**รายงานผลการเข้าฝึกอบรม**

**หลักสูตร**

# "การวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6 (ขั้นพื้นฐานและปานกลาง)"

**จัดโดย**

**ห้างหุ้นส่วนสามัญดาต้า คิวบ์**

**ณ โรงแรมเคยูโฮม มหาวิทยาลัยเกษตรศาสาตร์ วิทยาเขตบางเขน กรุงเทพฯ**

**วันที่ 25-27 มิถุนายน 2558**

**ผู้จัดทำ**

**รองศาสตราจารย์ ณัฐพร เห็นเจริญเลิศ**

**สาขาวิชาวิทยาศาสตร์และเทคโนโลยี**

**โครงการนี้ได้รับการสนับสนุนจากทุนพัฒนาบุคลากรประจำปีงบประมาณ 2558**

**1. ชื่อ** น.ส.ณัฐพร **นามสกุล** เห็นเจริญเลิศ **อายุ** 50 **ปี**

**ตำแหน่ง** รองศาสตราจารย์ **ระดับ** 9 **สังกัด** สาขาวิชาวิทยาศาสตร์และเทคโนโลยี

**ไปเข้าร่วมฝึกอบรมหลักสูตร** การวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6 (ขั้นพื้นฐานและปานกลาง)

**วันที่** 25-27 มิถุนายน 2558 รวมระยะเวลา 3 วัน

**2.วัตถุประสงค์ของการฝึกอบรม**

1) เพื่อศึกษาการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6

2) เพื่อนำความรู้ไปประยุกต์ใช้ในการวิจัย

**3.สรุปเนื้อหา** จากการฝึกอบรมการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6 บรรยายโดย อ.ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ดังนี้

ภาพรวมของหลักสูตร

โลกในยุคปัจจุบันได้ก้าวเข้าไปสู่ยุคที่เรียกว่า “Big Data” หรือ “ข้อมูลอภิมหาศาล” เนื่องจากในแต่ละวันมีข้อมูลเกิดขึ้นมากมาย อาทิเช่น ข้อมูลสมาชิกของ Facebook ข้อมูลการซื้อสินค้าจากในซุปเปอร์มาร์เกตต่างๆ และเพื่อให้เกิดประโยชน์มากที่สุดเราจำเป็นต้องนำข้อมูลอภิมหาศาลเหล่านี้มาทำการวิเคราะห์ (analyze) ซึ่งเทคนิคหนึ่งที่ได้รับการนิยมอย่างสูงในปัจจุบัน คือ เทคนิค Data Mining ซึ่งเป็นเทคนิคที่ค้นหาความสัมพันธ์ในข้อมูล เช่น ถ้าลูกค้าซื้อเบียร์แล้วลูกค้าจะซื้อผ้าอ้อมร่วมไปด้วย หรือถ้าเรากด Like หน้า Facebook page เราจะเห็นว่า Facebook มีระบบแนะนำ page อื่นๆ ที่เกี่ยวข้องมาให้ด้วย หรือการสร้างโมเดลเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น ทำนายยอดขายในไตรมาสถัดไป หรือ การทำนายว่าพนักงานคนไหนที่จะลาออกจากบริษัทในช่วง 3 เดือนข้างหน้า ตัวอย่างเหล่านี้ล้วนเป็นผลมาจากการวิเคราะห์ข้อมูลทางด้าน Data Mining

การวิเคราะห์ข้อมูลด้วย Data Mining นี้กำลังเป็นที่นิยมไปทั่วโลกด้วยแรงขับเคลื่อนอย่างหนึ่งคือการมีซอฟต์แวร์ที่ช่วยให้ทำการวิเคราะห์ได้ง่ายขึ้น แต่ซอฟต์แวร์ส่วนใหญ่จะเป็นซอฟต์แวร์เชิงพาณิชย์ (commercial software) เช่น SAS Enterprise Miner หรือ IBM Intelligent Miner ทว่าการลงทุนซื้อซอฟต์แวร์เชิงธุรกิจเหล่านี้มาใช้งานอาจจะไม่คุ้มค่าในการลงทุนสำหรับผู้ประกอบการวิสาหกิจขนาดกลางและขนาดย่อม (SMEs) หรืออาจารย์ นักวิจัย และ นักศึกษาระดับปริญญาโทและเอก ในมหาวิทยาลัยต่างๆ ดังนั้นวิธีการหนึ่งที่จะทำให้เราสามารถวิเคราะห์ข้อมูลเหล่านี้ได้คือการใช้ open source software ที่สามารถดาวน์โหลดมาใช้งานได้โดยไม่เสียค่าใช้จ่าย (ฟรี !!!) เช่น ซอฟต์แวร์ Weka ผมคลุกคลีกับ Weka มาเป็นเวลาหลายปี เคยเขียนคู่มือการใช้งาน Weka Explorer ลงในนิตยสาร OpenSource2Day สร้างหลักสูตรการอบรม

การใช้งาน Weka Explorer และอบรมการใช้งานซอฟต์แวร์ตัวนี้มาเป็นจำนวนเกือบยี่สิบรุ่น แม้ว่าซอฟต์แวร์นี้จะใช้งานได้ง่ายสำหรับผู้เริ่มต้นและสะดวกที่จะนำไปใช้ในการพัฒนา Web Application แต่ในหลายๆ ครั้งผมมักจะพบข้อจำกัดหรือความยากในการแสดงผลจากซอฟต์แวร์ตัวนี้ ดังนั้นผมจึงหันมาสนใจซอฟต์แวร์ตัวอื่นที่สามารถทดแทนหรือดีกว่าซอฟต์แวร์ Weka Explorer และผมก็พบกับซอฟต์แวร์ RapidMiner Studio 6 ซึ่งเป็นซอฟต์แวร์ทาง Data Mining ที่ได้รับการโหวตว่ามีผู้ใช้งานมากที่สุดจากเว็บไซต์ KDnuggets.com เมื่อปี 2013

ความรู้เกี่ยวกับข้อมูล และการทำเหมืองข้อมูล

ข้อมูลสามารถแบ่งตามแหล่งที่มาได้ดังนี้

1. ข้อมูลภายในองค์กร/บริษัท เช่น ข้อมูลเกี่ยวกับการซื้อขาย ข้อมูลประวัติลูกค้า ข้อมูลประวัติพนักงาน เป็น transaction 88% log data 73% Emails 57%
2. ข้อมูลภายนอกองค์กร/บริษัท เช่น ข้อมูลจาก social media ต่างๆ ข่าว ข้อมูลรูปภาพ เสียง

ซึ่งสามารถนำข้อมูลเหล่านี้มาวิเคราะห์ให้มีประโยชน์ต่อองค์กรได้


คำจำกัดความ ฐานข้อมูล (Database) ฐานข้อมูลใช้ในการจัดเก็บข้อมูล ลดความซ้ำซ้อนของข้อมูล เน้นการจัดเก็บ เพิ่ม แก้ไขและลบข้อมูลได้ถูกต้อง

คลังข้อมูล (Data Warehouse) คลังข้อมูลรวบรวมข้อมูลจากหลายๆ ฐานข้อมูล แปลงข้อมูลให้มีความเหมือนกัน เหมาะสำหรับการเรียกดู (view) เพื่อสร้างรายงานสรุปผลต่างๆ

การทำเหมืองข้อมูล (Data Mining) การวิเคราะห์ข้อมูลเพื่อค้นหาความสัมพันธ์หรือรูปแบบที่มีประโยชน์ในฐานข้อมูล

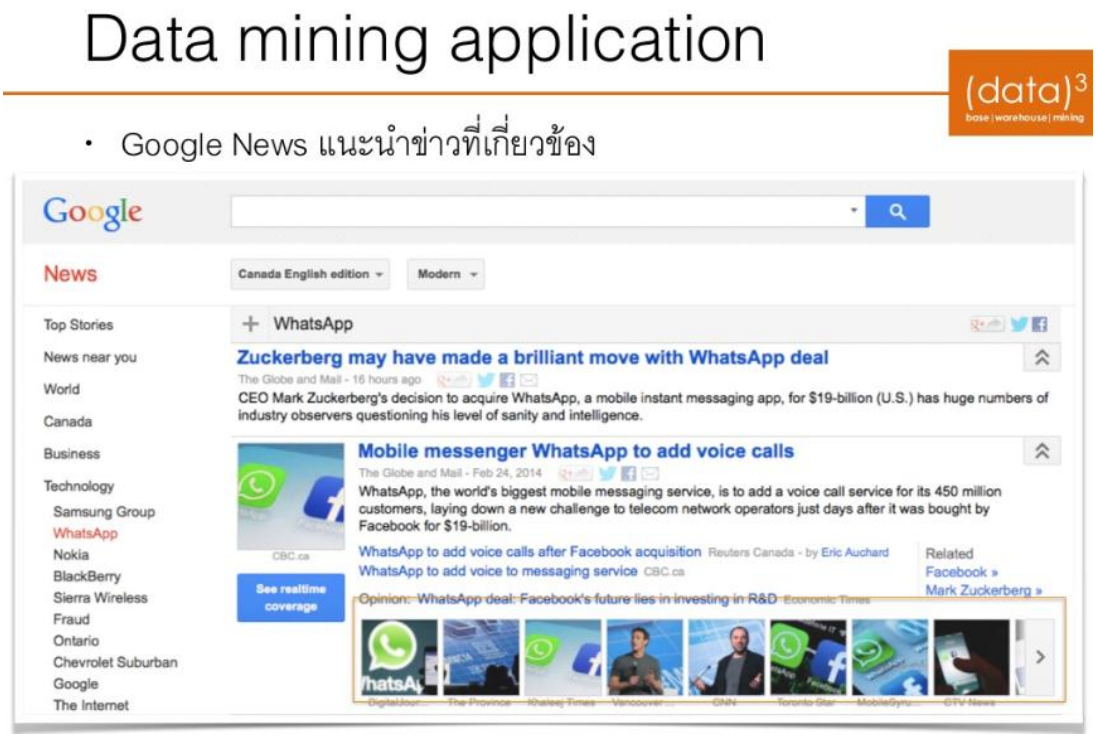
ตัวอย่างการนำ Data Mining ไปใช้งาน เช่น การทำบัตรสมาชิก (loyalty card) ของห้างร้านต่างๆ เพื่อ

- ติดตามพฤติกรรมการซื้อสินค้าของลูกค้าจากบัตร loyalty
- นำมาวิเคราะห์และนำเสนอเป็นโปรโมชันพิเศาให้แต่ละบุคคล
- เพิ่มโอกาสในการขายสินค้าให้กับลูกค้า
- กระตุ้นให้ลูกค้าได้ซื้อสินค้ามากขึ้น เช่น ซื้อสินค้าวันนี้ จะได้ส่วนลดพิเศษ ทำให้ลูกค้าเกิดการตัดสินใจซื้อทันที

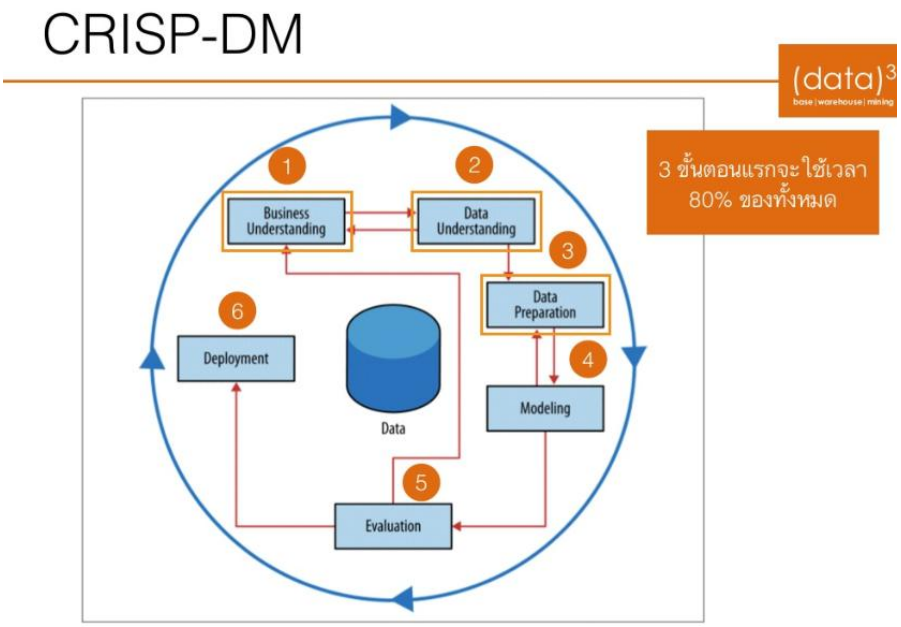
ตัวอย่างการวิเคราะห์ เช่น ห้าง Walmart พบว่าทุกวันศุกร์หลังบ่ายโมงจะมีลูกค้าเพศชายอายุระหว่าง 25-35 ปีซื้อสินค้าเบียร์และ Diapersมากที่สุด

ห้าง Target ทำการวิเคราะห์การซื้อสินค้าของลูกค้าเพศหญิงพบรูปแบบว่าถ้ามีการซื้อวิตามิน ซื้ออาหารบำรุง ซื้อตู้เตียงเพิ่ม ลูกค้าจะเริ่มตั้งครรภ์ Target จะส่งโปรโมชั่นที่เกี่ยวข้องไปให้ลูกค้าเหล่านั้น

การแนะนำสินค้าที่เกี่ยวข้องของเว็บไซต์ amazon.com ให้กับลูกค้าที่เคยเข้าสั่งซื้อหนังสือ หรือเว็บ Netflix แนะนำภาพยนต์ที่คล้ายกับที่เคยดู การทำนายอายุและเพศจากรูปภาพ การแนะนำข่าวที่เกี่ยวข้อง




การวิเคราะห์โดยทำเหมืองข้อมูลตาม methodology ของ CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งพัฒนาโดย 3 บริษัทคือ บริษัท SPSS บริษัท DaimlerChrysler และบริษัท NCR เป็น workflow มาตรฐานสำหรับทำ data mining ประกอบด้วย 6 ขั้นตอน




1. Business Understanding เป็นขั้นตอนแรกของ CRISP-DM โดยการทำความเข้ากับปัญหาหรือโอกาสเชิงธุรกิจ เป็นการตั้งเป้าหมาย ระบุ output หรือเป้าหมายที่ต้องการได้จากการวิเคราะห์ข้อมูลด้วย data mining ตัวอย่างเช่น
    - ทำอย่างไรถึงเพิ่มยอดขายให้กับสินค้าชนิดต่างๆ ได้
    - ต้องการแบ่งกลุ่มนักศึกษาออกตามความสนใจ
    - ทำอย่างไรให้ลูกค้ากลับมาซื้อสินค้าได้อีก
    - อยากทำนายปริมาณน้ำฝนที่ตกใน 2 วันถัดไป
    - อยากรู้ว่าลูกค้าคนใดบ้างมีโอกาสป่วยเป็นโรคมะเร็ง
2. Data Understanding ขั้นตอนนี้เป็นการรวบรวมข้อมูลที่เกี่ยวข้อง ซึ่งข้อมูลถูกต้องน่าเชื่อถือ ข้อมูลที่ได้มีปริมาณมากเพียงพอ ข้อมูลที่ได้มีความเหมาะสม มีรายละเอียดเพียงพอต่อการนำไปใช้ในการวิเคราะห์ ตัวอย่างเช่น ข้อมูลการซื้อสินค้าของแต่ละบุคคล ข้อมูลการลงทะเบียนและผลการศึกษาของนักศึกษา
3. Data Preparation ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากโมเดลที่ได้จากการทำดาต้าไมนิ่งจะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ แบ่งออกได้เป็น 3 ขั้นตอนย่อยคือ

    3.1 ทำการคัดเลือกข้อมูล (Data Selection) เป็นการกำหนดเป้าหมายก่อนว่าเราจะทำการวิเคราะห์อะไร เลือกใช้เฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่เราจะทำการวิเคราะห์

3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) เป็นการลบข้อมูลซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด เช่น ข้อมูลผิดรูปแบบ ข้อมูลที่หายไป ข้อมูลที่ outlier ที่แปลกแยกจากคนอื่น

3.3 แปลงรูปแบบของข้อมูล (Data transformation) เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ ตามอัลกอริธึมของ data mining ที่เลือกใช้

4. Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมนิ่ง เช่น

- classification สร้างโมเดลจากข้อมูลที่มีอยู่เพื่อทำนายอนาคต เช่น ทำนายปริมาณน้ำฝนที่ตกในวันถัดไป

- clustering เป็นการแบ่งข้อมูลหลายๆ กลุ่มตามความคล้ายคลึง เช่น แบ่งกลุ่มนักศึกษาตามคะแนนที่ได้

- association rules เป็นการหาความสัมพันธ์ของข้อมูลที่เกิดร่วมกัน เช่น ค้นหาสินค้าที่มีการซื้อร่วมกันบ่อยๆ

5. Evaluation เป็นการประเมินหรือวัดประสิทธิภาพของโมเดลวิเคราะห์ข้อมูลขั้นตอนก่อนหน้านั้น

6. Deployment นำโมเดลที่ได้หรือผลการวิเคราะห์ได้ไปใช้งานจริง

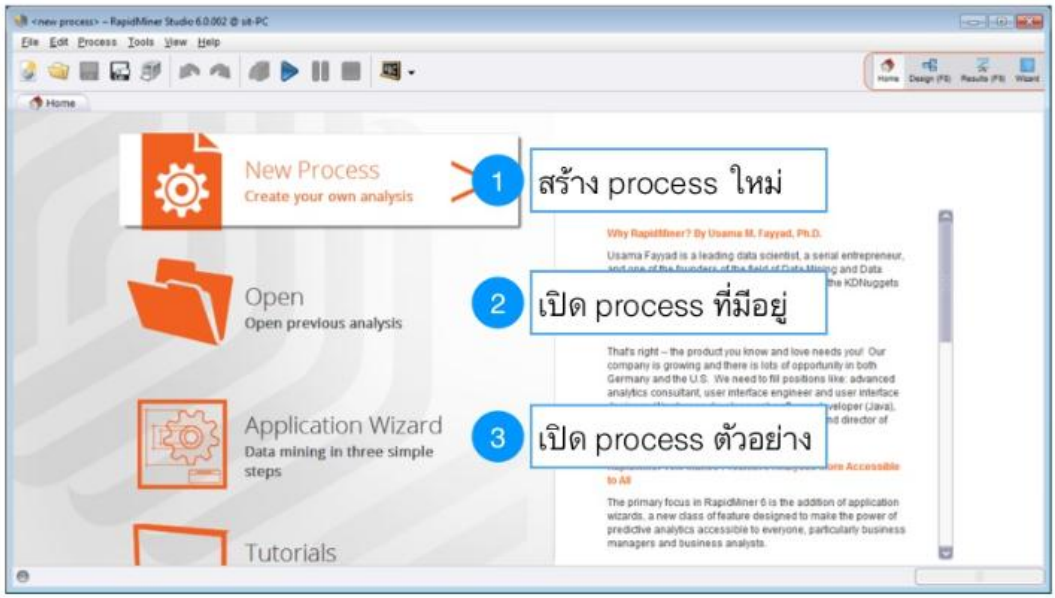
ตัวอย่าง CRISP-DM อ้างอิงจากงานวิจัยเรื่อง การใช้เทคนิคดาต้าไมน์นิ่งเพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์

1. Business Understanding ปัญหาคือ นิสิตคณะวิศวกรรมศาสตร์ ม.เกษตรศาสตร์จะเลือกภาควิชาเมื่อขึ้นชั้นปีที่ 2 นิสิตเลือกภาควิชาไม่ตรงกับความสามารถของตนเอง โดยเลือกตามเพื่อนหรือเลือกตามที่ผู้ปกครองแนะนำ ผลที่ตามมานิสิตบางคนได้ผลการเรียนตกต่ำและทำให้ต้องออกจากมหาวิทยาลัยกลางคัน
2. Data Understanding ข้อมูลคณะวิศวกรรมศาสตร์ ม.เกษตรศาสตร์ช่วงปี พ.ศ. 2535-2542 มีนิสิตประมาณ 10,000 คน ข้อมูลมีจำนวน 476,085 แถว โดยข้อมูลแบ่งเป็น 2 ส่วน ข้อมูลประวัติส่วนตัวของนิสิตประกอบด้วย เพศ ที่อยู่ GPA ระดับมัธยมปลาย GPA ชั้นปีที่ 1 และข้อมูลการลงทะเบียนของนิสิต ประกอบด้วยเกรดวิชาคณิตศาสตร์ เกรดวิชาฟิสิกส์ เกรดวิชาเคมี
3. Data Preparation คัดเลือกวิชาที่เกี่ยวข้องกับภาควิชาต่างๆ ในคณะวิศวกรรมศาสตร์ แปลงข้อมูลให้เหมาะสมกับการวิเคราะห์
4. Modeling แบ่งข้อมูลออกเป็น 2 ส่วน คือ 70% ของข้อมูลทั้งหมดใช้ในการสร้างโมเดล 30%ของข้อมูลทั้งหมดใช้ในการทดสอบประสิทธิภาพของโมเดล สร้างโมเดลด้วยเทคนิค Decision Tree วึ่งจะเป็นโมเดลที่เข้าใจได้ง่าย โมเดลแบ่งแยกตามภาควิชาต่างๆ เช่น ภาควิชาวิศวกรรมคอมพิวเตอร์ วิศวกรรมไฟฟ้า คำตอบ (class) จะแบ่งเป็น 2 ประเภท คือ Good หมายถึงนิสิตเรียนในภาควิชานี้แล้วจบมาได้ GPA อยู่ในช่วง 40% แรก (top 40%) Bad หมายถึง นิสิตเรียนในภาควิชานี้แล้วจบมาได้ GPA อยู่ในช่วง 40% จากท้าย (bottom 40%)
5. Evaluation ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้ แล้วคำนวณค่าความถูกต้อง

