ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ

FACTORS INFLUENCING THE CLASSIFICATION OF POVERTY AMONG INFORMAL WORKERS

ชนม์นิภา ตันหยง
CHONNIPA TANYONG

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติและการวิเคราะห์ธุรกิจ
ภาควิชาสถิติ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2568
KMITL-2025-SC-M-050-039

FACTORS INFLUENCING THE CLASSIFICATION OF POVERTY AMONG INFORMAL WORKERS

CHONNIPA TANYONG

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS AND BUSSINESS ANALYTICS

DEPARTMENT OF STATISTICS SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2025
KMITL--2025-SC-M-050-039

หัวข้อการค้นคว้าอิสระ ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ
ชื่อนักศึกษา นางสาวชนม์นิภา ตันหยง
รหัสประจำตัว 66056016
ปริญญา วิทยาศาสตรมหาบัณฑิต (สถิติและการวิเคราะห์ธุรกิจ)
ภาควิชา สถิติ
พ.ศ. 2568
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ ผู้ช่วยศาสตราจารย์ ดร.กนกกรรณ์ ลี้โรจนาประภา

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์ เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกกลุ่มความยากจนของแรงงานนอกระบบและเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆ ในการจำแนกสถานะความยากจน โดยใช้ข้อมูลทุติยภูมิจากการสำรวจแรงงานนอกระบบ พ.ศ. 2567 ของสำนักงานสถิติแห่งชาติ และข้อมูลสถิติดัชนีความก้าวหน้าของคน ปี 2566 ของสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ จำนวนข้อมูลทั้งหมด 42,662 คน เป็นแรงงานที่มีสถานะยากจน จำนวน 4,569 คน การศึกษาใช้เกณฑ์เส้นความยากจนระดับประเทศที่ 3,034 บาทต่อเดือนในการจำแนกสถานะความยากจนของแรงงานนอกระบบ งานวิจัยนี้ใช้แบบจำลอง 4 แบบจำลอง ได้แก่ Logistic Regression ด้วยวิธีทางสถิติ และแบบจำลองการเรียนรู้ของเครื่อง 3 แบบจำลอง คือ Logistic Regression, Random Forest และ XGBoost ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุล ได้แก่ Random Undersampling และ SMOTE-ENN ประสิทธิภาพของแบบจำลองประเมินด้วยค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) เป็นหลัก เนื่องจากการระบุกลุ่มแรงงานนอกระบบที่ยากจนได้อย่างครอบคลุมมีความสำคัญสำหรับการกำหนดนโยบายช่วยเหลือได้ตรงกลุ่มเป้าหมาย ผลการศึกษาพบว่าแบบจำลอง XGBoost ที่ใช้วิธี SMOTE-ENN ให้ประสิทธิภาพดีที่สุด โดยให้ค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.5303 และค่าความระลึก (Recall) เท่ากับ 0.6182 การวิเคราะห์อิทธิพลของตัวแปรด้วย SHAP จากแบบจำลอง XGBoost แสดงให้เห็นว่าตัวแปรประเภทค่าจ้างรูปแบบอื่นๆ โดยแรงงานนอกระบบที่มีค่าจ้างในลักษณะที่ไม่ใช่รายวัน รายเดือน รายสัปดาห์ หรือไม่เป็นตัวเงิน มีแนวโน้มยากจนมากกว่าแรงงานนอกระบบที่ได้รับค่าจ้างในลักษณะที่แน่นอน รองลงมาคือ จำนวนชั่วโมงทำงานต่อเดือน ซึ่งแรงงานนอกระบบที่มีชั่วโมงการทำงานน้อยมีแนวโน้มเสี่ยงต่อการตกอยู่ในสถานะยากจนสูงขึ้น และแรงงานในภาคบริการมีโอกาสตกอยู่ในภาวะยากจนน้อยกว่าภาคเกษตรกรรมที่มีรายได้ไม่แน่นอน

คำสำคัญ : ความยากจน แรงงานนอกระบบ การจัดการข้อมูลไม่สมดุล SHAP การเรียนรู้ของเครื่อง

Independent Study Title Factors Influencing the Classification of Poverty Among Informal Workers
Student Name Miss Chonnipa Tanyong
Student ID 66056016
Degree Master of Science (Statistics and Business Analytics)
Department Statistics
Year 2025
Independent Study Advisor Asst.Prof.Dr.Kanogkan Leerojanaprapa

Abstract

This study aims to analyze factors affecting poverty classification among informal workers and compare the performance of various models in classifying poverty status. The research utilized secondary data from the 2024 Informal Labor Force Survey by the National Statistical Office and the 2023 Human Achievement Index statistics, comprising 42,662 individuals, of which 4,569 were classified as poor using the national poverty line of 3,034 baht per month. This research applied four models: Logistic Regression using statistical methods and three machine learning models including Logistic Regression, Random Forest, and XGBoost, combined with imbalanced data handling techniques namely Random Undersampling and SMOTE-ENN. The findings revealed that the XGBoost model using SMOTE-ENN achieved the best performance, with an F1-Score of 0.5303 and recall of 0.6182. Variable influence analysis using SHAP demonstrated that informal workers receiving compensation in irregular forms (non-daily, monthly, weekly, or non-monetary payments) showed higher poverty tendencies. The second most important factor was monthly working hours, where fewer working hours increased poverty risks. Additionally, workers in the service sector had lower poverty chances compared to those in agriculture.

Keywords : Poverty, Informal workers, Imbalanced data handling, SHAP, Machine learning

กิตติกรรมประกาศ

การค้นคว้าอิสระนี้สำเร็จได้ เนื่องจากได้รับความเมตตาจากอาจารย์ที่ปรึกษาการค้นคว้าอิสระผู้ช่วยศาสตราจารย์ ดร.กนกกรรณ์ ลี้โรจนาประภา ที่เมตตาให้คำปรึกษา และยอมเสียสละเวลาอันมีค่าในการให้คำแนะนำ ตรวจสอบและให้ข้อเสนอแนะในการจัดทำการค้นคว้าอิสระนี้เป็นอย่างดีมาโดยตลอด จนกระทั่งการค้นคว้าอิสระนี้สำเร็จลุล่วงไปได้ด้วยดี ผู้วิจัยซาบซึ้งในความกรุณาจากท่านอาจารย์และกราบขอบพระคุณเป็นอย่างสูง ณ โอกาสนี้

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สิทธิชัย เจริญเศรษฐศิลป์ และ ดร.สกุณา ศรีอโณทัย ที่ให้เกียรติเป็นกรรมการสอบการค้นคว้าอิสระ และเสียสละเวลาอันมีค่าให้คำปรึกษา และชี้แนวทางการปรับปรุงแก้ไข ตลอดการจัดทำการค้นคว้าอิสระฉบับนี้

รวมถึงขอขอบพระคุณ สำนักงานสถิติแห่งชาติ ที่ให้ความอนุเคราะห์ข้อมูลสำรวจแรงงานนอกระบบ ในการศึกษาวิเคราะห์เพื่อให้เกิดประโยชน์สูงสุดกับผู้วิจัยและหน่วยงานที่เกี่ยวข้อง

สุดท้ายนี้ขอขอบคุณครอบครัว และเพื่อนๆ ที่คอยให้การสนับสนุนและส่งกำลังใจให้แก่ผู้วิจัยตลอดจนการจัดทำการค้นคว้าอิสระครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี ผู้วิจัยหวังเป็นอย่างยิ่งว่าการค้นคว้าอิสระนี้จะเป็นประโยชน์และจะเป็นแนวทางในการพัฒนาต่อไป

นางสาวชนม์นิภา ตันหยง

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ก
บทคัดย่อภาษาอังกฤษ ..... ข
กิตติกรรมประกาศ ..... ค
สารบัญ ..... ง
สารบัญตาราง ..... ช
สารบัญรูป ..... ซ
บทที่ 1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา 1
1.2 วัตถุประสงค์ของงานวิจัย 2
1.3 สมมติฐานการวิจัย 3
1.4 ขอบเขตของงานวิจัย 3
1.5 ประโยชน์ที่คาดว่าจะได้รับ 3
1.6 กรอบแนวคิด 4
1.7 นิยามศัพท์เฉพาะ 5
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 6
2.1 แนวคิดและทฤษฎีเกี่ยวกับการแบ่งเส้นความยากจน 6
2.2.1 แนวคิดเส้นความยากจน 6
2.2.2 เส้นความยากจนด้านอาหาร 6
2.2.3 เส้นความยากจนด้านอาหารในหมวดหมู่สินค้าที่ไม่ใช่อาหาร 7
2.2.4 การประหยัดจากขนาด 7
2.2.5 วิธีการคำนวณเส้นความยากจน (Poverty Line) 7
2.2 ปัจจัยที่เกี่ยวข้องกับความยากจน 8
2.2.1 แนวคิดปัจจัยส่วนบุคคล 8
2.2.2 ปัจจัยด้านเศรษฐกิจ 11
2.2.3 ปัจจัยด้านสภาพการทำงานและค่าตอบแทน 12
2.3 วิธีการจัดการข้อมูลที่ไม่สมดุล 13
2.4 แบบจำลองการจำแนกประเภทของข้อมูล 13
2.4.1 การวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติ 14
2.4.2 การวิเคราะห์การถดถอยลอจิสติก ด้วยการเรียนรู้ของเครื่อง 18
2.4.3 แบบจำลอง Random Forest 19
2.4.4 แบบจำลอง XGBoost 21
2.5 แนวคิดและทฤษฎีเกี่ยวกับการวัดประสิทธิภาพของแบบจำลอง 22
2.5.1 เมทริกซ์วัดประสิทธิภาพ (Confusion Matrix) 23
2.5.2 การวิเคราะห์ความแม่นตรง (Cross Validation) 24

สารบัญ (ต่อ)

หน้า

2.5.3 การทดสอบสมมติฐานด้วย McNemar’s Test 25
2.6 การตีความผลการทำนายของแบบจำลองด้วย SHAP 26
2.7 งานวิจัยที่เกี่ยวข้อง 27
บทที่ 3 วิธีดำเนินการวิจัย 32
3.1 ประชากรและกลุ่มตัวอย่าง 32
3.1.1 ประชากร 32
3.1.2 กลุ่มตัวอย่าง 32
3.2 เครื่องมือที่ใช้ในการวิจัย 32
3.3 การเก็บรวบรวมข้อมูล 33
3.4 การเตรียมข้อมูล 35
3.4.1 การทำความสะอาดข้อมูล 35
3.4.2 การเตรียมข้อมูล 35
3.4.3 แบ่งข้อมูลเป็นข้อมูลฝึกหัด (Train Data) และข้อมูลทดสอบ (Test Data) 37
3.5 การพัฒนาแบบจำลอง 38
3.6 การประเมินผลแบบจำลอง 39
บทที่ 4 ผลการวิจัยและอภิปรายผล 40
4.1 สถิติเชิงพรรณา (Descriptive Statistics) 40
4.2 ผลลัพธ์ของการพัฒนาแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ 48
4.3 ผลลัพธ์ของการพัฒนาแบบจำลองด้วยการเรียนรู้ของเครื่อง 57
4.4 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุล ทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test 65
4.5 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบของ ทั้ง 4 แบบจำลอง 67
4.5.1 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบของ ทั้ง 4 แบบจำลองด้วยค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) 67
4.5.2 การทดสอบ McNemar's Test สำหรับการเปรียบเทียบประสิทธิภาพแบบจำลอง 70
4.6 การวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP 70
4.7 อภิปรายผลการวิจัย 74
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 76
5.1 สรุปผลการเปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุลทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test 76
5.2 สรุปผลเปรียบเทียบประสิทธิภาพในการทำนายความยากจนของแรงงานนอกระบบของทั้ง 4 แบบจำลอง 77
5.3 สรุปผลการวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP 78

สารบัญ (ต่อ)

หน้า

5.4 ข้อเสนอแนะ 78
เอกสารอ้างอิง 80
ภาคผนวก 84
ประวัติผู้เขียน 88

สารบัญตาราง

ตารางที่ **หน้า**

ตารางที่ 2.1 การแปลความหมายค่าสัมประสิทธิ์ของ Cramer’s V 15
ตารางที่ 2.2 ข้อมูลที่ใช้ในการทดสอบ McNemar’s Test 25
ตารางที่ 3.1 ไลบรารีบน Python ที่ใช้ในงานวิจัย 33
ตารางที่ 3.2 ตัวแปรที่ใช้ในงานวิจัย 34
ตารางที่ 3.3 คำอธิบายตัวแปรอิสระหลังผ่านการเตรียมข้อมูล 35
ตารางที่ 4.1 สถิติเชิงพรรณาของตัวแปรเชิงปริมาณ 48
ตารางที่ 4.2 การแปลงตัวแปรเชิงปริมาณด้วย เทคนิค Box-Cox Transformation 51
ตารางที่ 4.3 ผลการวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ 52
ตารางที่ 4.4 การตรวจสอบระหว่างความสัมพันธ์ระหว่างตัวแปรอิสระ ด้วย VIF 54
ตารางที่ 4.5 การจำแนกประเภทของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ โดยใช้ชุดข้อมูลฝึกหัด 55
ตารางที่ 4.6 การจำแนกประเภทของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ (ปรับ Threshold = 0.2341) โดยใช้ชุดข้อมูลทดสอบ 56
ตารางที่ 4.7 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง Logistic Regression สำหรับแต่ละวิธีการปรับสมดุลข้อมูล 57
ตารางที่ 4.8 การจำแนกประเภทของแบบจำลอง Logistic Regression ชุดข้อมูลฝึกหัด 58
ตารางที่ 4.9 การจำแนกประเภทของแบบจำลอง Logistic Regression (Threshold = 0.8264) .. 59
ตารางที่ 4.10 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง Random Forest สำหรับแต่ละวิธีการปรับสมดุลข้อมูล 60
ตารางที่ 4.11 การจำแนกประเภทของแบบจำลอง Random Forest ชุดข้อมูลฝึกหัด 60
ตารางที่ 4.12 การจำแนกประเภทของแบบจำลอง Random Forest ชุดข้อมูลทดสอบ 61
ตารางที่ 4.13 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง XGBoost สำหรับแต่ละวิธีการปรับสมดุลข้อมูล 62
ตารางที่ 4.14 การจำแนกประเภทของแบบจำลอง XGBoost ชุดข้อมูลฝึกหัด 63
ตารางที่ 4.15 การจำแนกประเภทของแบบจำลอง XGBoost ชุดข้อมูลทดสอบ 63
ตารางที่ 4.16 ทดสอบสมมติฐานการเปรียบเทียบค่าเฉลี่ยของ F1-Score ในแต่ละแบบจำลองโดยวิธีการจัดการข้อมูลไม่สมดุลทั้ง 2 วิธี 65
ตารางที่ 4.17 ค่าเฉลี่ย F1-Score จากการทดสอบสมมติฐาน 66
ตารางที่ 4.18 การจำแนกประเภทของแบบจำลองทั้ง 4 แบบจำลองบนชุดข้อมูลทดสอบ 68
ตารางที่ 4.19 ผลการทดสอบ McNemar's Test เพื่อเปรียบเทียบประสิทธิภาพแบบจำลอง 70

สารบัญรูป

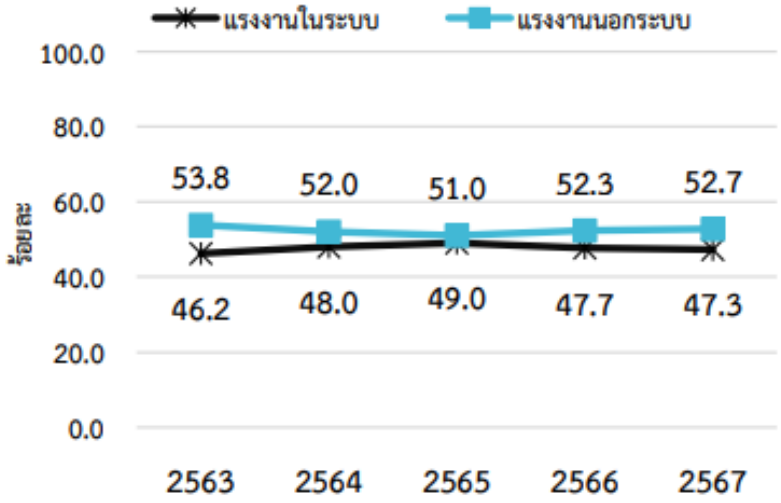
รูปที่ **หน้า**

รูปที่ 1.1 แนวโน้มของแรงงานในและนอกระบบ พ.ศ. 2563-2567 1
รูปที่ 2.1 ขั้นตอนและเทคนิคการคำนวณเส้นความยากจน 8
รูปที่ 2.2 ความเหลื่อมล้ำด้านรายได้ จำแนกรายภาค ปี พ.ศ. 2531 – 2566 11
รูปที่ 2.3 โครงสร้างของต้นไม้ตัดสินใจใน XGBoost 22
รูปที่ 2.4 Confusion Matrix 23
รูปที่ 2.5 K-fold Validation 24
รูปที่ 2.6 ตัวอย่างการใช้ SHAP อธิบายระดับผลกระทบของค่าตัวแปรต่อผลลัพธ์ของแบบจำลอง 26
รูปที่ 4.1 สถานะความยากจนของแรงงานนอกระบบ 41
รูปที่ 4.2 เพศของแรงงานนอกระบบ 41
รูปที่ 4.3 สภานภาพสมรสของแรงงานนอกระบบ 42
รูปที่ 4.4 ระดับการศึกษาสูงสุดของแรงงานนอกระบบ 42
รูปที่ 4.5 อาชีพของแรงงานนอกระบบ 43
รูปที่ 4.6 สถานะหัวหน้าครัวเรือนของแรงงานนอกระบบ 43
รูปที่ 4.7 ภาคที่อยู่อาศัยของแรงงานนอกระบบ 44
รูปที่ 4.8 เขตการปกครองที่อยู่อาศัยของแรงงานนอกระบบ 44
รูปที่ 4.9 ประเภทค่าจ้างที่ได้รับของแรงงานนอกระบบ 45
รูปที่ 4.10 กิจกรรมทางเศรษฐกิจของแรงงานนอกระบบ 45
รูปที่ 4.11 การมีปัญหาจากสภาพแวดล้อมการทำงานของแรงงานนอกระบบ 46
รูปที่ 4.12 การมีปัญหาจากการทำงานของแรงงานนอกระบบ 46
รูปที่ 4.13 การมีความเสี่ยงจากการทำงานของแรงงานนอกระบบ 47
รูปที่ 4.14 ประเภทสถานที่ทำงานของแรงงานนอกระบบ 47
รูปที่ 4.15 สถานประกอบการที่ทำงานของแรงงานนอกกระบบจดทะเบียนกับหน่วยงานรัฐ 48
รูปที่ 4.16 ขนาดความสัมพันธ์ของตัวแปรอิสระเชิงกลุ่ม ตัวแปรเชิงอันดับและสถานะความยากจน 49
รูปที่ 4.17 ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับตัวแปรตาม 50
รูปที่ 4.18 เมทริกซ์วัดประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ ปรับ Threshold = 0.2341 56
รูปที่ 4.19 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง Logistic Regression (Machine Learning) 59
รูปที่ 4.20 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง Random Forest 62
รูปที่ 4.21 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง XGBoost 64
รูปที่ 4.22 ค่า ค่าประสิทธิภาพโดยรวม (F1-Score) เปรียบเทียบประสิทธิภาพของแบบจำลอง 68
รูปที่ 4.23 ค่าความระลึก (Recall) เปรียบเทียบประสิทธิภาพของแบบจำลอง 69
รูปที่ 4.24 ผลการวิเคราะห์ความสำคัญของตัวแปรด้วยค่า SHAP 71

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

สถานการณ์แรงงานในประเทศไทย ปี 2567 จากการสำรวจแรงงานนอกระบบ พ.ศ.2567 ของสำนักงานสถิติแห่งชาติ พบว่าจำนวนแรงงานในประเทศไทยมีมากกว่า 40 ล้านคน โดยเป็นแรงงานในระบบ หรือผู้มีงานทำที่ได้รับความคุ้มครองหรือหลักประกันทางสังคมจากการทำงาน จำนวน 18.9 ล้านคน และแรงงานนอกระบบ หรือผู้มีงานทำที่ไม่ได้รับความคุ้มครอง หรือไม่มีหลักประกันทางสังคมจากการทำงาน จำนวน 21.1 ล้านคน และจากข้อมูลย้อนหลัง 5 ปีที่ผ่านมา



รูปที่ 1.1 แนวโน้มของแรงงานในและนอกระบบ พ.ศ. 2563-2567
ที่มา: สำนักงานสถิติแห่งชาติ (2567)

จากรูปดังกล่าวจะเห็นว่าร้อยละของจำนวนแรงงานนอกระบบในประเทศมีมากกว่าแรงงานในระบบ ซึ่งสะท้อนโครงสร้างตลาดแรงงานไทยที่ยังพึ่งพาแรงงานนอกระบบเป็นหลัก โดยแรงงานนอกระบบจำนวนมาก ทำงานอยู่ในภาคตะวันออกเฉียงเหนือ รองลงมาเป็นภาคเหนือ และน้อยที่สุดในกรุงเทพมหานคร ส่วนใหญ่อายุอยู่ระหว่าง 40-59 ปี และเป็นแรงงานชายมากกว่าแรงงานหญิง แรงงานนอกระบบมากกว่าครึ่งหนึ่งเป็นแรงงานในภาคเกษตรกรรม จำนวนกว่า 11.4 ล้านคน รองลงมาเป็นภาคการบริการและการค้า โดยภาคเกษตรถือเป็นภาคเศรษฐกิจที่สำคัญ เนื่องจากสามารถรองรับการเปลี่ยนแปลงที่รุนแรงต่อระบบเศรษฐกิจในช่วงที่เศรษฐกิจต้องมีการปรับตัวเพื่อตอบสนองต่อความผันผวนของสภาพแวดล้อมทางเศรษฐกิจ เช่น เหตุการณ์การแพร่ระบาดของไวรัสโควิด-19 ภาคเกษตรกรรมยังสามารถรักษาระดับและเพิ่มมูลค่าของผลผลิตในภาคเศรษฐกิจได้ อีกทั้งยังรองรับแรงงานที่ได้รับผลกระทบต่อการจ้างงานเฉียบพลันได้ (คณะพัฒนาการเศรษฐกิจ, 2565)

แม้แรงงานนอกระบบจะมีบทบาทสำคัญต่อเศรษฐกิจ แต่จากการสำรวจของสำนักงานสถิติแห่งชาติพบประเด็นที่น่าวิตกคือ มีแรงงานนอกระบบถึง 6.3 ล้านคน (ร้อยละ 29.9) ที่ประสบปัญหาจากการทำงาน โดยเฉพาะปัญหาค่าตอบแทนไม่เหมาะสม (ร้อยละ 47.7) และงานขาดความต่อเนื่อง

(ร้อยละ 19.0) (สำนักงานสถิติแห่งชาติ, 2567) ซึ่งสะท้อนถึงความเปราะบางทางเศรษฐกิจที่นำไปสู่วงจรความยากจนเรื้อรัง เนื่องจากไม่ได้รับการคุ้มครองทางสังคม ไม่มีสวัสดิการรองรับความเสี่ยงจากการเจ็บป่วยหรือสูญเสียรายได้ ทำให้แรงงานกลุ่มนี้มีคุณภาพชีวิตต่ำ โดยปัญหาด้านค่าตอบแทนพบต่อเนื่องติดต่อกันหลายปี ในขณะที่เส้นความยากจนซึ่งเป็นเกณฑ์ที่คำนวณมาจากระดับมาตรฐานการครองชีพขั้นต่ำที่สุดที่คนๆ หนึ่งมีชีวิตอยู่ได้โดยคำนึงถึงค่าใช้จ่ายด้านอาหารและไม่ใช่อาหาร (เช่น ที่อยู่อาศัย) ประเทศไทยใช้เส้นความยากจนในการแจงนับจำนวนคนจน ซึ่งจากรายงานการวิเคราะห์สถานการณ์ความยากจนและความเหลื่อมล้ำในประเทศไทย (สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2566) ระบุว่าในปัจจุบันเส้นความยากจนของประเทศไทยอยู่ที่ 3,043 บาทต่อคนต่อเดือน ซึ่งปรับตัวขึ้นเล็กน้อย อาจเป็นผลมาจากการขยายตัวของเศรษฐกิจภาคเกษตรกรรมมีสัดส่วนแรงงานยากจนสูงที่สุด ประกอบกับโครงการลงทะเบียนเพื่อสวัสดิการแห่งรัฐ ปี 2565 ซึ่งเปิดโอกาสให้คนจนสามารถเข้าถึงบริการและสวัสดิการต่างๆของภาครัฐมากขึ้น แต่ในกลุ่มของแรงงานนอกระบบยังพบว่ายังมีแรงงานนอกระบบจำนวนไม่น้อยมีรายได้ต่ำกว่าเส้นความยากจน สะท้อนถึงปัญหาความยากจนเชิงโครงสร้าง ซึ่งเกี่ยวข้องกับคุณภาพชีวิตของแรงงาน การเข้าถึงโอกาสทางเศรษฐกิจและประสิทธิภาพของมาตรการรัฐในการสนับสนุนกลุ่มแรงงานเปราะบาง

ปัญหาความยากจนของแรงงานนอกระบบเป็นประเด็นสำคัญที่ต้องการแก้ไขอย่างเร่งด่วน เนื่องจากส่งผลกระทบโดยตรงต่อคุณภาพชีวิตและความมั่นคงทางเศรษฐกิจ ดังนั้น การศึกษานี้จึงวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกสถานะความยากจนของแรงงานนอกระบบ และเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธีทางสถิติและการเรียนรู้ของเครื่องรวมทั้งสิ้น 4 แบบจำลอง ได้แก่ การวิเคราะห์การถดถอยลอจิสติกวิธีทางสถิติ และแบบจำลองที่พัฒนาด้วยการเรียนรู้ของเครื่องจำนวน 3 แบบจำลอง ประกอบด้วย แบบจำลอง Logistic Regression ,Random Forest และ XGBoost ซึ่งแต่ละแบบจำลองมีความสามารถที่แตกต่างกัน โดยการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติมีจุดเด่นในการให้ความสัมพันธ์ที่ตีความได้ชัดเจนระหว่างปัจจัยต่างๆ กับความน่าจะเป็นที่แรงงานจะตกอยู่ในสถานะยากจน โดยผลลัพธ์สามารถแสดงในรูป Odds Ratio ที่เข้าใจได้ง่าย (Stoltzfus, 2011) แบบจำลอง Random Forest เหมาะสำหรับข้อมูลที่มีความซับซ้อนและไม่สมดุล สามารถจับความสัมพันธ์แบบไม่เชิงเส้นและระบุความสำคัญของปัจจัยได้อย่างแม่นยำผ่านเทคนิค Ensemble Learning (Sohnesen & Stender, 2016) ส่วนแบบจำลอง XGBoost ใช้เทคนิค Gradient Boosting ที่ปรับปรุงประสิทธิภาพโดยการลดข้อผิดพลาดของแบบจำลองก่อนหน้าอย่างต่อเนื่อง พร้อมทั้งมีการควบคุม Overfitting (Qing Li et al., 2022) การเปรียบเทียบทั้ง 4 แบบจำลองจะช่วยคัดเลือกแบบจำลองที่มีความแม่นยำสูงสุดและเหมาะสมกับลักษณะข้อมูลแรงงานนอกระบบ เพื่อสนับสนุนหน่วยงานที่เกี่ยวข้องในการออกแบบมาตรการช่วยเหลือที่ตรงเป้าหมายและมีประสิทธิภาพ ซึ่งจะช่วยเพิ่มโอกาสในการเข้าถึงทรัพยากรและสิทธิประโยชน์ที่สำคัญ นำไปสู่การยกระดับรายได้ ลดภาระทางเศรษฐกิจ และพัฒนาคุณภาพชีวิตของแรงงานและครอบครัวในระยะยาว ส่งผลต่อการลดความเหลื่อมล้ำและสร้างเศรษฐกิจฐานรากที่เข้มแข็งอย่างยั่งยืน

1.2 วัตถุประสงค์ของงานวิจัย

1) ศึกษาปัจจัยส่วนบุคคล ปัจจัยด้านเศรษฐกิจ และปัจจัยสภาพการทำงานและค่าตอบแทนที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ

2) เปรียบเทียบประสิทธิภาพแบบจำลองที่แม่นยำในการจำแนกความยากจนของแรงงานนอกระบบ

1.3 สมมติฐานการวิจัย

ปัจจัยส่วนบุคคล ได้แก่ อายุ เพศ สถานภาพสมรส ระดับการศึกษา อาชีพ สถานะหัวหน้าครัวเรือน ภูมิภาคที่อาศัย เขตการปกครอง จำนวนสมาชิกในครัวเรือน **ปัจจัยด้านเศรษฐกิจ** ได้แก่ ค่าจ้างขั้นต่ำ ดัชนีความไม่เสมอภาคด้านรายได้ ดัชนีความก้าวหน้าของคน และ**ปัจจัยด้านสภาพการทำงานและค่าตอบแทน** ได้แก่ ประเภทของค่าตอบแทน กิจกรรมทางเศรษฐกิจ ปัญหาจากสภาพแวดล้อมการทำงาน ปัญหาจากการทำงาน ความเสี่ยงจากการทำงาน จำนวนชั่วโมงการทำงาน โบนัส ค่าล่วงเวลารายเดือน ผลประโยชน์ตอบแทนที่ไม่ได้เป็นตัวเงิน ประเภทสถานที่ทำงาน สถานประกอบการจดทะเบียนกับหน่วยงานรัฐ เป็นปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ

1.4 ขอบเขตของงานวิจัย

การศึกษาวิจัยครั้งนี้มุ่งเน้นเพื่อหาปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ โดยข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลทุติยภูมิ (Secondary Data) จาก 2 แหล่งข้อมูล ได้แก่

1) การสำรวจแรงงานนอกระบบ พ.ศ. 2567 ของสำนักงานสถิติแห่งชาติ โดยการสำรวจได้ดำเนินการพร้อมกันทั่วประเทศในระหว่างวันที่ 1-12 ของเดือนกรกฎาคม สิงหาคม และกันยายน พ.ศ. 2567 โดยใช้วิธีการสัมภาษณ์หัวหน้าครัวเรือนหรือสมาชิกในครัวเรือนตัวอย่าง ประกอบด้วย ตัวแปรปัจจัยด้านลักษณะส่วนบุคคล ตัวแปรปัจจัยด้านสภาพแวดล้อมการทำงานและค่าตอบแทน
2) ข้อมูลสถิติดัชนีความก้าวหน้าของคน ปี 2566 ของสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ประกอบด้วย ตัวแปรปัจจัยด้านเศรษฐกิจ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

หน่วยงานภาครัฐและหน่วยงานที่มีอำนาจหน้าที่เกี่ยวข้องกับการดูแลสวัสดิการของแรงงานสามารถนำผลการวิจัยในครั้งนี้ไปเป็นแนวทางออกแบบมาตรการสนับสนุนที่ตรงกับความต้องการของแรงงานได้อย่างมีประสิทธิภาพ และตรงกลุ่มเป้าหมาย เพื่อเพิ่มโอกาสในการเข้าถึงทรัพยากรและสิทธิประโยชน์ต่าง ๆ ที่จะช่วยยกระดับรายได้ ลดภาระทางเศรษฐกิจ และเพิ่มคุณภาพชีวิตของแรงงาน

1.6 กรอบแนวคิด

ตัวแปรอิสระ (Independent Variables)

1.7 นิยามศัพท์เฉพาะ

แรงงานนอกระบบ หมายถึง ผู้มีงานทำที่มีอายุ 15 ปีขึ้นไป ที่ไม่ได้รับความคุ้มครองตามกฎหมายและไม่มีหลักประกันทางสังคมจากการทำงาน

เส้นความยากจน หมายถึง หลักเกณฑ์ในการพิจารณาและจำแนกคนที่จนกับคนที่ไม่จนออกจากกันจะอาศัยแนวคิดพื้นฐานจากทฤษฎีอรรถประโยชน์สูงสุดของผู้บริโภค (Utility Theory)

สถานะยากจน หมายถึง ผู้ที่มีรายได้ทั้งสิ้นต่อเดือนน้อยกว่าเส้นความยากจน หรือ 3,034 บาทต่อเดือน

ปัจจัยส่วนบุคคล หมายถึง คุณลักษณะส่วนตัวของแรงงานนอกระบบ เช่น อายุ เพศ ระดับการศึกษาสูงสุด

ปัจจัยด้านสภาพการทำงานและค่าตอบแทน หมายถึง องค์ประกอบต่างๆ ที่เกี่ยวข้องกับสภาพแวดล้อมในการทำงาน เงื่อนไขการจ้างงาน และรูปแบบผลตอบแทนที่แรงงานได้รับ ซึ่งส่งผลต่อคุณภาพชีวิตและความมั่นคงทางเศรษฐกิจของผู้ทำงาน

ปัญหาจากการทำงาน หมายถึง ความยากลำบากหรืออุปสรรคที่เกิดขึ้นจากลักษณะงานและการบริหารจัดการงาน ได้แก่ ปัญหาด้านค่าตอบแทนที่ไม่เพียงพอหรือไม่เป็นธรรม งานที่มีความหนักเกินไป การทำงานที่ไม่เป็นไปตามเวลาปกติ งานที่ขาดความต่อเนื่องหรือไม่มั่นคง ชั่วโมงการทำงานที่มากเกินไป การไม่มีวันหยุด หรือการไม่สามารถลาหยุดหรือลาพักผ่อนได้ตามต้องการ รวมถึงการไม่มีสวัสดิการที่เหมาะสม

ปัญหาจากสภาพแวดล้อมการทำงาน หมายถึง ความไม่เหมาะสมของสถานที่แลลสภาพแวดล้อมทางกายภาพในการทำงาน ได้แก่ สถานที่ทำงานที่คับแคบ สกปรก อากาศไม่ถ่ายเทมีธรรมชาติของงานที่เต็มไปด้วยอิริยาบถที่ไม่เหมาะสมต่อสุขภาพ มีฝุ่นละออง ควัน กลิ่น เสียงดัง หรือแสงสว่างที่ไม่เหมาะสม รวมถึงปัญหาอื่นๆ ที่เกี่ยวข้องกับสภาพแวดล้อมทางกายภาพของสถานที่ทำงาน

ความเสี่ยงจากการทำงาน หมายถึง ความเสี่ยงและอันตรายที่อาจเกิดขึ้นในระหว่างการปฏิบัติงาน ได้แก่ การสัมผัสกับสารเคมีอันตราย การใช้เครื่องจักรหรือเครื่องมือที่เป็นอันตราย การเผชิญกับอันตรายต่อทุกส่วนของร่างกาย การทำงานในที่สูง/ใต้น้ำ/ใต้ดิน การเผชิญกับความไม่สงบหรือการก่อการร้าย รวมถึงความเสี่ยงด้านความปลอดภัยอื่นๆ ในการทำงาน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาเรื่อง “ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ” ผู้วิจัยได้ศึกษา ค้นคว้าแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง เพื่อเป็นแนวทางในการศึกษา ดังนี้

2.1 แนวคิดและทฤษฎีเกี่ยวกับการแบ่งเส้นความยากจน
2.2 ปัจจัยที่เกี่ยวข้องกับความยากจน
2.3 วิธีการจัดการข้อมูลที่ไม่สมดุล
2.4 แบบจำลองการจำแนกประเภทของข้อมูล
2.5 แนวคิดและทฤษฎีเกี่ยวกับการวัดประสิทธิภาพของแบบจำลอง
2.6 การตีความผลการทำนายของแบบจำลองด้วยค่า SHAP
2.7 งานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีเกี่ยวกับการแบ่งเส้นความยากจน

การแบ่งเส้นความยากจนเป็นแนวคิดสำคัญในการวัดและประเมินสถานะความยากจนในสังคม การกำหนดเส้นความยากจนจึงเป็นจุดเริ่มต้นที่สำคัญในการระบุกลุ่มประชากรที่ต้องการความช่วยเหลือและการออกแบบนโยบายเพื่อลดความยากจนอย่างมีประสิทธิภาพ โดยมีทฤษฎีและแนวทางหลัก ดังนี้

2.2.1 แนวคิดเส้นความยากจน

เส้นความยากจน (Poverty Line) คือ หลักเกณฑ์ในการพิจารณาและจำแนกคนที่จนกับคนที่ไม่จนออกจากกันจะอาศัยแนวคิดพื้นฐานจากทฤษฎีอรรถประโยชน์สูงสุดของผู้บริโภค (Utility Theory) โดยเส้นความยากจนจะสะท้อนมาตรฐานการครองชีพขั้นต่ำ (Minimum Standard Of Livine) ของสังคม เส้นความยากจนคำนวณจากมูลค่ารวมของค่าใช้จ่ายที่จำเป็นต่อการดำรงชีวิตในระดับพื้นฐาน ประกอบด้วย ค่าใช้จ่ายด้านอาหาร (เช่น พลังงานและสารอาหารที่จำเป็นต่อวัน) และค่าใช้จ่ายที่ไม่ใช่อาหาร (เช่น ที่อยู่อาศัย เสื้อผ้า การเดินทาง และค่ารักษาพยาบาล) (สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2566) ในประเทศไทยสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ (สศช.) เป็นหน่วยงานหลักในการกำหนดเส้นความยากจน โดยใช้ข้อมูลจากการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือน ซึ่งปรับค่าตามภาวะเงินเฟ้อและความแตกต่างของค่าครองชีพในแต่ละพื้นที่ เช่น เมืองหรือชนบท และในแต่ละภาคของประเทศ สำหรับปี 2566 เส้นความยากจนเฉลี่ยของประเทศไทยอยู่ที่ 3,043 บาทต่อคนต่อเดือน และใช้เป็นเกณฑ์ในการประเมินภาวะความยากจนและวางนโยบายสวัสดิการสังคม

2.2.2 เส้นความยากจนด้านอาหาร

เส้นความยากจนด้านอาหารถูกคำนวณโดยอิงจากต้นทุนสารอาหารและโปรตีน (Calorie Cost and Protein Cost) ของประชากร 10% ที่ยากจนที่สุด โดยใช้ข้อมูลราคาสินค้าในแต่ละภูมิภาคและเขตการปกครอง พร้อมถ่วงน้ำหนักด้วยสัดส่วนประชากรและปรับด้วยดัชนีราคาสินค้า

รายพื้นที่ (SPI) เพื่อสะท้อนต้นทุนเฉลี่ยระดับประเทศอย่างแม่นยำ การคำนวณนี้อิงจากทฤษฎีอรรถประโยชน์ที่ผู้บริโภคได้รับ ซึ่งเป็นฟังก์ชันเพิ่มขึ้นของต้นทุนสารอาหาร ทำให้สามารถเปรียบเทียบความเท่าเทียมของอรรถประโยชน์ระหว่างผู้บริโภคในพื้นที่และรูปแบบการบริโภคที่แตกต่างกันได้อย่างเป็นธรรม ทั้งนี้ เส้นความยากจนด้านอาหารควรมีการปรับทุก 10 ปี เพื่อให้สอดคล้องกับแบบแผนการบริโภคที่เปลี่ยนแปลง โดยข้อมูลด้านความต้องการสารอาหารขั้นต่ำของคนไทย ซึ่งกำหนดโดยคณะทำงานของกองโภชนาการ กระทรวงสาธารณสุข จะถูกนำมาใช้เป็นฐานข้อมูลสำคัญในการคำนวณ โดยพบว่าแนวโน้มของคนไทยในปัจจุบันต้องการแคลอรี่น้อยลงแต่ต้องการโปรตีนมากขึ้น (สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2566)

2.2.3 เส้นความยากจนด้านอาหารในหมวดหมู่สินค้าที่ไม่ใช่อาหาร

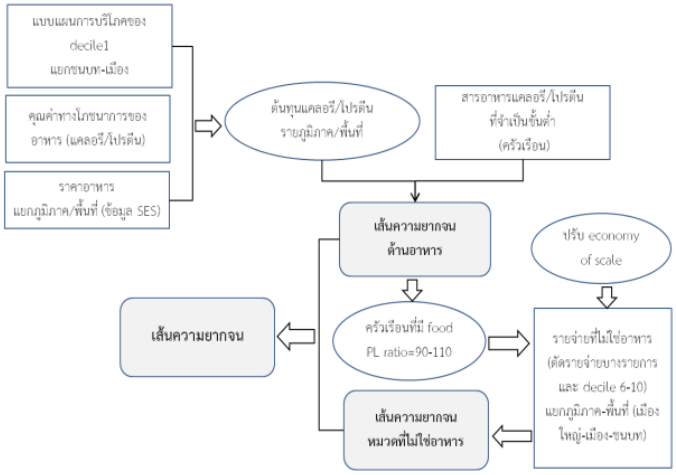
เส้นความยากจนในหมวดสินค้าที่ไม่ใช่อาหาร คำนวณโดยใช้แนวคิดเชิงอรรถประโยชน์ ซึ่งสะท้อนความสัมพันธ์ระหว่างค่าใช้จ่ายด้านอาหารและค่าใช้จ่ายรวม โดยเริ่มจากการกำหนดเส้นความยากจนด้านอาหาร แล้วหาค่าใช้จ่ายรวมที่ให้ระดับอรรถประโยชน์เท่ากัน เพื่อใช้เป็นเกณฑ์ในการคำนวณเส้นความยากจนรวม จากนั้นนำมาแยกส่วนเป็นค่าใช้จ่ายในหมวดที่ไม่ใช่อาหารผ่าน ขั้นตอน 3 ขั้น ได้แก่ (1) คำนวณค่า Food Welfare โดยเทียบค่าใช้จ่ายด้านอาหารกับเส้นความยากจนด้านอาหาร (2) คัดเลือกครัวเรือนที่มีค่า Food Welfare ระหว่าง 90–110 เพื่อให้ได้กลุ่มที่มีพฤติกรรมใกล้เคียงเกณฑ์ความยากจนด้านอาหาร และ (3) คำนวณค่าเฉลี่ยของค่าใช้จ่ายในหมวดสินค้าที่ไม่ใช่อาหารของกลุ่มนั้น ซึ่งผลลัพธ์จะสะท้อนเส้นความยากจนในหมวดสินค้าที่ไม่ใช่อาหารที่แท้จริง โดยหมวดสินค้าที่ไม่ใช่อาหารแบ่งออกเป็น 9 กลุ่ม ได้แก่ ค่าที่อยู่อาศัย ของใช้ในครัวเรือน ค่าจ้างผู้ให้บริการ ค่าเสื้อผ้า รองเท้า ค่าใช้จ่ายส่วนตัว ค่ารักษาพยาบาล การเดินทาง/การสื่อสาร และค่าใช้จ่ายด้านการศึกษา (สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2566)

2.2.4 การประหยัดจากขนาด

การประหยัดจากขนาด หมายถึง การอยู่อาศัยร่วมกันในครัวเรือนสามารถช่วยลดต้นทุนการดำรงชีวิตได้ โดยเฉพาะค่าใช้จ่ายในหมวดสินค้าที่ไม่ใช่อาหาร เช่น ค่าน้ำ ค่าไฟ ค่ารักษาพยาบาล หรือค่าซื้อของใช้บางอย่างที่สามารถใช้ร่วมกันได้ โดยไม่จำเป็นต้องซื้อแยกตามจำนวนคน เช่น ค่าเช่าบ้านไม่เพิ่มขึ้นตามจำนวนคนที่อยู่ (สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2566)

2.2.5 วิธีการคำนวณเส้นความยากจน (Poverty Line)

ขั้นตอนการคำนวณเส้นความยากจนจะใช้เริ่มต้นด้วยการคำนวณเส้นความยากจนด้านอาหาร แล้วจึงเส้นความยากจนในหมวดหมู่ที่ไม่ใช่อาหาร จากนั้นนำค่าทั้งสองมารวมกันจึงได้เป็นเส้นความยากจน มีขั้นตอนการคำนวณ ดังนี้



รูปที่ 2.1 ขั้นตอนและเทคนิคการคำนวณเส้นความยากจน
ที่มา: สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ (2566)

บทสรุปของการทบทวนวรรณกรรม เรื่อง แนวคิดและทฤษฎีเกี่ยวกับการแบ่งเส้นความยากจน เพื่อการวิเคราะห์ประเด็นสำคัญเข้าสู่งานวิจัยในครั้งนี้

เส้นความยากจน คือ เกณฑ์ค่าใช้จ่ายขั้นต่ำที่จำเป็นต่อการดำรงชีวิตอย่างมีคุณภาพ ซึ่งประกอบด้วยค่าใช้จ่ายด้านอาหารและสินค้าที่ไม่ใช่อาหาร โดยคำนวณจากต้นทุนโภชนาการพื้นฐานที่ปรับตามพฤติกรรมการบริโภคและค่าครองชีพในแต่ละพื้นที่ รวมถึงการพิจารณาการประหยัดจากขนาดครัวเรือนที่ใช้ทรัพยากรร่วมกันได้ เพื่อให้สะท้อนภาระค่าครองชีพที่แท้จริงอย่างแม่นยำ ในปัจจุบันเส้นความยากจนเฉลี่ยของประเทศไทยอยู่ที่ 3,043 บาทต่อคนต่อเดือน

2.2 ปัจจัยที่เกี่ยวข้องกับความยากจน

ปัจจัยที่เกี่ยวข้องกับความยากจนมีความซับซ้อนและเชื่อมโยงกันในหลายมิติ โดย**ปัจจัยด้านลักษณะส่วนบุคคล** เช่น อายุ เพศ ระดับการศึกษา และขนาดครัวเรือน ส่งผลต่อโอกาสในการหารายได้และการเข้าถึงทรัพยากร **ปัจจัยด้านเศรษฐกิจ** สะท้อนการกระจายรายได้ในสังคม ซึ่งกำหนดโครงสร้างโอกาสทางเศรษฐกิจโดยรวม ส่วน**ปัจจัยด้านสภาพแวดล้อมการทำงานและค่าตอบแทน** ครอบคลุมประเภทอาชีพ ความมั่นคงในการจ้างงาน ระดับค่าจ้าง และสวัสดิการจากการทำงาน ปัจจัยทั้ง 3 กลุ่มนี้มีปฏิสัมพันธ์และเสริมกันในการกำหนดสถานะความยากจนของบุคคลและครัวเรือน

2.2.1 แนวคิดปัจจัยส่วนบุคคล

ปัจจัยส่วนบุคคลเป็นลักษณะเฉพาะตัวของแต่ละบุคคลหรือครัวเรือนที่ส่งผลต่อขีดความสามารถในการสร้างรายได้และการเข้าถึงโอกาสทางเศรษฐกิจ มีแนวคิดที่เกี่ยวข้อง ดังนี้

สุพนิดา (2565) ได้ศึกษาความยากจนในผู้สูงอายุไทย : การเปลี่ยนแปลงและปัจจัยเสี่ยง จะมีปัจจัยต่างๆ ที่สำคัญ ดังนี้

อายุ เป็นปัจจัยสำคัญที่สัมพันธ์กับความยากจนในผู้สูงอายุ โดยพบแนวโน้มที่ชัดเจนว่ายิ่งมีอายุมากขึ้น โอกาสเผชิญกับความยากจนก็ยิ่งสูงขึ้น ผู้สูงอายุที่มีอายุ 70 ปีขึ้นไปมีความเสี่ยงต่อ

ความยากจนสูงกว่ากลุ่มอายุ 60-69 ปีอย่างมีนัยสำคัญ ทั้งนี้อาจเนื่องมาจากปัจจัยหลายประการ เช่น การเสื่อมถอยของสมรรถภาพทางร่างกาย ความสามารถในการทำงานหารายได้ที่ลดลง รวมถึงค่าใช้จ่ายด้านสุขภาพที่เพิ่มสูงขึ้นตามวัย

เพศ การศึกษาแสดงให้เห็นว่าผู้สูงอายุเพศหญิงมีโอกาสประสบกับความยากจนมากกว่าเพศชาย ปรากฏการณ์นี้อาจเกิดจากความเหลื่อมล้ำทางเพศที่สะสมมาตลอดชีวิต เช่น โอกาสทางการศึกษาที่น้อยกว่า การได้รับค่าตอบแทนที่ต่ำกว่า

จำนวนสมาชิกในครัวเรือน มีความสัมพันธ์เชิงบวกกับความยากจน กล่าวคือ ครัวเรือนที่มีสมาชิกอาศัยอยู่มากมักมีโอกาสเผชิญกับความยากจนสูงกว่า โดยเฉพาะในกรณีที่มีสัดส่วนของผู้พึ่งพิง (เด็กและผู้สูงอายุ) มากกว่าสมาชิกวัยแรงงาน ทำให้ภาระทางเศรษฐกิจตกอยู่กับผู้หารายได้น้อยราย

ระดับการศึกษา การศึกษาเป็นปัจจัยป้องกันความยากจนที่สำคัญ ผลการวิจัยพบว่าผู้สูงอายุยากจนส่วนใหญ่เป็นผู้ที่ไม่ได้รับการศึกษาหรือได้รับการศึกษาน้อย ในขณะที่ผู้สูงอายุที่จบการศึกษาระดับประถมศึกษามีโอกาสยากจนน้อยกว่าผู้ที่ไม่ได้รับการศึกษาอย่างมีนัยสำคัญ และยิ่งมีระดับการศึกษาสูงขึ้นเท่าใด โอกาสในการเผชิญกับความยากจนก็ยิ่งลดลง

สถานภาพสมรส มีผลต่อความเสี่ยงในการเผชิญกับความยากจนของผู้สูงอายุ จากการศึกษาพบว่า ผู้สูงอายุส่วนมากมีสถานภาพเป็นหม้าย หย่า หรือแยกกันอยู่ รองลงมาคือสถานภาพโสด และสมรสตามลำดับ ผลการวิเคราะห์ชี้ให้เห็นว่าผู้สูงอายุที่มีคู่สมรสหรือเคยสมรสแล้วหย่าร้างมักมีความเสี่ยงต่อความยากจนน้อยกว่าผู้ที่เป็นโสด

เขตที่อยู่อาศัย ผลการวิจัยแสดงให้เห็นว่าพื้นที่ที่อยู่อาศัยเป็นปัจจัยสำคัญที่สัมพันธ์กับความยากจนในผู้สูงอายุ โดยผู้สูงอายุส่วนใหญ่ (ร้อยละ 58.9) อาศัยอยู่นอกเขตเทศบาล หรือในพื้นที่ชนบท และมีแนวโน้มเผชิญกับความยากจนมากกว่าผู้ที่อาศัยอยู่ในเขตเทศบาล

ชูชิต (2561) กล่าวว่าความยากจนเกิดจากปัจจัยที่หลากหลายและมีความเชื่อมโยงกันในหลายระดับ ทั้งระดับพื้นที่ ชุมชน ครัวเรือน และบุคคล โดยเฉพาะในระดับครัวเรือนนั้น มีปัจจัยสำคัญที่ส่งผลต่อความยากจน ได้แก่ การมีหัวหน้าครัวเรือนที่มีอายุมาก ซึ่งอาจมีข้อจำกัดในการสร้างรายได้ การมีจำนวนสมาชิกที่ต้องดูแลมาก กรณีที่หัวหน้าครัวเรือนเป็นเพศหญิงซึ่งอาจเผชิญข้อจำกัดในการเข้าถึงโอกาสทางเศรษฐกิจและทรัพยากรบางประการ รวมถึงปัญหาภาวะหนี้สินที่เพิ่มขึ้น ในขณะที่รายรับกลับลดลงและรายจ่ายเพิ่มสูงขึ้น ส่งผลให้เกิดความเปราะบางทางเศรษฐกิจและนำไปสู่วงจรความยากจนที่ยากจะหลุดพ้น ในขณะที่ปัจจัยระดับบุคคลมีบทบาทสำคัญต่อสภาวะความขัดสน โดยเฉพาะการมีระดับการศึกษาที่ต่ำซึ่งจำกัดโอกาสในการเข้าถึงอาชีพที่มั่นคงและรายได้ที่เพียงพอ บุคคลที่ขาดโอกาสทางการศึกษามักประสบกับความยากไร้อย่างต่อเนื่อง เช่นเดียวกับผู้ที่ไม่ได้รับการฝึกอบรมทักษะทางวิชาชีพที่จำเป็น ทำให้ไม่สามารถแข่งขันในตลาดแรงงานที่เปลี่ยนแปลงอย่างรวดเร็วได้ นอกจากนี้ พฤติกรรมส่วนบุคคล เช่น ความเกียจคร้านในการประกอบอาชีพ การติดเครื่องดื่มแอลกอฮอล์จนสูญเสียเงินทองและความสามารถในการทำงาน การหมกมุ่นกับการพนันซึ่งนำไปสู่การสูญเสียทรัพย์สินอย่างรวดเร็ว การใช้จ่ายฟุ่มเฟือยโดยขาดการประหยัดอดออม และการก่อหนี้สินพอกพูนจนเกินความสามารถในการชำระคืน ล้วนเป็นปัจจัยที่ส่งผลให้บุคคลดังกล่าวยากจน ยากที่จะยกระดับคุณภาพชีวิตให้ดีขึ้นได้

นิภาพรรณ (2566) นิยามลักษณะของคนจน ไว้ดังนี้ โดยพิจารณาจากความเป็นอยู่และสาเหตุของความจน สามารถแบ่งออกเป็น 3 ประเภท ดังนี้

ประเภทที่ 1 คนจนพื้นฐานหรือคนจนเชิงกายภาพ คือ คนที่ขาดปัจจัยพื้นฐานในการดำรงชีวิต มักเป็นแรงงานที่มีทักษะขั้นพื้นฐาน หรือเป็นแรงงานรับจ้างหมุนเวียนในภาคการเกษตรหรือรับจ้างทั่วไป หรือเป็นแรงงานที่มีอายุน้อยไม่ได้รับการศึกษา ได้รับค่าจ้างขั้นต่ำแต่มีภาระค่าใช้จ่ายที่ต้องรับผิดชอบมาก เช่น มีจำนวนสมาชิกในครัวเรือนตั้งแต่ 3 คนขึ้นไป ซึ่งอยู่ภายใต้การดูแล หรือต้องดูแลเด็ก คนชรา ผู้พิการ

ประเภทที่ 2 คนเสี่ยงจน คนใกล้จน คนเกือบจน คือ คนที่มีโอกาสกลายเป็นคนจนได้อย่างง่าย โดยมีสาเหตุมาจากปัจจัยภายนอก เช่น อุบัติภัย การเสียชีวิตของหัวหน้าครัวเรือน โดยคนกลุ่มนี้มักเป็นเกษตรกรที่พอมีที่ดินทำกินแต่มีหนี้สิน ลูกจ้างที่มีรายได้พอสมควร แต่ไม่มั่นคง แรงงานที่ไม่มีทักษะแรงงานด้านเทคโนโลยีสมัยใหม่ เป็นแรงงานที่ทำงานโดยมีความเสี่ยงต่ออุบัติเหตุสูง

ประเภทที่ 3 คนจนเชิงเปรียบเทียบ คนจนเชิงสัมพันธ์ คนจนเชิงโครงสร้าง คือ กลุ่มคนจนกลุ่มนี้ไม่ได้ยากจนในเชิงขาดแคลนปัจจัยพื้นฐาน แต่ขาดโอกาสในการเข้าถึงข้อมูลข่าวสาร เทคโนโลยีและประโยชน์ที่ควรได้รับจากการพัฒนาต่าง ๆ

เอลวิส และจุฬาพรรณภรณ์ (2565) กล่าวว่า ทุนมนุษย์ คือ ความรู้ ทักษะ ความสามารถ ความชำนาญ และประสบการณ์ที่แต่ละคนสั่งสมไว้ในตนเอง ซึ่งสามารถรวมกันกลายเป็นศักยภาพขององค์กรหรือประเทศ ซึ่งทุนมนุษย์เป็นทรัพย์สินเฉพาะตัวของแต่ละคน ไม่สามารถโอนย้ายได้ ยิ่งเรียนรู้และสะสมประสบการณ์มาก ยิ่งเพิ่มมูลค่าให้ตัวเอง ส่งผลต่อโอกาสในอาชีพ รายได้ และการเลื่อนตำแหน่ง การมีทุนมนุษย์สูง ทำให้สามารถดำรงชีวิตและปรับตัวในยุคที่เปลี่ยนแปลงได้ดี

สุรศักดิ์ (2560) กล่าวว่า ทุนมนุษย์ หรือ Human Capital คือ คุณลักษณะต่างๆ และความสามารถที่อยู่ในตัวมนุษย์ เช่น ความรู้ ทักษะ ความชำนาญ ความสามารถ และประสบการณ์ ซึ่งเป็นประโยชน์ต่อการเพิ่มผลิตภาพงาน ซึ่งการพัฒนาทุนมนุษย์ส่งผลให้องค์กรมีความสามารถทางการแข่งขัน เนื่องจากมีทุนมนุษย์ที่มีคุณภาพ

สถาบันวิจัยเพื่อความเสมอภาคทางการศึกษา (2566) ทุนมนุษย์ วัดผลจากระดับการศึกษา รายได้ ระดับความซับซ้อนของงานที่ทำ และประสบการณ์ทำงาน โดยที่การมีทุนมนุษย์ต่ำ โดยเฉพาะในกลุ่มเด็กและเยาวชนจากครอบครัวยากจน เป็นตัวเร่งสำคัญที่ทำให้พวกเขาไม่สามารถหลุดพ้นจากความยากจนข้ามรุ่นได้ แม้จะจบการศึกษาในระดับหนึ่ง หากไม่มีทักษะชีวิตและความสามารถในการปรับตัว ก็จะไม่มีโอกาสเข้าถึงงานที่มีคุณภาพ รายได้ก็จะไม่พอเลี้ยงดูครอบครัว นำไปสู่กับดักความยากจนที่ส่งต่อไปยังรุ่นถัดไป

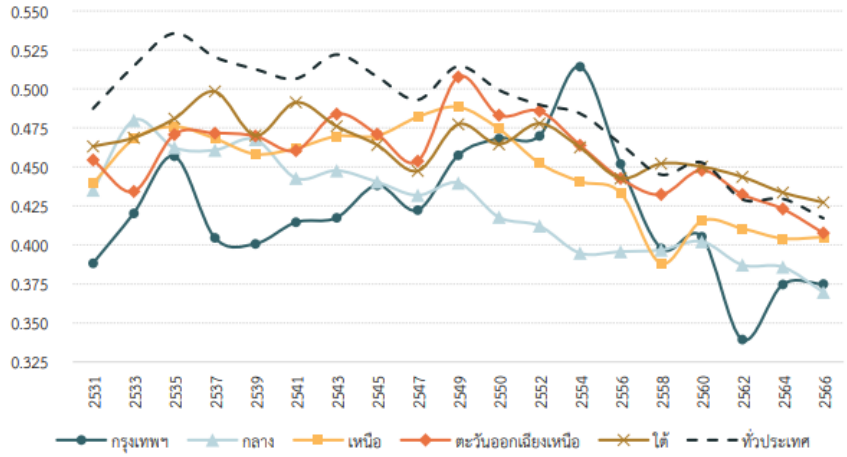
บทสรุปการทบทวนวรรณกรรม เรื่อง แนวคิดปัจจัยส่วนบุคคล เพื่อวิเคราะห์ประเด็นสำคัญเข้าสู่งานวิจัย ครั้งนี้

ปัจจัยส่วนบุคคลหลายประการมีความสัมพันธ์กับสถานะความยากจน เช่น เพศ อายุ สถานภาพสมรส ระดับการศึกษา พื้นที่อยู่อาศัย อาชีพ และจำนวนสมาชิกในครัวเรือน ลักษณะบุคคลที่มีโอกาสสูงในการตกอยู่ในภาวะความยากจน ได้แก่ เพศหญิง ผู้มีสถานภาพโสด ผู้สูงอายุ ผู้อาศัยนอกเขตเทศบาล และผู้มีการศึกษาน้อยหรือไม่ได้รับการศึกษา ซึ่งมักขาดทักษะวิชาชีพที่จำเป็น ทำให้เข้าถึงอาชีพที่มั่นคงได้ยากขึ้น ลักษณะเหล่านี้ล้วนเพิ่มความเสี่ยงต่อการเผชิญภาวะความยากจน นอกจากนี้ ครัวเรือนที่มีสมาชิกจำนวนมากยังมีแนวโน้มเผชิญภาวะความยากจนสูงขึ้นเช่นกัน ทั้งนี้ นอกเหนือจากปัจจัยส่วนบุคคลแล้ว ความเสี่ยงจากการประกอบอาชีพยังเป็นตัวแปรสำคัญในการจำแนกกลุ่มประชากรที่มีความยากจนอีกด้วย

2.2.2 ปัจจัยด้านเศรษฐกิจ

ดัชนีความก้าวหน้าของคน (Human Achievement Index :HAI) เป็นดัชนีที่ใช้วัดความก้าวหน้าของคนในระดับจังหวัด โดยครอบคลุม 8 ด้านสำคัญ เช่น รายได้ การศึกษา สุขภาพ และคุณภาพชีวิต ซึ่งหนึ่งในองค์ประกอบของ HAI คือ ข้อมูลด้านรายได้และความเหลื่อมล้ำ โดยดัชนีย่อยด้านรายได้เป็นดัชนีสำคัญที่สะท้อนให้เห็นถึงสถานะด้านการเงินหรือด้านเศรษฐกิจของคน ซึ่งถ้ามีรายได้เพียงพอต่อใช้จ่ายเพื่ออุปโภค บริโภคในสิ่งจำเป็นต่อการดำเนินชีวิตจะส่งผลให้คุณภาพชีวิตดีขึ้นและมีภูมิคุ้มกันต่อความยากจน (นิภาพรรณ, 2566) นอกจากดัชนีย่อยด้านรายได้ยังรวมถึงค่าดัชนีของความไม่เสมอภาคด้าน (Gini Coefficient) ด้วย ดังนั้น ดัชนี HAI จึงสามารถสะท้อนความยากจนในมิติต่าง ๆ ได้อย่างครอบคลุมมากกว่าการพิจารณาเฉพาะรายได้ ดังนั้น เมื่อใช้ HAI ควบคู่กับ Gini จะทำให้เข้าใจทั้ง "ช่องว่างของรายได้" และ "โอกาสในการเข้าถึงบริการพื้นฐาน" ของประชาชน โดยเฉพาะกลุ่มเปราะบาง ช่วยให้การกำหนดนโยบายลดความยากจนเป็นไปอย่างแม่นยำและรอบด้านมากยิ่งขึ้น

ดัชนีของความไม่เสมอภาคด้านรายได้ (Gini Coefficient หรือ Gini Index) เป็นตัวชี้วัดระดับความไม่เท่าเทียมในการกระจายรายได้ของประชากร โดยมีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งค่าที่ใกล้ 0 หมายถึง การกระจายรายได้อย่างเท่าเทียม ส่วนค่าที่ใกล้ 1 หมายถึงความเหลื่อมล้ำสูง การที่ค่า Gini อยู่ในระดับสูงสะท้อนว่ารายได้กระจุกตัวอยู่กับคนกลุ่มหนึ่ง ทำให้กลุ่มคนรายได้น้อยมีโอกาสหลุดพ้นจากความยากจนได้ยาก ในขณะที่ค่า Gini ที่ลดลงเป็นสัญญาณว่าการกระจายรายได้ดีขึ้น ซึ่งมักสัมพันธ์กับการลดลงของความยากจน



รูปที่ 2.2 ความเหลื่อมล้ำด้านรายได้ จำแนกรายภาค ปี พ.ศ. 2531 – 2566
ที่มา: สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ (2566)

จากรูปจะเห็นว่าความเหลื่อมล้ำด้านรายได้ ซึ่งวัดด้วยดัชนี Gini Coefficient ในแต่ละภาคมีความแตกต่างกัน โดยภูมิภาคที่มีความเหลื่อมล้ำด้านรายได้สูงกว่าภาคอื่นๆ ได้แก่ ภาคใต้ รองลงมาคือภาคตะวันออกเฉียงเหนือ ซึ่งทั้งสองภูมิภาคนี้มีจำนวนคนจนสูงกว่าภูมิภาคอื่นๆ แต่ในภาพรวมความเหลื่อมล้ำด้านรายได้ของประเทศไทยมีแนวโน้มลดลง

บทสรุปการทบทวนวรรณกรรม เรื่อง ปัจจัยด้านเศรษฐกิจ เพื่อวิเคราะห์ประเด็นสำคัญเข้าสู่งานวิจัย ครั้งนี้

ดัชนีของความไม่เสมอภาคด้านรายได้ (Gini Coefficient) สะท้อนระดับความเหลื่อมล้ำของรายได้ หากค่า Gini สูง แสดงว่ารายได้กระจุกตัว คนจนจึงมีโอกาสหลุดพ้นจากความยากจนได้ยากขึ้น ดัชนีความก้าวหน้าของคน (Human Achievement Index : HAI) วัดความก้าวหน้าด้านคุณภาพชีวิต เช่น การศึกษา สุขภาพ รายได้ หากค่า HAI ต่ำ แสดงว่าคนยังขาดโอกาสพัฒนาในหลายด้าน ซึ่งเป็นปัจจัยสำคัญที่ทำให้ความยากจนยั่งยืน

2.2.3 ปัจจัยด้านสภาพการทำงานและค่าตอบแทน

ภัทรียา (2559) กล่าวว่านโยบายการปรับเพิ่มอัตราค่าจ้างขั้นต่ำแม้จะมีเป้าหมายเพื่อยกระดับคุณภาพชีวิตแรงงานในระบบ แต่กลับส่งผลกระทบทางลบต่อแรงงานนอกระบบซึ่งเป็นกลุ่มใหญ่และเป็นกลุ่มคนจนของประเทศ โดยแรงงานนอกระบบไม่ได้รับประโยชน์จากค่าจ้างที่เพิ่มขึ้นตามนโยบาย ขณะเดียวกันกลับต้องเผชิญกับการลดลงของค่าจ้างจากแรงงานในระบบที่ถูกแทนที่และไหลเข้าสู่ตลาดแรงงานนอกระบบ ทำให้อุปทานแรงงานในกลุ่มนี้เพิ่มขึ้นและกดทับค่าจ้าง นอกจากนี้ การปรับขึ้นค่าจ้างขั้นต่ำยังส่งผลให้ราคาสินค้าและต้นทุนการผลิตเพิ่มสูงขึ้น ส่งผลให้แรงงานนอกระบบและครัวเรือนที่มีรายได้น้อยต้องเผชิญกับภาระค่าใช้จ่ายที่สูงขึ้น โดยไม่มีรายได้เพิ่มขึ้นตาม จึงสะท้อนว่านโยบายดังกล่าวไม่ได้ช่วยลดความยากจนหรือความเหลื่อมล้ำได้อย่างมีประสิทธิภาพสำหรับแรงงานกลุ่มนี้

ผกามาศ (2561) ได้ศึกษาค่าจ้างและความต้องการทำงานเพิ่มของแรงงานนอกระบบ: กรณีศึกษาประเทศไทย พบว่าแรงงานนอกระบบที่ยากจนมีแนวโน้มตอบสนองต่อการเพิ่มค่าจ้างอย่างชัดเจน โดยมักเลือกเพิ่มชั่วโมงทำงานหรือเข้าสู่ตลาดแรงงานมากขึ้น โดยเฉพาะกลุ่มที่ได้รับค่าจ้างต่ำและผู้ที่ได้รับค่าตอบแทนแบบรายชั่วโมงหรือรายวัน ซึ่งสะท้อนแนวคิด "Instant Gratification" ที่ต้องการเห็นผลตอบแทนที่รวดเร็ว ขณะที่แรงงานที่มีทักษะสูงก็แสดงความต้องการทำงานเพิ่มมากกว่าแรงงานทักษะต่ำ แต่หากต้องเผชิญปัญหาด้านความปลอดภัยหรือสภาพแวดล้อมการทำงานที่ไม่เหมาะสม ความต้องการทำงานจะลดลง พฤติกรรมการพยายามทำงานเพิ่มของกลุ่มแรงงานนอกระบบเหล่านี้สะท้อนถึงความจำเป็นในการดิ้นรนเพื่อให้มีรายได้เพียงพอ แสดงให้เห็นถึงความพยายามในการยกระดับคุณภาพชีวิตและความต้องการหลุดพ้นจากวงจรความยากจน

กุศล (2566) กล่าวว่าประเทศไทยเคยใช้การพัฒนาอุตสาหกรรมเพื่อดึงแรงงานจากภาคเกษตร ช่วยยกระดับคนจนชนบทสู่ชนชั้นกลาง แต่กระบวนการนี้เริ่มอิ่มตัวช่วงกลางทศวรรษ 2550 โดยแรงงานในภาคเกษตรยังคงอยู่ที่ประมาณ 1 ใน 3 ซึ่งสูงกว่าประเทศที่มีระดับการพัฒนาใกล้เคียงกัน ขณะที่สัดส่วนแรงงานในภาคอุตสาหกรรมไทยอิ่มตัวที่เพียง 17% และมีแนวโน้มลดลง ต่างจากประเทศที่ประสบความสำเร็จอย่างเกาหลีใต้และไต้หวันที่สามารถสร้างงานในภาคอุตสาหกรรมสัดส่วนไม่ต่ำกว่า 1 ใน 3 สาเหตุสำคัญมาจากแรงดึงดูดนอกภาคเกษตรไม่เพียงพอ นโยบายช่วยเหลือเกษตรกรที่ทำให้การทำงานในภาคเกษตรมีความมั่นคง และนโยบายพัฒนาเศรษฐกิจ "รอบด้าน" ของรัฐบาลที่สนับสนุนภาคการผลิตอื่นๆ รวมถึงการขาดความต่อเนื่องในนโยบายพัฒนาอุตสาหกรรม ส่งผลให้ตัวเลขความยากจนอยู่ในระดับคงตัว ประเทศไทยติดอยู่ในกับดักรายได้ปานกลาง และผลิตภาพในภาคเกษตรยังคงอยู่ในระดับต่ำ ทำให้ไม่สามารถยกระดับสู่ประเทศรายได้สูงได้เหมือนประเทศที่มีภาคอุตสาหกรรมที่แข็งแกร่งกว่า โดยประเทศที่พัฒนาแล้วมักมีภาคอุตสาหกรรมที่แข็งแกร่ง

สามารถดึงดูดแรงงานจากภาคเกษตรได้มากกว่า ซึ่งช่วยลดความยากจนและยกระดับรายได้ประชาชาติ ประเทศไทยจำเป็นต้องเร่งพัฒนาภาคอุตสาหกรรมที่อย่างมีเป้าหมายและต่อเนื่อง เพื่อเพิ่มโอกาสในการหลุดพ้นจากกับดักรายได้ปานกลาง

บทสรุปการทบทวนวรรณกรรม เรื่อง ปัจจัยด้านสภาพการทำงานและค่าตอบแทน เพื่อวิเคราะห์ประเด็นสำคัญเข้าสู่งานวิจัย ครั้งนี้

ปัจจัยที่มีอิทธิพลต่อความยากจนของแรงงาน ประกอบด้วย ประเภทของค่าจ้างที่ได้รับ โดยแรงงานที่ได้รับค่าจ้างรายวันหรือรายชั่วโมงมีแนวโน้มต้องการทำงานเพิ่มเพื่อให้มีรายได้เพียงพอในขณะที่ค่าจ้างขั้นต่ำที่แม้มีเป้าหมายยกระดับคุณภาพชีวิตแต่กลับส่งผลลบต่อแรงงานนอกระบบที่ต้องเผชิญต้นทุนชีวิตสูงขึ้นโดยไม่ได้รับประโยชน์ จำนวนชั่วโมงการทำงานที่เพิ่มขึ้นสะท้อนภาวะดิ้นรนเพื่อหลุดพ้นความยากจน และภาคอุตสาหกรรมที่ไม่พัฒนาอย่างต่อเนื่องทำให้ดูดซับแรงงานจากภาคเกษตรได้ไม่เพียงพอ ทำให้ความยากจนของแรงงานนอกระบบยังคงอยู่ในระดับเดิม นอกจากนี้ปัญหาสภาพแวดล้อมการทำงาน และความเสี่ยงจากการทำงานยังเป็นอุปสรรคสำคัญที่ทำให้แรงงานลดความต้องการทำงานลง ส่งผลต่อการสร้างรายได้และการยกระดับคุณภาพชีวิตในระยะยาว

2.3 วิธีการจัดการข้อมูลที่ไม่สมดุล

ชุดข้อมูลไม่สมดุล คือ ชุดข้อมูลที่มีการกระจายของกลุ่มหรือคลาสอย่างไม่เท่าเทียมกัน โดยที่จำนวนตัวอย่างในบางคลาสมีมากหรือน้อยกว่าคลาสอื่นอย่างชัดเจน ส่งผลให้แบบจำลองทางสถิติหรือการเรียนรู้ของเครื่องมีแนวโน้มที่จะเรียนรู้จากกลุ่มที่มีจำนวนมากกว่า และละเลยกลุ่มที่มีจำนวนน้อยกว่า ซึ่งอาจลดความแม่นยำของการทำนาย โดยเฉพาะในคลาสที่มีจำนวนน้อย ซึ่งปัจจุบันมีวิธีการจัดการข้อมูลที่ไม่สมดุล 4 วิธี ดังนี้ (อัจฉรา และสายชล, 2562)

1) วิธีการสุ่มเกิน (Over Sampling) คือ การสุ่มข้อมูลในคลาสที่มีจำนวนข้อมูลน้อยให้จำนวนข้อมูลในคลาสเพิ่มขึ้นใกล้เคียงหรือเท่ากับจำนวนข้อมูลในคลาสส่วนมาก โดยใช้วิธีการสุ่มตัวอย่างอย่างง่าย

2) วิธีการสุ่มลด (Under Sampling) คือ การปรับข้อมูลให้มีความสมดุล โดยการสุ่มลดจำนวนข้อมูลในคลาสส่วนมากให้ใกล้เคียงกับจำนวนข้อมูลในคลาสส่วนน้อย

3) วิธีการสุ่มแบบผสมผสาน (Hybrid Method) คือ วิธีการปรับสมดุลข้อมูลของทั้งสองคลาสให้มีจำนวนใกล้เคียงกัน โดยนำวิธีสุ่มเกินและสุ่มลดมาทำงานงานร่วมกัน โดยใช้วิธีสุ่มเกินในข้อมูลที่เป็นคลาสส่วนน้อย และใช้วิธีสุ่มลดในข้อมูลคลาสส่วนมาก

4) วิธีการสุ่มโดยใช้เทคนิค Synthetic Minority Oversampling Technique (SMOTE) คือ การสุ่มสร้างข้อมูลจากคลาสที่มีจำนวนข้อมูลน้อยให้มีจำนวนตามที่กำหนด โดยการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมใช้หลักการเพื่อนบ้านที่อยู่ใกล้ที่สุด ในการขยายขอบเขตการตัดสินใจของตัวแบบ (ภาณุภณ, 2564)

2.4 แบบจำลองการจำแนกประเภทของข้อมูล

แบบจำลองการจำแนกประเภท (Classification) คือ เทคนิคในการทำเหมืองข้อมูล (Data Mining) และการเรียนรู้ของเครื่อง (Machine Learning) ที่มีเป้าหมายในการทำนายหมวดหมู่ของข้อมูลใหม่โดยอิงจากการสังเกตในอดีต โดยมีเป้าหมายคือ การเรียนรู้แบบจำลองจากข้อมูลฝึกหัดที่สามารถทำนายคลาสของข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างแม่นยำ (GeeksforGeeks, 2023) การจำแนกประเภท

เป็นประเภทของการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ใช้อัลกอริทึมต่างๆ เพื่อสร้างแบบจำลองที่สามารถแยกแยะและจัดกลุ่มข้อมูลเข้าสู่หมวดหมู่ที่กำหนดไว้ล่วงหน้า

2.4.1 การวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติ

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นเทคนิคทางสถิติที่ใช้เมื่อตัวแปรตามเป็นข้อมูลเชิงคุณภาพ มีวัตถุประสงค์เพื่อสะท้อนความสัมพันธ์ระหว่างตัวแปรอิสระกับโอกาสที่เหตุการณ์จะเกิดขึ้น

สำหรับการวิเคราะห์การถดถอยลอจิสติกแบบทวิ (Binary Logistic Regression) ความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้น กรณีที่ตัวแปรอิสระมีมากกว่า 1 ตัว สามารถแสดงได้ดังสมการ (อรทัย, 2560)

$$P(y) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}} \quad (2.1)$$

เมื่อ P(y) คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (y=1)

Q(y) คือ ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ (y=0) หรือเท่ากับ 1 - P(y)

โดยทั่วไป หาก P(y) ≥ 0.5 สรุปว่าเกิดเหตุการณ์ที่สนใจ ในทางตรงกันข้ามจะสรุปว่าไม่เกิดเหตุการณ์ที่สนใจ สำหรับ 0.5 เป็นค่าความน่าจะเป็นที่ใช้เป็น Threshold ซึ่งอาจกำหนดเป็นค่าอื่นได้ แต่นิยมใช้เป็น 0.5

จากความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามที่ไม่เป็นเชิงเส้น จึงเปลี่ยนเป็นรูปแบบเชิงเส้นโดยการทำ logit transformation หรือ logit(Odds) ซึ่ง Odds หรือ Odds Ratio คือ โอกาสการเกิดเหตุการณ์ที่สนใจ (y=1) มีอัตราส่วนเป็นกี่เท่าของโอกาสที่ไม่สนใจ (y=0) ดังนี้

$$odds = \frac{P(y)}{Q(y)} = e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p} \quad (2.2)$$

$$log(odds) = log\left(\frac{P(y)}{Q(y)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p \quad (2.3)$$

สำหรับการประมาณค่าสัมประสิทธิ์การถดถอยของสมการถดถอยลอจิสติกจะใช้วิธีภาวะความน่าจะเป็นสูงสุด (Maximum Likelihood)

1) การทดสอบความสัมพันธ์ระหว่างตัวแปร

1.1) การวิเคราะห์ระดับความสัมพันธ์ด้วย Cramér's V

หากตัวแปรเชิงกลุ่มทั้งสองตัวแปรเป็นมาตรวัดนามบัญญัติ หรือตัวใดตัวหนึ่งเป็นตัวแปรเชิงอันดับ สามารถใช้ Cramér's V เพื่อวัดขนาดของความสัมพันธ์ (สำเริง, 2564) โดยใช้ร่วมกับการตรวจสอบความสัมพันธ์ระหว่างตัวแปรด้วยสถิติทดสอบ Chi-Square หากพบว่าตัวแปรทั้งสองมีความสัมพันธ์ต่อกัน สามารถใช้ Cramér's V ซึ่งเป็นสถิติที่ใช้วัดความสัมพันธ์ โดยใช้ค่า Chi-Square

เป็นฐานในการคำนวณ ซึ่งเหมาะสำหรับการวิเคราะห์ความสัมพันธ์ในตารางไขว้ (Contingency Table) ที่มีขนาดตั้งแต่ 2×2 ขึ้นไป โดยค่าของ Cramér's V จะอยู่ในช่วง 0 ถึง 1

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n(df_{smaller})}} \tag{2.4}$$

โดยที่ χ^2 คือ ค่าสถิติไคสแควร์

n คือ ขนาดตัวอย่าง

$df_{smaller}$ คือ องศาความเป็นอิสระของแถวหรือองศาความเป็นอิสระของคอลัมน์ ซึ่งเป็นค่าที่น้อยที่สุด

ตารางที่ 2.1 การแปลความหมายค่าสัมประสิทธิ์ของ Cramer's V

ค่า V	การแปลผลระดับความสัมพันธ์
0 - 0.10	ไม่มีความสัมพันธ์
0.10 - 0.20	มีความสัมพันธ์เพียงเล็กน้อย
0.20 - 0.40	มีความสัมพันธ์ปานกลาง
0.40 - 0.60	มีความสัมพันธ์ค่อนข้างมาก
0.60 - 0.80	มีความสัมพันธ์มาก
0.80 - 1.0 0	มีความสัมพันธ์อย่างมาก

1.2) Point Biserial correlation

Point Biserial Correlation เป็นเทคนิคทางสถิติที่ใช้สำหรับการวัดความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพที่มี 2 กลุ่มแบบตามธรรมชาติ (Dichotomous) กับตัวแปรเชิงปริมาณ (Interval scale หรือ Ratio scale) (Lee, 2025) ซึ่งเป็นกรณีเฉพาะของ Pearson's correlation โดยค่าความสัมพันธ์จะแสดงอยู่ในช่วงระหว่าง -1 ถึง 1

สูตรการคำนวณ Point Biserial Correlation คือ

$$r_{pb} = \frac{\overline{X_1} - \overline{X_0}}{s_X} \cdot \sqrt{\frac{n_1 n_0}{n^2}} \tag{2.5}$$

โดยที่ $\overline{X}_1$ และ $\overline{X}_0$ คือ ค่าเฉลี่ยของกลุ่มที่มีค่าของตัวแปรเชิงคุณภาพเป็น 1 และ 0 ตามลำดับ

s_X คือ ค่าเบี่ยงเบนมาตรฐานรวมของตัวแปรเชิงปริมาณ

n_1 และ n_0 คือ จำนวนตัวอย่างในแต่ละกลุ่ม

n คือ จำนวนรวมของตัวอย่างทั้งหมด

2) การตรวจสอบความเหมาะสมของแบบจำลอง

2.1) พิจารณาความเป็นไปได้ (Likelihood Value) จะพิจารณาค่า -2LL (-2 log Likelihood) ถ้ามีค่าลดลง แสดงว่าแบบจำลองมีความเหมาะสมมากที่สุด ในการทดสอบนัยสำคัญความเหมาะสมของแบบจำลองจะใช้สถิติ χ^2-test

2.2) พิจารณาสถิติทดสอบความเหมาะสมด้วย Hosmer and Lemeshow ซึ่งใช้ทดสอบความเหมาะสมของแบบจำลอง ดังนี้

$$P(y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p)}} \qquad (2.6)$$

สมมติฐานในการทดสอบ คือ

H_0: model เหมาะสม

H_1: model ไม่เหมาะสม

ในการทดสอบ χ^2 ไม่มีนัยสำคัญทางสถิติหรือ ถ้า p-value > 0.05 จะไม่สามารถปฏิเสธ H_0 นั่นคือ model มีความเหมาะสม

3) การทดสอบนัยสำคัญของสัมประสิทธิ์การถดถอยลอจิสติก

3.1) สถิติทดสอบของวอลด์ (Wald Statistic) ใช้สำหรับทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัว

สมมติฐานในการทดสอบ คือ

H_0: $\beta_i = 0$; i =1,2,...,p

H_1: $\beta_i \neq 0$

สถิติทดสอบของวอลด์จะมีการแจกแจงแบบ χ^2 และ df = p-1

สถิติทดสอบ คือ $Wald = \left[\frac{b_i}{SE(b_i)} \right]^2$ (2.7)

ถ้า p-value < 0.05 จะปฏิเสธ H_0 นั่นคือ ค่าสัมประสิทธิ์เป็นบวก แสดงว่าตัวแปรอิสระนั้นมีผลต่อการเพิ่มความน่าจะเป็นในการเกิดเหตุการณ์ แต่ถ้าหากสัมประสิทธิ์เป็นลบ แสดงว่าตัวแปรอิสระนั้นมีผลต่อการลดความน่าจะเกิดเหตุการณ์

3.2) สถิติทดสอบความสัมพันธ์

3.2.1) สถิติทดสอบ Cox & Snell R square เป็นการตรวจสอบความสอดคล้องของแบบจำลอง หรือสัดส่วนความผันแปรในตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระ ซึ่งค่า

Cox & Snell R square จะมีค่าน้อยกว่า 1 เสมอ (กาญจน์เขจร, 2561) สามารถคำนวณได้ดังสมการที่ 2.8

$$R_{cs}^{2} = 1 - e^{\left[-\frac{2}{n}(LL(new) - LL(baseline))\right]} \quad (2.8)$$

โดยที่ LL(new) คือ ค่า Log Likelihood ของแบบจำลองกรณีที่มีตัวแปรอิสระ

LL(baseline) คือ ค่า Log Likelihood ของแบบจำลองกรณีที่ไม่มีตัวแปรอิสระ

3.2.2) สถิติทดสอบ NagelKerke R square เกิดจากการนำสถิติ Cox & Snell R Square หารด้วย Log Likelihood ซึ่งค่าที่ได้จะมีค่าอยู่ในช่วง 0-1 และมีค่ามากกว่า Cox & Snell R Square เสมอ

4) ข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยลอจิสติกแบบทวิ

4.1) ตัวแปรอิสระเป็นตัวแปรที่ระดับข้อมูลอยู่ในระดับช่วง กรณีที่เป็นตัวแปรเชิงกลุ่มต้องแปลงให้เป็นตัวแปรหุ่น (Dichotomous Variable) และตัวแปรตามกำหนด 2 ค่า คือ 0 กับ 1

4.2) ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์หรือไม่มีความสัมพันธ์กัน

4.3) ตัวแปรอิสระไม่มีความสัมพันธ์กัน หรือไม่เกิดปัญหา Multicollinearity (ยุทธ, 2555)

5) การแปลผลค่าสัมประสิทธิ์การถดถอย

5.1) การแปลผลจากค่าสัมประสิทธ์การถดถอย (Coefficient) จากแบบจำลองการวิเคราะห์การถดถอยลอจิสติกแสดงถึงทิศทางและความแรงของความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับความน่าจะเป็นที่ตัวแปรตามจะเป็นคลาสบวก (เช่น ยากจน = 1) โดยมีการแปลผลดังนี้

หากค่าสัมประสิทธิ์ เป็นบวก แสดงว่า เมื่อค่าของตัวแปรอิสระเพิ่มขึ้น 1 หน่วย ความน่าจะเป็นที่ค่าตัวแปรตามจะเป็น 1 จะเพิ่มขึ้น

หากค่าสัมประสิทธิ์ เป็นลบ แสดงว่า เมื่อค่าของตัวแปรอิสระเพิ่มขึ้น 1 หน่วย ความน่าจะเป็นที่ค่าตัวแปรตามจะเป็น 1 จะลดลง

หากค่าสัมประสิทธิ์ที่ใกล้ศูนย์ หมายถึง ตัวแปรนั้นมีผลกระทบต่อการทำนายสถานะของตัวแปรตามน้อยหรือแทบไม่มีเลย

5.2) การแปลผลในรูปแบบอัตราส่วนความเป็นไปได้ (Odds Ratio) โดยทั่วไปนิยมแปลงค่าสัมประสิทธิ์ให้อยู่ในรูปของค่า Odds Ratio ซึ่งได้จากการยกกำลังฐานธรรมชาติ (exp) ของค่าสัมประสิทธิ์

สำหรับตัวแปรเชิงปริมาณ จะมีการแปลผลของค่า Odds Ratio ดังนี้

- หาก Odds Ratio > 1 หมายถึง เมื่อค่าตัวแปรอิสระเพิ่มขึ้น 1 หน่วย โอกาสที่ตัวแปรตามจะเกิด (class = 1) จะเพิ่มขึ้น
- หาก Odds Ratio < 1 หมายถึง เมื่อค่าตัวแปรอิสระเพิ่มขึ้น 1 หน่วย โอกาสที่ตัวแปรตามจะเกิดจะลดลง
- หาก Odds Ratio = 1 หมายถึง ตัวแปรอิสระไม่มีผลต่อโอกาสของตัวแปรตาม

สำหรับตัวแปรเชิงกลุ่ม จะมีการแปลผลของค่า Odds Ratio แตกต่างกัน ดังนี้ (สำนักงานสถิติแห่งชาติ, 2562)

- หาก Odds Ratio > 1 หมายถึง แรงงานนอกระบบที่อยู่ในกลุ่มที่สนใจมีโอกาสที่จะยากจน มากกว่า กลุ่มอ้างอิงมากกว่า 1 เท่า (หรือร้อยละ (Odds Ratio - 1)*100)
- หาก Odds Ratio < 1 หมายถึง แรงงานนอกระบบที่อยู่ในกลุ่มที่สนใจมีโอกาสที่จะยากจนน้อยกว่ากลุ่มอ้างอิง 1 เท่า (คิดเป็นร้อยละ (1 - Odds Ratio)*100)
- หาก Odds Ratio = 1 หมายถึง แรงงานนอกระบบที่อยู่ในกลุ่มที่สนใจมีโอกาสยากจนเท่ากับกลุ่มอ้างอิง หรือการเปลี่ยนแปลงของตัวแปรเชิงกลุ่มนั้น ไม่มีผลต่อโอกาสยากจนอย่างมีนัยสำคัญ

2.4.2 การวิเคราะห์การถดถอยลอจิสติก ด้วยการเรียนรู้ของเครื่อง

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression) เป็นหนึ่งในอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) ที่ได้รับความนิยมอย่างแพร่หลาย โดยเฉพาะในปัญหาการจำแนกประเภทแบบทวิภาค (Binary Classification) เช่น การทำนายว่าผู้ใช้จะกดโฆษณาหรือไม่ หรือการจำแนกกลุ่มผู้ป่วยเป็นผู้ที่เสี่ยงหรือไม่เสี่ยงต่อโรค (กิตติมศักดิ์, 2564) แม้ชื่อของอัลกอริทึมจะมีคำว่า “Regression” แต่การถดถอยลอจิสติกไม่ได้ใช้สำหรับการพยากรณ์ค่าต่อเนื่องเหมือน Linear Regression ทั่วไป หากแต่ใช้ฟังก์ชันลอจิสติก (Logistic Function) ในการแปลงผลรวมเชิงเส้นของตัวแปรอิสระให้เป็นความน่าจะเป็นที่อยู่ในช่วง 0 ถึง 1 ซึ่งสามารถนำไปใช้จำแนกประเภทได้ ในทางคณิตศาสตร์ความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่มเป้าหมาย (Positive Class) เช่น y=1 สามารถคำนวณได้จาก

$$P(y=1 \mid x_1, x_2, ..., x_p) = \sigma\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right) \qquad (2.9)$$

ขณะที่ความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่มตรงข้าม (negative class) เช่น y=0 จะสามารถคำนวณได้จาก

$$P(y=1 \mid x_1, x_2, ..., x_p) = 1 - \sigma\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right) \qquad (2.10)$$

โดยที่ $\sigma(z)$ คือ ฟังก์ชันซิกมอยด์ (Sigmoid Function) ซึ่งนิยามคือ $\sigma(z) = \frac{1}{1+e^{-z}}$

ทำหน้าที่เปลี่ยนค่าผลรวมเชิงเส้น z ให้อยู่ในรูปแบบของความน่าจะเป็นระหว่าง 0 ถึง 1 ทำให้สามารถใช้เกณฑ์ Threshold (เช่น 0.5) ในการตัดสินใจจำแนกประเภทได้อย่างชัดเจน (กิตติมศักดิ์, 2564)

ในการวิเคราะห์ข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยเฉพาะการจำแนกประเภทแบบทวิ (Binary Classification) ค่า Threshold มีบทบาทสำคัญอย่างยิ่ง เนื่องจากเป็นค่าที่ใช้กำหนดขอบเขตระหว่างกลุ่มที่จำแนก เช่น กลุ่ม “ยากจน” และ “ไม่ยากจน” หรือกลุ่ม “มีความเสี่ยง”

และ “ไม่มีความเสี่ยง” โดยทั่วไป เมื่อแบบจำลองสร้างค่าความน่าจะเป็น (Probability) ให้กับข้อมูลแต่ละรายการ ค่า Threshold ที่นิยมใช้คือ 0.5 ซึ่งหมายถึง หากค่าความน่าจะเป็นมากกว่าหรือเท่ากับ 0.5 จะถูกจัดอยู่ในกลุ่มเป้าหมาย (Positive Class) แต่หากต่ำกว่าจะถูกจัดอยู่ในกลุ่มอื่น (Negative Class) อย่างไรก็ตาม การใช้ Threshold ค่าคงที่อาจไม่เหมาะสมในบางกรณี โดยเฉพาะเมื่อข้อมูลมีความไม่สมดุลระหว่างกลุ่ม หรือเมื่อต้นทุนของความผิดพลาดมีความแตกต่างกัน เช่น ความผิดพลาดในการทำนายว่า "ไม่ยากจน" ในขณะที่ความจริงเป็น "ยากจน" อาจมีผลกระทบที่รุนแรงกว่าการกำหนด Threshold ที่เหมาะสมจึงควรพิจารณาให้สอดคล้องกับบริบทของปัญหา ทั้งนี้ หนึ่งในวิธีการกำหนดค่า Threshold ที่เหมาะสม คือ การเลือกค่า Threshold ที่ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงที่สุด ซึ่งค่าประสิทธิภาพโดยรวมเป็นค่าเฉลี่ยแบบฮาร์มอนิกระหว่าง ความแม่นยำ (Precision) และค่าความระลึก (Recall) โดยเฉพาะในกรณีที่ข้อมูลมีความไม่สมดุล ค่าประสิทธิภาพโดยรวม (F1-Score) จะสะท้อนประสิทธิภาพของแบบจำลองได้ดีกว่าความถูกต้อง (Accuracy) การเลือก Threshold โดยพิจารณาจากค่าประสิทธิภาพโดยรวม (F1-Score) ที่สูงที่สุดจึงช่วยเพิ่มความแม่นยำในการจำแนกกลุ่มเป้าหมาย พร้อมทั้งลดผลกระทบจากการจัดกลุ่มผิดพลาด (Iguazio, 2024)

ในบริบทของการเรียนรู้ของเครื่อง แบบจำลองนี้มักถูกฝึกด้วยวิธีการหาค่าพารามิเตอร์ที่เหมาะสม เช่น การไล่ระดับ (Gradient Descent) เพื่อหาค่า β ที่ทำให้ค่าความผิดพลาดของแบบจำลองต่ำที่สุด โดยใช้ฟังก์ชันการสูญเสีย (Loss Function) เช่น Cross-Entropy Loss ร่วมกับกระบวนการประเมินค่าด้วย K-Fold Cross-Validation เพื่อให้มั่นใจว่าแบบจำลองมีความสามารถในการ Generalize ข้อมูลใหม่ได้ดี ไม่เกิดปัญหา Overfitting (Miyazaki et al., 2024)

นอกจากนี้ยังสามารถใช้เทคนิค Regularization เช่น L1 (Lasso) และ L2 (Ridge) เพื่อควบคุมความซับซ้อนของแบบจำลอง และลดปัญหา Multicollinearity ในชุดข้อมูล การประเมินผลแบบจำลอง Logistic Regression ที่ใช้ในงาน Machine Learning มักพิจารณาจากตัวชี้วัดต่าง ๆ เช่น ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความระลึก (Recall), ค่าประสิทธิภาพโดยรวม (F1-Score) และ ROC-AUC ซึ่งเน้นประสิทธิภาพในการจำแนกมากกว่าการตีความตัวแปรเชิงสถิติแบบดั้งเดิม (Iwagami et al., 2024)

จากการศึกษาสามารถสรุปได้ว่า แบบจำลองการวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติเน้นการสร้างแบบจำลองที่สามารถตีความได้ โดยให้สมการชัดเจนในรูปของ Logit ซึ่งสะท้อนความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับ Log Odds ของความน่าจะเป็นในการเกิดเหตุการณ์ พร้อมทั้งใช้ค่าพารามิเตอร์ (β) และค่าสถิติต่าง ๆ เช่น p-value เพื่ออธิบายความสำคัญของแต่ละตัวแปร ส่วนแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีการเรียนรู้ของเครื่องมุ่งเน้นผลลัพธ์ในการทำนายมากกว่าการตีความสมการ โดยอาจใช้เทคนิคเพิ่มเติม เช่น การเลือกตัวแปรอัตโนมัติ การปรับพารามิเตอร์ และการประเมินผลผ่านความแม่นยำ แทนการพิจารณาค่าความมีนัยสำคัญเชิงสถิติของตัวแปรแต่ละตัว

2.4.3 แบบจำลอง Random Forest

Random Forest คือ อัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning) ที่เกิดจากการรวม Decision Tree หลาย ๆ ต้นเข้าด้วยกัน โดยแต่ละต้นจะถูกสร้างขึ้นจากคุณลักษณะของข้อมูล (Feature) ที่สุ่มมาเพียงบางส่วน เพื่อสร้างแบบจำลองที่มีประสิทธิภาพสูง โดยใช้หลักการของ Bagging (Bootstrap

Aggregating) เพื่อเพิ่มความแม่นยำและลดปัญหา Overfitting (ณัฐโชติ และสัจจาภรณ์, 2567) สรุปกระบวนการสร้าง Random Forest ได้ดังนี้

1. สุ่มตัวอย่างข้อมูลด้วยวิธี Bootstrap (การสุ่มแบบมีการทดแทน)

2. สร้าง Decision Tree จากข้อมูลที่สุ่มได้โดยในแต่ละ Node จะสุ่มเลือกคุณลักษณะ (Feature) มาใช้ในการแบ่งข้อมูล

3. ทำซ้ำขั้นตอนที่ 1 และ 2 จนได้จำนวน Tree ตามที่กำหนด

4. ในการทำนายใช้การทำนายเสียงส่วนใหญ่ (สำหรับ Classification) หรือค่าเฉลี่ย (สำหรับ Regression) จากผลลัพธ์ของ Tree ทุกต้น

Random Forest (RF) เป็นอัลกอริทึมการเรียนรู้ของเครื่องแบบ Ensemble ที่สามารถใช้ได้กับการวิเคราะห์การถดถอย (Regression) และการจำแนกประเภท (Classification) จุดแข็งของ Random Forest คือความสามารถในการลดปัญหา Overfitting ด้วยการเฉลี่ยผลลัพธ์จากต้นไม้หลายต้น ทำให้แบบจำลองมีความทนทานและมีความแม่นยำสูงกว่าต้นไม้ตัดสินใจเพียงต้นเดียว อีกทั้งอัลกอริทึมนี้ยังสามารถใช้งานได้กับข้อมูลหลายประเภท และมีประสิทธิภาพที่ดีเป็นพิเศษเมื่อใช้กับชุดข้อมูลขนาดใหญ่ที่มีจำนวนตัวแปรมาก (High-Dimensional Data) (Hsu et al., 2024)

สำหรับการเลือกจุดแบ่งภายในต้นไม้ อัลกอริทึมจะใช้เกณฑ์ ค่า Gini Index เพื่อวัดความบริสุทธิ์ของกลุ่มข้อมูล โดยสมการ Gini Index สามารถเขียนได้ดังนี้

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2.11}$$

โดยที่ p_i คือ สัดส่วนของข้อมูลในคลาสที่ i

c คือ จำนวนคลาสทั้งหมดในชุดข้อมูลนั้น

ค่า Gini จะมีค่าน้อยที่สุดเมื่อข้อมูลอยู่ในกลุ่มเดียวกันทั้งหมด (บริสุทธิ์) และจะมีค่าสูงขึ้นเมื่อข้อมูลมีการกระจายไปยังหลายกลุ่ม ซึ่งเป็นตัวช่วยในการแยกข้อมูลในแต่ละโหนดของต้นไม้

วิธีการคำนวณความสำคัญของตัวแปร

การคำนวณความสำคัญของตัวแปร (Feature Importance) ในวิธี Random Forest ใช้หลักการ Mean Decrease Impurity (MDI) หรือ Gini Importance โดยวัดจากความสามารถของตัวแปรในการลดความไม่บริสุทธิ์ (Impurity) ของข้อมูลภายในต้นไม้ เช่น ค่า Gini Impurity หรือ Entropy ในแต่ละโหนดที่ใช้ตัวแปรนั้นในการแบ่งข้อมูล ระบบจะคำนวณค่าการลดลงของ Impurity และถ่วงน้ำหนักตามจำนวนตัวอย่างที่เข้าสู่โหนดนั้น จากนั้นรวมค่าการลด Impurity ทั้งหมดจากทุกโหนดและทุกต้นไม้ในป่า เพื่อให้ได้ค่าความสำคัญรวมของแต่ละตัวแปร ตัวแปรที่ช่วยลด Impurity ได้มากและถูกใช้งานบ่อยจะมีค่าความสำคัญสูงกว่า กำหนดให้ตัวแปรตาม คือ Y และคำนวณค่าเฉลี่ยของตัวแปรที่เกี่ยวข้องกับตัวแปร X_i ด้วย N trees ได้สมการ ดังนี้

$$Imp(X_i) = \frac{1}{N} \sum_{T=1}^{N} \sum_{j \in T : \nu(s_j) = X_m} p(j) \Delta i(s_j, j) \tag{2.12}$$

เมื่อ Imp(X_i) คือ ค่า Feature Importance ของตัวแปร X_i

$p(j)\Delta i(s_j, j)$ คือ น้ำหนัก Purity ที่ลดลงของตัวแปร X_i ในโหนดทั้งหมดของ j

p(j) คือ ความน่าจะเป็นของตัวอย่างในแต่ละโหนด สามารถคำนวณได้จากสมการ 2.13

$$p(j) = \frac{N_j}{N} \tag{2.13}$$

โดยที่ N_j คือ จำนวนตัวอย่างที่เข้าโหนดนั้น

$i(s_j,j)$ เป็น Impurity Measure ที่โหนด j ด้วยตัวแบ่งที่โหนด j

ดังนั้น จะได้ว่า $\nu(s_j) = X_m$ ที่โหนด j, Split Identifier คือ ตัวแปร X_m (ฐิติพร อ้ายดี, 2564)

2.4.4 แบบจำลอง XGBoost

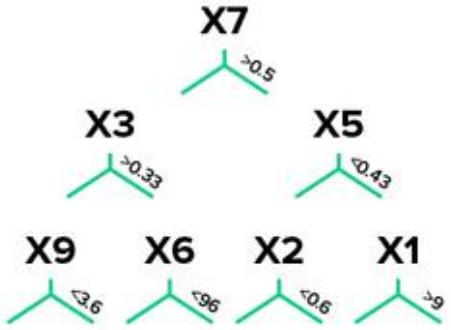
Boosting เป็นเทคนิคหนึ่งในกลุ่ม Ensemble Learning ที่ใช้ในการพัฒนาแบบจำลองหลายแบบจำลองเล็ก ๆ (Weak Learners) ให้กลายเป็นแบบจำลองที่มีประสิทธิภาพสูง (Strong Learner) และมีความสามารถในการจัดการข้อมูลสูญหายได้ดี อีกทั้งยังมีขั้นตอนการเตรียมข้อมูลเพื่อวิเคราะห์น้อยกว่าวิธีการวิเคราะห์ทางสถิติแบบดั้งเดิม (Machado and Holmer, 2022)

XGBoost (Extreme Gradient Boosting) ทำงานอยู่บนพื้นฐานของ Gradient Boosting ซึ่งเป็นเทคนิคการเรียนรู้แบบเสริมแรงที่เพิ่มแบบจำลองย่อยทีละชุดเพื่อลดข้อผิดพลาดสะสมจากแบบจำลองก่อนหน้า โดยแบบจำลองย่อยที่ XGBoost ใช้คือ ต้นไม้ตัดสินใจ (Decision Tree) ซึ่งจะถูกสร้างขึ้นทีละต้นเพื่อลดค่าความสูญเสีย (Loss Function) และจะหยุดสร้างเมื่อข้อผิดพลาดจากต้นไม้ก่อนหน้าไม่เหลือให้เรียนรู้แล้ว (Ali et al., 2023) ซึ่ง XGBoost เป็นเป็นอัลกอริทึมที่พัฒนาให้มีประสิทธิภาพสูง รองรับข้อมูลจำนวนมาก อีกทั้งยังสามารถจัดการข้อมูลสูญหาย (Missing Values) ได้เป็นอย่างดี โดยไม่จำเป็นต้องเติมค่าก่อนล่วงหน้า อัลกอริทึมจะกำหนดทิศทางเริ่มต้นให้กับค่าที่หายไป และค้นหาจุดแบ่งที่ช่วยลดความผิดพลาดในการฝึกแบบจำลอง โดยจะประมวลผลเฉพาะค่าที่หายไป ทำให้ทำงานได้เร็วกว่าวิธีการแบบดั้งเดิม และเพื่อลดปัญหาการเรียนรู้ที่ซับซ้อนเกินไปหรือ Overfitting ซึ่งมักเกิดในอัลกอริทึมแบบ ensemble อัลกอริทึม XGBoost มีฟังก์ชันค่าปรับ (regularization) เพื่อควบคุมโครงสร้างของแบบจำลอง ประกอบด้วย

L1 Regularization (LASSO) ช่วยลดจำนวนตัวแปรที่ไม่จำเป็น โดยส่งผลให้ค่าถ่วงน้ำหนักบางตัวแปรเป็นศูนย์

L2 Regularization (Ridge) ช่วยลดความซับซ้อนของแบบจำลองโดยถ่วงน้ำหนักค่าพารามิเตอร์ไม่ให้สูงเกินไป

ทำให้ XGBoost สามารถสร้างแบบจำลองที่ทั้งแม่นยำและไม่ซับซ้อนเกินความจำเป็น ตัวอย่างโครงสร้างของต้นไม้ตัดสินใจใน XGBoost เป็นดังนี้



รูปที่ 2.3 โครงสร้างของต้นไม้ตัดสินใจใน XGBoost
ที่มา: Aviv (2019)

จากรูปที่ 2.3 โครงสร้างของต้นไม้ตัดสินใจใน XGBoost เป็นแบบ Standard (Asymmetric) ซึ่งเป็นต้นไม้แบบทั่วไปที่แต่ละโหนดสามารถเลือกตัวแปร และ Threshold ต่างกันได้อย่างอิสระ ต้นไม้สามารถเติบโตลึกเฉพาะบางกิ่ง ซึ่งช่วยลดขนาดของแบบจำลองและเพิ่มความยืดหยุ่นในการจัดการข้อมูลที่มีความซับซ้อนหรือไม่สมดุล

วิธีการคำนวณความสำคัญของตัวแปร

XGBoost มีวิธีการคำนวณความสำคัญของตัวแปร 4 วิธี ดังนี้

1. Gain คือ การวัดระดับการมีส่วนร่วมของตัวแปรในการลดค่าความคลาดเคลื่อนของแบบจำลอง โดยค่าที่สูงแสดงว่าตัวแปรนั้นมีบทบาทสำคัญในการปรับปรุงความแม่นยำของแบบจำลอง ซึ่งเป็นวิธีเริ่มต้น **(**Default) ของ XGBoost

2. Weight คือ จำนวนครั้งที่ตัวแปรถูกเลือกใช้ในการแยกข้อมูลภายในต้นไม้ โดยถ้าตัวแปรถูกเลือกบ่อย ยิ่งแสดงถึงความสำคัญในเชิงโครงสร้างของแบบจำลอง

3. Coverage คือ การวัดสัดส่วนของข้อมูลที่ไหลผ่านตัวแปรนั้น ณ จุดที่ถูกใช้ในการแบ่งข้อมูล ซึ่งสะท้อนถึงขอบเขตของผลกระทบของตัวแปรที่มีต่อข้อมูลทั้งหมด

4. Frequency คือ แสดงความถี่ของการใช้งานตัวแปรนั้น ไม่เพียงแต่ในขั้นตอนการแบ่งข้อมูล แต่ยังรวมถึงการเข้าถึงโหนดปลายทาง (Leaf Node) ซึ่งเป็นจุดสุดท้ายของการตัดสินใจในต้นไม้ตัดสินใจ (เครือวัลย์, 2565)

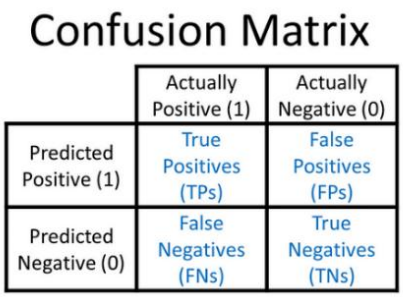
2.5 แนวคิดและทฤษฎีเกี่ยวกับการวัดประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของแบบจำลองเป็นกระบวนการประเมินความสามารถของแบบจำลองในการทำนายหรือจำแนกข้อมูลอย่างถูกต้อง โดยใช้เมตริกต่างๆ เช่น ความแม่นยำ (Accuracy) ความไว (Recall) และค่าประสิทธิภาพโดยรวม (F1-Score) สำหรับปัญหาการจำแนกประเภท นอกจากนี้ยังมีการใช้เทคนิค Cross-Validation เพื่อทดสอบความเสถียรของแบบจำลอง การเลือกใช้เมตริกซ์การวัดขึ้นอยู่กับลักษณะของปัญหาและวัตถุประสงค์ของการศึกษา

2.5.1 เมทริกซ์วัดประสิทธิภาพ (Confusion Matrix)

ในการวัดประสิทธิภาพของแบบจำลองสำหรับการแบ่งกลุ่มของข้อมูลจำนวน 2 กลุ่ม โดยเป็นข้อมูลที่เป็น Positive Class กลุ่มข้อมูลที่สนใจ และ Negative Class กลุ่มข้อมูลที่ไม่ได้สนใจ มีการวัดค่าประสิทธิภาพของแบบจำลองได้ดังนี้

Confusion Matrix คือ ตารางประเมินผลลัพธ์ของการทำนายของแบบจำลอง



รูปที่ 2.4 Confusion Matrix
ที่มา: Rachel (2019)

True Positive (TP) ข้อมูลที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง โดยแบบจำลองทำนายว่าจริงและสิ่งที่เกิดขึ้นก็คือจริง

True Negative (TN) ข้อมูลที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง โดยแบบจำลองทำนายว่าไม่จริงและสิ่งที่เกิดขึ้นก็คือไม่จริง

False Positive (FP) ข้อมูลที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง โดยแบบจำลองทำนายว่าจริงแต่สิ่งที่เกิดขึ้นคือไม่จริง

False Negative (FN) ข้อมูลที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง โดยแบบจำลองทำนายว่าไม่จริงแต่สิ่งที่เกิดขึ้นคือจริง

ค่าความถูกต้อง (Accuracy) คือ ค่าที่ถูกต้องในการทำนายของแบบจำลอง โดยการหาอัตราส่วนระหว่างกลุ่มที่ทำนายถูกต้องต่อผลการทำนายทั้งหมด

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2.14)$$

True Positive Rate หรือ **Recall** คือ การเทียบอัตราส่วนการทำนายที่ถูกต้องต่อจำนวนของที่เป็นจริงทั้งหมด

$$Recall = TP/(TP + FN) \quad (2.15)$$

ค่าความแม่นยำ (Precision) คือ เป็นการเปรียบเทียบการทำนายที่ถูกต้องว่าจริง และก็เกิดขึ้นจริง การทำนายว่าจริงแต่สิ่งที่เกิดขึ้น คือ ไม่จริง

$$Precision = TP/(TP/FP) \quad (2.16)$$

False Negative Rate (FNR) คือ อัตราการทำนายผิดเทียบกับข้อมูล Negative ทั้งหมด

$$FNR = FP/(TN + FP) \quad (2.17)$$

ค่าประสิทธิภาพโดยรวม (F1-Score) เป็นค่าที่สะท้อนค่า Recall กับ Precision พร้อมกัน

$$F\_1 = 2*(Precision*Recall)/(Precision+Recall) \quad (2.18)$$

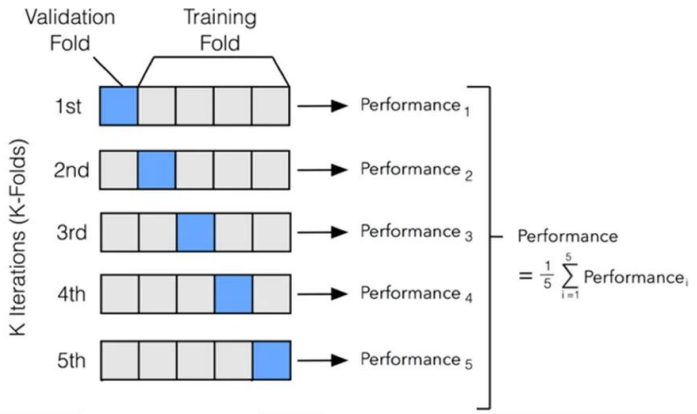
ค่าความถูกต้องสมดุล (Balanced Accuracy) สามารถใช้วัดประสิทธิภาพของแบบจำลองทั้งแบบการจำแนกประเภท 2 กลุ่มและมากกว่า 2 กลุ่ม มักจะถูกนำมาใช้ เมื่อข้อมูลมีความไม่สมดุลกันระหว่างคลาส

$$Balanced\ Accuracy = (Sensitivity+Specificity)/2 \quad (2.19)$$

โดยที่ Sensitivity = TP/(TP+FN) และ Specificity = TN/(TN+FP)

2.5.2 การวิเคราะห์ความแม่นตรง (Cross Validation)

การวิเคราะห์ความแม่นตรง (Cross Validation) คือ เป็นเครื่องมือที่ช่วยวิเคราะห์ความเที่ยงตรงของแบบจำลอง โดยลักษณะการตรวจสอบแบบไขว้กัน ซึ่งใช้การสุ่มตัวอย่างในการแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดข้อมูลตรวจสอบ วิธีการวิเคราะห์ความแม่นตรงที่ใช้บ่อยที่สุด คือ วิธี K-fold Cross Validation โดยแบ่งชุดข้อมูลออกเป็น K ชุดเท่า ๆ กัน และใช้ชุดข้อมูล K-1 ในการฝึกแบบจำลอง และอีก 1 ชุดข้อมูลที่เหลือใช้เพื่อทดสอบประสิทธิภาพของแบบจำลอง และทำซ้ำจำนวน K รอบ โดยเปลี่ยนชุดข้อมูลตรวจสอบทีละชุดข้อมูลไปเรื่อย ๆ จนครบ ดังรูปที่ 2.5



รูปที่ 2.5 K-fold Validation
ที่มา: Shen (2020)

จากรูปที่ 2.5 ใช้ 5 Fold Cross Validation โดยการแบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน ในการวนรอบครั้งแรก จะใช้ 4 ส่วนหลังเป็นชุดข้อมูลสำหรับฝึกสอน และอีก 1 ส่วนที่เหลือใช้เป็นชุดตรวจสอบ

และทำการประเมินประสิทธิภาพของแบบจำลอง เช่น ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) เป็นต้น ในรอบถัดไป จะเปลี่ยนส่วนที่ใช้เป็นชุดตรวจสอบโดยใช้ส่วนที่สองเป็นชุดตรวจสอบ และใช้ส่วนที่เหลืออีก 4 ส่วนเป็นชุดฝึกสอน แล้วทำการประเมินประสิทธิภาพของแบบจำลองอีกครั้ง กระบวนการนี้จะทำซ้ำไปเรื่อย ๆ จนครบทั้ง 5 ส่วน ซึ่งแต่ละส่วนจะถูกใช้เป็นชุดตรวจสอบหนึ่งครั้ง เมื่อสิ้นสุดการวนรอบทั้งหมด จะทำการหาค่าเฉลี่ยของค่าประสิทธิภาพที่ได้จากทุกรอบ เพื่อใช้เป็นค่าประมาณความสามารถของแบบจำลองในการทำงานกับข้อมูลใหม่ที่ไม่เคยใช้วิเคราะห์มาก่อน (Pramod, 2023)

2.5.3 การทดสอบสมมติฐานด้วย McNemar's Test

การทดสอบแมคเนมาร์ (McNemar's Test) เป็นการทดสอบทางสถิติที่ใช้สำหรับข้อมูลเชิงนามบัญญัติที่มีลักษณะจับคู่ (Paired Nominal Data) และมีสองกลุ่มย่อย (Dichotomous Variables) เช่น ใช่/ไม่ใช่ หรือ ผ่าน/ไม่ผ่าน การทดสอบนี้มุ่งเน้นการตรวจสอบว่าสัดส่วนของผลลัพธ์ในสองเงื่อนไขหรือสองช่วงเวลามีความแตกต่างกันหรือไม่ (Wikipedia, 2012) สมมติฐานของการทดสอบ คือ

H_0: ไม่มีความแตกต่างระหว่างสัดส่วนของผลลัพธ์ในสองเงื่อนไขหรือสองช่วงเวลา กล่าวคือ $b = c$

H_1: มีความแตกต่างระหว่างสัดส่วนของผลลัพธ์ในสองเงื่อนไขหรือสองช่วงเวลา กล่าวคือ $b \neq c$

โดยที่ b และ c คือ จำนวนของกรณีที่มีการเปลี่ยนแปลงจากสถานะหนึ่งไปยังอีกสถานะหนึ่งในข้อมูลที่จับคู่กัน

ข้อมูลที่ใช้ในการทดสอบ McNemar's Test มักนำเสนอในรูปแบบของตารางความถี่ขนาด 2 × 2 ดังนี้

ตารางที่ 2.2 ข้อมูลที่ใช้ในการทดสอบ McNemar's Test

A vs B		B	
		ใช่	**ไม่ใช่**
A	**ใช่**	a	b
	ไม่ใช่	c	d

ค่าที่สำคัญสำหรับการทดสอบ คือค่าของ b และ c ซึ่งเป็นจำนวนของกรณีที่มีการเปลี่ยนแปลงจากสถานะหนึ่งไปยังอีกสถานะหนึ่ง โดยที่ค่าสถิติของ McNemar's Test มีการแจกแจงแบบไคสแควร์ (Chi-Squared Distribution) กับ 1 องศาอิสระ (df = 1) อย่างไรก็ตาม หากจำนวนของ b + c มีค่าน้อย (เช่น น้อยกว่า 25) การใช้การแจกแจงแบบไคสแควร์อาจไม่เหมาะสม และควรพิจารณาใช้การทดสอบแบบ Exact หรือ Mid-p ซึ่งให้ค่าพีที่แม่นยำมากกว่าในกรณีที่ขนาดตัวอย่างมีน้อย

สูตรสำหรับคำนวณค่าสถิติไคสแควร์ของ McNemar's Test มีดังนี้

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{2.20}$$

โดยที่ b คือจำนวนกรณีที่เปลี่ยนจาก "ไม่ใช่" ⟶ "ใช่"
c คือจำนวนกรณีที่เปลี่ยนจาก "ใช่" ⟶ "ไม่ใช่"

หากทดสอบสมมติฐานทางสถิติ ($p < 0.05$) สามารถสรุปว่ามีความแตกต่างระหว่างสองสถานะหรือช่วงเวลานั้นอย่างมีนัยสำคัญ

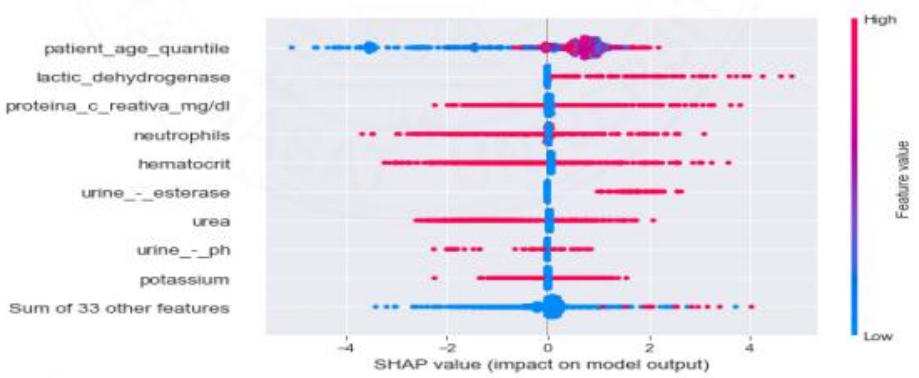
2.6 การตีความผลการทำนายของแบบจำลองด้วย SHAP

SHAP หรือ Shapley Additive Explanations เป็นเทคนิคสำหรับอธิบายการทำงานของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) โดยมีเป้าหมายเพื่อเพิ่มความโปร่งใส (transparency) และความเข้าใจในผลลัพธ์ที่แบบจำลองคาดการณ์ไว้ เทคนิคนี้พัฒนาจากแนวคิดของค่า Shapley ในทฤษฎีเกมแบบร่วมมือ (Cooperative Game Theory) ซึ่งใช้วิธีแบ่ง “ผลตอบแทน” อย่างยุติธรรมให้กับแต่ละตัวแปรที่มีส่วนร่วมในการตัดสินใจของแบบจำลอง

SHAP ช่วยให้ผู้ใช้สามารถทราบได้ว่า ตัวแปรแต่ละตัวมีอิทธิพลต่อผลลัพธ์มากน้อยเพียงใด และในทิศทางใด โดยสามารถอธิบายได้ทั้งในระดับภาพรวมของแบบจำลอง (Global Explanation) และระดับเฉพาะเจาะจงสำหรับแต่ละข้อมูล (Local Explanation) เช่น การแสดงว่าเพราะอะไรแบบจำลองจึงให้ผลลัพธ์แบบนี้กับข้อมูลบางรายการ และตัวแปรใดมีผลมากที่สุดในการตัดสินใจของแบบจำลองนั้น

จากงานวิจัยพบว่า การใช้ SHAP ช่วยให้สามารถเลือกตัวแปรที่สำคัญได้อย่างแม่นยำ ตัวอย่าง เช่น ในการวิเคราะห์ข้อมูลผู้ป่วยโควิด-19 งานวิจัยสามารถลดจำนวนตัวแปรจาก 42 เหลือเพียง 10 ตัวแปรที่จำเป็น โดยยังคงระดับความแม่นยำของแบบจำลองไว้ที่ประมาณ 90% และช่วยลดต้นทุนในการคำนวณได้ถึงร้อยละ 58.7 สิ่งนี้แสดงให้เห็นว่า SHAP ไม่เพียงแต่อธิบายผลลัพธ์ของแบบจำลอง แต่ยังช่วยในการคัดเลือกคุณลักษณะ (Feature Selection) ได้อย่างมีประสิทธิภาพ

SHAP จึงถือเป็นเครื่องมือสำคัญในกระบวนการพัฒนาแบบจำลองที่ต้องการความน่าเชื่อถือและสามารถอธิบายผลลัพธ์ได้ โดยเฉพาะในกรณีที่แบบจำลองมีความซับซ้อนหรือถูกใช้งานในบริบทที่ต้องการความโปร่งใส เช่น ด้านสุขภาพ การเงิน หรือการวิเคราะห์เชิงนโยบาย (ทินรัตน์, 2565)



รูปที่ 2.6 ตัวอย่างการใช้ SHAP อธิบายระดับผลกระทบของค่าตัวแปรต่อผลลัพธ์ของแบบจำลอง
ที่มา: ทินรัตน์ (2565)

SHAP Summary Plot มีลักษณะเป็นกราฟแบบจุดกระจาย (Dot Plot) ที่มีการจัดวางดังนี้

- แกน Y (แนวตั้ง) แสดงรายชื่อตัวแปรต่างๆ เรียงลำดับตามความสำคัญจากมากไปน้อย
- แกน X (แนวนอน) แสดงค่า SHAP Value ซึ่งบ่งบอกถึงผลกระทบของตัวแปรต่อการทำนาย โดยที่หาก SHAP Value เป็นบวก (+) ตัวแปรนั้นมีผลทำให้การทำนายเพิ่มขึ้น ในทางกลับกันหากค่า SHAP Value เป็นลบ (-) ตัวแปรนั้นมีผลทำให้การทำนายลดลง
- สีของจุด แสดงค่าของตัวแปรนั้นๆ (สีแดงแสดงค่าสูง สีน้ำเงินแสดงค่าต่ำ)

จากรูปที่ 2.6 แสดงภาพ SHAP Summary Plot ซึ่งเป็นเครื่องมือในการอธิบายผลลัพธ์ของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) โดยใช้ค่า SHAP (SHapley Additive exPlanations) เพื่อแสดงอิทธิพลของแต่ละตัวแปรที่มีต่อค่าทำนายของแบบจำลองในแต่ละข้อมูล (Instance) แกนนอนแสดงค่า SHAP ซึ่งบ่งบอกถึงขนาดและทิศทางของผลกระทบ กล่าวคือ ค่าบวกหมายถึงตัวแปรนั้นมีแนวโน้มผลักดันค่าทำนายให้สูงขึ้น ขณะที่ค่าลบหมายถึงมีแนวโน้มลดค่าทำนาย โดยแต่ละแถวในกราฟแสดงตัวแปรหนึ่งตัว โดยเรียงตามลำดับความสำคัญจากมากไปน้อย จุดแต่ละจุดแทนค่าจากแต่ละข้อมูลจริงที่ใช้ในการสร้างแบบจำลอง สีของจุดสะท้อนค่าของตัวแปรในแต่ละข้อมูล โดยสีแดงแสดงค่าที่สูง และสีน้ำเงินแสดงค่าที่ต่ำ การกระจายของจุดในแนวนอนแสดงให้เห็นว่าตัวแปรดังกล่าวมีผลกระทบมากน้อยเพียงใดในข้อมูลแต่ละชุด กราฟนี้จึงช่วยให้เห็นทั้งลำดับความสำคัญของตัวแปร และเข้าใจลักษณะของอิทธิพลที่แต่ละตัวแปรมีต่อการทำนายได้ในระดับเชิงลึก ทั้งในเชิงบวกและลบ ซึ่งเป็นประโยชน์อย่างยิ่งในการตีความการตัดสินใจของแบบจำลอง โดยเฉพาะในบริบทที่ต้องการความโปร่งใสและความเข้าใจที่มากกว่าเพียงค่าความแม่นยำของแบบจำลอง

2.7 งานวิจัยที่เกี่ยวข้อง

เครือวัลย์ (2565) ได้ทำการศึกษาเรื่อง “การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้” โดยมีวัตถุประสงค์เพื่อทำนายโอกาสที่ลูกหนี้จะผิดนัดชำระกับทางธนาคาร โดยแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มลูกหนี้ปกติ และกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร ใช้การวิเคราะห์ Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting โดยผู้วิจัยแบ่งข้อมูลออกเป็น Train Data 80% และ Test Data 20% ข้อมูลที่ใช้มีปัญหาข้อมูลไม่สมดุล ผู้วิจัยจึงเลือกจัดการด้วย Oversampling, Under Sampling และ Synthetic Minority Oversampling Technique (SMOTE) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองกับการจัดการข้อมูลด้วยวิธีที่แตกต่างกัน ในการสร้างแบบจำลองผู้วิจัยใช้ Cross Validation ในชุดข้อมูล Train เพื่อตรวจสอบประสิทธิภาพของแบบจำลอง โดยผลการวิจัยพบว่าการแก้ปัญหาไม่สมดุลของข้อมูลด้วย Under Sampling ทำให้แบบจำลองมีประสิทธิภาพที่ดีที่สุด และการใช้อัลกอริทึม Gradient Boosting มีค่าประสิทธิภาพโดยรวม (F1-Score), ค่าความระลึก (Recall), ค่าความถูกต้อง (Accuracy) สูงที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ และเมื่อพิจารณา Feature Importance ของการพัฒนาแบบจำลอง ระหว่าง XGBoostClassifier, Random Forest, Gradient Boosting แบบจำลองทั้งสามให้ความสำคัญกับตัวแปรที่เหมือนกัน ดังนี้ NAME_HOUSING_TYPE, AMT_ANNUITY, CODE_GENDER, NAME_TYPE_SUITE

บทสรุปจาการทบทวนงานวิจัย เรื่อง การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิตโดยการใช้อัลกอริทึมการเรียนรู้ เพื่อวิเคราะห์ประเด็นสำคัญเข้าสู่งานวิจัยครั้งนี้

งานวิจัยนี้เปรียบเทียบประสิทธิภาพประสิทธิภาพของอัลกอริทึม โดยตัวแปรตามมีลักษณะเป็นตัวแปรเชิงคุณภาพ 2 กลุ่ม ผู้วิจัยแบ่งชุดข้อมูลเพื่อฝึกแบบจำลองและทดสอบประสิทธิภาพของแบบจำลองด้วยอัตราส่วน และแบ่งข้อมูลฝึกด้วย Cross-Validation เพื่อตรวจสอบประสิทธิภาพของแบบจำลอง การจัดการข้อมูลไม่สมดุลสามารถทำได้หลายวิธี แต่จากงานวิจัยนี้วิธี Under Sampling ทำให้ประสิทธิภาพของแบบจำลองดีที่สุด และการวัดประสิทธิภาพของแบบจำลองผู้วิจัยพิจารณาค่าประสิทธิภาพโดยรวม (F1-Score), ค่าความระลึก (Recall), ค่าความถูกต้อง (Accuracy) ร่วมกัน โดยเลือกพิจารณาค่าความระลึก (Recall) เพราะธนาคารต้องการตรวจจับ “ลูกหนี้ที่มีความเสี่ยง” ให้ได้มากที่สุด

สุพนิดา (2565) ได้ทำการศึกษาเรื่อง “ความยากจนในผู้สูงอายุไทย: การเปลี่ยนแปลงและปัจจัยเสี่ยง” โดยมีวัตถุประสงค์เพื่อศึกษาการเปลี่ยนแปลงของความยากจนในผู้สูงอายุไทยและครัวเรือนผู้สูงอายุในช่วงปี พ.ศ. 2560 และ พ.ศ. 2564 รวมถึงการวิเคราะห์ปัจจัยที่มีความสัมพันธ์กับสถานะความยากจนโดยใช้ข้อมูลจากการสำรวจประชากรสูงอายุและการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือนของสำนักงานสถิติแห่งชาติ งานวิจัยนี้ใช้การวิเคราะห์ถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) เพื่อวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม คือ สถานะความยากจนของผู้สูงอายุ/ครัวเรือนผู้สูงอายุ (1 = ยากจน, 0 = ไม่ยากจน) กับตัวแปรอิสระ ได้แก่ อายุ, ขนาดครัวเรือน, ระดับการศึกษา, สถานภาพสมรส, เขตที่อยู่อาศัย, สถานภาพการทำงาน, สวัสดิการค่ารักษาพยาบาล, รูปแบบการอยู่อาศัย รวมถึงตัวแปรเพศและการพึ่งพิงรวม ซึ่งมีนัยสำคัญเฉพาะในบางปี ผลการวิเคราะห์ชี้ว่าปัจจัยสำคัญที่สัมพันธ์กับความยากจนอย่างมีนัยสำคัญในทั้งสองปี ได้แก่ อายุหัวหน้าครัวเรือน, จำนวนสมาชิกในครัวเรือน, ระดับการศึกษา, เขตที่อยู่อาศัย ปัจจัยดังกล่าวมีความสัมพันธ์กับความยากจนในทิศทางเดียวกัน ยกเว้นอายุมีความสัมพันธ์กันในทิศทางตรงกันข้าม และสถานภาพสมรสมีความสัมพันธ์กับความยากจนอย่างมีนัยสำคัญ เฉพาะการวิเคราะฆ์ข้อมูลปี 2560 โดยมีค่าความสามารถในการอธิบายแบบจำลอง $R^2 = 0.263$ แสดงถึงความสามารถของแบบจำลองในการอธิบายการแปรผันของสถานะความยากจน และใช้ ค่า Odds Ratio เพื่อแปลผลทิศทางของอิทธิพลของแต่ละตัวแปร

บทสรุปจาการทบทวนงานวิจัย เรื่อง ความยากจนในผู้สูงอายุไทย: การเปลี่ยนแปลงและปัจจัยเสี่ยง เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ปัจจัยที่ผู้วิจัยใช้ในการวิจัยนี้ ประกอบด้วย อายุ ขนาดครัวเรือน ระดับการศึกษา สถานภาพสมรส เขตที่อยู่อาศัย สถานภาพการทำงาน สวัสดิการค่ารักษาพยาบาล รูปแบบการอยู่อาศัย เพศ และการพึ่งพิงรวม ปัจจัยที่มีความสัมพันธ์กับความยากจนอย่างมีนัยสำคัญ ได้แก่ อายุ จำนวนสมาชิกในครัวเรือน ระดับการศึกษา เขตที่อยู่อาศัย สถานภาพสมรส การศึกษาข้อมูลแต่ละปีปัจจัยที่ส่งผลต่อความยากจนอาจแตกต่างกันซึ่งอาจจะเกิดขึ้นจากปัจจัยภายนอก

Adhikari (2020) ได้ทำการศึกษาเรื่อง “Factors Influencing the Income of Urban Informal Workers: Evidence from Nepal” โดยมีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่มีผลต่อรายได้ของแรงงานนอกระบบในเขตเมืองของประเทศเนปาล งานวิจัยนี้ใช้ข้อมูลจากกลุ่มตัวอย่างจำนวน

179 คน ใน 6 เมืองใหญ่ โดยใช้แบบสอบถาม และวิเคราะห์ด้วยวิธีการถดถอยพหุคูณ (Multiple Regression) เพื่อศึกษาความสัมพันธ์ระหว่างรายได้ประจำปีของแรงงานกับตัวแปรต่าง ๆ ซึ่งตัวแปรตามของงานวิจัยนี้คือ รายได้ต่อปีของแรงงานนอกระบบ ส่วนตัวแปรอิสระ ได้แก่ อายุ, จำนวนปีที่ศึกษา, ขนาดครัวเรือน, ประสบการณ์ทำงาน, จำนวนวันทำงานต่อเดือน และจำนวนเดือนทำงานต่อปี ผลการศึกษาแสดงให้เห็นว่า ตัวแปรที่มีอิทธิพลเชิงบวกและมีนัยสำคัญทางสถิติ ต่อรายได้ ได้แก่ อายุ จำนวนปีที่ศึกษา ขนาดครัวเรือน และจำนวนเดือนทำงาน ขณะที่จำนวนวันทำงานต่อเดือนและประสบการณ์ไม่มีนัยสำคัญทางสถิติ โดยแบบจำลองมีค่า R-squared เท่ากับ 0.287 และไม่พบปัญหา Multicollinearity จากค่า VIF ซึ่งอยู่เฉลี่ยเพียง 1.36

บทสรุปจากการทบทวนงานวิจัย เรื่อง Factors Influencing the Income of Urban Informal Workers: Evidence from Nepal เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ปัจจัยที่ผู้วิจัยใช้ในการวิจัยนี้ ประกอบด้วย อายุ จำนวนปีที่ศึกษา ขนาดครัวเรือน ประสบการณ์ทำงาน จำนวนวันทำงานต่อเดือน และจำนวนเดือนทำงานต่อปี ปัจจัยที่มีความสัมพันธ์กับรายได้ต่อปีของแรงงานนอกระบบอย่างมีนัยสำคัญ ได้แก่ อายุ จำนวนปีที่ศึกษา ขนาดครัวเรือน และจำนวนเดือนทำงาน

Roy และ Kundu (2020) ได้ทำการศึกษาเรื่อง “An Analysis of Poverty Among the Informal Workers of India” โดยมีวัตถุประสงค์เพื่อศึกษาระดับความยากจน ความรุนแรงของความยากจน และปัจจัยที่ส่งผลต่อความยากจนในหมู่แรงงานนอกระบบของอินเดีย โดยใช้ข้อมูลจากการสำรวจแรงงานและการว่างงานของ National Sample Survey Office ในปี 2011–2012 งานวิจัยใช้วิธีทางสถิติ Heckman 2-step Regression Model, OLS, และ Quantile Regression เพื่อวิเคราะห์ทั้งปัจจัยที่ส่งผลต่อความยากจน (Poverty Incidence) ความลึกของความยากจน (Poverty Gap) โดยตัวแปรตามมีค่าเป็น 0 เมื่อแรงงานนอกระบบมีรายได้สูงกว่าเส้นความยากจน และเป็น 1 เมื่อแรงงานนอกระบบมีรายได้ต่ำกว่าเส้นความยากจน ผลการศึกษาพบว่า ปัจจัยสำคัญที่เพิ่มโอกาสตกอยู่ในความยากจน ได้แก่ การไม่มีการศึกษาทางเทคนิค การสังกัดกลุ่มวรรณะล่าง การไม่มีบัญชีธนาคาร และการไม่พึงพอใจในงาน นอกจากนี้ แรงงานในภาคชนบทมีแนวโน้มยากจนมากกว่าในเมือง แต่มีระดับความรุนแรงของความยากจนน้อยกว่า โดยเฉพาะแรงงานที่ไม่มีสถานที่ทำงานถาวรหรือทำงานกลางแจ้ง เช่น ตามถนนหรือตลาด มีโอกาสจนสูงกว่ากลุ่มที่ทำงานในสำนักงานหรือที่อยู่อาศัยอย่างเป็นทางการ

บทสรุปจากการทบทวนงานวิจัย เรื่อง An Analysis of Poverty Among the Informal Workers of India เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ผู้วิจัยใช้เส้นความยากจนในการแบ่งคนจน โดยผู้ที่มีรายได้ต่ำกว่าเส้นความยากจน คือ คนจน เช่นเดียวกับการแบ่งกลุ่มความยากจนของประเทศไทย ปัจจัยที่มีผลให้แรงงานนอกระบบในประเทศอินเดียตกอยู่ในความยากจน ประกอบไปด้วย การศึกษา กลุ่มวรรณะ การไม่มีบัญชีธนาคาร การไม่พึงพอใจในงาน เขตที่อยู่อาศัย และสถานที่ทำงานของแรงงาน

Nurpratiwi et al. (2020) ได้ทำการศึกษาเรื่อง “Factors that Influence Wage Differences in Formal Sector on Male and Female Workers in Palembang City” โดยมีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อค่าจ้างแรงงานชายและหญิงในภาคทางการของเมืองปาเลมบัง ประเทศ

อินโดนีเซีย โดยใช้ข้อมูลจากกลุ่มตัวอย่างแรงงาน 164 คนจาก 6 อุตสาหกรรมเสี่ยง และวิเคราะห์ข้อมูลด้วย การวิเคราะห์การการถดถอยเชิงพหุคูณ (Multiple Linear Regression) แยกเพศ และการทดสอบความแตกต่างของพารามิเตอร์ (Different Parameter Test) ระหว่างแรงงานชายและหญิง ตัวแปรตามของการวิจัยคือ ระดับค่าจ้างรายเดือน ขณะที่ตัวแปรอิสระประกอบด้วย ระดับการศึกษา อายุ ชั่วโมงทำงาน ประสบการณ์ทำงาน และความเสี่ยงจากงาน ผลการศึกษาแสดงให้เห็นว่า สำหรับแรงงานชาย ตัวแปรการศึกษา ชั่วโมงทำงาน ประสบการณ์ทำงาน และความเสี่ยงมีผลบวกและมีนัยสำคัญทางสถิติต่อค่าจ้าง ($p<0.05$) โดยมีค่า Adjusted R^2 เท่ากับ 0.641 ขณะที่แรงงานหญิง ตัวแปรการศึกษา อายุ และความเสี่ยงมีผลบวกและมีนัยสำคัญทางสถิติ ($p<0.05$) โดยมีค่า Adjusted R^2 เท่ากับ 0.510 นอกจากนี้ยังพบว่า ค่าสัมประสิทธิ์ของการถดถอยของตัวแปรส่วนใหญ่แตกต่างกันระหว่างเพศ ยกเว้นตัวแปรความเสี่ยง ซึ่งไม่พบความแตกต่างอย่างมีนัยสำคัญ

บทสรุปจาการทบทวนงานวิจัย เรื่อง Factors that Influence Wage Differences in Formal Sector on Male and Female Workers in Palembang City เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ผู้วิจัยทำการวิเคราะห์ข้อมูล โดยการแบ่งข้อมูลเป็น 2 กลุ่มตามเพศ และศึกษาปัจจัยที่มีผลต่อค่าจ้างของแรงงาน ตัวแปรอิสระประกอบด้วย ระดับการศึกษา อายุ ชั่วโมงทำงาน ประสบการณ์ทำงาน และความเสี่ยงจากงาน ซึ่งปัจจัยที่มีผลต่อค่าจ้างของแรงงานชายและแรงงานหญิงแตกต่างกัน แต่ตัวแปรระดับการศึกษา และความเสี่ยงจากการทำงานมีความสัมพันธ์ต่อค่าจ้างของแรงงานทั้งเพศชายและเพศหญิง อย่างมีนัยสำคัญเช่นเดียวกัน

สุพิชชา และฐากูร (2562) ได้ทำการศึกษาเรื่อง “ภาวะความยากจนและคุณภาพชีวิตของครัวเรือนเกษตรกรในเขตจังหวัดเพชรบูรณ์” โดยมีวัตถุประสงค์เพื่อวิเคราะห์ระดับความยากจนและปัจจัยที่มีผลต่อความยากจนของครัวเรือนเกษตรในพื้นที่ดังกล่าว งานวิจัยนี้เก็บข้อมูลจากกลุ่มตัวอย่างจำนวน 400 ครัวเรือน และนำมาวิเคราะห์ด้วยวิธีทางสถิติ ได้แก่ ดัชนีวัดสัดส่วนและความรุนแรงของความยากจน (Head-count Ratio และ FGT Index) และการวิเคราะห์ถดถอย (Regression) ตัวแปรอิสระที่ใช้ในงานวิจัย ประกอบด้วยจำนวนผู้พึ่งพิงในครัวเรือน ขนาดที่ดินในภาคเกษตรกรรมของครัวเรือน ระดับการศึกษาของหัวหน้าครัวเรือน พื้นที่ทำการเกษตรของครัวเรือน กิจกรรมการเกษตรของครัวเรือน (การปลูกข้าว/ข้าวโพดเลี้ยงสัตว์, การปลูกผักผลไม้) หนี้สินของครัวเรือน และสินทรัพย์ของครัวเรือน ผลการศึกษาแสดงให้เห็นว่าครัวเรือนที่มีรายได้จากการเกษตรลดลง มีจำนวนผู้พึ่งพิงมาก หัวหน้าครัวเรือนมีระดับการศึกษาต่ำกว่าภาคบังคับ และประกอบอาชีพหลักในการปลูกข้าวหรือข้าวโพดเลี้ยงสัตว์ มีแนวโน้มที่จะยากจนมากกว่าครัวเรือนอื่น โดยตัวแปรตามของการวิจัยคือสถานะความยากจน (ยากจน/ไม่ยากจน) และแบบจำลองมีค่าความแม่นยำในการพยากรณ์ 45.26% และค่า McFadden R-Squared เท่ากับ 0.3281 ซึ่งสะท้อนว่าแบบจำลองสามารถอธิบายปัจจัยที่มีผลต่อความยากจนได้

บทสรุปจาการทบทวนงานวิจัย เรื่อง ภาวะความยากจนและคุณภาพชีวิตของครัวเรือนเกษตรกรในเขตจังหวัดเพชรบูรณ์ เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ผู้วิจัยทำการเก็บข้อมูลด้วยแบบสอบถาม ตัวแปรส่วนใหญ่ที่ใช้เป็นข้อมูลด้านการเกษตรของครัวเรือน แต่พบว่าลักษณะทางกิจกรรมการเกษตรที่แตกต่างกันก็ส่งผลต่อความยากจนแตกต่างกันระดับการศึกษา และจำนวนผู้พึ่งพิงหรือจำนวนสมาชิกในครัวเรือนที่ต้องดูแลมีผลต่อความยากจนของ

ครัวเรือนเช่นเดียวกัน

Alia El Mahdi (2010) ได้ทำการศึกษาเรื่อง “Poverty and Informality: A Restraining or Constructive Relationship” โดยมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างความยากจนกับการดำเนินกิจการขนาดเล็กที่ไม่เป็นทางการ (Informal MSEs) ในอียิปต์ โดยใช้ข้อมูลสำรวจแบบพาเนลระหว่างปี 1998–2006 พบว่า ครัวเรือนที่สามารถรักษากิจการไว้ได้มีแนวโน้มเพิ่มสถานะความมั่งคั่งมากกว่าครัวเรือนที่กิจการล้มเหลว ปัจจัยสำคัญที่ส่งผลต่อการรอดและการเปลี่ยนแปลงสถานะความมั่งคั่งของครัวเรือน ได้แก่ การดำเนินกิจการภายในสถานประกอบการ ตั้งอยู่ในเขตเมือง ขนาดกิจการที่ใหญ่ขึ้น การมีทุนมากขึ้น และการศึกษาของผู้ประกอบการ โดยใช้การวิเคราะห์ Logistic Regression และ Random Effects Model เพื่อวิเคราะห์การเปลี่ยนแปลงเชิงบวกในลักษณะของกิจการและผู้ประกอบการสามารถส่งผลโดยตรงต่อสถานะเศรษฐกิจของครัวเรือนได้อย่างมีนัยสำคัญ งานวิจัยชี้ให้เห็นว่าแม้ภาคกิจการไม่เป็นทางการจะช่วยดูดซับแรงงานและลดการว่างงาน แต่กลับพบว่าแรงงานส่วนใหญ่ในภาคนี้ยังคงติดอยู่ในกับดักความยากจน โดยเฉพาะผู้หญิงและผู้ที่ทำงานนอกสถานประกอบการอย่างไม่มั่นคง รายได้ต่ำ ขาดสิทธิและการคุ้มครอง การมี MSEs จึงไม่เพียงพอหากกิจการไม่สามารถเติบโตหรือเข้าสู่ระบบเศรษฐกิจทางการได้ ความยากจนจะยังคงอยู่ ในทางตรงกันข้าม หากสามารถยกระดับกิจการด้วยทักษะ ทุน และโครงสร้างพื้นฐานที่เหมาะสม ก็จะช่วยให้ครัวเรือนหลุดพ้นจากความยากจนได้ในระยะยาว

บทสรุปจากการทบทวนงานวิจัย เรื่อง Poverty and Informality: A Restraining or Constructive Relationship เพื่อวิเคราะห์ปัจจัยสำคัญเข้าสู่งานวิจัยครั้งนี้

ปัจจัยที่ส่งผลต่อการเปลี่ยนสถานะทางเศรษฐกิจของกิจการ หรือการเปลี่ยนแปลงสถานะความมั่งคั่งของครัวเรือน ประกอบด้วย สถานที่ทำงาน เขตที่ตั้ง ขนาดของกิจการ ระดับการศึกษา ต้นทุนการดำเนินกิจการ แต่เมื่อพิจารณาปัจจัยส่วนบุคคลของแรงงานพบว่า เพศ ระดับการศึกษา และสถานที่ทำงาน มีผลต่อความยากจนของแรงงาน

บทที่ 3
วิธีการดำเนินงานวิจัย

การวิจัยในครั้งนี้เป็นการศึกษาเรื่อง ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ เป็นการวิจัยเชิงปริมาณ โดยข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลจากการสำรวจแรงงานนอกระบบ พ.ศ. 2567 ของสำนักงานสถิติแห่งชาติ และข้อมูลสถิติดัชนีความก้าวหน้าของคน รายจังหวัด ปี 2566 ของสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ซึ่งผู้วิจัยสามารถสรุปวิธีการดำเนินงานวิจัยได้ดังนี้

3.1 ประชากรและกลุ่มตัวอย่าง
3.2 เครื่องมือที่ใช้ในการวิจัย
3.3 การเก็บรวบรวมข้อมูล
3.4 การเตรียมข้อมูล
3.5 การพัฒนาแบบจำลอง
3.6 การประเมินผล

3.1 ประชากรและกลุ่มตัวอย่าง

การกำหนดประชากรและกลุ่มตัวอย่างได้กำหนดขอบเขตของประชากรเป้าหมายและวิธีการเลือกกลุ่มตัวอย่างให้สอดคล้องกับวัตถุประสงค์การวิจัย เพื่อให้ได้ข้อมูลที่เป็นตัวแทนของแรงงานนอกระบบในประเทศไทยอย่างเหมาะสม มีรายละเอียด ดังนี้

3.1.1 ประชากร

ประชากรที่ใช้ในสำหรับการวิจัยในครั้งนี้ คือ ประชาชนที่อาศัยอยู่ในครัวเรือนส่วนบุคคล และครัวเรือนกลุ่มบุคคลประเภทครัวเรือนคนงานทุกครัวเรือนที่อาศัยอยู่ในเขตเทศบาลและนอกเขตเทศบาล ทุกจังหวัดทั่วประเทศ ยกเว้น ครัวเรือนชาวต่างประเทศที่มีเอกสิทธิ์ทางการฑูต ซึ่งเป็นผู้มีงานทำที่มีอายุ 15 ปีขึ้นไป ที่ไม่ได้รับความคุ้มครองตามกฎหมายและไม่มีหลักประกันทางสังคมจากการทำงาน ณ เดือนกรกฎาคม 2567

3.1.2 กลุ่มตัวอย่าง

วิธีการในการสุ่มตัวอย่างในโครงการสำรวจแรงงานนอกระบบของสำนักงานสถิติแห่งชาติใช้แผนการสุ่มตัวอย่างเป็นแบบ Stratified Two-Stage Sampling โดยมีจังหวัดเป็นสตราตัมเขตการแจงนับ (ในเขตเทศบาล และนอกเขตเทศบาล) เป็นหน่วยตัวอย่างขั้นที่หนึ่ง ครัวเรือนส่วนบุคคลและสมาชิกในครัวเรือนกลุ่มบบุคคลประเภทคนงาน เป็นหน่วยตัวอย่างขั้นที่ 2 งานวิจัยนี้ผู้วิจัยศึกษากลุ่มของแรงงานนอกระบบซึ่งเป็นผู้มีงานทำที่มีอายุ 15 ปีขึ้นไป และได้รับค่าตอบแทนจากการทำงานจึงมีจำนวนตัวอย่าง 43,703 คน

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล คือ โปรแกรม Python (version 3.12.2) โดยมีไลบรารีบน Python ที่ใช้ในงานวิจัย ดังนี้

ตารางที่ 3.1 ไลบรารีบน Python ที่ใช้ในงานวิจัย

ไลบรารี	คำอธิบาย
Statsmodels	ใช้ในการสร้างแบบจำลอง Logistic Regression ด้วยวิธีทางสถิติที่ให้ผลการวิเคราะห์เชิงสถิติโดยละเอียด เช่น p-value, Confidence Interval
Scikit-learn	ใช้ในการสร้างแบบจำลอง Logistic Regression และ Random Forest แบบ Machine Learning รวมถึงการประมวลผลข้อมูล (Preprocessing), การแบ่งข้อมูล, การประเมินผลด้วย Cross-Validation และการสร้าง Confusion Matrix
XGBoost	ใช้ในการสร้างแบบจำลอง XGBoost ซึ่งเป็น Gradient Boosting Algorithm ที่มีประสิทธิภาพสูงสำหรับงานจำแนกประเภท
Imbalanced-learn	ใช้ในการจัดการปัญหาข้อมูลไม่สมดุล โดยใช้วิธี SMOTEENN และ RandomUnderSampler รวมถึงสร้าง Pipeline ที่รองรับเทคนิคเหล่านี้
SHAP	ใช้ในการตีความและอธิบายแบบจำลอง (Model Interpretation) โดยแสดง Feature Importance และผลกระทบของแต่ละตัวแปรต่อการทำนาย
Pandas	ใช้ในการจัดการและประมวลผลข้อมูล เช่น อ่านไฟล์ CSV, สร้าง DataFrame เป็นต้น
NumPy	ใช้ในการคำนวณเชิงตัวเลข เช่น การทำ Log Transformation, Square Root Transformation และการจัดการ Array

3.3 การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้สำหรับการวิจัยครั้งนี้เป็นข้อมูลทุติยภูมิ (Secondary Data) จาก 2 แหล่งหลัก ได้แก่ ข้อมูลการสำรวจแรงงานนอกระบบ พ.ศ. 2567 ของสำนักงานสถิติแห่งชาติ ซึ่งดำเนินการสำรวจในระหว่างวันที่ 1-12 ของเดือนกรกฎาคม สิงหาคม และกันยายน พ.ศ. 2567 โดยใช้วิธีการสัมภาษณ์หัวหน้าครัวเรือนหรือสมาชิกในครัวเรือนตัวอย่าง ข้อมูลชุดนี้มีลักษณะเป็นข้อมูลรายบุคคล (Individual-Level Data) และข้อมูลสถิติดัชนีความก้าวหน้าของคน รายจังหวัด ปี 2566 ของสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ สำหรับข้อมูลที่นำมาใช้ในการวิจัยนี้ประกอบด้วย ตัวแปรดัชนีความไม่เสมอภาคด้านรายได้ และดัชนีความก้าวหน้าของคน ซึ่งเป็นข้อมูลระดับจังหวัด (Provincial-Level Data) ข้อมูลทั้งสองชุดถูกเชื่อมโยงผ่านรหัสจังหวัด โดยแรงงานแต่ละคนจะได้รับค่าดัชนีความก้าวหน้าของคนตามจังหวัดที่อาศัยอยู่ ทำให้สามารถวิเคราะห์ผลกระทบของปัจจัยระดับพื้นที่ต่อสถานะความยากจนของแรงงานนอกระบบได้ สำหรับความแตกต่างช่วงเวลาระหว่างข้อมูลทั้งสองชุดที่ห่างกัน 1 ปี (พ.ศ. 2566 และ 2567) ถือว่าสามารถยอมรับได้ เนื่องจากดัชนีความก้าวหน้าของคนเป็นตัวชี้วัดที่มีความเสถียรและเปลี่ยนแปลงค่อนข้างช้า ไม่เปลี่ยนแปลงอย่างรวดเร็วภายในระยะเวลาสั้น ดังนั้น งานวิจัยครั้งนี้มีตัวแปรอิสระทั้งหมด 23 ตัว ดังนี้

ตารางที่ 3.2 ตัวแปรที่ใช้ในงานวิจัย

ที่	ตัวแปร	คำอธิบาย	ชนิดของตัวแปร
	ปัจจัยด้านลักษณะส่วนบุคคล		
1	AGE	อายุ (ปี)	Ratio
2	SEX	เพศ	Nominal
3	MARITAL	สถานภาพสมรส	Nominal
4	EDU	ระดับการศึกษาสูงสุด	Ordinal
5	OCCUP	อาชีพ	Nominal
6	RELATION	สถานะหัวหน้าครัวเรือน	Nominal
7	REG	ภาค	Nominal
8	AREA	เขตการปกครอง	Nominal
9	MEMBERS	จำนวนสมาชิกในครัวเรือน (คน)	Ratio
	ปัจจัยด้านเศรษฐกิจ		
10	GINI_IDX	ดัชนีความไม่เสมอภาคด้านรายได้	Ratio
11	HAI	ดัชนีความก้าวหน้าของคน	Ratio
	ปัจจัยด้านสภาพการทำงานและค่าตอบแทน		
12	WAGE_TYPE	ประเภทค่าจ้างที่ได้รับ	Nominal
13	INDUS	ภาคกิจกรรมทางเศรษฐกิจ	Nominal
14	COND	การมีปัญหาจากสภาพแวดล้อมการทำงาน	Nominal
15	WORK_PROB	การมีปัญหาจากการทำงาน	Nominal
16	UNSAFE	การมีความเสี่ยงจากการทำงาน	Nominal
17	TOTAL_HR_MONTH	จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน (ชั่วโมง)	Ratio
18	BONUS	โบนัสรายปี (บาท)	Ratio
19	OT	ค่าล่วงเวลารายเดือน (บาท)	Ratio
20	OTH_THING	ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน (บาท)	Ratio
21	W_PLACE	ประเภทของสถานที่ทำงาน	Nominal
22	REGISTER	สถานประกอบการจดทะเบียนกับหน่วยงานรัฐ	Nominal
23	MIN_WAGE	ค่าจ้างขั้นต่ำต่อวัน (บาท)	Ratio

3.4 การเตรียมข้อมูล

การเตรียมข้อมูลเป็นขั้นตอนสำคัญในกระบวนการวิเคราะห์ข้อมูลที่มีผลต่อคุณภาพและความน่าเชื่อถือของผลการวิจัย ในการศึกษาครั้งนี้ ได้ดำเนินการตรวจสอบและจัดการข้อมูลอย่างเป็นระบบ เพื่อให้ข้อมูลมีความสมบูรณ์และเหมาะสมสำหรับการวิเคราะห์ด้วยวิธีทางสถิติและการเรียนรู้ของเครื่อง มีรายละเอียด ดังนี้

3.4.1 การทำความสะอาดข้อมูล

ข้อมูลทุติยภูมิที่ได้จากการสำรวจแรงงานนอกระบบ พ.ศ. 2567 ของสำนักงานสถิติแห่งชาติ พบว่าตัวแปรที่ใช้ในการวิจัยครั้งนี้มีข้อมูลในบางแถวมีข้อมูลขาดหาย (Missing Data) และบางตัวแปรมีข้อมูลที่เป็น Outlier ผู้วิจัยจึงเลือกตัดแถวดังกล่าวออก จึงทำให้มีจำนวนตัวอย่างทั้งสิ้น 42,662 คน

3.4.2 การเตรียมข้อมูล

หลังจากการทำความสะอาดข้อมูลแล้ว ได้จำนวนตัวอย่างทั้งสิ้น 42,662 คน โดยตัวแปรตามในการศึกษาครั้งนี้เป็นตัวแปรแบบไบนารี ซึ่งกำหนดจากเส้นความยากจนของสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติที่ระบุว่า แรงงานนอกระบบที่มีรายได้ต่อเดือนต่ำกว่าหรือเท่ากับ 3,043 บาทจะถูกจัดเป็นคนที่มีสถานะยากจน ดังนั้น ตัวแปรตามจึงมีค่าเป็น 1 เมื่อแรงงานนอกระบบมีรายได้ต่อเดือนน้อยกว่าหรือเท่ากับ 3,043 บาท และมีค่าเป็น 0 เมื่อมีรายได้ต่อเดือนมากกว่า 3,043 บาท

การจัดการตัวแปรอิสระ ในการศึกษาครั้งนี้มีตัวแปรอิสระแบ่งออกเป็น 2 ประเภทได้แก่ ตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ โดยตัวแปรเชิงคุณภาพจากตารางที่ 3.2 มีวิธีจัดการตัวแปรด้วยวิธีการที่แตกต่างกันตามลักษณะของข้อมูล ดังนี้

1) ตัวแปรอิสระเชิงคุณภาพที่เป็นข้อมูลนามบัญญัติ จัดการแปลงข้อมูลด้วยวิธี One-Hot Encoding ซึ่งเป็นการแปลงข้อมูลนามบัญญัติให้อยู่ในรูปแบบของข้อมูลเชิงตัวเลข โดยเปลี่ยนให้อยู่ในรูปแบบของ Binary Values ที่มีค่า 0 หรือ 1 เท่านั้น และใช้ค่าแรกของแต่ละตัวแปรเป็นตัวแปรอ้างอิง

2) ตัวแปรอิสระเชิงคุณภาพที่เป็นข้อมูลเชิงอันดับ ใช้การแปลงข้อมูลด้วยวิธี Ordinal Encoding เป็นการแปลงข้อมูลเชิงอันดับให้อยู่ในรูปแบบข้อมูลเชิงตัวเลขที่มีระยะห่างเท่ากัน จึงทำให้มีตัวแปรอิสระดังนี้

ตารางที่ 3.3 คำอธิบายตัวแปรอิสระหลังผ่านการเตรียมข้อมูล

ชื่อตัวแปร	ความหมาย	รหัสของตัวแปร
ปัจจัยด้านลักษณะส่วนบุคคล		
AGE	อายุ (ปี)	
SEX	เพศ	SEX0 = หญิง (ตัวแปรอ้างอิง) SEX1 = ชาย
MARITAL	สถานภาพสมรส	MARITAL0 = อื่นๆ (ตัวแปรอ้างอิง) MARITAL1 = โสด

ชื่อตัวแปร	ความหมาย	รหัสของตัวแปร
EDU	ระดับการศึกษาสูงสุด	ค่าของข้อมูล 1 = ไม่ได้เข้าศึกษา 2 = ต่ำกว่าระดับประถมศึกษา 3 = ประถมศึกษา 4 = มัธยมศึกษาตอนต้น 5 = มัธยมศึกษาตอนปลาย 6 = สูงกว่ามัธยมศึกษาตอนปลาย
OCCUP	อาชีพ	OCCUP1 = แรงงานฝีมือ (ตัวแปรอ้างอิง) OCCUP2 = แรงงานกึ่งฝีมือ OCCUP3 = แรงงานไร้ฝีมือ
RELATION	สถานะหัวหน้าครัวเรือน	RELATION0 = ไม่ใช่หัวหน้าครัวเรือน (ตัวแปรอ้างอิง) RELATION1 = หัวหน้าครัวเรือน
REG	ภาค	REG1 = ภาคตะวันออกเฉียงเหนือ (ตัวแปรอ้างอิง) REG2 = ภาคกลาง REG3 = ภาคเหนือ REG4 = ภาคใต้ REG5 = กรุงเทพมหานคร
AREA	เขตการปกครอง	AREA0 = นอกเขตเทศบาล (ตัวแปรอ้างอิง) AREA1 = ในเขตเทศบาล
MEMBERS	จำนวนสมาชิกในครัวเรือน (คน)	
ปัจจัยด้านเศรษฐกิจ		
GINI_IDX	ดัชนีความไม่เสมอภาคด้านรายได้	
HAI	ดัชนีความก้าวหน้าของคน	
ปัจจัยด้านสภาพการทำงานและค่าตอบแทน		
WAGE_TYPE	ประเภทค่าจ้างที่ได้รับ	WAGE_TYPE1 = รายวัน (ตัวแปรอ้างอิง) WAGE_TYPE2 = รายสัปดาห์ WAGE_TYPE3 = รายเดือน WAGE_TYPE4 = ไม่เป็นตัวเงิน WAGE_TYPE5 = อื่นๆ

ชื่อตัวแปร	ความหมาย	รหัสของตัวแปร
INDUS	ภาคกิจกรรมทางเศรษฐกิจ	INDUS1 = ภาคเกษตรกรรม (ตัวแปรอ้างอิง) INDUS2 = ภาคอุตสาหกรรม INDUS3 = ภาคบริการ
COND	การมีปัญหาจากสภาพแวดล้อมการทำงาน	COND0 = ไม่มีปัญหา (ตัวแปรอ้างอิง) COND1 = มีปัญหา
WORK_PROB	การมีปัญหาจากการทำงาน	WORK_PROB0 = ไม่มีปัญหา (ตัวแปรอ้างอิง) WORK_PROB1 = มีปัญหา
UNSAFE	การมีความเสี่ยงจากการทำงาน	UNSAFE0 = ไม่มีปัญหา (ตัวแปรอ้างอิง) UNSAFE1 = มีปัญหา
TOTAL_HR_MONTH	จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน (ชั่วโมง)	
BONUS	โบนัสรายปี (บาท)	
OT	ค่าล่วงเวลารายเดือน (บาท)	
OTH_THING	ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน (บาท)	
W_PLACE	ประเภทของสถานที่ทำงาน	W_PLACE1 = สถานประกอบการของนายจ้างหรือตนเอง (ตัวแปรอ้างอิง) W_PLACE2 = สถานที่ก่อสร้าง W_PLACE3 = ที่อยู่อาศัย W_PLACE4 = ไม่เป็นหลักแหล่ง W_PLACE5 = พื้นที่เพาะปลูก/พื้นที่เลี้ยงสัตว์
REGISTER	สถานประกอบการจดทะเบียนกับหน่วยงานรัฐ	REGISTER0 = ไม่จด (ตัวแปรอ้างอิง) REGISTER1 = จด
MIN_WAGE	ค่าจ้างขั้นต่ำต่อวัน (บาท)	

3.4.3 แบ่งข้อมูลเป็นข้อมูลฝึกหัด (Train Data) และข้อมูลทดสอบ (Test Data)

ในงานวิจัยครั้งนี้แบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลฝึกหัด (Train Data) เพื่อใช้สร้างแบบจำลอง และชุดข้อมูลทดสอบ (Test Data) เพื่อใช้ประเมินประสิทธิภาพของแบบจำลอง โดยแบ่งข้อมูลทั้ง 2 ส่วน ด้วยอัตราส่วน 80:20 ทำให้มีข้อมูลฝึกหัดจำนวน 34,129 คน และข้อมูลทดสอบ จำนวน 8,533 คน

3.4.4 การแก้ไขปัญหาข้อมูลไม่สมดุล

เนื่องจากชุดข้อมูลที่ใช้มีความไม่สมดุลกันของข้อมูลสูงมาก โดยจำนวนข้อมูลแรงงานนอก

ระบบที่มีสถานะไม่ยากจนมีมากกว่าจำนวนแรงงานนอกระบบที่มีสถานะยากจน ในการวิเคราะห์แบบจำลองด้วยการเรียนรู้ของเครื่อง (Machine Learning) จึงเลือกจัดการข้อมูลไม่สมดุลด้วย 2 วิธี ดังนี้

1) วิธี Random Undersampling เป็นการสุ่มลดจำนวนข้อมูลในกลุ่มแรงงานนอกระบบที่ไม่ยากจนให้มีสัดส่วนใกล้เคียงกับกลุ่มแรงงานนอกระบบที่ยากจน เพื่อป้องกันไม่ให้แบบจำลองเกิดอคติต่อกลุ่มที่มีข้อมูลมากกว่า

2) วิธี SMOTE-ENN เป็นวิธีสังเคราะห์ข้อมูลในกลุ่มแรงงานนอกระบบที่มีสถานะยากจนให้เพิ่มขึ้นใกล้เคียงกับกลุ่มแรงงานนอกระบบที่มีสถานะไม่ยากจน และทำการลดข้อมูลที่มีความผิดปกติหรือไม่สอดคล้องกับรูปแบบทั่วไปของข้อมูลที่อาจทำให้เกิดความสับสนออก

3.5 การพัฒนาแบบจำลอง

สำหรับงานวิจัยครั้งนี้ มีการพัฒนาแบบจำลองโดยใช้แนวทางที่แตกต่างกัน 2 วิธี ได้แก่ วิธีทางสถิติและวิธีการเรียนรู้ของเครื่อง เพื่อเปรียบเทียบประสิทธิภาพและเลือกแบบจำลองที่แม่นยำที่สุด

3.5.1 การพัฒนาแบบจำลองด้วยวิธีทางสถิติ

การพัฒนาแบบจำลองด้วยวิธีทางสถิติใช้การวิเคราะห์การถดถอยลอจิสติก มีขั้นตอนดังนี้

1) การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม
   1.1) การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงกลุ่ม (Dummy Variables) กับตัวแปรตาม ด้วย Chi-Square Test และ Cramer's V เพื่อวัดขนาดของความสัมพันธ์
   1.2) การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงปริมาณกับตัวแปรตาม ด้วย Point Biserial Correlation
2) การสร้างแบบจำลอง การคัดเลือกตัวแปรด้วยวิธี Stepwise Regression เพื่อหาชุดตัวแปรที่เหมาะสมที่สุด การประมาณค่าพารามิเตอร์ด้วย Maximum Likelihood Estimation การทดสอบนัยสำคัญทางสถิติของแต่ละตัวแปร
3) การตรวจสอบความเหมาะสมของแบบจำลอง และตรวจสอบข้อตกลงเบื้องต้น
4) ประเมินประสิทธิภาพของแบบจำลอง

3.5.2 การพัฒนาแบบจำลองด้วยการเรียนรู้ของเครื่อง (Machine Learning)

การพัฒนาแบบจำลองด้วยการเรียนรู้ของเครื่องใช้ 3 แบบจำลอง ได้แก่ Logistic Regression, Random Forest และ XGBoost ซึ่งมีขั้นตอนดังนี้

1) การเตรียมข้อมูล
   1.1) การประมวลผลข้อมูลเบื้องต้นที่ได้ดำเนินการไว้แล้ว (ข้อ 3.4.1-3.4.2)
   1.2) การแบ่งข้อมูลเป็นชุดข้อมูลฝึกหัด (Training Data) ชุดข้อมูลทดสอบ (Test Data)
   1.3) การจัดการข้อมูลไม่สมดุลในชุดข้อมูลฝึกหัดด้วย 2 วิธี ได้แก่ วิธี Random Undersampling และ SMOTE-ENN
   1.4) การปรับขนาดข้อมูลสำหรับตัวแปรเชิงปริมาณด้วย StandardScaler โดยแปลงให้มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1
2) การปรับแต่งพารามิเตอร์ นำเทคนิค Cross Validation มาใช้ในการค้นหาค่า Hyperparameter ที่เหมาะสม ใช้ GridSearchCV เพื่อหาพารามิเตอร์ที่ดีที่สุดสำหรับแต่ละแบบจำลอง และสำหรับแบบจำลอง Logistic Regression มีการปรับ Threshold

ที่เหมาะสม โดยอิงจากค่าประสิทธิภาพโดยรวม (F1-Score) บนชุดข้อมูลตรวจสอบ (Validation Data)

3) การพัฒนาแบบจำลองด้วยชุดข้อมูลฝึกหัด
4) การประเมินประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลทดสอบ จากนั้นทำการตรวจสอบปัญหา Overfitting

3.6 การประเมินผลแบบจำลอง

ในการศึกษาปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ การเลือกใช้เมตริกการประเมินที่เหมาะสมจึงมีความสำคัญอย่างยิ่ง การศึกษาครั้งนี้เลือกใช้ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F1-Score) เป็นเมตริกหลัก เนื่องจากกลุ่มแรงงานนอกระบบที่ยากจนเป็นกลุ่มเปราะบางที่ต้องการความช่วยเหลือจากภาครัฐ การที่แบบจำลองสามารถทำนายกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้อย่างครอบคลุม (ค่าความระลึกสูง) จึงมีความสำคัญมากกว่าการจำแนกที่แม่นยำแต่อาจพลาดกลุ่มแรงงานนอกระบบที่มีสถานะยากจนบางส่วน ในขณะที่ค่าประสิทธิภาพโดยรวม (F1-Score) ช่วยสร้างความสมดุลระหว่างความแม่นยำและค่าความระลึก เพื่อให้แบบจำลองสามารถระบุแรงงานนอกระบบที่ยากจนได้อย่างมีประสิทธิภาพ ซึ่งการใช้เมตริกเหล่านี้จะช่วยให้การวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกสถานะความยากจนมีความน่าเชื่อถือและนำไปสู่การวางแผนนโยบายที่มีประสิทธิภาพในการช่วยเหลือแรงงานนอกระบบได้อย่างเหมาะสม

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้ผู้วิจัยจะกล่าวถึงผลการวิเคราะห์การทำนายความยากจนของแรงงานนอกระบบ โดยใช้ชุดข้อมูลฝึกหัด (Train Data) ในการสร้างแบบจำลองด้วย 4 แบบจำลอง ได้แก่ แบบจำลอง Logistic Regression ด้วยวิธีทางสถิติ แบบจำลอง Logistic Regression ด้วยวิธีการเรียนรู้ของเครื่อง แบบจำลอง Random Forest และแบบจำลอง XGBoost จากนั้นนำชุดข้อมูลทดสอบ (Test Data) มาประเมินประสิทธิภาพของแบบจำลอง ซึ่งมีรายละเอียด ดังนี้

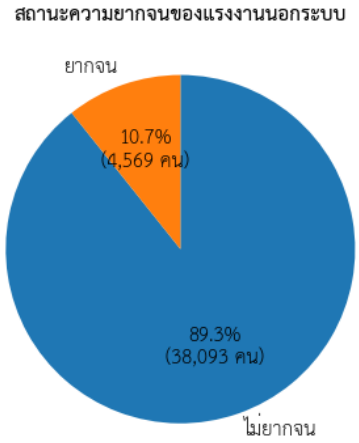
1. สถิติเชิงพรรณา (Descriptive Statistics)
2. ผลลัพธ์ของการพัฒนาแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ
3. ผลลัพธ์ของการพัฒนาแบบจำลองด้วยการเรียนรู้ของเครื่อง
   3.1) แบบจำลอง Logistic Regression
   3.2) แบบจำลอง Random Forest
   3.3) แบบจำลอง XGBoost
4. เปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุลทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test
5. เปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกความยากจนของแรงงานนอกระบบของทั้ง 4 แบบจำลอง
6. การวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP
7. อภิปรายผลการวิจัย

4.1 สถิติเชิงพรรณา (Descriptive Statistics)

ในการศึกษาครั้งนี้ มีการวิเคราะห์ข้อมูลแรงงานนอกระบบจำนวน 42,662 คน โดยกำหนดตัวแปรตาม คือ สถานะความยากจนของแรงงานนอกระบบ โดยแบ่งออกเป็น 2 กลุ่มด้วยเส้นความยากจน (3,043 บาทต่อเดือน) ดังนั้น ตัวแปรตามจึงมีค่าเป็น 1 เมื่อแรงงานนอกระบบมีรายได้ต่อเดือนไม่เกิน 3,043 บาท และมีค่าเป็น 0 เมื่อมีรายได้ต่อเดือนมากกว่า 3,043 บาท และตัวแปรอิสระประกอบด้วย อายุ เพศ สถานภาพสมรส ระดับการศึกษาสูงสุด อาชีพ สถานะหัวหน้าครัวเรือน ภาคเขตการปกครอง จำนวนสมาชิกในครัวเรือน ดัชนีความไม่เสมอภาคด้านรายได้ ดัชนีความก้าวหน้าของคน ประเภทค่าจ้างที่ได้รับ ภาคกิจกรรมทางเศรษฐกิจ การมีปัญหาจากสภาพแวดล้อมการทำงาน การมีปัญหาจากการทำงาน การมีความเสี่ยงจากการทำงาน จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน โบนัสรายปี ค่าล่วงเวลารายเดือน ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน ประเภทของสถานที่ทำงาน สถานประกอบการจดทะเบียนกับหน่วยงานรัฐ ค่าจ้างขั้นต่ำต่อวัน

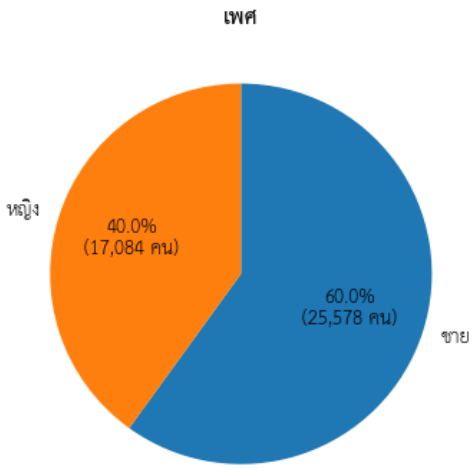
สถิติเชิงพรรณนา (Descriptive Statistics) ในส่วนนี้มีวัตถุประสงค์เพื่อแสดงให้เห็นถึงลักษณะทั่วไปและการกระจายตัวของแรงงานนอกระบบในประเทศไทย ทั้งในด้านคุณลักษณะส่วนบุคคล สภาพการทำงาน และสถานการณ์ทางเศรษฐกิจและสังคม ซึ่งจะช่วยให้เข้าใจบริบทและความหลากหลายของกลุ่มแรงงานนอกระบบก่อนทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต่างๆ

การนำเสนอข้อมูลในหัวข้อนี้จะจำแนกตามประเภทของตัวแปร โดยตัวแปรเชิงคุณภาพ (Qualitative Variables) ที่มีค่า 2-3 กลุ่ม จะนำเสนอในรูปแบบกราฟวงกลมพร้อมค่าร้อยละ เพื่อแสดงสัดส่วนของแต่ละกลุ่ม ส่วนตัวแปรเชิงคุณภาพที่มีหลายกลุ่มจะใช้กราฟแท่งแสดงจำนวนในแต่ละหมวดหมู่ สำหรับตัวแปรเชิงปริมาณ (Quantitative Variables) จะนำเสนอสถิติพื้นฐานในรูปแบบตาราง ประกอบด้วยค่าต่ำสุด ค่าสูงสุด ค่าเฉลี่ย และค่ามัธยฐาน เพื่อให้เห็นภาพรวมของการกระจายตัวของข้อมูลอย่างชัดเจน



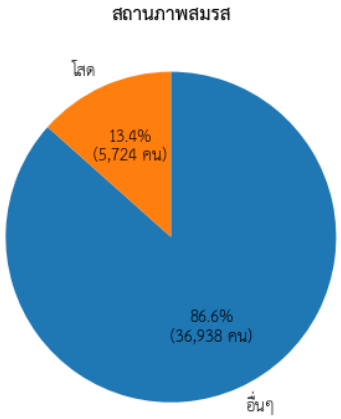
รูปที่ 4.1 สถานะความยากจนของแรงงานนอกระบบ

จากรูปที่ 4.1 แสดงสถานะของแรงงานนอกระบบ โดยแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่มแรงงานนอกระบบที่มีสถานะไม่ยากจน จำนวน 38,093 คน คิดเป็นร้อยละ 89.3 ของจำนวนแรงงานนอกระบบทั้งหมด และกลุ่มแรงงานที่มีสถานะยากจน จำนวน 4,569 คน คิดเป็นร้อยละ 10.7 ของจำนวนแรงงานนอกระบบทั้งหมด จากข้อมูลดังกล่าวจะเห็นได้ว่าในกลุ่มแรงงานนอกระบบซึ่งเป็นผู้ที่ไม่ได้รับความคุ้มครองตามกฎหมายและไม่มีหลักประกันทางสังคมจากการทำงาน ยังพบเจอปัญหาความยากจนซึ่งเป็นปัญหาสำคัญของการดำเนินชีวิต



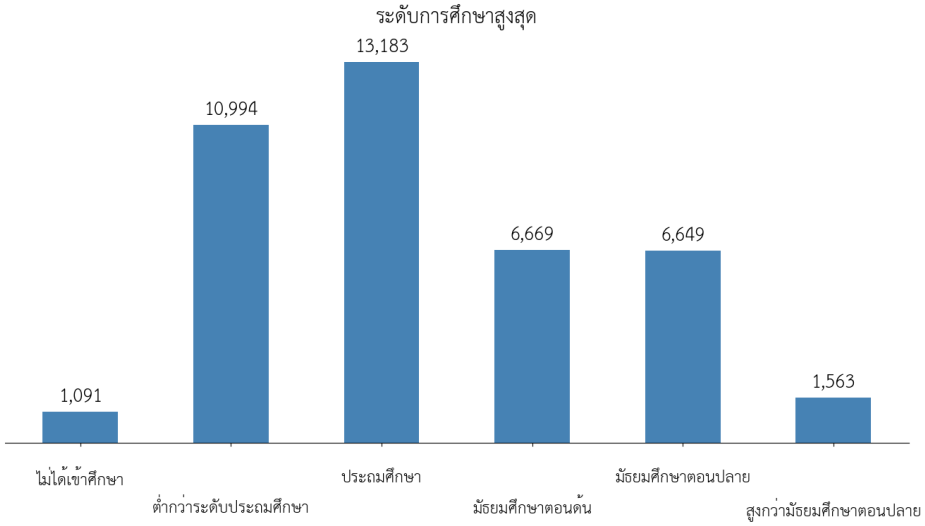
รูปที่ 4.2 เพศของแรงงานนอกระบบ

จากรูปที่ 4.2 เพศของแรงงานนอกระบบ แสดงให้เห็นว่า แรงงานนอกระบบในประเทศไทยมีสัดส่วนเพศชายมากกว่าเพศหญิง โดยเพศชาย คิดเป็น 60.0% (25,578 คน) และเพศหญิง คิดเป็น 40.0% (17,084 คน) จากกลุ่มตัวอย่างทั้งหมด 42,662 คน สะท้อนให้เห็นว่าแรงงานนอกระบบส่วนใหญ่เป็นเพศชาย ซึ่งอาจเกี่ยวข้องกับลักษณะงานที่ต้องใช้แรงงานทางกายภาพในภาคการก่อสร้าง เกษตรกรรม และอุตสาหกรรม



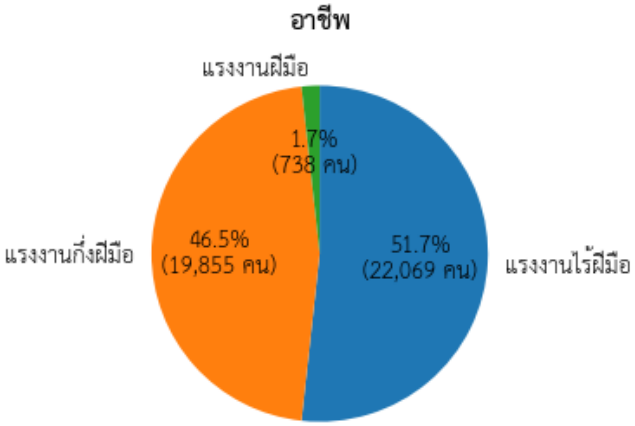
รูปที่ 4.3 สภานภาพสมรสของแรงงานนอกระบบ

จากงานวิจัยของ สุพนิดา (2565) ได้ศึกษาความยากจนในผู้สูงอายุไทย : การเปลี่ยนแปลงและปัจจัยเสี่ยง พบว่าผู้สูงอายุที่อยู่ในสถานะโสดจะมีโอกาสตกอยู่ในความยากจนมากกว่าสถานะอื่นๆ ผู้วิจัยจึงแบ่งข้อมูลสถานภาพการสมรสของแรงงานนอกระบบออกเป็น 2 กลุ่ม คือ กลุ่มที่มีสถานภาพสมรส โสด และสถานภาพสมรสอื่นๆ ซึ่งในที่นี้ หมายถึง สมรส ม่าย หย่า หรือแยกกันอยู่ โดยทั่วไปหมายถึงแต่งงานแล้วหรือเคยแต่งงาน จากรูปที่ 4.3 สถานภาพสมรสของแรงงานนอกระบบ แสดงให้เห็นว่าแรงงานนอกระบบในประเทศไทยส่วนใหญ่กว่า 86.6% มีสถานภาพสมรสเป็น “อื่นๆ” ขณะที่มีเพียง 13.4% ที่ยังโสด ข้อมูลนี้สะท้อนให้เห็นว่าแรงงานนอกระบบส่วนใหญ่อาจมีภาระครอบครัวและบทบาทในการเลี้ยงดูสมาชิกในครัวเรือน จึงมีความเสี่ยงสูงต่อความเปราะบางทางเศรษฐกิจจากรายได้ที่ไม่แน่นอน



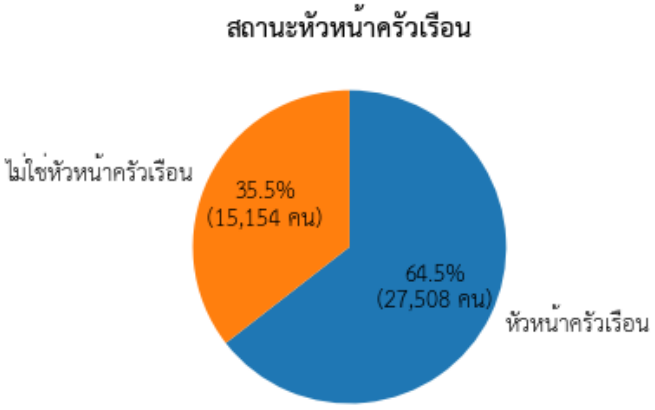
รูปที่ 4.4 ระดับการศึกษาสูงสุดของแรงงานนอกระบบ

จากรูปที่ 4.4 ระดับการศึกษาสูงสุดของแรงงานนอกระบบ พบว่า ระดับประถมศึกษามีจำนวนสูงสุด 13,183 คน ตามด้วย ต่ำกว่าประถมศึกษา 10,994 คน แสดงให้เห็นว่าแรงงานนอกระบบส่วนใหญ่มีระดับการศึกษาในระดับต่ำกว่าขั้นพื้นฐาน ขณะที่สูงกว่ามัธยมศึกษาตอนปลายมีเพียง 1,563 คน การกระจายนี้สะท้อนให้เห็นถึงความสัมพันธ์ระหว่างระดับการศึกษาและการเข้าสู่ตลาดแรงงานนอกระบบ โดยผู้ที่มีการศึกษาต่ำกว่าระดับการศึกษาขั้นพื้นฐานมักจะเข้าสู่งานนอกระบบเนื่องจากขาดทักษะหรือคุณวุฒิที่จำเป็นสำหรับการทำงานในระบบที่เป็นทางการ



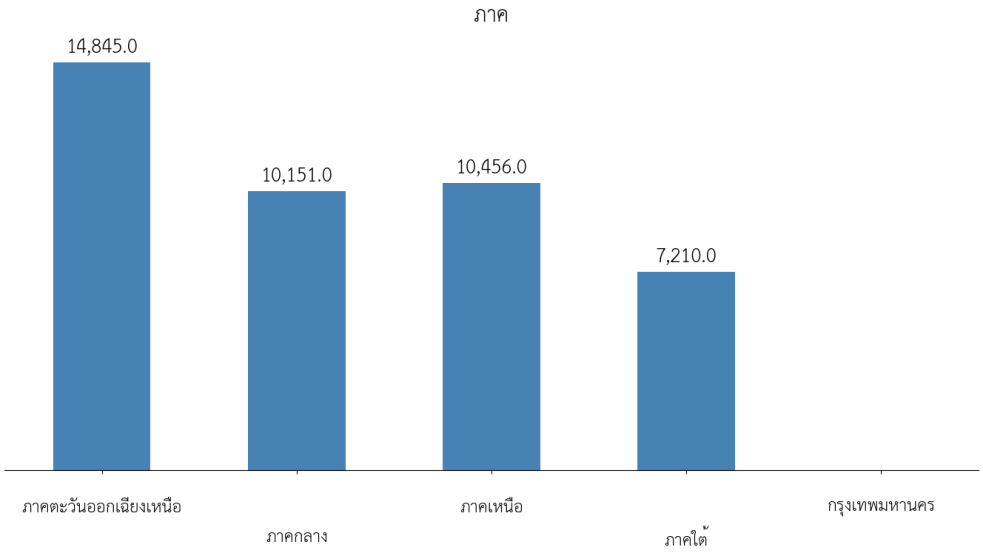
รูปที่ 4.5 อาชีพของแรงงานนอกระบบ

จากรูปที่ 4.5 อาชีพของแรงงานนอกระบบ แบ่งออกเป็น 3 กลุ่ม ได้แก่ แรงงานฝีมือ (ผู้จัดการ เสมียน เจ้าหน้าที่เทคนิค ผู้ประกอบวิชาชีพด้านต่างๆ) แรงงานกึ่งฝีมือ (พนักงานบริการ ผู้ประกอบอาชีพงานพื้นฐาน ผู้ควบคุมเครื่องจักรโรงงาน) และแรงงานไร้ฝีมือ (เกษตรกร ประมง) พบว่าแรงงานไร้ฝีมือมีสัดส่วนสูงสุด 51.7% (22,069 คน) ตามด้วยแรงงานกึ่งฝีมือ 46.5% (19,855 คน) และแรงงานฝีมือมีเพียง 1.7% (738 คน) การกระจายนี้สะท้อนให้เห็นว่าแรงงานนอกระบบของไทยส่วนใหญ่อยู่ในภาคเกษตรกรรมและงานที่ต้องใช้ทักษะระดับพื้นฐานถึงกลาง ขณะที่ผู้มีทักษะสูงมักเข้าสู่ตลาดแรงงานในระบบที่มีความมั่นคงมากกว่า



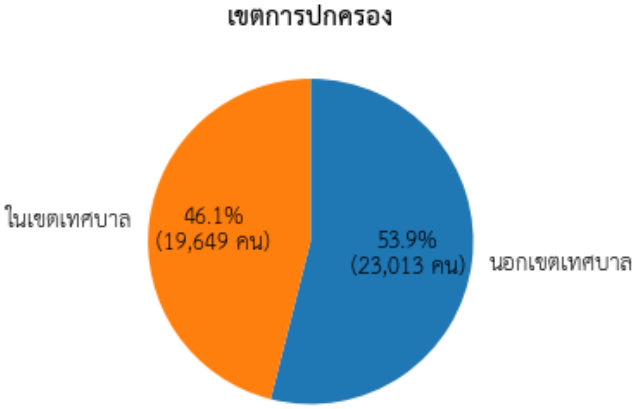
รูปที่ 4.6 สถานะหัวหน้าครัวเรือนของแรงงานนอกระบบ

จากรูปที่ 4.6 สถานะหัวหน้าครัวเรือนของแรงงานนอกระบบ พบว่า แรงงานนอกระบบที่เป็นหัวหน้าครัวเรือนมีสัดส่วน 64.5% (27,508 คน) ซึ่งมากกว่าแรงงานที่ไม่ใช่หัวหน้าครัวเรือน 35.5% (15,154 คน) การกระจายนี้สะท้อนให้เห็นว่าแรงงานนอกระบบส่วนใหญ่มีภาระรับผิดชอบหลักในการดูแลครอบครัว



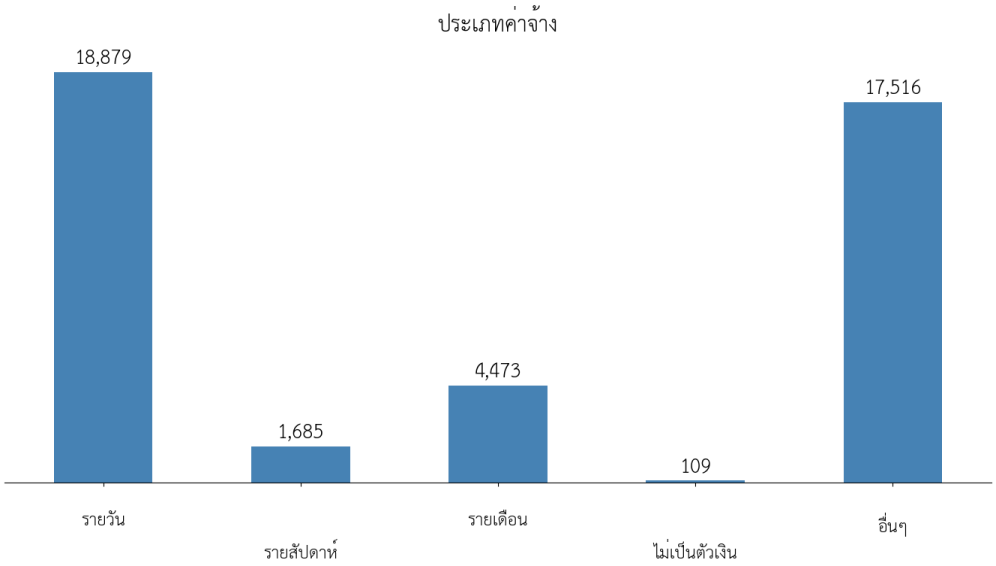
รูปที่ 4.7 ภาคที่อยู่อาศัยของแรงงานนอกระบบ

จากรูปที่ 4.7 ภาคที่อยู่อาศัยของแรงงานนอกระบบ แสดงการกระจายของแรงงานนอกระบบตามภูมิภาคต่างๆ ของประเทศไทย พบว่า ภาคตะวันออกเฉียงเหนือมีจำนวนแรงงานนอกระบบสูงสุด 14,845 คน ตามด้วยภาคกลาง 10,456 คน ภาคเหนือ 10,151 คน ภาคใต้ 7,210 คน และกรุงเทพมหานคร มีจำนวนน้อยที่สุด การกระจายนี้สะท้อนให้เห็นถึงความแตกต่างทางเศรษฐกิจและสังคมระหว่างภูมิภาค โดยภาคตะวันออกเฉียงเหนือซึ่งเป็นภูมิภาคที่มีฐานเศรษฐกิจหลักเป็นเกษตรกรรมจึงมีแรงงานนอกระบบมากที่สุด ขณะที่ภาคกรุงเทพมหานครเป็นพื้นที่ที่มีการพัฒนาอุตสาหกรรมและเศรษฐกิจที่เข้มแข็งกว่า จึงมีแรงงานนอกระบบน้อยที่สุด



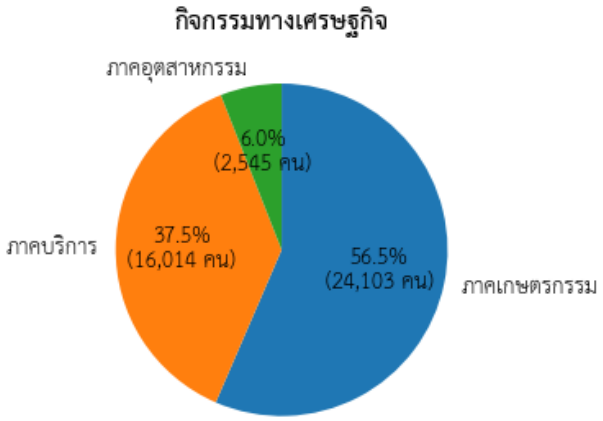
รูปที่ 4.8 เขตการปกครองที่อยู่อาศัยของแรงงานนอกระบบ

จากรูปที่ 4.8 เขตการปกครองที่อยู่อาศัยของแรงงานนอกระบบ แสดงการกระจายของแรงงานนอกระบบตามเขตการปกครองของประเทศไทย พบว่า แรงงานนอกระบบกระจายอยู่ทั้งในเขตเทศบาลและนอกเขตเทศบาลในสัดส่วนใกล้เคียงกัน



รูปที่ 4.9 ประเภทค่าจ้างที่ได้รับของแรงงานนอกระบบ

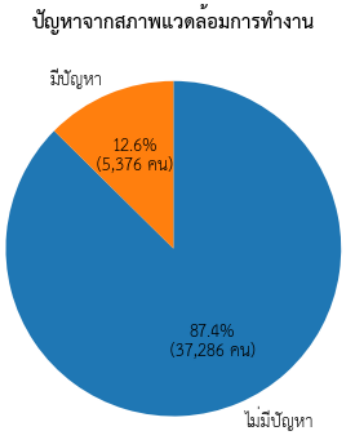
จากรูปที่ 4.9 ประเภทค่าจ้างที่ได้รับของแรงงานนอกระบบพบว่าแรงงานนอกระบบส่วนใหญ่ได้รับค่าจ้างเป็นรายวัน จำนวน 18,879 คน และอื่นๆ จำนวน 17,516 คน เนื่องจากแรงงานนอกระบบส่วนใหญ่ประกอบอาชีพเกษตรกร จึงทำให้ค่าตอบแทนของแรงงานนอกระบบส่วนใหญ่เป็นไปตามรอบฤดูกาลที่เก็บเกี่ยวผลผลิต จากรูปแสดงให้เห็นว่าแรงงานนอกระบบส่วนใหญ่มีลักษณะความหลากหลายและความไม่แน่นอนของระบบค่าตอบแทน และขาดรูปแบบรายได้ที่มีความต่อเนื่องและสวัสดิการที่มั่นคง



รูปที่ 4.10 กิจกรรมทางเศรษฐกิจของแรงงานนอกระบบ

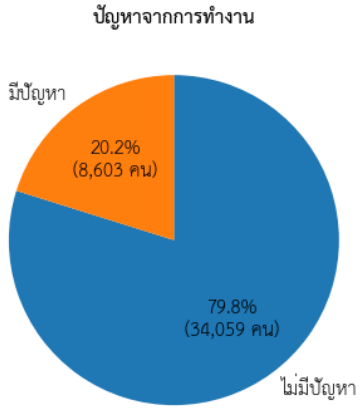
จากรูปที่ 4.10 กิจกรรมทางเศรษฐกิจของแรงงานนอกระบบ พบว่าแรงงานนอกระบบมากกว่าครึ่งนึงทำงานในภาคเกษตรกรรมคิดเป็นร้อยละ 56.5 ของแรงงานนอกระบบทั้งหมด รองลงมา คือ ภาคบริการ ร้อยละ 37.5 (16,014 คน) และภาคอุตสาหกรรมมีจำนวนเพียง 2,545 คน

หรือร้อยละ 6.0 เท่านั้น ข้อมูลนี้สะท้อนให้เห็นว่าแรงงานนอกระบบยังคงพึ่งพาภาคเกษตรกรรมเป็นหลัก ซึ่งเป็นภาคที่มีความเสี่ยงจากรายได้ไม่แน่นอนและความผันผวนของสภาพแวดล้อม



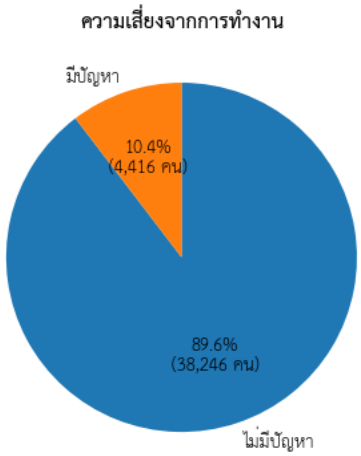
รูปที่ 4.11 การมีปัญหาจากสภาพแวดล้อมการทำงานของแรงงานนอกระบบ

จากรูปที่ 4.11 การมีปัญหาจากสภาพแวดล้อมการทำงานของแรงงานนอกระบบ พบว่าแรงงานนอกระบบที่ประสบปัญหาความไม่เหมาะสมของสถานที่และสภาพแวดล้อมทางกายภาพในการทำงาน ได้แก่ สถานที่ทำงานที่คับแคบ สกปรก อากาศไม่ถ่ายเท มีธรรมชาติของงานที่เต็มไปด้วยอิริยาบถที่ไม่เหมาะสมต่อสุขภาพ มีฝุ่นละออง ควัน กลิ่น เสียงดัง หรือแสงสว่างที่ไม่เหมาะสม รวมถึงปัญหาอื่นๆ ที่เกี่ยวข้องกับสภาพแวดล้อมทางกายภาพของสถานที่ทำงานมีจำนวน 5,376 คน คิดเป็นร้อยละ 12.6



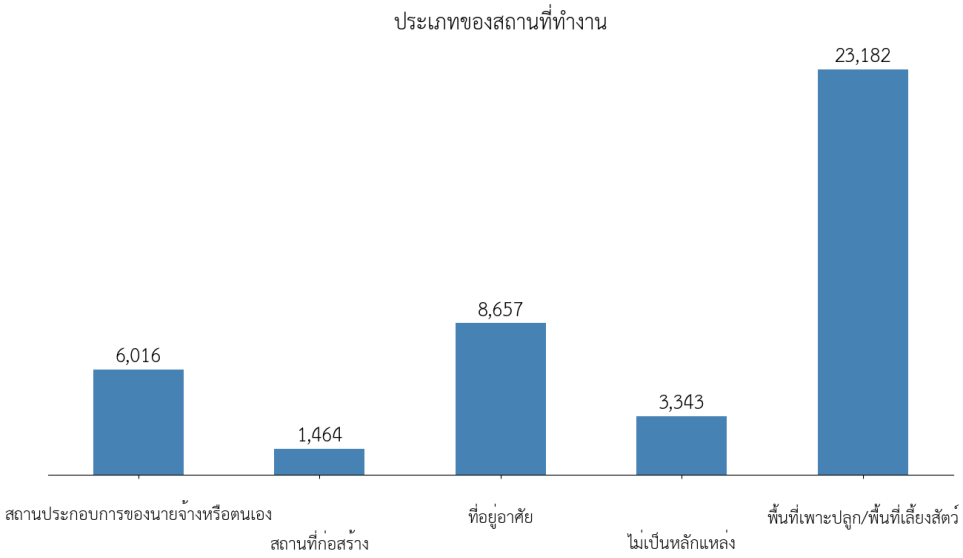
รูปที่ 4.12 การมีปัญหาจากการทำงานของแรงงานนอกระบบ

จากรูปที่ 4.12 การมีปัญหาจากการทำงานของแรงงานนอกระบบ พบว่าแรงงานนอกระบบที่ประสบปัญหา ความยากลำบากหรืออุปสรรคที่เกิดขึ้นจากลักษณะงานและการบริหารจัดการงาน ได้แก่ ปัญหาด้านค่าตอบแทนที่ไม่เพียงพอหรือไม่เป็นธรรม งานที่มีความหนักเกินไป การทำงานที่ไม่เป็นไปตามเวลาปกติ งานที่ขาดความต่อเนื่องหรือไม่มั่นคง ชั่วโมงการทำงานที่มากเกินไป การไม่มีวันหยุด หรือการไม่สามารถลาหยุดหรือลาพักผ่อนได้ตามต้องการ รวมถึงการไม่มีสวัสดิการที่เหมาะสมมีจำนวน 8,603 คน คิดเป็นร้อยละ 20.2



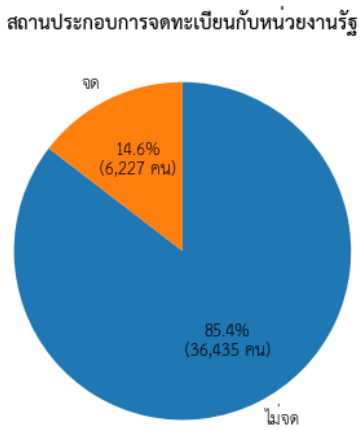
รูปที่ 4.13 การมีความเสี่ยงจากการทำงานของแรงงานนอกระบบ

จากรูปที่ 4.13 การมีความเสี่ยงจากการทำงานของแรงงานนอกระบบ พบว่าแรงงานนอกระบบที่ประสบปัญหา ความเสี่ยงและอันตรายที่อาจเกิดขึ้นในระหว่างการปฏิบัติงาน ได้แก่ การสัมผัสกับสารเคมีอันตราย การใช้เครื่องจักรหรือเครื่องมือที่เป็นอันตราย การเผชิญกับอันตรายต่อทุกส่วนของร่างกาย การทำงานในที่สูง/ใต้น้ำ/ใต้ดิน การเผชิญกับความไม่สงบหรือการก่อการร้าย รวมถึงความเสี่ยงด้านความปลอดภัยอื่นๆ ในการทำงาน มีจำนวน 4,416 คน คิดเป็นร้อยละ 10.4



รูปที่ 4.14 ประเภทสถานที่ทำงานของแรงงานนอกระบบ

จากรูปที่ 4.14 ประเภทสถานที่ทำงานของแรงงานนอก พบว่าแรงงานส่วนใหญ่ประกอบอาชีพในพื้นที่เพาะปลูกและพื้นที่เลี้ยงสัตว์ ซึ่งมีจำนวนสูงถึง 23,182 คน สะท้อนให้เห็นถึงบทบาทสำคัญของภาคเกษตรกรรมในกลุ่มแรงงานนอกระบบ รองลงมาคือกลุ่มที่ทำงานใน ที่อยู่อาศัยของตนเอง จำนวน 8,657 คน ซึ่งอาจรวมถึงงานค้าขายหรือการผลิตสินค้าภายในบ้าน ขณะที่แรงงานที่ทำงานในสถานประกอบการของนายจ้างหรือของตนเอง มีจำนวน 6,016 คน ส่วนกลุ่มที่ทำงานใน สถานที่ก่อสร้าง และสถานที่ไม่เป็นหลักแหล่ง มีจำนวนน้อยกว่า โดยอยู่ที่ 1,464 คน และ 3,343 คน ตามลำดับ



รูปที่ 4.15 สถานประกอบการที่ทำงานของแรงงานนอกกระบบจดทะเบียนกับหน่วยงานรัฐ

จากรูปที่ 4.15 สถานประกอบการที่ทำงานของแรงงานนอกกระบบจดทะเบียนกับหน่วยงานรัฐ พบว่าแรงงานนอกระบบกว่าร้อยละ 85 ทำงานในสถานประกอบการที่ไม่ได้จดทะเบียนกับหน่วยงานภาครัฐ

ตารางที่ 4.1 สถิติเชิงพรรณาของตัวแปรเชิงปริมาณ

ชื่อตัวแปร	ค่าเฉลี่ย	ค่ามัธยฐาน	ค่าต่ำสุด	ค่าสูงสุด
อายุ (AGE)	51.907	53.000	15.000	94.000
จำนวนสมาชิกในครัวเรือน (MEMBERS)	3.0877	3.000	1.000	15.000
ดัชนีความไม่เสมอภาคด้านรายได้ (GINI_IDX)	29.278	29.321	20.806	35.353
ดัชนีความก้าวหน้าของคน (HAI)	0.631	0.632	0.571	0.682
จำนวนชั่วโมงทำงานทั้งสิ้น (TOTAL_HR_MONTH)	177.400	173.200	8.660	424.340
โบนัส (BONUS)	16.509	0.000	0.000	300,000.000
ค่าล่วงเวลา (OT)	0.947	0.000	0.000	5,000.000
ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน (OTH_THING)	7.829	0.000	0.000	10000.000
ค่าจ้างขั้นต่ำ (MIN_WAGE)	343.778	343.000	330.000	370.000

4.2 ผลลัพธ์ของการพัฒนาแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ

ในการวิเคราะห์การถดถอยลอจิสติกเพื่อศึกษาปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ ผู้วิจัยได้แบ่งชุดข้อมูลออกเป็นสองชุด ได้แก่ ชุดข้อมูลฝึกหัด (Training Data) และชุดข้อมูลทดสอบ (Test Data) ในอัตราส่วน 80:20 โดยชุดข้อมูลฝึกหัดประกอบด้วยข้อมูลจำนวน 34,129 ราย และชุดข้อมูลทดสอบจำนวน 8,533 ราย เนื่องจากข้อมูลจริงมีลักษณะไม่สมดุล

จึงได้ใช้วิธีการสุ่มแบบมีการจัดชั้น (Stratified Sampling) เพื่อให้การแบ่งชุดข้อมูลทั้งสองสะท้อนสัดส่วนความยากจนและไม่ยากจนตามสภาพจริงอย่างเหมาะสม

ตัวแปรอิสระที่ใช้ในการวิเคราะห์ครั้งนี้ แบ่งออกเป็น 3 กลุ่ม ประกอบด้วย ตัวแปรอิสระเชิงกลุ่ม ตัวแปรเชิงอันดับ และตัวแปรเชิงปริมาณ ดังนี้

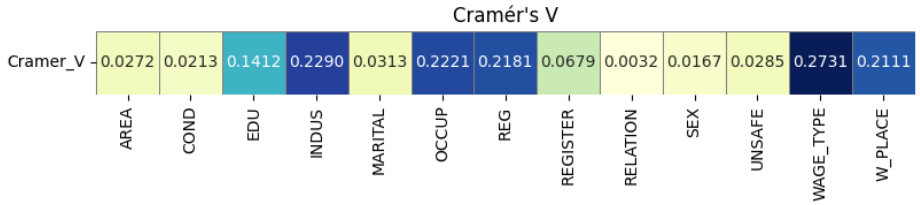
1) ตัวแปรอิสระเชิงกลุ่ม ประกอบด้วย เพศ (SEX) สถานภาพสมรส (MARITAL) อาชีพ (OCCUP) สถานะหัวหน้าครัวเรือน (RELATION) ภาค (REG) เขตการปกครอง (AREA) ประเภทค่าจ้างที่ได้รับ (WAGE_TYPE) ภาคกิจกรรมทางเศรษฐกิจ (INDUS) การมีปัญหาจากสภาพแวดล้อมการทำงาน (COND) การมีปัญหาจากการทำงาน (WORK_PROB) การมีความเสี่ยงจากการทำงาน (UNSAFE) ประเภทของสถานที่ทำงาน (W_PLACE) สถานประกอบการจดทะเบียนกับหน่วยงานรัฐ (REGISTER)
2) ตัวแปรเชิงอันดับ ประกอบด้วย ระดับการศึกษาสูงสุด (EDU)
3) ตัวแปรเชิงปริมาณ ประกอบด้วย อายุ (AGE) จำนวนสมาชิกในครัวเรือน (MEMBERS) ดัชนีความไม่เสมอภาคด้านรายได้ (GINI_IDX) ดัชนีความก้าวหน้าของคน (HAI) จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน (TOTAL_HR_MONTH) โบนัสรายปี (BONUS) ค่าล่วงเวลารายเดือน (OT) ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน (OTH_THING) ค่าจ้างขั้นต่ำต่อวัน (MIN_WAGE)

4.2.1 การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระกับสถานะความยากจน

1) การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงกลุ่ม (Dummy Variables) และตัวแปรเชิงอันดับกับสถานะความยากจน

สำหรับการการตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงกลุ่ม (Dummy Variables) และตัวแปรเชิงอันดับกับสถานะความยากจน ใช้สถิติทดสอบ Chi-Square Test และ Cramer's V เพื่อวัดขนาดของความสัมพันธ์ ผลการตรวจสอบความสัมพันธ์เป็น ดังนี้

จากการทดสอบ Chi-Square Test เพื่อทดสอบสมมติฐานว่าตัวแปรเชิงกลุ่ม ตัวแปรเชิงอันดับ มีความสัมพันธ์กับสถานะความยากจนของแรงงานนอกระบบหรือไม่ พบว่ามีตัวแปรการมีปัญหาจากการทำงาน (WORK_PROB) มีค่า p-value>0.05 ดังนั้น ตัวแปรการมีปัญหาจากการทำงานไม่มีความสัมพันธ์กับสถานะความยากจน ที่ระดับนัยสำคัญ 0.05 สำหรับตัวแปรอิสระอื่นๆ ที่มีความสัมพันธ์กับสถานะความยากจนวัดขนาดของความสัมพันธ์ (Effect Size) ได้ดังนี้

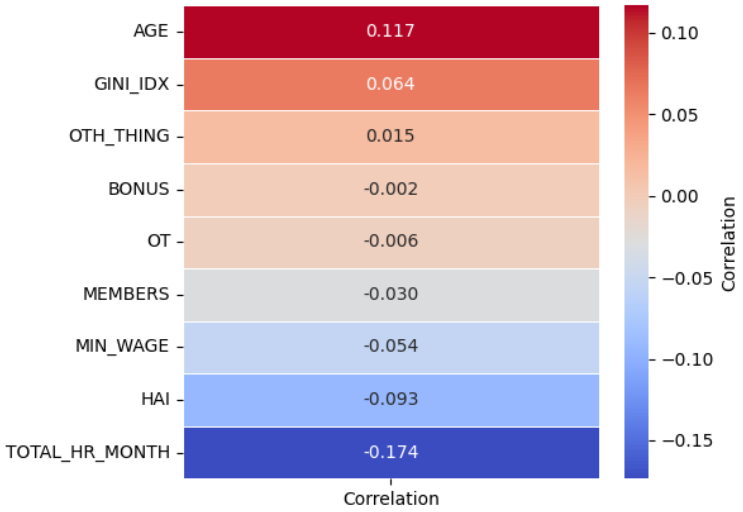


รูปที่ 4.16 ขนาดความสัมพันธ์ของตัวแปรอิสระเชิงกลุ่ม ตัวแปรเชิงอันดับและสถานะความยากจน

จากรูปที่ 4.16 จะเห็นว่าตัวแปรที่ขนาดของความสัมพันธ์กับสถานะความยากจนระดับปานกลาง (Cramér's V อยู่ระหว่าง 0.20 - 0.40) ประกอบด้วย ตัวแปรประเภทค่าจ้าง และตัวแปรอื่นๆ มีขนาดของความสัมพันธ์กับสถานะความยากจนน้อย

2) การตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงปริมาณกับสถานะความยากจน

สำหรับการตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระเชิงปริมาณกับสถานะความยากจน ใช้ Point Biserial Correlation เพื่อหาขนาดและทิศทางของความสัมพันธ์ ผลการตรวจสอบความสัมพันธ์เป็น ดังนี้



รูปที่ 4.17 ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับตัวแปรตาม

จากรูปที่ 4.17 สะท้อนให้เห็นว่า อายุ ดัชนีความไม่เสมอภาคด้านรายได้ ผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน มีความสัมพันธ์ในทิศทางเดียวกันกับตัวแปรตาม หรือถ้าตัวแปรเหล่านี้มีค่าเพิ่มขึ้น จะทำให้แรงงานนอกระบบตกอยู่ในสถานะยากจน ซึ่งตรงกันข้ามกับตัวแปรโบนัสรายปี ค่าล่วงเวลารายเดือน จำนวนสมาชิกในครัวเรือน ค่าจ้างขั้นต่ำต่อวัน ดัชนีความก้าวหน้าของคน จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน ถ้าตัวแปรเหล่านี้มีค่าเพิ่มขึ้นจะทำให้แรงงานนอกระบบมีโอกาสหลุดพ้นจากความยากจนมากขึ้น ซึ่งตัวแปรอิสระที่มีความสัมพันธ์กับสถานะความยากจนมากที่สุด คือ จำนวนชั่วโมงทำงานทั้งสิ้นรายเดือน รองลงมา คือ อายุ ดัชนีความก้าวหน้าของคน ดัชนีความไม่เสมอภาคด้านรายได้ ตามลำดับ

4.2.2 แปลงค่าตัวแปร

ในการวิเคราะห์การถดถอยลอจิสติกแบบทวิ (Binary Logistic Regression) ตามแนวคิดในสมการที่ (2.3) ตัวแปรอิสระเชิงปริมาณจำเป็นต้องมีความสัมพันธ์เชิงเส้นกับ Log(Odds) หรือค่า Logit ของตัวแปรตาม เพื่อให้สอดคล้องกับสมมติฐานพื้นฐานของแบบจำลองการถดถอยลอจิสติกที่แสดงในสมการที่ (2.1) จึงอาจจำเป็นต้องแปลงตัวแปรอิสระเพื่อให้เป็นไปตามข้อกำหนดดังกล่าว การศึกษานี้ใช้เทคนิค Box-Cox Transformation ในการแปลงตัวแปรเชิงปริมาณให้มีความสัมพันธ์เชิงเส้นกับ Logit และมีลักษณะการแจกแจงใกล้เคียงกับการแจกแจงแบบปกติมากขึ้น โดยพิจารณาจากค่าพารามิเตอร์ λ เพื่อเลือกวิธีการแปลงที่เหมาะสมกับลักษณะข้อมูลแต่ละตัวแปร อย่างไรก็ตาม Box-Cox Transformation สามารถประยุกต์ใช้ได้เฉพาะกับตัวแปรที่มีค่ามากกว่า 0 เท่านั้น ดังนั้น การศึกษานี้จึงเลือกใช้เทคนิคดังกล่าวเฉพาะกับตัวแปรที่ไม่มีค่าศูนย์หรือค่าติดลบ โดยผลการวิเคราะห์แสดงรายละเอียดในตารางที่ 4.2

ตารางที่ 4.2 การแปลงตัวแปรเชิงปริมาณด้วย เทคนิค Box-Cox Transformation

ตัวแปร	λ (Lambda)	วิธีการแปลง	ตัวแปรใหม่
อายุ (AGE)	1.4010	-	-
จำนวนสมาชิกในครัวเรือน (MEMBERS)	0.2916	รากที่สองของตัวแปร	รากที่สองของจำนวนสมาชิกในครัวเรือน (MEMBERS_sqrt)
จำนวนชั่วโมงทำงานทั้งสิ้น (TOTAL_HR_MONTH)	0.9332	-	-

จากการวิเคราะห์พบว่า ตัวแปรอายุ (AGE) และจำนวนชั่วโมงทำงานทั้งสิ้น (TOTAL_HR_MONTH) มีค่า λ ใกล้เคียงกับ 1 จึงไม่จำเป็นต้องแปลงค่าข้อมูล ขณะที่ตัวแปรจำนวนสมาชิกในครัวเรือน (MEMBERS) มีค่า λ เท่ากับ 0.2916 ซึ่งแตกต่างจาก 1 อย่างมีนัยสำคัญ ผู้วิจัยจึงเลือกแปลงตัวแปรดังกล่าวโดยใช้รากที่สองของค่าตัวแปร ทำให้ได้ตัวแปรใหม่คือ รากที่สองของจำนวนสมาชิกในครัวเรือน (MEMBERS_sqrt)

สำหรับตัวแปรระดับจังหวัด ได้แก่ ดัชนีความไม่เสมอภาคด้านรายได้ (GINI_IDX), ดัชนีความก้าวหน้าของคน (HAI) และค่าจ้างขั้นต่ำต่อวัน (MIN_WAGE) ซึ่งเป็นข้อมูลระดับจังหวัดที่ไม่มีค่าติดลบ และมีการกระจายของข้อมูลอยู่ในช่วงแคบโดยไม่แสดงลักษณะเบ้ที่รุนแรง จึงไม่ทำการแปลงข้อมูล

ขณะเดียวกัน ตัวแปรเชิงปริมาณที่มีค่าตั้งแต่ 0 ได้แก่ โบนัสรายปี (BONUS), ค่าล่วงเวลารายเดือน (OT) และผลประโยชน์ตอบแทนที่ไม่เป็นตัวเงิน (OTH_THING) จากตารางที่ 4.1 พบว่าตัวแปรเหล่านี้มีการแจกแจงแบบเบ้ขวา (Right-Skewed) อย่างชัดเจน ผู้วิจัยจึงทำการแปลงค่าด้วย log(x + 1) เพื่อช่วยลดความเบ้และเพิ่มประสิทธิภาพในการวิเคราะห์ โดยได้ตัวแปรใหม่คือ BONUS_log, OT_log และ OTH_THING_log

4.2.3 ผลการสร้างแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ

ในการวิเคราะห์ครั้งนี้ ได้ทำการสร้างแบบจำลองถดถอยลอจิสติกโดยใช้ชุดข้อมูลฝึกหัด (Training Data) โดยมีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลการจำแนกความยากจนของแรงงานนอกระบบ โดยใช้วิธีการคัดเลือกตัวแปรด้วยเทคนิค Stepwise Selection ซึ่งเป็นกระบวนการคัดเลือกตัวแปรแบบผสมระหว่าง Forward Selection และ Backward Elimination ภายใต้เกณฑ์ของค่าความมีนัยสำคัญทางสถิติ จากการวิเคราะห์โดยเริ่มต้นด้วยตัวแปรอิสระจำนวน 22 ตัวแปร ได้ผลลัพธ์ดังนี้

ตารางที่ 4.3 ผลการวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ

ตัวแปรอิสระ	B	S.E.	Wald	Sig	EXP(B)
const	-1.799	0.073	605.704	0.000	0.165
REG_2	-1.327	0.066	406.471	0.000	0.265
REG_3	-0.845	0.049	294.334	0.000	0.430
REG_4	-1.134	0.072	244.685	0.000	0.322
RELATION_1	-0.325	0.042	58.409	0.000	0.723
SEX_1	-0.346	0.041	72.791	0.000	0.708
WAGE_TYPE_4	3.247	0.238	186.751	0.000	25.708
WAGE_TYPE_5	1.439	0.051	808.801	0.000	4.217
W_PLACE_3	0.450	0.069	42.533	0.017	1.568
W_PLACE_4	0.200	0.097	4.270	0.039	1.221
UNSAFE_1	-0.658	0.069	90.394	0.000	0.518
EDU	-0.160	0.019	72.208	0.000	0.852
AGE	0.221	0.026	72.141	0.000	1.247
OTH_THING_log	0.094	0.014	43.510	0.000	1.098
TOTAL_HR_MONTH	-0.507	0.022	538.814	0.000	0.602
HAI	-0.149	0.020	54.949	0.000	0.861
GINI_IDX	0.263	0.021	160.325	0.000	1.301

หมายเหตุ Sig ใช้ในการพิจารณาสถิติทดสอบวอลด์

แบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ

$$\log(\text{odds}) = \log\left(\frac{P(y)}{Q(y)}\right) = -1.799 - 1.327\mathbf{REG\_2} - 0.845\mathbf{REG\_3} - 1.134\mathbf{REG\_4}$$
$$- 0.325\mathbf{RELATION\_1} - 0.346\mathbf{SEX\_1} + 3.247\mathbf{WAGE\_TYPE\_4}$$
$$+ 1.439\mathbf{WAGE\_TYPE\_5} + 0.450\mathbf{W\_PLACE\_3} + 0.200\mathbf{W\_PLACE\_4}$$
$$- 0.658\mathbf{UNSAFE\_1} - 0.160\mathbf{EDU} + 0.221\mathbf{AGE} + 0.094\mathbf{OTH\_THING\_log}$$
$$- 0.507\mathbf{TOTAL\_HR\_MONTH} - 0.149\mathbf{HAI} + 0.263\mathbf{GINI\_IDX}$$

สำหรับการแปลผลการวิเคราะห์ ตัวแปรอิสระเชิงกลุ่มได้ถูกแปลงเป็นตัวแปรดัมมี่ ซึ่งตัวแปรตัวแรกของแต่ละกลุ่มจะถูกกำหนดให้เป็น ตัวแปรอ้างอิง (Reference Group) ซึ่งไม่ได้นำเข้าสู่แบบจำลองโดยตรง เพื่อป้องกันปัญหา Multicollinearity โดยมีรายละเอียดของกลุ่มอ้างอิงในแต่ละตัวแปรดังนี้

REG (ภาค) ใช้ REG_1 (ภาคตะวันออกเฉียงเหนือ) เป็นกลุ่มอ้างอิง
REG2 (ภาคกลาง), REG3 (ภาคเหนือ), REG4 (ภาคใต้),
REG5 (กรุงเทพมหานคร)

SEX (เพศ) ใช้ SEX_0 (หญิง) เป็นกลุ่มอ้างอิง SEX1 = ชาย

WAGE_TYPE (ประเภทค่าจ้างที่ได้รับ) ใช้ WAGE_TYPE_1 (ค่าจ้างรายวัน) เป็นกลุ่มอ้างอิง

WAGE_TYPE2 (รายสัปดาห์), WAGE_TYPE3 (รายเดือน),
WAGE_TYPE4 (ไม่เป็นตัวเงิน), WAGE_TYPE5 (อื่นๆ)

W_PLACE (สถานที่ทำงาน) ใช้ W_PLACE_1 (สถานประกอบการของนายจ้างหรือตนเอง) เป็นกลุ่มอ้างอิง

W_PLACE2 (สถานที่ก่อสร้าง), W_PLACE3 (ที่อยู่อาศัย),
W_PLACE4 (ไม่เป็นหลักแหล่ง), W_PLACE5 (พื้นที่เพาะปลูก/พื้นที่เลี้ยงสัตว์)

UNSAFE (ความเสี่ยงจากการทำงาน) ใช้ UNSAFE_0 (ไม่มีความเสี่ยงจากการทำงาน) เป็นกลุ่มอ้างอิง และ UNSAFE1 (มีความเสี่ยงจากการทำงาน)

RELATION (สถานะหัวหน้าครัวเรือน) ใช้ RELATION_0 (ไม่ใช่หัวหน้าครัวเรือน) เป็นกลุ่มอ้างอิง และRELATION1 (หัวหน้าครัวเรือน)

จากแบบจำลองพบว่า ตัวแปรจำนวนมากมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 โดยเฉพาะตัวแปรกลุ่มภูมิภาค (REG) พบว่าผู้ที่อยู่ในภาคกลาง (REG_2), ภาคเหนือ (REG_3) และภาคใต้ (REG_4) มีโอกาสยากจนน้อยกว่าตะวันออกเฉียงเหนือ โดยมีอัตราส่วนความเป็นไปได้ (Exp(B)) เท่ากับ 0.265, 0.430 และ 0.322 ตามลำดับ

ในด้านความสัมพันธ์กับหัวหน้าครัวเรือน (RELATION_1) พบว่าผู้ที่เป็นหัวหน้าครัวเรือนมีโอกาสยากจนน้อยกว่าผู้ที่ไม่ใช่หัวหน้าครัวเรือนโดยมี Exp(B) = 0.723 เช่นเดียวกับเพศชาย (SEX_1) ซึ่งมีโอกาสยากจนน้อยกว่าเพศหญิง โดยมี Exp(B) = 0.708 อย่างมีนัยสำคัญ

สำหรับประเภทค่าจ้างที่ได้รับ (WAGE_TYPE) พบว่า ผู้ที่ได้รับค่าจ้างแบบไม่เป็นตัวเงิน (WAGE_TYPE_4) และผู้ที่มีรูปแบบค่าจ้างอื่น ๆ (WAGE_TYPE_5) มีโอกาสยากจนมากกว่าผู้ที่ได้รับค่าจ้างรายวันโดยมีค่า Exp(B) เท่ากับ 25.708 และ 4.217 ตามลำดับ ซึ่งอาจสะท้อนถึงลักษณะงานที่ไม่มั่นคงหรือความผันผวนของรายได้ในกลุ่มนี้

ในด้านสถานที่ทำงาน (W_PLACE) พบว่าแรงงานที่ทำงานภายนอกสถานประกอบการของตนเองหรือนายจ้าง ได้แก่ กลุ่มที่ทำงานในที่อยู่อาศัยของตนเอง (W_PLACE_3) และกลุ่มที่ทำงานไม่เป็นหลักแหล่ง (W_PLACE_4) มีโอกาสยากจนมากกว่าผู้ที่ทำงานภายในสถานประกอบการของนายจ้างหรือ โดยมีค่า Exp(B) = 1.568 และ 1.221 ตามลำดับ สำหรับแรงงานนอกระบบที่มีปัญหาความเสี่ยงจากการทำงาน (UNSAFE_1) โอกาสยากจนน้อยกว่าผู้ที่ทำงานไม่เสี่ยง โดยมีค่า Exp(B) = 0.518

ในส่วนของตัวแปรเชิงปริมาณ พบว่า อายุ (AGE) มี Exp(B) = 1.247 แสดงว่าเมื่ออายุเพิ่มขึ้น 1 ปี โอกาสยากจนจะเพิ่มขึ้นประมาณ 24.7% ส่วนจำนวนชั่วโมงทำงานทั้งสิ้นต่อเดือน TOTAL_HR_MONTH มี Exp(B) = 0.602 แสดงว่าเมื่อจำนวนชั่วโมงทำงานต่อเดือนเพิ่มขึ้น โอกาสยากจนจะลดลงอย่างมีนัยสำคัญ นอกจากนี้ ตัวแปรดัชนีความก้าวหน้าของคนมีแนวโน้มลดโอกาสความยากจน (Exp(B) = 0.861) ในขณะที่ดัชนีความไม่เสมอภาคด้านรายได้มีแนวโน้มเพิ่มโอกาสความยากจน (Exp(B) = 1.301)

นอกจากนี้ ตัวแปร OTH_THING_log (log ของผลตอบแทนที่ไม่ใช่เงิน), และระดับการศึกษาสูงสุด (EDU) ต่างก็มีนัยสำคัญในแบบจำลอง โดยเฉพาะตัวแปรระดับการศึกษาสูงสุดซึ่งมี Exp(B) = 0.852 แสดงว่าเมื่อระดับการศึกษาเพิ่มขึ้น โอกาสตกอยู่ในสถานะความยากจนจะลดลง

ผลการวิเคราะห์แสดงให้เห็นว่า ปัจจัยด้านบุคคล เศรษฐกิจ สภาพแวดล้อมในการทำงาน และค่าตอบแทน มีอิทธิพลอย่างชัดเจนต่อความเสี่ยงในการตกอยู่ในภาวะยากจนของแรงงานนอก

ระบบ ซึ่งผลการศึกษานี้สามารถนำไปใช้เป็นแนวทางในการกำหนดนโยบายที่เหมาะสมและตรงกลุ่มเป้าหมาย เพื่อสนับสนุนแรงงานนอกระบบที่มีแนวโน้มยากจนได้อย่างมีประสิทธิภาพ

4.2.4 การตรวจสอบความเหมาะสมของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก

ในการประเมินความเหมาะสมของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ใช้สถิติทดสอบ Hosmer–Lemeshow โดยผลการทดสอบพบว่า

H_0: แบบจำลองเหมาะสม

H_1: แบบจำลองไม่เหมาะสม

Hosmer–Lemeshow Test: χ^2 = 8.9150, df = 8, p = 0.3495

เนื่องจาก p-value >0.05 จึงยอมรับสมมติฐานหลัก แสดงว่า แบบจำลองมีความเหมาะสม ที่ระดับนัยสำคัญ 0.05

สถิติทดสอบวอลด์ (Wald Statistics) จากตารางที่ 4.3 ทุกตัวแปรที่อยู่ในแบบจำลองมีค่า Sig น้อยกว่า 0.05 นั่นคือ ตัวแปรอิสระในแบบจำลองมีผลต่อความยากจนของแรงงานนอกระบบ

สถิติทดสอบ Cox & Snell R^2 มีค่าเท่ากับ 0.1297 หมายความว่าแบบจำลองการวิเคราะห์การถดถอยลอจิสติกสามารถอธิบายความผันแปรของความยากจนของแรงงานนอกระบบได้ร้อยละ 12.97

สถิติทดสอบ Nagelkerke R^2 มีค่าเท่ากับ 0.2627 หมายความว่าแบบจำลองการวิเคราะห์การถดถอยลอจิสติกสามารถอธิบายความผันแปรของความยากจนของแรงงานนอกระบบได้ร้อยละ 26.27

4.2.5 การตรวจสอบข้อตกลงเบื้องต้นของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก

จากตารางที่ 4.4 การตรวจสอบระหว่างความสัมพันธ์ระหว่างตัวแปรอิสระ ด้วย VIF พบว่าตัวแปรเชิงปริมาณทั้งหมดมีค่า VIF น้อยกว่า 5 ดังนั้น จึงไม่เกิดปัญหา Multicollinearity

ตารางที่ 4.4 การตรวจสอบระหว่างความสัมพันธ์ระหว่างตัวแปรอิสระ ด้วย VIF

ตัวแปรอิสระ	AGE	OTH_THING_log	TOTAL_HR_MONTH	HAI	GINI_IDX
VIF	1.500	1.005	1.158	1.237	1.078

4.2.6 การประเมินประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติ

สำหรับการประเมินประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกบนชุดข้อมูลฝึกหัด ได้มีการวัดประสิทธิภาพของแบบจำลองภายใต้ 2 เงื่อนไข คือ Threshold = 0.5 และ Threshold ที่ทำให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงที่สุด (Threshold = 0.2341) โดยจะใช้ค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) ในการประเมินประสิทธิภาพของแบบจำลองเป็นหลัก

ตารางที่ 4.5 การจำแนกประเภทของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ โดยใช้ชุดข้อมูลฝึกหัด

แบบจำลอง	Logistic Regression Threshold = 0.5	Logistic Regression Threshold = 0.2341
Balanced Accuracy	0.5433	**0.7084**
Accuracy	**0.8950**	0.8543
F1-Score	0.1634	**0.4367**
Recall	0.0958	**0.5275**
Precision	**0.5573**	0.3726

จากตารางที่ 4.5 ซึ่งแสดงผลการจำแนกประเภทของแบบจำลองการถดถอยลอจิสติกด้วยวิธีทางสถิติ โดยใช้ชุดข้อมูลฝึกหัด พบว่า เมื่อเปรียบเทียบการตั้งค่า Threshold แบบคงที่ที่ 0.5 กับการปรับค่า Threshold เป็น 0.2038 ซึ่งเป็นค่าที่ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงสุดบนชุดข้อมูลฝึกหัด พบว่าการปรับค่า Threshold ดังกล่าวช่วยเพิ่มประสิทธิภาพของแบบจำลองได้อย่างชัดเจน โดยค่าความถูกต้องสมดุล (Balanced Accuracy) เพิ่มขึ้นจาก 0.5433 เป็น 0.7084 และค่าความระลึก (Recall) เพิ่มขึ้นจาก 0.0958 เป็น 0.5275 ซึ่งสะท้อนถึงความสามารถของแบบจำลองที่ดีขึ้นในการทำนายกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้อย่างถูกต้องมากขึ้น นอกจากนี้ ค่าประสิทธิภาพโดยรวม (F1-Score) ซึ่งเป็นค่าที่สะท้อนถึงความสมดุลระหว่างความแม่นยำ (Precision) และค่าความระลึก (Recall) เพิ่มขึ้นจาก 0.1632 เป็น 0.4367 แสดงให้เห็นว่าการปรับค่า Threshold ให้เหมาะสม สามารถเพิ่มประสิทธิภาพโดยรวมของแบบจำลอง ในการจำแนกแรงงานนอกระบบที่อยู่ในสถานะยากจนได้อย่างมีประสิทธิภาพ โดยเฉพาะในบริบทของข้อมูลที่ไม่สมดุล

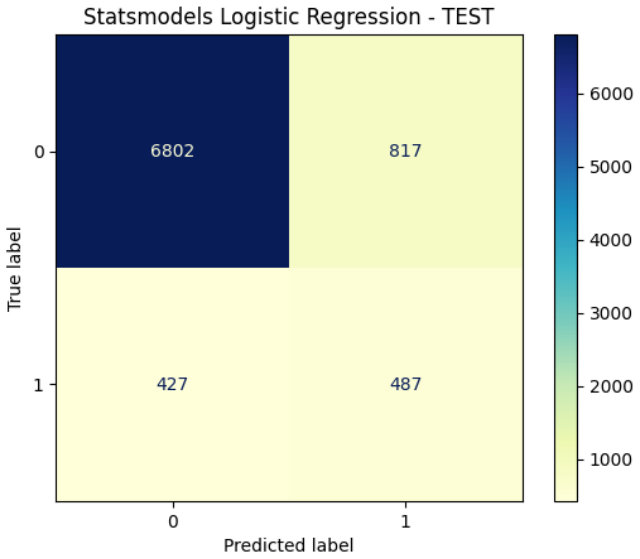
โดยสรุป การปรับค่า Threshold จากค่าเริ่มต้น 0.5 เป็น 0.2341 ในแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ส่งผลให้ประสิทธิภาพในการทำนายสถานะความยากจนของแรงงานนอกระบบดีขึ้นอย่างเห็นได้ชัด โดยเฉพาะค่าประสิทธิภาพโดยรวม (F1-Score) ที่เพิ่มขึ้นมากกว่า 27% เมื่อเปรียบเทียบกับการใช้ Threshold แบบเดิม ซึ่งแสดงให้เห็นว่าการปรับค่า Threshold ให้เหมาะสมช่วยเพิ่มความสามารถของแบบจำลองในการระบุกลุ่มแรงงานนอกระบบที่อยู่ในสถานะยากจนได้อย่างมีประสิทธิภาพมากขึ้น โดยเฉพาะภายใต้สถานการณ์ที่ข้อมูลมีความไม่สมดุลกัน

จากผลการศึกษาดังกล่าว จึงได้นำแบบจำลองการวิเคราะห์การถดถอยลอจิสติกที่ปรับค่า Threshold เป็น 0.2341 มาทดสอบกับชุดข้อมูลทดสอบเพื่อประเมินประสิทธิภาพของแบบจำลอง โดยผลลัพธ์ที่ได้นำเสนอไว้ในตารางที่ 4.6

ตารางที่ 4.6 การจำแนกประเภทของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ (ปรับ Threshold = 0.2341) โดยใช้ชุดข้อมูลทดสอบ

แบบจำลอง	**Logistic Regression Threshold = 0.2341**
Balanced Accuracy	0.7128
Accuracy	0.8542
F1-Score	0.4391
Recall	0.5328
Precision	0.3735

ผลการวิเคราะห์การจำแนกประเภท บนชุดข้อมูลทดสอบแสดงให้เห็นว่า ประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติ และปรับ Threshold เท่ากับ 0.2341 สามารถจำแนกกลุ่มสถานะความยากจนของแรงงานนอกระบบโดยมีค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.4391 (43.91%) และค่าความระลึก (Recall) มีค่า 0.5328 (53.29%) ซึ่งอยู่ในระดับปานกลาง แสดงให้เห็นว่าแบบจำลองสามารถทำนายแรงงานนอกระบบที่มีสถานะยากจนได้ถูกต้องประมาณครึ่งหนึ่งจากแรงงานนอกระบบที่มีสถานะยากจนทั้งหมด และแบบจำลองไม่เกิดปัญหา Overfitting เพราะค่าประสิทธิภาพโดยรวม (F1-Score) จากชุดข้อมูลฝึกหัดและทดสอบมีค่าใกล้เคียงกัน



รูปที่ 4.18 เมทริกซ์วัดประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีทางสถิติ ปรับ Threshold = 0.2341

จากรูปที่ 4.18 เมทริกซ์วัดประสิทธิภาพของแบบจำลองการวิเคราะห์การถดถอยลอจิสติก ด้วยวิธีทางสถิติ ปรับ Threshold = 0.2341 บนชุดข้อมูลทดสอบ พบว่าแบบจำลองทำนายกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้ถูกต้อง 487 คน จาก 914 คน และมีแรงงานนอกระบบที่มีสถานะยากจนจริงถูกทำนายว่าไม่ยากจน จำนวน 427 คน ในขณะที่แบบจำลองทำนายแรงงานนอกระบบที่จริงๆ แล้วไม่ได้มีสถานะยากจน ว่าเป็นแรงงานนอกระบบที่มีสถานะยากจนจำนวน 817 คน

4.3 ผลลัพธ์ของการพัฒนาแบบจำลองด้วยการเรียนรู้ของเครื่อง

ในการวิจัยครั้งนี้ ได้ประยุกต์ใช้แบบจำลองการเรียนรู้ของเครื่องจำนวน 3 แบบจำลอง ได้แก่ Logistic Regression, Random Forest และ XGBoost เพื่อวิเคราะห์และจำแนกสถานะความยากจนของแรงงานนอกระบบ โดยพิจารณาแนวทางที่เหมาะสมในการพัฒนาแบบจำลองภายใต้ข้อจำกัดของข้อมูลที่มีความไม่สมดุล

จากการตรวจสอบข้อมูลพบว่า กลุ่มแรงงานนอกระบบที่มีสถานะไม่ยากจนมีจำนวนมากกว่ากลุ่มที่มีสถานะยากจน ดังนั้น ผู้วิจัยจึงเลือกใช้วิธีการปรับสมดุลข้อมูล 2 วิธี ได้แก่ Random Undersampling และ SMOTE-ENN เพื่อแก้ไขปัญหาดังกล่าวในชุดข้อมูลฝึกหัด

ในการวิเคราะห์ข้อมูล ได้แบ่งข้อมูลทั้งหมดออกเป็น 2 ชุด โดยใช้สัดส่วน 80:20 ซึ่งได้ข้อมูลฝึกหัดจำนวน 34,129 คน และข้อมูลทดสอบจำนวน 8,533 คน โดยในชุดข้อมูลฝึกหัดมีแรงงานนอกระบบที่มีสถานะไม่ยากจนจำนวน 30,474 คน และแรงงานนอกระบบที่มีสถานะยากจนจำนวน 3,655 คน เมื่อปรับสมดุลข้อมูลฝึกด้วยวิธี Random Undersampling ทำให้จำนวนแรงงานนอกระบบที่มีสถานะไม่ยากจนลดลงเหลือ 3,655 คน ซึ่งเท่ากับจำนวนแรงงานนอกระบบที่มีสถานะยากจน ทำให้จำนวนข้อมูลที่ใช้ในการฝึกแบบจำลองมีทั้งสิ้น 7,310 คน ในขณะที่การปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ทำให้จำนวนแรงงานนอกระบบที่มีสถานะไม่ยากจนอยู่ที่ 20,776 คน และแรงงานนอกระบบที่มีสถานะยากจนเพิ่มขึ้นเป็น 29,150 คน รวมเป็นข้อมูลสำหรับฝึกจำนวนทั้งสิ้น 49,926 คน

หลังจากปรับสมดุลข้อมูลแล้ว ได้ดำเนินการจัดการข้อมูลตัวแปรอิสระเชิงคุณภาพตามรายละเอียดในหัวข้อ 3.4.2 และดำเนินการปรับขนาดข้อมูลตัวแปรอิสระเชิงปริมาณด้วยเทคนิค StandardScaler เพื่อให้ตัวแปรดังกล่าวมีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 จากนั้นได้ทำการปรับแต่งพารามิเตอร์ของแบบจำลองแต่ละตัว โดยใช้เทคนิค Cross Validation ร่วมกับ GridSearchCV เพื่อค้นหาพารามิเตอร์ที่เหมาะสมที่สุดสำหรับการเรียนรู้ของแต่ละแบบจำลองภายใต้โครงสร้างข้อมูลที่แตกต่างกันตามวิธีการปรับสมดุลที่เลือกใช้

4.3.1 แบบจำลอง Logistic Regression

1) พารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง Logistic Regression

ในการค้นหาพารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง Logistic Regression ได้กำหนดพารามิเตอร์ทั้งหมด 3 ตัว และค้นหาพารามิเตอร์ที่ดีที่สุดสำหรับการปรับสมดุลข้อมูลแต่ละวิธี ได้ผลลัพธ์ดังตารางที่ 4.7

ตารางที่ 4.7 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง Logistic Regression สำหรับแต่ละวิธีการปรับสมดุลข้อมูล

พารามิเตอร์	ค่าที่กำหนดให้พิจารณา	Random Undersampling	SMOTE-ENN
C	0.01, 0.1, 1, 10	10	10
penalty	l1, l2	l2	l1
solver	liblinear, bfgs, saga	liblinear	liblinear

2) ผลลัพธ์การพัฒนาแบบจำลอง Logistic Regression

หลังจากที่ได้พารามิเตอร์ที่ดีที่สุดสำหรับแต่ละวิธีการปรับสมดุลข้อมูล ได้นำชุดพารามิเตอร์ดังกล่าว มาพัฒนาแบบจำลอง Logistic Regression โดยใช้ข้อมูลฝึกหัดที่ผ่านการปรับสมดุลแล้ว เพื่อประเมินประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบ การพัฒนาแบบจำลองดำเนินการภายใต้ 2 เงื่อนไข ได้แก่ การใช้ Threshold คงที่ที่ค่า 0.5 ซึ่งเป็นค่ามาตรฐาน และการปรับ Threshold ให้เหมาะสมที่สุดจากชุดข้อมูลฝึกหัด โดยเลือก Threshold ที่ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงสุด เพื่อเพิ่มความสามารถของแบบจำลองในการจำแนกกลุ่มแรงงานที่มีสถานะยากจนได้แม่นยำยิ่งขึ้น

ตารางที่ 4.8 การจำแนกประเภทของแบบจำลอง Logistic Regression ชุดข้อมูลฝึกหัด

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	Random Undersampling	SMOTE-ENN	SMOTE-ENN
Threshold	**0.5**	**0.7131**	**0.5**	**0.8264**
Balanced Accuracy	0.7568	0.7213	**0.7590**	0.7315
Accuracy	0.7295	0.8313	0.6853	**0.8248**
F1-Score	0.3853	0.4247	0.3673	**0.4283**
Recall	0.7915	0.5814	**0.8528**	0.6129
Precision	0.2546	**0.3345**	0.2340	0.3291

จากตารางที่ 4.8 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง Logistic Regression ภายใต้การปรับสมดุลข้อมูลด้วยวิธี Random Undersampling และ SMOTE-ENN พบว่า ค่า Recall ของ SMOTE-ENN สูงกว่าวิธี Random Undersampling อย่างชัดเจน ซึ่งสะท้อนถึงความสามารถของแบบจำลองในการทำนายแรงงานนอกระบบที่มีสถานะยากจนได้มากกว่า นอกจากนี้ เมื่อพิจารณาเฉพาะภายในกลุ่มของ SMOTE-ENN พบว่า การปรับ Threshold ไปที่ค่า 0.8264 ซึ่งเป็นค่า Threshold ที่เหมาะสมที่สุด (Optimal Threshold) ส่งผลให้แบบจำลองมีค่าประสิทธิภาพโดยรวม (F1-Score) สูงที่สุด (0.4283) เมื่อเทียบกับกรณีใช้ Threshold คงที่ที่ 0.5 สะท้อนให้เห็นว่า การปรับ Threshold ร่วมกับเทคนิค SMOTE-ENN ช่วยเพิ่มประสิทธิภาพของแบบจำลองโดยรวมได้ดีที่สุด

จากผลการวิเคราะห์ดังกล่าวจึงสามารถสรุปได้ว่า แบบจำลอง Logistic Regression ร่วมกับการปรับสมดุลข้อมูลด้วย SMOTE-ENN เป็นวิธีการปรับสมดุลข้อมูลที่เหมาะสมสำหรับการจำแนกแรงงานนอกระบบที่มีสถานะยากจน และควรใช้ร่วมกับการปรับค่า Threshold เพื่อให้ได้ประสิทธิภาพสูงสุดในการจำแนกกลุ่มแรงงานนอกระบบที่มีสถานะยากจน

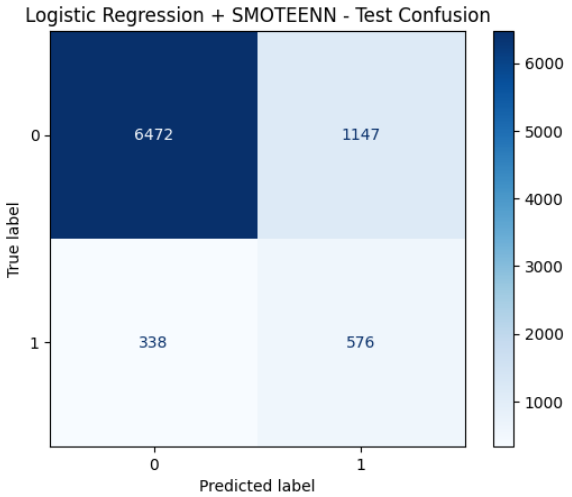
ดังนั้น แบบจำลอง Logistic Regression ที่ปรับสมดุลด้วยวิธี SMOTE-ENN และปรับ Threshold เท่ากับ 0.8264 เป็นแบบจำลองที่มีความแม่นยำ และมีประสิทธิภาพโดยรวมสมดุลที่สุด จากนั้นได้นำแบบจำลองดังกล่าวไปใช้กับชุดข้อมูลทดสอบ โดยผลลัพธ์แสดงรายละเอียดในตารางที่ 4.9

ตารางที่ 4.9 การจำแนกประเภทของแบบจำลอง Logistic Regression (Threshold = 0.8264)

เทคนิคการสุ่มตัวอย่าง	SMOTE-ENN
Balanced Accuracy	0.7398
Accuracy	0.8260
F1-Score	0.4369
Recall	0.6302
Precision	0.3343

จากตารางที่ 4.8 แสดงผลการประเมินประสิทธิภาพของแบบจำลอง Logistic Regression ซึ่งได้รับการปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN และกำหนดค่า Threshold เท่ากับ 0.8264 โดยทำการประเมินจากชุดข้อมูลทดสอบ ผลการประเมินพบว่า แบบจำลองมีค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.4369 ขณะเดียวกันค่าความระลึก (Recall) เท่ากับ 0.6302 แสดงถึงประสิทธิภาพของแบบจำลองในการทำนายแรงงานนอกระบบที่มีสถานะยากจนได้อย่างแม่นยำ ซึ่งสามารถทำนายได้ถูกต้องมากกว่าร้อยละ 63 ของจำนวนแรงงานนอกระบบที่มีสถานะยากจน เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลองระหว่างชุดข้อมูลฝึกหัดและชุดข้อมูลทดสอบ พบว่าเมทริกซ์ต่าง ๆ มีแนวโน้มใกล้เคียงกัน โดยเฉพาะค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) ซึ่งเป็นตัวชี้วัดสำคัญในบริบทของข้อมูลที่ไม่สมดุล แสดงให้เห็นว่าแบบจำลองมีเสถียรภาพในการนำไปใช้งานกับข้อมูลใหม่

ดังนั้น จึงสามารถสรุปได้ว่า แบบจำลอง Logistic Regression ที่ได้รับการปรับสมดุลด้วยเทคนิค SMOTE-ENN และกำหนดค่า Threshold ที่เหมาะสม สามารถนำไปประยุกต์ใช้กับข้อมูลใหม่ได้อย่างมีประสิทธิภาพ และตอบสนองต่อวัตถุประสงค์ของการจำแนกแรงงานนอกระบบที่มีสถานะยากจนได้อย่างเหมาะสม



รูปที่ 4.19 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง Logistic Regression (Machine Learning)

จากรูปที่ 4.19 แสดงเมทริกซ์วัดประสิทธิภาพของแบบจำลอง Logistic Regression ปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN และปรับ Threshold เท่ากับ 0.8264 แบบจำลองสามารถทำนาย

แรงงานนอกระบบที่มีสถานะยากจนได้ 576 คนจากทั้งหมด 914 คน ซึ่งถือว่าอยู่ในระดับที่น่าพอใจในบริบทที่ข้อมูลมีปัญหาไม่สมดุล ขณะเดียวกันมีแรงงานนอกระบบจำนวน 338 คนที่เป็นกลุ่มยากจนแต่ถูกทำนายผิดว่าไม่ยากจน และมีแรงงานจำนวน 1,147 คนที่ไม่ยากจนแต่ถูกทำนายผิดว่าเป็นกลุ่มยากจน โดยรวมแล้วแบบจำลองมีความสามารถในการจำแนกกลุ่มแรงงานได้อย่างมีประสิทธิภาพ และสะท้อนถึงศักยภาพของการใช้ SMOTE-ENN ในการจัดการกับปัญหาข้อมูลไม่สมดุล เพื่อเพิ่มความครอบคลุมในการทำนายกลุ่มแรงงานนอกระบบที่ยากจน

4.3.2 แบบจำลอง Random Forest

1) พารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง Random Forest

ในการค้นหาพารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง Random Forest ได้กำหนดพารามิเตอร์ทั้งหมด 5 ตัว และค้นหาพารามิเตอร์ที่ดีที่สุดสำหรับการปรับสมดุลข้อมูลแต่ละวิธี ได้ผลลัพธ์ดังตารางที่ 4.10

ตารางที่ 4.10 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง Random Forest สำหรับแต่ละวิธีการปรับสมดุลข้อมูล

พารามิเตอร์	ค่าที่กำหนดให้พิจารณา	Random Undersampling	SMOTE-ENN
n_estimators	100, 200, 300, 500	100	200
max_depth	5, 10, None	10	None
min_samples_split	2, 5	5	2
min_samples_leaf	1, 2	1	1
max_features	sqrt, log2	sqrt	sqrt

2) ผลลัพธ์การพัฒนาแบบจำลอง Random Forest

หลังจากที่ได้พารามิเตอร์ที่ดีที่สุดสำหรับแต่ละวิธีการปรับสมดุลข้อมูล ได้นำชุดพารามิเตอร์ดังกล่าว มาพัฒนาแบบจำลอง Random Forest โดยใช้ข้อมูลฝึกหัดที่ผ่านการปรับสมดุลแล้ว เพื่อประเมินประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบ

ตารางที่ 4.11 การจำแนกประเภทของแบบจำลอง Random Forest ชุดข้อมูลฝึกหัด

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	**0.7410**	0.7292
Accuracy	**0.8398**	0.8393
F1-Score	**0.4514**	0.4397
Recall	**0.6153**	0.5891
Precision	**0.3565**	0.3508

จากตารางที่ 4.11 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง Random Forest ระหว่างการใช้วิธี Random Undersampling และ SMOTE-ENN ในการปรับสมดุลข้อมูล พบว่าวิธี Random Undersampling มีประสิทธิภาพที่เหมาะสม สำหรับการระบุแรงงานนอกระบบที่มีสถานะยากจน โดยมีค่าความระลึก (Recall) สูงถึง 0.6153 (61.53%) เมื่อเทียบกับ SMOTE-ENN ที่มีค่า 0.5891 (58.91%) ซึ่งหมายความว่า Random Undersampling สามารถทำนายแรงงานยากจนได้มากกว่าประมาณ 2.6% นอกจากนี้ค่าประสิทธิภาพโดยรวม (F1-Score) ของวิธี Random Undersampling ยังสูงกว่าอย่างชัดเจน (0.4514 เทียบกับ 0.4397) ในขณะที่ความสามารถในการทำนายสถานะความยากจนของแรงงานนอกระบบในภาพรวมของวิธี Random Undersampling สูงกว่าวิธี SMOTE-ENN ในทุกเมทริกซ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ

ดังนั้น แบบจำลอง Random Forest ที่ปรับสมดุลด้วยวิธี Random Undersampling เป็นแบบจำลองที่มีความแม่นยำ และมีประสิทธิภาพโดยรวมสมดุลที่สุด จากนั้นได้นำแบบจำลองดังกล่าวไปใช้กับชุดข้อมูลทดสอบ โดยผลลัพธ์แสดงรายละเอียดในตารางที่ 4.10

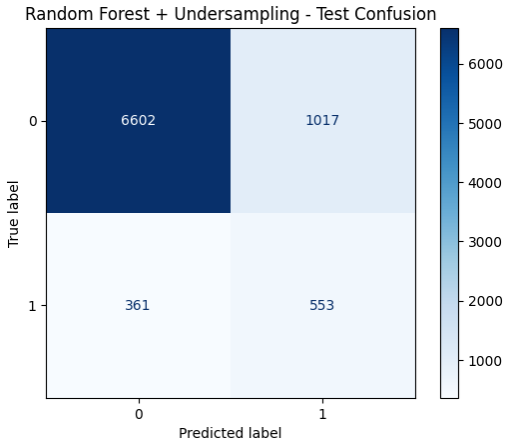
ตารางที่ 4.12 การจำแนกประเภทของแบบจำลอง Random Forest ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	**Random Undersampling**
Balanced Accuracy	0.7358
Accuracy	0.8385
F1-Score	0.4452
Recall	0.6050
Precision	0.3522

จากตารางที่ 4.12 แสดงผลการประเมินประสิทธิภาพของแบบจำลอง Random Forest เมื่อนำไปทดสอบกับชุดข้อมูลทดสอบ ผลการทดสอบแสดงให้เห็นว่าแบบจำลองที่ผ่านการ ปรับสมดุลข้อมูลด้วยวิธี Random Undersampling มีค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.4452 อย่างไรก็ตาม ในบริบทของการจำแนกแรงงานนอกระบบที่มีสถานะยากจน ค่าความระลึก (Recall) ถือเป็นตัวชี้วัดที่สำคัญที่สุด เนื่องจากแสดงถึงความสามารถของแบบจำลองในการทำนายกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้อย่างครอบคลุม ซึ่งมีความสำคัญอย่างยิ่งในการนำไปใช้ในเชิงนโยบายเพื่อหลีกเลี่ยงการมองข้ามแรงงานที่ควรได้รับการช่วยเหลือ โดยพบว่าแบบจำลองให้ค่าความระลึก (Recall) ที่ 0.6050 หรือ 60.50% หมายความว่าสามารถทำนายแรงงานนอกระบบที่เป็นคนจนได้ถูกต้องประมาณ 6 ใน 10 คน แม้ว่าจะยังมีโอกาสพลาดคนจนที่แท้จริงอีกประมาณ 40% แต่ผลลัพธ์นี้แสดงให้เห็นถึงศักยภาพของแบบจำลองในการสนับสนุนการตัดสินใจเชิงนโยบายสำหรับการช่วยเหลือแรงงานนอกระบบที่อยู่ในสภาวะยากจน

จากการเปรียบเทียบผลการจำแนกประเภทของแบบจำลอง Random Forest ปรับสมดุลข้อมูลด้วยวิธี Random Undersampling ทั้งในชุดข้อมูลฝึกหัด (ตารางที่ 4.9) และชุดข้อมูลทดสอบ (ตารางที่ 4.10) ค่าประสิทธิภาพของแบบจำลองทั้งสองชุดข้อมูลมีความใกล้เคียงกันในทุกตัวชี้วัดหลัก ได้แก่ ค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) สะท้อนว่าแบบจำลองมีประสิทธิภาพที่สม่ำเสมอระหว่างชุดข้อมูลฝึกหัดและชุดทดสอบไม่เกิดปัญหา Overfitting

โดยสรุป แบบจำลอง Random Forest พบว่าวิธี Random Undersampling เป็นวิธีการจัดการข้อมูลไม่สมดุลที่เหมาะสมที่สุด โดยผลการทำนายการจำแนกสถานะความยากจนของแรงงานนอกระบบของแบบจำลองดังกล่าวให้ผลลัพธ์ ดังนี้



รูปที่ 4.20 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง Random Forest

จากรูปที่ 4.20 แสดงเมทริกซ์วัดประสิทธิภาพของแบบจำลอง Random Forest ที่ปรับสมดุลข้อมูลด้วยวิธี Random Undersampling พบว่าแบบจำลองสามารถทำนายแรงงานนอกระบบที่มีสถานะยากจนได้อย่างถูกต้อง 553 คนจากทั้งหมด 914 คน ในขณะที่มีทำนายผิดโดยระบุว่าเป็นแรงงานนอกระบบที่มีสถานะไม่ยากจน 361 คน นอกจากนี้ยังพบว่าแรงงานที่ไม่ยากจนจำนวน 1,017 คน ถูกจำแนกผิดว่าเป็นแรงงานยากจน ส่วนการจำแนกกลุ่มไม่ยากจนที่ถูกต้องมีจำนวน 6,602 คน ผลลัพธ์ดังกล่าวแสดงให้เห็นว่าแบบจำลองสามารถจำแนกแรงงานทั้งสองกลุ่มได้ในระดับที่เหมาะสม

4.3.3. แบบจำลอง XGBoost

1) พารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง XGBoost

ในการค้นหาพารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง XGBoost ได้กำหนดพารามิเตอร์ทั้งหมด 5 ตัว และค้นหาพารามิเตอร์ที่ดีที่สุดสำหรับการปรับสมดุลข้อมูลแต่ละวิธี ได้ผลลัพธ์ดังตารางที่ 4.13

ตารางที่ 4.13 การตั้งค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง XGBoost สำหรับแต่ละวิธีการปรับสมดุลข้อมูล

พารามิเตอร์	ค่าที่กำหนดให้พิจารณา	Random Undersampling	SMOTE-ENN
n_estimators	100, 200, 300, 500	100	200
max_depth	3, 6, None	6	6
learning_rate	0.01, 0.1	0.1	0.1
subsample	0.8, 1.0	0.8	0.8
colsample_bytree	0.8, 1.0	0.8	0.8

2) ผลลัพธ์การพัฒนาแบบจำลอง XGBoost

หลังจากที่ได้พารามิเตอร์ที่ดีที่สุดสำหรับแต่ละวิธีการปรับสมดุลข้อมูล ได้นำชุดพารามิเตอร์ที่ดังกล่าว มาพัฒนาแบบจำลง XGBoost โดยใช้ข้อมูลฝึกหัดที่ผ่านการปรับสมดุลแล้ว เพื่อประเมินประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบ

ตารางที่ 4.14 การจำแนกประเภทของแบบจำลอง XGBoost ชุดข้อมูลฝึกหัด

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	0.8187	**0.8224**
Accuracy	0.8990	**0.9002**
F1-Score	0.6031	**0.6082**
Recall	0.7166	**0.7234**
Precision	0.5207	**0.5247**

จากตารางที่ 4.14 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง XGBoost ระหว่างการใช้วิธี Random Undersampling และ SMOTE-ENN ในการปรับสมดุลข้อมูล พบว่าวิธี SMOTE-ENN มีประสิทธิภาพที่เหมาะสมกว่าสำหรับการทำนายแรงงานนอกระบบที่มีสถานะยากจน โดยมีค่าความระลึก (Recall) สูงถึง 0.7234 (72.34%) เมื่อเทียบกับ Random Undersampling ที่มีค่า 0.7166 (71.66%) ซึ่งหมายความว่า SMOTE-ENN สามารถทำนายแรงงานยากจนได้ถูกต้องมากกว่าประมาณ 0.7% นอกจากนี้ค่าประสิทธิภาพโดยรวม (F1-Score) ของวิธี SMOTE-ENN ยังสูงกว่า (0.6082 เทียบกับ 0.6031) รวมถึงค่าความแม่นยำ (Precision) ที่ดีกว่า (0.5247 เทียบกับ 0.5207) และค่าความถูกต้องสมดุล (Balanced Accuracy) ที่สูงกว่า (0.8224 เทียบกับ 0.8187) ผลการเปรียบเทียบดังกล่าวแสดงให้เห็นว่า SMOTE-ENN มีประสิทธิภาพที่เหนือกว่าในทุกเมทริกซ์ ซึ่งมีความสำคัญสูงในบริบทการวิจัยนี้ เนื่องจากการพลาดการทำนายแรงงานนอกระบบที่มีสถานะยากจนอาจส่งผลให้นโยบายสวัสดิการไม่สามารถเข้าถึงกลุ่มแรงงานนอกระบบที่ต้องการความช่วยเหลืออย่างแท้จริง

ดังนั้น SMOTE-ENN จึงเป็นเทคนิคที่เหมาะสมที่สุดสำหรับการพัฒนาแบบจำลอง XGBoost ในการจำแนกกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้อย่างมีประสิทธิภาพ จากนั้นได้นำแบบจำลองดังกล่าวไปใช้กับชุดข้อมูลทดสอบ โดยผลลัพธ์แสดงรายละเอียดในตารางที่ 4.15

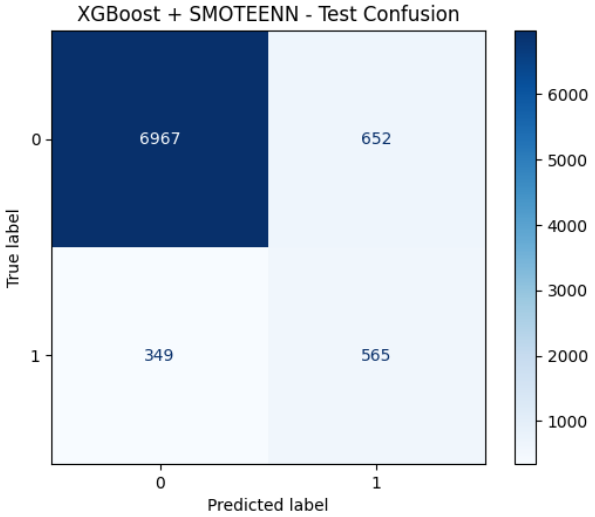
ตารางที่ 4.15 การจำแนกประเภทของแบบจำลอง XGBoost ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	SMOTE-ENN
Balanced Accuracy	0.7663
Accuracy	0.8827
F1-Score	0.5303
Recall	0.6182
Precision	0.4643

จากตารางที่ 4.15 แสดงผลการประเมินประสิทธิภาพของแบบจำลอง XGBoost ที่ใช้วิธี SMOTE-ENN ในการจัดการข้อมูลไม่สมดุลบนชุดข้อมูลทดสอบ ผลการทดสอบแสดงให้เห็นว่าแบบจำลองมีประสิทธิภาพที่ดีในหลายมิติ โดยมีค่าประสิทธิภาพโดยรวม (F1-Score) ที่ 0.5303 แสดงถึงความสมดุลที่ดีระหว่างความแม่นยำและความครอบคลุมในการทำนาย สำหรับค่าความระลึก (Recall) ที่ 0.6182 หรือ 61.82% หมายความว่าแบบจำลองสามารถทำนายแรงงานนอกระบบที่เป็นคนจนได้ประมาณ 6 ใน 10 คน ซึ่งแสดงถึงความสามารถในการระบุแรงงานนอกระบบที่ต้องการความช่วยเหลือได้อย่างครอบคลุมพอสมควร ผลลัพธ์นี้ชี้ให้เห็นว่าแบบจำลอง XGBoost ด้วยวิธี SMOTE-ENN มีศักยภาพในการสนับสนุนการตัดสินใจเชิงนโยบายสำหรับการช่วยเหลือแรงงานนอกระบบที่อยู่ในสภาวะยากจน

จากการเปรียบเทียบผลการจำแนกประเภทของแบบจำลอง XGBoost ปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ทั้งในชุดข้อมูลฝึกหัด (ตารางที่ 4.14) และชุดข้อมูลทดสอบ (ตารางที่ 4.15) ค่าประสิทธิภาพของแบบจำลองทั้งสองชุดข้อมูลมีความใกล้เคียงกันในทุกตัวชี้วัดหลักสะท้อนว่าแบบจำลองมีประสิทธิภาพที่สม่ำเสมอระหว่างชุดข้อมูลฝึกหัดและชุดทดสอบไม่เกิดปัญหา Overfitting

โดยสรุป วิธี SMOTE-ENN เป็นวิธีที่เหมาะสมที่สุดสำหรับการพัฒนาแบบจำลอง XGBoost ในการจำแนกกลุ่มแรงงานนอกระบบที่มีสถานะยากจนได้อย่างมีประสิทธิภาพ โดยผลการทำนายการจำแนกสถานะความยากจนของแรงงานนอกระบบของแบบจำลองดังกล่าวให้ผลลัพธ์ ดังนี้



รูปที่ 4.21 เมทริกซ์วัดประสิทธิภาพของแบบจำลอง XGBoost

จากรูปที่ 21 แสดงเมทริกซ์วัดประสิทธิภาพของแบบจำลอง XGBoost ที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ในชุดข้อมูลทดสอบ พบว่าแบบจำลองสามารถทำนายแรงงานนอกระบบที่มีสถานะยากจนได้อย่างถูกต้อง 565 คนจากทั้งหมด 914 คน ในขณะที่มีการจำแนกผิดพลาดโดยทำนายว่าไม่ยากจน 349 คน นอกจากนี้ยังพบว่าแรงงานที่ไม่ยากจนจำนวน 652 คน ถูกจำแนกผิดพลาดว่าเป็นแรงงานยากจน ผลลัพธ์ดังกล่าวแสดงให้เห็นว่าแบบจำลอง XGBoost ที่ใช้วิธี SMOTE-ENN มีความสามารถในการทำนายกลุ่มแรงงานยากจนได้ในระดับที่ดี โดยสามารถทำนายกลุ่มเป้าหมายได้มากกว่าครึ่งหนึ่ง (565 จาก 914 คน หรือ 61.82%) พร้อมทั้งมีความแม่นยำสูงในการ

จำแนกกลุ่มที่ไม่ยากจน ซึ่งสะท้อนความสมดุลในประสิทธิภาพของแบบจำลองที่เหมาะสมสำหรับการนำไปใช้ในทางปฏิบัติ

4.4 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุลทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test

ในการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้วิธีการจัดการข้อมูลไม่สมดุลระหว่าง Random Undersampling และ SMOTE-ENN การวิจัยครั้งนี้ได้ใช้เทคนิค K-Fold Cross-Validation (K = 5) เพื่อแบ่งชุดข้อมูลฝึกอบรมออกเป็น 5 ส่วนอย่างเท่าเทียมกัน โดยสลับใช้แต่ละส่วนเป็นชุดทดสอบภายในแต่ละรอบของการฝึกแบบจำลอง ซึ่งการปรับสมดุลข้อมูลทำให้ข้อมูลที่แบบจำลองแต่ละตัวเรียนรู้นั้นแตกต่างกัน ดังนั้นการเปรียบเทียบจึงเป็นแบบ Independent t-test

ในการประเมินประสิทธภาพของแบบจำลองจะเลือกใช้ค่าประสิทธิภาพโดยรวม (F1-Score) เป็นตัวชี้วัดหลักในการประเมินแบบจำลอง เนื่องจากสามารถสะท้อนสมดุลระหว่างค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้อย่างมีประสิทธิภาพ โดยเฉพาะในบริบทที่ข้อมูลมีความไม่สมดุลระหว่างกลุ่มเป้าหมาย ทั้งนี้ เพื่อทดสอบว่า วิธีการจัดการข้อมูลไม่สมดุลทั้งสองแบบให้ผลที่แตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ ได้ตั้งสมมติฐานไว้ดังนี้

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

โดยที่ μ_1 แทน ค่าเฉลี่ยของค่าประสิทธิภาพโดยรวม (F1-Score) ของการจัดการปัญหาข้อมูลไม่สมดุลด้วยวิธี Random Undersampling

μ_2 แทน ค่าเฉลี่ยของค่าประสิทธิภาพโดยรวม (F1-Score) ของการจัดการปัญหาข้อมูลไม่สมดุลด้วยวิธี SMOTE-ENN

ตารางที่ 4.16 ทดสอบสมมติฐานการเปรียบเทียบค่าเฉลี่ยของ F1-Score ในแต่ละแบบจำลองโดยวิธีการจัดการข้อมูลไม่สมดุลทั้ง 2 วิธี

แบบจำลอง	การปรับสมดุลข้อมูล	Shapiro-Wilk p-value	Levene's p-value	t-test (p-value)
Logistic Regression (Threshold = 0.5)	Random Undersampling	0.7425	0.9688	**5.4283 (0.0006)**
	SMOTE-ENN	0.0860		
Logistic Regression (Optimal Threshold)	Random Undersampling	0.3389	0.5918	0.0183 (0.9859)
	SMOTE-ENN	0.2125		
Random Forest	Random Undersampling	0.1120	0.5918	1.6433 (0.1390)
	SMOTE-ENN	0.5349		
XGBoost	Random Undersampling	0.7572	0.0841	**-12.0096 (0.0000)**
	SMOTE-ENN	0.8713		

ตารางที่ 4.17 ค่าเฉลี่ย F1-Score จากการทดสอบสมมติฐาน

แบบจำลอง	การปรับสมดุลข้อมูล	ค่าเฉลี่ย F1-Score
Logistic Regression (Threshold =0.5)	Random Undersampling	**0.3834**
	SMOTE-ENN	0.3658
Logistic Regression (Optimal Threshold)	Random Undersampling	0.4277
	SMOTE-ENN	**0.4278**
Random Forest	Random Undersampling	**0.3613**
	SMOTE-ENN	0.3560
XGBoost	Random Undersampling	0.4451
	SMOTE-ENN	**0.5170**

ตารางที่ 4.16 แสดงผลการทดสอบสมมติฐานการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้วิธีการจัดการข้อมูลไม่สมดุลทั้งสองวิธี โดยผลการทดสอบ Shapiro-Wilk และ Levene's Test พบว่า p-value มากกว่า 0.05 ในทุกแบบจำลอง แสดงว่าค่าประสิทธิภาพโดยรวม (F1-Score) มีการกระจายแบบปกติ และมีความแปรปรวนไม่แตกต่างกัน อย่างมีนัยสำคัญทางสถิติ จึงสามารถใช้สถิติ t-test ในการเปรียบเทียบได้อย่างเหมาะสม

นอกจากนี้ยังพบว่าแบบจำลอง Logistic Regression (Threshold = 0.5) ผลการทดสอบแสดงค่า t = 5.4283 และ p-value = 0.0006 เนื่องจาก p-value < 0.05 จึงปฏิเสธสมมติฐานหลัก แสดงว่ามีความแตกต่างอย่างมีนัยสำคัญทางสถิติ จากตารางที่ 4.17 แสดงให้เห็นว่าค่าเฉลี่ยของค่าประสิทธิภาพโดยรวม (F1-Score) ของวิธีการปรับสมดุล Random Undersampling มีประสิทธิภาพสูงกว่า SMOTE-ENN สำหรับ Logistic Regression ที่ใช้ Threshold เท่ากับ 0.5

แบบจำลอง Logistic Regression (Optimal Threshold = 0.8264) ค่า t = 0.0183 และ p-value = 0.9859 (> 0.05) ดังนั้น ไม่สามารถปฏิเสธสมมติฐานหลักได้ หมายความว่าไม่มีความแตกต่างของประสิทธิภาพค่าประสิทธิภาพโดยรวม (F1-Score) ระหว่าง Random Undersampling และ SMOTE-ENN อย่างมีนัยสำคัญทางสถิติ ทั้งสองวิธีให้ผลลัพธ์ที่ไม่แตกต่างกันสำหรับ Random Forest

แบบจำลอง Random Forest สถิติทดสอบ t = 1.6433 และ p-value = 0.1390 (> 0.05) ดังนั้น ไม่สามารถปฏิเสธสมมติฐานหลักได้ หมายความว่าไม่มีความแตกต่างของประสิทธิภาพค่าประสิทธิภาพโดยรวม (F1-Score) ระหว่าง Random Undersampling และ SMOTE-ENN อย่างมีนัยสำคัญทางสถิติ ทั้งสองวิธีให้ผลลัพธ์ที่ไม่แตกต่างกันสำหรับ Random Forest

แบบจำลอง XGBoost สถิติทดสอบ t = -12.0096 และ p-value = 0.0000 (< 0.05) ผลการทดสอบแสดงว่าปฏิเสธสมมติฐานหลัก จากตารางที่ 4.17 แสดงให้เห็นว่าค่าเฉลี่ยของค่าประสิทธิภาพโดยรวม (F1-Score) ของวิธีการปรับสมดุล SMOTE-ENN มีประสิทธิภาพสูงกว่า Random Undersampling สำหรับแบบจำลอง XGBoost อย่างมีนัยสำคัญทางสถิติ

ผลการวิเคราะห์ทางสถิติในการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้วิธีการจัดการข้อมูลไม่สมดุลแตกต่างกัน แสดงให้เห็นถึงความซับซ้อนของความสัมพันธ์ระหว่างวิธีการจัดการข้อมูลไม่สมดุลและประเภทของแบบจำลอง การศึกษานี้ชี้ให้เห็นว่าไม่มีวิธีการปรับสมดุลข้อมูลใดที่

เหนือกว่าในทุกสถานการณ์ แต่ประสิทธิภาพจะขึ้นอยู่กับลักษณะเฉพาะของแบบจำลองและการปรับแต่งพารามิเตอร์ที่เหมาะสม ซึ่งผลจากการทดสอบสมมติฐานสอดคล้องกับผลลัพธ์ที่ได้พัฒนาแบบจำลองในหัวข้อ 4.3

4.5 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบของ ทั้ง 4 แบบจำลอง

4.5.1 เปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจนของแรงงานนอกระบบของ ทั้ง 4 แบบจำลองด้วยค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall)

จากการวิเคราะห์ประสิทธิภาพของแบบจำลองทั้ง 4 แบบ ได้แก่ Logistic Regression ด้วยวิธีทางสถิติ, Logistic Regression ด้วยการเรียนรู้ของเครื่อง, Random Forest และ XGBoost ที่นำมาใช้ในการจำแนกสถานะความยากจนของแรงงานนอกระบบ พบว่าแบบจำลองแต่ละประเภทได้รับการปรับให้เหมาะสมกับลักษณะข้อมูลและปัญหาความไม่สมดุลของข้อมูล สำหรับแบบจำลอง Logistic Regression ทั้งในรูปแบบวิธีทางสถิติและการเรียนรู้ของเครื่อง ได้มีการปรับค่า Threshold เพื่อเพิ่มค่าประสิทธิภาพโดยรวม (F1-Score) ให้มีประสิทธิภาพสูงขึ้น ขณะเดียวกันแบบจำลองที่พัฒนาด้วยการเรียนรู้ของเครื่องทั้งหมดได้รับการปรับสมดุลข้อมูลด้วยวิธีการจัดการข้อมูลไม่สมดุล ได้แก่ Random Undersampling และ SMOTE-ENN โดยเลือกใช้วิธีที่ให้ผลลัพธ์ดีที่สุดในแต่ละกรณี ผลการวิเคราะห์พบว่า

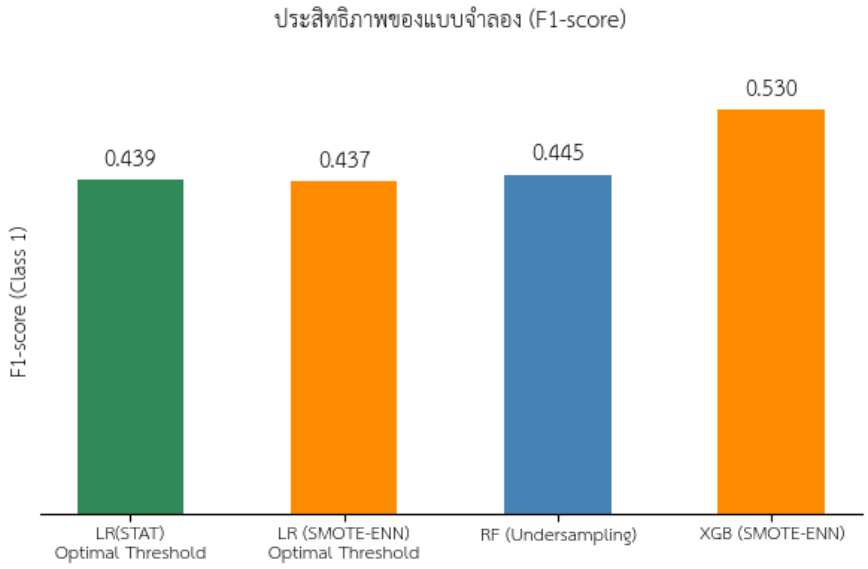
1) แบบจำลอง Logistic Regression ด้วยวิธีทางสถิติ ให้ผลลัพธ์ดีที่สุดเมื่อปรับ Threshold เป็น 0.2341
2) แบบจำลอง Logistic Regression ด้วยการเรียนรู้ของเครื่อง ให้ค่า ค่าประสิทธิภาพโดยรวม (F1-Score) สูงสุดเมื่อปรับ Threshold เป็น 0.8264 และปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN
3) แบบจำลอง Random Forest ให้ผลลัพธ์ดีที่สุดเมื่อใช้วิธี Random Undersampling
4) แบบจำลอง XGBoost ให้ประสิทธิภาพสูงสุดเมื่อใช้วิธี SMOTE-ENN

ดังนั้น ในการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมดในขั้นสุดท้าย จะเลือกใช้แบบจำลองที่ได้รับการปรับสมดุลข้อมูล และThreshold อย่างเหมาะสมที่สุดสำหรับแต่ละแบบจำลอง โดยพิจารณาค่าประสิทธิภาพโดยรวม (F1-Score) และค่าความระลึก (Recall) เป็นเกณฑ์หลักในการประเมิน

ตารางที่ 4.18 การจำแนกประเภทของแบบจำลองทั้ง 4 แบบจำลองบนชุดข้อมูลทดสอบ

แบบจำลอง	LR (STAT) Threshold = 0.2341	LR (SMOTE-ENN) Threshold = 0.8264	RF (Undersampling)	XGB (SMOTE-ENN)
Balanced Accuracy	0.7128	0.7398	0.7358	**0.7663**
Accuracy	0.8542	0.8260	0.8385	**0.8827**
F1-Score	0.4391	0.4369	0.4452	**0.5303**
Recall	0.5328	**0.6302**	0.6050	0.6182
Precision	0.3735	0.3343	0.3522	**0.4643**

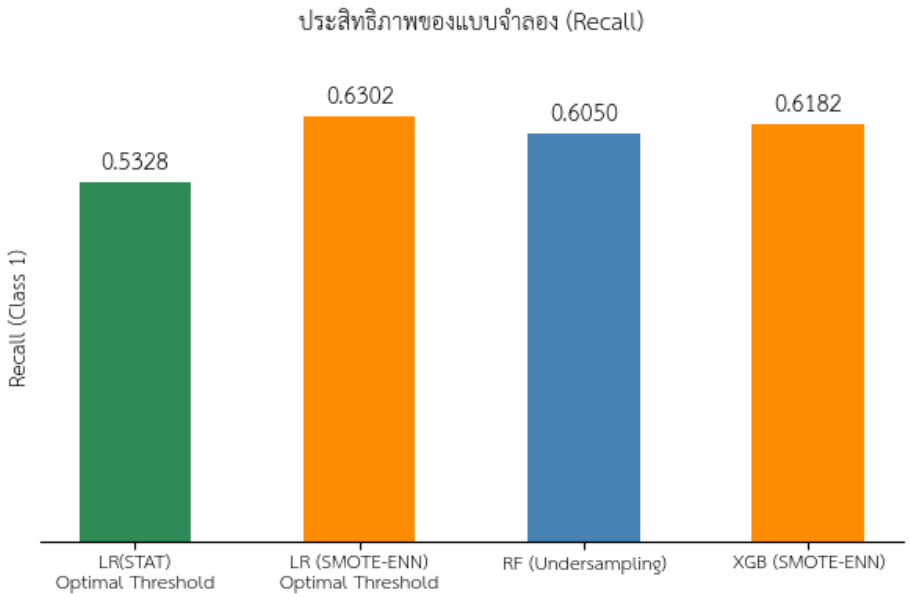
จากตารางที่ 4.18 แสดงให้เห็นว่าแบบจำลอง XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN มีประสิทธิภาพสูงสุดในการจำแนกสถานะความยากจนของแรงงานนอกระบบ โดยมีค่าประสิทธิภาพโดยรวม (F1-Score) สูงสุดที่ 0.5303 และค่าความถูกต้องสมดุล (Balanced Accuracy) สูงสุดที่ 0.7663 แม้ว่าจะมีค่าความระลึก (Recall) ต่ำกว่าแบบจำลอง Logistic Regression ร่วมกับวิธี SMOTE-ENN เล็กน้อย (0.6182 เทียบกับ 0.6302) แต่โดยรวมแบบจำลอง XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ให้ประสิทธิภาพโดยรวมที่สมดุลที่สุด นอกจากนี้จะสังเกตเห็นว่าแบบจำลอง Logistic Regression ทั้งสองวิธี (สถิติและการเรียนรู้ของเครื่อง) มีประสิทธิภาพใกล้เคียงกัน โดยแบบจำลอง Logistic Regression ด้วย SMOTE-ENN มีความสามารถในการทำนายกลุ่มแรงงานยากจนได้ถูกต้องมากกว่าแบบจำลอง Logistic Regression ด้วยวิธีทางสถิติประมาณ 10% (Recall เพิ่มจาก 0.5328 เป็น 0.6302)



รูปที่ 4.22 ค่า ค่าประสิทธิภาพโดยรวม (F1-Score) เปรียบเทียบประสิทธิภาพของแบบจำลอง

จากรูปที่ 4.22 แสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 แบบจำลอง ด้วยค่าประสิทธิภาพโดยรวม (F1-Score) ในการจำแนกสถานะความยากจนของแรงงานนอกระบบ พบว่า

แบบจำลอง XGBoost ที่ใช้วิธี SMOTE-ENN มีประสิทธิภาพสูงสุดด้วยค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.530 รองลงมาคือแบบจำลอง Random Forest ที่ใช้วิธี Random Undersampling ด้วยค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.445 ตามด้วยแบบจำลอง Logistic Regression ด้วยวิธีทางสถิติที่มีค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.439 และสุดท้ายคือแบบจำลอง Logistic Regression ด้วยการเรียนรู้ของเครื่องที่ใช้วิธี SMOTE-ENN ด้วยค่าประสิทธิภาพโดยรวม (F1-Score) เท่ากับ 0.437 ผลการเปรียบเทียบดังกล่าวแสดงให้เห็นว่าแบบจำลอง XGBoost มีความสามารถในการสร้างสมดุลระหว่างค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้ดีที่สุด โดยมีประสิทธิภาพสูงกว่าแบบจำลองอื่นๆ อย่างชัดเจน ซึ่งแสดงถึงความเหมาะสมในการนำไปใช้จำแนกกลุ่มแรงงานนอกระบบที่มีสถานะยากจน เพื่อการกำหนดนโยบายสวัสดิการที่มีประสิทธิภาพ ในขณะที่ความแตกต่างของประสิทธิภาพระหว่างแบบจำลอง Random Forest, Logistic Regression ด้วยวิธีทางสถิติ และ Logistic Regression ด้วยการเรียนรู้ของเครื่อง มีความใกล้เคียงกัน



รูปที่ 4.23 ค่าความระลึก (Recall) เปรียบเทียบประสิทธิภาพของแบบจำลอง

จากรูปที่ 4.23 แสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆ ในด้านค่าความระลึก (Recall) พบว่าทุกแบบจำลองมีค่าความระลึกที่ใกล้เคียงกัน โดยอยู่ในช่วง 0.5328-0.6302 ซึ่งแสดงให้เห็นว่าแบบจำลองแต่ละชนิดมีความสามารถในการระบุกลุ่มแรงงานนอกระบบที่ยากจนได้ในระดับที่ใกล้เคียงกันแบบจำลอง Logistic Regression (SMOTE-ENN) ที่ปรับ Optimal Threshold ให้ประสิทธิภาพสูงสุดด้วยค่าความระลึก 0.6302 ตามด้วย XGBoost (SMOTE-ENN) ที่ 0.6182 และ Random Forest (Undersampling) ที่ 0.6050 ส่วนแบบจำลอง Logistic Regression ด้วยวิธีทางสถิติ และปรับ Optimal Threshold ให้ค่าต่ำสุดที่ 0.5328 ความใกล้เคียงของค่าความระลึกนี้สะท้อนให้เห็นว่า แม้จะใช้เทคนิคการเรียนรู้ของเครื่องและวิธีการปรับสมดุลข้อมูลที่แตกต่างกัน แต่ทุกแบบจำลองมีศักยภาพในการค้นหาและระบุแรงงานนอกระบบที่มีสถานะยากจนได้ในระดับที่ไม่แตกต่างกันมาก

โดยสรุป แบบจำลองที่มีประสิทธิภาพสูงสุดในการทำนายสถานะความยากจนของแรงงานนอกระบบ คือ แบบจำลอง XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ซึ่งให้ผลลัพธ์ที่แม่นยำและสมดุลที่สุด

4.5.2 การทดสอบ McNemar's Test สำหรับการเปรียบเทียบประสิทธิภาพแบบจำลอง

เนื่องจากแบบจำลอง XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วย SMOTE-ENN เป็นแบบจำลองที่ดีที่สุด การทดสอบ McNemar's Test ถูกนำมาใช้ เพื่อประเมินความแตกต่างทางสถิติของประสิทธิภาพการทำนายระหว่างแบบจำลอง XGBoost กับแบบจำลองอีก 3 แบบจำลองที่มีประสิทธิภาพสูงสุดในแต่ละแบบจำลอง โดยทดสอบสมมติฐานดังนี้

H_0: แบบจำลอง XGBoost (SMOTE-ENN) มีประสิทธิภาพไม่แตกต่างจากแบบจำลองอื่นๆ

H_1: แบบจำลอง XGBoost (SMOTE-ENN) มีประสิทธิภาพแตกต่างจากแบบจำลองอื่นๆ

ผลการทดสอบแสดงในตารางที่ 4.19

ตารางที่ 4.19 ผลการทดสอบ McNemar's Test เพื่อเปรียบเทียบประสิทธิภาพแบบจำลอง

การเปรียบเทียบแบบจำลอง A vs B	แบบจำลอง A ถูก-B ผิด	แบบจำลอง A ผิด-B ถูก	p-value
XGBoost vs Logistic Regression (Stat)	603	155	0.000
XGBoost vs Logistic Regression (ML)	458	125	0.000
XGBoost vs Random Forest	450	138	0.000

หมายเหตุ: ** มีนัยสำคัญทางสถิติที่ระดับ $\alpha = 0.05$

จากตารางที่ 4.19 พบว่า การทดสอบ McNemar's Test แสดงผลลัพธ์ที่มีนัยสำคัญทางสถิติ (p-value < 0.05) สำหรับการเปรียบเทียบทั้ง 3 คู่ ซึ่งหมายความว่า สามารถปฏิเสธสมมติฐานหลักและยอมรับสมมติฐานทางเลือกได้ที่ระดับความเชื่อมั่น 95% ผลการทดสอบนี้ยืนยันว่า แบบจำลอง XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วย SMOTE-ENN มีประสิทธิภาพที่แตกต่างอย่างมีนัยสำคัญทางสถิติจากแบบจำลองอื่นๆ ทั้งหมด

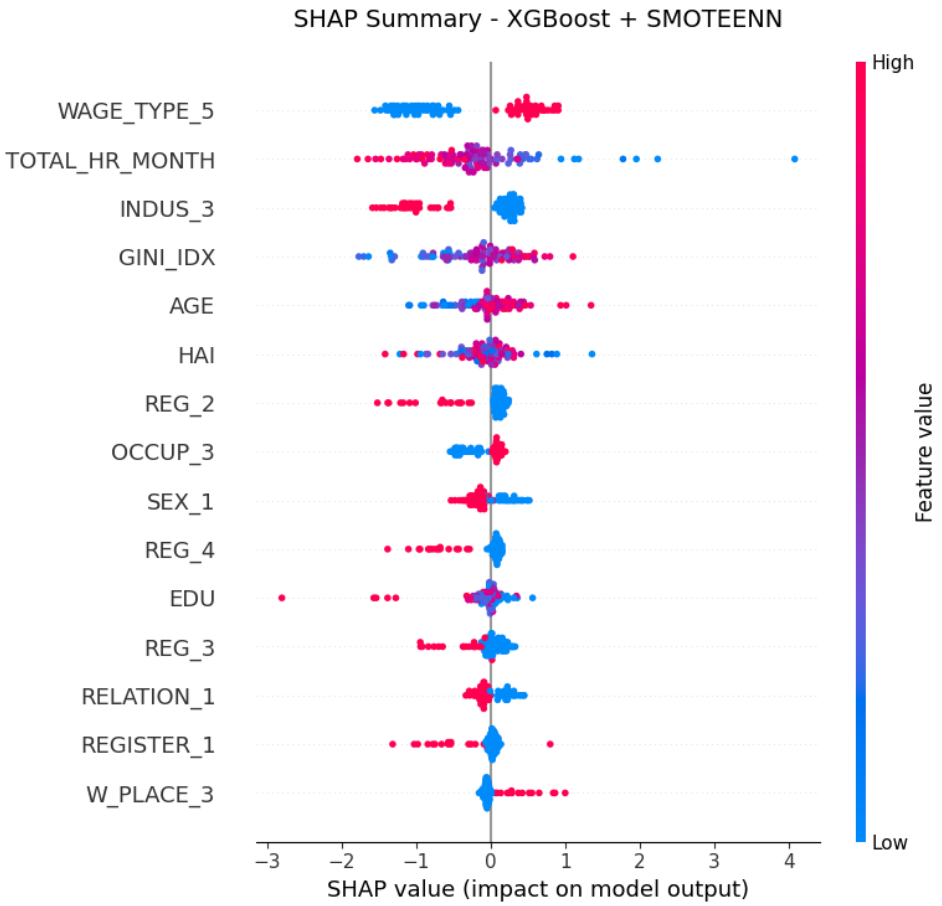
การทดสอบนี้จึงสนับสนุนการเลือกใช้แบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE-ENN เป็นแบบจำลองที่เหมาะสมที่สุดสำหรับการทำนายสถานะความยากจนของแรงงานนอกระบบ เนื่องจากมีประสิทธิภาพที่เหนือกว่าแบบจำลองอื่นๆ อย่างชัดเจนและมีนัยสำคัญทางสถิติ

4.6 การวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP

การศึกษาปัจจัยที่มีอิทธิพลต่อการจำแนกสถานะความยากจนของแรงงานนอกระบบสามารถดำเนินการได้ด้วยเทคนิค SHAP (SHapley Additive exPlanations) ซึ่งเป็นเครื่องมือที่พัฒนาจากแนวคิดทฤษฎีเกมของ Shapley value ที่ช่วยให้สามารถอธิบายผลกระทบของตัวแปรแต่ละตัวต่อผลการทำนายของแบบจำลองได้อย่างชัดเจนและครอบคลุม เนื่องจาก SHAP มีความสามารถในการวิเคราะห์ได้ทั้งในระดับรายบุคคล (Local explanation) เพื่อเข้าใจปัจจัยที่ส่งผลต่อการทำนายของแต่ละกรณี และในระดับภาพรวม (Global explanation) เพื่อทำความเข้าใจรูปแบบทั่วไปของอิทธิพลของตัวแปรต่างๆ ต่อแบบจำลอง ข้อได้เปรียบสำคัญของ SHAP เมื่อเปรียบเทียบกับวิธีการ

แบบดั้งเดิมคือความสามารถในการตีความได้ง่าย ความโปร่งใสในการอธิบาย และความสามารถในการประยุกต์ใช้กับแบบจำลองทุกประเภท

ในการศึกษาครั้งนี้ ได้นำเทคนิค SHAP มาประยุกต์ใช้ในการวิเคราะห์แบบจำลอง XGBoost ที่ได้ผ่านการปรับสมดุลข้อมูลด้วยเทคนิค SMOTE-ENN เหตุผลในการเลือกใช้แบบจำลองดังกล่าว เนื่องจากมีประสิทธิภาพสูงสุดในการทำนายสถานะความยากจนของแรงงานนอกระบบเมื่อเปรียบเทียบกับแบบจำลองอื่นๆ ที่ทำการทดสอบ การวิเคราะห์ด้วย SHAP จะช่วยให้เห็นภาพรวมของปัจจัยสำคัญที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ ซึ่งจะเป็นข้อมูลสำคัญในการเข้าใจกลไกที่ทำให้เกิดความยากจนในกลุ่มแรงงานนอกระบบดังกล่าว



รูปที่ 4.24 ผลการวิเคราะห์ความสำคัญของตัวแปรด้วยค่า SHAP

จากรูปที่ 4.24 แสดงผลการวิเคราะห์ความสำคัญของตัวแปรด้วยค่า SHAP จากแบบจำลอง XGBoost ที่ผ่านการปรับสมดุลข้อมูลด้วยเทคนิค SMOTE-ENN โดยแกนตั้งแสดงตัวแปรอิสระเรียงตามลำดับความสำคัญจากมากไปน้อย แกนนอนแสดงค่า SHAP ที่บ่งบอกผลกระทบต่อการทำนาย ซึ่งค่าบวกหมายถึงการมีแนวโน้มเพิ่มโอกาสการเป็นแรงงานระบบที่มีสถานะยากจน และค่าลบหมายถึงการลดโอกาสการเป็นแรงงานที่มีสถานะยากจน สีของจุดแสดงระดับค่าของตัวแปร โดยสีแดงแทนค่าสูงและสีน้ำเงินแทนค่าต่ำ

จากการวิเคราะห์ SHAP Summary Plot พบว่าตัวแปรต่างๆ มีอิทธิพลต่อการทำนายสถานะความยากจนของแรงงานนอกระบบในระดับที่แตกต่างกัน โดยเรียงลำดับตามความสำคัญดังนี้

1) ประเภทค่าจ้างรูปแบบอื่นๆ (WAGE_TYPE_5)

ตัวแปรนี้มีอิทธิพลสูงสุดต่อแบบจำลอง โดยจุดสีแดงที่แทนแรงงานที่ได้รับค่าจ้างในรูปแบบอื่นๆ กระจุกตัวทางด้านขวาอย่างชัดเจน แสดงว่าการได้รับค่าจ้างในรูปแบบอื่นๆ (นอกเหนือจากรายเดือน รายชิ้น หรือไม่เป็นตัวเงิน) มีโอกาสตกอยู่ในสภาวะยากจนสูงกว่าการได้รับค่าจ้างรายวัน (ตัวแปรอ้างอิง) อย่างมีนัยสำคัญ ซึ่งสะท้อนถึงความไม่มั่นคงทางรายได้และการขาดการคุ้มครองทางสังคมของแรงงานกลุ่มนี้

2) จำนวนชั่วโมงทำงานต่อเดือน (TOTAL_HR_MONTH)

การกระจายตัวของจุดแสดงความสัมพันธ์ที่ชัดเจนระหว่างชั่วโมงการทำงานกับสถานะความยากจน โดยจุดสีน้ำเงินที่แทนชั่วโมงการทำงานน้อยกระจุกตัวทางด้านขวา ในขณะที่จุดสีแดงที่แทนชั่วโมงการทำงานมากกระจุกตัวทางด้านซ้าย แสดงให้เห็นว่าแรงงานที่ทำงานน้อยชั่วโมงมีโอกาสเป็นแรงงานยากจนสูง ส่วนการทำงานที่จำนวนชั่วโมงมากช่วยลดโอกาสยากจน ซึ่งสะท้อนถึงความสำคัญของปริมาณงานต่อรายได้และความมั่นคงทางการเงิน

3) ภาคกิจกรรมทางเศรษฐกิจอยู่ในภาคบริการ (INDUS_3)

แรงงานที่ประกอบอาชีพในภาคบริการ (INDUS_3) มีค่า SHAP value เป็นลบ โดยจุดสีแดงที่แทนแรงงานในภาคบริการกระจุกตัวทางด้านซ้ายของกราฟอย่างชัดเจน แสดงให้เห็นว่าการทำงานในภาคบริการช่วยลดโอกาสตกอยู่ในสถานะยากจนเมื่อเปรียบเทียบกับภาคเกษตรกรรม (INDUS_1) ซึ่งเป็นตัวแปรอ้างอิง ผลลัพธ์นี้สะท้อนให้เห็นว่าแม้งานบริการในภาคนอกระบบจะมีความไม่แน่นอนในบางด้าน แต่ยังคงมีโอกาสในการหารายได้และเข้าถึงแหล่งรายได้หลากหลายได้มากกว่าภาคเกษตรกรรม ซึ่งมักประสบปัญหารายได้ไม่สม่ำเสมอตามฤดูกาล ความผันผวนของราคาผลผลิต และข้อจำกัดทางภูมิศาสตร์ในการเข้าถึงตลาดและบริการต่างๆ

4) ดัชนีความไม่เสมอภาคด้านรายได้ (GINI_IDX)

ตัวแปรนี้เป็นข้อมูลระดับจังหวัด ทำให้แรงงานในจังหวัดเดียวกันจะมีค่า GINI เท่ากัน จากกราฟแสดงการกระจายตัวของจุดที่ซับซ้อน โดยมีจุดทั้งสีแดงและน้ำเงินกระจายทั้งสองด้าน แสดงว่าอิทธิพลของความเหลื่อมล้ำรายได้ในระดับจังหวัดต่อความยากจนของแรงงานรายบุคคลไม่เป็นเชิงเส้นตรง โดยขึ้นอยู่กับบริบทของจังหวัดและลักษณะเฉพาะของแรงงานแต่ละคน การที่ข้อมูลเป็นระดับจังหวัดทำให้เกิดการจัดกลุ่มของค่า SHAP ตามแต่ละจังหวัด

5) อายุ (AGE)

จุดสีแดงที่แทนแรงงานอายุมากมีแนวโน้มกระจุกตัวทางด้านขวา แสดงให้เห็นว่าอายุที่เพิ่มขึ้นมีแนวโน้มเพิ่มโอกาสตกอยู่ในสภาวะยากจน อย่างไรก็ตาม การกระจายของจุดในทั้งสองทิศทางแสดงว่าอิทธิพลของอายุต่อความยากจนไม่ได้เป็นไปในทิศทางเดียวกันทั้งหมด โดยอาจขึ้นอยู่กับปัจจัยอื่นๆ ซึ่งสะท้อนให้เห็นถึงความซับซ้อนของความสัมพันธ์ระหว่างอายุกับสถานะทางเศรษฐกิจของแรงงานนอกระบบ

6) ดัชนีความก้าวหน้าของคน (HAI)

ตัวแปรนี้เป็นข้อมูลระดับจังหวัด ทำให้แรงงานในจังหวัดเดียวกันมีค่า HAI เหมือนกัน การกระจายตัวที่ซับซ้อนของจุดในกราฟแสดงว่าอิทธิพลของระดับการพัฒนาคนในจังหวัดต่อความยากจนของแรงงานรายบุคคลมีความแตกต่างกันไปตามบริบท แม้ว่าโดยทั่วไปค่า HAI ที่ต่ำจะมีแนวโน้มเพิ่มโอกาสยากจน แต่ความสัมพันธ์นี้ไม่เป็นเชิงเส้นตรงและขึ้นอยู่กับปัจจัยอื่นๆ ในระดับ

บุคคล การที่เป็นข้อมูลระดับจังหวัดทำให้เกิดผลกระทบแบบกลุ่มที่แรงงานในพื้นที่เดียวกันจะได้รับอิทธิพลจากสภาพแวดล้อมการพัฒนาคนในระดับเดียวกัน

7) ภาคกลาง (REG_2)

จุดสีแดงกระจุกตัวทางด้านซ้าย แสดงว่าแรงงานนอกระบบในภาคกลางมีโอกาสยากจนน้อยกว่าแรงงานในภาคตะวันออกเฉียงเหนือ (ตัวแปรอ้างอิง)

8) แรงงานไร้ฝีมือ (OCCUP_3)

จุดสีแดงที่แทนแรงงานไร้ฝีมือกระจุกตัวทางด้านขวา แสดงว่าแรงงานไร้ฝีมือมีโอกาสตกอยู่ในสภาวะยากจนสูงกว่าแรงงานมีฝีมือ (ตัวแปรอ้างอิง) อย่างชัดเจน ซึ่งสมเหตุสมผลเนื่องจากแรงงานไร้ฝีมือมักขาดทักษะเฉพาะทางที่จะสร้างมูลค่าเพิ่มให้กับงาน ทำให้ได้รับค่าตอบแทนที่ต่ำกว่าและมีความมั่นคงในการจ้างงานน้อยกว่าแรงงานที่มีฝีมือ

9) เพศชาย (SEX_1)

เพศชาย (SEX_1) จุดสีแดงกระจุกตัวทางด้านซ้าย แสดงว่าแรงงานเพศชายมีโอกาสยากจนต่ำกว่าแรงงานเพศหญิง (ตัวแปรอ้างอิง) ซึ่งอาจสะท้อนถึงความแตกต่างด้านโอกาสในการทำงานและรายได้

10) ภาคใต้ (REG_4)

จากกราฟจะเห็นว่า จุดสีแดงที่แสดงถึงการอาศัยอยู่ในภาคใต้ของแรงงานนอกระบบกระจุกตัวอยู่ทางซ้ายมือของกราฟ นั่นหมายความว่าแรงงานในภาคใต้มีโอกาสยากจนต่ำกว่าภาคตะวันออกเฉียงเหนือ (ตัวแปรอ้างอิง) ซึ่งอาจเป็นผลจากโครงสร้างเศรษฐกิจและต้นทุนการครองชีพในพื้นที่

11) ระดับการศึกษา (EDU)

จากกราฟจะเห็นว่าจุดทั้งสองสีมีการแบ่งแยกกันไม่ชัดเจน แสดงว่าผลกระทบของระดับการศึกษาต่อความยากจนมีความซับซ้อนและอาจขึ้นอยู่กับปัจจัยอื่นๆ ร่วมด้วย

12) ภาคเหนือ (REG_3)

จากกราฟจะเห็นว่า จุดสีแดงที่แสดงถึงการอาศัยอยู่ในภาคเหนือของแรงงานนอกระบบกระจุกตัวอยู่ทางซ้ายมือของกราฟ นั่นหมายความว่าแรงงานในภาคเหนือมีโอกาสยากจนต่ำกว่าภาคตะวันออกเฉียงเหนือ (ตัวแปรอ้างอิง) เช่นเดียวกัน

13) หัวหน้าครัวเรือน (RELATION_1)

การเป็นหัวหน้าครัวเรือนจากกราฟจะแสดงด้วยจุดสีแดง พบว่าช่วยลดโอกาสยากจนเมื่อเปรียบเทียบกับการไม่เป็นหัวหน้าครัวเรือน (ตัวแปรอ้างอิง) ซึ่งอาจสะท้อนถึงความรับผิดชอบและแรงจูงใจในการหารายได้

14) สถานประกอบการจดทะเบียน (REGISTER_1)

การทำงานในสถานประกอบการที่จดทะเบียนซึ่งจากกราฟจะแสดงด้วยจุดสีแดง พบว่าช่วยลดโอกาสยากจนเมื่อเปรียบเทียบกับสถานประกอบการไม่จดทะเบียน (ตัวแปรอ้างอิง) เนื่องจากมีความมั่นคงและการคุ้มครองมากกว่า

15) ใช้ที่อยู่อาศัยเป็นสถานที่ทำงาน (W_PLACE_3)

จากกราฟจะเห็นว่าจุดสีแดงกระจุกตัวทางด้านขวา แสดงว่าการทำงานในที่อยู่อาศัยมีโอกาสยากจนสูงกว่าการใช้สถานประกอบการของเจ้าของหรือตนเองเป็นที่ทำงาน (ตัวแปรอ้างอิง) ซึ่งอาจสะท้อนถึงลักษณะงานที่ไม่เป็นทางการและมีรายได้ไม่แน่นอน

การกระจายตัวของจุดข้อมูลในแต่ละตัวแปรแสดงให้เห็นว่าแบบจำลองพิจารณาปัจจัยหลายตัวร่วมกันในการตัดสินใจ ไม่ได้พึ่งพาตัวแปรใดตัวแปรหนึ่งเพียงอย่างเดียว ซึ่งสะท้อนถึงความซับซ้อนและหลากหลายของปัจจัยที่ส่งผลต่อการจำแนกสถานะความยากจนของแรงงานนอกระบบ

4.7 อภิปรายผลการวิจัย

ผลการศึกษาแสดงให้เห็นถึงความสำคัญของการเลือกเทคนิคการจัดการข้อมูลไม่สมดุลให้เหมาะสมกับลักษณะของแบบจำลองแต่ละประเภท ซึ่งสอดคล้องกับแนวคิดในงานวิจัยก่อนหน้า เช่น การศึกษาของ Fernández et al. (2018) ที่แสดงให้เห็นว่าไม่มีเทคนิคใดที่เหมาะสมกับทุกสถานการณ์ การเลือกเทคนิคที่เหมาะสมควรขึ้นอยู่กับลักษณะของข้อมูลและประเภทของแบบจำลองที่ใช้ โดยในกรณีของงานวิจัยนี้ แบบจำลอง Random Forest ให้ผลลัพธ์ที่ดีกว่าเมื่อใช้ Random Undersampling ขณะที่แบบจำลอง Logistic Regression และ XGBoost ให้ผลลัพธ์ดีกว่าเมื่อใช้ SMOTE-ENN สะท้อนถึงความทนทานต่อปัญหาความไม่สมดุลของข้อมูล

จากการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 แบบจำลอง พบว่า XGBoost ให้ประสิทธิภาพสูงสุดในการทำนายสถานะความยากจนของแรงงานนอกระบบ โดยมีค่าประสิทธิภาพโดยรวม (F1-Score) ที่ 53.03% ซึ่งสูงกว่าแบบจำลองอื่นๆ อย่างมีนัยสำคัญ ผลลัพธ์นี้สอดคล้องกับงานวิจัยของ Chen and Guestrin (2016) ที่แสดงให้เห็นว่า XGBoost มีความสามารถในการจัดการกับข้อมูลที่มีความซับซ้อนและความสัมพันธ์ที่ไม่เป็นเชิงเส้นได้ดีกว่าแบบจำลองแบบดั้งเดิม อย่างไรก็ตาม ประสิทธิภาพของแบบจำลอง XGBoost ไม่ได้เหนือกว่าแบบจำลองอื่นๆในทุกบริบท การศึกษาของ Song (2015) ที่เปรียบเทียบ XGBoost กับ Random Forest พบว่า Random Forest สามารถทำงานได้ดีกว่า XGBoost อย่างมีนัยสำคัญในบางกรณี โดยเฉพาะเมื่อ XGBoost ไม่ได้รับการปรับแต่งพารามิเตอร์อย่างเหมาะสม ซึ่งสะท้อนให้เห็นว่า XGBoost เป็นแบบจำลองที่ซับซ้อนกว่า Random Forest และต้องการการปรับแต่งพารามิเตอร์ที่ละเอียดกว่า โดยเฉพาะ Learning Rate และ Regularization Parameters ซึ่งการที่ XGBoost ในการศึกษานี้ให้ประสิทธิภาพที่เหนือกว่าอาจเกิดจากหลายปัจจัย เช่น ข้อมูลแรงงานนอกระบบมีความซับซ้อนและความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างตัวแปรต่างๆ ซึ่ง XGBoost มีความสามารถในการจับความสัมพันธ์เหล่านี้ได้ดีกว่าด้วยการใช้ Gradient Boosting Approach ที่สร้างต้นไม้แบบลำดับและแก้ไขข้อผิดพลาดจากต้นไม้ก่อนหน้า XGBoost พัฒนาต้นไม้ทีละต้นโดยแก้ไขความผิดพลาดที่เกิดจากต้นไม้ที่ได้รับการฝึกฝนก่อนหน้า ซึ่งแตกต่างจาก Random Forest ที่แต่ละต้นไม้ถูกสร้างขึ้นอย่างเป็นอิสระ สำหรับ Logistic Regression ทั้งวิธีทางสถิติและการเรียนรู้ของเครื่อง แม้จะมีข้อดีในการตีความได้ง่ายและความเรียบง่าย แต่มีข้อจำกัดในการจัดการกับความสัมพันธ์ที่ซับซ้อน Lou (2024) อธิบายว่าผลกระทบของการเตรียมข้อมูลต่อความแม่นยำของ Logistic Regression มีความสำคัญอย่างมาก เนื่องจากแบบจำลองนี้อาศัยสมมติฐานว่าตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกับ Log Odds ของผลลัพธ์ การเตรียมข้อมูลที่มีประสิทธิภาพ เช่น การปรับขนาดคุณลักษณะและการจัดการค่าผิดปกติ สามารถทำให้ข้อมูลสอดคล้องกับข้อสมมติเหล่านี้มากขึ้น ซึ่งมีความสำคัญต่อประสิทธิภาพของแบบจำลอง Logistic Regression แต่มีความสำคัญน้อยกว่าสำหรับแบบจำลอง Random Forest และ XGBoost ที่มีความทนทานต่อการกระจายข้อมูลที่แตกต่างกันตามธรรมชาติ ผลการศึกษานี้แสดงให้เห็นว่าปัจจัยที่ส่งผลต่อความยากจนของแรงงานนอกระบบมีความซับซ้อนที่ต้องการแบบจำลองที่สามารถจับ

ความสัมพันธ์ที่ไม่เป็นเชิงเส้นและการมีปฏิสัมพันธ์ระหว่างตัวแปรได้อย่างมีประสิทธิภาพ ซึ่ง XGBoost สามารถตอบสนองความต้องการนี้ได้ดีกว่าแบบจำลองอื่นๆ

การวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP จากแบบจำลอง XGBoost พบว่า ประเภทค่าจ้างรูปแบบอื่นๆ (WAGE_TYPE_5) เป็นปัจจัยสำคัญที่สุดในการทำนายสถานะความยากจนของแรงงานนอกระบบ ซึ่งแรงงานที่ได้รับค่าตอบแทนไม่แน่นอน มีโอกาสตกอยู่ในสภาวะยากจนสูงกว่าแรงงานที่ได้รับค่าจ้างรายวัน สะท้อนถึงความเสี่ยงด้านความไม่มั่นคงทางรายได้ของแรงงานกลุ่มนี้ ปัจจัยสำคัญอันดับสอง คือ จำนวนชั่วโมงทำงานต่อเดือน (TOTAL_HR_MONTH) ที่แสดงความสัมพันธ์เชิงลบกับความยากจน แรงงานที่ทำงานน้อยชั่วโมงมีโอกาสเป็นแรงงานยากจนสูงกว่า ซึ่งสอดคล้องกับการศึกษาของผกามาศ (2561) ที่พบว่าแรงงานนอกระบบที่ยากจนมักเลือกเพิ่มชั่วโมงทำงานเพื่อแสวงหารายได้เลี้ยงชีพ สำหรับด้านภาคกิจกรรมทางเศรษฐกิจ แรงงานในภาคบริการ (INDUS_3) มีโอกาสยากจนต่ำกว่าภาคเกษตรกรรม ซึ่งสะท้อนถึงความไม่สม่ำเสมอของรายได้ตามฤดูกาลและความผันผวนของราคาผลผลิตทางการเกษตร ในขณะที่ตัวแปรระดับบุคคลอื่นๆ เช่น การเป็นแรงงานไร้ฝีมือ (OCCUP_3) การใช้ที่อยู่อาศัยเป็นสถานที่ทำงาน (W_PLACE_3) และการทำงานในสถานประกอบการที่ไม่จดทะเบียน ล้วนแสดงแนวโน้มเพิ่มโอกาสความยากจน ผลการศึกษานี้สอดคล้องกับงานวิจัยของ Roy and Kundu (2020) ที่พบว่าแรงงานที่ไม่มีสถานที่ทำงานถาวรหรือทำงานกลางแจ้ง เช่น ตามถนนหรือตลาด มีโอกาสจนสูงกว่ากลุ่มที่ทำงานในสำนักงานหรือที่อยู่อาศัยอย่างเป็นทางการ และสำหรับปัจจัยระดับจังหวัด ดัชนีความไม่เสมอภาคด้านรายได้ (GINI_IDX) และดัชนีความก้าวหน้าของคน (HAI) ชี้ให้เห็นว่าการแก้ไขปัญหาความยากจนจำเป็นต้องพิจารณาทั้งปัจจัยระดับบุคคลและระดับพื้นที่อย่างครอบคลุม ซึ่งสอดคล้องกับการศึกษาของสุพนิดา (2565) ที่เน้นความสำคัญของปัจจัยหลากหลายมิติในการวิเคราะห์ความยากจนของกลุ่มประชากรเปราะบาง

ผลการวิเคราะห์นี้ชี้ให้เห็นว่า ปัจจัยที่ส่งผลต่อการจำแนกสถานะความยากจนมีหลากหลายมิติ ทั้งด้านสภาพแวดล้อมการทำงาน เช่น ประเภทค่าจ้างที่ได้รับและจำนวนชั่วโมงทำงาน ด้านโครงสร้างเศรษฐกิจ เช่น ความดัชนีความไม่เสมอภาคด้านรายได้ และดัชนีความก้าวหน้าของคน รวมถึงปัจจัยส่วนบุคคลอย่างอายุและการศึกษา ข้อค้นพบเหล่านี้มีนัยสำคัญต่อการกำหนดนโยบายแบบบูรณาการที่ต้องเชื่อมโยงทั้งมิติส่วนบุคคล เศรษฐกิจ และสังคมเข้าด้วยกันอย่างสอดคล้อง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการจำแนกความยากจนของแรงงานนอกระบบ และเปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกสถานะความยากจน กลุ่มเป้าหมายคือแรงงานนอกระบบที่มีอายุตั้งแต่ 15 ปีขึ้นไป และไม่ได้รับหลักประกันทางสังคมจากการทำงาน โดยเกณฑ์ในการจำแนกกลุ่มยากจนใช้เส้นความยากจนระดับประเทศที่ 3,043 บาทต่อคนต่อเดือนเป็นเกณฑ์ในการแบ่งกลุ่มแรงงานเป็นกลุ่มยากจนและไม่ยากจน การศึกษานี้มุ่งหวังให้หน่วยงานภาครัฐและองค์กรที่เกี่ยวข้องสามารถนำผลการวิเคราะห์ไปใช้ในการออกแบบมาตรการและนโยบายสวัสดิการที่ตอบโจทย์กลุ่มแรงงานนอกระบบได้อย่างมีประสิทธิภาพและตรงเป้าหมาย โดยผลการศึกษาจะช่วยให้สามารถระบุลักษณะของกลุ่มแรงงานที่มีความเสี่ยงต่อความยากจนได้แม่นยำยิ่งขึ้น ซึ่งจะส่งผลให้แรงงานนอกระบบได้รับการสนับสนุนที่เหมาะสม เช่น การเข้าถึงแหล่งรายได้ที่มั่นคง การฝึกอบรมทักษะอาชีพ การเข้าถึงบริการสาธารณสุขและการศึกษา ตลอดจนการมีโอกาสได้รับสวัสดิการหรือมาตรการช่วยเหลือที่เหมาะสมกับบริบทของตนเอง อันจะนำไปสู่การลดภาระทางเศรษฐกิจ ยกระดับรายได้ และพัฒนาคุณภาพชีวิตอย่างยั่งยืนในระยะยาว ซึ่งสามารถสรุปผลการวิจัยได้ดังต่อไปนี้

1. สรุปผลการเปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุลทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test
2. สรุปผลเปรียบเทียบประสิทธิภาพในการทำนายความยากจนของแรงงานนอกระบบของทั้ง 4 แบบจำลอง
3. สรุปผลการวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP
4. ข้อเสนอแนะ

5.1 สรุปผลการเปรียบเทียบประสิทธิภาพของแบบจำลองในการจัดการปัญหาข้อมูลไม่สมดุลทั้ง 2 วิธี ด้วยสถิติทดสอบ t-test

ในการศึกษานี้ได้ทำการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลไม่สมดุลสองวิธี ได้แก่ Random Undersampling และ SMOTE-ENN โดยประเมินผ่านเทคนิค k-fold Cross-Validation (k = 5) และทดสอบความแตกต่างของค่าประสิทธิภาพโดยรวม (F1-Score) ด้วยสถิติ Independent t-test สำหรับแบบจำลอง 3 แบบจำลอง ได้แก่ แบบจำลอง Logistic Regression, แบบจำลอง Random Forest และแบบจำลอง XGBoost

ผลการศึกษาพบว่าแบบจำลอง Logistic Regression (Threshold = 0.5) ที่ปรับสมดุลข้อมูลด้วยวิธี Random Undersampling ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงกว่าวิธี SMOTE-ENN อย่างมีนัยสำคัญทางสถิติ (p-value = 0.0006) แสดงให้เห็นว่าเมื่อปรับสมดุลข้อมูลด้วยการลดจำนวนข้อมูลแรงงานนอกระบบที่มีสถานะไม่ยากจนให้สมดุลกับแรงงานนอกระบบที่มีสถานะยากจนทำให้แบบจำลองนี้ให้ผลลัพธ์ที่ดีกว่า

สำหรับแบบจำลอง XGBoost ที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงกว่า Random Undersampling อย่างมีนัยสำคัญทางสถิติ (p-value = 0.0000) โดยมีค่าเฉลี่ย F1-Score เท่ากับ 0.5170 เทียบกับ 0.4451 จากวิธี Random Undersampling

แต่สำหรับแบบจำลอง Logistic Regression (Threshold = 0.8241) และ Random Forest ไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติระหว่างวิธี Random Undersampling และ SMOTE-ENN ซึ่งบ่งชี้ว่าทั้งสองวิธีให้ผลลัพธ์ใกล้เคียงกัน

โดยสรุป การเลือกวิธีการจัดการข้อมูลไม่สมดุลควรพิจารณาตามลักษณะของแบบจำลองที่ใช้ เนื่องจากไม่มีวิธีใดที่ให้ผลดีกว่าในทุกสถานการณ์

5.2 สรุปผลเปรียบเทียบประสิทธิภาพในการทำนายความยากจนของแรงงานนอกระบบของทั้ง 4 แบบจำลอง

ในการศึกษาครั้งนี้ได้พัฒนาแบบจำลองทำนายความยากจนของแรงงานนอกระบบ 4 แบบจำลอง ได้แก่ แบบจำลอง Logistic Regression ด้วยวิธีทางสถิติ และแบบจำลองที่พัฒนาด้วยการเรียนรู้ของเครื่อง 3 แบบจำลอง คือ Logistic Regression, Random Forest และ XGBoost โดยแบบจำลองที่พัฒนาด้วยการเรียนรู้ของเครื่องใช้เทคนิคการจัดการปัญหาข้อมูลไม่สมดุลสองวิธี ได้แก่ Random Undersampling และ SMOTE-ENN เพื่อแก้ไขปัญหาที่ข้อมูลแรงงานนอกระบบกลุ่มยากจนมีจำนวนน้อยกว่ากลุ่มไม่ยากจน นอกจากนี้ยังมีการปรับ Threshold สำหรับแบบจำลอง Logistic Regression เพื่อหาค่าที่เหมาะสมที่สุดในการจำแนกกลุ่ม

จากการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 4 แบบจำลอง โดยพิจารณาจากค่าประสิทธิภาพโดยรวม (F1-Score) พบว่าแบบจำลองที่มีประสิทธิภาพสูงสุดคือ XGBoost ที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN โดยใช้พารามิเตอร์ที่เหมาะสมที่สุด ได้แก่ n_estimators เท่ากับ 200 ต้น, max_depth ในระดับ 6, learning_rate ที่ 0.1, subsample และ colsample_bytree ที่ 0.8 ซึ่งให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงสุดอยู่ที่ 0.5303 และสามารถจำแนกสถานะความยากจนของแรงงานนอกระบบได้ถูกต้อง 565 คนจากทั้งหมด 914 คน อันดับที่สองคือแบบจำลอง Random Forest ที่ปรับสมดุลข้อมูลด้วยวิธี Random Undersampling ซึ่งให้ค่าประสิทธิภาพโดยรวม (F1-Score) อยู่ที่ 0.4452 อันดับที่สามเป็นแบบจำลอง Logistic Regression ด้วยวิธีทางสถิติที่ปรับ Threshold เท่ากับ 0.2341 สำหรับแบบจำลองที่มีประสิทธิภาพต่ำสุดคือ Logistic Regression ด้วยการเรียนรู้ของเครื่องที่ปรับ Threshold เป็น 0.8264 และปรับสมดุลข้อมูลด้วยวิธี SMOTE-ENN แม้ว่าจะใช้เทคนิคการเรียนรู้ของเครื่องและการปรับสมดุลข้อมูล แต่ก็ยังให้ประสิทธิภาพที่ต่ำกว่าแบบจำลองอื่นๆ ในการศึกษานี้

การเปรียบเทียบประสิทธิภาพของแบบจำลอง XGBoost กับแบบจำลองอื่นๆ การศึกษานี้ได้ทำการทดสอบทางสถิติด้วย McNemar's Test ตามที่แสดงในตารางที่ 4.19 ผลการทดสอบพบว่าแบบจำลอง XGBoost มีประสิทธิภาพเหนือกว่าแบบจำลองอื่นๆ ทั้ง 3 แบบจำลองอย่างมีนัยสำคัญทางสถิติ (p-value = 0.000 ในทุกการเปรียบเทียบ) โดยเฉพาะการเปรียบเทียบกับ Logistic Regression ด้วยวิธีทางสถิติ แบบจำลอง XGBoost สามารถทำนายแรงงานนอกระบบที่มีสถานะยากจนได้ถูกต้องมากกว่า 603 คน ในขณะที่แบบจำลอง Logistic Regression ด้วยวิธีทางสถิติทำนายถูกต้องมากกว่าเพียง 155 คน ส่วนการเปรียบเทียบกับ Logistic Regression ด้วยการเรียนรู้ของเครื่อง และแบบจำลอง Random Forest ก็แสดงผลในทิศทางเดียวกัน โดยแบบจำลอง

XGBoost มีประสิทธิภาพเหนือกว่า ซึ่งยืนยันได้ว่าแบบจำลอง XGBoost ที่ใช้เทคนิค SMOTE-ENN เป็นแบบจำลองที่มีประสิทธิภาพสูงสุดในการจำแนกสถานะความยากจนของแรงงานนอกระบบอย่างเป็นรูปธรรม

5.3 สรุปผลการวิเคราะห์อิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP

จากการทดลองสรุปได้ว่าแบบจำลอง XGBoost ที่ใช้วิธีปรับสมดุลของข้อมูลด้วย SMOTE-ENN ให้ค่าประสิทธิภาพโดยรวม (F1-Score) สูงที่สุดอยู่ที่ 0.5303 และให้ค่าความระลึก (Recall) สูงในระดับใกล้เคียงกับแบบจำลองอื่น ๆ แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกแรงงานนอกระบบที่มีสถานะยากจนได้อย่างแม่นยำ อีกทั้งยังให้ค่า Balanced Accuracy สูงที่สุดเท่ากับ 0.7663 ซึ่งสะท้อนถึงความสามารถในการทำนายได้ดีทั้งในกลุ่มแรงงานยากจนและไม่ยากจน

ผู้วิจัยจึงพิจารณาอิทธิพลของตัวแปรต่อแบบจำลอง ด้วย SHAP จากแบบจำลอง XGBoost เพียงแบบเดียว พบว่าตัวแปรที่ส่งผลต่อการทำนายมากที่สุด ได้แก่ ประเภทค่าจ้างรูปแบบอื่น (WAGE_TYPE_5) ซึ่งแรงงานนอกระบบที่มีค่าจ้างในลักษณะที่ไม่ใช่รายวัน รายเดือน รายสัปดาห์ หรือไม่เป็นตัวเงิน มีแนวโน้มยากจนมากกว่ากลุ่มอื่น สะท้อนถึงความไม่มั่นคงของรายได้และการขาดการคุ้มครองทางสังคม ซึ่งเป็นปัญหาโครงสร้างพื้นฐานของตลาดแรงงานนอกระบบ รองลงมาคือ จำนวนชั่วโมงทำงานต่อเดือน (TOTAL_HR_MONTH) แรงงานนอกระบบที่จำนวนชั่วโมงการทำงานน้อยมีแนวโน้มเพิ่มความเสี่ยงยากจน ตามมาด้วยภาคกิจกรรมทางเศรษฐกิจภาคบริการ (INDUS_3) ซึ่งแรงงานในภาคบริการมักมีโอกาสตกอยู่ในภาวะยากจนน้อยกว่าภาคเกษตรกรรมที่มีรายได้ขึ้นอยู่กับปริมาณผลผลิต และราคาของผลผลิตตามฤดูกาลนั้นๆ

นอกจากนี้ยังพบว่าความเหลื่อมล้ำด้านรายได้ (GINI_IDX) และอายุ (AGE) เมื่อมีค่าสูงขึ้นจะสัมพันธ์กับโอกาสยากจนที่เพิ่มขึ้น ในขณะที่ค่าดัชนีความก้าวหน้าของคน (HAI) หากลดต่ำลงจะเพิ่มโอกาสยากจนเช่นกัน รวมถึงตัวแปรด้านภูมิภาคที่อยู่อาศัย เช่น ภาคกลาง ภาคใต้ ภาคเหนือมีแนวโน้มสัมพันธ์กับโอกาสยากจนน้อยกว่าภาคตะวันออกเฉียงเหนือ

จากผลการวิเคราะห์นี้สามารถสรุปได้ว่า ปัจจัยด้านสภาพแวดล้อมการทำงาน เช่น ประเภทค่าจ้าง ภาคกิจกรรมทางเศรษฐกิจ และจำนวนชั่วโมงการทำงาน เป็นปัจจัยหลักที่ส่งผลโดยตรงต่อโอกาสความยากจนของแรงงานนอกระบบ รองลงมาคือปัจจัยด้านเศรษฐกิจ เช่น ดัชนีความไม่เสมอภาคด้านรายได้ และดัชนีความก้าวหน้าของคน ซึ่งสะท้อนถึงความสามารถในการเข้าถึงโอกาสทางเศรษฐกิจของแรงงานแต่ละคน ขณะเดียวกัน ปัจจัยด้านลักษณะส่วนบุคคล เช่น อายุและระดับการศึกษา ก็มีผลต่อศักยภาพในการหารายได้อย่างมีนัยสำคัญ และปัจจัยภูมิภาคหรือสภาพแวดล้อมที่อยู่อาศัย ก็สะท้อนถึงความเหลื่อมล้ำเชิงพื้นที่ที่มีผลต่อโอกาสยากจน

5.4 ข้อเสนอแนะ

1. การวิจัยครั้งนี้อาศัยข้อมูลทุติยภูมิในการวิเคราะห์ ซึ่งอาจมีข้อจำกัดในการสะท้อนภาพรวมความซับซ้อนของปัญหาความยากจนในแรงงานนอกระบบ เพื่อเพิ่มประสิทธิภาพของแบบจำลองในอนาคต ควรมีการสำรวจและรวบรวมตัวแปรเพิ่มเติม อาทิ ตัวแปรด้านสุขภาพ (ส่งผลต่อความสามารถในการทำงาน) การครอบครองทรัพย์สิน (สะท้อนฐานะทางเศรษฐกิจ) เป็นต้น

2. สำหรับหน่วยงานภาครัฐควรพิจารณาออกแบบนโยบายที่เหมาะสมกับกลุ่มแรงงานนอกระบบ โดยเฉพาะแรงงานที่ได้รับค่าจ้างในรูปแบบไม่แน่นอน หรือทำงานในสถานที่ที่ไม่ได้จดทะเบียน และมีระดับการศึกษาต่ำ ซึ่งเป็นกลุ่มที่มีความเสี่ยงสูงต่อความยากจน การส่งเสริมให้แรงงานเข้าถึงระบบสวัสดิการ ฝึกอบรมทักษะอาชีพ และการสนับสนุนให้มีชั่วโมงการทำงานที่มั่นคง จะช่วยลดความเสี่ยงต่อความยากจนได้อย่างเป็นรูปธรรม

3. การประยุกต์ใช้แบบจำลองในการบริหารจัดการนโยบาย เพื่อให้หน่วยงานที่เกี่ยวข้องสามารถนำแบบจำลองที่พัฒนาขึ้นมาใช้เป็นเครื่องมือในการทำนายกลุ่มแรงงานนอกระบบที่มีความเสี่ยงสูงต่อการตกอยู่ในภาวะความยากจน เพื่อกำหนดนโยบายเชิงป้องกันและแก้ไขปัญหาความยากจนอย่างมีประสิทธิภาพ ดังนี้

3.1) การประยุกต์ใช้แบบจำลอง XGBoost เพื่อสนับสนุนการคัดกรองกลุ่มแรงงานนอกระบบที่มีความเสี่ยงยากจนในกรณีข้อมูลรายได้ไม่ชัดเจน เนื่องจากแรงงานนอกระบบจำนวนมากมีรายได้ไม่แน่นอนหรือไม่สามารถระบุรายได้ที่แท้จริงได้ การใช้แบบจำลอง XGBoost ซึ่งพิจารณาปัจจัยอื่นที่เกี่ยวข้อง เช่น รูปแบบค่าจ้าง ชั่วโมงทำงาน และประเภทอาชีพ จึงสามารถช่วยประเมินความเสี่ยงจากโอกาสที่จะเป็นกลุ่มยากจนได้อย่างมีประสิทธิภาพ ซึ่งช่วยให้สามารถจัดสรรสวัสดิการและการดำเนินนโยบายช่วยเหลือ และสนับสนุนไปยังกลุ่มเป้าหมายได้อย่างความแม่นยำ ครอบคลุมและเป็นธรรมมากยิ่งขึ้น

3.2) การติดตามและกำหนดนโยบายรายพื้นที่ โดยหน่วยงานที่เกี่ยวข้องสามารถนำผลลัพธ์จากการวิเคราะห์อิทธิพลของตัวแปรด้วย SHAP มาใช้ในการติดตามและวิเคราะห์ปัจจัยเสี่ยงสำคัญในแต่ละพื้นที่ เช่น ประเภทค่าจ้างรูปแบบอื่น จำนวนชั่วโมงการทำงานต่อเดือน และภาคกิจกรรมทางเศรษฐกิจ ซึ่งช่วยให้เจ้าหน้าที่ทราบถึงปัจจัยที่ต้องกำหนดมาตรการช่วยเหลือให้ตรงจุด ทันเวลา และสนับสนุนการกำหนดนโยบายแบบมุ่งเป้าเฉพาะกลุ่มหรือพื้นที่ที่มีความเปราะบางสูง ได้อย่างมีประสิทธิภาพ เช่น

3.2.1) การส่งเสริมอาชีพเพื่อสร้างรายได้ประจำให้กับแรงงานนอกระบบในพื้นที่ เพื่อยกระดับรายได้ของแรงงานนอกระบบที่มีความเปราะบาง โดยเฉพาะกลุ่มที่มีค่าจ้างไม่แน่นอนหรือมีชั่วโมงทำงานต่ำ หน่วยงานภาครัฐ เช่น กรมพัฒนาฝีมือแรงงาน ควรจัดโครงการฝึกอบรมทักษะอาชีพที่ตอบโจทย์ตลาดแรงงานในพื้นที่ เช่น ทักษะด้านดิจิทัล งานบริการ หรืออาชีพอิสระที่สามารถทำจากที่บ้านได้ ในขณะเดียวกัน กรมการจัดหางาน ควรมีบทบาทในการเชื่อมโยงแรงงานกับโอกาสการจ้างงานในพื้นที่ ทั้งในภาครัฐและภาคเอกชน โดยให้ความสำคัญกับงานที่มีรายได้แน่นอนและต่อเนื่อง

3.2.2) การส่งเสริมความร่วมมือกับภาคเอกชนในการสร้างงานที่มั่นคงในพื้นที่ เพื่อให้แรงงานนอกระบบสามารถเข้าถึงงานที่มีค่าตอบแทนแน่นอนและชั่วโมงการทำงานที่เหมาะสม ควรมีการส่งเสริมความร่วมมือระหว่างภาครัฐกับ ภาคเอกชน ในการพัฒนารูปแบบการจ้างงานที่เหมาะสมกับบริบทของพื้นที่ เช่น งาน part-time งานตามฤดูกาล หรือการจ้างงานในรูปแบบสัญญาระยะยาวที่มีความมั่นคงมากขึ้นภาครัฐโดยเฉพาะ สำนักงานส่งเสริมวิสาหกิจขนาดกลางและขนาดย่อม (สสว.) และ สำนักงานคณะกรรมการส่งเสริมการลงทุน (BOI) ควรมีมาตรการจูงใจ เช่น การให้สิทธิประโยชน์ทางภาษีหรือการสนับสนุนเงินทุนแก่ภาคธุรกิจที่เข้าร่วมจ้างงานแรงงานนอกระบบในพื้นที่เปราะบาง

เอกสารอ้างอิง

กาญจน์เขจร ชูชีพ. 2561. การถดถอยโลจีสติก. *Remote Sensing Technical* Note. 5: 1-10.

กิตติมศักดิ์ ในจิต. 2564. **Logistic Regression จัดกลุ่มแบบ Binary Classification ใน Machine Learning ด้วย Python**. [Online]. เข้าถึงได้จาก : https://kittimasak.com/logistic-regression-binary-classification-machine-learning-python/

ชูชิต ชายทวีป. 2561. ปัจจัยสำเร็จของการลดปัญหาความยากจน. *วารสารสังคมศาสตร์บูรณาการ*. 3(2) : 188-214.

ณัฐวดี ศิริประสมทรัพย์. 2565. Interpretable machine learning for feature selection using SHAP values. วิทยาศาสตรมหาบัณฑิต. มหาวิทยาลัยธรรมศาสตร์.

ณัฐโชติ พรหมฤทธิ์ และสัจจาภรณ์ ไวจรรยา. 2567. **Random Forest**. [Online]. เข้าถึงได้จาก : https://blog.pjjop.org/random-forest/#:~:text=Random%20Forest%.

นิภาพรรณ เจนสันติกุล. 2566. ความยากจนและความเหลื่อมล้ำในสังคมไทย. *วารสารมนุษยศาสตร์และสังคมศาสตร์มหาวิทยาลัยราชภัฏเลย.* 6(1) : 61-74.

ทินรัตน์ อารมย์สุข. 2565. การตีความตัวแบบการเรียนรู้ของเครื่องด้วย SHAP (Shapley Additive Explanations): กรณีศึกษาแบบจำลองการจำแนกและแบบจำลองการถดถอย. วิทยาศาสตรมหาบัณฑิต. มหาวิทยาลัยธรรมศาสตร์.

ฐิติพร อ้ายดี. 2564. การศึกษาตัวจำแนกประเภทการเรียนรู้ของเครื่องเพื่อทำนายโรคหลอดเลือดสมอง. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมเครื่องกล. จุฬาลงกรณ์มหาวิทยาลัย.

ปวริศ เวชวรรณกิจกุล. 2565. การคาดการณ์เวลาเดินทางบนท้องถนนระหว่างพิกัดสองจุดในกรุงเทพมหานครด้วยวิธีการเรียนรู้ของเครื่อง. วิศวกรรมศาสตรมหาบัณฑิต. จุฬาลงกรณ์มหาวิทยาลัย.

ผกามาศ เลียงธนะฤกษ์. 2561. ค่าจ้างและความต้องการทำงานเพิ่มของแรงงานนอกระบบ: กรณีศึกษาในประเทศไทย. ศิลปศาสตรมหาบัณฑิต. จุฬาลงกรณ์มหาวิทยาลัย.

ภาณุภณ จิระอัมพร. 2564. การปรับปรุงวิธีการสุ่มตัวอย่างใหม่สำหรับข้อมูลไม่สมดุลด้วยเทคนิคแบบผสม DB2SM. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่. มหาวิทยาลัยธุรกิจบัณฑิตย์.

ภัทรียา นวลใย. 2559. **การประเมินความยากจนของแรงงานนอกระบบโดยใช้ Machine Learning**. [Online]. เข้าถึงได้จาก : https://www.car.chula.ac.th/display7.php?bib=2156489

ยุทธ ไกยวรรณ์. 2555. หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย. *วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย.* 4(1) : 1-12.

วิชญ์วิสิฐ เกษรสิทธิ์, วิชิต หล่อจีระชุณห์กุล และจิราวัลย์ จิตรถเวช. 2561. การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน. *วารสารวิจัยมหาวิทยาลัยขอนแก่น*. 18(3) : 11-12.

สุพนิดา จิระสินวรรธนะ. 2565. ความยากจนในผู้สูงอายุไทย: การเปลี่ยนแปลงและปัจจัยเสี่ยง. ปรัชญาดุษฎีบัณฑิต. จุฬาลงกรณ์มหาวิทยาลัย.

สุรศักดิ์ ชามะรักษ์. 2560. การจัดการทุนมนุษย์ (HCM): ความหมาย ความสำคัญ วิวัฒนาการ และแนวโน้ม. *วารสารวิจัยการและวิทยาลัยมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี สาขามนุษยศาสตร์และสังคมศาสตร์.* 2(1) : 56-67.

สำนักงานสถิติแห่งชาติ. 2562. **รายงานการวิเคราะห์ภาวะหนี้สินของครัวเรือนเกษตรกร พ.ศ. 2562**. [Online]. เข้าถึงได้จาก : https://www.nso.go.th/public/e-book/Analytical-Reports/Agri_Household_Deb62/5/

สำนักงานสถิติแห่งชาติ. 2567. **การสำรวจแรงงานนอกระบบ พ.ศ. 2567.** [Online]. เข้าถึงได้จาก : https://www.nso.go.th/nsoweb/storage/survey_detail/2025/20241125143625_37256.pdf

สำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ. 2566. **รายงานการวิเคราะห์สถานการณ์ความยากจนและความเหลื่อมล้ำในไทย**. [Online]. เข้าถึงได้จาก : https://www.nesdc.go.th/ewt_dl_link.php?nid=15744

สำเริง ไกยวงค์. 2565. การทดสอบไคสแควร์: สารสนเทศสำคัญอีก 2 ประเภทที่นักวิจัยควรเขียนรายงานให้ครบถ้วน. *วารสารปัญญาภิวัฒน์.* 14(1) : 333-340.

เครือวัลย์ เนตรพนา. 2565. การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง. วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล. มหาวิทยาลัยศรีนครินทรวิโรฒ.

เอลวิส โคตรชมภู และจุฬาพรรณภรณ์ ธนะแพทย์. 2565. การบริหารทุนมนุษญ์ในยุคศตวรรษที่ 21. *วารสาร มจร อุบลปริทรรศน์.* 7(4) : 1017-1028.

อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง. 2562. การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี. *Thai Journal of Science and technology.* 9(4) : 418-435.

Ali, Z.A., Abduljabbar, Z.H., Tahir, H.A., Sallow, A.B. and Almufti, S.M. 2023. Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. *Academic Journal of Nawroz University (AJNU)*. 12: 320-334.

Adhikari, D.B. 2020. Factors influencing the income of urban informal workers: Evidence from Nepal. *Economic Journal of Development Issues.* 29-30: 13-24.

Aviv Nahon. 2019. XGBoost, LightGBM or CatBoost — **Which boosting algorithm should I use?**. [Online]. Available : https://medium.com/riskified-technology/xgboost-lightgbm-or-catboost-which-boosting-algorithm-should-i-use-e7fda7bb36bc

Boonto, S. 2018. **Lecture 4: Training Models II**. [Online]. Available : https://inc.kmutt.ac.th/~sudchai.boo/Teaching/inc693d/lecture4_2017.pdf

Chen, T. and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785-794.

El Mahdi, A. 2010. Poverty and informality: A restraining or constructive relationship? Working Paper No. 569. The Economic Research Forum.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and Herrera, F. 2018. Learning from imbalanced data sets. Springer.

GeeksforGeeks. 2023. **Basic concept of classification (data mining).** [Online]. Available : https://www.geeksforgeeks.org/basic-concept-classification-data-mining/

Hsu, S.C., Sharma, A.K., Tanone, R. and Ye, Y.T. 2024. "Predicting Rainfall Using Random Forest and CatBoost Models". in **Proceedings of the 9th World Congress on Civil, Structural, and Environmental Engineering (CSEE 2024).** London, United Kingdom.

Iguazio. 2024. Classification threshold. [Online]. Available : https://www.iguazio.com/glossary/classification-threshold/

Iwagami, M., Anzai, D., Kido, A. and Tamiya, H. 2024. Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records. *PLOS Digital Health*. 3 : e0000578.

Lee, S. 2025. **Exploring Point-Biserial Correlation: 7 Key Stats for Analysis**. [Online]. Available : https://www.numberanalytics.com/blog/exploring-point-biserial-correlation-7-stats.

Li, Q., Yu, S., Échevin, D. and Fan, M. 2022. Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. *Socio-Economic Planning Sciences*. 81(1) : Article 101195.

Lou, J. 2024. Comparative Analysis of Logistic Regression, Random Forest, and XGBoost for Click-Through Rate Prediction in Digital Advertising. *Advances in Economics, Business and Management Research*. 300(1) : 462-470.

Machado, L. and Holmer, D. 2022. Credit risk modelling and prediction: Logistic regression versus machine learning boosting algorithms. Bachelor's thesis in Statistics. Uppsala Universitet.

Wikipedia. 2012. **McNemar's test**. [Online]. Available : https://en.wikipedia.org/wiki/McNemar%27s_test

Miyazaki, Y., Sato, Y. and Kamada, H. 2024. Logistic regression analysis and machine learning for predicting post-stroke gait independence. *Scientific Reports.* 14(1) : 721.

Nnaji, C. O. and Nwodo, D. U. 2022. Predicting customer churn in the telecommunication industry using machine learning algorithms: Performance comparison with logistic regression, random forest, and gradient boosting techniques. *Machine Learning*. 22(4) : 3-66

Nurpratiwi, I., Ak, S. and Yunisvita. 2020. Factors that influence wages differences in formal sector on male and female workers in Palembang City. *Theoretical and Applied Economics*. 27(1) : 147-158.

Pramod, O. 2023. **Cross-validation**. [Online]. Available : https://medium.com/@ompramod9921/cross-validation-623620ff84c2

Rachel Draelos. 2019. **Measuring performance: The confusion matrix.** [Online]. Available : https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

Roy, R. and Kundu, A. 2020. An analysis of poverty among the informal workers of India. *Theoretical and Applied Economics.* 27(1) : 87-104.

SDGmove. 2566. **SDG insights: Poverty line and well-being.** [Online]. เข้าถึงได้จาก : https://www.sdgmove.com/2023/09/11/sdg-insights-poverty-line-and-well-being/

Shen, Z. 2020. **Machine learning 101: Cross-validation**. [Online]. Available : https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/

Sohnesen, T. P. and Stender, N. 2016. "Is random forest a superior methodology for predicting poverty ? an empirical assessment". **Policy Research Working Paper Series 7612**. The World Bank.

Stoltzfus, J. C. 2011. Logistic regression: A brief primer. *Academic Emergency Medicine*. 18(10) : 1099-1104.

ภาคผนวก

ตัวแปร	ความยากจนในผู้สูงอายุไทย: การเปลี่ยนแปลงและปัจจัยเสี่ยง	Poverty and Informality: A Restraining or Constructive Relationship	Factors that influence wages differences in formal sector on male and female workers in Palembang City	ภาวะความยากจนและคุณภาพชีวิตของครัวเรือนเกษตรกรในเขตจังหวัดเพชรบูรณ์	An Analysis of Poverty Among the Informal Workers of India	Factors Influencing the Income of Urban Informal Workers: Evidence from Nepal
AGE	/		/			/
SEX	/	/	/			
MARITAL	/					
EDU	/	/	/	/	/	/
OCCUP					/	
RELATION				/		
REG	/	/			/	
AREA	/	/			/	
MEMBERS	/			/		/
GINI_IDX						
HAI						
WAGE_TYPE						
INDUS						
COND					/	
WORK_PROB						
UNSAFE			/			
TOTAL_HR_MONTH			/			/
BONUS						

ตัวแปร	ความยากจนในผู้สูงอายุไทย: การเปลี่ยนแปลงและปัจจัยเสี่ยง	Poverty and Informality: A Restraining or Constructive Relationship	Factors that influence wages differences in formal sector on male and female workers in Palembang City	ภาวะความยากจนและคุณภาพชีวิตของครัวเรือนเกษตรกรในเขตจังหวัดเพชรบูรณ์	An Analysis of Poverty Among the Informal Workers of India	Factors Influencing the Income of Urban Informal Workers: Evidence from Nepal
OT						
OTH_THING						
W_PLACE		/	/		/	
REGISTER						
MIN_WAGE						

ตารางแสดงการจำแนกประเภทของแบบจำลอง Logistic Regression (Threshold = 0.5) ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	**0.7631**	0.7600
Accuracy	**0.7300**	0.6831
F1-Score	**0.3898**	0.3670
Recall	0.8053	**0.8578**
Precision	**0.2572**	0.2335

ตารางแสดงการจำแนกประเภทของแบบจำลอง Logistic Regression (Optimal Threshold) ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	0.7332	**0.7398**
Accuracy	**0.8338**	0.8260
F1-Score	**0.4382**	0.4369
Recall	0.6050	**0.6302**
Precision	**0.3435**	0.3343

ตารางแสดงการจำแนกประเภทของแบบจำลอง Random Forest ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	**0.7358**	0.7269
Accuracy	**0.8385**	0.8382
F1-Score	**0.4452**	0.4366
Recall	**0.6050**	0.5853
Precision	**0.3522**	0.3481

ตารางแสดงการจำแนกประเภทของแบบจำลอง XGBoost ชุดข้อมูลทดสอบ

เทคนิคการสุ่มตัวอย่าง	Random Undersampling	SMOTE-ENN
Balanced Accuracy	0.7491	**0.7663**
Accuracy	0.8778	**0.8827**
F1-Score	0.5064	**0.5303**
Recall	0.5853	**0.6182**
Precision	0.4462	**0.4643**

ประวัติผู้เขียน

ชื่อ	นางสาวชนม์นิภ ตันหยง
วัน เดือน ปีเกิด	4 ธันวาคม 2541
ที่อยู่ปัจจุบัน	แจ้งวัฒนะ 12 แขวงทุ่งสองห้อง เขตหลักสี่ กรุงเทพมหานคร
ประวัติการศึกษา	(2564) วิทยาศาสตรบัณฑิต สาขาสถิติ เกรดเฉลี่ย 3.33 (มหาวิทยาลัยธรรมศาสตร์)