช่น การจัดกลุ่มลูกค้าของบริษัทประกันภัยที่มีลักษณะการเกิดอุบัติเหตุคล้ายคลึงกัน เพื่อวางแผนดำเนินการตามนโยบายหรือมาตรการที่เหมาะสม ทั้งหมดนี้ช่วยให้องค์กรสามารถบริหารจัดการข้อมูลและทรัพยากรได้อย่างมีประสิทธิภาพและนำไปสู่การพัฒนาธุรกิจอย่างยั่งยืน

2.8 เทคนิคที่ใช้ในเหมืองข้อมูล

เทคนิคที่ใช้ใน เหมืองข้อมูล (Data Mining Techniques) มีหลายวิธีขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล ซึ่งสามารถแบ่งออกเป็นหมวดหลัก ๆ ดังนี้:

🧠 1. การจำแนกประเภท (Classification)

ใช้ในการทำนายว่าข้อมูลใหม่จะอยู่ในกลุ่มใด โดยอาศัยข้อมูลในอดีตที่มีการจัดกลุ่มไว้แล้ว

ตัวอย่างเทคนิค:

- Decision Tree (ต้นไม้ตัดสินใจ)
- Naïve Bayes
- k-Nearest Neighbors (k-NN)
- Support Vector Machine (SVM)
- Artificial Neural Networks (ANN)

ตัวอย่างการใช้งาน: การทำนายว่าลูกค้าจะ “ยกเลิกการใช้บริการ” หรือไม่

📊 2. การจัดกลุ่ม (Clustering)

ใช้เพื่อจัดกลุ่มข้อมูลที่มีลักษณะคล้ายกัน โดยไม่รู้กลุ่มล่วงหน้า

ตัวอย่างเทคนิค:

- K-Means Clustering
- Hierarchical Clustering
- DBSCAN

ตัวอย่างการใช้งาน:

แบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้า

2.8 เทคนิคที่ใช้ในเหมืองข้อมูล

เทคนิคที่ใช้ใน เหมืองข้อมูล (Data Mining Techniques) มีหลายวิธีขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล ซึ่งสามารถแบ่งออกเป็นหมวดหลัก ๆ ดังนี้:

🔍 3. การค้นหาความสัมพันธ์ (Association Rule Mining)

ใช้ในการค้นหารูปแบบหรือความสัมพันธ์ระหว่างข้อมูล เช่น สินค้าที่มักจะถูกซื้อพร้อมกัน

ตัวอย่างเทคนิค:

- Apriori Algorithm
- FP-Growth Algorithm

ตัวอย่างการใช้งาน:

การวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis)

→ “หากลูกค้าซื้อขนมปัง มักจะซื้อนมด้วย”

📈 4. การทำนาย (Prediction)

ใช้เพื่อทำนายค่าตัวเลขในอนาคตโดยอิงจากข้อมูลในอดีต

ตัวอย่างเทคนิค:

- Linear Regression
- Nonlinear Regression
- Time Series Analysis
- Random Forest Regressor

ตัวอย่างการใช้งาน:

การทำนายยอดขายในเดือนถัดไป

2.8 เทคนิคที่ใช้ในเหมืองข้อมูล

เทคนิคที่ใช้ใน เหมืองข้อมูล (Data Mining Techniques) มีหลายวิธีขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล ซึ่งสามารถแบ่งออกเป็นหมวดหลัก ๆ ดังนี้:

🧩 5. การตรวจจับความผิดปกติ (Anomaly Detection / Outlier Detection)

ใช้เพื่อค้นหาข้อมูลที่มีค่าผิดปกติ หรือแตกต่างจากแนวโน้มปกติของข้อมูล

ตัวอย่างเทคนิค:

- **Z-score**
- **Isolation Forest**
- **DBSCAN (สำหรับ outlier)**

ตัวอย่างการใช้งาน:

ตรวจจับการทุจริตทางการเงิน หรือธุรกรรมที่น่าสงสัย

🕵️‍♂️ 6. การลดมิติ (Dimensionality Reduction)

ใช้เพื่อลดจำนวนตัวแปร โดยยังคงรักษาความสำคัญของข้อมูลไว้

ตัวอย่างเทคนิค:

- **Principal Component Analysis (PCA)**
- **t-SNE**
- **Autoencoder**

ตัวอย่างการใช้งาน:

การลดขนาดข้อมูลภาพ หรือข้อมูลที่มีตัวแปรจำนวนมากก่อนนำไปวิเคราะห์

2.8 เทคนิคที่ใช้ในเหมืองข้อมูล

เทคนิคที่ใช้ใน เหมืองข้อมูล (Data Mining Techniques) มีหลายวิธีขึ้นอยู่กับวัตถุประสงค์ของการวิเคราะห์ข้อมูล ซึ่งสามารถแบ่งออกเป็นหมวดหลัก ๆ ดังนี้:

📉 **7. การสรุปและทำเหมืองลำดับ (Summarization & Sequential Pattern Mining)**

ใช้เพื่อหาลำดับเหตุการณ์หรือแนวโน้มของข้อมูล

ตัวอย่างเทคนิค:

- **Sequential Pattern Analysis**
- **Time Series Pattern Mining**

ตัวอย่างการใช้งาน:

การวิเคราะห์ลำดับการซื้อสินค้าของลูกค้า หรือการวิเคราะห์พฤติกรรมในเว็บไซต์

2.8 เทคนิคที่ใช้ในเหมืองข้อมูล

Algorithm of Data Mining
Descriptive Modeling : Unsupervised Learning

1. Association Algorithm
2. Clustering Algorithm
3. Time Series Algorithm

Algorithm of Data Mining
Predictive Modeling : Supervised Learning
Classification

4. Decision Trees Algorithm
5. Naïve Bayes Algorithm
6. Neural Network Algorithm

Regression

7. Linear Regression Algorithm
8. Logistic Regression Algorithm

2.8.1 Descriptive Modeling : Unsupervised Learning

2.8.1 Descriptive Modeling : Unsupervised Learning

unsupervised learning เป็นกลุ่ม algorithm ที่ไม่มี label หรือการสอนอย่างชัดเจนว่าถ้าทำงานแล้วได้ผลลัพธ์แบบนี้หมายถึงถูกหรือผิด ตัวอย่างเช่นถ้ามีคน 100 คนในห้องแล้วให้เราทำการแบ่งกลุ่ม 100 คนนั้นเป็น 2 กลุ่ม ในกรณีนี้ไม่มีกฎชัดเจนว่าเราต้องแบ่งตามเพศ , อายุ , สีเสื้อหรืออื่นๆ แค่แบ่งออกมาให้ได้ 2 กลุ่ม ซึ่งเราอาจจะแบ่งว่าใครอยู่กลุ่มห้องส่วนหน้าคือกลุ่ม 1 ส่วนใครอยู่ครึ่งห้องหลังคือกลุ่ม 2

อีก 1 ตัวอย่างของ algorithm ในกลุ่ม unsupervised learning คือการลดรูปตัวแปร ตัวอย่างเช่นเรามีข้อมูลของการเข้าใช้งาน website ของ user แบ่งตามชั่วโมงการเข้าใช้งานในรูปแบบ tabular ดังนี้

เป็นการมีตัวอย่างของคำตอบไว้ให้ดูเป็นแนว

2.8.1 Descriptive Modeling : Unsupervised Learning

2.8.1 Descriptive Modeling : Unsupervised Learning

	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00
1	0	1	1	0	0	0	0	0
2	0	0	0	2	0	0	0	0
3	0	0	0	3	0	5	0	0
4	0	2	0	4	0	0	0	0
5	0	0	2	0	0	0	0	0
6	0	0	1	0	0	2	0	0
7	0	4	0	0	2	0	0	0
8	0	0	4	0	0	0	0	0
9	0	0	0	0	0	2	0	0
10	0	0	1	0	0	0	5	0
11	0	2	0	0	0	0	0	0
12	0	0	0	0	5	0	0	0
13	3	0	0	0	0	0	5	0
14	0	3	2	0	0	0	0	0
15	0	0	3	0	3	3	3	0

ตารางที่ 2.1 ตัวอย่างข้อมูลของการเข้าใช้งาน website ของ user แบ่งตามชั่วโมง

2.8.1 Descriptive Modeling : Unsupervised Learning

- **Descriptive Modeling : Unsupervised Learning**
- **ถ้าเราต้องการ plot ข้อมูลดังกล่าวในรูปแบบของ scatter plot เราจำเป็นต้องลดตัวแปรจาก 24 ตัว (ตามหลายชั่วโมง 00:00–23:00) ให้เหลือแค่ 2 ตัวแปรคือแกน x กับ y เพื่อใช้งานการ plot ผลที่ได้ก็จะออกมาตามดังนี้**
- **จะเห็นได้ว่าการทำงานแบบ unsupervised learning จะเป็นการทำงานแบบไม่มี label หรือกฎที่ตายตัว จากตัวอย่าง 2 ตัวอย่างข้างต้นจะเห็นว่าทั้งการแบ่งคนในห้องเป็น 2 กลุ่มหรือการลดรูปข้อมูลจาก 24 columns มาเป็น 2 columns เกิดจากการกำหนดตัวเลขขึ้นมาเองตามที่ผู้สร้างเห็นว่าเหมาะสมทั้งสิ้น**

ที่มา : https://medium.com/coeffest/ประเ ภทของ-machine-learning-f3159fee7b56

	x	y
1	5	37
2	-21	-6
3	45	37
4	-15	-45
5	34	-3
6	-19	-39
7	-9	30
8	26	-9
9	1	-38
10	50	-10
11	-45	-38
12	39	43
13	-44	-46
14	-23	-24
15	46	8

ตารางที่ 2.2 ตัวอย่างหลังจากลดตัวแปร (เป็นรูปจำลองที่ไม่ผ่านการประมวลลผลจริง)

