การเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม สำหรับข้อมูลที่มีลักษณะการจำแนกแบบไบนารี

น.ส.กรชนก ชมเชย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

A COMPARISON OF ENSEMBLE FEATURE SELECTION METHODS FOR BINARY CLASSIFICATION DATASETS

Miss Kornchanok Chomchoei

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2022

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่มสำหรับข้อมูลที่มีลักษณะการจำแนกแบบไบนารี
โดย	น.ส.กรชนก ชมเชย
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(ศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช)

..... กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

กรชนก ชมเชย : การเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม สำหรับข้อมูลที่มีลักษณะการจำแนกแบบไบนารี. (A COMPARISON OF ENSEMBLE FEATURE SELECTION METHODS FOR BINARY CLASSIFICATION DATASETS) อ.ที่ปรึกษาหลัก : ผศ. ดร.ณัตติฤดี เจริญรักษ์

งานศึกษานี้เปรียบเทียบวิธีการคัดเลือกตัวแปรแบบเดียว (Single-Feature Selection) และแบบรวมกลุ่ม (Ensemble Feature Selection) ซึ่งแบ่งเป็น 2 รูปแบบคือ รูปแบบการรวมลำดับความสำคัญของตัวแปรแล้วตามด้วยการเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุ (Design CT: Combination followed by Thresholding) และรูปแบบการการเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุแล้วตามด้วยการรวมเซตของตัวแปรที่มีความสำคัญดังกล่าว (Design TC: Thresholding followed by Combination) ผู้ศึกษาได้ใช้การคัดเลือกตัวแปรจากประเภท Filter Wrapper และ Embedded โดยใช้ 10-fold cross validation ในการเปรียบเทียบค่าเฉลี่ยของ F1-score แทนประสิทธิภาพการทำนายและค่าเบี่ยงเบนของ F1-score แทนค่าความเสถียรของการทำนาย ผ่านข้อมูล 3 ชุดได้แก่ Parkinson's Disease dataset (จำนวนตัวแปรต้น(P)=ขนาดข้อมูล(N)), LSVT Voice Rehabilitation dataset (P>N) และ Colon Cancer dataset (P>>N) ใช้ XGBoost เป็นตัวแบบทำนาย จากการศึกษาภายใต้ขอบเขตดังกล่าวพบว่า การคัดเลือกตัวแปรแบบวิธีเดียวด้วย RFE จะให้ผลดีในชุดข้อมูลที่มีมิติมาก P>>N ในเกณฑ์ 2.5% 5% และ 10% แต่การคัดเลือกแบบรวมกลุ่มจะให้ผลการทำนายที่ต่างกันภายใต้ลักษณะมิติของชุดข้อมูลและเกณฑ์ที่เลือกใช้ สำหรับการรวมลำดับความสำคัญของตัวแปรในรูปแบบ Design CT ด้วยค่ากลางและค่าเฉลี่ยเลขคณิตที่เกณฑ์ $\log_2(P)$ จะให้ผลการทำนายดีกว่าวิธีอื่นใน Design CT ในชุดข้อมูล P>>N แต่สำหรับชุดข้อมูล P=N และ P>N ผลการทำนายจากแต่ละวิธีใน Design CT เพิ่มประสิทธิภาพการทำนายเล็กน้อย และสำหรับ Design TC การรวมเซตของตัวแปรต้นที่มีความสำคัญด้วยวิธีอินเตอร์เซกและมัลติอินเตอร์เซกจะให้ผลดีกว่าวิธียูเนียน สำหรับชุดข้อมูล P>>N ในทุกเกณฑ์ การรวมวิธีมัลติอินเตอร์เซกใน $\log_2(P)$ ที่ให้ผลดีกว่าวิธีคัดเลือกแบบอื่น ๆ ในชุดข้อมูล P>>N

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6480386326 : MAJOR STATISTICS

KEYWORD: High-dimensional Data, Binary Classification, Ensemble Feature Selection

Kornchanok Chomchoei : A COMPARISON OF ENSEMBLE FEATURE SELECTION METHODS FOR BINARY CLASSIFICATION DATASETS. Advisor: Asst. Prof. NUTTIRUDEE CHAROENRUK

This research study compares single-feature selection and two ensemble feature selection methods to examine their predictive performance and stability. The first method, called Design Combination followed by Thresholding (Design CT), and the second, named Design Thresholding followed by Combination (Design TC), are selected from the Filter, Wrapper, and Embedded categories of feature selection methods. The study compares the performance (Average F1-score) and stability (Standard deviation F1-score) of these methods using 10-fold cross-validation with three datasets: the Parkinson's Disease (P=N), the LSVT Voice Rehabilitation (P>N), and the Colon Cancer (P>>N), with an XGBoost model used for each dataset. The results can be summarized in three key findings. Firstly, when using single-feature selection, RFE performed well in high-dimensional P>>N dataset at 2.5%, 5% and 10% thresholds. Secondly, the Design CT method, using median and arithmetic mean for combination at $\log_2(P)$ threshold, demonstrated better results than others Design CT methods in P>>N dataset. However, the results of the Design CT method resulted in only small improvements in average F1-scores for P=N and P>N datasets. Thirdly, the Design TC method, employing multi-intersection and intersection methods for combination, consistently provided superior results compared to the union method for P>>N dataset across all thresholds. Multi-intersection at $\log_2(P)$ threshold provided the best result.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study: Statistics Student's Signature .....

Academic Year: 2022 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้เนื่องจากได้รับความกรุณาอย่างสูงจาก ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์ อาจารย์ประจำภาควิชาสถิติ คณะพาณิชย์ศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ซึ่งเป็นอาจารย์ที่ปรึกษางานศึกษา ที่ได้กรุณาให้ความช่วยเหลือ คำแนะนำและชี้แนะแนวทางตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่ ผู้ศึกษาตระหนักถึงความเมตตาของอาจารย์ที่มีให้แก่ผู้ศึกษาเสมอมา จนทำให้ผู้ศึกษาประสบความสำเร็จในการศึกษานี้ ขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ประธานกรรมการ ผู้ช่วยศาสตราจารย์ ดร. นัท กุลวานิช และ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ให้ความอนุเคราะห์สละเวลาอันมีค่าเพื่อให้คำแนะนำ ตรวจสอบและแก้ไขงานวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วง และขอกราบขอบคุณอาจารย์ทุกท่านจากภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความรู้ เอื้อต่อการทำงานศึกษา

อนึ่ง ผู้ศึกษาหวังว่า งานศึกษาฉบับนี้จะมีประโยชน์อยู่ไม่น้อย จึงขอมอบส่วนดีทั้งหมดนี้ให้แก่เหล่าคณาจารย์ ที่ได้ประสิทธิประสาทวิชาจนทำให้งานวิทยานิพนธ์ฉบับนี้เป็นประโยชน์ต่อผู้ที่เกี่ยวข้อง และขอขอบคุณครอบครัวที่คอยผลักดันช่วยเหลือและเป็นกำลังสำคัญยิ่งในทุกด้าน ขอขอบคุณเพื่อน ๆ ที่ให้ข้อคิดเห็น คำปรึกษาเพิ่มเติมตลอดมา

กรชนก ชมเชย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญตาราง

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญรูป

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการวิเคราะห์ข้อมูลบางครั้งอาจเจอปัญหาของมิติข้อมูล (Curse of dimensionality) คือ การที่จำนวนตัวแปรต้นมากกว่าขนาดข้อมูลอยู่มาก (P >> N) (Sara & Hans, 2004; Trevor Hastie, 2009) ปัญหานี้พบมากในการวิเคราะห์ข้อมูลรหัสยีนส์ เช่น ในงานศึกษาของ Hastie & Tibshirani (2003) ใช้ข้อมูลรหัสยีนส์ พบว่าข้อมูลส่วนใหญ่จะมี N อยู่ระหว่าง 50 ถึง 100 และจำนวน P ที่แสดงถึงรหัสยีนส์มากถึง 5,000 ถึง 20,000 ตัวแปร ทำให้ส่งผลเสียต่อความแม่นยำของการทำนาย ดังนั้นจึงจำเป็นต้องลดมิติให้กับข้อมูลซึ่งสามารถทำโดยการคัดเลือกตัวแปรเข้าตัวแบบ

วิธีคัดเลือกตัวแปรเข้าตัวแบบ (Feature selection) แบ่งเป็น 3 ประเภทได้แก่ 1) ประเภท Filter เป็นการนำค่าของข้อมูลมาหาค่าสถิติ อาจนำไปใช้ในการจัดลำดับความสำคัญของตัวแปรต้นจากค่าความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม ด้วยค่าทดสอบทางสถิติ เช่น Correlation coefficient, Mutual Information 2) ประเภท Wrapper เป็นการนำการเรียนรู้ของเครื่อง (Machine learning) หาเซตของตัวแปรต้นที่ทำให้เกิดค่าความผิดพลาดการทำนายที่ต่ำสุด หรือประสิทธิภาพการทำนายสูงสุด เป็นการเรียนรู้แบบทำซ้ำ (The repeated learning steps) รวมถึงหาผลการทำนายของตัวแบบจากการที่หาเซตตัวแปรที่เป็นไปได้ทั้งหมด จึงต้องใช้ระยะเวลาการประมวลผลมากกว่าประเภท Filter 3) ประเภท Embedded การจัดลำดับความสำคัญของตัวแปรต้นจะได้ในระหว่างการฝึกฝนตัวแบบ (Model training) จากนั้นจึงเลือกตัวแปรต้นที่มีลำดับความสำคัญสูงสุด k ลำดับแรกเข้าสู่ตัวแบบ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วิธีคัดเลือกตัวแปรเข้าตัวแบบอาจทำด้วยการเลือกใช้เพียงวิธีเดียวหรือใช้หลายวิธีร่วมกัน Chandrashekar & Sahin (2014) พบว่าการคัดเลือกตัวแปรเข้าตัวแบบด้วยวิธีเดียว เช่น การใช้วิธี Correlation Criteria และ Mutual Information จากประเภท Filter จะทำให้ประสิทธิภาพของการทำนายของตัวแบบดีขึ้นกว่าการที่ไม่ได้ทำการคัดเลือกตัวแปร นอกจากการเลือกใช้เพียงวิธีเดียวแล้ว การคัดเลือกตัวแปรแบบรวมกลุ่มโดยใช้วิธีจาก ประเภท Filter มากกว่า 1 วิธี สามารถพัฒนาประสิทธิภาพการทำนายได้ดีขึ้นกว่าการใช้เพียงแค่วิธีเดียว (Bolón-Canedo et al., 2012; Wang et al., 2010) นอกจากนี้ได้มีงานศึกษาที่นำวิธีการคัดเลือกตัวแปรเข้าตัวแบบรวมกลุ่มจากหลากหลายวิธีมากขึ้น เช่น งานศึกษาของ Effrosynidis & Arampatzis (2021) นำ 12 วิธีการคัดเลือกตัวแปรจากทั้ง 3 ประเภท Filter Wrapper และ Embedded ด้วยตัวแบบ Random

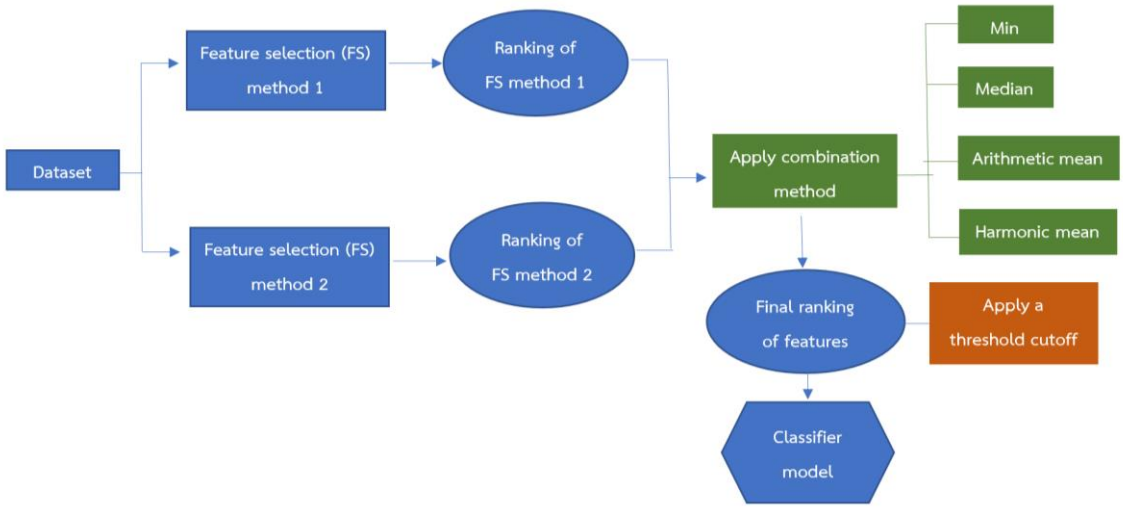
Forest และ LightGBM ผลการศึกษาพบว่า การเลือกใช้วิธีเดียวด้วย SHAP ซึ่งเป็นการคัดเลือกตัวแปรประเภท Wrapper จะให้ผลการทำนายมีความเสถียรที่สุดเมื่อเทียบกับวิธีเดียวอื่น อีกทั้งงานศึกษาดังกล่าวได้นำทั้ง 12 วิธีจาก Filter Wrapper และ Embedded มาคัดเลือกตัวแปรแบบรวมกลุ่ม พบว่าให้ประสิทธิภาพในการทำนายสูงสุดเมื่อเทียบกับวิธีอื่น

ในการคัดเลือกตัวแปรเข้าตัวแบบด้วยวิธีมากกว่า 1 วิธี มีการแบ่งขั้นตอนการรวมผลจากการคัดเลือกตัวแปรแบบรวมกลุ่ม (Aggregation step to combine) เป็น 2 รูปแบบ ดังนี้

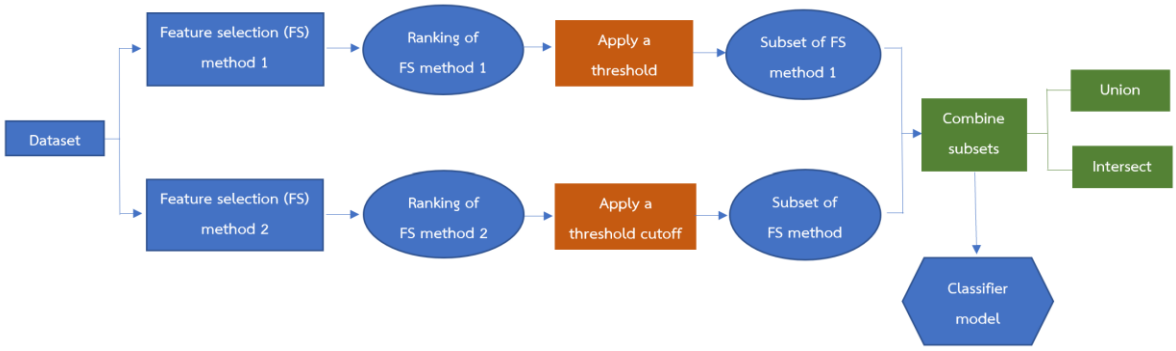
รูปแบบที่ 1 รูปแบบการรวมลำดับความสำคัญของตัวแปรแล้วตามด้วยการเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุ (Combination followed by Thresholding; Design CT) ดังรูปที่ 1 เป็นการคัดเลือกตัวแปรแบบรวมกลุ่มโดยการรวมลำดับความสำคัญของตัวแปรต้นก่อนแล้วจึงกำหนดเกณฑ์คัดเลือกตัวแปรเข้าตัวแบบภายหลัง เช่น เลือกตัวแปรต้นที่มีความสำคัญสูงที่สุดใน 5 ลำดับแรก โดยในขั้นตอนแรกที่รวมลำดับความสำคัญของตัวแปรต้นนั้นได้มีงานศึกษาพิจารณาการรวมลำดับความสำคัญของตัวแปรต้นโดยใช้ค่าต่ำสุด (Min) ค่ากลาง (Median) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) และ ค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) จากนั้นจึงตัดสินใจคัดเลือกตัวแปรเข้าตัวแบบตามเกณฑ์ที่กำหนด (Seijo-Pardo et al., 2016, 2018; Bolón-Canedo & Alonso-Betanzos, 2019; Effrosynidis & Arampatzis, 2021; Yi Yang, 2022) แต่ยังไม่มีงานศึกษาเพื่อเปรียบเทียบประสิทธิภาพทั้ง 4 วิธี ค่าต่ำสุด ค่ากลาง ค่าเฉลี่ยเลขคณิต และ ค่าเฉลี่ยฮาร์มอนิกด้วยกัน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

รูปแบบที่ 2 รูปแบบการการเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุแล้วตามด้วยการรวมเซตของตัวแปรที่มีความสำคัญดังกล่าว (Thresholding followed by Combination; Design TC) ดังรูปที่ 2 เป็นการกำหนดเกณฑ์คัดเลือกตัวแปรก่อน เช่น ในแต่ละวิธีการคัดเลือกตัวแปรเข้าตัวแบบจะเลือกตัวแปรต้นที่มีความสำคัญสูงที่สุดเป็น 5 ลำดับแรก จากนั้นจึงร่วมทำการตัดสินใจด้วยวิธีรวมตัวแปรต้นในรูปแบบของเซต เช่น ยูเนียนโดยเลือกตัวแปรต้นที่ปรากฏอยู่ในเซตใด ๆ และ อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่ทุกเซต



รูปที่ 1 ตัวอย่างการทำ Design CT (Design combination followed by thresholding) สำหรับวิธีการเลือกตัวแปรเข้าตัวแบบ 2 วิธี



รูปที่ 2 ตัวอย่างการทำ Design TC (Design thresholding followed by combination) สำหรับวิธีการเลือกตัวแปรเข้าตัวแบบ 2 วิธี

Seijo-Pardo et al. (2019) ได้ทำการเปรียบเทียบ Design CT และ Design TC สำหรับข้อมูลมิติสูงในรูปแบบการจำแนกไบนารีพบว่า Design TC โดยใช้วิธีอินเตอร์เซกและมัลติอินเตอร์เซกในการเลือกตัวแปรต้นเข้าตัวแบบจะให้ค่าความผิดพลาดในการทดสอบ (Test error percentage) น้อยกว่าวิธีอื่นในหลายชุดข้อมูลที่ทำการทดสอบ สำหรับ Design CT งานศึกษาได้ใช้วิธีการรวมลำดับความสำคัญของตัวแปรต้นโดยพิจารณาเพียงแค่ค่าต่ำสุด (Min) อีกทั้งไม่มีงานศึกษาการคัดเลือกตัวแปรแบบรวมกลุ่มในส่วนของวิธีคัดเลือกตัวแปรด้วยประเภท Wrapper โดย SHAP ซึ่งเป็นหนึ่งในวิธีการคัดเลือกตัวแปรเข้าตัวแบบในประเภท Wrapper ได้ถูกพบว่าเป็นวิธีที่ทำให้ผลการทำนายมี

ความเสถียรที่สุด เมื่อเทียบกับการคัดเลือกตัวแปรด้วยวิธีอื่น (Effrosynidis & Arampatzis, 2021) นอกจากนี้วิธีการคัดเลือกตัวแปรเข้าตัวแบบอาจมีประสิทธิภาพและความเสถียรของผลการทำนายที่ต่างกันภายใต้สถานการณ์ที่ต่างกัน

ดังนั้นผู้ศึกษาจึงมีความสนใจที่จะเปรียบเทียบการคัดเลือกตัวแปรเข้าตัวแบบ XGBoost ด้วย 3 วิธีดังนี้ วิธีคัดเลือกตัวแปรแบบวิธีเดียว (Single feature selection) และวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม (Ensemble feature selection) ด้วยวิธี Design CT และ วิธี Design TC โดยในส่วนของวิธีการคัดเลือกตัวแปรเข้าตัวแบบจะใช้ทั้งประเภท Filter Wrapper และ Embedded ในขั้นตอนการรวมลำดับความสำคัญของตัวแปรต้นใน Design CT จะพิจารณาการรวมลำดับความสำคัญตัวแปรต้นของแต่ละวิธีการคัดเลือกตัวแปรโดยใช้ค่า 1) ค่าต่ำสุด 2) ค่ากลาง 3) ค่าเฉลี่ยเลขคณิต และ 4) ค่าเฉลี่ยฮาร์มอนิก ในการจัดลำดับความสำคัญของตัวแปรต้นใหม่ และใน Design TC จะใช้วิธี 1) ยูเนียนโดยเลือกตัวแปรต้นที่ปรากฏอยู่ในเซตใด ๆ 2) มัลติอินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่อย่างน้อย 2 เซตพร้อมกัน และ 3) อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่ทั้ง 3 เซตพร้อมกัน

1.2 วัตถุประสงค์

1 เพื่อเปรียบเทียบประสิทธิภาพ (Efficiency) ของตัวแบบทำนายโดยวัดจากค่าเฉลี่ยของ F1-scoreและความเสถียร (Stability) ของตัวแบบทำนายโดยวัดจากค่าเบี่ยงเบนของ F1-score จากการคัดเลือกตัวแปรเข้าตัวแบบ 3 วิธีได้แก่ คัดเลือกตัวแปรแบบวิธีเดียว (Single feature selection) และคัดเลือกตัวแปรแบบรวมกลุ่ม (Ensemble feature selection) ด้วยวิธี Design CT (Combination followed by thresholding) และ วิธี Design TC (Thresholding followed by combination) โดยเลือกใช้วิธีการคัดเลือกตัวแปรเข้าตัวแบบจากทั้งประเภท Filter Wrapper และ Embedded

2 เพื่อเปรียบเทียบประสิทธิภาพ (Efficiency) และความเสถียร (Stability) ของตัวแบบทำนาย จากวิธีการรวมลำดับความสำคัญของตัวแปรต้นจากแต่ละวิธีการคัดเลือกตัวแปรเข้าตัวแบบ โดยรูปแบบ Design CT ใช้ 1) ค่าต่ำสุด (Min) 2) ค่ากลาง (Median) 3) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) และ 4) ค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) และรูปแบบ Design TC ใช้ 1) ยูเนียนโดยเลือกตัวแปรต้นที่ปรากฏอยู่ในเซตใด ๆ 2) มัลติ

อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอย่างน้อย 2 เซตพร้อมกัน และ 3) อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่ทั้ง 3 เซตพร้อมกัน

1.3 ขอบเขตของการศึกษา

1 ชุดข้อมูลมีลักษณะเป็นการจำแนกแบบไบนารี (Binary classification) ให้ P แทนจำนวนตัวแปรต้น และ N แทนขนาดข้อมูล ลักษณะชุดข้อมูลที่จะนำมาศึกษามี 3 ชุดดังนี้ 1) ชุดข้อมูล Parkinson's Disease ที่มีจำนวนตัวแปรเท่ากับขนาดข้อมูล (P = N; P = 753, N = 756) 2) ชุดข้อมูล LSVT Voice Rehabilitation ที่มีจำนวนตัวแปรมากกว่าขนาดข้อมูล (P > N; P = 312, N = 126) และ 3) ชุดข้อมูล Colon Cancer ที่มีจำนวนตัวแปรมากกว่าขนาดข้อมูลอยู่มาก (P >> N; P = 2,000, N = 62)

2 แปลงข้อมูลตัวแปรเชิงปริมาณด้วยวิธี Min-Max Scaler ก่อนเพื่อลดผลของหน่วยของตัวแปรต้นที่ไม่เหมือนกัน และตัวแปรเชิงคุณภาพให้อยู่ในรูปไบนารี [0,1]

3 เกณฑ์คัดเลือกจำนวนตัวแปรต้นเข้าตัวแบบ มี 4 เกณฑ์ได้แก่ $\log_2(P)$ 2.5% 5% และ 10% ลำดับแรกของตัวแปรต้นที่มีลำดับความสำคัญสูงสุด

4 ค่าประเมินของผลการทำนายที่ใช้ได้แก่ ประสิทธิภาพของตัวแบบผ่านค่าเฉลี่ย F1-score และ ความเสถียรของการทำนายผ่านค่าเบี่ยงเบนของของ F1-score ด้วยการทำ K-Fold Cross Validation ให้ k = 10

5 วิธีการคัดเลือกตัวแปรเข้าตัวแบบ ประเภท Filter พิจารณา 3 วิธีในประเภท Filter ได้แก่ 1) Mutual Information 2) Variance Threshold และ 3) MultiSURF

6 วิธีการคัดเลือกตัวแปรเข้าตัวแบบประเภท Wrapper พิจารณา 3 วิธีที่อยู่ในประเภท Wrapper ได้แก่ 1) SHAP 2) Recursive Feature Elimination (RFE) และ 3) Boruta

7 ในวิธีการจัดลำดับความสำคัญของตัวแปรต้นแบบรวมกลุ่ม แล้วจากนั้นกำหนดเกณฑ์คัดเลือกตัวแปร (Design Combination followed by Thresholding; Design CT) ใช้วิธีการร่วมจัดลำดับความสำคัญของตัวแปรต้นดังนี้ ค่าต่ำสุด (Min) ค่ากลาง (Median) ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) และ ค่าเฉลี่ยฮาร์มอนิก (Harmonic Mean)

8 ในวิธีการเลือกตัวแปรต้นตามเกณฑ์ก่อนแล้วจึงนำเซตของตัวแปรต้นดังกล่าวมาคัดเลือกตัวแปรแบบรวมกลุ่ม (Design Thresholding followed by Combination; Design TC) ใช้วิธีการร่วมกันจัดเซตของตัวแปรต้นดังนี้ 1) ยูเนียนโดยเลือกตัวแปรต้นที่ปรากฏอยู่ในเซต

ใดๆ 2) มัลติอินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่อย่างน้อย 2 เซตพร้อมกัน และ 3) อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่ทั้ง 3 เซตพร้อมกัน

9 ใช้ตัวแบบ XGBoost ในการเปรียบเทียบ และมีการปรับค่าพารามิเตอร์สำหรับตัวแบบทำนาย

1.4 วิธีดำเนินการศึกษา

1 เลือกข้อมูลที่จะนำมาทดสอบ 3 รูปแบบได้แก่ 1) จำนวนตัวแปรต้นเท่ากับขนาดข้อมูล (P=N) 2) จำนวนตัวแปรต้นมากกว่าขนาดข้อมูล (P > N) และ 3) จำนวนตัวแปรต้นมากกว่าขนาดข้อมูลอยู่มาก (P >> N) นำข้อมูลดังกล่าวทำการคัดเลือกตัวแปรเข้าตัวแบบ XGBoost จากนั้นจึงนำผลมาเปรียบเทียบกัน สำหรับการเปรียบเทียบผลจากตัวแบบนั้น ใช้ค่าประเมินที่ได้จากการทำ K-Fold Cross Validation ให้ k = 10 ด้วยตัวแบบ ดังนี้

ค่าประเมินที่1 คือ ค่าเฉลี่ย F1 Score เป็นตัววัดประสิทธิภาพ (Efficiency) ของการทำนาย

ตารางที่ 1 Confusion Matrix เพื่อใช้ประเมินผลลัพธ์การทำนายตัวแบบทำนาย

		Actual class	
		Positive (1)	Negative (0)
Predicted class	Positive (1)	True positive (TP)	Fault Positive (FP)
	Negative (0)	Fault Negative (FN)	True Negative (TN)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ให้ True Positive (TP) แทน ผู้ป่วยตรวจพบผลบวกจริง
Fault Positive (FP) แทน ผู้ป่วยตรวจพบผลบวกปลอม
Fault Negative (FN) แทน ผู้ป่วยตรวจพบผลลบปลอม
True Negative (TN) แทน ผู้ป่วยตรวจพบผลลบจริง

ค่า Precision แสดงถึงความแม่นยำของตัวแบบทำนาย และค่า Recall แสดงถึงความถูกต้องของการทำนายผ่านตัวแบบ นำมาซึ่งการหาค่า F1 Score ที่ได้จากการหาค่าเฉลี่ยฮาร์มอนิกของค่า Precision และค่า Recall

$$Precision = \frac{TP}{TP + FP}$$

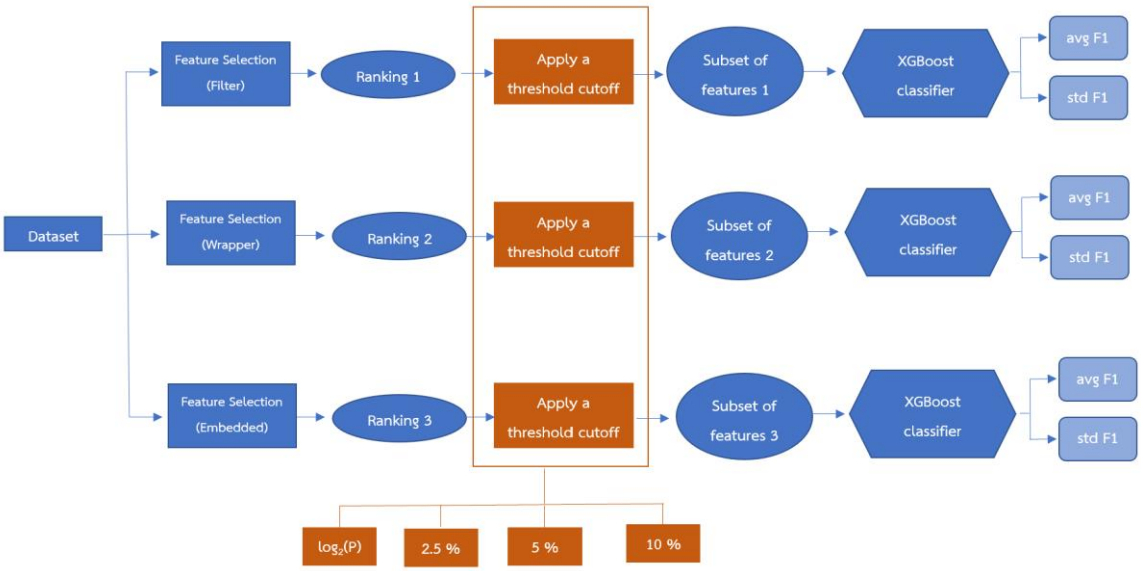
$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

<u>ค่าประเมินที่2</u> คือ ค่าเบี่ยงเบน (Standard Deviation) ของ F1-Score เป็นค่าวัดความเสถียร (Stability) ของการทำนาย

2 วิธีการทดสอบที่จะนำมาเปรียบเทียบผลลัพธ์มีทั้งหมด 3 วิธี ได้แก่

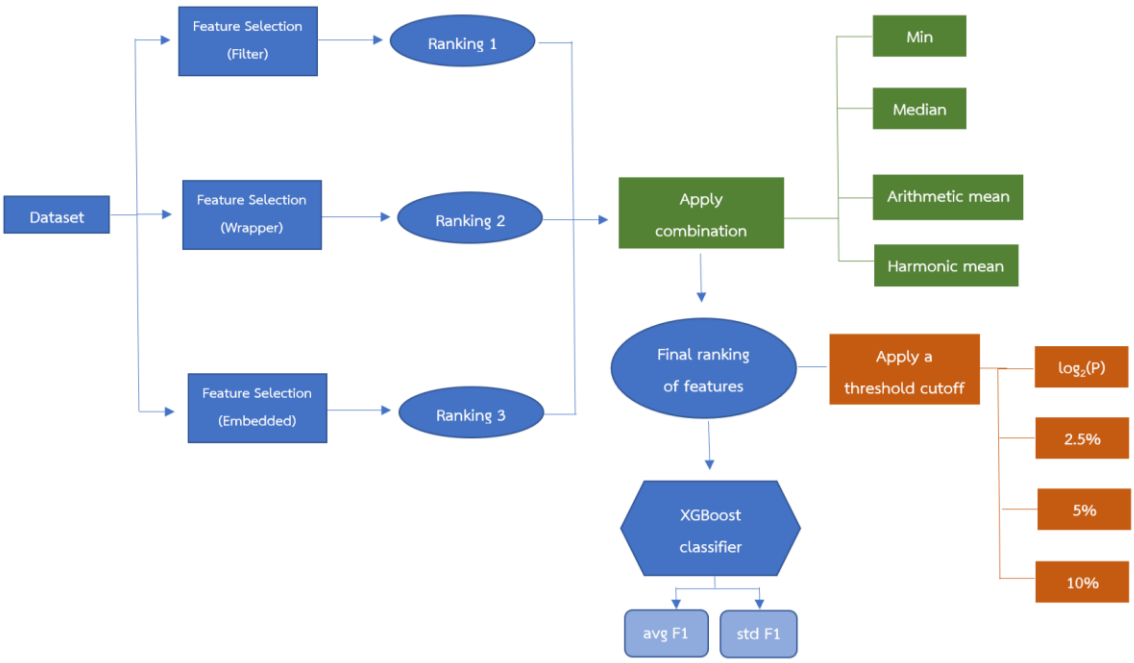
วิธีที่ 1 คัดเลือกตัวแปรแบบวิธีเดียว (Single feature selection) ทำการคัดเลือกตัวแปรเข้าตัวแบบด้วยวิธี 1) Mutual Information (MI) 2) SHAP และ 3) Embedded จากนั้นทำการเลือกจำนวนตัวแปรต้นที่ใช้ในตัวแบบตามเกณฑ์ $\log_2(P)$ 2.5% 5% และ 10% ลำดับแรกของตัวแปรต้นที่มีลำดับความสำคัญสูงสุด



CHULALONGKORN UNIVERSITY

รูปที่ 3 การคัดเลือกตัวแปรแบบวิธีเดียว (Single feature selection)

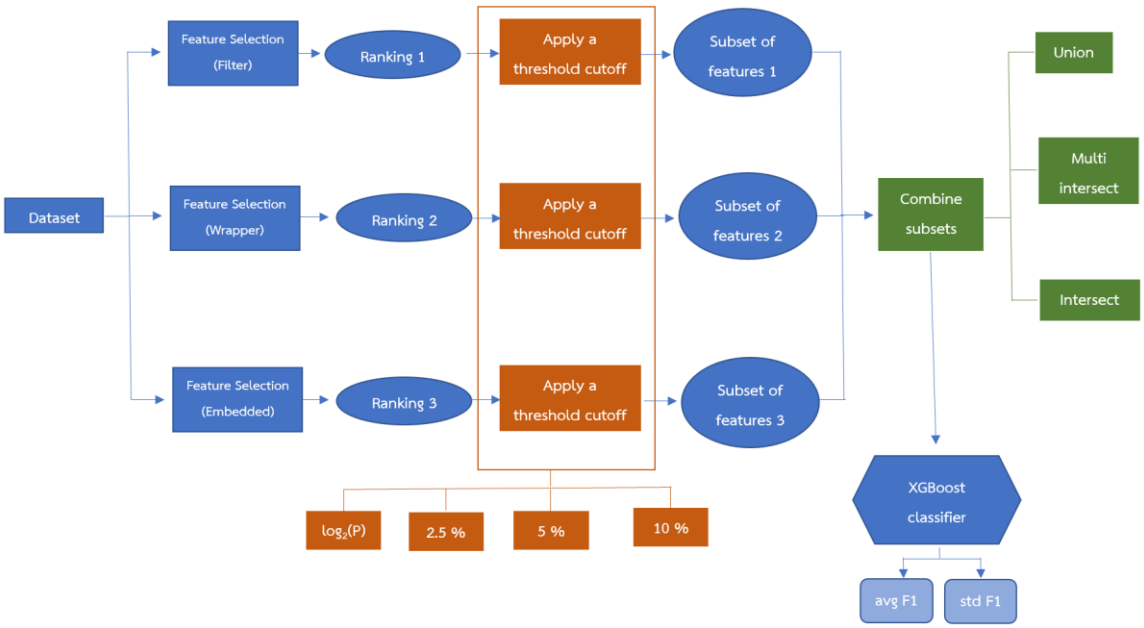
วิธีที่ 2 Design CT (Combination followed by thresholding) ทำการรวมลำดับความสำคัญของตัวแปรต้นจากวิธีที่ 1 ด้วยการคัดเลือกตัวแปรแบบรวมกลุ่ม โดยใช้ 1) ค่าต่ำสุด (Min) 2) ค่ากลาง (Med) 3) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) และ 4) ค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) ในการจัดลำดับความสำคัญของตัวแปรต้นใหม่ จากนั้นจึงเลือกจำนวนตัวแปรต้นที่ใช้ในตัวแบบตามเกณฑ์ $\log_2(P)$ 2.5% 5% และ 10% ลำดับแรกของตัวแปรต้นที่มีลำดับความสำคัญสูงสุด



รูปที่ 4 การคัดเลือกตัวแปรแบบรวมกลุ่ม รูปแบบ Design CT

วิธีที่ 3 Design TC (Thresholding followed by combination) โดยเลือกจำนวนตัวแปรต้นตามเกณฑ์ $\log_2(P)$ 2.5% 5% และ 10% ลำดับแรกของตัวแปรต้นที่มีลำดับความสำคัญสูงสุดจากวิธีที่ 1 จากนั้นจึงนำเซตของตัวแปรต้นที่เลือกดังกล่าวมาคัดเลือกตัวแปรแบบรวมกลุ่มด้วยวิธี 1) ยูเนียนโดยเลือกตัวแปรต้นที่ปรากฏอยู่ในเซตใด ๆ 2) มัลติอินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่อย่างน้อย 2 เซตพร้อมกัน และ 3) อินเตอร์เซกโดยเลือกตัวแปรต้นที่ปรากฏอยู่ทั้ง 3 เซตพร้อมกัน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



รูปที่ 5 การคัดเลือกตัวแปรแบบรวมกลุ่ม รูปแบบ Design TC

1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานศึกษา

สามารถเลือกวิธีการคัดเลือกตัวแปรเข้าตัวแบบที่เหมาะสมกับลักษณะของข้อมูลที่มีลักษณะหลายมิติและเป็นการจำแนกแบบไบนารี (Binary classification)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ข้อมูลหลายมิติพบมากในข้อมูลด้านดีเอ็นเอไมโครอะเรย์ (DNA microarray) มีงานศึกษาที่เกี่ยวกับการทำนายผู้ป่วยผ่านข้อมูลดีเอ็นเอไมโครอะเรย์ โดยมุ่งเน้นการหาวิธีที่ทำให้ได้ผลการทำนายที่ดีขึ้น เช่น การทำการคัดเลือกตัวแปร เนื่องจากยังคงสามารถอธิบายการทำนายผ่านตัวแปรต้นที่คงเหลือจากทำคัดเลือกตัวแปร ทั้งนี้งานศึกษาต่าง ๆ มีวิธีการคัดเลือกตัวแปรที่ต่างกัน นอกเหนือจากข้อมูลเอ็นเอไมโครอะเรย์แล้วยังมีข้อมูลอนุกรมเวลาที่แปลงให้อยู่ในรูปตาราง (Tabular) จะส่งผลให้จำนวนของตัวแปรต้นเพิ่มขึ้น เช่น เดิมข้อมูลมีตัวแปรต้น 5 ตัว ที่ถูกเก็บตามระยะเวลา 10 ช่วงระยะเวลา เมื่อแปลงเป็นรูปตารางจะทำให้จำนวนตัวแปรต้นเพิ่มขึ้นถึง 50 ตัว ตัวอย่างข้อมูลด้าน Environment ที่ใช้ในงานศึกษาของ Effrosynidis & Arampatzis (2021) งานศึกษานี้จึงสนใจการคัดเลือกตัวแปรแบบรวมกลุ่มทั้งประเภท Filter Wrapper และ Embedded ด้วยชุดข้อมูลดีเอ็นเอไมโครอะเรย์ โดยทฤษฎีที่เกี่ยวข้องได้แก่ ตัวแบบทำนายแบบจำแนกประเภท (Classifier model) การคัดเลือกตัวแปร (Feature selection) แบ่งเป็นประเภท Filter Wrapper และ Embedded

2.1 ทฤษฎีที่เกี่ยวข้อง

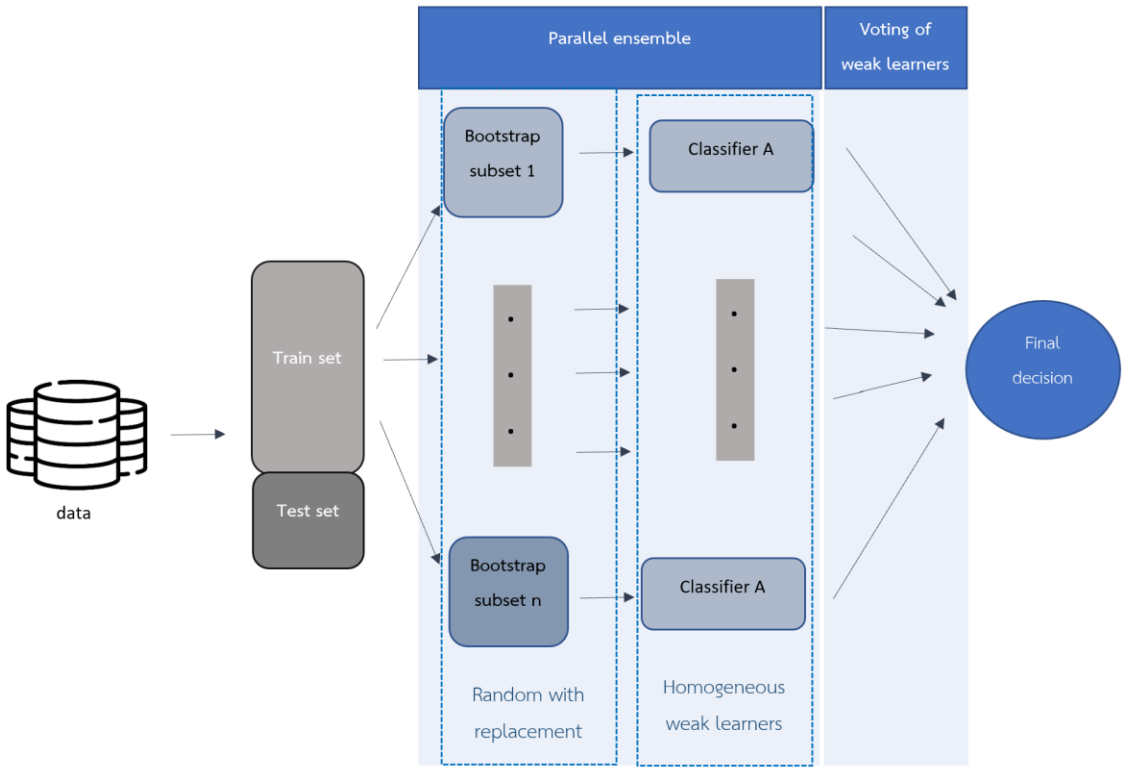
2.1.1 Ensemble Learning สำหรับตัวแบบ Tree-based

Ensemble คือการรวมการเรียนรู้ที่หลากหลายเข้าด้วยกัน เพื่อเพิ่มประสิทธิภาพของตัวแบบจากการลดขนาดของค่าเอนเอียง (Bias) และ ค่าความแปรปรวน (Variance) โดยการทำ Ensemble แบ่งได้เป็น 2 ประเภทได้แก่ 1) Homogeneous weak learners ใช้สำหรับ Bagging Ensemble และ Boosting Ensemble 2) Heterogeneous weak learners ใช้สำหรับ Stacking Ensemble

2.1.1.1 Bagging

Bootstrap Aggregating (Bagging) โดย Breiman เป็นการเรียนรู้แบบขนาน (Parallel ensemble) เริ่มจากการสุ่มชุดทดสอบเป็นหลายชุดด้วยวิธี Bootstrap แล้วจึงนำเข้าตัวแบบทำนายที่มีลักษณะ Homogeneous Weak Classifier ในขั้นตอนการเรียนรู้แบบขนาน (Parallel ensemble) จากนั้นนำผลของการทำนายที่ได้หลายชุดมารวมกัน (Aggregation) ด้วยวิธีหาค่าฐานนิยม (Mode) หรือ การโหวต (Voting) เพื่อนำมาสรุปผลเป็นผลการทำนายสุดท้าย สำหรับตัวแบบที่ใช้เทคนิค Bagging ensemble ได้แก่ Random Forest (RF) โดยมีตัวแบบ Decision Tree (DT) ใช้

สำหรับการทำการจำแนกประเภท (Classifier) ในขั้นตอนการเรียนรู้แบบขนาน (Parallel ensemble)

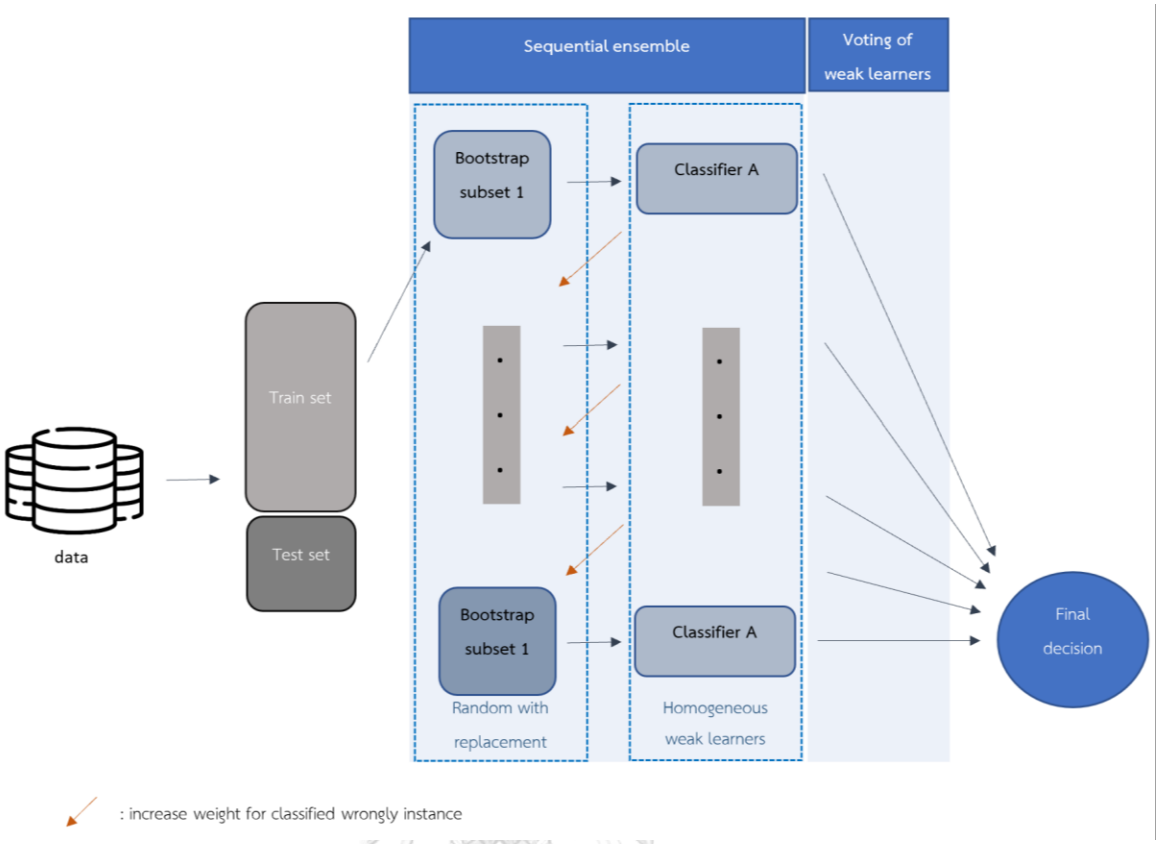


รูปที่ 6 แสดงขั้นตอนของการทำ Bagging Ensemble

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.1.1.2 Boosting

Schapire และ Freund ได้นำวิธีการทำ Boosting มาใช้กับ Adaptive Boosting (AdaBoost) ซึ่งเป็นตัวแบบแรกที่เริ่มใช้เทคนิค Boosting ในการทำนายข้อมูลที่มีลักษณะการจำแนกแบบไบนารี (Binary classification) เป็นการเรียนรู้ที่มีลักษณะการทำงานเป็นแบบลำดับขั้น (Sequential ensemble) โดยเรียนรู้จากข้อผิดพลาดการทำนายในตัวแบบก่อนหน้าเพื่อนำมาปรับปรุงในตัวแบบในขั้นถัดไป กล่าวคือเป็นการนำ Homogeneous Weak Classifiers หลายตัวมารวมกันเป็น Strong classifier การกระทำดังกล่าวจะทำให้ได้ผลการทำนายของตัวแบบที่มีประสิทธิภาพมากกว่าวิธี Bagging Ensemble และนอกเหนือจากตัวแบบ AdaBoost แล้วยังมีวิธีการทำ Boosting อีกชนิดคือ Gradient Boosting (GB) โดยมีตัวแบบที่พัฒนาต่อยอดมา ได้แก่ Extreme Gradient Boosting (XGBoost)

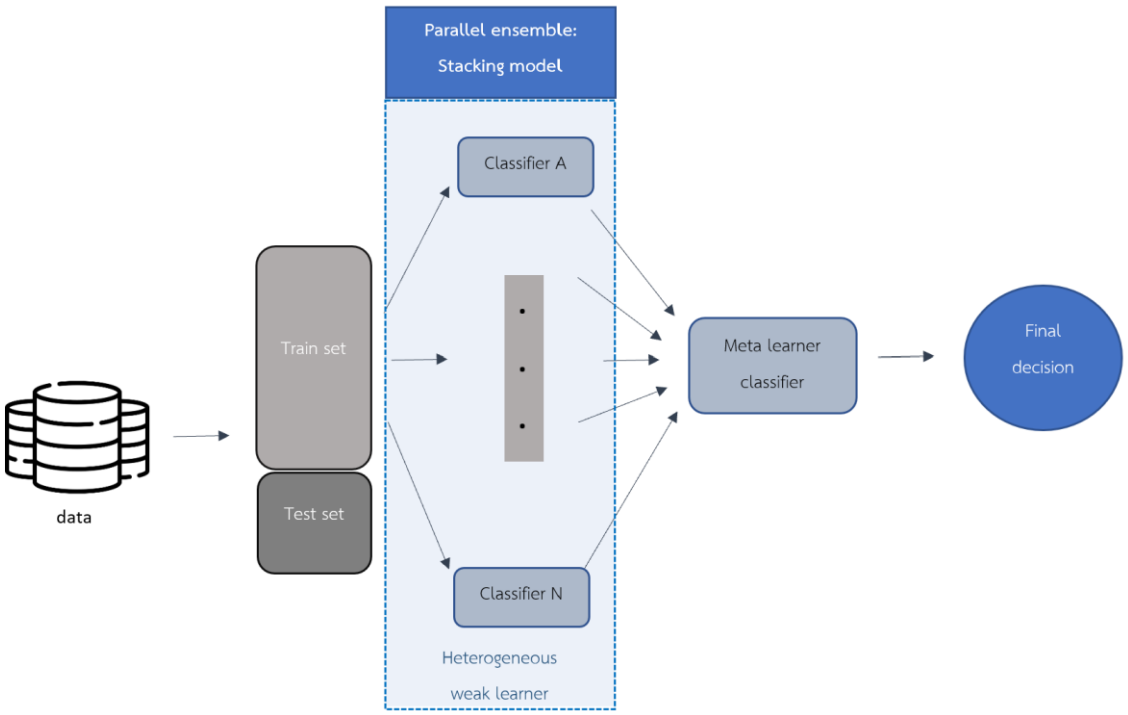


รูปที่ 7 แสดงขั้นตอนของการทำ Boosting Ensemble

2.1.1.3 Stacking

เป็นการเรียนรู้แบบขนาน (Parallel ensemble) โดยนำ Heterogeneous Weak Learner หลายตัวมาทำการเรียนรู้ร่วมกัน จากนั้นจึงส่งค่าทำนายไปยังตัวแบบที่อยู่ในขั้นตอนต่อไป หรือที่เรียกว่าขั้น Meta-Learner จากนั้นจึงนำผลจากตัวแบบในขั้น Meta-Learner ดังกล่าวเป็นผลของการทำนายขั้นสุดท้าย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

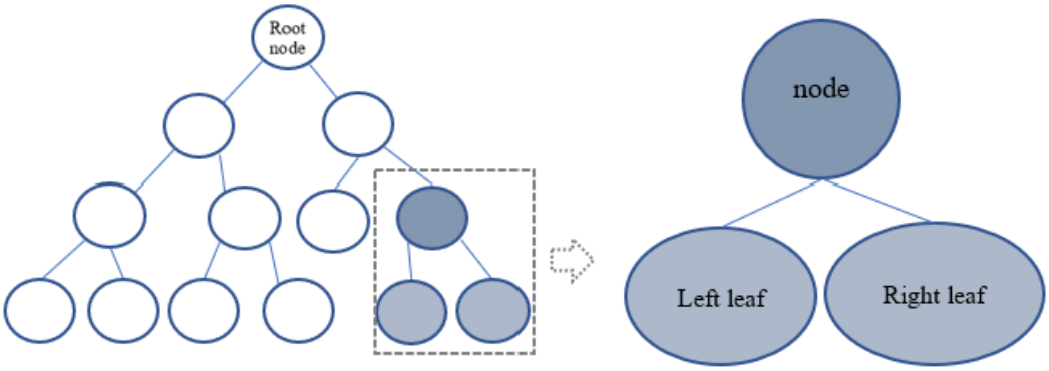


รูปที่ 8 แสดงขั้นตอนของการทำ Stacking Ensemble

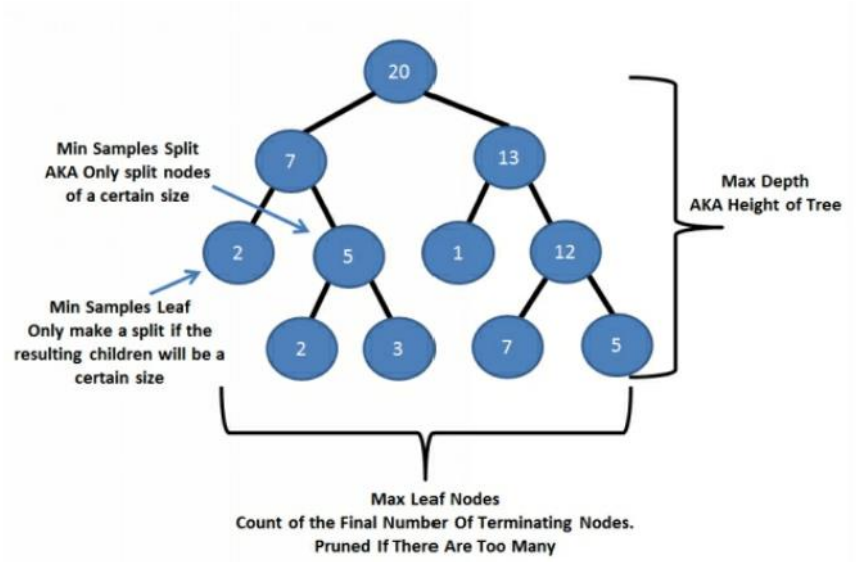
2.1.2 แบบจำลอง Tree-based

2.1.2.1 การเรียนรู้แบบต้นไม้ตัดสินใจ (Decision tree)

เป็นตัวแบบที่มีการตัดสินใจตามกฎ (Rule-based) ด้วยการสร้างกฎ if-else ของแต่ละค่าในฟีเจอร์ (Feature) หรือ ตัวแปรต้น ให้อยู่ในรูปของต้นไม้ตัดสินใจในการเรียนรู้ของเครื่อง โดยการตีความการตัดสินใจของต้นไม้นั้นสามารถทำได้ง่าย อีกทั้งยังใช้ได้ทั้งกับข้อมูลที่เป็นการจำแนกประเภท (Classification) และ การวิเคราะห์การถดถอย (Regression) ตัวแบบที่พัฒนามาจากต้นไม้การตัดสินใจ (Decision tree) ได้แก่ Random Forest (RF) XGBoost และอื่น ๆ



รูปที่ 9 แผนภาพแสดงต้นไม้การตัดสินใจ (Decision Tree) อย่างง่าย



CHULALONGKORN UNIVERSITY

รูปที่ 10 แผนภาพแสดงค่าพารามิเตอร์ในต้นไม้การตัดสินใจ (Decision tree)

ที่มา: (Hartshorn, 2017)

การแบ่งต้นไม้ตัดสินใจ (Decision tree) เริ่มจากแบ่งที่โหนดราก (Root node) การแบ่งโหนดในแต่ละครั้งนั้นจะเลือกวิธีแบ่งที่ทำให้ค่า Cost function ต่ำสุดเมื่อเปรียบเทียบกับการแบ่งโหนดในรูปแบบอื่น ๆ เพื่อให้ได้รูปแบบการแบ่งต้นไม้ที่เหมาะสม โดยค่า Cost Function สำหรับข้อมูลการจำแนกประเภท (Classification) จะแทนด้วยค่า Gini index / Gini impurity หรือ

Entropy สำหรับการวิเคราะห์การถดถอย (Regression) จะแทนด้วยค่า Residual Sum of Squares (RSS)

Gini index หรือ Gini impurity มีค่าอยู่ระหว่าง [0,1] ได้จากการในการแบ่งต้นไม้ตัดสินใจ หากจำนวนประเภทของตัวแปรตามที่อยู่ในโหนดใบ(Left leaf และ Right leaf ดังรูปที่ 9) มีจำนวนน้อย จะทำให้ค่า Gini index หรือ Gini impurity เข้าใกล้ 0 แสดงถึงการแบ่งโหนดดังกล่าวมีความสามารถที่ดีในการแบ่งประเภทของตัวแปรตามในต้นไม้การตัดสินใจ (Decision tree) ดังนั้นการเลือกรูปแบบการแบ่งโหนดในต้นไม้ตัดสินใจจะเลือกวิธีแบ่งที่ทำให้ค่า Gini Impurity มีค่าต่ำสุด โดยสามารถหาค่า Gini Impurity ของแต่ละโหนดใบได้ดังสมการ (2.1A) และค่า Gini Impurity ของแต่ละโหนดได้ดังสมการ (2.1B)

$$Gini\ Impurity\ for\ a\ Leaf = Gini(Z) = 1 - \sum_{i=1}^{g} {p_i}^2 \qquad (2.1A)$$

$$Total\ Impurity = weighted\ average\ of\ Gini\ Impurities\ for\ the\ Leaves \qquad (2.1B)$$

ให้ Z แทนชุดข้อมูลฝึกฝน (Training dataset)

g แทนจำนวนประเภทของตัวแปรตาม

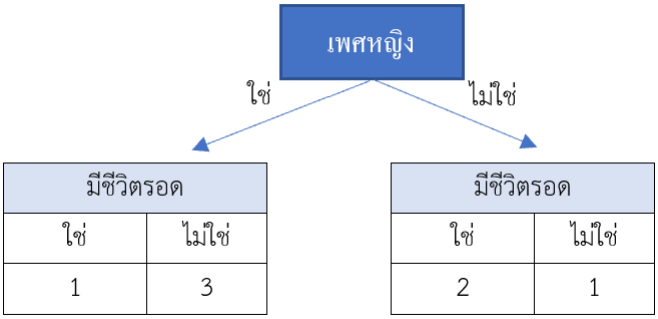
p_i แทนค่าความน่าจะเป็นที่ตัวแปรตามจัดอยู่ในประเภท (Class) ที่ i

ตัวอย่าง ให้ข้อมูลมี 2 ตัวแปร ได้แก่ ตัวแปรต้นคือเพศ และตัวแปรตามคือการมีชีวิตรอดแบ่งเป็น 2 ประเภทได้แก่ รอดชีวิตและไม่รอดชีวิต (g=2) นำมาสร้างต้นไม้ตัดสินใจได้ดังนี้

Chulalongkorn University

ตารางที่ 2 ตัวอย่างข้อมูลที่ใช้สำหรับสร้าง Decision tree

เพศ (Gender)	**การมีชีวิตรอด (Survived)**
หญิง	ไม่รอดชีวิต
หญิง	ไม่รอดชีวิต
ชาย	รอดชีวิต
ชาย	ไม่รอดชีวิต
หญิง	ไม่รอดชีวิต
หญิง	รอดชีวิต
ชาย	รอดชีวิต



มีชีวิตรอด	
ใช่	ไม่ใช่
1	3

มีชีวิตรอด	
ใช่	ไม่ใช่
2	1

คำนวณหาค่า Gini Impurity ของโหนดใบ (Leaf node) อ้างอิงสมการ (2.1A) ได้ดังนี้

ค่า Gini Impurity ของโหนดใบซ้ายมีค่าเท่ากับ

$$Gini\ Impurity_{left\ leaf} = 1 - \sum_{i=1}^{2} {p_i}^2$$

$$= 1 - ((\text{ความน่าจะเป็นของการรอดชีวิต})^2 + (\text{ความน่าจะเป็นของการไม่รอดชีวิต})^2)$$

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$$= 0.375$$

ค่า Gini Impurity ของโหนดใบขวามีค่าเท่ากับ

$$Gini\ Impurity_{right\ leaf} = 1 - \sum_{i=1}^{2} {p_i}^2$$

$$= 1 - ((\text{ความน่าจะเป็นของการรอดชีวิต})^2 + (\text{ความน่าจะเป็นของการไม่รอดชีวิต})^2)$$

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2$$

$$= 0.444$$

ผลรวมของค่า Gini Impurity จากการแบ่งโหนดตัวแปรเพศได้จากการคำนวณผ่านสมการ (2.1B) โดยค่า Gini Impurity ของโหนดใบได้จากการคำนวณผ่าน (2.1A) ได้ดังนี้

$Total\ Impurity =$ ค่าเฉลี่ยถ่วงน้ำหนักของค่า $Gini\ Impurities$ ของโหนดใบ

$$= \left(\frac{\text{เพศหญิง}}{\text{ทั้งหมด}}\right) Gini\ Impurity_{left\ leaf} + \left(\frac{\text{เพศชาย}}{\text{ทั้งหมด}}\right) Gini\ Impurity_{right\ leaf}$$

$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right) 0.444$$

$$= 0.405$$

ผลรวมของ Gini impurity ของการแบ่งโหนดด้วยตัวแปรเพศหญิงมีค่าเท่ากับ 0.405

2.1.2.2 Extreme Gradient Boosting (XGBoost)

พัฒนามาจาก Gradient Boosted Decision Tree (GBDT) มีลักษณะการเรียนรู้แบบลำดับขั้น (Sequential ensemble) เป็นการเรียนรู้จากความผิดพลาดของการทำนายในลำดับก่อนหน้า

ขั้นตอนการสร้างตัวแบบ XGBoost อย่างง่าย (มี 1 ต้นไม้ตัดสินใจ) สำหรับข้อมูลที่มีลักษณะการจำแนกแบบไบนารีมีดังนี้

ขั้นที่1 หาค่าความผิดพลาดจากการทำนาย (Residuals)

$$Residuals = Observed\ values - Pridicted\ Values \quad (2.1C)$$

ขั้นที่2 หาค่าความเหมือน (Similarity scores) หาค่า Similarity scores ของแต่ละโหนด(Node) และโหนดใบ (Leaf nodes) ดังนี้

$$\text{Similarity} = \frac{(\sum Residual_i)^2}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)] + \lambda} \quad (2.1D)$$

ให้ i = 1,..., N แทน จำนวนของค่าความแตกต่าง (Residual)

λ แทน ค่าพารามิเตอร์สำหรับ Regularization

ขั้นที่3 หาวิธีการแบ่งต้นไม้ตัดสินใจ (Prune the tree) จากการแบ่งโหนดเป็น 2 โหนดใบ ในที่นี้ให้ γ หรือแกมม่า แทนค่า Gain ขั้นต่ำที่ยอมรับได้ในการแบ่งโหนด หากค่า Gain - γ ที่ได้ใบมีค่าเป็นบวก แสดงถึงการแบ่งโหนดดังกล่าวสมควรทำ แต่ถ้าหากเป็นค่าลบหมายความว่าไม่ควรแบ่งโหนดเป็น 2 โหนดใบในรูปแบบดังกล่าว อีกทั้งการแบ่งโหนดสามารถทำได้หลายแบบ ดังนั้นจึงเลือกวิธีการแบ่งที่ทำให้เกิดค่า Gain - γ สูงสุด

$$Gain = LeftLeaf_{Similarity} + RightLeaf_{Similarity} - Node_{Similarity} \quad (2.1E)$$

ขั้นที่4 หาค่า Output value ของแต่ละโหนดใบ โดยแต่ละโหนดใบมีค่าความความผิดพลาด (Residual) เกิดขึ้นหลายค่า จึงต้องกำหนดค่า Output Value ให้เป็นค่าตัวแทนของแต่ละโหนดใบโดยมีเพียงค่าเดียว

$$\text{Output Value} = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)] + \lambda} \quad (2.1F)$$

ขั้นที่5 ทำนายผลโดยการหาค่าความน่าจะเป็น

$$log(odds)\ Prediction = \ Initial\ predicted + (\varepsilon \times \ Output\ Values) \quad (2.1G)$$

$$Probability = \frac{e^{Log(odds)}}{1+e^{Log(odds)}} = \frac{1}{1+e^{-(log(odds))}} \quad (2.1H)$$

ให้ ε แทน อัตราการเรียนรู้ (Learning rate)

ในกรณีตัวแบบ XGBoost ซับซ้อนมากขึ้น กล่าวคือจำนวนต้นไม้ตัดสินใจมีมากขึ้นในขั้นตอน Boosting ensemble จะได้ค่าของ log(odds) ดังนี้

$$log(odds)Prediction = Initialpredicted + (\varepsilon \times \ Output\ Value\ _1) + \cdots + (\varepsilon \times Output\ Value\ _T) \quad (2.1I)$$

ให้ T แทน จำนวนต้นไม้ตัดสินใจทั้งหมดในตัวแบบ XGBoost

ตัวอย่าง การสร้าง XGBoost โดยใช้ข้อมูลทำนายการมีชีวิตรอดของผู้โดยสารที่แสดงในตารางที่ 2

ขั้นที่1 หาค่าความผิดพลาดจากการทำนายครั้งที่ 1 (Residual_1)

ตารางที่ 3 ตัวอย่างข้อมูลการมีชีวิตรอดสำหรับแสดงขั้นตอนการสร้าง XGBoost

เพศ (Gender)	**อายุ (Age)**	**การมีชีวิตรอด (Survived)**	**ค่าความผิดพลาดครั้งที่ 1 (Residual_1)**
หญิง	52	ไม่รอดชีวิต	0 - 0. 875 = -0. 875
หญิง	60	ไม่รอดชีวิต	0 - 0. 875 = -0. 875
ชาย	40	รอดชีวิต	1 – 0. 875 = 0. 125
ชาย	70	ไม่รอดชีวิต	0 - 0. 875 = -0. 875
หญิง	45	ไม่รอดชีวิต	0 - 0. 875 = -0. 875
หญิง	25	รอดชีวิต	1 – 0. 875 = 0. 125
ชาย	30	รอดชีวิต	1 – 0. 875 = 0. 125
ความน่าจะเป็นในการมีชีวิตรอด =		$1 - \log\left(\frac{4}{3}\right)$ = 0.875	

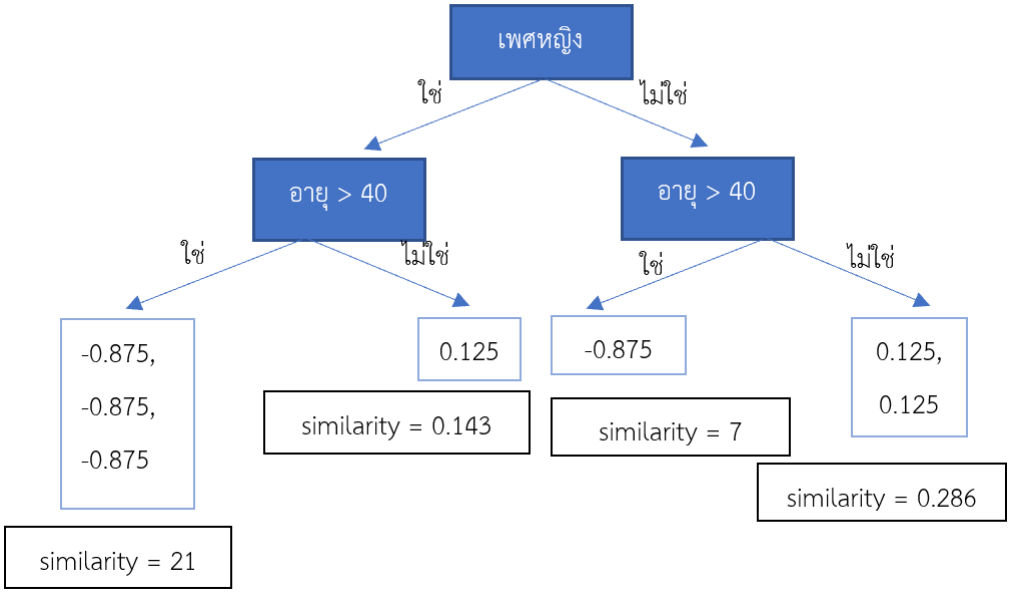
ขั้นที่2 หาค่าความเหมือน (Similarity scores) จากการแบ่งต้นไม้ได้ อ้างอิงจากสมการ (2.1D) เช่น หาค่า Similarity ของผู้หญิงที่มีอายุมากกว่า 40 ปี โดยกำหนดให้ $\lambda = 0$ ได้ดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$$\text{Similarity} = \frac{(\sum Residual_i)^2}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)] + \lambda}$$

$$= \frac{((-0.875)+(-0.875)+(-0.875))^2}{3(0.875*(1-0.875))}$$

$$= 21$$



<u>ขั้นที่3</u> หาวิธีการแบ่งต้นไม้ตัดสินใจ (Prune the tree) จากคำนวณค่า Gain - γ จากสมการ (2.1E)

$$Gain_{\text{ผู้}>40} = LeftLeaf_{Similarity} + RightLeaf_{Similarity} - Node_{Similarity}$$

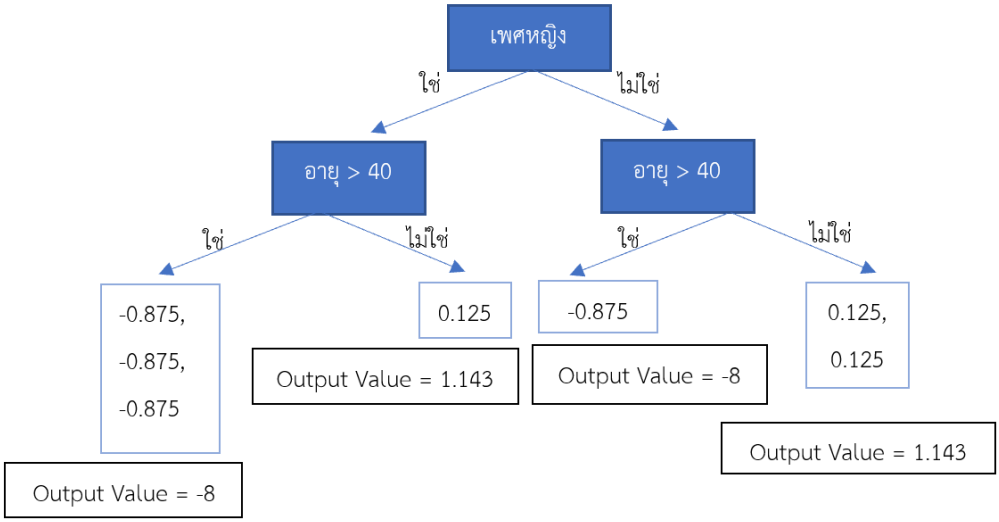
$$= 21 + 0.143 + 0$$

$$= 21.143$$

ค่า Gain ที่ได้จากการแบ่งโหนดผู้หญิงอายุมากกว่า 40 เท่ากับ 21.143 ในขั้นตอนนี้อาจมีการแบ่งด้วยเกณฑ์อื่น ๆ เช่น ผู้หญิงอายุมากกว่า 50 ปี ผู้หญิงอายุมากกว่า 60 ปีแล้วนำมาเปรียบเทียบด้วยค่า Gain ที่ได้จากการคำนวณดังตัวอย่างข้างต้น หากกำหนดให้ γ เท่ากับ 22 จะได้ว่า Gain - γ เป็นลบทำให้การแบ่งโหนดดังกล่าวจะถูกยกเลิก แต่หาก γ เท่ากับ 20 จะได้ว่า Gain - γ เป็นบวกหมายความว่าการแบ่งโหนดเช่นนั้นสามารถทำได้ โดยการแบ่งต้นไม้ตัดสินใจจะเลือกการแบ่งโหนดในต้นไม้ตัดสินใจที่ให้ค่า Gain - γ สูงสุด หากให้ค่าแกมม่าเท่ากับ 0 หมายความว่าการแบ่งโหนดจะพิจารณาจากเพียงค่า Gain เท่านั้น โดยจะเลือกการแบ่งโหนดที่ทำให้ค่า Gain สูงสุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

<u>ขั้นที่4</u> นำค่าความผิดพลาดจากการทำนายครั้งที่ 1 มาสร้างต้นไม้ตัดสินใจ และหาค่า Output ของแต่ละโหนดใบได้ดังรูป



หาค่า Output ครั้งที่ 1 อ้างอิงจากสมการ (2.1F) จะได้

$$\text{Output Value}_{\text{ญ}>40} = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)]}$$

$$= \frac{(-0.875)+(-0.875)+(-0.875)}{3(0.875(1-0.875))}$$

$$= -8$$

$$\text{Output Value}_{\text{ญ}\leq40} = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)]}$$

$$= \frac{0.125}{(0.875(1-0.875))}$$

$$= 1.143$$

$$\text{Output Value}_{\text{ช}>40} = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)]}$$

$$= \frac{(-0.875)}{(0.875(1-0.875))}$$

$$= -8$$

$$\text{Output Value}_{\text{ช}\leq40} = \frac{\sum Residual_i}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)]}$$

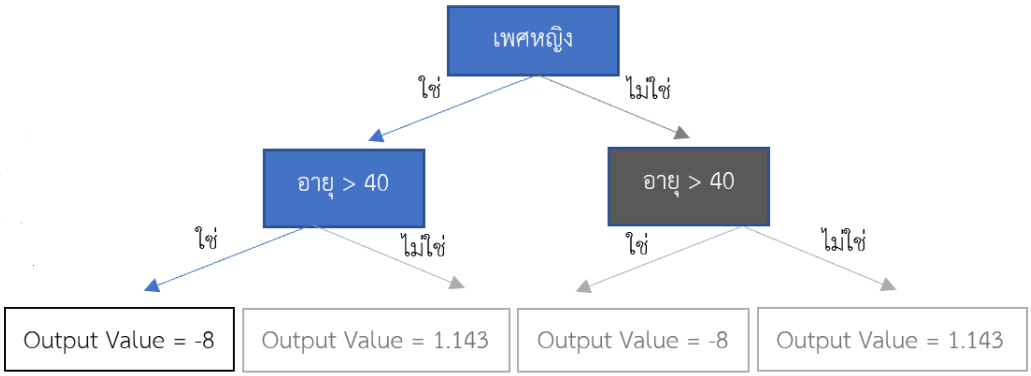
$$= \frac{0.125+0.125}{2(0.875(1-0.875))}$$

$$= 1.143$$

Chulalongkorn University

ขั้นที่5 หาค่าความน่าจะเป็นในการมีชีวิตรอดโดยใช้ค่า Output ที่ได้ในขั้นที่ 4 เพื่อใช้ในการหาค่าความผิดพลาดครั้งใหม่ (Residual_2)

ค่าความน่าจะเป็นในการมีชีวิตรอด เพื่อใช้ในการหาค่าความผิดพลาดใหม่ (Residual_2) ได้ดังตารางที่ 4 อ้างอิงค่า Output ที่ได้จากการคำนวณในตัวอย่างขั้นที่ 4 เช่น คนที่ 1 เพศหญิงอายุ 52 มีค่า Output เท่ากับ -8 นำมาหาค่าความน่าจะเป็นในการมีชีวิตรอดโดยอ้างอิงจากสมการที่ (2.1G) และ (2.1H) ได้ดังนี้



จากสมการที่ (2.1G) กำหนดอัตราการเรียนรู้เท่ากับ 0.8 ค่า log(odds) เท่ากับ

$$\begin{aligned} log(odds) &= Initial\ predicted + (\varepsilon \times Output\ Values) \\ &= \left(1 - \log\left(\tfrac{4}{3}\right)\right) + \left(0.8 \times (-8)\right) \\ &= -5.525 \end{aligned}$$

ความน่าจะเป็นของการมีชีวิตรอดของผู้โดยสารเพศหญิงอายุ 52 ปี อ้างอิงจากสมการที่ (2.1H) จะได้

$$\begin{aligned} Probability &= \frac{1}{1+e^{-(log(odds))}} \\ &= \frac{1}{1+e^{-(-5.525)}} \\ &= 0.004 \end{aligned}$$

คนที่ 6 คือเพศหญิงอายุ 25 ปี มีค่า Output เท่ากับ 1.143 นำมาหาค่าความน่าจะเป็นในการมีชีวิตรอดด้วยวิธีเช่นเดียวกับคนที่ 1 ได้ดังนี้

$$log(odds) = Initial\ predicted + (\varepsilon \times Output\ Values)$$

$$= \left(1 - \log\left(\frac{4}{3}\right)\right) + \left(0.8 \times (1.143)\right)$$

$$= 1.789$$

ความน่าจะเป็นของการมีชีวิตรอดของผู้โดยสารเพศหญิงอายุ 25 ปี อ้างอิงจากสมการที่ (2.1H) จะได้

$$Probability = \frac{1}{1+e^{-(log(odds))}}$$

$$= \frac{1}{1+e^{-(1.789)}}$$

$$= 0.857$$

ตารางที่ 4 หาค่าความผิดพลาด (Residual) ครั้งใหม่เพื่อนำไปสร้างต้นไม้รอบถัดไป

เพศ (Gender)	**อายุ (Age)**	**การมีชีวิตรอด (Survived)**	**ความน่าจะเป็นของการมีชีวิตรอดที่ได้จากค่า Output ครั้งที่ 1**	**ค่าความผิดพลาดครั้งที่ 2 (Residual_2)**
หญิง	52	ไม่รอดชีวิต	0.004	0 – 0.004 = - 0. 004
หญิง	60	ไม่รอดชีวิต	0. 004	0 – 0.004 = - 0. 004
ชาย	40	รอดชีวิต	0.857	1-0.857 = 0.143
ชาย	70	ไม่รอดชีวิต	0. 004	0 – 0.004 = - 0. 004
หญิง	45	ไม่รอดชีวิต	0. 004	0 – 0.004 = - 0. 004
หญิง	25	รอดชีวิต	0. 857	1-0.857 = 0. 143
ชาย	30	รอดชีวิต	0. 857	1– 0.857 = - 0. 143

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ค่าความผิดพลาดครั้งที่ 2 ในตารางที่ 4 นำไปสร้างต้นไม้ตัดสินใจแล้วทำซ้ำเหมือนในตัวอย่างขั้นที่ 1 เมื่อต้นไม้ถูกสร้างขึ้นมากกว่า 1 ต้น จะทำให้การหาค่าความน่าจะเป็นของการมีชีวิตรอดจากต้นไม้ตัดสินใจทั้งหมดจำนวน T ต้นในตัวแบบ XGBoost สามารถหาได้โดยอ้างอิงสมการที่ (2.1I)

2.1.3 การคัดเลือกตัวแปรเข้าตัวแบบ

2.1.3.1 Filter

การจัดลำดับความสำคัญของตัวแปร (Feature ranking) เพื่อใช้ในการคัดเลือกตัวแปรเข้าตัวแบบ ในประเภท Filter อาจทำโดยกาหาค่าความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม เช่น กำหนดว่าตัวแปรต้นที่มีค่าความพันธ์กับตัวแปรตามสูงสุด k ลำดับแรกจะใช้ในตัวแบบทำนาย ค่าทดสอบทางสถิติ (Statistical test) ในการจัดลำดับความสัมพันธ์ของตัวแปรต้นและตัวแปรตามจะต้องพิจารณาถึงประเภทของตัวแปรตามในชุดข้อมูลที่ใช้เป็นแบบกลุ่ม (Categorical) หรือแบบค่า

ต่อเนื่อง (Continuous) หากเลือกใช้การทดสอบทางสถิติ (Statistical test) ไม่เหมาะสมกับลักษณะข้อมูลอาจส่งผลต่อความถูกต้องของการจัดลำดับความสำคัญหรืออิทธิพลของตัวแปรต้นที่มีผลต่อตัวแปรตาม ตัวอย่างของการจัดลำดับความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามด้วยวิธี Filter เช่น Chi-square, ANOVA, Mutual Information (MI), MultiSURF และ Variance Threshold

Mutual Information (MI)

เป็นค่าวัดความสัมพันธ์ระหว่างตัวแปรต้นที่สนใจและตัวแปรตาม โดย MI สามารถใช้วัดค่าความสัมพันธ์ที่อยู่ในรูปแบบความสัมพันธ์เชิงเส้นตรง (Linear) และความสัมพันธ์แบบไม่เป็นเชิงเส้นตรง (Non-linear) มีค่าอยู่ระหว่าง [0, ∞) ค่า MI ที่มากขึ้นแสดงถึงตัวแปรต้นดังกล่าวมีความสัมพันธ์กับค่าตัวแปรตามยิ่งมากเช่นกัน การคัดเลือกตัวแปรต้นจะเลือกตัวแปรต้นที่มีค่า MI สูงสุด k ลำดับแรก ซึ่งค่า MI มีความต่างจากค่าสหสัมพันธ์ (Correlation) ในประเด็นที่ค่าสหสัมพันธ์เป็นการบ่งบอกความสัมพันธ์ที่อยู่ในรูปแบบความสัมพันธ์เชิงเส้นตรง (Linear) เท่านั้น มีค่าอยู่ระหว่าง [-1,1]

MI สามารถใช้ได้กับทั้งตัวแปรเชิงปริมาณ (Quantitative) และ เชิงคุณภาพ (Qualitative) ในการหาค่า MI ระหว่างตัวแปรที่มีลักษณะเชิงคุณภาพนั้น ให้ใช้การแทนค่าข้อมูลนามบัญญัติ (Nominal) ด้วยตัวแปรแบบค่าไม่ต่อเนื่อง (Discrete) ดังตัวอย่างในรูปที่ 11 และรูปที่ 12 การหาค่าของ MI มีแนวคิดเกี่ยวข้องกับค่าเอนโทรปี (Entropy) ค่า MI จะบ่งบอกว่าการที่รู้ค่าของตัวแปรหนึ่งจะช่วยลดค่าความไม่แน่นอนของอีกตัวแปรหนึ่งได้มากน้อยเพียงใด สำหรับค่าเอนโทรปีจะใช้วัดความไม่แน่นอนของการเกิดขึ้นของตัวแปรหนึ่ง โดยเอนโทรปีมีสมการดังนี้

$$H(X) = -\sum p(x) log_u\big(p(x)\big)$$

ให้ $H(X)$ แทน ค่าเอนโทรปี (Entropy)

$p(x)$ แทน ความน่าจะเป็นในประเภทตัวแปรตามที่สนใจ

u แทน หน่วยของ Entropy กรณีที่ u = 2 จะมีหน่วยเป็น bit ถ้าหาก u = e จะมีหน่วยเป็น nat และ u = 10 จะมีหน่วยเป็น Hartley

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

<u>ตัวอย่าง</u> การหาค่า MI จากการโยนเหรียญจำนวน 2 เหรียญ โดยเหรียญที่ 1 มีโอกาสออกด้านหัวเท่ากับ 0.5 และโอกาสออกด้านก้อยเท่ากับ 0.5 เปรียบเทียบกับเหรียญที่ 2 โอกาสออกด้านหัวเท่ากับ 0.3 และโอกาสออกด้านก้อยเท่ากับ 0.7 ให้ x = 0, 1 แทน เหรียญออกหัว และ ก้อย โดยให้ u=2 เนื่องจาก x มีความเป็นไปได้ 2 ค่าได้แก่ 0 และ 1

เหรียญที่ 1 นำมาหาค่า MI ได้ดังนี้

$H(X) = -(p(x=0)\ \log_2(p(x=0)) + p(x=1)\ \log_2(p(x=1)))$

$H(X) = -(0.5*\log_2(0.5) + 0.5*\log_2(0.5))$

$H(X) = -(-0.5 - 0.5) = 1$

เหรียญที่ 2 นำมาหาค่า MI ได้ดังนี้

$H(X) = -(p(x=0)\ \log_2(p(x=0)) + p(x=1)\ \log_2(p(x=1)))$

$H(X) = -(0.3*\log_2(0.3) + 0.7*\log_2(0.7))$

$H(X) = -(-0.52\ -0.36) = 0.88$

ค่าเอนโทรปีใช้วัดความไม่แน่นอน จากตัวอย่างข้างต้นแสดงให้เห็นว่าเหรียญที่ 1 มีโอกาสออกหัวก้อย (50:50) จะได้ค่าเอนโทรปีเท่ากับ 1 ซึ่งมีค่ามากกว่าเหรียญที่ 2 ที่โอกาสออกหัวก้อย (30:70) ที่มีค่าเอนโทรปีเท่ากับ 0.88 พบว่าค่าเอนโทรปีของเหรียญที่ 1 มากกว่าเนื่องจากมีความไม่แน่นอนมากกว่าเหรียญที่ 2

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

<u>การหาค่า MI ระหว่าง 2 ตัวแปรที่มีลักษณะเป็นตัวแปรไม่ต่อเนื่อง (Discrete variables)</u>

$$I(X,Y) = \sum_X \sum_Y \ p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (2.1J)$$

ให้ $I(X,Y)$ แทน ค่าของ MI ระหว่าง X และ Y ซึ่งจะมีค่ามากกว่า 0 และจะมีค่าเท่ากับ 0 เมื่อ

$$p(x,y) = p(x)p(y)$$

$p(x,y)$ แทน ความน่าจะเป็นร่วม (Joint probability) ระหว่างตัวแปร x และ y

$p(x)$, $p(y)$ แทน ความน่าจะเป็นตามขอบ (Marginal probabilities) ของตัวแปร x และ y ตามลำดับ

ตัวอย่าง การหาค่า MI จากตัวแปร 2 ตัวที่เป็นตัวแปรไม่ต่อเนื่อง (Discrete variables) แสดงผ่านชุดข้อมูลการมีชีวิตรอดของผู้โดยสารในเรือไททานิค แสดงความสัมพันธ์ของตัวแปรการมีชีวิตรอดและเพศในรูปแบบตารางการณ์จรดังรูปที่ 11 และแสดงในรูปแบบความน่าจะเป็นในรูปที่ 12 ในการหาค่า MI ระหว่าง 2 ตัวแปรแทนด้วย I(X,Y) จากสมการ (2.1J)

	Female	**Male**	**Total**
Not survived	89	483	571
Survived	230	112	342
Total	319	595	914

รูปที่ 11 ตารางการณ์จร (Contingency table) แสดงจำนวนการมีชีวิตรอดและเพศของผู้โดยสาร

	Female	**Male**	**Total**
Not survived	0.0974	0.5285	0.6258
Survived	0.2516	0.1225	0.3742
Total	0.3490	0.6510	1

รูปที่ 12 ตารางการณ์จร (Contingency table) แสดงความน่าจะเป็นของการมีชีวิตรอดและเพศของผู้โดยสาร

$$
\begin{aligned}
I(X,Y) = &\left(0.0974 \times log_2 \frac{0.0974}{0.6258 \times 0.3490}\right) \\
&+ \left(0.5285 \times log_2 \frac{0.5285}{0.6258 \times 0.6510}\right) \\
&+ \left(0.2516 \times log_2 \frac{0.2516}{0.3742 \times 0.3490}\right) \\
&+ \left(0.1225 \times log_2 \frac{0.1225}{0.3742 \times 0.6510}\right) \\
= &\ 0.2015
\end{aligned}
$$

ค่า MI เท่ากับ 0.2015 หมายถึงตัวแปรเพศของผู้โดยสารและการมีชีวิตรอดเกี่ยวข้องกัน เนื่องจาก MI มีค่ามากกว่า 0

<u>การหาค่า MI ระหว่าง 2 ตัวแปรที่มีลักษณะเป็นตัวแปรไม่ต่อเนื่อง (Discrete variables) และตัวแปรต่อเนื่อง (Continuous variables)</u>

$$I(X,Y) = \int_X \int_Y p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2.1K)$$

ให้ $I(X,Y)$ แทน ค่าของ MI ระหว่าง X และ Y ซึ่งจะมีค่ามากกว่า 0 และจะมีค่าเท่ากับ 0 เมื่อ

$$p(x,y) = p(x)p(y)$$

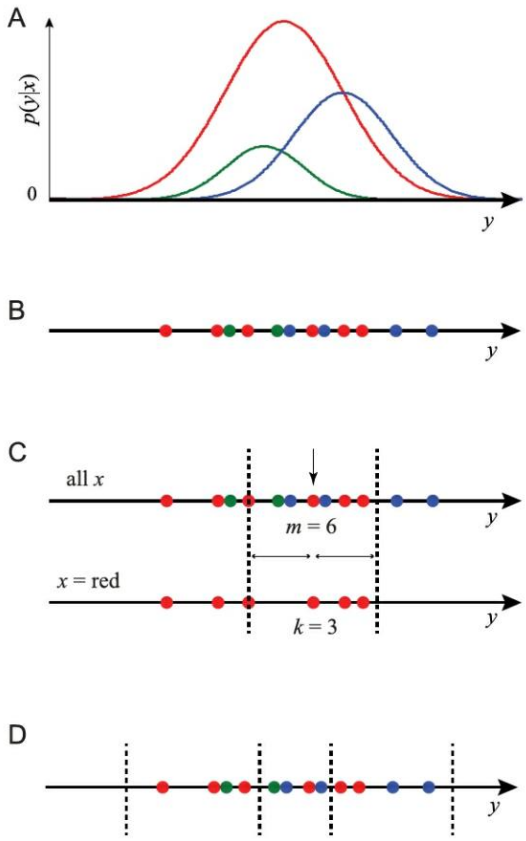
$p(x,y)$ แทน ความน่าจะเป็นร่วม (Joint probability) ระหว่างตัวแปร x และ y

$p(x)$, $p(y)$ แทน ความน่าจะเป็นตามขอบ (Marginal probabilities) ของตัวแปร x และ y ตามลำดับ

จากสมการ (2.1K) การหาค่า I(X,Y) จะทำได้ยากเมื่อตัวแปร X และ Y มีลักษณะความต่อเนื่อง (Continuous variables) อีกทางเลือกหนึ่งคือ การใช้วิธี Nearest-neighbors ในการประมาณค่า MI ซึ่งสามารถใช้ได้ในกรณีที่ตัวแปร X และ ตัวแปร Y ที่มีลักษณะต่อเนื่อง (Continuous variables) และไม่ต่อเนื่อง (Discrete variables) ด้วยการหาค่าความน่าจะเป็นร่วม (Joint probability) ของตัวแปร X และ Y

จากรูปที่ 13A แทนความน่าจะเป็นร่วม (Joint distribution) ของตัวแปร X ที่มีลักษณะไม่ต่อเนื่อง (Discrete variables) และตัวแปร Y ที่มีลักษณะต่อเนื่อง(Continuous variables)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



รูปที่ 13 Nearest-neighbor approach to estimate the MI

ที่มา: (Ross, 2014)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตัวอย่าง การคำนวณหาค่า MI ระหว่าง X เป็นตัวแปรไม่ต่อเนื่อง (Discrete variables) ที่มี 3 ค่า แทนด้วยสีแดง เขียว น้ำเงิน และ Y เป็นตัวแปรต่อเนื่อง (Continuous variables)

