การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง

PREDICTION OF POVERTY LEVEL ON CENSUS DATA USING MACHINE LEARNING

ศรราม หงษ์พรหม

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง

ศราม หงษ์พรหม

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563

PREDICTION OF POVERTY LEVEL ON CENSUS DATA USING MACHINE LEARNING

SORNRAM HONGPROM

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Information Technology)

Faculty of Science, Srinakharinwirot University

2020

สารนิพนธ์

เรื่อง

การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง

ของ

ศรราม หงษ์พรหม

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... ประธาน

(อาจารย์ ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศุภชัย ไทยเจริญ)

ชื่อเรื่อง	การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง
ผู้วิจัย	ศรราม หงษ์พรหม
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. จันตรี ผลประเสริฐ

งานวิจัยนี้นำเสนอการใช้การเรียนรู้ของเครื่องจักรในการวิเคราะห์ข้อมูลสำมะโนประชากร โดยนำเสนอการใช้กระบวนการปรับแต่งคุณลักษณะเฉพาะของข้อมูล (Feature Engineering) เพื่อสร้างคุณลักษณะเฉพาะของครัวเรือน ร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique) เพื่อใช้ในการทำนายความยากจนของประชากร ซึ่งความยากจนถูกแบ่งออกเป็น 4 ระดับคือ ขั้นรุนแรง, ปานกลาง, มีความเสี่ยงจะยากจน, ไม่มีความเสี่ยงจะยากจน โดยโมเดลการเรียนรู้ของเครื่องจักรที่งานวิจัยนี้นำมาใช้ในการทำนายความยากจนจากข้อมูลสำมะโนประชากรประกอบไปด้วย โครงข่ายประสาทเทียมแบบ Multilayer Perceptron ,การวิเคราะห์การจำแนกประเภทเชิงเส้น ,วิธีการเพื่อนบ้านใกล้ที่สุด , โมเดลป่าสุ่ม ,ต้นไม้ตัดสินใจจำนวนมาก หลังจากที่ได้ทดลองปรับไฮเปอร์พารามิเตอร์ (hyperparameter) ของแต่ละแบบจำลองเพื่อเพิ่มประสิทธิภาพในการทำนายแล้ว จากการทดลองพบว่าโมเดลการเรียนรู้ของเครื่องจักรแบบป่าสุ่มมีประสิทธิภาพดีที่สุดในการทำนายความยากจนจากข้อมูลสำมะโนประชากรโดยให้ค่าความถูกต้อง (Accuracy) เท่ากับ 0.63 , ความแม่นยำ (Precision) เท่ากับ 0.43 , ความครบถ้วน (Recall) เท่ากับ 0.42 และคะแนน F1 (macro F1) เฉลี่ยเท่ากับ 0.43 โดยจากการทดลองพบว่าเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมีส่วนสำคัญในการเพิ่มประสิทธิภาพของโมเดลในการระบุความยากจน โมเดลที่นำเสนอนี้มีคุณสมบัติที่สำคัญที่สุดสามประการที่มีผลต่อประสิทธิภาพของแบบจำลองอันประกอบไปด้วยอายุของประชากร,จำนวนปีในสถานศึกษาและจำนวนปีการศึกษาโดยเฉลี่ยสำหรับผู้ใหญ่ โดยมีค่าความสำคัญ (Feature Importance) เท่ากับ 0.066, 0.065 และ 0.059 ตามลำดับ

คำสำคัญ : การเรียนรู้ของเครื่องจักร, การทำนายความยากจน, การปรับแต่งคุณลักษณะเฉพาะของข้อมูล, การสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Title	PREDICTION OF POVERTY LEVEL ON CENSUS DATA USING MACHINE LEARNING
Author	SORNRAM HONGPROM
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Assistant Professor Chantri Polprasert , Ph.D.

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 in conjunction with the Synthetic Minority Over-sampling Technique (SMOTE) 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 the census data, including Multilayer Perceptron, Linear Discriminant Analysis, K-nearest neighbor, Random Forest and Extra Trees. After adjusting the hyperparameters of each model to improve prediction efficiency, the experimental results showed that Random Forest model had the best performance for poverty prediction from census data, with accuracy equal to 0.63, precision equal to 0.43, recall equal to 0.42 and the average macro F1 score equal to 0.43. The experimental results also revealed that SMOTE played a significant role in the optimization of the model of poverty identification. The models presented above possess three most important properties affecting the performance of the model, including age of the population, years in school and average years of education for adults with the feature importance was 0.066, 0.065 and 0.059, respectively.

Keyword : Machine learning, Predicting poverty, Feature engineering, Synthetic Minority Over-sampling Technique

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.จันตรี ผลประเสริฐ อาจารย์ที่ปรึกษา ที่ได้ให้ความเมตตากรุณาเป็นที่ปรึกษาและให้ความช่วยเหลือชี้แนะแนวทางในสิ่งที่เป็นประโยชน์ต่อการศึกษาและการทำสารนิพนธ์นี้ด้วยความเอาใจใส่ตลอดมา ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ที่ได้ให้คำแนะนำและข้อเสนอแนะสำหรับการปรับปรุงสารนิพนธ์ และการใช้เทคนิคการเรียนรู้ของเครื่องเป็นเครื่องมือในการวิเคราะห์และแก้ไขปัญหาทางข้อมูลต่อไป

ขอกราบขอบพระคุณคณาจารย์และกรรมการบริหารหลักสูตรสาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒทุกท่าน ที่ได้กรุณาประสิทธิประสาทความรู้ต่างๆ ให้แก่ผู้วิจัย ตลอดจนให้ความช่วยเหลือในการทำวิจัยครั้งนี้

ศรราม หงษ์พรหม

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ความยากจน หมายถึง สภาพความเป็นอยู่ของประชาชนที่ต่ำกว่ามาตรฐานและมีรายรับไม่เพียงพอที่จะใช้จ่ายในชีวิตประจำวันที่จำเป็นขั้นพื้นฐาน เช่น เครื่องนุ่งห่ม อาหาร ที่อยู่อาศัย ความยากจนจึงนับเป็นปัญหาที่ทุกประเทศทั่วโลกต่างให้ความสำคัญ เนื่องจากความยากจนเป็นสาเหตุอย่างหนึ่งที่ทำให้มีโอกาสที่จะก่อให้เกิดปัญหาอื่น ๆ เช่น ปัญหาครอบครัว ปัญหาอาชญากรรม ปัญหาสิ่งเสพติด ปัญหาความเหลื่อมล้ำของสังคม ตลอดจนปัญหาคุณภาพชีวิตของประชากรและรวมถึงการพัฒนาประเทศ ดังนั้นการแก้ปัญหาความยากจนจึงเป็นสิ่งสำคัญเพราะความยากจนนั้นเป็นสาเหตุที่ทำให้เกิดปัญหาต่างๆ เราต้องหาแนวทางในการแก้ไขปัญหาความยากจน ซึ่งมีหลากหลายวิธีเช่น การปรับเพิ่มสวัสดิการหรือการพัฒนาศักยภาพเพื่อให้เกิดการพัฒนาแก่บุคคลที่ยากจน แต่สิ่งที่สำคัญก็คือต้องสามารถระบุตัวตนของแต่ละบุคคลได้ว่าบุคลคนนั้นยากจนจริงหรือไม่และมีสิ่งใดบ้างเป็นปัจจัยที่ส่งผลกระทบต่อความยากจนเพื่อเป็นแนวทางในการแก้ไขปัญหาความยากจนต่อไป (สำนักงานราชบัณฑิตยสภา, 2552)

งานวิจัยนี้จึงพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคลและหาปัจจัยที่ส่งผลกระทบต่อความยากจน ด้วยการนำเทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ในการวิเคราะห์ข้อมูลสำมะโนประชากรเพื่อระบุระดับความยากจนในแต่ละบุคคล ซึ่งจะแบ่งระดับความยากจนเป็น 4 ระดับคือ บุคคลยากจนขั้นรุนแรง (Extreme poverty) , บุคคลยากจนปานกลาง (Moderate poverty) ,บุคคลเสี่ยงจะยากจน (Vulnerable households) และบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) โดยนำเสนอขั้นตอน 2 ส่วนหลักๆ โดยขั้นตอนที่หนึ่ง คือการทำความเข้าใจปัญหา (Problem Understanding) และศึกษาทำความเข้าใจข้อมูล (Data Understanding) เพื่อเป็นการวางเป้าหมายในการทำงานถือเป็นขั้นตอนแรกและเป็นขั้นตอนที่สำคัญในการกำหนดแนวทางของการทำงาน จากนั้นทำการเตรียมข้อมูล (Data Preparation) โดยเริ่มจากการทำความสะอาดข้อมูลเพื่อแก้ไขข้อมูลที่หายไปหรือข้อมูลที่ไม่ถูกต้อง ต่อมานำข้อมูลมาทำการวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis) เพื่อทำความเข้าใจรายละเอียดภายในของข้อมูล รวมถึงการทำวิศวกรรมข้อมูล (Feature Engineering) เพื่อแก้ปัญหาการแปรปวนของข้อมูล (Bias-Variance) โดยจะทำการจัดกลุ่มชุดข้อมูล โดยนำสมาชิกแต่ละบุคคลที่อยู่ในครอบครัวเดียวกันทำการรวมข้อมูลและหาค่าเฉลี่ย (Mean) , ค่าสูงสุด (Max) , ค่าต่ำสุด (Min) เพื่อให้ได้ชุดข้อมูลใหม่ โดยขั้นตอนสุดของการเตรียมข้อมูลคือการเลือก

คุณลักษณะ(Feature Selection) เป็นการลดขนาดของจำนวนคุณลักษณะและคัดเลือกคุณลักษณะที่มีความสำคัญต่อการพัฒนาแบบจำลอง โดยเลือกคุณลักษณะ (Feature) ที่มีคะแนนสูงสุด 35 อันดับแรก หลังจากการผ่านขั้นตอนการเตรียมข้อมูล (Data Preparation) เรียบร้อยแล้วจะได้ชุดข้อมูลใหม่ที่สามารถนำไปใช้ในขั้นตอนต่อไป ขั้นตอนที่ 2 คือ การประเมินอัลกอริทึมและสร้างแบบจำลองรวมทั้งวัดประสิทธิภาพ โดยนำชุดข้อมูลที่ได้ในขั้นตอนที่ 1 มาแบ่งชุดข้อมูลออกเป็น 2 ชุดคือ ชุดข้อมูลสำหรับฝึก(Train) และชุดข้อมูลสำหรับทดสอบ(Test) ในอัตราส่วน 70:30 จากนั้นนำชุดข้อมูลฝึก(Train) มาทำการประเมินอัลกอริทึม โดยเลือกใช้วิธีการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning) มาใช้ในการสร้างแบบจำลองทั้ง 5 เทคนิค คือ 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) โดยแต่ละเทคนิคจะใช้ค่าตัวแปรพารามิเตอร์แบบเริ่มต้น (Default) จาก Scikit-learn และทำการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคด้วยวิธีการตรวจสอบแบบไขว้โดยแบ่งข้อมูลเป็น 10 กลุ่ม (Stratified K-Fold Cross Validation) เพื่อเลือกอัลกอริทึมที่มีประสิทธิภาพดีที่สุดเพื่อนำไปใช้ในการสร้างแบบจำลอง จากชุดข้อมูลฝึก (Train) พบว่าชุดข้อมูลของงานวิจัยนี้จะเป็นข้อมูลที่ไม่สมดุล (Imbalance dataset) คือคุณลักษณะระบุระดับความยากจน (Target) ทั้ง 4 ระดับมีอัตราส่วนอยู่ที่ 2:4:3:16 ดังนั้นเราจะทำการแก้ไขปัญหาของข้อมูลที่ไม่สมดุล (Imbalance dataset) โดยการนำเทคนิคการปรับเพิ่มข้อมูล (Over-Sampling) ในแต่ละกลุ่มให้เท่ากันด้วยเทคนิค SMOTE , ADASYN และการปรับลดข้อมูล (Under-Sampling) ในแต่ละกลุ่มให้เท่ากันด้วยเทคนิค RandomUnder , Cluster มาทำการปรับปรุงข้อมูลเพื่อให้ได้ชุดข้อมูลใหม่ เมื่อทำการปรับปรุงแล้วจะได้ชุดข้อมูลทั้งหมด 5 กลุ่ม คือ ข้อมูลเดิม (Existing Data) , ข้อมูลที่มีการปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค SMOTE , ADASYN และข้อมูลที่มีการปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder , Cluster เมื่อทำการปรับปรุงข้อมูลเรียบร้อยแล้ว นำข้อมูลในแต่ละกลุ่มทั้ง 5 กลุ่มมาทำการสร้างแบบจำลองเพื่อเป็นการฝึกการทำนายระดับความยากจนของแต่ละบุคคลร่วมกับอัลกอริมทึมที่ได้ทำการเลือกในขั้นตอนก่อนหน้า และทำการปรับปรุงไฮเปอร์พารามิเตอร์(Hyperparameter) ด้วยเทคนิค Grid Search เพื่อให้แต่ละพารามิเตอร์เหมาะสมกับแต่ละแบบจำลอง จากนั้นทำการวัดประสิทธิภาพของแบบจำลองทั้ง 5 แบบ โดยการนำชุดข้อมูล Test มาทำการทดสอบกับแต่ละแบบจำลองและการวัดประสิทธิภาพด้วย คอนฟูชั่นแมททริก (Confusion Matrix) โดยเลือกใช้ตัวชี้วัดต่างๆ ดังนี้ ค่า

ความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความครบถ้วน(Recall), และค่าประสิทธิภาพโดยรวม (F1 Score) จากนั้นทำการเปรียบเทียบผลการทดลองของแต่ละแบบจำลองทั้ง 5 แบบเพื่อหาแบบจำลองที่มีประสิทธิภาพดีที่สุด ขั้นตอนสุดท้ายเลือกแบบจำลองที่มีประสิทธิภาพดีที่สุด นำมาหาคุณลักษณะที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคลด้วยเทคนิค Feature Importance โดยเลือกคุณลักษณะที่สำคัญที่สุด 3 อันดับแรก

1.2 วัตถุประสงค์ของการวิจัย

1.เพื่อศึกษาการวิเคราะห์ข้อมูลและคุณลักษณะของข้อมูลสำมะโนประชากร

2.เพื่อศึกษาและประยุกต์การใช้เทคนิคการเรียนรู้ของเครื่องมาสร้างแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคล

3.เพื่อศึกษาวิธีแก้ปัญหาการไม่สมดุลกันของข้อมูลสำมะโนประชากร

4.เพื่อศึกษาและเปรียบเทียบกลุ่มอัลกอริทึมที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคล

1.3 ขอบเขตของการวิจัย

1.ชุดข้อมูลในการวิจัยมาจาก www.kaggle.com ซึ่งเป็นข้อมูลสำมโนประชากรของประเทศคอสตาริกา

2.ทำการเปรียบเทียบประสิทธิภาพการแก้ปัญหาการไม่สมดุลกันของข้อมูลระหว่างชุดข้อมูลเดิม (Existing Data) , การสุ่มลดจำนวนข้อมูลกลุ่มหลัก (Under-Sampling) และการสุ่มเพิ่มจำนวนข้อมูลกลุ่มน้อย (Over-Sampling)

3.แบ่งระดับความยากจนเป็น 4 ระดับคือ บุคคลที่มีความยากจนขั้นรุนแรง (Extreme Poverty), บุคคลที่มีความยากจนปานกลาง (Moderate Poverty), คือบุคคลที่มีความเสี่ยงจะยากจน (Vulnerable Households), คือบุคคลที่ไม่มีความเสี่ยงจะยากจน (Non Vulnerable Households)

4.ใช้แบบจำลองการเรียนรู้ของเครื่องในรูปแบบการจำแนกประเภท (Classification) สำหรับการทำนายระดับความยากจนของแต่ละบุคคลโดยใช้ทั้งหมด 5 เทคนิค ได้แก่ โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP), การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear Discriminant Analysis) , การค้นหาเพื่อนบ้านใกล้สุด K อันดับ (K-Nearest Neighbors), การสุ่มป่าไม้ (Random Forest) , ต้นไม้ตัดสินใจจำนวนมาก (Extra Trees Classifier)

5.สร้างแบบจำลองเพื่อทำนายระดับความยากจนของแต่ละครอบครัว

6.การประเมินผลแบบจำลองใช้ค่า Confusion matrix, Stratified K-Fold Cross Validation ในการวัดประสิทธิภาพของแต่ละแบบจำลอง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.ได้ศึกษาทฤษฎีและแนวคิดเกี่ยวกับการวิเคราห์ข้อมูลเชิงสำรวจ

2.ได้ศึกษาทฤษฎีและแนวคิดเกี่ยวกับการแก้ปัญหาการไม่สมดุลกันของข้อมูล

3.ได้ศึกษาทฤษฎีและแนวคิดเกี่ยวกับการเรียนรู้ของเครื่อง

4.ได้ศึกษาขั้นตอนและเทคนิคที่ใช้ในการพัฒนาแบบจำลองที่เหมาะสมกับข้อมูล

5.เพื่อเป็นแนวทางในการศึกษาปัจจัยที่ส่งผลกระทบต่อความยากจน

6.สามารถทำนายระดับความยากจนได้อย่างแม่นยาเพื่อนำไปเป็นแนวทางการแก้ไขปัญหาความยากจนได้อย่างมีประสิทธิภาพ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ทำการศึกษาค้นคว้าทฤษฎีและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่างๆ ดังต่อไปนี้

2.1 องค์ความรู้เรื่อง Data Science

2.2 องค์ความรู้เรื่อง Machine Learning

2.3 อัลกอริทึมการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning Algorithms)

2.4 การเลือกคุณลักษณะ (Feature Selection)

2.5 องค์ความรู้เรื่องกับข้อมูลที่ไม่สมดุล (Imbalanced Datasets)

2.6 องค์ความรู้เรื่องกับการประเมินแบบจำลอง

2.7 องค์ความรู้เรื่องการวัดค่าประสิทธิภาพของแบบจำลอง

2.8 เครื่องมือที่ใช้ในการทำวิจัย

2.9 งานวิจัยที่เกี่ยวข้อง

2.1 **องค์ความรู้เรื่องวิทยาศาสตร์ข้อมูล** (Data Science)

2.1.1 **ทำความเข้าใจปัญหาและความต้องการ** (Problem Understanding)

การวางเป้าหมายในการทำงานถือเป็นขั้นตอนแรก และเป็นขั้นตอนที่สำคัญในการกำหนดแนวทางของการทำงาน โดยเริ่มจากปัญหาที่พบให้ชัดเจน หรือข้อสงสัยต่างๆ ที่ทำให้ต้องหาทางแก้ไข การทำความเข้าในส่วนงานต่างๆขององค์กรว่ามีส่วนใดที่สามารถเข้ามาทำให้เกิดการพัฒนาด้วยระบบดิจิทัล เพื่อสามารถต่อยอดการทำงานได้ อาจต้องมีการเพิ่มการสังเกตการณ์ (Observation) ไม่ว่าจะเป็นการเข้าร่วมทำงานกับสภาพแวดล้อมการทำงานจริง หรือแม้แต่การทำตัวเป็นผู้ใช้บริการ และการให้ผู้ที่มีส่วนร่วมกับการใช้งานของข้อมูลทั้งหมด เข้ามาช่วยกำหนดความต้องการเพื่อแปลงความต้องการทั้งหมด ออกมาเป็นปัจจัยหรือตัวแปลหนึ่งในแบบจำลอง เพื่อให้สามารถทำแบบจำลองที่สามารถใช้งานได้จริงได้ในโครงการ (ศูนย์เทคโนโลยีสารสนเทศ กรมการจัดหางาน, 2562)

2.1.2 **ทำความเข้าใจข้อมูล** (Data Understanding)

การทำความเข้าใจข้อมูลเป็นขั้นตอนที่ 2 ของการสร้างแบบจำลองเพื่อวิเคราะห์ข้อมูล ซึ่งหากทราบเป้าหมายในการใช้ข้อมูล ก็ไม่สามารถทราบได้ว่าต้องใช้ข้อมูลอะไรบ้างการทำ

ความเข้าใจข้อมูล ไม่ใช่แค่การทำความสะอาดข้อมูล (Cleansing Data) แต่ต้องทำความเข้าใจว่าข้อมูลใดมีความสำคัญ ในบางกรณีการทำความเข้าใจอาจจะเริ่มจากการนำข้อมูลทั้งหมดที่มีมารวมกัน และหาความสัมพันธ์ระหว่างข้อมูลดังนั้น สิ่งที่สำคัญที่สุดในการทำความเข้าใจข้อมูล คือ ค่านิยามของข้อมูล (Heading) แล้วตามมาด้วยรายละเอียดภายใน หากเป็นข้อมูลประเภทที่มีโครงสร้าง (Structured Data) สามารถเข้าใจง่าย แต่หากเป็นข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) จะต้องมีการแปลงโครงสร้างให้ชัดเจน ขั้นตอนนี้ขึ้นอยู่ที่มุมมองและประสบการณ์ของผู้วิจัย ในการคัดเลือกข้อมูลมาใช้งานที่ตอบโจทย์ในแต่ละโจทย์ (ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562)

2.1.3 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเพื่อทำการวิเคราะห์ถือว่าเป็นขั้นตอนที่ใช้เวลานานที่สุดของการทำ อาจจะใช้เวลาประมาณ 75-80 % ของการทำโครงการทั้งหมด โดยหลักการแล้วนักวิจัยจะเป็นคนที่มีบทบาทในส่วนนี้มากที่สุด เหตุผลที่ขั้นตอนนี้ใช้เวลานานที่สุด เพราะข้อมูลที่มีขนาดใหญ่มาจากการรวมกันจากหลายแหล่งข้อมูล และเจ้าของข้อมูลไม่เหมือนกัน ทำให้เกิดความยุ่งยากเกิดขึ้น การเตรียมข้อมูล แบ่งออกเป็นการทำความสะอาดข้อมูล (Cleansing Data) และการจัดรูปแบบข้อมูลให้พร้อมใช้งานนักวิจัยมีหน้าที่ในการทำให้ข้อมูลอยู่ในรูปแบบที่สามารถนำไปต่อยอดต่อได้หากเป็นข้อมูลที่ไม่มีโครงสร้าง เปลี่ยนแปลงให้เป็นข้อมูลที่มีโครงสร้าง เช่น การทำ Normalization หรือการทำ Meaning Extraction เพื่อจัดทำให้เป็นข้อมูลที่ได้มีโครงสร้างตามต้องการเพื่อนำไปสร้างเป็นแบบจำลองในขั้นตอนต่อไป นอกจากนี้ยังมีหน้าที่ทำความสะอาดข้อมูล การตรวจสอบข้อมูลที่มีความผิดปกติอีกด้วย นักวิจัยมีหน้าที่ในการกรองข้อมูลเพื่อนำข้อมูลไปเป็นต้นแบบในการทำแบบจำลองเช่น การหาข้อมูลที่อยู่นอกความเป็นไปได้ของข้อมูลที่เกิดขึ้น (Detect Outlier) การจัดการกับข้อมูลที่ขาดหายไป (Missing Value) การเลือกขนาดของข้อมูล การเลือกช่วงเวลาที่จะนำข้อมูลไปวิเคราะห์เป็นแบบจำลอง เป็นต้น (ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562)

2.1.4 การสร้างแบบจำลอง (Modeling)

การสร้างรูปแบบความสัมพันธ์ (Relational Pattern) จะอยู่ในรูปของแบบจำลองที่ทำงานบนระบบคอมพิวเตอร์ (Computer Model) หรือสมการความสัมพันธ์ (Equation) ก็ได้ การสร้างแบบจำลองถือเป็นขั้นตอนที่สำคัญที่สุด โดยนักวิจัยที่มีประสบการณ์จะสามารถพิจารณาได้ว่าแต่ละโจทย์จะเลือกใช้อัลกอริทึม (Algorithm) และสามารถเลือกตัวแปรที่เหมาะสมทำให้ใช้เวลาน้อยในการสร้างแบบจำลองหนึ่ง แต่การสร้างแบบจำลองเพียงแบบจำลองเดียวอาจไม่ตอบโจทย์เพราะฉะนั้นอาจจะต้องมีการสร้างแบบจำลองหลายแบบจำลอง เพื่อเปรียบเทียบหา

แบบจำลองที่ดีที่สุด ดังนั้นนักวิจัยส่วนใหญ่จะทำการสร้างแบบจำลองเพิ่มหรือปรับปรุงไปเรื่อยๆ จนกว่าจะหมดเวลาหรือต้องส่งมอบงาน และการสร้างแบบจำลองนั้นจะต้องเข้าใจรายละเอียดของผลลัพธ์ด้วย (ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562)

2.1.5 การประเมินผลแบบจำลอง (Evaluation)

หลังจากการที่ได้แบบจำลองแล้ว ต้องทำการประเมินว่าแบบจำลองนั้นมีความแม่นยำมากหรือน้อยเพียงใด โดยการประเมินผลอาจจะเป็นทั้งในกรณีที่สามารถวัดออกมาเป็นค่าได้ หรือเรียกว่า การวัดเชิงปริมาณ เช่นการทำนายยอดขาย เป็นต้น และในกรณีที่ไม่สามารถวัดค่าได้ หรือกรณีที่ไม่มีข้อเปรียบเทียบระหว่างค่าที่แบบจำลองทำนายได้กับค่าจริง อาจจะต้องทำการทดสอบโดยการจำลองสถานการณ์จริง เช่น การนำแบบจำลองไปประมวลผลข้อมูลจริงที่มีอยู่ เพื่อทำการวิเคราะห์ผลและเปรียบเทียบความถูกต้องออกมาเป็นอัตราร้อยละ หรือเรียกว่า การทดลองในระบบเสมือน (Simulation) เพื่อให้แน่ใจว่าแบบจำลองสามารถนำไปใช้งานได้จริง (ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562)

2.1.6 การนำแบบจำลองไปใช้งานจริง (Deployment)

หลังจากที่ได้แบบจำลองที่มีคุณภาพและมีความแม่นยำตามที่ต้องการ ก็สามารถนำแบบทดลองนั้นมาประยุกต์ใช้กับงานจริง โดยอาจจะต้องมีการปรับปรุงเพื่อให้เหมาะสมกับสภาวะจริง เพื่อนำแบบจำลองมาสร้างเป็นผลิตภัณฑ์ (Product) ต่างๆ เช่น การสร้างระบบ (Application) ทำนายความมีโอกาสในการได้งานทำหรืออาจร่วมกับระบบอื่น เพื่อให้การใช้แบบจำลองมีประโยชน์อย่างสูงสุดและมีความยั่งยืนอาจต้องมีปรับปรุงแบบจำลองใหม่เมื่อผลลัพธ์ไม่เป็นไปตามที่คาดหวัง หรือมีการปรับปรุงแบบจำลองเป็นระยะๆ อยู่เสมอเพราะว่าข้อมูลที่อยู่ภายในระบบข้อมูลขนาดใหญ่ (Big Data) และเทคโนโลยีต่างๆ มีการเปลี่ยนแปลงอยู่ตลอดเวลา (ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน, 2562)

2.2 องค์ความรู้เรื่องการเรียนรู้ของเครื่อง (Machine Learning)

Machine learning เริ่มต้นมาตั้งแต่ปี ค.ศ.1950 เมื่อนักวิทยาศาสตร์คอมพิวเตอร์คิดหาวิธีสอนคอมพิวเตอร์ให้เล่นหมากฮอส จากนั้นเมื่อวิวัฒนาการทางเทคโนโลยีระบบการคำนวณค่าต่างๆของคอมพิวเตอร์เพิ่มขึ้น ทำให้คอมพิวเตอร์สามารถเข้าใจ และจดจำรูปแบบของค่าต่างๆ ที่ซับซ้อนได้แล้วจึงประยุกต์ไปสู่การคาดการณ์สถานการณ์รวมไปถึงการแก้ปัญหาด้วยตัวเอง Machine learning คือ ระบบที่สามารถเรียนรู้จากตัวอย่างด้วยตนเอง โดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ความก้าวหน้าในครั้งนี้มาพร้อมกับความคิดที่ว่าเครื่องคอมพิวเตอร์

สามารถเรียนรู้เพียงแค่จากข้อมูลอย่างเดียวเพื่อที่จะผลิตผลลัพธ์ที่แม่นยำออกมาได้โดยสามารถแบ่งออกได้ 3 รูปแบบ (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562)

2.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Machine Learning)

ข้อมูลที่ถูกนำไปใช้ฝึกนั้น (Training Data) ได้ถูกนำมาจัดหมวดหมู่แต่ละประเภทของผลลัพธ์ด้วยการติดป้ายกำกับ (Label) แล้วจึงนำข้อมูลที่ติดป้ายกำกับไปใช้ในการฝึกการเรียนรู้ของเครื่องที่ทำงานร่วมกับอัลกอริทึมสำหรับสร้างแบบจำลองที่ใช้สำหรับการทำนายผลลัพธ์ เมื่อได้แบบจำลองที่ผ่านการฝึกแล้วก็จะนำมาทดลองกับข้อมูลใหม่(New Data) เพื่อให้แบบจำลองทำนาย(Predictive Model) ข้อมูลว่าคำตอบควรจะเป็นอย่างไร (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562)

2.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Machine Learning)

เป็นเทคนิคการเรียนรู้ด้วยข้อมูลที่ไม่ถูกจัดหมวดหมู่ หรือเป็นข้อมูลที่ถูกติดป้ายกำกับ (Label) การเรียนรู้ด้วยเทคนิคนี้คือเครื่องจะทำการอนุมานว่าข้อมูลที่ได้รับและทำความเข้าใจถึงโครงสร้างของข้อมูล ดังนั้นเทคนิคนี้จึงจะไม่ได้หาผลลัพธ์ที่ถูกต้อง แต่จะใช้วิธีสำรวจข้อมูลโครงสร้างของข้อมูลและอนุมานว่าข้อมูลควรจะเป็นข้อมูลอะไร (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562)

2.2.3 การเรียนรู้ด้วยอัลกอริทึมแบบกึ่งควบคุม (Semi-supervised Machine Learning)

เป็นเทคนิคการผสมผสานกันระหว่างการเรียนรู้แบบมีผู้สอน (Supervised) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised) เนื่องจากเป็นการนำข้อมูลที่มีการติดป้ายกำกับและไม่มีป้ายกำกับมาทำการฝึกให้เครื่องเรียนรู้ ซึ่งเทคนิคนี้สามารถเพิ่มความแม่นยำสำหรับการเรียนรู้ของเครื่องได้อย่างมีประสิทธิภาพ ส่วนใหญ่จะถูกนำมาใช้ในเนื่องจากข้อมูลที่มีการติดป้ายกำกับนั้นไม่สามารถวิเคราะห์ได้ด้วยวิธีการฝึกแบบปกติทั่วไป จึงจะต้องใช้เทคนิคนี้ในการวิเคราะห์เพิ่มเติม (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562)

2.2.4 เทคนิคการเรียนรู้ของเครื่องแบบเสริมกำลัง (Reinforcement Machine Learning Algorithms)

เป็นเทคนิคการเรียนรู้โดยจะการกำหนดเป้าหมายให้แก่เครื่อง ที่เรียกว่า “Reinforcement Signal” เพื่อให้เครื่องสามารถสร้างแนวทางเลือกสำหรับการตัดสินใจได้มากกว่าหนึ่งรูปแบบตามสภาพแวดล้อมที่หลากหลาย จากนั้นเครื่องจะรวบรวมข้อมูลการตัดสินใจในแต่ละแนวทางเลือกเพื่อเรียนรู้และฝึกการทำนายให้ได้ผลลัพธ์ ซึ่งวิธีนี้ช่วยให้เครื่องสามารถค้นหา

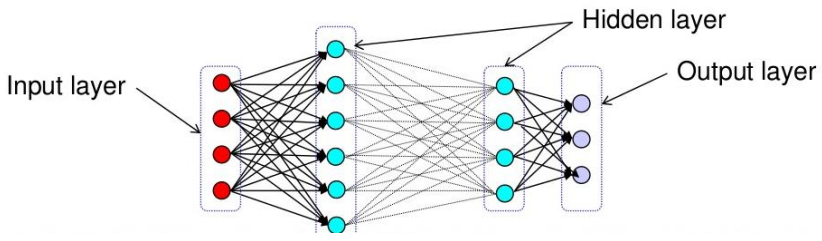
และทดสอบแนวทางเลือกที่ได้อย่างมีประสิทธิภาพ เพื่อให้บรรลุเป้าหมายโดยอัตโนมัติ (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562)

2.3 **อัลกอริทึมการเรียนรู้แบบมีผู้สอน** (Supervised Machine Learning Algorithms)

เทคนิคที่นำมาใช้ในการทดลองพัฒนาแบบจำลองมี ทั้งหมด 5 เทคนิค ได้แก่ โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP), การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear Discriminant Analysis) , การค้นหาเพื่อนบ้านใกล้สุด k อันดับ (K-Nearest Neighbors), ต้นไม้ตัดสินใจจำนวนมาก (Extra Trees Classifier) , การสุ่มป่าไม้ (Random Forest Classifier)

2.3.1 **โครงข่ายประสาทเทียมแบบ** Multilayer Perceptron (MLP)

MLP เป็นโครงข่ายประสาทเทียมที่มีโครงสร้างในการทำงานเป็นแบบชั้นๆ และใช้วิธีการส่งข้อมูลย้อนกลับ (Back propagation) สำหรับการทำงาน ซึ่งประกอบไปด้วย 2 ส่วนหลัก คือ การส่งข้อมูลไปข้างหน้า (Forward Pass) และการส่งข้อมูลย้อนกลับ (Backward Pass) สำหรับการส่งข้อมูลไปข้างหน้า ข้อมูลจะถูกส่งผ่านเข้าสู่โครงสร้างภายในของโครงข่ายประสาทเทียมในส่วนของข้อมูลขาเข้าและจะถูกส่งต่อไปยังส่วนต่างๆ ถัดไปจนกระทั่งถึงส่วนสุดท้ายคือข้อมูลขาออก ส่วนในการส่งข้อมูลย้อนกลับ จะทำการปรับปรุงค่าน้ำหนักของการเชื่อมต่อให้สอดคล้องกับ Error correction คือค่าส่วนต่างระหว่างคำตอบที่ทำนายได้ (Actual response) กับคำตอบที่เป้าหมายต้องการ (Target response) ค่าส่วนต่างนี้จะเรียกว่า Error signal ซึ่งจะถูกส่งย้อนกลับในทิศทางตรงกันข้ามกับการเชื่อมต่อและค่าน้ำหนักของการเชื่อมต่อจะถูกนำมาปรับปรุงพัฒนาจนกระทั่งคำตอบที่ทำนายได้จะใกล้เคียงผลตอบที่เป้าหมายต้องการ (target response) มากที่สุดดังภาพประกอบ 1 (ภรัณยา อำมฤครัตน์, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, & ณรงค์ โพธิ, 2552)



ภาพประกอบ 1 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP)

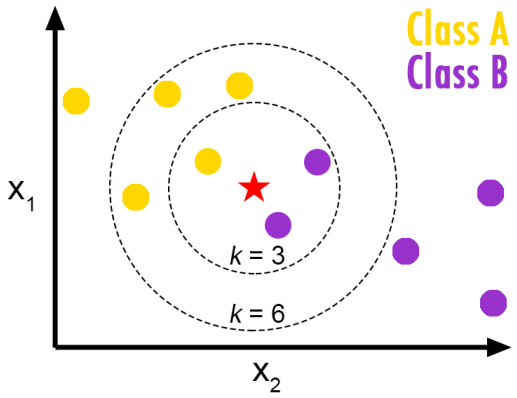
ที่มา ลักษนันท์ พลอยวัฒนาวงศ์, ปริญญา นาโท, & พยุง มีสัจ. (2560). *An Efficiency of Algorithms Machine Learning for Agricultural Product Classification*. สืบค้นจาก 203.158.224.208/05_ITS/วิจัย.ผศ.ลักษนันท์.pdf

2.3.2 การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear Discriminant Analysis)

การวิเคราะห์จำแนกประเภทของข้อมูลด้วยกระบวนการทางสถิติที่ใช้วิเคราะห์จำแนกประเภทของข้อมูลมากกว่า 2 ประเภทขึ้นไป โดยการวิเคราะห์จากตัวแปรหลัก 1 ตัวและตัวแปรอิสระจำนวน 1 ตัวขึ้นไป เป็นการแบ่งข้อมูลโดยใช้เทคนิคการลดมิติ (Reduce dimensional) จุดประสงค์ของวิธีนี้คือการลดมิติของรูปเพื่อให้ได้ตัวแบ่งข้อมูลที่ดีและยังลดการเกิด Over fitting ด้วย ซึ่งวิธีการของมันคือทำการ Maximize the distance between means และ การ Minimize the variation เพื่อสร้างแกนในการแบ่งอันใหม่ซึ่งมันมีประโยชน์กับข้อมูลที่มีหลาย Features เนื่องจากเราไม่สามารถที่จะพล็อตกราฟหลายมิติได้ เราจึงใช้ LDA ในการพล็อตกราฟกลุ่มของข้อมูลที่มีหลาย Features ให้อยู่ในภาพ 2 มิติ (ลักษนันท์ พลอยวัฒนาวงศ์, ปริญญา นาโท, & พยุง มีสัจ, 2560)

2.3.3 การค้นหาเพื่อนบ้านใกล้สุด k อันดับ (K-Nearest Neighbors)

K-Neighbors Classifier หรือ KNN เป็นหนึ่งในวิธีการแบ่งกลุ่มของข้อมูล อาศัยหลักการที่ว่าข้อมูลใหม่ที่ยังไม่ถูกติดฉลากจะมีคลาสคำตอบเป็นคลาสเดียวกันกับคลาสที่ปรากฏมากที่สุดของข้อมูลจำนวน K ตัวที่อยู่ใกล้กันมากที่สุด ซึ่ง การเลือกค่า K ให้เหมาะสมมีผลกระทบต่ออัลกอริทึม ค่า K ที่น้อยเกินไปจะทำให้ความแม่นยำลดลงเนื่องจากสนใจเพียงแค่ข้อมูลเพียงตัวเดียวที่อยู่ใกล้ที่สุด หรือค่า K ที่มากเกินไปก็จะทำให้ครอบคลุมพื้นที่ของข้อมูลที่ไม่เกี่ยวข้องมากตามไปด้วย จากภาพประกอบ 2 ถ้าเลือก K = 3 จะได้ว่ารูปดาวที่พิจารณามีคลาสคำตอบเป็นวงกลมสีม่วง แต่ถ้าเลือก K = 6 จะได้คลาสคำตอบเป็นสีเหลืองแทน เนื่องจากอยู่ใกล้วงกลมสีเหลืองสี่วง แต่อยู่ใกล้วงกลมสีม่วงเพียงสองวงเท่านั้น อัลกอริทึมจะเลือกคลาสที่ปรากฏมากที่สุดเป็นคลาสคำตอบ (THAIYATHUM, 2562)



ภาพประกอบ 2 การค้นหาเพื่อนบ้านใกล้สุด k อันดับ K-Nearest Neighbors

ที่มา THAIYATHUM, N. (2562). K-Nearest Neighbors สืบค้นจาก https://www.glurgeek.com/education/knn

2.3.4 การสุ่มป่าไม้ (Random Forest)

เทคนิคที่ใช้การสุ่มเลือกคุณลักษณะต่างๆ ของชุดข้อมูล จากนั้นนำเอาชุดข้อมูลและคุณลักษณะเหล่านี้มาทำการสร้างแบบจำลองการทำนายด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) หลายๆต้น และเลือกใช้แบบจำลองที่ได้ประสิทธิภาพดีที่สุด โดยเทคนิคการสุ่มป่าไม้ถูกพัฒนานำใช้งานปี ค.ศ. 1995 โดย Tin Kam โดยรูปแบบของต้นไม้ที่เป็นส่วนประกอบของเทคนิคการสุ่มป่าไม้จะถูกกำหนดด้วย 3 ปัจจัยหลักดังนี้

1. ต้นไม้ในทุกต้นจะถูกฝึกสอน (Train) ด้วยการนำข้อมูลย่อยจากข้อมูลหลัก
2. เมื่อต้นไม้มีขนาดใหญ่ขึ้นก็จะสามารถหาโหนด (Node) ในแต่ละโหนดที่อยู่กิ่งที่ดีที่สุดโดยใช้หลักการสุ่ม เลือกคุณลักษณะจาก N คุณลักษณะ
3. ต้นไม้แต่ละต้นจะไม่มีการทิ้ง แต่ทำให้ต้นไม้มีขนาดใหญ่ขึ้นไปเรื่อยๆ จนได้ผลลัพธ์ที่ดีที่สุดหลังจากการสร้างป่า แล้วทำการให้คะแนน (Vote) โดยต้นไม้ภายในป่า หากต้นไม้ต้นใดได้คะแนนสูงสุด ก็จะนำเอาต้นไม้นั้นออกมาสร้างเป็นโมเดล (ภูริพัทร์ ทองคำ, 2561)

2.3.5 ต้นไม้ตัดสินใจจำนานมาก (Extra Trees Classifier)

Extra Trees เป็นเหมือน Random Forest ซึ่งมันจะสร้างต้นไม้ตัดสินใจ (Decision Tree) หลายๆต้น และแยกโหนดโดยใช้คุณสมบัติแบบสุ่ม มีชื่ออีกชื่อคือ Extremely Randomized Trees แต่มีความแตกต่างที่สำคัญสองประการ คือ

1.การสร้างต้นไม้หลายต้นด้วยการที่ชุดตัวอย่างข้อมูลที่ถูกแบ่งโดยไม่ต้องเปลี่ยน

2.โหนดจะถูกแบ่งแบบสุ่มโดยเลือกจากคุณสมบัติโดยเลือกทุกโหนดไม่ได้เลือกเฉพาะโหนดที่ดีที่สุดแบบ Random Forrest (ภูริพัทร์ ทองคำ, 2561)

2.4 **การเลือกคุณลักษณะ** (Feature Selection)

การเลือกคุณลักษณะเป็นขั้นตอนในการเตรียมข้อมูลสำหรับการสร้างแบบจำลอง เป็นเทคนิคการลดขนาดของจำนวนคุณลักษณะ และคัดเลือกคุณลักษณะที่มีความสำคัญต่อการพัฒนาแบบจำลองและอาจกล่าวได้ว่าประสิทธิภาพของแบบจำลองขึ้นอยู่กับคุณลักษณะ (Feature) ที่นำมาใช้ ซึ่งนอกจากจะช่วยลดระยะเวลาในการสร้างตัวแบบจำลองให้เร็วขึ้นแล้ว ยังช่วยตัดคุณลักษณะที่ไม่จำเป็นต่อการสร้างตัวแบบจำลองได้อีกด้วย

2.4.1 **การทดสอบไคสแควร์** (Chi-Square Score)

เป็นกระบวนการเปรียบเทียบข้อมูลที่มีโครงสร้างเป็นแบบความถี่หรือโครงสร้างที่ถูกแบ่งเป็นสัดส่วน ซึ่งเป็นข้อมูลที่ถูกรวบรวมจากตัวแปรต่างๆที่เกี่ยวข้องและถูกนำมาจำแนกออกมาเป็นความถี่หรือเป็นสัดส่วน ถ้าหากต้องการศึกษาว่าการแจกแจงความถี่ของข้อมูลที่ได้จากตัวแปรหนึ่งเป็นไปในลักษณะใด หรือต้องการเปรียบเทียบตัวแปรตั้งแต่ 2 กลุ่มเป็นต้นไปว่ามีความสัมพันธ์กันหรือไม่ การทดสอบไคสแควร์จึงเป็นกระบวนการทางสถิติที่ถูกนำใช้เป็นอย่างมากในการเปรียบเทียบหรือทดสอบข้อมูลที่มีโครงสร้างเป็นแบบความถี่หรือโครงสร้างที่ถูกแบ่งเป็นสัดส่วน เช่น การใช้วิเคราะห์ข้อมูลที่ได้จากแบบสอบถามแบบมาตราส่วน (ภรัณยา อำมฤครัตน์ et al., 2552)

การทดสอบไคสแควร์โดยใช้จะมีอยู่ 2 กรณีคือ

1. การทดสอบกรณีตัวแปรเดียว (The one - variable case หรือบางครั้งอาจเรียกว่า การทดสอบภาวสารูปสนิทดี (Goodness of fit test)

2. การทดสอบกรณีสองตัวแปร (The two – variable case) เป็นการทดสอบเพื่อดูว่าตัวแปรสองตัวมีความสัมพันธ์หรือเกี่ยวข้องกันหรือไม่ ดังนั้นบางทีจึงเรียกว่า การทดสอบความเป็นอิสระ (The test for independence)

2.5 **องค์ความรู้เรื่องข้อมูลที่ไม่สมดุล** (Imbalanced Datasets)

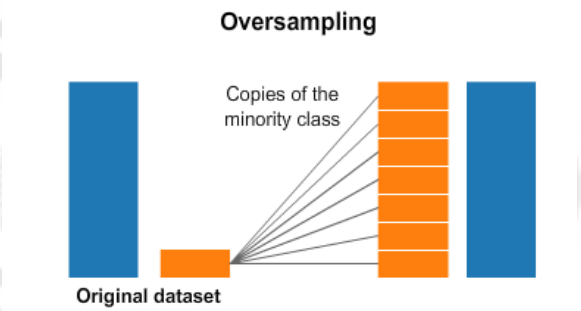
ข้อมูลที่ไม่สมดุล หมายถึง ข้อมูลที่มีการกระจายตัวที่ไม่เท่าเทียมกัน หรือข้อมูลซึ่งอัตราจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองมีจำนวนไม่เท่ากัน เช่น 100:1, 1000:1 หรือ 10000:1 เป็นต้น ตัวอย่างเช่นข้อมูลประชากรเกี่ยวกับยีนส์ที่ปกติกับยีนส์ที่ไม่ปกติ โดยปัญหานี้คำตอบที่เราต้องการ คือ คนมียีนส์ไม่ปกติ (Positive) หรือ คนที่มียีนส์ปกติ (Negative) ซึ่งข้อมูลประชากรที่มี

ยีนส์ปกติ (กลุ่มหลัก) อาจจะมีข้อมูลหลายหมื่นคน แต่ข้อมูลประชากรคนที่มียีนส์ไม่ปกติอาจมีเพียงประมาณหลักร้อยคน (กลุ่มรอง) ดังนั้นถ้านำข้อมูลทั้งสองกลุ่มมาแบ่งกลุ่มข้อมูลพร้อมกันทั้งหมด เราก็จะพบว่าคำตอบของการพยากรณ์เพื่อการแบ่งกลุ่มข้อมูลทุกๆ ข้อมูลจะถูกจัดเป็นกลุ่มประชากรที่มียีนส์ปกติทั้งหมด (Negative Class) เนื่องจากข้อมูลของคนที่มียีนส์ปกติมากกว่าคนที่มียีนส์ที่ผิดปกติ (ปณตทรง วัฒนศิริ, 2553)

การจัดการความไม่สมดุลของข้อมูล (Imbalanced datasets) เกิดขึ้นเมื่อข้อมูลที่ใช้งานมีจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมาก ทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก โดยการแก้ปัญหาที่นิยมใช้ส่วนมีดังนี้คือ

2.5.1 **เทคนิคการปรับเพิ่มข้อมูล** (Over-Sampling)

เทคนิคการปรับเพิ่มข้อมูล ซึ่งเป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อย ให้มีปริมาณข้อมูลใกล้เคียงกับข้อมูลกลุ่มมากดังภาพประกอบ 3



ภาพประกอบ 3 เทคนิคการปรับเพิ่มข้อมูล (Over-Sampling)

ที่มา ปณตทรง วัฒนศิริ. (2553). เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม. (ปริญญานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ). สืบค้นจาก http://161.200.145.125/bitstream/123456789/33369/1/panote_so.pdf

2.5.1.1 SMOTE (Synthetic Minority Over-sampling Technique)

เป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อยให้มีจำนวนเพิ่มขึ้น โดยการเพิ่มข้อมูลในกลุ่มน้อยนั้นทำให้การกระจายของกลุ่มข้อมูลมีความสมดุลมากขึ้น โดยทำการสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูลน้อยขึ้นมา 1 ค่า หลังจากนั้นพิจารณาค่าข้อมูลใกล้เคียงอีกจำนวน K ค่า (K-nearest neighbor) แล้วคำนวณค่าระยะทาง (Euclidean distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียงแต่ละค่า เพื่อหาค่าระยะทางที่น้อยที่สุดระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียง จากนั้นจึงสร้าง

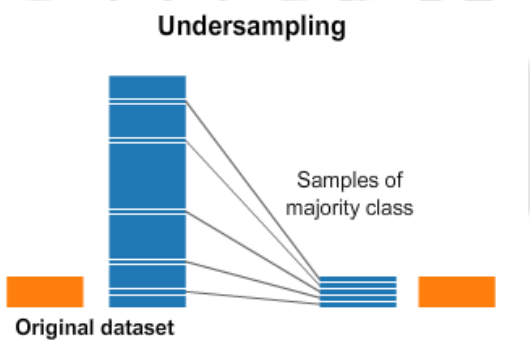
ข้อมูลเทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่ให้ค่าระยะทางที่น้อยที่สุด (ปณตทรง วัฒนศิริ, 2553)

2.5.1.2 ADASYN (Adaptive Synthetic)

เป็นวิธีที่ปรับปรุงการทำงานของ SMOTE ให้ดีขึ้น ซึ่งในขั้นตอนการสร้างข้อมูลเทียม (Synthetic data) นั้นไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดย ADASYN จะใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (Weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย โดยการสร้างข้อมูลเทียมซึ่งขึ้นอยู่กับความสำคัญของข้อมูลนั้น ๆ ถ้าข้อมูลตัวใดยากต่อการแบ่งกลุ่มก็จะให้ค่าน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาบริเวณนั้น ๆ ซึ่งจะทำให้มีการปรับขอบเขตของการตัดสินใจในการแบ่งกลุ่มดีขึ้น (ปณตทรง วัฒนศิริ, 2553)

2.5.2 **เทคนิคการปรับลดข้อมูล** (Under-Sampling)

เทคนิคการปรับลดข้อมูล ซึ่งเป็นการลดจำนวนข้อมูลของข้อมูลกลุ่มมาก ให้มีปริมาณข้อมูลใกล้เคียงกับข้อมูลกลุ่มน้อยดังภาพประกอบ 4



ภาพประกอบ 4 เทคนิคการปรับลดข้อมูล (Under-Sampling)

ที่มา ปณตทรง วัฒนศิริ. (2553). เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม. (ปริญญานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ). สืบค้นจาก http://161.200.145.125/bitstream/123456789/33369/1/panote_so.pdf

2.5.2.1 Random Under Sampling

เป็นเทคนิคการปรับลดข้อมูลด้วยวิธีสุ่มตรวจสอบและการสุ่มตัวอย่างซ้ำขึ้นมาเพื่อในการวิเคราะห์ข้อมูล ตัวอย่างเช่น สุ่มลบข้อมูลตัวอย่างในคลาสบวกที่มีจำนวนมาก โดยลบจนกระทั่งการกระจายตัวของคลาสบวกมีความสมดุลเท่ากันกับคลาสลด แต่วิธีการนี้ก็มีข้อเสียคือ

ในการลดข้อมูลตัวอย่างของแต่ละคลาสอาจจะทำให้ข้อมูลที่สำคัญ ๆ สูญหายจากชุดข้อมูลในระบบได้ (ปณตทรง วัฒนศิริ, 2553)

2.5.2.2 Cluster Centroids

เป็นเทคนิคการปรับลดข้อมูลด้วยวิธีการนี้ทำให้ข้อมูลกลุ่มมากให้มีจำนวนลดลงโดยแทนที่คลัสเตอร์ลงไปในตัวอย่างข้อมูลกลุ่มมาก ส่วนใหญ่วิธีนี้จะค้นหากลุ่มข้อมูลกลุ่มมากด้วยอัลกอริทึม K-Mean จากนั้นมันจะทำสร้างให้ข้อมูลแทนที่ในทำตำแหน่ง Centroids ของแต่ละกลุ่มคลัสเตอร์ โดยแต่ละคลัสเตอร์เป็นตัวอย่างกลุ่มข้อมูลกลุ่มมาก และจะแทนไปเรื่อยๆ จนกว่าปริมาณข้อมูลของแต่ละกลุ่มใกล้เคียงกัน (ปณตทรง วัฒนศิริ, 2553)

2.6 องค์ความรู้เรื่องการประเมินของแบบจำลอง

การพัฒนาแบบจำลองในแต่ละครั้งนั้น ผู้วิจัยจะพัฒนาแบบจำลองโดยเลือกใช้เทคนิคและกระบวนการที่แตกต่างกันออกไป ซึ่งหลักการในการเลือกเทคนิคต่างๆ นั้นมีหลากหลายวิธี โดยวิธีที่นำมาใช้ในการพัฒนาแบบจำลองในงานวิจัยนี้คือ

2.6.1 การตรวจสอบไขว้แบบแบ่งชั้นภูมิ (Stratified K-Fold Cross Validation)

