5ST-O15: การทำนายระดับความยากจนจากข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่องจักร

Prediction of poverty level on census data using machine learning

ศรราม หงษ์พรหม[1*] และ จันตรี ผลประเสริฐ[1]

Sornram Hongprom[1*] and Chantri Polprasert[1]

บทคัดย่อ

งานวิจัยนี้นำเสนอการใช้การเรียนรู้ของเครื่องจักรในการวิเคราะห์ข้อมูลสำมะโนประชากร โดยนำเสนอการใช้กระบวนการปรับแต่งคุณลักษณะเฉพาะของข้อมูล (Feature Engineering) เพื่อสร้างคุณลักษณะเฉพาะของครัวเรือน เพื่อใช้ในการทำนายความยากจนของประชากร ซึ่งความยากจนถูกแบ่งออกเป็น 4 ระดับคือ ขั้นรุนแรง (Extreme Poverty), ปานกลาง (Moderate Poverty), มีความเสี่ยงจะยากจน (Vulnerable Households), ไม่มีความเสี่ยงจะยากจน (Non Vulnerable Households) โดยแบบจำลองการเรียนรู้ของเครื่องจักรที่งานวิจัยนี้นำมาใช้ในการทำนายความยากจนจากข้อมูลสำมะโนประชากรประกอบไปด้วย โครงข่ายประสาทเทียมแบบ Multilayer Perceptron ,การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear discriminant analysis) ,วิธีการเพื่อนบ้านใกล้ที่สุด (K nearest neighbor) ,โมเดลป่าสุ่ม (Random Forest) ,ต้นไม้ตัดสินใจจำนวนมาก (Extra Trees) จากการทดลองพบว่าแบบจำลองการเรียนรู้ของเครื่องจักรแบบป่าสุ่มมีประสิทธิภาพดีที่สุดในการทำนายความยากจนจากข้อมูลสำมะโนประชากรโดยให้ความแม่นยำ (Precision) เท่ากับ 0.43 , ความครบถ้วน (Recall) เท่ากับ 0.46 , และคะแนน F1 (macro F1) เฉลี่ยเท่ากับ 0.43 โดยจากการทดลองพบว่าเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE) มีส่วนสำคัญในการเพิ่มประสิทธิภาพของแบบจำลองในการระบุความยากจน โดยค่า F1 (macro F1) เพิ่มขึ้นจาก 0.31 เป็น 0.43 แบบจำลองที่นำเสนอนี้มีคุณสมบัติที่สำคัญที่สุดสามประการที่มีผลต่อประสิทธิภาพของแบบจำลองอันประกอบไปด้วยจำนวนปีในสถานศึกษา,อายุของประชากรและระดับของการศึกษาโดยมีค่าความสำคัญ (feature importance) เท่ากับ 0.057, 0.055 และ 0.053 ตามลำดับ

คำสำคัญ: การเรียนรู้ของเครื่องจักร การทำนายความยากจน การปรับแต่งคุณลักษณะเฉพาะของข้อมูล การสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Abstract

The purpose of this research is to present the utilization of machine learning for analysis of census data by proposing a feature engineering process to create household characteristics for predicting population poverty. Poverty is divided into four levels: Extreme Poverty, Moderate Poverty, Vulnerable Households, and Non Vulnerable Households. The machine learning models used in the poverty prediction from census data included Multilayer Perceptron, Linear Discriminant Analysis, K nearest neighbor, Random Forest and Extra Trees. The experimental results shows that Random Forest model yields the best performance for poverty prediction from census data, with precision equal to 0.43, recall equal to 0.46 and the average macro F1 score equal to 0.43 The experimental results also revealed that SMOTE plays a significant role in the optimization of the model in poverty identification the value of macro F1 score increased from 0.31 to 0.43. The models presented above possess three most important properties affecting the model's performance, including years in school, age of the population and the level of education with the feature importance was 0.057, 0.055 and 0.053, respectively.

Keywords: machine learning, predicting poverty, feature engineering, synthetic minority over-sampling technique

[1] สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

[1] Department of Computer Science, Faculty of Science, Srinakharinwirot University

[*] Corresponding author. E-mail: sornram.hong@g.swu.ac.th

บทนำ

ความยากจน หมายถึง สภาพที่ประชาชนมีความเป็นอยู่ที่ต่ำกว่ามาตรฐาน กล่าวคือ ไม่มีรายได้เพียงพอที่จะใช้จ่ายในการซื้อสิ่งจำเป็นขั้นพื้นฐานในการครองชีพ ความยากจนจึงนับเป็นปัญหาที่ทุกประเทศทั่วโลกต่างให้ความสำคัญ เนื่องจากความยากจนเป็นสาเหตุอย่างหนึ่งที่ทำให้มีโอกาสที่จะก่อให้เกิดปัญหาอื่นๆ ตามมา (สำนักงานราชบัณฑิตยสภา, 2552) เพราะฉะนั้นการแก้ปัญหาความยากจนจึงเป็นสิ่งสำคัญ ดังนั้นจึงต้องหาแนวทางในการแก้ไขปัญหา แต่สิ่งที่สำคัญคือต้องสามารถระบุระดับความยากจนของแต่ละบุคคลได้และมีสิ่งใดบ้างเป็นปัจจัยที่ส่งผลกระทบต่อความยากจนเพื่อนำมาเป็นแนวทางในการแก้ปัญหาต่อไป จากการศึกษางานวิจัยที่เกี่ยวข้องเรื่อง Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification (Sani et al., 2018) และ Household poverty classification in data-scarce environments a machine learning approach (Kshirsagar et al., 2017) มีการนำข้อมูลของประชากรแต่ละประเทศมาทำการสร้างแบบจำลองการทำนายความยากจน

งานวิจัยนี้จึงพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคล ด้วยการนำเทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ในการวิเคราะห์ข้อมูลสำมะโนประชากรเพื่อระบุระดับความยากจน โดยนำชุดข้อมูลมาสร้างแบบจำลองและทำการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมทั้งหมด 5 แบบคือ โครงข่ายประสาทเทียมแบบ Multilayer Perceptron, การวิเคราะห์การจำแนกประเภทเชิงเส้น, การค้นหาเพื่อนบ้านใกล้สุด K อันดับ, การสุ่มป่าไม้, ต้นไม้ตัดสินใจจำนวนมาก จากนั้นทำการวัดประสิทธิภาพเพื่อหาแบบจำลองที่ดีที่สุด รวมทั้งคุณลักษณะสำคัญที่ส่งผลต่อการทำนายระดับความยากจน

วัตถุประสงค์

1. เพื่อศึกษาการวิเคราะห์ข้อมูลและคุณลักษณะของข้อมูลสำมะโนประชากร
2. เพื่อศึกษาและประยุกต์การใช้อัลกอริทึมการเรียนรู้ของเครื่องมาสร้างแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคล
3. เพื่อศึกษาวิธีแก้ปัญหาการไม่สมดุลกันของข้อมูลสำมะโนประชากร

วิธีการศึกษา

จุดประสงค์ของงานวิจัยนี้คือ การพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคล โดยทำการศึกษาและทดลองจากข้อมูลสำมโนประชากรของประเทศคอสตาริกา ซึ่งเป็นข้อมูลของ Inter-American Development Bank ในงานวิจัยนี้ใช้เครื่องมือจาก colab.research.google.com ซึ่งเป็นเครื่องมือที่ใช้สำหรับงานการเรียนรู้ของเครื่อง (Machine Learning) และการวิเคราะห์ข้อมูล(Data Analysis) โดยใช้ภาษา Python และชุดคำสั่งจาก Scikit-learn โดยมีขั้นตอนแบ่งเป็น 2 ส่วน ส่วนที่หนึ่งแบ่งข้อมูลเป็น 2 ชุดคือ ข้อมูลฝึก(Train) และข้อมูลทดสอบ(Test) จากนั้นการทำความเข้าใจข้อมูลและเตรียมข้อมูล โดยนำข้อมูลมาทำความสะอาดข้อมูล วิเคราะห์ข้อมูลเชิงสำรวจ ทำวิศวกรรมข้อมูลและเลือกคุณลักษณะ เพื่อให้ได้ชุดข้อมูลใหม่ ดังในตัวอย่าง Figure 1

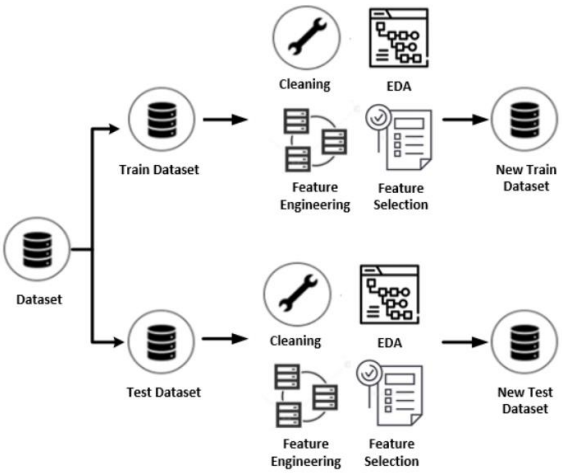

Figure 1 Steps 1 understanding data, 2 data preparation

ส่วนที่สองประเมินอัลกอริทึมและสร้างแบบจำลองรวมทั้งวัดประสิทธิภาพ โดยนำชุดข้อมูลฝึก (Train) ที่ได้ในขั้นตอนก่อนหน้านี้มาทำการประเมินอัลกอริทึมเพื่อเลือกอัลกอริทึมที่เหมาะสมและสร้างแบบจำลองจากนั้นวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Test) ดังในตัวอย่าง Figure 2

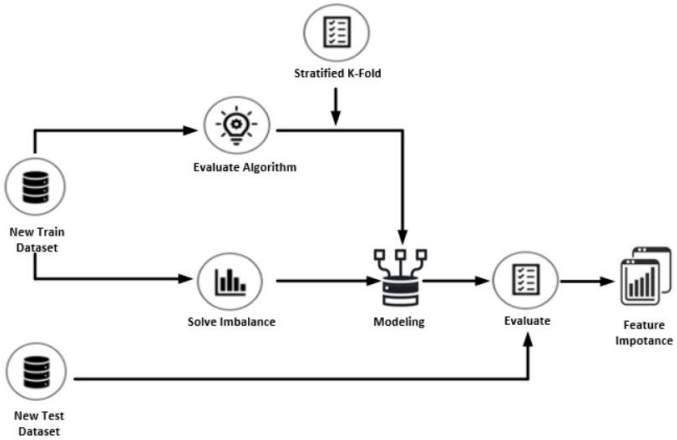

Figure 2 Step 3 evaluation of algorithm, step 4 modeling and evaluation.

1. การทำความเข้าใจข้อมูล (Understanding data)

1.1 ชุดข้อมูล

1.1.1 จำนวนข้อมูล 9,557 แถว แสดงถึงข้อมูลของแต่ละบุคคลที่ไม่ซ้ำกัน

1.1.2 จำนวนคอลัมน์ 143 คอลัมน์ แสดงถึงข้อมูลที่บ่งบอกถึงคุณลักษณะเฉพาะของแต่ละบุคคล ประกอบไปด้วย จำนวนเต็ม (Integer) 130 คอลัมน์ จำนวนจริง (Float) 8 คอลัมน์ และกลุ่มวัตถุ (Object) 5 คอลัมน์ ดังในตัวอย่าง Figure 3

	Id	v2a1	hacdor	rooms	hacapo	v14a	refrig	v18q	v18q1	r4h1	...	SQBescolari	SQBage	SQBhogar_total	SQBedjefe	SQBhogar_nin
0	ID_279628684	190000.0	0	3	0	1	1	0	NaN	0	...	100	1849	1	100	0
1	ID_f29eb3ddd	135000.0	0	4	0	1	1	1	1.0	0	...	144	4489	1	144	0
2	ID_68de51c94	NaN	0	8	0	1	1	0	NaN	0	...	121	8464	1	0	0
3	ID_d671db89c	180000.0	0	5	0	1	1	1	1.0	0	...	81	289	16	121	4
4	ID_d56d6f5f5	180000.0	0	5	0	1	1	1	1.0	0	...	121	1369	16	121	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9552	ID_d45ae367d	80000.0	0	6	0	1	1	0	NaN	0	...	81	2116	25	81	1
9553	ID_c94744e07	80000.0	0	6	0	1	1	0	NaN	0	...	0	4	25	81	1
9554	ID_85fc658f8	80000.0	0	6	0	1	1	0	NaN	0	...	25	2500	25	81	1
9555	ID_ced540c61	80000.0	0	6	0	1	1	0	NaN	0	...	121	676	25	81	1
9556	ID_a38c64491	80000.0	0	6	0	1	1	0	NaN	0	...	64	441	25	81	1

9557 rows × 143 columns

Figure 3 Example data set

1.2 คอลัมน์ Target

คือ ตัวแปรจำนวนเต็ม (Int64) ซึ่งจะถูกกำหนดให้เป็นผลลัพธ์ในการทำนายประกอบไปด้วย 4 กลุ่มดังนี้

1 บุคคลที่มีความยากจนขั้นรุนแรง (Extreme poverty) 755 แถว

2 บุคคลที่มีความยากจนปานกลาง (Moderate poverty) 1,597 แถว

3 บุคคลที่มีความเสี่ยงจะยากจน (Vulnerable households) 1,209 แถว

4 บุคคลที่ไม่มีความเสี่ยงจะยากจน (Non vulnerable households) 5,996 แถว

2. การเตรียมข้อมูล (Data Preparation)

นำชุดข้อมูลมาทำการแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train และ Test ในอัตราส่วน 70 : 30 เพื่อแยกในการทำขั้นตอนเตรียมข้อมูล โดย Train คือชุดข้อมูลการฝึกสำหรับสร้างแบบจำลอง และ Test คือชุดข้อมูลการทดสอบสำหรับการวัดประสิทธิภาพของแบบจำลอง

2.1 ทำความสะอาดข้อมูล (Cleansing Data)

การทำความสะอาดข้อมูล เป็นกระบวนการตรวจสอบและแก้ไขรายการข้อมูลที่ไม่ถูกต้องหรือไม่มีข้อมูลซึ่งไม่สามารถนำไปใช้สร้างแบบจำลองได้(ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562) โดยวิธีการแก้ไขมีดังนี้

2.1.1 หาค่าที่ขาดหายไปในข้อมูล (Data Missing)

ค่าที่ขาดหายไปจะต้องจัดการให้เรียบร้อยก่อนที่เราจะนำไปพัฒนาแบบจำลอง โดยใช้วิธีการเปรียบเทียบข้อมูลที่หายไปกับข้อมูลที่เกี่ยวข้อง แล้วทำการแทนค่าหรือในกรณีที่ไม่สามารถแทนค่าได้ก็จะทำการลบข้อมูลบุคคลนั้นออกไป

2.1.2 จำแนกคอลัมน์วัตถุ (Explain Column Object)

ชุดข้อมูลที่นำมาใช้ในการพัฒนาควรจะเป็นประเภทจำนวนเต็ม (Integer) และจำนวนจริง (Float) เพื่อให้สามารถนำไปคำนวณร่วมกับอัลกอริทึมได้ จึงต้องจำแนกข้อมูลที่ไม่สอดคล้องกับข้อมูลอื่นในแต่ละคอลัมน์ โดยใช้วิธีการเปรียบเทียบข้อมูลกับความหมายของแต่ละคอลัมน์ (Data description) เพื่อปรับปรุงข้อมูลให้ถูกต้อง

2.2 การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)

2.2.1 การกำหนดประเภทของกลุ่มคอลัมน์ (Define Column Categories) โดยแบ่งเป็น 4 กลุ่ม

1. กลุ่มข้อมูลระบุตัวตนซึ่งเป็นข้อมูลที่ระบุการไม่ซ้ำกันของแต่ละบุคคล เป็นข้อมูลสำคัญไม่ควรลบหรือแก้ไข

2. กลุ่มข้อมูลของแต่ละบุคคลเป็นข้อมูลส่วนตัวของแต่ละคน สามารถนำมาวิเคราะห์ได้

3. กลุ่มข้อมูลเกี่ยวกับครอบครัว เป็นข้อมูลเกี่ยวกับครอบครัวนั้นๆ ซึ่งบุคคลในครอบครัวเดียวกันจะมีค่าเท่ากัน สามารถนำมาวิเคราะห์ได้

4. กลุ่มข้อมูลยกกำลังสองซึ่งเป็นข้อมูลต่างๆ ที่อยู่ในชุดข้อมูลนี้ถูกนำมายกกำลังสอง

2.2.2 ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient)

เป็นการศึกษาความสัมพันธ์ของคอลัมน์ตั้งแต่ 2 คอลัมน์ขึ้นไป ว่ามีความคล้ายคลึงกันของข้อมูลมากน้อยเพียงใด(ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562) โดยจะนำกลุ่มคอลัมน์ที่ทำการแบ่งในขั้นตอนก่อนหน้านี้มาเปรียบเทียบกันแต่ละกลุ่ม เพื่อหาค่า Correlation Coefficient และจะทำการลบคอลัมน์ที่มีค่า Correlation Coefficient สูงมากกว่า 0.95 เพื่อลดขนาดของชุดข้อมูล

2.2.3 คอลัมน์ที่ซ้ำซ้อน (Duplicate Column)

ทำการจัดกลุ่มของชุดข้อมูลที่มีความหมายหรือคำอธิบายของคอลัมน์ (Data descriptions) เกี่ยวข้องหรือใกล้เคียงกันจะเป็นการลดขนาดคอลัมน์ของชุดข้อมูลเพื่อเพิ่มความเร็วในการประมวลผล โดยจะเหลือคอลัมน์ทั้งชุดข้อมูล Train และ Test อย่างละ 57 คอลัมน์

2.3 การทำวิศวกรรมข้อมูล (Feature Engineering)

ข้อมูลแต่ละบุคคลที่อยู่ในครอบครัวเดียวกันนั้นมีข้อมูลลักษณะคล้ายคลึงกันจึงทำให้เกิดการแปรปวนของข้อมูล (Bias-Variance) ดังนั้นจะทำการจัดกลุ่มชุดข้อมูล โดยนำสมาชิกแต่ละครอบครัวมาทำการรวมข้อมูลและหาค่าเฉลี่ย (Mean) , ค่าสูงสุด (Max) , ค่าต่ำสุด (Min) เพื่อสร้างชุดข้อมูลโดยแบ่งเป็น Train มีข้อมูล 2,074 แถว 198 คอลัมน์ และ Test มีข้อมูล 889 แถว 198 คอลัมน์ ดังในตัวอย่าง Figure 4

	v2a1			hacdor			rooms			hacapo			v14a			refrig			v18q			v18q1			r4h1		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max
idhogar																											
001ff74ca	0.0	0.0	0.0	0	0	0	6	6	6	0	0	0	1	1	1	1	1	1	1	1	1	1.0	1.0	1.0	0	0	0
003123ec2	0.0	0.0	0.0	0	0	0	3	3	3	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	2	2	2
004616164	0.0	0.0	0.0	0	0	0	4	4	4	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	0	0	0
004983866	0.0	0.0	0.0	0	0	0	5	5	5	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	0	0	0
005905417	0.0	0.0	0.0	0	0	0	8	8	8	0	0	0	1	1	1	0	0	0	0	0	0	0.0	0.0	0.0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
ff9343a35	0.0	0.0	0.0	0	0	0	5	5	5	0	0	0	1	1	1	1	1	1	1	1	1	1.0	1.0	1.0	0	0	0
ff9d5ab17	150000.0	150000.0	150000.0	0	0	0	5	5	5	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	0	0	0
ffae4a097	0.0	0.0	0.0	0	0	0	3	3	3	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	0	0	0
ffe90d46f	0.0	0.0	0.0	0	0	0	4	4	4	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	1	1	1
fff7d6be1	0.0	0.0	0.0	0	0	0	4	4	4	0	0	0	1	1	1	1	1	1	0	0	0	0.0	0.0	0.0	0	0	0

Figure 4 Amount of data after operation Feature Engineering

2.4 การเลือกคุณลักษณะ (Feature Selection)

โดยใช้วิธีการทดสอบไคสแควร์ (Chi-Square Score) เป็นการเปรียบเทียบตัวแปร 2 กลุ่มหรือมากกว่า 2 กลุ่มว่ามีความสัมพันธ์กันหรือไม่(ภรัณยา และคณะ, 2552) ซึ่งเลือกจากข้อมูลทั้งหมด 197 คอลัมน์ โดยแยกคอลัมน์ Target ออกมาเพราะเป็นคอลัมน์ที่ไว้สำหรับเป็นผลลัพธ์ในการทำทาย กำหนด K = 35 คือเลือกคุณลักษณะ (Feature) ที่มีคะแนนสูงสุด 35 อันดับแรก เมื่อนำไปรวมกับคุณสมบัติ Target จะเท่ากับ 36 คอลัมน์

3. การประเมินอัลกอริทึมสำหรับการสร้างแบบจำลอง (Evaluation of Algorithm)

จากขั้นตอนการเตรียมข้อมูล (Data Preparation) จะได้ชุดข้อมูล Train มีข้อมูล 2074 แถว 36 คอลัมน์ และ Test มีข้อมูล 889 แถว 36 คอลัมน์ จากนั้นจะทำการประเมินอัลกอริทึม ด้วยวิธีการนำข้อมูล Train มาสร้างแบบจำลอง โดยเลือกใช้อัลกอริทึมการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning) ทั้งหมด 5 อัลกอริทึม

ได้แก่ โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP), การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear Discriminant Analysis), การค้นหาเพื่อนบ้านใกล้สุด K อันดับ (K-Nearest Neighbors) กำหนดให้ K = 5 , การสุ่มป่าไม้ (Random Forest) , ต้นไม้ตัดสินใจจำนวนมาก (Extra Trees Classifier) โดยใช้ค่าตัวแปรพารามิเตอร์แบบค่าเริ่มต้นจากชุดคำสั่ง Scikit-learn ซึ่งจากการศึกษางานวิจัยที่ต่างๆ อัลกอริทึมดังกล่าวถูกนำมาใช้สร้างแบบจำลองการทำนายและสามารถทำนายได้ประสิทธิภาพที่แม่นยำ

จากนั้นทำการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึม ด้วยวิธีการตรวจสอบแบบไขว้โดยแบ่งข้อมูลเป็น 10 กลุ่ม (Stratified 10-Fold Cross Validation) โดยใช้คะแนน Accuracy ในการวัดประสิทธิภาพ

4.การสร้างแบบจำลองและวัดประสิทธิภาพ (Modeling and Evaluation)

Train คือชุดข้อมูลการฝึกสำหรับสร้างแบบจำลอง และ Test คือชุดข้อมูลการทดสอบสำหรับการวัดประสิทธิภาพของแบบจำลอง โดยมีการแบ่งข้อมูลได้ดังนี้

ข้อมูล Train จะมีข้อมูล 2,074 แถว 36 คอลัมน์โดยแบ่งเป็น

1 บุคคลที่มีความยากจนขั้นรุนแรง (Extreme poverty) 152 แถว

2 บุคคลที่มีความยากจนปานกลาง (Moderate poverty) 302 แถว

3 บุคคลที่มีความเสี่ยงจะยากจน (Vulnerable households) 257 แถว

4 บุคคลที่ไม่มีความเสี่ยงจะยากจน (Non vulnerable households) 1,363 แถว

ข้อมูล Test จะมีข้อมูล 889 แถว 36 คอลัมน์โดยแบ่งเป็น

1 บุคคลที่มีความยากจนขั้นรุนแรง (Extreme poverty) 65 แถว

2 บุคคลที่มีความยากจนปานกลาง (Moderate poverty) 129 แถว

3 บุคคลที่มีความเสี่ยงจะยากจน (Vulnerable households) 111 แถว

4 บุคคลที่ไม่มีความเสี่ยงจะยากจน (Non vulnerable households) 854 แถว

4.1 แก้ไขปัญหาของข้อมูลที่ไม่สมดุล (Imbalance Dataset)

จากชุดข้อมูลการฝึก(Train) พบว่าข้อมูลของงานวิจัยนี้เป็นข้อมูลที่ไม่สมดุล (Imbalance dataset) คือคุณลักษณะระบุระดับความยากจน (Target) ทั้ง 4 ระดับมีอัตราส่วนอยู่ที่ 2:4:3:16 โดยมีจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกันมาก เมื่อมีการสร้างแบบจำลองและการทำนายทำให้ผลลัพธ์การทำนายข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก (ปณตทรง, 2553) ดังนั้นจะต้องทำการแก้ไขปัญหาของข้อมูลที่ไม่สมดุลกัน

ซึ่งจะใช้วิธีนำข้อมูล Train มาทำการปรับเพิ่มหรือปรับลดข้อมูลในแต่ละกลุ่มให้มีอัตราส่วนของคุณลักษณะระบุระดับความยากจน (Target) มีจำนวนใกล้เคียงกัน โดยแบ่งชุดข้อมูล Train เป็นทั้ง 5 ชุดคือ ข้อมูลที่ไม่มีการปรับคือข้อมูลเดิม (Existing Data) , ข้อมูลที่มีการปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค SMOTE , ADASYN และข้อมูลที่มีการปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder , Cluster Centroids โดยทั้ง 5 ชุดข้อมูลใช้ชุดข้อมูลตั้งต้นเดียวกันคือชุดข้อมูล Train และค่าตัวแปรพารามิเตอร์แบบเริ่มต้นจากชุดคำสั่ง Scikit-learn เพื่อให้เป็นมาตรฐานเดียวกัน โดยผลลัพธ์ที่ได้จำนวนชุดข้อมูลดัง Table 1

Table 1 Amount of data for each technique after improve dataset

	Existing	SMOTE	ADASYN	RandomUnder	ClusterCentroids
Extreme poverty (1)	152	1,363	1,412	152	152
Moderate poverty (2)	302	1,363	1,479	152	152
Vulnerable households (3)	257	1,363	1,319	152	152
Non vulnerable households (4)	1,363	1,363	1,363	152	152

4.2 การปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter) ด้วยเทคนิค Grid Search

ก่อนที่จะนำชุดข้อมูลทั้ง 5 แบบมาสร้างแบบจำลองนั้น ควรจะมีการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter) เพื่อให้แต่ละพารามิเตอร์เหมาะสมกับแต่ละแบบจำลอง ด้วยเทคนิค Grid Search ร่วมกับอัลกอริทึม Random Forest Classifier ซึ่งเป็นอัลกอริทึมประสิทธิภาพดีที่สุดและกำหนดค่า F1 Macro เป็นค่าวัดประสิทธิภาพ โดยเลือกใช้ค่า Hyper-Parameter ดัง Table 2

Table 2 Hyper-Parameter of the model Random Forest Classifier

Parameter	Value
bootstrap	True, False
max_depth	100, 110, 130, 150
max_features	3, 4, 5, 6
min_samples_leaf	3, 4, 5, 6
min_samples_split	10, 12, 14,16
n_estimators	100, 200, 300, 1000

4.3 การสร้างแบบจำลองและวัดประสิทธิภาพ (Modeling and Evaluation)

เมื่อทำการปรับปรุงพารามิเตอร์แล้ว จึงนำข้อมูลทั้ง 5 แบบ มาทำการฝึกเพื่อสร้างแบบจำลองร่วมกับอัลกอริทึม Random Forest Classification และกำหนดพารามิเตอร์ตามผลลัพธ์ที่ได้ในขั้นตอนก่อนหน้านี้ จากนั้นทำการวัดประสิทธิภาพเพื่อหาแบบจำลองที่มีประสิทธิภาพดีที่สุด โดยการนำข้อมูล Test มาทำการทดสอบกับแต่ละแบบจำลอง และใช้ตัวชี้วัดเพื่อบ่งบอกถึงประสิทธิภาพในด้านต่างๆ ได้แก่ ค่าความครบถ้วน(Recall), ค่าความแม่นยำ (Precision), ค่าประสิทธิภาพโดยรวม (macro F1)

4.4 คุณลักษณะสำคัญที่มีผลต่อการทำนาย

ขั้นตอนสุดท้ายเลือกแบบจำลองที่มีประสิทธิภาพดีที่สุด นำมาหาคุณลักษณะที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคลด้วยเทคนิค Feature Importance คือการตรวจสอบแต่ละคอลัมน์ที่นำมาใช้สร้างแบบจำลองว่าคอลัมน์ใดถูกนำไปใช้ในอัตราส่วนเท่าใด โดยจะทำการเลือกคอลัมน์ที่มีอัตราส่วนมากที่สุด 3 อันดับแรกและการแสดงแผนภาพของข้อมูลในแต่ละคุณลักษณะโดยเลือกคุณลักษณะที่สำคัญที่สุด 3 อันดับแรก เพื่อวิเคราะห์ดูความแตกต่าง

ผลการศึกษา

1. ผลลัพธ์การประเมินอัลกอริทึมสำหรับการสร้างแบบจำลอง

การประเมินแต่ละอัลกอริทึม ผลการทดลองที่ได้ดัง Table 3 คืออัลกอริทึม Random Forest Classifier มีคะแนน Accuracy เฉลี่ยสูงสุดคือ 0.3745

Table 3 Results of each algorithm using Stratified 10-Fold Cross Validation

Algorithm	Average of Accuracy Score
Multilayer Perceptron (MLP)	0.3418
Linear Discriminant Analysis (LDA)	0.3338
K-Nearest Neighbors K=5 (KNN)	0.3484
Random Forest Classifier (RF)	0.3745
Extra Trees Classifier (EXT)	0.3673

2. ผลลัพธ์ของการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter)

จากการใช้เทคนิค Grid Search มาทำการปรับปรุงพารามิเตอร์ของแบบจำลอง Random Forest Classifier โดยการเลือกพารามิเตอร์ที่ได้ค่า F1 Macro ดีที่สุด ซึ่งผลลัพธ์ที่ได้ทั้ง 5 แบบดัง Table 4

Table 4 Results of Hyper-Parameter using Grid Search

Parameter	Existing	SMOTE	ADASYN	RandomUnder	ClusterCentroids
bootstrap	True	False	False	True	True
max_depth	80	100	150	100	130
max_features	2	4	4	3	3
min_samples_leaf	5	3	3	4	3
min_samples_split	8	10	10	12	14
n_estimators	400	100	400	100	100

3. ผลลัพธ์การวัดประสิทธิภาพของแบบจำลอง

ผลการวัดประสิทธิภาพของแบบจำลองทั้ง 5 แบบ ซึ่งใช้ตัวชี้วัดเพื่อบ่งบอกถึงประสิทธิภาพในด้านต่างๆ คือ ค่าความครบถ้วน(Recall), ค่าความแม่นยำ (Precision), ค่าประสิทธิภาพโดยรวม (macro F1) โดยในแต่ละตารางจะทำการแยกแสดงผลลัพธ์ในแต่ละกลุ่ม(Class) เพื่อเปรียบเทียบความแตกต่าง และทำการหาค่าเฉลี่ยทั้ง 4 กลุ่ม (Class) เพื่อสรุปประสิทธิภาพโดยรวมแต่ละแบบจำลอง โดยทั้ง 5 แบบจำลองได้ผลการทดลองดังนี้

3.1 ข้อมูลเดิม (Existing Data) ดัง Table 5

Table 5 Performance benchmarks of existing data

	Precision	Recall	macro F1
Class 1	0.38	0.09	0.15
Class 2	0.30	0.18	0.22
Class 3	0.21	0.03	0.05
Class 4	0.72	0.96	0.82
Average	0.40	0.31	0.31

3.2 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE ดัง Table 6

Table 6 Performance benchmarks of over-sampling using SMOTE

	Precision	Recall	macro F1
Class 1	0.31	0.40	0.35
Class 2	0.31	0.33	0.29
Class 3	0.29	0.30	0.64
Class 4	0.83	0.80	0.82
Average	0.43	0.46	0.43

3.3 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค ADASYN ดัง Table 7

Table 7 Performance benchmarks of over-sampling using ADASYN

	Precision	Recall	macro F1
Class 1	0.28	0.37	0.32
Class 2	0.28	0.27	0.27
Class 3	0.21	0.25	0.26
Class 4	0.83	0.81	0.81
Average	0.41	0.43	0.41

3.4 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค RandomUnder ดัง Table 8

Table 8 Performance benchmarks of under sampling using RandomUnder

	Precision	Recall	macro F1
Class 1	0.22	0.57	0.32
Class 2	0.30	0.33	0.11
Class 3	0.22	0.35	0.27
Class 4	0.89	0.62	0.72
Average	0.41	0.47	0.40

3.5 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค ClusterCentroids ดัง Table 9

Table 9 Performance benchmarks of under sampling using ClusterCentroids

	Precision	Recall	macro F1
Class 1	0.18	0.69	0.28
Class 2	0.17	0.19	0.18
Class 3	0.13	0.34	0.19
Class 4	0.93	0.33	0.48
Average	0.35	0.39	0.28

3.6 สรุปรวมผลการวัดประสิทธิภาพของแบบจำลองทั้ง 5 แบบ

โดยการหาค่าเฉลี่ยรวมของผลลัพธ์ตัวชี้วัดประสิทธิภาพในแต่ละแบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง โดยผลลัพธ์ที่ได้ดัง Table 10

Table 10 Summary Performance benchmarks of each technique

	Precision	Recall	macro F1
Existing Data	0.40	0.31	0.31
SMOTE	0.43	0.46	0.43
ADASYN	0.41	0.43	0.41
RandomUnder	0.41	0.47	0.40
ClusterCentroids	0.35	0.39	0.28

4. ผลลัพธ์ของคุณลักษณะสำคัญที่มีผลต่อการทำนาย

เลือกคุณลักษณะที่มีคะแนนสูง 3 อันดับแรกเพื่อระบุปัจจัยหรือสาเหตุที่ส่งผลกระทบต่อความยากจน ซึ่งได้ผลการทดลองดัง Table 11

Table 11 Result of features Importance

Feature	Importance
escolari-mean	0.057322
age-mean	0.055134
education-mean	0.053492

ผลการแสดงแผนภาพของข้อมูลในแต่ละคุณลักษณะโดยเลือกคุณลักษณะที่สำคัญที่สุด 3 อันดับแรก เพื่อวิเคราะห์ดูความแตกต่างระหว่างข้อมูลเดิม(Existing data)และข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE ผลการทดลองที่ได้ดัง Figure 5 , Figure 6 , Figure 7

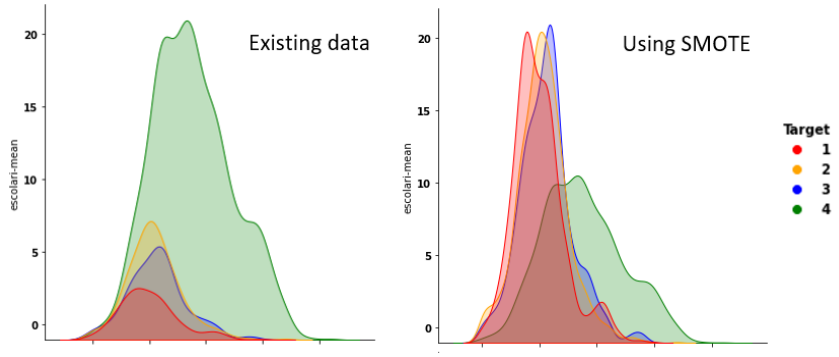

Figure 5 Compare features escolari-mean between existing data and SMOTE technique

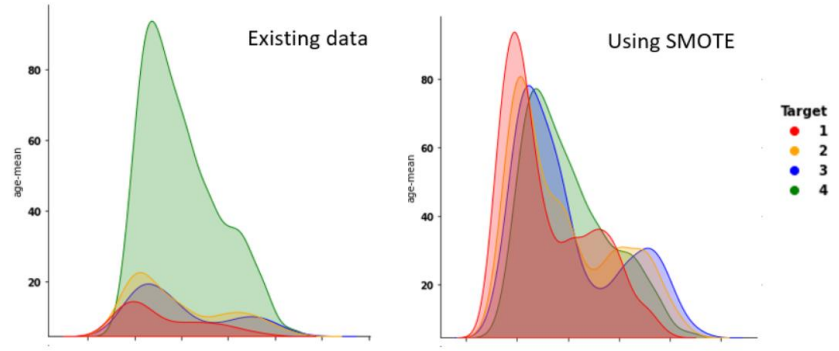

Figure 6 Compare features age-mean between existing data and SMOTE technique

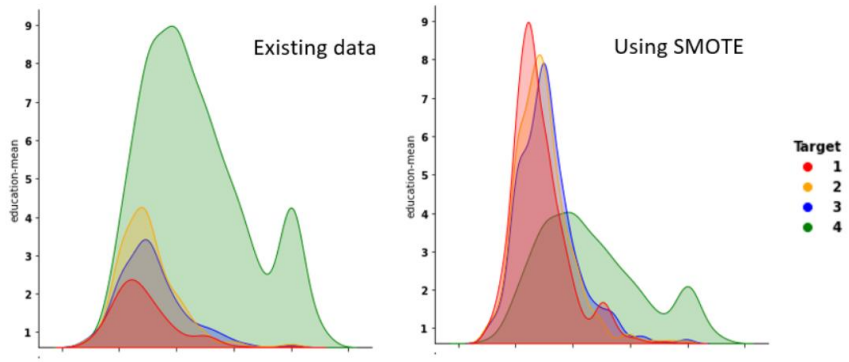

Figure 7 Compare features education-mean between existing data and SMOTE technique

อภิปรายผล

จาก Table 10 เมื่อทำการเปรียบเทียบผลการทดลองระหว่างชุดข้อมูลเดิม (Existing Data) กับชุดข้อมูลอื่นๆ จะพบว่าชุดข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE มีค่าเฉลี่ย macro F1 ดีที่สุดคือ 0.43 และเมื่อเปรียบเทียบผลการทดลองระหว่างชุดข้อมูลเดิมใน Table 5 และชุดข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE ในTable 6 เพื่อแยกดูผลการทดลองของแต่ละกลุ่ม พบว่าทั้งสองตารางในกลุ่มที่ 4 (Class 4) มีค่า macro F1 เท่ากัน ส่วนกลุ่มที่ 3 (Class 3) มีค่า macro F1 เพิ่มขึ้นเป็นอย่างมากจาก 0.05 เป็น 0.64 และ กลุ่มที่ 1 ,2 (Class 1,2) มีค่า macro F1 ดีขึ้นเล็กน้อย คือ กลุ่มที่ 1 เพิ่มขึ้นจาก 0.15 เป็น 0.35 และ กลุ่มที่ 2 เพิ่มขึ้นจาก 0.22 เป็น 0.29

จาก Figure 5 , Figure 6 , Figure 7 พบว่าเมื่อดูในส่วนของข้อมูลเดิม (Existing Data) ข้อมูลกลุ่มที่ 4 (สีเขียว) จะมีจำนวนมากกว่ากลุ่มอื่นๆ ทำให้ผลลัพธ์การทำนายข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก ซึ่งทำให้

ประสิทธิภาพในการทำนายไม่ดีเท่าที่ควรในตอนแรก แต่เพื่อเราทำการแก้ไขปัญหาข้อมูลที่ไม่สมดุลกัน (Imbalance Dataset) ด้วยวิธีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE จะทำให้ข้อมูลในกลุ่มที่ 1,2,3 (สีแดง,สีเหลือง,สีน้ำเงิน) เพิ่มขึ้นเป็นจำนวนใกล้เคียงกัน จึงทำให้ผลลัพธ์การทำนายระดับความยากจนของแต่ละบุคคลดีขึ้นไปด้วย

สรุป

1.สรุปผลการวิจัย

จากการทดลองพบว่าแบบจำลองการเรียนรู้ของเครื่องจักรแบบป่าสุ่ม (Random Forest Classifier) ที่พัฒนาร่วมกับข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE มีประสิทธิภาพดีที่สุดในการทำนายระดับความยากจนของแต่ละบุคคล โดยให้ความแม่นยำ (Precision) เท่ากับเฉลี่ย 0.43 , ความครบถ้วน (Recall) เฉลี่ยเท่ากับ 0.46 , และคะแนน F1 (macro F1) เฉลี่ยเท่ากับ 0.43 สรุปได้ว่าเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE) มีส่วนสำคัญในการเพิ่มประสิทธิภาพของกลุ่มข้อมูลกลุ่มที่ 1,2,3 บุคคลที่มีความยากจนขั้นรุนแรง ,บุคคลที่มีความยากจนปานกลาง , บุคคลที่มีความเสี่ยงจะยากจนตามลำดับ ซึ่งเป็นกลุ่มข้อมูลส่วนน้อยให้มีจำนวนเพิ่มขึ้นใกล้เคียงกับจำนวนกลุ่มข้อมูลส่วนมาก คือกลุ่มที่ 4 บุคคลที่ไม่มีความเสี่ยงจะยากจน โดยทำให้ ค่า F1 (macro F1) เพิ่มขึ้นจาก 0.31 เป็น 0.43 และคุณสมบัติที่สำคัญที่สุดสามประการที่มีผลต่อประสิทธิภาพของแบบจำลองประกอบไปด้วยจำนวนปีในสถานศึกษา (escolari) ,อายุของประชากร (age) และระดับของการศึกษาโดยมีค่าความสำคัญ (education) เท่ากับ 0.057, 0.055 และ 0.053 ตามลำดับ

2.ข้อเสนอแนะ

2.1 ในการวิจัยนี้ได้ใช้ข้อมูลในการศึกษาจาก www.kaggle.com เป็นชุดข้อมูลสำมโนประชากรของประเทศคอสตาริกา ซึ่งสามารถนำมาประยุกต์ใช้กับข้อมูลสำมโนประชากรของประเทศไทยได้

2.2 เนื่องจากชุดข้อมูลในการวิจัยนี้เป็นข้อมูลที่ไม่สมดุลกัน อาจจะใช้เทคนิคอื่นๆ นอกจากการปรับเพิ่มหรือปรับลดข้อมูล มาเป็นการปรับเพิ่มคุณลักษณะต่างๆ โดยใช้เทคนิค Feature Engineering เพื่อเพิ่มประสิทธิภาพของแบบจำลอง

คำขอบคุณ

ผู้วิจัยขอขอบคุณคณาจารย์ในสาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ให้การสนับสนุนคำปรึกษาและสถานที่ในการดำเนินการจัดทำวิจัยจนสำเร็จลุล่วงไปด้วยดี

เอกสารอ้างอิง

Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. arXiv preprint arXiv:1711.06813.

Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S., & Sarim, H. M. (2018). Machine learning approach for Bottom 40 Percent Households (B40) poverty classification. International Journal on Advanced Science, Engineering and Information Technology, 8(4-2), 1698-1705.

ปณตทรง วัฒนศิริ. (2553). เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม (ปริญญานิพนธ์ปริญญามหาบัณฑิต). จุฬาลงกรณ์มหาวิทยาลัย , กรุงเทพฯ.

ภรัณยา อำมฤครัตน์, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, & ณรงค์ โพธิ. (2552). A Comparative Efficiency of Feature Selection and Neural Network Classification. สืบค้น 10 เมษายน 2563, จาก http://www.nphothi.com/DrkLaNg/research/NCCIT2009-IT26-Neural.pdf

ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน. (2562). ความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning. สืบค้น 11 เมษายน 2563, จาก https://www.doe.go.th/prd/assets/upload/files/bkk_th /a323196dc548c204c85bc4a85b7bb46b.pdf

สำนักงานราชบัณฑิตยสภา. (2552). ความยากจน. สืบค้น 10 เมษายน 2563, จาก http://www.royin.go.th/?knowledges=%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1%E0%B8%A2%E0%B8%B2%E0%B8%81%E0%B8%88%E0%B8%99-%E0%B9%98-%E0%B9%80%E0%B8%A1%E0%B8%A9%E0%B8%B2%E0%B8%A2%E0%B8%99-%E0%B9%92%E0%B9%95%E0%B9%95%E0%B9%92