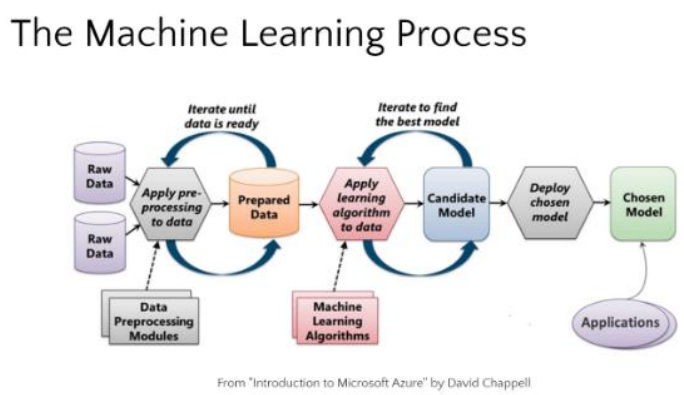
Data Preprocessing Checklist

Data Preprocessing คืออะไร?

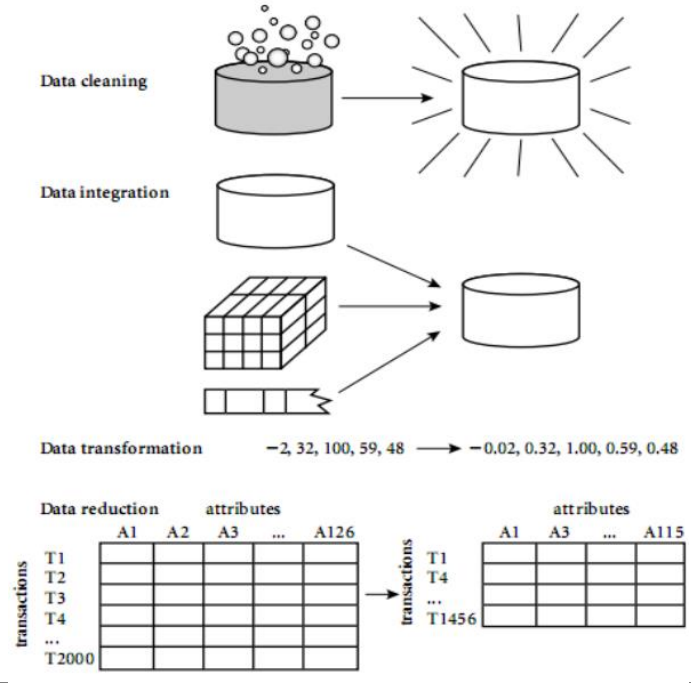
กระบวนการแปลงข้อมูลดิบให้อยู่ในรูปแบบที่สามารถเข้าใจได้ สำหรับการนำข้อมูลไปใช้ต่อในการทำ Data Mining, Machine Learning และงาน Data Science อื่นๆ



ทำไมถึงต้องทำ Data Preprocessing?

เพื่อให้ได้ข้อมูลที่มีคุณภาพ เนื่องจากประสิทธิภาพของโมเดลนั้นขึ้นอยู่กับคุณภาพของข้อมูล

ขั้นตอนของ Data Preprocessing



Data Quality Assessment - การกำหนดคุณภาพของข้อมูล

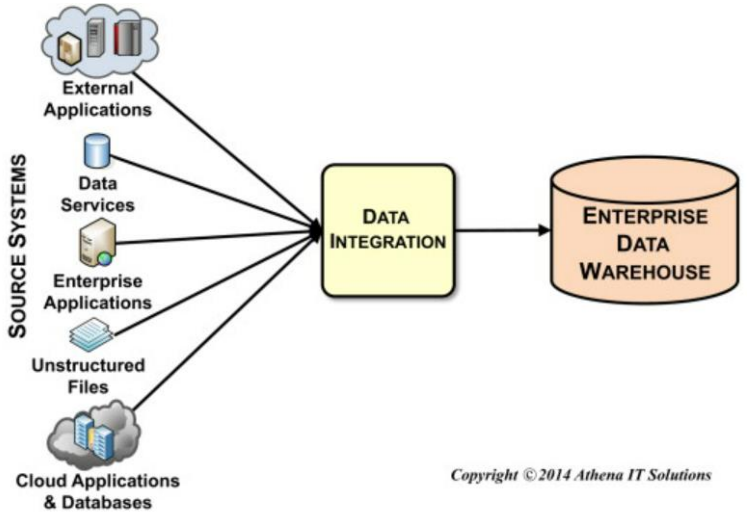
- ❑ ไม่มีข้อมูลสูญหาย
- ❑ ข้อมูลมีความถูกต้องและแหล่งข้อมูลมีความน่าเชื่อถือ
- ❑ ข้อมูลมีความสอดคล้องกัน
- ❑ ข้อมูลต้องไม่มีความซ้ำซ้อน

Data Cleaning - การ Clean ข้อมูลเพื่อให้ได้ข้อมูลที่ถูกต้องและสมบูรณ์ที่สุด

- ❑ การจัดการกับข้อมูลที่สูญหาย (Missing Data) ใช้วิธีเติมข้อมูลแบบ Manual หรือใช้วิธีการทำนายข้อมูลที่หายไป เช่น Mean, Regression Method เป็นต้น
- ❑ การจัดการกับข้อมูลที่เป็นสิ่งรบกวน (Noisy Data) เช่น ใช้วิธี Binning, Regression, Clustering เป็นต้น
- ❑ การจัดการกับ Outliers เช่น ใช้วิธี Clustering เป็นต้น

Data Integration

- ❑ การรวบรวมข้อมูลจากหลายแหล่งและจัดเก็บข้อมูลขนาดใหญ่เพียงแห่งเดียว เช่น คลังข้อมูล
- ❑ รวมถึงการลดความซ้ำซ้อนของข้อมูล และการจัดการกับข้อมูลที่ไม่สอดคล้องกันด้วย



Data Reduction - การแสดงชุดข้อมูลที่ลดลง แต่ยังให้ผลการวิเคราะห์ที่มีคุณภาพเท่ากัน

- ❑ Attribute Selection/ Feature Selection: การเลือกเฉพาะ Attribute ที่มีผลกับประสิทธิภาพของโมเดล
- ❑ Numerosity Reduction: การลดข้อมูลโดยแทนที่ข้อมูลเดิมด้วยการแสดงข้อมูลในรูปแบบที่เล็กกว่า แบ่งเป็น 2 เทคนิค

1) Parametric Methods: ข้อมูลจะถูกแสดงโดยใช้โมเดล แทนที่การเก็บข้อมูลจริง เช่น Regression เป็นต้น
2) Non-Parametric Methods: จัดเก็บการแสดงผลของข้อมูลที่ลดลง ได้แก่ Histograms, Clustering, Sampling เป็นต้น

- ❑ Dimensionality Reduction: เทคนิคที่ใช้เพื่อให้ได้ข้อมูลต้นฉบับที่ลดลง แบ่งออกเป็น 2 เทคนิค

1) Feature Selection: กระบวนการนำ Feature ที่ไม่เกี่ยวข้องหรือซ้ำซ้อนออก
2) Feature Extraction: กระบวนการลดจำนวน Feature ในชุดข้อมูล โดยการสร้าง Feature ใหม่จาก Feature เดิมที่มีอยู่ (แล้วละทิ้ง Feature เดิม)

Data Transformation - การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์

- ❑ Aggregation: การรวมข้อมูลทั้งหมดเข้าไว้ด้วยกัน โดยที่เป็นรูปแบบมาตรฐานเดียวกัน เช่น การรวมและแปลงข้อมูลเพื่อแสดงในรูปแบบรายเดือนและปี เป็นต้น
- ❑ Normalization: การปรับขนาดข้อมูลให้เป็นช่วงที่ทำให้เป็นมาตรฐาน เพื่อให้สามารถเปรียบเทียบได้แม่นยำยิ่งขึ้น เช่น Min-Max Normalization, Z-Score Normalization, Decimal Scaling Normalization เป็นต้น
- ❑ Feature Selection: คุณสมบัติใหม่ของข้อมูลถูกสร้างขึ้นจากแอตทริบิวต์ที่มีอยู่เพื่อช่วยในกระบวนการทำเหมืองข้อมูล
- ❑ Discretization: เหมาะสำหรับข้อมูลต่อเนื่องที่แบ่งออกเป็นช่วงเวลา เป็นการรวมข้อมูลให้เป็นช่วงที่เล็กลง ค่อนข้างคล้ายกับ Binning แต่มักจะเกิดขึ้นหลังจาก Cleaning ข้อมูลแล้ว
- ❑ Concept Hierarchy Generation: สามารถเพิ่มลำดับชั้นภายในและระหว่างคุณลักษณะที่ไม่มีอยู่ในข้อมูลต้นฉบับ ตัวอย่างเช่น หากเราวิเคราะห์ข้อมูลหมาป่า เราสามารถเพิ่มลำดับชั้นสำหรับ สกุล ได้

GBDi

Data Preprocessing

Training Course: Intermediate Data Engineering

Titirat Boonchuaychu (Data Engineer/ GBDi)