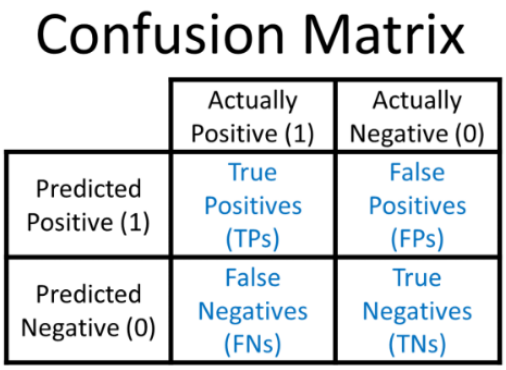
เป็นการแบ่งข้อมูลจำนวน K ส่วน ในแต่ละส่วนจะมีจำนวนข้อมูลเท่ากัน ซึ่งข้อมูลที่ถูกแบ่งในแต่ละส่วนนั้น จะมีจำนวนข้อมูลของกลุ่มต่างๆเฉลี่ยเท่ากันทุกชุดข้อมูล จากนั้นนำข้อมูล 1 ส่วนจากทั้งหมด K ส่วน ใช้เป็นชุดข้อมูลสำหรับทดสอบ (Testing Data) และข้อมูลที่เหลือ K-1 ส่วน ใช้เป็นข้อมูลสำหรับฝึกการทำนาย (Training Data) โดยจะแสดงรายละเอียดกระบวนการ K ครั้ง และนำค่าวัดประสิทธิภาพตัวแบบที่ได้จากการทดลอง K ครั้ง มาหาค่าเฉลี่ย (วิฎากร แซ่หว่อง , 2561)

2.7 องค์ความรู้เรื่องการวัดค่าประสิทธิภาพของแบบจำลอง

การเลือกตัวชี้วัดประสิทธิภาพของแบบจำลองจะต้องสามารถระบุถึงสิ่งที่ต้องการทราบได้ เนื่องจากตัวชี้วัดแต่ละค่าจะบอกถึงประสิทธิภาพของแบบจำลองในมุมมองที่ต่างกันออกไป

2.7.1 คอนฟูชั่นแมททริก (Confusion Matrix)

เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย หรือ Prediction ที่ทำนายจากแบบจำลองที่เราสร้างขึ้น โดยมีวิธีจากการวัดว่า สิ่งที่เราคิด สิ่งที่แบบจำลองทำนายกับสิ่งที่เกิดขึ้นจริง มีสัดส่วนดังภาพประกอบ 5 (วิฎากร แซ่หว่อง, 2561)



ภาพประกอบ 5 คอนฟูชั่นแมททริก (Confusion Matrix)

ที่มา วิฎากร แซ่หว่อง. (2561). การจำแนกเสียงโกรธในบทสนทนาของศูนย์ให้บริการข้อมูล. จุฬาลงกรณ์มหาวิทยาลัย , กรุงเทพฯ. (ปริญญานิพนธ์ปริญญามหาบัณฑิต). สืบค้นจาก http://161.200.145.125/bitstream/123456789/61238/1/5981539526.pdf

True Positive (TP) = สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง ในกรณีทำนายว่าจริงและสิ่งที่เกิดขึ้นก็คือ จริง

True Negative (TN) = สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณีทำนายว่าไม่จริงและสิ่งที่เกิดขึ้นก็คือ ไม่จริง

False Positive (FP) = สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นในกรณีทำนายว่ จริงแต่สิ่งที่เกิดขึ้นคือ ไม่จริง

False Negative (FN) = สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นในกรณีทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง (วิฎากร แซ่หว่อง, 2561)

2.7.1.1 **ค่าความถูกต้อง** (Accuracy)

เป็นการวัดประสิทธิภาพโดยรวมของแบบจำลอง ด้วยวิธีการคำนวณจากจำนวนคำตอบที่แบบจำลองได้ทำนายถูกต้องทั้งหมด โดยนำไปเทียบกับจำนวนข้อมูลทั้งหมดที่ต้องทำนาย (วิฎากร แซ่หว่อง, 2561)

2.7.1.2 **ค่าความครบถ้วน** (Recall)

เป็นการวัดประสิทธิภาพว่าแบบจำลองมีการครอบคลุมการทำนายข้อมูลของแต่ละกลุ่มนั้นมากน้อยเท่าไร โดยผลที่ได้จากการคำนวณจะทำให้สามารถหาแนวโน้มได้ว่าแบบจำลองนี้เหมาะสมกับการทำนายหรือไม่ (วิฎากร แซ่หว่อง, 2561)

2.7.1.3 **ค่าความแม่นยำ** (Precision)

เป็นการวัดประสิทธิภาพว่าแบบจำลองว่ามีความยากง่ายของการทำนายคำตอบของแต่ละกลุ่มเพียงใด โดยจะเปรียบเทียบกับจำนวนครั้งที่ทำนายทั้งหมด ซึ่งคำตอบที่แบบจำลองทำนายผิดจะถือเป็นอุปสรรคที่เกิดขึ้น ค่าความแม่นยำที่สูงจะบ่งบอกถึงว่าแบบจำลองสามารถทำนายคำตอบได้ถูกต้องของกลุ่มนั้น ๆ ได้ง่าย (วิฎากร แซ่ห่วง, 2561)

2.7.1.4 **ค่าประสิทธิภาพโดยรวม** (F1 Score)

เป็นการวัดประสิทธิภาพ โดยการหาค่าเฉลี่ยระหว่างค่าความความถูกต้องและค่าความแม่นยำของแต่ละกลุ่ม เพื่อแสดงถึงค่าเฉลี่ยการทำนายได้และความง่ายในการทำนายเมื่อเทียบกับอุปสรรค เป็นการวัดประสิทธิภาพของแบบจำลองโดยประเมินทั้งค่าความถูกต้องและค่าความแม่นยำ (วิฎากร แซ่ห่วง, 2561)

2.8 **เครื่องมือที่ใช้ในการทำวิจัย**

2.8.1 **จูปิเตอร์ โน๊ตบุ๊ค** (Jupyter Notebook)

Jupyter Notebook คือ เครื่องมือหนึ่งที่นิยมมากในกลุ่มคนที่ทำงานด้าน Data Science ซึ่งคนกลุ่มนี้ต้องทำงานที่เกี่ยวกับการจัดการข้อมูลเป็นจำนวนมาก แล้วยังต้องรายงานงานวิจัยที่ตนเองได้ทำขึ้นมา ซึ่งตัว Jupyter Notebook ก็ได้ออกแบบมาตรงตามจุดประสงค์การใช้งานไม่ว่าจะเป็น การเรียกใช้งาน library พร้อมทั้งเขียน Code และดูผลได้เลย Jupyter Notebook นั้นถูกออกแบบมาให้ทำงานและอ่านได้ง่ายกว่าการที่เรา Text editor รวมไปถึงเรายังสามารถที่จะพิมพ์ตัวหนังสือภาษาไทยลงไปได้เลย เหมือนเราพิมพ์รายงานใน Microsoft Word ได้เลยทีเดียว (iLog, 2562)

2.8.2 **ชุดคำสั่งไซคิทเลิร์น** (Scikit-learn)

เป็นเครื่องมือที่นิยมนำมาใช้ในงานด้านการวิเคราะห์ข้อมูล(Data Analysis) ซึ่งสามารถทำงานได้อย่างมีประสิทธิภาพที่ดี โดยมีการสร้างและพัฒนาขึ้นจากภาษาไพธอน (Python) มีเครื่องมือครอบคลุมสำหรับงานในด้านต่างๆ เช่น ด้านของเทคนิคที่ใช้ในการพัฒนาสร้างแบบจำลองมีให้เลือกใช้งานหลายประเภทคือ การจำแนกประเภท (Classification), การจำแนกแบบถดถอย (Regression) และด้านการวัดประสิทธิภาพของแบบจำลองคือ การประเมินผลลัพธ์ของการทำนาย คอนฟูชั่นแมททริก (Confusion Matrix), การตรวจสอบไขว้แบบแบ่งชั้นภูมิ (Stratified K-Fold Cross Validation) (Araya, 2562)

2.9 งานวิจัยที่เกี่ยวข้อง

บทความวิจัยที่ 1

เรื่อง POVERTY ANALYSIS USING MACHINE LEARNING METHODS (Nata Sa Plulikova, 2559)

งานวิจัยนี้มีวัตถุประสงค์เป้าหมายหลักคือการแนะนำแบบจำลองที่สามารถทำนายระดับความยากจนได้อย่างมีประสิทธิภาพ โดยนำเสนอวิธีการวัดความยากจนผ่านตัวชี้วัดความยากจนที่เราสร้างขึ้น โดยชุดข้อมูลที่ใช้ในงานวิจัยนำมาจากชุดข้อมูลประชากรประเทศอินโดนีเซีย จำนวน 6388 ครอบครัว ดังนั้นผลลัพธ์ของงานวิจัยนี้เป็นการสะท้อนระดับความยากจนในประเทศอินโดนีเซียเท่านั้น โดยผู้ทำการวิจัยคิดว่าความยากจนเป็นแนวคิดหลายมิติ โดยที่ไม่สามารถพิจารณาเรื่องรายได้ว่าเป็นตัวทำนายความยากจนเพียงเรื่องเดียว แต่นำข้อมูลต่างๆ มาพิจารณาประกอบกัน เช่น ระดับการศึกษาสุขภาพร่างกาย และสภาพครัวเรือน ซึ่งตัวแปรทั้งหมดเหล่านี้ล้วนเป็นปัจจัยสำคัญที่สามารถพิจารณาได้ว่าบุคคลดังกว่ามีระดับความยากจนในระดับใด โดยผู้วิจัยจึงได้ประยุกต์การใช้ Machine learning นำมาใช้ในงานวิจัย โดยข้อมูลทั้งหมดถูกแบ่งการวิเคราะห์ออกเป็น 2 กลุ่มส่วนคือ ส่วนแรกวิเคราะห์ข้อมูลและทำความเข้าใจข้อมูลโดยใช้รูปแบบ Multiple correspondence และใช้ K-Means ในการจัดกลุ่มและกำหนดระดับความยากจน ส่วนที่สอง จะใช้การเปรียบเทียบการทำนายของแต่ละอัลกอริทึมคือ Binary logistic regression , Neural networks , Decision trees และ Random forests และใช้ Sensitivity , Specificity และ Accuracy ในการประเมินประสิทธิภาพ ซึ่งผลลัพธ์จะเป็นการทำนายระดับความยากจนดังตาราง 1 เมื่อเปรียบเทียบการวิเคราะห์ในแต่ละอัลกอริทึมแล้ว ผู้วิจัยสรุปได้ว่า Neural Networks มีความแม่นยำมากที่สุด

ตาราง 1 ผลลัพธ์การทำนายการระดับความยากจนของบทความวิจัยที่ 1

Algorithm	Sensitivity	Specificity	Accuracy
Binary logistic regression	16%	97%	85%
Neural networks	24%	95%	84%
Decision trees	14	98	85
Random Forests	18	97	85

บทความวิจัยที่ 2

เรื่อง Satellite data and machine learning tools for predicting poverty in rural India (Subash, Kumar, & Aditya, 2018)

งานวิจัยนี้มีวัตถุประสงค์เป้าหมายหลักคือการทำนายความยากจนโดยใช้ภาพถ่ายดาวเทียม ซึ่งในปัจจุบันด้วยความสามารถของภาพถ่ายดาวเทียมที่พัฒนาไปอย่างมาก และการพัฒนาอย่างรวดเร็วในการประมวลผลข้อมูลภาพ ได้กระตุ้นความสนใจในหมู่นักเศรษฐศาสตร์ในการใช้ภาพถ่ายดาวเทียมนี้สำหรับการแปลผลในเชิงเศรษฐกิจให้ออกมามีความหมาย โดยใช้ข้อมูลภาพถ่ายไฟในเวลากลางคืนจากดาวเทียมเป็นตัวบ่งชี้ความยากจน เนื่องจากสถิติความยากจนในอินเดียได้รับการปล่อยตัวหนึ่งครั้งในห้าปี ความถี่ของแสงไฟในยามค่ำคืนสามารถใช้ในการทำนายความยากจนในปีนั้นๆ โดยที่ไม่ต้องพึ่งผลสถิติการทำนายจากทางการของประเทศอินเดีย ในงานวิจัยนี้ผู้จัดทำได้ทำการเปรียบเทียบข้อมูลสองชนิดคือ ภาพถ่ายดาวเทียมไฟในเวลากลางคืนและผลิตภัณฑ์โดยรวมภายในประเทศต่อหัว (GDP Per Capital) หรือ GDP ซึ่งใช้หลักการ Machine learning โดยการนำอัลกอริทึม Artificial Neural Network สร้างแบบจำลองทำนายความยากจนในชนบทระดับภูมิภาคของประเทศอินเดีย ภาพถ่ายดาวเทียมไฟในเวลากลางคืนและผลิตภัณฑ์รวมภายในประเทศต่อหัว (GDP) จากนั้นนำมาเปรียบเทียบด้วยวิธี RSME และ R^2 เพื่อวัดประสิทธิภาพการทำนายความยากจนที่แม่นยำที่สุด ผลลัพธ์จะเป็นการทำนายระดับความยากจนดังตาราง 2 พบว่าภาพถ่ายดาวเทียมไฟในเวลากลางคืนสามารถทำนายความยากจนได้ดีกว่าผลิตภัณฑ์รวมภายในประเทศต่อหัว (GDP)

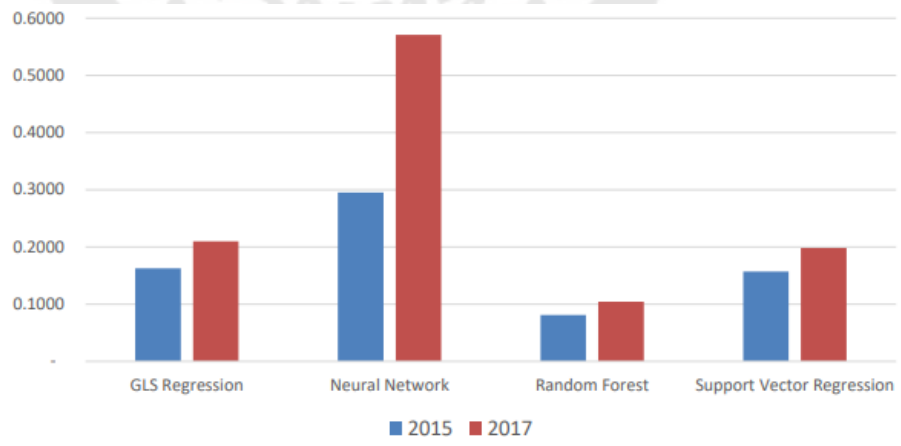
ตาราง 2 ผลลัพธ์การทำนายการระดับความยากจนของบทความวิจัยที่ 2

Model	Dataset	RSME	R2
Artificial neural network	Night Light	7.38	0.59
Artificial neural network	GDP	8.29	0.56

บทความวิจัยที่ 3

เรื่อง การวิเคราะห์ความเหลื่อมล้ำในประเทศไทยโดยใช้ข้อมูลดาวเทียมและภูมิสารสนเทศ (ณัฐพงษ์ พัฒนพงษ์ , Arturo M. Martinez Jr. , Ron Lester Santos Durante , & พิชญ์ จงวัฒนากุล, 2563)

งานวิจัยนี้ประกอบด้วย 2 วัตถุประสงค์หลัก โดยวัตถุประสงค์แรกเป็นการนำเสนอการบูรณาการข้อมูลซึ่งประยุกต์ใช้วิธีการทางภูมิเศรษฐมิติและกระบวนการ Machine Learning ร่วมกับข้อมูลดาวเทียมและภูมิสารสนเทศ วัตถุประสงค์ที่สองเป็นการสร้างแบบจำลองเพื่อพยากรณ์ระดับความยากจนในประเทศไทยโดยใช้ชุดข้อมูลดาวเทียมจาก Google Earth Engine และข้อมูลภูมิสารสนเทศจาก Open Street Map ผลลัพธ์ที่ได้จากการวิเคราะห์ด้วยวิธีการทางภูมิเศรษฐมิติและกระบวนการ Machine Learning แสดงให้เห็นว่า กรุงเทพฯและปริมณฑล เป็นศูนย์กลางแห่งเดียวของประเทศมีคุณลักษณะดึงดูดให้เกิดกิจกรรมทางเศรษฐกิจหนาแน่น โดยผลจากการพัฒนาแบบจำลองฯ โดยหาความสัมพันธ์จากตัวแปรที่เกี่ยวข้องกับความหนาแน่นของเมือง เช่น ความหนาแน่นของแสงสว่างกลางคืน ความหนาแน่นของอาคาร ดัชนีพืชพรรณ ความหนาแน่นของถนน ฯลฯ โดยการเปรียบเทียบประสิทธิภาพแต่ละอัลกอริทึมคือ GLS Regression , Neural Network , Random Forest, Support Vector Regression โดยการสร้างแบบจำลอง Machine Learning และวัดประสิทธิภาพด้วยเทคนิค RMSE ผลการทดลองดังภาพประกอบ 6 พบว่าวิธี Random Forest ให้ความแม่นยำสูงในระดับ สำหรับชุดข้อมูลปี 2015 และ 2017 ผลที่ได้นี้แสดงให้เห็นว่าสามารถนำข้อมูลดาวเทียมข้อมูลภูมิสารสนเทศ และกระบวนการ Machine Learning มาใช้ในการศึกษารายละเอียดของการกระจายตัวเชิงภูมิศาสตร์ของความยากจนและปัจจัยที่ส่งผล นอกจากนี้ผลลัพธ์และวิธีการวิเคราะห์ยังสามารถขยายขอบเขตไปสู่การประยุกต์ใช้ในสาขาอื่น ๆ ที่เกี่ยวข้องได้



ภาพประกอบ 6 ผลการวัดประสิทธิภาพของแบบจำลองด้วยเทคนิค RMSE บทความวิจัยที่ 3

ที่มา ณัฐพงษ์ พัฒนพงษ์ , Arturo M. Martinez Jr. , Ron Lester Santos Durante , & พิชญ์ จงวัฒนากุล. (2563). การวิเคราะห์ความเหลื่อมล้ำในประเทศไทยโดยใช้ข้อมูลดาวเทียมและภูมิสารสนเทศ. สืบค้นจาก http://www.econ.tu.ac.th/oldweb/doc/content/ 1850/ Discussion_Paper_No.57.pdf

บทความวิจัยที่ 4

เรื่อง Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification (Sani, Rahman, Bakar, Sahran, & Sarim, 2018)

งานวิจัยนี้มีวัตถุประสงค์เป้าหมายหลักคือการทำนายความยากจน โดยชุดข้อมูลที่ใช้ในงานวิจัยนำมาจากชุดข้อมูลประเทศมาเลเซียจากธนาคารข้อมูลความยากจนแห่งชาติที่เรียกว่า eKasih ที่ได้รับจากกรมสวัสดิการสังคมหน่วยประสานงานเพื่อการปฏิบัติงานของนายกรัฐมนตรี (ICU JPM) ประกอบด้วย 99,546 ครัวเรือนจาก 3 รัฐคือ ยะโฮร์, ตรังกานูและปะหัง แบ่งออกเป็นสามกลุ่มรายได้ที่แตกต่างกันซึ่ง ได้แก่ 20 เปอร์เซ็นต์สูงสุด (T20), 40 เปอร์เซ็นต์กลาง (M40), และ 40 เปอร์เซ็นต์ล่าง (B40) ทางผู้วิจัยได้พัฒนาสร้างแบบจำลองการทำนายความยากจนผ่านการจำแนกประเภทต่างๆ เพื่อระบุแบบจำลองการทำนายความยากจนที่ดีที่สุดโดยใช้อัลกอริทึมดังนี้ Naive Bayes, Decision Tree และ K-Neighbor Neighbor สำหรับการพัฒนาแบบจำลองการทำนายความยากจน มีการนำกระบวนการต่างๆของ Machine learning เช่น วิธีการเตรียมข้อมูลต่างๆ เช่น การล้างข้อมูล (Cleansing Data) แอตทริบิวต์สหสัมพันธ์ (Correlation), Feature Selection, Grid Search สำหรับปรับความเหมาะสมของพารามิเตอร์และตรวจสอบความถูกต้อง 10-Fold Cross-Validation เพื่อให้ได้ค่าที่ดีที่สุดก่อนที่จะเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความยากจนทั้ง 3 แบบจำลอง ผลการทดลองดังตาราง 3 แสดงให้เห็นว่าประสิทธิภาพโดยรวมของแบบจำลอง Decision Tree นั้นดีที่สุด

ตาราง 3 ผลลัพธ์การทำนายระดับความยากจนของบทความวิจัยที่ 4

Model	Score
Naive Bayes	97.27%
Decision Tree	99.27%
k-Neighbor Neighbor	96.80%

บทความวิจัยที่ 5

IDENTIFYING POVERTY-DRIVEN NEED BY AUGMENTING CENSUS AND COMMUNITY SURVEY DATA (Korivi, 2016)

งานวิจัยนี้มีวัตถุประสงค์เป้าหมายหลักคือการสร้างแบบจำลองเพื่อประเมินความต้องการพื้นฐานตามตัวแปรทางเศรษฐกิจและสังคมโดยรวม ที่บ่งชี้ถึงความยากจนแบบสัมบูรณ์และแบบสัมพัทธ์ โดยชุดข้อมูลที่ใช้ในงานวิจัยนำมาจากชุดข้อมูล American Community Survey (ACS) มาสร้างแบบจำลองเพื่อประเมินความต้องการและระบุระดับความยากจน โดยขั้นตอนเตรียมข้อมูลมีการทำ Missing Values, Feature Selection จากนั้นทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมต่างๆ คือ Random Forest, Linear Support Vector Machine และ Logistic Regression โดยใช้หลักการ Confusion Matrix เปรียบเทียบประสิทธิภาพ ผลการทดลองที่ได้ดังตารางที่ 4 คือ Random Forest มีประสิทธิภาพสูงที่สุดในเรื่องคะแนน และ Random Forest ยังใช้เวลาในการประมวลผลเร็วกว่าอัลกอริทึมอื่นๆ

ตาราง 4 ผลลัพธ์ประเมินความต้องการพื้นฐานตามตัวแปรทางเศรษฐกิจของบทความวิจัยที่ 5

Model	Accuracy	Precision	Recall	F1 score
Random Forest	0.79	0.77	0.72	0.73
Linear Support Vector Machine	0.75	0.71	0.71	0.71
Logistic Regression	0.78	0.79	0.68	0.69

บทความวิจัยที่ 6

Quarterly Multidimensional Poverty Predictions in Mexico using Machine Learning Algorithms (Rincon, 2562)

งานวิจัยนี้มีวัตถุประสงค์คือสร้างแบบจำลองเพื่อประเมินอัตราความยากจนในเม็กซิโกเป็นรายไตรมาส โดยชุดข้อมูลที่ใช้ในงานวิจัยนำมาจากชุดข้อมูลรายได้ของครัวเรือนและการใช้จ่าย ตั้งแต่ปี 2008 ถึงปี 2016 ของประเทศเม็กซิโก โดยการจำแนกว่าครอบครัวนี้จนหรือไม่ โดยการสร้างแบบจำลองตามกระบวนการ Machine Learning โดยใช้การเปรียบเทียบแต่ละอัลกอริทึม 3 แบบคือ Linear Discriminant Analysis (LDA), Random Forest (RF) and Support Vector Machines (SVM) และวัดประสิทธิภาพด้วยวิธี 10-Fold Cross-Validation และ Confusion

Matrix ผลการทดลองเปรียบเทียบประสิทธิภาพของแบบจำลองดังภาพประกอบ 7 ประสิทธิภาพโดยรวมของ Random Forest (RF) ได้อัตราเฉลี่ยดีที่สุดสำหรับชุดข้อมูล ตั้งแต่ปี 2008 ถึงปี 2016

	Accuracy	Recall	Specificity	Precision	NPV	F_1 score
			2008			
RF	0.944	0.943	0.945	0.927	0.957	0.935
SVM	0.890	0.884	0.894	0.860	0.913	0.872
Logit	0.871	0.863	0.877	0.839	0.896	0.850
LDA	0.815	0.734	0.875	0.813	0.817	0.772
			2010			
RF	0.941	0.947	0.936	0.928	0.953	0.938
SVM	0.889	0.898	0.880	0.868	0.908	0.883
Logit	0.867	0.878	0.857	0.844	0.889	0.861
LDA	0.809	0.766	0.847	0.815	0.805	0.790
			2012			
RF	0.938	0.948	0.929	0.922	0.952	0.935
SVM	0.879	0.889	0.870	0.859	0.898	0.874
Logit	0.860	0.873	0.848	0.836	0.882	0.854
LDA	0.804	0.766	0.837	0.807	0.801	0.786
			2014			
RF	0.934	0.939	0.930	0.919	0.948	0.929
SVM	0.872	0.867	0.876	0.855	0.886	0.861
Logit	0.856	0.863	0.850	0.829	0.880	0.846
LDA	0.798	0.749	0.840	0.798	0.798	0.773
			2016			
RF	0.939	0.935	0.941	0.917	0.954	0.926
Logit	0.853	0.830	0.868	0.814	0.880	0.822
SVM	0.837	0.726	0.914	0.854	0.827	0.785
LDA	0.803	0.683	0.887	0.808	0.800	0.740

ภาพประกอบ 7 ผลการวัดประสิทธิภาพของแบบจำลองของบทความวิจัยที่ 6

ที่มา Rincon, R. (2562). Quarterly Multidimensional Poverty Predictions in Mexico using Machine Learning Algorithms. Retrieved from https://sistemas.colmex.mx/Reportes/LACEALAMES/LACEA-LAMES2019_paper_754.pdf

บทความวิจัยที่ 7

A Statistical Approach to Adult Census Income Level Prediction (Chakrabarty, Navoneel, Biswas, & Sanket;, 2018)

งานวิจัยนี้มีวัตถุประสงค์คือสร้างแบบจำลองการทำนายรายรับต่อปีของบุคคลหนึ่งในสหรัฐอเมริกา โดยใช้กระบวนการ Machine Learning ซึ่งใช้ชุดข้อมูลจาก University of California Irvine (UCI) ประกอบไปด้วย 48,842 ข้อมูลและ 14 คุณลักษณะ โดยการทำนายว่า

บุคคลนั้นมีรายได้มากกว่า 50,000 ดอลลาร์หรือน้อยกว่า 50,000 ดอลลาร์ เพื่อแก้ปัญหาความเท่าเทียมกันทางรายได้ โดยการสร้างแบบจำลอง ขั้นตอนเริ่มจากการทำความเข้าใจ Feature และทำ Feature Selection จากนั้นทำการสร้างแบบจำลองโดยใช้อัลกอริทึม Gradient Boosting Classifier และใช้ Grid Search ในการปรับปรุงพารามิเตอร์ สุดท้ายวัดประสิทธิภาพของแบบจำลองด้วยคะแนน Accuracy Score , Recall , Precision, และ AUC ซึ่งผลการทดลองที่ได้ตามตาราง 5 พบว่าอัลกอริทึม Gradient Boosting Classifier สามารถทำนายรายได้บุคคลนั้นมีความถูกต้องคือ Accuracy 88.16 , Recall Precision 0.88 และ F1-Score เท่ากันทั้งคู่

ตาราง 5 ผลลัพธ์ประเมินความต้องการพื้นฐานตามตัวแปรทางเศรษฐกิจของบทความวิจัยที่ 7

Model	Accuracy	Precision	Recall	F1 score	AUC
Gradient Boosting Classifier	88.16%	00.88	0.88	0.73	0.93

จากงานวิจัยที่เกี่ยวข้องที่ได้ทำการศึกษาและวิเคราะห์เกี่ยวกับการทำนายระดับความยากจน ผู้วิจัยเห็นว่างานวิจัยเหล่านี้มีประโยชน์ในการใช้อ้างอิงจึงทำการเขียนรายละเอียดสรุปออกมา ดังตาราง 6 และตาราง 7 ละเพื่อให้ง่ายต่อการทำความเข้าใจ จึงได้จัดทำตารางคำอธิบายตัวย่ออัลกอริทึมในการสร้างแบบจำลอง(MODEL) ดังตาราง 8 และตาราง 9 และตารางคำอธิบายตัวย่อวิธีการประเมินแบบจำลอง(EVALUATION) ดังตาราง 10

ตาราง 6 สรุปงานวิจัยที่ศึกษาทั้งหมด

ลำดับ	ชื่องาน	ชื่อผู้จัดทำ	ปีที่พิมพ์	Model					Evaluation		
1	POVERTY ANALYSIS USING MACHINE LEARNING METHODS	1. NATAŠA PLULIKOVÁ	2016	BLR	ANN	DCT	RF	SST	SCP	ACC	
2	Satellite data and machine learning tools for predicting poverty in rural India	1. S P Subasha 2. Rajeev Ranjan Kumarb 3. K S Aditya	2018	ANN				RSME	R^2		

ตาราง 7 สรุปงานวิจัยที่ศึกษาทั้งหมด (ต่อ)

ลำดับ	ชื่องาน	ชื่อผู้จัดทำ	ปีที่พิมพ์	Model				Evaluation				
3	การวิเคราะห์ความเหลื่อมล้ำในประเทศไทยโดยใช้ข้อมูลดาวเทียมและภูมิสารสนเทศ	1.ณัฐพงษ์ พัฒนพงษ์ 2. Arturo M. Martinez Jr. 3. Ron Lester Santos Durante 4.พิชญ์ จงวัฒนากุล	2020	GLS	ANN	RF	SVM	RSME				
4	Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification	.1Nor Samsiah .2Sani Mariah Abdul Rahman .3Azuraliza Abu Bakar .4Shahnorbanun Sahran .5Hafiz Mohd Sarim	2018	NB	DCT	KNN		K-Fold				
5	IDENTIFYING POVERTY-DRIVEN NEED BY AUGMENTING CENSUS AND COMMUNITY SURVEY DATA	KEERTHI KORIVI	2012	LSVM	LG	RF		ACC	Re	PCS	F1	
6	Quarterly Multidimensional Poverty Predictions in Mexico using Machine Learning Algorithms	Ratzanyel Rincon	2019	LDA	RF	SVM		K-Fold	ACC	Re	PCS	F1
7	A Statistical Approach to Adult Census Income Level Prediction	1.Navoneel Chakrabarty 2.Sanket Biswas	2017	GBT				ACC	Re	PCS	F1	AUC

ตาราง 8 คำอธิบายสำหรับ MODEL

No.	MODEL	DESCRIPTION
1	BLR	Binary Logistic Regression
2	ANN	Artificial neural network
3	DCT	decision trees
4	RF	random forests
5	GLS	Generalized least squares Regression

ตาราง 9 คำอธิบายสำหรับ MODEL (ต่อ)

No.	MODEL	DESCRIPTION
6	NB	Naive Bayes
7	KNN	K-Neighbor Neighbor
8	SVM	Support Vector Machine
9	LSVM	Linear Support Vector Machine
10	LG	Logistic Regression
11	LDA	Linear Discriminant Analysis
12	GBT	Gradient Boosting Classifier

ตาราง 10 คำอธิบายสำหรับ EVALUATION

No.	EVALUATION	DESCRIPTION
1	SST	Sensitivity
2	SCP	Specificity
3	K-Fold	K-fold cross validation
4	RSME	Root Mean Square Error
5	R^2	R-Squared
6	Accuracy	Accuracy Score
7	Re	Recall Score
8	PCS	Precision Score
9	F1	F1 Score

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามกระบวนการและขั้นตอนต่างๆ ดังต่อไปนี้

3.1 ทำความเข้าใจปัญหาและความต้องการ

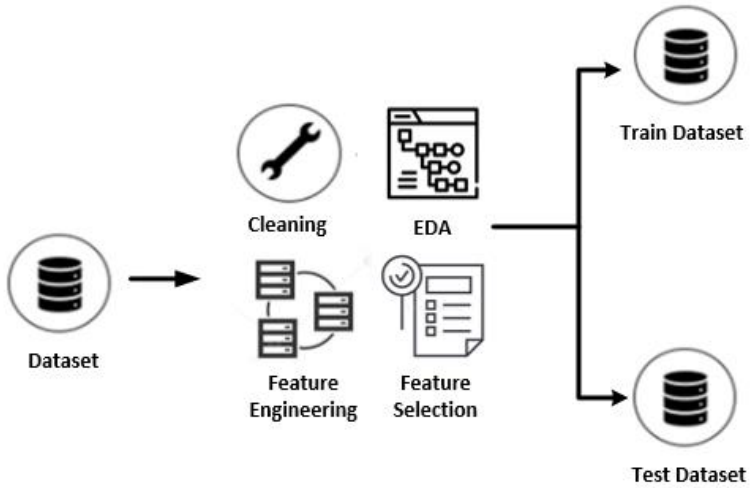
3.2 การทำความเข้าใจข้อมูล

3.3 การเตรียมข้อมูล

3.4 การสร้างแบบจำลอง

3.5 การประเมินผลแบบจำลอง

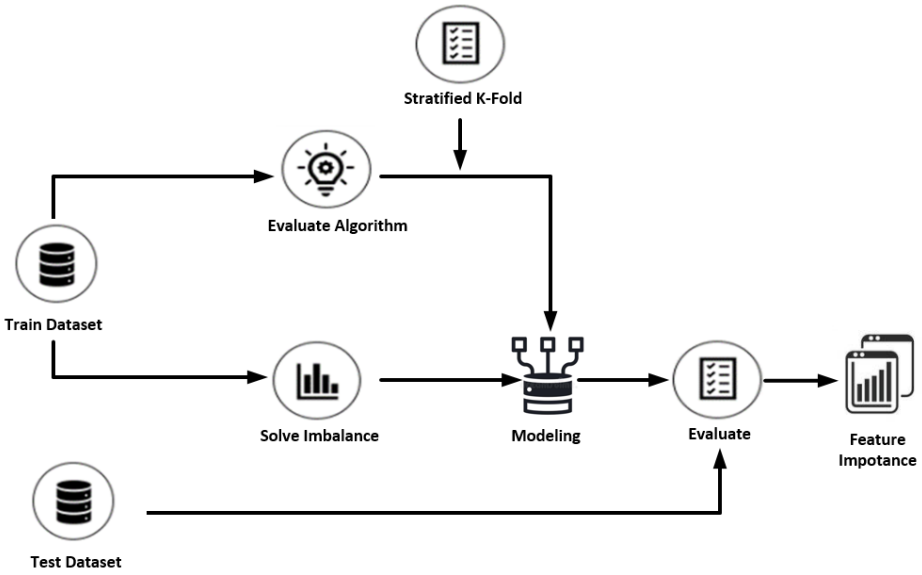
วัตถุประสงค์ในการทำการวิจัยคือ การพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคลด้วยการเรียนรู้ของเครื่อง มีขั้นตอนแบ่งเป็น 2 ส่วน ส่วนที่หนึ่งเป็นการทำความเข้าใจข้อมูลและเตรียมข้อมูล โดยนำข้อมูลมาผ่านกระบวนการต่างๆ เช่น ทำความสะอาด วิเคราะห์ข้อมูลเชิงสำรวจ ทำวิศวกรรมข้อมูลและเลือกคุณลักษณะ เพื่อให้ได้ชุดข้อมูลใหม่ จากนั้นแบ่งชุดข้อมูลเป็น 2 ชุดคือ ชุดข้อมูลฝึก(Train) และชุดข้อมูลทดสอบ(Test) ในอัตราส่วน 70 : 30 เพื่อนำไปสร้างแบบจำลองต่อไป ดังภาพประกอบ 8



ภาพประกอบ 8 การทำความเข้าใจข้อมูลและเตรียมข้อมูล

ส่วนที่สองประเมินอัลกอริทึมและสร้างแบบจำลองรวมทั้งวัดประสิทธิภาพ โดยนำชุดข้อมูลฝึก(Train) ที่ได้ในขั้นตอนก่อนหน้านี้มาทำการประเมินอัลกอริทึมเพื่อเลือกอัลกอริทึมที่

เหมาะสมและนำชุดข้อมูลมาทำการปรับปรุงข้อมูลที่ไม่สมดุลกัน จากนั้นนำข้อมูลที่ผ่านการปรับปรุง มาสร้างแบบจำลองและประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Test) ดังภาพประกอบ 9



ภาพประกอบ 9 ประเมินอัลกอริทึมและสร้างแบบจำลองรวมทั้งวัดประสิทธิภาพ

3.1 ทำความเข้าใจปัญหาและความต้องการ

การวางเป้าหมายในการทำงานถือเป็นขั้นตอนแรก และเป็นขั้นตอนที่สำคัญที่สุดในการกำหนดแนวทางของการทำงาน โดยเริ่มจากปัญหาที่พบให้ชัดเจน หรือข้อสงสัยต่างๆ ที่ทำให้ต้องหาทางแก้ไข การทำความเข้าในส่วนงานต่างๆ โดยมีจุดประสงค์ของงานวิจัยนี้คือ การพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคล โดยทำการศึกษาและทดลองจากตัวอย่างข้อมูลสำมโนประชากรของประเทศคอสตาริกา ซึ่งเป็นข้อมูลของ Inter-American Development Bank ซึ่งป็นการเก็บข้อมูลมาจากประชาชนชาวคอสตาริกา

3.2 การทำความเข้าใจข้อมูล

ชุดข้อมูลนี้คือข้อมูลสำมโนประชากรของประเทศคอสตาริกา เป็นข้อมูลของ Inter-American Development Bank ซึ่งเป็นการเก็บข้อมูลมาจากประชาชนชาวคอสตาริกา

3.2.1 ชุดข้อมูล

ประกอบไปด้วยข้อมูลจำนวน 9557 แถวและ 143 คอลัมน์ตัวอย่างดังภาพประกอบ 10 มีรายละเอียดดังนี้

1. จำนวนแถวข้อมูล 9557 แถว แสดงถึงข้อมูลของแต่ละบุคคลที่ไม่ซ้ำกัน

2. คอลัมน์จำนวน 143 คอลัมน์ ดังตาราง 11 ถึงตาราง 21 แสดงถึงข้อมูลที่บ่งบอกถึงคุณลักษณะเฉพาะของแต่ละบุคคล โดยประเภทของคอลัมน์ประกอบไปด้วย จำนวนเต็ม (Integer) 130 คอลัมน์ จำนวนจริง (Float) 8 คอลัมน์ และกลุ่มวัตถุ (Object) 5 คอลัมน์

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

9557 rows × 143 columns

ภาพประกอบ 10 ตัวอย่างชุดข้อมูล

ตาราง 11 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

No	Name	description	คำอธิบาย
1	Id	a unique identifier for each row.	ตัวระบุที่ไม่ซ้ำกันสำหรับแต่ละแถว
2	v2a1	Monthly rent payment	ค่าเช่ารายเดือน
3	hacdor	=1 Overcrowding by bedrooms	=1 ถ้ามีความแออัดยัดเยียดในห้องนอน
4	rooms	number of all rooms in the house	จำนวนห้องทั้งหมดในบ้าน

ตาราง 12 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 1)

5	hacapo	=1 Overcrowding by rooms	=1 ถ้ามีความแออัดยัดเยียดในห้องพัก
6	v14a	=1 has toilet in the household	=1 ถ้ามีห้องน้ำในครอบครัว
7	refrig	=1 if the household has refrigerator	=1 ถ้ามีตู้เย็นในครอบครัว
8	v18q	owns a tablet	จำนวนแท็บเล็ตที่เจ้าของบ้านเป็นเจ้าของ
9	v18q1	number of tablets household owns	จำนวนแท็บเล็ตที่ใช้ในครอบครัว
10	r4h1	Males younger than 12 years of age	จำนวนเพศชายอายุน้อยกว่า 12 ปี
11	r4h2	Males 12 years of age and older	จำนวนเพศชายอายุ 12 ปีขึ้นไป
12	r4h3	Total males in the household	จำนวนผู้ชายทั้งหมดในครอบครัว
13	r4m1	Females younger than 12 years of age	จำนวนผู้หญิงอายุน้อยกว่า 12 ปี
14	r4m2	Females 12 years of age and older	จำนวนเพศหญิงอายุ 12 ปีขึ้นไป
15	r4m3	Total females in the household	จำนวนหญิงในครอบครัวทั้งหมด
16	r4t1	persons younger than 12 years of age	จำนวนคนอายุน้อยกว่า 12 ปี
17	r4t2	persons 12 years of age and older	จำนวนคนที่อายุ 12 ปีขึ้นไป

ตาราง 13 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 2)

18	r4t3	Total persons in the household	จำนวนคนทั้งหมดในครอบครัว
19	tamhog	size of the household	ขนาดของครอบครัว
20	tamviv	TamViv	จำนวนคนที่อาศัยอยู่ในครัวเรือน
21	escolari	years of schooling	จำนวนปีในการศึกษา
22	rez_esc	Years behind in school	จำนวนปีการเข้าเรียนช้ากว่ากำหนด
23	hhsize	household size	ขนาดครัวเรือน
24	paredblolad	=1 if predominant material on the outside wall is block or brick	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นบล็อกหรืออิฐ
25	paredzocalo	=1 if predominant material on the outside wall is socket (wood, zinc or absbesto	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นไม้หรือ สังกะสี
26	paredpreb	=1 if predominant material on the outside wall is prefabricated or cement	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นแบบสำเร็จรูปหรือซีเมนต์
27	pareddes	=1 if predominant material on the outside wall is waste material	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นวัสดุเหลือทิ้ง
28	paredmad	=1 if predominant material on the outside wall is wood	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นไม้
29	paredzinc	=1 if predominant material on the outside wall is zink	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นสังกะสี

ตาราง 14 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 3)

30	paredfibras	=1 if predominant material on the outside wall is natural fibers	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นเส้นใยธรรมชาติ
31	paredother	=1 if predominant material on the outside wall is other	=1 ถ้าวัสดุเด่นบนผนังด้านนอกเป็นอื่น ๆ
32	pisomoscer	=1 if predominant material on the floor is mosaic, ceramic, terrazo	=1 ถ้าวัสดุเด่นบนพื้นคือกระเบื้องโมเสค, เซรามิก, หินขัด
33	pisocemento	=1 if predominant material on the floor is cement	=1 ถ้าวัสดุเด่นบนพื้นเป็นซีเมนต์
34	pisoother	=1 if predominant material on the floor is other	=1 ถ้าวัสดุเด่นบนพื้นเป็นอื่น ๆ
35	pisonatur	=1 if predominant material on the floor is natural material	=1 ถ้าวัสดุเด่นบนพื้นเป็นวัสดุธรรมชาติ
36	pisonotiene	=1 if no floor at the household	=1 ถ้าไม่มีพื้นบ้าน
37	pisomadera	=1 if predominant material on the floor is wood	=1 ถ้าวัสดุเด่นบนพื้นเป็นไม้
38	techozinc	=1 if predominant material on the roof is metal foil or zink	=1 ถ้าวัสดุเด่นบนหลังคาเป็นโลหะฟอยล์หรือสังกะสี
39	techoentrepiso	=1 if predominant material on the roof is fiber cement, mezzanine	=1 ถ้าวัสดุเด่นบนหลังคาคือไฟเบอร์ซีเมนต์, ชั้นลอย
40	techocane	=1 if predominant material on the roof is natural fibers	=1 ถ้าวัสดุเด่นบนหลังคาเป็นเส้นใยธรรมชาติ

ตาราง 15 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 4)

41	techootro	=1 if predominant material on the roof is other	=1 ถ้าวัสดุเด่นบนหลังคาเป็นอื่น ๆ
42	cielorazo	=1 if the house has ceiling	=1 ถ้าบ้านมีเพดาน
43	abastaguadentro	=1 if water provision inside the dwelling	=1 ถ้ามีน้ำภายในที่อยู่อาศัย
44	abastaguafuera	=1 if water provision outside the dwelling	=1 ถ้ามีน้ำนอกที่อยู่อาศัย
45	abastaguano	=1 if no water provision	=1 หากไม่มีการจัดสรรน้ำ
46	public	=1 electricity from CNFL, ICE, ESPH/JASEC	=1 ใช้ไฟฟ้าจาก CNFL, ICE, ESPH / JASEC
47	planpri	=1 electricity from private plant	=1 ใช้ไฟฟ้าจากโรงงานส่วนตัว
48	noelec	=1 no electricity in the dwelling	=1 ไม่มีไฟฟ้าใช้ในที่พักอาศัย
49	coopele	=1 electricity from cooperative	=1 ใช้ไฟฟ้าจากสหกรณ์
50	sanitario1	=1 no toilet in the dwelling	=1 ไม่มีห้องน้ำในบ้าน
51	sanitario2	=1 toilet connected to sewer or cesspool	=1 ห้องน้ำที่เชื่อมต่อกับท่อระบายน้ำหรือส้วมซึม
52	sanitario3	=1 toilet connected to septic tank	=1 ห้องน้ำเชื่อมต่อกับถังบำบัดน้ำเสีย
53	sanitario5	=1 toilet connected to black hole or latrine	=1 ห้องน้ำที่เชื่อมต่อกับหลุมดำหรือส้วม
54	sanitario6	=1 toilet connected to other system	=1 ห้องน้ำเชื่อมต่อกับระบบอื่น

ตาราง 16 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 5)

55	energcocinar1	=1 no main source of energy used for cooking (no kitchen)	=1 ไม่มีแหล่งพลังงานหลักที่ใช้สำหรับการปรุงอาหาร (ไม่มีครัว)
56	energcocinar2	=1 main source of energy used for cooking electricity	=1 แหล่งพลังงานหลักที่ใช้ในสำหรับการปรุงอาหารคือไฟฟ้า
57	energcocinar3	=1 main source of energy used for cooking gas	=1 แหล่งพลังงานหลักที่ใช้สำหรับการปรุงอาหารคือแก๊สหุงต้ม
58	energcocinar4	=1 main source of energy used for cooking wood charcoal	=1 แหล่งพลังงานหลักที่ใช้ใสำหรับการปรุงอาหารคือถ่านไม้
59	elimbasu1	=1 if rubbish disposal mainly by tanker truck	=1 ถ้าการกำจัดขยะส่วนใหญ่โดยรถบรรทุก
60	elimbasu2	=1 if rubbish disposal mainly by botan hollow or buried	=1 ถ้าการกำจัดขยะส่วนใหญ่โดยฝังกลบ
61	elimbasu3	=1 if rubbish disposal mainly by burning	=1 ถ้าการกำจัดขยะส่วนใหญ่โดยการเผาไหม้
62	elimbasu4	=1 if rubbish disposal mainly by throwing in an unoccupied space	=1 ถ้าการกำจัดขยะส่วนใหญ่โดยการทิ้งในพื้นที่ว่าง
63	elimbasu5	=1 if rubbish disposal mainly by throwing in river, creek or sea	=1 ถ้าการกำจัดขยะส่วนใหญ่โดยการโยนในแม่น้ำลำห้วยหรือทะเล
64	elimbasu6	=1 if rubbish disposal mainly other	=1 ถ้าการกำจัดขยะอื่น ๆ ส่วนใหญ่

ตาราง 17 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 6)

65	epared1	=1 if walls are bad	=1 ถ้ากำแพงไม่ได้มาตรฐาน
66	epared2	=1 if walls are regular	=1 ถ้าผนังเป็นปกติ
67	epared3	=1 if walls are good	=1 ถ้ากำแพงได้มาตรฐานดีมาตร
68	etecho1	=1 if roof are bad	=1 ถ้าหลังคาไม่ได้มาตรฐาน
69	etecho2	=1 if roof are regular	=1 ถ้าหลังคาเป็นได้ปกติ
70	etecho3	=1 if roof are good	=1 ถ้าหลังคาได้มาตรฐานดีมาก
71	eviv1	=1 if floor are bad	=1 ถ้าพื้นไม่ได้มาตรฐาน
72	eviv2	=1 if floor are regular	=1 ถ้าพื้นเป็นปกติ
73	eviv3	=1 if floor are good	=1 ถ้าพื้นได้มาตรฐานดีมาก
74	Dis	=1 if disable person	=1 ถ้าไม่มีข้อมูลคนนี้
75	male	=1 if male	=1 ถ้าเป็นผู้ชาย
76	female	=1 if female	=1 ถ้าเป็นผู้หญิง
77	estadocivil1	=1 if less than 10 years old	=1 ถ้าอายุน้อยกว่า 10 ปี
78	estadocivil2	=1 if free or coupled union	=1 ถ้าไม่มีความสัมพันธ์
79	estadocivil3	=1 if married	=1 ถ้าแต่งงานแล้ว
80	estadocivil4	=1 if divorced	=1 ถ้าหย่าร้าง
81	estadocivil5	=1 if separated	=1 ถ้าแยกทางกัน
82	estadocivil6	=1 if widow/er	=1 ถ้าเป็นแม่ม่าย / พ่อม่าย
83	estadocivil7	=1 if single	=1 ถ้าโสด
84	parentesco1	=1 if household head	=1 ถ้าเป็นหัวหน้าครอบครัว
85	parentesco2	=1 if spouse/partner	=1 ถ้าเป็นคู่สมรส / คู่
86	parentesco3	=1 if son/daughter	=1 ถ้าเป็นลูกชาย/ลูกสาว
87	parentesco4	=1 if stepson/daughter	=1 ถ้าเป็นลูกชาย/ลูกสาวของคู่สมรส

ตาราง 18 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 7)

88	parentesco5	=1 if son/daughter in law	=1 ถ้าเป็นลูกชาย/ลูกสาวของคู่สมรส
89	parentesco6	=1 if grandson/daughter	=1 ถ้าเป็นหลานชาย / ลูกสาว
90	parentesco7	=1 if mother/father	=1 ถ้าเป็นพ่อ/แม่
91	parentesco8	=1 if father/mother in law	=1 ถ้าเป็นพ่อ/แม่ คู่สมรส
92	parentesco9	=1 if brother/sister	=1 ถ้าเป็นพี่ชาย / น้องสาว
93	parentesco10	=1 if brother/sister in law	=1 ถ้าเป็นพี่ชาย/น้องสาว คู่สมรส
94	parentesco11	=1 if other family member	=1 ถ้าเป็นสมาชิกในครอบครัวสถานะอื่นๆ
95	parentesco12	=1 if other non family member	=1 ถ้าไม่ใช่สมาชิกในครอบครัว
96	idhogar	Household level identifier	ตัวระบุที่ไม่ซ้ำระดับครัวเรือน
97	hogar_nin	Number of children 0 to 19 in household	จำนวนเด็ก 0 ถึง 19 ในครัวเรือน
98	hogar_adul	Number of adults in household	จำนวนผู้ใหญ่ในครัวเรือน
99	hogar_mayor	# of individuals 65+ in the household	จำนวนของบุคคลที่อายุ 65+ ในครัวเรือน
100	hogar_total	# of total individuals in the household	จำนวนของบุคคลทั้งหมดในครัวเรือน
101	dependency	Dependency rate	อัตราการพึ่งพา((จำนวนสมาชิกในครัวเรือนอายุน้อยกว่า 19 ปีหรือมากกว่า 64 ปี) / (จำนวนสมาชิกในครัวเรือนระหว่าง 19 ถึง 64))

ตาราง 19 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 8)

102	edjefe	years of education of male head of household	จำนวนปีการศึกษาของหัวหน้าครอบครัวชาย
103	edjefa	years of education of female head of household	จำนวนปีการศึกษาของหัวหน้าครอบครัวหญิง
104	meaneduc	average years of education for adults (18+)	จำนวนปีการศึกษาโดยเฉลี่ยสำหรับผู้ใหญ่ (18+)
105	instlevel1	=1 no level of education	=1 ไม่มีระดับการศึกษา
106	instlevel2	=1 incomplete primary	=1 ไม่จบการศึกษาระดับประถม
107	instlevel3	=1 complete primary	=1 จบการศึกษาระดับประถม
108	instlevel4	=1 incomplete academic secondary level	=1 ไม่จบศึกษาระดับมัธยม
109	instlevel5	=1 complete academic secondary level	=1 จบศึกษาระดับมัธยม
110	instlevel6	=1 incomplete technical secondary level	=1 ไม่จบการศึกษาระดับมัธยมศึกษาทางเทคนิค
111	instlevel7	=1 complete technical secondary level	=1 จบการศึกษาระดับมัธยมศึกษาทางเทคนิค
112	instlevel8	=1 undergraduate and higher education	=1 จบการศึกษาระดับปริญญาตรี
113	instlevel9	=1 postgraduate higher education	=1 จบการศึกษาระดับสูงกว่าปริญญาตรี
114	bedrooms	number of bedrooms	จำนวนห้องนอน
115	overcrowding	# persons per room	จำนวนคนต่อห้อง
116	tipovivi1	=1 own and fully paid house	=1 บ้านของตัวเองและชำระเต็มจำนวน

ตาราง 20 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 9)

117	tipovivi2	=1 own, paying in installments	=1 เป็นเจ้าของจ่ายเป็นงวด
118	tipovivi3	=1 rented	=1 ให้เช่า
119	tipovivi4	=1 precarious	=1 ล่อแหลม
120	tipovivi5	=1 other(assigned, borrowed)	=1 อื่น ๆ (มอบหมายให้ยืม)
121	computer	=1 if the household has notebook or desktop computer	=1 ถ้าครัวเรือนมีคอมพิวเตอร์โน๊ตบุ๊คหรือคอมพิวเตอร์ตั้งโต๊ะ
122	television	=1 if the household has TV	=1 ถ้าครัวเรือนมีทีวี
123	mobilephone	=1 if mobile phone	=1 ถ้าโทรศัพท์มือถือ
124	qmobilephone	# of mobile phones	# ของโทรศัพท์มือถือ
125	lugar1	=1 region Central	=1 อยู่ในพื้นที่ Central
126	lugar2	=1 region Chorotega	=1 อยู่ในพื้นที่ Chorotega
127	lugar3	=1 region PacÃƒÆ'Ã‚Â‚Ã‚fico central	=1 อยู่ในพื้นที่ PacÃƒÆ'Ã‚Â‚Ã‚-fico central
128	lugar4	=1 region Brunca	=1 อยู่ในพื้นที่ Brunca
129	lugar5	=1 region Huetar AtlÃƒÆ'Ã‚Â‚Ã‚¡ntica	=1 อยู่ในพื้นที่ Huetar AtlÃƒÆ'Ã‚Â‚Ã‚¡ntica
130	lugar6	=1 region Huetar Norte	=1 อยู่ในพื้นที่ Huetar Norte
131	area1	=1 zona urbana	=1 อยู่ในบริเวณ urbana
132	area2	=2 zona rural	=2 อยู่ในบริเวณ rural
133	age	Age in years	อายุ(ปี)
134	SQBescolari	Escolari(years of schooling) squared	ค่าจำนวนปีในการศึกษายกกำลังสอง
135	SQBage	age squared	ค่าอายุ(ปี)ยกกำลังสอง

ตาราง 21 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description) (ต่อ 10)

136	SQBhogar_total	hogar_total(of total individuals in the household) squared	ค่าจำนวนของบุคคลทั้งหมดในครัวเรือนยกกำลังสอง
137	SQBedjefe	Edjefe(years of education of male head of household) squared	ค่าจำนวนปีการศึกษาของหัวหน้าครอบครัวชายยกกำลังสอง
138	SQBhogar_nin	hogar_nin(Number of children 0 to 19 in household) squared	ค่าจำนวนเด็ก 0 ถึง 19 ในครัวเรือนยกกำลังสอง
139	SQBovercrowding	Overcrowding(persons per room) squared	ค่าจำนวนคนต่อห้องยกกำลังสอง
140	SQBdependency	Dependency(Dependency rate) squared	ค่าอัตราการพึ่งพายกกำลังสอง
141	SQBmeaned	Meaned(average years of education for adults (18+)) squared	ค่าจำนวนปีการศึกษาโดยเฉลี่ยสำหรับผู้ใหญ่ (18+)ยกกำลังสอง
142	agesq	Age(Age in years) squared	ค่าอายุ(ปี)ยกกำลังสอง
143	Target	ordinal variable indicating groups of income level 1 = extreme poverty 2 = moderate poverty 3 = vulnerable households 4 = non vulnerable households	ระดับยากจนของแต่ละบุคคล 1 คือบุคคลยากจนขั้นรุนแรง 2 คือบุคคลยากจนปานกลาง 3 คือบุคคลเสี่ยงจะยากจน 4 คือบุคคลไม่เสี่ยงจะยากจน

3.2.3 **คอลัมน์สำคัญที่ใช้ในการพัฒนาแบบจำลอง**

3.2.3.1 **คอลัมน์** Target

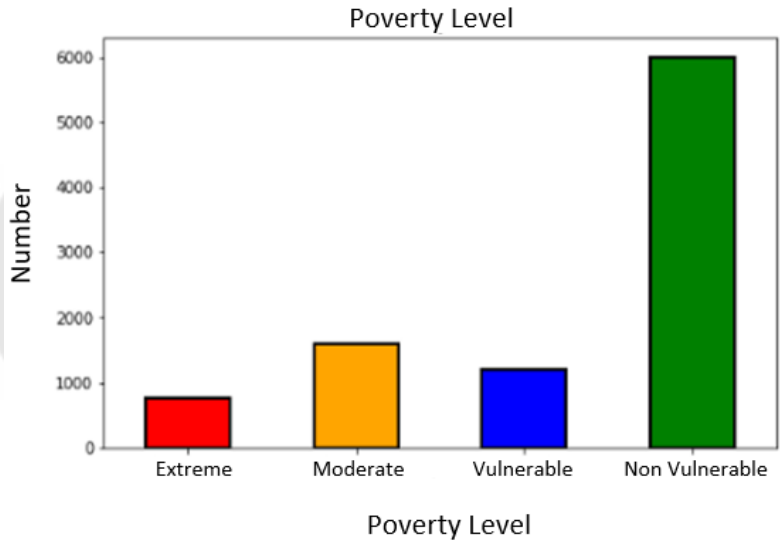
คือ ตัวแปรประเภทจำนวนเต็ม (Int64) ซึ่งถูกกำหนดเป็นผลลัพธ์ในการทำนาย ประกอบไปด้วยข้อมูล 4 กลุ่มแสดงข้อมูลระดับความยากจนในครัวเรือน ซึ่งถูกแบ่งออกเป็น 4 ระดับ ดังภาพประกอบ 11 ซึ่งแต่ละกลุ่มจำนวนข้อมูลและความหมายที่แตกต่างกันดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 755 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 1597 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 1209 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 5996 แถว



ภาพประกอบ 11 จำนวนข้อมูลแต่ละคลาสของคอลัมน์ Target

3.2.3.2 **คอลัมน์** idhogar

ตัวแปรประเภทจำนวนวัตถุ (Object) ซึ่งจะถูกกำหนดให้การระบุถึงแต่ลพครอบครัว โดยแต่ละคนในครอบครัวจะมี idhogar เหมือนกันดังภาพประกอบ 12

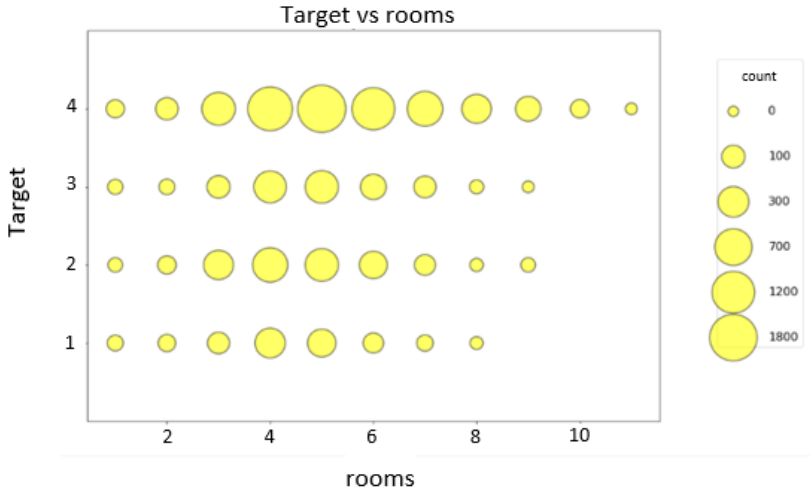
	Id	idhogar
0	ID_279628684	21eb7fcc1
1	ID_f29eb3ddd	0e5d7a658
2	ID_68de51c94	2c7317ea8
3	ID_d671db89c	2b58d945f
4	ID_d56d6f5f5	2b58d945f
5	ID_ec05b1a7b	2b58d945f
6	ID_e9e0c1100	2b58d945f
7	ID_3e04e571e	d6dae86b7
8	ID_1284f8aad	d6dae86b7
9	ID_51f52fdd2	d6dae86b7
10	ID_db44f5c59	d6dae86b7

ภาพประกอบ 12 ตัวอย่างชุดข้อมูลของคอลัมน์ idhogar

3.2.4 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และคอลัมน์อื่น ๆในชุดข้อมูล

3.2.4.1 คอลัมน์ Target และ rooms

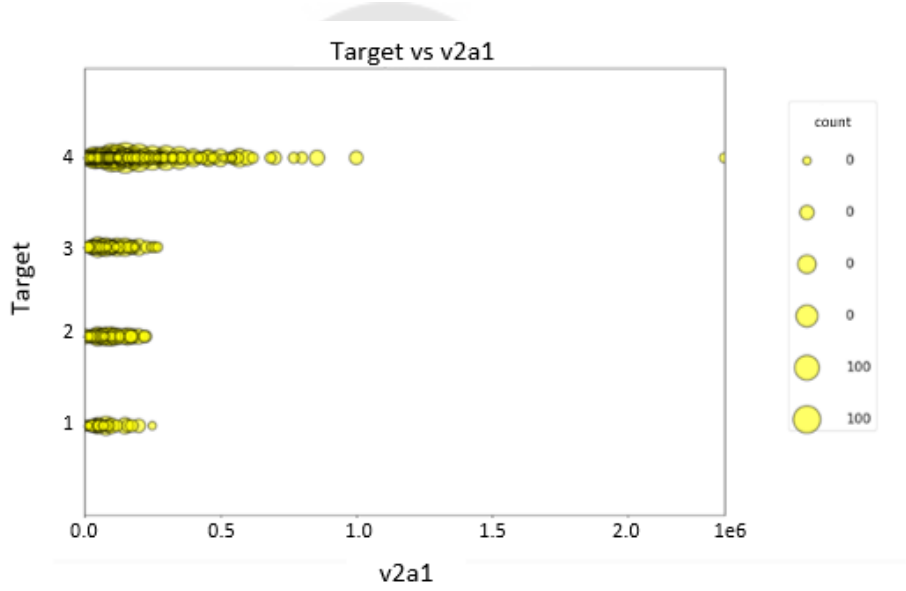
เป็นความสัมพันธ์ระหว่าง Target (Ordinal variable indicating groups of income level) คือระดับยากจนของแต่ละบุคคล กับคอลัมน์ Rooms (number of all rooms in the house) จำนวนห้องทั้งหมดในบ้าน จะเห็นได้ว่าในบุคคลยากจนขั้นรุนแรง (Target 1) มีจำนวนห้อง 2-8 ห้อง, บุคคลยากจนปานกลาง(Target 2) มีจำนวนห้อง 2-9 ห้อง, บุคคลเสี่ยงจะยากจน (Target 3) มีจำนวนห้อง 2-9 ห้อง และบุคคลไม่เสี่ยงจะยากจน (Target 4) มีจำนวนห้อง 2-11 ห้อง ดังภาพประกอบ 13



ภาพประกอบ 13 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และ Rooms

3.2.4.2 **คอลัมน์** Target **และ** v2a1

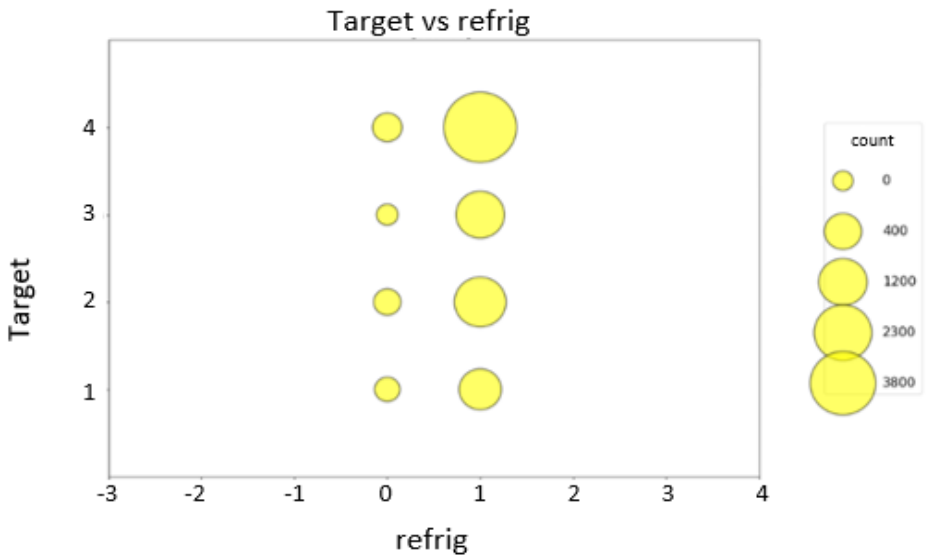
เป็นความสัมพันธ์ระหว่าง Target (Ordinal variable indicating groups of income level) คือระดับยากจนของแต่ละบุคคลกับคอลัมน์ v2a1 (Monthly rent payment) ค่าเช่าที่พักรายเดือน จะเห็นได้ว่าในบุคคลยากจนขั้นรุนแรง (Target 1) ,บุคคลยากจนปานกลาง (Target 2) บุคคลเสี่ยงจะยากจน (Target 3) มีค่าเช่าที่พักรายเดือนประมาณ 0 ถึง 300,000 และบุคคลไม่เสี่ยงจะยากจน (Target 4) ค่าเช่าที่พักรายเดือนประมาณ 0 ถึง 1,000,000 ดังภาพประกอบ 14



ภาพประกอบ 14 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และ V2a1

3.2.4.3 **คอลัมน์** Target **และ** refrig

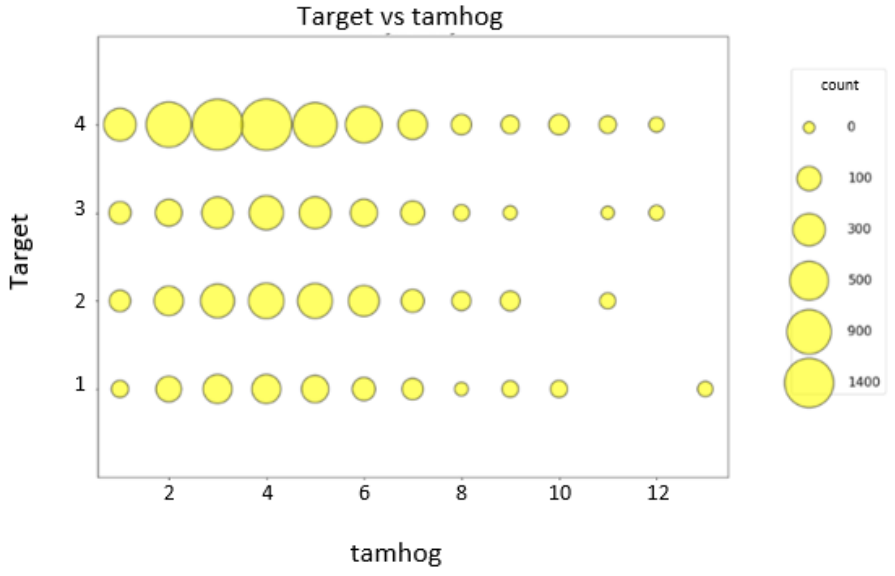
เป็นความสัมพันธ์ระหว่าง Target (Ordinal variable indicating groups of income level) คือระบุระดับยากจนของแต่ละบุคคลกับคอลัมน์ refrig (1 if the household has refrigerator) คือ 1 มีตู้เย็นในครอบครัว , 0 ไม่มีตู้เย็นในครอบครัว ซึ่งบุคคลทั้ง 4 ระดับความยากจน (Target 1-4) ครอบครัวมีจำนวนตู้เย็นมากกว่าครอบครัวที่ไม่มีจำนวนตู้เย็น ดังภาพประกอบ 15



ภาพประกอบ 15 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และ refrig

3.2.4.4 **คอลัมน์** Target **และ** tamhog

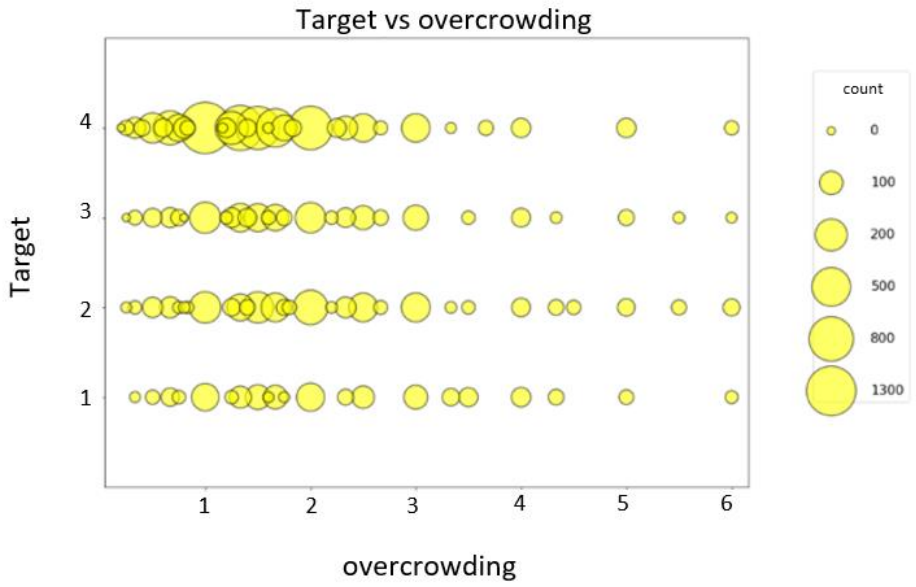
เป็นความสัมพันธ์ระหว่าง Target (Ordinal variable indicating groups of income level) คือระบุระดับยากจนของแต่ละบุคคลกับคอลัมน์ tamhog (size of the household) ขนาดของครอบครัว ซึ่งบุคคลทั้ง 4 ระดับความยากจน (Target 1-4) มีจำนวนสมาชิกในครอบอยู่ระหว่าง 1 ถึง 12 คน ดังภาพประกอบ 16



ภาพประกอบ 16 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และ tamhog

3.2.4.5 **คอลัมน์** Target **และ** overcrowding

เป็นความสัมพันธ์ระหว่าง Target (Ordinal variable indicating groups of income level) คือระบุระดับยากจนของแต่ละบุคคลกับคอลัมน์ overcrowding (persons per room) คือจำนวนคนต่อห้องซึ่งบุคคลทั้ง 4 ระดับความยากจน (Target 1-4) มีจำนวนสมาชิกที่พักอาศัยในห้องเฉลี่ยอยู่ที่ 0.1 ถึง 12 คน ดังภาพประกอบ 17



ภาพประกอบ 17 การแสดงความสัมพันธ์ระหว่างคอลัมน์ Target และ overcrowding

3.3 การเตรียมข้อมูล

การเตรียมข้อมูล แบ่งออกเป็นการทำความสะอาดข้อมูล (Cleansing Data), การจัดรูปแบบข้อมูล และการวิเคราะห์ข้อมูล โดยทำให้ชุดข้อมูลสามารถนำไปใช้งานได้

3.3.1 การทำความสะอาดข้อมูล (Cleansing Data)

เป็นขั้นตอนการตรวจสอบและการแก้ไขข้อมูลที่ไม่ถูกต้อง การทำความสะอาดข้อมูลนั้นเกิดขึ้นเนื่องจากมีความไม่สอดคล้องกันระหว่างข้อมูลและคุณลักษณะ ซึ่งเกิดจากข้อผิดพลาดของการบันทึกข้อมูล การส่งข้อมูลหรือการให้ความหมายของข้อมูลที่แตกต่างกัน

3.3.1.1 การหาค่าที่ขาดหายไปในข้อมูล (Data Missing)

การหาค่าที่ขาดหายไปในข้อมูลและกำหนดวิธีจัดการกับข้อมูลเหล่านี้ ค่าที่ขาดหายไปจะต้องจัดการให้เรียบร้อยก่อนที่เราจะนำไปใช้ในการพัฒนาแบบจำลอง โดยจากภาพประกอบ18 จะพบว่ามีคอลัมน์ที่มีข้อมูลขาดหายคือ rez_esc,(Years behind in school) , v2a1 (Monthly rent payment) ,v18q1 (Number of tablets)

	total	percent
rez_esc	7928	0.829549
v18q1	7342	0.768233
v2a1	6860	0.717798
SQBmeaned	5	0.000523
meaneduc	5	0.000523
Id	0	0.000000
hogar_adul	0	0.000000
parentesco10	0	0.000000
parentesco11	0	0.000000
parentesco12	0	0.000000

ภาพประกอบ 18 แสดงค่าที่ขาดหายไปในข้อมูล (Data Missing)

rez_esc (years behind in school) คือคอลัมน์ที่บ่งบอกเกี่ยวกับจำนวนปีการเข้าเรียนช้ากว่ากำหนด จากการวิเคราะห์ข้อมูล เมื่อเปรียบเทียบระหว่าง rez_esc และ age (Age in years) ดังภาพประกอบ 19 จะพบว่าข้อมูลที่มีอยู่จะอยู่ในช่วงอายุ 7 ถึง 17 ปี ดังนั้นข้อมูลที่หายไปคือบุคคลที่มีอายุน้อยกว่า 7 ปี และมากกว่า 17 ปี คือบุคคลที่ยังไม่ถึงเกณฑ์เข้าศึกษาหรือจบการศึกษาไปแล้วดังนั้นจะทำการแทนค่า 0 ลงไปแทนที่ข้อมูลที่หายไป

```
count    1629.000000
mean       12.258441
std         3.218325
min         7.000000
25%         9.000000
50%        12.000000
75%        15.000000
max        17.000000
Name: age, dtype: float64
```

ภาพประกอบ 19 แสดงเปรียบเทียบระหว่างคอลัมน์ rez_esc และคอลัมน์ age

v2a1 (Monthly rent payment) คือคอลัมน์ที่บ่งบอกเกี่ยวกับค่าเช่ารายเดือนของที่พัก จากการวิเคราะห์ข้อมูลเมื่อเปรียบเทียบระหว่างข้อมูลที่หายไปในคอลัมน์ v2a1 และ tipovivi1 (own and fully paid house), tipovivi2 (paying in installments), tipovivi3 (rented), tipovivi4 (precarious), tipovivi5 (other) จะพบว่าข้อมูลที่หายไปในคอลัมน์ v2a1 จะตรงกับ tipovivi1, tipovivi4 , tipovivi5 ดังภาพประกอบ 20 ดังนั้นจะทำการแทนค่า 0 ลงไปแทนที่ข้อมูลที่หายไปเพราะถือว่าบุคคล ไม่ได้มีการจ่ายค่าเช่ารายเดือนของที่พักเพราะอาจจะเป็นเจ้าของได้

```
tipovivi1    5911
tipovivi2       0
tipovivi3       0
tipovivi4     163
tipovivi5     786
dtype: int64
```

ภาพประกอบ 20 แสดงการเปรียบเทียบข้อมูลที่หายไปในคอลัมน์ v2a1 และ tipovivi1-5

v18q1 (Number of tablets) คือคอลัมน์ที่บ่งบอกเกี่ยวกับจำนวนของ Tablets ของแต่ละบุคคล จากการวิเคราะห์ข้อมูลโดยทำการตรวจสอบจำนวน Tablets จะพบว่าจำนวนของ Tablets เป็นจำนวนเต็มตั้งแต่ 1-6 ดังภาพประกอบ 21 ดังนั้นข้อมูลที่หายไปคือบุคคลที่ไม่มี Tablets จึงจะทำการแทนค่า 0 ลงไปแทนที่ข้อมูลที่หายไป

```
1.0    1586
2.0     444
3.0     129
4.0      37
5.0      13
6.0       6
Name: v18q1, dtype: int64
```

ภาพประกอบ 21 แสดงผลรวมจำนวนของ Tablets

Meaneduc (average years of education for adults (18+)) คือคอลัมน์ที่บ่งบอกเกี่ยวกับค่าเฉลี่ยอายุของการศึกษาของผู้ใหญ่ จากการวิเคราะห์ข้อมูลโดยการเปรียบเทียบระหว่าง Meaneduc และ Age ข้อมูลที่หายไปจะอยู่ในช่วงอายุ 18 และ 19 ปี ดังภาพประกอบ 22 ซึ่งไม่สามารถกำหนดค่าข้อมูลที่หายไปได้ เพราะข้อมูลที่มีในคอลัมน์ Meaneduc มีการกระจายของข้อมูลอยู่กับทุกช่วงค่าอายุดังภาพประกอบ 23 ดังนั้นจะทำการลบข้อมูลที่หายไป

```
count     5.000000
mean     18.400000
std       0.547723
min      18.000000
25%      18.000000
50%      18.000000
75%      19.000000
max      19.000000
Name: age, dtype: float64
```

ภาพประกอบ 22 แสดงการเปรียบเทียบระหว่างคอลัมน์ Meaneduc และ Age

```
0.000000     71
0.333333      3
0.500000     19
0.666667     10
1.000000     42
             ..
27.000000     2
28.000000     5
29.000000     7
32.000000     2
37.000000     3
Name: meaneduc, Length: 155, dtype: int64
```

ภาพประกอบ 23 การกระจายของข้อมูลของคอลัมน์ Meaneduc

3.3.1.2 **การหาแจกแจงคอลัมน์วัตถุ** (Explain Column Object)

ในการพัฒนาแบบจำลอง ชุดของข้อมูลที่นำมาใช้ในการพัฒนาควรจะเป็นประเภทจำนวนเต็ม (Integer) และจำนวนจริง (Float) เพื่อให้ง่ายต่อการนำไปคำนวนร่วมกับอัลกอริทึม จึงต้องทำการแจกแจงชุดข้อมูลที่มีความไม่สอดคล้องกับความหมายของแต่ละคอลัมน์ (data description) โดยการแสดงชนิดของคอลัมน์ที่ไม่ใช่จำนวนเต็ม (Integer) และจำนวนจริง (Float) จะพบว่าชุดข้อมูลนี้ประกอบไปด้วยข้อมูลที่เป็น Object จำนวน 5 คอลัมน์ คือ Id , idhogar , dependency , edjefe , edjefa ดังภาพประกอบ 24

	Id	idhogar	dependency	edjefe	edjefa
0	ID_279628684	21eb7fcc1	no	10	no
1	ID_f29eb3ddd	0e5d7a658	8	12	no
2	ID_68de51c94	2c7317ea8	8	no	11
3	ID_d671db89c	2b58d945f	yes	11	no
4	ID_d56d6f5f5	2b58d945f	yes	11	no

ภาพประกอบ 24 แสดงคอลัมน์ที่เป็น Object

คอลัมน์ id (a unique identifier for each row) คือ ตัวระบุที่ไม่ซ้ำกันสำหรับแต่ละแถว และ idhogar (a unique identifier for each household) คือ เป็นตัวระบุที่ไม่ซ้ำกันสำหรับแต่ละครอบ ครัว ซึ่งข้อมูลของทั้งสองคอลัมน์เป็นตัวอักษรทั้งหมด จึงเหมาะสมที่จะเป็นชนิดของคอลัมน์ Object อย่างไรก็ตามคอลัมน์อื่น ๆ ที่เหลือจะเป็นการผสมผสานระหว่างตัวอักษรและตัวเลข ดังนั้นจึงจะจัดการแก้ไขข้อมูลให้ถูกต้อง

Dependency (Dependency rate (number younger than 19 or older than 64)/ (between 19 and 64) คอลัมน์ที่เกี่ยวกับอัตราส่วนระหว่างบุคคลภายในครอบครัวที่อายุน้อยกว่า 19 หรือมากกว่า 64 ต่อบุคคลที่อายุระหว่าง 19 ถึง 64 ข้อมูลคอลัมน์นี้ประกอบไปด้วยจำนวนจริง และ ข้อความ (Yes ,No) ถ้าเป็นอัตราส่วนดังนั้นควรจะเป็นจำนวนเต็มเท่านั้นไม่ควร

เป็นข้อความ จากภาพประกอบ 25 เป็นการเปรียนเทียบระหว่างคอลัมน์ที่มีความเกี่ยวข้องกันกับ dependency คือ age และ idhogar พบว่า No คือ ครอบครัวที่มีจำนวนสามารถที่ไม่สามารถนำมาคำนวนอัตราส่วนได้ เช่น มีสมาชิกคนเดียว หรือ สมาชิกอยู่ในช่วงอายุ 19 ถึง 64 ทุกคน และ Yes คือครอบครัวที่อัตราส่วนของสมาชิกที่อายุน้อยกว่า19 หรือมากกว่า 64 เท่ากันกับอายุระหว่าง 19 ถึง 64 ดังนั้นจะทำการแทนค่า Yes = 1 และ No = 0

	idhogar	dependency	age
0	21eb7fcc1	no	43
1	0e5d7a658	8	67
2	2c7317ea8	8	92
3	2b58d945f	yes	17
4	2b58d945f	yes	37
5	2b58d945f	yes	38
6	2b58d945f	yes	8
7	d6dae86b7	yes	7
8	d6dae86b7	yes	30
9	d6dae86b7	yes	28
10	d6dae86b7	yes	11
11	bb2094100	yes	18
12	bb2094100	yes	34

ภาพประกอบ 25 การเปรียนเทียบระหว่างคอลัมน์ dependency , age และ idhogar

edjefe (years of education of male head of household) ปีการศึกษาหัวหน้าครัวเรือนชาย , edjefa (years of education of female head of household) ปีการศึกษาหัวหน้าครัวเรือนหญิง จากตัวอย่างข้อมูลดังภาพประกอบ 26 จะพบว่าข้อมูลระหว่าง 2 คอลัมน์นี้จะมีความสอดคล้องกัน ถ้าคอลัมน์ใดมีข้อมูลอีกคอลัมน์จะเป็น No เพราะทั้ง 2 คอลัมน์นี้เป็นข้อมูลหัวหน้าครอบครัวของบุคคลนั้นๆ ซึ่งแต่ละครอบครัวจะมีหัวหน้าครอบครัวเพียงคนเดียว ดังนั้นจะทำการแทนค่าที่เป็น No = 0

	edjefe	edjefa
0	10	no
1	12	no
2	no	11
3	11	no
4	11	no
5	11	no
6	11	no
7	9	no
8	9	no
9	9	no

ภาพประกอบ 26 การเปรียนเทียบระหว่างคอลัมน์ edjefe กับ edjefa

3.3.2 **การวิเคราะห์ข้อมูลเชิงสำรวจ** (Exploratory Data Analysis)

การวิเคราะห์ข้อมูลเพื่อทำความเข้าใจรายละเอียดภายในของข้อมูล ก่อนที่จะเข้าสู่กระบวนการนำข้อมูลไปพัฒนาแบบจำลองในลำดับต่อไป

3.3.2.1 **การกำหนดประเภทของกลุ่มคอลัมน์** (Define Column Categories)

เป็นการแยกข้อมูลเป็นกลุ่มๆ โดยแต่ละกลุ่มจะมีคุณลักษณะที่มีความคล้ายคลึงกัน เพื่อให้ง่ายในการจัดการข้อมูลหรือลบกลุ่มข้อมูลที่ไม่มีความจำเป็นในการพัฒนาแบบจำลอง โดยแบ่งเป็น 4 กลุ่ม คือ กลุ่มข้อมูลระบุตัวตน กลุ่มข้อมูลของแต่ละบุคคล กลุ่มข้อมูลเกี่ยวกับครอบครัว กลุ่มสุดท้ายกลุ่มข้อมูลยกกำลังสอง

1. กลุ่มข้อมูลระบุตัวตน เช่น Id, idhogar, Target เป็นต้น ซึ่งเป็นข้อมูลที่ระบุความเป็นตัวตนที่ไม่ซ้ำกันของแต่ละบุคคล จึงเป็นข้อมูลสำคัญไม่ควรลบหรือแก้ไข

2. กลุ่มข้อมูลของแต่ละบุคคล เช่น male, female, age เป็นต้น ซึ่งเป็นข้อมูลของแต่ละบุคคล จึงสามารถนำมาวิเคราะห์หรือจัดการข้อมูลก่อนนำไปใช้งานได้

3. กลุ่มข้อมูลเกี่ยวกับครอบครัว เช่น rooms, area1, bedrooms, refrig เป็นต้น ซึ่งเป็นข้อมูลเกี่ยวกับครอบครัวนั้นๆ จึงสามารถนำมาวิเคราะห์หรือจัดการข้อมูลก่อนนำไปใช้งานได้

4. กลุ่มข้อมูลยกกำลังสอง เช่น SQBmeaned, agesq, SQBdependency เป็นต้น ซึ่งเป็นข้อมูลต่างๆ ที่เป็นจำนวนจริงในชุดข้อมูลมายกกำลังสอง

3.3.2.2 **ค่าสัมประสิทธิ์สหสัมพันธ์** (Correlation coefficient)

การวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ เป็นขั้นตอนเปรียบเทียบความสัมพันธ์ของตัวแปรตั้งแต่ 2 ตัวขึ้นไป ว่ามีความสัมพันธ์กันมากน้อยเพียงใด

1. การหาค่าสัมประสิทธิ์สหสัมพันธ์ในกลุ่มข้อมูลของแต่ละบุคคล จะพบว่าข้อมูล female มีค่าสัมประสิทธิ์สหสัมพันธ์สูงมากกว่า 0.95 ดังนั้นจึงนำข้อมูล female หาความสัมพันธ์กับข้อมูลอื่นๆ พบว่า ข้อมูล female มีค่าความสัมพันธ์ตรงข้ามกับ male ดังภาพประกอบ 27 เนื่องจากบุคคลนั้นจะมีเพศได้เพียง 1 เพศเท่านั้น จึงทำการลบข้อมูล male ออก และทำการเปลี่ยนชื่อคอลัมน์ female เป็น gender และกำหนด ให้ 1 = female, 0 = male

	male	female
male	1.0	-1.0
female	-1.0	1.0

ภาพประกอบ 27 ความสัมพันธ์ระหว่างคอลัมน์ female และ male

2. การหาค่าสัมประสิทธิ์สหสัมพันธ์ในกลุ่มข้อมูลเกี่ยวกับครอบครัว จะพบว่ามีข้อมูล coopele, area2, tamhog, hhsize, hogar_total มีค่าสัมประสิทธิ์สหสัมพันธ์สูงมากกว่า 0.95 ดังนั้นจึงนำข้อมูลเหล่านี้มา หาความสัมพันธ์กับข้อมูลอื่นๆ

ข้อมูล coopele (electricity from cooperative) ค่าความสัมพันธ์ตรงข้ามกับข้อมูล public (electricity from CNFL, ICE, ESPH/JASEC") ดังภาพประกอบ 28 แต่เมื่อตรวจสอบความหมายของข้อมูล (data descriptions) ของทั้ง coopele และ public ซึ่งทั้ง 2 คอลัมน์เป็นข้อมูลเกี่ยวกับแหล่งกำเนิดไฟฟ้า (electricity) ยังพบว่ามีข้อมูลอื่นๆ ที่อยู่ในกลุ่มข้อมูลนี้อีก ดังนั้นจึงไม่สามารถแก้ไขข้อมูลกลุ่มนี้ได้ เพราะอาจจะส่งผลกระทบต่อชุดข้อมูล

	public	coopele
public	1.000000	-0.979822
coopele	-0.979822	1.000000

ภาพประกอบ 28 ความสัมพันธ์ระหว่างคอลัมน์ coopele และ public

ข้อมูล area2 (zona rural) ค่าความสัมพันธ์ตรงข้ามกับข้อมูล area1 (zona urbana) ดังภาพประกอบ 29 ซึ่งทั้ง 2 คอลัมน์เป็นข้อมูลเกี่ยวกับภูมิภาคที่อยู่อาศัย โดยแบ่งเป็น 2 ส่วนคือ area1,area2 จึงทำการลบข้อมูล area2 ออกและทำการเปลี่ยนชื่อคอลัมน์ area1 เป็น area และกำหนด ให้ 1 = zona urbana , 0 = zona rural

	area1	area2
area1	1.0	-1.0
area2	-1.0	1.0

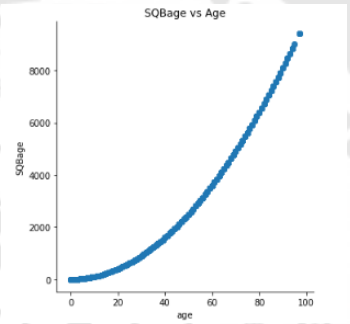
ภาพประกอบ 29 ความสัมพันธ์ระหว่างคอลัมน์ area1 และ area2

ข้อมูล tamhog (size of the household) มีค่าความสัมพันธ์กับข้อมูล r4t3 (Total persons in the household), tamhog (size of the household) , hhsize (household size), hogar_total (total individuals in the household) โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ สูงมากกว่า 0.95 ข้อมูลเหล่านี้มีความสัมพันธ์กันอยู่ ดังภาพประกอบ 30 ซึ่งทั้ง 5 คอลัมน์เป็นข้อมูลเกี่ยว กับจำนวนสมาชิกในครอบครัว ดังนั้นจะทำการลบข้อมูล r4t3 , hhsize , hogar_total ออกจากชุดข้อมูลและคงเหลือ tamhog ไว้เท่านั้น

	r4t3	tamhog	hhsize	hogar_total
r4t3	1.000000	0.998106	0.998106	0.998106
tamhog	0.998106	1.000000	1.000000	1.000000
hhsize	0.998106	1.000000	1.000000	1.000000
hogar_total	0.998106	1.000000	1.000000	1.000000

ภาพประกอบ 30 ความสัมพันธ์ระหว่างคอลัมน์ r4t3, tamhog ,hhsize ,hogar_total

3. การหาค่าสัมประสิทธิ์สหสัมพันธ์ในกลุ่มข้อมูลยกกำลังสอง ซึ่งเป็นข้อมูลต่างๆ ที่เป็นจำนวนจริงในชุดข้อมูลมายกกำลังสอง โดยการหาความสัมพันธ์จะพบว่าข้อมูลกลุ่มนี้จะมีค่าสัมประสิทธิ์สหสัมพันธ์กับค่าที่ถูกนำมายกกำลังสอง ดังภาพประกอบ 31 เป็นข้อมูลเกี่ยวกับ Age จะสัมพันธ์กับข้อมูล SQBage ดังนั้นจะทำการลบข้อมูลกลุ่มยกกำลังสองออกจากชุดข้อมูล



ภาพประกอบ 31 ค่าสัมประสิทธิ์สหสัมพันธ์ของกลุ่มข้อมูลยกกำลังสอง

หลังจากผ่านขั้นตอนหาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) จะได้ชุดข้อมูลจำนวน 9551 แถวและ 129 คอลัมน์ ดังภาพประกอบ 32

	Id	v2a1	hacdor	rooms	hacapo	v14a	refrig	v18q	v18q1	r4h1	r4h2	r4h3	r4m1	r4m2	r4m3	r4t1	r4t2	tamviv
0	ID_279628684	190000.0	0	3	0	1	1	0	0.0	0	1	1	0	0	0	0	1	1
1	ID_f29eb3ddd	135000.0	0	4	0	1	1	1	1.0	0	1	1	0	0	0	0	1	1
2	ID_68de51c94	0.0	0	8	0	1	1	0	0.0	0	0	0	0	1	1	0	1	1
3	ID_d671db89c	180000.0	0	5	0	1	1	1	1.0	0	2	2	1	1	2	1	3	4
4	ID_d56d6f5f5	180000.0	0	5	0	1	1	1	1.0	0	2	2	1	1	2	1	3	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9552	ID_d45ae367d	80000.0	0	6	0	1	1	0	0.0	0	2	2	1	2	3	1	4	5
9553	ID_c94744e07	80000.0	0	6	0	1	1	0	0.0	0	2	2	1	2	3	1	4	5
9554	ID_85fc658f8	80000.0	0	6	0	1	1	0	0.0	0	2	2	1	2	3	1	4	5
9555	ID_ced540c61	80000.0	0	6	0	1	1	0	0.0	0	2	2	1	2	3	1	4	5
9556	ID_a38c64491	80000.0	0	6	0	1	1	0	0.0	0	2	2	1	2	3	1	4	5

9551 rows × 129 columns

ภาพประกอบ 32 ชุดข้อมูลหลังจากผ่านขั้นตอนหาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient)

3.3.2.3 **การทำวิศวกรรมข้อมูล** (Feature Engineering)

1.การจัดกลุ่มครอบครัว (Group by household)

จากชุดข้อมูลแต่ละบุคคลในที่อยู่ในครอบครัวเดียวกันนั้นมี ข้อมูลลักษณะคล้ายคลึงกันอาจจะทำให้เกิดการแปรปวนของข้อมูล (Bias-Variance) และข้อมูลรั่วไหล(Data Leak) ในขั้นการทำนาย เนื่องจากในครอบครัวนั้นๆ อาจจะถูกแบ่งไปฝึกการทำนายและทดสอบการทำนายซึ่งเป็นวิธีที่ไม่ถูกต้อง เพราะจะเป็นการนำข้อมูลเดิมไปทำการฝึกและทำนาย ดังนั้นจะทำการแก้ไขด้วยการโดยนำสมาชิกแต่ละบุคคลที่อยู่ในครอบครัวเดียวกันมาทำการยุบรวมเพื่อให้เหลือเพียงครอบครัวละ 1 บุคคลเท่านั้น โดยวิธีคือรวมข้อมูลและหาค่าเฉลี่ย (Mean) , ค่าสูงสุด (Max) , ค่าต่ำสุด (Min) ในแต่ละคุณลักษณะของแต่ละครอบครัวซึ่งจะได้ชุดข้อมูลจำนวน 2985 แถวและ 381 คอลัมน์ดังภาพประกอบ 33

	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
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
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
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
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
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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
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
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
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
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

2985 rows × 381 columns

ภาพประกอบ 33 ชุดข้อมูลหลังผ่านการเตรียมข้อมูล

จากการตรวจสอบชุดข้อมูลหลังจากการจัดกลุ่มครอบครัว (Group by household) จะพบว่าคอลัมน์ Target มีการเพิ่มคอลัมน์ Mean , Max , Min ดังภาพประกอบ 34 จากนั้นทำการหาค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ของคอลัมน์ Target-mean , Target-min , Target-max ดังภาพประกอบ 35 ผลที่ได้แสดงให้เห็นว่าทั้ง 3 คอลัมน์มีข้อมูลเหมือนกันดังนั้นจะทำการลบ คอลัมน์ , Target-min , Target-max และเปลี่ยนชื่อ Target-mean เป็น Target ซึ่งได้ชุดข้อมูลจำนวน 2985 แถวและ 196 คอลัมน์

idhogar	Target-mean	Target-min	Target-max
001ff74ca	4	4	4
003123ec2	2	2	2
004616164	2	2	2
004983866	3	3	3
005905417	2	2	2
...	...	...	...
ff9343a35	4	4	4
ff9d5ab17	4	4	4
ffae4a097	4	4	4
ffe90d46f	1	1	1
fff7d6be1	4	4	4

2985 rows × 3 columns

ภาพประกอบ 34 คอลัมน์ Target

	Target-mean	Target-min	Target-max
Target-mean	1.0	1.0	1.0
Target-min	1.0	1.0	1.0
Target-max	1.0	1.0	1.0

ภาพประกอบ 35 ค่าสัมประสิทธิ์สหสัมพันธ์ของคอลัมน์ Target

3.3.2.4 **การเลือกคุณลักษณะ** (Feature Selection)

การเลือกคุณลักษณะเป็นขั้นตอนสำคัญในการเตรียมข้อมูลก่อนการทำเหมืองข้อมูล เป็นเทคนิคการลดขนาดของจำนวนคุณลักษณะ และคัดเลือกคุณลักษณะที่มีความสำคัญต่อการพัฒนาแบบจำลอง โดยเลือกใช้วิธีการทดสอบไคสแควร์ (Chi-Square Score) เป็นการการเปรียบเทียบตัวแปรตั้งแต่ 2 กลุ่มเป็นต้นไปว่ามีความสัมพันธ์กันมากน้อยเพียงใด โดยกำหนด K = 35 คือเลือกคุณลักษณะ (Feature) ที่มีคะแนนสูงสุด 35 อันดับแรก โดยผลการทดลองที่ได้ทั้ง 3 คอลัมน์คือ v2a1-mean', 'v2a1-min', 'v2a1-max', 'v18q1-mean', 'v18q1-min', 'v18q1-max', 'r4m1-mean', 'r4m1-min', 'r4m1-max', 'r4t1-mean', 'r4t1-min', 'r4t1-max', 'escolari-mean', 'escolari-min', 'escolari-max', 'rez_esc-max', 'hogar_nin-mean', 'hogar_nin-min', 'hogar_nin-max', 'dependency-mean', 'dependency-min', 'dependency-max', 'edjefe-mean', 'edjefe-min', 'edjefe-max', 'edjefa-mean', 'edjefa-min', 'edjefa-max', 'meaneduc-mean', 'meaneduc-min', 'meaneduc-max', 'instlevel8-mean', 'instlevel8-max', 'age-mean', 'age-min' ดังภาพประกอบ 36

```
Index(['v2a1-mean', 'v2a1-min', 'v2a1-max', 'v18q1-mean', 'v18q1-min',
       'v18q1-max', 'r4m1-mean', 'r4m1-min', 'r4m1-max', 'r4t1-mean',
       'r4t1-min', 'r4t1-max', 'escolari-mean', 'escolari-min', 'escolari-max',
       'rez_esc-max', 'hogar_nin-mean', 'hogar_nin-min', 'hogar_nin-max',
       'dependency-mean', 'dependency-min', 'dependency-max', 'edjefe-mean',
       'edjefe-min', 'edjefe-max', 'edjefa-mean', 'edjefa-min', 'edjefa-max',
       'meaneduc-mean', 'meaneduc-min', 'meaneduc-max', 'instlevel8-mean',
       'instlevel8-max', 'age-mean', 'age-min'],
      dtype='object')
```

ภาพประกอบ 36 คุณลักษณะ (Feature) ที่มีคะแนนสูงสุด 35 อันดับแรก

จากนั้นสร้างชุดข้อมูลใหม่ที่ประกอบไปด้วยคุณลักษณะ (Feature) ที่มีคะแนนสูงสุด 35 อันดับแรกและคุณลักษณะระดับความยากจน (Target) 1 คอลัมน์ ซึ่งได้ชุดข้อมูลจำนวน 2963 แถวและ 36 คอลัมน์ดังภาพประกอบ 37

	v2a1-mean	v2a1-min	v2a1-max	v18q1-mean	v18q1-min	v18q1-max	r4m1-mean	r4m1-min	r4m1-max	r4t1-mean	r4t1-min	r4t1-max	escolari-mean
idhogar													
001ff74ca	0.0	0.0	0.0	1.0	1.0	1.0	1	1	1	1	1	1	8.000000
003123ec2	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	2	2	2	3.250000
004616164	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0	7.000000
004983866	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0	7.500000
005905417	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	1	1	1	5.666667
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ff9343a35	0.0	0.0	0.0	1.0	1.0	1.0	0	0	0	0	0	0	8.000000
ff9d5ab17	150000.0	150000.0	150000.0	0.0	0.0	0.0	1	1	1	1	1	1	9.000000
ffae4a097	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0	8.500000
ffe90d46f	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	1	1	1	3.000000
fff7d6be1	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0	8.500000

2985 rows × 36 columns

ภาพประกอบ 37 ชุดข้อมูลใหม่หลังจากทำการเลือกคุณลักษณะ (Feature Selection)

3.4 การสร้างแบบจำลอง

หลังจากการผ่านขั้นตอนการเตรียมข้อมูล (Data Preparation) จะได้ชุดข้อมูลจำนวน 2985 แถวและ 36 คอลัมน์ จากนั้นทำการแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train และ Test

3.4.1 การแบ่งข้อมูลสำหรับฝึกและทดสอบ (Train and Test Split)

การแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train และ Test ในอัตราส่วน 70 : 30 โดย Train คือชุดข้อมูลการฝึกสำหรับสร้างแบบจำลอง และ Test คือชุดข้อมูลสำหรับการทดสอบประสิทธิภาพของแบบจำลอง

3.4.1.1 **ข้อมูล** Train

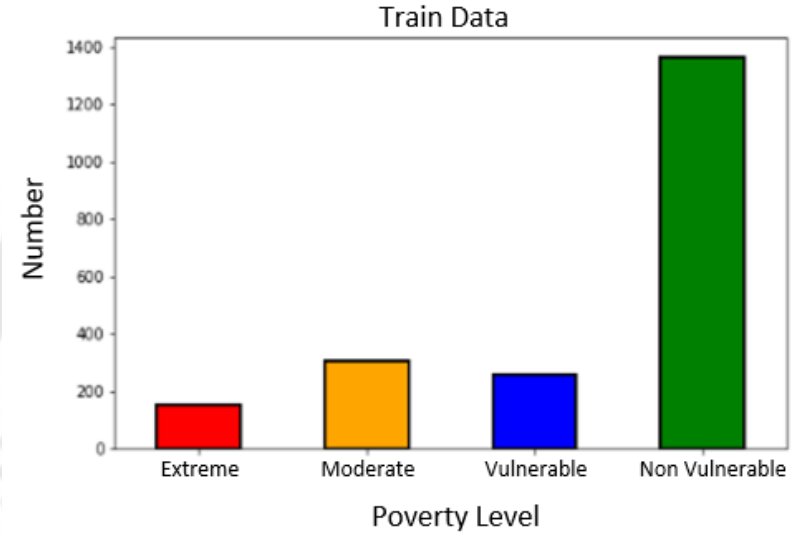
ดังภาพประกอบ 38 จะมีข้อมูล 2089 แถว 36 คอลัมน์โดยแบ่งเป็น

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 156 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 309 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 258 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 1366 แถว



ภาพประกอบ 38 ชุดข้อมูล Train แต่ละระดับความยากจน

3.4.1.2 **ข้อมูล** Test

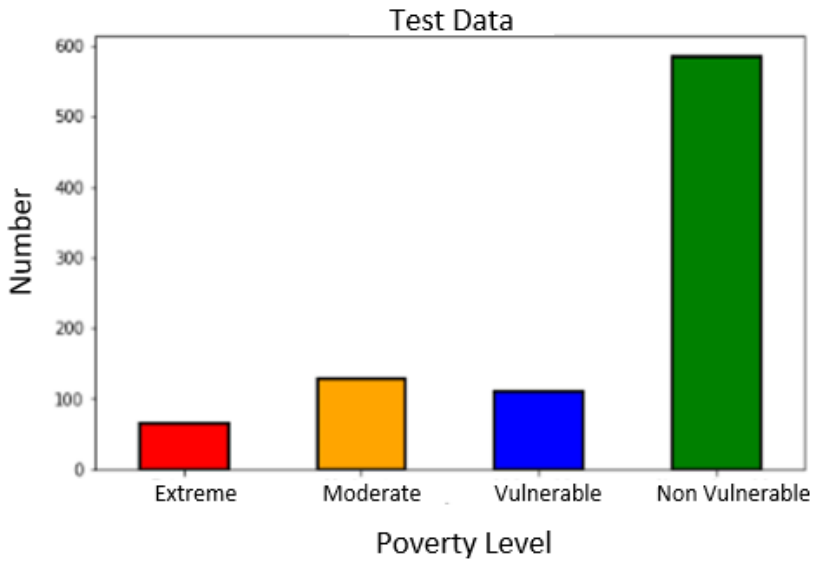
ดังภาพประกอบ 39 จะมีข้อมูล 896 แถว 36 คอลัมน์โดยแบ่งเป็น

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 66 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 133 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 111 แถว

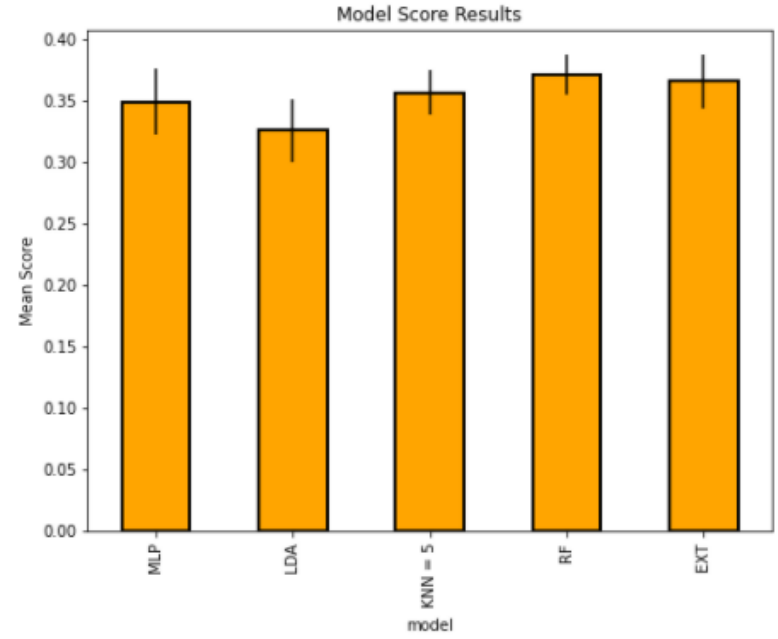
4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 586 แถว



ภาพประกอบ 39 ชุดข้อมูล Test แต่ละระดับความยากจน

ก่อนที่นำชุดข้อมูลไปใช้งานนั้น จะกำหนดคุณลักษณะในการทำนายระดับความยากจนคือคอลัมน์ Target และสร้างแบบจำลองโดยเลือกใช้วิธีการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning) มาใช้ในการสร้างแบบจำลองทั้ง 5 เทคนิคคือ 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) เนื่องจากในการศึกษางานวิจัยที่เกี่ยวข้องพบว่าเทคนิคเหล่านี้มีประสิทธิภาพที่ดีและเหมาะสมในการทำนายระดับความยากจน

