บทที่ 2
แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้อง

การวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด ในบทนี้เป็นการนำเสนอเกี่ยวกับ แนวคิด ทฤษฎี เครื่องมือและวรรณกรรมที่เกี่ยวข้องของการวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด และการแสดงผลข้อมูลบนเว็บไซต์ ซึ่งได้รวบรวมการศึกษา เอกสารงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล เพื่อใช้เป็นแนวทางการศึกษาประกอบด้วยรายละเอียดตามลำดับ ดังนี้

2.1 แนวคิด

- 2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data Analytics)
- 2.1.2 แนวคิดเกี่ยวกับการทำเหมืองข้อมูล (Data Mining)
- 2.1.3 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data Visualization)

2.2 ทฤษฎี

- 2.2.1 หลักการทำเหมืองข้อมูล
- 2.2.2 การจำแนกประเภทข้อมูล (Classification)
- 2.2.3 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์
- 2.2.4 ปัจจัยเสี่ยงหลักที่มีผลต่อการเกิดโรคเบาหวาน

2.3 เครื่องมือในการออกแบบและวิเคราะห์ข้อมูล

- 2.3.1 โปรแกรม Rapid Miner Studio
- 2.3.2 กระบวนการวิเคราะห์ข้อมูลด้วย (CRISP–DM)
- 2.3.3 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)
- 2.3.4 เทคนิคนาอีฟเบย์ (Naïve Bayes)
- 2.3.5 เทคนิคป่าสุ่ม (Random Forest)
- 2.3.6 เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)
- 2.3.7 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

2.4 วรรณกรรมที่เกี่ยวข้อง

2.5 บทสรุป

2.1 แนวคิด

2.1.1 แนวคิดเกี่ยวกับการวิเคราะห์ข้อมูล (Data Analytics)

Data Analytics คือ การวิเคราะห์ข้อมูลที่มีอยู่ตั้งแต่อดีตจนถึงปัจจุบัน เพื่อทำนายอนาคต ที่เป็นประโยชน์ในการพัฒนาการตลาดให้ตรงใจลูกค้ามากยิ่งขึ้น Data Analytics เป็นเครื่องมือสำหรับธุรกิจ (Business Intelligence) การทำ Data Analytics นี้ไม่จำเป็นต้องเป็นธุรกิจขนาดใหญ่เท่านั้น แต่ธุรกิจขนาดกลางและเล็กก็สามารถทำได้เช่นกัน สำหรับรูปแบบของการวิเคราะห์ข้อมูล (Data Analytics) สามารถแบ่งได้ดังนี้

1) การวิเคราะห์แบบเชิงบรรยาย (Descriptive Analytics) เพื่อแสดงผลที่เกิดขึ้นหรือกำลังจะเกิดขึ้น จากข้อมูลในอดีต ในลักษณะที่เข้าใจง่ายสามารถสร้างขึ้นได้ด้วยตนเอง เช่น รายงาน แผนภูมิ กราฟ ตาราง เป็นต้น ซึ่งจะช่วยให้เข้าใจการเปลี่ยนแปลงที่เกิดขึ้นกับองค์กรได้ดียิ่งขึ้น เช่น การเปลี่ยนแปลงของราคาสินค้ารายปี การเติบโตของยอดขายรายเดือน การเปรียบเทียบยอดขายในแต่ละสาขาหรือแต่ละช่องทาง การเปรียบเทียบจำนวนผู้ใช้งานเว็บไซต์ในแต่ละช่วงเวลา ซึ่ง Descriptive Analytics คือการอธิบายสิ่งที่เกิดขึ้นตามวัตถุประสงค์และช่วงเวลาที่กำหนด

2) การวิเคราะห์แบบเชิงวินิจฉัย (Diagnostic Analytics) เป็นการอธิบายถึงสาเหตุของสิ่งที่เกิดขึ้น ปัจจัยต่าง ๆ และความสัมพันธ์ของปัจจัยหรือตัวแปรต่างๆที่มีความสัมพันธ์ต่อกันของสิ่งที่เกิดขึ้น ตัวอย่างเช่น ความสัมพันธ์ระหว่างยอดขายต่อกิจกรรมทางการตลาดแต่ละประเภท ซึ่งเป็นก้าวใหม่ที่ช่วยเสริมให้ตัดสินใจไปในทางที่ถูกต้อง

3) การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) เป็นการวิเคราะห์เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูลที่ได้เกิดขึ้นแล้วกับแบบจำลองทางสถิติ หรือ เทคโนโลยีปัญญาประดิษฐ์ต่างๆ (Artificial intelligence) ตัวอย่างเช่น การพยากรณ์ยอดขาย การพยากรณ์ผลประชามติ

4) การวิเคราะห์แบบให้คำแนะนำ (Prescriptive Analytics) เป็นการวิเคราะห์ข้อมูลที่มีความซับซ้อนที่สุด เป็นทั้งการพยากรณ์สิ่งต่างๆ ที่จะเกิดขึ้น ข้อดี ข้อเสีย สาเหตุ และระยะเวลาของสิ่งที่จะเกิดขึ้น ร่วมถึงการให้คำแนะนำทางเลือกต่างๆ ที่มีอยู่ และผลของแต่ละทางเลือก (BLENDATA, 2024)

2.1.2 แนวคิดเกี่ยวกับการทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases หรือ KDD) ซึ่งเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึง

ความรู้ หรือ สิ่งที่สำคัญออกมาเพื่อใช้ในการวิเคราะห์ หรือทำนายสิ่งต่างๆที่จะเกิดขึ้น ในการค้นหาความรู้นั้นความจริงที่แฝงอยู่ในข้อมูล (Knowledge Discovery) ซึ่งเป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่เรามีอยู่เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล กระบวนการ KDD แบ่งออกเป็น 6 ขั้นตอน ดังนี้

1) แนะนำการทำเหมืองข้อมูล (Introduction to Data Mining) ในชีวิตประจำวันของเราทุกคนจะต้องข้องเกี่ยวกับข้อมูลต่าง ๆ มากมายที่เราจำเป็นต้องจดจำและจดบันทึกลงบนกระดาษหรือบนอุปกรณ์ช่วยจำ ตั้งแต่อดีตจนถึงปัจจุบันมนุษยชาติ มีการบันทึกข้อมูลเรื่องราวต่าง ๆ อย่างต่อเนื่องเพื่อเก็บไว้เป็นข้อมูลทางสถิติหรือข้อมูลทางประวัติศาสตร์ เพื่อนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ต่อการวางแผนการทำงาน การกำหนดทิศทางการดำเนินงาน หรือเพื่อสนับสนุนการตัดสินใจในเรื่องต่าง ๆ เช่น การทำนายผลประกอบการของบริษัท การวางแผนงานเชิงรุกของบริษัท เป็นต้น ถ้าเราลองพิจารณาถึงข้อมูลส่วนบุคคลต่าง ๆ ที่เราต้องจัดเก็บตั้งแต่เกิด จะประกอบด้วยข้อมูลมากมาย เช่น วันเกิด น้ำหนักแรกเกิด ความสูง น้ำหนัก โรคภัย วุฒิการศึกษา ประวัติการทำงาน อายุ เงินเดือน วันแต่งงาน บันทึกค่าใช้จ่าย วันตาย เป็นต้น ข้อมูลเหล่านี้เป็นเพียงตัวอย่าง อันเล็กน้อยของข้อมูลที่มีการจดบันทึกและจัดเก็บจริงของคนคนเดียว แต่ถ้าลองคิดดู คนบนโลกใบนี้ ที่มีจำนวนกว่าหมื่นล้านคนจะมีปริมาณข้อมูลจำนวนมากมายมหาศาลเพียงใด และนอกเหนือ จากข้อมูลส่วนบุคคลแล้ว ยังมีข้อมูลแวดล้อมอื่น ๆ อีกมากมายที่อยู่รอบตัวเรา เช่น ราคาอาหาร ราคาน้ำมัน ราคาทอง ปริมาณน้ำฝนและอุณหภูมิจากสถานีวัด ภาพถ่ายจากดาวเทียม ข่าวสาร ในแต่ละวัน

2) การเตรียมข้อมูล (Data Preprocessing) ในแต่ละวันเราจะได้รับข้อมูลและสารสนเทศมากมาย โดยข้อมูลเหล่านี้อาจจะเป็นข้อมูลที่ผ่านมาและผ่านไปโดยที่เราไม่ได้สนใจ หรือบางทีอาจเป็นข้อมูลที่มีความสำคัญที่เราจะต้อง จดจำและรับทราบเอาไว้ หรือเป็นข้อมูลที่เราต้องเก็บมาวิเคราะห์ สังเคราะห์ เพื่อนำไปใช้ให้เกิดประโยชน์ต่อไป

ข้อมูล (Data) คือ ข้อเท็จจริงเกี่ยวกับเรื่องที่เราสนใจ ซึ่งอาจเป็นการจัดเก็บแบบ จดบันทึกรายวัน หรือเป็นการจัดเก็บอย่างมีระบบระเบียบในลักษณะของฐานข้อมูล ซึ่งในที่นี้ จะอธิบายข้อมูลในมุมมองของกลุ่มของค่าของข้อมูลที่อยู่รวมกัน ซึ่งจะเรียกว่า ลักษณะประจำ (Attributes) หรือตัวแปร (Variable) โดยความหมาย ลักษณะประจำ (Attributes) คือ คุณสมบัติหรือลักษณะประจำของ ข้อมูลหรือวัตถุหรือสิ่งที่เราสนใจ เช่น ลักษณะประจำอายุ ลักษณะประจำเพศ ลักษณะประจำสีตา เป็นต้น ซึ่งจะมีลักษณะและค่าแตกต่างกันไป

3) เทคนิคการจำแนก (Classification) เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูลที่ใช้เพื่อทำนายคำตอบที่เป็น ค่าเชิงคุณภาพ (Qualitative Value) หรือค่าเต็มหน่วย (Discrete Value)

หรือค่าแบบแค็ตตาล็อก (Catalogue Value) เช่น ใช่หรือไม่ใช่ ซื้อหรือไม่ซื้อ คำตอบ ก ข ค ง ระดับความพึงพอใจ ดีมาก ดี พอใช้ เป็นต้น โดยใช้หลักการการนำชุดข้อมูลที่มีอยู่มาพัฒนาโมเดลเพื่อการจำแนก และประยุกต์ ใช้หาคำตอบหรือทำนายคำตอบของข้อมูลชุดใหม่ (Unseen Objects) ที่เข้ามาโดยเทคนิคนี้ได้รับความนิยมอย่างมาก และถูกนำมาประยุกต์ใช้เพื่อสนับสนุน การตัดสินใจทางธุรกิจและทางวิทยาศาสตร์ เพราะการพยากรณ์เพื่อจำแนกว่าข้อมูลใหม่ที่เข้ามาควรจะถูกจัดหรือจำแนกให้เป็นหมวดใดเป็นสิ่งที่นำมาใช้เพื่อการวางแผนและการตัดสินใจ ในการดำเนินกิจการต่างๆ ได้ ตัวอย่างของการประยุกต์ใช้การจำแนก

เทคนิคการจำแนก (Classification) เป็นเทคนิคที่อยู่ในกลุ่มของการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่เน้นการทำนายหรือพยากรณ์ ใช้ชุดข้อมูลที่มีการกำหนดป้ายกำกับ (label) หรือจัดกลุ่มประเภทไว้แล้วเพื่อฝึกฝนโมเดล โดยอัลกอริธึมจะวิเคราะห์ข้อมูลป้อนเข้าและป้ายกำกับที่สัมพันธ์กัน และเปรียบเทียบผลลัพธ์ที่ได้กับคำตอบที่ถูกต้อง เพื่อระบุข้อผิดพลาดและปรับปรุงความแม่นยำ (ALPHASEC, 2567)

4) การวิเคราะห์การจัดกลุ่ม (Cluster Analysis) เป็นอีกหนึ่งเทคนิคของเหมืองข้อมูล ที่ได้รับความนิยมใช้ในงานด้านต่าง ๆ อย่างแพร่หลาย เช่น การจัดกลุ่มลูกค้าของบริษัท การจัดกลุ่มเอกสาร การจัดกลุ่มผู้ป่วย เป็นต้น

การจัดกลุ่มข้อมูลเป็นเทคนิคที่อยู่ในกลุ่มของการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่เน้นการบรรยายลักษณะข้อมูลมากกว่าการทำนายหรือพยากรณ์ ที่จัดเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ส่วนใหญ่งานด้านนี้มีไว้เพื่อลดขนาดหรือมิติของข้อมูลให้เป็นกลุ่มหรือคลัสเตอร์ โดยมีจุดประสงค์เพื่อรวมกลุ่มของสิ่งที่มีความคล้ายกันให้อยู่กลุ่มเดียวกัน เพื่อจะได้ทำให้ง่ายต่อการดำเนินการทางการทำธุรกิจ หรือการวิเคราะห์ปัจจัยได้เจาะจงยิ่งขึ้น เช่น การสร้างโปรไฟล์การตลาดท่องเที่ยวด้วยการวิเคราะห์การจัดกลุ่ม การวิเคราะห์การจัดกลุ่มของลูกค้าที่มีลักษณะหรือพฤติกรรมการบริโภคที่คล้ายคลึงกัน การจัดกลุ่มเอกสาร ที่มีสาระหลักหรือสาระสำคัญที่คล้ายคลึงกัน เป็นต้น

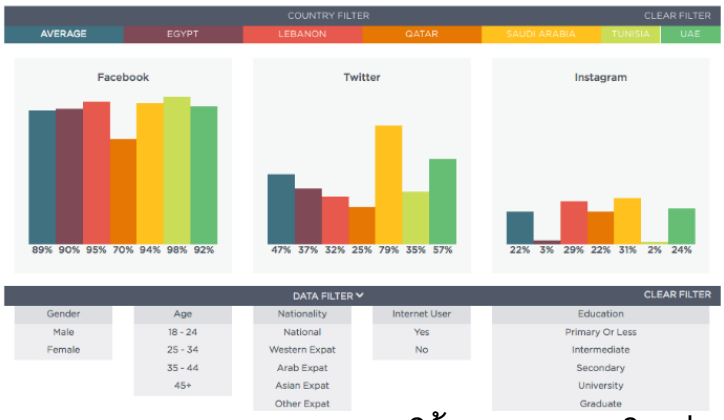
5) การวิเคราะห์ความสัมพันธ์ (Association Analysis) กฎความสัมพันธ์ (Association Rules) การวิเคราะห์กฎความสัมพันธ์เป็นการศึกษาหาลักษณะบางอย่างที่ไปในทิศทางเดียวกันหรือมีความเกี่ยวข้องกัน (Affinity) โดยมีจุดเริ่มต้นจากการวิเคราะห์ข้อมูลการซื้อสินค้า หรือที่รู้จักกันดีในชื่อการวิเคราะห์ตะกร้าซื้อสินค้า (Market basket analysis) ซึ่งคือการวิเคราะห์รายการทั้งหมดที่ลูกค้าซื้อสินค้าต่อครั้ง การวิเคราะห์กฎความสัมพันธ์เป็นการค้นหาความสัมพันธ์เชิงปริมาณระหว่างลักษณะประจำตั้งแต่ 2 ตัว เป็นต้น โดย antecedent หมายถึงสิ่งที่มาก่อน และ consequent หมายถึงผลที่จะเกิดตามมา โดยการที่จะได้กฎความสัมพันธ์จาก

ชุดข้อมูล ซึ่งโดยมากจะเป็นข้อมูลรายการเปลี่ยนแปลง (Transaction Data) โดยใช้เครื่องวัดหรือเกณฑ์การวัดที่เรียกว่า ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence)

6) การพยากรณ์ (Prediction) เป็นการนำข้อมูลมาทำนายคำตอบเช่นเดียวกับการจำแนกที่อธิบายไว้ในบทที่ 2 เพียงแต่ค่าของการพยากรณ์หรือการทำนายจะเป็นค่าแบบต่อเนื่อง (Continuous Value) ซึ่งแตกต่างจากเทคนิคการจำแนกที่คำตอบของการทำนายจะเป็นค่าเต็มหน่วย (Discrete Value) หรือที่เรียกว่า คลาส (Class) ที่เป็นการสื่อถึงค่าคำตอบแบบเต็มหน่วย ขั้นตอนการพัฒนาตัวพยากรณ์จะมีความคล้ายคลึงกับการพัฒนาตัวจำแนก โดยจะมีการแบ่งข้อมูล เป็นข้อมูลฝึกสอนและข้อมูลทดสอบเหมือนกัน แต่สิ่งที่แตกต่างกันคือการวัดประสิทธิภาพ ของการพยากรณ์หรือความแม่นยำในการพยากรณ์ (Predicted Accuracy) ซึ่งจะใช้เกณฑ์การวัดค่าความแม่นยำอีกลักษณะหนึ่งที่ไม่ใช่การวัดร้อยละการจำแนกที่ถูกต้องและเมทริกซ์สับสนเหมือนเทคนิคการจำแนก โดยเกณฑ์การวัดประสิทธิภาพที่นิยมใช้กัน เช่น รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) (Pregibons, 2554)

2.1.3 แนวคิดเกี่ยวกับการแสดงผลข้อมูล (Data Visualization)

Data Visualization คืออาจเป็นข้อมูลหรือข้อมูลที่มาจากการตรวจสอบต่างๆ มาวิเคราะห์อย่างละเอียดแล้วนำเสนอที่คำอธิบายและเข้าใจได้ด้วยตา ตัวอย่างระดับรูปภาพกราฟแสดงระดับการควบคุมของอินโฟกราฟิก แดชบอร์ด ฯลฯ การจดบันทึกการทำ Data Visualization คือแหล่งข้อมูลให้เข้าใจง่ายอ่านข้อมูลสามารถเข้าใจได้ทันทีว่าองค์ประกอบสื่อสารอะไรชี้จุดเน้นเนื้อหาและชี้ให้เห็นข้อมูลเชิงลึกเน้นการบริโภคและการบริโภคของจุดศูนย์กลางของข้อมูลต่างๆ



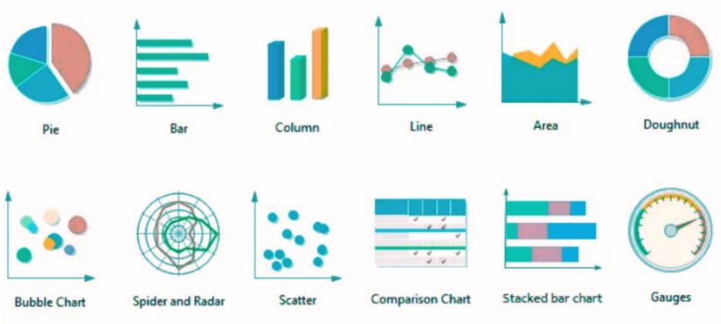
ภาพที่ 2.1 Data Visualization แสดงการใช้งานของบุคคลในแต่ละประเทศ

(ที่มา: blog.hubspot.com)

การแสดงข้อมูล มีรูปแบบที่หลากหลายและไม่จำกัดว่าต้องใช้รูปแบบตามลำดับในการสำรวจข้อมูลเท่านั้น เพราะแต่ละรูปแบบก็มีฟังก์ชั่นเฉพาะของแหล่งข้อมูลบางรูปแบบข้อมูลใช้วิเคราะห์ข้อมูลแต่ละชุดได้ดี บางรูปแบบช่วยให้รูปแบบต่าง ๆ ของตลาดในบางรูปแบบเล่าถึงไกลตัวให้เข้าใจโดยหาข้อมูลโดยหาข้อมูลบางอย่างสิ่งที่จะนำไปใช้ได้ และ 6 การทำ Data Visualization ขึ้นอยู่กับพื้นฐานพื้นฐานเพื่อให้นำเสนอข้อมูลอย่างที่ใช้กันบ่อย ๆ

1. แผนภูมิ (Charts)

การแสดงข้อมูลในตอนแรกคือ แผนภูมิ โดยที่รูปแบบต่าง ๆ ที่พบบ่อยกันมากที่สุดและส่วนมากที่มีความหลากหลายของคุณสมบัติเฉพาะของข้อมูลโมดูล เช่น แผนภูมิวงกลมช่วยให้มองเห็นปริมาณที่เห็นได้ชัดได้ชัดเจน แผนภูมิเปรียบเทียบ คุณสมบัติหลาย ๆ อย่างไม่จำเป็นเพิ่มเติมช่วยให้เห็นการวิจารณ์หรือน้ำหนัก

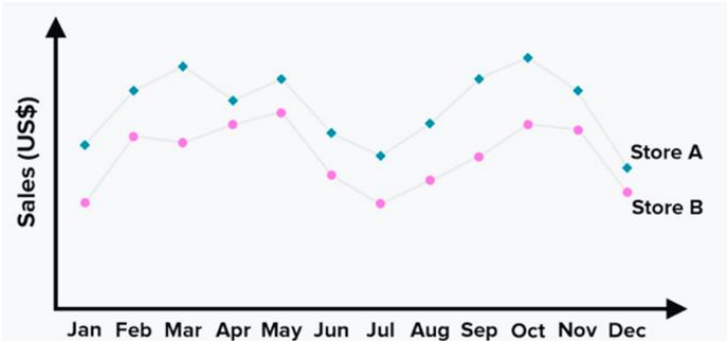


ภาพที่ 2.2 ตัวอย่างแผนภูมิ (Charts)

(ที่มา: medium.com/@Lynia_Li/)

2. กราฟ (Graphs)

กราฟ (Graphs) คือเซตย่อยหรือประเภทต่างๆ ในระดับปานกลาง โดยกราฟิกจะแสดงผลความสัมพันธ์ระหว่างข้อมูล 2 ส่วนต่างๆ (ผ่านแกน X) และแกนเมนบอร์ด (แกน Y) ตรวจสอบสภาพร่างกายประกอบกับระบบการควบคุมความร้อน



ภาพที่ 2.3 ตัวอย่างกราฟ (Graphs)

(ที่มา: mindtools.com)

3. ตาราง (Tables)

ตาราง (Tables) จะเป็นอีกรูปแบบที่ใช้กันมากเพื่อนำเสนอข้อมูลให้ออกมาดูง่ายระดับหนึ่ง 2 ส่วนต่างๆ ภายในองค์กรและแถวซึ่งช่วยจัดการข้อมูลให้ดำเนินการตามปกติและความสัมพันธ์ของข้อมูลหลาย ๆ ชุดตรวจสอบและแถวซึ่งช่วยจัดการข้อมูลให้ดำเนินการตามปกติและความสัมพันธ์ของข้อมูลหลาย ๆ ชุดตรวจสอบและแถว

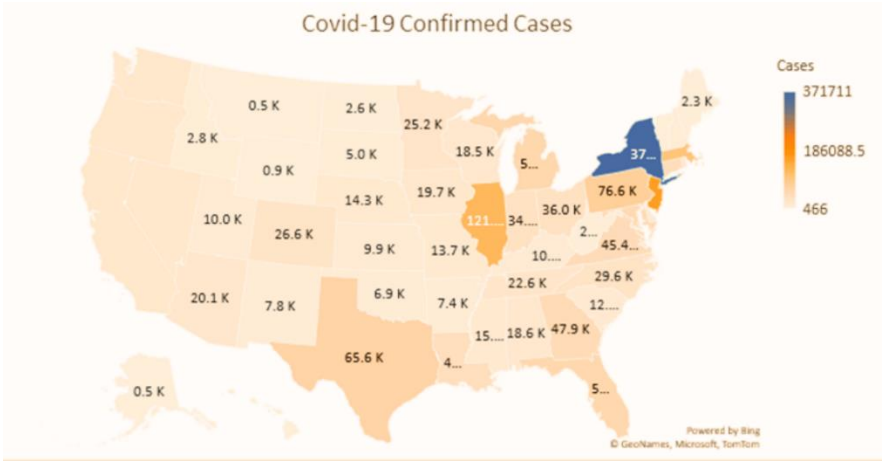
Marks	Number of Students		Total
	Males	Females	
30 – 40	8	6	14
40 – 50	16	10	26
50 – 60	14	16	30
60 – 70	12	8	20
70 – 80	6	4	10
Total	56	44	100

ภาพที่ 2.4 ตัวอย่างตาราง (Tables)

(ที่มา: embibe.com)

4. แผนที่ (Maps)

แผนที่ (Maps)เป็นข้อมูลอ้างอิงไปยังพื้นที่ต่างๆ เก็บข้อมูลข้อมูลยอดของ Covid-19 อย่างต่อเนื่องซึ่งนอกเหนือไปจากการใส่ข้อมูลลงไปยังพื้นที่ต่างๆ แล้วยังคงสามารถใช้สีสันเพื่อบอกช่วงรัฐหรือความเข้มข้นของจุดเริ่มต้นได้

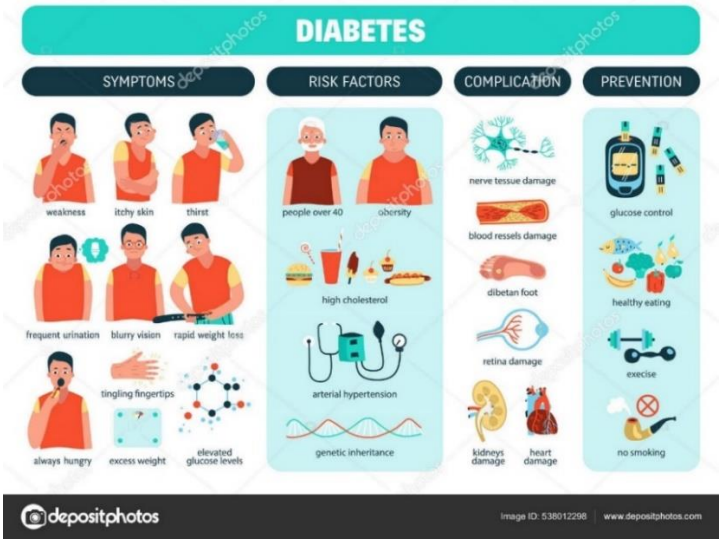


ภาพที่ 2.5 ตัวอย่างแผนที่ (Maps)

(ที่มา: Spreadsheetweb.com)

5. อินโฟกราฟฟิก (Infographics)

คือ ข้อมูลสารสนเทศ ด้วยภาพกราฟิก เพื่อรองรับภาพแทนการเรียนรู้ข้อมูลเข้าใจข้อมูลหรือสามารถเข้าใจผ่านภาพ เพื่อเพิ่มประสิทธิภาพของอินโฟกราฟิก มีเทคนิคการเล่าเรื่อง (Storytelling) หาข้อมูลที่ไม่เคยทราบมาก่อนนำเสนอเนื้อหาความรู้หรือเป็นสื่อข้อมูล



ภาพที่ 2.6 ตัวอย่างอินโฟกราฟฟิก (Infographics)

(ที่มา: https://depositphotos.com/th/vector/diabetes-flat-infographics-538012298.html)

6. แดชบอร์ด (Dashboards)

แดชบอร์ด (Dashboards) คือข้อมูลต่างๆ มาเรียบเรียงและสรุปเป็นภาพปกติในระดับและกราฟิคต่างๆ ที่มานำเสนอที่นี่เป็น Data Visualization ตามที่อธิบายข้อมูลแบบ Real-time ผ่านเครื่องมือหรือเครื่องมือวิเคราะห์และวิเคราะห์ข้อมูลต่างๆ (Stcraft, 2020)

ภาพที่ 2.7 ตัวอย่างแดชบอร์ด (Dashboards)

(ที่มา: octoboard.com)

2.2 ทฤษฎี

2.2.1 หลักการทำเหมืองข้อมูล

1) การทำความสะอาดข้อมูล (Data Cleaning): การตรวจสอบและกำจัดค่าผิดปกติ (outliers) หรือข้อมูลที่ขาดหายไป

2) การรวบรวมข้อมูล (Data Integration): การรวมข้อมูลจากหลายแหล่ง

3) การแปลงข้อมูล (Data Transformation): การปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ได้ง่ายขึ้น

4) การเลือกข้อมูล (Data Selection): การเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับเป้าหมายของการวิเคราะห์

5) การวิเคราะห์แบบจำแนกประเภท (Classification): การแบ่งกลุ่มข้อมูลตามลักษณะเฉพาะ

6) การทำนาย (Prediction): ใช้แบบจำลองเพื่อทำนายค่าหรือแนวโน้มของข้อมูลในอนาคต

7) การสรุปข้อมูล (Summarization): การสรุปข้อมูลที่ซับซ้อนให้เข้าใจง่ายขึ้น

2.2.2 การจำแนกประเภทข้อมูล (Classification)

การจำแนกประเภทของข้อมูล (Classification) เป็นหนึ่งในเทคนิคที่สำคัญของเหมืองข้อมูล (Data Mining) ซึ่งใช้สำหรับจัดกลุ่มข้อมูลออกเป็นประเภทหรือคลาสที่กำหนดไว้ล่วงหน้า โดยอาศัยตัวแปรอิสระ (Features) ที่มีอยู่ในข้อมูลในการพยากรณ์หรือทำนายค่าของตัวแปรตาม (Target Class หรือ Label) หลักการทำงานของการจำแนกประเภทของข้อมูล การจำแนกประเภทของข้อมูลมี 2 กระบวนการหลัก ได้แก่

1. กระบวนการเรียนรู้ (Training Phase)

1.1 ใช้ชุดข้อมูลที่มีป้ายกำกับ (Labeled Data)

1.2 อัลกอริทึมเรียนรู้จากตัวอย่างที่ให้มาเพื่อสร้างโมเดลจำแนกประเภท

2. กระบวนการทำนาย (Prediction Phase)

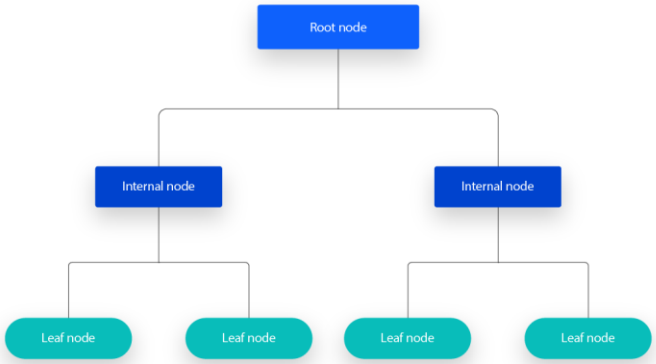
2.1 โมเดลที่ได้จะถูกนำไปใช้ทำนายคลาสของข้อมูลใหม่

2.2 ระบบจะคำนวณและระบุประเภทของข้อมูลที่ยังไม่มีป้ายกำกับมาก่อน

1) เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

Decision Tree Algorithm เป็นโมเดลในรูปแบบของแผนภูมิต้นไม้ (Tree Diagram) ที่มีคอนเซปในการเลียนแบบการตัดสินใจของมนุษย์ มาตัดสินใจในข้อมูลที่ไม่เคย

เห็น 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



ภาพที่ 2.8 โครงสร้างต้นไม้ (Tree Structure)

(ที่มา: https://www.ibm.com/topics/decision-trees)

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

2) เทคนิค Naïve Bayes ใช้ทฤษฎีความน่าจะเป็นในการจำแนกประเภท

Naive Bayes เป็นอีกหนึ่ง Classification Model ที่ใช้สำหรับทำนายว่าข้อมูลอยู่ในกลุ่มไหน ซึ่งเป็นโมเดลที่ต้องอาศัยการกำหนด Label หรือแยกหมวดหมู่ไว้แล้วในการสอน โดยใช้หลักการความน่าจะเป็นทางสถิติ (Probability) ของ Bayes เพื่อทำนายว่าข้อมูลอยู่ในกลุ่มไหน

$$\begin{aligned} \llbracket P(\mathrm{C}\rrbracket \_i \,|A) &= P(\vdash A\dashv \,|\, C\_i\,)P(C\_i\,)/P(A) \\ P(\vdash A\dashv \,|\, C\_j\,) \quad &= -\bigcap\nolimits\_(k=1)\hat{}n\,(\vdash A\_k\dashv \,|Ci) \\ &= P(\vdash A\_1\dashv \,|\, C\_i\,)\times P(\vdash A\_2\dashv \,|\, C\_i\,)\times \ldots \times P(\vdash A\_n\dashv \,|\, C\_i\,) \end{aligned}$$

โดยการคำนวณของ Naive Bayes Algorithm ด้วยสมการตามภาพด้านบน เพื่อคำนวณหาค่า P(CIA) หรือความน่าจะเป็นที่ข้อมูลที่มี Attribute เป็น A จะมีคลาส C ซึ่งจาก

คลาสทั้งหมด ตัวไหนที่มีค่าความน่าจะเป็นนี้มากกว่าก็จะได้ผลลัพธ์ออกมาเป็นค่านั้น ซึ่งตัวโมเดลนี้ก็สามารถนำไปใช้ได้หลากหลายทั้งการทำ Sentiment Analysis Spam Filtering ไปจนถึงเป็นพื้นฐานของ Language Model

3) เทคนิค Support Vector Machine (SVM)

ใช้เส้นแบ่งเขตแดน (Hyperplane) เพื่อจำแนกข้อมูล SVM หรือ Support Vector Machine ถือเป็นวิธีคลาสสิคที่น่าเรียนรู้มากทีเดียว เพราะไอเดียจากโมเดลนี้เป็นหนึ่งในรากฐานสำคัญที่ทำให้เข้าใจโมเดลใหญ่ ๆ ในปัจจุบันมากขึ้น

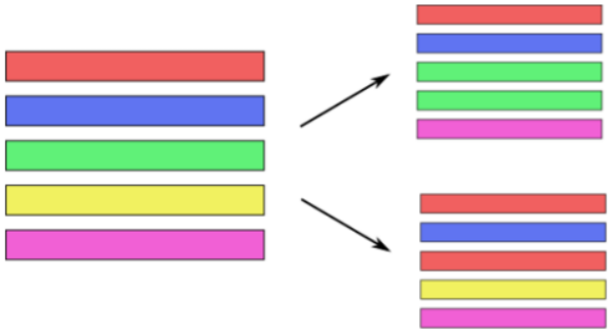
Support Vector Machine (SVM) เป็นอัลกอริธึมสำหรับการจำแนกประเภท (classification) และการถดถอย (regression) ที่ได้รับความนิยมสูง เพราะสามารถจัดการกับข้อมูลที่มีความซับซ้อนและมีมิติสูงได้ดี หลักการของ SVM คือการหาเส้นแบ่ง (Hyperplane) ที่มีระยะห่างมากที่สุดระหว่างกลุ่มข้อมูล เพื่อให้การจำแนกมีความแม่นยำและสามารถทำงานกับข้อมูลใหม่ได้ดี โดยข้อมูลที่อยู่ใกล้เส้นแบ่งมากที่สุดเรียกว่า Support Vectors ซึ่งมีผลต่อการกำหนดตำแหน่งของเส้นแบ่ง

SVM สามารถจัดการกับข้อมูลที่ไม่สามารถแบ่งด้วยเส้นตรงได้ผ่าน Kernel Trick ซึ่งช่วยเปลี่ยนข้อมูลให้อยู่ในมิติที่สูงขึ้น ทำให้สามารถใช้เส้นแบ่งในมิติใหม่ได้อย่างมีประสิทธิภาพ และช่วยให้โมเดลทำงานได้ดีแม้ข้อมูลมีความซับซ้อน นอกจากนี้ SVM ยังมีข้อดีคือสามารถลด Overfitting ได้ เพราะเลือกใช้เฉพาะข้อมูลที่มีความสำคัญต่อการจำแนก และยังทนทานต่อสัญญาณรบกวนหรือ Outlier

4) เทคนิค Random Forest ใช้หลายต้นไม้ตัดสินใจมาช่วยกันจำแนกข้อมูล

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 ตัวอย่างการแบ่งข้อมูลออกเป็น Tree แต่ละต้น คล้ายกับ bagging ดังภาพใต้ล่างนี้

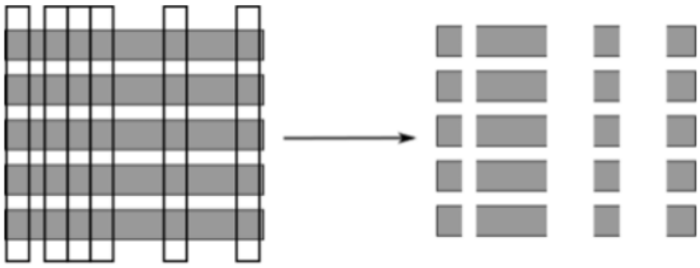
Bagging จะมีการแบ่งข้อมูลออกเป็น Tree หลายๆ ต้นแต่การทำ Bagging จะมีปัญหาเรื่องความไม่เป็นอิสระของข้อมูล เนื่องจากต่อให้แยกออกไปหลายๆ Tree แต่ก็คือข้อมูลเดียวกัน Random Forest จึงเข้ามาแก้ปัญหาตรงนี้ โดยการทำ Random Sample Feature



ภาพที่ 2.9 ตัวอย่างการแบ่งข้อมูลออกเป็น Tree แต่ละต้น

(ที่มา: https://shorturl.asia/94wAy)

Random Sample Feature คือ นอกจากจะแบ่งเป็น Tree หลายๆ ต้นแล้ว ยังแบ่ง Feature ของ Tree แต่ละต้นจะมี Feature ที่ไม่เหมือนกัน เพื่อทำให้แต่ละ Tree มีความหลากหลายและมีความอิสระกันมากขึ้น ตัวอย่างการทำ Random Sample Feature ดังภาพใต้ล่างนี้



ภาพที่ 2.10 ตัวอย่างการทำ Random Sample Feature

(ที่มา: https://shorturl.asia/94wAy)

2.2.3 ทฤษฎีเกี่ยวข้องกับการสร้างเว็บไซต์

2.2.3.1 การสร้างเว็บไซต์สิ่งสำคัญอยู่ที่การออกแบบเว็บ เพราะเว็บไซต์ที่มีรูปแบบสวยงามจะสามารถดึงดูดความสนใจจากผู้คนได้ดีกว่าทำให้ผู้คนเกิดความรู้สึกประทับใจอยากกลับมาใช้งานเว็บไซต์อีกครั้งในอนาคต ดังนั้นเริ่มแรกก่อนทำเว็บไซต์จึงจำเป็นต้องทำความเข้าใจกับหลักการออกแบบ และรูปแบบโครงสร้างของเว็บก่อนการออกแบบเว็บไซต์เพื่อให้มีประสิทธิภาพ และสามารถดึงดูดความสนใจของผู้คนได้ดีจะต้องมีองค์ประกอบของเว็บไซต์อย่างครบถ้วน (วันบีลีฟ, 2560)

1) ความเรียบง่ายเข้าใจง่าย การออกแบบเว็บไซต์ที่ดี จะต้องเน้นที่ความเรียบง่ายเป็นหลัก โดยเลือกนำเสนอเฉพาะสิ่งที่ต้องการนำเสนอจริงๆ ในรูปแบบที่หลากหลาย โดยอาจจะเป็นสีสัน กราฟิก ภาพเคลื่อนไหวหรือตัวอักษร ที่สำคัญจะต้องมีการนำเสนอที่ไม่ดูรก หน้าเว็บจนเกินไป เพื่อไม่ให้เกิดความรู้สึกรกสายตา หรือสร้างความเบื่อหน่าย น่ารำคาญ ให้กับผู้ที่เข้าชมเว็บไซต์ มีตัวอย่างเว็บไซต์ที่มีการออกแบบโดยเน้นความเรียบง่ายได้ดี

2) ความสม่ำเสมอ ไม่สับสนควรออกแบบเว็บไซต์ด้วยความสม่ำเสมอ คือ จะต้องมีรูปแบบ กราฟิก โทนสีและการตกแต่งต่างๆ ให้แต่ละหน้าบนเว็บไซต์มีความคล้ายคลึงกัน และเป็นแนวเดียวกันไปตลอดทั้งเว็บไซต์ ดังตัวอย่างเว็บไซต์ทั่ว ๆ ไปที่จะสังเกตเห็นได้ว่าทุกหน้าของเว็บไซต์นั้น จะเน้นการตกแต่งในรูปแบบเดียวกันทั้งหมด ต่างก็แค่การนำเสนอของ แต่ละหน้าเท่านั้น

3) สร้างความโดดเด่น เป็นเอกลักษณ์การออกแบบเว็บไซต์เพื่อให้สามารถสื่อถึงจุดประสงค์ในการนำเสนอเว็บได้ดี จะต้องมีการสร้างความเป็นเอกลักษณ์และจุดเด่นให้กับเว็บไซต์ เพื่อให้สามารถสะท้อนถึงลักษณะขององค์กรได้มากที่สุด โดยการสร้างเอกลักษณ์ดังกล่าวนั้นอาจใช้ชุดสี รูปภาพ ตัวอักษรหรือกราฟิก นอกจากนี้ก็ต้องขึ้นอยู่กับว่าเป็นเว็บไซต์ แบบบทางการหรือไม่ เพื่อจะได้ออกแบบได้อย่างเหมาะสมที่สุด

4) เนื้อหาต้องดีครบถ้วนเนื้อหาเป็นสิ่งที่สำคัญที่สุดของการสร้างเว็บไซต์ เพราะสิ่งที่ทำให้ผู้คนเกิดความสนใจ และหมั่นติดตามเว็บไซต์เหล่านั้นอยู่เสมอ คือเนื้อหาที่มีความสมบูรณ์และน่าสนใจ นอกจากนี้จะต้องมีการปรับปรุง พัฒนาเนื้อหาบนเว็บให้มีความ ทันสมัยอยู่เสมอ รวมถึงข้อมูลต้องมีความถูกต้องที่สุด

5) ระบบเนวิเกชั่น เป็นเสมือนป้ายบอกทางเพื่อให้ผู้ใช้งานไม่เกิดความสับสน ในขณะใช้งานเว็บไซต์ ซึ่งการออกแบบเนวิเกชั่นก็จะต้องเน้นที่ความเรียบง่ายใช้งานสะดวก และมีความเข้าใจได้ง่าย ที่สำคัญจะต้องมีตำแหน่งการวางที่สม่ำเสมอเพื่อให้ดูเป็นแนวทางเดียวกัน ทำให้ผู้ใช้งานหรือผู้ชมรู้สึกประทับใจ และจดจำเว็บไซต์ได้ง่ายขึ้นส่วนใครที่มีการนำกราฟิกมาใช้ในระบบเนวิเกชั่น จะต้องเลือกกราฟิกที่สามารถสื่อความหมายได้ดีเช่นกัน

6) คุณภาพของเว็บไซต์ที่ดีจะต้องมีคุณภาพ ทั้งสิ่งที่ปรากฏให้เห็นบนเว็บไซต์ กราฟิก ชนิดตัวอักษร รูปภาพหรือสีสันที่ใช้ เนื้อหาที่นำมาแสดงผล ซึ่งหากเว็บไซต์มีคุณภาพก็จะสร้างความน่าเชื่อถือ และเป็นจุดเด่นที่ทำให้ผู้คนส่วนใหญ่เกิดความสนใจได้ดี เพราะฉะนั้นห้ามละเลยในส่วนของคุณภาพเด็ดขาด

7) ความสะดวกในการเข้าใช้งานเว็บไซต์ควรให้ความสะดวกสบายแก่ผู้ใช้งานได้ดี คือจะต้องมีการแสดงผลได้ในทุกระบบปฏิบัติการ ไม่ว่าจะเป็นเว็บเบราว์เซอร์คอมพิวเตอร์ โน้ตบุ๊ค หรือบนโทรศัพท์มือถือ ที่สำคัญจะต้องมีความละเอียดของการแสดงผลและสามารถใช้งานได้โดยไม่มีปัญหาด้วย

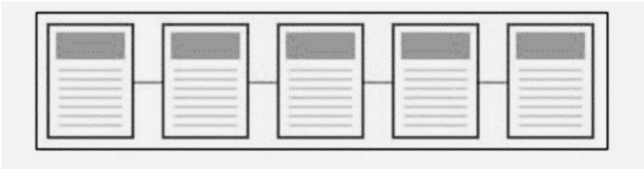
8) ความคงที่ของการออกแบบ การออกแบบเว็บไซต์ควรจะมีความคงที่ในการออกแบบด้วยการสร้างเว็บไซต์ด้วยแบบแผนเดียวกัน และมีการเรียบเรียงเนื้อหาอย่างรอบคอบ ทำให้เว็บมีความน่าเชื่อถือ และดูมีคุณภาพช่วยสร้างความประทับใจให้กับผู้ใช้งานได้เป็นอย่างดี

9) ความคงที่ของการทำงานระบบการทำงานบนเว็บไซต์จะต้องมีความคงที่ และสามารถใช้งานได้ดี ซึ่งนอกจากการออกแบบระบบการทำงานให้มีความทันสมัยและสร้างสรรค์แล้วจะต้องหมั่นตรวจสอบอยู่เสมอ เพราะหากระบบการใช้งานมีความผิดปกติจะได้แก้ปัญหาได้ทัน นอกจากนี้อาจมีการอัพเดตดีไซน์ให้ทันสมัยขึ้นบ่อยๆ เพื่อให้ผู้ใช้งานรู้สึกสนุกไปกับการใช้งานเว็บไซต์3. Density–Based Spatial Clustering of Application with Noise เป็นประเภท Density–Based clustering คือ การแบ่งกลุ่มแบบความหนาแน่น Bagging จะมีการแบ่งข้อมูลออกเป็น Tree หลายๆ ต้นแต่การทำ Bagging จะมีปัญหาเรื่องความไม่เป็นอิสระของข้อมูล เนื่องจากต่อให้แยกออกไปหลายๆ Tree แต่ก็คือข้อมูลเดียวกัน Random Forest จึงเข้ามาแก้ปัญหาตรงนี้ โดยการทำ Random Sample Feature

2.2.3.2 รูปแบบโครงสร้างของเว็บไซต์

การออกแบบโครงสร้างของเว็บไซต์ สามารถทำได้หลากหลายแบบ ซึ่งก็ขึ้นอยู่กับความชอบและความถนัดของแต่ละบุคคลนอกจากนี้ยังขึ้นอยู่กับกลุ่มเป้าหมายที่ต้องการนำเสนอ เพราะจะต้องออกแบบให้เหมาะกับการใช้งานของกลุ่มเป้าหมายมากที่สุด โดยโครงสร้างของเว็บไซต์ส่วนใหญ่ก็จะประกอบไปด้วย 4 รูปแบบดังนี้

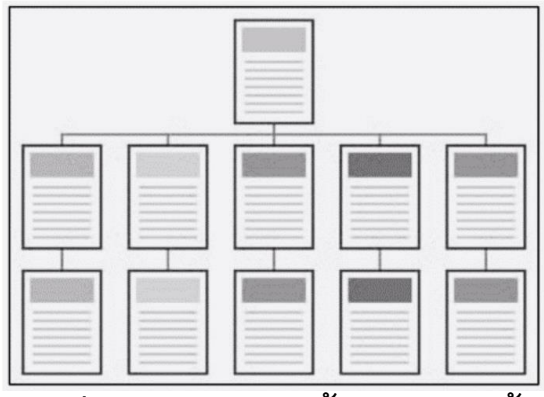
1) โครงสร้างเว็บไซต์แบบเรียงลำดับ จะเป็นโครงสร้างแบบธรรมดาที่นิยมใช้งานกันมากที่สุด เนื่องจากมีความง่ายต่อการจัดระบบข้อมูล และสามารถนำเสนอเรื่องราวตามลำดับได้เป็นอย่างดีเหมาะกับเว็บไซต์ที่มีขนาดเล็กมีเนื้อหาที่ไม่ซับซ้อน ส่วนใหญ่จะเป็นพวกเว็บไซต์ที่ให้ความรู้หรือเว็บไซต์องค์กรขนาดย่อม โดยลักษณะการลิงค์เนื้อหาจะลิงค์ไปทีละหน้ามีทิศทางการเข้าสู่เนื้อหาต่างๆ ในแบบเส้นตรง ใช้ปุ่มเดินหน้า ถอยหลังในการกำหนดทิศทาง จึงทำให้การใช้งานเป็นไปอย่างง่าย แต่โครงสร้างเว็บไซต์แบบเรียงลำดับมีข้อเสีย คือจะทำให้ผู้ใช้งานต้องเสียเวลาในการเข้าสู่เนื้อหาเพราะไม่สามารถกำหนดทิศทางการเข้าสู่เนื้อหาด้วยตัวเองได้



ภาพที่ 2.11 แสดงโครงสร้างเว็บไซต์แบบเรียงลำดับ

(ที่มา: https://www.1belief.com/article/website-design/)

2) โครงสร้างแบบลำดับขั้น นิยมใช้กับเว็บที่มีความซับซ้อนของข้อมูล เพื่อให้สามารถเข้าถึงข้อมูลต่างๆ ได้ง่ายขึ้น โดยจะมีการแบ่งเนื้อหาออกเป็นส่วนๆ และมีการนำเสนอรายละเอียดย่อยๆ ที่ลดหลั่นกันมา ทำให้สามารถทำความข้าใจกับโครงสร้างเนื้อหาได้ง่ายขึ้น โดยจะมีโฮมเพจเป็นจุดเริ่มต้น และจุดร่วมจุดเดียวที่จะนำไปสู่การเชื่อมโยงเนื้อหาเป็นลำดับจากบนลงล่าง



ภาพที่ 2.12 แสดงโครงสร้างแบบลำดับขั้น

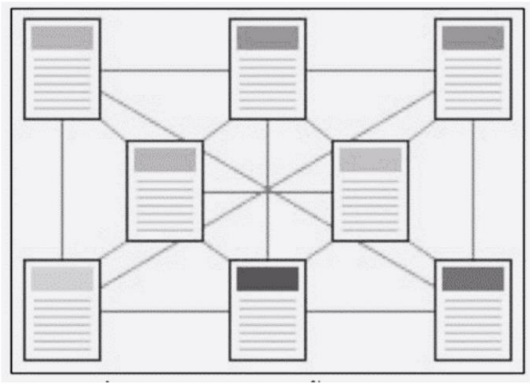
(ที่มา: https://www.1belief.com/article/website-design/)

3) โครงสร้างแบบตาราง เป็นโครงสร้างการออกแบบเว็บไซต์ที่มีความซับซ้อน แต่ก็มีความยืดหยุ่นในระดับหนึ่ง เพื่อให้ผู้ใช้งานสามารถเข้าสู่เนื้อหาต่างๆ ได้ง่ายขึ้น การ ออกแบบในลักษณะนี้จะมีการเชื่อมโยงเนื้อหาในแต่ละส่วนซึ่งกันและกัน ทำให้ผู้ใช้งานสามารถ เปลี่ยนทิศทาง หรือกำหนดทิศทางในการเข้าสู่เนื้อหาด้วยตัวเองได้ จึงไม่ทำให้เสียเวลาแถมยังทำให้เว็บไซต์มีความทันสมัยขึ้น

ภาพที่ 2.13 แสดงโครงสร้างแบบตาราง

(ที่มา: https://www.1belief.com/article/website-design/)

4) โครงสร้างแบบใยแมงมุม เป็นโครงสร้างที่ได้รับความนิยมเป็นอย่างมาก เพราะมีความยืดหยุ่นมากที่สุด โดยทุกหน้าเว็บจะมีการเชื่อมโยงถึงกันหมดทำให้สามารถเข้าถึงหน้าเว็บเพจต่างๆ ที่ต้องการได้อย่างง่าย และมีความอิสระมากขึ้น นอกจากนี้สามารถเชื่อมโยงไปสู่เว็บไซต์ภายนอกได้ดี



ภาพที่ 2.14 แสดงโครงสร้างแบบใยแมงมุม

(ที่มา: https://www.1belief.com/article/website-design/)

2.2.3.3 ส่วนประกอบสำคัญบนหน้าเว็บเพจ บนหน้าเว็บเพจ จะมีส่วนประกอบสำคัญที่จำเป็นต้องมีอยู่ 3 ส่วน

1) ส่วนหัวของหน้า (Header) อยู่ตอนบนสุดของหน้าและเป็นส่วนที่สำคัญที่สุด โดยจะต้องทำให้สามารถดึงดูดผู้ชมให้รู้สึกอยากติดตามเนื้อหาในเว็บไซต์ต่อไป ซึ่งส่วนใหญ่ก็ มักจะมีการใส่ภาพกราฟิกให้ดูสวยงาม สิ่งสำคัญหลักๆ เลย ก็คือ โลโก้ ชื่อเว็บไซต์ และเมนู หลักที่สามารถลิงค์ไปยังเนื้อหาในหน้าเว็บเพจต่างๆ ได้

2) ส่วนของเนื้อหา (body) อยู่บริเวณตอนกลางของหน้าเว็บ โดยจะแสดง ข้อมูลเกี่ยวกับเนื้อหาบนเว็บแบบคร่าวๆ ซึ่งก็จะมีข้อความ กราฟิก ตารางข้อมูลหรือวิดีโอ ประกอบอยู่ และหากมีเมนูแบบเฉพาะกลุ่มก็จะถูกจัดไว้ในหน้านี้เช่นกัน และที่สำคัญเนื้อหาใน ส่วนนี้ควรจะมีความกระชับ เข้าใจง่าย มีการใช้รูปแบบตัวอักษรแบบเรียบง่ายและเป็นระเบียบ

3) ส่วนท้ายของหน้า (Footer) อยู่ล่างสุดของหน้าเว็บ ซึ่งจะมีหรือไม่มีก็ได้ ส่วน นี้จะแสดงถึงข้อมูลต่างๆ เพิ่มเติมเข้าไป เช่น ข้อความที่แสดงถึงการเป็นลิขสิทธิ์ ข้อมูลเจ้าของ เว็บไซต์ วิธีการติดต่อและคำแนะนำต่างๆ เกี่ยวกับการใช้งานเว็บไซต์อย่างถูกต้อง

2.2.3.4 วิธีการเลือกใช้สีสำหรับการออกแบบเว็บไซต์

การเลือกใช้สีในการออกแบบ เว็บไซต์มีความสำคัญเป็นอย่างมาก เพราะสีสามารถกำหนดอารมณ์ ความรู้สึกและกระตุ้น การรับรู้ทางด้านจิตใจของมนุษย์ได้ดี ดังนั้นสีที่ใช้จึงต้องมีความสอดคล้องกับเนื้อหาและ จุดประสงค์ของเว็บ ว่าต้องการให้ผู้เข้าชมรู้สึกอย่างไรต่อเนื้อหาที่ได้อ่าน โดยรูปแบบของสีที่ สายตาของมนุษย์สามารถมองเห็นได้ก็แบ่งออกเป็น 3 กลุ่ม ดังนี้

1) สีโทนร้อน (Warm Colors) เป็นสีแห่งความอบอุ่น ปลอบโยนและกระตุ้น ความสุขได้ดี ซึ่งจะทำให้ผู้เข้าชมรู้สึกมีชีวิตชีวาและมีแรงผลักดันมากขึ้น อีกทั้งยังช่วยดึงดูดให้ ผู้ชมรู้สึกอยากติดตามเนื้อหามากขึ้น

2) สีโทนเย็น (Cool Colors) เป็นสีแห่งความสุภาพและความอ่อนโยน ทำให้ผู้ชมรู้สึกผ่อนคลายและเพลิดเพลินมากขึ้น และยังสามารถใช้โน้มน้าวจากในระยะไกลได้อีกด้วย

3) สีโทนกลาง (Neutral Colors) สีเหล่านี้มักจะถูกนำไปผสมกับสีอื่นๆ เพื่อให้ เกิดสีที่เป็นกลางมากขึ้น และให้ความรู้สึกที่เป็นธรรมชาติ

2.2.4 ปัจจัยเสี่ยงหลักที่มีผลต่อการเกิดโรคเบาหวาน

คือ กลุ่มโรคที่เกิดจากการขาดฮอร์โมนอินซูลิน หรือประสิทธิภาพของอินซูลินลดลง ทำให้ร่างกายมีระดับน้ำตาลในเลือดสูงเป็นเวลานาน ซึ่งมีผลต่อการทำงานของระบบต่าง ๆ โดยเฉพาะ Type 2 DM มีสาเหตุมาจากภาวะดื้ออินซูลิน ร่วมกับความผิดปกติในการหลั่งอินซูลินของตับอ่อนอะไรทำให้เป็นเช่นนั้น สภาวะหนึ่งที่ทำให้เกิดภาวะดื้อต่ออินซูลินเนื่องมาจากการสลายให้กรดไขมันเพิ่มขึ้นในกระแสเลือด กรดไข มีผลทำให้การตอบสนองของอินซูลินลดลง ที่เรียกว่า insulin resistance การกักเก็บน้ำตาล นำน้ำตาลไปใช้จึงลดลง กอปร

กับตับที่ได้รับกรดไขมันและกรดแลคติกที่ไหลล้น จะมีการสร้างน้ำตาลเพิ่มขึ้น ซึ่งทั้งหมดล้วนเป็นตัวการที่ทำให้น้ำตาลในเลือดสูง เมื่อเกิดขึ้นในระยะแรกๆ จะยังไม่พบความผิดปกติ สิ่งที่จะตรวจพบก่อน คือ ตรวจพบน้ำตาลหลังอาหารสูง คนกลุ่มนี้ควรได้รับการ ดูแลให้มีการควบคุมอาหาร แต่ถ้าจะให้เร็วกว่านั้นคือ การป้องกันไม่ให้เกิดภาวะอ้วน แต่หากเกิดน้ำตาลไหลล้นในกระแสเลือดนานๆ beta cell จะเสื่อมและตายทำให้ปริมาณอินซูลินลดลง จนกลายเป็นโรคเบาหวานในที่สุด โรคเบาหวานแบ่งตามสาเหตุของการเกิดโรคได้ 4 ชนิด ได้แก่

1) เบาหวานชนิดที่ 1 (Type 1 DM) เกิดจากการทำลายเบตาเซลล์ของตับอ่อน ทำให้ร่างกายขาดอินซูลิน โดยส่วนใหญ่มีสาเหตุจากภูมิคุ้มกันทำลายตัวเอง (autoimmune) และจำเป็นต้องใช้อินซูลินเพื่อป้องกันภาวะคีโตอะซิโดซิส (DKA)

2) เบาหวานชนิดที่ 2 (Type 2 DM) เกิดจากภาวะดื้ออินซูลินร่วมกับความผิดปกติในการหลั่งอินซูลินของตับอ่อน มักพบในผู้ที่มีภาวะอ้วนหรือมีปัจจัยเสี่ยงจากพฤติกรรมการใช้ชีวิต

3) เบาหวานชนิดอื่นๆ เกิดจากปัจจัยทางพันธุกรรม ความผิดปกติของตับอ่อน ฮอร์โมน ยา หรือสารเคมีที่ส่งผลต่อระดับน้ำตาลในเลือด

4) เบาหวานขณะตั้งครรภ์ ภาวะน้ำตาลในเลือดสูงที่ตรวจพบครั้งแรกในระหว่างตั้งครรภ์ ซึ่งอาจเพิ่มความเสี่ยงต่อภาวะแทรกซ้อนทั้งแม่และทารก

2.2.4.1 สาเหตุหลักของการเกิดโรคเบาหวาน

เกิดจากความผิดปกติของฮอร์โมนอินซูลิน ซึ่งเป็นฮอร์โมนที่ผลิตจากตับอ่อน และทำหน้าที่ควบคุมระดับน้ำตาลในเลือดให้สมดุล เมื่อร่างกายไม่สามารถผลิตอินซูลินได้เพียงพอหรือเซลล์ในร่างกายตอบสนองต่ออินซูลินได้ลดลง จะทำให้ระดับน้ำตาลในเลือดสูงขึ้น จนเกิดภาวะเบาหวานขึ้นมา ซึ่งปัจจัยที่ส่งผลให้เกิดโรคนี้อาจแบ่งออกเป็น 2 ปัจจัย ได้แก่ ปัจจัยทางพันธุกรรมและปัจจัยทางสิ่งแวดล้อม โดยปัจจัยทางพันธุกรรมการมีประวัติครอบครัวที่เป็นเบาหวานซึ่งอาจเพิ่มความเสี่ยงให้เกิดโรคนี้ได้ง่ายขึ้น ส่วนปัจจัยทางสิ่งแวดล้อม เช่น พฤติกรรมการใช้ชีวิตที่ไม่เหมาะสมอย่างการรับประทานอาหารที่มีน้ำตาลหรือไขมันสูงเป็นประจำ การขาดการออกกำลังกาย ภาวะอ้วน หรือความเครียดสะสม ก็เป็นสาเหตุที่ทำให้เกิดภาวะดื้อต่ออินซูลินและนำไปสู่โรคเบาหวานในที่สุด นอกจากนี้อายุที่เพิ่มขึ้นก็เป็นอีกหนึ่งปัจจัยที่ทำให้ร่างกายมีประสิทธิภาพในการควบคุมระดับน้ำตาลลดลง ส่งผลให้ความเสี่ยงในการเกิดโรคเบาหวานสูงขึ้นตามไปด้วย

2.2.4.2 ความรู้เกี่ยวกับโรคเบาหวาน

Metabolic syndrome คือภาวะที่เกิดจากการกินเกิน ส่งผลให้ร่างกายมีสารอาหารสะสมมากเกินไป นำไปสู่กลุ่มอาการหลัก ได้แก่ โรคอ้วนลงพุง เบาหวาน ไขมันในเลือดสูง และความดันโลหิตสูง ซึ่งล้วนเป็นปัจจัยเสี่ยงของโรคหัวใจ ภาวะนี้ทำให้ระดับสารอาหารบางชนิดในร่างกาย เช่น คอเลสเตอรอล ไตรกลีเซอไรด์ และน้ำตาล สูงขึ้นเกินปกติ รวมถึงสารอื่นๆ อย่างกรดอะมิโน กรดแลคติก และกรดไขมันอิสระ (FFA) ที่อาจไม่ถูกตรวจพบ ในภาวะทั่วไปร่างกายมีระบบจัดระเบียบสารอาหารที่ดูดซึมจากลำไส้เล็กและสังเคราะห์ขึ้น โดยจะกักเก็บไว้ในที่เก็บสารอาหาร แต่หากเก็บจนเต็มแล้ว สารอาหารส่วนเกินจะไหลล้นเข้าสู่กระแสเลือดและอวัยวะต่างๆ ซึ่งร่างกายจะพยายามจัดการด้วยการเปลี่ยนเป็นพลังงานหรือสังเคราะห์เป็นสารอื่น เช่น น้ำตาลส่วนเกินถูกแปรเป็นไกลโคเจน ไขมัน หรือถูกขับออกทางปัสสาวะ เมื่อกระบวนการจัดการไม่สามารถรองรับได้อีก จะทำให้ระดับน้ำตาลในเลือดสูงจนเกิดโรคเบาหวาน นอกจากนี้ ไขมันและคอเลสเตอรอลส่วนเกินที่ไหลเวียนในเลือดยังเป็นสาเหตุสำคัญของภาวะหลอดเลือดอุดตัน ซึ่งเพิ่มความเสี่ยงต่อโรคหัวใจขาดเลือดในทางการแพทย์ การวัดภาวะอ้วนใช้ค่า BMI โดยทั่วไป BMI เกิน 25 ถือว่าอ้วน แต่สำหรับชาวเอเชีย WHO กำหนดค่าปกติไม่เกิน 23 นอกจากนี้ การวัดรอบเอวก็เป็นเกณฑ์สำคัญ โดยผู้ชายไม่ควรเกิน 90 เซนติเมตร และผู้หญิงไม่เกิน 80 เซนติเมตร แต่ในบางคนกินเยอะแต่ไม่อ้วน ในขณะที่บางคนกินไม่มากแต่กลับอ้วน ซึ่งแบ่งออกเป็น 2 กลุ่มหลัก ได้แก่ MONW (Metabolically Obese, Normal-Weight) หรือคนที่มีน้ำหนักและรอบเอวปกติแต่มีไขมันและน้ำตาลในเลือดสูง มักเกิดจากความสามารถในการกักเก็บสารอาหารต่ำ เช่น ผู้สูงอายุหรือผู้ที่เคยขาดสารอาหารตอนเด็ก และ MINO (Metabolically Normal but Obese) หรือคนที่อ้วนตามเกณฑ์แต่สุขภาพดี ไม่มีปัญหาระดับน้ำตาลหรือไขมันสูง เนื่องจากมีเซลล์ไขมันมากพอในการเก็บสารอาหาร มักพบในคนที่อ้วนมาตั้งแต่วัยเด็ก

Diabetic Ketoacidosis (DKA) คือภาวะกรดคั่งในเลือดจากสารคีโตน ซึ่งเกิดจากการขาดอินซูลิน ทำให้ร่างกายเผาผลาญน้ำตาลได้น้อยลงและหันไปสลายไขมันและโปรตีนแทน ส่งผลให้น้ำตาลในเลือดสูง น้ำตาลถูกขับออกทางปัสสาวะ ทำให้ปัสสาวะบ่อยและมาก ร่างกายสูญเสียน้ำและเกลือแร่ ผู้ป่วยอาจมีอาการคลื่นไส้ อาเจียน ความดันโลหิตลดลง หายใจหอบลึก และมีกลิ่นปากคล้ายผลไม้ หากไม่ได้รับการรักษา อาจนำไปสู่ภาวะหมดสติหรือโคม่า

ภาวะดื้ออินซูลิน (Insulin Resistance) ทำให้ระดับไตรกลีเซอไรด์สูง ไขมันดี (HDL-C) ต่ำ และความดันโลหิตสูง นอกจากนี้ ยังส่งผลให้ LDL มีขนาดเล็กลงและหนาแน่นขึ้น

สามารถฝังตัวในผนังหลอดเลือด ก่อให้เกิดการอักเสบ และถูก macrophage จับจนกลายเป็น foam cell นำไปสู่การเกิด fatty streak และ endothelial dysfunction ซึ่งเป็นสาเหตุของโรคหลอดเลือดแดงแข็ง (atherosclerosis) และตีบตัน ซึ่งอาจเกิดขึ้นตั้งแต่ระยะก่อนเป็นเบาหวาน

โอกาสที่ทำให้เป็นเบาหวาน จากรายงานการศึกษาพัฒนาดัชนี Diabetes Risk Score ของ รศ. นพ. วิชัยเอกพลากร ซึ่งลักษณะการศึกษา ใช้ข้อมูลการศึกษาทางระบาดวิทยาในกลุ่มพนักงานการไฟฟ้าแห่งประเทศไทย (EGAT study) เริ่มต้นการศึกษา ในปี 2528ในพนักงานที่ไม่มีภาวะเบาหวาน 2,677 คน ใช้เวลาติดตาม 12 ปี สรุปเป็นปัจจัยเสี่ยงและคะแนนความเสี่ยง ดังนี้

ตารางที่ 2.1 ปัจจัยเสี่ยงและคะแนนความเสี่ยง

ปัจจัย	**คะแนน**
1. อายุ	
34 – 39	0
40 – 44	0
45 – 49	1
> 50	2
2. เพศ	
ผู้หญิง	0
ผู้ชาย	2
3. ดัชนีมวลกาย	
< 23	0
23 – 27.5	3
> 27.5	5
4. ความยาวเส้นรอบเอว	
< 90 เซนติเมตร (ผู้ชาย) และ < 80 เซนติเมตร (ผู้หญิง)	0
> 90 เซนติเมตร (ผู้ชาย) และ > 80 เซนติเมตร (ผู้หญิง)	2
5. เป็นโรคความดันเลือดสูง	
ไม่เป็นโรคความดันเลือดสูง	0
เป็นโรคความดันเลือดสูง	2
6. ประวัติเบาหวานในครอบครัว	

ปัจจัย	คะแนน
ไม่มีประวัติ	0
มีประวัติ	4

สำหรับการแปลผลค่าคะแนน เป็นการทำนายความเสี่ยงของการเกิดเบาหวานในอีก 12 ปีข้างหน้า ดังนี้

ตารางที่ 2.2 การทำนายความเสี่ยงของการเกิดเบา หวานในอีก 12 ปีข้างหน้า

ผลรวมคะแนน	ความเสี่ยงต่อเบาหวานใน 12 ปี
<= 2	< 5 %
3 – 5	5 – 10 %
6 – 8	11 – 20 %
9 – 10	21 – 30 %
>= 11	> 30 %

ปัจจัยเสี่ยงที่นอกเหนือจาก 6 ข้อในตาราง ที่ควรพิจารณาควบคู่กันไป ได้แก่ ตับอ่อนได้รับความกระทบกระเทือน เช่น อักเสบเรื้อรังจากการดื่มสุรา หรือบอบช้ำจากการประสบอุบัติเหตุ การติดเชื้อไวรัสบางชนิด ยาบางชนิด เช่น ยาขับปัสสาวะ ยาคุมกำเนิด การตั้งครรภ์ (เนื่องจากฮอร์โมนที่รกสังเคราะห์ขึ้นมีผลยับยั้งการทำงานของอินซูลิน) มีประวัติคลอด บุตรน้ำหนักเกิน 4 กิโลกรัม และไขมันในเลือดผิดปกติ (Triglyceride > 250 มิลลิกรัม/เดซิลิตร HDL–cholesterol < 35 มิลลิกรัม/เดซิลิตร)

IGT และ IFG หมายถึงอะไร และสำคัญต่อสุขภาพอย่างไรเราคงเคยได้ยินคำว่า 'ความทนต่อกลูโคส' กันมาบ้าง ลองมาทำความเข้าใจว่าหมายถึงอะไรและเกี่ยวข้องสำคัญต่อสุขภาพอย่างไร ความทนต่อ กลูโคส หรือ Impaired glucose tolerance : IGT และ Impaired fasting glucose : IFG หรือระดับน้ำตาลในเลือดหลังอดอาหารผิดปกติ ทั้งสองอย่างเป็นความผิดปกติก่อนเกิดเบาหวาน และยังทำให้มีความเสี่ยงต่อโรคระบบหัวใจและหลอดเลือดสูงขึ้นอีกด้วย IGT คือ ภาวะที่มีน้ำตาลในเลือดสูงหลังได้รับกลูโคส ค่าระดับน้ำตาลในเลือดจะอยู่ระหว่างค่าของคนปกติและคนที่เป็นเบาหวาน พบว่ามีอยู่อย่างน้อย 200 ล้านคนทั่วโลก ในกรณี IFG ค่าระดับน้ำตาลในเลือดจะไม่สูงเท่า IGT การตรวจพบ IFG จะช่วยทำให้การคัดกรองคนที่มีความเสี่ยงซึ่งจำเป็นต้องทดสอบความทนกลูโคสกระทำได้ง่ายขึ้น ผู้ที่มี IGT, IFG มีความ

เสี่ยงต่อการเกิดโรคเบาหวานในอนาคต แต่การมี IGT จะทำนายการเกิดโรคเบาหวานได้ดีกว่าคนกลุ่ม นี้ประมาณ 40 % จะมีการดำเนินโรคไปสู่การเป็นเบาหวานภายใน 5–10 ปี บางส่วนอาจกลับเป็นปกติหรือยังคงมี IGT อยู่ต่อไป

การทดสอบความทนต่อน้ำตาลกลูโคส (Oral Glucose Tolerance Test : OGTI) ทำได้โดยเจาะเลือดวัดระดับน้ำตาลในเลือดก่อนรับประทานอาหารเช้า จากนั้นให้ดื่มสารละลายกลูโคสจำนวน 75 กรัมในน้ำ 1 แก้ว รอ 2 ชั่วโมง เจาะเลือดวัดระดับน้ำตาลในเลือดอีกครั้ง แต่สำหรับหญิงตั้งครรภ์ ให้ดื่มสารละลายกลูโคส 100 กรัม แล้วเจาะเลือด 3 ครั้ง คือ ภายหลังการดื่มชั่วโมงที่ 1. 2 และ 3สำหรับการแปลผล ให้ยึดค่าน้ำตาลในเลือดชั่วโมงที่ 2 เป็นหลัก ดังนี้ < 140 มก. /เดซิลิตร ถือว่า ปกติ 140 – 199 มิลลิกรัม/เดซิลิตร มีความบกพร่องต่อการควบคุมน้ำตาล 200 มิลลิกรัม/เดซิลิตร เป็นเบาหวาน

เมื่อใดจะวินิจฉัยว่าเป็น 'เบาหวาน' ทำไมถึงต้องเป็น 126 mg/dl เป็นที่ทราบกันดีว่าเกณฑ์วินิจฉัยเบาหวานเปลี่ยนจากระดับน้ำตาลในเลือดก่อนอาหาร 140 มิลลิกรัม/เดซิลิตร 2 ครั้งหรือน้ำตาลในเลือดไม่ว่าที่เวลาใดๆ มากกว่าหรือเท่ากับ 200 มิลลิกรัม/เดซิลิตร ร่วมกับมีอาการใดอาการหนึ่ง เช่น ปัสสาวะบ่อย คอแห้ง กระหายน้ำ กินจุและน้ำหนักลด ถือว่าเป็นเบาหวานมาเป็น 126 มิลลิกรัม/เดซิลิตรใน ปัจจุบัน อันเนื่องมาจากสมาคมโรคเบาหวานแห่งสหรัฐอเมริกา (ADA : the American Diabetes Association) ค้นพบว่า มีบุคคลบางกลุ่มแม้นระดับ น้ำตาลในเลือดไม่สูงแต่กลับมีภาวะแทรกซ้อนเช่นเดียวกับผู้ที่ได้รับการวินิจฉัย ว่าเป็นเบาหวาน จึงปรับค่าการวินิจฉัยใหม่ให้ลดน้อยลงกว่าเดิม เกณฑ์การวินิจฉัยว่าเป็นเบาหวาน ระดับกลูโคสในพลาสมาขณะอดอาหารอย่างน้อย 8 ชั่วโมง (fasting plasma glucose หรือ FPG บางแห่งใช้ fasting blood sugar) มากกว่าหรือเท่ากับ 126 มิลลิกรัม/เดซิลิตร 2 ครั้ง ระดับกลูโคสในพลาสมาเมื่อเวลาใดก็ได้ (Tandom plasma glucose) มากกว่าหรือเท่ากับ 200 มิลลิกรัม/เดซิลิตร ร่วมกับมีอาการของโรคเบาหวาน ระดับกลูโคสในพลาสมาที่ 2 ชั่วโมงหลังอาหารเมื่อเวลาใดก็ได้ (random plasma glucose) มากกว่าหรือเท่ากับ 200 มิลลิกรัม/เดซิลิตร ร่วมกับมีอาการของโรคเบาหวาน เกณฑ์สำหรับคนปกติทั่วไป ในผู้ใหญ่ ระดับกลูโคสในพลาสมาขณะอดอาหารอย่างน้อย 8 ชั่วโมง (Tasting plasma glucose หรือ FPG) น้อยกว่า 100 มิลลิกรัม/เดซิลิตร (ถ้าอยู่ระหว่าง 100–125 เป็น IGT) ในเด็ก ระดับกลูโคสในพลาสมาขณะอดอาหารอย่างน้อย 8 ชั่วโมง (fasting plasma glucose หรือ FPG) น้อยกว่า 130 มิลลิกรัม/เดซิลิตร ในหญิงมีครรภ์ ระดับกลูโคสในพลาสมาขณะอดอาหารอย่างน้อย 8 ชั่วโมง (fasting plasma glucose หรือ FPG บางแห่งใช้ fasting blood sugar) ไม่เกิน 105 มิลลิกรัม/เดซิลิตร

ตารางที่ 2.3 ความแตกต่างของเบาหวานชนิดที่ 1 และเบาหวานชนิดที่ 2

	เบาหวานชนิดที่ 1	**เบาหวานชนิดที่ 2**
กลุ่มอายุ	พบในเด็ก และคนที่มีอายุ < 40 ปี	มักเกิดในคนที่มีอายุ > 40 ปี
น้ำหนักตัว	ผอม	อ้วน
การทำงานของตับอ่อน	ไม่สามารถผลิตอินซูลินได้ หรือผลิตได้น้อย	ยังสามารถผลิตอินซูลินได้บ้างหรือเป็นปกติแต่ประสิทธิภาพของอินซูลินลดลง
อาการแรกพบ	มักเกิดอาการรุนแรง	อาจมีอาการเล็กน้อย รุนแรง หรือไม่มีอาการเลย
การรักษา	จำเป็นต้องใช้อินซูลิน	อาจควบคุมอาหารอย่างเดียวโดยไม่ต้องใช้อินซูลินหรือใช้ยารับประทาน หรือบางรายอาจต้องใช้อินซูลินฉีดด้วย

2.2.4.3 ผลกระทบของเบาหวานต่อการดำรงชีวิต

นับเนื่องจากเมื่อได้รับการวินิจฉัยว่าเป็นเบาหวาน ผู้ป่วยบางรายยอมรับได้กับการเป็นโรค แต่มีผู้ป่วยอีกหลายรายกลับไม่เป็นเช่นนั้น กล่าวคือ มีความ วิตกกังวล กลัว รู้สึกว่าตนเองป่วย รู้สึกว่าตนเองเป็นโรค บางรายถึงกับหยุดการประกอบอาชีพ หรือทำงานลดลงกว่าเดิมและยังพบว่าส่วนใหญ่ในครั้งแรกที่ได้รับข้อมูลและข้อแนะนำจากบุคลากรทางการแพทย์จะมีความพยายามทำตามเกือบทุกประการ แต่ได้เพียงระยะหนึ่งก็พบว่าเป็นสิ่งที่ขัดกับวิถีชีวิตที่เป็นจริงได้และยังรู้สึกว่าตนเองไม่ได้ดีขึ้นเลย ดังนั้นผู้เป็นเบาหวานจะพยายามแสวงหาวิธีรักษาอื่นๆ นอกเหนือจากการรักษาทางการแพทย์ที่ได้รับจากสถานพยาบาล ทั้งนี้เพื่อให้บรรลุความคาดหวังของตนเอง คือ การหายจากโรค กระบวนการที่ผู้เป็นเบาหวานใช้คือการทดลองทำ เป็นการลองผิดลองถูก จากนั้นเฝ้าดูผลที่เกิด ขึ้นกับตนเอง หากพบว่าไม่ดีขึ้นก็จะเลิกวิธีนั้นแล้วแสวงหาวิธีใหม่ โดยมีช่องทางของการรับข้อมูลเพื่อการดูแลและรักษาตนเอง ช่องทางหลักๆ ส่วนใหญ่ได้จากเพื่อนทั้งที่อยู่ในสภาวะเดียวกัน คือเป็นโรคแบบเดียวกัน และเพื่อนที่ไม่ได้ป่วยแต่ได้รับรู้มาจากการบอกต่อๆ กันมา สื่อโฆษณาโดยเฉพาะทางวิทยุ หนังสือพิมพ์ และบุคลากรสาธารณสุข (สถาบันวิจัยและพัฒนาระบบสุขภาพชุมชน, 2550)

2.3 เครื่องมือในการออกแบบ และวิเคราะห์ข้อมูล

2.3.1 โปรแกรม Rapid Miner Studio

เป็นโปรแกรมที่ออกแบบมาสำหรับการวิเคราะห์ข้อมูล การทำเหมืองข้อมูล (Data mining) เป็นกระบวนการที่กระทำกับข้อมูลขนาดใหญ่เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นโดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์เพื่อให้ได้สารสนเทศที่เราไม่รู้ออกมา โดยสารสนเทศที่ได้จะมีเหตุผล และสามารถนำไปใช้ประโยชน์ได้ โปรแกรม Rapid Miner สามารถช่วยให้เรียกดูข้อมูล และสร้างแบบจำลองเพื่อระบุแนวโน้มได้อย่างง่ายดาย เมื่อสามารถจัดการกับฐานข้อมูลขนาดใหญ่การระบุการเชื่อมต่อระหว่างสองเหตุการณ์อาจเป็นเรื่องยากหรือเป็นไปไม่ได้ เนื่องจากมีหลายธุรกิจที่ต้องพึ่งพาข้อมูลที่มีอยู่เพื่อทำการตัดสินใจที่สำคัญ นักวิเคราะห์ข้อมูลจึงใช้แอพพลิเคชั่นพิเศษเพื่อให้เห็นภาพ และเข้าใจข้อมูล โดยโปรแกรม Rapid Miner Studio มีวัตถุประสงค์เพื่อมอบเครื่องมือที่ใช้งานง่าย และมีประสิทธิภาพสำหรับการวิเคราะห์ข้อมูลจากแหล่งต่างๆ ช่วยให้เราสามารถโหลดข้อมูลที่ต้องการจากไฟล์ข้อความธรรมดา เอกสาร Office หรือเซิร์ฟเวอร์ฐานข้อมูลต่างๆ เช่น Oracle, MySQL หรือ PostgreSQL

2.3.2 กระบวนการวิเคราะห์ข้อมูลด้วย CRISP–DM

CRISP–DM (Cross Industry Standard Process for Data Mining) เป็นกระบวนการมาตรฐานที่ใช้สำหรับการทำเหมืองข้อมูล (Data Mining) และการวิเคราะห์ข้อมูล (Data Analytics) โดยมีโครงสร้างที่ชัดเจน ช่วยให้ธุรกิจนำข้อมูลมาใช้ได้อย่างมีประสิทธิภาพ

การทำ CRISP–DM มีอยู่ 6 ขั้นตอน คือ

1) รู้จักและเข้าใจในธุรกิจ (Business understanding) เป็นขั้นตอนแรกของกระบวนการ ที่มุ่งเน้นไปที่การทำความเข้าใจกระบวนการทางธุรกิจโดยรวม หัวหน้าโครงการหรือที่ปรึกษาด้านการวางระบบวิเคราะห์ข้อมูล จะต้องทำการสัมภาษณ์หรือรับฟังปัญหาหรือความต้องการจากผู้บริหารองค์กรและหน่วยงานต่างๆ ที่จะนำผลการวิเคราะห์ข้อมูลไปใช้ประโยชน์ โดยความต้องการทั้งหมดจะนำมาจัดลำดับความสำคัญ และกำหนดวัตถุประสงค์ที่จะนำไปสู่รูปแบบการวิเคราะห์ข้อมูลขององค์กร เช่น ผู้บริหารห้างสรรพสินค้า ต้องการรู้ว่าอะไรเป็นเหตุปัจจัยที่ทำให้ลูกค้าเป้าหมายตัดสินใจและเลือกที่จะเข้าห้าง ไม่ว่าจะเพื่อการจับจ่ายซื้อของ ใช้เป็นสถานที่นัดพบหรือพักผ่อน หรือหาอาหารรับประทาน ร้านขายสินค้าออนไลน์อยากรู้ว่าผู้คนกำลังให้ความสนใจในสินค้าหรือบริการประเภทใดอยู่ แหล่งข้อมูลออนไลน์ใดที่มีอิทธิพลต่อการตัดสินใจเลือกซื้อสินค้า เป็นต้น

2) สร้างฐานข้อมูลให้ครบ (Data understanding) ขั้นตอนการจัดเก็บและรวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ โดยเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนในการวิเคราะห์ให้สอดรับกับวัตถุประสงค์ที่กำหนดไว้ ในอดีตการศึกษาหาแนวโน้มความต้องการตลาด หรือพฤติกรรมผู้บริโภคในการตัดสินใจซื้อสินค้าเป็นเรื่องที่ยุ่งยากและต้องว่าจ้างบริษัทวิจัยสำรวจภาพรวม ควบคู่กับการพิจารณารายการสั่งซื้อสินค้าที่เก็บไว้ในฐานข้อมูลของบริษัท แต่ด้วยความก้าวหน้าทางด้านเทคโนโลยีในปัจจุบันและการทำธุรกรรมผ่านเครือข่ายอินเทอร์เน็ต ทำให้ข้อมูลมากมายมหาศาลวิ่งผ่านไปมาอยู่ในระบบเว็บไซต์หรือแอพที่เป็นช่องทางในการทำธุรกรรมต่างๆ จึงเป็นแหล่งข้อมูลสำคัญ อีกทั้งยังได้ข้อมูลความสนใจของคนที่พร้อมยอมให้อย่างเต็มใจจากห้องแชทต่างๆ ที่มีการพูดคุยหารือกันปัจจุบันการแกะรอยหรือสะกดรอยตามคนได้ดีที่สุดเกิดขึ้นได้ง่ายมากจากออนไลน์ ไม่ว่าจะเป็นพิกัดตำแหน่งที่อยู่ของเราที่อนุญาตให้แอพต่างๆ เข้าถึง

3) เตรียมข้อมูลให้พร้อมใช้ (Data preparation) ขั้นตอนการแปลงข้อมูลที่ได้รวบรวมมาและเลือกไว้ให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปวิเคราะห์ในขั้นตอนต่อไปได้ โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data cleaning) มักใช้เวลาค่อนข้างมากกระบวนการรับข้อมูลป้อนเข้าสู่ระบบที่ทันสมัยในปัจจุบันจะลดการคีย์ข้อมูลจากคนให้น้อยที่สุด แต่จะใช้วิธีการสแกน การติ๊กเลือก เพื่อลดความผิดพลาดให้น้อยที่สุด เพราะขั้นตอนใช้เวลามากกว่า 50% ของเวลารวมทั้งหมด การลดข้อผิดพลาดของข้อมูลได้มากเท่าใดก็จะยิ่งมีประสิทธิภาพมากขึ้น

4) จัดทำและเลือกโมเดลที่ใช่ (Modeling) ขั้นตอนการสร้างตัวแบบทางคณิตศาสตร์และสถิติเพื่อการวิเคราะห์ข้อมูล โดยสามารถใช้เทคนิควิธีการต่างๆ ผสมผสานกัน เช่น การจำแนก (Classification) การแบ่งกลุ่ม (Clustering) และการสร้างความสัมพันธ์ (Association rule) ในร้านสะดวกซื้อ จะนำข้อมูลการซื้อสินค้าของลูกค้าแต่ละรายมาหาความสัมพันธ์ เช่น คนที่ซื้อเครื่องดื่มแต่ละชนิดมักจะซื้อขนมหรือของกินอะไรร่วมอยู่ด้วย การใช้จ่ายของแต่ละคนจะอยู่ที่ประมาณกี่บาท คนส่วนใหญ่ที่เข้ามาจะซื้อสินค้ากี่ชิ้นต่อคน และเพื่อให้ทราบข้อมูลของผู้ซื้อ ร้านค้ามักจะใช้การออกบัตรเติมเงินที่จูงใจให้ใช้จากส่วนลดหรือสะสมแต้ม ทำให้สามารถติดตามประวัติการใช้จ่ายได้ง่ายขึ้น ซึ่งปัจจุบันมีการนำกล้องจับภาพผู้ซื้อในการแยกแยะเพศ อายุ และไลฟ์สไตล์ของคน

5) ประเมินผลก่อนตัดสินใจ (Evaluation) เป็นขั้นตอนก่อนนำผลลัพธ์ที่ได้จากขั้นตอนที่ 4 ไปใช้งานด้วยการวัดประสิทธิผลของผลลัพธ์ที่ได้กับวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรก ว่ามีนัยสำคัญหรือความน่าเชื่อถือมากน้อยเพียงใด ทั้งนี้อาจต้องกลับไปทบทวนขั้นตอนที่ 2 – 4 ซ้ำอีกครั้ง ในกรณีที่ผลลัพธ์ไม่มีความน่าเชื่อถือเพียงพอ หรือไม่สอดรับกับวัตถุประสงค์

6) เผยแพร่ผลวิเคราะห์ (Deployment) ขั้นตอนการนำผลลัพธ์ที่ได้ไปใช้งานเป็นการทั่วไป อาจจัดทำเป็นรูปแบบของรายงาน (Report) หรือแผนภาพ (Dashboard) ที่พร้อมให้ฝ่ายต่างๆ นำไปใช้ประโยชน์ในการวางแผน กำหนดกลยุทธ์ และดำเนินการต่างๆ ในทางธุรกิจต่อไป (สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2558)

กระบวนการทั้ง 6 ขั้นตอนนี้ไม่ได้ดำเนินการทำอยู่เป็นประจำ แต่จะนำโมเดลที่ได้ไปใช้ในการวิเคราะห์ข้อมูลทางธุรกิจ จนกว่าสภาพแวดล้อมทางธุรกิจเปลี่ยนไปอย่างเห็นได้ชัด หรือโมเดลที่ได้จัดทำไว้เริ่มมีความแม่นยำน้อยลง ซึ่งนั่นหมายความว่าอาจจะมีตัวแปรใหม่ๆ หรือปัจจัยบางอย่างที่มีความสำคัญน้อยลงไป สำหรับผู้ที่สนใจศึกษาเพิ่มเติมได้ในหนังสือเกี่ยวกับ Data Mining และ Data Analytics ส่วนซอฟท์แวร์ที่ช่วยในการจัดการข้อมูล อาทิ R programming Python และ RapidMiner Studio (ขุนพลแก้ว, 2562)

2.3.3 แบบจำลองต้นไม้ตัดสินใจ (Decision Tree)

2.3.3.1 ความหมายเทคนิคต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ช่วยในการตัดสินใจ (Decision Tree) เป็นวิธีหนึ่งที่สำคัญในเทคนิค Classification ซึ่งเป็นเทคนิค การเรียนรู้แบบมีผู้สอน (Supervised learning) นับว่าเป็นวิธีการเรียนรู้ที่นิยมใช้กันมากที่สุดวิธีหนึ่ง ในการเรียนรู้ของเครื่อง การเรียนรู้แบบนี้เป็นการเรียนรู้โดยการแยกแยะหรือแบ่งกลุ่มข้อมูลออกเป็นกลุ่ม (lass) ต่างๆ โดยใช้คุณสมบัติ (Attribute) ของข้อมูลในการแยกแยะหรือแบ่งกลุ่มและคุณสมบัติแต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกันอย่างไรบ้าง ซึ่งเป็นประโยชน์ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น โดยมีโครงสร้างเป็นต้นไม้ แตกแขนงไปตามเงื่อนไขหรือข้อมูลที่ได้คาดคะเนไว้ว่าจะเกิดขึ้น

2.3.3.2 โครงสร้างของต้นไม้ตัดสินใจ

a. โหนดราก (Root Node): เป็นจุดเริ่มต้นของต้นไม้ที่มีข้อมูลทั้งหมด

b. โหนดภายใน (Internal Node): เป็นจุดแตกแขนงที่ใช้เงื่อนไขในการแยกข้อมูลออกเป็นกลุ่มย่อย

c. โหนดใบไม้ (Leaf Node): เป็นจุดสิ้นสุดของต้นไม้ที่แสดงผลลัพธ์ของการจำแนกประเภทหรือการทำนายค่า

2.3.3.3 หลักการทำงานของต้นไม้ตัดสินใจ

a. การเลือกตัวแปรในการแบ่งข้อมูล: ใช้เกณฑ์ต่าง ๆ เช่น

b. Gini Index: ใช้สำหรับวัดความบริสุทธิ์ของข้อมูลในโหนด

c. Entropy และ Information Gain: ใช้ในอัลกอริทึม ID3 และ C4.5

d. Mean Squared Error (MSE): ใช้ในปัญหาการพยากรณ์เชิงตัวเลข

e. การแบ่งข้อมูลซ้ำ ๆ จนกว่าจะได้โหนดใบไม้: ต้นไม้จะแตกแขนงไปเรื่อย ๆ จนกว่าข้อมูลที่อยู่ในแต่ละโหนดมีความบริสุทธิ์เพียงพอ

f. การตัดแต่งต้นไม้ (Pruning): เพื่อลดความซับซ้อนของต้นไม้และป้องกันปัญหา Overfitting

2.3.3.4 ข้อดีและข้อเสียของต้นไม้ตัดสินใจ

ข้อดีของเทคนิคต้นไม้ตัดสินใจ คือ เข้าใจง่าย และสามารถแสดงผลเป็นแผนผังที่มนุษย์อ่านเข้าใจได้ รองรับข้อมูลที่มีทั้งตัวเลขและตัวอักษร ไม่ต้องการการปรับแต่งค่ามาก (Hyperparameter Tuning) เมื่อเทียบกับโมเดลอื่นๆ

ส่วนข้อเสีย คือ อาจมีปัญหา Overfitting หากต้นไม้มีความลึกมากเกินไป มีความอ่อนไหวต่อข้อมูลที่ไม่สมดุล และอาจต้องมีการปรับแต่งหรือใช้เทคนิค Bagging (เช่น Random Forest) เพื่อเพิ่มความแม่นยำ

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) ใช้สมการทางคณิตศาสตร์ในการแยกข้อมูล โดยมีสมการสำคัญที่ใช้คำนวณ ได้แก่

1) สมการการคำนวณ Entropy (ความไม่แน่นอนของข้อมูล) Entropy ใช้ในอัลกอริทึม ID3 และ C4.5 เพื่อวัดระดับความบริสุทธิ์ของกลุ่มข้อมูลในโหนด

Entropy(S)=−∑i=1cpilog f_0 2piEntropy(S) = – \sum_{i=1}^{c} p_i \log_2 p_i

SS คือชุดข้อมูลในโหนด

cc คือจำนวนคลาสที่เป็นไปได้

pip_i คือสัดส่วนของข้อมูลในคลาสที่ ii

ถ้าข้อมูลทั้งหมดอยู่ในคลาสเดียวกัน Entropy=0Entropy = 0 (ข้อมูลบริสุทธิ์ 100%) ถ้าข้อมูลกระจายตัวเท่าๆ กันระหว่างคลาส EntropyEntropy จะมีค่าสูงสุด

2) สมการ Information Gain (IG) ใช้เพื่อเลือกตัวแปรที่ดีที่สุดในการแบ่งข้อมูล

IG(S,A)=Entropy(S)−∑v∈Values(A)∣Sv∣∣S∣Entropy(Sv)IG(S, A) = Entropy(S) – \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)

AA คือคุณลักษณะที่ใช้ในการแบ่งข้อมูล

SvS_v คือเซตย่อยของ SS เมื่อแบ่งตามค่า vv ของ AA

∣Sv∣/∣S∣|S_v| / |S| คือสัดส่วนของข้อมูลที่ถูกแบ่งไปยังแต่ละกลุ่ม

ค่าของ IGIG ที่มากที่สุดบ่งชี้ว่าตัวแปร AA ควรถูกใช้เป็นเกณฑ์ในการแยกข้อมูล

3) สมการ Gini Index (ใช้ใน CART Algorithm) การวัดความบริสุทธิ์ของโหนดคือ Gini Index ซึ่งใช้ในอัลกอริทึม CART

Gini(S)=1−∑i=1cpi2Gini(S) = 1 – \sum_{i=1}^{c} p_i^2

ค่าของ Gini Index อยู่ระหว่าง 0 ถึง 1 ถ้าโหนดมีเฉพาะข้อมูลจากคลาสเดียว Gini=0Gini = 0 (บริสุทธิ์ที่สุด) ถ้าข้อมูลกระจายตัวเท่า ๆ กันระหว่างหลายคลาส Gini มีค่าสูง

4) สมการ Mean Squared Error (MSE) สำหรับ Regression Tree ในกรณีที่ต้นไม้ตัดสินใจใช้สำหรับ การทำนายค่าเชิงตัวเลข (Regression Tree) ใช้ Mean Squared Error (MSE) เป็นเกณฑ์ในการแบ่งข้อมูล

MSE = 1n∑I = 1n(yi−y^)2

MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i – \hat{y})^2

yiy_i คือค่าจริงของตัวแปรเป้าหมาย

y^\hat{y} คือค่าทำนายของโหนดนั้น

nn คือจำนวนข้อมูลในโหนดนั้น

5) สมการ Reduction in Variance (อีกวิธีหนึ่งสำหรับ Regression Tree) ใช้วัดคุณภาพของการแบ่งข้อมูลสำหรับต้นไม้ที่ใช้พยากรณ์ตัวเลขคือ Reduction in Variance

Reduction=Variance(S)−∑v∈Values(A)|Sv||S|Variance(Sv)Reduction= Variance(S) – \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Variance(S_v)

ค่าของ Reduction ที่สูงขึ้นหมายความว่าการแบ่งข้อมูลทำให้ความแปรปรวนลดลง ซึ่งถือเป็นสิ่งที่ดี

2.3.4 เทคนิคนาอีฟเบย์ (Naïve Bayes)

2.3.4.1 ความหมายเทคนิคนาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์ (Naïve Bayes) เป็นอัลกอริทึมการจำแนกประเภทในสาขาการเรียนรู้ของเครื่อง (Machine Learning) ที่อิงตามทฤษฎีบทของเบย์ (Bayes' Theorem) โดยมีสมมติฐานว่าแต่ละคุณลักษณะ (attribute) ของข้อมูลเป็นอิสระต่อกัน แม้สมมติฐานนี้อาจไม่สอดคล้องกับความเป็นจริงเสมอไป แต่ในหลายกรณีนาอีฟเบย์ยังคงให้ผลลัพธ์ที่มีประสิทธิภาพ อัลกอริทึมนี้เหมาะสำหรับการจำแนกข้อมูลที่มีขนาดใหญ่และมีคุณลักษณะจำนวนมาก เนื่องจากมีความเรียบง่ายและประสิทธิภาพในการคำนวณ นาอีฟเบย์ถูกนำไปใช้ในงานต่าง ๆ เช่น การกรองอีเมลสแปม การวิเคราะห์ความรู้สึก (sentiment analysis) และการจำแนกข้อความ

2.3.4.2 หลักการทำงานของ Naïve Bayes

มีการคำนวณความน่าจะเป็นของแต่ละคลาสจากข้อมูลฝึกสอน คำนวณความน่าจะเป็นแบบมีเงื่อนไขของแต่ละคุณลักษณะภายใต้แต่ละคลาส มีการใช้กฎของเบย์ เพื่อคำนวณค่าความน่าจะเป็นรวม และเลือกคลาสที่มีค่าความน่าจะเป็นสูงสุด เป็นคำตอบ และการทำงานของ Naïve Bayes มักใช้งานบ่อยในการจำแนกอีเมลสแปม การวิเคราะห์ข้อความและการทำนายหมวดหมู่ข้อมูล

2.3.4.3 สมการที่เกี่ยวข้องกับ Naïve Bayes

1. ทฤษฎีบทของเบย์ (Bayes' Theorem)

P(C|X)=P(X|C)P(C)P(X)P(C \mid X) = \frac{P(X \mid C) P(C)}{P(X)}P(C|X)=P(X)P(X|C)P(C)

P(C|X)P(C \mid X)P(C|X) = ความน่าจะเป็นที่ข้อมูล XXXอยู่ในคลาส CCC (Posterior Probability)

P(X|C)P(X \mid C)P(X|C) = ความน่าจะเป็นของข้อมูล XXX ภายใต้คลาส CCC (Likelihood)

P(C)P(C)P(C) = ความน่าจะเป็นของคลาส CCC (Prior Probability)

P(X)P(X)P(X) = ความน่าจะเป็นของข้อมูล XXX (Evidence)

เมื่อใช้สำหรับการจำแนกประเภท เราสามารถตัด P(X)P(X)P(X) ออกไปได้ เพราะมันเป็นค่าคงที่สำหรับทุกคลาส

2. สมการของ Naïve Bayes (ภายใต้สมมติฐานความเป็นอิสระ)

ถ้า X=(X1,X2,...,Xn)X = (X_1, X_2, ..., X_n)X=(X1,X2,...,Xn) เป็นชุดของ nnn คุณลักษณะ (features) และสมมติว่าแต่ละ XiX_iXi เป็นอิสระต่อกันเมื่อพิจารณาจากคลาส CCC

P(C|X1,X2,...,Xn)=P(C)∏i=1nP(Xi|C)P(X1,X2,...,Xn)P(C \mid X_1, X_2, ..., X_n) = \frac{P(C) \prod_{i=1}^{n} P(X_i \mid C)}{P(X_1, X_2, ..., X_n)}P(C|X1,X2,...,Xn)=P(X1,X2,...,Xn)P(C)∏i=1nP(Xi|C)

เนื่องจาก P(X1,X2,...,Xn)P(X_1, X_2, ..., X_n)P(X1,X2,...,Xn) เป็นค่าคงที่สำหรับทุกคลาส CCC ในการพิจารณาการจำแนกประเภท เราสามารถใช้ Maximum A Posteriori (MAP) Estimation

C*=arg⁡max⁡CP(C)∏I=1nP(Xi|C)C^*=\arg\max_{C}P(C)\prod_{i=1}^{n} P(X_i\midC)C*=argCmaxP(C)i=1∏nP(Xi|C)

3. การประมาณค่า P(Xi|C)P(X_i \mid C)P(Xi|C) สำหรับประเภทของข้อมูลต่างๆ Naïve Bayes มีหลายเวอร์ชัน ขึ้นอยู่กับประเภทของข้อมูลที่ใช้

3.1 Gaussian Naïve Bayes (สำหรับข้อมูลตัวเลขแบบต่อเนื่อง)

สมมติว่าแต่ละคุณลักษณะ XiX_iXi มีการแจกแจงแบบปกติ (Normal Distribution)

P(Xi∣C)=12πσC2exp⁡(−(Xi−μC)22σC2)P(X_i \mid C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(X_i - \mu_C)^2}{2\sigma_C^2}\right)P(Xi ∣C)=2πσC21exp(−2σC2(Xi−μC)2)

โดยที่ μC\mu_CμC และ σC2\sigma_C^2σC2 เป็นค่าเฉลี่ย (mean) และความแปรปรวน (variance) ของ XiX_iXi สำหรับคลาส CCC

3.2 Multinomial Naïve Bayes (สำหรับข้อมูลข้อความ เช่น การจำแนกอีเมลสแปม) ใช้สำหรับข้อมูลประเภทที่เป็นจำนวนครั้งของการปรากฏของคำในเอกสาร เช่น Bag of Words (BoW) ใน NLP

P(Xi∣C)=count(Xi,C)+α∑jcount(Xj,C)+αNP(X_i \mid C) = \frac{count(X_i, C) + \alpha}{\sum_{j} count(X_j, C) + \alpha N}P(Xi∣C)=∑jcount(Xj ,C)+αNcount(Xi,C)+α

โดยที่ count (Xi,C)count(X_i, C)count(Xi,C) = จำนวนครั้งที่คำ XiX_iXi ปรากฏในคลาส CCC

NNN = จำนวนคำศัพท์ทั้งหมดในคลาส

α\alphaα = ตัวแปร Laplace Smoothing (ป้องกันค่า 000, โดยปกติ α=1\alpha = 1α=1)

3.3 Bernoulli Naïve Bayes (สำหรับข้อมูลไบนารี เช่น การตรวจสอบว่าคำปรากฏในเอกสารหรือไม่) ใช้ในกรณีที่คุณลักษณะมีค่า 0 หรือ 1 เช่น การมีอยู่ของคำศัพท์ในเอกสาร

P(Xi∣C)=P(Xi=1∣C)iX(1−P(Xi=1∣C))(1−Xi)P(X_i \mid C) = P(X_i = 1 \mid C)^X_i (1 - P(X_i = 1 \mid C))^{(1 - X_i)}P(Xi∣C)=P(Xi=1∣C)iX(1−P(Xi=1∣C))(1−Xi)

โดยที่ P(Xi=1∣C)P(X_i = 1 \mid C)P(Xi=1∣C) = ความน่าจะเป็นที่คำ XiX_iXi ปรากฏในคลาส CCC

4. การคำนวณ Prior Probability P(C)P(C)P(C) Prior Probability สามารถประมาณค่าได้จากข้อมูลที่มีอยู่

P(C)=จำนวนตัวอย่างในคลาส Cจำนวนตัวอย่างทั้งหมดP(C) = \frac{\text{จำนวนตัวอย่างในคลาส } C}{\text{จำนวนตัวอย่างทั้งหมด}}P(C)=จำนวนตัวอย่างทั้งหมดจำนวนตัวอย่างในคลาส C

5. การตัดสินใจเลือกคลาส (Prediction)

เมื่อนำทุกค่ามาคำนวณ เราเลือกคลาสที่มีค่าความน่าจะเป็นสูงสุด

C*=arg⁡max⁡CP(C)∏i=1nP(Xi|C)C^*=\arg\max_{C} P(C) \prod_{i=1}^{n} P(X_i \mid C)C*=argCmaxP(C)i=1∏nP(Xi|C)

2.3.4.4 ข้อดีและข้อเสียของ Naïve Bayes

ข้อดีของ Naïve Bayes คือ ทำงานรวดเร็วและมีประสิทธิภาพ เหมาะสำหรับงานจำแนกข้อความ รองรับ Multi-class Classification ทนต่อ Overfitting และจัดการกับข้อมูลที่ขาดหายไปได้

ส่วนข้อเสียคือ ไม่สามารถจัดการกับข้อมูลที่มีค่าเป็นศูนย์ได้ดี (Zero Probability Problem) อาจไม่ทำงานได้ดีเมื่อต้องการพิจารณาความสัมพันธ์ระหว่าง Feature มีปัญหากับข้อมูลที่เป็นตัวเลขต่อเนื่อง ไม่สามารถเรียนรู้การโต้ตอบของ Feature ได้

2.3.5 เทคนิคป่าสุ่ม (Random Forest)

2.3.5.1 ความหมายเทคนิคป่าสุ่ม (Random Forest)

Random Forest เป็นอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้หลักการของ Ensemble Learning โดยการรวมแบบจำลอง Decision Tree หลายต้นเข้าด้วยกันเพื่อเพิ่มความแม่นยำและลดปัญหา Overfitting ซึ่งพบในต้นไม้ตัดสินใจ (Decision Tree) แบบเดี่ยวๆ

2.3.5.2 หลักการทำงานของ Random Forest

1) Bootstrap Aggregating (Bagging)

a. ใช้เทคนิค Bootstrap Sampling โดยการสุ่มตัวอย่างข้อมูลจากชุดข้อมูลฝึกอบรมแบบมีการคืนตัวอย่าง (Sampling with Replacement) เพื่อสร้างชุดข้อมูลย่อย (Subsets)

b. ฝึกต้นไม้ตัดสินใจ (Decision Tree) หลายต้นโดยใช้ชุดข้อมูลเหล่านี้

c. ต้นไม้แต่ละต้นจะได้รับชุดข้อมูลฝึกอบรมที่แตกต่างกัน

2) eature Randomness (Random Subspace Method)

a. เมื่อสร้างแต่ละต้นไม้ จะเลือก ชุดของตัวแปร (Features) แบบสุ่ม เพื่อใช้ในการแยกข้อมูลแทนที่จะใช้ทุกตัวแปร

b. เทคนิคนี้ช่วยลดความสัมพันธ์ระหว่างต้นไม้แต่ละต้น ทำให้แบบจำลองมีความเป็นอิสระมากขึ้น

3) การรวมผลลัพธ์ (Aggregation of Results)

a. กรณีการจำแนกประเภท (Classification): ใช้วิธี Voting โดยให้แต่ละต้นไม้ทำนายคลาส และใช้ เสียงข้างมาก (Majority Vote) เป็นผลลัพธ์สุดท้าย

b. กรณีการพยากรณ์ค่าเชิงตัวเลข (Regression): ใช้ค่าเฉลี่ยของค่าทำนายจากทุกต้นไม้ข้อดีของเทคนิคต้นไม้ตัดสินใจ คือ เข้าใจง่าย และสามารถแสดงผลเป็นแผนผังที่มนุษย์อ่านเข้าใจได้ รองรับข้อมูลที่มีทั้งตัวเลขและตัวอักษร ไม่ต้องการการปรับแต่งค่ามาก (Hyperparameter Tuning) เมื่อเทียบกับโมเดลอื่นๆ

2.3.5.3 สมการที่เกี่ยวข้องกับ Random Forest

1) การเลือกตัวอย่างข้อมูลด้วย Bootstrap Sampling เลือกตัวอย่างข้อมูล mm ชุดจากข้อมูลต้นฉบับ nn ตัวอย่าง โดยมีการคืนตัวอย่าง

Di={X1,X2,...,Xm},*โดยที่* Xi∼D,i=1,2,...,mD_i = \{ X_1, X_2, ..., X_m \}, \quad \text{*โดยที่* } X_i \sim D, \quad i = 1, 2, ..., m

DD คือชุดข้อมูลต้นฉบับขนาด nn

DiD_i คือชุดข้อมูลฝึกที่สุ่มมาจำนวน mm ตัวอย่าง (โดยปกติ m≈nm \approx n)

2) การคำนวณ Gini Index สำหรับการแบ่งข้อมูล Gini Index ใช้ในอัลกอริทึม CART (Classification and Regression Trees) ซึ่งเป็นรากฐานของ Random Forest ค่าที่ต่ำหมายความว่าโหนดนั้นมีข้อมูลที่บริสุทธิ์มากขึ้น

Gini(S)=1−∑i=1cpi2Gini(S)=1\sum_{i=1}^{c} p_i^2

SS คือชุดข้อมูลปัจจุบันในโหนดที่กำลังแบ่ง

pip_i คือสัดส่วนของข้อมูลที่อยู่ในคลาสที่ ii

3) การคำนวณ Entropy และ Information Gain (ใช้ใน ID3, C4.5) Entropy ใช้คำนวณระดับความไม่แน่นอนของข้อมูล

Entropy(S)=−∑i=1cpilog⁡2piEntropy(S) = – \sum_{i=1}^{c} p_i \log_2 p_i

จากนั้นคำนวณ Information Gain เพื่อตัดสินใจเลือกตัวแปรที่ดีที่สุด ค่าของ IG ที่มากกว่าจะถูกใช้เป็นเกณฑ์ในการแบ่งข้อมูล

IG(S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy(Sv)IG(S, A) = Entropy(S) – \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)

4) การรวมผลลัพธ์ของ Random Forest

1. สำหรับปัญหาการจำแนกประเภท (Classification)

$$y^ = \arg\max_{k} \sum_{i = 1}^{T} I(h_i(x) = k)\hat{y} = \arg\max_{k} \sum_{i = 1}^{T} I(h_i(x) = k)$$

hi(x)h_i(x) คือผลลัพธ์ของต้นไม้ที่ ii

I(hi(x)=k)I(h_i(x) = k) เป็นฟังก์ชันตัวบ่งชี้ (Indicator Function) ที่มีค่าเป็น 1 ถ้าต้นไม้ที่ ii ทำนายว่าเป็นคลาส kk

TT คือต้นไม้ทั้งหมดในป่า

2. สำหรับปัญหาการพยากรณ์ค่าเชิงตัวเลข (Regression) คำนวณค่าเฉลี่ยของผลลัพธ์จากทุกต้นไม้ตัดสินใจ

$$y^ = \frac{1}{T} \sum_{i = 1}^{T} h_i(x)\hat{y} = \frac{1}{T} \sum_{i = 1}^{T} h_i(x)$$

2.3.5.4 ข้อดีและข้อเสียของ Random Forest

ข้อดีของ Random Fprest คือ ลดปัญหา Overfitting ของ Decision Tree รองรับข้อมูลที่มีมิติสูง และใช้ได้กับทั้ง Classification และ Regression มีความทนทานต่อข้อมูลที่ขาดหายและค่าผิดปกติ (Outliers) สามารถขนานกันทำงาน (Parallel Processing) ได้ดี

ส่วนข้อเสีย คือ ใช้เวลาฝึกอบรมมากกว่าต้นไม้ตัดสินใจต้นเดียว อาจใช้ทรัพยากรมากเนื่องจากต้องสร้างต้นไม้หลายต้น ตีความได้ยากกว่า Decision Tree (แม้ว่า Feature Importance จะช่วยได้)

2.3.6 แบบซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

2.3.6.1 ความหมายของ Support Vector Machine

เป็นอัลกอริธึมทาง Machine Learning ที่ใช้สำหรับ การจำแนกประเภท (Classification) และ การถดถอย (Regression) แต่โดยทั่วไปนิยมใช้สำหรับงาน Classification มากกว่า

2.3.6.2 หลักการทำงานของ Support Vector Machine

SVM ทำงานโดยค้นหา เส้นแบ่งเขต (Hyperplane) ที่สามารถแยกกลุ่มข้อมูลออกจากกันได้ดีที่สุด ซึ่งเส้นแบ่งเขตที่ดีต้องมีระยะห่าง (Margin) จากกลุ่มข้อมูลทั้งสองข้างมากที่สุด

a. Support Vectors คือ จุดข้อมูลที่อยู่ใกล้เส้นแบ่งเขตมากที่สุด และมีบทบาทสำคัญในการกำหนดตำแหน่งของเส้นแบ่ง

b. Margin คือ ระยะห่างระหว่างเส้นแบ่งเขตกับ Support Vectors ทั้งสองฝั่ง

c. Kernel Trick ใช้สำหรับแปลงข้อมูลที่ไม่สามารถแยกได้ในมิติเดิมให้ไปอยู่ในมิติที่สูงขึ้น ซึ่งสามารถแยกข้อมูลออกจากกันได้ง่ายขึ้น

2.3.6.3 สมการของ Support Vector Machine

1) ฟังก์ชันเส้นแบ่งเขต (Hyperplane Equation) โดยทั่วไป Hyperplane ในมิติที่ nnn สามารถเขียนได้ในรูปสมการเส้นตรง

$$wTx + b = 0w^T x + b = 0wTx + b = 0$$

w = เวกเตอร์น้ำหนัก (Weight Vector)

x = เวกเตอร์ข้อมูล

b = ค่าไบแอส (Bias)

เส้นแบ่งเขตนี้ใช้จำแนกข้อมูลเป็นสองคลาส

หาก $wTx+b>0w^T x + b > 0wTx+b>0$ ⟶ ข้อมูลอยู่ฝั่งหนึ่งของเส้น

หาก $wTx+b<0w^T x + b < 0wTx+b<0$ ⟶ ข้อมูลอยู่อีกฝั่งของเส้น

2) เงื่อนไขของ Support Vector สำหรับข้อมูลที่เป็น Support Vectors ต้องอยู่บนเส้นขอบของ Margin ดังนั้นจึงมีเงื่อนไข

$$yi(wTxi + b) = 1y\_i\ (w\text{^}T\ x\_i\ +\ b)\ =\ 1yi(wTxi + b) = 1$$

โดยที่ yiy_iyi เป็นค่าป้ายกำกับของข้อมูล (+1+1+1 หรือ −1−1−1)

3. การเพิ่ม Margin (Optimization Problem)

เป้าหมายของ SVM คือการหาค่า www และ bbb ที่ทำให้ Margin กว้างที่สุด ซึ่งหมายถึงการ ลดขนาดของ www ให้เล็กที่สุด โดยเงื่อนไขที่กำหนดให้ข้อมูลถูกจำแนกได้อย่างถูกต้อง

$$\min\ 12\ ||\ w\ ||\ 2\backslash min\ \backslash frac\{1\}\{2\}\ ||w||\text{^}2min21\ ||\ w\ ||\ 2$$

ภายใต้เงื่อนไขว่า

$$yi(wTxi + b) \geq 1, \forall iy\_i\ (w\text{^}T\ x\_i\ +\ b)\ \backslash geq\ 1,\backslash quad\ \backslash forall\ iyi(wTxi + b) \geq 1, \forall i$$

4. การใช้ Lagrange Multiplier แก้ปัญหา Optimization สามารถใช้ Lagrange Multiplier เพื่อเปลี่ยนปัญหาให้อยู่ในรูปของฟังก์ชันลากรองจ์ (Lagrangian)

$$L(w, b, \alpha) = 12\ ||\ w\ ||\ 2 - \sum i = 1N\alpha i[yi(wTxi + b) - 1]L(w, b,\backslash alpha)$$
$$= \backslash frac\{1\}\{2\}\ ||w||\text{^}2\ -\backslash sum\_\{i$$
$$= 1\}\text{^}\{N\}\ \backslash alpha\_i\ [y\_i\ (w\text{^}T\ x\_i\ +\ b)\ -\ 1]L(w, b, \alpha) = 21\ ||\ w\ |$$
$$|\ 2 - i = 1\sum N\alpha i[yi(wTxi + b) - 1]$$

โดยที่ αi\alpha_iαi เป็นพารามิเตอร์ของ Lagrange Multiplier ซึ่งสามารถใช้ KKT Conditions ในการแก้หา www และ bbb

5. Kernel Trick สำหรับข้อมูลที่ไม่สามารถแยกได้เชิงเส้น

หากข้อมูลไม่สามารถแยกได้ด้วยเส้นตรง สามารถใช้ Kernel Trick เพื่อเปลี่ยนข้อมูลไปอยู่ในมิติที่สูงขึ้น โดยใช้ Kernel Function K(xi,xj)K(x_i, x_j)K(xi,xj) แทนการคำนวณจุดข้อมูลโดยตรง

$$\mathrm{K(xi, xj) = \phi(xi)T\phi(xj)K(x\_i, x\_j) = \phi(x\_i)\hat{}T\ \phi(x\_j)K(xi, xj) = \phi(xi)T\phi(xj)}$$

2.3.6.4 ข้อดีและข้อเสียของ Support Vector Machine

ข้อดีของ Support Vector Machine คือ มีความแม่นยำสูง ทำงานได้ดีในปัญหาที่มีขนาดข้อมูลเล็ก ทนต่อ Overfitting ใช้งานได้ทั้ง Classification และ Regression และสามารถจัดการกับข้อมูลที่ไม่ได้แยกกันอย่างเป็นระเบียบได้ดี

ส่วนข้อเสียคือ ใช้เวลาประมวลผลนาน เลือก Kernel ที่เหมาะสมยาก ตีความผลลัพธ์ยาก ไม่สามารถอธิบายการตัดสินใจได้ง่ายเหมือน Decision Tree หรือ Logistic Regression และไม่เหมาะกับข้อมูลที่มี Noise มาก โดยเฉพาะถ้าข้อมูลมี Overlapping ระหว่างคลาส

2.3.7 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

2.3.7.1 ความหมายของ Artificial Neural Network

Artificial Neural Network (ANN) หรือโครงข่ายประสาทเทียม เป็นหนึ่งในเทคนิคของ Machine Learning ที่ได้รับแรงบันดาลใจจากการทำงานของสมองมนุษย์ ANN เป็นระบบที่ประกอบด้วยโครงสร้างของ "เซลล์ประสาทเทียม (Artificial Neurons)" ซึ่งเชื่อมโยงกันและทำหน้าที่ประมวลผลข้อมูล โดยอาศัยหลักการเรียนรู้เพื่อสร้างแบบจำลองที่สามารถใช้ทำนายหรือจำแนกข้อมูลได้

Artificial Neural Network เป็นโมเดลทางคณิตศาสตร์และคอมพิวเตอร์ที่จำลองการทำงานของสมองมนุษย์ในการเรียนรู้ข้อมูล ประมวลผล และตอบสนองต่อปัญหาต่างๆ ANN ใช้หลักการของการเชื่อมโยงของเซลล์ประสาทที่มีโครงสร้างเป็นชั้น (Layers) ซึ่งแต่ละชั้นจะประกอบด้วยโหนด (Nodes) หรือเซลล์ประสาท (Neurons) ที่ทำหน้าที่ประมวลผลข้อมูล

2.3.7.2 หลักการทำงานของ Artificial Neural Network

Artificial Neural Network (ANN) ทำงานโดยการเลียนแบบโครงสร้างและหน้าที่ของเซลล์ประสาทในสมองมนุษย์ โดยมีการเรียนรู้ข้อมูลผ่านชุดของ "ชั้น" (Layers) ที่ประกอบด้วย "นิวรอน" (Neurons) หลายตัวที่เชื่อมต่อกัน

โครงสร้างพื้นฐานของ Artificial Neural Network (ANN) ประกอบด้วย 3 ชั้นหลัก ได้แก่

1) Input Layer (ชั้นนำเข้า)

เป็นชั้นแรกของ ANN ที่ใช้รับข้อมูลนำเข้า เช่น รูปภาพ ตัวเลข ข้อความ ฯลฯ อีกทั้งยังไม่มีการประมวลผลข้อมูลในชั้นนี้ มีเพียงการส่งต่อไปยังชั้นถัดไป และจำนวนของ Neuron ในชั้นนี้ยังขึ้นอยู่กับจำนวน Feature ของข้อมูล เช่น หากป้อนภาพขนาด 28×28 พิกเซล จะมี 784 Neuron (28×28 = 784)

2) Hidden Layer (ชั้นซ่อน)

เป็นชั้นที่ทำหน้าที่ประมวลผลข้อมูลโดยใช้ค่าน้ำหนัก (Weights) และฟังก์ชันกระตุ้น (Activation Function) อาจมีหลายชั้น เรียกว่า Deep Neural Network (DNN) เพื่อเพิ่มความสามารถในการเรียนรู้ และแต่ละ Neuron ในชั้นนี้จะรับค่าจากชั้นก่อนหน้า คำนวณผลลัพธ์ แล้วส่งต่อไปยังชั้นถัดไป

3) Output Layer (ชั้นส่งออก)

เป็นชั้นสุดท้ายที่ให้ผลลัพธ์ตามที่ต้องการ จำนวนของ Neuron ในชั้นนี้ขึ้นอยู่กับประเภทของปัญหา เช่น การจำแนกประเภท (Classification) ใช้ Softmax หรือ Sigmoid และการพยากรณ์ค่า (Regression) ใช้ Linear Activation Function เป็นต้น

2.3.7.3 สมการของ Artificial Neural Network

ANN ใช้สมการทางคณิตศาสตร์เพื่อประมวลผลข้อมูล ปรับค่าพารามิเตอร์ และเรียนรู้จากข้อมูลที่ได้รับ สมการที่สำคัญของ ANN มีดังนี้

1) สมการของเซลล์ประสาท (Neuron Equation)

แต่ละ Neuron ใน ANN คำนวณผลลัพธ์โดยใช้สมการพื้นฐานนี้

$$Y = f(\sum i = 1nwixi + b)Y = f\left(\sum_{i = 1}^{n} w_i x_i + b \right)Y = f(i = 1\sum nwixi + b)$$

xix_ixi = ข้อมูลนำเข้า (Input)

wiw_iwi = ค่าน้ำหนัก (Weight)

bbb = ค่าไบแอส (Bias)

f(x)f(x)f(x) = ฟังก์ชันกระตุ้น (Activation Function)

YYY = ผลลัพธ์ของ Neuron

2) สมการของ Forward Propagation

เป็นกระบวนการที่ข้อมูลไหลผ่านเครือข่ายจาก Input ⟶ Hidden Layers ⟶ Output โดยแต่ละชั้นจะคำนวณค่าใหม่ตามสมการนี้

$$\begin{aligned} Z(l) &= W(l)A(l-1) + B(l)Z\hat{}\{(l)\} \\ &= W\hat{}\{(l)\}A\hat{}\{(l-1)\} + B\hat{}\{(l)\}Z(l) \\ &= W(l)A(l-1) + B(l)A(l) = f(Z(l))A\hat{}\{(l)\} \\ &= f(Z\hat{}\{(l)\})A(l) = f(Z(l)) \end{aligned}$$

Z(l)Z^{(l)}Z(l) = ผลรวมเชิงเส้นของชั้นที่ lll

W(l)W^{(l)}W(l) = เมทริกซ์น้ำหนักของชั้นที่ lll

A(l−1)A^{(l−1)}A(l−1) = เอาต์พุตของชั้นก่อนหน้า

B(l)B^{(l)}B(l) = ไบแอสของชั้นที่ lll

f(x)f(x)f(x) = ฟังก์ชันกระตุ้น (Activation Function)

A(l)A^{(l)}A(l) = เอาต์พุตของชั้นที่ lll

3) สมการของฟังก์ชันกระตุ้น (Activation Functions)

ฟังก์ชันกระตุ้นใช้เพื่อกำหนดว่าข้อมูลควรผ่านต่อไปยังชั้นถัดไปหรือไม่

a. Sigmoid Function ใช้ในปัญหาที่ต้องการค่าระหว่าง 0 และ 1 เช่น การจำแนกแบบ Binary

$$f(x) = 1/(1 + e\hat{}(-x))$$

b. ReLU (Rectified Linear Unit) ใช้ใน *Deep Learning* เพราะช่วยลดปัญหา *Gradient Vanishing*

$$f(x) = \max(0, x)f(x)$$

c. Tanh (Hyperbolic Tangent) ให้ค่าระหว่าง *−1* ถึง *1* ทำให้การเรียนรู้มีประสิทธิภาพขึ้น

$$f(x) = (e\hat{}x - e\hat{}(-x))/(e\hat{}x + e\hat{}(-x))$$

4) สมการของ *Loss Function (*ค่าความผิดพลาด)

Loss Function ใช้เพื่อคำนวณความแตกต่างระหว่างค่าที่คาดการณ์และค่าจริง

a. Mean Squared Error (MSE) ใช้ใน *Regression*

$$E = 1/N \sum\_(i = 1)\^N(Y\_predicted - Y\_tr\ ue\ )\^2$$

b. Binary Cross–Entropy ใช้ในปัญหา *Binary Classification*

$$E = -1/N \sum\_(i = 1)\^N \left[\left[Y\_true\ \log (Y\_predicted\ )\right]\right] + (1 - Y\_true\ )\ \log (1 - Y\_predicted\ )$$

c. Categorical Cross–Entropy ใช้ใน *Multi–Class Classification*

$$E = -\sum\_(1\^1 = 1)\^N\sum\_(j = 1)\^C(Y_{tr\ ue^{(i,j)}\log} \left[\left(\left(Y^{(i,j)}_{predicted}\right)\right)\right.$$

5) สมการของ *Backpropagation (*การปรับค่าพารามิเตอร์)

Backpropagation เป็นกระบวนการที่ *ANN* ใช้ปรับค่าน้ำหนักและไบแอสเพื่อลดค่าความผิดพลาด โดยใช้ *Gradient Descent* ดังนี้

a. Gradient Descent Algorithm

$$W = W - n\ \partial E/\ \partial \omega$$

b. การปรับ Bias

$$b = b - n\ \partial E/\ \partial b$$

w = ค่าน้ำหนัก

b = ค่าไบแอส

η = อัตราการเรียนรู้ (Learning Rate)

∂E\∂w = อนุพันธ์ของค่าความผิดพลาดต่อค่าพารามิเตอร์

2.3.7.4 ข้อดีและข้อเสียของ Artificial Neural Network

ข้อดีของ Artificial Neural Network คือ สามารถเรียนรู้รูปแบบที่ซับซ้อนได้ สามารถจับแพทเทิร์นที่ซับซ้อนในข้อมูลได้ดี ทำงานกับข้อมูลที่ไม่มีโครงสร้างแน่นอนได้ ใช้ได้ทั้งรูปภาพ ข้อความ และเสียง และ สามารถเรียนรู้และพัฒนาได้เอง ปรับค่าตัวเองให้เหมาะสมกับข้อมูลที่ได้รับ

ส่วนข้อเสียคือ ต้องใช้ข้อมูลจำนวนมาก ต้องการข้อมูลฝึกสอนจำนวนมากเพื่อให้ได้ผลลัพธ์ที่แม่นยำ มีกระบวนการคำนวณที่ซับซ้อน ใช้ทรัพยากรในการคำนวณสูง ต้องใช้ GPU หรือ TPU อาจเกิด Overfitting ถ้า ANN ซับซ้อนเกินไป อาจเรียนรู้เฉพาะข้อมูลที่ฝึกสอนและไม่สามารถใช้กับข้อมูลใหม่ได้ดี

2.4 วรรณกรรมที่เกี่ยวข้อง

งานวิจัยของวนิดา พงษ์สงวน (2561) พัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วย Decision Tree เพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน พัฒนาแบบจำลองด้วยขั้นตอนวิธีเจสี่สิบแปด (J48) เพื่อประเมินประสิทธิภาพของ แบบจำลองด้วยค่าความแม่นตรง ผลการวิจัยพบว่าแบบจำลองที่พัฒนาให้ประสิทธิภาพที่มีค่าความ แม่นตรงร้อยละ 76.14 และสามารถสร้างกฎการจำแนกจากต้นไม้ตัดสินใจทั้งสิ้น 97 กฎ ซึ่งพบว่า ปัจจัยเสี่ยงที่อาจก่อให้เกิดโรคเบาหวาน ได้แก่ อายุ เพศ สถานะภาพ ที่อยู่ อาชีพ ประวัติความดัน โลหิตเกินมาตรฐาน ประวัติค่าดัชนีมวลกายเกินมาตรฐาน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา และประวัติครอบครัวเป็นเบาหวาน

งานวิจัยของ Talha Mahboob Alam et al. (2019) ได้สร้างต้นแบบการทำนายโรคเบาหวานในระยะเริ่มต้น ซึ่งทำนายโรคเบาหวานโดยใช้คุณลักษณะที่มีนัยสำคัญ และความสัมพันธ์ของ คุณลักษณะที่แตกต่างกัน จากผลการวิจัยบ่งชี้ความเกี่ยวข้องของโรคเบาหวานกับดัชนีมวลกาย (BMI) และระดับกลูโคส ซึ่งสกัดคุณลักษณะด้วยขั้นตอนวิธี Apriori method, Artificial neural network (ANN), Random Forest (RF) และ K-means clustering สำหรับการทำนาย โรคเบาหวาน ผลการวิจัยพบว่า Artificial neural network (ANN) ให้ความแม่นยำสูงสุดร้อยละ 75.7ข้อดีของ DBSCAN คือ สามารถจัดกลุ่มข้อมูลที่มีรูปทรงไม่เป็นทรงกลม เช่น รูปตัว U หรือเส้นโค้ง ไม่ต้องกำหนดจำนวนกลุ่ม KK ล่วงหน้า และสามารถตรวจจับ Noise หรือ Outliers ได้โดยอัตโนมัติ

งานวิจัยของ Sidong Wei et al. (2018) เป็นการสำรวจที่ครอบคลุมรวมถึงอัลกอริทึมการ เรียนรู้ของเครื่องสำหรับการระบุโรคเบาหวาน ทำการสำรวจที่ครอบคลุม และใช้

อัลกอริทึมที่ได้รับ ความนิยมมากที่สุด เช่น Deep Neural Network (DNN), Support Vector Machine (SVM) ที่ใช้ในการระบุโรคเบาหวานและวิธีการประมวลผลข้อมูลล่วงหน้า หลังการประเมินโดย 10-fold cross-validation พบว่าอัลกอริทึมที่มีประสิทธิภาพดีที่สุดคือ Deep Neural Network (DNN) ให้ความแม่นยำร้อยละ 77.86

นางสาวเมธาพร ผ่องยิ่ง (2565) งานวิจัยนี้มีเป้าหมายเพื่อเปรียบเทียบประสิทธิภาพของเทคนิค Machine Learning ในการจำแนกการเป็นโรคเบาหวาน โดยพิจารณาและไม่พิจารณาอิทธิพลร่วม โดยใช้ 4 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) ป่าสุ่ม (Random Forest) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) และวิธีเคใกล้เคียงที่สุด K-Nearest Neighbor (KNN) ผลการศึกษาพบว่าแบบจำลองที่พิจารณาอิทธิพลร่วมมีประสิทธิภาพสูงกว่าแบบจำลองที่ไม่พิจารณาอิทธิพลร่วม คือเทคนิค Random Forest ให้ค่าความถูกต้องสูงสุดที่ 97.5% ในกรณีที่พิจารณาอิทธิพลร่วม และ 88.2% ในกรณีที่ไม่พิจารณาอิทธิพลร่วม

อรุณรักษ์ ตันพานิช และคณะ (2562) ทำการวิเคราะห์ข้อมูลที่ใช้ได้มาจากโรงพยาบาลส่งเสริมสุขภาพประจำตำบลท่าจีน จังหวัดสงขลา จำนวนทั้งสิ้น 300 ชุด โดยมีกลุ่มตัวอย่างจำนวน 2 กลุ่ม ได้แก่ กลุ่มผู้ป่วยโรคเบาหวานชนิดที่ 2 ที่มีภาวะชาปลายเท้า และผู้ป่วยปกติ เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ ได้แก่ ค่าความถูกต้อง ส่วนเบี่ยงเบนมาตรฐาน (SD) และระยะเวลาในการประมวลผล เทคนิคในการทำเหมืองข้อมูล (Data Mining Techniques) โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน Support Vector Machine (SVM) เทคนิคการเรียนรู้แบบอัตโนมัติด้วยการเลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ Deep Learning และเทคนิคป่าสุ่ม Random Forest

ธรรญชนก นิลมณี และ จรัญ แสนราช (2564) ผู้วิจัยได้นำข้อมูลภาวะการมีงานทำของบัณฑิตและผลการเรียนตลอดหลักสูตรของนักศึกษาระดับปริญญาตรี คณะบริหารธุรกิจของมหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ ที่มีการจัดการศึกษาประกอบด้วย 3 พื้นที่ คือ พื้นที่บพิตรพิมุข จักรวรรดิ พื้นที่ศาลายา และพื้นที่วิทยาเขตวังไกลกังวล ในปีการศึกษา พ.ศ. 2557-2561จำนวน 6,185 ระเบียน มาศึกษาเทคนิคการทำนายความสอดคล้องระหว่างอาชีพที่ทำงาน กับสาขาที่ตนเองเรียนมา โดยใช้เทคนิคดาต้าไมน์นิ่ง ในการจำแนกข้อมูล (Classification) เพื่อเป็นการเพิ่ม ดังนั้นผู้วิจัยจึงได้นำข้อมูลภาวะการมีงานทำของบัณฑิต โดยเลือกใช้เทคนิควิธีจำแนกข้อมูลที่เหมาะสมกับข้อมูลและได้ค่าความแม่นยำที่อยู่ในระดับที่ยอมรับได้ ด้วยการใช้คุณลักษณะทางด้านผลการเรียนของผู้สำเร็จการศึกษาและ ด้านตำแหน่งงาน โดยเทคนิคดาต้าไมน์นิ่งวิธีการ 4 วิธี ที่ผู้วิจัยเลือกใช้ได้แก่ วิธีต้นไม้ตัดสินใจ

(Decision tree) วิธีแบบเบย์ (Naive Bayes) วิธีแรนดอมฟอร์เรส (Random forest) และ วิธีการเรียนรู้เชิงลึก (Deep learning) นำแต่ละวิธีมาเปรียบเทียบประสิทธิภาพ พบว่า วิธีการเรียนรู้เชิงลึกให้ค่าความถูกต้องสูงที่สุด คิดเป็นร้อยละ 83.70 รองลงมาคือ วิธีแรนดอมฟอร์เรส คิดเป็นร้อยละ 83.20 วิธีต้นไม้ตัดสินใจ คิดเป็นร้อยละ 83.00 และวิธีแบบเบย์ คิดเป็นร้อยละ 80.90 ตามลำดับ

บีสุดา ดาวเรือง จรัญ แสนราช และ อนิราช มิ่งขวัญ (2564) การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกตัวแบบทำนายแขนงวิชาเรียนของ นักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ วิทยาเขตปราจีนบุรี โดยใช้เทคนิคเหมืองข้อมูล 3 วิธี คือวิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ และวิธีฐานกฎนำมาเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกหาตัวแบบที่เหมาะสมเพื่อใช้ทำนายแขนงวิชาเรียนที่เหมาะสมกับนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ โดยรวบรวมข้อมูลของนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ วิทยาเขตปราจีนบุรี ตั้งแต่ปีการศึกษา 2548–2559 จำนวน 759 ระเบียน วิเคราะห์ข้อมูลบนพื้นฐานของวิธี 10–Fold Cross Validation โดยใช้โปรแกรม Rapid Miner Studio 7 ในการสร้างแบบจำลอง ผลการศึกษาพบว่าวิธีฐานกฎมีประสิทธิภาพสูงสุด ให้ค่าความถูกต้องเท่ากับ 78.66% รองลงมาคือวิธีต้นไม้ตัดสินใจที่ให้ค่าความถูกต้องเท่ากับ 78.26% และวิธีแบบเบย์ให้ค่าความถูกต้องเท่ากับ 78.12% จากผลการเปรียบเทียบประสิทธิภาพในครั้งนี้ สามารถนำวิธีฐานกฎไปใช้ในการทำนายแขนงวิชาเพื่อให้นักศึกษาเลือกเรียนให้เหมาะสม ซึ่งเป็นแนวทางในการเสนอแนะแขนงวิชาเรียนแก่นักศึกษาให้เรียนได้อย่างมีประสิทธิภาพ และสามารถจบการศึกษาได้ในระยะเวลาตามหลักสูตร

กาญจน์ ณ ศรีธะ (2560) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกด้วยเทคนิคการ เรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลผู้ป่วยโรคมะเร็งเต้านม (Breast Cancer) ชุดข้อมูล ผู้ป่วยโรคเบาหวานประเภทที่ 2 (Pima Indian Diabetes) และชุดข้อมูลผู้ป่วยโรคเท้าเบาหวาน (Diabetic foot) โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลประกอบด้วย วิธีสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Over–Sampling Technique) วิธีสุ่มลด (Under Sampling) และเทคนิคการสุ่ม ตัวอย่างซ้ำ เทคนิคการจำแนกที่ใช้ประกอบด้วย ต้นไม้ตัดสินใจ

เทคนิคป่าสุ่ม เครื่องสนับสนุนเวกเตอร์และโครงข่ายประสาทเทียม และเทคนิคที่ใช้ช่วยเพิ่มประสิทธิภาพการจำแนก ได้แก่ เทคนิคเอดาบูทและเทคนิคถุงจำแนก ผลจากงานวิจัยพบว่าเมื่อพิจารณาที่ชุดข้อมูล40 ผู้ป่วยโรคมะเร็งเต้านม เทคนิคป่าสุ่มโดยปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการสุ่มตัวอย่าง ซ้ำมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 86.38 ค่าการเรียกคืนร้อยละ 86.64 ค่าประสิทธิภาพร้อยละ 86.40 เมื่อพิจารณาที่ชุดข้อมูลผู้ป่วยโรคเบาหวานประเภทที่ 2 เทคนิคป่าสุ่มโดยปรับปรุงข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการสุ่มตัวอย่างซ้ำมีประสิทธิภาพในการ จำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 91.25 ค่าการเรียกคืนร้อยละ 91.28 ค่าประสิทธิภาพร้อย ละ 91.19 และเมื่อพิจารณาที่ชุดข้อมูลผู้ป่วยโรคเท้าเบาหวาน เทคนิคป่าสุ่มที่ทำงานร่วมกับเทคนิคเอ ดาบูทมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 86.46 ค่าการเรียกคืนร้อยละ 86.41 ค่าประสิทธิภาพร้อยละ 85.91

สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณ์ทอง (2560) ผู้วิจัยจึงได้มีความสนใจที่จะนำข้อมูลเกี่ยวกับโรคไตเรื้อรังมาวิเคราะห์ในการทำวิจัยครั้งนี้ ซึ่งได้มีการ ค้นคว้าข้อมูลจาdฐานข้อมูล UCI Machine Learning Repository จำนวน 1 ชุด คือ ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย (Chronic Kidney Disease of Apollo Hospitals, India) รวบรวมข้อมูลผู้ป่วยโรคไตเรื้อรังโดย Rubiniนักวิชาการจากมหาวิทยาลัยอลากัปปา (L.Jerlin Rubini Research Scholar Alagappa University) บริจาคข้อมูล ณ วันที่ 3 กรกฎาคม 2558 ซึ่งได้มีการเก็บรวบรวมข้อมูลเป็นระยะเวลา 2 เดือนจำนวนข้อมูล ที่เก็บรวบรวมทั้งหมด 400 ระเบียน ประกอบด้วยตัวแปรอิสระ 24 ตัวแปร และตัวแปรตามศึกษาและหาข้อมูลเพื่อนำมาเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนก ด้วยวิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) วิธีต้นไม้ตัดสินใจ (Decision tree) วิธีโครงข่ายประสาทเทียม (Artificial neural network) วิธีซัพพอร์ทเวกเตอร์-แมชชีน (Support vector machine) วิธีฐานกฎ (rule-based) วิธีการถดถอยลอจิสติก (Logistic regression) และวิธีนาอีฟเบย์ (Naive Bayes) โดยใช้หลักการของการทำเหมืองข้อมูลมาใช้ในการจำแนกข้อมูล ซึ่งสรุปผลได้ดังนี้ วิธีความใกล้เคียงกันมากที่สุดมีค่าความถูกต้อง คือ 94.17 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0579 วิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง คือ 100% และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0059 วิธีโครงข่ายประสาทเทียมมีค่าความถูกต้อง คือ 98.33 % ซึ่งวิธีการจำแนกกลุ่มที่มีประสิทธิภาพในการจำแนกดีที่สุดสำหรับข้อมูลโรคไตเรื้อรัง คือ วิธีต้นไม้ตัดสินใจการ

สรุปผลการวิจัยครั้งนี้ ดังนั้นผู้วิจัยจึงเลือกวิธีต้นไม้ตัดสินใจ สำหรับการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโรงพยาบาลอพอลโลประเทศอินเดีย เพราะว่ามีค่าความถูกต้องอยู่ในระดับสูงที่สุดและความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดผลการสรุปดังกล่าวใกล้เคียงกับในช่วงที่ผ่านมา

รัชนีวรรณ ไพศาลวรเกียรติ (2564) งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเป็นโรคเบาหวานและเปรียบเทียบตัวแบบพยากรณ์การเป็นโรคเบาหวานด้วยวิธีการถดถอยลอจิสติกทวิภาค ต้นไม้ตัดสินใจ ด้วยอัลกอริทึม J48 และ LMT และเทคนิคโครงข่ายประสาทเทียม โดยใช้ข้อมูลผู้ป่วยที่เข้ามารับบริการในโรงพยาบาลมหาวิทยาลัยนเรศวร จำนวน 5,081 ชุด โดยแบ่งข้อมูลเป็นข้อมูลเรียนรู้ และข้อมูลทดสอบด้วยสัดส่วน 70:30 และ 80:20 ทำการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์บนข้อมูลทดสอบด้วยค่าความถูกต้อง และค่าความถูกต้องในการจำแนกแบบสมดุล จากการศึกษาพบว่า ปัจจัยที่ส่งผลต่อการเป็นโรคเบาหวาน คือ ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลาย ตัวอัตราการเต้นของหัวใจ น้ำหนัก ความสูง และระดับน้ำตาลในเลือด นอกจากนี้จากผลการศึกษา พบว่า เทคนิคโครงข่ายประสาทเทียมมีประสิทธิภาพในการพยากรณ์ดีที่สุดในชุดข้อมูลทั้ง 2 แบบ โดยในชุดข้อมูลแบบที่ 1 (70:30) ให้ค่าความถูกต้อง และค่าความถูกต้องในการจำแนกแบบสมดุล เท่ากับ 81.7824% และ 73.9704% ตามลำดับ และในชุดข้อมูลแบบที่ 2 (80:20) ให้ค่าความถูกต้อง และค่าความถูกต้องในการจำแนกแบบสมดุล เท่ากับ 81.4159% และ 73.7482% ตามลำดับ โดยเทคนิคที่มีประสิทธิภาพรองลงมา คือเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม J48 และ LMT และ เทคนิคการถดถอยลอจิสติกทวิภาค ตามลำดับ

A. Rahman Khan,M. Rahman,J. Us Salehin,S. Islam and F. Rabbi. (2021) ทำวิจัยเรื่องเทคนิคการขุดข้อมูลอย่างมีประสิทธิภาพสำหรับโรคหัวใจการทำนายและการวิเคราะห์เปรียบเทียบของอัลกอริทึมการจำแนกประเภท งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์เปรียบเทียบการทำนายโรคหัวใจโดยใช้เทคนิคนาอีฟเบย์ เทคนิคต้นไม้ตัดสินใจ เทคนิคเครือข่ายประสาทเทียมและเทคนิคการสุ่มป่าไม้ การวิจัยใช้ชุดข้อมูลดิบและเครื่องมือเวก้าสำหรับการวิเคราะห์ประสิทธิภาพ พบว่าเทคนิคดีที่สุดเป็นเทคนิคการสุ่มป่าไม้ โดยให้ค่าความถูกต้องร้อยละ 97.7049

สายชล สินสมบูรณ์ทอง (2561) ทำการศึกษาเปรียบเทียบประสิทธิภาพในการพยากรณ์การเป็นโรคเบาหวานของโรงพยาบาลแห่งหนึ่ง ข้อมูลตั้งแต่เดือนมกราคม – เมษายน พ.ศ. 2559 จำนวน 1,233 ชุด โดยใช้เทคนิคการจำแนกประเภทข้อมูล 6 เทคนิค ประกอบด้วย

เทคนิคความใกล้เคียงกันมากที่สุดโดยใช้อัลกอริทึม IBk เทคนิคต้นไม้ตัดสินใจโดยใช้อัลกอริทึมชนิด J48 เทคนิคโครงข่าย ประสาทเทียมโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น เทคนิคซัพพอร์ตเวกเตอร์แมชชีน โดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล เทคนิคการถดถอยโลจิสติกทวิภาค และเทคนิคนาอีฟเบย์ โดยพิจารณาค่าความถูกต้อง ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ จากผลการศึกษา พบว่าเทคนิคโครงข่ายประสาทเทียมมีประสิทธิภาพในการพยากรณ์ดีที่สุด โดยให้ค่าความถูกต้อง ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 95.94 เปอร์เซ็นต์ 0.0491 และ 0.0396 ตามลำดับ

Luís Chaves และ Gonçalo Marques (2021) งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูล 6 วิธี ได้แก่ Naive Bayes, Neural Network, AdaBoost, k–Nearest Neighbors (KNN), Random Forest และ Support Vector Machine (SVM) สำหรับการวินิจฉัยโรคเบาหวานระยะเริ่มต้น โดยใช้ชุดข้อมูลผู้ป่วยจากโรงพยาบาล Sylhet Diabetes Hospital ในบังคลาเทศ จำนวน 520 ราย อายุระหว่าง 16–90 ปี และ 16 ตัวแปรอาการทางคลินิก ผลการทดสอบด้วย 10–Fold Cross–Validation พบว่า Neural Network ให้ผลแม่นยำสูงสุดที่ 98.1% (AUC 0.983, F1–Score 0.984, Precision 0.984, Sensitivity 0.984, Specificity 0.975) รองลงมาคือ KNN และ AdaBoost ที่ 97.31% และ Random Forest 96.9% ส่วน Naive Bayes ทำได้ต่ำสุดที่ 86.9% โดยตัวแปรเสี่ยงสำคัญที่สุดคือ Polyuria, Polydipsia และ Gender งานวิจัยนี้ชี้ให้เห็นว่าเทคนิค Neural Network เหมาะสมที่สุดสำหรับการวินิจฉัยโรคเบาหวานจากข้อมูลลักษณะทางคลินิก และแนะนำให้เพิ่มข้อมูลด้านพฤติกรรมและพันธุกรรมในอนาคตเพื่อเพิ่มความแม่นยำ

Alain Hennebelle, Huned Materwala และ Leila Ismail (2023) งานวิจัยนี้นำเสนอ HealthEdge กรอบงานระบบการดูแลสุขภาพอัจฉริยะที่ใช้ Machine Learning สำหรับพยากรณ์โรคเบาหวานชนิดที่ 2 โดยเชื่อมต่อระบบ IoT, Edge และ Cloud Computing พร้อมทดสอบเปรียบเทียบประสิทธิภาพของ Random Forest และ Logistic Regression บนชุดข้อมูล PIMA Indian และ Sylhet Diabetes Dataset โดยใช้เทคนิค SMOTE และ Recursive Feature Elimination (RFECV) ช่วยจัดสมดุลและคัดเลือกตัวแปรเสี่ยง ผลการทดลองด้วย 10–Fold Cross–Validation พบว่า Random Forest ให้ผลแม่นยำสูงกว่า Logistic Regression ในทุกกรณี โดยบนชุด PIMA RF+FS+Balancing ได้ Accuracy สูงสุดที่ 81.85% (F1–Score 0.8085, AUC 0.8100) ส่วนบนชุด Sylhet RF+FS+Balancing ได้ Accuracy สูงถึง 98.13% (F1–Score 0.9788, AUC 0.9800)

ซึ่งดีกว่า LR ทุกค่า ชี้ให้เห็นว่าการใช้ Random Forest ร่วมกับการเลือกตัวแปรและจัดสมดุลข้อมูลเหมาะสมที่สุดสำหรับการพยากรณ์เบาหวานชนิดที่ 2 ในระบบ Smart Healthcare Framework

Maydanchi et al. (2024) ผู้วิจัยได้นำ Machine Learning มาทดสอบและเปรียบเทียบความแม่นยำของ 9 อัลกอริธึม ในการทำนายเบาหวานจากข้อมูลสุขภาพของผู้ป่วย ใช้ชุดข้อมูลเบาหวานจาก Kaggle ขนาด 100,000 ตัวอย่าง ตัวแปรมีทั้งประเภทเชิงหมวดหมู่ (Gender, Hypertension, Heart Disease, Smoking) และตัวเลข (Age, BMI, HbA1c, Glucose Level) มีการปรับค่าพารามิเตอร์ของแต่ละโมเดลด้วย GridSearchCV และเปรียบเทียบประสิทธิภาพด้วย Accuracy, F1-score, AUC และ Runtime งานวิจัยนี้เปรียบเทียบ 9 ML โมเดลทำนายเบาหวานจากข้อมูลขนาดใหญ่ พบว่า Gradient Boosting, XGBoost และ AdaBoost ให้ผลแม่นยำสูงสุด (f1-score & AUC) แต่ใช้เวลา train นานกว่าโมเดลอื่นๆ ขณะที่ Naïve Bayes และ KNN ทำงานเร็วมากแต่ความแม่นยำต่ำกว่า เหมาะกับงานที่ต้องการ real-time

Shweta Yadu, Rashmi Chandra และ Vivek Kumar Sinha (2024) งานวิจัยนี้มุ่งเน้นการใช้เทคนิค Machine Learning (ML) เพื่อทำนายโรคเบาหวานในระยะเริ่มต้น โดยใช้ข้อมูลจากโรงพยาบาลในบังกลาเทศจำนวน 520 รายการ ที่มีลักษณะเฉพาะ 16 รายการ (เช่น เพศ, อายุ, อาการต่างๆ) โดยใช้โมเดล Random Forest, AdaBoost, K-Nearest Neighbor (KNN) และ Bagging ผลการทดสองพบว่าโมเดล Random Forest มีค่าความแม่นยำสูงสุดเท่ากับ 97.03 ทั้งและผลจาก Cross-Validation พบว่าโมเดล Random Forest มีค่าความแม่นยำสูงสุดเท่ากับ 95.03

Chun-Yang Chou, Ding-Yang Hsu และ Chun-Hung Chou (2023) งานวิจัยนี้ศึกษาการใช้แมชชีนเลิร์นนิงเพื่อพยากรณ์การเกิดโรคเบาหวานในผู้หญิง 15,000 คน อายุ 20–80 ปี จากข้อมูลตรวจสุขภาพผู้ป่วยนอกของโรงพยาบาลในไทเปช่วงปี 2018–2022 โดยใช้ข้อมูล 8 ตัวแปร ได้แก่ จำนวนการตั้งครรภ์ ระดับกลูโคส ความดันเลือด ไขมันใต้ผิวหนัง อินซูลิน BMI พันธุกรรมเบาหวาน และอายุ โดยเปรียบเทียบ 4 โมเดล ได้แก่ Logistic Regression, Neural Network, Decision Jungle และ Boosted Decision Tree พบว่าโมเดล Two-Class Boosted Decision Tree ให้ผลดีที่สุดด้วย Accuracy 95.3%, Precision 92.7%, Recall 93.1%, F1-score 92.9% และ AUC 0.991 ซึ่งเหนือกว่าโมเดลอื่นและค่ามาตรฐานในวรรณกรรมก่อนหน้า แสดง

ให้เห็นถึงศักยภาพในการใช้โมเดลนี้ช่วยแพทย์วินิจฉัยโรคเบาหวานในผู้ป่วยแม้ไม่มีอาการได้อย่างแม่นยำ

นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา (2565) งานวิจัยนี้ผู้วิจัยมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวาน โดยใช้เทคนิคการเลือกคุณลักษณะเด่นไปข้างหน้า (Forward Feature Selection) ร่วมกับวิธีการตัดสินใจแบบรวม (Ensemble Decision-Making) ซึ่งประกอบด้วยการรวมผลการตัดสินใจจากโมเดลต้นไม้ตัดสินใจ (Decision Tree) 3 แบบ ได้แก่ Random Forest, Gradient Boosted Decision Tree และ Voting Tree เพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูล ข้อมูลที่ใช้ในการวิจัยมาจากเวชระเบียนผู้ป่วยโรคเบาหวานของโรงพยาบาลสมเด็จพระยุพราชบ้านดุง ในช่วงปี พ.ศ. 2557-2561 ซึ่งเป็นข้อมูลที่มีมิติสูง (High-Dimensional Data) การเลือกคุณลักษณะเด่นช่วยลดความซ้ำซ้อนของข้อมูลและเพิ่มความแม่นยำในการจำแนกข้อมูล ผลลัพธ์การทดลองพบว่าโดยใช้วิธี 10- Fold Cross Validation ผลการเปรียบเทียบพบว่า วิธีต้นไม้ตัดสินใจให้สัมประสิทธิภาพสูงสุด โดยมีค่าความถูกต้อง 93.73% วิธีนาอีฟเบย์ค่าความถูกต้อง 88.92% วิธีความใกล้เคียงกันที่สุดและวิธีซัพพอร์ตเวกเตอร์แมชชีนค่าความถูกต้อง 86.97% และ 86.13% ตามลำดับ จะพบว่า วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการสร้างแบบจำลองมากที่สุด

Savitesh Kushwaha, Rachana Srivastava, Rachita Jain, Vivek Sagar, Arun Kumar Aggarwal, Sanjay Kumar Bhadada, Poonam Khanna (2565) งานวิจัยนี้พัฒนาเครื่องมือคัดกรองภาวะก่อนเบาหวานในเด็กและวัยรุ่นอายุ 5–19 ปี โดยใช้โมเดลแมชชีนเลิร์นนิงแบบไม่รุกล้ำจากข้อมูลประชากรจำนวน 26,567 คน (ปกติ 23,777 คน, ก่อนเบาหวาน 2,790 คน) โดยใช้ 8 ตัวแปรที่ไม่ใช้สารชีวภาพ เช่น น้ำหนัก รอบเอว ความสูง และอายุ ผ่านการปรับสมดุลข้อมูลด้วย SVM-SMOTE และ Random Under Sampler จากนั้นเปรียบเทียบประสิทธิภาพของ 6 โมเดล พบว่า XGBoost มีความแม่นยำสูงสุด (Accuracy 94.24%, 10-fold Cross-validation 90.13%, Cohen's Kappa 88.48%, AUC 0.959) และถูกนำไปพัฒนาเป็นเครื่องมือคัดกรองแบบออนไลน์ในเวลาจริง ซึ่งมีศักยภาพในการใช้งานในชุมชนเพื่อป้องกันการลุกลามของเบาหวานในเด็กได้อย่างมีประสิทธิภาพ

Wenhui Zhou, Xiaomin Liu, Hongtao Bai, Lili He (2024) งานวิจัยนี้เสนอโมเดล Interpretable Predictor with Deep Convolutional Fuzzy Neural Network (IP-DCFNN) สำหรับ

วินิจฉัยและรักษาโรคเบาหวาน โดยผสานความสามารถของโครงข่ายประสาทเทียมเชิงลึกและตรรกะฟัซซีเข้าด้วยกัน เพื่อให้ได้ผลการพยากรณ์ที่แม่นยำและสามารถตีความได้ โมเดลนี้ถูกทดสอบกับ 3 ชุดข้อมูล ได้แก่ Pima Indians, Iraqi Patient และ NHANES ซึ่งให้ผลแม่นยำเฉลี่ยสูงกว่าโมเดลเปรียบเทียบ (ANFIS และ DNN) ถึง 7.4% โดยเฉพาะในชุดข้อมูล Iraqi Patient ความแม่นยำสูงถึง 90.95% (Precision 93.26%, Recall 96.51%, F1-score 94.86%) ทั้งนี้ระบบสามารถให้คำอธิบายเบื้องหลังผลลัพธ์ผ่านกฎฟัซซีที่ตีความได้ เช่น ผลกระทบของจำนวนการตั้งครรภ์และค่าดัชนีมวลกาย (BMI) ต่อความเสี่ยงโรคเบาหวาน พร้อมคำแนะนำเฉพาะบุคคลเช่นการควบคุมน้ำหนักและการเฝ้าระวังระดับน้ำตาลในเลือด เพื่อใช้ในการดูแลผู้ป่วยเชิงป้องกันและสนับสนุนการวินิจฉัยของแพทย์ในอนาคต

จากการศึกษาจิจัยพบว่า มีหลายปัจจัยเสี่ยงที่ทำให้เกิดโรคเบาหวานได้ ไม่ว่าจะเป็นอายุ เพศ ประวัติความดัน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา และประวัติครอบครัวเป็นเบาหวาน ซึ่งโรคเบาหวานนั้นเกิดจากระดับน้ำตาลในเลือดมีมากจนเกินค่า เป็นโรคที่ไม่สามารถรักษาให้หายขาดได้โดยสิ้นเชิง แต่สามารถควบคุมให้อยู่ในระดับปกติได้ โดยเฉพาะในโรคเบาหวานชนิดที่ 2 ต้องผ่านการปรับเปลี่ยนพฤติกรรมและการดูแลสุขภาพอย่างเคร่งครัด ทั้งนี้ได้มุ่งเน้นการใช้เทคนิค Machine Learning และ Data Mining ในการวิเคราะห์และจำแนกข้อมูลด้านสุขภาพและการศึกษา โดยมีการศึกษาปัจจัยที่ส่งผลต่อโรคเบาหวาน เช่น อายุ เพศ ดัชนีมวลกาย (BMI) และพฤติกรรมสุขภาพ ผ่านอัลกอริทึมต่างๆ เช่น Decision Tree, Random Forest, Support Vector Machine (SVM), Deep Neural Network (DNN) และ Artificial Neural Network (ANN) ซึ่งพบว่า DNN และ Random Forest ให้ความแม่นยำสูงสุดในหลายกรณี อีกทั้งยังมีการนำเทคนิคการทำเหมืองข้อมูล เช่น Apriori Method และ K-Means Clustering มาใช้เพื่อสกัดคุณลักษณะสำคัญ นอกจากนี้มีงานวิจัยที่วิเคราะห์ข้อมูลด้านอาชีพของบัณฑิตและการได้รับทุนการศึกษา โดยใช้เทคนิค Decision Tree, Naive Bayes, Deep Learning และ Rule-Based Classification เพื่อทำนายความสอดคล้องของอาชีพและสาขาที่เรียน รวมถึงโอกาสได้รับทุน ซึ่งพบว่า Random Forest และ Deep Learning มักให้ผลลัพธ์ที่แม่นยำที่สุด งานวิจัยเหล่านี้สะท้อนถึงศักยภาพของ AI และ Machine Learning ในการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อสนับสนุนการตัดสินใจในด้านสุขภาพและการศึกษา

จากการทบทวนงานวิจัยด้านการวิเคราะห์และพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล พบว่าองค์ประกอบความเสี่ยงหลักที่ถูกนำมาใช้ในการสร้างแบบจำลองและ

วิเคราะห์ข้อมูลมีความหลากหลายและครอบคลุมปัจจัยสำคัญทางสุขภาพและพฤติกรรมของบุคคล โดยปัจจัยที่พบได้บ่อย ได้แก่ อายุ เพศ ดัชนีมวลกาย (BMI) ความดันโลหิต ประวัติการสูบบุหรี่ การดื่มสุรา ประวัติโรคเบาหวานในครอบครัว ระดับน้ำตาลในเลือด ระดับไขมันในเลือด รอบเอว น้ำหนัก ส่วนสูง จำนวนครั้งที่ตั้งครรภ์ และอาการทางคลินิก เช่น ปัสสาวะบ่อย กระหายน้ำ น้ำหนักลด กล้ามเนื้ออ่อนแรง เป็นต้น ปัจจัยเหล่านี้ล้วนมีความสัมพันธ์กับความเสี่ยงในการเกิดโรคเบาหวานและถูกนำมาใช้เป็นตัวแปรสำคัญในการสร้างแบบจำลองพยากรณ์และคัดกรองโรคเบาหวานในงานวิจัยต่างๆ

ในด้านเทคนิคเหมืองข้อมูลที่ใช้ในการวิเคราะห์และพยากรณ์โรคเบาหวาน งานวิจัยส่วนใหญ่ได้เปรียบเทียบประสิทธิภาพของเทคนิคต่าง ๆ เพื่อค้นหาเทคนิคที่เหมาะสมและให้ผลลัพธ์ดีที่สุด เทคนิคที่ได้รับความนิยมและให้ประสิทธิภาพสูง ได้แก่ Random Forest, Artificial Neural Network (ANN) หรือ Deep Neural Network (DNN), Gradient Boosting, XGBoost, AdaBoost, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), และ Decision Tree (J48) โดยเทคนิค Random Forest และ ANN/DNN มักให้ค่าความแม่นยำสูง เช่น Random Forest มีค่าความถูกต้องสูงถึง 97.5% ในบางกรณี และ ANN/DNN ให้ค่าความถูกต้องสูงสุดถึง 98.1% ในบางชุดข้อมูล นอกจากนี้ เทคนิค Gradient Boosting และ XGBoost ก็ให้ F1-score และ Accuracy สูง เช่น 81% และ 94.24% ตามลำดับ ขณะที่เทคนิคใหม่อย่าง Deep Convolutional Fuzzy Neural Network (IP-DCFNN) ก็แสดงศักยภาพด้วย Accuracy สูงสุดถึง 90.95% ในบางชุดข้อมูล

สำหรับแนวทางการทำวิจัยต่อไป ควรให้ความสำคัญกับการเลือกใช้ปัจจัยเสี่ยงที่เกี่ยวข้องกับสุขภาพและพฤติกรรมของกลุ่มเป้าหมายอย่างรอบด้าน รวมถึงการใช้เทคนิคเหมืองข้อมูลที่มีประสิทธิภาพสูง เช่น Random Forest, ANN/DNN, Gradient Boosting, XGBoost และเทคนิคใหม่ ๆ อย่าง IP-DCFNN เพื่อเพิ่มความแม่นยำในการพยากรณ์และคัดกรองโรคเบาหวาน นอกจากนี้ ควรพิจารณาการประยุกต์ใช้เทคนิคเหล่านี้กับข้อมูลขนาดใหญ่และหลากหลายกลุ่มประชากร เพื่อให้ได้แบบจำลองที่มีความแม่นยำและสามารถนำไปใช้ในงานคัดกรองหรือวินิจฉัยโรคเบาหวานในเชิงปฏิบัติได้อย่างมีประสิทธิภาพ อีกทั้งควรศึกษาการผสมผสานเทคนิคหลายรูปแบบ และการปรับแต่งพารามิเตอร์ให้เหมาะสมกับลักษณะข้อมูลเพื่อเพิ่มศักยภาพของแบบจำลองให้สูงขึ้น รวมถึงการพัฒนาเครื่องมือหรือโปรแกรมสำหรับคัดกรองและวินิจฉัยโรคเบาหวานที่ใช้งานง่ายและเข้าถึงกลุ่มเป้าหมายได้อย่างกว้างขวางในอนาคต

2.5 บทสรุป

จากแนวคิด ทฤษฎี เครื่องมือ และวรรณกรรมที่เกี่ยวข้อง จากที่ได้กล่าวมาในข้างต้นมีเนื้อหาเกี่ยวข้องกับการวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวาน จากชุดข้อมูลทุติยภูมิแบบซึ่งนำมาดำเนินการวิเคราะห์ตามขั้นตอนของกระบวนการ CRISP-DM (Cross Industry Standard Process for Data Mining) จากนั้นจึงเลือกใช้เทคนิคการทำเหมืองข้อมูลแบบจำแนกประเภท (Classification) โดยใช้โมเดลต้นไม้ตัดสินใจ (Decision Tree) นาอีฟเบย์ (Naïve Bayes) ป่าสุ่ม (Random Forest) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และโครงข่ายประสาทเทียม (Artificial Neural Network)เพื่อช่วยในการทำนายผลลัพธ์และจำแนกกลุ่มข้อมูลได้อย่างแม่นยำ โดยใช้โปรแกรม RapidMiner Studio เป็นเครื่องมือหลักในการสร้างแบบจำลอง เพื่อให้สามารถวิเคราะห์หาปัจจัยที่ส่งผลต่อการเป็นโรคเบาหวานได้อย่างชัดเจน พร้อมทั้งนำผลการวิเคราะห์มาแสดงผลในรูปแบบ Visualization ผ่านแดชบอร์ดและสร้างส่วนติดต่อประสานระหว่างผู้ใช้ (User Interface) กับ ระบบการทำเหมืองข้อมูล และเผยแพร่ผ่านเว็บไซต์