สรุปบทเรียนจากการพัฒนาความรู้

หลักสูตร
Data Analytics

โดย พนัสบดี ธัชโอภาส

เนื่องจากเทคโนโลยีทำให้เกิดข้อมูลมหาศาล ในการนำข้อมูลเหล่านี้มาใช้อย่างมีประสิทธิภาพ Data Analytics เป็นเครื่องมือในการวิเคราะห์เพื่อให้เห็นคุณค่าที่แท้จริงของข้อมูล เพื่อนำไปใช้ให้เกิดประโยชน์ได้สูงสุด

Data Analytics เป็น Business Intelligence ซึ่งเป็นศาสตร์ของการใช้ข้อมูลต่าง ๆ จำนวนมาก (Big Data) มาวิเคราะห์ร่วมกัน และแสดงผลเพื่อสนับสนุนการทำงานต่าง ๆ ตามวัตถุประสงค์ที่ต้องการ มักเกี่ยวข้องกับการศึกษาข้อมูลในอดีต เพื่อหาแนวโน้มที่อาจเป็นไปได้ในการวิจัย เพื่อวิเคราะห์ผลกระทบของการตัดสินใจ เป็นการสร้าง model ทางคณิตศาสตร์จากข้อมูลที่มีอยู่เพื่อช่วยประกอบการตัดสินใจ

ประเภทของการวิเคราะห์

๑. Prescriptive Analytics - ศึกษาข้อมูลและวิเคราะห์ผลลัพธ์ที่อาจจะเป็นไปได้ ให้คำแนะนำว่าควรทำอะไร

๒. Predictive Analytics - ศึกษาข้อมูลเดิมเพื่อพยากรณ์อนาคตโดยดูจาก pattern ของข้อมูลในอดีต

๓. Descriptive Analytics - ให้ข้อมูลอธิบายตัวเองว่ามันคืออะไร มีความหมายอย่างไร เพื่อใช้วิเคราะห์ ประกอบการตัดสินใจในขั้นตอนต่อไป

ระดับของ Data Analytics

๑. Query Processing - เรียกดูข้อมูลที่ต้องการ

๒. Summary Statistics - ประมวลผลได้เป็นข้อมูลทางสถิติ (ผมรวม ค่าเฉลี่ย ค่าสูงสุด ต่ำสุด)

๓. Exploration - ดูข้อมูลเชิงลึก นำเสนอเป็นภาพเพื่อให้เข้าใจง่ายขึ้น

๔. Modeling - ให้เครื่องเรียนรู้ข้อมูลและสร้างเป็น Model เพื่อให้เห็นรูปแบบสำหรับการทำนายอนาคต แบ่งเป็น Supervised Machine Learning (มีข้อมูลและเฉลยให้เครื่องเรียนรู้) Unsupervised Machine Learning (ให้เครื่องเรียนรู้จากข้อมูลเพียงอย่างเดียว)

KDD (Knowledge Discovery in Databases) Process

๑. Selection - เลือกข้อมูลที่ต้องการ

๒. Preprocessing - ทำความสะอาดข้อมูล ตัดข้อมูลที่ไม่จำเป็น ข้อมูลที่สุดโต่ง

๓. Transformation - แปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะแก่การประมวลผล เช่น แทนค่าด้วยรหัส/ตัวเลข จัดหมวดหมู่เข้าด้วยกัน

๔. Data Mining - สกัดเพื่อให้ได้ Patterns ของข้อมูล

๕. Interpretation/Evaluation - ตีความ/ประเมินความหมาย/ความสำคัญของข้อมูล

ทีมงานที่ร่วมวิเคราะห์ข้อมูลมีอย่างน้อย ๓ กลุ่ม

๑. Business Analyst/Expert - ผู้เชี่ยวชาญในเรื่องที่ต้องการวิเคราะห์ เช่น หมอ วิศวกร นักลงทุน นักกฎหมาย ฯลฯ

๒. Machine-Learning Expert/Data Scientist/Data Analyst - ผู้มีความรู้ทางสถิติและการสร้าง Data Model ที่จะออกแบบกระบวนการวิเคราะห์ข้อมูล

๓. Data Engineer - ผู้เชี่ยวชาญในการรวบรวม จัดเก็บ และประมวลผลข้อมูล

Data Understanding and Processing คือกระบวนการทำความสะอาดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์

Unstructured Data - ข้อมูลแบบไม่มีโครงสร้าง เช่น ภาพถ่ายดาวเทียม, ข้อมูลจากการสังเกตทางวิทยาศาสตร์, ภาพถ่ายและวีดิโอ, ข้อมูลจากเรดาร์หรือโซนาร์, ข้อความในเอกสาร, ข้อมูลจากโซเชียลมีเดีย, ข้อมูลในโทรศัพท์, เนื้อหาเว็บไซต์ ฯลฯ เป็นข้อมูลแบบที่จัดการยาก จำเป็นต้องมีการแปลงค่าก่อนจึงจะสามารถนำไปใช้งานได้

Semi-Structured Data - ข้อมูลแบบกึ่งมีโครงสร้าง มีคำอธิบายความหมายรวมอยู่กับข้อมูล เช่น ไฟล์แบบ JSON, XML สามารถนำไปใช้งานได้ง่ายกว่าแบบแรก

ข้อมูลแบ่งออกได้เป็น ๒ ประเภท

๑. Categorical เป็นข้อมูลที่ไม่ใช่ตัวเลข ต้องนำไปแปลงก่อนจึงจะใช้ได้ มี ๓ แบบ

๑.๑ Ordinal เป็นข้อมูลที่สามารถเรียงลำดับได้ บอกได้ว่าตัวไหนมากกว่าน้อยกว่า เช่น > ๕ (มากกว่า ๕), ๐< . . . <๑๐ (มากกว่า ๐ แต่น้อยกว่า ๑๐)

๑.๒ Binary เป็นข้อมูลที่เป็นไปได้เพียง ๒ ทาง มี/ไม่มี ชาย/หญิง

๑.๓ Nominal เป็นข้อมูลแบบข้อความทั่วไป

๒. Continuous เป็นข้อมูลตัวเลขที่สามารถนับต่อเนื่องได้ เช่น ๑๐, ๑๓๒, ๕๔๕๑๒

การจัดเตรียมข้อมูล

Continuous -> Ordinal เป็นการแปลงข้อมูลตัวเลขให้เป็นกลุ่ม เช่น เปลี่ยนข้อมูล "อายุ" ให้กลายเป็น "ช่วงอายุ" มีวิธีแปลง ๒ แบบคือ

๑. Equal Interval คือกำหนดให้แต่ละช่วงมีระยะห่างเท่ากัน เช่น ๑ - ๑๐, ๑๑ - ๒๐, ๒๑ - ๓๐, ...

๒. Equal Frequency ซึ่งแบ่งให้แต่ละช่วงมีจำนวนข้อมูลเท่ากัน โดยที่ไม่คำนึงถึงระยะห่างของข้อมูลในแต่ละช่วง เช่น ๑ - ๒๐, ๒๑ - ๓๐, ๓๑ - ๔๕, ๔๖ - ๘๐, ...

Nominal -> Binary ใช้เทคนิค One Hot Encoder โดยแตกข้อมูลแต่ละแบบออกเป็น Field ต่าง ๆ แล้วกำหนดให้เป็นตัวเลือกว่า มี/ไม่มี (๑/๐ - Binary) เช่น ข้อมูลดิบ Field หนึ่งคือ พืชที่ปลูก ซึ่งมีข้อมูลเป็น ข้าว, อ้อย, ถั่ว, ... ให้แตก Field เป็น ข้าว, อ้อย, ถั่ว, ... โดยแต่ละ Field เก็บค่า มี/ไม่มี (๑/๐ - Binary)

Missing Values คือการที่ไม่มีข้อมูล แบ่งเป็น ๓ แบบ

๑. ไม่มีข้อมูลจริง ๆ เช่น Field วันที่เสียชีวิต ในกรณีที่เจ้าของข้อมูลใน Record นั้นยังไม่ตาย

๒. ไม่เปิดเผย คือเจ้าของข้อมูลเจตนาไม่บอกให้ทราบ

๓. บันทึกข้อมูลผิด

สามารถเลือกวิธีจัดการได้ ๓ วิธี

๑. แทนที่

๑.๑ แทนที่ด้วยข้อมูลทางสถิติของข้อมูลเท่าที่มี เช่น แทนที่ด้วยค่าเฉลี่ยหรือฐานนิยม

๑.๒ แทนที่ด้วยข้อมูลของกลุ่มตัวอย่างคล้ายกัน เช่น ถ้า Record อื่นที่มีอายุ เพศ การศึกษาเหมือนกันตอบอะไร ก็เอาข้อมูลนั้นมาแทน

๑.๓ ใช้โมเดลทางคณิตศาสตร์คำนวณว่าข้อมูลที่หายไปคืออะไร

๒. ลบทิ้ง - ใช้ในกรณีที่ Field นั้นไม่มีนัยสำคัญ หรือมีข้อมูลที่หายไปมากเกินไป

๓. เก็บไว้อย่างนั้น คือมองว่าการไม่มีข้อมูลเป็นข้อมูลอย่างหนึ่งที่มีความหมายในการวิเคราะห์

Outlier - ข้อมูลสุดโต่ง เป็นข้อมูลที่มากหรือน้อยกว่าข้อมูลอื่น ๆ มาก แบ่งได้เป็น ๒ แบบ

๑. Valid Observations - มันสุดโต่ง แต่เป็นแบบนั้นจริง ๆ เช่น เงินเดือนผู้บริหารที่สูงกว่าเงินเดือนของพนักงานทั่วไปอย่างมาก จัดการได้โดย

๑.๑ Truncation ใช้หลักการคำนวณทางคณิตศาสตร์เพื่อทอนข้อมูลสุดโต่งลง เช่น กำหนดให้ข้อมูลสุดโต่งมีค่าเท่ากับ ๑.๕ เท่าของค่าเฉลี่ย วิธีนี้ข้อมูลที่เป็น Outliner จะมีค่าเท่ากันหมด

๑.๒ Capping เป็นการแปลงค่าของ Outliner ให้ไม่เกินขอบเขตที่กำหนด วิธีนี้ข้อมูลที่เป็น Outliner จะมีค่าไม่เท่ากันขึ้นกับความมากน้อยของข้อมูล

๒. Invalid Observations - เป็นข้อมูลสุดโต่งที่รู้ว่าเป็นข้อมูลผิดพลาด เช่น อายุ ๒๐๐ ปี จัดการเสมือนว่าเป็น Missing Values

Standardization ใช้เมื่อข้อมูลมีหน่วยไม่เท่ากันหรือมีค่าแตกต่างกันมาก จนทำให้บางข้อมูลมีอิทธิพลต่อการคำนวณมากกว่าข้อมูลอื่น (เช่น ตัวเลขของเงินเดือนที่สูงกว่าตัวเลขอายุ - ความต่าง ๒๐ แทบไม่มีผลในกรณีของเงินเดือน แต่เป็นความแตกต่างที่ชัดเจนเมื่อเป็นอายุ) จึงต้องหาวิธีทำให้แต่ละข้อมูลมีค่าอยู่ในสัดส่วนเดียวกันที่สามารถเปรียบเทียบกันได้

มีวิธีการ Standardization เช่น

๑. วิธี Max/Min ทอนสัดส่วนให้แต่ละข้อมูลมีค่าต่ำสุด/สูงสุดเท่ากัน

๒. ใช้ Z-Score (เป็นวิธีที่นิยมที่สุด) แปลงค่าของข้อมูลแต่ละตัวให้มีค่าเฉลี่ยเท่ากับ ๐ และมีส่วนเบี่ยงเบนมาตรฐานเท่ากัน

๓. Decimal Scaling นำข้อมูลหารด้วย ๑๐ ยกกำลังจำนวนหลักของข้อมูล เช่น เงินเดือน ๑๕,๐๐๐ (ตัวเลข ๕ หลัก) แปลงโดยให้มีค่าเท่ากับ ๑๕,๐๐๐/๑๐^๕

Data Exploration and Visualization

ข้อมูลแต่ละชุดอาจจะมีรายละเอียดจำนวนมาก การนำเสนอข้อมูลแต่ละแบบไม่สามารถนำเสนอข้อมูลทุกอย่างได้ครบถ้วนในคราวเดียว แต่ละแบบมีจุดเด่นจุดด้อยที่แตกต่างกัน และทำให้เห็นข้อมูลในแง่มุมที่ไม่เหมือนกัน นักวิเคราะห์จำเป็นต้องเลือกวิธีการนำเสนอที่เหมาะสมเพื่อให้สามารถทำความเข้าใจข้อมูลได้อย่างถูกต้องเพื่อเลือกใช้งานให้ได้ถูกต้องตามวัตถุประสงค์

Machine Learning

คือการทำให้คอมพิวเตอร์เรียนรู้จากข้อมูล โดยแบ่งได้เป็น ๒ แบบ ได้แก่

๑. Supervised Machine Learning - ข้อมูลที่ส่งให้เรียนรู้มี Label ระบุไว้ว่าเป็นข้อมูลอะไร คอมพิวเตอร์จะอ่านข้อมูลเทียบกับ Label เพื่อสร้างเป็น Model (สมการทางคณิตศาสตร์) ที่เป็นเกณฑ์การประเมินว่ามีองค์ประกอบใดบ้างที่ทำให้ข้อมูลจัดอยู่ในประเภทตามที่ระบุไว้ที่ Label

๒. Unsupervised Machine Learning - ข้อมูลที่ส่งให้เรียนรู้ไม่มี Label ระบุไว้ว่าเป็นข้อมูลอะไร คอมพิวเตอร์จะอ่านข้อมูลและสร้างเป็น Model (สมการ) ที่ใช้จัดประเภทข้อมูลที่มีคุณสมบัติใกล้เคียงเข้าด้วยกัน เป็นหน้าที่ของผู้ใช้งานที่จะประเมินว่าแต่ละประเภทหมายถึงอะไร

หลังจากที่เรียนรู้จนสามารถสร้างเป็น Model ได้แล้ว เมื่อได้รับข้อมูลชุดใหม่เข้ามา คอมพิวเตอร์จะใช้ Model ที่สร้างขึ้นประเมินว่าข้อมูลนั้นอยู่ในประเภทใด

คอมพิวเตอร์สามารถสร้าง Model ได้หลายตัวจากการเรียนรู้ เราจำเป็นจะต้องประเมินว่า Model ตัวใดที่ทำงานได้ดีที่สุด โดยทำได้ด้วยการแบ่งข้อมูลออกเป็น ๒ ชุด แบ่งชุดหนึ่งไว้ให้เรียนรู้และสร้างเป็น Model ต่าง ๆ และอีกชุดหนึ่งใช้สำหรับทดสอบ โดย Model ตัวใดที่ใช้กับข้อมูลทดสอบแล้วได้ผลที่ถูกต้องที่สุดก็จะถูกเลือกนำไปใช้

แต่บางครั้งหากการเรียนรู้จากกลุ่มตัวอย่างได้ดีเกินไป จะทำให้เกิดสิ่งที่เรียกว่า Overfitting ขึ้นได้ นั่นคือการที่สร้าง Model ที่สามารถจัดประเภทของข้อมูลของชุดฝึกได้ถูกต้องเกินไป ทั้งที่ข้อมูลนั้นเป็น Noise ทำให้ได้ Model ที่ไม่ถูกต้องอย่างที่ควรเมื่อใช้กับข้อมูลจริง

การเลี่ยง Overfitting ทำได้โดยการกำหนดจุดที่จะเลิกฝึก โดยใช้วิธีแบ่งข้อมูลเป็นชุด Validation หรือการใช้สูตรคำนวณเพื่อหาจุดเลิกฝึก

Descriptive Analytics

คือการประมวลผลเพื่อให้ข้อมูลอธิบายตัวมันเองว่ามีความหมายว่าอย่างไร แบ่งได้เป็นหลายระดับ ได้แก่

- การทำความเข้าใจข้อมูล เช่น ประเภทของข้อมูล การวิเคราะห์เชิงสถิติ การนำเสนอข้อมูลด้วยภาพ

- การเตรียมข้อมูล เช่น การจัดการ Missing Values, Outlier

- การทำ Data Modelling เช่น วิเคราะห์การกระจายตัวของข้อมูล วิเคราะห์ความสัมพันธ์ของข้อมูล การจัดกลุ่มข้อมูล

การเตรียมข้อมูลและการทำ Data Modelling จะต้องมีกระบวนการวัดความเหมือนความต่างของข้อมูล (Similarity Measurement) การจัดกลุ่มข้อมูล (Segmentation) การตรวจจับสิ่งผิดปกติ (Anomaly Detection Strategy)

Similarity Measurement ทำให้รู้ว่าข้อมูลแต่ละตัวมีความเหมือนหรือแตกต่างจากข้อมูลอื่น ๆ เท่าใด

Segmentation กำหนดเกณฑ์ในการเลือกว่าข้อมูลใดบ้างที่มีความเหมือนมากพอที่จะถือว่าอยู่ในกลุ่มเดียวกัน และควรแบ่งเป็นกี่กลุ่ม โดยจะมีจุดศูนย์กลางที่ถือเป็นตัวแทนที่จะอธิบายลักษณะของแต่ละกลุ่ม

Anomaly Detection Strategy คือวิธีการจัดการกับข้อมูลที่ผิดปกติ

กระบวนการข้างต้นจะช่วยตอบคำถาม ๔ ข้อดังนี้

๑. จะแบ่งกลุ่มด้วยวิธีใด (เลือกวิธีการทางสถิติที่เหมาะสม)

๒. ควรแบ่งเป็นกี่กลุ่ม (ตัดสินจากผลลัพธ์ที่ได้)

๓. จะเข้าใจคุณสมบัติของแต่ละกลุ่มได้อย่างไร (ดูจากจุดศูนย์กลางที่ถือเป็นตัวแทนของกลุ่ม)

๔. เมื่อมีข้อมูลใหม่เข้ามา ควรจะจัดเข้าไปอยู่ในกลุ่มใด (จัดเข้าอยู่ในกลุ่มที่มีระยะห่างจากจุดศูนย์กลางของกลุ่มน้อยที่สุด)

Predictive Analytics

ใช้กระบวนการที่เรียกว่า Regression ซึ่งเป็นการทำนายข้อมูลที่มีลักษณะเป็นข้อมูลต่อเนื่อง (Continuous Value) เช่น หาราคาบ้านจากพื้นที่ หาร้อยละของไขมันในร่างกายจากค่าดัชนีมวลกาย หาราคาหุ้นตัวหนึ่งจากราคาหุ้นตัวอื่น ๆ โดยเราจะต้องเอาข้อมูลที่มีอยู่แล้วในอดีตมาให้คอมพิวเตอร์เรียนรู้และสร้างออกมาเป็น Model ซึ่งส่วนมากจะเป็นสมการทางวิทยาศาสตร์ โดยหลักการคือใส่ค่าถ่วงน้ำหนักให้ตัวแปรแต่ละตัวเพื่อคำนวณออกมาเป็นค่าที่ต้องการ โดยให้คอมพิวเตอร์เรียนรู้จากข้อมูลเก่าเพื่อให้ได้ค่าถ่วงน้ำหนักที่เหมาะสมที่จะให้ค่า Error ต่ำที่สุด

Predictive Analytics Classification

Predictive Analytics Classification คือการทำนายข้อมูลที่ผลลัพธ์ไม่ได้ออกมาเป็นตัวเลขที่ต่อเนื่องแต่จะเป็นค่าที่จำกัด เช่น ถูก/ผิด สูง/กลาง/ต่ำ เมื่อนำข้อมูลเก่ามาฝึกจนได้เป็น Model ก็จะมีวิธีในการประเมินความถูกต้องของ Model ได้หลายอย่าง

ตัวอย่างเครื่องมือสำหรับประเมินความถูกต้องของ Model

The Confusion Matrix

ใช้กับกรณีที่ผลการทำนายออกมาได้เป็น ๒ กรณี เช่น ถูก/ผิด สูง/ต่ำ ดี/เสีย โดยนำโมเดลมาใช้กับข้อมูลทดสอบที่รู้ผลลัพธ์แล้วและแบ่งผลที่ได้ออกมาเป็น ๔ กลุ่ม ได้แก่

๑. True Positive (TP) - ทำนายถูกว่าผลลัพธ์เป็น True

๒. True Negative (TN) - ทำนายถูกว่าผลลัพธ์เป็น False

๓. False Positive (FP) - ทำนายผิด ทำนายว่าผลลัพธ์เป็น True แต่ความจริงเป็น False

๔. False Negative (FN) - ทำนายผิด ทำนายว่าผลลัพธ์เป็น False แต่ความจริงเป็น True

ค่าความถูกต้องคือกรณีที่เป็น True Positive (TP) บวกกับ True Negative (TN) หารจำนวนทั้งหมด

แต่ในกรณีที่ข้อมูล ๒ กลุ่มมีปริมาณต่างกันมาก ๆ ค่าความถูกต้องที่ได้อาจจะไม่สะท้อนความน่าเชื่อถือของ Model เช่น ถ้าข้อมูลจริงเป็น T = ๙๐% และ F = ๑๐% หาก Model ทำนายว่าทุกตัวคือ T ก็เท่ากับมีค่าความถูกต้อง ๙๐% แล้ว

จึงมีการใช้ค่าอีก ๒ ตัวมาใช้ประกอบการประเมิน ได้แก่

๑. Recall คือความแม่นยำในการหา T - เมื่อข้อมูลจริงเป็น T จะทำนายได้ถูกต้องเท่าไร - Recall = TP/(TP+FN)

๒. Precision คือความแม่นยำของการทำนาย เมื่อทำนายว่าเป็น T จะทำนายได้ตรงกับความเป็นจริงเท่าไร - Precision = TP/(TP+FP)

เราสามารถนำค่าทั้ง Recall และ Precision มาคำนวณร่วมกันเพื่อได้ค่าค่าเดียวสำหรับประเมินความถูกต้องได้ เรียกว่า F-Measure

F = (๒ * Precision * Recall)/(Recall + Precision)

F = (๒ * TP)/(๒ * TP + FP + FN)

<u>Predictive Analytics Classification Model</u>

ตัวอย่างของ Predictive Analytics Classification Model

๑. Decision Trees เป็นการสร้างเงื่อนไขการตัดสินใจเป็นชั้น ๆ จนกระทั่งสามารถทำนายได้ว่าตัวอย่างนั้นอยู่ในกลุ่มใด

๒. Logistic Classification ใช้สมการทางคณิตศาสตร์ในการแบ่งกลุ่ม ได้ผลลัพธ์ตั้งแต่ ๐ - ๑

๓. Neural Network เลียนแบบเครือข่ายประสาทของมนุษย์ หลักการคือการรับค่า Input จากนั้นก็คูณด้วยค่าถ่วงน้ำหนัก แล้วนำเข้าสมการบางอย่าง จนได้เป็นผลของการทำนาย ทั้งนี้อาจจะเอาผลที่ได้มาใช้เป็น Input สำหรับ Neural Network อื่นเพื่อใช้ทำนายสิ่งอื่น ๆ ต่อไปได้อีก

Recommendation System – ระบบให้คำแนะนำ

เป็นการประเมินว่าเมื่อเกิดเหตุการณ์หนึ่งขึ้น จะมีโอกาสเกิดอีกเหตุการณ์หนึ่งมากน้อยแค่ไหน ด้วยความเชื่อมั่นเท่าไร เช่น ระบบให้คำแนะนำสินค้า เมื่อลูกค้าเลือกของชิ้นหนึ่ง ระบบจะประเมินว่ามีสินค้าใดบ้างที่มักจะซื้อพร้อมกับสินค้านั้นและให้คำแนะนำมา

- Association Rules คือการกำหนดกฎหรือเงื่อนไข เช่น เมื่อลูกค้าซื้อสปาเก็ตตี้ จะมีโอกาส ๗๐% จะซื้อไวน์แดงด้วย

มีขั้นตอนในการกำหนดกฎดังนี้

๑. Frequent Itemset Generation เป็นการสร้าง Itemset ที่มีความถี่ตามที่ต้องการ โดยมีระดับการสนับสนุนสูงกว่าค่า minsup (เรียกว่า Frequent Itemsets)

๒. Rule Generation เป็นการสร้างกฎ โดยมีค่าความเชื่อมั่นสูงกว่าค่า minconf

ทั้งนี้จะต้องมีการกำหนดค่า minsup และ minconf ไว้ตั้งแต่ต้น

- Collaborative Filtering ใช้ "Wisdom of the Crowd" ในการให้คำแนะนำ เช่น ลูกค้าที่ในอดีตมีรสนิยมคล้ายกัน ในอนาคตก็จะมีรสนิยมคล้ายกันด้วย

ใช้ในกรณี เช่น มีข้อมูลการประเมินความพึงพอใจสินค้าต่าง ๆ ของกลุ่มตัวอย่าง ระบบจะทำนายว่ากลุ่มตัวอย่างจะมีความพึงพอใจสินค้าที่ไม่ประเมินในระดับใด โดยอ้างอิงจากกลุ่มตัวอย่างอื่น ๆ ที่มีรสนิยมใกล้เคียงกัน

แหล่งที่มา

สรุปบทเรียนจากการเรียนรู้ผ่านสื่ออิเล็กทรอนิกส์ (e-Learning)
สำนักงานคณะกรรมการข้าราชการพลเรือน (สำนักงาน ก.พ.)
หลักสูตร : Data Analytics
บรรยายโดย : ดร.ไพรสันต์ ผดุงเวียง