

ดาวน์โหลด Summary Note บทที่ 2 เรื่อง CRISP-DM



Summary Note จากหลักสูตร

Practical Data Mining with RapidMiner Studio

Chapter 2

- ในการทำ project ทางด้าน Data Mining หรือ Machine Learning มีแนวทางอย่างหนึ่งที่ใช้กันอย่างแพร่หลาย นั่นคือ **CRISP-DM**

- CRISP-DM ย่อมาจากคำว่า **C**ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

- กระบวนการนี้มีทั้งหมด 6 ขั้นตอน คือ

1. Business Understanding
2. Data Understanding

[facebook.com/datacube.th](https://datacube.th)

- กระบวนการนี้มีทั้งหมด 6 ขั้นตอน คือ
 3. Data Preparation
 4. Modeling
 5. Evaluation
 6. Deployment
- โดยส่วนใหญ่เวลาจะหมดไปกับ 3 ขั้นตอนแรกซึ่งเป็นเรื่องของการจัดการข้อมูล โดยอาจจะใช้เวลาประมาณ 70% - 80% ของโครงการทั้งหมด

- ขั้นตอนแรกของ CRISP-DM คือ **Business Understanding** เป็นการดูว่าจะ**ประยุกต์ใช้การวิเคราะห์ข้อมูลมาช่วยงานอย่างไรได้บ้าง** ตัวอย่างเช่น
 - การคาดการณ์ปริมาณสินค้าที่ต้องผลิตสำหรับวางจำหน่ายในอีก 3 เดือนข้างหน้า หรือเรียกว่า **Demand Forecasting**
 - การคาดการณ์ว่าลูกค้าคนใดบ้างมีโอกาสจะยกเลิกการใช้บริการในอีก 90 วันถัดไป หรือเรียกว่า **Churn Prevention**
 - การคาดการณ์ว่าลูกค้าคนใดบ้างควรได้รับสินเชื่อที่ขอกู้มา หรือเรียกว่า **Credit Scoring**

- การคาดการณ์ว่าเครื่องจักรใดบ้างมีโอกาสจะทำงานผิดปกติในอีก 24 หรือ 48 ชั่วโมงถัดไป หรือเรียกว่า **Predictive Maintenance**
- การคาดการณ์ว่าลูกค้าคนใดบ้างจะมีโอกาสซื้อสินค้าที่นำเสนอไป หรือเรียกว่า **Propensity to Buy** หรือ **Response Modeling**
- การแบ่งกลุ่มลูกค้าตามพฤติกรรม การซื้อสินค้าต่างๆ หรือเรียกว่า **Customer Segmentation**
- การแนะนำสินค้าที่ลูกค้าน่าจะสนใจ ให้ หรือเรียกว่า **Cross-Sell**

- ขั้นตอนที่ 2 ของ CRISP-DM คือ **Data Understanding** เป็นการพิจารณาว่า**มีข้อมูลอะไรบ้างที่จะนำมาวิเคราะห์เพื่อตอบโจทย์**ในขั้นตอน Business Understanding ได้บ้าง
- ข้อมูลมักจะ**กระจัดกระจายอยู่ตามตารางต่างๆ** ดังนั้นในขั้นตอนนี้คือ
 - ดูว่ามีข้อมูลเก็บไว้ที่ใดบ้าง
 - ข้อมูลมีความหมายอะไรบ้าง เช่น ข้อมูลจากเซ็นเซอร์ (sensor) ได้มาทุก 15 นาที หรือ รหัสที่ใช้ในข้อมูลนี้หมายความว่าอย่างไร
 - ข้อมูลมีปริมาณมากเพียงพอไหม
 - ข้อมูลมีปัญหาอะไรบ้าง

- ขั้นตอนที่ 3 ของ CRISP-DM คือ **Data Preparation** เป็นขั้นตอนหลังจากเราทำความเข้าใจข้อมูลแล้วว่าอยู่ที่ไหน มีตัวแปรอะไรบ้าง ขั้นตอนนี้แบ่งได้เป็น 3 ขั้นตอนย่อยๆ คือ
 - ขั้นตอนที่ 3.1 คือ **การเลือกข้อมูล (Data Selection)** เป็นการเลือกข้อมูลหรือตัวแปรเพื่อใช้ในการวิเคราะห์ข้อมูล
 - ข้อมูลบางอย่างมีการเกิดขึ้นค่อนข้างบ่อย เช่น ข้อมูลจาก เซนเซอร์ ข้อมูลการซื้อขาย ข้อมูลการใช้งานบัตรเครดิต ข้อมูลการใช้งานโทรศัพท์

- ข้อมูลเหล่านี้ควรจะย้อนหลังไปไม่มากนักอาจจะสัก 1-2 ปีก็มีข้อมูลที่มากเพียงพอสำหรับการวิเคราะห์แล้ว เพราะถ้าใช้ข้อมูลย้อนหลังมากเกินไปอาจจะเจอเทคโนโลยีที่ไม่ใช้งานแล้วก็เป็นได้
- แต่สำหรับข้อมูลบางประเภทที่นานๆจะมีเกิดขึ้นสักครั้ง เช่น การซื้อบ้าน ซื้อรถยนต์ ข้อมูลในลักษณะนี้เราอาจจะต้องใช้ประวัติการซื้อสินค้าย้อนหลังไปนานเสียน้อย อาจจะ เป็นสัก 5-10 ปีเลยทีเดียว
- ขั้นตอนที่ 3.2 คือ **การแก้ไขข้อมูลที่ผิดพลาด (Data Cleansing)** เช่น แก้ไขข้อมูลที่เป็นค่าว่าง (missing value) หรือข้อมูลที่เป็น Outlier

- ขั้นตอนที่ 3.3 คือ **การแปลงข้อมูล (Data Transformation)** โดยปกติแล้วข้อมูลจะไม่ได้อยู่ในรูปแบบที่นำไปใช้งานต่อได้ จึงต้องมีการสร้างตัวแปรใหม่ เช่น การคำนวณครั้งในการซื้อสินค้า หรือ การทำ pivot เพื่อให้การซื้อสินค้าแต่ละครั้งเป็นแถวเดียวกัน
- ทั้งสามขั้นตอนนี้ใช้เวลาค่อนข้างนาน
- ขั้นตอนที่ 4 ของ CRISP-DM คือ **Modeling** เป็น**การประยุกต์ใช้เทคนิคการวิเคราะห์ข้อมูล** โดยวิธีการวิเคราะห์ข้อมูลนี้จะแบ่งได้เป็น 2 แบบหลักๆ ได้แก่

- **Unsupervised Learning** เป็นการวิเคราะห์ข้อมูลแบบที่**ไม่มีคำตอบ (class หรือ label) มาให้** ตัวอย่างของเทคนิคแบบนี้คือ การจัดกลุ่มข้อมูลที่มีลักษณะคล้ายกันไว้ด้วยกัน
- **Supervised Learning** เป็นการวิเคราะห์ข้อมูลที่มี**คำตอบ (class หรือ label) มาให้แล้ว** ตัวอย่าง เช่น การคาดการณ์ว่าลูกค้าคนใดบ้างจะยกเลิกการใช้บริการ (Churn)
- ขั้นตอนที่ 5 ของ CRISP-DM คือ **Evaluation** เป็น**การวัดผล** ซึ่งจะแบ่งเป็นการวัดประสิทธิภาพเพื่อเปรียบเทียบว่าเทคนิคใดดีกว่ากัน และเป็นการย้อนกลับไปดูว่าสามารถตอบโจทย์ในขั้นตอนแรกได้แล้วหรือยัง

- ขั้นตอนที่ 6 ของ CRISP-DM คือ **Deployment** เป็น**ขั้นตอนการนำผลที่ได้ไปใช้งานต่อ** โดยอาจจะเป็น
 - การนำไปใช้โดยการ**สร้างรายงาน**
 - การส่งชื่อ เบอร์โทรศัพท์หรือ email ไปให้ฝ่ายอื่นๆ ทำงานต่อ
 - การพัฒนาเป็นระบบ **web application**
- ในบทนี้จะแสดงตัวอย่างการใช้งาน 2 ตัวอย่างได้แก่
 - การแนะนำนักศึกษาให้เลือกรเรียนในภาควิชาที่เหมาะสม
 - การคาดการณ์ว่าลูกค้าคนใดบ้างจะยกเลิกการใช้บริการ