บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง รวมถึงเอกสารและงานเขียนอื่น ๆ ที่เกี่ยวข้องกับงานวิจัยโดยในบทนี้จะแบ่งเนื้อหาหลัก ๆ ออกเป็น 2 หัวข้อ ประกอบด้วย

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การทำเหมืองข้อมูล (Data Mining)

2.1.1.1 การทำเหมืองข้อมูล จำแนกออกเป็น 2 ประเภท

2.1.2 ต้นไม้ตัดสินใจ (Decision Tree) C4.5

2.1.3 Feature selection การคัดเลือกคุณสมบัติ

2.1.3.1 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection

2.1.4 ตัววัดประสิทธิภาพของโมเดล

2.1.4.1 การวัดประสิทธิภาพโมเดลด้วย Cross Validation

2.1.5 กระบวนการวิเคราะห์ข้อมูล CRISP-DM (Cross-Industry Standard Process For Data Mining)

2.1.6 โปรแกรม RapidMiner Studio

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

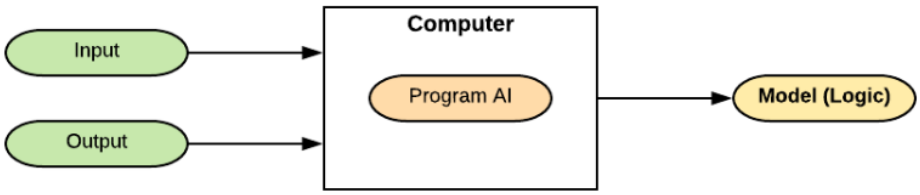
2.1.1 การทำเหมืองข้อมูล (Data Mining)

เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์ Data Mining เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์ เช่น ลูกค้าส่วนใหญ่ที่ซื้อเบียร์มักจะซื้อผ้าอ้อมด้วย จะเห็นว่าข้อมูลนี้เป็นข้อมูลที่ไม่เคยคิดว่ามีความสัมพันธ์กัน และเมื่อได้ความรู้แบบนี้ก็อาจจะนำเป็นออกโปรโมชั่นหรือช่วยในการจัดวางชั้นสินค้า หรือเป็นแนวทางในการสั่งซื้อสินค้าในซุปเปอร์มาร์เก็ตต่อไปได้ นอกจากนี้ Data Mining ยังมีเทคนิคในการประยุกต์ใช้งานได้อย่างดี (หนึ่งหทัย ชัยอาภร, 2559: ออนไลน์)

2.1.1.1 การทำเหมืองข้อมูล จำแนกออกเป็น 2 ประเภท คือ

1) Unsupervised Learning การสร้างโมเดลโดยใช้ข้อมูล input เพียงอย่างเดียวไม่มี target การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่อง โดยการสร้างโมเดลที่เหมาะสมกับข้อมูล การเรียนรู้แบบนี้แตกต่างจากการเรียนรู้แบบมีผู้สอน คือ จะไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน การเรียนรู้แบบนี้จะพิจารณาวัตถุเป็นเซตของตัวแปรสุ่ม แล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูลการเรียนรู้แบบไม่มีผู้สอนสามารถนำไปใช้ร่วมกับการอนุมานแบบเบย์ เพื่อหาความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรสุ่มโดยกำหนดตัวแปรที่เกี่ยวข้องให้ นอกจากนี้ยังสามารถนำไปใช้ในการบีบอัดข้อมูล ซึ่งโดยพื้นฐานแล้ว ขั้นตอนวิธีการบีบอัดข้อมูลจะขึ้นอยู่กับ การแจกแจงความน่าจะเป็นของข้อมูลไม่อย่างชัดแจ้งก็โดยปริยาย (สุพรรณ ฟ้าหยง, 2562)

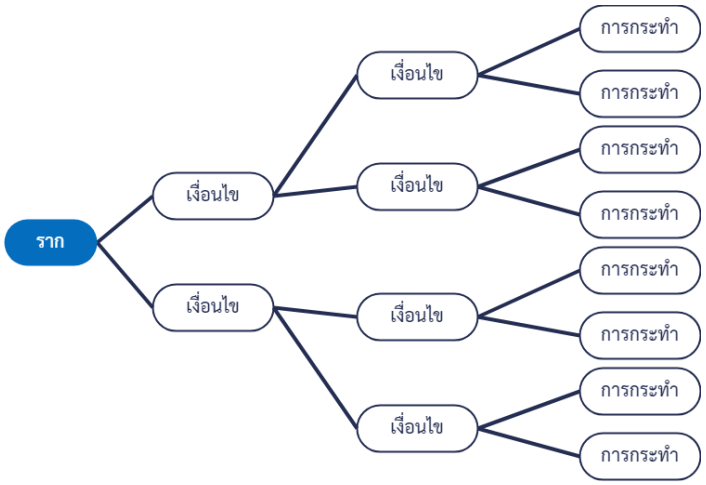
2) Supervised Learning เป็นการเรียนรู้ข้อมูลต่าง ๆ โดยมีผู้สอน อาศัยข้อมูลในการฝึกฝน เพื่อช่วยให้ตัวเทคโนโลยีสามารถเรียนรู้ผล และคาดคะเนผลลัพธ์ต่าง ๆ ได้อย่างแม่นยำมากยิ่งขึ้น โดยการเรียนรู้ในรูปแบบนี้มักถูกนำมาใช้งานในเชิงธุรกิจทั้งการคำนวณราคาบ้าน การคาดคะเนค่าเงิน หรือแม้แต่การวิเคราะห์ผลการแข่งขันต่าง ๆ เป็นต้น กระบวนการสร้าง model เรียกว่าการ เทรน ซึ่งสามารถกินเวลาได้ตั้งแต่หลักวินาทีจนถึงหลาย ๆ วัน แล้วแต่ความซับซ้อนของโจทย์ที่เราต้องการแก้ และพลังในการประมวลผลของเครื่องคอมพิวเตอร์ที่ใช้เทรน (Phuri Chalermkiatsakul, 2563. : ออนไลน์)



ภาพที่ 2.1 แสดงกระบวนการเทรน เพื่อให้ได้ model ที่ต้องการ
ที่มา : Phuri Chalermkiatsakul (2563: ออนไลน์)

2.1.2 ต้นไม้ตัดสินใจ (Decision Tree) C4.5

ต้นไม้การตัดสินใจ เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่าง ๆ แตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ ต้นไม้การตัดสินใจมีประโยชน์ในการสรุปการตัดสินใจที่มีความซับซ้อนให้ง่ายต่อความเข้าใจ ปัจจุบันต้นไม้การตัดสินใจเป็นที่นิยมใช้ในงานหลายอย่าง เช่น การแพทย์ ธุรกิจ การเขียนโปรแกรม การสร้างเครื่องที่เรียนรู้ได้เอง การสร้างระบบผู้เชี่ยวชาญ ฯลฯ (ครรชิต มาลัยวงศ์, 2553: ออนไลน์)



ภาพที่ 2.2 แสดงโครงสร้างต้นไม้การตัดสินใจ
ที่มา: ดัดแปลงจาก Nuthdanai wangpratham (2564: ออนไลน์)

เทคนิคต้นไม้ตัดสินใจ จะมีลักษณะคล้ายโครงสร้างต้นไม้ที่แต่ละโหนดแสดงคุณลักษณะ แต่ละกิ่งแสดงเงื่อนไขในการทดสอบ และโหนดปลายแสดงกลุ่มที่กำหนดไว้ ต้นไม้ตัดสินจ ประกอบด้วย

- โหนดภายใน (Internal Node) คือ คุณสมบัติต่าง ๆ ของข้อมูล ใช้ในการ ตัดสินใจว่าข้อมูลจะไปอยู่ในกรณีไหน โดยโหนดภายในที่เป็นโหนดเริ่มต้น เรียกว่า โหนดราก

- กิ่ง (Branch,Link) เป็นค่าคุณสมบัติหรือเงื่อนไขของคุณสมบัติในโหนดที่ใช้ในการจำแนกข้อมูล ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดภายใน

- โหนดใบ (Leaf Node) คือคลาสต่าง ๆ ซึ่งเป็นผลลัพธ์

เกณฑ์ที่ช่วยตัดสินใจ ในการเลือก โหนดราก (Root Node) คือการทดลองเลือกคุณลักษณะแต่ละตัวมาทำหน้าที่เป็นโหนดราก แล้วหาค่า Gain ratio ซึ่งเป็นค่าที่ใช้บอกคุณลักษณะแต่ละตัวมาทำหน้าที่เป็นโหนดราก ดังแสดงในสมการดังนี้

$$Gain(X)=info(T)-info_x(T) \quad \text{----------(2.1)}$$

โดยที่ T แทน เซตของ Training Set

X แทน คุณลักษณะที่ถูกเลือกให้เป็นตัวจำแนกข้อมูล

Info(T) เป็นฟังก์ชันที่ระบุปริมาณข้อมูลที่ต้องการเพื่อให้สามารถจำแนกคุณลักษณะที่ต้องการได้

(Info)_X(T) หรือ Entropy คือ ฟังก์ที่ระบุปริมาณข้อมูลที่ต้องการเพื่อการจำแนกคลาสของข้อมูลโดยใช้คุณลักษณะ X เป็นตัวตรวจสอบเพื่อแยกข้อมูล (เอกสิทธิ์ พัชรวงศ์ศักดา และสิริวรรณ แต้วิจิตร. 2553)

Info(T) เป็นฟังก์ชันที่ระบุปริมาณข้อมูลที่ต้องการเพื่อให้สามารถจำแนกคุณลักษณะที่ต้องการได้ จากสมการนี้

$$Info(T)=-\sum_{j=1 \text{ to } k}[freq\ (C_{j,T})\div |T|]\times \log_2 [freq(C_{j,}T)\div |T|]bits \quad \text{----------(2.2)}$$

โดยที่ |T| คือ จำนวนข้อมูลทั้งหมดใน Training Datasets

(Freq(C)_(j,)T) คือ ความถี่ที่ข้อมูลใน T ปรากฏเป็นคลาส C_(j)

(Info)_X(T) หรือ Entropy คือ ฟังก์ที่ระบุปริมาณข้อมูลที่ต้องการเพื่อการจำแนกคลาสของข้อมูลโดยใช้คุณลักษณะ X เป็นตัวตรวจสอบเพื่อแยกข้อมูล

$$Info(T)=-\sum_{j=1 \text{ to } n}\left(\frac{|T_i|}{T}\right)x\ info(T)bits \quad \text{----------(2.3)}$$

โดยที่ i คือ จำนวนค่าที่เป็นไปได้ของคุณลักษณะ X

(|T)i| คือ จำนวนค่าที่เป็นไปได้ของคุณลักษณะ X=1

2.1.3 การคัดเลือกคุณสมบัติ (Feature Selection)

การคัดเลือกคุณสมบัติเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในตัวแบบพยากรณ์ อาจกระทำเพื่อเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว หรือเลือกกลุ่มของตัวแปรที่มีความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณสมบัติเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลของการทำเหมืองข้อมูล เพื่อทำให้การสร้างตัวแบบพยากรณ์มีประสิทธิภาพ เพราะจะช่วยลดมิติของข้อมูลและอาจช่วยให้การเรียนรู้วิธีการ พยากรณ์ดำเนินการได้เร็วขึ้นและมีประสิทธิภาพมากขึ้น ในงานวิจัยนี้ ทดลองใช้การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection (นิภาพร ชนะมาร และพรรณี สิทธิเดช, 2557)

2.1.3.1 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection เป็นวิธีคัดเลือกตัวแปรโดยมี หลักการเช่นเดียวกับการเลือกตัวแปรของการสร้างต้นไม้ตัดสินใจ เพื่อให้ได้ตัวแปรที่เป็นตัวแบ่งข้อมูล ออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด (Homogeneous) ด้วยมาตรวัดการได้ ประโยชน์จากการแบ่งกลุ่มย่อยเรียกว่าอัตราส่วนเกน (Gain Ratio) ค่าอัตราส่วนเกนจะเป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยที่พัฒนามาจากค่าเกนความรู้ โดยเมื่อเราใช้ค่าเกนความรู้ในการแบ่งชุดข้อมูลจะทำให้เกิดความเอนเอียงเกิดขึ้นเมื่อแอทริบิวที่ทำการพิจารณามีค่าที่เกิดขึ้นเป็นจำนวนมาก โดยในการใช้ค่าเกนความรู้มักจะทำการเลือกแอทริบิวที่มีค่าที่เกิดขึ้นเป็นจำนวนมาก (โกเมศ อัมพวัน, 2563: ออนไลน์) ซึ่งอัตราส่วนของค่าเกน (Gain หรือ Information Gain) กับ ค่าสารสนเทศการแบ่งกลุ่ม (Split Info) อันเป็นการลดอิทธิพลของตัวแปรที่มีค่าหลายค่า ผลที่ได้รับจากการใช้เทคนิคนี้จะได้ลำดับของตัวแปร ซึ่งตัวแปรที่อยู่ลำดับแรก ๆ จะถือว่ามีอิทธิพลในการพยากรณ์ตัวแปรเป้าหมายมากกว่าตัวแปรในลำดับถัดไป ทำให้เราสามารถพิจารณาเลือกจำนวนตัวแปรที่เหมาะสมได้อย่างมีประสิทธิภาพ (Tan, Steinbach and Kumar. 2006, Asha, Manjunath and Jayaram. 2010) เกนเรโช (GR) เป็นการประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ในแต่ละคลาสการคำนวณ GR โดยใช้ค่า SplitINFO ในสมการที่ 1 และการคำนวณค่าการวัด Gain Ratio ดังสมการที่ 2 (วีระยุธ พิมพาพร และพยุง มีสัจ, 2557)

$$SplitINFO = \sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n} \text{----------(2.4)}$$

หลังจากทำการคำนวณหาค่า SplitINFO แล้วเราจะสามารถคำนวณหาค่าอัตราส่วนเกนได้ดังนี้

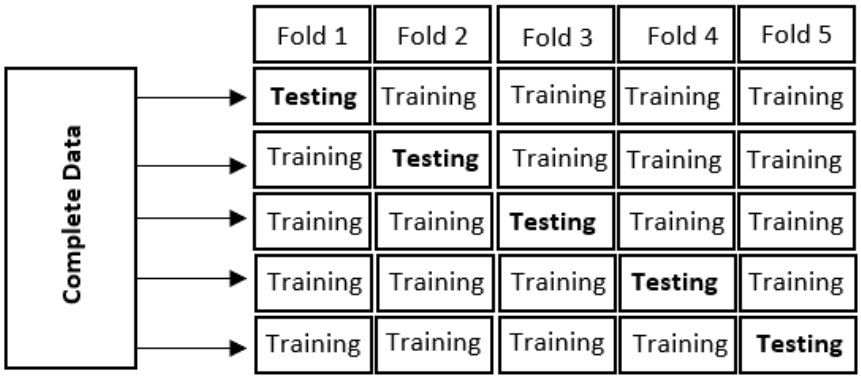
$$GainRatio = \frac{\Delta\ INFO}{SplitINFO} \text{----------(2.5)}$$

เมื่อทำการคำนวณหาค่าอัตราส่วนเกนของทุกแอทริบิวที่ทำการพิจารณาแล้ว เราจะทำการเลือกแอทริบิวที่มีค่าอัตราส่วนเกนสูงที่สุดเพื่อเป็นแอทริบิวสำหรับแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยต่อไป

2.1.4 ตัววัดประสิทธิภาพของโมเดล

2.1.4.1 การวัดประสิทธิภาพโมเดลด้วย Cross Validation

การวัดประสิทธิภาพโมเดลการจำแนกข้อมูลด้วย Cross Validationได้กำหนดค่า K =5 และ K =10 โดยการทดสอบแบบเคโฟลครอสเวลิเดชันนั้น จะเลือกสุ่มข้อมูลออกเป็น K ชุดเท่า ๆ กันในการทดลองครั้งแรกข้อมูลชุดที่ 1 เป็นข้อมูลสำหรับทดสอบและข้อมูล ชุดที่เหลือเป็นข้อมูลชุดสอน ในการทดลองครั้งที่สองข้อมูลชุดที่ 2 เป็นข้อมูลทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอนวนซ้ำจะกระทั้ง ข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบ (พรนภา ชุมเชื้อ, 2562)



ภาพที่ 2.3 ตัวอย่างการแบ่งสัดส่วนทดสอบประสิทธิ์ภาพด้วยวิธี 5- Folds Cross Validation
ที่มา: พรนภา ชุมเชื้อ (2562)

การประเมินประสิทธิภาพการจำแนกของโมเดล และแสดงผลด้วยเมตริกซ์เป็นส่วนสำคัญในขั้นตอนสุดท้ายของการทำเหมืองข้อมูล เนื่องจากการวัดประสิทธิภาพของการวิเคราะห์ปัจจัยของข้อมูลจะบอกถึงความน่าเชื่อถือของโมเดล โดยใช้รูปแบบของตาราง Confusion Matrix เป็นเครื่องมือในการคำนวณการวัดประสิทธิภาพ (ธรรมสรณ์ นูมหันต์, 2558)

ตารางที่ 2.1 ตาราง Confusion Matrix ของข้อมูล Weather ซึ่งมี 2 คลาส

Predicted/Actual	Yes	No
Yes	TP	FP
No	FN	TN

จากตารางที่ 2.1 ค่าที่แสดงในช่องต่าง ๆ ของตารางประกอบด้วย

True Positive (TP) คือ จำนวนข้อมูลทดสอบที่เป็นคลาส Positive และโมเดลจำแนกได้ถูกต้องว่าเป็น Positive

False Positive (FP) คือ จำนวนข้อมูลทดสอบที่ไม่ใช่คลาส Positive แต่โมเดลทำนายผิดว่าเป็นคลาส Positive

True Negative (TN) คือ จำนวนข้อมูลทดสอบที่เป็นคลาส Negative และโมเดลทำนายได้ถูกต้องว่าเป็นคลาส Negative

False Negative (FN) คือ จำนวนข้อมูลทดสอบที่เป็นคลาส Positive และโมเดลทำนายผิดว่าเป็นคลาส Negative

การนำโมเดลไปใช้งานจริงได้นั้นจำเป็นจะต้องทราบประสิทธิภาพของโมเดล ทั่วไปจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่าง ๆ (ธาดา จันตะคุณ, 2559) เพื่อหาค่าความถูกต้องในการจำแนกประเภท ซึ่งใช้การหาค่า Accuracy ซึ่งเป็นหาค่าความถูกต้องจากการวัดอัตราส่วนของผลการจำแนกประเภทเหตุการณ์ฯ ที่ถูกต้องทั้งหมดต่อผลการจำแนกประเภทเหตุการณ์ฯ ทั้งหมด เมื่อแต่ละเทคนิคได้ค่า Accuracy จะนำมเปรียบเทียบหาเทคนิคที่ให้ค่าความถูกต้องสูงสุด

จากผลลัพธ์ที่ได้ สามารถนิยามการวัดประสิทธิภาพ ได้ดังนี้ (ภรัณยา ปาลวิสุทธิ์, อภินันท์ จุ่นกรณ์, มงคล รอดจันทร์ และธานิล ม่วงพูล, 2563) แสดงรายละเอียดดังนี้

1) ความแม่นยำ (Accuracy)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{----------(2.6)}$$

จากสมการที่ 6 คือ ความแม่นยำ (Accuracy) ค่าความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส สามารถหาได้จากจำนวนการพยากรณ์ทั้ง Positive และ Negative ได้ถูกต้องหารด้วยจำนวนข้อมูลทั้งหมด

2) เป็นการวัดค่าความแม่นยำของโมเดล (Precision)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \text{----------(2.7)}$$

จากสมการที่ 7 คือ ค่าความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส สามารถหาได้จากจำนวนการพยากรณ์ที่ถูกต้อง Positive หารด้วยส่วนที่เป็นจริง Positive ทั้งหมดของข้อมูลจริง

3) Recall เป็นการวัดค่าความถูกต้องของโมเดล

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \text{----------(2.8)}$$

จากสมการที่ 8 คือ การวัดความสามารถในการค้นหาข้อมูลที่เป็นคลาส Positive คำนวณได้จากการหาอัตราส่วนของการพยากรณ์ว่าถูกต้องเมื่อเทียบกับข้อมูลที่ถูกต้องจริง (TP) หารด้วยค่าทีพยากรณ์ว่าถูกต้องทั้งหมด (TP+FN)

4) F-measure เป็นการวัดค่า Precision และ Recall

$$\text{F-measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \text{----------(2.9)}$$

จากสมการที่ 9 คือ F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

5) (Root Mean Square Error, RMSE) ค่าคลาดเคลื่อนกำลังสองเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{Error})^2} \text{----------(2.10)}$$

จากสมการที่ 10 คือ (root mean square error, RMSE) ค่าคลาดเคลื่อนกำลังสองเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย

2.1.5 กระบวนการวิเคราะห์ข้อมูล CRISP-DM (Cross-Industry Standard Process For Data Mining)

ภาพที่ 2.4 CRISP-DM Process Model
ที่มา: Thapanee Boonchob (2564: ออนไลน์)

กระบวนการมาตรฐานที่ใช้สำหรับการทำเหมืองข้อมูล เพื่อทำการวิเคราะห์และนำไปใช้ประโยชน์ มีอยู่ 6 ขั้นตอน คือ

1) การทำความเข้าใจโจทย์ (Business Understanding) ขั้นตอนแรกมุ่งไปที่การทำความเข้าใจข้อมูลปัญหาและวัตถุประสงค์ของโครงการจากมุมมองข้อมูล จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

2) การทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนนี้เริ่มต้นด้วยการรวบรวมข้อมูล จากนั้นทำความเข้าใจ ตรวจสอบคุณภาพ และเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์ขั้นตอนที่ 1 และ 2 สามารถทำกลับไปมาได้ เนื่องจากการทำความเข้าใจข้อมูลทำให้เราเข้าใจข้อมูลมากขึ้น และการเข้าใจข้อมูลก็ทำให้เราเข้าใจข้อมูลมากขึ้นเช่นกัน

3) การเตรียมข้อมูล (Data Preparation) ขั้นตอนการเตรียมข้อมูล หมายถึงขั้นตอนทั้งหมดที่จะทำเพื่อให้ข้อมูลดิบที่เรารวบรวมมา กลายเป็นข้อมูลสมบูรณ์ที่พร้อมจะเข้าสู่โมเดลในขั้นตอนที่ 4 เช่น การสร้างตาราง การลบข้อมูลที่ไม่ต้องการออก การแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ

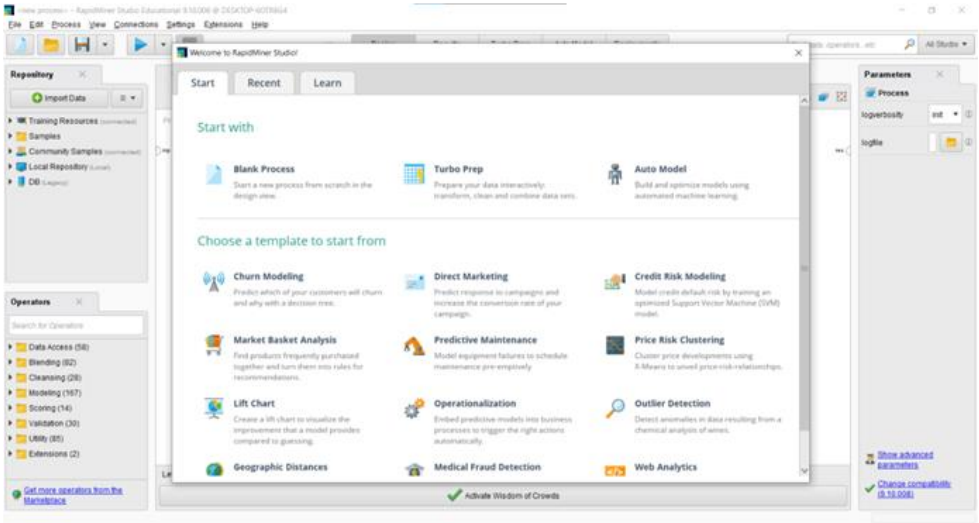
4) การสร้างโมเดล (Modeling) ในขั้นตอนนี้ เราจะเลือกและทดสองสร้างโมเดลหลาย ๆ แบบที่น่าจะสามารถแก้ไขปัญหาที่ต้องการได้ จากนั้นค่อย ๆ ปรับค่าพารามิเตอร์ในแต่ละโมเดล เพื่อให้ได้โมเดลที่เหมาะสมที่สุดมาใช้ในการแก้ไขปัญหา

5) การวัดประสิทธิภาพของโมเดล (Evaluation) เราจะทำการวัดประสิทธิภาพของโมเดลที่ได้จากขั้นตอนที่ 4 เพื่อวัดว่าโมเดลมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานแล้วหรือไม่ ซึ่งโมเดลแต่ละประเภทก็จะมีตัววัดประสิทธิภาพที่แตกต่างกันออกไป

6) การนำโมเดลไปใช้งานจริง (Deployment) เป็นการนำโมเดลที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ (Thapanee Boonchob, 2563)

2.1.6 โปรแกรม RapidMiner Studio

ภาพที่ 2.5 โปรแกรม RapidMiner
ที่มา: mypccrack (2565 : ออนไลน์)



ภาพที่ 2.6 Interface RapidMiner Studio
ที่มา : ผู้วิจัย (2565)

RapidMiner คือซอฟต์แวร์ Data Science ใช้สำหรับการเตรียมข้อมูล การเรียนรู้เครื่อง การเรียนรู้ลึกการทำเหมืองข้อความ และการวิเคราะห์การทำนาย (Predictive analysis) เป็นซอฟต์แวร์ที่ช่วยในการจัดส่งข้อมูล และลดข้อผิดพลาดจนแทบจะไม่จำเป็นต้องเขียนโค้ดเพิ่ม แต่ที่ทำให้เป็นเครื่องมือที่ Data Scientist นิยมเลือกใช้เป็นเพราะว่า RapidMiner มีขั้นตอนพร้อมสำหรับการทำ Data mining (ขุดข้อมูล) และ Machine Learning ซึ่งรวมไปถึงการโหลดและการแปลงข้อมูล (ETL) การประมวลผลล่วงหน้าและการวาดภาพจากข้อมูล การวิเคราะห์เชิงพยากรณ์และการสร้างแบบจำลองทางสถิติ การประเมินผลและการปรับใช้ ต่าง ๆ ล้วนเป็นสิ่งที่ Data Scientist จำเป็นต้องทำในการเข้าใจข้อมูลมากขึ้น (Achieve. Plus, 2563: ออนไลน์)

2.2 งานวิจัยที่เกี่ยวข้อง

นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ได้ศึกษาการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ มีวัตถุประสงค์เพื่อการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการพยากรณ์ผลสัมฤทธิ์ ทางการเรียนของนิสิต โดยใช้เทคนิคการคัดเลือกคุณสมบัติที่สำคัญแล้วสร้างตัวแบบการพยากรณ์ด้วย เทคนิค BPNN และเทคนิค SVMs จากข้อมูลที่คัดเลือกซึ่งเป็นปัจจัยการเรียนรู้ที่สำคัญ ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลข้อมูลของนิสิตที่ศึกษาหลักสูตรปริญญาตรี สาขาวิชาวิทยาการ คอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2548 จำนวน 180 ระเบียน ประกอบด้วยคุณสมบัติ 23 ตัวแปร แบ่งเป็น ตัวแปรอิสระ 22 ตัวแปร ผู้วิจัยได้ทดลองสร้างตัวแบบการพยากรณ์จากข้อมูลทั้งหมดที่มีตัวแปรอิสระ 22 ตัวแปร ด้วย เทคนิค BPNN และเทคนิค SVMs ได้ผลการพยากรณ์ที่มีค่ารากที่สองของกำลังสองของข้อผิดพลาด (Root Mean Square Error: RMSE) เท่ากับ 0.2444 และ 0.1246 ตามลำดับ หลังจากนั้น จึงทำการวิเคราะห์ปัจจัยการเรียนรู้ ด้วยการคัดเลือกคุณสมบัติที่สำคัญ โดยใช้เทคนิคการคัดเลือกคุณสมบัติ 3 วิธี ได้แก่ การคัดเลือกคุณสมบัติ แบบ Correlation-based Feature Selection การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection และ การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection ผลการทดลองทั้งสามเทคนิคสามารถลดจำนวน ของคุณสมบัติจาก 22 ตัวแปร เหลือ 9 ตัวแปร 10 ตัวแปร และ 11 ตัวแปร ตามลำดับ ผลของงานวิจัยนี้ให้ประโยชน์ในการ วิเคราะห์ปัจจัยการเรียนรู้และการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิตซึ่งจะช่วยให้นิสิตสามารถ พยากรณ์ผลการเรียนของตนเองและปรับพฤติกรรมการเรียน ได้เช่น การเพิ่มถอนรายวิชาให้เหมาะสมกับ ศักยภาพตนเอง

วรายุทธ พลาศรี (2556) ได้ศึกษาปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบท กรณีศึกษาจังหวัดมหาสารคาม การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาสภาพเศรษฐกิจของครัวเรือนในชนบท สถานการณ์ความยากจน ลักษณะของครัวเรือนที่ยากจน และปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบทจังหวัดมหาสารคาม โดยกลุ่มประชากรที่ใช้ในการศึกษาคือ ครัวเรือนที่อยู่ ในเขตพื้นที่ชนบทจังหวัดมหาสารคาม จำนวน 180,328 ครัวเรือน ขนาดกลุ่มตัวอย่างเท่ากับ 400 ครัวเรือน โดยได้เลือกวิธีการสุ่ม ตัวอย่างแบบหลายขั้นตอน (Multi-stage sampling method) และ

ได้ทำการเก็บรวบรวมข้อมูลโดยการใช้แบบสอบถามเป็นเครื่องมือ หลังจากนั้นจึงนำข้อมูลที่ได้มาวิเคราะห์โดยการใช้สถิติพรรณนาและสถิติอนุมาน การศึกษาครั้งนี้ได้ใช้เส้นความยากจนของครัวเรือนภาคตะวันออกเฉียงเหนือในเขตพื้นที่ชนบท ปี 2553 ที่คำนวณโดยสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ซึ่งจากการคำนวณได้เส้นความยากจนเท่ากับ 1,565 บาทต่อคนต่อเดือน เป็นเกณฑ์ในการแบ่งกลุ่มครัวเรือนยากจนกับกลุ่มครัวเรือนที่ไม่ยากจน ผลการศึกษาพบว่า ครัวเรือนในชนบทของจังหวัดมหาสารคามมีจำนวนสมาชิกในครัวเรือนเฉลี่ยครัวเรือนละ 4.16 คน มีจำนวนแรงงานในครัวเรือน เฉลี่ยครัวเรือนละ 3.15 คน และจำนวนสมาชิกที่มีรายได้ในครัวเรือนเฉลี่ยครัวเรือนละ 2.13 คน ระดับการศึกษาของหัวหน้าครัวเรือนของกลุ่มตัวอย่างส่วนใหญ่สำเร็จการศึกษาในระดับประถมศึกษาร้อยละ 49.3 อาชีพของหัวหน้าครัวเรือนส่วนใหญ่ประกอบอาชีพ เกษตรกรรมคิดเป็นร้อยละ 57.5 ครัวเรือนมีรายได้รวมเฉลี่ยครัวเรือนละ 16,036.89 บาท ต่อเดือน และมีค่าใช้จ่ายสำหรับการอุปโภค บริโภค เฉลี่ยครัวเรือนละ 7,666 บาทต่อเดือน เมื่อคิดเป็นอัตราส่วนร้อยละของค่าใช้จ่ายอุปโภคบริโภคต่อรายได้จะเท่ากับ 47.80 มี หนี้สินเฉลี่ยครัวเรือนละ 187,530.38 บาท และครัวเรือนมีการเก็บออมคิดเป็นร้อยละ 76.5 ของจำนวนครัวเรือนตัวอย่างทั้งหมด สถานการณ์ความยากจนและลักษณะของครัวเรือนที่ยากจนพบว่า ครัวเรือนตัวอย่างในเขตพื้นที่ชนบทของจังหวัดมหาสารคามมี สัดส่วนของครัวเรือนที่ยากจนคิดเป็นร้อยละ 31.2 โดยครัวเรือนที่ยากจนในเขตชนบทจะมีลักษณะร่วมคือ หัวหน้าครัวเรือนมีระดับ การศึกษาต่ำ มีครัวเรือนขนาดใหญ่ มีระดับรายได้ต่ำ มีขนาดพื้นที่ที่ใช้ในการประกอบอาชีพการเกษตรน้อย มีระดับความมั่งคั่งต่ำ และมีหนี้สิน ส่วนปัจจัยที่มีผลต่อความยากจนของครัวเรือน ได้แก่ ระดับการศึกษาของหัวหน้าครัวเรือน ขนาดของครัวเรือน ขนาดพื้นที่ที่ใช้ ในการประกอบอาชีพ ความมั่งคั่งและหนี้สินของครัวเรือน

ภัทร์พงศ์ พงศ์ภัทรกานต์, วิชัย พัวรุ่งโรจน์, คมยุทธ ไชยวงษ์, สุชาดา พรหมโคตร และปาริชาติ แสงระชัฏ (2560) ได้ศึกษาการใช้เทคนิคเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา งานวิจัยนี้นำเสนอการทดสอบวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา มหาวิทยาลัยราชภัฏเลย โดยใช้ข้อมูลการเข้าใช้บริการผ่านประตูอัตโนมัติในช่วงเดือนกุมภาพันธ์ถึง ตุลาคม 2559 ที่มี 9 ปัจจัยพื้นฐาน คือ วันที่เข้าใช้บริการ ช่วงเวลา เพศ คณะ ชั้นปี จังหวัดที่เกิด หมู่ เลือด จำนวนพี่น้อง และเกรดเฉลี่ยสะสม จำนวน 79,953 ชุดข้อมูล ทำการประมวลผลด้วยอัลกอริทึม C5.0, Neural Network และ CART เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการคัดแยกข้อมูล ผลการศึกษาพบว่าอัลกอริทึม C5.0 ให้ค่าความถูกต้อง 97.78% และใช้ระยะเวลาในการประมวลผล น้อยกว่าอัลกอริทึมที่นำมาเปรียบเทียบ ผลจากการวิเคราะห์ด้วยอัลกอริทึม C5.0 พบว่ามี 3 ปัจจัยที่มี อิทธิพลต่อการใช้บริการห้องสมุดของนักศึกษาที่ส่งผลตามคณะ คือ เกรดเฉลี่ยสะสม มีอิทธิพลสูงสุด ร้อยละ 93.8% เพศ มีอิทธิพลร้อยละ 6.0 และช่วงเวลา มีอิทธิพลร้อยละ 0.2 ซึ่งนำมาสร้างความสัมพันธ์ ได้ 21 ระดับ ซึ่งเป็นแนวทางในการประชาสัมพันธ์และส่งเสริมนักศึกษาเข้ามาใช้บริการห้องสมุดผ่าน คณะที่สังกัดได้ โดยเฉพาะเกรดเฉลี่ยมีผลอย่าง

มากในการเข้ามาใช้บริการ นักศึกษาที่มีเกรดสูงมี แนวโน้มการเข้าใช้ห้องสมุดมากกว่านักศึกษาที่มีเกรดต่ำ ดังนั้น ห้องสมุดควรเน้นไปที่การเปิดบริการ หรือเชิญชวนให้นักศึกษาที่มีเกรดน้อยเข้าห้องสมุดมากขึ้น ห้องสมุดควรคิดกิจกรรมส่งเสริมใหม่เพิ่ม มากขึ้นเพื่อให้นักศึกษามีความสนใจในการเข้าใช้ห้องสมุด

รัชพล กลัดชื่น และจรัญ แสนราช (2561) เปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยทำการศึกษาข้อมูลนักศึกษาระดับประกาศนียบัตรวิชาชีพ จำนวน 5,100 ระเบียน ตั้งแต่ปีการศึกษา 2550 -2559 9 สาขาวิชา 27 คุณลักษณะ โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ Decision Tree : J48graft, Naïve Bayes และ Rule Induction ทำการเปรียบเทียบประสิทธิภาพตัวแบบการทำนาย ระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะแบบ Forward Select ทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation โดยใช้โปรแกรม Rapid Miner Studio 8 จากนั้นนำผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด 2 ค่า มาทำการเปรียบเทียบด้วยวิธี T-Test ผลการศึกษาพบว่าการใช้เทคนิค Decision Tree : J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection และ การเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องเท่ากับ 83.08% และ 81.71% ตามลำดับ และทดสอบด้วยวิธี T-Test พบว่าการทดสอบทั้งสองแบบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จากผลการเปรียบเทียบประสิทธิภาพในครั้งนี้ สามารถนำเทคนิค Decision Tree : J48graft ไปใช้ในการทำนายผลสัมฤทธิ์ทางการเรียน และเป็นแนวทางในการสอนเสริมหรือแนะแนวให้กับนักศึกษาต่อไป

ประเสริฐ บัวทอง (2560) ได้ศึกษาปัจจัยที่มีผลต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในตำบลอ่างคีรี อำเภอมะขาม จังหวัดจันทบุรี มีวัตถุประสงค์ ประการแรกเพื่อศึกษาปัจจัยส่วนบุคคลที่มีผล ต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในพื้นที่ ตำบลอ่างคีรี อำเภอมะขาม จังหวัดจันทบุรี และประการที่สองเพื่อศึกษาปัจจัยทางเศรษฐกิจและปัจจัยด้านกายภาพที่มีผลต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในพื้นที่ ตำบลอ่างคีรีอำเภอมะขาม จังหวัดจันทบุรี โดยมีกลุ่มตัวอย่าง คือ เกษตรกรสวนทุเรียนที่ขึ้นทะเบียนการปลูกทุเรียนในตำบลอ่างคีรี อำเภอมะขามจังหวัดจันทบุรี จำนวน 300 ครัวเรือน และสัมภาษณ์เชิงลึก จำนวน 10 ครัวเรือน ผลการวิจัยเป็นไปตามสมมติฐานที่ตั้งไว้พบว่า เกษตรกรส่วนใหญ่เป็นเพศชายอายุ 31-40 ปี ระดับการศึกษาประถมศึกษารายได้ต่อปีของครอบครัว 750,000-1,000,000 บาท มีจำนวนแรงงานในการปลูกทุเรียน 6-10 คน ต้นทุนในการปลูกทุเรียน 100,001-200,000 บาท สายพันธุ์ทุเรียนที่ปลูกหมอนทอง ลักษณะของดินเป็นดินร่วน ลักษณะพื้นที่เป็นที่ราบสูงขนาดของแหล่งน้ำ มีขนาดตั้ง แต่1ไร่ลงมาการคมนาคมสะดวกสบายเป็นช่วงขนาดพื้นที่ปลูกทุเรียนมีขนาด 26-50ไร่ ประสบการณ์ในการปลูกทุเรียน 3-6 ปี การสัมภาษณ์

เกษตรกรส่วนใหญ่ให้เหตุผลว่า ทำไมถึงตัดสินใจปลูกทุเรียน เพราะทุเรียนเป็นผลไม้ที่มีความต้องการของตลาดสูง

วีระยุธ พิมพาพร และพยุง มีสัจ (2557) ได้ศึกษาการวิเคราะห์องค์ประกอบของชุดข้อมูลที่ซับซ้อนด้วยวิธีการเลือกคุณลักษณะสำคัญแบบพลวัต มีวัตถุประสงค์เพื่อศึกษากระบวนการวิเคราะห์องค์ประกอบ (Factor Analysi) บนชุดข้อมูลที่ทับซ้อนด้วยวิธีการ เลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) โดยประยุกต์ใช้กระบวนการเลือกตัวแปร (Feature Selection) และการวิเคราะห์กลุ่ม (Clustering analysis) ข้อมูลในการประมวลผลเกิดจากกิจกรรมต่าง ๆ ในระบบการเรียนออนไลน์ (E-Learning) โดยเน้นปัจจัยที่ส่งผลโดยตรงต่อผลสัมฤทธิ์ทางการเรียน ผลการวิจัยพบว่าประสิทธิภาพโดยรวมของกระบวนการวิเคราะห์องค์ประกอบ โดยใช้อัลกอริทึมการเลือกลักษณะสำคัญ แบบพลวัต ให้ค่าความถูกต้องสูงสุดที่ 45.17% โดยใช้ 3 ตัวแปร สำหรับกระบวนการวิเคราะห์องค์ประกอบโดยวิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio ให้ค่าความถูกต้องสูงสุดที่ 44.80% โดยใช้ตัวแปร 7 ตัวแปร จากผลการวิจัยสามารถสรุปได้ว่า อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัตมีค่าความถูกต้องสูงกว่า และใช้จำนวนตัวแปรที่น้อยกว่า วิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio

อัจจิมา มณฑาพันธุ์ (2562) งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการ ปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่าง ๆ จำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation Based Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection หลังจากคัดเลือกคุณลักษณะ ที่สำคัญจึงนำผลที่ได้จากแต่ละเทคนิคมาคำนวณหาค่าประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านมโดยการใช้ทคนิคซัพพอร์ต เวกเตอร์แมชชีน ผลการทดลองพบว่าร้อยละของความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จากจำนวนคุณลักษณะของ ข้อมูลทั้งหมด 30 คุณลักษณะเท่ากับ 91.39% ขณะที่เทคนิค Evolutionary Selection ให้ผลดีที่สุดโดยสามารถลดคุณลักษณะ ที่สำคัญเหลือเพียง 16 คุณลักษณะ และให้ผลการวัดค่าความถูกต้องในการพยากรณ์ได้ดีถึงร้อยละ 95.26%

จากการศึกษางานวิจัยพบว่า ลักษณะข้อมูลที่มีจำนวนมากส่วนใหญ่จะเลือกใช้การคัดเลือกคุณสมบัติ และการคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection มาใช้ในการคัดเลือกวิเคราะห์ปัจจัยที่สำคัญและมีประสิทธิภาพในการคัดเลือกวิเคราะห์ปัจจัยที่เหมาะสม และต้นไม้ตัดสินใจ (Decision Tree) ที่มีผู้เชี่ยวชาญหลายท่านเลือกใช้เพราะให้ผลลัพธ์ ออกมาในระดับดี