การจำแนกการเป็นโรคเบาหวานโดยใช้เทคนิค Machine learning

โดย

นางสาวเมธาพร ผ่องยิ่ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

สาขาวิชาสถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต

ภาควิชาสถิติ

มหาวิทยาลัยศิลปากร

ปีการศึกษา 2565

การจำแนกการเป็นโรคเบาหวานโดยใช้เทคนิค Machine learning

โดย
นางสาวเมธาพร ผองยิ่ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาสถิติ
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2565

DIABETES CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

By

MISS Methaporn PHONGYING

A Thesis Submitted in Partial Fulfillment of the Requirements

for Master of Science (APPLIED STATISTICS)

Department of STATISTICS

Silpakorn University

Academic Year 2022

หัวข้อ	การจำแนกการเป็นโรคเบาหวานโดยใช้เทคนิค Machine learning
โดย	นางสาวเมธาพร ผ่องยิ่ง
สาขาวิชา	สถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. ศศิประภา หิริโอตป์

คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

.....คณบดีคณะวิทยาศาสตร์
(ผู้ช่วยศาสตราจารย์ ดร. นรงค์ ฉิมพาลี)

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ดร. กรรณิกาณ์ หิรัญกสิ)

.....อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร. ศศิประภา หิริโอตป์)

.....ผู้ทรงคุณวุฒิภายนอก
(ผู้ช่วยศาสตราจารย์ ดร. พรรณนภา ช่างเพ็ชร)

630720073 : สถิติประยุกต์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
คำสำคัญ : การเรียนรู้ด้วยเครื่อง, โรคเบาหวาน, ต้นไม้ตัดสินใจ, ต้นไม้ป่าสุ่ม, ซัพพอร์ตเวกเตอร์แมชชีน, เพื่อนบ้านใกล้ที่สุด

นางสาว เมธาพร ผ่องยิ่ง: การจำแนกการเป็นโรคเบาหวานโดยใช้เทคนิค Machine learning อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. ศศิประภา หิริโอตป์

ปัจจุบันเทคนิค Machine learning ได้เข้ามามีบทบาททางการแพทย์ในการวินิจฉัยโรคมากขึ้น เนื่องจากเราสามารถใช้เทคนิค Machine learning ในการวิเคราะห์ข้อมูลขนาดใหญ่ทางการแพทย์ เพื่อค้นหารูปแบบหรือข้อเท็จจริงบางอย่างที่ยากต่อการอธิบาย ซึ่งมีส่วนช่วยให้การวินิจฉัยโรคทำได้อย่างแม่นยำมากยิ่งขึ้น โดยในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคที่ใช้ในการสร้างแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม 4 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision tree) เทคนิคต้นไม้ป่าสุ่ม (Random Forest) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) โดยมีเกณฑ์ที่ใช้ในการทดสอบประสิทธิภาพของการจำแนก คือ ค่าความถูกต้อง (accuracy) ค่าความเที่ยง (precision) ค่าความครบถ้วน (recall) และค่าคะแนน F1 (F1-score) ที่ให้ค่ามากที่สุด ซึ่งผลของการวิจัยพบว่าแบบจำลองกรณีที่พิจารณาอิทธิพลร่วมมีประสิทธิภาพการจำแนกดีกว่าแบบจำลองกรณีที่ไม่พิจารณาอิทธิพลร่วมทั้ง 4 เทคนิค โดยที่แบบจำลองกรณีที่พิจารณาอิทธิพลร่วม เทคนิค Random forest มีประสิทธิภาพการจำแนกดีที่สุด ซึ่งให้ค่าความถูกต้องในการจำแนก 97.5% มีค่าความแม่นยำที่ 97.4% มีค่าความครบถ้วนที่ 96.6% และค่าคะแนน F1 ที่ 97% ในทางเดียวกัน แบบจำลองกรณีที่ไม่พิจารณาอิทธิพลร่วม เทคนิค Random forest มีประสิทธิภาพการจำแนกดีที่สุด ซึ่งให้ค่าความถูกต้องในการจำแนก 88.2% มีค่าความแม่นยำที่ 92.2% มีค่าความครบถ้วนที่ 89.3% และค่าคะแนน F1 ที่ 90.7% โดยผลการวิจัยที่ได้นี้สามารถนำไปใช้เป็นแนวทางในการพัฒนาโปรแกรมสำหรับการคัดกรองผู้ป่วยโรคเบาหวานได้อย่างมีประสิทธิภาพต่อไปในอนาคต

630720073 : Major (APPLIED STATISTICS)
Keyword : machine learning, diabetes, Decision tree, Random forest, Support Vector Machine, K-Nearest neighbor

MISS Methaporn PHONGYING : DIABETES CLASSIFICATION USING MACHINE LEARNING TECHNIQUES Thesis advisor : Assistant Professor Sasiprapa Hiriote

Nowadays, Machine learning techniques play an increasingly prominent role in medical diagnosis because using these techniques can be analyzed to find patterns or facts that are difficult to explain, which contributes to making the diagnosis more accurate. The purpose of this research is to compare the efficiency of diabetic classification models with and without interaction using four machine learning techniques including Decision tree, Random forest, Support Vector Machine and K-Nearest neighbor. These models are compared base on accuracy, precision, recall, and F1-score. The results of this research showed that the models with interaction have better classification performance than those without interaction for all 4 machine learning techniques. Among models with interaction, Random forest classifiers had the best performance with 97.5% accuracy, 97.4% precision, 96.6% recall, and 97% F1-score. In the same way, Random forest also had the best classification performance among models without interaction with 88.2% accuracy, 92.2% precision, 89.3% recall, and 90.7% F1-score. The findings from this research can be further developed into a program to effectively screen diabetes patients.

กิตติกรรมประกาศ

การดำเนินงานวิจัยและการเรียบเรียงวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้เป็นอย่างดี เนื่องจากได้รับความอนุเคราะห์จากผู้ช่วยศาสตราจารย์ ดร. ศศิประภา หิริโอตป์ ผู้เป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาให้คำแนะนำ คำปรึกษา แนวคิด องค์ความรู้ รวมถึงการตรวจทานและแก้ไขข้อบกพร่องต่าง ๆ ให้ข้าพเจ้าเป็นอย่างดี ผู้วิจัยจึงขอกราบขอบพระคุณอาจารย์เป็นอย่างสูงด้วยความซาบซึ้ง

ขอขอบคุณอาจารย์กรรณิกาณ์ หิรัญกสิ ที่กรุณาเป็นประธานกรรมการในการสอบวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร. พรรณนภา ช่างเพ็ชร ที่กรุณาเป็นผู้ทรงคุณวุฒิ สำหรับการให้คำแนะนำการตรวจสอบความถูกต้อง และชี้แนะแนวทางทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากรทุกท่านที่ได้มอบองค์ความรู้ ให้คำแนะนำ และให้กำลังใจตลอดระยะเวลาในการศึกษา รวมทั้งบุคลากรภาควิชาสถิติ สายสนันสนุน คุณนงลักษณ์ เอี๊ยวเจริญ ที่ให้ความช่วยเหลือและอำนวยความสะดวกด้านงานเอกสารและการดำเนินงานวิจัยแก่ผู้วิจัย

ขอบคุณรุ่นพี่และเพื่อน ๆ ภาควิชาสถิติที่ให้คำแนะนำ ให้ความช่วยเหลือ และเป็นกำลังใจให้ในการทำวิทยานิพนธ์นี้สำเร็จลุล่วงไปด้วยดี

สุดท้ายนี้ ขอขอบพระคุณคุณพ่อ คุณแม่ที่เป็นผู้ปกครองของผู้ทำวิจัยเป็นอย่างสูง รวมถึงขอบคุณครอบครัวและคนสนิททุก ๆ ท่านที่สนันสนุนการศึกษาการทำวิจัยครั้งนี้ ตลอดจนให้ความรัก กำลังใจ แรงผลักดันและแรงสนันสนุนในทุกด้านแก่ผู้ทำวิจัย

นางสาว เมธาพร ผ่องยิ่ง

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1

บทนำ

ความสำคัญของปัญหา

โรคเบาหวาน (Diabetes Mellitus) เป็นโรคที่เกิดจากความผิดปกติในการทำงานของฮอร์โมนที่ชื่อว่า อินซูลิน (Insulin) ซึ่งเป็นฮอร์โมนที่ผลิตขึ้นจากเบต้าเซลล์ในตับอ่อน ทำหน้าที่ช่วยให้กลูโคสจากกระแสเลือดเข้าสู่เซลล์ของร่างกาย โดยที่กลูโคสจะถูกเปลี่ยนให้เป็นพลังงาน ซึ่งปกติแล้วร่างกายของคนเราจำเป็นต้องมีอินซูลิน เพื่อนำน้ำตาลในกระแสเลือดไปเลี้ยงอวัยวะต่าง ๆ ของร่างกาย โดยเฉพาะสมองและกล้ามเนื้อ หากร่างกายเกิดภาวะที่อินซูลินมีความผิดปกติ ไม่ว่าจะเป็นการลดลงของปริมาณอินซูลินในร่างกาย หรือการที่อวัยวะต่าง ๆ ของร่างกายตอบสนองต่ออินซูลินลดลง (หรือเรียกว่า ภาวะดื้ออินซูลิน) จะทำให้ร่างกายไม่สามารถใช้อินซูลินได้อย่างมีประสิทธิภาพ นอกจากนี้อินซูลินยังจำเป็นสำหรับการเผาผลาญโปรตีนและไขมันในร่างกาย ซึ่งการขาดอินซูลินหรือการที่ร่างกายไม่สามารถนำน้ำตาลที่อยู่ในกระแสเลือดไปใช้ได้อย่างเต็มประสิทธิภาพ ส่งผลให้มีปริมาณน้ำตาลคงเหลือในกระแสเลือดมากกว่าปกติ (Hyperglycemia) หากปล่อยทิ้งไว้โดยไม่ได้รับการรักษาอย่างถูกวิธี อาจทำให้เกิดความเสียหายต่ออวัยวะต่าง ๆ ของร่างกาย ซึ่งนำไปสู่ภาวะแทรกซ้อนที่ร้ายแรงต่อสุขภาพตามมาในที่สุด

การระบุชนิดของโรคเบาหวาน อาศัยผลการจากห้องปฏิบัติการเป็นหลัก ซึ่งโรคเบาหวานสามารถแบ่งได้เป็น 4 ชนิดตามสาเหตุของการเกิดโรค ได้แก่ โรคเบาหวานชนิดที่ 1 (Type 1 diabetes mellitus, T1DM) เป็นผลจากการทำลายเบต้าเซลล์ที่ตับอ่อนจากภูมิคุ้มกันของร่างกาย โดยส่วนใหญ่มักพบในกลุ่มคนอายุน้อย รูปร่างไม่อ้วน อาการของโรคเบาหวานชนิดที่ 1 คือ ปัสสาวะมาก กระหายน้ำมาก ดื่มน้ำมาก อ่อนเพลีย น้ำหนักลด อาจจะเกิดขึ้นได้อย่างรวดเร็วและรุนแรง ซึ่งในบางกรณีพบภาวะเลือดเป็นกรดจากสารคีโตน (ketoacidosis) โรคเบาหวานชนิดที่ 2 (Type 2diabetes mellitus, T2DM) เป็นผลมาจากการมีภาวะดื้อต่ออินซูลิน (insulin deficiency) ร่วมกับความบกพร่องในการผลิตอินซูลินที่เหมาะสม (Relative insulin deficiency) ส่วนมากพบในกลุ่มคนอายุ 30 ปีขึ้นไป รูปร่างท้วมหรืออ้วน โรคเบาหวานชนิดที่ 3 คือ โรคเบาหวานขณะตั้งครรภ์ (gestational diabetes mellitus, GDM) เกิดจากการที่ร่างกายมีภาวะดื้อต่ออินซูลินมากขึ้นในระหว่างการตั้งครรภ์ เป็นผลจากปัจจัยที่เกิดจากรกและตับอ่อนของมารดาไม่สามารถผลิตอินซูลินให้เพียงพอกับความต้องการได้ และโรคเบาหวานชนิดที่ 4 คือ โรคเบาหวานที่มีสาเหตุจำเพาะ (Specific types of diabetes due to other causes) เป็นโรคเบาหวาน ที่มีสาเหตุชัดเจน ได้แก่

โรคเบาหวานที่เกิดจากความผิดปกติทางพันธุกรรม เช่น MODY (Maturity-Onset Diabetes of the Young) โรคเบาหวานที่เกิดจากโรคของตับอ่อน โรคเบาหวานที่เกิดจากความผิดปกติของต่อมไร้ท่อ โรคเบาหวานที่เกิดจากยา โรคเบาหวานที่เกิดจากการติดเชื้อ โรคเบาหวานที่เกิดจากปฏิกิริยาภูมิคุ้มกัน หรือโรคเบาหวานที่พบร่วมกับกลุ่มอาการต่าง ๆ ผู้ป่วยจะมีลักษณะจำเพาะของโรคหรือกลุ่มอาการนั้น ๆ หรือมีอาการและอาการแสดงของโรคที่ทำให้เกิดเบาหวาน (สมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี, 2560)

โรงพยาบาลในสังกัดสำนักการแพทย์ กรุงเทพมหานคร มีจำนวนผู้ป่วยที่เข้ารับการรักษาด้วยโรคเบาหวานเป็นจำนวนมากอย่างต่อเนื่อง พบว่าในปี 2564 มีจำนวนผู้ป่วยโรคเบาหวาน 72,958 ราย คิดเป็นร้อยละ 68.07 ของผู้ป่วยทั้งหมด และคาดการณ์ว่าในอนาคต จำนวนผู้ป่วยโรคเบาหวานจะเพิ่มสูงขึ้น ซึ่งจะเห็นว่าจำนวนผู้ป่วยโรคเบาหวานของโรงพยาบาลในสังกัดสำนักการแพทย์ กรุงเทพมหานคร เพิ่มขึ้นอย่างต่อเนื่อง ดังภาพที่ 1

ภาพที่ 1 จำนวนผู้ป่วยโรคเบาหวานของโรงพยาบาลในสังกัดสำนักการแพทย์ ระหว่างปี พ.ศ.2554 - 2564

จากการรายงานการสำรวจสุขภาพประชากรไทยโดยการตรวจร่างกายครั้งที่ 6 ในปี พ.ศ. 2562-2563 พบว่า ร้อยละ 30.6 ของผู้ที่เป็นเบาหวาน ไม่ทราบว่าตนเองป่วยเป็นโรคเบาหวานมาก่อน ส่วนผู้ที่เป็นเบาหวานมีร้อยละ 13.9 ซึ่งไม่ได้รับการรักษา (วิชัย เอกพลากร, หทัยชนก พรรคเจริญ, & วราภรณ์ เสถียรนพเก้า, 2564) ดังนั้นการตรวจคัดกรอง (screening test) จึงมีประโยชน์ในการ

ค้นหาผู้ซึ่งไม่มีอาการ เพื่อการวินิจฉัยและให้การรักษาตั้งแต่ระยะเริ่มแรก เพราะโรคในระยะที่เริ่มเป็นสามารถควบคุมได้ตามเป้าหมายและป้องกันการเกิดโรคแทรกซ้อนได้ง่าย อีกทั้งผู้ที่มีความเสี่ยงที่จะเป็นโรคเบาหวานสามารถป้องกันหรือชะลอการเกิดของโรคแทรกซ้อนต่าง ๆ จากเบาหวานในระยะยาวได้ ซึ่งในปัจจุบันการตรวจคัดกรองเบาหวานในประชากรทั่วไปทุก ๆ คนนั้น มีค่าใช้จ่ายที่ค่อนข้างสูงและอาจไม่คุ้มค่าสำหรับการประเมินความเสี่ยงของการเกิดโรคเบาหวานสำหรับผู้ซึ่งไม่มีอาการ ดังนั้นจึงมีผู้คิดค้นวิธีการประเมินความเสี่ยงการเกิดโรคเบาหวานโดยใช้ข้อมูลจากการศึกษาปัจจัยเสี่ยงหลายอย่างที่สามารถประเมินได้ง่ายด้วยแบบสอบถามและการตรวจร่างกายเบื้องต้นโดยที่ไม่ต้องเจาะเลือดตรวจ ดังตารางที่ 1 แล้วนำข้อมูลมาคำนวณเป็นคะแนน (risk score) เพื่อเปรียบเทียบกับเกณฑ์การแปลผลคะแนนความเสี่ยงที่ได้ต่อการเกิดโรคเบาหวาน และข้อแนะนำเพื่อการปฏิบัติ ดังตารางที่ 2 โดยเมื่อนำคะแนนของแต่ละปัจจัยเสี่ยงมารวมกัน คะแนนจะอยู่ในช่วง 0-17 คะแนน หากคะแนนความเสี่ยงที่ประเมินได้มีค่าตั้งแต่ 6 คะแนนขึ้นไป หมายความว่ามีความเสี่ยงสูงที่จะเกิดโรคเบาหวาน ซึ่งการประเมินความเสี่ยงโดยวิธีนี้ สามารถนำมาใช้เป็นแนวปฏิบัติเพื่อการคัดกรองโรคเบาหวานซึ่งมีข้อจำกัดทางด้านงบประมาณได้ แต่ยังมีข้อจำกัดในหลาย ๆ ด้าน เช่น ผู้ป่วยไม่สะดวกทำแบบประเมินด้วยตนเอง ภาระงานของบุคลากรทางการแพทย์ที่มากจึงไม่ได้มีการประเมินความเสี่ยงให้ผู้รับบริการ และความไม่เพียงพอของเครื่องมือที่จะนำมาใช้ในการทำแบบประเมิน เป็นต้น

ตารางที่ 1 ปัจจัยเสี่ยงของโรคเบาหวาน และคะแนนความเสี่ยง

ปัจจัยเสี่ยง	คะแนนความเสี่ยง
อายุ	
• 34 - 39 ปี	0
• 40 - 44 ปี	0
• 45 - 49 ปี	1
• ตั้งแต่ 50 ปีขึ้นไป	2
เพศ	
• หญิง	0
• ชาย	2
ดัชนีมวลกาย	
• ต่ำกว่า 23 กก./ม.^{2}	0
• ตั้งแต่ 23 ขึ้นไปแต่น้อยกว่า 27.5 กก./ม.^{2}	3
• ตั้งแต่ 27.5 กก./ม.^{2} ขึ้นไป	5
รอบเอว	
• ผู้ชายน้อยกว่า 90 ซม. ผู้หญิงน้อยกว่า 80 ซม.	0
• ผู้ชายตั้งแต่ 90 ซม. ขึ้นไป ผู้หญิงตั้งแต่ 80 ซม. ขึ้นไป	2
ความดันโลหิตสูง	
• ไม่มี	0
• มี	2
ประวัติโรคเบาหวานในญาติสายตรง	
(พ่อ แม่ พี่ หรือน้อง)	0
• ไม่มี	4
• มี	

ที่มา : วารสารแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี

ตารางที่ 2 การแปลผลคะแนนความเสี่ยงของโรคเบาหวาน และข้อแนะนำ

ผลรวมคะแนน	ระดับความเสี่ยง	ข้อแนะนำ
น้อยกว่าหรือเท่ากับ 2 คะแนน	น้อย	• ออกกำลังกายสม่ำเสมอ • ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม • ตรวจความดันโลหิต • ควรประเมินความเสี่ยงซ้ำทุก 3 ปี
3 - 5 คะแนน	ปานกลาง	• ออกกำลังกายสม่ำเสมอ • ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม • ตรวจความดันโลหิต • ควรประเมินความเสี่ยงซ้ำทุก 1-3 ปี
6 - 8 คะแนน	สูง	• ควบคุมอาหารและออกกำลังกายสม่ำเสมอ • ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม • ตรวจความดันโลหิต • ตรวจระดับน้ำตาลในเลือด • ควรประเมินความเสี่ยงซ้ำทุก 1-3 ปี
มากกว่า 8 คะแนน	สูงมาก	• ควบคุมอาหารและออกกำลังกายสม่ำเสมอ • ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม • ตรวจความดันโลหิต • ตรวจระดับน้ำตาลในเลือด • ควรประเมินความเสี่ยงซ้ำทุก 1 ปี

ที่มา : วารสารแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี

Machine learning เป็นสาขาหนึ่งของเทคโนโลยีด้านปัญญาประดิษฐ์ (artificial intelligence) โดยจุดมุ่งหมาย คือ การออกแบบและพัฒนาอัลกอริธึมที่อนุญาตให้คอมพิวเตอร์ปรับปรุงประสิทธิภาพด้วยข้อมูลอย่างอัตโนมัติ ซึ่ง Machine learning เกี่ยวข้องกับการวิเคราะห์ข้อมูลที่ผ่านมาเพื่อค้นหารูปแบบหรือข้อเท็จจริงบางอย่างที่ยากต่อการอธิบาย เช่น ข้อมูลจำนวนมากอาจมีความสัมพันธ์ที่ซ่อนอยู่ โดยที่มนุษย์ไม่สามารถอธิบายได้ เนื่องจากความสามารถในการจัดเก็บและประมวลผลที่จำกัด เป็นต้น (Viviana & Andrei, 2009)

Machine learning ได้เข้ามามีบทบาททางการแพทย์ในการวินิจฉัยโรคมากขึ้น ปัจจุบันการเก็บข้อมูลการรักษาต่าง ๆ สามารถรวบรวมและวิเคราะห์ข้อมูลได้จำนวนมากผ่านเทคนิค Machine learning ซึ่งช่วยให้การวินิจฉัยโรคทำได้อย่างแม่นยำมากยิ่งขึ้น ดังนั้นการนำข้อมูลจำนวนมากเหล่านี้มาใช้ในการวิเคราะห์เพื่อช่วยในการวินิจฉัย จึงเป็นทางเลือกหนึ่งแทนการใช้วิธีการทางการแพทย์ที่ต้องมีการตรวจทางคลินิกหรือการรอผลจากห้องปฏิบัติการหลายขั้นตอน ซึ่งมีระยะเวลาค่อนข้างนานและค่าใช้จ่ายสูง

การประเมินความเสี่ยงการเกิดโรคเบาหวานเบื้องต้น มักใช้การประเมินปัจจัยเสี่ยงต่าง ๆ ผ่านแบบประเมินคะแนนความเสี่ยง ซึ่งอาจไม่สะดวกสำหรับกลุ่มผู้สูงอายุเนื่องจากข้อจำกัดทางด้านเทคโนโลยี อีกทั้งอาจใช้ระยะเวลานานในการทำแบบประเมินด้วยตนเอง และเป็นการเพิ่มภาระงานของบุคลากรทางการแพทย์ในการจัดตั้งจุดคัดกรองประเมินความเสี่ยงดังกล่าว ดังนั้นการจำแนกประเภทด้วยเทคนิค Machine learning ซึ่งสามารถสร้างแบบจำลองในการจำแนกได้อย่างแม่นยำผ่านการเรียนรู้ข้อมูลในอดีตจำนวนมาก จึงเป็นทางเลือกหนึ่งที่ควรนำมาใช้ในการประเมินความเสี่ยงการเกิดโรคเบาหวานเพื่อหลีกเลี่ยงข้อจำกัดทางด้านต่าง ๆ ของการประเมินความเสี่ยงผ่านแบบประเมินคะแนนความเสี่ยงเดิม

ปัจจุบันการวิเคราะห์ข้อมูลด้วยเทคนิค Machine learning ได้รับความสนใจอย่างมากในการทำนายและวินิจฉัยโรค อย่างไรก็ตาม วิธีการเหล่านี้มีความซับซ้อนมากและต้องการข้อมูลจำนวนมากที่ใช้ในการวิเคราะห์ (Viviana & Andrei, 2009) ซึ่งเทคนิคการจำแนกประเภทนั้นมีหลายวิธี จากการศึกษางานวิจัยทางการแพทย์ในการจำแนกประเภทของผู้ป่วยโดยใช้เทคนิค Machine learning พบว่าวิธี K-Nearest neighbor, วิธี Decision tree, วิธี Random forest และวิธี Support Vector Machine เป็นแบบจำลองที่ใช้กันอย่างแพร่หลาย เช่น การวิเคราะห์ประสิทธิภาพของแบบจำลองในการทำนายโรคเบาหวาน โดยใช้เทคนิคเหมืองข้อมูล และทดสอบกับชุดข้อมูลที่ไม่มีข้อมูลที่ผิดพลาดอยู่ พบว่าเทคนิค K-Nearest neighbor กรณีที่ k=1 และเทคนิค Random forest มีประสิทธิภาพในการทำนายสูงสุด และให้ค่าความถูกต้องเท่ากับ 100% (Kandhasam & Balamurali, 2015) การสร้างแบบจำลองในการจำแนกโรคเบาหวานจากข้อมูลของโรงพยาบาลศูนย์สวรรค์ประชารักษ์จำนวน 48,763 ชุด พบว่าแบบจำลองการจำแนกจากเทคนิค Bagging ร่วมกับ

อัลกอริทึม Decision tree มีความแม่นยำในการจำแนกสูงสุด 95.31% (Nai-arun & Sittidech, 2014) การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผู้ป่วยโรคเบาหวาน กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี พบว่าเทคนิคป่าสุ่ม (Random Forest) ให้ค่าความถูกต้องในการทำนายผลการเป็นโรคเบาหวานมากที่สุด 88.03% มีค่าความแม่นยำ 88.22% และค่าวัดประสิทธิภาพโดยรวม 89.28% (ปพนน์ศรณ์ สิ่วสำแดงเดช, 2565) และนอกจากนี้ยังมีการศึกษาการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล และการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล พบว่าเทคนิค Support vector machine มีประสิทธิภาพการทำนายสูงสุด และให้ค่าความถูกต้อง เท่ากับ 76.95% (รุ่งโรจน์ บุญมา & นิเวศ จิระวิชิตชัย, 2562) โดยที่เทคนิคทั้ง 4 วิธีดังกล่าว มีแนวคิดและหลักการที่แตกต่างกัน

K-Nearest neighbor เป็นวิธีการจำแนกประเภท โดยมีแนวคิด คือ การค้นหาข้อมูลที่มีระยะทางที่ใกล้เคียงที่สุดระหว่างข้อมูลที่ประเมินและจำนวน K (เพื่อนบ้าน) ที่ใกล้เคียงที่สุดกับข้อมูลชุดฝึกฝน (training set) (Dimas & Naqshauliza, 2020) จากนั้นรวบรวมข้อมูลที่ใกล้เคียงที่สุด K ตัว แล้วเลือกกลุ่มของข้อมูลที่มีสมาชิกในกลุ่ม K มากที่สุดให้กับสมาชิกใหม่ โดยอัลกอริทึมนี้สามารถเรียนรู้ได้ง่าย สะดวด และรวดเร็ว รวมถึงมีประสิทธิภาพในการวิเคราะห์ข้อมูลจำนวนมาก (Mutrofin, Izzah, Kurniawardhani, & Masrur, 2014)

Decision tree เป็นการนำข้อมูลมาสร้างแบบจำลองการทำนายที่มีลักษณะคล้ายกับต้นไม้ โดยจะมีการสร้างกฎต่าง ๆ ขึ้นเพื่อใช้ในการตัดสินใจ ซึ่ง Decision tree นั้นมีการทำงานแบบ Supervised learning คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ก่อน (training set) ได้อย่างอัตโนมัติ และสามารถทำนายกลุ่มของข้อมูลที่ยังไม่ทราบหมวดหมู่ได้ (Ding, Ding, & Perrizo, 2002; Quadri & Kalyankar, 2021)

Random forest เป็นเทคนิคการสร้างโมเดลด้วยวิธีการ Decision tree ขึ้นมาหลาย ๆ โมเดลอย่างสุ่ม จากนั้นนำผลลัพธ์ที่ได้ของแต่ละโมเดลมารวมกัน แล้วนับจำนวนผลลัพธ์ที่มีจำนวนซ้ำกันมากที่สุด เพื่อสกัดออกมาเป็นผลลัพธ์สุดท้าย ซึ่งข้อดีของ Random forest คือ การให้ผลการทำนายที่แม่นยำและเกิดปัญหา overfitting น้อย (Breiman, 2001)

Support Vector Machine (SVM) เป็นหนึ่งในตัวแบบ Machine learning สำหรับใช้ในการจำแนกประเภทข้อมูล โดยใช้อัลกอริทึม SVM ในการค้นหาเส้นที่ใช้แบ่งข้อมูล (hyperplane) ที่ดีที่สุด ซึ่งหลักการของ SVM คือการหาเส้นแบ่งที่มีระยะขอบมากที่สุด (maximum margin) ที่สามารถแบ่งกลุ่มข้อมูลออกจากกันได้ดีที่สุด ข้อได้เปรียบของ SVM คือ มีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมาก นอกจากนี้ การใช้ฟังก์ชันเคอร์เนล (kernel function) เพื่อแปลงข้อมูลไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ (feature space) สามารถจำแนกข้อมูลที่มีความคลุมเครือได้อย่างมีประสิทธิภาพ (Setiyorini & Asmono, 2020)

การวิเคราะห์ข้อมูลด้วยเทคนิค Machine learning มีหลายปัจจัยที่ส่งผลต่อประสิทธิภาพของแบบจำลองการจำแนกที่สร้างขึ้น ซึ่งปัจจัยหนึ่งที่มีความสำคัญอย่างมาก คือการกำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุด (Hyperparameters optimization) กับข้อมูล (Elgeldawi, Sayed, Galal, & Zaki, 2021) โดยงานวิจัยส่วนใหญ่ มีการสร้างแบบจำลองการจำแนกจากเทคนิค Machine learning โดยไม่ได้กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมกับข้อมูล เช่น การสร้างแบบจำลองการจำแนกโดยการใช้ค่าไฮเปอร์พารามิเตอร์เริ่มต้นที่โปรแกรมกำหนด หรือการสร้างแบบจำลองการจำแนกโดยการกำหนดค่าไฮเปอร์พารามิเตอร์เพียงค่าเดียวโดยผู้วิจัย เป็นต้น ดังนั้นในงานวิจัยนี้ เล็งเห็นถึงความสำคัญของการกำหนดค่าไฮเปอร์พารามิเตอร์ที่มีความเหมาะสมกับข้อมูล เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพสูงสุด โดยใช้วิธีการค้นหาแบบกริด (Grid Search) ซึ่งเป็นการสร้างแบบจำลองจากค่าไฮเปอร์พารามิเตอร์ที่กำหนดไว้ทุกชุด และประเมินประสิทธิภาพของแบบจำลองแต่ละชุดเพื่อนำมาเปรียบเทียบ โดยพิจารณาจากค่าความถูกต้องของการจำแนกสูงสุด ซึ่งวิธีการดังกล่าวนี้จะทำให้ค่าไฮเปอร์พารามิเตอร์แต่ละชุดถูกนำมาพิจารณาและประเมินประสิทธิภาพ เพื่อหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดสำหรับข้อมูล

จากการศึกษาปัจจัยเสี่ยงที่ผลส่งต่อการเกิดโรคเบาหวาน พบว่าประวัติคนในครอบครัวป่วยด้วยโรคเบาหวาน และดัชนีมวลกาย เป็นปัจจัยเสี่ยงที่มีความสำคัญต่อการเกิดโรคเบาหวาน (Tsenkova, Karlamangla, & Ryff, 2016) ดังนั้นการนำปัจจัยดังกล่าวมาพิจารณาอิทธิพลร่วมกับปัจจัยอื่น ๆ อาจส่งผลให้ประสิทธิภาพของแบบจำลองเพิ่มขึ้น โดยงานวิจัยที่เกี่ยวข้องซึ่งศึกษาการสร้างแบบจำลองการจำแนกประเภทด้วยเทคนิค Naïve Bayes โดยการพิจารณาอิทธิพลร่วม พบว่าทำให้ความถูกต้องในการจำแนกเพิ่มขึ้น (Changpetch, Pitpeng, Hiriote, & Yuangyai, 2021) ดังนั้นในงานวิจัยนี้จึงสนใจศึกษาแบบจำลองการจำแนกประเภทกรณีการพิจารณาอิทธิพลร่วม เพื่อปรับปรุงประสิทธิภาพของแบบจำลองการจำแนกจากเทคนิค Machine learning ให้ดียิ่งขึ้น

จากปัญหาจำนวนผู้ป่วยโรคเบาหวานที่เพิ่มขึ้นอย่างต่อเนื่อง ผู้วิจัยจึงตระหนักถึงความสำคัญของการประเมินการเป็นโรคเบาหวาน และเพื่อให้โรงพยาบาลในสังกัดสำนักการแพทย์กรุงเทพมหานคร สามารถประเมินการเป็นโรคเบาหวานได้อย่างรวดเร็วและแม่นยำ ผู้วิจัยจึงได้นำปัจจัยเสี่ยงดังกล่าวมาสร้างแบบจำลองการจำแนกด้วยเทคนิค Machine learning ซึ่งเป็นเครื่องมือที่สามารถวิเคราะห์ได้อย่างถูกต้องและแม่นยำและสามารถนำมาประยุกต์ใช้ในการจำแนกการเป็นโรคเบาหวานของผู้รับบริการในโรงพยาบาลเบื้องต้นได้อย่างมีประสิทธิภาพ โดยจุดมุ่งหมายของการศึกษานี้ คือการศึกษาและเปรียบเทียบประสิทธิภาพแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวานเพื่อนำแบบจำลองการจำแนกการเป็นโรคเบาหวานที่มีประสิทธิภาพสูงที่สุด มาใช้ในการประเมินความเสี่ยงของผู้รับบริการในโรงพยาบาลสังกัดสำนักการแพทย์กรุงเทพมหานคร ต่อไป

วัตถุประสงค์ของการศึกษา

เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคที่ใช้ในการสร้างแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวานกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

ประโยชน์ที่คาดว่าจะได้รับ

โรงพยาบาลในสังกัดสำนักการแพทย์ กรุงเทพมหานคร มีการรับรักษาผู้ป่วยโรคเบาหวานเป็นจำนวนมาก ด้วยเหตุนี้ ผลจากการศึกษาจะเป็นประโยชน์ต่อการนำไปประยุกต์ใช้ในการจำแนกการเป็นโรคเบาหวานเบื้องต้นของผู้เข้ารับบริการ นำไปสู่การตรวจเช็คร่างกายที่ละเอียด เพื่อป้องกันเฝ้าระวัง และลดอัตราการเกิดโรคแทรกซ้อนต่าง ๆ จากโรคเบาหวาน รวมถึงช่วยสนับสนุนการตรวจและการวางแผนการรักษาของแพทย์ในโรงพยาบาลได้อีกทางหนึ่ง

ขอบเขตของการศึกษา

การศึกษาการเปรียบเทียบประสิทธิภาพแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน ข้อมูลที่ใช้ในการสร้างแบบจำลองเป็นข้อมูลการประเมินความเสี่ยงการเป็นโรคเบาหวานของผู้ป่วยในโรงพยาบาลสังกัดสำนักการแพทย์ กรุงเทพมหานคร ประกอบด้วย 8 โรงพยาบาล ได้แก่ โรงพยาบาลกลาง โรงพยาบาลตากสิน โรงพยาบาลเจริญกรุงประชารักษ์ โรงพยาบาลหลวงพ่อทวีศักดิ์ ชุตินฺธโร อุทิศ โรงพยาบาลเวชการุณย์รัศมิ์ โรงพยาบาลลาดกระบังกรุงเทพมหานคร โรงพยาบาลราชพิพัฒน์ และโรงพยาบาลสิรินธร ซึ่งเก็บข้อมูลตั้งแต่ปี 2562 ถึง 2564 จำนวนรวมทั้งสิ้น 20,227 ราย

ศึกษาปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวาน ประกอบไปด้วย

1. อายุ
2. เพศ
3. น้ำหนัก
4. ส่วนสูง
5. ดัชนีมวลกาย
6. ค่าความดันขณะหัวใจบีบตัว
7. ค่าความดันขณะหัวใจคลายตัว
8. อัตราการเต้นของหัวใจ
9. ประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง)

เพื่อสร้างแบบจำลองที่เหมาะสมที่สุดในการนำมาใช้ในการจำแนกการเป็นโรคเบาหวานด้วยเทคนิค Machine learning สำหรับการจำแนกประเภท 4 วิธี คือ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor

การพิจารณาค่าไฮเปอร์พารามิเตอร์ของแบบจำลองสำหรับการจำแนกประเภทด้วยเทคนิค Machine learning โดยการกำหนดขอบเขตค่าไฮเปอร์พารามิเตอร์ของเทคนิคต่าง ๆ จากงานจัยที่อ้างอิง ร่วมกับการพิจารณาความคงที่ของค่าความถูกต้องในการจำแนกที่ได้จากการสร้างแบบจำลอง ดังนี้

1. เทคนิคต้นไม้ตัดสินใจ (Decision tree) ขอบเขตของการกำหนดค่าไฮเปอร์พารามิเตอร์อ้างอิงจากงานวิจัย An empirical study on hyperparameter tuning of decision trees (Mantovani et al., 2018)

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
confidenceFactor	0.25, 0.5, 0.75
minNumObj	1, 3, 5, 7, 9

2. เทคนิคต้นไม้ป่าสุ่ม (Random forest) ขอบเขตของการกำหนดค่าไฮเปอร์พารามิเตอร์อ้างอิงจากงานวิจัย Tropical Mangrove Species Classification Using Random Forest Algorithm and Very High-Resolution Satellite Imagery (Intarat & Sillaparat, 2019)

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
numIterations	10, 20, ... , 100
maxDepth	3, 5, 10, 20, none

3. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ขอบเขตของการกำหนดค่าไฮเปอร์พารามิเตอร์อ้างอิงจากงานวิจัยการเรียนรู้ของเครื่องจักรเพื่อการตรวจจับการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย (ธนพล เริ่มปลูก, 2562)

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
kernel	polykernel (exponent=1), polykernel (exponent=2, ... , 5), rbf
C	5, 10, 15, ... , 50
exponent	2, 3, 4, 5
gamma	0.05, 0.1, 0.2, 0.5, 1

4. เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor) ขอบเขตของการกำหนดค่าไฮเปอร์พารามิเตอร์อ้างอิงจากงานวิจัยการจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง (Chuchuepruksaphan & Thanosawan, 2020)

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
K	1, 3, ... , 31
distanceFunction	Euclidean, Manhattan
DistanceWeighting	No distance weighting, Weight by 1/distance

นิยามศัพท์

โรคเบาหวาน เป็นโรคที่ระดับน้ำตาลในเลือดสูงมากกว่าปกติ (hyperglycemia) ต่อเนื่องกันและเป็นระยะเวลานาน มีสาเหตุจากตับอ่อนไม่สามารถสร้างฮอร์โมนอินซูลิน (insulin) ได้อย่างเพียงพอ หรือเกิดจากการที่อวัยวะต่าง ๆ ของร่างกายตอบสนองต่ออินซูลินลดลง (ภาวะดื้ออินซูลิน) ซึ่งอินซูลินเป็นฮอร์โมนที่เกี่ยวข้องกับการควบคุมสมดุลน้ำตาลในกระแสเลือด ทำให้นำน้ำตาลเข้าสู่เซลล์ต่าง ๆ ของร่างกาย เพื่อนำไปใช้เป็นแหล่งพลังงาน ซึ่งโรคเบาหวานสามารถแบ่งได้เป็น 4 ชนิด ตามสาเหตุของการเกิดโรค และสามารถยืนยันชนิดของโรคเบาหวานได้ด้วยผลตรวจทางห้องปฏิบัติการ

Machine learning เป็นรูปแบบหนึ่งของการวิเคราะห์ข้อมูลที่ดำเนินการวิเคราะห์ด้วยแบบจำลองอย่างอัตโนมัติ ซึ่งเป็นสาขาหนึ่งของเทคโนโลยีด้าน AI (artificial intelligence) โดยที่ระบบต่าง ๆ นั้นสามารถที่จะเรียนรู้และมีปฏิสัมพันธ์กับชุดข้อมูลต่าง ๆ รวมถึงสามารถระบุ และทราบรูปแบบต่าง ๆ ที่เกิดขึ้น เพื่อนำไปสู่การตัดสินใจได้เองอย่างมีประสิทธิภาพมากขึ้นและไม่จำเป็นต้องพึ่งพามนุษย์

วิธีการค้นหาแบบกริด (Grid search) เป็นเทคนิคที่ใช้ในการหาค่าไฮเปอร์พารามิเตอร์ด้วยการลองใช้ไฮเปอร์พารามิเตอร์ที่กำหนดไว้ล่วงหน้าทุกชุดและประเมินประสิทธิภาพของแบบจำลองแต่ละชุดโดยแบบจำลองที่มีชุดไฮเปอร์พารามิเตอร์ที่ให้ความถูกต้องในการจำแนกสูงสุดจะถือว่าเป็นแบบจำลองที่ดีที่สุด

โรงพยาบาลในสังกัดสำนักการแพทย์ กรุงเทพมหานคร เป็นหน่วยงานในสังกัดกรุงเทพมหานครที่มีอำนาจหน้าที่เกี่ยวกับการให้บริการตรวจรักษาพยาบาล โดยมีโรงพยาบาลในสังกัดทั้งสิ้น 8 แห่ง ประกอบด้วย โรงพยาบาลกลาง โรงพยาบาลตากสิน โรงพยาบาลเจริญกรุงประชารักษ์ โรงพยาบาลหลวงพ่อทวีศักดิ์ ชุตินฺธโร อุทิศ โรงพยาบาลเวชการุณย์รัศมิ์ โรงพยาบาลลาดกระบัง กรุงเทพมหานคร โรงพยาบาลราชพิพัฒน์ และโรงพยาบาลสิรินธร

บทที่ 2

แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

ในการศึกษาเรื่อง การเปรียบเทียบประสิทธิภาพแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน ได้ทำการรวบรวมแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง เพื่อทำให้เกิดความเข้าใจในเรื่องที่จะทำการศึกษาอย่างชัดเจน ประกอบด้วยหัวข้อดังต่อไปนี้

- ความรู้เกี่ยวกับโรคเบาหวาน
- การประเมินความเสี่ยงการเกิดโรคเบาหวานโดยใช้การคำนวณคะแนนความเสี่ยง (risk score)
- เทคนิคการเรียนรู้ของเครื่อง (Machine learning)
- ต้นไม้ตัดสินใจ (Decision tree)
- ต้นไม้ป่าสุ่ม (Random forest)
- ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
- เพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor)
- การประเมินประสิทธิภาพแบบจำลอง
- เครื่องมือที่ใช้ในงานวิจัย
- งานวิจัยที่เกี่ยวข้อง

ความรู้เกี่ยวกับโรคเบาหวาน

โรคเบาหวาน (Diabetes mellitus) หมายถึง ภาวะที่ร่างกายมีระดับน้ำตาลในเลือดสูงกว่าปกติ เกิดจากเบต้าเซลล์ในกลุ่มเซลล์แลงเกอร์แฮนของตับอ่อนสร้างฮอร์โมนอินซูลินได้น้อย หรือสร้างไม่ได้เลย ซึ่งฮอร์โมนอินซูลินนี้ มีหน้าที่ช่วยให้ร่างกายเผาผลาญน้ำตาลมาใช้เป็นพลังงาน เมื่ออินซูลินในร่างกายไม่เพียงพอกับความต้องการ จะส่งผลให้กระบวนการดูดซึมน้ำตาลในเลือดให้เป็นพลังงานของเซลล์ในร่างกายมีความผิดปกติหรือทำงานได้ไม่เต็มประสิทธิภาพ เมื่อน้ำตาลไม่ถูกนำไปใช้เป็นพลังงาน ทำให้เกิดการคั่งของน้ำตาลในเลือดจนมีน้ำตาลสะสมในเลือดปริมาณมาก (ค่าปกติของน้ำตาลในเลือด คือ 70 - 120 มิลลิกรัมต่อเลือด 100 มิลลิลิตรในขณะอดอาหาร) เมื่อน้ำตาลในเลือดคั่งมาก ๆ จะถูกไตกรองออกมาในปัสสาวะ โดยปกตินำ้ตาลมีประโยชน์ต่อร่างกาย และร่างกาย

พยายามสงวนไว้ไม่ขับทิ้งโดยง่าย ซึ่งไตสามารถกรองน้ำตาลที่ผ่านไตได้ระดับสูงสุด คือ 160 มิลลิกรัมต่อเลือด 100 มิลลิลิตร และสามารถดูดซึมน้ำตาลที่ผ่านการกรองของไตได้นาทีละ 200 มิลลิกรัม แต่เมื่อระดับน้ำตาลในเลือดสูงเกิน 160 มิลลิกรัม ไตจะไม่สามารถดูดซึมน้ำตาลได้มากกว่า 200 มิลลิกรัม ดังนั้นน้ำตาลจึงถูกขับออกมาทางปัสสาวะทำให้ปัสสาวะมีน้ำตาล จึงเรียกโรคนี้ว่าโรคเบาหวาน

สาเหตุการเกิดโรคเบาหวานมักมีส่วนเกี่ยวข้องกับกรรมพันธุ์ กล่าวคือ มีพ่อ แม่ หรือญาติพี่น้อง เป็นโรคเบาหวานด้วย นอกจากนี้ยังมีสาเหตุอื่น เช่น อ้วนเกินไป การใช้ยา เช่น สเตียรอยด์ ยาขับปัสสาวะ ยาเม็ดคุมกำเนิด หรืออาจพบร่วมกับโรคอื่น ๆ เช่น ตับอักเสบเรื้อรัง มะเร็งตับอ่อนระยะสุดท้าย คอพอกเป็นพิษ โรคซิคุชชิ่ง ซินโดรม เป็นต้น หากปล่อยให้ร่างกายอยู่ในสภาวะนี้เป็นเวลานานโดยไม่ได้รับการรักษาอย่างถูกวิธี จะทำให้อวัยวะต่าง ๆ เสื่อมลง และอาจเกิดภาวะแทรกซ้อนที่ร้ายแรงตามมา

อาการของโรคเบาหวานที่พบบ่อย คือ ปัสสาวะบ่อย กระหายน้ำมาก หิวมากกว่าปกติ น้ำหนักลด อ่อนเพลีย สมาธิไม่มี ชาปลายมือปลายเท้า ตามัว ป่วยบ่อย คลื่นไส้ เวียนหัว หงุดหงิด ขบคิดปัญหาง่าย ๆ ได้ไม่ดี และคันตามผิวหนัง อาการที่พบบ่อยนี้จะเริ่มสังเกตเห็นได้เมื่อระดับน้ำตาลในเลือดสูงกว่า 200 มิลลิกรัมต่อเดซิลิตร

ชนิดของโรคเบาหวาน

โรคเบาหวานแบ่งเป็น 4 ชนิดตามสาเหตุของการเกิดโรค ได้แก่

1. โรคเบาหวานชนิดที่ 1 (Type 1 diabetes mellitus : T1DM) เป็นผลจากการทำลายเบต้าเซลล์ที่ตับอ่อนจากภูมิคุ้มกันของร่างกายโดยผ่านขบวนการ cellular mediated ส่วนใหญ่โรคเบาหวานชนิดที่ 1 พบในกลุ่มคนอายุน้อย รูปร่างไม่อ้วน มีอาการปัสสาวะมาก กระหายน้ำ ดื่มน้ำมาก อ่อนเพลีย น้ำหนักลด อาจจะเกิดขึ้นได้อย่างรวดเร็วและรุนแรง (ในวัยเด็ก) ซึ่งในบางกรณีพบภาวะเลือดเป็นกรดจากสารคีโตน (ketoacidosis) เป็นอาการแสดงแรกของโรค หรือมีการแสดงอาการของโรคอย่างช้า ๆ จากระดับน้ำตาลในเลือดที่สูงปานกลางแล้วเกิดภาวะ ketoacidosis เมื่อมีการติดเชื้อหรือได้รับสิ่งกระตุ้นชนิดอื่น โดยในกรณีนี้มักพบในผู้ใหญ่

2. โรคเบาหวานชนิดที่ 2 (Type 2 diabetes mellitus : T2DM) เป็นชนิดที่พบบ่อยที่สุดในคนไทย โดยพบประมาณร้อยละ 95 ของผู้ป่วยโรคเบาหวานทั้งหมด เป็นผลจากการมีภาวะดื้อต่ออินซูลิน (insulin resistance) ร่วมกับความบกพร่องในการผลิตอินซูลิน (relative insulin deficiency) โรคเบาหวานชนิดที่ 2 มักพบในคนอายุ 30 ปีขึ้นไป รูปร่างท้วมหรืออ้วน (ดัชนีมวลกายมากกว่าหรือเท่ากับ 23 กก./ม.^{2}) อาจไม่มีอาการผิดปกติ หรือมีอาการของโรคเบาหวานซึ่งไม่รุนแรง โดยที่อาการแสดงของโรคเบาหวานชนิดที่ 2 อาจคล้ายกับโรคเบาหวานชนิดที่ 1 เช่น การเกิดภาวะ diabetic ketoacidosis โดยที่ความเสี่ยงต่อการเกิดโรคเบาหวานชนิดที่ 2 นี้พบมากเมื่อมีประวัติโรคเบาหวานชนิดนี้ของบุคคลในครอบครัว เช่น พ่อ แม่ พี่หรือน้อง มีอายุสูงขึ้น มีน้ำหนักตัวเพิ่มขึ้น ขาดการออกกำลังกาย และพบมากในหญิงที่มีประวัติการเป็นโรคเบาหวานขณะตั้งครรภ์
3. โรคเบาหวานขณะตั้งครรภ์ (Gestational Diabetes Mellitus : GDM) เกิดจากการที่มีภาวะดื้อต่ออินซูลิน (insulin resistance) ในระหว่างการตั้งครรภ์ ซึ่งมีปัจจัยจากรกและตับอ่อนของมารดาไม่สามารถผลิตอินซูลินให้เพียงพอกับความต้องการของร่างกายได้ สามารถตรวจพบโรคเบาหวานชนิดนี้จากการทำ Oral Glucose Tolerrance Test (OGTT) ในหญิงตั้งครรภ์ไตรมาสที่ 2 หรือ 3 โดยจะทำการตรวจครั้งเดียวโดยการใช้ 75 กรัม OGTT (one-step) หรือจะใช้การตรวจด้วย 50 กรัม glucose challenge test และทำการยืนยันด้วย 100 กรัม OGTT (two-step) โดยปกติแล้วหลังการคลอด โรคเบาหวานจะหายไป
4. โรคเบาหวานที่มีสาเหตุจำเพาะ (Specific types of diabetes due to other causes) เป็นโรคเบาหวานที่มีสาเหตุชัดเจน ได้แก่ โรคเบาหวานที่เกิดจากความผิดปกติทางพันธุกรรม เช่น MODY (Maturity-Onset Diabetes of the Young) โรคเบาหวานที่เกิดจากโรคของตับอ่อน จากความผิดปกติของต่อมไร้ท่อ การติดเชื้อ ปฏิกิริยาภูมิคุ้มกัน หรือโรคเบาหวานที่พบร่วมกับกลุ่มอาการต่าง ๆ โดยผู้ป่วยจะมีอาการแสดงจำเพาะของโรคหรือกลุ่มอาการนั้น ๆ ซึ่งโรคเบาหวานที่มีสาเหตุจำเพาะ มีหลายประเภท ได้แก่

4.1. โรคเบาหวานที่เกิดจากความผิดปกติบนสายพันธุกรรมเดี่ยวที่ควบคุมการทำงานของเบต้าเซลล์ คือ Maturity-Onset Diabetes in the Young (MODY) หลากหลายรูปแบบและความผิดปกติของ Mitochondrial DNA เช่น MODY 3 เป็นความผิดปกติของโครโมโซมที่ 12 ที่ HNF-1 alpha, MODY 2 เป็นความผิดปกติของโครโมโซมที่ 7 ที่ glucokinases และ MODY 1 มีความผิดปกติของโครโมโซมที่ 20 ที่ HNF-4 alpha เป็นต้น

4.2. โรคเบาหวานที่เกิดจากความผิดปกติบนสายพันธุกรรมที่ควบคุมการทำงานของอินซูลิน เช่น Type A insulin resistance, Leprechaunism, Lipoatrophic diabetes และ Rabson-Mendenhall syndrome

4.3. โรคเบาหวานที่เกิดจากโรคของตับอ่อน เช่น Hemochromatosis, Cystic fibrosis และ Fibrocalous pancreatopathy เป็นต้น

4.4. โรคเบาหวานที่เกิดจากโรคของต่อมไร้ท่อ เช่น Acromegaly, Cushing syngrome, Pheochromocytoma, Hyperthyroidism, Glucagonoma และ Aldosteronoma

4.5. โรคเบาหวานที่เกิดจากยาหรือสารเคมีบางชนิด เช่น Pentamidine, Glucocorticoids, Phenytoin, Gamma-interferon, Nicotinic acid และ Diazoxide

4.6. โรคเบาหวานที่เกิดจากโรคติดเชื้อ เช่น Congenital rubella และ Cytomegalovirus

4.7. โรคเบาหวานที่เกิดจากปฏิกิริยาภูมิคุ้มกันที่พบไม่บ่อย เช่น Anti-insulin receptor antibodies และ Stiff-man syndrome

4.8. โรคเบาหวานที่พบร่วมกับกลุ่มอาการต่าง ๆ เช่น Turner syndrome, Prader-Willi syndrome, Friedrich ataxia, Huntington chorea และ Porphyria

ภาวะแทรกซ้อนของโรคเบาหวาน

ภาวะหรือโรคแทรกซ้อนที่เกิดจากโรคเบาหวาน สามารถแบ่งได้เป็น 2 ประเภทใหญ่ ๆ ได้แก่

1. ภาวะแทรกซ้อนเฉียบพลันจากโรคเบาหวาน (Acute diabetes complication)
    1.1. ภาวะที่เกิดจากการมีระดับน้ำตาลในเลือดสูง (Hyperlycemia) การที่ร่างกายมีระดับน้ำตาลในเลือดสูง สามารถทำให้เกิดภาวะ Diabetic Ketoacidosis (DKA) และ Hyperosmolar Hyperglycemic State (HHS) ซึ่งเป็นภาวะแทรกซ้อนเฉียบพลันที่จะเกิดตามมา โดยทั้ง 2 ภาวะนี้สามารถพบได้ในผู้ป่วยโรคเบาหวานชนิดที่ 1 และชนิดที่ 2 แต่ภาวะ DKA มักพบได้บ่อยครั้งในผู้ป่วยโรคเบาหวานชนิดที่ 1 ส่วนภาวะ HHS มักพบได้บ่อยครั้งในผู้ป่วยโรคเบาหวานชนิดที่ 2
    1.2. ภาวะที่มีระดับน้ำตาลในเลือดต่ำ (hypoglycemia) เป็นภาวะแทรกซ้อนเฉียบพลัน ที่สามารถพบได้ในผู้ป่วยโรคเบาหวาน โดยทั่วไประดับน้ำตาลในเลือดต่ำเมื่อมีระดับกลูโคสในพลาสมาต่ำกว่า 40-50 มก./ดล. ร่วมกับการมีอาการของภาวะน้ำตาลในเลือดต่ำ ได้แก่ มองภาพไม่ชัด พูดไม่ชัด และมีอาการชัก เป็นต้น
2. โรคแทรกซ้อนเรื้อรังจากโรคเบาหวาน (Chronic diabetes complication) ผู้ป่วยโรคเบาหวานส่วนใหญ่มักเสียชีวิตด้วยโรคแทรกซ้อนเรื้อรัง มักเกิดขึ้นในผู้ป่วยที่เป็นโรคเบาหวานมานานอย่างน้อย 5 ปีขึ้นไป ซึ่งโรคแทรกซ้อนเรื้อรังจะเกิดขึ้นอย่างช้า ๆ โดยที่ผู้ป่วยนั้นไม่รู้ตัว และจะขึ้นอยู่กับระยะเวลาการเป็นโรคเบาหวานของผู้ป่วย โดยเฉพาะในผู้ที่ควบคุมระดับน้ำตาลในเลือดไม่ได้ตามเกณฑ์ที่กำหนด โดยภาวะแทรกซ้อนเรื้อรังสามารถแบ่งออกเป็น
    2.1. ภาวะแทรกซ้อนที่หลอดเลือดขนาดเล็ก
    2.2. ภาวะแทรกซ้อนที่จอประสาทตา
    2.3. ภาวะแทรกซ้อนที่ไต
    2.4. ภาวะแทรกซ้อนเรื้อรังที่เส้นประสาท
    2.5. ภาวะแทรกซ้อนที่หลอดเลือดขนาดใหญ่
    2.6. โรคหลอดเลือดหัวใจ

2.7. โรคหลอดเลือดสมอง

2.8. โรคหลอดเลือดส่วนปลายอุดตัน ซึ่งเป็นปัจจัยส่งเสริมให้เกิดแผลที่เท้าในผู้ป่วยเบาหวาน

การประเมินความเสี่ยงการเกิดโรคเบาหวานโดยใช้การคำนวณคะแนนความเสี่ยง (risk score)

การประเมินความเสี่ยงการเกิดโรคเบาหวานเพื่อทำนายการเกิดโรค สามารถใช้ข้อมูลจากการศึกษาปัจจัยเสี่ยงหลายอย่างที่สามารถประเมินได้ง่ายด้วยแบบสอบถามและการตรวจร่างกายเบื้องต้นโดยไม่ต้องเจาะเลือด ดังตารางที่ 3 แล้วนำข้อมูลมาคำนวณเป็นคะแนนความเสี่ยง (risk score) สามารถใช้ทำนายความเสี่ยงในการเกิดโรคเบาหวานในอนาคตได้แม่นยำพอสมควรในคนไทย การประเมินโดยวิธีนี้สามารถนำมาใช้เป็นแนวทางปฏิบัติเพื่อการคัดกรองโรคเบาหวานในชุมชนซึ่งมีข้อจำกัดทางด้านงบประมาณ

รายละเอียดการแปลผลคะแนนความเสี่ยงที่ได้จากการประเมินความเสี่ยงการเกิดโรคเบาหวานและข้อแนะนำเพื่อการปฏิบัติดังในตารางที่ 4 เมื่อนำคะแนนของแต่ละปัจจัยเสี่ยงมารวมกัน คะแนนจะอยู่ในช่วง 0-17 คะแนน โดยอาจทำการตรวจคัดกรองเบาหวานเฉพาะผู้ที่มีคะแนนความเสี่ยงตั้งแต่ 6 คะแนนขึ้นไป เป็นต้น

โดยสรุป การประเมินความเสี่ยงเพื่อตรวจคัดกรองผู้ป่วย นอกจากจะช่วยค้นหาผู้ที่มีโอกาสเสี่ยงที่จะเป็นโรคเบาหวานในอนาคต แล้วยังช่วยให้สามารถตรวจพบผู้ที่เป็นเบาหวาน โดยไม่แสดงอาการเพื่อป้องกันการเกิดโรคแทรกซ้อนจากโรคเบาหวาน และให้ได้รับการรักษาตั้งแต่เริ่มต้นได้อีกทางหนึ่ง

ตารางที่ 3 ปัจจัยเสี่ยงของโรคเบาหวานและคะแนนความเสี่ยง

ปัจจัยเสี่ยง	คะแนนความเสี่ยง
อายุ	
• 34 - 39 ปี	0
• 40 - 44 ปี	0
• 45 - 49 ปี	1
• ตั้งแต่ 50 ปีขึ้นไป	2
เพศ	
• หญิง	0
• ชาย	2
ดัชนีมวลกาย	
• ต่ำกว่า 23 กก./ม.^{2}	0
• ตั้งแต่ 23 ขึ้นไปแต่น้อยกว่า 27.5 กก./ม.^{2}	3
• ตั้งแต่ 27.5 กก./ม.^{2} ขึ้นไป	5
รอบเอว	
• ผู้ชายน้อยกว่า 90 ซม. ผู้หญิงน้อยกว่า 80 ซม.	0
• ผู้ชายตั้งแต่ 90 ซม. ขึ้นไป ผู้หญิงตั้งแต่ 80 ซม. ขึ้นไป	2
ความดันโลหิตสูง	
• ปกติ	0
• ผิดปกติ	2
ประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง)	
• ไม่มี	0
• มี	4

ตารางที่ 4 การแปลผลคะแนนความเสี่ยงของโรคเบาหวานและข้อแนะนำ

ผลรวมคะแนน	ระดับความเสี่ยง	ข้อแนะนำ
น้อยกว่าหรือเท่ากับ 2 คะแนน	น้อย	- ออกกำลังกายสม่ำเสมอ - ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม - ตรวจวัดความดันโลหิต - ควรประเมินความเสี่ยงซ้ำทุก 3 ปี
3 - 5 คะแนน	ปานกลาง	- ออกกำลังกายสม่ำเสมอ - ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม - ตรวจวัดความดันโลหิต - ควรประเมินความเสี่ยงซ้ำทุก 1-3 ปี
6 - 8 คะแนน	สูง	- ควบคุมอาหารและออกกำลังกายสม่ำเสมอ - ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม - ตรวจวัดความดันโลหิต - ตรวจระดับน้ำตาลในเลือด - ควรประเมินความเสี่ยงซ้ำทุก 1-3 ปี
มากกว่า 8 คะแนน	สูงมาก	- ควบคุมอาหารและออกกำลังกายสม่ำเสมอ - ควบคุมน้ำหนักตัวให้อยู่ในเกณฑ์ที่เหมาะสม - ตรวจวัดความดันโลหิต - ตรวจระดับน้ำตาลในเลือด - ควรประเมินความเสี่ยงซ้ำทุก 1 ปี

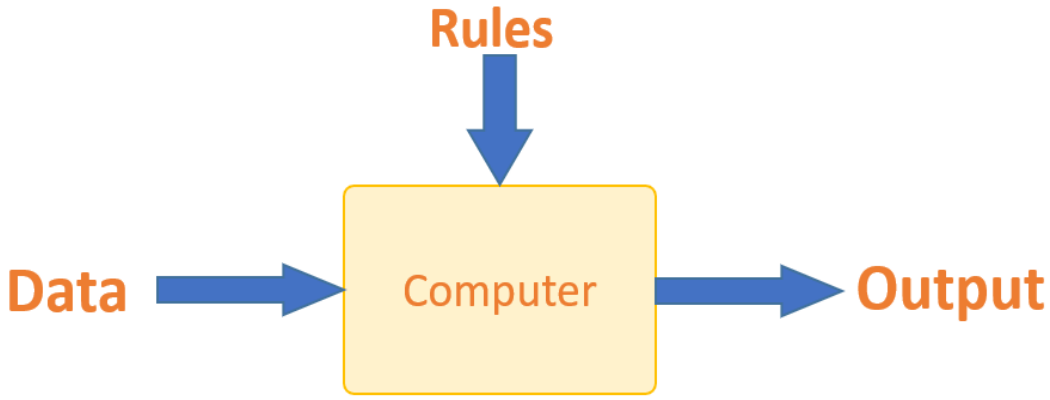
ที่มา : วารสารแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี

จากการศึกษาปัจจัยเสี่ยงการเกิดโรคเบาหวานดังกล่าว ผู้วิจัยจึงนำปัจจัยเสี่ยง คือ อายุ เพศ น้ำหนัก ส่วนสูง ดัชนีมวลกาย ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลายตัว อัตราการเต้นของหัวใจ และประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง) มาใช้ในการสร้างแบบจำลอง Machine learning เนื่องจากเป็นปัจจัยเสี่ยงที่สามารถประเมินได้ง่ายและมีความสำคัญซึ่งส่งผลต่อการเกิดโรคเบาหวาน

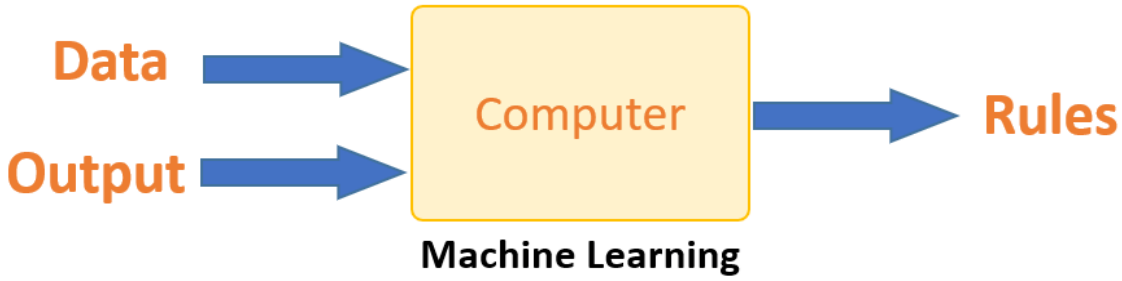
เทคนิคการเรียนรู้ของเครื่อง (Machine learning)

Machine learning คือ การเรียนรู้ (Learning) และการอนุมาน (Inference) โดยที่เครื่อง (Machine) จะเรียนรู้ผ่านการค้นพบรูปแบบหรือแบบแผนซ้ำ ๆ ผ่านชุดข้อมูลฝึกฝน และใช้อัลกอริทึมเพื่อสร้างแบบจำลอง (Model) โดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ จากนั้นนำไปทดสอบประสิทธิภาพกับข้อมูลชุดทดสอบ เพื่อนำแบบจำลองที่มีประสิทธิภาพสูงสุดที่ได้ไปใช้ในการทำนายผลลัพธ์ของชุดข้อมูลใหม่

การเขียนโปรแกรมสมัยก่อนแตกต่างอย่างมากกับ Machine learning เนื่องจากการเขียนโปรแกรมในสมัยก่อน (ภาพที่ 1) นั้นโค้ดทั้งหมดจะต้องถูกกำหนดแนวทางไว้อย่างชัดเจนจากผู้เชี่ยวชาญ โดยแต่ละกฎจะขึ้นอยู่กับพื้นฐานความเข้าใจด้านตรรกศาสตร์ (Logic Foundation) เครื่องจะทำงานและส่งผลลัพธ์ออกมาตามคำสั่ง (Logical statement) เมื่อแนวทางการประมวลผลเริ่มซับซ้อนมากขึ้น ทำให้ต้องมีการเขียนกฎมากขึ้น ดังนั้นในปัจจุบันมีการแก้ปัญหาการเขียนโปรแกรมสมัยก่อนโดยใช้ Machine learning (ภาพที่ 2) ซึ่งเครื่องจะเรียนรู้ความเกี่ยวข้องกันระหว่างข้อมูลขาเข้าและข้อมูลขาออก เพื่อที่จะเขียนกฎแสดงความเกี่ยวข้องนั้นขึ้นมา โดยที่โปรแกรมเมอร์ไม่จำเป็นต้องเขียนกฎใหม่ทุกครั้งเมื่อมีข้อมูลใหม่ อัลกอริทึมของเครื่องจะปรับเข้ากับข้อมูลใหม่เพื่อปรับปรุงประสิทธิภาพในการประมวลผล



ภาพที่ 2 การเขียนโปรแกรมแบบดั้งเดิม



ภาพที่ 3 Machine learning

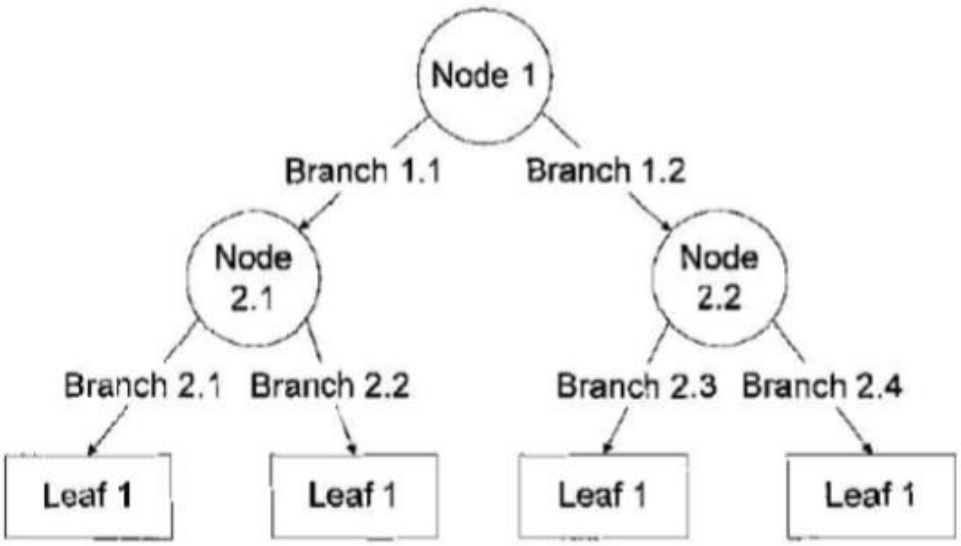
Machine learning แบ่งออกเป็นการเรียนรู้ 2 แบบใหญ่ ๆ ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised learning) และ การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) ซึ่งการเรียนรู้แบบมีผู้สอน เป็นกลุ่มของอัลกอริทึมที่เน้นสอนคอมพิวเตอร์ โดยการศึกษาจากข้อมูลตัวอย่าง เพื่อทำให้คอมพิวเตอร์สามารถหาคำตอบได้ด้วยตัวเอง หลังจากเรียนรู้จากชุดข้อมูลตัวอย่างที่ได้ป้อนให้ไปแล้วระยะหนึ่ง โดยหลักการ Supervised learning สามารถนำไปประยุกต์ใช้แก้ปัญหาได้ 2 รูปแบบ คือ Regression และ Classification ส่วนการเรียนรู้แบบไม่มีผู้สอน ไม่จำเป็นต้องมีค่าเป้าหมายของแต่ละข้อมูลตัวอย่างในระหว่างการเรียนรู้ ซึ่งการเรียนรู้ประเภทนี้จะเป็นการระบุกลุ่มของข้อมูลที่ใส่เข้าไป โดยจะอิงกับวิธีการจัดกลุ่มซึ่งได้เรียนรู้จากข้อมูลที่เคยพบมา หลักการ Unsupervised learning สามารถนำไปประยุกต์ใช้แก้ปัญหาประเภท Clustering

ในปัจจุบันมีอัลกอริทึมของ Machine learning อยู่มากมาย โดยการเลือกอัลกอริทึมจะขึ้นอยู่กับวัตถุประสงค์ของการนำไปใช้งาน ในงานวิจัยนี้ขอเน้นเพียงอัลกอริทึมของ Machine learning สำหรับการ Classification โดยเทคนิค Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor

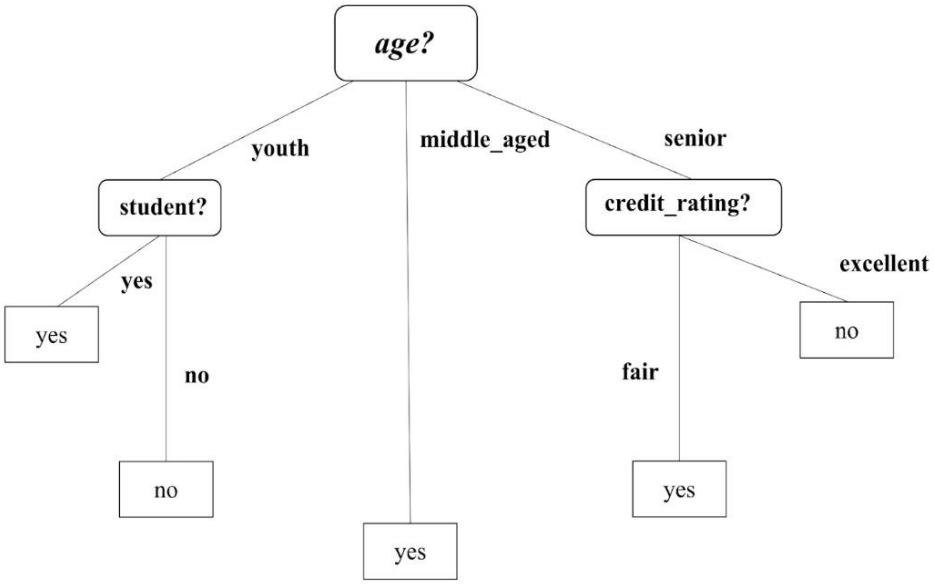
ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจ (Decision tree) เป็นการนำข้อมูลมาสร้างแบบจำลอง (Sahin, 2011) มีลักษณะเหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised learning) คือ สร้างแบบจำลองขึ้นมาจากข้อมูลที่นำมาใช้เรียนรู้ (Training set) และสามารถทำนายกลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้โดยผลการทำนายจะขึ้นอยู่กับตัวแปรต้น

รูปแบบโครงสร้างต้นไม้ตัดสินใจ ประกอบด้วย โหนดราก (Root node) ซึ่งจะแตกออกเป็นโหนดลูก โดยจะมีกิ่งของต้นไม้ (Branch) ในการเชื่อมระหว่างโหนดต่าง ๆ และโหนดลูกระดับสุดท้ายเรียกว่า โหนดใบ (Leaf node) โดยที่แต่ละโหนดของโหนดรากและโหนดลูกจะแสดงค่าคุณลักษณะ (Attribute) ของโหนดที่แตกออกมาใช้ในการทดสอบข้อมูล โดยจำนวนของกิ่งจะเท่ากับจำนวนค่าของคุณลักษณะในโหนดนั้น ส่วนโหนดใบจะแสดงกลุ่ม (Class) ของข้อมูลที่กำหนด ซึ่งสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ ดังภาพที่ 3 การทำนายกลุ่มของข้อมูลการทำงานจะเริ่มต้นจากโหนดราก ซึ่งจะนำค่าคุณลักษณะต่าง ๆ ของข้อมูลไปเปรียบเทียบกับคุณลักษณะของโหนด จากนั้นทำการตัดสินใจในการจำแนกกลุ่มของข้อมูลโดยการเปรียบเทียบคุณลักษณะไปเรื่อย ๆ จนกระทั่งถึงโหนดใบจึงได้กลุ่มของข้อมูลที่ถูกกำหนด



ภาพที่ 4 ส่วนประกอบต้นไม้ตัดสินใจ



ภาพที่ 5 ตัวอย่างต้นไม้ตัดสินใจ
ที่มา : (Jingfen Han et al., 2012)

จากภาพที่ 4 เป็นตัวอย่างของต้นไม้ตัดสินใจที่มีผลลัพธ์ คือ yes และ no โดยโหนดราก คือ age ซึ่งประกอบไปด้วยกิ่ง 3 กิ่ง คือ กิ่ง youth กิ่ง middle_aged และ กิ่ง senior ในส่วนของกิ่ง youth มีโหนดภายใน คือ student ประกอบไปด้วยกิ่ง 2 กิ่ง คือ yes และ no ในกิ่งมีใบที่เป็นคำตอบ คือ yes และ no ตามลำดับ กิ่ง middle_aged มีใบที่เป็นคำตอบ คือ yes และส่วนสุดท้าย กิ่ง senior มีโหนดภายใน คือ credit_rating ประกอบไปด้วยกิ่ง 2 กิ่ง คือ fair และ excellent ในกิ่งมีใบที่เป็นคำตอบ คือ yes และ no ตามลำดับ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) เริ่มจากการแบ่งข้อมูลออกเป็นโครงสร้างต้นไม้ โดยที่แผนภาพนั้นจะเป็นโครงสร้างที่มีกฎต่าง ๆ เกิดขึ้นตามเป้าหมายของการใช้งาน เมื่อได้โครงสร้างต้นไม้แล้วจะสามารถนำโครงสร้างที่ได้นั้นไปใช้งานกับข้อมูลอื่น ๆ ได้โดยนำข้อมูลนั้นผ่านกฎการตัดสินใจ (Decision rules) ขั้นตอนการสร้างต้นไม้ตัดสินใจ มีดังนี้ (Han, Kamber, & Pei, 2012)

1. ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกฝน (training set)
2. ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนกันอยู่ จะต้องวัดค่าคุณลักษณะ (attribute measurement) ของแต่ละคุณลักษณะเพื่อใช้เป็นเกณฑ์ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด ซึ่งจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจ
4. กิ่งของต้นไม้จะถูกสร้างขึ้นจากค่าของคุณลักษณะต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบ และจะแบ่งข้อมูลออกตามกิ่งของต้นไม้ตัดสินใจที่สร้างขึ้น
5. วนซ้ำกระบวนการเดิมเพื่อหาคุณลักษณะที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุดสำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่ง เพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับถัดไป
6. แบ่งแยกข้อมูลและแตกกิ่งของต้นไม้ตัดสินใจไปเรื่อย ๆ โดยวนซ้ำกระบวนการเดิม ซึ่งจะสิ้นสุดกระบวนการก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งเป็นจริง

อัลกอริทึม ID3 (Iterative Dichotomiser 3) (ศุภชัย ประคองศิลป์, 2551) เป็นอัลกอริทึมพื้นฐานที่ใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่งข้อมูล โดยวิธีกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลำดับของค่า Gain สูงที่สุดเป็นข้อมูลเริ่มต้นและข้อมูลถัดไปมีค่าลดลงตามลำดับ

อัลกอริทึม C4.5 (J48) (พวงทิพย์ แท่นแสง & ลือพล พิพานเมฆาภรณ์, 2550) เป็นอัลกอริทึมที่ใช้ในการสร้างกฎจากต้นไม้ตัดสินใจ (Decision tree) ถูกออกแบบโดย Quinlan (1992) ซึ่งพัฒนากระบวนการเพิ่มเติมจากอัลกอริทึม ID3 ดังนี้

1. อัลกอริทึม C4.5 จะหลีกเลี่ยงการสร้างต้นไม้ตัดสินใจที่ใหญ่เกินไป เนื่องจากการมีข้อมูลจำนวนมาก อย่างไรก็ตามขึ้นอยู่กับการกำหนดความลึกเมื่อมีการเจริญเติบโตของต้นไม้ตัดสินใจ
2. ความผิดพลาดลดลง เนื่องจากมีการตัดทอนความผิดพลาดออกไป (prunning node)
3. มีกระบวนการสร้างกฎหลังจากการตัดข้อมูลที่มีความผิดพลาดออก
4. ใช้กับข้อมูลต่อเนื่อง (continuous attributes) ที่เป็นตัวเลขได้ เช่น อายุ จำนวนเงิน เป็นต้น
5. การเลือก attributes โดยการพิจารณาค่าวัดคุณลักษณะที่เหมาะสม
6. สามารถใช้กับชุดข้อมูล training data ที่มีค่าผิดพลาด (missing attribute)

การวัดค่าคุณลักษณะ (Attribute measurement)

ค่าเกนความรู้ (Information gain) คือการประเมินค่าซึ่งถูกนำมาใช้ในการแบ่งข้อมูล โดยที่การคำนวณค่า Gain สำหรับแต่ละมิติข้อมูล หากมิติของข้อมูลใดมีค่า Gain สูงสุด ข้อมูลนั้นจะถูกเลือกเป็นกลุ่มย่อยที่มีอำนาจในการจำแนก โดยการคำนวณค่า Entropy ของตัวแปรตาม และตัวแปรอิสระแต่ละตัว แสดงดังสมการที่ 1 และ 2 ตามลำดับ

$$\mathrm{Entropy(Y)} = -\sum_{i=1}^{N} p_i \log_2(p_i) \quad (1)$$

โดยที่ $\mathrm{Entropy(Y)}$ คือ ค่า Entropy ของตัวแปรตาม

p_i คือ ค่าความน่าจะเป็นที่จะเกิดคลาสที่ i ในตัวแปรตาม

N คือ จำนวนคลาสทั้งหมดในตัวแปรตาม

จากนั้นหาเกนสารสนเทศของตัวแปรอิสระแต่ละตัว ดังสมการที่ 2

$$Entropy(A) = \sum_{K=1}^{M} \left[P_{K,A} \times Entropy(A_K) \right] \quad (2)$$

โดยที่ $\mathrm{Entropy(A)}$ คือ ค่า Entropy ของตัวแปรอิสระ A

$\mathrm{Entropy}(A_K)$ คือ ค่า Entropy ของกลุ่มที่ K ในตัวแปรอิสระ A

$P_{K,A}$ คือ ค่าความน่าจะเป็นที่จะเกิดกลุ่มที่ K ในตัวแปรอิสระ A

M คือ จำนวนกลุ่มทั้งหมดในตัวแปรอิสระ A

เมื่อได้ค่า Entropy ของตัวแปรตามและตัวแปรอิสระทั้งหมดแล้ว ขั้นตอนต่อไป คือ หาค่า Gain ของตัวแปรอิสระแต่ละตัว ดังสมการที่ 3 จากนั้นจึงเลือกตัวแปรอิสระที่มีค่า Gain สูงสุดเป็นตัวจำแนกชุดข้อมูล

$$\mathrm{Gain(A)} = \mathrm{Entropy(Y)} - \mathrm{Entropy(A)} \quad (3)$$

โดยที่ $\mathrm{Gain(A)}$ คือ ค่า Gain ของตัวแปรอิสระ A

$\mathrm{Entropy(Y)}$ คือ ค่า Entropy ของตัวแปรตาม

$\mathrm{Entropy(A)}$ คือ ค่า Entropy ของตัวแปรอิสระ A

ตารางที่ 5 ตัวอย่างข้อมูล

No	Age	Income	Student	Credit_rating	Class: Buy_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

ที่มา : การสร้างต้นไม้ตัดสินใจ (Jingfen Han et al., 2012)

จากตารางที่ 5 แสดงตัวอย่างข้อมูลการตัดสินใจซื้อคอมพิวเตอร์ ซึ่งมีข้อมูลทั้งหมด 14 ชุด 4 ตัวแปรอิสระ ได้แก่ age, income, student และ credit_rating และตัวแปรตาม ได้แก่ การตัดสินใจซื้อคอมพิวเตอร์ (Buy_computer) สามารถแสดงขั้นตอนการคำนวณค่าวัดคุณลักษณะโดยวิธีค่าเกนความรู้ (Information gain) ได้ดังนี้

ขั้นตอนที่ 1 คำนวณค่า Entropy ของตัวแปรตาม จากสมการที่ 1

$$\text{Entropy(Y)} = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

ขั้นตอนที่ 2 คำนวณค่า Entropy ของตัวแปรอิสระแต่ละตัว จากสมการที่ 2

$$\begin{aligned}\text{Entropy(age} &= \frac{5}{14}\times\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)\\&+\frac{4}{14}\times\left(-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right)\\&+\frac{5}{14}\times\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)\\&=\ 0.694\end{aligned}$$

$$\begin{aligned}\text{Entropy(income)} &= \frac{4}{14}\times\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)\\&+\frac{6}{14}\times\left(-\frac{4}{6}\log_2\left(\frac{4}{6}\right)-\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)\\&+\frac{4}{14}\times\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\\&=\ 0.911\end{aligned}$$

$$\begin{aligned}\text{Entropy(student)} &= \frac{7}{14}\times\left(-\frac{6}{7}\log_2\left(\frac{6}{7}\right)-\frac{1}{7}\log_2\left(\frac{1}{7}\right)\right)\\&+\frac{7}{14}\times\left(-\frac{3}{7}\log_2\left(\frac{3}{7}\right)-\frac{4}{7}\log_2\left(\frac{4}{7}\right)\right)\\&=\ 0.787\end{aligned}$$

$$\begin{aligned}\text{Entropy(credit\_rating)} &= \frac{8}{14}\times\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right)-\frac{2}{8}\log_2\left(\frac{2}{8}\right)\right)\\&+\frac{6}{14}\times\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right)-\frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)\\&=\ 0.892\end{aligned}$$

ขั้นตอนที่ 3 คำนวณค่า Gain ของตัวแปรอิสระแต่ละตัว จากสมการที่ 3

$$\text{Gain(age)} = 0.940 - 0.694 \ = \ 0.246$$

$$\text{Gain(income)} = 0.940 - 0.911 \ = \ 0.029$$

$$\text{Gain(student)} = 0.940 - 0.787 \ = \ 0.151$$

$$\text{Gain(credit\_rating)} = 0.940 - 0.892 \ = \ 0.048$$

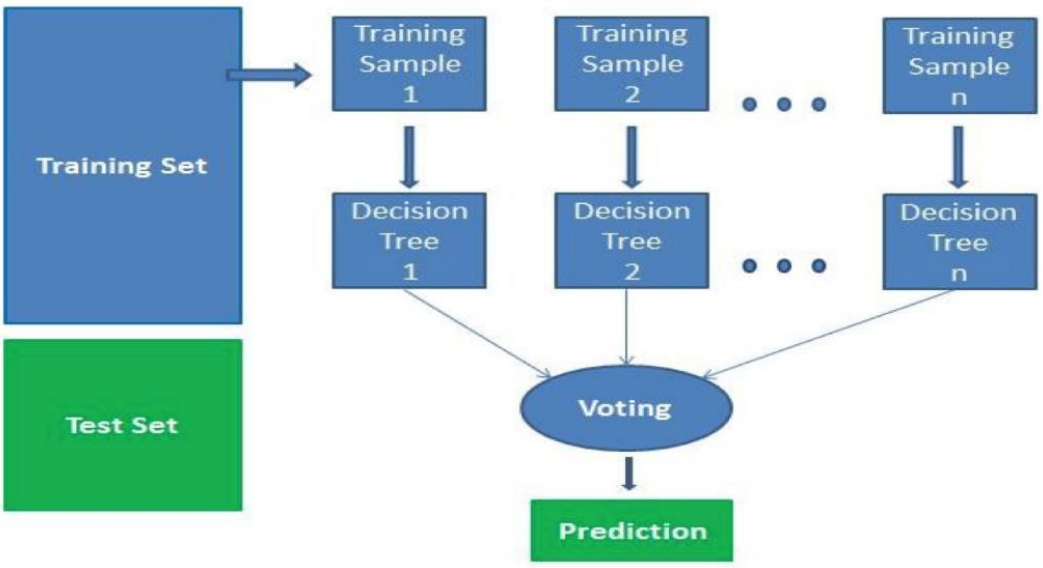
หลักการพิจารณาค่า Gain คือเลือกค่า Gain ที่มีค่าสูงสุดเป็นโหนดเริ่มต้นในการสร้างต้นไม้ตัดสินใจจากขั้นตอนที่ 3 ค่า Gain สูงสุด คือ age ดังนั้นจึงเลือก age เป็นโหนดเริ่มต้น หรือโหนดรากในการสร้างต้นไม้ตัดสินใจ สำหรับคุณลักษณะอื่น ๆ ที่มีค่า Entropy ไม่เท่ากับศูนย์จะถูกคำนวณค่า Gain และเลือกค่าที่มากที่สุดเพื่อกำหนดเป็นโหนดต่อมา จนกระทั่งมีค่า Entropy เท่ากับศูนย์จึงหยุดกระบวนการสร้างต้นไม้ตัดสินใจ

ต้นไม้ป่าสุ่ม (Random forest)

เทคนิคต้นไม้ป่าสุ่ม (Random forest) เป็นเทคนิคการเรียนรู้แบบมีผู้สอนแบบหนึ่งที่ใช้งานง่าย ถูกพัฒนามาจากเทคนิคต้นไม้ตัดสินใจ (Decision tree) สำหรับงานด้านการจำแนกกลุ่มข้อมูล ต้นไม้ป่าสุ่มเกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ตัดสินใจ (Hartshorn, 2016) ซึ่งค่าความคลาดเคลื่อนโดยรวมของต้นไม้ป่าสุ่มจะถูกเปลี่ยนให้เป็นค่าลิมิต ทำให้จำนวนของต้นไม้ในป่าเพิ่มขึ้น ค่าความคลาดเคลื่อนโดยรวมจะขึ้นกับความมั่นคงของต้นไม้แต่ละต้น โดยจะใช้วิธีการสุ่มเลือกคุณสมบัติเพื่อการแบ่งแยกโหนด ทำให้ค่าความผิดพลาดลดลง

หลักการที่สำคัญของเทคนิคต้นไม้ป่าสุ่ม คือการสร้างต้นไม้ตัดสินใจอย่างง่ายจำนวนมากในขั้นตอนการฝึกฝน และใช้วิธีการทำ Majority vote หรือการลงคะแนนเสียงข้างมากช่วยในการตัดสินใจผลการจำแนกกลุ่มของข้อมูล โดยนำผลการจำแนกของต้นไม้แต่ละต้นมาร่วมกันตัดสินใจ จากนั้นจึงเลือกผลลัพธ์การจำแนกที่ได้รับการโหวตมากที่สุด

ในขั้นตอนการฝึกฝนตัวจำแนก จะใช้เทคนิคที่เรียกว่า Bagging (Bootstrap Aggregation) คือ การสร้างข้อมูลย่อยหลายชุดจากชุดฝึกฝนด้วยวิธีสุ่มแบบใส่คืน (Sampling with Replacement) และสร้างต้นไม้ตัดสินใจขึ้นมาหลาย ๆ ต้นจากชุดข้อมูลดังกล่าว ซึ่งต้นไม้ตัดสินใจแต่ละต้นจะไม่มีการตัดแต่ง (prunning) โดยการสุ่มข้อมูลแต่ละครั้งจะมีข้อมูลส่วนหนึ่งที่ไม่ถูกเลือกมาสร้างต้นไม้ตัดสินใจ เรียกว่า out-of-bag ซึ่งเป็นข้อมูลที่สามารถนำไปใช้ในการตรวจสอบความแม่นยำของต้นไม้ตัดสินใจแต่ละต้นที่สร้างขึ้นได้ จากการคำนวณค่าความผิดพลาดของข้อมูล out-of-bag



ภาพที่ 6 การทำงานของอัลกอริทึม Random forest

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine :SVM) จัดเป็น Machine learning ประเภทแบบจำลอง ที่ได้รับความนิยมในงานการจำแนกประเภทข้อมูล เนื่องจาก SVM สามารถทำงานได้ดีกับข้อมูลที่มีมิติสูงหรือข้อมูลที่มีขนาดมิติมากกว่าจำนวนตัวอย่าง (Huwaidah, Adiwijaya, & Faraby, 2021) ซึ่งสามารถจำแนกกลุ่มข้อมูลโดยใช้ไฮเปอร์เพลน (hyperplane) ในการแบ่งกลุ่มข้อมูลออกจากกัน

ในปริภูมิ p มิติ ไฮเปอร์เพลนย่อยที่มี p-1 มิติ เช่น ระนาบของปริภูมิ 2 มิติ ไฮเปอร์เพลนจะเป็นปริภูมิย่อยที่มี 1 มิติ ซึ่งคือเส้นตรง เขียนได้ในรูป

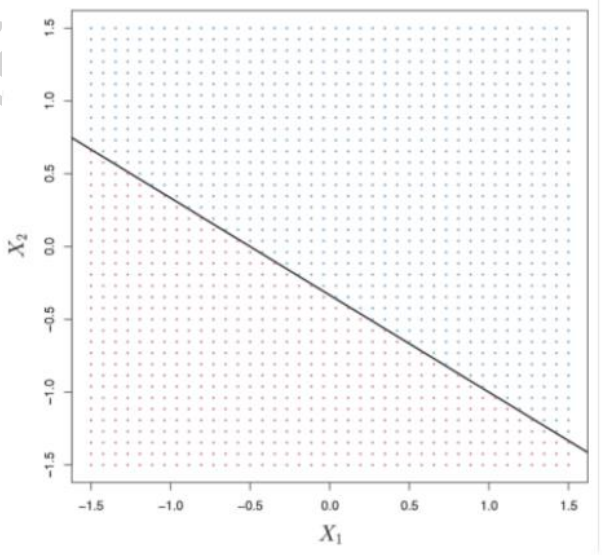
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{4}$$

โดยทั่วไปแล้ว ในปริภูมิ p มิติ ไฮเปอร์เพลนจะเขียนได้ดังในรูป

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0 \tag{5}$$

หาก $\mathbf{x} = \left(X_1, X_2, \ldots, X_p\right)^T$ ทำให้สมการที่ 5 เป็นจริง ดังนั้นแสดงว่า X จะอยู่บนไฮเปอร์เพลน หาก X ไม่อยู่บนไฮเปอร์เพลน X จะอยู่ในฝั่งใดฝั่งหนึ่งของไฮเปอร์เพลน นั่นคือ

$$\begin{aligned} &\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0 \\ &\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0 \end{aligned} \quad \text{หรือ} \tag{6}$$



ภาพที่ 7 ตัวอย่างไฮเปอร์เพลนในปริภูมิ 2 มิติ

ที่มา : (James, Witten, Hastie, & Tibshirani, 2017)

สำหรับการจำแนกประเภทของข้อมูลที่มีตัวแปรอิสระ p ตัว ดังนั้นจะต้องไฮเปอร์เพลนสำหรับการแบ่งข้อมูลปริภูมิ p มิติ นั่นคือ หากมีข้อมูลจำนวน n คู่ คือ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ โดย x_i เป็นตัวแปรอิสระ $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$; $x_i \in \mathbb{R}^p$, $\forall_i = 1, \dots, n$ และ y_i เป็นตัวแปรตามซึ่งใช้ในการระบุ class ของข้อมูล (โดยตัวอย่างนี้ จะพิจารณากรณีที่มี 2 class) ซึ่ง $y_i \in \{-1, 1\}$ ดังนั้นไฮเปอร์เพลนที่ใช้แบ่งข้อมูล นิยามโดย

$$\{x : f(x) = \beta_0 + x^T\beta = 0\} \tag{7}$$

เมื่อ $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ และเป็นเวกเตอร์หนึ่งหน่วย $\left(\|\beta\| = \sum_{j=1}^{p} \beta_j^2 = 1\right)$ จากอสมการที่ 7 ทำให้สามารถจำแนก class ของข้อมูลโดยใช้ไฮเปอร์เพลนในการแบ่งข้อมูล ซึ่งจะพิจารณาจากเครื่องหมายของ $f(x)$ นั่นคือ

$$G(x) = \text{sign}[\beta_0 + x^T\beta] \tag{8}$$

โดย $G(x_i)$ จะเป็นบวก เมื่อ $y_i = 1$ และเป็นลบ เมื่อ $y_i = -1$ อีกนัยหนึ่ง คือ

$$\beta_0 + x^T\beta > 0 \qquad \text{ถ้า} \qquad y_i = 1 \tag{9}$$

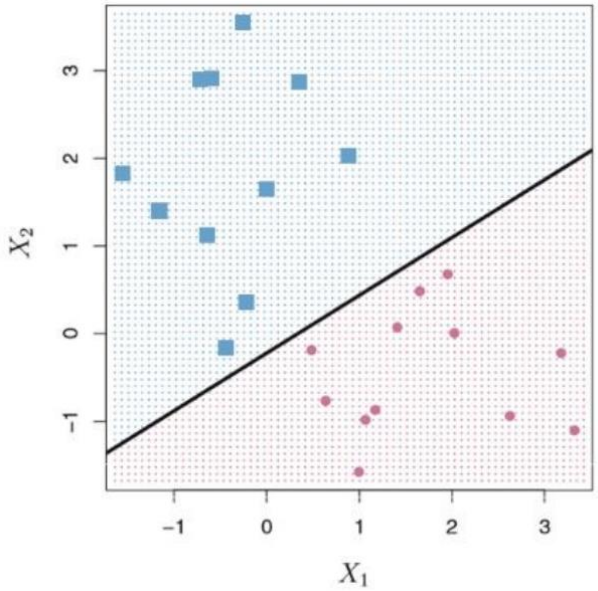
$$\beta_0 + x^T\beta < 0 \qquad \text{ถ้า} \qquad y_i = -1 \tag{10}$$

ซึ่งสามารถรวมเป็นอสมการเดียวได้ว่า

$$y_i(\beta_0 + {x_i}^T\beta) > 0 \ , \ \forall_i = 1, \dots, n \tag{11}$$

ในขณะที่เครื่องหมายของ $f(x)$ ระบุ class ของข้อมูล ขนาดของ $f(x)$ จะบอกถึงระยะทางระหว่างจุดของข้อมูลและไฮเปอร์เพลน หาก $f(x_i)$ มีค่าใกล้เคียง 0 แสดงว่า x_i นั้นอยู่ใกล้กับไฮเปอร์เพลนที่ใช้ในการแบ่งข้อมูล

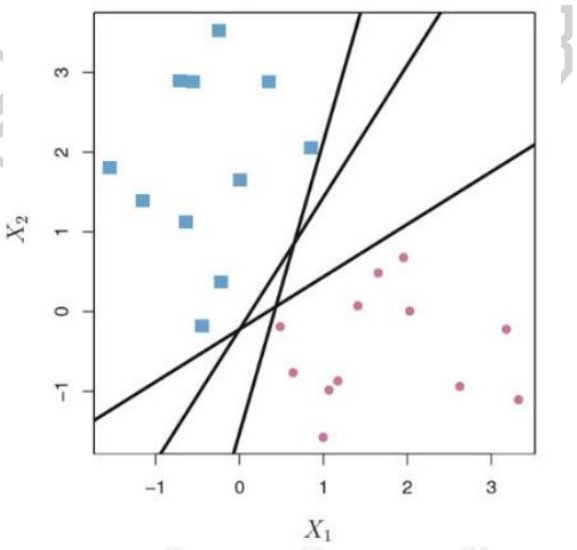
ตัวอย่างการใช้ไฮเปอร์เพลนในการแบ่งข้อมูล กรณีข้อมูลที่มีตัวแปรอิสระ 2 ตัว และมีจำนวน class เท่ากับ 2 แสดงดังภาพที่ 8



ภาพที่ 8 ตัวอย่างการใช้ไฮเปอร์เพลนในการแบ่งประเภทข้อมูล กรณีที่มีตัวแปรอิสระ 2 ตัวและมี 2 class ซึ่งแต่ละฝั่งของไฮเปอร์เพลนที่ใช้ในการแบ่งข้อมูลจะแทนแต่ละ class ของข้อมูล

ที่มา : (Jingfen Han et al., 2012)

จะเห็นได้ว่า หากข้อมูลสามารถแบ่งได้ด้วยไฮเปอร์เพลนแล้ว ไฮเปอร์เพลนที่ใช้ในการแบ่งข้อมูลจะมีจำนวนเป็นอนันต์ ดังภาพที่ 9

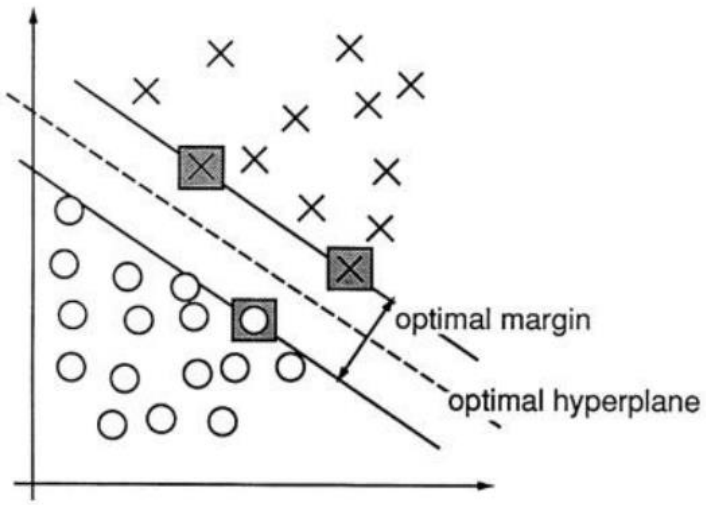


ภาพที่ 9 ตัวอย่างไฮเปอร์เพลนที่สามารถแบ่งประเภทข้อมูลได้

ที่มา : (Jingfen Han et al., 2012)

เวกเตอร์ที่มีผลต่อไฮเปอร์เพลนที่ใช้แบ่งข้อมูล เรียกว่าซัพพอร์ตเวกเตอร์ (support vector) โดยกรณีที่สามารถแบ่งข้อมูลได้ด้วยไฮเปอร์เพลน ซัพพอร์ตเวกเตอร์นั้นจะหมายถึงเวกเตอร์ของข้อมูลที่อยู่ใกล้กับไฮเปอร์เพลนมากที่สุด ซึ่งในแต่ละ class จะมีไฮเปอร์เพลนที่ตัดผ่านซัพพอร์ตเวกเตอร์และลากขนานกับไฮเปอร์เพลนเพียงแค่ 1 เส้น ระยะห่างระหว่างไฮเปอร์เพลนที่ตัดผ่านซัพพอร์ตเวกเตอร์ของ class หนึ่งกับอีก class หนึ่งนั้น เรียกว่า มาร์จิน (margin)

ไฮเปอร์เพลนที่สามารถแบ่งข้อมูลออกจากกันได้ดีที่สุด (optimal hyperplane) คือ ไฮเปอร์เพลนที่มีมาร์จินกว้างที่สุด (Cortes & Vapnik, 1995) เรียกว่า maximal margin hyperplane (หรือ optimal separating hyperplane) และเรียกตัวจำแนกประเภทนี้ว่า maximal margin classifier



ภาพที่ 10 ตัวอย่างของ maximal margin classifier โดยที่ optimal hyperplane แสดงด้วยเส้นประ ซัพพอร์ตเวกเตอร์แสดงด้วยข้อมูลที่อยู่ในกรอบสี่เหลี่ยมสีเทา ระยะห่างระหว่างไฮเปอร์เพลนที่ทับ ซัพพอร์ตเวกเตอร์ (เส้นทึบ) ของแต่ละ class คือ มาร์จิน แสดงด้วยลูกศร

ที่มา : (Cortes & Vapnik, 1995)

หากกำหนดให้ความกว้างของมาร์จินยาวเท่ากับ 2M แล้ว maximal margin classifier คือ $f(x) = \beta_0 + x^T\beta$ โดยพารามิเตอร์ β_0 และ β หาได้จากการแก้ปัญหาค่าสูงสุด ดังนี้

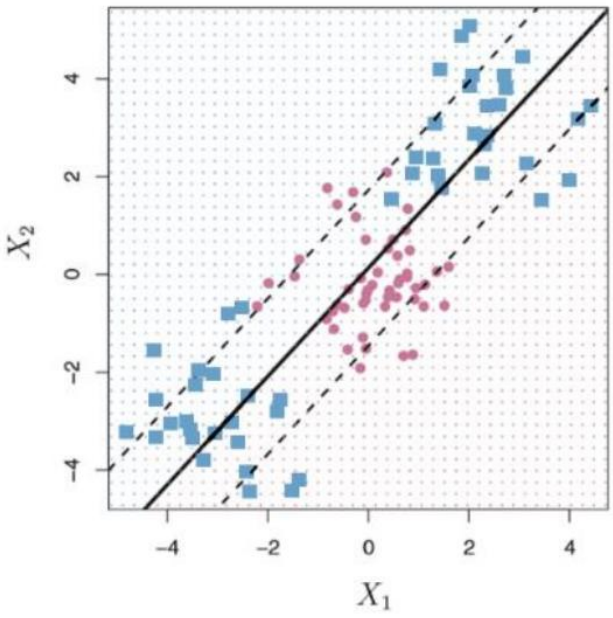
$$\max_{\beta_0,\beta} M$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 = 1\,, \qquad (12)$$

$$y_i(\beta_0 + x_i{}^T\beta) \geq M \;,\; \forall_i = 1,\ldots,n$$

ข้อจำกัด $y_i(\beta_0 + x_i{}^T\beta) \geq M$ จะเป็นการบังคับว่า ทุกเวกเตอร์จะต้องอยู่นอกบริเวณของมาร์จินใน class ที่ถูกต้อง

กรณีที่ข้อมูลไม่สามารถแบ่งได้ด้วยไฮเปอร์เพลนอย่างสมบูรณ์ การใช้ไฮเปอร์เพลนแบ่งประเภทโดยที่มีข้อมูลบางเวกเตอร์ไม่อยู่ใน class ที่ถูกต้อง แสดงดังภาพที่ 11 จะเรียกตัวจำแนกประเภทนี้ว่า support vector classifier หรือ soft margin classifier



ภาพที่ 11 ตัวอย่างของ support vector classifier

ที่มา : (Jingfen Han et al., 2012)

เมื่อไม่สามารถแบ่งประเภทของข้อมูลได้อย่างสมบูรณ์ จึงมีข้อมูลบางเวกเตอร์ที่ไม่ได้อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้อง กล่าวคือ อาจอยู่ในบริเวณมาร์จิน แต่อยู่ใน class ที่ถูกต้องหรือไม่

ถูกต้องก็ได้ หรืออาจจะอยู่บริเวณมาร์จินและอยู่ใน class ที่ไม่ถูกต้องก็ได้เช่นกัน เมื่อกำหนดให้ x_i^* เป็นเวกเตอร์ที่ไม่ได้อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้อง กำหนด $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ เมื่อ ξ_i คือ ระยะห่างจาก x_i^* ไปยังไฮเปอร์เพลนที่เป็นขอบของมาร์จินใน class ของ x_i^* เมื่อเทียบกับความกว้างของมาร์จิน ซึ่งพารามิเตอร์ของ support vector classifier สามารถหาได้จากการแก้ปัญหาค่าสูงสุด ดังนี้

$$\max_{\beta_0,\beta,\xi} M$$

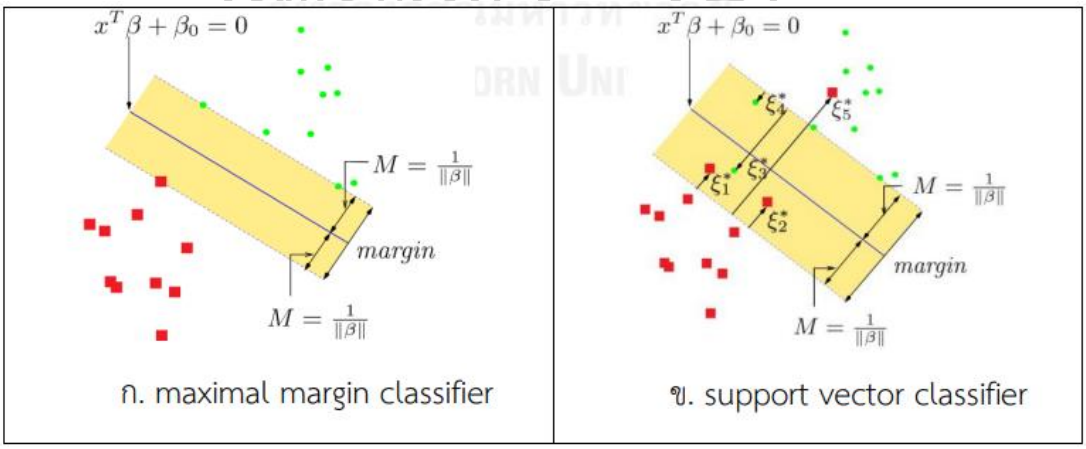
$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \,,$$

$$y_i(\beta_0 + x_i{}^T\beta) \geq M(1-\xi_i) \quad , \forall_i = 1,\ldots,n \qquad (13)$$

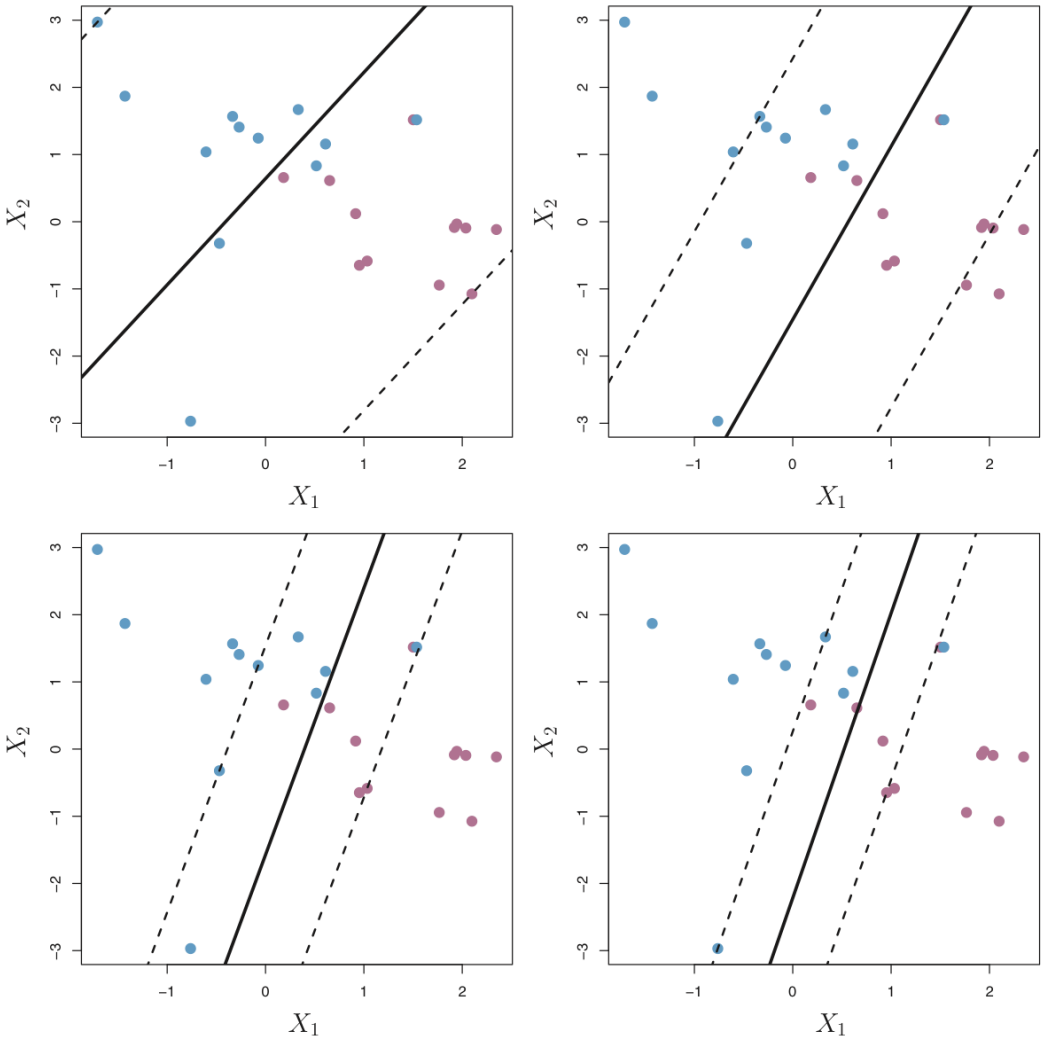
$$\sum_{i=1}^{n} \xi_i \leq C \,,$$

$$\xi_i \geq 0 \quad , \forall_i = 1,\ldots,n$$

เมื่อ C หรือพารามิเตอร์ที่ใช้กำหนดระดับของการไม่ได้อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้องที่รับได้ นั่นคือ ยิ่ง C มีค่ามาก มาร์จินก็จะยิ่งกว้าง ในกรณีที่ C > 0 และเป็นจำนวนเต็ม จะหมายถึงจำนวนสูงสุดของเวกเตอร์ที่ support vector classifier อนุญาตให้อยู่ใน class ที่ไม่ถูกต้องได้ กรณีที่ C = 0 support vector classifier จะเทียบเท่ากับ maximal margin classifier



ภาพที่ 12 เปรียบเทียบตัวแบ่งประเภท maximal margin classifier กับ support vector classifier
ที่มา : (Hastie, Tibshirani, & Friedman, 2008)



ภาพที่ 13 ตัวแบ่งประเภท support vector classifier ที่กำหนดพารามิเตอร์ C ที่แตกต่างกัน เมื่อ C มีค่ามาก จะมีความคลาดเคลื่อนสูงสำหรับข้อมูลที่อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้อง ดังนั้นระยะขอบจะกว้าง และเมื่อ C มีค่าน้อย จะทำให้มีข้อมูลที่อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้องลดลง ดังนั้นระยะขอบจะแคบ

ที่มา : (James et al., 2017)

สำหรับ support vector classifier ซัพพอร์ตเวกเตอร์จะหมายถึงเวกเตอร์ทั้งหมดที่มิได้อยู่นอกบริเวณมาร์จินใน class ที่ถูกต้อง

สำหรับข้อมูลที่มีตัวแปรอิสระทั้งหมด p ตัว และมีข้อมูลจำนวน n คู่ เมื่อกำหนดให้ $\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$ แล้ว support vector classifier ซึ่งเป็นตัวจำแนกประเภทเชิงเส้น สามารถเขียนได้ในรูป

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i y_i \langle x, x_i \rangle \qquad (14)$$

โดย $x_i = \left(X_{i1}, X_{i2}, \ldots, X_{ip}\right)^T$ คือ เวกเตอร์ของข้อมูลชุดฝึกสอนที่ใช้สร้างตัวแบบตัวที่ i

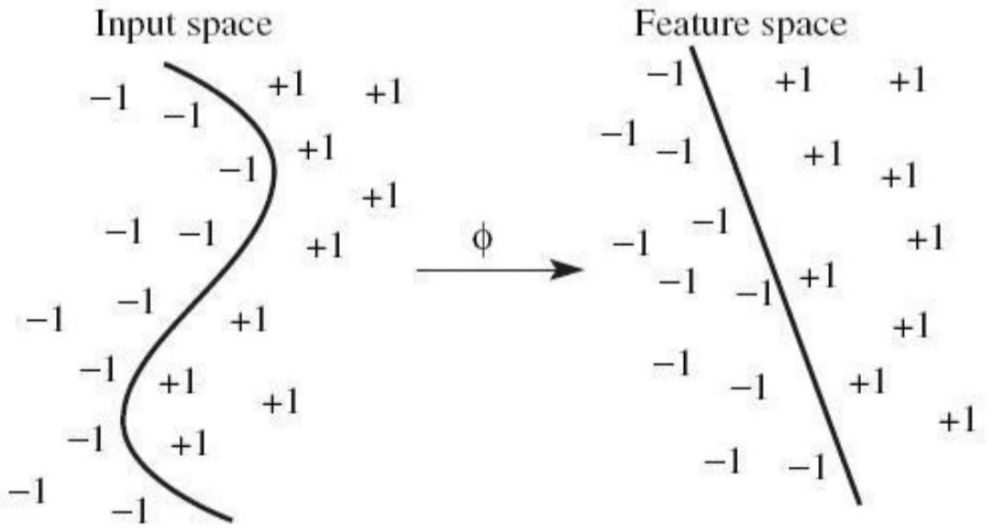
$x = \left(X_i, X_2, \ldots, X_p\right)^T$ คือ เวกเตอร์ของข้อมูลชุดทดสอบที่ใช้ทดสอบประสิทธิภาพตัวแบบสร้างตัวแบบ

$\langle x_i, x_{i'} \rangle$ คือ ผลคูณภายใน (inner product) ของข้อมูล x_i และ $x_{i'}$ กล่าวคือ

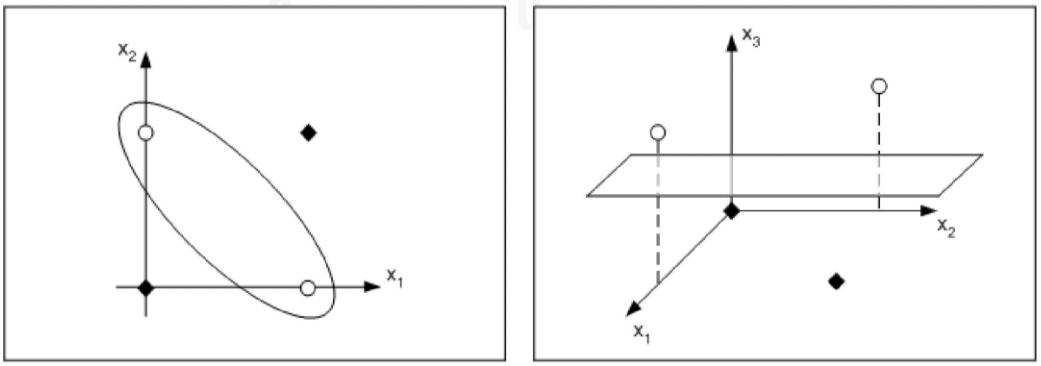
$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

ซึ่งการประมาณค่าพารามิเตอร์ $\alpha_1, \ldots, \alpha_n$ และ β_0 จำเป็นต้องคำนวณผลคูณภายในของข้อมูลที่ใช้สร้างตัวแบบทุกคู่

ในกรณีส่วนใหญ่ การหาไฮเปอร์เพลนเชิงเส้นที่เหมาะสมไม่สามารถทำได้ จึงต้องใช้เทคนิคการแมปข้อมูลตัวอย่างไปยังปริภูมิอันดับสูงโดยใช้ฟังก์ชันการแมป (Mapping Function : ϕ) แล้วจึงค่อยทำการฝึกและจำแนกกลุ่มข้อมูล (ดังภาพที่ 13 และ 14)



ภาพที่ 14 การแมปข้อมูลจากปริภูมินำเข้าไปยังมิติปริภูมิอันดับสูง



(ก) (ข)

ภาพที่ 15 ตัวอย่างระนาบหลายมิติสำหรับแบ่งแยก (ก) ปริภูมินำเข้า (ข) มิติปริภูมิอันดับสูง

คุณสมบัติที่ดีอีกประการของซัพพอร์ตเวกเตอร์แมชชีน คือ ไม่จำเป็นต้องรู้รูปแบบที่ชัดเจนของ ϕ โดยสิ่งที่ต้องพิจารณาคือ นิยามผลคูณภายในมิติปริภูมิอันดับสูง ซึ่งเรียกว่า ฟังก์ชันเคอร์เนล (Kernel function) ฟังก์ชันเคอร์เนล $K(x_i, x_{i,})$ เป็นฟังก์ชันที่แก้ปัญหาภายใต้เงื่อนไขของ Mercer's ซึ่งมีค่าเท่ากับการคูณกันของสองเวกเตอร์ x_i และ $x_{i,}$ ในพื้นที่คุณลักษณะ $\phi(x_i)$ และ $\phi(x_{i,})$ ดังนั้น เคอร์เนลเชิงเส้น (Linear Kernel) สามารถเขียนได้ด้วยสมการ ดังนี้

$$K(x_i, x_{i,}) = \phi(x_i) \cdot \phi(x_{i,}) \quad (15)$$

โดยที่ ϕ คือ ฟังก์ชันการแปลงแบบไม่เป็นเชิงเส้น (Nonlinear Projection Function) ซึ่งฟังก์ชันเคอร์เนลหลายตัวได้ถูกนำมาใช้กับซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เป็นเชิงเส้น ซึ่งจะได้สมการของซัพพอร์ตเวกเตอร์แมชชีนอยู่ในรูป

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i y_i K\langle x, x_i \rangle \quad (16)$$

เมื่อ $K\langle x, x_i \rangle$ คือ ฟังก์ชันเคอร์เนล โดยฟังก์ชันเคอร์เนลที่เป็นที่นิยม ได้แก่

1. เคอร์เนลพหุนาม (polynomial kernel)

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \quad (17)$$

เมื่อ d คือ จำนวนเต็มบวกใด ๆ ที่แสดงถึงอันดับของพหุนาม หาก d = 1 จะกลายเป็นเคอร์เนลเชิงเส้น (linear kernel) ซึ่งหมายถึง support vector classifier

2. เคอร์เนลเรเดียล (radial basis function kernel) หรือเคอร์เนลเกาส์เซียน (Gaussian kernel)

$$K(x_i, x_{i'}) = \exp\left(\frac{\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2}{2\sigma^2}\right) \quad (18)$$

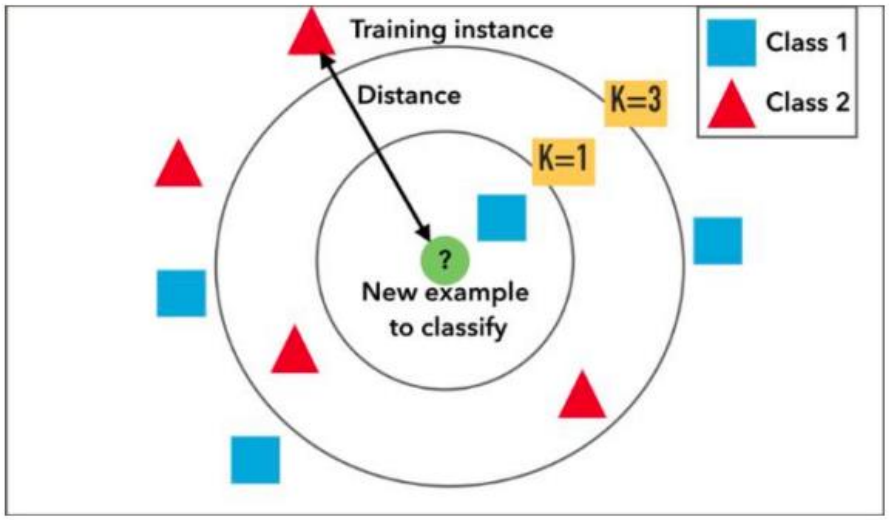
โดยที่ σ เป็นพารามิเตอร์อิสระ โดยทั่วไปแล้ว จะกำหนดให้ $\gamma = \frac{1}{2\sigma^2}$ ซึ่งจะเขียนเคอร์เนลใหม่ได้เป็น

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2) \text{ สำหรับ } \forall\gamma > 1 \quad (19)$$

เพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor)

เพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor : K-NN) (Jingfen Han et al., 2012) โดยหลักการของวิธีนี้จะจำแนกประเภทของข้อมูลโดยขึ้นกับข้อมูลที่มีคุณลักษณะใกล้เคียงกันมากที่สุด K ตัวจากชุดข้อมูลตัวอย่าง ซึ่งจะวัดจากระยะทางที่น้อยที่สุดระหว่างสมาชิกใหม่หรือข้อมูลที่ป้อนถามกับชุดข้อมูลตัวอย่างฝึกฝน และจะคำนวณหาเพื่อนบ้านใกล้ที่สุด K ตัว หลังจากนั้นจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัว แล้วเลือกคลาสที่มีจำนวนหรือความถี่มากที่สุดใน K ตัวดังกล่าวให้กับสมาชิกใหม่

ข้อมูลการจำแนก โดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยคุณลักษณะหลายตัวแปร X_i ซึ่งจะนำมาใช้ในการจำแนกประเภท Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ K-NN ได้แก่ 1-NN, 2-NN, 3-NN, ... , K-NN ซึ่งตัวอย่าง 3-NN หมายถึง อัลกอริทึมเพื่อนบ้านใกล้ที่สุดจะค้นหา 3 กรณีที่มีคุณลักษณะที่ใกล้เคียงกับกรณีใหม่ (3 nearest cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝนมาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (distance) ใกล้เคียงที่สุดมา K ตัวโดยใช้การวัดระยะทางระหว่างสองวัตถุ ถ้าหากวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายคลึงกันน้อย และถ้าวัตถุห่างกันน้อยแสดงว่ามีความคล้ายคลึงกันมาก



ภาพที่ 16 การจำแนกประเภทด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด

ตัวอย่างการจำแนกข้อมูลโดยเทคนิคเพื่อนบ้านใกล้ที่สุด จากภาพที่ 16 เป็นการหาว่าจุดสีเขียว (new example) เป็น class 1 หรือ class 2 ด้วยการกำหนดค่า k และเมื่อกำหนด k=1 จะทำนายว่าเป็น class 1 เพราะระยะห่างของข้อมูลที่ใกล้เคียงจุดสีเขียวที่สุด 1 ตัว คือสี่เหลี่ยมสีฟ้า (class 1) แต่เมื่อกำหนด k=3 จะทำนายว่าเป็น class 2 เพราะระยะห่างของข้อมูลที่ใกล้เคียงจุดสีเขียวที่สุด 3 ตัวโดยเรียงจากมากไปน้อย คือ สี่เหลี่ยมสีฟ้า (class 1) 1 รูป และมีสามเหลี่ยมสีแดง (class 2) 2 รูป โดยวิธีการคำนวณระยะห่างระหว่างข้อมูลที่นิยมใช้ เช่น ยูคลิเดียน (Euclidean distance) และแมนฮัตตัน (Manhattan distance) เป็นต้น

วิธีการคำนวณระยะทาง

การวัดระยะทางยูคลิเดียน (Euclidean distance) เป็นวิธีที่นิยมใช้มากที่สุด ซึ่งเป็นการวัดค่าความห่างระหว่างข้อมูล 2 ข้อมูล ในระบบพิกัดคาร์ทีเซียน ที่มาจากทฤษฎีปีทาโกรัส ซึ่งถ้าข้อมูล 2 ตัวมีความคล้ายกันมาก แสดงว่าข้อมูลแต่ละตัวจะอยู่ใกล้กันมาก ซึ่งทำให้ค่า Euclidean มีค่าน้อยจนเข้าใกล้ศูนย์ คำนวณได้จากสมการ ดังนี้

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \qquad (20)$$

โดยที่ $\mathrm{dist}(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับ ตัวอย่าง x_j
d คือ จำนวนคุณลักษณะทั้งหมดของตัวอย่าง
$x_{i,k}$ คือ คุณลักษณะตัวที่ k ของตัวอย่าง x_i
$x_{j,k}$ คือ คุณลักษณะตัวที่ k ของตัวอย่าง x_j

การวัดระยะทางแมนฮัตตัน (Manhattan distance) เป็นการวัดระยะห่างโดยวัดตามระบบพิกัดฉาก X และ Y ระหว่างข้อมูลทั้งสอง หรือข้อมูลศูนย์กลางของพื้นที่ คำนวณได้ตามสมการ ดังนี้

$$\mathrm{dist}(x_i, x_j) = \sum_{k=1}^{d}|x_{i,k} - x_{j,k}| \qquad (21)$$

โดยที่ $\mathrm{dist}(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับ ตัวอย่าง x_j
d คือ จำนวนคุณลักษณะทั้งหมดของตัวอย่าง
$x_{i,k}$ คือ พิกัดตำแหน่งของจุดหรือศูนย์กลางของพื้นที่ x_i
$x_{j,k}$ คือ พิกัดตำแหน่งของจุดหรือศูนย์กลางของพื้นที่ x_j

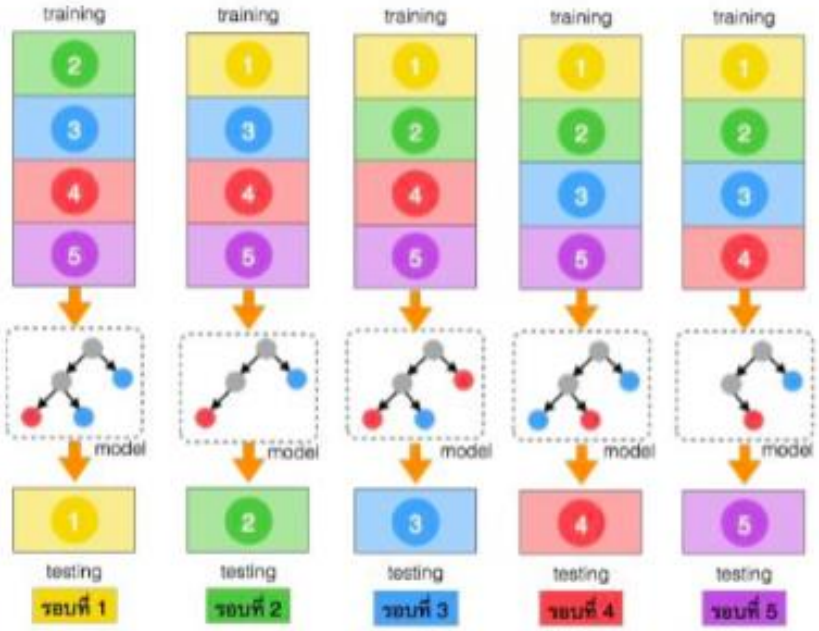
การประเมินประสิทธิภาพแบบจำลอง

การประเมินแบบจำลองสำหรับการจำแนกประเภทข้อมูล เป็นขั้นตอนที่มีความสำคัญ เพื่อให้ทราบถึงประสิทธิภาพของแบบจำลองที่ถูกสร้างขึ้นมารวมทั้งสามารถเปรียบเทียบแบบจำลองเพื่อเลือกแบบจำลองที่ดีที่สุดในการนำไปใช้ประมวลผลข้อมูล โดยมีวิธีการวัดประสิทธิภาพของแบบจำลอง ดังนี้

1. การประเมินประสิทธิภาพด้วยวิธีการตรวจสอบไขว้แบบ k รอบ (k-fold cross validation)

วิธีการตรวจสอบไขว้แบบ k รอบ (k-fold cross validation) เป็นวิธีที่นิยมใช้ในการทดสอบประสิทธิภาพของแบบจำลอง ซึ่งผลที่ได้มีความน่าเชื่อถือ เนื่องจากข้อมูลทุกชุดจะมีโอกาสได้เป็นชุดสอนและชุดทดสอบ โดยการวัดประสิทธิภาพด้วยวิธี cross validation นี้ จะทำการแบ่งข้อมูลออกเป็นหลายส่วนเท่า ๆ กัน เพื่อใช้เป็นข้อมูลชุดสอน (training set) และข้อมูลชุดทดสอบ (test set) เช่น 5-fold cross validation โดยข้อมูลทั้งหมดจะถูกแบ่งเป็น 5 ส่วนเท่า ๆ กัน ใช้ 4 ส่วน เป็นชุดสอน และ 1 ส่วน เป็นชุดทดสอบ ซึ่งจะทำเช่นนี้ไปเรื่อย ๆ จนครบทั้ง 5 ส่วน แสดงดังภาพที่ 16 จากนั้นประเมินประสิทธิภาพด้วยตัววัดประสิทธิภาพของแบบจำลองการจำแนกประเภท

ข้อดีของวิธีการนี้ คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด k-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดขนาดของข้อมูลและจำนวนรอบในการทดสอบได้เหมาะสำหรับการประมวลผลในการทดสอบข้อมูลที่มีขนาดมิติจำนวนมาก



ภาพที่ 17 ตัวอย่างการแบ่งข้อมูลแบบ 5-fold cross validation
ที่มา: (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

2. การประเมินประสิทธิภาพด้วยตารางสรุปผลลัพธ์การทำนาย (Confusion matrix)

การวัดประสิทธิภาพด้วยตารางสรุปผลลัพธ์การทำนาย (Confusion matrix) เป็นวิธีการแทนจำนวนของข้อมูลในแต่ละประเภทที่ถูกจำแนกโดยแบบจำลองด้วยเทคนิคต่าง ๆ ลงในตารางที่ 6

ตารางที่ 6 ตารางสรุปผลลัพธ์การทำนาย (Confusion Matrix)

		Actual values (ผลลัพธ์จริง)	
		Yes	No
Predicted values (ผลลัพธ์การทำนาย)	Yes	TP	FP
	No	FN	TN

จากตารางที่ 6 ค่าที่แสดงในตารางสรุปผลลัพธ์การทำนาย ประกอบด้วย
True Positive (TP) คือ จำนวนตัวอย่างที่ผลลัพธ์การทำนายและผลลัพธ์จริงเป็นจริง
True Negative (TN) คือ จำนวนตัวอย่างที่ผลลัพธ์การทำนายและผลลัพธ์จริงเป็นเท็จ
False Positive (FP) คือ จำนวนตัวอย่างที่ผลลัพธ์การทำนายเป็นจริง แต่ผลลัพธ์จริงเป็นเท็จ
False Positive (FN) คือ จำนวนตัวอย่างที่ผลลัพธ์การทำนายเป็นเท็จ แต่ผลลัพธ์จริงเป็นจริง

ตารางสรุปผลลัพธ์การทำนายสำหรับการจำแนกประเภทแบบไบนารี ค่าบนเส้นทแยงมุมหมายถึงจำนวนตัวอย่างที่แบบจำลองทำนายได้ถูกต้อง ได้แก่ จำนวนตัวอย่างที่ผลลัพธ์การทำนายและผลลัพธ์จริงเป็นจริง (TP) และจำนวนตัวอย่างที่ผลลัพธ์การทำนายและผลลัพธ์จริงเป็นเท็จ (TN) ส่วนค่าที่อยู่นอกเส้นทแยงมุม หมายถึงจำนวนตัวอย่างที่แบบจำลองทำนายผิด ได้แก่ จำนวนตัวอย่างที่ผลลัพธ์การทำนายเป็นเท็จ แต่ผลลัพธ์จริงเป็นจริง (FN) และจำนวนตัวอย่างที่ผลลัพธ์การทำนายเป็นจริง แต่ผลลัพธ์จริงเป็นเท็จ (FP)

ในการประเมินประสิทธิภาพจะพิจารณาจากหลาย ๆ ค่าประกอบกัน หากพิจารณาค่าความถูกต้องเพียงอย่างเดียว อาจทำให้การประเมินบางส่วนผิดพลาดไป สำหรับชุดข้อมูลที่ไม่สมดุลกัน (Powers, 2011) ดังนั้นจึงได้ใช้ตัววัดประสิทธิภาพอื่น ๆ จากตารางสรุปผลลัพธ์การทำนาย (Confusion matrix) เช่น ค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 มาช่วยในการประเมินประสิทธิภาพ เพื่อช่วยลดความผิดพลาดในการประเมิน โดยมีรายละเอียดดังนี้

- ค่าความถูกต้อง (Accuracy)

ค่าความถูกต้องเป็นการประเมินแบบหนึ่งที่นิยมใช้มากที่สุดสำหรับการประเมินประสิทธิภาพการจำแนกประเภท โดยคำนวณจากผลรวมของตัวเลขบนเส้นทแยงมุมในตารางสรุปผลลัพธ์การทำนายหารด้วยจำนวนตัวอย่างทั้งหมด เมื่อค่าความถูกต้องสูง กล่าวคือ ค่าการทำนายนั้นสามารถทำนายได้ถูกต้องใกล้เคียงกับค่าจริง ดังสมการที่ 22

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (22)$$

- ค่าความเที่ยง (Precision)

ค่าความเที่ยงหรือค่าทำนายผลบวก (Positive Predictive Value: PPV) คือ อัตราส่วนระหว่างผลลัพธ์การทำนายและผลลัพธ์จริงเป็นจริง (TP) ต่อจำนวนตัวอย่างที่ผลลัพธ์ทำนายเป็นจริงทั้งหมด เมื่อค่าความเที่ยงสูง กล่าวคือ ค่าการทำนายนั้นสามารถทำนายได้แม่นยำใกล้เคียงกับค่าจริง ดังสมการที่ 23

$$\text{PPV} = \text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (23)$$

- ค่าความครบถ้วน (Recall)

ค่าความครบถ้วนหรืออัตราผลบวกจริง (True Positive Rate: TPR) คือจำนวนตัวอย่างที่เป็นบวกที่ทำนายได้ถูกต้องเทียบกับจำนวนตัวอย่างที่เป็นบวกทั้งหมด เมื่อค่าความครบถ้วนสูง กล่าวคือ ค่าการทำนายนั้นสามารถทำนายได้อย่างครบถ้วนใกล้เคียงกับค่าจริง ดังสมการที่ 24

$$TPR = \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (24)$$

- คะแนน F1 (F1-Score)

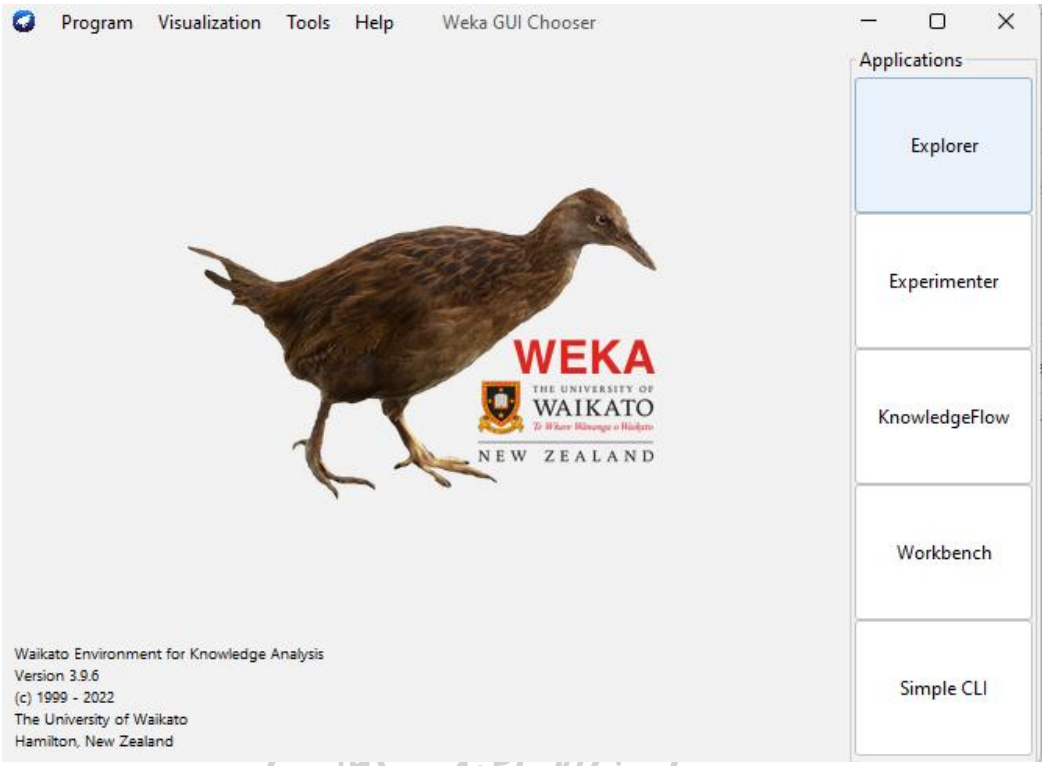
คะแนน F1 คือค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ของ Precision และ Recall เมื่อคะแนน F1 มีค่าสูง หมายความว่ามีประสิทธิภาพในการจำแนกประเภทสูง แสดงดังสมการที่ 25

$$\text{F1} - \text{Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}\right) \quad (25)$$

เครื่องมือที่ใช้ในการวิจัย

โปรแกรม Weka (Waikato Environment for Knowledge Analysis) เริ่มพัฒนามาตั้งแต่ปี 1997 โดยมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ โดยโปรแกรม Weka เป็นซอฟต์แวร์สำเร็จประเภทฟรีแวร์อยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม Weka ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด เขียนขึ้นมาโดยเน้นกับงานทางด้าน Machine learning โปรแกรมจะประกอบไปด้วยโมดูลย่อย ๆ สำหรับใช้ในการจัดการข้อมูลและเป็นโปรแกรมที่สามารถใช้ Graphic User Interface (GUI) และใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผล และสามารถรัน (run) ได้หลายระบบปฏิบัติการ รวมไปถึงการพัฒนาต่อยอดโปรแกรมได้ เป็นเครื่องมือที่ใช้ทำงานในด้าน Machine learning ที่รวบรวมแนวคิดอัลกอริทึมมากมาย ซึ่งอัลกอริทึมสามารถเลือกใช้งานโดยตรงได้จาก 2 ทาง คือ จากชุดเครื่องมือที่มีอัลกอริทึมมาให้ หรือเลือกใช้จากอัลกอริทึมที่ได้เขียนเป็นโปรแกรมลงไปเป็นชุดเครื่องมือเพิ่มเติมและชุดเครื่องมือที่มีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูล ได้แก่ Pre-Processing, Classification, Regression, Clustering, Association rules, Selection และ Visualization (มณีรัตน์ ภารนันท์, 2555)

ในงานวิจัยนี้ ผู้วิจัยเลือกใช้โปรแกรม Weka ในการสร้างแบบจำลอง เนื่องจากเป็นซอฟต์แวร์ที่สามารถดาวน์โหลดและติดตั้งได้ฟรี สามารถทำงานได้ทุกระบบปฏิบัติการ ง่ายต่อการใช้งานในการสร้างแบบจำลองด้วยเทคนิคการจำแนกข้อมูล (Classification) และสามารถกำหนดค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ในแต่ละเทคนิคการจำแนกได้ โดยมีอัลกอริทึมที่ผู้วิจัยสนใจครบถ้วนให้เลือกใช้งาน นอกจากนี้ยังมีการนำเสนอข้อมูลด้วยรูปภาพ (Visualization) เพื่อแสดงผลการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจ (Decision tree)

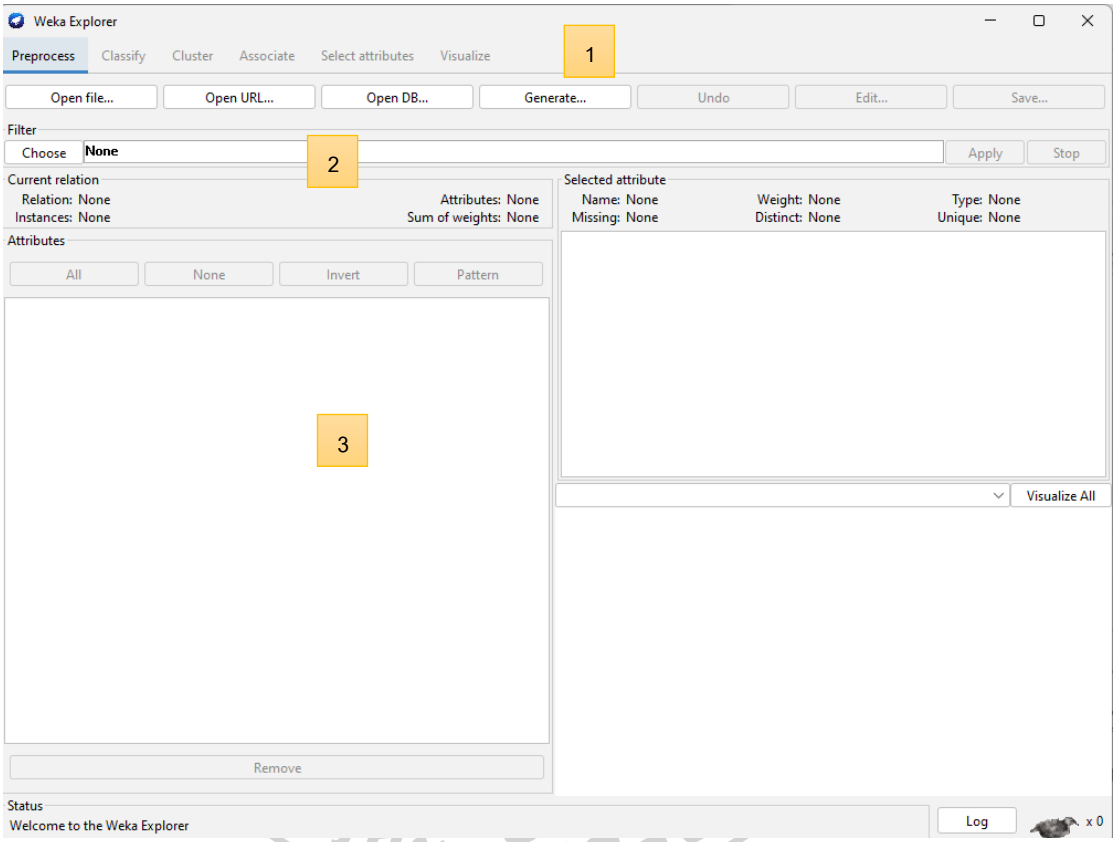


ภาพที่ 18 หน้าจอหลักโปรแกรม Weka

จากภาพที่ 18 หน้าจอหลักโปรแกรม Weka มี 2 ส่วน คือ ส่วนแรกเป็นแถบเครื่องมือ และส่วนที่สองเป็นโปรแกรมประยุกต์ ซึ่งจะแบ่งเป็น 4 เมนูหลัก รายละเอียดของเมนูต่าง ๆ มีดังนี้

1. เมนู Explorer เป็นส่วนที่เหมาะสำหรับผู้เริ่มต้นใช้งาน โดยสามารถเรียกใช้ฟังก์ชันการทำงานต่าง ๆ ของโปรแกรมผ่านทางหน้าจอ GUI และขั้นตอนการใช้งานที่ง่าย เช่น สามารถดูค่าทางสถิติของพารามิเตอร์ในชุดข้อมูลได้ทางหน้าจอ เป็นต้น
2. เมนู Experimenter เป็นส่วนที่ผู้ใช้งานสามารถทดลองปรับเปลี่ยนค่าไฮเปอร์พารามิเตอร์ต่าง ๆ เพื่อค้นหาคำตอบที่ดีที่สุดสำหรับคำถามหรือปัญหาที่ต้องการ โดยสามารถทำการเปรียบเทียบผลลัพธ์ที่ได้ในแต่ละเทคนิคได้ เช่น เปรียบเทียบผลลัพธ์ระหว่างเทคนิคการจำแนกข้อมูลกับเทคนิคการถดถอย
3. เมนู Knowledge Flow เป็นส่วนที่ผู้ใช้สามารถออกแบบการไหลของข้อมูลร่วมกับเทคนิค Machine learning ในส่วนที่เมนู Explorer ไม่ได้จัดเตรียมไว้ การออกแบบจะเป็นในลักษณะ drag and drop เครื่องมือ (Computers) และนำมาผูกกันเป็นกระบวนการ (Process) ทำงานอัตโนมัติได้

4. เมนู Simple CLI เป็นส่วนที่ผู้ใช้สามารถเรียกใช้ฟังก์ชันการทำงานผ่านทาง Command line ได้ ซึ่งช่วยให้ผู้ใช้สามารถเข้าใจการทำงานของฟังก์ชันที่อยู่เบื้องหลังหน้าจอ GUI และยังสามารถนำไปประยุกต์ใช้ในการเขียนโปรแกรมเพื่อเรียกใช้งานฟังก์ชันจากโปรแกรม Weka ได้อีกด้วย



ภาพที่ 19 หน้าหลักในการทำงานของโปรแกรม Weka

จากภาพที่ 19 แสดงส่วนประกอบของหน้าจอหลักการทำงานของโปรแกรม Weka ดังนี้

ส่วนที่ 1 ส่วนบนสุดเป็นแท็บ (tab) ซึ่งมีทั้งหมด 6 แท็บ เป็นเมนูให้ผู้ใช้สามารถใช้งานเทคนิคต่าง ๆ ของ Weka ได้

ส่วนที่ 2 ส่วนที่อยู่ตรงกลางซึ่งจะเปลี่ยนไปตามการกดเมนูการใช้งานต่าง ๆ เป็นส่วนของการเลือก option ในการวิเคราะห์ข้อมูลและส่วนการแสดงผลลัพธ์หลังจากทำการวิเคราะห์ข้อมูล

ส่วนที่ 3 ส่วนที่อยู่ด้านล่างสุด จะเป็นส่วนที่บอกสถานะของการทำงานในแต่ละขั้นตอน

การทำงานของเมนูต่าง ๆ ในโปรแกรม Weka

1. Preprocess เป็นส่วนที่ใช้ในการเลือกไฟล์ข้อมูลสำหรับเป็นการนำเข้าข้อมูล (input) เพื่อการวิเคราะห์ข้อมูล
2. Classify ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการจำแนกข้อมูล (classification) หรือทำนายข้อมูล (prediction) ซึ่งจะมีอัลกอริทึมต่าง ๆ ให้เลือกมากมาย
3. Cluster เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการจัดกลุ่มข้อมูล (clustering) โดยจะจัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันหรือมีความสัมพันธ์กันเข้าไว้ด้วยกัน
4. Associate เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูลด้วยวิธีการหาความสัมพันธ์ของข้อมูล
5. Select attributes เป็นส่วนที่คล้ายกับส่วน Preprocess แต่จะเน้นที่การหาว่าตัวแปรไหนที่สำคัญและไม่สำคัญในชุดข้อมูล ซึ่งตัวแปรที่ไม่สำคัญนี้จะถูกกำจัดทิ้งไปก่อนที่จะวิเคราะห์ข้อมูลด้วยอัลกอริทึมต่าง ๆ
6. Visualize เป็นส่วนของการ plot จุดข้อมูลในรูปแบบ 2 มิติ

เมื่อพิจารณาแบบจำลองในการจำแนกประเภทจากงานวิจัยที่เกี่ยวข้องจำนวน 14 งานวิจัย สามารถสรุปภาพรวมของแบบจำลองในการจำแนกประเภทได้ดังนี้

ตารางที่ 7 ภาพรวมเทคนิคการจำแนกประเภทจากงานวิจัยที่เกี่ยวข้อง

ลำดับ	เทคนิคการจำแนกประเภท	ความถี่														รวม
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	Decision tree	✓	✓	✓	✓	✓	✓					✓	✓	✓		9
2	Support Vector Machine		✓			✓	✓	✓	✓	✓	✓					7
3	Random forest		✓		✓		✓				✓		✓	✓	✓	7
4	K-Nearest neighbor	✓	✓			✓	✓			✓			✓	✓		7
5	Naïve Bayes	✓		✓			✓					✓	✓	✓		6
6	Deep learning												✓			1

ที่มา : จากการรวบรวมของผู้วิจัย

จากการศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง พบว่าเทคนิค Decision tree, เทคนิค Support Vector Machine, เทคนิค Random forest และเทคนิค K-Nearest neighbor เป็นเทคนิคที่เป็นที่นิยมและมีประสิทธิภาพการจำแนกที่ดี ดังนั้นผู้วิจัยจึงตัดสินใจเลือกเทคนิคดังกล่าวมาใช้ในการสร้างแบบจำลองสำหรับการจำแนกการเป็นโรคเบาหวาน

งานวิจัยที่เกี่ยวข้อง

(Nai-arun & Sittidech, 2014) เสนอเทคนิคเหมืองข้อมูลเพื่อปรับปรุงประสิทธิภาพและความน่าเชื่อถือในการจำแนกโรคเบาหวาน จากข้อมูลจำนวน 48,763 ราย และ 16 ตัวแปรที่รวบรวมจากโรงพยาบาลศูนย์สวรรค์ประชารักษ์ ประเทศไทย โดยใช้เทคนิค Gain ratio ในการคัดเลือกคุณลักษณะจากทั้งหมด 15 ตัวแปร เหลือ 13 ตัวแปรที่นำมาใช้ในการสร้างแบบจำลอง จากนั้นใช้เทคนิค Boosting และ Bagging ร่วมกับอัลกอริทึม Naïve bayes, K-Nearest neighbors และ Decision tree ในการสร้างแบบจำลองการจำแนก ซึ่งผลการศึกษาพบว่า แบบจำลองจากเทคนิค Bagging ร่วมกับอัลกอริทึม Decision tree มีความแม่นยำสูงสุด (95.31%)

(Kandhasam & Balamurali, 2015) ศึกษาการวิเคราะห์ประสิทธิภาพของแบบจำลองในการทำนายโรคเบาหวาน โดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมที่ใช้ในการทำนายโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล ซึ่งจะเปรียบเทียบวิธีการจำแนกจากตัวแบบ Machine learning ได้แก่ วิธี J48 Decision tree, K-Nearest neighbors, Random forest และ Support Vector Machine เพื่อจำแนกผู้ป่วยโรคเบาหวาน ซึ่งใช้ตัวแปรในการสร้างแบบจำลองทั้งหมด 9 ตัวแปร คือ จำนวนครั้งที่ตั้งครรภ์ ความเข้มข้นของกลูโคสในเลือด ความดันโลหิต ความหนาของผิวหนัง Triceps ระดับอินซูลินใน 2 ชั่วโมง ดัชนีมวลกาย ประวัติโรคเบาหวานของคนในครอบครัว อายุ และประเภทของผู้ป่วย (เป็นเบาหวาน หรือไม่เป็นเบาหวาน) จากผลการศึกษาพบว่าแบบจำลอง J48 Decision tree มีความแม่นยำสูงถึง 73.82% เมื่อเทียบกับแบบจำลองอื่น ๆ แต่เมื่อประมวลผลกับข้อมูลเดิมที่ตัดข้อมูลที่มีความผิดพลาดออก พบว่า ทั้งวิธี KNN(k=1) และ Random forest มีประสิทธิภาพดีกว่าวิธีอื่น ๆ และให้ความแม่นยำ 100% ทำให้ทราบได้ว่าหลังจากการตัดข้อมูลที่มีความผิดพลาดออกจากชุดข้อมูล ส่งผลต่อผลลัพธ์ที่ดีสำหรับความถูกต้องของแบบจำลองในการจำแนกประเภทผู้ป่วยโรคเบาหวาน และจากเทคนิคที่ใช้ในการสร้างแบบจำลองดังกล่าว พบว่ามีประสิทธิภาพในการทำนายสูง

(สิตา ธานี, 2559) ศึกษาการพัฒนาตัวแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่นโดยเทคนิคนาอีฟเบย์และเทคนิคต้นไม้ตัดสินใจ มีวัตถุประสงค์ 1) เพื่อสังเคราะห์แบบจำลองการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่น 2) เพื่อวิเคราะห์รูปแบบการพยากรณ์ความเสี่ยง

การเกิดโรคซึมเศร้าในวัยรุ่น 3) เพื่อพัฒนาตัวแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่น โดยผลการศึกษา พบว่าการสังเคราะห์แบบจำลองการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่นได้ผลลัพธ์เป็นกรอบแนวคิดในการพัฒนาต้นแบบการพยากรณ์ซึ่งแบ่งการทำงานออกเป็น 4 โมดูล ผลการวิเคราะห์รูปแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่นโดยใช้เทคนิคนาอีฟเบย์และเทคนิคต้นไม้ตัดสินใจ ได้ผลลัพธ์เป็นกฎพื้นฐานโดยเทคนิคที่มีผลการทำนายแม่นยำมากที่สุดคือ เทคนิคต้นไม้ตัดสินใจซึ่งมีค่าความถูกต้อง 93.09% นำมาแปลงเป็นกฎพื้นฐานได้ทั้งหมด 64 กฎ ผลการพัฒนาตัวแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่นได้ผลลัพธ์เป็นตัวแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่นที่สามารถวิเคราะห์หารูปแบบแนวโน้มการเกิดภาวะซึมเศร้าและปัจจัยที่ก่อให้เกิดภาวะซึมเศร้าในวัยรุ่น ซึ่งได้รับผลการประเมินจากผู้เชี่ยวชาญอยู่ในระดับเหมาะสมมาก และได้รับการประเมินความพึงพอใจในการใช้งานตัวแบบจากวัยรุ่นกลุ่มตัวอย่างอยู่ในระดับความพึงพอใจมาก

(ภรัณยา ปาลวิสุทธิ์, 2559) ศึกษาการเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique : SMOTE) สำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต มีวัตถุประสงค์เพื่อพัฒนาตัวแบบสำหรับพยากรณ์การเป็นโรคติดอินเทอร์เน็ตในเยาวชน โดยใช้วิธีต้นไม้ตัดสินใจ J48, ID3, LMT, CART และ Random forest และ ใช้ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ และพบว่าตัวแบบที่พัฒนาโดยเทคนิค Random forest สามารถพยากรณ์ได้ดีกว่า J48 ID3 LMT และ CART ซึ่งมีค่าความแม่นยำร้อยละ 87.15 ค่าความไวร้อยละ 85.89 และค่าความจำเพาะร้อยละ 87.53

(Saravananathan & Velmurugan, 2016) ได้นำการทำเหมืองข้อมูลมาวินิจฉัยผู้ป่วยโรคหัวใจโดยใช้อัลกอริทึมการจำแนกและวัดประสิทธิภาพการจำแนกของข้อมูลเพื่อหาเทคนิคการจำแนกข้อมูลที่ดีที่สุด จากการทดลองพบว่า อัลกอริทึม J48 มีค่าความถูกต้องสูงสุด 67.15% รองลงมา คือ Support Vector Machine มีค่าความถูกต้อง 65.04%, การจำแนกและสมการถดถอยแบบต้นไม้ (Classification and Regression tree : CART) มีค่าความถูกต้อง 62.28% และ K-Nearest neighbor มีค่าความถูกต้อง 53.39% ตามลำดับ

(ประยุทธ์ศิลป์ ชัยนาม, 2562) ศึกษาการสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลและวิชวลไลเซชั่น โดยมีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบโมเดลสำหรับจำแนกผู้ป่วยโรคไตเรื้อรัง ซึ่งใช้เทคนิคการทำเหมืองข้อมูล เลือกใช้วิธีต้นไม้ตัดสินใจ วิธีต้นไม้ตัดสินใจแบบสุ่ม วิธีความใกล้เคียงกันด้วยค่าเค วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีนาอีฟเบย์ จากผลการศึกษาพบว่าแบบจำลองนาอีฟเบย์ แบบ MultinomialNB มีความสามารถในการพยากรณ์ได้

แม่นยำที่สุด รองลงมา คือ แบบจำลองต้นไม้ตัดสินใจ แบบจำลองป่าไม้แบบสุ่ม แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน และแบบจำลองค่าเคใกล้เคียง ตามลำดับ

(อรุณรักษ์ ตันพานิช, ดุษณี ศุภวรรธนะกุล, พิเชฐ บัญญัติ, & จรูญ จันแทน, 2562) ศึกษาการเปรียบเทียบโมเดลการเรียนรู้ของเครื่องสำหรับคัดกรองผู้ป่วยเบาหวานที่มีภาวะชาปลายเท้า โดยมีวัตถุประสงค์เพื่อสังเคราะห์ข้อมูลผู้ป่วยโรคเบาหวานที่มีภาวะชาปลายเท้าและเปรียบเทียบโมเดลการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner Studio ซึ่งปัจจัยส่วนบุคคลที่ส่งผลให้มีโอกาสเกิดความเสี่ยงในการเป็นโรคเบาหวานเพิ่มขึ้น คือกลุ่มคนอายุมากกว่า 55 ปี และมีประวัติเกี่ยวกับความดันโลหิตสูง การใช้โปรแกรม RapidMiner Studio สำหรับการเรียนรู้ของเครื่องนั้น พบว่าโมเดลแบบ Support Vector Machine (SVM) มีความเชื่อมั่นสูงสุดและใช้ระยะเวลาในการประมวลผลเร็วกว่าวิธีอื่น หากเพิ่มจำนวนกลุ่มตัวอย่างให้มากขึ้น และการเก็บข้อมูลเพิ่มจำนวนครั้งสำหรับการเก็บข้อมูลซ้ำ จะทำให้การเรียนรู้ของเครื่องมีความแม่นยำยิ่งขึ้น โดยในงานวิจัยนี้จำนวนข้อมูลที่นำมาใช้ในการสร้างแบบจำลองมีเพียง 100 คน โดยแบ่งเป็นผู้ป่วยที่โรคเบาหวานเบาหวานที่มีภาวะชาปลายเท้าจำนวน 50 คน และคนปกติที่ไม่เป็นโรคเบาหวานจำนวน 50 คน ดังนั้นจึงมีข้อเสนอแนะให้เพิ่มจำนวนกลุ่มตัวอย่างให้มากขึ้น เพื่อให้การสร้างแบบจำลอง Machine learning มีความถูกต้องและแม่นยำยิ่งขึ้น

(รุ่งโรจน์ บุญมา & นิเวศ จิระวิชิตชัย, 2562) ศึกษาการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล และการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล โดยมีวัตถุประสงค์เพื่อสร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูลและการเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยในงานวิจัยนี้ใช้ข้อมูลจาก UCI ซึ่งเป็นข้อมูลเปิดที่มีความสมบูรณ์เนื่องจากเป็นข้อมูลมีการกรองมาแล้วจากทางผู้ดูแลเว็บไซต์สำหรับการนำไปทดสอบประสิทธิภาพด้านเหมืองข้อมูล ซึ่งเป็นฐานข้อมูล Diabetes Data ศึกษาตัวแปรทั้งหมด 9 แอททริบิวต์ ได้แก่ การตั้งครรภ์ กลูโคส ความดันโลหิต ความหนาของชั้นผิวหนัง อินซูลิน ค่าดัชนีมวลกาย ครอบครัวมีประวัติเป็นโรคเบาหวาน อายุ และผลการจำแนกผู้ป่วยโรคเบาหวาน โดยผลการศึกษาพบว่า Support Vector Machine มีประสิทธิภาพการทำนายสูงสุด

(Dimas & Naqshauliza, 2020) ศึกษาการเปรียบเทียบความแม่นยำของ Support Vector Machine (SVM) และ K-Nearest neighbors (KNN) ในการทำนายโรคหัวใจ โดยมีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการจำแนกประเภทระหว่าง Support Vector Machine (SVM) และ K-Nearest neighbors (KNN) ซึ่งเป็นหนึ่งในเทคนิคของ Machine learning เพื่อค้นหาวิธีที่สามารถนำมาใช้ในการทำนายข้อมูลโรคหัวใจได้อย่างแม่นยำ ผลการศึกษาพบว่าการทดสอบอัลกอริทึม Support Vector Machine ด้วยการปรับให้เป็นค่ามาตรฐานมีผลการจำแนกที่แม่นยำกว่าเมื่อเทียบกับอัลกอริทึม KNN ไม่ว่าจะในกรณีการปรับหรือไม่ปรับค่าให้เป็นค่ามาตรฐาน

(Abdulqadir, Abdulazeez, & Zebari, 2021) ศึกษาการจำแนกประเภทโดยใช้เทคนิคเหมืองข้อมูลสำหรับการทำนายโรคหลอดเลือดสมอง มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของวิธีการที่ใช้ในการจำแนกประเภท คือ Random forest และ Support Vector Machine โดยพิจารณาจากความถูกต้องในการจำแนก ซึ่งผลการศึกษาพบว่า วิธี Random forest มีความถูกต้องในการจำแนกร้อยละ 75.78 ซึ่งสูงกว่าวิธี Support Vector Machine ที่มีความถูกต้องในการจำแนกร้อยละ 65.10

(Changpetch et al., 2021) ศึกษาการรวมเทคนิคการทำเหมืองข้อมูลสำหรับการจำแนกประเภท Naïve Bayes โดยการประยุกต์ใช้กับชุดข้อมูลทางการแพทย์ 3 ชุด คือ ไทรอยด์ เบาหวาน และไส้ติ่งอักเสบ โดยมีวัตถุประสงค์เพื่อเลือกตัวแปรทำนายและอิทธิพลร่วมโดยวิธีต้นไม้ตัดสินใจแบบจำแนก (classification trees) และวิธีการวิเคราะห์กฎความสัมพันธ์ (association rules analysis) เพื่อปรับปรุงประสิทธิภาพแบบจำลองการจำแนกประเภทของเทคนิค Naïve Bayes ซึ่งผลการศึกษาพบว่า การพิจารณาอิทธิพลร่วมทำให้ประสิทธิภาพของแบบจำลองการจำแนกของเทคนิค Naïve Bayes เพิ่มขึ้น

(พีรศุษม์ ทองพ่วง & จรัญ แสนราช, 2021) ศึกษาการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายการได้รับทุนการศึกษาของนักศึกษาปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล โดยมีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการได้รับทุนการศึกษาของนักศึกษาระดับปริญญาตรี และเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของตัวแบบด้วยเทคนิควิธีต้นไม้ตัดสินใจ (Decision tree) วิธีแบบเบย์ (Naïve Bayes) วิธีเคเนียร์เรสเนเบอร์ (K-Nearest neighbors) วิธีการเรียนรู้เชิงลึก (Deep learning) และวิธีต้นไม้ป่าสุ่ม (Random forest) มีข้อมูลจำนวน 1,155 ชุด 15 แอตทริบิวต์ และใช้โปรแกรม Rapid Miner 9.3 ในการวิเคราะห์ปัจจัยที่เกี่ยวข้องและทดสอบประสิทธิภาพด้วยวิธี 10-fold cross validation กับตัวแบบ ซึ่งผลการศึกษาพบว่า ข้อมูลที่มีความเกี่ยวข้องกันมีจำนวน 11 แอตทริบิวต์ และการจำแนกประเภทโดยวิธีต้นไม้ป่าสุ่มเป็นวิธีที่มีค่าความถูกต้องสูงสุด รองลงมา คือ วิธีการเรียนรู้เชิงลึก วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ และวิธีเคเนียร์เรสเนเบอร์ ตามลำดับ

(ปพนน์ศรณ์ สิ่วสำแดงเดช, 2565) ศึกษาการจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี ได้ประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผู้ป่วยโรคเบาหวานด้วยเทคนิคต้นไม้การตัดสินใจ (Decision tree) เทคนิคนาอีฟ เบย์ (Naïve Bayes) เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor) เทคนิคโหวตร่วม (Vote ensemble) และเทคนิคป่าสุ่ม (Random forest) เพื่อสร้างแบบจำลองในการพยากรณ์ผู้ป่วย ซึ่งคุณลักษณะที่ใช้สำหรับงานวิจัยนี้ได้จากฐานข้อมูลของโรงพยายาลศูนย์อุดรธานี ซึ่งประกอบด้วยตัวแปรทั้งหมด 13 ตัวแปร คือ เลขประจำตัวผู้ป่วย อายุ เพศ น้ำหนัก ส่วนสูง ค่าล่างความดันเลือด ค่าบนความดันเลือด ชีพจร การสูบบุหรี่ การดื่มแอลกอฮอล์ ดัชนีมวลกาย ระดับคอลเรสเตอรอล ระดับกลูโคส ผลการศึกษาพบว่า เทคนิคป่าสุ่มให้ค่าความถูกต้องในการทำนายผลการเป็นโรคเบาหวานได้ดีที่สุด

(Nandhini & Dharmarajan, 2022) ศึกษาการคาดการณ์โรคเบาหวานโดยใช้ชุดข้อมูล Pima Indian ประกอบด้วย 1,500 ราย และ 10 ตัวแปร ซึ่งการศึกษานี้มุ่งเน้นไปที่ผลการจำแนกของเทคนิค Random forest ร่วมกับการคัดเลือกคุณลักษณะด้วยวิธี Wrapper ที่แตกต่างกัน 4 วิธี ได้แก่ Forward feature selection, Backward feature selection, Exhaustive feature selection และ Recursive feature elimination โดยแต่ละเทคนิคทำการปรับค่า hyperparameter ที่เหมาะสมโดยใช้วิธี Grid search ซึ่งผลการศึกษาพบว่า การปรับค่า hyperparameter ที่เหมาะสมทำให้ความแม่นยำในการจำแนกเพิ่มขึ้น และการใช้วิธี Exhaustive feature selection ในการคัดเลือกคุณลักษณะจากทั้งหมด 9 ตัวแปร เหลือ 6 ตัวแปร เพื่อใช้ในการสร้างแบบจำลองโดยใช้เทคนิค Random forest มีความถูกต้องสูงสุด (90.66%)

จากงานวิจัยที่กล่าวมาข้างต้น ผู้วิจัยจึงเลือกใช้เทคนิคต้นไม้ตัดสินใจ (Decision tree) ต้นไม้ป่าสุ่ม (Random forest) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor) ในการสร้างแบบจำลองสำหรับการจำแนกการเป็นโรคเบาหวาน ซึ่งพบว่าเทคนิคดังกล่าวเป็นเทคนิคที่มีประสิทธิภาพในการจำแนกสูง โดยจะสร้างแบบจำลองทั้งกรณีพิจารณาและไม่พิจารณาอิทธิพลร่วม ทำการปรับปรุงข้อมูลให้มีคุณภาพมากยิ่งขึ้น โดยการตัดข้อมูลที่มีความซ้ำซ้อนออกและประมาณค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ไม่ได้สูญหาย เพื่อให้ได้ข้อมูลที่มีความสมบูรณ์มากที่สุด และปรับปรุงประสิทธิภาพของแบบจำลองการจำแนกโดยการกำหนดค่าไฮเปอร์พารามิเตอร์ที่มีความเหมาะสมกับข้อมูลจากวิธีการค้นหาแบบกริด (Grid search) เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพสูงสุด

บทที่ 3

ระเบียบวิธีวิจัย

การศึกษาเรื่องการเปรียบเทียบประสิทธิภาพแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน ในส่วนระเบียบวิธีวิจัยจะเป็นการอธิบายขั้นตอนของวิธีการดำเนินงานวิจัย โดยแบ่งหัวข้อออกเป็นดังนี้

- ข้อมูลที่ใช้ในการศึกษา
- วิธีการดำเนินงานวิจัย
- ขั้นตอนการสร้างแบบจำลอง และการหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม
  1. เทคนิคต้นไม้ตัดสินใจ (Decision tree)
  2. เทคนิคต้นไม้ป่าสุ่ม (Random forest)
  3. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
  4. เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor)
- เกณฑ์การเปรียบเทียบประสิทธิภาพของแบบจำลอง
- เครื่องมือที่ใช้ในการวิจัย

ข้อมูลที่ใช้ในการศึกษา

ขอบเขตด้านข้อมูล ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลทุติยภูมิ (Secondary data) ระหว่างปี พ.ศ. 2562 - 2564 รวมระยะเวลา 3 ปี โดยรวบรวมข้อมูลต่าง ๆ จากการซักประวัติผู้ป่วยที่เข้ารับบริการในโรงพยาบาลสังกัดสำนักการแพทย์ กรุงเทพมหานคร ทั้งในรูปแบบของออนไลน์ (ระบบสารสนเทศโรงพยาบาล) หรือออฟไลน์ (รูปแบบของเอกสาร) เพื่อใช้ในการสร้างแบบจำลองสำหรับการจำแนกการเป็นโรคเบาหวาน

ขอบเขตด้านคุณลักษณะ คุณลักษณะที่ใช้ในการศึกษาวิจัยครั้งนี้ มีจำนวนทั้งสิ้น 10 ตัวแปร ประกอบด้วย คุณลักษณะที่เป็นตัวแปรอิสระ จำนวน 9 ตัวแปร ได้แก่ เพศ อายุ น้ำหนัก ส่วนสูง ดัชนีมวลกาย ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลายตัว อัตราการเต้นของหัวใจ และประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง) และคุณลักษณะที่เป็นตัวแปรตามเพื่อใช้ในการแบ่งกลุ่ม 1 ตัวแปร คือ การเป็นโรคเบาหวาน โดยแบ่งเป็น 2 กลุ่ม ได้แก่ เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวาน

ตัวแปรที่นำมาใช้พิจารณาอิทธิพลร่วม

จากการศึกษาของ (Amput, Srithawong, Sittitan, Wongphon, & Sangkarit, 2016) พบว่า ค่าดัชนีมวลกายเฉลี่ยเท่ากับ 26.16 ± 3.16 กิโลกรัมต่อตาราเมตร ซึ่งถือว่าอยู่ในกลุ่มที่มีน้ำหนักเกินเกณฑ์มาตรฐาน มีความสัมพันธ์ในระดับสูง (r=0.91) กับระดับน้ำตาลในเลือดอย่างมีนัยสำคัญทางสถิติ และจากการศึกษาของ (Techasuwanna, Chottanapun, Chamroonsawasd, Sornpaisar, & Tunyasitthisundhorn, 2020) พบว่า การมีพ่อ แม่ และญาติสายตรงเป็นโรคเบาหวานก่อให้เกิดความเสี่ยงต่อการเป็นโรคเบาหวานมากกว่าปกติถึง 3 เท่า ดังนั้นปัจจัยเสี่ยงดังกล่าว จึงเป็นปัจจัยเสี่ยงที่มีความสำคัญต่อการเกิดโรคเบาหวาน ในงานวิจัยนี้จึงพิจารณาอิทธิพลร่วมของปัจจัยดังกล่าว ดังนี้

1. ดัชนีมวลกาย*ประวัติโรคเบาหวานในญาติสายตรง
2. ดัชนีมวลกาย*เพศ
3. ดัชนีมวลกาย*อายุ
4. ดัชนีมวลกาย*ค่าความดันขณะหัวใจบีบตัว
5. ดัชนีมวลกาย*ค่าความดันขณะหัวใจคลายตัว
6. ดัชนีมวลกาย*อัตราการเต้นของหัวใจ
7. ประวัติโรคเบาหวานในญาติสายตรง*เพศ
8. ประวัติโรคเบาหวานในญาติสายตรง*อายุ
9. ประวัติโรคเบาหวานในญาติสายตรง*ค่าความดันขณะหัวใจบีบตัว
10. ประวัติโรคเบาหวานในญาติสายตรง*ค่าความดันขณะหัวใจคลายตัว
11. ประวัติโรคเบาหวานในญาติสายตรง*อัตราการเต้นของหัวใจ

วิธีการดำเนินงานวิจัย

การวิจัยในครั้งนี้เป็นการศึกษาเพื่อเปรียบเทียบแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน โดยเลือกสร้างแบบจำลองจาก 4 เทคนิค คือ เทคนิคต้นไม้ตัดสินใจ (Decision tree) เทคนิคต้นไม้ป่าสุ่ม (Random forest) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor) ซึ่งกระบวนการในการสร้างแบบจำลองสำหรับการจำแนกประเภทและการเปรียบเทียบประสิทธิภาพ แบ่งเป็นขั้นตอนดังต่อไปนี้

1. การทำความเข้าใจข้อมูล

จากการศึกษาเอกสารที่เกี่ยวข้องและปัจจัยเสี่ยงของโรคเบาหวาน (สมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี, 2560) ประกอบด้วยตัวแปรอิสระ 9 ตัวแปร ได้แก่ เพศ อายุ น้ำหนัก ส่วนสูง ดัชนีมวลกาย ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลายตัว อัตราการเต้นของหัวใจ และประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง) และ ตัวแปรตาม คือ การเป็นโรคเบาหวาน โดยที่ตัวแปรแต่ละตัวมีรายละเอียด ดังตารางที่ 8

ตารางที่ 8 รายละเอียดตัวแปรที่ใช้ในงานวิจัย

ตัวแปร	คุณลักษณะ	หน่วย
Y	การเป็นโรคเบาหวาน	0 = ไม่เป็นโรคเบาหวาน 1 = เป็นโรคเบาหวาน
X_1	เพศ	0 = ชาย (Male) 1 = หญิง (Female)
X_2	อายุ	ปี (Year)
X_3	น้ำหนัก	กิโลกรัม (kg.)
X_4	ส่วนสูง	เซนติเมตร (cm.)
X_5	ดัชนีมวลกาย	(กิโลกรัม/เมตร^2)
X_6	ค่าความดันขณะหัวใจบีบตัว	มิลลิเมตรปรอท (mmHg.)
X_7	ค่าความดันขณะหัวใจคลายตัว	มิลลิเมตรปรอท (mmHg.)
X_8	อัตราการเต้นของหัวใจ	ครั้งต่อนาที (bpm)
X_9	ประวัติโรคเบาหวาน ในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง)	0 = ไม่มี 1 = มี

2. การจัดเตรียมข้อมูล คือการปรับปรุงข้อมูลให้มีคุณภาพมากยิ่งขึ้น โดยดำเนินการดังนี้

   1) การทำความสะอาดข้อมูลเป็นขั้นตอนที่ต้องใช้เวลามากที่สุด ซึ่งในการเตรียมข้อมูลพบว่ามีข้อมูลบางส่วนที่มีความซ้ำซ้อนและสูญหาย ซึ่งปัญหานี้อาจส่งผลต่อประสิทธิภาพของแบบจำลอง จึงจำเป็นต้องมีการทำความสะอาดข้อมูล ดังนั้นผู้วิจัยจึงตัดข้อมูลที่มีความซ้ำซ้อนออก และประมาณค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ไม่ได้สูญหาย ดังนั้นเหลือข้อมูลทั้งสิ้น 20,227 ราย ซึ่งแบ่งเป็นข้อมูลผู้ป่วยที่ไม่เป็นโรคเบาหวานจำนวน 11,662 ราย และข้อมูลผู้ป่วยที่เป็นโรคเบาหวานจำนวน 8,565 ราย จะเป็นชุดข้อมูลที่ผู้วิจัยจะใช้ในการสร้างแบบจำลองสำหรับการจำแนกการเป็นโรคเบาหวาน โดยรายละเอียดของข้อมูลเป็น ดังตารางที่ 9

ตารางที่ 9 แสดงรายละเอียดข้อมูลผู้ป่วยจำนวน 20,227 ราย โดยแยกเป็นกรณีผู้ป่วยที่ไม่เป็นโรคเบาหวาน จำนวน 11,662 ราย และผู้ป่วยที่เป็นโรคเบาหวานจำนวน 8,565 ราย

ตัวแปร	ผู้ป่วยที่ไม่เป็นโรคเบาหวาน			ผู้ป่วยที่เป็นโรคเบาหวาน		
	จำนวน	Min	Max	จำนวน	Min	Max
เพศ	ชาย 4,089 หญิง 7,573			ชาย 3,573 หญิง 4,992		
อายุ		31	97		31	94
น้ำหนัก		30	167		31	167
ส่วนสูง		100	195		136	195
ดัชนีมวลกาย		12.49	45.18		12.66	45.79
ค่าความดันขณะหัวใจบีบตัว		77	198		80	237
ค่าความดันขณะหัวใจคลายตัว		37	128		35	150
อัตราการเต้นของหัวใจ		36	161		36	151
ประวัติโรคเบาหวานในญาติสายตรง	มี 2,700 ไม่มี 8,962			มี 6,321 ไม่มี 2,244		

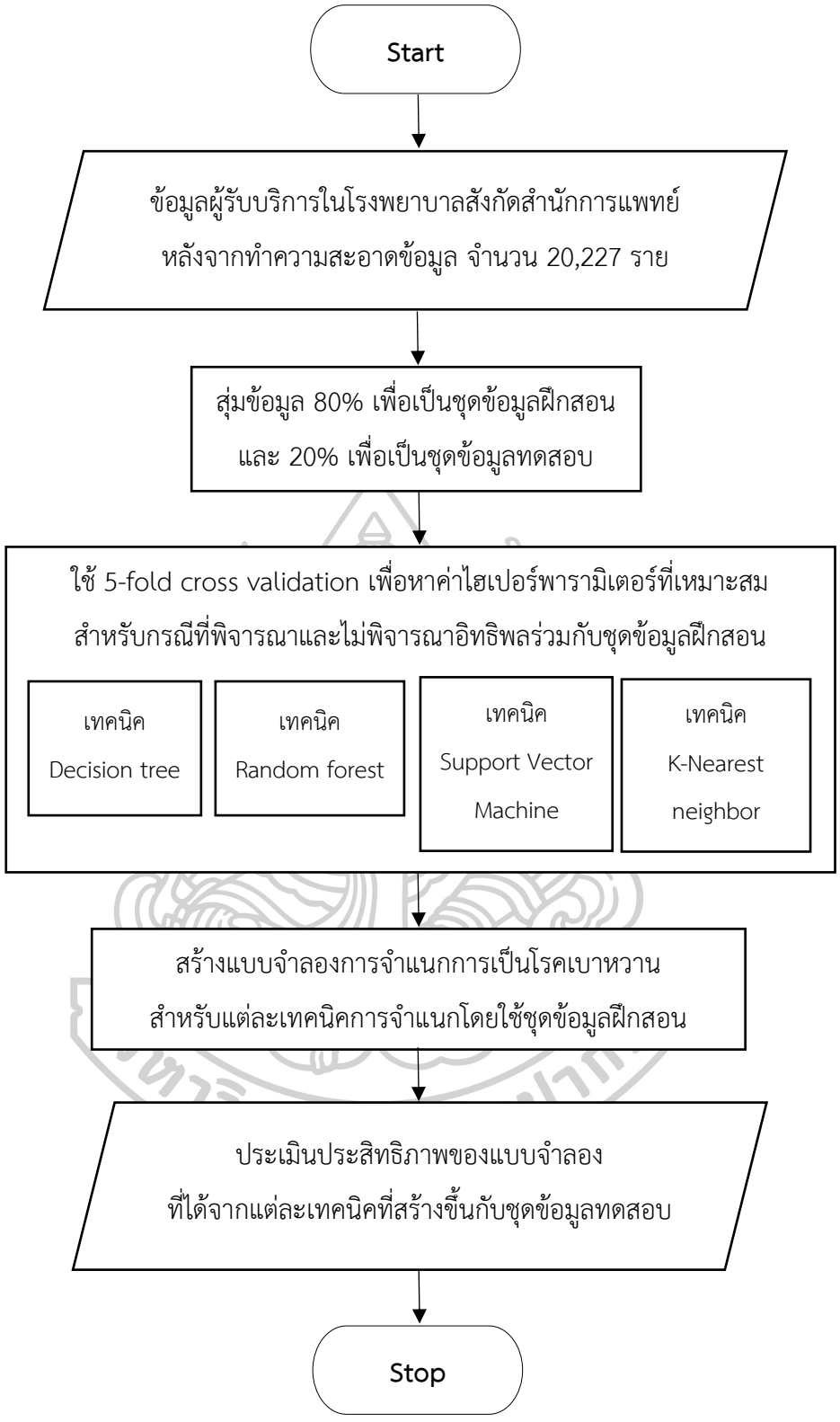
2) การแปลงข้อมูล เป็นการแปลงข้อมูลเพื่อให้แต่ละตัวแปรอิสระมีค่าที่อยู่ในรูปการทำให้เป็นปกติ (Normalization) โดยในที่นี้ใช้การแปลงข้อมูลของแต่ละตัวแปรอิสระให้อยู่ในช่วง [0,1] เพื่อให้ตัวแปรอิสระแต่ละตัวมีความเท่าเทียมกัน โดยใช้สูตรการคำนวณ คือ

$$x^* = \frac{X - \min(x)}{\max(x) - \min(x)} \quad (26)$$

โดยที่ x^* คือ ค่าที่ได้จากการแปลงค่าอยู่ในรูปการทำให้เป็นปกติ
X คือ ค่าปัจจุบันที่นำมาแปลงค่าอยู่ในรูปการทำให้เป็นปกติ
$\min(x)$ คือ ค่าข้อมูลที่มีค่าน้อยที่สุดในชุดข้อมูล
$\max(x)$ คือ ค่าข้อมูลที่มีค่ามากที่สุดในชุดข้อมูล

ขั้นตอนการสร้างแบบจำลอง

1. จากข้อมูลทั้งสิ้นจำนวน 20,227 ราย สุ่มข้อมูล 80% เพื่อเป็นชุดข้อมูลฝึกสอน ซึ่งจะได้ข้อมูลในส่วนนี้จำนวน 16,181 ราย และ 20% เพื่อเป็นชุดข้อมูลทดสอบ ซึ่งจะได้ข้อมูลในส่วนนี้จำนวน 4,046 ราย
2. ใช้ 5-fold cross validation กับชุดข้อมูลฝึกสอน เพื่อหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมสำหรับเทคนิค Machine learning ทั้ง 4 วิธี คือ ต้นไม้ตัดสินใจ ต้นไม้ป่าสุ่ม ซัพพอร์ตเวกเตอร์แมชชีน และเพื่อนบ้านใกล้ที่สุด
3. สร้างแบบจำลองการจำแนกการเป็นโรคเบาหวานโดยใช้เทคนิค Machine learning กับชุดข้อมูลฝึกสอน โดยการกำหนดค่าไฮเปอร์พารามิเตอร์ที่ได้จากข้อ 2
4. นำแบบจำลองการจำแนกการเป็นโรคเบาหวานที่ได้จากข้อ 3 มาประเมินประสิทธิภาพของแบบจำลองด้วยค่าวัดประสิทธิภาพต่าง ๆ กับชุดข้อมูลทดสอบ



ภาพที่ 20 แผนผังแสดงขั้นตอนการสร้างแบบจำลอง

การหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม

การสร้างแบบจำลองโดยกำหนดไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุด ผู้วิจัยใช้วิธีการค้นหาแบบกริด (Grid search) ซึ่งเป็นการสร้างแบบจำลองจากค่าของไฮเปอร์พารามิเตอร์ที่กำหนดไว้ทุกชุด โดยกำหนดให้แบ่งชุดข้อมูลฝึกสอนออกเป็น 5 กลุ่ม สำหรับการตรวจสอบแบบไขว้ (cross validation) จากนั้นประเมินประสิทธิภาพของแบบจำลองด้วยค่าวัดความถูกต้อง (accuracy) และเลือกค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมจากแบบจำลองที่ให้ค่าความถูกต้องสูงสุด โดยทดลองหาค่าไฮเปอร์พารามิเตอร์กับทั้ง 4 เทคนิคการจำแนกข้อมูลทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม ซึ่งแต่ละเทคนิคมีรายละเอียดดังนี้

1. เทคนิคต้นไม้ตัดสินใจ (Decision tree)

เทคนิคต้นไม้ตัดสินใจ (Decision tree) เป็นการใช้คุณลักษณะต่าง ๆ ของข้อมูลในการสร้างต้นไม้ตัดสินใจในลักษณะจากบนลงล่าง โดยที่การสร้างต้นไม้ตัดสินใจนั้นจะเป็นโครงสร้างที่มีกฎต่าง ๆ ซึ่งนำมาใช้ในการจำแนกกลุ่มข้อมูล ทำได้โดยการพิจารณาต้นไม้ตัดสินใจจากรากไปจนถึงใบหนึ่ง ๆ ซึ่งเส้นทางที่ได้จะต้องผ่านโหนดที่เก็บค่าคุณลักษณะ และกิ่งของต้นไม้ซึ่งเก็บค่าที่เป็นไปได้ของโหนดนั้น ๆ โดยกฎที่ได้จากต้นไม้ตัดสินใจในแต่ละโหนด จะมีจำนวนที่ผ่านโหนดและกิ่งของต้นไม้ต่าง ๆ ไม่เท่ากัน เมื่อนำข้อมูลที่ต้องการทดสอบเข้ามาตรวจสอบกับกฎข้อต่าง ๆ ทีละข้อ หากข้อมูลที่นำมาทดสอบนี้มีค่าคุณลักษณะที่ตรงกับกฎทุกคุณลักษณะข้อมูลที่นำมาทดสอบนั้นก็จะมีกลุ่มตามกฎข้อนั้น ๆ หากไม่ตรงทั้งหมด ก็จะทำการเปรียบเทียบกับกฎข้อต่อ ๆ ไป

เทคนิคต้นไม้ตัดสินใจในโปรแกรม Weka ใช้อัลกอริทึม J48 ซึ่งสามารถใช้ได้กับข้อมูลทั้งแบบไม่ต่อเนื่องและแบบต่อเนื่อง ต้นไม้ตัดสินใจที่เหมาะสมสามารถหาได้จากการปรับค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ด้วยวิธีการค้นหาแบบกริด ไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา มีดังนี้

- confidenceFactor คือ ช่วงความเชื่อมั่นที่ใช้ในการแตกกิ่งต้นไม้
- minNumObj คือ จำนวนข้อมูลเรียนรู้ขั้นต่ำในโหนดใบ (leaf node) หากมีการแบ่งแล้วมีข้อมูลเรียนรู้ที่ต่ำกว่า minNumObj จะถูกตัดออกจากการตัดสินใจ

การหาค่าที่เหมาะสมสำหรับไฮเปอร์พารามิเตอร์ดังกล่าวทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วมด้วยวิธีการค้นหาแบบกริด ทำโดยกำหนดค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ดังตารางที่ 10

ตารางที่ 10 ไฮเปอร์พารามิเตอร์เทคนิคต้นไม้ตัดสินใจสำหรับการค้นหาแบบกริด

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
confidenceFactor	0.25, 0.5, 0.75
minNumObj	1, 3, 5, 7, 9

2. เทคนิคต้นไม้ป่าสุ่ม (Random forest)

เทคนิคต้นไม้ป่าสุ่ม (Random forest) ใช้หลักการสุ่มคุณลักษณะหลาย ๆ รูปแบบเพื่อสร้างต้นไม้ตัดสินใจขึ้นมาหลาย ๆ ต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้นนั้นจะมีกฎที่แตกต่างกันเพื่อนำมาใช้เป็นเกณฑ์ในการจำแนกกลุ่มข้อมูล ซึ่งผลการจำแนกกลุ่มที่ได้จากต้นไม้ตัดสินใจแต่ละต้นจะถูกนำมาคิดเป็นผลการโหวต โดยผลการจำแนกกลุ่มที่ได้ผลโหวตมากที่สุดจะใช้ในการระบุกลุ่มของข้อมูลทดสอบ

เทคนิคต้นไม้ป่าสุ่มในโปรแกรม Weka ใช้อัลกอริทึมชนิด RandomForest ซึ่งในการหาต้นไม้ที่เหมาะสมนั้น สามารถหาได้จากการปรับค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ให้เหมาะสมด้วยวิธีการค้นหาแบบกริด ไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา มีดังนี้

- numIterations คือ จำนวนต้นไม้ทั้งหมดที่ต้องการสร้าง
- maxDepth คือ ระดับความลึกที่มากที่สุดของต้นไม้ตัดสินใจ

การหาค่าที่เหมาะสมสำหรับไฮเปอร์พารามิเตอร์ดังกล่าวทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วมด้วยวิธีการค้นหาแบบกริด ทำโดยกำหนดค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ดังตารางที่ 11

ตารางที่ 11 ไฮเปอร์พารามิเตอร์เทคนิคต้นไม้ป่าสุ่มสำหรับการค้นหาแบบกริด

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
numIterations	10, 20, ... , 100
maxDepth	3, 5, 10, 20, none

3. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นเทคนิคที่อาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นหรือระนาบ (hyperplane) สำหรับแบ่งกลุ่มข้อมูลออกจากกันได้ดีที่สุด โดยเส้นแบ่งที่ดีที่สุดพิจารณาจากเส้นแบ่งที่มีระยะขอบ (margin) กว้างมากที่สุด ซึ่งจะถูกนำมาใช้เป็นเกณฑ์ในการจำแนกกลุ่มข้อมูล การจำแนกกลุ่มด้วยเทคนิคซัพพอร์ตเวกเตอร์ แมชชีนมีทั้งแบบเชิงเส้น (Linear SVM) ซึ่งสามารถจำแนกกลุ่มข้อมูลออกจากกันได้โดยเส้นตรงและแบบไม่เชิงเส้น (Nonlinear SVM) ซึ่งจำเป็นต้องอาศัยการแปลงข้อมูลให้อยู่ในมิติที่สูงขึ้นโดยใช้ฟังก์ชันเคอร์เนล (kernel) เพื่อให้สามารถจำแนกกลุ่มข้อมูลออกจากกันได้

เทคนิคซัพพอร์ตเวกเตอร์แมชชีนในโปรแกรม Weka ใช้อัลกอริทึมชนิด SMO ซึ่งอาศัยไฮเปอร์พารามิเตอร์ต่าง ๆ ในการกำหนดเงื่อนไขสำหรับสร้างเส้นแบ่ง (hyperplane) โดยไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา มีดังนี้

- C คือ ไฮเปอร์พารามิเตอร์ที่กำหนดให้เป็นมาตรฐาน (Regularization parameter) เมื่อ C มีค่ามาก ระยะขอบของเส้นแบ่งจะลดลง ทำให้ตัวจำแนกมีความซับซ้อนมากขึ้น และเมื่อ C มีค่าน้อย ระยะขอบของเส้นแบ่งจะเพิ่มขึ้น ทำให้ตัวจำแนกมีความเรียบง่ายมากขึ้น
- kernel คือ ประเภทของฟังก์ชันทางคณิตศาสตร์ในรูปแบบต่าง ๆ ได้แก่ เคอร์เนลเชิงเส้น (polykernel (exponent=1)) เคอร์เนลพหุนาม (polykernel (exponent=2, ... , 5)) และเคอร์เนลฟังก์ชันฐานรัศมี (rbf)
- exponent คือ ค่ายกกำลังของเคอร์เนลพหุนาม
- gamma คือ ไฮเปอร์พารามิเตอร์ที่กำหนดอิทธิพลข้อมูลชุดเรียนรู้ของเคอร์เนลฟังก์ชันฐานรัศมี โดยค่า gamma จะมีผลต่อรูปร่างและความซับซ้อนในการจำแนกข้อมูลของแบบจำลอง เมื่อ gamma มีค่าน้อย Hyperplane จะมีการโค้งน้อยหรือเกือบจะเป็นเส้นตรงหากมีค่าที่น้อยมาก และเมื่อ gamma มีค่ามาก Hyperplane จะมีความโค้งมากและจะมีความจำเพาะกับข้อมูลได้ดี

การหาค่าที่เหมาะสมสำหรับไฮเปอร์พารามิเตอร์ดังกล่าวทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วมด้วยวิธีการค้นหาแบบกริด ทำโดยกำหนดค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ดังตารางที่ 12

ตารางที่ 12 ไฮเปอร์พารามิเตอร์เทคนิคซัพพอร์ตเวกเตอร์แมชชีนสำหรับการค้นหาแบบกริด

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
kernel	polykernel (exponent=1), polykernel (exponent=2, ... , 5), rbf
C	5, 10, 15, ... , 50
exponent	2, 3, 4, 5
gamma	0.05, 0.1, 0.2, 0.5, 1

4. เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor)

เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor) เป็นวิธีการที่ใช้ในการจำแนกกลุ่มข้อมูล โดยการวัดระยะห่างของข้อมูลทดสอบกับข้อมูลที่มีอยู่ด้วยสมการหาระยะทางระหว่างจุดเพื่อหาจุดข้อมูลใกล้เคียงที่สุด ซึ่งจะใช้ค่า K หรือจำนวนข้อมูลในระยะใกล้เคียงที่กำหนดมาใช้เป็นเกณฑ์ในการตัดสินใจจำแนกกลุ่มของข้อมูลทดสอบ โดยจะกำหนดกลุ่มให้กับข้อมูลทดสอบตามกลุ่มส่วนใหญ่ของสมาชิก K ตัวที่มีความใกล้เคียงที่สุด

เทคนิคเพื่อนบ้านใกล้ที่สุดในโปรแกรม Weka ใช้อัลกอริทึมชนิด IBk ซึ่งจำเป็นต้องอาศัยไฮเปอร์พารามิเตอร์ต่าง ๆ ในการกำหนดเงื่อนไขของจุดเพื่อนบ้านที่จะนำมาพิจารณา ดังนี้

- K คือ จำนวนจุดเพื่อนบ้านที่ใช้ (K)
- distanceFunction คือ ฟังก์ชันการหาระยะทางสำหรับค้นหาเพื่อนบ้าน
- DistanceWeighting คือ ฟังก์ชันถ่วงน้ำหนักที่ใช้ ได้แก่
  - ทุกจุดเพื่อนบ้านไม่ถูกถ่วงน้ำหนัก (No distance weighting)
  - ทุกจุดเพื่อนบ้านถูกถ่วงน้ำหนักตามระยะทางโดยใช้ฟังก์ชันถ่วงน้ำหนัก 1/distance (Weight by 1/distance)

การหาค่าที่เหมาะสมสำหรับไฮเปอร์พารามิเตอร์ดังกล่าวทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วมด้วยวิธีการค้นหาแบบกริด ทำโดยกำหนดค่าไฮเปอร์พารามิเตอร์ต่าง ๆ ดังตารางที่ 13

ตารางที่ 13 ไฮเปอร์พารามิเตอร์เทคนิคเพื่อนบ้านใกล้ที่สุดสำหรับการค้นหาแบบกริด

ไฮเปอร์พารามิเตอร์	ค่าไฮเปอร์พารามิเตอร์
K	1, 3, ... , 31
distanceFunction	Euclidean, Manhattan
DistanceWeighting	No distance weighting, Weight by 1/distance

เกณฑ์การเปรียบเทียบประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลองด้วยเกณฑ์การพิจารณาประสิทธิภาพ 4 เกณฑ์ กับชุดข้อมูลทดสอบ ดังนี้

1. ค่าความถูกต้อง (accuracy)
2. ค่าความแม่นยำ (precision)
3. ค่าความระลึก (recall)
4. ค่า F-score

เครื่องมือที่ใช้ในการวิจัย

ด้านโปรแกรมคอมพิวเตอร์ หรือซอฟต์แวร์

1 โปรแกรม Microsoft Excel เพื่อใช้ในการคัดกรองข้อมูล

2 โปรแกรม Weka version 3.9.6 เพื่อใช้ในการสร้างแบบจำลองทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม และเลือกอัลกอริทึมที่ใช้ในการจำแนกประเภท ดังนี้

1) เทคนิค Decision tree ใช้อัลกอริทึม J48
2) เทคนิค Random forest ใช้อัลกอริทึม RandomForest
3) เทคนิค Support Vector Machine ใช้อัลกอริทึม SMO
4) เทคนิค K – Nearest neighbor ใช้อัลกอริทึม IBk

บทที่ 4

ผลการวิเคราะห์ข้อมูล

วัตถุประสงค์ของงานวิจัย คือเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคที่ใช้ในการสร้างแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวานทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม โดยเทคนิคที่ใช้ในการสร้างแบบจำลอง ได้แก่ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor ซึ่งผู้วิจัยได้ประมวลผลข้อมูล โดยมีรายละเอียดและผลการวิเคราะห์ ดังนี้

4.1 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจข้อมูล

4.2 ผลการหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองการจำแนกทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

4.3 ผลการวิเคราะห์ในขั้นตอนการประเมินประสิทธิภาพของแบบจำลอง

ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจข้อมูล

ข้อมูลที่ใช้ในการวิจัยได้มาจากฐานข้อมูล (database) ของโรงพยาบาลในสังกัดสำนักการแพทย์ กรุงเทพมหานคร ซึ่งเป็นข้อมูลผลการตรวจของผู้เข้ารับบริการในโรงพยาบาลจำนวนทั้งสิ้น 20,227 ราย ประกอบไปด้วย เพศ อายุ น้ำหนัก ส่วนสูง ดัชนีมวลกาย ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลายตัว อัตราการเต้นของหัวใจ ประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง) และผลการตรวจโรคเบาหวาน โดยเก็บข้อมูลตั้งแต่ปี 2562 ถึง 2564 และสามารถนำเสนอให้อยู่ในรูปแบบของตารางและแผนภูมิ ดังต่อไปนี้

ตารางที่ 14 ข้อมูลผลการตรวจโรคเบาหวานของผู้รับบริการ

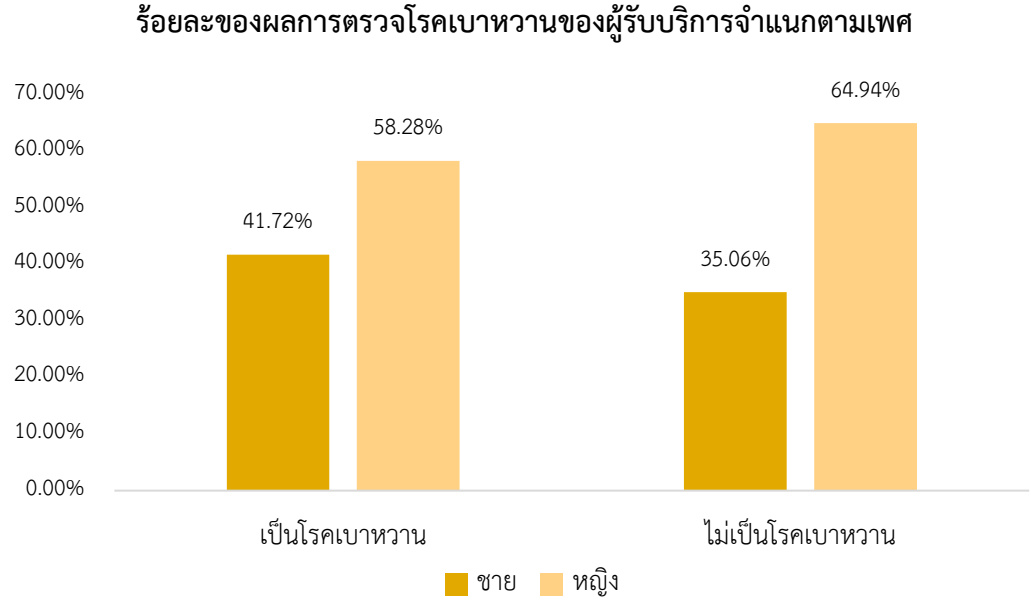
ผลการตรวจโรคเบาหวาน	จำนวน	ร้อยละ
เป็นโรคเบาหวาน	8,565	42.34
ไม่เป็นโรคเบาหวาน	11,662	57.66
รวม	**20,227**	**100**

จากตารางที่ 14 แสดงจำนวนผลการตรวจโรคเบาหวานของผู้รับบริการในโรงพยาบาลสังกัดสำนักการแพทย์ กรุงเทพมหานคร ตั้งแต่ปี 2562 ถึง 2564 พบว่า ผู้รับบริการที่มีผลการตรวจว่าไม่เป็นโรคเบาหวานมีร้อยละ 57.66 ซึ่งมากกว่าผู้รับบริการที่มีผลการตรวจว่าเป็นโรคเบาหวานซึ่งมีร้อยละ 42.34

ตารางที่ 15 ข้อมูลเพศของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

เพศ	จำนวน (ร้อยละ)		รวม
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน	
ชาย	3,573 (46.63)	4,089 (53.37)	**7,662 (100)**
หญิง	4,992 (39.73)	7,573 (60.27)	**12,565 (100)**

จากตารางที่ 15 แสดงจำนวนและร้อยละของเพศของผู้รับบริการจำนวน 20,227 ราย พบว่าผู้รับบริการเพศชายที่เป็นโรคเบาหวานร้อยละ 46.63 ซึ่งมากกว่าผู้รับบริการเพศหญิงที่เป็นโรคเบาหวานร้อยละ 39.73



ภาพที่ 21 แผนภูมิแสดงร้อยละของผลการตรวจโรคเบาหวานของผู้รับบริการจำแนกตามเพศ

จากแผนภูมิแสดงร้อยละของผลการตรวจโรคเบาหวานจำแนกตามเพศ เมื่อพิจารณาผลการตรวจโรคเบาหวาน พบว่าร้อยละของผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นเพศหญิงร้อยละ 58.28 และร้อยละของผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นเพศหญิงร้อยละ 64.94

ตารางที่ 16 ข้อมูลอายุของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	64	10.96	31	94
ไม่เป็นโรคเบาหวาน	62	12.33	31	97

จากตารางที่ 16 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของอายุของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยอายุของผู้รับบริการที่มีผลตรวจว่าเป็นโรคเบาหวาน คือ 64 ปี ($\bar{X}$ = 64 , S.D. = 10.96) ซึ่งมากกว่าค่าเฉลี่ยอายุของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 62 ปี ($\bar{X}$ = 62 , S.D. = 12.33) โดยที่ผู้รับบริการทั้งกลุ่มที่เป็นและไม่เป็นโรคเบาหวานมีอายุต่ำสุด 31 ปี และผู้รับบริการที่เป็นโรคเบาหวานมีอายุมากสุด 97 ปี ซึ่งมากกว่าผู้รับบริการที่ไม่เป็นโรคเบาหวานที่มีอายุมากสุด 94 ปี

ตารางที่ 17 ข้อมูลน้ำหนักของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	67	14.71	31	167
ไม่เป็นโรคเบาหวาน	62	13.69	30	167

จากตารางที่ 17 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของน้ำหนักของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยน้ำหนักของผู้รับบริการที่มีผลตรวจว่าเป็นโรคเบาหวาน คือ 67 กิโลกรัม ($\bar{X}$ = 67 , S.D. = 14.71) ซึ่งมากกว่าค่าเฉลี่ยน้ำหนักของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 62 กิโลกรัม ($\bar{X}$ = 62 , S.D. = 13.69) โดยที่ผู้รับบริการทั้งกลุ่มที่เป็นและไม่เป็นโรคเบาหวานมีน้ำหนักสูงสุด 167 กิโลกรัม และผู้รับบริการที่ไม่เป็นโรคเบาหวานมีน้ำหนักน้อยสุด 30 กิโลกรัม ซึ่งน้อยกว่าผู้รับบริการที่เป็นโรคเบาหวานที่มีน้ำหนักน้อยสุด 31 กิโลกรัม

ตารางที่ 18 ข้อมูลส่วนสูงของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	159	8.46	136	195
ไม่เป็นโรคเบาหวาน	158	8.49	100	195

จากตารางที่ 18 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของส่วนสูงของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยส่วนสูงของผู้รับบริการที่มีผลตรวจว่าเป็นโรคเบาหวาน คือ 159 เซนติเมตร ($\bar{X}$ = 159 , S.D. = 8.46) ซึ่งมากกว่าค่าเฉลี่ยส่วนสูงของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 158 เซนติเมตร ($\bar{X}$ = 158 , S.D. = 8.49) โดยที่ผู้รับบริการทั้งกลุ่มที่เป็นและไม่เป็นโรคเบาหวานมีส่วนสูงสูงสุด 195 เซนติเมตร และผู้รับบริการที่ไม่เป็นโรคเบาหวานมีส่วนสูงน้อยสุด 100 เซนติเมตร ซึ่งน้อยกว่าผู้รับบริการที่เป็นโรคเบาหวานที่มีส่วนสูงน้อยสุด 136 เซนติเมตร

ตารางที่ 19 ข้อมูลดัชนีมวลกายของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	26.16	4.97	12.66	45.79
ไม่เป็นโรคเบาหวาน	24.69	4.75	12.49	45.18

จากตารางที่ 19 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของดัชนีมวลกายของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยดัชนีมวลกายของผู้รับบริการที่มีผลตรวจว่าเป็นโรคเบาหวาน คือ 26.16 กิโลกรัมต่อตารางเมตร ($\bar{X}$ = 26.16 , S.D. = 4.97) ซึ่งมากกว่าค่าเฉลี่ยดัชนีมวลกายของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 24.69 กิโลกรัมต่อตารางเมตร ($\bar{X}$ = 24.69 , S.D. = 4.75) โดยที่ผู้รับบริการที่เป็นโรคเบาหวานมีดัชนีมวลกายมากสุด 45.79 กิโลกรัมต่อตารางเมตร ซึ่งมากกว่าผู้รับบริการที่ไม่เป็นโรคเบาหวานที่มีดัชนีมวลกายมากสุด 45.18 กิโลกรัมต่อตารางเมตร และผู้รับบริการที่ไม่เป็นโรคเบาหวานมีดัชนีมวลกายน้อยสุด 12.49 กิโลกรัมต่อตารางเมตร ซึ่งน้อยกว่าผู้รับบริการที่เป็นโรคเบาหวานที่มีดัชนีมวลกายน้อยสุด 12.66 กิโลกรัมต่อตารางเมตร

ตารางที่ 20 ข้อมูลความดันขณะหัวใจคลายตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	71.83	12.49	35	150
ไม่เป็นโรคเบาหวาน	73.90	11.86	37	128

จากตารางที่ 20 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของความดันขณะหัวใจคลายตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยความดันขณะหัวใจคลายตัวของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 73.90 มิลลิเมตรปรอท ($\bar{X}$ = 73.90 , S.D. = 11.86) ซึ่งมากกว่าค่าเฉลี่ยความดันขณะหัวใจคลายตัวของผู้รับบริการที่มีผลตรวจว่าเป็นโรคเบาหวาน คือ 71.83 มิลลิเมตรปรอท ($\bar{X}$ = 71.83 , S.D. = 12.49) โดยที่ผู้รับบริการที่เป็นโรคเบาหวานมีความดันขณะหัวใจคลายตัวสูงสุด 150 มิลลิเมตรปรอท ซึ่งมากกว่าผู้รับบริการที่ไม่เป็นโรคเบาหวานที่มีความดันขณะหัวใจคลายตัวมากสุด 128 มิลลิเมตรปรอท และผู้รับบริการที่เป็นโรคเบาหวานมีความดันขณะหัวใจคลายตัวต่ำสุด 35 มิลลิเมตรปรอท ซึ่งต่ำกว่าผู้รับบริการที่ไม่เป็นโรคเบาหวานที่มีความดันขณะหัวใจคลายตัวต่ำสุด 37 มิลลิเมตรปรอท

ตารางที่ 21 ข้อมูลความดันขณะหัวใจบีบตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	135.56	18.33	80	237
ไม่เป็นโรคเบาหวาน	132.51	17.70	77	198

จากตารางที่ 21 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของความดันขณะหัวใจบีบตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยความดันขณะหัวใจบีบตัวของผู้รับบริการที่เป็นโรคเบาหวาน คือ 135.56 มิลลิเมตรปรอท ($\bar{X}$ = 135.56 , S.D. = 18.33) ซึ่งมากกว่าค่าเฉลี่ยความดันขณะหัวใจบีบตัวของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 132.51 มิลลิเมตรปรอท ($\bar{X}$ = 132.51 , S.D. = 17.70) โดยที่ผู้รับบริการที่เป็นโรคเบาหวานมีความดันขณะหัวใจบีบตัวสูงสุด 237 มิลลิเมตรปรอท ซึ่งมากกว่าผู้รับบริการที่ไม่เป็นโรคเบาหวานที่มีความดันขณะหัวใจบีบตัวมากสุด 198 มิลลิเมตรปรอท และผู้รับบริการที่ไม่เป็นโรคเบาหวานมีความดัน

ขณะหัวใจบีบตัวต่ำสุด 77 มิลลิเมตรปรอท ซึ่งต่ำกว่าผู้รับบริการที่เป็นโรคเบาหวานที่มีความดันขณะหัวใจบีบตัวต่ำสุด 80 มิลลิเมตรปรอท

ตารางที่ 22 ข้อมูลอัตราการเต้นของหัวใจของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

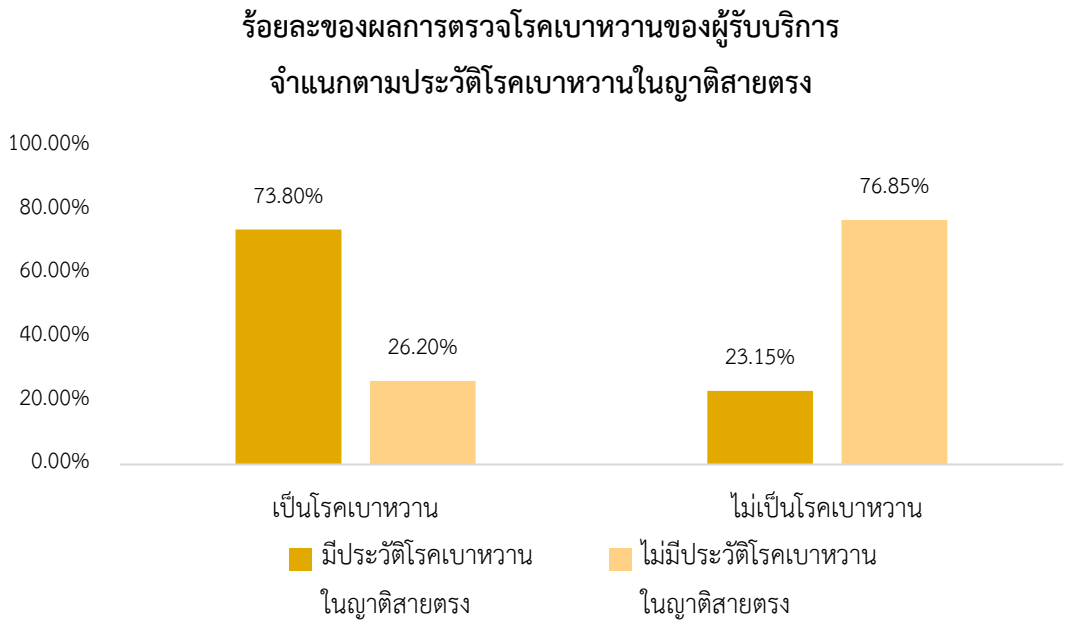
ผลการตรวจโรคเบาหวาน	ค่าเฉลี่ย $(\bar{x})$	ส่วนเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด	ค่าสูงสุด
เป็นโรคเบาหวาน	83.20	14.09	36	151
ไม่เป็นโรคเบาหวาน	82.43	13.59	36	161

จากตารางที่ 22 แสดงข้อมูลค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าต่ำสุด และค่าสูงสุดของอัตราการเต้นของหัวใจของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน พบว่าค่าเฉลี่ยอัตราการเต้นของหัวใจของผู้รับบริการที่เป็นโรคเบาหวาน คือ 83.20 ครั้งต่อนาที ($\bar{\mathbf{X}}$ = 83.20 , S.D. = 14.09) ซึ่งมากกว่าค่าเฉลี่ยอัตราการเต้นของหัวใจของผู้รับบริการที่ไม่เป็นโรคเบาหวาน คือ 82.43 ครั้งต่อนาที ($\bar{\mathbf{X}}$ = 82.43 , S.D. = 13.59) โดยที่ผู้รับบริการทั้งกลุ่มที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวานมีอัตราการเต้นของหัวใจตัวต่ำสุด 36 ครั้งต่อนาที และผู้รับบริการที่ไม่เป็นโรคเบาหวานมีอัตราการเต้นของหัวใจสุงสุด 161 ครั้งต่อนาที ซึ่งสูงกว่าผู้รับบริการที่เป็นโรคเบาหวานที่มีอัตราการเต้นของหัวใจสูงสุด 151 ครั้งต่อนาที

ตารางที่ 23 ข้อมูลประวัติโรคเบาหวานในญาติสายตรงของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง	จำนวน (ร้อยละ)		รวม
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน	
มี	6,321 (70.07)	2,700 (29.93)	**567 (100)**
ไม่มี	2,244 (20.02)	8,962 (79.98)	**17,695 (100)**

ตารางที่ 23 แสดงจำนวนและร้อยละของประวัติโรคเบาหวานในญาติสายตรงของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงที่เป็นโรคเบาหวานร้อยละ 70.07 ซึ่งมากกว่าผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงที่เป็นโรคเบาหวานร้อยละ 20.02



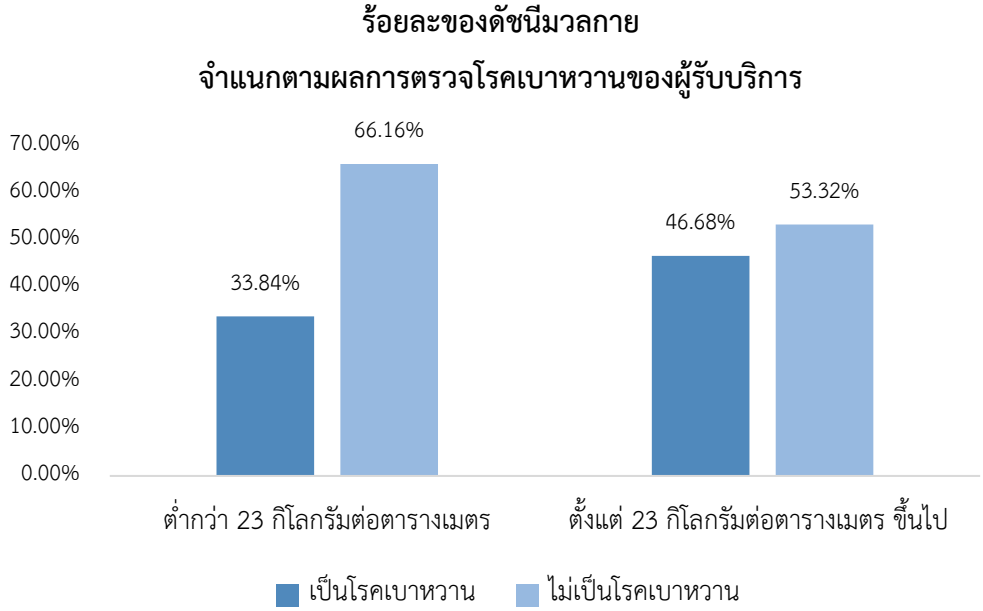
ภาพที่ 22 แผนภูมิแสดงร้อยละของผลการตรวจโรคเบาหวานของผู้รับบริการ จำแนกตามประวัติโรคเบาหวานในญาติสายตรง

จากแผนภูมิแสดงร้อยละของผลการตรวจโรคเบาหวานจำแนกตามประวัติโรคเบาหวานในญาติสายตรง เมื่อพิจารณาผลการตรวจโรคเบาหวาน พบว่าร้อยละของผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้ที่มีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 73.80 และร้อยละของผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้ที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 76.85

กรณีที่พิจารณาอิทธิพลร่วม

จากการศึกษาของ (Techasuwanna et al., 2020) ซึ่งศึกษาปัจจัยเสี่ยงที่ส่งผลต่อการเป็นโรคเบาหวาน พบว่ามีปัจจัยเสี่ยงดังนี้ ผู้ที่มีดัชนีมวลกายมากเกินเกณฑ์มาตรฐาน อายุมาก ความดันโลหิตสูง ไขมัน HDL ต่ำ และไตรกลีเซอไรด์สูง ซึ่งส่งผลทำให้อัตราการเต้นของหัวใจผิดปกติตามมา ดังนั้นในงานวิจัยนี้จึงพิจารณาอิทธิพลร่วมของปัจจัยเสี่ยงดังกล่าว ดังนี้

1. การพิจารณาปัจจัยดัชนีมวลกายแบ่งเป็น 2 กลุ่ม[1] ตามเกณฑ์การแปลผลค่าดัชนีมวลกาย คือ กลุ่มที่มีค่าดัชนีมวลกายเกินเกณฑ์ (ค่าดัชนีมวลกายตั้งแต่ 23 กิโลกรัมต่อตารางเมตร ขึ้นไป) และกลุ่มที่มีค่าดัชนีมวลกายปกติ (ค่าดัชนีมวลกายน้อยกว่า 23 กิโลกรัมต่อตาราง)

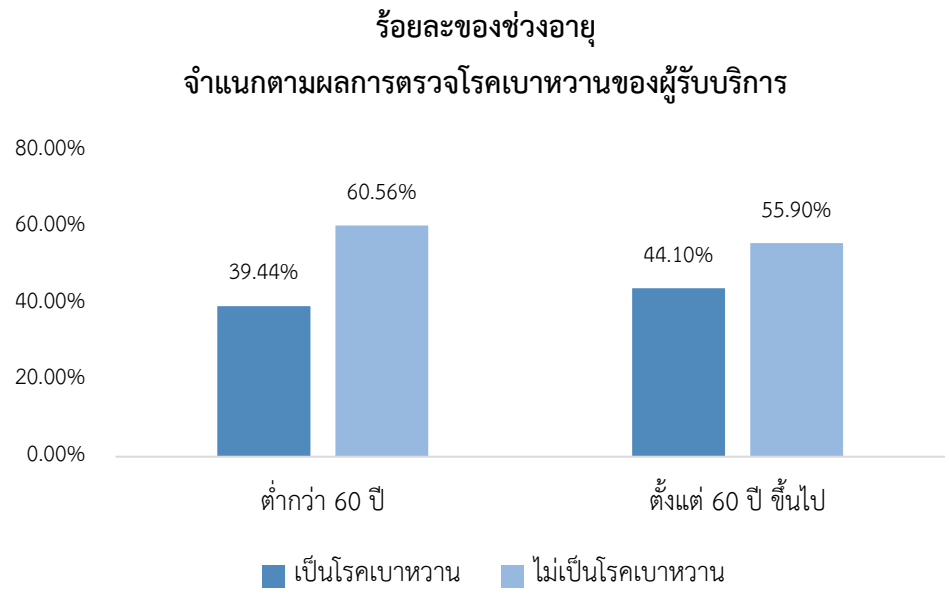


ภาพที่ 23 แผนภูมิแสดงร้อยละของดัชนีมวลกายจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ

จากแผนภูมิแสดงร้อยละของดัชนีมวลกายจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการพบว่าผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์ที่เป็นโรคเบาหวานร้อยละ 46.68 ซึ่งมากกว่าผู้รับบริการที่มีดัชนีมวลปกติที่เป็นโรคเบาหวานร้อยละ 33.84

[1] (นันท์พัสพร สุขสานต์ & จิราพร เกศพิชญวัฒนา, 2560)

2. การพิจารณาปัจจัยอายุแบ่งเป็น 2 กลุ่ม[2] คือ กลุ่มผู้สูงอายุ (อายุตั้งแต่ 60 ปีขึ้นไป) และกลุ่มที่ไม่ใช่ผู้สูงอายุ (อายุน้อยกว่า 60 ปี) โดยอ้างอิงจากการศึกษาของ (วิชัย เอกพลากร, 2557) ซึ่งพบว่าโรคเบาหวานพบบ่อยในผู้สูงอายุที่มีอายุตั้งแต่ 60 ปีขึ้นไป เนื่องจากมีการเสื่อมของตับอ่อนที่ทำหน้าที่ในการผลิตฮอร์โมนอินสุลินที่ใช้ในการควบคุมระดับน้ำตาลในเลือด

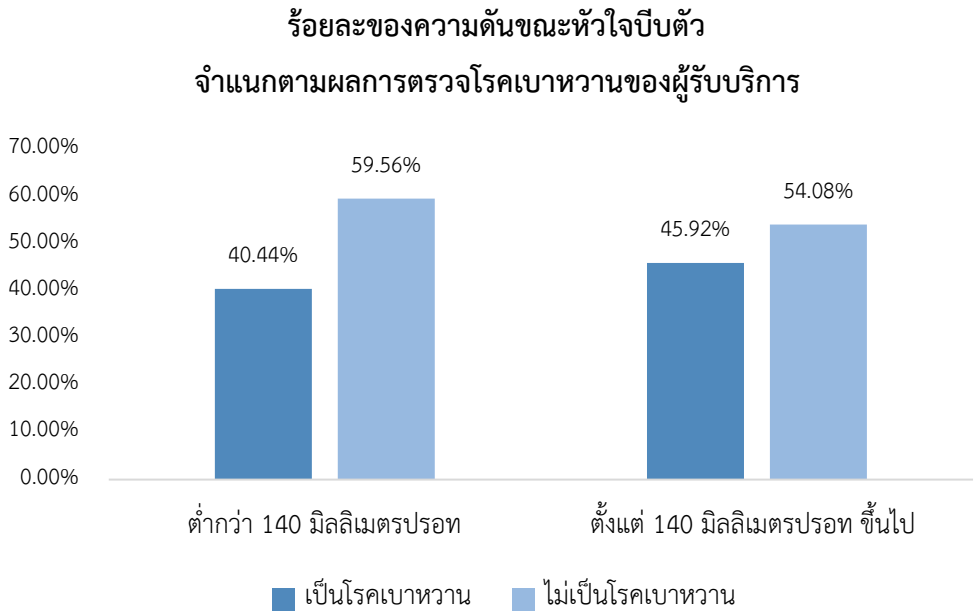


ภาพที่ 24 แผนภูมิแสดงร้อยละของช่วงอายุจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ

จากแผนภูมิแสดงร้อยละของช่วงอายุจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ พบว่าผู้รับบริการที่เป็นผู้สูงอายุที่เป็นโรคเบาหวานร้อยละ 44.10 ซึ่งมากกว่าผู้รับบริการที่ไม่ใช่ผู้สูงอายุที่เป็นโรคเบาหวานร้อยละ 39.44

[2] (นันท์พัสพร สุขสานต์ & จิราพร เกศพิชญวัฒนา, 2560)

3. การพิจารณาปัจจัยความดันขณะหัวใจบีบตัวแบ่งเป็น 2 กลุ่ม[3] ตามเกณฑ์การแปลผลค่าความดันขณะหัวใจบีบตัว คือ กลุ่มที่มีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ (ค่าความดันขณะหัวใจบีบตัวตั้งแต่ 140 มิลลิเมตรปรอท ขึ้นไป) และกลุ่มที่มีค่าความดันขณะหัวใจบีบตัวปกติ (ค่าความดันขณะหัวใจบีบตัวน้อยกว่า 140 มิลลิเมตรปรอท)

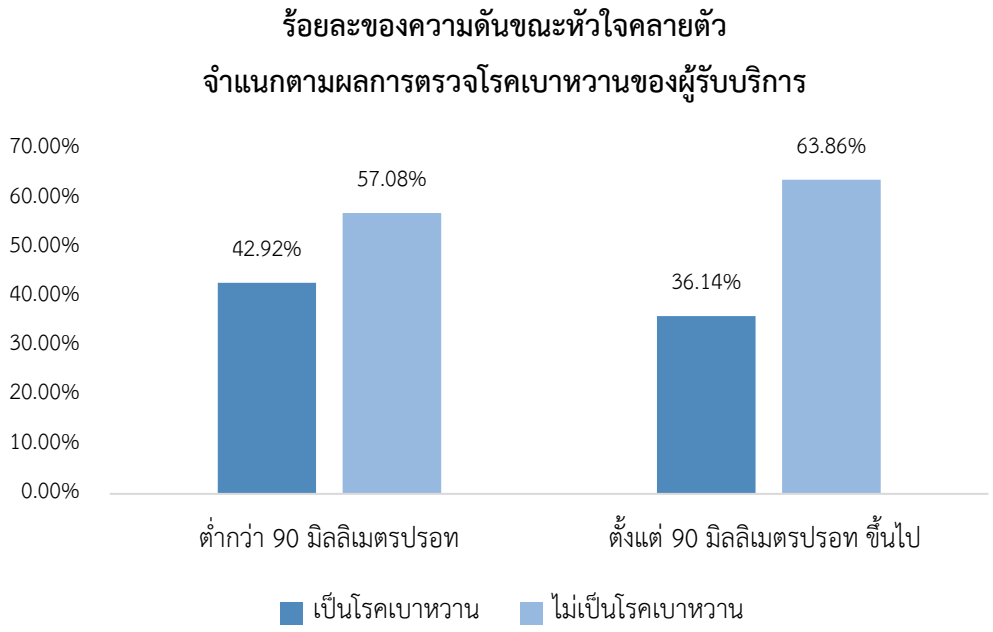


ภาพที่ 25 แผนภูมิแสดงร้อยละของความดันขณะหัวใจบีบตัว
จำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ

จากแผนภูมิแสดงร้อยละของความดันขณะหัวใจบีบตัวจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการพบว่าผู้รับบริการที่มีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ที่เป็นโรคเบาหวานร้อยละ 45.92 ซึ่งมากกว่าผู้รับบริการที่มีค่าความดันขณะหัวใจบีบตัวปกติที่เป็นโรคเบาหวานร้อยละ 40.44

[3] (สมาคมความดันโลหิตสูงแห่งประเทศไทย, 2562)

4. การพิจารณาปัจจัยความดันขณะหัวใจคลายตัวแบ่งเป็น 2 กลุ่ม[4] ตามเกณฑ์การแปลผลค่าความดันขณะหัวใจคลายตัว คือ กลุ่มที่มีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ (ค่าความดันขณะหัวใจคลายตัวตั้งแต่ 90 มิลลิเมตรปรอท ขึ้นไป) และกลุ่มที่มีค่าความดันขณะหัวใจคลายตัวปกติ (ค่าความดันขณะหัวใจคลายตัวน้อยกว่า 90 มิลลิเมตรปรอท)

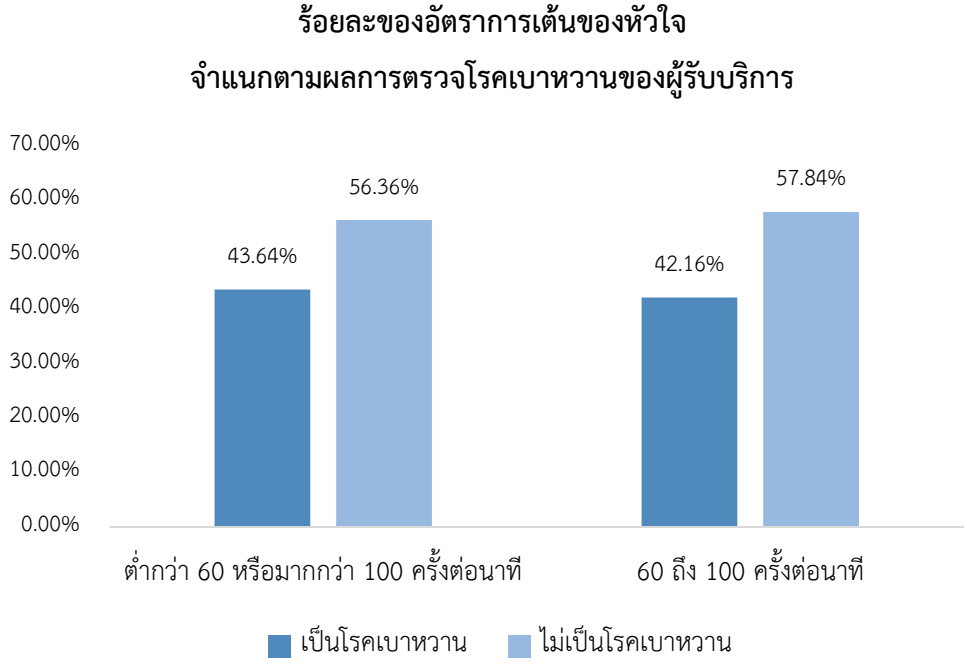


ภาพที่ 26 แผนภูมิแสดงร้อยละของความดันขณะหัวใจคลายตัว
จำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ

จากแผนภูมิแสดงร้อยละของความดันขณะหัวใจคลายตัวจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ พบว่าผู้รับบริการที่มีค่าความดันขณะหัวใจคลายตัวปกติที่เป็นโรคเบาหวานร้อยละ 42.92 ซึ่งมากกว่าผู้รับบริการที่มีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ที่เป็นโรคเบาหวานร้อยละ 36.14

[4] (สมาคมความดันโลหิตสูงแห่งประเทศไทย, 2562)

5. การพิจารณาปัจจัยอัตราการเต้นของหัวใจแบ่งเป็น 2 กลุ่ม[5] ตามเกณฑ์การแปลผลค่าอัตราการเต้นของหัวใจ คือ กลุ่มที่มีค่าอัตราการเต้นของหัวใจผิดปกติ (ค่าอัตราการเต้นของหัวใจน้อยกว่า 60 ครั้งต่อนาที หรือ มากกว่า 100 ครั้งต่อนาที) และกลุ่มที่มีค่าอัตราการเต้นของหัวใจปกติ (ค่าอัตราการเต้นของหัวใจระหว่าง 60 ถึง 100 ครั้งต่อนาที)



ภาพที่ 27 แผนภูมิแสดงร้อยละของอัตราการเต้นของหัวใจ
จำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ

จากแผนภูมิแสดงร้อยละของอัตราการเต้นของหัวใจจำแนกตามผลการตรวจโรคเบาหวานของผู้รับบริการ พบว่าผู้รับบริการที่มีอัตราการเต้นของหัวใจผิดปกติที่เป็นโรคเบาหวานร้อยละ 43.64 ซึ่งมากกว่าผู้รับบริการที่มีค่าอัตราการเต้นของหัวใจปกติที่เป็นโรคเบาหวานร้อยละ 42.16

[5] (โรงพยาบาลรามคำแหง, 2563)

สำหรับกรณีที่พิจารณาอิทธิพลร่วมสามารถนำเสนอให้อยู่ในรูปแบบของตารางและแผนภูมิ ได้ดังนี้

1. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและประวัติโรคเบาหวานในญาติสายตรง

ตารางที่ 24 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและประวัติโรคเบาหวานในญาติสายตรงของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*ประวัติโรคเบาหวานในญาติสายตรง	**จำนวน (ร้อยละ)**	
	เป็นโรคเบาหวาน	**ไม่เป็นโรคเบาหวาน**
ดัชนีมวลกายปกติและไม่มีประวัติโรคเบาหวานในญาติสายตรง	2,112 (24.66)	5,381 (46.14)
ดัชนีมวลกายปกติและมีประวัติโรคเบาหวานในญาติสายตรง	1,748 (20.41)	1,433 (12.29)
ดัชนีมวลกายเกินเกณฑ์และไม่มีประวัติโรคเบาหวานในญาติสายตรง	132 (1.54)	3,581 (30.71)
ดัชนีมวลกายเกินเกณฑ์และมีประวัติโรคเบาหวานในญาติสายตรง	4,573 (53.39)	1,267 (10.86)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 24 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและประวัติโรคเบาหวานในญาติสายตรงของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์ร่วมกับการมีประวัติโรคเบาหวานในญาติสายตรง ซึ่งมีร้อยละ 53.39 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายปกติร่วมกับการไม่มีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 24.66 ผู้รับบริการที่มีดัชนีมวลกายปกติร่วมกับการมีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 20.41 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์ร่วมกับการไม่มีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 1.54 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติร่วมกับการไม่มีประวัติโรคเบาหวานในญาติสายตรง ซึ่งมีร้อยละ 46.14 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์ร่วมกับการไม่มีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 30.71 ผู้รับบริการที่มีดัชนีมวลกายปกติร่วมกับการมีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 12.29 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์ร่วมกับการมีประวัติโรคเบาหวานในญาติสายตรงร้อยละ 10.86 ตามลำดับ

2. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและเพศ

ตารางที่ 25 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและเพศของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*เพศ	**จำนวน (ร้อยละ)**	
	เป็นโรคเบาหวาน	**ไม่เป็นโรคเบาหวาน**
ดัชนีมวลกายปกติและเพศชาย	1,634 (19.08)	2,451 (21.02)
ดัชนีมวลกายปกติและเพศหญิง	2,226 (25.99)	4,363 (37.41)
ดัชนีมวลกายเกินเกณฑ์และเพศชาย	1,939 (22.64)	1,638 (14.04)
ดัชนีมวลกายเกินเกณฑ์และเพศหญิง	2,766 (32.29)	3,210 (27.52)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 25 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและเพศของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเพศหญิง ซึ่งมีร้อยละ 32.29 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นเพศหญิง ร้อยละ 25.99 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเพศชาย ร้อยละ 22.64 และผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นเพศชายร้อยละ 19.08 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นเพศหญิง ซึ่งมีร้อยละ 37.41 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเพศหญิงร้อยละ 27.52 ผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นเพศชายร้อยละ 21.02 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเพศชายร้อยละ 14.04 ตามลำดับ

3. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและอายุ

ตารางที่ 26 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและอายุของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*อายุ	**จำนวน (ร้อยละ)**	
	เป็นโรคเบาหวาน	**ไม่เป็นโรคเบาหวาน**
ดัชนีมวลกายปกติและไม่ใช่ผู้สูงอายุ	1,040 (12.14)	2,438 (20.91)
ดัชนีมวลกายปกติและเป็นผู้สูงอายุ	2,738 (31.97)	4,376 (37.52)
ดัชนีมวลกายเกินเกณฑ์และไม่ใช่ผู้สูงอายุ	1,967 (22.97)	2,179 (18.68)
ดัชนีมวลกายเกินเกณฑ์และเป็นผู้สูงอายุ	2,820 (32.92)	2,669 (22.88)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 26 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและอายุของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเป็นผู้สูงอายุ ซึ่งมีร้อยละ 32.92 รองลงมาคือ ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นผู้สูงอายุร้อยละ 31.97 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และไม่ใช่ผู้สูงอายุ ร้อยละ 22.97 และผู้รับบริการที่มีดัชนีมวลกายปกติและไม่ใช่ผู้สูงอายุร้อยละ 12.14 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและเป็นผู้สูงอายุ ซึ่งมีร้อยละ 37.52 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และเป็นเป็นผู้สูงอายุร้อยละ 22.88 ผู้รับบริการที่มีดัชนีมวลกายปกติและไม่ใช่ผู้สูงอายุร้อยละ 20.91 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และไม่ใช่ผู้สูงอายุร้อยละ 18.68 ตามลำดับ

4. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะหัวใจบีบตัว

ตารางที่ 27 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะหัวใจบีบตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*ค่าความดันขณะหัวใจบีบตัว	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ดัชนีมวลกายปกติและค่าความดันขณะหัวใจบีบตัวปกติ	2,735 (31.93)	5,164 (44.28)
ดัชนีมวลกายปกติและค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์	1,125 (13.13)	1,650 (14.15)
ดัชนีมวลกายเกินเกณฑ์และค่าความดันขณะหัวใจบีบตัวปกติ	2,972 (34.70)	3,112 (26.68)
ดัชนีมวลกายเกินเกณฑ์และค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์	1,733 (20.23)	1,736 (14.89)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 27 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะหัวใจบีบตัวของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจบีบตัวปกติ ซึ่งมีร้อยละ 34.70 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจบีบตัวปกติร้อยละ 31.93 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 20.23 และผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 13.13 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจบีบตัวปกติ ซึ่งมีร้อยละ 44.28 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจบีบตัวปกติร้อยละ 26.68 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 14.89 และผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 14.15 ตามลำดับ

5. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะหัวใจคลายตัว

ตารางที่ 28 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะหัวใจคลายตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*ค่าความดันขณะหัวใจคลายตัว	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ดัชนีมวลกายปกติและค่าความดันขณะหัวใจคลายตัวปกติ	3,682 (42.99)	6,338 (54.35)
ดัชนีมวลกายปกติและค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์	178 (2.08)	476 (4.08)
ดัชนีมวลกายเกินเกณฑ์และค่าความดันขณะหัวใจคลายตัวปกติ	4,329 (50.54)	4,373 (37.50)
ดัชนีมวลกายเกินเกณฑ์และค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์	376 (4.39)	475 (4.07)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 28 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าความดันขณะคลายตัวของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจคลายตัวปกติ ซึ่งมีร้อยละ 50.54 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจคลายตัวปกติร้อยละ 42.99 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 4.39 และผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 2.08 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจคลายตัวปกติ ซึ่งมีร้อยละ 54.35 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจคลายตัวปกติร้อยละ 37.50 ผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 4.08 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 4.07 ตามลำดับ

6. การพิจารณาอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าอัตราการเต้นของหัวใจ

ตารางที่ 29 ข้อมูลอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าอัตราการเต้นของหัวใจของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ดัชนีมวลกาย*อัตราการเต้นของหัวใจ	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ดัชนีมวลกายปกติและค่าอัตราการเต้นของหัวใจปกติ	3,398 (39.66)	6,001 (51.47)
ดัชนีมวลกายปกติและค่าอัตราการเต้นของหัวใจผิดปกติ	465 (5.43)	810 (6.95)
ดัชนีมวลกายเกินเกณฑ์และค่าอัตราการเต้นของหัวใจปกติ	4,065 (47.44)	4,234 (36.32)
ดัชนีมวลกายเกินเกณฑ์และอัตราการเต้นของหัวใจผิดปกติ	640 (7.47)	614 (5.27)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 29 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างดัชนีมวลกายและค่าอัตราการเต้นของหัวใจของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าอัตราการเต้นของหัวใจปกติ ซึ่งมีร้อยละ 47.44 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าอัตราการเต้นของหัวใจปกติละ 39.66 ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีอัตราการเต้นของหัวใจผิดปกติร้อยละ 7.47 และผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 5.43 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าอัตราการเต้นของหัวใจปกติ ซึ่งมีร้อยละ 51.47 รองลงมาคือ ผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีค่าอัตราการเต้นของหัวใจปกติร้อยละ 36.32 ผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 6.95 และผู้รับบริการที่มีดัชนีมวลกายเกินเกณฑ์และมีอัตราการเต้นของหัวใจผิดปกติร้อยละ 5.27 ตามลำดับ

7. การพิจารณาอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและเพศ

ตารางที่ 30 ข้อมูลอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและเพศของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง*เพศ	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศชาย	2,662 (31.08)	868 (7.44)
มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิง	3,659 (42.72)	1,832 (15.71)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศชาย	911 (10.64)	3,221 (27.62)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิง	1,333 (15.56)	5,741 (49.23)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 30 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและเพศของผู้รับบริการจำนวน 20,227 ราย พบว่าผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิง ซึ่งมีร้อยละ 42.72 รองลงมาคือ ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศชายร้อยละ 31.08 ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิงร้อยละ 15.56 และผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิงร้อยละ 10.64 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีดัชนีมวลกายปกติและมีค่าอัตราการเต้นของหัวใจปกติ ซึ่งมีร้อยละ 49.23 รองลงมาคือ ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศชายร้อยละ 27.62 ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศหญิงร้อยละ 15.71 และผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นเพศชายร้อยละ 7.44 ตามลำดับ

8. การพิจารณาอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและอายุ

ตารางที่ 31 ข้อมูลอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและอายุของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง*อายุ	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุ	4,094 (47.80)	1,630 (13.98)
มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุ	2,227 (26.00)	1,070 (9.18)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุ	1,464 (17.09)	5,415 (46.43)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุ	780 (9.11)	3,547 (30.42)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 31 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและอายุของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุ ซึ่งมีร้อยละ 47.80 รองลงมาคือ ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุร้อยละ 26 ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุร้อยละ 17.09 และผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุร้อยละ 9.11 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุ ซึ่งมีร้อยละ 46.43 รองลงมาคือ ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุร้อยละ 30.42 ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและเป็นผู้สูงอายุร้อยละ 13.98 และผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและไม่ใช่ผู้สูงอายุร้อยละ 9.18 ตามลำดับ

9. การพิจารณาอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัว

ตารางที่ 32 ข้อมูลอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง*ค่าความดันขณะหัวใจบีบตัว	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวปกติ	4,096 (47.82)	1,887 (16.18)
มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์	2,225 (25.98)	813 (6.97)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวปกติ	1,611 (18.81)	6,389 (54.78)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์	633 (7.39)	2,573 (22.06)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 32 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจบีบตัวของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวปกติ ซึ่งมีร้อยละ 47.82 รองลงมาคือ ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 25.98 ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวปกติร้อยละ 18.81 และผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 7.39 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวปกติ ซึ่งมีร้อยละ 54.78 รองลงมาคือ ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 22.06 ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวปกติร้อยละ 16.18 และ

ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจบีบตัวเกินเกณฑ์ร้อยละ 6.97 ตามลำดับ

10. การพิจารณาอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัว

ตารางที่ 33 ข้อมูลอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง*ค่าความดันขณะหัวใจคลายตัว	จำนวน (ร้อยละ)	
	เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวปกติ	5,878 (68.63)	2,472 (21.20)
มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์	443 (5.17)	228 (1.96)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวปกติ	2,133 (24.90)	8,239 (70.65)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์	111 (1.30)	723 (6.20)
รวม	**8,565 (100)**	**11,662 (100)**

ตารางที่ 33 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าความดันขณะหัวใจคลายตัวของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวปกติ ซึ่งมีร้อยละ 68.63 รองลงมาคือ ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจ คลายตัวปกติร้อยละ 24.90 ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 5.17 และผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและ มีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 1.30 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวปกติ ซึ่งมีร้อยละ 70.65 รองลงมาคือ ผู้รับบริการ

ที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวปกติ ร้อยละ 21.20 ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ ร้อยละ 6.20 และผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและมีค่าความดันขณะหัวใจคลายตัวเกินเกณฑ์ร้อยละ 1.96 ตามลำดับ

11. การพิจารณาอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจ

ตารางที่ 34 ข้อมูลอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจของผู้รับบริการจำแนกตามผลการตรวจโรคเบาหวาน

ประวัติโรคเบาหวานในญาติสายตรง*ค่าอัตราการเต้นของหัวใจ	**จำนวน (ร้อยละ)**	
	เป็นโรคเบาหวาน	**ไม่เป็นโรคเบาหวาน**
มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติ	5,460 (63.75)	2,389 (20.49)
มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติ	861 (10.05)	311 (2.67)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติ	2,000 (23.35)	7,846 (67.28)
ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติ	244 (2.85)	1,116 (9.57)
รวม	**8,565 (100)**	**11,662 (100)**

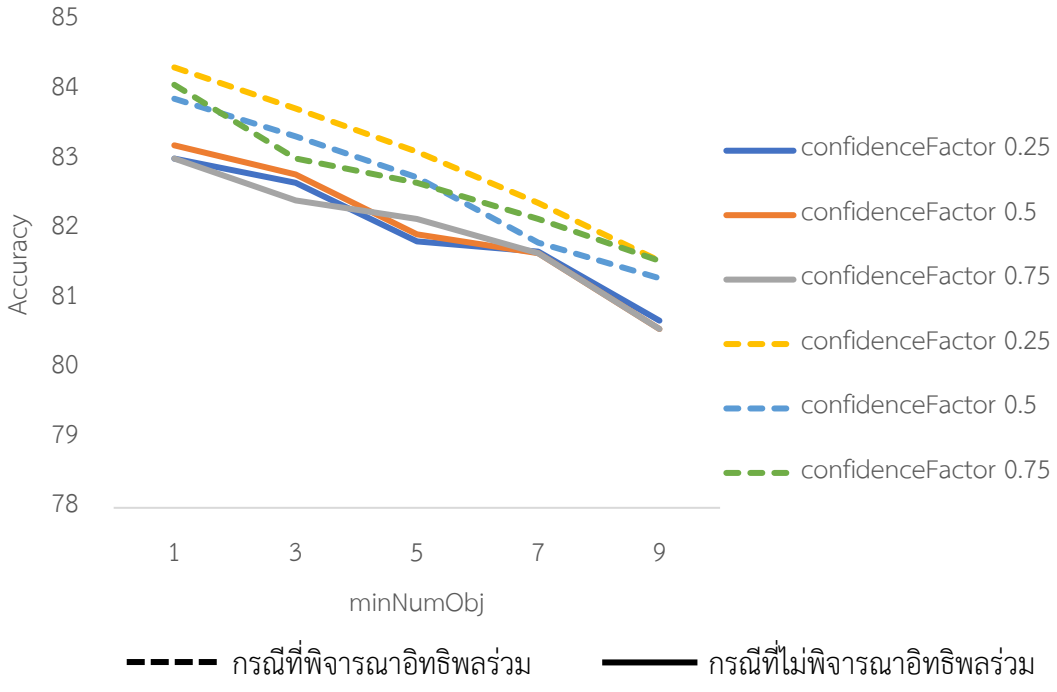
ตารางที่ 34 แสดงจำนวนและร้อยละของอิทธิพลร่วมระหว่างประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจของผู้รับบริการจำนวน 20,227 ราย พบว่า ผู้รับบริการที่เป็นโรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติ ซึ่งมีร้อยละ 63.75 รองลงมาคือ ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติร้อยละ 23.35 ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 10.05 และผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 2.85 ตามลำดับ ส่วนผู้รับบริการที่ไม่เป็น

โรคเบาหวานส่วนใหญ่เป็นผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติ ซึ่งมีร้อยละ 67.28 รองลงมาคือ ผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจปกติร้อยละ 20.49 ผู้รับบริการที่ไม่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 9.57 และผู้รับบริการที่มีประวัติโรคเบาหวานในญาติสายตรงและค่าอัตราการเต้นของหัวใจผิดปกติร้อยละ 2.67 ตามลำดับ

ผลการหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองการจำแนกทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

1. เทคนิค Decision tree

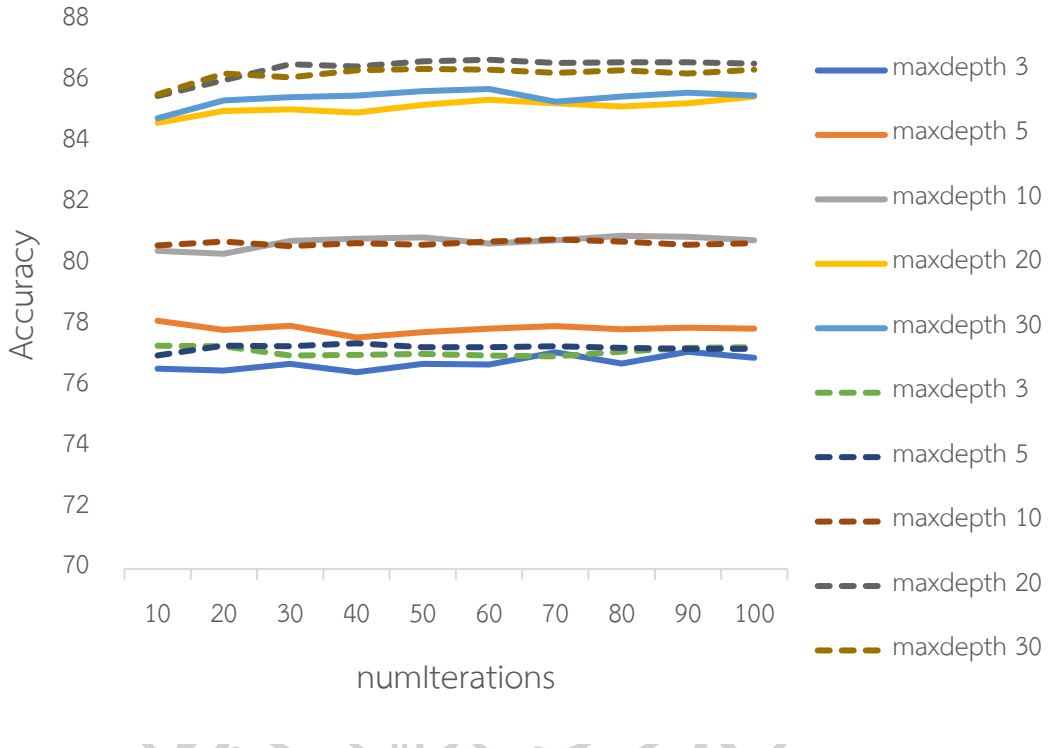
เมื่อพิจารณาค่าความถูกต้องจากการตรวจสอบแบบไขว้กรณีที่ไม่พิจารณาอิทธิพลร่วม จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุด คือ confidenceFactor = 0.5 และ minNumObj = 1 และกรณีที่พิจารณาอิทธิพลร่วม จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุด คือ confidenceFactor = 0.25 และ minNumObj = 1 ดังภาพที่ 28



ภาพที่ 28 ค่าความถูกต้องของแต่ละค่า minNumObj ที่แตกต่างกัน
เมื่อกำหนดค่า confidenceFactor ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

2. เทคนิค Random forest

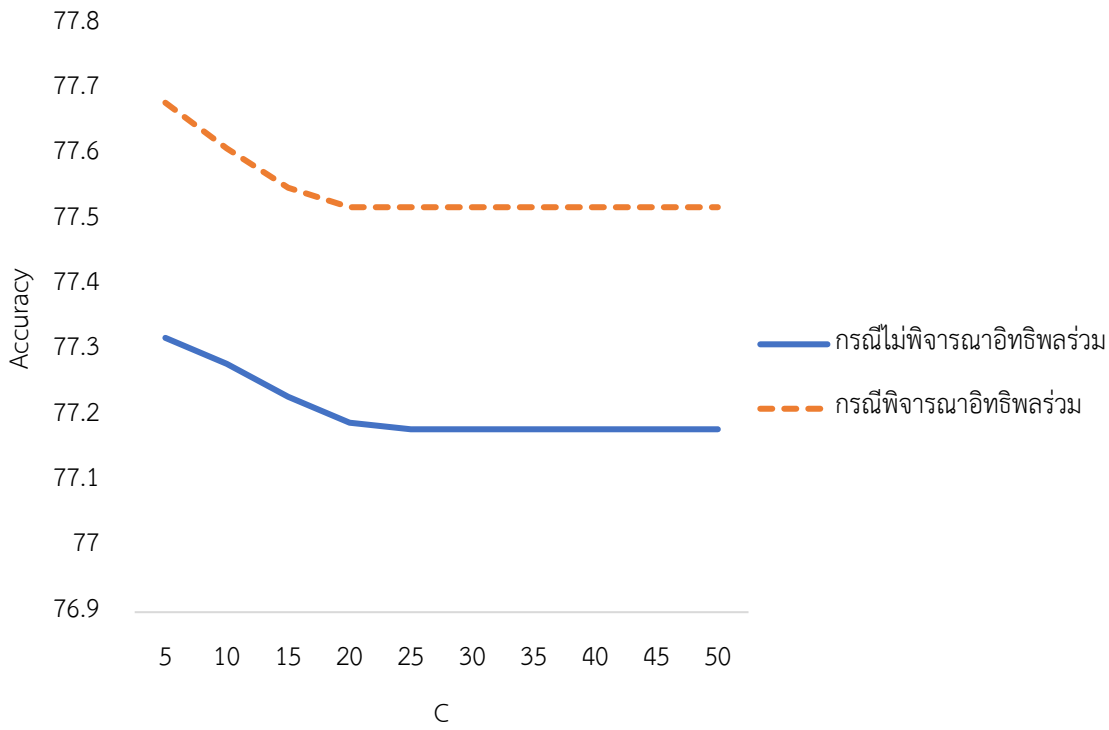
เมื่อพิจารณาค่าความถูกต้องจากการตรวจสอบแบบไขว้กรณีที่ไม่พิจารณาอิทธิพลร่วม จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุด คือ numIterations = 60 และ maxDepth = 30 และกรณีที่พิจารณาอิทธิพลร่วม จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุด คือ numIterations = 60 และ maxDepth = 20 ดังภาพที่ 29



ภาพที่ 29 ค่าความถูกต้องของแต่ละค่า numIterations ที่แตกต่างกัน เมื่อกำหนดค่า maxDepth ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

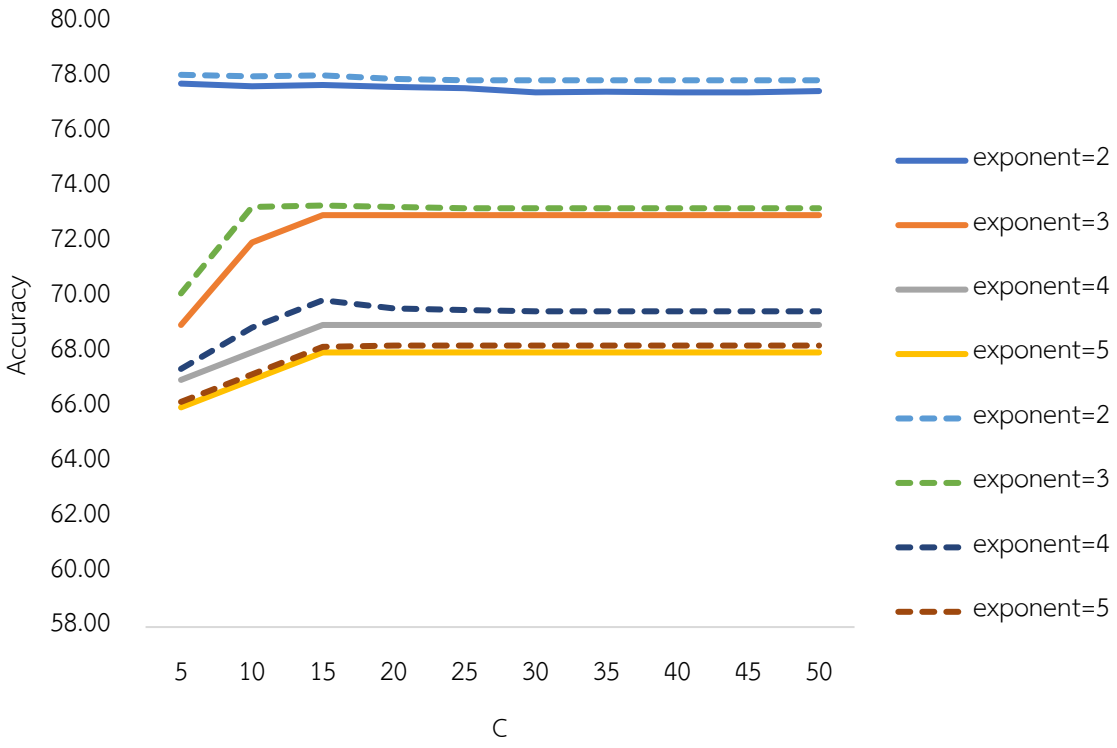
3. เทคนิค Support Vector Machine

จากการค้นหาแบบกริด เมื่อพิจารณาค่าความถูกต้องจากการตรวจสอบแบบไขว้ พบว่า kernel ที่แตกต่างกัน จะมีค่า C ที่เหมาะที่สุดแตกต่างกัน



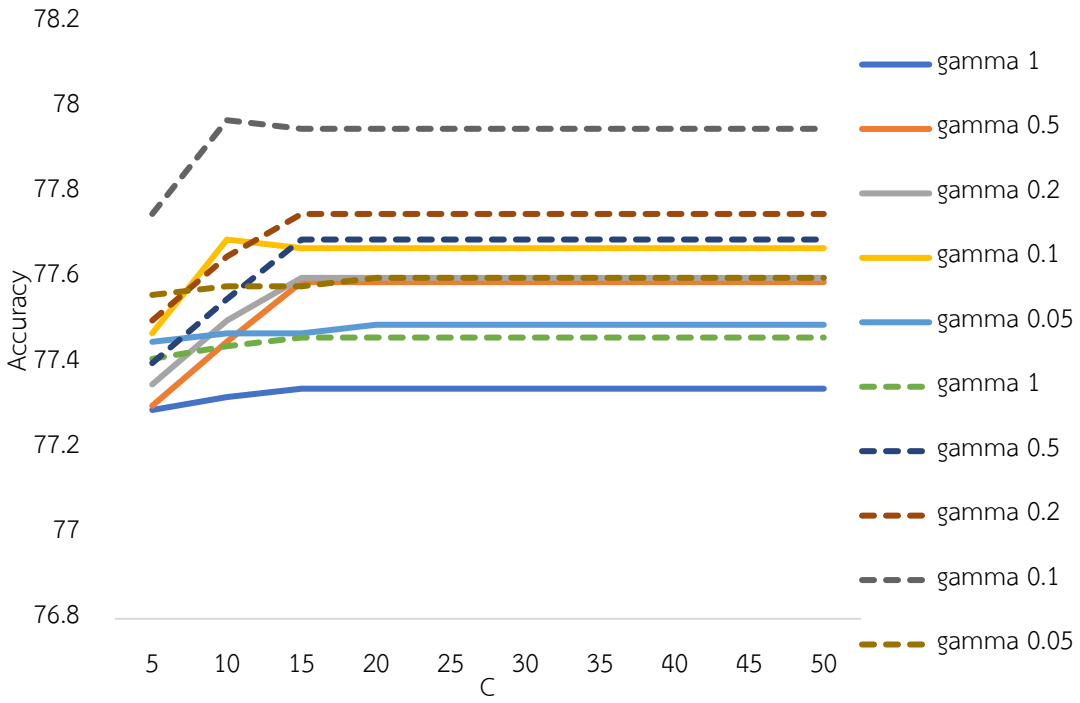
ภาพที่ 30 ค่าความถูกต้องของเคอร์เนลเชิงเส้นในแต่ละค่า C ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 30 แสดงค่าความถูกต้องของเคอร์เนลเชิงเส้น เมื่อกำหนดให้ C มีค่า 5, 10, 15, ... , 50 จะได้ว่าค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุดทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม คือ C=5



ภาพที่ 31 ค่าความถูกต้องของเคอร์เนลพหุนามในแต่ละค่า C ที่แตกต่างกัน เมื่อกำหนดค่า exponent ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 31 แสดงค่าความถูกต้องของเคอร์เนลพหุนามของค่า exponent ที่แตกต่างกันจำนวน 4 ค่า ได้แก่ 2, 3, 4 และ 5 เมื่อกำหนดให้ C มีค่า 5, 10, 15, ... , 50 จะได้ว่าค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุดทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม คือ exponent=2 และ C=5

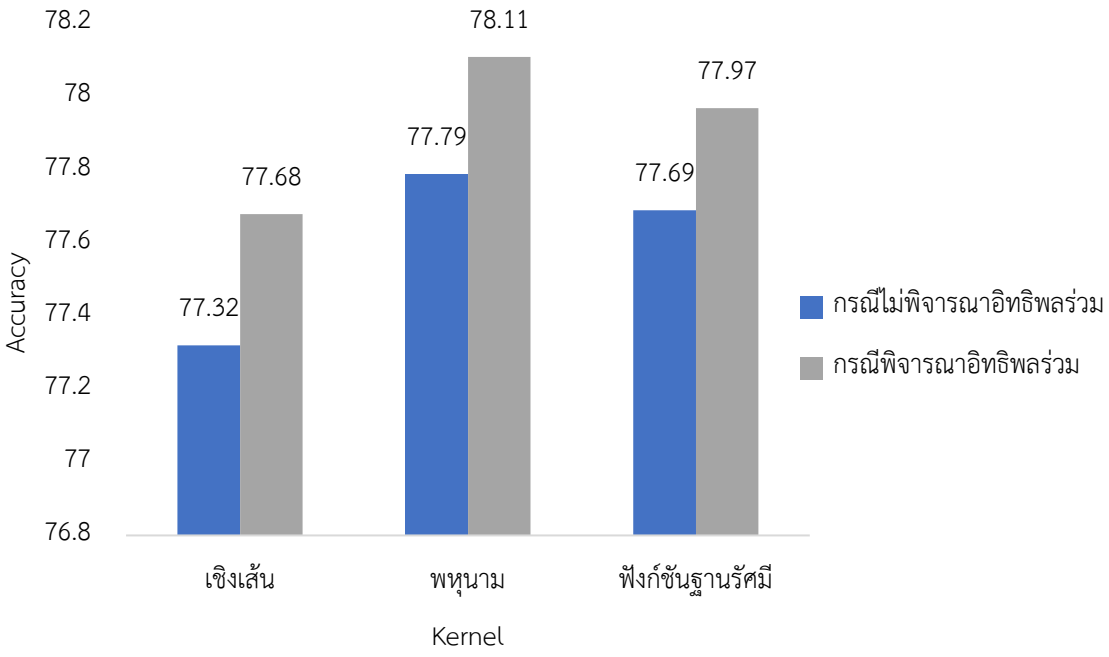


ภาพที่ 32 ค่าความถูกต้องของเคอร์เนลฟังก์ชันฐานรัศมีในแต่ละค่า C ที่แตกต่างกัน เมื่อกำหนดค่า gamma ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 32 แสดงค่าความถูกต้องของเคอร์เนลฟังก์ชันฐานรัศมีของค่า gamma ที่แตกต่างกันจำนวน 5 ค่า ได้แก่ 0.05, 0.1, 0.2, 0.5 และ 1 เมื่อกำหนดให้ C มีค่า 5, 10, 15, ... , 50 จะได้ว่าค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุดทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม คือ gamma=0.1 และ C ที่มีค่าตั้งแต่ 10 ขึ้นไป ดังนั้นเลือก C ที่มีค่าไม่สูงมาก เพราะอาจให้ตัวจำแนกมีความเรียบง่ายเกินไป จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมสำหรับทั้ง 2 กรณี คือ C=10 และ gamma=0.1

จากการค้นหาแบบกริด พบว่าสำหรับเคอร์เนลที่แตกต่างกัน จะมีค่า C ที่เหมาะสมแตกต่างกันทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม ได้ผลลัพธ์ดังนี้

- kernel = polykernel (exponent=1), C = 5
- kernel = polykernel (exponent=2), C = 5
- kernel = rbf, C = 10, gamma = 0.1



ภาพที่ 33 ค่าความถูกต้องของไฮเปอร์พารามิเตอร์เคอร์เนลประเภทต่าง ๆ กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 33 แสดงค่าความถูกต้องของไฮเปอร์พารามิเตอร์เคอร์เนลทั้ง 3 ประเภท โดยใช้ C ที่เหมาะสมที่สุดแตกต่างกัน จะเห็นได้ว่าไฮเปอร์พารามิเตอร์ที่ให้ค่าความถูกต้องสูงสุดกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม คือ เคอร์เนลฟังก์ชันพหุนาม ดังนั้นจึงกำหนดไฮเปอร์พารามิเตอร์ kernel = polykernel (exponent=2) และ C = 5

4. เทคนิค K-Nearest neighbor

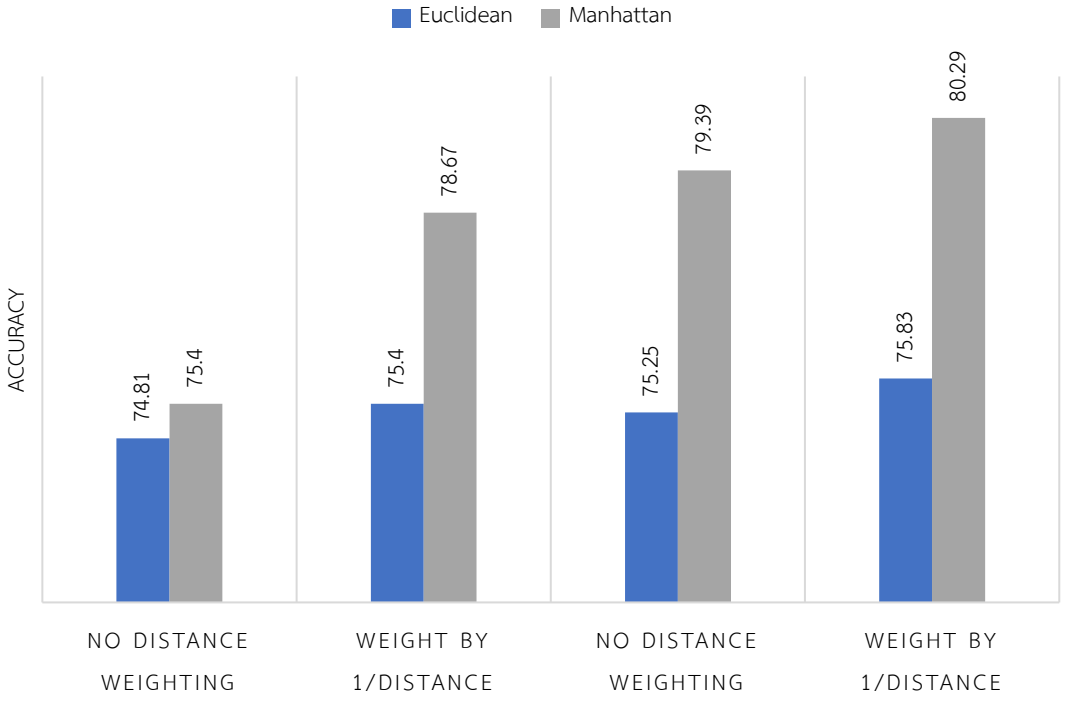
จากการค้นหาแบบกริด เมื่อพิจารณาค่าความถูกต้องจากการตรวจสอบแบบไขว้ พบว่า พบว่าสำหรับ distanceFunction ที่แตกต่างกัน จะมี DistanceWeighting และ K ที่เหมาะที่สุดแตกต่างกัน ได้ผลลัพธ์ดังนี้

กรณีไม่พิจารณาอิทธิพลร่วม

- distanceFunction = Euclidean, DistanceWeighting = Weight by 1/distance, K = 17
- distanceFunction = Manhattan, DistanceWeighting = Weight by 1/distance, K = 21

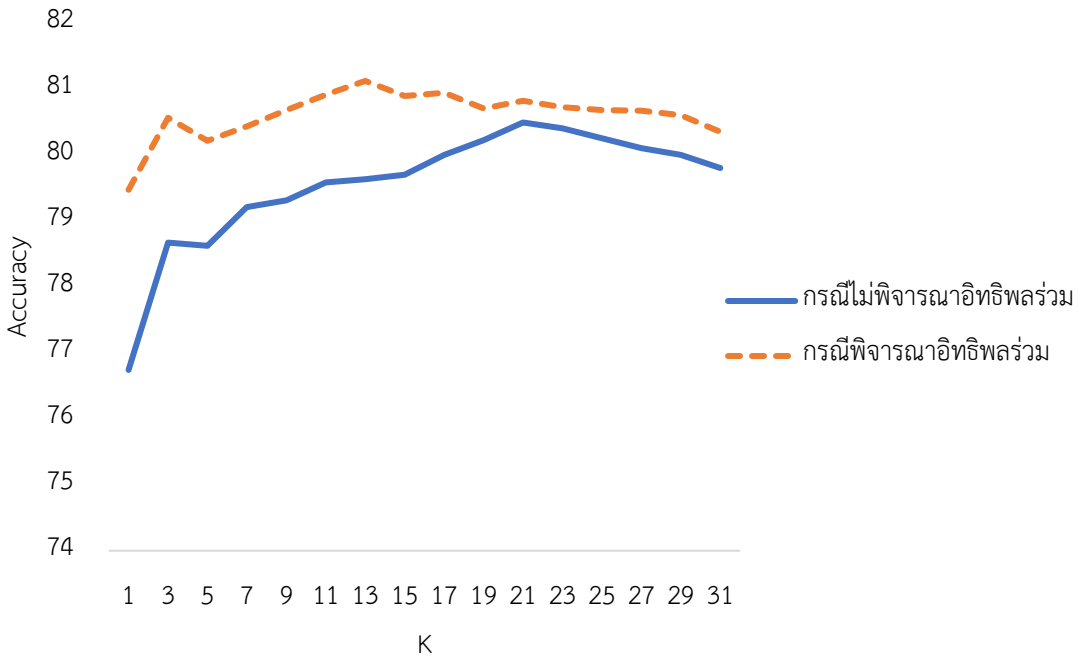
กรณีพิจารณาอิทธิพลร่วม

- distanceFunction = Euclidean, DistanceWeighting = Weight by 1/distance, K = 11
- distanceFunction = Manhattan, DistanceWeighting = Weight by 1/distance, K = 13



ภาพที่ 34 ค่าความถูกต้องของไฮเปอร์พารามิเตอร์ distanceFunction ที่แตกต่างกัน เมื่อกำหนดไฮเปอร์พารามิเตอร์ DistanceWeighting ที่แตกต่างกัน กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 34 แสดงค่าความถูกต้องของไฮเปอร์พารามิเตอร์ distanceFunction ทั้ง 2 แบบ โดยใช้ DistanceWeighting และ K ที่แตกต่างกัน เมื่อพิจารณาค่าความถูกต้องจากการตรวจสอบแบบไขว้ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม จะได้ไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุด คือ distanceFunction = Manhattan และ DistanceWeighting = Weight by 1/distance ดังนั้นจึงกำหนดไฮเปอร์พารามิเตอร์ distanceFunction = Manhattan และ DistanceWeighting = Weight by 1/distance สำหรับทั้ง 2 กรณี



ภาพที่ 35 ค่าความถูกต้องของ distanceFunction = Manhattan และ DistanceWeighting = Weight by 1/distance เมื่อกำหนด K ให้มีค่า 1, 3, ... , 31 กรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากกราฟในภาพที่ 35 แสดงค่าความถูกต้องของ distanceFunction = Manhattan และ DistanceWeighting = Weight by 1/distance เมื่อกำหนด K ให้มีค่า 1, 3, ... , 31 จะได้ว่าค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ทำให้ค่าความถูกต้องมีค่าสูงที่สุดกรณีที่พิจารณาอิทธิพลร่วม คือ K = 13 และกรณีที่ไม่พิจารณาอิทธิพลร่วม คือ K = 21

จากการประเมินประสิทธิภาพของแบบจำลองการจำแนกทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วมด้วยค่าวัดความถูกต้อง พบว่า ค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ให้ค่าความถูกต้องสูงสุดสำหรับทั้ง 4 เทคนิคการจำแนก ดังนี้

1. เทคนิคต้นไม้ตัดสินใจ (Decision tree)
    - กรณีไม่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ confidenceFactor = 0.5 และ minNumObj = 1 ซึ่งให้ค่าความถูกต้อง 83.02%
    - กรณีที่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ confidenceFactor = 0.25 และ minNumObj = 1 ซึ่งให้ค่าความถูกต้อง 84.08%

2. เทคนิคต้นไม้ป่าสุ่ม (Random forest)
    - กรณีไม่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ numIterations = 60 และ maxDepth = 30 ซึ่งให้ค่าความถูกต้อง 85.76%
    - กรณีที่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ numIterations = 60 และ maxDepth = 20 ซึ่งให้ค่าความถูกต้อง 86.72%
3. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
    - ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ kernel = polykernel (exponent=2) และ C = 5 ซึ่งกรณีที่ไม่พิจารณาอิทธิพลร่วมให้ค่าความถูกต้อง 77.79% และกรณีที่พิจารณาอิทธิพลร่วมให้ค่าความถูกต้อง 78.11%
4. เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest neighbor)
    - กรณีไม่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ distanceFunction = Manhattan, DistanceWeighting = Weight by 1/distance และ K = 21 ซึ่งให้ค่าความถูกต้อง 80.49%
    - กรณีที่พิจารณาอิทธิพลร่วม กำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม คือ distanceFunction = Manhattan, DistanceWeighting = Weight by 1/distance และ K = 13 ซึ่งให้ค่าความถูกต้อง 81.12%

ผลการวิเคราะห์ในขั้นตอนการประเมินประสิทธิภาพของแบบจำลอง

ผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลองที่สร้างขึ้นจากเทคนิคการจำแนก 4 วิธี ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม จากการใช้แบบจำลองการจำแนกการเป็นโรคเบาหวานจากค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่ได้จากชุดข้อมูลฝึกสอนซึ่งมีจำนวน 16,181 ราย มาทำการทดสอบกับชุดข้อมูลทดสอบซึ่งมีจำนวน 4,046 ราย แสดงผลการประเมินประสิทธิภาพของแบบจำลอง ดังนี้

1. เทคนิค Decision tree

ตารางที่ 35 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Decision tree กรณีที่ไม่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,397	356
	ไม่เป็นโรคเบาหวาน	401	1,892

จากตารางที่ 35 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Decision tree กรณีที่ไม่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,397 ราย คิดเป็น 77.7 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 1,892 ราย คิดเป็น 84.16 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,289 ราย คิดเป็น 81.29 %

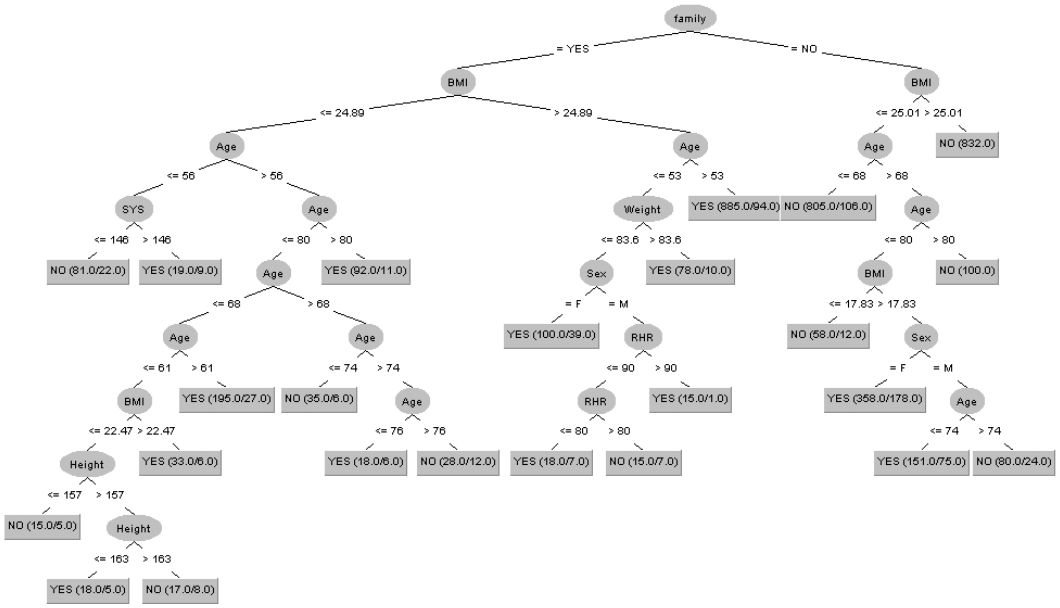
ตารางที่ 36 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Decision tree กรณีที่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,626	87
	ไม่เป็นโรคเบาหวาน	87	2,246

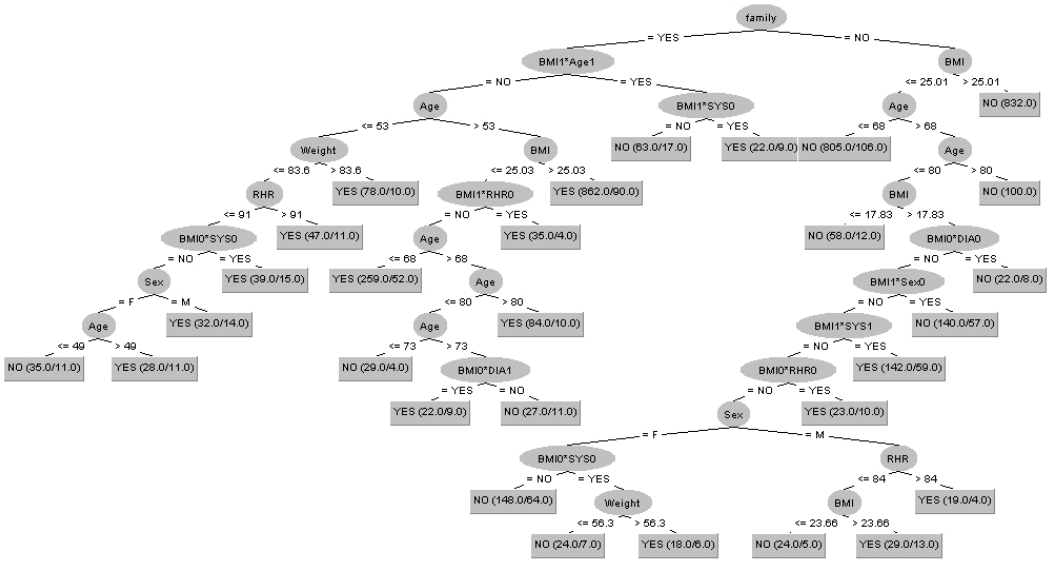
จากตารางที่ 36 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Decision tree กรณีที่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,626 ราย คิดเป็น

94.92 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 2,246 ราย คิดเป็น 96.27% ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,872 ราย คิดเป็น 95.7 %

ต้นไม้ตัดสินใจที่สร้างขึ้นโดยใช้เทคนิค Decision tree จากแบบจำลองกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม แสดงดังภาพที่ 36 และ 37 ตามลำดับ



ภาพที่ 36 ต้นไม้ตัดสินใจจากแบบจำลองกรณีที่ไม่พิจารณาอิทธิพลร่วม



ภาพที่ 37 ต้นไม้ตัดสินใจจากแบบจำลองกรณีที่พิจารณาอิทธิพลร่วม

2. เทคนิค Random forest

ตารางที่ 37 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Random forest กรณีที่ไม่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	2,328	197
	ไม่เป็นโรคเบาหวาน	279	1,242

จากตารางที่ 37 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Random forest กรณีที่ไม่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 2,328 ราย คิดเป็น 89.3 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 1,242 ราย คิดเป็น 86.31 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,570 ราย คิดเป็น 88.24 %

ตารางที่ 38 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Random forest กรณีที่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,668	45
	ไม่เป็นโรคเบาหวาน	58	2,275

จากตารางที่ 38 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Random forest กรณีที่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,668 ราย คิดเป็น 96.64 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 2,275 ราย คิดเป็น 98.06 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,943 ราย คิดเป็น 97.45 %

3. เทคนิค Support Vector Machine

ตารางที่ 39 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Support Vector Machine กรณีที่ไม่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,286	397
	ไม่เป็นโรคเบาหวาน	375	2,069

จากตารางที่ 39 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Support Vector Machine กรณีที่ไม่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,286 ราย คิดเป็น 77.42 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 2,069 ราย คิดเป็น 89.9 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,355 ราย คิดเป็น 81.29 %

ตารางที่ 40 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค Support Vector Machine กรณีที่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,561	233
	ไม่เป็นโรคเบาหวาน	183	2,069

จากตารางที่ 40 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค Support Vector Machine กรณีที่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,561 ราย คิดเป็น 89.51 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 2,069 ราย คิดเป็น 89.88 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,630 ราย คิดเป็น 89.72 %

4. เทคนิค K-Nearest neighbor

ตารางที่ 41 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค K-Nearest neighbor กรณีที่ไม่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,333	368
	ไม่เป็นโรคเบาหวาน	373	1,972

จากตารางที่ 41 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค K-Nearest neighbor กรณีที่ไม่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,333 ราย คิดเป็น 78.14 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 1,972 ราย คิดเป็น 84.27 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,305 ราย คิดเป็น 81.69 %

ตารางที่ 42 ตารางสรุปผลลัพธ์การทำนายโดยเทคนิค K-Nearest neighbor กรณีที่พิจารณาอิทธิพลร่วม

		ผลลัพธ์จริง	
		เป็นโรคเบาหวาน	ไม่เป็นโรคเบาหวาน
ผลลัพธ์การจำแนก	เป็นโรคเบาหวาน	1,552	61
	ไม่เป็นโรคเบาหวาน	83	2,350

จากตารางที่ 42 พบว่า แบบจำลองการจำแนกที่สร้างขึ้นโดยเทคนิค K-Nearest neighbor กรณีที่พิจารณาอิทธิพลร่วม สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานถูกต้อง 1,552 ราย คิดเป็น 94.92 % และจำแนกผู้รับบริการที่ไม่เป็นโรคเบาหวานถูกต้อง 2,350 ราย คิดเป็น 97.47 % ซึ่งสามารถจำแนกผลการตรวจโรคเบาหวานโดยรวมถูกต้อง 3,902 ราย คิดเป็น 96.44 %

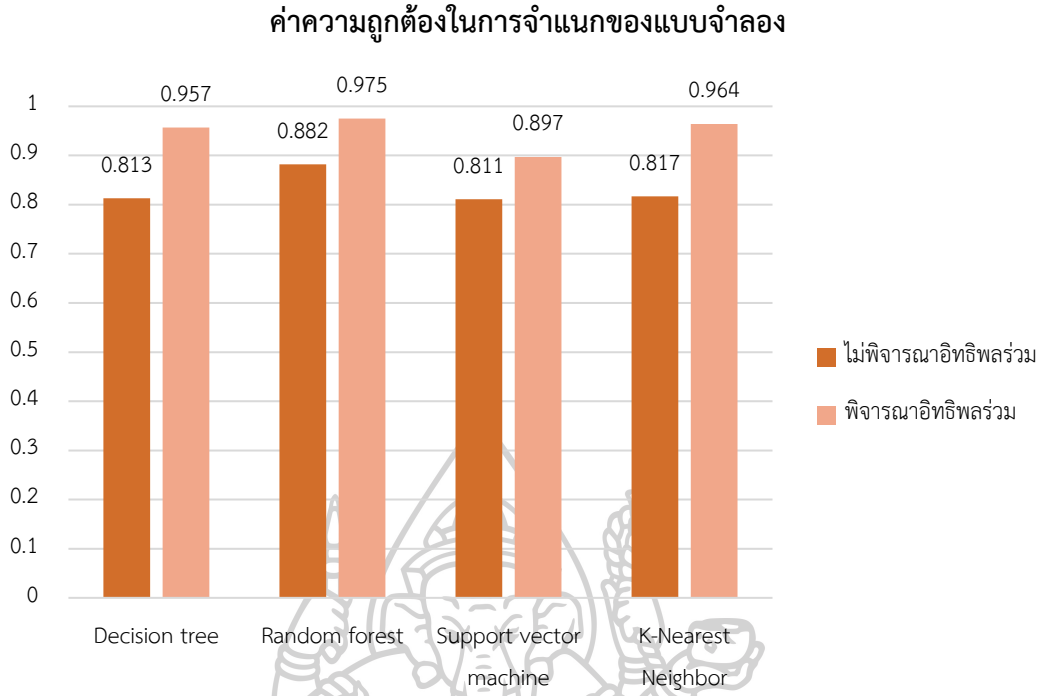
ตารางที่ 43 ตารางแสดงค่าวัดประสิทธิภาพการจำแนกของแบบจำลอง

กรณีพิจารณา	เทคนิคการจำแนก	ค่าวัดประสิทธิภาพ			
		Accuracy	Precision	Recall	F1-Score
ไม่พิจารณาอิทธิพลร่วม	Decision tree	0.813	0.797	0.777	0.783
	Random forest	**0.882**	**0.922**	**0.893**	**0.907**
	Support Vector Machine	0.811	0.764	0.774	0.769
	K-Nearest neighbor	0.817	0.784	0.781	0.787
พิจารณาอิทธิพลร่วม	Decision tree	0.957	0.949	0.945	0.949
	Random forest	**0.975**	**0.974**	**0.966**	**0.970**
	Support Vector Machine	0.897	0.870	0.895	0.882
	K-Nearest neighbor	0.964	0.962	0.949	0.956

จากตารางที่ 43 พบว่า แบบจำลองการจำแนกกรณีที่ไม่พิจารณาอิทธิพลร่วม เทคนิค Random forest ให้ค่าความถูกต้องสูงที่สุด มีค่าเท่ากับ 0.882 และเมื่อพิจารณาค่าความเที่ยง ค่าความครบถ้วน และค่าคะแนน F1 พบว่าเทคนิค Random forest มีค่าสูงที่สุด ดังนั้นแบบจำลองการจำแนกกรณีที่ไม่พิจารณาอิทธิพลร่วม เทคนิค Random forest มีประสิทธิภาพการจำแนกดีที่สุด

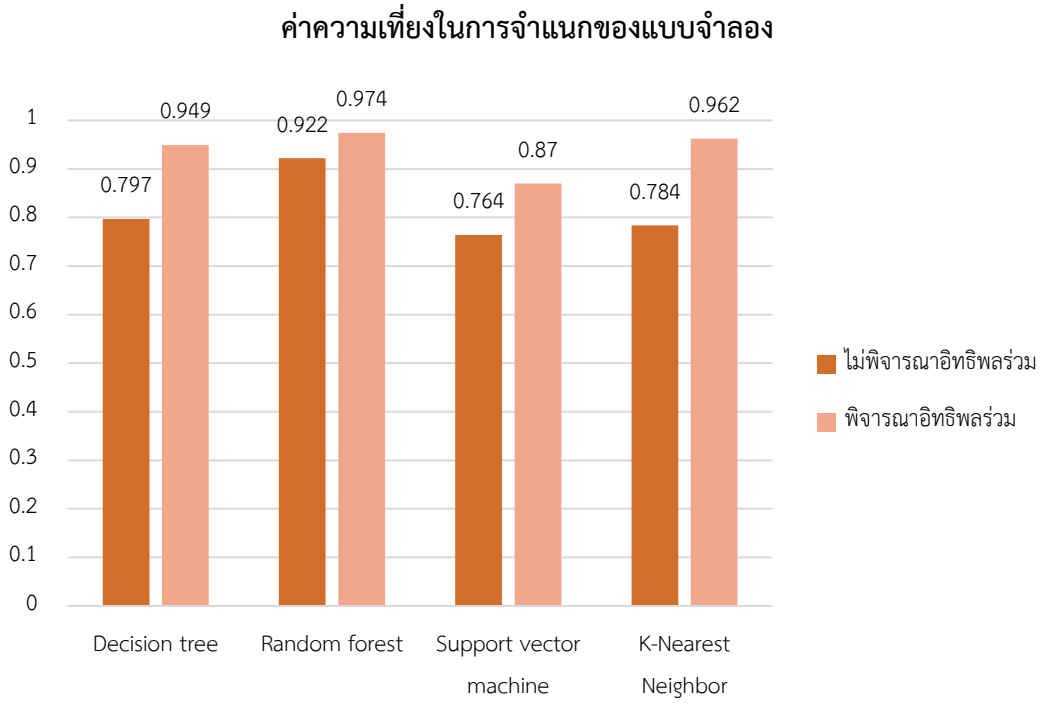
แบบจำลองการจำแนกกรณีที่พิจารณาอิทธิพลร่วม เทคนิค Random forest ให้ค่าความถูกต้องสูงที่สุด มีค่าเท่ากับ 0.975 และเมื่อพิจารณาค่าความเที่ยง ค่าความครบถ้วน และค่าคะแนน F1 พบว่าเทคนิค Random forest มีค่าสูงที่สุด ดังนั้นแบบจำลองการจำแนกกรณีที่พิจารณาอิทธิพลร่วม เทคนิค Random forest มีประสิทธิภาพการจำแนกดีที่สุด

เมื่อเปรียบเทียบประสิทธิภาพการจำแนกของแบบจำลองทั้ง 2 กรณี พบว่า แบบจำลองกรณีที่พิจารณาอิทธิพลร่วมให้ผลการจำแนกที่ดีกว่าแบบจำลองกรณีที่ไม่พิจารณาอิทธิพลร่วมสำหรับทุกเทคนิคการจำแนก



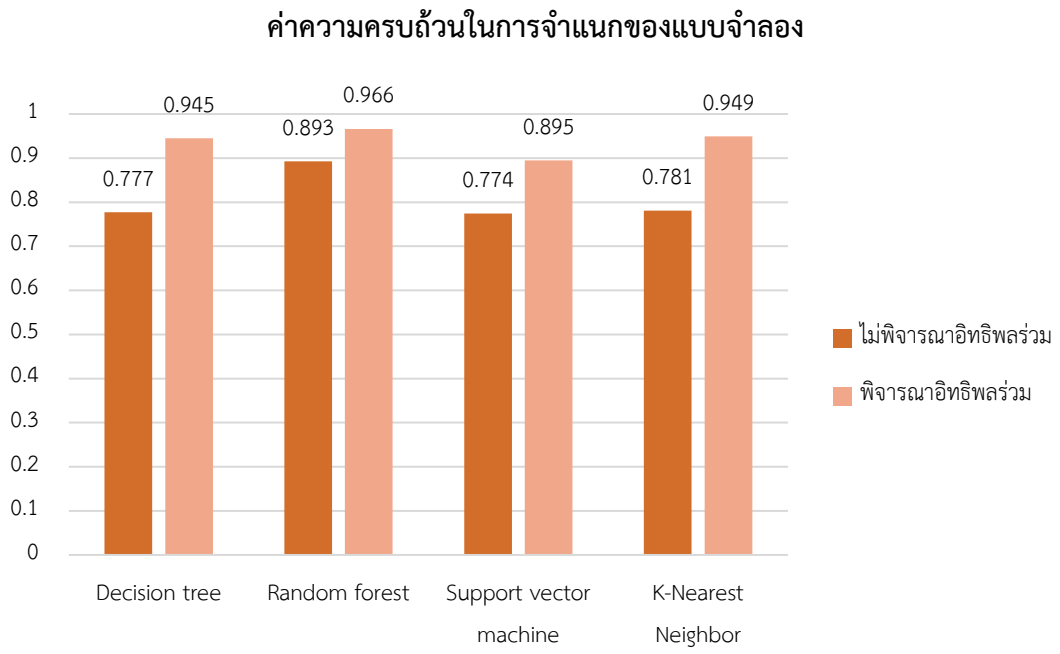
ภาพที่ 38 แผนภูมิแสดงค่าความถูกต้องในการจำแนกของแบบจำลองทั้ง 4 เทคนิค ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากแผนภูมิในภาพที่ 38 พบว่า ค่าความถูกต้องในการจำแนกของแบบจำลองกรณีที่พิจารณาอิทธิพลร่วม มีค่ามากกว่ากรณีที่ไม่พิจารณาอิทธิพลร่วม ทั้ง 4 เทคนิค คือ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor โดยที่เทคนิค Random forest ให้ค่าความถูกต้องสูงสุด ซึ่งหมายถึงเป็นวิธีที่สามารถจำแนกกลุ่มได้ถูกต้องมากที่สุด



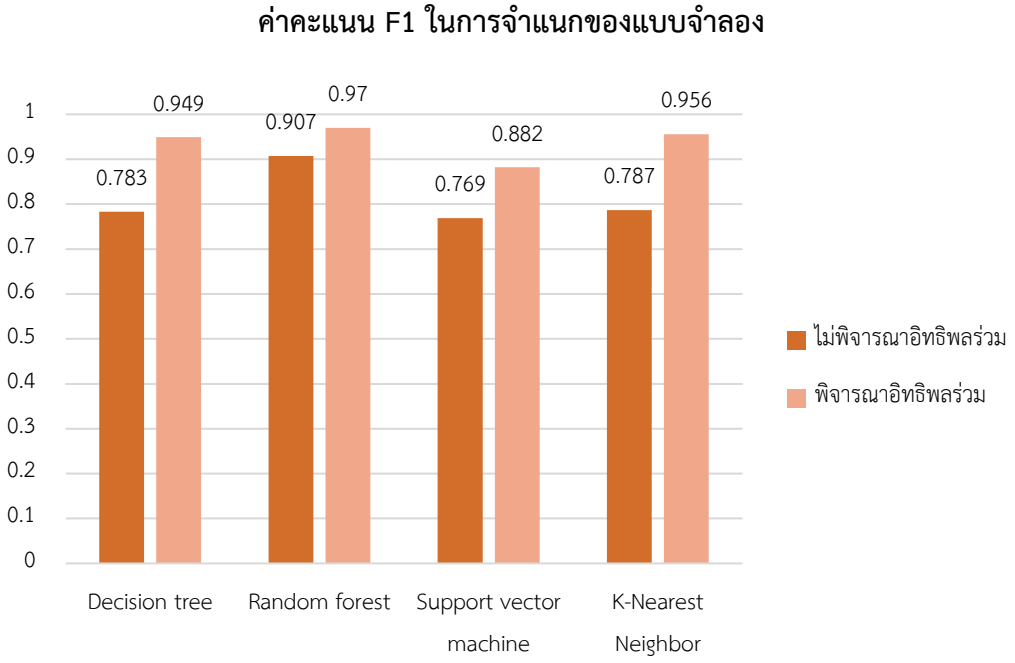
ภาพที่ 39 แผนภูมิแสดงค่าความเที่ยงในการจำแนกของแบบจำลองทั้ง 4 เทคนิค ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากแผนภูมิในภาพที่ 39 พบว่า ค่าความเที่ยงในการจำแนกของแบบจำลองกรณีที่พิจารณาอิทธิพลร่วม มีค่ามากกว่ากรณีที่ไม่พิจารณาอิทธิพลร่วม ทั้ง 4 เทคนิค คือ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor โดยที่เทคนิค Random forest ให้ค่าความเที่ยงสูงสุด ซึ่งหมายถึงเป็นวิธีที่สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานได้แม่นยำที่สุด



ภาพที่ 40 แผนภูมิแสดงค่าความครบถ้วนในการจำแนกของแบบจำลองทั้ง 4 เทคนิค ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากแผนภูมิในภาพที่ 40 พบว่า ค่าความครบถ้วนในการจำแนกของแบบจำลองกรณีที่พิจารณาอิทธิพลร่วม มีค่ามากกว่ากรณีที่ไม่พิจารณาอิทธิพลร่วม ทั้ง 4 เทคนิค คือ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor โดยที่เทคนิค Random forest ให้ค่าความครบถ้วนสูงสุด ซึ่งหมายถึงเป็นวิธีที่สามารถจำแนกผู้รับบริการที่เป็นโรคเบาหวานได้ครบถ้วนที่สุด



ภาพที่ 41 แผนภูมิแสดงค่าคะแนน F1 ในการจำแนกของแบบจำลองทั้ง 4 เทคนิค ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

จากแผนภูมิในภาพที่ 41 พบว่า ค่าคะแนน F1 ในการจำแนกของแบบจำลองกรณีที่พิจารณาอิทธิพลร่วม มีค่ามากกว่ากรณีที่ไม่พิจารณาอิทธิพลร่วม ทั้ง 4 เทคนิค คือ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor โดยที่เทคนิค Random forest ให้ค่าคะแนน F1 สูงสุด ซึ่งหมายถึงเป็นวิธีที่มีประสิทธิภาพในการจำแนกสูงที่สุด

บทที่ 5

สรุป อภิปรายผล และข้อเสนอแนะ

จากการวิจัยเรื่อง การเปรียบเทียบประสิทธิภาพแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวาน มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคที่ใช้ในการสร้างแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวานกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม โดยใช้ข้อมูลผลการตรวจเบื้องต้นของผู้รับบริการในโรงพยาบาลสังกัดสำนักการแพทย์ กรุงเทพมหานคร ตั้งแต่ปี 2562 ถึง 2564 ซึ่งตัวแปรอิสระที่ใช้ในงานวิจัยนี้มีทั้งหมด 9 ตัวแปร ได้แก่ อายุ เพศ น้ำหนัก ส่วนสูง ดัชนีมวลกาย ค่าความดันขณะหัวใจบีบตัว ค่าความดันขณะหัวใจคลายตัว อัตราการเต้นของหัวใจ และประวัติโรคเบาหวานในญาติสายตรง (พ่อ แม่ พี่ หรือน้อง) และตัวแปรตาม คือ การเป็นโรคเบาหวาน จากนั้นนำข้อมูลตามตัวแปรที่กล่าวมาข้างต้น มาหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมด้วยวิธี 5-fold cross validation สำหรับทั้ง 4 เทคนิค ได้แก่ Decision tree, Random forest, Support Vector Machine และ K-Nearest neighbor ทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม เพื่อใช้ในการสร้างแบบจำลองการจำแนกการเป็นโรคเบาหวาน จากนั้นเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกด้วยค่าวัดประสิทธิภาพต่าง ๆ คือ ความถูกต้อง ค่าความเที่ยง ค่าความครบถ้วน และค่าคะแนน F1 โดยสามารถสรุปผลและอภิปรายผลการวิจัย

สรุปผลการวิจัย

การหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมสำหรับทั้ง 4 เทคนิคการจำแนกข้อมูลทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม โดยพิจารณาจากค่าความถูกต้องในการจำแนกข้อมูลสูงสุด สรุปผลการหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมได้ดังตารางที่ 44

ตารางที่ 44 ตารางแสดงค่าความถูกต้องในการจำแนกข้อมูลจากการกำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม

กรณีพิจารณา	เทคนิคการจำแนก	ค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม	ค่าความถูกต้องในการจำแนก
ไม่พิจารณาอิทธิพลร่วม	Decision tree	confidenceFactor = 0.5 minNumObj = 1	83.02%
	Random forest	numIterations = 60 maxDepth = 30	85.76%
	Support Vector Machine	kernel = polykernel exponent = 2 C = 5	77.69%
	K-Nearest neighbor	distanceFunction = Manhattan DistanceWeighting = Weight by 1/distance K = 2	80.49%
พิจารณาอิทธิพลร่วม	Decision tree	confidenceFactor = 0.25 minNumObj = 1	84.08%
	Random forest	numIterations = 60 maxDepth = 20	86.72%
	Support Vector Machine	kernel = polykernel exponent = 2 C = 5	77.97%
	K-Nearest neighbor	distanceFunction = Manhattan DistanceWeighting = Weight by 1/distance K = 1	81.12%

การสร้างแบบจำลอง Machine learning สำหรับการจำแนกการเป็นโรคเบาหวานทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม จากการเก็บข้อมูลผู้รับบริการในโรงพยาบาลสังกัดสำนักการแพทย์ กรุงเทพมหานคร ตั้งแต่ปี 2562 ถึง 2564 เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองสำหรับการจำแนกการเป็นโรคเบาหวาน โดยการพิจารณาค่าวัดประสิทธิภาพ ดังนี้ ค่าความถูกต้อง ค่าความเที่ยง ค่าความครบถ้วน และค่าคะแนน F1 สรุปผลการวิจัยได้ดังตารางที่ 45

ตารางที่ 45 ตารางเปรียบเทียบค่าวัดประสิทธิภาพการจำแนกทั้งกรณีที่พิจารณาและไม่พิจารณาอิทธิพลร่วม

กรณีพิจารณา	เทคนิคการจำแนก	ค่าวัดประสิทธิภาพ			
		Accuracy	Precision	Recall	F1-Score
ไม่พิจารณาอิทธิพลร่วม	Decision tree	0.813	0.797	0.777	0.783
	Random forest	**0.882**	**0.922**	**0.893**	**0.907**
	Support Vector Machine	0.811	0.764	0.774	0.769
	K-Nearest neighbor	0.817	0.784	0.781	0.787
พิจารณาอิทธิพลร่วม	Decision tree	0.957	0.949	0.945	0.949
	Random forest	**0.975**	**0.974**	**0.966**	**0.970**
	Support Vector Machine	0.897	0.870	0.895	0.882
	K-Nearest neighbor	0.964	0.962	0.949	0.956

จากตารางที่ 45 พบว่ากรณีที่พิจารณาอิทธิพลร่วม ผลการทดสอบประสิทธิภาพสำหรับเทคนิคการจำแนกทั้ง 4 เทคนิคดีกว่ากรณีที่ไม่พิจารณาอิทธิพลร่วม

แบบจำลอง Machine learning กรณีที่ไม่พิจารณาอิทธิพลร่วม เทคนิคที่ให้ผลการทดสอบประสิทธิภาพที่ดีที่สุด คือการจำแนกข้อมูลด้วยเทคนิค Random forest โดยให้ค่าความถูกต้อง 88.2% ค่าความเที่ยง 92.2% ค่าความครบถ้วน 89.3% และค่าคะแนน F1 90.7% ส่วนเทคนิคที่ให้ผลการทดสอบประสิทธิภาพรองลงมา คือเทคนิค K-Nearest neighbor , Decision tree และ Support Vector Machine ตามลำดับ

แบบจำลอง Machine learning กรณีที่พิจารณาอิทธิพลร่วม เทคนิคที่ให้ผลการทดสอบประสิทธิภาพที่ดีที่สุด คือการจำแนกข้อมูลด้วยเทคนิค Random forest โดยให้ค่าความถูกต้อง 97.5% ค่าความเที่ยง 97.4% ค่าความครบถ้วน 96.6% และค่าคะแนน F1 97% ส่วนเทคนิคที่ให้ผลการทดสอบประสิทธิภาพรองลงมา คือเทคนิค K-Nearest neighbor , Decision tree และ Support Vector Machine ตามลำดับ

อภิปรายผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพการจำแนกของแบบจำลองทั้ง 4 เทคนิคสำหรับกรณีที่ไม่พิจารณาอิทธิพลร่วม ซึ่งมีการปรับปรุงข้อมูลให้มีคุณภาพมากยิ่งขึ้นโดยการตัดข้อมูลที่มีความผิดพลาดออก พบว่าเทคนิค Random forest มีประสิทธิภาพดีที่สุด ซึ่งสอดคล้องกับงานวิจัยของ (Kandhasam & Balamurali, 2015) นอกจากนี้ สอดคล้องกับงานวิจัยของ (Nandhini & Dharmarajan, 2022) เนื่องจากมีการปรับค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมกับข้อมูลจึงทำให้ความถูกต้องในการจำแนกเพิ่มขึ้น

การสร้างแบบจำลองโดยพิจารณาอิทธิพลร่วมของปัจจัยเสี่ยงที่ส่งผลต่อการเกิดโรคเบาหวาน สำหรับในงานวิจัยนี้ พิจารณาอิทธิพลร่วมของปัจจัยดัชนีมวลกาย และประวัติโรคเบาหวานในญาติสายตรงร่วมกับปัจจัยอื่น ๆ โดยผลการทดสอบพบว่าแบบจำลองกรณีพิจารณาอิทธิพลร่วมสามารถเพิ่มประสิทธิภาพการจำแนกการเป็นโรคเบาหวานได้มากยิ่งขึ้น ซึ่งสอดคล้องกับงานวิจัยของ (Pannapa, Apasiri, Sasiprapa, & Chumpol, 2021) ซึ่งการที่แบบจำลองกรณีพิจารณาอิทธิพลร่วมมีประสิทธิภาพการจำแนกที่ดีกว่าแบบจำลองกรณีที่ไม่พิจารณาอิทธิพลร่วม เนื่องจากในงานวิจัยนี้ ผู้วิจัยพิจารณาอิทธิพลร่วมจากปัจจัยสำคัญที่ส่งผลต่อการเกิดโรคเบาหวาน คือดัชนีมวลกายและประวัติโรคเบาหวานในญาติสายตรง ดังนั้นเมื่อเพิ่มตัวแปรอิทธิพลร่วมดังกล่าวที่มีความสำคัญต่อการเกิดโรคเบาหวานเข้ามาในการสร้างแบบจำลอง จึงทำให้แบบจำลองมีประสิทธิภาพการจำแนกดียิ่งขึ้น ซึ่งผลการวิจัยในครั้งนี้สามารถนำเทคนิคการจำแนกดังกล่าวเป็นแนวทางในการพัฒนาโปรแกรมสำหรับการคัดกรองผู้ป่วยโรคเบาหวานต่อไป

ข้อเสนอแนะ

1.1 ในฐานข้อมูลของระบบสารสนเทศโรงพยาบาล พบว่ามีการบันทึกข้อมูลไม่ครบถ้วน อันเนื่องมาจากสาเหตุต่าง ๆ เช่น เจ้าหน้าที่หรือพยาบาลที่ซักประวัติของผู้รับบริการบันทึกข้อมูลในระบบสารสนเทศของโรงพยาบาลไม่ครบถ้วน ความเร่งรีบในการบันทึกข้อมูลทำให้เกิดความผิดพลาดในการบันทึก ดังนั้นในการบันทึกข้อมูล ควรมีการตรวจสอบความถูกต้องและความสมบูรณ์ครบถ้วนของข้อมูลที่บันทึก เพื่อผู้ใช้ข้อมูลจะได้นำไปใช้ในการวิเคราะห์ให้ได้ผลใกล้เคียงความจริงมากที่สุด

1.2 ในการวิจัยครั้งหน้า อาจพิจารณาการสร้างอิทธิพลร่วมในรูปแบบอื่น เช่น แบบ Information gain, Information gain ratio หรือ Chi-square เป็นต้น เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกการเป็นโรคเบาหวาน

1.3 ในการวิจัยครั้งหน้า อาจพิจารณาการหาค่าไฮเปอร์พารามิเตอร์ด้วยเทคนิคอื่น ๆ เช่น Random search, Manual search หรือ Bayesian optimization เป็นต้น เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกการเป็นโรคเบาหวาน

รายการอ้างอิง

Abdulqadir, H. R., Abdulazeez, A. M., & Zebari, D. A. (2021). Data Mining Classification Techniques for Diabetes Prediction. *Qubahan Academic Journal,* 1(2), 125-133. doi:10.48161/qaj.v1n2a55

Amput, P., Srithawong, A., Sittitan, M., Wongphon, S., & Sangkarit, N. (2016). The assessment of balance ability in person with type 2 diabetes mellitus. *Bull Chiang Mai Assoc Med Sci,* 49(2), 338-343. doi:10.14456/jams.2016.35

Breiman, L. (2001). Random Forests. *Machine Learning,* 45, 5-32.

Changpetch, P., Pitpeng, A., Hiriote, S., & Yuangyai, C. (2021). Integrating Data Mining Techniques for Naïve Bayes Classification: Applications to Medical Datasets. *Computation.* 2021, 9(9), 99. doi:https://doi.org/10.3390/computation9090099

Chuchuepruksaphan, S., & Thanosawan, I. (2020). Classifying Thai News Dialoaues into Topic Types Using Machine Learning Technique.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning,* 20, 273-297. doi:10.1007/BF00994018

Dimas, A. A., & Naqshauliza, D. K. (2020). Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. *International Journal of Emerging Trends in Engineering Research,* 8, 1689-1694.

Ding, Q., Ding, Q., & Perrizo, W. (2002). Decision tree classification ofspatial data streams using Peano Count Trees. *Proceedings of the* 2002 *ACM symposium on Applied computing*, 413-417. doi:10.1145/508791.508870

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics,* 8(4), 79. doi:10.3390/informatics8040079

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques. 3.

Han, J., Saavedra, D. M., Luebbering, N., Singh, A., Sibbet, G., Ferguson, M. A. J., & Cleghon, V. (2012). Deep evolutionary conservation of an intramolecular protein kinase activation mechanism. *Research Support, N.I.H., Extramural*.

doi:10.1371/journal.pone.0029702

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Stanford, California.

Huwaidah, A., Adiwijaya, K., & Faraby, S. A. (2021). Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital. *Procedia Computer Science* 179(4), 407-415. doi:10.1016/j.procs.2021.01.023

Intarat, K., & Sillaparat, S. (2019). Tropical Mangrove Species Classification Using Random Forest Algorithm and Very High-Resolution Satellite Imagery. *BURAPHA SCIENCE JOURNAL,* 24(2), 12.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*.

Kandhasam, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science,* 47, 45-51. doi:10.1016/J.PROCS.2015.03.182

Mantovani, R. G., Cerri, R., Vanschoren, J., Horváth, T., Junior, S. B., & Carvalho, A. C. P. L. F. d. (2018). An empirical study on hyperparameter tuning of decision trees. 1, 36. doi:10.48550/arXiv.1812.02207

Mutrofin, S., Izzah, A., Kurniawardhani, A., & Masrur, M. (2014). Optimasi teknik klasifikasi modified k nearest neighbor menggunakan algoritma genetika. Retrieved from http://ejournal.umm.ac.id/index.php/gamma/article/view/2493/2698. Retrieved 1-Feb-2019 http://ejournal.umm.ac.id/index.php/gamma/article/view/2493/2698.

Nai-arun, N., & Sittidech, P. (2014). Ensemble Learning Model for Diabetes Classification. *Advanced Materials Research*(931-932), 5. doi:10.4028/www.scientific.net/AMR.931-932.1427

Nandhini, A. U., & Dharmarajan, K. (2022). Diabetes Prediction using Random Forest Classifier with Different Wrapper Methods. *International Conference on Edge Computing and Applications*, 6. doi:10.1109/ICECAA55415.2022.9936172

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC,

informedness, markedness and correlation. *International Journal of Machine Learning Technology,* 2, 27. doi:10.48550/arXiv.2010.16061

Quadri, M. N., & Kalyankar, N. V. (2021). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology* 10(2), 2-5.

Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. *Indian Journal of Science and Technology,* 9(43). doi:10.17485/ijst/2016/v9i43/93874

Setiyorini, T., & Asmono, R. T. (2020). Implementation of gain ratio and K-Nearest Neighbor for classification of student rerformance. *Jurnal Pilar Nusa Mandiri,* 16(1), 19-24. doi:10.33480/pilar.v16i1.813

Techasuwanna, R., Chottanapun, S., Chamroonsawasd, K., Sornpaisar, B., & Tunyasitthisundhorn, P. (2020). Risk factors for type 2 diabetes mellitus in Thai population. *Disease Control Journa,* 46(3), 12. doi:10.14456/dcj.2020.26

Tsenkova, V., Karlamangla, A. S., & Ryff, C. D. (2016). Parental History of Diabetes, Positive Affect, and Diabetes Risk in Adults: Findings from MIDUS. *Annals of Behavioral Medicine,* 60(6).

Viviana, A., & Andrei, D. (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. *Journal of Applied Computer Science & Mathematics,* 5(3), 4.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). An Introduction to Data Mining Techniques.

โรงพยาบาลรามคำแหง. (2563). ภาวะหัวใจเต้นผิดจังหวะ Cardiac Arrhythmia. Retrieved from https://www.ram-hosp.co.th/news_detail/144

ธนพล เริ่มปลูก. (2562). การเรียนรู้ของเครื่องจักรเพื่อการตรวจจับการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย.

นันท์พัสพร สุขสานต์, & จิราพร เกศพิชญวัฒนา. (2560). ผลของโปรแกรมส่งเสริมการกำกับตนเองต่อพฤติกรรมการบริโภคและขนาดรอบเอวของผู้ป่วยสูงอายุโรคเบาหวานที่มีภาวะอ้วน. วารสารแพทย์นาวี, 44(3), 18.

ปพนน์ศรณ์ สิ่วสำแดงเดช. (2565). การจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี.

ประยุทธ์ศิลป์ ชัยนาม. (2562). การสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมือง

ขอมูลและวิชวลไลเซชั่น.

พวงทิพย์ แท่นแสง, & ลือพล พิพานเมฆาภรณ์. (2550). การทดสอบประสิทธิภาพการท างานของอัลกอริทึมการไมนิ่งกฎสำหรับจำแนก.

พีรศุษม์ ทองพ่วง, & จรัญ แสนราช. (2021). การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายการได้รับทุนการศึกษาของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล. *Journal of Professional Routine to Research,* 8, 44-52.

ภรัณยา ปาลวิสุทธิ์. (2559). การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต. วารสารเทคโนโลยีสารสนเทศ, 12(1), 10.

มณีรัตน์ ภารนันท์. (2555). WEKA โปรแกรมทำเหมืองข้อมูล. Retrieved from https://maneerat-paranan.blogspot.com/2012/02/weka.html

รุ่งโรจน์ บุญมา, & นิเวศ จิระวิชิตชัย. (2562). การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล. วารสารวิชาการซายน์เทค มหาวิทยาลัยราชภัฏภูเก็ต, 3(2), 11-19.

วิชัย เอกพลากร, หทัยชนก พรรคเจริญ, & วราภรณ์ เสถียรนพเก้า. (2564). รายงานการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 6 พ.ศ.2562-2563.

ศุภชัย ประคองศิลป์. (2551). การออกแบบและพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการโดยใช้เทคนิคต้นไม้ตัดสินใจ กรณีศึกษา มูลนิธิที่อยู่อาศัยเพื่อมนุษยชาต.

สมาคมโรคเบาหวานแห่งประเทศไทย ในพระราชูปถัมภ์ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี. (2560). แนวทางเวชปฏิบัติสำหรับโรคเบาหวาน 2560.

สมาคมความดันโลหิตสูงแห่งประเทศไทย. (2562). แนวทางการรักษาโรคความดันโลหิตสูง ในเวชปฏิบัติทั่วไป พ.ศ. 2562.

สิตา ธานี. (2559). การพัฒนาต้นแบบการพยากรณ์ความเสี่ยงการเกิดโรคซึมเศร้าในวัยรุ่น โดยเทคนิคนาอีฟเบย์และเทคนิคต้นไม้ตัดสินใจ.

อรุณรักษ์ ตันพานิช, ดุษณี ศุภวรรธนะกุล, พิเชฐ บัญญัติ, & จรูญ จันแทน. (2562). การเปรียบเทียบโมเดลการเรียนรู้ของเครื่องสำหรับคัดกรองผู้ป่วยเบาหวานที่มีภาวะชาปลายเท้า การประชุมหาดใหญ่วิชาการระดับชาติและนานาชาติ ครั้งที่ 10, 14.

มหาวิทยาลัยศิลปากร

ประวัติผู้เขียน

ชื่อ-สกุล	เมธาพร ผ่องยิ่ง
วัน เดือน ปี เกิด	21 พฤศจิกายน 2538
สถานที่เกิด	สมุทรสาคร ประเทศไทย
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต (วท.บ.) สถิติ มหาวิทยาลัยศิลปากร
ที่อยู่ปัจจุบัน	15 ถนนสุคนธวิท 10 ตำบลตลาดกระทุ่มแบน อำเภอกระทุ่มแบน จังหวัสมุทรสาคร 74110