วิชา Statistics for computer engineering

รหัสวิชา 01204314

จัดทำโดย
นายก้องภพ ไพเราะ
รหัสนิสิต 6610505268

คณะวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์

อาจารย์ผู้สอน
อาจารย์ สุภาพร เอื้อจงมานี

มหาวิทยาลัยเกษตรศาสตร์

ภาคต้น ปีการศึกษา 2568

Part 1: Gender Ratio Class Analysis

Part 1.0: Exploratory Data Analysis (EDA)

วัตถุประสงค์

วัตถุประสงค์ของส่วน Exploratory Data Analysis (EDA) คือการทำความเข้าใจลักษณะของข้อมูล ตรวจสอบคุณภาพของข้อมูล และเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์และการสร้างโมเดลในขั้นตอนถัดไป

ข้อมูลที่ใช้

โครงงานนี้ใช้ข้อมูล 2 ชุด ได้แก่

1. **Dataset A (worldbank_gender)** ซึ่งประกอบด้วยตัวแปรด้านเศรษฐกิจ สังคม แรงงาน และตัวแปรเป้าหมายคือ *Gender Ratio Class*

2. **Dataset B (country regions)** ซึ่งให้ข้อมูลภูมิภาค (Region) ของแต่ละประเทศ

ข้อมูลทั้งสองชุดถูกนำมารวมกันโดยใช้ชื่อประเทศเป็นตัวเชื่อม เพื่อเพิ่มบริบทเชิงพื้นที่สำหรับการวิเคราะห์ในขั้นตอนถัดไป

การตรวจสอบและจัดการข้อมูล

จากการสำรวจข้อมูลพบว่า Dataset A มีจำนวนตัวแปรจำนวนมาก และหลายตัวแปรมีค่าข้อมูลสูญหาย (missing values) ในสัดส่วนที่สูงมาก โดยเฉพาะตัวแปรที่ไม่มีการรายงานข้อมูลในหลายประเทศ

เพื่อหลีกเลี่ยงการใช้ตัวแปรที่ไม่ให้ข้อมูลที่เป็นประโยชน์ จึงได้ดำเนินการ:

- คำนวณสัดส่วนของ missing values ในแต่ละตัวแปร

- ลบตัวแปรที่มี missing values **มากกว่า 70%** ออกจากชุดข้อมูล

หลังจากนั้น ได้ตรวจสอบตัวแปรเป้าหมาย (*Gender Ratio Class*) และลบแถวข้อมูลที่ไม่มีค่าของตัวแปรเป้าหมายออก เพื่อให้ข้อมูลพร้อมสำหรับการสร้างโมเดล

ผลลัพธ์ของ EDA ทำให้ได้ชุดข้อมูลที่มีคุณภาพดีขึ้น มีจำนวนตัวแปรที่เหมาะสม และสามารถนำไปใช้ในขั้นตอน modeling ได้อย่างมีประสิทธิภาพ

Part 1.1: Modeling – Data Science Workflow

วัตถุประสงค์

ส่วนนี้มีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบ Data Science Workflow หลายรูปแบบ โดยแสดงให้เห็นว่าการปรับเปลี่ยนขั้นตอน preprocessing, feature selection และการเลือกโมเดลสามารถนำไปสู่การปรับปรุงประสิทธิภาพของโมเดลได้อย่างมีนัยสำคัญ

ตัวแปรเป้าหมายคือ Gender Ratio Class ซึ่งเป็นปัญหาแบบ multi-class classification (5 classes)

โครงสร้างของ Data Science Workflow

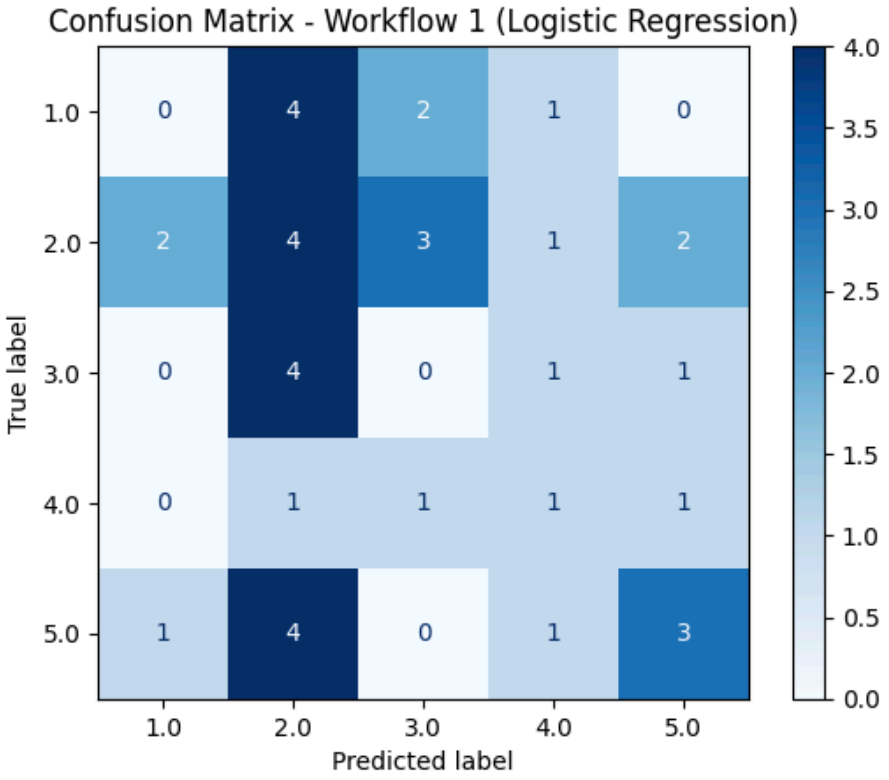
ในแต่ละ workflow ประกอบด้วยขั้นตอนหลัก 4 ขั้นตอน ได้แก่

