บทที่ 2
แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้อง

โครงงานเรื่องการประยุกต์ใช้เทคนิคการวิเคราะห์การจัดกลุ่มเพื่อทำความเข้าใจความหลากหลายของ Data Scientist ในบทนี้เป็นการนำเสนอเกี่ยวกับ แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้องของการวิเคราะห์ข้อมูลการจัดกลุ่มเพื่อทำความเข้าใจความหลากหลายของ Data Scientist ซึ่งได้รวบรวมการศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล เพื่อใช้เป็นแนวทางการศึกษาประกอบด้วยรายละเอียดตามลำดับ ดังนี้

2.1 แนวคิด

2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data Analytic)

2.1.2 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data Visualization)

2..2 ทฤษฎี

2.2.1 ทฤษฎีเกี่ยวกับการจัดการข้อมูลขนาดใหญ่

2.2.2 ทฤษฎีเกี่ยวกับการทำเหมืองข้อมูล

2.2.3 ทฤษฎีเกี่ยวกับการออกแบบโครงสร้างเว็บไซต์

2.2.4 ทฤษฎีเกี่ยวกับชุดคำสั่ง Python

2.2.5 ทฤษฎีเกี่ยวกับการทำ Classification

2.2.6 ทฤษฎีเกี่ยวกับการ Visualization

2.2.7 ทฤษฎีเกี่ยวกับการการทำ Decision Tree

2.2.8 ทฤษฎีเกี่ยวกับการการทำ Random Forest

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

2.3.1 RapibMiner Studio

2.3.2 Visual Studio Code

2.3.3 HTML สำหรับสร้างหน้าเว็บไซต์

2.3.4 Figma (สำหรับออกแบบหน้าเว็บและ Layout)

2.4 วรรณกรรมที่เกี่ยวข้อง

2.5 บทสรุป

2.1 แนวคิด

2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data Analytic) เป็นศาสตร์การวิเคราะห์ข้อมูลจากการนำข้อมูล หรือ Big data ที่มีอยู่มากมายนำมาใช้ประโยชน์ในการตัดสินใจ แก้ไขปัญหา และสร้างมูลค่าแก่ธุรกิจ โดยจะนำข้อมูลเหล่านั้นมาประมวลผล วิเคราะห์ เชื่อมโยงข้อมูลต่าง ๆ เข้าด้วยกัน เพื่อทำความเข้าใจและแปลงข้อมูลออกมาเป็นข้อมูลสารสนเทศที่พร้อมใช้งาน เข้าใจง่าย ไม่ยุ่งยาก ซึ่งจะช่วยให้องค์กรได้ข้อมูลที่มีประสิทธิภาพอย่างสูงสุด Data Analytics มีรูปแบบการวิเคราะห์ข้อมูลทั้งหมด 4 รูปแบบ ซึ่งแต่ละรูปแบบมีวัตถุประสงค์และการใช้งานที่แตกต่างกันออกไป ดังนี้

1) การวิเคราะห์ข้อมูลพื้นฐาน หรือที่เรียกว่า การวิเคราะห์เชิงบรรยาย (Descriptive Analytics) Data Analytics เพื่อวิเคราะห์ข้อมูลสถานะปัจจุบันขององค์กรธุรกิจ โดยจะใช้เครื่องมือทางสถิติสร้างรายงานและแสดงผลข้อมูลในรูปแบบกราฟหรือสถิติเชิงลำดับ ซึ่งจะช่วยในการจัดการและตัดสินใจระดับบริหาร

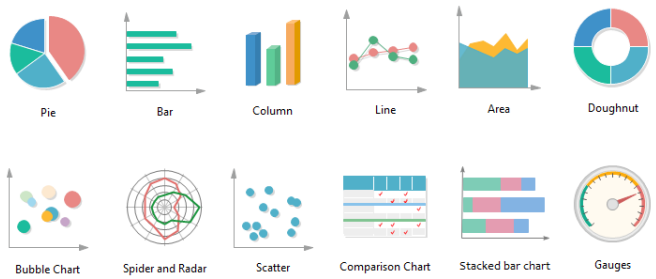
2) การวิเคราะห์เชิงวินิจฉัย (Diagnostic Analytics)รูปแบบ Data Analytics เชิงวินิจฉัยมีวัตถุประสงค์เพื่อค้นหาสาเหตุของปัญหาหรือวิเคราะห์สถานการณ์ต่าง ๆ ที่เกิดขึ้น โดยจะใช้เทคนิคและเครื่องมือเชิงสถิติ เพื่อหาปัจจัยหลักที่ส่งผลต่อการดำเนินงาน

3) การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics)Data Analytics ในรูปแบบการพยากรณ์เป็นการใช้ข้อมูลเก่าหรือข้อมูลประวัติศาสตร์ เพื่อสร้างและทดสอบโมเดลที่จะทำนายสถานการณ์ในอนาคต Data Analytics ตัวอย่างเช่น การนำข้อมูลยอดขายในอดีตมาวิเคราะห์ข้อมูลเพื่อทำนายยอดขายในอนาคต, การแบ่งกลุ่มลูกค้า

4) การวิเคราะห์เชิงแนะนำ (Prescriptive Analytics)Data Analytics ในรูปแบบเชิงแนะนำ จะเป็นการใช้ข้อมูลและโมเดลเพื่อแนะนำกลยุทธ์หรือการดำเนินงานที่เหมาะสมเพื่อแก้ไขปัญหาหรือช่วยให้การดำเนินงานเป็นไปอย่างราบรื่นมากยิ่งขึ้น ตัวอย่างเช่น การวางแผนการจัดการทรัพยากรของระบบการผลิต หรือการแนะนำสินค้าและบริการแก้ไขปัญหาทางธุรกิจ เป็นต้น

2.1.2 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data Visualization) Data Visualization คือการนำข้อมูลหรือ Data ที่ได้มาจากแหล่งข้อมูลต่างๆ มาวิเคราะห์ประมวลผลแล้วนำเสนอ

ออกมาในรูปแบบที่มองเห็นและทำความเข้าใจได้ด้วยตา เช่น แผนภูมิ รูปภาพ แผนที่ กราฟแสดงเทรนด์ ตาราง วิดีโอ อินโฟกราฟิก (Infographic) แดชบอร์ด (dashboard) ฯลฯ จุดประสงค์สำคัญของการทำ Data Visualization คือ การนำเสนอข้อมูลให้เข้าใจง่าย ผู้อ่านข้อมูลสามารถเข้าใจได้ทันทีว่าตัวชิ้นงาน (media) ต้องการสื่อสารอะไร ชี้จุดสำคัญของเนื้อหา และชี้ Insight ข้อเปรียบเทียบให้เห็นอย่างชัดเจน ช่วยให้สังเกตเห็นจุดที่น่าสนใจของข้อมูลได้ง่ายขึ้น

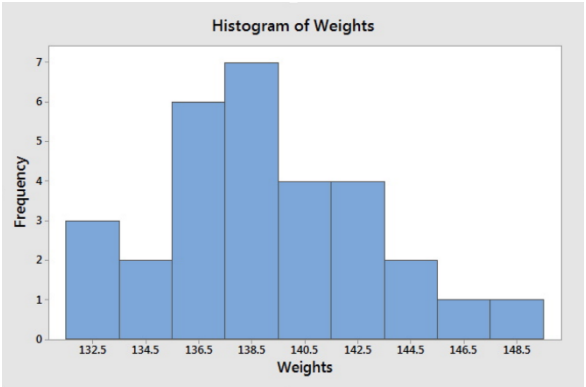


ภาพที่ 2.1 แสดง Data Visualization

(ที่มา: https://1stcraft.com)

พื้นฐานของการสร้าง Data Visualization คือ การ Mapping ส่วนข้อมูลกับส่วนของGraphic เข้าด้วยกัน ซึ่งตอนนี้มีโปรแกรมสำเร็จรูปในการสร้าง Data Visualization หลากหลาย โปรแกรมมีฟังก์ชันการใช้งานที่เข้าใจง่าย เช่น การสร่างฟิลเตอร์ การออกแบบเพื่อให้งานการวิเคราะห์ข้อมูลมีความยืดหยุ่นเป็นต้น ตัวอย่างรูปแบบ Data Visualization ที่นิยมใช้กันมีดังนี้

1) Distribution Histogram กราฟแท่งแบบเฉพาะที่แสดงความสัมพันธ์ระหว่างข้อมูลเป็นหมวดหมู่

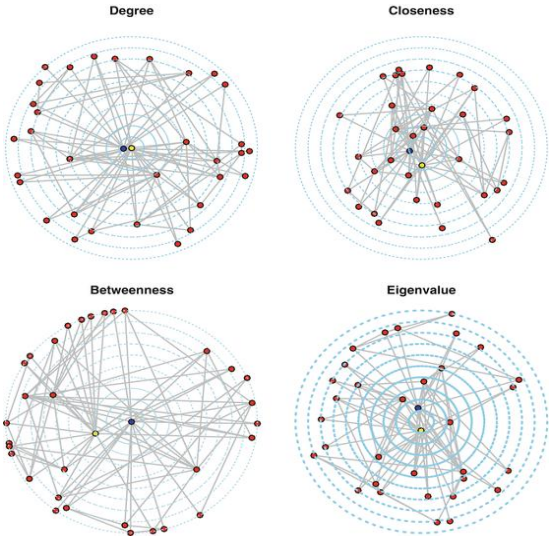


ภาพที่ 2.2 แสดง Histogram

(ที่มา: https://online.stat.psu.edu/stat500/book/export/html/539)

2) NETWORK/FLOW

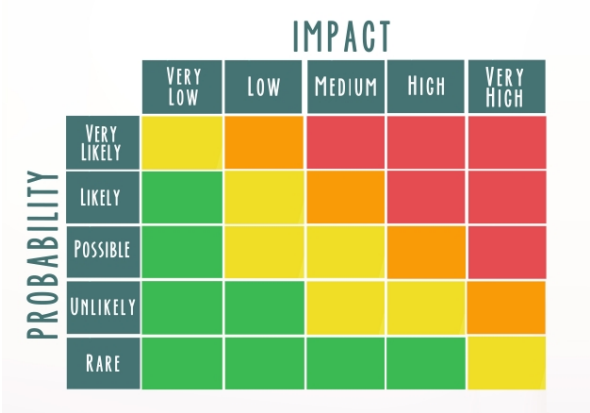
Network Graph ใช้แสดงความเชื่อมโยงของเครือข่ายหรือความสัมพันธ์ในกลุ่ม



ภาพที่ 2.3 แสดง Network Graph

(ที่มา: https://link.springer.com/)

3) RelationshipHeatmap ใช้แสดงรูปแบบความสัมพันธ์ของข้อมูล โดยจะแสดงออกมาในรูปแบบของ"สี" ซึ่งแต่ละสีจะบ่งบอกถึงระดับความถี่ของพฤติกรรม แต่ใช้อ่านค่าความแตกต่างเล็กน้อยได้ยากที

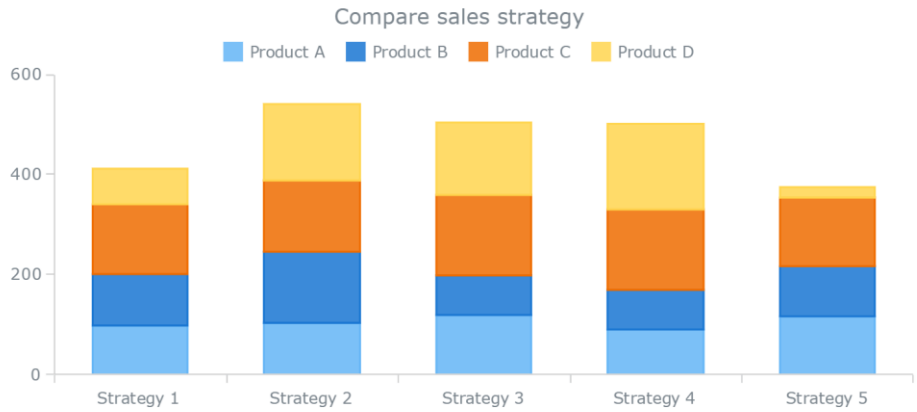


ภาพที่ 2.4 แสดง Heatmap

(ที่มา: https://www.neugroup.com/)

4) Comparison

Stacked Bar Chart ใช้เปรียบเทียบค่าผลรวมและสัดส่วนจากข้อมูลหลายกลุ่ม อาจอ่านสัดส่วนยากเมื่อมีข้อมูลหลายกลุ่มมากเกินไป

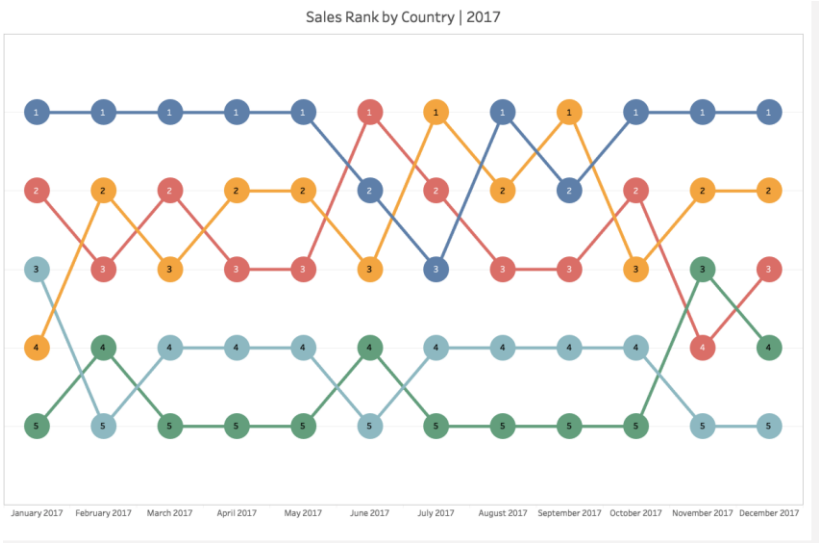


ภาพที่ 2.5 แสดง Stacked Bar Chart

(ที่มา: https://www.smashingmagazine.com/)

5) Ranking

Bump Chart ใช้แสดงการเปลี่ยนแปลงลำดับของข้อมูลในหลาบช่วงเวลา โดยใช้สีเพื่อช่วยให้อ่านข้อมูลได้ง่ายขึ้น

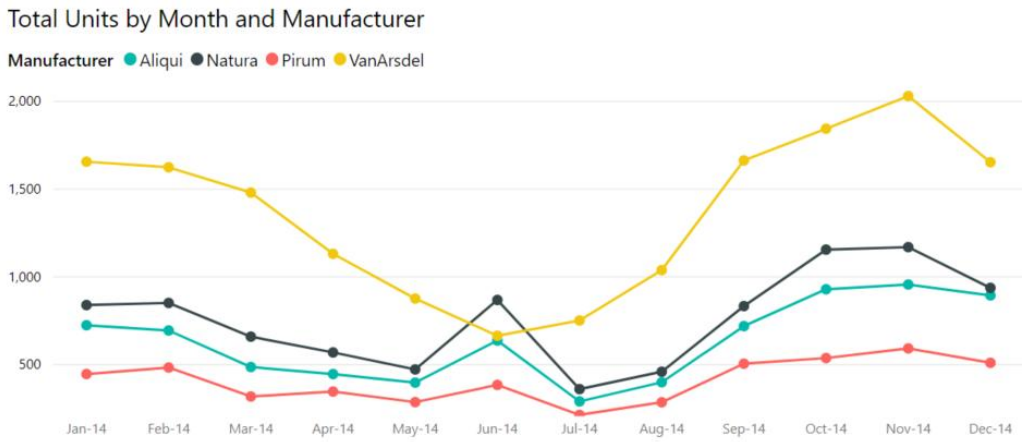


ภาพที่ 2.6 แสดง Bump Chart

(ที่มา: https://vizright.wordpress.com/)

6) Time–Series

Line Chart ใช้เปรียบเทียบข้อมูลเพื่อดูแนวโน้ม (Trends) โดยอาจจะเทียบกับมิติของเวลา Time



ภาพที่ 2.7 แสดง Line Chart

(ที่มา: https://learn.microsoft.com/)

จากที่ได้กล่าวไปจะเห็นได้ว่าข้อมูลแบบ Data Visualization มีพลังทางการสื่อสารอย่างมากเพราะสามารถแสดงผลและเพิ่มความสนใจได้เป็นอย่างดี หากองค์กรได้ลองนำการแสดงข้อมูลแบบ Data Visualization มาใช้ก็จะช่วยในการจัดการระดับสายงานแผนกอื่นๆ จะส่งผลให้การวิเคราะห์ข้อมูลและตัดสินใจในสายงานต่างๆ ได้ง่าย สามารถช่วยให้มีประสิทธิภาพที่ดีต่อธุรกิจและทีม หรือนำมาใช้ในหน้าแรกของเว็บไซต์ เพื่อทำการโปรโมทโปรโมชั่นหรือข้อเสนอทางการตลาดของธุรกิจและจะพบว่า Data Visualization ช่วยให้เกิดการตอบสนองกับข้อความมากขึ้น

2.2 ทฤษฎี

2.2.1 ทฤษฎีเกี่ยวกับการจัดการข้อมูลขนาดใหญ่

2.2.1.1 ข้อมูลขนาดใหญ่ (Big Data) ชุดข้อมูลใดๆ ก็ตามที่มีขนาดใหญ่และถูกเก็บบันทึกไว้ผ่านวิธีการต่างๆ ยกตัวอย่างเช่นฐานข้อมูลภายในองค์กรของคุณเอง ไม่ว่าจะเป็นข้อมูล

การทำ transaction ต่างๆ หรือข้อมูลพฤติกรรมของลูกค้าปัจจุบันโลกกำลังถูกขับเคลื่อนด้วยเทคโนโลยีดิจิทัล ชุดข้อมูล (Data) จำนวนมากถูกสร้างขึ้นใหม่ทุกวัน และสามารถทำการเก็บรวบรวมได้ง่ายขึ้น ไม่ว่าจะเป็นพฤติกรรมการค้นหาข้อมูลทางอินเทอร์เน็ต การรับชมสื่อออนไลน์หรือธุรกรรมสำคัญต่างๆ บริษัท International Data Corporation (IDC) คาดการณ์ว่าปริมาณข้อมูล (Data) ที่เกิดขึ้นบน Digital Platform ภายในปี 2020 จะมีจำนวนมากถึง 40 Zettabyte หรือเทียบเท่ากับ 40 ล้านล้าน Gigabyte เลยทีเดียวชุดข้อมูลมหาศาลเหล่านี้มีประโยชน์อย่างมากต่อการทำ Customer Insight โดยเป็นส่วนสำคัญสำหรับนำไปพัฒนาระบบ AI เพื่อให้เข้าถึงลูกค้าได้อย่างตรงจุด ตัวอย่างเช่น Chatbot บนแพลตฟอร์ม Facebook เพื่อนำเสนอคำตอบที่ตรงใจ และชักจูงให้ลูกค้าสนใจสินค้าหรือบริการมากที่สุดดังนั้น ถึงเวลาแล้วที่แต่ละองค์กรควรจะเริ่มทำการเก็บข้อมูลและนำมาใช้ต่อยอดธุรกิจ การมีข้อมูลที่มากกว่าจะทำให้คุณทำความเข้าใจในสถานการณ์ต่างๆ ได้ดียิ่งกว่าคู่แข่ง ทำให้คุณสามารถแข่งขันและขึ้นไปยืนอยู่บนจุดสูงสุดของการทำธุรกิจได้ หรืออาจกล่าวได้ว่า ข้อมูลยิ่งมีมากเท่าใดโอกาสที่จะประสบความสำเร็จขององค์กรก็มีมากเท่านั้น โดย Big Data นั้นสามารถแบ่งเป็นหมวดหมู่ตามโครงสร้างของชุดข้อมูลได้ดังนี้

1) ข้อมูลที่มีโครงสร้างชัดเจน (Structured Data) หมายถึง ชุดข้อมูลที่มีการจัดเรียงโครงสร้างอย่างเป็นระเบียบ มีความชัดเจน หรือระบุได้ด้วยตัวเลข พร้อมใช้งานได้ทันที เช่น จำนวนการซื้อขายกับลูกค้า เปอร์เซ็นต์ความเคลื่อนไหวภายในตลาดหุ้น ฯลฯ

2) ข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) หมายถึง ชุดข้อมูลที่มีโครงสร้างไม่ชัดเจนหรือไม่สามารถระบุความแน่นอนของข้อมูลนั้นๆ ได้ ยังไม่สามารถประมวลผลเพื่อนำไปใช้ได้ทันที อย่างเช่น บทสนทนาโต้ตอบกับลูกค้าทาง Social Media

3) ข้อมูลกึ่งมีโครงสร้าง (Semi–Structured Data) หมายถึง ชุดข้อมูลที่มีโครงสร้างระดับหนึ่งแต่ยังไม่สมบูรณ์ เช่น สเตตัสใน Social Media เป็นข้อมูลที่ไม่มีโครงสร้าง แต่ในกรณีที่มี Hashtag เข้ามาช่วยในการจัดหมวดหมู่ จะทำให้ข้อมูลมีความเป็นระเบียบขึ้นมาเล็กน้อย

คุณลักษณะของ Big Data ที่มีประสิทธิภาพ มี 6 ประการประกอบด้วย

1) ข้อมูลที่มีปริมาณมาก (Volume) หมายถึง มีปริมาณข้อมูลอยู่มาก มีขนาดใหญ่ สามารถนับรวมได้ทั้งข้อมูลแบบออนไลน์และแบบออฟไลน์ โดยข้อมูลต้องมีขนาดใหญ่เกินกว่า Terabyte

2) ข้อมูลที่มีความหลากหลาย (Variety)หมายถึง ข้อมูลแต่ละชนิดนั้นมีความหลากหลาย รวมกันทั้งรูปแบบมีโครงสร้าง ไม่มีโครงสร้าง และกึ่งโครงสร้าง

3) ข้อมูลที่มีการเพิ่มขึ้นอย่างรวดเร็ว (Velocity)หมายถึง ข้อมูลที่มีการเพิ่มขึ้นและเกิดความเปลี่ยนแปลงอย่างรวดเร็ว ทำให้เกิดข้อมูลแบบ Real-time มากมาย อย่างเช่นข้อมูลการจราจร ซึ่ง Google Map ก็ได้ใช้ประโยชน์จากการเข้าถึง GPS ของผู้ที่สัญจรไปมาบนท้องถนนเพื่อวิเคราะห์และนำเสนอเส้นทางที่การจราจรคล่องตัวที่สุดให้กับผู้ใช้งาน

4) ข้อมูลที่สร้างประโยชน์นำไปใช้ในทางธุรกิจได้ (Value)หมายถึง ข้อมูลที่มีคุณค่าต่อการนำไปใช้งาน สามารถก่อให้เกิดประโยชน์ทางธุรกิจได้เป็นอย่างดี ยกตัวอย่างเช่น พฤติกรรมการค้นหาข้อมูลผ่าน Google ที่ทำให้สามารถทราบถึงความสนใจของผู้คนในช่วงเวลานั้นๆ ได้

5) ข้อมูลต้องมีความถูกต้องชัดเจน (Veracity) เนื่องจาก Big Data นั้นรวบรวมข้อมูลไว้เป็นจำนวนมหาศาล เพราะฉะนั้น สิ่งที่สำคัญที่สุดก็คือความถูกต้องชัดเจนของข้อมูล ซึ่งจะเป็นส่วนสำคัญที่จะสามารถนำข้อมูลเหล่านั้นมาประมวลผลเพื่อการใช้งานต่อในอนาคตได้

6) ข้อมูลต้องมีความเชื่อมโยงกัน (Complexity) การจะใช้ประโยชน์จาก Big Data ได้นั้น มีอีกหนึ่งปัจจัยสำคัญนั่นก็คือความเชื่อมโยงกันของข้อมูล หากสิ่งที่รวบรวมมานั้นไม่สามารถหาจุดเชื่อมโยงกันได้ ข้อมูลเหล่านั้นก็ไร้ประโยชน์ การเก็บ Data ที่มีประสิทธิภาพนั้นจึงต้องคำนึงถึงความสัมพันธ์กันของข้อมูลด้วย

2.2.1.2 Big Data Analytics กระบวนการนำข้อมูลดิบที่มีปริมาณมหาศาลและกระจัดกระจายมาจัดเรียงและวิเคราะห์ข้อมูล เพื่อให้ธุรกิจสามารถเห็นภาพรวมของข้อมูลชุดต่าง ๆ หรือความเชื่อมโยงข้อมูลกับบริบทที่มีอยู่ได้อย่างชัดเจน

1) การรวบรวมและจัดเก็บข้อมูล (Storage)การจัดเก็บข้อมูลขนาดใหญ่ (Big Data) คือ จำเป็นต้องใช้โซลูชันการจัดเก็บข้อมูลที่มีประสิทธิภาพและสามารถปรับขนาดได้ แพลตฟอร์มการจัดเก็บข้อมูลบนคลาวด์ ตามบทบาทของ Big Data ในด้านต่างๆเช่น Amazon S3 และ Google Cloud Storage มีโครงสร้างพื้นฐานที่จำเป็นในการจัดเก็บชุดข้อมูลขนาดใหญ่

2) การประมวลผลข้อมูล (Processing)Big Data Analysis คือ เกี่ยวข้องกับการทำความสะอาด การแปลง และการจัดระเบียบข้อมูลเพื่อการวิเคราะห์ เครื่องมืออย่าง Apache

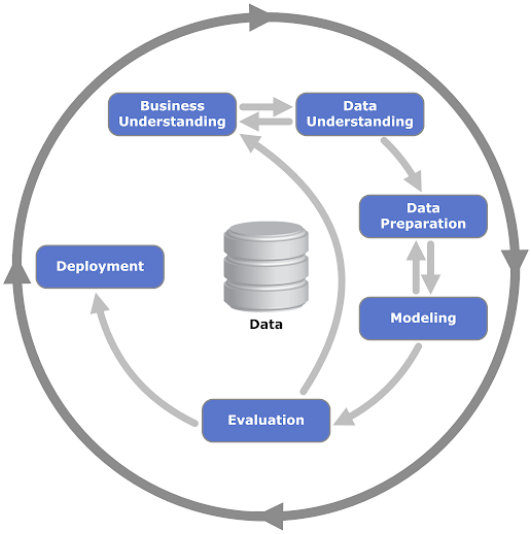
Hadoop และ Spark มักถูกใช้สำหรับการประมวลผล Big Data ช่วยให้ธุรกิจต่างๆ จัดการและวิเคราะห์ชุดข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ

3) การวิเคราะห์และนำเสนอข้อมูล (Analyst)การวิเคราะห์ Big Data เกี่ยวข้องกับการใช้การวิเคราะห์ขั้นสูงและอัลกอริทึมการเรียนรู้ของเครื่องเพื่อดึงข้อมูลเชิงลึกที่นำไปใช้งานได้จริง แพลตฟอร์มต่างๆ เช่น Tableau และ Power BI ช่วยแสดงข้อมูลให้เห็นภาพ ทำให้ธุรกิจต่างๆ เข้าใจและดำเนินการกับข้อมูลได้ง่ายขึ้น

2.2.2 ทฤษฎีเกี่ยวกับการทำเหมืองข้อมูล

การทำเหมืองข้อมูล (data mining) เป็นกระบวนการในการค้นหารูปแบบในชุดข้อมูลขนาดใหญ่ โดยใช้วิธีการของการเรียนรู้ของเครื่อง สถิติ และระบบฐานข้อมูลการทำเหมืองข้อมูลเป็นขั้นตอนวิธีการในการ"การค้นหาความรู้ในฐานข้อมูล" (knowledge discovery in databases – KDD) การทำเหมืองข้อมูลเป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และ การรู้จำแบบ หรือในอีกนิยามหนึ่งการทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูล(โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ การประยุกต์ใช้การทำเหมืองข้อมูลได้แก่ การขายปลีกและขายส่ง การธนาคาร การประดิษฐ์และการผลิต การประกันภัย การทำงานของตำรวจ การดูแลสุขภาพ การตลาด การใช้งานอินเทอร์เน็ต การศึกษา เป็นต้น

ขั้นตอนในการวิเคราะห์ข้อมูลได้ใช้หลักการของกระบวนการหาความรู้แบบ Cross–industry standard process for data mining ซึ่งเป็นแนวทางในการดำเนินงาน CRISP–DM ประกอบไปด้วย 6 ขั้นตอน



ภาพที่ 2.8 แสดง CRISP–DM

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

1) Business understanding การทำความเข้าใจธุรกิจ ปัญหาและวัตถุประสงค์ของโครงการจากมุมมองทางธุรกิจ จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

2) Data understanding การทำความเข้าใจข้อมูล รวบรวมข้อมูล จากนั้นทำความเข้าใจ ตรวจสอบคุณภาพ และเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์

3) Data preparation การเตรียมข้อมูล ทำความสะอาดข้อมูล แปลงข้อมูลให้เป็นรูปแบบที่เหมาะสมสำหรับการวิเคราะห์

4) Modeling การสร้างโมเดล เลือกเทคนิคการทำเหมืองข้อมูลที่เหมาะสมกับโจทย์และข้อมูล จากนั้นสร้างโมเดลและประเมินประสิทธิภาพของโมเดล

5) Evaluation การประเมินผลโมเดล ทดสอบโมเดลกับข้อมูลทดสอบเพื่อประเมินประสิทธิภาพของโมเดล

6) Deployment การนำผลการวิเคราะห์ไปใช้ นำเสนอผลการวิเคราะห์ต่อผู้บริหารหรือผู้ที่เกี่ยวข้อง เพื่อให้สามารถนำไปประยุกต์ใช้ให้เกิดประโยชน์ต่อธุรกิจ

2.2.3 ทฤษฎีเกี่ยวกับการออกแบบโครงสร้างเว็บไซต์

การออกแบบโครงสร้างเว็บไซต์ คือ การวางแผนการจัดลำดับ เนื้อหาสาระของเว็บไซต์ออกเป็นหมวดหมู่ เพื่อจัดทำเป็นโครงสร้างในการจัดวางหน้าเว็บเพจทั้งหมด เปรียบเสมือนแผนที่ที่ทำให้เห็นโครงสร้างทั้งหมดของเว็บไซต์ ช่วยในนักออกแบบเว็บไซต์ไม่ให้หลงทาง การจัดโครงสร้างของเว็บไซต์ มีจุดมุ่งหมายสำคัญคือ การที่จะทำให้ผู้เข้าเยี่ยมชม สามารถค้นหาข้อมูลในเว็บเพจได้อย่างเป็นระบบ ซึ่งถือว่าเป็นขั้นตอนที่สำคัญ ที่สามารถสร้างความสำเร็จให้กับผู้ที่ทำหน้าที่ในการออกแบบและพัฒนาเว็บไซต์ (Webmaster) การออกแบบโครงสร้างหรือจัดระเบียบของข้อมูลที่ชัดเจน แยกย่อยเนื้อหาออกเป็นส่วนต่าง ๆ ที่สัมพันธ์กันและให้อยู่ในมาตรฐานเดียวกันจะช่วยให้น่าใช้งานและง่าย ต่อการเข้าอ่านเนื้อหาของผู้ใช้เว็บไซต์

2.2.3.1 รูปแบบโครงสร้างของเว็บไซต์

การออกแบบโครงสร้างเว็บไซต์ สามารถทำได้หลากหลายแบบซึ่งก้ขึ้นอยู่กับความชอบ และความถนัดของแต่ละบุคคล นอกจากนี้ยังขึ้นอยู่กับกลุ่มเป้าหมายที่ต้องการนำเสนอ เพราะจะต้องออกแบบให้เหมาะกับการใช้งานของกลุ่มเป้าหมายมากที่สุด โดยโครงสร้างของเว็บไซต์ส่วนใหญ่ก็จะประกอบไปด้วย 4 รูปแบบดังนี้

1) เว็บที่มีโครงสร้างแบบเรียงลำดับ (Sequential Structure)เป็นโครงสร้างแบบธรรมดาที่ใช้กันมากที่สุดเนื่องจากง่ายต่อการจัดระบบข้อมูล ข้อมูลที่นิยม จัดด้วยโครงสร้างแบบนี้มักเป็นข้อมูลที่มีลักษณะเป็นเรื่องราวตามลำดับของเวลา เช่น การเรียงลำดับตามตัวอักษร ดรรชนี สารานุกรม หรืออภิธานศัพท์ โครงสร้างแบบนี้ เหมาะกับเว็บไซต์ที่มีขนาดเล็ก เนื้อหาไม่ซับซ้อนใช้การลิงก์ (Link) ไปทีละหน้า ทิศทางของการเข้าสู่เนื้อหา (Navigation) ภายในเว็บจะเป็นการดำเนินเรื่องในลักษณะเส้นตรง โดยมี ปุ่มเดินหน้า–ถอยหลังเป็นเครื่องมือหลักในการกำหนดทิศทาง ข้อเสียของโครงสร้างระบบนี้คือ ผู้ใช้ไม่สามารถกำหนดทิศทางการเข้าสู่เนื้อหาของตนเองได้ ทำให้เสียเวลาเข้าสู่เนื้อ

ภาพที่ 2.9 แสดงโครงสร้างแบบเรียงลำดับ

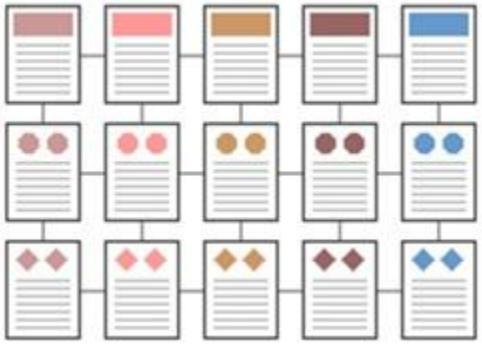
(ที่มา: https://www.g-tech.ac.th)

2) เว็บที่มีโครงสร้างแบบลำดับขั้น (Hierarchical Structure) เป็นวิธีที่ดีที่สุดวิธีหนึ่งในการจัดระบบโครงสร้างที่มีความซับซ้อนของข้อมูล โดยแบ่งเนื้อหา ออกเป็นส่วนต่างๆ และมีรายละเอียดย่อยๆ ในแต่ละส่วนลดหลั่นกันมาในลักษณะแนวคิดเดียวกับ แผนภูมิองค์กร จึงเป็นการง่ายต่อการทำความเข้าใจกับโครงสร้างของเนื้อหาในเว็บลักษณะนี้ ลักษณะเด่นเฉพาะของ เว็บประเภทนี้คือการมีจุดเริ่มต้นที่จุดร่วมจุดเดียว นั่นคือ โฮมเพจ (Homepage) และเชื่อมโยงไปสู่เนื้อหา ในลักษณะเป็นลำดับจากบนลงล่าง

ภาพที่ 2.11 แสดงโครงสร้างแบบเรียงลำดับขั้น

(ที่มา: https://www.g-tech.ac.th)

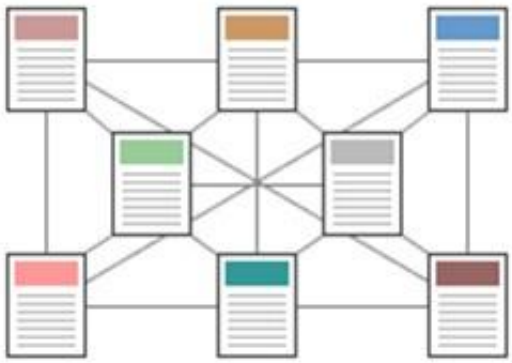
3) เว็บที่มีโครงสร้างแบบตาราง (Grid Structure) โครงสร้างรูปแบบนี้มีความซับซ้อนมากกว่ารูปแบบที่ผ่านมา การออกแบบเพิ่มความยืดหยุ่น ให้แก่การเข้าสู่เนื้อหาของผู้ใช้ โดยเพิ่มการเชื่อมโยงซึ่งกันและกันระหว่างเนื้อหาแต่ละส่วน เหมาะแก่ การแสดงให้เห็นความสัมพันธ์กันของเนื้อหา การเข้าสู่เนื้อหาของผู้ใช้จะไม่เป็นลักษณะเชิงเส้นตรง เนื่องจากผู้ใช้สามารถเปลี่ยนทิศทางการเข้าสู่เนื้อหาของตนเองได้



ภาพที่ 2.12 แสดงโครงสร้างแบบตาราง

(ที่มา: https://www.g-tech.ac.th)

4) เว็บที่มีโครงสร้างแบบใยแมงมุม (Web Structure)โครงสร้างประเภทนี้จะมีความยืดหยุ่นมากที่สุด ทุกหน้าในเว็บสามารถจะเชื่อมโยงไปถึงกัน ได้หมด เป็นการสร้างรูปแบบการเข้าสู่เนื้อหาที่เป็นอิสระ ผู้ใช้สามารถกำหนดวิธีการเข้าสู่เนื้อหาได้ด้วย ตนเอง การเชื่อมโยงเนื้อหาแต่ละหน้าอาศัยการโยงใยข้อความที่มีมโนทัศน์ (Concept) เหมือนกัน ของแต่ละหน้าในลักษณะของไฮเปอร์เท็กซ์หรือไฮเปอร์มีเดีย โครงสร้างลักษณะนี้จัดเป็นรูปแบบที่ ไม่มีโครงสร้างที่แน่นนอนตายตัว (Unstructured) นอกจากนี้การเชื่อมโยงไม่ได้จำกัดเฉพาะเนื้อหา ภายในเว็บนั้นๆ แต่สามารถเชื่อมโยงออกไปสู่เนื้อหาจากเว็บภายนอกได้



ภาพที่ 2.13 แสดงโครงสร้างแบบใยแมงมุน

(ที่มา: https://www.g-tech.ac.th)

2.2.3.2 องค์ประกอบที่ดีของการออกแบบเว็บไซต์

1) โครงสร้างที่ชัดเจน ผู้ออกแบบเว็บไซต์ควรจัดโครงสร้างหรือจัดระเบียบของข้อมูลที่ชัดเจน แยกย่อยเนื้อหาออกเป็นส่วนต่าง ๆ ที่สัมพันธ์กันและให้อยู่ในมาตรฐานเดียวกัน จะช่วยให้น่าใช้งานและง่าย ต่อการอ่านเนื้อหาของผู้ใช้

2) การใช้งานที่ง่าย ลักษณะของเว็บที่มีการใช้งานง่ายจะช่วยให้ผู้ใช้รู้สึกสบายใจต่อการอ่านและสามารถทำความเข้าใจกับเนื้อหาได้อย่างเต็มที่ โดยไม่ต้องมาเสียเวลาอยู่กับการทำความเข้าใจ การใช้งานที่สับสนด้วยเหตุนี้ผู้ออกแบบจึงควรกำหนดปุ่มการใช้งานที่ชัดเจนเหมาะสม โดยเฉพาะปุ่มควบคุมเส้นทางการเข้าสู่เนื้อหา (Navigation) ไม่ว่าจะเป็นเดินหน้า ถอยหลัง หากเป็นเว็บไซต์ที่มีเว็บเพจจำนวนมาก ควรจะจัดทำแผนผังของเว็บไซต์ (Site Map) ที่ช่วยให้ผู้ใช้ทราบว่า ตอนนี้อยู่ ณ จุดใด หรือเครื่องมือสืบค้น (Search Engine) ที่ช่วยในการค้นหาหน้าที่ที่ต้องการ

3) การเชื่อมโยงที่ดี ลักษณะไฮเปอร์เท็กซ์ที่ใช้ในการเชื่อมโยง ควรอยู่ในรูปแบบที่เป็นมาตรฐาน ทั่วไปและต้องระวังเรื่องของตำแหน่งในการเชื่อมโยง การที่จำนวนการเชื่อมโยงมากและกระจัดกระจายอยู่ทั่วไปในหน้าอาจก่อให้เกิดความสับสน นอกจากนี้คำที่ใช้สำหรับการเชื่อมโยงจะต้องเข้าใจง่ายมีความชัดเจนและไม่สั้นจนเกินไป นอกจากนี้ในแต่ละเว็บเพจที่สร้างขึ้นมาควรมี จุดเชื่อมโยงกลับมายังหน้าแรกของเว็บไซต์ที่กำลังใช้งานอยู่ด้วย ทั้งนี้เผื่อว่าผู้ใช้เกิดหลงทาง และไม่ทราบว่าจะทำอย่างต่อไปดีจะได้มีหนทางกลับมาสู่จุดเริ่มต้นใหม่ ระวังอย่าให้มีหน้าที่ไม่มีการเชื่อมโยง (Orphan Page) เพราะจะทำให้ผู้ใช้ไม่รู้จะทำอย่างไรต่อไป

4) ความเหมาะสมในหน้าจอ เนื้อหาที่นำเสนอในแต่ละหน้าจอควรสั้น กระชับ และทันสมัย หลีกเลี่ยงการใช้หน้าจอที่มีลักษณะการเลื่อนขึ้นลง (Scrolling) แต่ถ้าจำเป็นต้องมี ควรจะให้ข้อมูลที่มี ความสำคัญอยู่บริเวณด้านบนสุดของหน้าจอ หลีกเลี่ยงการใช้กราฟิกด้านบนของหน้าจอ เพราะถึงแม้จะดูสวยงาม แต่จะทำให้ผู้ใช้เสียเวลาในการได้รับข้อมูลที่ต้องการ แต่หากต้องมีการใช้ภาพประกอบก็ควรใช้เฉพาะที่มีความสัมพันธ์กับเนื้อหาเท่านั้น นอกจากนี้การใช้รูปภาพเพื่อเป็นพื้นหลัง (Background) ไม่ควรเน้นสีสันที่ฉูดฉาดมากนัก เพราะอาจจะไปลดความเด่นชัดของเนื้อหาลง ควรใช้ภาพที่มีสีอ่อน ๆ ไม่สว่างจนเกินไปรวมไปถึงการใช้เทคนิคต่าง ๆ เช่น ภาพเคลื่อนไหว หรือตัวอักษรวิ่ง (Marquees) ซึ่งอาจจะเกิดการรบกวนการอ่านได้ ควรใช้เฉพาะที่

จำเป็นจริง ๆ เท่านั้นตัวอักษรที่นำมาแสดงบนจอภาพควรเลือกขนาดที่อ่านง่าย ไม่มีสีสันและลวดลายมากเกินไป

5) ความรวดเร็ว ความรวดเร็วเป็นสิ่งสำคัญประการหนึ่งที่ส่งผลต่อการเรียนรู้ ผู้ใช้จะเกิดอาการเบื่อหน่ายและหมดความสนใจกับเว็บที่ใช้เวลาในการแสดงผลนาน สาเหตุสำคัญที่จะทำให้การแสดงผลนานคือการใช้ภาพกราฟิกหรือภาพเคลื่อนไหว ซึ่งแม้ว่าจะช่วยดึงดูดความสนใจได้ดี ฉะนั้นในการออกแบบจึงควรหลีกเลี่ยงการใช้ภาพขนาดใหญ่ หรือภาพเคลื่อนไหวที่ไม่จำเป็น และพยายามใช้กราฟิกแทนตัวอักษรธรรมดาให้น้อยที่สุด โดยไม่ควรใช้มากเกินกว่า 2 – 3 บรรทัดในแต่ละหน้าจอ

2.2.4 ทฤษฎีเกี่ยวกับชุดคำสั่ง Python

ภาษาไพทอน (Python programming language) เป็นภาษาโปรแกรมแบบอินเทอร์พรีเตอร์ที่สร้างโดย กีโด ฟาน รอสซัม (Guido van Rossum) ในพ.ศ. 2533 ปัจจุบันดูแลโดย มูลนิธิซอฟต์แวร์ไพทอนและมีจุดเด่นของภาษาไพทอนเป็นภาษาสคริปต์ ทำให้ใช้เวลาในการเขียนและคอมไพล์ไม่มาก ทำให้เหมาะกับงานด้านการดูแลระบบ (System administration) เป็นอย่างยิ่ง ได้มีการสนับสนุนภาษาไพทอนโดยเป็นส่วนหนึ่งของระบบปฏิบัติการยูนิกซ์, ลินุกซ์ และสามารถติดตั้งให้ทำงานเป็นภาษาสคริปต์ของวินโดวส์ ผ่านระบบ Windows Script Host ได้อีกด้วย และ Python เองก็ได้ถูกนำมาพัฒนา Web application อย่างแพร่หลาย ซึ่งมี Framework สำหรับทำเว็บของ Python ที่ได้รับความนิยมอย่างมากคือ Django

2.2.4.1 ไวยากรณ์ของ Python ไวยากรณ์ของไพทอนได้กำจัดการใช้สัญลักษณ์ที่ใช้ในการแบ่งบล็อกของโปรแกรม และใช้การย่อหน้าแทน ทำให้สามารถอ่านโปรแกรมที่เขียนได้ง่าย นอกจากนั้นยังมีการเขียน docstring ซึ่งเป็นข้อความสั้นๆ ที่ใช้อธิบายการทำงานของฟังก์ชัน, คลาส , และโมดูลอีกด้วย

2.2.5 ทฤษฎีเกี่ยวกับการทำ Classification

Classification หรือเทคนิคการจำแนก เป็นหนึ่งใน Machine Learning ที่ช่วยในการทำเหมืองข้อมูล อยู่ในกลุ่มของ Supervised Learning โดยเพื่อน ๆ จำเป็นต้องกำหนดวัตถุประสงค์ เพื่อให้โมเดลสามารถเรียนรู้จากข้อมูลที่เรานำเข้า และให้ผลลัพธ์ที่ตอบคำถามตามวัตถุประสงค์ได้ ซึ่งคำตอบที่ได้จาก Classification Model จะอยู่ในรูปแบบค่าที่ไม่ต่อเนื่อง (Discrete Value) ทำให้

สามารถตอบคำถามที่กำหนดคำตอบไว้ล่วงหน้าได้ และประเภทของ Classification Model มี 3 ประเภทด้วยกัน ดังนี้

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

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

3) Multi-Label Classification (จำแนกแบบหลายเลเบล) Multi-Class และ Multi-Label ในส่วนของ Multi- Class เป็นการจำแนกข้อมูลที่มีคุณสมบัติเหมือนกัน แต่ให้ผลลัพธ์ที่แตกต่างกันหลายประเภท ส่วน Multi-Label เป็นการจำแนกข้อมูลที่อาจมีคุณสมบัติเหมือนกันแต่ผลลัพธ์สามารถแตกต่างกันได้ ซึ่ง Multi-Label จะมีความซับซ้อนในการวิเคราะห์มากกว่า

2.2.6 ทฤษฎีเกี่ยวกับการ Visualization

Visualization หรือ การแสดงข้อมูลเชิงภาพ คือ การนำข้อมูลที่เป็นตัวเลข สถิติ หรือข้อมูลที่ซับซ้อน มาแปลงให้อยู่ในรูปแบบของภาพ กราฟ แผนภูมิ หรือสื่อที่มองเห็นได้ เพื่อให้เข้าใจข้อมูลได้ง่าย รวดเร็ว และมีประสิทธิภาพมากขึ้น Visualization เข้าใจง่ายขึ้นมนุษย์เราเข้าใจภาพได้เร็วกว่าตัวเลข ทำให้การตัดสินใจและวิเคราะห์ข้อมูลทำได้ง่ายขึ้น เห็นภาพรวมช่วยให้เห็นภาพรวมของข้อมูลทั้งหมด รวมถึงความสัมพันธ์ของข้อมูลต่างๆ ได้อย่างชัดเจนค้นพบแนวโน้มช่วยให้สังเกตเห็นแนวโน้ม หรือรูปแบบต่างๆ ของข้อมูลได้ง่ายขึ้นสื่อสารได้มีประสิทธิภาพสามารถนำเสนอข้อมูลให้ผู้อื่นเข้าใจได้ง่ายและรวดเร็วขึ้น

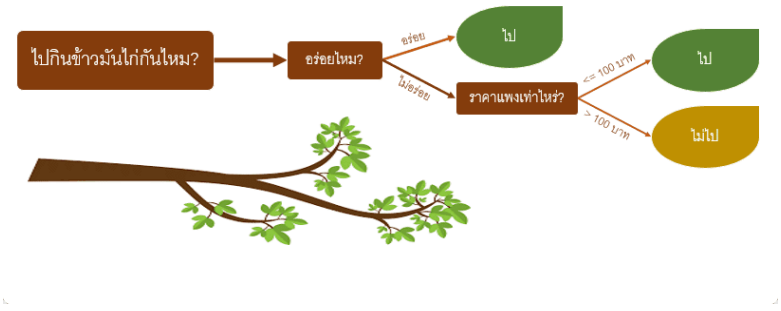
2.2.5.1 วิธีการ Visualization

การใช้ Visual เพื่อทำการค้นหาข้อมูลนั้นผู้ใช้จะทำขั้นตอนหลักๆอยู่ 3 ขั้นตอน คือ Overview First, Zoom and Filter และ Detail on Demand โดยอันดับแรก ผู้ใช้ต้องการที่จะดูข้อมูลภาพรวมทั้งหมดซึ่งหลังจากดูภาพรวมทั้งหมดแล้วผู้ใช้ก็จะทำการตัดสินใจเลือกรูปแบบหรือกลุ่มข้อมูลที่สนใจซึ่งก็จะมาถึงขั้นตอนในการวิเคราะห์ข้อมูลผู้ใช้ก็จะทำการเจาะลึกถึงข้อมูลใน

รายละเอียดซึ่ง Visualization Technology ก็จะอ้างอิงหรือพัฒนาจากขั้นตอนเหล่านี้ซึ่ง Visualization Technique มีประโยชน์มากในการแสดงภาพรวมหรือแสดงข้อมูลย่อยที่ผู้ใช้ต้องการโดยอาจจะใช้หลายๆวิธีการรวมกันเพื่อให้ข้อมูลที่ผู้ใช้ต้องการ ซึ่งช่วยลดช่องว่างของกิจกรรมที่ใช้ในการดึงข้อมูลต่างๆ ไปใช้ซึ่งลักษณะของข้อมูลที่สามารถนำมาผ่านกระบวนการของ Visualization มีลักษณะต่างๆ มากมายดังนี้ ข้อมูล 1D ได้แก่ ข้อมูลที่เกี่ยวข้องกับช่วงเวลาใดเวลาหนึ่ง,ข้อมูล 2D ได้แก่ ข้อมูลที่เกี่ยวข้องกับแผนที่ภูมิศาสตร์, Multi Dimensional ได้แก่ Relation Table, Text และ Hypertext ได้แก่ ข้อมูลหัวข้อข่าวต่างๆ และ Web Document, Hierarchies และ Graph ได้แก่ หมายเลขโทรศัพท์ และ Web Document, Algorithms และ Software ได้แก่ Debugging Operation ซึ่งแต่ละข้อมูลก็จะมีวิธีการที่ช่วยในการจัดการแสดงผลข้อมูลที่หลากหลาย

2.2.7 ทฤษฎีเกี่ยวกับการการทำ Decision Tree

2.2.7.1 Decision Tree หรือ ต้นไม้ตัดสินใจ คือ การจำลองวิธีการตัดสินใจของมนุษย์ ซึ่งการตัดสินใจแต่ละครั้งของมนุษย์ มนุษย์จะแตกโจทย์หลักออกเป็นโจทย์ย่อยหลาย ๆ โจทย์ก่อนเพื่อที่จะได้ง่ายต่อการตัดสินใจ หรือจะนำเอาปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการตัดสินใจหรือเกี่ยวข้องกับโจทย์หลักมาตั้งเป็นคำถามใหม่หรือแตกเป็นโจทย์ย่อย และถามตัวเองใหม่อีกครั้ง



รูปภาพที่ 2.14 Decision Tree หรือ ต้นไม้ตัดสินใจ

(ที่มา: www.borntodev.com)

2.2.7.2 Model Decision Tree คือ Rule-Based Model ที่จะสร้างเงื่อนไข If-else ขึ้นมาจากข้อมูลในตัวแปร เพื่อที่จะแบ่งข้อมูลออกเป็นกลุ่มใหม่ที่สามารถอธิบาย Target ได้ดีที่สุด โดยการสร้างเงื่อนไข If-else ในแต่ละตัวแปร จะถูกกำหนดด้วย Objective Function ซึ่ง Model Decision Tree

มี Objective Function อยู่หลายตัว ตามประเภทของ Decision Tree นั้น ๆDecision Tree จะแบ่งออกเป็น 2 ประเภท คือRegression Tree คือ Decision Tree ที่ใช้สำหรับการทำโจทย์ Regression โดยมีค่า Residual sum of squares (RSS) เป็น Objective Function ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point) จากการ Minimize ให้ RSS มีค่าน้อยที่สุด – Residual (e_i) คือ ค่าความคลาดเคลื่อนหรือค่า Error ระหว่าง y ทุก ๆ จุดในข้อมูล กับ y_hat ที่ได้มาจากการประมาณค่าขึ้นมาการคำนวณ Residual ของข้อมูลตัวที่ i

$$e_i = y_i - \hat{y}_i$$

รูปภาพที่ 2.15 Residual (e_i) คือ ค่าความคลาดเคลื่อน

(ที่มา: www.borntodev.com)

– Residual sum of squares (RSS) คือ การวัดค่า Residual หรือ ค่า Error ของทุก ๆ จุดในชุดข้อมูล และนำมายกกำลังสอง เพื่อให้ค่า Residual เป็นบวก และเป็นการทำ Normalize ด้วย เนื่องจากถ้า y_hat มีค่ามากกว่า yi จะทำให้ค่า Residual ติดลบการคำนวณ Residual sum of squares

$$RSS = {e_1}^2 + {e_2}^2 + {e_3}^2 + \cdots + {e_n}^2$$

รูปภาพที่ 2.16 Residual sum of squares

(ที่มา: www.borntodev.com)

2.2.7.3 Classification Tree คือ Decision Tree ที่ใช้สำหรับการทำ Classification โดยจะใช้ Gini Impurity หรือ Entropy เป็น Objective Function ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point)

– Gini Impurity คือ การวัดค่า Impurity หรือ ค่าความไม่บริสุทธิ์ในการอธิบาย Target ของกลุ่มที่ถูกแบ่งออกมาจากตัวแปร นั่นหมายความว่าถ้าค่า Impurity ยิ่งต่ำก็ยิ่งแบ่งข้อมูลออกมาได้ดีนั่นเอง

– การคำนวณ Gini Impurity คือ การนำเอาผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจมาคูณด้วย (1 ลบ ค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจ)

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

รูปภาพที่ 2.17 การคำนวณ Gini Impurity

(ที่มา: www.borntodev.com)

หลังจากได้ค่า Gini Impurity ของแต่ละกลุ่มในทุก ๆ ตัวแปรแล้ว จะทำการหาค่า Weighted Gini Impurity เพื่อเลือกตัวแปรที่มีค่า Weighted Gini Impurity ต่ำที่สุดมาใช้ในการตัดสินใจก่อน เพราะสามารถ Split ข้อมูลได้ดีที่สุด การคำนวณ Weighted Gini Impurity คือ การนำเอาผลรวมของค่า Gini ในเหตุการณ์ที่เราสนใจ คูณกับจำนวนข้อมูลของ Class ที่ i ในตัวแปรที่เราสนใจ และหารด้วยจำนวนข้อมูลทั้งหมดในตัวแปรที่เราสนใจ

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

รูปภาพที่ 2.18 Weighted Gini Impurity

(ที่มา: www.borntodev.com)

2.2.7 ทฤษฎีเกี่ยวกับการการทำ Random Forest

2.2.7.1 Random Forest เป็นอัลกอริทึมประเภท Ensemble Learning ที่ใช้สำหรับงาน Classification และ Regression โดยมีแนวคิดหลักคือการรวมผลลัพธ์จาก หลาย ๆ

ต้นไม้ตัดสินใจ (Decision Trees) เข้าด้วยกัน เพื่อเพิ่มความแม่นยำและลดปัญหาการเรียนรู้แบบเฉพาะเจาะจงกับข้อมูล (Overfitting)

2.2.7.2 หลักการของ Random Forest คือการสร้างต้นไม้ตัดสินใจหลายต้นขึ้นจากชุดข้อมูลฝึก (Training Data) ที่ถูกสุ่มออกมาโดยใช้วิธี Bootstrap Sampling ซึ่งหมายถึงการสุ่มข้อมูลบางส่วนจากชุดข้อมูลเดิม (อาจมีการซ้ำได้) เพื่อใช้สร้างต้นไม้แต่ละต้น จากนั้นในแต่ละขั้นตอนของการแบ่งโหนด (Node Splitting) ระบบจะสุ่มเลือกเฉพาะบางตัวแปร (Attributes) จากทั้งหมดมาใช้ในการตัดสินใจเพื่อลดความเอนเอียงของโมเดลเมื่อสร้างต้นไม้ครบทุกต้นแล้ว การทำนายผลลัพธ์จะอาศัยวิธี Voting (ในกรณี Classification) หรือ Averaging (ในกรณี Regression) โดยผลลัพธ์สุดท้ายจะเป็นค่าที่ได้จากเสียงข้างมากหรือค่าเฉลี่ยจากทุกต้นไม้ในป่า (Forest)

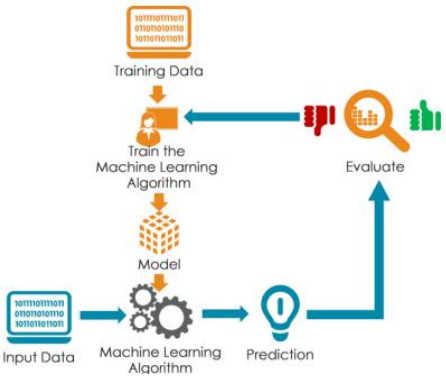
2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

2.3.1 RapibMiner Studio

2.3.1.1 RapidMiner เป็นโปรแกรมช่วยวิเคราะห์ข้อมูลสำหรับงาน Data Science ที่ไม่จำเป็นต้องมีพื้นฐานการเขียนโปรแกรม ใช้งานง่ายและมีประสิทธิภาพ ครอบคลุมตั้งแต่การเก็บข้อมูล การทำความสะอาดและเตรียมข้อมูล การวิเคราะห์ สร้างโมเดล ประเมินผล จนถึงการนำไปใช้งานจริง สามารถค้นหา Insights ที่เป็นประโยชน์ต่อการตัดสินใจหรือแก้ปัญหาเฉพาะด้าน โปรแกรม RapidMiner ประกอบด้วยหลายส่วน ทั้ง RapidMiner Studio สำหรับสร้างและประมวลผลโมเดลบนเดสก์ท็อป RapidMiner Server สำหรับจัดการและเผยแพร่โมเดลบนเซิร์ฟเวอร์ RapidMiner AI Hub ที่เป็นแพลตฟอร์มสำหรับสร้าง จัดการ และเผยแพร่โมเดลและ AI และ RapidMiner Go ซึ่งเป็นแอปพลิเคชันบนเว็บที่ช่วยให้ผู้ใช้สามารถสร้าง จัดการ และเผยแพร่โมเดล รวมถึงทำงานร่วมกับผู้อื่นได้

โปรแกรมมีส่วนประกอบหลักที่สำคัญ ได้แก่ Repository สำหรับจัดการไฟล์ข้อมูลและ Process Operators เป็นโมดูลฟังก์ชันต่างๆ ที่สามารถเชื่อมต่อเป็น Process เพื่อทำงานด้าน Machine Learning Parameters สำหรับปรับค่าต่างๆ ของแต่ละ Operator และ Help ที่ให้คำอธิบายเครื่องมือใน Process

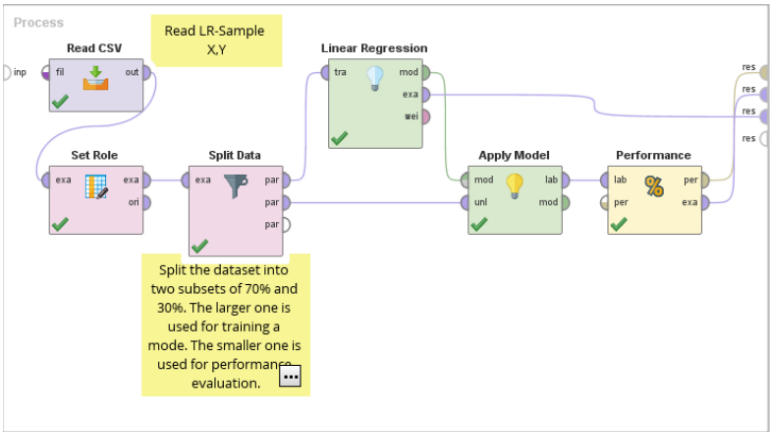
2.3.1.2 ขั้นตอนการทำงานของโปรแกรม RapidMiner ใช้งานในรูปแบบ Workflow-Based Data Science โดยผู้ใช้สร้างกระบวนการวิเคราะห์ข้อมูลผ่านการลากและวาง Operators เชื่อมต่อกันเป็นเครือข่าย (Process Flow) โดยทั่วไปเริ่มจากการนำเข้าข้อมูล ทำความสะอาดและเตรียมข้อมูล สร้างและฝึกโมเดล ประเมินผลโมเดล และสุดท้ายนำไปใช้งานจริง



รูปภาพที่ 2.19 ขั้นตอนการทำงานของโปรแกรม RapidMiner

(ที่มา: https://orgweb.ldd.go.th/)

2.3.1.3 RapidMiner มีโมเดล Machine Learning หลากหลายประเภท ได้แก่ Supervised Learning สำหรับการทำนาย เช่น Decision Tree, Random Forest, SVM และ Deep Learning, Unsupervised Learning สำหรับการจำแนกกลุ่ม เช่น K-Means Clustering, DBSCAN, PCA และ Text Mining สำหรับวิเคราะห์ข้อความ เช่น Sentiment Analysis หรือ Topic Modeling การสร้างโมเดลเริ่มจากการใช้ Operator Split Data เพื่อแบ่งข้อมูลเป็น Training Set และ Testing Set จากนั้นเลือกโมเดลที่ต้องการจาก Operators Panel แล้วเชื่อมต่อกับโมดูล Train Model เพื่อฝึกโมเดลกับ Training Set เช่น Linear Regression สำหรับข้อมูลเชิงปริมาณ หรือ Decision Tree/Random Forest สำหรับข้อมูลเชิงคุณภาพ หลังจากฝึกเสร็จ ใช้ Operator Apply Model ทดสอบโมเดลกับ Testing Set แล้วปรับพารามิเตอร์ด้วย Optimize Parameters สุดท้ายกด Run เพื่อฝึกโมเดลและสร้างผลลัพธ์



รูปภาพที่ 2.20 การสร้างโมเดล

(ที่มา: https://orgweb.ldd.go.th/)

2.3.2 Visual Studio Code

เป็นโปรแกรมแก้ไขซอร์สโค้ดที่พัฒนาโดยไมโครซอฟท์สำหรับ Windows, Linux และ macOS[5] มีการสนับสนุนสำหรับการดีบัก การควบคุม Git ในตัวและ GitHub การเน้นไวยากรณ์ การเติมโค้ดอัจฉริยะ ตัวอย่าง และ code refactoring มันสามารถปรับแต่งได้หลายอย่างให้ผู้ใช้สามารถเปลี่ยนธีม แป้นพิมพ์ลัด การตั้งค่า และติดตั้งส่วนขยายที่เพิ่มฟังก์ชันการทำงานเพิ่มเติม ซอร์สโค้ดนั้นฟรีและโอเพนซอร์สและเผยแพร่ภายใต้สิทธิ์การใช้งาน MIT[6] ไบนารีที่คอมไพล์แล้วเป็นฟรีแวร์และฟรีสำหรับการใช้ส่วนตัวหรือเพื่อการค้า

คุณสมบัติหลักของ VS Code

1) น้ำหนักเบาและรวดเร็ว เป็นโปรแกรมแก้ไขโค้ดที่ทำงานได้รวดเร็วและตอบสนองได้ดี

2) รองรับหลายภาษา มาพร้อมการรองรับ JavaScript, TypeScript, และ Node.js ในตัว และมีส่วนขยายสำหรับภาษาอื่นๆ เช่น C++, C#, Java, Python, PHP, Go

3) คุณสมบัติการเขียนโค้ด มีการเน้นไวยากรณ์, การเติมโค้ดอัจฉริยะ (IntelliSense) การรีแฟกเตอร์โค้ด, และการดีบักแบบอินเทอร์แอคทีฟ

4) รองรับ Git มีการควบคุม Git และ GitHub ในตัว ช่วยให้นักพัฒนาสามารถทำงานกับระบบควบคุมเวอร์ชันโดยไม่ต้องออกจากโปรแกรม

5) ปรับแต่งได้สูง ผู้ใช้สามารถเปลี่ยนธีม, แป้นพิมพ์ลัด, และติดตั้งส่วนขยายเพิ่มเติมเพื่อเพิ่มฟังก์ชันการทำงาน

6) ฟรีและโอเพนซอร์ส ซอร์สโค้ดเป็นโอเพนซอร์สภายใต้สิทธิ์การใช้งาน MIT และไบนารีที่คอมไพล์แล้วเป็นฟรีแวร์

7) ทำงานข้ามแพลตฟอร์ม สามารถดาวน์โหลดและใช้งานได้บนระบบปฏิบัติการ Windows, macOS, และ Linux

2.3.3 HTML สำหรับสร้างหน้าเว็บไซต์

HTML ย่อมาจาก HyperText Markup Language เป็นภาษามาร์กอัปที่ใช้กำหนดโครงสร้างของหน้าเว็บไซต์ อธิบายให้เข้าใจง่ายขึ้น ภาษา HTML คือ ภาษาคอมพิวเตอร์ที่ถูกเขียนขึ้น ทำให้ Web Browser คือ โปรแกรมที่ใช้แปลงภาษา HTML สามารถค้นหา ดึงข้อมูล และแสดงเนื้อหาบนเว็บไซต์ในรูปแบบที่ผู้พัฒนาออกแบบไว้ได้ ไม่ว่าจะเป็นในรูปแบบของข้อมูลตัวอักษร ภาพ วิดีโอ เสียง และไฟล์ต่าง ๆ รวมถึงการเชื่อมโยงลิงก์ไปยังหน้าอื่นในเว็บไซต์ด้วย

2.3.3.1 โครงสร้างของ HTML ประกอบไปด้วยอะไรบ้าง

1.<!DOCTYPE html> เป็นคำสั่งที่ใช้บอกเว็บเบราว์เซอร์ว่า เอกสารนี้ใช้ HTML เวอร์ชันไหน โดยจะแจ้งให้เบราว์เซอร์ทราบว่าเอกสารนี้ใช้ HTML5 (เวอร์ชันล่าสุด) และป้องกันปัญหาการแสดงผลผิดพลาด (Browser Rendering Mode) ซึ่งโค้ด <!DOCTYPE html> ต้องอยู่บรรทัดแรกของไฟล์ HTML เท่านั้น

2. <html> เป็นโครงสร้างหลักของ HTML ทำหน้าที่เป็นตัวครอบองค์ประกอบทั้งหมดของเว็บเพจ โดยทุกองค์ประกอบของเว็บต้องอยู่ภายใน <html> และควรกำหนด lang=”th” หรือ lang=”en” ตามภาษาของเว็บไซต์ เพื่อช่วยให้ SEO และ Accessibility (การเข้าถึงสำหรับผู้ใช้) ดียิ่งขึ้น

3. <head> เป็นส่วนที่ใช้เก็บข้อมูลที่ไม่แสดงผลโดยตรงบนหน้าเว็บ แต่มีความสำคัญอย่างมากสำหรับ SEO, การโหลดเว็บ และการแสดงผล โดยองค์ประกอบที่อยู่ใน <head> จะมีอยู่หลายส่วนด้วยกัน

4. <body> เป็นส่วนเนื้อหาหลักของหน้าเว็บไซต์ที่แสดงทุกอย่างที่ผู้ใช้สามารถมองเห็นและโต้ตอบได้ เช่น ข้อความ, รูปภาพ, วิดีโอ, ฟอร์ม, ลิงก์, ปุ่ม เป็นต้น โดยองค์ประกอบหลักที่อยู่ภายใน <body> จะมีอยู่ด้วยกันหลายส่วน

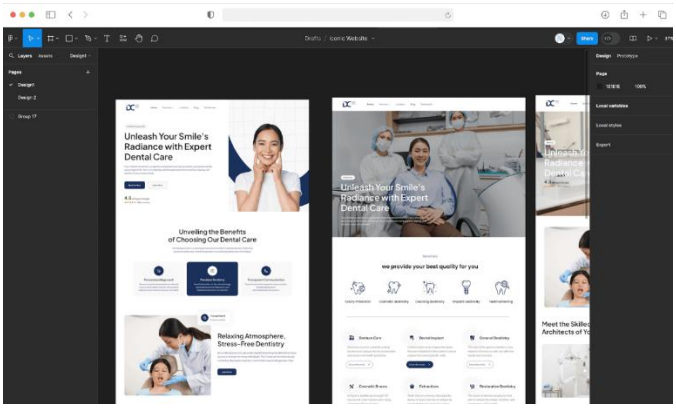
2.3.4 Figma

Figma คือแพลตฟอร์มออกแบบกราฟิกที่ให้บริการการออกแบบอินเตอร์เฟซและประสบการณ์ผู้ใช้งาน (UI/UX) ผ่านระบบคลาวด์ ทำให้การเข้าถึงและการทำงานร่วมกันเป็นไปได้อย่างราบรื่น ผู้ใช้สามารถเริ่มต้นออกแบบจากการวาดโครงร่างหรือ Wireframe ไปจนถึงสร้าง Design Systems ที่สมบูรณ์ และ Prototype ที่สามารถทดลองใช้และโต้ตอบได้จริง มันเป็นเครื่องมือที่ทำให้การออกแบบไม่ต้องจำกัดอยู่เพียงในสตูดิโอหรือหน้าจอคอมพิวเตอร์ของนักออกแบบเท่านั้น แต่ยังรวมไปถึงการทำงานร่วมกันแบบเรียลไทม์ที่สามารถแสดงความคิดเห็นและปรับเปลี่ยนได้ตลอดเวลาโดย Figma สามารถทำได้โดยใช้ฟีเจอร์หลากหลายของแพลตฟอร์มนี้

1) Interface Figma มีอินเตอร์เฟซที่ใช้งานง่าย ทำให้คุณสามารถเริ่มต้นโครงการและออกแบบ UX/UI ได้อย่างลื่นไหล ขั้นตอนแรกคือการสร้าง Frames สำหรับแต่ละหน้าของเว็บหรือแอพพลิเคชัน โดยเลือกขนาดที่เหมาะสมตามอุปกรณ์ที่จะแสดงผล

2) Prototype และ Interactions เมื่อมีหน้าต่างๆ สามารถเริ่มสร้างโปรโตไทป์ได้โดยการเชื่อมการโต้ตอบระหว่าง Frames เช่น การคลิก การเลื่อน หรือการเปลี่ยนสถานะ นี่จะช่วยให้สามารถแสดงการไหลของการใช้งานและประสบการณ์ผู้ใช้ได้อย่างชัดเจน

3) Smart Animate Figma ช่วยให้สร้าง Animation ระหว่าง Frames ที่แตกต่างกันได้ง่ายดาย เพิ่มความสมจริงให้แอพพลิเคชัน และช่วยให้ผู้ใช้เข้าใจการโต้ตอบแบบต่างๆ



ภาพที่ 2.19 แสดงตัวอย่างหน้าโปรแกรม

(ที่มา: https://insights.nconnect.asia/)

2.4 วรรณกรรมที่เกี่ยวข้อง

เจนจิรา สุกใส และ นิภาพร ชนะมาร (2566) ศึกษาการจัดกลุ่มข้อมูลเศรษฐกิจครัวเรือนด้วยการประยุกต์ใช้เทคนิคการจัดกลุ่ม แบบ K–Means Clustering มีการเก็บข้อมูลออกเป็น 10 ส่วน คือ 1) ข้อมูลทั่วไปครัวเรือน 2) ทรัพย์สินของครัวเรือน 3) อาชีพและรายได้ของครัวเรือน 4) รายจ่ายของครัวเรือน 5) หนี้สินของครัวเรือน 6) ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID – 19) 7) การใช้เทคโนโลยีสารสนเทศ 8) การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน 9) การเข้าร่วมโครงการที่ผ่านมา ย้อนหลัง 3 ปี และ 10) ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม ซึ่งผู้วิจัยมีความสนใจที่จะนำข้อมูลมาทำการแบ่งระดับ เศรษฐกิจครัวเรือน เพื่อส่งเสริมงานวิจัยและใช้ข้อมูลประกอบการตัดสินใจ ในการวางแผนพัฒนาเศรษฐกิจครัวเรือนในชุมชน โดยใช้ K–Means Clustering พบว่า มีการจัดกลุ่มได้ 3 กลุ่ม ตามระดับเศรษฐกิจกลุ่มที่มีเศรษฐกิจน้อย 304 ครัวเรือนกลุ่มที่มีเศรษฐกิจปานกลาง 544 ครัวเรือน กลุ่มที่มีเศรษฐกิจสูง 903 ครัวเรือนวิจัยนี้ใช้พัฒนาด้วยโปรแกรม RapidMiner เวอร์ชัน 9.10 โดยอิงจากกระบวนการ CRISP–DM รวมถึงการทำความสะอาดข้อมูล, คัดเลือกข้อมูล, ลดมิติข้อมูล, และแปลงรูปแบบข้อมูล เพื่อให้ได้ข้อมูลทดลองที่มีคุณภาพ

สุชาวดี ปลั่งศรี (2563) ศึกษาการจัดกลุ่มผลิตภัณฑ์สำหรับออกแบบขนาดกล่องบรรจุภัณฑ์ เพื่อเป็นแนวทางการลดต้นทุนการสั่งซื้อ โดยเริ่มจากการศึกษาข้อมูลผลิตภัณฑ์เพื่อให้ทราบถึงปัจจัยที่ส่งผลต่อการออกแบบขนาดกล่องบรรจุภัณฑ์ ซึ่งมีข้อมูล 3 ชุดข้อมูล ชุดข้อมูลละ 437 ข้อมูล หลังจากนั้นใช้การแบ่งกลุ่มข้อมูลแบบเคมีน (K–means Clustering Analysis) เพื่อจัดกลุ่มผลิตภัณฑ์และออกแบบขนาดกล่องบรรจุภัณฑ์จากนั้น วิเคราะห์หาจำนวนกลุ่มที่เหมาะสมที่สุดโดยใช้เรื่องต้นทุนบรรจุภัณฑ์เป็นเกณฑ์เพื่อจัดกลุ่มผลิตภัณฑ์ครั้งละ1 ชุดข้อมูล โดยใช้โปรแกรมทางสถิติ SPSS และทำการเก็บข้อมูลจากกระบวนการผลิตเพื่อหาระยะเผื่อจากการห่อของพนักงาน เพื่อออกแบบขนาดกล่องบรรจุภัณฑ์ใหม่ จากนั้นวิเคราะห์หาปริมาณผลิตภัณฑ์ที่มากที่สุดที่สามารถซ้อนทับกันบนพาเลก สุดท้ายจะศึกษาเรื่อง การบรรจุหีบห่อในการขนถ่ายวัสดุ (Packaging in Material Handling) เพื่อหาปริมาณผลิตภัณฑ์ที่ มากที่สุดที่สามารถเรียงซ้อนทับกันบนพาเลทได้เพื่อเป็นข้อมูลให้บริษัทสำหรับนำไปพัฒนาระบบการ ขนส่งและการจัดเก็บให้มีประสิทธิภาพมาก

ณัฏฐ์ฏาพร สายคำวงษ์ (2563) ศึกษาการจัดกลุ่มหลักทรัพย์ด้วยอัลกอริทึมการจัดกลุ่มวิธี K-means ซึ่งสามารถนำมาประยุกต์ใช้ในการจัดกลุ่มหลักทรัพย์ที่มีการเปลี่ยนแปลงคล้ายกันได้ โดยทำการเก็บรวบรวมข้อมูลย้อนหลังของหลักทรัพย์ ใน SET 50 ของตลาดหลักทรัพย์แห่งประเทศไทย ซึ่งเป็นหุ้นสามัญ 50 บริษัท โดยส่วนใหญ่เป็นหุ้นที่มีปัจจัยพื้นฐานดี มีอัตราการเติบโตของบริษัทอย่างต่อเนื่อง โดยข้อมูลที่นำมาวิเคราะห์ประกอบด้วย ราคาสูงสุด ราคาต่ำสุด ราคาเปิดตลาด และราคาปิดตลาดจากการศึกษาพบว่า วิธีการดังกล่าวสามารถจัดกลุ่มหลักทรัพย์ที่มีความคล้ายคลึงกันได้ซึ่งกลุ่มหลักทรัพย์ที่จัดอยู่ด้วยกันจะถูกนำไปกระจายความเสี่ยงในการจัดพอร์ตการลงทุน

ธนะวัฒน์ วรรณประภา (2564) ศึกษาการการใช้เทคนิคเหมืองข้อมูลเพื่อคัดกรองบุคคลที่มีแนวโน้มสัมฤทธิ์ผลในการศึกษาระดับปริญญาตรี สาขาเทคโนโลยีการศึกษากระบวนการคัดเลือกบุคคลเพื่อศึกษาต่อในระดับปริญญาตรีคัดเลือกบุคคลที่มีความพร้อมทั้งความรู้ และความพร้อมทางด้านทักษะ ด้วยเทคนิค Decision Tree เก็บข้อมูลจากบัณฑิตที่สำเร็จการศึกษาระหว่างปีการศึกษา 2554-2558 จากระบบกองทะเบียนและประมวลผลการศึกษา จำนวน 738 คน ซึ่งถูกใช้เป็นชุดข้อมูลที่ใช้ในการศึกษา โดยแบ่งชุดข้อมูลสำหรับการสอนและทดสอบด้วยอัตราส่วน 70:30 75:25 และ 80:20 ได้ใช้วิธีการวิเคราะห์ความถดถอยเพื่อคัดกรองจำนวนแอททริบิวต์เพื่อได้ค่าที่ดีสุด

ไพชยนต์ คงไชย (2564) ศึกษาการพัฒนาวิธีเพื่อจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์แบบคลุมเครือที่กะทัดรัด การทำเหมืองข้อมูลด้วยวิธีการจำแนกประเภทข้อมูลเพื่อนำโมเดล (Model) ที่ได้มาทำกระบวนการเรียนรู้มาใช้ในการทำนายข้อมูลในอนาคต ซึ่งให้ความสนใจที่จะพัฒนาประสิทธิภาพขั้นตอนการจำแนกประเภทข้อมูล เพื่อให้ความแม่นยำและโมเดลที่ดี มีการนำเทคนิคการทำเหมืองข้อมูลนำมาหาความกฎความสัมพันธ์ (Association Rule Mining) จำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์ (Associative Classification) และมีประโยชน์อย่างมากในการระบุปัจจัยที่จะทำให้เกิดบางสิ่งในอนาคตเช่น ปัจจัยที่จะทำให้เกิดไฟไหม้ป่าปัจจัยที่ทำให้เกิดแผ่นดินไหว โดยเป็นอัลกอริทึมที่นำเทคนิคหากฎความสัมพันธ์มาช่วยในการวิเคราะห์หาความสัมพันธ์ของข้อมูลและใช้เทคนิคฟัซซีเซตมาช่วยในการแก้ปัญหาข้อมูลตัวเลขนำเข้าที่มี

ลักษณะเป็นค่าต่อเนื่อง อีกทั้งเทคนิคดังกล่าวยังเพิ่มความถูกต้องในการจำแนกข้อมูลที่มีลักษณะซ้อนทับกันมากอีกด้วย

2.5 บทสรุป

จากแนวคิด ทฤษฎี เครื่องมือ และงานวิจัยที่เกี่ยวข้องทั้งหมด ผู้วิเคราะห์ข้อมูลได้เลือกใช้กระบวนการ CRISP-DM ในการวิเคราะห์ข้อมูลใช้เทคนิค Decision Tree และ Random Forest (Classification) ผ่าน RapidMiner เพื่อจำแนกระดับเงินเดือนของ Data Scientist ตามปัจจัยสำคัญ เช่น ประสบการณ์ ระดับการศึกษา สถานที่ทำงาน และทักษะเฉพาะด้านข้อมูลที่ได้ถูกนำมาพัฒนาเป็น เว็บแอปพลิเคชัน พร้อม Visualization ในรูปแบบกราฟและตาราง เพื่อให้ผู้ใช้งานสามารถเข้าใจความแตกต่างของระดับเงินเดือนและลักษณะเฉพาะของแต่ละกลุ่มได้อย่างชัดเจนผลลัพธ์ช่วยให้ผู้ใช้งานทั่วไปสามารถเข้าถึงข้อมูล วิเคราะห์แนวโน้มตลาดแรงงาน วางแผนพัฒนาทักษะ และใช้ประกอบการตัดสินใจเชิงกลยุทธ์ได้อย่างมีประสิทธิภาพ