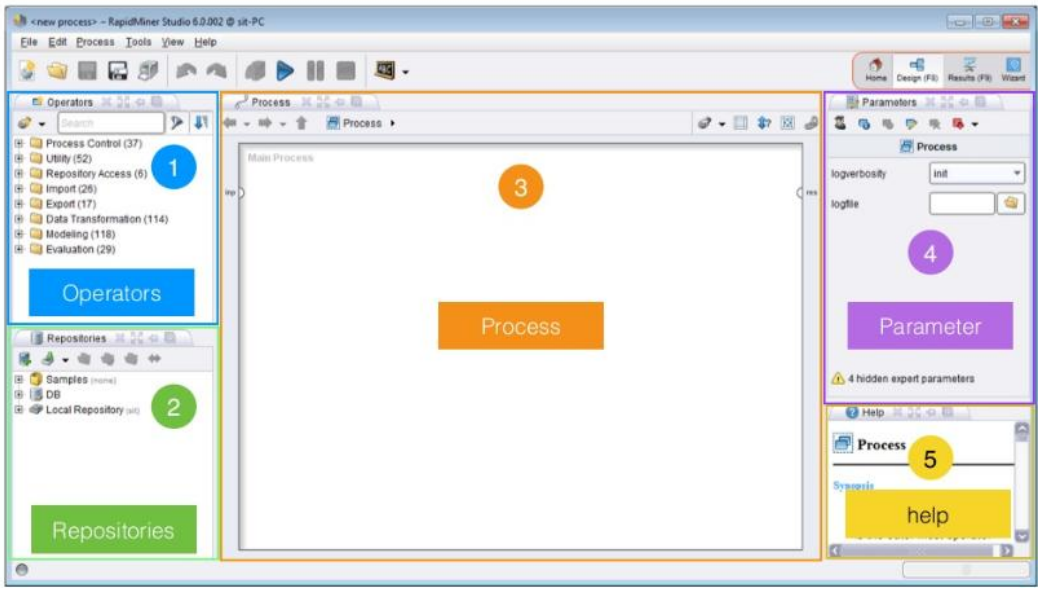
6. Deployment นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา พิจารณาจากเกรดตามโมเดลที่สร้างได้

ส่วนการใช้ซอฟต์แวร์ RapidMiner

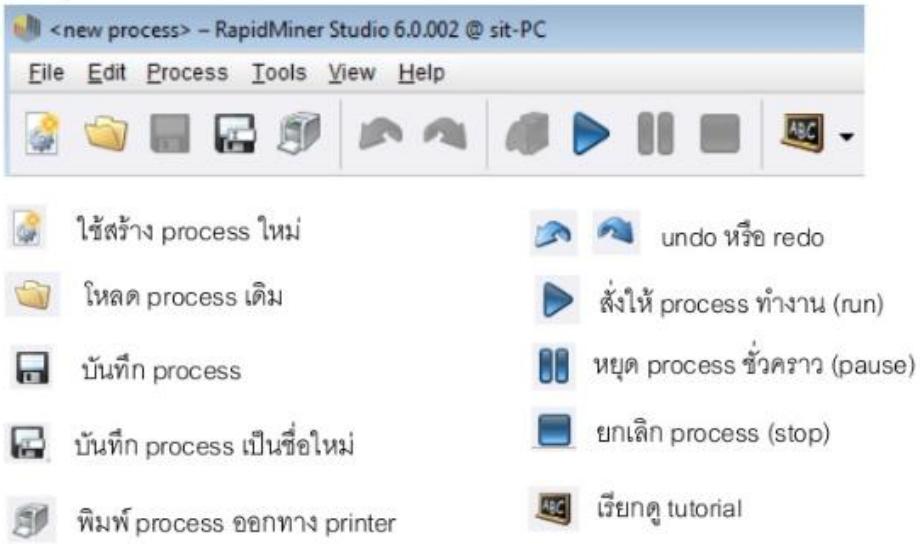
RapidMiner ในตอนแรกใช้ชื่อบริษัทว่า Rapid-I ก่อนตั้งขึ้นเมื่อปี 2006 ในช่วงแรกบริษัทตั้งอยู่ที่ประเทสเยอรมันนี ปี 2013 ได้เปลี่ยนชื่อบริษัทเป็น RapidMiner หลังจากได้รับเงินลงทุนจำนวน 5 ล้านเหรียญและย้ายบริษัทมาอยู่ที่บอสตัน ประเทศสหรัฐอเมริกา ผลิตภัณฑ์หลักของบริษัทคือ RapidMiner Studio 6 บริษัทชั้นนำต่างๆ เช่น PayPal ใช้ซอฟต์แวร์ RapidMiner Studio 6 โดยสามารถดาวน์โหลดได้จาก http://rapidminer.com/download-rapidminer/ เมื่อเริ่มใช้งานจะพบกับหน้าต่าง Home Screen ดังรูป




- องค์ประกอบของ RapidMiner Studio 6



- เมนูใน RapidMiner Studio 6




การเตรียมข้อมูล

ข้อมูลแสดงในรูปแบบของตาราง แถวเรียกว่า ตัวอย่าง (example) คอลัมน์เรียกว่า แอตทริบิวต์ (attribute) มี 3 หน้าที่ที่ใช้งานบ่อย

- ไอดี (ID) เป็นแอตทริบิวต์ที่แสดงหมายเลขของข้อมูลหรือ primary key ในฐานข้อมูล
- แอตทริบิวต์ทั่วไป (attribute) เป็นแอตทริบิวต์ปกติที่จะใช้ในการสร้างโมเดลหรือเรียกว่าเป็นฟีเจอร์ (feature) หรือตัวแปรต้น (independent variable)
- ลาเบล (label) เป็นแอตทริบิวต์ชนิดพิเศษที่มักจะใช้แสดงคำตอบของสิ่งที่เราต้องการจะสร้างโมเดลทำนาย หรือเรียกว่าคลาส (class) หรือตัวแปรตาม (dependent variable)

ประเภทของข้อมูลที่เก็บไว้ในแต่ละแอตทริบิวต์

- Nominal ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่ามากกว่า 2 ค่าขึ้นไป
- Binomial ข้อมูลประเภท category (ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าเพียง 2 ค่าเท่านั้น
- Numeric ข้อมูลประเภทตัวเลข
- Text ข้อมูลประเภทข้อความ

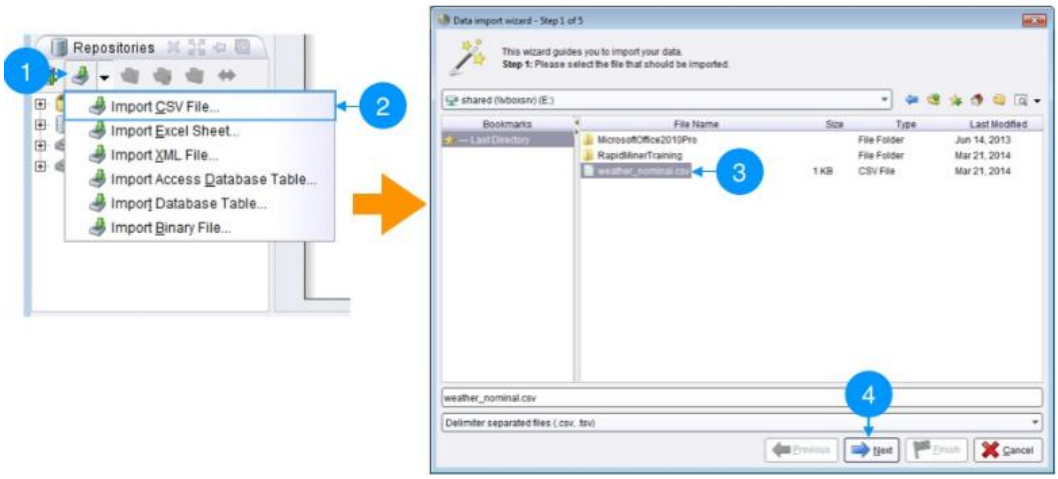
- Repository
  - เป็นที่เก็บข้อมูลและ process เพื่อใช้งานใน RapidMiner Studio 6
  - ทำให้ไม่ต้องโหลดข้อมูลจากไฟล์ใหม่ทุกครั้ง
- องค์ประกอบในส่วน Repository
  - ส่วนที่ 1
    - สำหรับสร้าง Repository ใหม่
    - โหลดไฟล์ประเภทต่างๆ เข้าไปไว้ใน Repository
    - สร้างโฟลเดอร์ใหม่
  - ส่วนที่ 2
    - ข้อมูลและ process Sample ที่ RapidMiner Studio 6 เตรียมไว้ให้
    - ข้อมูลที่เก็บอยู่ในแต่ละ Repository


ไฟล์ประเภท CSV ย่อมาจาก Comma Separated Value ใช้เครื่องหมาย , (comma) คั่นระหว่างแอตทริบิวต์ ไฟล์ CSV สามารถ export ได้จาก Excel หรือ database ต่างๆ

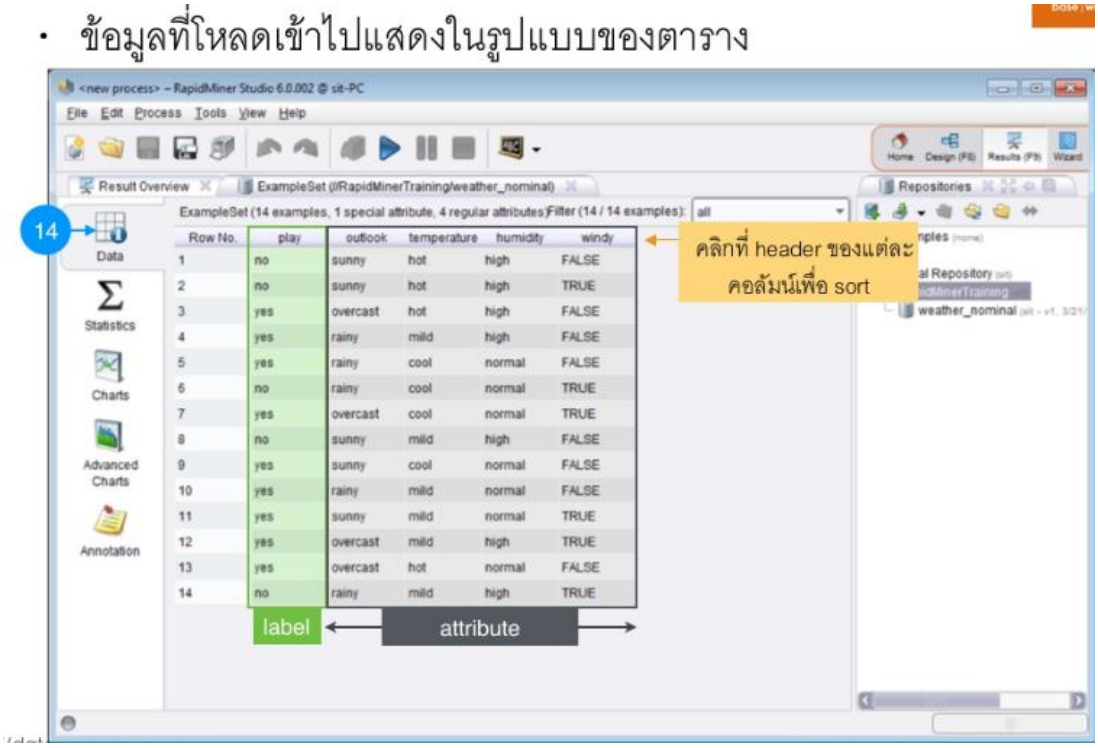
การโหลดไฟล์ CSV เข้าไปใช้ใน RapidMiner Studio 6 ทำได้ 2 แบบ

- ใช้การ import ในส่วนของ Repositories โหลดเข้ามาเก็บไว้และใช้งานได้ตลอด ถ้าข้อมูลในไฟล์ CSV มีการเปลี่ยนแปลงจะไม่ update ต้องทำการโหลดใหม่
- ใช้โอเปอร์เรเตอร์ Read CSV โหลดเข้ามาใช้งานโดยการอ่านจากไฟล์ CSV ทุกครั้ง เมื่อไฟล์ update ข้อมูลจะเปลี่ยนตาม

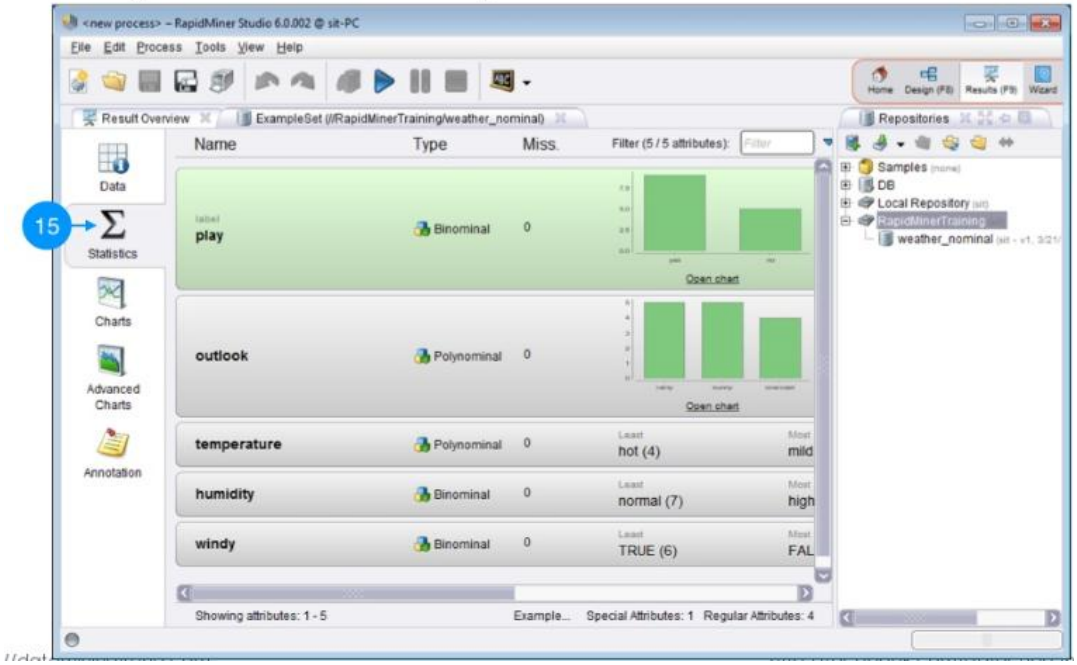
- โหลดไฟล์ csv เข้าไปไว้ใน Repository
  - ในส่วน Repositories เลือก Import CSV File…
  - เลือกไฟล์ weather_nominal.csv ใน thumb drive



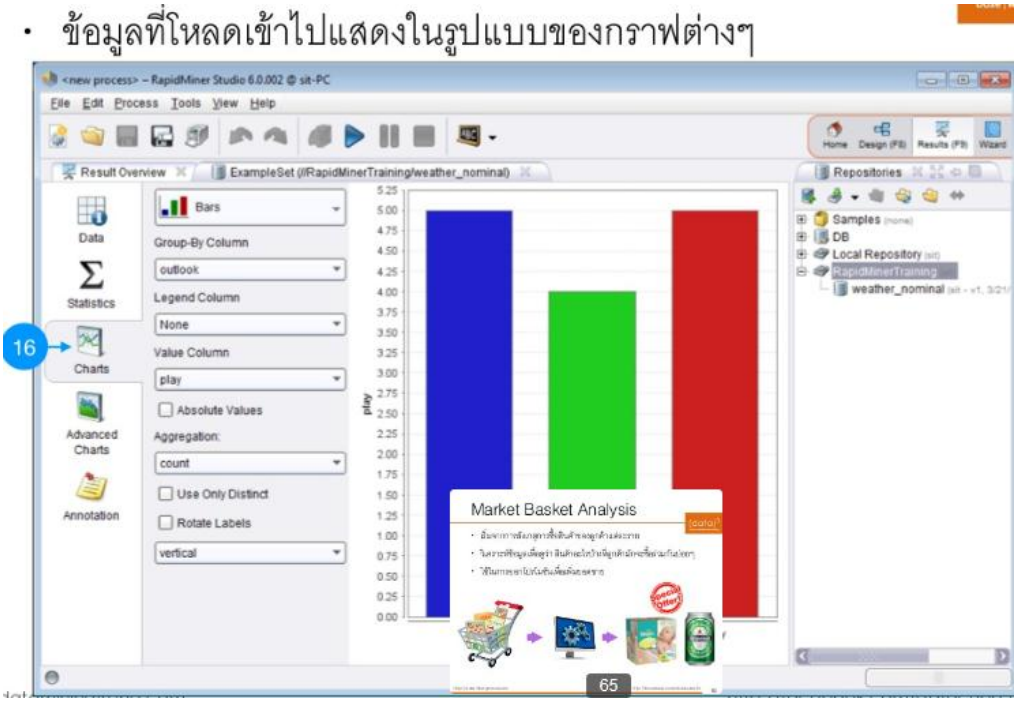
- ข้อมูลที่โหลดเข้าไปแสดงในรูปแบบของตาราง




- ข้อมูลที่โหลดเข้าไปแสดงในรูปแบบของค่าสถิติ



- ข้อมูลที่โหลดเข้าไปแสดงในรูปแบบของกราฟต่างๆ




ตัวอย่างการใช้ Association Rules โดยการทำ Market Basket Analysis เริ่มจากการสังเกตุการซื้อสินค้าของลูกค้าแต่ละราย วิเคราะห์ข้อมูลเพื่อดูว้า สินค้าอะไรบ้างที่ลูกค้ามักจะซื้อร่วมกันบ่อยๆ เพื่อใช้ในการออกโปรโมชั่นเพื่อเพิ่มยอดขาย โดยนำข้อมูลที่ได้จากการซื้อสินค้าแต่ละครั้งของลูกค้าที่จัดเก็บไว้ใน database มาทำการแปลงข้อมูลจาก POS database เป็น transaction database โดย group by ตามเวลาที่ซื้อสินค้า

- นับจำนวนครั้งการซื้อสินค้าแต่ละชนิดคิดเป็น % ของการซื้อสินค้า
- วิเคราะห์เพื่อหาสินค้าที่มีการซื้อมากกว่าหรือเท่ากับ 50%

| Items | Transaction ID | | | | Support |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| Apple | 1 | 0 | 1 | 0 | 2/4 = 50% |
| Beer | 0 | 1 | 1 | 1 | 3/4 = 75% |
| Cereal | 1 | 1 | 1 | 0 | 3/4 = 75% |
| Diapers | 1 | 0 | 0 | 0 | 1/4 = 25% |
| Eggs | 0 | 1 | 1 | 1 | 3/4 = 75% |

| Items | Transaction ID | | | | Support |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| Apple | 1 | 0 | 1 | 0 | 2/4 = 50% |
| Beer | 0 | 1 | 1 | 1 | 3/4 = 75% |
| Cereal | 1 | 1 | 1 | 0 | 3/4 = 75% |
| ~~Diapers~~ | ~~1~~ | ~~0~~ | ~~0~~ | ~~0~~ | ~~1/4 = 25%~~ |
| Eggs | 0 | 1 | 1 | 1 | 3/4 = 75% |

- กฎความสัมพันธ์ (association rules)
  - สร้างจากสินค้าที่ลูกค้าซื้อบ่อยๆ
  - รูปแบบของกฎความสัมพันธ์ คือ **LHS ⇒ RHS**
    - LHS คือ Left Hand Side สินค้าที่ซื้อพร้อมกันบ่อยๆ ด้านซ้ายของกฎ
    - RHS คือ Right Hand Side สินค้าที่ซื้อพร้อมกันบ่อยๆ ด้านขวาของกฎ

| Frequent itemset | Support | Size |
|---|---|---|
| {Apple, Cereal} | 2/4 = 50% | 2 |
| {Beer, Cereal} | 2/4 = 50% | 2 |
| {Beer, Eggs} | 3/4 = 75% | 2 |
| {Cereal, Eggs} | 2/4 = 50% | 2 |
| {Beer, Cereal, Eggs} | 2/4 = 50% | 3 |

Apple ⇒ Cereal
Beer ⇒ Cereal
Beer ⇒ Eggs
Cereal ⇒ Apple
Cereal ⇒ Eggs
Cereal, Eggs ⇒ Beer
Eggs ⇒ Beer

การประยุกต์ใช้ในการเพิ่มยอดขายโดยแนะนำสินค้าที่ลูกค้ามักจะซื้อพร้อมกันบ่อยๆ (cross-selling) และใช้ในการจัดสินค้าในร้านโดย การวางสินค้าที่ลูกค้ามักซื้อพร้อมกันไว้ใกล้ๆ กัน หรือวางสินค้าที่ลูกค้ามักซื้อพร้อมกันไว้ไกลๆ กัน

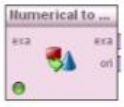
การหากฎความสัมพันธ์ (association rules) มี 2 ขั้นตอน

ขั้นตอนที่ 1 หา frequent itemset ซึ่งใช้เวลานานกว่าขั้นตอนที่ 2 มี 2 เทคนิคที่นิยมใช้คือ

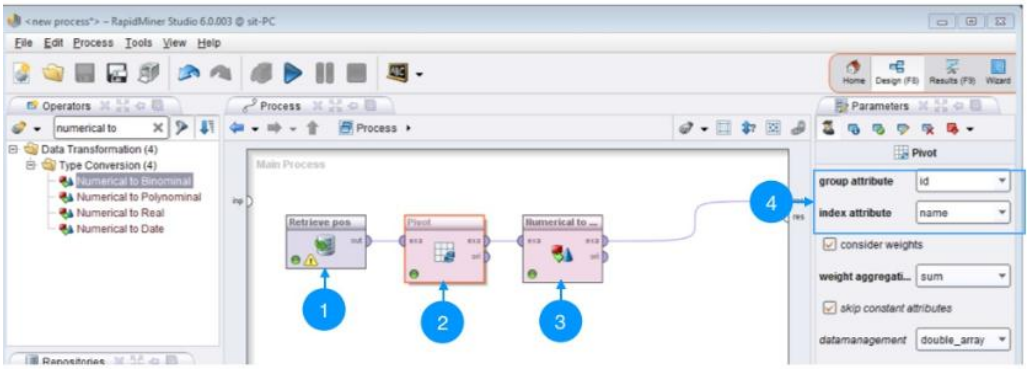
- เทคนิค Apriori (Agrawal and Srikant,1994)
  - สร้างรูปแบบของสินค้า (itemset) ที่มีจำนวนเพิ่มทีละ 1
  - คำนวณค่า support จากในฐานข้อมูล
  - ข้อเสียคือต้องดึงข้อมูลจากฐานข้อมูลหลายรอบทำให้ทำงานช้า
- เทคนิค FP-Growth ( Han, Pei and Yin, 2000)
  - อ่านข้อมูลในฐานข้อมูลและสร้าง FP-tree
  - คำนวณค่า support จาก FP-tree
  - ทำงานได้เร็วกว่าวิธี Apriori

การใช้งานโปรแกรม

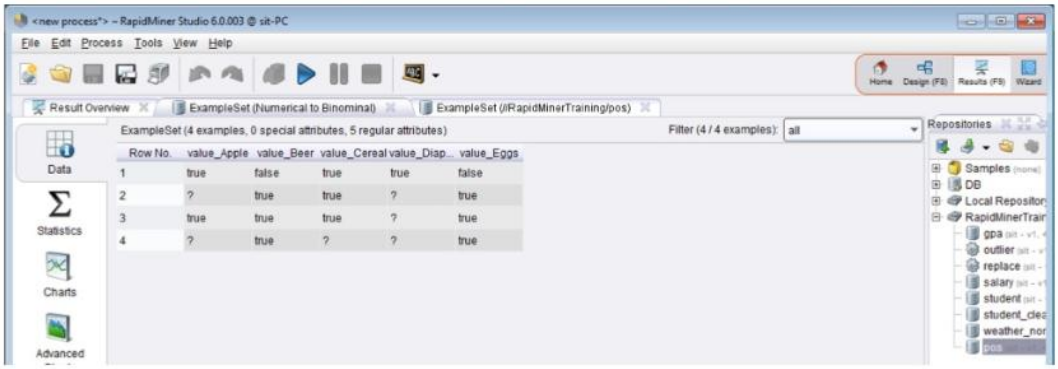
- โอเปอเรเตอร์ที่เกี่ยวข้อง

| โอเปอเรเตอร์ | | คำอธิบาย |
|---|---|---|
|  | Retrieve | ใช้สำหรับดึงข้อมูลออกมาจาก Repositories |
|  | Pivot | ใช้สำหรับสร้างตารางในรูปแบบ Pivot |
|  | Numerical to Binominal | ใช้สำหรับแปลงข้อมูลตัวเลขให้เป็นประเภท Binominal |

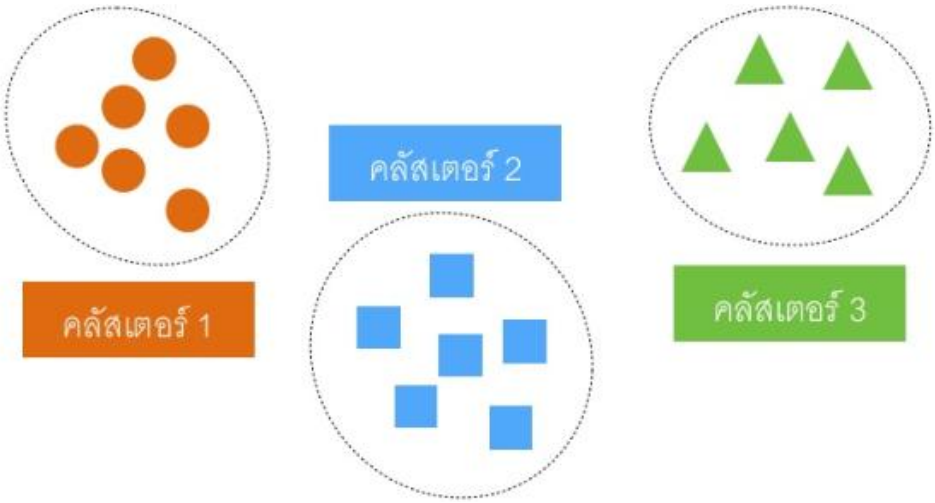
- เปลี่ยนพารามิเตอร์ของโอเปอเรเตอร์ Pivot ดังนี้
  - group attribute เป็นค่า id
  - index attribute เป็นค่า name




- ผลการแปลงจาก POS database



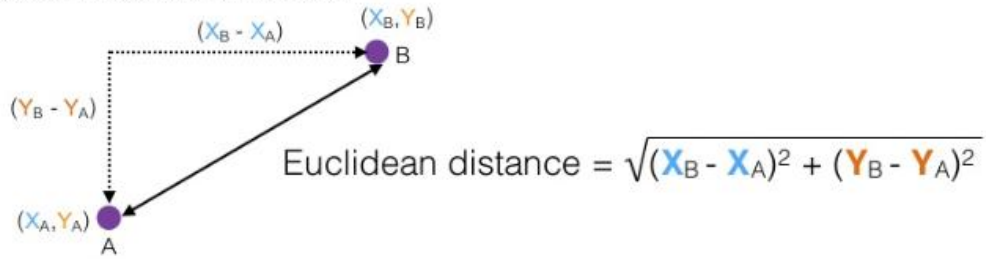
การทำ Clustering คือการแบ่งกลุ่มข้อมูล ข้อมูลที่มีลักษณะคล้ายๆ กันอยู่กลุ่มเดียวกัน ข้อมูลที่อยู่คนละกลุ่มจะมีลักษณะที่แต่ต่างกันมากๆ แต่ละกลุ่มจะเรียกว่าคลัสเตอร์ (cluster)




การจัดข้อมูลให้อยู่ในกลุ่มต่างๆ จะต้องมีการวัดค่าความคล้ายคลึง (similarity) หรือค่าระยะห่าง (distance) ระหว่างข้อมูลแต่ละตัว

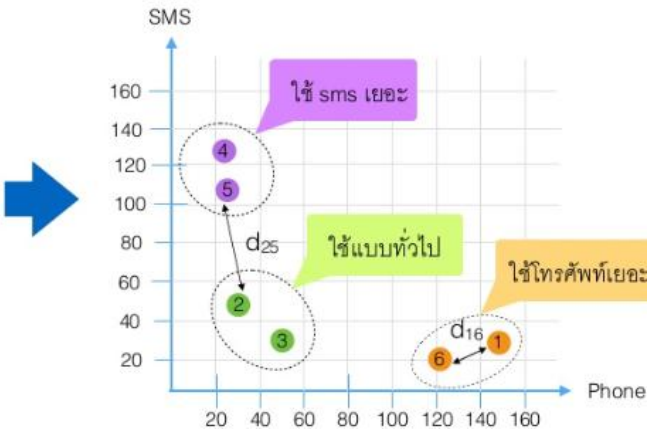
ค่าระยะห่างที่นิยมใช้ เช่น ระยะห่างยูคลิเดียน (Euclidean distance)

- ผมรวมของระยะห่างกำลังสอง




ตัวอย่างการทำ clustering ลูกค้าตามพฤติกรรมการใช้ SMS และ Phone

| ID | SMS | Phone |
|---|---|---|
| 1 | 27 | 144 |
| 2 | 32 | 44 |
| 3 | 41 | 30 |
| 4 | 125 | 21 |
| 5 | 105 | 23 |
| 6 | 20 | 121 |



การประยุกต์ใช้ เช่น แบ่งข้อมูลลูกค้าออกเป็นกลุ่มย่อยๆ เพื่อจะได้เข้าใจพฤติกรรมการบริโภคของลูกค้าได้ดีขึ้น ส่งโปรโมชันได้ตรงกับความต้องการของลูกค้าแต่ละกลุ่ม ช่วยในการจัดกลุ่มเอกสารตามความคล้ายคลึง

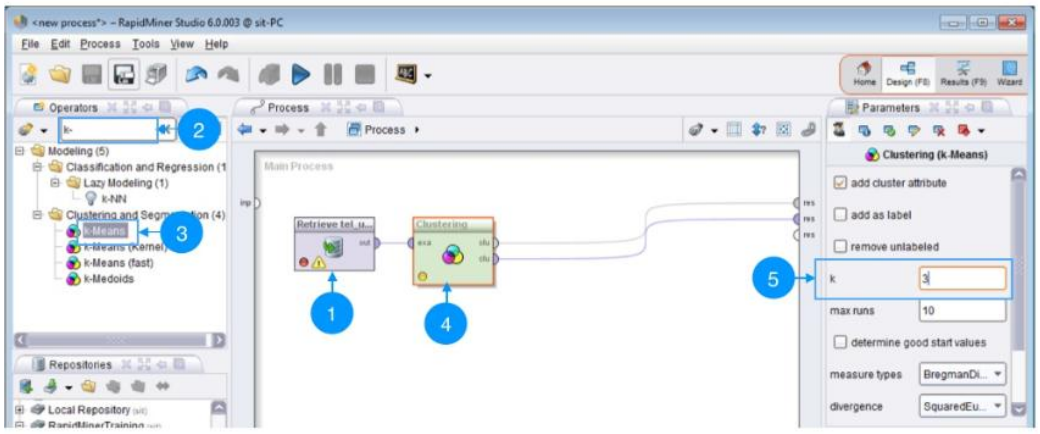
เทคนิคที่นิยมใช้ในการทำ clustering

- K-Means เป็นประเภท partitional clustering ที่นิยมมากที่สุด
- ต้องกำหนดจำนวนกลุ่ม หรือคลัสเตอร์ที่ต้องการจะแบ่งในตัวแปร k
- วิธีการทำงาน
  - กำหนดจุดศูนย์กลางของแต่ละคลัสเตอร์
  - หาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง (mean) ของแต่ละคลัสเตอร์
  - กำหนดให้ข้อมูลที่อยู่ในคลัสเตอร์ใกล้ที่สุด
  - คำนวณหาจุดศูนย์กลางของแต่ละคลัสเตอร์ใหม่
  - ทำซ้ำจนข้อมูลอยู่ในคลัสเตอร์เดิมไม่มีการเปลี่ยนแปลง

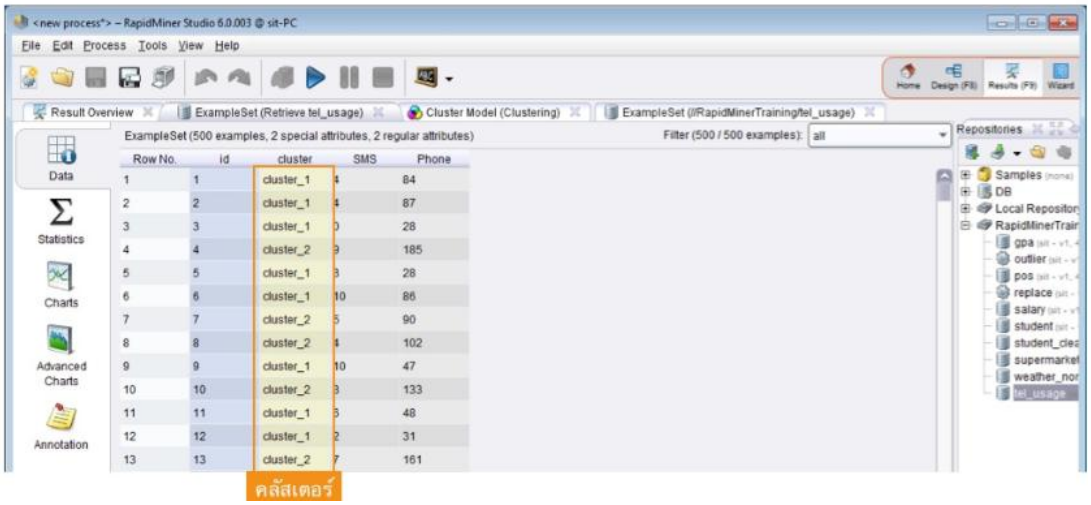
- โอเปอเรเตอร์ที่เกี่ยวข้องกับ K-Means

| โอเปอเรเตอร์ | | คำอธิบาย |
|---|---|---|
|  | Retrieve | ใช้สำหรับดึงข้อมูลออกมาจาก Repositories |
|  | K-Means | ใช้สำหรับแบ่งกลุ่มข้อมูลด้วยเทคนิค K-Means |

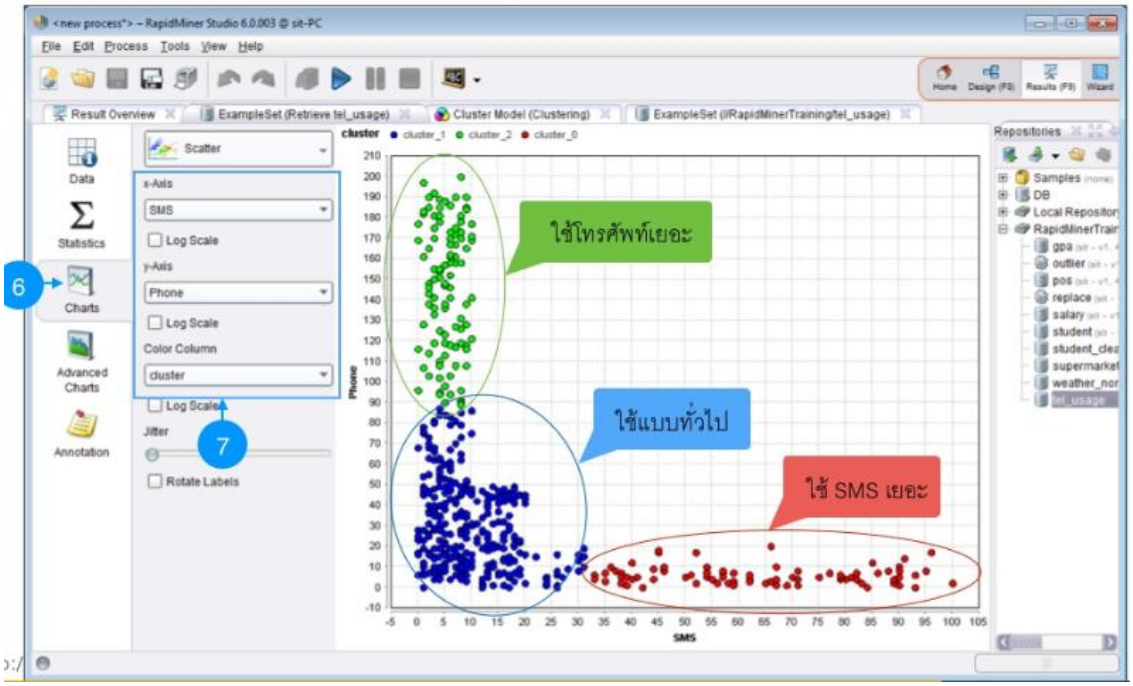
- โหลด tel_usage จาก Repositories
- กำหนดจำนวนกลุ่มของ K-Means ให้เป็น 3 กลุ่ม



- ผลการแบ่งกลุ่มข้อมูลออกเป็น 3 คลัสเตอร์

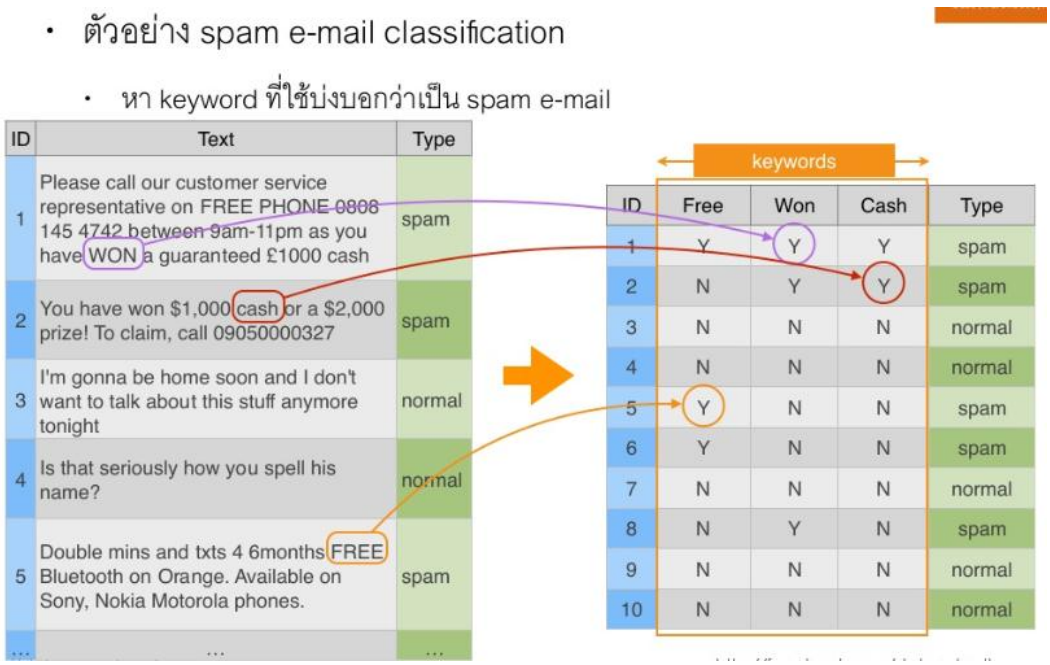



- แสดงการกระจายตัวของข้อมูลในแต่ละคลัสเตอร์

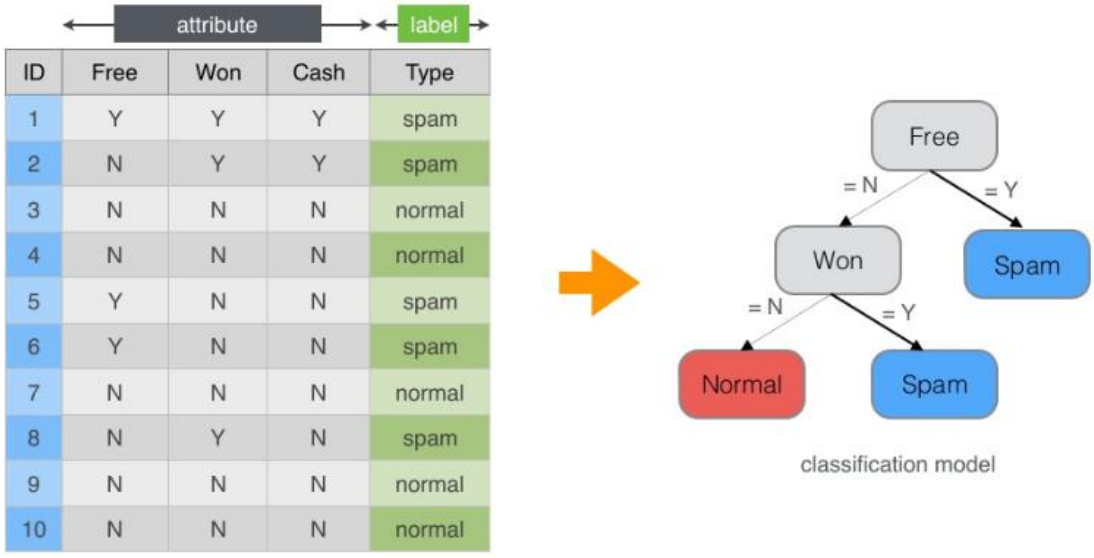



- การทำ Agglomerative clustering เป็นประเภท hierarchical clustering
- การทำ DBSCAN (Density Based Spatial Clustering of Applications with Noise) เป็นประเภท density-base clustering

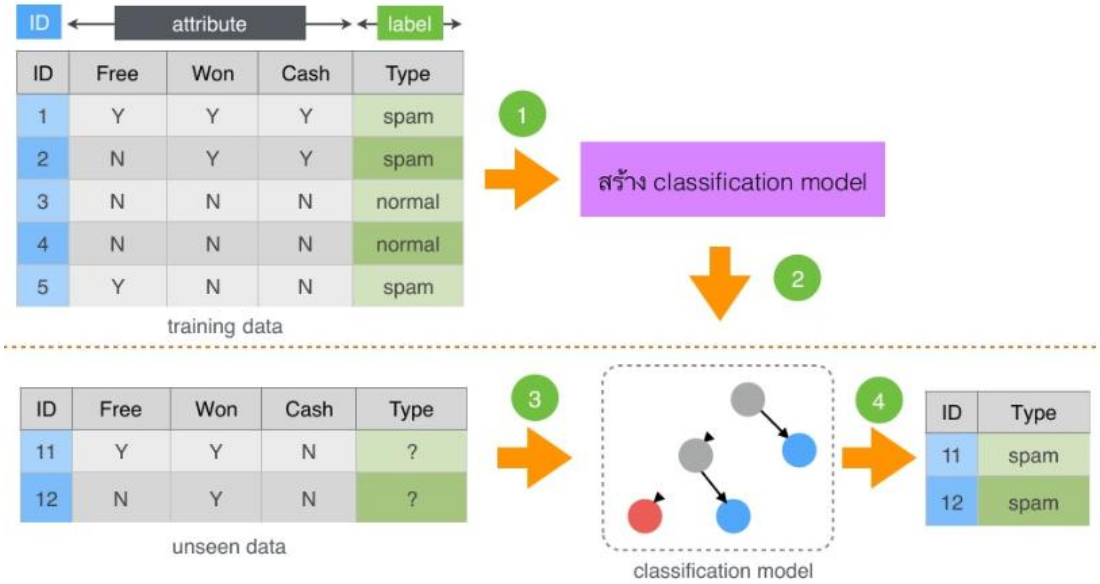
การทำ Classification ตัวอย่างเช่น การทำ spam e-mail classification โดยระบุว่าอีเมลไหนเป็น spam อีเมลไหนเป็นเมลปกติ



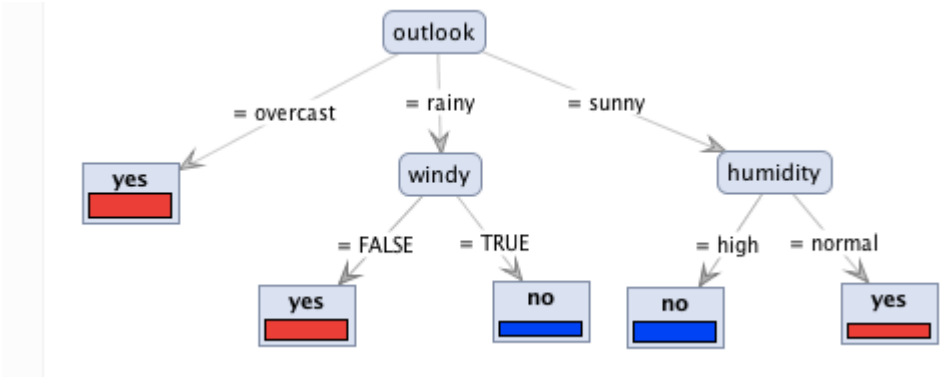

ตัวอย่าง spam e-mail classification สร้างดมเดลจาก training data ซึ่งมีลาเบล (lable)




นำข้อมูลใหม่ (unseen data ) มาทำนายโดยใช้โมเดล




Classification Techniques โดยใช้ Decision Tree สร้างกฎได้จาก Decision Tress โดยการใส่ไปตามแต่ละ Path ของ Tree



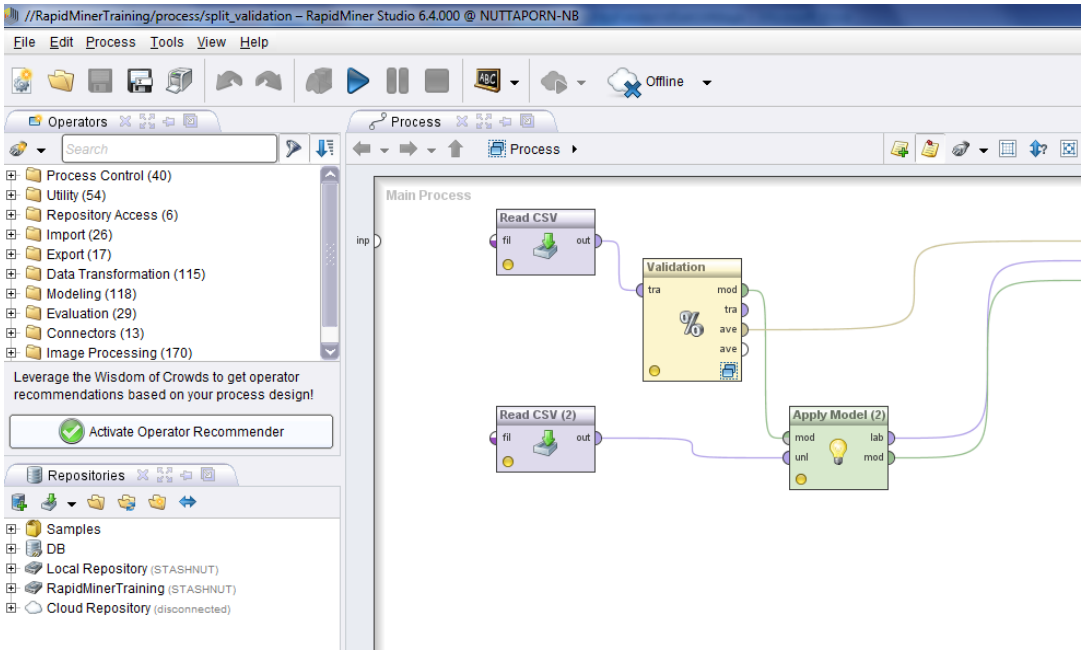

Decision Tree เป็นเทคนิคที่นิยมใช้ในการทำ Classification ขั้นตอนการสร้าง decision tree จะเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมาใช้งาน

- คำนวณค่า Entropy และ Information Gain (IG)

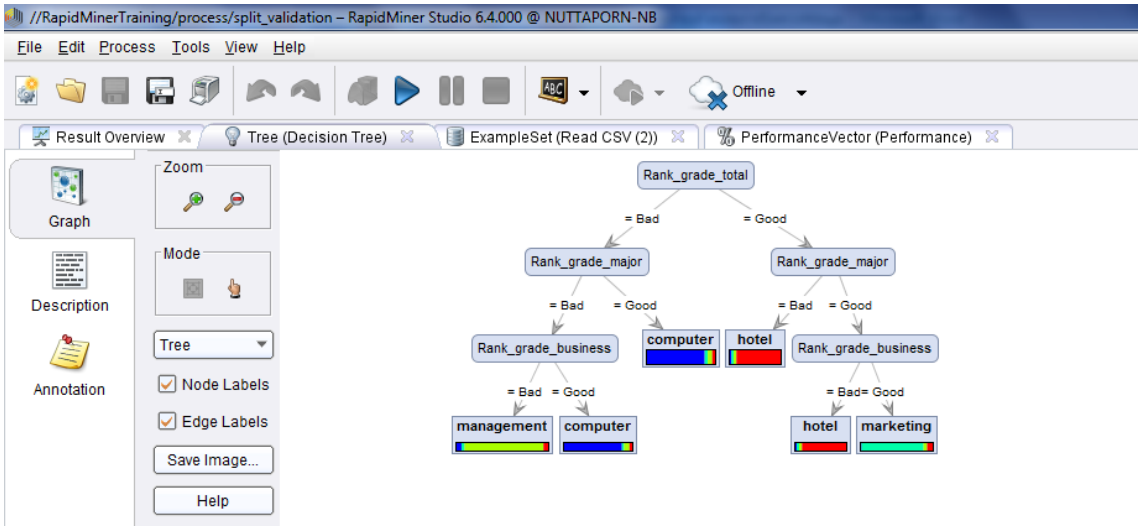
  Entropy (c1) = - p(c1) log p(c1)

  IG (parent, child) = Entropy(parent )-[p(c1) x Entropy(c1) + p(c2) x Entropy (c2)+…]

ตัวอย่างการใช้โมเดล Decision Tree ที่สร้างได้เพื่อทำนายผลการเลือกสาขาวิชาของนักศึกษา




ผลลัพธ์ที่ได้จากการ run




เทคนิคการใช้ Naïve Bayes จะใช้คำนวณค่าความน่าจะเป็น (probability)

โอกาสที่เกิดเหตุการณ์จากเหตุการณ์ทั้งหมด ใช้สัญลักษณ์ P() หรือ Pr() เช่น

โยนเหรียญบาท (มีหัวและก้อย) โอกาสได้หัว มีความน่าจะเป็น ½ = 0.5

โอกาสได้ก้อย มีความน่าจะเป็น ½ = 0.5

ตัวอย่าง ความน่าจะเป็นของการพบ spam mail เมื่อมี email ทั้งหมด 100 ฉบับ มี spam email ทั้งหมด 20 ฉบับ มี normal email ทั้งหมด 80 ฉบับ โอกาสที่ email จะเป็น spam มีความน่าจะเป็น 20/100 = 0.2 หรือ P(spam)= 0.2 โอกาสที่ email จะเป็น normal มีความน่าจะเป็น 80/100 = 0.8 หรือ P(normal) = 0.8

Naïve Bayes จะใช้หลักการของความน่าจะเป็นแบบมีเงื่อนไข

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

โดยการประยุกต์เป็น $P(C|A) = \frac{P(A|C) \times P(C)}{P(A)}$

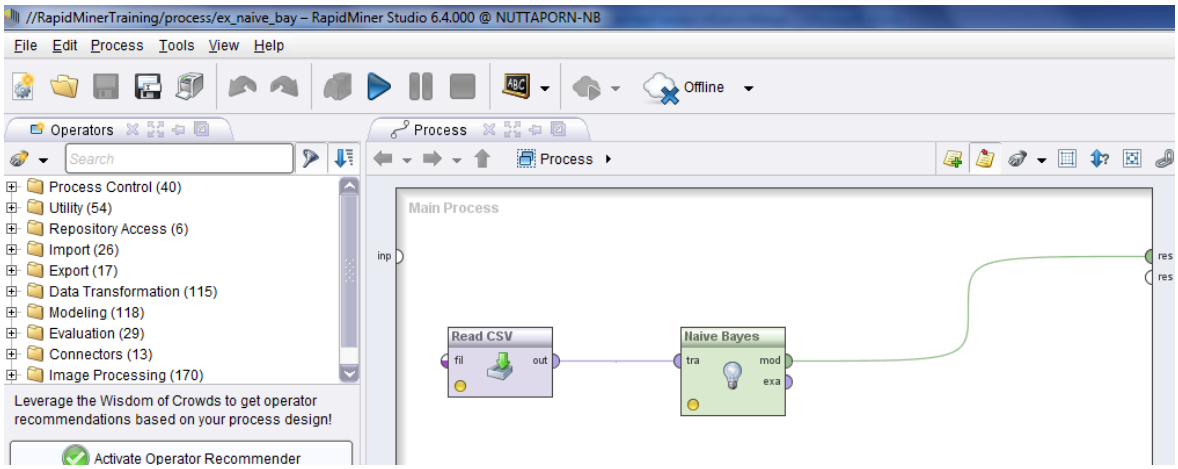
เมื่อ P(C|A) คือความน่าจะเป็นของข้อมูลที่มีแอตทริบิวต์ A จะมีคลาส C

P(A|C) คือ ความน่าจะเป็นของข้อมูลใน training data ที่มีแอตทริบิวต์ A และมีคลาส C

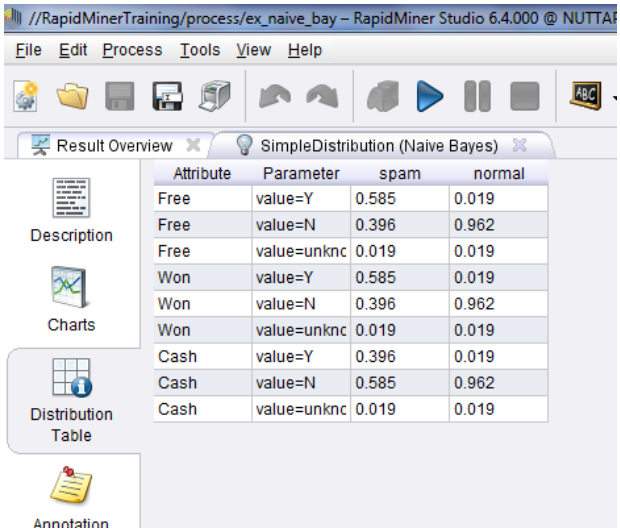
$P(A|C) = P(a_1 \cap a_2 \cap a_3 \ldots \cap a_m|C)$

$P(A|C) = P(a_1|C) \times P(a_2|C) \times P(a_3|C) \times \ldots P(a_m|C)$

P(C) หรือ P(A) คือ ความน่าจะเป็นของคลาส C หรือแอตทริบิวต์ A



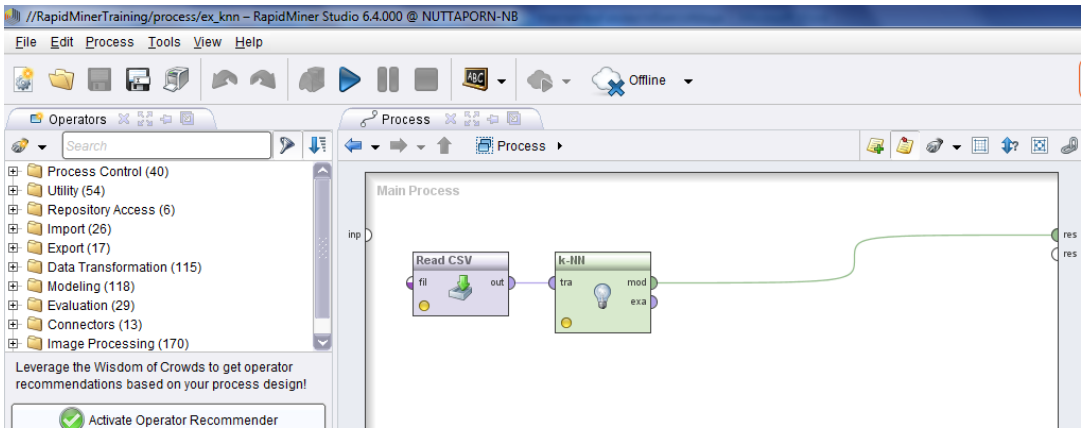

ผลจากการ run Naïve Bayes



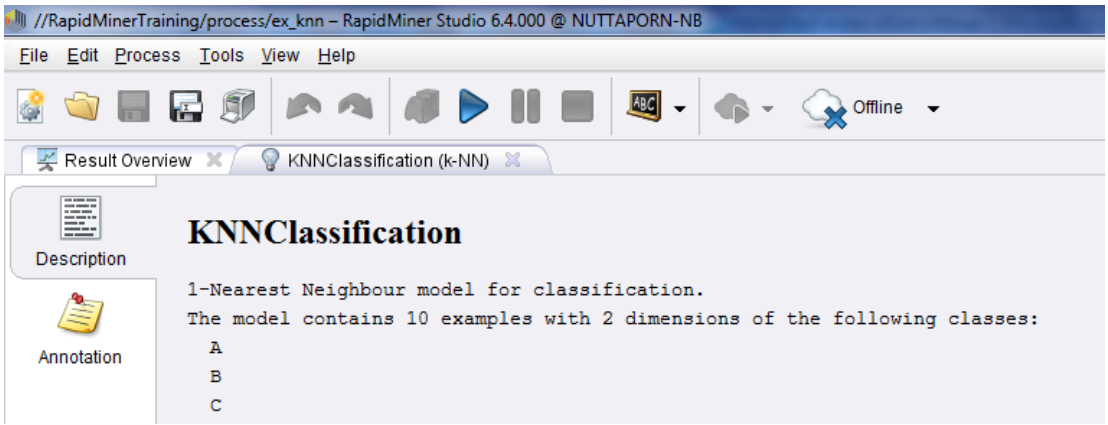

| Attribute | Parameter | spam | normal |
|---|---|---|---|
| Free | value=Y | 0.585 | 0.019 |
| Free | value=N | 0.396 | 0.962 |
| Free | value=unknc | 0.019 | 0.019 |
| Won | value=Y | 0.585 | 0.019 |
| Won | value=N | 0.396 | 0.962 |
| Won | value=unknc | 0.019 | 0.019 |
| Cash | value=Y | 0.396 | 0.019 |
| Cash | value=N | 0.585 | 0.962 |
| Cash | value=unknc | 0.019 | 0.019 |

การ Classification โดยใช้ K-nearest Neighbors (kMN) เป็นการจำแนกข้อมูลที่มีลักษณะคล้ายๆ กันให้อยู่ใน class เดียวกัน โดย K เป็นจำนวนข้อมูล ตัวอย่างการให้ยาของคนไข้โดยดูจากอายุและค่า Na/K

| ID | Na/k | Age | Type |
|---|---|---|---|
| 1 | 10 | 20 | A |
| 2 | 25 | 15 | A |
| 3 | 10 | 15 | A |
| 4 | 50 | 15 | B |
| 5 | 55 | 20 | B |
| 6 | 60 | 30 | B |
| 7 | 35 | 40 | C |
| 8 | 25 | 50 | C |
| 9 | 30 | 60 | C |
| 10 | 40 | 55 | C |




ผลลัพธ์จากการ Run K-nearest Neighbors (kMN)



การทำ Text Mining เบื้องต้น

ข้อมูลแบ่งเป็นแบบมีโครงสร้าง (structure) และไม่มีโครงสร้าง (unstructure)

ข้อมูลแบบมีโครงสร้าง เช่นข้อมูลเก็บในรูปแบบตาราง

| ID | outlook | humidity | windy | play |
|---|---|---|---|---|
| 1 | Sunny | High | FALSE | No |
| 2 | Sunny | High | TRUE | No |
| 3 | Overcast | Normal | FALSE | yes |

←numeric →← nominal →←———— binominal ————→

ข้อมูลแบบไม่มีโครงสร้าง เช่น ข้อมูลที่เป็นข้อความ ข้อมูลที่เป็นรูปภาพ ข้อมูลที่ไม่มีโครงสร้างที่เก็บอยู่ในรูปแบบข้อความ รูปภาพ เสียงมีจำนวนมากถึง 80%ของข้อมูลทั้งหมด เช่น ข้อความใน Social Network ต่างๆ เช่น Facebook, Twitter, LinkedIN ข้อความในเอกสารต่างๆ เช่น email, SMS , รายงานต่างๆ ข้อความในข่าวต่างๆ เช่น หนังสือพิมพ์, Google News

การประยุกต์ใช้ข้อมูลประเภทข้อความ เช่น การหาข่าวที่ใกล้เคียง การวิเคราะห์ทัศนคติในแง่ต่างๆ จากสังคมออนไลน์

ในการวิเคราะห์ข้อมูลข้อความต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่มีโครงสร้าง เช่น การนับความถี่ของคำ (ที่สนใจจะแปลความหมาย) ที่เกิดขึ้นในข้อความ ในที่นี้เป็นข้อความภาษาอังกฤษ โดยทำการแปลงคำให้เป็นนรากศัพท์ (root) เช่น finding แปลงเป็น find ตัดคำเชื่อมหรือคำที่เป็นบุพบททิ้ง และนับคำที่เกิดขึ้นในแต่ละเอกสร ถ้ามีคำนั้นให้เป็น 1 ถ้าไม่เกิดขึ้นให้เป็น 0 แล้วพิจารณาความถี่หรือจำนวนครั้งที่คำนั้นเกิดขึ้นในเอกสารทั้งหมด (Term Frequency)

-TF-IDF คือจำนวนครั้งของคำที่เกิดขึ้นคูณกับจำนวนคำที่เกิดขึ้นเฉพาะเอกสารในคลาส

- จำนวนคำท่พิจารณาต่อกัน N ตัว

unigram พิจารณาการเกิดของแต่ละคำ

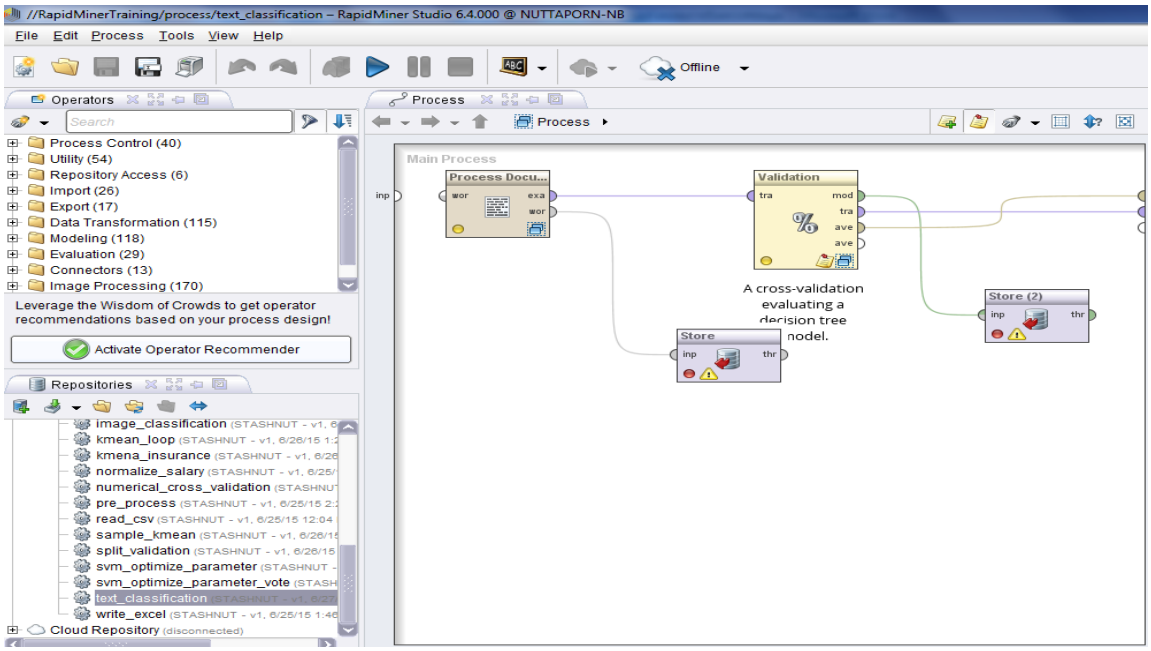
bi-gram พิจารณาการเกิดขึ้นของคำที่ติดกัน 2 คำ เช่น so good

tri-gram พิจารณาการเกิดขึ้นของคำที่ติดกัน 3 คำ เช่น smell so good

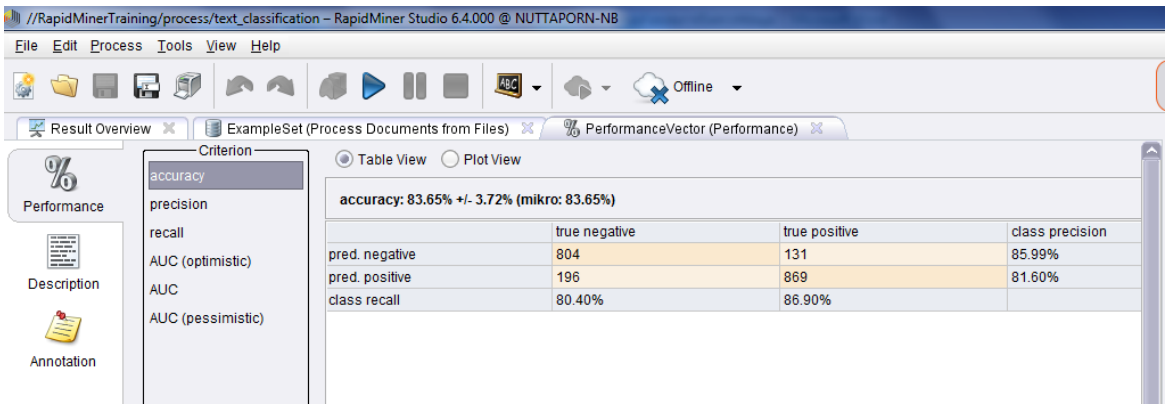
วิธีการ โดยการติดตั้ง text mining plugin ลงใน RapidMiner Studio 6 โดยเลือกเมนู Help>Updates and Extension (Marketplace)… ค้นหา plugin ที่ชื่อว่า text mining โดยมีกระบวนการดังนี้

| โอเปอเรเตอร์ | คำอธิบาย |
|---|---|
| Process Document from files | อ่านข้อความต่างๆ และคำนวนค่า term representation |
| X-Validation | แบ่งข้อมูลสำหรับสร้างและทดสอบโมเดลแบบ 10-flod |

| | |
|---|---|
| | cross-validation |
| Store | สำหรับบันทึกผลลงใน Repositories |
| Tokenize | สำหรับตัดข้อความ text ออกเป็นคำศัพท์ต่างๆ (term) |
| Filter Token (by Length) | สำหรับกรองคำศัพท์ที่มีความยาวน้อยกว่าหรือมากกว่าที่กำหนด |
| Stem (Porter) | สำหรับแปลงคำให้อยู่ในรูปของรากศัพท์ |
| Filter Stopwords (English) | สำหรับลบคำที่เป็น Stopwords เช่น an, an, the |
| SVM(linear) | สร้างโมเดล Support Vector Machines ด้วย Linear Kernal |




## ผลลัพธ์ของโมเดล



**4. ประโยชน์ที่ได้รับ**

1. ได้ความรู้และทักษะในการใช้โปรแกรม RapidMiner Studio 6 ในการทำ Data Mining
2. นำความรู้และประสบการณ์ที่ได้ไปผลิต/ปรับปรุงชุดวิชาทางของแขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร
3. นำความรู้และประสบการณ์ที่ได้ไปใช้ในงานวิจัย

**คำชี้แจงการใช้เอกสาร**

ขอขอบคุณที่ท่านให้ความสนใจศึกษาเอกสารเผยแพร่ความรู้ (KM) ของสาขาวิชาวิทยาศาสตร์และเทคโนโลยี มสธ. ซึ่งจัดทำขึ้นเพื่อเผยแพร่ให้เกิดประโยชน์เชิงวิชาการในวงกว้าง ทั้งนี้ หากท่านนำข้อมูลจากเอกสารนี้ไปใช้ประโยชน์ ขอให้อ้างอิงแหล่งที่มาของเราด้วย พร้อมทั้งแจ้งให้เราทราบถึงแหล่งที่ท่านนำไปใช้อ้างอิง โดยแจ้งมาทางอีเมล stoffice@stou.ac.th เพื่อประโยชน์ในการบูรณาการข้อมูลร่วมกัน