บทที่ 2
แนวคิด ทฤษฎีเอกสารที่เกี่ยวข้อง

โครงงานเรื่อง การเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการพยากรณ์การใช้น้ำประปาด้วยเทคนิคเหมืองข้อมูล และนำเสนอข้อมูลผ่านเว็บไซต์ ในบทนี้เป็นการนำเสนอเกี่ยวกับ แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้องของการเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการพยากรณ์การใช้น้ำประปาด้วยเทคนิคเหมืองข้อมูล และนำเสนอข้อมูลผ่านเว็บไซต์ ซึ่งได้รวบรวมการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล เพื่อใช้เป็นแนวทางการศึกษาประกอบด้วยรายละเอียดตามลำดับดังนี้

2.1 แนวคิด

2.1.1 การวิเคราะห์ข้อมูล (Data analytic)

2.1.2 การพยากรณ์ข้อมูล (Forecasting data)

2.1.3 การเปรียบเทียบโมเดล (Model Comparison Concepts)

2.1.4 การแสดงผลข้อมูล (Data visualization)

2.1.5 การทำความสะอาดข้อมูล (Data Cleaning)

2.2 ทฤษฎี

2.2.1 การทำเหมืองข้อมูล (Data Mining)

2.2.2 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

2.3.1 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

2.3.2 เทคนิคการวิเคราะห์ข้อมูลถดถอยเชิงเส้นหรือ (Linear Regression)

2.3.3 เทคนิคแรนดอมฟอเรสต์ (Random Forest)

2.3.4 เทคนิคต้นไม้เสริมกำลังแบบไล่ระดับ (Gradient Boosted Trees)

2.3.5 การทดสอบประสิทธิภาพของตัวแบบโดยใช้ค่าเฉลี่ยของกำลังสองของความคลาดเคลื่อน (Mean Absolute Error: MAE)

2.3.6 ค่าเฉลี่ยของรากที่สองของกำลังสองของความคลาดเคลื่อน (Root Mean Square Error : RMSE)

2.3.7 กระบวนการวิเคราะห์ข้อมูลด้วย (CRISP–DM)

2.4 วรรณกรรมที่เกี่ยวข้อง

2.5 บทสรุป

2.1 แนวคิด

2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data analytic)

ในการดำเนินงานเรื่องการเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการพยากรณ์การใช้น้ำประปาด้วยเทคนิคเหมืองข้อมูล ทางผู้วิเคราะห์ข้อมูลได้ศึกษาหลักการ และทฤษฎีต่าง ๆ องค์ประกอบหนึ่งที่สำคัญคือการวิเคราะห์ข้อมูล ซึ่งมีรายละเอียด ดังนี้

การวิเคราะห์ข้อมูล (Data Analytics) คือ การรวบรวมข้อมูลขนาดใหญ่ (Big Data) มาจัดเรียงให้เป็นระบบ และนำมาวิเคราะห์หาข้อมูลเชิงลึก (Insight) เพื่อสรุปผลและใช้สนับสนุนการตัดสินใจทางธุรกิจได้อย่างมีประสิทธิภาพ เนื่องจากความสามารถในการนำข้อมูลมาวิเคราะห์เพื่อแก้ไขปัญหาหรือช่วยในการตัดสินใจทางธุรกิจ และวัตถุประสงค์อื่นๆ ที่ต้องการเริ่มต้นจากขั้นตอนพื้นฐานของการนำข้อมูลมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการประมวลผล ทั้งนี้ การวิเคราะห์ข้อมูลยังขึ้นอยู่กับเทคโนโลยีและชุดคำสั่งที่ใช้ สำหรับรูปแบบของการวิเคราะห์ข้อมูล (Data Analytics) สามารถแบ่งได้ดังนี้

1.การวิเคราะห์ข้อมูลพื้นฐาน (Descriptive Analytics) เป็นการวิเคราะห์ข้อมูลเชิงพรรณนา เพื่อตอบคำถามว่าเหตุการณ์ หรือกิจกรรมต่าง ๆ ที่เกิดขึ้นในอดีตเป็นอย่างไร มีอะไรผิดปกติ หรือเกิด ในลักษณะที่ง่ายต่อการเข้าใจและตัดสินใจ เช่น รายงานการขายและผลดำเนินงาน

2.การวิเคราะห์แบบเชิงวินิจฉัย (Diagnostic Analytics) การวิเคราะห์ข้อมูลนี้เน้นเจาะลึกลงไปถึงสาเหตุของสิ่งที่เกิดขึ้นมีสาเหตุมาจากอะไร มีปัจจัยใดบ้างที่ส่งผลต่อความสัมพันธ์ของข้อมูลที่เกิดขึ้น โดยจะมีความลึกกว่าการวิเคราะห์ข้อมูลแบบพื้นฐาน ลงลึกในรายละเอียดเพิ่มเติมขากข้อมูลแบบ การวิเคราะห์ข้อมูลแบบพื้นฐานจนไปถึงการตั้งสมมติฐานเบื้องต้นว่าทำไมเหตุการณ์เหล่านี้จึงเกิดขึ้น

3.การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) การวิเคราะห์เชิงทำนาย โดยนำข้อมูลที่มีในอดีตมาวิเคราะห์และทำนายสิ่งที่จะเกิดขึ้นในอนาคต หรือมีความเป็นไปได้ว่าจะเกิดขึ้น ทำให้สามารถวิเคราะห์หาโอกาสและความเสี่ยงต่าง ๆ ที่จะเกิดขึ้นในอนาคตได้ โดยการใช้แบบจำลองทางสถิติ

4.การวิเคราะห์ข้อมูลแบบให้คำแนะนำ (Prescriptive Analytics) การวิเคราะห์ข้อมูลที่ต่อเนื่องมาจากการวิเคราะห์พยากรณ์ ซึ่งเป็นการวิเคราะห์ข้อมูลแบบที่มีความซับซ้อนมากที่สุด ซึ่งเป็นการพยากรณ์สิ่งที่กำลังจะเกิดขึ้น เพื่อหาสาเหตุ ข้อดี ข้อสีย รวมไปถึงระยะเวลา

ของสิ่งที่จะเกิดขึ้น เพื่อหาว่าควรปรับปรุง พัฒนา หรือแก้ไขปัญหาอะไรบ้าง หรือใช้เพื่อระบุแนวโน้มเทรนด์ต่าง ๆ ที่จะเกิดขึ้น และสามารถใช้วิเคราะห์ได้ว่าทางเลือกแต่ละแนวทางที่มีจะให้ผลลัพธ์แบบใดได้บ้าง ซึ่งช่วยให้สามารถตัดสินใจได้แม่นยำ (Survey, 2565)

2.1.2 แนวคิดเกี่ยวกับการพยากรณ์ข้อมูล (Forecasting data)

การพยากรณ์ (Forecasting) คือการคาดการณ์โดยศึกษาจากข้อมูลเก่าและรูปแบบต่าง ๆ ในอดีต หลายธุรกิจใช้เครื่องมือและระบบซอฟต์แวร์เพื่อวิเคราะห์ข้อมูลจำนวนมากที่เก็บรวบรวมมาเป็นระยะเวลานาน จากนั้นซอฟต์แวร์จะคาดการณ์ความต้องการและแนวโน้มในอนาคต การพยากรณ์มีความสำคัญในการช่วยให้การตัดสินใจมีความถูกต้องและการวางแผนต่าง ๆ สำหรับอนาคตขององค์กรหรือดำเนินงาน กระบวนการนี้อาศัยข้อมูลปัจจุบันและอดีตรวมทั้งการใช้วิจารณญาณ ความรู้และประสบการณ์ของบุคคลเพื่อกำหนดการประมาณการให้มีประโยชน์ในการตัดสินใจ ซึ่งการพยากรณ์มีกระบวนการพยากรณ์ (Forecasting Process) 5 ขั้นตอนดังนี้

1.ระบุวัตถุประสงค์ของการพยากรณ์ เพื่อให้สามารถเลือกเทคนิคการพยากรณ์ที่เหมาะสมกับวัตถุประสงค์ของผู้ใช้

2.กำหนดช่วงเวลาที่ต้องการพยากรณ์

2.1 การพยากรณ์ระยะสั้น (Short-term Forecasting) ไม่เกิน 1 ปี

2.2 การพยากรณ์ระยะปานกลาง (Medium-term Forecasting) 1-3 ปี

2.3 การพยากรณ์ระยะยาว (Long-term Forecasting) 3 ปีขึ้นไป

3.เลือกเทคนิคการพยากรณ์ที่เหมาะสม (Forecasting Techniques) กับวัตถุประสงค์ของการพยากรณ์ ข้อมูลที่ต้องการ ระยะเวลาที่ต้องการและต้นทุนในการพยากรณ์

4.เก็บข้อมูลที่ต้องการใช้ในการพยากรณ์

5.นำข้อมูลที่ได้ไปทำการพยากรณ์

2.1.3 แนวคิดเกี่ยวกับการเปรียบเทียบโมเดล (Model Comparison Concepts)

การเปรียบเทียบประสิทธิภาพระหว่างโมเดลวิเคราะห์ข้อมูลคือกระบวนการที่ใช้เพื่อทดสอบและวัดผลประสิทธิภาพของโมเดลเพื่อหาว่าแบบจำลองใดมีประสิทธิภาพที่ดีกว่าในการแก้ไขปัญหาหรือทำนายผลข้อมูลที่ให้มา การเปรียบเทียบประสิทธิภาพของโมเดลอาจมีหลายแง่มุมต่าง ๆ และต้องพิจารณาในรูปแบบต่าง ๆ ตามลักษณะของงานและข้อมูลที่มีอยู่ด้วยกัน ดังนี้

1) ประสิทธิภาพในการทำนาย (Prediction Performance): เมื่อเปรียบเทียบโมเดลสองระบบ โมเดลที่ให้ผลการทำนายที่แม่นยำและความคลาดเคลื่อนต่ำจะถือว่ามี ประสิทธิภาพ

ดีกว่า ค่าเชิงคณิตศาสตร์ที่ใช้ในการวัดอาจมีค่าเช่น Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Rsquared (R2) หรืออื่น ๆ โดยทั่วไป ค่าที่น้อยกว่าคือดีกว่า

2) ประสิทธิภาพในการจัดกลุ่ม (Classification Performance: ในงานการจัดกลุ่ม (classification) การใช้ค่าคลาดเคลื่อนรวม (accuracy), precision, recall, F1-score, ROC-AUC, หรือค่าอื่น ๆ เป็นวิธีการวัดประสิทธิภาพที่พบบ่อย เพื่อประเมิน ความสามารถในการแยกแยะคลาสต่าง ๆ โมเดลที่ให้ผลดีในเรื่องนี้จะมีค่าใกล้เคียง 1 หรือใกล้เคียง 100% ในค่าที่เหมาะสม

3) ประสิทธิภาพในการใช้งาน (Operational Performance): การประเมินประสิทธิภาพที่มี ผลต่อการใช้งานจริงของโมเดล เช่น ความเร็วในการสร้างโมเดล (training time), ความเร็วในการทำนาย (inference time), และการใช้ทรัพยากรคอมพิวเตอร์ เป็นต้น โมเดลที่มีประสิทธิภาพดีแต่ใช้เวลานานหรือทรัพยากรมากอาจไม่เหมาะสมสำหรับงาน บางประเภท

4) ประสิทธิภาพในการจัดการข้อมูล (Data Handling Performance): การควบคุมความเสี่ยงและการจัดการข้อมูลที่หายไป (missing data) หรือข้อมูลที่ผิดปกติ (outliers) โดย การปรับปรุงและปรับแต่งข้อมูลอาจส่งผลต่อประสิทธิภาพของโมเดล

5) ความคงเส้นคงวง (Robustness): โมเดลที่มีความสามารถในการทำงานแม้ในสภาวะที่ ข้อมูลอาจมีความผิดปกติหรือเงื่อนไขที่ซับซ้อน หรือโมเดลที่ไม่ถูกกระทบโดยมากจาก 11 ความเปลี่ยนแปลงของข้อมูล (แยกจากการเทรนอาจเกิดการเรียนรู้เกินไปจากข้อมูล เฉพาะ)

6) การใช้งานจริง (Practical Usability): การคำนึงถึงการทำงานในสภาพแวดล้อมที่จริง โดยพิจารณาปัญหาเชิงปฏิบัติที่อาจเกิดขึ้นในการใช้งานจริง เช่น ปัญหาความเร็วใน การรับข้อมูลใหม่ หรือปัญหาความปลอดภัยและความเป็นส่วนตัว

ในการเปรียบเทียบประสิทธิภาพของโมเดลวิเคราะห์ข้อมูลสองโมเดล ควรพิจารณาแง่มุมที่ เหมาะสมกับงานและวัตถุประสงค์ของคุณ และใช้เครื่องมือทางสถิติและการทดสอบเพื่อวัดผล ประสิทธิภาพที่ถูกต้องและทางวิเคราะห์ต่อไป

2.1.4 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data visualization)

Data visualization คือ การนำข้อมูลหรือ Data ที่ได้มาจากแหล่งข้อมูลต่างๆ มาวิเคราะห์ประมวลผลแล้วนำเสนอออกมาในรูปแบบที่มองเห็นและทำความเข้าใจได้ด้วยตา เช่น แผนภูมิ รูปภาพ แผนที่ กราฟแสดงเทรนด์ ตาราง วิดีโอ อินโฟกราฟิก (Infographic) แดชบอร์ด (dashboard) จุดประสงค์สำคัญของการทำ Data Visualization คือ การนำเสนอข้อมูลให้เข้าใจง่าย ผู้อ่านข้อมูลสามารถเข้าใจได้ทันทีว่าตัวชิ้นงาน (media) ต้องการสื่อสารอะไร ชี้จุดสำคัญ

ของเนื้อหา และชี้ Insight ข้อเปรียบเทียบให้เห็นอย่างชัดเจน ช่วยให้สังเกตเห็นจุดที่น่าสนใจของข้อมูลได้ง่ายขึ้น

Data Visualization คล้ายกับ การทำ Presentation ที่ต้องอาศัยศาสตร์แห่งศิลป์เข้ามาช่วยในการนำเสนอข้อมูลให้ดูน่าสนใจ ดังนั้น องค์ประกอบต่างๆ ของการทำ Data Visualization จึงเป็นหนึ่งในขั้นตอนการเตรียมข้อมูลในการนำเสนอ เพิ่มเติมคือรายละเอียดและความซับซ้อนของข้อมูลที่มากกว่า โดยองค์ประกอบของ Data Visualization มีดังนี้

1.ข้อมูล (Information): ข้อมูลถือเป็นองค์ประกอบตั้งต้นของการทำ Data Visualization เมื่อมีข้อมูลแล้วจำเป็นต้องนำมาจำแนกแยกย่อย เพื่อเลือกข้อมูลสำคัญและตอบโจทย์ในสิ่งที่ต้องการ จากนั้นก็ทำการจัดเตรียมข้อมูลให้ชัดเจนก่อนนำไปใช้ในการทำ Data Visualization ต่อไป

2.เรื่องราว (Story): เรื่องราวในที่นี้หมายถึงเรียงร้อยข้อมูลให้ออกมาเป็นลำดับขั้นตอนในการนำเสนอ อย่างการนำเสนอข้อมูลที่เป็นภาพรวมก่อนจะย่อยลงไปข้อมูลย่อยในส่วนต่างๆ และมุมต่างๆ เพื่อให้ผู้อ่านข้อมูลเข้าใจได้ง่าย

3.เป้าหมาย (Goal): จุดประสงค์คือการตั้งคำถามว่า การทำ Data Visualization ครั้งนี้ทำเพื่ออะไร ต้องการหาคำตอบเกี่ยวกับอะไร หรือต้องการนำเสนอข้อมูลในมุมไหน ซึ่งองค์ประกอบนี้ของ Data Visualization ต้องชัดเจน เป็นเหมือนแกนกลางยึดอีกสามองค์ประกอบเอาไว้

4.รูปแบบการนำเสนอ (Visual Form): รูปแบบการนำเสนอ เป็นอีกหนึ่งส่วนประกอบของการทำ Data Visualization เช่น การเลือกนำเสนอในรูปแบบ Infographic ชิ้นเดียว หรือนำเสนอผ่านแดชบอร์ดหลายหน้า ฯลฯ และข้อมูลแต่ละส่วนจะนำเสนอรูปแบบใด เช่น กราฟแท่ง กราฟวงกลม ตาราง ตัวเลขจริง

Data Visualization มีหลากหลายรูปแบบและไม่จำกัดว่าต้องใช้รูปแบบต่อไปนี้ในการนำเสนอข้อมูลเท่านั้น เพราะแต่ละรูปแบบก็มีฟังก์ชันเฉพาะของการนำเสนอข้อมูล บางรูปแบบใช้เปรียบเทียบข้อมูลแต่ละชุดได้ดี บางรูปแบบช่วยให้มองเห็นเทรนด์ได้ง่าย บางรูปแบบช่วยเล่าข้อมูลที่ไกลตัวให้เข้าใจได้ง่ายโดยการเปรียบเทียบให้สอดคล้องกับสิ่งที่คุ้นเคยในชีวิตประจำวัน

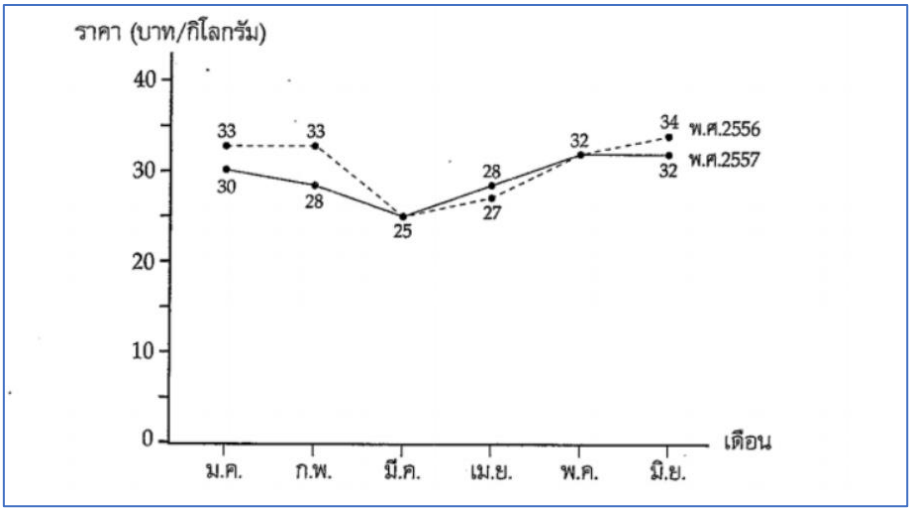
1.ตาราง (Table) เป็นรูปแบบการนำเสนอข้อมูลที่ประกอบด้วย Row หรือแถวแนวนอน และ Column หรือแถวแนวนอน เป็นการนำเสนอข้อมูลในการเปรียบเทียบตัวแปรหลากหลายตัว ใช้นำเสนอข้อมูลจำนวนมากแบบมีโครงสร้างได้

	Jun 2011	Jun 2012	Jun 2013	Jun 2014	Jun 2015	Mar 2016	Mar 2017	Mar 2018	Mar 2019	Mar 2020	Mar 2021	Mar 2022	TTM
Sales +	15,730	20,831	25,581	32,144	36,701	31,136	47,568	50,569	60,427	70,676	75,379	85,651	97,447
Expenses +	13,275	17,135	19,901	24,108	28,215	24,482	37,178	39,323	46,501	53,360	55,331	65,122	75,630
Operating Profit	**2,456**	**3,695**	**5,680**	**8,035**	**8,486**	**6,654**	**10,390**	**11,246**	**13,926**	**17,316**	**20,048**	**20,529**	**21,817**
OPM %	16%	18%	22%	25%	23%	21%	22%	22%	23%	24%	27%	24%	22%
Other Income +	300	206	332	677	1,126	871	1,069	1,230	943	589	927	1,067	1,222
Interest	160	143	106	114	91	74	89	69	174	505	511	319	324
Depreciation	460	549	637	681	404	410	828	1,383	2,073	3,420	4,611	4,326	4,102
Profit before tax	**2,135**	**3,210**	**5,270**	**7,917**	**9,117**	**7,041**	**10,542**	**11,024**	**12,622**	**13,980**	**15,853**	**16,951**	**18,613**
Tax %	23%	24%	23%	18%	20%	20%	18%	21%	20%	21%	30%	20%	
Net Profit	**1,647**	**2,423**	**4,040**	**6,528**	**7,342**	**5,602**	**8,606**	**8,722**	**10,120**	**11,057**	**11,169**	**13,523**	**14,463**
EPS in Rs	5.98	8.74	14.49	23.25	26.02	19.86	30.16	31.32	37.31	40.75	41.07	49.74	53.29
Dividend Payout %	31%	34%	21%	11%	58%	40%	40%	19%	11%	25%	24%	84%	

ภาพที่ 2.1 สรุปผลประกอบการทางการเงิน

(ที่มา: https://www.reddit.com/r/IndianStreetBets/)

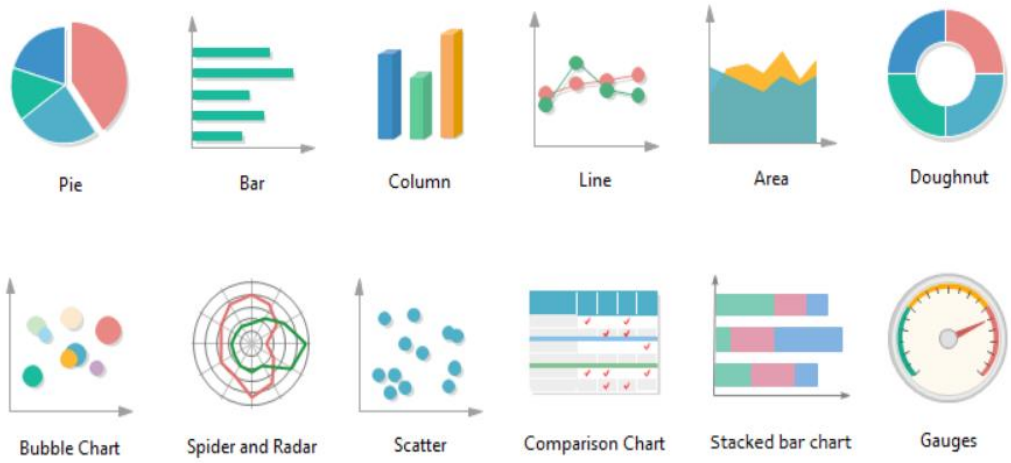
2.กราฟ (Graphs) subset หรือประเภทหนึ่งของแผนภูมิ โดยกราฟจะทำหน้าที่แสดงความสัมพันธ์ระหว่างข้อมูล 2 ตัวแปร ผ่านแกนแนวนอน (แกน X) และแกนแนวตั้ง (แกน Y) ช่วยให้เห็นเทรนด์สถานการณ์ประกอบกับบริบทได้เป็นอย่างดี



ภาพที่ 2.2 แสดงกราฟการเปรียบเทียบราคาส้มเขียวหวานระหว่างปี พ.ศ.2556 – 2567

(ที่มา: https://nockacademy.com/)

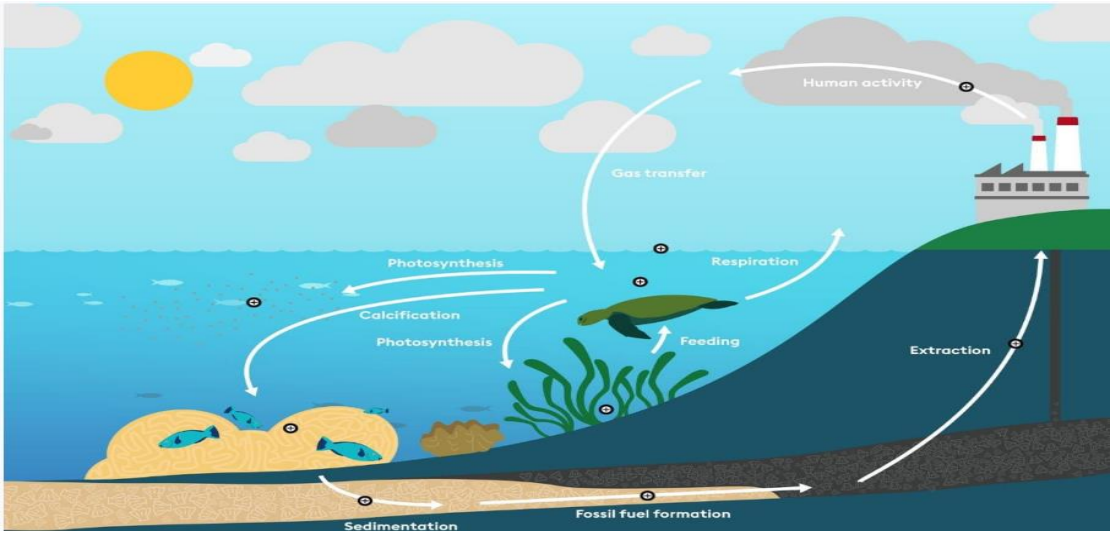
3.แผนภูมิ (Charts) ซึ่งเป็นรูปแบบที่น่าจะคุ้นเคยกันมากที่สุด และเป็นรูปแบบที่มีหลากหลายชนิดที่เหมาะกับการนำเสนอข้อมูลที่แตกต่างกันไปตามวัตถุประสงค์ เช่น Pie chart จะช่วยให้เราเห็นปริมาณความแตกต่างได้ชัดเจน, Comparison chart เหมาะสำหรับการเปรียบเทียบคุณสมบัติหลายๆ ข้อ, มาตรวัด (Gauges) จะช่วยให้เห็นความเข้มข้น ความรุนแรง หรือน้ำหนัก



ภาพที่ 2.3 แสดงตัวอย่างแผนภูมิรูปแบบต่าง ๆ หลายรูปแบบ

(ที่มา: https://www.niwat.blog/graph)

4.อินโฟกราฟิก (Infographic) การนำเสนอสารสนเทศ (Info: information) ด้วยภาพกราฟิก (Graphic) เป็นรูปแบบการนำเสนอข้อมูลที่ใช้ภาพสื่อแทน ทำให้ผู้อ่านข้อมูลเข้าใจข้อมูลได้ง่ายหรือสามารถทำความเข้าใจผ่านภาพแทนที่คุ้นเคย นอกจากนี้ อินโฟกราฟิกยังเป็นรูปแบบการนำเสนอข้อมูลที่น่าสนใจ มีการนำเทคนิคการเล่าเรื่อง (Storytelling) มาใช้ ทำให้ข้อมูลน่าสนใจ น่าดึงดูด จึงมักจะใช้เพื่อนำเสนอเนื้อหา ความรู้ หรือเป็นสื่อการเรียนการสอน



ภาพที่ 2.4 แสดงแบบภาพอินโฟกราฟิก

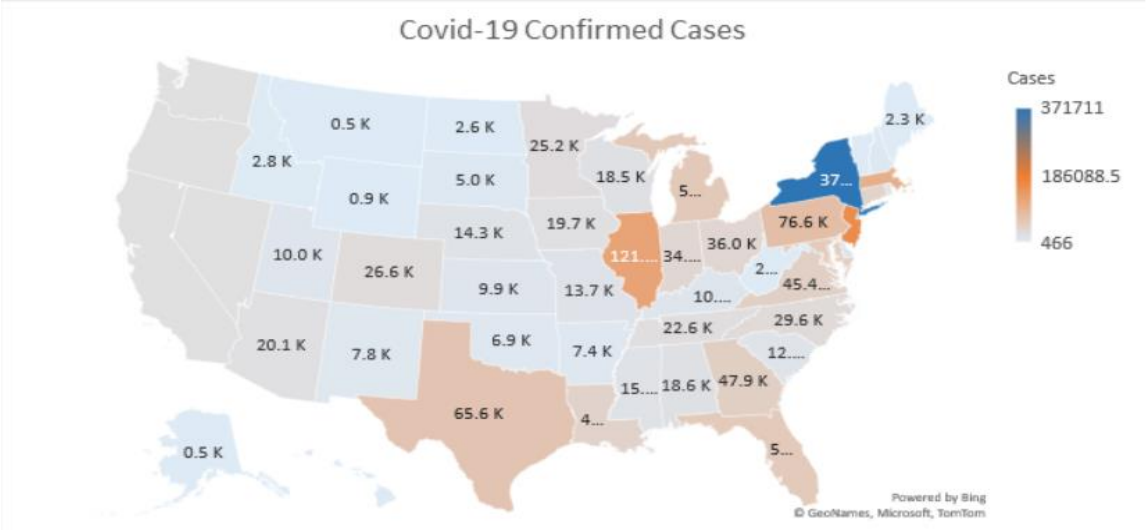
(ที่มา: https://blog.mandalasystem.com/)

5.แดชบอร์ด (Dashboards) การนำข้อมูลต่างๆ มาเรียบเรียงและสรุปเป็นภาพ โดยใช้แผนภูมิและกราฟต่างๆ มาใช้นำเสนอ ปัจจุบันแดชบอร์ดเป็น Data Visualization ที่นิยมใช้กับการนำเสนอข้อมูลแบบ Real-time ผ่านซอฟต์แวร์หรือเครื่องมือจัดการและวิเคราะห์ข้อมูลต่างๆ เช่น เครื่องมือการตลาด เครื่องมือบริหารจัดการข้อมูล เครื่องมือติดตามและดูแลเว็บไซต์

ภาพที่ 2.5 แสดงตัวอย่างหน้าแดชบอร์ด

(ที่มา: https://www.fusionsol.com/)

6.แผนที่ (Maps) เป็นการนำเสนอข้อมูลบนแผนที่เพื่อแสดงข้อมูลเกี่ยวกับพื้นที่ต่างๆ ยกตัวอย่างเช่น การนำเสนอข้อมูลยอดผู้ติดเชื้อ Covid-19 ในแต่ละรัฐของประเทศสหรัฐอเมริกา ซึ่งนอกจากการใส่ข้อมูลลงไปยังพื้นที่ต่างๆ แล้ว ยังสามารถใช้สีสันเพื่อบอกช่วงปริมาณหรือความหนาแน่นของผู้ติดเชื้ออีกด้วย



ภาพที่ 2.6 แสดงตัวอย่างแบบแผนที่

(ที่มา: spreadsheetweb.com)

2.1.5 การทำความสะอาดข้อมูล (Data Cleaning)

Data Cleaning หรือการทำความสะอาดข้อมูล เป็นกระบวนการในการจัดการข้อมูลดิบ (raw data) ที่มีปัญหาหรือข้อผิดพลาด เช่น ข้อมูลที่ขาดหาย ข้อมูลที่ซ้ำซ้อน หรือข้อมูลที่มีรูปแบบไม่สอดคล้องกัน เป้าหมายของการทำ Data Cleaning คือการทำให้ข้อมูลมีความสมบูรณ์ ถูกต้อง และพร้อมสำหรับการวิเคราะห์หรือการใช้ในโมเดล Machine Learning

Cleaning Data มีความสำคัญในการทำธุรกิจจำเป็นจะต้องใช้ข้อมูลที่มีความถูกต้องแม่นยำ มาใช้ในการวิเคราะห์เพื่อให้ได้ผลลัพธ์ที่สมบูรณ์และสามารถนำไปวางแผนทางการตลาดได้หากขาดการเตรียมข้อมูลที่คุณภาพ ไม่มีการคัดกรองข้อมูลด้วย Data Cleaning อาจทำให้การวิเคราะห์ผิดพลาด เกิดการตัดสินใจที่ผิดพลาดและส่งผลกระทบเป็นวงกว้างต่อธุรกิจ ซึ่งข้อดีของ Data Cleansing นั้นมีอีกหลายประการ ไม่ว่าจะเป็น ช่วยให้ได้ Insight หรือรายงาน (Report) ที่แม่นยำ ทำให้ตัดสินใจได้รวดเร็วขึ้น ช่วยให้ดึงข้อมูลออกมาใช้ได้ทันที และข้อมูลอยู่ในรูปแบบที่สมบูรณ์ การทำ Data Cleansing อาจหมายถึง การล้าง

ข้อมูลที่หมดอายุ ซึ่งเกี่ยวข้องกับ พ.ร.บ.ข้อมูลส่วนบุคคล (PDPA) ซึ่งลักษณะของข้อมูลที่ต้องผ่านการ Data Cleaning ก่อนนำไปใช้ประโยชน์ มีดังนี้

1. ข้อมูลที่ไม่ได้อยู่ในรูปแบบเดียวกัน ในกรณีนี้อาจเกิดจากการที่มีข้อมูลจากหลาย Database ทำให้ข้อมูลที่รวบรวมมามีไฟล์คนละนามสกุลกัน เช่น .pdf, .doc, .xls หรือ .pptx เป็นต้น ทำให้ไม่สามารถใช้ในการประมวลผลด้วยกัน จึงต้องมีการแปลงไฟล์ให้อยู่ในนามสกุลเดียวกันเพื่อสามารถประมวลผลได้ และลดพื้นที่ในการจัดเก็บชุดข้อมูล

2. ข้อมูลที่ไม่ได้จัดเก็บในรูปแบบที่ต้องการ เป็นข้อมูลที่ต้องทำให้อยู่ในรูปแบบที่สามารถนำไปใช้ในการวิเคราะห์ได้ บางข้อมูลที่ถูกรวบรวมมาอาจอยู่ในรูปแบบของรูปภาพ เช่น .jpg, .png, .tiff หรือ .bmp จึงต้องมีการแปลงไฟล์รูปภาพให้อยู่ในรูปแบบของไฟล์ข้อความหรือสคริปต์ เช่น .csv, .tsv, .json, และ .xml เป็นต้น

3. ข้อมูลที่ไม่ถูกต้อง โดยส่วนใหญ่แล้วการรวบรวมข้อมูลโดยคนอาจเกิดการผิดพลาด (Human Errors) เป็นเรื่องปกติ เช่น กรอกข้อมูลเกินความจริง กรอกข้อมูลในช่องที่ผิด หรือสะกดชื่อข้อมูลไม่ถูกต้อง ทำให้วิเคราะห์ข้อมูลเชิงลึก หรือ Insight ออกมาผิดพลาดไม่แม่นยำ จึงต้องทำ Data Cleansing เพื่อแก้ไขข้อมูลให้ถูกต้อง

2.2 ทฤษฎี

2.2.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) หรือ การค้นหาองค์ความรู้ในฐานข้อมูล (Knowledge Discovery In Databases หรือ KDD) เป็นเทคนิคการค้นหาองค์ความรู้ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่เพื่อค้นหาแนวโน้ม รูปแบบรวม ความสัมพันธ์หรือความรู้ใหม่อื่น ๆ โดยอาศัยข้อมูลในอดีตความรู้ที่ได้ทำให้เข้าใจและทราบปัจจัยที่ทำ ให้เกิดลักษณธบางอย่างของข้อมูล ซึ่งจะช่วยให้สามารถทำนายแนวโน้มของข้อมูลใหม่ที่จะเกิดขึ้นในอนาคตได้ ผลที่ได้จะมีลักษณะของข้อมูลอยู่ 3 แบบคือข้อมูลแบบที่ไม่ทราบมาก่อน (Unknown) ไม่มีความชัดเจนและไม่สามารถตั้งสมมติฐานก่อนได้ว่าควรเป็นแบบใด ข้อมูลแบบที่มีเหตุผล (Valid) และข้อมูลแบบที่สามารถนำไปใช้ได้ (Actionable)

ขั้นตอนการทำเหมืองข้อมูล (Data Mining)มีรายละเอียดดังนี้

1. การทำความสะอาดข้อมูล (Data Cleaning) เป็นการแก้ไขข้อมูลให้ถูกต้องและสมบูรณ์ เช่น การแก้ไขค่าว่างของข้อมูลโดยอาจใส่ค่า 0 ลงไป หรืออาจไม่นำข้อมูลแถวนั้นมาใช้ในการประมวลผลขึ้นอยู่กับการตัดสินใจของผู้ดูแลระบบ

2. การจัดรูปแบบข้อมูล (Data Transaction Identification) เป็นการจัดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมก่อน ที่นิยมใช้กันมากคือการทำข้อมูลให้อยู่ในรูปแบบตาราง (Table) มีลักษณะเป็น “แถว” และ “คอลัมน์” ที่สัมพันธ์กัน

3. การรวบรวมข้อมูล (Data Integration) เป็นการรวบรวมข้อมูลทั้งหมดที่ต้องการ ซึ่งอาจอยู่ในหลายฐานข้อมูลหลายระบบปฏิบัติการให้อยู่ในฐานข้อมูลเดียวกันหรือ ตารางเดียวก่อน อาจใช้ในลักษณะของคลังข้อมูล (Data Warehouses) ในการรวบรวมข้อมูล

4. การแปลงข้อมูล (Data Transformation) เป็นการปรับเปลี่ยนข้อมูลให้มีค่าที่เหมาะสมในการตัดสินใจ เช่น ข้อมูลของสินค้าเป็นข้อมูลที่มีค่า “Coke” และ “Pepsi” มีการเปลี่ยนค่าให้เป็น “น้ำอัดลม” เพื่อความเหมาะสมในการตัดสินใจมากขึ้น

5. การค้นหารูปแบบ (Pattern Discovery) เป็นการกำหนดรูปแบบในการค้นหา เพื่อให้ได้ผลลัพธ์ที่ต้องการสามารถแบ่งเป็นรูปแบบการวิเคราะห์ (Path Analysis), กฎที่สัมพันธ์กัน 6 (Association Rules), รูปแบบการทำงานตามลำดับ (Sequential Patterns), การจัดกลุ่มและการจำแนกกฎ (Cluster & Classification Rules) เป็นต้น

6. การวิเคราะห์รูปแบบ (Patten Analysis) เป็นรูปแบบการนำผลลัพธ์จากการค้นหามาทำการวิเคราะห์เพื่อช่วยในการตัดสินใจหรือการวางแผนทางธุรกิจ

2.2.2 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์

ชัยมงคล เทพวงษ์(2550) การออกแบบเว็บไซต์ที่มีประสิทธิภาพนั้นต้องคำนึงถึงองค์ประกอบสำคัญดังต่อไปนี้

1. ความเรียบง่าย (Simplicity) หมายถึง การจากองค์ประกอบเสริมให้เหลือเฉพาะองค์ประกอบหลัก กล่าวคือในการสื่อสารเนื้อหากบผู้ใช้นั้นเราต้องเลือกเสนอสิ่งที่เราต้องการนำเสนอจริง ๆ ออกมาในส่วนของ กราฟิกสีสันตัวอักษรและภาพเคลื่อนไหวต้องเลือกให้พอเหมาะ ถ้าหากมีมากเกินไปจะรบกวนสายตาและสร้างความคำราญต่อผู้ใช้ตัวอย่างเว็บไซต์ที่ได้รับการออกแบบที่ดี ได้แก่ เว็บไซต์ของ บริษัทใหญ่ ๆ อยางเช่น Apple Adobe Microsoft หรือ Kokia ที่มีการออกแบบเว็บไซต์ในรูปแบบที่ เรียบง่าย ไม่ซับซ้อน และใช้งานอย่างสะดวก

2. ความสม่ำเสมอ (Consistency) หมายถึง การสร้างความสม่ำเสมอให้เกิดขึ้นตลอดทั้งเว็บไซต์ โดยอาจเลือกใช้รูปแบบเดียวกันตลอดทั้งเว็บไซต์ก็ได้ เพราะถ้าหากว่าแต่ละหน้าในเว็บไซต์นั้นมีความแตกต่างกันมากจนเกินไปอาจทำให้ผู้ใช้เกิดความสับสนและไม่แน่ใจว่ากำลังอยู่ในเว็บไซต์เดิมหรือไม่ เพราะฉะนั้นการออกแบบเว็บไซต์ในแต่ละหน้าควรที่จะมีรูปแบบสไตล์ของกราฟิก ระบบเนวิเกชั่น (Navigation) และโทนสีที่มีความคล้ายคลึงกันตลอดทั้งเว็บไซต์

3. ความเป็นเอกลักษณ์ (Identity) ในการออกแบบเว็บไซต์ต้องคำนึงถึงลักษณะขององค์กรเป็นหลัก เนื่องจากเว็บไซต์จะสะท้อนถึงเอกลักษณ์และลักษณะขององค์กร การเลือกใช้ตัวอักษร ชุดสี รูปภาพหรือกราฟิก จะมีผลต่อรูปแบบของเว็บไซต์เป็นอย่างมาก ตัวอย่างเช่น ถ้าเราต้องออกแบบเว็บไซต์ของธนาคารแต่เรากลับเลือกสีสันและกราฟิกมากมายอาจทำให้ผู้ใช้คิดว่าเป็นเว็บไซต์ของสวนสนุกซึ่งส่งผลต่อความเชื่อถือขององค์กรได้

4. เนื้อหา (Useful Content) ถือเป็นสิ่งสำคัญที่สุดในเว็บไซต์เนื้อหาในเว็บไซต์ต้องสมบูรณ์และได้รับการปรับปรุงพัฒนาให้ทันสมัยอยู่เสมอผู้พัฒนาต้องเตรียมข้อมูลและเนื้อหาที่ผู้ใช้ต้องการให้ถูกต้องและ สมบูรณ์เนื้อหาที่สคำญที่สุดคือเนื้อหาที่ทีมผู้พัฒนาสร้างสรรค์ขึ้นมาเอง และไม่ไปซ้ำกับเว็บอื่น เพราะจะถือเป็นสิ่งที่ดึงดูดผู้ใช้ให้เข้ามาเว็บไซต์ได้เสมอ แต่ถ้าเป็นเว็บที่ลิงค์ข้อมูลจากเว็บอื่น ๆ มาเมื่อใดก็ตามที่ผู้ใช้ทราบว่า ข้อมูลนั้นมาจากเว็บใดผู้ใช้ก็ไม่จำเป็นต้องกลับมาใช้งานลิงค์เหล่านั้นอีก

5. ระบบเนวิเกชัน (User–Friendly Navigation) เป็นส่วนประกอบที่มีความสำคัญต่อเว็บไซต์มาก เพราะจะช่วยไม่ให้ผู้ใช้เกิดความสับสนระหว่างดูเว็บไซต์ระบบเนวิเกชั่นจึงเปรียบเสมือนป้ายบอกทางดังนั้นการออกแบบเนวิเกชั่นจึงควรให้เข้าใจง่าย ใช้งานได้สะดวก ถ้ามีการใช้กราฟิกก็ควรสื่อความหมายตำแหน่งของการวางเนวิเกชั่นก็ควรวางให้สม่ำเสมอ เช่น อยู่ตำแหน่งบนสุดของทุกหน้าเป็นต้น ซึ่งถ้าจะให้ดีเมื่อมีเนวิเกชั่นที่เป็นกราฟิกก็ควรเพิ่มระบบเนวิเกชั่นที่เป็นตัวอักษรไว้ส่วนล่างด้วย เพื่อช่วยอำนวยความสะดวกให้กับผู้ใช้ที่ยกเลิกการแสดงผลภาพกราฟิกบนเว็บเบราเซอร์

6. คุณภาพของสิ่งที่ปรากฏให้เห็นในเว็บไซต์ (Visual Appeal) ลักษณะที่น่าสนใจของเว็บไซต์นั้น ขึ้นอยู่กับความชอบส่วนบุคคลเป็นสำคัญ แต่โดยรวมแล้วก็สามารถสรุปได้ว่าเว็บไซต์ที่น่าสนใจนั้นส่วนประกอบต่าง ๆ ควรมีคุณภาพ เช่น กราฟิกควรสมบูรณ์ไม่มีรอยหรือขอบขั้นบันได้ให้เห็นชนิดตัวอักษรอ่านง่ายสบายตามีการเลือกใช้โทนสีที่เข้ากันอย่างสวยงาม เป็นต้น

7. ความสะดวกของการใช้ในสภาพต่าง ๆ (Compatibility) การใช้งานของเว็บไซต์นั้นไม่ควรมีขอบจำกัด กล่าวคือ ต้องสามารถใช้งานได้ดีในสภาพแวดล้อมที่หลากหลายไม่มีการบังคับให้ผู้ใช้ต้องติดตั้งโปรแกรมอื่นใดเพิ่มเติมนอกเหนือจากเว็บบราวเซอร์ควรเป็นเว็บที่แสดงผลได้ดีในทุกระบบปฏิบัติการสามารถแสดง ผลได้ในทุกความละเอียดหน้าจอ ซึ่งหากเป็นเว็บไซต์ที่มีผู้ใช้บริการมากและกลุ่มเป้าหมายหลากหลายควรให้ความสำคัญกบเรื่องนี้ให้มาก

8. ความคงที่ในการออกแบบ (Design Stability) ถ้าต้องการให้ผู้ใช้งานรู้สึกวาเว็บไซต์มีคุณภาพ ถูกต้อง และเชื่อถือได้ควรให้ความสำคัญกับการออกแบบเว็บไซต์เป็นอย่างมาก

ต้องออกแบบวางแผนและเรียบเรียงเนื้อหาอยางรอบคอบ ถ้าเว็บที่จัดขึ้นลวก ๆ ไม่มีมาตรฐานการออกแบบและระบบการจัดการข้อมูล ถ้ามีปัญหามากขึ้นอาจส่งผลให้เกิดปัญหาและทำให้ผู้ใช้หมดความเชื่อถือ

9. ความคงที่ของการทำงาน (Function Stability) ระบบการทำงานต่าง ๆ ในเว็บไซต์ควรมีความถูกต้องแน่นอน ซึ่งต้องได้รับการออกแบบสร้างสรรค์และตรวจสอบอยู่เสมอ ตัวอย่างเช่น ลิงค์ต่าง ๆ ในเว็บไซต์ ต้องตรวจสอบวายังสามารถลิงค์ข้อมูลได้ถูกต้องหรือไม่ เพราะเว็บไซต์อื่นอาจมีการเปลี่ยนแปลงได้ตลอดเวลา ปัญหาที่เกิดจากลิงค์ก็คือลิงค์ขาด ซึ่งพบได้บ่อยเป็นปัญหาที่สร้างความรำคาญกับผู้ใช้เป็นอย่างมาก

โครงสร้างเว็บไซต์ (Site Structure) เป็นแผนผังของการลำดับเนื้อหาหรือการจัดวางตำแหน่งเว็บเพจทั้งหมด ซึ่งจะทำให้เรารู้ว่าทั้งเว็บไซต์ประกอบไปด้วยเนื้อหาอะไรบ้าง และมีเว็บเพจหน้าไหนที่เกี่ยวข้องเชื่อมโยงถึงกัน ดังนั้นการออกแบบโครงสร้างเว็บไซต์จึงเป็นเรื่องสำคัญ เปรียบเสมือนกับการเขียนแบบอาคารก่อนที่จะลงมือสร้าง เพราะจะทำให้เรามองเห็นหน้าตาของเว็บไซต์เป็นรูปธรรมมากขึ้นสามารถออกแบบระบบเนวิเกชันได้เหมาะสมและเป็นแนวทางการทำงานที่ชัดเจนสำหรับขั้นตอนต่อ ๆ ไป นอกจากนี้โครงสร้างเว็บไซต์ที่ดียังช่วยให้ผู้ชมไม่สับสนและค้นหาข้อมูลที่ต้องการได้อย่างรวดเร็ว โดยโครงสร้างของเว็บไซต์ส่วนใหญ่ก็จะประกอบไปด้วย 4 รูปแบบดังนี้

1.โครงสร้างเว็บไซต์แบบเรียงลำดับ (Sequential Structure) โครงสร้างเว็บไซต์ที่จะนำเสนอเนื้อหาเป็นลำดับๆ ทีละหัวข้อๆ ซึ่งบ้างเรียกว่า Sequential Structure หรือโครงสร้างแบบตามลำดับวิธีกาออกแบบจะเริ่มจาก Main Page หรือหน้า Home ซึ่งเป็นหน้าแรกที่เจ้าของเว็บไซต์อยากให้ผู้คนเข้ามาเจอก่อนจากนั้นเมนูหลักของเว็บไซต์ Navigator จะพาไปดูเว็บเพจต่างๆ ไปตามลำดับ โครงสร้างเว็บไซต์ประเภทนี้ เหมาะกับสินค้าหรือบริการที่นำเสนอเป็นลำดับขั้น 1-2-3 ไปเรื่อย ๆ จนจบ เช่น Online Course ที่จะไล่เรียงจากเนื้อหาบทที่ 1, 2, 3 ต่อไปเรื่อย ๆ หรือ e book

ภาพที่ 2.7 แสดงโครงสร้างเว็บไซต์แบบเรียงลำดับ

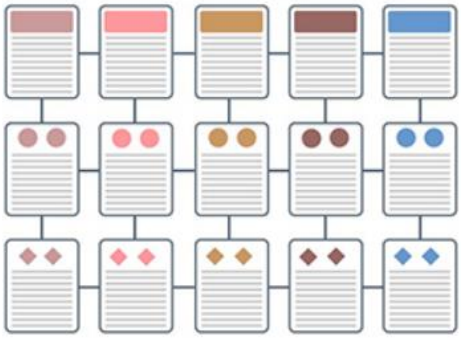
(ที่มา: https://enfete.co.th/th/website-structure/)

2.โครงสร้างแบบลำดับขั้น (Hierarchical Structure) โครงสร้างเว็บไซต์ที่แบ่งเนื้อหาออกเป็นหมวดหมู่และย่อยหมวดหมู่ตามลำดับชั้น โดยหน้าหลักจะเป็นหน้าแรกที่ผู้ใช้จะเห็น จากนั้นจึงจะมีหน้าย่อยและหน้าย่อยย่อยตามลำดับ โครงสร้างแบบลำดับชั้นช่วยให้ผู้ใช้สามารถเข้าใจเนื้อหาของเว็บไซต์ได้ง่ายขึ้น โดยสามารถเลื่อนลงตามลำดับชั้น เพื่อค้นหาเนื้อหาที่ต้องการ

ภาพที่ 2.8 แสดงโครงสร้างแบบลำดับขั้น

(ที่มา: https://enfete.co.th/th/website-structure/)

3.โครงสร้างแบบตาราง (Grid Structure) โครงสร้างเว็บไซต์ที่มีความซับซ้อนแต่ก็มีความยืดหยุ่นในระดับหนึ่ง เพื่อให้ผู้ใช้งานสามารถเข้าสู่เนื้อหาต่าง ๆ ได้ง่ายขึ้น การออกแบบในลักษณะนี้จะมีการเชื่อมโยงเนื้อหาในแต่ละส่วนเข้าหาซึ่งกันและกัน ทำให้ผู้ใช้งานสามารถเปลี่ยนทิศทาง หรือกำหนดทิศทางในการเข้าสู่เนื้อหาด้วยตัวเองได้ จึงไม่ทำให้เสียเวลาแถมยังทำให้เว็บไซต์มีความทันสมัยขึ้น



ภาพที่ 2.9 แสดงโครงสร้างแบบตาราง

(ที่มา: https://enfete.co.th/th/website-structure/)

4.โครงสร้างแบบใยแมงมุม (Web Structure) โครงสร้างเว็บไซต์ที่เชื่อมโยงกันทุกหน้าโดยไม่คำนึงถึงลำดับชั้น โครงสร้างแบบใยแมงมุมช่วยให้ผู้เข้าชมเว็บไซต์สามารถค้นหาข้อมูลที่ต้องการได้อย่างอิสระ แต่อาจทำให้ยากต่อการติดตามลำดับของข้อมูล

ภาพที่ 2.10 แสดงโครงสร้างแบบใยแมงมุม

(ที่มา: https://enfete.co.th/th/website-structure/)

การใช้สีในการออกแบบเว็บไซต์ การสร้างสีสันบนหน้าเว็บเป็นสิ่งที่สื่อความหมายของเว็บไซต์ได้อย่างชัดเจนการเลือกใช้สีให้เหมาะสม กลมกลืนไม่เพียงแต่จะสร้างความพึงพอใจให้กบผู้ใช้แต่ยังสามารถทำให้เห็นถึงความแตกต่างระหว่างเว็บไซต์ได้ สีเป็นองค์ประกอบหลักสำหรับการตกแต่งเว็บ จึงจำเป็นอยางยิ่งที่จะต้องทำความเข้าใจเกี่ยวกับการใช้สี ระบบสีที่แสดงบนจอคอมพิวเตอร์ มีระบบการแสดงผลผานหลอดลำแสงที่เรียกว่า CRT (Cathode ray tube) โดยมีลักษณะระบบสีแบบพวกอาศัยการผสมของของแสงสีแดง สีเขียว และสีน้ำเงิน หรือระบบสี RGB สามารถกำหนดค่าสีจาก 0 ถึง 255 ได้ จากการรวมสีของแม่สีหลักจะทำให้เกิดแสงสีขาวมีลักษณะเป็นจุดเล็ก ๆ บนหน้าจอไม่สามารถมองเห็นด้วยตาเปล่าได้ จะมองเห็นเป็นสีที่ถูกผสมเป็นเนื้อสีเดียวกันแล้วจุดตาละจุดหรือพิกเซล (Pixel) เป็นส่วนประกอบของภาพบนหน้าจอคอมพิวเตอร์ โดยจำนวนบิตที่ใช้ในการกำหนดความสามารถของการแสดงสีต่าง ๆ เพื่อ สร้างภาพบนจอนั้นเรียกว่า บิตเด็ป (Bit-depth) ในภาษา HTML มีการกำหนดสีด้วยระบบเลขฐานสิบหก ซึ่งมีเครื่องหมาย (#) อยู่ด้านหน้าและตามด้วยเลขฐานสิบหกจำนวนอักษรอีก 6 หลัก โดยแต่ละไบต์ (byte) จะมีตัวอักษรสองตัว แบ่งออกเป็น 3 กลุ่ม เช่น #FF12AC การใช้ตัวอักษรแต่ ละไบต์นี้เพื่อกาหนดระดับความเข้มของแม่สีแต่ละสีของชุดสี RGB โดย 2 หลักแรกแสดงถึงความ เข้มของสีแดง 2 หลักต่อมา แสดงถึงความเข้มของสีเขียว 2 หลักสุดท้ายแสดงถึงความเข้มของสีน้ำเงินสีมีอิทธิพลในเรื่องของอารมณ์การสื่อความหมายที่เด่นชัดกระตุ้นการ

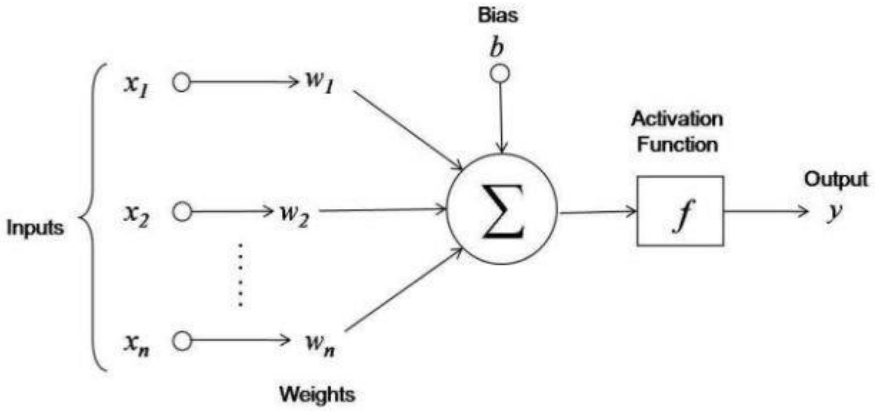
รับรู้ทางด้านจิตใจมนุษย์ สีแต่ละสีให้ความรู้สึกอารมณ์ที่ไม่เหมือนกันสีบางสีให้ความรู้สึกสงบ บางสีให้ความรู้สึกตื่นเต้นรุนแรง สีจึงเป็นปัจจัยสำคัญอย่างยิ่งต่อการออกแบบเว็บไซต์ ดังนั้น การเลือกใช้โทนสีภายในเว็บไซต์เป็นการแสดงถึงความแตกต่างของสี ที่แสดงออกทางอารมณ์ มีชีวิตชีวาหรือเศร้าโศกรูปแบบของสีที่สายตาของมนุษย์มองเห็น สามารถแบ่งออกเป็น 3 กลุ่ม คือ 1. สีโทนร้อน (Warm Colors) เป็นกลุ่มสี ที่แสดงถึงความสุข ความปลอบโยน ความ อบอุ่น และดึงดูดใจ สีกลุ่มนี้เป็นกลุ่มสีที่ช่วยให้หายจากความเฉื่อยชา มีชีวิตชีวามากยิงขึ้น 2. สีโทนเย็น (Cool Colors) แสดงถึงความที่ดูสุภาพ อ่อนโยน เรียบร้อย เป็นกลุ่มสีที่มีคนชอบมากที่สุด สามารถโน้มนาวในระยะไกลได้ 3. สีโทนกลาง (Neutral Colors) สีที่เป็นกลางประกอบด้วย สีดำ สีขาว สีเทา และสีน้ำตาล กลุ่มสีเหล่านี้คือสีกลางที่สามารถนำไปผสมกบสีอื่น ๆ เพื่อให้เกิดสีกลางขึ้นมาสิ่งที่สำคัญต่อผู้ออกแบบเว็บคือการเลือกใช้สีสำหรับเว็บนอกจากจะมีผลต่อการแสดงออกของเว็บแล้วยังเป็นการสร้างความรู้สึกที่ดีต่อผู้ใช้บริการ ดังนั้นจะเห็นว่าสีแต่ละสีสามารถสื่อความหมายของเว็บได้อย่างชัดเจนความแตกต่างความสัมพันธ์ที่เกิดขึ้นยอมส่งผลให้เว็บมีความน่าเชื่อถือมากยิ่งขึ้น ชุดสีแต่ละชุดมีความสำคัญต่อเว็บถ้าเลือกใช้สีไม่ตรงกับวัตถุประสงค์หรือเป้าหมายอาจจะทำให้เว็บไม่น่าสนใจผู้ใช้บริการจะไม่กลับมาใช้บริการอีกภายหลัง ฉะนั้นการใช้สีอย่างเหมาะสมเพื่อสื่อความหมายของเว็บต้องเลือกใช้สีที่มีความกลมกลืนกัน

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

2.3.1 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

เทคนิคโครงข่ายประสาทเทียม เป็นการศึกษาระบบการทำงานของเซลล์ประสาทภายในสมองที่ประกอบด้วยเซลล์ประสาท (Neuron) และเส้นประสาทโดยที่เซลล์ประสาทจะเชื่อมต่อกันในรูปแบบโครงข่าย ซึ่งการวิเคราะห์ และประมวลผลข้อมูลของระบบประสาทนั้นจะส่งข้อมูลผ่านระบบโครงข่ายของเซลล์ประสาท และทำงานในลักษณะขนานคือ ทำกิจกรรมหรืองานหลายอย่างได้ในเวลาเดียวกันให้ได้มาซึ่งผลลัพธ์ที่ต้องการ โดยการทำงานของสมองในรูปแบบที่กล่าวมาในข้างต้นนั้นมีความสามารถหลายประการเช่น การสังเกต เรียนรู้ จดจำ ทำซ้ำและแยกแยะสิ่งต่าง ๆ ซึ่งโครงข่ายประสาทเทียมได้จำลองรูปแบบการท้างาน และโครงสร้างการเชื่อมต่อดังกล่าวมา เพื่อท้าให้ระบบคอมพิวเตอร์สามารถท้างานที่สมองทำได้ซึ่งสามารถเพิ่มประสิทธิภาพในการทำงานให้มากขึ้น และช่วยลดข้อผิดพลาดที่เกิดขึ้นจากการทำงาน โครงข่ายประสาทเทียมถูกสร้างขึ้นเพื่อการจำลองลักษณะการประมวลผลของระบบประสาทมนุษย์ด้วยแบบจำลองเชิงคณิตศาสตร์และสถิติ (Mathematical and Statistical Model) ซึ่งประกอบด้วยส่วนของการประมวลผลที่เรียกว่า นิวรอน (Neuron) ทุกๆ นิวรอนสามารถมี

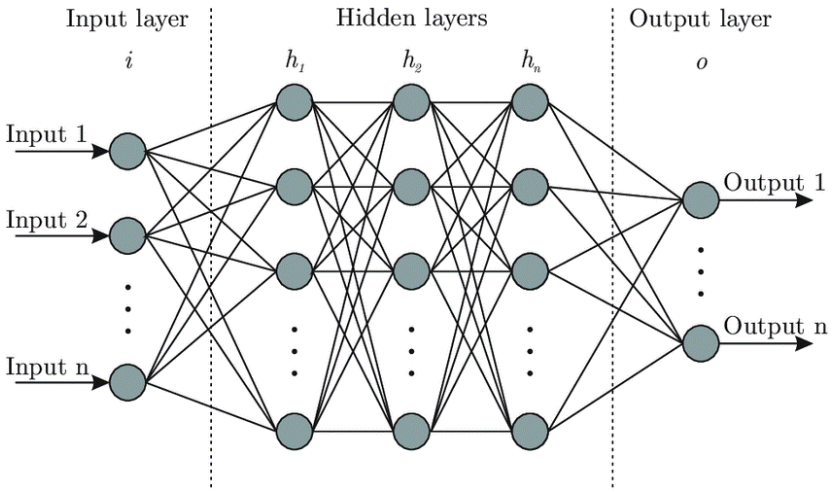
ข้อมูลป้อนเข้า (Input) ได้หลายค่า แต่ข้อมูลส่งออก (Output) มีได้เพียงค่าเดียว และทุกๆข้อมูลส่งออก (Output) จะเชื่อมโยงไปยังข้อมูลป้อนเข้าของนิวรอน (Input) อื่นๆภายในโครงข่าย สำหรับการเชื่อมโยงกันภายในระหว่างนิวรอนทุกๆข้อมูลป้อนเข้า (Input) จะมีค่าน้ำหนัก (Bias) เป็นตัวกำหนดกำลังของการเชื่อมโยงภายในนิวรอนจะมีฟังก์ชันกำหนดสัญญาณส่งออกที่เรียกว่า ฟังก์ชันถ่ายโอน (Transfer Function)



ภาพที่ 2.11 Artificial Neural Network

(ที่มา: weather4thai.kmitl.ac.th)

โครงข่ายประสาทเทียมประกอบด้วยนิวรอนจำนวนมากเชื่อมต่อกัน ซึ่งการเชื่อมต่อแบ่งออกเป็นกลุ่มย่อย เรียกว่า ชั้น (Layer) ชั้นแรกเป็นชั้นข้อมูลป้อนเข้า (Input Layer) ชั้นสุดท้ายเป็นชั้นข้อมูลส่งออก (Output Layer) ส่วนชั้นที่อยู่ ระหว่างชั้นข้อมูลป้อนเข้าและชั้นข้อมูลส่งออก เรียกว่าชั้นซ่อน (Hidden Layer) ซึ่งโดยทั่วไปชั้นซ่อนอาจมีมากกว่า 1ชั้นก็ได้ ด้วยเหตุนี้จึงสามารถแบ่งประเภทตามโครงสร้างของโครงข่ายประสาทเทียมได้2แบบ คือ โครงข่ายประสาทเทียมแบบ ชั้นเดียว (Single Layer) ดังแสดงในรูปที่ 5 และโครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer)



ภาพที่ 2.12 Multilayer

(ที่มา: weather4thai.kmitl.ac.th)

โดยทั่วไปการทำงานของโครงข่ายประสาทเทียมก็คือ การสอนให้โครงข่ายทำการคำนวนข้อมูลส่งออก (Output) พร้อมกับการปรับปรุงค่าน้ำหนัก (Bias) โดยอาศัยกระบวนการทำซ้ำ (Iterative) แบ่งออกเป็น 3 ประเภท คือ

1.การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบมีผู้สอน (Supervised Learning) คือการสอนโครงข่ายโดยใช้ข้อมูลป้อนเข้า(Input)และข้อมูลส่งออก (Output)เป็นชุดฝึกสอนควบคู่(Training pair) โดยการสอนโครงข่ายนั้นจะใช้ชุดฝึดสอนหลายคู่จึงทำให้ข้อมูลส่งออกจริงกับข้อมูลส่งออก(Output)มีความคลาดเคลื่อนกัน โดยโครงข่ายจะต้องมีการปรับค่าน้ำหนัก(Bias)เพื่อลดค่าความแตกต่าง(Error)ระหว่างข้อมูลส่งออกจริงกับข้อมูลส่ง(Output)

2.การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือการสอนโครงข่ายโดยใช้ข้อมูลป้อนเข้า(Input)และหลักการทางสถิติหาค่าทางสถิติของชุดฝึกสอน ทำการจัดกลุ่มข้อมูลออกเป็นระดับต่างๆโดยโครงข่ายประสาทเทียมจะหาค่าข้อมูลออก(Output)

3.การเรียนรู้เชิงบังคับ (Reinforcement Learning) การเรียนรู้เชิงบังคับ (Reinforcement Learning) คือการเรียนรู้แบบมีผู้สอนและไม่มีผู้สอนโดยจะใช้การเรียนรู้แบบไม่มีผู้สอนสำหรับข้อมูลป้อนเข้า (Input) และจะใช้การเรียนรู้แบบมีผู้สอนเมื่อได้ข้อมูลส่งออก (Output)แล้วสำหรับฟังก์ชันถ่ายโอน (Transfer Function) หรือฟังก์ชันการกระตุ้น (Activation Function)ถูกแบ่งเป็น 4 ประเภทได้แก่ (1) ฟังก์ชันเชิงเส้น (Linear function) (2) ฟังก์ชันไม่เชิงเส้น (Non-linear function) (3) ฟังก์ชันสมมาตร (Symmetrical function) และ (4) ฟังก์ชันไม่

สมมาตร (Non-symmetrical function) ฟังก์ชัน Sigmoid เป็นฟังก์ชันการกระตุ้น (Activation Function) แบบฟังก์ชันไม่เชิงเส้น(Non-linear function) ซึ่งจะมีฟังก์ชันอยู่ในรูปแบบ

$$f(x) = \frac{1}{1+e^{-x}}$$

โครงข่ายประสาทเทียม (ANN) ถูกแบ่งตามการเชื่อมต่อของปมประสาทและฟังก์ชันกระตุ้นแบ่งออกเป็น 5 ประเภท 1.ประเภทการคาดเดา (Prediction) คือการใช้ข้อมูลนำเข้า (Input) เพื่อเดาข้อมูลส่งออก(Output) เช่น Back-propagation, Delta Bar Delta, Extended Delta Bar Delta, Directed Random Search, Higher Order Neural Networks และ Self-organizing map into Back-propagation 2.ประเภทการจัดหมวดหมู่ (Classification) คือการใช้ข้อมูลนำเข้าเพื่อกำหนดการจัดหมวดหมู่ เช่น Learning Vector Quantization, Counter-propagation และ Probabilistic Neural Networks 3.ประเภทการเชื่อมโยงข้อมูล (Data Association) คือการใช้ข้อมูลนำเข้าเพื่อกำหนดการจัดหมวดหมู่แต่จะจดจำข้อมูลที่มีค่า Error เช่น Hopfield, Boltzamann Machine, Hamming Network และ Bidirectional associative Memory 4.ประเภทกระบวนการสร้างความคิด(Data Conceptualization) คือการวิเคราะห์ข้อมูลนำเข้า(Input)เพื่อจัดกลุ่ม เช่น Adaptive Resonance Network และ Self Organization Map 5.ประเภทการกลั่นกรองข้อมูล (Data Filtering) คือการทำให้ข้อมูลนำเข้ามีความสม่ำเสมอเช่น Recirculation

2.3.2 การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression Analysis)

การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติอย่างหนึ่งที่ใช้ในการตรวจสอบลักษณะของความสัมพันธ์ระหว่างตัวแปร ตั้งแต่ 2 ตัวขึ้นไป โดยแบ่งเป็นตัวแปรอิสระ (Independent variable) และตัวแปรตาม (Dependentvariable) ผลของการศึกษาจะให้ทราบถึง

1. ขนาดของความสัมพันธ์ระหว่างตัวแปรอิสระ ที่มีตอตัวแปรตาม
2. แบบจำลองความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม

ในการวิเคราะห์การถดถอย มักเรียกตัวแปรอิสระวาตัวทำนาย (predictor) หรือตัวแปรกระตุ้น (stimulus variable) สวนตัวแปรตามมักเรียกวา ตัวแปรตอบสนอง (response variable) หรือตัวแปรเกณฑ์ (criterion variable)

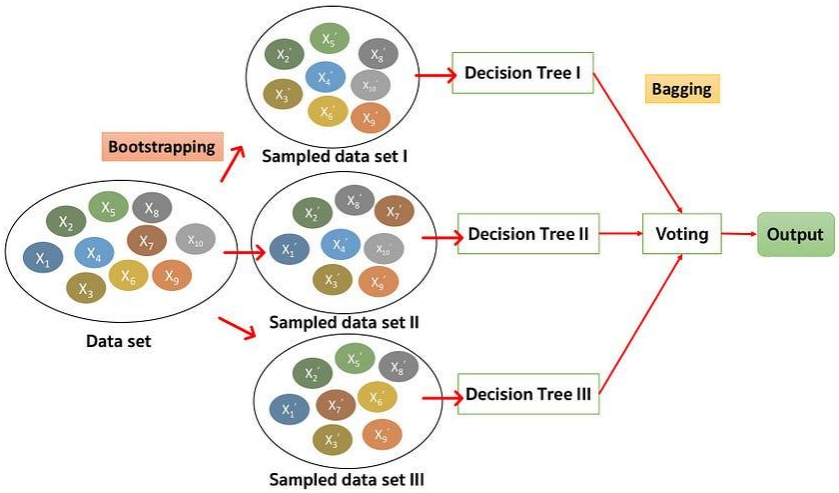
ชนิดของการวิเคราะห์การถดถอย การวิเคราะห์การถดถอยมีหลายชนิดขึ้นกับลักษณะของตัวแปรตาม รูปแบบ ความสัมพันธ์ และการกำหนดตัวแปรอิสระ (ตัวแปรตน) ซึ่งโดยทั่วไปแบงการวิเคราะห์การถดถอยได้เป็น 2 ประเภท คือ

– การวิเคราะห์การถดถอยเชิงเสน (Linear regression analysis) เป็นการวิเคราะห์การถดถอยที่ตัวแปรอิสระส่วนใหญ่เป็นตัวแปรเชิงปริมาณ ส่วนตัวแปรตามจะต้องเป็นตัวแปรเชิงปริมาณเท่านั้น รูปแบบของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม สามารถแทนได้ด้วยสมการทางคณิตศาสตร์ที่เป็นเชิงเส้น (Linear model)

– การวิเคราะห์การถดถอยแบบไม่เป็นเชิงเสน (Non linear regression) เป็นการวิเคราะห์การถดถอย ที่รูปแบบของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม สามารถแทนได้ด้วยสมการทางคณิตศาสตร์ที่ไม่เป็นเชิงเสน (non – Linear model)

2.3.3 เทคนิคแรนดอมฟอเรสต์ (Random Forest)

Random Forest คือ Algorithm การเรียนรู้ของเครื่อง (Machine Learning) ที่เกิดจากการรวม Decision Tree หลาย ๆ ต้นเข้าด้วยกัน โดยแต่ละต้นจะถูกสร้างขึ้นจากคุณลักษณะของข้อมูล (Feature) ที่สุ่มมาเพียงบางส่วน Random Forest เป็นเทคนิคการรวม Model หลาย ๆ Model เพื่อสร้าง Model ที่มีประสิทธิภาพสูงขึ้น (Ensemble Learning) ซึ่งใช้หลักการของ Bagging (Bootstrap Aggregating) เพื่อเพิ่มความแม่นยำและลดปัญหา Overfitting โดยการสุ่มตัวอย่าง (Random Sampling) แบบใส่คืนเพื่อสร้างชุดข้อมูลย่อยหลายชุดจากชุดข้อมูลต้นฉบับ ทำให้แต่ละชุดข้อมูลย่อยอาจมีตัวอย่างซ้ำกันได้ สำหรับนำมาสร้าง Model หลาย ๆ Model โดยแต่ละ Model จะใช้ชุดข้อมูลย่อยที่สุ่มมาในการสร้าง



ภาพที่ 2.13 หลักการทำ Random Forest

(ที่มา: medium.com)

หลักการของ Random Forest คือ สร้าง model จาก Decision Tree หลายๆ model ย่อย ๆ (ตั้งแต่ 10 model ถึง มากกว่า 1000 model) โดยแต่ละ model จะได้รับ data set ไม่เหมือนกัน ซึ่งเป็น subset ของ data set ทั้งหมด ตอนทำ prediction ก็ให้แต่ละ Decision Tree

ทำ prediction ของใครของมัน และคำนวณผล prediction ด้วยการ vote output ที่ ถูกเลือกโดย Decision Tree มากที่สุด (กรณี classification) หรือ หาค่า mean จาก output ของแต่ละ Decision Tree (กรณี regression)

2.3.4 เทคนิคต้นไม้เสริมกำลังแบบไล่ระดับ (Gradient Boosted Trees)

Gradient Boosted Trees (GBTs) หรือที่เรียกว่า Gradient Boosting Decision Trees (GBDT) เป็นเทคนิคการเรียนรู้ของเครื่องที่ใช้ในการสร้างแบบจำลองการทำนายโดยการรวมเอาโมเดลการทำนายที่อ่อนแอหลาย ๆ โมเดลเข้าด้วยกัน โดยทั่วไปแล้ว โมเดลเหล่านี้มักเป็นต้นไม้ตัดสินใจ (Decision Trees) ที่มีความซับซ้อนต่ำ GBTs ถูกนำมาใช้ในงานต่าง ๆ เช่น การจัดอันดับในเครื่องมือค้นหาเว็บ การวิเคราะห์ข้อมูลในฟิสิกส์พลังงานสูง และการประเมินคุณภาพของหินทรายในธรณีวิทยา

อย่างไรก็ตาม ควรระมัดระวังเรื่องการปรับแต่งโมเดลมากเกินไป (overfitting) ซึ่งอาจทำให้ประสิทธิภาพของโมเดลลดลงเมื่อนำไปใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน

2.3.5 การทดสอบประสิทธิภาพของตัวแบบโดยใช้ค่าเฉลี่ยของกำลังสองของความคลาดเคลื่อน (Mean Absolute Error: MAE)

ค่า Mean Absolute Error (MAE) หรือ "ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์" เป็นตัวชี้วัดที่ใช้ประเมินความแม่นยำของแบบจำลองการทำนาย โดยคำนวณจากค่าเฉลี่ยของค่าสัมบูรณ์ของความคลาดเคลื่อนระหว่างค่าที่ทำนายกับค่าจริง

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{true,i} - y_{pred,i}\right|$$

2.3.6 ค่าเฉลี่ยของรากที่สองของกำลังสองของความคลาดเคลื่อน (Root Mean Square Error: RMSE)

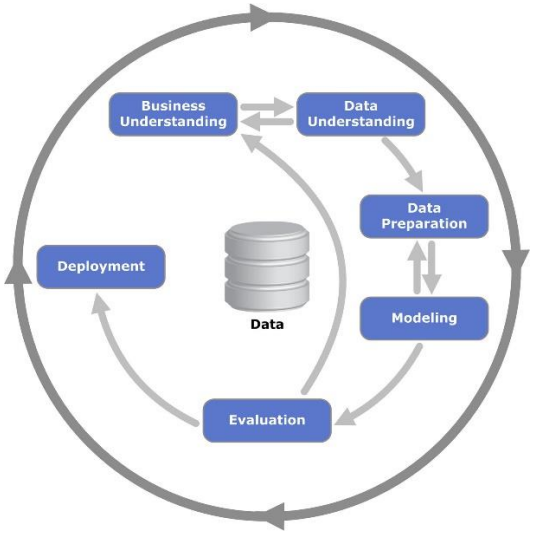
คือ การวัดความคลาดเคลื่อนเพื่อเปรียบเทียบความแตกต่างระหว่างค่าจากการพยากรณ์และค่าจริงเฉลี่ยกำลังสอง

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

จากค่าที่ได้ ถ้าค่า MAE และ RMSE ต่ำแสดงว่าค่าพยากรณ์มีความใกล้เคียงกับค่าจริง ซึ่งหมายถึงแบบจำลองมีประสิทธิภาพในการพยากรณ์สูง

2.3.7 กระบวนการวิเคราะห์ข้อมูลด้วย (CRISP-DM)

CRISP-DM ย่อมาจาก Cross-industry standard process for data mining หรือกระบวนการมาตรฐานข้ามอุตสาหกรรมสำหรับการทำเหมืองข้อมูล เป็นรูปแบบกระบวนการมาตรฐานแบบเปิดที่อธิบายแนวทางทั่วไปที่ผู้เชี่ยวชาญด้านการทำเหมืองข้อมูลใช้ เป็นรูปแบบการวิเคราะห์ที่ใช้กันอย่างแพร่หลายที่สุด



ภาพที่ 2.14 ขั้นตอนในกระบวนการ CRISP-DM

(ที่มา: https://kamboonchob.medium.com/)

1. การทำความเข้าใจธุรกิจ (Business Understanding) ขั้นตอนแรกมุ่งไปที่การทำความเข้าใจธุรกิจ ปัญหาและวัตถุประสงค์ของโครงการจากมุมมองทางธุรกิจ จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

2. การทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนนี้เริ่มต้นด้วยการรวบรวมข้อมูล จากนั้นทำความเข้าใจ ตรวจสอบคุณภาพ และเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์ ขั้นตอนที่ 1 และ 2 สามารถทำกลับไปมาได้ เนื่องจากการทำความเข้าใจธุรกิจทำให้เราเข้าใจข้อมูลมากขึ้น และการเข้าใจข้อมูลก็ทำให้เราเข้าใจธุรกิจมากขึ้นเช่นกัน

3. การเตรียมข้อมูล (Data Preparation) ขั้นตอนการเตรียมข้อมูล หมายถึง ขั้นตอนทั้งหมดที่จะทำเพื่อให้ข้อมูลดิบที่เรารวบรวมมา กลายเป็นข้อมูลสมบูรณ์ที่พร้อมจะเข้าสู่โมเดลในขั้นตอนที่ 4 เช่น การสร้างตาราง การลบข้อมูลที่ไม่ต้องการออก การแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ

4. การสร้างโมเดล (Modeling) ในขั้นตอนนี้ เราจะเลือกและทดสองสร้างโมเดลหลายๆแบบที่น่าจะสามารถแก้ไขปัญหาที่ต้องการได้ จากนั้นค่อยๆปรับค่าพารามิเตอร์ในแต่ละโมเดล เพื่อให้ได้โมเดลที่เหมาะสมที่สุดมาใช้ในการแก้ไขปัญหา

5. การวัดประสิทธิภาพของโมเดล (Evaluation) เราจะทำการวัดประสิทธิภาพของโมเดลที่ได้จากขั้นตอนที่ 4 เพื่อวัดว่าโมเดลมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานแล้วหรือไม่ ซึ่งโมเดลแต่ละประเภทก็จะมีตัววัดประสิทธิภาพที่แตกต่างกันออกไป

6. การนำโมเดลไปใช้งานจริง (Deployment) เป็นการนำโมเดลที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ

2.4 วรรณกรรมที่เกี่ยวข้อง

วีรศักดิ์ ฟองเงิน,วรปภา อารีราษฎร์และเผด็จ พรหมสาขา ณ สกลนคร (2561) ได้ทำการศึกษาวิจัย การพยากรณ์ปริมาณน้ำในเขื่อน โดยใช้เทคนิคเหมืองข้อมูล การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาเทคนิคเหมืองข้อมูลที่เหมาะสมในการพยากรณ์ปริมาณน้ำในเขื่อน และ 2) เปรียบเทียบผลการพยากรณ์ปริมาณน้ำรายเดือนในเขื่อนกิ่วลม จังหวัดลำปาง โดยใช้เทคนิคเหมืองข้อมูล งานวิจัยนี้ได้นำข้อมูลที่เป็นปัจจัยที่มีผลต่อการเปลี่ยนแปลงระดับน้ำประกอบด้วย ปริมาณน้ำไหลเข้าเขื่อน ปริมาณน้ำในเขื่อน ปริมาณการปล่อยน้ำและอัตราการระเหย โดยรวบรวมข้อมูลรายวัน ตั้งแต่ปี พ.ศ.2535 – พ.ศ.2559 รวม 25 ปี จำนวน 9,300 รายการ โดยมีการแยกข้อมูลรายเดือนเพื่อนำมาพยากรณ์ด้วยเทคนิคการพยาการณ์ ผลการวิจัยพบว่า 1) เทคนิคเหมืองข้อมูลที่เหมาะสมในการพยากรณ์ปริมาณน้ำในเขื่อน ประกอบด้วย 4 เทคนิค ได้แก่ เทคนิควิธีการวิเคราะห์การถดถอย (Regression Analysis) และวิธีโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) วิธีแบบจำลองต้นไม้เอ็มไฟว์พี (Model Tree: M5P) และ วิธีเทคนิคซัพพอร์ตเวกเตอร์แมกชีนส์ (SVM) และ 2) ผลการเปรียบเทียบการพยากรณ์ปริมาณน้ำรายเดือนในเขื่อนกิ่วลม จังหวัดล้าปาง โดยใช้เทคนิคเหมืองข้อมูลทั้ง 4 เทคนิค พบว่า วิธีแบบจำลองต้นไม้เอ็มไฟว์พี มีค่าสัมบูรณ์ของความคลาดเคลื่อนต่ำสุด ที่ 10.56 และเป็นวิธีที่เหมาะสมที่สุดสำหรับนำไปพัฒนาระบบพยากรณ์น้ำในเขื่อน ทั้งนี้เมื่อพิจารณาค่าสัมบูรณ์ของความคลาดเคลื่อนแต่ละเทคนิค พบว่าวิธีแบบจำลอง

ต้นไม้เอ็มไฟว์พีวิธีซัพพอร์ตเวกเตอร์ชีนส์วิธีวิเคราะห์การถดถอยและวิธีโครงข่ายประสาทเทียม มีค่าความคาดเคลื่อนเท่ากับ10.56, 10.84, 11.12 และ 12.53 ตามลำดับ

ธนกร สุทธิสนธ์ (2562) ได้ทำการศึกษาวิจัย การหาตัวแบบที่เหมาะสมเพื่อพยากรณ์ปริมาณการใช้น้ำประปาในจังหวัดอุบลราชธานี ปริมาณการใช้น้ำประปาของประชาชนในจังหวัดอุบลราชธานี ปัจจุบันมีแนวโน้มเพิ่มมากขึ้น การพยากรณ์ปริมาณการใช้น้ำประปาของประชาชนโดยใช้ตัวแบบที่เหมาะสมและมีความแม่นยำจึงมีความสำคัญและเป็นประโยชน์อย่างยิ่งในวางแผนและหามาตรการรองรับกับสถานการณ์การใช้น้ำประปาที่เพิ่มขึ้น การวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อสร้างตัวแบบที่เหมาะสมกับอนุกรมเวลา ปริมาณการใช้น้ำประปาของประชาชนในจังหวัดอุบลราชธานี จำนวน 4 สาขา ได้แก่ สาขาอุบลราชธานี สาขาพิบูลมังสาหาร สาขาเดชอุดม และสาขาเขมราฐ โดยใช้ข้อมูลจากเว็บไซต์ของการประปาส่วนภูมิภาค ตั้งแต่เดือนมกราคม พ.ศ. 2547 ถึงเดือนธันวาคม พ.ศ. 2560 จำนวน 168 ค่า ข้อมูลถูกแบ่งออกเป็น 2 ชุด คือ ข้อมูลชุดที่ 1 ตั้งแต่เดือนมกราคม พ.ศ. 2547 ถึงเดือนธันวาคม พ.ศ. 2559 จำนวน 156 ค่า ใช้สำหรับสร้างตัวแบบการพยากรณ์แบบอนุกรมเวลาเชิงเดี่ยวตามวิธีการของบอกซ์- เจนกินส์ และตัวแบบผสมระหว่างเทคนิคของบอกซ์-เจนกินส์กับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน ข้อมูลชุดที่ 2 ตั้งแต่เดือนมกราคม ถึงเดือน ธันวาคม พ.ศ.2560 จำนวน 12 ค่า ใช้สำหรับเปรียบเทียบความแม่นยำของการพยากรณ์แต่ละสาขา โดยใช้เกณฑ์ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) และเกณฑ์ร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percent Error : MAPE) ผลการวิจัย พบว่า ตัวแบบผสมมีความแม่นยำในการพยากรณ์มากกว่าตัวแบบการพยากรณ์เชิงเดี่ยว เนื่องจากมีค่า MAE และ MAPE ต่ำสุด ดังนั้น ตัวแบบผสมสามารถใช้เป็นเครื่องมือในการพยากรณ์ปริมาณการใช้น้ำประปาของประชาชนในจังหวัดอุบลราชธานีได้อย่างเหมาะสม และสามารถใช้ประกอบการตัดสินใจสำหรับการวางแผนการจัดการน้ำประปาให้เพียงพอต่อความต้องการของประชาชนในอนาคตได้

คชินทร์ โกกนุทาภรณ์ (2563) ได้ทำการศึกษาวิจัย การเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการพยากรณ์จำนวนผู้ใช้น้ำประปาของการประปาส่วนภูมิภาค สาขาปทุมธานี การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาหาตัวแบบพยากรณ์ที่เหมาะสมสำหรับการพยากรณ์จำนวนผู้ใช้น้ำประปาของการประปาส่วนภูมิภาค สาขาปทุมธานี ในการศึกษาครั้งนี้ได้นำเทคนิคการพยากรณ์มาช่วยในการ วิเคราะห์ข้อมูล ซึ่งประกอบด้วยวิธีสมการแนวโน้มเชิงเส้น วิธีสมการแนวโน้มกำลังสอง วิธีสมการแนวโน้มกำลังสาม วิธีสมการแนวโน้มเอ็กซ์โพเนนเชียล และวิธีสมการแนวโน้มกำลัง ข้อมูลที่ใช้เป็นข้อมูลทุติยภูมิที่รวบรวมจากกอง ศูนย์ข้อมูลและ

แผนเทคโนโลยีสารสนเทศ การประปาส่วนภูมิภาค สาขาปทุมธานี ลักษณะข้อมูลจำแนกเป็นรายเดือน ระหว่างเดือนมกราคม พ.ศ. 2555 ถึงเดือนกรกฎาคม พ.ศ. 2562 จำนวน 91 ค่า ผู้วิจัยได้แบ่งข้อมูลออกเป็น 2 ชุด ข้อมูลชุดที่ 1 ตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงเดือนธันวาคม พ.ศ. 2561 จำนวน 84 ค่า สำหรับเปรียบเทียบหาวิธีการพยากรณ์ที่เหมาะสมที่สุด โดยใช้เกณฑ์พิจารณาค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (MAD) และค่าเฉลี่ยเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (MAPE) ที่ต่ำที่สุด จากนั้นเลือกวิธีการพยากรณ์ที่เหมาะสมที่สุดคำนวณหาช่วงการพยากรณ์ล่วงหน้า โดยใช้ข้อมูลชุดที่ 2 คือ ตั้งแต่เดือนมกราคม พ.ศ. 2562 ถึงเดือนกรกฎาคมพ.ศ. 2562 จำนวน 7 ค่า โดยใช้เกณฑ์ค่าเฉลี่ยเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (MAPE) ที่ต่ำที่สุด ผลการวิจัยพบว่าวิธีการพยากรณ์ที่มีความเหมาะสมที่สุด คือ การพยากรณ์โดยวิธีสมการแนวโน้มกำลังสาม โดยมีตัวแบบพยากรณ์ คือ $\hat{Y}_3(t) = -0.007t^3 - 1.01t^2 + 243.173t + 43423.657$ เมื่อ t แทนช่วงเวลา จากรูปแบบ ดังกล่าวนำมาคำนวณหาช่วงการพยากรณ์ล่วงหน้าที่เหมาะสมที่สุด 3 เดือน 5 เดือน และ 7 เดือน พบว่าวิธีนี้เหมาะสำหรับการพยากรณ์ล่วงหน้า 7 เดือน

จิรโรจน์ ตอสะสุกุล และสุพิชชา ขัดธิพงษ์ (2565) ได้ทำการศึกษาวิจัย การเปรียบเทียบ ประสิทธิภาพของตัวแบบการพยากรณ์ปริมาณน้ำในเขื่อนด้วยเทคนิคการทำเหมืองข้อมูล เขื่อนภูมิพลถือว่าเป็นเขื่อนกักเก็บน้ำขนาดใหญ่ที่สุดของประเทศ นอกจากมีความสำคัญต่อการผลิตกระแสไฟฟ้าเพื่อสนองความต้องการของประเทศแล้ว ยังมีความสำคัญในการกักเก็บน้ำสำหรับการอุปโภคและบริโภครวมถึงช่วยในการจัดสรรน้ำในการเกษตรให้เพียงพอกับความต้องการในช่วงฤดูแล้ง โดยงานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพตัวแบบสำหรับการพยากรณ์ปริมาณน้ำในเขื่อนภูมิพล จังหวัดตาก โดยใช้ข้อมูลรายวันจากรายงานสถานภาพน้ำในเขื่อนของคลังข้อมูลน้ำแห่งชาติและข้อมูลทางอุตุนิยมวิทยาของสถานีกรมอุตุนิยมวิทยาภาคเหนือ ตั้งแต่วันที่ 1 มกราคม พ.ศ. 2554 ถึง 31 พฤษภาคม พ.ศ. 2564 จำนวนทั้งสิ้น 3,804 รายการ และวิเคราะห์ตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล โดยเปรียบเทียบประสิทธิภาพในการพยากรณ์ของเทคนิคการทำเหมืองข้อมูล 6 เทคนิค ได้แก่ เทคนิคการถดถอยเชิงเส้นพหุคูณ เทคนิคแรนดอมฟอร์เรส เทคนิคโครงข่ายประสาทเทียม เทคนิคจำลองต้นไม้เอ็มไฟว์พี เทคนิคเพื่อนบ้านใกล้ที่สุด และเทคนิคซัพพอร์ตเวกเตอร์แมกชีนส์ และในแต่ละเทคนิคได้เปรียบเทียบการแบ่งข้อมูลด้วยวิธีการตรวจสอบไขว้ ซึ่งผลการวิจัยพบว่า ตัวแบบการพยากรณ์ปริมาณน้ำในเขื่อนภูมิพลของเทคนิคแรนดอมฟอร์เรส มีประสิทธิภาพมากที่สุด โดยให้ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด เท่ากับ 341.945 และ 593.790 ตามลำดับ

ซึ่งตัวแบบการพยากรณ์ที่ได้ดังกล่าวสามารถนำไปใช้ในการประกอบการตัดสินใจและกำหนดแนวทางการพัฒนาระบบการบริหารจัดการน้ำสำหรับการผลิตกระแสไฟฟ้า การอุปโภค และบริโภคอย่างเพียงพอและเหมาะสม อีกทั้งยังสามารถช่วยในการวางแผนเพิ่มผลผลิตทางการเกษตรและกำหนดชนิดของพืชให้สอดคล้องกับปริมาณน้ำในเขื่อนได้ต่อไปในอนาคต

สุภารัตน์ พิลางาม (2560) ได้ทำการศึกษาวิจัย การใช้น้ำประปาและการคาดการณ์การใช้น้ำของโรงแรมในกรุงเทพมหานคร กรุงเทพมหานครเป็นเมืองศูนย์กลางเศรษฐกิจและยังเป็นเมืองอันดับหนึ่งที่เป็นจุดหมายปลายทางของนักท่องเที่ยว ทำให้ธุรกิจโรงแรมในกรุงเทพฯ ขยายตัวเพิ่มสูงขึ้นเป็นผลความต้องการใช้น้ำประปาในโรงแรมมากขึ้น จุดประสงค์ของการทำวิจัยนี้เพื่อศึกษาปริมาณการใช้น้ำของโรงแรมในกรุงเทพฯ โดยเก็บข้อมูลอาคารตัวอย่าง 34 แห่ง การศึกษา ชุดข้อมูลออกเป็นสองส่วน ส่วนแรกเป็นความสัมพันธ์ของปัจจัยที่มีผลต่อการใช้น้ำในโรงแรมจากการสัมภาษณ์กลุ่มตัวอย่าง และส่วนที่สองเป็นข้อมูลปริมาณการใช้น้ำปี พ.ศ.2558 ของกลุ่มตัวอย่างจากการประปานครหลวง ผลการศึกษาพบว่าปริมาณการใช้น้ำของโรงแรมในกรุงเทพฯ มีค่าเฉลี่ยเท่ากับ 1,210 ลิตร/ห้องพัก/วัน ในการสร้างแบบจำลองเพื่อคาดการณ์การใช้น้ำประปา ได้ใช้วิธีวิเคราะห์ถดถอยพหุคูณโดยคัดเลือกตัวแปรแบบขั้นตอนได้สมการคือ 26.01 (จำนวนห้องที่ขายได้) + 0.07 (พื้นที่ใช้สอยของอาคาร + 1,593.49 (ระดับดาว) – 6,239.41 โดยมีค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันของสมการคือ 0.92 เมื่อนำสมการมาทำการทดสอบความแม่นยำพบว่ามีค่าความคลาดเคลื่อน 3.73% สามารถนำไปวิเคราะห์ค่าการใช้น้ำในอนาคต ในการพยากรณ์การใช้น้ำในอนาคตเป็นการช่วยจัดสรรน้ำในอาคาร เป็นแนวทางให้ผู้ออกแบบโรงแรมในกรุงเทพฯ มีเข้าใจในระบบการจัดการน้ำใช้เพื่อช่วยประหยัดน้ำในอาคารได้

2.5 บทสรุป

จากแนวคิด ทฤษฎี เครื่องมือ และวรรณกรรมที่เกี่ยวข้องที่ได้กล่าวมาในข้างต้นที่เกี่ยวกับข้องกับการเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการพยากรณ์การใช้น้ำประปาด้วยเทคนิคเหมืองข้อมูล ผู้ศึกษาได้เลือกใช้กระบวนการ CRISP-DM โดยเปรียบเทียบตัวแบบ 4 ตัวแบบ คือเทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network : ANN) เทคนิคการวิเคราะห์ข้อมูลถดถอยเชิงเส้นหรือ (Linear Regression) เทคนิคแรนดอมฟอเรสต์ (Random Forest) และเทคนิคต้นไม้เสริมกำลังแบบไล่ระดับ (Gradient Boosted Trees) เพื่อศึกษาตัวแบบที่เหมาะสมสำหรับการพยากรณ์การใช้น้ำประปาในอนาคต เพื่อพยากรณ์การใช้น้ำประปาในอนาคต และเพื่อเผยแพร่ข้อมูลผ่านเว็บไซต์ จากนั้นนำข้อมูลมาทำการแสดงผลแบบ Visualization ในรูปแบบของแดชบอร์ดและเผยแพร่ผ่านเว็บไซต์