1. Data preprocessing
2. Feature selection
3. Modeling
4. Performance evaluation

Workflow 1: Baseline Model (Logistic Regression)

Workflow แรกถูกออกแบบให้เป็น baseline model โดยใช้ Logistic Regression ซึ่งเป็นโมเดลเชิงเส้น ประกอบด้วยขั้นตอนการจัดการ missing values ด้วย median imputation และการปรับสเกลข้อมูลด้วย StandardScaler

ผลลัพธ์จาก workflow นี้พบว่าโมเดลมีประสิทธิภาพค่อนข้างต่ำ ซึ่งใกล้เคียงกับการทำนายแบบสุ่ม แสดงให้เห็นว่าความสัมพันธ์ระหว่างตัวแปรและ Gender Ratio Class มีความซับซ้อนและไม่สามารถอธิบายได้ด้วยเส้นตรงเพียงอย่างเดียว อย่างไรก็ตาม workflow นี้มีความสำคัญในฐานะจุดอ้างอิงสำหรับการเปรียบเทียบกับ workflow อื่น ๆ



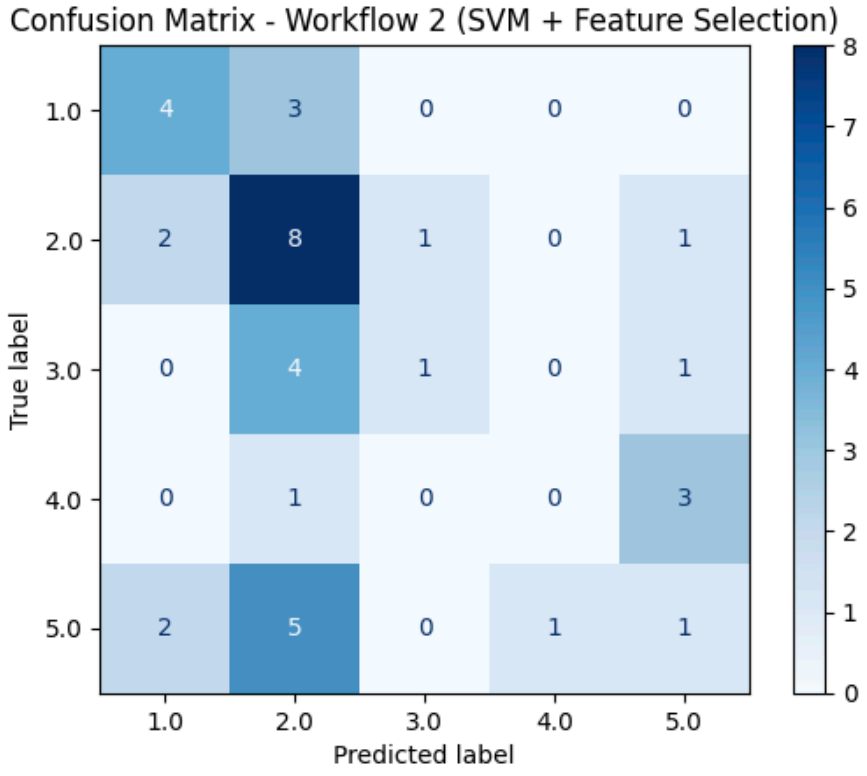

```
Workflow 1 - Logistic Regression

Accuracy: 0.21052631578947367

F1 (macro): 0.17461685823754788
```

Workflow 2: Feature Selection และ Non-linear Model (SVM)

Workflow ที่สองได้เพิ่มขั้นตอนการคัดเลือกตัวแปร โดยใช้ Variance Threshold เพื่อลบตัวแปรที่ไม่มีความแปรปรวน และ SelectKBest (ANOVA F-test) เพื่อเลือกตัวแปรที่มีความสัมพันธ์กับ Gender Ratio Class มากที่สุด จากนั้นใช้ Support Vector Machine (SVM) แบบ RBF kernel ซึ่งสามารถจับความสัมพันธ์แบบไม่เชิงเส้นได้

ใน workflow นี้ได้มีการนำขั้นตอน feature selection มาใช้เพื่อลดจำนวนตัวแปร โดยเริ่มจากการลบตัวแปรที่ไม่มีความแปรปรวน และเลือกเฉพาะตัวแปรที่มีความสัมพันธ์กับ Gender Ratio Class มากที่สุด การลดจำนวนตัวแปรช่วยลด noise ในข้อมูล และทำให้โมเดลสามารถเรียนรู้รูปแบบที่สำคัญได้ดีขึ้น ส่งผลให้ประสิทธิภาพของโมเดลเพิ่มขึ้นเมื่อเทียบกับ workflow แรก

ผลลัพธ์จาก workflow นี้แสดงให้เห็นว่าเมื่อมีการลดจำนวนตัวแปรและใช้โมเดลที่เหมาะสมกับลักษณะข้อมูล ประสิทธิภาพของโมเดลเพิ่มขึ้นอย่างชัดเจนเมื่อเทียบกับ workflow แรก สะท้อนให้เห็นถึงความสำคัญของ feature selection ในงาน data science



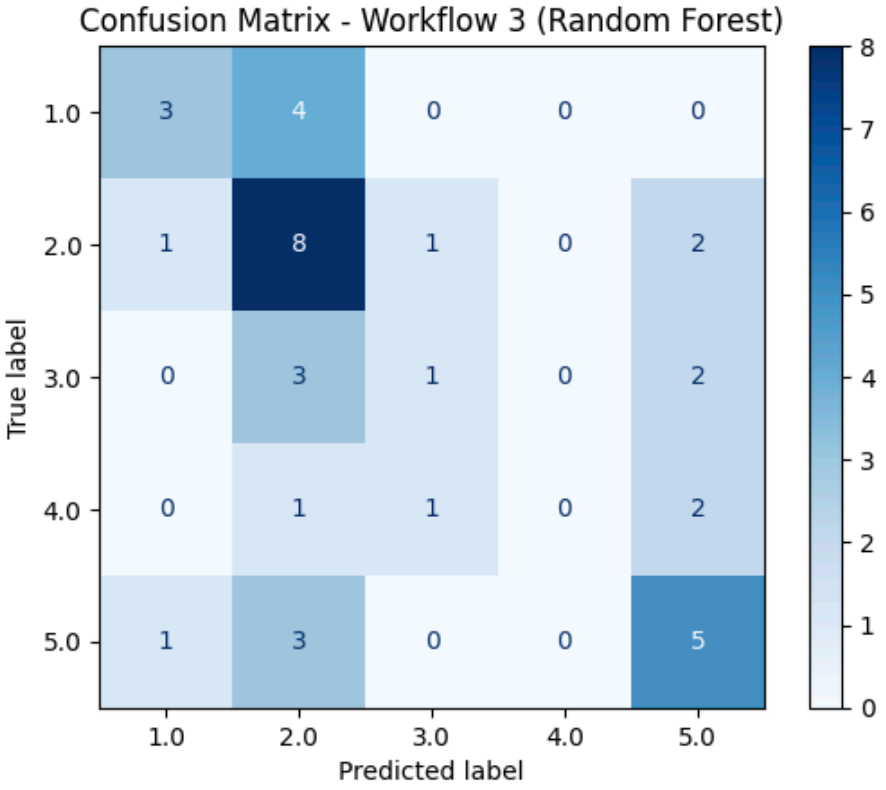

```
Workflow 2 - SVM + Feature Selection

Accuracy: 0.3684210526315789

F1 (macro): 0.2803030303030303
```

Workflow 3: Ensemble Model (Random Forest)

Workflow ที่สามใช้ Random Forest ซึ่งเป็น ensemble model ที่รวมการตัดสินใจจาก decision tree จำนวนมาก โมเดลนี้สามารถจับความสัมพันธ์ที่ซับซ้อนและปฏิสัมพันธ์ระหว่างตัวแปรได้ดี และมีความทนทานต่อ noise ในข้อมูล

แม้จะไม่ได้มีการลดจำนวนตัวแปรอย่างชัดเจนก่อนการสร้างโมเดล แต่ Random Forest สามารถทำการคัดเลือกตัวแปรได้โดยอัตโนมัติผ่านโครงสร้างของ decision tree และการรวมผลลัพธ์จากหลายต้นไม้ ทำให้โมเดลสามารถให้ความสำคัญกับตัวแปรที่มีผลต่อการทำนายมากที่สุด และลดผลกระทบจากตัวแปรที่ไม่สำคัญ

ผลการประเมินพบว่า Random Forest ให้ประสิทธิภาพสูงที่สุดในบรรดาทั้งสาม workflow ทั้งในแง่ของ Accuracy และ F1-score (macro) จึงถูกเลือกเป็น final model สำหรับการตีความผลลัพธ์ในขั้นตอนถัดไป




```
Workflow 3 - Random Forest

Accuracy: 0.4473684210526316

F1 (macro): 0.34767025089605735
```

Part 1.1.2: Result Interpretation

การประเมินประสิทธิภาพของโมเดล

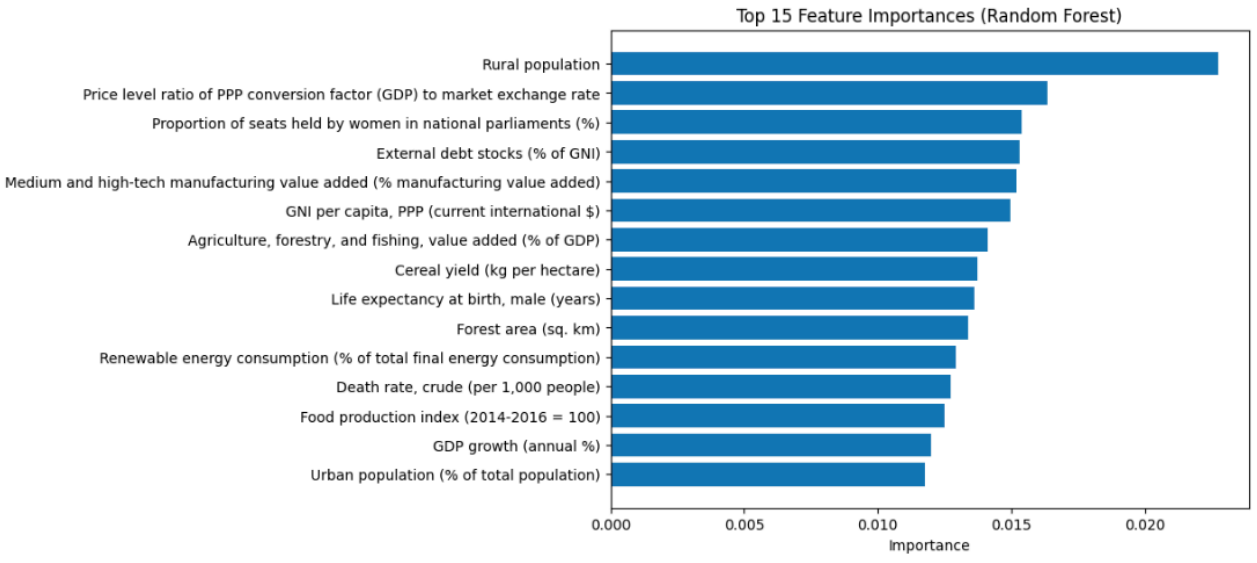
ประสิทธิภาพของ final model (Random Forest) ถูกประเมินโดยใช้ Accuracy (macro), F1-score (macro) และ confusion matrix เพื่อให้การประเมินมีความยุติธรรมต่อทุก class ในปัญหาแบบ multi-class

ผลลัพธ์ที่ได้แสดงให้เห็นว่าโมเดลมีประสิทธิภาพในระดับปานกลางถึงดี โดยสามารถทำนายได้ดีกว่าการสุ่มอย่างมีนัยสำคัญ confusion matrix ของ final model สามารถทำนาย Gender Ratio Class ในบางกลุ่มได้ค่อนข้างดี โดยเฉพาะกลุ่มที่มีลักษณะเด่นชัด อย่างไรก็ตาม ในบาง class โมเดลยังเกิดความสับสนระหว่างกลุ่มที่มีลักษณะใกล้เคียงกัน ซึ่งสะท้อนให้เห็นว่าข้อมูลของบางประเทศมีคุณลักษณะที่ทับซ้อนกันระหว่าง Gender Ratio Class ส่งผลให้การแยกกลุ่มทำได้ยากขึ้น แม้โมเดลจะมีประสิทธิภาพโดยรวมดีกว่า workflow ก่อนหน้า

การตีความตัวแปรสำคัญ

การตีความโมเดลทำโดยใช้ feature importance จาก Random Forest พบว่าตัวแปรที่เกี่ยวข้องกับโครงสร้างทางเศรษฐกิจและภาคเกษตร เช่น สัดส่วนมูลค่าเพิ่มจากภาคเกษตร และตัวชี้วัดด้านผลผลิตทางการเกษตร มีบทบาทสำคัญในการทำนาย Gender Ratio Class

ผลลัพธ์นี้ชี้ให้เห็นว่าโครงสร้างเศรษฐกิจและรูปแบบการใช้ทรัพยากรอาจมีความสัมพันธ์กับความแตกต่างด้านสัดส่วนแรงงานชายและหญิงในประเทศต่าง ๆ



การวิเคราะห์เชิงสถิติ: Region และ Gender Ratio Class

เพื่อศึกษาความสัมพันธ์ระหว่างภูมิภาคและ Gender Ratio Class ได้มีการใช้การทดสอบทางสถิติ 2 วิธี ได้แก่

- **Analysis of Variance (ANOVA)** เพื่อทดสอบว่าค่า Gender Ratio Class มีความแตกต่างกันระหว่างภูมิภาคหรือไม่

  จาก Analysis of Variance (ANOVA) พบว่า ค่า Gender Ratio Class มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติระหว่างภูมิภาค ($F = 3.68$, p-value < 0.001) แสดงให้เห็นว่าภูมิภาคมีผลต่อระดับ Gender Ratio Class ของประเทศ

- **Chi-square test of independence** เพื่อทดสอบความเป็นอิสระระหว่าง Region และ Gender Ratio Class

  จาก Chi-square test of independence แสดงให้เห็นว่า Region และ Gender Ratio Class มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติ ($\chi^2 = 98.37$, p-value < 0.001) ซึ่งบ่งชี้ว่า Gender Ratio Class ไม่เป็นอิสระจากภูมิภาค

ผลการทดสอบทางสถิติแสดงให้เห็นว่าภูมิภาคมีความสัมพันธ์กับ Gender Ratio Class อย่างมีนัยสำคัญทางสถิติ ซึ่งสนับสนุนแนวคิดว่าปัจจัยเชิงพื้นที่และบริบททางภูมิภาคมีบทบาทต่อโครงสร้างแรงงานชายและหญิง

Part 1.2: Feature Extraction

วัตถุประสงค์

ส่วน Feature Extraction มีวัตถุประสงค์เพื่อลดมิติของข้อมูลและอธิบายโครงสร้างที่ซ่อนอยู่ในตัวแปรจำนวนมาก โดยไม่มุ่งเน้นการเพิ่มประสิทธิภาพของโมเดลโดยตรง แต่เพื่อเพิ่มความเข้าใจเชิงโครงสร้างของข้อมูล

Principal Component Analysis (PCA)

PCA ถูกนำมาใช้เพื่อสร้างแกนข้อมูลใหม่ (Principal Components) ที่อธิบายความแปรปรวนของข้อมูลได้มากที่สุด ผลการวิเคราะห์พบว่า principal components แรก ๆ ถูกขับเคลื่อนโดยตัวแปรที่เกี่ยวข้องกับภาคเกษตร การใช้ที่ดิน และผลผลิตทางการเกษตร ซึ่งสะท้อนมิติด้านโครงสร้างเศรษฐกิจเกษตรของประเทศ

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...
Year	7.681803e-18	2.334773e-18	-6.336367e-18	-6.000941e-18	-2.049702e-17	-1.477782e-17	8.200864e-18	-1.326430e-18	1.883091e-17	7.566959e-18	...
Agricultural land (% of land area)	5.384150e-02	3.843797e-02	6.256265e-02	8.741548e-02	-1.280472e-01	-1.354887e-01	8.657909e-03	2.059272e-01	-3.686067e-02	6.524144e-02	...
Agriculture, forestry, and fishing, value added (% of GDP)	1.524610e-01	4.703353e-02	-7.397025e-03	9.316823e-02	1.450180e-03	-4.609286e-02	5.811121e-04	-9.980533e-02	1.497101e-03	1.395321e-02	...
Arable land (% of land area)	4.373193e-03	2.965068e-02	4.959370e-02	1.668074e-01	-1.234172e-01	-1.901800e-01	-6.153450e-02	-2.992795e-03	-5.363824e-02	1.569931e-01	...
Cereal yield (kg per hectare)	-9.691404e-02	-2.949840e-03	-9.971203e-02	-9.352466e-03	-1.350331e-03	2.121712e-02	9.870989e-03	-9.268204e-02	1.981993e-02	-3.873799e-02	...

ในการวิเคราะห์ด้วย Principal Component Analysis (PCA) ได้กำหนดจำนวน principal components โดยใช้เกณฑ์ cumulative explained variance อย่างน้อย 80% ของข้อมูล (n_components = 0.8) ส่งผลให้ได้จำนวน principal components ที่เพียงพอในการอธิบายโครงสร้างหลักของข้อมูล พร้อมทั้งยังสามารถตีความความหมายของแต่ละ component ได้จากค่า PCA loading

ค่า PCA loading ถูกนำมาใช้ในการวิเคราะห์ว่าตัวแปรใดมีบทบาทสำคัญในการสร้างแต่ละ principal component โดยพิจารณาจากค่า absolute ของ loading

Linear Discriminant Analysis (LDA)

LDA ถูกใช้เพื่อหาทิศทางที่สามารถแยก Gender Ratio Class ได้ดีที่สุด โดยใช้ข้อมูล class โดยตรง ผลการวิเคราะห์แสดงให้เห็นว่าตัวแปรด้านบทบาทภาคเกษตรและประสิทธิภาพการผลิตมีบทบาทสำคัญในการแยกกลุ่ม Gender Ratio Class

	LD1	LD2	LD3	LD4	LD5
Year	-2.201017e-10	7.608270e-11	2.806319e-11	1.172397e-11	4.641601e-11
Agricultural land (% of land area)	7.412304e-01	-2.708884e-01	7.474781e-01	8.225415e-01	-1.122360e+00
Agriculture, forestry, and fishing, value added (% of GDP)	2.791075e+00	-1.063661e+00	1.144167e+00	1.576483e-01	-1.656592e+00
Arable land (% of land area)	4.358785e-01	1.589379e-01	1.483776e-01	2.232681e+00	-1.662106e+00
Cereal yield (kg per hectare)	-7.982260e-01	-1.327362e+00	1.695044e-01	5.633935e-02	2.191315e+00

เนื่องจากตัวแปรเป้าหมาย Gender Ratio Class มีทั้งหมด 5 กลุ่ม จึงสามารถสร้าง linear discriminants ได้สูงสุด 4 แกน (LD1–LD4) โดยในการวิเคราะห์นี้ได้พิจารณา linear discriminants แรก ๆ ซึ่งสามารถอธิบายความแตกต่างระหว่างกลุ่ม Gender Ratio Class ได้ชัดเจนที่สุด

ค่า LDA coefficients ถูกใช้เพื่อระบุว่าตัวแปรใดมีบทบาทสำคัญต่อการแยกกลุ่ม โดยพิจารณาจากค่า absolute ของ coefficient และทิศทางของค่าเพื่ออธิบายลักษณะของการแบ่งกลุ่ม

Part 2: Exploratory Data Analysis, Clustering, Modeling, and Visualization

คำอธิบายข้อมูล (Data Explanation)

งานนี้ใช้ข้อมูลจาก 2 ชุดข้อมูล ได้แก่

Dataset C: ชุดข้อมูลจาก World Bank ซึ่งประกอบด้วยตัวแปรด้านเศรษฐกิจ สังคม สุขภาพ และการศึกษา ของหลายประเทศในหลายช่วงเวลา
Dataset B: ชุดข้อมูลภูมิภาคของประเทศ (Country Regions) ซึ่งเป็นชุดข้อมูลเดียวกับ Dataset B ใน Part 1

ทั้งสองชุดข้อมูลถูกนำมารวมกันโดยใช้ชื่อประเทศเป็นตัวเชื่อม เพื่อให้สามารถวิเคราะห์ความสัมพันธ์ระหว่างตัวชี้วัดทางเศรษฐกิจและข้อมูลเชิงภูมิภาคได้อย่างครบถ้วน ข้อมูลที่รวมแล้วถูกนำไปใช้ในทุกส่วนของ Part 2 (2.0–2.3)

Part 2.0: Exploratory Data Analysis (EDA)

การทำความสะอาดและเตรียมข้อมูล (Data Cleaning and Preparation)

ก่อนเริ่มการวิเคราะห์ ข้อมูลถูกเตรียมและทำความสะอาดตามขั้นตอนดังนี้:

1. เลือกใช้ข้อมูลเฉพาะปี **2017** เพื่อหลีกเลี่ยงความซ้ำซ้อนจากหลายปี และเพื่อให้การเปรียบเทียบประเทศเป็นไปในช่วงเวลาเดียวกัน

2. เลือกเฉพาะตัวแปรที่เกี่ยวข้องกับการวิเคราะห์ เช่น GNI per capita, ตัวแปรด้านการศึกษา สุขภาพ และเศรษฐกิจ

3. เปลี่ยนชื่อคอลัมน์ให้สั้นและสอดคล้องกัน เพื่อความสะดวกในการประมวลผลและการรวมข้อมูล

4. จัดการค่าที่ขาดหาย (missing values) โดยลบแถวที่ไม่มีค่าของตัวแปรเป้าหมาย (GNI per capita)

5. รวม Dataset C กับ Dataset B เพื่อเพิ่มข้อมูลภูมิภาคของประเทศ

จากการทำ EDA พบว่า **GNI per capita มีการกระจายแบบเบ้ขวา (right-skewed)** แสดงให้เห็นถึงความเหลื่อมล้ำของรายได้ระหว่างประเทศ ซึ่งส่งผลต่อการเลือกเทคนิคในขั้นตอนการวิเคราะห์ถัดไป เช่น การแปลงข้อมูลด้วย log

Part 2.1: Clustering

Clustering Method

ในส่วนนี้ใช้ **K-Means Clustering** เพื่อจัดกลุ่มประเทศตามตัวแปร

GNI per capita, PPP (current international $)

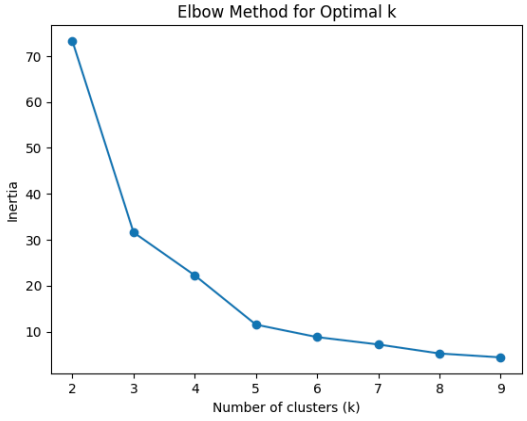
K-Means ถูกเลือกเนื่องจาก:

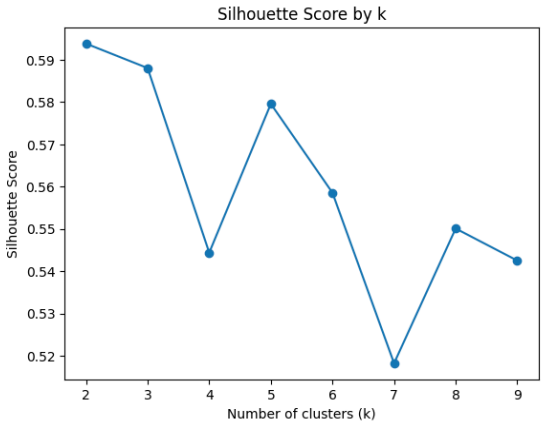
- เป็นอัลกอริทึมที่เหมาะกับข้อมูลเชิงตัวเลข
- สามารถตีความผลลัพธ์ได้ง่าย
- เหมาะกับการแบ่งประเทศตามระดับรายได้

Number of Clusters

จำนวนกลุ่มถูกกำหนดโดยใช้:

Elbow Method เพื่อพิจารณาความคุ้มค่าของการเพิ่มจำนวนกลุ่ม
Silhouette Score เพื่อประเมินคุณภาพของการจัดกลุ่ม





จากทั้งสองวิธี เลือกจำนวนกลุ่มสุดท้ายเป็น **3 กลุ่ม** ซึ่งให้ความสมดุลระหว่างความเรียบง่ายและคุณภาพของการจัดกลุ่ม

ลักษณะของแต่ละกลุ่มประเทศ

Cluster 0 (รายได้ต่ำ)
ประเทศในกลุ่มนี้มีค่า GNI per capita ต่ำ มีข้อจำกัดด้านโครงสร้างพื้นฐาน การศึกษา และสุขภาพ

Cluster 1 (รายได้ปานกลาง)
ประเทศในกลุ่มนี้มีรายได้ระดับกลาง มีการพัฒนาด้านการศึกษาและเศรษฐกิจในระดับหนึ่ง

Cluster 2 (รายได้สูง)
ประเทศในกลุ่มนี้มีค่า GNI per capita สูง ระบบการศึกษาและสุขภาพมีคุณภาพสูง และมีเศรษฐกิจที่มั่นคง

Part 2.2: Modeling

วัตถุประสงค์ - สร้างโมเดลเพื่อทำนาย

GNI per capita, PPP (current international $)

โดยใช้ตัวแปรอื่น ๆ และไม่ใช้ GNI เป็นตัวแปรอิสระ

Workflow 1: Baseline Model

(1.A) Data Preprocessing

- เลือกข้อมูลปี 2017
- ลบข้อมูลที่มี missing values
- ใช้ตัวแปรด้านเศรษฐกิจ สุขภาพ และการศึกษา

(1.B) Feature Selection

เลือกตัวแปรที่มีเหตุผลเชิงทฤษฎี เช่น:

- GDP (current US$)
- Life expectancy at birth
- School enrollment (secondary)
- Unemployment rate
- Urban population (%)

(1.C) Modeling

ใช้ **Linear Regression** เป็นโมเดลพื้นฐาน เนื่องจากเข้าใจง่ายและใช้เป็น baseline สำหรับเปรียบเทียบ

(1.D) Performance Evaluation

ประเมินโมเดลด้วย:

- RMSE
- MAPE
- R-squared

ผลลัพธ์จาก Workflow แรกแสดงให้เห็นว่าโมเดลสามารถอธิบายแนวโน้มได้ในระดับหนึ่ง แต่ยังมีข้อจำกัดในการจับความสัมพันธ์ที่ไม่เป็นเชิงเส้น

Workflow 2: Improved Model

(1.A) Data Preprocessing

- แปลงค่า GNI per capita ด้วย log เพื่อแก้ปัญหาการกระจายแบบเบ้
- ทำ scaling ให้ตัวแปรอยู่ในช่วงเดียวกัน

(1.B) Feature Selection

ใช้ชุดตัวแปรเดียวกับ Workflow แรก แต่ลดจำนวนตัวแปรให้เหลือเฉพาะตัวแปรที่สำคัญ

(1.C) Modeling

ใช้ **Random Forest Regressor** เพื่อจับความสัมพันธ์ที่ไม่เป็นเชิงเส้น

(1.D) Performance Evaluation

ประเมินด้วย:

- RMSE
- MAPE
- R-squared

ผลลัพธ์พบว่า Workflow ที่สองให้ประสิทธิภาพที่ดีกว่าอย่างชัดเจน โดยมีค่า R-squared ประมาณ **0.86** และ MAPE ต่ำกว่า 5% ซึ่งถือว่าดีมากสำหรับข้อมูลระดับประเทศ

การตีความผลลัพธ์ของโมเดลสุดท้าย

จากการวิเคราะห์ Feature Importance พบว่า:

Life expectancy at birth, total (years)	0.735210
School enrollment, secondary (% gross)	0.158401
Unemployment, total (% of total labor force) (modeled ILO estimate)	0.037762
GDP (current US$)	0.037420
Urban population (% of total population)	0.031207

- **Life expectancy at birth** มีผลต่อ GNI per capita มากที่สุด
- รองลงมาคือ **School enrollment (secondary)**
- ตัวแปรด้าน GDP และโครงสร้างประชากรมีผลรองลงมา

ผลลัพธ์สะท้อนให้เห็นว่าปัจจัยด้าน **ทรัพยากรมนุษย์ (Human Capital)** มีบทบาทสำคัญต่อรายได้ประเทศ

Part 2.3: Visualization

วัตถุประสงค์ของ Dashboard

Dashboard ถูกออกแบบมาเพื่อแสดงความสัมพันธ์ระหว่าง

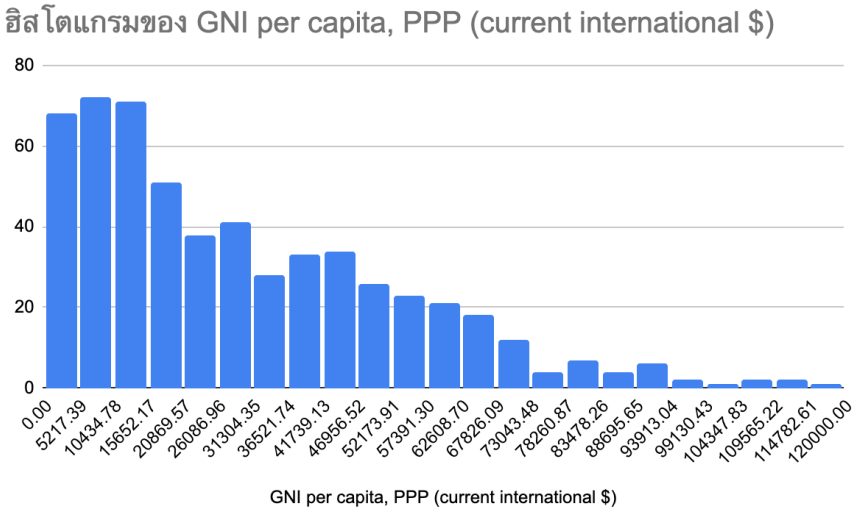
GNI per capita และตัวแปรด้านการศึกษา

ในมุมมองของการกระจาย ความสัมพันธ์ โครงสร้าง และแนวโน้มตามเวลา

คำอธิบายกราฟแต่ละประเภท

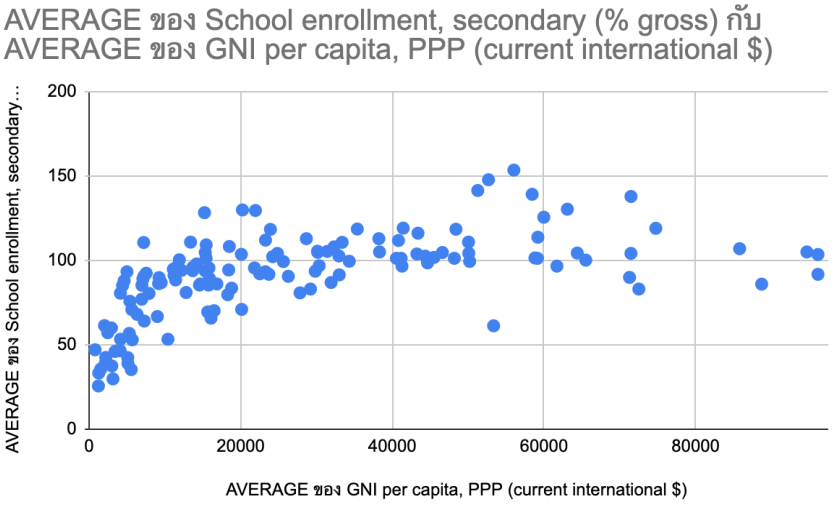
1. **Distribution (Histogram)**

   ใช้แสดงการกระจายของ GNI per capita เพื่อสะท้อนความเหลื่อมล้ำของรายได้ระหว่างประเทศ



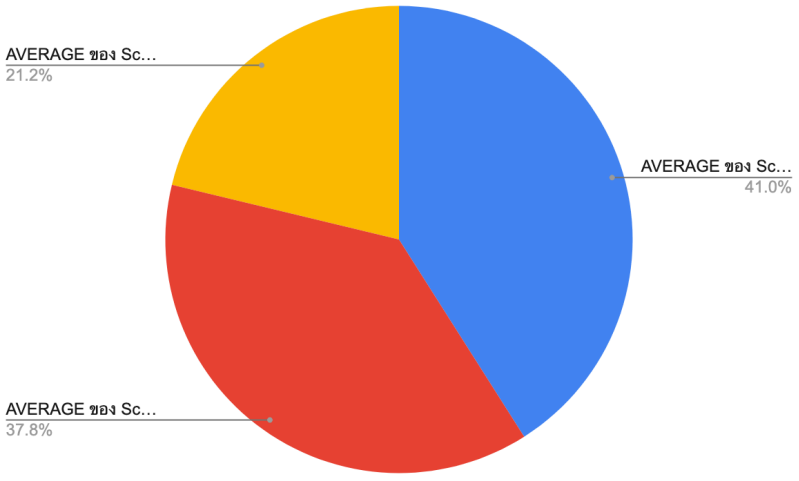
2. **Correlation (Scatter Chart)**

แสดงความสัมพันธ์ระหว่าง GNI per capita และ School enrollment (secondary) เพื่อศึกษาความเชื่อมโยงระหว่างรายได้และการศึกษา



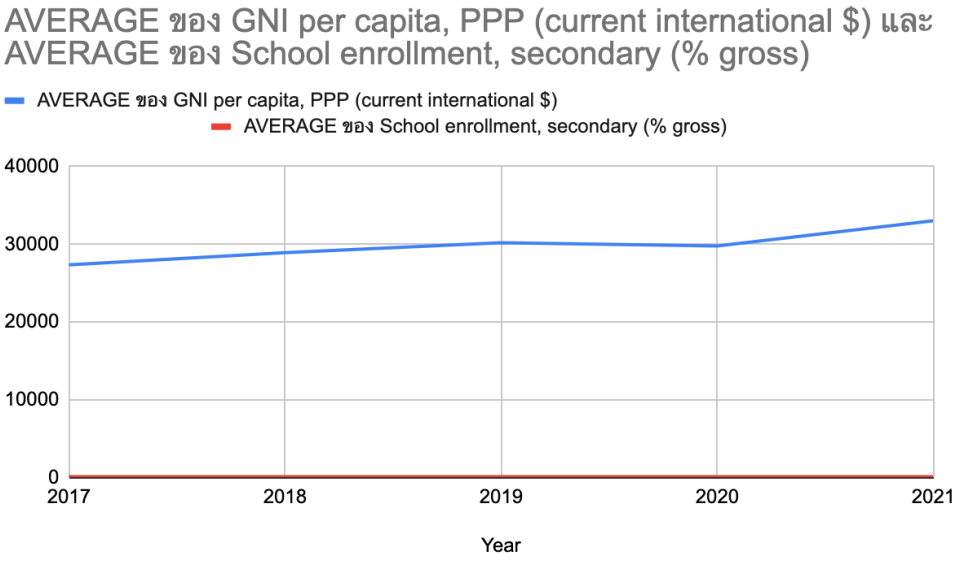
3. **Part-to-whole (Pie Chart)**

แสดงสัดส่วนการเข้าเรียนในระดับ Primary, Secondary และ Tertiary เพื่อดูโครงสร้างระบบการศึกษา



4. **Timeseries (Line Chart with Secondary Axis)**

แสดงแนวโน้มของ GNI per capita และ School enrollment ตามเวลา โดยใช้แกนรองเพื่อจัดการกับหน่วยที่แตกต่างกัน



Narrative Visual Structure

Dashboard ถูกออกแบบให้เล่าเรื่องข้อมูลจากภาพรวมไปสู่รายละเอียด โดยเริ่มจากการแสดงการกระจายของรายได้ประเทศ (Distribution) เพื่อสะท้อนความเหลื่อมล้ำ จากนั้นนำเสนอความสัมพันธ์ระหว่างรายได้และการศึกษา (Correlation) ต่อด้วยโครงสร้างของระบบการศึกษาในแต่ละระดับ (Part-to-whole) และปิดท้ายด้วย แนวโน้มการเปลี่ยนแปลงของรายได้และการศึกษาตามเวลา (Timeseries) โครงสร้างนี้ช่วยให้ผู้อ่านเข้าใจทั้งภาพรวม ความสัมพันธ์ และพัฒนาการของข้อมูลได้อย่างเป็นลำดับ