บทที่ 2
แนวคิด ทฤษฎีเอกสารที่เกี่ยวข้อง

ในบทนี้เป็นการนำเสนอเกี่ยวกับ แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้องของการวิเคราะห์ปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน เพื่อทำนายข้อมูลการจำแนกประเภทด้วยโมเดล Classification และ เพื่อทำนายโมเดลสำหรับการพยากรณ์ด้วยโมเดล Multiple Linear Regression ซึ่งได้รวบรวมการศึกษาเอกสารงานวิจัยที่เกี่ยวข้อง กับการวิเคราะห์ข้อมูล เพื่อใช้เป็นแนวทางการศึกษาประกอบด้วยรายละเอียดตามลำดับ ดังนี้

2.1 แนวคิดเกี่ยวกับการวิเคราะห์ปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน
- 2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data analytic)
- 2.1.2 แนวคิดการทำความสะอาดข้อมูล (Data Cleansing)
- 2.1.3 แนวคิดการพยากรณ์ข้อมูล (Forecasting data)
- 2.1.4 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data visualization)

2.2 ทฤษฎีที่เกี่ยวข้อง
- 2.2.1 ทฤษฎีเหมืองข้อมูล (Data Mining)
- 2.2.2 ทฤษฎีการจำแนกประเภทของข้อมูล (Classification)
- 2.2.3 ทฤษฎีการถดถอยพหุคูณ (Multiple Regression)
- 2.2.4 หลักในการทำเหมืองข้อมูล
- 2.2.5 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์
- 2.2.6 ทฤษฎีเกี่ยวกับปัจจัยหลักที่มีผลต่อความล่าช้าของสายการบิน
- 2.2.7 ทฤษฎีการสุ่มตัวอย่าง (Sampling)

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล
- 2.3.1 กระบวนการวิเคราะห์ข้อมูลด้วย (CRISP–DM)
- 2.3.2 โปรแกรม RapidMiner Studio หรือ Altair® AI Studio™
- 2.3.3 โปรแกรม Tableau Public
- 2.3.4 เทคนิคทางสถิติแมทริกซ์สหสัมพันธ์ (Correlation Matrix)
- 2.3.5 แบบจำลองการจำแนกประเภทของข้อมูล (Classification)
- 2.3.6 เทคนิคการวิเคราะห์ข้อมูลการถดถอยพหุคูณ (Multiple Linear Regression)

2.4 วรรณกรรมที่เกี่ยวข้อง

2.5 บทสรุป

2.1 แนวคิดเกี่ยวกับการวิเคราะห์ปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน

2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data analytic)

ในการดำเนินงานเรื่องการเปรียบเทียบประสิทธิภาพโมเดลที่เหมาะสมกับการการวิเคราะห์ปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน ทางผู้วิเคราะห์ข้อมูลที่ศึกษาหลักการ และทฤษฎีต่าง ๆ องค์ประกอบหนึ่งที่สำคัญคือการวิเคราะห์ข้อมูล ซึ่งมีรายละเอียดดังนี้

2.1.1.1 การวิเคราะห์ข้อมูล (Data Analytics) คือกระบวนการนำข้อมูลมาจัดระเบียบ ตรวจสอบ แปลง และสรุปผลเพื่อนำไปใช้ประโยชน์ในด้านต่าง ๆ เช่น การตัดสินใจทางธุรกิจ การพยากรณ์แนวโน้ม หรือการค้นหาโอกาสใหม่ ๆ ในการพัฒนาองค์กร กระบวนการนี้อาจใช้สถิติ คณิตศาสตร์ เทคโนโลยีคอมพิวเตอร์ และปัญญาประดิษฐ์ (AI) ในการประมวลผลข้อมูล ซึ่งช่วยให้สามารถมองเห็นภาพรวม ค้นหาความสัมพันธ์ หรือคาดการณ์สิ่งที่จะเกิดขึ้นได้อย่างมีประสิทธิภาพมากขึ้น ประเภทของการวิเคราะห์ข้อมูล สามารถแบ่งได้ดังนี้

ก. การวิเคราะห์แบบพรรณนา (Descriptive Analytics) คือการวิเคราะห์ข้อมูลแบบพื้นฐาน เพื่อแสดงผลที่เกิดขึ้น หรือกำลังจะเกิดขึ้น จากข้อมูลในอดีต ในลักษณะที่เข้าใจง่ายสามารถสร้างขึ้นได้ด้วยตนเอง เช่น รายงาน แผนภูมิ กราฟ ตาราง เป็นต้น ซึ่งจะช่วยให้เข้าใจการเปลี่ยนแปลงที่เกิดขึ้นกับองค์กรได้ดียิ่งขึ้น เช่น การเปลี่ยนแปลงของราคาสินค้ารายปี การเติบโตของยอดขายรายเดือน การเปรียบเทียบยอดขายในแต่ละสาขาหรือแต่ละช่องทาง การเปรียบเทียบจำนวนผู้ใช้งานเว็บไซต์ในแต่ละช่วงเวลา เป็นต้น ซึ่ง Descriptive Analytics คือการอธิบายสิ่งที่เกิดขึ้นตามวัตถุประสงค์และช่วงเวลาที่กำหนด

ข. การวิเคราะห์แบบวินิจฉัย (Diagnostic Analytics) คือการวิเคราะห์เชิงวินิจฉัย ซึ่งเป็นรูปแบบหนึ่งของการวิเคราะห์ขั้นสูงแบบเจาะลึก โดยการวิเคราะห์ข้อมูลที่ได้จากการทำ Descriptive Analytics เพื่อหาคำตอบว่าทำไมจึงเกิดสิ่งนั้น ๆ หรืออธิบายปัจจัยและตัวแปรที่เป็นสาเหตุของการเกิดสิ่งนั้น ๆ ขึ้น ซึ่งจะต้องอาศัยเทคนิคต่าง ๆ เข้ามาช่วย เช่น การทำ Data discovery หรือ Data mining เป็นต้น ตัวอย่างการวิเคราะห์ข้อมูลเชิงวินิจฉัย เช่น อธิบายสาเหตุที่ทำให้ยอดขายเพิ่มขึ้น จำนวน ผู้ใช้ที่เข้าชมเว็บไซต์เพิ่มขึ้น จำนวนลูกค้าที่เข้าใช้บริการที่หน้าร้านลดลง โปรโมชันที่ไม่ค่อยได้รับความนิยม เนื้อหาโฆษณาที่ได้ CTR% มากกว่าเนื้อหาอื่น ๆ หรือวิเคราะห์การทำงานของเครื่องคอมพิวเตอร์เพื่อตรวจจับความผิดปกติ เป็นต้น ข้อมูลเหล่านี้จะทำให้องค์กรรู้ถึงความต้องการของตลาด เข้าใจพฤติกรรมของ

ลูกค้า รู้สาเหตุของปัญหาด้านเทคโนโลยี รวมถึงสามารถปรับปรุงวัฒนธรรมองค์กร เพื่อการทำงานที่ดีขึ้น

ค. การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) คือการวิเคราะห์ข้อมูลทั้งข้อมูลในอดีตและปัจจุบันออกมาในเชิงคาดการณ์ ทำนาย หรือการพยากรณ์ เพื่อหาแนวโน้มที่จะเกิดสิ่งต่าง ๆ ขึ้นตามวัตถุประสงค์ที่กำหนด โดยการสร้างแบบจำลองทางสถิติ บวกกับการนำเทคโนโลยีปัญญาประดิษฐ์มาใช้ ซึ่งสามารถสร้างประโยชน์ได้มากมายในหลายแง่มุม เช่น การคาดการณ์ความเสี่ยงและโอกาสยอดขาย ภัยไซเบอร์ สภาพอากาศ การลงทุน หุ้น หรือผลการเลือกตั้ง เป็นต้น อย่างไรก็ตามการทำ Predictive Analytics ที่ถูกต้องและแม่นยำนั้น ขึ้นอยู่กับคุณภาพของข้อมูล ซึ่งเป็นสิ่งที่องค์กรควรให้ความสำคัญเป็นอันดับแรก โดยการเตรียมข้อมูลให้มีคุณภาพที่ดีและเหมาะสม ก่อนนำไปใช้วิเคราะห์เพื่อให้เกิดผลลัพธ์ที่มีประสิทธิภาพ ลดข้อผิดพลาด

ง. การวิเคราะห์แบบแนะนำ (Prescriptive Analytics) คือการวิเคราะห์แบบให้คำแนะนำ ซึ่งเป็นการวิเคราะห์ที่มีความซับซ้อนมากที่สุด ต่อเนื่องจากการทำ Predictive Analytics กล่าวคือ เมื่อได้ข้อมูลแนวโน้มที่จะเกิดบางสิ่งขึ้นแล้ว การทำ Prescriptive Analytics จะช่วยแนะนำแนวทางการดำเนินการในขั้นตอนต่อไปที่เหมาะสมที่สุด และวิเคราะห์ไปถึงผลที่จะเกิดขึ้นถ้าหากเลือกปฏิบัติตามแนวทางนั้น ๆ หรือแม้แต่แนะนำแนวทางในการรับมือและแก้ไขปัญหา การวิเคราะห์แบบให้คำแนะนำจึงถือเป็นเครื่องมือที่สำคัญอย่างมากสำหรับการตัดสินใจที่ขับเคลื่อนด้วยข้อมูล การวิเคราะห์แบบให้คำแนะนำคือทำงานร่วมกันระหว่าง Big data อัลกอริทึมของ Machine learning และเทคโนโลยี AI เพื่อช่วยในการวิเคราะห์ข้อมูลจำนวนมหาศาลที่มีความซับซ้อนเกินกว่าที่มนุษย์จะทำได้ ซึ่งการทำ Prescriptive Analytics ยังช่วยเพิ่มประสิทธิภาพในการตัดสินใจด้านต่าง ๆ (BLENDATA, 2025)

2.1.2 แนวคิดการทำความสะอาดข้อมูล (Data Cleansing)

การทำความสะอาดข้อมูล (Data Cleansing) กระบวนการตรวจสอบ สะสาง แก้ไข หรือจัดรูปแบบข้อมูลให้อยู่ในสภาพที่พร้อมใช้งานที่สุด รวมไปถึงคัดกรองข้อมูลที่ไม่ถูกต้องหรือไม่จำเป็นออกไปจากชุดข้อมูลที่จะใช้วิเคราะห์หรือประมวลผล เพื่อให้ชุดข้อมูลที่จะใช้มีความสมบูรณ์ มีคุณภาพ พร้อมนำไปวิเคราะห์และใช้ประโยชน์ ซึ่งอาจเรียกอีกอย่างว่าเป็นการทำ “re-organize” ข้อมูลใหม่ก็ได้ ส่วนสาเหตุของการที่เราต้องทำความสะอาดข้อมูลก่อนใช้ นั่นก็เพื่อหลีกเลี่ยงผลลัพธ์ที่ไม่ถูกต้องแม่นยำ จากปัจจัยหลายอย่าง ไม่ว่าจะเป็นความ

ผิดพลาดของการบันทึกข้อมูล การพิมพ์ผิด รูปแบบข้อมูลที่แตกต่างกัน ชุดข้อมูลไม่สอดคล้องกับคำถาม ข้อมูลที่ไม่เป็นความจริง ข้อมูลที่ไม่สามารถอ้างอิงในระบบได้ เป็นต้น (บริษัท ริคโค จำกัด, 2025)

การทำความสะอาดข้อมูลเป็นขั้นตอนสำคัญก่อนนำข้อมูลไปใช้ เนื่องจากข้อมูลที่ไม่มีคุณภาพอาจส่งผลกระทบต่อผลลัพธ์ที่ได้ ดังนี้

- ช่วยให้ข้อมูลมีความแม่นยำ ลดข้อผิดพลาดที่อาจเกิดขึ้นจากข้อมูลที่ไม่ถูกต้อง
- เพิ่มประสิทธิภาพของโมเดลการวิเคราะห์ โดยเฉพาะในการเรียนรู้ของเครื่อง (Machine Learning) ที่ต้องการข้อมูลที่สะอาดและมีความสม่ำเสมอ
- ลดความซ้ำซ้อน ป้องกันปัญหาข้อมูลซ้ำซ้อนที่อาจทำให้การคำนวณผิดเพี้ยน
- ช่วยให้การตัดสินใจมีความน่าเชื่อถือ ลดโอกาสของการตัดสินใจผิดพลาดจากข้อมูลที่ไม่สมบูรณ์

2.1.2.1 ขั้นตอน Data Cleansing ที่ช่วยให้ข้อมูลมีคุณภาพ

ก. ลบข้อมูลที่ซ้ำซ้อน จากการรวบรวมข้อมูลจากหลายแหล่ง ส่งผลให้อาจมีการดึงข้อชุดข้อมูลที่ซ้ำซ้อน ส่งผลให้ชุดข้อมูลหนักขึ้นและประมวลผลช้า หากใช้โมเดลเป็น Machine Learning ก็อาจทำให้เกิดการให้น้ำหนักกับข้อมูลซ้ำซ้อนมากเกินไป จนไม่สะท้อนความจริง ดังนั้นจึงควรเช็กว่าข้อมูลของเรามีความซ้ำซ้อนหรือไม่และทำการลบ

ข. ลบข้อมูลที่ไม่เกี่ยวข้อง ในการวิเคราะห์ข้อมูล (Data Analytics) จะมีการกำหนดคำถามและสมมติฐานเพื่อระบุสิ่งที่เราอยากรู้ หากมีตัวแปรที่ไม่เกี่ยวข้องอยู่มากเกินไปก็อาจทำให้ผลลัพธ์ที่ได้มีความคลาดเคลื่อน จึงควรเข้าใจคำถามและจุดประสงค์ของการวิเคราะห์เพื่อหาข้อมูลที่ไม่เกี่ยวข้องแล้วทำการลบออก

ค. ระบุข้อมูลแทนที่ข้อมูลเดิม เมื่อมีการลบข้อมูลที่ผิดพลาดหรือไม่เกี่ยวข้องออกไปแล้ว จะต้องมีการทดแทนข้อมูลเดิม ซึ่งกระบวนการนี้ขึ้นอยู่กับการพิจารณาของแต่ละบุคคล อาจจะใช้วิธีดึงข้อมูลจากฐานข้อมูลมาสันนิษฐานและระบุแทนที่ชุดข้อมูลเก่า เพื่อให้ข้อมูลมีความสอดคล้องกัน หรืออาจจะลบข้อมูลไปเฉย ๆ ไม่มีการเพิ่มเติมอะไรเลยก็ได้เช่นกัน

ง. บำรุงรักษาข้อมูลอยู่เสมอ หลังจากจัดเก็บข้อมูลมาในระยะเวลาหนึ่งข้อมูลอาจเกิดการสูญหายตามกาลเวลาจึงต้องมีการบำรุงรักษาข้อมูลอยู่เสมอ เพื่อให้ข้อมูลยังคงสมบูรณ์ไม่สูญหาย

จ. ตรวจสอบความถูกต้อง การตรวจสอบความถูกต้องของข้อมูล เป็นขั้นตอนสุดท้ายที่ตรวจสอบว่าชุดข้อมูลทั้งหมดที่ผ่านกระบวนการ Data Cleansing มีความถูกต้องหรือไม่ (dittothailand, 2025)

2.1.3 แนวคิดการพยากรณ์ข้อมูล (Forecasting data)

การพยากรณ์ข้อมูล (Forecasting data) คือ การประมาณ หรือ การคาดคะเนว่าอะไรจะเกิดขึ้นในอนาคต การพยากรณ์มีบทบาทสำคัญกับทุกด้าน ทั้งหน่วยงานของรัฐบาลและเอกชน รัฐบาลต้องประมาณ หรือ พยากรณ์รายได้รายจ่ายในปีหน้า เพื่อนำมาวางแผน เอกชนต้องพยากรณ์ยอดขาย เพื่อนำมาวางแผนการผลิต สินค้าคงคลัง แรงงาน ฯลฯ

2.1.3.1 การพยากรณ์เชิงปริมาณ (Quantitative methods) เป็นการพยากรณ์ที่ใช้ข้อมูลเชิงปริมาณ (ตัวเลข) ใน อดีตเพื่อนำมาพยากรณ์ค่าในอนาคต โดยสร้างตัวแบบทางคณิตศาสตร์ การพยากรณ์ประเภทนี้แบ่งออกเป็น 2 เทคนิคย่อย คือ

2.1.3.2 การพยากรณ์ความสัมพันธ์ (Casual Forecasting) เป็นเทคนิคที่ใช้ปัจจัยที่คาดว่าจะมีความสัมพันธ์ กับตัวแปรที่จะพยากรณ์ เช่น ถ้าต้องการพยากรณ์ยอดขาย จะพิจารณาหาความสัมพันธ์ระหว่างยอด ขายกับค่าโฆษณา รายได้ของประชากร สภาพสินค้า ฯลฯ การหาความสัมพันธ์ดังกล่าวจะใช้เทคนิคที่ เรียกว่า การวิเคราะห์ความถดถอย และสหสัมพันธ์

2.1.3.3 การพยากรณ์อนุกรมเวลา (Time series Forecasting) เป็นเทคนิคที่ใช้เฉพาะข้อมูลในอดีตของตัวแปร ที่ต้องการพยากรณ์ เพื่อพยากรณ์ค่าของตัวแปรนั้นในอนาคต เช่น ใช้ข้อมูลยอดขายปี 2530–2541 เพื่อพยากรณ์ยอดขายปี 2542 เป็นการพยากรณ์ที่ใช้ผู้ที่มีประสบการณ์ ความรู้ ความสามารถ เป็นผู้พยากรณ์ โดยไม่ใช้ตัวแบบทางคณิตศาสตร์ จึงตรวจสอบความแม่นยำของการพยากรณ์ได้ยากกว่าการพยากรณ์เชิงปริมาณ การพยากรณ์เชิงคุณภาพประกอบด้วย

ก. การคาดคะเน หรือ ประมาณการ (Judgement) วิธีนี้มักใช้กับธุรกิจขนาดเล็กที่มีเจ้าของคนเดียว หรือหน่วยงานขนาดเล็กที่หัวหน้ามีอำนาจเต็ม เจ้าของหรือหัวหน้างานจะคาดการณ์ยอดขาย หรือ สิ่งที่จะเกิดขึ้นในอนาคต โดยอาศัยประสบการณ์ที่ทำงานในด้านนั้นๆ มาเป็นระยะเวลานานพอ

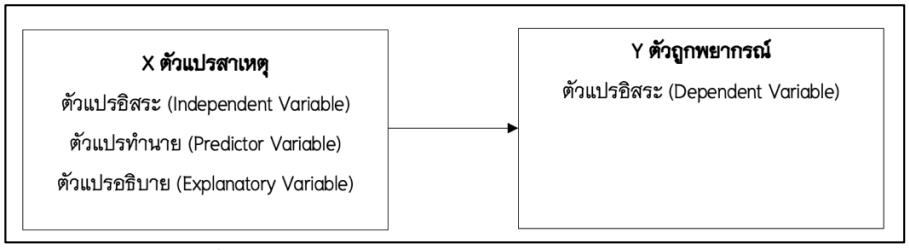
ข. การระดมความคิด (Jury of Executive Operation) วิธีนี้เป็นการระดมความคิด หรือประชุมกลุ่มผู้บริหารของบริษัท เช่น ประชุมคณะกรรมการบริหาร เพื่อให้ทุกคนออกความคิดเห็นเกี่ยวกับสิ่งที่จะเกิดในอนาคต เช่น ยอดขายปีหน้า จะเป็นเท่าใด ควรพัฒนาผลิตภัณฑ์ใหม่หรือไม่ และผลสรุป จะได้เสียงส่วนใหญ่ของการประชุม แต่วิธีนี้จะมีข้อเสียตรงที่อาจเกิดความเอนเอียง หรือ เกรงใจทำให้ไม่กล้าออกความคิดเห็น ถ้าความคิดเห็นไม่ตรงกับคนอื่นๆหรือไม่ตรงกับความคิดเห็นของผู้มีอำนาจมากกว่าหรือผู้ถือหุ้นใหญ่และมักจะเห็นด้วยกับความคิดเห็นของผู้มีอำนาจหรือผู้ถือหุ้นใหญ่

2.1.3.4 การพยากรณ์ยอดขาย (Sale Force Composite Forecasts) เป็นการพยากรณ์โดยให้แต่ละฝ่าย เช่น ให้หัวหน้าฝ่ายขายตามภาคต่างๆ ประมาณยอดขาย แล้วนำมารวมกันทุกภาคกลายเป็นค่าพยากรณ์ยอดขายรวมของบริษัท หรือให้ตัวแทนขายแต่ละคนประมาณยอดขายของตนเองแล้วนำมารวมกันเป็นยอดขายรวมของบริษัท การพยากรณ์ยอดขายโดยวิธีนี้ค่อนข้างจะแม่นยำ เนื่องจากตัวแทนขายแต่ละคน/หน่วยจะใกล้ชิดกับลูกค้า/ตลาดมาก ทำให้คาดคะเนได้ถูกต้อง

2.1.3.5 พยากรณ์โดยการสำรวจตลาด (Survey of Expectations and Anticipations) เป็นการพยากรณ์ยอดขายโดยทำการสำรวจลูกค้าหรือผู้ที่คาดว่าจะเป็นลูกค้าเพื่อตรวจสอบว่าในอนาคตลูกค้าต้องการสินค้าอะไรบ้าง จำนวนเท่าใด ด้วยการทำวิจัยตลาดซึ่งอาจใช้การสัมภาษณ์ตัวต่อตัว โทรศัพท์หรือจดหมาย เป็นต้น

2.1.3.6 การพยากรณ์ด้วยเทคนิคเดลไฟ (Delphi) เทคนิคเดลไฟเป็นเทคนิคที่แก้ข้อเสียของวิธีระดมความคิด ซึ่งอาจก่อให้เกิดความเอนเอียง หรือคล้อยตามผู้อื่น เทคนิคเดลไฟ จึงแก้ปัญหาโดยการไม่ให้ผู้บริหารพบปะกัน หรือมาประชุมกัน หรือระดมความคิดเห็นกันซึ่งๆหน้า แต่จะส่งคำถามเกี่ยวกับสิ่งที่ต้องการพยากรณ์ให้ผู้บริหารทุกคนเขียนตอบมาพร้อมทั้งระบุเหตุผล เช่น ยอดขายปีหน้าควรเป็นเท่าใด ควรออกผลิตภัณฑ์ใหม่หรือไม่ เพราะเหตุใด ดังนั้น โดยวิธีนี้จะได้ความคิดเห็นของทุกคน และไม่มีการชี้นำ เมื่อได้คำตอบจากทุกคนแล้วให้นำมารวมกัน ซึ่งมักจะพบว่าจะมีความคิดเห็นที่แตกต่างกันออกไป ผู้รวบรวมจะต้องสรุป แล้วส่งกลับไปให้ผู้บริหารทุกคนเป็นรอบที่ 2 เพื่อให้แสดงความคิดเห็นเพิ่มเติม เป็นเช่นนี้ไปเรื่อย ๆ จนได้ข้อสรุปเป็นหนึ่งเดียว

2.1.3.7 การพยากรณ์ข้อมูลเชิงเหตุผล เทคนิคการพยากรณ์ข้อมูลเชิงเหตุผลที่ได้รับความนิยม คือ การวิเคราะห์การถดถอย (Regression Model) ซึ่งเป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปร อิสระ(Independent Variable) กับตัวแปรตาม (Dependent Variable) โดยมีวัตถุประสงค์เพื่อ รักษาความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม และศึกษาปัจจัยหรือตัวแปรอิสระ ที่ร่วมกันทำนายหรือพยากรณ์ตัวแปรตามการวิเคราะห์การถดถอยเพื่อหาความสัมพันธ์หรือสร้างสมการทำนายหรือพยากรณ์ตัวแปรตาม (Y) หนึ่งตัวจากกลุ่มตัวแปรอิสระ (X) ตัวเดียวหรือหลายตัวนั้น ตัวแปรอิสระที่นำมาวิเคราะห์จะต้องมีหลักฐานตามทฤษฎีหรือรายงานการวิจัยที่เกี่ยวข้องว่าเป็นตัวแปรต้นเหตุที่ส่งผลต่อตัวแปรตาม (kalasin, 2025)



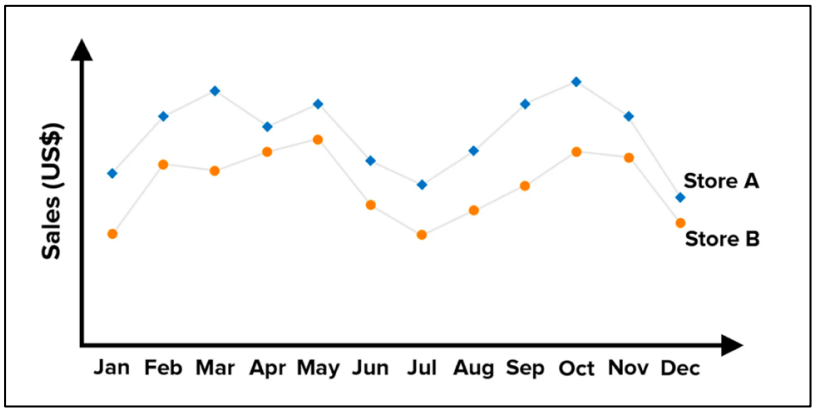
ภาพที่ 2.1 กรอบแนวคิดในการวิเคราะห์ Regression
ที่มา : ethesisarchive.library.tu.ac.th (2556)

2.1.4 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data visualization)

การแสดงผลข้อมูล (Data Visualization) เป็นส่วนประกอบสำคัญใน Cognitive System ซึ่ง เป็นส่วนในการแสดงข้อมูลหรือผลลัพธ์ต่าง ๆ ในระหว่างคอมพิวเตอร์และผู้ใช้งาน ในรูปแบบของ ภาพ โดยผู้ใช้สามารถเรียนรู้และจดจำข้อมูลผ่านการมองได้มากกว่าการใช้ประสาทสัมผัสอื่น ๆ หรือจะกล่าวว่า Visualization ก็คือ การสร้างมโนภาพของสิ่งต่าง ๆ ที่เราสนใจขึ้นมาในใจ ซึ่งต่อมา ได้กลายเป็นการนำภาพมาใช้กับการนำเสนอหรือนามาเป็นกรอบความคิด ซึ่งได้นำไปใช้ในการ สนับสนุนการตัดสินใจ หากไม่มีการทำ Data Visualization แล้วอาจทำให้เราไม่สามารถค้นพบ นัยยะของข้อมูลในแง่ของแนวโน้ม, รูปแบบพฤติกรรม, และความสัมพันธ์เชื่อมโยงได้

การเลือกรูปแบบ Visualization ให้เหมาะสมกับข้อมูล ในปัจจุบันเป็นยุคที่เทคโนโลยีเข้าถึง ทุกคน ทำให้การรับรู้ข่าวสาร ข้อมูลต่างๆ เป็นไปได้ง่าย และรวดเร็วมากขึ้น คนที่นำเสนอข้อมูลจึง ต้องนำเสนอข้อมูลให้น่าสนใจ เข้าใจง่าย และรวดเร็ว จึงเกิดการสร้าง Data Visualization ขึ้นมา Data Visualization เป็นการใช้ภาพเพื่อแสดงข้อมูลในเชิงปริมาณที่วัดได้ ซึ่งอาจนำเสนอออกมาใน รูปแบบ แผนภูมิ กราฟ กราฟิก และอื่นๆ อีกมากมาย เพื่อให้เข้าใจได้ง่าย (Data Wow Co., 2025)

กราฟ (Graph) เป็นรูปแบบการนำเสนอข้อมูลที่แสดงตำแหน่งของข้อมูล และการลากเส้นเชื่อมในแต่ละจุด เพื่อดูแนวโน้มของข้อมูล หรือการเปลี่ยนแปลงของข้อมูลเมื่อเวลาผ่านไป subset หรือประเภทหนึ่งของแผนภูมิ โดยกราฟจะทำหน้าที่แสดงความสัมพันธ์ระหว่างข้อมูล 2 ตัวแปร ผ่านแกนแนวนอน (แกน X) และแกนแนวตั้ง (แกน Y) ช่วยให้เห็นเทรนด์สถานการณ์ประกอบกับบริบท (1stCraft, 2025)

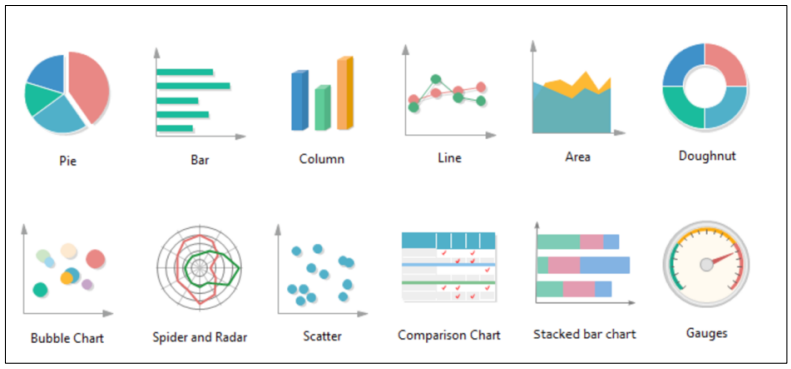


ภาพที่ 2.2 กราฟ (Graphs)

ที่มา : mindtools.com

แผนภูมิ (Charts) ซึ่งเป็นรูปแบบที่น่าจะคุ้นเคยกันมากที่สุด และเป็นรูปแบบที่มีหลากหลายชนิดที่เหมาะกับการนำเสนอข้อมูลที่แตกต่างกันไปตามวัตถุประสงค์ เช่น Pie chart จะช่วยให้เราเห็นปริมาณความแตกต่างได้ชัดเจน, Comparison chart เหมาะสำหรับการ

เปรียบเทียบคุณสมบัติหลายๆ ข้อ, มาตรวัด (Gauges) จะช่วยให้เห็นความเข้มข้น ความรุนแรง หรือน้ำหนัก (1stCraft, 2025)



ภาพที่ 2.3 แผนภูมิ (Charts)

ที่มา : medium.com/@Lynia_Li

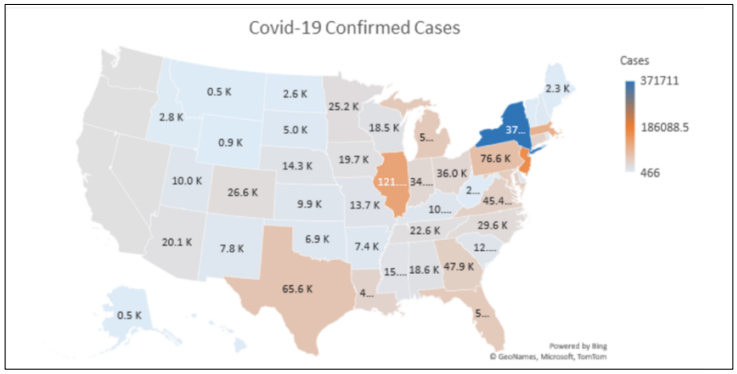
ตาราง (Tables) เป็นรูปแบบที่ใช้กันมากเพื่อนำเสนอข้อมูลให้ออกมาดูง่าย ตารางประกอบไปด้วย 2 ส่วน ได้แก่ คอลัมน์และแถว ซึ่งช่วยจัดการข้อมูลให้เรียบร้อย ช่วยให้มองเห็นบริบทและความสัมพันธ์ของข้อมูลหลายๆ ชุด (1stCraft, 2025)

Marks	Number of Students		Total
	Males	Females	
30 – 40	8	6	14
40 – 50	16	10	26
50 – 60	14	16	30
60 – 70	12	8	20
70 – 80	6	4	10
Total	56	44	100

ภาพที่ 2.4 ตาราง (Tables)

ที่มา : embibe.com

แผนที่ (Maps) เป็นการนำเสนอข้อมูลบนแผนที่เพื่อแสดงข้อมูลเกี่ยวกับพื้นที่ต่างๆ ซึ่งนอกจากการใส่ข้อมูลลงไปยังพื้นที่ต่างๆ แล้ว ยังสามารถใช้สีสันเพื่อบอกช่วงปริมาณหรือความหนาแน่นของผู้ติดเชื้ออีกด้วย (1stCraft, 2025)



ภาพที่ 2.5 แผนที่ (Maps)

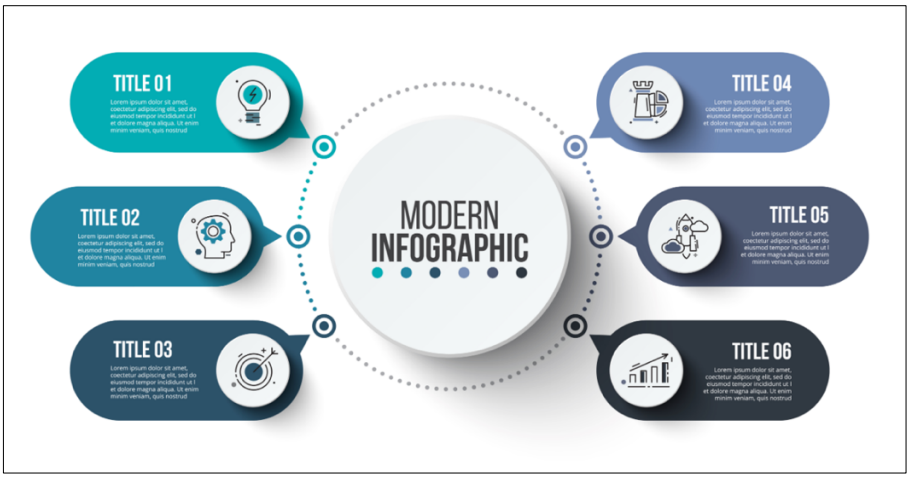
ที่มา : spreadsheetweb.com

แดชบอร์ด (Dashboards) คือ การนำข้อมูลต่างๆ มาเรียบเรียงและสรุปเป็นภาพ โดยใช้แผนภูมิและกราฟต่างๆ มาใช้นำเสนอ ปัจจุบันแดชบอร์ดเป็น Data Visualization ที่นิยมใช้กับการนำเสนอข้อมูลแบบ Real-time ผ่านซอฟต์แวร์หรือเครื่องมือจัดการและวิเคราะห์ข้อมูลต่างๆ เช่น เครื่องมือการตลาด เครื่องมือบริหารจัดการข้อมูล เครื่องมือติดตามและดูแลเว็บไซต์ (1stCraft, 2025)

ภาพที่ 2.6 แดชบอร์ด (Dashboards)

ที่มา : datawow.co.th

อินโฟกราฟิก (Infographics) เป็นการนำเสนอข้อมูลโดยการใช้เครื่องมือด้านกราฟิกเข้ามาช่วย สามารถใส่เทคนิค Storytelling เพื่อให้ข้อมูลน่าสนใจและดึงดูดมากยิ่งขึ้น ดูแล้วเข้าใจง่ายในเวลารวดเร็ว (1stCraft, 2025)



ภาพที่ 2.7 อินโฟกราฟิก (Infographics)

ที่มา : datawow.co.th

การทำ Data Visualization สามารถสรุปข้อมูลที่ซับซ้อนให้อยู่ในรูปแบบที่เข้าใจง่าย ตรงประเด็น สามารถสื่อสารออกไปให้เข้าใจตรงกันทั้งผู้ส่งสารและผู้รับสาร และยังมีประโยชน์อื่น ๆ ในการช่วยให้ธุรกิจของคุณทำงานได้อย่างมีคุณภาพ (1stCraft, 2025)

2.2 ทฤษฎีที่เกี่ยวข้อง

2.2.1 ทฤษฎีเหมืองข้อมูล (Data Mining)

เหมืองข้อมูล (Data Mining) คือ กระบวนการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อค้นหารูปแบบ แนวโน้ม และความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูล โดยอาศัยเทคนิคทางสถิติ ปัญญาประดิษฐ์ และการเรียนรู้ของเครื่อง (Machine Learning) เพื่อช่วยในการตัดสินใจและคาดการณ์แนวโน้มในอนาคต เหมืองข้อมูลเป็นส่วนหนึ่งของกระบวนการ การค้นพบความรู้ในฐานข้อมูล (Knowledge Discovery in Databases: KDD) ซึ่งครอบคลุมตั้งแต่การรวบรวมข้อมูล การทำความสะอาด การแปลงข้อมูล การวิเคราะห์ข้อมูล และการนำผลลัพธ์ไปใช้

2.2.1.1 กระบวนการทำเหมืองข้อมูล

ก. การเก็บรวบรวมข้อมูล (Data Collection) รวบรวมข้อมูลจากแหล่งต่างๆ เช่น ฐานข้อมูล ระบบสารสนเทศ หรือข้อมูลบนอินเทอร์เน็ต

ข. การทำความสะอาดข้อมูล (Data Cleaning) กำจัดข้อมูลที่ซ้ำซ้อน ข้อมูลที่ขาดหาย และค่าผิดปกติ เพื่อให้ข้อมูลมีความสมบูรณ์และเชื่อถือได้

ค. การแปลงข้อมูล (Data Transformation) ปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ เช่น การทำให้ข้อมูลอยู่ในช่วงค่าที่ใกล้เคียงกัน (Normalization)

ง. การเลือกข้อมูล (Data Selection) คัดเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับเป้าหมายของการวิเคราะห์

จ. การทำเหมืองข้อมูล (Data Mining) ใช้เทคนิคและอัลกอริทึมในการวิเคราะห์ข้อมูล เช่น การจำแนกประเภท การจัดกลุ่ม และการทำนายแนวโน้ม

ฉ. การประเมินผลและการนำเสนอข้อมูล (Evaluation & Interpretation) ตรวจสอบความถูกต้องของโมเดล และนำเสนอผลลัพธ์ในรูปแบบที่เข้าใจง่าย เช่น รายงานหรือการแสดงผลข้อมูล

2.2.1.2 การจำแนกประเภทของข้อมูล (CLASSIFICATION) ใช้สำหรับแบ่งข้อมูลออกเป็นกลุ่ม เช่น การพิจารณาว่าลูกค้าจะซื้อสินค้าหรือไม่ ตัวอย่างอัลกอริทึมที่ใช้ ได้แก่ DECISION TREE, NAÏVE BAYES และ SUPPORT VECTOR MACHINE (SVM)

2.2.1.3 การจัดกลุ่มข้อมูล (CLUSTERING) ใช้แบ่งข้อมูลออกเป็นกลุ่มโดยไม่มีการกำหนดกลุ่มล่วงหน้า เช่น การแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้า ตัวอย่างอัลกอริทึมที่ใช้ ได้แก่ K-MEANS CLUSTERING และ HIERARCHICAL CLUSTERING

2.2.1.4 การค้นหากฎความสัมพันธ์ (Association Rule Mining) ใช้เพื่อค้นหาความสัมพันธ์ระหว่างข้อมูล เช่น การวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) ตัวอย่างอัลกอริทึมที่ใช้ ได้แก่ Apriori Algorithm และ FP-Growth Algorithm

2.2.1.5 การวิเคราะห์ค่าผิดปกติ (Anomaly Detection) ใช้ตรวจสอบข้อมูลที่แตกต่างจากปกติ เช่น การตรวจจับการฉ้อโกงทางการเงิน ตัวอย่างอัลกอริทึมที่ใช้ ได้แก่ Isolation Forest และ Local Outlier Factor (LOF)

2.2.1.6 การพยากรณ์ข้อมูล (Prediction & Forecasting) ใช้ข้อมูลในอดีตเพื่อคาดการณ์ค่าที่อาจเกิดขึ้นในอนาคต เช่น การพยากรณ์ยอดขายสินค้าในเดือนถัดไป ตัวอย่างอัลกอริทึมที่ใช้ ได้แก่ Linear Regression และ Time Series Analysis (ARIMA, LSTM)

เหมืองข้อมูลเป็นกระบวนการวิเคราะห์ข้อมูลที่ช่วยให้สามารถค้นพบรูปแบบความสัมพันธ์ และแนวโน้มที่ซ่อนอยู่ ซึ่งสามารถนำไปใช้ในหลายอุตสาหกรรมเพื่อช่วยในการตัดสินใจและพยากรณ์แนวโน้มในอนาคต ด้วยเทคนิคต่าง ๆ เช่น การจำแนกประเภท การจัดกลุ่ม และการวิเคราะห์ความสัมพันธ์ ทำให้เหมืองข้อมูลเป็นเครื่องมือสำคัญในการจัดการและใช้ประโยชน์จากข้อมูลขนาดใหญ่

2.2.2 ทฤษฎีการจำแนกประเภทของข้อมูล (Classification)

การจำแนกประเภทของข้อมูล (Classification) หรือเทคนิคการจำแนก เป็นหนึ่งใน Machine Learning ที่ช่วยในการทำเหมืองข้อมูล อยู่ในกลุ่มของ Supervised Learning โดยเพื่อน ๆ จำเป็นต้องกำหนดวัตถุประสงค์ เพื่อให้โมเดลสามารถเรียนรู้จากข้อมูลที่เรานำเข้าและให้ผลลัพธ์ที่ตอบคำถามตามวัตถุประสงค์ได้ ซึ่งคำตอบที่ได้จาก Classification Model จะอยู่ในรูปแบบค่าที่ไม่ต่อเนื่อง (Discrete Value)

2.2.2.1 ประเภทของ Classification Model

ก. Binary Classification (จำแนกแบบไบนารี่) การจำแนกแบบ 2 ตัวแปร ผลลัพธ์ที่ได้จากข้อมูลจะมีเพียงแค่ 2 หมวดหมู่ที่เรากำหนดไว้เท่านั้น เช่น ข้อมูลลักษณะของลูกค้าที่เข้าร้านค้าซื้อสินค้าหรือไม่ คำตอบที่ได้จะมีเพียง “ซื้อ” หรือ “ไม่ซื้อ”

ข. Multi-Class Classification (จำแนกแบบหลายคลาส) การจำแนกแบบหลายหมวดหมู่ ลักษณะการวิเคราะห์จะใกล้เคียงกับ Binary แต่ผลลัพธ์ที่ได้จะมีมากกว่า 2 คำตอบ เป็นการทำนายสิ่งที่ใกล้เคียงกับข้อมูลที่ป้อน เช่น การจำแนกรูปภาพว่าเป็น “รถ” หรือ “เรือ” หรือไม่ใช่ทั้งคู่

ค. Multi-Label Classification (จำแนกแบบหลายเลเบล) เพื่อน ๆ หลายคนอาจจะเข้าใจสับสนหระหว่าง Multi-Class และ Multi-Label ในส่วนของ Multi- Class เป็นการจำแนกข้อมูลที่มีคุณสมบัติเหมือนกัน แต่ให้ผลลัพธ์ที่แตกต่างกันหลายประเภท ส่วน Multi-Label เป็นการจำแนก

ข้อมูลที่อาจมีคุณสมบัติเหมือนกันแต่ผลลัพธ์สามารถแตกต่างกันได้ ซึ่ง Multi-Label จะมีความซับซ้อนในการวิเคราะห์มากกว่า

2.2.2.2 ขั้นตอนการทำ Classification

ก. กำหนดวัตถุประสงค์ ในขั้นตอนการสร้างโมเดล เพื่อน ๆ จำเป็นต้องรู้ถึงวัตถุประสงค์ในการวิเคราะห์เสียก่อน จะได้รู้ว่าสร้างโมเดลขึ้นมาเพื่ออะไร ต้องการผลลัพธ์แบบไหน และจะเลือกใช้ Algorithm ตัวไหนในการทำ

ข. การจัดเตรียมข้อมูล การจะได้โมเดล Classification ที่ดีนั้นจะมีสิ่งที่เพื่อน ๆ จะต้องคำนึงถึงอยู่ สิ่งแรกเลยก็คือการเตรียมข้อมูล (Data Preprocessing) ให้ดีไม่ว่าจะเป็นการทำความสะอาดข้อมูล หรือการเปลี่ยนรูปข้อมูล ให้พร้อมต่อการนำไปใช้สร้างโมเดล เพราะการที่เตรียมข้อมูลที่ไม่ดีพออาจจะส่งผลให้การวิเคราะห์ข้อมูลไม่แม่นยำได้

ค. แบ่งข้อมูลชุด Train และ Test ในการสร้าง Classification Model จำเป็นต้องมีการแบ่งชุดข้อมูลออกเป็น Training และ Test ในส่วนนี้หากเพื่อน ๆ แบ่งสัดส่วนข้อมูลไม่เหมาะสม อาจจะทำให้เกิดการ Overfitting ที่โมเดลทำงานได้ดีมากเฉพาะบน Training Set หรือ Underfitting ที่ข้อมูลไม่เพียงพอจนทำให้โมเดลได้

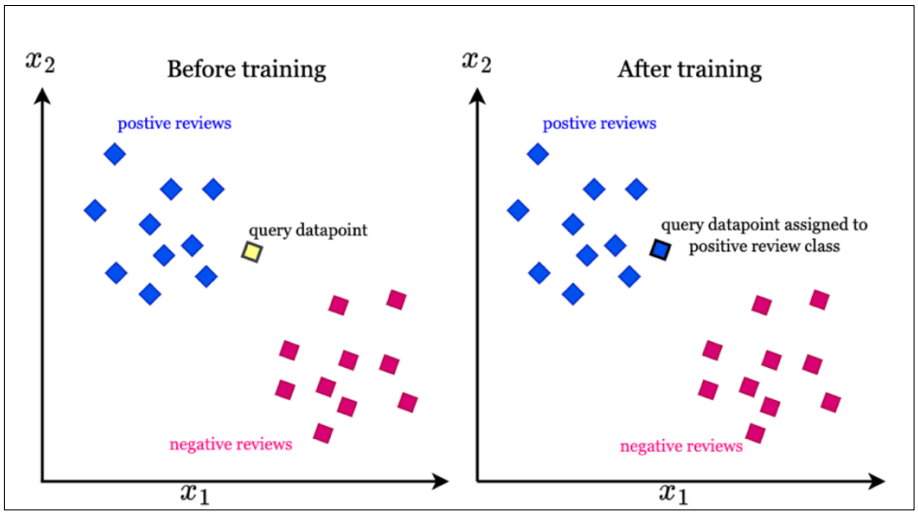
ง. Training และ Testing โมเดล หลังจากแบ่งข้อมูลแล้ว ให้เพื่อน ๆ Input ข้อมูลชุด Training เข้าไปสู่โมเดล เพื่อให้โมเดลเรียนรู้ความสัมพันธ์ต่าง ๆ ภายในชุดข้อมูล และส่งต่อโมเดลเพื่อทดสอบในข้อมูลชุด Test เพื่อทดสอบประสิทธิภาพของโมเดล

จ. การวัดผลโมเดล เมื่อทดสอบโมเดลเสร็จเรียบร้อยแล้วก่อนจะนำโมเดลออกไปใช้งานจริง จำเป็นต้องมีการวัดผลประสิทธิภาพของโมเดล โดย Classification Model แบบ 2 ตัวแปร จะใช้ Confusion Matrix ในการวัดผลและแสดงค่าความแม่นยำออกมา หากค่าความแม่นยำต่ำ อาจจำเป็นต้องทำซ้ำและปรับเปลี่ยนเพื่อทดสอบให้ได้โมเดลที่ดีที่สุด

2.2.2.3 Classification Algorithm

K-NN Algorithm หรือ K-nearest Neighbors เป็นหนึ่งในวิธีการจำแนกข้อมูล ที่เหมาะกับการวิเคราะห์และแก้ปัญหากับชุดข้อมูลที่เพื่อน ๆ รู้จำนวนการแบ่งหมวดหมู่ที่แน่นอนอยู่แล้วเพื่อนำไปจำแนกข้อมูลที่ไม่สามารถระบุหมวดหมู่ได้ โดยอาศัยหลักการการเปรียบเทียบข้อมูลในตำแหน่งที่สนใจกับข้อมูลอื่น ๆ ที่ระบุหมวดหมู่ไว้แล้วว่าใกล้เคียงกับข้อมูลไหนมากที่สุด ตัวโมเดลก็จะให้ผลลัพธ์เป็นคลาสของข้อมูล หลักการทำงานของ K-NN คือการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลใกล้เคียงโดยจำนวนที่โมเดลจะตรวจสอบข้อมูล เพื่อน ๆ

สามารถกำหนดได้เอง โดยใช้ค่าที่เรียกค่า K ที่เปรียบเสมือนค่าที่จะกำหนดขอบเขตในการตัดสินใจ เช่น K = 2 โมเดลก็จะทำการตรวจสอบข้อมูล 2 ตัวที่ใกล้กับเป้าหมายมากที่สุด



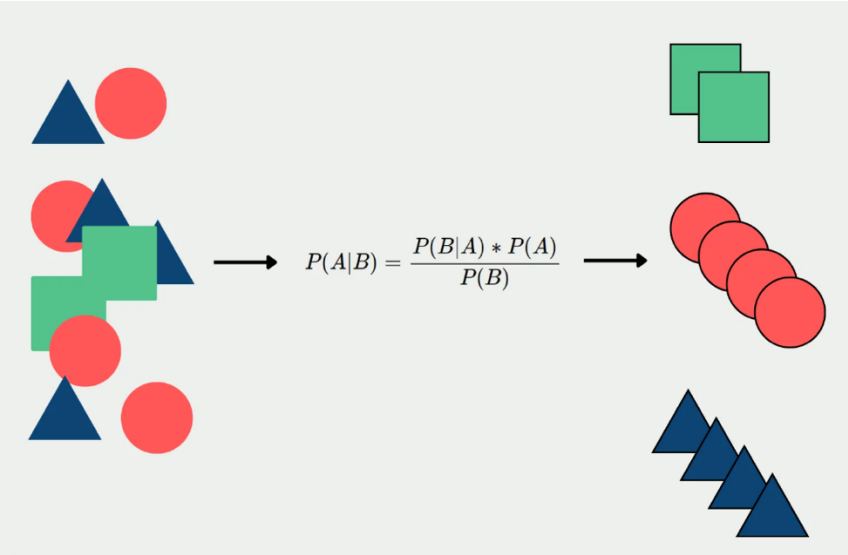
ภาพที่ 2.8 การใช้อัลกอริทึม K-nearest Neighbors

ที่มา : intuitivetutorial.com

Naive Bayes เป็น Classification Model ที่ใช้สำหรับทำนายว่าข้อมูลอยู่ในกลุ่มไหน ซึ่งเป็นโมเดลที่ต้องอาศัยการกำหนด Label หรือแยกหมวดหมู่ไว้แล้วในการสอน โดยใช้หลักการความน่าจะเป็นทางสถิติ (Probability) ของ Bayesเพื่อทำนายว่าข้อมูลอยู่ในกลุ่มไหน โดยการคำนวนของ Naive Bayes Algorithm ด้วยสมการตามภาพด้านบน เพื่อคำนวนหาค่า P(CIA) หรือความน่าจะเป็นที่ข้อมูลที่มี Attribute เป็น A จะมีคลาส C ซึ่งจากคลาสทั้งหมดตัวไหนที่มีค่าความน่าจะเป็นนี้มากกว่า ก็จะได้ผลลัพธ์ออกมาเป็นค่านั้น ซึ่งตัวโมเดลนี้ก็สามารถนำไปใช้ได้หลากหลายทั้งการทำ Sentiment Analysis, Spam Filtering ไปจนถึงเป็นพื้นฐานของ Language Model

$$\boldsymbol{P(A|B) = \frac{P(B|A) * P(A)}{P(B)}}$$

- P(BIA) = ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น ถ้าเหตุการณ์ A เกิดขึ้นแล้ว
- P(A) = ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้น
- P(B) = ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น

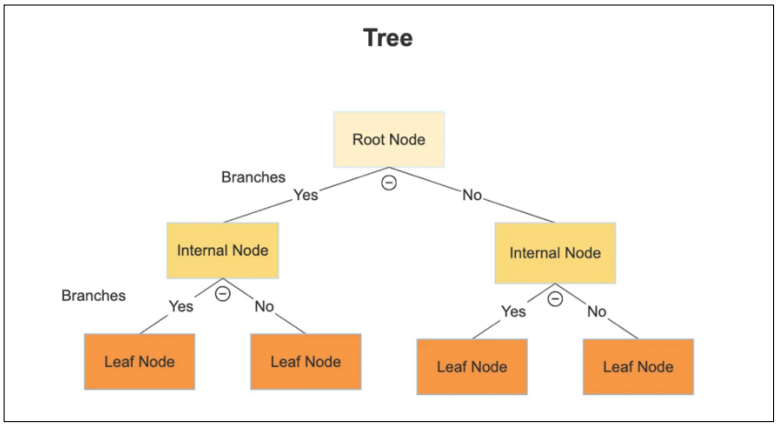


ภาพที่ 2.9 การใช้อัลกอริทึม Naive Bayes

ที่มา : databasecamp.de

Decision Tree Algorithm เป็นโมเดลในรูปแบบของแผนภูมิต้นไม้ (Tree Diagram) ที่มีคอนเซปในการเลียนแบบการตัดสินใจของมนุษย์ มาตัดสินใจในข้อมูลที่ไม่เคยเห็น ซึ่งเป็น Rule-Based Model ที่ใช้การสร้างเงื่อนไข If-else จาก Attribute ต่าง ๆ ภายในชุดข้อมูล โดยมีวัตถุประสงค์ในการแบ่งข้อมูลออกเป็นแต่ละประเภทเพื่อที่จะอธิบายกลุ่มข้อมูลได้ดีที่สุด ซึ่งโมเดล Decision Tree แบ่งออกได้ 2 ประเภท คือ Regression Tree ที่ใช้ในการแก้ปัญหาโจทย์ Regression และ Classification Tree สำหรับจำแนกประเภท

Decision Tree นั้นประกอบด้วย Node โดยโหนดแรกจะเรียกว่า Root Node และโหนดสุดท้ายเรียกว่า Leaf Node โดยสิ่งที่เป็นหลักสำคัญในการสร้างโมเดล Decision Tree คือ การ Split ค่า Feature หรือ Attribute แต่ละครั้งต้อง Minimize ค่า Cost Function โดยวัดออกมาด้วยค่า Gini Impurity คือการวัดค่าความผิดพลาดของการแบ่ง (Impurity) ที่ยิ่งค่าน้อยก็จะแสดงให้เห็นว่าโมเดลจำแนกได้ดี



ภาพที่ 2.10 การใช้อัลกอริทึม Decision Tree

ที่มา : medium.com

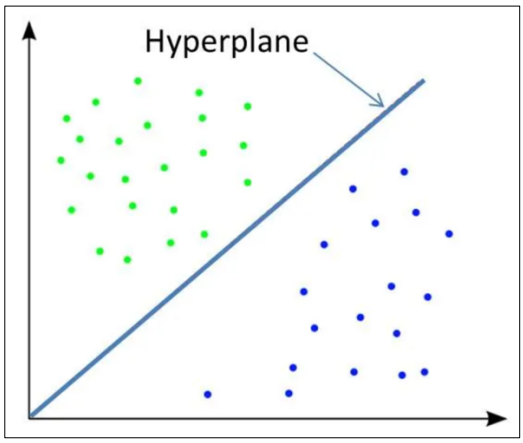
ต้นไม้ตัดสินใจคือความง่ายในการเข้าใจและการตีความ เนื่องจากโครงสร้างของมันสามารถแสดงผลในรูปแบบที่ชัดเจนและเป็นธรรมชาติ อย่างไรก็ตาม ต้นไม้ตัดสินใจอาจมีความซับซ้อนเกินไปหากไม่มีการควบคุม ทำให้เกิดปัญหาการฟิตเกิน (Overfitting) ดังนั้น การตัดแต่ง (Pruning) ต้นไม้จึงเป็นสิ่งจำเป็นเพื่อขจัดโหนดที่ไม่จำเป็นและปรับปรุงประสิทธิภาพของโมเดล ต้นไม้ตัดสินใจสามารถใช้ได้ทั้งในงานการจำแนกประเภท (Classification) และการถดถอย (Regression)

Support Vector Machine – SVM) เป็นในโมเดล Machine Learning ที่ใช้ในการจำแนกข้อมูล หรือแบ่งกลุ่มข้อมูลโดยจะสร้างเส้นตรงที่ใช้แบ่งกลุ่มข้อมูล (Hyperplane) และหาเส้นที่ดีที่สุด ถือเป็นวิธีคลาสสิคที่น่าเรียนรู้มากทีเดียว เพราะไอเดียจากโมเดลนี้ก็เป็นหนึ่งในรากฐานสำคัญที่ทำให้เข้าใจโมเดลใหญ่ๆในปัจจุบันมากขึ้น อีกทั้งหนึ่งในผู้ที่คิดค้นวิธี Vladimir Naumovich Vapnik ซึ่งเป็นชาวรัสเซีย และยังเป็นนักวิทยาศาสตร์ที่ยิ่งใหญ่ในวงการ Machine Learning สำหรับปัญหาการจำแนกประเภทแบบสองกลุ่ม (Binary Classification) SVM จะสร้าง เส้นแบ่งเขตที่ดีที่สุด ซึ่งเรียกว่า Hyperplane เพื่อแยกกลุ่มของข้อมูลออกจากกัน โดยพยายามให้ ระยะห่าง (Margin) ระหว่างข้อมูลของแต่ละกลุ่มกับเส้น Hyperplane มีค่ามากที่สุด เส้น Hyperplane ในปัญหาการจำแนกแบบสองมิติสามารถเขียนได้ในรูปของสมการเชิงเส้นดังนี้

$$w \cdot x + b = 0$$

โดยที่

- w = เวกเตอร์น้ำหนัก (Weight Vector) ซึ่งกำหนดทิศทางของ Hyperplane
- x = คือเวกเตอร์ของจุดข้อมูล
- b = คือค่าคงที่หรือตัวตัดแกน (Bias)
-



ภาพที่ 2.11 เส้น Hyperplane
ที่มา : medium.com

เงื่อนไขของ SVM ในปัญหาการจำแนกประเภทแบบสองกลุ่ม ข้อมูลที่อยู่คนละกลุ่มจะมีป้ายกำกับ (Label) เป็น y_i ซึ่งมีค่าเป็น +1 หรือ −1

SVM กำหนดเงื่อนไขให้ข้อมูลแต่ละจุดต้องอยู่ด้านที่ถูกต้องของ Hyperplane ตามสมการ

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i$$

- ถ้า y_i = + 1 (อยู่ฝั่งบวก) ต้องมีค่า $w \cdot x_i + b \geq 1$
- ถ้า y_i = – 1 (อยู่ฝั่งลบ) ต้องมีค่า $w \cdot x_i + b \leq -1$

เป้าหมายของ SVM คือการทำให้ Margin กว้างที่สุด ซึ่ง Margin มีค่าเป็น

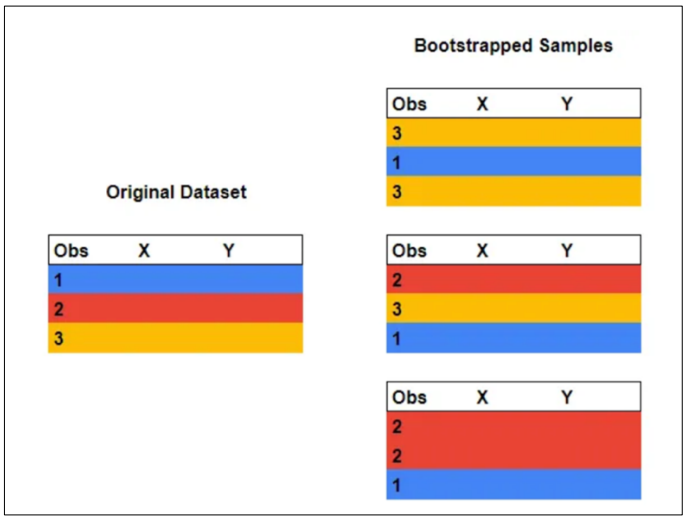
$$\frac{2}{||w||}$$

ดังนั้น เราสามารถหา และ ที่เหมาะสมได้โดยการแก้ปัญหาการหาค่าต่ำสุดของฟังก์ชันวัตถุประสงค์ต่อไปนี้

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

ซึ่งเป็นปัญหาการหาค่าต่ำสุดแบบกำหนดเงื่อนไข (Constrained Optimization Problem) และสามารถแก้ได้โดยใช้ ตัวคูณของลากรังจ์ (Lagrange Multipliers) SVM เป็นหนึ่งในอัลกอริทึมที่มีความแม่นยำสูงและใช้กันอย่างแพร่หลายในการเรียนรู้ของเครื่อง เช่น งานด้านการจดจำรูปภาพ การวิเคราะห์ข้อความ และการตรวจจับอีเมลสแปม

Random Forest เป็น Model ประเภทหนึ่งของ Machine Learning ถูกพัฒนาขึ้นจาก Decision Tree ต่างกันที่ Random Forest เป็นการเพิ่มจำนวน Tree เป็น Tree หลายๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้น แม่นยำมากขึ้น ซึ่งโมเดล Random Forest เป็นโมเดลที่ได้รับความนิยมไปอย่างมากในการใช้ Machine Learning

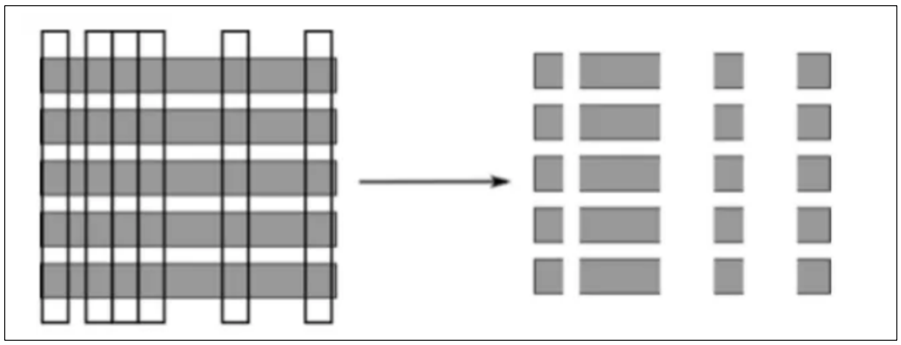


ภาพที่ 2.12 ตัวอย่างการแบ่งข้อมูลออกเป็น Tree แต่ละต้น

ที่มา : medium.com

Bagging จะมีการแบ่งข้อมูลออกเป็น Tree หลายๆ ต้นแต่การทำ Bagging จะมีปัญหาเรื่องความไม่เป็นอิสระของข้อมูลเนื่องจากต่อให้เราแยกออกไปหลายๆ Tree ก็จริงแต่มันก็คือข้อมูลเดียวกัน Random Forest จึงเข้ามาแก้ปัญหาตรงนี้ โดยการทำ Random Sample Feature

Random Sample Feature คือ นอกจากจะแบ่งเป็น Tree หลายๆ ต้นแล้ว ยังแบ่ง Feature ของ Tree แต่ละต้นจะมี Feature ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้แต่ละ Tree มีความหลากหลายและมีความอิสระกันมากขึ้น

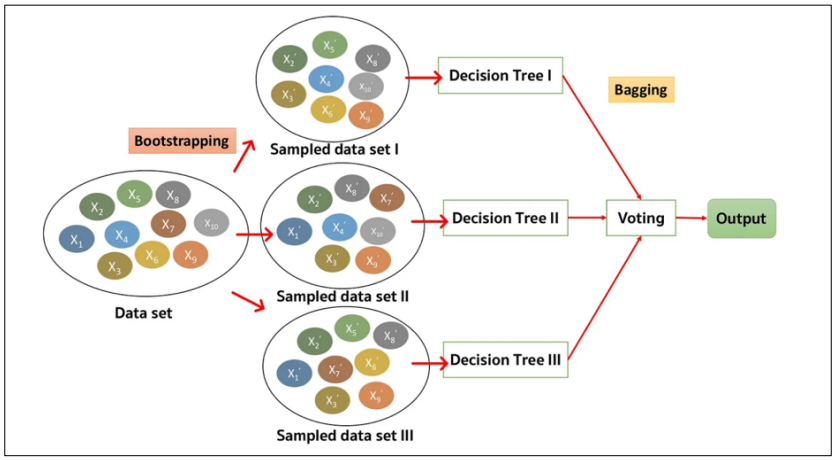


ภาพที่ 2.13 ตัวอย่างการทำ Random Sample Feature

ที่มา : medium.com

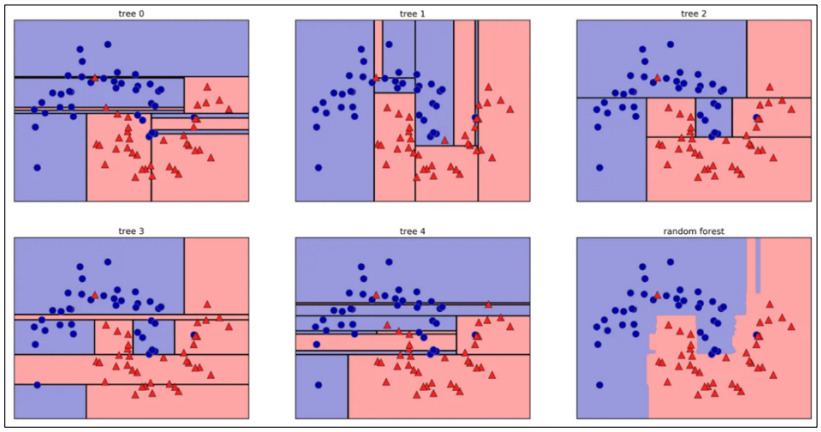
หลักการ Random Forest

- sample ข้อมูล (bootstrapping) จาก data set ทั้งหมด ให้ได้ข้อมูลออกมา n ชุด ที่ไม่เหมือนกัน ตามจำนวน Decision Tree ใน Random Forest เช่น data set ตั้งต้นมีอยู่ 10 features (X1, X2,...,X10) แต่ละ Decision Tree จะได้ feature ไปไม่เหมือนกัน และ จะได้ข้อมูลไม่ครบทุก row ด้วยจาก data set ทั้งหมดด้วย (X1 –> X1',X2–>X2',...)
- สร้าง model Decision Tree สำหรับแต่ละชุดข้อมูล
- ทำ aggregation ผลลัพธ์ จากแต่ละ model (bagging) เช่น voting ในกรณี classification หรือ หาค่า mean ในกรณี regression



ภาพที่ 2.14 หลักการทำ Random Forest

ที่มา : medium.com



ภาพที่ 2.15 ตัวอย่าง Trees แต่ละต้น

ที่มา : medium.com

Random Forest Error Rate

- Correlation ระหว่าง Tree แต่ละต้น ยิ่ง Correlation กันมากยิ่งแย่
- Single trees ยิ่งสูงยิ่งดี

2.2.2.4 การวัดประสิทธิภาพ (Evaluation Metrics) สำหรับการจำแนกประเภท (Classification) มีดังนี้

ก. Accuracy (ความแม่นยำ) Accuracy หมายถึง จำนวนของตัวอย่างที่ได้ตอบถูกเป็นเปอร์เซ็นต์ของตัวอย่างทั้งหมดในชุดข้อมูล นับตั้งแต่คำตอบที่ได้ตรงกับคำตอบจริงถึงคำตอบที่ได้ตรงกับคำตอบผิดสูตรคำนวณคือ Accuracy = (จำนวนตัวอย่างที่ตอบถูก / จำนวนตัวอย่างทั้งหมด) x 100%

ข. Precision (ความแม่นยำทางเชิงบวก) Precision หมายถึง ความสามารถในการตอบถูกที่เป็นบวก (Positive) เมื่อโมเดลทำนายว่าเป็นบวก สูตรคำนวณคือ Precision = (จำนวน True Positive / จำนวนที่โมเดลทำนายว่าเป็น Positive) x 100%

ค. Recall (ความสามารถในการตอบถูกที่เป็นบวกทั้งหมด) Recall หมายถึง ความสามารถในการตอบถูกที่เป็นบวกทั้งหมด (ที่มีในชุดข้อมูล) เมื่อโมเดลทำนายวางเป็นบวก สูตรคำนวณคือRecall = (จำนวน True Positive / จำนวนที่เป็น Positive จริงๆ) x 100%

ง. F1-Score F1-Score เป็นค่าเฉลี่ยความสอดคล้องของค่า Precision และ Recall ซึ่งจะให้ค่าดีเมื่อทั้ง Precision และ Recall มีค่าสูงเท่ากัน สูตรคำนวณคือ F1-Score = 2 x (Precision x Recall) / (Precision + Recall) Confusion Matrix (เมทริกซ์การสับเปลี่ยน)

จ. Confusion Matrix เป็นเมทริกซ์ที่แสดงผลการทำนายของโมเดล โดยแบ่งเป็น 4 ช่อง ได้แก่ True Positive (TP), False Positive (FP), False Negative (FN), และ True Negative (TN) โดยที่

- True Positive (TP) คือ จำนวนที่โมเดลทำนายว่าเป็น Positive และเป็น Positive จริงๆ
- False Positive (FP) คือ จำนวนที่โมเดลทำนายว่าเป็น Positive แต่เป็น Negative จริงๆ
- False Negative (FN) คือ จำนวนที่โมเดลทำนายว่าเป็น Negative แต่เป็น Positive จริงๆ
- True Negative (TN) คือ จำนวนที่โมเดลทำนายว่าเป็น Negative และเป็น Negative จริงๆ

โดยการใช้ Confusion Matrix จะช่วยในการวิเคราะห์ปัญหาและปรับปรุงโมเดลให้ดีขึ้นได้ ซึ่งสามารถนำมาคำนวณเป็น Precision, Recall, และ Accuracy ได้ดังนี้

- Accuracy = (TP + TN) / (TP + FP + FN + TN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)

2.2.3 ทฤษฎีการถดถอยพหุคูณ (Multiple Regression)

การถดถอยพหุคูณ (Multiple Regression) เป็นเทคนิคทางสถิติที่ใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) กับตัวแปรอิสระหลายตัว (Independent Variables) โดยมีสมมติฐานว่าตัวแปรอิสระเหล่านี้มีผลต่อการเปลี่ยนแปลงของตัวแปรตาม

2.2.3.1 สมมติฐานของการถดถอยพหุคูณ

ก. ความเป็นเส้นตรง (Linearity) ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระต้องเป็นเส้นตรง

ข. ความเป็นอิสระของข้อผิดพลาด (Independence): ข้อผิดพลาดของแต่ละการสังเกตต้องไม่มีความสัมพันธ์กัน

ค. ความแปรปรวนคงที่ (Homoscedasticity) ข้อผิดพลาดมีความแปรปรวนคงที่ในทุกค่าของตัวแปรอิสระ

ง. การแจกแจงแบบปกติของข้อผิดพลาด (Normality) ข้อผิดพลาดต้องมีการแจกแจงแบบปกติ

จ. ไม่มีปัญหาสหสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) ตัวแปรอิสระไม่ควรมีความสัมพันธ์กันมากเกินไป

2.2.3.2 การวัดคุณภาพของโมเดล

ก. ค่าสัมประสิทธิ์กำหนด (R-Squared,) วัดว่าตัวแปรอิสระสามารถอธิบายความแปรปรวนของตัวแปรตามได้มากน้อยเพียงใด

ข. ค่าทางสถิติ F-Test ใช้ทดสอบว่าสมการถดถอยโดยรวมมีความเหมาะสมหรือไม่

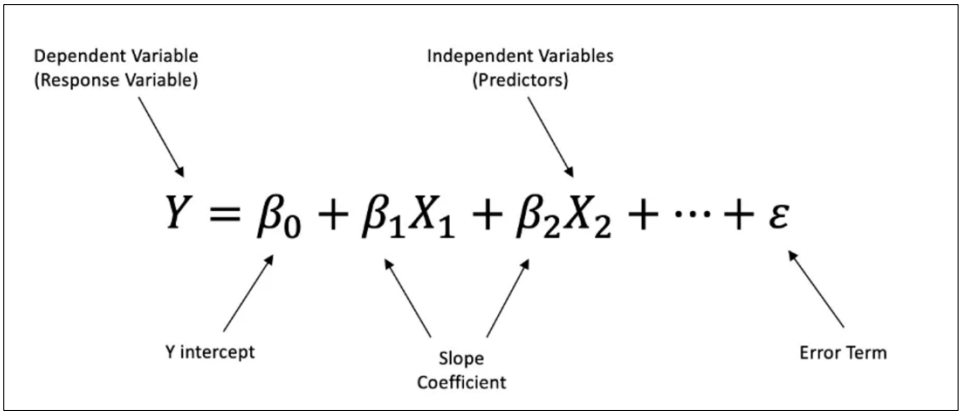
ค. ค่า p-value ของค่าสัมประสิทธิ์แต่ละตัว ใช้ทดสอบว่าสัมประสิทธิ์ของตัวแปรอิสระมีนัยสำคัญทางสถิติหรือไม่

2.2.3.3 ข้อจำกัดของการถดถอยพหุคูณ

ก. ไม่สามารถใช้กับข้อมูลที่มีความสัมพันธ์แบบไม่เป็นเชิงเส้น

ข. มีความอ่อนไหวต่อค่าผิดปกติ (Outliers)

ค. ต้องระมัดระวังปัญหาสหสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity)



ภาพที่ 2.16 ตัวแบบการถดถอยเชิงเส้นพหุคูณ

ที่มา : medium.com

การถดถอยพหุคูณเป็นเครื่องมือทางสถิติที่ช่วยวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรหลายตัว ทำให้สามารถพยากรณ์และทำความเข้าใจปัจจัยที่มีผลต่อผลลัพธ์ได้ อย่างไรก็ตาม การใช้งานต้องพิจารณาสมมติฐานและข้อจำกัดของโมเดลเพื่อให้ได้ผลลัพธ์ที่ถูกต้องและแม่นยำ

2.2.4 หลักในการทำเหมืองข้อมูล

เหมืองข้อมูล (Data Mining) เป็นกระบวนการสกัดข้อมูลเชิงลึกจากชุดข้อมูลขนาดใหญ่ โดยใช้เทคนิคทางสถิติ คณิตศาสตร์ และปัญญาประดิษฐ์ เพื่อค้นหารูปแบบ แนวโน้ม และความสัมพันธ์ที่ซ่อนอยู่ในข้อมูล

2.2.4.1 หลักการพื้นฐานของเหมืองข้อมูล

ก. ความเข้าใจในข้อมูล (Understanding Data) ก่อนเริ่มกระบวนการเหมืองข้อมูล ต้องทำความเข้าใจกับข้อมูลที่มี รวมถึงประเภทของข้อมูล แหล่งที่มา และคุณภาพของข้อมูล

ข. การเตรียมข้อมูล (Data Preparation) ข้อมูลต้องถูกทำความสะอาดและแปลงให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ ซึ่งรวมถึง

- การจัดการค่าที่หายไป (Handling Missing Data)
- การกำจัดค่าผิดปกติ (Outlier Detection and Removal)
- การลดขนาดข้อมูล (Dimensionality Reduction)

ค. การเลือกเทคนิคที่เหมาะสม (Choosing the Right Technique) การเลือกเทคนิคเหมืองข้อมูลที่เหมาะสมขึ้นอยู่กับวัตถุประสงค์ เช่น

- การจำแนกประเภท (Classification) ใช้สำหรับจำแนกข้อมูลออกเป็นกลุ่ม เช่น การพยากรณ์โรคจากข้อมูลผู้ป่วย
- การวิเคราะห์การถดถอย (Regression Analysis) ใช้พยากรณ์ค่าของตัวแปรเป้าหมาย เช่น การทำนายราคาหุ้น
- การจัดกลุ่มข้อมูล (Clustering) ใช้ในการแบ่งกลุ่มข้อมูลที่มีลักษณะคล้ายกัน
- การหากฎความสัมพันธ์ (Association Rule Mining) ใช้ในการค้นหาความสัมพันธ์ระหว่างข้อมูล

ง. การตรวจสอบและประเมินผล (Model Evaluation and Validation) เพื่อให้มั่นใจว่าแบบจำลองมีความแม่นยำ ควรใช้วิธีการตรวจสอบ

- Cross-Validation การแบ่งชุดข้อมูลออกเป็นกลุ่มย่อยเพื่อทดสอบโมเดล
- Confusion Matrix การวิเคราะห์ผลลัพธ์ของการจำแนกประเภท
- Mean Squared Error (MSE) ใช้สำหรับประเมินความคลาดเคลื่อนของโมเดลการพยากรณ์

จ. การนำผลไปใช้ (Deployment and Interpretation) ผลลัพธ์จากเหมืองข้อมูลต้องสามารถนำไปใช้ประโยชน์ได้จริง โดยต้องแปลความหมายของผลลัพธ์ให้สอดคล้องกับบริบททางธุรกิจหรือการวิจัย

ฉ. ข้อจำกัดของเหมืองข้อมูล

- ปัญหาคุณภาพข้อมูล ข้อมูลที่ไม่สมบูรณ์หรือมีค่าผิดปกติอาจทำให้ผลลัพธ์ไม่น่าเชื่อถือ
- ความซับซ้อนของแบบจำลอง แบบจำลองที่ซับซ้อนเกินไปอาจทำให้ยากต่อการตีความและนำไปใช้จริง
- ข้อกฎหมายและจริยธรรม ต้องคำนึงถึงความเป็นส่วนตัวและความปลอดภัยของข้อมูล

หลักในการทำเหมืองข้อมูลประกอบด้วยการทำความเข้าใจข้อมูล การเตรียมข้อมูล การเลือกเทคนิคที่เหมาะสม การประเมินผล และการนำไปใช้งานอย่างถูกต้อง แม้ว่าจะมีข้อจำกัด แต่หากดำเนินการอย่างเป็นระบบก็สามารถช่วยให้ได้ข้อมูลเชิงลึกที่มีคุณค่าและนำไปใช้ประโยชน์ในหลายด้าน เช่น ธุรกิจ การแพทย์ และวิทยาศาสตร์ข้อมูล

2.2.5 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์

เว็บไซต์ (website, web site, หรือ web site) หมายถึง หน้าเว็บเพจหลายหน้าซึ่งเชื่อมโยงกันผ่านทางไฮเปอร์ลิงก์ ส่วนใหญ่จัดทำขึ้นเพื่อนำเสนอข้อมูลผ่านคอมพิวเตอร์โดยถูกจัดเก็บไว้ในเวิลด์ไวด์เว็บ หน้าแรกของเว็บไซต์ที่เก็บไว้ที่ชื่อหลักจะเรียกว่าโฮมเพจ เว็บไซต์โดยทั่วไปจะให้บริการต่อผู้ใช้ฟรี แต่ในขณะเดียวกันบางเว็บไซต์จำเป็นต้องมีการสมัครสมาชิกและเสียค่าบริการเพื่อที่จะดูข้อมูลในเว็บไซต์นั้น ซึ่งได้แก่ข้อมูลทางวิชาการ ข้อมูลตลาดหลักทรัพย์ หรือข้อมูลสื่อต่างๆ ผู้ทำเว็บไซต์มีหลากหลายระดับ ตั้งแต่สร้างเว็บไซต์ส่วนตัวจนถึงระดับเว็บไซต์สำหรับนักธุรกิจหรือองค์กรต่างๆ การเรียกดูเว็บไซต์โดยทั่วไปนิยมเรียกดูผ่านซอฟต์แวร์ในลักษณะของ เว็บเบราเซอร์

2.2.5.1 องค์ประกอบของการออกแบบเว็บไซต์

ก. ความเรียบง่าย (Simplicity)หมายถึง การจำกัดองค์ประกอบเสริมให้เหลือเฉพาะองค์ประกอบหลัก กล่าวคือในการสื่อสารเนื้อหากับผู้ใช้นั้น เราต้องเลือกเสนอสิ่งที่เราต้องการนำเสนอออกมาในส่วนของกราฟิก สีสัน ตัวอักษรและภาพเคลื่อนไหวต้องเลือกให้พอเหมาะ ถ้าหากมีมากเกินไปจะรบกวนสายตาและสร้างความคำราญต่อผู้ใช้ตัวอย่างเว็บไซต์ที่ได้รับการออกแบบที่ดี ได้แก่ เว็บไซต์ของบริษัทใหญ่ ๆ อย่างเช่น Apple Adobe Microsoft หรือ Kokia ที่มีการออกแบบเว็บไซต์ในรูปแบบที่เรียบง่าย ไม่ซับซ้อน และใช้งานอย่างสะดวก

ข. ความสม่ำเสมอ (Consistency) หมายถึง การสร้างความสม่ำเสมอให้เกิดขึ้นตลอดทั้งเว็บไซต์ โดยอาจเลือกใช้รูปแบบเดียวกันตลอดทั้งเว็บไซต์ก็ได้ เพราะถ้าหากว่าแต่ละหน้าในเว็บไซต์นั้นมีความแตกต่างกันมากจนเกินไป อาจทำให้ผู้ใช้เกิดความสับสนและไม่แน่ใจว่ากำลังอยู่ในเว็บไซต์เดิมหรือไม่ เพราะฉะนั้นการออกแบบเว็บไซต์ในแต่ละหน้าควรที่จะมี

รูปแบบ สไตล์ของกราฟิก ระบบเนวิเกชั่น (Navigation) และโทนสีที่มีความคล้ายคลึงกันตลอดทั้งเว็บไซต์

ค. ความเป็นเอกลักษณ์ (Identity) ในการออกแบบเว็บไซต์ต้องคำนึงถึงลักษณะขององค์กรเป็นหลัก เนื่องจากเว็บไซต์จะสะท้อนถึงเอกลักษณ์และลักษณะขององค์กร การเลือกใช้ตัวอักษร ชุดสี รูปภาพหรือกราฟิก จะมีผลต่อรูปแบบของเว็บไซต์เป็นอย่างมาก ตัวอย่างเช่น ถ้าเราต้องออกแบบเว็บไซต์ของธนาคารแต่เรากลับเลือกสีสันและกราฟิกมากมาย อาจทำให้ผู้ใช้คิดว่าเป็นเว็บไซต์ของสวนสนุกซึ่งส่งผลต่อความเชื่อถือขององค์กรได้

ง. เนื้อหา (Useful Content) ถือเป็นสิ่งสำคัญที่สุดในเว็บไซต์ เนื้อหาในเว็บไซต์ต้องสมบูรณ์และได้รับการปรับปรุงพัฒนาให้ทันสมัยอยู่เสมอ ผู้พัฒนาต้องเตรียมข้อมูลและเนื้อหาที่ผู้ใช้ต้องการให้ถูกต้องและสมบูรณ์ เนื้อหาที่สำคัญที่สุดคือเนื้อหาที่ทีมผู้พัฒนาสร้างสรรค์ขึ้นมาเอง และไม่ไปซ้ำกับเว็บอื่น เพราะจะถือเป็นสิ่งที่ดึงดูดผู้ใช้ให้เข้ามาเว็บไซต์ได้เสมอ แต่ถ้าเป็นเว็บที่ลิงค์ข้อมูลจากเว็บอื่น ๆ มาเมื่อใดก็ตามที่ผู้ใช้ทราบว่า ข้อมูลนั้นมาจากเว็บใด ผู้ใช้ก็ไม่จำเป็นต้องกลับมาใช้งานลิงค์เหล่านั้นอีก

จ. ระบบเนวิเกชั่น (User–Friendly Navigation) เป็นส่วนประกอบที่มีความสำคัญต่อเว็บไซต์มาก เพราะจะช่วยไม่ให้ผู้ใช้เกิดความสับสนระหว่างดูเว็บไซต์ ระบบเนวิเกชั่นจึงเปรียบเสมือนป้ายบอกทาง ดังนั้นการออกแบบเนวิเกชั่น จึงควรให้เข้าใจง่าย ใช้งานได้สะดวก ถ้ามีการใช้กราฟิกก็ควรสื่อความหมาย ตำแหน่งของการวางเนวิเกชั่นก็ควรวางให้สม่ำเสมอ เช่น อยู่ตำแหน่งบนสุดของทุกหน้าเป็นต้น ซึ่งถ้าจะให้ดีเมื่อมีเนวิเกชั่นที่เป็นกราฟิกก็ควรเพิ่มระบบเนวิเกชั่นที่เป็นตัวอักษรไว้ส่วนล่างด้วย เพื่อช่วยอำนวยความสะดวกให้กับผู้ใช้ที่ยกเลิกการแสดงผลภาพกราฟิกบนเว็บเบราเซอร์

ฉ. คุณภาพของสิ่งที่ปรากฏให้เห็นในเว็บไซต์ (Visual Appeal) ลักษณะที่น่าสนใจของเว็บไซต์นั้น ขึ้นอยู่กับความชอบส่วนบุคคลเป็นสำคัญ แต่โดยรวมแล้วก็สามารถสรุปได้ว่าเว็บไซต์ที่น่าสนใจนั้นส่วนประกอบต่าง ๆ ควรมีคุณภาพ เช่น กราฟิกควรสมบูรณ์ไม่มีรอยหรือขอบขั้นบันได้ให้เห็น ชนิดตัวอักษร

อ่านง่ายสบายตา มีการเลือกใช้โทนสีที่เข้ากันอย่างสวยงาม เป็นต้น

ช. ความสะดวกของการใช้ในสภาพต่าง ๆ (Compatibility) การใช้งานของเว็บไซต์นั้นไม่ควรมีขอบจำกัด กล่าวคือ ต้องสามารถใช้งานได้ดีในสภาพแวดล้อมที่หลากหลาย ไม่มีการบังคับให้ผู้ใช้ต้องติดตั้งโปรแกรมอื่นใดเพิ่มเติม นอกเหนือจากเว็บบราวเซอร์ ควรเป็นเว็บที่แสดงผลได้ดีในทุกระบบปฏิบัติการ สามารถแสดงผลได้ในทุกความละเอียดหน้าจอ ซึ่งหากเป็นเว็บไซต์ที่มีผู้ใช้บริการมากและกลุ่มเป้าหมายหลากหลายควรให้ความสำคัญกับเรื่องนี้ให้มาก

ซ. ความคงที่ในการออกแบบ (Design Stability) ถ้าต้องการให้ผู้ใช้งานรู้สึกว่าเว็บไซต์มีคุณภาพ ถูกต้อง และเชื่อถือได้ ควรให้ความสำคัญกับการออกแบบเว็บไซต์เป็นอย่างมาก ต้องออกแบบวางแผนและเรียบเรียงเนื้อหาอย่างรอบคอบ ถ้าเว็บที่จัดทำขึ้นอย่างลวก ๆ ไม่มีมาตรฐานการออกแบบและระบบการจัดการข้อมูล ถ้ามีปัญหามากขึ้นอาจส่งผลให้เกิดปัญหาและทำให้ผู้ใช้หมดความเชื่อถือ

ฌ. ความคงที่ของการทำงาน (Function Stability) ระบบการทำงานต่าง ๆ ในเว็บไซต์ควรมีความถูกต้องแน่นอน ซึ่งต้องได้รับการออกแบบสร้างสรรค์และตรวจสอบอยู่เสมอ ตัวอย่างเช่น ลิงค์ต่าง ๆ ในเว็บไซต์ ต้องตรวจสอบว่ายังสามารถลิงค์ข้อมูลได้ถูกต้องหรือไม่ เพราะเว็บไซต์อื่นอาจมีการเปลี่ยนแปลงได้ตลอดเวลา ปัญหาที่เกิดจากลิงค์ ก็คือ ลิงค์ขาด ซึ่งพบได้บ่อยเป็นปัญหาที่สร้างความรำคาญกับผู้ใช้เป็นอย่างมาก

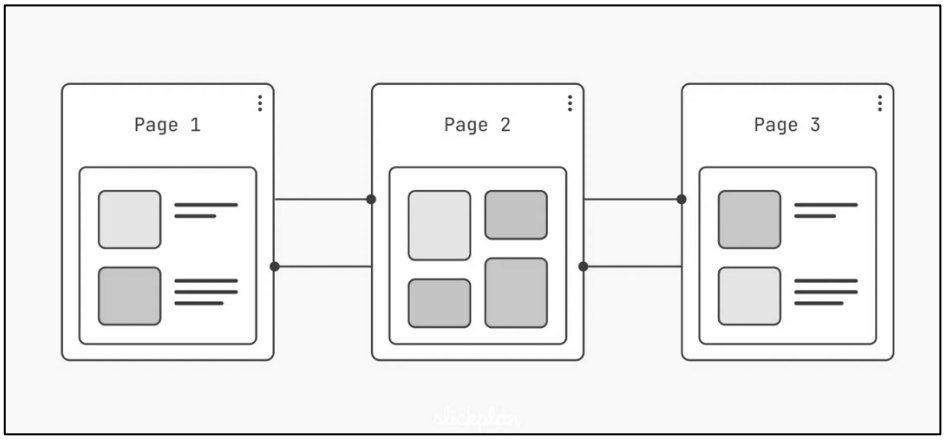
2.2.5.2 การออกแบบเว็บไซต์

ในการออกแบบเว็บไซต์นั้นประกอบด้วยกระบวนการต่าง ๆ มากมาย เช่น การออกแบบโครงสร้าง ลักษณะหน้าตา หรือการเขียนโปรแกรม แต่มีหลายคนที่พัฒนาเว็บไซต์ โดยขาดการวางแผนและทำงานไม่เป็นระบบ ตัวอย่างเช่น การลงมือออกแบบโดยการใช้โปรแกรมช่วยสร้างเว็บ เนื้อหาและรูปแบบก็เป็นไปตามที่นึกขึ้นได้ขณะนั้น และเมื่อเห็นว่าดูดีแล้วก็เปิดตัวเลย ทำให้เว็บนั้นมีเป้าหมายและแนวทางที่ไม่แน่นอน ผลลัพธ์ที่ได้จึงเสี่ยงกับความล้มเหลวค่อนข้างมาก ความล้มเหลวที่พบเห็นได้ทั่วไป ได้แก่ เว็บที่แสดงข้อความว่าอยู่ระหว่างการก่อสร้าง (Under Construction หรือ Coming soon) ซึ่งแสดงให้เห็นถึงการขาดการวางแผนที่ดีบางเว็บถือได้ว่าตายไปแล้ว เนื่องจากข้อมูลไม่ทันสมัย ขาดการพัฒนาปรับปรุงเทคโนโลยีล้าสมัย ลิงค์ผิดพลาด สิ่งเหล่านี้แสดงให้เห็นถึงการขาดการดูแล ตรวจสอบและพัฒนาให้

ทันสมัยอยู่เสมอ การออกแบบเว็บไซต์อย่างถูกต้องจะช่วยลดความผิดพลาดเหล่านี้ และช่วยลดความเสี่ยงที่จะทำให้เว็บประสบความล้มเหลว การออกแบบเว็บไซต์ที่ดีต้องอาศัยการออกแบบและจัดระบบข้อมูลอย่างเหมาะสม กระบวนการแรกของการออกแบบเว็บไซต์คือการกำหนดเป้าหมายของเว็บไซต์กำหนดกลุ่มผู้ใช้ ซึ่งการจะให้ได้มาซึ่งข้อมูล ผู้พัฒนาต้องเรียนรู้ผู้ใช้ หรือจำลองสถานการณ์ สิ่งเหล่านี้จะช่วยให้เราสามารถออกแบบเนื้อหาและการใช้งานเว็บไซต์ได้อย่างเหมาะสม ตรงกับความต้องการของผู้ใช้อย่างแท้จริง

2.2.5.3 โครงสร้างของเว็บไซต์

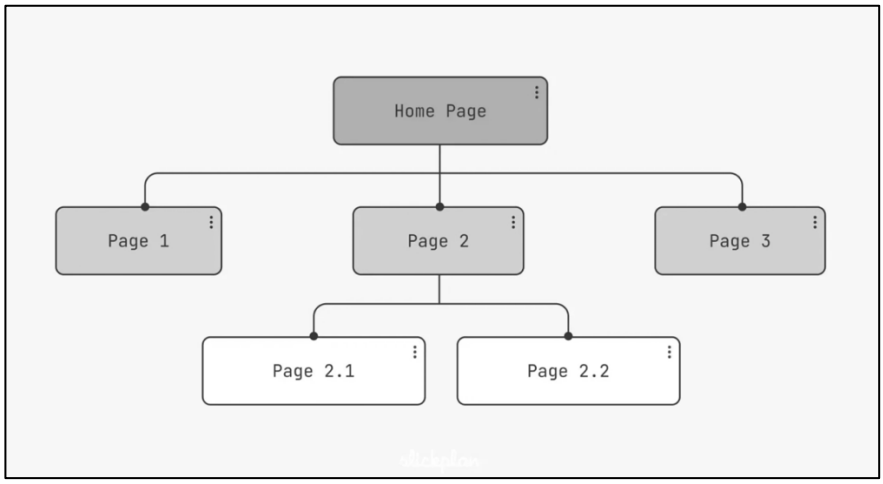
Sequential Structure หรือ Linear Structure เป็นโครงสร้างเว็บไซต์แบบเส้นตรง ซึ่งเป็นรูปแบบที่เรียบง่ายที่สุด เหมาะสำหรับเว็บไซต์ที่ต้องการนำเสนอข้อมูลเป็นลำดับขั้นตอน เช่น คอร์สเรียนออนไลน์หรือ e-book โดยผู้ใช้จะต้องเข้าชมเนื้อหาตามลำดับที่กำหนดไว้ผ่านปุ่ม “ถัดไป” หรือ “ย้อนกลับ” ข้อดีคือผู้ใช้สามารถเรียนรู้เนื้อหาได้อย่างเป็นระบบ แต่อาจไม่ยืดหยุ่นสำหรับการค้นหาข้อมูลเฉพาะ



ภาพที่ 2.17 Sequential Structure

ที่มา : Slickplan.com

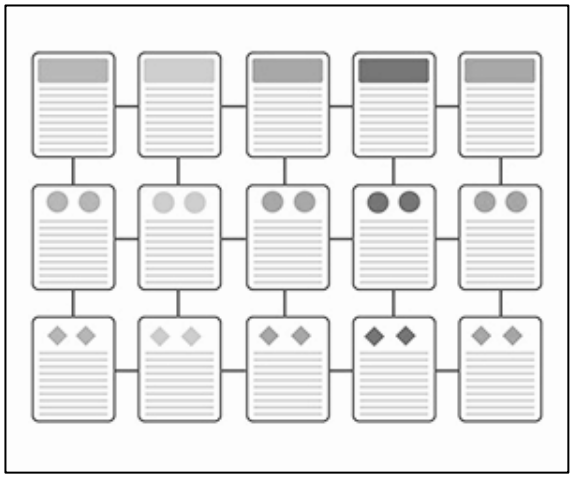
Hierarchical Structure คือโครงสร้างเว็บไซต์แบบลำดับชั้น รูปแบบนี้ได้รับความนิยมมากที่สุดเพราะเหมาะกับเว็บไซต์ทุกขนาด ตั้งแต่เว็บเล็ก ๆ ไปจนถึง e-commerce ขนาดใหญ่ ลักษณะจะคล้ายแผนผังต้นไม้ที่แตกกิ่งก้านจากหน้าหลักลงไปเป็นหมวดหมู่ย่อยต่าง ๆ ทำให้จัดการข้อมูลได้เป็นระบบและ Google สามารถเข้าใจโครงสร้างเว็บไซต์ได้ง่าย



ภาพที่ 2.18 Hierarchical Structure

ที่มา : Slickplan.com

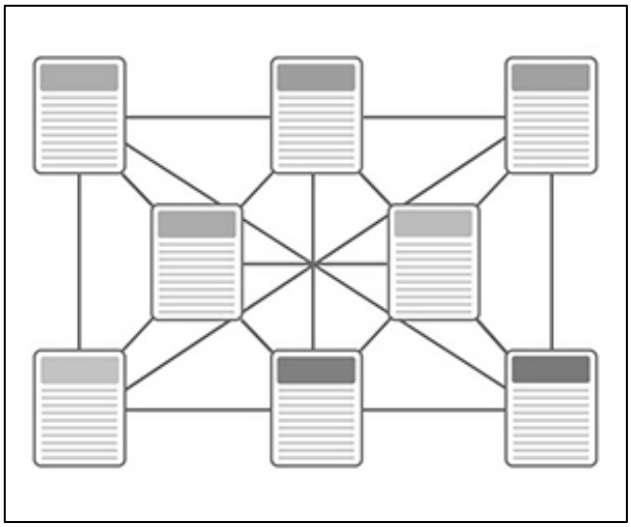
Grid Structure เป็นโครงสร้างเว็บไซต์แบบตาราง ข้อดีคือมีความยืดหยุ่นสูง เหมาะสำหรับเว็บไซต์ที่มีเนื้อหาหลากหลายและต้องการให้ผู้ใช้สามารถเข้าถึงข้อมูลได้หลายทาง เช่น เว็บไซต์ข่าวหรือบล็อก แต่ละหมวดหมู่สามารถเชื่อมโยงถึงกันได้อย่างอิสระ ทำให้ผู้ใช้สามารถค้นหาข้อมูลที่สนใจได้หลายวิธี



ภาพที่ 2.19 Grid Structure

ที่มา : enfete.co.th

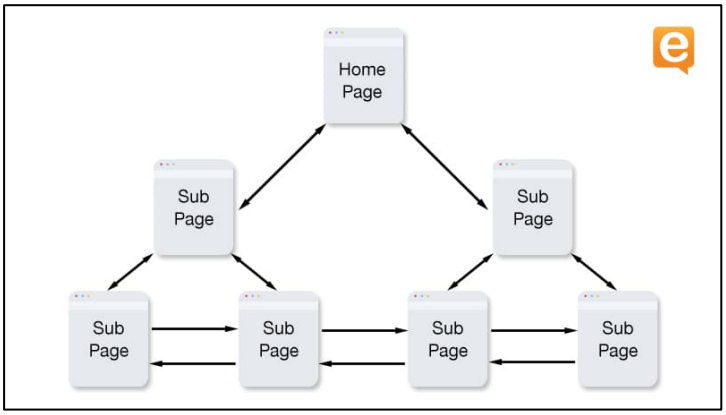
Web Structure เป็นโครงสร้างเว็บไซต์แบบใยแมงมุมให้อิสระในการเชื่อมโยงเนื้อหามากที่สุด ทุกหน้าสามารถเชื่อมถึงกันได้โดยตรง เหมาะกับเว็บไซต์ที่ต้องการความยืดหยุ่นสูงในการนำทาง อย่างไรก็ตามหากมีหน้าเว็บจำนวนมาก อาจทำให้ผู้ใช้สับสนและ Google อาจเข้าใจโครงสร้างเว็บไซต์ได้ยาก



ภาพที่ 2.20 Web Structure

ที่มา : enfete.co.th

Hybrid Structure คือโครงสร้างเว็บไซต์แบบผสม โดยจะเป็นการผสมผสานจุดเด่นของโครงสร้างหลายรูปแบบเข้าด้วยกัน มักใช้โครงสร้างแบบลำดับชั้น (Hierarchical) เป็นพื้นฐาน แล้วเพิ่มการเชื่อมโยงรูปแบบอื่นตามความเหมาะสม เช่น การใช้โครงสร้างแบบเส้นตรงสำหรับส่วนที่ต้องการนำเสนอข้อมูลเป็นลำดับขั้น หรือเพิ่มการเชื่อมโยงแบบใยแมงมุมในบางส่วนเพื่อให้ผู้ใช้เข้าถึงข้อมูลได้สะดวกขึ้น โครงสร้างแบบผสมนี้เหมาะสำหรับเว็บไซต์ขนาดใหญ่ที่มีเนื้อหาหลากหลายในการนำเสนอข้อมูล

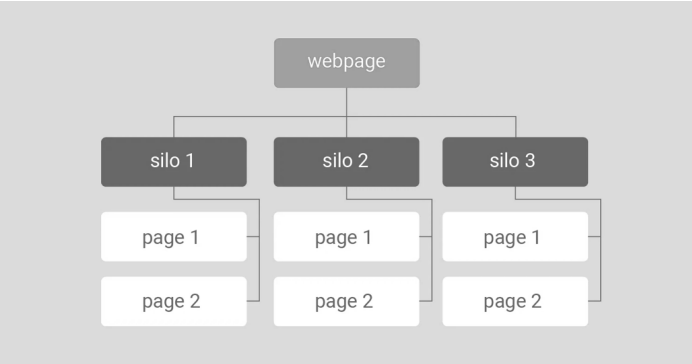


ภาพที่ 2.21 Hybrid Structure

ที่มา : ehowme.com

Silo Structure เป็นโครงสร้างเว็บไซต์แบบไซโล ที่ออกแบบมาเพื่อเพิ่มประสิทธิภาพการทำ SEO โดยเฉพาะ หลักการสำคัญคือการจัดกลุ่มเนื้อหาที่เกี่ยวข้องให้อยู่ในหมวดหมู่เดียวกันอย่างชัดเจน และสร้าง Internal Link เฉพาะระหว่างเนื้อหาในกลุ่มเดียวกันเท่านั้น วิธีนี้ช่วยให้ Google เข้าใจความเชื่อมโยงของเนื้อหาได้ดีขึ้น ซึ่งสอดคล้องกับหลักการของกลยุทธ์ Semantic SEO ที่เน้นการสร้างความสัมพันธ์ของเนื้อหาอย่างมีความหมาย ส่งผลให้เว็บไซต์มีโอกาสติดอันดับสูงในคำค้นหาที่เฉพาะเจาะจงมากขึ้น การเลือกใช้โครงสร้างแบบ

ไซโลนี้เหมาะสำหรับเว็บไซต์ที่มีเนื้อหาเชิงลึกในแต่ละหมวดหมู่ เช่น เว็บไซต์ข้อมูลเฉพาะทาง บล็อกความรู้ หรือเว็บไซต์ e-commerce ที่แบ่งสินค้าเป็นหมวดหมู่ชัดเจน อย่างไรก็ตาม การทำ Silo Structure ต้องวางแผนอย่างรอบคอบตั้งแต่เริ่มต้น เพราะการปรับเปลี่ยนโครงสร้างในภายหลังอาจส่งผลกระทบต่ออันดับ SEO ได้



ภาพที่ 2.22 Silo Structure

ที่มา : Slickplan.com

2.2.5.4 องค์ประกอบของโครงสร้างเว็บไซต์

ก. หน้าหลัก (Homepage) เป็นหน้าแรกที่ผู้ใช้จะพบเมื่อเข้าสู่เว็บไซต์ แสดงภาพรวมของเว็บไซต์และจุดเด่นที่สำคัญ พร้อมกับมีเมนูนำทางไปยังส่วนต่าง ๆ ของเว็บไซต์

ข. 2.2.5.4.2 เมนูนำทาง (Navigation Menu) เป็นส่วนที่ช่วยให้ผู้ใช้สามารถเข้าถึงเนื้อหาต่างๆ ภายในเว็บไซต์ได้อย่างสะดวก โดยควรมีการจัดหมวดหมู่ที่ชัดเจน เข้าใจง่าย และรองรับการแสดงผลบนทุกอุปกรณ์

ค. หน้าเนื้อหา (Content Pages) เป็นส่วนที่นำเสนอข้อมูลหลักของเว็บไซต์ โดยมีการแบ่งเป็นหมวดหมู่ตามประเภทของเนื้อหา และมีการเชื่อมโยงระหว่างหน้าที่มีความเกี่ยวข้องกัน

ง. ส่วนท้ายเว็บ (Footer) เป็นพื้นที่สำหรับรวบรวมลิงก์สำคัญ ข้อมูลการติดต่อ ลิขสิทธิ์ นโยบายต่าง ๆ และอาจมีแผนผังเว็บไซต์แบบย่อเพื่อช่วยในการนำทาง

จ. การเชื่อมโยงภายใน (Internal Link) คือการสร้างลิงก์เชื่อมต่อระหว่างหน้าต่างๆ ภายในเว็บไซต์ ช่วยในการกระจายค่าความน่าเชื่อถือและสร้างเส้นทางการนำทางที่เป็นธรรมชาติสำหรับผู้ใช้

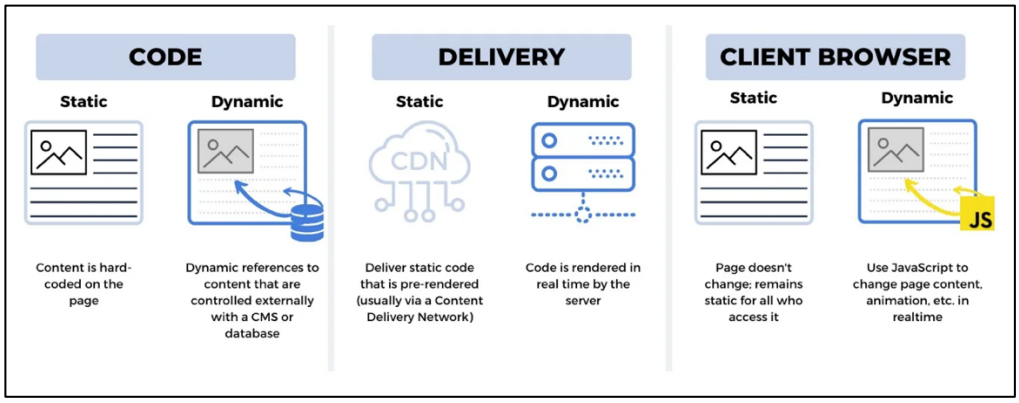
ฉ. แผนผังเว็บไซต์ (Sitemap) เป็นเอกสารที่แสดงโครงสร้างทั้งหมดของเว็บไซต์ มีทั้งรูปแบบ HTML สำหรับผู้ใช้งานทั่วไป

และ XML สำหรับเครื่องมือค้นหา ช่วยให้ทั้งผู้ใช้และ Googlebot เข้าใจโครงสร้างเว็บไซต์

2.2.5.5 รูปแบบของเว็บไซต์

ก. Static Website เป็นเว็บไซต์ที่สร้างขึ้นด้วยภาษา HTML พื้นฐาน เหมาะสำหรับเว็บไซต์ขนาดเล็กที่มีจำนวนหน้าไม่มาก และไม่ต้องการปรับเปลี่ยนเนื้อหาบ่อย เช่น เว็บไซต์แนะนำองค์กร เว็บไซต์ประวัติส่วนตัว หรือเว็บไซต์ Portfolio ข้อมูลถูกจัดเก็บในรูปแบบไฟล์ .html แยกเป็นแต่ละหน้า การแก้ไขเนื้อหาจำเป็นต้องเข้าไปแก้ไขโค้ด HTML โดยตรง

ข. Dynamic Website เป็นเว็บไซต์ที่ออกแบบมาเพื่อรองรับการเปลี่ยนแปลงเนื้อหาอย่างสม่ำเสมอ มีระบบจัดการเนื้อหา (CMS) ที่ช่วยให้ผู้ดูแลสามารถปรับปรุงข้อมูลได้สะดวกผ่านหน้าจัดการระบบ โดยไม่ต้องแก้ไขโค้ด HTML เหมาะสำหรับเว็บไซต์ที่มีเนื้อหาจำนวนมาก เช่น เว็บข่าว บล็อก หรือเว็บ E-commerce เนื้อหาจะถูกจัดเก็บในฐานข้อมูลและดึงมาแสดงผลตามความต้องการ



ภาพที่ 2.23 รูปแบบของเว็บไซต์

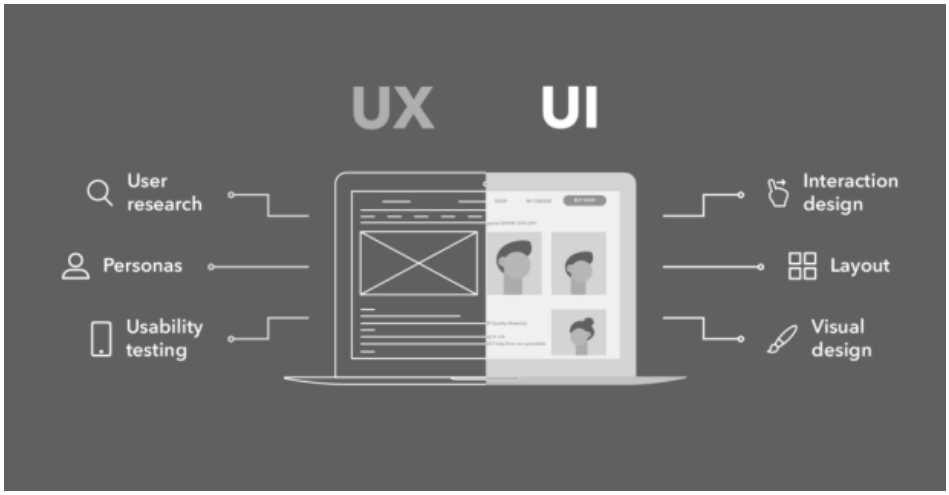
ที่มา : Zesty.com

2.2.5.6 ทฤษฎีด้านประสบการณ์ผู้ใช้ (User Experience : UX) เป็นศาสตร์ที่เกี่ยวข้องกับการออกแบบเว็บไซต์ให้ตอบสนองความต้องการของผู้ใช้ ซึ่งมีหลักสำคัญ ได้แก่

ก. การทดสอบการใช้งาน (Usability Testing) ตรวจสอบว่าเว็บไซต์ใช้งานง่ายหรือไม่

ข. การออกแบบที่ยึดผู้ใช้เป็นศูนย์กลาง (User Centered Design : UCD) พัฒนาเว็บไซต์โดยคำนึงถึงประสบการณ์ของผู้ใช้เป็นหลัก

ค. การออกแบบอินเทอร์เฟซผู้ใช้ (User Interface Design : UI Design) การออกแบบองค์ประกอบต่าง ๆ ให้ใช้งานง่ายและสวยงาม



ภาพที่ 2.24 ทฤษฎีด้านประสบการณ์ผู้ใช้
ที่มา : marketeeronline.co

การสร้างเว็บไซต์เป็นกระบวนการที่ต้องอาศัยทฤษฎีและแนวคิดจากหลายสาขาวิชา ตั้งแต่การออกแบบและพัฒนาไปจนถึงการจัดการฐานข้อมูลและการรักษาความปลอดภัย ความเข้าใจในหลักการเหล่านี้จะช่วยให้สามารถพัฒนาเว็บไซต์ที่มีประสิทธิภาพ ใช้งานง่าย และตอบสนองความต้องการของผู้ใช้ได้เป็นอย่างดี

2.2.6 ทฤษฎีเกี่ยวกับปัจจัยหลักที่มีผลต่อความล่าช้าของสายการบิน

ความล่าช้าของสายการบิน (Flight Delay) เป็นปัญหาสำคัญในอุตสาหกรรมการบิน ซึ่งอาจเกิดจากปัจจัยหลายประการที่ส่งผลต่อการดำเนินงานของเที่ยวบิน โดยสามารถแบ่งออกเป็นปัจจัยหลักดังนี้

2.2.6.1 ปัจจัยด้านสภาพอากาศ (Weather Factors) สภาพอากาศเป็นปัจจัยสำคัญที่มีผลต่อความล่าช้าของเที่ยวบิน โดยเฉพาะในกรณีที่เกิดสภาพอากาศแปรปรวน เช่น หมอกหนา (Fog) ส่งผลต่อทัศนวิสัยในการบินและการลงจอด พายุฝนฟ้าคะนอง (Thunderstorms) ทำให้ต้องเปลี่ยนเส้นทางบินหรือรอให้สภาพอากาศดีขึ้น หิมะและน้ำแข็ง (Snow and Ice) ทำให้ต้องใช้เวลาในการกำจัดน้ำแข็งจากปีกเครื่องบิน ลมแรง (Strong Winds) อาจทำให้เครื่องบินต้องรอคิวในการขึ้นหรือลงจอด

2.2.6.2 ปัจจัยด้านการจราจรทางอากาศ (Air Traffic Factors) ความหนาแน่นของเที่ยวบินในบางสนามบินอาจทำให้เกิดความล่าช้า เช่น ความแออัดของสนามบิน (Airport Congestion) ความล่าช้าในการอนุญาตให้บินขึ้นหรือลงจอด การควบคุมการจราจรทางอากาศ (Air Traffic Control) อาจมีข้อจำกัดในการบริหารเที่ยวบิน

2.2.6.3 ปัจจัยด้านเครื่องบินและการบำรุงรักษา (Aircraft and Maintenance Factors) เครื่องบินแต่ละลำต้องได้รับการตรวจสอบและบำรุงรักษาก่อนออกเดินทาง หากพบ

ปัญหาทางเทคนิค อาจต้องใช้เวลาแก้ไข ส่งผลให้เที่ยวบินล่าช้า ปัจจัยที่เกี่ยวข้อง ได้แก่ ปัญหาทางเทคนิคของเครื่องบิน (Mechanical Issues) การซ่อมบำรุงตามรอบ (Scheduled Maintenance) การขาดแคลนอะไหล่หรืออุปกรณ์สำรอง

2.2.6.4 ปัจจัยด้านสายการบิน (Airline Operational Factors) การบริหารจัดการของสายการบินเองอาจส่งผลต่อความล่าช้า เช่น การหมุนเวียนของเครื่องบิน (Aircraft Turnaround Time) การจัดตารางเที่ยวบินที่ไม่สอดคล้องกับความสามารถในการรองรับของสนามบิน ปัญหาการบริหารลูกเรือ (Crew Scheduling)

2.2.6.5 ปัจจัยด้านผู้โดยสาร (Passenger-Related Factors) ผู้โดยสารสามารถส่งผลต่อความล่าช้าของเที่ยวบิน เช่น การเช็กอินและตรวจสอบสัมภาระล่าช้า ผู้โดยสารมาขึ้นเครื่องไม่ตรงเวลา กระบวนการรักษาความปลอดภัยที่ใช้เวลานานขึ้นในบางกรณี

2.2.6.5 ปัจจัยด้านกฎระเบียบและข้อบังคับ (Regulatory and Security Factors) มาตรการด้านความปลอดภัยและกฎระเบียบของหน่วยงานการบินพลเรือนอาจทำให้เที่ยวบินล่าช้า เช่น การตรวจสอบความปลอดภัยที่เข้มงวด การปฏิบัติตามกฎระเบียบของประเทศปลายทาง

2.2.7 ทฤษฎีการสุ่มตัวอย่าง (Sampling)

การสุ่มตัวอย่างข้อมูล (Sampling) ไว้ว่าเป็นหนึ่งในเทคนิคการลดขนาดข้อมูล (Data Reduction) ที่มีบทบาทสำคัญในกระบวนการทำเหมืองข้อมูล โดยมีวัตถุประสงค์เพื่อลดปริมาณข้อมูลที่ต้องประมวลผล ขณะเดียวกันยังคงรักษาคุณลักษณะสำคัญและความเป็นตัวแทนของข้อมูลต้นฉบับให้มากที่สุด การสุ่มตัวอย่างเป็นวิธีที่ช่วยให้การประมวลผลรวดเร็วขึ้น ใช้ทรัพยากรระบบอย่างมีประสิทธิภาพ และสามารถใช้ในการสร้างแบบจำลองหรือทดสอบโมเดลได้แม้เมื่อข้อมูลต้นฉบับมีขนาดใหญ่มาก วิธีการสุ่มตัวอย่างที่สำคัญประกอบด้วย การสุ่มแบบง่าย (Simple Random Sampling) ซึ่งเลือกตัวอย่างจากประชากรข้อมูลโดยไม่พิจารณาโครงสร้างข้อมูล, การสุ่มแบบแบ่งชั้น (Stratified Sampling) ที่แบ่งข้อมูลออกเป็นกลุ่มย่อย (strata) ตามคุณลักษณะบางประการ แล้วสุ่มจากแต่ละกลุ่มเพื่อคงสัดส่วนให้ใกล้เคียงกับข้อมูลเดิม และการสุ่มแบบกลุ่ม (Cluster Sampling) ที่สุ่มทั้งกลุ่มข้อมูลเข้ามาวิเคราะห์ Han และคณะเน้นว่าการเลือกวิธีสุ่มต้องพิจารณาความสมดุลระหว่างความถูกต้องของการเป็นตัวแทนข้อมูลกับข้อจำกัดด้านทรัพยากรคอมพิวเตอร์ (Jiawei Han, 2012)

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

2.3.1 กระบวนการวิเคราะห์ข้อมูลด้วย (CRISP-DM)

CRISP-DM (Cross-Industry Standard Process for Data Mining) เป็นมาตรฐานที่ใช้ในกระบวนการวิเคราะห์ข้อมูลและทำเหมืองข้อมูล ซึ่งถูกพัฒนาเพื่อให้สามารถใช้ได้กับอุตสาหกรรมที่หลากหลาย โดย CRISP-DM แบ่งกระบวนการออกเป็น 6 ขั้นตอนหลัก ดังนี้

CRISP-DM (Cross-industry standard process for data mining) หมายถึง กระบวนการมาตรฐานที่ใช้สำหรับการทำเหมืองข้อมูล เพื่อทำการวิเคราะห์และนำไปใช้

ประโยชน์ทางธุรกิจ ประกอบด้วย 6 ขั้นตอน แต่ละขั้นตอนในรูปจะเป็นขั้นตอนที่ต่อเนื่องกัน นั่นคือขั้นตอนถัดไปจะรอผลลัพธ์จากขั้นตอนก่อนหน้า ซึ่งแสดงด้วยลูกศรที่เชื่อมระหว่างกล่องสี่เหลี่ยมแต่ละกล่อง ตัวอย่างเช่น เมื่อได้ผลลัพธ์จากขั้นตอนการเตรียมข้อมูล (Data Preparation) แล้วจะนำไปสร้างโมเดลจำแนกประเภทข้อมูลในขั้น Modeling และหลังจากนั้นอาจจะย้อนกลับมาเปลี่ยนแปลงข้อมูลให้ถูกต้องมากขึ้นเพื่อหวังว่าจะได้โมเดลที่ให้ความถูกต้องมากขึ้น

2.3.1.1 ขั้นตอนการทำกระบวนการวิเคราะห์ข้อมูลด้วย (CRISP-DM)

ก. การทำความเข้าใจธุรกิจ (Business Understanding) มุ่งไปที่การทำความเข้าใจธุรกิจ ปัญหาและวัตถุประสงค์ของโครงการจากมุมมองทางธุรกิจ จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

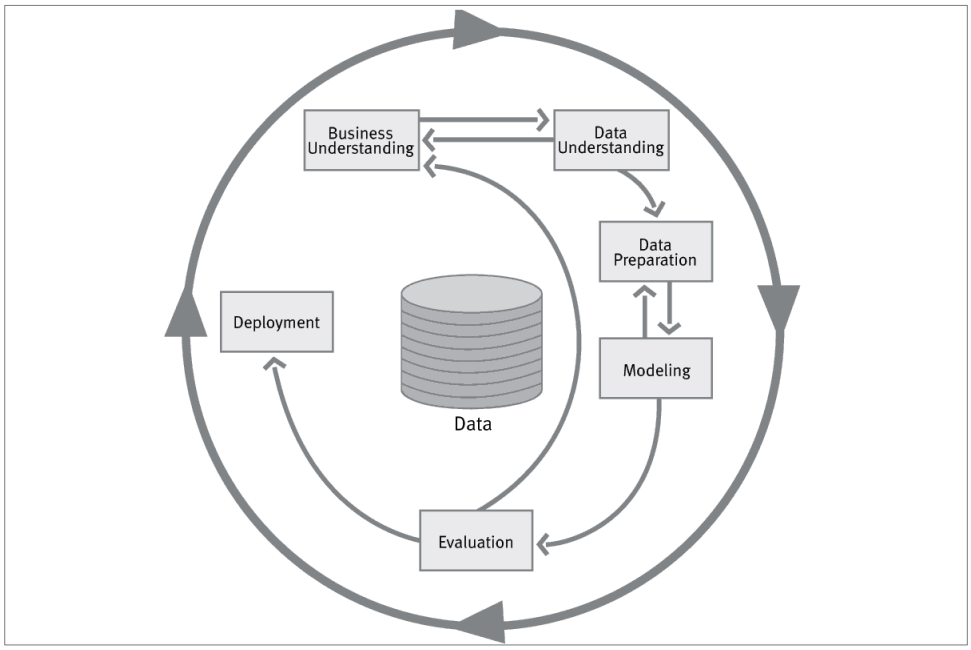
ข. การทำความเข้าใจข้อมูล (Data Understanding) เริ่มต้นด้วยการรวบรวมข้อมูล จากนั้นทำความเข้าใจ ตรวจสอบคุณภาพและเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์

ค. การเตรียมข้อมูล (Data Preparation) ขั้นตอนการเตรียมข้อมูลหมายถึง ขั้นตอนทั้งหมดที่จะทำเพื่อให้ข้อมูลดิบที่เรารวบรวมมา กลายเป็นข้อมูลสมบูรณ์ที่พร้อมจะเข้าสู่โมเดลในขั้นตอนที่ 4 เช่น การสร้างตาราง การลบข้อมูลที่ไม่ต้องการออก การแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ

ง. การสร้างโมเดล (Modeling) เลือกและทดสองสร้างโมเดลหลายๆแบบที่น่าจะสามารถแก้ไขปัญหาที่ต้องการได้ จากนั้นค่อยๆปรับค่าพารามิเตอร์ในแต่ละโมเดล เพื่อให้ได้โมเดลที่เหมาะสมที่สุดมาใช้ในการแก้ไขปัญหา

จ. การวัดประสิทธิภาพของโมเดล (Evaluation) ทำการวัดประสิทธิภาพของโมเดลที่ได้จากขั้นตอนที่ 4 เพื่อวัดว่าโมเดลมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานแล้วหรือไม่ ซึ่งโมเดลแต่ละประเภทก็จะมีตัววัดประสิทธิภาพที่แตกต่างกันออกไป

ฉ. นำโมเดลไปใช้งานจริงในกระบวนการทางธุรกิจ ซึ่งอาจอยู่ในรูปแบบของระบบอัตโนมัติ รายงานเชิงวิเคราะห์ หรือการสร้าง API เพื่อใช้งานร่วมกับระบบอื่น เช่น นำโมเดลไปใช้งานในสภาพแวดล้อมจริง เฝ้าติดตามผลลัพธ์และปรับปรุงโมเดล



ภาพที่ 2.25 CRISP–DM model

ที่มา : datasolut.com

CRISP–DM เป็นกระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูลและการวิเคราะห์ข้อมูลที่สามารถนำไปปรับใช้ได้กับทุกอุตสาหกรรม โดยมีลักษณะเป็น กระบวนการแบบวนซ้ำ (Iterative Process) ซึ่งสามารถย้อนกลับไปปรับปรุงในแต่ละขั้นตอนได้ตามความเหมาะสม เพื่อให้ผลลัพธ์ที่ได้มีความแม่นยำและสอดคล้องกับวัตถุประสงค์ทางธุรกิจมากที่สุด

2.3.2 โปรแกรม RapidMiner Studio หรือ Altair® AI Studio™

โปรแกรม RapidMiner Studio หรือ Altair® AI Studio™ เครื่องมือสำหรับการวิเคราะห์ข้อมูลและสร้างโมเดลทางด้าน AI เป็นซอฟต์แวร์ที่ถูกพัฒนาเพื่อรองรับการทำงานด้าน Data Science และ Artificial Intelligence (AI) โดยครอบคลุมกระบวนการทำงานตั้งแต่การเตรียมข้อมูล (Data Preparation), การวิเคราะห์ข้อมูล (Data Analysis), ไปจนถึงการสร้างโมเดลด้วยเทคนิค Machine Learning ซึ่งช่วยให้ผู้ใช้สามารถดำเนินการวิเคราะห์และพัฒนาโมเดลได้โดยไม่จำเป็นต้องเขียนโค้ดเพิ่มเติม

ซอฟต์แวร์นี้ได้รับการออกแบบมาเพื่อช่วยให้ Data Scientists และ Data Analysts สามารถทำงานได้อย่างสะดวกและรวดเร็วผ่าน Graphical User Interface (GUI) ที่ใช้งานง่าย รองรับการทำงานแบบ Drag & Drop ทำให้สามารถสร้างกระบวนการวิเคราะห์ข้อมูลได้อย่างเป็นระบบและมีประสิทธิภาพ

นอกจากนี้ RapidMiner ยังสามารถเชื่อมต่อกับเครื่องมือวิเคราะห์ข้อมูลอื่น ๆ เช่น SAS และ Python รองรับการใช้งานทั้งบน Desktop และ Cloud ช่วยให้ผู้ใช้งานสามารถประยุกต์ใช้ในการวิเคราะห์ข้อมูลเชิงลึก และพัฒนาโมเดล AI ได้อย่างมีประสิทธิภาพ จึงเป็นหนึ่งในเครื่องมือที่ได้รับความนิยมอย่างแพร่หลายในระดับสากล

2.3.3 โปรแกรม Tableau Public

Tableau คือ ซอฟต์แวร์หรือแพลตฟอร์มที่เป็นเครื่องมือในการสร้างภาพและการวิเคราะห์ (Data Visualization and Analytics Tools) สำหรับระบบธุรกิจอัจฉริยะ (Business Intelligence) มีเทคโนโลยีด้านข้อมูลและการวิเคราะห์ธุรกิจที่ทันสมัย เพื่อหาคำตอบในลักษณะของข้อมูล (Data Insight) ทางธุรกิจได้อย่างรวดเร็วและง่ายดาย โดย Tableau จาก Salesforce เป็นหนึ่งไงซอฟต์แวร์หรือแพลตฟอร์มด้าน Visualization and Analytics ชั้นนำที่ได้รับความเชื่อถือ และถูกจัดอยู่ในกลุ่มผู้นำด้าน Analytics and BI กลุ่มเดียวกันกับ Power BI ของ Microsoft เหมาะสำหรับองค์กรที่ต้องการดำเนินธุรกิจสู่อนาคตบนวิถีทางใหม่ซึ่งมีข้อมูลเป็นตัวขับเคลื่อน (Data-Driven Culture) เพื่อสร้างการตัดสินใจ มองหาโอกาสใหม่ หรือการเปลี่ยนแปลงที่ทรงพลัง ซึ่งจำเป็นต้องบ่มเพาะบุคลากรในองค์กรให้มีความเชื่อมั่น มองเห็นคุณค่า และมีการพัฒนาทักษะการใช้ประโยชน์จากข้อมูล จนกระทั่งข้อมูลสามารถนำไปสู่ความสำเร็จขององค์กวอย่านที่จึงสามารถวัดความสำเร็จได้ Tableau แพลตฟอร์มข้อมูลและการวิเคราะห์แบบครบวงจร ซึ่งเป็นเครื่องมือสำคัญในการ ขับเคลื่อนผลลัพธ์ทางธุรกิจให้ดียิ่งขึ้น ด้วยระบบการจัดการข้อมูล (Data Management), การวิเคราะห์ด้วยภาพ (Visual Analytics) และการเล่าเรื่องข้อมูล (Data Storytelling) ตลอดจนการทำงานร่วมกัน (Collaboration) โดยบริการซอฟต์แวร์ทั้งหมดมี "Einstein" ซึ่งเป็นเทคโนโลยี AI/ML ที่พัฒนาโดย Salesforce ขับเคลื่อนอยู่เบื้องหลัง (บริษัทไอดีซีพรีเมียร์จำกัด, 2566)

2.3.4 เทคนิคทางสถิติเมทริกซ์สหสัมพันธ์ (Correlation Matrix)

เมทริกซ์สหสัมพันธ์ (Correlation Matrix) เป็นเครื่องมือทางสถิติที่ใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลข (Numerical Attributes) ในชุดข้อมูล โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient, r) ซึ่งมีค่าตั้งแต่ –1 ถึง 1 เพื่อตรวจสอบว่าตัวแปรใดมีผลกระทบต่อกันอย่างไรในซอฟต์แวร์ RapidMiner Studio สามารถใช้ Correlation Matrix Operator เพื่อคำนวณและแสดงค่าความสัมพันธ์ระหว่างตัวแปรได้อย่างง่ายดาย โดย RapidMiner จะช่วยให้ผู้ใช้สามารถวิเคราะห์และตีความข้อมูลได้สะดวกผ่าน Graphical User Interface (GUI) แบบ Drag & Drop โดยไม่จำเป็นต้องเขียนโค้ด

2.3.4.1 ขั้นตอนการใช้งาน Correlation Matrix ใน RapidMiner

- นำเข้าชุดข้อมูล
  - ใช้ Retrieve Operator เพื่อนำเข้าข้อมูลที่ต้องการวิเคราะห์
- เพิ่ม Correlation Matrix Operator
  - ไปที่ Operators และค้นหา "Correlation Matrix"
  - ลาก Operator นี้มาเชื่อมต่อกับชุดข้อมูล
- ตั้งค่าพารามิเตอร์
  - กำหนดให้คำนวณค่าความสัมพันธ์ระหว่างตัวแปรทั้งหมด (All Attributes) หรือเลือกเฉพาะตัวแปรที่สนใจ

- รันกระบวนการ (Run Process)
  - RapidMiner จะคำนวณค่า Correlation และแสดงผลลัพธ์เป็น Matrix Table
  - สามารถใช้ "Correlation Matrix to Example Set" Operator เพื่อนำผลลัพธ์ไปใช้ในกระบวนการวิเคราะห์อื่น ๆ
- วิเคราะห์ผลลัพธ์
  - ตรวจสอบค่าความสัมพันธ์ของตัวแปร หากพบว่ามีตัวแปรที่มีความสัมพันธ์สูง อาจเลือกใช้ในการสร้างโมเดล หรือหากพบว่าตัวแปรมีความสัมพันธ์ต่ำ อาจพิจารณาลบออกเพื่อลดความซับซ้อนของชุดข้อมูล

2.3.5 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

Decision tree เป็น Algorithm ที่เป็นที่นิยม ใช้ง่าย เข้าใจง่าย ได้ผลดี และเป็นฐานของ Random Forest ซึ่งเป็นหนึ่งใน Algorithm ที่ดีที่สุดในปัจจุบัน หลักการพยากรณ์ด้วย Decision tree นั้นเข้าใจง่ายมาก ให้นึกว่า Decision tree คือต้นไม้กลับหัว โดยบนสุดคือราก และส่วนล่างลงมาที่ไม่สามารถแตกไปไหนได้แล้วก็คือใบ เราจะเริ่มด้วยการพิจารณาเริ่มแรกบนจุดเริ่มต้นที่เรียกว่า Root node ถ้าข้อมูลที่พบเป็นไปตามเงื่อนไขนั้น การตัดสินใจก็จะวิ่งไปทางซ้ายของ Root node ไปที่จุดที่เรียกว่า Child node ซึ่งถ้าข้อมูลที่มาตามเส้นทางนี้ตรงตามเงื่อนไขของ Child node นี้ ก็จะถือว่าสิ้นสุด เราเรียกว่า Node สิ้นสุดว่า Leaf node ย้อนกลับไปยัง Root node ถ้าข้อมูลที่พิจารณาไม่เป็นไปตามเงื่อนไข การตัดสินใจจะวิ่งไปอีกทาง คือทางขวา ไปพบ Child node อีกอันซึ่งก็จะตั้งเงื่อนไขคำถามต่อไป การตัดสินใจก็จะวิ่งไปทางที่ตรงตามเงื่อนไข ทำอย่างนี้ไปเรื่อย ๆ จนได้คำตอบ สำหรับกลไกการเทรนโมเดลเพื่อสร้าง Decision tree จะอธิบายด้านล่าง เมื่อได้อธิบายเรื่องค่า Gini ชี้ให้เห็นว่า Decision tree นั้นไม่ต้องใช้ข้อมูลที่ทำ Feature scaling เพราะไม่ได้มี Optimisation algorithm แบบทั่วไป จึงใช้งานสะดวกมากAlgorithm ค่า gini คำนวณตามสูตรนี้

$$Gini(t) = 1 - \sum_{i=1}^{c} {p_i}^2$$

โดย pi,k คือสัดส่วนว่าจากจำนวนรายการข้อมูลใน Node ที่ i นั้นอยู่ใน Class k กี่รายการ

Gini impurity เป็นการวัดความไม่บริสุทธิ์ หรือความไม่เพียวของ class ในแต่ละกลุ่มข้อมูลที่แบ่งตามแต่ละ split point... สำหรับปัญหา classification แบบ binary ที่มี target variable เป็น 0 หรือ 1 การ split ที่ดี ควรจะได้กลุ่มข้อมูลออกมา 2 กลุ่มที่สามารถแยก class 0 กับ class 1 ออกมาได้ชัดเจนในแต่ละกลุ่ม ยิ่งสามารถแบ่งแยก class ของ target variable ออกมาได้ดี ค่า Gini impurity ก็จะยิ่งต่ำ (Daroontham, 2018)

scikit–learn จะใช้ Algorithm ที่ชื่อ Classification and Regression Tree (CART) ซึ่งทำงานตามลำดับดังนี้

2.3.5.1 แบ่ง Train set ออกเป็น 2 ส่วนโดยเลือก Class k และเงื่อนไข tk เช่น petal length <= 2.45 โดยค้นหาคู่ของ kและ tk ที่จะได้ Node ที่ "บริสุทธิ์" ที่สุด นั่นคือมีค่า Gini ต่ำที่สุดนั่นเอง เราสามารถแสดง Cost function ที่สอดคล้องกับเงื่อนไขนี้ได้ดังนี้

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

- Gleft/right คือ gini ของข้อมูลชุดซ้ายและขวาที่ถูกแบ่ง
- mleft/right คือจำนวนรายการขอมูลในชุดซ้ายและขวา
- m คือจำนวนรายการข้อมูลทั้งหมดใน Node นั้น

2.3.5.2 แยกข้อมูลแต่ละชุดย่อยออกเป็นสองชุดและทำซ้ำข้อ 1) เรื่อย ๆ จนกระทั่งถึงความลึก max depth ที่กำหนด หรือจนกระทั่งไม่พบคา k และ tk ที่จะลดความไม่บริสุทธิ์ได้อีกต่อไป

ต้นไม้ตัดสินใจเป็นเครื่องมือที่มีประโยชน์ในด้านการจำแนกประเภทและการพยากรณ์ข้อมูล สามารถใช้งานได้ง่ายและมีความยืดหยุ่น อย่างไรก็ตาม จำเป็นต้องมีการปรับปรุงโครงสร้างของต้นไม้เพื่อลดปัญหาการเกิด Overfitting และเพิ่มความแม่นยำของโมเดล (กิตตินราดร, ชิตพงษ์ กิตตินราดร, 2019)

2.3.6 เทคนิคป่าสุ่ม (Random Forest)

Random forest เป็นหนึ่งในกลุ่มของโมเดลที่เรียกว่า Ensemble learning ที่มีหลักการคือการเทรนโมเดลที่เหมือนกันหลายๆ ครั้ง (หลาย Instance) บนข้อมูลชุดเดียวกัน โดยแต่ละครั้งของการเทรนจะเลือกส่วนของข้อมูลที่เทรนไม่เหมือนกัน แล้วเอาการตัดสินใจของโมเดลเหล่านั้นมาโหวตกันว่า Class ไหนถูกเลือกมากที่สุดกลไกการรวมการตัดสินใจของผู้ตัดสินใจจำนวนมากเข้าด้วยกันมักจะให้ผลการตัดสินใจที่แม่นยำมากกว่าการพึงพาการตัดสินใจจากแหล่งเดียว ปรากฏการณ์นี้เป็นจริงในหลายมิติ เช่นในทางสังคม เราเรียกว่า "ปัญญาของฝูงชน" (Wisdom of the crowd)การเรียนรู้แบบ Ensemble นี้จะทำงานได้ดีบนเงื่อนไขที่ว่า โมเดลผู้ทำนายแต่ละตัวจะต้องเรียนรู้อย่างเป็นอิสระต่อกันให้มากที่สุด เหมือนกับเงื่อนไขของปัญญาของฝูงชน ว่าคนแต่ละคนจะต้องตัดสินใจด้วยตนเองให้มากที่สุดโดยไม่ได้รับข้อมูลจากคนอื่นหรือนำเอาข้อมูลจากคนอื่นมาเป็นส่วนในการตัดสินใจ ใน Machine learning algorithm เรามีวิธีการที่ทำให้การตัดสินใจของแต่ละโมเดลเป็นอิสระต่อกัน โดยการใช้ Algorithm เดียวกัน แต่ให้แต่ละ Instance เรียนรู้จากส่วนของข้อมูลที่ไม่เหมือนกันโดยใช้การสุ่มเลือก กลไกนี้เรียกว่า Bagging และ Pasting โดยสิ่งที่ต่างกันคือ Bagging สามารถสุ่มเลือกข้อมูลรายการเดียวกันได้ แต่ Pasting ไม่อนุญาตให้สุ่มรายการซ้ำกันได้เลย ในทางปฏิบัติ

Bagging จะลด Variance ของโมเดลได้ดีกว่า เพราะมีการเลือกรายการข้อมูลซ้ำ ทำให้ได้โมเดลที่เสถียรกว่าและมักจะแม่นยำกว่า Pasting (กิตตินราดร, 2019)

การใช้ป่าสุ่ม (Random Forests) ในการจำแนกประเภทข้อมูล เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่องที่ได้รับความนิยมอย่างมาก เนื่องจากมีความสามารถในการจำแนกข้อมูลที่มีประสิทธิภาพสูง และสามารถลดปัญหาการฟิตมากเกินไป (over-fitting) ได้อย่างมีประสิทธิภาพ แนวคิดของป่าสุ่มถูกพัฒนาโดย Breiman (2001) ซึ่งเป็นการผสานเทคนิคก่อนหน้านี้ที่เขามีส่วนร่วม เช่น การสร้างต้นไม้การจำแนกประเภท (classification trees) และการใช้เทคนิค bagging

หลักการทำงานของป่าสุ่ม ป่าสุ่มเป็นการรวมกลุ่มของต้นไม้ตัดสินใจ (decision trees) หลายต้นเข้าด้วยกัน โดยใช้กระบวนการ bootstrap sampling ซึ่งเป็นการสุ่มตัวอย่างข้อมูลจากชุดข้อมูลต้นฉบับแบบมีการคืนตัวอย่าง (with replacement) สำหรับแต่ละต้นไม้ที่สร้างขึ้นจะถูกฝึกให้เติบโตจนกระทั่งเกิดการฟิตมากเกินไปโดยไม่มีการตัดแต่ง (pruning) และจากนั้นต้นไม้แต่ละต้นจะทำการลงคะแนนเสียงแบบคณะกรรมการ (ensemble voting) เพื่อกำหนดผลลัพธ์สุดท้ายของป่าสุ่มสิ่งที่แตกต่างจากการใช้ต้นไม้การจำแนกประเภทแบบดั้งเดิมคือ ในแต่ละขั้นตอนของการเติบโตของต้นไม้ จะมีการเลือกใช้ตัวแปรอิสระบางตัว (features) แบบสุ่มแทนที่จะใช้คุณลักษณะทั้งหมดในการสร้างกฎการแบ่งข้อมูล โดยปกติแล้ว จำนวนคุณลักษณะที่เลือกใช้ในการแบ่งข้อมูลแต่ละชั้นจะอยู่ที่ประมาณ ของคุณลักษณะทั้งหมด วิธีการนี้ช่วยลดความสัมพันธ์ระหว่างต้นไม้แต่ละต้น ทำให้ป่าสุ่มมีความสามารถในการจำแนกข้อมูลที่มีความแม่นยำสูงและลดปัญหาการฟิตมากเกินไป ความทนทานต่อการฟิตมากเกินไปของป่าสุ่ม มีการพิสูจน์ทางคณิตศาสตร์ว่าป่าสุ่มมีความทนทานต่อปัญหาการฟิตมากเกินไป กล่าวคือ การเพิ่มจำนวนต้นไม้ ในป่าอย่างไม่มีขีดจำกัดจะไม่ทำให้ป่าสุ่มเกิดการฟิตมากเกินไป แม้ว่าจำนวนพารามิเตอร์ที่ใช้ในการอธิบายป่าจะเพิ่มขึ้นเป็นอนันต์ก็ตาม Breiman (2001) ได้แสดงให้เห็นว่าเมื่อ ความเสี่ยงของป่าสุ่มจะเข้าใกล้ค่าความน่าจะเป็นที่ข้อมูลใหม่ที่ได้จากการแจกแจงข้อมูลต้นฉบับจะถูกจัดประเภทผิดโดยต้นไม้ต้นเดียวที่ใช้กฎการเลือกตัวแปรอิสระแบบสุ่ม กล่าวคือ ในทางปฏิบัติไม่มีความจำเป็นต้องกังวลเกี่ยวกับการกำหนดค่า ที่สูงเกินไป เว้นแต่ต้นทุนในการคำนวณที่เพิ่มขึ้น

ความใกล้ชิดของป่าสุ่ม (Random Forest Proximity)

Breiman ได้เสนอแนวคิดเกี่ยวกับ "ความใกล้ชิดของป่าสุ่ม" (Random Forest Proximity) ซึ่งเป็นค่าที่ใช้วัดความสัมพันธ์ระหว่างจุดข้อมูลสองจุดในป่าสุ่ม โดยพิจารณาจากสัดส่วนของต้นไม้ที่ทำให้จุดข้อมูลทั้งสองตกอยู่ในโหนดปลายทาง (leaf node) เดียวกัน ความใกล้ชิดนี้สามารถใช้ในการวิเคราะห์โครงสร้างของข้อมูล เช่น การจัดกลุ่ม (clustering) หรือการค้นหาความสัมพันธ์ของข้อมูลในแต่ละคลาส วิธีการดังกล่าวช่วยให้ป่าสุ่มไม่เพียงแต่ใช้ในการจำแนกข้อมูล แต่ยังสามารถนำมาใช้เพื่อสำรวจลักษณะโครงสร้างของข้อมูลเชิงลึกได้อีกด้วย ป่าสุ่มเป็นเทคนิคที่มีความแข็งแกร่งและมีความสามารถในการจำแนกข้อมูลได้อย่างมีประสิทธิภาพ เนื่องจาก

สามารถลดความสัมพันธ์ระหว่างต้นไม้ในป่า ส่งผลให้มีความแม่นยำสูงและสามารถต้านทานปัญหาการฟิตมากเกินไป นอกจากนี้ ป่าสุ่มยังมีความสามารถในการวัดความใกล้ชิดของข้อมูล ซึ่งเป็นประโยชน์อย่างมากในการวิเคราะห์โครงสร้างข้อมูลและค้นหารูปแบบในชุดข้อมูลที่ซับซ้อน การเลือกใช้ป่าสุ่มในงานวิจัยจึงเป็นทางเลือกที่น่าสนใจสำหรับปัญหาการจำแนกประเภทข้อมูลและการวิเคราะห์ข้อมูลในเชิงลึก (walter A. Shewhart and Samuel S. Wilks, 2018)

2.3.7 เทคนิคการวิเคราะห์ข้อมูลการถดถอยพหุคูณ (Multiple Linear Regression)

การถดถอยพหุคูณ (Multiple Regression Analysis) เป็นเทคนิคทางสถิติที่ใช้วิเคราะห์ความสัมพันธ์ระหว่าง ตัวแปรตาม (Dependent Variable, Y) และ ตัวแปรอิสระหลายตัว (Independent Variables, X1, X2, X3, ...) โดยมีสมการทั่วไปดังนี้ (Montgomery et al., 2021)

$$\mathcal{Y} = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \ldots + \beta_n \chi_n + \varepsilon$$

โดยที่

- $\mathcal{Y}$ คือค่าของตัวแปรตาม
- $\chi_1, \chi_2, \ldots, \chi_n$ คือตัวแปรอิสระ
- β_0 ค่าคงที่ (Intercept)
- $\beta_1, \beta_2 \ldots \beta_n$ คือสัมประสิทธิ์ของตัวแปรอิสระ
- ε คือข้อผิดพลาด (Error Term)

2.3.7.1 ขั้นตอนการวิเคราะห์

ก. ตรวจสอบสมมติฐานเบื้องต้น

- ความเป็นเส้นตรง (Linearity) ควรตรวจสอบว่าความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเป็นเชิงเส้น
- ไม่มีมัลติโคลิเนียร์ริตี (No Multicollinearity) ใช้ Variance Inflation Factor (VIF) เพื่อตรวจสอบว่าตัวแปรอิสระมีความสัมพันธ์ซ้อนทับกันหรือไม่
- ความเป็นอิสระของข้อผิดพลาด (Independence of Errors) ทดสอบด้วย Durbin–Watson Test
- ความเป็นปกติของข้อผิดพลาด (Normality of Residuals) ใช้การทดสอบ Shapiro–Wilk หรือ Q–Q Plot
- ความคงที่ของความแปรปรวน (Homoscedasticity) ตรวจสอบด้วย Scatter Plot หรือ Breusch–Pagan Test

ข. การประมาณค่าสัมประสิทธิ์ ใช้วิธี Ordinary Least Squares (OLS) เพื่อหาค่าของสัมประสิทธิ์ที่เหมาะสมที่สุด โดยพยายามลดค่าผลรวมของกำลังสองของข้อผิดพลาดให้ต่ำที่สุด

ค. การทดสอบนัยสำคัญทางสถิติ

- ทดสอบนัยสำคัญของโมเดล (Overall Model Significance) ใช้ ค่า F-test เพื่อตรวจสอบว่าโมเดลมีนัยสำคัญหรือไม่
- ทดสอบนัยสำคัญของตัวแปรอิสระแต่ละตัว (Individual Variable Significance) ใช้ ค่า t-test เพื่อตรวจสอบว่าสัมประสิทธิ์ของตัวแปรอิสระมีนัยสำคัญทางสถิติหรือไม่

ง. การวัดความเหมาะสมของโมเดล

- ค่า R-Squared (R^2) ใช้ในการประเมินว่าสัดส่วนของความแปรปรวนของตัวแปรตามสามารถอธิบายได้โดยตัวแปรอิสระ
- Adjusted R-Squared เป็นค่า R^2 ที่ปรับให้เหมาะสมเมื่อเพิ่มตัวแปรอิสระมากขึ้น

2.3.7.2 การแปลผลและการนำไปใช้

หลังจากวิเคราะห์แล้ว สามารถนำผลลัพธ์ไปใช้ในการพยากรณ์หรือวิเคราะห์ปัจจัยที่ส่งผลต่อปรากฏการณ์ที่สนใจ โดยควรระมัดระวังเกี่ยวกับการตีความความสัมพันธ์เชิงเหตุและผล รวมถึงข้อจำกัดของโมเดล เช่น การละเมิดสมมติฐาน และปัจจัยที่ไม่ได้รวมอยู่ในโมเดล (Field, 2013)

2.4 วรรณกรรมที่เกี่ยวข้อง

2.4.1 การวิเคราะห์และเปรียบเทียบปัจจัยการล่าช้าและการตรงต่อเวลาของสายการบินในภูมิภาคเอเชีย (บุญญวัฒน์ อักษรกิตติ์ และคณะ, 2563)

งานวิจัยนี้ใช้ การวิเคราะห์ความถดถอยพหุคูณ (Multiple Linear Regression Analysis) เพื่อหาความสัมพันธ์ระหว่างปัจจัยที่ส่งผลต่อความล่าช้า เช่น การจัดการจราจรทางอากาศ สภาพอากาศ และการซ่อมบำรุง นอกจากนี้ ยังใช้ การจำลองสถานการณ์ (Simulation) เพื่อทดสอบแนวทางการแก้ปัญหา โดยผลการวิจัยแสดงว่าการปรับปรุงระบบควบคุมจราจรทางอากาศและการลงทุนในเทคโนโลยี AI สามารถช่วยลดความล่าช้าได้อย่างมีประสิทธิภาพ

2.4.2 การพยากรณ์ความล่าช้าในการมาถึงของอากาศยานโดยใช้วิธีการเรียนรู้ของเครื่อง (ร.อ.ภัทร ชาญวานิชบริการ, 2566)

งานวิจัยนี้มุ่งศึกษาการพยากรณ์ความล่าช้าของเที่ยวบินโดยใช้วิธีการเรียนรู้ของเครื่องเพื่อจัดการกับปัญหาความล่าช้า ซึ่งส่งผลต่อประสิทธิภาพและต้นทุนในอุตสาหกรรมการบินของสหรัฐฯ โดยใช้ข้อมูลการบินระหว่างปี 2019-2020 มาพัฒนาและทดสอบแบบจำลองการคาดการณ์ผ่านวิธีการ Classification ได้แก่ Random Forest, Cat Boost, Gradient Boosting และ AdaBoost โดย Random Forest แสดงผลลัพธ์ได้ดีที่สุดด้วยความแม่นยำ 83% และ F1-Score

ที่ 0.56 การวิจัยยังชี้ให้เห็นข้อจำกัดในการเข้าถึงข้อมูลแบบเรียลไทม์ เช่น สภาพอากาศ ระหว่างทางและการจราจรทางอากาศ พร้อมเสนอการบูรณาการข้อมูลเชิงลึกเพิ่มเติมเพื่อปรับปรุงความแม่นยำในอนาคต ทั้งนี้ ผลการวิจัยสามารถช่วยสายการบินปรับปรุงตารางเที่ยวบิน ลดต้นทุน เพิ่มความพึงพอใจของผู้โดยสาร และสนับสนุนการลดการปล่อยก๊าซเรือนกระจก ซึ่งเป็นประโยชน์ต่อทั้งธุรกิจการบินและสิ่ง

2.4.3 การศึกษาผลกระทบของเที่ยวบินล่าช้าขาออกภายในประเทศ กรณีศึกษาท่าอากาศยานเชียงใหม่ (ดลณา พานอ่อง, 2566)

งานวิจัยนี้ใช้ การวิเคราะห์เชิงสถิติ (Statistical Analysis) และ การวิเคราะห์ข้อมูลเชิงคุณภาพ (Qualitative Data Analysis) เพื่อศึกษาผลกระทบของความล่าช้าต่อผู้โดยสาร ผลการวิเคราะห์พบว่าความล่าช้าส่งผลต่อความพึงพอใจและค่าใช้จ่ายเพิ่มเติม เช่น การพลาดเที่ยวบินต่อ งานวิจัยแนะนำให้ใช้ การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics) เพื่อคาดการณ์ความล่าช้าและจัดทำแผนสำรอง

2.4.4 การศึกษาเปรียบเทียบความล่าช้าเที่ยวบินระหว่างประเทศของสายการบินต้นทุนต่ำ (เมธี เกตุรักษา, 2564)

งานวิจัยนี้ใช้ โปรแกรม Easy Fit เป็นเครื่องมือหลักในการวิเคราะห์โมเดลการแจกแจง (Distribution Model) ของข้อมูลความล่าช้าในเที่ยวบิน โดยทำการเปรียบเทียบโมเดลต่าง ๆ เพื่อหาการแจกแจงที่เหมาะสมที่สุดกับข้อมูล เช่น Normal Distribution, Exponential Distribution และ Weibull Distribution โมเดลที่ได้จากการวิเคราะห์ถูกนำมาใช้ในการระบุรูปแบบความล่าช้าที่พบได้บ่อยและระยะเวลาความล่าช้าที่เป็นไปได้มากที่สุดในช่วงเวลาที่กำหนดโมเดล Weibull Distribution แสดงผลลัพธ์ที่เหมาะสมที่สุดสำหรับข้อมูลความล่าช้าของสายการบินต้นทุนต่ำในสถานการณ์ที่สนามบินมีความหนาแน่นสูง โดยโมเดลนี้สามารถสะท้อนพฤติกรรมของความล่าช้าได้อย่างแม่นยำในช่วงเวลาสำคัญ เช่น วันหยุดยาวและฤดูกาลท่องเที่ยว

2.4.5 การวิเคราะห์การแจกแจงความล่าช้าของเที่ยวบินขาออก (นิษฐเนตร์ ขจรเทวาวงศ์ และคณะ, 2563)

งานวิจัยนี้ใช้ การวิเคราะห์เชิงสถิติ (Statistical Analysis) เพื่อศึกษาการแจกแจงความล่าช้าของเที่ยวบินขาออกภายในประเทศ โดยใช้ข้อมูลเที่ยวบินระหว่างปี 2560–2561 และวิเคราะห์ด้วยโปรแกรม Easy Fit เวอร์ชัน 5.6 พร้อมทดสอบความเหมาะสมด้วยสถิติแอนเดอร์สัน–ดาร์ลิง ผลการศึกษาแสดงให้เห็นว่าความล่าช้าส่วนใหญ่อยู่ในช่วงไม่เกิน 30 นาที และพบว่าเที่ยวบินจากจังหวัดพิษณุโลกและกรุงเทพฯ ไปยังภาคตะวันออกเฉียงเหนือในช่วงฤดูฝนมีความล่าช้ากว่า 6 ชั่วโมง การแจกแจงที่เหมาะสมที่สุดของความล่าช้า ได้แก่ การแจกแจงล็อกเพียร์สันประเภทที่ 3 และการแจกแจงเบอร์ คิดเป็นร้อยละ 25.71 เท่ากัน งานวิจัยเสนอแนะให้ขยายระยะเวลาการเก็บข้อมูลและนำผลการวิเคราะห์ไปใช้พัฒนาตารางการบิน ลดความล่าช้า และปรับปรุงกรมธรรม์ประกันภัยการเดินทางให้เหมาะสมกับเส้นทางบิน

2.4.6 ปัจจัยที่มีผลต่อการผันแปรของหุ้นกลุ่มเทคโนโลยีโดยใช้เทคนิคการทำเหมืองข้อมูล(ยศสยา แสงหิรัญ และ สมชาย เล็กเจริญ,2561)

งานวิจัยนี้ศึกษาปัจจัยที่มีผลกระทบต่อการเปลี่ยนแปลงของราคาหุ้นในกลุ่มเทคโนโลยี โดยใช้ เทคนิคการทำเหมืองข้อมูล (Data Mining) เพื่อวิเคราะห์และเปรียบเทียบประสิทธิภาพของโมเดลการพยากรณ์ที่แตกต่างกัน ซึ่งได้เลือกใช้ Decision Tree (J48), Random Forest, K-Nearest Neighbor (KNN), และ Neural Network ผลการวิเคราะห์พบว่า ปัจจัยที่มีผลกระทบต่อราคาหุ้นมากที่สุด ได้แก่ มูลค่าตามบัญชีต่อหุ้น (BVPS), ผลตอบแทนต่อสินทรัพย์ (ROA), และอัตราผลตอบแทนผู้ถือหุ้น (ROE) สำหรับประสิทธิภาพของแบบจำลอง KNN เป็นโมเดลที่แม่นยำที่สุด โดยมีค่าความแม่นยำ 100% และค่าความคลาดเคลื่อนเฉลี่ย 0.00 รองลงมาคือ Random Forest (ความแม่นยำ 100%, ค่าคลาดเคลื่อน 0.14), Neural Network (81.11%, 0.24), และ Decision Tree (76.66%, 0.35)

2.4.7 การนิยามเที่ยวบินล่าช้าในอุตสาหกรรมการบินตามมาตรฐาน OAG (OAG, 2025)

งานวิจัยนี้อ้างอิงนิยามของเที่ยวบินล่าช้าตามมาตรฐานที่กำหนดโดย OAG ซึ่งเป็นผู้ให้บริการข้อมูลการบินระดับโลก โดย OAG กำหนดว่า เที่ยวบินจะถือว่า “ตรงต่อเวลา” (On-Time Performance: OTP) ก็ต่อเมื่อเที่ยวบินนั้นออกเดินทางหรือมาถึงปลายทางภายในระยะเวลาไม่เกิน 15 นาทีจากเวลาที่กำหนดในตาราง หากเที่ยวบินมีความล่าช้ามากกว่า 15 นาที จะถือว่าเป็นเที่ยวบิน “ล่าช้า” (Delayed Flight) นิยามนี้ถูกนำมาใช้เป็นเกณฑ์ในการสร้างตัวแปรเป้าหมาย (Target Variable) สำหรับการจำแนกเที่ยวบินล่าช้าในแบบจำลองการเรียนรู้ของเครื่อง ซึ่งช่วยให้สามารถประเมินและพยากรณ์ความล่าช้าได้อย่างแม่นยำตามมาตรฐานอุตสาหกรรมที่ได้รับการยอมรับในระดับสากล

2.4.8 การวิเคราะห์ผลกระทบของช่วงเวลาออกเดินทางต่อความล่าช้าของเที่ยวบิน (Arora et al., 2020)

งานวิจัยนี้ศึกษาปัจจัยที่ส่งผลต่อความล่าช้าของเที่ยวบินในสหรัฐอเมริกา โดยใช้เทคนิคการวิเคราะห์ข้อมูลขนาดใหญ่จากข้อมูลเที่ยวบินเชิงประจักษ์ เพื่อวิเคราะห์ความสัมพันธ์ระหว่างช่วงเวลาออกเดินทางกับความถี่ของความล่าช้า ผลการศึกษาพบว่าเที่ยวบินที่มีกำหนดการออกเดินทางในช่วงบ่ายถึงเย็น มีแนวโน้มเกิดความล่าช้ามากกว่าเที่ยวบินที่ออกเดินทางในช่วงเช้าอย่างมีนัยสำคัญ เนื่องจากเป็นช่วงเวลาที่เกิดการสะสมของความล่าช้าจากเที่ยวบินก่อนหน้า (delay propagation) และภาระงานของสนามบินอยู่ในระดับสูง งานวิจัยนี้จึงเสนอให้สายการบินและผู้วางแผนการเดินทางพิจารณาใช้ข้อมูลช่วงเวลานี้ในการทำนายโอกาสเกิดความล่าช้าและวางแผนล่วงหน้าอย่างเหมาะสม

2.4.9 การทำนายความล่าช้าของอากาศยานด้วยวิธีการเรียนรู้ของเครื่อง (Charnvanichborikarn, 2023)

งานวิจัยนี้มุ่งเน้นการพัฒนาแบบจำลองเพื่อทำนายความล่าช้าในการมาถึงของเที่ยวบิน โดยใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อรับมือกับความท้าทายใน

อุตสาหกรรมการบินสหรัฐอเมริกา ผู้วิจัยได้ใช้ชุดข้อมูลจาก Kaggle ที่ประกอบด้วยข้อมูลการบินและสภาพอากาศที่สนามบินปลายทาง พร้อมทั้งทำการประมวลผลและทำความสะอาดข้อมูลเพื่อเพิ่มความแม่นยำของแบบจำลอง

แบบจำลองที่นำมาทดสอบประกอบด้วย Random Forest, CatBoost, Gradient Boosting และ AdaBoost โดยใช้เทคนิค GridSearchCV เพื่อปรับค่าพารามิเตอร์ให้เหมาะสมที่สุด ผลการทดลองพบว่าแบบจำลอง Random Forest มีประสิทธิภาพดีที่สุด โดยได้ค่าความแม่นยำ (Accuracy) 83% และค่า F1–score เท่ากับ 0.56 ซึ่งสูงกว่าแบบจำลองอื่น ๆ ได้แก่ Gradient Boosting (Accuracy 81.1%, F1–score 0.47), CatBoost (Accuracy 81%, F1–score 0.46) และ AdaBoost (Accuracy 72.6%, F1–score 0.45)

ข้อค้นพบสำคัญของงานวิจัยนี้คือ ความล่าช้าของเที่ยวบินส่วนใหญ่มักเกิดจากสภาพอากาศและปัจจัยด้านการจราจรทางอากาศ จึงเสนอให้มีการบูรณาการข้อมูลสภาพอากาศแบบเรียลไทม์เข้ากับแบบจำลองทำนาย เพื่อเพิ่มประสิทธิภาพในการคาดการณ์ความล่าช้า ผลลัพธ์ดังกล่าวชี้ให้เห็นว่าเทคนิคการเรียนรู้ของเครื่อง โดยเฉพาะ Random Forest สามารถนำมาใช้เป็นเครื่องมือสนับสนุนการตัดสินใจในการจัดการตารางบิน และช่วยลดผลกระทบจากความล่าช้าต่อผู้โดยสารและสายการบินได้อย่างมีนัยสำคัญ

2.5 บทสรุป

บทนี้ได้กล่าวถึงแนวคิดและทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลที่ส่งผลต่อความล่าช้าของสายการบิน โดยครอบคลุมตั้งแต่การวิเคราะห์ข้อมูล (Data Analytics), การทำความสะอาดข้อมูล (Data Cleansing), การพยากรณ์ข้อมูล (Forecasting Data), และการแสดงผลข้อมูล (Data Visualization) ซึ่งเป็นองค์ประกอบสำคัญในการจัดการข้อมูลเพื่อใช้ประโยชน์สูงสุด ด้านทฤษฎีได้กล่าวถึงแนวคิดการทำเหมืองข้อมูล (Data Mining), การจำแนกประเภทข้อมูล (Classification), การถดถอยพหุคูณ (Multiple Regression), และหลักการของการทำเหมืองข้อมูล ซึ่งเป็นแนวทางหลักที่ใช้ในกระบวนการวิเคราะห์ข้อมูล นอกจากนี้ ยังมีการกล่าวถึงปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน ทั้งจากปัจจัยภายในและภายนอก เช่น สภาพอากาศ, การจราจรทางอากาศ, การบำรุงรักษาเครื่องบิน, และการบริหารจัดการภายในสายการบิน เครื่องมือที่ใช้ในการออกแบบและวิเคราะห์ข้อมูล เช่น กระบวนการ CRISP–DM, เทคนิคการถดถอยพหุคูณ, และเครื่องมือ Correlation Matrix ซึ่งเป็นองค์ประกอบสำคัญในการวิจัยเพื่อคาดการณ์และวิเคราะห์ข้อมูลเชิงลึก