บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

โครงงานวิจัยเรื่อง พีพี'เว็บแอปพลิเคชันคัดกรองบุคคลทั่วไปที่มีความเสี่ยงภาวะซึมเศร้าโดยเทคนิคการทำเหมืองเว็บข้อมูล ผู้วิจัยได้ทำการศึกษาและรวบรวมทฤษฎีหรืองานวิจัยที่เกี่ยวข้องสำหรับการทำโครงงานวิจัย ดังนี้

2.1 ภาวะซึมเศร้า (Depression) (pobpad, 2559: ออนไลน์)

ภาวะซึมเศร้า เป็นภาวะความผิดปกติทางอารมณ์ที่ผู้ป่วยอาจรู้สึกเศร้า สิ้นหวัง ช่วยเหลือตัวเองไม่ได้ หรือรู้สึกว่าตนด้อยค่า แม้ความรู้สึกและอารมณ์เหล่านี้จะเกิดขึ้นได้กับทุกคนเป็นครั้งคราว แต่อาการของภาวะซึมเศร้านั้นมีความรุนแรงและยาวนานกว่ามาก จนถึงขั้นส่งผลกระทบต่อการใช้ชีวิตในด้านต่างๆของผู้ป่วย โรคซึมเศร้าเกิดขึ้นได้กับคนทุกเพศทุกวัยแต่โดยมากมักเริ่มตั้งแต่ช่วงอายุ 20-30 ปี โรคความผิดปกติทางอารมณ์ส่วนใหญ่จะเริ่มพัฒนามาจากช่วงวัยรุ่นที่มีความเครียดและความวิตกกังวลสูง ดังนั้น ยิ่งในช่วงวัยรุ่นประสบกับความกังวลมากเท่าไหร่ ก็ยิ่งมีความเสี่ยงต่อการเกิดโรคซึมเศร้าในวัยผู้ใหญ่มากขึ้นเท่านั้น จากสถิติทั่วโลกพบว่ามีผู้ป่วยโรคซึมเศร้าประมาณ 350 ล้านคน มีความชุกราว 2-10 เปอร์เซ็นต์ และเป็นในเพศหญิงมากกว่าเพศชาย ส่วนในไทยมีจำนวนผู้ป่วยซึมเศร้าเพศหญิงเป็นอันดับที่ 3 และเพศชายเป็นอันดับที่ 8 ทั้งนี้กรมสุขภาพจิตของไทยได้คัดกรองกลุ่มเสี่ยงจำนวน 12 ล้านคน ในจำนวนนี้มีแนวโน้มป่วยโรคซึมเศร้า 6 ล้านคน ได้รับการวินิจฉัยแล้วว่าเป็นโรคซึมเศร้า 5 แสนคน มีแนวโน้มที่จะฆ่าตัวตาย 6 แสนคน และคาดว่าคนไทยน่าจะมีภาวะซึมเศร้าถึงประมาณ 1.2 ล้านคน ซึ่งในประเทศไทย โรคซึมเศร้านั้นถือเป็นอีกหนึ่งปัญหาด้านสุขภาพของคนไทยที่มีความสำคัญและน่าเป็นห่วงอย่างมาก โดยสังเกตได้จากสังคมในปัจจุบันนี้ที่มักมีข่าวเกี่ยวกับปัญหาการฆ่าตัวตาย รวมทั้งปัญหาการทำร้ายร่างกายของตัวเองและคนรอบข้าง ซึ่งเหตุการณ์เหล่านี้ก็นับเป็นเรื่องที่น่าสลดใจไม่น้อยทีเดียว เพราะฉะนั้น ต้นตอสาเหตุจากการที่ผู้ป่วยประสบปัญหาชีวิต ปัญหาครอบครัว การงาน การเงินหรือพบเจอความล้มเหลวสูญเสียในชีวิตอย่างรุนแรง ทุกปัญหาเหล่านี้สามารถนำมาสู่การเกิดโรคซึมเศร้าได้หมดทั้งสิ้น

2.1.1 โรคซึมเศร้าแบ่งออกได้เป็น 2 ประเภทใหญ่

1) โรคซึมเศร้าชนิดรุนแรง (Major Depression) เป็นอาการซึมเศร้าที่ส่งผลกระทบถึงชีวิตการทำงานหรือการเรียน รวมไปถึงการนอนหลับและการกินอยู่ได้อย่างเป็นปกติสุขอย่างรุนแรง

2) โรคซึมเศร้าเรื้อรัง (Persistent Depressive Disorder) แม้จะมีอาการและความรุนแรงของอาการน้อยกว่า แต่ภาวะซึมเศร้าชนิดนี้จะคงอยู่กับผู้ป่วยยาวนานกว่ามาก เป็นเวลาอย่างน้อยตั้งแต่ 2 ปีขึ้นไป ซึ่งผู้ที่ป่วยเป็นโรคซึมเศร้าเรื้อรังก็อาจมีบางช่วงเวลาที่ต้องเผชิญภาวะซึมเศร้าชนิดรุนแรงร่วมด้วย

2.1.2 สาเหตุของโรคซึมเศร้า

สาเหตุที่จะกระตุ้นการเกิดโรคซึมเศร้าที่พบบ่อยก็คือ การมีทั้งความเสี่ยงทางพันธุกรรม ทางสภาพจิตใจ และการเผชิญกับสถานการณ์เลวร้าย ร่วมกันทั้ง 3 ปัจจัย

1) โรคซึมเศร้าเกิดจากความเครียด แต่ทั้งนี้คนที่ไม่มีญาติเคยป่วยก็อาจเกิดเป็นโรคนี้ได้ มักพบว่าผู้ป่วยโรคนี้จะมีความผิดปกติของระดับสารเคมี ที่เซลล์สมองสร้างขึ้น เพื่อรักษาสมดุลของอารมณ์

2) สภาพทางจิตใจที่เกิดจากการเลี้ยงดู ก็เป็นปัจจัยที่เสี่ยงอีกประการหนึ่งต่อการเกิดโรคซึมเศร้าเช่นกัน คนที่ขาดความภูมิใจในตนเองมองตนเองและโลกที่เขาอยู่ในแง่ลบตลอดเวลา หรือเครียด ง่ายเมื่อเจอกับมรสุมชีวิต ล้วนทำให้เขาเหล่านั้นมีโอกาสป่วยง่ายขึ้น

3) การเผชิญกับสถานการณ์เลวร้าย เช่น หากชีวิตพบกับการสูญเสียครั้งใหญ่ต้องเจ็บป่วยเรื้อรัง ความสัมพันธ์กับคนใกล้ชิดไม่ราบรื่น หรือต้องมีการเปลี่ยนแปลงในทางที่ไม่ปรารถนาก็อาจกระตุ้นให้โรคซึมเศร้ากำเริบได้

2.1.3 อาการของโรคซึมเศร้า โรคซึมเศร้ามีอาการรู้สึกเศร้าใจ หม่นหมอง หงุดหงิด หรือรู้สึกกังวลใจ ไม่สบายใจ ขาดความสนใจต่อสิ่งแวดล้อมรอบข้าง หรือสิ่งที่เคยให้ความสนุกสนานในอดีต น้ำหนักลดลง หรือเพิ่มขึ้น ความอยากอาหารเปลี่ยนแปลงไป นอนไม่หลับ หรือนอนมากเกินกว่าปกติ คนที่เป็นโรคซึมเศร้า จะรู้สึกผิด สิ้นหวัง หรือรู้สึกว่าตนเองไร้ค่า ไม่มีสมาธิ ไม่สามารถตัดสินใจเองได้ ความจำแย่ลง อ่อนเพลีย เมื่อยล้า ไม่มีเรี่ยวแรง กระวนกระวาย ไม่อยากทำกิจกรรมใด ๆ คิดถึงแต่ความตาย และอยากที่จะฆ่าตัวตาย

2.1.4 การป้องกันโรคซึมเศร้า ภาวะซึมเศร้าไม่มีวิธีการป้องกันที่แน่นอน เนื่องจากสาเหตุอาจเกิดจากโรคบางประการ เช่น ความผิดปกติในสมอง อาการเจ็บป่วย หรือการใช้ยาที่เกิดภาวะซึมเศร้าแทรกซ้อนได้ ถือว่าเป็นปัจจัยที่อยู่เหนือการควบคุม แต่การสร้างพฤติกรรมทางสุขภาพที่ดีด้วยการเลือกรับประทานอาหารและออกกำลังกาย การรักษาสภาวะอารมณ์ให้แจ่มใสด้วยการทำกิจกรรมเพื่อความสนุกสนานและผ่อนคลาย นับเป็นส่วนหนึ่งในการช่วยลดความเสี่ยงจากภาวะซึมเศร้าได้

2.2 การทำเหมืองข้อมูล (Data Mining) (ศจี วานิช, 2558: ออนไลน์)

เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า “เหมืองข้อมูล” ซึ่งเป็นคำศัพท์ที่ใช้เปรียบกับการขุดเหมืองแร่ทั่วๆไป โดยในการขุดเหมืองแร่นั้นสิ่งที่ต้องการก็คือแร่ที่มีค่า เช่น เพชร พลอย เป็นต้น ในขั้นตอนการทำเหมืองแร่นั้นจะต้องระเบิดภูเขาใหญ่หลายๆ ลูกเพื่อค้นหาแร่ที่ต้องการ ซึ่งแร่ที่พบนั้นก็ได้ออกมาน้อยมากเมื่อเทียบกับหินที่โดนระเบิดจากภูเขา เช่นเดียวกันเมื่อในองค์กรหรือบริษัทมีภูเขาของข้อมูลที่มีขนาดมหาศาล บริษัทจึงต้องการขุดค้นหาลงไปในข้อมูลเหล่านี้เพื่อให้ได้สิ่งที่มีค่าซึ่งอยู่ในข้อมูลเหล่านี้ เป็นการเปรียบเปรยให้เห็นลักษณะที่คล้ายกันระหว่างการขุดเหมืองแร่และการขุดเหมืองข้อมูลสามารถสรุปได้ว่า เหมืองข้อมูลคือการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่างๆ เป็นต้น ซึ่งข้อมูลนี้จะเก็บรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้งโดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลแล้วจะได้สิ่งที่มีประโยชน์ เช่น ลูกค้าส่วนใหญ่ที่ซื้อเบียร์มักจะซื้อผ้าอ้อมด้วย เป็นต้น จะเห็นได้ว่าข้อมูลนี้เป็นข้อมูลที่ไม่เคยคิดว่ามีความสัมพันธ์กันและไม่เคยรู้มาก่อน เมื่อได้ความรู้แบบนี้ออกมาแล้วอาจจะนำไปออกโปรโมชันหรือช่วยในการจัดวางชั้นสินค้าในซุปเปอร์มาร์เก็ตต่อไปได้

การทำเหมืองข้อมูล คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงาน

หลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่ง ในการจัดเก็บและตีความหมายข้อมูลจากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

2.2.1 ประเภทของข้อมูลที่ใช้ในการทำเหมืองข้อมูล

1) ข้อมูลที่มาจากฐานข้อมูลเชิงสัมพันธ์ (relational databases)

2) ข้อมูลจากคลังข้อมูล (data warehouses)

3) ข้อมูลจากฐานข้อมูลรายการปรับปรุง (transactional databases)

4) จากฐานข้อมูลพิเศษหรือที่เก็บข่าวสารพิเศษซึ่งได้แก่ ฐานข้อมูลเชิงวัตถุ ข้อมูลเกี่ยวกับเวลา ฐานข้อมูลข้อความ (text databases) และฐานข้อมูลมัลติมีเดีย ฐานข้อมูลแบบเก่าในอดีตหรือข้อมูลที่มาจากต่างฐานข้อมูลกัน

2.2.2 ขั้นตอนการทำเหมืองข้อมูล (Data Mining)

1) การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป

2) การรวมข้อมูล (Data Integration) เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน

3) การเลือกข้อมูล (Data Selection) เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้

4) การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน

5) การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่

6) การประเมินรูปแบบ (Pattern Evaluation) เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล

7) การนำเสนอความรู้ (Knowledge Representation) เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ

2.2.3 ประโยชน์ของเหมืองข้อมูล

1) ช่วยชี้แนวทางการตัดสินใจและคาดการณ์ผลลัพธ์ที่จะได้จากการตัดสินใจ

2) เพิ่มความเร็วในการวิเคราะห์ฐานข้อมูลขนาดใหญ่

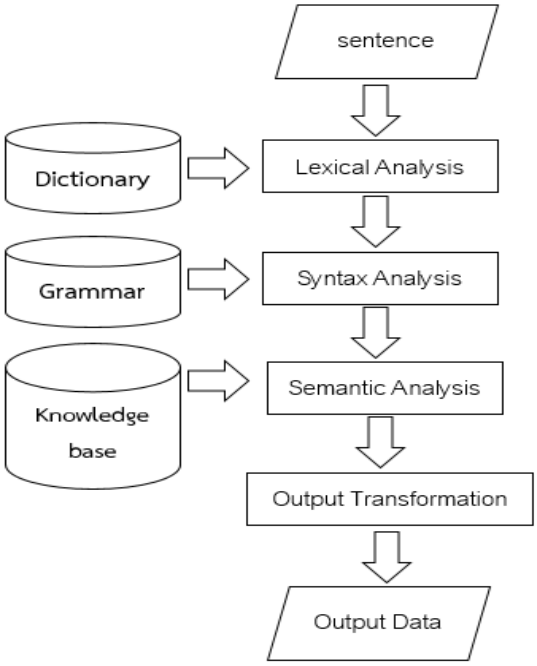
3) ค้นหาส่วนประกอบที่ซ่อนอยู่ภายในเอกสาร รวมถึงความสัมพันธ์ของส่วนประกอบต่างๆด้วย

4) การจัดกลุ่มข้อมูล เช่น จัดกลุ่มลูกค้าทั้งหมดของบริษัทประกันภัยที่ประสบอุบัติเหตุลักษณะเดียวกันเพื่อดำเนินการต่าง ๆ ตามนโยบายของบริษัท

2.3 การประมวลผลภาษาธรรมชาติ (Natural Language Processing) (The AI Midnight, 2562: ออนไลน์)

การประมวลผลภาษาธรรมชาติ เป็นการประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ คำอธิบายที่เรียบง่าย คือ ทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ รวมไปถึงการประมวลผลที่ไม่ใช่แค่ทำให้คอมพิวเตอร์เข้าใจเรา แต่รวมถึงไปการวิเคราะห์ทางด้านภาษาศาสตร์ การตีความจากข้อความ ตัวอย่างเช่น การวิเคราะห์และตอบสนองความต้องการของมนุษย์ด้วยกัน เป็นต้น จึงทำให้ Natural Language Processing มีความสำคัญอย่างมากมาย

2.3.1 การทำงานของการประมวนผลภาษาธรรมชาติ แสดงดังรูปที่ 1 มีขั้นตอนการทำงานของการประมวนผลภาษาธรรชาติดังนี้ นำเข้าประโยค (Sentence) มายังระบบคอมพิวเตอร์ วิเคราะห์คำ (Lexical Analysis) โดยใช้ข้อมูลจากพจนานุกรม (Dictionary) ของภาษา วิเคราะห์ประโยคและโครงสร้าง โดยใช้ข้อมูลไวยากรณ์ (Grammar) วิเคราะห์ความหมาย (Semantic Analysis) โดยอิงจากฐานความรู้ (Knowledge base) ที่รวบรวมไว้ จากนั้นทำการแปลงข้อมูล (Output Transformation) ที่คอมพิวเตอร์เข้าใจและแสดงผลลัพธ์ (Output Data) แก่ผู้ใช้ ดังรูปที่ 2.1



รูปที่ 2.1 การทำงานของการประมวนผลภาษาธรรมชาติ

2.4 กฎความสัมพันธ์ (Association rule) (เอกสิทธิ์ พัชรวงค์ศักดา, 2557)

แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุ ที่เกิดขึ้นพร้อมกัน ตัวอย่างของการประยุกต์ใช้กฎเชื่อมโยง เช่น การวิเคราะห์ข้อมูลการขายสินค้า โดยเก็บข้อมูลจากระบบ ณ จุดขายหรือร้านค้าออนไลน์ แล้วพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกัน เช่น ถ้าพบว่าคนที่ซื้อเทปวิดีโอมักจะซื้อเทปกาวด้วย ร้านค้าก็อาจจะจัดร้านให้สินค้าสองอย่างอยู่ใกล้กัน เพื่อเพิ่มยอดขาย หรืออาจจะพบว่า

หลังจากคนซื้อหนังสือ ก แล้ว มักจะซื้อหนังสือ ข ด้วย ก็สามารถนำความรู้นี้ไปแนะนำผู้ที่กำลังจะซื้อหนังสือ ก ได้

2.4.1 กฎความสัมพันธ์และการประยุกต์ใช้งาน การสืบค้นกฎความสัมพันธ์เป็นการวิเคราะห์พฤติกรรมการซื้อสินค้าของผู้บริโภคเพื่อค้นหาว่าสินค้าชนิดใดบ้างที่ลูกค้ามักจะซื้อไปด้วยพร้อมกัน เช่น“เมื่อลูกค้าซื้อขนมปังแล้วจะซื้อแยมด้วย” หรือ “เมื่อลูกค้าซื้อเบียร์แล้วจะซื้อผ้าอ้อมไปด้วย” เมื่อนำข้อมูลที่ได้มานี้มาเขียนให้อยู่ในรูปของกฎความสัมพันธ์จะได้เป็น

กฎความสัมพันธ์ที่ 1: ลูกค้าซื้อขนมปัง => ลูกค้าซื้อแยม
กฎความสัมพันธ์ที่ 2: ลูกค้าซื้อเบียร์ => ลูกค้าซื้อผ้าอ้อม

กฎความสัมพันธ์จะประกอบไปด้วย 2 ส่วน คือส่วนที่อยู่ด้วนซ้าย (Left Hand Side: LHS) ของเครื่องหมาย => ซึ่งเราจะเรียกว่าข้ออ้าง (premise) (Right Hand Side: RHS) คือ ข้อสรุป (conclusion) นอกจากนี้จำนวนครั้งที่ลูกค้าซื้อสินค้าจะเรียกว่า ค่าสนับสนุน (support) ดังรูปที่ 2.2

การซื้อสินค้าครั้งที่	สินค้าที่ลูกค้าซื้อ
1	Apple, Cereal, Diapers
2	Beer, Cereal, Eggs
3	Apple, Beer, Cereal, Eggs
4	Beer, Eggs

