

ใบความรู้รายวิชา

รายวิชา: การวิเคราะห์ข้อมูล

รหัสวิชา: 31910-1003

ระดับ: ประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.)

หน่วยที่ 3: การทำ Data Cleansing (การทำความสะอาดข้อมูล)

จุดประสงค์การเรียนรู้

- อธิบายความสำคัญของการทำ Data Cleansing ได้
- เข้าใจวิธีการตรวจสอบและจัดการข้อมูลที่ผิดพลาด ขาดหาย หรือซ้ำซ้อน
- ปฏิบัติกรทำความสะอาดข้อมูลเบื้องต้นได้ด้วยเครื่องมือวิเคราะห์ข้อมูล เช่น Excel, Python (Pandas), หรือโปรแกรมอื่น ๆ

สาระสำคัญของหน่วย

ข้อมูล (Data) ที่ได้จากหลายแหล่งมักมีข้อผิดพลาด เช่น ข้อมูลที่หายไป (Missing Data), ข้อมูลซ้ำ (Duplicate Data) หรือข้อมูลไม่ถูกต้อง (Invalid Data) ซึ่งจะทำให้การวิเคราะห์ข้อมูลผิดพลาดได้ ดังนั้นจึงจำเป็นต้องทำ **Data Cleansing** หรือ “การทำความสะอาดข้อมูล” ก่อนการวิเคราะห์ เพื่อให้ข้อมูลมีคุณภาพ ถูกต้อง และพร้อมใช้งาน

เนื้อหา

3.1 การตรวจสอบข้อมูลที่หายไปหรือขาดหาย (Missing Data)

ความหมาย:

ข้อมูลบางส่วนไม่มีค่า เช่น ช่องว่างในตาราง หรือค่าที่เป็น “NULL” หรือ “NaN”

วิธีตรวจสอบ:

- ใช้ฟังก์ชัน `isnull()` หรือ `isna()` ใน Python (Pandas)
- ใช้คำสั่ง “Filter” หรือ “Conditional Formatting” ใน Excel เพื่อหาช่องว่าง

แนวทางแก้ไข:

- เติมค่าที่เหมาะสม (Imputation) เช่น ค่าเฉลี่ย (Mean) หรือค่ากลาง (Median)
- ลบแถวหรือคอลัมน์ที่ขาดข้อมูลมากเกินไป
- ตรวจสอบแหล่งข้อมูลเพิ่มเติมเพื่อกรอกค่าที่หายไปให้ถูกต้อง

3.2 การตรวจสอบข้อมูลที่ซ้ำกัน (Duplicate Data)

ความหมาย:

ข้อมูลที่ถูกบันทึกซ้ำ เช่น รายชื่อผู้สมัครที่ซ้ำกัน หรือรหัสสินค้าซ้ำ

วิธีตรวจสอบ:

- ใช้ฟังก์ชัน `drop_duplicates()` ใน Python
- ใช้ “Remove Duplicates” ใน Excel

แนวทางแก้ไข:

- ลบข้อมูลที่ซ้ำออก

- รวบรวมข้อมูลซ้ำให้เหลือเพียงแถวเดียว โดยเก็บข้อมูลสำคัญที่สุดไว้

3.3 การแก้ไขข้อมูลที่ไม่ถูกต้อง (Invalid Data)

ความหมาย:

ข้อมูลที่กรอกผิด เช่น อายุ = 250 ปี หรือวันที่ไม่ตรงรูปแบบ

แนวทางการแก้ไข:

- ตรวจสอบค่าที่อยู่นอกช่วง (Outliers)
- ปรับรูปแบบข้อมูลให้อยู่ในรูปแบบเดียวกัน (เช่น รูปแบบวันที่ YYYY-MM-DD)
- ตรวจสอบความสอดคล้องกับเงื่อนไข เช่น รหัสนักศึกษา ต้องมีตัวเลข 10 หลัก

3.4 การลบข้อมูลที่ไม่เกี่ยวข้อง (Irrelevant Data)

ความหมาย:

ข้อมูลที่ไม่เกี่ยวข้องกับการวิเคราะห์ เช่น หมายเลขโทรศัพท์ในข้อมูลการขายสินค้า

แนวทางการแก้ไข:

- วิเคราะห์วัตถุประสงค์ของการใช้ข้อมูล
- ลบคอลัมน์หรือแถวที่ไม่จำเป็น
- จัดกลุ่มข้อมูลใหม่ให้ตรงกับเป้าหมายการวิเคราะห์

3.5 การทำความสะอาดข้อมูลที่ไม่สมบูรณ์ (Incomplete Data)

ความหมาย:

ข้อมูลที่มีบางส่วนครบ แต่บางส่วนไม่สมบูรณ์ เช่น ข้อมูลที่ไม่มีหน่วยวัด หรือไม่มีคำอธิบาย

แนวทางการแก้ไข:

- ตรวจสอบและเติมเต็มส่วนที่ขาด
- ใช้ข้อมูลจากแหล่งอื่นมาช่วยเติมเต็ม
- ทำเครื่องหมายระบุข้อมูลที่ไม่สมบูรณ์เพื่อไม่ให้เกิดความสับสน

สรุป

การทำ Data Cleansing เป็นขั้นตอนสำคัญในการเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์ ช่วยลดข้อผิดพลาด เพิ่มความน่าเชื่อถือ และทำให้ผลการวิเคราะห์มีคุณภาพมากขึ้น