ขั้นที่1 สุ่มเลือกค่าที่สนใจมา 1 จุด จากรูปที่ 13C จุดที่เลือกคือจุดที่มีลูกศรบ่งชี้ จากนั้นทำการหาจำนวนข้อมูลที่อยู่ใกล้และมีค่า x เดียวกัน (N_{xi}) จากรูปที่ 13C พบว่ามี 3 จุด (k = 3)

ขั้นที่2 คำนวณหาระยะห่าง (d) ระหว่างค่าสังเกตสีแดงที่เลือกมา และค่าสังเกตที่มีค่า x เดียวกันที่อยู่ตำแหน่งที่ k

ขั้นที่3 นับจำนวนค่าสังเกต(m_i) ในระยะ d จากค่าที่สนใจ จากรูป $m_i = 6$ จะได้ค่า MI ของค่าสังเกตแต่ละตัว i = 1,...,N มีค่าเท่ากับ

$$I_i = \psi(N) - \psi(N_{xi}) + \psi(k) - \psi(m_i)$$

หาค่า MI ของชุดข้อมูล โดยการหาค่าเฉลี่ย I_i ในกรณีที่ X เป็นตัวแปรไม่ต่อเนื่อง (Discrete variables) และ Y เป็นตัวแปรต่อเนื่อง (Continuous variables) จะได้ค่า MI เท่ากับ

$$I(X,Y) =< I_i >= \psi(N) - \psi(N_x) + \psi(k) - \psi(m)$$

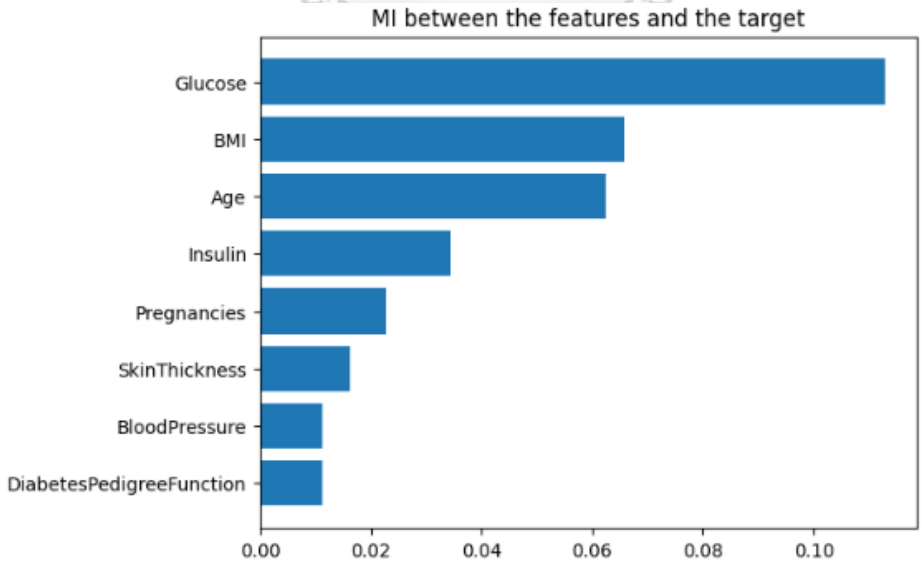
ให้ ψ แทน ฟังก์ชันไดแกมมา (Digamma function) เป็นอนุพันธ์ลอการิทึม (Logarithmic derivative) ของฟังก์ชันแกมมา $\psi(q) = \frac{d}{dq} ln\Gamma(q) = \frac{\Gamma'(q)}{\Gamma(q)}$, ในกรณีที่ n เป็นจำนวนเต็มบวกจะได้

$$\Gamma(q) = (q-1)!$$

k แทน จำนวนค่าสังเกตข้างเคียง (Number of Neighbors)

N แทน จำนวนค่าสังเกตทั้งหมด (Total number of observations)

<u>ตัวอย่าง</u> การหาค่า MI ระหว่างตัวแปรต้นที่มีค่าต่อเนื่อง (Continuous features) และตัวแปรตามมีลักษณะตัวแปรกลุ่ม (Categorical target) โดยใช้ชุดข้อมูล Diabetes ขนาด (N) 768 ตัวอย่าง จำนวนตัวแปรต้น (P) 8 ตัว และตัวแปรตามมีลักษณะการจำแนกแบบไบนารีผู้ป่วยที่ไม่เป็นเบาหวาน (y = 0) และ ผู้ป่วยที่เป็นเบาหวาน (y = 1) ได้ดังรูปที่ 14



รูปที่ 14 ค่า MI ระหว่างตัวแปรต้นและตัวแปรตามในชุดข้อมูล Diabetes

ใช้ค่า MI ที่ได้โดยเรียงลำดับจากมากไปน้อย ในการเลือกคัดเลือกตัวแปรต้นที่มีค่า MI สูงสุด k ลำดับแรก เช่น หากต้องการใช้ตัวแปรต้น 3 ลำดับแรกจะได้ว่าตัวแปร Glucose BMI และ Age จะถูกนำเข้าตัวแบบตามลำดับ

Variance Threshold

การคัดเลือกตัวแปรด้วยวิธี Variance Threshold จะพิจารณาเพียงค่าความแปรปรวนของตัวแปรต้น ในการจัดลำดับความสำคัญของตัวแปรต้นผ่านค่าความแปรปรวนดังกล่าว กล่าวคือตัวแปรที่มีแปรปรวนต่ำจะถูกลดความสำคัญลง เนื่องจากอาจไม่มีความสำคัญต่อการทำนายค่าตัวแปรตาม

วิธี Variance threshold สามารถใช้ได้กับตัวแปรเชิงปริมาณ (Quantitative) และ เชิงคุณภาพ (Qualitative) สำหรับตัวแปรเชิงปริมาณในชุดข้อมูลที่ใช้สามารถหาค่าความแปรปรวน (s_X^2) มีค่าเท่ากับ $\frac{\sum(X-\bar{x})^2}{N-1}$ และสำหรับตัวแปรเชิงคุณภาพที่ต้องทำให้อยู่ในรูปไบนารีก่อน ค่าความแปรปรวนของตัวแปรรูปแบบไบนารี X (s_X^2) มีค่าเท่ากับ $p(1-p)$ ให้ p แทนความน่าจะเป็นที่ x = 1 และ N แทนขนาดตัวอย่างของ X

การกำหนดค่า Threshold สำหรับวิธีนี้ขึ้นอยู่กับชุดข้อมูลที่ใช้ โดยอาจทำการเลือกทดสอบหลายเกณฑ์เช่น 0.05 0.1 0.15 แล้วเลือกผลที่ดีที่สุด สำหรับใน sklearn ค่า Threshold ค่าเริ่มต้น (default) เท่ากับ 0 กล่าวคือ ตัวแปรต้นใดในชุดข้อมูลที่เป็นค่าคงที่ในทุก N ตัวอย่าง จะถูกคัดออก

ตัวอย่าง ตัวแปรต้น C และ D จะเป็นตัวแปรต้นที่มีความแปรปรวนต่ำ กล่าวคือมีความสำคัญต่อการทำนายค่าตัวแปรตามน้อยกว่าตัวแปรต้น A และ B ดังตารางที่ 5

ตารางที่ 5 ตัวอย่างข้อมูลที่ใช้สำหรับการคัดเลือกตัวแปรด้วยวิธี Variance threshold

A	B	C	D
1	10	0	1
5	2	0	1
3	8	0	1
6	3	0	1

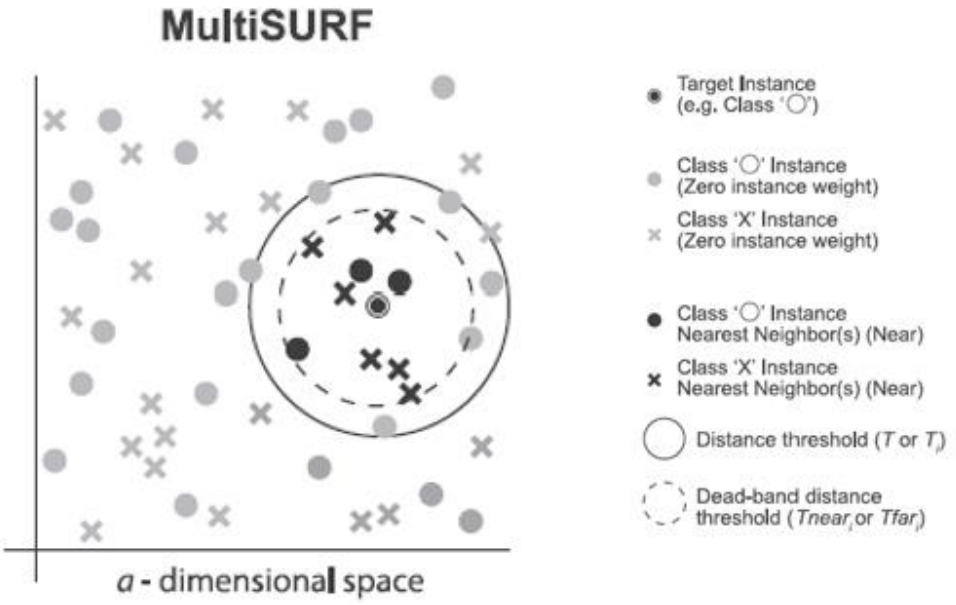
ขั้นตอนการคัดเลือกตัวแปรด้วย Variance Threshold มีดังนี้

ขั้นที่1 หาค่าความแปรปรวนของตัวแปรต้นทุกตัว

ขั้นที่2 หาค่าความแปรปรวนของตัวแปรต้นแต่ละตัว จากนั้นทำการคัดเลือกตัวแปรต้นจำนวน k ลำดับแรกที่มีค่าความแปรปรวนสูงสุด เนื่องจากตัวแปรต้นที่มีความแปรปรวนต่ำจะมีความสำคัญต่อการทำนายตัวแปรตามน้อยกว่าตัวแปรต้นที่มีความแปรปรวนสูง

MultiSURF

การคัดเลือกตัวแปรด้วยวิธี MultiSURF เหมาะสำหรับข้อมูลการจำแนกประเภท (Classification) สามารถใช้ได้กับตัวแปรเชิงปริมาณ (Quantitative) และตัวแปรเชิงคุณภาพ (Qualitative) ถ้าหากเป็นข้อมูลเชิงคุณภาพนั้นจะต้องทำการแปลงให้อยู่ในรูปไบนารีก่อน แล้วจึงหาค่าความสำคัญของตัวแปรต้น



รูปที่ 15 ตัวอย่าง MultiSURF ให้ a แทนจำนวนตัวแปรต้นในชุดข้อมูล และ k = 3

ที่มา: (Urbanowicz et al., 2018)

การหาความสำคัญของตัวแปรต้นด้วยวิธี MultiSURF มีแนวคิดที่ว่า ค่าของตัวแปรต้นดังกล่าวจะต้องทำให้เกิดการแบ่งประเภท (Class) ของตัวแปรตามได้อย่างชัดเจน และตัวแปรต้นที่มีค่าใกล้เคียงกันจะมีตัวแปรตามที่อยู่ในประเภทเดียวกัน ตัวอย่างการหาค่าความสำคัญของตัวแปรต้น

จากข้อมูลที่มีขนาดตัวอย่างเท่ากับ N และมีจำนวนตัวแปรต้นเท่ากับ P ให้ค่าของตัวแปรตาม y มีลักษณะไบนารี [0,1] มีขั้นตอนดังนี้

ขั้นที่1 ให้น้ำหนักเริ่มต้น (Initialize weight) หรือคะแนน (Score) มีค่าเท่ากับศูนย์ โดยเก็บในรูปแบบของอาร์เรย์ (Array) ที่มีขนาดเท่ากับ (1,P) ดังตารางที่ 6 เพื่อใช้ในการจัดลำดับความสำคัญของตัวแปรต้นแต่ละตัว

ขั้นที่2 ใช้ Nearest-Neighbors ในการหาจำนวนค่าสังเกตุที่มีตัวแปรตามเหมือนและต่างกันในบริเวณรอบค่าเป้าหมาย (Target instance$_i$) ให้ i มีค่าเท่ากับ 1,..,N

ขั้นที่3 นำค่าตัวอย่างจากชุดข้อมูลมาจำนวน 2 ค่าคือ Xi_1 และ Xi_2 ให้ i = 1,..,N

ขั้นที่4 หาระยะห่างระหว่างค่า Xi_1 และ Xi_2 ในขั้นที่ 3

ขั้นที่5 หาค่าเฉลี่ยสำหรับระยะห่างระหว่าง Xi_1 และ Xi_2 ในขั้นที่ 4

ขั้นที่6 หาค่าเบี่ยงเบนมาตรฐานของระยะห่างระหว่างค่า Xi_1 และ Xi_2 ขั้นที่ 4

ขั้นที่7 หาค่า Dead-band distance threshold (เส้นประวงกลม) จากผลต่างของค่าในขั้นที่ 5 และ 6 เพื่อนำไปใช้ในการประเมินการทำนาย

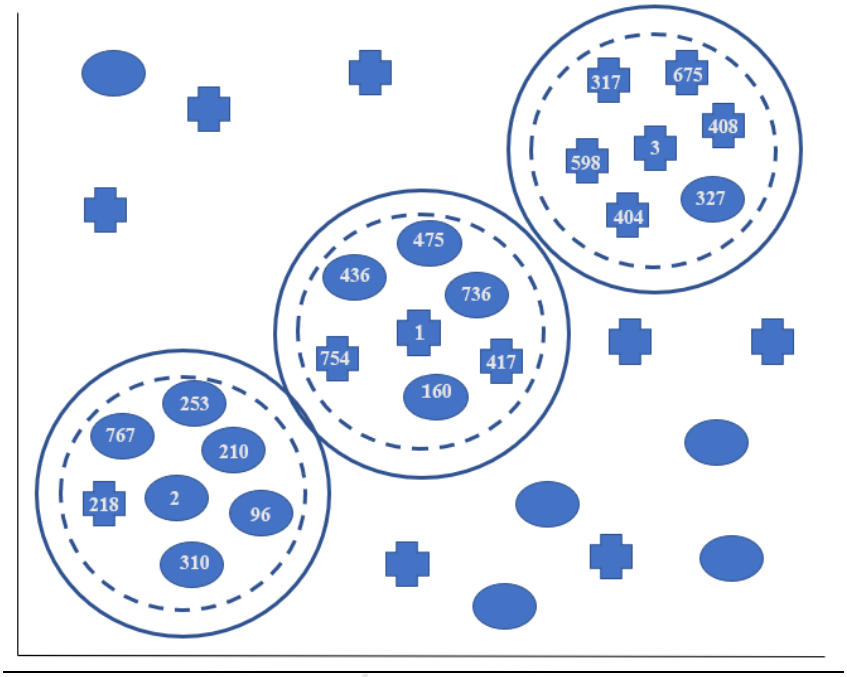
ขั้นที่8 นำค่าสังเกตมาจำนวน 1 ค่า หรือเรียกว่าค่าเป้าหมาย (Target instance$_i$) เมื่อ i = 1,...,N จากนั้น หาระยะห่างจากค่าสังเกตข้างเคียงในระยะที่สนใจ ในที่นี้กำหนดให้ใช้ระยะ k = 7 โดยทำการพิจารณาทีละคู่ค่าสังเกต กล่าวคือถ้าหากระยะห่างระหว่างค่าสังเกตคู่นั้นมีค่าน้อยกว่าค่า Dead-band Distance Threshold และทั้ง 2 ค่านั้นมีประเภทของตัวแปรตามชนิดเดียวกัน ค่าสังเกตตัวดังกล่าวจะถูกจัดเก็บใน Hit ในกรณีที่ระยะห่างของ 2 จุดดังกล่าวมีค่าน้อยกว่า Dead-band distance Threshold แต่มีประเภทของตัวแปรตามต่างชนิดกัน เช่น $Target\ instance_{i=1}$ มีค่า y เท่ากับ 1 แต่ค่าสังเกตอีกตัวหนึ่งมีค่า y เท่ากับ 0 ค่าสังเกตดังกล่าวจะถูกเก็บใน Miss

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตัวอย่าง หา Hit และ Miss สำหรับชุดข้อมูล Diabetes มีขนาด (N) 768 ตัวอย่าง และตัวแปรต้น (P) 8 ตัว ตัวอย่างค่าของข้อมูลแสดงดังรูปที่ 17 พบว่า $\mathrm{Target\ instance}_{i=1}$ มีค่า y เท่ากับ 1

```
miss {0: [436, 475, 160, 756], 1: []}
hit [754, 417]
```

จากที่กำหนดให้ข้อมูลข้างเคียง 7 ตัว (k=7) ใน 7 ตัวดังกล่าว มี 2 ตัวที่ค่า y เท่ากับ 1 เช่นกันจึงถูกจัดเก็บใน Hit ได้แก่ $\mathbf{Instance}_{i=754,417}$ และมี 4 ตัว ที่ค่า y เท่ากับ 0 ซึ่งต่างกับค่า $\mathbf{Target\ instance}_{i=1}$ จึงถูกจัดกับใน Miss ได้แก่ $\mathbf{Instance}_{i=436,475,160,756}$ ดังแสดงในรูปที่ 16



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

รูปที่ 16 การระบุตัวแปรเป็น Hit และ Miss ใน MultiSURF

ขั้นที่9 นำ Hit และ Miss ที่ได้จากทุกๆ ค่าเป้าหมาย ($\mathbf{Target\ instance}_i$) ในการหาผลต่างระหว่าง $\mathbf{Target\ instance}_i$ และค่าสังเกตข้างเคียงที่อยู่ใน Hit จะได้ $\frac{diff(A,R_i,H)}{k}$ หรือ Hit term ดังสมการที่ (2.1L) สำหรับผลต่างระหว่าง $\mathbf{Target\ instance}_i$ และค่าสังเกตข้างเคียงที่อยู่ใน Miss จะได้ $\frac{diff(A,R_i,S)}{N \cdot k}$ หรือ Miss Term ดังสมการที่ (2.1L)

ให้ A แทน ตัวแปรต้น f_j, j = 1,...,P

R_i แทน ค่าเป้าหมาย (Target instance) ดังตัวอย่างข้างต้น รูปที่ 16 และ ตารางที่ 6 R_i คือ $\mathbf{Target\ instance}_i$, i = 1,2,3

H แทน ค่าที่อยู่รอบระยะ k Nearest Neighbors โดยมีประเภทของตัวตามชนิดเดียวกัน (Hit)

S แทน ค่าที่อยู่รอบระยะ k Nearest Neighbors โดยมีประเภทของตัวตามต่างชนิดกัน (Miss)

N แทน จำนวนค่าสังเกตทั้งหมด

k แทน จำนวนค่าสังเกตข้างเคียงที่กำหนดไว้ ในตัวอย่างนี้ k = 7

ขั้นที่10 นำ Hit term และ Miss term จากขั้นที่ 9 มาปรับน้ำหนักที่กำหนดไว้ในขั้นที่ 1 ใหม่ได้ดังนี้

$$W[A] = W[A] - \frac{diff(A,R_i,H)}{k} + \frac{diff(A,R_i,S)}{N \cdot k} \quad (2.1L)$$

ให้ W[A] แทน ค่าน้ำหนักของตัวแปรต้น “A”

สำหรับตัวแปรมีลักษณะค่าไม่ต่อเนื่อง (Discrete)

$$diff(A, I_1, I_2) = \begin{cases} 0, & when value(A, I_1) = value(A, I_2) \\ 1, & otherwise \end{cases}$$

สำหรับตัวแปรมีลักษณะค่าต่อเนื่อง (Continuous)

$$diff(A, I_1, I_2) = |value(A, I_1) - value(A, I_2)|$$

ให้ $I_1 = R_i$ และ $I_2 = H$ หรือ S

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตัวอย่าง การหาน้ำหนักของตัวแปรจากการเลือกค่าสังเกต i = 1 จากชุดข้อมูล Diabetes ขนาด (N) 768 และตัวแปรต้น (P) 8 ตัว ดังรูปที่ 16,17 และตารางที่ 6

I) Hit และ Miss เมื่อ $\text{Target instance}_{i=1}$

```
miss {0: [436, 475, 160, 756], 1: []}
hit [754, 417]
```

II) ค่าของ Hit Term $\frac{diff(A,R_i,H)}{k}$ ในสมการ (2.1L) เมื่อ $\text{Target instance}_{i=1}$ มีค่าเท่ากับ

[0.571 1.429 2.286 0.857 0. 0.871 0.037 2.571]

โดยเรียงลำดับตามตัวแปรต้น Pregnancies, Glucose, Blood Pressure, Skin Thinkness, Insulin, BMI, Diabetes Pedigree Function, Age ตามลำดับ เช่น ค่าแรก 0.571 ได้จากค่า hit term ของตัวแปร Pregnancies ดัง [3] ในตารางที่ 6 หารด้วย k โดยที่ k=7 ค่า Hit term สำหรับทุกตัวแปรต้นแสดงดัง [9] ในตารางที่ 6

III) ค่าของ Miss Term $\frac{diff(A,R_i,S)}{N \cdot k}$ ในสมการ (2.1L) เมื่อ $\text{Target instance}_{i=1}$ มีค่าเท่ากับ

[0.003 0.006 0.011 0.004 0.0 0.003 0.000 0.008]

เช่น ค่าแรก 0.003 ได้จากค่า miss term ของตัวแปร Pregnancies ดัง [8] ในตารางที่ 6 หารด้วย N*k โดยที่ N = 768, k=7 ค่า miss term สำหรับทุกตัวแปรแสดงดัง [10] ในตารางที่ 6

VI) ค่าน้ำหนักที่ปรับในรอบที่ i = 1 มีค่าเท่ากับ น้ำหนัก เริ่มต้น (Initialize weight) ที่มีค่าศูนย์นำมาลบด้วย Hit Term จาก II) และบวกด้วย Miss Term จาก III) อ้างอิงสมการที่ (2.1L) ได้น้ำหนักของตัวแปรต้นแต่ละตัวใหม่ดังนี้

[-0.568 -1.422 -2.274 -0.854 0.0 -0.869 -0.036 -2.563]

เมื่อ $\text{Target instance}_{i=2}$ ให้ทำซ้ำใน I) โดยนำค่าน้ำหนักที่ได้จาก VI) มาเป็นตัวตั้งต้นในการปรับน้ำหนักใหม่ต่อจากขั้นตอนที่ II-III) และทำซ้ำจนกว่าครบ i = N

ขั้นที่11 จากขั้นที่ 10 เมื่อทำซ้ำครบ N ครั้ง ค่าน้ำหนัก (W) ของตัวแปรต้นแต่ละตัว จะถูกนำไปใช้ในการจัดลำดับความสำคัญของตัวแปร

ตารางที่ 6 ตัวอย่างการจัดลำดับความสำคัญของตัวแปรต้นในชุดข้อมูล Diabetes เมื่อมี Target Instance 3 ตัว

	Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**
Initialize Weight	0	0	0	0	0	0	0	0
Target Instance i =1, y=1								
hit [754,417]								
miss {0: [436, 475, 160, 756], 1:[ ]}								
for hit [754]								
X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h, :]	8	154	78	32	0	32.4	0.443	45
[1] = abs (X[i, :] - X[h, :])	2	6	6	3	0	1.2	0.184	5
for hit [417]								
X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h, :]	4	144	82	32	0	38.5	0.554	37
[2] = abs (X[i, :] - X[h, :])	2	4	10	3	0	4.9	0.073	13
[3] hit term = [1]+[2]	4	10	16	6	0	6.1	0.257	18
for miss [436]								
X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h,:]	12	140	85	33	0	37.4	0.244	41
[4] = abs (X [i,:] - X[h, :])	6	8	13	2	0	3.8	0.383	9
for miss [475]								
X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h, :]	0	137	84	27	0	27.3	0.231	59
[5] = abs (X[i, :] - X[h, :])	6	11	12	8	0	6.3	0.396	9
for miss [160]								

จุฬาลงกรณ์มหาวิทยาลัย

X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h, :]	4	151	90	38	0	29.7	0.294	36
[6] = abs (X [i, :] - X[h, :])	2	3	18	3	0	3.9	0.333	14
for miss [756]								
X [i, :]	6	148	72	35	0	33.6	0.627	50
X [h, :]	7	137	90	41	0	32	0.391	39
[7] = abs (X[i, :] - X[h, :])	1	11	18	6	0	1.6	0.236	11
[8] miss term (Almayyan) = [4] + [5] + [6] + [7]	15	33	61	19	0	15.6	1.348	43
[9] diff (A, Ri, H)/k = [3]/k, k=7	0.571	1.429	2.286	0.857	0	0.871	0.037	2.571
[10] diff (A, Ri, S)/N*k = [8] /768*7, N=768, k=7	0.003	0.006	0.011	0.004	0	0.003	0	0.008
[11] score = Initialize Weight - [9] + [10]	**-0.568**	**-1.423**	**-2.275**	**-0.853**	**0**	**-0.868**	**-0.037**	**-2.563**
Target Instance i =2, y=0								
hit [253, 210, 96, 310, 767]								
miss {0: [ ], 1: [218]}								
hit term (calculation same as [1] to [3])	8	25	16	14	0	19.5	0.469	37
miss term (Araújo et al.) (calculation same as [4] to [8])	4	0	8	7	0	2.4	0.873	1
[12] = diff (A, Ri, H)/k	1.143	3.571	2.286	2.000	0.000	2.786	0.067	5.286
[13] = diff (A, Ri, S)/N*k	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000

[14] Score = [11]-[12] + [13]	**-1.710**	**-4.994**	**-4.559**	**-2.852**	**0.000**	**-3.653**	**-0.104**	**-7.849**
Target Instance i =3, y=1								
hit [317, 675, 408, 404, 598]								
miss {0: [327], 1: []}								
hit term (calculation same as [1] to [3])	17	52	36	0	0	40.5	2.311	26
miss term (Almayyan) (calculation same as [4] to [8])	2	4	6	0	0	11.8	0.472	5
[15] = diff (A, Ri, H) /k	2.429	7.429	5.143	0.000	0.000	5.786	0.330	3.714
[16] = diff (A, Ri, S) /N*k	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000
[17] Score = [14]-[15] + [16]	**-4.138**	**-12.423**	**-9.701**	**-2.850**	**0.000**	**-9.439**	**-0.434**	**-11.563**
Result Feature Rank	**4**	**8**	**6**	**3**	**1**	**5**	**2**	**7**

หมายเหตุ ในขั้นตอนการหา Hit Term และ Miss Term ใน Target Instance = 2 และ 3 ทำเช่นเดียวกับในการหา Hit term และ Miss Term ใน Target Instance = 1

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.1.3.2 Wrapper

นำการเรียนรู้ของเครื่อง (Machine learning) มาใช้ประเมินผลการทำนายที่เกิดจากเซตของตัวแปรต้นในรูปแบบต่าง ๆ ในกรณีที่ตัวแปรต้นในชุดข้อมูลมีจำนวนน้อยจะสามารถสร้างตัวแบบด้วยการเพิ่มทีละ 1 ตัวแปร (Forward feature selection) หรือสร้างตัวแบบด้วยการลดทีละ 1 ตัวแปร (Backward elimination) วิธีคัดเลือกตัวแปรอื่น ๆ ที่อยู่ในประเภท Wrapper ได้แก่ RFE, Boruta

สำหรับตัวแบบที่มีความซับซ้อนสูง (Black-Box model) จะทำให้เกิดปัญหาการตีความเกี่ยวกับขนาดของอิทธิพลของตัวแปรต้นที่มีผลต่อตัวแปรตามนั้นทำได้ยาก ทำให้ SHAP เป็นวิธีที่ถูกใช้ในการอธิบายการทำนายตัวแบบที่มีความซับซ้อนสูงเหล่านั้น ทั้งในระดับ Local และ Global ทำให้การอธิบายการทำนายในระดับ Global นั้นสามารถนำมาจัดลำดับความสำคัญของตัวแปรต้นได้

ตามความสามารถในการอธิบายการทำนายที่เกิดจากแต่ละตัวแปรต้น ทำให้ SHAP เป็นวิธีการคัดเลือกตัวแปรที่อยู่ในประเภท Wrapper เช่นกัน

Shapley Additive Explanations (SHAP)

การหาค่า Shapley ของตัวแปรต้นอ้างอิงจากทฤษฎีเกมแบบร่วมมือกัน (Cooperative game) เป็นการร่วมมือกันระหว่างผู้เล่น เพื่อให้ได้ผลประโยชน์ (Payout) สูงสุด ถูกนำมาใช้กับการอธิบายการทำนายของตัวแบบโดยการหาว่าตัวแปรต้นแต่ละตัวส่งผลต่อทำนายของตัวแปรตามอย่างไร โดยให้ผู้เล่น (Player in game) แทนตัวแปรต้นในชุดข้อมูล และผลประโยชน์ (Payout) แทนผลการทำนายผ่านตัวแบบ ($\hat{f}(X_i), i = 1, \ldots, N$)

ตัวอย่างชุดข้อมูลการจำแนกแบบไบนารี ใช้สำหรับการอธิบายการทำนายของตัวแบบระดับ Local ด้วยค่า Shapley ในการอธิบายการทำนาย $\hat{f}(X_i)$, i = 1, ..., N เช่น ในการทำนาย $\hat{f}(X_1)$ สามารถอธิบายได้ด้วยด้วยค่า Shapley ของตัวแปรต้นแต่ละตัว ($\hat{\phi}_{j,1}$, j = 1,...,P) ดังตารางที่ 7

ตารางที่ 7 ให้ตัวอย่างชุดข้อมูลที่มีขนาด (N) 5 ตัวอย่าง และมีจำนวนตัวแปรต้น (P) 4 ตัว ($x_i^{[j]}$, i=1, ... ,5 และ j=1, ..., 4)

ค่าทำนายตัวแปรตาม	ตัวแปรต้นที่1	ตัวแปรต้นที่2	ตัวแปรต้นที่3	ตัวแปรต้นที่4
$\hat{f}(X_1) = 1$	$x_1^{[1]}$	$x_1^{[2]}$	$x_1^{[3]}$	$x_1^{[4]}$
$\hat{f}(X_2) = 0$	$x_2^{[1]}$	$x_2^{[2]}$	$x_2^{[3]}$	$x_2^{[4]}$
$\hat{f}(X_3) = 1$	$x_3^{[1]}$	$x_3^{[2]}$	$x_3^{[3]}$	$x_3^{[4]}$
$\hat{f}(X_4) = 0$	$x_4^{[1]}$	$x_4^{[2]}$	$x_4^{[3]}$	$x_4^{[4]}$
$\hat{f}(X_5) = 1$	$x_5^{[1]}$	$x_5^{[2]}$	$x_5^{[3]}$	$x_5^{[4]}$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ในการหาค่า Shapley ของตัวแปรต้นแต่ละตัว ($\emptyset_j$, j = 1,...,4) ที่ใช้ในการอธิบายทำนาย $\hat{f}(X_i)$ มีขั้นตอนดังนี้

ขั้นที่1 สร้างรูปแบบการรวมกลุ่มที่เป็นไปได้ทั้งหมด (Coalition) ซึ่งแต่ละกลุ่มจะไม่นับรวมตัวแปรที่ต้องการหาค่า Shapley ให้จำนวนกลุ่มทั้งหมดด้วยแทนด้วย C มีจำนวนเท่ากับ 2^{P-1} รูปแบบ เช่น การอธิบายการทำนาย $\hat{f}(X_1)$ จากค่า Shapley ของตัวแปรต้นที่ 4 $(\hat{\phi}_4)$ จะได้ตัวแปรต้นที่ถูกนำมา

สร้างการรวมกลุ่มที่เป็นไปได้ทั้งหมดมี 3 ตัว คือ $x_1^{[1]}$, $x_1^{[2]}$, $x_1^{[3]}$ จำนวนการรวมกลุ่ม (Coalition) ที่เกิดขึ้นมีค่าเท่ากับ 2^{4-1} หรือ C เท่ากับ 8 รูปแบบ c = 1, ..., C ดังตาราง 8

ตารางที่ 8 การสร้างรูปแบบการรวมกลุ่มที่เป็นไปได้ (Coalition) สำหรับการหาค่า Shapley ของตัวแปรต้นที่ 4

c = 1	{ }
c = 2	$\{x_i^{[1]}\}$
c = 3	$\{x_i^{[2]}\}$
c = 4	$\{x_i^{[3]}\}$
c = 5	$\{x_i^{[1]}, x_i^{[2]}\}$
c = 6	$\{x_i^{[1]}, x_i^{[3]}\}$
c = 7	$\{x_i^{[2]}, x_i^{3]}\}$
c = 8	$\{x_i^{[1]}, x_i^{[2]}, x_i^{[3]}\}$

<u>ขั้นที่2</u> หาความต่างของค่าทำนาย (Marginal contribution) ในกรณีที่มีตัวแปรต้นที่ j และไม่มีตัวแปรต้นที่ j ในรูปแบบการรวมกลุ่ม (Coalitional) ที่ c = 1,...,C

$$\phi_j^c = \hat{f}\left(x_{+j}^c\right) - \hat{f}\left(x_{-j}^c\right) \qquad (2.1M)$$

ให้ $\hat{f}\left(x_{+j}^c\right)$ แทน ค่าทำนายผ่านเซตตัวแปรที่อยู่ในกลุ่ม c แบบมีตัวแปร j

$\hat{f}\left(x_{-j}^c\right)$ แทน ค่าทำนายผ่านเซตตัวแปรที่อยู่ในกลุ่ม c แบบไม่มีตัวแปร j

จากนั้นทำการหาค่า Shapley ของตัวแปร j จากค่าเฉลี่ยของ ϕ_j^c, c = 1, ..., C

$$\hat{\phi}_j = \frac{1}{C}\sum_{c=1}^{C}\left(\hat{f}\left(x_{+j}^c\right) - \hat{f}\left(x_{-j}^c\right)\right) \qquad (2.1N)$$

เช่น ค่า Shapley ของตัวแปรต้นที่ 4 $(\hat{\phi}_4)$ ที่ใช้อธิบายการทำนายค่า $\hat{f}(X_1)$ จากขั้นที่1 และ 2 แสดงดังตารางที่ 9

ตารางที่ 9 การหาค่าความต่างของค่าทำนายในแต่ละรูปแบบการรวมกลุ่ม กรณีตัวแปรต้น j = 4

การรวมกลุ่ม (Coalition)	**มีตัวแปร j**	**ไม่มีตัวแปร j**	**ความต่างของค่าทำนาย**
c = 1	$\{ x_1^{[4]} \}$	{ }	$\hat{f}\left(x_{+j=4}^{c=1}\right)$ - $\hat{f}\left(x_{-j=4}^{c=1}\right)$ = $\emptyset_{j=4}^{c=1}$
c = 2	$\{x_1^{[1]}, x_1^{[4]}\}$	$\{x_1^{[1]}\}$	$\hat{f}\left(x_{+j=4}^{c=2}\right)$ - $\hat{f}\left(x_{-j=4}^{c=2}\right)$ = $\emptyset_{j=4}^{c=2}$

c = 3	$\{x_1^{[2]}, x_1^{[4]}\}$	$\{x_1^{[2]}\}$	$\hat{f}(x_{+j=4}^{c=3})$ - $\hat{f}(x_{-j=4}^{c=3})$ = $\emptyset_{j=4}^{c=3}$
c = 4	$\{x_1^{[3]}, x_1^{[4]}\}$	$\{x_1^{[3]}\}$	$\hat{f}(x_{+j=4}^{c=4})$ - $\hat{f}(x_{-j=4}^{c=4})$ = $\emptyset_{j=4}^{c=4}$
c = 5	$\{x_1^{[1]}, x_1^{[2]}, x_1^{[4]}\}$	$\{x_1^{[1]}, x_1^{[2]}\}$	$\hat{f}(x_{+j=4}^{c=5})$ - $\hat{f}(x_{-j=4}^{c=5})$ = $\emptyset_{j=4}^{c=5}$
c = 6	$\{x_1^{[1]}, x_1^{[3]}, x_1^{[4]}\}$	$\{x_1^{[1]}, x_1^{[3]}\}$	$\hat{f}(x_{+j=4}^{c=6})$ - $\hat{f}(x_{-j=4}^{c=6})$ = $\emptyset_{j=4}^{c=6}$
c = 7	$\{x_1^{[2]}, x_1^{3]}, x_1^{[4]}\}$	$\{x_1^{[2]}, x_1^{3]}\}$	$\hat{f}(x_{+j=4}^{c=7})$ - $\hat{f}(x_{-j=4}^{c=7})$ = $\emptyset_{j=4}^{c=7}$
c = 8	$\{x_1^{[1]}, x_1^{[2]}, x_1^{[3]}, x_1^{[4]}\}$	$\{x_1^{[1]}, x_1^{[2]}, x_1^{[3]}\}$	$\hat{f}(x_{+j=4}^{c=8})$ - $\hat{f}(x_{-j=4}^{c=8})$ = $\emptyset_{j=4}^{c=8}$

จากตารางที่ 9 ค่า Shapley ของตัวแปรต้นที่ 4 ในการทำนาย $\hat{f}(X_1)$ เท่ากับ $\hat{\phi}_4 = \frac{1}{8}\sum_{c=1}^{8} \emptyset_{j=4}^{c}$ ในกรณีที่หาค่า Shapley ของตัวแปรต้นที่ j = 1, ..., 3 ($\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$) ให้ทำเช่นเดียวกับกรณีที่หา $\hat{\phi}_4$ ที่แสดงข้างต้น และเมื่อหาค่า Shapley ของตัวแปร j = 1, ..., 4 จะได้ตารางที่แสดงผลค่า Shapley ดังตารางที่ 10

ตารางที่ 10 ค่า Shapley ของตัวแปรต้นใช้อธิบายการทำนายตัวแปรตาม $(\hat{\phi}_{j,i})$, j = 1, ..., 4 และ i =1, ..., 5

ค่าทำนายตัวแปรตาม	**ตัวแปรต้นที่1**	**ตัวแปรต้นที่2**	**ตัวแปรต้นที่3**	**ตัวแปรต้นที่4**
$\hat{f}(X_1)$	$\hat{\phi}_{1,1}$	$\hat{\phi}_{2,1}$	$\hat{\phi}_{3,1}$	$\hat{\phi}_{4,1}$
$\hat{f}(X_2)$	$\hat{\phi}_{1,2}$	$\hat{\phi}_{2,2}$	$\hat{\phi}_{3,2}$	$\hat{\phi}_{4,2}$
$\hat{f}(X_3)$	$\hat{\phi}_{1,3}$	$\hat{\phi}_{2,3}$	$\hat{\phi}_{3,3}$	$\hat{\phi}_{4,3}$
$\hat{f}(X_4)$	$\hat{\phi}_{1,4}$	$\hat{\phi}_{2,4}$	$\hat{\phi}_{3,4}$	$\hat{\phi}_{4,4}$
$\hat{f}(X_5)$	$\hat{\phi}_{1,5}$	$\hat{\phi}_{2,5}$	$\hat{\phi}_{3,5}$	$\hat{\phi}_{4,5}$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ค่า Shapley ดังกล่าวสามารถนำมาอธิบายการทำนายตัวแปรตาม ($\hat{f}(X_i)$) ถึงความต่างจากค่า Base Value $(E[\hat{f}(X)])$ โดย Base Value เป็นค่าทำนายของตัวแปรตามกรณีที่ไม่พิจารณาค่าของตัวแปรต้นใด ๆ ร่วมด้วย การอธิบายการทำนายในลักษณะนี้สำหรับการเรียนรู้ของเครื่องเรียกว่า SHAP (Shapley additive explanations)

จากตัวอย่างการอธิบายการทำนายตัวแปรตาม $\hat{f}(X_1)$ ได้ค่า Shapley ดังนี้

$$\sum_{j=1}^{4} \phi_{j,i}(\hat{f}) = \hat{f}(x) - E[\hat{f}(X)] \qquad (2.10)$$

ค่า Shapley ที่ได้แสดงถึงความสามารถในการอธิบายการทำนายตัวแปรตามในระดับ Local ซึ่งเป็นการอธิบายการทำนายตัวแปรตามแบบแยกกัน $\hat{f}(X_i)$, i = 1, ..., 5 หากต้องการหาค่า

ความสำคัญของตัวแปรต้นในการอธิบายการทำนายระดับ Global จะใช้วิธีค่าเฉลี่ยค่าสัมบูรณ์ค่า Shapley ของตัวแปรต้น j = 1, ..., 4 แสดงตัวอย่างในตารางที่ 11

ตารางที่ 11 การหาค่าเฉลี่ยค่าสัมบูรณ์ค่า Shapley ของตัวแปรต้น j = 1,...,4

ค่าทำนายตัวแปรตาม	ตัวแปรต้นที่1	ตัวแปรต้นที่2	ตัวแปรต้นที่3	ตัวแปรต้นที่4
$\hat{f}(X_1)$	$\hat{\phi}_{1,1}$	$\hat{\phi}_{2,1}$	$\hat{\phi}_{3,1}$	$\hat{\phi}_{4,1}$
$\hat{f}(X_2)$	$\hat{\phi}_{1,2}$	$\hat{\phi}_{2,2}$	$\hat{\phi}_{3,2}$	$\hat{\phi}_{4,2}$
$\hat{f}(X_3)$	$\hat{\phi}_{1,3}$	$\hat{\phi}_{2,3}$	$\hat{\phi}_{3,3}$	$\hat{\phi}_{4,3}$
$\hat{f}(X_4)$	$\hat{\phi}_{1,4}$	$\hat{\phi}_{2,4}$	$\hat{\phi}_{3,4}$	$\hat{\phi}_{4,4}$
$\hat{f}(X_5)$	$\hat{\phi}_{1,5}$	$\hat{\phi}_{2,5}$	$\hat{\phi}_{3,5}$	$\hat{\phi}_{4,5}$
ค่าเฉลี่ยค่าสัมบูรณ์ของค่า Shapley =	$\frac{1}{5}\sum_{i=1}^{5}\lvert\hat{\phi}_{1,i}\rvert$	$\frac{1}{5}\sum_{i=1}^{5}\lvert\hat{\phi}_{2,i}\rvert$	$\frac{1}{5}\sum_{i=1}^{5}\lvert\hat{\phi}_{3,i}\rvert$	$\frac{1}{5}\sum_{i=1}^{5}\lvert\hat{\phi}_{4,i}\rvert$

ตัวอย่าง การหาค่าความสำคัญของตัวแปรต้นในชุดข้อมูล Diabetes ขนาดข้อมูล (N) 768 ตัวอย่าง จำนวนตัวแปรต้น (P) 8 ตัว ($x_i^{[j]}$, i = 1, ..., 768 และ j = 1, ..., 8) และตัวแปรตาม Class ที่มีลักษณะการจำแนกแบบไบนารีสำหรับผู้ป่วยที่ไม่เป็นเบาหวาน (y = 0) และ ผู้ป่วยที่เป็นเบาหวาน (y = 1) ดังรูปที่ 17 และเมื่อทำการหาค่า Shapley ของทุกตัวแปรต้นที่ใช้อธิบายการทำนายตัวแปร Class ของผู้ป่วยแต่ละคนได้ดังรูปที่ 18

i	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	class
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.340	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

รูปที่ 17 ชุดข้อมูล Diabetes ขนาด (N) 768 ตัวอย่าง จำนวนตัวแปรต้น (P) 8 ตัว

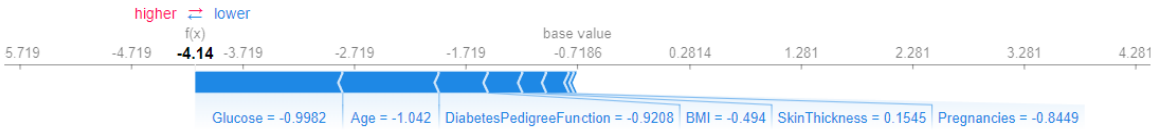
i	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	-0.031201	0.787301	0.103203	0.110871	0.173925	0.310910	0.257932	0.732017
2	-0.246517	-1.841380	0.070312	-0.027001	0.071407	-0.667965	0.075647	0.350570
3	0.494159	2.497543	0.139928	-0.034944	0.063308	-1.063982	0.204836	0.370197
4	-0.201568	-1.324166	-0.026719	-0.214264	-0.044517	-0.309722	-0.438906	-0.864595
5	-0.076191	0.666866	0.372333	0.128129	0.157150	0.762862	-0.004802	0.322181
...	...	...	...	...	...	...	...	...
764	0.331721	-0.726437	-0.109286	-0.181109	0.318510	0.227188	-0.823163	0.033232
765	-0.172094	-0.454295	0.031527	-0.083502	0.159475	0.143563	-0.080323	-0.352028
766	-0.062777	-0.494434	-0.105149	-0.048003	-0.060294	-1.351798	-0.092793	0.209353
767	-0.207881	0.270428	0.122236	0.080975	0.202650	0.393478	0.207647	0.848472
768	-0.211335	-1.400332	-0.029391	-0.042367	0.088087	0.109765	-0.299557	-0.754366

768 rows × 8 columns

รูปที่ 18 ค่า Shapley ของตัวแปรต้นแต่ละตัว ใช้อธิบายการทำนาย Class ของผู้ป่วยแต่ละคน

<u>การอธิบายการทำนายในระดับ Local</u>

ตัวอย่างที่ i = 4 จากรูปที่ 17 ตัวแปรต้นมีค่าดังนี้ Pregnancies=1, Glucose=89, Blood Pressure=66, Skin Thickness=23, Insulin=94, BMI=28.1, Diabetes Pedigree Function=0.167, Age=21 ทำนายว่าผู้ป่วยคนนี้มีโอกาสที่จะไม่เป็นเบาหวานผ่านค่า Shapley ดังรูปที่ 19



รูปที่ 19 อธิบายการทำนายตัวแปรตาม $\hat{f}(X_i)$ ที่ i=4 ด้วยค่า SHAP

อ้างอิงจากสมการ (2.10) ค่า $\mathrm{E}[\hat{\mathrm{f}}(\mathrm{X})]$ หรือ Base Value มีค่าเท่ากับ -0.7186 และ $\hat{\mathrm{f}}(\mathrm{X}_4)$ เท่ากับ -4.14 จะได้ค่า SHAP เท่ากับ

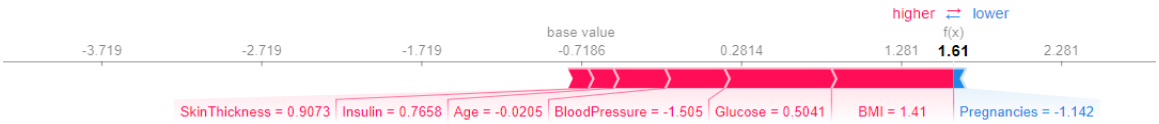
$$\begin{aligned}\sum_{j=1}^{8} \phi_{j,4}(\hat{f}) &= \hat{f}(X_4) - \mathrm{E}[\hat{f}(X_4)] \\ &= -4.14 - (-0.71) \\ &= -3.43\end{aligned}$$

ซึ่งมีค่าเท่ากับผลรวมของค่า Shapley ภายในแถวที่ i = 4 ในรูปที่ 18 จากค่าทำนาย $\hat{f}(X_4) = log(odds) = -4.14$ อ้างอิงสมการ (2.1H) จะได้ความน่าจะเป็นที่ผู้ป่วยคนนี้จะเป็นเบาหวานเท่ากับ

$$\begin{aligned} \text{Probability} &= \frac{1}{1+ e^{-(log(odds))}} \\ &= \frac{1}{1+ e^{-(-4.14)}} \\ &= 0.0157 \end{aligned}$$

ความน่าจะเป็นที่ผู้ป่วยคนนี้จะไม่เป็นเบาหวานเท่ากับ 1- 0.0157 = 0.9843

ตัวอย่างที่ i = 5 จากข้อมูลรูปที่ 17 ตัวแปรต้นมีค่าดังนี้ Pregnancies=0, Glucose=137, Blood Pressure=40, Skin Thickness=35, Insulin=168, BMI=43.1, Diabetes Pedigree Function=2.288, Age=33 ทำนายว่าผู้ป่วยคนนี้มีโอกาสเป็นเบาหวานผ่านค่า Shapley ดังรูปที่ 20



รูปที่ 20 อธิบายการทำนายตัวแปรตาม $\hat{f}(X_i)$ ที่ i = 5 ด้วยค่า SHAP

อ้างอิงจากสมการ(2.10) ค่า $\mathrm{E}[\hat{\mathrm{f}}(\mathrm{X})]$ หรือ Base Value มีค่าเท่ากับ -0.7186 และ $\hat{\mathrm{f}}(\mathrm{X}_5)$ เท่ากับ 1.61 จะได้ค่า Shapley เท่ากับ

$$\begin{aligned} \textstyle\sum_{j=1}^{8} \phi_{j,5}(\hat{f}) &= \hat{\mathrm{f}}(\mathrm{X}_5) - \mathrm{E}[\hat{\mathrm{f}}(\mathrm{X}_5)] \\ &= 1.61 - (-0.71) \\ &= 2.32 \end{aligned}$$

ซึ่งมีค่าเท่ากับผลรวมของค่า Shapley ภายในแถวที่ i = 5 ในรูปที่ 18 จากค่าทำนาย $\hat{\mathrm{f}}(\mathrm{X}_5) = log(odds) = 1.61$ อ้างอิงสมการ (2.1H) จะได้ความน่าจะเป็นที่ผู้ป่วยคนนี้จะเป็นเบาหวานเท่ากับ

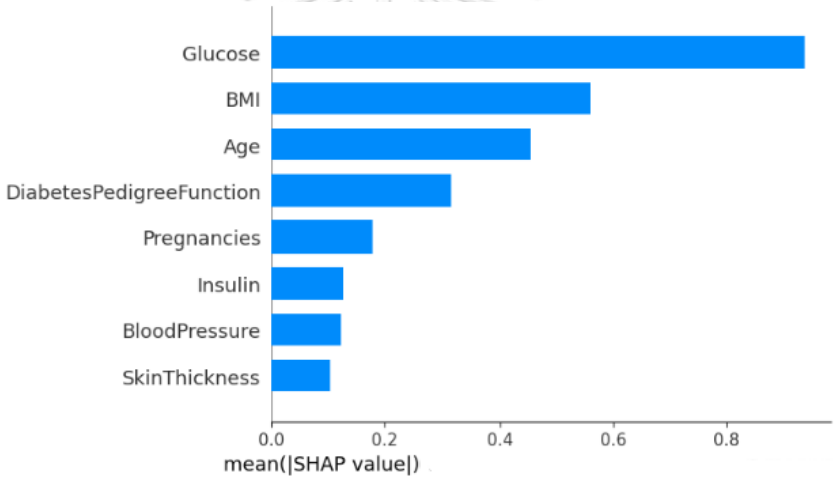
$$\begin{aligned} \text{Probability} &= \frac{1}{1+ e^{-(log(odds))}} \\ &= \frac{1}{1+ e^{-(1.61)}} \end{aligned}$$

$$= 0.8334$$

ความน่าจะเป็นที่ผู้ป่วยคนนี้จะไม่เป็นเบาหวานเท่ากับ 1- 0.8334 = 0.1666

การอธิบายการทำนายในระดับ Global

ในการอธิบายความสำคัญของตัวแปรต้นได้จากการนำค่าเฉลี่ยค่าสัมบูรณ์ค่า Shapley ของตัวแปรต้นแต่ละตัว ดังตัวอย่างในตารางที่ 11 เพื่ออธิบายอิทธิพลของการทำนายผล (impact) ต่อตัวแปรตามด้วยชุดข้อมูล Diabetes พบว่าตัวแปร Glucose มีอิทธิพลต่อการอธิบายการทำนายมากที่สุด และ SkinThickness มีอิทธิพลต่อการอธิบายการทำนายน้อยที่สุด ดังรูปที่ 21



รูปที่ 21 ลำดับความสำคัญ (เรียงลำดับจากความสำคัญมากที่สุดไปน้อยที่สุด) ของตัวแปรต้นที่มีอิทธิพลต่อการอธิบายการทำนายผ่านตัวแบบด้วยวิธี SHAP

Recursive Feature Elimination (RFE)

การคัดเลือกตัวแปรด้วย RFE อาศัยการเรียนรู้ของเครื่องที่เป็นการเรียนรู้แบบทำซ้ำ เพื่อหาตัวแปรต้นหรือกลุ่มของตัวแปรต้น ที่ทำให้ตัวแบบมีประสิทธิภาพทำนายสูงสุด โดยวิธี RFE สามารถใช้ได้กับตัวแปรเชิงปริมาณ (Quantitative) และ เชิงคุณภาพ (Qualitative) การลดจำนวนตัวแปรต้นในชุดข้อมูลด้วยวิธีนี้ทำให้ลดความเสี่ยงที่ทำให้ตัวแบบเกิดปัญหา Overfitting อีกทั้งยังสามารถเพิ่มประสิทธิภาพการทำนายของตัวแบบได้ ขั้นตอนการจัดลำดับความสำคัญของตัวแปรต้นด้วย RFE โดย Sklearn มีดังนี้

ขั้นที่1 แบ่งชุดข้อมูลเป็นชุดข้อมูลฝึกฝนและทดสอบ จากนั้นจึงใช้ข้อมูลชุดฝึกฝนในการสร้างตัวแบบ XGBoost

ขั้นที่2 ใช้ตัวแบบ XGBoost หาค่าความสำคัญจากค่า Gini Importance หรือ Mean Decrease in Gini Impurity (MDI) โดยใช้ klearn.feature_selection.RFE ในจัดลำดับความสำคัญของตัวแปรต้น โดยอัลกอริทึมจะเลือกตัดตัวแปรต้นที่ตัวแบบ XGBoost ระบุค่าความสำคัญน้อยที่สุดออกในแต่ละรอบ

ตัวอย่าง การหาค่าความสำคัญของตัวแปรต้นด้วยวิธี RFE ด้วยชุดข้อมูล Diabetes เช่นเดียวกับในวิธี SHAP กำหนดชื่อตัวแปรดังตารางที่ 12

ตารางที่ 12 ชื่อตัวแปรสำหรับข้อมูลที่ใช้ในการคัดเลือกตัวแปรต้นด้วยวิธี RFE

ตัวแปร	ชื่อตัวแปร
y	Class
f_1	Pregnancies
f_2	Glucose
f_3	Blood Pressure
f_4	Skin Thickness
f_5	Insulin
f_6	BMI
f_7	Diabetes Pedigree Function
f_8	Age

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เริ่มต้นมีตัวแปรดังนี้ : $[f_1\ f_2\ f_3\ f_4\ f_5\ f_6\ f_7\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 1 : $[f_1\ f_2\ f_3\ f_4\ f_5\ f_6\ f_7\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 2 : $[f_1\ f_2\ f_4\ f_5\ f_6\ f_7\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 3 : $[f_2\ f_4\ f_5\ f_6\ f_7\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 4 : $[f_2\ f_4\ f_5\ f_6\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 5 : $[f_2\ f_5\ f_6\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 6 : $[f_2\ f_6\ f_8]$

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 7 : [f_2 f_6]

ตัวแปรต้นที่เหลือจากการทดสอบรอบที่ 8 : [f_2]

ตัวแปรต้นที่ตัวแบบ XGBoost ระบุค่าความสำคัญน้อยสุดจะถูกตัดออกในแต่ละรอบมีลำดับดังนี้ f_3, f_1, f_7, f_4, f_5, f_8, f_6, f_2 ตามลำดับก่อนไปหลัง จะได้ว่าตัวแปรที่มีความสำคัญมากสุดคือตัวแปรที่ถูกตัดออกตัวสุดท้าย ดังนั้นลำดับความสำคัญของตัวแปรต้นที่ถูกจัดผ่านวิธี RFE ได้ดังรูปที่ 22

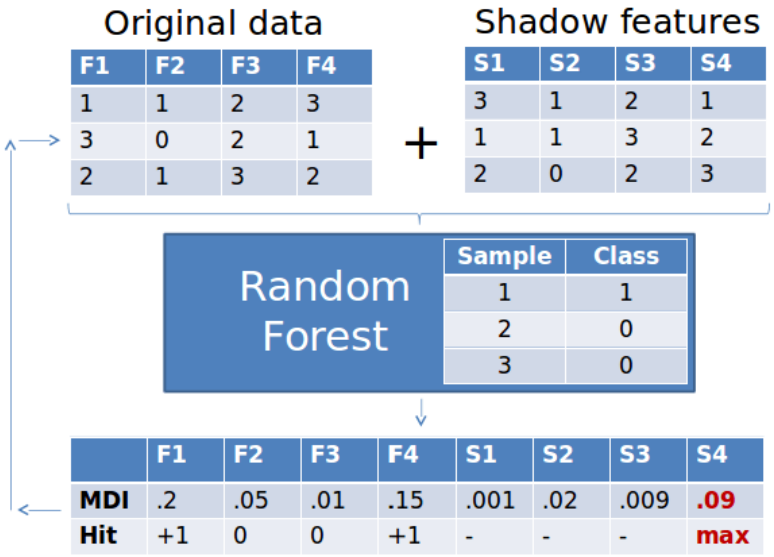
Feature	value_rfe
Glucose	1
BMI	2
Age	3
Insulin	4
SkinThickness	5
DiabetesPedigreeFunction	6
Pregnancies	7
BloodPressure	8

รูปที่ 22 ลำดับความสำคัญของตัวแปรต้นที่มีอิทธิพลต่อการอธิบายการทำนายผ่านตัวแบบด้วยวิธี RFE

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Boruta

การคัดเลือกตัวแปรด้วยวิธี Boruta เหมาะสำหรับตัวแบบที่อยู่ในรูปแบบต้นไม้ตัดสินใจเช่น Random Forest และ XGBoost ซึ่งการคัดเลือกตัวแปรด้วยวิธีนี้สามารถใช้ได้กับตัวแปรเชิงปริมาณ (Quantitative) และ เชิงคุณภาพ (Qualitative ขั้นตอนการจัดลำดับความสำคัญของตัวแปรต้นด้วย Boruta มีดังนี้



รูปที่ 23 ขั้นตอนการทำ Boruta

ที่มา: (DataCamp Team, 2018)

ขั้นที่1 สร้างตัวแปรเงา (Shadow features) ดังรูปที่ 23 ตัวแปร S1-S4 ได้จากตัวแปรต้น F1-F4 ในชุดข้อมูลเดิม (Original Data) จากนั้นทำการสลับตำแหน่ง (Shuffling) ภายในคอลัมน์ของตัวแปรเงา

ขั้นที่2 สร้างตัวแบบ XGBoost จากนั้นหาค่าความสำคัญของตัวแปรต้นในชุดข้อมูลขั้นที่ 1 ด้วย Gini Importance หรือ Mean Decrease in Gini Impurity (MDI) ค่าที่มากหมายถึงยิ่งมีความสำคัญมาก จากนั้นแปลงค่าความสำคัญดังกล่าวเป็นค่า Z-score

ขั้นที่3 อัลกอริทึมจะทำการหาค่าความสำคัญของตัวแปรต้นดั้งเดิม แล้วเลือกตัวแปรต้นใดมีค่า Z-score ของค่าความสำคัญดังกล่าวที่มีค่ามากกว่า ค่า Z-score ที่มากที่สุดของตัวแปรเงา จากนั้นเก็บผลลัพธ์ไว้ในรูปอาร์เรย์ Hit ดัง hit_reg ในตารางที่ 15

ขั้นที่4 ในแต่ละรอบการทำซ้ำ (Iteration) อัลกอริทึมจะเปรียบเทียบค่า Z-score ของตัวแปรเงา และค่า Z-score ของตัวแปรดั้งเดิม ถ้าหาก Z-score ของตัวแปรดั้งเดิมมากกว่าของตัวแปรเงาอย่างมีนัยสำคัญ หมายความว่าตัวแปรนั้นมีความสำคัญ จากรูปที่ 23 พบว่าตัวแปร F1 และ F4 มีความสำคัญ จึงถูกเก็บไว้ใน Hit โดยความต่างของค่า Z-score ที่ยิ่งมากหมายถึงตัวแปรต้นดังกล่าวยิ่งมีความสำคัญมาก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

<u>ตัวอย่าง</u> การหาค่าความสำคัญของตัวแปรต้นโดยใช้วิธี Boruta โดยใช้ชุดข้อมูล Diabetes เช่นเดียวกับวิธี SHAP กำหนดให้ iteration = 10 มีขั้นตอนดังนี้

I) จากตัวแปรต้นในข้อมูลดังเดิมมีทั้งหมด 8 ตัว และตัวแปรเงาที่สร้างขึ้นมาอีก 8 ตัว จึงได้จำนวนตัวแปรต้นรวมทั้งหมด 16 ตัว

II) ค่าความสำคัญของตัวแปรต้นตั้งเดิมแต่ละตัวที่ได้จากตัวแบบ XGBoost ในรอบที่ 1 ถึง 9 ได้ดังนี้

ตารางที่ 13 ค่าความสำคัญของตัวแปรตั้งต้นสำหรับการคัดเลือกตัวแปรด้วยวิธี Boruta

iteration	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
1	0.0546**	0.1719**	0.0564**	0.0543**	0.0686**	0.0961**	0.0547**	0.0811**
2	0.0490	0.1773**	0.0486	0.0589	0.0625	0.0849**	0.0605	0.0729**
3	0.0462	0.193**	0.0490	0.0433	0.0619**	0.0967**	0.0595**	0.0738**
4	0.0407	0.1724**	0.0556	0.0516	0.065**	0.102**	0.0537	0.0851**
5	0.0562	0.1833**	0.0499	0.0596**	0.0684**	0.0907**	0.0531	0.0812**
6	0.0337	0.1633**	0.0513	0.0591**	0.0650**	0.0999**	0.0672**	0.0904**
7	0.0437	0.1853**	0.0462	0.0454	0.0608**	0.0973**	0.0561	0.0883**
8	0.0556	0.1701**	0.0533	0.0684**	0.0655**	0.0937**	0.0580	0.0834**
9	0.0462	0.1617**	0.0463	0.0475	0.0450	0.1107**	0.0523	0.088**

**แสดงถึงค่าความสำคัญของตัวแปรต้นดั้งเดิมที่มีค่ามากกว่าค่ามากสุดของตัวแปรเงาในแต่ละรอบ (* ดังแสดงในตารางที่ 14) ซึ่งจะถูกจัดเก็บใน hit_reg ในตารางที่ 15

ตารางที่ 14 ค่าความสำคัญของตัวแปรเงาสำหรับการคัดเลือกตัวแปรด้วยวิธี Boruta

iteration	Shadow Pregnancies	Shadow Glucose	Shadow Blood Pressure	Shadow Skin Thickness	Shadow Insulin	Shadow BMI	Shadow Diabetes Pedigree Function	Shadow Age
1	0.0523	0.0406	0.0495	0.0448	0.0535*	0.0318	0.0463	0.0435
2	0.0473	0.0645	0.0695*	0.0249	0.0529	0.0498	0.0412	0.0352

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3	0.0400	0.0377	0.0491	0.0576*	0.0435	0.0505	0.0511	0.0471
4	0.0413	0.0472	0.0541	0.0390	0.0471	0.0391	0.0477	0.0585*
5	0.0430	0.0356	0.0580*	0.0518	0.0340	0.0521	0.0354	0.0477
6	0.0533*	0.0476	0.0489	0.0417	0.0427	0.0438	0.0501	0.0419
7	0.0597*	0.0506	0.0449	0.0365	0.0439	0.0382	0.0574	0.0460
8	0.0446	0.0334	0.0413	0.0402	0.0426	0.0428	0.0448	0.0625*
9	0.0525	0.0481	0.0557	0.0588	0.0616*	0.0452	0.0447	0.0357

*แสดงถึงค่าความสำคัญของตัวแปรต้นมากที่สุดในตัวแปรเงาในแต่ละรอบ

III) จัดประเภทของตัวแปรเข้าสู่ Confirmed, Tentative และ Rejected ในที่นี้จะแสดงตัวอย่างการจัดประเภทเฉพาะ iteration = 1 (เริ่มต้น) และ iteration = 8 (ขั้นตอนที่มีการเพิ่มขึ้นของ Confirmed)

ตารางที่ 15 การจัดประเภทของตัวแปร Confirmed, Tentative และ Rejected ในวิธี Boruta

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
iteration = 1, hit_reg	1	1	1	1	1	1	1	1
[1.1] to_accept	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
[1.2] to_reject	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
[1.3] = sorted [1.1]	[0.5000 0.5000 0.5000 0.5000 0.5000 0.5000 0.5000 0.5000], index = [0 1 2 3 4 5 6 7]							
[1.4] = ecdffactor	[0.125 0.25 0.375 0.5 0.625 0.75 0.875 1]							
[1.5] to confirm = [1.3]/[1.4]	[4.0000 2.0000 1.3333 1.0000 0.8000 0.6667 0.5714 0.5000], index = [0 1 2 3 4 5 6 7]							
[1.6] = sorted [1.2]	[1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000], index = [0 1 2 3 4 5 6 7]							

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

[1.7] to reject = [1.6]/[1.4]	[8.000 4.000 2.6667 2.0000 1.6 1.3333 1.1429 1.0000], index = [0 1 2 3 4 5 6 7]							
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 2, hit_reg	1	2	1	1	1	2	1	2
to_accept	0.7500	0.2500	0.7500	0.7500	0.7500	0.2500	0.7500	0.2500
to_reject	0.7500	1.0000	0.7500	0.7500	0.7500	1.0000	0.7500	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 3, hit_reg	1	3	1	1	2	3	2	3
to_accept	0.8750	0.1250	0.8750	0.8750	0.5000	0.1250	0.5000	0.1250
to_reject	0.5000	1.0000	0.5000	0.5000	0.8750	1.0000	0.8750	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 4, hit_reg	1	4	1	1	3	4	2	4
to_accept	0.9375	0.0625	0.9375	0.9375	0.3125	0.0625	0.6875	0.0625
to_reject	0.3125	1.0000	0.3125	0.3125	0.9375	1.0000	0.6875	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 5, hit_reg	1	5	1	2	4	5	2	5
to_accept	0.9688	0.0313	0.9688	0.8125	0.1875	0.0313	0.8125	0.0313
to_reject	0.1875	1.0000	0.1875	0.5000	0.9688	1.0000	0.5000	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 6, hit_reg	1	6	1	3	5	6	3	6
to_accept	0.9844	0.0156	0.9844	0.6563	0.1094	0.0156	0.6563	0.0156
to_reject	0.1094	1.0000	0.1094	0.6563	0.9844	1.0000	0.6563	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 7, hit_reg	1	7	1	3	6	7	3	7

CHULALONGKORN UNIVERSITY

to_accept	0.9922	0.0078	0.9922	0.7734	0.0625	0.0078	0.7734	0.0078
to_reject	0.0625	1.0000	0.0625	0.5000	0.9922	1.0000	0.5000	1.0000
update	tentative	tentative	tentative	tentative	tentative	tentative	tentative	tentative
iteration = 8, hit_reg	1	8	1	4	7	8	3	8
[8.1] to_accept	0.9961	0.0039	0.9961	0.6367	0.0352	0.0039	0.8555	0.0039
[8.2] to_reject	0.0352	1.0000	0.0352	0.6367	0.9961	1.0000	0.3633	1.0000
[8.3] = sorted [8.1]	[0.0039 0.0039 0.0039 0.0352 0.6367 0.8555 0.9961 0.9961], index = [1 5 7 4 3 6 0 2]							
[8.4] = ecdffactor	[0.125 0.25 0.375 0.5 0.625 0.75 0.875 1]							
[8.5] to confirm = [8.3]/[8.4]	[0.0312* 0.0156* 0.0104* 0.0703 1.0188 1.1406 1.1384 0.9961], index = [1* 5* 7* 4 3 6 0 2]							
[8.6] = sorted [8.2]	[0.0352 0.0352 0.3633 0.6367 0.9961 1.0000 1.0000 1.0000], index = [0 2 6 3 4 1 5 7]							
[8.7] to reject = [8.6]/[8.4]	[0.2813 0.1406 0.9688 1.2734 1.5938 1.3333 1.1429 1.0000], index = [0 2 6 3 4 1 5 7]							
update	tentative	confirmed	tentative	tentative	tentative	confirmed	tentative	confirmed
iteration = 9, hit_reg	1	9	1	4	7	9	3	9
to_accept	0.9980	0.0020	0.9980	0.7461	0.0898	0.0020	0.9102	0.0020
to_reject	0.0195	1.0000	0.0195	0.5000	0.9805	1.0000	0.2539	1.0000
update	tentative	confirmed	tentative	tentative	tentative	confirmed	tentative	confirmed

สำหรับ ecdffactor เป็นการหาช่วงความน่าจะเป็น [0,1] ออกเป็น 8 ช่วงค่า (ตามจำนวนตัวแปรต้น) สำหรับค่า to_accept และ to_reject หาได้จากความน่าจะเป็นสะสมของ hit_reg ที่เกิดขึ้น ตัวอย่าง ตัวแปรต้น Pregnancies ใน iteration = 2 พบว่า hit_reg มีค่าเท่ากับ 1 จะได้ค่า to_accept หาได้จากการคำนวณใน Excel จะได้1-BINOM.DIST (hit_reg(=1) -1,iteration = 2,

p = 0.5, cumulative) และค่า to_reject หาได้จากสูตรการคำนวณใน Excel จะได้ BINOM.DIST (hit_reg(=1), iteration = 2, p = 0.5, cumulative)

ใน [8.5] *คือค่าความน่าจะเป็นสะสมที่น้อยกว่า 0.05 จะถูกจัดให้อยู่ใน Confirmed หรือ Rejected ดัง iteration = 8 ค่าของตัวแปรต้นลำดับที่ 1 5 และ 7 น้อยกว่า 0.05 จึงถูกจัดให้อยู่ใน Confirmed แต่สำหรับการทดสอบ Rejected ไม่พบค่าที่น้อยกว่า 0.05 ทำให้ตัวแปรตัวอื่น ๆ ยังคงสถานะ Tentative เช่นเดิม

VI) จากค่าความสำคัญของตัวแปรตั้งต้นและค่ามากสุดของความสำคัญตัวแปรเงาในแต่ละรอบ จาก II) นำมาหาค่ากลาง (Median) ได้ดังตารางที่ 16

ตารางที่ 16 หาค่ากลางจากค่าความสำคัญของตัวแปรตั้งต้นในวิธี Boruta

iteration	**Pregnancies**	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**
1	0.0546	0.1719	0.0564	0.0543	0.0686	0.0961	0.0547	0.0811
2	0.0490	0.1773	0.0486	0.0589	0.0625	0.0849	0.0605	0.0729
3	0.0462	0.1930	0.0490	0.0433	0.0619	0.0967	0.0595	0.0738
4	0.0407	0.1724	0.0556	0.0516	0.0650	0.1020	0.0537	0.0851
5	0.0562	0.1833	0.0499	0.0596	0.0684	0.0907	0.0531	0.0812
6	0.0337	0.1633	0.0513	0.0591	0.0650	0.0999	0.0672	0.0904
7	0.0437	0.1853	0.0462	0.0454	0.0608	0.0973	0.0561	0.0883
8	0.0556	0.1701	0.0533	0.0684	0.0655	0.0937	0.0580	0.0834
9	0.0462	0.1617	0.0463	0.0475	0.0450	0.1107	0.0523	0.0880
10, median =	0.0462	0.1724**	0.0499	0.0543	0.065**	0.0967**	0.0561	0.0834**
From Table 15, update	rejected	confirmed	rejected	rejected	tentative	confirmed	rejected	confirmed

**ใน iteration = 10 แสดงถึงค่ากลางของค่าความสำคัญของตัวแปรต้นดั้งเดิมที่มากกว่าค่ากลางของค่าความสำคัญมากสุดในแต่ละรอบของตัวแปรเงา (>0.0585) ดังตารางที่ 17

ตารางที่ 17 หาค่ากลางของค่าความสำคัญมากสุดในแต่ละรอบของตัวแปรเงา ในวิธี Boruta

iteration	1	2	3	4	5	6	7	8	9	median
Shadow max	0.0535	0.0695	0.0576	0.0585	0.058	0.0533	0.0597	0.0625	0.0616	0.0585

V) จาก IV ตัวแปรที่มีค่ากลางของค่าความสำคัญตัวแปรน้อยกว่า 0.0585 จะถูกจัดอยู่ในตัวแปรที่ไม่ถูกเลือก หรือ not_selected เพื่อใช้สำหรับการจัดลำดับความสำคัญแยกจากตัวแปรที่มีค่ามากกว่า 0.0585 แสดงการจัดลำดับความสำคัญของตัวแปรใน not_selected ได้ดังตารางที่ 18

ตารางที่ 18 การจัดลำดับความสำคัญของตัวแปรเฉพาะตัวแปรไม่ถูกคัดเลือกใน IV ในวิธี Boruta

iteration	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
1	-0.0546		-0.0564	-0.0543			-0.0547	
ranked_1	3		1	4			2	
2	-0.0490		-0.0486	-0.0589			-0.0605	
ranked_2	3		4	2			1	
3	-0.0462		-0.0490	-0.0433			-0.0595	
ranked_3	3		2	4			1	
4	-0.0407		-0.0556	-0.0516			-0.0537	
ranked_4	4		1	3			2	
5	-0.0562		-0.0499	-0.0596			-0.0531	
ranked_5	2		4	1			3	
6	-0.0337		-0.0513	-0.0591			-0.0672	
ranked_6	4		3	2			1	
7	-0.0437		-0.0462	-0.0454			-0.0561	
ranked_7	4		2	3			1	
8	-0.0556		-0.0533	-0.0684			-0.0580	
ranked_8	3		4	1			2	

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

9	-0.0462		-0.0463	-0.0475			-0.0523	
ranked_9	4		3	2			1	
[1] median of rank_1to9	3		3	2			1	
[2]	5 (=3+2)		5 (=3+2)	4 (=2+2)			3 (=1+2)	
[3] Selected feature (from Table 16 iteration10)		1 (confirmed)			2 (tentative)	1 (confirmed)		1 (confirmed)
[4] Importance all feature = [2] + [3]	5	1	5	4	2	1	3	1
update	rejected	confirmed	rejected	rejected	tentative	confirmed	rejected	confirmed

หมายเหตุ [2] เกิดจาก [1] บวกด้วย 3 หากตัวแปรดังกล่าวเดิมจัดอยู่ในประเภท tentative หากตัวแปรดังกล่าวจัดอยู่ใน rejected จะได้ [2] เกิดจาก [1] บวกด้วย 2

พบว่าการจัดลำดับความสำคัญของตัวแปรด้วยวิธี Boruta อาจเกิดการซ้ำกันของลำดับความสำคัญตัวแปรต้น สังเกตได้ว่าตัวแปรที่ถูกจัดอยู่ใน Confirmed จะมีค่าเท่ากับ 1 สำหรับตัวแปรที่ถูกจัดอยู่ใน Tentative จะมีค่าเท่ากับ 2 และตัวแปรอื่น ๆ จะมีลำดับ 3 เป็นต้นไป จากข้อมูล Diabetes ตัวอย่างจะได้ความสำคัญของตัวแปรต้นดังรูปที่ 24

Feature	value_boruta
Glucose	1
BMI	1
Age	1
Insulin	2
DiabetesPedigreeFunction	3
SkinThickness	4
Pregnancies	5
BloodPressure	5

รูปที่ 24 ลำดับความสำคัญของตัวแปรต้นที่มีอิทธิพลต่อการอธิบายการทำนายผ่านตัวแบบด้วยวิธี Boruta

2.1.3.3 Embedded

สำหรับการจำแนกประเภท (Classification) การแบ่งโหนดในต้นไม้ตัดสินใจจะทำการหาจากจุดแบ่งที่เหมาะสม หรือการหาวิธีที่ทำให้อัลกอริทึมให้ค่า Gini impurity หรือ Entropy ที่ต่ำสุด แต่ถ้าหากเป็นการวิเคราะห์การถดถอย (Regression) คือการหาวิธีที่ทำให้อัลกอริทึมให้ค่าเฉลี่ยข้อผิดพลาดยกกำลังสอง (Mean Squared Error; MSE) หรือ ค่าสัมบูรณ์ของค่าเฉลี่ยข้อผิดพลาด (Mean Absolute Error; MAE) มีค่าต่ำสุด

ตัวแบบ XGBoost ประกอบด้วยต้นไม้การตัดสินใจ (Classification trees) ทำให้ค่าความสำคัญของตัวแปรต้นเป็นค่าเฉลี่ยจากทุกต้นไม้การตัดสินใจดังสมการที่ (2.1R)

ค่าความสำคัญของตัวแปรต้นที่สนใจ ($Importance_A$) หาจากค่าความสำคัญแต่ละรูปแบบการแบ่งโหนดในต้นไม้การตัดสินใจ ดังรูปที่ 9 สามารถหาได้ดังนี้

$$Importance_{rA} = [\left(\%sample_{node_{r_A}} \times Gini\ Impurity_{node_{r_A}}\right) - \left(\%sample_{LeftLeaf} \times Gini\ Impurity_{LeftLeaf}\right) - (\%sample_{RightLeaf} \times Gini\ Impurity_{RightLeaf})]/100 \quad (2.1P)$$

$$Importance_A = \frac{\sum Importaance_{rA}}{\sum Importaance_r} \quad (2.1Q)$$

$$\widehat{Importance}_A = \frac{1}{T}\sum Importance_A \quad (2.1R)$$

ให้ A แทน ตัวแปรต้นที่ทำการหาค่าความสำคัญ f_j, i=1,...,P

$Importance_{rA}$ แทน ค่าความสำคัญของโหนดที่เป็นตัวแทนตัวแปรต้น A

$node_{rA}$ แทน โหนดของตัวแปรต้น A

T แทน จำนวนของต้นไม้ตัดสินใจที่อยู่ในตัวแบบ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตัวอย่าง การหาค่าความสำคัญตัวแปรต้นเพศหญิง อ้างอิงจากข้อ 2.1.2.1 แบ่งโหนดต้นไม้การตัดสินใจได้ดังรูปที่ 25 และอ้างอิงสมการที่ (2.1P) จะได้ค่าความสำคัญของโหนดการแบ่งด้วยเพศหญิงดังนี้

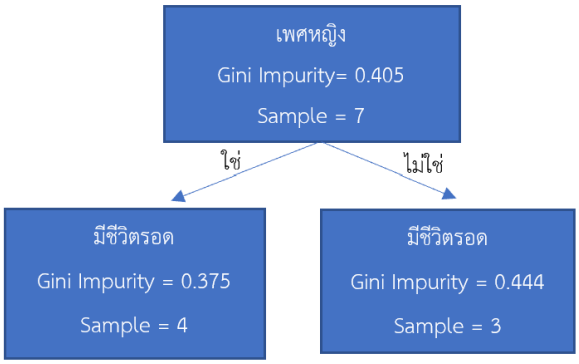
$$Importance_{rA} = [\left(\%sample_{node_{r_A}} \times Gini\ Impurity_{node_{r_A}}\right) - \left(\%sample_{LeftLeaf} \times Gini\ Impurity_{LeftLeaf}\right) - (\%sample_{RightLeaf} \times Gini\ Impurity_{RightLeaf})]/100$$

$$= [\left(\left(100 \times \frac{7}{7}\right) \times 0.405\right) - \left(\left(100 \times \frac{4}{7}\right) \times 0.375\right) - \left(\left(100 \times \frac{3}{7}\right) \times 0.444\right]/100$$

$$= (40.5 - 21.429 - 19.029)/100$$

$$= 0.00042$$

ค่าความสำคัญของโหนดเพศหญิงเท่ากับ 0.00042

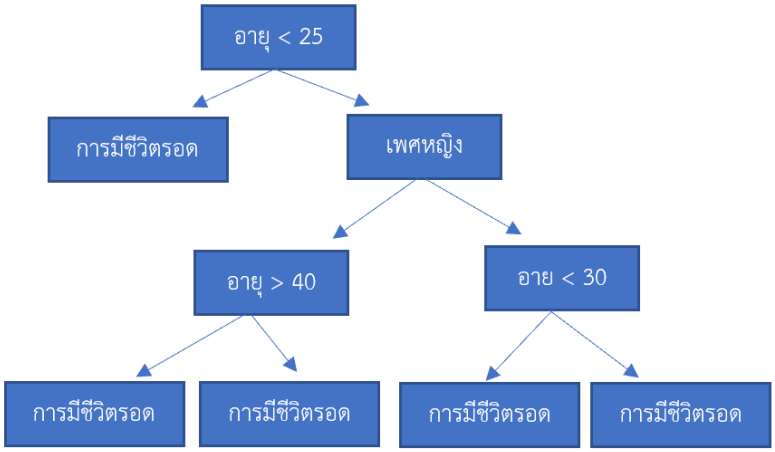


รูปที่ 25 การแบ่งโหนดตัวแปรเพศหญิง ($node_r$) ในต้นไม้การตัดสินใจ

สำหรับต้นไม้ตัดสินใจที่มีความซับซ้อนมากขึ้น (มีจำนวนต้นไม้ตัดสินใจมากขึ้น) ดังรูปที่ 26 การหาความสำคัญของตัวแปร ทำได้โดยอ้างอิงจากสมการที่ (2.1Q) เช่น การหาค่าความสำคัญของตัวแปรอายุได้ดังนี้

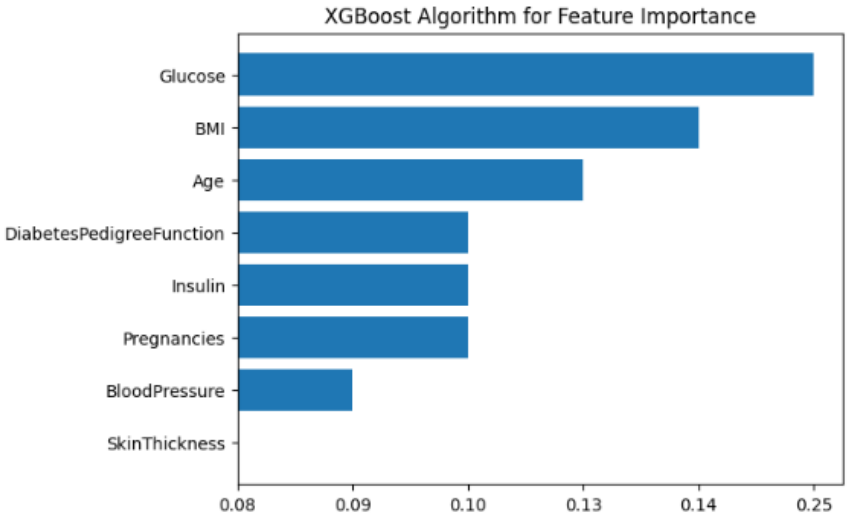
$$Importance_A = \frac{\sum Importaance_{r_A}}{\sum Importaance_r}$$

$$= \frac{Importaance_{\text{อายุ}<25} + Importaance_{\text{อายุ}<30} + Importaance_{\text{อายุ}>40}}{Importaance_{\text{อายุ}<25} + Importaance_{\text{อายุ}<30} + Importaance_{\text{อายุ}>40} + Importaance_{\text{เพศหญิง}}}$$



รูปที่ 26 ตัวอย่างต้นไม้ตัดสินใจที่มีความซับซ้อนมากขึ้น

การจัดลำดับความสำคัญของตัวแปรต้นด้วยชุดข้อมูล Diabetes ดังตัวอย่างการจัดลำดับความสำคัญของตัวแปรต้นผ่านประเภท Filter ด้วยวิธี Mutual Information (ในหัวข้อที่ 2.1.3.1) และประเภท Wrapper ด้วยวิธี SHAP (ในหัวข้อที่ 2.1.3.2) เมื่อนำมาใช้ในการจัดลำดับความสำคัญของตัวแปรผ่านอัลกอริทึม XGBoost ในประเภท Embedded จะได้ลำดับความสำคัญของตัวแปรดังรูปที่ 27



รูปที่ 27 ลำดับความสำคัญของตัวแปรต้นที่มีความสำคัญต่อการอธิบายการทำนายผ่านตัวแบบ XGBoost เรียงลำดับจากความสำคัญมากที่สุดไปน้อยที่สุด

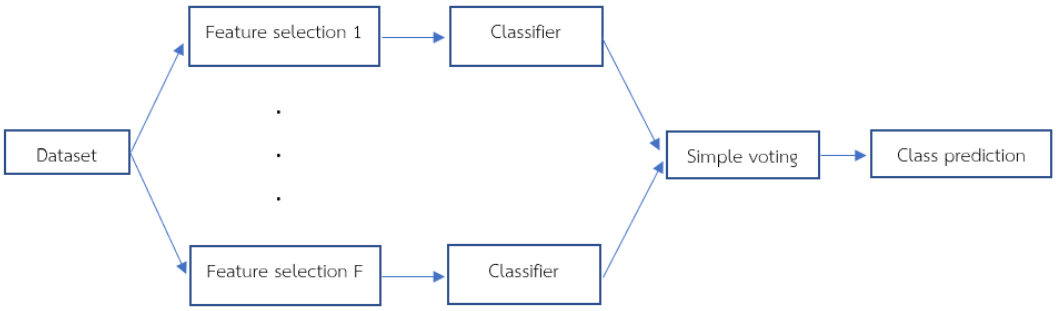
2.2 งานวิจัยที่เกี่ยวข้อง

ชุดข้อมูลที่ใช้ในงานศึกษานี้มี 3 ชุด ได้แก่ Parkinson's Disease, LSVT Voice Rehabilitation และ Colon Cancer ซึ่งมีเงื่อนไขขนาดของข้อมูลและจำนวนตัวแปรต้นที่แตกต่างกัน ที่ผ่านมามีงานศึกษาเกี่ยวกับการคัดเลือกตัวแปรโดยเลือกใช้ข้อมูลดังกล่าว รายละเอียดดังนี้

สำหรับข้อมูล Parkinson's Disease งานศึกษาของ Polat (2019) ใช้ข้อมูลเสียงพูดของผู้ป่วยพาคินสันเพื่อใช้ในการวินิจฉัยอาการ โดยใช้ตัวแบบ Random Forest ได้ Accuracy เท่ากับ 87.037% และเมื่อทำ SMOTE เพื่อลดปัญหาข้อมูลไม่สมมาตร ทำให้ได้ค่า Accuracy ขึ้นมาสูงถึง 94.89% จากปัญหาเรื่องมิติของข้อมูลในชุดข้อมูล Parkinson's Disease งานศึกษาของ Gunduz (2021) ใช้ตัวแบบ SVM ในการทำนาย หากไม่ทำการคัดเลือกตัวแปรจะได้ Accuracy เท่ากับ 84.5% และ F1-score เท่ากับ 0.902 แต่เมื่อทำการคัดเลือกตัวแปรจำนวน 60 ตัว ด้วยวิธี Relief เข้าตัวแบบได้ Accuracy เท่ากับ 87.3% และ F1-score เท่ากับ 0.917 อีกทั้งยังมีงานศึกษาของ Liu et al. (2022) ใช้ข้อมูล Parkinson's Disease ศึกษาเกี่ยวกับการคัดเลือกตัวแปรเข้าตัวแบบ พบว่าวิธี SHAP ให้ผลการทำนายดีกว่าวิธีการคัดเลือกตัวแปรอื่นที่นำมาเปรียบเทียบได้แก่ F-score, Analysis of variance (Anova-F) และ Mutual information สำหรับตัวแบบทำนายที่นำมาเปรียบเทียบมีทั้งหมด 4 ตัวแบบได้แก่ Deep forest (gcForest), Extreme gradient boosting (XGBoost), light gradient boosting machine (lightGBM) และ Random forest (RF) ผลพบว่า เมื่อใช้ gcForest ร่วมกับ SHAP ในการคัดเลือกตัวแปรเข้าตัวแบบ เมื่อใช้ตัวแปรต้นจำนวน 150 ตัว จะได้ Accuracy เท่ากับ 91.78% และ F1-score เท่ากับ 0.945 เมื่อใช้ตัวแบบ lightGBM ร่วมกับ SHAP ในการคัดเลือกแปรเข้าตัวแบบ เมื่อใช้ตัวแปรต้นจำนวน 50 ตัว จะได้ Accuracy เท่ากับ 91.62% และ F1-score เท่ากับ 0.945

สำหรับข้อมูล LSVT งานศึกษาของ Araújo et al. (2016) เปรียบเทียบวิธีการคัดเลือกตัวแปรทั้งแบบวิธีเดียวและแบบรวมกลุ่ม โดยใช้วิธีการคัดเลือกตัวแปรใน Mutual Information based ได้แก่ Maximum relevance (maxRel), Minimum redundancy maximum relevance (MRMR), Minimum redundancy (minRed), Quadratic programming feature selection (QPFS), Mutual information quotient (MIQ), Maximum relevance minimum total redundancy (MRMTR), Spectral relaxation global Conditional Mutual Information (SPEC CMI), Conditional mutual information minimization (CMIM), Conditional Infomax Feature Extraction (CIFE) สำหรับการคัดเลือกตัวแปรรวมกลุ่ม ใช้วิธี Simple Voting ของผลจาก

ตัวแบบทำนายทั้งหมด ดังรูปที่ 28 ตัวอย่างถ้าหากใช้การคัดเลือกตัวแปรเข้าตัวแบบทั้งหมด 4 วิธี เข้าสู่ตัวแบบ SVM จะได้ค่าความน่าจะเป็นที่ y=1 มาทั้งหมด 4 ค่า นำค่าความน่าจะเป็นดังกล่าวมาทำ Simple Voting โดยหาค่าเฉลี่ยเพื่อสรุปเป็นความน่าจะเป็นที่ y=1 ในครั้งสุดท้าย ในชุดข้อมูล LSVT การคัดเลือกตัวแปรต้นจำนวน $\sqrt{n} \approx$ 11 ตัว (n=126) ด้วยวิธี MRMTR เข้าตัวแบบ k-NN จะได้ Accuracy เท่ากับ 86.5% และเมื่อใช้การคัดเลือกตัวแปรแบบรวมกลุ่ม โดยคัดเลือกตัวแปรต้นให้เหลือจำนวน 3 ตัว จะได้ Accuracy เท่ากับ 84.8% พบว่าวิธีการคัดเลือกตัวแปรทั้ง 2 วิธีข้างต้นสามารถทำให้ Accuracy เพิ่มขึ้นจากการไม่คัดเลือกตัวแปรเข้าตัวแบบ k-NN ที่มี Accuracy เท่ากับ 75.88% แต่เมื่อเปรียบเทียบการคัดเลือกตัวแปรในหลายชุดข้อมูล ไม่พบวิธีการคัดเลือกตัวแปรที่ดีที่สุดสำหรับการใช้ในทุกชุดข้อมูล สำหรับการคัดเลือกตัวแปรแบบรวมกลุ่มทำให้ได้ผลการทำนายที่แม่นยำ (Robust performance) นอกจากนี้งานศึกษา Almayyan (2020) ใช้ชุดข้อมูล LSVT เพื่อคัดเลือกตัวแปร โดยแบ่งเป็น 2 ขั้นตอน ขั้นตอนแรกใช้วิธี Tabu และ ขั้นตอนที่สองใช้วิธี minimum Redundancy-Maximum Relevance (mRMR) จะได้ตัวแปรต้นที่ใช้ในตัวแบบมีจำนวน 8 ตัว จากนั้นทำ Oversampling เพื่อใช้ในตัวแบบ Random Forest ได้ Accuracy เท่ากับ 84.1% ซึ่งเท่ากับ Accuracy ของตัวแบบ Random Forest ที่ไม่ทำการคัดเลือกตัวแปร แต่เมื่อทำการการสุ่มตัวอย่างซ้ำ (Resampling) ในขั้นตอนที่ 2 ของการลดจำนวนตัวแปรต้นเพื่อให้จำนวนประเภท (Class) ของตัวแปรตามมีจำนวนเท่ากับจะทำให้ Accuracy สูงขึ้นมาถึง 95.2%



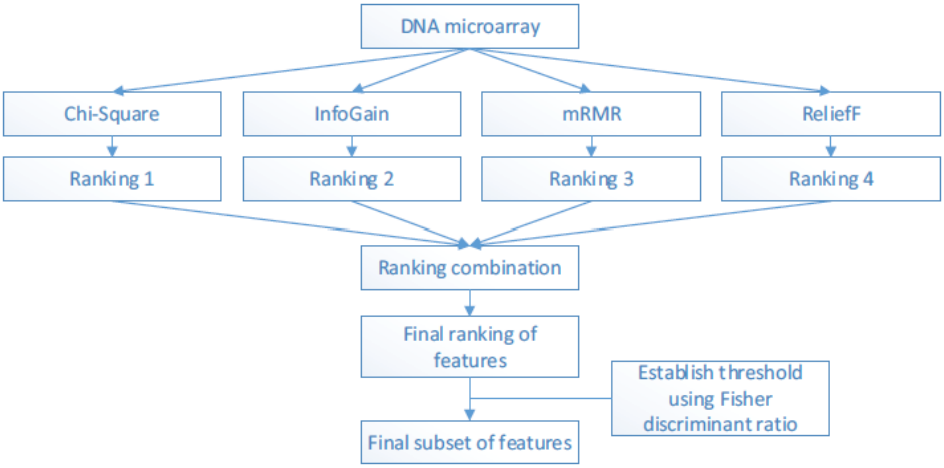
รูปที่ 28 การรวมผลการทำนายแบบ Simple Voting

สำหรับข้อมูล Colon cancer งานศึกษา Ruiz et al.(2006) ได้ศึกษาพบว่าเมื่อใช้ตัวแบบ C4.5 โดยไม่ทำการคัดเลือกตัวแปรได้ Accuracy เท่ากับ 82.14% แต่เมื่อทำการคัดเลือกตัวแปรจำนวนประมาณ 15 ตัว ด้วยวิธี Fast correlation-based filter (FCBF) จะได้ Accuracy เพิ่มขึ้นเป็น 88.33% ต่อมางานศึกษาของ Bolón-Canedo et al.(2012) ใช้การคัดเลือกตัวแปร 5 วิธีจาก

ประเภท Filter ได้แก่ 1) Correlation-based Feature Selection (CFS) 2) Consistency-based 3) INTERACT algorithm 4) Information Gain 5) ReliefF มาทำการคัดเลือกตัวแปรแบบรวมกลุ่มด้วย Simple Voting เช่นเดียวกับงานของ Araújo et al. (2016) ผ่านผลการทำนายของแต่ละตัวแบบหลังทำการคัดเลือกตัวแปรทั้ง 5 วิธี ผลการศึกษาสำหรับชุดข้อมูล Microarray พบว่า เมื่อใช้ตัวแบบ C4.5 ในการคัดเลือกตัวแปรแบบรวมกลุ่มแบบ Simple Voting ได้ค่าความผิดพลาดของการทำนายจากตัวแบบ 13.10 % ซึ่งน้อยกว่าผลการทำนายจากการไม่คัดเลือกตัวแปรที่มีค่าความผิดพลาดของการทำนายจากตัวแบบ 23.81% และดีกว่าการคัดเลือกตัวแปรแบบเดียวเกือบทุกวิธีที่ใช้สำหรับการคัดเลือกตัวแปรแบบรวมกลุ่ม Simple Voting จาก 2 วิธีการคัดเลือกตัวแปรจากประเภท Filter ได้แก่ ReliefF และ Information gain ผ่านตัวแบบ C4.5 จะให้ค่าความผิดพลาดการทำนาย 12.86% ซึ่งน้อยกว่าการคัดเลือกตัวแปรแบบรวมกลุ่มที่ใช้ทั้ง 5 วิธีในประเภท Filter ด้วยตัวแบบ C4.5 ต่อมางานศึกษาของ Seijo-Pardo et al. (2016a) ใช้ 4 วิธีจากประเภท Filter ได้แก่ 1) Chi-Square 2) Information Gain 3) ReliefF 4) mRMR ทำการคัดเลือกตัวแปรแบบรวมกลุ่มโดยปรับเปลี่ยนรูปแบบการรวมกลุ่มแบบ Simple Voting ของแต่ละตัวแบบทำนาย มาเป็นการรวมลำดับความสำคัญของตัวแปรที่ได้จากการคัดเลือกตัวแปร 4 วิธี ดังรูปที่ 29 โดยงานศึกษาดังกล่าวแบ่งการรวมลำดับความสำคัญเป็นหลายวิธีได้แก่ Min, Median, Mean, GeomMean, Stuart และ RRA ผลพบว่ามีวิธีการคัดเลือกตัวแปรแบบรวมกลุ่มอย่างน้อย 1 วิธี ที่สามารถลดความผิดพลาดจากการทำนายได้จาก Baseline ในชุดข้อมูล Colon, DLBCL, Lung และ Ovarian สำหรับชุดข้อมูล Colon การรวมลำดับความสำคัญของตัวแปรต้นด้วยวิธี Min และ Median เพื่อคัดเลือกตัวแปรต้นจำนวน 10 และ 5 ตัวตามลำดับ เข้าตัวแบบ SVM จะได้ความผิดพลาดในการทำนาย 15.0% ต่อมางานศึกษาของ Bolón-Canedo & Alonso-Betanzos (2019) ใช้ 4 วิธีการคัดเลือกจากประเภท Filter ได้แก่ 1) Chi-Square 2) Information Gain 3) ReliefF 4) mRMR และใช้ 2 วิธีจาก Embedded ได้แก่ 1) SVM-RFE 2) Feature Selection-Perceptron (FS-P) มาทำการคัดเลือกตัวแปรแบบรวมกลุ่มโดยการรวมลำดับความสำคัญของตัวแปร แบ่งได้เป็น 2 รูปแบบได้แก่ Design CT และ Design TC อีกทั้งยังศึกษาเกี่ยวกับการกำหนดเกณฑ์แบบอัตโนมัติ (Automatic threshold) ผลพบกว่า Design TC วิธีมัลติอินเตอร์เซกอย่างน้อย 2 เซตตัวแปรต้นที่ได้จากการคัดเลือกตัวแปรร่วมกับการใช้เกณฑ์แบบอัตโนมัติ จะให้ค่าความผิดพลาดของการทำนาย (≈15%) และจำนวนตัวแปรต้น(≈15) ที่คัดเลือกเข้าสู่ตัวแปรน้อยกว่าวิธีอื่น ๆ โดยเปรียบเทียบ สำหรับเกณฑ์อัตโนมัติที่

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

กล่าวถึงจะมีจำนวนตัวแปรต้นเริ่มต้นเท่ากับ $\log_2(nF)$ ให้ nF แทนด้วยจำนวนตัวแปรต้นในชุดข้อมูล (Seijo-Pardo et al., 2016b)



รูปที่ 29 การรวมลำดับความสำคัญของตัวแปรต้น

ที่มา: (Seijo-Pardo et al., 2016)

ในงานศึกษาของ Effrosynidis & Arampatzis (2021) ใช้การคัดเลือกตัวแปรจาก 3 ประเภท 1. Filter ได้แก่ 1.1) Chi-square 1.2) Mutual Information 1.3) Anova 1.4) F-value 1.5) Variance Threshold Fisher-Score 1.6) MultiSURF 2. Wrapper ได้แก่ 2.1) Recursive Feature Elimination 2.2) Permutation Importance 2.3) SHAP 2.4) Boruta และ 3. Embedded ได้แก่ 3.1) Random Forest 3.2) LightGBM โดยงานศึกษาได้นำทั้ง 12 วิธีดังกล่าวมาเปรียบเทียบกับวิธีทำการคัดเลือกตัวแปรแบบรวมกลุ่ม ซึ่งแบ่งเป็น 6 วิธีการคัดเลือกตัวแปรแบบรวมกลุ่มได้แก่ 1. Borda Count 2. Reciprocal Ranking 3. Condorcet 4. Coombs 5. Instant Runoff 6. Bucklin เกณฑ์ที่ใช้ได้แก่ 5% 10% 20% และ 30% ผลการศึกษาพบว่า ในการคัดเลือกตัวแปรแบบวิธีเดียว ด้วย SHAP ให้ผลดีที่สุดเมื่อเปรียบเทียบกับวิธีเดียวอื่น และสำหรับวิธีทำการคัดเลือกตัวแปรแบบรวมกลุ่ม (Ensemble) ด้วยการเลือกใช้วิธี Reciprocal Ranking จะให้ผลดีที่สุดเมื่อเทียบกับวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม (Ensemble) แบบอื่น

วิธีการคัดเลือกตัวแปรเข้าตัวแบบนั้นมี 3 ประเภท ที่ผ่านมามีงานศึกษาเกี่ยวกับการคัดเลือกตัวแปรเข้าตัวแบบด้วยวิธีที่ต่างกัน ทั้งการเลือกใช้การคัดเลือกตัวแปรแบบวิธีเดียวและการคัดเลือกตัว

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

แปรแบบรวมกลุ่ม อีกทั้งยังเลือกใช้ตัวแบบทำนายที่ต่างกัน ผลการศึกษาที่ผ่านมาไม่มีวิธีการคัดเลือกใดหรือตัวแบบใดที่ให้ผลการทำนายที่ดีที่สุดในทุกชุดข้อมูลไมโครอาเรย์ โดยงานศึกษาของ Bolón-Canedo et al.(2012) ได้นำ 5 วิธีการคัดเลือกตัวแปรจากประเภท Filter ได้แก่ Correlation-based Feature Selection (CFS), Consistency-based, INTERACT algorithm, Information Gain, ReliefF มาทำการคัดเลือกตัวแปรแบบรวมกลุ่มด้วยวิธี Simple Voting เข้าตัวแบบ C4.5 ต่อมางานศึกษา Seijo-Pardo et al. (2016) ใช้ 4 วิธีการคัดเลือกตัวแปรจากประเภท Filter ได้แก่ Chi-Square, Information Gain, ReliefF, mRMR และเปลี่ยนรูปแบบการคัดเลือกตัวแปรแบบรวมกลุ่มจาก Simple Voting เป็นการรวมลำดับความสำคัญของตัวแปรเป็น Min, Median, Mean, Geometric Mean, Stuart และ RRA เข้าตัวแบบ SVM ต่อมางานศึกษา Bolón-Canedo & Alonso-Betanzos (2019) เลือกใช้ 2 วิธีการคัดเลือกตัวแปรในประเภท Embedded ได้แก่ SVM-RFE และ Feature Selection-Perceptron (FS-P) เพิ่มเติมมาจาก 4 วิธีประเภท Filter ที่ปรากฏในงานปี 2016 ข้างต้น โดยทำการแบ่งวิธีการคัดเลือกตัวแปรแบบรวมกลุ่มออกเป็น 2 รูปแบบได้แก่ Design CT ที่ใช้วิธีค่าต่ำสุดในการรวมลำดับ และ Design TC ที่รวมตัวแปรในรูปแบบของเซต อีกทั้งยังศึกษาเกี่ยวกับเกณฑ์อัตโนมัติ (Automatic threshold) สำหรับกำหนดจำนวนตัวแปรต้นเข้าตัวแบบ SVM โดยพิจารณาเปรียบเทียบผลของการคัดเลือกตัวแปร เฉพาะการคัดเลือกตัวแปรแบบรวมกลุ่ม สำหรับชุดข้อมูล Environment ในรูปแบบของอนุกรมเวลาที่ปรากฏในงานศึกษาของ Effrosynidis & Arampatzis (2021) ศึกษาเกี่ยวกับการคัดเลือกตัวแปร โดยเปรียบเทียบทั้งการคัดเลือกตัวแปรแบบวิธีเดียวและแบบวิธีรวมกลุ่ม พบว่าการคัดเลือกตัวแปรแบบวิธีเดียวด้วย SHAP ซึ่งอยู่ในประเภท Wrapper ให้ผลลัพธ์ที่ดีเช่นกันเมื่อเทียบกับการคัดเลือกแบบรวมกลุ่มด้วยวิธี Reciprocal จากการทบทวนงานวิจัยที่ผ่านมาพบว่า ยังมิได้ศึกษาการคัดเลือกตัวแปรแบบรวมกลุ่มในรูปแบบ Design CT และ Design TC ที่พิจารณาการคัดเลือกตัวทั้ง 3 ประเภทได้แก่ Filter Wrapper และ Embedded อีกทั้งพิจารณาการรวมลำดับความสำคัญของตัวแปรต้นใน Design CT ที่นอกเหนือจากวิธีค่าต่ำสุด รวมถึงการเปรียบเทียบการคัดเลือกตัวแปรแบบวิธีเดียวกับการคัดเลือกตัวแปรแบบรวมกลุ่มทั้ง Design CT และ Design TC

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 3
วิธีการดำเนินการวิจัย

3.1 ชุดข้อมูลทดสอบ

3.1.1 ชุดข้อมูลทดสอบที่ใช้

ชุดข้อมูล Parkinson's Disease จาก UCI Machine Learning Repository (Sakar et al., 2019) จุดประสงค์เพื่อวินิจฉัยโรคพาร์กินสัน (Parkinson's Disease: 1= patients with PD, 0 = healthy patients) โดยมีกลุ่มตัวอย่าง 756 ตัวอย่าง และ 753 ตัวแปรต้น

ชุดข้อมูล LSVT Voice Rehabilitation จาก UCI Machine Learning Repository (Tsanas et al., 2014) จุดประสงค์เพื่อประเมินการรักษาการพูดเสียงเบาหรือ LSVT (Lee Silverman voice treatment: 1= unacceptable, 0= acceptable) นำไปสู่ผลที่ยอมรับได้หรือไม่ โดยมีกลุ่มตัวอย่าง 126 ตัวอย่าง และ 312 ตัวแปรต้น

ชุดข้อมูล Colon Cancer (Shafi et al., 2020) จุดประสงค์เพื่อวินิจฉัยมะเร็งลำไส้ (Colon Cancer: 1= abnormal (tumor biopsies), 0= normal) โดยมีกลุ่มตัวอย่าง 62 ตัวอย่าง และ 2,000 ตัวแปรต้น

3.1.2 ขั้นตอนการเตรียมข้อมูล

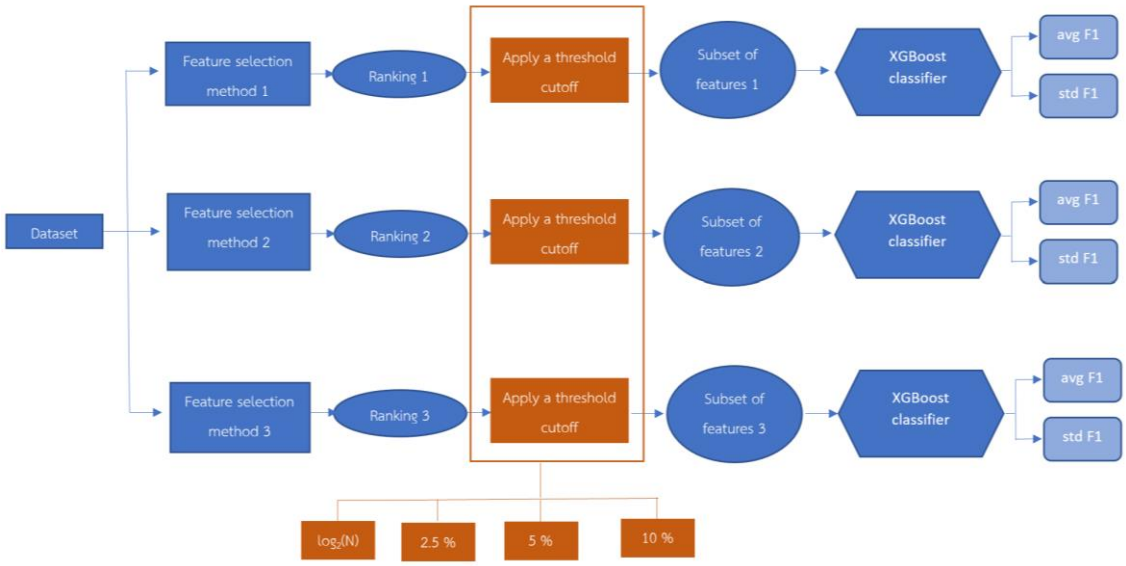
ทำการปรับข้อมูลเชิงปริมาณ (Quantitative data) ให้อยู่ในช่วงค่า 0 ถึง 1 ด้วยการทำ Min-Max Scaler และปรับข้อมูลเชิงคุณภาพ (Qualitative data) ให้อยู่ในรูปไบนารี (Binary) ด้วยการทำ One-hot encoding

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.2 การเลือกวิธีการคัดเลือกตัวแปรจากแต่ละประเภท

ในการคัดเลือกตัวแปรสามารถแบ่งได้เป็น 3 ประเภทได้แก่ Filter Wrapper และ Embedded สำหรับวิธีการคัดเลือกตัวแปรประเภท Filter จะใช้การทดสอบเปรียบเทียบระหว่างวิธี Mutual Information, Variance Threshold และ MultiSURF สำหรับประเภท Wrapper จะใช้การทดสอบเปรียบเทียบระหว่างวิธี SHAP, RFE และ Boruta เลือกวิธีการคัดเลือกตัวแปรที่ดีที่สุดจากแต่ละประเภท โดยการทดสอบเริ่มจากการจัดลำดับความสำคัญของตัวแปรต้นในชุดข้อมูลที่ได้จากแต่ละวิธีการคัดเลือกตัวแปรที่กล่าวมาข้างต้น จากนั้นจึงเลือกตัวแปรต้นที่มีความสำคัญสูงสุด k ลำดับแรกตามเกณฑ์ (Threshold) ที่กำหนดได้แก่ $\log_2(P)$ 2.5% 5% และ 10% แล้วทำการเปรียบเทียบผลของแต่ละวิธีด้วยการวัดผลการทำนายของตัวแบบผ่านการทำ 10-fold cross

validation เพื่อหาค่าเฉลี่ยของ F1-score และค่าเบี่ยงเบนของ F1-score ดังรูปที่ 30 แสดงขั้นตอนการเปรียบเทียบการคัดเลือกตัวแปรต้น 3 วิธีในประเภท Filter และ Wrapper สำหรับประเภท Embedded นั้นจะใช้การทำงานของอัลกอริทึม XGBoost ในการจัดลำดับความสำคัญของตัวแปร



รูปที่ 30 ขั้นตอนการเปรียบเทียบวิธีการคัดเลือกตัวแปรทั้ง 3 วิธี ในประเภท Filter และ Wrapper

3.3 การเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม

การคัดเลือกตัวแปรแบบวิธีเดียวจากข้อ 3.2 เมื่อนำ 3 วิธีการคัดเลือกตัวแปรดังกล่าวเข้าสู่กระบวนการเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่ม โดยแบ่งเป็น 2 รูปแบบคือ รูปแบบแรกคือ Design CT คือ การรวมลำดับความสำคัญของตัวแปรด้วยวิธีค่าต่ำสุด (Min) ค่ากลาง (Median) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) หรือค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) ดังตารางที่ 19 ก่อนแล้วตามด้วยการเลือกจำนวนตัวแปรมีความสำคัญตามเกณฑ์ที่ระบุ ดังรูปที่ 4 รูปแบบที่สองคือ Design TC เป็น การเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุก่อนแล้วตามด้วยการรวมเซตของตัวแปรที่มีความสำคัญดังกล่าวด้วยวิธียูเนียน มัลติอินเตอร์เซก และ อินเตอร์เซก จากตารางที่ 19 สมมติให้ใช้เกณฑ์ 5 ลำดับแรกจะได้ set1 และ set2 คือตัวแปรต้น 5 ลำดับแรกที่มีความสำคัญจากการคัดเลือกตัวแปรด้วยวิธีที่ 1 และวิธีที่ 2 ตามลำดับ ดังรูปที่ 31 จะได้ว่าการรวมด้วยวิธียูเนียนคือการเลือกตัวแปรที่ปรากฏในพื้นที่ใดพื้นที่หนึ่งใน 1 2 หรือ 3 วิธีอินเตอร์เซกคือการเลือกตัวแปรที่ปรากฏในพื้นที่ 2 เท่านั้น กรณีที่ใช้วิธีการคัดเลือกตัวแปรมากกว่า 2 วิธี ดังรูปที่ 32 การรวมเซตตัวแปรที่มีความสำคัญด้วยวิธีมัลติอินเตอร์เซกคือการคัดเลือกตัวแปรที่ปรากฏในพื้นที่ 4 5 6 และ 7

ขั้นตอนของ Design TC ดังรูปที่ 5 สำหรับงานศึกษานี้ใช้ 4 เกณฑ์ในการกำหนดลำดับการเลือกตัวแปรต้นได้แก่ $\log_2(P)$ 2.5% 5% และ 10% นอกเหนือจากการเปรียบเทียบวิธีการคัดเลือกตัวแปรแบบรวมกลุ่มแล้วยังมีการแสดงผลของการคัดเลือกตัวแปรแบบวิธีเดียว (Single feature selection) มีขั้นตอนดังรูปที่ 3

3.3.1 Design CT

จากรูปที่ 1 เริ่มจากการรวมลำดับความสำคัญของตัวแปรต้นก่อน แล้วจึงกำหนดเกณฑ์คัดเลือกตัวแปรเข้าตัวแบบ เช่น เลือกตัวแปรต้นที่ลำดับความสำคัญสูงสุดใน 5 ลำดับแรก โดยการรวมลำดับความสำคัญของตัวแปรต้นแบ่งเป็น 4 รูปแบบได้แก่ ค่าต่ำสุด (Min) ค่ากลาง (Median) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) และ ค่าเฉลี่ยฮาร์มอนิก (Harmonic mean)

ตัวอย่าง กำหนดให้เลือกใช้การคัดเลือกตัวแปรเข้าตัวแบบ 2 วิธี การคัดเลือกตัวแปรเข้าตัวแบบด้วยวิธีที่ 1 จะให้ค่าลำดับความสำคัญของตัวแปรต้น f_1 f_2 f_3 f_4 และ f_5 ตามลำดับ สำหรับวิธีที่ 2 จะให้ค่าลำดับความสำคัญของตัวแปรต้น f_4 f_1 f_2 f_5 และ f_3 ตามลำดับ ผลของการจัดลำดับใหม่จาก 2 วิธีดังกล่าวด้วยการรวมลำดับในรูปแบบต่าง ๆ ให้ผลดังตารางที่ 19

ตารางที่ 19 ตัวอย่างผลของการร่วมกันคำนวณลำดับความสำคัญของตัวแปรต้นในรูปแบบ Design CT

Feature	Feature Ranking for method 1	Feature Ranking for method 2	Min	Median	Arithmetic mean	Harmonic mean
f_1	1	2	1	1.50	(1+2)/2 = 1.50	2 / [(1/1] + (1/2)] = 1.33
f_2	2	3	2	2.50	(2+3)/2 = 2.50	2 / [(1/2] + (1/3)] = 2.40
f_3	3	5	3	4.00	(3+5)/2 = 4.00	2 / [(1/3] + (1/5)] = 3.75
f_4	4	1	1	1.00	(4+1)/2 = 2.50	2 / [(1/4] + (1/1)] = 1.60
f_5	5	4	4	4.00	(5+4)/2 = 4.50	2 / [(1/5] + (1/4)] = 4.44

ตัวอย่างวิธีการรวมลำดับความสำคัญของตัวแปรต้นแสดงตัวอย่างของ X_1 ดังนี้

ค่าต่ำสุด (Min) ของค่า f_1 เท่ากับ 1 เนื่องจากการคัดเลือกตัวแปรวิธีที่ 1 ให้ลำดับความสำคัญของ f_1 ที่ 1 ในขณะที่การคัดเลือกตัวแปรวิธีที่ 2 ให้ลำดับที่ 2 จึงเลือกใช้ผลลำดับความสำคัญที่ดีกว่า

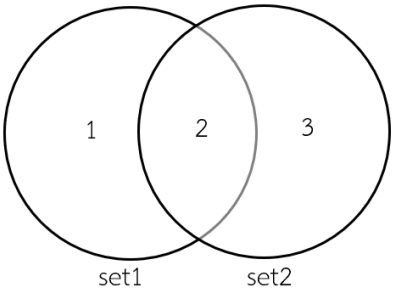
ค่ากลาง (Median) ของค่า f_1 เท่ากับ 1.5 จากตัวอย่างที่ให้มาพบว่า f_1 ถูกจัดลำดับความสำคัญด้วยการคัดเลือกตัวแปรวิธีที่ 1 อยู่ลำดับ 1 และวิธีที่ 2 อยู่ลำดับ 2 ค่ากลางของ 1 และ 2 มีค่าเท่ากับ 1.5

ค่าเฉลี่ยเลขคณิต (Arithmetic mean) ของค่า f_1 เท่ากับ 1.5 จากตัวอย่างที่ให้มาพบว่า f_1 ถูกจัดลำดับความสำคัญด้วยการคัดเลือกตัวแปรวิธีที่ 1 อยู่ลำดับ 1 และวิธีที่ 2 อยู่ลำดับ 2 ค่าเฉลี่ยเลขคณิต ของ 1 และ 2 มีค่าเท่ากับ 1.5

ค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) ของค่า f_1 เท่ากับ 1.33 จากตัวอย่างที่ให้มาพบว่า f_1 ถูกจัดลำดับความสำคัญด้วยการคัดเลือกตัวแปรวิธีที่ 1 อยู่ลำดับ 1 และวิธีที่ 2 อยู่ลำดับ 2 ค่าเฉลี่ยฮาร์มอนิกของ 1 และ 2 มีค่าเท่ากับ 1.33

3.3.2 Design TC

จากรูปที่ 2 เริ่มจากการกำหนดเกณฑ์คัดเลือกตัวแปรก่อน เช่น ในแต่ละวิธีการคัดเลือกตัวแปรเข้าตัวแบบจะเลือกตัวแปรที่มีความสำคัญสูงที่สุดเป็น 5 ลำดับแรกของแต่ละวิธี ได้เซตที่ 1 และ 2 ซึ่งภายในเซตคือตัวแปรต้นที่มีลำดับความสำคัญสูงสุด 5 ลำดับแรก แล้วจึงรวมเซตของตัวแปรที่คัดเลือกมาแล้วตามเกณฑ์ เช่น ในรูปที่ 31 จะได้ว่า วิธียูเนียนจะเลือกตัวแปรต้นที่ปรากฏอยู่ในเซตใดๆ กล่าวคือเลือกตัวแปรต้นที่ปรากฏอยู่ในพื้นที่ 1 2 หรือ 3 วิธีอินเตอร์เซกจะเลือกตัวแปรต้นที่ปรากฏอยู่ทุกเซตพร้อมกัน กล่าวคือเลือกตัวแปรต้นที่ปรากฏอยู่ในพื้นที่ 2

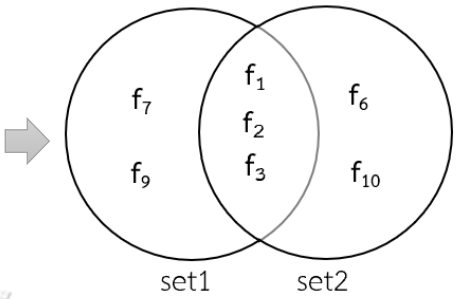


รูปที่ 31 ตัวอย่างการรวมเซตของตัวแปรที่มีความสำคัญ 5 ลำดับแรกใน Design TC เมื่อใช้การคัดเลือกตัวแปรด้วยกัน 2 วิธี

ตัวอย่าง การคัดเลือกตัวแปรต้น 2 วิธี ได้ผล 5 ลำดับแรกของความสำคัญตัวแปรต้นดังตารางที่ 20

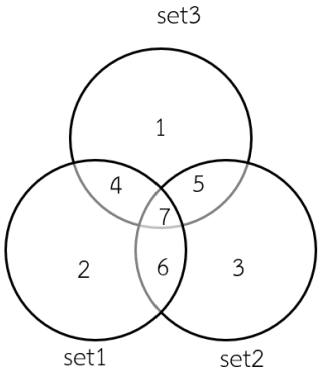
ตารางที่ 20 ผลการรวมเซตของตัวแปรต้นที่มีลำดับความสำคัญในรูปแบบ Design TC

Rank	Feature ranking for FS Method 1	Feature ranking for FS Method 2
1	f_3	f_1
2	f_2	f_3
3	f_1	f_2
4	f_7	f_6
5	f_9	f_{10}



จากตัวอย่างได้ว่า ตัวแปรต้นที่จะนำไปใช้ในตัวแบบ หลังจากทำการรวมลำดับความสำคัญด้วยวิธียูเนียนได้แก่ f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 และ f_{10} สำหรับการรวมลำดับความสำคัญด้วยวิธีอินเตอร์เซกมีตัวแปรต้นได้แก่ f_1 f_2 และ f_3

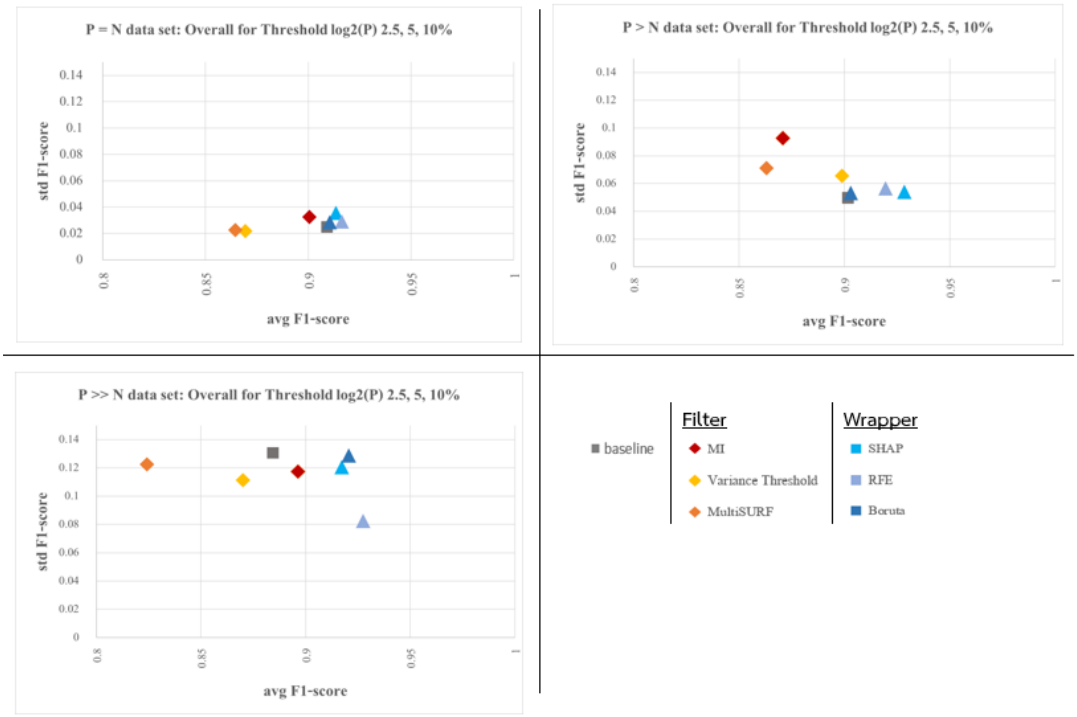
กรณีที่ใช้วิธีการคัดเลือกตัวแปรมากกว่า 2 วิธี การรวมเซตของตัวแปรต้นที่มีความสำคัญอาจใช้วิธีมัลติอินเตอร์เซกร่วมด้วย เช่น ถ้าหากใช้การคัดเลือกตัวแปรเข้าตัวแบบ 3 วิธี จะได้เซตของตัวแปรต้นที่ความสำคัญสูงสุด 5 ลำดับแรกจากแต่ละวิธีการคัดเลือกตัวแปรเป็นจำนวน 3 เซต รูปที่ 32 เมื่อใช้วิธีมัลติอินเตอร์เซกจะได้ตัวแปรต้นที่อยู่ในพื้นที่ 4 5 6 หรือ 7 จะถูกใช้ในตัวแบบทำนาย



รูปที่ 32 ตัวอย่างการรวมเซตของตัวแปรที่มีความสำคัญ 5 ลำดับแรกใน Design TC เมื่อใช้การคัดเลือกตัวแปรด้วยกัน 3 วิธี

บทที่ 4
การทดลองและผลการทดลอง

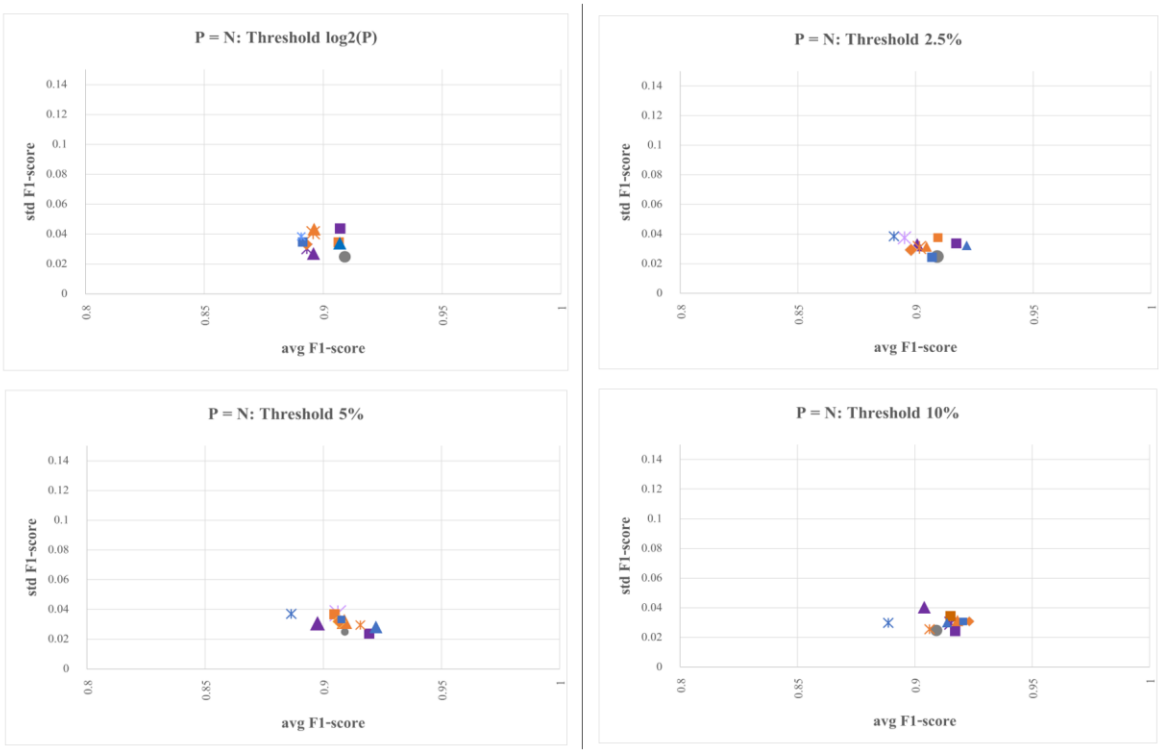
ในส่วนนี้เป็นการวิเคราะห์ผลการศึกษาสำหรับการเปรียบเทียบการคัดเลือกตัวแปรแบบรวมกลุ่ม และการคัดเลือกตัวแปรแบบวิธีเดียวเพื่อพิจารณาผลลัพธ์ที่ได้จากการทำนาย พบว่าโดยรวมจากทุกเกณฑ์ที่เลือกใช้ในแต่ละชุดข้อมูล วิธี Mutual Information จากประเภท Filter และวิธี RFE จากประเภท Wrapper ให้ผลดีกว่าวิธีอื่น ๆ ที่อยู่ในประเภทเดียวกันดังรูปที่ 33 ในการจัดลำดับความสำคัญของตัวแปรในขั้นตอนการคัดเลือกตัวแปรแบบรวมกลุ่ม ผู้ศึกษาจึงเลือกใช้ Mutual Information, RFE และ อัลกอริทึม XGBoost จากประเภท Filter Wrapper และ Embedded ตามลำดับ แสดงผลการศึกษาผ่านแผนภาพการกระจาย (Scatter plot) ให้แกนนอนแทนค่าเฉลี่ยของ F1-score ใช้วัดประสิทธิภาพการทำนายของตัวแบบ และให้แกนตั้งแทนค่าเบี่ยงเบนของ F1-score ใช้วัดความเสถียรของผลการทำนายที่ได้จากตัวแบบ ทำให้ผลลัพธ์ที่คาดหวังอยู่ในบริเวณมุมขวาล่างของแผนภาพ ที่จะแสดงถึงการทำนายที่มีประสิทธิภาพสูง (avg F1-score สูง) และมีความเสถียรมาก (std F1-score ต่ำ) ผลการศึกษาพบว่า



รูปที่ 33 เลือกวิธีการคัดเลือกตัวแปรแบบวิธีเดียวจากประเภท Filter และ ประเภท Wrapper

4.1 ชุดข้อมูล P=N

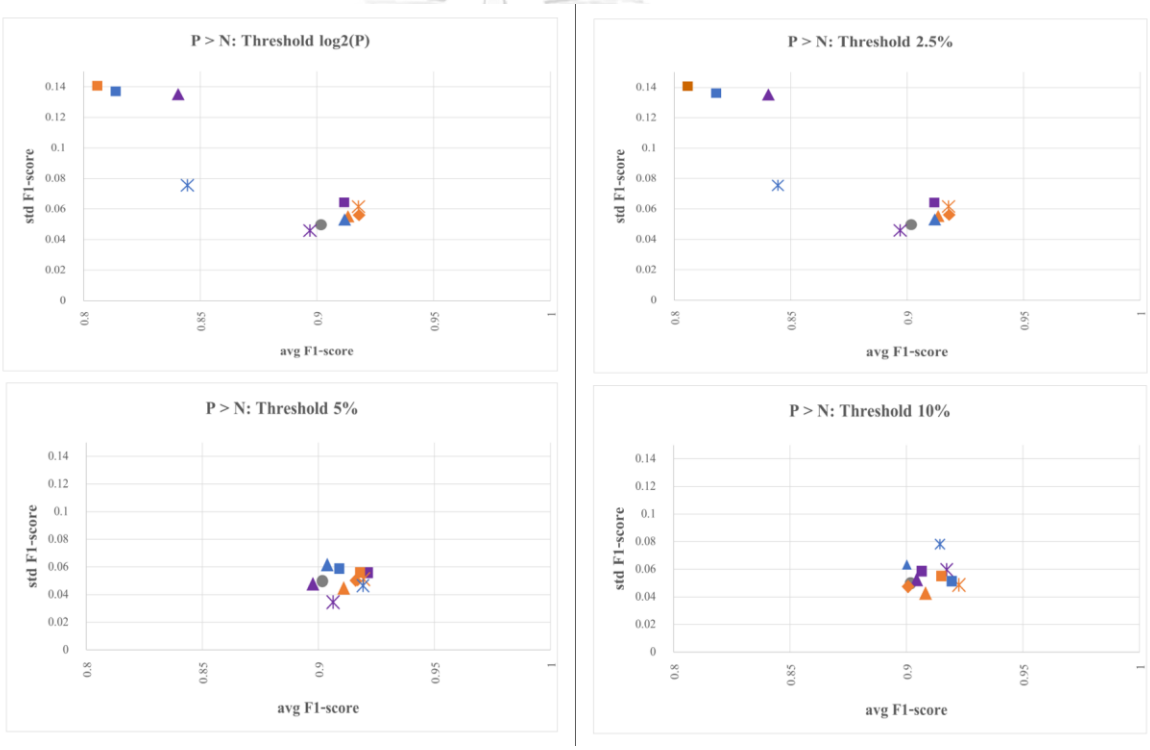
ชุดข้อมูล Parkinson's Disease (P=N) การไม่ทำการคัดเลือกตัวแปรจะให้ ค่าเฉลี่ย F1-score ที่ 0.91 และ ค่าเบี่ยงเบน F1-score ที่ 0.03 การรวมลำดับความสำคัญของตัวแปรใน Design CT ด้วยวิธีค่าเฉลี่ยฮาร์มอนิกที่เกณฑ์ 10% จะทำให้ค่าเฉลี่ย F1-score เพิ่มเป็น 0.92 จากการไม่คัดเลือกตัวแปร โดยเพิ่มมากกว่าการรวมด้วยวิธีค่าต่ำสุด แต่สำหรับค่าเบี่ยงเบน F1-score ไม่ต่างกันมาก สำหรับการรวมเซตตัวแปรที่สำคัญใน Design TC ด้วยวิธียูเนียนที่เกณฑ์ 5% จะทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.92 ดังรูปที่ 34



รูปที่ 34 เปรียบเทียบวิธีการคัดเลือกตัวแปรเข้าตัวแบบในชุดข้อมูล Parkinson's Disease

4.2 ชุดข้อมูล P>N

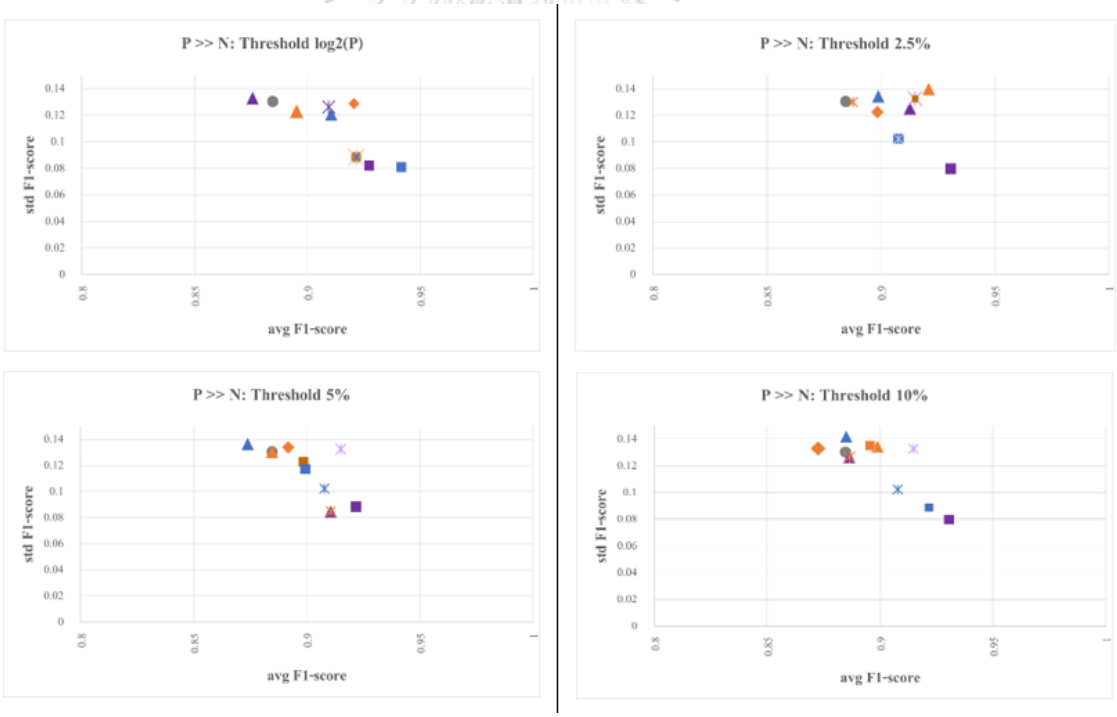
ชุดข้อมูล LSVT Voice Rehabilitation (P>N) พบว่าการไม่ทำการคัดเลือกตัวแปรจะให้ค่าเฉลี่ย F1-score ที่ 0.90 และ ค่าเบี่ยงเบน F1-score ที่ 0.05 การรวมลำดับความสำคัญของตัวแปรใน Design CT ด้วยวิธีค่าเฉลี่ยเลขคณิตที่เกณฑ์ 5% และ 10% และวิธีค่าเฉลี่ยฮาร์มอนิกที่เกณฑ์ 2.5% จะให้ผลดีกว่าการรวมด้วยวิธีอื่นใน Design CT โดยทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.92 สำหรับการรวมเซตตัวแปรที่สำคัญใน Design TC ด้วยวิธีอินเตอร์เซกที่เกณฑ์ 5% และวิธีมัลติอินเตอร์เซกที่เกณฑ์ 10% จะทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.92 เช่นกัน ในขณะที่ค่าเบี่ยงเบน F1-score ไม่ต่างจากเดิม อีกทั้งยังพบว่าการคัดเลือกตัวแปรแบบรวมกลุ่มบางวิธีให้ผลดีกว่าการคัดเลือกแบบวิธีเดียวเล็กน้อย ในทุกเกณฑ์ ดังรูปที่ 35



รูปที่ 35 เปรียบเทียบวิธีการคัดเลือกตัวแปรเข้าตัวแบบในชุดข้อมูล LSVT Voice Rehabilitation

4.3 ชุดข้อมูล P>>N

ชุดข้อมูล Colon Cancer (P>>N) พบว่าการไม่ทำการคัดเลือกตัวแปรจะให้ ค่าเฉลี่ย F1-score ที่ 0.88 และ ค่าเบี่ยงเบน F1-score ที่ 0.13 การรวมลำดับความสำคัญของตัวแปรใน Design CT ด้วยวิธีค่ากลางและค่าเฉลี่ยเลขคณิตที่เกณฑ์ $\log_2(P)$ ทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.92 และลดค่าเบี่ยงเบน F1-score ลงเหลือ 0.09 เมื่อเทียบกับการไม่คัดเลือกตัวแปร สำหรับค่าเฉลี่ยเลขคณิตที่เกณฑ์ 5% ให้ค่าเฉลี่ย F1-score เพิ่มเป็น 0.91 และลดค่าเบี่ยงเบน F1-score ลงเหลือ 0.08 เมื่อเทียบกับการไม่คัดเลือกตัวแปร สำหรับการรวมเซตตัวแปรที่สำคัญในรูปแบบ TC ด้วยวิธีอินเตอร์เซกและมัลติอินเตอร์เซกให้ผลดีกว่าวิธียูเนียนในทุกเกณฑ์โดยวิธีที่ให้ผลดีที่สุดจากรูปแบบ TC คือ วิธีมัลติอินเตอร์เซกที่เกณฑ์ $\log_2(P)$ ที่ทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.94 และทำให้ค่าเบี่ยงเบน F1-score ลงลงเหลือ 0.08 และสำหรับการเลือกตัวแปรแบบวิธีเดียวด้วย RFE จะให้ผลดีในเกณฑ์ที่ 2.5% 5% และ 10% ซึ่ง RFE ให้ผลดีที่สุดที่เกณฑ์ 2.5% และ 10% โดยจะทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นเป็น 0.93 และทำให้ค่าเบี่ยงเบน F1-score ลงลงเหลือ 0.08 ดังรูปที่ 36



รูปที่ 36 เปรียบเทียบวิธีการคัดเลือกตัวแปรเข้าตัวแบบในชุดข้อมูล Colon Cancer

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษา

งานศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดเลือกตัวแปรเข้าตัวแบบสำหรับข้อมูลลักษณะหลายมิติด้วยวิธีเดียวและวิธีแบบรวมกลุ่มที่มี 2 รูปแบบคือ รูปแบบการรวมลำดับความสำคัญของตัวแปรก่อนด้วยวิธีค่าต่ำสุด (Min) ค่ากลาง (Median) ค่าเฉลี่ยเลขคณิต (Arithmetic mean) และค่าเฉลี่ยฮาร์มอนิก (Harmonic mean) แล้วตามด้วยการเลือกจำนวนตัวแปรมีความสำคัญตามเกณฑ์ที่ระบุ (Design Combination followed by Thresholding; Design CT) และรูปแบบการการเลือกจำนวนตัวแปรที่มีความสำคัญตามเกณฑ์ที่ระบุก่อนแล้วตามด้วยการรวมเซตของตัวแปรที่มีความสำคัญดังกล่าวด้วยวิธียูเนียน มัลติอินเตอร์เซก และ อินเตอร์เซก (Design Thresholding followed by Combination; Design TC) ผ่านตัวแบบ XGBoost ผลปรากฏว่าการคัดเลือกตัวแปรแบบรวมกลุ่ม (Ensemble Feature Selection) อาจไม่เป็นวิธีการที่ดี ทั้งนี้ควรพิจารณาการใช้วิธีการคัดเลือกตัวแปรแบบวิธีเดียวเช่น RFE เนื่องจากผลของข้อมูลที่นำมาศึกษาพบว่าการคัดเลือกแบบวิธีเดียวด้วย RFE จะให้ผลที่ดีสำหรับชุดข้อมูลที่มีมิติมาก P>>N ในเกณฑ์ 2.5% 5% และ 10% ซึ่งเป็นชุดข้อมูลที่การไม่คัดเลือกตัวแปรจะให้ประสิทธิภาพการทำนายมีค่าน้อยกว่าหรือเท่ากับ 0.9 แต่สำหรับข้อมูล P=N พบว่า ประสิทธิภาพการทำนายโดยไม่ทำการคัดเลือกตัวแปรมีค่าสูงกว่า 0.9 และเมื่อทำการคัดเลือกตัวแปรแล้วประสิทธิภาพการทำนายที่วัดค่าผ่านค่าเฉลี่ย F1-score เพิ่มขึ้นเล็กน้อย ในขณะความเสถียรของการทำนายที่วัดผ่านค่าเบี่ยงเบน F1-score ไม่ต่างจากเดิม การคัดเลือกตัวแปรเข้าตัวแบบจึงให้ผลดีเมื่อข้อมูลมีมิติมากขึ้นปรากฏในชุดข้อมูล P>N และ P>>N

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

การคัดเลือกแบบรวมกลุ่มนั้นจะให้ผลการทำนายที่ต่างกัน ภายใต้ชุดข้อมูลและเกณฑ์ที่เลือกใช้ สำหรับชุดข้อมูล P=N การรวมลำดับความสำคัญตัวแปรรูปแบบ Design CT ด้วยวิธีค่าเฉลี่ยฮาร์มอนิกที่เกณฑ์ 10% จะทำให้ค่าเฉลี่ย F1-score เพิ่มจาก 0.91 เป็น 0.92 จากการไม่คัดเลือกตัวแปร และทำได้ดีกว่าวิธีค่าต่ำสุดใน Design CT สำหรับชุดข้อมูล P>N วิธีค่าเฉลี่ยเลขคณิตที่เกณฑ์ 5% และ 10% และวิธีค่าเฉลี่ยฮาร์มอนิกที่เกณฑ์ 2.5% จะให้ผลดีกว่าการรวมด้วยวิธีอื่น โดยทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้นจาก 0.90 เป็น 0.92 สำหรับชุดข้อมูล P>>N การรวมด้วยวิธีค่ากลางและค่าเฉลี่ยเลขคณิตที่เกณฑ์ $\log_2(P)$ ทำให้ค่าเฉลี่ย F1-score เพิ่มขึ้น จาก 0.88 เป็น 0.92 และลดค่าเบี่ยงเบน F1-score จาก 0.13 ลงเหลือ 0.09 จึงได้ว่าในชุดข้อมูลที่ P=N และ P>N ไม่พบวิธีการรวมลำดับใดในรูปแบบ CT ที่ให้ผลดีกว่าวิธีอื่นในทุกเกณฑ์

การรวมเซตตัวแปรที่สำคัญใน Design TC ด้วยวิธียูเนียนจะให้ประสิทธิภาพการทำนายที่เพิ่มมาจากการไม่คัดเลือกตัวแปร โดยเพิ่มขึ้นมากกว่าวิธีอินเตอร์เซกและมัลติอินเตอร์เซกในชุดข้อมูล P=N ที่ทุกเกณฑ์ การรวมด้วยวิธีอินเตอร์เซกและมัลติอินเตอร์เซกจะให้ผลดีกว่าวิธียูเนียนในชุดข้อมูล P>>N ที่ทุกเกณฑ์และชุดข้อมูล P>N ที่เกณฑ์ 5% และ 10% ดังนั้นในกรณีที่จำนวนตัวแปรต้นมากกว่าขนาดข้อมูลอยู่มาก (P>>N) ควรเลือกใช้วิธีมัลติอินเตอร์เซกหรืออินเตอร์เซกแทนวิธียูเนียน

5.2 ข้อเสนอแนะที่ได้จากงานศึกษา

งานศึกษานี้มีผลลัพธ์เฉพาะเจาะจงจากการเลือกใช้ข้อมูล 3 ชุด ได้แก่ Parkinson's Disease (P=N), LSVT Voice Rehabilitation (P>N) และ Colon Cancer (P>>N) เพื่อประเมินประสิทธิภาพและความเสถียรของการทำนาย ผ่านตัวแบบ XGBoost การคัดเลือกตัวแปรแบบรวมกลุ่ม แบ่งเป็น 2 รูปแบบคือ Design CT และ Design TC อีกทั้งยังนำผลการทำนายเปรียบเทียบกับการคัดเลือกตัวแปรแบบวิธีเดียว ผลการทดลองแสดงให้เห็นว่าการคัดเลือกตัวแปรบางวิธีอาจให้ผลการทำนายที่แย่กว่าการไม่คัดเลือกตัวแปรในชุดข้อมูล P=N แต่สำหรับชุดข้อมูล P>N และ P>>N มีหลายวิธีการคัดเลือกตัวแปรที่ให้ผลลัพธ์ที่ดีขึ้นกว่าการไม่ทำการคัดเลือกตัวแปรเข้าตัวแบบ เช่น การคัดเลือกตัวแปรแบบรวมกลุ่มด้วยวิธีค่าเฉลี่ยเลขคณิตใน Design CT ที่เกณฑ์ $\log_2(P)$ ในชุดข้อมูล P>>N และวิธีมัลติอินเตอร์เซกใน Design TC ที่เกณฑ์ $\log_2(P)$ ในชุดข้อมูล P>>N ให้ผลการทำนายที่ดีขึ้นจากการไม่คัดเลือกตัวแปรและมากกว่าวิธีอื่น ๆ การใช้วิธีการคัดเลือกแบบรวมกลุ่มนั้นต้องการระยะเวลาประมวลผลมากกว่าแบบวิธีเดียว ทำให้การเลือกใช้วิธีเดียวจะทำให้ระยะเวลาการทดสอบลดลง โดยการคัดเลือกตัวแปรแบบวิธีเดียวด้วย RFE เป็นวิธีที่ให้ผลดีเช่นกันเปรียบเทียบกับวิธีอื่น ๆ จะให้ผลการทำนายที่ดีอีกทั้งยังสามารถช่วยลดระยะเวลาการทำงาน ซึ่งขึ้นอยู่กับเกณฑ์ที่เลือกใช้ ผลการศึกษาโดยรวมสอดคล้องกับงานศึกษาอื่น ๆ ในประเด็นเรื่องการเลือกใช้ตัวแปรต้นในจำนวนที่น้อยลงจากชุดข้อมูลดั้งเดิมจะทำให้ประสิทธิภาพและความเสถียรของการทำนายที่วัดจาก 10-fold cross validation มีผลดีขึ้น จากงานศึกษา Almayyan (2020) ที่ได้นำการสุ่มตัวอย่างซ้ำ (Resampling) มาช่วยเรื่องการเพิ่มประสิทธิภาพของการทำนาย ผู้ศึกษาจึงคาดหวังว่าหากต้องการเพิ่มประสิทธิภาพของการทำนายที่นอกเหนือจากทำการคัดเลือกตัวแปรแล้วในอนาคตอาจศึกษาเพิ่มเกี่ยวกับการนำเทคนิคการสุ่มตัวอย่างซ้ำ (Resampling) มาใช้ร่วมพิจารณา อีกทั้งตัวแบบ XGBoost ที่ใช้ในงานศึกษานี้ไม่มีข้อจำกัดในเรื่องของปัญหา Multicollinearity หากเลือกใช้ตัวแบบทำนายอื่นที่มีข้อจำกัดของปัญหาดังกล่าวอาจต้องระวังเรื่องตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบ

บรรณานุกรม

Almayyan, W. (2020). A Modified Maximum Relevance Minimum Redundancy Feature Selection Method Based on Tabu Search For Parkinson's Disease Mining. *International Journal of Artificial Intelligence & Applications*.

Araújo, D., Jesus, J., Neto, A. D., & Martins, A. (2016, 2016//). A Combination Method for Reducing Dimensionality in Large Datasets. Artificial Neural Networks and Machine Learning – ICANN 2016, Cham.

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, *52*, 1-12. https://doi.org/https://doi.org/10.1016/j.inffus.2018.11.008

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2012). An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition*, *45*(1), 531-539. https://doi.org/https://doi.org/10.1016/j.patcog.2011.06.006

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1007/BF00058655

Bühlmann, P. (2012). Bagging, Boosting and Ensemble Methods.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28. https://doi.org/https://doi.org/10.1016/j.compeleceng.2013.11.024

Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics*, *61*, 101224. https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101224

Effrosynidis, D., Tsikliras, A., Arampatzis, A., & Sylaios, G. (2020). Species Distribution Modelling via Feature Engineering and Machine Learning for Pelagic Fishes in the Mediterranean Sea. *Applied Sciences*, *10*(24), 8900. https://www.mdpi.com/2076-3417/10/24/8900

Galli, S. (2022). *Feature Selection in Machine Learning with Python*. Lulu.com.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531-537. https://doi.org/10.1126/science.286.5439.531

Gunduz, H. (2021). An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomedical Signal Processing and Control*, *66*, 102452. https://doi.org/https://doi.org/10.1016/j.bspc.2021.102452

Hartshorn, S. (2017). *Machine Learning With Boosting: A Beginner's Guide*.

Hastie, T. J., & Tibshirani, R. (2003). Expression Arrays and the p n Problem.

Liu, Y., Liu, Z., Luo, X., & Zhao, H. (2022). Diagnosis of Parkinson's disease based on SHAP value feature selection. *Biocybernetics and Biomedical Engineering*, *42*(3), 856-869. https://doi.org/https://doi.org/10.1016/j.bbe.2022.06.007

Polat, K. (2019, 24-26 April 2019). A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT),

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLoS One*, *9*(2), e87357. https://doi.org/10.1371/journal.pone.0087357

Ruiz, R., Riquelme, J. C., & Aguilar-Ruiz, J. S. (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, *39*(12), 2383-2392. https://doi.org/https://doi.org/10.1016/j.patcog.2005.11.001

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., & Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, *74*, 255-263. https://doi.org/https://doi.org/10.1016/j.asoc.2018.10.022

Sara, A. V. D. G., & Hans, C. V. H. (2004). High-dimensional data: p >> n in mathematical statistics and bio-medical applications. *Bernoulli*, *10*(6), 939-943. https://doi.org/10.3150/bj/1106314843

Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2016a). Using a feature selection ensemble on DNA microarray datasets. The European Symposium on

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Artificial Neural Networks,

Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2016b, 2016//). Using Data Complexity Measures for Thresholding in Feature Selection Rankers. Advances in Artificial Intelligence, Cham.

Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2018). On developing an automatic threshold applied to feature selection ensembles. *Inf. Fusion*, *45*, 227-245.

Shafi, A. S. M., Molla, M. M. I., Jui, J. J., & Rahman, M. M. (2020). Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques. *SN Applied Sciences*, *2*(7), 1243. https://doi.org/10.1007/s42452-020-3051-2

Smith, C. (2017). *Decision Trees and Random Forests: A Visual Introduction for Beginners*. Blue Windmill Media. https://books.google.co.th/books?id=Hj_CtAEACAAJ

Starmer, J. (2022). *The StatQuest Illustrated Guide To Machine Learning*. (Qurate Books Private Limited)

Team, D. (2018). Feature Selection in R with the Boruta R Package.

Trevor Hastie, R. T., Jerome Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Tsanas, A., Little, M. A., Fox, C., & Ramig, L. O. (2014). Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease. *IEEE Trans Neural Syst Rehabil Eng*, *22*(1), 181-190. https://doi.org/10.1109/tnsre.2013.2293575

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, *85*, 168-188. https://doi.org/https://doi.org/10.1016/j.jbi.2018.07.015

Wang, H., Khoshgoftaar, T. M., & Napolitano, A. (2010). A Comparative Study of Ensemble Feature Selection Techniques for Software Defect Prediction. *2010 Ninth International Conference on Machine Learning and Applications*, 135-140.

Yi Yang, W. L., Tingting Zeng, Linhan Guo, Yong Qin, Xue Wang. (2022). An Improved Stacking Model for Equipment Spare Parts Demand Forecasting Based on

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Scenario Analysis. *Scientific Programming*, *2022*, Article 5415702. https://doi.org/https://doi.org/10.1155/2022/5415702

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวกรชนก ชมเชย
วัน เดือน ปี เกิด	6 พฤษภาคม 2539
วุฒิการศึกษา	ระดับปริญญาตรี เศรษฐศาสตรบัณฑิต มหาวิทยาลัยธรรมศาสตร์ ระดับปริญญาโท วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติและวิทยาการข้อมูล จุฬาลงกรณ์มหาวิทยาลัย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY