การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค

วิทยานิพนธ์

ของ

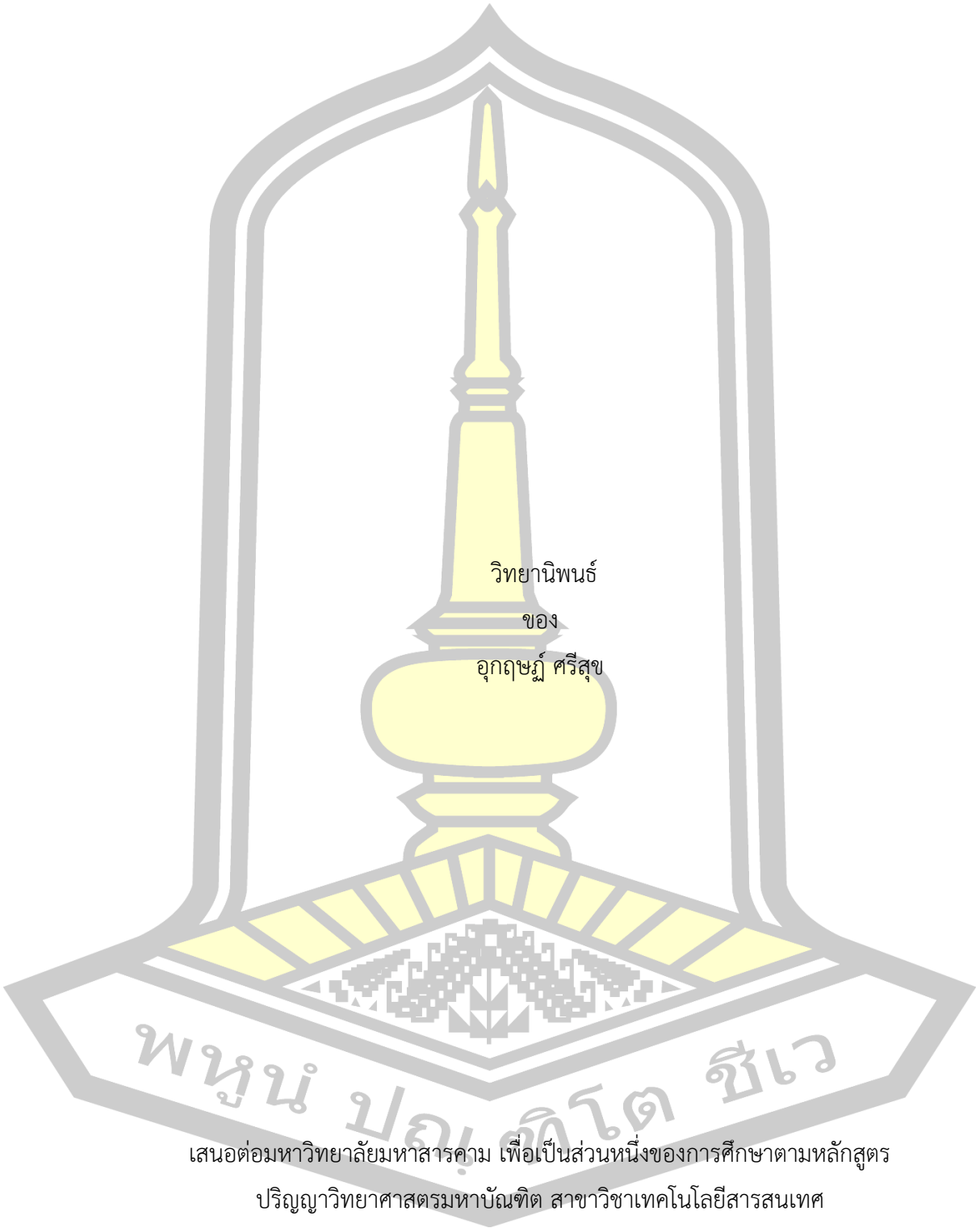
อุกฤษฎ์ ศรีสุข

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มิถุนายน 2564

การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค



วิทยานิพนธ์

ของ

อุกฤษฎ์ ศรีสุข

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มิถุนายน 2564

The performance comparison of data mining techniques for patient incidence

Ukrit Srisuk

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

June 2021

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายอุกฤษฎ์ ศรีสุข แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. สิทธิชัย บุษหมั่น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จารี ทองคำ)

.....กรรมการ
(ผศ. ดร. แกมกาญจน์ สมประเสริฐศรี)

.....กรรมการ
(ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค

ผู้วิจัย อุกฤษฏ์ ศรีสุข

อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จารี ทองคำ

ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2564

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูลในข้อมูลที่หลากหลาย ข้อมูลในงานวิจัยนี้ประกอบด้วยข้อมูลผู้ป่วยโรคมะเร็งเต้านม ผู้ป่วยโรคเบาหวาน และผู้ป่วยโรคไฮเปอร์ไทรอยด์ โดยข้อมูลทั้งหมดถูกรวบรวมมาจากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล มาทำการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks, เทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการวัดประสิทธิภาพแบบพยากรณ์ต่างๆ ผู้วิจัยได้ใช้ 10-fold cross validation ได้ถูกนำมาใช้ในการแบ่งข้อมูลออกเป็นกลุ่มฝึกสอน, กลุ่มทดสอบ และวัดค่าความถูกต้อง ค่าความไว และค่าจำเพาะ หลักการของ Wrapper และ หลักการของ Gain Ratio หลังจากที่นำเข้ามาในการคัดเลือกตัวแปรแล้ว ผลการทดลองพบว่าชุดข้อมูลที่ทำการคัดเลือกตัวแปรด้วยหลักการของ Wrapper โดยใช้เทคนิค Random Forest มีประสิทธิภาพสูงที่สุดคือเทคนิค Random Forest ได้ค่าความถูกต้องในการพยากรณ์โรคไฮโปไทรอยด์ โดยให้ค่าความถูกต้องร้อยละ 99.90 ค่าความไวร้อยละ 99.89 และค่าจำเพาะร้อยละ 100

คำสำคัญ : การทำเหมืองข้อมูล, โรคมะเร็งเต้านม, โรคเบาหวาน, โรคไฮโปไทรอยด์

TITLE The performance comparison of data mining techniques for patient incidence

AUTHOR Ukrit Srisuk

ADVISORS Assistant Professor Jaree Thongkam , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2021

ABSTRACT

This research aims to study the performance of data mining techniques in medical dataset. The data in this research contains data of patients with breast cancer, diabetics and patients with hyperthyroidism. All dataset were collected from UCI databases. This research has been used machine learning in particular Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest, Deep Learning techniques and the Gain Ratio principle to create the models of disease Breast cancer, diabetes and hypothyroidism prediction models. In order to measure the performance of prediction models, 10-fold cross validation was utilized to divide the data into training and testing sets. Accuracy, sensitivity and specificity of the prediction models were used to compare the prediction performance of each model. The experimental results showed that the Random Forest technique was the best technique in modeling the prognosis of hypothyroidism. It provided 99.90% accuracy, 99.89% sensitivity and 100 % specificity.

Keyword : Data Mining, Breast cancer, Diabetics, Hypothyroid

กิตติกรรมประกาศ

การศึกษาค้นคว้าวิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.สิทธิชัย บุษหมั่น ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร.แกมกาญจน์ สมประเสริฐศรี และ ดร.สาธิต แสงประดิษฐ กรรมการสอบ

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ ที่ถ่ายทอดวิชาความรู้ตลอดจนคอยพร่ำสอนศิษย์ด้วยจิตเมตรา ผู้ซึ่งมีจิตวิญญาณของความเป็นครูโดยแท้จริงและแก้ไขข้อบกพร่องต่างๆด้วยความเอาใจใส่ทุกขั้นตอน เพื่อให้วิทยานิพนธ์ฉบับนี้ สมบูรณ์ที่สุด

ขอขอบพระคุณคณาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามที่ให้ความรู้และคำแนะนำในการศึกษา จนสำเร็จการศึกษาในครั้งนี้

ขอขอบพระคุณ บิดามารดาและเพื่อนๆ ที่คอยสนันสนุนและเป็นกำลังใจ จนทำให้งานการศึกษาค้นคว้าวิทยานิพนธ์ครั้งนี้ได้สำเร็จไปด้วยดี

และสุดท้ายขอขอบพระคุณขอเว็บไซต์ UCI ที่ให้ข้อมูลในการทำวิจัยครั้งนี้

อุกฤษฏ์ ศรีสุข

สารบัญ

หน้า

สารบัญภาพ

สารบัญตาราง

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

มะเร็งเต้านม เป็นมะเร็งที่พบมากที่สุดเป็นอันดับ 1 ของผู้หญิงไทย และเป็นสาเหตุของการเสียชีวิตอันดับต้นๆ ในผู้หญิง แนวโน้มคนไทยป่วยเป็นโรคมะเร็งสูงขึ้นทุกปี แต่อัตราการเป็นโรคน้อยกว่าประเทศทางตะวันตก หญิงไทยมีอัตราการพบมะเร็งประมาณ 40 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ซึ่งถ้าเทียบกับประเทศตะวันตกพบมะเร็งเต้านมได้มากกว่า 100 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ในผู้ชายพบมะเร็งเต้านมได้เช่นกัน แต่ไม่บ่อยนัก โดยมีอุบัติการณ์ของโรคนี้น้อยกว่าผู้หญิงเกือบ 100 เท่า [1] ส่วนโรคเบาหวาน เกิดจากเซลล์ร่างกายมีความผิดปกติในขบวนการเปลี่ยนน้ำตาลในเลือดให้เป็นพลังงาน โดยขบวนการนี้เกี่ยวข้องกับอินซูลินซึ่งเป็นฮอร์โมนที่สร้างจากตับอ่อนเพื่อใช้ควบคุมระดับน้ำตาลในเลือด เมื่อน้ำตาลไม่ได้ถูกใช้จึงทำให้ระดับน้ำตาลในเลือดสูงขึ้นกว่าระดับปกติ [2] ยิ่งไปกว่านั้นโรคไฮเปอร์ไทรอยด์ หมายถึง ภาวะที่ต่อมไทรอยด์ มีการหลั่งฮอร์โมนไทรอยด์ออกมามากเกินไป กระตุ้นให้อวัยวะทั่วร่างกายมีการเผาผลาญสูงกว่าปกติ เป็นสาเหตุทำให้เกิดอาการเจ็บป่วย ๆ ต่างขึ้นตามมา เช่น เหนื่อยง่าย ใจสั่น ขี้ร้อนง่าย เหงื่อออกมาก หงุดหงิด นอนไม่หลับ น้ำหนักตัวลดลงอย่างรวดเร็วแบบผิดปกติ เป็นต้น สาเหตุของไทรอยด์เป็นพิษเกิดจากการที่ต่อมไทรอยด์ทำงานมากผิดปกติ จนทำให้ร่างกายมีปริมาณของของฮอร์โมนไทรอยด์มากเกินความต้องการของร่างกายและมีสภาวะเป็นพิษจนส่งผลต่อร่างกายในด้านต่าง ๆ [3]

เหมืองข้อมูล (Data Mining) คือ กระบวนการวิเคราะห์ข้อมูล เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ๆ ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น การจำแนกพันธุ์ต้นไม้ การจำแนกผู้ใช้บัตรเครดิต รวมถึงการจำแนกผู้ป่วยเพื่อพยากรณ์การณ์การอุบัติของโรคต่าง ๆ เช่น โรคมะเร็งเต้านม โรคเบาหวาน และโรคอื่น ๆ เป็นต้น เทคนิคที่นิยมนำมาใช้ในการจำแนกได้แก่ Decision Tree, Naive Bayes, Neural Networks, Random Forest, Deep Learning มีนักวิจัยหลายท่านที่ได้ทำการจำแนกโรคมะเร็งเต้านม เช่น Fan, Zhu, และ Yin [1] ได้ศึกษาการทำนายการกลับเป็นซ้ำของมะเร็งเต้านม ด้วยเทคนิค C4.5, CHAID, QUEST, CART, ANN พบว่า เทคนิค C4.5 มีประสิทธิภาพที่ดีที่สุดที่ 71.17% ส่วน Balpande และ Wajgi [2] ได้ศึกษาการคาดคะเนและการประมาณความรุนแรงของโรคเบาหวานโดยใช้เทคนิคการขุดข้อมูล ด้วยเทคนิค CHAID, Naïve Bayes, K-Nearest, Decision Tree ตัวแปรที่มีผลต่อการคาดคะเน คือ Age, Gender, BMI พบว่า เทคนิค Decision Tree มีประสิทธิภาพที่ดีที่สุดที่ 72% จะเห็นได้ว่าแต่ละเทคนิคมีประสิทธิภาพที่ไม่แน่นอน และในบางครั้งเทคนิคมีประสิทธิภาพต่ำ

การเลือกตัวแปร เป็นหลักการที่ใช้ในการเลือกตัวแปรที่มีความสำคัญต่อการพยากรณ์ การคัดเลือกตัวแปรมีหลายวิธี เช่น การใช้สถิติ Gain Raito และ Chi-square [4] [5] ซึ่งเหมาะต่อข้อมูลที่มีชนิดเป็นแบบ nominal ยิ่งไปกว่านั้นวิธีการ Wrapper เป็นหลักการที่นำเอาเทคนิคในการจำแนกมาทำการวิเคราะห์เลือกตัวแปรที่มีความสำคัญต่อการพยากรณ์มาใช้ เช่น ภูบุญอบ ได้ศึกษาการเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะ 2 แบบ ได้แก่ 1) การเลือกคุณลักษณะแบบควบรวม (Wrapper) โดยใช้วิธี Backward Elimination และ 2) การเลือกคุณลักษณะแบบกรอง (Filter) ด้วยวิธี Gain Ratio และจำแนกข้อมูลด้วยเทคนิค Neural Network เพื่อสร้างแบบจำลองในการพยากรณ์การออกกลางคันของนักศึกษาใน ระดับประกาศนียบัตรวิชาชีพ (ปวช.) และนักศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) วิทยาลัยอาชีวศึกษามหาสารคาม ผลการวิจัยพบว่าวิธี Backward Elimination และ Neural Network ให้ค่าประสิทธิภาพในการจำแนกข้อมูล หรืออัตราการถูกต้องในการทำนายมากที่สุด 100% ประสิทธิภาพของ Gain ratio และ Neural Network 99.93% ส่วนผลลัพธ์ที่ได้จากการใช้อัลกอริทึม Neural Network ให้ค่าประสิทธิภาพในการจำแนกข้อมูลคิดเป็นร้อยละ 99.89%

ดังนั้นงานวิจัยนี้ผู้วิจัยมีความสนใจที่จะศึกษาเทคนิคการคัดเลือกตัวแปร (attribute selection) ด้วยหลักการ Wrapper ในการเพิ่มประสิทธิภาพของเทคนิค Decision Tree, Naive Bayes, Neural Networks, Random Forest, Deep Learning ในการสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮเปอร์ไทรอยด์ ในการวัดประสิทธิภาพแบบจำลอง 10-fold cross validation ใช้ในการแบ่งข้อมูลออกเป็นข้อมูลชุดสอน และชุดข้อมูลทดสอบ เพื่อแสดงค่าความถูกต้อง ค่าจำเพาะ และความแม่นยำ

1.2 วัตถุประสงค์ของการวิจัย

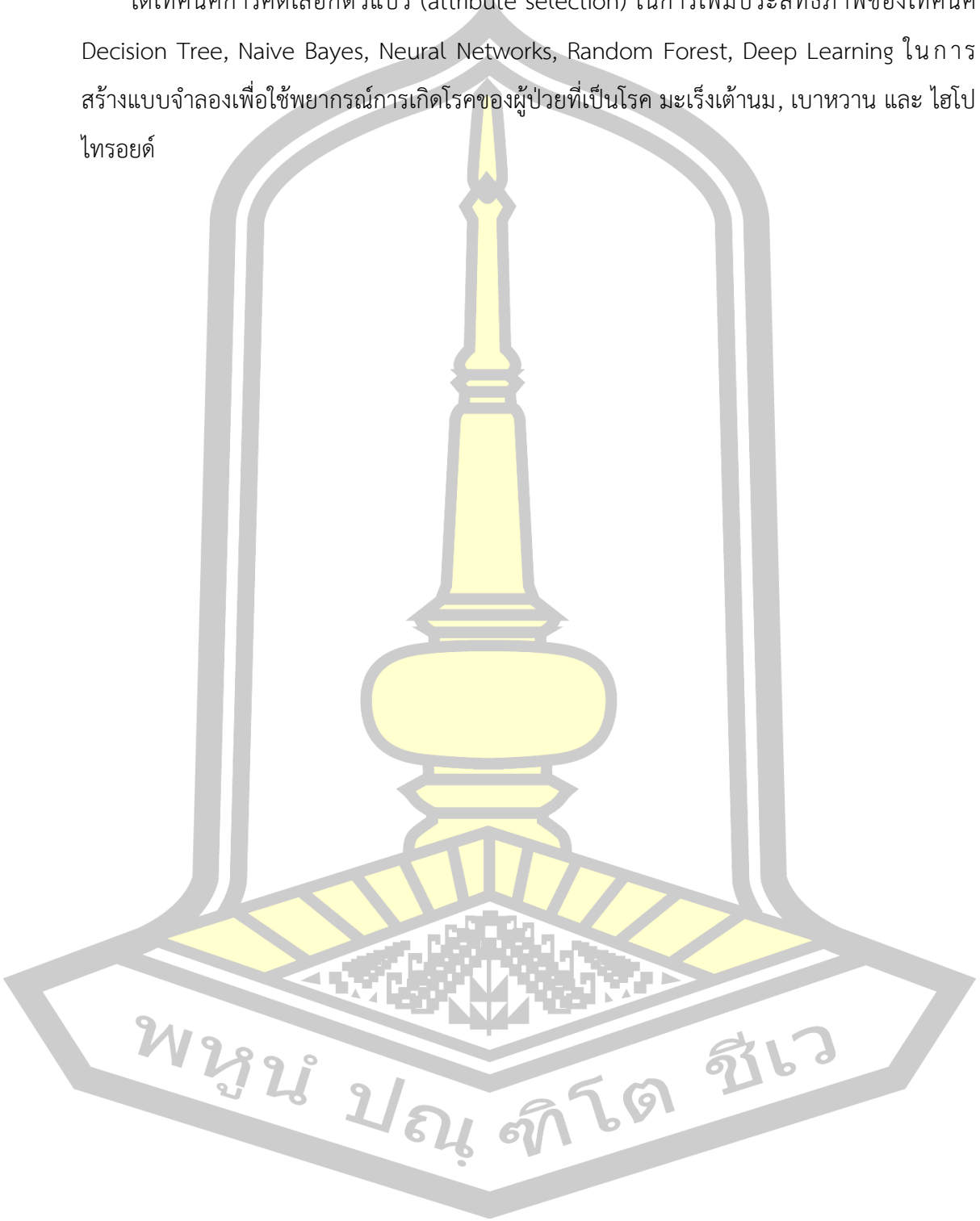
เพื่อศึกษาเทคนิคการคัดเลือกตัวแปร ในการเพิ่มประสิทธิภาพของแบบจำลองในการพยากรณ์การเกิดโรคของผู้ป่วยที่เป็นโรค มะเร็งเต้านม, เบาหวาน และ ไฮโปไทรอยด์

1.3 ขอบเขตงานวิจัย

1. รวบรวมข้อมูลผู้ป่วยโรค มะเร็งเต้านม เบาหวาน และ ไฮโปไทรอยด์จาก ฐานข้อมูล UCI
2. คัดเลือกตัวแปรที่มีความสำคัญต่อการสร้างแบบจำลอง

1.4 ผลที่คาดว่าจะได้รับ

ได้เทคนิคการคัดเลือกตัวแปร (attribute selection) ในการเพิ่มประสิทธิภาพของเทคนิค Decision Tree, Naive Bayes, Neural Networks, Random Forest, Deep Learning ในการสร้างแบบจำลองเพื่อใช้พยากรณ์การเกิดโรคของผู้ป่วยที่เป็นโรค มะเร็งเต้านม, เบาหวาน และ ไฮโปไทรอยด์



บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึง โรคมะเร็งเต้านม โรคเบาหวาน โรคมะเร็งเต้านม เทคนิคการเลือกตัวแปร เทคนิคเหมืองข้อมูลที่ใช้ในงานวิจัย การวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 โรคมะเร็งเต้านม

มะเร็งเต้านม เกิดจากความผิดปกติของเซลล์ที่อยู่ภายในท่อน้ำนมหรือต่อมน้ำนม เซลล์เหล่านี้มีการแบ่งตัวผิดปกติไม่สามารถควบคุมได้อาจมีการแพร่กระจายไปตามทางเดินน้ำเหลือง ไปสู่ต่อมน้ำเหลือง หรือแพร่กระจายไปสู่อวัยวะที่อยู่ห่างไกลเช่น กระดูก ปอด ตับ เป็นต้น การตรวจพบมะเร็งในระยะแรกจะช่วยให้การรักษามีโอกาสประสบความสำเร็จได้สูง มะเร็งเต้านม แบ่งออกเป็น 3 ระยะ คือ

1. มะเร็งเต้านมระยะเริ่มแรก มะเร็งยังคงอยู่เฉพาะในเต้านม ต่อน้ำเหลืองที่รักแร้ยังไม่มีการลุกลามไปอย่างสำคัญ หมายถึง มักจะยังคลำก้อนผิดปกติที่เต้านมไม่ได้ และยังไม่มีการกระจายไปที่ใด

2. มะเร็งเต้านมระยะลุกลาม ระยะนี้โรคมักจะเป็นค่อนข้างมาก เช่นก้อนมีขนาดใหญ่มากขึ้น มากกว่า 5 เซ็นติเมตรขึ้นไป หรือมีการลุกลามไปยังต่อมน้ำเหลืองที่รักแร้ เช่น คลำ ได้ก้อนที่รักแร้ ยิ่งถ้าใหญ่หรือติดกับเนื้อเยื่อข้างใต้ แต่ยังไม่มีการแพร่กระจายไปยังอวัยวะห่างไกล

3. มะเร็งเต้านมระยะแพร่กระจาย คือ มีการกระจายของมะเร็งไปยังอวัยวะห่างไกล เช่น ปอด ตับ กระดูก สมอง เป็นต้น

2.1.2 โรคเบาหวาน

โรคเบาหวาน เกิดจากเซลล์ร่างกายมีความผิดปกติในขบวนการเปลี่ยนน้ำตาลในเลือดให้เป็นพลังงาน โดยขบวนการนี้เกี่ยวข้องกับอินซูลินซึ่งเป็นฮอร์โมนที่สร้างจากตับอ่อนเพื่อใช้ควบคุมระดับน้ำตาลในเลือด เมื่อน้ำตาลไม่ได้ถูกใช้จึงทำให้ระดับน้ำตาลในเลือดสูงขึ้นกว่าระดับปกติ โรคเบาหวานแบ่งเป็น 4 ชนิด ตามสาเหตุของการเกิดโรค

1. โรคเบาหวานชนิดที่ 1 (type 1 diabetes mellitus, T1DM) เกิดจากเซลล์ตับอ่อนถูกทำลายจากภูมิคุ้มกันของร่างกาย ทำให้ขาดอินซูลิน มักพบในเด็ก

2. โรคเบาหวานชนิดที่ 2 (type 2 diabetes mellitus, T1DM) เป็นชนิดที่พบบ่อยที่สุด ร้อยละ 95 ของผู้ป่วยเบาหวานทั้งหมด เกิดจากภาวะดื้อต่ออินซูลิน มักพบในผู้ใหญ่ที่มีน้ำหนักเกินหรืออ้วนร่วมด้วย

3. โรคเบาหวานขณะตั้งครรภ์ (gestational diabetes mellitus, GDM) เป็นโรคเบาหวานที่เกิดขึ้นขณะตั้งครรภ์ มักเกิดเมื่อไตรมาส 2-3 ของการตั้งครรภ์

4. โรคเบาหวานที่มีสาเหตุจำเพาะ (specific types of diabetes due to other causes) มีได้หลายสาเหตุ เช่น โรคทางพันธุกรรม โรคของตับอ่อน โรคทางต่อมไร้ท่อ ยาบางชนิด เป็นต้น

2.1.3 โรคไฮเปอร์ไทรอยด์

โรคไฮเปอร์ไทรอยด์ หมายถึง ภาวะที่ต่อมไทรอยด์ มีการหลั่งฮอร์โมนไทรอยด์ออกมามากเกินไป กระตุ้นให้อวัยวะทั่วร่างกายมีการเผาผลาญสูงกว่าปกติ เป็นสาเหตุทำให้เกิดอาการเจ็บป่วย ๆ ต่างขึ้นตามมา เช่น เหนื่อยง่าย ใจสั่น ขี้ร้อนง่าย เหงื่อออกมาก หงุดหงิด นอนไม่หลับ น้ำหนักตัวลดลงอย่างรวดเร็วแบบผิดปกติ เป็นต้น สาเหตุของไฮโปไทรอยด์ต่อมไทรอยด์ทำหน้าที่ผลิตฮอร์โมนไทรอยด์ ซึ่งเป็นฮอร์โมนที่ส่งผลต่อสุขภาพอย่างมาก รวมทั้งกระทบต่อกระบวนการเมตาบอลิซึม หากต่อมไทรอยด์ผลิตฮอร์โมนไทรอยด์ออกมาไม่เพียงพอ สามารถก่อให้เกิดภาวะขาดไทรอยด์ โดยอาจเกิดได้จากสาเหตุต่อไปนี้

1. โรคภูมิคุ้มกันทำลายตนเอง (Autoimmune Disease) โรคนี้คือการที่ระบบภูมิคุ้มกันในร่างกายผลิตแอนติบอดี้ขึ้นมาทำลายเนื้อเยื่อภายในร่างกายตัวเอง ที่พบได้บ่อยที่สุดคือ ไทรอยด์อักเสบฮาชิโมโต (Hashimoto's Thyroiditis) เมื่อต่อมไทรอยด์อักเสบและผลิตแอนติบอดีขึ้นมาทำลายเนื้อเยื่อของตัวเอง ย่อมส่งผลต่อการทำงานของต่อมไทรอยด์ที่ผลิตฮอร์โมนไทรอยด์ออกมาได้ไม่เพียงพอจนนำไปสู่ภาวะขาดไทรอยด์ ยังไม่ปรากฏหลักฐานชี้ชัดว่าเหตุใดร่างกายจึงผลิตสารภูมิคุ้มกันขึ้นมาทำลายเนื้อเยื่อภายในร่างกายตัวเอง โดยสันนิษฐานว่าอาจมาจากเชื้อไวรัส แบคทีเรีย หรือความผิดปกติของยีน อย่างไรก็ตาม โรคภูมิคุ้มกันทำลายตนเองเกิดจากหลายปัจจัยรวมกัน

2. การรักษาต่อมไทรอยด์เป็นพิษ (Hyperthyroidism) ต่อมไทรอยด์เป็นพิษหรือไฮเปอร์ไทรอยด์ (Hyperthyroidism) คือภาวะที่ร่างกายผลิตฮอร์โมนไทรอยด์มากเกินไป โดยผู้ป่วยโรคนี้อาจจะได้รับการรักษาด้วยการฉายรังสี (Radioactive Lodine) หรือยาต้านไทรอยด์เพื่อลดและปรับระดับฮอร์โมนดังกล่าวให้เป็นปกติ อย่างไรก็ดี รังสีที่ใช้รักษานั้นจะทำลายเซลล์ในต่อมไทรอยด์ จนนำไปสู่ภาวะขาดไทรอยด์ได้

3. การผ่าตัดต่อมไทรอยด์ การผ่าตัดนำส่วนต่าง ๆ เกือบทุกส่วนของต่อมไทรอยด์ออกไปนั้นจะทำให้ร่างกายลดหรือหยุดการผลิตฮอร์โมน หากเกิดกรณีดังกล่าวขึ้น ผู้ป่วยจำเป็นต้องรับการให้ฮอร์โมนไทรอยด์ไปตลอดชีวิต

4. การรักษาด้วยรังสี ผู้ป่วยมะเร็งบางชนิดต้องได้รับการรักษาด้วยการฉายรังสีที่คอหรือหัว โดยรังสีจะทำลายเซลล์ภายในต่อมไทรอยด์ ทำให้ไม่สามารถผลิตฮอร์โมนออกมาได้เพียงพอ

2.1.4 เหมืองข้อมูล

การวิเคราะห์ข้อมูลจากข้อมูลจำนวนมาก เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกประเภท รูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กัน และหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ มีขั้นตอนการทำงานดังนี้

1. การเตรียมข้อมูล คือ โดยจะเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่ต้องการวิเคราะห์ ทำ การลบข้อมูลที่ซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด แปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์

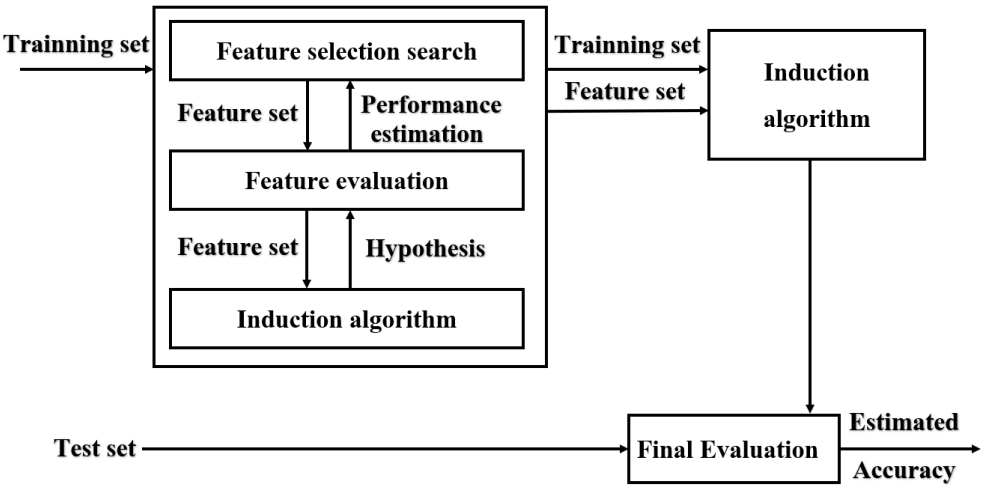
2. ขั้นตอนก่อนการสร้างแบบจำลอง เป็นขั้นตอนเลือกข้อมูล ทำความสะอาดข้อมูลที่ไม่สมบูรณ รวมถึงการเลือกตัวแปรที่เป็นอิสระจากตัวแปรตาม

3. การสร้างแบบจำลอง คือ กระบวนการในการนำเทคนิคต่าง ๆ ในเหมืองข้อมูลมาทำการวิเคราะห์ข้อมูล เพื่อหารูปแบบในการจำแนก เพื่อใช้ในการพยากรณ์

4. การวัดประสิทธิภาพของแบบจำลอง คือ การทดลองเพื่อหาค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (specificity)

2.1.5 การคัดเลือกตัวแปร

การคัดเลือกตัวแปรด้วยวิธีแรปเปอร์ (Wrapper) เป็นวิธีคัดเลือกตัวแปรที่สำคัญ โดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มาสร้างเซตของตัวแปรใหม่โดยการเพิ่มหรือลดจำนวนตัวแปรจากเซตเดิม อาศัยขั้นตอนวิธีการเรียนรู้ในการประเมินค่าชุดข้อมูลฝึกฝน ซึ่งจะทำให้ได้ชุดข้อมูลฝึกฝนที่มีความแม่นยำในการจำแนกประเภทมากกว่าการใช้ค่าจากมาตรวัดอื่นในการประเมินค่า ขั้นตอนการทำงาน ดังภาพที่ 2.1 ขั้นที่ 1 เป็นการคัดเลือกชุดข้อมูลฝึกฝนของคุณลักษณะซึ่งเลือกชุดข้อมูลฝึกฝนที่ดีที่สุดโดยดูจากความแม่นยำของตัวจำแนกประเภท ขั้นที่ 2 เป็นการเรียนรู้และการทดสอบ โดยนำชุดข้อมูลฝึกฝนที่ดีที่สุดที่ได้จากการคัดเลือกคุณลักษณะในขั้นแรกมาเรียนรู้เพื่อสร้างตัวแบบบนข้อมูลฝึกฝน และทำการทดสอบตัวแบบที่ได้บนข้อมูลทดสอบ เมื่อแต่ละเซตย่อยถูกสร้าง ตัวแบบจะถูกสร้างจากข้อมูลของชุดข้อมูลฝึกฝนนั้น และคำนวณหาค่าประมาณความถูกต้อง [6] [7] [8]



ภาพที่ 2.1 การทำงานของ Wrapper

เทคนิค Gain Ratio เป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยที่ พัฒนามาจาก Information Gain เนื่องจากการใช้ Information Gain ในการแบ่งชุดข้อมูลจะมีโอกาสทำให้เกิดความ เอนเอียงขึ้น เมื่อคุณลักษณะที่ทำการพิจารณาได้ค่า gain ที่สูงเป็นจำนวนมาก ทำให้คุณลักษณะที่ ถูกคัดเลือกไม่ถูกต้อง ตัวอย่างเช่นพิจารณาคุณลักษณะที่ทำหน้าที่เป็นตัวระบุเฉพาะ เช่นรหัส ผลิตภัณฑ์ การแยกรหัสผลิตภัณฑ์จะส่งผลให้มีชุดข้อมูลย่อยจำนวนมาก แต่ละชุดข้อมูลย่อยมีเพียง หนึ่ง record ซึ่งเมื่อนำมาหาค่า Information Gain จะได้ค่าที่สูงจำนวนมากนั่นเอง จากความเอนเอียง ทำให้มีการพัฒนาตัวชี้วัดการแบ่งข้อมูลใหม่ที่ชื่อเรียกว่า Gain Ratio โดย การประยุกต์ใช้การทำนอร์มัลไลซ์ค่า Information Gain ด้วยการใช้ค่า Split Information ซึ่ง สามารถคำนวณได้ดังสมการที่ 2.1

$$\text{SplitInfo}_A(\text{D}) = -\sum_{j=1}^{m} \frac{|D_j|}{|D|} \text{x} \log_2 \frac{|D_j|}{|D|} \quad (2.1)$$

โดยค่า $SplitInfo_A(D)$ หมายถึงปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งข้อมูลในชุดข้อมูล *D* ออกเป็น *m* ชุดข้อมูลย่อยตามค่าคุณลักษณะ A โดยหลังจากทำการคำนวณหาค่า $SplitInfo_A(D)$ แล้วเราจะสามารถคำนวณหาค่า *Gain Ratio* ได้ดังสมการ *2.2*

$$\text{GainRatio(A)} = \frac{\text{Gain(A)}}{\text{SplitInfo(A)}} \quad (2.2)$$

2.1.6 เทคนิคเหมืองข้อมูลที่ใช้ในงานวิจัย

2.1.6.1 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นวิธีหนึ่งที่จะประมาณฟังก์ชันที่มีค่าที่ไม่ต่อเนื่อง (discrete-value function) ด้วย แผนผังต้นไม้ อาจประกอบด้วยเซตของกฎต่าง ๆแบบถ้า-แล้ว (if-then) เพื่อให้มนุษย์สามารถอ่านแล้วเข้าใจการตัดสินใจของต้นไม้ได้ ข้อดีอย่างหนึ่งของการใช้ Decision Tree คือเลือกตัวแปรที่มีความสำคัญที่ช่วยแบ่งแยกข้อมูลออกมาได้ ตัวอย่างงานวิจัย ชณิดาภา บุญประสม ได้ศึกษาตัวแปรที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิควิธี Decision Tree, K-Nearest Neighbors, Naive Bayes โดยใช้ข้อมูลจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานีของนักศึกษาระดับปริญญาตรี มีจำนวน 11 แอททริบิวต์และ 13,729 ชุดข้อมูล และได้วิเคราะห์ค่าน้ำหนักของแอททริบิวต์ ด้วยวิธีการ Information Theory ซึ่งมีตัวแปรที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาจำนวน 8 ตัวแปร เพื่อนำตัวแปรที่ได้มาทำการสร้างเป็นโมเดลทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วย ค่า Accuracy เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลพบว่าโมเดลที่ Decision Tree มีค่าเฉลี่ยความถูกต้อง 93.52 % [9] [10] [11]

2.1.6.2 เทคนิคการจำแนกข้อมูลแบบจำลองเบย์

เทคนิคการจำแนกข้อมูลแบบจำลองเบย์ (Bayesian) เป็นการจำแนกประเภทโดยใช้กฎของเบย์หรือเป็นการจำแนกประเภทโดยใช้หลักสถิติในการพยากรณ์ความน่าจะเป็นของสมาชิกเรียกว่า ทฤษฎีของเบย์ (Bayesian theorem) เป็นการเรียนรู้เพิ่มได้ตัวอย่างใหม่ที่ได้มาถูกนามาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มลดความน่าจะเป็นทำให้มีการเรียนรู้ที่เปลี่ยนไปวิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มีการทำนายค่าคลาสเป้าหมายของตัวอย่างใช้ความน่าจะเป็นมากที่สุดของทุกสมมติฐาน ตัวอย่างงานวิจัย Raheela Asif ,Agathe Merceron ,Syed Abbas Ali ,Najmi Ghani Haider [12] ได้วิเคราะห์ประสิทธิภาพของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิค Decision Tree with Gini Index , Decision Tree with Information Gain , Decision Tree with Accuracy , Rule Induction with Information Gain , 1-Nearest Neighbour , Naive Bayes , Neural Networks, Random Forest Trees with Gini Index, Random Forest Trees with Accuracy พบว่าตัวแปรที่มีผลต่อการสำเร็จการศึกษา คือเกรดเฉลี่ยของวิชาที่สำคัญ เช่น คณิตศาสตร์,ฟิสิกส์,เคมี โดยการใช้เครื่องมือเหมืองข้อมูลแบบจำแนกกลุ่ม โดยใช้ขั้นตอนในการดำเนินงานการทำเหมืองข้อมูล แบบจำแนกกลุ่ม พบว่า Naïve Bayes ให้ประสิทธิภาพสูงสุดที่ 83.65 % และ Maninder Kaur, Akshay Girdhar ได้วิเคราะห์ประสิทธิภาพ

ของการสอนรูปแบบต่างๆ โดยใช้เทคนิค OneR, ZeroR, J48, IBK, Naïve Bayes วิธีการสอนที่นำว่าวิเคราะห์ คือ ครูสอนในห้องเรียน, ให้นักเรียนคิดวิเคราะห์แก้ปัญหาด้วยตัวเอง โดยใช้ขั้นตอนในการดำเนินงานการทำเหมืองข้อมูล พบว่า Naïve Bayes ให้ประสิทธิภาพสูงสุดที่ 100 % [13] [14]

2.1.6.3 เทคนิคโครงข่ายประสาทเทียม

เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ หลักการสำคัญของโครงข่าย ประสาทเทียม คือ ความพยายามที่จะลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์เพื่อทำงานได้อย่าง มีประสิทธิภาพ ลักษณะทั่วไปของโครงข่ายประสาทเทียม คือ การที่โหนด (node) ต่าง ๆ จำลองมาจากไซแนปส์ (synapse) ของเซลล์ประสาทระหว่าง เดนไดรต์ (dendrite) และแอกซอน (axon) โดยมีฟังก์ชันเป็นตัวกำหนด สัญญาณส่งออก (activation function or transfer function) นั่นเอง ลักษณะของโครงข่ายประสาทเทียม สามารถแบ่งได้ 2 แบบ คือ 1) โครงข่ายประสาทเทียมแบบ ชั้นเดียว (single layer) ซึ่งจะมีเพียงชั้นสัญญาณ ประสาทขาเข้า และชั้นสัญญาณประสาทขาออก เท่านั้น เช่น โครงข่ายเพอเซบตรอนอย่างง่าย (simple perceptron) และโครงข่ายโฮปฟิลด์ (Hopfield networks) เป็นต้น และ 2) โครงข่ายประสาทเทียมแบบหลายชั้น (multilayer) ซึ่งมีลักษณะเช่นเดียวกับโครงข่ายประสาทเทียมแบบชั้นเดียว แต่จะมีชั้นแอบแฝง (hidden) เพิ่มขึ้น โดยอยู่ส่วนกลางระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งขอมูลออกทั้งนี้ชั้นแอบแฝงอาจมีมากกว่า 1 ชั้น ตัวอย่างงานวิจัย เสกสรรค์ วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์ และดวงดาว วิชาดากุล ศึกษาการใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียน โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกาแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา ใช้กระบวนการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งใช้วิธี Correlation-based FeatureSelection (CFS) และวิธี Information Gain (IG) แล้วใช้เทคนิคเหมืองข้อมูลแบบโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน (MLP) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และต้นไม้ตัดสินใจ (Decision Tree) มาสร้างตัวแบบพยากรณ์และเปรียบเทียบตัวแบบ ด้วยการทดสอบประสิทธิภาพแบบ 10-Fold Cross Validation [15] ทิพย์หทัย ทองธรรมชาติ ศึกษาการสร้างโมเดลสำหรับพยากรณ์ผลการเรียนของนักศึกษา ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร และสาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด ด้วยเทคนิคโครงข่ายประสาทเทียมและเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ผลการเปรียบเทียบพบว่า เทคนิคโครงข่ายประสาทเทียม ให้ค่าความถูกต้อง ร้อยละ 85.71 ซึ่งมากกว่าเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ซึ่งมีค่าร้อยละ 76.62 [16]

2.1.6.4 เทคนิคต้นไม้ป่าสุ่ม

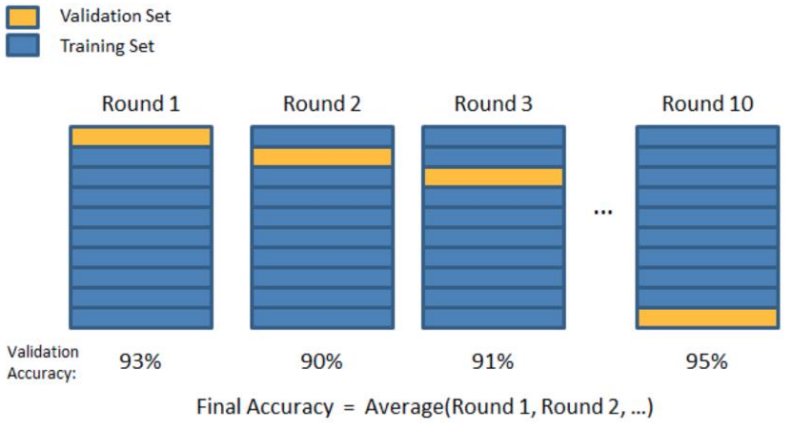
เทคนิคต้นไม้ป่าสุ่ม (Random Forest) เป็นเทคนิคที่สร้างแบบจำลองเป็นต้นไม้ตัดสินใจหลายๆ ต้น โดยการสุ่มเลือกแอตทริบิวต์ ที่สำคัญออกมาเป็นหลายๆ ชุดแล้วนำไปสร้างต้นไม้ตัดสินใจ โดยมีข้อดีคือ ช่วยในการลดค่า correlation ระหว่าง Tree ได้ เนื่องจาก function Random Attribute Subsets [17] [18]

2.1.6.5 เทคนิคการเรียนรู้เชิงลึก

เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นส่วนหนึ่งของวิธีการการเรียนรู้ของเครื่องบนพื้นฐานของโครงข่ายประสาทเทียมและการเรียนเชิงคุณลักษณะ การเรียนรู้สามารถเป็นได้ทั้งแบบการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอน [19] [20] [21]

2.1.7 การวัดประสิทธิภาพแบบจำลอง

ในการวัดแบบจำลองในการทำเหมืองข้อมูลโดยส่วนใหญ่ใช้ 10-fold cross validation คือ การเลือกสุ่มข้อมูลแบบความเที่ยงตรง ซึ่งเป็นวิธีที่นิยมในการทำงานวิจัยเพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลอง เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัด ประสิทธิภาพด้วยวิธี 10-fold cross-validation จะทำการเลือกสุ่มข้อมูลออกเป็น 10 ชุดเท่าๆ กัน จากนั้นจะทำการทดลองครั้งแรกด้วยข้อมูลชุดที่ 1 ซึ่งเป็นข้อมูลทดสอบและกำหนดให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั้งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ โดยวิธีการทดสอบประสิทธิภาพแบบ 10-fold cross validationมีข้อเสียคือ จะต้องทำการเริ่มทดสอบใหม่โดยจะต้องทำทั้งหมด 10 รอบ ดังภาพที่ 2.2



ภาพที่ 2.2 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

ในการวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (specificity) ดังสมการ

2.1 ค่าความถูกต้อง (Accuracy) คือ ค่าที่แบบจำลองสามารถจำแนกข้อมูลผู้ป่วยที่เกิดโรค และไม่เกิดโรคได้อย่างถูกต้องต่อข้อมูลทั้งหมด ดังสมการที่ 2.3

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.3}$$

2.2 ค่าความไว (Sensitivity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่เกิดโรค ได้อย่างถูกต้องต่อผู้ป่วยที่เกิดโรคจริง ดังสมการที่ 2.4

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.4}$$

2.3 ค่าความจำเพาะ (Specificity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่ไม่เกิดโรค ได้อย่างถูกต้องต่อผู้ป่วยที่พยากรณ์ว่าเกิดโรค ดังสมการที่ 2.5

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

เมื่อ TP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้อย่างถูกต้อง

TN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การไม่เกิดโรคได้อย่างถูกต้อง

FP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้ไม่ถูกต้อง

FN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การไม่เกิดโรคได้ไม่ถูกต้อง

2.2 งานวิจัยที่เกี่ยวข้อง

Fan, Zhu และ Yin [22] ได้ศึกษาการทำนายการกลับเป็นซ้ำของมะเร็งเต้านม ด้วยเทคนิค C5.0, CHAID, QUEST, C&RT, ANN พบว่า เทคนิค C5.0 มีประสิทธิภาพที่ดีที่สุด ที่ 71.17%

Balpande และ Wajgi [23] ได้ศึกษาการคาดคะเนและการประมาณความรุนแรงของโรคเบาหวานโดยใช้เทคนิคการขุดข้อมูล ด้วยเทคนิค CHAID, Naïve Bayes, K-Nearest, Decision Tree ตัวแปรที่มีผลต่อการคาดคะเน คือ Age, Gender, BMI พบว่า เทคนิค Decision Tree มีประสิทธิภาพที่ดีที่สุด ที่ 72%

Mousavi, Somayeh, Zanjireh, Morteza และ Marzieh [24] ได้ศึกษาการจำแนกประเภทเชิงคำนวณเพื่อวินิจฉัยภาวะไทรอยด์ทำงานผิดปกติของทารก ด้วยเทคนิค CHAID, ID3, MLP, SVM พบว่า เทคนิค SVM มีประสิทธิภาพที่ดีที่สุด ที่ 99.58 %

Ojha และ Goel [25] ได้ศึกษาเกี่ยวกับการทำนายการกลับเป็นซ้ำของมะเร็งเต้านมโดย ใช้เทคนิคการขุดข้อมูล ด้วยเทคนิค C4.5, KNN, Naïve Bayes, SVM, พบว่า เทคนิค C4.5 และ SVM มีประสิทธิภาพที่ดีที่สุดเท่ากันที่ 81.03%

นุ้ยเพียร และ มีสัจ [26] ได้ศึกษาการเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรอง (Filter Approach) และ การควบรวม (Wrapper Approach) ของการทำเหมืองข้อความเพื่อการจำแนกข้อความ สรุปได้ว่าใช้วิธีการจำแนกประเภทแบบซัพพอร์ตเวกเตอร์แมชชีน โดยใช้เคอร์เนลฟังชันเรเดียลเบสิสฟังก์ชัน (SVMR) ให้ผลการวัดประสิทธิภาพโดยรวมสูงที่สุดคือ 92.2% นาอีฟเบย์ 91.7% และ เบย์เซียนเน็ต 91.4% ตามลำดับ

ศรีเปารยะ และ สินสมบูรณ์ทอง [27] ได้ศึกษาปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุดวิธีต้นไม้ตัดสินใจวิธีโครงข่ายประสาทเทียมวิธีซัพพอร์ตเวกเตอร์แมชชีนวิธีฐานกฎวิธีการถดถอยลอจิสติกและวิธีนาอีฟเบย์เพื่อวัดประสิทธิภาพการจำแนกกลุ่มโดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโลประเทศอินเดีย พบว่า วิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้องคือ 100%

บทที่ 3
วิธีการดำเนินการวิจัย

ในบทนี้วิธีการดำเนินงานวิจัยประกอบด้วย 4 ขั้นตอนคือ การเตรียมข้อมูล ขั้นตอนก่อนการสร้างแบบจำลอง การสร้างแบบจำลอง และการวัดประสิทธิภาพของแบบจำลอง

3.1 การเตรียมข้อมูล

การเตรียมข้อมูลในงานวิจัยนี้ข้อมูลประกอบด้วยชุดข้อมูลที่มีชนิดตัวแปรที่แตกต่างกันจำนวน 3 ชุดข้อมูล คือ ชุดที่ 1 คือโรคมะเร็งเต้านม มีชนิดตัวแปรเป็น Nominal ทั้งหมด จำนวน 10 ตัวแปร ชุดที่ 2 คือโรคเบาหวาน มีชนิดตัวแปรเป็น Numeric ทั้งหมด จำนวน 9 ตัวแปร และชุดที่ 3 โรคไฮโปไทรอยด์ มีชนิดตัวแปรเป็น Nominal และ Numeric จำนวน 30 ตัวแปร ดังตารางต่อไปนี้

ตารางที่ 3.1 ตัวแปรที่ใช้ในงานวิจัย โรคมะเร็งเต้านม

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
1	age	อายุ	Nominal
2	menopause	สตรีวัยหมดประจำเดือน	Nominal
3	tumor-size	ขนาดของเนื้องอก	Nominal
4	inv-nodes	ช่วงของมะเร็ง	Nominal
5	node-caps	การกระจายตัวของมะเร็ง	Nominal
6	deg-malig	ระดับความร้ายแรง	Nominal
7	breast	ตำแหน่งของมะเร็ง	Nominal
8	breast-quad	ตำแหน่งของมะเร็ง	Nominal
9	Irradiat	การฉายรังสี	Nominal
10	class	เหตุการณ์	Nominal

ตารางที่ 3.2 ตัวแปรที่ใช้ในงานวิจัย โรคเบาหวาน

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
1	preg	จำนวนครั้งที่ตั้งครรภ์	Numeric
2	plas	ระดับน้ำตาลในเลือด	Numeric
3	pres	ความดันโลหิต	Numeric
4	skin	ความหนาของไขมันใต้ผิวหนังกล้ามแขนส่วนหลัง	Numeric
5	insu	ปริมาณอินซูลินที่ได้รับภายใน 2 ชั่วโมง	Numeric
6	mass	ดัชนีมวลกาย	Numeric
7	pedi	การติดต่อทางสายเลือด	Numeric
8	age	อายุ	Numeric
9	class	ตัวแปรคลาส	Numeric

ตารางที่ 3.3 ตัวแปรที่ใช้ในงานวิจัย โรคไฮโปไทรอยด์

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
1	age	อายุ	Numeric
2	sex	เพศ	Nominal
3	on thyroxine	ฮอร์โมนที่หลั่งออกมาจากต่อมไทรอยด์	Nominal
4	query on thyroxine	คำถามของฮอร์โมนที่หลั่งออกมาจากต่อมไทรอยด์	Nominal
5	on antithyroid medicate	การใช้ยาต้านไทรอยด์	Nominal
6	sick	อาการไข้	Nominal
7	pregnant	การตั้งครรภ์	Nominal
8	thyroid surgery	การตัดต่อมไทรอยด์บางส่วน	Nominal
9	I131 treatment	การรักษาด้วยแร่ไอโอดีน 131	Nominal
10	query hypothyroid	โรคไทรอยด์ชนิดอ้วน	Nominal

ตารางที่ 3.4 ตัวแปรที่ใช้ในงานวิจัย โรคไฮโปไทรอยด์

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
11	query hyperthyroid	ภาวะที่ต่อมไทรอยด์สร้างและปล่อยฮอร์โมนออกมามากเกินไป	Nominal
12	lithium	การใช้ยาทางจิตเวชที่ใช้รักษาอาการคลุ้มคลั่ง	Nominal
13	goitre	ภาวะที่มีการโตขึ้นของต่อมไทรอยด์	Nominal
14	tumor	เนื้องอก	Nominal
15	hypopituitary	ภาวะที่ร่างกายขาดฮอร์โมนที่สร้างจากต่อมใต้สมอง	Nominal
16	psych	อัตราชีพจร	Nominal
17	TSH measured	ค่าของฮอร์โมนกระตุ้นต่อมไทรอยด์	Nominal
18	TSH	ฮอร์โมนกระตุ้นต่อมไทรอยด์	Numeric
19	T3 measured	ค่าของไทรอยด์ฮอร์โมนที่ได้จากสองแหล่ง	Nominal
20	T3	ไทรอยด์ฮอร์โมนที่ได้จากสองแหล่ง	Numeric
21	TT4 measured	ค่าของฮอร์โมนทั้งส่วนที่จับกับโปรตีน	Nominal
22	TT4	ฮอร์โมนทั้งส่วนที่จับกับโปรตีน	Numeric
23	T4U measured	ค่าของการใช้ไทรอกซิน	Nominal
24	T4U	การใช้ไทรอกซิน	Numeric
25	FTI measured	ค่าของฮอร์โมนไทรอกซินชนิดอิสระ	Nominal
26	FTI	ฮอร์โมนไทรอกซินชนิดอิสระ	Numeric

ตารางที่ 3.5 ตัวแปรที่ใช้ในงานวิจัย โรคไฮโปไทรอยด์

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
27	TBG measured	ค่าของโปรตีนที่สร้างจากตับ	Nominal
28	TBG	โปรตีนที่สร้างจากตับ	Numeric
29	referral source	แหล่งอ้างอิง	Nominal
30	Class	ตัวแปรคลาส	Nominal

3.2 ขั้นตอนก่อนการสร้างแบบจำลอง

ในขั้นตอนก่อนการสร้างแบบจำลองผู้วิจัยได้ทำการปรับเปลี่ยนข้อมูล เนื่องจากข้อมูลทั้งเป็นตัวเลขและข้อมูลที่เป็นตัวอักษรที่ไม่อยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ได้จึงได้ทำการแทนค่าข้อมูลให้อยู่ในรูปแบบที่สามารถนำมาวิเคราะห์ได้และทำการเลือกตัวแปรต้นที่เป็นอิสระต่อกัน แต่มีความสัมพันธ์กับตัวแปรตาม (คลาส) โดยใช้หลักการ Wrapper และหลักการของ Gain Ratio ได้ดังตาราง

ตารางที่ 3.6 ผลการคัดเลือกตัวแปรของโรคมะเร็งเต้านม

Wrapper					Gain Ratio
Random Forest	Decision Tree C4.5	Naïve bayes	Neural Networks	Deep Learning	
Node-caps	Node-caps	Node-caps	Node-caps	menopause	Deg-malig
Deg-malig	Deg-malig	Deg-malig	Deg-malig	Node-caps	Inv-nodes
	breast			Deg-malig	Node-caps
				Breast-quad	Tumor-size
				irradiat	irradiat
					age
					Breast-quad
					Breast
					menopause

จากตารางที่ 3.4 แสดงผลการคัดเลือกตัวแปรของโรคมะเร็งเต้านม เป็นการคัดเลือกตัวแปรด้วยหลักการ Wrapper เป็นการคัดเลือกตัวแปรโดยใช้เทคนิค Random Forest เทคนิค Decision Tree C4.5 เทคนิค Naïve bayes เทคนิค Neural Networks และ เทคนิค Deep Learning จากข้อมูลเดิมมีตัวแปรอยู่ที่ 10 ตัวแปร เหลือ 2 ตัวแปร 3 ตัวแปร 2 ตัวแปร 2 ตัวแปร และ 5 ตัวแปร ตามลำดับ ส่วน หลักการของ Gain Ratio เหลือตัวแปรอยู่ที่ 9 ตัว

ตารางที่ 3.7 ผลการคัดเลือกตัวแปรของโรคเบาหวาน

Wrapper					Gain Ratio
Random Forest	Decision Tree C4.5	Naïve bayes	Neural Networks	Deep Learning	
Preg	age	Plas	age	Plas	Plas
Plas	Plas	Pres	Plas	Skin	Age
Pres	Pres	Mass	Pres	Mass	Mass
Insu	Mass	pedi	pedi	pedi	Insu
Mass		preg	Mass		Skin
Pedi					Preg
age					Pedi
					pres

จากตารางที่ 3.5 แสดงผลการคัดเลือกตัวแปรของโรคเบาหวาน เป็นการคัดเลือกตัวแปรด้วยหลักการ Wrapper เป็นการคัดเลือกตัวแปรโดยใช้เทคนิค Random Forest เทคนิค Decision Tree C4.5 เทคนิค Naïve bayes เทคนิค Neural Networks และ เทคนิค Deep Learning จากข้อมูลเดิมมีตัวแปรอยู่ที่ 9 ตัวแปร เหลือ 7 ตัวแปร 4 ตัวแปร 5 ตัวแปร 5 ตัวแปร และ 4 ตัวแปร ตามลำดับ ส่วน หลักการของ Gain Ratio เหลือตัวแปรอยู่ที่ 8 ตัว

ตารางที่ 3.8 ผลการคัดเลือกตัวแปรของโรคไฮโปไทรอยด์

Wrapper					Gain Ratio
Random Forest	Decision Tree C4.5	Naïve bayes	Neural Networks	Deep Learning	
Age	On thyroxine	Age	Age	age	TSH
On thyroxine	Thyroid surgery	On thyroxine	On thyroxine	On thyroxine	FTI
Thyroid surgery	TSH measured	Sick	Sick	Query on thyroxine	TT4
TSH measured	TSH	Thyroid surgery	Thyroid surgery	Sick	On thyroxine
TSH	TT4	Query hypothyroid	Query hypothyroid	Thyroid surgery	TSH measured
TT4 measured		TSH measured	TSH measured	Query hypothyroid	Query hypothyroid
TT4		TSH	TSH	Tumor	T3
		T3	T3	TSH measured	TT4 measured
		TT4	TT4 measured	TSH	Sex
		FTI	TT4	T3	Referral source
			FTI	TT4 measured	Pregnant
				TT4	Thyroid surgery
				FTI	Goitre
					Sick
					T4U measured

ตารางที่ 3.9 ผลการคัดเลือกตัวแปรของโรคไฮโปไทรอยด์

Wrapper					Gain Ratio
Random Forest	Decision Tree C4.5	Naïve bayes	Neural Networks	Deep Learning	
					On antithyroid medication
					FTI measured
					Psych
					T3 measured
					Tumor
					Hypopituitary
					Query on thyroxine
					Query hyperthyroid
					Lithium
					I131 treatment
					TBG
					T4U
					TBG measured
					age

จากตารางที่ 3.6 แสดงผลการคัดเลือกตัวแปรของโรคไฮโปไทรอยด์ เป็นการคัดเลือกตัวแปรด้วยหลักการ Wrapper เป็นการคัดเลือกตัวแปรโดยใช้เทคนิค Random Forest เทคนิค Decision Tree C4.5 เทคนิค Naïve bayes เทคนิค Neural Networks และ เทคนิค Deep Learning จากข้อมูลเดิมมีตัวแปรอยู่ที่ 30 ตัวแปร เหลือ 2 ตัวแปร 3 ตัวแปร 2 ตัวแปร 2 ตัวแปร และ 5 ตัวแปร ตามลำดับ ส่วน หลักการของ Gain Ratio เหลือตัวแปรอยู่ที่ 29 ตัว

3.3 การสร้างแบบจำลอง

ทำการสร้างแบบจำลองเพื่อใช้ในการพยากรณ์ โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning

3.4 การวัดประสิทธิภาพของแบบจำลอง

เมื่อทำการสร้างแบบจำลองเสร็จแล้วนำแบบบจำลองมาทดสอบประเมินประสิทธิภาพด้วยวิธีการของ 10-fold cross validation โดยการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่าๆ กันและทำการเปรียบเทียบค่าด้วยการจำแนกกลุ่มข้อมูล คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity)

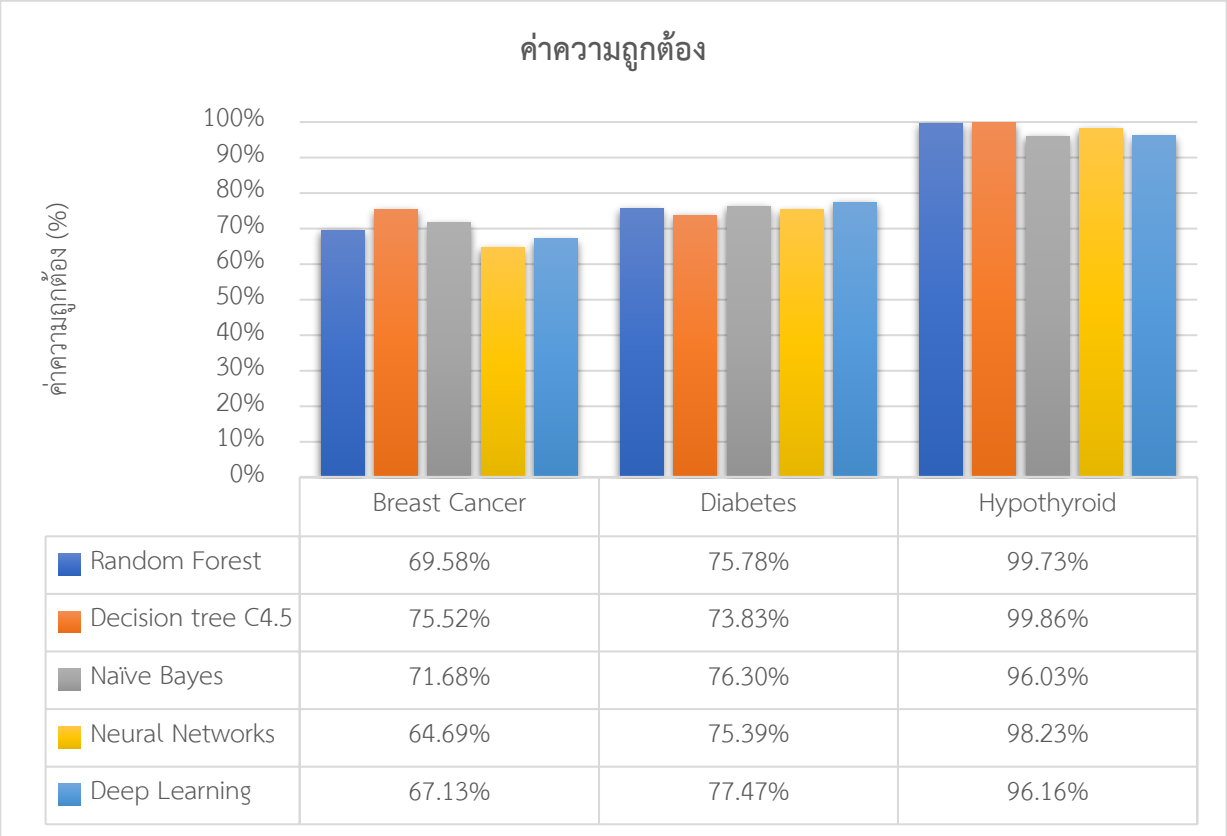
บทที่ 4
ผลการดำเนินการวิจัย

ผลการดำเนินการวิจัยประกอบด้วย ประสิทธิภาพของแบบจำลองก่อนการคัดเลือกตัวแปร ประสิทธิภาพของแบบจำลองหลังการคัดเลือกตัวแปร และการเพิ่มประสิทธิภาพแบบจำลองด้วยเทคนิค Decision Tree, Naive Bayes, Neural Networks, Random Forest, Deep Learning โดยในการทดทดลองแต่ละครั้งได้ใช้หลักการ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดฝึกสอนและชุดทดสอบโดยใช้ชุดฝึกในการสร้างแบบจำลองและชุดทดสอบนำมาทดสอบแบบจำลองด้วยค่าความถูกต้อง ค่าความไว และค่าจำเพาะ

4.1 ประสิทธิภาพของแบบจำลองก่อนการคัดเลือกตัวแปร

ประสิทธิภาพของแบบจำลองก่อนการคัดเลือกตัวแปร จากการทดลองโดยทำการแยกข้อมูลออกเป็น 10 ชุดเท่า ๆ กันนำชุดข้อมูล 9 ชุดมาเป็นชุดฝึกสอน ส่วนชุดที่เหลือเป็นชุดทดสอบ และทำการทดสอบจำนวน 10 รอบโดยแต่ละรอบชุดทดสอบจะทดลองไม่ซ้ำกัน ทำให้การทดลองใช้ข้อมูลที่ไม่อยู่ในชุดฝึกสอนในชุดฝึกสอนมาทำการทดสอบหา ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ ดังภาพที่ 4.1 ภาพที่ 4.2 และ ภาพที่ 4.3 ตามลำดับ

ค่าความถูกต้องของแบบจำลองก่อนคัดเลือกตัวแปร

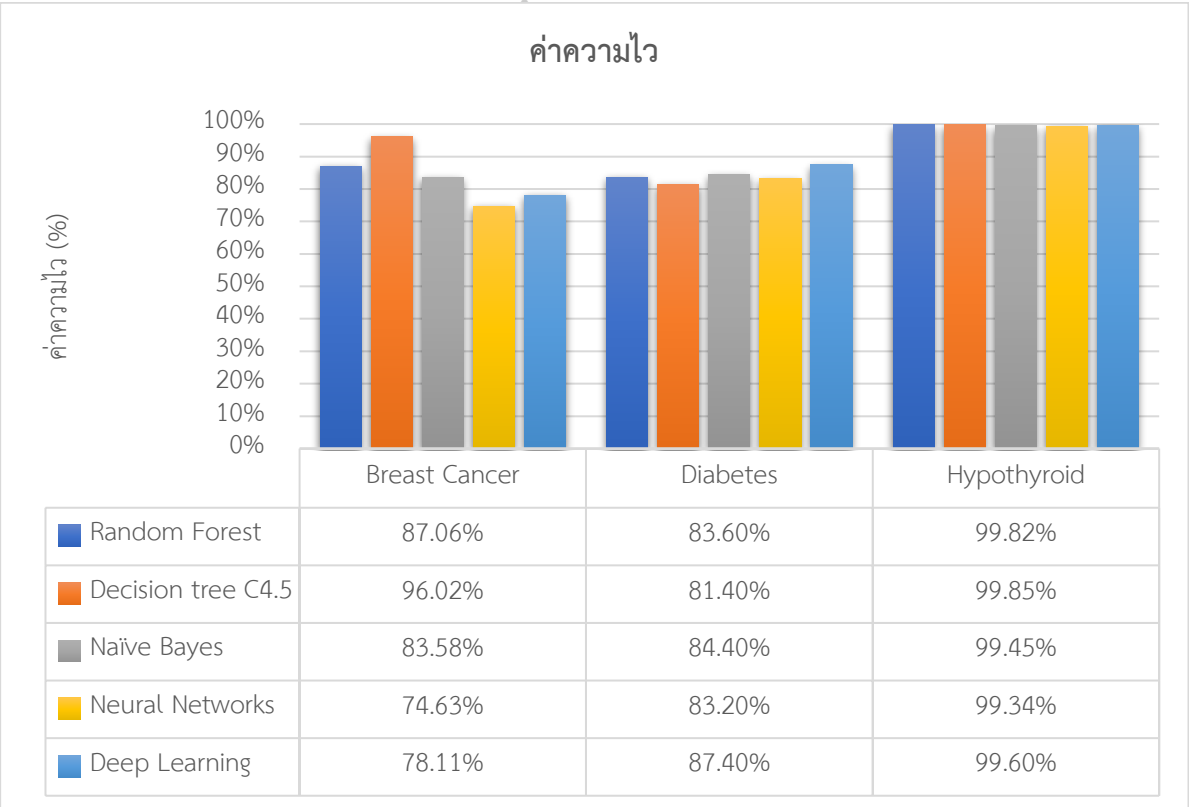


	Breast Cancer	Diabetes	Hypothyroid
Random Forest	69.58%	75.78%	99.73%
Decision tree C4.5	75.52%	73.83%	99.86%
Naïve Bayes	71.68%	76.30%	96.03%
Neural Networks	64.69%	75.39%	98.23%
Deep Learning	67.13%	77.47%	96.16%

ภาพที่ 4.1 ค่าความถูกต้องของแบบจำลองก่อนคัดเลือกตัวแปร

จากภาพที่ 4.1 แสดงค่าความถูกต้องก่อนคัดเลือกตัวแปร (Accuracy) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 75.52% และน้อยที่สุดใน เทคนิค Artificial Neural Networks ให้ค่าความถูกต้องที่ 64.69% ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 77.47% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าความถูกต้องที่ 73.83% และโรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 99.86% และน้อยที่สุดในเทคนิค Naïve Bayes ให้ค่าความถูกต้อง ที่ 96.03%

ค่าความไวก่อนคัดเลือกตัวแปร

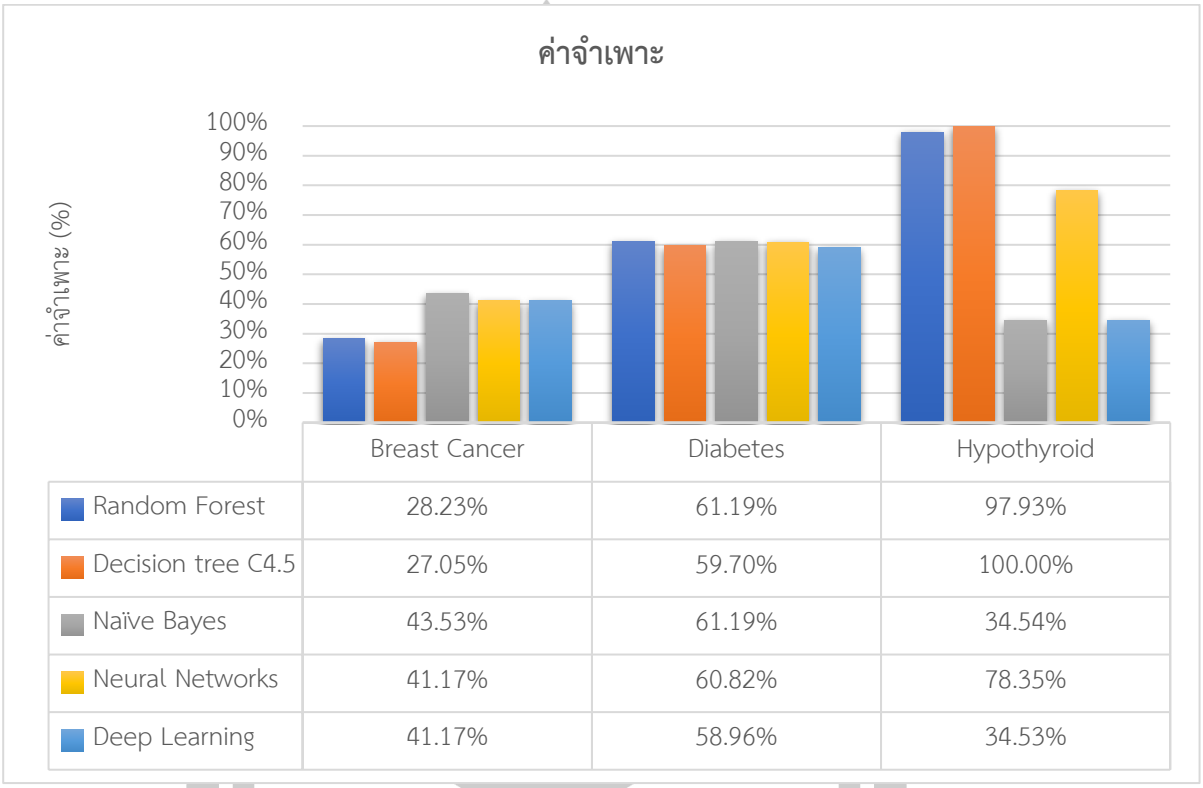


	Breast Cancer	Diabetes	Hypothyroid
Random Forest	87.06%	83.60%	99.82%
Decision tree C4.5	96.02%	81.40%	99.85%
Naïve Bayes	83.58%	84.40%	99.45%
Neural Networks	74.63%	83.20%	99.34%
Deep Learning	78.11%	87.40%	99.60%

ภาพที่ 4.2 ค่าความไวก่อนคัดเลือกตัวแปร

จากภาพที่ 4.2 แสดงค่าความไว (Sensitivity) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 96.02% และน้อยที่สุดในเทคนิค Artificial Neural Networks ให้ค่าความไวที่ 74.63% ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 87.40% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าความไวที่ 81.40% และ โรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 99.85% และน้อยที่สุดในเทคนิค Artificial Neural Networks ให้ค่าความไวที่ 99.34%

ค่าจำเพาะก่อนคัดเลือกตัวแปร



	Breast Cancer	Diabetes	Hypothyroid
Random Forest	28.23%	61.19%	97.93%
Decision tree C4.5	27.05%	59.70%	100.00%
Naïve Bayes	43.53%	61.19%	34.54%
Neural Networks	41.17%	60.82%	78.35%
Deep Learning	41.17%	58.96%	34.53%

ภาพที่ 4.3 ค่าจำเพาะก่อนคัดเลือกตัวแปร

จากภาพที่ 4.3 แสดงค่าจำเพาะก่อนคัดเลือกตัวแปร (Specificity) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Naïve Bayes ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดถึง 43.53% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะ ที่ 27.05% ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Random Forest กับ เทคนิค Naïve Bayes ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดเท่ากันที่ 61.19% และน้อยที่สุดในเทคนิค Deep Learning ให้ค่าจำเพาะที่ 58.96% และ โรคไฮโปไทรอยด์ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดถึง 100% และน้อยที่สุดในเทคนิค Deep Learning ให้ค่าจำเพาะที่ 34.53%

4.2 ประสิทธิภาพของแบบจำลองหลังการคัดเลือกตัวแปร

4.2.1 ผลค่าความถูกต้องหลังการคัดเลือกตัวแปร

ค่าความถูกต้องของแบบพยากรณ์การเกิดโรคมะเร็งเต้านม หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.1

ตารางที่ 4.1 ค่าความถูกต้องหลังคัดเลือกตัวแปร โรคมะเร็งเต้านม

วิธีการคัดเลือกตัวแปร	ค่าความถูกต้อง				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	69.58	75.52	71.68	64.69	67.13
Wrapper + Random Forest	75.17	75.17	75.17	75.87	70.97
Wrapper + Decision Tree C4.5	74.13	75.87	73.78	75.52	72.03
Wrapper + Naïve Bayes	75.17	75.17	75.17	75.87	70.97
Wrapper + Neural Networks	75.17	75.17	75.17	75.87	70.97
Wrapper + Deep Learning	69.58	75.87	74.48	72.03	73.78
Gain Ratio	75.87	76.57	75.17	69.58	75.52

จากตารางที่ 4.1 ค่าความถูกต้องหลังคัดเลือกตัวแปร ของโรคมะเร็งเต้านม โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Decision Tree C4.5 และ วิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Random Forest ให้ค่าความถูกต้องสูงที่สุดเท่ากันที่ร้อยละ 75.87 และให้ค่าความถูกต้องน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ที่ร้อยละ 69.58

ค่าความถูกต้องของแบบพยากรณ์การเกิดโรคเบาหวาน หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.2

ตารางที่ 4.2 ค่าความถูกต้องหลังคัดเลือกตัวแปร โรคเบาหวาน

วิธีการคัดเลือกตัวแปร	ค่าความถูกต้อง				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	75.78	73.83	73.60	75.39	77.47
Wrapper + Random Forest	76.17	73.96	76.82	75.65	77.34
Wrapper + Decision Tree C4.5	73.05	75.78	76.69	77.08	76.95

ตารางที่ 4.3 ค่าความถูกต้องหลังคัดเลือกตัวแปร โรคเบาหวาน

วิธีการคัดเลือกตัวแปร	ค่าความถูกต้อง				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
Wrapper + Naïve Bayes	74.22	74.74	77.73	75.26	77.18
Wrapper + Neural Networks	75.78	75.91	77.08	76.82	77.68
Wrapper + Deep Learning	72.79	74.61	76.56	75.00	77.21
Gain Ratio	73.83	75.78	76.30	75.39	77.47

จากตารางที่ 4.2 ค่าความถูกต้องหลังคัดเลือกตัวแปร ของโรคเบาหวาน โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคเบาหวาน ซึ่ง เป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Naïve Bayes ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 77.73 และให้ค่าความถูกต้องน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ที่ร้อยละ 72.79

ค่าความถูกต้องของแบบพยากรณ์การเกิดโรคไฮโปไทรอยด์หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.3

ตารางที่ 4.4 ค่าความถูกต้องหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์

วิธีการคัดเลือกตัวแปร	ค่าความถูกต้อง				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	99.73	99.86	96.03	98.23	96.16
Wrapper + Random Forest	99.89	99.86	95.59	98.59	95.95
Wrapper + Decision Tree C4.5	99.84	99.86	95.59	98.78	96.00
Wrapper + Naïve Bayes	99.86	99.86	95.97	98.56	96.08
Wrapper + Neural Networks	99.86	99.86	95.95	98.31	96.08
Wrapper + Deep Learning	99.89	99.86	95.97	98.50	96.11
Gain Ratio	99.73	99.86	96.03	98.23	96.16

จากตารางที่ 4.3 ค่าความถูกต้องหลังคัดเลือกตัวแปร ของโรคไฮโปไทรอยด์ โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ให้ค่าความถูกต้องสูงที่สุดเท่ากันที่ร้อยละ 99.89 และให้ค่าความถูกต้องน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Naïve Bayes และ Wrapper + Decision Tree C4.5 ด้วยเทคนิค Naïve Bayes เท่ากันที่ร้อยละ 95.59

4.2.2 ผลค่าความไวหลังการคัดเลือกตัวแปร

ค่าความไวของแบบพยากรณ์การเกิดโรคมะเร็งเต้านม หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.4

ตารางที่ 4.5 ค่าความไวหลังคัดเลือกตัวแปร โรคมะเร็งเต้านม

วิธีการคัดเลือกตัวแปร	ค่าความไว				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	87.06	96.02	83.58	74.63	78.11
Wrapper + Random Forest	95.52	95.52	95.52	95.02	80.59
Wrapper + Decision Tree C4.5	93.53	96.52	92.54	93.03	80.10
Wrapper + Naïve Bayes	95.52	95.52	95.52	95.02	80.59
Wrapper + Neural Networks	95.52	95.52	95.52	95.02	80.59
Wrapper + Deep Learning	85.57	96.52	88.06	87.06	84.08
Gain Ratio	27.94	27.94	36.76	30.88	41.18

จากตารางที่ 4.4 ค่าความไวต้องหลังคัดเลือกตัวแปร ของโรคมะเร็งเต้านม โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูล

โรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Decision Tree C4.5 และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Decision Tree C4.5 ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 96.52 และให้ค่าความไวน้อยที่สุดในวิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Random Forest และ Decision Tree C4.5 เท่ากันที่ร้อยละ 27.94

ค่าความไวของแบบพยากรณ์การเกิดโรคเบาหวาน หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.5

ตารางที่ 4.6 ค่าความไวหลังคัดเลือกตัวแปร โรคเบาหวาน

วิธีการคัดเลือกตัวแปร	ค่าความไว				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	83.60	81.40	84.40	83.20	87.40
Wrapper + Random Forest	84.00	81.20	85.20	83.20	87.60
Wrapper + Decision Tree C4.5	80.80	84.20	86.60	86.20	87.20
Wrapper + Naïve Bayes	83.20	85.20	88.00	84.60	87.40
Wrapper + Neural Networks	84.00	86.20	87.00	84.60	87.80
Wrapper + Deep Learning	82.40	85.00	88.00	85.80	87.80
Gain Ratio	81.40	83.60	84.40	83.20	87.40

จากตารางที่ 4.5 ค่าความไวต้องหลังคัดเลือกตัวแปร โรคเบาหวาน โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Naïve Bayes และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Naïve Bayes ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 88.00 และให้ค่าความไวน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Random Forest ที่ร้อยละ 80.80

ค่าความไวของแบบพยากรณ์การเกิดโรคไฮโปไทรอยด์ หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.6

ตารางที่ 4.7 ค่าความไวหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์

วิธีการคัดเลือกตัวแปร	ค่าความไว				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	99.82	99.85	99.45	99.34	99.60
Wrapper + Random Forest	99.89	99.86	99.48	99.14	99.57
Wrapper + Decision Tree C4.5	99.83	99.86	99.48	99.25	99.60
Wrapper + Naïve Bayes	99.86	99.86	99.48	99.28	99.60
Wrapper + Neural Networks	99.86	99.86	99.45	98.99	99.60

ตารางที่ 4.8 ค่าความไวหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์

วิธีการคัดเลือกตัวแปร	ค่าความไว				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
Wrapper + Deep Learning	99.89	99.86	99.45	99.43	99.60
Gain Ratio	99.83	99.86	99.45	99.34	99.60

จากตารางที่ 4.6 ค่าความไวต้องหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์ โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric

วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และวิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 99.89 และให้ค่าความไวน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Neural Networks ที่ร้อยละ 98.99

4.2.3 ผลค่าจำเพาะหลังการคัดเลือกตัวแปร

ค่าจำเพาะของแบบพยากรณ์การเกิดโรคมะเร็งเต้านม หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.7

ตารางที่ 4.9 ค่าจำเพาะหลังคัดเลือกตัวแปร โรคมะเร็งเต้านม

วิธีการคัดเลือกตัวแปร	ค่าจำเพาะ				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	28.23	27.05	43.53	41.17	41.17
Wrapper + Random Forest	27.05	27.05	27.05	30.58	48.23
Wrapper + Decision Tree C4.5	28.24	27.06	29.41	34.12	52.94
Wrapper + Naïve Bayes	27.05	27.05	27.05	30.58	48.23
Wrapper + Neural Networks	27.05	27.05	27.05	30.58	48.23
Wrapper + Deep Learning	31.76	27.06	42.35	36.47	49.41
Gain Ratio	90.83	91.74	87.16	81.65	86.24

จากตารางที่ 4.7 ค่าจำเพาะหลังคัดเลือกตัวแปร ของโรคมะเร็งเต้านม โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านม ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะสูงที่สุดเท่ากันที่ร้อยละ 91.74 และให้ค่าจำเพาะน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest, Decision Tree C4.5 และ Naïve Bayes, Wrapper + Naïve Bayes ด้วยเทคนิค Random Forest, Decision Tree C4.5 และ Naïve Bayes และ Wrapper + Neural Networks ด้วยเทคนิค Random Forest, Decision Tree C4.5 และ Naïve Bayes,เท่ากันที่ร้อยละ 27.05

ค่าจำเพาะของแบบพยากรณ์การเกิดโรคเบาหวาน หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.8

ตารางที่ 4.10 ค่าจำเพาะหลังคัดเลือกตัวแปร โรคเบาหวาน

วิธีการคัดเลือกตัวแปร	ค่าจำเพาะ				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	61.19	59.70	61.19	60.82	58.96
Wrapper + Random Forest	61.57	60.45	61.19	61.57	58.21
Wrapper + Decision Tree C4.5	58.58	60.07	58.21	60.07	57.84
Wrapper + Naïve Bayes	57.46	55.22	58.58	57.84	58.05
Wrapper + Neural Networks	60.45	56.72	58.58	62.31	58.65
Wrapper + Deep Learning	54.85	55.22	55.22	54.85	57.46
Gain Ratio	59.70	61.19	61.19	60.82	58.96

จากตารางที่ 4.8 ค่าจำเพาะหลังคัดเลือกตัวแปร โรคเบาหวาน โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคเบาหวาน ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks

ด้วยเทคนิค Neural Networks ให้ค่าจำเพาะสูงที่สุดที่ร้อยละ 62.31 และให้ค่าจำเพาะน้อยที่สุดในวิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ที่ร้อยละ 54.85

ค่าจำเพาะของแบบพยากรณ์การเกิดโรคไฮโปไทรอยด์ หลังการคัดเลือกตัวแปรด้วยวิธีการ Wrapper ที่ร่วมกับเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning และหลักการของ Gain Ratio แล้วนำมาสร้างแบบพยากรณ์ด้วยเทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ได้ผลดังตารางที่ 4.9

ตารางที่ 4.11 ค่าจำเพาะหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์

วิธีการคัดเลือกตัวแปร	ค่าจำเพาะ				
	Random Forest	Decision Tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
ข้อมูลก่อนการคัดเลือกตัวแปร	97.93	100.00	34.54	78.35	34.53
Wrapper + Random Forest	100.00	100.00	25.77	88.66	30.93
Wrapper + Decision Tree C4.5	100.00	100.00	25.77	90.21	31.44
Wrapper + Naïve Bayes	100.00	100.00	32.99	85.57	31.94
Wrapper + Neural Networks	100.00	100.00	32.99	86.08	32.99
Wrapper + Deep Learning	100.00	100.00	33.51	81.96	33.51
Gain Ratio	97.94	100.00	34.54	78.35	34.54

จากตารางที่ 4.9 ค่าจำเพาะหลังคัดเลือกตัวแปร โรคไฮโปไทรอยด์ โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคไฮโป

ไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest และ Decision Tree C4.5 และ วิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะสูงที่สุดที่ร้อยละ 100.00 และให้ค่าจำเพาะน้อยที่สุดใน วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Naïve Bayes และ วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Naïve Bayes เท่ากันที่ร้อยละ 25.77

บทที่ 5
สรุปผลและข้อเสนอแนะการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูลในข้อมูลที่หลากหลาย โดยการสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ จากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล ด้วยเทคนิค Random Forest เทคนิค Decision Tree C4.5 เทคนิค Naïve Bayes เทคนิค Artificial Neural Networks และเทคนิค Deep Learning จากการทดลองพบว่า ชุดข้อมูลก่อนทำการคัดเลือกตัวแปร ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึงร้อยละ 75.52 ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึงร้อยละ 77.47 และ โรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึงร้อยละ 99.86

วิธีการของ Wrapper และ หลักการของ Gain Ratio หลังจากที่นำเข้ามาในการคัดเลือกตัวแปรแล้ว ผลการทดลองพบว่าชุดข้อมูลโรคมะเร็งเต้านมที่ทำการคัดเลือกตัวแปร วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Neural Networks, วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Decision Tree C4.5 และ วิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Random Forest ให้ค่าความถูกต้องสูงที่สุดเท่ากันที่ร้อยละ 75.87 เมื่อเปรียบเทียบกับข้อมูลก่อนการคัดเลือกตัวแปร ค่าความถูกต้องเพิ่มขึ้นจากค่าสูงสุดของร้อยละ 0.35

วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Decision Tree C4.5 และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Decision Tree C4.5 ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 96.52 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปร ร้อยละ 0.50

วิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะสูงที่สุดเท่ากันที่ร้อยละ 91.74 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 48.21

ชุดข้อมูลโรคเบาหวานที่ทำการคัดเลือกตัวแปร วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Naïve Bayes ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 77.73 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 0.26

วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Naïve Bayes และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Naïve Bayes ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 88.00 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 0.60

วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Neural Networks ให้ค่าจำเพาะสูงที่สุดที่ร้อยละ 62.31 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 3.35

ชุดข้อมูลโรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และ วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ให้ค่าความถูกต้องสูงที่สุดเท่ากันที่ร้อยละ 99.89 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 0.16

วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และวิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest ให้ค่าความไวสูงที่สุดเท่ากันที่ร้อยละ 99.89 เพิ่มขึ้นจากค่าสูงสุดของข้อมูลก่อนการคัดเลือกตัวแปรร้อยละ 0.04

วิธีการคัดเลือกตัวแปร Wrapper + Random Forest ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Decision Tree C4.5 ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Naïve Bayes ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Neural Networks ด้วยเทคนิค Random Forest และ Decision Tree C4.5, วิธีการคัดเลือกตัวแปร Wrapper + Deep Learning ด้วยเทคนิค Random Forest และ Decision Tree C4.5 และวิธีการคัดเลือกตัวแปร Gain Ratio ด้วยเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะสูงที่สุดที่ร้อยละ 100.00

5.2 ข้อเสนอแนะ

งานวิจัยนี้เป็นการสร้างแบบจำลองเพื่อศึกษาเทคนิคของการคัดเลือกตัวแปร ในการเพิ่มประสิทธิภาพของแบบจำลองในการพยากรณ์การเกิดโรคของผู้ป่วยที่เป็นโรค มะเร็งเต้านม, เบาหวาน และ ไฮโปไทรอยด์ หากจะนำไปใช้ อาจจะต้องใช้เทคนิคเหมืองข้อมูลเทคนิคอื่นๆ เข้ามาเปรียบเทียบเพื่อให้ได้ผลการพยากรณ์ที่มีประสิทธิภาพมากขึ้น

บรรณานุกรม

บรรณานุกรม

[1] ณัฏฐพร นันทิวัฒนา. (25/09/2020). มะเร็งเต้านม. Available: https://www.sikarin.com/content/detail/461/

[2] พิมพ์ใจ อันทานนท์. (11/6/2020). โรคเบาหวาน. Available: https://www.dmthai.org/index.php/knowledge/for-normal-person/health-information-and-articles/health-information-and-articles-old-3/846-2019-04-20-01-49-18

[3] เมดไทย. (25/09/2020). ไทรอยด์เป็นพิษ. Available: https://medthai.com

[4] อัจจิมา มณฑาพันธุ์. (2/12/2020). การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม. Available: http://www.dspace.spu.ac.th/bitstream/123456789/6198/1

[5] D. K. B. N. Hoque, J.K. Kalita,. (2/12/2020). *MIFS-ND: A mutual information-based feature selection method*. Available: https://www.sciencedirect.com/science/article/pii/S0957417414002164

[6] X. Liu, Y. Liang, S. Wang, Z. Yang, and H. Ye, "A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection," *IEEE Access,* vol. 6, pp. 22863-22874, 2018.

[7] N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification," in 2019 *International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1-6.

[8] N. S.-M. V. Bolón-Canedo, A. Alonso-Betanzos, J.M. Benítez, F. Herrera. (2014, 2/12/2020). *A review of microarray datasets and applied feature selection methods*. Available:https://www.sciencedirect.com/science/article/pii/S0020025514006021

[9] ชณิดาภา บุญประสม. (25/9/2020). การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล. Available: http://research.fte.kmutnb.ac.th/download.php?filename=620701000056&filepath=20190701155744.pdf

[10] H. L. Han, H. Y. Ma, and Y. Yang, "Study on the Test Data Fault Mining Technology Based on Decision Tree," *Procedia Computer Science,* vol. 154, pp. 232-237, 2019/01/01/ 2019.

[11] M. Czajkowski and M. Kretowski, "Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach," *Expert Systems with Applications,* vol. 137, pp. 392-404, 2019/12/15/ 2019.

[12] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education,* vol. 113, pp. 177-194, 2017/10/01/ 2017.

[13] A. G. Maninder Kaur. *A Framework for the Indirect Assessment Tool for Outcome Based Education Using Data Mining*. Available: https://ieeexplore. ieee.org/document/8782336

[14] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems,* vol. 192, p. 105361, 2020/03/15/ 2020.

[15] ว. เ. เสกสรรค์ วิลัยลักษณ์, ดวงดาว วิชาดากุล,. (28/11/2020). การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา. Available: https://ph01.tci-thaijo.org/index.php/VESTSU/article/view/45633/37766

[16] ทิพย์หทัย ทองธรรมชาติ. (2017, 25/9/2020). การคัดเลือกคุณลักษณะเพื่อสร้างโมเดลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล. Available: https://research.kpru.ac.th/sac/fileconference/10912018-05-01

[17] M. Rastgou, H. Bayat, M. Mansoorizadeh, and A. S. Gregory, "Estimating the soil water retention curve: Comparison of multiple nonlinear regression approach and random forest data mining technique," *Computers and Electronics in Agriculture,* vol. 174, p. 105502, 2020/07/01/ 2020.

[18] Y. Zhao, S.-K. Otto, N. Brandt, M. Selzer, and B. Nestler, "Application of Random Forests in ToF-SIMS Data," *Procedia Computer Science,* vol. 176, pp. 410-419, 2020/01/01/ 2020.

[19] S. Komatsu *et al.*, "Deep learning-assisted literature mining for in vitro radiosensitivity data," *Radiotherapy and Oncology,* vol. 139, pp. 87-93, 2019/10/01/ 2019.

[20] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications,* vol. 166, p. 114060, 2021/03/15/ 2021.

[21] F. O. Gallego and R. Corchuelo, "A deep-learning approach to mining conditions," *Knowledge-Based Systems,* vol. 193, p. 105422, 2020/04/06/ 2020.

[22] F. Qi, Z. Chang-jie, and Y. Liu, "Predicting breast cancer recurrence using data mining techniques," in 2010 *International Conference on Bioinformatics and Biomedical Technology*, 2010, pp. 310-311.

[23] V. R. Balpande and R. D. Wajgi, "Prediction and severity estimation of diabetes using data mining technique," in 2017 *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2017, pp. 576-580.

[24] S. S. Zarin Mousavi, M. Mohammadi Zanjireh, and M. Oghbaie, "Applying computational classification methods to diagnose Congenital Hypothyroidism: A comparative study," *Informatics in Medicine Unlocked,* vol. 18, p. 100281, 2020/01/01/ 2020.

[25] U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," in 2017 7*th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, 2017, pp. 527-530.

[26] วาทินี นุ้ยเพียร. (2/12/2020). การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ Available: http://j.cit.kmutnb.ac.th/storage/attachments/020/2-9

[27] สุรวัชร ศรีเปารยะ. (30/11/2020). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. Available: https://li01.tci-thaijo.org/index.php/tstj/article/view/85101/67778

ประวัติผู้เขียน

ชื่อ	อุกฤษฏ์ ศรีสุข
วันเกิด	วันศุกร์ ที่ 9 สิงหาคม พ.ศ.2539
สถานที่เกิด	โรงพยาบาลอำนาจเจริญ
สถานที่อยู่ปัจจุบัน	129 หมู่ 5 ตำบล บุ่ง อำเภอ เมือง จังหวัด อำนาจเจริญ 37000
ประวัติการศึกษา	พ.ศ.2558 ปริญญาบริหารธุรกิจบัณฑิต (บธ.บ) สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม พ.ศ. 2564 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม