

การวิเคราะห์การจัดกลุ่มข้อมูลเพื่อระบุบุคคลไทยที่มีความเสี่ยงสูงต่อความรุนแรงในครอบครัวและปัญหาสุขภาพจิต โดยใช้เทคนิคการเรียนรู้ของเครื่อง

Data clustering analysis to identify high-risk Thai individuals for domestic violence and mental health problems using machine learning

วุฒิพงษ์ เชื้อนดิน และชนิดา แก้วเพชร^{1*}

Wutthiphong Khuandin¹ and Chanida Kaewphet^{1*}

Received: 17 March 2025 ; Revised: 8 May 2025 ; Accepted: 13 June 2025

บทคัดย่อ

จากสถานการณ์ความรุนแรงในครอบครัวและปัญหาสุขภาพจิตในประเทศไทยยังคงเป็นปัญหาสำคัญที่ส่งผลกระทบต่อคุณภาพชีวิตและความมั่นคงของสังคม โดยเฉพาะอย่างยิ่งในช่วงวิกฤตเศรษฐกิจและการแพร่ระบาดของโรค COVID-19 ที่มีแนวโน้มการเพิ่มขึ้นของเหตุการณ์ความรุนแรงภายในครอบครัว อย่างไรก็ตาม งานวิจัยที่ผ่านมายังขาดการวิเคราะห์เชิงลึกโดยใช้เทคนิควิทยาการข้อมูลเพื่อจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูง การวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกกลุ่มบุคคลไทยที่มีความเสี่ยงสูงต่อปัญหาสุขภาพจิตและความรุนแรงในครอบครัว โดยประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ด้วยวิธีการจัดกลุ่มข้อมูลแบบไม่มีการกำกับ (Unsupervised Learning) ได้แก่ K-means, Hierarchical Clustering (HC) และ Gaussian Mixture Models (GMMs) และเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคด้วยตัวชี้วัดภายในเพื่อหาวิธีการที่เหมาะสมที่สุดในการแยกกลุ่มความเสี่ยง การวิจัยครั้งนี้ใช้ข้อมูลผู้กระทำความรุนแรงในครอบครัวจำนวน 1,162 รายการที่ได้จากแหล่งข้อมูลสาธารณะ สำนักงานพัฒนารัฐบาลดิจิทัล (DGA) ผลการวิเคราะห์พบว่า HC มีประสิทธิภาพดีที่สุด (Silhouette Score = 0.429, Calinski-Harabasz Index = 61.790, Davies-Bouldin Index = 1.034) สามารถจำแนกกลุ่มความเสี่ยงได้ชัดเจน โดยกลุ่มเสี่ยงสูงมีลักษณะสำคัญคือ เพศชาย อายุช่วงวัยกลางคน มีปัญหาสุขภาพจิต มีการใช้สารเสพติด และความเครียดทางเศรษฐกิจ ผลการวิจัยนี้สะท้อนให้เห็นถึงศักยภาพของการใช้เทคนิคการเรียนรู้ของเครื่องในการจำแนกกลุ่มเสี่ยงเพื่อสนับสนุนการออกแบบมาตรการป้องกันเฉพาะกลุ่มได้อย่างมีประสิทธิภาพ หรือหน่วยงานที่เกี่ยวข้องสามารถนำผลการศึกษานี้ไปพัฒนาระบบแจ้งเตือนความเสี่ยงเชิงป้องกันหรือวางแผนเชิงนโยบายเพื่อป้องกันความรุนแรงในครอบครัวและส่งเสริมสุขภาพจิตของประชาชนอย่างยั่งยืน

คำสำคัญ: สุขภาพจิต, ความรุนแรงในครอบครัว, การจัดกลุ่มข้อมูล

Abstract

Domestic violence and mental health problems remain critical public health concerns in Thailand, with an increasing prevalence observed during economic crises and the COVID-19 pandemic. Despite their significant societal impact, previous research has lacked in-depth analytical approaches utilizing data science to identify high-risk populations. This study aims to classify high-risk Thai individuals vulnerable to mental health issues and domestic violence by applying unsupervised machine learning techniques, specifically K-means, Hierarchical Clustering (HC), and Gaussian Mixture Models (GMMs). Internal evaluation metrics, including the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index, were used to assess clustering performance. The dataset comprised 1,162 records of domestic violence offenders obtained from the Digital Government Development Agency (DGA) in Thailand.

¹ สาขาวิชาระบบสารสนเทศและคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ

¹ Major of Information System and Business Computer, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Suvarnabhumi

* Corresponding author; E-mail: chanida.k@rmutsb.ac.th

The findings indicate that HC achieved the highest performance (Silhouette Score = 0.429, Calinski-Harabasz Index = 61.790, and Davies-Bouldin Index = 1.034), effectively differentiating risk groups. The high-risk group was predominantly characterized by middle-aged males with mental health issues, substance abuse, and economic stress. This study demonstrates the potential of machine learning for identifying vulnerable populations. It provides insights that can inform the development of targeted prevention strategies, early warning systems, and evidence-based policymaking to mitigate domestic violence and promote sustainable mental well-being.

Keywords: Mental health, domestic violence, data clustering

บทนำ

องค์การอนามัยโลกพบว่าความรุนแรงของคู่สมรสและการใช้ความรุนแรงในครอบครัวเป็นปัญหาสุขภาพ ปัญหาทางสังคมและปัญหาด้านสาธารณสุขที่สำคัญทั่วโลก ซึ่งส่งผลกระทบต่อสุขภาพร่างกายและจิตใจของผู้ที่ตกเป็นเหยื่อ (World Health Organization, 2024) โดยสถานการณ์การระบาดของโรค COVID-19 ส่งผลกระทบต่อภาวะเศรษฐกิจและสังคม ทำให้หลายครอบครัวมีรายได้ลดลง เกิดความเครียดและเสี่ยงต่อการใช้ความรุนแรงในครอบครัวเพิ่มขึ้น นอกจากนี้ การเปลี่ยนแปลงรูปแบบการใช้ชีวิตประจำวันอย่างกะทันหัน การทำงานที่บ้าน การตกงาน และการกักตัว ทำให้สมาชิกในครอบครัวมีโอกาสอยู่ร่วมกันมากขึ้น ซึ่งอาจนำไปสู่ความขัดแย้งและความรุนแรงได้มากขึ้น เหยื่อส่วนใหญ่เป็นเพศหญิง (85%) มีอายุระหว่าง 36-59 ปี (45%) มีความสัมพันธ์กับผู้กระทำในฐานคู่สมรส (60%) และผู้กระทำไม่ได้ใช้สารเสพติด (70%) ไม่มีปัญหาสุขภาพจิต (65%) ไม่มีปัญหาการพนัน (80%) และไม่มีปัญหาเศรษฐกิจ (75%) นอกจากนี้ ผลการศึกษายังพบว่าส่วนใหญ่เคยถูกผู้กระทำความรุนแรงในครอบครัวอย่างน้อย 1 ครั้ง (70%) โดยรูปแบบความรุนแรงที่พบบ่อยที่สุดคือ ความรุนแรงทางร่างกาย (60%) รองลงมาคือ ความรุนแรงทางจิตใจ (50%) และความรุนแรงทางเพศ (30%) และสำหรับประเทศไทยในช่วงสถานการณ์การแพร่ระบาดของโรค COVID-19 ความรุนแรงในครอบครัวที่เกิดขึ้นที่จังหวัดกรุงเทพมหานคร มีแนวโน้มเพิ่มสูงขึ้นเมื่อเทียบกับช่วงก่อนการระบาด โดยจากการสำรวจในช่วงการระบาดครั้งที่ 1 (มกราคม - พฤษภาคม 2563) พบว่า ความรุนแรงในครอบครัวมีความชุกอยู่ที่ 42.9% และแม้จะลดลงเล็กน้อยเหลือ 37.4% ในช่วงการระบาดครั้งที่ 2 (กรกฎาคม 2563 - มกราคม 2564) แต่ยังคงถือว่าสูงกว่าอัตราความรุนแรงในครอบครัวก่อนเกิดการระบาดซึ่งอยู่ที่ 26% สะท้อนให้เห็นถึงผลกระทบทางเศรษฐกิจและความเครียดในครอบครัวอันเป็นปัจจัยสำคัญที่เกี่ยวข้องกับการเกิดความรุนแรงในครอบครัวในช่วงวิกฤตโรคระบาด (พิทยา สังข์แก้ว และคณะ, 2565)

จากงานวิจัยที่ผ่านมาพบว่ายังขาดการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลเชิงลึก เพื่อวิเคราะห์ปัจจัยเสี่ยงของความรุนแรงในครอบครัวในบริบทของประเทศไทย แม้ว่าการวิจัยที่ผ่านมาจะมีการวิเคราะห์ปัจจัยที่เกี่ยวข้องกับพฤติกรรมความรุนแรงบ้างแล้ว แต่ส่วนใหญ่มักใช้วิธีการวิเคราะห์เชิงสถิติพื้นฐานโดยไม่ได้ใช้เทคนิคเชิง Machine Learning (ML) เพื่อจำแนกกลุ่มเสี่ยงอย่างเป็นระบบ (อรรถพล ดำเนินผล, 2564) นอกจากนี้ ยังมีจำนวนน้อยมากที่ใช้ข้อมูลจริงจากฐานข้อมูลสาธารณะเพื่อสร้างแบบจำลองการประเมินความเสี่ยงอย่างเป็นรูปธรรม ในประเทศไทยสำนักงานพัฒนารัฐบาลดิจิทัล (สำนักงานพัฒนารัฐบาลดิจิทัล, 2567) ได้เผยแพร่ข้อมูลเปิดเกี่ยวกับผู้กระทำความรุนแรงในครอบครัว ซึ่งเป็นแหล่งข้อมูลที่มีคุณค่าในการนำมาศึกษาและวิเคราะห์รูปแบบปัจจัยเสี่ยงเพื่อจำแนกกลุ่มบุคคลที่มีแนวโน้มในการก่อความรุนแรง ข้อมูลนี้ประกอบด้วยรายละเอียดด้านประชากรศาสตร์ พฤติกรรมเสี่ยง สถานะทางจิตใจ และปัจจัยเศรษฐกิจซึ่งสามารถนำมาประยุกต์ใช้ร่วมกับเทคนิคทางวิทยาการข้อมูลเพื่อสร้างแบบจำลองที่แม่นยำในการระบุความเสี่ยง

ความสำคัญของงานวิจัยนี้คือการประยุกต์ใช้ Machine Learning ในการจำแนกกลุ่มข้อมูล สำหรับการวิเคราะห์ปัจจัยเสี่ยงของบุคคลที่มีแนวโน้มก่อความรุนแรงในครอบครัวแบบดั้งเดิมมักใช้วิธีการสถิติพื้นฐาน เช่น การวิเคราะห์การถดถอย ซึ่งถึงแม้จะสามารถระบุความสัมพันธ์ของตัวแปรได้ดี แต่ยังมีข้อจำกัดหลายด้าน ได้แก่ (1) ไม่สามารถค้นหารูปแบบที่ซับซ้อนได้ วิธีการเชิงสถิติอาจไม่สามารถแยกแยะความสัมพันธ์เชิงซ้อนของตัวแปรต่าง ๆ ได้ดีเท่ากับ Machine Learning (2) ความสามารถในการปรับตัว เนื่องจากสามารถเรียนรู้จากข้อมูลใหม่ ๆ และสามารถปรับปรุงความแม่นยำได้โดยอัตโนมัติ และสุดท้ายคือการจัดกลุ่มข้อมูลโดยไม่ต้องมีข้อมูลป้ายกำกับ (Unsupervised Learning) เป็นเทคนิคการจัดกลุ่มช่วยให้สามารถจำแนกกลุ่มความเสี่ยงได้แม้ไม่มีข้อมูลป้ายกำกับล่วงหน้า

จากข้อได้เปรียบข้างต้น งานวิจัยนี้จึงเลือกใช้ Machine Learning แบบ Unsupervised Learning ได้แก่ K-means,

Hierarchical Clustering (HC) และ Gaussian Mixture Models (GMMs) เพื่อจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูงและสามารถนำไปใช้พัฒนามาตรการป้องกันในเชิงนโยบายได้อย่างมีประสิทธิภาพ (ณัฐวุฒิ แถมเงิน และคณะ, 2567) และมีจุดมุ่งหมาย เพื่อเติมเต็มช่องว่างดังกล่าว โดยการนำเทคนิคการจัดกลุ่มข้อมูลทั้งสามรูปแบบมาเปรียบเทียบประสิทธิภาพการจัดกลุ่ม พร้อมประเมินด้วยดัชนีที่ได้รับการยอมรับ เช่น Silhouette Score, Calinski-Harabasz Index และ Davies-Bouldin Index เพื่อให้ได้แบบจำลองที่สามารถนำไปใช้ในการกำหนดนโยบายและมาตรการป้องกันได้อย่างมีประสิทธิภาพ สำหรับวัตถุประสงค์ของการวิจัยคือ (1) เพื่อจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูงต่อการใช้ความรุนแรงในครอบครัวและปัญหาสุขภาพจิตด้วยการจัดกลุ่มข้อมูลแบบ Unsupervised Learning และ (2) เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจัดกลุ่มแต่ละแบบโดยใช้เกณฑ์การประเมินโมเดล เพื่อเลือกวิธีการที่เหมาะสมที่สุดในการแยกกลุ่มความเสี่ยง การศึกษานี้คาดว่าจะมีส่วนช่วยให้หน่วยงานที่เกี่ยวข้องสามารถใช้ข้อมูลเชิงลึกในการออกแบบแผนการแทรกแซงอย่างเหมาะสมกับแต่ละกลุ่มความเสี่ยง เพื่อป้องกันและลดปัญหาความรุนแรงในครอบครัวได้อย่างยั่งยืน

วิธีดำเนินการวิจัย

ประกอบไปด้วย 6 ขั้นตอน (Figure 1) ดังนี้

1. วิธีการเก็บรวบรวมข้อมูล (Data Collection)

ข้อมูลที่ใช้ในการวิจัยครั้งนี้คือข้อมูลผู้กระทำความรุนแรงในครอบครัว จำแนกตามปัญหาหรือสาเหตุรายประเด็น ที่เผยแพร่รูปแบบสาธารณะในเว็บไซต์ของสำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน) (สพร.) ในกลุ่มชุดข้อมูลสังคมและสวัสดิการ ตั้งแต่วันที่ 28 มีนาคม 2567 ถึง 24 กุมภาพันธ์ 2568 จำนวน 1,162 รายการ จากเว็บไซต์ <https://data.go.th/> สืบค้นเมื่อ 17 มกราคม 2568 สำหรับการนำข้อมูลเข้า ผู้วิจัยทำการดาวน์โหลดข้อมูลจากเว็บไซต์ ซึ่งข้อมูลถูกจัดเก็บอยู่ในรูปแบบไฟล์ csv โดยใช้ไลบรารี pandas ในภาษา Python

2. การเตรียมข้อมูล (Data preparation)

เพื่อตรวจสอบความถูกต้องของข้อมูลและทำให้ข้อมูลพร้อมสำหรับการวิเคราะห์ข้อมูล (Georgiou, 2024) ผู้วิจัยเลือกใช้ scikit-learn ในภาษา Python

การคัดเลือกแอตทริบิวต์ที่เป็นการพิจารณาความถูกต้อง ครบถ้วนของแต่ละข้อมูลที่จะถูกนำมาใช้ในงานวิจัยนี้ เพราะข้อมูลที่ถูกเลือกนำมาใช้มีค่าในแต่ละแอตทริบิวต์เป็นข้อมูลเชิงคุณภาพ (Qualitative data) จากข้อมูลทั้งหมด

1,162 รายการ 19 แอตทริบิวต์ ผู้วิจัยพิจารณาเลือกใช้ปัจจัยสำหรับจำแนกกลุ่มจำนวน 12 แอตทริบิวต์

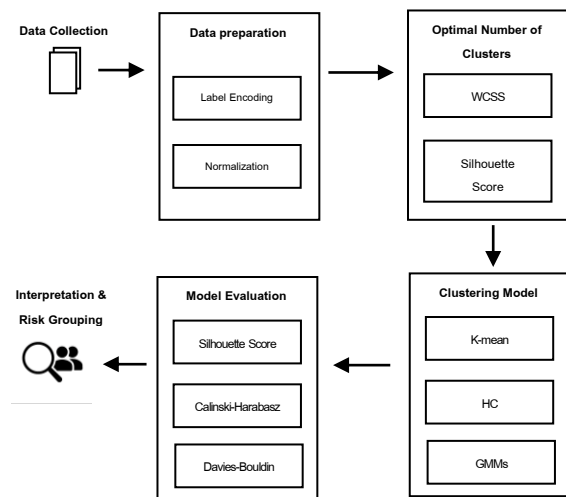


Figure 1 Cluster Analysis Pipeline

จากการศึกษาข้อมูลเชิงสถิติจาก Table 1 พบว่าปัจจัยเสี่ยงต่อปัญหาสุขภาพจิตและความรุนแรงในครอบครัวของคนไทยสามารถจำแนกออกเป็น 5 กลุ่มหลัก ได้แก่ ปัจจัยด้านประชากรศาสตร์ พบว่าผู้ก่อเหตุเป็นชายและหญิงในสัดส่วนใกล้เคียงกัน โดยมีอายุเฉลี่ย 38.32 ปี สะท้อนว่าปัญหานี้สามารถเกิดขึ้นได้ในทุกเพศและวัยกลางคนเป็นช่วงที่พบมากที่สุด ปัจจัยพฤติกรรมเสี่ยง การดื่มแอลกอฮอล์ และการใช้สารเสพติด เป็นปัจจัยที่สัมพันธ์กับพฤติกรรมรุนแรงและปัญหาสุขภาพจิตอย่างมีนัยสำคัญ นอกจากนี้ปัจจัยด้านครอบครัวและอารมณ์ พบว่าความโกรธที่รุนแรง และความหึงหวงมีความเกี่ยวข้องกับพฤติกรรมความรุนแรง ขณะที่การหย่าร้างไม่ใช่ปัจจัยสำคัญ ปัจจัยด้านสุขภาพและปัญหาสุขภาพจิตมีความสัมพันธ์กับพฤติกรรมเสี่ยงมากกว่าสุขภาพกาย การคิดการพนันพบน้อยมากที่สุดในปัจจัยทั้งหมด และสุดท้ายคือปัจจัยด้านเศรษฐกิจ ความเครียดทางเศรษฐกิจอาจเป็นตัวกระตุ้นให้เกิดปัญหายาในครอบครัวและภาวะทางจิตที่นำไปสู่ความรุนแรงได้เช่นกัน

2.1 การแปลงข้อมูลประเภทหมวดหมู่ให้เป็นข้อมูลเชิงตัวเลขด้วยวิธี LabelEncoder ซึ่งเป็นเครื่องมือที่สามารถนำไปใช้ในการวิเคราะห์ข้อมูลและการสร้างโมเดลการเรียนรู้ของเครื่องได้ง่ายขึ้น (Hasan *et al.*, 2020)

2.2 การจัดการข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการใช้งาน ด้วยการใช้พารามิเตอร์ MinMaxScaler ทำ Normalize เพื่อให้ข้อมูลทุกตัวแปรอยู่ในช่วง 0 - 1 ช่วยลดผลกระทบจากสเกลที่ไม่เท่ากันในโมเดลที่อิงกับระยะห่าง เพราะงานวิจัยนี้ใช้โมเดล K-means, HC และ GMMs ซึ่งจะ

คำนวณระยะห่างสำหรับจัดกลุ่ม เพื่อให้ผลลัพธ์ของการจัดกลุ่มมีความแม่นยำมากขึ้น

Table 1 Risk and Demographic Profiles

Factor	Data Proportion
Gender	Female: 581 (50.04%), Male: 578 (49.78%), Not specified: 2 (0.17%)
Age	Average age: 38.32 years
Alcohol	Alcohol consumption: 788 (67.87%), No alcohol consumption: 378 (32.13%)
Drug Use	Drug use: 728 (62.70%), No drug use: 433 (37.30%)
Authoritative Behavior	Present: 301 (25.93%), Not present: 860 (74.07%)
Rage	Present: 820 (70.63%), Not present: 341 (29.37%)
Jealousy	Present: 239 (20.59%), Not present: 922 (79.41%)
Divorce	Not divorced: 1,123 (96.73%), Divorced: 38 (3.27%)
Health Problems	No health problems: 1,123 (96.73%), Health problems: 38 (3.27%)
Mental Health Problems	No mental health problems: 975 (83.98%), Mental health problems: 186 (16.02%)
Gambling Addiction	Not addicted: 1,135 (97.76%), Addicted: 26 (2.24%)
Economic Stress	No stress: 1,008 (86.82%), Economic stress: 153 (13.18%)

3. การวิเคราะห์เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด (Optimal Number of Clusters)

เป็นกระบวนการวิเคราะห์ข้อมูลเพื่อกำหนดจำนวนกลุ่ม (k) ที่เหมาะสมที่สุดในการจัดกลุ่มข้อมูล โดยไม่มากเกินไปส่งผลต่อคุณภาพของการจัดกลุ่ม ผู้วิจัยได้ทำการวิเคราะห์เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด โดยใช้วิธีดังนี้

3.1 Elbow Method เป็นกระบวนการวิเคราะห์เพื่อเลือกจำนวนกลุ่มที่เหมาะสมที่สุดในงานจัดกลุ่มข้อมูล โดยเฉพาะการประยุกต์ใช้ใน K-means clustering (BuiltIn, 2024) วิธีการนี้อาศัยการคำนวณค่า Within-Cluster Sum of Squares (WCSS) หรือค่า Sum of Squared Errors (SSE) ดังสมการที่ (1) ซึ่งเป็นผลรวมของระยะทางกำลังสองระหว่างข้อมูลแต่ละจุดกับ centroid ของกลุ่มที่ข้อมูลนั้นสังกัดอยู่ (Analytics Vidhya, 2021) จากนั้นทำการวาดกราฟแสดงความสัมพันธ์ระหว่างค่า WCSS และจำนวนกลุ่มที่กำหนด เพื่อค้นหาจุดที่การลดลงของ WCSS เริ่มชะลอตัวหรือจุดหัก

ศอก(GeeksforGeeks, 2020) จุดดังกล่าวสะท้อนถึงจำนวนกลุ่มที่เหมาะสมที่สุด เนื่องจากการเพิ่มจำนวนกลุ่มหลังจากจุดนี้จะไม่ทำให้ประสิทธิภาพการจัดกลุ่มดีขึ้นอย่างมีนัยสำคัญอีกต่อไป และได้ผลการคำนวณค่าดัง Figure 2 (a) พบว่าจำนวนกลุ่มที่ดีที่สุดคือ 3 กลุ่ม

$$WCSS = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (1)$$

โดยที่:

K คือ จำนวนกลุ่มทั้งหมดที่กำหนด

C_k คือ เซตของข้อมูลในกลุ่มที่ k

x_i คือ เวกเตอร์ของจุดข้อมูลลำดับที่ i

μ_k คือ เวกเตอร์ของจุดศูนย์กลางของกลุ่มที่ k

$\|\cdot\|$ คือ ระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง

(Euclidean Distance)

3.2 ค่าเฉลี่ย Silhouette เป็นตัวชี้วัดการจัดกลุ่มข้อมูลที่ประเมินความใกล้เคียงระหว่างข้อมูลในกลุ่มเดียวกันกับข้อมูลในกลุ่มอื่น ๆ ดังสมการที่ (2) งานวิจัยนี้เลือกใช้พารามิเตอร์ silhouette_score โดยมีค่าตั้งแต่ -1 ถึง 1 สำหรับค่าใกล้ 1 หมายถึงการจัดกลุ่มที่ดี โดยข้อมูลภายในกลุ่มมีความคล้ายคลึงกันมากและข้อมูลต่างกลุ่มมีความแตกต่างกันมาก สำหรับค่าใกล้ -1 หมายถึงการจัดกลุ่มที่ไม่ดี โดยข้อมูลภายในกลุ่มมีความคล้ายคลึงกันน้อยและข้อมูลต่างกลุ่มมีความคล้ายคลึงกันมาก (Wang *et al.*, 2017, scikit-learn developers, 2023) ดัง Figure 2 (b) พบว่าจำนวนกลุ่มที่เท่ากับ 3 มีค่าที่ใกล้เคียงค่า 1 มากที่สุด

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

โดยที่

$s(i)$ คือ Silhouette Score ของจุดข้อมูล i ซึ่งอยู่ในช่วงระหว่าง -1 ถึง 1

$a(i)$ คือ ค่าเฉลี่ยของระยะทาง (Distance) ระหว่างจุดข้อมูล i กับจุดข้อมูลอื่น ๆ ที่อยู่ในกลุ่มเดียวกัน (intra-cluster distance)

$b(i)$ คือ ค่าเฉลี่ยของระยะทาง ระหว่างจุดข้อมูล i กับจุดข้อมูลในกลุ่มที่ใกล้ที่สุด (nearest other cluster) ซึ่งจุดข้อมูล i ไม่ได้เป็นสมาชิก

หาก $s(i) \approx 1$ คือ การจัดกลุ่มมีความแม่นยำสูง จุด i อยู่ใกล้กับจุดในกลุ่มเดียวกันและห่างจากจุดในกลุ่มอื่น

$s(i) \approx 0$ คือ จุด i อยู่บริเวณขอบเขตของกลุ่ม ใกล้กับกลุ่มอื่นพอสมควร

$s(i) \approx -1$ คือ การจัดกลุ่มที่ขาดความแม่นยำ จุด i ใกล้กับจุดในกลุ่มอื่นมากกว่ากลุ่มของตัวเอง

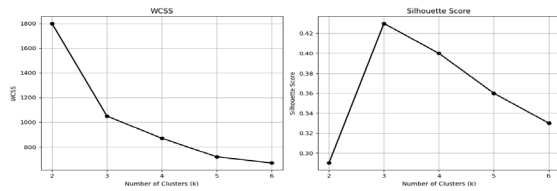


Figure 2 WCSS and Silhouette Score by Number of Clusters (k)

4. การสร้างแบบจำลองเพื่อจัดกลุ่ม (Clustering Model)

เป็นขั้นตอนที่จะนำข้อมูลที่ถูกเตรียมไว้ในขั้นตอนก่อนหน้า เข้าสู่กระบวนการสร้างโมเดลเพื่อจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูงต่อปัญหาสุขภาพจิตและปัญหาความรุนแรงในจากขั้นตอนการวิเคราะห์เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุดก่อนหน้านี้ จึงกำหนดให้ $k = 3$ เนื่องจากเป็นค่าที่ดีที่สุดที่สามารถแยกแยะกลุ่มความเสี่ยงได้อย่างมีประสิทธิภาพ

4.1 K-means Clustering มีหลักการทำงานแบบวนซ้ำเพื่อแบ่งข้อมูลออกเป็นกลุ่มย่อยตามลักษณะความใกล้เคียงกันของข้อมูล (Supakorn, 2021) มีการนำไปประยุกต์ใช้กับงานวิจัยในด้านต่างๆ เช่น นำเสนอการประยุกต์ใช้ K-Means ร่วมกับ SVM เพื่อพยากรณ์ภาวะเบาหวาน (Arora et al., 2022) การดำเนินงานของ K-means (Rizman Žalik & Žalik, 2023) มีขั้นตอนสำคัญดังนี้ (1) กำหนดจำนวนกลุ่มล่วงหน้า ดังนั้นจึงกำหนดพารามิเตอร์ $n_clusters = 3$ (2) สุ่มเลือกจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่ม ทำการสุ่มเลือกข้อมูลจำนวน 3 จุดจากชุดข้อมูลทั้งหมดเพื่อใช้เป็นจุดเริ่มต้น (Initial Centroids) ของแต่ละกลุ่ม และใช้พารามิเตอร์ $init='k-means++'$ วิธีการนี้โมเดลจะค่อย ๆ เลือกจุด Centroid ที่ละจุดโดยใช้หลักการว่า จุดใหม่ควรอยู่ห่างจาก Centroid ที่เลือกไว้ก่อนหน้านี้ให้มากที่สุด Centroid ที่ได้จะกระจายตัวได้ดีตั้งแต่ต้น ช่วยให้กลุ่มสุดท้ายมีคุณภาพดีกว่า และลดจำนวนรอบที่ต้องรัน K-means ลง จากนั้นกำหนดค่า $n_init=10$ เป็นการกำหนดจำนวนครั้งที่โมเดล K-means จะสุ่มจุดเริ่มต้น แล้วเลือกผลลัพธ์ที่ดีที่สุดคือที่มี WCSS ต่ำที่สุดและเป็นค่าที่ทดลองแล้วทำให้ผลลัพธ์ที่เสถียรในหลายกรณี (3) จัดกลุ่มข้อมูลตามความใกล้เคียงกับจุดศูนย์กลาง โดยการคำนวณระยะทางระหว่างข้อมูลแต่ละจุดกับจุดเริ่มต้นของแต่ละกลุ่ม โดยปกติมักใช้ระยะทางแบบ Euclidean Distance ซึ่งสามารถคำนวณได้จากสมการ (3) แล้วทำการจัดกลุ่มข้อมูลให้เป็นสมาชิกของกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^n (x_{ij} - \mu_{kj})^2} \quad (3)$$

โดยที่:

x_i คือ เวกเตอร์ของจุดข้อมูลลำดับที่ i

μ_k คือ เวกเตอร์ของจุดศูนย์กลางของกลุ่มที่ k

x_{ij} และ μ_{kj} คือ ค่าคุณลักษณะลำดับที่ j ของ

x_{ij} และ μ_{kj} ตามลำดับ

n คือ จำนวนคุณลักษณะ (Features) ของข้อมูล

(4) หลังจากจัดกลุ่มข้อมูลแล้ว จะคำนวณจุดศูนย์กลางใหม่ของแต่ละกลุ่มโดยใช้ค่าเฉลี่ยของข้อมูลทั้งหมดในกลุ่มนั้น โดยสมการของจุดศูนย์กลางใหม่ดังสมการที่ (4)

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \quad (4)$$

โดยที่:

x_i คือ เวกเตอร์ของจุดข้อมูลลำดับที่ i

μ_k คือ จุดศูนย์กลางใหม่ของกลุ่มที่ k

C_k คือ เซตของข้อมูลที่อยู่ในกลุ่ม k

$|C_k|$ คือ จำนวนข้อมูลในกลุ่ม k

(5) ทำซ้ำขั้นตอนที่ 3 และ 4 ไปเรื่อย ๆ

จนกว่าจุดศูนย์กลางของแต่ละกลุ่มจะไม่เปลี่ยนแปลง หรือมีการเปลี่ยนแปลงเพียงเล็กน้อย ซึ่งถือว่าอัลกอริทึมเข้าสู่ภาวะสมดุล (Stable State) ขั้นตอนนี้ใช้พารามิเตอร์ max_iter ที่กำหนดไว้ในคลาส KMeans เพื่อใช้ควบคุมจำนวนรอบสูงสุดของการทำงาน โดยกำหนดพารามิเตอร์ $max_iter=300$ เพื่อควบคุมจำนวนรอบสูงสุดของการปรับ centroid เป็นค่าพื้นฐานที่แนะนำในเอกสารทางเทคนิคของไลบรารีของ scikit-learn

4.2 HC เป็นการจัดกลุ่มแบบลำดับขั้นโดยการรวมกลุ่มที่ใกล้เคียงกันที่สุดไปเรื่อย จากนั้นกำหนดวิธีการเชื่อมโยงแบบ Ward ($linkage='ward'$) เป็นวิธีที่ลดความแปรปรวนภายในกลุ่ม ช่วยให้กลุ่มที่มีความแตกต่างกันชัดเจนมากขึ้น ในแต่ละขั้นตอนของการรวมกลุ่มจะเลือกการรวมกลุ่มที่ส่งผลให้ผลรวมของความแปรปรวนภายในกลุ่มเพิ่มขึ้นน้อยที่สุด (Hemad et al., 2022) โดยการคำนวณของวิธี Ward จะยึดหลักการของการวิเคราะห์ความแปรปรวน (Analysis of Variance; ANOVA) สามารถแสดงในรูปของฟังก์ชันการวัดระยะห่าง (Müllner, 2020) ได้ดังสมการที่ (5)

$$D_{(A,B)} = \frac{n_A n_B}{n_A + n_B} \cdot \|\bar{x}_A - \bar{x}_B\|^2 \quad (5)$$

โดยที่

$D_{(A,B)}$ คือ ค่าระยะห่างระหว่างกลุ่ม A และกลุ่ม B
 n_A และ n_B คือ จำนวนข้อมูลในกลุ่ม A และ B ตามลำดับ

$\bar{x}_A$ และ $\bar{x}_B$ คือ เวกเตอร์ค่าเฉลี่ยของข้อมูลในกลุ่ม A และ B ตามลำดับ

$\|\bar{x}_A - \bar{x}_B\|^2$ คือ ระยะห่างแบบ Euclidean Distance ยกกำลังสองระหว่างค่าเฉลี่ยของทั้งสองกลุ่ม

จากสมการจะเห็นได้ว่า วิธีการนี้จะเลือกจับคู่กลุ่มที่มีค่าเฉลี่ยใกล้เคียงกันและมีจำนวนสมาชิกที่เหมาะสมเพื่อให้การรวมกลุ่มนั้นส่งผลต่อความแปรปรวนภายในกลุ่มน้อยที่สุด วิธีนี้เหมาะสำหรับข้อมูลเชิงปริมาณในกรณีที่มีตัวแปรต้นมีลักษณะเป็นข้อมูลเชิงคุณภาพต้องทำการแปลงให้อยู่ในรูปเชิงตัวเลขก่อนเพื่อให้สามารถนำไปใช้ในการคำนวณระยะห่างและการวิเคราะห์กลุ่มได้อย่างถูกต้อง (Murtagh & Legendre, 2014)

4.3 GMMs เป็นเทคนิคที่สามารถรองรับข้อมูลที่มีการกระจายตัวในลักษณะที่ไม่จำเป็นต้องมีขอบเขตแบบเชิงเส้น (non-linear boundaries) และสามารถจำแนกกลุ่มข้อมูลที่มีความซับซ้อนสูงได้อย่างมีประสิทธิภาพ หลักการ GMMs เป็นการจำลองการกระจายของข้อมูลโดยสมมติว่าข้อมูลทั้งหมดเกิดจากการผสมกันของการแจกแจงแบบ Gaussian หลายตัว จำนวนกลุ่มเช่นเดียวกับทั้ง 2 วิธีข้างต้น จึงกำหนดให้ $n_components = 3$ ในคลาสของ GaussianMixture โดยแต่ละกลุ่ม k จะมีการแจกแจงปกติของตัวเอง (Huang *et al.*, 2023) ดังนั้นฟังก์ชันความหนาแน่นร่วมของ GMMs สามารถเขียนได้ดังสมการที่ (6)

$$p(x) = \sum_{k=1}^K \pi_k \cdot N(x | \mu_k, \Sigma_k) \quad (6)$$

โดยที่

x คือ เวกเตอร์ของข้อมูล (เช่น อายุ การดื่มแอลกอฮอล์ การใช้สารเสพติด ความก้าวร้าว ฯลฯ)

K คือ จำนวนกลุ่ม (clusters) ที่ต้องการจำแนก เช่น กลุ่มความเสี่ยงต่ำ ปานกลาง และสูง

π_k คือ น้ำหนักของกลุ่มที่ k (mixing coefficient) ($\sum_{k=1}^K \pi_k = 1, 0 < \pi_k < 1$)

μ_k คือ เวกเตอร์ค่าเฉลี่ยของกลุ่มที่ k

Σ_k คือ เมทริกซ์ความแปรปรวนร่วม (covariance matrix) ของกลุ่มที่ k

โดยการแจกแจงแบบ Gaussian ในแต่ละกลุ่มสามารถนิยามได้ดังสมการ (7)

$$N(x | \mu_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \quad (7)$$

โดยที่

x คือ เวกเตอร์ข้อมูลที่ต้องการคำนวณค่าความน่าจะเป็น ซึ่งในบริบทของการวิจัยอาจประกอบด้วยตัวแปร เช่น อายุ การดื่มแอลกอฮอล์ การใช้สารเสพติด ความก้าวร้าว ฯลฯ

μ_k คือ เวกเตอร์ค่าเฉลี่ยของกลุ่มที่ k

Σ_k คือ เมทริกซ์ความแปรปรวนร่วม (covariance matrix) ของกลุ่มที่ k

$|\Sigma_k|$ คือ Determinant ของเมทริกซ์ Σ_k ซึ่งใช้เป็นตัวปรับขนาดความน่าจะเป็นในหลายมิติ

Σ_k^{-1} คือ เมทริกซ์ผกผันของ Σ_k ใช้ในการคำนวณระยะห่างแบบ Mahalanobis (AntixK, 2024)

d คือ จำนวนมิติของข้อมูลหรือตัวแปรต้น

$\exp$ คือ ฟังก์ชันเอ็กซ์โพเนนเชียล

ประเภทของ covariance_type ใน GMMs งานวิจัยนี้เลือกใช้ covariance_type='full' ซึ่งช่วยให้การแจกแจงสามารถสะท้อนความสัมพันธ์ระหว่างตัวแปรต้นได้อย่างสมบูรณ์ ในบริบทของการจำแนกกลุ่มความเสี่ยง เช่น ความเสี่ยงต่ำ ปานกลาง และสูง จากข้อมูลสุขภาพจิตและพฤติกรรม ช่วยให้โมเดลสามารถแยกกลุ่มที่มีความซับซ้อนได้ดียิ่งขึ้น เนื่องจากสามารถอธิบายโครงสร้างความสัมพันธ์ของข้อมูลได้ครบถ้วน (Budiarto *et al.*, 2021)

5. ประเมินประสิทธิภาพของโมเดล (Model Evaluation)

สำหรับการวิเคราะห์ข้อมูลและการตีความผลลัพธ์ของโมเดลการจัดกลุ่มแต่ละแบบ มีวิธีการวิเคราะห์ข้อมูลรูปแบบเดียวกันคือ คำนวณค่าเฉลี่ยของปัจจัยต่าง ๆ ในแต่ละกลุ่ม เช่น อายุ การดื่มแอลกอฮอล์ การใช้สารเสพติด ฯลฯ และวิเคราะห์ความคล้ายคลึงกันและความแตกต่างระหว่างกลุ่ม ตรวจสอบว่าแต่ละกลุ่มมีลักษณะเฉพาะอย่างไร และมีปัจจัยเสี่ยงใดที่เด่นชัดในแต่ละกลุ่ม ผู้วิจัยมีวิธีการประเมินดังนี้

5.1 Silhouette Score เป็นวิธีการเดียวกันกับการวิเคราะห์เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุดที่กล่าวไว้ก่อนหน้านี้คือการหาค่าเฉลี่ย Silhouette

5.2 Calinski-Harabasz Index เป็นตัวชี้วัดการจัดกลุ่มข้อมูลที่ใช้ประเมินความแข็งแกร่งของการจัดกลุ่มโดยคำนวณจากอัตราส่วนระหว่างความแปรปรวนของข้อมูลภายในกลุ่มกับความแปรปรวนของข้อมูลต่างกลุ่ม ดังสมการ

ที่ (8) โดยใช้พารามิเตอร์ `calinski_harabasz_score` ค่าสูงหมายถึงการจัดกลุ่มที่ดี โดยความแปรปรวนของข้อมูลภายในกลุ่มมีค่าน้อยและความแปรปรวนของข้อมูลต่างกลุ่มมีค่ามาก (Lima & Cruz, 2020)

$$CH = \frac{SSB / (k - 1)}{SSW / (N - k)} \quad (8)$$

โดยที่

SSB คือ ผลรวมของความแปรปรวนระหว่างกลุ่ม
 SSW คือ ผลรวมของความแปรปรวนภายในกลุ่ม
 k คือ จำนวนกลุ่ม
 N คือ จำนวนข้อมูลทั้งหมด

5.3 Davies-Bouldin Index เป็นตัวชี้วัดการจัดกลุ่มข้อมูลที่ใช้ประเมินความคล้ายคลึงระหว่างกลุ่มข้อมูล โดยคำนวณจากอัตราส่วนของความคล้ายคลึงกันภายในกลุ่มกับความคล้ายคลึงกันระหว่างกลุ่ม ค่าต่ำหมายถึงการจัดกลุ่มที่ดี โดยข้อมูลภายในกลุ่มมีความคล้ายคลึงกันมากและข้อมูลต่างกลุ่มมีความแตกต่างกันมาก (Hassan *et al.*, 2021) โดยใช้พารามิเตอร์ `davies_bouldin_score` ดังสมการที่ (9)

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (9)$$

โดยที่:

k คือ จำนวนกลุ่มทั้งหมด
 S_i คือ ค่าเฉลี่ยของระยะห่างระหว่างแต่ละจุดข้อมูลในกลุ่มที่ i กับเซนทรอยด์ของกลุ่ม
 M_{ij} คือ ระยะห่างระหว่างเซนทรอยด์ของกลุ่ม i และกลุ่ม j

6. การตีความผลและการจัดกลุ่มความเสี่ยง (Interpretation & Risk Grouping)

ขั้นตอนสุดท้ายของกระบวนการวิเคราะห์การจัดกลุ่มข้อมูล ผู้วิจัยได้ดำเนินการตีความผลลัพธ์ที่ได้จากแบบจำลองเพื่อนำไปสู่การจัดกลุ่มบุคคลตามระดับความเสี่ยงต่อปัญหาความรุนแรงในครอบครัวและสุขภาพจิต

เริ่มจากการวิเคราะห์คุณลักษณะเชิงประชากรและพฤติกรรมของสมาชิกในแต่ละกลุ่มที่ได้จากการจัดกลุ่มโดยใช้ข้อมูลเชิงสถิติ เช่น ค่าเฉลี่ยของตัวแปรต่าง ๆ (เช่น อายุ การดื่มแอลกอฮอล์ การใช้สารเสพติด พฤติกรรมก้าวร้าว ปัญหาทางอารมณ์ และสุขภาพจิต) เพื่อสรุปลักษณะเฉพาะของแต่ละกลุ่ม ซึ่งช่วยให้สามารถจำแนกได้ว่าแต่ละกลุ่มมีระดับความเสี่ยงอยู่ในระดับต่ำ ปานกลาง หรือสูง

ผู้วิจัยได้แบ่งกลุ่มออกเป็น 3 กลุ่ม ตามขั้นตอนก่อนหน้า ได้แก่ (1) กลุ่มเสี่ยงต่ำ (Cluster0: Low Risk) ซึ่งมีปัจจัยเสี่ยงในระดับต่ำหรือไม่มีเลย (2) กลุ่มเสี่ยงปานกลาง (Cluster1: Medium Risk) ซึ่งมีบางปัจจัยเสี่ยงร่วมกัน เช่น ความเครียดทางเศรษฐกิจหรือความก้าวร้าวระดับปานกลาง และ (3) กลุ่มเสี่ยงสูง (Cluster2: High Risk) ซึ่งมีหลายปัจจัยเสี่ยงร่วมกันอย่างชัดเจน เช่น การใช้สารเสพติดควบคู่กับปัญหาสุขภาพจิตและพฤติกรรมรุนแรง

นอกจากนี้ ผู้วิจัยได้วิเคราะห์กลุ่มบุคคลในกลุ่มเสี่ยงต่ำที่มีพฤติกรรมผิดปกติ ที่บ่งชี้ถึงความเสี่ยงที่ซ่อนอยู่ กลุ่มนี้เรียกว่า กลุ่มเสี่ยงเงียบ (Silent Risk Group)

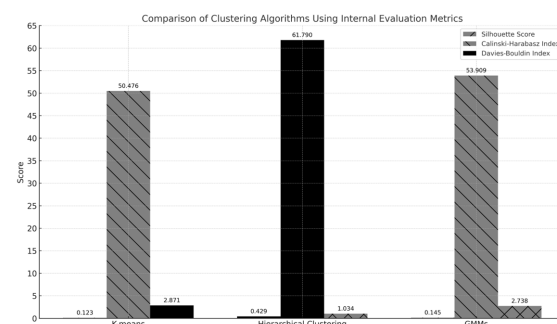


Figure 3 Comparison of Clustering Model

ผลการทดลองและอภิปรายผล

จาก Table 2 แสดงการจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูงต่อปัญหาสุขภาพจิตและปัญหาความรุนแรงในครอบครัว โดยใช้เทคนิคการจัดกลุ่มข้อมูล 3 เทคนิค พบว่าแต่ละเทคนิคสามารถแบ่งกลุ่มตามระดับความเสี่ยงได้ชัดเจน โดยมีลักษณะสำคัญดังนี้ (1) K-means แบ่งกลุ่มได้ใกล้เคียงกันในกลุ่ม Low และ Medium Risk โดยมีอายุเฉลี่ยประมาณ 36 ปี แต่ในกลุ่ม High Risk อายุสูงขึ้นเป็น 42 ปี โดยกลุ่ม High Risk มีจำนวนน้อยที่สุดและมีอัตราการหย่าร้างสูงถึง 1.0 (2) HC ให้ผลลัพธ์คล้ายกับ K-means แต่กลุ่ม Low Risk มีอายุเฉลี่ยต่ำที่สุด และกลุ่ม High Risk มีอัตราการหย่าร้างสูงเช่นกัน ซึ่งสะท้อนถึงความเสี่ยงในกลุ่มนี้อย่างชัดเจน และ (3) GMMs สามารถจำแนกปัจจัยความเสี่ยงได้ละเอียดกว่า โดยพบว่าในทุกกลุ่มความเสี่ยงมีปัจจัยด้านอารมณ์ และสุขภาพจิตเข้ามาเกี่ยวข้องชัดเจน เช่น กลุ่ม High Risk มีระดับความเสี่ยงด้านสุขภาพจิตสูงสุด และมีแนวโน้มปัญหาสุขภาพ และการพนันสูง โดยรวม GMMs แสดงความแตกต่างของกลุ่มได้เด่นชัดกว่าทั้ง K-means และ HC เพราะสามารถตรวจจับลักษณะเสี่ยงซ้อนกันหลายปัจจัยได้ดีกว่า เช่น ความก้าวร้าว ความหึงหวง และพฤติกรรมเสี่ยงอื่น ๆ พร้อมกันในแต่ละกลุ่ม แม้ว่าผลการจัดกลุ่มด้วยทั้ง 3 เทคนิคจะสามารถแบ่งกลุ่มผู้มี

ความเสี่ยงต่อความรุนแรงในครอบครัวได้อย่างมีประสิทธิภาพ แต่เมื่อวิเคราะห์เชิงลึกลงไปด้วยเทคนิค Local Outlier Factor (LOF) พบว่ากลุ่มเสี่ยงต่ำมี บุคคลบางรายที่มีลักษณะผิดปกติ (outliers) ซึ่งสะท้อนความเสี่ยงที่ซ่อนอยู่ บุคคลทั้งหมดในกลุ่ม outlier ที่อยู่ในกลุ่มเสี่ยงต่ำ มีประวัติการใช้สารเสพติด แม้จะไม่มีปัจจัยเสี่ยงอื่นที่มักปรากฏร่วม เช่น ปัญหาสุขภาพจิต ความเครียดทางเศรษฐกิจ หรือประวัติการหย่าร้าง นอกจากนี้ ยังพบพฤติกรรมบางรายที่เกี่ยวข้องกับการใช้อำนาจในครอบครัว และภาวะอารมณ์โกรธรุนแรง แม้ไม่มีปัญหาทางจิตเวชอย่างชัดเจน ผลดังกล่าวชี้ให้เห็นว่า การประเมินความเสี่ยงโดยอิงเฉพาะค่ากลางหรือปัจจัยพื้นฐานที่ปรากฏเด่นชัด อาจไม่สามารถตรวจจับกลุ่มบุคคลที่มีความเสี่ยงเชิงพฤติกรรมที่ซ่อนอยู่ได้อย่างเพียงพอ โดยเฉพาะผู้ที่ไม่มีปัจจัยเสี่ยงตามแบบแผน เช่น สุขภาพจิตหรือสถานะเศรษฐกิจ แต่กลับมีพฤติกรรมเสี่ยงที่เป็นอิสระ เช่น การเสพยา หรือการใช้อารมณ์รุนแรง

ข้อค้นพบนี้จึงชี้ให้เห็นถึงความสำคัญของการพัฒนาระบบคัดกรองที่สามารถตรวจจับ “กลุ่มเสี่ยงเจียบ” ได้อย่างมีประสิทธิภาพ ควรใช้แนวทางการผสมผสานระหว่างเทคนิคการจัดกลุ่ม และเทคนิคการตรวจจับความผิดปกติ (Anomaly detection) เพื่อสนับสนุนการวางแผนแทรกแซงเชิงป้องกันที่ครอบคลุมกลุ่มเสี่ยงที่ตรวจจับยากได้มากยิ่งขึ้น โดยเฉพาะสำหรับข้อมูลที่มีลักษณะซับซ้อนและมีพฤติกรรมแฝงที่หลากหลาย

สรุปผลการทดลองและข้อเสนอแนะ

จาก Figure 3 แสดงข้อมูลการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค พบว่า HC ให้ผลลัพธ์ที่ดีที่สุด บ่งชี้ถึงการจัดกลุ่มที่สามารถแยกความแตกต่างระหว่างกลุ่มได้ชัดเจนกว่าวิธีอื่น สอดคล้องกับงานวิจัยของ Zhang *et al.* (2025) และ Amer *et al.* (2024) สาเหตุที่ HC ให้ผลลัพธ์ที่ดีกว่า ได้แก่ (1) สามารถสะท้อนโครงสร้างข้อมูลได้ดีกว่า โดยเฉพาะในข้อมูล

ที่มีแนวโน้มเป็นลำดับชั้น (2) จัดการกับข้อมูลที่ไม่สมดุลได้ดีกว่า K-means และ GMMs (3) ไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า ทำให้เหมาะกับข้อมูลที่มีการกระจายตัวที่ไม่แน่นอน อย่างไรก็ตาม ข้อเสียของ HC คือใช้ทรัพยากรคำนวณสูงและอ่อนไหวต่อค่าผิดปกติ ดังนั้นการใช้ร่วมกับเทคนิคอื่น เช่น การลดขนาดข้อมูลหรือการกำจัดค่าผิดปกติล่วงหน้าช่วยเพิ่มประสิทธิภาพของการจัดกลุ่ม

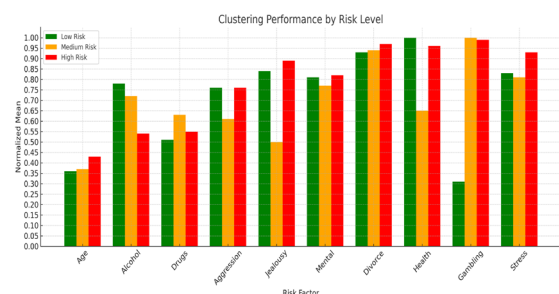
เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของทั้ง 2 โมเดลที่เหลือ พบว่า GMMs ให้ค่า Silhouette Score สูงกว่า แสดงถึงความสามารถในการสร้างกลุ่มที่มีความเหมือนกันของข้อมูลภายในกลุ่มและมีความแตกต่างระหว่างกลุ่มได้ดีกว่า นอกจากนี้ GMMs ยังมีค่า Calinski-Harabasz Index สูงกว่า บ่งชี้ว่ากลุ่มที่ได้จาก GMMs มีความแตกต่างระหว่างกลุ่มที่ชัดเจนและภายในกลุ่มมีความสอดคล้องกันมากกว่า K-mean สำหรับค่า Davies-Bouldin Index ค่าที่ต่ำกว่าจะสะท้อนถึงการจัดกลุ่มที่มีคุณภาพ พบว่า GMMs มีค่าต่ำกว่า K-means แสดงว่า GMMs สามารถสร้างกลุ่มที่มีข้อมูลในกลุ่มที่เหมือนกันดี และข้อมูลต่างกลุ่มแยกออกจากกันชัดเจน กล่าวคือ GMMs สามารถจำแนกกลุ่มบุคคลเหล่านี้ได้อย่างมีประสิทธิภาพ โดยทำให้สมาชิกที่มีพฤติกรรมเสี่ยงสูงอยู่รวมกัน ได้แก่ พฤติกรรมการดื่มแอลกอฮอล์ในระดับสูง, ความก้าวร้าวสูง และปัญหาสุขภาพกายในระดับสูงสุด ซึ่งแสดงถึงความเสี่ยงเชิงพฤติกรรมที่เกิดขึ้นร่วมกันอย่างชัดเจน แม้ข้อมูลบางส่วนมีความแปรปรวนเล็กน้อยระหว่างบุคคล ตัวอย่างเช่น บุคคลที่มีพฤติกรรมเสี่ยงเพียงบางด้านแต่มีคะแนนความก้าวร้าวหรือสุขภาพกายสูงก็ยังสามารถถูกรวมอยู่ในกลุ่มเสี่ยงสูงได้อย่างเหมาะสม ด้วยความสามารถของ GMMs ในการใช้แนวทางการจัดกลุ่มแบบความน่าจะเป็น ซึ่งข้อมูลแต่ละตัวมีความน่าจะเป็นเกิดขึ้นได้ในหลายกลุ่ม (Soft assignment) ต่างจาก K-mean ที่ใช้แนวทางการจัดกลุ่มแบบจัดกลุ่มแบบเด็ดขาด ข้อมูลแต่ละตัวอยู่ในกลุ่มเดียวแน่นอน (Hard Assignment)

Table 2 Comparative Characteristics of Risk Groups Identified by Clustering Techniques

Model	Risk	Age	Alcohol	Drugs	Aggression	Jealousy	Mental	Divorce	Health	Gambling	Stress	N
K-means	Low (Cluster0)	36.13	0.33	0.5	0.88	-	-	0.96	0.94	-	-	214
	Medium (Cluster1)	36.55	0.88	0.7	0.61	-	-	0.97	0.98	-	-	370
	High (Cluster2)	42.54	0.77	0.38	0.77	-	-	0.92	1.0	-	-	13
HC	Low (Cluster0)	30.53	0.63	0.53	0.47	-	-	0.95	0.0	-	-	19
	Medium (Cluster1)	36.59	0.68	0.63	0.72	-	-	0.97	1.0	-	-	565
	High (Cluster2)	42.54	0.77	0.38	0.77	-	-	0.92	1.0	-	-	13
GMMs	Low (Cluster0)	29.3	0.65	0.65	0.74	1.0	0.75	1.0	0.95	1.0	0.98	285
	Medium (Cluster1)	33.29	0.8	0.75	0.74	1.0	0.92	0.94	1.0	0.95	0.66	189
	High (Cluster2)	37.52	0.61	0.53	0.68	0.0	0.93	1.0	0.96	0.98	0.92	123

ผลการวิเคราะห์นี้สอดคล้องกับแนวคิดของ Liu *et al.* (2021) ที่เน้นถึงความสำคัญของการใช้ตัวชี้วัดภายในในการประเมินคุณภาพของการจัดกลุ่ม โดยเฉพาะ Silhouette Score, Calinski-Harabasz Index, และ Davies-Bouldin Index ซึ่งสามารถสะท้อนความสอดคล้องภายในกลุ่มและความแตกต่างระหว่างกลุ่มได้อย่างชัดเจน การจัดกลุ่มสะท้อนให้เห็นถึงลักษณะเฉพาะของแต่ละกลุ่มความเสี่ยง เช่น กลุ่มเสี่ยงสูงมีความเกี่ยวข้องกับปัจจัยเชิงประชากรศาสตร์บางประการ ได้แก่ เพศชาย อายุช่วงกลางคน การมีภาวะสุขภาพจิต หรือ ประวัติการใช้สารเสพติด กลุ่มเสี่ยงสูงมีลักษณะดังนี้เป็นส่วนใหญ่เป็น เพศชายและมีอายุอยู่ในช่วง 35-50 ปี มีประวัติการใช้สารเสพติดและแอลกอฮอล์สูงกว่ากลุ่มอื่น ๆ แสดงพฤติกรรมก้าวร้าวและมีระดับความเครียดทางเศรษฐกิจสูง สำหรับปัจจัยหลักที่ทำให้เกิดความรุนแรงในครอบครัวคือ มีปัญหาสุขภาพจิต เช่น ภาวะซึมเศร้าและความโกรธที่รุนแรง ปัจจัยที่มีความสัมพันธ์กับพฤติกรรมรุนแรงในครอบครัว ส่วนใหญ่คือการใช้สารเสพติด เพิ่มโอกาสเกิดความรุนแรงสูงถึง 65% ความเครียดทางเศรษฐกิจเป็นบุคคลที่มีปัญหาทางการเงินมีแนวโน้มเกิดความรุนแรงสูงขึ้น 30% และสุดท้ายปัญหาสุขภาพจิต ผู้ที่มีภาวะซึมเศร้าหรือความโกรธสูงมีแนวโน้มก่อความรุนแรงมากกว่ากลุ่มทั่วไป 45% ซึ่งสอดคล้องกับงานวิจัยของ Jewkes *et al.* (2015) ที่ระบุว่าความเสี่ยงต่อพฤติกรรมรุนแรงในครอบครัวมักสัมพันธ์กับปัจจัยหลายมิติ ทั้งด้าน

เศรษฐกิจและจิตสังคม ดัง Figure 4 และเมื่อเปรียบเทียบกับงานวิจัยของ Lima and Cruz (2020) ที่ใช้เพียงเทคนิค HC ในการจำแนกกลุ่มความเสี่ยงของผู้ป่วย พบว่า งานวิจัยฉบับนี้ซึ่งใช้การเปรียบเทียบหลากหลายเทคนิคการจัดกลุ่มทั้ง K-means, HC และ GMMs ช่วยเพิ่มความน่าเชื่อถือของผลลัพธ์ อีกทั้งยังเปิดโอกาสให้นักวิจัยสามารถเลือกใช้โมเดลที่เหมาะสมกับบริบทของข้อมูลได้มากยิ่งขึ้น แสดงให้เห็นถึงข้อได้เปรียบของการใช้การวิเคราะห์เปรียบเทียบแบบหลายโมเดลร่วมกัน

**Figure 4** Clustering Performance by Risk Level

อย่างไรก็ตาม งานวิจัยนี้ยังมีข้อจำกัดบางประการที่ควรพิจารณา ได้แก่ ความไม่สมดุลของข้อมูล (Imbalanced Data) จำนวนข้อมูลในแต่ละกลุ่มอาจมีความแตกต่างกันมาก เช่นปัญหาสุขภาพจิต (16.02%) และการติดการพนัน (2.24%) เป็นต้น เนื่องจากโมเดลมีแนวโน้มเรียนรู้จากกลุ่มข้อมูลที่มี

จำนวนมากและอาจมองข้ามกลุ่มข้อมูลที่มีจำนวนน้อย ผลที่เกิดขึ้นอาจทำให้โมเดลไม่สามารถแยกแยะหรือตรวจจับกลุ่มเสี่ยงที่แท้จริงได้อย่างแม่นยำ โดยเฉพาะในกลุ่มที่มีปัญหาเฉพาะ เช่น กลุ่มที่มีปัญหาสุขภาพจิตหรือการติดการพนัน แม้ว่าจะเป็นปัจจัยที่สัมพันธ์กับพฤติกรรมความรุนแรงในครอบครัวก็ตาม สำหรับแนวทางในการแก้ไขเพื่อเพิ่มประสิทธิภาพความน่าเชื่อถือของผลลัพธ์ในการจำแนก อาจใช้เทคนิค SMOTE (Synthetic Minority Over-sampling Technique) เพื่อเพิ่มตัวอย่างของกลุ่มที่มีจำนวนน้อยโดยการสร้างตัวอย่างเทียม (Synthetic Sample) เพิ่มเข้าไปในกลุ่มย่อยเหล่านั้น ช่วยให้โมเดลมีโอกาสรียนรู้จากกลุ่มข้อมูลจำนวนน้อยได้มากขึ้นโดยไม่สูญเสียข้อมูลจากกลุ่มอื่น หรือปรับค่าการให้น้ำหนักกับจุดข้อมูลที่อยู่ในกลุ่มจำนวนน้อย เช่น บุคคลที่มีปัญหาสุขภาพจิต อาจกำหนดน้ำหนักเพิ่มขึ้นเพื่อส่งผลกระทบต่อค่าคำนวณ Centroid หรือการกระจายของกลุ่ม ใช้ค่าถ่วงน้ำหนักเพื่อให้ตัวอย่างบางตัวมีอิทธิพลมากกว่าตัวอื่นในการกำหนดตำแหน่ง centroid เป็นต้น สำหรับ K-means ใช้วิธี Weighted Centroid เพื่อคำนวณค่าของ Centroid ช่วยให้ Centroid ที่สะท้อนกลุ่มเสี่ยงข้างน้อยมากขึ้น และ GMMs ใช้ Weighted EM เพื่อช่วยให้โมเดลระบุตำแหน่งของกลุ่มได้ใกล้เคียงความจริงมากขึ้น และค่าการกระจายของกลุ่มที่สามารถสะท้อนลักษณะของกลุ่มนั้นได้ชัดเจน

การศึกษาวิจัยในอนาคต ผู้วิจัยมีแนวโน้มที่จะขยายผลการศึกษาโดยการเปรียบเทียบโมเดลทั้งสามในด้านอื่นเช่น ความเร็วในการประมวลผล และการใช้ทรัพยากรระบบ เพื่อสะท้อนศักยภาพของแต่ละโมเดลได้อย่างครอบคลุมยิ่งขึ้นและเป็นประโยชน์ต่อการเลือกใช้งานในสถานการณ์ที่มีข้อจำกัดด้านเวลาและทรัพยากร

กิตติกรรมประกาศ

การศึกษานี้สำเร็จลุล่วงไปได้ด้วยความร่วมมือและการสนับสนุนจากหลายฝ่าย ผู้วิจัยขอแสดงความขอบคุณเป็นอย่างยิ่งต่อสำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน) (สพร.) ที่ให้ข้อมูลเกี่ยวกับผู้กระทำความรุนแรงในครอบครัว เป็นแหล่งข้อมูลสำคัญในการวิเคราะห์และจัดกลุ่มเพื่อจำแนกกลุ่มบุคคลที่มีความเสี่ยงสูงต่อปัญหาความรุนแรงในครอบครัว ผู้วิจัยขอขอบคุณวารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคามที่ให้เกียรติรับพิจารณาและตีพิมพ์ผลงานวิจัยฉบับนี้ ซึ่งเป็นโอกาสสำคัญในการเผยแพร่ความรู้และแนวทางการศึกษาที่อาจเป็นประโยชน์ต่อการแก้ไขปัญหาความรุนแรงในครอบครัวในอนาคต

เอกสารอ้างอิง

- ณัฐวุฒิ แกมเงิน, ปกรณ์ ล่องทอง, พงศรัณย์ ทองหนูญ, กนกวรรณ ละอองศรี, อนามัย เทศกะทีก, พีรพล ศิริพงศ์ วุฒิกร, และณฐนนท์ เทพตะขบ, และวิริยะ มหิกุล. (2567). การเรียนรู้ของเครื่องเพื่อทำนายระดับความรุนแรงของความผิดปกติของความยืดหยุ่นปอดของพนักงานโรงงาน. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม*, 43(2), 84–95.
- พิทยา สังข์แก้ว, รุ่งทิวา เสาวนีย์, และอารยา หาอุปละ. (2565). การสำรวจความรุนแรงในครอบครัวในสถานการณ์การแพร่ระบาดของโรคโควิด 19 กรุงเทพฯ. *Rama Medical Journal*, 45(3), 33–41. <https://doi.org/10.33165/rmj.2022.45.3.257258>
- สำนักงานพัฒนารัฐบาลดิจิทัล. (2567). ข้อมูลผู้กระทำความรุนแรงในครอบครัว [ชุดข้อมูล]. ศูนย์กลางข้อมูลเปิดภาครัฐ. <https://gdcatalog.go.th/dataset/gdpublish-dwf-pb-dmv01-050507-04> (เข้าถึงเมื่อ 3 มีนาคม 2568)
- อรรถพล ดำเนินผล. (2564). บทบาทศูนย์พัฒนาครอบครัวในชุมชนกับปัญหาความรุนแรงในครอบครัว: กรณีศึกษา ศูนย์พัฒนาครอบครัวในชุมชน ตำบลท่าทราย จังหวัดนนทบุรี (สารนิพนธ์ปริญญารัฐประศาสนศาสตรมหาบัณฑิต). จุฬาลงกรณ์มหาวิทยาลัย, Chula Digital Collections. <https://digital.car.chula.ac.th/chulaetd/8052>
- Amer, A. A., Al-Razgan, M., Abdalla, H. I., Al-Asaly, M., Alfakih, T., & Al-Hammadi, M. (2024). Neighboring-aware hierarchical clustering: A new algorithm and extensive evaluation. *International Journal on Semantic Web and Information Systems*, 20(1). <https://doi.org/10.4018/IJSWIS.346377>
- Analytics Vidhya. (2021, January 12). *In-depth intuition of K-means clustering algorithm in machine learning*. <https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/>
- AntixK. (2024, January 12). *Extending Mahalanobis distance to Gaussian mixtures*. https://antixk.netlify.app/blog/gmm_mahalanobis/
- Arora, N., Singh, A., Al-Dabagh, M. Z. N., & Maitra, S. K. (2022). A novel architecture for diabetes patients' prediction using K-means clustering and SVM. *Mathematical Problems in Engineering*, 2022, 4815521. <https://doi.org/10.1155/2022/4815521>

- Budiarto, A., Mahesworo, B., Hidayat, A. A., Nurlaila, I., & Pardamean, B. (2021). Gaussian mixture model (GMMs) implementation for population stratification estimation from genomics data. *Procedia Computer Science*, 179, 202–210. <https://doi.org/10.1016/j.procs.2020.12.026>
- BuiltIn. (2024). *The elbow method for clustering explained*. <https://builtin.com/data-science/elbow-method>
- GeeksforGeeks. (2020, May 8). *Elbow method for optimal value of k in KMeans*. <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>
- Georgiou, K. (2024). Thematic analysis: A practical guide. *European Journal of Psychotherapy & Counselling*. <https://doi.org/10.1080/13642537.2024.2391666>
- Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, 76516–76531. <https://doi.org/10.1109/ACCESS.2020.2989857>
- Hassan, B. A., Rashid, T. A., & Hamarashid, H. K. (2021). A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star. *Computers in Biology and Medicine*, 138, Article 104866. <https://doi.org/10.1016/j.combiomed.2021.104866>
- Hemad, B. A., Ibrahim, N. M. A., Fayad, S. A., & Talaat, H. E. A. (2022). Hierarchical clustering-based framework for interconnected power system contingency analysis. *Energies*, 15(15), Article 5631. <https://doi.org/10.3390/en15155631>
- Huang, H., Liao, Z., Wei, X., & Zhou, Y. (2023). Combined Gaussian mixture model and pathfinder algorithm for data clustering. *Entropy*, 25(6), Article 946. <https://doi.org/10.3390/e25060946>
- Jewkes, R., Flood, M., & Lang, J. (2015). From work with men and boys to changes of social norms and reduction of inequities in gender relations: A conceptual shift in prevention of violence against women and girls. *The Lancet*, 385(9977), 1580–1589. [https://doi.org/10.1016/S0140-6736\(14\)61683-4](https://doi.org/10.1016/S0140-6736(14)61683-4)
- Lima, S. P., & Cruz, M. D. (2020). A genetic algorithm using Calinski–Harabasz index for automatic clustering problem. *Revista Brasileira de Computação Aplicada*, 12(3), 97–106. <https://doi.org/10.5335/rbca.v12i3.11117>
- Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2021). Understanding of internal clustering validation measures. *Information Sciences*, 575, 346–364. <https://doi.org/10.1016/j.ins.2021.07.021>
- Müllner, D. (2020). *Modern hierarchical, agglomerative clustering algorithms*. arXiv. <https://arxiv.org/pdf/2005.03197>
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, 31(2), 274–295. <https://doi.org/10.1007/s00357-014-9161-z>
- Rizman Žalik, K., & Žalik, M. (2023). Comparison of K-means, K-means++, X-means and single value decomposition for image compression. In *27th International Conference on Circuits, Systems, Communications and Computers (CSCC)* (pp. 295–301). IEEE. <https://doi.org/10.1109/CSCC58962.2023.00055>
- Scikit-learn developers. (2023). *sklearn.metrics.silhouette_score*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- Supakorn, S. (2021, December 14). *K-means clustering: การแบ่งกลุ่มย่อยเพื่อหาศูนย์กลาง!* Medium. https://medium.com/@si.supakorn_st/k-means-clustering-67a018f8cfb6
- Wang, F., Franco-Penya, H.-H., Kelleher, J. D., Pugh, J., & Ross, R. (2017). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In P. Perner (Ed.), *Machine learning and data mining in pattern recognition. MLDM 2017. Lecture notes in computer science* (Vol. 10358, pp. 291–305). Springer. https://doi.org/10.1007/978-3-319-62416-7_21
- World Health Organization. (2024, March 25). *Violence against women*. <https://www.who.int/news-room/fact-sheets/detail/violence-against-women>
- Zhang, T., Ju, L., Singh, P., & Toor, S. (2025). *InfoHier: Hierarchical information extraction via encoding and embedding*. arXiv. <https://arxiv.org/abs/2501.08717>