การปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับสำหรับการจำแนกความยากจนระดับครัวเรือนในประเทศไทย

นายธารินทร์ สุขเนาว์

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

IMPROVING BINARY PREDICTION USING ADAPTIVE WEIGHTS META-LEARNING WITH AN APPLICATION TO POVERTY CLASSIFICATION IN THAILAND

Mr. Taryn Suknao

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	การปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับสำหรับการจำแนกความยากจนระดับครัวเรือนในประเทศไทย
โดย	นายธารินทร์ สุขเนาว์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.นัจชลี ศรีมณีกาญจน์)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ธารินทร์ สุขเนาว์ : การปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับสำหรับการจำแนกความยากจนระดับครัวเรือนในประเทศไทย. (IMPROVING BINARY PREDICTION USING ADAPTIVE WEIGHTS META-LEARNING WITH AN APPLICATION TO POVERTY CLASSIFICATION IN THAILAND) อ.ที่ปรึกษาหลัก : รศ. ดร.วิฐรา พึ่งพาพงศ์

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือนและเสนอวิธีการเปรียบเทียบและปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับจากการคำนวนค่าถ่วงน้ำหนักวิธีที่ดีที่สุดสำหรับการจำแนกความยากจนระดับครัวเรือนในประเทศไทย โดยนำเสนอวิธีการสองขั้นตอน คือนำตัววัดประสิทธิภาพการทำนายมาใช้ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ซึ่งนำมาใช้เสมือนเป็นค่าถ่วงน้ำหนักเริ่มต้นที่ให้กับแต่ละตัวแบบ จากนั้นจึงทำนายผลด้วยวิธีการวิเคราะห์การถดถอยลอจิสติกอีกขั้นตอนหนึ่ง งานวิจัยนี้ศึกษาการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนายใน 3 กรณี ได้แก่ 1. การใช้ค่า AUC 2. การใช้ค่า F1-Score โดยพิจารณาจุดตัด 0.5 และ 3. การใช้ค่า F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด นอกจากนี้เนื่องจากชุดข้อมูลสำรวจประชากรรายครัวเรือนในระดับพื้นที่มีความไม่สมดุลของระดับความยากจน จึงใช้เทคนิค SMOTE ในการจัดการกับข้อมูลที่ไม่สมดุล ทั้งนี้ ผู้วิจัยได้ทำการเปรียบเทียบผลลัพธ์จากชุดข้อมูลก่อนและหลังใช้เทคนิค SMOTE ผลการศึกษาพบว่า ปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือนสูงมีหลายปัจจัย อาทิ อายุของหัวหน้าครัวเรือน จำนวนผู้ที่ได้รับบัตรสวัสดิการแห่งรัฐในครัวเรือน,ค่าใช้จ่ายเพื่อการบริโภคในครัวเรือน เป็นต้น และวิธีการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพ F1-Score ที่จุดตัด 0.5 มีประสิทธิภาพสูงสุดจากการพิจารณาด้วยค่าความแม่นยำในชุดข้อมูลตั้งต้นก่อนใช้เทคนิค SMOTE อย่างไรก็ตาม จากการทดสอบในชุดข้อมูลที่มีการจัดการกับข้อมูลที่ไม่สมดุลด้วยวิธี SMOTE พบว่า ประสิทธิภาพในการทำนายไม่ปรากฏว่าวิธีการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพแบบใดแบบหนึ่งที่มีประสิทธิภาพสูงสุดอย่างชัดเจน

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380156226 : MAJOR STATISTICS

KEYWORD: Poverty Prediction, Binary classification, Machine learning, Stacking, Meta-learning

Taryn Suknao : IMPROVING BINARY PREDICTION USING ADAPTIVE WEIGHTS META-LEARNING WITH AN APPLICATION TO POVERTY CLASSIFICATION IN THAILAND. Advisor: Assoc. Prof. Dr. VITARA PUNGPAPONG

This research is aimed to study factors related to poverty at the household level, compare and explore the best method-weighted computation to improve binary prediction using adaptive weights meta-learning with an application to poverty classification in Thailand. This paper is to introduce a two-step meta-learning approach which adaptive weights are calculated by the predictive evaluation metrics. These weights are used as initial assigned to each model. Then, a final prediction is obtained from a logistic regression model. Here, we compare the results from three methods to calculate adaptive weights including 1. AUC Score, 2. F1-Score at the cutoff probability at 0.5, and 3. F1-Score at an optimal cut-off value from the highest Youden's Index. Due to an imbalanced dataset, the SMOTE technique is employed to manage. Comparing results of using SMOTE also included in this research. The results showed that there are many factors highly related to household poverty such as the age of the household head, the number of persons receiving the state welfare card, household consumption expenditures. Furthermore, adaptive weights computed from the F1-Score with cutoff value 0.5 yielded highest prediction accuracy based on the original dataset. However, It was unclear which method to compute adaptive weights give the best results from the dataset with the SMOTE imbalance data handling.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

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

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

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จสมบูรณ์ได้ด้วยดีเพราะได้รับความอนุเคราะห์ เมตตา และเอาใจใส่เป็นอย่างดียิ่งจาก รองศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้ความกรุณารับเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ช่วยให้คำปรึกษา แนะนำ แก้ไขปรับปรุงข้อบกพร่อง รวมทั้งให้องค์ความรู้ แนวทางในการศึกษาค้นคว้า ตลอดจนให้ความช่วยเหลือและอบรมสั่งสอนมาโดยตลอด ผู้วิจัยขอขอบพระคุณท่านอาจารย์เป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์ และผู้ช่วยศาสตราจารย์ ดร.ทพญ.นัจชลี ศรีมณีกาญจน์ กรรมการสอบวิทยานิพนธ์ ท่านอาจารย์ทั้งสามท่านได้กรุณาสละเวลาเป็นกรรมการสอบในครั้งนี้ ตลอดจนช่วยให้ความรู้ คำแนะนำที่มีประโยชน์อย่างยิ่งในการเขียนวิทยานิพนธ์นี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้ง ขอขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ที่ได้อบรม สั่งสอนวิชาความรู้มาจนสำเร็จการศึกษาในครั้งนี้

ขอขอบพระคุณ คุณเสกสรร สุขเนาว์ และคุณจิรพันธ์ สุขเนาว์ ผู้เป็นบิดา และมารดาของผู้วิจัย ซึ่งเป็นผู้สนับสนุน มอบโอกาสในการศึกษาที่ดี และเป็นผู้อยู่เบื้องหลังความสำเร็จของผู้วิจัยเสมอมา รวมถึงครอบครัวที่คอยเป็นกำลังใจให้ผู้วิจัยสามารถจัดทำวิทยานิพนธ์เล่มนี้จนสำเร็จลุล่วงด้วยดี และขอขอบคุณกัลยาณมิตรทุกท่านที่ช่วยเหลือ ให้คำแนะนำ และเป็นกำลังใจ แรงสนับสนุนที่ดีตลอดมา ผู้วิจัยมีความซาบซึ้งในความกรุณาของทุกท่านทั้งที่ได้กล่าวถึงและไม่ได้กล่าวถึงในที่นี้ ทุกท่านได้มีส่วนช่วยเหลือและสนับสนุนด้วยดีตลอดมา จึงขอกราบขอบพระคุณทุกท่านด้วยความจริงใจ

อนึ่ง ผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะก่อให้เกิดประโยชน์แก่ผู้ที่สนใจศึกษาค้นคว้า ในเรื่องดังกล่าว และขอมอบคุณประโยชน์ใดที่เกิดขึ้นจากวิทยานิพนธ์เล่มนี้เป็นกตเวทิตาคุณแด่บิดา มารดา ครูอาจารย์ และผู้เกี่ยวข้อทุกท่านที่ได้กล่าวมาทั้งหมดนี้ ตลอดจนท่านผู้เขียนตำราที่ผู้วิจัยนำมาอ้างอิงและเรียบเรียงเป็นวิทยานิพนธ์เล่มนี้ หากวิทยานิพนธ์เล่มนี้มีข้อบกพร่อง หรือผิดพลาดประการใด ผู้วิจัยขอน้อมรับไว้เพียงผู้เดียวและขออภัยไว้ ณ โอกาสนี้

ธารินทร์ สุขเนาว์

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญภาพ ..... ฌ

สารบัญตาราง ..... ญ

บทที่ 1 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 3

1.3 สมมติฐานการวิจัย 3

1.4 การนำเสนอ 3

1.5 ประโยชน์ที่คาดว่าจะได้รับ 3

บทที่ 2 5

แนวคิดและทฤษฎี 5

2.1 การจำแนกประเภท (Classification) 5

2.2 เทคนิคการวิเคราะห์การจำแนกประเภท 5

2.2.1 การถดถอยลอจิสติก 5

2.2.2 ตัวแบบต้นไม้ (Tree-based Model) 6

2.2.3 โครงข่ายประสาทเทียม (Artificial Neural Network : ANN) 10

2.3 ตัวชี้วัดประสิทธิภาพในการจำแนกประเภทแบบไบนารี 11

2.4 การเรียนรู้เมตา 14

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.5 การเรียนรู้เมตาโดยใช้ค่าถ่วงน้ำหนักแบบปรับ (Adaptive Weight) 16
เอกสารและงานวิจัยที่เกี่ยวข้อง 18
บทที่ 3 23
3.1 ขอบเขตของการวิจัย 23
3.1.1 การคาดการณ์ 23
3.1.2 ขอบเขตของข้อมูล 23
3.1.3 ตัวแบบในการคาดการณ์ที่ใช้ในการวิจัย 23
3.1.4 การปรับปรุงความสามารถในการคาดการณ์ 26
3.1.5 เครื่องมือ 26
3.2 วิธีการดำเนินการวิจัย 26
บทที่ 4 30
4.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่าง ๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย ในแต่ละรอบ 33
4.2 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 34
4.3 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย ในแต่ละรอบ 36
4.4 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 37

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 5 39
5.1 สรุปผลการวิจัย 39
5.2 สรุปและอภิปรายผล 41
5.3 ข้อเสนอแนะ 43
บรรณานุกรม 45
ภาคผนวก 47
ประวัติผู้เขียน 58

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญภาพ

ภาพที่ 2.1 เส้นโค้งการถดถอยแบบโลจิสติกส์ 6

ภาพที่ 2.2 การเรียนรู้แบบต้นไม้ตัดสินใจ 7

ภาพที่ 2.3 เทคนิค Bagging 8

ภาพที่ 2.4 กระบวนการป่าไม้แบบสุ่ม 8

ภาพที่ 2.5 เทคนิค Boosting 9

ภาพที่ 2.6 Level-wise and Leaf-wise tree growth 9

ภาพที่ 2.7 โครงสร้างการเรียนรู้ของโครงข่ายประสาทเทียม 10

ภาพที่ 2.8 ตารางเมทริกซ์ความสับสน 11

ภาพที่ 2.9 เส้นโค้ง AUC-ROC 13

ภาพที่ 2.10 ประสิทธิภาพที่วัดได้จากเส้นโค้ง AUC-ROC 13

ภาพที่ 2.11 ดัชนีของโยเดน 14

ภาพที่ 2.12 เทคนิคสแต็กกิ้ง (Sikora & Al-laymoun, 2014) 15

ภาพที่ 2.13 ตัวอย่าง 5-fold Cross Validation 16

ภาพที่ 2.14 การใช้เทคนิคสแต็กกิ้งร่วมกับการถ่วงน้ำหนักในหนึ่งรอบ (Ren et al., 2022) 17

ภาพที่ 3.1 ขั้นตอนการทำงานในรอบที่ 1 27

ภาพที่ 4.1 กราฟเปรียบเทียบค่าคะแนนคุณลักษณะที่สำคัญ 30

Chulalongkorn University

สารบัญตาราง

ตารางที่ 2.1 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (Alsharkawi et al., 2021) 19

ตารางที่ 2.2 ประสิทธิภาพผลลัพธ์การคาดการณ์ระดับความยากจนของ (Pluliková, 2015) 20

ตารางที่ 2.3 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (S Sani et al., 2018)..... 21

ตารางที่ 2.4 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (Surawatchayotin & Paireekreng, 2021)..... 22

ตารางที่ 3.1 พารามิเตอร์สำหรับสร้างตัวแบบการถดถอยลอจิสติก..... 24

ตารางที่ 3.2 พารามิเตอร์สำหรับสร้างตัวแบบต้นไม้ตัดสินใจ 24

ตารางที่ 3.3 พารามิเตอร์สำหรับสร้างตัวแบบ Bagging Tree 24

ตารางที่ 3.4 พารามิเตอร์สำหรับสร้างตัวแบบ Random Forest 25

ตารางที่ 3.5 พารามิเตอร์สำหรับสร้างตัวแบบ Light GBM..... 25

ตารางที่ 3.6 พารามิเตอร์สำหรับสร้างตัวแบบโครงข่ายประสาทเทียม 25

ตารางที่ 4.1 แสดงค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจนสูงสุด 24 อันดับแรก 31

ตารางที่ 4.2 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ..... 33

ตารางที่ 4.3 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย..... 33

ตารางที่ 4.4 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ย..... 34

ตารางที่ 4.5 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด..... 34

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 4.6 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 35

ตารางที่ 4.7 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 35

ตารางที่ 4.8 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ 36

ตารางที่ 4.9 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย 36

ตารางที่ 4.10 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ 37

ตารางที่ 4.11 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 37

ตารางที่ 4.12 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 38

ตารางที่ 4.13 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ 38

ตารางที่ 5.1 แสดงค่าความแม่นยำเฉลี่ยจากการทดลองของสองชุดข้อมูล และการใช้ตัววัดประสิทธิภาพการทำนายแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด 40

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตารางที่ 5.2 แสดงค่าความแม่นยำเฉลี่ยจากการทดลองของสองชุดข้อมูล และการใช้ตัววัดประสิทธิภาพการทำนายแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย ในแต่ละรอบ..... 41

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันปัญหาความยากจน เป็นปัญหาที่เกิดขึ้นในหลากหลายภูมิภาคทั่วโลก โดยเฉพาะอย่างยิ่งในประเทศกำลังพัฒนา รวมทั้งสังคมไทยที่ประสบปัญหาความยากจนสะสมมาอย่างยาวนาน ปัญหาความยากจนนั้นไม่เพียงจำกัดเฉพาะคนที่ขัดสนด้านรายได้เพียงอย่างเดียว แต่ครอบคลุมถึงความยากจนเชิงโครงสร้างที่เกิดจากความขัดสนในมิติอื่นๆ ด้วย อาทิ ความขัดสนในมิติสุขภาพ สภาพแวดล้อมความเป็นอยู่ และการศึกษา โดยอาจเกิดได้จากปัจจัยภายในตัวบุคคลเองและปัจจัยภายนอกตัวบุคคล ก่อให้เกิดความยากลำบากในการดำเนินชีวิตทั้งในระดับบุคคลและครัวเรือน เป็นอุปสรรคสำคัญที่ส่งผลต่อการพัฒนาประเทศ

จากรายงานการจัดทำดัชนีความยากจนหลายมิติ (Multidimensional Poverty Index : MPI) โดยโครงการพัฒนาแห่งสหประชาชาติ (UNDP) ร่วมกับโครงการพัฒนาทรัพยากรมนุษย์และแก้ไขความยากจนแห่งมหาวิทยาลัยอ็อกซ์ฟอร์ด (Oxford Poverty and Human Development Initiative : OPHI) ในปี 2564 ได้รวบรวมข้อมูลจาก 109 ประเทศ ครอบคลุมประชากรรวม 5.9 พันล้านคน พบว่าประชากรประมาณ 1.3 พันล้านคนต้องเผชิญกับปัญหาความยากจนหลายมิติ ซึ่งร้อยละ 67 อาศัยอยู่ในประเทศรายได้ปานกลาง รวมถึงประเทศไทย สำหรับรายงานภาพรวมคนจนในปี 2565 จากระบบบริหารจัดการข้อมูลการพัฒนาคนแบบชี้เป้า (Thai People Map and Analytics Platform : TPMAP) พบว่าจากประชากรสำรวจ 36,103,806 คน มีคนจนเป้าหมายจำนวน 1,025,782 คน คิดเป็นร้อยละ 2.84 ของประชากรสำรวจ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากการประชุมประเทศสมาชิกของสหประชาชาติได้มีข้อตกลงให้การแก้ไขปัญหาความยากจนเป็นเป้าหมายหลักอันดับ 1 ของการพัฒนาอย่างยั่งยืน (Sustainable Development : SDGs) ขณะเดียวกัน ในปี 2563 ประเทศจีนได้ประสบความสำเร็จในการแก้ไขปัญหาความยากจนอย่างเบ็ดเสร็จ โดยใช้มาตรการการขจัดความยากจนอย่างตรงจุด (Targeted Poverty Eradication : TPE) ซึ่งหนึ่งในปัจจัยสำคัญที่ทำให้มาตรการดังกล่าวประสบผลสำเร็จคือ การออกแบบมาตรการอย่างเฉพาะเจาะจงกับกลุ่มเป้าหมาย มีมาตรการที่แตกต่างกันตามบริบทของครัวเรือน ส่วนหนึ่งมีการใช้ฐานข้อมูลประชากรร่วมเป็นส่วนหนึ่งในการกำหนดใช้มาตรการได้อย่างตรงกลุ่มเป้าหมาย และในประเทศไทยนั้น ได้มีความพยายามอย่างต่อเนื่องในการดำเนินมาตรการในการแก้ไขปัญหาความยากจนแบบพุ่งเป้าแต่ละครัวเรือน เพื่อให้เกิดการแก้ไขปัญหาได้อย่างสอดคล้องกับศักยภาพและความต้องการของประชากรในแต่ละพื้นที่

(Alsharkawi et al., 2021) การวิเคราะห์คุณลักษณะเฉพาะและพัฒนาตัวแบบสำหรับการจำแนกกลุ่มความยากจน (Poverty Classification) ระดับครัวเรือนในประเทศจอร์แดน โดยใช้ข้อมูลจากการสำรวจประชากรรายครัวเรือนภายในประเทศจอร์แดน ซึ่งมีลักษณะเฉพาะทั้งสิ้น 47 คุณลักษณะมาทำการเปรียบเทียบตัวแบบสำหรับการจัดกลุ่มความยากจนโดยใช้ตัวแบบจากการเรียนรู้ของเครื่องทั้งหมด 16 ตัวแบบ ร่วมกับการใช้เทคนิคการจัดการข้อมูลที่ไม่สมดุล (Imbalanced Data) เนื่องจากข้อมูลที่จัดเก็บมานั้นมีความแตกต่างกันของจำนวนข้อมูลครัวเรือนยากจนและครัวเรือนปกติเป็นจำนวนมาก จากการศึกษาพบว่าตัวแบบที่มีประสิทธิภาพสูงสุดในการจัดกลุ่มความยากจน 2 อันดับแรก ได้แก่ตัวแบบ Light GBM และ Bagged Decision Trees แม้ว่างานของ (Alsharkawi et al., 2021) จะใช้ตัวแบบการเรียนรู้ของเครื่องจำนวนมากถึง 16 ตัวแบบ แต่ยังไม่มีการนำการเรียนรู้เมตา (Meta Learning) มาใช้ในการนำผลพยากรณ์จากหลากหลายตัวแบบ

สแต็กกิ้ง (Stacking) เป็นหนึ่งในวิธีการเรียนรู้แบบรวม (Ensemble learning) เพื่อปรับปรุงและพัฒนาประสิทธิภาพของผลการทำนายจากหลากหลายตัวแบบผ่านการเรียนรู้เมตา การเรียนรู้เมตามักจะหาค่าถ่วงน้ำหนักของผลลัพธ์ที่ได้จากแต่ละตัวแบบเพื่อให้ได้มาซึ่งค่าพยากรณ์ที่ดีที่สุด วิธีการพื้นฐานที่มักใช้ในการเรียนรู้แบบเมตาคือใช้การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression) ในการคำนวณค่าถ่วงน้ำหนัก (Ren et al., 2022) นำเสนอวิธีการสองขั้นตอน โดยขั้นตอนแรกจะนำตัววัดประสิทธิภาพการทำนาย (Predictive Evaluation Metric) ซึ่งคือค่า AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ (Adaptive Weight) โดยจะนำมาใช้เสมือนเป็นค่าถ่วงน้ำหนักเริ่มต้นที่ให้กับแต่ละตัวแบบ จากนั้นจึงคำนวณเพื่อหาค่าถ่วงน้ำหนักด้วยวิธีการวิเคราะห์การถดถอยลอจิสติกอีกขั้นตอนหนึ่ง แม้ว่าค่า AUC จะเป็นตัววัดที่ใช้กันอย่างแพร่หลายในการวัดประสิทธิภาพการทำนายสำหรับตัวแปรตามแบบไบนารี อย่างไรก็ดี มีงานวิจัยหลายงานที่กล่าวถึงข้อเสียของการใช้ค่า AUC ในการวัดประสิทธิภาพโดยเฉพาะกับกรณีที่ข้อมูลไม่สมดุล (Chicco & Jurman, 2020; Hanczar et al., 2010; Lobo et al., 2008; Muschelli, 2020) ในงานวิจัยนี้จะนำเสนอการใช้ตัววัดอีกตัวหนึ่งที่นิยมใช้กันในการวัดประสิทธิภาพตัวแบบพยากรณ์แบบไบนารีมาใช้ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ คือค่า F1-Score โดยใช้ค่า F1-score จากจุดตัดที่แตกต่างกันในการเปรียบเทียบประสิทธิภาพร่วมกับค่า AUC

ในปัจจุบันงานวิจัยเกี่ยวกับการวิเคราะห์คุณลักษณะที่ส่งผลต่อความยากจนในระดับครัวเรือน และการพยากรณ์ความยากจนในระดับครัวเรือนของประเทศไทยนั้น จะใช้ข้อมูลเชิงสถิติเป็นส่วนใหญ่ในการวิเคราะห์ ซึ่งจนถึงปัจจุบันยังไม่มีงานวิจัยชิ้นใดที่นำเสนอโดยการนำข้อมูลระดับครัวเรือนที่ได้จากการลงพื้นที่สำรวจมาวิเคราะห์และคาดการณ์ความยากจนร่วมกับวิธีการเรียนรู้เมตา สำหรับการวิจัยนี้มีการศึกษาเพื่อหาลักษณะความสัมพันธ์และวิเคราะห์คุณลักษณะที่ส่งผลต่อความยากจนในระดับครัวเรือน พัฒนาตัวแบบสำหรับการคาดการณ์ความยากจนในระดับครัวเรือน

เปรียบเทียบและปรับปรุงความสามารถในการคาดการณ์ด้วยการเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับจากการคำนวณค่าถ่วงน้ำหนักวิธีที่ดีที่สุด และใช้ผลลัพธ์การคาดการณ์จากหลากหลายตัวแบบเพื่อพยากรณ์ตัวแปรตามในขั้นสุดท้าย โดยใช้ข้อมูลสำรวจประชากรรายครัวเรือนในระดับพื้นที่ ซึ่งได้รับความอนุเคราะห์ข้อมูลจากสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ผลการวิจัยดังกล่าวสามารถนำไปใช้ประโยชน์ทางนโยบาย การจัดทำโครงการ หรือกำหนดมาตรการเพื่อแก้ไขปัญหาความยากจนได้หลากหลายมิติ อาทิ การให้ความช่วยเหลือในกลุ่มประชากรยากจนได้อย่างตรงเป้าหมาย การจัดลำดับความสำคัญของครัวเรือนที่ควรได้รับความช่วยเหลือเร่งด่วน การกำหนดหลักเกณฑ์ในการรับสวัสดิการภาครัฐ รวมทั้งการช่วยเหลือจากภาคเอกชนที่สามารถให้ความช่วยเหลือได้อย่างเฉพาะเจาะจงกับกลุ่มเป้าหมายเพื่อการขจัดปัญหาความยากจนตามเป้าหมายหลักของการพัฒนาอย่างยั่งยืน (SDGs)

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือน
2. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการทำนายของวิธีคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนวิธีต่าง ๆ

1.3 สมมติฐานการวิจัย

การคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนโดยใช้ค่า F1-Score ในขั้นตอนการคำนวณค่าถ่วงน้ำหนักมีประสิทธิภาพในการทำนายไม่น้อยกว่าการใช้ค่า AUC ในขั้นตอนการคำนวณค่าถ่วงน้ำหนัก

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1.4 การนำเสนอ

นำเสนอข้อมูลในรูปแบบตารางเพื่อเปรียบเทียบว่าการเรียนรู้เมตาทั้ง 3 กรณี แบบใดที่มีประสิทธิภาพสูงสุด

1.5 ประโยชน์ที่คาดว่าจะได้รับ

สามารถใช้การเรียนรู้เมตาด้วยวิธีการถ่วงน้ำหนักวิธีที่ดีที่สุดจากการทดลองเพื่อปรับปรุงประสิทธิภาพในการคาดการณ์ให้มีประสิทธิภาพสูงยิ่งขึ้น ทราบถึงพารามิเตอร์ที่เหมาะสมเพื่อเป็นแนวทางในการสร้างตัวแบบเพื่อคาดการณ์ความยากจนในระดับครัวเรือน และทราบถึงคุณลักษณะ

สำคัญที่ส่งผลให้เกิดความยากจนในระดับครัวเรือนจากชุดข้อมูลที่มีคุณลักษณะแตกต่างจากชุดข้อมูลที่บ่งบอกสถานะความยากจน

ข้อมูลดังกล่าวสามารถนำไปใช้ประโยชน์ทางนโยบายของรัฐได้หลายด้าน อาทิ นโยบายในการแก้ไขปัญหาความยากจน การให้ความช่วยเหลือในประชากรยากจนได้อย่างตรงเป้าหมาย การจัดลำดับความสำคัญของครัวเรือนเป้าหมายที่ควรได้รับความช่วยเหลือเร่งด่วน จัดลำดับความสำคัญของปัจจัยที่ควรได้รับความช่วยเหลือ หรือการออกมาตรการเพื่อช่วยเหลือ เยียวยาปัญหาความยากจนในมิติต่าง ๆ รวมทั้งการช่วยเหลือจากภาคเอกชนที่สามารถให้ความช่วยเหลือได้อย่างเฉพาะเจาะจงกับกลุ่มเป้าหมายและมิติความขัดสน เพื่อให้เกิดการพัฒนาอย่างยั่งยืนในอนาคตได้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

แนวคิดและทฤษฎี

งานวิจัยนี้ได้ศึกษาทฤษฎีและนำเสนอแนวทางการสร้างแบบจำลองในการคาดการณ์ความยากจนของประชากรรายครัวเรือนโดยใช้การเรียนรู้ของเครื่อง และการปรับปรุงประสิทธิภาพของการคาดการณ์โดยใช้แบบจำลองการเรียนรู้แบบรวม ซึ่งมีแนวคิดและทฤษฎีที่เกี่ยวข้อง ดังต่อไปนี้

2.1 การจำแนกประเภท (Classification)

การจำแนกประเภทเป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง (Machine learning) ในรูปแบบของการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งเป็นวิธีการที่ผู้ใช้งานหรือผู้สอนนั้น นำชุดข้อมูลการฝึกและผลลัพธ์ (Training data) เข้าสู่เครื่อง โดยเครื่องจะทำการเรียนรู้ความสัมพันธ์ของข้อมูลและผลลัพธ์นั้น ให้ได้มาซึ่งข้อกำหนดคำสั่ง กฎหรือเงื่อนไขต่างๆ ในรูปของสมการทางคณิตศาสตร์ หรือที่เรียกว่าอัลกอริทึม ซึ่งเมื่อมีการนำเข้าข้อมูลชุดใหม่ก็จะใช้อัลกอริทึมนั้นในการคาดการณ์ผลลัพธ์ โดยการเรียนรู้เพื่อจำแนกประเภทนั้นจะพยากรณ์ข้อมูลที่มีผลลัพธ์แบบไม่ต่อเนื่อง (Discrete) โดยจำแนกประเภทข้อมูลออกเป็นประเภทต่างๆ ตามที่ได้ป้อนข้อมูลให้เครื่องได้เรียนรู้ผลลัพธ์ ซึ่งจะให้ผลลัพธ์เป็นระดับ (Class) ตามสิ่งที่เคยเรียนรู้มาเท่านั้น ไม่สามารถให้คำตอบที่นอกเหนือจากผลลัพธ์ที่เคยป้อนไปได้ อาทิ การจำแนกเพศจากลักษณะเด่นของข้อมูลคน การจำแนกลูกค้าที่มีโอกาสยกเลิกการใช้งานบัตรเครดิตจากข้อมูลพฤติกรรมของลูกค้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2 เทคนิคการวิเคราะห์การจำแนกประเภท

2.2.1 การถดถอยลอจิสติก

ในการวิเคราะห์การถดถอยอย่างง่าย แสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามอยู่ในรูปเชิงเส้น ดังสมการที่ (1)

$$y = X\beta + \varepsilon \qquad \text{...(1)}$$

เมื่อ y คือ เวกเตอร์ของตัวแปรตามขนาด n

X คือ เมทริกซ์ของตัวแปรอิสระขนาด n x p

β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด p

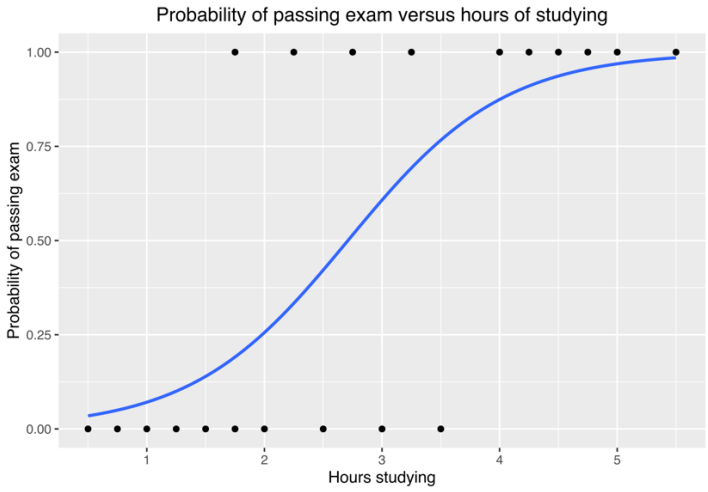
ε คือ เวกเตอร์ของค่าความคาดเคลื่อนขนาด n โดยที่ $E(\varepsilon_i) = 0$ และ $Var(\varepsilon_i)X = \sigma^2 I_n$

แต่ในการวิเคราะห์ลอจิสติกแบบทวิจะใช้เทคนิคในการวิเคราะห์ตัวแปรที่มีวัตถุประสงค์เพื่อประมาณค่าหรือคาดการณ์เหตุการณ์ที่สนใจว่าจะเกิดขึ้นหรือไม่เกิดเหตุการณ์นั้นโดยการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม โดยตัวแปรตามเป็นข้อมูลคุณภาพมีได้สองค่าคือ 0 กับ 1 อาทิ ข้อมูลที่มีค่า“จริง”กับ“ไม่จริง”ซึ่งความสัมพันธ์ของตัวแปรของการวิเคราะห์ลอจิสติกจะอยู่ในรูปคล้ายตัว S ดังภาพที่ 2.1 ทั้งนี้ตัวแปรอิสระ(ตัวแปรทำนาย) จะเป็นได้ทั้งข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณที่อาจจะมีตัวเดียวหรือหลายตัวก็ได้ การวิเคราะห์ลอจิสติกเป็นการประมาณค่าความน่าจะเป็นของเหตุการณ์ แสดงสมการดังสมการที่ (2)

$$P_y = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_p X_p)}} \quad \text{...(2)}$$

เมื่อ P_y คือ ความน่าจะเป็นของการเกิดเหตุการณ์ y

e คือ ฟังก์ชันเอ็กซ์โพเนนเชียล (e = 2.71828)



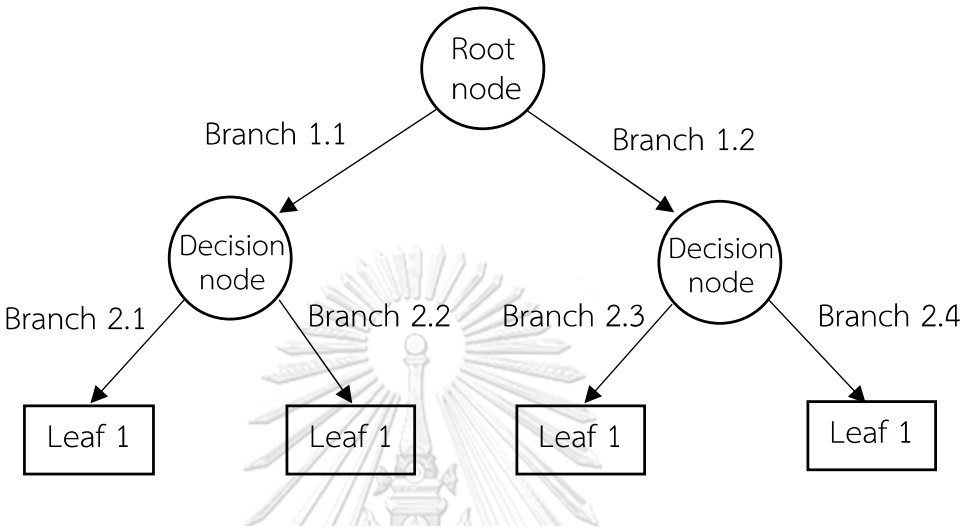
ภาพที่ 2.1 เส้นโค้งการถดถอยแบบโลจิสติกส์
(https://en.wikipedia.org/wiki/Logistic_regression)

2.2.2 ตัวแบบต้นไม้ (Tree-based Model)

1. ต้นไม้ตัดสินใจ (Decision Tree)

การเรียนรู้แบบต้นไม้ตัดสินใจ เป็นพื้นฐานแนวคิดในการสร้างตัวแบบการเรียนรู้ด้วยเครื่องหลายชนิด มีแนวคิดโดยการสร้างต้นไม้แบบกลับหัวขึ้นเพื่อตัดสินใจจากข้อมูลที่มีหมวดหมู่ต้นไม้ตัดสินใจประกอบด้วยบัพ (node) ลักษณะต่าง ๆ ที่ถูกใช้ในการแสดงเงื่อนไขหรือแอทริบิวหนึ่งของข้อมูล โดยแต่ละกิ่ง (link or branch) ของบัพจะหมายถึงค่าที่เป็นไปได้จากการทดลอง

กับแอทริบิวนั้น และประกอบด้วยบัพใบ (leaf node) ที่จะจัดเก็บหมวดหมู่ของข้อมูล ในแต่ละใบแสดงถึงผลลัพธ์การตัดสินใจ โดยต้นไม้ตัดสินใจมีจุดเริ่มต้นที่เรียกว่า บัพราก (root node) จากนั้นจะทำการสร้างบัพทีละบัพเพื่อตรวจสอบคุณลักษณะและย่อยลงไปตามกิ่งต่าง ๆ แสดงดังภาพที่ 2.2



ภาพที่ 2.2 การเรียนรู้แบบต้นไม้ตัดสินใจ

ในการสร้างบัพของต้นไม้ตัดสินใจนั้นจะถูกสกัดโดยการวัดค่าข้อมูลที่ได้รับ (Gain) เพื่อนำมาคัดแยกคุณลักษณะไปยังแต่ละบัพของต้นไม้ โดยค่าดังกล่าวสามารถคำนวณได้จากค่าเอ็นโทรปี (Entropy) ดังสมการที่ (3)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$E(S) = -\sum_{i=1}^{n} P_i log_2 P_i \quad \text{...(3)}$$

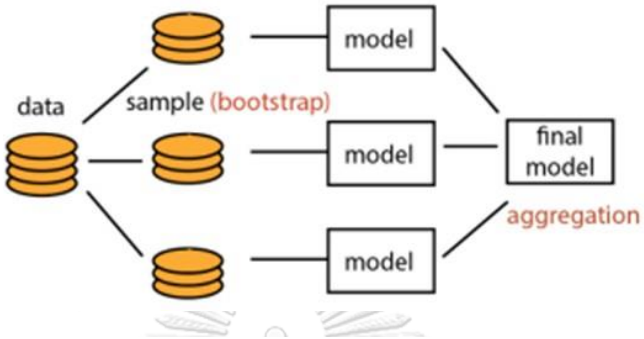
เมื่อ $E(S)$ คือ ค่าเอ็นโทรปีของเซตข้อมูลทั้งหมด

P_i คือ ค่าความน่าจะเป็นของข้อมูล

2. Bagged Tree

เป็นวิธีการที่ต่อยอดมาจากการเรียนรู้แบบต้นไม้ตัดสินใจโดยจะใช้เทคนิค Bagging มาพัฒนาประสิทธิภาพของตัวแบบ โดยการใช้เทคนิคบูทสแตรป (Bootstrap Aggregation) ซึ่งเป็นเทคนิคที่ช่วยลดความแปรปรวนของผลลัพธ์ โดยการสร้างชุดข้อมูลใหม่จากการสุ่มข้อมูลมาใช้แทนข้อมูลเดิม แบ่งออกเป็นชุด ๆ หรือที่เรียกว่าต้นไม้หลายๆ ต้น (Subset) ซึ่งในวิธี Bagged นี้ แต่ละต้นจะใช้ค่าคุณลักษณะของข้อมูลตั้งต้นครบทุกหลัก ทำการทดลองในแต่ละตัวแบบและตัดสินใจผลลัพธ์

สุดท้ายร่วมกันด้วยการหาค่าเฉลี่ย (Averaging) หรือค่าที่มากที่สุด (Voting) แสดงดังภาพที่ 2.3 ข้อดีของเทคนิค Bagging คือ การเรียนรู้ของเครื่องในแต่ละรูปแบบสามารถทำงานกันได้พร้อมกันแบบคู่ขนาน เร็วและไม่ซับซ้อน

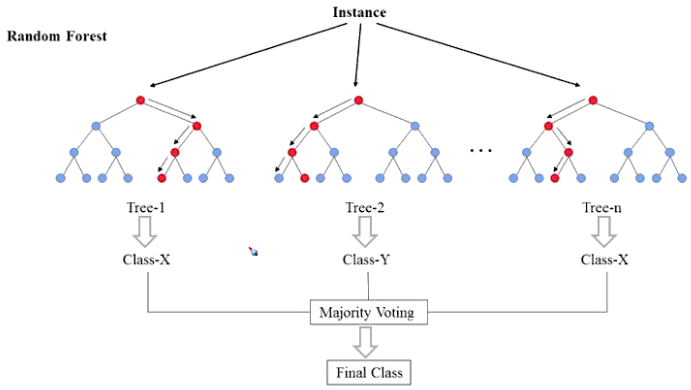


ภาพที่ 2.3 เทคนิค Bagging
(https://tupleblog.github.io/bagging-boosting/)

3. ป่าไม้แบบสุ่ม (Random Forest)

วิธีการป่าไม้แบบสุ่มเป็นวิธีการที่ต่อยอดมาจากการเรียนรู้แบบต้นไม้ตัดสินใจและใช้เทคนิคเดียวกันกับ Bagged Tree ข้างต้น แตกต่างกันที่ในการใช้เทคนิคบูทสแตรปในการแบ่งชุดข้อมูลโดยวิธีการป่าไม้แบบสุ่ม ในแต่ละชุดข้อมูลนั้นจะไม่ใช้ค่าคุณลักษณะของข้อมูลครบทุกตัว จะเป็นการสุ่มค่าคุณลักษณะของข้อมูลพร้อมกับการสุ่มข้อมูลมาสร้างเป็นชุดข้อมูลย่อยด้วย สามารถช่วยลดโอกาสการเกิด overfit ของตัวแบบ การทำงานของป่าไม้แบบสุ่มแสดงดังภาพที่ 2.4

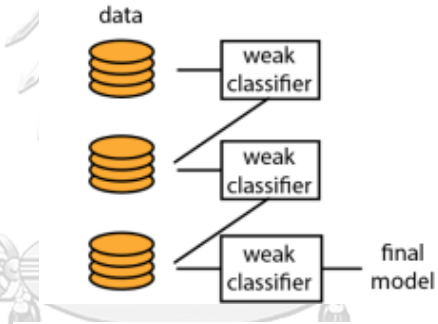
จุฬาลงกรณ์มหาวิทยาลัย



ภาพที่ 2.4 กระบวนการป่าไม้แบบสุ่ม
(https://gaussian37.github.io/ml-concept-RandomForest/)

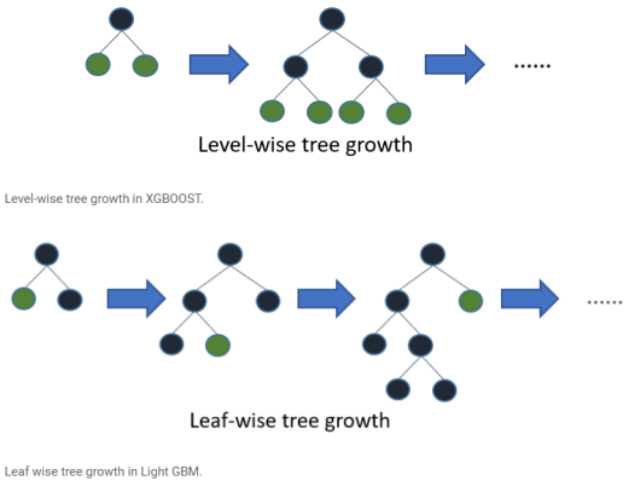
4. Light Gradient Boosting Machine (Light GBM)

Light GBM พัฒนาขึ้นจากเทคนิค Boosting ซึ่งเป็นเทคนิคที่ช่วยลดความผิดพลาดของผลลัพธ์ เริ่มต้นโดยการให้ชุดข้อมูลมีค่าน้ำหนัก (weight) เท่ากัน และทำการปรับค่าน้ำหนักหลังจากระบบได้มีการเรียนรู้ หากมีการจำแนกผิดค่าน้ำหนักก็จะน้อยลง เป็นการเรียนรู้ข้อผิดพลาดจากการกระทำก่อนหน้า ทำให้มีความแม่นยำที่สูงขึ้น โดยไม่ได้เป็นการเอาผลลัพธ์ที่ตัวแบบตัวสุดท้ายมาเป็นตัวตัดสิน แต่จะใช้ตัวแบบทั้งหมดที่เรียนรู้มาใช้ แสดงดังภาพที่ 2.5 ข้อดีของเทคนิค Boosting คือสามารถลดความอคติ (Bias) ของตัวแบบได้ดีและสามารถประยุกต์ใช้กับตัวแบบที่หลากหลาย ซึ่ง Light GMB นั้นถูกพัฒนาขึ้นเพื่อลดเวลาในการใช้งานตัวแบบ มีลักษณะการทำงานโดยการสร้างต้นไม้ตัดสินใจ ต่อจากจุดที่เลือกไว้และสร้างต่อจากผลที่ได้เป็นขั้น เป็นการเติบโตแนวตั้ง โดยไม่ต้องคำนวณจากต้นไม้ตัดสินใจทั้งหมดที่สร้างขึ้นพร้อมกันใน XGboost โดยมีอัลกอริทึมในการใช้ฮิสโตแกรม ในการจัดกลุ่มข้อมูล ช่วยให้ Light GBM สามารถทำงานได้เร็วขึ้น ดังภาพที่ 2.6 แสดงความแตกต่างระหว่าง XGboost และ Light GBM



ภาพที่ 2.5 เทคนิค Boosting
(https://tupleblog.github.io/bagging-boosting/)

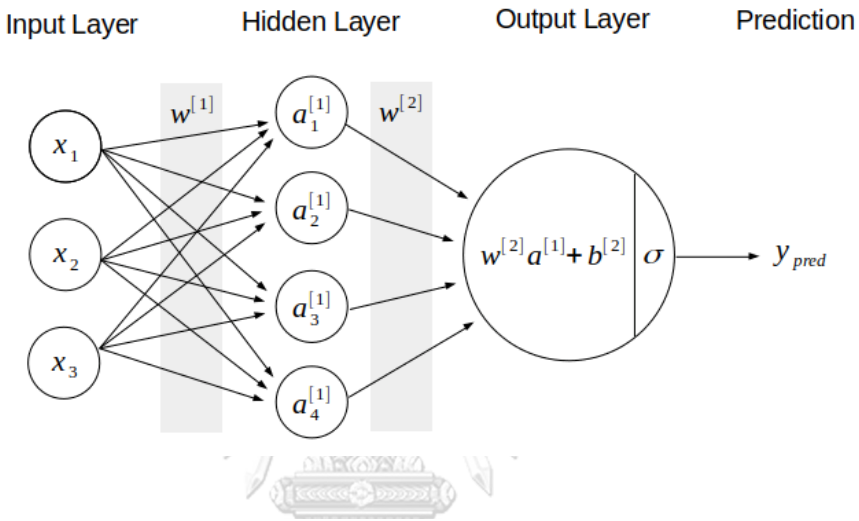
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



ภาพที่ 2.6 Level-wise and Leaf-wise tree growth

2.2.3 โครงข่ายประสาทเทียม (Artificial Neural Network : ANN)

โครงข่ายประสาทเทียมเป็นรูปแบบการทำงานที่จำลองมาจากการทำงานของเซลล์ประสาท โดยหนึ่งหน่วยการทำงานของโครงข่ายประสาทเทียมเทียบได้กับหนึ่งเซลล์ประสาท การทำงานของโครงข่ายประสาทเทียมนั้นจะเป็นการรวบรวมแต่ละหน่วยของประสาทเทียมมาประมวลผลร่วมกัน เมื่อมีการป้อนข้อมูลเพิ่มขึ้น จำนวนของหน่วยประสาทเทียมและชั้นประมวลผลที่ซ่อนอยู่ (Hidden Layer) ก็จะมีเพิ่มมากขึ้นด้วย แสดงดังภาพที่ 2.7



ภาพที่ 2.7 โครงสร้างการเรียนรู้ของโครงข่ายประสาทเทียม
(https://guopai.github.io/ml-blog14.html)

จากภาพที่ 2.7 โครงสร้างการเรียนรู้ของโครงข่ายประสาทเทียมนั้นมีความแตกต่างจากตัวแบบการเรียนรู้ของเครื่องแบบทั่วไป คือมีส่วนของชั้นประมวลผลที่ซ่อนอยู่ซึ่งสามารถมีได้หลายชั้น โดยในแต่ละชั้นจะมีหน่วยประมวลผลย่อย คือ Neuron ($a_i^{[1]}$) ทำหน้าที่รับข้อมูลทุกตัวจากชั้นอินพุตก่อนหน้ามาประมวลผลร่วมกับค่าน้ำหนัก (ω^1) โดยใช้การถดถอยเชิงเส้น ได้มาซึ่งค่า $a^{[1]}$ ในชั้นประมวลผลที่ซ่อนอยู่ จากนั้นในชั้นเอาพุต (Output Layer) จะทำการประมวลผล $a^{[1]}$ ร่วมกับค่าน้ำหนัก (ω^2) ที่ได้จากชั้นก่อนหน้า ได้ผลเป็น $a^{[2]}$ จากนั้นนำผล $a^{[2]}$ มาตัดสินใจเพื่อทำนายผลตามสมการดังนี้

$$\hat{y}^{(i)} = \sigma(a^{[2](i)}) \quad \text{...(4)}$$

ในการพยากรณ์ค่าแบบไบนารีอาจกำหนด σ เป็นฟังก์ชัน Sigmoid เพื่อให้ค่า $\hat{y}^{(i)}$ เป็นค่า 0 – 1 เป็นต้น

$$y_{predict}^{(i)} = \begin{cases} 1 \ if \ a^{[2](i)} \ \geq \ \tau \\ 0 \ if \ a^{[2](i)} \ < \ \tau \end{cases} \quad ...(5)$$

เมื่อ τ คือค่าจุดตัด (Threshold) ที่ต้องการ

โครงข่ายประสาทเทียมนั้น สามารถประยุกต์ใช้งานได้หลากหลายครอบคลุมทุกรูปแบบ ทั้งการถดถอย การจำแนกข้อมูล ทั้งการเรียนรู้แบบไม่มีผู้สอนและการเรียนรู้แบบมีผู้สอน ข้อมูลที่ไม่มีโครงสร้างหรือมีค่าคุณลักษณะจำนวนมากได้

2.3 ตัวชี้วัดประสิทธิภาพในการจำแนกประเภทแบบไบนารี

เมทริกซ์สับสน (Confusion Matrix) คือ ตารางในการประเมินความสามารถการเรียนรู้ของเครื่องในการแก้ปัญหาการจำแนกประเภท โดยเป็นการแสดงสัดส่วนของค่าจริงของข้อมูลและค่าที่ได้จากการคาดการณ์ ในกรณีงานวิจัยนี้ ผลลัพธ์ของการคาดการณ์เป็นได้สองค่า จึงใช้เมททริกซ์ขนาด 2x2 แสดงในภาพที่ 2.8

	ข้อมูลจริง บวก (1)	ข้อมูลจริง ลบ (0)
ข้อมูลคาดการณ์ บวก (1)	True Positives (TPs)	False Positives (FPs)
ข้อมูลคาดการณ์ ลบ (0)	False Negatives (FNs)	True Negatives TNs

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาพที่ 2.8 ตารางเมทริกซ์ความสับสน

รายละเอียด ดังนี้

- True Positive (TP) คือ สิ่งที่คาดการณ์ตรงกับข้อมูลโดยคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริง
- True Negative (TN) คือ สิ่งที่คาดการณ์ตรงกับข้อมูลโดยคาดการณ์ว่าไม่จริงและข้อมูลมีค่าเป็นไม่จริง
- False Positive (FP) คือ สิ่งที่คาดการณ์ไม่ตรงกับข้อมูลโดยคาดการณ์ว่าจริงแต่ข้อมูลมีค่าเป็นไม่จริง
- False Negative (FN) คือ สิ่งที่คาดการณ์ไม่ตรงกับข้อมูลโดยคาดการณ์ว่าไม่จริงแต่ข้อมูลมีค่าเป็นจริง

โดยนำไปใช้ในการคำนวณดังนี้

2.3.1 ค่าความเที่ยงตรง (Precision)

เป็นการประเมินความสามารถของแบบจำลองโดยการเปรียบเทียบสัดส่วนของจำนวนการคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริง เทียบกับจำนวนการคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริงรวมกับจำนวนการคาดการณ์ว่าจริงแต่ข้อมูลเป็นไม่จริง ดังสมการที่ (6)

$$\text{ค่าความเที่ยงตรง (Precision)} = \frac{\text{TPs}}{(\text{TPs}+\text{FPs})} \quad \ldots(6)$$

2.3.2 ค่าการเรียกคืน (Sensitivity, Recall)

เป็นการประเมินความสามารถของแบบจำลองโดยการเปรียบเทียบสัดส่วนของจำนวนการคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริง เทียบกับจำนวนการคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริงรวมกับจำนวนการคาดการณ์ว่าไม่จริงแต่ข้อมูลเป็นจริง ดังสมการที่ (7)

$$\text{ค่าการเรียกคืน (Sensitivity, Recall)} = \frac{(\text{TPs})}{(\text{TPs}+\text{FNs})} \quad \ldots(7)$$

2.3.3 ค่าความแม่นยำ (Accuracy)

เป็นการประเมินความสามารถของแบบจำลองโดยการเปรียบเทียบระหว่างสัดส่วนของจำนวนการคาดการณ์ว่าจริงและข้อมูลมีค่าเป็นจริงรวมกับจำนวนการคาดการณ์ว่าไม่จริงและข้อมูลเป็นไม่จริง เทียบกับผลรวมของจำนวนทั้งหมดในตาราง ดังสมการที่ (8)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\text{ค่าความแม่นยำ (Accuracy)} = \frac{(\text{TPs}+\text{TNs})}{(\text{TPs}+\text{TNs}+\text{FPs}+\text{FNs})} \quad \ldots(8)$$

2.3.4 ค่าความถูกต้องของ F1 (F1-Score)

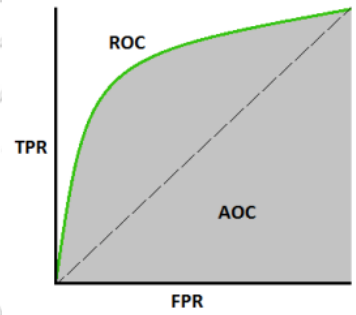
เป็นการประเมินความสามารถของแบบจำลองโดยเป็นการหาค่าเฉลี่ยแบบฮาร์โมนิกระหว่างค่าความเที่ยงตรง (Precision) และค่าการเรียกคืน (Recall) ดังสมการที่ (9)

$$\text{ค่าความถูกต้องของ F1 (F1-Score)} = \frac{2 \times (\text{Precision}\times\text{Recall})}{(\text{Precision}\times\text{Recall})} \quad \ldots(9)$$

2.3.5 เส้นโค้ง AUC-ROC

เป็นการวัดประสิทธิภาพสำหรับปัญหาการจำแนกประเภท โดยการเปรียบเทียบพื้นที่ใต้โค้ง หรือ Area Under The Curve (AUC) ภายใต้เส้นโค้ง หรือ Receiver Operating Characteristics (ROC) โดย AUC แสดงถึงระดับหรือการวัดความสามารถในการแยกประเภท และ ROC คือ เส้นกราฟความน่าจะเป็นโดยค่าดังกล่าวจะเป็นการบอกประสิทธิภาพของแบบจำลองว่าสามารถจำแนกประเภทได้เพียงใด โดยเส้นโค้ง ROC จะเป็นการพล็อตค่าการเรียกคืน หรือ True Positive Rate (TPR) ในแนวแกน y เทียบกับ False Positive Rate (FPR) ดังสมการที่ (10) ในแนวแกน x ดังภาพที่ 2.9

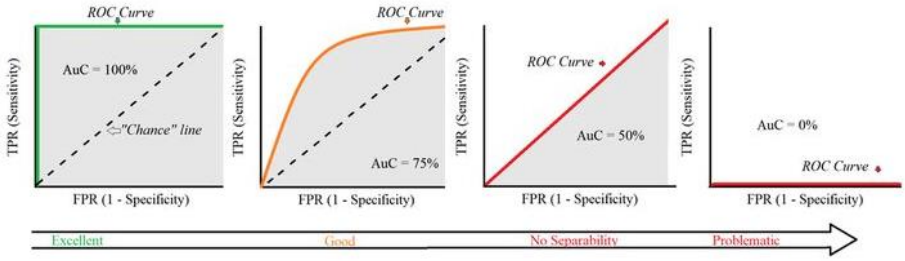
$$FPR = \frac{FPs}{(TNs+FPs)} \quad ...(10)$$



ภาพที่ 2.9 เส้นโค้ง AUC-ROC
(https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ค่า AUC จะมีค่าอยู่ระหว่าง 0-1 โดยเมื่อ AUC มีค่าสูงเข้าใกล้ 1 แสดงว่าแบบจำลองมีประสิทธิภาพสูงสามารถคาดการณ์ได้ถูกต้องแม่นยำมากขึ้น และเมื่อ AUC มีค่าน้อยหรือต่ำกว่า 0.5 แสดงถึงแบบจำลองคาดการณ์นั้นไม่มีประสิทธิภาพ ไม่ต่างจากการเดาสุ่ม แสดงดังภาพที่ 2.10

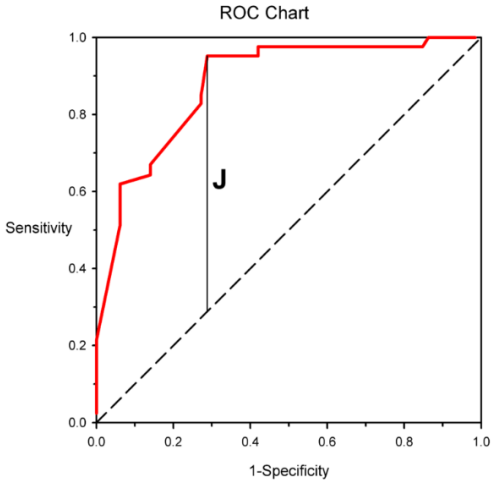


ภาพที่ 2.10 ประสิทธิภาพที่วัดได้จากเส้นโค้ง AUC-ROC
(https://www.datasciencecentral.com/profiles/blogs/roc-curve-explained-in-one-picture)

2.3.6 ดัชนีของโยเดน (Youden's Index)

เป็นดัชนีที่ใช้เพื่อสรุปประสิทธิภาพของการทดสอบ มักใช้ร่วมกับการวิเคราะห์ด้วยเส้นโค้ง ROC ซึ่งดัชนีจะถูกกำหนดไว้สำหรับทุกจุดของเส้นโค้งสามารถคำนวณได้ดังสมการที่ (11) และอาจใช้ค่าสูงสุดของดัชนีเป็นเกณฑ์ในการเลือกจุดตัดที่เหมาะสมที่สุด

$$J = \frac{TPs}{TPs + FNs} + \frac{TNs}{TNs + FPs} - 1 \qquad ...(11)$$



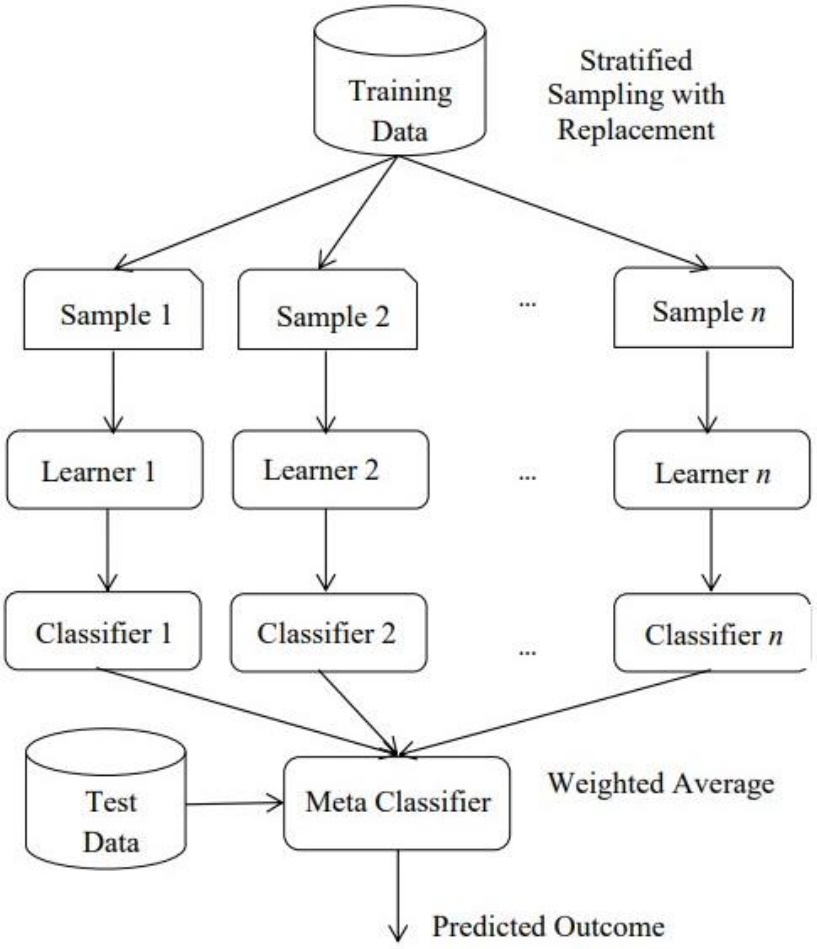
ภาพที่ 2.11 ดัชนีของโยเดน
(https://commons.wikimedia.org/wiki/File:ROC_Curve_Youden_J.png)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.4 การเรียนรู้เมตา

การเรียนรู้เมตาเป็นส่วนหนึ่งของวิธีการที่ใช้เพื่อพัฒนาและปรับปรุงประสิทธิภาพผลการพยากรณ์ของการเรียนรู้แบบรวมในเทคนิคสแต็กกิ้ง ซึ่งเป็นเทคนิคที่ช่วยลดความแปรปรวนของผลลัพธ์ โดยการสร้างชุดข้อมูลใหม่จากการสุ่มข้อมูลมาใช้แทนข้อมูลเดิมแบ่งออกเป็นชุด ๆ โดยใช้วิธีการสุ่มแบบใส่คืน เพื่อให้ชุดข้อมูลใหม่ที่สุ่มออกมานั้นมีลักษณะคล้ายกับชุดข้อมูลหลัก และนำแต่ละชุดข้อมูลไปเข้าแบบจำลองชนิดต่าง ๆ จากนั้นนำผลลัพธ์ที่ได้ในแต่ละแบบจำลองมาสร้างเป็นข้อมูลชุดใหม่ร่วมกับการถ่วงน้ำหนัก (Weighted Average) จากการวัดประสิทธิภาพของแบบจำลอง แล้วส่งต่อให้ตัวแบบเมตาเรียนรู้และทำนายเป็นผลลัพธ์สุดท้าย ดังภาพที่ 2.12 ซึ่งมีข้อดีคือการเรียนรู้ของเครื่องในแต่ละรูปแบบสามารถทำงานพร้อมกันได้แบบคู่ขนานเร็วและไม่ซับซ้อน

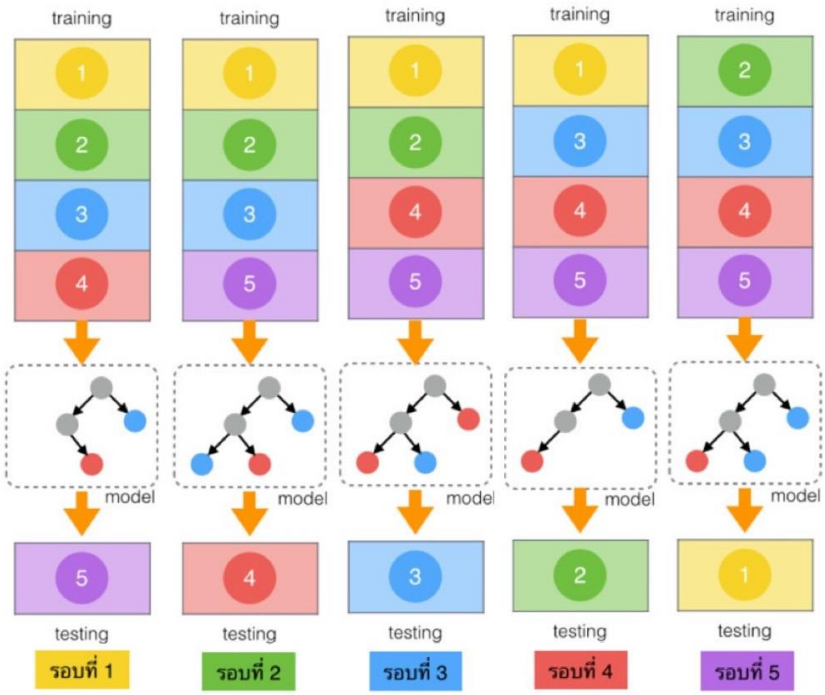
การใช้เทคนิคสแต็กกิ้งร่วมกับการแบ่งข้อมูลแบบตรวจสอบไขว้ (Cross Validation) ซึ่งเป็นเทคนิคของการสุ่มตัวอย่าง (Resampling) โดยจะแบ่งข้อมูลออกเป็นส่วน ๆ ส่วนละเท่า ๆ กัน และใช้ข้อมูลบางส่วนเพื่อสร้างตัวแบบและทดสอบประสิทธิภาพ



ภาพที่ 2.12 เทคนิคสแต็กกิ้ง (Sikora & Al-laymoun, 2014)

การแบ่งข้อมูลแบบ K-fold Cross Validation คือการแบ่งข้อมูลเป็น K ชุด ชุดละเท่า ๆ กัน ทำการทดสอบ K รอบ โดยในแต่ละรอบนั้นจะใช้ชุดข้อมูลจำนวน K-1 ชุดสำหรับการสร้างตัวแบบ หรือเรียกว่าชุดข้อมูลการฝึก (Training Set) และใช้ข้อมูลอีกหนึ่งชุดที่เหลือ เพื่อทดสอบประสิทธิภาพ หรือเรียกว่าชุดข้อมูลทดสอบ (Testing Set) ดังตัวอย่างในภาพที่ 2.13 เป็นตัวอย่างการแบ่งชุดข้อมูลเป็น 5 ส่วน หรือ 5-fold Cross Validation จากนั้นทำการทดสอบประสิทธิภาพของตัวแบบ 5 ครั้ง ดังนี้

รอบที่ 1 ใช้ข้อมูลส่วนที่ 1,2,3 และ 4 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 5
รอบที่ 2 ใช้ข้อมูลส่วนที่ 1,2,3 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 4
รอบที่ 3 ใช้ข้อมูลส่วนที่ 1,2,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 3
รอบที่ 4 ใช้ข้อมูลส่วนที่ 1,3,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 2
รอบที่ 5 ใช้ข้อมูลส่วนที่ 2,3,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 1



จุฬาลงกรณ์มหาวิทยาลัย

ภาพที่ 2.13 ตัวอย่าง 5-fold Cross Validation

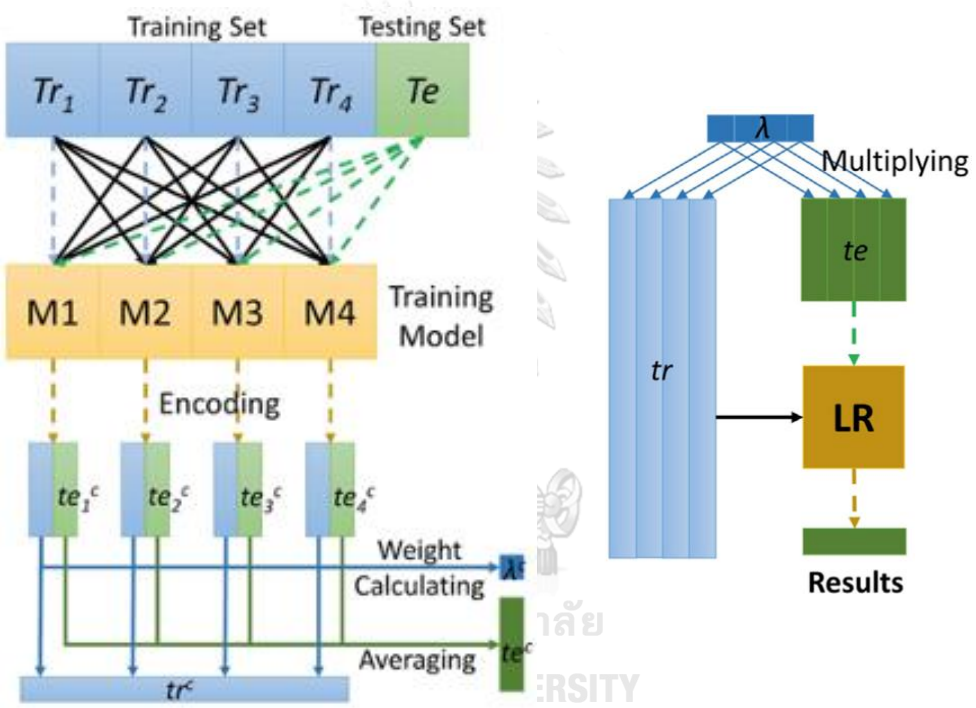
(https://dataminingtrend.com)

จากนั้นนำข้อมูลในแต่ละรอบไปใช้ในเทคนิคสแต็กกิ้งเพื่อให้ได้มาซึ่งผลลัพธ์สุดท้ายจากตัวแบบการเรียนรู้เมตา

2.5 การเรียนรู้เมตาโดยใช้ค่าถ่วงน้ำหนักแบบปรับ (Adaptive Weight)

(Ren et al., 2022) นำเสนอแนวทางการดำเนินการการเรียนรู้เมตาโดยใช้ตัววัดประสิทธิภาพในการคำนวณค่าถ่วงน้ำหนักแบบปรับ โดยเริ่มจากการแบ่งส่วนข้อมูลด้วย K-fold Cross Validation เช่นในกรณีที่ K=5 จะได้ชุดข้อมูลที่ถูกแบ่งเป็นห้าส่วนข้างต้นแล้วนั้น หนึ่งส่วนจะถูกเก็บไว้เพื่อใช้เป็นชุดทดสอบสุดท้ายเพื่อประเมินผลลัพธ์ที่ได้จะตัวแบบการเรียนรู้เมตา

ส่วนที่เหลืออีกสี่ส่วนจะถูกใช้ในการเรียนรู้ของหลากหลายตัวแบบ โดยจากสี่ส่วนนั้นจะถูกแบ่งหนึ่งส่วนให้เป็นข้อมูลสำหรับตรวจสอบ (Validation) และสามส่วนที่เหลือ ใช้สำหรับการเรียนรู้ของตัวแบบ ทำให้มีรอบการทำงานย่อยอีกสี่รอบ กล่าวคือ จากการแบ่งข้อมูลเป็น 5-fold Cross Validation ทำให้มีการทำงานห้ารอบ และในแต่ละรอบจะมีการทำงานอีกสี่รอบย่อย เมื่อทำซ้ำครบสี่รอบย่อยแล้ว สามารถคำนวณค่าเฉลี่ยของชุดข้อมูลทดสอบ ค่าเฉลี่ยการถ่วงน้ำหนัก ได้ดังสมการที่ (12) และ (13) เพื่อสร้างชุดข้อมูลการฝึกและชุดข้อมูลทดสอบใหม่จากการถ่วงน้ำหนักนำเข้าตัวแบบการเรียนรู้เมตาโดยใช้การถดถอยลอจิสติกเพื่อพยากรณ์ผลลัพธ์ขั้นสุดท้ายต่อไป ดังภาพที่ 2.14



ภาพที่ 2.14 การใช้เทคนิคสแต็กกิ้งร่วมกับการถ่วงน้ำหนักในหนึ่งรอบ (Ren et al., 2022)

ค่าเฉลี่ยของชุดข้อมูลทดสอบ (te^c)

$$te^c = \frac{1}{k}\sum_{i=1}^{k} te_i^c \qquad \text{...(12)}$$

เมื่อ k คือจำนวนของตัวแบบทั้งหมด

ค่าเฉลี่ยการถ่วงน้ำหนัก (λ^c)

$$\lambda^c = -\frac{1}{(1-\frac{1}{(\omega^c)^2})k} \quad \text{...(13)}$$

$$\omega^c = \frac{1}{k}\sum_{i=1}^{k}\omega_i^c \quad \text{...(14)}$$

เมื่อ ω_i^c คือค่า AUC ของการพยากรณ์ในแต่ละตัวแบบต่อชุดข้อมูลการฝึกรอบที่ i

ทั้งนี้ผู้วิจัยตั้งข้อสังเกตุว่าในการคำนวณค่าเฉลี่ยของชุดข้อมูลทดสอบตามสมการที่ (12) นั้นมีความไม่สมบูรณ์ เนื่องจากในการหาค่าเฉลี่ยชุดข้อมูลทดสอบ เกิดจากการนำชุดข้อมูลทดสอบในแต่ละรอบย่อยมารวมกัน การใช้จำนวนของรอบย่อยเป็นตัวหารเพื่อหาค่าเฉลี่ยของชุดข้อมูลทดสอบอาจจะให้ผลลัพธ์ที่มีความถูกต้องสมบูรณ์มากยิ่งขึ้นกว่าการหารด้วยจำนวนของตัวแบบทั้งหมด ในงานวิจัยนี้ จึงได้มีการนำเสนอวิธีการคำนวณค่าเฉลี่ยข้องชุดข้อมูลทดสอบตามข้อสังเกตดังกล่าวเพื่อเปรียบเทียบประสิทธิภาพอีกด้วย

เอกสารและงานวิจัยที่เกี่ยวข้อง

การคาดการณ์มีการพัฒนามาอย่างต่อเนื่องตั้งแต่อดีตจนปัจจุบัน และได้มีการพัฒนาตัวแบบการคาดการณ์ขึ้นมามากมายจากการศึกษาวิจัย สำหรับงานวิจัยนี้จะทำการศึกษาเกี่ยวกับการคาดการณ์ด้วยวิธีการจำแนกประเภทด้วยการเรียนรู้ของเครื่อง และการเรียนรู้แบบรวมการประยุกต์ใช้การเรียนรู้ของเครื่อง โดยได้ศึกษางานวิจัยที่เกี่ยวข้องดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1. Poverty Classification Using Machine Learning : The Case of Jordan

งานวิจัยของ (Alsharkawi et al., 2021) ศึกษาการคาดการณ์ความยากจนระดับครัวเรือนในประเทศจอร์แดน เพื่อจำแนกความยากจนที่สามารถคาดการณ์ข้อมูลประชากรในอนาคตได้ ในการศึกษานี้ใช้ข้อมูลสำรวจประชากรรายครัวเรือนในประเทศจอร์แดนที่มีรายได้แตกต่างกัน 5 กลุ่ม มีข้อมูลรวมทั้งสิ้น 63,211 ข้อมูล จาก 5 ปี คือ ปี 2002 จำนวน 10,558 ข้อมูล ปี 2006 จำนวน 11,494 ข้อมูล ปี 2008 จำนวน 10,956 ข้อมูล ปี 2010 จำนวน 10,987 ข้อมูล และปี 2017 จำนวน 19,216 ข้อมูล โดยมีค่าคุณลักษณะทั้งสิ้น 47 คุณลักษณะ แบ่งเป็น ข้อมูลเชิงปริมาณ (Quantative data) จำนวน 30 คุณลักษณะ และข้อมูลเชิงคุณภาพ (Qualitative data) จำนวน 17 คุณลักษณะ โดยมีคุณลักษณะสำคัญ อาทิ รายจ่ายประจำปีของสมาชิกในครัวเรือน, รายได้ประจำปี

ของสมาชิกในครัวเรือน, พื้นที่ที่อยู่อาศัย, ค่าเช่าบ้าน, ขนาดครัวเรือน,จำนวนห้อง, จำนวนเครื่องใช้ไฟฟ้า, จำนวนสิ่งอำนวยความสะดวกภายในบ้าน, ลักษณะกายภาพของหัวหน้าครัวเรือน และสถานการณ์เป็นครัวเรือนยากจน(ใช่หรือไม่ใช่) เป็นต้น ทีมวิจัยได้ใช้แบ่งชุดข้อมูลเพื่อทำการสอน (Train data) เป็นร้อยละ 90 ของข้อมูลทั้งหมด และแบ่งชุดทดสอบ (Test data) เป็นร้อยละ 10 ของข้อมูลทั้งหมด พร้อมกับใช้เทคนิคปรับช่วงขอบเขตของข้อมูลการทำให้เป็นมาตรฐานแซด (Z-Score standardization) สำหรับข้อมูลเชิงปริมาณ และเทคนิคการเข้ารหัสข้อมูล (One-hot encoding) สำหรับข้อมูลเชิงคุณภาพ นอกจากนี้ยังมีการใช้เทคนิควิธีการสุ่มเกิน (Random Oversampling), วิธีการสุ่มลด (Random undersampling), การจำแนกแบบถ่วงน้ำหนัก (Class weights), SMOTE เพื่อรับมือกับความไม่สมดุลของข้อมูล โดยได้ประยุกต์ใช้ร่วมกับการเรียนรู้ของเครื่องจำนวน 16 ตัวแบบ อาทิ การถดถอยลอจิสติก, ต้นไม้ตัดสินใจ, ป่าไม้แบบสุ่ม, การถดถอยริดจ์ (Ridge Regression), เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), ต้นไม้ตัดสินใจจำนวนมาก (Extra Tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), เบย์อย่างง่าย (Naive Bayes), เอดาบูท (AdaBoost), Bagged Decision Trees, Gradient Boosting Machine, Light GBM ใช้การประเมินประสิทธิภาพตัวแบบด้วยค่าความถูกต้องของ F1 (F1-Score) พบว่าตัวแบบ Light GBM และ Bagged Decision Trees มีค่าความถูกต้องสูงสุดสองอันดับแรกและให้ค่าความถูกต้องมากกว่าร้อยละ 80 โดยตัวแบบ Light BGM มีค่าคะแนนความถูกต้องของ F1 สูงสุดที่ร้อยละ 0.81 เมื่อทำงานร่วมกับเทคนิควิธีการสุ่มเกิน และวิธี Bagged Decision Trees มีค่าคะแนนความถูกต้องของ F1 สูงสุดที่ร้อยละ 0.80 เมื่อทำงานร่วมกับเทคนิคการจำแนกแบบถ่วงน้ำหนัก (การค้นหาแบบกริด (grid search)) ดังตารางที่ 2.1

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 2.1 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (Alsharkawi et al., 2021)

กระบวนการ	**ค่าคะแนนความถูกต้องของ F1 (F1-Score)**	
	LightGBM	**Bagged Decision Trees**
การจำแนกแบบถ่วงน้ำหนัก (สมดุล)	0.75	0.79
การจำแนกแบบถ่วงน้ำหนัก (การค้นหาแบบกริด)	0.80	0.80
วิธีการสุ่มเกิน	0.81	0.79
วิธีการสุ่มลด	0.80	0.79
SMOTE	0.80	0.79

2. Poverty Analysis Using Machine Learning Method

งานวิจัยของ (Pluliková, 2015) ศึกษาการสร้างแบบจำลองที่สามารถทำนายระดับความยากจนได้อย่างมีประสิทธิภาพ ศึกษาวิธีการประเมินความยากจนผ่านดัชนีความยากจน ในการศึกษานี้ใช้ข้อมูลสำรวจประชากรรายครัวเรือนในประเทศอินโดนีเซีย มีข้อมูลรวมทั้งสิ้น 6,388 ข้อมูล ซึ่งมีการสำรวจในปี 1993/1994, 1997/1998, 2000 และ 2007/2008 ครอบคลุมปัจจัยต่างๆ อาทิ สภาพเศรษฐกิจและข้อจำกัดของครัวเรือน, สภาพการศึกษาและสุขภาพรายบุคคล และข้อมูลชุมชน เป็นต้น โดยใช้ตัวแปรทั้งหมดเหล่านี้มาวิเคราะห์ร่วมกันให้ได้มาซึ่งระดับความยากจน ร่วมกับการประยุกต์ใช้การเรียนรู้ของเครื่อง โดยมีการแบ่งข้อมูลทั้งหมดออกเป็น 2 กลุ่ม ในกลุ่มที่ 1 ทำการวิเคราะห์โดยใช้รูปแบบการวิเคราะห์การสมนัยพหุคูณ (Multiple correspondence) และใช้เทคนิคการจัดกลุ่มข้อมูลแบบเคมีน (K-Means clusters respondents) ในการจัดกลุ่มระดับความยากจน ในกลุ่มที่ 2 จะใช้การเปรียบเทียบการทำนายของแต่ละตัวแบบ คือ การถดถอยลอจิสติก, โครงข่ายประสาทเทียม, ต้นไม้ตัดสินใจ และ ป่าไม้แบบสุ่ม โดยใช้การประเมินประสิทธิภาพตัวแบบด้วยค่าความไว, ค่าความจำเพาะ (Specificity) และค่าความแม่นยำ ซึ่งได้ประสิทธิภาพของผลลัพธ์การคาดการณ์ความยากจนดังตารางที่ 2.2 เมื่อเปรียบเทียบประสิทธิภาพในแต่ละตัวแบบแล้ว ผู้วิจัยสรุปได้ว่าตัวแบบโครงข่ายประสาทเทียม มีประสิทธิภาพในการคาดการณ์สูงที่สุด

ตารางที่ 2.2 ประสิทธิภาพผลลัพธ์การคาดการณ์ระดับความยากจนของ (Pluliková, 2015)

ตัวแบบ	ค่าความไว	ค่าความจำเพาะ	ค่าความแม่นยำ
การถดถอยลอจิสติก	0.16	0.97	0.85
โครงข่ายประสาทเทียม	0.24	0.95	0.84
ต้นไม้ตัดสินใจ	0.14	0.98	0.85
ป่าไม้แบบสุ่ม	0.18	0.97	0.85

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3. Machine Learning Approach for bottom 40 Percent Households (B40) Poverty Classification

งานวิจัยของ (S Sani et al., 2018) ศึกษาการสร้างแบบจำลองที่สามารถคาดการณ์ความยากจนในกลุ่มประชากรรายครัวเรือนที่มีรายได้ต่ำในประเทศมาเลเซียได้อย่างมีประสิทธิภาพ ในประเทศมาเลเซียนั้นได้มีการแบ่งกลุ่มประชากรตามความแตกต่างของรายได้เป็น 3 กลุ่มคือ รายได้สูงร้อยละ 20 (T20), รายได้ปานกลาง (M40), รายได้ต่ำ (B40) วิจัยนี้เป็นการศึกษาในกลุ่ม

ของประชากรรายครัวเรือนที่มีรายได้ต่ำหรือ B40 โดยใช้ข้อมูลจาก National Poverty Data Bank หรือ eKasih ซึ่งได้รับจากกรมสวัสดิการในส่วนการประสานการดำเนินงานของสำนักนายกรัฐมนตรี (ICU JPM) มีข้อมูลรวมทั้งสิ้น 99,546 ข้อมูล มีค่าคุณลักษณะทั้งสิ้น 16 คุณลักษณะ อาทิ พื้นที่อยู่อาศัย(ตัวเมืองหรือชนบท), ชาติพันธุ์, สถานการณ์แต่งงาน,อายุ, งาน, การศึกษา, สถานการณ์เป็นเจ้าของที่อยู่อาศัย, ขนาดของครัวเรือน ,รายได้ครัวเรือน และสถานะความยากจน โดยข้อมูลทั้งหมดเป็นข้อมูลของครัวเรือนที่อาศัยอยู่ใน 3 รัฐของมาเลเซีย คือ รัฐยะโฮร์, ตรังกานู และ ปะหัง เพื่อคาดการณ์ความจากจนรายครัวเรือนโดยประยุกต์ใช้ตัวแบบการเรียนรู้ของเครื่อง 3 ตัวแบบ คือ เบย์อย่างง่าย, ตัวแบบต้นไม้ตัดสินใจ และตัวแบบเพื่อนบ้านใกล้ที่สุด ร่วมกับเทคนิควิเคราะห์ความสัมพันธ์ (Correlation), หาคุณลักษณะที่สำคัญของข้อมูล (Feature Selection) และการค้นหาแบบกริด สำหรับปรับความเหมาะสมของพารามิเตอร์และเทคนิค 10-Fold Cross-Validation เพื่อตรวจสอบความถูกต้องด้วย โดยหลังจากใช้เทคนิคปรับความเหมาะสมนั้นทำให้เหลือคุณลักษณะ 8 คุณลักษณะสำหรับเข้าตัวแบบ งานวิจัยนี้ใช้การประเมินประสิทธิภาพตัวแบบด้วยค่าความแม่นยำ ซึ่งได้ประสิทธิภาพของผลลัพธ์การคาดการณ์ความยากจนดังตารางที่ 2.3 เมื่อเปรียบเทียบประสิทธิภาพในแต่ละตัวแบบแล้ว ผู้วิจัยสรุปได้ว่าตัวแบบต้นไม้ตัดสินใจมีประสิทธิภาพในการคาดการณ์สูงที่สุด

ตารางที่ 2.3 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (S Sani et al., 2018)

ตัวแบบ	ค่าความแม่นยำ
เบย์อย่างง่าย	97.27 %
ตัวแบบต้นไม้ตัดสินใจ	99.27 %
เพื่อนบ้านใกล้ที่สุด	96.80 %

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4. The predictive model of higher education guidance using integrated techniques for imbalanced data of learner group

งานวิจัยของ (Surawatchayotin & Paireekreng, 2021) ศึกษาการสร้างแบบจำลองที่สามารถทำนายในการเลือกศึกษาต่อในระดับอุดมศึกษา โดยการใช้เทคนิคแบบบูรณาการในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุลของกลุ่มผู้เรียน เพื่อสร้างตัวแบบการทำนายในการเลือกศึกษาต่อในระดับอุดมศึกษา โดยใช้ข้อมูลนักเรียน จาก 3 แผนการเรียนคือ ศิลป์คำนวณ, วิทย์คณิต และศิลป์ภาษา ตั้งแต่ปี 2560 – 2562 จากโรงเรียนเซนต์ฟรังซีสเซเวียร์ จำนวน 312 คน โดยใช้ตัวแบบต้นไม้ตัดสินใจ, ป่าไม้แบบสุ่ม, ต้นไม้ตัดสินใจจำนวนมาก, เอดาบูท, Gradient Boosting และ XGBoost ร่วมกับการเพิ่มประสิทธิภาพความแม่นยำของตัวแบบการทำนายด้วยเทคนิคของแบบจำลองการ

เรียนรู้แบบรวมด้วย 3 เทคนิค คือ แบ็กกิ้ง, บูทติ้ง และสแต็กกิ้ง และได้เปรียบเทียบประสิทธิภาพการคาดการณ์ของ 3 เทคนิคดังกล่าว ปรากฏว่า เทคนิค สแต็กกิ้ง โดยเทคนิคแบ็กกิ้ง งานวิจัยนี้ใช้การประเมินประสิทธิภาพตัวแบบด้วยค่าความแม่นยำ ซึ่งได้ประสิทธิภาพของผลลัพธ์การทำนายดังตารางที่ 2.4 เมื่อเปรียบเทียบประสิทธิภาพในแต่ละเทคนิคแล้ว ผู้วิจัยสรุปได้ว่าเทคนิคสแต็กกิ้งมีประสิทธิภาพในการทำนายสูงที่สุด

ตารางที่ 2.4 ประสิทธิภาพผลลัพธ์การคาดการณ์ความยากจนของ (Surawatchayotin & Paireekreng, 2021)

เทคนิคแบบจำลองการเรียนรู้แบบรวม	ค่าความแม่นยำ
แบ็กกิ้ง	75 %
บูทติ้ง	72 %
สแต็กกิ้ง	77 %

5. Small-sample precision of ROC-related estimates

งานวิจัยของ (Hanczar et al., 2010) ศึกษาความน่าเชื่อถือของการวัดประสิทธิภาพของการจำแนกประเภทของตัวแบบด้วยการใช้ เส้นโค้ง AUC-ROC ในกลุ่มตัวอย่างที่มีขนาดเล็ก โดยทำการทดสอบด้วยชุดข้อมูลสังเคราะห์ขนาด 50, 100, 200, 500 และ 1000 และชุดข้อมูลไมโครอาร์เรย์จริง พร้อมทั้งเปรียบเทียบกับการวัดประสิทธิภาพด้วย AUC ,อัตราผลบวกจริง และอัตราผลลบเทียม กับรากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสอง (Root Mean Square Error: RMSE) พบว่า การวัดประสิทธิภาพด้วยค่า AUC ,อัตราผลบวกจริง และอัตราผลลบเทียม นั้นมีค่าความผิดพลาดสูงกว่าการใช้ค่าเฉลี่ยความผิดพลาดกำลังสอง โดยเฉพาะอย่างยิ่งในการทดลองกับชุดข้อมูลที่มีจำนวนตัวอย่างน้อย การใช้ค่า AUC มีความไม่แม่นยำและเกิดความผิดพลาดได้สูง เนื่องด้วยในการคำนวณค่า AUC ขึ้นอยู่กับอัตราผลบวกจริงและอัตราผลลบจริงเป็นหลัก

ดังนั้นในงานวิจัยนี้ ซึ่งมีการคาดการณ์ผลลัพธ์แบบไบนารี่บนชุดข้อมูลกลุ่มตัวอย่างขนาดเล็ก ในขั้นตอนการปรับปรุงความสามารถของการคาดการณ์ด้วยการใช้การเรียนรู้เมตา โดยมีขั้นตอนในการถ่วงน้ำหนักด้วยการวัดประสิทธิภาพของตัวแบบนั้น จากเดิมงานวิจัยของ Ren et al. (2022) เสนอให้ใช้ค่า AUC ในการคำนวนค่าถ่วงน้ำหนัก ผู้วิจัยจึงเสนอการถ่วงน้ำหนักโดยใช้ค่าจากตัววัดประสิทธิภาพ F1-Score ร่วมด้วย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 3
วิธีการดำเนินการศึกษา

3.1 ขอบเขตของการวิจัย

3.1.1 การคาดการณ์

การคาดการณ์ความยากจนโดยการจำแนกประเภทครัวเรือนเป็น 2 ประเภท คือ ครัวเรือนยากจนและครัวเรือนปกติ

3.1.2 ขอบเขตของข้อมูล

ข้อมูลทุติยภูมิจากการลงพื้นที่สำรวจประชากรรายครัวเรือนในจังหวัดสระบุรี ปี 2565 จากสำนักงานสภาพัฒนาการณ์เศรษฐกิจและสังคมแห่งชาติ โดยเป็นข้อมูลระดับครัวเรือน จำนวน 352 ครัวเรือน ครอบคลุมทุกอำเภอในจังหวัดสระบุรี มีค่าคุณลักษณะทั้งสิ้น 136 คุณลักษณะ แบ่งเป็นข้อมูลเชิงคุณภาพ 24 คุณลักษณะ และข้อมูลเชิงปริมาณ 112 คุณลักษณะ รายละเอียดตามภาคผนวก ตารางที่ 1 แบ่งเป็นครัวเรือนยากจน จำนวน 80 ครัวเรือน และครัวเรือนปกติ จำนวน 272 ครัวเรือน โดยมีอัตราส่วนความไม่สมดุลของข้อมูล (Imbalance ratio) ที่ 3.4 ซึ่งค่าคุณลักษณะที่แสดงสถานะของครัวเรือนยากจนนั้นได้จากการประเมินเกณฑ์ดัชนีความยากจนหลายมิติ (MPI) จากข้อมูลในระบบบริหารจัดการข้อมูลการพัฒนาคนแบบชี้เป้า (TPMAP) ซึ่งเป็นการประเมินข้อมูลความจำเป็นพื้นฐาน (จปฐ.) ของกระทรวงมหาดไทยร่วมกับข้อมูลผู้ลงทะเบียนสวัสดิการแห่งรัฐของกระทรวงการคลัง จึงเป็นข้อมูลที่มีคุณลักษณะที่แตกต่างกันกับชุดข้อมูลที่ใช้ในงานวิจัย ครั้งนี้ กล่าวคือ ใช้ค่าคุณลักษณะจากชุดข้อมูลสำรวจประชากรรายครัวเรือนเป็นตัวแปรต้น และใช้สถานะของครัวเรือนยากจนจากระบบ TPMAP เป็นตัวแปรตามในการทดลอง

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.1.3 ตัวแบบในการคาดการณ์ที่ใช้ในการวิจัย

ในงานวิจัยนี้ใช้ตัวแบบสำหรับการคาดการณ์รวม 6 ตัวแบบ ทำการหาพารามิเตอร์ที่ดีที่สุดของตัวแบบด้วยการใช้ RandomizedSearchCV ภายในไลบารี sklearn ด้วยข้อมูลทั้งหมด และใช้ค่าพารามิเตอร์ที่ดีที่สุดนี้ ในการสร้างตัวแบบเพื่อทำนายผลและวัดประสิทธิภาพเพื่อนำไปคำนวณค่าถวงน้ำหนัก และสร้างตัวแบบเมตาในขั้นตอนสุดท้าย โดยขอบเขตของการหาพารามิเตอร์ที่ดีที่สุดของแต่ละตัวแบบ มีรายละเอียดดังนี้

1. ตัวแบบการถดถอยลอจิสติกกำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.1

ตารางที่ 3.1 พารามิเตอร์สำหรับสร้างตัวแบบการถดถอยลอจิสติก

ลำดับ	พารามิเตอร์	ขอบเขต
1	solver	[“newton-cg” , “lbfgs”, “liblinear”, ”sag”, ”saga”]
2	penalty	[“none”, “l1”, “l2”, “elasticnet”]
3	c	[100, 10, 1.0, 0.1, 0.01]

2. ตัวแบบต้นไม้ตัดสินใจ กำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.2

ตารางที่ 3.2 พารามิเตอร์สำหรับสร้างตัวแบบต้นไม้ตัดสินใจ

ลำดับ	พารามิเตอร์	ขอบเขต
1	random_state	[“instance” , “None”]
2	criterion	[“gini”, “entropy”]
3	max_features	[“auto”, “sqrt”, “log2”]
4	min_samples_split	[2,5]
5	min_samples_leaf	[1,4]
6	splitter	[“best”, “random”]

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3. ตัวแบบ Bagging Tree กำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.3

ตารางที่ 3.3 พารามิเตอร์สำหรับสร้างตัวแบบ Bagging Tree

ลำดับ	พารามิเตอร์	ขอบเขต
1	n_estimators	[0, 500, step=1]
2	max_features	[“auto”, “sqrt”]
3	max_depth	[10, 110, step=10]
4	min_samples_split	[2,5]
5	min_samples_leaf	[1,4]
6	bootstrap	[“True”, “False”]

4. ตัวแบบป่าไม้แบบสุ่ม กำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.4

ตารางที่ 3.4 พารามิเตอร์สำหรับสร้างตัวแบบ Random Forest

ลำดับ	พารามิเตอร์	ขอบเขต
1	n_estimators	[0, 500, step=1]
2	max_features	[“auto”, “sqrt”]
3	max_depth	[10, 110, step=10]
4	min_samples_split	[2,5]
5	min_samples_leaf	[1,4]
6	bootstrap	[“True”, “False”]

5. ตัวแบบ Light GBM กำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.5

ตารางที่ 3.5 พารามิเตอร์สำหรับสร้างตัวแบบ Light GBM

ลำดับ	ชื่อพารามิเตอร์	ขอบเขต
1	boosting_type	gbdt, dart
2	learning_rate	[0.01, 1.00, step=0.05]
3	n_estimators	[0, 500, step=1]
4	subsample	[0.1, 1.0, step=0.1]
5	min_samples_split	[2,5]
6	min_samples_leaf	[1,5]
7	max_depth	[2,5]

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6. ตัวแบบโครงข่ายประสาทเทียม กำหนดขอบเขตของพารามิเตอร์สำหรับสร้างตัวแบบ ดังตารางที่ 3.6

ตารางที่ 3.6 พารามิเตอร์สำหรับสร้างตัวแบบโครงข่ายประสาทเทียม

ลำดับ	ชื่อพารามิเตอร์	ขอบเขต
1	hidden_layer_sizes	[16,4], [32,4], [64,4], [128,4], [256,4]
2	activation	[‘logistic’, ‘tanh’, ‘relu’]
3	solver	[‘lbfgs’, ‘sgd’]
4	learning_rate_int	[0.01, 1.00, step=0.05]
5	max_iter	[50, 500, step=50]

3.1.4 การปรับปรุงความสามารถในการคาดการณ์

ใช้แบบจำลองการเรียนรู้เมตาในการปรับปรุงความสามารถในการคาดการณ์ โดยใช้ตัวแบบการถดถอยลอจิสติกในการพยากรณ์ขั้นสุดท้ายร่วมกับเทคนิค 5-fold cross validation ด้วยตัวแบบการเรียนรู้ของเครื่องหลากหลายชนิดข้างต้น

3.1.5 เครื่องมือ

ใช้โปรแกรมจูปิเตอร์โน้ตบุค (Jupyter Notebook) ในการเขียนโปรแกรมเพื่อจัดการข้อมูล รวมถึงการสร้างแบบจำลอง และใช้ภาษาคอมพิวเตอร์ไพธอน (Python)

3.2 วิธีการดำเนินการวิจัย

1. ศึกษา ค้นคว้าเอกสารงานวิจัย ทฤษฎีและกรอบแนวคิดที่เกี่ยวข้อง
2. ศึกษาและทำความเข้าใจคุณลักษณะของข้อมูลที่จะนำเข้าแบบจำลอง
3. เตรียมข้อมูล (Data Preprocess)

3.1 เติมข้อมูลว่าง (null) และกำจัดคุณลักษณะที่ไม่จำเป็น โดยข้อมูลเชิงปริมาณทำการเติมเต็มข้อมูลว่างด้วย “0” และข้อมูลเชิงคุณภาพทำการเติมเต็มข้อมูลว่างด้วยค่าฐานนิยม (Mode) และละทิ้งคุณลักษณะที่เป็นข้อมูลว่างในทุกแถว

3.2 การเข้ารหัสข้อมูลแบบสร้างตัวแปรหุ่น (Dummy Variable Encoding) สำหรับข้อมูลเชิงคุณภาพ คือการแปลงข้อมูลเชิงคุณภาพให้อยู่ในรูปของตัวเลขโดยการเพิ่มหลัก (Column) ของข้อมูล ซึ่งจะมีข้อมูลที่เพิ่มมา n-1 หลัก เมื่อ n คือจำนวนของค่าที่แตกต่างกันในหลักที่ต้องการเข้ารหัส

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.3 ปรับช่วงขอบเขตของข้อมูล (Feature Scaling) สำหรับข้อมูลเชิงปริมาณ ให้อยู่ในช่วงเดียวกัน โดยใช้เทคนิคการทำให้เป็นปกติน้อยที่สุด-มากที่สุด (Min-Max Normalization) ซึ่งเป็นการเตรียมข้อมูลให้อยู่ในรูปมาตรฐานเดียวกัน คือมีค่า 0 ถึง 1 ดังสมการที่ (15)

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{...(15)}$$

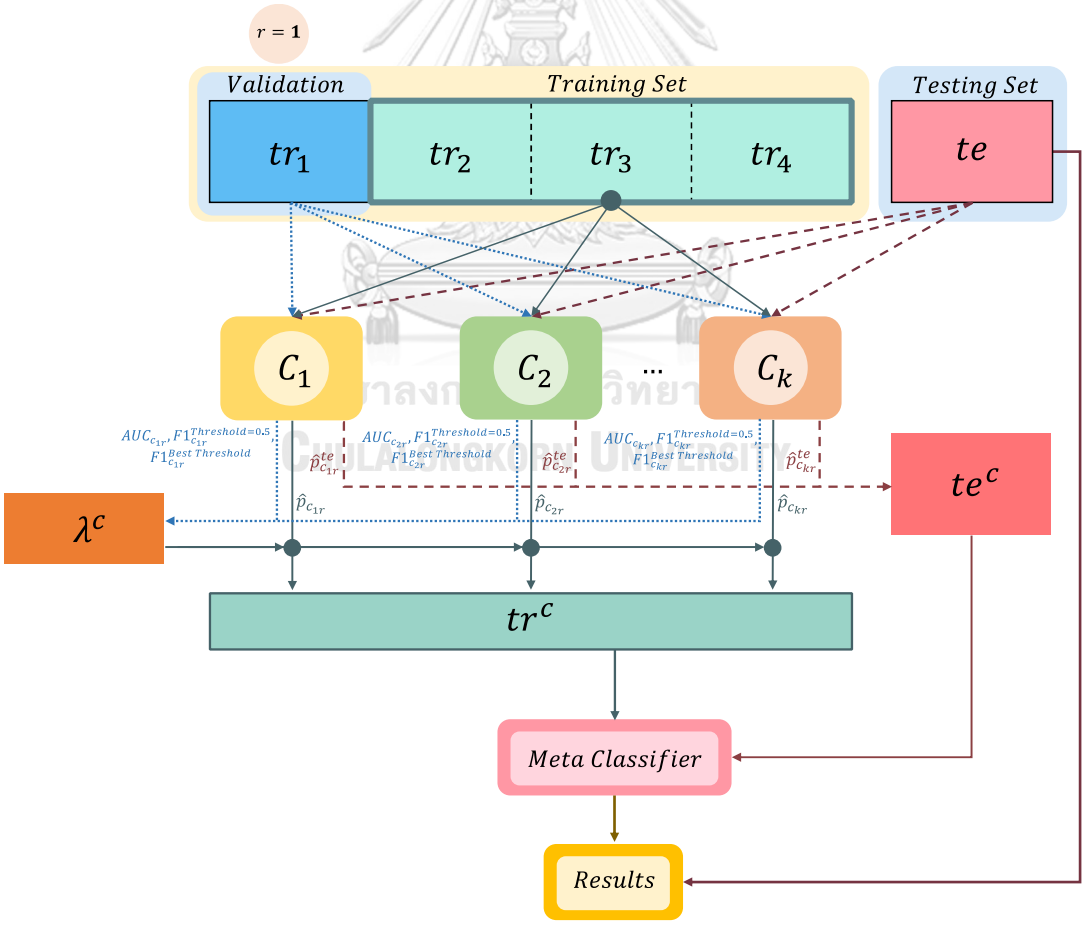
โดย X' คือ ค่าที่ทำให้เป็นปกติมาตรฐานของตัวแปร X

X_{max} คือ ค่าสูงสุดของตัวแปร X

X_{min} คือ ค่าต่ำสุดของตัวแปร X

4. หาค่าคุณลักษณะที่สำคัญของข้อมูล (Feature Importance) ด้วยการตรวจสอบ สำรวจข้อมูลเบื้องต้น (Exploratory Data Analysis : EDA) ร่วมกับการใช้ตัวแบบต้นไม้ในการตัดสินใจ

5. แบ่งข้อมูลทั้งหมดออกเป็น 5 ส่วน ส่วนละเท่า ๆ กันโดยใช้เทคนิค 5-fold Cross Validation แบ่งเป็นชุดข้อมูลการฝึก (tr) 4 ส่วน และชุดข้อมูลทดสอบ (te) 1 ส่วน ในรอบที่ 1 (Fold 1) รอบย่อยที่ 1 $(r = 1)$ ใช้ข้อมูลส่วนที่ 1 (tr_1) เป็นข้อมูลตรวจสอบไขว้ ใช้ข้อมูลส่วนที่ 2,3 และ 4 เป็นชุดข้อมูลการฝึก เพื่อสร้างตัวแบบ (C_k) จำนวน k แบบ และใช้ชุดข้อมูลทดสอบสำหรับทดสอบตัวแบบ ทำซ้ำเพื่อสลับชุดข้อมูลตรวจสอบไขว้จนครบสี่รอบ และทำการประเมินประสิทธิภาพตัวแบบจากชุดข้อมูลการฝึกด้วยค่า AUC และ F1-Score เพื่อหาค่าเฉลี่ยการถ่วงน้ำหนัก (λ^c) ของแต่ละตัวแบบตามสมการที่ (13) และ (14) เมื่อให้ $\hat{p}_{c_{1r}}$ คือค่าความน่าจะเป็นที่ได้จากการพยากรณ์ด้วยชุดข้อมูลตรวจสอบของตัวแบบที่ 1 (C_1) ในรอบย่อยที่ 1 และ $\hat{p}^{te}_{c_{1r}}$ คือ ค่าความน่าจะเป็นที่ได้จากการพยากรณ์ด้วยชุดข้อมูลทดสอบของตัวแบบที่ 1 ได้มาซึ่งชุดข้อมูลใหม่สำหรับนำเข้าตัวแบบเมตาดังภาพที่ 3.1



ภาพที่ 3.1 ขั้นตอนการทำงานในรอบที่ 1

6. ดำเนินการข้อที่ 5 จนครบ 4 รอบย่อยและหาค่าเฉลี่ยชุดข้อมูลทดสอบ (te^c) ตามสมการที่ (12)

7. สร้างตัวแบบการเรียนรู้เมตาโดยใช้ชุดข้อมูลทดสอบจากข้อที่ 5 และ 6 ในการคาดการณ์ผลลัพธ์และค่าความน่าจะเป็น รวมถึงการประเมินประสิทธิภาพตัวแบบ ซึ่งในการสร้างตัวแบบการเรียนรู้แบบเมตานั้น จะแบ่งเป็น 3 กรณีเพื่อทำการเปรียบเทียบประสิทธิภาพ ดังนี้

กรณีที่ 1 ตัวแบบการถดถอยลอจิสติก โดยกำหนดให้เวกเตอร์ของสัมประสิทธิ์การถดถอยเป็นค่าเฉลี่ยการถ่วงน้ำหนักและค่าเฉลี่ยของความน่าจะเป็น ดังสมการที่ (16) เมื่อให้ $\hat{\boldsymbol{p}}_{c_1}$ คือค่าความน่าจะเป็นของตัวแบบที่ 1

$$logit(\hat{\pi}_{tr}) = \beta_0 + \beta_1\lambda^c\hat{p}_{c_1} + \beta_2\lambda^c\hat{p}_{c_2} + \ldots + \beta_k\lambda^c\hat{p}_{c_k} \qquad \ldots(16)$$

กรณีที่ 2 จากกรณีที่ 1 ในการคำนวณค่าเฉลี่ยการถ่วงน้ำหนัก (λ^c) ของแต่ละตัวแบบตามสมการที่ (13) และ (14) จะใช้ค่า F1-Score ที่จุดตัด (Threshold) 0.5 แทนค่า AUC ในการคำนวณโดยแทน ω_i^c ในสมการที่ (14) ด้วยค่า F1-Score ในการพยากรณ์ของแต่ละตัวแบบต่อชุดข้อมูลการฝึกรอบที่ i โดยใช้ค่าจุดตัด (Threshold) ที่ 0.5

กรณีที่ 3 แทน ω_i^c ในสมการที่ (14) ด้วยค่า F1-Score ในการพยากรณ์ของแต่ละตัวแบบต่อชุดข้อมูลการฝึกรอบที่ i โดยใช้ค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีของโยเดนที่สูงสุด

8. นำผลการพยากรณ์ที่ได้จากตัวแบบเมตาที่ใช้ค่าถ่วงน้ำหนักแบบปรับทั้ง 3 วิธีมาคำนวณค่า AUC และ F1-Score โดยอ้างอิงจากชุดข้อมูลทดสอบ

9. ดำเนินการข้อที่ 5,6,7 และ 8 ซ้ำจนกว่าจะครบ 5 รอบ ดังนี้

   รอบที่ 2 ใช้ข้อมูลส่วนที่ 1,2,3 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 4

   รอบที่ 3 ใช้ข้อมูลส่วนที่ 1,2,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 3

   รอบที่ 4 ใช้ข้อมูลส่วนที่ 1,3,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 2

   รอบที่ 5 ใช้ข้อมูลส่วนที่ 2,3,4 และ 5 สร้างตัวแบบและใช้ตัวแบบทำนายข้อมูลส่วนที่ 1

10. สร้างเมทริกซ์การประเมินจากผลในข้อที่ 9 เพื่อวิเคราะห์และเปรียบเทียบ

11. ใช้ข้อมูลอีกหนึ่งชุดทำการทดลอง กล่าวคือหลังจากดำเนินการในข้อ 3.3 แล้ว ทำการจัดการกับข้อมูลที่ไม่สมดุล(Imbalance Data) ด้วยเทคนิค SMOTE และดำเนินการข้อที่ 4-10 ซ้ำ

12. วิเคราะห์และเปรียบเทียบผลที่ได้จากชุดข้อมูลชุดแรกและชุดข้อมูลที่ใช้เทคนิค SMOTE

13. เนื่องจากผู้วิจัยตั้งข้อสังเกตุว่า งานวิจัยของ (Ren et al., 2022) ในการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบตามสมการที่ (12) นั้นมีความไม่สมบูรณ์ จึงเสนอให้มีการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบอีกวิธีหนึ่งตามสมการที่ (17) กล่าวคือ ดำเนินการข้อที่ 5-12 ซ้ำ โดยในข้อที่ 6 เปลี่ยนวิธีการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ (te^c) ตามสมการที่ (17)

$$te^c = \frac{1}{r}\sum_{i=1}^{k} te_i^c \qquad \text{...(17)}$$

เมื่อ r คือจำนวนของรอบย่อยในแต่ละ fold โดยในงานวิจัยนี้มีค่าเท่ากับ 4

14. วิเคราะห์และเปรียบเทียบผลที่ได้จากการใช้วิธีการคำนวณค่าเฉลี่ยชุดทดสอบที่แตกต่างกันในข้อ 5 และ 13

15. สรุปผลการดำเนินงานวิจัย

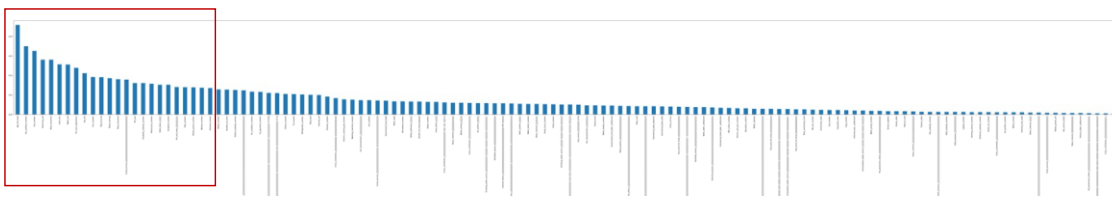
จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 4
ผลการวิจัย

ปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือน

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือน เพื่อนำไปเป็นชุดข้อมูลสำหรับคาดการณ์ความยากจน โดยใช้ข้อมูลทุติยภูมิจากการลงพื้นที่สำรวจประชากรรายครัวเรือนในจังหวัดสระบุรี ปี 2565 โดยเป็นข้อมูลระดับครัวเรือน จำนวน 352 ครัวเรือน ครอบคลุมทุกอำเภอในจังหวัดสระบุรี มีค่าคุณลักษณะทั้งสิ้น 136 คุณลักษณะ แบ่งเป็นข้อมูลเชิงคุณภาพ 24 คุณลักษณะ และข้อมูลเชิงปริมาณ 112 คุณลักษณะ แบ่งเป็นครัวเรือนยากจน จำนวน 80 ครัวเรือน และครัวเรือนปกติ จำนวน 272 ครัวเรือน โดยหลังจากขั้นตอนการเตรียมข้อมูลในขั้นตอนการเติมข้อมูลว่างนั้น ทำการละทิ้งคุณลักษณะที่เป็นข้อมูลว่างในทุกแถว จำนวน 7 คุณลักษณะ ได้แก่ จำนวนหนี้สินคงค้างร้านค้าอุปโภค บริโภคและปัจจัยการผลิต, จำนวนเงินในการแบ่งชำระคืนหนี้สินคงค้างร้านค้าอุปโภค บริโภคและปัจจัยการผลิต ต่อเดือน, จำนวนรถบัส, จำนวนควาย, จำนวนแกะ, จำนวนม้า และจำนวนผู้ที่ได้รับเงินช่วยเหลือจากกองทุนคุ้มครองเด็ก จากนั้นในข้อตอนการเข้ารหัสข้อมูลแบบสร้างตัวแปรหุ่นสำหรับข้อมูลเชิงคุณภาพ ทำให้มีค่าคุณลักษณะเพิ่มขึ้น รวมทั้งสิน 172 คุณลักษณะ

การศึกษาครั้งนี้ทำการหาค่าคุณลักษณะที่สำคัญของข้อมูล โดยการใช้ตัวแบบต้นไม้ตัดสินใจ ได้ค่าคะแนนความสำคัญของคุณลักษณะทั้งหมดดังภาพที่ 4.1 โดยผู้วิจัยทำการเลือกคุณลักษณะที่สำคัญภายในสี่เหลี่ยมสีแดงดังภาพ เนื่องจากเป็นกลุ่มที่มีค่าคะแนนสูงที่สุด โดยมีจำนวน 24 คุณลักษณะ โดยค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจน 24 อันดับแรกนั้นมีน้ำหนักรวมมากกว่าร้อยละ 50 จากค่าความสำคัญของคุณลักษณะทั้งหมด (จากผลรวมของค่าความสำคัญของคุณลักษณะทั้งหมดเป็น 1) รายละเอียดดังตารางที่ 4.1



ภาพที่ 4.1 กราฟเปรียบเทียบค่าคะแนนคุณลักษณะที่สำคัญ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 4.1 แสดงค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจนสูงสุด 24 อันดับแรก

คุณลักษณะ	ค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจน
อายุของหัวหน้าครัวเรือน	0.045912
จำนวนผู้ที่ได้รับบัตรสวัสดิการแห่งรัฐในครัวเรือน	0.034988
จำนวนพัดลม	0.032591
ค่าไฟ	0.028061
ค่าใช้จ่ายเพื่อการบริโภคในครัวเรือน	0.028056
พื้นที่รวมที่ดินที่ถือครอง	0.025650
ค่าน้ำ	0.025611
จำนวนสมาชิกทั้งหมดของครัวเรือน	0.023925
ค่าเชื้อเพลิง	0.021095
จำนวนเตารีด	0.019140
รายได้เฉลี่ยต่อเดือน	0.019123
เงินออมรวมทั้งครัวเรือน	0.018556
ค่าใช้จ่ายเพื่อการอุปโภค	0.018044
ความสามารถสามารถหาซื้อและมีกำลังทรัพย์เพียงพอในการซื้ออุปกรณ์ที่จำเป็นในการป้องกันตนเองจากโรคระบาด	0.017853
จำนวนสมาชิกทั้งหมดของครัวเรือน	0.016066
จำนวนผู้ที่ได้รับเบี้ยคนพิการ	0.016061
จำนวนจักรยานยนต์	0.015752
การมีหนีสินของครัวเรือน_ไม่มี	0.015168
จำนวนสมาชิกครัวเรือนที่พิการ	0.015150
เขตการปกครองของที่อยู่อาศัย_ในเขตเทศบาล	0.014037
จำนวนสมาชิกในครัวเรือนเพศชาย	0.013981
จำนวนโทรศัพท์มือถือในครัวเรือน	0.013888
จำนวนสมาชิกในครัวเรือนเพศหญิง	0.013699
จำนวนสมาชิกที่ใช้อินเทอร์เน็ต	0.013507

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประสิทธิภาพในการทำนายของวิธีคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนวิธีต่าง ๆ

จากวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการทำนายของวิธีคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนวิธีต่างๆ นั้น การศึกษาครั้งนี้ได้ทำการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนายใน 3 กรณี ได้แก่ 1. การใช้ค่า AUC 2. การใช้ค่า F1-Score โดยพิจารณาจุดตัด 0.5 และ 3. การใช้ค่า F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ร่วมกับการเปรียบเทียบข้อมูลจาก 2 ชุดข้อมูล คือ 1.ชุดข้อมูลสำรวจประชากรรายครัวเรือนในระดับพื้นที่ (ชุดข้อมูลก่อนใช้เทคนิค SMOTE) และ 2.ชุดข้อมูลสำรวจประชากรรายครัวเรือนในระดับพื้นที่ข้างต้นที่มีการจัดการกับข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE (ชุดข้อมูลหลังใช้เทคนิค SMOTE) อีกทั้งผู้วิจัยได้เสนอวิธีในการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ (te^C) 2 วิธี คือ 1.หาค่าเฉลี่ยด้วยจำนวนของตัวแบบทั้งหมด ดังวิธีการดำเนินการวิจัย ข้อที่ 6 และหาค่าเฉลี่ยด้วยจำนวนของรอบย่อยในแต่ละรอบ ดังวิธีการดำเนินการวิจัย ข้อที่ 13

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบโดยแบ่งเป็น 4 ส่วน ได้แก่

ส่วนที่ 1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ส่วนที่ 2 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ส่วนที่ 3 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ส่วนที่ 4 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

4.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่าง ๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ตารางที่ 4.2 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.970	0.895	1.000
	2	0.882	0.750	0.833
	3	0.803	0.435	0.909
	4	0.875	0.562	0.818
	5	0.878	0.636	0.933
	ค่าเฉลี่ย	**0.881**	**0.656**	**0.898**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.059**	**0.176**	**0.075**

ตารางที่ 4.3 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.940	0.933	0.824
	2	0.869	0.737	0.778
	3	0.817	0.455	0.909
	4	0.917	0.692	0.818
	5	0.905	0.722	0.867
	ค่าเฉลี่ย	**0.890**	**0.708**	**0.839**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.048**	**0.170**	**0.050**

ตารางที่ 4.4 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ย

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.970	0.895	1.000
	2	0.882	0.727	0.889
	3	0.789	0.417	0.909
	4	0.847	0.500	0.909
	5	0.851	0.583	0.933
	ค่าเฉลี่ย	**0.868**	**0.624**	**0.928**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.066**	**0.190**	**0.043**

4.2 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ตารางที่ 4.5 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.896	0.696	1.000
	2	0.838	0.560	1.000
	3	0.704	0.308	0.727
	4	0.889	0.696	0.941
	5	0.851	0.545	0.923
	ค่าเฉลี่ย	**0.835**	**0.561**	**0.918**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.078**	**0.159**	**0.112**

ตารางที่ 4.6 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.910	0.812	0.812
	2	0.838	0.571	0.857
	3	0.718	0.420	0.727
	4	0.930	0.778	0.824
	5	0.892	0.632	0.923
	ค่าเฉลี่ย	**0.858**	**0.643**	**0.829**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.085**	**0.160**	**0.071**

ตารางที่ 4.7 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	1	0.896	0.696	1.000
	2	0.809	0.519	1.000
	3	0.654	0.268	0.727
	4	0.875	0.654	1.000
	5	0.797	0.462	0.923
	ค่าเฉลี่ย	**0.806**	**0.520**	**0.930**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.095**	**0.170**	**0.118**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.3 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ตารางที่ 4.8 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.791	0.696	0.889
	2	0.805	0.702	0.917
	3	0.742	0.617	0.853
	4	0.756	0.604	0.906
	5	0.898	0.914	0.972
	ค่าเฉลี่ย	**0.798**	**0.706**	**0.907**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.061**	**0.124**	**0.043**

ตารางที่ 4.9 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.721	0.604	0.914
	2	0.701	0.579	0.943
	3	0.652	0.509	0.906
	4	0.656	0.518	0.879
	5	0.906	0.908	0.972
	ค่าเฉลี่ย	**0.727**	**0.624**	**0.923**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.104**	**0.164**	**0.036**

ตารางที่ 4.10 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.709	0.604	0.889
	2	0.690	0.579	0.917
	3	0.607	0.492	0.853
	4	0.600	0.492	0.829
	5	0.906	0.908	0.972
	ค่าเฉลี่ย	**0.702**	**0.615**	**0.892**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.124**	**0.171**	**0.056**

4.4 ผลการเปรียบเทียบประสิทธิภาพในการทำนายที่ได้จากการใช้ตัววัดประสิทธิภาพแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ตารางที่ 4.11 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ AUC ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.698	0.593	0.889
	2	0.713	0.623	0.868
	3	0.629	0.492	0.906
	4	0.688	0.580	0.879
	5	0.891	0.874	0.993
	ค่าเฉลี่ย	**0.724**	**0.632**	**0.907**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.099**	**0.144**	**0.050**

ตารางที่ 4.12 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับ ด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.605	0.533	0.842
	2	0.644	0.541	0.917
	3	0.573	0.468	0.853
	4	0.533	0.453	0.806
	5	0.865	0.863	0.972
	ค่าเฉลี่ย	**0.644**	**0.572**	**0.878**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.130**	**0.167**	**0.066**

ตารางที่ 4.13 แสดงประสิทธิภาพในการทำนาย ที่ได้จากการใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในการคำนวณค่าถ่วงน้ำหนักแบบปรับด้วยชุดข้อมูลหลังใช้เทคนิค SMOTE และคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ

ชุดข้อมูล	รอบ	ค่าความแม่นยำ	ค่าความเที่ยงตรง	ค่าการเรียกคืน
ชุดข้อมูลหลังใช้เทคนิค SMOTE	1	0.570	0.492	0.889
	2	0.632	0.532	0.917
	3	0.562	0.453	0.879
	4	0.500	0.42	0.853
	5	0.865	0.869	0.965
	ค่าเฉลี่ย	**0.629**	**0.553**	**0.901**
	ส่วนเบี่ยงเบนมาตรฐาน	**0.143**	**0.181**	**0.043**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ทำการศึกษาปัจจัยที่มีความสัมพันธ์กับความยากจนในระดับครัวเรือนจากข้อมูลทุติยภูมิจากการลงพื้นที่สำรวจประชากรรายครัวเรือน ที่มีค่าคุณลักษณะทั้งสิ้น 136 คุณลักษณะ โดยเสนอวิธีการหาค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจนด้วยการใช้ตัวแบบต้นไม้ตัดสินใจ พร้อมทั้งศึกษาและเปรียบเทียบประสิทธิภาพในการทำนายของวิธีคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนวิธีต่าง ๆ จากการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนายใน 3 กรณี ได้แก่ 1. การใช้ค่า AUC 2. การใช้ค่า F1-Score โดยพิจารณาจุดตัด 0.5 และ 3. การใช้ค่า F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ร่วมกับการเปรียบเทียบจากชุดข้อมูลสำรวจประชากรรายครัวเรือนในระดับพื้นที่ และชุดข้อมูลชุดหลังที่มีการจัดการกับข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE พร้อมทั้งเสนอวิธีการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบใน 2 วิธี ได้แก่ 1.คำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด และ 2.คำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ โดยมีการสรุปผลการวิจัยดังนี้

5.1 สรุปผลการวิจัย

จากตารางที่ 4.1 เมื่อทำการหาค่าความสำคัญของคุณลักษณะที่ส่งผลต่อความยากจนด้วยการใช้ตัวแบบต้นไม้ตัดสินใจพบว่า ปัจจัยที่มีค่าความสำคัญสูงสุด คือ อายุของหัวหน้าครัวเรือน ที่มีค่าความสำคัญอยู่ที่ 0.046 และปัจจัยที่มีค่าความสำคัญสูงกว่า 0.0135 มีจำนวน 24 คุณลักษณะ ซึ่งมีค่าความสำคัญรวมกันมากกว่าร้อยละ 50 จากค่าความสำคัญของคุณลักษณะทั้งหมดรวมกัน

การเปรียบเทียบประสิทธิภาพของวิธีการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนโดยใช้ชุดข้อมูลชุดข้อมูลก่อนใช้เทคนิค SMOTE และชุดข้อมูลหลังใช้เทคนิค SMOTE ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมดปรากฏดังตารางต่อไปนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 5.1 แสดงค่าความแม่นยำเฉลี่ยจากการทดลองของสองชุดข้อมูล และการใช้ตัววัดประสิทธิภาพการทำนายแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด

ชุดข้อมูล	ประสิทธิภาพในการทำนายโดยเฉลี่ย	ตัววัดประสิทธิภาพที่ใช้ในการคำนวณค่าถ่วงน้ำหนัก		
		AUC	F1	Optimal F1
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	ค่าความแม่นยำ	0.835	**0.858**	0.806
	ค่าความเที่ยงตรง	0.561	**0.643**	0.520
	ค่าการเรียกคืน	0.918	0.829	**0.930**
ชุดข้อมูลหลังใช้เทคนิค SMOTE	ค่าความแม่นยำ	0.724	0.644	0.629
	ค่าความเที่ยงตรง	0.632	0.572	0.553
	ค่าการเรียกคืน	0.907	0.878	0.901

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุดจากการประเมินในแต่ละตัววัดประสิทธิภาพ

ผลการวิเคราะห์ตารางที่ 5.1 ได้ดังนี้

1. การวิจัยพบว่าเมื่อใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE ร่วมกับการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมดนั้น มีค่าความแม่นยำ และค่าความเที่ยงตรง สูงที่สุด

2. ค่าการเรียกคืนมีประสิทธิภาพสูงสุดเมื่อทดสอบด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE โดยในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน โดยใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

การเปรียบเทียบประสิทธิภาพของวิธีการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนโดยใช้ชุดข้อมูลตั้งและชุดข้อมูลที่มีการจัดการกับข้อมูลที่ไม่สมดุล ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบปรากฏดังตารางต่อไปนี้

ตารางที่ 5.2 แสดงค่าความแม่นยำเฉลี่ยจากการทดลองของสองชุดข้อมูล และการใช้ตัววัดประสิทธิภาพการทำนายแบบต่างๆ ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบ

ชุดข้อมูล	ประสิทธิภาพในการทำนายโดยเฉลี่ย	ตัววัดประสิทธิภาพที่ใช้ในการคำนวณค่าถ่วงน้ำหนัก		
		AUC	F1	Optimal F1
ชุดข้อมูลก่อนใช้เทคนิค SMOTE	ค่าความแม่นยำ	0.881	**0.890**	0.868
	ค่าความเที่ยงตรง	0.656	**0.708**	0.624
	ค่าการเรียกคืน	0.898	0.839	**0.928**
ชุดข้อมูลหลังใช้เทคนิค SMOTE	ค่าความแม่นยำ	0.798	0.727	0.702
	ค่าความเที่ยงตรง	0.706	0.624	0.615
	ค่าการเรียกคืน	0.907	0.923	0.892

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุดจากการประเมินในแต่ละตัววัดประสิทธิภาพ

ผลการวิเคราะห์ตารางที่ 5.2 ได้ดังนี้

1. การวิจัยพบว่าเมื่อใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE ร่วมกับการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมดนั้นมีค่าความแม่นยำ และค่าความเที่ยงตรง สูงที่สุด

2. ค่าการเรียกคืนมีประสิทธิภาพสูงสุดเมื่อทดสอบด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE โดยในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอนโดยใช้ตัววัดประสิทธิภาพ F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

5.2 สรุปและอภิปรายผล

การปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับยังคงเป็นประเด็นที่ท้าทายและซับซ้อน ในงานวิจัยนี้ได้นำเสนอวิธีคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนายด้วยค่า F1-Score โดยได้รับแนวคิดมาจากการปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้เมตาแบบถ่วงน้ำหนักแบบปรับโดยการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนายด้วยค่า AUC ซึ่งนำเสนอโดย

Ren et al. (2022) งานวิจัยนี้ผู้วิจัยตั้งสมมติฐานว่าการคำนวณค่าถ่วงน้ำหนักสำหรับการเรียนรู้เมตาแบบสองขั้นตอนโดยใช้ค่า F1-Score ในขั้นตอนการคำนวณค่าถ่วงน้ำหนักมีประสิทธิภาพในการทำนายไม่น้อยกว่าการใช้ค่า AUC ในขั้นตอนการคำนวณค่าถ่วงน้ำหนัก

เมื่อพิจารณาในภาพรวมพบว่า การคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนาย F1-Score โดยพิจารณาจุดตัด 0.5 มีประสิทธิภาพการทำนายในค่าความแม่นยำและความเที่ยงตรงสูงที่สุดเมื่อทำการทดสอบด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE ในทั้งสองกรณีของการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ ส่วนประสิทธิภาพการทำนายในค่าการเรียกคืนนั้นให้ค่าการเรียกคืนสูงสุดเมื่อทดสอบด้วยชุดข้อมูลที่ตั้งต้นร่วมกับการคำนวณค่าถ่วงน้ำหนักแบบปรับจากตัววัดประสิทธิภาพการทำนาย F1-Score โดยพิจารณาค่าจุดตัดที่เหมาะสมที่สุดจากดัชนีโยเดนที่สูงสุด ในทั้งสองกรณีของการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบ สอดคล้องกับงานวิจัยของ (Hanczar et al., 2010; Muschelli, 2020) ที่ทำการทดสอบความน่าเชื่อถือของการของการวัดประสิทธิภาพด้วยค่า AUC ที่การคำนวณขึ้นอยู่กับอัตราผลบวกจริงและอัตราผลลบจริงเป็นหลัก โดยงานวิจัยดังกล่าวได้แสดงให้เห็นว่าในการใช้ค่า AUC วัดประสิทธิภาพในการทดลองที่มีกลุ่มตัวอย่างจำนวนน้อยนั้นมีความไม่แม่นยำและเกิดความผิดพลาดได้สูง รวมทั้งเสนอให้ใช้ค่า AUC ในการวัดประสิทธิภาพในการทดลองในกรณีเฉพาะที่ชุดข้อมูลมีความสมดุลของข้อมูลด้วย และในงานวิจัยของ Lobo et al., (2008) ที่แสดงให้เห็นถึงความผิดพลาดของการใช้ค่า AUC ในการวัดประสิทธิภาพการทำนายของตัวแบบด้วยเช่นกัน

อนึ่ง ประสิทธิภาพในการทำนายด้วยชุดข้อมูลที่มีการจัดการกับข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE นั้น ไม่ได้ให้ผลลัพธ์ที่ดีกว่าการทำนายด้วยชุดข้อมูลก่อนใช้เทคนิค SMOTE ซึ่งสอดคล้องกับงานวิจัยของ (Blagus & Lusa, 2013; Elor & Averbuch-Elor, 2022) ที่ทำการทดลองประสิทธิภาพของการจัดการชุดข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE สำหรับการจำแนกประเภทแบบไบนารี่ จากชุดข้อมูลที่มีความหลากหลายของข้อมูล 73 ชุด โดยมีอัตราส่วนความไม่สมดุลของข้อมูลตั้งแต่ 1.82 ถึง 129.44 ทำการเปรียบเทียบผลการทำนายด้วยการเรียนรู้ของเครื่องจาก 4 ตัวแบบ คือ ต้นไม้ตัดสินใจLight GBM, XGBoost และ CatBoost โดยใช้ตัววัดประสิทธิภาพการทำนายด้วยค่า logloss และ AUC พบว่าในภาพรวมการใช้เทคนิค SMOTE ในการจัดการกับข้อมูลที่ไม่สมดุลนั้นไม่ได้ส่งผลต่อการปรับปรุงประสิทธิภาพของการทำนายเช่นกันเมื่อใช้เทคนิค SMOTE ร่วมกับตัวจำแนกแข็งแรง (Strong Classifier) โดยงานวิจัยดังกล่าวได้มีข้อเสนอว่าการใช้เทคนิค SMOTE เพื่อเพิ่มประสิทธิภาพของการทำนายนั้นจะเกิดประโยชน์สูงสุดเมื่อใช้ร่วมกับตัวจำแนกอ่อนแอ (Weak Classifier)

เมื่อพิจารณาประสิทธิภาพในการทำนายโดยเปรียบเทียบวิธีการคำนวณค่าเฉลี่ยของชุดข้อมูลทดสอบ พบว่า การคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของรอบย่อยในแต่ละรอบนั้น

มีประสิทธิเฉลี่ยสูงกว่าการคำนวณค่าเฉลี่ยชุดข้อมูลทดสอบด้วยการหารจำนวนของตัวแบบทั้งหมด โดยผู้วิจัยตั้งข้อสังเกตว่า ในการหาค่าเฉลี่ยชุดข้อมูลทดสอบนั้น เกิดจากการนำชุดข้อมูลทดสอบในแต่ละรอบย่อยมารวมกัน จึงสมเหตุสมผลกว่าในการใช้จำนวนของรอบย่อยเป็นตัวหารเพื่อหาค่าเฉลี่ยของชุดข้อมูลทดสอบ

จากการทำวิจัยครั้งนี้ การคำนวณค่าถ่วงน้ำหนักแบบปรับจากการตัววัดประสิทธิภาพการทำนาย F1-Score โดยพิจารณาจุดตัด 0.5 มีประสิทธิภาพการทำนายสูงที่สุดเมื่อทดสอบบนข้อมูลตั้งต้นที่ไม่ได้ผ่านกระบวนการจัดการกับข้อมูลที่ไม่สมดุล ดังนั้นจึงนำเสนอการใช้ตัววัดประสิทธิภาพการทำนาย F1-Score โดยพิจารณาจุดตัด 0.5 ในการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการปรับปรุงความสามารถในการพยากรณ์แบบไบนารี่โดยใช้การเรียนรู้แบบเมตาในงานวิจัยแขนงอื่นอีกด้วย

5.3 ข้อเสนอแนะ

จากงานวิจัยชิ้นนี้ผู้สนใจอาจจะนำไปศึกษาต่อได้อีกในกรณีดังนี้

1. ขอบเขตของข้อมูล เนื่องจากในการคาดการณ์ความยากจนระดับครัวเรือนในงานวิจัยนี้ใช้ข้อมูลจากการลงพื้นที่สำรวจประชากรรายครัวเรือนในจังหวัดสระบุรีเพียงจังหวัดเดียว โดยเป็นข้อมูลระดับครัวเรือนจำนวน 352 ครัวเรือน ผู้ที่สนใจอาจใช้ข้อมูลที่มีการสำรวจครอบคลุมทุกจังหวัดในประเทศ และมีจำนวนครัวเรือนที่ใช้ศึกษาเพิ่มขึ้นเพื่อเพิ่มประสิทธิภาพในการคาดการณ์ความยากจน

2. ขั้นตอนการหาคุณลักษณะที่สำคัญของข้อมูลเพื่อนำเข้าตัวแบบ ผู้วิจัยเลือกใช้ตัวแบบต้นไม้ตัดสินใจเพื่อหาค่าคะแนนคุณลักษณะที่สำคัญที่ส่งผลต่อค่าความยากจนและทำการคัดเลือกจากคุณลักษณะที่มีค่าคะแนนสูงสุดตามลำดับ และรวมคะแนนได้มากกว่าร้อยละ 50 ของคะแนนรวมทั้งหมด ผู้ที่สนใจอาจใช้วิธีการคัดจากการคัดเลือกตัวแปรด้วยวิธีอื่น ๆ อาทิ วิธีเพิ่มตัวแปร (Forward Selection), วิธีลดตัวแปร (Backward Elimination) หรือวิธีเพิ่มตัวแปรอิสระแบบขั้นต้น (Stepwise Selection) ร่วมด้วย

3. การจัดการกับข้อมูลที่ไม่สมดุล ในงานวิจัยนี้ผู้วิจัยได้ใช้เทคนิค SMOTE เพื่อจัดการกับข้อมูลที่ไม่สมดุลในขั้นตอนก่อนการแบ่งข้อมูลด้วยเทคนิค 5-fold Cross Validation ซึ่งทำให้เป็นการสุ่มข้อมูลข้างน้อยเพิ่มขึ้นทั้งในชุดข้อมูลการฝึกและชุดข้อมูลทดสอบ เนื่องจากข้อจำกัดทางด้านการเขียนโปรแกรมคอมพิวเตอร์ ผู้ที่สนใจอาจเลือกใช้เทคนิค SMOTE ในขั้นตอนหลังจากการแบ่งข้อมูล และเลือกใช้เทคนิค SMOTE เฉพาะชุดข้อมูลทดสอบเพื่อความแม่นยำของการคาดการณ์มากยิ่งขึ้น

4. วิธีการทดสอบสมมติฐานของการคำนวณค่าถ่วงน้ำหนักแบบปรับสำหรับการเรียนรู้เมตาแบบสองขั้นตอน ในงานวิจัยนี้ ผู้วิจัยเลือกใช้ตัววัดประสิทธิภาพ 3 ตัวในการเปรียบเทียบ โดยในความเป็นจริงยังมีตัววัดประสิทธิภาพอื่น ๆ ที่น่าสนใจ ผู้ที่สนใจอาจเลือกตัววัดประสิทธิภาพอื่น ๆ พิจารณาร่วมด้วย

5. งานวิจัยนี้ในขั้นตอนการเรียนรู้เมตา ผู้วิจัยเลือกใช้ตัวแบบ 6 ตัวแบบในการคาดการณ์ความยากจน เพื่อได้มาซึ่งข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นผู้ที่สนใจอาจใช้เทคนิคการวิเคราะห์การจำแนกประเภทแบบอื่น ๆ ร่วมด้วยในการเรียนรู้แบบเมตา อาทิ ซัพพอร์ตเวกเตอร์แมชชีน, เพื่อนบ้านใกล้ที่สุด, เอดาบูท หรืออื่น ๆ รวมถึงในขั้นตอนการหาค่าพารามิเตอร์ที่ดีที่สุดนั้น ผู้วิจัยได้ทำการหาค่าพารามิเตอร์โดยใช้ข้อมูลทั้งหมดก่อนการแบ่งข้อมูลด้วยเทคนิค 5-fold Cross Validation เพื่อหาค่าพารามิเตอร์ที่ดีที่สุดของแต่ละตัวแบบก่อนนำตัวแบบดังกล่าว มาใช้ร่วมกันในการเรียนรู้เมตา ผู้ที่สนใจอาจเลือกใช้ขั้นตอนการหาพารามิเตอร์ที่ดีที่สุดนี้ในทุกรอบของการทดลองย่อยหลังจากการแบ่งข้อมูล กล่าวคือ ใช้ผลจากชุดข้อมูลตรวจสอบ (Validation) ของแต่ละรอบเพื่อหาคาพารามิเตอร์ที่ดีที่สุดใหม่ในทุกรอบ และใช้วิธีเดียวการนี้กับตัวแบบการถดถอยลอจิสติกส์ในขั้นสุดท้ายด้วย เพื่อเพิ่มความแม่นยำของตัวแบบและอาจทำให้ประสิทธิภาพของการคาดการณ์โดยรวมดียิ่งขึ้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บรรณานุกรม

1. Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty Classification Using Machine Learning: The Case of Jordan. *Sustainability*, *13*(3), 1412. https://www.mdpi.com/2071-1050/13/3/1412
2. Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics*, *14*(1), 1-16.
3. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7
4. Elor, Y., & Averbuch-Elor, H. (2022). *To SMOTE, or not to SMOTE?*
5. Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*(6), 822-830. https://doi.org/10.1093/bioinformatics/btq037
6. Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145-151. https://doi.org/https://doi.org/10.1111/j.1466-8238.2007.00358.x
7. Muschelli, J. (2020). ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of classification*, *37*(3), 696-708.
8. Pluliková, N. (2015). Poverty analysis using machine learning methods. In.
9. Ren, Z. H., Yu, C. Q., Li, L. P., You, Z. H., Guan, Y. J., Li, Y. C., & Pan, J. (2022). SAWRPI: A Stacking Ensemble Framework With Adaptive Weight for Predicting ncRNA-Protein Interactions Using Sequence Information. *Front Genet*, *13*, 839540. https://doi.org/10.3389/fgene.2022.839540
10. S Sani, N., Abdul Rahman, M., Bakar, A., Sahran, S., & Sarim, H. (2018). Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification. *International Journal on Advanced Science, Engineering and Information Technology*, *8*, 1698. https://doi.org/10.18517/ijaseit.8.4-2.6829
11. Sikora, R., & Al-laymoun, O. l. H. (2014). A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms. *Journal of International Technology and Information Management*, *23*, 1.
12. Surawatchayotin, A., & Paireekreng, W. (2021). The predictive model of higher education guidance using integrated techniques for imbalanced data of learner groups. *JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY*, *11*(1), 65-74. https://doi.org/10.14456/jist.2021.8

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาคผนวก

ตารางที่ 1 คำอธิบายโดยละเอียดของคุณลักษณะของข้อมูล

ชื่อคุณลักษณะ	ประเภทของข้อมูล	คำอธิบาย
HH_status	ข้อมูลเชิงคุณภาพ	สถานะครัวเรือนยากจนเป้าหมาย 1 = ใช่ , 0 = ไม่ใช่
HH_unit_area	ข้อมูลเชิงคุณภาพ	เขตการปกครองของที่อยู่อาศัยในเขตหรือนอกเขตเทศบาล 1=นอกเขตเทศบาล , 0=ในเขตเทศบาล
HH_size	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกทั้งหมดของครัวเรือน (มีชื่ออยู่ในทะเบียนบ้าน หรือไม่ก็ได้) ที่อยู่อาศัยในอยู่ในบ้าน/สถานที่เดียวกัน จัดหา หรือใช้สิ่ง อุปโภคบริโภคอันจำเป็นแก่การครองชีพร่วมกัน โดยบุคคลเหล่านั้นอาจเป็น ญาติ หรือไม่เป็นญาติก็ได้ แต่ไม่นับสมาชิกที่ไปอยู่อาศัยที่อื่นและไม่กลับมาพักนอนที่บ้าน นานเกิน 3 เดือนในรอบ 12 เดือนที่ผ่านมานับจากวัน สัมภาษณ์ โดยในกรณีที่จากไปเกิน 3 เดือนแต่ไม่มีที่อยู่อาศัยประจำที่อื่น
HH_size_registration	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกตามทะเบียนบ้าน
Gender_hh_head	ข้อมูลเชิงคุณภาพ	เพศของหัวหน้าครัวเรือน 0=ผู้ชาย, 1=ผู้หญิง
Age_hh_head	ข้อมูลเชิงปริมาณ	อายุของหัวหน้าครัวเรือน หน่วยเป็นปี
Edu_level_hh_head	ข้อมูลเชิงคุณภาพ	ระดับการศึกษาของหัวหน้าครัวเรือน 1=ไม่เคยได้รับการศึกษา, 2=ประถมศึกษา, 3=มัธยมศึกษาตอนต้น, 4=มัธยมศึกษาตอนปลาย, 5=ปวช., 6=ปวส., 7=อนุปริญญา, 8=ปริญญาตรี, 9=สูงกว่าปริญญาตรี
Woman_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกในครัวเรือนเพศหญิง
Man_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกในครัวเรือนเพศชาย
Child_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกที่อายุต่ำกว่า 15 ปี
Adults_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกที่อายุ 16 – 59 ปี
Eldery_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกที่อายุมากกว่า 60 ปี

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Disabled_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกครัวเรือนที่พิการ ทุพพลภาพ โดย ไม่พิการ=0, พิการทางปัญญา=1, พิการทาง ร่างกาย=2, พิการทางร่างกายและสติปัญญา=3
HH_characteristics	ข้อมูลเชิงคุณภาพ	ประเภทที่อยู่อาศัย 1.บ้านเดี่ยว (บ้านที่ไม่มีผนังส่วนใดติดกับบ้านหลัง อื่น หรือ มีพื้นที่ว่างรอบตัวบ้าน) 2.ห้องแถว/ตึกแถว/อาคารพาณิชย์ 3.ทาวน์เฮ้าส์/บ้านแฝด 4.ห้องชุด (บ้านเช่า อพาร์ทแม้น แฟลต หรือ คอนโดมิเนียม) 5.เพิงพักพิงชั่วคราว 6.อื่น ๆ
HS_possession_status	ข้อมูลเชิงคุณภาพ	สถานภาพการครอบครองที่อยู่อาศัย 1.ไม่มีบ้านพักอาศัยเป็นหลักแหล่ง (เช่น อาศัย ตามที่สาธารณะเพื่อพักแรม) 2.อาศัยอยู่กับผู้อื่นโดยไม่เสียค่าเช่า (แยก ครัวเรือนกัน และมีสถานะเป็นผู้อาศัย) 3.เช่าบ้านอยู่ 4.อาศัยในบ้านพักสวัสดิการตามสิทธิที่ตนเอง หรือสมาชิกในครัวเรือนได้รับ 5.อาศัยหรือปลูกบ้านในที่ดินผู้อื่น อาทิ ญาติ ที่ดิน สาธารณะ ที่ดินของหน่วยงานภาครัฐ ภาคเอกชน ศาสนสถาน 6.มีบ้านและที่ดินเป็นของตนเอง (หมายถึงมี เอกสารสิทธิตามกฎหมายที่ดิน) แต่อยู่ระหว่างการ ผ่อนชำระ 7.มีบ้านและที่ดินเป็นของตนเอง (หมายถึงมี เอกสารสิทธิตามกฎหมายที่ดิน) แต่ติดจำนอง หรือค้ำประกัน 8.มีบ้านและที่ดินเป็นของตนเอง (หมายถึงมี เอกสารสิทธิตามกฎหมายที่ดิน) โดยไม่ต้องผ่อน ชำระ และไม่ติดจำนองหรือค้ำประกัน 9.อื่น ๆ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

HS_characteristics	ข้อมูลเชิงคุณภาพ	สภาพของที่อยู่อาศัย 1.บ้าน/อาคาร/ห้องเช่า ที่มีสภาพทรุดโทรม ไม่มีความปลอดภัย หรือใช้วัสดุก่อสร้างประเภทไม่คงทนถาวร เช่น ไม้อัด ไม้ไผ่ หลังคามุงจาก ผ้าใบ หรือวัสดุตามธรรมชาติ หรือไม่ได้มีการปรับปรุงและซ่อมแซมมากกว่า 20 ปี 2.บ้าน/อาคาร/ห้องเช่า ที่ใช้วัสดุแข็งแรงปานกลาง เช่น ไม้ หลังคามุงสังกะสี หรืออาจมีลักษณะ ครึ่งปูน ครึ่งไม้ หรือมีการปรับปรุงและซ่อมแซมบ้าน ให้อยู่ในสภาพพร้อมใช้งาน ในช่วงระยะเวลา 10-20ปี 3.บ้าน/อาคาร/ห้องเช่า ที่มีสภาพมั่นคงแข็งแรงและใช้วัสดุทนทาน เช่น ปูน อิฐบล็อก เหล็ก หลังคามุงกระเบื้อง หลังคาเมทัลชีท หรือมีการปรับปรุงและซ่อมแซมบ้าน ให้อยู่ในสภาพพร้อมใช้งาน ภายในช่วงระยะเวลาไม่เกิน 10 ปี
Number_of_room	ข้อมูลเชิงปริมาณ	จำนวนห้องที่กั้นเป็นสัดส่วนชัดเจน โดยไม่นับรวมห้องน้ำ
Safety_bathroom	ข้อมูลเชิงคุณภาพ	บ้านที่อยู่อาศัยมีห้องน้ำที่ปิดมิดชิด ปลอดภัย และอยู่ในตัวบ้าน 1 = ใช่ , 0 = ไม่ใช่
drain_system	ข้อมูลเชิงคุณภาพ	บ้านที่อยู่อาศัยมีระบบระบายน้ำได้ดี ไม่มีท่ออุดตัน และไม่มีน้ำขัง 1 = ใช่ , 0 = ไม่ใช่
Air_system	ข้อมูลเชิงคุณภาพ	บ้านที่อยู่อาศัยสามารถถ่ายเท และระบายอากาศได้ดี 1 = ใช่ , 0 = ไม่ใช่
Waste_sorting	ข้อมูลเชิงคุณภาพ	บ้านที่อยู่อาศัยมีการคัดแยกขยะ เพื่อเพิ่มมูลค่านำกลับไปใช้ประโยชน์ใหม่ หรือสะดวกต่อการกำจัด 1 = ใช่ , 0 = ไม่ใช่
Electricity_dwelling	ข้อมูลเชิงคุณภาพ	สถานะการมีไฟฟ้าใช้ 1.ไม่มีไฟฟ้าใช้ (ใช้แสงสว่างอื่น อาทิ เทียน ตะเกียง แบตเตอรี่ เครื่องปั่นไฟ โซลาร์เซลล์) 2.มีไฟฟ้าใช้ แต่ต้องชำระค่าไฟผ่านเพื่อนบ้าน/

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

		เอกชน/คนกลาง 3.มีไฟฟ้าใช้ และสามารถชำระค่าไฟโดยตรงกับหน่วยงานของรัฐ (การไฟฟ้านครหลวง / การไฟฟ้าส่วนภูมิภาค)
Consumption_water_source	ข้อมูลเชิงคุณภาพ	แหล่งที่มาน้ำสำหรับการอุปโภค (น้ำใช้) 1.น้ำจากระบบประปา (นครหลวง ภูมิภาค หมู่บ้าน ต่อพ่วงคนอื่น) 2.น้ำจากแหล่งน้ำตามธรรมชาติ (เช่น แม่น้ำ ลำคลอง อ่างเก็บน้ำ น้ำฝน บาดาล) 3.ซื้อน้ำ (น้ำถัง น้ำขวด ตู้กดน้ำ)
Drinking_water_source	ข้อมูลเชิงคุณภาพ	แหล่งที่มาน้ำสำหรับการบริโภค (น้ำดื่ม) 1.น้ำจากระบบประปา (นครหลวง ภูมิภาค หมู่บ้าน ต่อพ่วงคนอื่น) 2.น้ำจากแหล่งน้ำตามธรรมชาติ (เช่น แม่น้ำ ลำคลอง อ่างเก็บน้ำ น้ำฝน บาดาล) 3.ซื้อน้ำ (น้ำถัง น้ำขวด ตู้กดน้ำ)
Gas_source	ข้อมูลเชิงคุณภาพ	แหล่งที่มาของเชื้อเพลิง 1.แก๊สหุงต้ม (LPG) 2.ไฟฟ้า 3.เศษพืช/หญ้า/ฟาง/กองไม้ใบหญ้า 4.ถ่าน/ไม้ฟืน 5.อื่นๆ
Internet_status	ข้อมูลเชิงคุณภาพ	การเชื่อมต่ออินเทอร์เน็ต 1.ไม่สามารถเข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยเนื่องจากไม่มีเครือข่ายรองรับ 2.ไม่สามารถเข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยเนื่องจากไม่มีกำลังทรัพย์เพียงพอ 3.ไม่สามารถเข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยเนื่องจากไม่ต้องการใช้งาน 4.เข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยจาก Wifi ส่วนตัว

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

		5.เข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยจาก Wifi สาธารณะ 6.เข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยจากเครือข่ายโทรศัพท์มือถือ 7.เข้าถึงสัญญาณอินเทอร์เน็ตภายในที่อยู่อาศัยได้มากกว่า 1 ช่องทาง 8.อื่น ๆ
Internet_number	ข้อมูลเชิงปริมาณ	จำนวนสมาชิกที่ใช้อินเทอร์เน็ต
Mobile_phone_number	ข้อมูลเชิงปริมาณ	จำนวนโทรศัพท์มือถือในครัวเรือน
Tablet_number	ข้อมูลเชิงปริมาณ	จำนวนแท็บเล็ตในครัวเรือน
Computer_number	ข้อมูลเชิงปริมาณ	จำนวนคอมพิวเตอร์ตั้งโต๊ะถือในครัวเรือน
Laptop_number	ข้อมูลเชิงปริมาณ	จำนวนคอมพิวเตอร์โน๊ตบุ๊คในครัวเรือน
Internet_purpose	ข้อมูลเชิงคุณภาพ	วัตถุประสงค์หลักของการใช้งานอินเทอร์เน็ต 1.เพื่อการศึกษา/เรียนรู้/ค้นคว้าหาข้อมูลที่จำเป็น 2.เพื่อการทำงาน/ประกอบอาชีพ/หารายได้ 3.เพื่อความบันเทิง อาทิ ดูหนัง ฟังเพลง เล่นเกมส์ ใช้งานโซเชียลมีเดีย (ที่ไม่ก่อให้เกิดรายได้) 4.เพื่อเข้าถึงบริการ/ความช่วยเหลือของรัฐ 5.เพื่อเข้าถึงข้อมูลข่าวสาร 6.อื่น ๆ
Total_income	ข้อมูลเชิงปริมาณ	รายได้เฉลี่ยต่อเดือน หน่วยเป็นบาท/เดือน
Total_consume	ข้อมูลเชิงปริมาณ	ค่าใช้จ่ายเพื่อการบริโภค หน่วยเป็นบาท/เดือน
Total_consumer	ข้อมูลเชิงปริมาณ	ค่าใช้จ่ายเพื่อการอุปโภค หน่วยเป็นบาท/เดือน
Water_bill	ข้อมูลเชิงปริมาณ	ค่าน้ำ หน่วยเป็นบาท/เดือน
Electricity_bill	ข้อมูลเชิงปริมาณ	ค่าไฟ หน่วยเป็นบาท/เดือน
Gas_bill	ข้อมูลเชิงปริมาณ	ค่าเชื้อเพลิง หน่วยเป็นบาท/เดือน
Total_hs_rent	ข้อมูลเชิงปริมาณ	ค่าเช่าบ้าน หน่วยเป็นบาท/เดือน
Internet_bill	ข้อมูลเชิงปริมาณ	ค่าบริการอินเทอร์เน็ตไวไฟ หน่วยเป็นบาท/เดือน
Mobile_phone_bill	ข้อมูลเชิงปริมาณ	ค่าโทรศัพท์ หน่วยเป็นบาท/เดือน
Education_fees	ข้อมูลเชิงปริมาณ	ค่าเล่าเรียน/ค่าธรรมเนียมการศึกษา/ค่าบำรุงการศึกษา หน่วยเป็นบาท/ปีการศึกษา

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Education_uniform	ข้อมูลเชิงปริมาณ	ค่าเครื่องแบบ หน่วยเป็นบาท/ปีการศึกษา
School_supplies	ข้อมูลเชิงปริมาณ	ค่าสมุดหนังสือ/อุปกรณ์การเรียน หน่วยเป็นบาท/ปีการศึกษา
School_food	ข้อมูลเชิงปริมาณ	ค่าขนม/อาหารไปโรงเรียน หน่วยเป็นบาท/ปีการศึกษา
Totoal_savings	ข้อมูลเชิงปริมาณ	เงินออมรวมทั้งครัวเรือน จำนวนเงินที่ออม หน่วยเป็นบาท
Total_cash	ข้อมูลเชิงปริมาณ	เงินสด และทรัพย์สิน (อาทิ ทอง เพชร พลอย ของสะสมที่มีมูลค่า) จำนวนเงิน หน่วยเป็นบาท
Commercial_bank_deposits	ข้อมูลเชิงปริมาณ	เงินฝากกับธนาคารพาณิชย์ และสถาบันการเงิน จำนวนเงิน หน่วยเป็นบาท
Total_community_fund	ข้อมูลเชิงปริมาณ	เงินฝากกับกองทุนหมู่บ้าน สหกรณ์ กลุ่มออมทรัพย์ กองทุนชุมชน จำนวนเงิน หน่วยเป็นบาท
Government_bonds	ข้อมูลเชิงปริมาณ	พันธบัตรรัฐบาล สลากออมทรัพย์ (อาทิ ออมสิน ธกส.) กองทุนรวมตลาดเงิน จำนวนเงิน หน่วยเป็นบาท
National_savings_fund	ข้อมูลเชิงปริมาณ	กองทุนการออมแห่งชาติ (กอช.) จำนวนเงิน หน่วยเป็นบาท
Funeral_fund	ข้อมูลเชิงปริมาณ	ฌาปนกิจสงเคราะห์ จำนวนเงิน หน่วยเป็นบาท
Crisis_survive	ข้อมูลเชิงปริมาณ	หากเกิดสภาวะวิกฤติ เงินออมที่มีอยู่สามารถใช้จ่ายเพื่อยังชีพครอบครัวได้กี่เดือน จำนวนเดือน
Debt_status	ข้อมูลเชิงคุณภาพ	การมีหนีสินของครัวเรือน 1 = มี , 0 = ไม่มี
NonDebt_reason	ข้อมูลเชิงคุณภาพ	เหตุผลกรณีไม่มีหนี้สิน 1.ไม่มีหลักทรัพย์/บุคคลค้ำประกัน 2.ไม่มีความสามารถในการชำระคืน 3.ไม่ต้องการเป็นหนี้/ไม่มีความจำเป็นในการก่อหนี้ 4.อื่น ๆ
Family_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง เครือข่ายส่วนตัว อาทิ ญาติ/เพื่อน/เพื่อนบ้าน จำนวนเงิน หน่วยเป็นบาท
Family_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืนเครือข่ายส่วนตัว อาทิ ญาติ/เพื่อน/เพื่อนบ้าน หน่วยเป็นบาท/เดือน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Community_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง กองทุนการเงินของชุมชน (สหกรณ์ กลุ่มออมทรัพย์ และกลุ่มกองทุน) จำนวนเงิน หน่วยเป็นบาท
Community_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน กองทุนการเงินของชุมชน (สหกรณ์ กลุ่มออมทรัพย์ และกลุ่มกองทุน) หน่วยเป็นบาท/เดือน
Government_fund_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง กองทุนการเงินที่รัฐสนับสนุน (กองทุนหมู่บ้าน กองทุนเงินล้าน อื่น ๆ จำนวนเงิน หน่วยเป็นบาท
Government_fund_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน กองทุนการเงินที่รัฐสนับสนุน (กองทุนหมู่บ้าน กองทุนเงินล้าน อื่น ๆ) หน่วยเป็นบาท/เดือน
BAAC_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง ธนาคารเพื่อการเกษตรและสหกรณ์ (ธกส.) จำนวนเงิน หน่วยเป็นบาท
BAAC_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน ธนาคารเพื่อการเกษตรและสหกรณ์ (ธกส.) หน่วยเป็นบาท/เดือน
Commercial_bank_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง ธนาคารพาณิชย์ (รวมทุกแห่ง)
Commercial_bank_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน ธนาคารพาณิชย์ (รวมทุกแห่ง) หน่วยเป็นบาท/เดือน
Finance_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง สถาบันการเงินเอกชน (ไฟแนนซ์/บัตรกดเงินสด/บัตรผ่อนชำระสินค้า) จำนวนเงิน หน่วยเป็นบาท
Finance_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน สถาบันการเงินเอกชน (ไฟแนนซ์/บัตรกดเงินสด/บัตรผ่อนชำระสินค้า) หน่วยเป็นบาท/เดือน
Grocery_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง ร้านค้าอุปโภค บริโภค และปัจจัยการผลิต (อาทิ ปุ๋ย ยา เครื่องใช้ไฟฟ้า อื่น ๆ) จำนวนเงิน หน่วยเป็นบาท
Grocery_debt_settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน ร้านค้าอุปโภค บริโภค และปัจจัยการผลิต (อาทิ ปุ๋ย ยา เครื่องใช้ไฟฟ้า อื่น ๆ) หน่วยเป็นบาท/เดือน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Informal_loan_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง เงินกู้นอกระบบ จำนวนเงิน หน่วยเป็นบาท
Informal_loan _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน เงินกู้นอกระบบ หน่วยเป็นบาท/เดือน
SLF_debt	ข้อมูลเชิงปริมาณ	หนี้สินคงค้าง กองทุนเงินให้กู้ยืมเพื่อการศึกษา (กยศ.) จำนวนเงิน หน่วยเป็นบาท
SLF_debt _settlement	ข้อมูลเชิงปริมาณ	แบ่งชำระคืน กองทุนเงินให้กู้ยืมเพื่อการศึกษา (กยศ.) หน่วยเป็นบาท/เดือน
Debt_purpose	ข้อมูลเชิงคุณภาพ	วัตถุประสงค์หลักในการก่อหนี้ 1.ใช้เพื่อการอุปโภคบริโภค 2.ใช้เพื่อลงทุนประกอบอาชีพเช่นซื้อรถเข็นขายสินค้าซื้ออุปกรณ์ทำอาหารเดินทางไปต่างประเทศ 3.ใช้เพื่อซื้อ/ก่อสร้าง/ปรับปรุง/ซ่อมแซมที่อยู่อาศัยและสิ่งปลูกสร้าง 4.ใช้เพื่อการรักษาพยาบาล 5.ใช้เพื่อการศึกษา 6.ใช้เพื่อซื้อยานพาหนะ 7.ใช้เพื่อชำระหนี้เก่าที่คงค้าง 8.ใช้เพื่อกิจกรรมทางสังคม อาทิ งานบวช งานแต่ง งานศพ 9.ใช้เพื่อวัตถุประสงค์อื่น ๆ
Land_status	ข้อมูลเชิงคุณภาพ	การถือครองทรัพย์สินประเภทที่ดิน 1 = มี , 0 = ไม่มี
Total_land	ข้อมูลเชิงปริมาณ	จำนวนที่ดินที่ถือครอง จำนวนแปลง
Land_size	ข้อมูลเชิงปริมาณ	พื้นที่รวมที่ดินที่ถือครอง ตารางวา
Motorcycle_number	ข้อมูลเชิงปริมาณ	จำนวนจักรยานยนต์
Bigbike_number	ข้อมูลเชิงปริมาณ	จำนวนรถจักรยานยนต์ขนาดใหญ่
Small_car_number	ข้อมูลเชิงปริมาณ	จำนวนถยนต์ขนาดเล็ก (สามล้อ รถอีแต๋น รถกระป๋อ)
Car_number	ข้อมูลเชิงปริมาณ	จำนวนรถยนต์นั่งส่วนบุคคล (รถเก๋ง รถตู้)
Small_truck_number	ข้อมูลเชิงปริมาณ	จำนวนรถบรรทุกขนาดเล็ก (รถกระบะ)
Middle_truck_number	ข้อมูลเชิงปริมาณ	จำนวนถบรรทุกขนาดกลาง (4-6 ล้อ)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Truck_number	ข้อมูลเชิงปริมาณ	จำนวนรถบรรทุกขนาดใหญ่ (10 ล้อ ขึ้นไป)
Bus_number	ข้อมูลเชิงปริมาณ	จำนวนรถบัส
Rowboat_number	ข้อมูลเชิงปริมาณ	จำนวนเรือพาย
Motorboat_number	ข้อมูลเชิงปริมาณ	จำนวนเรือยนต์
Backhoe_number	ข้อมูลเชิงปริมาณ	จำนวนรถขุด/รถแบคโฮ/รถแม็คโคร
Trackter_number	ข้อมูลเชิงปริมาณ	จำนวนรถแทรกเตอร์
Farming_vehicle_number	ข้อมูลเชิงปริมาณ	จำนวนรถไถ/รถเกี่ยวนวดข้าว/รถดำนา
Agriculture_vehicle_number	ข้อมูลเชิงปริมาณ	จำนวนรถสำหรับการทำเกษตรกรรม อื่น ๆ
TV_number	ข้อมูลเชิงปริมาณ	จำนวนโทรทัศน์
Radio_number	ข้อมูลเชิงปริมาณ	จำนวนวิทยุ
Electric_cooking_pot_number	ข้อมูลเชิงปริมาณ	จำนวนหม้อหุงข้าวไฟฟ้า
Refrigerator_number	ข้อมูลเชิงปริมาณ	จำนวนตู้เย็น
Fan_number	ข้อมูลเชิงปริมาณ	จำนวนพัดลม
Air_condition_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องปรับอากาศ
Washing_machine_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องซักผ้า
Iron_number	ข้อมูลเชิงปริมาณ	จำนวนเตารีด
Water_boiler_number	ข้อมูลเชิงปริมาณ	จำนวนกาต้มน้ำไฟฟ้า
Electric_pot_pan_number	ข้อมูลเชิงปริมาณ	จำนวนเตา/กระทะไฟฟ้า
Microwave_number	ข้อมูลเชิงปริมาณ	จำนวนเตาอบ/ไมโครเวฟ
Water_pump_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องสูบน้ำไฟฟ้า(ปั๊มน้ำ)
Water_heater_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องทำน้ำอุ่น
Air_purifier_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องฟอกอากาศ
Vacuum_cleaner_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องดูดฝุ่น
Wifi_router_number	ข้อมูลเชิงปริมาณ	จำนวนเครื่องกระจายสัญญาณอินเทอร์เน็ต
Beef_cattle_number	ข้อมูลเชิงปริมาณ	จำนวนโคเนื้อ
Dairy_cattle_number	ข้อมูลเชิงปริมาณ	จำนวนโคนม
Buffalo_number	ข้อมูลเชิงปริมาณ	จำนวนควาย
Goat_number	ข้อมูลเชิงปริมาณ	จำนวนแพะ
Sheep_number	ข้อมูลเชิงปริมาณ	จำนวนแกะ
Pig_number	ข้อมูลเชิงปริมาณ	จำนวนหมู

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Chicken_number	ข้อมูลเชิงปริมาณ	จำนวนไก่
Duck_number	ข้อมูลเชิงปริมาณ	จำนวนเป็ด
Goose_number	ข้อมูลเชิงปริมาณ	จำนวนห่าน
Horse_number	ข้อมูลเชิงปริมาณ	จำนวน ม้า/ลา/ล่อ
Dog_number	ข้อมูลเชิงปริมาณ	จำนวนสุนัข
Cat_number	ข้อมูลเชิงปริมาณ	จำนวนแมว
Bird_number	ข้อมูลเชิงปริมาณ	จำนวนนก
Fish_number	ข้อมูลเชิงปริมาณ	จำนวนปลา นับเป็นขนาดบ่อต่อลูกบาศก์เมตร
Gov_welfare_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับบัตรสวัสดิการแห่งรัฐในครัวเรือน
Childcare_subsidy_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเงินอุดหนุนเพื่อการเลี้ยงดูเด็กแรกเกิด
Elderly_subsidy_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเบี้ยยังชีพผู้สูงอายุ
Disability_subsidy_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเบี้ยคนพิการ
AIDS_subsidy_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเงินสงเคราะห์เบี้ยยังชีพผู้ป่วยเอดส์
Edu_subsidy_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับทุนการศึกษาแบบไม่ต้องใช้คืน
Child_Protect_fund_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเงินช่วยเหลือจากกองทุนคุ้มครองเด็ก
We_win_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับโครงการเราชนะ
We_love_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับโครงการเรารักกัน
Co_payment_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับโครงการคนละครึ่ง
Spm_gm_number	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับโครงการยิ่งใช้ยิ่งได้
BAAC_subsidy	ข้อมูลเชิงปริมาณ	จำนวนผู้ที่ได้รับเงินช่วยเหลือเกษตรกรจาก ธกส
News_channel	ข้อมูลเชิงคุณภาพ	ช่องทางการรับรู้ข่าวสารเกี่ยวกับการบริการและให้ความช่วยเหลือ ของภาครัฐ 1.ไม่ได้รับข่าวสาร 2.โทรทัศน์ 3.วิทยุ 4.เว็บไซต์/โซเชียลมีเดีย/ไลน์ 5.หนังสือพิมพ์ 6.ผู้นำชุมชน 7.เพื่อนบ้าน/ญาติ 8.รถประชาสัมพันธ์เคลื่อนที่ 9.ป้ายติดประกาศ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Crisis_contributor	ข้อมูลเชิงคุณภาพ	กรณีที่ครัวเรือนได้รับความเดือดร้อน เผชิญกับภาวะวิกฤต จะได้รับ ความช่วยเหลือจากใครหรือหน่วยงานใดมากที่สุด 1.หน่วยงานราชการ ส่วนกลาง/ส่วนภูมิภาค 2.องค์กรปกครองส่วนท้องถิ่น (อบจ. อบต. อื่น ๆ) 3.ภาคเอกชน (บริษัท/ห้าง/ร้านค้า) 4.องค์กรภาคประชาสังคม (มูลนิธิ/กลุ่มอาสา) 5.สถาบันและองค์กรทางศาสนา 6.เครือข่ายชุมชนท้องถิ่น 7.ญาติพี่น้อง/เพื่อน/เพื่อนบ้าน 8.เครือข่ายสังคมออนไลน์ 9.เครือข่ายส่วนตัวอื่น ๆ
Covid_survive	ข้อมูลเชิงคุณภาพ	ความสามารถสามารถหาซื้อและมีกำลังทรัพย์เพียงพอในการซื้ออุปกรณ์ที่จำเป็นในการ ป้องกันตนเองจากโรคระบาด หมอกควัน ฝุ่นละออง และวิกฤตทางสุขภาพอื่น ๆ 1.สามารถหาแหล่งซื้อได้และมีกำลังทรัพย์เพียงพอ 2.สามารถหาแหล่งซื้อได้แต่ไม่มีกำลังทรัพย์เพียงพอ 3.ไม่สามารถหาแหล่งซื้อได้แต่มีกำลังทรัพย์เพียงพอ 4.ไม่สามารถหาแหล่งซื้อได้และไม่มีกำลังทรัพย์ 5.ไม่มีความจำเป็นต้องหาซื้อ เนื่องจากได้รับแจกอย่างเพียงพอ 6.อื่น ๆ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียน

ชื่อ-สกุล	ธารินทร์ สุขเนาว์
วัน เดือน ปี เกิด	22 กันยายน 2538
สถานที่เกิด	จังหวัดนครศรีธรรมราช
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ที่อยู่ปัจจุบัน	77/124 ถ.พิบูลสงคราม ม.8 ต.สวนใหญ่ อ.เมืองนนทบุรี จ.นนทบุรี 11000

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY