การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง

นางสาวคุณาภรณ์ พันธ์เพียร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
หลักสูตรสาขาวิชานวัตกรรม วิศวชีวการแพทย์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2564

DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING

KUNAPORN PUNPAIN

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Engineering in Biomedical Innovation Engineering
Suranaree University of Technology
Academic Year 2021

การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

.....

(ผศ. ดร.สุขเกษม วัชรมัยสกุล)
ประธานกรรมการ

.....

(ผศ. ดร.เจษฎา ตัณฑนุช)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

.....

(ผศ. ดร.พิสมัย กิตติภูมิ)
กรรมการ

.....

(รศ. ดร.ฉัตรชัย โชติษฐยางกูร)
รองอธิการบดีฝ่ายวิชาการและประกันคุณภาพ

.....

(รศ. ดร.พรศิริ จงกล)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

คุณาภรณ์ พันธ์เพียร : การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง (DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ตัณฑนุช, 57 หน้า.

คำสำคัญ : โรคเบาหวาน, วิศวกรรมคุณลักษณะ, การเรียนรู้ของเครื่อง

จุดมุ่งหมายของการศึกษานี้คือ เพื่อศึกษาปัจจัยเสี่ยงที่มีผลกระทบต่อการทำให้เกิดโรคเบาหวานโดยใช้วิธีวิศวกรรมคุณลักษณะ แล้วนำไปพัฒนาแบบจำลองเพื่อจำแนกประเภทของผู้ป่วยโรคเบาหวานและไม่เป็นโรคเบาหวาน ซึ่งการศึกษาครั้งนี้ได้นำข้อมูลจากการตอบแบบสำรวจจำนวน 70,692 รายการ โดยได้ข้อมูลจากเว็บไซต์ Kaggle แล้วนำมาใช้วิธีวิศวกรรมคุณลักษณะเพื่อหาคุณลักษณะที่ดีที่สุดออกมา จากนั้นนำไปสร้างแบบจำลองด้วย 5 ขั้นตอนวิธี ได้แก่ เทคนิคป่าสุ่ม เทคนิคต้นไม้ตัดสินใจ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคเกรเดียนท์บูตทรี และซัพพอร์ทเวกเตอร์แมชชีน แล้วทำการเปรียบเทียบประสิทธิภาพและความแม่นยำของแบบจำลอง ผลการศึกษาพบว่าประสิทธิภาพการทำงานของบางแบบจำลองเมื่อเปรียบเทียบระหว่างชุดข้อมูลต้นฉบับกับชุดข้อมูลที่ผ่านการสกัดคุณลักษณะด้วยวิธีวิศวกรรมคุณลักษณะมีค่าลดลงแต่ยังคงมีค่าใกล้เคียงกันมาก ซึ่งได้แก่ แบบจำลองที่สร้างโดยเทคนิคป่าสุ่มและเทคนิคต้นไม้ตัดสินใจ แต่แบบจำลองที่สร้างโดยเทคนิคเกรเดียนท์บูตทรี เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ตเวกเตอร์แมชชีน มีประสิทธิภาพที่มากขึ้นกว่าเดิมอย่างชัดเจน ทั้งนี้พบว่าการใช้วิธีวิศวกรรมคุณลักษณะร่วมกับการสร้างแบบจำลองด้วยเทคนิคป่าสุ่ม สามารถให้ประสิทธิภาพการทำงานโดยภาพรวมอยู่ในเกณฑ์ที่ดีกว่าแบบจำลองอื่น ๆ โดยคุณลักษณะที่สกัดออกมาได้นั้นเหลือเพียง 7 คุณลักษณะ จากทั้งหมด 21 คุณลักษณะ ได้แก่ 1) ภาวะความดันโลหิต 2) ภาวะคลอเรสเตอรอลสูง 3) ค่าดัชนีมวลกาย 4) ระดับสุขภาพโดยทั่วไป 5) เพศ 6) ระดับช่วงอายุ และ 7) ระดับเงินเดือน โดยมีผลการวัดประสิทธิภาพดังนี้ Accuracy = 73.35% Precision = 68.53% Specificity = 60.17% Sensitivity = 87.02% F1 score = 76.55% ROC AUC = 0.815 และ Kappa = 0.469 ดังนั้นในการศึกษาครั้งนี้พบว่าการนำวิธีวิศวกรรมคุณลักษณะเข้ามาทำงานร่วมกับการสร้างแบบจำลองด้วยเทคนิคป่าสุ่ม ทำให้สามารถสร้างแบบจำลองเหมาะสมในการคัดกรองผู้ป่วยที่เป็นโรคเบาหวานได้ดีที่สุด

สาขาวิชานวัตกรรม วิศวชีวการแพทย์ ลายมือชื่อนักศึกษา ...คุณาภรณ์ พันธ์เพียร...

ปีการศึกษา 2564 ลายมือชืออาจารย์ที่ปรึกษา .....

KUNAPORN PUNPAIN: DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING. THESIS ADVISOR : ASST. PROF. JESSADA TANTHANUCH, Ph.D., 57 PP.

Keyword : DIABETES, FEATURE ENGINEERING, MACHINE LEARNING

The aim of this study is to study the risk factors affecting the incidence of diabetes using feature engineering method and then to develop a model to classify the types of diabetic and non-diabetic patients. This study used data from 70,692 survey responses from Kaggle website. First the feature engineering method was applied to extract the best features and then created models by 5 algorithms, which were random forest technique (RFT), decision tree technique (DTT), nearest neighbor technique (NNT), gradient boot tree technique (BGT), and support vector Machine (SVM). All models were evaluated the performance and the accuracy. The results showed that the performance of some models compared between the original dataset and the feature extracted by feature engineering method was lower but still very close, i.e. the model generated by RFT and DTT. However, for the model created by GBT, NNT, and SVM, feature engineering method clearly helped in increasing the efficiency. It was found that the use of feature engineering method together with RFT for creating the model provided the better overall performance than other methods. For the model mentioned, only 7 of the 21 features extracted were available, namely: 1) high blood pressure 2) high cholesterol 3) BMI 4) general health 5) gender 6) age and 7) salary. The efficiency measurement results were as follows: Accuracy = 73.35% Precision = 68.53% Specificity = 60.17% Sensitivity = 87.02% F1 score = 76.55% ROC AUC = 0.815 and Kappa = 0.469. Hence, the result of the study presents that the use of feature engineering method together with RFT offers a suitable model for the best screening of patients with diabetes.

School of Biomedical Innovation Engineering
Academic Year 2021

Student's Signature คุณากรณ์ พันธ์เพียร
Advisor's Signature ....J. Tanthanuch

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงด้วยดี เนื่องจากได้รับความกรุณาให้คำปรึกษา แนะนำ ชี้แนะแนวทางการดำเนินการวิจัย ข้าพเจ้าขอกราบขอบพระคุณบุคคลดังต่อไปนี้

ผู้ช่วยศาสตราจารย์ ดร.สุขเกษม วัชรมัยสกุล ประธานกรรมการที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาแนะนำ ช่วยเหลือ อย่างดียิ่ง ทั้งด้านการศึกษาและด้านการดำเนินการวิจัย ตลอดจนคำแนะนำในการรับทุนการศึกษาในระดับบัณฑิตศึกษา

ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ตัณฑนุช อาจารย์ประจำสาขาวิชาคณิตศาสตร์ (อาจารย์ที่ปรึกษาวิทยานิพนธ์) ที่ได้กรุณาให้คำปรึกษาแนะนำ ช่วยเหลือ และแก้ไขปรับปรุงข้อบกพร่องในด้านการดำเนินการวิจัย ตลอดจนคำแนะนำในการเขียน การตรวจแก้ไขวิทยานิพนธ์เล่มนี้จนเสร็จสมบูรณ์

ผู้ช่วยศาสตราจารย์ ดร.เบญจวรรณ โรจนดิษฐ์ อาจารย์ประจำสาขาวิชาคณิตศาสตร์ ที่ได้กรุณาให้คำปรึกษาแนะนำในการเขียน และการตรวจแก้ไขในด้านการดำเนินการวิจัย

คณาจารย์ทุกท่านในหลักสูตรสาขาวิชานวัตกรรม วิศวชีวการแพทย์มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ให้คำแนะนำด้านองค์ความรู้ แนวทางในการศึกษาค้นคว้ามาโดยตลอด เพื่อให้ผู้วิจัยสามารถดำเนินการวิจัยได้เป็นอย่างดี

คุณจักรกฤษณ์ พลรบ นักศึกษาบัณฑิตสาขาวิชาคณิตศาสตร์ประยุกต์ ที่ให้คำปรึกษาแนะนำ ข้อเสนอแนะต่างๆ และช่วยเหลือในการดำเนินการวิจัย

มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ได้มอบทุนการศึกษาประเภทบัณฑิตศึกษาประจำปี 2563 ให้แก่ผู้วิจัยและทุนสนับสนุนในการนำเสนอและเผยแพร่งานวิจัย ทำให้การศึกษาในระดับปริญญาโทสำเร็จลุล่วงไปได้ด้วยดี

ผู้วิจัยมีความซาบซึ้งในความกรุณาของทุกท่านที่ได้กล่าวมาข้างต้นและบุคคลอื่นที่ผู้วิจัยมิได้กล่าวถึง ซึ่งได้มีส่วนช่วยเหลือและให้การสนับสนุนงานวิจัยในครั้งนี้ จึงขอกราบขอบพระคุณทุกท่านด้วยความจริงใจ

คุณาภรณ์ พันธ์เพียร

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหาการวิจัย

ปัจจุบันโรคเบาหวานเป็นปัญหาทางสาธารณสุขที่สำคัญโรคหนึ่งของทั่วโลกรวมทั้งประเทศไทยด้วยโดยโรคเบาหวานเป็นโรคที่มีความผิดปกติเกี่ยวกับการนำน้ำตาลไปใช้ประโยชน์อันเกี่ยวเนื่องกับความบกพร่องของฮอร์โมนอินซูลิน หรือประสิทธิภาพการทำงานของฮอร์โมนอินซูลินลดลง ส่งผลให้กระบวนการดูดซึมน้ำตาลในเลือดให้เป็นพลังงานของเซลล์ในร่างกายทำงานได้ไม่เต็มประสิทธิภาพ ทำให้ร่างกายมีระดับน้ำตาลในเลือดสูงกว่าปกติ หากปล่อยให้ร่างกายอยู่ในสภาวะนี้เป็นเวลานานจะทำให้อวัยวะต่าง ๆ เสื่อมลงก่อให้เกิดอาการและภาวะแทรกซ้อนต่าง ๆ ตามมาที่ส่งผลให้เกิดอันตรายได้ เช่น โรคความดันโลหิตสูง โรคหัวใจขาดเลือดและกล้ามเนื้อหัวใจตายเฉียบพลัน อัมพฤกษ์และอัมพาตจากหลอดเลือดสมองตีบ โรคแทรกซ้อนทางตา เช่น อาการตามัวเบาหวานขึ้นตา (retinopathy) โรคแทรกซ้อนทางไต ทำให้ไตเสื่อม ไตวาย (กรมควบคุมโรค, 2564)

สถานการณ์โรคเบาหวานในปี 2564 พบว่าทั่วโลกมีผู้ป่วยจำนวน 463 ล้านคน และคาดการณ์ว่าในปี 2588 จะมีผู้ป่วยโรคเบาหวานเพิ่มขึ้นถึง 629 ล้านคน สำหรับประเทศไทยพบอุบัติการณ์โรคเบาหวานมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง มีผู้ป่วยรายใหม่เพิ่มขึ้นประมาณ 3 แสนคนต่อปี และมีผู้ป่วยโรคเบาหวานอยู่ในระบบทะเบียนของกระทรวงสาธารณสุขจำนวน 3.2 ล้านคน (กรมควบคุมโรค, 2564) แม้ว่าโรคเบาหวานจะเป็นโรคเรื้อรังที่ยังไม่มีทางรักษาให้หายขาด แต่สามารถป้องกันและชะลออาการได้ อย่างเช่น การลดน้ำหนัก การรับประทานอาหารเพื่อสุขภาพ การออกกำลังกายอย่างสม่ำเสมอ คอยควบคุมระดับน้ำตาลในเลือดและคอเลสเตอรอลให้อยู่เกณฑ์ปกติ และการได้รับการรักษาพยาบาลสามารถบรรเทาอันตรายจากโรคนี้ในผู้ป่วยจำนวนมากได้

การนำเทคโนโลยีมาประยุกต์ใช้ในการวินิจฉัยโรคต่าง ๆ รวมไปถึงการใช้เทคโนโลยีทางด้านการเรียนรู้ของเครื่อง (Machine Learning: ML) และ ปัญญาประดิษฐ์ (Artificial Intelligence: AI) มาช่วยในการวินิจฉัยโรคเบาหวาน เพื่อพัฒนาให้การรักษามีประสิทธิภาพมากขึ้น เนื่องจากมุ่งเน้นการใช้ตัวอย่างหรือประสบการณ์เพื่อการเรียนรู้งาน ทำให้มีระดับความแม่นยำในการวินิจฉัยโรคที่สูง ซึ่งจะเป็นกำลังสำคัญเข้ามาช่วยสนับสนุนการทำงานของแพทย์ ดังนั้นผลลัพธ์ที่ได้จากการทำงานร่วมกันคือเพิ่มทั้งความเร็วในกระบวนการคัดกรองผู้ป่วย การตรวจสอบรักษา โดยผู้ป่วยก็จะไม่เสียโอกาสได้เข้ารับการรักษาตั้งแต่ช่วงระยะแรกที่เริ่มมีแนวโน้มเสี่ยงต่อการเกิดโรค นอกจากนี้หากได้เข้ารับการรักษาที่รวดเร็วยังช่วยลดการเกิดภาวะแทรกซ้อน และลดการสูญเสีย (คลังความรู้ SciMath, 2561)

จากข้อมูลที่กล่าวมาข้างต้นผู้วิจัยพบว่าการวินิจฉัยอาการหรือตรวจพบว่าเป็นโรคเบาหวานตั้งแต่ระยะแรกเริ่ม สามารถนำไปสู่การเปลี่ยนแปลงวิถีชีวิตและทำให้ได้รับการรักษาที่มีประสิทธิภาพมากขึ้น ดังนั้นผู้วิจัยจึงมีแนวคิดในการสร้างแบบจำลองสำหรับการทำนายผู้ที่มีแนวโน้มเสี่ยงต่อการเป็นโรคเบาหวาน โดยใช้เทคนิควิศวกรรมคุณลักษณะ (Feature Engineering) เพื่อสกัดคุณลักษณะที่เสี่ยงต่อการเกิดโรคเบาหวาน เช่น เพศ อายุ น้ำหนัก และหลังจากนั้นจะนำเข้าสู่กระบวนการการเรียนรู้ของเครื่อง (Machine Learning) เพื่อให้แบ่งประเภท (Classification) สำหรับผู้ป่วยโรคเบาหวาน และไม่ป่วยโรคเบาหวาน โดยผลลัพธ์ที่ได้สามารถนำไปประยุกต์ใช้สำหรับการประเมินผู้ที่มีแนวโน้มเป็นโรคเบาหวาน รวมไปถึงช่วยแพทย์ในการตัดสินใจสำหรับการเลือกวิธีการรักษาให้เหมาะสมกับลักษณะของผู้ป่วยได้อย่างรวดเร็วขึ้น

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยเสี่ยงที่มีผลกระทบต่อการทำให้เกิดโรคเบาหวาน โดยใช้วิธีวิศวกรรมคุณลักษณะ แล้วนำไปพัฒนาแบบจำลองของผู้ป่วยโรคเบาหวาน โดยการจำแนกประเภทของผู้เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน และเปรียบเทียบประสิทธิภาพ และความแม่นยำของแบบจำลองที่ได้พัฒนาขึ้น

1.3 ข้อตกลงเบื้องต้น

1.3.1 ร้างแบบจำลองสำหรับการทำนายผู้ที่มีแนวโน้มเสี่ยงต่อการเป็นโรคเบาหวาน โดยใช้เทคนิควิศวกรรมคุณลักษณะร่วมกับซัพพอร์ทเวกเตอร์แมชชีน, ต้นไม้ตัดสินใจ, เทคนิคเพื่อนบ้านใกล้ที่สุด, เกรเดียนท์บูตทรี และป่าสุ่ม

1.3.2 ออกแบบและพัฒนาแบบจำลองของผู้ป่วยโรคเบาหวาน โดยการจำแนกประเภทของผู้เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน

1.3.3 เปรียบเทียบประสิทธิภาพ และความแม่นยำของแบบจำลองที่ได้พัฒนาขึ้นโดยใช้ตัววัดประสิทธิภาพ Confusion Matrix, Accuracy, Precision, Recall, F1 Score, AUC of ROC และ Cohen Kappa Coefficient

1.4 ขอบเขตของงานวิจัย

1.4.1 ข้อมูลได้จากการตอบแบบสำรวจจำนวน 70,692 รายการ โดยผู้ตอบแบบสำรวจแบ่งเป็นคนที่เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวาน ซึ่งมีจำนวนเท่า ๆ กัน โดยได้ข้อมูลจาก https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset

1.4.2 สกัดคุณลักษณะโดยใช้เทคนิควิศวกรรมคุณลักษณะ

1.4.3 สร้างแบบจำลองโดยใช้ขั้นตอนวิธี ดังนี้

- Support Vector Machine (SVM)
- Decision Tree
- k – Nearest Neighbor (kNN)
- Gradient Boosted Tree
- Random Forest

1.4.4 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี โดยใช้ตัววัดประสิทธิภาพ ดังนี้

- Confusion Matrix
- Accuracy
- Precision, Recall and F1 Score
- AUC of ROC
- Cohen Kappa Coefficient

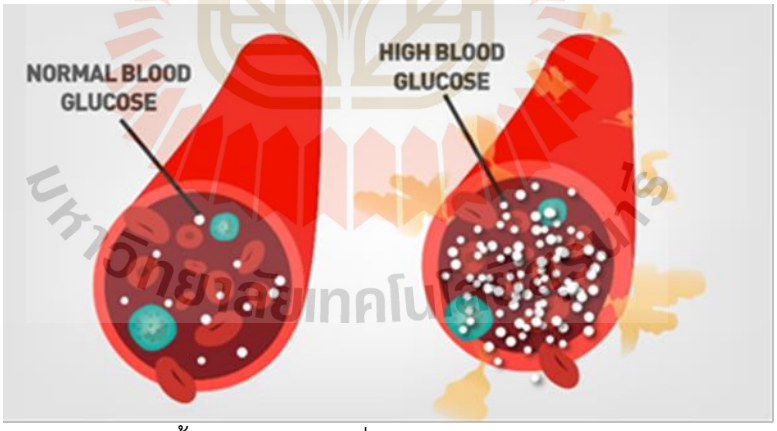
1.5 ประโยชน์ที่คาดว่าจะได้รับ

ผลลัพธ์ของแบบจำลองสามารถจำแนกผู้ป่วยที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวานได้ โดยสามารถนำไปประยุกต์ใช้สำหรับการประเมินผู้ที่มีแนวโน้มเป็นโรคเบาหวาน รวมไปถึงช่วยแพทย์ในการตัดสินใจสำหรับการเลือกวิธีการรักษาให้เหมาะสมกับลักษณะของผู้ป่วยได้อย่างรวดเร็วขึ้น

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 ความรู้เกี่ยวกับโรคเบาหวาน

โรคเบาหวานเป็นโรคที่มีความผิดปกติเกี่ยวกับการนำน้ำตาลไปใช้ประโยชน์อันเกี่ยวเนื่องกับความบกพร่องของฮอร์โมนอินซูลิน หรือประสิทธิภาพการทำงานของฮอร์โมนอินซูลินลดลง ส่งผลให้กระบวนการดูดซึมน้ำตาลในเลือดให้เป็นพลังงานของเซลล์ในร่างกายทำงานได้ไม่เต็มประสิทธิภาพ ทำให้ร่างกายมีระดับน้ำตาลในเลือดสูงกว่าปกติ หากน้ำตาลในกระแสเลือดสูงมากขึ้นถึงระดับหนึ่ง จะทำให้ไตซึ่งปกติจะมีหน้าที่ดูดกลับน้ำตาลจากสารที่ถูกกรองจากหน่วยไตไปใช้ ดูดกลับน้ำตาลได้ไม่หมด ส่งผลให้มีน้ำตาลรั่วออกมากับปัสสาวะ จึงเป็นที่มาของคำว่า “โรคเบาหวาน” ผู้ที่เป็นเบาหวานมักจะมีประวัติคนในครอบครัว พ่อแม่หรือญาติพี่น้องสายตรงเป็นโรคนี้ด้วย ร่วมกับมีพฤติกรรมการใช้ชีวิตที่ไม่ดีต่อสุขภาพ ก็จะมีโอกาสเป็นโรคเบาหวานได้มากขึ้น หากเราปล่อยให้เกิดภาวะเช่นนี้ไปนาน ๆ โดยไม่ได้รับการรักษาอย่างถูกวิธีจะทำให้เกิดภาวะแทรกซ้อนตามมา



รูปที่ 2.1 แสดงลักษณะน้ำตาลในเลือด (ที่มา : https://www.howtorelief.com)

ชนิดของโรคเบาหวาน มีดังนี้

- **เบาหวานชนิดที่ 1** เกิดจากการขาดอินซูลิน เนื่องจากตับอ่อนไม่สามารถหลั่งอินซูลินได้ ทำให้เกิดภาวะขาดอินซูลิน เบาหวานชนิดนี้มักพบในเด็กและผู้ที่มีอายุน้อยกว่า 40 ปี

- **เบาหวานชนิดที่ 2** เกิดจากการที่เซลล์ของร่างกายต่อสนองต่ออินซูลินได้ไม่ดีหรือที่เรียกว่าภาวะดื้อต่ออินซูลิน ทำให้ร่างกายเหมือนขาดอินซูลินไประดับหนึ่ง พบมากในคนส่วนใหญ่ สาเหตุ ได้แก่ พันธุกรรม ความอ้วน และการไม่ออกกำลังกาย
- **เบาหวานชนิดที่ 3** เป็นเบาหวานชนิดที่มีสาเหตุชัดเจน เช่น โรคตับอ่อนอักเสบ มะเร็งตับอ่อน การรับประทานยาบางชนิด เช่น ยาสเตียรอยด์ การได้รับสารเคมีบางชนิด ความผิดปกติของฮอร์โมนจากต่อมหมวกไต
- **เบาหวานชนิดที่ 4** เป็นเบาหวานที่เกิดขึ้นขณะตั้งครรภ์ และหายไปได้หลังคลอดบุตร แต่มีโอกาสเสี่ยงที่จะเกิดโรคเบาหวานชนิดที่ 2 ในอนาคต (พญ.นพวรรณ กิติวัฒน์, 2559)

อาการของโรคเบาหวาน

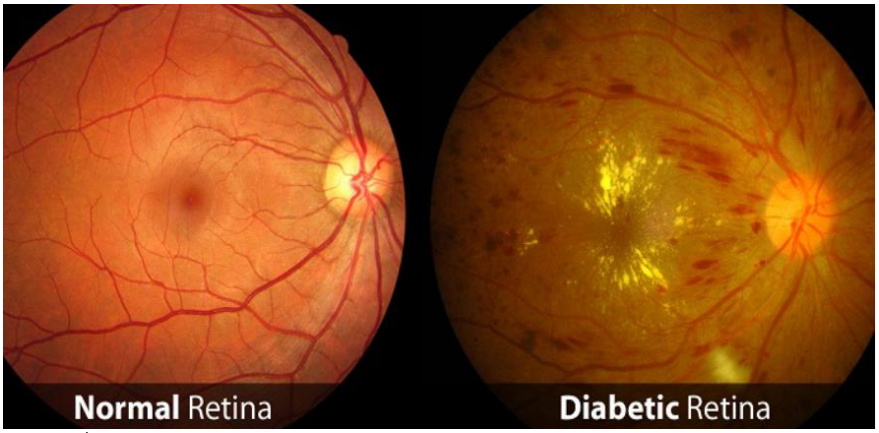
โรคเบาหวานในระยะแรกจะไม่แสดงอาการผิดปกติ บางรายอาจตรวจพบโรคเบาหวานเมื่อพบภาวะแทรกซ้อนขึ้นแล้ว ซึ่งอาการที่พบส่วนใหญ่ คือ กระหายน้ำมาก ปากแห้ง ปัสสาวะบ่อย หิวบ่อย น้ำหนักลดหรือเพิ่มผิดปกติ สายตาพร่ามัว รู้สึกเหนื่อยง่าย ชาบริเวณปลายมือปลายเท้า บาดแผลหายยาก หย่อนสมรรถภาพทางเพศ หากมีอาการเหล่านี้เกิดขึ้น ควรรีบปรึกษาแพทย์ทันที ไม่ควรรอจนอาการต่าง ๆ เหล่านี้เกิดขึ้น เพราะบางครั้งกว่าจะเกิดอาการเหล่านี้ ระดับน้ำตาลในเลือดก็อาจสูงเกินไปแล้ว ทางที่ดีควรเข้ารับการตรวจสุขภาพประจำปี โดยเฉพาะบุคคลที่มีแนวโน้มเสี่ยงต่อโรคเบาหวานควรได้รับการตรวจคัดกรอง เช่น ผู้ที่มีอายุ 45 ปีขึ้นไป คนในครอบครัว พ่อแม่หรือญาติพี่น้องสายตรง มีประวัติเป็นโรคเบาหวาน ผู้ที่มีน้ำหนักเกินเกณฑ์มาตรฐาน (มีค่าดัชนีมวลกายหรือ BMI มากกว่า 25) ผู้ที่เป็นเบาหวานขณะตั้งครรภ์ ผู้ที่มีระดับไขมันในเลือดผิดปกติ ผู้มีโรคหัวใจและหลอดเลือด

ภาวะแทรกซ้อนของโรคเบาหวาน

ภาวะโรคแทรกซ้อนจากโรคเบาหวานเกิดจากภาวะน้ำตาลในเลือดสูงผิดปกติเป็นระยะเวลานาน จะทำให้เกิดภาวะอักเสบ และภาวะหลอดเลือดอุดตันได้ง่ายกว่าคนปกติ นอกจากนี้ยังส่งผลต่อระบบภูมิคุ้มกันของร่างกาย ทำให้เม็ดเลือดขาวซึ่งเป็นส่วนสำคัญของระบบภูมิคุ้มกันของร่างกายมีหน้าที่หลักในการกำจัดเชื้อโรค มีประสิทธิภาพในการต่อสู้กับเชื้อโรคต่าง ๆ ได้ลดลง ทำให้เกิดภาวะติดเชื้อต่าง ๆ ได้ง่าย และส่งผลต่อการเกิดโรคแทรกซ้อนได้

ภาวะแทรกซ้อนหลัก ๆ ที่พบได้ในผู้ป่วยเบาหวาน มีดังนี้

- **ภาวะแทรกซ้อนทางตา** หรือที่เรียกกันว่า ภาวะเบาหวานขึ้นตา หรือ จอประสาทตาเสื่อม ภาวะน้ำตาลในเลือดสูงเรื้อรังส่งผลต่อจอประสาทตาทำให้เกิดจอประสาทเสื่อม ทำให้เกิดอาการตามัวและตาบอดได้
- **ภาวะแทรกซ้อนทางไต** หรือที่เรียกกันว่า ภาวะเบาหวานลงไต ส่งผลให้เกิดโรคไตเสื่อม (Nephropathy) หรือไตวายเรื้อรัง (Chronic renal failure) หากไม่ได้รับการรักษาอย่างถูกวิธี การทำงานของไตจะเสื่อมลง และอาจดำเนินไปถึงภาวะไตวายเรื้อรังจนอาจต้องทำการฟอกไตหรือผ่าตัดปลูกถ่ายไต
- **ภาวะแทรกซ้อนทางเส้นประสาท** มีอาการระบบประสาทเสื่อม (Neuropathy) ซึ่งเกิดจากหลอดเลือดแดงขนาดเล็กที่มาเลี้ยงระบบประสาทเกิดการแข็งและตีบ โดยผู้ป่วยมักมีอาการชาปลายมือปลายเท้า เมื่อเกิดบาดแผลก็จะมีโอกาสติดเชื้ออักเสบเนื่องจากภูมิคุ้มกันต่ำ และเนื่องจากมีภาวะขาดเลือดจากภาวะหลอดเลือดแดงแข็งและตีบ จึงทำให้แผลหายได้ยาก
- **เส้นเลือดแดงใหญ่อุดตัน** อาการที่พบได้บ่อยคือ มีอาการปวดขามากเมื่อเดินหรือวิ่ง หากไม่ได้รับการรักษาเป็นเวลานานอาจทำให้เกิดภาวะเส้นเลือดอุดตันจนปลายเท้าขาดเลือด ติดเชื้อ และอาจต้องตัดนิ้วเท้าหรือขาทิ้งได้
- **เส้นเลือดหัวใจตีบ** เป็นภาวะแทรกซ้อนที่พบได้บ่อยในผู้ป่วยโรคเบาหวาน และเป็นภาวะแทรกซ้อนที่รุนแรง มีอาการหัวใจวาย ความดันโลหิตต่ำ หัวใจเต้นผิดจังหวะ และเสียชีวิตอย่างเฉียบพลันได้
- **เส้นเลือดสมองตีบ** เป็นภาวะแทรกซ้อนที่รุนแรงเช่นกัน เมื่อเกิดภาวะเส้นเลือดสมองตีบ ทำให้การทำงานของสมองและเส้นประสาทบริเวณที่ขาดเลือดลดลงหรือไม่ทำงาน ส่งผลให้เกิดอัมพฤกษ์อัมพาต



รูปที่ 2.2 แสดง (ซ้าย) จอประสาทตาปกติ (ขวา) เบาหวานจอประสาทตา

(แพทย์หญิงอุษณีย์ สีพงษ์พันธ์, 2022)

การวินิจฉัยโรคเบาหวาน

ปัจจุบันการวินิจฉัยโรคเบาหวานในประเทศไทย แพทย์สามารถวินิจฉัยอาการเบื้องต้นของโรคเบาหวานได้จากประวัติอาการ ประวัติการเจ็บป่วยต่าง ๆ ประวัติการเป็นโรคเบาหวานของคนในครอบครัว และที่สำคัญมากที่สุดคือการตรวจเลือดเพื่อดูปริมาณน้ำตาลในเลือด (ผู้ป่วยโรคเบาหวานจะพบปริมาณน้ำตาลในเลือดมากกว่าคนปกติ) และมีการตรวจอื่น ๆ ประกอบไปด้วยตามความเหมาะสม เช่น การตรวจปัสสาวะเพื่อดูน้ำตาลในปัสสาวะ (ซึ่งจะไม่พบในคนปกติ) การตรวจเลือดเพื่อดูการทำงานของไต (เพราะโรคเบาหวานมักส่งผลต่อการเกิดแทรกซ้อนเช่น โรคไตเรื้อรัง) และการตรวจสุขภาพตาโดยจักษุแพทย์ (เพื่อเฝ้าระวังภาวะแทรกซ้อนของโรคเบาหวานต่อจอตา)

การรักษาหรือป้องกันโรคเบาหวาน

โรคเบาหวานเป็นโรคเรื้อรัง และพบการเกิดโรคเพิ่มขึ้นทุกวัน ถึงแม้จะเป็นโรคที่ก่อให้เกิดภาวะแทรกซ้อนต่าง ๆ ตามมาได้มากมาย แต่ก็เป็นโรคที่สามารถควบคุมได้ เช่น ควบคุมระดับน้ำตาลให้อยู่ในเกณฑ์ปกติ การออกกำลังกายอย่างสม่ำเสมอ ควบคุมน้ำหนักให้อยู่ในเกณฑ์คงที่ รับประทานอาหารให้ครบ 5 หมู่ ในสัดส่วนที่เหมาะสม และหลีกเหลี่ยงการดื่มเครื่องดื่มแอลกอฮอล์ หากผู้ป่วยตระหนักถึงความสำคัญของโรค และให้ความร่วมมือในการรักษาและติดตามอาการอยู่เสมอ ก็จะทำให้ผู้ป่วยมีคุณภาพชีวิตที่ดีสามารถใช้ชีวิตประจำวันและทำกิจกรรมต่าง ๆ รวมถึงทำงานประจำได้ตามปกติ

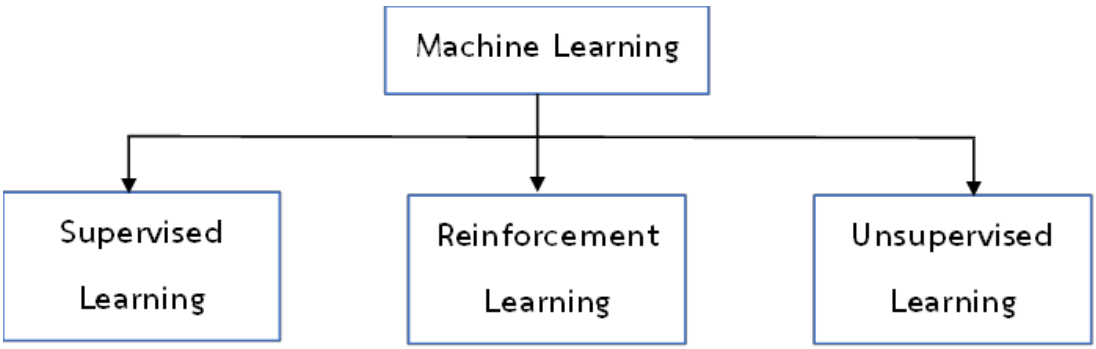
2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือส่วนการเรียนรู้ของเครื่องคอมพิวเตอร์ เป็นเครื่องมือหนึ่งของปัญญาประดิษฐ์ ถูกใช้งานเสมือนเป็นสมองในการสร้างความฉลาด โดยมุ่งเน้นการสอนให้ระบบคอมพิวเตอร์สามารถเรียนรู้ด้วยตนเองในการใช้ตัวอย่างหรือประสบการณ์เพื่อการเรียนรู้งาน มนุษย์มีส่วนร่วมเพียงการออกแบบระบบเท่านั้น หลังจากนั้นระบบจะสกัดสาระสำคัญจากตัวอย่างหรือประสบการณ์เหล่านี้เอง หลังจากการเรียนรู้เสร็จสิ้นด้วยตัวอย่างหรือประสบการณ์จำนวนหนึ่งอย่างเพียงพอเพื่อให้ทำงานนั้น ๆ ได้อย่างมีประสิทธิภาพ อีกทั้งยังสามารถเพิ่มประสิทธิภาพได้จากการเรียนรู้จากตัวอย่างหรือประสบการณ์ที่เพิ่มขึ้นได้

ประเภทของการเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่องแบ่งออกเป็น 3 ประเภทดังนี้

- การเรียนรู้แบบมีการสอน (Supervised Learning)
- การเรียนรู้แบบไม่ต้องมีการสอน (Unsupervised Learning)
- การเรียนรู้อาศัยการป้อนกลับผลลัพธ์ (Reinforcement Learning)



รูปที่ 2.3 ประเภทของการเรียนรู้ของเครื่อง

การเรียนรู้แบบมีการสอน (Supervised Learning)

เป็นการเรียนรู้แบบที่ต้องมีการสอน (Train) คือ ให้คอมพิวเตอร์เรียนรู้จากข้อมูลที่เข้าไปสอน (ข้อมูลที่นำเข้าไปสอน เรียกว่า Training data หรือ Training set) เพื่อที่คอมพิวเตอร์สามารถวิเคราะห์สิ่งต่าง ๆ ได้ โดยคอมพิวเตอร์จะทำการเรียนรู้ จดจำ เมื่อพบสิ่งที่มีลักษณะคล้ายกับสิ่งที่เคยถูกกำหนดไว้ ถ้าเกิดระบุผิดพลาดทางคอมพิวเตอร์ก็จะทำการระบุว่าสาเหตุใดจึงผิด และทำการเรียนรู้ต่อไปเรื่อย ๆ เพื่อให้เกิดความแม่นยำมากขึ้น

การเรียนรู้แบบมีการสอนแบ่งออกเป็น 2 ประเภทดังนี้

- **การจำแนกประเภท** (Classification) คือการจำแนก การแบ่งแยกประเภท คัดแยก เช่น ระบบการ คัดแยก Spam mail และ ระบบคัดแยกภาพเนื้อเยื่อว่าป่วยหรือไม่ป่วย
- **การถดถอย** (Regression) เป็นการคำนวณเพื่อทำนายค่าโดยมีการวัดเป็นตัวเลขได้ เช่น ทำนายการขึ้นลงของราคาหุ้น และ ทำนายค่าความดันโลหิตจากข้อมูลด้านสุขภาพ

การเรียนรู้แบบไม่ต้องมีการสอน (Unsupervised Learning)

เป็นการเรียนรู้แบบที่ไม่ต้องมีการสอน ไม่จำเป็นต้องใช้ข้อมูลที่มี Output หรือไม่ต้องมีข้อมูล Label/Target โดยลักษณะการทำงานคือป้อนข้อมูลที่ต้องการทำนาย จากนั้นระบบจะทำการประมวลผลข้อมูลให้เอง

การเรียนรู้แบบไม่ต้องมีการสอนแบ่งออกเป็น 2 ประเภทดังนี้

- **การจัดกลุ่ม** (Clustering) เป็นกระบวนการจัดกลุ่ม (grouping) สิ่งที่มีลักษณะคล้ายกันให้อยู่ด้วยกัน
- **การหาความสัมพันธ์** (Association) เป็นขั้นตอนวิธีในการหาความสัมพันธ์ซ่อน (hidden relation) ของชุดข้อมูลที่สนใจ

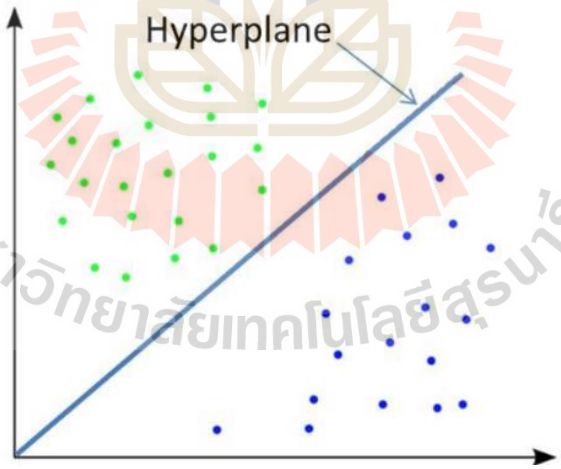
การเรียนรู้อาศัยการป้อนกลับผลลัพธ์ (Reinforcement Learning)

เป็นระบบการเรียนรู้ที่อาศัยการป้อนกลับแล้วให้ระบบเรียนรู้แล้วปรับปรุงตัวเอง โดยมีลักษณะการใช้งานเฉพาะด้าน ซึ่งเป็นการเรียนรู้ที่มีกลไกการเสริมแรงเพื่อให้คอมพิวเตอร์มีพฤติกรรมที่เราต้องการ

2.3 แบบจำลองสำหรับการจำแนกประเภท (Models for Classification)

2.3.1 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine : SVM)

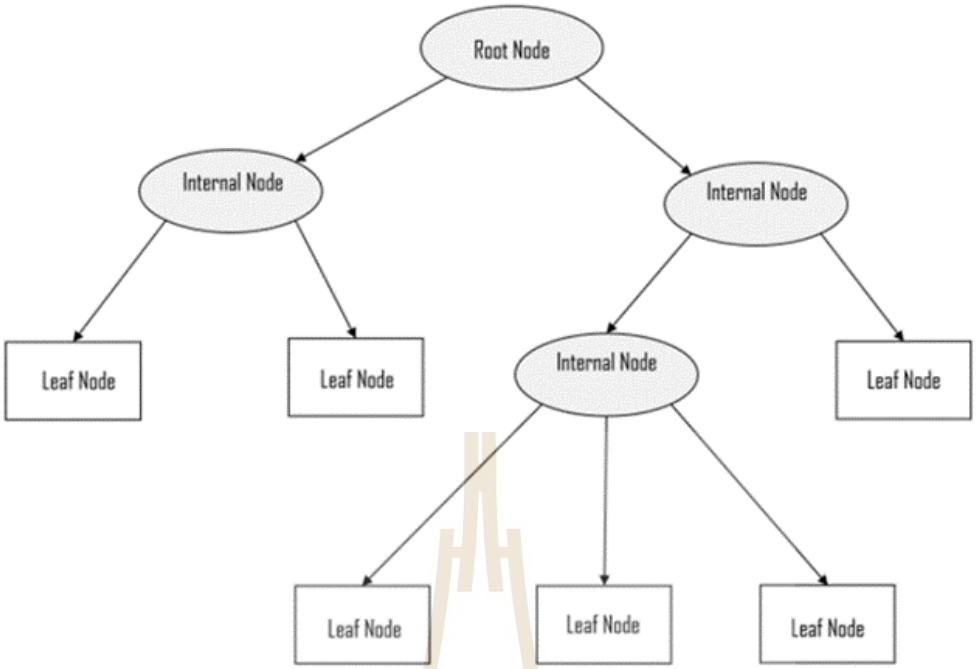
ซัพพอร์ทเวกเตอร์แมชชีน เป็นหนึ่งในขั้นตอนวิธีการเรียนรู้ของเครื่องที่ได้รับความนิยมมากที่สุดโดยใช้เทคนิคการเรียนรู้แบบมีการสอน สามารถใช้กับงานลักษณะการจำแนกประเภท และการถดถอยได้ อย่างไรก็ตามโดยหลักแล้วจะใช้สำหรับปัญหาการจำแนกประเภท ซึ่งเป็นขั้นตอนวิธีที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล และการวิเคราะห์ข้อมูล ข้อได้เปรียบคือมีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมากได้ หลักการทำงานคือการนำค่าของกลุ่มข้อมูลมาพล็อตลงในพื้นที่เรียกว่าฟีเจอร์สเปซ (Feature Space) จากนั้นหาแนวขอบเขตการตัดสินใจที่ดีที่สุดเรียกว่าไฮเปอร์เพลน (Hyper plan) ที่ใช้แบ่งข้อมูลกลุ่มออกจากกัน โดยจะสร้างแนวที่ดีที่สุดในการแบ่งกลุ่ม



รูปที่ 2.4 Support Vector Machine (SVM) (PradyaSin, 2019)

2.3.2 ต้นไม้ตัดสินใจ (Decision Tree)

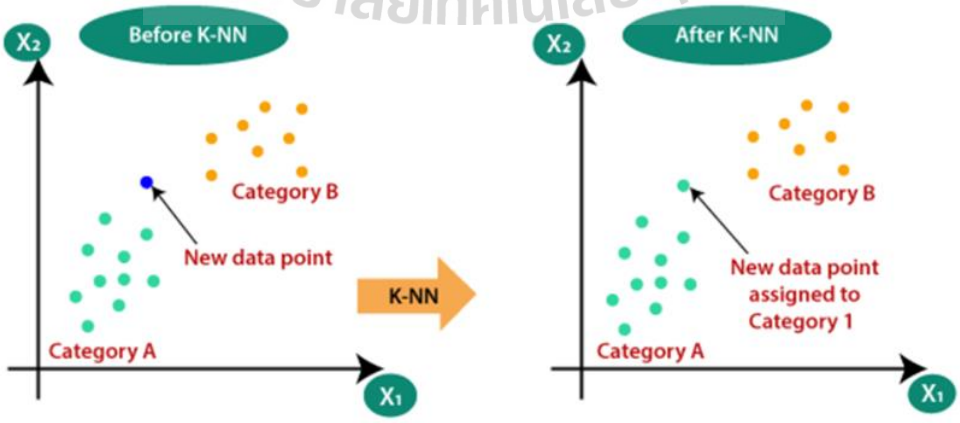
ต้นไม้ตัดสินใจ คือโครงสร้างต้นไม้เป็นขั้นตอนวิธีพื้นฐานที่สำคัญตัวหนึ่งของการเรียนรู้ของเครื่อง ใช้เทคนิคการเรียนรู้แบบมีการสอน เหมาะสำหรับงานลักษณะการจำแนกประเภทว่าข้อมูลจัดอยู่ในกลุ่มใด หลักการทำงานคล้ายกับการถามตอบ เป็นการเรียนรู้จากคุณลักษณะของข้อมูล (Attributes) แล้วสร้างผังการตัดสินใจคล้ายกับต้นไม้ จึงเรียกว่าต้นไม้ตัดสินใจ โดยผลลัพธ์จะมีเพียงสองกลุ่ม หรือมากกว่าก็ได้



รูปที่ 2.5 โครงสร้างของต้นไม้ตัดสินใจ (Huawei, 2021)

2.3.3 เทคนิคเพื่อนบ้านใกล้ที่สุด (k – Nearest Neighbors : kNN)

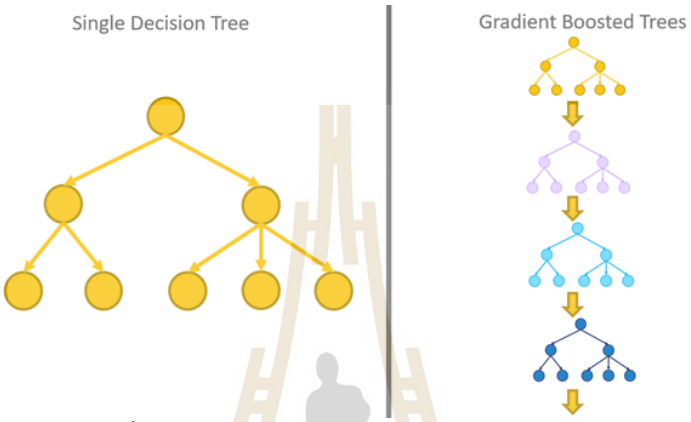
เทคนิคเพื่อนบ้านใกล้ที่สุด เป็นขั้นตอนวิธีประเภทหนึ่งของการเรียนรู้ของเครื่องที่ง่ายที่สุดโดยใช้เทคนิคการเรียนรู้แบบมีการสอน สามารถใช้กับงานลักษณะการจำแนกประเภท และการถดถอยได้ แต่ส่วนใหญ่จะใช้สำหรับปัญหาการจำแนกประเภท หลักการทำงานคือการเทียบหาข้อมูลจุดใหม่ ถ้าพบว่าอยู่ใกล้กับกลุ่มใด ก็จะจัดให้ข้อมูลใหม่อยู่ในกลุ่มนั้น หมายความว่าเมื่อมีข้อมูลใหม่ปรากฏขึ้น ก็สามารถจำแนกประเภทข้อมูลได้ง่ายโดยใช้ขั้นตอนวิธี k-NN (ชื่อ Nearest Neighbors หมายถึงจัดให้เข้ากลุ่มกับเพื่อนบ้านที่ใกล้ที่สุด)



รูปที่ 2.6 เทคนิคเพื่อนบ้านใกล้ที่สุด (javaTpoint, 2021)

2.3.4 เกรเดียนท์บูตทรี (Gradient Boosted Trees)

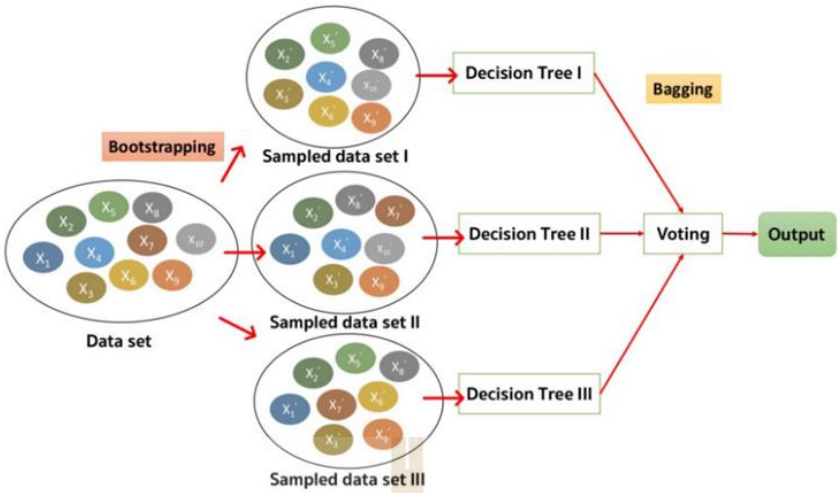
เกรเดียนท์บูตทรี เป็นขั้นตอนวิธีประเภทหนึ่งของการเรียนรู้ของเครื่องที่สามารถใช้กับงานลักษณะการจำแนกประเภท และการถดถอยได้ เป็นขั้นตอนวิธีที่มีพื้นฐานมาจากต้นไม้ตัดสินใจซึ่งเป็นการปรับปรุงประสิทธิภาพให้มีค่าสูงขึ้น โดยการสุ่มสร้างต้นไม้ตัดสินใจหลายร้อยแบบจำลอง และประเมินผลแต่ละแบบจำลองจนกว่าจะได้ต้นไม้ตัดสินใจที่สมบูรณ์



รูปที่ 2.7 เกรเดียนท์บูตทรี (thedatascientist)

2.3.5 ป่าสุ่ม (Random Forest)

ป่าสุ่มเป็นขั้นตอนวิธีประเภทหนึ่งของการเรียนรู้ของเครื่องที่ได้รับความนิยมอย่างมาก ซึ่งถูกพัฒนาขึ้นจากต้นไม้ตัดสินใจ ต่างกันที่ป่าสุ่มเป็นการเพิ่มจำนวน ต้นไม้เป็นหลาย ๆ ต้น ทำให้มีประสิทธิภาพในการทำงานสูงขึ้น แม่นยำมากขึ้น ซึ่งขั้นตอนวิธีป่าสุ่มเป็นหนึ่งในเทคนิคที่เรียกว่า Bagging (ช่วยลดปัญหาการเกิด Overfit ของขั้นตอนวิธีต้นไม้ตัดสินใจ) หลักการป่าสุ่มคล้ายต้นไม้หรือต้นไม้ตัดสินใจปกติ แต่จะสุ่มเอาข้อมูล (Instance) ไปสร้างเป็นต้นไม้หลาย ๆ ต้น แต่ละต้นเรียกว่า Subset เหมือนกับว่ามีป่าที่มีต้นไม้จำนวนมาก ๆ ซึ่งแต่ละต้นจะมีรูปแบบสุ่มไม่เหมือนกัน ในตอนทำงานจะให้แต่ละต้นทำนาย และคำนวณผลการทำนายด้วยการโหวตผลลัพธ์ที่ถูกเลือกมากที่สุด



รูปที่ 2.8 โครงสร้างของป่าสุ่ม (Watchapong Daroontham, 2018)

2.4 ตัวชี้วัดประสิทธิภาพสำหรับระบบการเรียนรู้ของเครื่องที่ทำงานในลักษณะจำแนกประเภท

2.4.1 ตาราง Confusion Matrix

ตาราง Confusion Matrix คือตารางที่ใช้ประเมินผลลัพธ์การทำนายหรือผลลัพธ์จากโปรแกรม (Prediction) เปรียบเทียบกับค่าจริง ๆ (Actual)

True Positive (TP): ทำนายถูกต้อง สำหรับผลในทางบวก (บอกว่าเป็น Yes หรือ ใช่)

True Negative (TN): ทำนายถูกต้อง สำหรับผลในทางลบ (บอกว่าเป็น No หรือ ไม่ใช่)

False Positive (FP): ทำนายผิด โดยให้ผลเป็นบวก (ของจริงไม่ใช่ แต่บอกใช่)

False Negative (FN): ทำนายผิด โดยให้ผลเป็นลบ (ของจริงใช่ แต่บอกไม่ใช่)

ตารางที่ 2.1 ตาราง Confusion Matrix ของระบบการจำแนกประเภทแบบ 2 Classes

		ทำนาย (Prediction)	
		Negative (0)	**Positive (1)**
ของจริง (Actual)	**Negative**	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

2.4.2 ความแม่น (Accuracy)

ความแม่น คือสัดส่วนเปอร์เซ็นต์ความถูกต้อง คำนวณได้จากอัตราส่วนจำนวนที่ทำนายถูกต่อจำนวนทั้งหมด

$$Accuracy = \frac{True\ Positive\ (TP)\ +\ True\ Negative\ (TN)}{All} \quad (2.1)$$

2.4.3 ความเที่ยง (Precision)

ความเที่ยง เป็นการสนใจว่าในการทำนายผลนั้น การทำนายที่ถูกต้องสำหรับผลที่เป็นบวกมีความคงเส้นคงวาแค่ไหน คำนวณโดยสัดส่วนของการทำนายที่ถูกต้องสำหรับผลที่เป็นบวกหารด้วยจำนวนการทำนายเฉพาะผลที่เป็นบวกทั้งหมด

$$Precision\ = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Positive\ (FP)} \quad (2.2)$$

2.4.4 ค่าเรียกคืน (Recall)

ค่าเรียกคืน เรียกอีกอย่างหนึ่งว่า ความไว (Sensitivity) หรือ อัตราผลบวกจริง (True Positive Rate: TPR) คือการวัดความแม่นยำอีกมิติหนึ่ง ที่สนใจด้านการทำนายที่ถูกต้องเป็นหลัก เป็นการพิจารณาการทำนายที่ถูกต้องสำหรับผลที่เป็นบวกเทียบกับการทำนายที่ถูกต้องทั้งหมดของผลที่เป็นทั้งบวกและลบ ถ้าการทำนายมีค่านี้สูง แสดงว่ามีความสามารถในการตรวจจับหรือทำนายค่าที่สนใจได้ดี

$$Recall\ (Positive) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Negative\ (FN)} \quad (2.3)$$

2.4.5 อัตราผลลบจริง (True Negative Rate: TNR)

อัตราผลลบจริง เรียกอีกอย่างว่า ความเฉพาะเจาะจง (Specificity) หรือ บางครั้งอาจจะถูกเรียกว่า Recall (Negative) เป็นการพิจารณาการทำนายที่ถูกต้องสำหรับผลที่เป็นลบเทียบกับการทำนายที่ถูกต้องทั้งหมดของผลที่เป็นทั้งบวกและลบ

$$TNR\ = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN)\ \ +\ False\ Positive\ (FP)} \quad (2.4)$$

2.4.6 อัตราผลบวกเท็จ (False Positive Rate: FPR)

อัตราผลบวกเท็จ หรือ FPR คือ ค่าที่บอกว่าโปรแกรมทำนายว่าไม่จริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมดมีความสัมพันธ์กับอัตราผลลบจริงคือ $FPR = 1 - TNR$ หรือคำนวณได้จาก

$$FPR\ = \frac{False\ Positive\ (FP)}{True\ Negative\ (TN)\ \ +\ False\ Positive\ (FP)} \quad (2.5)$$

2.4.7 F1 score

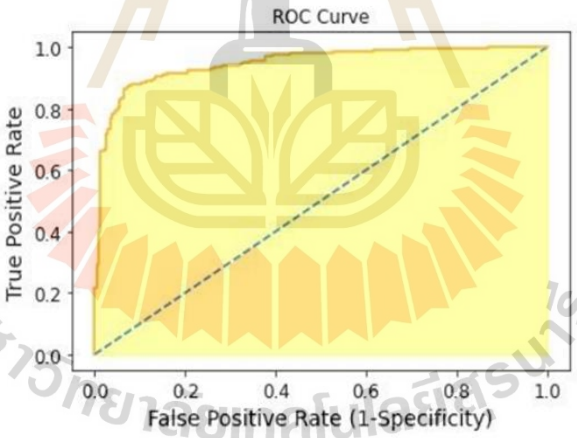
F1 score คือค่าที่แสดงประสิทธิภาพ โดยการนำความเที่ยงกับค่าเรียกคืนมาคำนวณหาค่าเฉลี่ยแบบ ฮาร์มอนิก (Harmonic mean) ซึ่งค่า F1 score ที่สูงจะแสดงว่าแบบจำลองดังกล่าวมีประสิทธิภาพดี

$$F1\ score\ =\ \frac{2 \times (Precision \times Recall)}{(Precision\ +\ Recall)} \tag{2.6}$$

2.4.8 เส้นโค้ง ROC (ROC curve) และ AUC

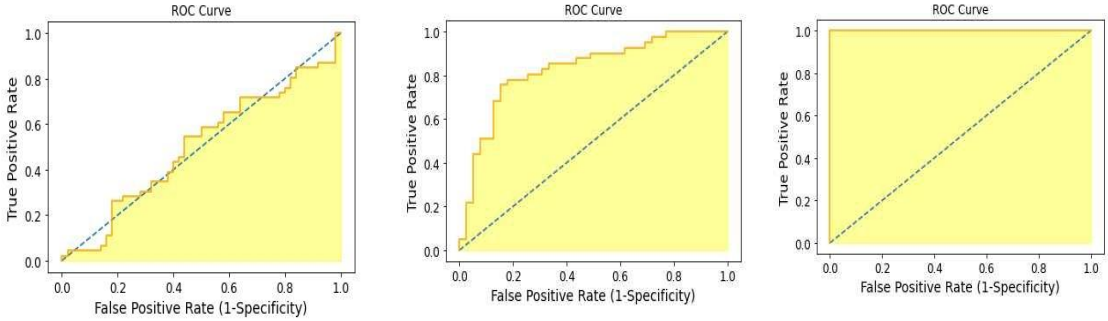
เส้นโค้งของ ROC หรือ ROC curve ซึ่งย่อมาจากคำว่า Receiver Operating Characteristic curve เป็นตัวชี้วัดที่บอกความสามารถว่าแบบจำลองสามารถจำแนกแยกแยะ (Classify) 2 กลุ่มออกจากกันได้ดีแค่ไหน เช่น กลุ่มผู้ที่ป่วยและกลุ่มที่ไม่ป่วย และยังสามารถใช้บอกจุดตัด (Cut-off) ที่เหมาะสมที่สุดเพื่อให้ได้ผลการจำแนกหรือทำนายได้แม่นยำที่สุด

สำหรับ AUC มาจากคำว่า Area Under the Curve คือค่าพื้นที่ใต้เส้น ROC curve ใช้เป็นตัวชี้วัดบอกประสิทธิภาพของ Model



รูปที่ 2.9 แสดงลักษณะ ROC curve และ AUC

กราฟ ROC curve ได้จากการพล็อตแกน x คืออัตรา False Positive Rate (FPR ทำนายผิดว่าเป็น Positive) และแกน y คือ True Positive Rate (TPR ทำนายถูกต้องว่าเป็น Positive)



รูปที่ 2.10 เปรียบเทียบประสิทธิภาพแบบจำลองในการทำนายโดยพิจารณาผ่านกราฟ ROC curve ซ้าย:แย่ กลาง:ดีปานกลาง ขวา:ดีที่สุด

2.4.9 Cohen's kappa Coefficient (**K**)

เป็นค่าสัมประสิทธิ์ตัวชี้วัดทางสถิติเพื่อใช้ประเมินความสอดคล้องระหว่างผู้ประเมิน โดยค่าสัมประสิทธิ์จะมีค่าระหว่าง -1.0 ถึง 1.0 หากค่าสัมประสิทธิ์มีค่าเป็นลบ แสดงถึงความเห็นระหว่างผู้ประเมินมีความสอดคล้องในลักษณะตรงกันข้าม ซึ่งในทางทฤษฎีมักพิจารณาค่าสัมประสิทธิ์ควรมีค่า ตั้งแต่ 0.0 แต่ไม่เกิน 1.0

$$\kappa = \frac{p_a - p_e}{1 - p_e} = 1 - \frac{1 - p_a}{1 - p_e} \qquad (2.7)$$

p_a = ความน่าจะเป็นที่ผู้ประเมินเห็นสอดคล้องกัน

p_e = ความน่าจะเป็นที่ผู้ประเมินเห็นไม่สอดคล้องกัน

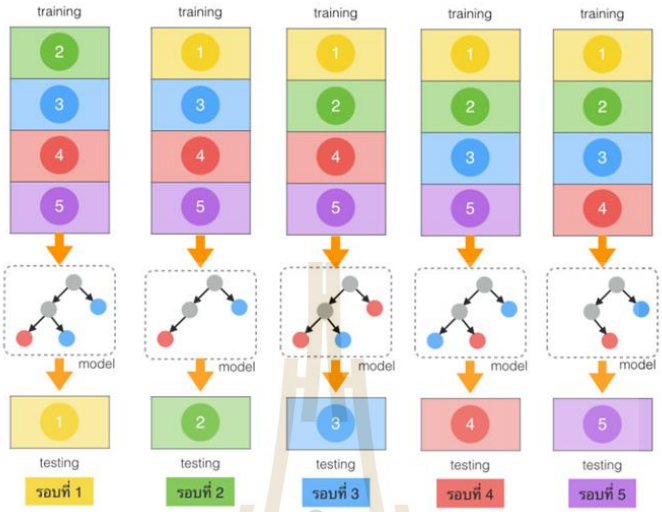
ตารางที่ 2.2 ระดับความสอดคล้องของ Cohen's kappa Coefficient (**K**)

ค่าสัมประสิทธิ์ของ K	**ระดับความสอดคล้องระหว่างผู้ประเมิน**
0.81 – 1.00	ความสอดคล้องดีมาก (Almost Perfect)
0.61 – 0.80	ความสอดคล้องดี (Substantial)
0.41 – 0.60	ความสอดคล้องปานกลาง (Moderate)
0.21 – 0.40	ความสอดคล้องพอใช้ (Fair)
0.00 – 0.20	ความสอดคล้องเล็กน้อย (Slight)
น้อยกว่า 0.00	ไม่มีความสอดคล้อง (Poor)

2.5 การทดสอบแบบไขว้ (K-fold Cross validation)

วิธีการทดสอบ K-fold Cross validation เป็นวิธีที่นิยมในการทำงานวิจัย เพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลองเนื่องจากผลที่ได้มีความน่าเชื่อถือ ลักษณะการทำงานคือแบ่งข้อมูลออกเป็นหลายส่วน มักจะแสดงด้วยค่า K เช่น 5- fold Cross validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลส่วนหนึ่งจะใช้เป็น

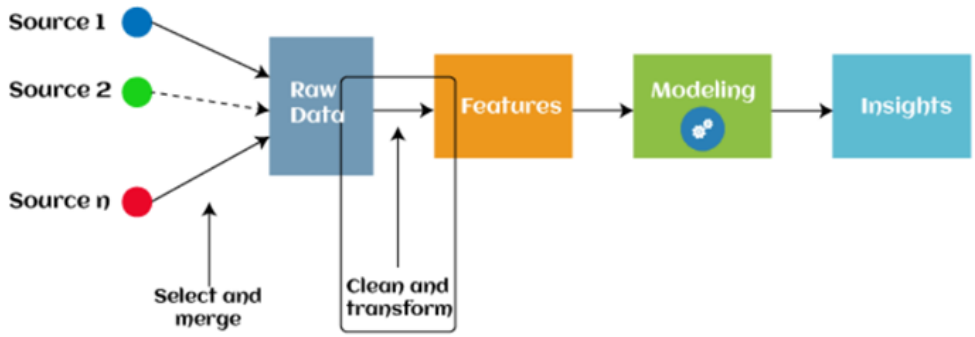
ตัวทดสอบประสิทธิภาพของแบบจำลอง ทดสอบวนซ้ำไปเช่นนี้จนครบจำนวนที่แบ่งไว้ แสดงดังรูปที่ 2.11



รูปที่ 2.11 การทดสอบแบบจำลองด้วยวิธี 5- fold Cross validation (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

2.6 วิศวกรรมคุณลักษณะ (Feature Engineering)

วิศวกรรมคุณลักษณะเป็นขั้นตอนก่อนการประมวลผลของการเรียนรู้ของเครื่อง ซึ่งจะดึงคุณลักษณะจากข้อมูลดิบ ช่วยแสดงปัญหาพื้นฐานของแบบจำลองการคาดการณ์ในทางที่ดีขึ้น ซึ่งจะส่งผลให้สามารถปรับปรุงความถูกต้องของแบบจำลองสำหรับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) แบบจำลองการทำนายประกอบด้วยตัวแปรทำนายและตัวแปรผลลัพธ์ และในขณะที่กระบวนการวิศวกรรมคุณลักษณะจะเลือกตัวแปรการทำนายที่มีประโยชน์ที่สุดสำหรับแบบจำลอง (javaTpoint, 2021)



รูปที่ 2.12 กระบวนการทำงานของวิศวกรรมคุณลักษณะ (javaTpoint, 2021)

วิศวกรรมคุณลักษณะในการเรียนรู้ของเครื่อง ประกอบด้วย 6 กระบวนการหลัก ดังนี้

- การเตรียมข้อมูล (Data Preparation)
- การสร้างคุณลักษณะ (Feature Generation)
- การแปลงคุณลักษณะ (Feature Transformation)
- การแยกคุณลักษณะ (Feature Extraction)
- การเลือกคุณลักษณะ (Feature Selection)
- วิศวกรรมคุณลักษณะอัตโนมัติ (Automatic Feature Engineering)

2.6.1 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลคือขั้นตอนแรกของกระบวนการ ในขั้นตอนนี้จะได้รับข้อมูลดิบจากแหล่งทรัพยากรต่าง ๆ ซึ่งจะถูกจัดเตรียมเพื่อให้อยู่ในรูปแบบที่เหมาะสม เพื่อให้สามารถใช้ในแบบจำลองการเรียนรู้ของเครื่องได้

การเตรียมข้อมูลหลัก ๆ มีดังนี้

- **การประเมินค่า (Imputation)** เกี่ยวข้องกับข้อมูลที่ไม่เหมาะสม ค่าที่หายไป ข้อผิดพลาดทั่วไป แหล่งข้อมูลไม่เพียงพอ ซึ่งส่งผลกระทบอย่างมากต่อประสิทธิภาพของขั้นตอนวิธี และเพื่อจัดการกับค่าเหล่านี้ มีการใช้เทคนิคการประเมินค่ามีหน้าที่จัดการกับความผิดปกติภายในชุดข้อมูล สามารถแบ่งออกเป็น 2 ประเภท คือ Categorical Imputation และ Numerical Imputation
- **การจัดการค่าผิดปกติ (Handling Outliers)** เกี่ยวข้องกับค่าผิดปกติคือค่าเบี่ยงเบนหรือจุดข้อมูลที่สังเกตได้ห่างจากจุดข้อมูลอื่นมากเกินไปในลักษณะที่ส่งผลเสียต่อประสิทธิภาพของแบบจำลอง สามารถใช้ค่าเบี่ยงเบนมาตรฐาน (Standard deviation) เพื่อระบุค่าผิดปกติได้
- **การแปลงลอการิทึม (Log transform)** ช่วยในการจัดการข้อมูลที่บิดเบี้ยว และทำให้การกระจายใกล้เคียงกับปกติมากขึ้นหลังการแปลง นอกจากนี้ยังลดผลกระทบของค่าผิดปกติต่อข้อมูล
- **การแบ่งข้อมูล (Binning)** เกี่ยวข้องกับการใส่ข้อมูลมากเกินไปเป็นหนึ่งในปัญหาหลักที่ทำให้ประสิทธิภาพของแบบจำลองลดลง และเกิดขึ้นเนื่องจากพารามิเตอร์จำนวนมากขึ้นและข้อมูลที่มีเสียงรบกวน สามารถใช้การแบ่งข้อมูลเพื่อทำให้ข้อมูลที่มีเสียงดังเป็นปกติได้
- **การแยกคุณลักษณะ (Feature Split)** เกี่ยวข้องกับการแยกคุณลักษณะออกเป็นสองส่วนหรือมากกว่า และดำเนินการเพื่อสร้างคุณลักษณะใหม่ การแยกคุณลักษณะช่วยให้คุณลักษณะใหม่สามารถจัดกลุ่มและรวมเข้าด้วยกัน ช่วยให้ขั้นตอนวิธีเข้าใจและเรียนรู้รูปแบบในชุดข้อมูลได้ดีขึ้น
- **การเข้ารหัส (One hot encoding)** เกี่ยวข้องกับการแปลงข้อมูลในรูปแบบหมวดหมู่เพื่อให้ขั้นตอนวิธีการเรียนรู้ของเครื่องสามารถเข้าใจได้ง่าย ช่วยให้การจัดกลุ่มข้อมูลแบบหมวดหมู่โดยไม่สูญเสียข้อมูลใด ๆ ส่งผลให้แบบจำลองสามารถทำนายได้ดีขึ้น

2.6.2 การสร้างคุณลักษณะ (Feature Generation)

การสร้างคุณลักษณะคือการค้นหาตัวแปรที่มีประโยชน์ที่สุดเพื่อใช้ในแบบจำลอง การสร้างคุณลักษณะใหม่นี้สร้างขึ้นโดยการผสมคุณลักษณะที่มีอยู่ซึ่งสามารถทำได้โดยการดำเนินการทางคณิตศาสตร์ เช่น การบวก การลบ การคูณ และการหาร

2.6.3 การแปลงคุณลักษณะ (Feature Transformation)

การแปลงคุณลักษณะคือขั้นตอนที่เกี่ยวข้องกับการปรับตัวแปร โดยยังคงคุณสมบัติของข้อมูลเดิมไว้ ช่วยให้มั่นใจว่าคุณลักษณะทั้งหมดอยู่ในช่วงที่ยอมรับได้เพื่อหลีกเลี่ยงข้อผิดพลาดในการคำนวณ เช่น การทำให้ข้อมูลมีการกระจายแบบปกติ (normal distribution), การทำให้ข้อมูลกระจายอยู่ในช่วง 0 - 1 (min – max normalization) ทำให้แบบจำลองเข้าใจง่ายขึ้น และปรับปรุงประสิทธิภาพและความแม่นยำของแบบจำลอง

2.6.4 การแยกคุณลักษณะ (Feature Extraction)

การแยกคุณลักษณะเป็นกระบวนการทางวิศวกรรมคุณลักษณะอัตโนมัติที่สร้างตัวแปรใหม่โดยการดึงข้อมูลจากข้อมูลดิบ เป้าหมายหลักของขั้นตอนนี้คือการลดปริมาณข้อมูลเพื่อให้สามารถใช้และจัดการสำหรับการสร้างแบบจำลองข้อมูลได้อย่างง่ายดาย วิธีการแยกคุณลักษณะประกอบด้วยการวิเคราะห์กลุ่ม (Cluster Analysis) การวิเคราะห์ข้อความ (Text Analytics) ขั้นตอนวิธีการตรวจจับขอบ (Edge Detection Algorithms) และการวิเคราะห์ส่วนประกอบหลัก (Principal Components Analysis: PCA) (javaTpoint, 2021)

2.6.5 การเลือกคุณลักษณะ (Feature Selection)

การเลือกคุณลักษณะคือการระบุหรือการเลือกคุณลักษณะที่เหมาะสมที่สุดในชุดข้อมูลเพื่อให้มีประโยชน์สำหรับการสร้างแบบจำลอง และลบคุณลักษณะที่ไม่เกี่ยวข้องหรือมีความสำคัญน้อย เป็นเทคนิคในการลดจำนวนคุณลักษณะลง หรือป้องกันการ Overfitting ของข้อมูลเนื่องจากมิติที่มากเกินไป ส่งผลให้ประสิทธิภาพโดยรวมและความถูกต้องของแบบจำลองเพิ่มขึ้น

2.6.6 วิศวกรรมคุณลักษณะอัตโนมัติ (Automatic Feature Engineering)

วิศวกรรมคุณลักษณะอัตโนมัติเป็นกระบวนการสร้างแบบจำลองที่สามารถเพิ่มประสิทธิภาพ และความแม่นยำ รวมถึงสามารถลดความซับซ้อนของแบบจำลอง ขั้นตอนวิธีที่อยู่เบื้องหลังของวิธีนี้ เรียกว่า Deep Feature Synthesis (DFS) (Kanter and Veeramachaneni, 2015) โดยแนวคิดของวิศวกรรมคุณลักษณะอัตโนมัติ คือการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม ซึ่งใช้ตัวดำเนินการ ทางคณิตศาสตร์ และหลังจากได้คุณลักษณะเหล่านี้แล้วจะทำการเลือกคุณลักษณะที่สำคัญต่อการสร้างแบบจำลอง

2.7 แนวคิดและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาและสืบค้นงานวิจัยที่เกี่ยวข้องในการทำวิทยานิพนธ์เรื่องนี้ ผู้วิจัยได้ศึกษาเอกสารงานวิจัยที่เกี่ยวข้อง ดังนี้

งานวิจัยของ Neha Prerna Tigga, Shruti Garg (2020) ได้ทำนายความเสี่ยงของโรคเบาหวานประเภท 2 โดยใช้การเรียนรู้ของเครื่องในการจำแนกประเภท มีการจำแนกประเภทโดยใช้ Logistic regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes และ Random Forest จากผลการวิจัยพบว่า Random Forest มีประสิทธิภาพในการจำแนกดีที่สุด

งานวิจัยของ Sisodia et al. (2018) ได้ออกแบบแบบจำลองที่สามารถทำนายแนวโน้มของโรคเบาหวานในผู้ป่วยเพื่อตรวจหาโรคเบาหวานในระยะเริ่มแรกได้อย่างแม่นยำสูงสุด โดยมีการใช้ขั้นตอนวิธีการจำแนกประเภท ได้แก่ Decision Tree, Support Vector Machine (SVM) และ Naïve Bayes ผลลัพธ์ที่ได้แสดงให้เห็นว่า Naive Bayes มีประสิทธิภาพความแม่นยำสูงสุดร้อยละ 76.30 เมื่อเปรียบเทียบกับขั้นตอนวิธีอื่น ๆ

งานวิจัยของวนิดา พงษ์สงวน (2561) พัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วย Decision Tree เพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน พัฒนาแบบจำลองด้วยขั้นตอนวิธีเจสี่สิบแปด (J48) เพื่อประเมินประสิทธิภาพของแบบจำลองด้วยค่าความแม่นตรง ผลการวิจัยพบว่าแบบจำลองที่พัฒนาให้ประสิทธิภาพที่มีค่าความแม่นตรงร้อยละ 76.14 และสามารถสร้างกฎการจำแนกจากต้นไม้ตัดสินใจทั้งสิ้น 97 กฎ ซึ่งพบว่าปัจจัยเสี่ยงที่อาจก่อให้เกิดโรคเบาหวาน ได้แก่ อายุ เพศ สถานะภาพ ที่อยู่ อาชีพ ประวัติความดันโลหิตเกินมาตรฐาน ประวัติค่าดัชนีมวลกายเกินมาตรฐาน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา และประวัติครอบครัวเป็นเบาหวาน

งานวิจัยของศรัณย์ชัย ศิลปะศร และคณะ (2564) ได้สร้างแบบจำลองที่ใช้ในการทำนายคะแนนแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังผ่าตัดเปลี่ยนข้อเข่าด้วยวิศวกรรมคุณลักษณะและเทคนิคเกรเดียนท์บูตทรี และนำมาเปรียบเทียบประสิทธิภาพกับแบบจำลองเชิงเส้นวางนัยทั่วไป ผลการศึกษาพบว่าการใช้เทคนิคแรนดอมฟอร์เรสและกลุ่มคุณลักษณะที่ 3 มีประสิทธิภาพดีที่สุด โดยให้ค่ารากที่สองของคลาดเคลื่อนกำลังสองเฉลี่ย ค่าคลาดเคลื่อนกำลังสองและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเป็น 5.915±0.715, 35.447±8.438 และ 4.198± 0.416

งานวิจัยของ Huda et al. (2019) ได้หาแนวทางการปรับปรุงสำหรับการตรวจหาภาวะเบาหวานขึ้นจอตาโดยใช้ความสำคัญของคุณลักษณะและอัลกอริทึมการเรียนรู้ของเครื่อง เช่น ความจำเพาะของรอยโรค (lesion-specific) (microaneurysms, exudates) การมีอยู่ของการตกเลือด (presence of hemorrhages) ของชุดข้อมูล Diabetic Retinopathy และได้ใช้อัลกอริทึม Decision Tree, Logistic Regression และ Support Vector Machine (SVM) เพื่อทำนายการมี

อยู่ของภาวะเบาหวานขึ้นจอตา พบว่าได้ผลลัพธ์ที่มีความแม่นยำโดยรวมร้อยละ 88 โดยมีค่า precision และ recall ร้อยละ 97 และร้อยละ 92 ตามลำดับ เมื่อเทียบกับผลลัพธ์ที่มีอยู่ร้อยละ72 และ 63 ตามลำดับ โดยเฉลี่ยแล้วผลลัพธ์เพิ่มขึ้น ร้อยละ 25

งานวิจัย Talha Mahboob Alam et al. (2019) ได้สร้างต้นแบบการทำนายโรคเบาหวานในระยะเริ่มต้น ซึ่งทำนายโรคเบาหวานโดยใช้คุณลักษณะที่มีนัยสำคัญ และความสัมพันธ์ของคุณลักษณะที่แตกต่างกัน จากผลการวิจัยบ่งชี้ความเกี่ยวข้องของโรคเบาหวานกับดัชนีมวลกาย (BMI) และระดับกลูโคส ซึ่งสกัดคุณลักษณะด้วยขั้นตอนวิธี Apriori method, Artificial neural network (ANN), Random Forest (RF) และ K-means clustering สำหรับการทำนายโรคเบาหวาน ผลการวิจัยพบว่า Artificial neural network (ANN) ให้ความแม่นยำสูงสุดร้อยละ 75.7

งานวิจัยของ Tawfik Beghriche et al. (2021) ได้นำเสนอระบบการตัดสินใจทางการแพทย์ที่มีประสิทธิภาพสำหรับการทำนายโรคเบาหวานโดยใช้อัลกอรึทึม Deep Neural Network (DNN) ผลลัพธ์ที่ได้พบว่าการใช้อัลกอรึทึม Deep Neural Network (DNN) ให้ความแม่นยำร้อยละ 99.75 และคะแนน F1 ร้อยละ 99.66

งานวิจัยของ Sidong Wei et al. (2018) เป็นการสำรวจที่ครอบคลุมรวมถึงอัลกอรึทึมการเรียนรู้ของเครื่องสำหรับการระบุโรคเบาหวาน ทำการสำรวจที่ครอบคลุม และใช้อัลกอรึทึมที่ได้รับความนิยมมากที่สุด เช่น Deep Neural Network (DNN), Support Vector Machine (SVM) ที่ใช้ในการระบุโรคเบาหวานและวิธีการประมวลผลข้อมูลล่วงหน้า หลังการประเมินโดย 10-fold cross-validation พบว่าอัลกอรึทึมที่มีประสิทธิภาพดีที่สุดคือ Deep Neural Network (DNN) ให้ความแม่นยำร้อยละ 77.86

งานวิจัยของ Karan Bhatia et al. (2016) เป็นนำเสนอการตัดสินใจเกี่ยวกับการปรากฏตัวของโรคภาวะเบาหวานขึ้นจอตาโดยใช้ชุดอัลกอริทึมการจำแนกประเภทของการเรียนรู้ของเครื่องกับคุณลักษณะที่ดึงออกมา (Features Extracted) จากเอาต์พุตของอัลกอริทึม เช่น การตรวจจอประสาทตา, ความจำเพาะของรอยโรค (microaneurysms, exudates), ระดับภาพ (การคัดกรองล่วงหน้า, AM/FM, การประเมินคุณภาพ) การทำนายการปรากฏตัวของภาวะเบาหวานขึ้นจอตาดำเนินการโดยใช้อัลกอริทึม Decision Tree, adaBoost, Naive Bayes, Random Forest และ Support Vector Machines (SVM)

บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้จะกล่าวถึงขั้นตอนการสกัดคุณลักษณะ โดยใช้วิธีวิศวกรรมคุณลักษณะร่วมกับเทคนิคต่าง ๆ และการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner studio

3.1 เครื่องมือที่ใช้ในการทำวิจัย

เครื่องมือและอุปกรณ์ที่ใช้ในการวิจัยครั้งนี้ คือ โปรแกรม RapidMiner Studio Educational version 9.10.001 ซึ่งทำงานบนเครื่องคอมพิวเตอร์ LAPTOP-309E40FG มีหน่วยประมวลผลกลาง Intel(R) Core(TM) I5-9300H CPU @ 2.40GHz หน่วยความจำขนาด 16 กิกะไบต์ ระบบปฏิบัติการ Windows 11

3.2 ข้อมูลที่ใช้ในการศึกษา

ข้อมูลที่ใช้ในการศึกษาได้มากจาก https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset โดยเป็นข้อมูลที่ได้จากการตอบแบบสำรวจจำนวน 70,692 รายการ ผู้ตอบแบบสำรวจแบ่งเป็นคนที่เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวาน ซึ่งมีจำนวนเท่า ๆ กัน พิจารณาได้ว่าข้อมูลมีความสมดุล ชุดข้อมูลนี้มีตัวแปรคุณลักษณะ 21 ตัวดังนี้

1. มีภาวะความดันโลหิตสูงหรือไม่
2. มีภาวะคลอเรสเตอรอลสูงหรือไม่
3. มีการตรวจคลอเรสเตอรอลหรือไม่
4. ค่าดัชนีมวลกาย (BMI)
5. สูบบุหรี่หรือไม่
6. มีภาวะเป็นโรคหลอดเลือดสมองหรือไม่
7. มีภาวะโรคหลอดเลือดหัวใจหรือไม่
8. มีการออกกำลังกายใน 30 วันที่ผ่านมาหรือไม่
9. การรับประทานผลไม้
10. การรับประทานผัก
11. ปริมาณการดื่มเครื่องดื่มแอลกอฮอล์

12. การมีประกันสุขภาพ
13. เหตุไม่ไปพบแพทย์เพราะค่าใช้จ่าย
14. ระดับของสุขภาพโดยทั่วไป
15. ภาวะสุขภาพจิต
16. ภาวะสุขภาพทางกาย
17. ภาวะการเดินขึ้นบันได
18. เพศ
19. ระดับช่วงอายุ
20. ระดับการศึกษา และ
21. ระดับเงินเดือน (สกุลเงิน US Dollar)

ตัวแปรคุณลักษณะมีรายละเอียดดังตารางที่ 3.1 อีกทั้งยังสามารถแสดงการแจกแจงของข้อมูลในแต่ละคุณลักษณะได้ดังรูปที่ 3.1-3.21

ตารางที่ 3.1 คุณลักษณะที่เกี่ยวข้องในชุดข้อมูลต้นฉบับ

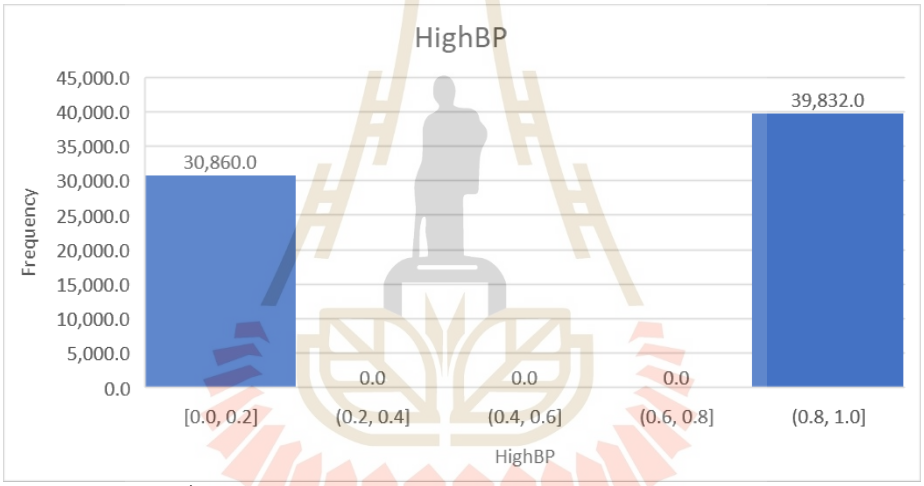
ลำดับ	คุณลักษณะ	รายละเอียด
1.	มีภาวะความดันโลหิตสูงหรือไม่	0 = ความดันโลหิตไม่สูง 1 = ความดันโลหิตสูง
2.	มีภาวะคลอเรสเตอรอลสูงหรือไม่	0 = คอเลสเตอรอลไม่สูง 1 = คอเลสเตอรอลสูง
3.	มีการตรวจคลอเรสตอรอลหรือไม่	0 = ไม่ตรวจโคเลสเตอรอลใน 5 ปี 1 = ตรวจโคเลสเตอรอลใน 5 ปี
4.	ค่าดัชนีมวลกาย (BMI)	เป็นค่าสากลที่ใช้เพื่อคำนวณเพื่อหาน้ำหนักตัวที่ควรจะเป็น
5.	สูบบุหรี่หรือไม่	คุณสูบบุหรี่อย่างน้อย 100 มวนตลอดชีวิตหรือไม่ ? [หมายเหตุ: 5 ซอง = 100 มวน] 0 = ไม่ใช่ 1 = ใช่
6.	มีภาวะเป็นโรคหลอดเลือดสมองหรือไม่	คุณเคยเป็นโรคหลอดเลือดสมองหรือไม่ ? 0 = ไม่ใช่ 1 = ใช่
7.	มีภาวะโรคหลอดเลือดหัวใจหรือไม่	โรคหลอดเลือดหัวใจ (CHD) หรือกล้ามเนื้อหัวใจตาย (MI) 0 = ไม่ใช่ 1 = ใช่
8.	มีการออกกำลังกายใน 30 วันที่ผ่านมาหรือไม่	การออกกำลังกายใน 30 วันที่ผ่านมา 0 = ไม่ใช่ 1 = ใช่

ตารางที่ 3.1 คุณลักษณะที่เกี่ยวข้องในชุดข้อมูลต้นฉบับ (ต่อ)

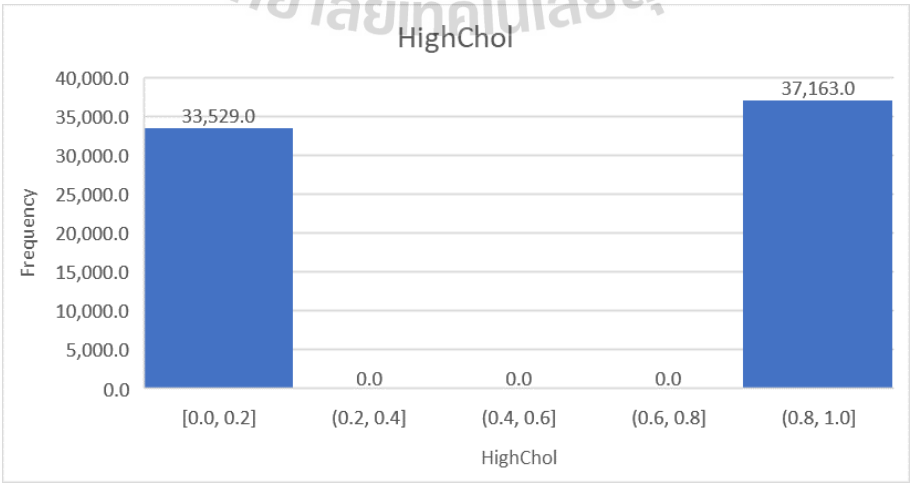
ลำดับ	คุณลักษณะ	รายละเอียด
9.	การรับประทานผลไม้	การรับประทานผลไม้ 1 ครั้งหรือมากกว่า 1 ครั้งต่อวัน 0 = ไม่ใช่ 1 = ใช่
10.	การรับประทานผัก	การรับประทานผัก 1 ครั้งหรือมากกว่า 1 ครั้งต่อวัน 0 = ไม่ใช่ 1 = ใช่
11.	ปริมาณการดื่มแอลกอฮอล์	เป็นผู้ชายที่เป็นผู้ใหญ่ที่ดื่มมากกว่า 14 แก้วต่อสัปดาห์ และผู้หญิงที่เป็นผู้ใหญ่ที่ดื่มมากกว่า 7 แก้วต่อสัปดาห์) 0 = ไม่ใช่ 1 = ใช่
12.	การมีประกันสุขภาพ	มีประกันสุขภาพแบบใดแบบหนึ่ง ทั้งประกันสุขภาพ แบบเติมเงิน เช่น HMO เป็นต้น 0 = ไม่ใช่ 1 = ใช่
13.	เหตุไม่ไปพบแพทย์เพราะค่าใช้จ่าย	มีเวลาหรือไม่ในช่วง 12 เดือนที่ผ่านมาที่คุณต้องไปพบแพทย์แต่ทำไม่ได้เพราะมีค่าใช้จ่าย? 0 = ไม่ใช่ 1 = ใช่
14.	ระดับของสุขภาพโดยทั่วไป	โดยทั่วไปสุขภาพของคุณคือ ระดับ 1-5 1 = ดีเยี่ยม 2 = ดีมาก 3 = ดี 4 = พอใช้ 5 = แย่
15.	ภาวะสุขภาพจิต	เรื่องสุขภาพจิตซึ่งรวมถึงความเครียด ความซึมเศร้า และปัญหาทางอารมณ์ ในช่วงเวลา 30 วันที่ผ่านมา
16.	ภาวะสุขภาพทางกายภาพ	สุขภาพกายซึ่งรวมถึงความเจ็บป่วยทางกายและการบาดเจ็บในช่วงเวลา 30 วันที่ผ่านมา
17.	ภาวะการเดินขึ้นบันได	คุณมีปัญหาร้ายแรงในการเดินหรือขึ้นบันไดหรือไม่ ? 0 = ไม่ใช่ 1 = ใช่
18.	เพศ	0 = หญิง 1 = ชาย
19.	ระดับอายุ	หมวดหมู่อายุ 13 ระดับ 1 = 18-24, 9 = 60-64, 13 = 80 ขึ้นไป

ตารางที่ 3.1 คุณลักษณะที่เกี่ยวข้องในชุดข้อมูลต้นฉบับ (ต่อ)

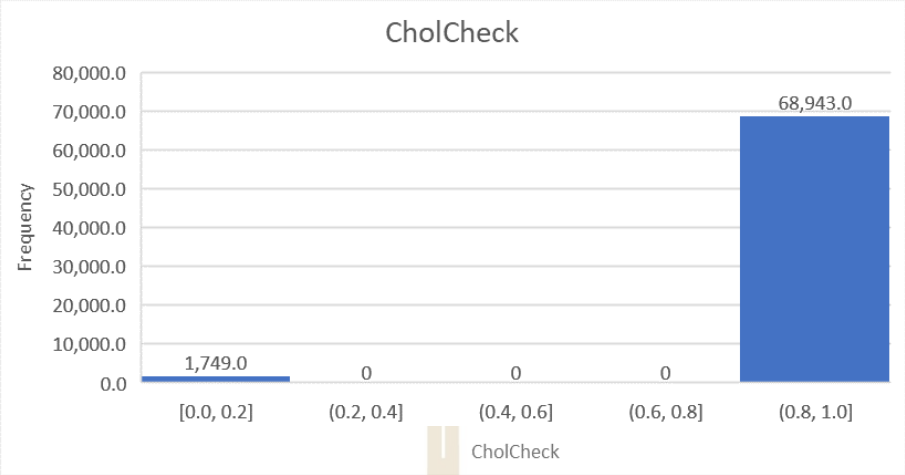
ลำดับ	คุณลักษณะ	รายละเอียด
20.	ระดับการศึกษา	ระดับการศึกษา ระดับ 1-6 1 = ไม่เคยเข้าเรียนหรือแค่อนุบาล 2 = เกรด 1 ถึง 8
21.	ระดับเงินเดือน	มาตราส่วนรายได้ มาตราส่วน 1-8 1 = น้อยกว่า 10,000 ดอลลาร์ 5 = น้อยกว่า 35,000 ดอลลาร์ 8 = 75,000 ดอลลาร์ขึ้นไป



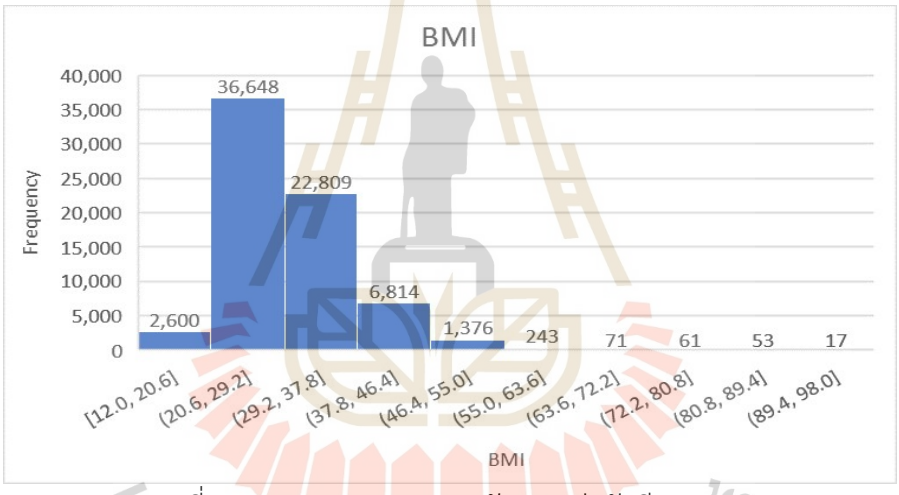
รูปที่ 3.1 การแจกแจงของคุณลักษณะภาวะความดันโลหิต



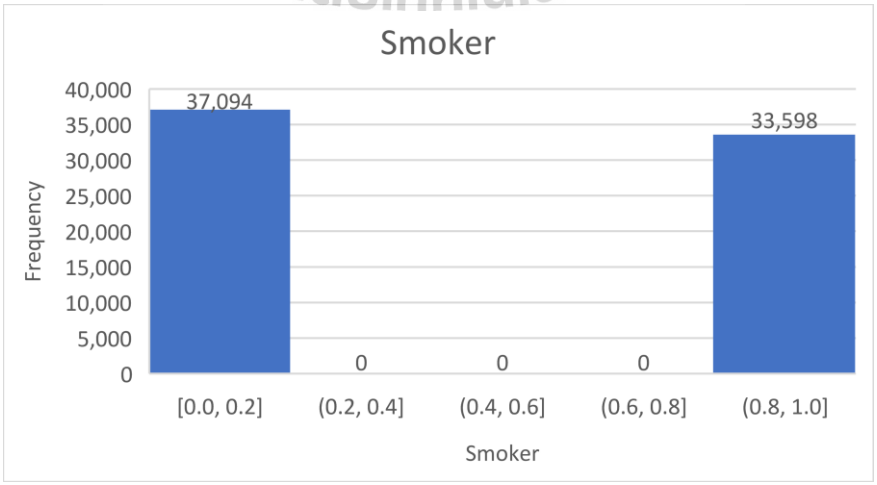
รูปที่ 3.2 การแจกแจงของคุณลักษณะภาวะคลอเรสเตอรอล



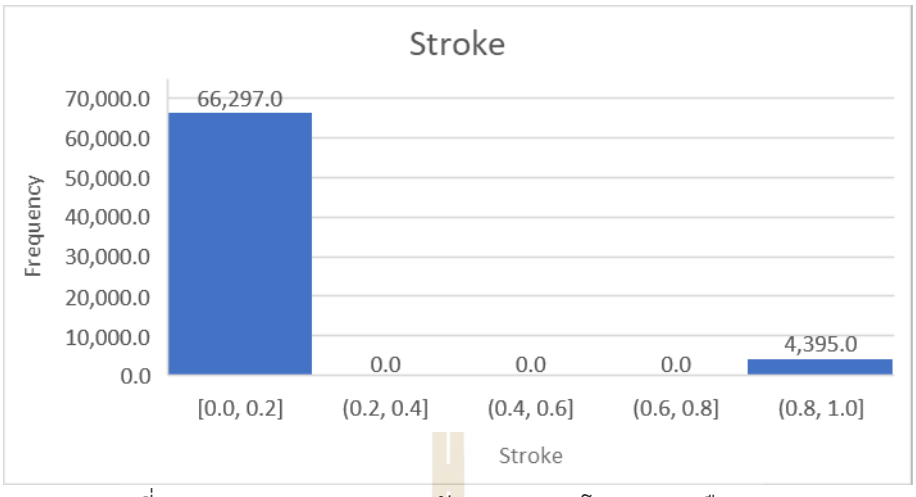
รูปที่ 3.3 การแจกแจงของคุณลักษณะการตรวจคลอเรสเตอรอล



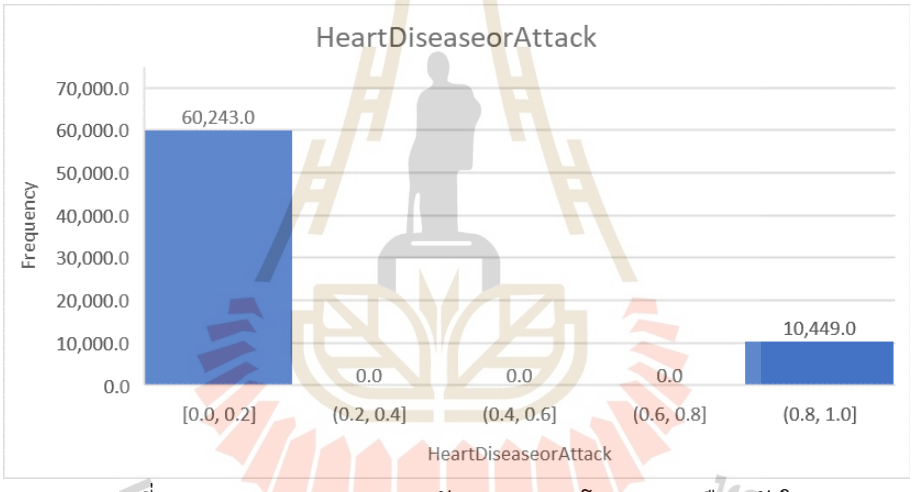
รูปที่ 3.4 การแจกแจงของคุณลักษณะค่าดัชนีมวลกาย



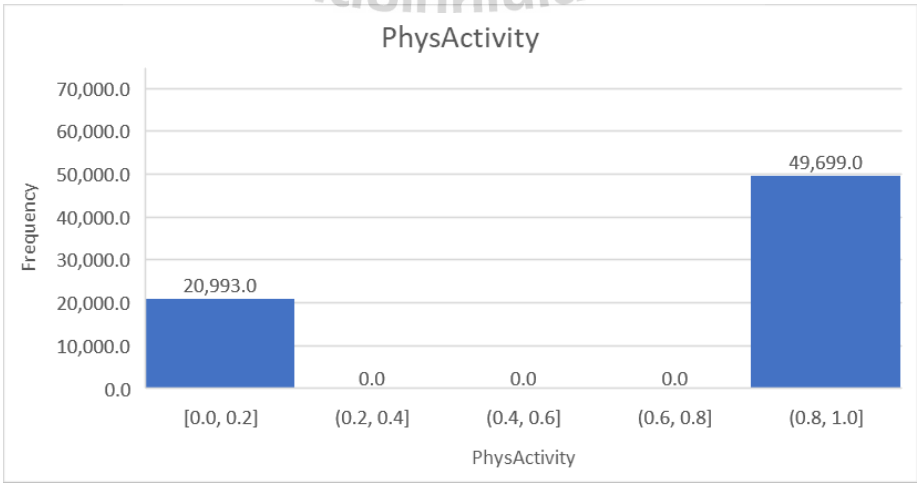
รูปที่ 3.5 การแจกแจงของคุณลักษณะการสูบบุหรี่



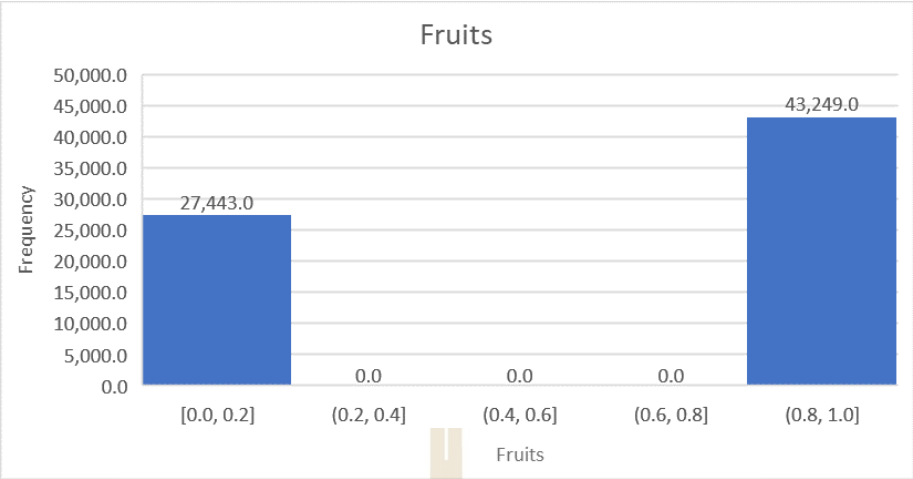
รูปที่ 3.6 การแจกแจงของคุณลักษณะภาวะโรคหลอดเลือดสมอง



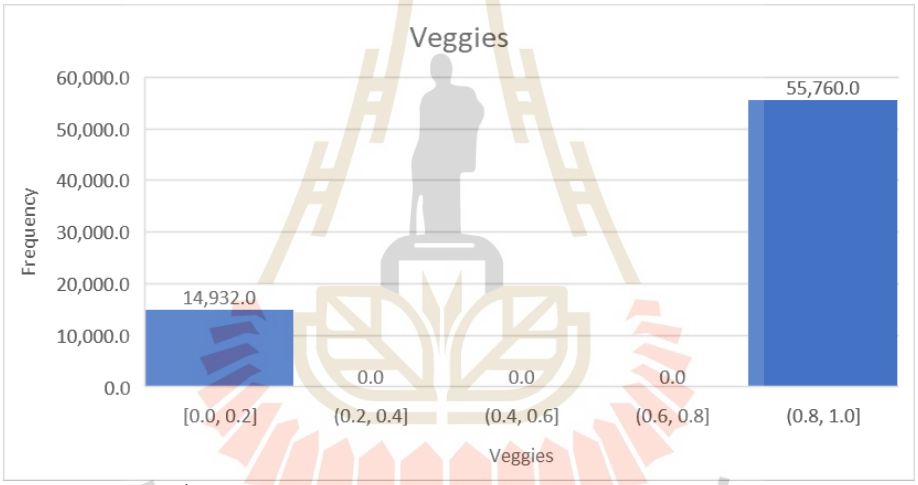
รูปที่ 3.7 การแจกแจงของคุณลักษณะภาวะโรคหลอดเลือดหัวใจ



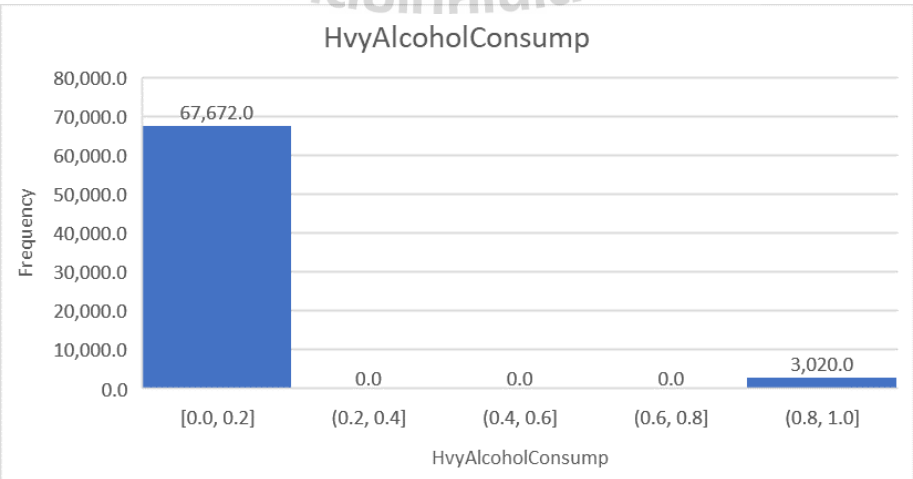
รูปที่ 3.8 การแจกแจงของคุณลักษณะการออกกำลังกาย



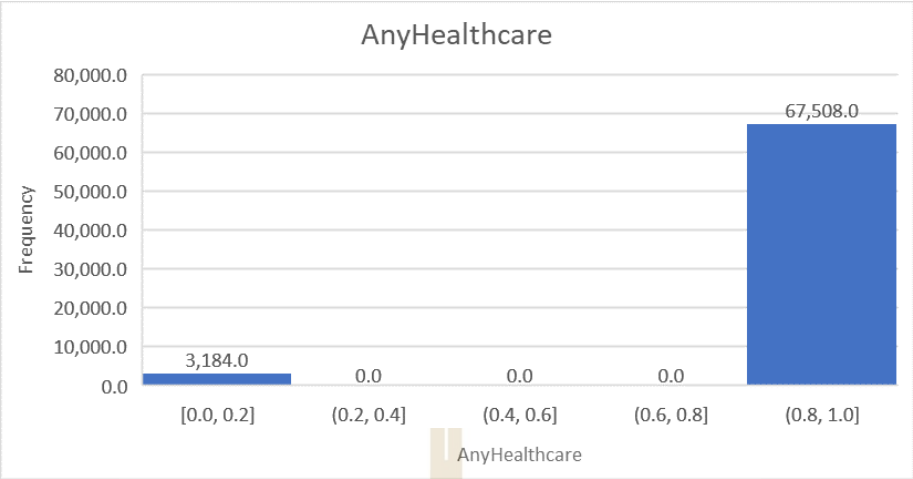
รูปที่ 3.9 การแจกแจงของคุณลักษณะการรับประทานผลไม้



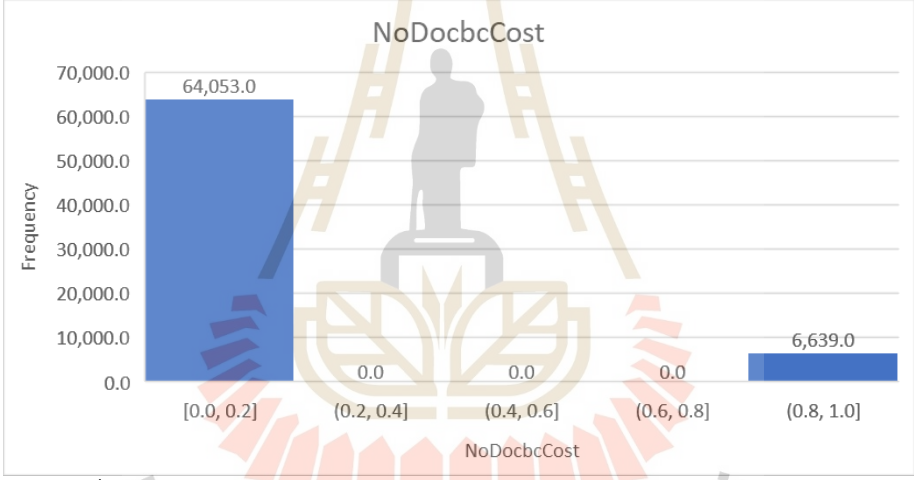
รูปที่ 3.10 การแจกแจงของคุณลักษณะการรับประทานผัก



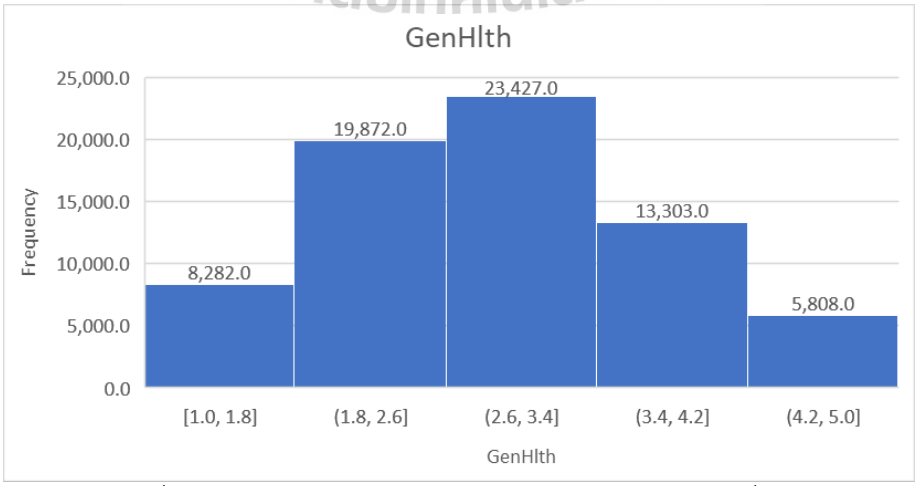
รูปที่ 3.11 การแจกแจงของคุณลักษณะปริมาณการดื่มแอลกอฮอล์



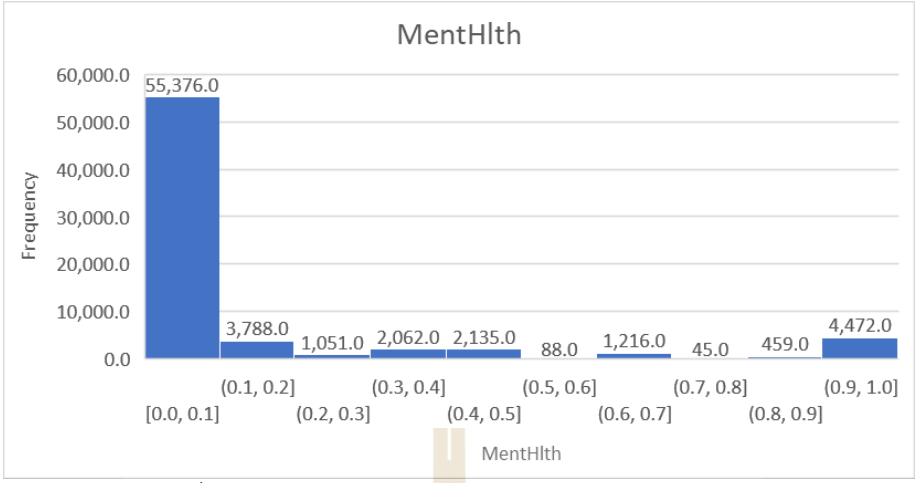
รูปที่ 3.12 การแจกแจงของคุณลักษณะการมีประกันสุขภาพ



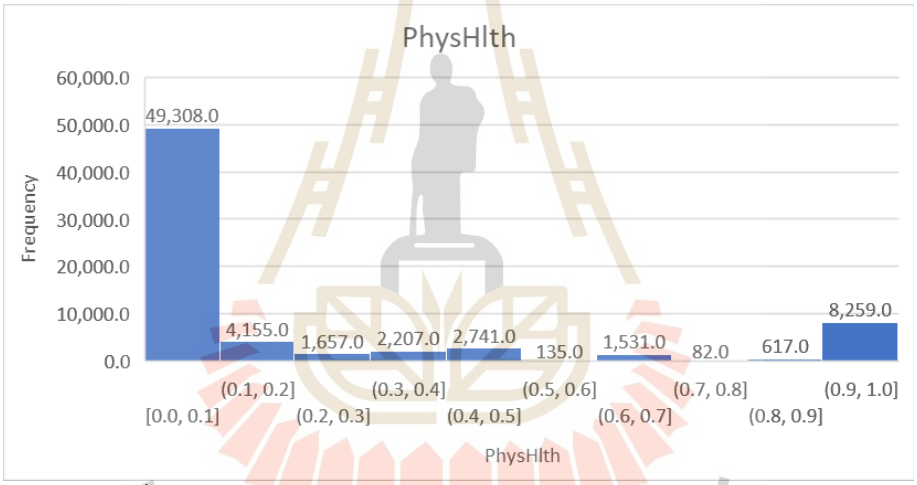
รูปที่ 3.13 การแจกแจงของคุณลักษณะเหตุไม่ไปพบแพทย์เพราะค่าใช้จ่าย



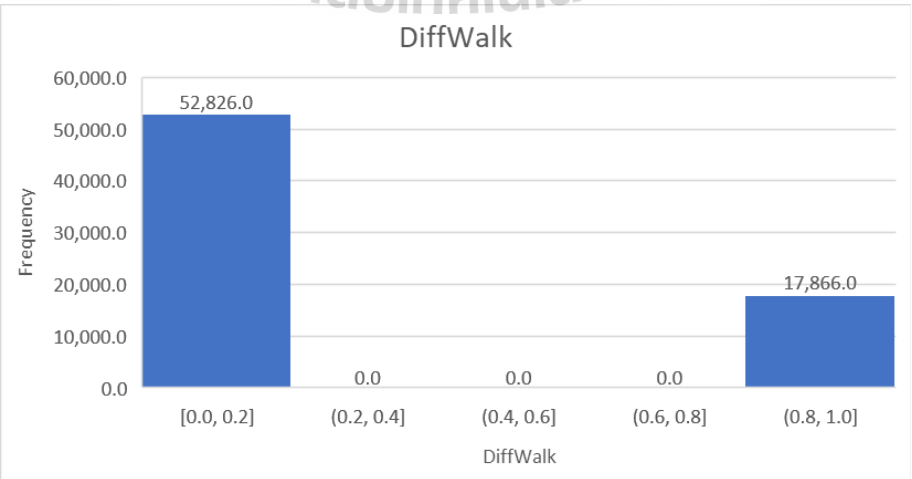
รูปที่ 3.14 การแจกแจงของคุณลักษณะระดับสุขภาพโดยทั่วไป



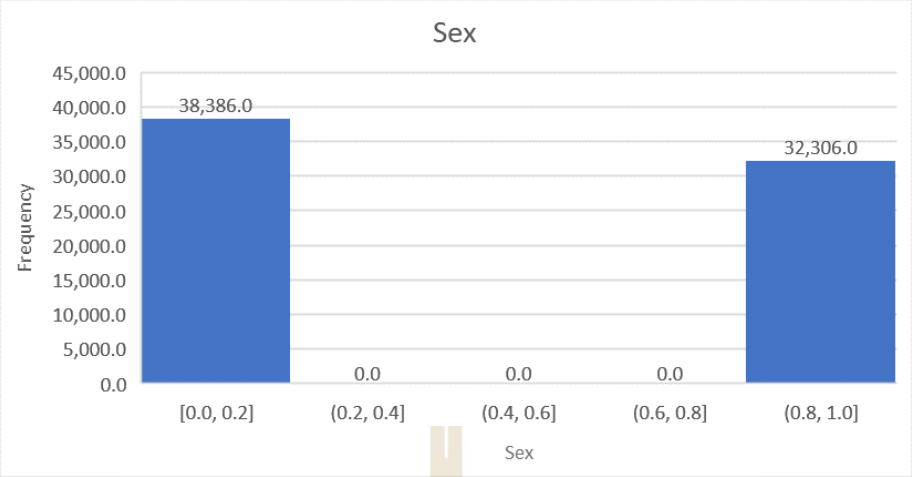
รูปที่ 3.15 การแจกแจงของคุณลักษณะภาวะสุขภาพจิต



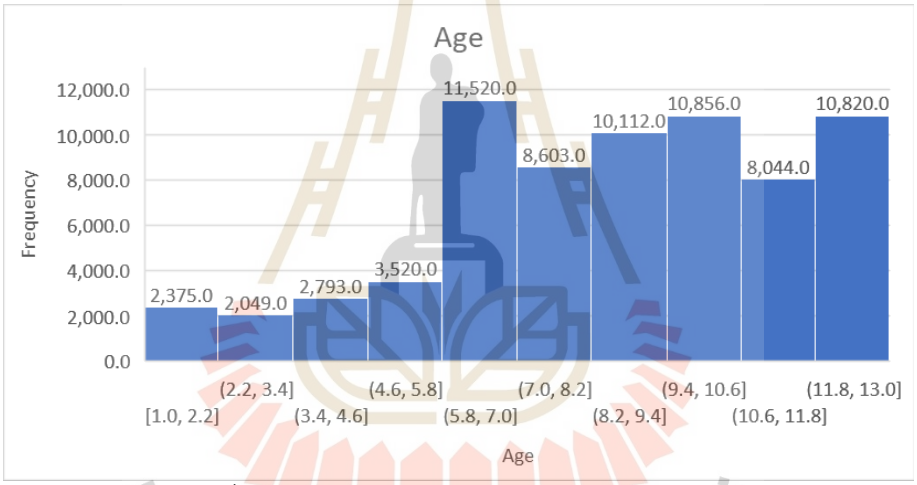
รูปที่ 3.16 การแจกแจงของคุณลักษณะภาวะสุขภาพทางกายภาพ



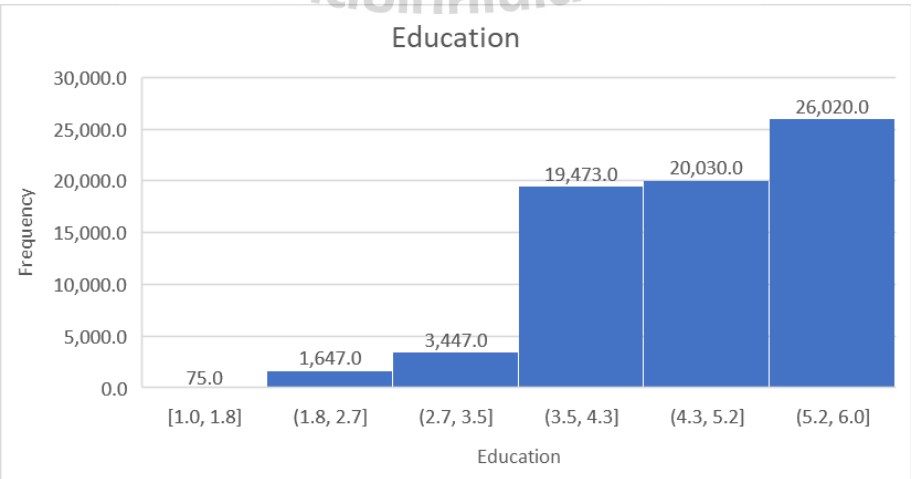
รูปที่ 3.17 การแจกแจงของคุณลักษณะภาวะการเดินขึ้นบันได



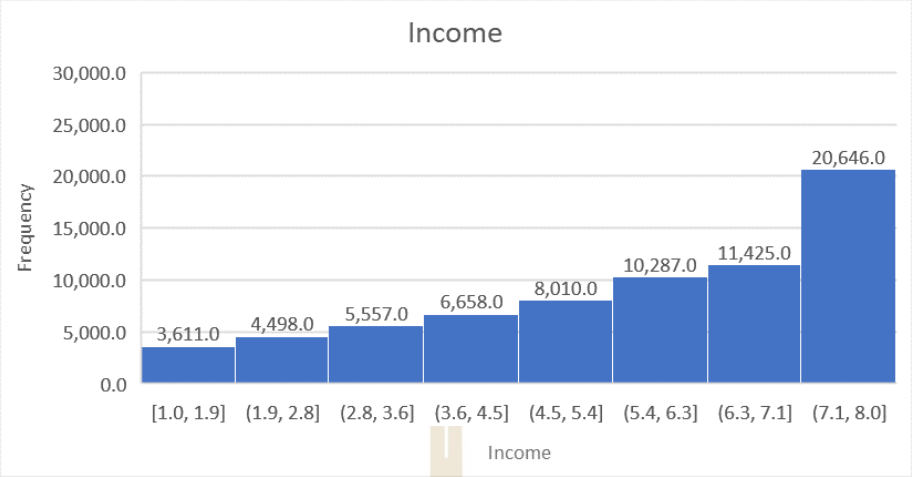
รูปที่ 3.18 การแจกแจงของคุณลักษณะทางเพศ



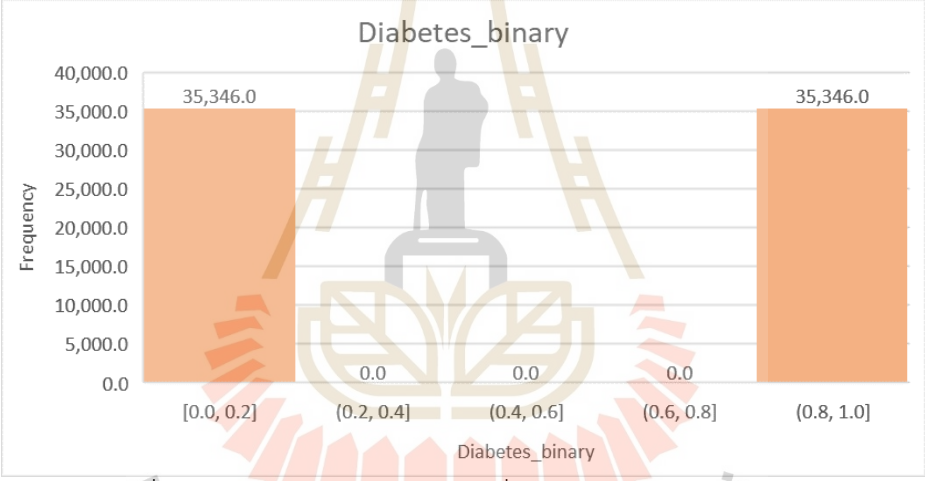
รูปที่ 3.19 การแจกแจงของคุณลักษณะระดับอายุ



รูปที่ 3.20 การแจกแจงของคุณลักษณะระดับการศึกษา



รูปที่ 3.21 การแจกแจงของคุณลักษณะระดับเงินเดือน



รูปที่ 3.22 แสดงการแจกแจงผู้ป่วยที่เป็นโรคเบาหวานและไม่เป็น

3.3 ขั้นตอนการดำเนินงานวิจัย

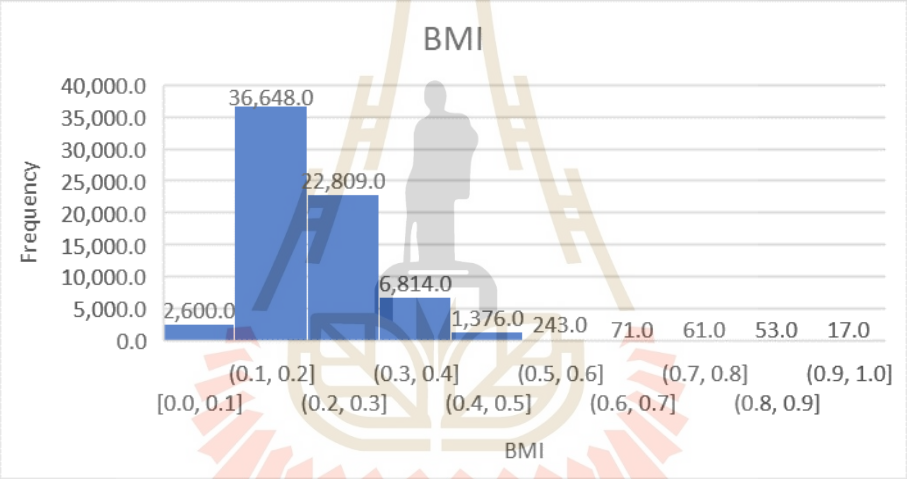
สำหรับขั้นตอนวิธีการดำเนินงานวิจัยแบ่งออกเป็น 4 ขั้นตอนหลัก ๆ คือ

- เตรียมข้อมูล (Data preparation)
- วิศวกรรมคุณลักษณะ (Feature Engineering)
- การสร้างแบบจำลอง (Modeling) และ
- การประเมินประสิทธิภาพ (Evaluation)

โดยมีรายละเอียดขั้นตอนดังต่อไปนี้

3.3.1 เตรียมข้อมูล (Data Preparation)

ในขั้นตอนการเตรียมข้อมูลจะเริ่มจากการตรวจสอบว่ามีข้อมูลที่ขาดหายไปหรือไม่ ซึ่งข้อมูลที่นำมาใช้นั้นไม่มีข้อมูลที่สูญหาย และจากการศึกษาชุดข้อมูลต้นฉบับจะเห็นได้ว่าส่วนใหญ่จะเป็นข้อมูลเชิงคุณภาพ อย่างไรก็ตาม คุณลักษณะค่าดัชนีมวลกายเป็นข้อมูลเชิงปริมาณ ฉนั้นเพื่อให้การคำนวณในสร้างแบบจำลองของการเรียนรู้ของเครื่องเป็นได้อย่างรวดเร็วและตีความง่าย จึงทำการแปลงข้อมูลค่าดัชนีมวลกาย ให้อยู่ในช่วง 0 ถึง 1 ซึ่งวิธีนี้เรียกว่า Min-Max Normalization การทำเช่นนี้จะทำให้ช่วงของข้อมูลมีขนาดเล็กลง แต่ก็ยังคงมีลักษณะการกระจายตัวของข้อมูลที่เหมือนเดิมโดยลักษณะการกระจายตัวดังกล่าวได้แสดงไว้ดังรูปที่ 3.23



รูปที่ 3.23 การแจกแจงของคุณลักษณะค่าดัชนีมวลกาย (Min-Max Normalization)

3.3.2 วิศวกรรมคุณลักษณะ (Feature Engineering)

สำหรับการทำงานของเทคนิควิศวกรรมคุณลักษณะนั้นจำเป็นต้องมีแบบจำลองตั้งต้นเพื่อประเมินประเมินว่าคุณลักษณะแบบใดจึงจะมีความเหมาะสมที่สุด โดยค่าวัตถุประสงค์ในการหาคุณลักษณะที่ดีที่สุดคือค่าความถูกต้อง โดยในการวิจัยครั้งนี้ได้ใช้ 5 แบบจำลองในการประเมินคุณลักษณะ คือ 1) เทคนิคป่าสุ่ม 2) เทคนิคต้นไม้ตัดสินใจ 3) เทคนิคเกรเดียนต์บูตทรี 4) เทคนิคเพื่อนบ้านใกล้ที่สุด และ 5) ซัพพอร์ทเวกเตอร์แมชชีน ซึ่งค่าพารามิเตอร์ตั้งต้นของแต่ละแบบจำลองแสดงดังตารางที่ 3.2

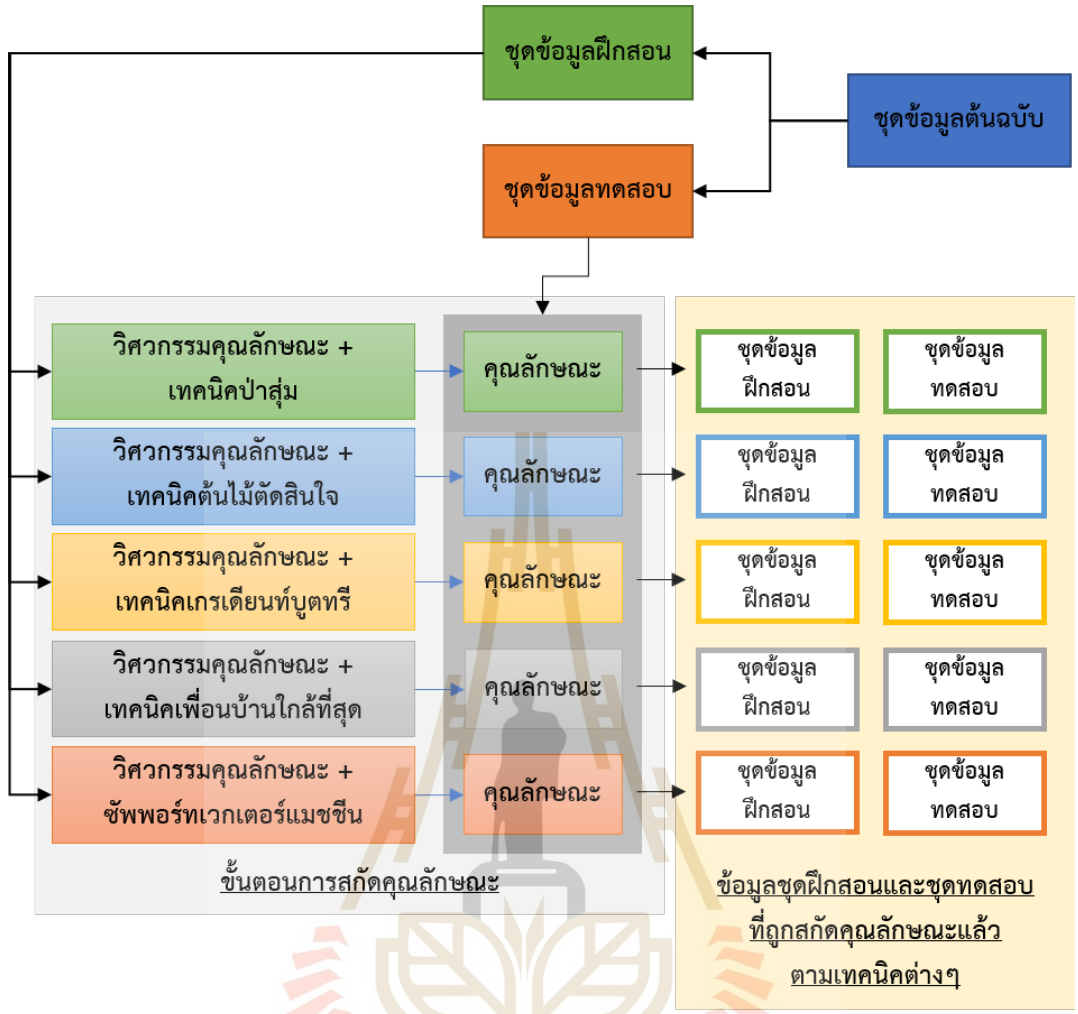
ขั้นตอนในการหาค่าคุณลักษณะที่สุด มีดังต่อไปนี้

1. เริ่มจากแบ่งชุดข้อมูลออกเป็น 2 ส่วน โดยส่วนที่หนึ่ง คือ ชุดฝึกสอนและส่วนที่สอง คือ ชุดทดสอบ ซึ่งจะแบ่งในอัตราส่วน 75/25 (ชุดฝึกสอน/ชุดทดสอบ)
2. นำชุดข้อมูลฝึกสอนไปสกัดคุณลักษณะ โดยวิศวกรรมคุณลักษณะร่วมกับ 5 เทคนิค ดังต่อไปนี้ 1) เทคนิคป่าสุ่ม 2) เทคนิคต้นไม้ตัดสินใจ 3) เทคนิคเกรเดียนต์บูตทรี 4) เทคนิคเพื่อนบ้านใกล้ที่สุด และ 5) ซัพพอร์ทเวกเตอร์แมชชีน
3. หลังจากกระบวนการสกัดคุณลักษณะเสร็จสิ้น ให้ทำการนำข้อมูลชุดฝึกสอนและชุดทดสอบมาเลือกคุณลักษณะที่ได้ ตามแต่ละแบบจำลอง
4. จะได้ชุดข้อมูลฝึกสอนและชุดทดสอบทั้งหมดอย่างละ 5 ชุด โดยข้อมูลชุดใหม่นี้ได้เกิดจากคุณลักษณะตามแต่ละแบบจำลองที่ได้ทำงานร่วมกับเทคนิควิศวกรรมคุณลักษณะ

ขั้นตอนที่กล่าวมาข้างต้นสามารสรุปได้ดังรูปที่ 3.24 ต่อไปจะเป็นการนำชุดข้อมูลที่ถูกสกัดจากวิธีข้างต้น นำมาสร้างแบบจำลองเพื่อวัดประสิทธิภาพและเทียบผลลัพธ์ต่อไป

ตารางที่ 3.2 แสดงค่าพารามิเตอร์ตั้งต้นสำหรับวิศวกรรมคุณลักษณะ

ค่าพารามิเตอร์ตั้งต้นสำหรับวิศวกรรมคุณลักษณะ		
Models	**Parameters**	
Random Forest	criterion	Information_gain
	number of trees	100
	maximal depth	10
Decision Tree	criterion	Information_gain
	maximal depth	10
	minimal size for split	4
Gradient Boost Tree	number of trees	50
	maximal depth	4
	learning rate	0.01
K-nearest neighbors	K	5
Support vector machine	Solver	L2 SVM Dual
	Kernel	Linear
	C	1



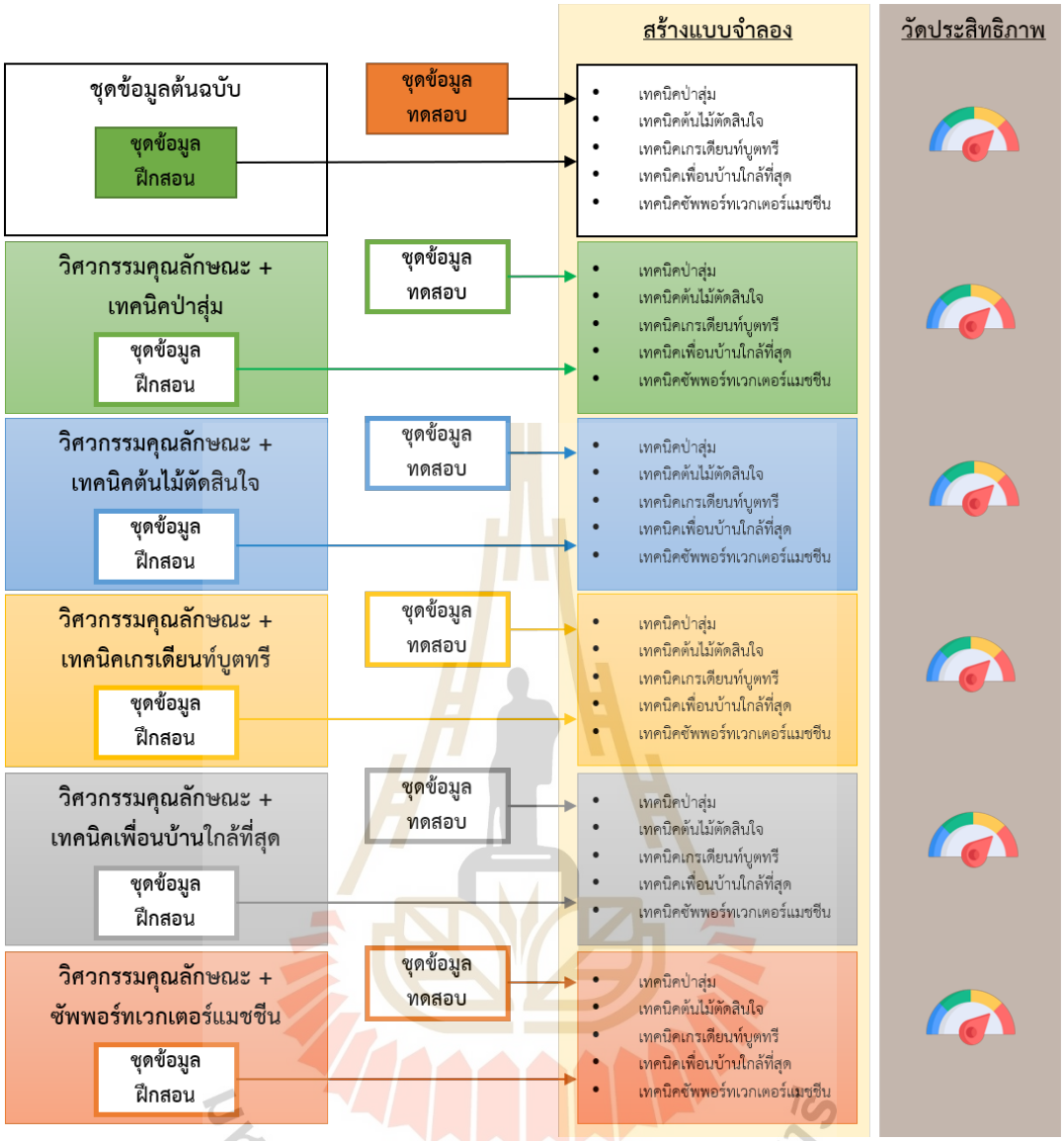
รูปที่ 3.24 แสดงขั้นตอนการสกัดคุณลักษณะร่วมกับเทคนิคต่าง ๆ

3.3.3 การสร้างแบบจำลอง (Modeling)

ในการสร้างแบบจำลองเราจะนำข้อมูลที่ถูกสกัดคุณลักษณะตามวิธีต่างๆที่ได้อธิบายไว้ในหัวข้อ 3.3.2 และชุดข้อมูลต้นฉบับ นำมาสร้างแบบจำลองชุดข้อมูลละ 5 แบบจำลอง ซึ่งแสดงว่าจะต้องสร้างแบบจำลองทั้งหมด 30 แบบจำลอง เพื่อวัดประสิทธิภาพการทำงาน โดยขั้นตอนการสร้างแบบจำลองได้แสดงดังรูปที่ 3.25

3.4 การประเมินประสิทธิภาพ (Evaluation)

ในการวิจัยครั้งนี้ เราจะประเมินประสิทธิภาพของแบบจำลองโดยใช้ตัวชี้วัดประสิทธิภาพดังแสดงไว้ในตารางที่ 3.3



รูปที่ 3.25 แสดงขั้นตอนการสร้างแบบจำลอง

ตารางที่ 3.3 การประเมินประสิทธิภาพโดยใช้ตัวชี้วัดประสิทธิภาพ 7 ชนิด

ลำดับ	ตัวชี้วัดประสิทธิภาพ
1	ความแม่น (Accuracy)
2	ความเที่ยง (Precision)
3	ความเฉพาะเจาะจง (Specificity)
4	ความไว (Sensitivity)
5	F1 Score
6	พื้นที่ใต้เส้นโค้ง ROC (ROC AUC Score)
7	ค่าสัมประสิทธิ์ตัวชี้วัดทางสถิติ (Kappa)

บทที่ 4
ผลการวิจัยและอภิปรายผล

ในบทนี้จะนำเสนอผลการทดลอง การวิเคราะห์และการอภิปรายผลแบบจำลองการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner studio โดยเปรียบเทียบผลลัพธ์ระหว่างการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม, เทคนิคต้นไม้ตัดสินใจ, เทคนิคเกรเดียนท์บูตทรี, เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ทเวกเตอร์แมชชีน โดยแสดงค่าความแม่น (Accuracy), ความเที่ยง (Precision), ความเฉพาะเจาะจง (Specificity), ความไว (Sensitivity), ค่า F1 Score, พื้นที่ใต้เส้นโค้ง ROC และค่าสัมประสิทธิ์ตัวชี้วัดทางสถิติ (Kappa) เพื่อจำแนกประเภทของผู้เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน

4.1 ผลการสกัดคุณลักษณะด้วยวิศวกรรมคุณลักษณะ

จากการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม, เทคนิคต้นไม้ตัดสินใจ, เทคนิคเกรเดียนท์บูตทรี, เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ทเวกเตอร์แมชชีน นั้นได้จำนวนคุณลักษณะที่สกัดได้คือ 7 คุณลักษณะ, 6 คุณลักษณะ, 7 คุณลักษณะ, 8 คุณลักษณะ และ 17 คุณลักษณะ ตามลำดับ ผลการสกัดคุณลักษณะที่สำคัญในการจำแนกประเภทของผู้ที่เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน มีรายละเอียดตามตารางที่ 4.1 จากการสังเกตพบว่าจะมีคุณลักษณะที่สำคัญที่จะปรากฏในทุก ๆ เทคนิคนั้น คือ ค่าดัชนีมวลกาย, ระดับช่วงอายุ, ภาวะคลอเรสเตอรอล และระดับสุขภาพโดยทั่วไป ซึ่งแสดงว่าคุณลักษณะเหล่ามีอิทธิพลในการสร้างแบบจำลองเป็นอย่างมาก และเป็นคุณลักษณะที่เป็นพื้นฐานของทุกเทคนิคอีกด้วย

4.2 ผลการทดสอบประสิทธิภาพแบบจำลองด้วยการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคต่าง ๆ

จากการสกัดคุณลักษณะด้วยจำนวนข้อมูลทั้งหมด 70,692 รายการ มีคุณลักษณะ 21 คุณลักษณะ และมีตัวแปรตาม คือผู้ป่วยที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวาน ผลการทดสอบประสิทธิภาพแบบจำลองด้วยการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม, เทคนิคต้นไม้ตัดสินใจ, เทคนิคเกรเดียนท์บูตทรี, เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ทเวกเตอร์แมชชีน ด้วยการประเมินแบบ 5-การทดสอบแบบไขว้ ซึ่งประเมินประสิทธิภาพของแบบจำลองในการจำแนกประเภทของผู้เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน มีรายละเอียดตามตารางที่ 4.2 – 4.6

ตารางที่ 4.1 ผลการสกัดคุณลักษณะด้วยวิศวกรรมคุณลักษณะร่วมกับเทคนิคต่างๆ

ชุดข้อมูลต้นฉบับ	**วิศวกรรมคุณลักษณะ ร่วมกับเทคนิคป่าสุ่ม**	**วิศวกรรมคุณลักษณะ ร่วมกับเทคนิคต้นไม้ ตัดสินใจ**	**วิศวกรรมคุณลักษณะ ร่วมกับเทคนิคเกรเดียนท์ บูตทรี**	**วิศวกรรมคุณลักษณะ ร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะ ร่วมกับเทคนิค ซัพพอร์ทเวกเตอร์แมชชีน**
1.ภาวะความดันโลหิต	1.ภาวะความดันโลหิต	1.ภาวะความดันโลหิต	1.ภาวะความดันโลหิต	1.ภาวะคลอเรสเตอรอล	1.ภาวะความดันโลหิต
2.ภาวะคลอเรสเตอรอล	2.ภาวะคลอเรสเตอรอล	2.ภาวะคลอเรสเตอรอล	2.ภาวะคลอเรสเตอรอล	2.ค่าดัชนีมวลกาย	2.ภาวะคลอเรสเตอรอล
3.การตรวจคลอเรสเตอรอล	3.ค่าดัชนีมวลกาย	3.ค่าดัชนีมวลกาย	3.ค่าดัชนีมวลกาย	3.การรับประทานผลไม้	3.ค่าดัชนีมวลกาย
4.ค่าดัชนีมวลกาย	4.ระดับสุขภาพโดยทั่วไป	4.ระดับสุขภาพโดยทั่วไป	4.ระดับสุขภาพโดยทั่วไป	4.การรับประทานผัก	4.สูบบุหรี่
5.สูบบุหรี่	5.เพศ	5.ภาวะการเดินขึ้นบันได	5.ภาวะการเดินขึ้นบันได	5.ระดับสุขภาพโดยทั่วไป	5.ภาวะโรคหลอดเลือดหัวใจ
6.ภาวะโรคหลอดเลือดสมอง	6.ระดับช่วงอายุ	6.ระดับช่วงอายุ	6.ระดับช่วงอายุ	6.ภาวะสุขภาพทากายภาพ	6.การออกกำลังกาย
7.ภาวะโรคหลอดเลือดหัวใจ	7.ระดับเงินเดือน		7.[การตรวจคลอเรสเตอรอล] × [ค่าดัชนีมวลกาย]	7.ระดับช่วงอายุ	7.การรับประทานผัก
8.การออกกำลังกาย				8.ระดับเงินเดือน	8.การมีประกันสุขภาพ
9.การรับประทานผลไม้					9.ไม่ไปพบแพทย์แพราะค่าใช้จ่าย
10.การรับประทานผัก					10.ระดับสุขภาพโดยทั่วไป
11.การดื่มแอลกอฮอล์					11.ภาวะสุขภาพจิต
12.การมีประกันสุขภาพ					12.ภาวะสุขภาพทางกายภาพ
13.ไม่ไปพบแพทย์แพราะค่าใช้จ่าย					13.ภาวะการเดินขึ้นบันได
14.ระดับสุขภาพโดยทั่วไป					14.เพศ
15.ภาวะสุขภาพจิต					15.ระดับช่วงอายุ
16.ภาวะสุขภาพทางกายภาพ					16.ระดับการศึกษา
17.ภาวะการเดินขึ้นบันได					17.ระดับเงินเดือน
18.เพศ					
19.ระดับช่วงอายุ					
20.ระดับการศึกษา					
21.ระดับเงินเดือน					

ตารางที่ 4.2 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคป่าสุ่ม

	คุณลักษณะ					
ตัววัดประสิทธิภาพ	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนท์บูตทรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ทเวกเตอร์แมชชีน**
Accuracy	74.07***	73.68**	73.68**	73.64*	72.18	73.61
Precision	71.76***	71.38	71.35	71.50*	68.92	71.55**
Specificity	68.76**	68.31	68.21	68.66	63.57*	68.84***
Sensitivity	79.37**	79.05	79.15*	78.61	80.76***	78.37
F1 Score	75.37***	75.02*	75.05**	74.89	74.39	74.81
ROC AUC	0.816***	0.815**	0.811	0.814*	0.796	0.811
Kappa	0.481***	0.474**	0.474**	0.473	0.444	0.472

ตารางที่ 4.3 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ

	คุณลักษณะ					
ตัววัดประสิทธิภาพ	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนท์บูตทรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
Accuracy	73.37***	73.10**	72.80	72.93*	72.67	72.10
Precision	71.61**	71.64***	70.49	70.76	70.17	71.33*
Specificity	69.31*	69.71**	67.17	67.71	66.46	70.29***
Sensitivity	77.43	76.49	78.43**	78.15*	78.89***	73.90
F1 Score	74.41**	73.99	74.25*	74.27**	74.27**	72.59
ROC AUC	0.806***	0.800**	0.796	0.797*	0.789	0.792
Kappa	0.467***	0.462**	0.456	0.459*	0.453	0.442

ตารางที่ 4.4 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคเกรเดียนท์บูตทรี

	คุณลักษณะ					
ตัววัดประสิทธิภาพ	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนท์บูตทรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
Accuracy	72.77**	73.35***	72.31*	72.35	69.50	71.14
Precision	72.34***	68.33*	68.19	68.00	68.51	69.56**
Specificity	71.81***	59.68	60.97	60.28	66.82*	67.09**
Sensitivity	73.73	87.02***	83.66*	84.43**	72.18	75.19
F1 Score	73.03	76.55***	75.13*	75.33**	70.30	72.26
ROC AUC	0.801**	0.815***	0.797*	0.797*	0.770	0.785
Kappa	0.467***	0.462**	0.456	0.459*	0.453	0.442

ตารางที่ 4.5 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด

	คุณลักษณะ					
ตัววัดประสิทธิภาพ	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนท์บูตทรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
Accuracy	73.05	73.59***	72.79	73.48**	72.31	73.36*
Precision	70.84	71.54*	73.08***	71.98**	70.25	71.14
Specificity	67.75	68.83*	73.42***	70.07**	67.24	68.11
Sensitivity	78.35**	78.35**	72.16	76.89*	77.38	78.61***
F1 Score	74.40*	74.79***	72.62	74.35	73.64	74.69**
ROC AUC	0.804	0.810**	0.803	0.813***	0.795	0.809*
Kappa	0.461	0.472***	0.456	0.470**	0.446	0.467*

ตารางที่ 4.6 ผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีน

	คุณลักษณะ					
ตัววัดประสิทธิภาพ	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนท์บูตทรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
Accuracy	68.52	73.45***	72.45**	71.80*	69.84	71.46
Precision	61.94	68.53*	76.35**	65.48	76.58***	65.15
Specificity	40.93	60.17*	79.84**	51.39	82.51***	50.61
Sensitivity	96.11***	86.72	65.06	92.20*	57.18	92.32**
F1 Score	75.33	76.56**	70.26	76.58***	65.47	76.39*
ROC AUC	0.818***	0.817**	0.814	0.816*	0.803	0.816
Kappa	0.370	0.469***	0.449**	0.436*	0.397	0.429

(หมายเหตุ : ตารางที่ 4.2 - 4.6 สัญลักษณ์ *** ** และ * หมายถึง คะแนนมาเป็นอันดับที่ หนึ่ง สอง และสาม ตามลำดับ)

4.3 วิจารณ์และอภิปรายผลการวิจัย

วัตถุประสงค์สำหรับงานวิจัยเพื่อออกแบบและพัฒนาแบบจำลองของผู้ป่วยโรคเบาหวาน โดยการจำแนกประเภทของผู้เสี่ยงต่อโรคเบาหวาน และผู้ป่วยโรคเบาหวาน ซึ่งอภิปรายผลการวิจัยได้ดังต่อไปนี้

1. จากผลการทดสอบประสิทธิภาพแบบจำลองด้วยเทคนิคป่าสุ่ม พบว่า การใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม และการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคต้นไม้ตัดสินใจ สามารถให้ผลลัพธ์ของประสิทธิภาพที่ค่อนข้างใกล้เคียงกับต้นฉบับมาก แม้จะถูกลดจำนวนคุณลักษณะจาก 21 คุณลักษณะ เหลือเพียง 7 คุณลักษณะและ 6 คุณลักษณะ ตามลำดับ อย่างไรก็ตามในการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิค 4 เทคนิคที่เหลือก็ให้ค่าที่ใกล้เคียงกับข้อมูลต้นฉบับเช่นเดียวกัน และมีบางเทคนิคที่ให้ประสิทธิภาพของบางตัววัดที่โดดเด่นกว่าต้นฉบับอีกด้วย เช่น ค่า Sensitivity ของการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุด เท่ากับ 80.76% และ ค่า Specificity ของการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ทเวกเตอร์แมชชีน เท่ากับ 68.84%
2. จากผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ พบว่า การใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม สามารถให้ผลลัพธ์ของประสิทธิภาพที่ค่อนข้างใกล้เคียงกับต้นฉบับ อย่างไรก็ตามมีบางเทคนิคที่ให้ประสิทธิภาพของบางตัววัดที่โดดเด่นกว่าต้นฉบับอีกด้วย เช่น ค่า Sensitivity ของการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิค

เพื่อนบ้านใกล้ที่สุด เท่ากับ 78.89% และ ค่า Specificity ของการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ทเวกเตอร์แมชชีน เท่ากับ 70.29%

3. จากผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคเกรเดียนท์บูตทรี พบว่า การใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม สามารถให้ผลลัพธ์ของประสิทธิภาพที่ดีกว่าชุดข้อมูลต้นฉบับในหลายๆตัววัด และมีค่าของตัววัดประสิทธิภาพที่โดดเด่นหลายค่า นั้นคือ ค่า Sensitivity เท่ากับ 87.02% ค่า Accuracy เท่ากับ 73.35% ค่า F1 score เท่ากับ 76.55% และ ค่า ROC AUC เท่ากับ 0.815
4. จากผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด พบว่า การใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่มและเทคนิคเกรเดียนท์บูตทรี ให้ผลประสิทธิภาพโดยภาพรวมนั้นดีกว่าข้อมูลต้นฉบับอย่างชัดเจน โดยจำนวนคุณลักษณะที่ลดลงอย่างมาก คือ 7 คุณลักษณะเท่ากัน ซึ่งการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่มให้ผลการวัดประสิทธิภาพที่โดดเด่นกับตัววัดประสิทธิเหล่านี้ คือ ค่า Accuracy เท่า 73.59% ค่า F1 score เท่ากับ 74.79% และ ค่า kappa เท่ากับ 0.472 อีกวิธีคือการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคเกรเดียนท์บูตทรี ให้ผลการวัดประสิทธิภาพที่โดดเด่นกับตัววัดประสิทธิเหล่านี้ คือ ค่า ROC AUC เท่ากับ 0.813
5. จากผลการทดสอบประสิทธิภาพของแบบจำลองด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีน พบว่า การใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม ให้ภาพรวมของตัววัดประสิทธิภาพที่ดีกว่าข้อมูลต้นฉบับ โดยมีตัววัดประสิทธิภาพที่โดดเด่น คือ ค่า Accuracy เท่ากับ 73.45% และ ค่า kappa เท่ากับ 0.469 อย่างไรก็ตามมีบางเทคนิคที่ให้ประสิทธิภาพของบางตัววัดที่โดดเด่นที่สุดเมื่อเทียบกับเทคนิคอื่นๆอีกด้วย เช่น วิศวกรรมคุณลักษณะร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพ ค่า Precision เท่ากับ 76.58% ค่า Specificity เท่ากับ 82.51% และ วิศวกรรมคุณลักษณะร่วมกับเทคนิคเกรเดียนท์บูตทรี ให้ประสิทธิภาพ ค่า F1 score ดีที่สุด เท่ากับ 76.58%

จากผลการทดสอบตามแบบจำลองทั้ง 5 แบบ พบว่าโดยภาพรวมแล้วการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่มที่มีคุณลักษณะ 7 คุณลักษณะ คือ 1) ภาวะความดันโลหิต 2) ระดับสุขภาพโดยทั่วไป 3) ค่าดัชนีมวลกาย 4) ระดับช่วงอายุ 5) ระดับเงินเดือน 6) ภาวะคลอเรสเตอรอล และ 7) เพศ ให้ผลโดยภาพรวมค่อนข้างดีกว่าข้อมูลต้นฉบับและข้อมูลที่ใช้เทคนิคอื่น ๆ และด้วยคุณลักษณะที่น้อยลงทำให้เวลาในการสร้างแบบจำลองน้อยลงตามไปด้วย อย่างไรก็ตามถ้าสนใจตัววัดประสิทธิภาพตัวใดตัวหนึ่งเป็นพิเศษอาจจะมีการเลือกการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคอื่น ๆ ที่ต่างออกไป

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จุดมุ่งหมายของการศึกษานี้คือ เพื่อศึกษาปัจจัยเสี่ยงที่มีผลกระทบต่อการทำให้เกิดโรคเบาหวาน โดยใช้วิธีวิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม, เทคนิคต้นไม้ตัดสินใจ, เทคนิคเกรเดียนท์บูตทรี, เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ทเวกเตอร์แมชชีน แล้วนำไปพัฒนาแบบจำลองของผู้ป่วยโรคเบาหวาน และเปรียบเทียบประสิทธิภาพ และความแม่นยำของแบบจำลองที่ได้พัฒนาขึ้น

จากผลการวิจัยที่ได้ในบทที่ 4 พบว่าการสกัดคุณลักษณะโดยใช้วิธีวิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม, เทคนิคต้นไม้ตัดสินใจ, เทคนิคเกรเดียนท์บูตทรี, เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ทเวกเตอร์แมชชีน ได้จำนวนคุณลักษณะที่สกัดได้คือ 7 คุณลักษณะ, 6 คุณลักษณะ, 7 คุณลักษณะ, 8 คุณลักษณะ และ 17 คุณลักษณะ ตามลำดับ และสังเกตพบว่าจะมีคุณลักษณะที่สำคัญที่จะปรากฏในทุก ๆ เทคนิคนั้น คือ ค่าดัชนีมวลกาย, อายุ, คลอเรสเตอรอลสูง และสุขภาพโดยทั่วไป และเมื่อพิจารณาจากภาพรวมแล้วการใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่มที่มีการสกัดคุณลักษณะให้เหลือเพียง 7 คุณลักษณะ คือ 1) ความดันโลหิตสูง 2) สุขภาพโดยทั่วไป 3) ค่าดัชนีมวลกาย 4) อายุ 5) เงินเดือน 6) คลอเรสเตอรอลสูง และ 7) เพศ ให้ผลโดยภาพรวมค่อนข้างดีกว่าข้อมูลต้นฉบับและข้อมูลที่ใช้กับเทคนิคอื่น ๆ ดังนั้นสามารถสรุปผลการวิจัยได้ว่าการนำวิศวกรรมคุณลักษณะเข้ามาทำงานร่วมกับวิธีป่าสุ่ม ทำให้สามารถคัดกรองคุณลักษณะที่สำคัญที่ส่งผลกระทบต่อการเกิดโรคเบาหวานได้ โดยมีความซับซ้อนของการสร้างแบบจำลองลดลงเพราะคุณลักษณะลดลงจาก 21 เหลือเพียง 7 คุณลักษณะ โดยผลลัพธ์ที่ได้สามารถนำไปประยุกต์ใช้สำหรับการประเมินผู้ที่มีแนวโน้มเป็นโรคเบาหวาน รวมไปถึงช่วยแพทย์ในการการตัดสินใจสำหรับการเลือกวิธีการรักษาให้เหมาะสมกับลักษณะของผู้ป่วยได้อย่างรวดเร็วขึ้น นอกจากนี้ผู้ป่วยก็จะไม่เสียโอกาสได้เข้ารับการรักษาตั้งแต่ช่วงระยะแรกที่เริ่มมีแนวโน้มเสี่ยงต่อการเกิดโรคด้วย

5.2 ข้อเสนอแนะ

1. ในการศึกษาครั้งนี้มีการศึกษาเพื่อทำนายผู้ป่วยที่เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวานเพียงเท่านั้น ไม่สามารถจำแนกชนิดของโรคเบาหวานได้ ดังนั้นในการพัฒนาต่อไปใน

การจำแนกชนิดของโรคเบาหวานควรจะมีชุดข้อมูลที่มีความจำเพาะเจาะจงในแต่ละชนิดของโรคเบาหวาน เพื่อนำมาสร้างแบบจำลองที่มีประสิทธิภาพในการทำนายต่อไป

2. ในการพัฒนาต่อไปสามารถเพิ่มเทคนิคเกี่ยวกับการเรียนรู้เชิงลึก (Deep Learning) และเทคนิคโครงข่ายประสาทเทียม (Artificial neural networks: ANN) ร่วมกับวิศวกรรมคุณลักษณะเพื่อสกัดคุณลักษณะที่สำคัญจากข้อมูล

3. สามารถสร้างแบบจำลองโดยการใช้เทคนิคอื่น ๆ เช่น การเรียนรู้เชิงลึก โครงข่ายประสาทเทียม การถดถอยโลจิสติค (Logistic Regression) และ Extreme Gradient Boosting (XGBoost) เป็นต้น

4. ในการนำไปใช้ทำนายผู้ป่วยโรคเบาหวานในประเทศไทย อาจจะต้องมีการเก็บข้อมูลหรือใช้ข้อมูลแบบเฉพาะเจาะจงสำหรับคนไทย เพื่อนำมาสร้างแบบจำลองที่มีประสิทธิภาพต่อไป อย่างไรก็ตามในการเก็บข้อมูลเกี่ยวกับผู้ป่วยโรคเบาหวานอาจพบเจอกับปัญหาข้อมูลไม่สมดุล (Imbalance Data) นั่นคืออาจทำให้แบบจำลองเกิดการทำนายที่ผิดพลาดได้ ดังนั้นควรใช้เทคนิคต่าง ๆ ในการแก้ปัญหาข้อมูลไม่สมดุลก่อนการสร้างแบบจำลองต่อไป เช่น การสุ่มลดข้อมูล การสุ่มเพิ่มข้อมูล และ Synthetic Minority Oversampling Technique (SMOTE)

รายการอ้างอิง

รายการอ้างอิง

กฤษณา คำลอยฟ้า. (2554). พฤติกรรมการดูแลสุขภาพตนเองของผู้ป่วยโรคเบาหวานในคลินิกโรคเบาหวาน โรงพยาบาลแก้งสนามนาง อำเภอแก้งสนามนาง จังหวัดนครราชสีมา. *Journal of Health and Nursing Education, 17*(1), 17-30.

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI : Machine Learning ด้วย Python* กรุงเทพฯ: หสม มีเดีย เนทเวิร์ค.

ชิตพงษ์ กิตตินราดร. (2562). *Machine Learning: บทนำ.* https://guopai.github.io/ml-blog01.html

ประสพชัย พสุนันท์. (2558). การประเมินความเชื่อมั่นระหว่างผู้ประเมินโดยใช้สถิติแคปปา. *วารสารวิชาการศิลปศาสตร์ประยุกต์. 2*-20.

ปริญญา สงวนสัตย์. (2562). *Artificial Intelligence with Machine Learning AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง.* นนทบุรี: บริษัท ไอดีซี พรีเมียร์ จำกัด

พบแพทย์. (2559). *โรคเบาหวาน.* https://www.pobpad.com/

เพ็ญพันธุ์ ภูริปัญโญ. (2564). *โรคเบาหวาน (Diabetes mellitus).* https://www.thonburihospital.com/DM.html

โรงพยาบาลกรุงเทพ. (2564). *โรคเบาหวาน...รู้จักเพื่อป้องกัน รู้ทันเพื่อควบคุม.* https://www.bangkokhospital-chiangmai.com/สาระสุขภาพและกิจกรรม/โรคเบาหวาน

โรงพยาบาลพญาไท. (2563). *โรคเบาหวาน คืออะไร.* https://www.phyathai.com/article_detail/2705/th/โรคเบาหวาน_คืออะไร

โรงพยาบาลบำรุงราษฎร์. (2559). *เบาหวาน เรื่องหวาน ๆ ที่ไม่ควรเสี่ยง.* https://www.bumrungrad.com/th/health-blog/May-2016/diabetes-risk

วนิดา พงษ์สงวน. (2561). การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ ตัดสินใจ. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม, 1*(1), 1-9.

ศรัณย์ชัย ศิลปะศร, ชนิกานต์ นิกูลรัมย์, เจษฎา ตัณฑนุช, บุระ สินธุภากร และเบญจวรรณ โรจนดิษฐ.(2564). *การสร้างแบบจำลองที่ใช้ในการทำนายคะแนนแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังผ่าตัดเปลี่ยนข้อเข่าด้วยวิศวกรรมคุณลักษณะและเทคนิคเกรเดียนท์บูตทรี.* [วิทยานิพนธ์ปริญญามหาบัณฑิต]. มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

สมาคมโปรแกรมเมอร์ไทย. (2561). *อะไรคือ การเรียนรู้ของเครื่อง (Machine Learning) (ฉบับมือใหม่).* https://www.thaiprogrammer.org/2018/12/

Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., . . . Ibrar, S. (2019). *A model for early prediction of diabetes*. Informatics in Medicine Unlocked, 16, 100204.

Alex Teboul. (2021). *Diabetes Health Indicators Dataset*. https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset

Beghriche, T., Djerioui, M., Brik, Y., Attallah, B., & Belhaouari, S. B. (2021). An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. Complexity, 2021. Bhatia, K., Arora, S., & Tomar, R. (2016). *Diagnosis of diabetic retinopathy using machine learning classification algorithm.* Paper presented at the 2016 2nd international conference on next generation computing technologies (NGCT).

Codecademy. (2022). *Article Normalization*. https://www.codecademy.com/article/normalization

Huda, S. A., Ila, I. J., Sarder, S., Shamsujjoha, M., & Ali, M. N. Y. (2019). *An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms.* Paper presented at the 2019 7th International Conference on Smart Computing & Communications (ICSCC).

javaTpoint. (2021). *Feature Engineering for Machine Learning*. https://www.javatpoint.com/feature-engineering-for-machine-learning

javatpoint. (2021). *K-Nearest Neighbor (KNN) Algorithm for Machine Learning*. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

MedThai. (2017). *โรคเบาหวาน (Diabetes) อาการ, สาเหตุ, การรักษา, วิธีป้องกัน ฯลฯ*. https://medthai.com/

PTT ExpresSo. (2562). *บทบาทของ Machine Learning ในการส่งเสริมนวัตกรรมพลังงาน*. https://blog.pttexpresso.com/machine-learning/

Raymond Cheng. (2020). *Performance Metrics for Classification Machine Learning Problems*. https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science, 132*, 1578-1585.

Sky B.T. Williams. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Statistics How To. (2022). *Cohen's Kappa Statistic*. https://www.statisticshowto.com/cohens-kappa-statistic/

Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia computer science, 167*, 706-716.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med, 37 (5),* 360-363.

Wei, S., Zhao, X., & Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. *Paper presented at the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT).*

ภาคผนวก ก
พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลอง

ตารางที่ ก.1. พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองเทคนิคป่าสุ่ม

พารามิเตอร์	ชุดข้อมูลต้นฉบับ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม	วิศวกรรมคุณลักษณะร่วมกับเทคนิคต้นไม้ตัดสินใจ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเกรเดียนบูททรี	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุด	วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน
criterion	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio
number of trees	74	74	63	63	14	74
maximal depth	96	96	76	76	96	96

ตารางที่ ก.2. พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองเทคนิคต้นไม้ตัดสินใจ

พารามิเตอร์	ชุดข้อมูลต้นฉบับ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม	วิศวกรรมคุณลักษณะร่วมกับเทคนิคต้นไม้ตัดสินใจ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเกรเดียนบูททรี	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุด	วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน
criterion	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio	Gain Ratio
maximal depth	48	12	73	73	73	73
minimal size for split	93	10	96	96	96	96

ตารางที่ ก.3. พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองเทคนิคเกรเดียนบูททรี

พารามิเตอร์	ชุดข้อมูลต้นฉบับ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคป่าสุ่ม	วิศวกรรมคุณลักษณะร่วมกับเทคนิคต้นไม้ตัดสินใจ	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเกรเดียนบูททรี	วิศวกรรมคุณลักษณะร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุด	วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน
number of trees	841	841	489	489	489	488
maximal depth	93	9	93	93	93	95
learning rate	0.0014	0.1248	0.1245	0.1172	0.1320	0.1292

ตารางที่ ก.4. พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองเทคนิคเพื่อนบ้านใกล้ที่สุด

พารามิเตอร์	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนบูททรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
k	31	38	38	74	77	77

ตารางที่ ก.5. พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองเทคนิคซัพพอร์ตเวกเตอร์แมชชีน

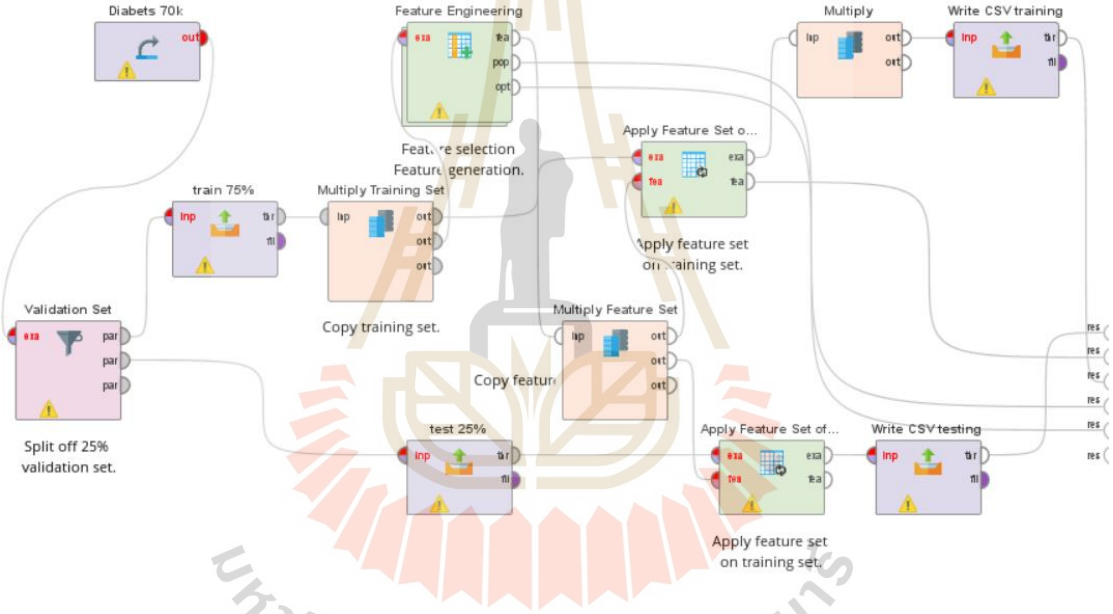
พารามิเตอร์	**ชุดข้อมูลต้นฉบับ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ป่าสุ่ม**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ต้นไม้ตัดสินใจ**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เกรเดียนบูททรี**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค เพื่อนบ้านใกล้ที่สุด**	**วิศวกรรมคุณลักษณะร่วมกับเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน**
Kernel	Linear	Linear	Linear	Linear	Linear	Linear
C	6.1085	6.1256	6.1324	6.1330	6.1624	6.1436

ภาคผนวก ข

กระบวนการใน RapidMiner studio

ขั้นตอนการสกัดคุณลักษณะ ในโปรแกรม RapidMiner studio

ในการใช้เทคนิควิศวกรรมคุณลักษณะบนโปรแกรม RapidMiner studio นั้น จะใช้เครื่องมือที่ชื่อว่า Automatic Feature Engineering ซึ่งมีการรวมสองกระบวนการเข้าด้วยกัน คือ Feature selection และ Feature generation ซึ่งในเครื่องมือนี้จะมีการต่อแบบจำลองเพื่อใช้ในการประเมินประสิทธิภาพว่าคุณลักษณะใดเหมาะสมที่สุด ซึ่งแบบจำลองตั้งต้นนั้นมีทั้งหมด 5 แบบ และได้แสดงรายละเอียดการตั้งค่าพารามิเตอร์ตามที่แสดงไว้ในบทที่ 3 และได้แสดงการเชื่อมต่อเครื่องมือต่างๆดังนี้ด้านล่างนี้

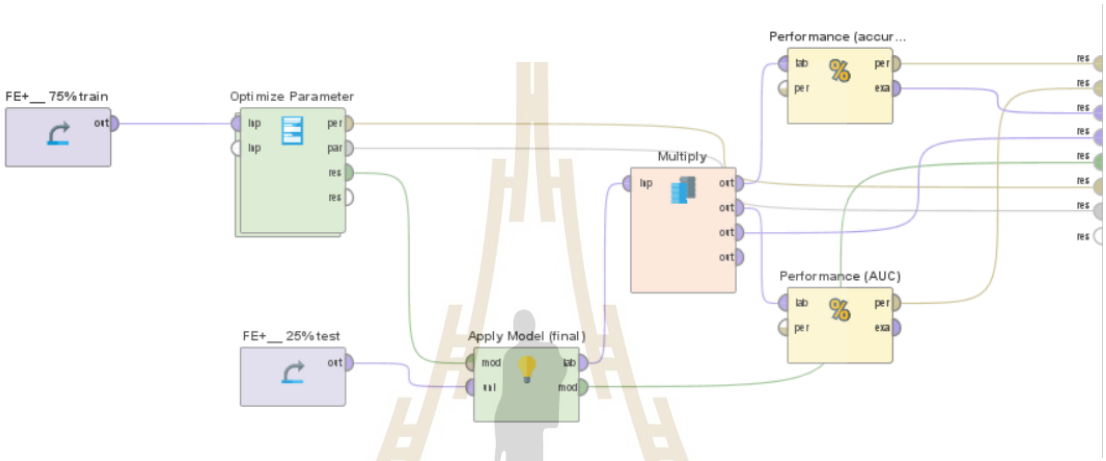


รูปที่ ข.1. ขั้นตอนการสกัดคุณลักษณะ ในโปรแกรม RapidMiner studio

(หมายเหตุ: เครื่องหมายตกใจ(!) ไม่มีผลต่อการรันโปรแกรมเนื่องจาก มีการจัดรูปแบบการวางเครื่องมือให้สวยงามเพียงเท่านั้น ทำให้การเชื่อมต่อจุดใหม่ไม่ได้รับการอัพเดตจึงขึ้นเครื่องหมายเตือน)

ขั้นตอนการสร้างแบบจำลอง ในโปรแกรม Rapidminer Studio

จากขั้นตอนการสกัดคุณลักษณะข้างต้น ผู้วิจัยได้ทำการเก็บข้อมูลการสกัดคุณลักษณะต่างๆ ในรูปแบบไฟล์ .csv ไว้ทั้งหมดตามแต่ละเทคนิคเรียบร้อยแล้ว หลังนี้จะนำข้อมูลที่ได้มาหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละตัวแบบ ซึ่งในเครื่องมือ Optimize parameter นั้นข้างในจะมีการเปลี่ยนแบบจำลองให้ครบทั้ง 5 แบบ ตามแต่วิศวกรรมคุณลักษณะร่วมกับเทคนิคต่างๆ เพื่อวัดประสิทธิภาพ ซึ่งแสดงการทำงานดังรูปด้านล่าง



รูปที่ ข.2. ขั้นตอนการสร้างแบบจำลอง ในโปรแกรม RapidMiner studio

ภาคผนวก ค
บทความทางวิชาการที่ได้รับการนำเสนอในระหว่างการศึกษา

รายชื่อบทความที่ได้รับการนำเสนอในระหว่างการศึกษา

- การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง

- DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING

- https://amm2022.sut.ac.th/

AMM 2022

CERTIFICATE
OF PARTICIPATION

THIS IS TO CERTIFY THAT

Kunaporn Punpian, Jakkrit Polrob and Jessada Tanthanuch

HAS PARTICIPATED WITH A PAPER ENTITLED

การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับ ขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง

in the 26th Annual Meeting in Mathematics and
the 1st International Annual Meeting in Mathematics (AMM 2022)
May 18th - 20th, 2022

J. Tanthanuch
(Asst. Prof. Dr. Jessada Tanthanuch)
General Chair of AMM 2022

(Prof. Dr. Santi Maensiri)
Dean, Institute of Science,
Suranaree University of Technology

ประวัติผู้เขียน

นางสาวคุณาภรณ์ พันธ์เพียร เกิดเมื่อวันที่ 26 เดือนเมษายน พ.ศ. 2541 ณ จังหวัดราชบุรี สำเร็จ การศึกษาระดับมัธยมศึกษาจากโรงเรียนนารีวุฒิ อำเภอบ้านโป่ง จังหวัดราชบุรี ในปีการศึกษา 2559 เข้ารับการศึกษาในระดับอุดมศึกษา ณ มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา จนสำเร็จการศึกษาระดับปริญญาตรีวิศวกรรมศาสตรบัณฑิต (วิศวกรรมอากาศยาน) จากมหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา เมื่อ พ.ศ. 2563

หลังจากนั้นในปีการศึกษา 2563 ได้เข้ารับการศึกษาระดับปริญญาโท หลักสูตรสาขาวิชานวัตกรรม วิศวชีวการแพทย์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี โดยได้รับทุนอุดหนุนการวิจัยจากทางมหาวิทยาลัยเทคโนโลยีสุรนารี ได้แก่ ทุนศักยภาพบัณฑิต นอกจากนี้ยังมีผลงานทางวิชาการระดับนานาชาติที่ได้รับการตีพิมพ์เผยแพร่ ดังปรากฏในภาคผนวก ค