2.8.1 Descriptive Modeling : Unsupervised Learning

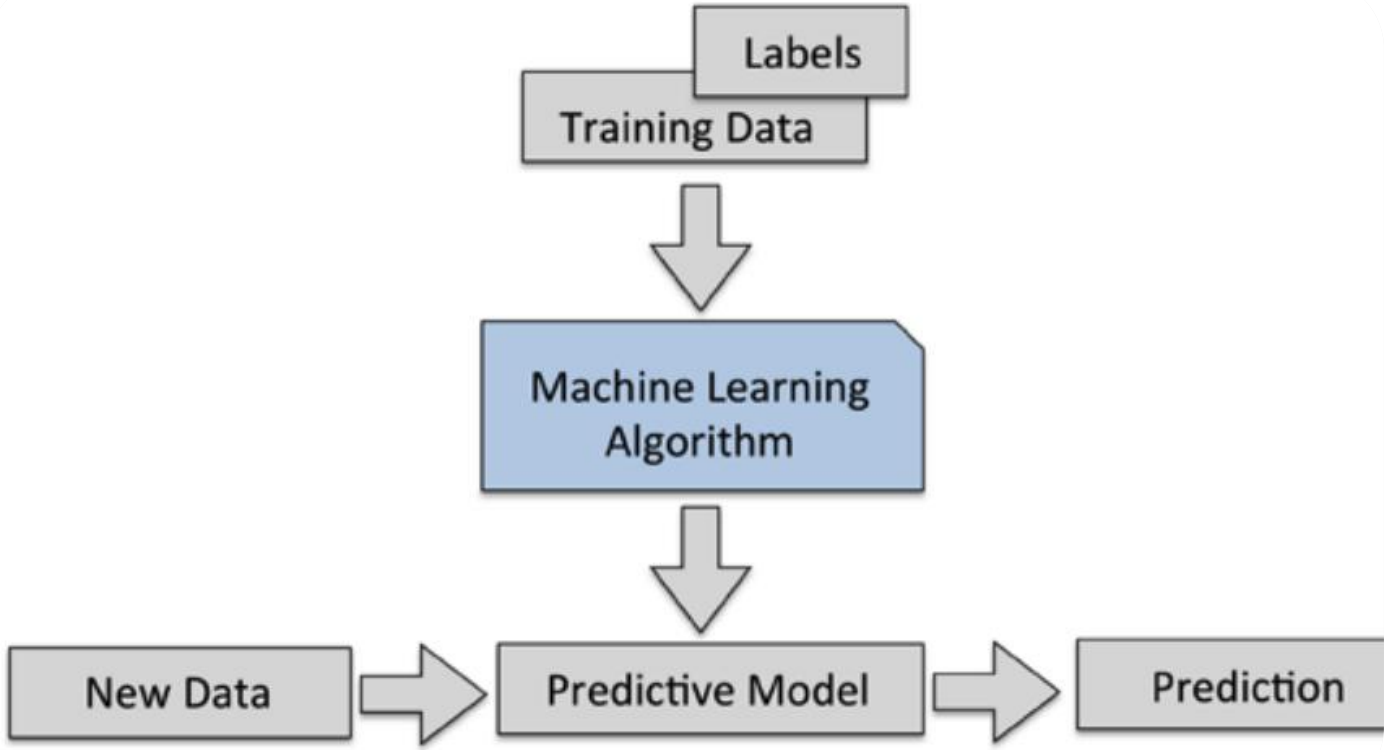
1. Association Algorithm กฎความสัมพันธ์ เป็นอัลกอริทึมการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ (Big Data) เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่าง ๆ หรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้าที่เรียกว่า “Market Basket Analysis” โดยนำ transaction การซื้อสินค้ามาทำการค้นหาวิเคราะห์ ว่าลูกค้าใช้อะไรคู่กับสินค้าอะไรบ่อย ๆ ทำให้สามารถออกโปรโมชั่นแคมเปญคู่กัน ทำให้มีราคาถูกลง เพื่อเพิ่มมูลค่าให้กับสินค้า ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์ ”(Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

2. Clustering Algorithm เป็นเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูลใหม่ที่มีลักษณะคล้ายกันไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (similarity) หรือ ความใกล้ชิด (proximity) โดยคำนวณจากการวัดระยะระหว่างเวกเตอร์ของข้อมูลเข้า ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ กลุ่มผู้มีรายได้สูง (>80,000 บาท) , กลุ่มผู้มีรายได้ปานกลาง (25,000 ถึง 80,000 บาท), กลุ่มผู้มีรายได้ต่ำ (less than 25,000 บาท) หรือจำแนกกลุ่มลูกค้าช่วงตามช่วงอายุ,เพศ วิเคราะห์ปัจจัยเสี่ยงที่มีโอกาศเกิดโรคต่าง ๆ เพื่อจัดแคมเปญเสนอขายประกันคุ้มครองชีวิตได้ตรงกลุ่มเป้าหมาย

3. Time Series Algorithm เป็นวิธีการพยากรณ์แบบข้อมูลอนุกรมเวลา ซึ่งถ้าจะเริ่มศึกษา algorithm นี้ก็จะเน้นไปในเรื่องการพยากรณ์การขาย (Sales forecasting) ก็คือ การประมาณ หรือ การคาดคะเนว่าอะไรจะเกิดขึ้นในอนาคต เช่น การพยากรณ์ยอดขายของ 3 ปีข้างหน้า การพยากรณ์มีบทบาทสำคัญกับทุกด้าน ทั้งหน่วยงานของรัฐบาล และเอกชน รัฐบาลต้องประมาณ หรือ พยากรณ์รายได้รายจ่ายในปีหน้า เพื่อนำมาวางแผน เอกชนต้องพยากรณ์ยอดขาย เพื่อนำมาวางแผนการผลิต สินค้าคงคลัง แรงงาน เป็นต้น

2.8.2 (Predictive Modeling : Supervised Learning)

2) การสร้างตัวแบบทำนายผล(Predictive Modeling : Supervised Learning) หรือเรียกอีกอย่างหนึ่งว่า การเรียนรู้แบบมีผู้สอน คือการนำเอ่ข้อมูลในอดิดมาสร้างตัวแบบ (Model) เพื่อการทำนายผลในอนาคต โดยมีการใช้ชุดข้อมูลเรียนรู้ (Training Data) ซึ่งข้อมูลทุกตัวจะมีคุณสมบัติที่ใช้ในการทำนายผลข้อมูลอัลกอริทึมประเภทนี้จะใช้การแบ่งข้อมูลออกเป็นกลุ่มตามคุณสมบัติ



ภาพที่ 2. การเรียนรู้แบบมีผู้สอน

ที่มา : https://medium.com/@natthawatphongchit/machine-learning-basics-2b38700cb10b

2.8.2 (Predictive Modeling : Supervised Learning)

การเรียนรู้แบบมีผู้สอน (Supervised Learning)

รูปแบบการเรียนรู้แบบมีผู้สอนโดยทั่วไป จะอยู่ในลักษณะการทำนายผลลัพธ์ เช่น จะทำนายว่าตัวอักษรในภาพคืออะไร หรือรายได้ที่ควรคาดหวังจากการลงทุนเป็นเท่าใด

“การมีผู้สอน” ในที่นี้หมายความว่า ในข้อมูลที่ใช้ฝึก (Training Data) เราจะมีมนุษย์มาคอยแยกประเภทหรือบอกผลลัพธ์ (Label) ที่ควรจะเป็นไปไว้ด้วย จากนั้นเราจะนำข้อมูลที่ใช้ฝึก (ที่มี Label) ไปผ่านอัลกอริทึมสำหรับสร้างโมเดลที่ไว้ทำนายผลลัพธ์

เมื่อเราได้โมเดลที่ไว้ทำนายแล้ว เราจะนำข้อมูลใหม่ที่เครื่องไม่เคยเห็น กล่าวคือไม่ใช่ข้อมูลชุดเดียวกันกับข้อมูลฝึกหัด เครื่องจะต้องทำนาย (Predict) ว่าคำตอบที่ได้ควรจะเป็นอะไร

2.8.2 (Predictive Modeling : Supervised Learning)

การเรียนรู้แบบมีผู้สอน (Supervised Learning)

“ภาพด้านบนคือ รูปแบบพื้นฐานของการเรียนรู้แบบมีผู้สอน เรามาดูกันว่าแต่ละส่วนคืออะไร

1. Labels เป็นส่วนที่ผู้สอนกำหนดให้ว่าจากข้อมูลที่ให้นั้นคำตอบที่ถูกคืออะไร
2. Predictive Model ผลจากการฝึกคือโมเดลที่ใช้ในการทำนายผลลัพธ์
3. New Data ข้อมูลที่เครื่องไม่เคยเห็น (ใส่เข้ามาตอนใช้งานจริง หรือตอนที่ต้องการวัดประสิทธิภาพของโมเดล)
4. Prediction ผลที่เครื่องทำนายออกมา

- “มีผู้สอน” ในที่นี้หมายความว่า ในข้อมูลที่ใช้ฝึก (Training Data) เราจะมีมนุษย์มาคอยแยกประเภทหรือบอกผลลัพธ์ (Label) ที่ควรจะเป็นไปไว้ด้วย จากนั้นเราจะนำข้อมูลที่ใช้ฝึก (ที่มี Label) ไปผ่านอัลกอริทึมสำหรับสร้างโมเดลที่ไว้ทำนายผลลัพธ์
- เมื่อเราได้โมเดลที่ไว้ทำนายแล้ว เราจะนำข้อมูลใหม่ที่เครื่องไม่เคยเห็น กล่าวคือไม่ใช่ข้อมูลชุดเดียวกันกับข้อมูลฝึกหัด เครื่องจะต้องทำนาย (Predict) ว่าคำตอบที่ได้ควรจะเป็นอะไร

2.8.2 (Predictive Modeling : Supervised Learning)

- *Classification*

4. Decision Trees Algorithm เป็นการแยกข้อมูล (Classification) ออกเป็นกลุ่มโดยใช้คุณสมบัติของข้อมูล (Attribute) เป็นตัวกำหนด ซึ่งประกอบไปด้วย โหนดภายใน (Internal node), กิ่ง (Link), โหนดใบ (Leaf node) วิธีการวิเคราะห์แบบต้นไม้ตัดสินใจเป็นการค้นหาจากบนลงล่าง(Top-down)โดยเริ่มจากการเลือกคุณสมบัติที่ดีที่สุดมาเป็นโหนดราก (Root node) และวนสร้างโหนดลูกและเส้นเชื่อมไปเรื่อยๆจนกว่าข้อมูลที่ได้จะถูกจัดไว้เป็นกลุ่มเดียวกันเราถึงจะหยุดสร้างต้นไม้ แนะนำอัลกอริทึมที่ใช้สร้าง decision tree ได้แก่ ID3, C4.5, C5.0, CART algorithm

5. Naive Bayes Algorithm

6. Neural Network Algorithm เป็นแนวคิดที่ได้มาจากการจำลองการทำงานของเซลล์สมองของมนุษย์ ซึ่งมีโครงสร้างประกอบด้วย Input layer, Hidden layer, Output layer มีหน่วยย่อยเรียกว่า Perceptron ซึ่งเทียบเท่าได้กับเซลล์สมองของมนุษย์หนึ่ง Neuronโดยหลักการของ neural network จะมีการกำหนดค่าน้ำหนัก weight และ threshold ให้แก่ input แต่ละตัวโดยใช้ back-propagation algorithm ในการคำนวณ และในการสร้างโมเดล Neural network สามารถทำได้ทั้ง 2 วิธีคือ Supervised Learning และ Unsupervised Learning ซึ่งเราสามารถนำมาประยุกต์ใช้กับงานด้านต่าง ๆ อาทิเช่น การพยากรณ์ การจดจำใบหน้า เรียนรู้จำลายมือ ลายเซนต์ ใช้ในทางการแพทย์

- *Regression*

7. Linear Regression Algorithm
8. Logistic Regression Algorithm

2.9 การประยุกต์ใช้งานเหมืองข้อมูล (Data Mining)

- ปัจจุบันได้มีการนำ เทคนิคการทำเหมืองข้อมูล (Data Mining) มาใช้ในธุรกิจอย่างหลากหลายสามารถทำให้การบริหารจัดการเป็นไปอย่างง่ายดาย และสามารถวางแผนวิเคราะห์การตลาดเพื่อให้ได้ทราบสิ่งที่เป็นประโยชน์ อาทิเช่น

- ธุรกิจค้าปลีกสามารถใช้งาน Data Mining ในการพิจารณาหากลยุทธ์ให้เป็นที่สนใจกับผู้บริโภคในรูปแบบต่าง ๆ เช่น ที่ว่างในชั้นวางของจะจัดการอย่างไรถึงจะเพิ่มยอดขายได้ เช่นที่ Midas ซึ่งเป็นผู้แทนจำหน่ายอะไหล่สำหรับอุตสาหกรรมรถยนต์ งานที่ต้องทำคือการจัดการกับข้อมูลที่ได้รับจากสาขาทั้งหมด ซึ่งจะต้องทำการรวบรวมและวิเคราะห์อย่างทันท่วงที

- กิจการโทรคมนาคม เช่นที่ Bouygues Telecom ได้นำมาใช้ตรวจสอบการโกงโดยวิเคราะห์รูปแบบการใช้งานของสมาชิกลูกข่ายในการใช้งานโทรศัพท์ เช่น คาบเวลาที่ใช้จุดหมายปลายทาง ความถี่ที่ใช้ ฯลฯ และคาดการณ์ข้อบกพร่องที่เป็นไปได้ในการชำระเงิน เทคนิคนี้ยังได้ถูกนำมาใช้กับลูกค้าโทรศัพท์เคลื่อนที่ซึ่งระบบสามารถตรวจสอบได้ว่าที่ใดที่เสี่ยงที่จะสูญเสียลูกค้าสูงในการแข่งขัน France Telecom ได้ค้นหาวิธีรวมกลุ่มผู้ใช้ให้เป็นหนึ่งเดียวด้วยการสร้างแรงดึงดูดในเรื่องค่าใช้จ่ายและพัฒนาเรื่องความจงรักภักดีต่อตัวสินค้า

2.9 การประยุกต์ใช้งานเหมืองข้อมูล (Data Mining)

การตลาด

- การทำนายผลการตอบสนองกับการเปิดตัวสินค้าใหม่ เปิดตัว iPhone15 proMax
- การทำนายยอดขายเมื่อมีการลดราคาสินค้า
- การทำนายกลุ่มลูกค้าที่น่าจะใช้สินค้าของเรา

การเงินการธนาคาร

- การคาดการณ์ถึงโอกาสในการชำระหนี้ของลูกค้าว่าสูงเท่าไหร่?
- ค้นหาลูกค้าขาดคุณภาพ เพื่อหลีกเลี่ยงความเสี่ยงในการปล่อยกู้
- ค้นหาลูกค้าชั้นดี เพื่อเสนอการปล่อยกู้
- ทำนายแนวโน้มของพฤติกรรมการใช้บัตรเครดิต

สถานีโทรทัศน์หรือวิทยุ

- ค้นหารายการที่ดีและเหมาะสมต่อช่วงเวลาที่สุด เพื่อวางผังรายการในแต่ละเดือน

ฮาร์ดแวร์และซอฟต์แวร์คอมพิวเตอร์

- ค้นหาช่วงเวลาที่เหมาะสมกับการผลิตชิพคอมพิวเตอร์ตัวใหม่ เพื่อป้อนสู่ตลาด
- การทำนายอายุการใช้งานของ Disk Drive หรือ อุปกรณ์ต่าง ๆ

2.9 การประยุกต์ใช้งานเหมืองข้อมูล (Data Mining)

- การวิเคราะห์ผลิตภัณฑ์ เก็บรวบรวมลักษณะและราคาของผลิตภัณฑ์ทั้งหมดสร้างโมเดลด้วยเทคนิค Data Mining และใช้โมเดลในการทำนายราคาผลิตภัณฑ์ตัวอื่น ๆ

- การวิเคราะห์บัตรเครดิต
    - ช่วยบริษัทเครดิตการ์ดตัดสินใจในการที่จะให้เครดิตการ์ดกับลูกค้าหรือไม่
    - แบ่งประเภทของลูกค้าว่ามีความเสี่ยงในเรื่องเครดิต ต่ำ ปานกลาง หรือสูง
    - ป้องกันปัญหาเรื่องการทุจริตบัตรเครดิต

- การวิเคราะห์ลูกค้า
    - ช่วยแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะผลิตและเสนอสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม
    - ทำนายว่าลูกค้าคนใดจะเลิกใช้บริการจากบริษัทภายใน 6 เดือนหน้า

- การวิเคราะห์การขาย
    - พบว่า 70 % ของลูกค้าที่ซื้อโทรทัศน์แล้วจะซื้อวิดีโอตามมา ดังนั้นผู้จัดการจึงควรมุ่งไป ลูกค้าที่ซื้อโทรทัศน์ แล้วจึงส่งเมล์ไปยังลูกค้าเหล่านั้นเพื่อที่จะเชิญชวน หรือให้ข้อเสนอที่ดี เพื่อให้ลูกค้ามาซื้อวิดีโอในครั้งต่อไป
    - ช่วยในการโฆษณาสินค้าได้อย่างเหมาะสมและตรงตามเป้าหมาย
    - ช่วยในการจัดวางสินค้าได้อย่างเหมาะสม

- Text Mining

    เป็นการปรับใช้ Data Mining มาอยู่ในรูปของข้อมูลตัวอักษรซึ่งเป็นรูปแบบของภาษาเครื่อง SDP Infoware ตัวอย่างของงานคือใช้เป็นเครื่องมือตรวจระดับความพึงพอใจของผู้ที่เข้าชมนิทรรศการโดยผ่านการประมวลผลจากแบบสอบถาม

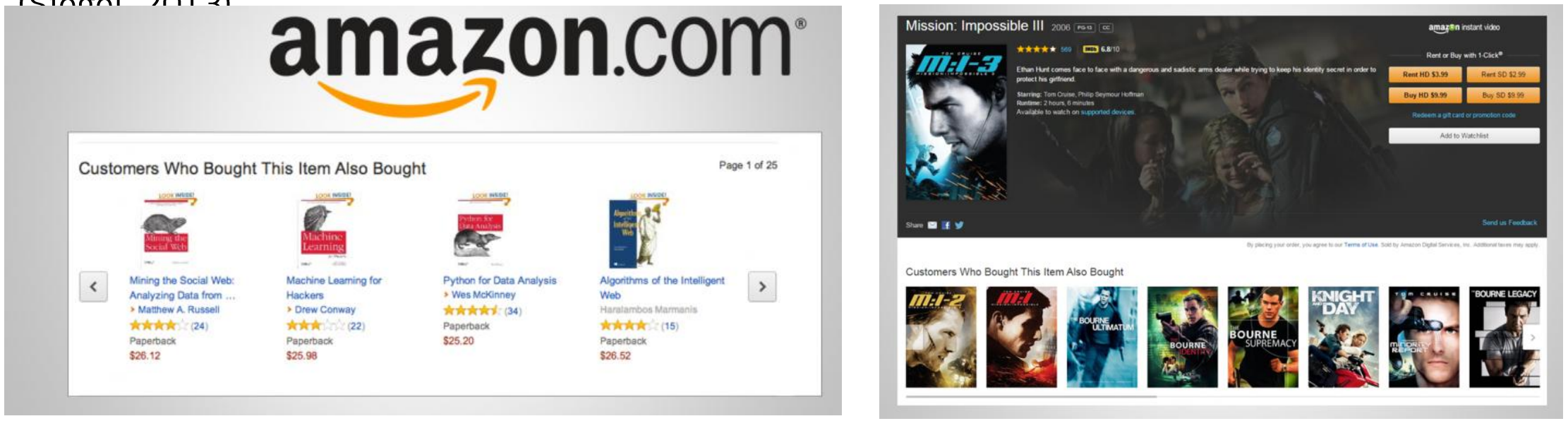
- E-Commerce
    - ช่วยให้เข้าใจพฤติกรรมของลูกค้า เช่น ลูกค้ามักเข้าไปที่ web ใดตามลำดับก่อนหลัง
    - ช่วยในการปรับปรุง web site เช่น พิจารณาว่าส่วนใดของ web ที่ควรปรับปรุงหรือควรเรียงลำดับการเชื่อมโยงในแต่ละหน้าอย่างไรเพื่อให้สะดวกกับผู้เข้าเยี่ยมชม

ตัวอย่าง การประยุกต์ใช้งานเหมืองข้อมูล (Data Mining)

ภาพที่ 3 : แสดงบัตรสมาชิกของห้างสรรพสินค้า เพื่อใช้ในการติดตามพฤติกรรมการบริโภคสินค้าของลูกค้า
(ที่มา http://open-miner.com/2009/11/03/introduction-datamining/)

ตัวอย่างธุรกิจที่นำเทคนิค Data Mining ไปประยุกต์ใช้งานเพื่อสร้างความได้เปรียบทางการแข่งขัน (Competitive Advantage)

amazon.com เป็นตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือ,หนังให้กับลูกค้าของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่ง ๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้านเรียกว่า Product Recommendations “amazon นั้นได้ส่วนแบ่งจากการทำการแนะนำสินค้าแบบนี้กว่าร้อยละ 35 จากยอดขายทั้งหมด” (Siegel, 2013)



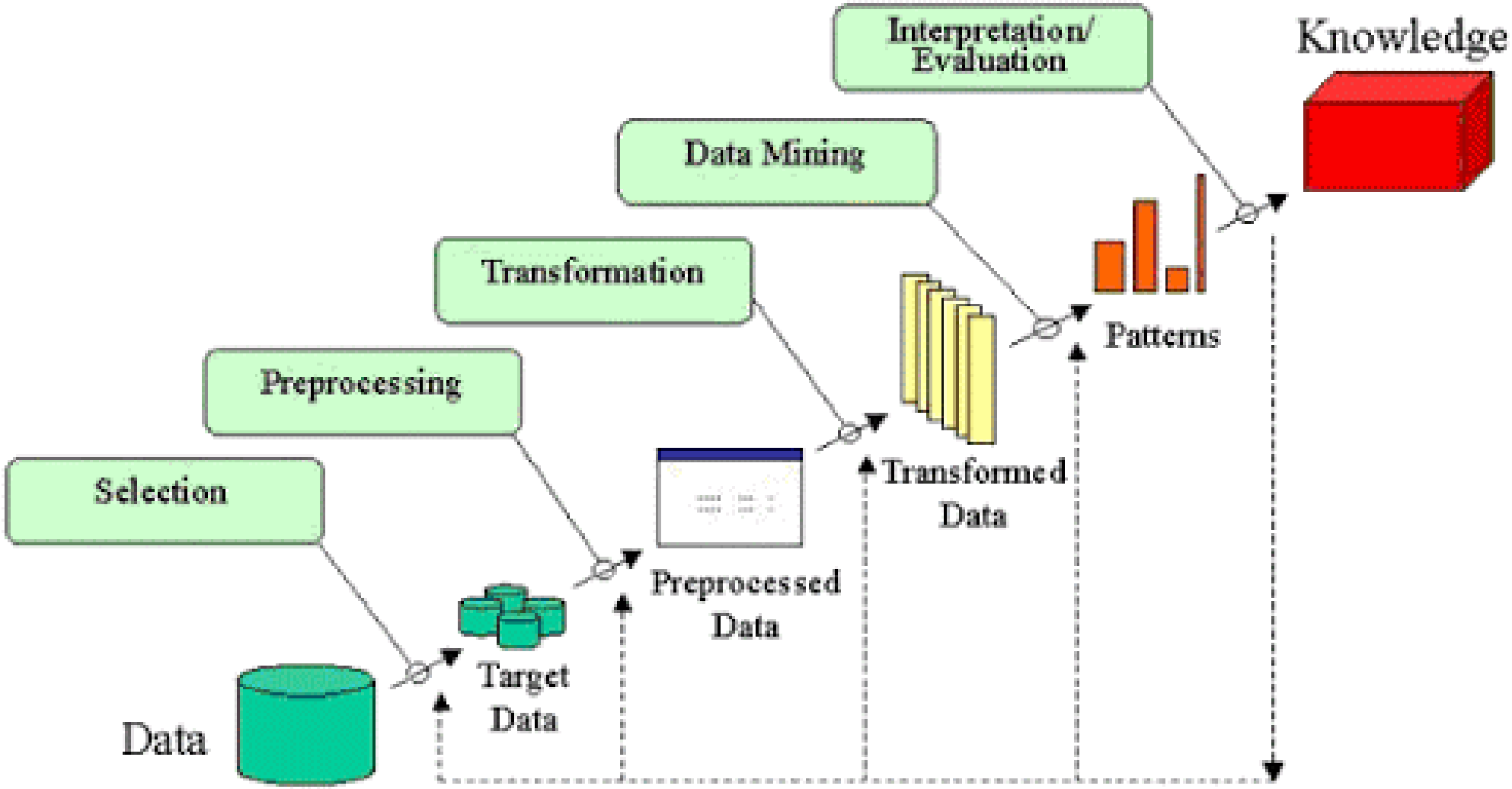
ภาพที่ 4. ตัวอย่างธุรกิจที่นำเทคนิค Data Mining

http://www.autosoft.in.th/data-warehouse/เหมืองข้อมูล-data-mining/

2.10 ขั้นตอนการทำเหมืองข้อมูล

1. Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป
2. Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน
3. Data Selection เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้
4. Data Transformation เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน
5. Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่
6. Pattern Evaluation เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล
7. Knowledge Representation เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ

2.10 ขั้นตอนการทำเหมืองข้อมูล



ภาพที่ 5. ขั้นตอนการทำเหมืองข้อมูล

2.11 รู้จักกับข้อมูล

คำว่า "ข้อมูล" ใช้เพื่ออธิบายข้อมูลทุกประเภทที่ถูกเก็บรวบรวม หรือประมวลผลเพื่อให้ได้ความหมายหรือความรู้บางประการ ข้อมูลมีหลายรูปแบบและสามารถเป็นตัวเลข, ข้อความ, ภาพ, เสียง, หรือรูปแบบอื่น ๆ ได้ตามลักษณะของข้อมูลนั้น ๆ

ข้อมูลมีความสำคัญมากในชีวิตประจำวันและในสายงานต่าง ๆ เพราะมันช่วยให้เราสามารถทำความเข้าใจเกี่ยวกับโลกที่รอบตัวได้มากขึ้น ข้อมูลที่ถูกประมวลผลอาจจะให้ข้อมูลที่มีประโยชน์เพื่อตัดสินใจหรือแก้ไขปัญหาต่าง ๆ ได้ดีขึ้น

ตัวอย่างของประเภทข้อมูล:

- ข้อมูลทางวิทยาศาสตร์: เช่น ข้อมูลที่ได้จากการวิจัย, การทดลอง
- ข้อมูลทางเศรษฐศาสตร์: เช่น ข้อมูลการเงิน, สถิติเศรษฐกิจ
- ข้อมูลทางสังคม: เช่น ข้อมูลประชากร, ข้อมูลสังคม
- ข้อมูลทางเทคโนโลยี: เช่น ข้อมูลเกี่ยวกับเทคโนโลยีสารสนเทศ
- ข้อมูลทางการแพทย์: เช่น ประวัติการรักษา, ข้อมูลสุขภาพ

การทำนุบำรุง, การจัดเก็บ, และการใช้ประโยชน์จากข้อมูลมีความสำคัญมากในยุคปัจจุบันที่เต็มไปด้วยข้อมูลจำนวนมากและทันสมัยที่สามารถนำมาใช้ในหลายแวดวงการได้

2.11.1 ข้อมูลในรูปแบบต่าง ๆ

หลังจากที่เราได้เห็นตัวอย่างการนำเทคนิคดาต้าไมน์นิงไปใช้งานกันบ้างแล้ว ถัดมาขอแนะนำเรื่องเกี่ยวกับข้อมูลและคำศัพท์ที่เกี่ยวข้องในการวิเคราะห์ข้อมูลที่จะได้พบในบทถัดไป โดยปกติแล้วข้อมูลที่สามารถนำมาวิเคราะห์ได้จะต้องเป็นแบบมีโครงสร้าง (structured data) หรืออยู่ในรูปแบบตาราง เช่น ข้อมูลของสมาชิก หรือ ข้อมูลการซื้อขายสินค้าต่าง ๆ แต่ในปัจจุบันเราก้าวเข้าสู่ยุคของบิ๊กดาต้า ซึ่งมีข้อมูล ขนาดมหาศาลแต่ข้อมูลเหล่านี้มักจะไม่มีโครงสร้าง (unstructured data) เช่น ข้อความต่าง ๆ ไม่ว่าจะอยู่ในอีเมล์ (e-mail) หรือเครือข่ายสังคมออนไลน์ (social network) ต่าง ๆ ถ้าเราต้องการนำข้อมูลที่ไม่มีโครงสร้าง นี้มาทำการวิเคราะห์ จำเป็นจะต้องทำการแปลงรูปแบบข้อมูลเหล่านี้ให้อยู่ในรูปแบบที่มีโครงสร้าง หรือ รูป แบบตารางเสียก่อน ในหนังสือเล่มนี้ผมจะเน้นไปที่ข้อมูลที่มีโครงสร้างแต่ขอแนะนำวิธีการแปลงข้อมูลที่ไม่มี โครงสร้างสักเล็กน้อย เรามาลองดูคำศัพท์ต่าง ๆ ที่เกี่ยวข้องกับข้อมูลทั้งสองแบบกันก่อนครับ

- ข้อมูลแบบมีโครงสร้าง (structured data)
- ข้อมูลแบบไม่มีโครงสร้าง (unstructured data)

2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)

ข้อมูลแบบมีโครงสร้างเป็นข้อมูลทั่วๆ ไปที่เรามักจะพบเห็นกันในรูปแบบของตาราง เช่น ข้อมูลที่เก็บรายละเอียดของสมาชิก หรือ การซื้อขายสินค้า โดยปกติแล้วข้อมูลเหล่านี้มักจะเก็บอยู่ในไฟล์ประเภท Excel หรือในฐานข้อมูลต่าง ๆ ตัวอย่างเช่น ข้อมูลสมาชิกที่แสดงในตารางที่ 2.3 ซึ่งประกอบด้วย 5 แถวและ 5 คอลัมน์ ในหนังสือที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงส่วนใหญ่จะเรียกข้อมูลแต่ละ "**แถว**" ว่า "**ตัวอย่าง** (example)" หรือ "**อินสแตนซ์** (instance)" และข้อมูลแต่ละ "**คอลัมน์**" ว่า "**แอตทริบิวต์** (attribute)" หรือ "**ฟีเจอร์** (feature)" ซึ่งในหนังสือเล่มนี้ผมขอเรียกข้อมูลในแต่ละแถวว่า **ตัวอย่าง** และแต่ละคอลัมน์ว่า **แอตทริบิวต์** ครับ และบรรทัดแรกในตารางที่ 1-1 คือชื่อของแต่ละ แอตทริบิวต์ ดังนั้นเราสามารถอ่านได้ว่าลูกค้าที่มีหมายเลขสมาชิกเป็น 2014-00001 ชื่อว่า สมชาย เป็นเพศ ชาย อายุ 33 ปีและมีรายได้ 33,000 บาท เป็นต้น

Column / Field / Attribute

Row / Record / Tuple

หมายเลขสมาชิก	ชื่อ	เพศ	อายุ	รายได้
2014-00001	สมชาย	ชาย	33	33,000
2014-00002	สมหญิง	หญิง	35	35,000
2014-00003	ปิติ	ชาย	23	23,000
2014-00004	มานี	หญิง	22	22,000
2014-00005	ชูใจ	หญิง	23	23,000

ตารางที่ 2.3 แสดงข้อมูลรายละเอียดของลูกค้า

2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)

ข้อมูลแบบมีโครงสร้าง (Structured Data) หมายถึงข้อมูลที่มีการจัดเรียงและจัดระเบียบตามรูปแบบที่กำหนดไว้อย่างชัดเจน มักจะอยู่ในรูปแบบของตาราง หรือฐานข้อมูลเชิงสัมพันธ์ (Relational Database) ประกอบด้วยแถว (Row) และคอลัมน์ (Column) ที่มีความหมายเฉพาะ ตัวอย่างของข้อมูลแบบมีโครงสร้าง ได้แก่

1. ข้อมูลพนักงาน

   - ชื่อ นามสกุล เพศ วันเกิด ตำแหน่ง เงินเดือน

2. ข้อมูลยอดขาย

   - วันที่ รหัสสินค้า ชื่อสินค้า จำนวนสินค้า ราคาต่อหน่วย ยอดขาย

3. ข้อมูลนักศึกษา

   - รหัสนักศึกษา ชื่อ-นามสกุล คณะ สาขาวิชา เกรดเฉลี่ย

4. ข้อมูลผู้ป่วย

   - ชื่อ-สกุล เพศ อายุ โรค อาการ ยาที่ได้รับ

5. ข้อมูลบัญชีธนาคาร

   - เลขที่บัญชี ชื่อเจ้าของบัญชี ประเภทบัญชี วันเปิดบัญชี ยอดคงเหลือ

ลักษณะสำคัญของข้อมูลแบบมีโครงสร้างคือ มีการกำหนดชนิดข้อมูล รูปแบบ และความสัมพันธ์ระหว่างข้อมูลต่างๆ ไว้อย่างชัดเจน ทำให้ง่ายต่อการค้นหา จัดเก็บ ปรับปรุง และวิเคราะห์ข้อมูล

2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)

ข้อมูลแบบมีโครงสร้าง (structured data) เป็นข้อมูลที่มีการจัดเก็บไว้ในรูปแบบหรือโครงสร้างที่ชัดเจน นักวิทยาศาสตร์ข้อมูลสามารถนำไปใช้ได้ง่าย เช่น ข้อมูลที่อยู่ในรูปของไฟล์ CSV หรือไฟล์ JSON

ตัวอย่างของข้อมูลแบบมีโครงสร้าง ได้แก่

- ข้อมูลในตาราง Excel เช่น ชื่อ, อายุ, น้ำหนัก, ส่วนสูง
- ข้อมูลในฐานข้อมูล เช่น ชื่อตาราง, ชื่อคอลัมน์, ค่าข้อมูล
- ข้อมูลในไฟล์ CSV เช่น ชื่อ, นามสกุล, อายุ, เพศ
- ข้อมูลในไฟล์ JSON เช่น { "ชื่อ": "John Doe", "นามสกุล": "Doe", "อายุ": 30, "เพศ": "ชาย" }

2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)

ตัวอย่างข้อมูลแบบมีโครงสร้างเพิ่มเติม ได้แก่

- **ข้อมูลการสั่งซื้อสินค้า เช่น รหัสสินค้า, จำนวนสินค้า, ราคาสินค้า, วันที่สั่งซื้อ**
- **ข้อมูลการจองตั๋วเครื่องบิน เช่น หมายเลขเที่ยวบิน, ชื่อผู้โดยสาร, ที่นั่ง, วันที่เดินทาง**
- **ข้อมูลการสมัครงาน เช่น ชื่อผู้สมัคร, ที่อยู่, เบอร์โทรศัพท์, อีเมล, ทักษะ**
- **ข้อมูลสภาพอากาศ เช่น อุณหภูมิ, ความชื้น, ทิศทางลม, ฝน**

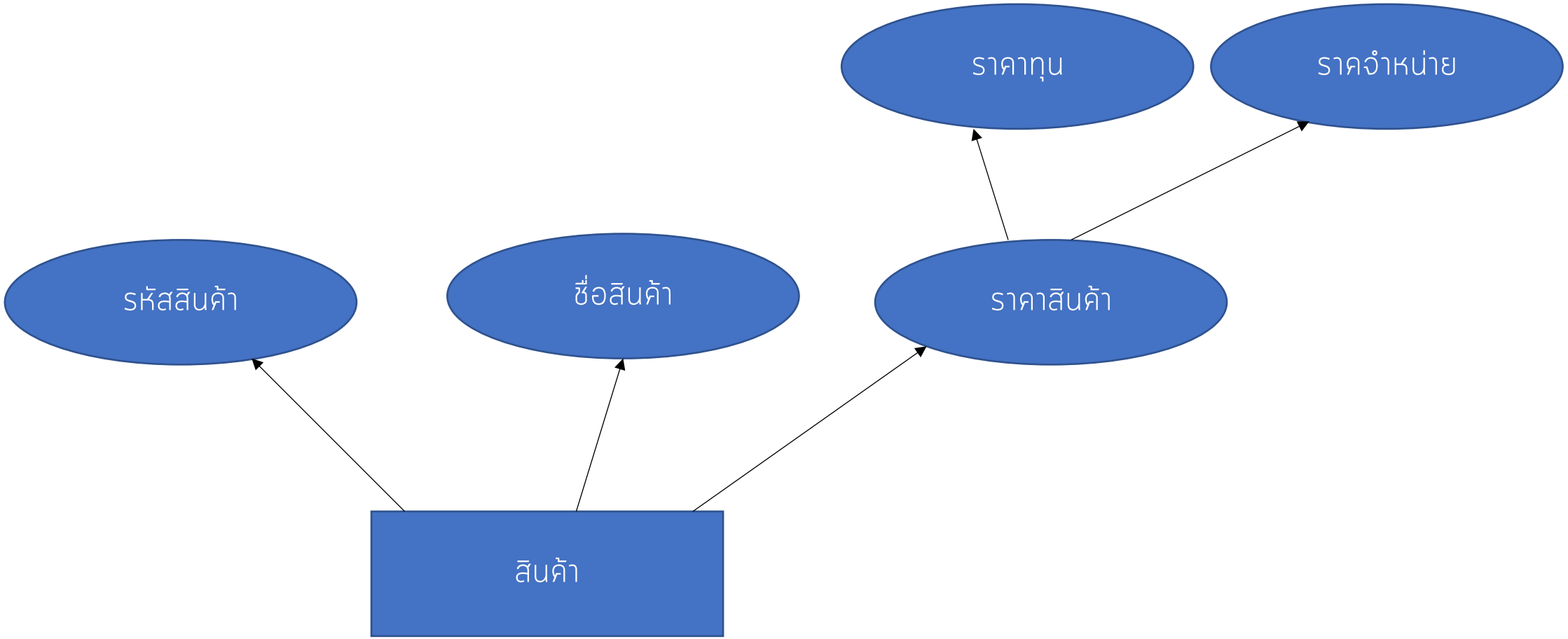
2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)

ข้อมูลแบบมีโครงสร้างมีประโยชน์ในการประมวลผลและวิเคราะห์ข้อมูล เนื่องจากสามารถจัดเก็บและเข้าถึงข้อมูลได้อย่างมีประสิทธิภาพ นอกจากนี้ยังช่วยให้สามารถระบุความสัมพันธ์ระหว่างข้อมูลต่างๆ ได้ง่ายขึ้น

- ข้อดีของข้อมูลแบบมีโครงสร้าง ได้แก่
  - ง่ายต่อการประมวลผลและวิเคราะห์
  - สามารถระบุความสัมพันธ์ระหว่างข้อมูลต่างๆ ได้ง่ายขึ้น
  - สามารถจัดเก็บและเข้าถึงข้อมูลได้อย่างมีประสิทธิภาพ

- ข้อเสียของข้อมูลแบบมีโครงสร้าง ได้แก่
  - อาจไม่เพียงพอสำหรับข้อมูลที่มีความซับซ้อน
  - อาจไม่ยืดหยุ่นเพียงพอสำหรับข้อมูลที่มีการเปลี่ยนแปลงบ่อยครั้ง

2.11.2 ข้อมูลแบบมีโครงสร้าง (structured data)



2.11.3 ข้อมูลแบบไม่มีโครงสร้าง (unstructured data)

ดังที่ได้กล่าวไปแล้วว่าข้อมูลส่วนใหญ่จะเป็นข้อมูลแบบที่ไม่มีโครงสร้าง เช่น ข้อความ หรือ รูปภาพ ต่าง ๆ แต่ข้อมูลเหล่านี้ก็มีความสำคัญ เช่น ตัวอย่างของการนำไปประยุกต์เพื่อหาทัศนคติต่างๆ ของลูกค้าที่เกิดขึ้น ในหัวข้อนี้ผมขอยกตัวอย่างการแปลงข้อมูลที่เป็นรูปแบบข้อความให้เป็นข้อมูลในรูปแบบตารางโดยใช้วิธีการแปลงข้อมูล (data transformation) เช่น ข้อความข่าวในตารางที่ 2.3-2.4

เอกสาร	ข้อความในเอกสาร
1	ชาวนาหลายจังหวัดอีสานปลื้มทหาร หลังได้รับเงินจำนำข้าว
2	ชาวนาโคราช แห่รับเงินจำนำข้าวจาก ทหาร-ธ.ก.ส. พร้อมขอบคุณ ค.ส.ช.
3	เกษตรกร เฮ ผู้ว่าฯปราจีนบุรี นำเงินจำนำข้าวคืนชาวนาครบทุกราย

ตารางที่ 2.4 แสดงข้อความข่าวต่างๆ

2.11.3 ข้อมูลแบบไม่มีโครงสร้าง (unstructured data)

จากข้อความในตารางที่ 2.3-2.4 เราสามารถตัดข้อความออกมาเป็นคำต่างๆ ได้ เช่น ชาวนา หลายจังหวัดอีสาน ปลื้ม ทหาร หลัง เป็นต้น หลังจากนั้นคัดเลือกคำที่เป็นคำสำคัญและนำมาเป็นชื่อของแต่ละแอตทริบิวต์ และพิจารณาว่าในแต่ละเอกสารมีคำใดปรากฏขึ้นบ้าง ซึ่งถ้าเอกสารมีคำนั้นปรากฏขึ้นจะมีค่าเป็น Y และไม่มีปรากฏขึ้นจะมีค่าเป็น N ดังแสดงในตารางที่ 1-3 ซึ่งแปลความหมายได้ว่า เอกสารที่ 1 มีคำว่า "ชาวนา","ปลื้ม", "รับเงิน", "จำนำข้าว", "ทหาร" เกิดขึ้นแต่ไม่มีคำว่า "ขอบคุณ" ปรากฏอยู่เลย

เอกสาร	ชาวนา	ปลื้ม	รับเงิน	จำนำข้าว	ทหาร	ขอบคุณ
1	Y	Y	Y	Y	Y	N
2	Y	N	Y	Y	Y	Y
3	Y	N	N	Y	N	N

ตารางที่ 2.5 แสดงข้อความที่ได้ทำการแปลงมาเป็นข้อมูลในรูปแบบตาราง

จากตัวอย่างนี้เป็นการแปลงข้อมูลในรูปแบบที่ไม่มีโครงสร้าง คือ ข้อความให้เป็นข้อมูลในรูปแบบที่มีโครงสร้าง คือ ตาราง ถัดจากนี้ผมจะแนะนำให้ท่านผู้อ่านรู้จักกับเทคนิคในการวิเคราะห์ข้อมูลด้วยดาต้าไมน์นิงซึ่งมี 2 ประเภทหลักดังจะได้อธิบายในหัวข้อถัดไปครับ

2.12 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

- **การเตรียมข้อมูลสำหรับ Data Mining**

สิ่งแรกที่ต้องทำคือ เราต้องมาคิดกันก่อนว่าจะนำเทคนิคดาต้าไมน์นิงไปประยุกต์กับด้านใด เพราะเหตุใด และต้องการจะหาความรู้แบบใดออกมาจากการทำดาต้าไมน์นิง สมมุติว่าเราต้องการนำเทคนิคดาต้าไมน์นิงไปประยุกต์ใช้กับด้านการศึกษา เนื่องมาจากเราได้เล็งเห็นว่าในปัจจุบันตามสถาบันการศึกษาส่วนใหญ่มีข้อมูลต่าง ๆ นิสิตที่ได้ถูกจัดเก็บไว้เป็นเวลานาน แต่ข้อมูลส่วนใหญ่จะได้นำมาใช้ประโยชน์ตอนที่นิสิตศึกษาอยู่เท่านั้น เมื่อนิสิตจบการศึกษาไปแล้วข้อมูลก็จะได้รับการจัดเก็บไว้เป็นอย่างดี โดยที่ไม่ได้นำมาใช้ให้เกิดประโยชน์เท่าที่ควร เมื่อเราคิดได้แล้วว่าเราต้องการนำเทคนิคดาต้าไมน์นิงไปประยุกต์ใช้กับการศึกษา ต่อมาเราต้องหาเป้าหมาย (Mining Objective) ว่าเราต้องการสืบค้นความรู้แบบใดจากการทำดาต้าไมน์นิงกับข้อมูลนิสิตนี้บ้าง ถ้าเราต้องการนำเทคนิคดาต้าไมน์นิงเพื่อนำมาช่วยนิสิตในการเลือกสาขาวิชา เช่น สำหรับนิสิตที่คณะวิศวกรรมศาสตร์ จะเห็นได้ว่ามีสาขาวิชาต่าง ๆ มากมายกว่า 10 สาขาวิชา ซึ่งจะเห็นได้ว่า นิสิตส่วนใหญ่เมื่อเข้ามาศึกษาในคณะวิศวกรรมศาสตร์แล้ว พอถึงเวลาที่ต้องเลือกสาขาวิชา นิสิตจะไม่ทราบว่าความสามารถตนเองควรจะเข้าเรียนในสาขาวิชาใดจึงจะมีโอกาสประสบความสำเร็จมากที่สุด ดังนั้น เราจึงเห็นว่าสมควรอย่างยิ่งที่จะนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้กับฐานข้อมูลนิสิต โดยความรู้ (knowledge) ที่ได้จากการทำดาต้าไมน์นิงสามารถนำมาใช้ในการช่วยนิสิตเลือกสาขาวิชาได้เมื่อเราได้เป้าหมายในการทำดาต้าไมน์นิงแล้ว เราก็ต้องมาหาข้อมูลนิสิตกัน สมมุติว่าเราได้ข้อมูลนิสิตย้อนหลังทั้งหมด 10 ปี มีทั้งหมด 2 ส่วน คือ ข้อมูลประวัติส่วนตัวนิสิตดังตารางที่ 1 และข้อมูลการลงทะเบียนเรียนในแต่ละรายวิชาของนิสิตดังตารางที่ 2

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

2.12.1 Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป

Customer ข้อมูลจากแหล่งที 1

CID	Name	Street	City	Sex
11	Kristen Smith	2 Hurley Pl	South Fork, MIN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Male , Female , M ,F , ชาย , หญิง

Client ข้อมูลจากแหล่งที่ 2

CNo	Lastname	FirstName	Gender	Address	Phone/Fax
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	6542/333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

ตารางที่ 2.6 แสดงตารางที่ทำความสะอาดข้อมูล

2.12.2 Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน

Customers ของข้อมูลที่นำมารวมกันและทำความสะอาดแล้ว

No	LName	FName	Gender	Street	City	State	ZIP	Phone	Fax	CID	CNo
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

(แหล่งที่มา: Rahm, Erhard and Do, Hong-Hai) [1]

ตารางที่ 2.7 แสดงตารางที่ทำความสะอาดข้อมูลแล้ว

2.12.3 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ID	Sex	ชื่อ	Address	SchoolGPA	...	Major	GPA
1	นาย	วิโรจน์ พัฒนากุล	86/9 หมู่ 2	2.5	.....	ไฟฟ้า	2.3
2	นส.	ดวงพร เอี่ยมสุข	54/2 หมู่ 7	3.4	.....	โยธา	3.2

ตารางที่ 2.8 ตัวอย่างข้อมูลประวัติส่วนตัวนิสิต

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

สิ่งสำคัญในการทำเหมืองข้อมูล

1. ทำความเข้าใจธุรกิจ ...ร้านขายกาแฟ ....สิ่งที่ต้องการพฤติกรรมของผู้บริโภค ส่วนใหญ่แล้วผู้เวลาสั่งอาหารชนิดหนึ่งแล้ว มักจะสั่งอีกสิ่งหนึ่งด้วยพร้อมกัน ...ใช้เทคนิคอะไร ? ....association rule <u>การหากฎของความสัมพันธ์</u>

   ตัวอย่าง การจงรักภักดีของต่อสินค้า ของผู้บริโภค ควรใช้เทคนิคอะไร .....decision tree

2. ทำความความเข้าใจข้อมูล

2.12.3 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

จากตารางที่ 2.8 เป็นตัวอย่างข้อมูลประวัติส่วนตัวต่าง ๆ ของนิสิต เช่น รหัสประจำตัวนิสิต ชื่อ เพศ สัญชาติ ที่อยู่ วันเกิด สถานภาพทางครอบครัว คะแนนสอบเข้า ผลการเรียนระดับมัธยม สาขาวิชาที่นิสิตศึกษาอยู่ เกรดเฉลี่ยสะสมจนถึงปีปัจจุบัน ฯลฯ

ID	Subject	Section	Term	Year	Grade
1	001	1	1	2537	C+
1	002	1	1	2537	D
1	005	1	1	2537	B+

ตารางที่ 2.9 ตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิต

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

2.12.3 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

จากตารางที่ 2.9 เป็นตารางข้อมูลการลงทะเบียนของนิสิตในแต่ละรายวิชา ในแต่ละภาคการศึกษา พร้อมทั้งหมู่ที่เรียน และผลการเรียนในรายวิชานั้น ๆ ของนิสิตแต่ละคน

เมื่อเราได้ข้อมูลทั้งหมดแล้ว ขั้นต่อมาก็คือ การเตรียมข้อมูลเพื่อให้พร้อมที่จะนำไปทำดาต้าไมน์นิง ซึ่งแบ่งเป็นขั้นต่าง ๆ ได้ดังนี้

การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ข้อมูลที่ได้มานั้น เป็นข้อมูลที่ยังไม่สมบูรณ์ที่จะสามารถนำไปใช้ผ่านกระบวนการดาต้าไมน์นิงได้ จึงต้องมีการจัดการข้อมูล การเตรียมข้อมูลเบื้องต้นมีวิธีการดังนี้ เลือกเฉพาะคอลัมน์สำคัญที่คาดว่าจะสามารถนำมาใช้ประโยชน์ได้ และเป็นคอลัมน์ที่มีข้อมูลค่อนข้างครบถ้วนเมื่อเทียบกับจำนวนนิสิต เช่น จากในตารางที่ 2.9 คอลัมน์สำคัญที่มีข้อมูลค่อนข้างมาก ได้แก่ ข้อมูลรหัสนิสิต ที่อยู่ อายุ เพศ ประวัติครอบครัวโรงเรียน เกรดเฉลี่ยที่จบการศึกษาในมหาวิทยาลัย เป็นต้น ส่วนในบางคอลัมน์ที่มีความสำคัญ แต่มีข้อมูลน้อยมากนั้นจะไม่นำมาพิจารณา เช่น ข้อมูลคะแนนสอบเอ็นทรานซ์ในแต่ละวิชา เหตุผลในการสอบเข้า เป็นต้น

สำหรับคอลัมน์ที่มีค่าสำหรับทุกแถวเป็นค่าเดียวกัน เช่น “สัญชาติไทย” จะเป็นข้อมูลที่ไม่สามารถแยกความแตกต่างของแต่ละแถวได้เลย ดังนั้นในการทำดาต้าไมน์นิงจะไม่สามารถใช้ประโยชน์จากคอลัมน์นี้ ดังนั้น จึงไม่นำคอลัมน์นี้มาพิจารณา คอลัมน์ที่มีค่าที่ไม่ซ้ำกันเลย จากตารางที่ 1 ได้แก่ ชื่อผู้ปกครอง หมายเลขโทรศัพท์ เป็นต้น ข้อมูลเหล่านี้ไม่สามารถหาแถวที่มีข้อมูลสัมพันธ์กันได้เลย การทำดาต้าไมน์นิงจึงไม่สามารถนำข้อมูลเหล่านี้มาใช้ประโยชน์ได้ ดังนั้นในการทำดาต้าไมน์นิงควรกำจัดคอลัมน์ที่มีข้อมูลไม่ซ้ำกันเลยออก แก้ไขข้อมูลให้ถูกต้องสมบูรณ์ ได้แก่ การแก้ไขค่าว่างของข้อมูล ซึ่งสามารถแก้ไขได้หลายวิธี เช่น แก้ไขโดยกำจัดข้อมูลที่ในแถวเป็นค่าว่าง (NULL) ยกตัวอย่างเช่น จากในตารางที่ 2 ข้อมูลบางแถวค่าในคอลัมน์ Grade หายไป ซึ่งจะเห็นได้ว่าถ้ามีแต่รหัสนิสิตและวิชาที่ลงทะเบียน โดยที่ไม่มีข้อมูลเกรดแล้ว เราก็ไม่สามารถจะนำแถวนั้นพิจารณาเพื่อหาความสัมพันธ์ที่น่าสนใจได้

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

2.12.3 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ปรับเปลี่ยนข้อมูลให้มีค่าเหมาะสมในการตัดสินใจ เช่น จากตารางที่ 1 ข้อมูลที่เป็นที่อยู่นั้นไม่สามารถที่จะนำมาใช้โดยตรงได้ เพราะจะเป็นปัญหาดังข้อ 1.3 คือ ข้อมูลที่อยู่ของนิสิตแต่ละคนไม่ซ้ำกันเลย ดังนั้นจึงต้องปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่จะสามารถนำไปใช้ได้ ในกรณีนี้จะปรับข้อมูลในคอลัมน์ที่อยู่ของนิสิตให้เป็น Bangkok และ Non-Bangkok อย่างใดอย่างหนึ่ง เป็นต้นการจัดกลุ่มข้อมูลเพื่อลดการกระจาย (Binning Data) ทั้งนี้เนื่องมาจากข้อมูลของนิสิตมีจำนวนไม่มาก แต่เกรดในแต่ละวิชาที่สามารถมีได้นั้นมีจำนวนมากถึง 10 ตัวด้วยกันคือ {A, B+, B, C+, C, D+ ,D, F, W, I} ดังนั้นเพื่อลดการกระจายของข้อมูลเกรดของนิสิตที่มีมากเมื่อเทียบกับจำนวนนิสิต จึงได้จัดกลุ่มเกรดของนิสิตเป็น 3 กลุ่ม ดังนี้ คือ เกรด {A, B+, B} เป็น High, เกรด{C+, C} เป็น Medium และ เกรด {D+, D, F, W, I} เป็น Low

จากตารางที่ 1 ที่เป็นข้อมูลประวัตินิสิต เราได้นำมาปรับเปลี่ยนข้อมูลบางส่วนเพื่อให้สมบูรณ์ขึ้น ได้แก่

-การตัดคอลัมน์ที่ไม่จำเป็นในการทำดาต้าไมน์นิงออก เช่น คอลัมน์ชื่อนิสิต เพราะชื่อนิสิตแต่ละคนไม่สามารถนำมาทำดาต้าไมน์นิงได้

-คัดเลือกเฉพาะคอลัมน์ที่คาดว่าจะสามารถนำมาทำดาต้าไมน์นิงได้ เช่น คัดเลือกคอลัมน์โรงเรียน แต่เนื่องจากชื่อโรงเรียนของนิสิตแต่ละคนมีมากมาย เราจึงต้องปรับข้อมูลโรงเรียนให้เป็นกลุ่มอย่างสมดุลเพื่อที่จะได้สามารถนำไปใช้ในการทำดาต้าไมน์นิงได้ เช่น แบ่งข้อมูลโรงเรียนเป็น 2 กลุ่ม คือ สอบเทียบและจบจากมัธยมศึกษาปีที่ 6 โดยกำหนดว่าSchool = 0 คือจบการศึกษาจากมัธยมศึกษาปีที่ 6 และ School = 1 คือสอบเทียบ เป็นต้น

- ปรับเปลี่ยนข้อมูลในบางคอลัมน์เพื่อให้สามารถนำไปไมน์นิงได้ เช่น คอลัมน์ที่อยู่

ปรับข้อมูลให้เป็นกลุ่มว่านิสิตอยู่ในกรุงเทพฯ หรือไม่ เป็นต้น

ผลที่ได้จากการทำข้อมูลจากตารางที่ 2.9 ให้สมบูรณ์แสดงดังตารางที่ 2.10

ID	Sex	Address	School	...	Major	GPA
1	Female	Bangkok	1	....	ELEC	2.3
2	Male	Non-Bangkok	0	....	CIVIL	3.2

ตารางที่ 2.10 ตัวอย่างข้อมูลประวัตินิสิตที่ทำให้สมบูรณ์

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

2.12.3 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

จากตารางที่ 2.10 ที่เป็นตารางข้อมูลการลงทะเบียนเรียนของนิสิต เราได้ปรับข้อมูลบางส่วนให้สมบูรณ์ขึ้น ได้แก่

- **การตัดบางคอลัมน์ที่ไม่น่าสนใจที่จะนำมาทำดาต้าไมน์นิงออก เช่น คอลัมน์หมู่การเรียน**
- **จับกลุ่มข้อมูลในคอลัมน์เกรดเพื่อลดการกระจายของข้อมูล เป็นต้น**

ผลที่ได้จากการทำข้อมูลในตารางที่ 2.10 ให้สมบูรณ์แสดงดังตารางที่ 2.11

ID	Subject	Term	Year	Grade
1	001	1	2537	Medium
1	002	1	2537	Low
1	005	1	2537	High

ตารางที่ 2.11 ตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิตที่ทำให้สมบูรณ์

อ้างอิง http://llem0nz.blogspot.com/2012/03/data-mining_6511.html

2.12 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ตัวอย่างของขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning) โดยใช้ภาษา Python และไลบรารี Pandas:

Data_clearning.py

```
import pandas as pd
import numpy as np

# สร้าง DataFrame ตัวอย่าง
data = {'ชื่อ': ['John', 'Jane', 'Bob', 'Alice',
'Charlie'],
        'อายุ': [25, 28, np.nan, 22, 30],
        'เงินเดือน': [50000, 60000, 75000, np.nan,
90000],
        'เพศ': ['ชาย', 'หญิง', 'ชาย', 'หญิง', 'ชาย']}
df = pd.DataFrame(data)

# 1. ตรวจสอบข้อมูลที่หายไป (Missing Data)
print("ข้อมูลที่หายไป:")
print(df.isnull())
# 2. การจัดการข้อมูลที่หายไป (Handling Missing Data)
# เติมค่าเฉลี่ยของคอลัมน์ 'อายุ' และ 'เงินเดือน' ในข้อมูลที่หายไป
df['อายุ'].fillna(df['อายุ'].mean(), inplace=True)
df['เงินเดือน'].fillna(df['เงินเดือน'].mean(), inplace=True)
# 3. ตรวจสอบข้อมูลที่ซ้ำซ้อน (Handling Duplicates)
print("ข้อมูลที่ซ้ำซ้อน:")
print(df.duplicated())
# ลบข้อมูลที่ซ้ำซ้อน
df.drop_duplicates(inplace=True)
# 4. การแปลงข้อมูล (Data Transformation)
# ในตัวอย่างนี้ไม่ได้ทำการแปลงข้อมูล
# 5. แสดง DataFrame หลังจากทำความสะอาดข้อมูล
print("DataFrame หลังจากทำความสะอาดข้อมูล:")
print(df)
```

2.13 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ตัวอย่างของขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning) โดยใช้ภาษา Python และไลบรารี Pandas:

ในตัวอย่างนี้:

เริ่มต้นด้วยการตรวจสอบข้อมูลที่หายไป (missing data) และใช้ fillna เพื่อเติมค่าเฉลี่ยของคอลัมน์ในที่ข้อมูลที่หายไป.

ตรวจสอบและลบข้อมูลที่ซ้ำซ้อน.

ตามด้วยการแสดง DataFrame หลังจากทำความสะอาดข้อมูล.

นั่นคือขั้นตอนพื้นฐานของการทำความสะอาดข้อมูล แต่ละโปรเจกต์อาจมีความซับซ้อนมากขึ้นและต้องปรับแต่งตามความต้องการของโปรเจกต์.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  SERIAL MONITOR

4  False  False      False  False
ข้อมูลที่ซ้ำซ้อน:
0    False
1    False
2    False
3    False
4    False
dtype: bool
DataFrame หลังจากทำความสะอาดข้อมูล:
      ชื่อ   อายุ  เงินเดือน   เพศ
0     John  25.00    50000.0   ชาย
1     Jane  28.00    60000.0  หญิง
2      Bob  26.25    75000.0   ชาย
3    Alice  22.00    68750.0  หญิง
4  Charlie  30.00    90000.0   ชาย
PS C:\AppServ\www\Python_DataMining>
```

2.13 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

Pandas เป็นไลบรารี (library) ในภาษา Python ที่ถูกพัฒนาขึ้นสำหรับการวิเคราะห์ข้อมูลและการจัดการข้อมูลที่มีโครงสร้างตาราง (tabular data) อย่างมีประสิทธิภาพ. ไลบรารี Pandas มีโครงสร้างข้อมูลหลักสำหรับการทำงานกับข้อมูลที่เรียกว่า DataFrame และ Series.

1. DataFrame:

DataFrame เป็นโครงสร้างข้อมูลที่คล้ายกับตารางฐานข้อมูล มีแถวและคอลัมน์ ทำให้ง่ายต่อการจัดการข้อมูลที่มีโครงสร้าง. คุณสามารถนำเข้าข้อมูลจากไฟล์ CSV, Excel, SQL database, หรือแหล่งข้อมูลอื่น ๆ เข้ามาใน DataFrame และทำการจัดการข้อมูลได้ตามต้องการ.

2. Series:

Series เป็นโครงสร้างข้อมูลที่ประกอบด้วยข้อมูลในลักษณะแถวเดียวของ DataFrame ซึ่งสามารถเป็นข้อมูลของประเภทเดียวกันหรือไม่ก็ได้.

2.13 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

Pandas เป็นไลบรารี (library) ในภาษา Python ที่ถูกพัฒนาขึ้นสำหรับการวิเคราะห์ข้อมูลและการจัดการข้อมูลที่มีโครงสร้างตาราง (tabular data) อย่างมีประสิทธิภาพ. ไลบรารี Pandas มีโครงสร้างข้อมูลหลักสำหรับการทำงานกับข้อมูลที่เรียกว่า DataFrame และ Series.

Pandas มีฟังก์ชันที่มีประสิทธิภาพสำหรับการทำงานกับข้อมูลที่หลากหลาย, เช่น:

การอ่านและเขียนข้อมูล: Pandas สามารถอ่านและเขียนข้อมูลจากและไปยังหลายรูปแบบได้, เช่น CSV, Excel, SQL, JSON, HTML, และอื่น ๆ.

การจัดการข้อมูลที่หายไป: มีฟังก์ชันที่ช่วยในการจัดการข้อมูลที่หายไป, เช่น fillna เพื่อเติมค่าในที่ข้อมูลที่หายไป.

การจัดการข้อมูลที่ซ้ำซ้อน: สามารถตรวจสอบและลบข้อมูลที่ซ้ำซ้อนได้.

การกระทำทางสถิติ: Pandas มีฟังก์ชันสถิติที่สามารถใช้ในการวิเคราะห์ข้อมูลทางสถิติได้, เช่น mean, median, std, และอื่น ๆ.

การกระทำทางคณิตศาสตร์: มีฟังก์ชันคณิตศาสตร์ที่ช่วยในการทำงานกับข้อมูลตัวเลข, เช่น sum, mean, max, min, และอื่น ๆ.

นอกจากนี้, Pandas ยังมีฟังก์ชันอื่น ๆ ที่ช่วยในการจัดการข้อมูลอย่างมีประสิทธิภาพ, ทำให้เป็นไลบรารีที่สำคัญสำหรับงานวิเคราะห์ข้อมูลในภาษา Python.

2.13 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

ก่อนที่จะเริ่มต้น, คุณต้องติดตั้งไลบรารี pandas ถ้ายังไม่ได้ทำตามนี้:

pip install pandas

ต่อมา, นำเข้าไลบรารี pandas และสร้างไฟล์ Excel ด้วยข้อมูล 5 รายการ:

Data_clearning02.py

```
import pandas as pd
# สร้าง DataFrame
data = {
    'ชื่อ': ['John', 'Jane', 'Bob', 'Alice', 'Charlie'],
    'อายุ': [25, 28, 22, 30, 26],
    'เงินเดือน': [50000, 60000, 75000, 90000, 80000],
    'เพศ': ['ชาย', 'หญิง', 'ชาย', 'หญิง', 'ชาย']
}
```

ตอนนี้, ไฟล์ Excel ที่มีชื่อว่า "example_data.xlsx" ถูกสร้างขึ้นมาแล้ว และมีข้อมูล 5 รายการ.
ต่อมา, เราจะอ่านข้อมูลจากไฟล์ Excel นี้และแสดงผลใน DataFrame:

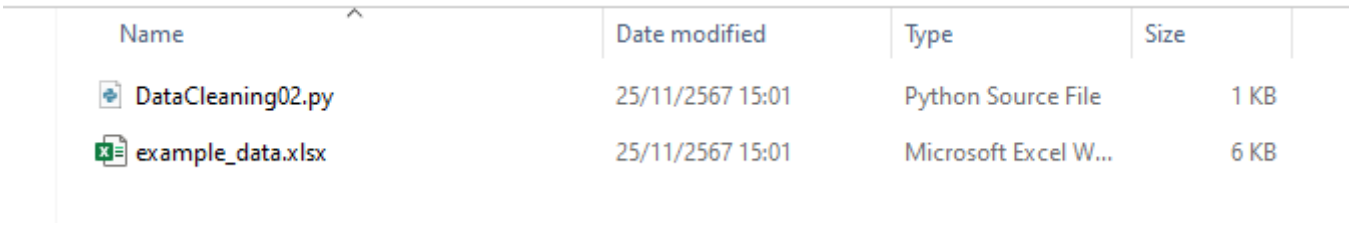
```
df = pd.DataFrame(data)
# บันทึก DataFrame เป็นไฟล์ Excel
df.to_excel('example_data.xlsx', index=False)
# อ่านข้อมูลจากไฟล์ Excel เข้า DataFrame
read_df = pd.read_excel('example_data.xlsx')
# แสดง DataFrame
print(read_df)
```

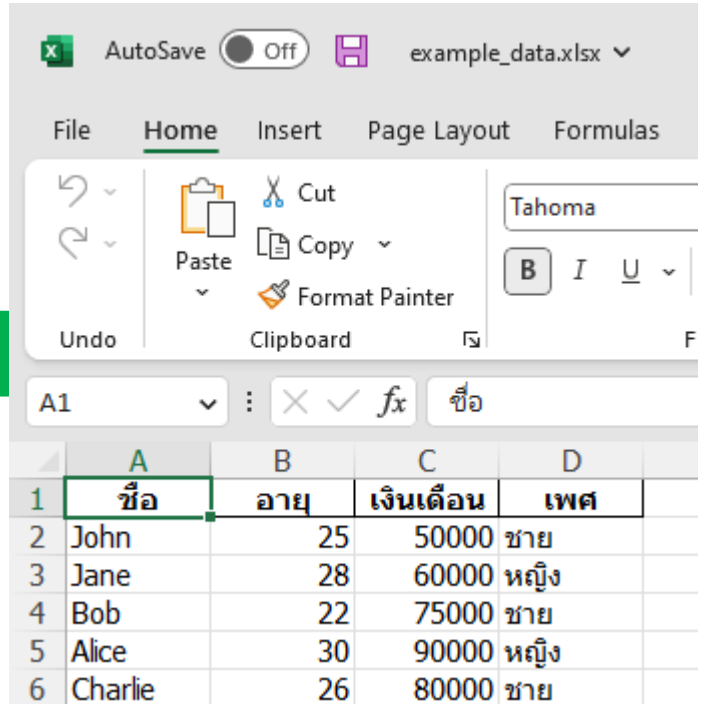
2.13 การเตรียมข้อมูลสำหรับ Data Mining และ การทำข้อมูลให้สมบูรณ์ (Data Cleaning)

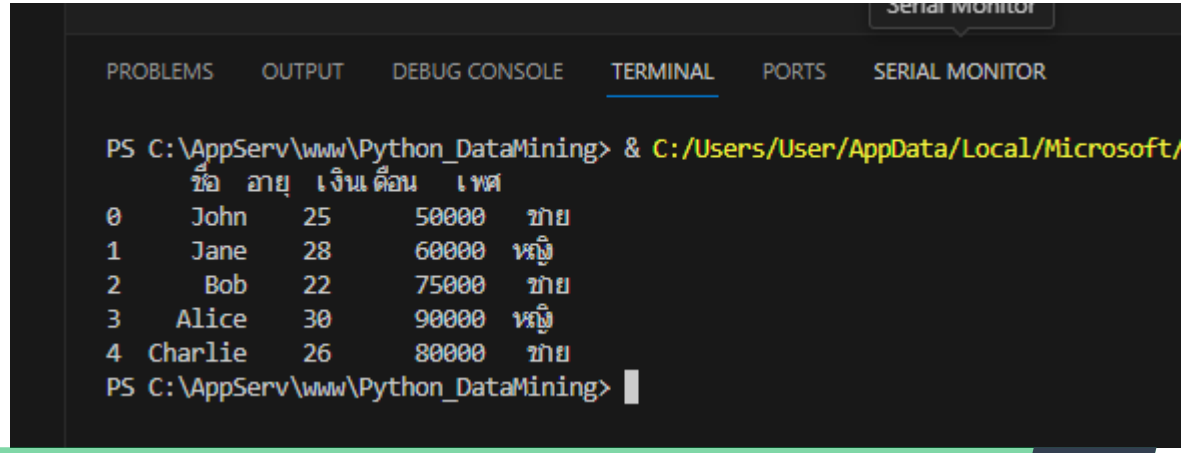
ก่อนที่จะเริ่มต้น, คุณต้องติดตั้งไลบรารี pandas ถ้ายังไม่ได้ทำตามนี้:

```
pip install pandas
```

เมื่อคุณรันโค้ดนี้, คุณจะได้รับผลลัพธ์ที่แสดง DataFrame ที่ถูกอ่านมาจากไฟล์ Excel:

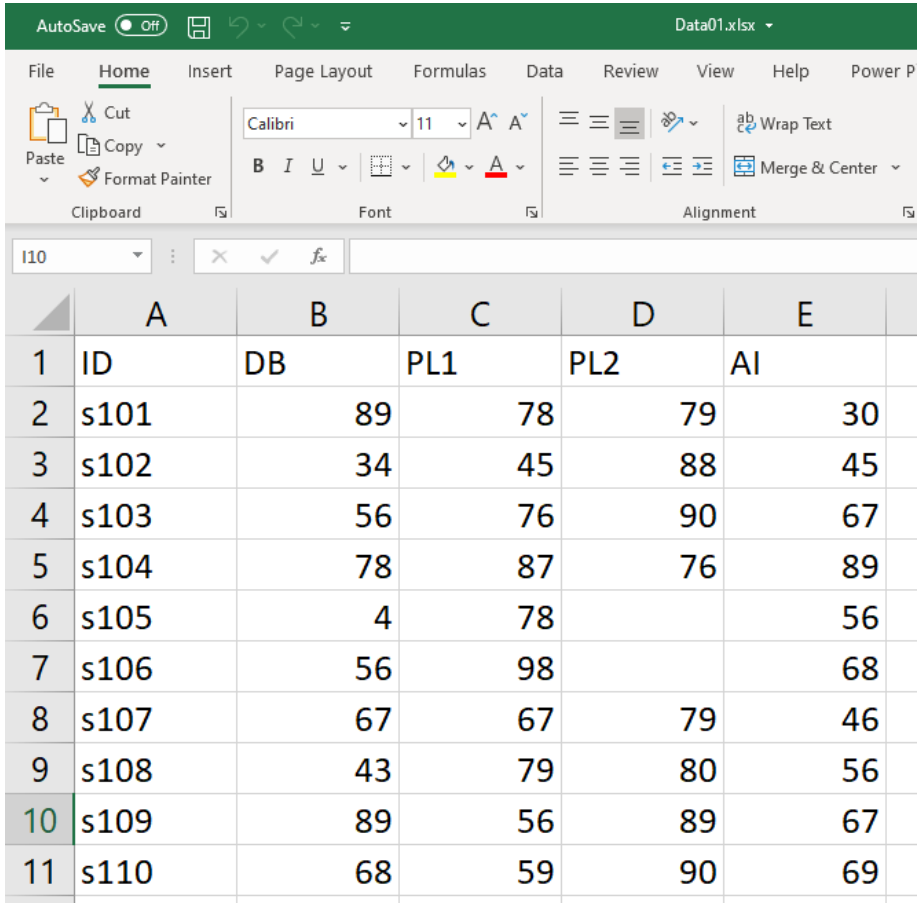






แบบฝึกหัด

- แสดงข้อมูลทั้งหมดของนักศึกษาจากตาราง Student
- แสดงข้อมูลชื่อ (First) ของนักศึกษาจากตาราง Student
- แสดงข้อมูลนามสกุล (Last) ของนักศึกษาจากตาราง Student
- แสดงข้อมูลชื่อ (First) และ นามสกุล (Last) ของนักศึกษาจากตาราง Student
- แสดงข้อมูลนามสกุล (Last) ของนักศึกษาที่แตกต่างกันจากตาราง Student
- แสดงข้อมูลทั้งหมดของนักศึกษาจากตาราง Student ที่มีชื่อว่า John
- แสดงข้อมูลทั้งหมดของนักศึกษาจากตาราง Student ยกเว้นนักศึกษาที่ชื่อ John
- แสดงข้อมูลทั้งหมดของนักศึกษาจากตาราง Grade ที่มีคะแนนมากกว่าหรือเท่ากับ 60 คะแนน
- แสดงข้อมูลชื่อ (First) นามสกุล (Last) และคะแนน (Mark) จากตาราง Grade ที่มีคะแนนน้อยกว่าหรือเท่ากับ 60 คะแนน



	A	B	C	D	E
1	ID	DB	PL1	PL2	AI
2	s101	89	78	79	30
3	s102	34	45	88	45
4	s103	56	76	90	67
5	s104	78	87	76	89
6	s105	4	78		56
7	s106	56	98		68
8	s107	67	67	79	46
9	s108	43	79	80	56
10	s109	89	56	89	67
11	s110	68	59	90	69

อ้างอิง

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Larose, D. T. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd ed.). Wiley.

Piatetsky-Shapiro, G. (1991). Knowledge discovery in databases. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Advances in knowledge discovery and data mining* (pp. 1–28). MIT Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Pearson Education.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

Online Resources

RapidMiner Documentation. (n.d.). Retrieved December 26, 2024, from https://rapidminer.com/

Weka Documentation. (n.d.). Retrieved December 26, 2024, from https://www.cs.waikato.ac.nz/ml/weka/

2.2 องค์ประกอบของเหมืองข้อมูล (ต่อ)

อ้างอิง

- Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.* Elsevier
- Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Databases: *Advances in Knowledge Discovery and Data Mining.* MIT Press
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining.* Pearson Education
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann
- [Weka Documentation](https://www.cs.waikato.ac.nz/ml/weka/)
- [RapidMiner Documentation](https://rapidminer.com/)
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine.
- Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining.* Wiley.