การทำนายความยากจนจากข้อมูลสำมะโนครัวประชากร ด้วยเทคนิคการเรียนรู้ของเครื่องจักร

รุจิรัตน์ สุธิมนต์รัตน์, ภัทรนันท์ ทุนทวีทรัพย์ และ ผศ.ดร. จันตรี ผลประเสริฐ
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

ที่มาและความสำคัญ

ความยากจนเป็นหนึ่งในปัจจัยสำคัญที่ส่งผลต่อสุขภาพและส่งผลต่อการเจริญเติมโตของประชากร การขจัดความยากจนยังคงเป็นความท้าทายที่สำคัญและเป็นเป้าหมายหลักของการพัฒนาอย่างยั่งยืน ซึ่งทางรัฐมีความพยายามที่จะแก้ปัญหานี้โดยการใช้ข้อมูลสำมะโนครัวประชากรมาทำการวิเคราะห์ เพื่อระบุผู้ยากไร้ในระดับครัวเรือนหรือชุมชนเพื่อหาวิธีการที่เหมาะสมในการแก้ไขปัญหาความยากจนได้อย่างยั่งยืน งานวิจัยนี้นำเสนอการใช้เทคโนโลยีการเรียนรู้ของเครื่องจักรแบบ Gradient boosting มาทำการวิเคราะห์ข้อมูลสำมะโนครัวประชากรเพื่อระบุความยากจน โดยใช้ precision, recall และ f1-score เป็นตัวชี้วัดความแม่นยำของโมเดล

องค์ความรู้ที่เกี่ยวข้อง

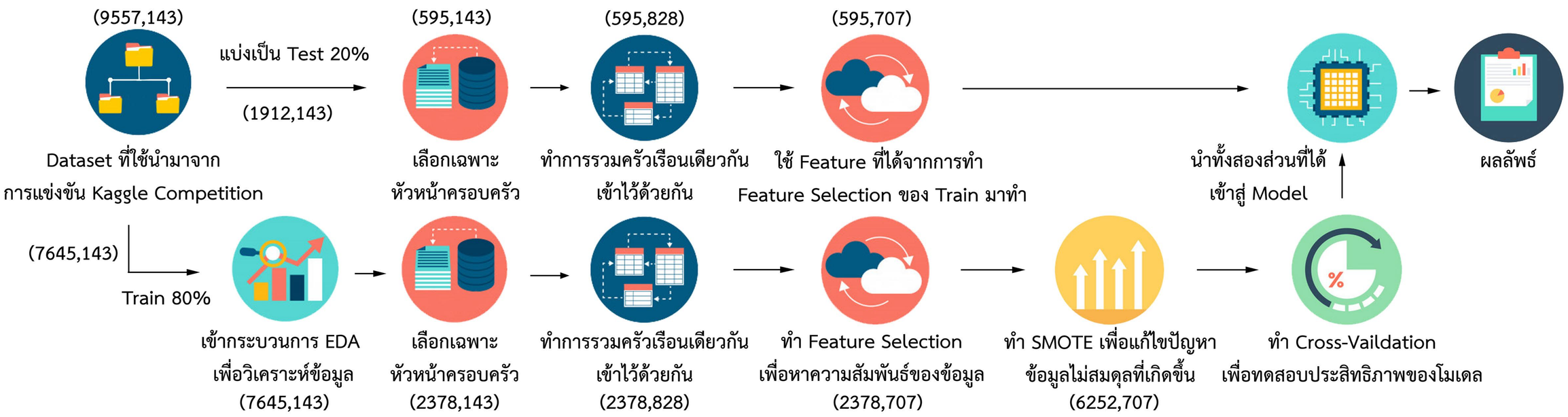
Machine Learning (เทคนิคการเรียนรู้ของเครื่องจักร)

พัฒนามาจากการศึกษาการรู้จำแบบและการสร้างอัลกอริทึมที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้

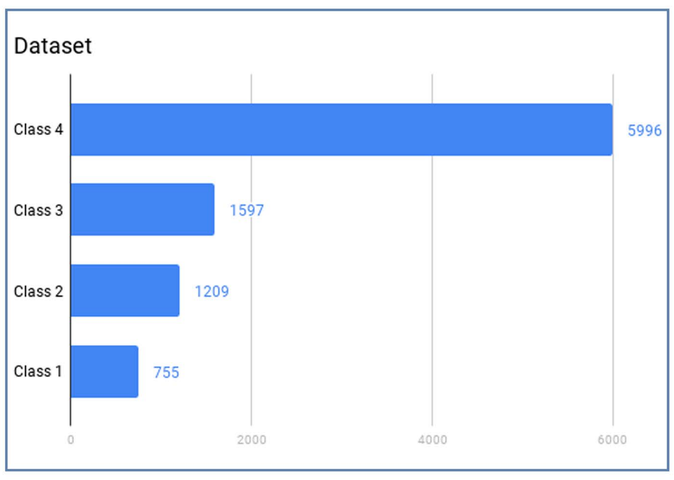
Gradient Boosting

การเรียนแบบเป็นลำดับ โดย learner ก่อนหน้าเรียนแล้วนำเอาข้อผิดพลาดของตัวเอง มาปรับปรุง learner ต่อๆไป เพื่อลด error จาก learner ก่อนหน้า

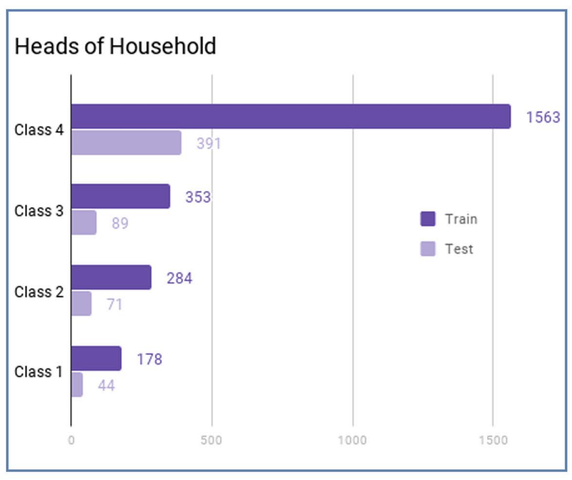
กระบวนการดำเนินงาน



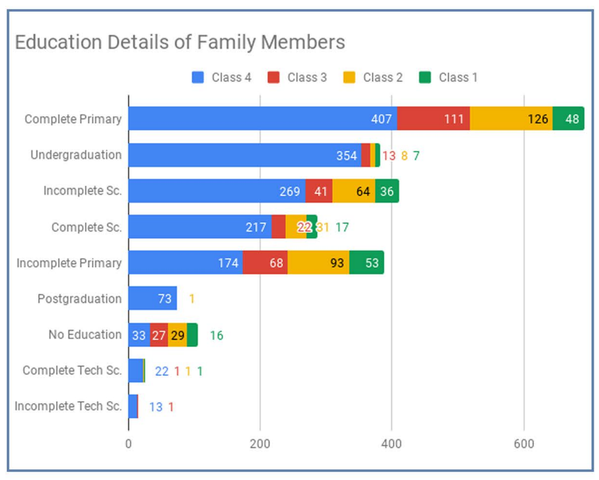
Exploratory Data Analysis (EDA)



ข้อมูลทั้งหมดจาก Kaggle Competition*



แบ่งข้อมูลเป็น Train 80%, Test 20% และเลือกเฉพาะ row ที่เป็นหัวหน้าครอบครัว



ข้อมูลระดับการศึกษาของหัวหน้าครอบครัว

ระดับของความยากจน

Class4 คือ ครัวเรือนที่ไม่มีความเสี่ยงที่จะยากจน

Class3 คือ ครัวเรือนที่มีความเสี่ยงเล็กน้อยที่จะยากจน

Class2 คือ ครัวเรือนที่มีความเสี่ยงที่จะยากจนพอสมควร

Class1 คือ ครัวเรือนที่มีความเสี่ยงสูงที่จะยากจน

*W. Koehrsen, "Costa Rican Household Poverty Level Prediction," สิงหาคม 2561. [Online]. Available: https://www.kaggle.com/c/costa-rican-household-poverty-prediction/data.

สรุปผล

ประสิทธิภาพของ Gradient Boosting	
10 fold CV F1-Score	0.88
precision	0.46
recall	0.42
F1-Score	0.43
Accuracy	0.68

คุณลักษณะที่สำคัญ 3 ลำดับแรก ของ Gradient Boosting	
คุณภาพของหลังคาบ้าน	0.024
จำนวนปีในสถานศึกษา	0.023
คุณภาพของพื้นบ้าน	0.022

จากผลการทดลองพบว่าโมเดล Gradient Boosting สามารถทำนายความยากจนทั้ง 4 ระดับได้อย่างแม่นยำโดยสามารถทำนาย Class ที่ 1-3 ได้แม่นยำกว่าโมเดลอื่น (KNN, RandomForest , Adaboost) โดยมีค่า Accuracy = 0.68, precision = 0.46, recall = 0.42, F1-score = 0.43 (คะแนน F1-score สูงสุดสูงสุดจากการแข่งขัน Kaggle Competition คือ 0.44 เกณฑ์การให้คะแนนในการแข่งขันคือ จะคิดคะแนนจากการนำ row ที่เป็นหัวหน้าครอบครัวมาวิเคราะห์เท่านั้น) โดยพบว่าสาเหตุหลักที่ส่งผลต่อความยากจนคือคุณภาพบ้านและการศึกษา