**Received:** Jun 10, 2023 **Revised:** Aug 24, 2023 **Accepted:** Aug 28, 2023

# การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน

# COMPARISON OF FEATURE SELECTION METHODS TO IMPROVE DIABETES PREDICTIONS

สรรภัส สังเกตุ, และปราลี มณีรัตน์
สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม
Sanpath Sunggad, Paralee Maneerat
Information Technology, Sripathum University
sanpath.sug@spumail.net

**บทคัดย่อ**

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน เพื่อพัฒนาการคัดกรองโรคเบาหวานจากข้อมูลผู้เข้ารับคัดกรองโรค เนื่องจากเบาหวานเป็นหนึ่งในกลุ่มโรคไม่ติดต่อ ซึ่งเป็นปัญหาสุขภาพอันดับหนึ่งของโลกทั้งจำนวนของการเสียชีวิต จากการรายงานข้อมูลขององค์การอนามัยโลก (WHO) พบประชากรทั่วโลกเสียชีวิตจากโรคกลุ่มโรคไม่ติดต่อมีแนวโน้มเพิ่มสูงขึ้น ซึ่งในแต่ละปีพบผู้เสียชีวิตจากโรคไม่ติดต่อในกลุ่มอายุ 30-69 ปี มากถึง 15 ล้านคน ในประเทศไทยพบเพศชายมีอัตราการเสียชีวิตสูงกว่า เพศหญิง อาการของโรคเบาหวานที่พบบ่อย มักพบอาการเหล่านี้ร่วมกัน อาการปัสสาวะบ่อย, อาการกระหายน้ำหิวน้ำบ่อย, อาการ น้ำหนักลดโดยไม่ทราบสาเหตุ, อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง, อาการกินจุ หิวบ่อย, อาการโรคเชื้อราในช่องคลอด, สายตาพร่ามัวมองไม่ชัดเจน, คันตามผิวหนัง, อาการหงุดหงิดง่าย, อาการเป็นแผลง่ายแผลหายยาก, อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน, อาการกล้ามเนื้อหดเกร็ง, อาการการร่วงของผมเป็นหย่อม และอาการสะสมไขมันในส่วนต่างๆ ของร่างกายเกินปกติ ผู้วิจัยได้ทำการนำคุณลักษณะของโรคเบาหวานทั้ง 16 คุณลักษณะ มาศึกษาเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญ ที่ใช้ในการปรับปรุงการพยากรณ์โรคเบาหวาน

**คำสำคัญ:** โรคเบาหวาน, การคัดกรองโรคเบาหวานเบาหวาน, การคัดเลือกคุณลักษณะ

**Abstract**

This research aims to compare methods for selecting key characteristics to enhance the prognosis of diabetes. Its objective is to develop a diabetes screening process based on data obtained from individuals undergoing diabetes screening. Diabetes ranks as a leading non-communicable disease and represents a primary global health concern, according to the World Health Organization (WHO). There's a growing trend in fatalities due to non-communicable diseases globally. Annually, approximately 15 million deaths occur within the 30-69 age group due to non-communicable diseases. In Thailand, males exhibit higher mortality rates than females. Commonly observed symptoms among individuals with diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, increased hunger, fungal infections, blurred vision, itchy skin, irritability, slow-healing wounds, numbness, muscle stiffness, hair loss, and abnormal fat accumulation in various body parts. The research compared the 16 characteristics of diabetes to refine methods for selecting crucial features that significantly improve disease prognosis. The aim is to enhance understanding regarding the most suitable feature selection methods for diabetes screening. Furthermore, the study endeavors to provide valuable information for developing future

strategies in managing and treating this condition effectively. Moreover, the research aims to increase the efficiency of diabetes screening to reduce risks and enable effective early-stage treatment.
**Keywords:** Diabetes, Diabetes screening, Feature Selection.

## บทนำ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน เพื่อพัฒนาการคัดกรองโรคเบาหวานจากข้อมูลผู้เข้ารับคัดกรองโรค เนื่องจากเบาหวานเป็นหนึ่งในกลุ่มโรคไม่ติดต่อ ซึ่งเป็นปัญหาสุขภาพอันดับหนึ่งของโลกทั้งจำนวนของการเสียชีวิต จากการรายงานข้อมูลขององค์การอนามัยโลก (WHO) พบประชากรทั่วโลกเสียชีวิตจากโรคกลุ่มโรคไม่ติดต่อมีแนวโน้มเพิ่มสูงขึ้น ซึ่งในแต่ละปีพบผู้เสียชีวิตจากโรคไม่ติดต่อในกลุ่มอายุ 30-69 ปี มากถึง 15 ล้านคน ในประเทศไทยพบเพศชายมีอัตราการเสียชีวิตสูงกว่า เพศหญิง อาการของโรคเบาหวานผู้ป่วยมักมีอาการเหล่านี้ร่วมกัน อาการปัสสาวะบ่อย, อาการกระหายน้ำหิวน้ำบ่อย, อาการ น้ำหนักลดโดยไม่ทราบสาเหตุ, อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง, อาการกินจุ หิวบ่อย, อาการโรคเชื้อราในช่องคลอด, สายตาพร่ามัวมองไม่ชัดเจน, คันตามผิวหนัง, อาการหงุดหงิดง่าย, อาการเป็นแผลง่ายแผลหายยาก, อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน, อาการกล้ามเนื้อหดเกร็ง, อาการการร่วงของผมเป็นหย่อม และอาการสะสมไขมันในส่วนต่างๆ ของร่างกายเกินปกติ ผู้วิจัยได้ทำการนำคุณลักษณะของโรคเบาหวานทั้ง 16 คุณลักษณะ มาศึกษาเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญ ที่ใช้ในการปรับปรุงการพยากรณ์โรคเบาหวาน งานวิจัยนี้ผู้วิจัยมุ่งหวังที่ผลลัพธ์จากการศึกษานี้จะช่วยให้มีความเข้าใจมากขึ้นเกี่ยวกับวิธีการคัดเลือกคุณลักษณะที่เหมาะสมที่สุด สำหรับการคัดกรองโรคเบาหวาน และเป็นข้อมูลที่มีคุณค่าในการพัฒนาแนวทางในการจัดการและการรักษาโรคนี้ในอนาคต นอกจากนี้งานวิจัยนี้ยังมีเป้าหมายที่จะช่วยเพิ่มประสิทธิภาพของการคัดกรองโรคเบาหวานเพื่อลดความเสี่ยงและสามารถรักษาโรคเบาหวานเพื่อลดความเสี่ยงและสามารถรักษาโรคในระยะต้น ๆ ได้อย่างมีประสิทธิภาพ

## วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญในการพยากรณ์โรคเบาหวาน
2. เพื่อพัฒนากระบวนการการคัดกรองโรคเบาหวานในกลุ่มที่ยังไม่ครอบคลุมทุกกลุ่มอายุ
3.เปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน

## วิธีดำเนินการวิจัย

งานวิจัยนี้เป็นการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน เพื่อการปรับปรุงคุณภาพกระบวนการคัดกรองโรคเบาหวานในกลุ่มที่มีอาการแรกเริ่มของการเป็นโรคเบาหวานเพื่อให้กระบวนการคัดกรองครอบคลุมในประชากรทุกกลุ่มอายุ เป้าหมายของการวิจัยเพื่อการศึกษาวิธีการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน จากการเลือกคุณลักษณะที่สำคัญของอาการของโรคเบาหวานสามารถลดจำนวนมิติของคุณลักษณะของข้อมูลลง เพื่อการพัฒนาไปสู่การใช้งานจริงในกรณีที่มีข้อจำกัดของข้อมูลและเวลาในการประมวลผล

## ขั้นตอนการศึกษา

การศึกษาปัจจัยที่ส่งผลต่อการเป็นโรคเบาหวานพบว่าปัจจัยที่ส่งผลให้เป็นโรคเบาหวานมีอยู่หลายปัจจัยที่สามารถสรุปได้ดังนี้

**อายุ** เป็นปัจจัยหนึ่งที่เกี่ยวข้องกับการเป็นโรคเบาหวานพบว่าผู้ป่วยที่อายุมากขึ้นมีปัจจัยเสี่ยงทีจะเป็นโรคเบาหวาน

**เพศ** เป็นปัจจัยหนึ่งที่เกี่ยวข้องกับการเป็นโรคเบาหวานโดยพบว่าเพศชายกับเพศหญิงจะมีความเสี่ยงในการ

เป็นโรคเบาหวานที่แตกต่างกัน

**อาการแรกเริ่มของผู้ป่วยโรคเบาหวาน** เป็นปัจจัยหนึ่งที่เกี่ยวข้องกับการเป็นโรคเบาหวาน พบว่าผู้ป่วยที่ไปพบแพทย์มักไปด้วยอาการเหล่านี้ร่วมกันอาการปัสสาวะบ่อย, อาการกระหายน้ำหิวน้ำบ่อย, อาการ น้ำหนักลดโดยไม่ทราบสาเหตุ, อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง, อาการกินจุ หิวบ่อย, อาการโรคเชื้อราในช่องคลอด, สายตาพร่ามัวมองไม่ชัดเจน , คันตามผิวหนัง, อาการหงุดหงิดง่าย, อาการเป็นแผลง่ายแผลหายยาก, อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน, อาการกล้ามเนื้อหดเกร็ง, อาการการร่วงของผมเป็นหย่อม และอาการสะสมไขมันในส่วนต่างๆ ของร่างกายเกินปกติ

## ศึกษาวิธีการคัดเลือกคุณลักษณะ

**การคัดเลือกคุณลักษณะ (Feature Selection)**

เป็นการลดขนาดหรือมิติของข้อมูลและยังคงคุณลักษณะที่สำคัญของข้อมูลจากการจำแนกประเภทข้อมูล (classification) จะพบว่าบางครั้งจำนวน คุณลักษณะ (Attribute) หรือ ฟีเจอร์ (Feature) นั้นมีจำนวน ซึ่งจำเป็นต้องทำการคัดเลือกคุณลักษณะ ที่สำคัญมาใช้งาน ขั้นตอนนี้เรียกว่าการคัดเลือกคุณลักษณะ (Feature Selection) โดยสามารถแบ่งได้เป็น 2 กลุ่มดังนี้

**1. Filter approach**

เป็นการคัดเลือกฟีเจอร์โดยใช้การคำนวณหาค่าน้ำหนักซึ่งอาจจะเป็นค่าความสัมพันธ์ระหว่างแต่ละฟีเจอร์และคลาสต่างๆและจะเลือกฟีเจอร์โดยเรียงลำดับตามค่าน้ำหนักที่คำนวณได้แล้วเลือกฟีเจอร์ที่มีค่าน้ำหนักมากกว่าที่ต้องการมาใช้งานต่อไปวิธีการนี้จะไม่มีการสร้างโมเดลเพื่อคัดเลือกฟีเจอร์เทคนิคในการคำนวณค่าน้ำหนักของฟีเจอร์ต่างๆ มีหลายวิธี เช่น Correlation Based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR), Chi Square, Evolutionary

**1.1 Correlation Based Feature Selection (CFS)**

เป็นการหากลุ่มคุณลกัษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์โดยคุณลักษณะที่ถูกคัดเลือกใช้สำหรับการจำแนกประเภทของข้อมูล

**1.2 Information Gain (IG)**

เป็นการพิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญในการจำแนกกลุ่มได้ดีที่สุด

**1.3 Gain Ratio (GR)**

เป็นเทคนิคเพื่อประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ค่าเกนความรู้ในการแบ่งชุดข้อมูลจะทำให้เกิดความเอนเอียงเกิดขึ้นเมื่อแอทริบิวต์ที่ทำการ พิจารณามีค่าที่เกิดขึ้นเป็นจานวนมาก โดยในการใช้ค่าเกณฑ์ความรู้มักจะทำการเลือกแอทริบิวที่มีค่าที่เกิดขึ้น จากความโน้มเอียงที่อาจเกิดขึ้นจากการใช้ค่าเกนความรู้ การจะลดทอนความเอนเอียงลงด้วยตัวชี้วัดการแบ่งข้อมูลใหม่ เรียกว่า อัตราส่วนเกน (gain ratio) ที่จะประยุกต์ใช้การทำนอร์มัลไลซ์ค่าเกนความรู้ด้วยการใช้ค่า Split Information

**1.4 Chi Square**

เป็นการวัดโดยการใช้สถิติประมาณความสัมพันธ์ร่วมระหว่างคุณลักษณะเฉพาะกับคลาสของคุณลักษณะเฉพาะ การทดสอบผลรวมของสัดส่วนกำลังสองของผลต่าง ระหว่างความถี่ของค่าที่สังเกตุกับค่าความถี่ของค่าคาดหวัง หรือใช้ทดสอบการแจกแจงของข้อมูล มักใช้กับข้อมูลที่แจกแจงแบบไม่ต่อเนื่อง (Discontinuous data)

**1.5 Evolutionary Selection**

เป็นการสุ่มเลือกคุณลักษณะซึ่งเป็นตัวแปรการพยากรณ์เข้ามาในสมการทีละตัวและทำการทดสอบหาประสิทธิภาพในการพยากรณ์คำตอบหากค่าประสิทธิภาพในการคาดการณ์สูงขึ้นจะเก็บคุณลักษณะนั้นไว้ แล้วทำการสุ่มเลือกคุณลักษณะอื่นเข้าไปเพิ่ม หากค่าประสิทธิภาพของโมเดลลงลดลงก็จะตัดคุณลักษณะนั้นออก

**2. Wrapper approach**

เป็นการคัดเลือกฟีเจอร์ด้วยการสร้างโมเดล (Classification model) ขึ้นมาจากเซตของฟีเจอร์ที่กำหนดไว้และวัดประสิทธิภาพการทำงานของโมเดล และเลือกเซตของฟีเจอร์ที่ทำให้โมเดลมีประสิทธิภาพมากที่สุดมาใช้งาน เช่น โมเดลที่ให้ค่าความถูกต้อง (Accuracy) มากที่สุดการคัดเลือกฟีเจอร์ด้วยวิธีการนี้แบ่งย่อยได้เป็น 2 แบบ

**2.1 Forward Selection**

เป็นการเพิ่มคุณลักษณะครั้งละหนึ่งคุณลักษณะถ้าคุณลักษณะที่ใส่เพิ่มให้ประสิทธิภาพของโมเดลที่ดีจะถูกเลือกเก็บไว้และเลือกคุณลักษณะอื่นๆมาเพิ่มต่อจนประสิทธิภาพของโมเดลไม่มีการเพิ่มขึ้นจากเดิมจึงหยุดทำงาน

**2.2 Backward Elimination**

เป็นการตัดคุณลักษณะออกโดยเริ่มต้นจากคุณลักษณะทั้งหมดและตัดคุณลักษณะที่ไม่สำคัญออกทีละคุณลักษณะถ้าคุณลักษณะที่ตัดออกให้ประสิทธิภาพของโมเดลที่ดีให้ตัดคุณลักษณะอื่นๆต่อไปจนกว่าประสิทธิภาพของโมเดลลดลงอย่างมีนัยสำคัญ

**ศึกษาวิธีการวัดประสิทธิภาพตัวแบบการพยากรณ์**

การวัดประสิทธิภาพตัวแบบการพยากรณ์โดยใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling ซึ่งประกอบด้วยค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าประสิทธิภาพโดยรวม (F–Measure) และค่าความถูกต้อง (Accuracy) ซึ่งมีค่าอยู่ระหว่าง 0 - 1 ซึ่ง 1 หมายถึง ประสิทธิภาพดี สามารถคำนวณค่า Accuracy หรือ ค่าความถูกต้องที่แสดงถึงประสิทธิภาพของ Model ที่โมเดลทายถูกทั้งหมดผลการพยากรณ์แบบ True Positive และ True Negative ดังสมการด้านล่าง

$$\text{Accurancy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive False Negative}}$$

สามารถคำนวณค่า Precision หรือ ค่าความแม่นยำที่แสดงถึงประสิทธิภาพของ Model ที่แสดงถึงประสิทธิภาพของ Model เมื่อผลการพยากรณ์แบบ False Positive เช่น ในกรณีของการพยากรณ์ว่าเป็น เบาหวาน หรือไม่เป็นเบาหวาน ผลการพยากรณ์แบบ False Positive หมายถึงกลุ่มคนที่เป็นปกติแต่พยากรณ์ว่าเป็นเบาหวานเป็นต้น

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

สามารถคำนวณค่า Recall หรือ Sensitivity หรือ True Positive Rate ที่แสดงถึงประสิทธิภาพของ Model ผลการพยากรณ์แบบ False Negative ดังสมการด้านล่าง

$$\text{Recall หรือ Sensitivity หรือ True Positive Rate} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

สามารถคำนวณค่า F1-score เพื่อวัดประสิทธิภาพของ Model จากค่า F1-score ซึ่งเป็นค่าเฉลี่ยฮาร์โมนิก (Harmonic Mean) ของ Precision และ Recall

$$\text{F1-score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

**ศึกษาวิธีการทดสอบประสิทธิภาพของโมเดล**

การวัดค่าประสิทธิภาพของเทคนิควิธีต่างๆ จะต้องทำการเลือกข้อมูลสาหรับเรียนรู้ (Training Set) และ ข้อมูลสาหรับทดสอบ (Testing Set) ในงานวิจัยนี้เลือกใช้วิธีสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง k กลุ่ม (k-Fold Cross Validation) โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วนๆ ให้เท่าๆ กัน ต่อจากนั้นให้นาข้อมูลบางส่วนมาทำการ เรียนรู้ และนำข้อมูลบางส่วนมาทำการทดสอบแบบจำลองที่ได้จากการเรียนรู้ โดยในการทำงานจะทาการเลือกสุ่มข้อมูล ออกเป็น k ชุดที่เท่าๆ กัน ในการทดลองครั้งแรก ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุด เรียนรู้ ในการทดลองครั้งที่ 2 ข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ทำจนกระทั่ง ข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบและข้อมูลชุดเรียนรู้ ซึ่งจะมีการทดลองทั้งหมด k ครั้ง ในงานวิจัยนี้ได้ เลือกใช้ค่า k = 10

**แผนการดำเนินงาน**

การเตรียมข้อมูลโรคเบาหวาน ผู้วิจัยได้คัดเลือกชุดข้อมูลความเสี่ยงเบาหวานระยะเริ่มต้น (Early stage diabetes risk prediction dataset) ข้อมูลที่รวบรวมมาจากโรงพยาบาลโรคเบาหวานซิลเฮต (Sylhet Diabetic Hospital) ในประเทศบังกลาเทศ ซึ่งมีประชาการอยู่อาศัยในเมืองราว 20 ล้านคน สาเหตุที่เลือกใช้ชุดข้อมูลความเสี่ยงเบาหวานระยะ

เริ่มต้นจากโรงพยาบาลโรคเบาหวานซิลเฮตประเทศบังกลาเทศเนื่องจากเป็นโรงพยาบาลที่ให้บริการดูแลผู้ป่วยเบาหวานโดยเฉพาะซึ่งให้บริการดูแลรักษาผู้ป่วยป่วยมาแล้วไม่น้อยกว่า 77,000 รายและมีแนวโน้มเพิ่มสูงขึ้นอย่างต่อเนื่องและมีการจัดทำโปรแกรมการรับรู้การป้องกันโรคเบาหวานและค่ายสุขภาพและจากวัฒนธรรมการกินอาหารประเภทข้าวผักและแป้งที่คล้ายคลึงกับประเทศไทยซึ่งมีข้อมูลที่เหมาะสมสำหรับนำมาดำเนินการวิจัย สำหรับแหล่งที่มาของข้อมูลผู้วิจัยได้เลือกใช้จากแหล่งข้อมูลของ UCI Machine Learning Repository ซึ่งเป็นเว็บไซต์ที่ รวบรวมฐานข้อมูล ทฤษฎีโดเมน สำหรับการวิเคราะห์เชิงประจักษ์ของอัลกอริทึมการเรียนรู้ของเครื่อง ถูกใช้งานโดยนักเรียนนักศึกษานักการศึกษาและนักวิจัยทั่วโลกได้ใช้กันอย่างแพร่หลายในฐานะแหล่งข้อมูลหลักของชุดข้อมูล การเรียนรู้ของเครื่อง ติดตั้งโปรแกรม RapidMiner สำหรับใช้วิเคราะห์ชุดข้อมูล งานวิจัยนี้ได้เสนอวิธีการทำโดยใช้โปรแกรม RapidMiner Studio Version 9.10 นำชุดข้อมูลที่มาวิเคราะห์ประเภทของข้อมูลที่อยู่ในแต่ละคุณลักษณะ

งานวิจัยนี้เป็นการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน เพื่อการปรับปรุงคุณภาพกระบวนการคัดกรองโรคเบาหวานในกลุ่มที่มีอาการแรกเริ่มของการเป็นโรคเบาหวานเพื่อให้กระบวนการคัดกรองครอบคลุมในประชากรทุกกลุ่มอายุ เป้าหมายของการวิจัยเพื่อการศึกษาวิธีการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน จากการเลือกคุณลักษณะที่สำคัญของอาการของโรคเบาหวานสามารถลดจำนวนมิติของคุณลักษณะของข้อมูลลง เพื่อการพัฒนาไปสู่การใช้งานจริงในกรณีที่มีข้อจำกัดของข้อมูลและเวลาในการประมวลผล งานวิจัยนี้สามารถแบ่งตามประเด็นได้เป็น 7 ส่วน ดังต่อไปนี้

**ส่วนที่ 1 ลักษณะของกลุ่มตัวอย่าง**

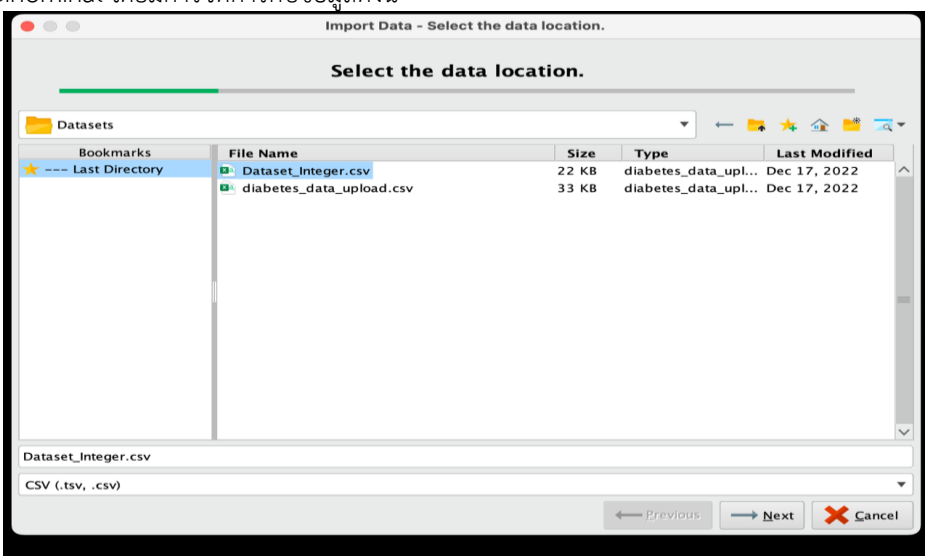
ผู้วิจัยได้คัดเลือกชุดข้อมูลความเสี่ยงเบาหวานระยะเริ่มต้น (Early stage diabetes risk prediction dataset) เป็นข้อมูลที่รวบรวมมาจากโรงพยาบาลโรคเบาหวานซิลเฮต (Sylhet Diabetic Hospital) ในประเทศบังกลาเทศ มีจำนวน 520 คน ซึ่งมีอาการของโรคเบาหวานที่พบบ่อยจำแนกได้เป็นคุณลักษณะที่ใช้สำหรับการดำเนินงานวิจัยดังต่อไปนี้

**ตารางที่ 1 รายละเอียดคุณลักษณะที่ใช้สำหรับการดำเนินงานวิจัย**

| ลำดับ | แอททริบิวต์ | ชนิดข้อมูล | คำอธิบาย |
|---|---|---|---|
| 1 | Age | Integer | อายุ |
| 2 | Gender | Polynomial | เพศ |
| 3 | Polyuria | Polynomial | อาการปัสสาวะบ่อย |
| 4 | Polydipsia | Polynomial | อาการกระหายน้ำหิวน้ำบ่อย |
| 5 | Sudden weight loss | Polynomial | อาการน้ำหนักลดโดยไม่ทราบสาเหตุ |
| 6 | Weakness | Polynomial | อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง |
| 7 | Polyphagia | Polynomial | อาการกินจุ หิวบ่อย |
| 8 | Genital thrush | Polynomial | อาการโรคเชื้อราในช่องคลอด |
| 9 | Visual blurring | Polynomial | สายตาพร่ามัวมองไม่ชัดเจน |
| 10 | Itching | Polynomial | คันตามผิวหนัง |
| 11 | Irritability | Polynomial | อาการหงุดหงิดง่าย |
| 12 | Delayed healing | Polynomial | อาการเป็นแผลง่ายแผลหายยาก |
| 13 | Partial paresis | Polynomial | อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน |
| 14 | Muscle stiffness | Polynomial | อาการกล้ามเนื้อหดเกร็ง |
| 15 | Alopecia | Polynomial | อาการการร่วงของผมเป็นหย่อม |
| 16 | Obesity | Polynomial | อาการสะสมไขมันในส่วนต่างๆ ของร่างกายเกินปกติ |
| 17 | Class | Polynomial | การเป็นโรคเบาหวาน |

**ส่วนที่ 2 ปรับรูปแบบของข้อมูลให้เหมาะสมสำหรับนำไปใช้วิเคราะห์ข้อมูล**

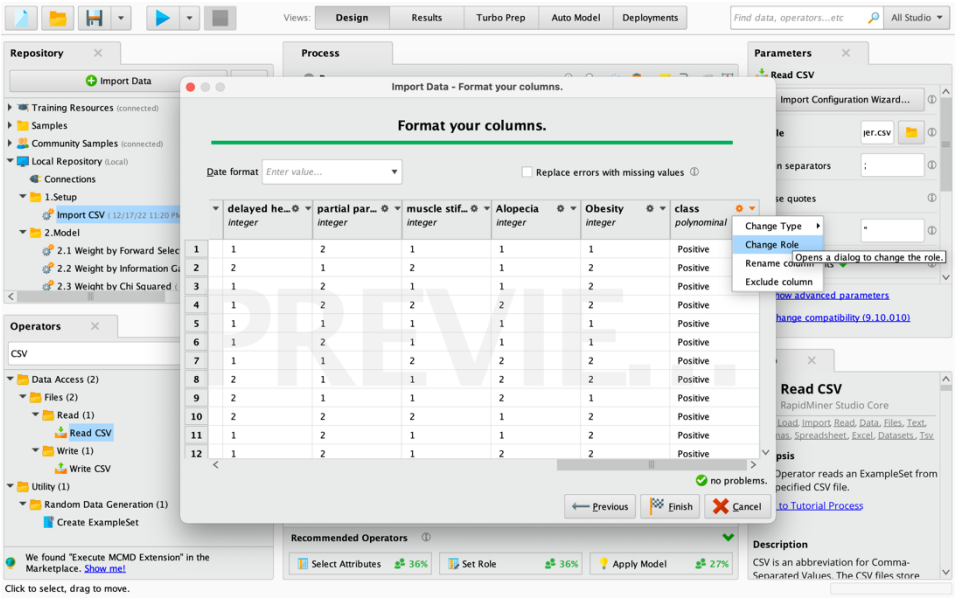
ในการวิจัยนี้มีการเตรียมข้อมูลในรูปแบบ Excel นามสกุล (.xlsx) เป็นหลัก โดยนำข้อมูลตั้งต้นมากำหนดค่าให้เหมาะสมก่อนนำเข้าโปรแกรม RapidMiner ดังนี้ กำหนดค่าให้เหมาะสมกับการวัดประสิทธิภาพโมเดลด้วยเทคนิค ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) โดยแปลงประเภทของข้อมูลที่เก็บไว้ในแต่ละคุณลักษณะจาก Polynomial เป็นประเภท Binominal โดยมีการจัดการกับข้อมูลดังนี้




**ภาพที่ 1** นำเข้าข้อมูลโดยใช้โปรแกรม RapidMiner Studio

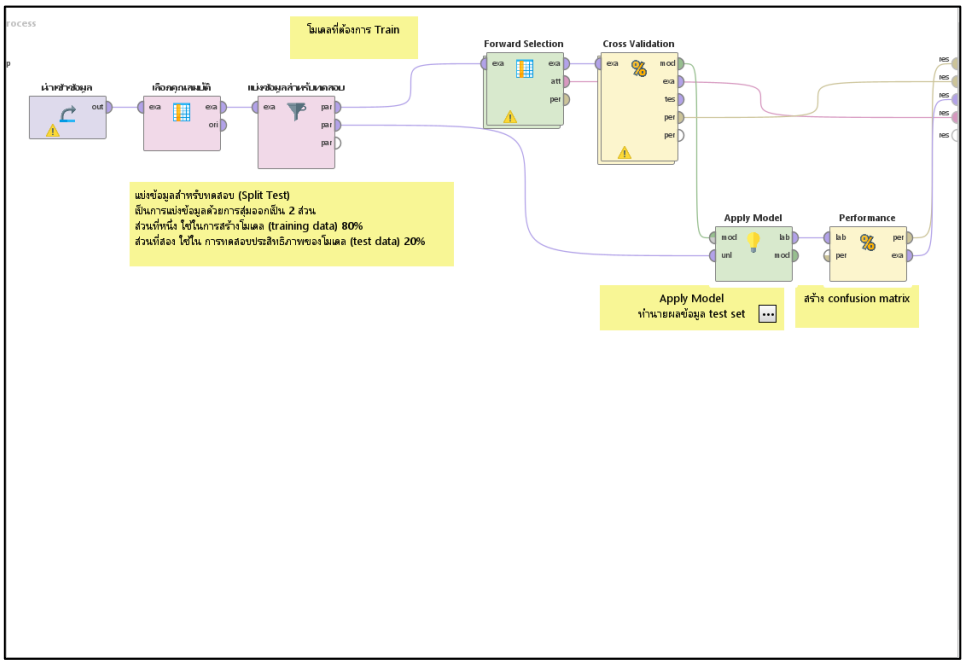
**ส่วนที่ 3 กำหนดชนิดของคุณลักษณะสำหรับใช้เป็นคำตอบ**

การวัดประสิทธิ์ภาพการพยากรณ์จำเป็นต้องมีการสร้างคุณลักษณะที่เป็นคำตอบที่จะต้องการจะสร้างโมเดลขึ้นมาพยากรณ์ หรือ เรียกว่า คลาส (Class) หรือตัวแปรตาม (Dependent variable)




ภาพที่ 2 เลือกคุณลักษณะสำหรับใช้เป็นคำตอบ

**ส่วนที่ 4 การสร้างโมเดล (Modeling)**




ภาพที่ 3 สร้างโมเดล

การสร้างโมเดลทำการกำหนดคุณสมบัติต่างๆ ในหน้าต่าง Process ดังนี้

4.1 ส่วนนำเข้าข้อมูล (Data set)

4.2 ส่วนเลือกคุณสมบัติ (Feature)

4.3 ส่วนแบ่งข้อมูล (Split Data) ทำการการแบ่งข้อมูลเป็นสองส่วน กำหนดให้ข้อมูลส่วนที่ 1 จำนวน 80% เป็นข้อมูลสำหรับสร้างโมเดล (Training Set) และข้อมูลส่วนที่ 2 จำนวน 20% เป็นข้อมูลสำหรับทดสอบประสิทธิภาพของโมเดล (Testing Set) กำหนดเงื่อนไขสำหรับการสุ่มตัวอย่างสำหรับใช้ทดสอบ ให้กำหนด local random seed = 1992

**ส่วนที่ 5 วัดประสิทธิภาพความแม่นยำของการพยากรณ์**

การวัดประสิทธิภาพความแม่นยำของการพยากรณ์จำเป็นต้องทดสอบความถูกต้องของโมเดลวิเคราะห์ข้อมูล ด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) เพื่อตรวจสอบความประสิทธิภาพความแม่นยำของ การพยากรณ์การเป็นโรคเบาหวานใช้ Confusion Matrix มาคำนวณ การประเมินประสิทธิภาพของการพยากรณ์ปัจจัยที่ใช้ในการวัดประสิทธิภาพประกอบด้วย

**ความแม่นยำของข้อมูล (Precision)**

ค่าที่พยากรณ์ว่าเป็นโรคเบาหวานได้ถูกต้อง (True Positive) เทียบกับค่าที่พยากรณ์ว่าเป็นโรคเบาหวานได้ถูกต้อง (True Positive) และไม่ถูกต้อง (False Negative)

**Recall (ความถูกต้องของการพยากรณ์)**

ค่าความถูกต้องของการพยากรณ์ว่าจะเป็น จริง (True Positive) เทียบกับ จำนวนครั้งของเหตุการณ์ ที่เกิดขึ้นว่า จริง (True Positive) และ สิ่งที่พยากรณ์ไม่ตรงกับที่เกิดขึ้นจริง (False Negative) อธิบายได้ดังนี้ ค่าของการพยากรณ์ว่าเป็นโรคเบาหวาน โมเดลพยากรณ์ว่า เป็นโรคเบาหวาน เทียบกับ ค่าของการพยากรณ์ว่าเป็นโรคเบาหวาน โมเดลพยากรณ์ว่าเป็นโรคเบาหวาน กับ ค่าของการพยากรณ์ว่าเป็นโรคเบาหวาน โมเดลพยากรณ์ว่า ไม่เป็นโรคเบาหวาน

**ความถูกต้อง (Accuracy)** คือ การวัดความถูกต้องของการพยากรณ์ถูก โดยพิจารณารวมทุกกรณี อธิบายได้ดังนี้ โมเดลพยากรณ์ว่าเป็นโรคเบาหวานหรือไม่เป็นโรคเบาหวานได้ถูกต้องทั้งหมด เทียบกับค่าที่โมเดลพยากรณ์ว่าเป็นโรคเบาหวานถูกต้องและไม่ถูกต้อง

**ความถ่วงดุล (F-measure)** คือ ค่าเฉลี่ยของความแม่นยำและความถูกต้องของข้อมูลการพยากรณ์

**True Positive** คือ สิ่งที่การพยากรณ์ตรงกับสิ่งที่เกิดขึ้นจริง อธิบายได้ดังนี้ เป็นโรคเบาหวาน โมเดลพยากรณ์ว่าเป็นโรคเบาหวาน

**True Negative** คือ สิ่งที่การพยากรณ์ตรงกับสิ่งที่เกิดขึ้นจริง อธิบายได้ดังนี้ ไม่เป็นโรคเบาหวาน โมเดลพยากรณ์ว่า ไม่เป็นโรคเบาหวาน

**False Positive** คือ สิ่งที่การพยากรณ์ไม่ตรงกับสิ่งที่เกิดขึ้นจริง อธิบายได้ดังนี้ ไม่เป็นโรคเบาหวาน โมเดลพยากรณ์ว่า เป็นโรคเบาหวาน

**False Negative** คือ สิ่งที่การพยากรณ์ไม่ตรงกับสิ่งที่เกิดขึ้น อธิบายได้ดังนี้ เป็นโรคเบาหวาน โมเดลพยากรณ์ว่า ไม่เป็นโรคเบาหวาน

**ส่วนที่ 6 การประเมินผล**

จากการดำเนินการลดคุณลักษณะของข้อมูลโดยใช้เทคนิควิธีต่าง ๆ การประเมินผล (Evaluation) เป็นขั้นตอนของการนำผลลัพธ์ที่ได้จากกระบวนการพยากรณ์มาทำการทดสอบความถูกต้องแม่นยำและความน่าเชื่อถือของโมเดล โดยงานวิจัยครั้งนี้นำข้อมูลจากโมเดลการพยากรณ์ที่ผ่านกระบวนการลดมิติของข้อมูลโดยใช้การเลือกจากค่า Weight ที่มีความสัมพันธ์กันของข้อมูลน้อยที่สุดออกที่ละค่า และผ่านการทดสอบประสิทธิภาพการพยากรณ์ของโมเดลด้วยวิธีการแบ่งข้อมูลออกเป็น 2 ชุด กำหนดให้ข้อมูลส่วนที่ 1 จำนวน 80% เป็นข้อมูลสำหรับสร้างโมเดล (Training Set) และข้อมูลส่วนที่ 2 จำนวน 20% เป็นข้อมูลสำหรับทดสอบประสิทธิภาพของโมเดล (Testing Set) การประเมินผลทำโดยการวัดประสิทธิภาพของการพยากรณ์โดยใช้ Confusion Matrix แสดงผลได้ดังตารางต่อไปนี้

**ตารางที่ 2 ผลประสิทธิภาพการพยากรณ์ของโมเดล หลังการลดจำนวนคุณลักษณะจากชุดข้อมูลสำหรับทดสอบประสิทธิภาพของโมเดล (Testing Set)**

| โมเดล | จำนวนคุณลักษณะ | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 1. Forward Selection | 3 | 90.38% | 92.19% | 92.19% | 92.19% |
| 2. Information Gain Ratio (GR) | 10 | 93.27% | 92.54% | 96.88% | 94.66% |
| 3. Chi Square | 10 | 93.27% | 92.54% | 96.88% | 94.66% |
| 4. Correlation Based Feature Selection (CFS) | 13 | 94.23% | 93.94% | 96.88% | 95.38% |
| 5. Information Gain (IG) | 13 | 94.23% | 93.94% | 96.88% | 95.38% |
| 6. Evolutionary Selection | 8 | 95.19% | 98.36% | 93.75% | 96.00% |
| 7. Backward Elimination | 16 | 94.23% | 95.31% | 95.31% | 95.31% |

**ส่วนที่ 7 สรุปผลการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน**

สามารถสรุปผลประสิทธิภาพการพยากรณ์ของโมเดล จากการปรับปรุงการพยากรณ์โรคเบาหวาน ผลการเปรียบเทียบการพยากรณ์ของโมเดล จากการเปรียบเทียบค่าความน่าเชื่อถือของโมเดลหลังลดจำนวนคุณลักษณะลง ซึ่งได้ค่าความน่าเชื่อถือของโมเดลไม่น้อยกว่าค่าเดิมก่อนดำเนินการลดจำนวนคุณลักษณะ

**ตารางที่ 3 ผลประสิทธิภาพการพยากรณ์ของโมเดล ก่อนและหลังการลดจำนวนคุณลักษณะ**

| โมเดล | ก่อนลดจำนวนคุณลักษณะ | | หลังลดจำนวนคุณลักษณะ | |
|---|---|---|---|---|
| | จำนวนคุณลักษณะ | F1-Score | จำนวนคุณลักษณะ | F1-Score |
| 1. Forward Selection | 16 | 91.59% | 3 | 92.19% |
| 2. Information Gain Ratio (GR) | 16 | 94.19% | 10 | 94.66% |

| โมเดล | ก่อนลดจำนวนคุณลักษณะ | | หลังลดจำนวนคุณลักษณะ | |
|---|---|---|---|---|
| | จำนวนคุณลักษณะ | F1-Score | จำนวนคุณลักษณะ | F1-Score |
| 3. Chi Square | 16 | 94.19% | 10 | 94.66% |
| 4. Correlation Based Feature Selection (CFS) | 16 | 94.19% | 13 | 95.38% |
| 5. Information Gain (IG) | 16 | 94.19% | 13 | 95.38% |
| 6. Evolutionary Selection | 16 | 94.44% | 13 | 94.49% |
| 7. Backward Elimination | 16 | 93.68% | 16 | 95.31% |

## ผลการวิจัย

จากการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวานพบว่า โมเดล Forward Selection ให้ประสิทธิภาพในการลดจำนวนมิติของข้อมูลได้มากที่สุดจากการลดจำนวนมิติของคุณลักษณะของโรคเบาหวานลงเหลือ 3 คุณลักษณะ จากคุณลักษณะของโรคเบาหวานทั้งหมด 16 คุณลักษณะ โดยยังคงความถูกต้องในการพยากรณ์ไว้ ได้ผลประสิทธิภาพการพยากรณ์โรคเบาหวาน เป็นร้อยละ 92.19

## อภิปรายผลการวิจัย

งานวิจัยนี้ผู้วิจัยพบว่าวิธีการคัดเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับการคัดกรองโรคเบาหวาน หลังจากการลดจำนวนมิติของคุณลักษณะลงคุณลักษณะของโรคเบาหวานโดยทั้งหมด 16 คุณลักษณะ มีคุณลักษณะแต่ละคุณลักษณะที่มีค่าน้ำหนัก (Weight) ที่ตัวแปรต้นที่ส่งผลต่อตัวแปรเป้าหมายที่ผู้วิจัยศึกษา ในงานวิจัยนี้คือ ผลการพยากรณ์การเป็นโรคเบาหวานและไม่เป็นโรคเบาหวาน พบว่าข้อมูลคุณลักษณะต่าง ๆ ที่นำมาวิเคราะห์จะมีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์การเป็นโรคเบาหวานและไม่เป็นโรคเบาหวานพบว่าคุณลักษณะที่มีค่าน้ำหนัก (Weight) ที่ส่งผลต่อตัวแปรเป้าหมายที่พบในแต่ละโมเดลเรียงตามค่าน้ำหนักจากมีค่ามากที่สุดไปน้อยที่สุดสามารถแจกแจงได้ดังนี้

โมเดล Forward Selection

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 3 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ เพศ (Gender), อาการปัสสาวะบ่อย (Polyuria), อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia)

โมเดล Information Gain Ratio (GR)

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 10 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), อาการปัสสาวะบ่อย (Polyuria), เพศ (Gender), อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis), อายุ (Age), อาการหงุดหงิดง่าย (Irritability), อาการกินจุ หิวบ่อย (Polyphagia), อาการการร่วงของผมเป็นหย่อม (Alopecia), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring)

โมเดล Chi Square

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 10 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), , อาการปัสสาวะบ่อย (Polyuria), เพศ (Gender), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis), อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss), อาการกินจุ หิวบ่อย (Polyphagia), อาการการร่วงของผมเป็นหย่อม (Alopecia), อาการหงุดหงิดง่าย (Irritability), อายุ (Age), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring)

โมเดล Correlation Based Feature Selection (CFS)

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 13 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), , อาการปัสสาวะบ่อย (Polyuria), เพศ (Gender), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis), อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss), อาการกินจุ

หิวบ่อย (Polyphagia), อาการการร่วงของผมเป็นหย่อม (Alopecia), อาการหงุดหงิดง่าย (Irritability), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring), อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง (Weakness), อาการกล้ามเนื้อหดเกร็ง (Muscle stiffness), อาการโรคเชื้อราในช่องคลอด (Genital thrush), อายุ (Age)

โมเดล Information Gain (IG)

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 13 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), , อาการปัสสาวะบ่อย (Polyuria), เพศ (Gender), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis), อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss), อาการกินจุ หิวบ่อย (Polyphagia), อาการการร่วงของผมเป็นหย่อม (Alopecia), อาการหงุดหงิดง่าย (Irritability), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring), อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง (Weakness), อายุ (Age), อาการกล้ามเนื้อหดเกร็ง (Muscle stiffness), อาการโรคเชื้อราในช่องคลอด (Genital thrush)

โมเดล Evolutionary Selection

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 13 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อายุ (Age), เพศ (Gender), อาการปัสสาวะบ่อย (Polyuria), อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss), อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง (Weakness), อาการกินจุ หิวบ่อย (Polyphagia), อาการโรคเชื้อราในช่องคลอด (Genital thrush), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring), อาการคันตามผิวหนัง (Itching), อาการหงุดหงิดง่าย (Irritability), อาการเป็นแผลง่ายแผลหายยาก
(Delayed healing), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis)

โมเดล Backward Elimination

จากการลดจำนวนมิติของคุณลักษณะลงเหลือ 16 คุณลักษณะ พบว่าคุณลักษณะ ที่มีค่าน้ำหนักที่ส่งผลต่อการพยากรณ์มีดังนี้ อายุ (Age), เพศ (Gender), อาการปัสสาวะบ่อย (Polyuria), อาการกระหายน้ำหิวน้ำบ่อย (Polydipsia), อาการอ่อนเพลียเหนื่อยง่ายไม่มีแรง (Weakness), อาการกินจุ หิวบ่อย (Polyphagia), อาการโรคเชื้อราในช่องคลอด (Genital thrush), สายตาพร่ามัวมองไม่ชัดเจน (Visual blurring), อาการคันตามผิวหนัง (Itching), อาการหงุดหงิดง่าย (Irritability), อาการเป็นแผลง่ายแผลหายยาก (Delayed healing), อาการเหน็บชาหรือภาวะอัมพาตเฉพาะส่วน (Partial paresis), อาการกล้ามเนื้อหดเกร็ง (Muscle stiffness), อาการการร่วงของผมเป็นหย่อม (Alopecia), อาการสะสมไขมันในส่วนต่างๆ ของร่างกายเกินปกติ (Obesity) ,อาการน้ำหนักลดโดยไม่ทราบสาเหตุ (Sudden weight loss)

## ข้อเสนอแนะ

งานวิจัยครั้งนี้ผู้วิจัยได้ทำการศึกษาการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์โรคเบาหวาน โดยมีข้อเสนอเพิ่มเติมดังนี้

1. ต้นแบบของการการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญการพยากรณ์โรคเบาหวาน สามารถนำไปประยุกต์ใช้ในการพัฒนาและออกแบบซอฟต์แวร์ที่ใช้งานในการคัดกรองโรคได้จริง

2. การกำหนดคุณลักษณะของการพยากรณ์โรคเบาหวาน อาจมีการนำผลของวิธีที่ได้ผลลัพธ์ดีที่สุดมาพัฒนาคัดกรองผู้ป่วยแรกเริ่มโรคเบาหวานสำหรับบุคลากรทางการแพทย์ได้อย่างมีศักยภาพ

## เอกสารอ้างอิง

อัจจิมา มณฑาพันธุ์. (2562). *การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม.* ผลงานวิจัย คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม.

ภานุวัฒน์ เมฆะ, พฤติพงศ์ มุสิกอง, ณัฐภาส ผลากอง, พาสน์ ปราโมกข์ชน และ พยุงศักดิ์ เกษมสำราญ. (2566). การเปรียบเทียบประสิทธิภาพของโมเดลจำแนกภาพสำหรับโรคใบข้าวโพด. *วารสารแม่โจ้เทคโนโลยีสารสนเทศและนวัตกรรม,* 9(2), 1-16.

กฤตกนก ศรีพิมพ์สอ และ กิตติพล วิแสง. (2566). การพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล. *วารสารวิชาการการจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม,* 10(1), 51-63.

กองโรคไม่ติดต่อ กรมควบคุมโรค กระทรวงสาธารณสุข. (2564). สืบค้นจาก http://www.thaincd.com/2016/mission/documents.php?tid=32&gid=1-020

สมาคมโรคเบาหวานแห่งประเทศไทย. (2564). *วารสารเบาหวาน,* 53(1).

นงเยาว์ ในอรุณ. (2564). การเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ความเสี่ยงโรคหัวใจ และหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม*, 40 (2), 137-147.

สาธิมา สุระธรรม และ พรศรี ศรีอัษฎาพร. (2564). การให้ความรู้และสนับสนุนช่วยเหลือเพื่อการจัดการตนเองของผู้เป็นเบาหวาน: กรณีศึกษาวัยรุ่นเบาหวานชนิดที่ 1. *วารสารพยาบาลสภากาชาดไทย,* 14(2), 25-36.

Early stage diabetes risk prediction dataset. (2020). *UCI Machine Learning Repository*. Retrieved from https://doi.org/10.24432/C5VG8H.

รณิดา เตชะสุวรรณา, สุทัศน์ โชตนะพันธ์, กนิษฐา จำรูญสวัสดิ์, บัณฑิต ศรไพศาล และ ประวิช ตัญญสิทธิสุนทร. (2563). ปัจจัยเสี่ยงต่อการเกิดโรคเบาหวานชนิดที่สองในคนไทย. *วารสารควบคุมโรค*, 46(3), 268-279.

สำเภา แก้วโบราณ, นิภาวรรณ สามารถกิจ และ เขมารดี มาสิงบุญ. (2562). ปัจจัยพยากรณ์พฤติกรรมป้องกันโรคเบาหวานในวัยรุ่นที่มีภาวะเสี่ยงต่อโรคเบาหวาน ในจังหวัดสมุทรปราการ. *วารสารการพยาบาลและการดูแลสุขภาพ,* 37(2), 218-27.

ธวัชระพงษ์ วงศ์สกุล. (2563). การกำหนดค่าน้ำหนักหลักเกณฑ์เพื่อการตัดสินใจ. *วารสารวิชาการเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏบุรีรัมย์*, 1(2), 63-71.

วรพรรณ เจริญขา. (2563). การตรวจสอบค่านอกเกณฑ์ในตัวอย่างสุ่มจากประชากรที่มีการแจกแจงปรกติโดยใช้สัมประสิทธิ์ความเบ้. *วารสารวิทยาศาสตร์บูรพา,* 25(1), 236-245.

พนิดา สมบัติมาก, ภัสสร จันทร์หอม, ศุภกร รัศมี, โอฬาร รุ่งมณีธรรมคุณ และ สายชล สินสมบูรณ์ทอง. (2562). การเปรียบเทียบประสิทธิภาพในการจำแนกเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล. *วารสารวิทยาศาสตร์และเทคโนโลยี*, 27(6), 975-985.

กรมควบคุมโรค กระทรวงสาธารณสุข. (2562). *แนวทางการเฝ้าระวังโรค ของกระทรวงสาธารณสุข*. กรุงเทพฯ: องค์การรับส่งสินค้าและวัสดุภัณฑ์.

กระทรวงสาธารณสุข. (2561). *รายงานผลการสำรวจพฤติกรรมเสี่ยงโรคไม่ติดต่อและการบาดเจ็บ พ.ศ. 2561*. นทบุรี.

ธีร์ธวัช แก้ววิจิตร, นิตยา เกิดประสพ และ กิตติศักดิ์ เกิดประสพ. (2559). การเพิ่มประสิทธิภาพซัพพอร์ตเวกเตอร์รีเกรสชันในการพยากรณ์อนุกรมเวลา. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม,* 36(4), 452-458.

จิราพร เดชมา, วนิดา ดุรงค์ฤทธิชัย และ วิชุดา กิจธรธรรม. (2556). การศึกษาปัจจัยพยากรณ์ภาวะแทรกซ้อนให้ผู้เป็นเบาหวานในชุมชนภายใต้ทฤษฎีการพยาบาลของคิง. *วารสารพยาบาลสาธารณสุข,* 27(2), 63-80.

Maeda-Gutiérrez, V., Galván-Tejada, C. E., Cruz, M., Galván-Tejada, J. I., Gamboa-Rosales, H., García-Hernández, A., Luna-García, H., Gonzalez-Curiel, I., & Martínez-Acuña, M. (2021). Risk-Profile and Feature Selection Comparison in Diabetic Retinopathy. *Journal of personalized medicine*, 11(12), 1327.

Kulkarni, A., Chong, D., & Batarseh, Feras A. (2020). 5 - Foundations of data imbalance and solutions for a data democracy, *Data Democracy*, 83-106.

Nai-arun, N., & Moungmai, M. (2020). Diagnostic Prediction Models for Cardiovascular Disease Risk using Data Mining Techniques. *Journal of ECTI Transactions on Computer and Information Technology*, 14(2), 113-121.

Lukmanto, Rian B., Suharjito., Nugroho, A., & Akbar, H. (2019). Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine. *Procedia Computer Science*, 157, 46-54.

Rubaiat, S. Y., Rahman, M. M., & Hasan, M. K. (2018). Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection. *International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh,* 1-6.

Suksawatchon, U., Suksawatchon, J., & Lawang, W. (2018). Health Risk Analysis Expert System for Family Caregiver of Person with Disabilities using Data Mining Techniques. *ECTI Transactions on Computer and Information Technology,* 12(1), 62–72

Assari, R., Azimi, P., & Taghva, M. R. (2017). Heart Disease Diagnosis Using Data Mining Techniques. International. *Journal of Economics & Management Science,* 6(3), 72-79.

Tuso P. (2014). Prediabetes and lifestyle modification: time to prevent a preventable disease. *Perm J. Summer,* 18(3), 88-93.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision tree. *International Journal of Computer Science,* 9(5), 272-278.

Breiman, L. (2001). Random forests. *Journal of Machine Learning*, 5–32.

## Translated Thai References

Montaphan, A. (2019). *Comparison of Feature Selection Methods to Improve Breast Cancer Prediction. Name of Researcher.* School of Information Technology, Sripatum University.

Mekha, P., Musikong, P., Palakong, N., Pramokchon, P., & Kasemsumran, P. (2023). Performance Comparison of Image Classification Models for Corn Leaf Disease. *Maejo Information Technology and Innovation Journal (MITIJ),* 9(2), 1-16.

Sripimsor, K., & Wisaeng, K. (2023). Diabetes Mellitus by Using Data Mining Techniques. *Journal of Technology Management Rajabhat Maha Sarakham University*, 10(1), 51-63.

Division of Non Communicable Disease, Department of Disease Control, Ministry of Public Health. (2021). Retrieved from http://www.thaincd.com/2016/mission/documents.php?tid=32&gid=1-020.

Diabetes Association of Thailand. (2021). *Thai Diabetes Bulletin,* 53(1).

Nai-arun, N. (2021). The Performance Comparison of Cardiovascular Risk Prediction Models using Data Mining Algorithms. *Journal of Science and Technology Mahasarakham University*, 40(2), 137-147.

Suratham, S., & Sriussadaporn, P. (2021), Diabetes Self-Management Education and Support: Case Study of Type 1 Diabetes Adolescents. *Thai Red Cross Nursing Journal,* 14(2), 25-36.

Techasuwanna, R., & et al. (2020). Risk Factors for Developing Type 2 Diabetes in Thai People. *Disease Control Journal,* 46(3), 268-279.

Kaewboran, S., Samartkit, N., & Masingboon, K. (2019). Factors Predicting Diabetes Prevention Behaviors Among Adolescents at Risk for Type 2 Diabetes in Samutprakan Province. *Journal of Nursing and Therapeutic Care*, 37(2), 218-27.

Wongsakul, T. (2020). Determining Weights of Criteria for Decision Making. *Journal of Industrial Technology Buriram Rajabhat University,* 1(2), 63-71.

Jareankam, W. (2020). A Detection of Outliers in Random Sample from Normally Distributed Population Using Coefficient of Skewness. Burapha Science Journal, 25(1) 236-245.

Sombatmak, P., Janhom, P., Ratsamee, S., Rungmaneethummakun, O., & Sinsomboonthong, S. (2019). Performance Comparison of Data Mining's Classification Methods on Data Set with Outliers. *Thai Science and Technology Journal,* 27(6), 975-985.

Department of Disease Control, Ministry of Public Health. (2019). *Disease surveillance of the Ministry of Public Health.* Bangkok: Express Transportation Organization.

Ministry of Public Health. (2018). *Report on the results of a survey of risky behaviors for non-communicable diseases and injuries 2018*. Nonthaburi.

Kaewwijit, T., & et al. (2016). The Improvement of Support Vector Regression to Forecast Time Series. *Journal of Science and Technology Mahasarakham University,* 36(4), 452-458.

Dechma, J., & et al. (2016). The Study of Predictive Factors' Complications of Diabetes Mellitus Client in Community Under King's Nursing Theory. *The Journal of Public Health Nursing,* 27(2), 63-80.