

หลักการวิเคราะห์ข้อมูลเบื้องต้น

คือ กระบวนการที่นำมาใช้ในการพิจารณาค้นหาความหมายของข้อมูลเพื่อช่วยในการตัดสินใจในองค์การ และสามารถใช้หลักการวิเคราะห์ข้อมูลเบื้องต้น ในชีวิตประจำวัน เช่น ข้อมูลการเงิน ข้อมูลค่าใช้จ่าย ซึ่งประกอบด้วยขั้นตอนดังนี้

๑. ระบุปัญหา (Define the Problem)
๒. เก็บรวบรวมข้อมูลที่เกี่ยวข้อง (Collect the data)
๓. เตรียมข้อมูลเบื้องต้น เพื่อให้ข้อมูลมีความสมบูรณ์มีคุณภาพ (Pre-Process the data)
๔. วิเคราะห์ข้อมูล โดยใช้หลักสถิติเบื้องต้น หรือ Machine Learning (Analyze the data)
๕. Interpret the results : ตีความหมายของผลลัพธ์ที่ได้ สามารถตอบปัญหาที่ตั้งไว้ได้หรือไม่ ถ้ายังไม่สามารถตอบได้ อาจย้อนกลับไปขั้นตอนการแรก โดยระบุปัญหาใหม่ หรือ เก็บข้อมูลเพิ่มเติม ทำกระบวนการทั้งหมดซ้ำอีกครั้ง

เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่

๑. Spreadsheets เช่น Excel หรือ Google sheet เหมาะกับข้อมูลขนาดเล็กถึงขนาดกลาง
๒. Statistical software วิเคราะห์ข้อมูลโดยใช้สถิติขั้นสูง
๓. Data mining tools เครื่องมือการทำเหมืองข้อมูล วิทยาการข้อมูล
๔. Business intelligence tools (BI) เช่น Power BI หรือ QlikView เครื่องมือช่วยในการสร้างกราฟในลักษณะ Dashboard
๕. Programming languages การเขียนโปรแกรมโดยใช้ภาษาคอมพิวเตอร์ เช่น ภาษา Python หรือ ภาษา R ใช้ในการวิเคราะห์ข้อมูลเชิงลึก ในรูปแบบ OpenSource

ชุดข้อมูลและตัวแปร

ชุดข้อมูล (Dataset) ประกอบไปด้วย รายการที่อธิบายวัตถุ ตั้งแต่ ๑ วัตถุขึ้นไป (วัตถุ=สิ่ง/เหตุการณ์ที่เราสนใจ)
ตัวแปร : คุณลักษณะที่อธิบายวัตถุที่เราสนใจ โดยประเภทของตัวแปร เป็นตัวกำหนดค่าที่เป็นไปได้ทั้งหมด (ขอบเขต) ของตัวแปร โดยแบ่งออก ๒ ประเภท

๑. ตัวแปรเชิงคุณภาพ คือ ตัวแปรที่ค่าที่เป็นไปได้ใช้บอกลักษณะการแบ่งเป็นประเภท เช่น เชื้อชาติ ,อาชีพ
๒. ตัวแปรเชิงปริมาณ คือ ตัวแปรที่ค่าเป็นตัวเลข แสดงความมากมาย นำไปคำนวณได้ (บอกความแตกต่างเป็นตัวเลขได้ ค่าเฉลี่ยมีความหมาย เช่น จำนวนบุตร ,รายได้

สถิติเพื่อการอธิบายข้อมูลเบื้องต้น

จะใช้สถิติเชิงพรรณนา (Descriptive Statistics) เพื่ออธิบายคุณลักษณะเบื้องต้นของข้อมูล ซึ่งแบ่งออกเป็น ๒ ประเภท ได้แก่

๑. การวัดค่าแนวโน้มเข้าสู่ส่วนกลาง (Central tendency) การหาค่าของข้อมูลเพื่อเป็นตัวแทนของข้อมูลทั้งหมด เป็นค่าที่อยู่กึ่งกลางระหว่างข้อมูลทั้งหมด ใช้เป็นตัวแทนชุดข้อมูลสำหรับตัวแปรแต่ละตัว ประเภทของของค่าแนวโน้มเข้าสู่ส่วนกลางที่สำคัญได้แก่ ค่าเฉลี่ยเลขคณิต(Arithmetic Mean) มัธยฐาน(Median) ฐานนิยม(Mode)

๒. การวัดการกระจายตัวของข้อมูล (Data Dispersion) การวัดว่าข้อมูลแต่ละตัวมีค่าแตกต่างกันมากน้อยเพียงใด การวัดแนวโน้มเข้าสู่ส่วนกลางเพียงอย่างเดียว เป็นการอธิบายลักษณะข้อมูลไม่สมบูรณ์ มีโอกาสคลาดเคลื่อนได้ ควรนำลักษณะการกระจายตัวของกลุ่มข้อมูลมาพิจารณาควบคู่ด้วย ค่าสถิติที่ใช้ในการวัดการวัดการกระจายตัวของข้อมูล ได้แก่ ค่าพิสัย (Range) (ค่าสูงสุด-ค่าต่ำสุด) และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

การเตรียมข้อมูล

ทำไมต้องเตรียมข้อมูล เนื่องจากถ้ามีข้อมูลที่ไม่ถูกต้อง และถูกนำไปวิเคราะห์ ผลลัพธ์ที่ได้ อาจทำให้การตัดสินใจผิดพลาด แต่ถ้าข้อมูลมีความถูกต้องเมื่อนำไปวิเคราะห์ ผลลัพธ์ที่ได้จะช่วยให้การตัดสินใจถูกต้อง โดยลักษณะข้อมูลที่ถูกต้องควรมี ๕ ลักษณะ คือ

๑. Consistency ความสอดคล้องกัน : ค่าของข้อมูลสอดคล้องกันในทุกแหล่งข้อมูล ข้อมูลที่อ้างอิงถึงสิ่งเดียวกันต้องมีค่าเหมือนกัน เช่น ที่อยู่ รวมทั้งใช้หน่วยของข้อมูลเดียวกัน เช่น สกุลเงินเดียวกัน
๒. Accuracy ความถูกต้อง : ค่าของข้อมูลจัดเก็บตามจริง และรูปแบบข้อมูลถูกต้อง
๓. Completeness ความสมบูรณ์ : ไม่มีข้อมูลสูญหาย มีข้อมูลครบถ้วน
๔. Auditability สามารถตรวจสอบได้ : ต้องมีกลไกช่วยให้ตรวจสอบได้
๕. Timeliness ทันสมัย : ข้อมูลพร้อมใช้เมื่อต้องการ

การจัดการข้อมูลเบื้องต้น

๑. การทำความสะอาดข้อมูล (Data Cleaning) ข้อมูลไม่สะอาดส่งผลกระทบต่อความถูกต้องของการวิเคราะห์ข้อมูล ข้อมูลที่ไม่สะอาด ผู้ใช้จะลังเลที่จะเชื่อผลที่ได้จากการวิเคราะห์ข้อมูลดังกล่าว การทำความสะอาดข้อมูล เป็นขั้นตอนแรกที่มีความสำคัญ โดยแบ่งได้ ดังนี้

- การจัดการค่าข้อมูลสูญหาย ได้แก่ ลบรายการที่มีข้อมูลสูญหาย (ไม่ควรเกิน ๕% ของรายการทั้งหมด) หรือ การเติมค่าที่เหมาะสม เช่น ค่า Null หรือ ค่าเฉลี่ย
- การจัดการค่าสุดโต่ง ค่าของข้อมูลที่สูง/ต่ำ กว่าข้อมูลอื่น เพราะจะมีผลให้ค่าเฉลี่ยผิดเพี้ยน โดยใช้เครื่องมือแผนภาพกระจายข้อมูล หรือ เครื่องมือ BOX Plot ในการวิเคราะห์ แล้วแก้ไขโดยลบข้อมูล หรือ แก้ไขค่าผิดปกติ

๒.การรวบรวมข้อมูล การนำข้อมูลจากหลายแหล่งมารวมกัน เพื่อลดความซ้ำซ้อนของข้อมูลและลดความไม่สอดคล้องของข้อมูลจากหลายแหล่ง สิ่งที่เราควรคำนึงเมื่อรวบรวมข้อมูล ได้แก่

- การระบุตัวแปรเดียว ในการรวมข้อมูล มักมีการนำข้อมูลจากหลายแหล่งที่ตัวแปรอ้างอิงถึงสิ่งเดียวกันมารวมกัน ชื่อตัวแปรต่างกัน แต่แทนค่าสิ่งเดียวกัน เราสามารถใช้ข้อมูลของข้อมูล (Metadata) มาช่วยจับคู่ข้อมูลได้

- การระบุค่าตัวแปรเดียวกัน การระบุค่าของตัวแปรที่ต่างกัน มีความหมายเหมือนกัน
- การตรวจสอบตัวแปรซ้ำซ้อน คือ ตัวแปรที่ไม่จำเป็น สามารถคำนวณค่าได้ผ่านตัวแปรอื่น และตัวแปรที่ปรากฏซ้ำหลายตาราง การระบุตัวแปรที่ไม่จำเป็น อาจทำได้โดยสอบถามผู้เชี่ยวชาญ หรือวิเคราะห์ความสัมพันธ์

- การแก้ไขความไม่สอดคล้องกัน ในการรวบรวมข้อมูลจากหลายแหล่ง อาจเกิดข้อมูลที่มีค่าไม่สอดคล้องกันได้หลายรูปแบบ เช่น การใช้หน่วยของตัวแปรที่ต่างกัน หรือ การใช้ความหมายของตัวแปรที่ต่างกัน

๓. การลดขนาดข้อมูล การลดขนาดของข้อมูลช่วยลดพื้นที่ในการเก็บข้อมูล ยังช่วยให้อัลกอริทึมวิเคราะห์ข้อมูลทำงานได้เร็วขึ้น บริหารจัดการข้อมูลง่ายขึ้นทำได้โดย

- การลดจำนวนตัวแปร เป็นการคัดเลือกเฉพาะตัวแปรที่สำคัญเก็บไว้ในคลังข้อมูล โดยมากมักเป็นตัวแปรที่มีผลต่อการวิเคราะห์ข้อมูล เช่น ตัวแปรเพศ และ ตัวแปรอายุ มีผลต่อการพยากรณ์ยอดขาย โดยใช้เทคนิคการเลือกแบบเดินหน้า หรือ เทคนิคการเลือกแบบถอยหลัง (ผลลัพธ์ที่ได้จาก ๒ เทคนิค อาจไม่เหมือนกัน)

- การแทนค่าที่เล็กลง เป็นการแปลงข้อมูลก่อนจัดเก็บลงคลังข้อมูล แต่สาระของข้อมูลอาจจะหายไป

๔. แปลงข้อมูล การแปลงข้อมูล มีเพื่อช่วยให้กระบวนการวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น ช่วยให้โปรแกรม/อัลกอริทึมทำงานเร็วขึ้น ใช้พื้นที่ความจำลดลง ค้นพบรูปแบบในข้อมูลง่ายขึ้น เทคนิคการแปลงข้อมูล ได้แก่

- เทคนิคการลดขนาดข้อมูล เช่น การแทนที่ด้วยค่าที่เล็กลง
- การทำให้เป็นมาตรฐาน

การสร้างโมเดลเพื่องานจำแนกประเภท

๑. การเรียนรู้ของเครื่อง (Machine Learning) คือ กระบวนการสอนคอมพิวเตอร์โดยใช้ข้อมูลตัวอย่าง เพื่อให้คอมพิวเตอร์เรียนรู้ได้ด้วยตนเอง และสกัดเอาสาระออกมาจากข้อมูลได้ แบ่งออกเป็น ๒ ประเภท ได้แก่

- การเรียนรู้แบบมีผู้สอน คือ การเรียนรู้แบบมีผลลัพธ์ที่ถูกต้อง เพื่อให้เครื่องเรียนรู้รูปแบบความสัมพันธ์ระหว่างวัตถุและผลลัพธ์ที่ถูกต้อง โดยข้อมูลแบ่งออกเป็น ๒ ชุด ได้แก่ ข้อมูลฝึกสอน และข้อมูลทดสอบ

- การเรียนรู้แบบไม่มีผู้สอน คือ การเรียนรู้แบบที่ไม่มีผลลัพธ์ที่ถูกต้อง ใช้ในการจัดกลุ่มตามลักษณะความคล้ายกัน

๒. หลักการจำแนกประเภท เป็นการเรียนรู้แบบมีผู้สอน โดยคอมพิวเตอร์ใช้คุณลักษณะ (Features) ในการจำแนกวัตถุ

๓. ต้นไม้ตัดสินใจ (Decision tree) บางทีเรียกต้นไม้จำแนกประเภท (Classification tree) เป็นโมเดลจำแนกประเภท ฟิเจอร์ที่อยู่ด้านบนของต้นไม้มีความสำคัญ (ช่วยจำแนกข้อมูลได้ดีกว่า) มากกว่าฟิเจอร์ที่อยู่ด้านล่าง โดย ฟิเจอร์สำคัญ คือ ฟิเจอร์ที่เมื่อนำมาใช้จำแนกข้อมูลแล้วได้กลุ่มวัตถุในคลาสเดียวกันหมดหรือเกือบหมด

เกณฑ์ที่ใช้วัดความสำคัญของฟิเจอร์ ได้แก่

- ๑.Information gain เลือกฟิเจอร์ที่นำมาแบ่งข้อมูลแล้วได้กลุ่มที่คล้ายกันมากกว่า (ยิ่งสูงยิ่งดี)
- ๒.Chi-square เลือกฟิเจอร์ที่มีความสัมพันธ์กับตัวแปรเป้าหมายมากกว่า (ยิ่งสูงยิ่งดี)

หลักการสร้างต้นไม้ตัดสินใจ

๑. เลือกฟีเจอร์ที่สำคัญที่สุดบนชุดข้อมูลที่พิจารณาแล้วทำการแบ่งโดยใช้ฟีเจอร์ดังกล่าว
๒. ทำซ้ำบนชุดข้อมูลย่อย
๓. หยุดการแบ่งข้อมูลเมื่อถึงเกณฑ์ที่กำหนด

เกณฑ์วัดประสิทธิภาพ

๑. อัตราความถูกต้อง (Accuracy rate)
๒. อัตราความผิดพลาด (Error rate)

การใช้งานโปรแกรม RapidMiner

RapidMiner เป็นเครื่องมือที่ช่วยในการเตรียมข้อมูล วิเคราะห์ข้อมูล การสร้างโมเดลด้วยเทคนิคการเรียนรู้ของเครื่อง เน้นด้านงานวิชาการข้อมูล หรือ Machine Learning โดยโปรแกรมมีรูปแบบการใช้งานแบบ Drag and Drop คำสั่งจะอยู่ในรูปกล่อง ไม่จำเป็นต้องเขียนคำสั่ง และสามารถทดลองใช้งานได้ฟรี

๑. การนำเข้าข้อมูล : โปรแกรมรองรับการนำเข้าข้อมูลจากภายนอก เช่น ไฟล์ CSV, Excel ใช้ โมดูลในการนำเข้า เช่น “Read CSV” หรือ “Read Excel”
๒. การจัดการค่าที่ขาดหาย ใช้โมดูล “Replace Missing Value” เพื่อแทนที่ค่าที่หายไปด้วยค่าที่เหมาะสม
๓. การกรองข้อมูล ใช้โมดูล “Filter Examples” หรือ “Filter Examples Range” เพื่อกรองข้อมูลตามเงื่อนไข
๔. การจัดการกับข้อมูลซ้ำซ้อน ใช้โมดูล “Remove Duplicate” เพื่อลบแถวที่มีค่าซ้ำกันในชุดข้อมูล
๕. การตรวจสอบข้อมูลสุดท้าย ใช้โมดูล “Data Overview” หรือ “Statistics” เพื่อให้แน่ใจว่าข้อมูลมีความสมบูรณ์พร้อมใช้งาน
๖. การสร้างต้นไม้ตัดสินใจ เป็นวิธีในการจำแนกประเภทข้อมูลและสร้างโมเดลการเรียนรู้ของเครื่อง (Machine Learning Model)
๗. การนำเสนอผลลัพธ์ ใช้โมดูล “Decision Tree Visualization” เพื่อแสดงภาพของต้นไม้ตัดสินใจในรูปแบบกราฟฟิก



ข้อสรุปและแนวทางในการประยุกต์ใช้พัฒนางาน

จากการอบรมหลักสูตร การใช้โปรแกรมดิจิทัลเพื่อการวิเคราะห์ข้อมูล ทำให้ได้เรียนรู้หลักการวิเคราะห์ข้อมูลเบื้องต้น เรียนรู้การเตรียมข้อมูล การจัดการค่าสูญหาย เพื่อให้ได้ข้อมูลที่มีคุณภาพมากยิ่งขึ้น เพราะข้อมูลที่มีคุณภาพและมีความถูกต้อง ช่วยให้เราสามารถวิเคราะห์ข้อมูลเพื่อการตัดสินใจได้อย่างมีประสิทธิภาพ

โดยวิทยากรได้แนะนำการใช้งานโปรแกรม RapidMiner เบื้องต้น ซึ่งทำให้เห็นถึงความสามารถของโปรแกรมที่ช่วยในการจำแนกประเภทของข้อมูล การสร้างโมเดลการเรียนรู้ของเครื่อง (Machine Learning Model) และการสร้างต้นไม้ตัดสินใจ ซึ่งเป็นสามารถนำมาประยุกต์ใช้ในการวิเคราะห์ข้อมูลในอนาคต

แหล่งที่มา

หลักสูตร : การใช้โปรแกรมดิจิทัลเพื่อการวิเคราะห์ข้อมูล

ด้านการพัฒนา (ระบุ ✓) : ☒ ดิจิทัล

☐ ความรู้เฉพาะด้านเพื่อใช้ในการปฏิบัติงาน

☐ เพิ่มศักยภาพตนเอง/ประสิทธิภาพในการทำงาน

บรรยายโดย : ผศ.ดร.สิรินดา พลหาญ อาจารย์ประจำคณะวิทยาศาสตร์และเทคโนโลยี ม.หอการค้าไทย

หน่วยงานผู้รับผิดชอบ สำนักงาน ก.พ.

วิธีการพัฒนาดตนเอง : e-learning (OCSC Learning Portal - สำนักงาน ก.พ.)

วันที่ได้รับการฝึกอบรม : 18 ก.พ. 2568

สถานที่ : กรมพัฒนาที่ดิน