รูปที่ 2.2 แสดงข้อมูลสินค้าที่ลูกค้าซื้อในแต่ละครั้ง

จากข้อมูลในรูปที่ 2 จะพบว่ามีการซื้อเบียร์ (Beer) จำนวน 3 ครั้ง นั่นคือค่าสนับสนุน (support) ของเบียร์จึงเท่ากับ 75% (3/4) และไข่ (Eggs) เองก็ถูกซื้อไปเป็นจำนวน 3 ครั้งเช่นกัน ดังนั้นค่าสนับสนุนของไข่ก็เท่ากับ 75% (3/4) เช่นกัน

1) ค่าความมั่นใจ (Confidence) คือ ตัววัดประสิทธิภาพของกฎความสัมพันธ (Association rule) แสดงความเชื่อมั่นของกฎความสัมพันธ์ที่เมื่อรูปแบบในด้านซ้ายของกฎความสัมพันธ์ (LHS) เกิดขึ้นแล้วรูปแบบในด้านขวาของกฎความสัมพันธ์ (RHS) จะเกิดขึ้นด้วยเป็นจำนวนกี่เปอร์เซ็นต์ การคำนวณค่า confidence หาได้จาก

$$\text{Confidence}(LHS => RHS) = \frac{support\ (LHS,RHS)}{\text{support (LHS)}} \qquad (2.1)$$

โดยที่ support (LHS,RHS) คือ ค่าสนับสนุนที่รูปแบบ LHS และ RHS ของกฎความสัมพันธ์ที่เกิดขึ้นพร้อม ๆกัน และ support (LHS) คือ ค่าสนับสนุนของรูปแบบที่อยู่ด้านซ้ายของกฎความสัมพันธ์

2) ค่าสนับสนุน (Support) คือ ตัววัดประสิทธิภาพสำหรับสินคา (Item) นับจำนวนครั้งการซื้อสินค้าแต่ละชนิดแล้วคิดเป็น % ของการซื้อสินค้า หาได้จาก

$$\text{Support}\ (S) = \frac{\text{จำนวนรายการที่ซื้อสินค้า}}{\textbf{Transaction (T)}} \qquad (2.2)$$

2.4.2 เทคนิคในการหากฎความสัมพันธ์ด้วยวิธี Apriori นั้นเป็นอัลกอริทึมพื้นฐานที่ใช้ในการหาความสัมพันธ์ของข้อมูลโดยใช้หลักการค้นหาแบบวงกว้างก่อนนับทรานแซคชัน ซึ่งจะทำการสร้างและตรวจสอบเซตไอเท็มที่เกิดขึ้นบ่อยทีละชั้น โดยเริ่มจากเซ็ตไอเท็มที่มีจำนวนสมาชิกเท่ากับหนึ่งถ้าเซ็ตไอเท็มใดมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนที่กำหนดก็จะตัดเซตไอเท็มนั้นออก ไม่นำไปสร้างเซ็ตไอเท็มในชั้นต่อไป การทำงานของอัลกอริทึมจะวนไปเรื่อย ๆ จนกระทั่งไล่ทุกระดับชั้นหรือไม่เหลือเซ็ตไอเท็มในชั้นต่อไป ในการนับจำนวนทรานแซคชันอัลกอริทึม Apriori จะทำการไล่ทรานแซคชันครั้งเดียวในแต่ละระดับชั้น ในการตรวจดูว่าทรานแซคชันนั้นบรรจุเซตไอเท็มใดบ้าง เพื่อความรวดเร็วจะเก็บเซตไอเท็มในแต่ละระดับชั้นทั้งหมดไว้ในโครงสร้าง Hash Tree จุดเด่นของอัลกอริทึมนี้อยู่ที่ความสามารถในความเร็วของการค้นหาไอเท็มเซตที่ปรากฏบ่อย ด้วยการละเว้นการพิจารณาไอเท็มเซตที่ปรากฏซ้ำด้วยความถี่ที่ต่ำกว่าเกณฑ์

1) ข้อดีของ Apriori ช่วยให้ทราบพฤติกรรมของเป้าหมายได้ โดยการใช้อัลกอริทึมจัดการเชื่อมความสัมพันธ์ของเหตุการณ์ต่างๆ ที่เราต้องการหาความสัมพันธ์ของเป้าหมาย คัดกรองข้อมูลออกมาตามความสัมพันธ์ วิเคราะห์ข้อมูลมาจนมีความน่าเชื่อถือและนำไปใช้ได้จริง

2) ข้อเสียของ Apriori อัลกอริทึม Apriori ถือเป็นอัลกอริทึมที่นิยมใช้ในการหากฎความสัมพันธ์ของข้อมูล แต่ถ้าฐานข้อมูลมีการเพิ่มข้อมูลเข้ามาหรือเกิดมีการเปลี่ยนแปลงข้อมูล อัลกอริทึม Apriori จะต้องนำข้อมูลทั้งหมดมารวมกันก่อน แล้วจึงจะสามารถนำข้อมูลทั้งหมดไปค้นหากฎความสัมพันธ์ใหม่ทั้งหมด โดยไม่สามารถนำกฎความสัมพันธ์ที่หาได้จากกลุ่มข้อมูลเก่าก่อนหน้ามาใช้ให้เกิดประโยชน์ได้ ทำให้เสียเวลาในการทำงานเพื่อค้นหากฎความสัมพันธ์ใหม่ทั้งหมด

2.4.3 เทคนิค FP-Growth ใช้หลักการสร้างต้นไม้ (FP-tree) การทำซ้ำแบบ divide-and-conquer ในการสร้างกฎความสัมพันธ์ที่พบบ่อย เทคนิคนี้สามารถสร้างกฎความสัมพันธ์ได้อย่างรวดเร็ว แต่มีข้อจำกัดเกี่ยวกับข้อมูลที่ใช้ต้องอยู่ในรูปของไบนารี และการใช้หน่วยความจำจำนวนมากเมื่อต้นไม้มีขนาดใหญ่ ในขั้นตอนแรกรากของโครงสร้างต้นไม้จะถูกสร้างขึ้นโดยจัดให้เป็นค่าว่าง (null) จากนั้นข้อมูลจะถูกจัดเรียงเป็นลำดับ โดยแต่ละค่าของตัวแปรจะแทนด้วยโหนด

1) ข้อดีของ FP-Growth ชวยลดจำนวนการอ่านข้อมูลจากฐานข้อมูลสำหรับการค้นหาข้อมูลที่ปรากฏรวมกันบอยเหลือเพียง 2 ครั้ง และกระบวนการทำงานใชหลักการทำงานแบบพลวัตร (Dynamic Programming) ทำใหการทำงานมีประสิทธิภาพ เหมาะสมกับฐานขอมูลที่มีขนาดเล็กและขนาดใหญ มีจำนวนชิ้นขอมูลในฐานขอมูลนอย และลักษณะขอมูลที่เหมาะสมตองมีความหนาแนนของขอมูลสูง คืออัตราสวนของจำนวนชิ้นขอมูลที่ปรากฏอยูในรายการขอมูลมีมาก การทำงานสามารถทำงานไดดีหากกำหนดคาสนับสนุนขั้นต่ำมีคามาก ๆ เพราะจะใชเวลาในการทองไปยังแตละโหนดสำหรับการคนหากลุ่มข้อมูลที่ปรากฏรวมกันบอยไดเร็ว และลดการใชเนื้อที่ในการสรางตนไม FP-Tree สำหรับจัดเก็บขอมูล

2) ข้อเสียของขั้นตอนวิธี FP-Growth หากฐานข้อมูลที่ใช้ในการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยมีจำนวนชิ้นข้อมูลในฐานข้อมูลมากแล้ว การทำงานของขั้นตอนวิธี FP-Growth จะต้องใช้เนื้อที่หน่วยความจำเป็นจำนวนมาก เปลืองเนื้อที่ในการจัดเก็บข้อมูลในระหว่างการประมวลผลเนื่องจากต้นไม้ FP-Tree ที่สร้างขึ้นจะมีขนาดใหญ่ ซึ่งเกิดจากที่ต้องสร้างโหนดแทนชิ้นข้อมูลในฐานข้อมูลเป็นจำนวนมาก อีกทั้งใช้เวลาในการท่องไปยังโหนดที่ต้องการนาน
(ฟูไดละห์ ดือมอง, 2553)

2.5 การจำแนกประเภทข้อมูล (Classification) (Ryoma, 2556: ออนไลน์)

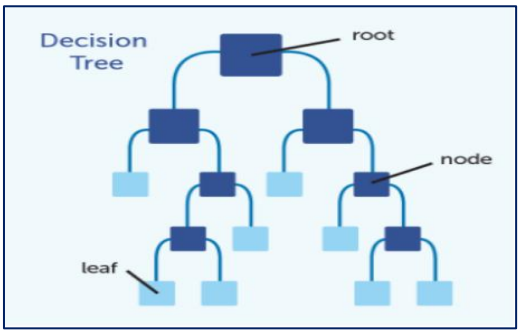
การจำแนกประเภทข้อมูลนั้นเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นความแตกต่างระหว่างคลาสหรือกลุ่มของข้อมูลได้ และเพื่อทำนายว่าข้อมูลนี้ควรจัดอยู่ในคลาสใด ซึ่งโมเดลที่ใช้จำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ จะขึ้นอยู่กับการวิเคราะห์เซตของข้อมูลทดลอง (training data) โดยนำ training data มาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในคลาสเดียวกันบ้าง ผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลจัดประเภทข้อมูล (Classifier model) โมเดลนี้ สามารถแทนได้ในหลายรูปแบบ เช่น Classification (IF-THEN) rules Decision Tree Mathematical formulae หรือ Neural networks และจะนำข้อมูลส่วนที่เหลือจาก training data เป็นข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งเป็นกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง โดยเราจะปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลนี้ได้

2.5.1 ต้นไม้ตัดสินใจ (Decision Tree) เป็นหนึ่งในเทคนิคการทำเหมืองข้อมูลในรูปแบบวิธีการจัดหมวดหมู่ที่รู้จักกันดีที่สุด โดยมักใช้ตรวจสอบข้อมูลและสร้างต้นไม้เพื่อการพยากรณ์ สำหรับโครงสร้างของต้นไม้ตัดสินใจ จะมีลักษณะคล้ายโครงสร้างต้นไม้ทั่วไป โดยการแตกแขนงไปตามเงื่อนไขหรือเส้นทางของกิ่งไม้และข้อมูลที่คาดคะเนไว้ว่าจะเกิดขึ้น ซึ่งจะใช้กฎในรูปแบบ “ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์)” (If-then Rule) มาประกอบการสร้างโครงสร้างต้นไม้ตัดสินใจ สำหรับโครงสร้างต้นไม้ตัดสินใจจะประกอบด้วย

1) โหนด (Node) คือโหนดที่แสดงถึงคุณลักษณะ (Feature) ที่นำมาใช้ในการแบ่งกลุ่มของข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งมีโหนดราก (Root Node) อยู่บนสุดของโครงสร้าง ซึ่งเป็นโหนดที่มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุด

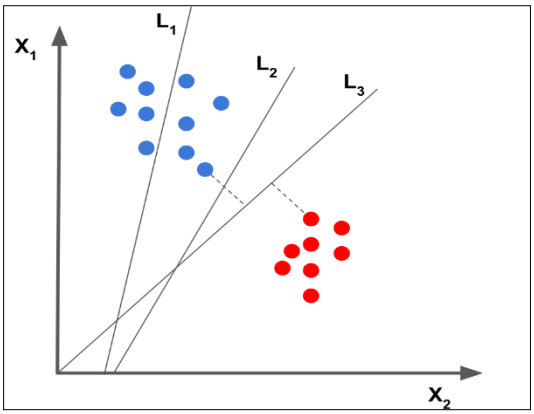
2) กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนดที่ใช้เป็นเงื่อนไขหรือทางเลือกของการกระทำซึ่งมาจากผลลัพธ์แต่ละตัวของทุกตัวทำนาย (Predictor) หรือคุณสมบัติ (Feature)

3) โหนดใบ (Leaf Node) เป็นโหนดที่แสดงผลลัพธ์ของเงื่อนไข หรือการทำตามเงื่อนไขที่เกิดขึ้น สำหรับการสร้างต้นไม้ตัดสินใจของแต่ละอัลกอริทึมนั้นจะมีลักษณะที่คล้ายกันคือ เริ่มต้นทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้ (Root Node) หลังจากนั้นจะทำการแตกกิ่งแอตทริบิวต์ออกไปเรื่อย ๆ จนสามารถแบ่งข้อมูลออกเป็นคลาสได้ชัดเจน ดังรูปที่ 2.3



รูปที่ 2.3 ต้นไม้ตัดสินใจ
[**ที่มา:** https://medium.com/@panumars124/data-mining]

2.5.2 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) เป็นอัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด ซึ่งแนวความคิดของซัพพอร์ตเวกเตอร์แมชชีน เกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในพื้นที่ จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกันโดยจะสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกันนั้น เส้นตรงใดเป็นเส้นที่ดีที่สุด ดังรูปที่ 2.4



รูปที่ 2.4 ซัพพอร์ตเวกเตอร์แมชชีน
[**ที่มา:** https://www.glurgeek.com/education/support-vector-machine/]

1) สิ่งที่ใช้พิจารณาในการแบ่งก็คือ การหาช่องว่างที่จะทำให้ทั้ง 2 กลุ่มแยกห่างออกจากกันมากที่สุด สร้างช่องว่างระหว่าง 2 กลุ่มให้มากที่สุด

2) วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบ (margin) ให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลในพื้นที่ที่ใกล้ที่สุด ดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุด และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ตเวกเตอร์” (mindphp, 2555: ออนไลน์)

2.5.3 โครงข่ายประสาทเทียม (Neural Network หรือ NN) เป็นหนึ่งในเทคนิคการทำเหมืองข้อมูลที่ใช้โมเดลทางคณิตศาสตร์หรือโมเดลทางคอมพิวเตอร์มาประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (connectionist) โดยโครงข่ายประสาทเทียมเป็นแขนงหนึ่งของปัญญาประดิษฐ์ที่มีโครงสร้างการทำงานคล้ายคลึงกับการทำงานของเซลล์สมอง หรือระบบประสาทของมนุษย์ ซึ่งพบว่าสามารถแก้ปัญหาที่มีความซับซ้อน หรือใช้ในการทำนายหรือพยากรณ์พฤติกรรมที่มีลักษณะไม่เป็นเชิงเส้น (Nonlinear) ได้ดี และในปัจจุบันนิยมนำโครงข่ายประสาทเทียมมาประยุกต์ใช้แก้ปัญหางานจริงได้อย่างหลากหลายด้าน เช่น การเงินการธนาคาร อวกาศ ระบบป้องกันประเทศ ระบบรักษาความปลอดภัย การแพทย์ ระบบสื่อสาร ระบบขนส่ง การบันเทิง ทางด้านวิศวกรรม รวมทั้งงานทางด้านการเกษตรซึ่งพบบ่อยมากขึ้น โดยนำโครงข่ายประสาทเทียมมาใช้ด้านเพื่อการจัดหมวดหมู่และแยกแยะวัสดุทางการเกษตร การพยากรณ์ผลลัพธ์ของผลผลิต การประมาณค่าความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม การควบคุมสภาวะที่มีการเปลี่ยนแปลง การจดจำรูปแบบการจัดกลุ่ม เป็นต้น

คุณสมบัติที่สำคัญโครงข่ายประสาทเทียม คือความสามารถในการเรียนรู้จากตัวอย่าง โดยการพยายามคำนวณหาความสัมพันธ์ระหว่างปัจจัยนำเข้าและผลลัพธ์ การเรียนรู้จะเริ่มจากสุ่มค่าน้ำหนักและค่าเบี่ยงเบนเริ่มต้นค่าผลลัพธ์ที่ได้จากค่าเริ่มต้นจะถูกนำมาเปรียบเทียบกับผลลัพธ์จริง ค่าที่แตกต่างจะถูกนำมาปรับค่าน้ำหนักและค่าเบี่ยงเบนโดยวิธีลองผิดลองถูก จนได้ผลลัพธ์ที่ใกล้เคียงหรือตรงกับผลลัพธ์จริง ค่าน้ำหนักและค่าเบี่ยงเบนสุดท้ายจะถูกนำมาใช้ในการพยากรณ์ผลลัพธ์ที่เกิดจากข้อมูลใหม่

2.5.4 เคเนียเรสเนเบอร์ (K-Nearest Neighbors หรือ KNN) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดเป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาสและกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด ซึ่งการนำเทคนิคของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดไปใช้นั้น เป็นการหาระยะห่างระหว่างแต่ละตัวแปรในข้อมูล จากนั้นก็คำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น อย่างเช่น ถ้าเป็นเรื่องของสี เราจะใช้อะไรวัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว ต่อจากนั้นเราต้องมีวิธีในการรวมค่าระยะห่างของตัวแปรทุกค่าที่วัดมาได้ เมื่อสามารถคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่างๆ ได้จากนั้นก็เลือกชุดของเงื่อนไขที่ใช้จัดคลาส มาเป็นฐานสำหรับการจัดคลาสในเงื่อนไขใหม่ๆ ได้แล้วเรา

จะตัดสินได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้น ควรมีขนาดใหญ่เท่าไร และอาจมีการตัดสินใจได้ด้วยว่าจะนับจำนวนจุดข้างเคียงตัวมันได้อย่างไร

1) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดมีขั้นตอนโดยสรุป ดังนี้

- กำหนดขนาดของ K (ควรกำหนดให้เป็นเลขคี่)
- คำนวณระยะห่าง (Distance) ของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลตัวอย่าง
- จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน K ที่กำหนดไว้
- พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่ม (class) ไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด
- กำหนด class ให้กับจุดที่พิจารณา (class) ที่ใกล้จุดพิจารณามากที่สุด

วิธี k-Nearest Neighbors ทำให้เราทราบประเภทข้อมูลของสิ่งของที่เรายังไม่เคยจำแนกมาก่อนได้ ในการใช้งานจริงยังมีเงื่อนไขบางอย่างที่เราต้องพิจารณาเป็นพิเศษ (wikipedia, 2562: ออนไลน์)

2.6 การแบ่งกลุ่มข้อมูล (Clustering) (coralline, 2561: ออนไลน์)

เป็น Machine Learning Model ประเภทการเรียนรู้แบบไม่มีผู้สอนหรือไม่มีต้นแบบ (target) ของผลลัพธ์ ซึ่งเป็นโมเดลที่เอาไว้ใช้การจัดกลุ่มจัดก้อนของข้อมูล ที่ไม่เคยมีการจัดกลุ่มก้อนมาก่อน ตัวอย่างเช่น การจัดกลุ่มของผลิตภัณฑ์สินค้า ซึ่งอาจมีหลายคนสงสัยว่า ทำไมต้องใช้ Clustering โมเดลทั้ง ๆที่สินค้าก็จัดหมวดหมู่ด้วยประเภทของสินค้าได้เองอยู่แล้ว ก็ต้องบอกว่า การจัดกลุ่มด้วยการใช้ Model จะจัดกลุ่มตามพฤติกรรมที่ลูกค้ามีต่อสินค้า โดยใช้ข้อมูลการสั่งซื้อสินค้า เช่น จัดกลุ่มจากความถี่ในการซื้อ จัดกลุ่มจากปริมาณของการซื้อ เป็นต้น โดยการจัดกลุ่มแบบคลัสเตอร์ที่ได้จากข้อมูลจะทำให้สามารถแบ่งกลุ่มของผลิตภัณฑ์ตามความประสงค์ของลูกค้า ซึ่งอาจจะมีจำนวนกลุ่มน้อยกว่าการจัดกลุ่มด้วยประเภทผลิตภัณฑ์

2.6.1 การหาระยะห่างระหว่างข้อมูล (distance function) เป็นการจัดข้อมูลที่มีลักษณะคล้ายกันเอาไว้เป็นกลุ่มเดียวกัน ดังนั้นเราจึงมีวิธีการวัดความคล้ายคลึงระหว่างข้อมูล ซึ่งในทางดาต้าไมน์นิงเราจะเรียกความแตกต่างระหว่างข้อมูลว่า ระยะห่างระหว่างข้อมูล (distance) โดยมีฟังก์ชันในการคำนวณหาระยะห่างระหว่างข้อมูลที่ใช้กันอย่างแพร่หลายอยู่หลายแบบ เช่น การหาระยะห่างด้วยวิธี City block หรือ Manhattan (City block/Manhattan distance) การหาระยะห่างด้วยวิธี Euclidean (Euclidean distance) และการหาระยะห่างด้วยวิธี Jaccard (Jaccard distance) โดยวิธีที่ 1 และ 2 เหมาะสำหรับข้อมูลที่มีแอตทรีบิวต์เป็นค่าตัวเลข และวิธีที่ 3 เหมาะสำหรับแอตทริบิวต์ที่มีค่าเป็นแบบนอมินอล

- การหาระยะห่างด้วยวิธี City block หรือ Manhattan เป็นวิธีการหาระยะห่างแบบพื้นฐาน โดยระยะห่าง city block เกิดจากผลต่างระหว่างแอตทริบิวต์ต่างๆ ดังสมการ

$$D_{City-block} = |x_1 - y_1| + |x_2 - y_2| + .. + |x_L - y_L| \quad (2.3)$$

โดยที่ x_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลจุดที่ 1 และ y_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลชุดที่ 2 โดยข้อมูลทั้งสองตัว (x และ y) มีจำนวนแอตทริบิวต์เท่ากับ L

- การหาระยะห่างด้วยวิธี Euclidean เป็นวิธีการหาระยะห่างที่นิยมใช้กันอย่างแพร่หลาย โดยระยะห่าง Euclidean เกิดจากรากที่สองของผลต่างระหว่างแอตทริบิวต์ต่างๆ ยกกำลังสอง ดังสมการ

$$D_{Euclideanx} = \sqrt{(x_1 - y_1)^2} + (x_2 - y_2)^2 + .. + (x_L - y_L)^2 \quad (2.4)$$

โดยที่ x_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลจุดที่ 1 และ y_1 คือ แอตทริบิวตีที่ 1 ของข้อมูลชุดที่ 2 โดยข้อมูลทั้งสองตัว (x และ y) มีจำนวนแอตทริบิวต์เท่ากับ L

- การหาระยะห่างด้วยวิธี Jaccard โดยการหาระยะห่างสองวิธีแรกที่แนะนำไปใช้กับแอตทริบิวต์ที่มีข้อมูลที่เป็นตัวเลข ถ้าแอตทริบิวต์ที่เป็นค่านอมินอล หรือ ไบนารี (binary) จะใช้วิธีการ Jaccard โดยวิธีการนี้เป็นการนับจำนวนค่าที่เหมือนกันในแต่ละแอตทริบิวต์แล้วหารด้วยจำนวนค่าของแอตทริบิวต์ทั้งหมด ดังสมการ

$$D_{Jaccard} = 1\frac{|X \cap Y|}{\mathrm{X} \cap \mathrm{Y}} \quad (2.5)$$

โดยที่ X คือข้อมูลจุดที่ 1 ซึ่งประกอบด้วยแอตทริบิวต์ $x_1, x_2, \ldots x_L$ และ Y คือข้อมูลจุดที่ 2 ซึ่งประกอบด้วยแอตทริบิวต์ $y_1, y_2, \ldots y_L$ โดยข้อมูลทั้งสองตัว (X และ Y) มีจำนวนแอตทริบิวต์เท่ากับ L

2.6.2 เทคนิคในการแบ่งกลุ่มข้อมูลด้วยวิธี K-Means เป็นเทคนิคที่นิยมใช้ในการแบ่งกลุ่มข้อมูลซึ่งจะแบ่งกลุ่มออกเป็นจำนวน K คลัสเตอร์ตามที่ผู้ใช้งานกำหนดขึ้นมา โดยใช้การวัดระยะห่างระหว่างข้อมูลแต่ละตัวกับจุดศูนย์กลาง (centroid) ของแต่ละคลัสเตอร์ ซึ่งจริง ๆแล้วจุดศูนย์กลางของแต่ละคลัสเตอร์คือ ค่าเฉลี่ยของแต่ละแอตทริบิวต์ของข้อมูลที่อยู่ในคลัสเตอร์นั้นเอง

1) สรุปขั้นตอนการแบ่งกลุ่มด้วยวิธี K-Means เป็น 4 ขั้นตอนดังนี้

- กำหนดจุดศูนย์กลางของแต่ละคลัสเตอร์โดยทำการสุ่มให้มีตามจำนวนที่กำหมด
- คำนวณระยะห่างระหว่างข้อมูลแต่ละตัวกับจุดศูนย์กลางของแต่ละคลัสเตอร์ที่ได้จากขั้นตอนก่อนหน้า และกำหนดให้อยู่ในคลัสเตอร์ที่ใกล้กับจุดศูนย์กลางของคลัสเตอร์นั้น
- คำนวณหาจุดศูนย์กลางของแต่ละคลัสเตอร์ใหม่อีกครั้ง
- ทำขั้นตอนที่ 2-3 จนกว่าข้อมูลทุกตัวอยู่ในคลัสเตอร์เดิมหรือทำงานจนถึงจำนวนรอบที่กำหนดไว้

2.6.3 เทคนิคในการแบ่งกลุ่มข้อมูลด้วยวิธี Agglomerative Clustering เป็นการแบ่งกลุ่มแบบ Hierarchical Clustering หรือการแบ่งกลุ่มแบบเป็นระดับชั้น โดยจะทำการรวมข้อมูลสองตัวที่มีระยะห่างใกล้กันมากที่สุดเป็นหนึ่งคลัสเตอร์ก่อน หลังจากนั้นจึงค่อยรวมข้อมูลคู่อื่นตามมา ในการวัดระยะห่างระหว่างข้อมูลสองตัวจะใช้วิธีวัดแบบ Euclidean แต่ถ้าต้องการวัดระยะห่างระหว่างข้อมูลแต่ละตัวกับข้อมูลที่จัดอยู่ในคลัสเตอร์แล้วหรือ วัดระยะห่างระหว่างสองคลัสเตอร์จะมีวิธีการวัดอยู่ 3 แบบ คือ

1) Single Link เป็นการวัดระยะห่างเทียบกับข้อมูลที่อยู่ใกล้สุดในคลัสเตอร์

2) Complete Link เป็นการวัดระยะห่างเทียบกับข้อมูลที่อยู่ไกลสุดในคลัสเตอร์

3) Average Link เป็นการวัดระยะห่างเทียบกับข้อมูลทุกจุดที่อยู่ในคลัสเตอร์ แล้วจึงหาค่าเฉลี่ย

2.7 โปรแกรมแรบพิทไมเนอร์ (Rapid Miner) (เอกสิทธิ์ พัชรวงค์ศักดา, 2557)

เป็นโปรแกรมที่นิยมใช้ในปัจจุบัน ใช้ในเชิงของการวิเคราะห์ข้อมูลและเชิงของวิทยาศาสตร์ข้อมูล (Data science) นิยมนำ Rapid Miner มาใช้ในการวิเคราะห์ข้อมูลเยอะมากในปัจจุบัน ซึ่ง Rapid Miner เป็นโปรแกรมที่สามารถนำเข้าข้อมูลได้หลายลักษณะ เช่น การเชื่อมโยงจากฐานข้อมูลโดยตรง ไฟล์ Excel ไฟล์ CSV การเขียนไฟล์ให้อยู่ในรูปแบบของ Excel และ CSV หรือการแสดงข้อมูลในกราฟแบบต่างๆ เช่น scatter plot time series และสามารถจัดการข้อมูลได้ตั้งแต่การเตรียมข้อมูล (Data Preparation) สร้างโมเดล (Model & Validate) ไปจนถึงนำไปใช้งานใน production (Operationalize) ซึ่งจะแยกส่วนได้ดังนี้

2.7.1 Rapid Miner Radoop เป็นเวอร์ชันที่ทำงานบน Hadoop (ซึ่งเป็นการนำคอมพิวเตอร์หลายๆ เครื่องมาช่วยประมวลผล) ทำให้สามารถรองรับการทำงานกับข้อมูลที่มีขนาดใหญ่ๆ หลาย (ร้อย) ล้านเรคคอร์ดได้

2.7.2 Rapid Miner Studio เป็นเวอรชันที่ทำงานบนเครื่องคอมพิวเตอร์ของเราเอง (อาจจะเป็น PC หรือ Notebook ก็ได้) เป็นตัวหลักในการออกแบบโพรเซส (process) หรือ workflow เพื่อใช้ในการวิเคราะห์ข้อมูลต่างๆ เช่น สร้างโพรเซสในการคาดการณ์ว่าลูกค้าคนใดจะยกเลิกการใช้บริการ (churn) ด้วยโมเดล Decision Tree

2.7.3 Rapid Miner Server เป็นเวอร์ชันที่ทำงานบนเครื่องคอมพิวเตอร์และรองรับการทำงานที่มีผู้ใช้งานหลายๆ คนพร้อมกัน โดยเวอร์ชันนี้สามารถสร้างกราฟในลักษณะของ BI (Business Intelligence) ตั้งเวลาให้ทำงาน (scheduler) และสร้าง web service เพื่อให้โปรแกรมต่างๆ มาติดต่อได้ด้วย

2.7.4 ฟังก์ชันหลักๆของ Rapid Miner Go มีดังนี้

1) ทำงานผ่าน web browser ได้เลย โดยไม่ต้องติดตั้งโปรแกรม

2) upload ข้อมูลขึ้นไปสร้างโมเดลได้ง่ายๆ

3) สร้างโมเดล classification แบบต่างๆ ได้ เช่น Decision Tree, NaiveBayes, Deep Learning, Random Forest, Gradient Boosted Tree (GBT), Support Vector Machines (SVM)

4) หาคำน้ำหนักหรือความสำคัญของแอตทริบิวต์ (หรือ feature ต่างๆ ได้)

5) สามารถแสดงผลในรูปแบบของ GUI

6) สร้าง web service สำหรับการทำ deploy ได้แบบอัตโนมัติ

2.8 โปรแกรมวิชวลสตูดิโอโค้ด (visual studio code) (Mindphp, 2560: ออนไลน์)

Visual Studio Code หรือ VSCode เป็นโปรแกรม Code Editor ที่ใช้ในการแก้ไขและปรับแต่งโค้ด จากค่ายไมโครซอฟท์ มีการพัฒนาออกมาในรูปแบบของ Open Source จึงสามารถนำมาใช้งานได้แบบฟรีๆ ที่ต้องการความเป็นมืออาชีพ

ซึ่ง Visual Studio Code นั้น เหมาะสำหรับนักพัฒนาโปรแกรมที่ต้องการใช้งานข้ามแพลตฟอร์ม รองรับการใช้งานทั้งบน Windows, macOS และ Linux สนับสนุนทั้งภาษา JavaScript, TypeScript และ Node.js สามารถเชื่อมต่อกับ Git ได้ นำมาใช้งานได้ง่ายไม่ซับซ้อน มีเครื่องมือส่วนขยายต่าง ๆ ให้เลือกใช้อย่างมากมาก ไม่ว่าจะเป็น 1.การเปิดใช้งานภาษาอื่น ๆ ทั้งภาษา C++, C#, Java, Python, PHP หรือ Go 2.Themes 3.Debugger 4.Commands เป็นต้น (“mindphp.com”, 2560: ออนไลน์)

2.9 ไมโครซอฟเอ็กเซล (Microsoft Excel) (ปานระพี รพิพันธุ์, 2561: ออนไลน์)

คือ โปรแกรมตัวนึงในชุดโปรแกรม Microsoft Office ซึ่ง Excel นั้นเป็นโปรแกรมยอดฮิต มีความสามารถรอบด้าน แต่เก่งมากด้านการวิเคราะห์ คำนวณ และการจัดการข้อมูลในรูปแบบตารางที่เรียกว่า Spreadsheet รวมถึงนำข้อมูลในตารางมาแสดงผลในรูปแบบที่ทำให้เราเข้าใจข้อมูลนั้นลึกซึ้งมากยิ่งขึ้น เช่น สร้างกราฟ หรือจะตารางที่ให้เราลองเปลี่ยนมุมมองไปมาได้อย่างง่ายดายก็ยังได้ และ Microsoft Excel ยังสามารถทำเป็นไฟล์ CSV ได้อีกด้วย ซึ่ง CSV นั้นย่อมาจาก Comma Separated Value เป็นไฟล์ข้อความประเภทหนึ่งที่ใช้สำหรับเก็บข้อมูลในรูปแบบตาราง ใช้เครื่องหมายจุลภาค หรือคอมม่า (,) ในการแบ่งแต่ละคอลัมน์ โดยปกติเราสามารถบันทึกไฟล์จาก Microsoft Excel ออกมาเป็น CSV ไฟล์ได้โดยตรง หรือ อาจได้ไฟล์ CSV จากการ export ไฟล์จากระบบฐานข้อมูลอื่น ๆ โดยปกติ สำหรับผู้ใช้งานทั่วไป มักจะใช้โปรแกรม Microsoft Excel ในการเปิด

เพื่อให้แสดงผลในรูปแบบตาราง และทำให้ดูและอ่านออกได้ง่าย และสะดวกมากขึ้น แต่เรายังสามารถใช้โปรแกรมอื่น ๆ เปิดไฟล์ CSV ได้อีกด้วย เช่น Notepad , Edit plus , Word และ Rapid Miner

2.9.1 จุดเด่นของไฟล์ CSV

1) รองรับการใช้งานกับโปรแกรมฐานข้อมูลต่างๆ รวมทั้ง Microsoft Excel
2) ไฟล์ที่ได้มีขนาดเล็กมาก
3) รองรับการเปิดไฟล์ด้วยโปรแกรม Text Editor รวมทั้ง Microsoft Word

2.10 เว็บแอปพลิเคชัน (Web Application) (Mdsoft, 2562: ออนไลน์)

คือ แอปพลิเคชัน (Application) ที่ถูกเขียนขึ้นมาเพื่อเป็น Browser (เบราว์เซอร์) สำหรับการใช้งาน Webpage (เว็บเพจ) ต่างๆ ซึ่งถูกปรับแต่งให้แสดงผลแต่ส่วนที่จำเป็น เพื่อเป็นการลดทรัพยากรในการประมวลผล ของตัวเครื่องสมาร์ทโฟน หรือ แท็บเล็ต ทำให้โหลดหน้าเว็บไซต์ได้เร็วขึ้น อีกทั้งผู้ใช้งานยังสามารถใช้งานผ่าน Internet (อินเทอร์เน็ต)และ Intranet (อินทราเน็ต) ในความเร็วต่ำได้

2.10.1 ข้อดีของ เว็บแอปพลิเคชัน (Web Application) นั้น คือ ในส่วนของการใช้งานที่สามารถใช้งานได้ง่าย สะดวกทุกที่ ทุกเวลา ถ้าหากไม่มีเครื่องคอมพิวเตอร์ แต่ต้องการใช้ Web browser (เว็บเบราว์เซอร์) ก็สามารถใช้แอพพลิเคชั่นประเภทนี้ได้ รวมถึงมีการอัพเดท แก้ไขข้อผิดพลาดต่างๆ อยู่ตลอดเวลา และใช้งานได้ทุกแพลตฟอร์ม

2.11 เฟรมเวิร์ค บูสแตก (Framework Bootstrap) (Softmelt, 2562: ออนไลน์)

Bootstrap เป็น Front-end Framework ที่ช่วยให้เราสามารถสร้างเว็บแอพลิเคชันได้อย่างรวดเร็ว และ สวยงาม ตัว Bootstrap เองมีทั้ง CSS Component และ JavaScript Plugin ให้เราได้เรียกใช้งานได้อย่างหลากหลาย ตัว Bootstrap ถูกออกแบบมาให้รองรับการทำงานแบบ Responsive Web ซึ่งทำให้เราเขียนเว็บแค่ครั้งเดียวสามารถนำไปรันผ่านเบราว์เซอร์ได้ทั้งบน มือถือ แท็บเล็ต และพีซีทั่วไป โดยที่ไม่ต้องเขียนใหม่

Bootstrap ถูกพัฒนาขึ้นด้วยกลุ่มนักพัฒนาจากทั่วทุกหนแห่งในโลก มีการอัปเดทอยู่ตลอดเวลา เพื่อรองรับการทำงานได้อย่างทันสมัย และ การแก้ไขปัญหาต่างๆ หรือ Bug ก็ทำได้เร็ว ดังนั้น ผู้เขียนเอง จึงได้เลือกที่จะใช้ Bootstrap ในการนำมาช่วยพัฒนาโปรเจค ทั้งเว็บแอพลิเคชัน App บนมือถือ

Bootstrap เป็นเครื่องมือที่ช่วยให้เราสามารถพัฒนาเว็บแอพลิเคชันได้อย่างรวดเร็วและดูสวยงาม UI (User Interface) นั้นถูกออกแบบมาเพื่อให้ทันสมัยตลอดเวลา สามารถนำไปใช้ได้กับเว็บที่ทั่วไป และ เว็บสำหรับมือถือ (โดยใช้ Responsive utilities) ในการเรียนรู้ Bootstrap นั้นง่ายมาก เราไม่จำเป็นต้องเก่ง CSS ก็สามารถสร้างเว็บที่สวยงามได้ ไม่ว่าจะเป็นปุ่ม (Buttons) สีต่างๆ ฟอร์ม คอนโทรลต่างๆ, ตาราง, ไอคอน, เมนูบาร์, Dropdown, เมนู, หน้าต่าง Popup (Modal) และ อีกหลายๆ รายการที่พร้อมให้เราเลือกใช้งาน ซึ่งจะได้อธิบายในหัวข้อต่อ ๆ ไป การใช้งาน

2.12 ภาษาพีเฮชพี (Personal Home Page, PHP) (wikibooks, 2562: ออนไลน์)

พีเอชพี (PHP) ย่อมาจาก (PHP Hypertext Preprocessor) PHP คือ ภาษาคอมพิวเตอร์จำพวก scripting language ภาษาจำพวกนี้คำสั่งต่างๆจะเก็บอยู่ในไฟล์ที่เรียกว่า script และเวลาใช้งานต้องอาศัยตัวแปรชุดคำสั่ง ตัวอย่างของภาษาสคริปก็เช่น JavaScript , Perl เป็นต้น ลักษณะของ PHP ที่แตกต่างจากภาษาสคริปต์แบบอื่น ๆ คือ PHP ได้รับการพัฒนาและออกแบบมา เพื่อใช้งานในการสร้างเอกสารแบบ HTML โดยสามารถสอดแทรกหรือแก้ไขเนื้อหาได้โดยอัตโนมัติ ดังนั้นจึงกล่าวว่า PHP เป็นภาษาที่เรียกว่า server-side หรือ HTML-embedded scripting language นั้นคือในทุก ๆ ครั้งก่อนที่เครื่องคอมพิวเตอร์ซึ่งให้บริการเป็น Web server จะส่งหน้าเว็บเพจที่เขียนด้วย PHP ให้เรา มันจะทำการประมวลผลตามคำสั่งที่มีอยู่ให้เสร็จเสียก่อน แล้วจึงค่อยส่งผลลัพธ์ที่ได้ให้เรา ผลลัพธ์ที่ได้นั้นก็คือเว็บเพจที่เราเห็นนั่นเอง ถือได้ว่า PHP เป็นเครื่องมือที่สำคัญชนิดหนึ่งที่ช่วยให้เราสามารถสร้าง Dynamic Web pages (เว็บเพจที่มีการโต้ตอบกับผู้ใช้) ได้อย่างมีประสิทธิภาพและมีลูกเล่นมากขึ้น PHP เป็นผลงานที่เติบโตมาจากกลุ่มของนักพัฒนาในเชิงเปิดเผยรหัสต้นฉบับ หรือ Open Source ดังนั้น PHP จึงมีการพัฒนาไปอย่างรวดเร็ว และแพร่หลายโดยเฉพาะอย่างยิ่งเมื่อใช้

ร่วมกับ Apache Web server ระบบปฏิบัติอย่างเช่น Linux หรือ FreeBSD เป็นต้น ซึ่งในปัจจุบัน PHP สามารถใช้ร่วมกับ Web Server หลายๆตัวบนระบบปฏิบัติการ

2.13 ภาษาเคสคาดดิง สไตล์ ชีทส์ (Cascading Style Sheets, CSS) (Kipakapron, 2561: ออนไลน์)

คือ ภาษาที่ใช้สำหรับตกแต่งเอกสาร HTML/XHTML ให้มีหน้าตา สีสัน ระยะห่าง พื้นหลัง เส้นขอบและอื่น ๆ ตามที่ต้องการ CSS ย่อมาจาก Cascading Style Sheets มีลักษณะเป็นภาษาที่มีรูปแบบในการเขียน Syntax แบบเฉพาะและได้ถูกกำหนดมาตรฐานโดย W3C เป็นภาษาหนึ่งในการตกแต่งเว็บไซต์ ได้รับความนิยมอย่างแพร่หลาย

1.13.1 ประโยชน์ของ CSS

1) ช่วยให้เนื้อหาภายในเอกสาร HTML มีความเข้าใจได้ง่ายขึ้นและในการแก้ไขเอกสารก็สามารถทำได้ง่ายกว่าเดิม เพราะการใช้ CSS จะช่วยลดการใช้ภาษา HTML ลงได้ในระดับหนึ่ง และแยกระหว่างเนื้อหากับรูปแบบในการแสดงผลได้อย่างชัดเจน

2) ทำให้สามารถดาวน์โหลดไฟล์ได้เร็ว เนื่องจาก code ในเอกสาร HTML ลดลง จึงทำให้ไฟล์มีขนาดเล็กลง

3) สามารถกำหนดรูปแบบการแสดงผลจากคำสั่ง style sheet ชุดเดียวกัน ให้มีการแสดงผลในเอกสารแบบเดียวทั้งหน้าหรือในทุก ๆ หน้าได้ ช่วยลดเวลาในการปรับปรุงและทำให้การสร้างเอกสารบนเว็บมีความรวดเร็วยิ่งขึ้น นอกจากนี้ยังสามารถควบคุมการแสดงผล ให้คล้ายหรือเหมือนกันได้ในหลาย Web Browser

4) ช่วยในการกำหนดการแสดงผลในรูปแบบที่มีความเหมาะกับสื่อต่างๆ ได้เป็นอย่างดี

5) ทำให้เว็บไซต์มีความเป็นมาตรฐานมากขึ้นและมีความทันสมัย สามารถรองรับการใช้งานในอนาคตได้ดี

2.14 ภาษาเฮชทีเอ็มแอล (Hypertext Markup Language, HTML) (mindphp, 2560: ออนไลน์)

คือ ภาษาหลักที่ใช้ในการเขียนเว็บเพจ โดยใช้ Tag ในการกำหนดการแสดงผล การสร้างเว็บเพจ โดยใช้ภาษา HTML สามารถทำโดยใช้โปรแกรม Text Editor ต่างๆ เช่น Notepad, Edit Plus หรือจะอาศัยโปรแกรมที่เป็นเครื่องมือช่วยสร้างเว็บเพจ เช่น Microsoft FrontPage, Dream Weaver ซึ่งอำนวยความสะดวกในการสร้างหน้า HTML มีข้อเสียคือ โปรแกรมเหล่านี้มัก generate code ที่เกินความจำเป็นมากเกินไป ทำให้ไฟล์ HTML มีขนาดใหญ่ และแสดงผลช้า ดังนั้นหากเรามีความเข้าใจภาษา HTML จะเป็นประโยชน์ให้เราสามารถแก้ไข code ของเว็บเพจได้ตามความต้องการ และยังสามารถนำ script มาแทรก ตัดต่อ สร้างลูกเล่นสีสันให้กับเว็บเพจของเราได้

2.15 ภาษาจาวาสคริปต์ (JavaScript) (mindphp, 2560: ออนไลน์)

คือ ภาษาคอมพิวเตอร์สำหรับการเขียนโปรแกรมบนระบบอินเทอร์เน็ต ที่กำลังได้รับความนิยมอย่างสูง Java JavaScript เป็น ภาษาสคริปต์เชิงวัตถุ (ที่เรียกกันว่า "สคริปต์" (script) ซึ่งในการสร้างและพัฒนาเว็บไซต์ (ใช่ร่วมกับ HTML) เพื่อให้เว็บไซต์ของเราดูมีการเคลื่อนไหว สามารถตอบสนองผู้ใช้งานได้มากขึ้น ซึ่งมีวิธีการทำงานในลักษณะ "แปลความและดำเนินงานไปทีละคำสั่ง" (interpret) หรือเรียกว่า อ็อบเจ็กโอเรียลเต็ด (Object Oriented Programming) ที่มีเป้าหมายในการ ออกแบบและพัฒนาโปรแกรมในระบบอินเทอร์เน็ต สำหรับผู้เขียนด้วยภาษา HTML สามารถทำงานข้ามแพลตฟอร์มได้ โดยทำงานร่วมกับ ภาษา HTML และภาษา Java ได้ทั้งทางฝั่งไคลเอนต์ (Client) และทางฝั่งเซิร์ฟเวอร์ (Server)

2.16 งานวิจัยที่เกี่ยวข้อง

รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย (2562: 11) การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล วัตถุประสงค์ของงานวิจัยนี้คือการสร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูลและทำการเปรียบเทียมประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท ได้แก่ เนอีฟเบย์, เคเนียเรสเนเบอร์, ต้นไม้ตัดสินใจ และซับพอร์ทเวกเตอร์แมชชีน จากการทดลองพบว่าซัพพอร์ทเวกเตอร์แมชชีนมีประสิทธิภาพการทำนายสูงสุด คิดเป็น 76.95 สามารถนำผลที่ได้จากงานวิจัยนี้ไปประยุกต์ใช้ในการคัดกรองและสร้างระบบสนับสนุนการตัดสินใจในส่วนของแนวทางการรักษาของแพทย์ต่อไป

จารี ทองคำ, วาทินี สุขมาก และภีมพศ สุขมาก (2561) การเปรียบเทียบประสิทธิภาพของเทคนิค Apriori และ FP-Growth ในการสร้างกฎความสัมพันธ์ของโรคมะเร็งต่อมลูกหมาก ปัจจุบันอัตราการเกิดมะเร็งต่อมลูกหมากมีเพิ่มมากขึ้น ดังนั้นการทราบระยะเวลาของการรอดชีวิตขอลผู้ป่วยโรคมะเร็งต่อมลูกหมากจึงมีความสำคัญสำหรับแพทย์และผู้ป่วยเป็นอย่างยิ่ง เนื่องจากแพทย์สามารถนำมาวางแนวทางการรักษาผู้ป่วยได้ถูกต้องและเกิดประโยชน์สูงสุดต่อผู้ป่วยแต่ละราย การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิค Apriori และ FP-Growth ในการสร้างกฎความสัมพันธ์ของโรคมะเร็งต่อมลูกหมาก รวบรวมข้อมูลจากฐานข้อมูล SEER ระหว่างเดือนมกราคม พ.ศ. 2547 ถึง พ.ศ. 2557 จำนวน 2,308 ระเบียน ข้อมูลทั้งหมดได้ถูกนำมาสร้างกฎความสัมพันธ์ด้วยเทคนิค Apriori และเทคนิค FP-Growth ผลการศึกษาพบว่า เทคนิค FP-Growth มีความสามารถในการสร้างกฎความสัมพันธ์ได้มากกว่าเทคนิค Apriori และค่าความเชื่อมั่นของกฎความสัมพันธ์จากเทคนิค FP-Growth สูงกว่าเทคนิค Apriori ในช่วงสนุบสนุนระหว่าง 80-84.9% ค่าความเชื่อมั่นที่ 96.00%

ทวีศักดิ์ คงตุก (2560) การเปรียบเทียบประสิทธิภาพอัลกอริทึมสำหรับค้นหาไอเท็มเซตที่ปรากฏร่วมกันบ่อย การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษา และเปรียบเทียบอัลกอริทึมสำหรับค้นหาไอเท็มเซตที่ปรากฏร่วมกันบ่อย (Frequent Item sets) ซึ่งเป็นขั้นตอนหนึ่งในกระบวนการทำเหมืองข้อมูลกฎความสัมพันธ์ (Association Rule Mining) 2) ศึกษาชุดข้อมูลที่ใช้สำหรับการทดสอบอัลกอริทึมสำหรับค้นหาไอเท็มเซตที่ปรากฏร่วมกันบ่อย และ 3) สรุปผลได้ว่าอัลกอริทึมใด เหมาะกับ

ชุดข้อมูลลักษณะใด และอัลกอริทึมใดทำงานได้เร็วที่สุดหรือใช้หน่วยความจำน้อยที่สุด ผลการวิจัยพบว่า 1) อัลกอริทึมสำหรับค้นหาไอเท็มเซตที่ปรากฏร่วมกันบ่อย มีข้อดีและข้อเสียแตกต่างกันไป ดังนั้นแต่ละอัลกอริทึมจะเหมาะสำหรับการใช้วิเคราะห์ชุดข้อมูลที่แตกต่างกัน 2) อัลกอริทึมที่ทำงานได้เร็วที่สุดสำหรับชุดข้อมูลขนาดใหญ่และมีความหนาแน่นมาก คืออัลกอริทึม FP-Growth, Apriori และ PrePost+ 3) อัลกอริทึมที่ทำงานได้เร็วที่สุดสำหรับชุดข้อมูลขนาดใหญ่และมีความหนาแน่นน้อย คืออัลกอริทึม LCMFreq 4) อัลกอริทึมที่ทำงานได้เร็วที่สุดสำหรับชุดข้อมูลขนาดเล็กและมีความหนาแน่นน้อย คืออัลกอริทึม LCMFreq และ 5) อัลกอริทึมที่ทำงานได้เร็วที่สุดสำหรับชุดข้อมูลขนาดเล็กและมีความหนาแน่นมาก คืออัลกอริทึม PrePost+,LCMFreq

พรพิมล ชัยวุฒิศักดิ์ และยุวดี กล่อมวิเศษ (2562: 43) การพัฒนากฎการทำนายผลการเรียนของนักศึกษาชั้นปีที่ 1 โดยใช้เทคนิคการทำเหมืองข้อมูล งานวิจัยนี้มีวัตถุประสงค์เพื่อนำความรู้การทำเหมืองข้อมูลมาวิเคราะห์ผลการเรียนของนักศึกษาในรายวิชาต่างๆ ของแผนการศึกษาชั้นปีที่ 1 ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง และสร้างกฎความสัมพันธ์ระหว่างผลการเรียนของรายวิชาและเกรดเฉลี่ยสะสมต่ำกว่า 2.00 โดยใช้กฎความสัมพันธ์(Association Rules)ด้วยอัลกอริทึมอพริโอริ (Apriori algorithm) และกฎการตัดสินใจสำหรับจำแนกข้อมูล (Data Classification) ด้วยเทคนิค J48 เพื่อจะได้นำมาวางแผนการเรียนของนักศึกษา จากการศึกษาพบว่ากฎที่ใช้ในการจำแนกผลการเรียนของนักศึกษาชั้นปีที่ 1 กลุ่มที่เกรดเฉลี่ยสะสมต่ำกว่า 2.00 และ กลุ่มที่ได้เกรดเฉลี่ยสูงกว่า 2.00 ด้วยเทคนิค J48 ให้ค่าความถูกต้องสูงถึง 91% และจำนวนกฎความสัมพันธ์ของรายวิชาที่มีผลต่อเกรดเฉลี่ยสะสมต่ำกว่า 2.00 ของนักศึกษาชั้นปีที่ 1 มีจำนวนเท่ากับ 5 ด้วยความเชื่อมั่นที่ 1.00 และค่าสหสัมพันธ์มากกว่า 1.00

ณัฐวดี หงษ์บุญมี และประภาสิริ ตรีพาณิชกุล (2562: 41) การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูล งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกข้อมูลด้วยอัลกอริธึมเหมืองข้อมูล 3 แบบคือ โครงข่ายประสาทเทียม การเรียนรู้แบบเบย์และต้นไม้ตัดสินใจเพื่อให้ได้อัลกอลิทึมที่มีประสิทธิภาพสูงสุดที่จะถูกนำมาวิเคราะห์หาปัจจัยที่ส่งผลต่อความเสี่ยงการเกิดโรคไฮเปอร์ไทรอยด์โดยการลดการนำเข้าทีละปัจจัย ซึ่งข้อมูลที่นำมาใช้ในการทดลองเป็นข้อมูลจากโรงพยาบาลในจังหวัดพิษณุโลกจำนวน 323 ชุดข้อมูล ข้อมูลสำหรับการวิเคราะห์มีจำนวน 12 ปัจจัย ผลการเปรียบเทียบพบว่าการจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียมให้ค่าประสิทธิภาพสูงสุดโดยมีค่าความถูกต้อง 82.97% ซึ่งมากกว่าต้นไม้ตัดสินใจและการเรียนรู้แบบเบย์ที่มีค่าประสิทธิภาพความถูกต้อง 79.87%และ 68.11% ตามลำดับ ผลการค้นหาปัจจัยที่ส่งผลต่อความเสี่ยงโรคไฮเปอร์ไทรอยด์ พบว่าปัจจัยลักษณะอาการที่มีความสำคัญคือ อารมณ์แปรปรวนและเหนื่อยง่าย ส่วนปัจจัยส่วนบุคคลที่มีความสำคัญคือเพศ นอกจากการค้นหาปัจจัยแล้วงานวิจัยนี้ยังสามารถนำแบบจำลองการจำแนกข้อมูลที่ได้มาพัฒนาระบบการพยากรณ์ความเสี่ยงโรคไฮเปอร์ไทรอยด์บนสมาร์ทโฟน เพื่อช่วยสนับสนุนการตัดสินใจในส่วนของการวิเคราะห์ความ

เสี่ยงโรคไฮเปอร์ไทรอยด์ช่วยคัดกรองด้วยตัวเองเบื้องต้นและสามารถแนะแนวทางการรักษาของแพทย์และผู้ป่วยได้ต่อไป

2.17 สรุปทฤษฎีและงานวิจัยที่เกี่ยวข้อง

จากการศึกษา ค้นคว้า รวบรวมทฤษฎีและวิเคราะห์งานวิจัยที่เกี่ยวข้องพบว่า การใช้เทคนิคการทำเหมืองข้อมูลนั้นสามารถเข้ามาช่วยในการจัดเก็บข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ได้จนถึงสามารถค้นพบองค์ความรู้ที่ซ่อนอยู่ในข้อมูลนั้น ๆ และการทำเหมืองข้อมูลยังสามารถใช้อัลกอริทึมและกฎความสัมพันธ์ต่างๆเพื่อคาดการณ์ความแม่นยำของข้อมูลได้เป็นอย่างดี โดยการนำโปรแกรมแรบพิทไมเนอร์เข้ามาช่วยจัดการในข้อมูนั้น ๆ เป็นการใช้กฎความสัมพันธ์เพื่อที่จะทำให้โมเดลมีความถูกต้องและแม่นยำที่สุด และการทำเว็บแอพพลิเคชั่น (Web Application) ก็เป็นการแสดงผลในส่วนที่จำเป็น และเป็นการลดทรัพยากรในการประมวลผลทำให้โหลดหน้าเว็บไซต์ได้เร็วขึ้น โดยผู้ใช้ยังสามารถใช้งานได้ง่าย สะดวกทุกที่ ทุกเวลา หากไม่มีเครื่องคอมพิวเตอร์ใช้เบราว์เซอร์ก็สามารถเข้าใช้งานได้