โดยแต่ละเทคนิคจะใช้ค่าตัวแปรพารามิเตอร์แบบเริ่มต้น (Default) จาก Scikit-learn และจะทำการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคด้วยวิธีการตรวจสอบแบบไขว้โดยแบ่งข้อมูลเป็น 10 กลุ่ม (Stratified K-Fold Cross Validation) โดยใช้คะแนน Accuracy ในการวัดประสิทธิภาพ ผลการทดลองดังภาพประกอบ 40



ภาพประกอบ 40 เปรียบเทียบประสิทธิภาพของแต่ละเทคนิคด้วยวิธีการตรวจสอบไขว้โดยแบ่งข้อมูลเป็น 10 กลุ่ม (Stratified K-Fold Cross Validation)

ตาราง 22 ผลการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค

Model	F1 Score
Multilayer Perceptron (MLP)	0.3488
Linear Discriminant Analysis (LDA)	0.3254
K-Nearest Neighbors K=5 (KNN, K=5)	0.3563
Random Forest Classifier (RF)	0.3712
Extra Trees Classifier (EXT)	0.3655

จากตาราง 22 เมื่อเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค จะพบว่า Random Forest Classifier มีคะแนน F1 Score สูงที่สุด ดังนั้นจึงเลือกเทคนิค Random Forest Classifier เพื่อนำไปใช้ในขั้นตอนต่อไป

3.5 การประเมินผลแบบจำลอง

3.5.2 การสร้างแบบจำลองและวัดประสิทธิภาพ (Modeling and Evaluation)

จากชุดข้อมูล Train จะพบว่าชุดข้อมูลของงานวิจัยนี้จะเป็นข้อมูลที่ไม่สมดุล (Imbalance dataset) คือคุณลักษณะระบุระดับความยากจน (Target) ทั้ง 4 ระดับมีอัตราส่วนอยู่ที่ 2:4:3:16 ดังนั้นเราจะทำการแก้ไขปัญหาของข้อมูลที่ไม่สมดุล (Imbalance dataset) โดยการนำเทคนิคต่างๆ มาช่วยพัฒนาการสร้างแบบจำลอง เพื่อให้ได้ประสิทธิภาพ

3.5.1 การปรับปรุงข้อมูลที่ไม่สมดุล (Imbalance dataset)

โดยนำข้อมูล Train มาทำการปรับปรุงข้อมูลโดยแบ่งเป็นทั้ง 3 กลุ่ม คือ ข้อมูลเดิม (Existing Data) , ข้อมูลที่มีการปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค SMOTE , ADASYN และข้อมูลที่มีการปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder , Cluster Centroids โดยแต่ละเทคนิคมีจำนวนข้อมูลดังนี้

3.5.1.1 ข้อมูลเดิม (Existing Data)

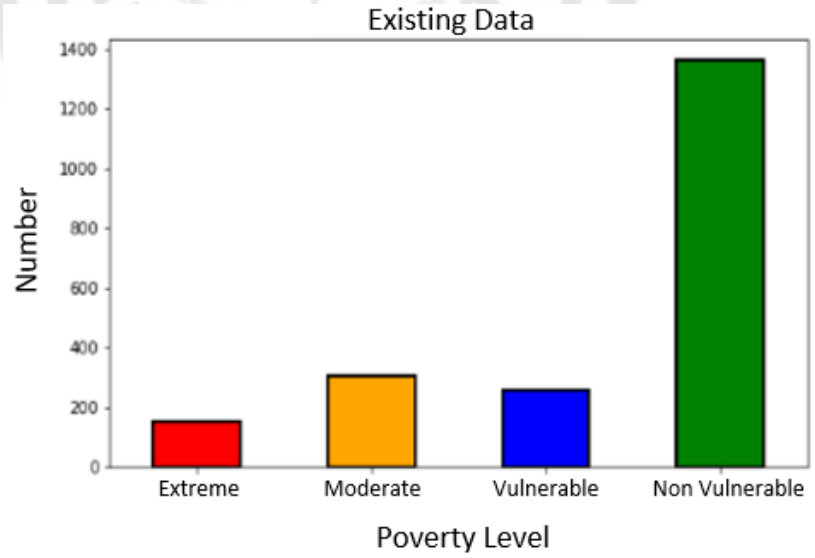
ภาพประกอบ 41 มีจำนวนข้อมูลดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 156 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 309 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 258 แถว

4 คือบุคคลเสี่ยงจะยากจน (Non vulnerable households) 1366 แถว



ภาพประกอบ 41 ข้อมูลเดิม (Existing Data)

3.5.1.2 **การปรับเพิ่มข้อมูล** (Over-Sampling) **ด้วยเทคนิค** SMOTE

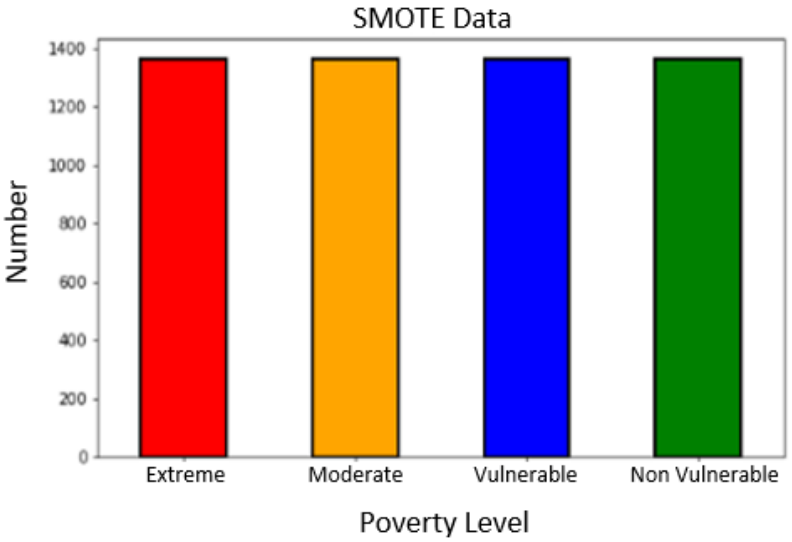
ภาพประกอบ 42 มีจำนวนข้อมูลดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 1366 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 1366 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 1366 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 1366 แถว



ภาพประกอบ 42 การปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค SMOTE

3.5.1.3 **การปรับเพิ่มข้อมูล** (Over-Sampling) **ด้วยเทคนิค** ADASYN

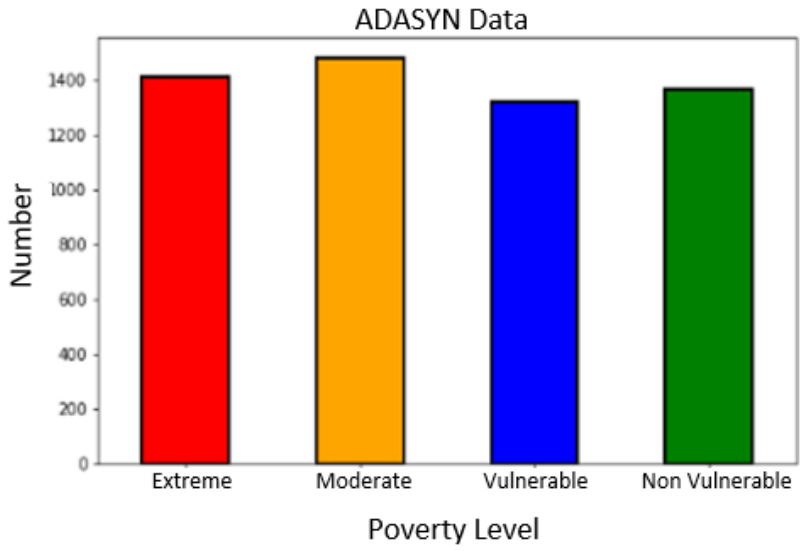
ภาพประกอบ 43 มีจำนวนข้อมูลดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 1424 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 1403 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 1348 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 1366 แถว



ภาพประกอบ 43 การปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค ADASYN

3.5.1.4 การปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder

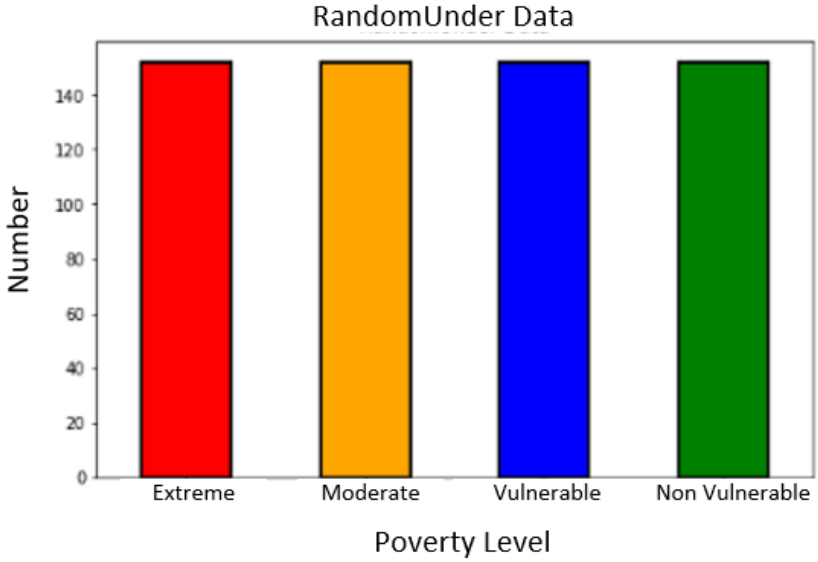
ภาพประกอบ 44 มีจำนวนข้อมูลดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 156 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 156 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 156 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 156 แถว



ภาพประกอบ 44 การปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder

3.5.1.5 **การปรับลดข้อมูล** (Under-Sampling) **ด้วยเทคนิค** ClusterCentroids

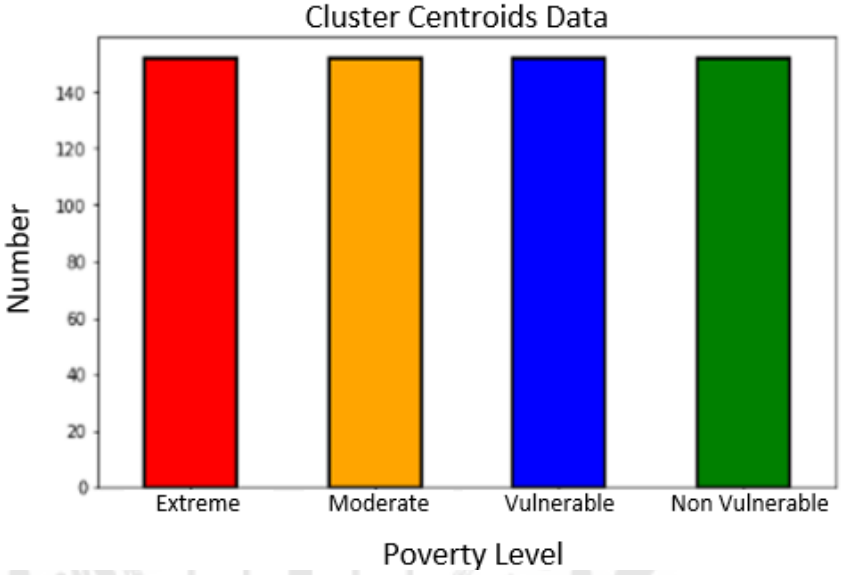
ภาพประกอบ 45 มีจำนวนข้อมูลดังนี้

1 คือบุคคลยากจนขั้นรุนแรง (Extreme poverty) 156 แถว

2 คือบุคคลยากจนปานกลาง (Moderate poverty) 156 แถว

3 คือบุคคลเสี่ยงจะยากจน (Vulnerable households) 156 แถว

4 คือบุคคลไม่เสี่ยงจะยากจน (Non vulnerable households) 156 แถว



ภาพประกอบ 45 การปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค Cluster Centroids

3.5.2 **การปรับปรุงไฮเปอร์พารามิเตอร์** (Hyper-Parameter) **ด้วยเทคนิค** Grid Search

ก่อนที่จะนำชุดข้อมูลทั้ง 5 แบบมาสร้างแบบจำลองนั้น ควรจะมีการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter) เพื่อให้แต่ละพารามิเตอร์เหมาะสมกับแต่ละแบบจำลอง ด้วยเทคนิค Grid Search ร่วมกับอัลกอริทึม Random Forest Classifier และกำหนดค่า F1-Score เป็นค่าวัดประสิทธิภาพ โดยเลือกใช้ค่า Hyper-Parameter ดังตาราง 23

ตาราง 23 ชื่อไฮเปอร์พารามิเตอร์ (Hyper-Parameter)ของแบบจำลอง Random Forest Classifier

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

3.5.3 **การสร้างแบบจำลองและวัดประสิทธิภาพ (Modeling and Evaluation)**

เมื่อทำการปรับปรุงพารามิเตอร์แล้ว จึงนำข้อมูลทั้ง 5 แบบคือ ข้อมูลเดิม (Existing Data) , ข้อมูลที่มีการปรับเพิ่มข้อมูล (Over-Sampling) ด้วยเทคนิค SMOTE , ADASYN และข้อมูลที่มีการปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder , ClusterCentroids มาทำการฝึกการทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลองทั้ง 5 แบบ โดยการนำข้อมูล Test มาทำการทดสอบกับแบบจำลองและการวัดประสิทธิภาพด้วย คอนฟูชั่นแมททริก (Confusion Matrix) , ค่าความถูกต้อง (Accuracy), ค่าความครบถ้วน(Recall), ค่าความแม่นยำ (Precision), ค่าประสิทธิภาพโดยรวม (F1 Score) จากนั้นทำการเปรียบเทียบกันระหว่าง 5 แบบจำลองเพื่อหาแบบจำลองที่มีประสิทธิภาพดีที่สุด

3.5.4 **คุณลักษณะสำคัญที่มีผลต่อการทำนาย**

ขั้นตอนสุดท้ายเลือกแบบจำลองที่มีประสิทธิภาพดีที่สุด นำมาหาคุณลักษณะที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคลด้วยเทคนิค Feature Importance คือการตรวจสอบแต่ละคอลัมน์ที่นำมาใช้สร้างแบบจำลองว่าคอลัมน์ใดถูกนำไปใช้ในอัตราส่วนเท่าใด โดยจะทำการเลือกคอลัมน์ที่มีอัตราส่วนมากที่สุด 3 อันดับแรกและการแสดงแผนภาพของข้อมูลในแต่ละคุณลักษณะโดยเลือกคุณลักษณะที่สำคัญที่สุด 3 อันดับแรก เพื่อวิเคราะห์ดูความแตกต่าง

บทที่ 4
การทดลอง

ในงานวิจัยการทำนายระดับความยากจนของแต่ละบุคคล โดยใช้ชุดข้อมูลสำมโนประชากรของประเทศคอสตาริกา ด้วยเทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนต่าง ๆ ตลอดจนการวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ของการวิจัยที่ได้กำหนดไว้ ได้ดังนี้

4.1 ผลลัพธ์ของการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter)

4.2 ผลลัพธ์การวัดประสิทธิภาพของแบบจำลอง

4.3 ผลลัพธ์ของคุณลักษณะสำคัญที่มีผลต่อการทำนาย

4.1 ผลลัพธ์ของการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter)

ในการสร้างแบบจำลองนั้นควรจะมีการปรับปรุงไฮเปอร์พารามิเตอร์ (Hyper-Parameter) เพื่อเพิ่มประสิทธิภาพในการทำงานโดยในที่นี้จะใช้เทคนิค Grid Search โดยเลือกปรับปรุงพาราเตอร์ของแบบจำลอง Random Forest Classifier ซึ่งผลลัพธ์ของการปรับปรุงพารามิเตอร์ (Parameter) ทั้ง 5 แบบดังตาราง 24

ตาราง 24 ผลลัพธ์ของการปรับปรุงพารามิเตอร์ (Parameter)

Parameter	Existing	SMOTE	ADASYN	RandomUnder	ClusterCentroids
bootstrap	True	False	False	True	Ture
max_depth	110	110	150	100	150
max_features	3	4	4	4	3
min_samples_leaf	5	3	3	4	4
min_samples_split	12	10	10	12	14
'n_estimators	600	100	100	200	100

4.2 ผลลัพธ์การวัดประสิทธิภาพของแบบจำลอง

เมื่อทำการปรับปรุงไฮเปอร์พารามิเตอร์แล้ว จึงนำข้อมูลทั้ง 5 แบบคือ ข้อมูลเดิม (Existing Data) , ข้อมูลที่มีการปรับเพิ่มข้อมูล (Over-sampling) ด้วยเทคนิค SMOTE , ADASYN และข้อมูลที่มีการปรับลดข้อมูล (Under-Sampling) ด้วยเทคนิค RandomUnder , ClusterCentroids มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลองทั้ง 5 แบบ โดยใช้ตัวชี้วัดดังนี้

1.ค่าความถูกต้อง (Accuracy)

2.ค่าความครบถ้วน (Recall)

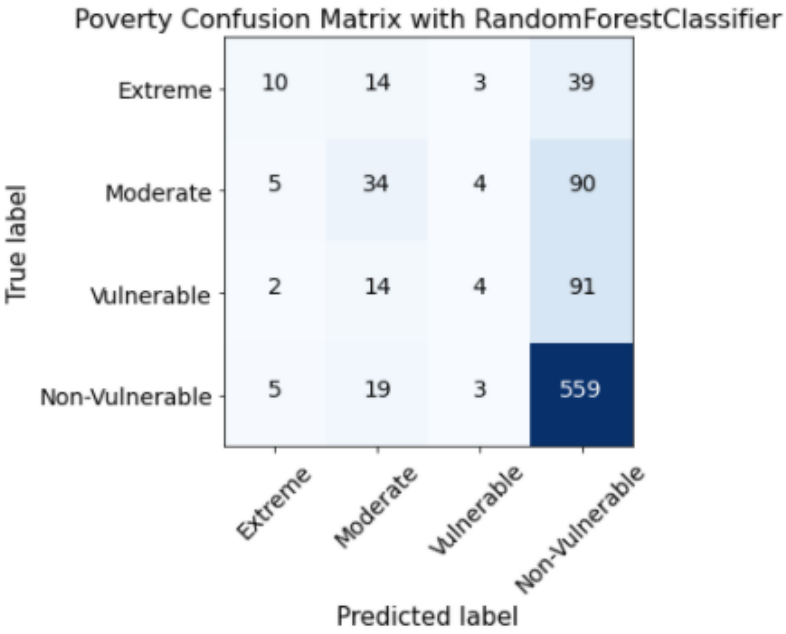
3.ค่าความแม่นยำ (Precision)

4.ค่าประสิทธิภาพโดยรวม (F1 Score)

ซึ่งได้ผลการทดลองดังนี้

4.2.1 ข้อมูลเดิม (Existing Data)

นำชุดข้อมูลเดิม (Existing Data) มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ดังภาพประกอบ 46 และสามารถสรุปได้ดังตาราง 25



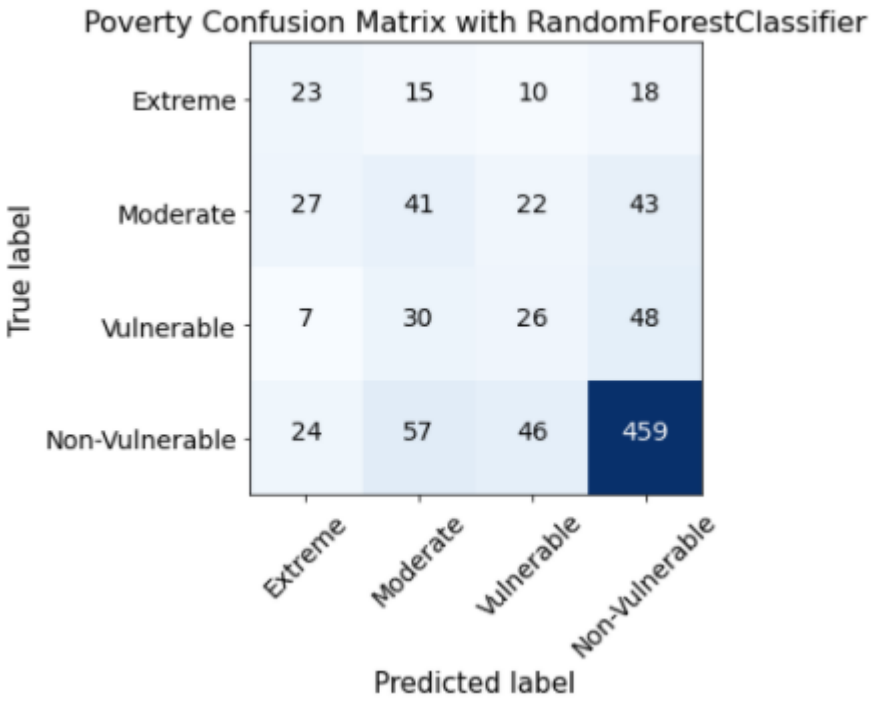
ภาพประกอบ 46 Confusionmatrix ของข้อมูลเดิม (Existing Data)

ตาราง 25 ผลการทดลองข้อมูลเดิม (Existing Data)

	Precision	Recall	F1-Score	Accuracy
Class 1	0.39	0.11	0.17	
Class 2	0.39	0.20	0.27	
Class 3	0.20	0.05	0.07	
Class 4	0.72	0.96	0.82	
Average	0.42	0.33	0.33	0.67

4.2.2 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE

นำข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ดังภาพประกอบ 47 และสามารถสรุปได้ดังตาราง 26



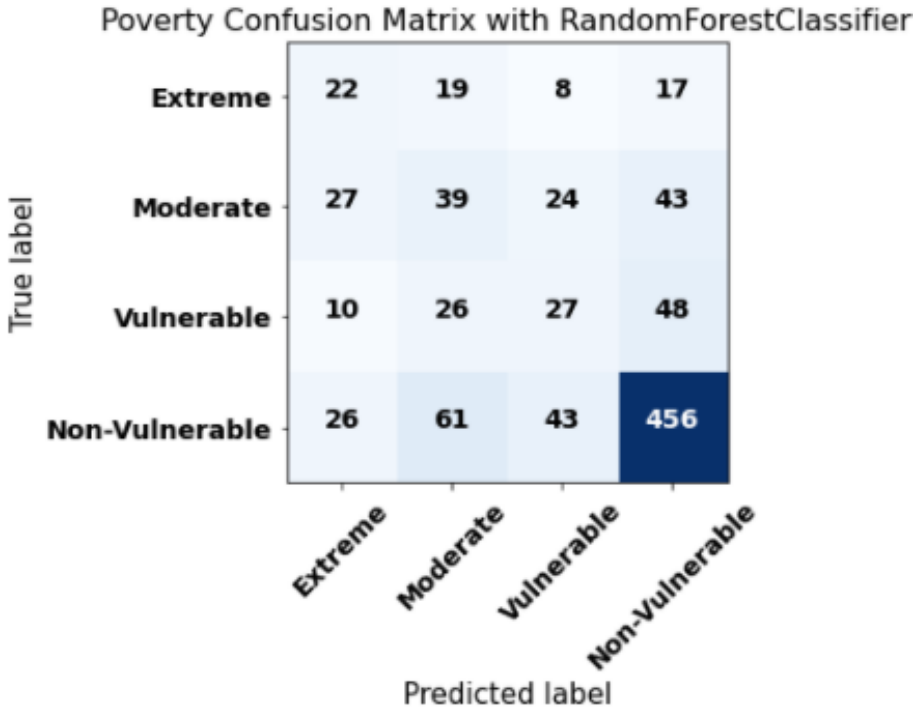
ภาพประกอบ 47 confusion matrix ของข้อมูลปรับเพิ่มด้วยเทคนิค SMOTE

ตาราง 26 ผลการทดลองข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE

	Precision	Recall	F1-Score	Accuracy
Class 1	0.32	0.29	0.30	
Class 2	0.33	0.36	0.35	
Class 3	0.25	0.25	0.25	
Class 4	0.80	0.80	0.80	
Average	0.43	0.42	0.43	0.63

4.2.3 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค ADASYN

นำข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค ADASYN มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ดังภาพประกอบ 48 และสามารถสรุปได้ดังตาราง 27



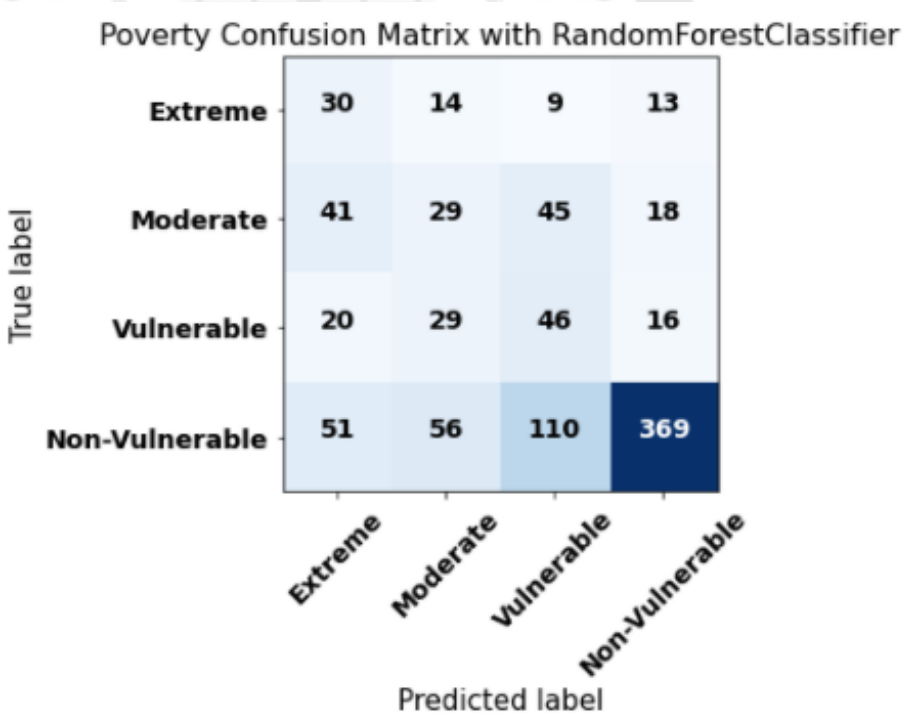
ภาพประกอบ 48 Confusion matrix ของข้อมูลปรับเพิ่มด้วยเทคนิค ADASYN

ตาราง 27 ผลการทดลองข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค ADASYN

	Precision	Recall	F1-Score	Accuracy
Class 1	0.28	0.27	0.28	
Class 2	0.29	0.31	0.30	
Class 3	0.18	0.18	0.18	
Class 4	0.80	0.79	0.80	
Average	0.39	0.39	0.39	0.61

4.2.4 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค RandomUnder

ข้อมูลที่มีการปรับลดด้วยเทคนิค Random Under มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ดังภาพประกอบ 49 และสามารถสรุปได้ดังตาราง 28



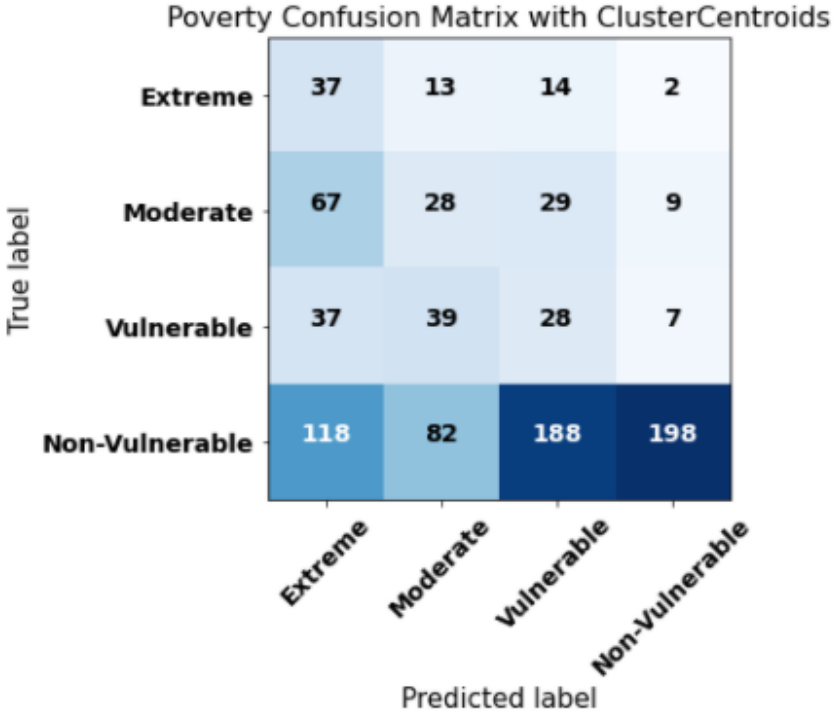
ภาพประกอบ 49 Confusion matrix ของข้อมูลปรับลดด้วยเทคนิค Random Under

ตาราง 28 ผลการทดลองข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค Random Under

	Precision	Recall	F1-Score	Accuracy
Class 1	0.26	0.42	0.32	
Class 2	0.27	0.32	0.29	
Class 3	0.23	0.41	0.29	
Class 4	0.87	0.65	0.75	
Average	0.41	0.45	0.41	0.56

4.2.5 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค Cluster Centroids

ข้อมูลที่มีการปรับลดด้วยเทคนิค Cluster Centroids มาทำการฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ดังภาพประกอบ 50 และสามารถสรุปได้ดังตาราง 29



ภาพประกอบ 50 Confusion matrix ของข้อมูลปรับลดด้วยเทคนิค Cluster Centroids

ตาราง 29 ผลการทดลองข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค ClusterCentroids

	Precision	Recall	F1-Score	Accuracy
Class 1	0.13	0.52	0.21	
Class 2	0.23	0.29	0.25	
Class 3	0.14	0.30	0.19	
Class 4	0.91	0.36	0.51	
Average	0.35	0.36	0.29	0.35

การฝึกทำนายข้อมูลกับแบบจำลองที่สร้างด้วยเทคนิค Random Forest Classification จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง ซึ่งผลการวัดประสิทธิภาพที่ได้ของแบบจำลองทั้ง 5 แบบสามารถสรุปรวมได้ดังตาราง 30

ตาราง 30 สรุปผลการทดลองของแต่ละแบบจำลอง

	Precision	Recall	F1-Score	Accuracy
Existing Data	0.42	0.33	0.33	0.67
SMOTE	0.43	0.42	0.43	0.63
ADASYN	0.39	0.39	0.39	0.61
RandomUnder	0.41	0.45	0.41	0.56
ClusterCentroids	0.35	0.36	0.29	0.35

4.3 ผลลัพธ์ของคุณลักษณะสำคัญที่มีผลต่อการทำนาย

เมื่อการสร้างแบบจำลองทั้ง 5 แบบและวัดประสิทธิภาพเรียบร้อยแล้ว จากนั้นทำการหาคุณลักษณะสำคัญที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคลด้วยเทคนิค Feature Impotence โดยจะเลือกคุณลักษณะที่มีคะแนนสูง 10 อันดับแรกซึ่งแต่ละแบบจำลองได้ผลการทดลองดังนี้

4.3.1 ข้อมูลเดิม (Existing Data)

จากแบบจำลองที่สร้างด้วยชุดข้อมูลเดิม (Existing Data) พบว่าคุณลักษณะสำคัญที่มีผลต่อการทำนาย 10 อันดับแรกผลการทดลองที่ได้ดังตาราง 31

ตาราง 31 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลเดิม (Existing Data)

Feature	Importance
escolari-mean	0.094779
age-mean	0.066320
escolari-max	0.064953
meaneduc-min	0.061133
meaneduc-max	0.060318
meaneduc-mean	0.059888
age-min	0.050351
escolari-min	0.038363
dependency-min	0.038237
dependency-mean	0.035602

4.3.2 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE

จากแบบจำลองที่สร้างด้วยข้อมูลที่ปรับเพิ่มด้วยเทคนิค SMOTE พบว่าคุณลักษณะสำคัญที่มีผลต่อการทำนาย 10 อันดับแรกผลการทดลองที่ได้ดังตาราง 32

ตาราง 32 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลที่ปรับเพิ่มด้วยเทคนิค SMOTE

Feature	Importance
age-mean	0.066546
escolari-max	0.065201
meaneduc-max	0.059740
escolari-mean	0.057657
age-min	0.042458
meaneduc-min	0.041288
meaneduc-mean	0.040783
dependency-min	0.038765
instlevel8-max	0.037774

4.3.3 ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค ADASYN

จากแบบจำลองที่สร้างด้วยข้อมูลที่ปรับเพิ่มด้วยเทคนิค ADASYN พบว่าคุณลักษณะสำคัญที่มีผลต่อการทำนาย 10 อันดับแรกผลการทดลองที่ได้ดังตาราง 33 และตาราง 34

ตาราง 33 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลที่ปรับเพิ่มด้วยเทคนิค ADASYN

Feature	Importance
age-mean	0.066071
age-min	0.060120
escolari-mean	0.059078
escolari-max	0.058799
meaneduc-mean	0.046620
meaneduc-min	0.043471
meaneduc-max	0.043452
instlevel8-max	0.039829
dependency-mean	0.038938
escolari-min	0.038822

4.3.4 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค RandomUnder

จากแบบจำลองที่สร้างด้วยข้อมูลที่ปรับลดด้วยเทคนิค RandomUnder พบว่าคุณลักษณะสำคัญที่มีผลต่อการทำนาย 10 อันดับแรกผลการทดลองที่ได้ดังตาราง 35

ตาราง 34 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลที่ปรับลดด้วยเทคนิค RandomUnder

Feature	Importance
escolari-mean	0.088763
age-mean	0.083926
age-min	0.068121
escolari-max	0.064580

ตาราง 35 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลที่ปรับลดด้วยเทคนิค RandomUnder(ต่อ)

escolari-max	0.064580
meaneduc-max	0.053435
meaneduc-min	0.052936
meaneduc-mean	0.050518
dependency-mean	0.037859
dependency-max	0.037538

4.3.5 ข้อมูลที่มีการปรับลด (Under-Sampling) ด้วยเทคนิค ClusterCentroids

จากแบบจำลองที่สร้างด้วยข้อมูลที่ปรับลดด้วยเทคนิค ClusterCentroids พบว่าคุณลักษณะสำคัญที่มีผลต่อการทำนาย 10 อันดับแรกผลการทดลองที่ได้ดังตาราง 36

ตาราง 36 คุณลักษณะที่มีผลต่อการทำนายโดยใช้ข้อมูลที่ปรับลดด้วยเทคนิค ClusterCentroids

Feature	Importance
v2a1-min	0.070269
v2a1-mean	0.060743
v2a1-max	0.058560
escolari-mean	0.055234
education-mean	0.044345
age-mean	0.042311
escolari-max	0.039963
meaneduc-min	0.038614
meaneduc-max	0.034475
age-min	0.034170

จากผลการทดลองคุณลักษณะที่มีผลต่อการทำนายระดับความยากจนของแต่ละบุคคล ซึ่งแบบจำลองทั้ง 5 แบบสามารถสรุปรวมได้ดังตาราง 37 และตาราง 38

ตาราง 37 สรุปคุณลักษณะสำคัญที่มีผลต่อการทำนาย

No.	Existing Data	SMOTE	ADASYN	RandomUnder	ClusterCentroids
1	escolari-mean	age-mean	age-mean	escolari-mean	v2a1-min
2	age-mean	escolari-max	age-min	age-mean	v2a1-mean
3	escolari-max	meaneduc-max	escolari-mean	age-min	v2a1-max
4	meaneduc-min	escolari-mean	escolari-max	escolari-max	escolari-mean
5	meaneduc-max	age-min	meaneduc-mean	escolari-max	education-mean
6	meaneduc-mean	meaneduc-min	meaneduc-min	meaneduc-max	age-mean
7	age-min	meaneduc-mean	meaneduc-max	meaneduc-min	escolari-max
8	escolari-min	dependency-min	instlevel8-max	meaneduc-mean	meaneduc-min
9	dependency-min	instlevel8-max	dependency-mean	dependency-mean	meaneduc-max
10	dependency-mean	dependency-max	escolari-min	dependency-max	age-min

จากภาพรวมสรุปได้ว่าการทำนายระดับความยากจนของแต่ละบุคคล โดยใช้ชุดข้อมูลสำมโนประชากรของประเทศคอสตาริกา ด้วยวิธีการเรียนรู้ของเครื่องพบว่า ข้อมูลที่มีการปรับเพิ่ม (Over-Sampling) ด้วยเทคนิค SMOTE จะได้ประสิทธิที่ภาพดีที่สุด และมีปัจจัยที่ส่งผลกระทบต่อความยากจน 3 อันดับแรกคือ age-mean (อายุ) , escolari-max (จำนวนปีในการศึกษา), meaneduc-max (จำนวนปีการศึกษาโดยเฉลี่ยสำหรับผู้ใหญ่)

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยการทำนายระดับความยากจนของแต่ละบุคคล โดยใช้ชุดข้อมูลสำมโนประชากรของประเทศคอสตาริกา ด้วยเทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแต่ละแบบจำลองเพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

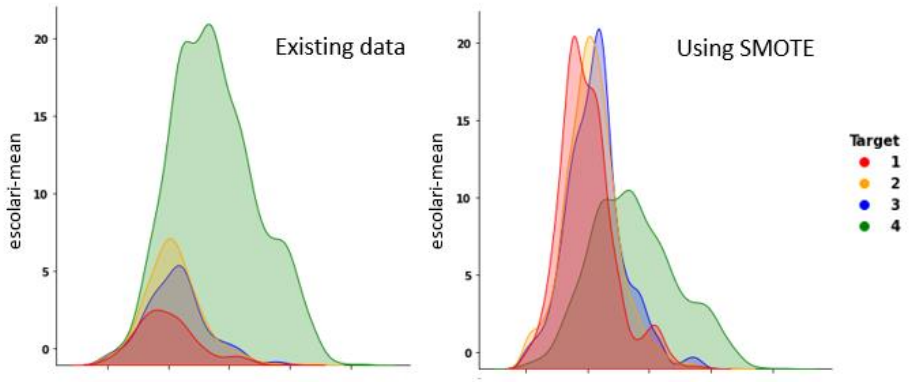
ความยากจน หมายถึง สภาพความเป็นอยู่ของประชาชนที่ต่ำกว่ามาตรฐานและมีรายรับไม่เพียงพอที่จะใช้จ่ายในชีวิตประจำวันที่จำเป็นขั้นพื้นฐาน ความยากจนจึงนับเป็นปัญหาที่ทุกประเทศทั่วโลกต่างให้ความสำคัญ เนื่องจากความยากจนเป็นสาเหตุอย่างหนึ่งที่ทำให้มีโอกาสที่จะก่อให้เกิดปัญหาอื่น ๆ ตามมา เพราะฉะนั้นการแก้ปัญหาความยากจนจึงเป็นสิ่งสำคัญ ดังนั้นจึงต้องหาแนวทางในการแก้ไขปัญหาความยากจน แต่สิ่งที่สำคัญก็คือต้องสามารถระบุตัวตนของแต่ละบุคคลได้ว่าบุคลคนนั้นยากจนจริงหรือไม่และมีสิ่งใดบ้าง ที่เป็นปัจจัยที่ส่งผลกระทบต่อความยากจนเพื่อเป็นแนวทางในการแก้ไขปัญหาความยากจนต่อไป

ในวิจัยนี้จึงพัฒนาแบบจำลองการทำนายระดับความยากจนของแต่ละบุคคลและปัจจัยที่ส่งผลกระทบต่อความยากจน ด้วยวิธีการนำเทคโนโลยีการเรียนรู้ของเครื่องมาใช้ในการวิเคราะห์ข้อมูลสำมะโนประชากรเพื่อระบุระดับความยากจน โดยนำชุดข้อมูลมาสร้างแบบจำลองและทำการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมทั้งหมด 5 แบบคือ โครงข่ายประสาทเทียมแบบ Multilayer Perceptron, การวิเคราะห์การจำแนกประเภทเชิงเส้น , การค้นหาเพื่อนบ้านใกล้สุด K อันดับ, การสุ่มป่าไม้ , ต้นไม้ตัดสินใจจำนวนมาก เพื่อหาอัลกอริทึมดีที่สุด จากนั้นทำการแก้ไขปัญหาข้อมูลที่ไม่สมดุลกัน (Imbalance Dataset) โดยใช้วิธีการปรับเพิ่มด้วยเทคนิค SMOTE และ ADYSYN และวิธีการปรับลดด้วยเทคนิค RandomUnder และ ClusterCentroids จากนั้นนำชุดข้อมูลต่างๆมาสร้างแบบจำลองร่วมกับอัลกอริทึมทีหาได้ก่อนหน้านี้เพื่อหาแบบจำลองที่ดีที่สุด โดยวัดประสิทธิภาพของแต่ละแบบจำลองได้ตามตาราง 30 รวมถึงหาคุณลักษณะสำคัญที่ส่งผลต่อการทำนายระดับความยากจนตามตาราง 37

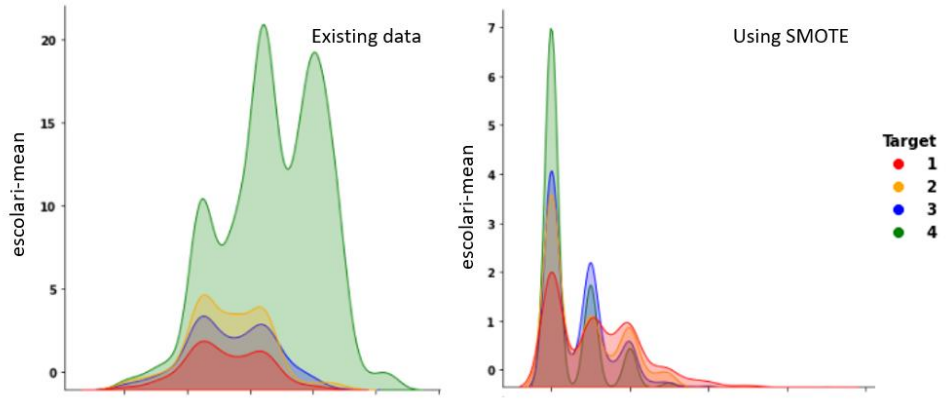
จากผลการทดลองสรุปได้ว่าแบบจำลองที่มีประสิทธิภาพดีที่สุดในการแก้ไขปัญหาข้อมูลที่ไม่สมดุลนั้นคือ แบบจำลองการสุ่มป่าไม้ที่พัฒนาร่วมกับข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE โดยค่าเฉลี่ยที่ได้จากการทำนายระดับความยากจนคือ Precision = 0.43 , Recall = 0.42 , F1-Score = 0.43 , Accuracy = 0.63 และปัจจัยที่ส่งผลกระทบต่อความยากจน 3 อันดับแรกคือ age-mean (อายุ) , escolari-max (จำนวนปีในการศึกษา), meaneduc-max (จำนวนปีการศึกษาโดยเฉลี่ยสำหรับผู้ใหญ่)

5.2. **อภิปรายผลการวิจัย**

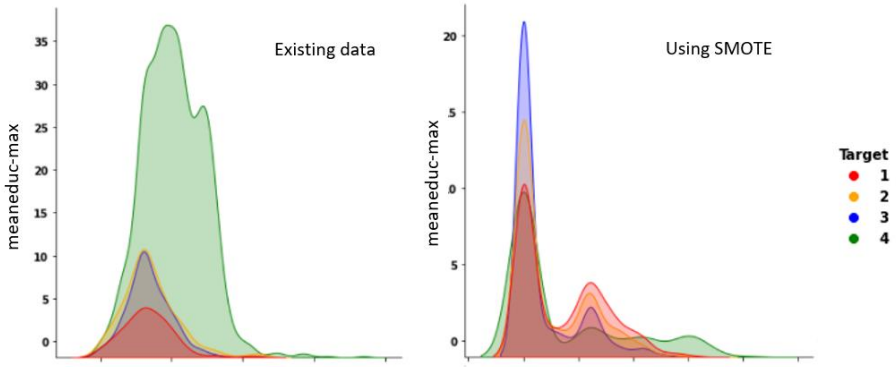
จากตาราง 30 สรุปผลการทดลองของแต่ละแบบจำลอง เมื่อทำการเปรียบเทียบผลการทดลองระหว่างชุดข้อมูลเดิม (Existing Data) และชุดข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE ซึ่งในงานวิจัยนี้จุดประสงค์คือต้องการแก้ไขปัญหาข้อมูลที่ไม่สมดุลกัน (Imbalance Dataset) พบว่าค่าเฉลี่ยของการทำนายระดับความยากจนที่ได้นั้นประสิทธิภาพของทั้ง 2 แบบจำลองนั้นไม่ได้แตกต่างกันมาก เมื่อทำการเปรียบเทียบผลการทดลองระหว่างชุดข้อมูลเดิมในตาราง 25 และชุดข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE ในตาราง 26 เพื่อแยกดูผลการทดลองของแต่ละกลุ่ม พบว่าในกลุ่มที่ 4 คือบุคคลไม่เสี่ยงจะยากจน (Non Vulnerable Households) มีค่าเฉลี่ยของการทำนายระดับความยากจนที่ได้นั้นประสิทธิภาพค่อนข้างสูง แต่ในกลุ่มที่ 1,2,3 คือบุคคลยากจนขั้นรุนแรง (Extreme Poverty), บุคคลยากจนปานกลาง (Moderate Poverty),บุคคลเสี่ยงจะยากจน (Vulnerable Households) ตามลำดับ มีค่าเฉลี่ยของการทำนายระดับความยากจนที่ได้นั้นประสิทธิภาพไม่ดีเท่าที่ควร ดังนั้นทางผู้วิจัยได้ทำการแสดงแผนภาพของข้อมูลในแต่ละคุณลักษณะโดยเลือกคุณลักษณะที่สำคัญที่สุด 3 อันดับแรก เพื่อดูความแตกต่างระหว่างข้อมูลเดิมและข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE ดังภาพประกอบ 51 ถึง ภาพประกอบ 53



ภาพประกอบ 51 คุณลักษณะ age-mean เปรียบเทียบระหว่างข้อมูลเดิมและข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE



ภาพประกอบ 52 คุณลักษณะ escolari-max เปรียบเทียบระหว่างข้อมูลเดิมและข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE



ภาพประกอบ 53 คุณลักษณะ meaneduc-max เปรียบเทียบระหว่างข้อมูลเดิมและข้อมูลที่มีการปรับเพิ่มด้วยเทคนิค SMOTE

พบว่าเมื่อดูในส่วนของข้อมูลเดิม (Existing Data) ข้อมูลกลุ่มที่ 4 (สีเขียว) จะมีจำนวนมากกว่ากลุ่มอื่นๆ ทำให้ผลลัพธ์การทำนายข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก ซึ่งทำให้ประสิทธิภาพในการทำนายไม่ดีเท่าที่ควร แต่เพื่อเราทำการแก้ไขปัญหาข้อมูลที่ไม่สมดุลกัน (Imbalance Dataset) ด้วยวิธีการปรับเพิ่มด้วยเทคนิค SMOTE จะทำให้ข้อมูลในกลุ่มที่ 1,2,3 (สีแดง,สีเหลือง,สีน้ำเงิน) เพิ่มขึ้นเป็นจำนวนใกล้เคียงกัน จึงทำให้ผลลัพธ์การทำนายระดับความยากจนของแต่ละบุคคลดีขึ้นไปด้วย

5.3 ข้อเสนอแนะ

5.3.1 ในการวิจัยนี้ได้ใช้ข้อมูลในการศึกษาจาก www.kaggle.com เป็นชุดข้อมูลสำมโนประชากรของประเทศคอสตาริกา ซึ่งสามารถนำมาประยุกต์ใช้กับข้อมูลสำมโนประชากรของประเทศไทยได้

5.3.2 เนื่องจากชุดข้อมูลในการวิจัยนี้เป็นข้อมูลที่ไม่สมดุลกันและข้อมูลในกลุ่มที่ 1,2,3 คือบุคคลยากจนขั้นรุนแรง (Extreme Poverty), บุคคลยากจนปานกลาง (Moderate Poverty), บุคคลเสี่ยงจะยากจน (Vulnerable Households) ตามลำดับ มีจำนวนค่อนข้างน้อย อาจจะใช้เทคนิคอื่นๆ นอกจากการปรับเพิ่มหรือปรับลดข้อมูล มาเป็นการปรับเพิ่มคุณลักษณะต่างๆ โดยใช้เทคนิค Feature Engineering อาจช่วยให้แบบจำลองนั้นสามารถฝึกฝนชุดข้อมูลได้เพิ่มเติม รวมถึงสามารถทำนายค่าออกมาได้มีมีประสิทธิภาพดีกว่าเดิม

บรรณานุกรม

Araya. (2562). 11 framework ที่เปิดให้ใช้ฟรีสำหรับโมเดลของ AI และ Machine Learning. สืบค้นจาก https://www.yod.net/framework-for-ai-and-ml/

Chakrabarty, Navoneel, Biswas, & Sanket;. (2018). A Statistical Approach to Adult Census Income Level Prediction.

iLog. (2562). Jupyter Notebook คือ เครื่องมือสำหรับ Data Science.

Korivi, K. (2016). Identifying poverty-driven need by augmenting census and community survey data.

Nata Sa Plulikova. (2559). POVERTY ANALYSIS USING MACHINE LEARNING METHODS. Retrieved from http://www.iam.fmph.uniba.sk/institute/stehlikova/BC/2016-plulikova.pdf?fbclid=IwAR1nDdAWizqsJNQ-nWMiHbBPQWGyLzmnjoqJPRxCWLQ9adIu0XrdiUXiKpE

Rincon, R. (2562). Quarterly Multidimensional Poverty Predictions in Mexico using Machine Learning Algorithms. Retrieved from https://sistemas.colmex.mx/Reportes/LACEALAMES/LACEA-LAMES2019_paper_754.pdf

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.

Subash, S., Kumar, R., & Aditya, K. S. (2018). Satellite data and machine learning tools for predicting poverty in rural India.

THAIYATHUM, N. (2562). K-Nearest Neighbors สืบค้นจาก https://www.glurgeek.com/education/knn

ณัฐพงษ์ พัฒนพงษ์ , Arturo M. Martinez Jr. , Ron Lester Santos Durante , & พิชญ์ จงวัฒนากุล. (2563). การวิเคราะห์ความเหลื่อมล้ำในประเทศไทยโดยใช้ข้อมูลดาวเทียมและภูมิสารสนเทศ. สืบค้นจาก http://www.econ.tu.ac.th/oldweb/doc/content/1850/Discussion_Paper_No.57.pdf

ปณตทรง วัฒนศิริ. (2553). เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม. (ปริญญานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย , กรุงเทพฯ). สืบค้นจาก http://161.200.145.125/bitstream/123456789/33369/1/panote_so.pdf

ภรัณยา อำมฤครัตน์, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, & ณรงค์ โพธิ. (2552). *A Comparative Efficiency of FeatureSelection and Neural Network Classification*. สืบค้นจาก http://www.nphothi.com/DrkLaNg/research/NCCIT2009-IT26-Neural.pdf

ภูริพัทร์ ทองคำ. (2561). อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล. (ปริญญานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์). สืบค้นจาก http://ethesisarchive.library.tu.ac.th/thesis/2016/TU_2016_5809035198_6887_4831.pdf

ลักษนันท์ พลอยวัฒนาวงศ์, ปริญญา นาโท, & พยุง มีสัจ. (2560). *An Efficiency of Algorithms Machine Learning for Agricultural Product Classification*. สืบค้นจาก 203.158.224.208/05_ITS/วิจัย.ผศ.ลักษนันท์.pdf

วิฎากร แซ่หว่อง. (2561). การจำแนกเสียงโกรธในบทสนทนาของศูนย์ให้บริการข้อมูล. จุฬาลงกรณ์มหาวิทยาลัย , กรุงเทพฯ. (ปริญญานิพนธ์ปริญญามหาบัณฑิต). สืบค้นจาก http://161.200.145.125/bitstream/123456789/61238/1/5981539526.pdf

ศูนย์เทคโนโลยีสารสนเทศกรมการจัดหางาน. (2562). ความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning.

สืบค้นจาก https://www.doe.go.th/prd/assets/upload/files/bkk_th/a323196dc548c204c85bc4a85b7bb46b.pdf

สำนักงานพัฒนารัฐบาลดิจิทัล. (2562). เทคโนโลยีปัญญาประดิษฐ์. สืบค้นจาก https://www.dga.or.th/upload/download/file_e4db016970b7f8b6764f4289c5e9a83f.pdf

สำนักงานราชบัณฑิตยสภา. (2552). ความยากจน. สืบค้นจาก 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

ประวัติผู้เขียน

ชื่อ-สกุล	ศรราม หงษ์พรหม
วัน เดือน ปี เกิด	15 ตุลาคม 2535
สถานที่เกิด	นครสวรรค์
วุฒิการศึกษา	พ.ศ.2558 วิศวกรรมศาสตรบัณฑิต วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์
ที่อยู่ปัจจุบัน	532 หมู่ 7 ตำบล ท่าตะโก อำเภอ ท่าตะโก จังหวัด นครสวรรค์