การจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี

วิทยานิพนธ์

ของ

ปพนน์ศรณ์ สิ่วสำแดงเดช

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์
มกราคม 2565

การจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี

วิทยานิพนธ์

ของ

ปพนน์ศรณ์ สิ่วสำแดงเดช

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

มกราคม 2565

Diabetes Miletus Classification Using Voting Ensemble Case Study: Udon Thani Hospital

Phaponsorn Siwsamdangdet

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Engineering (Electrical and Computer Engineering)

January 2022

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายปพนน์ศรณ์ สิ่วสำแดง เดช แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. อนันต์ เครือทรัพย์ถาวร)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รศ. ดร. วรวัฒน์ เสงี่ยมวิบูล)

.....กรรมการ
(ผศ. ดร. นวรัตน์ พิลาแดง)

.....กรรมการ
(ผศ. ดร. ชัยยงค์ เสริมผล)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....
(รศ. ดร. เกียรติศักดิ์ ศรีประทีป)
คณบดีคณะวิศวกรรมศาสตร์

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

ชื่อเรื่อง การจำแนกผู้ป่วยเบาหวานโดยใช้เทคนิคการโหวตรวม กรณีศึกษา: โรงพยาบาลศูนย์อุดรธานี
ผู้วิจัย ปพนน์ศรณ์ สิ่วสำแดงเดช
อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. วรวัฒน์ เสงี่ยมวิบูล
ปริญญา วิศวกรรมศาสตรมหาบัณฑิต **สาขาวิชา** วิศวกรรมไฟฟ้าและคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2565

บทคัดย่อ

งานวิจัยนี้ได้ประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผู้ป่วยโรคเบาหวานด้วยเทคนิคต้นไม้ในการตัดสินใจ (Decision Tree) เทคนิคนาอีฟ เบย์ (Naïve Bayes) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) เทคนิคโหวตร่วม (Vote Ensemle) เทคนิคป่าสุ่ม (Random Forest) เพื่อสร้างแบบจำลองในการพยากรณ์ผู้ป่วย และนำค่าวัดประสิทธิภาพของการจำแนกประเภทข้อมูลมาเปรียบเทียบ โดยค่าความถูกต้อง (Accuracy) ที่ให้ค่ามากที่สุด ผลของการวิจัยพบว่าเทคนิคป่าสุ่ม (Random Forest) ให้ค่าความถูกต้องในการทำนายผลเป็นโรคเบาหวนมากที่สุดที่อยู่ที่ 88.03% มีค่าความแม่นยำ (Precision) ที่ 88.22% ค่าความครบถ้วน (Recall) ได้ 90.36% และค่าวัดประสิทธิภาพโดยรวม (F-Measure) 89.28% สามารถนำผลลัพธ์ที่ได้ จังหวัดอุดรธานี ไปใช้ในการประกอบการรักษาผู้ป่วยโรคเบาหวานต่อไปในอนาคต

คำสำคัญ : การทำเหมืองข้อมูล, ผู้ป่วยโรคเบาหวาน

TITLE Diabetes Miletus Classification Using Voting Ensemble Case Study: Udon Thani Hospital

AUTHOR Phaponsorn Siwsamdangdet

ADVISORS Associate Professor Worawat Sa-Ngiamvibool , Ph.D.

DEGREE Master of Engineering **MAJOR** Electrical and Computer Engineering

UNIVERSITY Mahasarakham University **YEAR** 2022

ABSTRACT

This research aims to apply data mining techniques to forecast patients who were diagnosed as having Diabetes by using Decision Tree, Naïve Bayes, K-Nearest Neighbor, Vote Ensemble, and Random Forest for developing patient forecasting models and comparing classification performances by their accuracy. Results show that Random Forest indicates Diabetes with the highest accuracy rate of 88.03% and indicates the precision rate of 88.22%, additionally with the 90.36% Recall rate and F-Measure rate of 89.28% The outcomes can be applied in the treatment of diabetic patients in the future

Keyword : Data Mining, Diabetes

กิตติกรรมประกาศ

ในการศึกษาการพยากรณ์ผู้ป่วยโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษาโรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี มีจุดประสงค์เพื่อศึกษาสาเหตุการเกิดโรคในการสร้างแบบจำลองการพยากรณ์ผู้ป่วย เพื่อที่จะได้นำแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวานนำไปประกอบการรักษาให้มีประสิทธิภาพให้มากขึ้น

ต้องขอขอบพระคุณคณะอาจารย์สาขาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ที่ได้มอบความรู้และประสบการณ์ในการทำวิจัยเกี่ยวกับการวิเคราะห์ระบบ การวิเคราะห์ข้อมูล มาประยุกต์ใช้กับการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจโดยใช้เทคนิคการทำเหมืองข้อมูล

ขอขอบพระคุณที่เป็นรศ. ดร. วรวัฒน์ เสงี่ยมวิบูล ที่ปรึกษา การทำวิจัยการพยากรณ์ผู้ป่วยโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล และผู้ชี้ทางนำพาสู่ความสำเร็จ เพื่อที่จะได้นำความรู้ไปศึกษาต่อยอดต่อไป

ขอขอบพระคุณโรงพยาบาลศูนย์อุดรธานีที่ได้ให้โอกาสในการฝึกงานและอนุเคราะห์ข้อมูลผู้เข้ารับประกันสังคมและผู้ป่วยเบาหวานตั้งแต่ปีพ.ศ. 2558-2564

ขอขอบพระคุณ พ่อ แม่ ญาติพี่น้อง โดยเฉพาะอย่างยิ่งอาจารย์ ดร. ภัชษร สิ่วสำแดงเดชที่ได้ฟูมฟัก เลี้ยงดู และมอบการศึกษาที่ดี และสุดท้าย ขอขอบคุณเพื่อน ๆ ที่เป็นกำลังใจและฟันฝ่าอุปสรรคมาด้วยกัน

ปพนน์ศรณ์ สิ่วสำแดงเดช

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมา

ในปัจจุบันวิถีชีวิตของมนุษย์เราได้มีการเปลี่ยนแปลงไปมาก อันเนื่องจากมีสิ่งอำนวยความสะดวกในการใช้ชีวิตทำให้กิจกรรมต่างๆในแต่ละวันไม่ว่าจะเป็นการทำงาน การเดินทาง และเครื่องมือเครื่องใช่ต่างๆสามารถตอบสนองความต้องการได้ทันที อีกทั้งเทคโนโลยีการผลิตอาหารในปัจจุบันทำให้สามารถผลิตอาหารได้ตามจำนวนที่ต้องการ โดยเฉพาะอย่างยิ่งอาหารจำพวกคาร์โบไฮเดรต โปรตีน หรือไขมันที่สามารถให้พลังงานแก่ผู้บริโภคได้หลายแคลอรี่ ด้วยปัจจัยเหล่านี้ทำให้เกิดการละเลยความใส่ใจในสุขภาพของตนเองจนนำมาสู่การเกิดโรคไม่ติดต่อต่างๆ ขึ้นมา หนึ่งในโรคดังกล่าวคือโรคเบาหวานซึ่งเป็นโรคที่อยู่ในอันดับโรคยอดนิยมในประเทศไทย

โรคเบาหวานเป็นโรคที่เกิดจากภาวะน้ำตาลในเลือดสูงในระยะยาวซึ่งแบ่งเป็น 2 ระยะ ซึ่งผู้ป่วยจะมีอาการปัสสาวะบ่อย มีกรดในกระเพาะและมีอาการหิว ซึ่งสามารถมีโรคแทรกซ้อนอื่น ๆ เกิดขึ้นได้ เช่น โรคหลอดเลือดหัวใจ โรคความดันโลหิตสูง รวมถึงอาการแทรกซ้อนต่างๆ เช่น ภาวะคีโตซีส ภาวะช็อคจากน้ำตาลในเลือดสูง อ้างอิงจากรายงานโรคเบาหวานทั่วโลกขององค์การอนามัยโลก ในปีพ.ศ. 2557 มีจำนวนประชาการวัยผู้ใหญ่ประมาณ 422 ล้านคนเป็นโรคเบาหวานอยู่ เมื่อเทียบกับปี พ.ศ. 2523 ที่มีจำนวน 108 ล้านคนพบว่าเพิ่มขึ้นจาก 4.7% เป็น 8.5% ในประชากรผู้ใหญ่ ในประเทศไทย ในปีพ.ศ. 2562 [1] การป้องกันและการรักษาเกี่ยวข้องกับอาหารเพื่อสุขภาพ การออกกำลังกายการไม่ใช้ยาสูบและการมีน้ำหนักตัวปกติ การควบคุมความดันโลหิตและการดูแลเท้าที่เหมาะสมเป็นสิ่งสำคัญสำหรับผู้ที่เป็นโรค โรคเบาหวานประเภท 1 ต้องได้รับการจัดการด้วยการฉีดอินซูลิน โรคเบาหวานประเภท 2 อาจได้รับการรักษาด้วยยาที่มีหรือไม่มีอินซูลิน อินซูลินและยารักษาโรคบางชนิดอาจทำให้น้ำตาลในเลือดต่ำ การผ่าตัดลดน้ำหนักในผู้ที่มีโรคอ้วนเป็นวิธีการวัดที่มีประสิทธิภาพสำหรับผู้ป่วยโรคเบาหวานชนิดที่ 2 เบาหวานขณะตั้งครรภ์มักจะหายไปหลังจากการคลอดของทารก

การพยากรณ์คือการทำนายเหตุการณ์ในอนาคตซึ่งอาจจะนำข้อมูลในอดีตมาพยากรณ์โดยใช้หลักการณ์ทางคณิตศาสตร์ ใช้หลักดุลยพินิจของผู้พยากรณ์หรืออาจใช้หลายวิธีมารวมกันเพื่อให้การพยากรณ์นั้นแม่นยำที่สุด การวิเคราะห์จำนวนผู้ป่วยโรคเบาหวานนั้น จะได้จากการประมาณ หรือการวิเคราะห์เชิงสถิติของเจ้าหน้าที่กระทรวงสาธารณสุข ซึ่งอาจทำให้เกิดความคลาดเคลื่อนของจำนวนผู้ป่วยโรคหัวใจและส่งผลกระทบต่อการวางแผนในการรักษาโรคหัวใจเกิดความคลาดเคลื่อนไปด้วย

ดังนั้นการนำเทคโนโลยีสารสนเทศมาช่วยการพยากรณ์จำนวนผู้ป่วยโรคหัวใจเป็นแนวทางหนึ่งที่ช่วยให้การการวิเคราะห์เกิดความแม่นยำมากขึ้น ซึ่งมีความรวดเร็วในการประมวลผลจะทำให้ได้ข้อมูลที่ใกล้เคียงความเป็นจริงและน่าเชื่อถือมากขึ้น

ด้วยเหตุดังกล่าวผู้วิจัยจึงมีแนวคิดในการพัฒนาแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวานด้วยเทคนิคการทำเหมืองข้อมูล (data mining) ซึ่งเป็นกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบ แนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นมาสร้างโมเดลสำหรับพยากรณ์โอกาสการเป็นโรคของผู้ป่วยโรคเบาหวาน โดยใช้ข้อมูลจากระบบฐานข้อมูลจาก โรงพยาบาลศูนย์อุดรธานีมาใช้ในการวิเคราะห์ในครั้งนี้ ผลลัพธ์ที่ได้จากการวิเคราะห์ สนับสนุนการตัดสินใจในการรักษาผู้ป่วยโรคเบาหวานได้ในอนาคต

1.2 วัตถุประสงค์การวิจัย

1. เพื่อพัฒนาแบบจำลองสำหรับการพยากรณ์ผู้ป่วยโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล
2. เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองที่เหมาะสมที่สุดสำหรับการพยากรณ์ผู้ป่วยโรคเบาหวาน

1.3 ขอบเขตของการวิจัย

1. ข้อมูลผู้ป่วยโรคเบาหวานจากระบบฐานข้อมูลจากโรงพยาบาลศูนย์อุดรธานี ซึ่งเก็บข้อมูลตั้งแต่ปี 2558-2562 มีจำนวน 70,421 แถว
2. ข้อมูลสุขภาพของผู้เข้ารับสิทธิตามประกันสังคม ซึ่งเก็บข้อมูล ตั้งแต่ปี 2558-2564 มีจำนวน 80,201 แถว
3. พื้นที่ที่ใช้ในการวิจัยคือโรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี
4. ระยะเวลาในการเก็บข้อมูลตั้งแต่ 22 กรกฎาคม พ.ศ. 2563 ถึง 22 พฤศจิกายน พ.ศ. 2563

1.4 ความสำคัญของการวิจัย

โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานีนั้นได้มีการรับผู้ป่วยโรคเบาหวานจำนวนมาก ด้วยเหตุนี้เองจึงนำแบบจำลองที่ใช้ในการพยากรณ์ผู้ป่วยโรคเบาหวานในโรงพยาบาลศูนย์อุดรธานีไปใช้เปรียบเทียบประสิทธิภาพแบบจำลองการพยารณ์ผู้ป่วยและประกอบการรตัดสินใจในการรักษาเพื่อให้สอดคล้องต่อจำนวนผู้ป่วยในอนาคตที่มีการเป็นโรคนี้มากขึ้น รวมถึงลดระดับความเสี่ยงต่อการเสียชีวิต ป้องกันเฝ้าระวังแก้ไขเพื่อลดอัตราการเกิดโรคต่อไป

1.5 นิยามศัพท์เฉพาะ

1. เหมืองข้อมูล คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูป แบบและความสัมพันธ์ ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

2. โรคเบาหวาน โรคที่ระดับน้ำตาลในเลือดสูงมากกว่าปกติ (hyperglycemia) ติดต่อกันเป็นระยะเวลานานและต่อเนื่อง มีสาเหตุจากตับอ่อน (pancreases) ไม่สามารถสร้างฮอร์โมนอินซูลิน (insulin) ได้อย่างเพียงพอหรือเกิดจากเซลล์ต่าง ๆ ในร่างกายไม่ตอบสนองต่อฮอร์โมนอินซูลิน (insulin tolerance) อินซูลินเป็นฮอร์โมนที่เกี่ยวข้องกับการควบคุมสมดุลของน้ำตาลในเลือด (โดยเฉพาะน้ำตาลกลูโคส) อินซูลินทำให้มีการนำน้ำตาลเข้าสู่เซลล์ต่าง ๆ เพื่อนำไปใช้เป็นแหล่งพลังงาน กระบวนการเมแทบอลิซึม และทำให้ระดับน้ำตาลในเลือดลดลงเข้าสู่ภาวะปกติ

3. การพยากรณ์ คือการคาดคะเนหรือการทานายลักษณะการเกิดของเหตุการณ์หรือสภาพการณ์ในอนาคต โดยศึกษารูปแบบเหตุการณ์หรือสภาพการณ์จากข้อมูลที่รวบรวมอย่างมีระบบและ/หรือ จากความสามารถ ประสบการณ์ และวิจารญาณของผู้พยากรณ์

บทที่ 2
ปริทัศน์เอกสารข้อมูล

การวิจัยเรื่องการจำแนกผู้ป่วยโรคเบาหวานด้วยกระบวนการโหวต กรณีศึกษา โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องเพื่อเป็นแนวทางในการศึกษา ดังรายละเอียดต่อไปนี้

2.1 โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี

2.2 โรคเบาหวาน

2.3 การทำเหมืองข้อมูล

2.4 เทคนิคการทำเหมืองข้อมูลในงานวิจัย

2.5 เทคโนโลยีที่ใช้ในการพัฒนาแบบจำลอง

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี

2.1.1 ประวัติโรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี

โรงพยาบาลอุดรธานี ตั้งอยู่ในเขตเทศบาลนครอุดรธานี อยู่ห่างจากศาลากลางจังหวัดประมาณ 1.2 กิโลเมตร เดิมที่ตั้งของโรงพยาบาลอุดรธานี เป็นที่ของกระทรวงศึกษาธิการ ใช้เป็นบ้านพักของธรรมการมณฑลอุดร ต่อมาทางราชการยุบมณฑลอุดรและบ้านพักธรรมการได้ชำรุดผุพัง ที่ดินจึงวางเปล่าไม่ได้ใช้ประโยชน์ทางราชการจึงได้จัดตั้งเป็นสถานพยาบาลขึ้นในปี พ.ศ.2494 โดย พ.ต.อ.ขุนศุภกิจเลขการ ซึ่งขณะนั้นดำรงตำแหน่งข้าหลวงประจำจังหวัดอุดรธานี ได้เห็นความจำเป็นในด้านสุขอนามัย และการรักษาพยาบาลเกี่ยวกับการเจ็บป่วยของประชาชนจังหวัดอุดรธานีเป็นสำคัญ จึงได้เสนอของบประมาณสำหรับจัดสร้างโรงพยาบาล แต่งบประมาณที่ได้รับจัดสรรมาใหม่จำนวนน้อย ทางจังหวัดจึงสนับสนุนจัดหาเงินสมทบจากงานประจำปีทุ่งศรีเมือง ตลอดทั้งชักชวนร่วมสมทบบริจาคอีกด้วย และในเวลาต่อมาวันที่ 29 มิถุนายน 2496 ได้มีการประกอบพิธีวางศิลาฤกษ์ก่อสร้างตึกอานวยการหลังแรกของโรงพยาบาลอุดรธานีขึ้น และวันที่ 24 เมษายน 2497 ขุนบริบาลบรรพตเขตต์ผู้ว่าราชการจังหวัดในขณะนั้น ได้ทำพิธีเปิดตึกอำนวยการเพื่อเปิดรับบริการผู้ป่วยอย่างเป็นทางการเป็นครั้งแรก ตลอดระยะเวลาที่ผ่านมาโรงพยาบาลอุดรธานีได้ดำเนินการปรับปรุงพัฒนาคุณภาพและขยายการให้บริการอย่างสม่ำเสมอมาตามลำดับจนถึงปัจจุบัน ภายใต้การบริหารงานโดย

มีวิสัยทัศน์และพันธกิจด้วยความมุ่งมั่นจนได้ผลงานบริการด้านสุขภาพที่มีการพัฒนาอย่างต่อเนื่อง จากผลการปฏิบัติงานด้วยความเสียสละและมีความตั้งใจทำงานในหน้าที่ราชการเป็นอย่างดีของ ผู้อำนวยการ แพทย์ ทันต-แพทย์ เภสัชกร พยาบาล และเจ้าหน้าที่อื่น ๆ ทั้งในอดีตและปัจจุบันจน เป็นที่ยอมรับของประชาชนในจังหวัดอุดรธานี และจังหวัดใกล้เคียง ซึ่งได้สนับสนุนทางด้านกำหลัง ทรัพย์วัสดุอุปกรณ์ ทางการแพทย์ ตลอดจนผู้บริหารชั้นผู้ใหญ่ในกระทรวงสาธารณสุขที่ให้การ สนับสนุนด้านเงินงบประมาณ ทำให้โรงพยาบาลอุดรธานี ได้มีการพัฒนาจนเจริญก้าวหน้าดั้งที่เป็นอยู่ ในปัจจุบันและจะเจริญก้าวหน้า ทั้งในด้านการบริหาร และการบริการยิ่งขึ้นต่อไปในอนาคต [2]

2.1.2 พันธกิจ (Mission)

1) ให้บริการสาธารณสุขแบบองค์รวมอย่างมีคุณภาพ ครอบคลุมในด้านการดูแล รักษา/ส่งเสริมป้องกันและฟื้นฟูสภาพ

2) เป็นโรงพยาบาลแม่ข่ายในการรับส่งต่อการดูแลรักษาระดับตติยภูมิ และตติยภูมิ ระดับสูง ใน 4 ด้าน ได้แก่ Trauma, Cardio, Cardiac Arrest (CA), Newborn

3) เป็นสถาบันรวมผลิต และฝึกอบรมแพทย์และบุคลากรสาธารณสุข

4) ส่งเสริมการพัฒนาบุคลากร การศึกษาวิจัยและพัฒนาระบบบริการสุขภาพอย่าง ต่อเนื่อง

5) ส่งเสริมและบูรณาการการดูแลสุขภาพโดยสร้างการมีส่วนร่วมของภาคเครือข่าย สุขภาพ ภารกิจ ของหน่วยงานในสังกัดกระทรวงสาธารณสุข

2.1.3 วิสัยทัศน์ (Vision)

เป็นโรงพยาบาลเชี่ยวชาญระดับสูงชั้นนำของอนุภูมิภาคลุ่มน้ำโขง ที่ให้การดูแลระบบบริการ สุขภาพแบบองค์รวม และเป็นสถาบันฝึกอบรมด้านการแพทย์และสาธารณสุขที่มีคุณภาพ

2.2 โรคเบาหวาน

โรคเบาหวาน (Diabetes Mellitus: DM) เป็นโรคที่ทีระดับน้ำตาลในเลือดสูงในระยะยาวซึ่ง ทำให้มีอาการปัสสาวะบ่อยและอาการหิวที่เพิ่มขึ้น หากไม่ได้รับการรักษาอาจจะทำให้เกิดโรคแทรก ซ้อนตามมาภายหลังรวมถึงภาวะแทรกซ้อนเฉียบพลันคือภาวะคิโตซิสและภาวะน้ำตาลในเลือดสูงและ โรคแทรกซ้อนในระยะยาวได้แก่โรคหัวใจ โรคหลอดลือดตีบตัน ภาวะตับล้มเหลว เป็นแผลที่เท้าและ การบาดเจ็บทางตา สาเหตุของการเกิดโรคเบาหวานคือไตไม่สามารถผลิตอิลซูลินได้มากพอหรือมี เซลล์ที่ไม่สามารถตอบสนองอิลซูลินแบบปกติ [3]

2.2.1 อาการและสัญญาณเกิดโรค

อาการโดยพื้นฐานของโรคเบาหวานจะมีน้ำหนักลดลง ปัสสาวะบ่อย มีความหิวและกระหายที่มากขึ้น โดยอาการเหล่านี้จะเกิดขึ้นอย่างต่อเนื่อง (ในระยะเวลาเดือนหรือสัปดาห์) ในเบาหวานประเภทที่ 1 ในขณะที่อาการเหล่านี้เกิดขึ้นช้าหรือไม่เกิดเลยในโรคเบาหวานประเภทที่ 2 หลายๆ อาการและสัญญาณเกิดโรคสามารถระบุการเป็นโรคเบาหวานได้ จากข้างต้นทั้งหมดไม่ได้ระบุในโรคเบาหวานซึ่งรวมถึงอาการ สายตาพล่ามัว ปวดหัว เมื่อยล้า บาดแผลหายช้า มีอาการคันตามผิวหนัง หากมีอาการน้ำตาลในเลือดสูงเป็นเวลานานจะส่งผลต่อการดูดซับน้ำตาลในนัยน์ตาได้ ทำให้มีการเปลี่ยนรูปร่างส่งผลต่อการมองเห็น จำนวนการเกิดอาการคันตามผิวหนังเกิดขึ้นในโรคเบาหวานโดยรวมรู้จักกันคือโรคผิวหนังในผู้เป็นโรคเบาหวาน

2.2.2 สาเหตุการเกิดโรคเบาหวาน

โรคเบาหวานสามารถจำแนกได้เป็น 4 แบบ คือเบาหวานชนิดที่ 1 , เบาหวานชนิดที่ 2 , เบาหวานขณะตั้งครรภ์และอื่น ๆ โรคเบาหวานอื่น ๆ นั้นรวมถึงสาเหตุทั้งหมด โรคเบาหวานที่ไม่คุณสมบัติไม่ครบก็หมายถึงโรคเบาหวานเช่นกัน

1. เบาหวานชนิดที่ 1

โรคเบาหวานชนิดที่ 1 จะมีอาการสูญเสียการผลิตอิลซูลินเบต้าเซลของบริเวณของตับอ่อนที่มีเซลล์ไร้ท่อที่มีความผิดปกติ เบาหวานแบบนี้จะตรวจสอบได้ทันทีหรือไม่มีการตอบสนอง โดยธรรมชาติของโรคเบาหวานประเภทที่ 1 ภูมิคุ้มกันอัตโนมัติ T – cell จะโจมตีส่วนที่เบต้าเซลล์เสียหายไปจนถึงอิลซูลิน ประมาณ 10% ของผู้ป่วยโรคเบาหวานในอเมริกาเหนือและยุโรปได้รับผลกระทบส่วนใหญ่มีสุขภาพแข็งแรงและมีน้ำหนักที่ดีเมื่อเริ่มมีอาการความไวและการตอบสนองต่ออินซูลินเป็นปกติโดยเฉพาะอย่างยิ่งในระยะแรก โรคเบาหวานประเภท 1 สามารถส่งผลกระทบต่อเด็กหรือผู้ใหญ่ แต่โดยทั่วไปเรียกว่า "โรคเบาหวานสำหรับเด็ก" เพราะส่วนใหญ่ของผู้ป่วยโรคเบาหวานเหล่านี้อยู่ในเด็ก

เบาหวานที่ควบคุมได้ยาก (Brittle diabetes) หรือที่รู้จักกันในชื่อ labile diabetes คือการที่ระดับน้ำตาลกลูโคสไม่คงที่ซึ่งเกิดขึ้นบ่อยโดยไม่ทราบสาเหตุกับโรคเบาหวาน ชนิดพึ่งอินซูลิน อย่างไรก็ตามคำนี้ไม่มีพื้นฐานทางชีววิทยาและไม่ควรใช้ ถึงกระนั้นเบาหวานชนิดที่ 1 ก็สามารถมาพร้อมกับภาวะน้ำตาลในเลือดสูงผิดปกติและคาดเดาไม่ได้มักจะมีคีโตซีสและบางครั้งก็มีภาวะน้ำตาลในเลือดสูง ภาวะแทรกซ้อนอื่น ๆ ได้แก่ การตอบสนองต่อภาวะน้ำตาลในเลือดต่ำการติดเชื้อ gastroparesis (ซึ่งนำไปสู่การดูดซึมคาร์โบไฮเดรตคาร์โบไฮเดรตที่ไม่แน่นอน) และ

endocrinopathies (เช่นโรคแอดดิสัน) ปรากฏการณ์เหล่านี้เชื่อว่าจะเกิดขึ้นไม่บ่อยกว่าใน 1% ถึง 2% ของผู้ที่เป็นโรคเบาหวานประเภท 1

โรคเบาหวานประเภท 1 นั้นได้รับการถ่ายทอดบางส่วนพร้อมยีนหลายชนิดรวมถึงยีน HLA บางชนิดที่ทราบกันดีว่ามีอิทธิพลต่อความเสี่ยงของโรคเบาหวาน ในคนที่มีความอ่อนไหวทางพันธุกรรมการเริ่มต้นของโรคเบาหวานสามารถเกิดขึ้นได้จากปัจจัยทางสภาพแวดล้อมหนึ่งหรือหลายอย่างเช่นการติดเชื้อไวรัสหรืออาหาร มีหลักฐานบางอย่างที่ชี้ให้เห็นความสัมพันธ์ระหว่างโรคเบาหวานประเภท 1 และไวรัสคอกซากีกี้บี 4 ซึ่งแตกต่างจากโรคเบาหวานประเภท 2, การโจมตีของโรคเบาหวานประเภท 1 ไม่เกี่ยวข้องกับวิถีชีวิต

2. เบาหวานชนิดที่ 2

เบาหวานชนิดที่ 2 นั้นมีความต้านทานต่ออินซูลินซึ่งอาจรวมกับการหลั่งอินซูลินที่ลดลงได้ การตอบสนองที่บกพร่องของเนื้อเยื่อในร่างกายต่ออินซูลินเชื่อว่าเกี่ยวข้องกับตัวรับอินซูลิน อย่างไรก็ตามยังไม่ทราบข้อบกพร่องเฉพาะ กรณีโรคเบาหวานเนื่องจากข้อบกพร่องที่รู้จักกันจะถูกจัดแยกต่างหาก โรคเบาหวานประเภท 2 เป็นโรคที่พบได้บ่อยที่สุด ในระยะแรกของประเภท 2 ความผิดปกติที่เด่นชัดจะลดความไวของอินซูลิน ในขั้นตอนนี้ภาวะน้ำตาลในเลือดสูงสามารถย้อนกลับได้ด้วยมาตรการและยาหลายอย่างที่ช่วยเพิ่มความไวของอินซูลินหรือลดการผลิตกลูโคสจากตับ โรคเบาหวานประเภท 2 เกิดจากปัจจัยการดำเนินชีวิตและพันธุกรรม จำนวนของปัจจัยการดำเนินชีวิตเป็นที่รู้จักกันว่ามีความสำคัญต่อการพัฒนาของโรคเบาหวานประเภท 2 รวมถึงโรคอ้วน (กำหนดโดยดัชนีมวลกายมากกว่าสามสิบ), ขาดการออกกำลังกาย, อาหารที่ไม่ดี, ความเครียด, และการใช้ชีวิตในเมือง ไขมันในร่างกายส่วนเกินมีความสัมพันธ์กับ 30% ของผู้ป่วยในเชื้อสายจีนและญี่ปุ่น, 60-80% ของผู้ป่วยในเชื้อสายยุโรปและแอฟริกาและ 100% ของ ชาวพิมาอินเดียนและหมู่เกาะแปซิฟิก ซึ่งคนที่ไม่อ้วนมักจะมีอัตราส่วนเอวสะโพกสูง ปัจจัยด้านอาหารมีผลต่อความเสี่ยงในการเกิดโรคเบาหวานประเภท 2 ด้วย การบริโภคเครื่องดื่มที่มีน้ำตาลหวานมากเกินไปนั้นเกี่ยวข้องกับความเสี่ยงที่เพิ่มขึ้น ชนิดของไขมันในอาหารก็มีความสำคัญเช่นกันด้วยไขมันอิ่มตัวและกรดไขมันทรานส์เพิ่มความเสี่ยงและไขมันไม่อิ่มตัวเชิงซ้อนและไขมันไม่อิ่มตัวเชิงเดี่ยวลดความเสี่ยง การรับประทานอาหาร ข้าวขาวจำนวนมากดูเหมือนจะมีบทบาทในการเพิ่มความเสี่ยง อีกทั้งการขาดการออกกำลังกายก็ยังเชื่อว่าเป็นสาเหตุก่อให้เกิดโรคมากขึ้น 7%

3. โรคเบาหวานขณะตั้งครรภ์

โรคเบาหวารขณะตั้งครรภ์คล้ายกับเบาหวานชนิดที่ 2 ในหลายประการที่เกี่ยวข้องกับการรวมกันของการหลั่งอินซูลินที่ค่อนข้างไม่เพียงพอและการตอบสนอง มันเกิดขึ้นประมาณ 2-10% ของการตั้งครรภ์ทั้งหมดและอาจดีขึ้นหรือหายไปหลังคลอด อย่างไรก็ตามหลังจากตั้งครรภ์ประมาณ 5-10% ของผู้หญิงที่เป็นโรคเบาหวานขณะตั้งครรภ์จะพบว่ามีโรคเบาหวานส่วนใหญ่เป็นประเภทที่ 2 เบาหวานขณะตั้งครรภ์สามารถรักษาได้อย่างสมบูรณ์ แต่ต้องมีการดูแลทางการแพทย์อย่างระมัดระวังตลอดการตั้งครรภ์ การจัดการอาจรวมถึงการเปลี่ยนแปลงอาหารการตรวจสอบระดับน้ำตาลในเลือดและในบางกรณีอาจจำเป็นต้องใช้อินซูลิน แม้ว่ามันอาจจะเป็นเพียงชั่วคราวแต่ทว่าโรคเบาหวานขณะตั้งครรภ์ที่ไม่ได้รับการรักษาสามารถทำลายสุขภาพของทารกในครรภ์หรือแม่ได้ ความเสี่ยงต่อทารกรวมถึง macrosomia (น้ำหนักแรกเกิดสูง), หัวใจพิการ แต่กำเนิดและความผิดปกติของระบบประสาทส่วนกลางและความผิดปกติของกล้ามเนื้อโครงร่าง อินซูลินของทารกในครรภ์ที่เพิ่มขึ้นอาจยับยั้งการผลิตสารลดแรงตึงผิวของทารกในครรภ์และทำให้เกิดอาการหายใจลำบาก Hyperbilirubinemia อาจเกิดจากการทำลายเซลล์เม็ดเลือดแดง ในกรณีที่รุนแรงอาจมีการเสียชีวิตจากปริกำเนิดโดยทั่วไปมักจะเป็น ผลของการกระจายรกที่ไม่ดีเนื่องจากการด้อยค่าของหลอดเลือด การเหนี่ยวนำแรงงานอาจถูกระบุด้วยฟังก์ชันรกลดลง การผ่าซีซาร์ (Caesarean) อาจดำเนินการหากมีการกำหนด fetal distress ของทารกในครรภ์หรือเพิ่มความเสี่ยงของการบาดเจ็บที่เกี่ยวข้องกับ macrosomia เช่น shoulder dystocia เป็นต้น

2.2.3 อาการแทรกซ้อน

โรคเบาหวานทุกรูปแบบเพิ่มความเสี่ยงของภาวะแทรกซ้อนในระยะยาวซึ่งจะก่อเกิดขึ้นในระยะเวลาหลายปี (ประมาณ 10 – 20 ปี) ซึ่งอาจจะไม่มีผู้ป่วยได้รับการตรวจก่อนหน้านี้ อาการแทรกซ้อนในระยะยาวอาจส่งผลกระทบต่อเส้นเลือดได้จึงทำให้มีความเสี่ยงต่อการเป็นโรคหลอดเลือดหัวใจจนถึงแก่ชีวิต 75% โรคหลอดเลือดหัวใจใหญ่และหลอดเลือดอื่น ๆ นั้นมีส่วนเกี่ยวข้องกับโรคทางหลอดเลือดเช่นกัน อาการแทรกซ้อนหลักของโรคหลอดเลือดหัวใจกระทบต่อดวงตา ไต และประสาท การบาดเจ็บทางตารู้จักกันในชื่อเบาหวานเข้าจอประสาทตาซึ่งมีสาเหตุจากการบาดเจ็บของเส้นเลือดในดวงตาทำให้การมองเห็นลดลงเรื่อย ๆ มีแนวโน้มสู่อาการตาบอดได้ อาการบาดเจ็บที่ไตเรียกว่าภาวะแทรกซ้อนทางไตของผู้ป่วยเบาหวานทำให้เป็นแผลเป็นเนื้อเยื่อ สูญเสียโปรตีนในปัสสาวะ และเป็นโรคไตเรื้อรังเช่นกัน ซึ่งจำเป็นต้องมีการฝอกหรือเปลี่ยนถ่ายไต ผลกระทบที่ระบบประสาทของร่างกายเรียกว่าโรคเส้นประสาทเหตุเบาหวานที่เป็นอาการแทรกซ้อนของโรคเบาหวานทั่วไป อาการของโรคนี้คือมีการชา ปวด และอาการปวดอื่น ๆ ที่ส่งผลกระทบต่อผิวหนัง โรคเบาหวานที่ส่งผล

กระทบต่อเท้าที่อาจเกิดขึ้นและยากต่อการรักษาซึ่งอาจได้ตัดอวัยวะออก เบาหวานใกล้ปลายประสาทส่งผลให้สูญเสียกล้ามเนื้อและอ่อนแรงอย่างรุนแรง

2.2.4 การวินิจฉัย

เบาหวานเป็นลักษณะของน้ำตาลในเลือดสูงกำเริบหรือถาวรและเป็นการวินิจฉัยโดยแสดงให้เห็นถึงสิ่งใดสิ่งหนึ่งต่อไปนี้:

1) ระดับน้ำตาลกลูโคสในพลาสมาที่เผาผลาญ≥ 7.0 mmol / l (126 mg / dl)

2) พลาสม่ากลูโคส≥ 11.1 มิลลิโมล / ลิตร (200 มก. / ดล.) สองชั่วโมงหลังจาก 75 กรัมกลูโคสในช่องปาก โหลดในการทดสอบความทนทานต่อกลูโคส

3) อาการน้ำตาลในเลือดสูงและระดับน้ำตาลในเลือดในพลาสมา≥ 11.1 มิลลิโมล / ลิตร (200 มก. / ดล)

จากผลลัพธ์ที่เป็นบวกในกรณีที่ไม่มีภาวะน้ำตาลในเลือดสูงที่ไม่ชัดเจนควรได้รับการยืนยันโดยการทำซ้ำของวิธีการอื่น ๆ ข้างต้นในวันที่แตกต่างกัน การวัดระดับกลูโคสในการอดอาหารเป็นเรื่องที่ดีกว่าเพราะความสะดวกในการวัดและความมุ่งมั่นในการทดสอบน้ำตาลกลูโคสอย่างเป็นทางการซึ่งใช้เวลานานสองชั่วโมงและไม่มีข้อได้เปรียบด้านการพยากรณ์โรคในการทดสอบการอดอาหาร ตามคำจำกัดความปัจจุบันการตรวจระดับกลูโคสในการอดอาหารสองครั้งที่สูงกว่า 126 มก. / ดล. (7.0 มิลลิโมล / ลิตร) ถือเป็นการวินิจฉัยโรคเบาหวาน

2.2 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นการวิเคราะห์ชุดข้อมูลเชิงสังเกตการณ์ (ข้อมูลมักจะมีขนาดใหญ่) เพื่อค้นหาความสัมพันธ์ที่ไม่น่าสงสัยและเพื่อสรุปข้อมูลในรูปแบบใหม่ที่ทั้งเข้าใจและมีประโยชน์ต่อเจ้าของข้อมูล การพัฒนาเทคโนโลยีสารสนเทศได้สร้างฐานข้อมูลจำนวนมากและข้อมูลขนาดใหญ่ในพื้นที่ต่างๆ การวิจัยในฐานข้อมูลและเทคโนโลยีสารสนเทศได้ก่อให้เกิดวิธีการจัดเก็บและจัดการกับข้อมูลที่มีค่านี้เพื่อการตัดสินใจต่อไป การทำเหมืองข้อมูลเป็นกระบวนการของการดึงข้อมูลที่มีประโยชน์และรูปแบบจากข้อมูลขนาดใหญ่ มันถูกเรียกว่าเป็นกระบวนการค้นหาความรู้, การขุดความรู้จากข้อมูล, การดึงความรู้หรือการวิเคราะห์ข้อมูลหรือรูปแบบข้อมูล ความสัมพันธ์และบทสรุปที่ได้จากการทำเหมืองข้อมูลมักเรียกว่าแบบจำลองหรือรูปแบบ ตัวอย่างรวมถึงสมการเชิงเส้นกฎกลุ่ม กราฟโครงสร้างต้นไม้และรูปแบบที่เกิดซ้ำในอนุกรมเวลา การทำเหมืองข้อมูลมักตั้งอยู่ในบริบทที่กว้างขึ้นของการค้นหาความรู้ในฐานข้อมูลหรือ Knowledge Discovery in Databases หรือ KDD

คำนี้เกิดขึ้นในสาขาวิจัยปัญญาประดิษฐ์ (Artificial Intelligence: AI) ขั้นตอน KDD เกี่ยวข้องกับหลายขั้นตอนเช่น การเลือกข้อมูลเป้าหมายประมวลผลข้อมูลล่วงหน้าแปลงข้อมูลหากจำเป็นดำเนินการทำเหมืองข้อมูลเพื่อแยกรูปแบบและความสัมพันธ์จากนั้นตีความและประเมินโครงสร้างที่ค้นพบ [4]

2.2.2 วิธีการทำเหมืองข้อมูล

การทำเหมืองข้อมูลจำเป็นต้องมีแนวทางมาตรฐานซึ่งจะช่วยแปลปัญหาทางธุรกิจให้เป็นงานขุดข้อมูลแนะนำการแปลงข้อมูลและเทคนิคการทำเหมืองข้อมูลที่เหมาะสมและจัดหาวิธีการประเมินประสิทธิภาพของผลลัพธ์และจัดเก็บผลลัพธ์ กระบวนการมาตรฐานสำหรับการขุดข้อมูล (CRoss Industry Standard Process for Data Mining: CRISP-DM) ได้แก้ไขปัญหาเหล่านี้โดยการกำหนดรูปแบบกระบวนการซึ่งเป็นกรอบสำหรับการดำเนินโครงการเหมืองข้อมูลซึ่งไม่ขึ้นอยู่กับทั้งภาคอุตสาหกรรมและเทคโนโลยีที่ใช้ แบบจำลองกระบวนการ CRISP-DM มีจุดมุ่งหมายเพื่อสร้างโครงการขุดข้อมูลขนาดใหญ่ต้นทุนน้อยลงเชื่อถือได้มากขึ้นทำซ้ำได้มากขึ้นจัดการได้มากขึ้นและเร็วขึ้น ประกอบไปด้วย 6 ขั้นตอนได้แก่ [5]

1) การเข้าใจปัญหา (Business Understanding)

ขั้นตอนนี้นี้มุ่งเน้นไปที่การทำความเข้าใจวัตถุประสงค์ของโครงการและข้อกำหนดจากมุมมองทางธุรกิจแล้วแปลงความรู้นี้เป็นกำหนดปัญหาการทำเหมืองข้อมูลและแผนการเบื้องต้นที่ออกแบบมาเพื่อให้บรรลุวัตถุประสงค์

2) การเข้าใจข้อมูล (Data Understanding)

ขั้นตอนการทำความเข้าใจข้อมูลเริ่มต้นด้วยการรวบรวมข้อมูลเบื้องต้นและดำเนินการกับกิจกรรมต่างๆเพื่อทำความคุ้นเคยกับข้อมูลระบุปัญหาด้านคุณภาพของข้อมูลค้นหาข้อมูลเชิงลึกแรกของข้อมูลหรือตรวจหาชุดย่อยที่น่าสนใจเพื่อสร้างสมมติฐานสำหรับข้อมูลที่ซ่อนอยู่ ขั้นตอนนี้มีความเชื่อมโยงอย่างใกล้ชิดระหว่างความเข้าใจปัญหาในการกำหนดปัญหาการทำเหมืองข้อมูลและแผนโครงการอย่างน้อยต้องมีความเข้าใจเกี่ยวกับข้อมูลที่มีอยู่ด้วย

3) การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลครอบคลุมกิจกรรมทั้งหมดในการสร้างชุดข้อมูลสุดท้าย (ข้อมูลที่จะป้อนเข้าในเครื่องมือการสร้างแบบจำลอง) จากข้อมูลดิบเริ่มต้น การจัดเตรียมข้อมูลมีแนวโน้มที่จะดำเนินการหลายครั้งและไม่เป็นไปตามลำดับที่กำหนด งานประกอบด้วยการเลือกตารางบันทึกและแอตทริบิวต์ (Attribute) การล้างข้อมูลการสร้างแอตทริบิวต์ใหม่และการแปลงข้อมูลสำหรับเครื่องมือสร้างแบบจำลอง

4) การทำแบบจำลอง (Modeling)

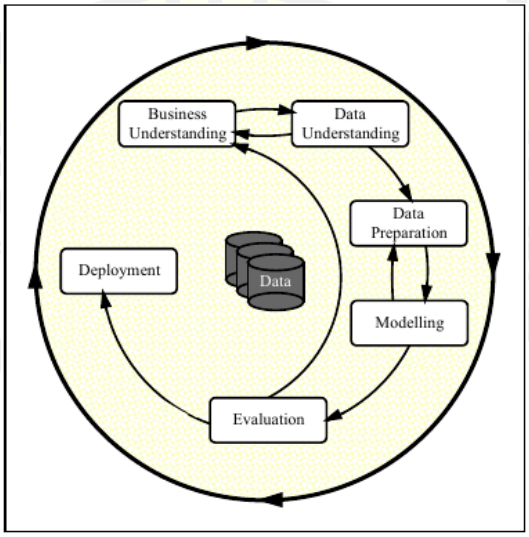
ในขั้นตอนนี้เทคนิคการสร้างแบบจำลองต่างๆจะถูกเลือกและนำไปใช้และพารามิเตอร์ปรับเทียบเป็นค่าที่เหมาะสมที่สุด โดยทั่วไปมีเทคนิคหลายแบบสำหรับประเภทปัญหา การทำเหมืองข้อมูลเดียวกัน เทคนิคบางอย่างต้องการรูปแบบข้อมูลเฉพาะ ขั้นตอนนี้จะมีความเชื่อมโยงกับการเตรียมข้อมูลกันอย่างใกล้ชิด

5) การประเมินผล (Evaluation)

ก่อนที่จะดำเนินการปรับใช้แบบจำลองขั้นสุดท้ายคือสิ่งสำคัญคือต้องประเมินโมเดลอย่างละเอียดถี่ถ้วนมากขึ้นและทบทวนขั้นตอนที่ดำเนินการเพื่อสร้างแบบจำลองเพื่อให้แน่ใจว่าบรรลุวัตถุประสงค์ทางธุรกิจอย่างเหมาะสม วัตถุประสงค์หลักคือเพื่อตรวจสอบว่ามีปัญหาทางธุรกิจที่สำคัญบางอย่างที่ไม่ได้รับการพิจารณาอย่างเพียงพอหรือไม่ ในตอนท้ายของขั้นตอนนี้ควรมีการตัดสินใจเกี่ยวกับการใช้ผลการทำเหมืองข้อมูล

6) การนำไปใช้ (Deployment)

การสร้างแบบจำลองโดยทั่วไปไม่ใช่จุดสิ้นสุดของโครงการ โดยปกติแล้วความรู้ที่ได้รับจากการทำเหมืองข้อมูลจะต้องมีการจัดระเบียบและนำเสนอในรูปแบบที่ลูกค้าสามารถใช้งานได้ ขึ้นอยู่กับข้อกำหนดขั้นตอนการปรับใช้อาจทำได้ง่ายเพียงแค่สร้างรายงานหรือเป็นการนำกระบวนการขุดข้อมูลที่ทำซ้ำได้ ในหลาย ๆ กรณีจะเป็นผู้ใช้งาน ไม่ใช่นักวิเคราะห์ข้อมูลซึ่งจะดำเนินการตามขั้นตอนการนำไปใช้ ไม่ว่าในกรณีใดสิ่งสำคัญคือต้องทำความเข้าใจล่วงหน้าว่าจะต้องดำเนินการใดจึงจะสามารถใช้ประโยชน์จากแบบจำลองที่สร้างขึ้นได้จริง



ภาพประกอบ 1 แบบจำลองจองของกระบวนการ CRISP-DM สำหรับการทำเหมืองข้อมูล ที่มา [5]

2.2.3 การนำไปใช้

การทำเหมืองข้อมูลเป็นเทคโนโลยีที่ค่อนข้างใหม่ที่ยังไม่ได้พัฒนาเต็มที่ อย่างไรก็ตามเรื่องนี้มีจำนวนของอุตสาหกรรมที่ใช้อยู่แล้วเป็นประจำ องค์กรเหล่านี้บางแห่งรวมถึงร้านค้าปลีกโรงพยาบาล ธนาคารและบริษัทประกันภัย องค์กรเหล่านี้หลายแห่งกำลังรวมการทำเหมืองข้อมูลเข้ากับสิ่งต่าง ๆ เช่นสถิติการจดจำรูปแบบและเครื่องมือสำคัญอื่น ๆ การขุดข้อมูลสามารถใช้เพื่อค้นหารูปแบบและการเชื่อมต่อที่หาได้ยาก เทคโนโลยีนี้ได้รับความนิยมจากหลายธุรกิจเพราะช่วยให้พวกผู้ประกอบ การเรียนรู้เพิ่มเติมเกี่ยวกับลูกค้าและทำการตัดสินใจทางการตลาดอย่างชาญฉลาด เป็นต้น

2.3 เทคนิคการทำเหมืองข้อมูลในงานวิจัย

2.3.1 เทคนิคต้นไม้การตัดสินใจ

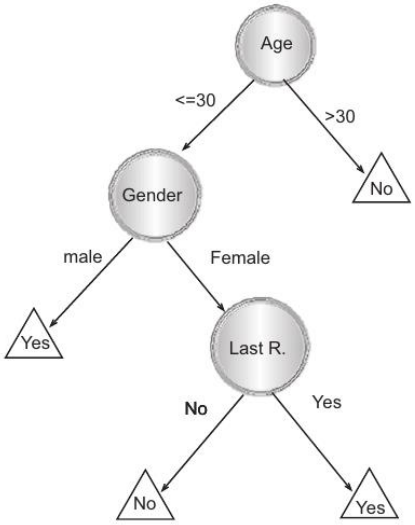
อัชฌาพร กว้างสวาสดและคณะ [6] ได้กล่าวว่าต้นไม้ตัดสินใจเป็นแบบจำลองทางคณิตศาสตร์ เพื่อหาทางเลือกที่ดีที่สุด โดยการนำเอาข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างแบบต้นไม้โดยจะมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างแบบจำลองการจัดหมวดหมู่ (Clustering) ได้จากกลุ่มตัวอย่างที่มีการกำหนดข้อมูลไว้ล่วงหน้า (Training Set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย

วิชญ์วิสิฐ เกสรสิทธิ์และคณะ [7] ได้กล่าวว่าต้นไม้การตัดสินใจ (decision tree) เป็นกฎการตัดสินใจเพื่อหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาวิเคราะห์เพื่อสร้างกฎการตัดสินใจในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้แนะนำการสอนการสร้างโมเดลด้วยวิธีต้นไม้การตัดสินใจจะคัดเลือกตัวแปรที่มีความสัมพันธ์กับกลุ่มมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้หลังจากนั้นก็จะหาตัวแปรที่มีความสัมพันธ์ถัดไปเรื่อย ๆ เพื่อสร้างกิ่งและโหนดใบต่อไป ในการหาความสัมพันธ์ของตัวแปรนี้ใช้ตัววัดที่เรียกว่าอัตราส่วนของค่าเกน (information gain, IG)

ภูมิพัฒน์ ดวงกลาง [8] ได้กล่าวว่า Decision Treeคือเทคนิคการสร้างแบบจำลองที่ใช้ช่วยในการตัดสินใจ โดยแบบจำลองจะอยู่ในรูปแบบโครงสร้างต้นไม้ (Tree) โดยมีใบ (leaf) เป็นสิ่งที่อยู่ล่างสุด บ่งบอกถึงชุดข้อมูลคำตอบ (class) ซึ่งเป็นผลลัพธ์สุดท้ายที่ผ่านการทดสอบเงื่อนไขตามคุณลักษณะของแต่ละโหนด (node)ตามเส้นทางของกิ่งต้นไม้(branch)

จากการศึกษาพบว่าเทคนิคต้นไม้การตัดสินใจ (Decision Tree) สร้างการจำแนกประเภทหรือแบบจำลองการถดถอยในรูปแบบของโครงสร้างต้นไม้โดยจะแบ่งชุดข้อมูลออกเป็นชุดย่อยที่เล็กลงและเล็กลง ในขณะเดียวกันก็มีการพัฒนาโครงสร้างการตัดสินใจที่เกี่ยวข้องขึ้นทีละส่วน ผลลัพธ์

สุดท้ายคือต้นไม้ที่มีโหนดการตัดสินใจและโหนดปลายสุด โหนดการตัดสินใจ มีสองแขนงขึ้นไป โหนดต้นไม้แสดงถึงการจำแนกประเภทหรือการตัดสินใจ โหนดการตัดสินใจบนสุดในแผนภาพต้นไม้ซึ่งสอดคล้องกับตัวทำนายที่ดีที่สุดที่เรียกว่าโหนดราก ต้นไม้การตัดสินใจรองรับได้ทั้งข้อมูลที่เป็นหมวดหมู่และตัวเลข อัลกอริทึมหลักสำหรับการสร้างแผนผังการตัดสินใจที่เรียกว่า ID3 ซึ่งใช้การค้นหาจากบนลงล่างผ่านพื้นที่ของกิ่งก้านที่เป็นไปได้โดยไม่มีการย้อนกลับ ID3 ใช้เอนโทรปีและการรับข้อมูลเพื่อสร้างแผนผังการตัดสินใจ แบบต้นไม้การตัดสินใจจะถูกสร้างขึ้นจากบนลงล่างจากโหนดรูทและเกี่ยวข้องกับการแบ่งส่วนข้อมูลออกเป็นชุดย่อยที่มีอินสแตนซ์ที่มีค่าใกล้เคียงกันหรือหมือนกัน อัลกอริทึม ID3 ใช้เอนโทรปีในการคำนวณความเป็นเนื้อเดียวกันของตัวอย่าง ถ้าตัวอย่างเป็นเนื้อเดียวกันโดยสมบูรณ์ เอนโทรปีจะเป็นศูนย์ และถ้าตัวอย่างมีการแบ่งเท่าๆ กัน จะมีเอนโทรปีเป็นหนึ่ง แบบต้นไม้การตัดสินใจสามารถเปลี่ยนเป็นชุดของกฎได้อย่างง่ายดายโดยการทำแผนที่จากโหนดรากไปยังโหนดปลายสุดทีละรายการ



ภาพประกอบ 2 ตัวอย่างการทำงานของต้นไม้การตัดสินใจ ที่มา [9]

โดยเอนโทรปีโดยใช้ตารางความถี่ของแอตทริบิวต์เดียวจะเขียนสมการเป็น

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2.1}$$

โดยเอนโทรปีโดยใช้ตารางความถี่ของสองแอตทริบิวต์จะเขียนสมการเป็น

$$E(T,X) = \sum_{c \in X} P(c)E(c) \quad (2.2)$$

การรับข้อมูล (Information Gain) ขึ้นอยู่กับการลดลงของเอนโทรปีหลังจากชุดข้อมูลถูกแบ่งตามแอตทริบิวต์ การสร้างโครงสร้างการตัดสินใจเป็นเรื่องเกี่ยวกับการค้นหาแอตทริบิวต์ที่ส่งข้อมูลที่คืนได้รับสูงสุด

$$Gain(T,X) = Entropy(T) - Entropy(T,X) \quad (2.3)$$

2.3.2 เทคนิคนาอีฟ เบย์

อับดุลเลาะ บากาและคณะ [10] ได้กล่าวว่านาอีฟเบย์เป็นเทคนิคการจำแนกข้อมูล โดยมีการตั้งสมมติฐานเพื่อกำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน ซึ่งจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นของแต่ละความสัมพันธ์

จุฑาทิพย์ ทิพย์พูลและ นิเวศ จิระวิชิตชัย [11] ได้กล่าวว่านาอีฟเบย์เป็นวิธีการเรียนรู้ที่ใช้หลักของความน่าจะเป็นตามกฎทฤษฎีของนาอีฟเบย์เข้ามาช่วยในการเรียนรู้เพื่อหาสมมติฐานหนึ่งๆ ร่วมกับข้อมูล การเรียนรู้แบบนาอีฟเบย์อาศัยหลักการของการคำนวณความน่าจะเป็นของแต่ละสมมติฐาน โดยการเรียนรู้แบบนาอีฟเบย์เป็นการเรียนรู้เพิ่มได้ เนื่องจากตัวอย่างใหม่ที่ได้มาถูกนำมาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นทำให้มีการเรียนรู้ที่เปลี่ยนไป

ชลิตา เจริญเนตร และคณะ [12] ได้กล่าวว่า Naive bay คือ เทคนิควิธีการจำแนกที่ไดรับความนิยมและถูกนำมาใช้ อย่างแพร่หลายในงานจำแนกหมวดหมู่เอกสาร เนื่องจากความเรียบงายของขั้นตอนวิธีและให้ประสทธิ ภาพการจำแนกที่ดี Naive bay เป็นขั้นตอนวิธิที่มี พื้นฐานมาจากทฤษฎีเบยส์ (Bayes' Theorem) ซึ่งอาศัยหลักความน่าจะเปนในการทำนายผลลัพธโดยการวิเคราะหความสัมพันธระหวางตัวแปรเพื่อใชในการสรางเงื่อนไขความน่าจะเปนสำหรับแตละความสัมพันธ

วีรยุทธ์ พิมพาภรณ์และพยุง มีสัจ [13] ได้กล่าวว่า Naive bay เป็นการคำนวณหาความหน้าจะเป็นของกลุ้มข้อมูลในรูปแบบของคลาส (class) เมื่อมีการกำหนดแอททิบิวท์ (attribute) และค่าความหน้าจะเป็นของทุกๆ คลาส (class membership probabilities) มาเปรียบเทียบกันแล้วเลือกค่าความน่าจะเป็นที่สูงสุดของคลาสใดๆ มาเป็นผลของการทำนายเพียงค่าเดียวโดยถือว่าค่าคุณสมบัติแต่ละตัวขึ้นต่อกันกับค่าคุณสมบัติอื่น (class conditional independence)

จากการศึกษาพบว่าเทคนิคนาอีฟ เบย์ (Naive Bayes) เป็นกลุ่มของเทคนิคจำแนกประเภทที่รวดเร็วและเรียบง่าย ซึ่งมักจะเหมาะสำหรับชุดข้อมูลที่มีมิติสูงมาก เนื่องจากมีความรวดเร็วและมีพารามิเตอร์ที่ปรับแต่งได้น้อยมาก จึงมีประโยชน์อย่างมากในฐานะพื้นฐานที่รวดเร็วและสกปรก สำหรับปัญหาการจำแนกประเภท เทคนิค Naive Bayes คำนวณโดยใช้ความน่าจะเป็นส่วนหลัง P(y|x) จาก P(c), P(x) และ P(y|c) เทคนิคจำแนกประเภท Naive Bayes ถือว่าผลกระทบของค่าของตัวทำนาย (x) ในคลาสที่กำหนด (y) นั้นไม่ขึ้นกับค่าของตัวทำนายอื่นๆ สมมติฐานนี้เรียกว่าความเป็นอิสระแบบมีเงื่อนไขของคลาส

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{2.4}$$

$$P(y|X) = P(x_1|y) \times P(x_2|y) \times \cdots P(x_n|y) \times P(y) \tag{2.5}$$

โดย

- P(y|x) คือความน่าจะเป็นหลังของคลาสเป้าหมาย ที่กำหนดแอตทริบิวต์ทำนาย
- P(y) คือความน่าจะเป็นก่อนหน้าของคลาส
- P(x|y) คือความน่าจะเป็นของตัวทำนายที่กำหนดคลาส
- P(x) คือความน่าจะเป็นก่อนหน้าของการทำนาย

2.3.3 เทคนิคเพื่อนบ้านใกล้ที่สุด

พัชรียา ทองพูลและคณะ [14] ได้กล่าวว่า เป็นวิธีไม่มีการสร้างตัวแบบจากข้อมูลเรียนรู้เก็บไว้ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลเรียนรู้จำนวน k ตัวที่อยู่ใกล้เคียงกันมากที่สุด ใช้คำตอบของข้อมูลฝึกหัดที่อยู่ใกล้เคียงกันมากที่สุด k ตัว ที่พบมากที่สุดเป็นคำตอบ

จักรกฤษณ์ หงส์เวียงจันทร์และคณะ [15] ได้กล่าวว่าเทคนิคเพื่อนบ้านใกล้ที่สุดเป็นเทคนิคที่ใช้สำหรับการจัดกลุ่มของข้อมูลโดยคำนวณจากระยะห่างของแต่คุณลักษณะในข้อมูล (Data) ซึ่งวิธีนี้จะเหมาะสมกับข้อมูลที่เป็นเชิงตัวเลข การจัดข้อมูลที่อยู่ใกล้กันให้เป็นกลุ่มเดียวกันจะตรวจสอบจากเงื่อนไขของข้อมูลจะตรวจสอบจากจำนวน K ที่กำหนดไว้แต่เทคนิคนี้จะใช้ระยะเวลาในการประมวลผลที่นาน ถ้าข้อมูลมีปริมาณมากอาจเกิดปัญหาในการคำนวณและใช้ปริมาณทรัพยากรใน

การประมวลผลสูงมากเนื่องจากจะใช้เวลาสำหรับการประมวลผลเพิ่มขึ้นตามจำนวนข้อมูลที่เพิ่มขึ้นทั้งหมด

จิตกานต์ จันทราชและคณะ [16] ได้กล่าวว่าเทคนิคเพื่อนบ้านใกล้ที่สุดเป็นการหาระยะห่างระหว่างแต่ละตัวแปรในข้อมูล ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลเชิงตัวเลขและสามารถใช้กับตัวแปรไม่ต่อเนื่อง สามารถจัดในลักษณะพิเศษเพิ่มขึ้นเช่น สี สามารถใช้วิธีเพื่อนบ้านใกล้สุด k ตัว วัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว หลังจากนั้นต้องมีวิธีในการรวมค่าระยะห่างของคุณลักษณะ (attribute) ทุกค่า เพื่อคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่าง ๆ จากนั้นเลือกชุดของเงื่อนไขที่ใช้จัดกลุ่ม (class) มาเป็นฐานสำหรับการจัดกลุ่มในเงื่อนไขใหม่ ๆ จะตัดสินใจได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้นควรมีขนาดใหญ่ และอาจมีการตัดสินใจได้ด้วยการนับจำนวนจุดข้างเคียง

อิทธิพล ดวงแก้ว และ สายยัญ สายยศ [17] ได้กล่าวว่าเทคนิคเพื่อนบ้านใกล้ที่สุดจะตัดสินใจว่า คลาสใดที่สามารถแทนเงื่อนไขหรือกรณีใหม่ๆ ได้โดยวิธีการตรวจสอบจำนวนบางจำนวนk ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุดโดยจะหาผลรวม (count up) ของจำนวนเงื่อนไขหรือกรณีต่างๆ สำหรับแต่ละคลาสแล้วกำหนดเงื่อนไขใหม่ ๆ ให้กับคลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุดเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor Classification หรือ k-NN) เป็นเทคนิคที่ใช้ในการพิจารณาโดยใช้เพื่อนบ้านของ k ที่ใกล้ที่สุดในการพิจารณาคลาส เนื่องจากจำเป็นต้องมีตัวอย่างการฝึกอบรมในขณะรันไทม์นั่นคือต้องอยู่ในหน่วยความจำในขณะรันไทม์ บางครั้งจึงเรียกอีกอย่างว่า Memory-Based Classification เนื่องจากการเหนี่ยวนำล่าช้าในการทำงานจึงถือเป็นเทคนิค Lazy Learning เนื่องจากการจำแนกจะขึ้นอยู่กับตัวอย่างการฝึกอบรมโดยตรงจึงเรียกอีกอย่างว่าการจำแนกตามตัวอย่าง (Example-Based Classification) หรือการจำแนกตามกรณี (Case-Based Classification)

จากการศึกษาพบว่าเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor (KNN)) เป็นเทคนิคที่สามารถใช้ได้ทั้งปัญหาสมการถดถอยและการจำแนกประเภท เทคนิคเพื่อนบ้านใกล้ที่สุดตรวจสอบคลาสกำกับของจุดข้อมูลจำนวนที่เลือกไว้รอบๆ จุดข้อมูลเป้าหมาย เพื่อทำการคาดการณ์เกี่ยวกับคลาสที่จุดข้อมูลเข้า โดยกำหนดแต่ละตัวอย่างในชุดการเรียนรู้เป็นเวกเตอร์สุ่มใน Rn แต่ละจุดมีคำอธิบายเป็น x =< a1(x), a2(x), a3(x),.., an(x) > โดยที่ ar(x) หมายถึงค่า l ของแอตทริบิวต์ rth ar(x) สามารถเป็นได้ทั้งตัวแปรเชิงปริมาณหรือเชิงคุณภาพ เพื่อกำหนดคลาสของจุดสืบค้น xq แต่ละจุดที่ใกล้ที่สุด k x1,...,xk ถึง xq ดำเนินการโหวตคลาสของ xq สอดคล้องกับคลาสส่วนใหญ่ ในการทำงานของเทคนิคเพื่อนบ้านใกล้ที่สุดในขั้นตอนแรกเลือก K และกำหนดตัวเทคนิคในพิจารณาจำนวน k (เพื่อนบ้านที่มีจุดข้อมูลรอบข้างกี่จุด) เมื่อตัดสินเกี่ยวกับกลุ่มตัวอย่าง ที่เป็นของเป้าหมาย ในขั้นตอนที่สองแบบจำลองจะตรวจสอบระยะห่างระหว่างตัวอย่างเป้าหมายกับทุกตัวอย่างในชุดข้อมูล ระยะทางจะถูกเพิ่มลงในรายการและจัดเรียง หลังจากนั้น รายการที่เรียงลำดับจะถูกตรวจสอบและป้ายกำกับสำหรับองค์ประกอบ K ด้านบนจะถูกส่งกลับ กล่าวอีกนัยหนึ่ง หากตั้งค่า K เป็น 5 โมเดล

จะตรวจสอบป้ายกำกับของจุดข้อมูลที่ใกล้เคียงที่สุด 5 อันดับแรกไปยังจุดข้อมูลเป้าหมาย เมื่อแสดงการคาดการณ์เกี่ยวกับจุดข้อมูลเป้าหมาย สิ่งสำคัญคืองานนั้นเป็นงานการถดถอยหรือการจัดหมวดหมู่ สำหรับงานถดถอย ใช้ค่าเฉลี่ยของป้ายกำกับ K ระดับบนสุด ในขณะที่โหมดของป้ายกำกับ K ระดับบนสุดจะใช้ในกรณีของการจำแนกประเภท การดำเนินการทางคณิตศาสตร์ที่แน่นอนที่ใช้ในการดำเนินการ KNN จะ Euclidean เป็นหลักในการวัดเมตริกระยะทาง

$$\sqrt{\sum_{i=1}^{k}(x_1 - y_1)^2} \tag{2.6}$$

2.3.4 กระบวนการโหวตรวม (Voting Ensemble Method)

ธนพัฒน์ ทองมา [18] ได้กล่าวว่า Voting ensemble เป็นเทคนิคการทำนายข้อมูลโดยการสร้างโมเดลจำแนกประเภทข้อมูลหลายโมเดลโดยใช้ชุดข้อมูลฝึกฝนชุดเดียวกันด้วยเทคนิคการวิเคราะห์มากกว่าหนึ่งเทคนิค เพื่อช่วยกันทำนายกลุ่มประเภทข้อมูลด้วยวิธีการโหวตเสียงข้างมาก เทคนิคนี้มีประสิทธิภาพดีกว่าการใช้โมเดลจำแนกประเภทข้อมูลเพียงโมเดลเดียว

จิราภา เลาหะวรนันท์และคณะ [19] ได้กล่าวว่า Voting ensemble เป็นการสร้างโมเดลด้วยเทคนิคต่างๆกันโดยใช้ชุดข้อมูลสำหรับสอน (Training Data) ชุดเดียวกัน ซึ่งเป็นวิธีการที่ใช้ชุดข้อมูลสอนชุดเดียวกัน มาสร้างโมเดลจากเทคนิคการพยากรณ์จากเทคนิคการจำแนก เพื่อให้โมเดลมีความหลากหลายมากขึ้น

นรินทร์ พนาวาส [20] ได้กล่าวว่า Voting ensemble คือการรวมการทำนายของแต่ละเทคนิคการจำแนกเพื่อให้ได้ผลลัพธ์การทำนายที่ดีที่สุดโดยเหลือเทคนิคเดียว โดยแต่ละเทคนิคจะมีความเชี่ยวชาญในแต่ละพื้นที่การทำนายจะปฏิบัติต่างกันภายใต้ความเอนเอียงทางทฤษฎี เทคนิคเหล่านั้นจะประกอบไปด้วยแนวทางการสร้างการจำแนกโหวตรวมที่กฎที่เหนือกว่าในแต่ละกฎ

จากการศึกษาพบว่าเทคนิคโหวตร่วม (Voting Ensemble) เป็นแบบจำลองการเรียนรู้ของเครื่องทั้งมวลที่รวมการทำนายจากแบบจำลองอื่น ๆ ไว้ด้วยกัน เป็นเทคนิคที่อาจนำไปใช้ในการปรับปรุงประสิทธิภาพของโมเดล เพื่อให้ได้ประสิทธิภาพที่ดีกว่าแบบจำลองเดียวที่ใช้ในแบบจำลองร่วม การโหวตร่วมทำงานโดยการรวมการทำนายจากแบบจำลองต่างๆ แบบจำลองนี้สามารถใช้สำหรับแบบจำลองการจำแนกประเภทหรือสมการถดถอย ในกรณีของการถดถอย จะเกี่ยวข้องกับการคำนวณค่าเฉลี่ยของการคาดคะเนจากตัวแบบ ในกรณีของการจัดประเภท การคาดคะเนสำหรับแต่ละคลาสกำกับจะถูกรวมเข้าด้วยกัน และคาดการณ์คลาสกำกับที่มีคะแนนเสียงข้างมาก โดยโหวตจะแบ่งเป็น 2 แบบ คือโหวตแบบยาก (hard vote) และโหวตแบบง่าย (soft vote) โหวตแบบยาก

จะเกี่ยวข้องกับการสรุปผลการทำนายสำหรับแต่ละคลาสกำกับและการทำนายคลาสกำกับที่มีการโหวตมากที่สุดจากสมการด้านล่าง โดย C(X) คือ การออกเสียงส่วนใหญ่ของแต่ละเทคนิคการจำแนก h_x [นรินทร์ พนาวาส]

$$C(X) = mode\{h_1(X), h_2(X), h_3(X)\} \quad (2.7)$$

การโหวตไม่รับประกันว่าจะให้ประสิทธิภาพที่ดีกว่าแบบจำลองเดียวที่ใช้ในในการโหวตร่วม หากแบบจำลองใดแบบจำลองหนึ่งที่ใช้ในการโหวตร่วมทำงานได้ดีกว่าชุดลงคะแนนเสียง ควรใช้แบบจำลองนั้นแทนชุดลงคะแนนเสียง การโหวตสามารถให้ค่าความแปรปรวนที่ต่ำกว่าในการคาดการณ์ที่ทำขึ้นในแต่ละแบบจำลอง สามารถเห็นได้ในความแปรปรวนที่ต่ำกว่าในข้อผิดพลาดในการทำนายสำหรับการทำแบบจำลองสมการถดถอย นอกจากนี้ยังสามารถเห็นได้ในความแปรปรวนที่ต่ำกว่าในด้านความแม่นยำสำหรับงานจำแนกประเภท ความแปรปรวนที่ต่ำกว่านี้อาจส่งผลให้การโหวตมีประสิทธิภาพเฉลี่ยลดลง การโหวตมีประโยชน์อย่างยิ่งสำหรับโมเดลแมชชีนเลิร์นนิงที่ใช้อัลกอริธึมการเรียนรู้แบบสุ่ม และส่งผลให้โมเดลสุดท้ายแตกต่างกันในแต่ละครั้งที่ได้รับการฝึกในชุดข้อมูลเดียวกัน เมื่อการโหวตร่วมคือเมื่อรวมอัลกอริธึมการเรียนรู้ของเครื่องเดียวกันหลายชุดเข้ากับไฮเปอร์พารามิเตอร์ที่แตกต่างกันเล็กน้อย การโหวตร่วมจะมีประสิทธิภาพสูงสุดเมื่อการรวมหลายชุดของแบบจำลองที่ได้รับการฝึกฝนโดยใช้อัลกอริธึมการเรียนรู้แบบสุ่มและการผสมผสานของแบบจำลองที่มีไฮเปอร์พารามิเตอร์ต่างกัน ข้อจำกัดของการโหวตร่วมคือจะปฏิบัติต่อโมเดลทั้งหมดเหมือนกัน หมายความว่าโมเดลทั้งหมดมีส่วนร่วมในการทำนายอย่างเท่าเทียมกัน นี่เป็นปัญหาบางแบบจำลองบางแบบดีในบางสถานการณ์และไม่ดีในบางสถานการณ์

2.3.5 เทคนิคป่าสุ่ม

Md. Aminul Islam Nusrat Jahan [21] กล่าวว่าเทคนิคต้นไม้ป่าสุ่ม (Random Forest) เกิดจากการรวมตัวกันของการทำนายของต้นไม้การตัดสินใจขึ้นอยู่กับค่าการสุ่มเวคเตอร์อย่างอิสระกับการกระจายของต้นไม้โดยมีส่วนสำคัญในการเพิ่มประสิทธิภาพการนำแนกข้อมูลมีความแม่นยำส่งผลให้มีมีการเพิ่มจำนวนของต้นไม้การตัดสินใจที่ใช้ในการโหวตคลาสที่โดดเด่นที่สุด จำลองจะสร้างเวกเตอร์สุ่มที่ควบคุมการสร้างแบบต้นไม้การตัดสินใจแต่ละต้นในกลุ่ม เทคนิคต้นไม้ป่าสุ่มจะมีกลไก

การลงคะแนนสำหรับการเลือกคลาสที่ได้รับความนิยมมากที่สุดหลังจากสร้างแบบต้นไม้การตัดสินใจจำนวนมาก

Austin Haynesworth [22] กล่าวว่าเทคนิคต้นไม้ป่าสุ่ม (Random Forest) เป็นเทคนิคประเมินค่าประชากรทางสถิติโดยแทนค่าชุดข้อมูลด้วยการสุ่มโดยเฉพาะแบบต้นไม้การตัดสินใจในเทคนิคป่าสุ่มที่มีพื้นฐานจากชุดข้อมูลบูตสแตรปและสุ่มตัวแปรการทำนายจนถึงผลลัพธ์ก การจำแนกครั้งสุดท้ายโดยจะกำหนดการออกเสียงของแบบต้นไม้การตัดสินใจทั้งหมด

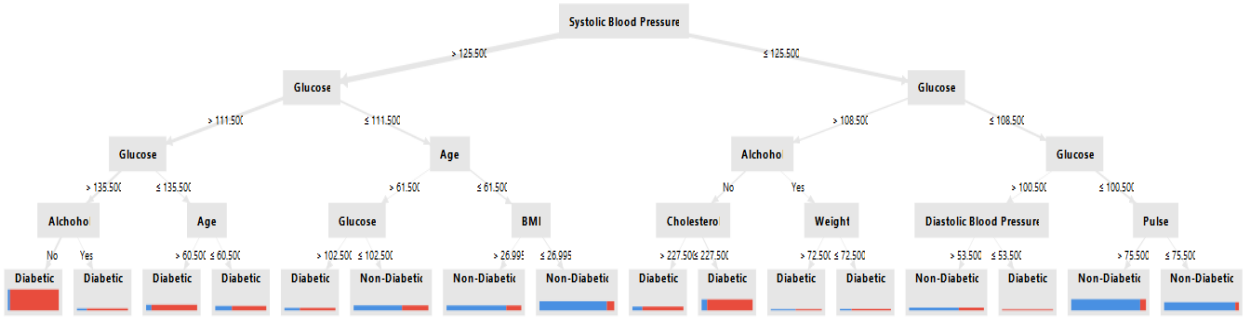
IFTIKHAR AHMAD และคณะ [23] กล่าวว่าเทคนิคต้นไม้ป่าสุ่ม (Random Forest) เป็นเทคนิคที่ใช้ในการวิเคราะห์การจำแนกและสมการถดถอย ทำงานโดยการสร้างต้นไม้การตัดสินใจในการฝึกข้อมูลและผลิตออกเป็นคลาสกำกับที่เป็นการออกเสียงส่วนใหญ่ เทคนิคต้นไม้ป่าสุ่มสามารถสร้างผลลัพธ์การจำแนกที่มีความแม่นยำสูงและรองรับปัญหาข้อมูลสุดโต่งและข้อมูลไม่สะอาดได้

WEIWEI LIN และคณะ [24] กล่าวว่าเทคนิคต้นไม้ป่าสุ่ม (Random Forest) ถูกสร้างขึ้นจากแบบต้นไม้การตัดสินใจหลากหลายโดยไม่ลดทอนกระบวนการ โดยจำนวนแบบต้นไม้การตัดสินใจมากจะทำให้ความแม่นยำแบบจำลองมีมากขึ้นและไม่มีปัญหาการโอเวอร์ฟิต เทคนิคต้นไม้ป่าสุ่มจะทำการประมาณการโดยรวมและมีข้อดีของระบบการเลือกคุณลักษณะอัตโนมัติฯลฯ

JIANGTAO MA และคณะ [25] กล่าวว่าเทคนิคต้นไม้ป่าสุ่ม (Random Forest) ถูกสร้างขึ้นจากแบบต้นไม้การตัดสินใจจำนวนนมากในเทคนิคต้นไม้ป่าสุ่มโดยไม่สัมพันธ์กัน เมื่อนำเข้าข้อมูลสู่เทคนิคต้นไม้ป่าสุ่มแบบต้นไม้การตัดสินใจแต่ละแบบจะตัดสินคลาสกำกับถูกแบ่งและจำแนกข้อมูลเพื่อโหวตตามคลาสกำกับ

เทคนิคป่าสุ่ม (Random Forest) ถูกนำเสนอโดย Leo Breiman [26] ถูกสร้างขึ้นมาเพื่อปัญหาการจำแนกและสมการถดถอย มีลักษณะคล้ายเทคนิคต้นไม้การตัดสินใจอันเนื่องมาจากการรวมกันของต้นไม้อันเกิดจากการฝึกข้อมูลของบรรจุถุง (Bagging) หรือบูตสแตรป (bootstrap) ความแตกต่างหลักระหว่างเทคนิคต้นไม้การตัดสินใจและเทคนิคป่าสุ่มคือ การสร้างโหนดรูทและการแยกโหนดจะทำแบบสุ่มในภายหลัง พบว่าเทคนิคป่าสุ่มจะใช้วิธีการบรรจุถุงเพื่อสร้างการคาดการณ์ที่จำเป็น บรรจุถุงจะมีส่วนช่วยในการสุ่มชุดข้อมูลฝึกจำนวนหนึ่งจะประกอบด้วยการสังเกตและคุณลักษณะที่ใช้ในการทำนายเพื่อผลิตผลลัพธ์ในรูปแบบต้นไม้การตัดสินใจโดยขึ้นอยู่กับชุดข้อมูลฝึกที่ป้อนเข้าไปในเทคนิคป่าสุ่มซึ่งจะเรียงลำดับผลลัพธ์ที่สูงที่สุดจะเป็นผลลัพธ์สุดท้าย ในเทคนิคป่าสุ่มของเทคนิคการจำแนกประเภทใช้วิธีการทั้งมวล (Ensemble) เพื่อให้ได้ผลลัพธ์ ชุดข้อมูลฝึกจะถูกป้อนเพื่อฝึกแบบต้นไม้การตัดสินใจต่างๆ ซึ่งข้อมูลชุดนี้ประกอบด้วยการสังเกตและคุณลักษณะที่จะถูกเลือกแบบสุ่มในระหว่างการแยกโหนด ระบบป่าฝนอาศัยต้นไม้ตัดสินใจต่างๆ แผนผังการตัดสินใจทั้งหมดประกอบด้วยโหนดการตัดสินใจ โหนดปลายสุด และโหนดราก โหนดปลายสุดของแผนผังต้นไม้แต่ละต้นเป็นผลลัพธ์สุดท้ายที่สร้างโดยแผนผังการตัดสินใจ การเลือกผลลัพธ์สุดท้ายเป็นไปตาม

ระบบการลงคะแนนเสียงข้างมาก ในกรณีนี้ผลลัพธ์ที่เลือกโดยต้นไม้ตัดสินใจส่วนใหญ่จะกลายเป็นผลลัพธ์สุดท้าย



ภาพประกอบ 3 ตัวอย่างการทำงานของเทคนิคป่าสุ่ม

2.3.8 การวัดประสิทธิภาพแบบจำลอง

การวัดประสิทธิภาพแบบจำลองคือมาตรการในการประเมินว่าอัลกอรึทึมการจำแนกมีความถูกต้อง เพียงใดในการทำนายลาเบลคลาสของแถวข้อมูลในตาราง หากพิจารณากรณีที่มีการกระจายตัวของคลาสมากหรือน้อยอย่างเท่าเทียมกันรวมทั้งกรณีที่คลาสไม่สมดุลกัน ซึ่งรวมถึงค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) ค่า และค่าประสิทธิภาพโดยรวม (F-Measure) การใช้ข้อมูลฝึกเพื่อหาค่าจำแนกและประเมินความถูกต้องของแบบจำลองที่เรียนรู้ที่เกิดขึ้นอาจส่งผลให้เกิดการประมาณการที่เข้าใจผิดมากเกินไปเนื่องจากอัลกอริทึมการเรียนรู้ที่มีความเชี่ยวชาญมากเกินไป แต่จะดีกว่าถ้าวัดความแม่นยำของการจำแนกในชุดทดสอบซึ่งประกอบด้วยลาเบลกำกับคลาสซึ่งไม่ได้ใช้ในการฝึกโมเดล ในปัญหาการจำแนกแหล่งที่มาหลักของการวัดประสิทธิภาพคือตาราง Confusion Matrix รูปที่ 2.9 แสดงตารางสำหรับปัญหาการจำแนกของตาราง Confusion Matrix 2 ระดับที่ใช้บ่อยที่สุดซึ่งสามารถคำนวณได้จากตาราง Confusion Matrix มีดังต่อไปนี้ [27]

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Positive (FN)	True Negative (TN)

ภาพประกอบ 4 ตาราง Confusion Matrix ที่มา [27]

จากภาพกรอบซ้ายบนไปขวาล่างแสดงถึงการตัดสินใจที่ถูกต้องและกรอบที่อยู่ด้านนอกแสดงถึงข้อผิดพลาด อัตราการทำนายถูกต้อง (True Positive Rate: TPR) ของการจำแนกประเภทข้อมูลจะหาค่าโดยการหารผลบวกที่ถูกจำแนกประเภทข้อมูลอย่างถูกต้องของ TP ด้วยจำนวน TP ทั้งหมด อัตราการทำนายผิด (False Positive Rate: FPR) ของตัวจำแนกถูกหาค่าโดยการหารเชิงลบที่จำแนกประเภทข้อมูลประเภทไม่ถูกต้องของ FPR ด้วยผลลบทั้งหมด ความแม่นยำโดยรวมของการจำแนกประเภทข้อมูลหาค่าโดยการหารการทำนายถูกและผิดที่ถูกจำแนกประเภทอย่างถูกต้องด้วยจำนวนชุดข้อมูลทั้งหมด การวัดประสิทธิภาพอื่น ๆ ในการคำนวณประสิทธิภาพแบบจำลองด้วยเช่นกัน

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \tag{2.8}$$

$$\text{True Negetive Rate} = \frac{TN}{TN+FP} \tag{2.9}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.10}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2.12}$$

$$F - \text{measure} = \frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}} \tag{2.13}$$

โดยที่

- True Positive (TP): หมายถึงคลายเป้าหมายเป็นถูกและแบบจำลองทำนายเป็นถูก
- True Negatives (TN): หมายถึงคลายเป้าหมายเป็นถูกและแบบจำลองทำนายเป็นผิด
- False Positive (FP): หมายถึงคลายเป้าหมายเป็นผิดและแบบจำลองทำนายเป็นผิด
- False Negative (FN): หมายถึงคลายเป้าหมายเป็นผิดและแบบจำลองทำนายเป็นถูก

2.5 งานวิจัยที่เกี่ยวข้อง

Vanishri Arun และคณะ (2017) [28] ได้ใช้การทำเหมืองข้อมูลในการพยากรณ์และจำแนกผู้ป่วยโรคเบาหวานโดยนำเทคนิคการวิเคราะห์องค์ประกอบพื้นฐานและแบบจำลองการจำแนกผสมในการค้นหาจำนวนขั้นต่ำของของแอททริบิวท์เพื่อปรับปรุงให้มีความเร็วมากขึ้นผ่านเทคนิคจำแนกข้อมูล จากการทดลองพบว่า Naïve Bayes ให้ค่าความถูกต้อง 75.52%, Linear regression, Quadratic Discriminant Analysis ให้ค่าความถูกต้อง 55.94%, Support Vector Machine ให้ค่าความถูกต้อง 76.95%, K-Nearest Neighbor ให้ค่าความถูกต้อง 68.23%, ต้นไม้การตัดสินใจให้ค่าความถูกต้อง 74.35% และ Hierarchical Majority Voting (HMV) โดยให้ค่าความถูกต้องที่ 75.52%

Ramya Akula และคณะ (2019) [29] ได้ทำการวินิจฉัยผู้ป่วยโรคเบาหวานระยะที่ 2 โดยใช้ข้อมูลผู้ป่วยเบาหวานจำนวน 10,000 คน ตั้งแต่ปี 2009 ถึง 2012 ประกอบไปด้วยตัวแปร อายุ ความดันโลหิต diastolic และ systolic เพศ ความสูงและน้ำหนัก จากการทดลองพบว่า K-Nearest neighbor ให้ค่าความถูกต้อง 69.41%, Support Vector Machines ให้ค่าความถูกต้อง 34.27% ,Decision Tree ให้ค่าความถูกต้อง 72.60%, Random Forest ให้ค่าความถูกต้อง 71.23% Gradient Boosting ให้ค่าความถูกต้อง 68.49%, Neural Network ให้ค่าความถูกต้อง 34.25%,

Naive Bayes ให้ค่าความถูกต้อง 68.03% รวมถึงแบบจำลอง Voting Ensemble ทำให้การทำนายผู้ป่วยเบาหวานชนิดที่ 2 มีความแม่นยำถึง 85%

Kawsar Ahmed และ Tasnuba Jesmin (2014) [30] ได้เปรียบเทียบการวิเคราะห์ของเทคนิคการจำแนกเหมืองข้อมูลในการพยากรณ์ข้อมูลโรคเบาหวานชนิดที่ 2 ในประชากรบังกลาเทศ ในข้อมูลที่ใช้ประกอบไปด้วยข้อมูลผู้ป่วย 400 คน แบ่งเป็น 200 คนที่เป็นเบาหวานและ 200 คนที่ไม่ได้เป็นโรคเบาหวาน อัลกอริทมที่ใช้ในการพยากรประกอบไปด้วย Bayes Classifiers, Trees Classifiers, Rules Classifiers, Functions Classifiers, Lazy Classifiers, Miscellaneous classifiers, Metalearning Classifiers จากการทดสอบได้ใช้ 20 อัลกอริทึมที่ใช้ในการการจำแนกข้อมูลผู้ป่วยโรคเบาหวาน พบว่ามี 5 อัลกอริทึมใน 3 กรณีของชุดข้อมูลฝึกทั้งหมดได้แบ่งเปอร์เซ็ตต์และทดสอบประสิทธิภาพ 10 fold cross validation พบว่าอัลกกอริทึมการจำแนก Bagging มีค่าความถูกต้อง 80.82% Logistic Regression และ Multiclass Classifier มีค่าความถูกต้อง 79.17% และ Random Tree มีค่าความถูกต้อง 80%

Ashok Kumar และ R. Govindasamy (2015) [31] ได้ใช้เทคนิคการจำแนกในการทดสอบประสิทธิภาพจากฐานข้อมูลผู้ป่วยเบาหวาน UCI 768 คน โดยใช้การเลือกคุณสมบัติพิเศษชุดย่อย (ข้อมูล) ของผู้ป่วยโรคเบาหวานจากชุดข้อมูลผู้ป่วยเบาหวานทั้งหมดจะได้รับซึ่งประกอบด้วยเฉพาะคุณลักษณะที่สำคัญจากนั้นใช้เทคนิคการจำแนกประเภทที่ได้รับชุดข้อมูลย่อยที่มีนัยสำคัญ จากการทดลองพบว่า Support Vector Machine มีค่าความถูกต้อง 77.73% Regression มีค่าความถูกต้อง 77.60% Bayesian Network มีค่าความถูกต้อง 78.25% Decision Table มีค่าความถูกต้อง 79.81% และ Naïve Bayes มีค่าความถูกต้อง 77.60%

V.Karthikeyani และคณะ (2014) [32] ใช้การทำเหมืองข้อมูลในเปรียบเทียบอัลกอริทึมการจำแนกแบบมีผู้สอนและไม่มีผู้สอนโดยนำข้อมูลผู้ป่วยโรคเบาหวาน Pima Indian จำนวน 768 คน จากนั้นจึงวัดประสิทธิภาพด้วย 10-fold Cross Validation จากการทดลองพบว่า Support Vector Machine มีค่าความถูกต้อง 74.80%, Prototype Neural Network มีค่าความถูกต้อง 67%, Logistic regression มีค่าความถูกต้อง 75% และ Multinomial Regression มีค่าความถูกต้อง 75%

K. Saravananathan และ T. Velmurugan (2016) [33] ได้นำการทำเหมืองข้อมูลมาใช้วินิจฉัยผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมการจำแนกและวัดประสิทธิภาพการจำแนกประเภทของข้อมูลเพื่อหาเทคนิคการจำแนกข้อมูลที่ดีที่สุด จากการทดลองพบว่าอัลกอริทึม J48 มีค่าความถูกต้อง 67.15%, การจำแนกและสมการทดถอยแบบต้นไม้ (Classification and Regression Tree (CART)) มีค่าความถูกต้อง 62.28%, Support Vector Machines มีค่าความถูกต้อง 65.04%, k-Nearest Neighbor มีค่าความถูกต้อง 53.39%

Ratna Patil และ Sharavari Tamane (2018) [34] เสนอการศึกษาเชิงทดลองเกี่ยวกับอัลกอริทึมต่างๆซึ่งจำแนกข้อมูลของโรคเบาหวานได้อย่างมีประสิทธิผลโดยใช้เทคนิคการจำแนกเพื่อระบุข้อดีและข้อ เสีย อีกทั้งประเมินประสิทธิภาพของอัลกอริทึมที่มีอยู่จะดำเนินการเพื่อกำหนดแนวทางที่ดีที่สุด จากการทดลองพบว่าอัลกอริทึมสมการถดถอยและ Gradient Boost มีค่าความถูกต้อง 79%, k-Nearest Neighbor มีค่าความถูกต้อง 74%, Linear Support Vector Machine มีค่าความถูกต้อง 67.79%, Decision tree มีค่าความถูกต้อง 73.16%, Multilayer Perception มีค่าความถูกต้อง 64%, Random Forest มีค่าความถูกต้อง 76.19%, Gaussian Naïve Bayes มีค่าความถูกต้อง 76%

G.Visalatchi และคณะ (2014) [35] ได้ใช้การทำเหมืองข้อมูลในการทดสอบประสิทธิภาพข้อมูลผู้ป่วยโรคเบาหวานโดยใช้เทคนิคในการจำแนกแบบมีผู้สอนเพื่อเปรียบเทียบประสิทธิภาพความแม่นยำของอัลกอริทึม จากการทดลองพบว่าอัลกอริทึมต้นไม้การตัดสินใจ C4.5 มีค่าความถูกต้อง 86%, Support Vector Machine มีค่าความถูกต้อง 75%, K-Nearest Neighbor มีค่าความถูกต้อง 78%, Naïve Bay มีค่าความถูกต้อง 76%, Apriori มีค่าความถูกต้อง 75% จากการทดลองอัลกอริทึมต้นไม้การตัดสินใจ C4.5 มีค่าความถูกต้องสูงสุดจึงสามารถนำไปปรับปรุงประสิทธิภาพการจำแนกต่อไป

Nilesh Jagdish Vispute และคณะ (2015) [36] ได้ทำเหมืองข้อมูลโดยใช้โปรกแกรม WEKA นำชุดข้อมูลผู้ป่วยโรคเบาหวานมาทำการจำแนกและเปรียบเทียบประสิทธิภาพในแต่ละเทคนิค อีกทั้งวัดประสิทธิภาพข้อมูลเพื่อผลความแม่นยำในการทำนายผลของอัลกอริทึม จากการทดลองพบว่า Naïve Bayes มีค่าความถูกต้อง 76.30%, ต้นไม้การตัดสินใจ J48 มีค่าความถูกต้อง 73.82%, Sequential minimal optimization (SMO) มีค่าความถูกต้อง 77.34%, REP Tree มีค่าความถูกต้อง 75.26% และ Random Tree มีค่าความถูกต้อง 68.09%

S.Selvakumar และคณะ (2017) [37] ใช้การทำเหมืองข้อมูลในการทำนายการป่วยโรคเบาหวานโดยนำอัลกอริทึมการจำแนกและเปรียบเทียบประสิทธิภาพของค่าความถูกต้อง ข้อมูลที่ใช้ศึกษาเป็น UCI machine learning repository จำนวน 89 เรคคอร์ด การทดสอบโรคเบาหวาน 318 รายถูกแยกเพื่อบันทึกผู้ป่วยอีกทั้งได้นำวิธี Adaptive Neuro Fuzzy Inference System (ANFIS) และ Rough Setเพื่อกำหนดปริมาณแผน จากการทดลองพบว่า Binary Logistic Regression มีค่าความถูกต้อง 69%, Multilayer Perceptron มีค่าความถูกต้อง 71% และ K-Nearest Neighbor มีค่าความถูกต้อง 80%

Sonu Bala Garga และคณะ (2017) [38] ใช้การทำเหมืองข้อมูลในการตรวจหาผู้ป่วยเบาหวานโดยใช้อัลกอริทึมการจำแนกและข้อมูลเบาหวานจำนวน 768 ตัวอย่าง เพื่อหาอัลกอริทึมที่ดีที่สุดในการจำแนกผู้ป่วยจากการวัดประสิทธิภาพ 10-cross validation และวิธีแยก Percentage

จากการทดลองโดยใช้วิธีการวัดประสิทธิภาพ 10-cross validation พบว่า Decision table มีค่าความถูกต้อง 71.22%, Naïve Bay มีค่าความถูกต้อง 76.30%, Bayes Net มีค่าความถูกต้อง 74.34%, J48 Tree มีค่าความถูกต้อง 73.82%, Multilayer Perceptron มีค่าความถูกต้อง 75.39%, SMO มีค่าความถูกต้อง 77.34%, Random Forest มีค่าความถูกต้อง 75.78% และเมื่อใช้วิธีแยก Percentage พบว่า Decision table มีค่าความถูกต้อง 82%, Naïve Bay มีค่าความถูกต้อง 77%, Bayes Net มีค่าความถูกต้อง 78.16%, J48 Tree มีค่าความถูกต้อง 76.24%, Multilayer Perceptron มีค่าความถูกต้อง 74.32%, SMO มีค่าความถูกต้อง 79.31%, Random Forest มีค่าความถูกต้อง 78.54%

รุ่งโรจน์ บุญมา และ นิเวศ จิระวิชิตชัย (2562) [39] ได้สร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานจากกลุ่มตัวอย่าง 768 คนจากฐานข้อมูล UCI โดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูลและทำการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท จากการทดลองพบว่า Support Vector Machine มีค่าความถูกต้อง 76.95% Decision Tree มีค่าความถูกต้อง 68.62% K-Nearest Neighbor มีค่าความถูกต้อง 75.79% และ Naïve Bay มีค่าความถูกต้อง 75.79% สามารถนำผลที่ได้จากงานวิจัยนี้ไปประยุกต์ใช้ในการคัดกรองและสร้างระบบสนับสนุนการตัดสินใจในส่วนของแนวทางการรักษาของแพทย์ต่อไป

จากการทบทวนวรรณกรรมข้างต้น สามารถสรุปกลุ่มการทำเหมืองข้อมูลผู้ป่วยโรคเบาหวานได้ดังนี้ การวิจัยในต่างประเทศ เช่น ประเทศอินเดียและบังกลาเทศ นักวิจัยใช้การทำเหมืองข้อมูลในการเปรียบเทียบอัลกอริทึมการจำแนกแบบมีผู้สอนและไม่มีผู้สอน โดยใช้วิธี Support Vector Machine, Neural Network, Logistic Regression และ Multinomial Regression เพื่อวัดประสิทธิภาพด้วย 10-fold Cross Validation โดยใช้ข้อมูลผู้ป่วยโรคเบาหวาน Pima Indian จำนวน 768 คน เช่นเดียวกันกับนักวิจัยในประเทศบังกลาเทศการจำแนกข้อมูลผู้ป่วยจำนวน 400 คน ที่เป็นโรคเบาหวาน 200 คนและไม่เป็นโรคเบาหวาน 200 คน Algorithm ที่ใช้ในการพยากรณ์ประกอบไปด้วย Bayes Classifiers, Trees Classifiers, Rules Classifiers, Functions Classifiers, Lazy Classifiers, Miscellaneous classifiers, Meta Learning Classifiers และทดสอบประสิทธิภาพด้วย 10-fold Cross Validation (V.Karthikeyani และคณะ (2014); Kawsar Ahmed และ Tasnuba Jesmin (2014))

ในประเทศไทยมีนักวิจัยที่ใช้การทำเหมืองข้อมูลในการพยากรณ์และจำแนกผู้ป่วยโรคเบาหวานโดยนำเทคนิคการวิเคราะห์องค์ประกอบพื้นฐานและแบบจำลองการจำแนกผสมในการค้นหาจำนวนขั้นต่ำของของAttribute เพื่อปรับปรุงให้มีความเร็วมากขึ้นผ่านเทคนิคจำแนกข้อมูลเพื่อระบุข้อดีและข้อเสีย อีกทั้งประเมินประสิทธิภาพของ Algorithm ที่มีอยู่จะดำเนินการเพื่อกำหนด

แนวทางที่ดีที่สุด Vanishri Arun และคณะ (2017); Ramya Akula และคณะ (2019) Sonu Bala Garg และคณะ (2017) K. Saravananathan (2016) และ T. Velmurugan Ratna Patil (2016) และ Sharavari Tamane (2018) G.Visalatchi และคณะ (2014) Nilesh Jagdish Vispute และคณะ (2015)

Ashok Kumar และ R. Govindasamy (2015) S. Selvakumar และคณะ (2017) และ รุ่งโรจน์ บุญมา และ นิเวศ จิระวิชิตชัย (2562) ใช้ Algorithm การจำแนกและเปรียบเทียบในการทดสอบประสิทธิภาพของค่าความถูกต้องจากฐานข้อมูลผู้ป่วยโรคเบาหวาน UCI 768 คน และ UCI Machine Learning Repository จำนวน 89 เรคคอร์ด ตามลำดับ โดยใช้การเลือกคุณสมบัติพิเศษชุดย่อย (ข้อมูล) ของผู้ป่วยโรคเบาหวานจากชุดข้อมูลผู้ป่วยโรคเบาหวานทั้งหมดจะได้รับซึ่งประกอบด้วยเฉพาะคุณลักษณะที่สำคัญจากนั้นใช้ algorithm จำแนกประเทที่ได้รับชุดข้อมูลย่อยที่มีนัยสำคัญ

วิทยานิพนธ์เล่มนี้ผู้วิจัยมีวัตถุประสงค์ในการพัฒนาและวัดประสิทธิภาพแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวาน เพื่อนำไปใช้ในการจำแนกผู้ป่วยโรคเบาหวาน เทคนิคการทำเหมืองข้อมูล (Data Mining) จะถูกนำมาประยุกต์ใช้ในกระบวนการจัดการกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และ ความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น มาสร้างโมเดลสำหรับการพยากรณ์โอกาส (สาเหตุและปัจจัยที่ส่งผลต่อการเป็นโรคเบาหวาน) ข้อมูลขนาดใหญ่ที่ผู้วิจัยนำใช้ในการทำเหมืองข้อมูลคือจำนวนผู้ป่วยโรคเบาหวานจากระบบฐานข้อมูลของโรงพยาบาลศูนย์อุดรธานี จ. อุดรธานี ที่เข้ารับบริการการรักษา ในช่วงระยะเวลา 5 ปีย้อนหลัง ตั้งแต่วันที่ 1 ตุลาคม 2558 ถึงวันที่ 30 กันยายน 2563 มาใช้ในการวิเคราะห์ในครั้งนี้ ผลลัพธ์ที่ได้จากการวิเคราะห์ สามารถนำไปวางแผนและจัดการการรักษาผู้ป่วยโรคเบาหวานในโรงพยาบาลศูนย์อุดรธานีให้มีประสิทธิภาพมากขึ้น ทั้งนี้ยังสามารถนำมาใช้ประกอบการการรักษาแพทย์ เพื่อให้รองรับจำนวนผู้ป่วยที่จะเพิ่มขึ้นในอนาคตได้

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินงานวิทยานิพนธ์เล่มนี้เป็นการศึกษาและการประยุกต์ใช้โมเดลเทคนิคการทำเหมืองข้อมูลในการพยากรณ์สถานการณ์เกี่ยวกับผู้ป่วยโรคเบาหวาน ที่เข้ารับการตรวจและรักษาในโรงพยาบาลศูนย์อุดรธานี โดยนำเทคนิคต้นไม้ตัดสินใจ (Decision Trees) เทคนิคนาอิฟ เบย์ (Naïve Bayes) เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) เทคนิคโครงข่ายประสาทเทียม (Neural Network) ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) เทคนิคป่าสุ่ม (Random Forest) และเพิ่มประสิทธิภาพด้วยเทคนคโหวตร่วม (Voting Ensemble) มีวัตถุประสงค์เพื่อการพัฒนาแบบจำลองและสกัดข้อมูลและเพื่อนำไปใช้ประโยชน์ในการพยากรณ์เกี่ยวกับผู้ป่วยโรคเบาหวานโดยใช้ข้อมูลผลการตรวจวินิจฉัยทางการแพทย์

3.1 กรอบแนวคิดในการวิจัย

ในวิทยานิพนธ์นี้แสดงกระบวนการและกรอบแนวคิดในการวิจัยในการพยากรณ์ผู้ป่วยโรคเบาหวาน โรงพยาบาลศูนย์อุดรธานี

1. การเข้าใจปัญหา (Business Understanding) ทำการศึกษาปัญหาเกี่ยวกับการพยากรณ์ผู้ป่วยโรคเบาหวานและงานวิจัยที่เกี่ยวข้อง

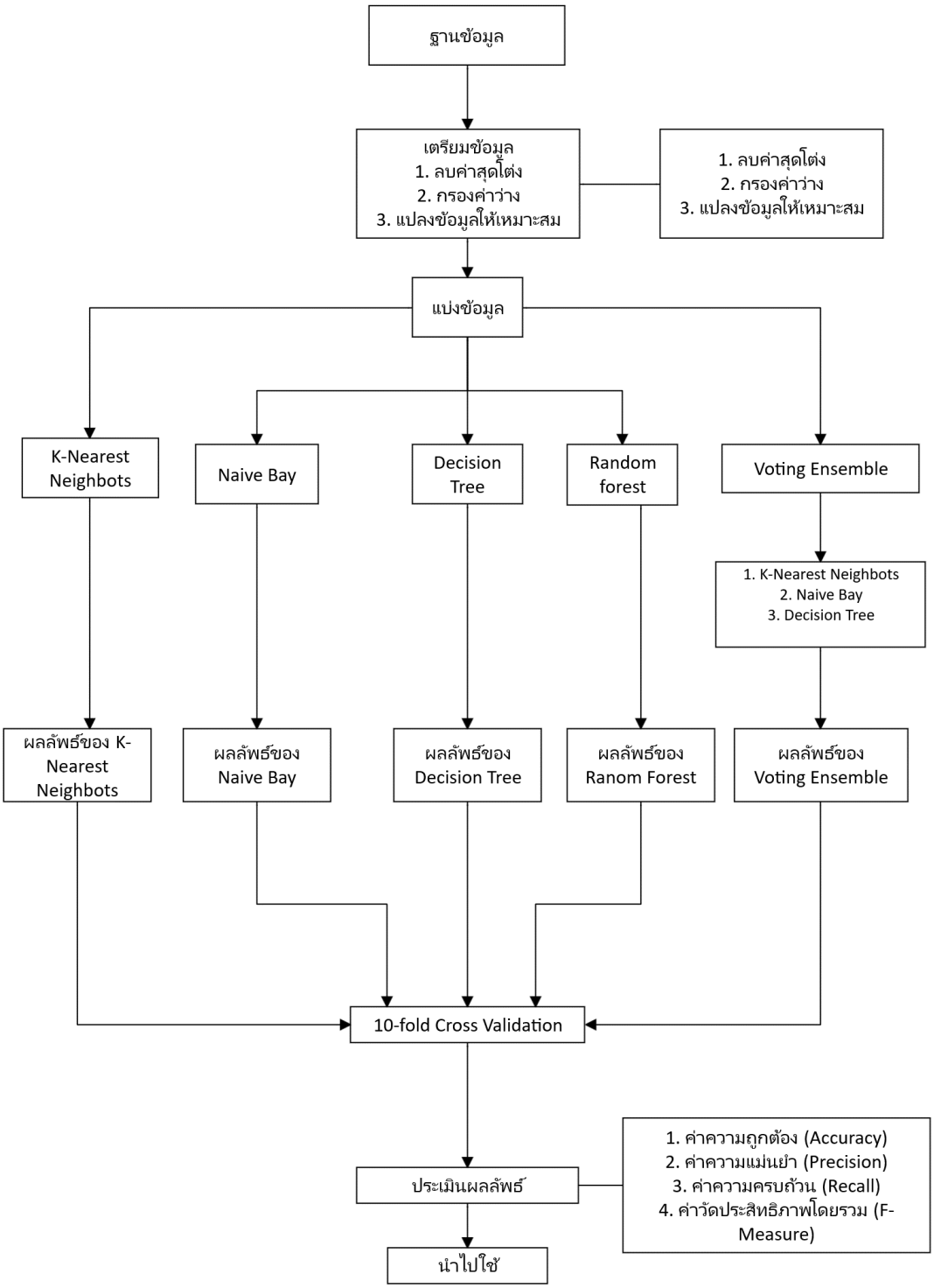
2. การทำความเข้าใจข้อมูล (Data Understanding) ทำการรวบรวมข้อมูลผู้ป่วยเบาหวานจากฐานข้อมูลโรงพยาบาลศูนย์อุดรธานี

3. การเตรียมข้อมูล (Data Preparation) ข้อมูลที่ได้นั้นยังไม่สามารถใช้งานได้ทันที่จึงต้องทำการแปลงค่าให้สมบูรณ์ และจัดทำคลังข้อมูล (Data Warehouse)

4. การทำแบบจำลอง (Modeling) นำข้อมูลที่เตรียมไว้เข้าสู่การบวนการทำเหมืองข้อมูล ได้แก่ เทคนิคต้นไม้การตัดสินใจ, เทคนิคนาอีฟ เบย์, เทคนิคเพื่อนบ้านใกล้ที่สุด, เทคนิคโหวตร่วม และ และเทคนิคป่าสุ่ม เพื่อสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวาน

5. การวัดผล (Evaluation) วัดผลแบบจำลองโดยใช้ 10-fold Cross Validation และเพิ่มประสิทธิภาพด้วย Voting Ensemble

6. การนำไปใช้ (Deployment) นำผลลัพธ์แบบจำลองที่ได้ไปใช้พยากรณ์ผู้ป่วยโรคเบาหวานและส่งแบบจำลองโรงพยาบาลศูนย์อุดรธานี



ภาพประกอบ 5 กรอบแนวคิดการวิจัย

3.2 เข้าใจปัญหา (Business Understanding)

ผู้วิจัยได้ศึกษางานวิจัยที่เกี่ยวข้องกับการรักษาของผู้ป่วยโรคเบาหวาน ปัญหาการตรวจรับการรักษาของโรงพยาบาลศูนย์อุดรธานี อำเภอเมือง จังหวัดอุดรธานี ค้นพบว่าทางโรงพยาบาลมีผู้ป่วยจำนวนมากอันเนื่องมาจากกโรงพยาบาลประจาอำเภอมีบุคลากรทางการแพทย์และเครื่องมือไม่เพียงพอในการตรวจโรคดังกล่าวจึงต้องส่งต่อผู้ป่วยมายังโรงพยาบาลศูนย์อุดรธานี และด้วยเหตุนี้เองทางโรงพยาบาลจึงจำเป็นต้องปรับปรุงในส่วนของอุปกรณ์ ยา บุคลากรและนโยบาย เพื่อเพิ่มประสิทธิภาพในการรักษาและการบริการที่มากขึ้น และนำเสนอขึ้นสู่ผู้อำนวยการต่อไป

3.3 การทำความเข้าใจข้อมูล (Data Understanding)

ผู้วิจัยได้รวบรวมข้อมูลที่เกี่ยวข้องสำหรับทำการวิเคราะห์ด้วยเทคนิคการเหมืองข้อมูล (Data Mining) โดยเริ่มจากการเก็บรวบรวมข้อมูลประจำตัวผู้ป่วย ข้อมูลที่ได้จากการตรวจรักษาของผู้ป่วยโรคหลอดเลือดหัวใจของผู้ป่วยใน IPD: (Inpatient Department) โดยเจ้าหน้าที่ดูแลระบบฐานข้อมูลกลุ่มภารกิจเทคโนโลยีสารสนเทศคอมพิวเตอร์ โรงพยาบาลศูนย์อุดรธานี ทำการส่งข้อมูลมาให้ซึ่งอยู่ในรูปแบบไฟล์ .xlsx หรือ Microsoft Excel Worksheet เพื่อนำข้อมูลที่ได้ ไปใช้งานต่อไป ซึ่งมีรายละเอียดข้อมูลดังต่อไปนี้

3.3.1 ข้อมูลประจำตัวผู้ป่วยเบาหวาน ที่เข้ามารับการรักษาในโรงพยาบาลศูนย์อุดรธานี อำเภอเมือง จังหวัดอุดรธานี ประกอบไปด้วย ตามรายละเอียดแสดงในตารางที่ 3.1

ตาราง 1 ตารางข้อมูลผู้ป่วย

ลำดับ	ชื่อตัวแปร	ความหมาย
1	Hn	เลขประจำตัวผู้ป่วย
2	Reg. no.	รหัสหน่วยตรวจ
3	regis_date	วันที่เข้ารับการตรวจ
4	Age	อายุ
5	Sex	เพศ
6	Dept. Code	คลินิกตรวจโรค
7	Vital Sign	สัญญาณชีพ
8	Weight	น้ำหนัก

ลำดับ	ชื่อตัวแปร	ความหมาย
9	Height	ความสูง
10	Lbloodpress	ค่าล่างความดันเลือด
11	Hbloodpress	ค่าบนความดันเลือด
12	temperature	อุณหภูมิ
13	Pulse	ชีพจร
14	Breathe	การหายใจ
15	Treatment	การรักษา
16	Smoke	การสูบบุหรี่
17	Alchohol	การดื่มแอลกอฮอล์
18	WHR	สัดส่วนรอบเอว
19	BMI	ดัชนีมวลร่างกาย
20	Cholesterol	ค่าโคเรสเตอรอล
21	Glucose	ค่ากลูโคส

3.1.2 ข้อมูลผู้เข้ารับการตรวจสุขภาพตามสิทธิประกันสังคมที่โรงพยาบาลศูนย์อุดรธานี อำเภอเมือง จังหวัดอุดรธานี ประกอบไปด้วย ตามรายละเอียดแสดงในตารางที่ 3.2

ตาราง 2 ตารางข้อมูลผู้เข้ารับการตรวจสุขภาพตามสิทธิประกันสังคม

ลำดับ	ชื่อตัวแปร	ความหมาย
1	hn	เลขประจำตัวผู้ป่วย
2	regNo	รหัสหน่วยตรวจ
3	VisitDate	วันที่เข้ารับการตรวจ
4	TAmount	จำนวนครั้งที่รับการรักษา
5	Age	อายุ
6	sex	เพศ
7	Weight	น้ำหนัก
8	Height	ส่วนสูง
9	Lbloodpress	ค่าล่างความดันเลือด

ลำดับ	ชื่อตัวแปร	ความหมาย
10	Hbloodpress	ค่าบนความดันเลือด
11	Pulse	ชีพจร
12	Smoke	การสูบบุหรี่
13	Alcohol	การดื่มแอลกอฮอล์
14	BMI	ดัชนีมวลร่างกาย
15	waistline	สัดส่วนรอบเอว
16	Cholesterol	ค่าโคเรสเตอรอล
17	Glucose	ค่ากลูโคส

3.4 การเตรียมข้อมูล (Data Preparation)

เมื่อทำการคัดเลือกข้อมูลเสร็จแล้วมาถึงทำข้อมูลให้อยู่ในรูปแบบคลังของข้อมูล (Data Warehouse) เพื่อเป็นการจัดเก็บข้อมูลให้ถูกต้อง เป็นระเบียบเรียบร้อยสะดวกต่อการนำไปใช้งาน หรือนำไปประมวลผลเพื่อพยากรณ์ผู้ป่วยโรคเบาหวาน ด้วยเทคนิคการทำเหมืองข้อมูล กรณีศึกษาโรงพยาบาลศูนย์อุดรธานี อำเภอเมือง จังหวัดอุดรธานี ประกอบไปด้วย

3.4.1 คัดเลือกข้อมูล โดยการคัดเลือกข้อมูลให้มีความถูกต้อง เพื่อให้ตัวแบบมีความแม่นยำ โดยใช้ข้อมูลจากฐานข้อมูลของโรงพยาบาลศูนย์อุดรธานี อำเภอเมือง จังหวัดอุดรธานี เพื่อใช้ในการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจซึ่งประกอบไปด้วยแอททริบิวต์ (Attributes) ดังตารางรายละเอียดในตารางที่ 3.2

ตาราง 3 รายละเอียดแอททริบิวต์ (Attributes) ของข้อมูลที่ใช้ในงานวิจัย

ลำดับ	แอททริบิวต์	คำอธิบาย
1	Hn	เลขประจำตัวผู้ป่วย
2	age	อายุ
3	sex	เพศ
4	Weight	น้ำหนัก
5	Height	ส่วนสูง
6	Lbloodpress	ค่าล่างความดันเลือด

ลำดับ	แอททริบิวต์	คำอธิบาย
7	Hbloodpress	ค่าบนความดันเลือด
8	Pulse	ชีพจร
9	Smoke	การสูบบุหรี่
10	Alcohol	การดื่มแอลกอฮอล์
11	BMI	ดัชนีมวลร่างกาย
12	Cholesterol	ระดับคลอเรสเตอรอล
13	Glucose	ระดับกลูโคส
14	Disease	โรค

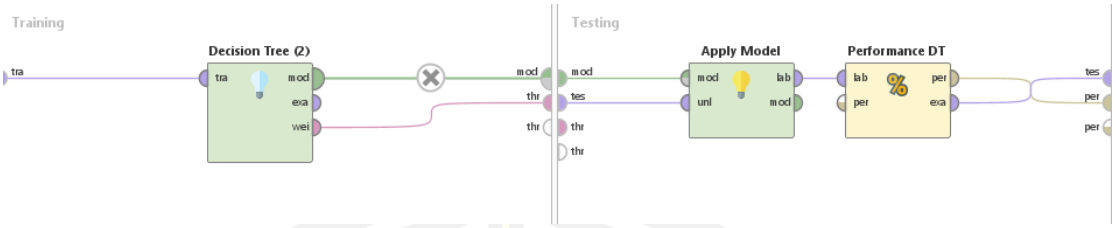
3.4.2 ทำความสะอาดข้อมูลหลังจากสำรวจข้อมูลแล้วพบว่าข้อมูลยังไม่ครบสมบูรณ์ เช่น ค่าว่าง (Missing Value) และมีสิ่งรบกวน (Noisy Data) แก้ไขโดยการแทนค่าข้อมูลที่ผิดปกติดังกล่าว

3.4.3 แปลงข้อมูล (Transform Data) เนื่องจากข้อมูลมีทั้ง ตัวเลข ตัวอักษร ซึ่งอาจเป็นคำที่แทนความหมายใดความหนึ่งที่อาจไม่ทราบความหมาย จำเป็นต้องหาว่าข้อความที่เป็น คำเฉพาะหรือความหมายของค่านั้นๆ และข้อมูลบางส่วนก็อาจจะไม่สามารถนำมาวิเคราะห์ได้

3.5 การสร้างแบบจำลอง (Modeling)

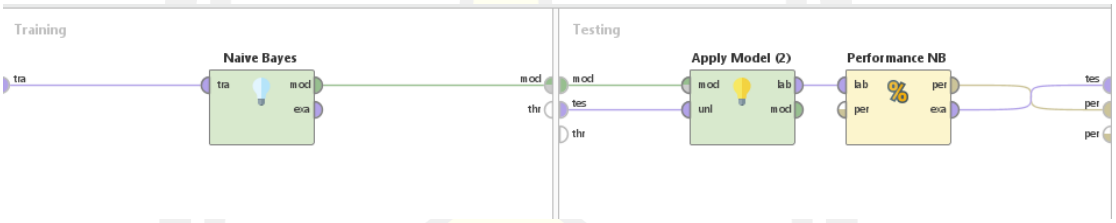
การจัดทำแบบจำลองพยากรณ์ภาวะแทรกซ้อนของโรคอื่นหลังจากที่ผู้ป่วยเป็นโรคหลอดเลือดหัวใจ โรคเบาหวาน และ โรคความดันโลหิตสูง ใช้โปรแกรม RapidMiner Studio เพื่อใช้สร้างตัวแบบในการวิเคราะห์ข้อมูลเพื่อพยากรณ์ ผู้ป่วยเป็นโรคหลอดเลือดหัวใจ โรคเบาหวาน และ โรคความดันโลหิตสูง โดยใช้เทคนิค ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟ เบย์ (Naïve Bay) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN: K-Nearest Neighbor) แล้วนำผลพยากรณ์มาเปรียบเทียบหาค่าความถูกต้องแม่นยำโดยใช้ผลที่ มีความเชื่อถือมากที่สุด

3.5.1 เทคนิคต้นไม้การตัดสินใจ (Decision Tree) เป็นเทคนิคที่นำข้อมูลแต่ละโนด (Node) ของแอททริบิวท์ (Attribute) มาทำการตัดสินใจจากนั้นจะแสดงข้อมูลออกมาเป็นกิ่ง (Branch) และแสดงค่าออกมาเป็นใบ (Leaf) โดยใช้ information Gain มาหาความสัมพันธ์ในแต่ละโนดและทำให้ต้นไม้การตัดสินใจมีความซับซ้อนไม่มาก



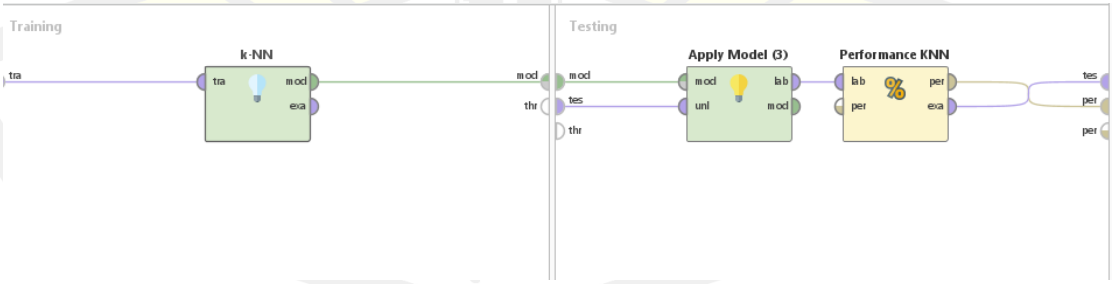
ภาพประกอบ 6 แบบจำลองเพื่อพยากรณ์ผู้ป่วยด้วยเทคนิคต้นไม้การตัดสินใจ (Decision Tree)

3.5.2 เทคนิคนาอีฟ เบย์ (Naïve Bayes) เป็นเทคนิคที่จำแนกประเภทที่อาศัยหลักการความน่าจะเป็นโดยอาศัยหลัก Naïve Bayesian Classification ใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อนแล้วที่เป็นอิสระต่อกัน จากนั้นจึงเพิ่มรอบการคำนวนให้มากขึ้น



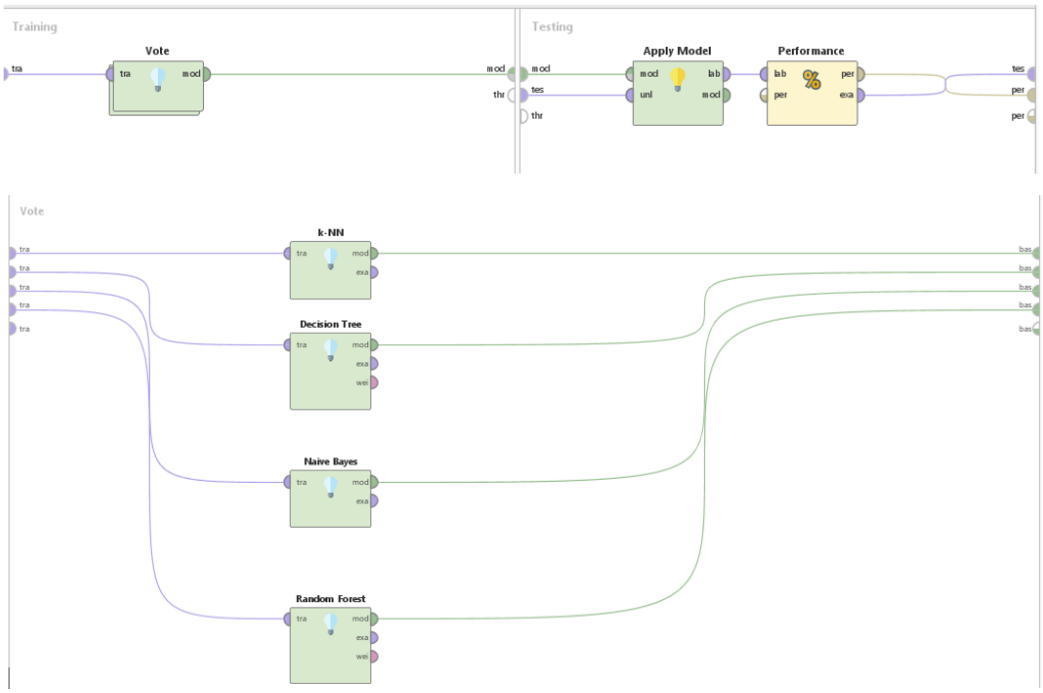
ภาพประกอบ 7 แบบจำลองเพื่อพยากรณ์ผู้ป่วยด้วยเทคนิค นาอีฟ เบย์ (Naïve Bayes)

3.5.3 เทคนิคเพื่อนบ้านใกล้ที่สุด (k-NN: k-Nearest Neighbor) เป็นวิธีการในการจัดแบ่งคลาส เพื่อแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวน K ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไขหรือกรณี ต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด



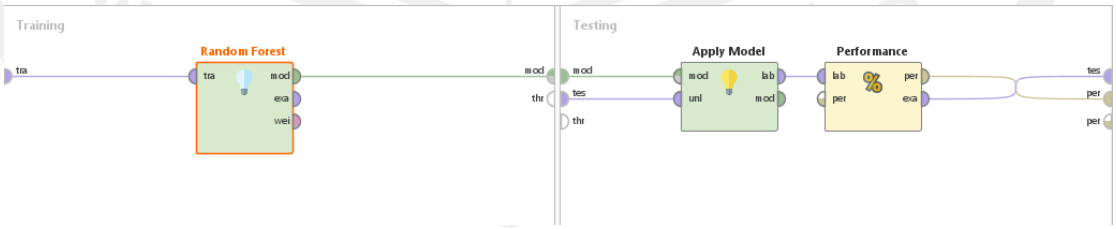
ภาพประกอบ 8 แบบจำลองเพื่อพยากรณ์ผู้ป่วยด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด (k-NN: k-Nearest Neighbor)

3.5.4 เทคนิคโหวตร่วม (Vote Ensemble) เป็นโมเดลที่ฝึกกับโมเดลต่างๆและคาดการณ์เอาต์พุตตามความน่าจะเป็นสูงสุดของคลาสที่เลือกเป็นเอาต์พุต โดยรวบรวมผลการค้นพบของแต่ละเทคนิค (เทคนิคต้นไม้การตัดสินใจ เทคนิคนาอีฟ เบย์เทคนิคเพื่อนบ้านใกล้ที่สุด และเทคนิคป่าสุ่ม) ที่ส่งผ่านไปยังตัวจำแนกการลงคะแนนเสียง (Voting Classifier) และคาดการณ์ระดับผลลัพธ์ตามเสียงส่วนใหญ่โหวตสูงสุด



ภาพประกอบ 9 แบบจำลองเพื่อพยากรณ์ผู้ป่วยด้วยเทคนิคโหวตร่วม (Vote Ensemble)

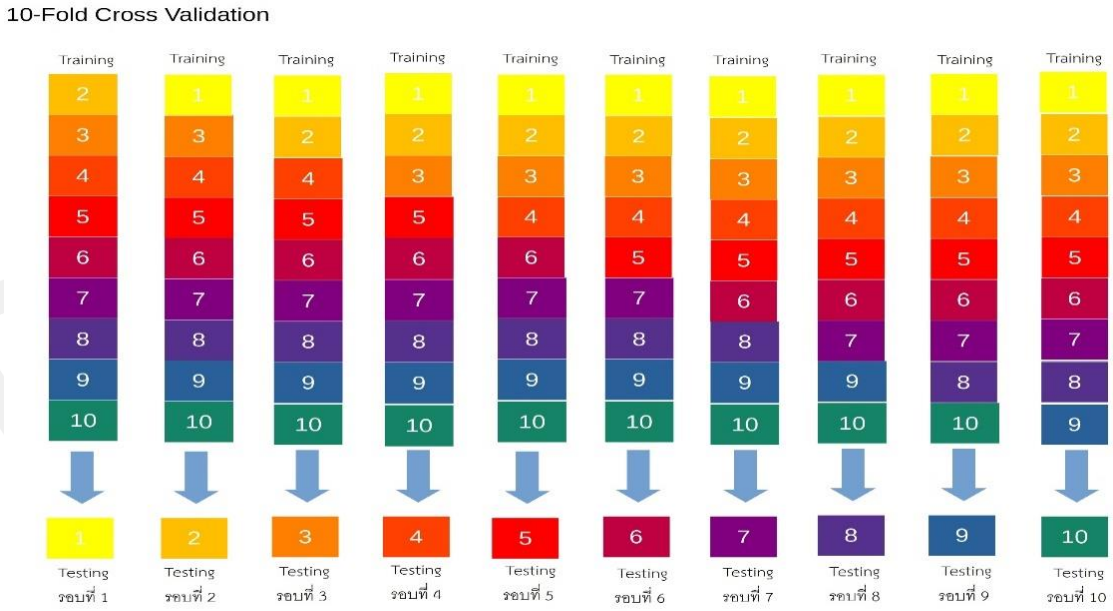
3.5.5 เทคนิคป่าสุ่ม (Random Forest) เป็นเทคนิคที่มีลักษณะคล้ายต้นไม้การตัดสินใจ โดยเทคนิคนี้จะสร้างแบบต้นไม้การตัดสินใจหลายแบบเพื่อทำนายผลลัพธ์เพื่อสร้างรูปแบบการตัดสินใจที่ไม่เหมือนกันด้วยการลงคะแนนเสียงผลลัพธ์ที่ถูกเลือกมากที่สุด



ภาพประกอบ 10 แบบจำลองเพื่อพยากรณ์ผู้ป่วยด้วยเทคนิคป่าสุ่ม (Random Forest)

3.6 การประเมินผล (Evaluation Phase)

เป็นการประเมินผลลัพธ์จากแบบจำลองและอัลกอริทึมทั้ง 2 ที่ใช้วิเคราะห์ข้อมูลว่า ผลลัพธ์สามารถครอบคลุมและสามารถตอบวัตถุประสงค์ที่กำหนดไว้หรือไม่ จากนั้นจึงนำแบบ จำลองจากการทำเหมืองข้อมูล มาประเมินพิจารณาถึงความเหมาะสมต่อการนำแบบจำลองและการทำนายผลพยากรณ์มีความแม่นยำมากน้อย เพียงใด โดยในการประเมินแบบจำลองนั้น จะแบ่งตามลักษณะของการทำเหมืองข้อมูล กฎ ความสัมพันธ์สามารถประเมินผลที่ได้จากการพิจารณาค่าความเชื่อมั่น และค่าสนับสนุน ประเมินผลโดยการเปรียบเทียบผลพยากรณ์ที่ได้กับข้อมูลจริงที่เกิดขึ้นซึ่งในการวัดประสิทธิภาพในการทำเหมืองข้อมูลนั้นจะประกอบไปด้วย ค่าความถูกต้อง (Accuracy) คือจำนวนข้อมูลที่ทำนายถูกของคลาส ค่าความแม่นยำ (Precision) คือค่าตัวแบบทำนายที่ถูกต้อง ค่าความครบถ้วน (Recall) คือค่าของตัวแบบที่ตรงกับค่าความเป็นจริง และค่าวัดประสิทธิภาพโดยรวม (F-Measure) คือค่าที่เกิดจากการเปรียบเทียบระหว่างค่าความแม่นยำและค่าความครบถ้วน โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงาน นั้นคือ 10-Fold Cross Validation ใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัด ประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล



ภาพประกอบ 11 การประเมินผลตามแบบ 10-Fold Cross Validation

3.6 การนำแบบจำลองไปใช้งาน (Deployment)

นำแบบจำลองที่ใช้ในการพยากรณ์ผู้ป่วยโรคเบาหวานในโรงพยาบาลศูนย์อุดรธานีไปใช้ประกอบการตัดสินใจในการรักษาเพื่อให้สอดคล้องต่อจำนวนผู้ป่วยในอนาคตที่มีการเป็นโรคนี้มากขึ้น รวมถึงลดระดับความเสี่ยงต่อการเสียชีวิต ป้องกันเฝ้าระวังแก้ไขเพื่อลดอัตราการเกิดโรคในอนาคต

บทที่ 4
ผลการวิจัยและการอภิปราย

วัตถุประสงค์ของงานวิจัยคือการศึกษาการสร้างตัวแบบสำหรับการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจ โรคความดันโลหิตสูง โรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษาโรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี เทคนิคที่ใช้ในการทำเหมืองข้อมูล ได้แก่ ต้นไม้การตัดสินใจ (Decision Tree) ทฤษฎีของเบย์ (Naïve Bay) และเทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbor (k-NN)) เพิ่มประสิทธิภาพด้วยวิธีการโหวตร่วม (Voting Ensemble) และเทคนิคป่าสุ่ม (Random Forest) ผู้วิจัยได้ประมวลผลข้อมูลตามขั้นตอนของมาตรฐานการทำเหมืองข้อมูล (CRISP-DM) ซึ่งมีรายละเอียดและผลการวิเคราะห์แบ่งออกเป็นดังนี้

4.1 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจธุรกิจ (Business Understanding)

4.2 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจข้อมูล (Data Understanding)

4.3 ผลการวิเคราะห์ในขั้นตอนการเตรียมข้อมูล (Data Preparation)

4.4 ผลการวิเคราะห์ในขั้นตอนการสร้างแบบจำลอง (Modeling)

4.5 ผลการวิเคราะห์ในขั้นตอนการประเมินผล (Evaluation)

4.6 ผลการวิเคราะห์ในขั้นตอนการนำไปใช้งาน (Deployment)

4.1 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจธุรกิจ (Business Understanding)

โรงพยาบาลอุดรธานี ตั้งอยู่ในเขตเทศบาลนครอุดรธานี อยู่ห่างจากศาลากลางจังหวัดประมาณ 1.2 กิโลเมตร ตลอดระยะเวลาที่ผ่านมาโรงพยาบาลอุดรธานี ได้ดำเนินการปรับปรุงพัฒนาคุณภาพและขยายการให้บริการอย่างสม่ำเสมอมาตามลำดับจนถึงปัจจุบัน ภายใต้การบริหารงานโดยมีวิสัยทัศน์และพันธกิจด้วยความมุ่งมั่นจนได้ผลงานบริการด้านสุขภาพที่มีการพัฒนาอย่างต่อเนื่องจากผลการปฏิบัติงานด้วยความเสียสละและมีความตั้งใจทำงานในหน้าที่ราชการเป็นอย่างดีของผู้อำนวยการ แพทย์ ทันตแพทย์ เภสัชกร พยาบาล และเจ้าหน้าที่อื่น ๆ ทั้งในอดีตและปัจจุบันจนเป็นที่ยอมรับของประชาชนในจังหวัดอุดรธานี และจังหวัดใกล้เคียง ซึ่งได้สนับสนุนทางด้านกำหลังทรัพย์ วัสดุอุปกรณ์ ทางการแพทย์ ตลอดจนผู้บริหารชั้นผู้ใหญ่ในกระทรวงสาธารณสุขที่ให้การสนับสนุนด้านเงินงบประมาณ ทำให้โรงพยาบาลอุดรธานี ได้มีการพัฒนาจนเจริญก้าวหน้าดั้งที่เป็นอยู่ในปัจจุบันและจะเจริญก้าวหน้า ทั้งในด้านการบริหาร และการบริการยิ่งขึ้นต่อไปในอนาคต โดยปัญหาที่เกิดขึ้นในโรงพยาบาลศูนย์อุดรธานีเกิดขึ้นจากจำนวนผู้ป่วยที่มากขึ้น โดยแต่ละวันจะมีผู้ป่วยเข้ามา

ใช้บริการมากในแผนกกว่า 1,500 คนต่อวัน อีกทั้งยังบุคลากรที่ไม่เพียงพอและยังมีการจัดบุคลากรที่ไม่เหมาะสมกับตำแหน่งจึงทำให้ประสิทธิภาพการบริการลดลง เวชภัณฑ์และอุปกรณ์สำหรับการรักษาบางส่วนได้รับจากบริจากเป็นเวลานานซึ่งทำให้อุปกรณ์เริ่มมีสภาพชำรุดขึ้นทุก ๆ วัน และเครื่องมือที่ทางรัฐสนับสนุนก็ไม่เพียงพอต่อการรักษาเช่นกัน ดังนั้น โรงพยาบาลศูนย์อุดรธานีจึงต้องค้นหากลยุทธ์เพื่อเพิ่มประสิทธิภาพการบริหารทางการแพทย์เพื่อรองรับการเพิ่มขึ้นของผู้ป่วยในอนาคต

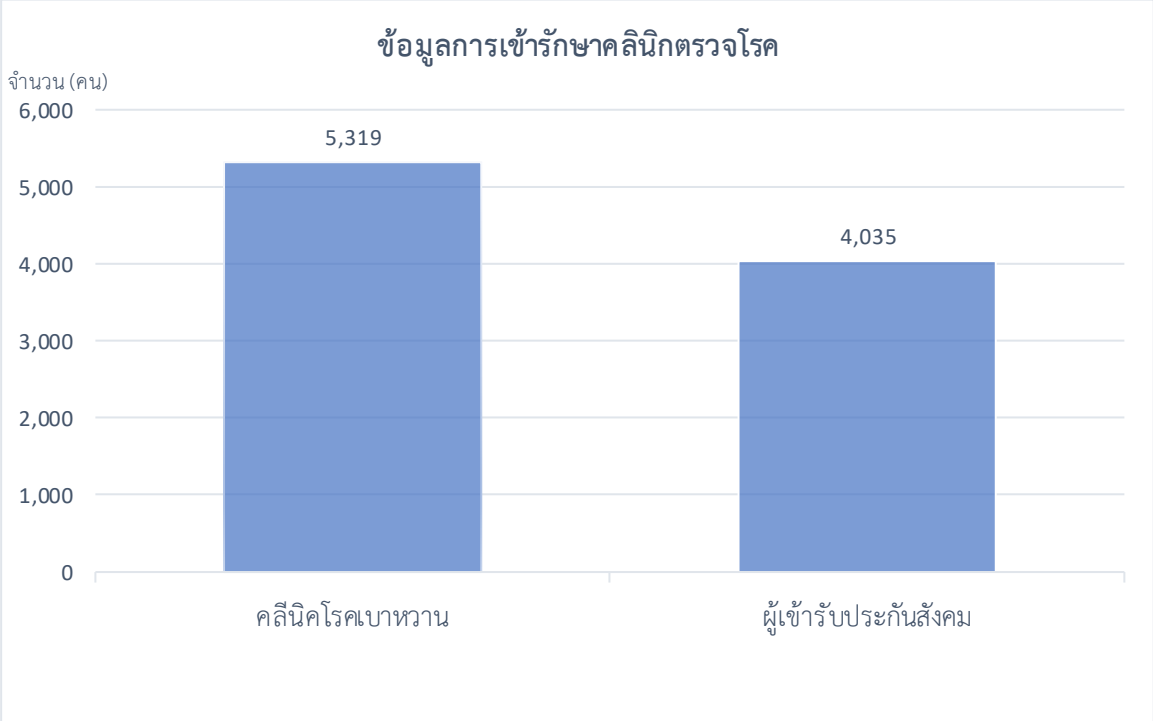
4.2 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในการวิจัยได้มาจากฐานข้อมูล (Database) จากฝ่ายคอมพิวเตอร์ของโรงพยาบาลศูนย์อุดรธานี ข้อมูลเหล่านี้เป็นการแสกน (scan) เก็บหลังจากผู้ป่วยโรคเบาหวานที่เข้ารับการตรวจและรักษาโรคของคลินิกประเภทผู้ป่วยนอก หรือ OPD (Out Patient Department) จำนวน 70,420 แถว และข้อมูลผู้เข้ารับการตรวจสุขภาพตามสิทธิประกันสังคมจำนวน 82,007 ข้อมูลเหล่านี้ประกอบไปด้วย รหัสผู้ป่วย จำนวนครั้งในการตรวจ วันเข้ารับการตรวจ อายุ เพศ สัญญาณชีพ น้ำหนัก ส่วนสูง ความดันเลือดช่วงล่าง ความดันเลือดช่วงบน อุณหภูมิ การเต้นของชีพจร การหายใจ การรักษา การสูบบุหรี่ การดื่มแอลกอฮอล์ สัดส่วนรอบเอว ค่าดัชนีมวลกาย (BMI) ระดับคอลเลสเตอรอล และระดับกลูโคลส โดยข้อมูลชุดนี้ทำการเก็บข้อมูลตั้งแต่ปี พ.ศ. 2558-2564 ซึ่งถูกจัดเก็บไว้ในโปรแกรม Microsoft Excel และได้นำเสนออยู่ในรูปแบบตารางและแผนภูมิดังต่อไปนี้

ตาราง 4 ข้อมูลผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคม

ห้องตรวจ	จำนวน	ร้อยละ
คลินิกโรคเบาหวาน	14,153	55.1
ผู้เข้ารับประกันสังคม	11,507	44.9
รวม	25,660	100.0

ตาราง 4 แสดงจำนวนผู้ป่วยที่เข้ารับการรักษาในคลินิกห้องตรวจโรคประเภทผู้ป่วยนอก ของโรงพยาบาลศูนย์อุดรธานีตั้งแต่ปี พ.ศ. 2558 – 2564 และข้อมูลผู้เข้ารับประกันสังคม แบ่งเป็นผู้ที่มารับการรักษากับคลินิกโรคเบาหวาน 14,153 คน คิดเป็นร้อยละ 55.1 และข้อมูลผู้เข้ารับประกันสังคม 11,507 คน คิดเป็นร้อยละ 44.9 ดังแสดงในภาพประกอบ 12



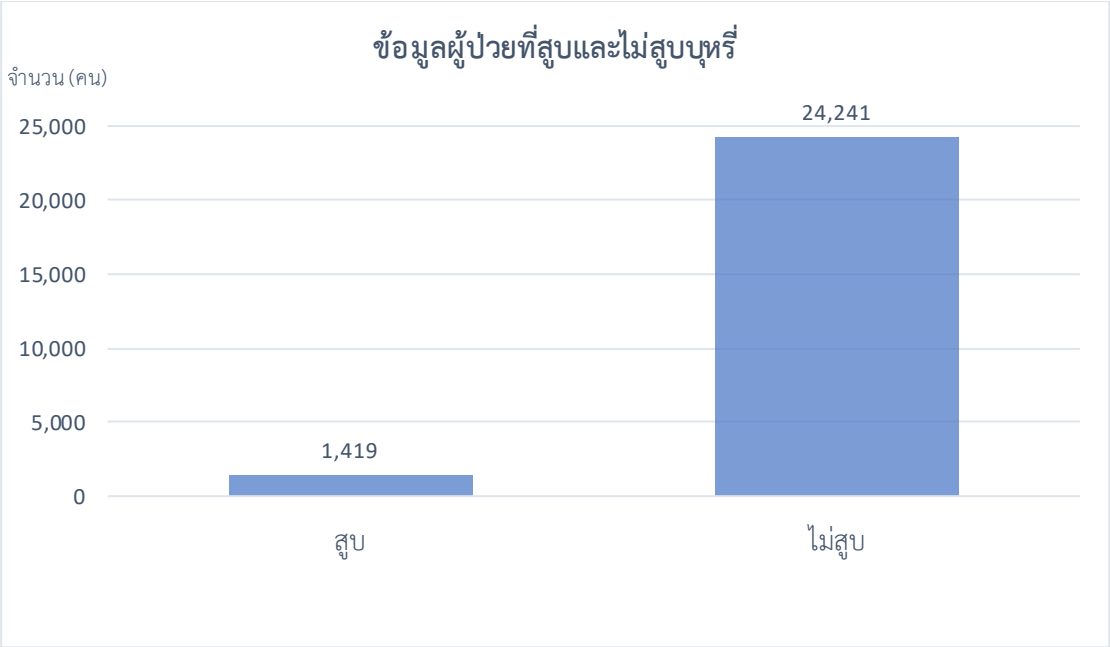
ภาพประกอบ 12 แผนภูมิแสดงจำนวนผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคม

จากฐานข้อมูลของโรงพยาบาลศูนย์อุดรธานีในช่วงปี พ.ศ. 2558-2564 ของผู้ป่วยโรคเบาหวานและผู้เข้ารับประกันสังคมพบว่ามีปัจจัยร่วมอื่น ๆ ที่อาจส่งผลกระทบต่อการเป็นโรคดังแสดงรายละเอียดในตารางต่อไปนี้

ตาราง 5 ข้อมูลผู้ป่วยที่สูบบุหรี่

การสูบบุหรี่	จำนวน	ร้อยละ
สูบบุหรี่	1,419	5.5
ไม่สูบบุหรี่	24,241	94.5
รวม	25,660	100.0

ตาราง 5 แสดงถึงจำนวนผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คน แบ่งเป็นผู้ป่วยที่สูบบุหรี่มีจำนวน 1,419 คน คิดเป็นร้อยละ 5.5 และผู้ป่วยที่ไม่สูบมีจำนวน 24,241 คน 3.4 คิดเป็นร้อยละ 94.5 ดังแสดงในภาพประกอบ 13

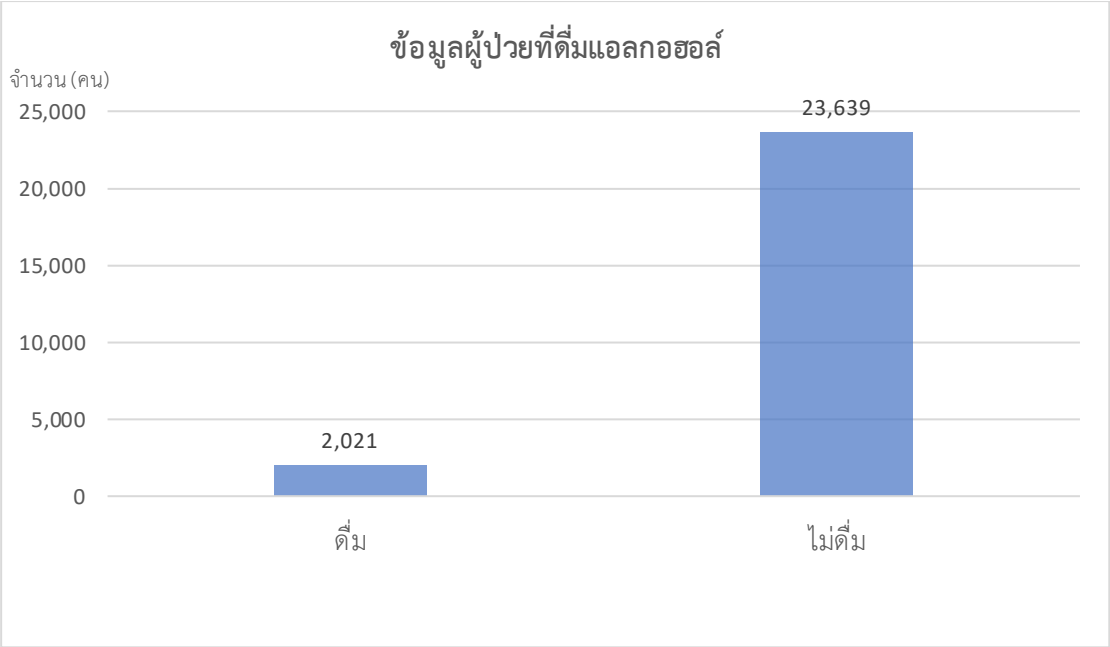


ภาพประกอบ 13แผนภูมิข้อมูลผู้ป่วยที่สูบและไม่สูบบุหรี่

ตาราง 6ข้อมูลผู้ป่วยที่ดื่มและไม่ดื่มแอลกอฮอล์

การดื่มแอลกอฮอล์	จำนวน	ร้อยละ
ดื่ม	2,021	7.9
ไม่ดื่ม	23,639	92.1
รวม	25,660	100.0

ตาราง 6 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คน แบ่งเป็นผู้ป่วยที่ดื่มแอลกอฮอล์มีจำนวน 2,021 คน คิดเป็นร้อยละ 7.9 และผู้ป่วยที่ไม่ดื่มแอลกอฮอล์มีจำนวน 23,639 คน คิดเป็นร้อยละ 92.1 ดังแสดงในภาพประกอบ 14

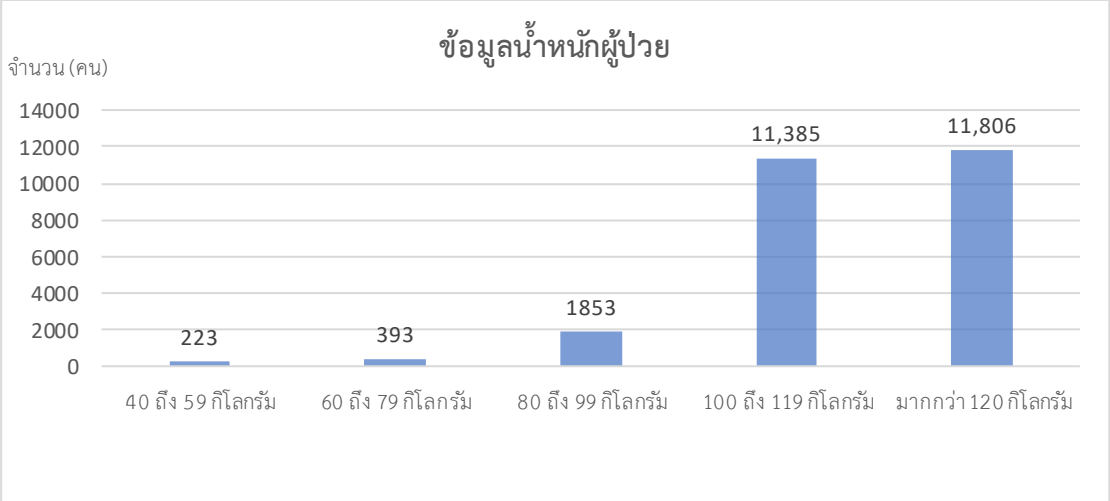


ภาพประกอบ 14 แผนภูมิข้อมูลของผู้ป่วยที่ดื่มและไม่ดื่มแอลกอฮอล์

ตาราง 7 ข้อมูลน้ำหนักผู้ป่วย

น้ำหนัก	จำนวน	ร้อยละ
40 ถึง 59 กิโลกรัม	223	0.87
60 ถึง 79 กิโลกรัม	393	1.53
80 ถึง 99 กิโลกรัม	1,853	7.22
100 ถึง 119 กิโลกรัม	11,385	44.37
มากกว่า 120 กิโลกรัม	11,806	46.01
รวม	25,660	100.0

ตาราง 7 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คน แบ่งเป็นผู้ป่วยที่มีน้ำหนักระหว่าง 40 กิโลกรัมถึง 59 กิโลกรัมมีจำนวน 223 คน คิดเป็นร้อยละ 0.87 ผู้ป่วยที่มีน้ำหนัก 60 กิโลกรัม ถึง 79 กิโลกรัมมีจำนวน 393 คน คิดเป็นร้อยละ 1.53ผู้ป่วยที่มีน้ำหนัก 80 กิโลกรัมถึง 99 กิโลกรัมมีจำนวน 1,853 คน คิดเป็นร้อยละ 7.22 ผู้ป่วยที่มีน้ำหนัก 100 กิโลกรัม ถึง 119 กิโลกรัมมีจำนวน 11,385 คน คิดเป็นร้อยละ 44.37 และผู้ป่วยที่มีน้ำหนักมากกว่า 120 กิโลกรัม มีจำนวน 11,806 คน คิดเป็นร้อยละ 46.01 ดังแสดงในภาพประกอบ 15

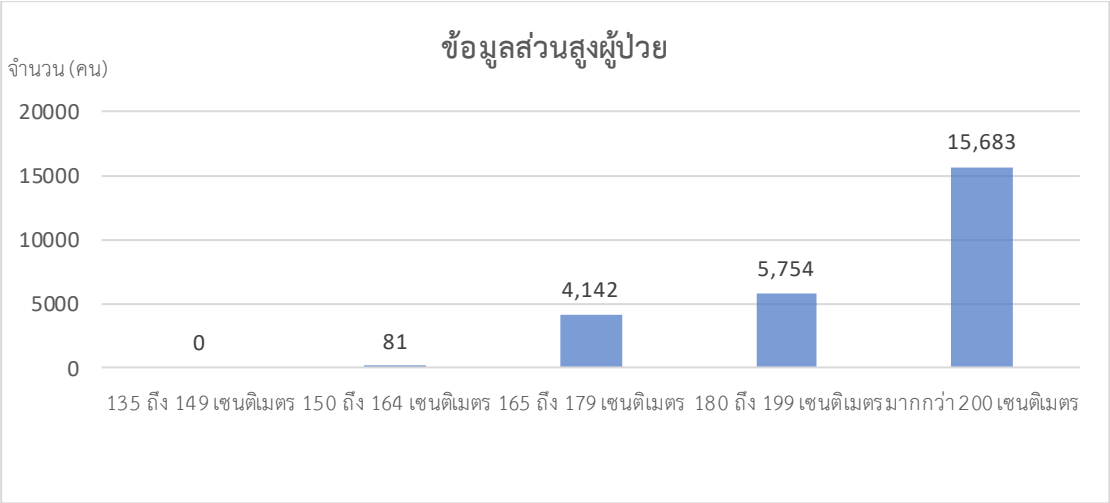


ภาพประกอบ 15 แผนภูมิข้อมูลน้ำหนักผู้ป่วย

ตาราง 8 ข้อมูลส่วนสูงผู้ป่วย

ส่วนสูง	จำนวน	ร้อยละ
135 ถึง 149 เซนติเมตร	0	0.0
150 ถึง 164 เซนติเมตร	81	0.3
165 ถึง 179 เซนติเมตร	4,142	16.1
180 ถึง 199 เซนติเมตร	5,754	22.4
มากกว่า 200 เซนติเมตร	15,683	61.1
รวม	25,660	100.0

ตาราง 8 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คน แบ่งเป็นผู้ป่วยที่ที่มีส่วนสูง 135 เซนติเมตร ถึง 149 เซนติเมตรมีจำนวน 0 คน ผู้ป่วยที่มีส่วนสูง 150 เซนติเมตร ถึง 164 เซนติเมตรมีจำนวน 81 คนคิดเป็นร้อยละ 0.3 ผู้ป่วยที่มีส่วนสูง 165 เซนติเมตร ถึง 179 เซนติเมตรมีจำนวน 4,142 คนคิดเป็นร้อยละ 16.1 ผู้ป่วยที่มีส่วนสูง 180 เซนติเมตร ถึง 199 เซนติเมตรมีจำนวน 5,754 คนคิดเป็นร้อยละ 22.4 และผู้ป่วยที่มีส่วนสูงมากกว่า 200 เซนติเมตรมีจำนวน 15,683 คนคิดเป็นร้อยละ 66.1 ดังแสดงในภาพประกอบ 16

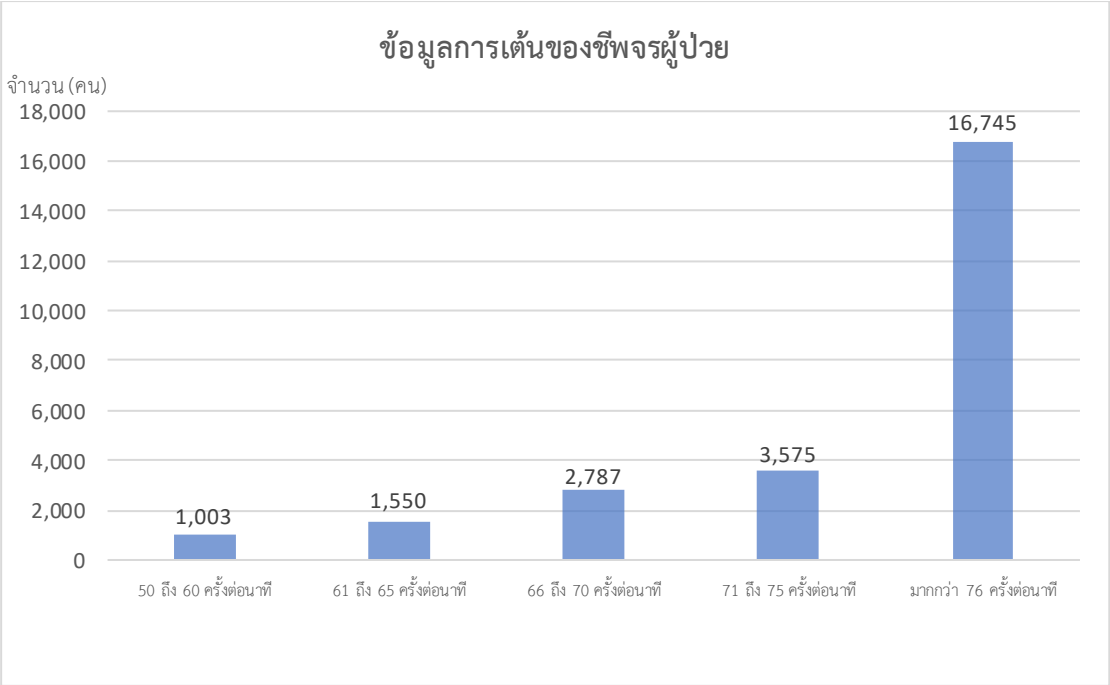


ภาพประกอบ 16 แผนภูมิข้อมูลส่วนสูงผู้ป่วย

ตาราง 9 ข้อมูลการเต้นของชีพจรผู้ป่วย

การเต้นของชีพจร	**จำนวน**	**ร้อยละ**
50 ถึง 60 ครั้งต่อนาที	1,003	3.9
61 ถึง 65 ครั้งต่อนาที	1,550	6.0
56 ถึง 70 ครั้งต่อนาที	2,787	10.9
71 ถึง 75 ครั้งต่อนาที	3,575	13.9
มากกว่า 76 ครั้งต่อนาที	16,745	65.3
รวม	25,660	100

ตาราง 9 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คนแบ่งเป็นผู้ป่วยที่มีการเต้นของชีพจร 50 ถึง 60 ครั้งต่อนาทีมีจำนวน คน ผู้ป่วยที่มีการเต้นของชีพจร 61 ถึง 65 ครั้งต่อนาทีมีจำนวน คน ผู้ป่วยที่มีการเต้นของชีพจร 66 ถึง 70 ครั้งต่อนาทีมีจำนวน คน ผู้ป่วยที่มีการเต้นของชีพจร 50 ถึง 60 ครั้งต่อนาทีมีจำนวน คน และผู้ป่วยที่มีการเต้นของชีพจรมากกว่า 76 ครั้งต่อนาทีมีจำนวน คน คิดเป็นร้อยละ และ ตามลำดับ ดังแสดงในภาพประกอบ 17

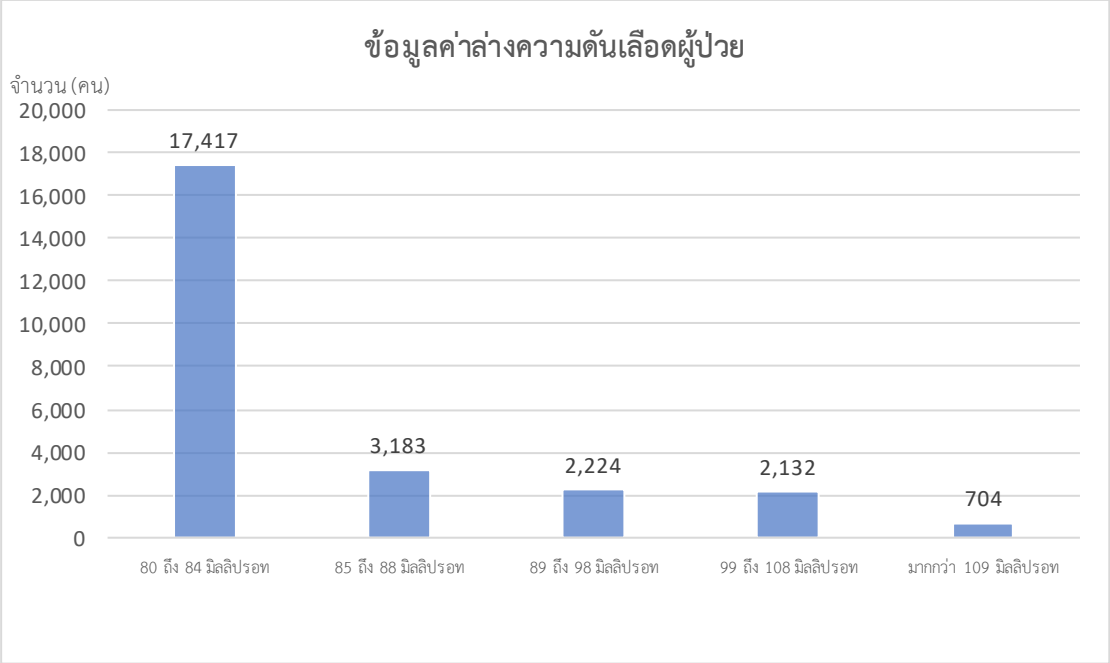


ภาพประกอบ 17 แผนภูมิข้อมูลการเต้นของชีพจรผู้ป่วย

ตาราง 10 ข้อมูลค่าล่างความดันเลือดผู้ป่วย

ความดันเลือด	จำนวน	ร้อยละ
80 ถึง 84 มิลลิปรอท	17,417	67.9
85 ถึง 88 มิลลิปรอท	3,183	12.4
89 ถึง 98 มิลลิปรอท	2,224	8.7
99 ถึง 108 มิลลิปรอท	2,132	8.3
มากกว่า 109 มิลลิปรอท	704	2.7
รวม	25,660	100

ตาราง 10 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คน แบ่งเป็นผู้ป่วยที่มีล่างความดันเลือด 80 ถึง 84 มิลลิปรอทมีจำนวน 17,417 คน คิดเป็นร้อยละ 67.9 ผู้ป่วยที่มีค่าล่างความดันเลือด 85 ถึง 88 มิลลิปรอทมีจำนวนคน 3,183 คิดเป็นร้อยละ 12.4 ผู้ป่วยที่มีค่าล่างความดันเลือด 89 ถึง 98 มิลลิปรอทมีจำนวน 2,224 คน คิดเป็นร้อยละ 8.7 ผู้ป่วยที่มีค่าล่างความดันเลือด 99 ถึง 108 มิลลิปรอทมีจำนวน 2,132 คนคิดเป็นร้อยละ 8.3 และผู้ป่วยที่ค่าล่างความดันเลือดมากกว่า 109 มิลลิปรอทมีจำนวน 704 คน คิดเป็นร้อยละ 2.7 ดังแสดงในภาพประกอบ 18

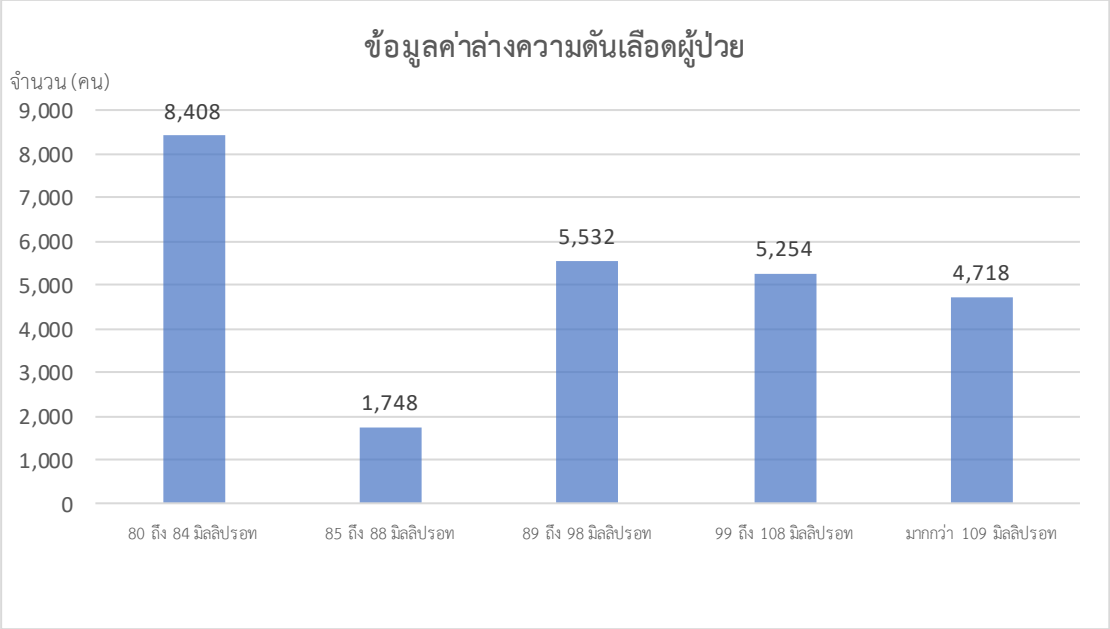


ภาพประกอบ 18 แผนภูมิข้อมูลค่าล่างความดันเลือดผู้ป่วย

ตาราง 11 ข้อมูลค่าบนความดันเลือดผู้ป่วย

ความดันเลือด	จำนวน	ร้อยละ
120 ถึง 129 มิลลิปรอท	8,408	32.8
130 ถึง 138 มิลลิปรอท	1,748	6.8
139 ถึง 158 มิลลิปรอท	5,532	21.6
159 ถึง 178 มิลลิปรอท	5,254	20.5
มากกว่า 178 มิลลิปรอท	4,718	18.4
รวม	25,660	100

ตาราง 11 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คนแบ่งเป็นผู้ป่วยที่มีค่าบนความดันเลือด 120 ถึง 129 มิลลิปรอทมีจำนวน 8,408 คน คิดเป็นร้อยละ 32.8 ผู้ป่วยที่มีค่าบนความดันเลือด 130 ถึง 138 มิลลิปรอทมีจำนวน 1,748 คน คิดเป็นร้อยละ 6.8 ผู้ป่วยที่มีค่าบนความดันเลือด 139 ถึง 158 มิลลิปรอทมีจำนวน 5,532 คน คิดเป็นร้อยละ 21.6 ผู้ป่วยที่มีค่าบนความดันเลือด 159 ถึง 178 มิลลิปรอทมีจำนวน 5,532 คน คิดเป็นร้อยละ 20.5 และผู้ป่วยที่ค่าบนความดันเลือดมากกว่า 178 มิลลิปรอทมีจำนวน 4,718 คน คิดเป็นร้อยละ 18.4 ดังแสดงในภาพประกอบ 19

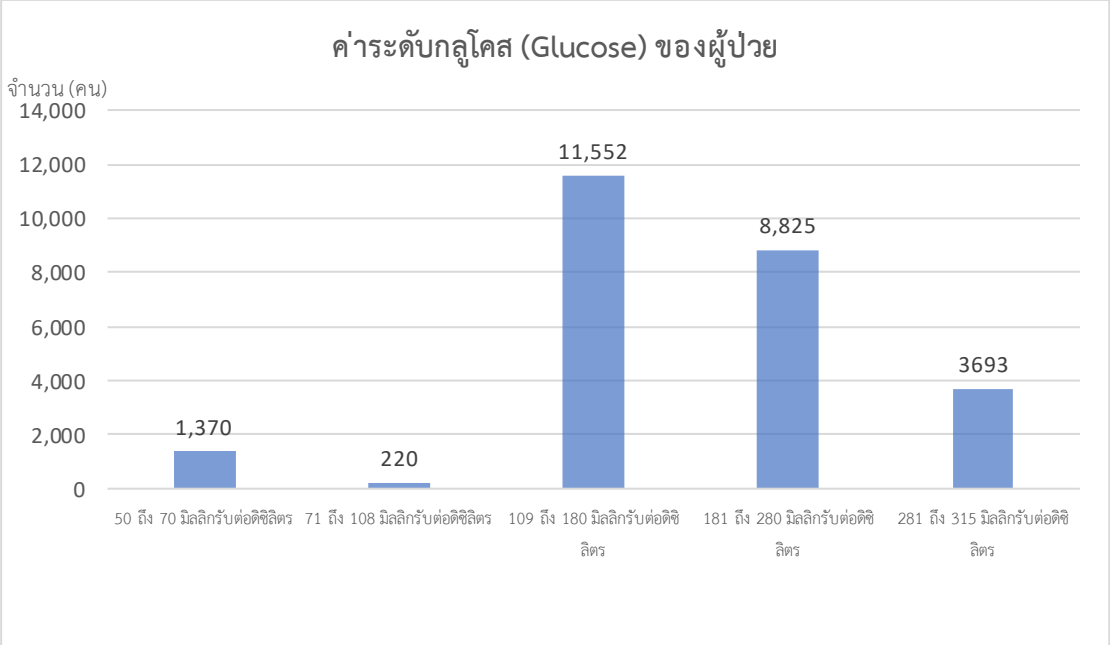


ภาพประกอบ 19 แผนภูมิข้อมูลค่าล่างความดันเลือดผู้ป่วย

ตาราง 12 ข้อมูลระดับกลูโคส (*Glucose*) ผู้ป่วย

ระดับกลูโคส (Glucose)	จำนวน	ร้อยละ
50 ถึง 70 มิลลิกรัมต่อดิซิลิตร	1,370	5
71 ถึง 108 มิลลิกรัมต่อดิซิลิตร	220	1
109 ถึง 180 มิลลิกรัมต่อดิซิลิตร	11,552	45
181 ถึง 280 มิลลิกรัมต่อดิซิลิตร	8,825	34
281 ถึง 315 มิลลิกรัมต่อดิซิลิตร	3693	14
รวม	25,660	100.0

ตาราง 33 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คนแบ่งเป็นผู้ป่วยที่มีระดับกลูโคส 50 ถึง 70 มิลลิกรัมต่อดิซิลิตรมีจำนวน 1,370 คน คิดเป็นร้อยละ 5 ผู้ป่วยที่มีระดับกลูโคส 71 ถึง 108 มิลลิกรัมต่อดิซิลิตรมีจำนวน 220 คนคิดเป็นร้อยละ 1 ผู้ป่วยที่มีระดับกลูโคส 109 ถึง 180 มิลลิกรัมต่อดิซิลิตรมีจำนวน 11,552 คนคิดเป็นร้อยละ 45 ผู้ป่วยที่มีระดับกลูโคส 181 ถึง 280 มิลลิกรัมต่อดิซิลิตรมีจำนวน 34 คนคิดเป็นร้อยละ 34 และผู้ป่วยที่มีระดับกลูโคส 281 ถึง 315 มิลลิกรัมต่อดิซิลิตรมีจำนวน 14 คน คิดเป็นร้อยละ 14 ดังแสดงในภาพประกอบ 20

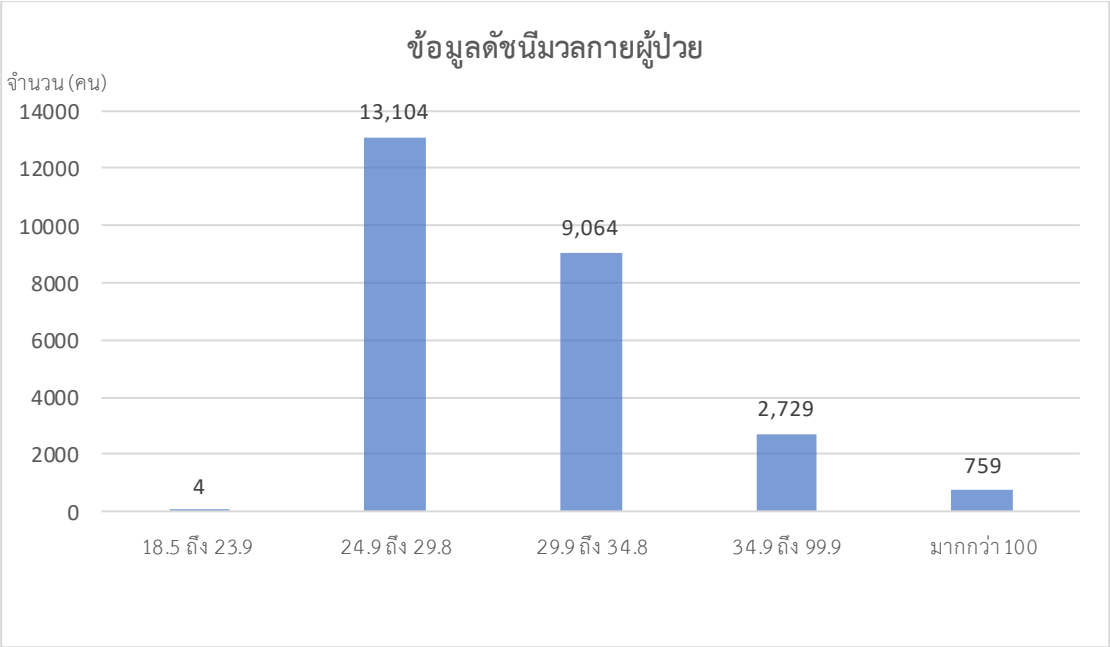


ภาพประกอบ 20 แผนภูมิค่าระดับกลูโคส (*Glucose)* ของผู้ป่วย

ตาราง 13 ข้อมูลดัชนีมวลกายผู้ป่วย

ดัชนีมวลกาย	จำนวน	ร้อยละ
18.5 ถึง 23.9	4	0
24.9 ถึง 29.8	13,104	51
29.9 ถึง 34.8	9,064	35
34.9 ถึง 99.9	2,729	11
มากกว่า 100	759	3
รวม	25,660	100.0

ตาราง 13 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คนแบ่งเป็นผู้ป่วยที่มีดัชนีมวลกาย 18.5 ถึง 23.9 มีจำนวน 4 คนคิดเป็นน้อยละ 0 ผู้ป่วยที่มีดัชนีมวลกาย 24.9 ถึง 29.8 จำนวน 13,104 คนคิดเป็นร้อยละ 51 ผู้ป่วยที่มีดัชนีมวลกาย 29.9 ถึง 34.8 จำนวน 9,064 คนคิดเป็นร้อยละ 35 ผู้ป่วยที่มีดัชนีมวลกาย 34.9 ถึง 99.9 มีจำนวน 2,729 คน คิดเป็นร้อยละ 11 และ ผู้ป่วยที่มีดัชนีมวลกายมากกว่า 100 มีจำนวน 759 คน คิดเป็นร้อยละ 3 ดังแสดงในภาพประกอบ 21

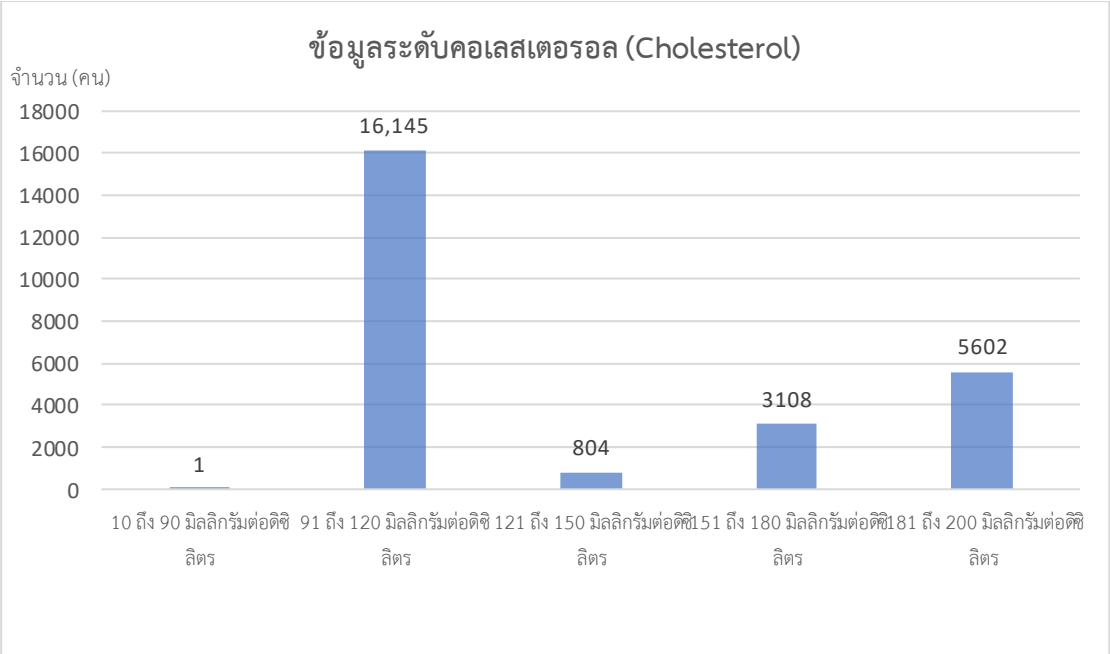


ภาพประกอบ 21 แผนภูมิข้อมูลดัชนีมวลกายผู้ป่วย

ตาราง 14 ข้อมูลคอเลสเตอรอล

คอเลสเตอรอล	จำนวน	ร้อยละ
10 ถึง 90 มิลลิกรัมต่อดิซิลิตร	1	0
91 ถึง 120 มิลลิกรัมต่อดิซิลิตร	16,145	63
121 ถึง 150 มิลลิกรัมต่อดิซิลิตร	804	3
151 ถึง 180 มิลลิกรัมต่อดิซิลิตร	3108	12
181 ถึง 200 มิลลิกรัมต่อดิซิลิตร	5602	22
รวม	**25,660**	**100.0**

ตาราง 14 แสดงถึงข้อมูลของผู้ป่วยโรคเบาหวานและข้อมูลผู้เข้ารับประกันสังคมจำนวน 25,660 คนแบ่งเป็นผู้ป่วยที่มีคอเลสเตอรอล 10 ถึง 90 มิลลิกรัมต่อดิซิลิตรมีจำนวน 1 คน คิดเป็นร้อยละ 0 ผู้ป่วยที่มีคอเลสเตอรอล 91 ถึง 120 มิลลิกรัมต่อดิซิลิตรมีจำนวน 16,145 คน คิดเป็นร้อยละ 63 ผู้ป่วยที่มีคอเลสเตอรอล 121 ถึง 150 มิลลิกรัมต่อดิซิลิตรมีจำนวน 804 คนคิดเป็นร้อยละ 3 ผู้ป่วยที่มีคอเลสเตอรอล 151 ถึง 180 มิลลิกรัมต่อดิซิลิตรมีจำนวน 3,108 คนคิดเป็นร้อยละ 12 และผู้ป่วยที่มีคอเลสเตอนอล 181 ถึง 200 มิลลิกรัมต่อดิซิลิตรมีจำนวน 5,602 คนคิดเป็นร้อยละ 22 ดังแสดงในภาพประกอบ 22

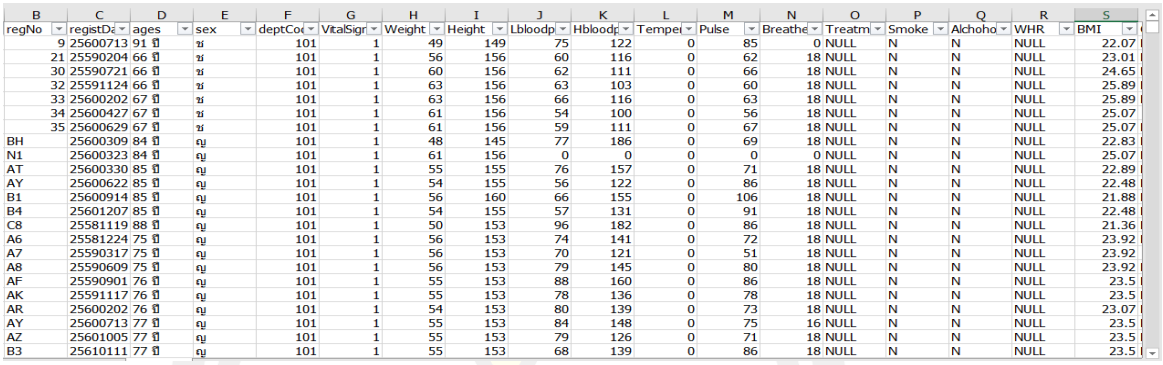


ภาพประกอบ 22 แผนภูมิข้อมูลคอเลสเตอรอล (Cholesterol)

4.3 ผลการวิเคราะห์ในขั้นตอนการเตรียมข้อมูล (Data Preparation)

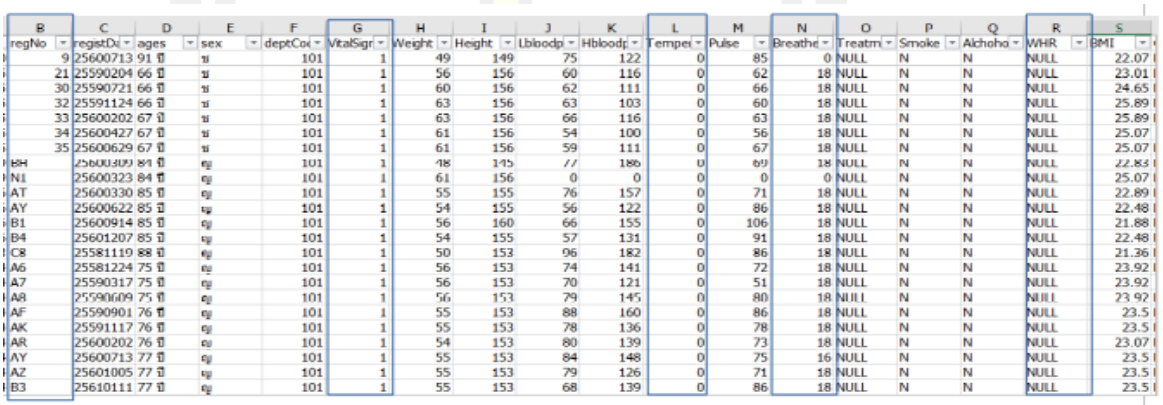
ในขั้นตอนการเตรียมข้อมูล ผู้วิจัยได้ทำการคัดเลือกข้อมูลผู้ป่วยโรคเบาหวานผู้เข้ารับประกันสังคมของโรงพยาบาลศูนย์อุดรธานี ตั้งแต่ปี พ.ศ. 2558 ถึง พ.ศ. 2564 มีจำนวนข้อมูลผู้ป่วยโรคเบาหวานจำนวน 70,420 แถว โดยข้อมูลแบ่งออกเป็น รหัสผู้ป่วย จำนวนครั้งในการเข้าพบแพทย์ วันที่เข้ารับการรักษา อายุ เพศ รหัสโรค สัญญาณชีพ น้ำหนัก ส่วนสูง ระดับค่าล่างความดันเลือด ระดับค่าบนความดันเลือด อุณหภูมิ การเต้นของชีพจร การหายใจ การรักษา การสูบบุหรี่ แอลกอฮอล์ สัดส่วนรอบเอว ดัชนีมวลกาย (BMI) ระดับคอเลสเตอรอล และระดับกลูโคส

ในส่วนของการทำความสะอาดข้อมูล ผู้วิจัยได้จัดการข้อมูลให้อยู่ในสภาพพร้อมใช้ ได้แก่ การตัดคอลัมน์ (Remove Attribute) จำนวนครั้งที่เข้ารักษา (RegNo) อุณหภูมิ (Tempurature) การรักษา(Treatment) สัดส่วนรอบเอว (WHR) การหายใจ (Breath) เหตุผลคือข้อมูลนั้นมีรูปแบบที่ไม่สามารถอ่านค่าได้และเก็บค่าเป็น null เสียส่วนใหญ่ อีกทั้งไม่มีความจำเป็นต่อการวินิจฉัย จากนั้นจึงทำการแก้ไขชื่อคอลัมน์เพื่อให้เหมาะสมแก่การประมวลผลและมีความเข้าใจง่าย ต่อไปได้ทำการนำค่าว่างออกจากแอททริบิวท์ ระดับค่าล่างความดันเลือด (Lbloodpress) ระดับค่าบนความดันเลือด (Hbloodpress) ค่าการเต้นของชีพจร (Pulse) ระดับค่าคอเลสเตอรอล (Cholesterol) และระดับกลูโคส (Glucose)



regNo	registDa	ages	sex	deptCo	VitalSigr	Weight	Height	Lbloodp	Hbloodp	Tempe	Pulse	Breathe	Treatm	Smoke	Alchoho	WHR	BMI
9	25600713	91	ช	101	1	49	149	75	122	0	85	0	NULL	N	N	NULL	22.07
21	25590204	66	ช	101	1	56	156	60	116	0	62	18	NULL	N	N	NULL	23.01
30	25590721	66	ช	101	1	60	156	62	111	0	66	18	NULL	N	N	NULL	24.65
32	25591124	66	ช	101	1	63	156	63	103	0	60	18	NULL	N	N	NULL	25.89
33	25600202	67	ช	101	1	63	156	66	116	0	63	18	NULL	N	N	NULL	25.89
34	25600427	67	ช	101	1	61	156	54	100	0	56	18	NULL	N	N	NULL	25.07
35	25600629	67	ช	101	1	61	156	59	111	0	67	18	NULL	N	N	NULL	25.07
BH	25600309	84	ญ	101	1	48	145	77	186	0	69	18	NULL	N	N	NULL	22.83
N1	25600323	84	ญ	101	1	61	156	0	0	0	0	0	NULL	N	N	NULL	25.07
AT	25600330	85	ญ	101	1	55	155	76	157	0	71	18	NULL	N	N	NULL	22.89
AY	25600622	85	ญ	101	1	54	155	56	122	0	86	18	NULL	N	N	NULL	22.48
B1	25600914	85	ญ	101	1	56	160	66	155	0	106	18	NULL	N	N	NULL	21.88
B4	25601207	85	ญ	101	1	54	155	57	131	0	91	18	NULL	N	N	NULL	22.48
C8	25581119	88	ญ	101	1	50	153	96	182	0	86	18	NULL	N	N	NULL	21.36
A6	25581224	75	ญ	101	1	56	153	74	141	0	72	18	NULL	N	N	NULL	23.92
A7	25590317	75	ญ	101	1	56	153	70	121	0	51	18	NULL	N	N	NULL	23.92
A8	25590609	75	ญ	101	1	56	153	79	145	0	80	18	NULL	N	N	NULL	23.92
AF	25590901	76	ญ	101	1	55	153	88	160	0	86	18	NULL	N	N	NULL	23.5
AK	25591117	76	ญ	101	1	55	153	78	136	0	78	18	NULL	N	N	NULL	23.5
AR	25600202	76	ญ	101	1	54	153	80	139	0	73	18	NULL	N	N	NULL	23.07
AY	25600713	77	ญ	101	1	55	153	84	148	0	75	16	NULL	N	N	NULL	23.5
AZ	25601005	77	ญ	101	1	55	153	79	126	0	71	18	NULL	N	N	NULL	23.5
B3	25610111	77	ญ	101	1	55	153	68	139	0	86	18	NULL	N	N	NULL	23.5

ภาพประกอบ 23 ข้อมูลผู้ป่วยตั้งแต่ปี พ.ศ. 2558-2564



regNo	registDa	ages	sex	deptCo	VitalSigr	Weight	Height	Lbloodp	Hbloodp	Tempe	Pulse	Breathe	Treatm	Smoke	Alchoho	WHR	BMI
9	25600713	91	ช	101	1	49	149	75	122	0	85	0	NULL	N	N	NULL	22.07
21	25590204	66	ช	101	1	56	156	60	116	0	62	18	NULL	N	N	NULL	23.01
30	25590721	66	ช	101	1	60	156	62	111	0	66	18	NULL	N	N	NULL	24.65
32	25591124	66	ช	101	1	63	156	63	103	0	60	18	NULL	N	N	NULL	25.89
33	25600202	67	ช	101	1	63	156	66	116	0	63	18	NULL	N	N	NULL	25.89
34	25600427	67	ช	101	1	61	156	54	100	0	56	18	NULL	N	N	NULL	25.07
35	25600629	67	ช	101	1	61	156	59	111	0	67	18	NULL	N	N	NULL	25.07
BH	25600309	84	ญ	101	1	48	145	77	186	0	69	18	NULL	N	N	NULL	22.83
N1	25600323	84	ญ	101	1	61	156	0	0	0	0	0	NULL	N	N	NULL	25.07
AT	25600330	85	ญ	101	1	55	155	76	157	0	71	18	NULL	N	N	NULL	22.89
AY	25600622	85	ญ	101	1	54	155	56	122	0	86	18	NULL	N	N	NULL	22.48
B1	25600914	85	ญ	101	1	56	160	66	155	0	106	18	NULL	N	N	NULL	21.88
B4	25601207	85	ญ	101	1	54	155	57	131	0	91	18	NULL	N	N	NULL	22.48
C8	25581119	88	ญ	101	1	50	153	96	182	0	86	18	NULL	N	N	NULL	21.36
A6	25581224	75	ญ	101	1	56	153	74	141	0	72	18	NULL	N	N	NULL	23.92
A7	25590317	75	ญ	101	1	56	153	70	121	0	51	18	NULL	N	N	NULL	23.92
A8	25590609	75	ญ	101	1	56	153	79	145	0	80	18	NULL	N	N	NULL	23.92
AF	25590901	76	ญ	101	1	55	153	88	160	0	86	18	NULL	N	N	NULL	23.5
AK	25591117	76	ญ	101	1	55	153	78	136	0	78	18	NULL	N	N	NULL	23.5
AR	25600202	76	ญ	101	1	54	153	80	139	0	73	18	NULL	N	N	NULL	23.07
AY	25600713	77	ญ	101	1	55	153	84	148	0	75	16	NULL	N	N	NULL	23.5
AZ	25601005	77	ญ	101	1	55	153	79	126	0	71	18	NULL	N	N	NULL	23.5
B3	25610111	77	ญ	101	1	55	153	68	139	0	86	18	NULL	N	N	NULL	23.5

ภาพประกอบ 24 การตัดคอลัมน์ข้อมูลที่ไม่จำเป็นออก

เพื่อให้ข้อมูลเหมาะสมสำหรับโมเดลต่าง ๆ จึงต้องแปลงข้อมูลให้เหมาะสม โดยเริ่มจากการเปลี่ยนค่าแอทริบิวท์ sex ซึ่งหมายถึงเพศจากค่าเดิมคือ ช กับ ญ ซึ่งหมายถึงผู้ป่วยเพศชายและหญิงจึงเปลี่ยนเป็น male และ female ในส่วนแอททริบิวท์ smoke ซึ่งหมายถึงการสูบบุหรี่ซึ่งมีค่าเป็น สูบ และ ไม่สูบ ได้แปลงค่าเป็น Yes แทนสูบและ No แทนไม่สูบ ในส่วนแอททริบิวท์ Alcohol ซึ่งหมายถึงการดื่มแอลกอฮอล์มีค่าในแอททริบิวท์เป็น No และ Yes ซึ่งหมายถึง ไม่ดื่มและดื่มตามลำดับได้ที่การแก้ไขเป็น no กับ yes ในส่วนของค่าว่างในตารางแอททริบิวท์ ระดับค่าบนความดันเลือด (Lbloodpress) ระดับค่าล่างความดันเลือด (Hbloodpress) ค่าการเต้นของชีพจร (Pulse) ระดับกลูโคส (Glucose) และ ระดับค่าคอเลสเตอรอล (Cholesterol) นั้นได้ทำการกำจัดค่าว่างเพื่อให้สามารถใช้ในการทำแบบจำลองต่อไป ในส่วนของการกำหนดหน้าที่ใหม่ ได้กำหนดค่า hn ให้ทำหน้าที่เป็น ID และในแอททริบิวท์อายุ (age) ในการกำจัดค่าสุดโต่ง (Outliner) ผู้วิจัยได้ใช้วิธีการค้นหาค่าสุดโต่งในข้อมูลของแต่ละโรค ซึ่งได้กำหนดจำนวน k เท่ากับ 4 และค่าสุดโต่งเท่ากับ 40 และใช้เทคนิคการวัดระยะทางในการนั้นจึงเทคการวัดระยะทาง (Euclidian distance) ในการค้นหาค่าสุดโต่ง ในส่วนของแอททริบิวท์ดัชนีมวลกาย (BMI) ระดับค่าบนความดันเลือด (Hbloodpress) ระดับ

ค่าล่างความดันเลือด (Lbloodpress) ความสูง (Height) ค่าการเต้นของชีพจร (Pulse) และน้ำหนัก (Weight) มีค่าว่างอยู่จึงได้กำจัดออกเพื่อให้สามารถใช้ทำแบบจำลองต่อไป จากนั้นจึงกรองข้อมูลสุดโต่งและคัดออกไปในที่สุด สุดท้ายผู้วิจัยได้สร้างแอททิบิวท์ใหม่ขึ้นมาคือ Disease หรือโรคเพื่อใช้ในการทำนายการเกิดโรคโดยข้อมูลผู้ป่วยเบาหวานจะแทนค่า Diabetes หรือเป็นโรคเบาหวาน และข้อมูลผู้ที่เข้ารับประกันสังคมจะแทนค่าเป็น Non-Diabetes หรือไม่เป็นโรคเบาหวาน ในการแปลงช่วงข้อมูลเพื่อเตรียมข้อมูลเพื่อสร้างแบบจำลองต้นไม้การตัดสินใจ (Decision Tree) ได้แปลงค่าในแอททริบิวท์ดังนี้

ดัชนีมวลกาย (BMI) ได้ทำการแบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 18 แบ่งเป็นช่วงน้ำหนักต่ำกว่าเกณฑ์

18.5 – 24.5 แบ่งเป็นช่วงปกติ

25 – 29.9 แบ่งเป็นช่วงน้ำหนักเกิน

30 – 34.9 แบ่งเป็นช่วงอ้วน

มากกว่า 35 แบ่งเป็นช่วงอ้วนมาก

ค่าคอเลสเตอรอล (Cholesterol) ได้ทำการแบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 50 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงปกติ

91 – 120 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงดีที่สุด

121 – 150 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงเส้นแบ่ง

151 – 180 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงสูง

181 – 200 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงสูงแบบอันตราย

ระดับกลูโคส (Glucose) ได้ทำการแบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 70 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงต่ำ

71 – 108 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงปกติ

109 – 180 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงเส้นแบ่ง

181 – 280 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงสูง

281 – 315 มิลลิกรัมต่อเดซิลิตร แบ่งเป็นช่วงสุงแบบอันตราย

อายุ (Age) ได้ทำการแบ่งช่วงข้อมูลเป็น 4 ช่วงคือ

18 – 35 ปี แบ่งเป็นช่วงวัยผู้ใหญ่ตอนต้น

36 – 59 ปี แบ่งเป็นช่วงวัยกลางคน

60 – 79 ปี แบ่งเป็นช่วงสูงวัย

มากกว่า 80 ปี แบ่งเป็นช่วงปลายอายุ

ค่าบนความดันเลือด (Hbloodpress) ได้แบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 120 มิลลิปรอท แบ่งเป็นช่วงดี

121 – 130 มิลลิปรอท แบ่งเป็นช่วงปกติ

131 – 139 มิลลิปรอท แบ่งเป็นช่วงปกติค่อนข้างสูง

140 – 159 มิลลิปรอท แบ่งเป็นช่วงระยะที่ 1

160 – 179 มิลลิปรอท แบ่งเป็นช่วงระยะที่ 2

ค่าล่างความดันเลือด (Hbloodpress) ได้แบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 80 มิลลิปรอท แบ่งเป็นช่วงดี

81 – 85 มิลลิปรอท แบ่งเป็นช่วงปกติ

86 – 89 มิลลิปรอท แบ่งเป็นช่วงปกติค่อนข้างสูง

90 – 99 มิลลิปรอท แบ่งเป็นช่วงระยะที่ 1

100 – 109 มิลลิปรอท แบ่งเป็นช่วงระยะที่ 2

การเต้นของชีพจร (Pulse) ได้แบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 60 ครั้งต่อนาที แบ่งเป็นช่วงยอดเยี่ยม

61 – 65 ครั้งต่อนาที แบ่งเป็นช่วงดี

66 – 70 ครั้งต่อนาที แบ่งเป็นช่วงสูงกว่ามาตรฐาน

71 – 75 ครั้งต่อนาที แบ่งเป็นช่วงมาตรฐาน

76 – 80 ครั้งต่อนาที แบ่งเป็นช่วงต่ำกว่ามาตรฐาน

ความสูง (Height) ได้แบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 135 เซนติเมตร แบ่งเป็นช่วงเตี้ยมาก

136 – 150 เซนติเมตร แบ่งเป็นช่วงเตี้ย

151 – 165 เซนติเมตร แบ่งเป็นช่วงมาตรฐาน

166 – 180 เซนติเมตร แบ่งเป็นช่วงสูง

181 – 200 เซนติเมตร แบ่งเป็นช่วงสูงที่สุด

น้ำหนัก (Weight) ได้แบ่งช่วงข้อมูลเป็น 5 ช่วงคือ

น้อยกว่า 40 เซนติเมตร แบ่งเป็นช่วงต่ำกว่าเกณฑ์

41 – 60 เซนติเมตร แบ่งเป็นช่วงมาตรฐาน

61 – 80 เซนติเมตร แบ่งเป็นช่วงอวบ

81 – 100 เซนติเมตร แบ่งเป็นช่วงอ้วน

101 – 120 เซนติเมตร แบ่งเป็นช่วงอ้วนมาก

4.4 ผลการวิเคราะห์ในขั้นตอนการสร้างแบบจำลอง (Modeling)

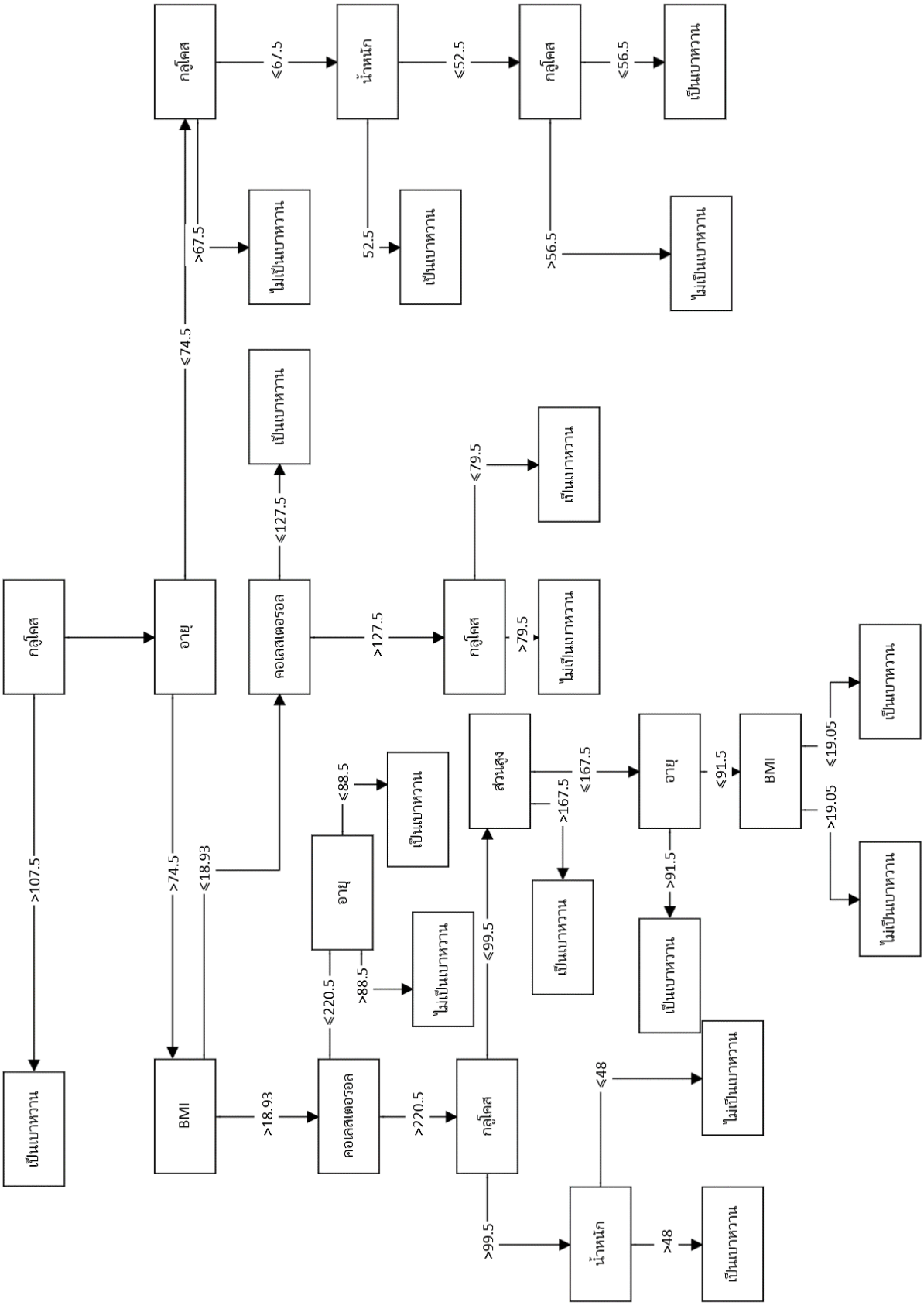
ในส่วนของการสร้างแบบจำลองในการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจ โรคความดันโลหิตสูง โรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี ผู้วิจัยได้ใช้เทคนิค Classification 3 เทคนิคอันได้แก่ ต้นไม้การตัดสินใจ (Decision Tree), ทฤษฎีนาอิฟ เบย์ (Naïve Bay) และ เทคนิคเพื่อนใกล้มากที่สุด ((k-NN) K-Nearest Neighbor) และ 2 เทคนิคเพิ่มเติมคือเทคนิคโหวตร่วม (Vote Ensemble) และเทคนิคป่าสุ่ม (Random Forest) ซึ่งผลลัพธ์จะออกมาเป็นดังนี้

4.4.1 ผลลัพธ์เทคนิคต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจจะทำการจัดกลุ่มชุดข้อมูลโดยนำเข้าในแต่ละกรณี แต่ละบัพของต้นไม้การตัดสินใจคือตัวแปรต่าง ๆ ของชุดข้อมูลและมีตัวแปรตาม ซึ่งแต่ละตัวแปรนั้นมีค่าของตัวเองเกิดเป็นชุดค่าของตัวแปร การทำนายประเภทด้วยต้นไม้จะเริ่มจากบัพราก โดยทดสอบตัวแปรค่าบัพรากโดยทดสอบจากค่าของบัพ ผู้วิจัยได้กำหนดค่าพารามิเตอร์ โดยกำหนดชั้นของต้นไม้เป็น 5 ชั้น แล้วจึงตามกิ่งของต้นไม้ที่กำหนดค่า เพื่อไปยังบัพลูกถัดไป โดยผลการทำนายจะเกิดเป็นดังนี้

Tree
กลูโคส > 107.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =1744, เป็นเบาหวาน =12609}
กลูโคส ≤ 107.500
| อายุ > 74.500
| | BMI > 18.930
| | | คอเรสโตรอล > 220.500
| | | | กลูโคส > 99.500
| | | | | น้ำหนัก > 48: เป็นเบาหวาน {ไม่เป็นเบาหวาน =5, เป็นเบาหวาน =20}
| | | | | น้ำหนัก ≤ 48: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =2, เป็นเบาหวาน =0}
| | | | กลูโคส ≤ 99.500
| | | | | ส่วนสูง > 167.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =0, เป็นเบาหวาน =4}
| | | | | ส่วนสูง ≤ 167.500
| | | | | | อายุ > 91.500: Diabetic {ไม่เป็นเบาหวาน =0, เป็นเบาหวาน =2}
| | | | | | อายุ ≤ 91.500
| | | | | | | BMI > 19.050: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =38, เป็นเบาหวาน =7}

| | | | | | | BMI ≤ 19.050: เป็นเบาหวาน {ไม่เป็นเบาหวาน =0, เป็นเบาหวาน =2}
| | | โคเรสเตอรอล≤ 220.500
| | | | อายุ > 88.500: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =2, เป็นเบาหวาน =0}
| | | | อายุ ≤ 88.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =65, เป็นเบาหวาน =179}
| | BMI ≤ 18.930
| | | โคเรสเตอรอล> 127.500
| | | | กลูโคส > 79.500: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =29, เป็นเบาหวาน =5}
| | | | กลูโคส ≤ 79.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =1, เป็นเบาหวาน =3}
| | | โคเรสเตอรอล ≤ 127.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =1, เป็นเบาหวาน =4}
| อายุ ≤ 74.500
| | กลูโคส > 67.500: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =9657, เป็นเบาหวาน =1344}
| | กลูโคส ≤ 67.500
| | | น้ำหนัก > 52.500: เป็นเบาหวาน {ไม่เป็นเบาหวาน =13, เป็นเบาหวาน =75}
| | | น้ำหนัก ≤ 52.500
| | | | กลูโคส > 56.500: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =13, เป็นเบาหวาน =2}
| | | | กลูโคส ≤ 56.500: ไม่เป็นเบาหวาน {ไม่เป็นเบาหวาน =2, ไม่เป็นเบาหวาน =7}



ภาพประกอบ 25 แผนภาพต้นไม้การตัดสินใจ

จากผลลัพธ์การพยากรณ์ หากผู้ป่วยมีกลูโคสมากกว่า 107.5 มิลลิกรัมผลลัพธ์คือเป็นโรคเบาหวาน หากผู้ป่วยกลูโคสน้อยกว่าเท่ากับ 107.5 มิลลิกรัม มีอายุมากกว่า 74 ปี มี BMI มากกว่า 18.9 มีคอเลสเตอรอลมากกว่า 220.5 มิลลิกรัม มีกลูโคสมากกว่า 99.5 มิลลิกรัมแล้วมีน้ำหนักมากกว่า 48 กิโลกรัมผลลัพธ์คือเป็นโรคเบาหวาน แต่ถ้าผู้ป่วยมีน้ำหนักน้อยกว่าเท่ากับ 48 กิโลกรัมผลลัพธ์คือเป็นไม่โรคเบาหวาน ถ้ามีกลูโคสน้อยกว่าเท่ากับ 99.5 มิลลิกรัม มีความสูงมากกว่า 167.5 เซนติเมตรผลลัพธ์คือเป็นโรคเบาหวาน แต่ถ้าผู้ป่วยมีความสูงน้อยกว่าหรือเท่ากับ 167.5 เซนติเมตรแต่มีอายุมากกว่า 91 ปีผลลัพธ์คือเป็นโรคเบาหวาน แต่มีอายุน้อยกว่าหรือเท่ากับ 91 ปี แล้วมี BMI มากกว่า 19 ปีผลลัพธ์คือไม่เป็นโรคเบาหวาน แต่ถ้ามี BMI น้อยกว่าหรือเท่ากับ 19 ผลลัพธ์คือเป็นโรคเบาหวาน ถ้าผู้ป่วยมีคอลเลสเตอรอลน้อยกว่าหรือเท่ากับ 220.5 มิลลิกรัมแล้วมีอายุมากกว่า 88 ปี ผลลัพธ์คือไม่เป็นโรคเบาหวาน แต่ถ้ามีอายุน้อยกว่าหรือเท่ากับ 88 ปี ผลลัพธ์คือเป็นโรคเบาหวาน ถ้าผู้ป่วยมี BMI น้อยกว่าหรือเท่ากับ 18.9 แล้วมีคอเลสเตอรอลมากกว่า 127.5 มิลลิกรัม มีกลูโคสมากกว่า 79.5 มิลลิกรัมผลลัพธ์คือไม่เป็นโรคเบาหวาน แต่ถ้ามีกลูโคสน้อยกว่าหรือเท่ากับ 79.5 มิลลิกรัมผลลัพธ์คือเป็นโรคเบาหวาน แต่ถ้าคอเลสเตอรอลน้อยกว่าหรือเท่ากับ 127.5 มิลลิกรัมผลลัพธ์คือเป็นโรคเบาหวาน ผู้ป่วยที่มีอายุน้อยกว่าหรือเท่ากับ 74 ปี มีกลูโคสมากกว่า 67.5 มิลลิกรัมผลลัพธ์คือไม่เป็นโรคเบาหวาน แต่ถ้ามีกลูโคสน้อยกว่าหรือเท่ากับ 67.5 มิลลิกรัมแล้วมีน้ำหนักมากกว่า 52.5 กิโลกรัมผลลัพธ์คือเป็นโรคเบาหวาน แต่ถ้าผู้ป่วยมีน้ำหนักน้อยกว่าหรือเท่ากับ 52.5 กิโลกรัมแล้วมีกลูโคสมากกว่า 56.5 มิลลิกรัมผลลัพธ์คือไม่เป็นโรคเบาหวาน แต่ถ้ามีกลูโคสน้อยกว่าหรือเท่ากับ 56.5 มิลลิกรัมผลลัพธ์คือเป็นโรคเบาหวาน

4.4.2 ผลลัพธ์เทคนิคนาอิฟ เบย์ (Naïve Bay)

Naïve Bay คือโมเดลที่ใช้ในการจัดกลุ่มโดยอาศัยหลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Beyes' Theorem และสมมุติฐานที่กำหนดให้เกิดความเป็นอิสระต่อกัน ซึ่งการเรียนรู้จำแนกด้วยกระบวนการของ Naïve Bayes นี้ได้ถูกนำมาใช้ในการคำนวณที่ไม่ซับซ้อนแต่ทว่าทำงานได้อย่างมีประสิทธิภาพ ในการจัดกลุ่มโดยอาศัย Naïve Bay อาจจะเกิดเหตุการณ์และปัจจัยในการแบ่งกลุ่มมากกว่า 1 ชนิดเมื่อทำการประยุกต์ใช้งานร่วมกับ Beyes' Theorem แล้วเกิดการคำนวณที่ซับซ้อนมากขึ้น เนื่องจากการเกิดขึ้นของเหตุการณ์นั้นเป็นอิสระต่อกันจะทำให้เกิดการคำนวณที่มีจำนวนรอบมากขึ้น ซึ่งผลการทำนายจะเกิดขึ้นดังนี้

SimpleDistribution

Distribution model for label attribute Disease

Class ไม่เป็นโรคเบาหวาน (0.448)

12 distributions

Class เป็นโรคเบาหวาน (0.552)

ภาพประกอบ 26 ผลลัพธ์การจำแนกประเภทข้อมูลด้วยเทคนิคเทคนิคนาอิฟ เบย์ (Naïve Bay)

จากาการประมวลผลเทคนิคนาอิฟ เบย์ (Naïve Bay) มีโอกาสที่จะไม่เป็นโรคเบาหวาน 44.8% มีโอกาสที่จะเป็นโรคเบาหวาน 55.2%

ตาราง 15 ความน่าจะเป็นของผลลัพธ์การจำแนกประเภทข้อมูลด้วยเทคนิคนาอิฟ เบย์ *(Naïve Bay)*

แอททริบิวท์	การวัดค่า	ไม่เป็นเบาหวาน	เป็นเบาหวาน
อายุ	ค่าเฉลี่ย	50.79	59.61
อายุ	ส่วนเบี่ยงเบนมาตรฐาน	11.76	11.60
เพศ	ค่า=เพศหญิง	0.65	0.62
เพศ	ค่า=เพศชาย	0.35	0.38
น้ำหนัก	ค่าเฉลี่ย	61.42	65.89
น้ำหนัก	ส่วนเบี่ยงเบนมาตรฐาน	11.68	22.36
ส่วนสูง	ค่าเฉลี่ย	159.16	158.14
ส่วนสูง	ส่วนเบี่ยงเบนมาตรฐาน	8.02	15.27
ความดันเลือดค่าล่าง	ค่าเฉลี่ย	77.74	76.73
ความดันเลือดค่าล่าง	ส่วนเบี่ยงเบนมาตรฐาน	84.49	19.49

แอททริบิวท์	การวัดค่า	ไม่เป็นเบาหวาน	เป็นเบาหวาน
ความดันเลือดค่าบน	ค่าเฉลี่ย	129.65	135.65
ความดันเลือดค่าบน	ส่วนเบี่ยงเบนมาตรฐาน	226.40	33.88
ชีพจร	ค่าเฉลี่ย	81.12	84.09
ชีพจร	ส่วนเบี่ยงเบนมาตรฐาน	85.46	18.41
การสูบบุหรี่	ค่า=ไม่สูบ	0.92	0.96
การสูบบุหรี่	ค่า=สูบ	0.08	0.04
การดื่มแอลกอฮอล์	ค่า=ไม่ดื่ม	0.87	0.96
การดื่มแอลกอฮอล์	ค่า=ดื่ม	0.13	0.04
BMI	ค่าเฉลี่ย	24.25	26.38
BMI	ส่วนเบี่ยงเบนมาตรฐาน	4.62	33.98
คอเลสเตอรอล	ค่าเฉลี่ย	210.11	194.27
คอเลสเตอรอล	ส่วนเบี่ยงเบนมาตรฐาน	45.50	57.37
กลูโคส	ค่าเฉลี่ย	98.95	173.83
กลูโคส	ส่วนเบี่ยงเบนมาตรฐาน	34.90	75.05

4.4.3 ผลลัพธ์เทคนิคเพื่อนใกล้มากที่สุด ((KNN) K-Nearest Neighbor)

เทคนิคเพื่อนใกล้มากที่สุด ((KNN) K-Nearest Neighbor) เป็นวิธีการที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะมีเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่าง ๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยผลลัพธ์การพยากรณ์จะเกิดขึ้นดังนี้

KNNClassification

Weighted 5-Nearest Neighbour model for classification.

The model contains 25835 examples with 12 dimensions of the following classes:

Non-Diabetic (ไม่เป็นโรคเบาหวาน)

Diabetic (เป็นโรคเบาหวาน)

ภาพประกอบ 27 ผลลัพธ์การจำแนกประเภทข้อมูลด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด *((*k-NN) K-Nearest Neighbor)

จากการประมวลผลแบบจำลองเทคนิคเพื่อนบ้านใกล้ที่สุด ((k-NN) K-Nearest Neighbor) มี 29 แถวที่เข้าใกล้มากที่สุด ผู้ป่วยที่ทำนายเป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1 ผู้ป่วยที่ทำนายไม่เป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1

Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse
1	Non-Diabetic	Diabetic	0.598	0.402	71	Male	56	163	68	123	91
2	Non-Diabetic	Non-Diabetic	0	1	49	Female	44	158	63	108	75
3	Non-Diabetic	Non-Diabetic	0	1	44	Female	54	140	60	108	73
4	Non-Diabetic	Non-Diabetic	0	1	41	Female	56	157	81	128	87
5	Non-Diabetic	Diabetic	1	0	74	Male	65	174	87	166	78
6	Non-Diabetic	Diabetic	0.599	0.401	48	Female	60	147	76	124	80
7	Non-Diabetic	Non-Diabetic	0.206	0.794	60	Female	52	151	69	138	63
8	Non-Diabetic	Non-Diabetic	0.197	0.803	61	Male	65	160	90	135	76
9	Non-Diabetic	Non-Diabetic	0	1	60	Female	44	162	60	90	67
10	Non-Diabetic	Non-Diabetic	0	1	34	Female	63	165	79	107	80
11	Non-Diabetic	Diabetic	1	0	78	Female	65	165	79	160	90
12	Non-Diabetic	Non-Diabetic	0	1	43	Female	65	150	71	112	84
13	Non-Diabetic	Non-Diabetic	0.410	0.590	56	Male	65	165	76	121	81
14	Non-Diabetic	Diabetic	0.600	0.400	81	Female	54	150	70	140	77
15	Non-Diabetic	Non-Diabetic	0.201	0.799	73	Female	58	148	83	116	98
16	Non-Diabetic	Diabetic	0.804	0.196	52	Male	71	170	87	130	108
17	Non-Diabetic	Non-Diabetic	0.198	0.802	49	Female	60	156	53	117	77
18	Non-Diabetic	Non-Diabetic	0	1.000	63	Female	57	153	76	118	74
19	Non-Diabetic	Diabetic	0.800	0.200	55	Male	74	165	79	117	93
20	Non-Diabetic	Non-Diabetic	0	1	58	Female	56	145	77	135	83
21	Non-Diabetic	Non-Diabetic	0	1	63	Female	42	148	79	129	96
22	Non-Diabetic	Non-Diabetic	0.413	0.587	51	Female	95	150	85	127	95
23	Non-Diabetic	Non-Diabetic	0	1	37	Female	58	160	72	120	74
24	Non-Diabetic	Non-Diabetic	0	1	49	Female	62	150	73	130	71
25	Non-Diabetic	Non-Diabetic	0	1	51	Female	65	155	80	119	75

ExampleSet (25,660 examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 28 เทคนิคเพื่อนบ้านใกล้ที่สุด ((KNN) K-Nearest Neighbor)

4.4.4 ผลลัพธ์เทคนิคการโหวตร่วม (Vote Ensmeble)

เทคนิคการโหวตร่วม (Vote Ensemble) เป็นเทคนิคการฝึกข้อมูลชุดเดียวกัน ทำให้เกิดความหลากหลายมากขึ้น ในการทดลองครั้งนี้ได้นำเทคนิคการจำแนก 3 เทคนิคได้แก่เทคนิคเพื่อนบ้านใกล้ที่สุด ((KNN) K-Nearest Neighbor) เทคนิคเทคนิคนาอิฟ เบย์ (Naïve Bay) และเทคนิคต้นไม้การตัดสินใจ (Decision Tree) และเทคนิคป่าสุ่ม (Random Forest) ฝึกร่วมกันทำให้เกิด Unseen data ทำให้เกิดผลลัพธ์การทำนายใหม่พบว่าเป็นโรคเบาหวาน โดยผู้ป่วยที่ทำนายเป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1 ผู้ป่วยที่ทำนายไม่เป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1

Attribute Based Voting

Using the majority of the following attributes for prediction:

base_prediction0

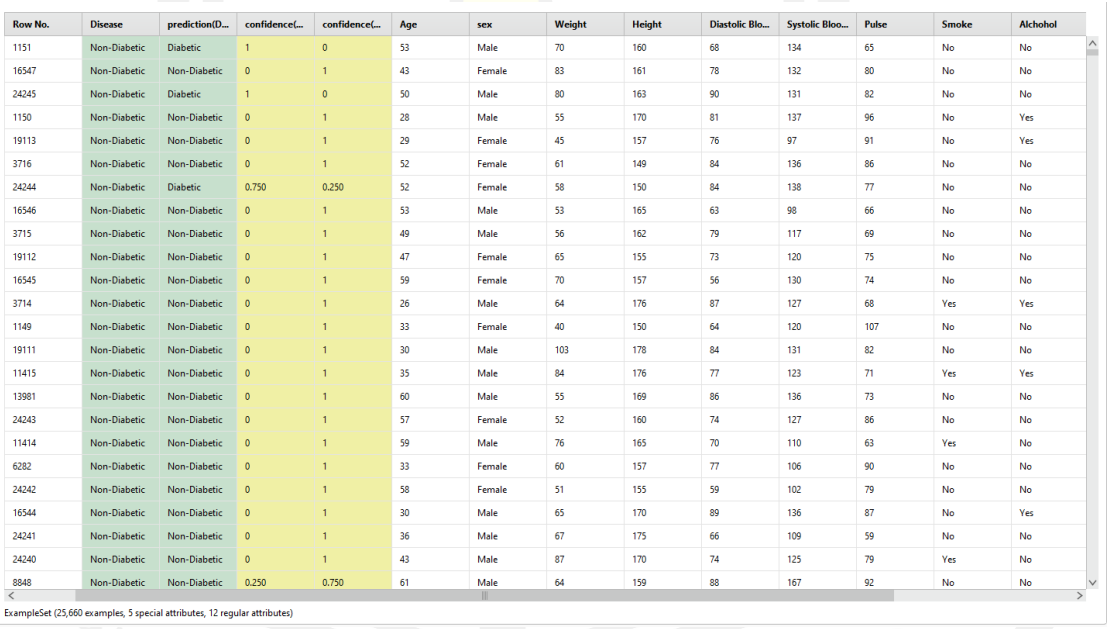
base_prediction1

base_prediction2

base_prediction3

The default value is Diabetic

ภาพประกอบ 29 ผลลัพธ์การจำแนกประเภทข้อมูลด้วยเทคนิคการโหวตร่วม (Vote Ensemble)



Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol
1151	Non-Diabetic	Diabetic	1	0	53	Male	70	160	68	134	65	No	No
16547	Non-Diabetic	Non-Diabetic	0	1	43	Female	83	161	78	132	80	No	No
24245	Non-Diabetic	Diabetic	1	0	50	Male	80	163	90	131	82	No	No
1150	Non-Diabetic	Non-Diabetic	0	1	28	Male	55	170	81	137	96	No	Yes
19113	Non-Diabetic	Non-Diabetic	0	1	29	Female	45	157	76	97	91	No	Yes
3716	Non-Diabetic	Non-Diabetic	0	1	52	Female	61	149	84	136	86	No	No
24244	Non-Diabetic	Diabetic	0.750	0.250	52	Female	58	150	84	138	77	No	No
16546	Non-Diabetic	Non-Diabetic	0	1	53	Male	53	165	63	98	66	No	No
3715	Non-Diabetic	Non-Diabetic	0	1	49	Male	56	162	79	117	69	No	No
19112	Non-Diabetic	Non-Diabetic	0	1	47	Female	65	155	73	120	75	No	No
16545	Non-Diabetic	Non-Diabetic	0	1	59	Female	70	157	56	130	74	No	No
3714	Non-Diabetic	Non-Diabetic	0	1	26	Male	64	176	87	127	68	Yes	Yes
1149	Non-Diabetic	Non-Diabetic	0	1	33	Female	40	150	64	120	107	No	No
19111	Non-Diabetic	Non-Diabetic	0	1	30	Male	103	178	84	131	82	No	No
11415	Non-Diabetic	Non-Diabetic	0	1	35	Male	84	176	77	123	71	Yes	Yes
13981	Non-Diabetic	Non-Diabetic	0	1	60	Male	55	169	86	136	73	No	No
24243	Non-Diabetic	Non-Diabetic	0	1	57	Female	52	160	74	127	86	No	No
11414	Non-Diabetic	Non-Diabetic	0	1	59	Male	76	165	70	110	63	Yes	No
6282	Non-Diabetic	Non-Diabetic	0	1	33	Female	60	157	77	106	90	No	No
24242	Non-Diabetic	Non-Diabetic	0	1	58	Female	51	155	59	102	79	No	No
16544	Non-Diabetic	Non-Diabetic	0	1	30	Male	65	170	89	136	87	No	Yes
24241	Non-Diabetic	Non-Diabetic	0	1	36	Male	67	175	66	109	59	No	No
24240	Non-Diabetic	Non-Diabetic	0	1	43	Male	87	170	74	125	79	Yes	No
8848	Non-Diabetic	Non-Diabetic	0.250	0.750	61	Male	64	159	88	167	92	No	No

ExampleSet (25,660 examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 30 ชุดข้อมูล Unseen data ของด้วยเทคนิคการโหวตร่วม (Vote Ensemble)

4.4.5 ผลลัพธ์เทคนิคป่าสุ่ม (Random Forest)

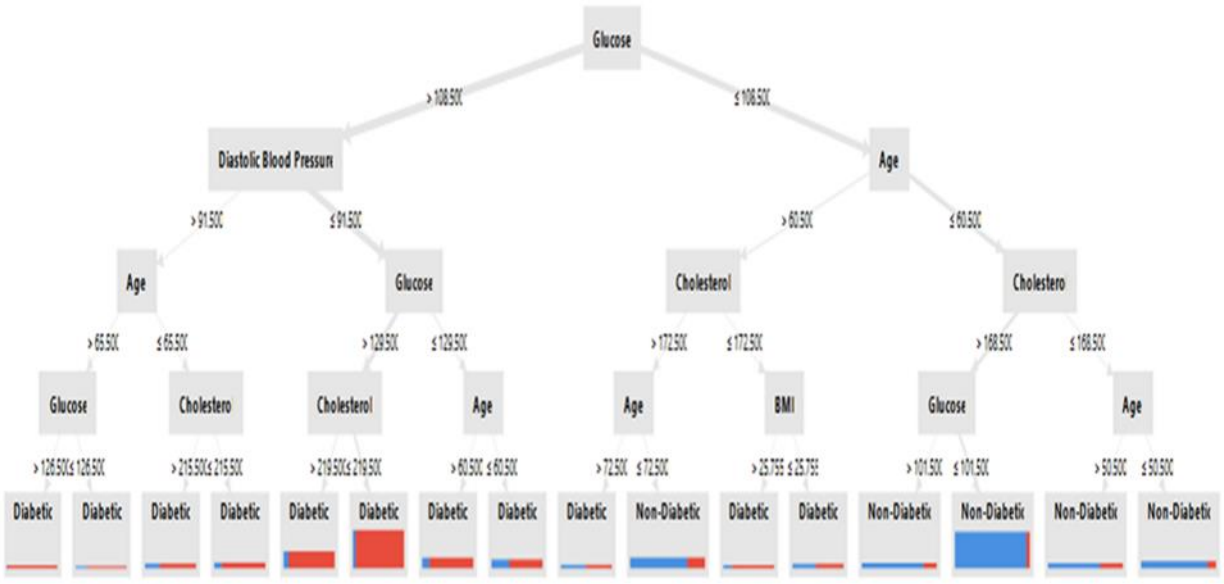
เทคนิคป่าสุ่มจะใช้เทคนิคต้นไม้การตัดสินใจเป็นพื้นฐานการจำแนกข้อมูลเป็นต้นไม้แต่ละต้นถูกสร้างขึ้นจากตัวอย่างบูตสแตรปจากชุดข้อมูลดั้งเดิมเพื่อคำนวนผลโหวตจากผลลัพธ์เดิมเพื่อหาคลาสที่ถูกโหวตมากที่สุด ในการทดลองนี้ได้ปรับพารามิเตอร์ในเทคนิคโดยให้มีรูปแบบของต้นไม้การตัดสินใจอยู่ 250 แบบ กำหนดเกณฑ์การจำแนกเป็น Information Gain และความลึกสูงสุดที่ 5 ทำ

ให้เกิด Unseen data และรูปแบบการเกิดผลลัพธ์การทำนายใหม่เป็นโรคเบาหวานแบบแผนผังต้นไม้ 250 แบบ โดยผู้ป่วยที่ทำนายเป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1 ผู้ป่วยที่ทำนายไม่เป็นโรคเบาหวานมีค่าความเชื่อมั่นสูงสุดคือ 1

Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol
1	Non-Diabetic	Non-Diabetic	0.466	0.534	72	Female	46	150	65	118	78	No	No
2	Non-Diabetic	Diabetic	0.838	0.162	38	Female	60	150	79	128	87	No	No
3	Non-Diabetic	Non-Diabetic	0.496	0.504	61	Male	86	167	83	137	62	No	No
4	Non-Diabetic	Non-Diabetic	0.104	0.896	35	Female	45	158	66	118	93	No	No
5	Non-Diabetic	Non-Diabetic	0.184	0.816	53	Female	62	165	65	119	77	No	No
6	Non-Diabetic	Non-Diabetic	0.363	0.637	67	Female	49	145	61	122	69	No	No
7	Non-Diabetic	Non-Diabetic	0.271	0.729	64	Female	53	150	77	129	78	No	No
8	Non-Diabetic	Non-Diabetic	0.171	0.829	53	Female	64	152	91	120	78	No	No
9	Non-Diabetic	Non-Diabetic	0.328	0.672	63	Female	80	156	81	132	81	No	No
10	Non-Diabetic	Non-Diabetic	0.155	0.845	49	Female	67	160	78	124	66	No	No
11	Non-Diabetic	Non-Diabetic	0.160	0.840	35	Female	52	154	71	117	103	No	No
12	Non-Diabetic	Diabetic	0.537	0.463	74	Male	65	174	87	166	78	No	No
13	Non-Diabetic	Non-Diabetic	0.108	0.892	50	Female	55	160	64	112	76	No	No
14	Non-Diabetic	Non-Diabetic	0.325	0.675	65	Female	72	160	72	130	64	No	No
15	Non-Diabetic	Non-Diabetic	0.107	0.893	34	Female	63	165	79	107	80	No	No
16	Non-Diabetic	Diabetic	0.710	0.290	56	Female	82	150	80	136	95	No	No
17	Non-Diabetic	Non-Diabetic	0.156	0.844	51	Female	57	153	74	138	88	No	No
18	Non-Diabetic	Non-Diabetic	0.109	0.891	49	Female	60	156	53	117	77	No	No
19	Non-Diabetic	Diabetic	0.584	0.416	55	Male	74	165	79	117	93	No	No
20	Non-Diabetic	Non-Diabetic	0.258	0.742	53	Female	57	149	94	157	75	No	No
21	Non-Diabetic	Non-Diabetic	0.100	0.900	30	Female	37	155	74	110	84	No	No
22	Non-Diabetic	Diabetic	0.732	0.268	55	Female	78	165	85	114	80	No	No
23	Non-Diabetic	Non-Diabetic	0.314	0.686	62	Female	61	155	80	145	80	No	No
24	Non-Diabetic	Non-Diabetic	0.319	0.681	76	Male	70	165	67	136	75	No	Yes

ExampleSet (25,660 examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 31 ชุดข้อมูล Unseen data ของด้วยเทคนิคป่าสุ่ม (Random Forest)



ภาพประกอบ 32 หนึ่งในแผนผังต้นไม้การตัดสินใจจากเทคนิคป่าสุ่ม (Random Forest)

4.5 ผลการวิเคราะห์ในขั้นตอนการประเมินผล (Evaluation)

ผู้วิจัยได้ประเมินผลแบบจำลองที่ได้จากการทำเหมืองข้อมูลเพื่อพิจารณาถึงการนำแบบจำลองไปประยุกต์ใช้กับโรงพยาบาลศูนย์อุดรธานีโดยคำนึงถึงความแม่นยำนั้นมีมากน้อยเพียงใดขึ้นอยู่กับลักษณะการทำเหมืองข้อมูลนั้น ๆ

ตาราง 16 เทคนิคต้นไม้การตัดสินใจ (Decision Tree)

Accuracy: 87.39% +/-0.65% (micro average: 87.39%)			
	True ไม่เป็นเบาหวาน	**True เป็นเบาหวาน**	**Class precision**
Pred. ไม่เป็นเบาหวาน	9723	1408	87.35%
Pred. เป็นเบาหวาน	1849	12855	87.43%
Class recall	84.02%	90.13%	

ตาราง 16 คลาสเป้าหมายคือไม่เป็นเบาหวานทำนายถูกเป็น 9723 ทำนายผิดพลาดเป็นโรคเบาหวานเป็น 1408 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.39% ค่าความแม่นยำ (Precision) เป็น 87.44% ค่าความครบถ้วน (Recall) เป็น 90.13% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 88.76%

คลาสเป้าหมายคือเป็นโรคเบาหวานทำนายถูกเป็น 12855 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 1849 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.39% ค่าความแม่นยำ (Precision) เป็น 87.44% ค่าความครบถ้วน (Recall) เป็น 90.13% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 88.76%

ตาราง 17 เทคนิคนาอิฟเบย์ (Naïve Bay)

Accuracy: 71.41% +/-5.52% (micro average: 71.41%)			
	True ไม่เป็นเบาหวาน	**True เป็นเบาหวาน**	**Class precision**
Pred. ไม่เป็นเบาหวาน	4846	660	88.01%
Pred. เป็นเบาหวาน	6726	13603	66.91%
Class recall	41.88%	95.37%	

ตาราง 17 คลาสเป้าหมายคือไม่เป็นเบาหวานทำนายถูกเป็น 4846 ทำนายผิดพลาดเป็นโรคเบาหวานเป็น 660 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 71.41% ค่าตัวแปรในการทำนายถูก (Precision) เป็น 67.60% ค่าจำนวนการกระทำด้วยกันแบบที่ตรงกับความเป็นจริง (Recall) เป็น 95.37% ค่า f-measure คือค่าเปรียบเทียบระหว่างค่า Precision กับ Recall เป็น 78.82%

คลาสเป้าหมายคือเป็นเบาหวานทำนายถูกเป็น 13603 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 6726 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 71.41% ค่าความแม่นยำ (Precision) เป็น 67.60% ค่าความครบถ้วน (Recall) เป็น 95.37% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 78.82%

ตาราง 18 เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbor (k-NN))

Accuracy: 87.82% +/-0.51% (micro average: 87.82%)			
	True ไม่เป็นเบาหวาน	**True เป็นเบาหวาน**	**Class precision**
Pred. ไม่เป็นเบาหวาน	9898	1472	87.05%
Pred. เป็นเบาหวาน	1674	12791	88.43%
Class recall	85.53%	89.68%	

ตาราง 18 คลาสเป้าหมายคือไม่เป็นเบาหวานทำนายถูกเป็น 9898 ทำนายผิดพลาดเป็นโรคเบาหวานเป็น 1472 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.82% ค่าความแม่นยำ (Precision) เป็น 88.43% ค่าความครบถ้วน (Recall) เป็น 89.68% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.05%

คลาสเป้าหมายคือเป็นเบาหวานทำนายถูกเป็น 12791 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 1674 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.82% ค่าความแม่นยำ (Precision) เป็น 88.43% ค่าความครบถ้วน (Recall) เป็น 89.68% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.05%

ตาราง 19 เทคนิคการโหวตร่วม (Vote Ensemble)

Accuracy: 87.94% +/- 0.45% (micro average: 87.94%)			
	True ไม่เป็นเบาหวาน	**True เป็นเบาหวาน**	**Class precision**
Pred. ไม่เป็นเบาหวาน	10074	1662	85.84%
Pred. เป็นเบาหวาน	1433	12491	89.71%
Class recall	87.55%	88.26%	

ตาราง 19 คลาสเป้าหมายคือไม่เป็นเบาหวานทำนายถูกเป็น 10074 ทำนายผิดพลาดเป็นโรคเบาหวานเป็น 1662 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.94% ค่าความแม่นยำ (Precision) เป็น 85.84% ค่าความครบถ้วน (Recall) เป็น 87.55% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 88.98%

คลาสเป้าหมายคือเป็นเบาหวานทำนายถูกเป็น 12491 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 1433 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 87.94% ค่าความแม่นยำ (Precision) เป็น 89.71% ค่าความครบถ้วน (Recall) เป็น 88.26% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 88.98%

ตาราง 20 เทคนิคป่าสุ่ม (Random Forest)

Accuracy: 88.03% +/- 0.51% (micro average: 88.03%)			
	True ไม่เป็นเบาหวาน	**True เป็นเบาหวาน**	**Class precision**
Pred. ไม่เป็นเบาหวาน	9799	1364	87.78%
Pred. เป็นเบาหวาน	1708	12789	88.22%
Class recall	85.16%	90.36%	

ตาราง 20 คลาสเป้าหมายคือไม่เป็นเบาหวานทำนายถูกเป็น 9799 ทำนายผิดพลาดเป็นโรคเบาหวานเป็น 1364 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 88.03% ค่าความแม่นยำ (Precision) เป็น 87.78% ค่าความครบถ้วน (Recall) เป็น 85.16% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.28%

คลาสเป้าหมายคือเป็นเบาหวานทำนายถูกเป็น 12789 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 1708 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 88.03% ค่าความแม่นยำ (Precision) เป็น 88.22% ค่าความครบถ้วน (Recall) เป็น 90.36% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.28%

4.5 ผลการวิเคราะห์ในขั้นตอนการนำไปใช้งาน (Deployment)

นำแบบจำลองที่ใช้ในการพยากรณ์ผู้ป่วยโรคเบาหวานของโรงพยาบาลศูนย์อุดรธานีไปใช้ในการประกอบการวางแผนบุคลากร อุปกรณ์การแพทย์ และวางแผนการรักษา เพื่อให้สอดคล้องต่อจำนวนผู้ป่วยในอนาคตที่มีการเป็นโรคนี้มากขึ้น รวมถึงลดระดับความเสี่ยงต่อการเสียชีวิต ป้องกันเฝ้าระวังแก้ไขเพื่อลดอัตราการเกิดโรคในอนาคต

บทที่ 5
การสรุป อภิปรายผล และข้อเสนอแนะ

การวิจัยเรื่องการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจโดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา : โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานี สามารถสรุปผลและอภิปรายผลการวิจัยแบ่งออกเป็นหัวข้อดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการทำวิจัย

5.3 ข้อเสนอแนะ

5.4 การนำแบบจำลองไปใช้กับการจัดทำนโยบายสำหรับรองรับผู้ป่วยที่จะเกิดขึ้นในอนาคต

5.1 สรุปผลการวิจัย

จากการศึกษาการเก็บข้อมูลในฐานข้อมูลผู้ป่วยโรงพยาบาลมหาวิทยาลัยมหาสารคาม ตั้งแต่ปี พ.ศ. 2558-2561 จากนั้นจึงประมวลผลตามขั้นตอน CRISP-DM

5.1.1 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าธุรกิจ (Business Understanding)

เป้าหมายในงานวิจัยครั้งนี้คือสร้างและเปรียบแบบจำลองเพื่อหาแบบจำลองที่ดีที่สุดสำหรับการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจ โรคเบาหวาน และโรคความดันโลหิตสูง โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา โรงพยาบาลศูนย์อุดรธานี จังหวัดอุดรธานีและเพิ่มประสิทธิภาพและการบริการแก่ผู้ป่วยมีสุขภาพที่ดีต่อไป

5.1.2 ผลการวิเคราะห์ในขั้นตอนการทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในการวิจัยได้มาจากการตรวจของคลินิกผู้ป่วยนอกหรือ OPD (Out Patient Department) และและข้อมูลการตรวจสุขภาพผู้เข้ารับประกันสังคม ประกอบไปด้วย รหัสผู้ป่วย จำนวนครั้งในการเข้าตรวจ วันเข้ารับการตรวจ อายุ เพศ สัญญาณชีพ น้ำหนัก ส่วนสูง ระดับค่าล่างความดันเลือด ระดับค่าบนความดันเลือด อุณหภูมิ สัญญาณชีพจร การหายใจ การรักษา การสูบบุหรี่ การดื่มแอลกอฮอล์ สัดส่วนรอบเอว ค่าดัชนีมวลกาย (BMI) ระดับคอลเลสเตอรอล และระดับกลูโคลส โดยข้อมูลชุดนี้ทำการเก็บข้อมูลตั้งแต่ปี พ.ศ. 2558-2564

5.1.3 ผลการวิเคราะห์ในขั้นตอนการเตรียมข้อมูล (Data Preparation)

ในเตรียมข้อมูลคลินิกผู้ป่วยนอกหรือ OPD (Out Patient Department) และข้อมูลผู้เข้ารับการตรวจสุขภาพตามสิทธิประกันสังคมเพื่อใช้ในการสร้างแบบจำลองการพยากรณ์การพยากรณ์ผู้ป่วย โรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล จะต้องเตรียมข้อมูลให้มีความถูกต้องและสามารถประมวลผลได้อย่างแม่นยำ โดยข้อมูลที่สำคัญที่สุดในการสร้างแบบจำลองคือตัวแปรคือแอททริบิวท์การสูบบุหรี่และการดื่มแอลกอฮอล์ ซึ่งทำหน้าที่เป็นตัวแปรต้น ส่วนแอททริบิวท์การเข้ารักษาคลินิกตรวจโรคเป็นตัวแปรตาม และมีการตัดข้อมูลที่ไม่จำเป็นต่อการสร้างแบบจำลอง ได้แก่ จำนวนครั้งที่เข้ารักษา (RegNo) อุณหภูมิ (Temperatures) การรักษา (Treatment) สัดส่วนรอบเอว (WHR) การหายใจ (Breath) ข้อมูลถูกจัดเก็บอยู่ในรูปแบบ Microsoft Excel 2016 เพื่อเป็นข้อมูลการสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจในโรงพยาบาลศูนย์อุดรธานี

5.1.4 ผลการวิเคราะห์ในขั้นตอนการสร้างแบบจำลอง (Modeling)

การสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวาน โรงพยาบาลศูนย์อุดรธานีในการพยากรณ์ผู้ป่วยทั้ง 3 โรค โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา โรงพยาบาลศูนย์อุดรธานี โดยวิธีการจำแนกข้อมูล (Classification) และเทคนิคที่นำมาใช้ได้แก่ เทคนิคการจำแนกข้อมูลด้วยเทคนิคต้นไม้การตัดสินใจ (Decision Tree) เทคนิคการจำแนกข้อมูลด้วยวิธีนาอีฟเบย์ (Naive Bay) เทคนิคการจำแนกข้อมูลด้วยเทคนิคเพื่อนใกล้บ้านมากที่สุด (k-NN: k-Nearest Neighbor) และเทคนิคการโหวตร่วม (Vote Ensemble) และเทคนิคป่าสุ่ม (Random Forest)

5.1.5 ผลการวิเคราะห์ในขั้นตอนการประเมินผล (Evaluation)

เมื่อทำการสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจ โรคเบาหวาน โรคความดันโลหิตสูง โรงพยาบาลศูนย์อุดรธานี เพื่อเปรียบเทียบหาค่าการทดสอบประสิทธิภาพของการจำแนกข้อมูล ได้แก่การวัดค่าความถูกต้อง (Accuracy) ค่าวัดความถูกต้องของการพยากรณ์ (Precision) ค่าความครบถ้วน (Recall) และ ค่าวัดประสิทธิภาพโดยรวม (F-Measure)

ตาราง 21 เปรียบเทียบค่าทดสอบประสิทธิภาพของการจำแนกข้อมูล

เทคนิคสำหรับการจำแนกประเภทข้อมูล	**ค่าทดสอบประสิทธิภาพ**			
	Accuracy	Precision	Recall	F-Measure
Random Forest	88.03%	88.22%	90.36%	89.28%
Vote Ensemble	87.94%	87.71%	88.26%	88.98%
k-NN	87.82%	88.43%	89.68%	89.05%
Decision Tree	87.39%	87.44%	90.13%	88.76%
Naïve Bayes	71.41%	67.60%	95.37%	78.82%

ตาราง 21 เปรียบเทียบค่าทดสอบประสิทธิภาพของการจำแนกข้อมูลโรคความเบาหวาน เทคนิคที่ผลลัพธ์การทดสอบประสิทธิภาพที่ดีที่สุดคือการจำแนกข้อมูลด้วยเทคนิคป่าสุ่ม (Random Forest) โดยให้ค่าความถูกต้อง (Accuracy) เป็น 88.03% ค่าความแม่นยำ (Precision) เป็น 88.82% ค่าความครบถ้วน (Recall) เป็น 90.36% และค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.28% ส่วนเทคนิคที่ให้ผลลัพธ์การทดสอบประสิทธิภาพรองลงมาคือเทคนิคเทคนิคการจำแนกข้อมูลด้วยเทคนิคการจำแนกข้อมูลด้วยเทคนิคเทคนิคการโหวตร่วม (Vote Ensemble) เทคนิคเพื่อนใกล้บ้านมากที่สุด (k-NN: k-Nearest Neighbor) เทคนิคต้นไม้การตัดสินใจ (Decision Tree) และเทคนิคนาอิฟ เบย์ (Naïve Bayes) ตามลำดับ

5.1.6 ผลการวิเคราะห์ในขั้นตอนการนำไปใช้งาน (Deployment)

นำแบบจำลองที่ใช้ในการพยากรณ์ผู้ป่วยโรคเบาหวาน โรงพยาบาลศูนย์อุดรธานี ไปใช้ในการสนับสนุนการตัดสินใจในการรักษาผู้ป่วยโรคเบาหวานได้เพื่อให้สอดคล้องต่อจำนวนผู้ป่วยในอนาคตที่มีการเป็นโรคนี้มากขึ้น

5.2 อภิปรายผลวิจัย

ผลการสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคหลอดเลือดหัวใจ โรงพยาบาลศูนย์อุดรธานี จากข้อมูลการตรวจของคลินิกผู้ป่วยนอกหรือ OPD (Out Patient Department) และและข้อมูลผู้เข้ารับการตรวจสุขภาพตามสิทธิประกันสังคมมาสร้างแบบจำลองการจำแนกประเภทข้อมูล (Classification) โดยใช้เทคนิคการจำแนกข้อมูลด้วยวิธีการเพื่อนใกล้บ้านมากที่สุด (k-NN: k-Nearest Neighbor) และ เทคนิคการจำแนกข้อมูลด้วยเทคนิคต้นไม้การตัดสินใจ (Decision Tree) และเทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคการโหวตร่วม (Vote Ensemble) และเทคนิคป่าสุ่ม (Random Forest) พบว่าเทคนิคการจำแนกข้อมูลด้วยเทคนิคป่าสุ่ม (Random Forest) ให้ผลลัพธ์ที่ดีที่สุดโดยทำนายคลาสเป้าหมายคือเป็นโรคเบาหวานทำนายถูกเป็น 12789 ทำนายผิดพลาดเป็นไม่เป็นโรคเบาหวานเป็น 1708 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 88.03% ค่าความแม่นยำ (Precision) เป็น 88.22% ค่าความครบถ้วน (Recall) เป็น 90.36% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.28% หากคลาสเป้าหมายคือไม่เป็นโรคเบาหวานทำนายถูกเป็น 9799 ทำนายผิดพลาดเป็นเป็นโรคเบาหวานเป็น 1364 การทดสอบประสิทธิภาพเทคนิคให้ค่าความถูกต้อง (Accuracy) เป็น 88.03% ค่าความแม่นยำ (Precision) เป็น 88.22% ค่าความครบถ้วน (Recall) เป็น 90.36% ค่าวัดประสิทธิภาพโดยรวม (F-Measure) เป็น 89.28% ดังนั้นเทคนิคที่ให้ผลลัพธ์ที่ดีที่สุดเทคนิคการจำแนกข้อมูลด้วยเทคนิคป่าสุ่ม (Random Forest) เพราะมีค่าการทดสอบประสิทธิภาพที่ดีและน่าเชื่อถือมากในบรรดาเทคนิคที่กล่าวมา แบบจำลองการพยากรณ์ที่ดีที่สุดคือเทคนิคการจำแนกข้อมูลด้วยเทคนิคเทคนิคป่าสุ่ม (Random Forest)

5.3 ข้อเสนอแนะ

5.3.1 ในการบันทึกข้อมูล พยาบาลมักจะบันทึกลงกระดาษเสียส่วนใหญ่อันเนื่องมาจากพยาบาลส่วนใหญ่มีอายุประมาณ 40 ถึง 50 กว่าปีขึ้นไปไม่มีความชำนาญในการใช้ระบบบันทึกการตรวจโรคจึงนิยมใช้การจดบันทึกที่ถนัดกว่า ข้อมูลจึงไม่มีตรงตามเวลาจริง จึงอยากให้พยาบาลใส่ใจกับการฝึกฝนในการพิมพ์ให้มากขึ้น

5.3.2 ข้อมูลที่ได้มามีความผิดพลาดอันเนื่องมาจากความเร่งรีบในการดูแลผู้ป่วยจำนวนมาก อีกทั้งพยาบาลที่ทำการตรวจโรคนั้นนี้อายุพอสมควร จึงอยากให้ทางโรงพยาบาลพิจารณาในเรื่องการจัดบุคลากรที่มีประสิทธิภาพ สามารถทำการตรวจและใช้ระบบบันทึกการตรวจโรคได้พร้อมกัน เพื่อจะได้สามารถบันทึกข้อมูลได้แม่นยำขึ้น

5.3.3 ในการจัดเก็บข้อมูล ข้อมูลที่เป็นปัจจัยส่งผลต่อโรคบางส่วนไม่ได้จัดเก็บไว้ เช่น การออกกำลังกาย อาชีพ ความเครียด พันธุกรรม การตรวจปัสสาวะ การหายใจ ฯลฯ จึงอยากจะให้มีการพัฒนาการจัดเก็บข้อมูลให้ยิ่งขึ้นสำหรับผู้วิจัยที่ต้องการข้อมูลดังกล่าวสามารถทำการวิจัยได้ดียิ่งขึ้น

5.3.4 เทคนิคเพื่อนใกล้บ้านมากที่สุด (k-NN: k-Nearest Neighbor) และเทคนิคการโหวตร่วม (Vote Ensemble) นั้นสามารถพัฒนาประสิทธิภาพของตัวเทคนิคเพื่อเพิ่มความแม่นยำได้แต่ข้อมูลมีขนาดใหญ่ใช้เวลาประมวลผลนานพอสมควรจึงไม่ได้เพิ่มประสิทธิภาพในการทดลองครั้งนี้ การวิจัยครั้งหน้าควรจัดหาคอมพิวเตอร์ที่ใช้ในการทำเหมืองข้อมูลที่มีประสิทธิภาพสูงเพื่อที่จะได้ผลลัพธ์ที่ดีขึ้นในการวิจัยในอนคต

5.3.4 แม้เทคนิคต้นไม้การตัดสินใจ (Decision Tree) จะให้ผลลัพธ์น้อยกว่าเทคนิคเพื่อนใกล้บ้านมากที่สุด แต่เป็นเทคนิคให้ความเข้าใจแก่ผู้ได้ดีในระดับหนึ่ง ด้วยผลลัพธ์ดังกล่าวสามารถใช้เป็นแนวทางในการรักษาสุขภาพโดยการรักษาระดับความดันโลหิตให้อยู่ในช่วง < 120 และ < 80 มม. ปรอท การรักษาค่าคอเลสเตอรอล (Cholesterol) ให้อยู่ในระดับ 91 – 120 มิลลิกรัมต่อเดซิลิตร การรักษาระดับกลูโคส (Glucose) 71 – 108 มิลลิกรัมต่อเดซิลิตร การรักษาการเต้นของชีพจร (Pulse) ให้อยู่ในช่วง 61 – 65 ครั้งต่อนาที เพื่อให้ผู้ป่วยมีสุขภาพที่ดีขึ้นและลดค่าใช้จ่ายในการรักษาอีกทั้งยังเป็นภาระต่อโรงพยาบาลต่อไป

บรรณานุกรม

[1] กระทรวงสาธาณสุข กรมควบคุมโรค. "กรมควบคุมโรค เตือนประชาชนใส่ใจดูแลสุขภาพตนเองและคนในครอบครัว ระวังป่วยโรคเบาหวาน เผยจากการสำรวจสุขภาพประชาชนไทยอายุ 15 ปีขึ้นไป พบผู้ป่วยเบาหวาน 4.8 ล้านคน." [ออนไลน์] 2563. [สืบค้นเมื่อ 20 ธันวาคม 2563] ; https://pr.moph.go.th/?url=pr/detail/2/02/134288.

[2] ระบบสารสนเทศภูมิศาสตร์และทรัพยากรสุขภาพ. "ประวัติสถานพยาบาล โรงพยาบาลอุดรธานี." [ออนไลน์] [สืบค้นเมื่อ 15 ธันวาคม 2563)]; http://gishealth.moph.go.th/healthmap/info_history.php?maincode=10671.

[3] Hossam A. Shouip. "Diabetes mellitus." [ออนไลน์] 2014. [สืบค้นเมื่อ 15 ธันวาคม 2020]; https://www.researchgate.net/publication/270283336_Diabetes_mellitus

[4] Bharati Mahadev Ramageri, "Data mining techniques and applications". Indian Journal of Computer Science and Engineering. 2010; 1(4): 301-305.

[5] Rüdiger Wirth, Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, vol. 1: Springer-Verlag London, UK.

[6] อัชฌา พรกว้างสวาสดิ์, เพียงฤทัย หนูสวัสดิ์, วราลี คงเหมาะ, ปวีณา ทิพยากุลรักษ์, และบุษกร สังขนันท์. ระบบทำนายระดับความครียดด้วยเทคนิคต้นไม้ การตัดสินใจ. Rattanakosin Journal of Science and Technology 2019; 1(2): 13-26.

[7] วิชญ์วิสิฐ เกษรสิทธิ์, จิราวัลย์ จิตรถเวช และ วิชิต หล่อจีระชุณห์กุล, การลดจำนวนกลุ่มในการจำแนกแบบหลายกลุ่มเป็นสองกลุ่มสำหรับการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน. Thai Science and Technology Journal 2563; 28(1): 41-51.

[8] ภูมิพัฒน์ ดวงกลาง, รัจนา เครือแก้ว, แบบจำลองการทำนาย แบบอากาศยานจาก ข้อมูลเป้าหมายไม่ทราบฝ่ายอัตโนมัติ. NKRAFA JOURNAL OF SCIENCE AND TECHNOLOGY 2562, 15: 1-8.

[9] Charu C. Aggarwal. "Data mining: the textbook. Springer". 1st edition. Switzerland: Springer International Publishing; 2015.

[10] อับดุลเลาะ บากา, อรรถพล อดุลยศาสน์, อิสมาแอ ล่าเตะเกะ, สุลัยมาน เภอโส๊ะ, จีรวุธ มุนินทร์นพมาศ, อิมรอน แวมง, มูฮัมหมัด ปุ, "แบบจำลองพยากรณ์ผลการเรียนของนักศึกษา

จากพฤติกรรมการใช้งานอินเทอร์เน็ตโดยใช้เทคนิคการทำเหมืองข้อมูลกรณีศึกษามหาวิทยาลัยราชภัฏยะลา," การประชุมวิชาการระดับ ชาติเครือข่ายวิจัยสถาบันอุดมศึกษาทั่ว ประเทศ ครั้งที่ 11 “เครือข่ายวิจัยอุดมศึกษา สานพลังประชารัฐ” มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา, 2559; หน้า 187-196.

[11] จุฑาทิพย์ ทิพย์พูล, นิเวศ จิระวิชิตชัย, การจำแนกจดหมาย อิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคเหมืองข้อมูล, Progress in Applied Science and Technology 2559; 6(1): 102-109.

[12] ชลิตา เจริญเนตร, จารี ทองคำ, สิทธิชัย บุษหมั่น, การเปรียบเทียบเทคนิคเหมืองข้อมูลในการจำแนกใบหน้า, วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม 2558; 34(3): 263-269.

[13] วีระยุทธ พิมพาภรณ์, พยุง มีสัจ, เทคนิคเหมืองข้อมูลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการศึกษา ด้วยวิธีการจัดการเรียนการสอนแบบผสมผสาน, *Sripatum Review of Science and Technology* 2555; 4(1): 107-115.

[14] พัชรียา ทองพูล, พิมพ์ชนก จำเรือง, รมย์นลิน บุญฤทธิ์, สายชล สินสมบูรณ์ทอง, การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล, *Thai Journal of Science and Technology* 2562; 8(6): 565-584.

[15] จักรกฤษณ์ หงส์เวียงจันทร์, นิติมา ลักขณานุรักษ์, ไก้รุ่ง เฮงพระพรหม "การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลโรคออทิสติกด้วยเทคนิคเหมืองข้อมูล," งานประชุมวิชาการระดับชาติครั้งที่ 11 มหาวิทยาลัยราชภัฏนครปฐม, จังหวัดนครปฐม, 11 – 12 กรกฎาคม 2562, หน้า 321-326.

[16] จิตกานต์ จันทราชม, มนทิราลัย ชัยมงคล, รัตนชัย แซ่โง้ว, สายทิพย์ พลอยสัมฤทธิ์, สายชล สินสมบูรณ์ทอง, การเปรียบเทียบประสิทธิภาพการทำนายผลการจำแนกกรณีข้อมูลสูญหายด้วยเทคนิคการทำ เหมืองข้อมูล. Thai Journal of Science and Technology 2563; 9(1): 1-15.

[17] อิทธิพล ดวงแก้ว, สายยัญ สายยศ, การวิเคราะห์ปัจจัยที่มีผลต่อพัฒนาการล่าช้าตามช่วงอายุในเด็กปฐมวัยด้วยเทคนิคเหมืองข้อมูล, JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY 2562; 9(2): 44-55.

[18] ธนพัฒน์ ทองมา, "การพัฒนาโมเดลทำนายแผนการเรียนในการศึกษาต่อระดับมัธยมศึกษาตอนปลายของนักเรียนโรงเรียนสาธิตจุฬาลงกรณ์มหาวิทยาลัย ฝ่ายมัธยม: การประยุกต์ใช้เทคนิค

เอ็นเซ็มเบิ้ลโหวตร่วมกันระหว่างเครือข่ายใยประสาท ซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้ตัดสินใจ, วารสารครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย 2563; 48(3): 125-143.

[19] จิราภา เลาหะวรนันท์, รชต ลิ้มสุทธิวันภูมิ, บัณฑิต ฐานะโสภณ, พรฤดี เนติโสภากุล, การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือก แขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ, วารสารเทคโนโลยี สารสนเทศลาดกระบัง 2018; 4(2).

[20] นรินทร์ พนาวาส, THAI SENTIMENT ANALYSIS ON SOCIAL MEDIA USING MAJORITY VOTING-BASED ENSEMBLE METHOD, วารสารวาชาการศรีปทุม ชลบุรี 2561; 15(1): 51-67.

[21] Aminul Islam, Nusrat Jahan. Prediction of onset diabetes using machine learning techniques, International Journal of Computer Applications 2017; 180(5): 7-11.

[22] Austin Haynesworth. Usage of Electronic Health Record Phenotyping in American Adult Patients with Schizophrenia to Improve Detection of Type II Diabetes Mellitus. University of California, Los Angeles, 2020.

[23] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection, IEEE access 2018; 6: 33789-33795.

[24] Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen, Jin Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," IEEE access 2017; 5: 16568-16575.

[25] Jiangtao Ma, Yaqiong Qiao, Guangwu Hu, Yongzhong Huang, Arun Kumar sangaiah, CHAOQIN ZHANG, YANJUN WANG, RUI ZHANG. De-anonymizing social networks with random forest classifier, IEEE Access 2017; 6: 10139-10150.

[26] Adele Cutler, Richard D Cutler, John R. Stevens. Random forests in Ensemble Machine Learning: Springer, 2012; 157-175.

[27] David L. Olson, Dursun Delen. Advanced Data Mining Techniques. Springer Science & Business Media, 2008.

[28] Vanishri Arun, Arunkumar B V, Padma S K, Shyam V. Disease Classification and Prediction using Principal Component Analysis and Ensemble Classification Framework, International Journal of Control Theory and Applications 2017; vol.

10(14).

[29] Ramya Akula, Ni Nguyen, and Ivan Garibay. Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes, *SoutheastCon*, 2019: IEEE; 1-8.

[30] Kawsar Ahmed, Tasnuba Jesmin. Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using Weka Approach, International Journal of Science and Engineering 2014; 7(2): 155-160.

[31] D. Ashok Kumar, R. Govindasamy. Performance And Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies 2015; 6(2): 1312-1319.

[32] V. Karthikeyani, I. Parvin Begum, K. Tajudin, I. Shahina Begam. Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction. International Journal of Computer Applications 2012; 60(12).

[33] K. Saravananathan, T. Velmurugan.A Analyzing Diabetic Data using Classification Algorithms in Data Mining.Indian Journal of Science and Technology 2016; 9(43): 1-6.

[34] Ratna Patil, Sharavari Tamane. A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes. International Journal of Electrical and Computer Engineering 2018; 8(5): 3966.

[35] G. Visalatchi, S. Gnanasoundhari, M. Balamurugan, A Survey on Data Mining Methods and Techniques for Diabetes Mellitus, International Journal of Computer Science and Mobile Applications 2014; 2(2): 100-105.

[36] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput. An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set.
International Journal of Computer Applications 2015; 131(2): 6-11.

[37] S. Selvakumar, K. Senthamarai Kannan, S. GothaiNachiyar. Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques. International Journal of Statistics and Systems 2017; 12(2): 183-188.

[38] Sonu Bala Garg, Ajay Kumar Mahajan, T.S.Kamal. An Approach for Diabetes Detection using Data Mining Classification Techniques. International Journal of Engineering Sciences, 2017; 202-218.

[39] รุ่งโรจน์ บุญมา, นิเวศ จิระวิชิตชัย การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล และการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล, PKRU SciTech Journal 2562; 3(2): 11-19.

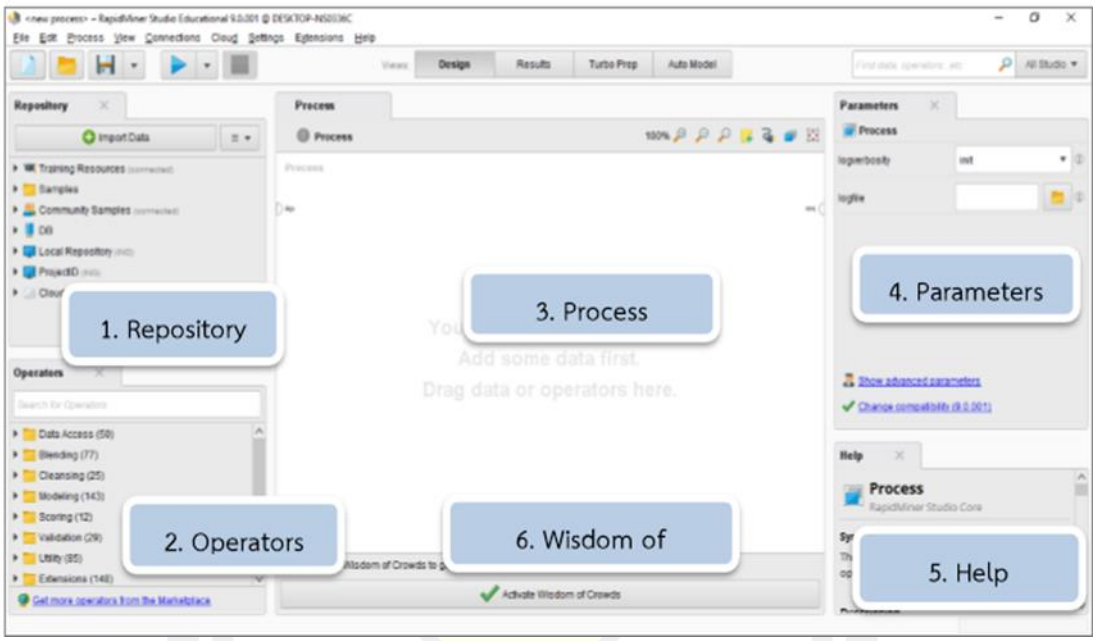
ภาคผนวก

ภาคผนวก ก
การใช้งานโปรแกรม RapidMiner Studio

การใช้งานโปรแกรม RapidMiner Studio

แรกเริ่มพัฒนาขึ้นจากบริษัทที่ชื่อว่า Rapid-I ในประเทศเยอรมนี และเมื่อช่วงปลายปี 2013 ที่ผ่านมาได้รับทุนก้อนโตจากนักลงทุนในประเทศสหรัฐอเมริกา จึงเปลี่ยนชื่อบริษัทจาก Rapid-I เป็น RapidMiner แทนและได้ย้ายสำนักงานใหญ่มาอยู่ประเทศสหรัฐอเมริกา

องค์ประกอบของหน้าต่าง Design ใน RapidMiner Studio



ภาพประกอบ 33 แสดงองค์ประกอบหลักของหน้าต่าง Design ในโปรแกรม RapidMiner Studio

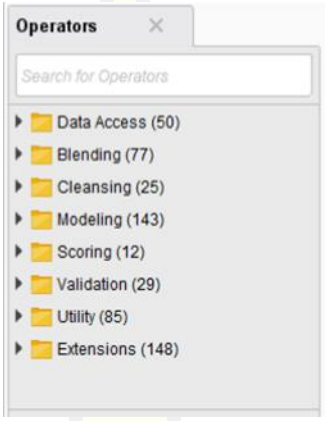
หน้าต่างโปรแกรม RapidMiner Studio ประกอบด้วย 6 ส่วนหลักๆ ได้แก่

1. Repository ส่วนนี้เป็นส่วนจัดการไฟล์ RapidMiner Studio จะจัดการข้อมูลจาก 3 แหล่ง คือ Data Base (ฐานข้อมูล), Local (ในเครื่องคอมพิวเตอร์ที่ใช้อยู่) และ Cloud Repository (ในคลาวด์) ในการบันทึกข้อมูลเราจะทำการเก็บไฟล์ Data Set, Process และ Model ต่างๆ แยก เก็บคนละโฟลเดอร์กัน

2. Operators เป็นส่วนที่ใช้เก็บ ตัวโอเปอร์เรเตอร์ ที่ใช้ในการท างานทั้งหมด ซึ่งจัดเป็นกลุ่มๆ โดยกลุ่มที่ใช้งานคล้ายคลึงกันจะจัดอยู่ในกลุ่มเดียวกัน มี 8 กลุ่ม โอเปอร์เรเตอร์ คือ

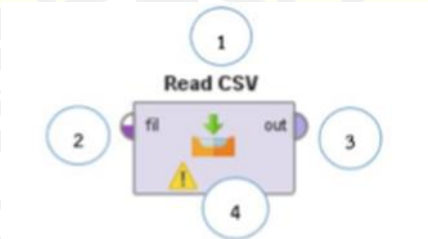
1. Data Access
2. Blending

3. Cleansing
4. Modeling
5. Scoring
6. Validation
7. Utility
8. Extentions



ภาพประกอบ 34 แสดงกลุ่มของ Operators

เช่น โอเปอร์เรเตอร์สำหรับการอ่านข้อมูลจากไฟล์ประเภท CSV จะอยู่ในหมวด Import และ หมวดย่อย Data นอกจากนี้ในส่วนของโอเปอร์เรเตอร์นี้ ยังมีส่วนในการค้นหาชื่อของโอเปอร์เรเตอร์ ต่างๆ ได้ โดยโอเปอร์เรเตอร์แต่ละตัวจะประกอบด้วย 4 ส่วน คือ



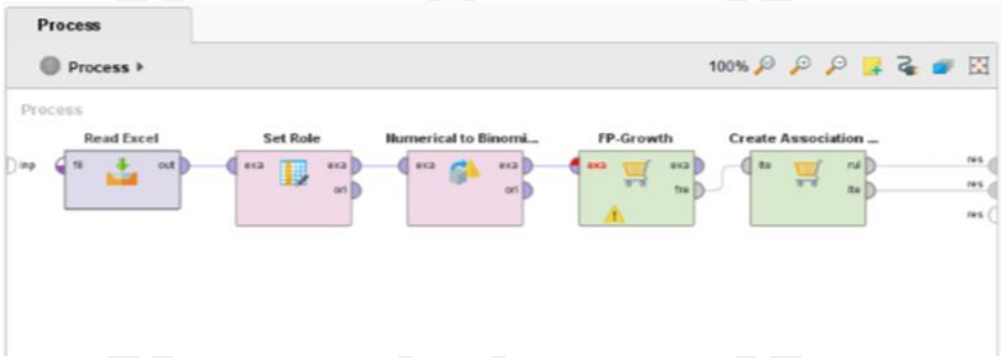
ภาพประกอบ 35 แสดงส่วนประกอบของ Operators Read CSV

1. Operators Name ชื่อของ Operators
2. Input port พอร์ตนำเข้าข้อมูล

3. Output port พอร์ตส่งออกข้อมูล

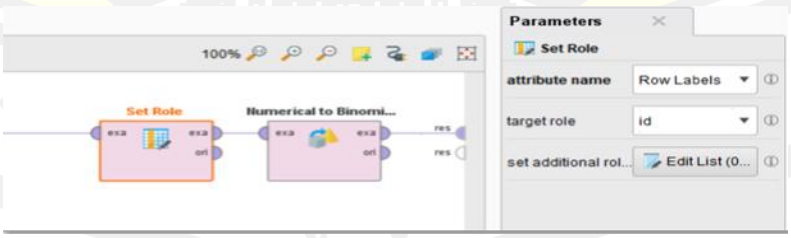
4. Operators Status สถานะของ Operators

3. Process เป็นหน้าหลักในการทำงานในการสร้างโพรเซสสำหรับทำ Machine Learning ของซอฟต์แวร์นี้ โดยจะนำโอเปอร์เรเตอร์มาประกอบเพื่อสร้างโพรเซสขึ้นมาตามวัตถุประสงค์ของโจทย์ที่ตั้งไว้



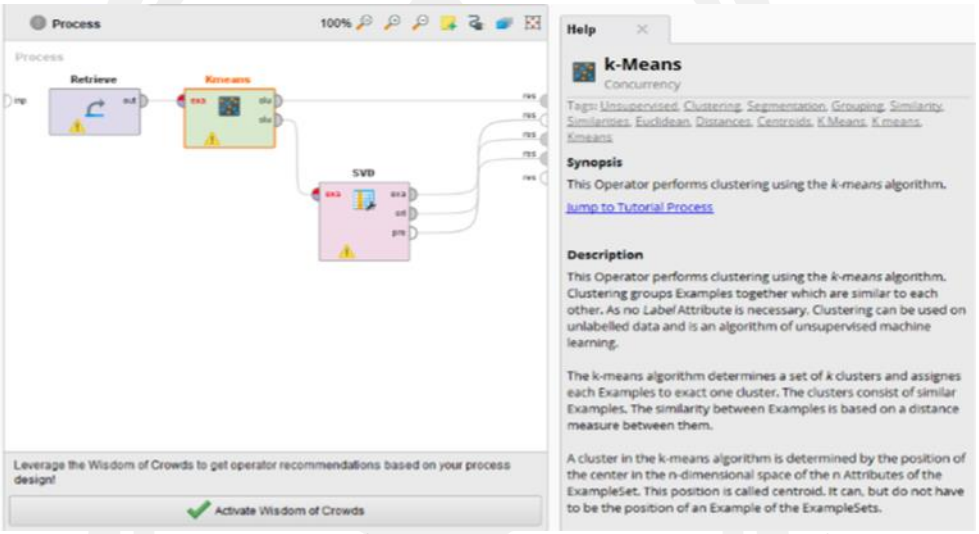
ภาพประกอบ 36 แสดงโพรเซสสำหรับการทำ *Machine Learning* ของโปรแกรม

4. Parameters เป็นส่วนของ Configuration, Option และกำหนดค่าพารามิเตอร์ที่เป็นรายละเอียดของ โอเปอร์เรเตอร์ที่เลือกใช้งาน เช่น โอเปอร์เรเตอร์ Set Role เป็น Operator ที่มีพารามิเตอร์ที่เกี่ยวข้องสองพารามิเตอร์คือ attribute name ซึ่งเป็นพารามิเตอร์ที่ใช้กำหนดเอาท์พุทของโพรเซสาจะเลือกอันใหน ซึ่งจากตัวอย่าง เลือก Row Labels เป็นเอาท์พุทของโพรเซส และพารามิเตอร์ target role เพื่อระบุว่าพารามิเตอร์เอาท์พุทเป็น id เท่านั้น



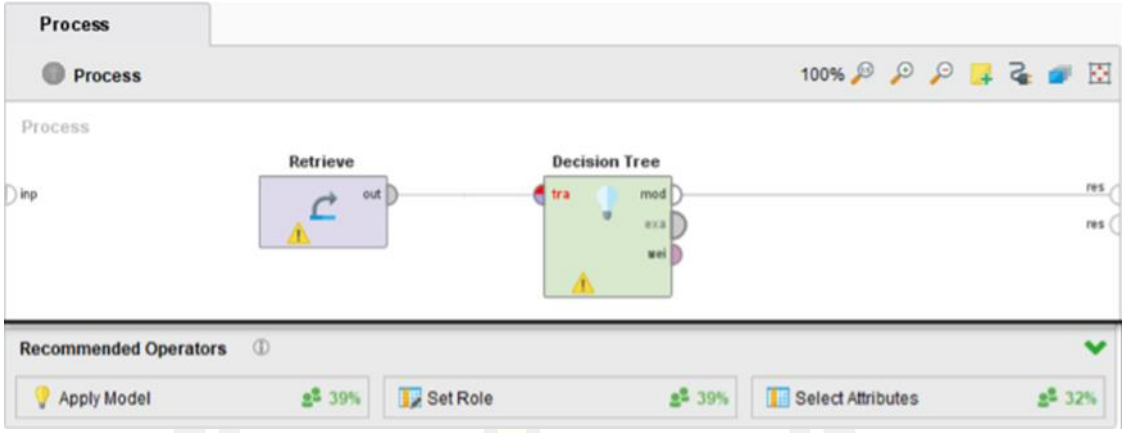
ภาพประกอบ 37 แสดงส่วนของ *Configuration, Option* และกำ หนดค่าพารามิเตอร์ที่เป็นรายละเอียดของโอเปอร์เรเตอร์ที่เลือกใช้งาน

5. Help เป็นส่วนช่วยเหลือซึ่งจะแสดงรายละเอียดของตัวโอเปอร์เรเตอร์ที่เลือกใช้งานอยู่ ส่วนช่วยเหลือของ RapidMiner Studio จะบอกเพียงหน้าที่และรายละเอียดคร่าวๆของโอเปอร์เรเตอร์ หากต้องการดูรายละเอียดมากกว่านี้ต้องไปที่ Jump to Tutorial Process ซึ่งจะ Link ไปยังเว็บไซต์ที่มีรายละเอียดเกี่ยวกับ Operator ที่ใช้อยู่ เช่น โอเปอร์เรเตอร์ชื่อ KMeans ในหน้า Help ก็จะบอกว่าเป็น โอเปอร์เรเตอร์ที่ใช้ในการจัดกลุ่มข้อมูล โดยใช้วิธี k-Means algorithm



ภาพประกอบ 38 แสดงส่วนช่วยเหลือซึ่งจะแสดงรายละเอียดของโอเปอร์เรเตอร์ที่เลือกใช้งานอยู่

6. Wisdom of Crowds เป็นส่วนที่ช่วยให้การสร้างเวิร์กไฟล์สำหรับการวิเคราะห์ในรูปแบบที่ ง่ายที่สุด กล่าวคือ โปรแกรมจะแนะนำเครื่องมือให้เหมาะสมสำหรับการวิเคราะห์นั้นๆ ซึ่งจะมีการแนะนำเครื่องมือที่เหมาะสมสำหรับขั้นตอนถัดไป โดยอาศัยการวิเคราะห์และการเรียนรู้ด้วยตัวเครื่อง เองจากประสบการณ์ของกลุ่มคนที่ใช้งาน RapidMiner ที่คล้ายกานมากที่สุด มาช่วยในการตัดสินใจ หมายเหตุ: จะใช้ส่วนนี้ได้จะต้องทำการ Activate Wisdom of Crowds



ภาพประกอบ 39 แสดงส่วนที่ช่วยให้การสร้างเวิร์กไฟล์สำหรับการวิเคราะห์ในรูปแบบที่ง่ายที่สุด

นอกจากส่วนประกอบของโปรแกรมทั้ง 6 ส่วนที่ได้อธิบายไปแล้วนั้น ยังมีส่วนเมนูด้านบนใต้เมนูบาร์เพิ่มเติมดังนี้

ภาพประกอบ 40 แสดงเมนูด้านบนใต้เมนูบาร์

1. เมนูสำหรับการสร้างโพรเซสใหม่
2. เมนูสำหรับการโหลดไฟล์ต่างๆ จาก repository
3. เมนูสำหรับการบันทึกโพรเซส
4. เมนูสำหรับสั่งให้โพรเซสทำงาน (run Process)
5. เมนูสำหรับยกเลิกการทำงานโพรเซส (stop)

ส่วนตรงกลางจะเป็นเมนูสำหรับเปลี่ยนหน้าจอ (perspective) ลักษณะต่างๆ ซึ่งแต่ละเมนู ทำหน้าที่ดังนี้

ภาพประกอบ 41 แสดงเมนูสำหรับเปลี่ยนหน้าจอ

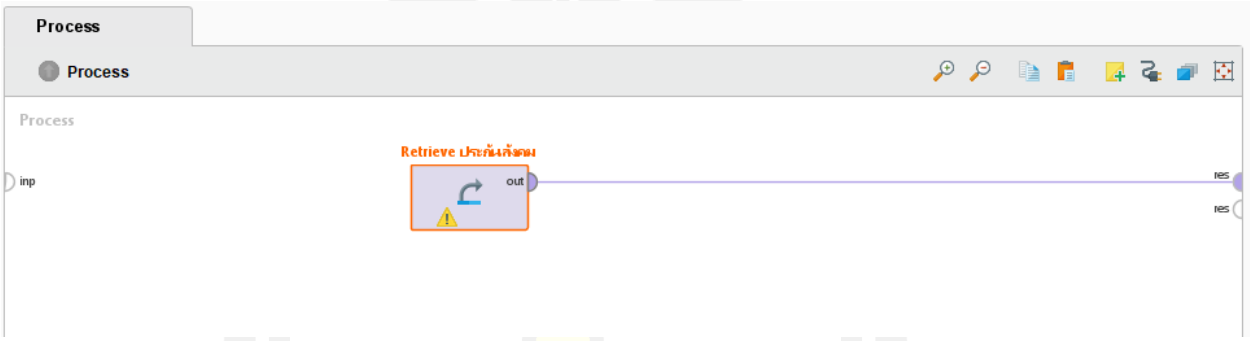
1. แสดงหน้าจอการออกแบบ
2. แสดงหน้าจอผลลัพธ์การทำงาน
3. แสดงหน้าจอสำหรับรันโมเดลของ Classification หลายๆ โมเดลและเปรียบเทียบ ประสิทธิภาพของในแต่ละโมเดล

ภาคผนวก ข

การสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวาน

การสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคเบาหวาน

1. นำเข้าข้อมูลด้วยโอเปอร์เรเตอร์ที่ชื่อว่า Retrieve เพื่อตรวจสอบข้อมูล

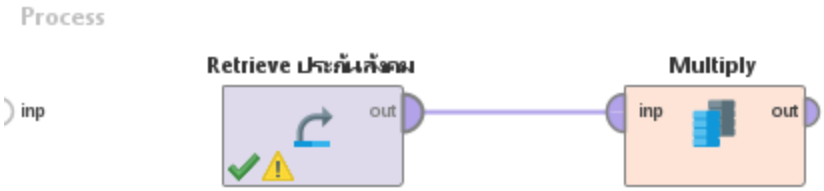


ภาพประกอบ 42 แสดงการนำเข้าไฟล์ข้อมูล *Retrieve* ประกันสังคม

Row No.	Disease	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol	BMI	Cholesterol	Glucose
1	Non-Diabetic	68	Female	47	146	65	120	55	No	No	22.050	200	83
2	Non-Diabetic	38	Female	48	152	54	108	100	No	No	20.780	213	81
3	Non-Diabetic	56	Male	91	165	90	126	79	No	No	33.430	237	93
4	Non-Diabetic	58	Male	67	160	81	131	91	No	No	26.170	182	124
5	Non-Diabetic	77	Female	54	164	68	146	85	No	No	20.080	157	214
6	Non-Diabetic	66	Female	58	159	72	132	78	No	No	25.780	237	92
7	Non-Diabetic	59	Male	57	155	73	127	75	No	Yes	23.730	211	111
8	Non-Diabetic	37	Female	46	150	59	105	66	No	No	20.440	146	86
9	Non-Diabetic	41	Female	69	150	86	134	107	No	No	30.670	261	115
10	Non-Diabetic	65	Female	53	149	69	117	63	No	No	23.870	250	93
11	Non-Diabetic	72	Female	46	150	65	118	78	No	No	20.440	153	97
12	Non-Diabetic	51	Female	83	160	59	110	67	No	No	32.420	191	85
13	Non-Diabetic	51	Female	69	158	85	143	97	No	No	27.640	164	79
14	Non-Diabetic	58	Female	60	160	86	127	76	No	No	23.440	248	268
15	Non-Diabetic	40	Female	88	158	66	133	99	No	No	35.250	201	87
16	Non-Diabetic	58	Female	55	155	86	158	109	No	No	22.890	190	84
17	Non-Diabetic	27	Male	80	165	67	120	59	Yes	Yes	29.380	183	85
18	Non-Diabetic	56	Female	57	150	82	132	64	No	No	25.330	187	118
19	Non-Diabetic	70	Female	50	155	78	138	64	No	No	20.810	264	103
20	Non-Diabetic	42	Female	70	163	67	107	92	No	No	26.350	279	77
21	Non-Diabetic	71	Female	56	150	82	134	79	No	No	24.890	222	90
22	Non-Diabetic	37	Female	39	150	72	107	106	No	No	17.330	185	86
23	Non-Diabetic	49	Female	83	170	91	156	103	No	No	28.720	191	83
24	Non-Diabetic	50	Female	70	160	81	147	98	No	No	27.340	224	76
25	Non-Diabetic	52	Female	55	155	61	103	66	No	No	22.890	212	84

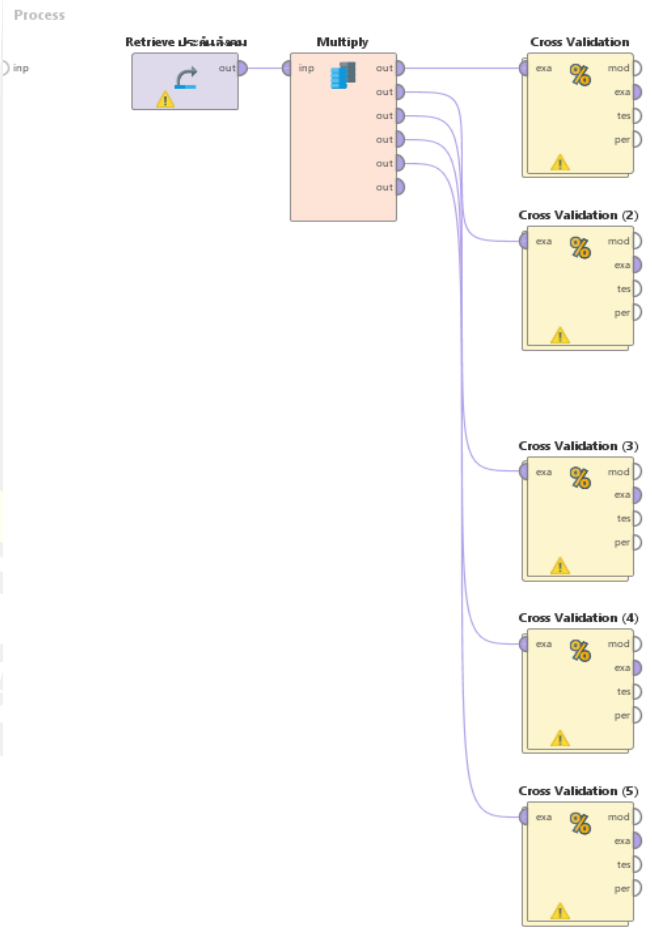
ภาพประกอบ 43 แสดงผลลัพธ์ของไฟล์ข้อมูล Retrieve Data_for_RUN

2. นำเขาโอเปอร์เรเตอร์ที่ชื่อว่า Multiply เพื่อแบ่งข้อมูลในการสร้างบบจำลอง



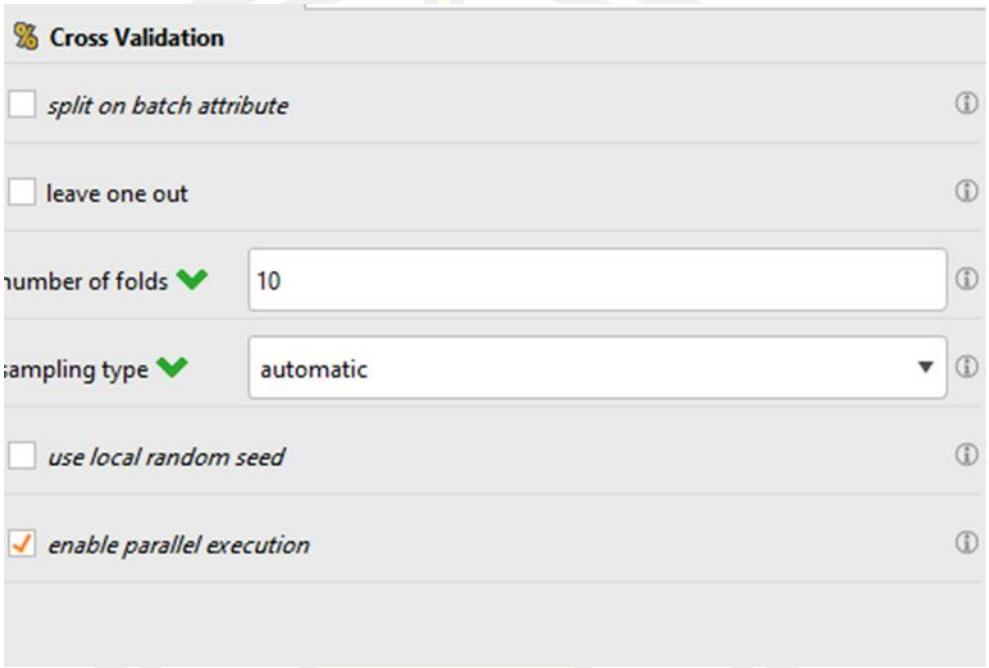
ภาพประกอบ 44 โอเปอร์เรเตอร์ Multiply

3. นำเขาโอเปอร์เรเตอร์ที่ชื่อว่า Cross Validation จำนวน 5 โอเปอร์เรเตอร์เพื่อสร้างแบบจำลอง



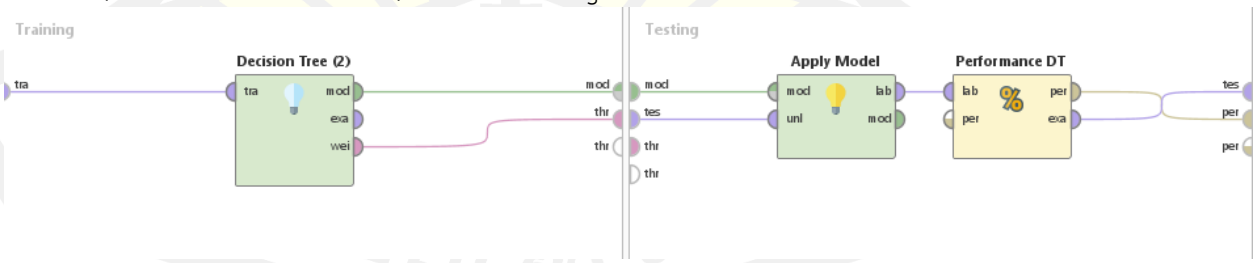
ภาพประกอบ 45 โอเปอร์เรเตอร์ Cross Validation

4. ปรับจำนวน fold ในโอเปอร์เรเตอร์ Cross Validation ทุกตัวให้เป็น 10 เพื่อใช้ในการทดสอบ 10 fold Cross Validation



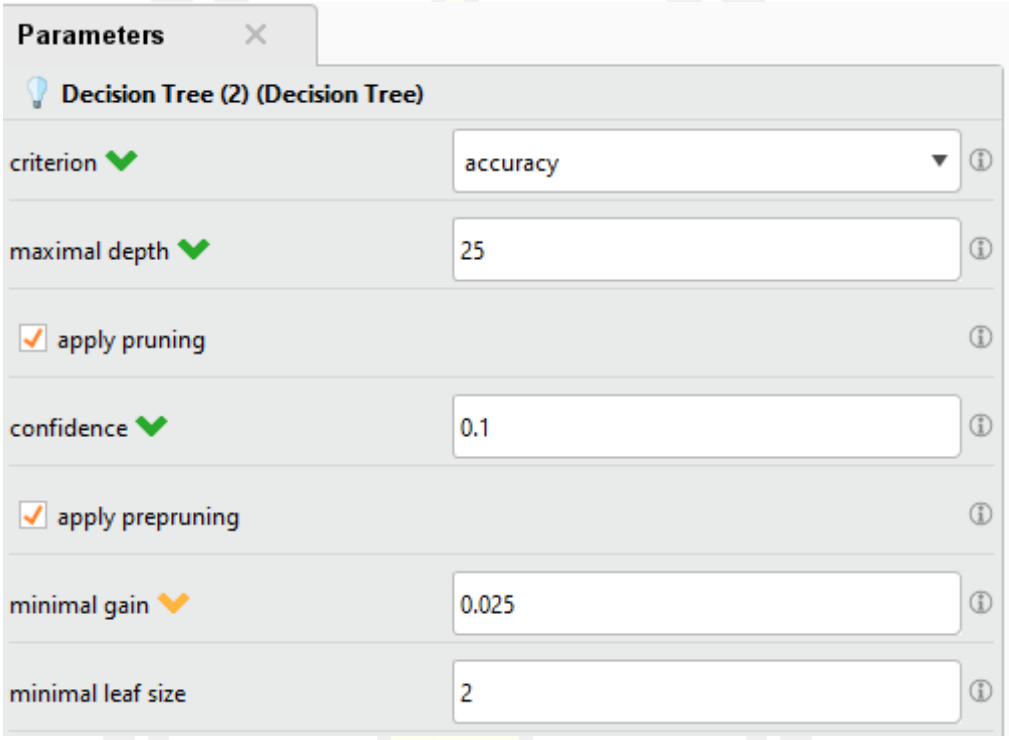
ภาพประกอบ 46 พารามีเตอร์ในโอเปอร์เรเตอร์ Cross Validation

5. เข้าไปในโอเปอร์เรเตอร์ Cross Validation อันแรกแล้วนำเข้าโอเปอร์เรเตอร์ Decision Tree ในพื้นที่ Training นำเข้าโอเปอร์เรเตอร์ Apply Model และ Performance (Binomial Classification) ในพื้นที่ Testing



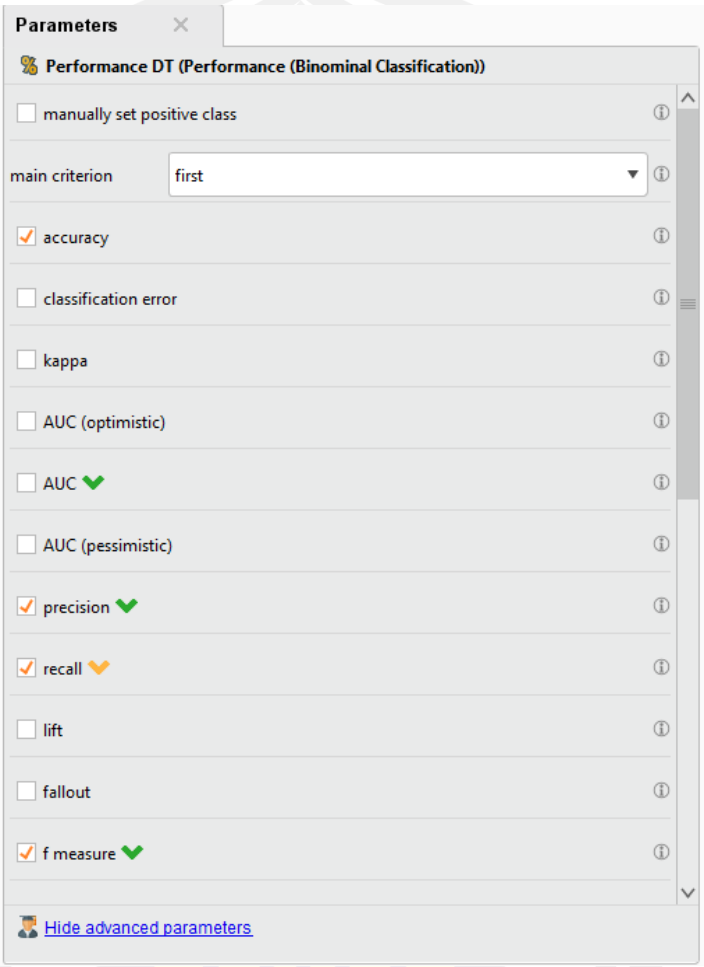
ภาพประกอบ 47 โอเปอร์เรเตอร์ Decision Tree, Apply Model และ Performance (Binomial Classification)

6. ตั้งค่าในพารามีเตอร์ ใน Decision Tree โดยตั้ง criterion เป็น Accuracy maximum depth เป็น 25 confident เป็น 1 minimal gain เป็น 0.01 และ minimal leaf size เป็น 2



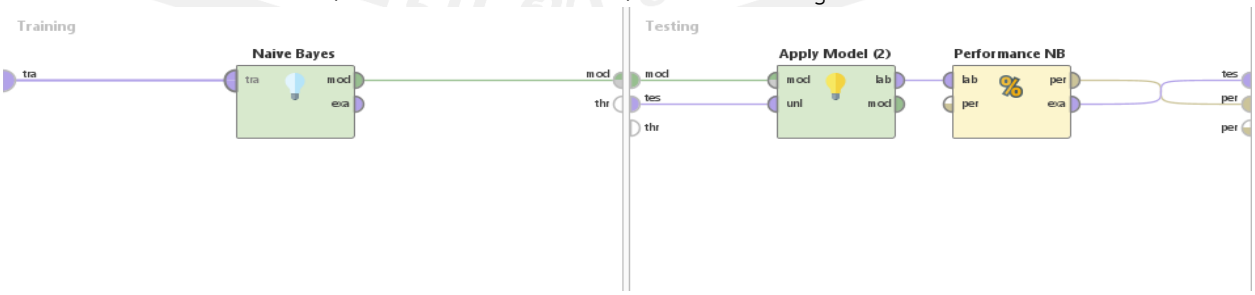
ภาพประกอบ 48 การตั้งค่าในพารามีเตอร์ Decision Tree

7. ตั้งค่าในพารามีเตอร์ Cross Validation โดยตั้ง main criterion เป็น first และเลือก accuracy, precision, recall และ f-measure



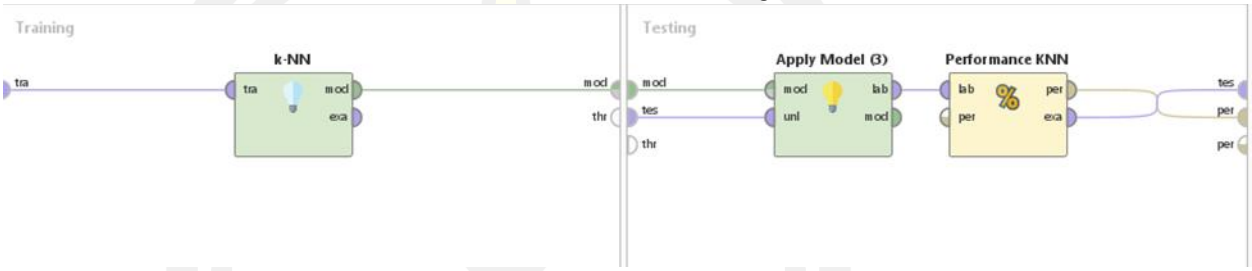
ภาพประกอบ 49 พารามีเตอร์ Cross Validation

8. ออกจากเข้าโอเปอร์เรเตอร์เดิมแล้วไปในโอเปอร์เรเตอร์ Cross Validation อันต่อมาแล้วนำเข้าโอเปอร์เรเตอร์ Naïve Bay ในพื้นที่ Training นำเข้าโอเปอร์เรเตอร์ Apply Model และ Performance (Binomial Classification) ในพื้นที่ Testing มาแล้วเชื่อมดังภาพ



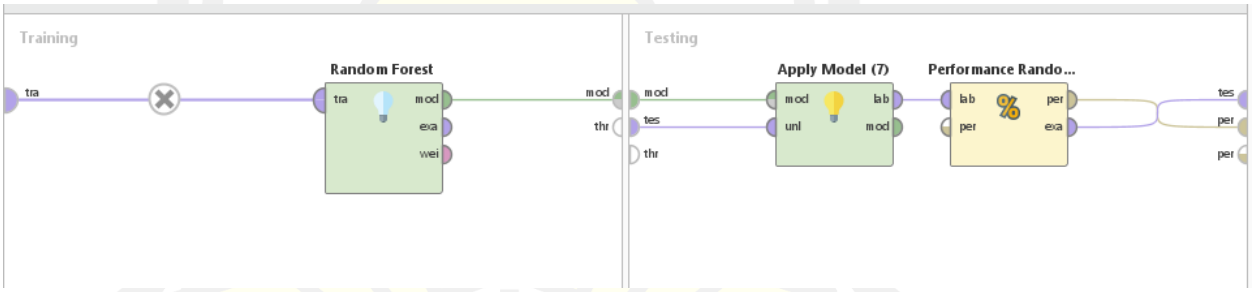
ภาพประกอบ 50 โอเปอร์เรเตอร์ Naïve Bay, Apply Model และ Performance (Binomial Classification)

9. ออกจากเข้าโอเปอร์เรเตอร์เดิมแล้วไปในโอเปอร์เรเตอร์ Cross Validation อันต่อมาแล้วนำเข้าโอเปอร์เรเตอร์ k-NN ในพื้นที่ Training นำเข้าโอเปอร์เรเตอร์ Apply Model และ Performance (Binomial Classification) ในพื้นที่ Testing มาแล้วเชื่อมดังภาพ



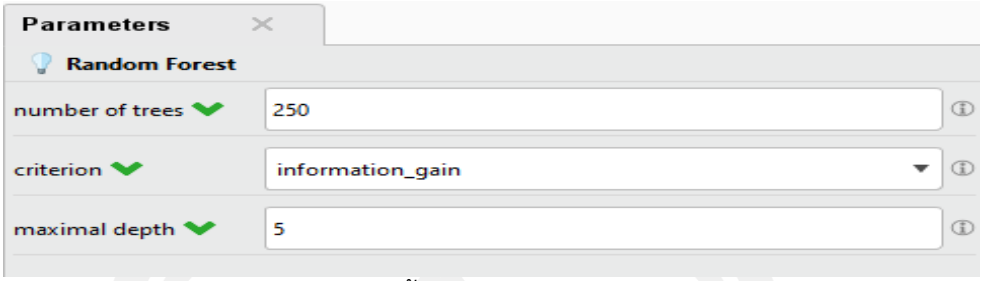
ภาพประกอบ 51 โอเปอร์เรเตอร์ k-NN, Apply Model *และ* Performance (Binomial Classification)

10. ออกจากเข้าโอเปอร์เรเตอร์เดิมแล้วไปในโอเปอร์เรเตอร์ Cross Validation อันต่อมาแล้วนำเข้าโอเปอร์เรเตอร์ Random Forest ในพื้นที่ Training นำเข้าโอเปอร์เรเตอร์ Apply Model และ Performance (Binomial Classification) ในพื้นที่ Testing มาแล้วเชื่อมดังภาพ



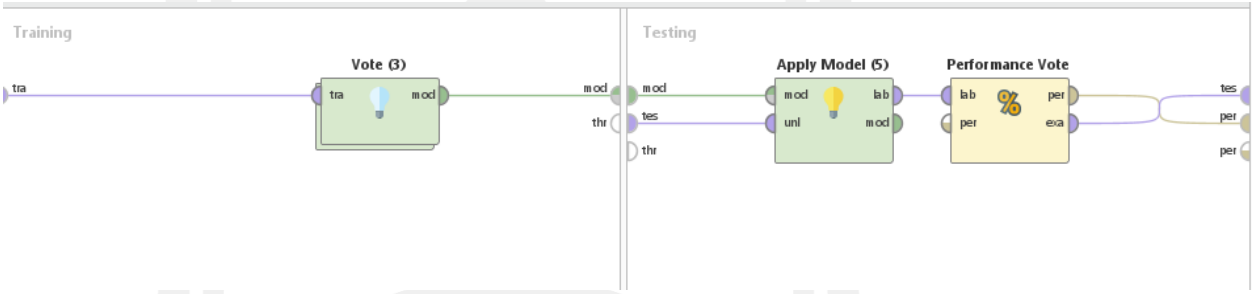
ภาพประกอบ 52 โอเปอร์เรเตอร์ Random Forest, Apply Model และ Performance (Binomial Classification)

11. โอเปอร์เรเตอร์ Random Forest ให้ทำการคลิ๊กที่โอเปอร์เรเตอร์ Random Forest แล้วตั้งค่า number of trees เป็น 250 criterion เป็น information gain maximal depth เป็น 5



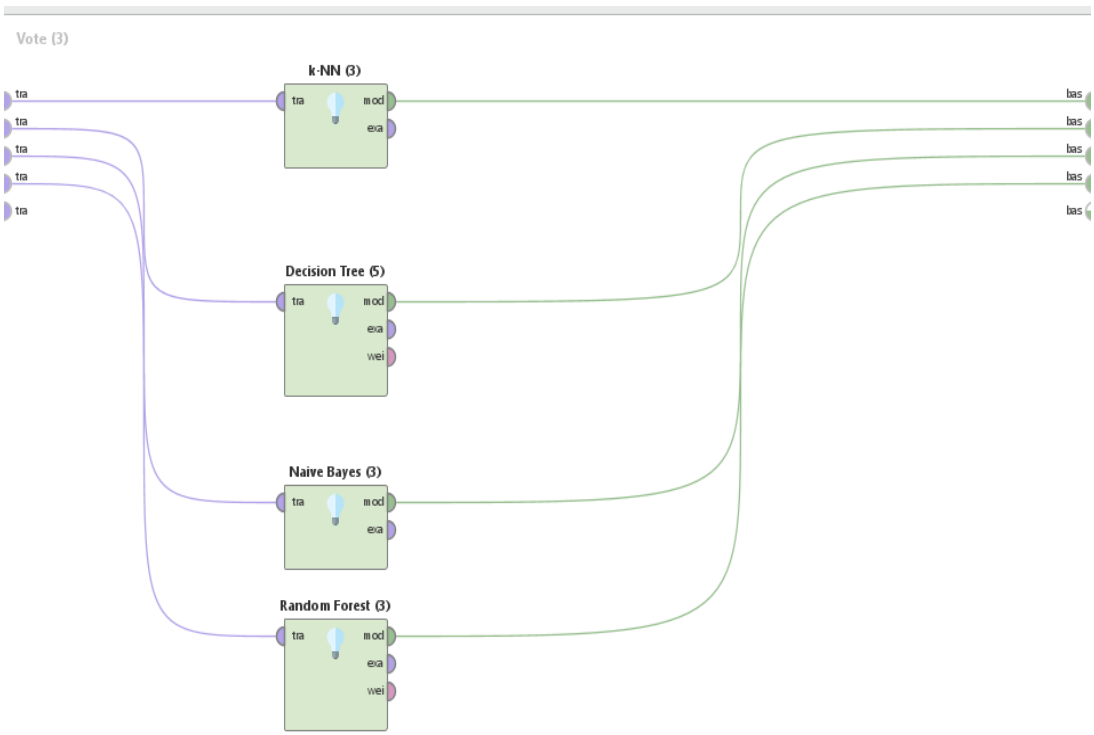
ภาพประกอบ 53 การตั้งค่าในพารามีเตอร์ Random Forest

12. ออกจากเข้าโอเปอร์เรเตอร์เดิมแล้วไปในโอเปอร์เรเตอร์ Cross Validation อันต่อมาแล้วนำเข้าโอเปอร์เรเตอร์ Vote ในพื้นที่ Training นำเข้าโอเปอร์เรเตอร์ Apply Model และ Performance (Binomial Classification) ในพื้นที่ Testing มาแล้วเชื่อมดังภาพ



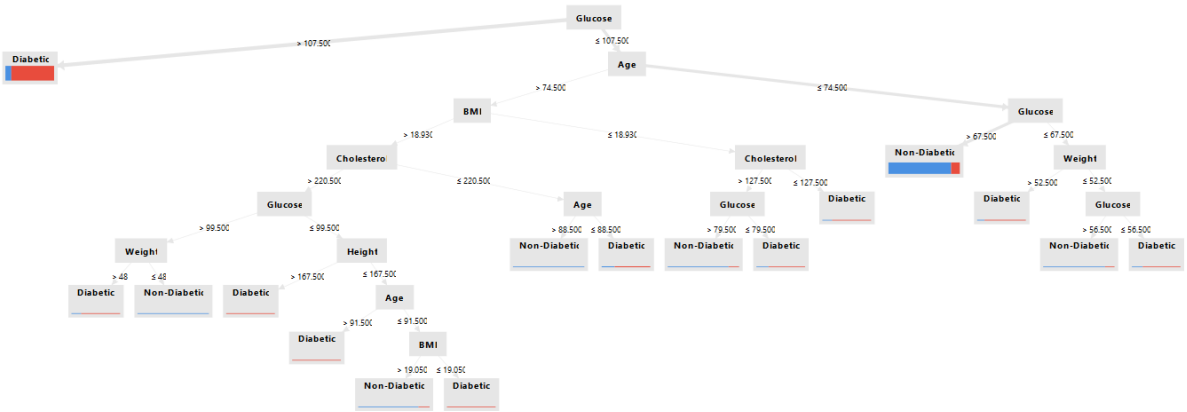
ภาพประกอบ 54 โอเปอร์เรเตอร์ Vote, Apply Model และ Performance (Binomial Classification)

13. ในโอเปอร์เรเตอร์ Vote ให้ดับเบิ้ลคลิกเข้าไปในโอเปอร์เรเตอร์แล้วเข้าโอเปอร์เรเตอร์ Decision Tree, Naïve Bay, k-NN และ Random Forest แล้วตั้งค่าพารามีเตอร์ดังที่แล้วมาแล้วเชื่อมดังภาพ



ภาพประกอบ 55 โอเปอร์เรเตอร์ภายในของ Vote

พอเมื่อกลับมาหน้าหลักแล้วให้กดรันผลลัพธ์แล้วจะได้ผลลัพธ์ในแต่ละเทคนิค



ภาพประกอบ 56 แผนภาพ Decision Tree

Tree

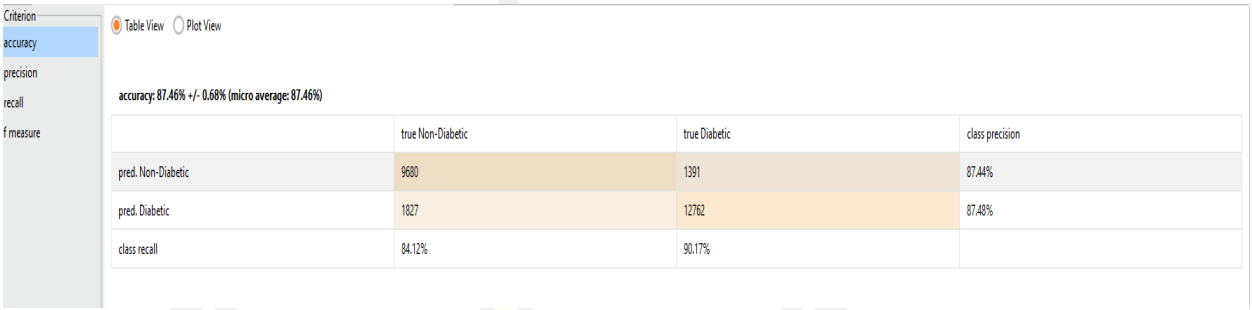
```
Glucose > 107.500: Diabetic {Non-Diabetic=1729, Diabetic=12517}
Glucose ≤ 107.500
|   Age > 74.500
|   |   BMI > 18.930
|   |   |   Cholesterol > 220.500
|   |   |   |   Glucose > 99.500
|   |   |   |   |   Weight > 48: Diabetic {Non-Diabetic=5, Diabetic=20}
|   |   |   |   |   Weight ≤ 48: Non-Diabetic {Non-Diabetic=2, Diabetic=0}
|   |   |   |   Glucose ≤ 99.500
|   |   |   |   |   Height > 167.500: Diabetic {Non-Diabetic=0, Diabetic=4}
|   |   |   |   |   Height ≤ 167.500
|   |   |   |   |   |   Age > 91.500: Diabetic {Non-Diabetic=0, Diabetic=2}
|   |   |   |   |   |   Age ≤ 91.500
|   |   |   |   |   |   |   BMI > 19.050: Non-Diabetic {Non-Diabetic=38, Diabetic=7}
|   |   |   |   |   |   |   BMI ≤ 19.050: Diabetic {Non-Diabetic=0, Diabetic=2}
|   |   |   Cholesterol ≤ 220.500
|   |   |   |   Age > 88.500: Non-Diabetic {Non-Diabetic=2, Diabetic=0}
|   |   |   |   Age ≤ 88.500: Diabetic {Non-Diabetic=65, Diabetic=179}
|   |   BMI ≤ 18.930
|   |   |   Cholesterol > 127.500
|   |   |   |   Glucose > 79.500: Non-Diabetic {Non-Diabetic=29, Diabetic=5}
|   |   |   |   Glucose ≤ 79.500: Diabetic {Non-Diabetic=1, Diabetic=3}
|   |   |   Cholesterol ≤ 127.500: Diabetic {Non-Diabetic=1, Diabetic=4}
|   Age ≤ 74.500
|   |   Glucose > 67.500: Non-Diabetic {Non-Diabetic=9607, Diabetic=1329}
|   |   Glucose ≤ 67.500
|   |   |   Weight > 52.500: Diabetic {Non-Diabetic=13, Diabetic=72}
|   |   |   Weight ≤ 52.500
|   |   |   |   Glucose > 56.500: Non-Diabetic {Non-Diabetic=13, Diabetic=2}
|   |   |   |   Glucose ≤ 56.500: Diabetic {Non-Diabetic=2, Diabetic=7}
```

ภาพประกอบ 57 รายละเอียดของ Decision Tree

Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol	BMI	Cholesterol	Glucose
1	Non-Diabetic	Non-Diabetic	0.123	0.877	56	Male	91	165	90	126	79	No	No	33.430	237	93
2	Non-Diabetic	Diabetic	0.879	0.121	41	Female	69	150	86	134	107	No	No	30.670	261	115
3	Non-Diabetic	Non-Diabetic	0.123	0.877	51	Female	83	160	59	110	67	No	No	32.420	191	85
4	Non-Diabetic	Diabetic	0.879	0.121	58	Female	60	160	86	127	76	No	No	23.440	248	268
5	Non-Diabetic	Non-Diabetic	0.123	0.877	73	Female	63	160	78	124	88	No	No	24.610	163	92
6	Non-Diabetic	Non-Diabetic	0.123	0.877	48	Female	55	155	62	115	72	No	No	22.890	205	88
7	Non-Diabetic	Non-Diabetic	0.123	0.877	64	Female	52	156	77	134	65	No	No	21.370	209	84
8	Non-Diabetic	Non-Diabetic	0.123	0.877	54	Female	63	150	68	125	85	No	No	28	229	97
9	Non-Diabetic	Non-Diabetic	0.123	0.877	36	Male	76	175	80	113	76	No	Yes	24.820	191	92
10	Non-Diabetic	Diabetic	0.879	0.121	71	Male	56	163	68	123	91	No	No	21	204	141
11	Non-Diabetic	Non-Diabetic	0.123	0.877	54	Female	65	160	70	108	67	No	No	25.390	248	83
12	Non-Diabetic	Non-Diabetic	0.123	0.877	42	Female	53	158	63	105	70	No	No	21.230	224	86
13	Non-Diabetic	Non-Diabetic	0.123	0.877	74	Male	65	174	87	166	78	No	No	21.470	99	94
14	Non-Diabetic	Non-Diabetic	0.123	0.877	61	Male	65	160	90	135	76	No	No	25.390	186	99
15	Non-Diabetic	Non-Diabetic	0.123	0.877	64	Female	48	149	66	128	85	No	No	21.620	203	97
16	Non-Diabetic	Diabetic	0.879	0.121	56	Male	65	165	76	121	81	No	No	23.880	277	117
17	Non-Diabetic	Non-Diabetic	0.123	0.877	60	Female	57	156	46	110	66	No	No	23.420	214	92
18	Non-Diabetic	Diabetic	0.879	0.121	61	Male	52	164	70	96	96	No	No	19.330	256	112
19	Non-Diabetic	Non-Diabetic	0.123	0.877	55	Male	58	160	78	136	80	No	No	22.850	239	86
20	Non-Diabetic	Non-Diabetic	0.123	0.877	53	Female	69	155	86	132	93	No	No	28.720	169	80
21	Non-Diabetic	Non-Diabetic	0.123	0.877	54	Female	67	160	82	138	70	No	No	26.170	212	89
22	Non-Diabetic	Non-Diabetic	0.123	0.877	64	Female	51	152	72	130	76	No	No	22.070	199	83
23	Non-Diabetic	Non-Diabetic	0.123	0.877	57	Female	41	145	78	135	89	No	No	19.500	195	89
24	Non-Diabetic	Non-Diabetic	0.123	0.877	48	Female	56	150	85	119	75	No	No	24.890	184	84
25	Non-Diabetic	Non-Diabetic	0.123	0.877	62	Male	72	161	82	152	68	No	No	27.780	163	85

examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 58 Example set *แสดงค่า* unseen data *และค่า* confident *ของ* Decision Tree



Criterion
accuracy
precision
recall
f measure

Table View Plot View

accuracy: 87.46% +/- 0.68% (micro average: 87.46%)

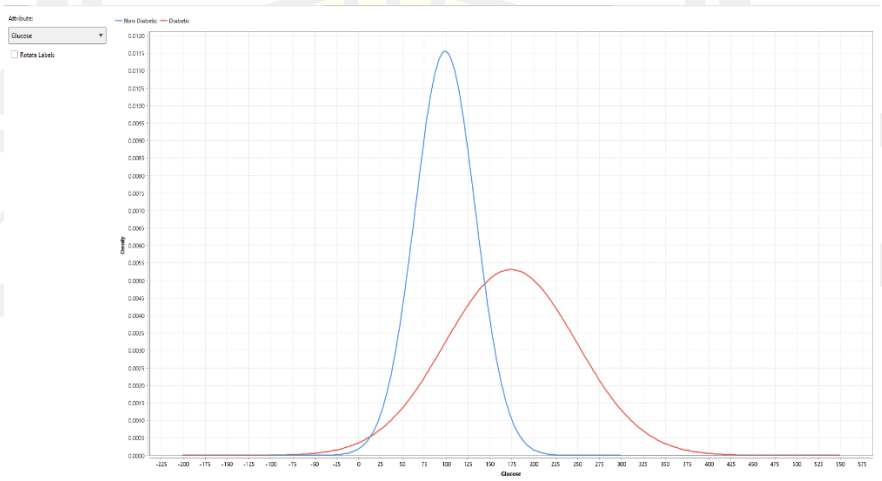
	true Non-Diabetic	true Diabetic	class precision
pred. Non-Diabetic	9680	1391	87.44%
pred. Diabetic	1827	12762	87.48%
class recall	84.12%	90.17%	

ภาพประกอบ 59 ตารางการวัดประสิทธิภาพของ Decision Tree

SimpleDistribution

```
Distribution model for label attribute Disease


Class Non-Diabetic (0.448)
12 distributions

Class Diabetic (0.552)
12 distributions
```

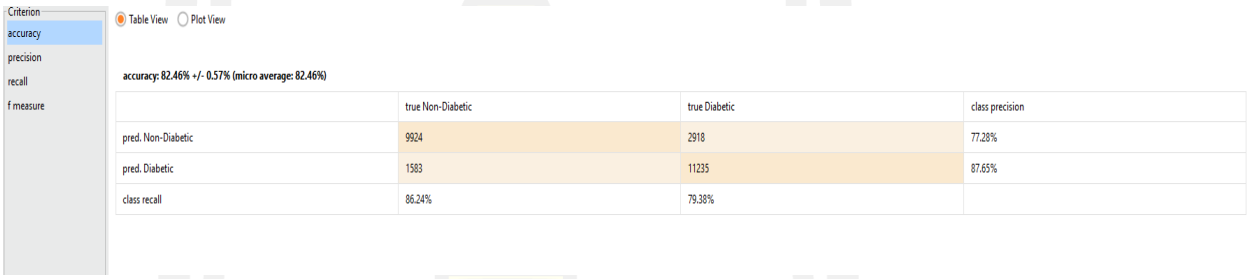
ภาพประกอบ 60 แสดงผลลัพธ์ของการกลุ่มข้อมูลด้วยเทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Naïve Bayes



ภาพประกอบ 61 แสดงกราฟ ผลลัพธ์ของการกลุ่มข้อมูลด้วยเทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Naïve Bayes

Attribute	Parameter	Non-Diabetic	Diabetic
Age	mean	50.786	59.621
Age	standard deviation	11.768	11.594
sex	value=Female	0.649	0.624
sex	value=Male	0.351	0.376
sex	value=unknown	0.000	0.000
Weight	mean	61.448	65.457
Weight	standard deviation	11.652	12.976
Height	mean	159.198	158.180
Height	standard deviation	7.944	8.091
Diastolic Blood Pressure	mean	76.519	76.473
Diastolic Blood Pressure	standard deviation	12.178	12.136
Systolic Blood Pressure	mean	125.216	135.077
Systolic Blood Pressure	standard deviation	17.399	19.401

ภาพประกอบ 62 แสดงผลลัพธ์ Distribution Table ของค่า Attribute , ค่า Parameter , ค่าความน่าจะค่าความน่าจะเป็นโรคเบาหวานและเป็นโรคเบาหวาน



ภาพประกอบ 63 ตารางการวัดประสิทธิภาพของ Naïve Bay

KNNClassification

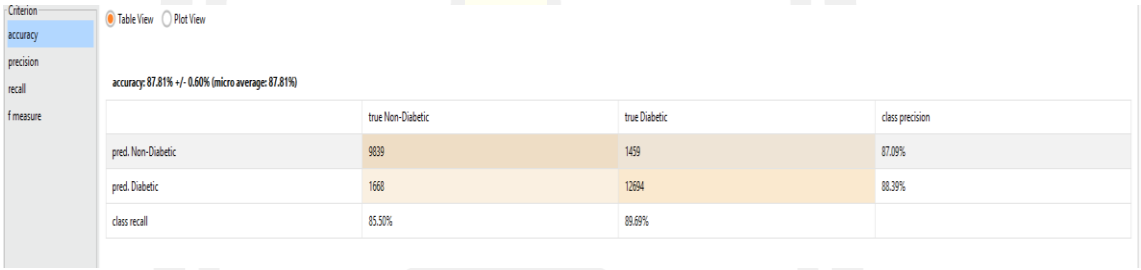
```
Weighted 5-Nearest Neighbour model for classification.
The model contains 25660 examples with 12 dimensions of the following classes:
  Non-Diabetic
  Diabetic
```

ภาพประกอบ 64 แสดงผลลัพธ์ของการกลุ่มข้อมูลด้วยเทคนิคการจำแนกประเภทข้อมูลด้วยวิธี K-Nearest Neighbor

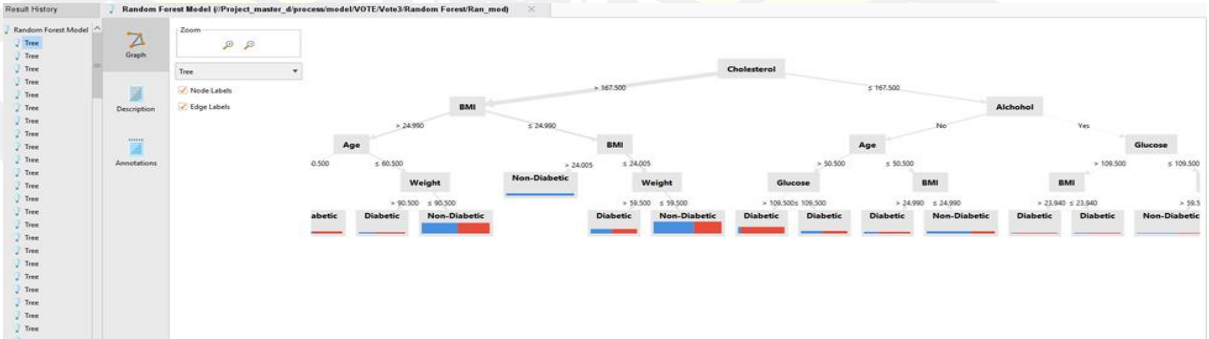
Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol	BMI	Cholesterol	Glucose
1	Non-Diabetic	Diabetic	0.598	0.402	71	Male	56	163	68	123	91	No	No	21	204	141
2	Non-Diabetic	Non-Diabetic	0	1	49	Female	44	158	63	108	75	No	Yes	17.630	228	96
3	Non-Diabetic	Non-Diabetic	0	1	44	Female	54	140	60	108	73	No	Yes	27.550	180	87
4	Non-Diabetic	Non-Diabetic	0	1	41	Female	56	157	81	128	87	No	No	22.720	189	78
5	Non-Diabetic	Diabetic	1	0	74	Male	65	174	87	166	78	No	No	21.470	99	94
6	Non-Diabetic	Diabetic	0.599	0.401	48	Female	60	147	76	124	80	No	No	27.770	267	129
7	Non-Diabetic	Non-Diabetic	0.206	0.794	60	Female	52	151	69	138	63	No	No	22.810	212	91
8	Non-Diabetic	Non-Diabetic	0.197	0.803	61	Male	65	160	90	135	76	No	No	25.390	186	99
9	Non-Diabetic	Non-Diabetic	0	1	60	Female	44	162	60	90	67	No	No	16.770	261	94
10	Non-Diabetic	Non-Diabetic	0	1	34	Female	63	165	79	107	80	No	No	23.140	210	87
11	Non-Diabetic	Diabetic	1	0	78	Female	65	165	79	160	90	No	No	23.880	184	108
12	Non-Diabetic	Non-Diabetic	0	1	43	Female	65	150	71	112	84	No	No	28.890	209	87
13	Non-Diabetic	Non-Diabetic	0.410	0.590	56	Male	65	165	76	121	81	No	No	23.880	277	117
14	Non-Diabetic	Diabetic	0.600	0.400	81	Female	54	150	70	140	77	No	No	24	230	91
15	Non-Diabetic	Non-Diabetic	0.201	0.799	73	Female	58	148	83	116	98	No	No	26.750	176	101
16	Non-Diabetic	Diabetic	0.804	0.196	52	Male	71	170	87	130	108	Yes	Yes	24.570	206	145
17	Non-Diabetic	Non-Diabetic	0.198	0.802	49	Female	60	156	53	117	77	No	No	24.650	191	91
18	Non-Diabetic	Non-Diabetic	0	1.000	63	Female	57	153	76	118	74	No	No	24.350	255	82
19	Non-Diabetic	Diabetic	0.800	0.200	55	Male	74	165	79	117	93	No	No	27.180	169	109
20	Non-Diabetic	Non-Diabetic	0	1	58	Female	56	145	77	135	83	No	No	26.630	248	97
21	Non-Diabetic	Non-Diabetic	0	1	63	Female	42	148	79	129	96	No	No	19.170	184	99
22	Non-Diabetic	Non-Diabetic	0.413	0.587	51	Female	95	150	85	127	95	No	No	42.220	236	110
23	Non-Diabetic	Non-Diabetic	0	1	37	Female	58	160	72	120	74	No	No	22.660	201	84
24	Non-Diabetic	Non-Diabetic	0	1	49	Female	62	150	73	130	71	No	No	27.560	193	78
25	Non-Diabetic	Non-Diabetic	0	1	51	Female	65	155	80	119	75	No	No	27.060	240	92

examples, 5 special attributes, 12 regular attributes)

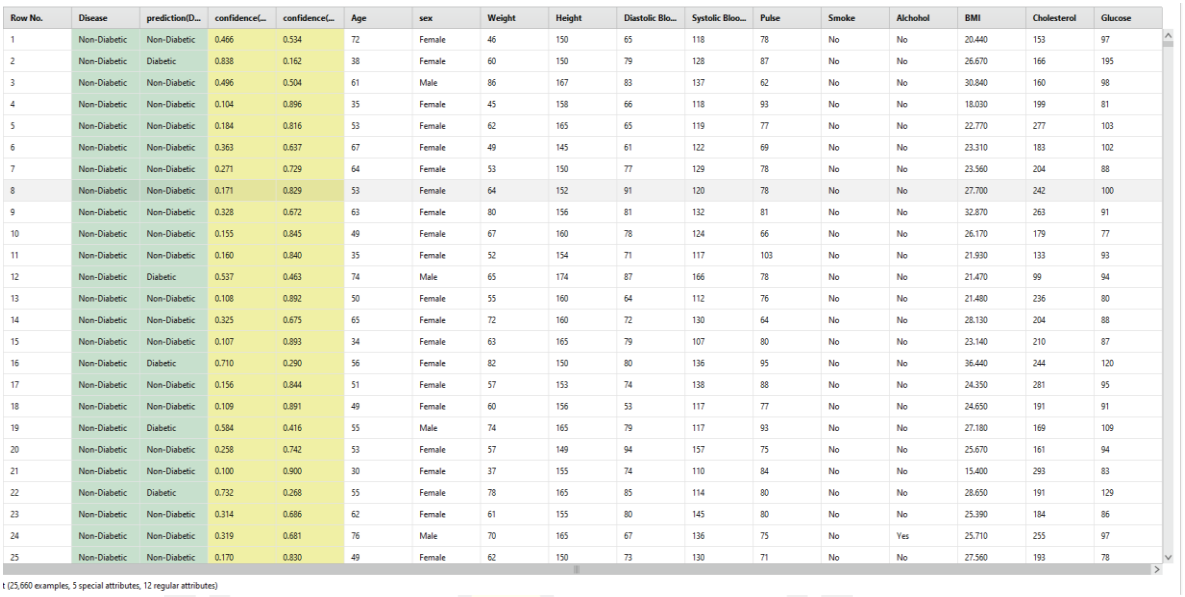
ภาพประกอบ 65 Example set แสดงค่า unseen data และค่า confident ของ K- Nearest Neighbor



ภาพประกอบ ข.21: ตารางการวัดประสิทธิภาพของ K- Nearest Neighbor



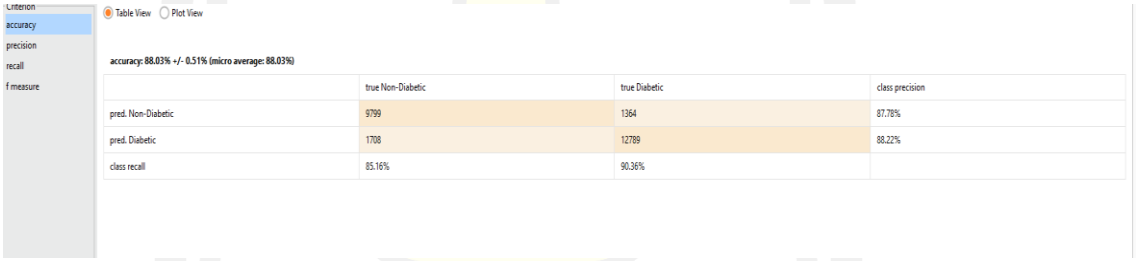
ภาพประกอบ 66 ตัวอย่างหนึ่งในแผนภาพของ Random Forest จาก 250 แบบ



Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol	BMI	Cholesterol	Glucose
1	Non-Diabetic	Non-Diabetic	0.466	0.534	72	Female	46	150	65	118	78	No	No	20.440	153	97
2	Non-Diabetic	Diabetic	0.838	0.162	38	Female	60	150	79	128	87	No	No	26.670	166	195
3	Non-Diabetic	Non-Diabetic	0.496	0.504	61	Male	86	167	83	137	62	No	No	30.840	160	98
4	Non-Diabetic	Non-Diabetic	0.104	0.896	35	Female	45	158	66	118	93	No	No	18.030	199	81
5	Non-Diabetic	Non-Diabetic	0.184	0.816	53	Female	62	165	65	119	77	No	No	22.770	277	103
6	Non-Diabetic	Non-Diabetic	0.363	0.637	67	Female	49	145	61	122	69	No	No	23.310	183	102
7	Non-Diabetic	Non-Diabetic	0.271	0.729	64	Female	53	150	77	129	78	No	No	23.560	204	88
8	Non-Diabetic	Non-Diabetic	0.171	0.829	53	Female	64	152	91	120	78	No	No	27.700	242	100
9	Non-Diabetic	Non-Diabetic	0.328	0.672	63	Female	80	156	81	132	81	No	No	32.870	263	91
10	Non-Diabetic	Non-Diabetic	0.155	0.845	49	Female	67	160	78	124	66	No	No	26.170	179	77
11	Non-Diabetic	Non-Diabetic	0.160	0.840	35	Female	52	154	71	117	103	No	No	21.930	133	93
12	Non-Diabetic	Diabetic	0.537	0.463	74	Male	65	174	87	166	78	No	No	21.470	99	94
13	Non-Diabetic	Non-Diabetic	0.108	0.892	50	Female	55	160	64	112	76	No	No	21.480	236	80
14	Non-Diabetic	Non-Diabetic	0.325	0.675	65	Female	72	160	72	130	64	No	No	28.130	204	88
15	Non-Diabetic	Non-Diabetic	0.107	0.893	34	Female	63	165	79	107	80	No	No	23.140	210	87
16	Non-Diabetic	Diabetic	0.710	0.290	56	Female	82	150	80	136	95	No	No	36.440	244	120
17	Non-Diabetic	Non-Diabetic	0.156	0.844	51	Female	57	153	74	138	88	No	No	24.350	281	95
18	Non-Diabetic	Non-Diabetic	0.109	0.891	49	Female	60	156	53	117	77	No	No	24.650	191	91
19	Non-Diabetic	Diabetic	0.584	0.416	55	Male	74	165	79	117	93	No	No	27.180	169	109
20	Non-Diabetic	Non-Diabetic	0.258	0.742	53	Female	57	149	94	157	75	No	No	25.670	161	94
21	Non-Diabetic	Non-Diabetic	0.100	0.900	30	Female	37	155	74	110	84	No	No	15.400	293	83
22	Non-Diabetic	Diabetic	0.732	0.268	55	Female	78	165	85	114	80	No	No	28.650	191	129
23	Non-Diabetic	Non-Diabetic	0.314	0.686	62	Female	61	155	80	145	80	No	No	25.390	184	86
24	Non-Diabetic	Non-Diabetic	0.319	0.681	76	Male	70	165	67	136	75	No	Yes	25.710	255	97
25	Non-Diabetic	Non-Diabetic	0.170	0.830	49	Female	62	150	73	130	71	No	No	27.560	193	78

t (25,660 examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 67 Example set แสดงค่า unseen data และค่า confident ของ Random Forest



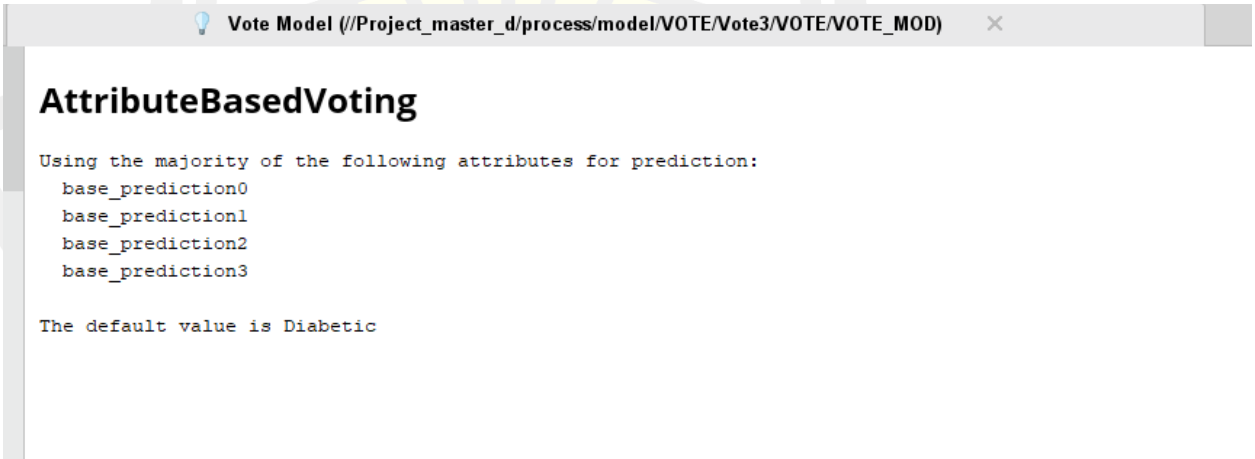
Criterion
accuracy
precision
recall
f measure

Table View Plot View

accuracy: 88.03% +/- 0.51% (micro average: 88.03%)

	true Non-Diabetic	true Diabetic	class precision
pred. Non-Diabetic	9799	1364	87.78%
pred. Diabetic	1708	12789	88.22%
class recall	85.16%	90.36%	

ภาพประกอบ 68 ตารางการวัดประสิทธิภาพของ Random Forest



Vote Model (//Project_master_d/process/model/VOTE/Vote3/VOTE/VOTE_MOD)

AttributeBasedVoting

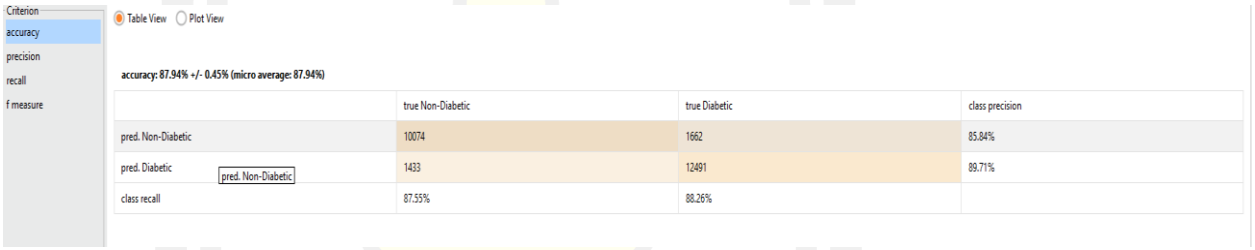
```
Using the majority of the following attributes for prediction:
  base_prediction0
  base_prediction1
  base_prediction2
  base_prediction3

The default value is Diabetic
```

ภาพประกอบ 69 ผลลัพธ์การจำแนกประเภทข้อมูลด้วยเทคนิค Vote Ensemble

Row No.	Disease	prediction(D...	confidence(...	confidence(...	Age	sex	Weight	Height	Diastolic Blo...	Systolic Bloo...	Pulse	Smoke	Alchohol	BMI	Cholesterol	Glucose
1	Non-Diabetic	Non-Diabetic	0	1	37	Female	46	150	59	105	66	No	No	20.440	146	86
2	Non-Diabetic	Non-Diabetic	0	1	50	Female	70	160	81	147	98	No	No	27.340	224	76
3	Non-Diabetic	Non-Diabetic	0.250	0.750	73	Female	63	160	78	124	88	No	No	24.610	163	92
4	Non-Diabetic	Non-Diabetic	0	1	48	Female	50	158	64	99	75	No	No	20.030	235	87
5	Non-Diabetic	Non-Diabetic	0.250	0.750	41	Female	95	160	64	127	74	No	No	37.110	206	104
6	Non-Diabetic	Non-Diabetic	0	1	50	Female	48	165	52	84	85	No	No	17.630	276	89
7	Non-Diabetic	Non-Diabetic	0	1	52	Female	55	145	75	125	98	No	No	26.160	175	105
8	Non-Diabetic	Diabetic	1	0	66	Female	77	160	80	120	85	No	No	30.080	178	112
9	Non-Diabetic	Non-Diabetic	0.500	0.500	70	Male	69	169	71	165	46	No	No	24.160	190	89
10	Non-Diabetic	Non-Diabetic	0	1	50	Female	55	160	64	112	76	No	No	21.480	236	80
11	Non-Diabetic	Non-Diabetic	0	1	60	Female	52	151	69	138	63	No	No	22.810	212	91
12	Non-Diabetic	Non-Diabetic	0	1	70	Female	52	157	75	120	84	No	No	21.100	213	96
13	Non-Diabetic	Non-Diabetic	0	1	49	Female	75	160	90	132	84	No	No	29.300	243	95
14	Non-Diabetic	Non-Diabetic	0	1	53	Female	45	150	65	102	67	No	No	20	192	88
15	Non-Diabetic	Non-Diabetic	0	1	34	Female	63	165	79	107	80	No	No	23.140	210	87
16	Non-Diabetic	Diabetic	0.750	0.250	56	Female	82	150	80	136	95	No	No	36.440	244	120
17	Non-Diabetic	Non-Diabetic	0.500	0.500	61	Male	80	165	69	131	60	No	No	29.380	200	109
18	Non-Diabetic	Non-Diabetic	0.500	0.500	71	Female	49	158	77	128	65	No	No	19.630	318	116
19	Non-Diabetic	Non-Diabetic	0	1	43	Female	65	150	71	112	84	No	No	28.890	209	87
20	Non-Diabetic	Non-Diabetic	0	1	51	Female	61	155	75	122	78	No	No	25.390	263	86
21	Non-Diabetic	Non-Diabetic	0	1	35	Female	72	160	68	118	80	No	Yes	28.130	155	85
22	Non-Diabetic	Non-Diabetic	0.250	0.750	58	Female	53	150	80	130	78	No	No	23.560	185	101
23	Non-Diabetic	Non-Diabetic	0	1	62	Female	57	150	89	141	83	No	No	25.330	256	96

0 examples, 5 special attributes, 12 regular attributes)

ภาพประกอบ 70 Example set แสดงค่า unseen data และค่า confident ของ Vote Ensemble



ภาพประกอบ 71 ตารางการวัดประสิทธิภาพของ Vote Ensemble

ประวัติผู้เขียน

ชื่อ	ปพนน์ศรณ์ สิ่วสำแดงเดช
วันเกิด	11 พฤศจิกายน พ.ศ.2539
สถานที่เกิด	ต.หมากแข้ง อ.เมือง จ. อุดรธานี
สถานที่อยู่ปัจจุบัน	93/108 ถนนศรีชมชื่น ซอยสุรทักษ์ 6 ตำบลหมากแข้ง อำเภอเมือง จังหวัดอุดรธานี รหัสไปรษณีย์ 41000
ประวัติการศึกษา	พ.ศ. 2555 มัธยมศึกษาตอนต้น โรงเรียนดอนบอสโกวิทยา อุดรธานี พ.ศ. 2558 มัธยมศึกษาตอนปลาย โรงเรียนมัธยมเทศบาล 6 นครอุดรธานี พ.ศ. 2562 ปริญญาบริหารธุรกิจบัณฑิต (บธ.บ) สาขาเทคโนโลยีสารสนเทศธุรกิจ มหาวิทยาลัยมหาสารคาม พ.ศ. 2565 วิศวกรรมศาสตรมหาบัณฑิต (วศ.ม.) วิศวกรรมไฟฟ้าและคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม