วิธีการแบบผสมผสานโดยใช้อัลกอริทึมการจัดกลุ่มและกฎความสัมพันธ์สำหรับการแนะนำสินค้าในธุรกิจค้าปลีก

A HYBRID APPROACH USING CLUSTERING ALGORITHMS AND ASSOCIATION RULES FOR PRODUCT RECOMMENDATION IN RETAIL BUSINESSES

ประทีป ปัญญนนทกานต์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

วิธีการแบบผสมผสานโดยใช้อัลกอริทึมการจัดกลุ่มและกฎความสัมพันธ์สำหรับการแนะนำสินค้าในธุรกิจค้าปลีก

ประทีป ปัญญนนทกานต์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

A HYBRID APPROACH USING CLUSTERING ALGORITHMS AND ASSOCIATION RULES FOR PRODUCT RECOMMENDATION IN RETAIL BUSINESSES

PRATEEP PANYANONTAKARN

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

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

สารนิพนธ์

เรื่อง

วิธีการแบบผสมผสานโดยใช้อัลกอริทึมการจัดกลุ่มและกฎความสัมพันธ์

สำหรับการแนะนำสินค้าในธุรกิจค้าปลีก

ของ

ประทีป ปัญญนนทกานต์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	วิธีการแบบผสมผสานโดยใช้อัลกอริทึมการจัดกลุ่มและกฎความสัมพันธ์สำหรับการแนะนำสินค้าในธุรกิจค้าปลีก
ผู้วิจัย	ประทีป ปัญญนนทกานต์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. นุวีย์ วิวัฒนวัฒนา

ในยุคที่อุตสาหกรรมร้านค้าปลีกต้องเผชิญกับการแข่งขันที่รุนแรง ธุรกิจจึงจำเป็นต้องปรับตัวทั้งด้านช่องทางการจัดจำหน่าย และการทำความเข้าใจพฤติกรรมผู้บริโภคที่มีความหลากหลายมากยิ่งขึ้น การวิเคราะห์และจำแนกกลุ่มลูกค้าจึงเป็นปัจจัยสำคัญที่ช่วยสนับสนุนการวางกลยุทธ์ทางการตลาด และการเลือกสรรสินค้าให้เหมาะสมกับกลุ่มเป้าหมายได้อย่างมีประสิทธิภาพ งานวิจัยนี้นำข้อมูลธุรกรรมของร้านค้าปลีกมาวิเคราะห์และจัดกลุ่มลูกค้าด้วยแบบจำลองจากอัลกอริทึมการเรียนรู้ของเครื่องแบบไม่มีผู้สอน ได้แก่ K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM) และ Agglomerative Clustering โดยใช้ข้อมูลที่ได้จากการวิเคราะห์ผ่านแบบจำลอง Recency, Frequency, Monetary (RFM) มาสร้างเป็นข้อมูลนำเข้าในการจัดกลุ่ม พร้อมเปรียบเทียบผลลัพธ์จากข้อมูลสองรูปแบบ คือ ข้อมูลที่ผ่านการลดมิติด้วยเทคนิค Principal Component Analysis (PCA) และข้อมูลที่ไม่มีการลดมิติ รวมถึงปรับค่าพารามิเตอร์ของแต่ละอัลกอริทึม และประเมินประสิทธิภาพด้วยค่า Silhouette Score ผลการทดลอง พบว่า อัลกอริทึม DBSCAN ที่ใช้ข้อมูลทั้ง 2 รูปแบบ เมื่อกำหนดค่า Epsilon เท่ากับ 1 และค่า min_samples เท่ากับ 10 ให้ผลลัพธ์ดีที่สุดเป็นสองอันดับแรก โดยมีค่า Silhouette Score เท่ากับ 0.9133 และ 0.8871 และสามารถจำแนกกลุ่มลูกค้าได้เป็น 2 และ 3 กลุ่ม ตามลำดับ จากนั้นได้นำผลการจัดกลุ่มดังกล่าวมาวิเคราะห์ตะกร้าสินค้าเพื่อค้นหากฎความสัมพันธ์ของสินค้าภายในแต่ละกลุ่ม โดยใช้อัลกอริทึม Frequent Pattern Growth (FP-Growth) โดยกำหนดค่า min_confidence เท่ากับ 0.01 และค่า Lift มากกว่าหรือเท่ากับ 1 พร้อมทั้งปรับค่า min_support ในระดับที่ต่างกัน ได้แก่ 0.01, 0.002 และ 0.001 ผลการวิเคราะห์ แสดงให้เห็นว่า ค่าของ min_support มีผลต่อจำนวนกฎความสัมพันธ์ที่ค้นพบ โดยเฉพาะในกลุ่มคลัสเตอร์ -1 ซึ่งเป็นกลุ่มของข้อมูลที่ไม่สามารถจำแนกได้อย่างชัดเจน หรือเป็นข้อมูล Noise/Outlier พบว่า มีรูปแบบการซื้อร่วมกันของสินค้าจำนวนมาก นอกจากนี้ คลัสเตอร์ที่ 1 ซึ่งได้จากข้อมูลของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ผ่านการลดมิติ ยังสามารถพบกฎความสัมพันธ์ได้มากที่สุด แม้มีการเพิ่มค่าของ min_support แล้วก็ตาม

คำสำคัญ : การเรียนรู้ของเครื่อง, การเรียนรู้ของเครื่องแบบไม่มีผู้สอน, การจัดกลุ่มลูกค้า, การวิเคราะห์ตะกร้าสินค้า, กฎความสัมพันธ์

Title	A HYBRID APPROACH USING CLUSTERING ALGORITHMS AND ASSOCIATION RULES FOR PRODUCT RECOMMENDATION IN RETAIL BUSINESSES
Author	PRATEEP PANYANONTAKARN
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	Assistant Professor Dr. Nuwee Wiwatwattana

Amid growing competition in the retail sector, businesses must adapt in both their distribution channels and their understanding of increasingly diverse consumer behavior. Customer analysis and segmentation are key factors that support effective marketing strategies and product selection appropriate for target groups. This study analyzes retail transaction data to perform customer segmentation using unsupervised machine learning algorithms, including K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), and Agglomerative Clustering from the input data were created using Recency, Frequency, and Monetary (RFM) analysis. The study compares clustering results using two datasets: one with dimensionality reduced using Principal Component Analysis (PCA), and one with original dimensions. Each algorithm underwent parameter tuning, and clustering performance was evaluated using the Silhouette Score. Experimental results show that DBSCAN, applied to both datasets with Epsilon set to 1 and min_samples set to 10, produced the two highest Silhouette Scores: 0.9133 and 0.8871, successfully segmenting customers into two and three clusters, respectively. These clustering results were subsequently used in market basket analysis to identify association rules within each customer group using the Frequent Pattern Growth (FP-Growth) algorithm. Parameters included a minimum confidence of 0.01 and a lift value of at least 1, with minimum support thresholds adjusted to 0.01, 0.002, and 0.001. The findings indicate that the minimum support value significantly affects the number of association rules identified. Notably, Cluster -1, which represents unclassified or noise/outlier data. This cluster showed a large number of co-purchased product patterns. Additionally, Cluster 1, derived from DBSCAN with PCA-reduced data, yielded the highest number of association rules even when the minimum support was increased.

Keyword : Machine learning, Unsupervised learning, Customer segmentation, Market Basket Analysis, Association rules

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยการสนับสนุนและความอนุเคราะห์จาก ผศ. ดร. นุวีย์ วิวัฒนวัฒนา อาจารย์ที่ปรึกษา ซึ่งได้ให้คำแนะนำและข้อคิดเห็นอันเป็นประโยชน์ในทุกขั้นตอนของการดำเนินงาน รวมถึงช่วยตรวจสอบความถูกต้องของเนื้อหาและสนับสนุนข้อมูลทางวิชาการ ทำให้ข้าพเจ้ามีความเข้าใจเชิงลึกเกี่ยวกับผลการทดลองและแนวทางการวิเคราะห์มากยิ่งขึ้น

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ และ ผศ. ดร. ศิริสรรพ เหล่าหะเกียรติ ที่ให้เกียรติเป็นกรรมการสอบ พร้อมทั้งให้คำแนะนำและข้อเสนอแนะที่เป็นประโยชน์อย่างยิ่งในการปรับปรุงสารนิพนธ์ให้มีความสมบูรณ์และถูกต้องมากยิ่งขึ้น

สุดท้ายนี้ขอกล่าวขอบพระคุณครอบครัว ที่เป็นกำลังใจสำคัญ และให้การสนับสนุนด้านค่าเล่าเรียนตลอดระยะเวลาการศึกษา รวมถึงขอขอบคุณเพื่อนร่วมสาขาวิชา ที่ให้ความช่วยเหลือและคำแนะนำในการเรียนและการจัดทำสารนิพนธ์จนสำเร็จลุล่วง

ประทีป ปัญญนนทกานต์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ภูมิหลัง

ในปัจจุบัน เทคโนโลยีมีบทบาทสำคัญในทุกด้านของชีวิตผู้คน ตั้งแต่กิจวัตรประจำวัน การศึกษา ตลอดจนถึงการทำงาน และการดำเนินธุรกิจ ความก้าวหน้าของเทคโนโลยีนี้ทำให้ความต้องการของผู้บริโภคมีความหลากหลายและซับซ้อนมากขึ้น อีกทั้งยังเปลี่ยนแปลงวิธีการดำเนินธุรกิจให้แตกต่างจากรูปแบบเดิมอย่างมาก จากที่ในอดีต หากต้องการซื้อสินค้า ผู้บริโภคจำเป็นต้องเดินทางไปยังห้างสรรพสินค้าหรือร้านค้าที่อยู่ห่างไกล แต่ในยุคนี้ การซื้อสินค้ากลายเป็นเรื่องง่ายและสะดวกสบายมากขึ้น เพราะมีช่องทางในการซื้อที่หลากหลายทั้งออนไลน์และออฟไลน์ การเปลี่ยนแปลงนี้ช่วยให้ผู้บริโภคประหยัดเวลา ไม่ต้องเดินทางไกลเพื่อซื้อสินค้า และยังมีเวลาในการพิจารณาและเปรียบเทียบสินค้าได้ละเอียดขึ้น ทั้งในด้านราคาและความคุ้มค่า ทำให้การตัดสินใจเลือกซื้อสินค้าจากแหล่งจำหน่ายต่างๆ เป็นไปได้อย่างมีประสิทธิภาพ

การเพิ่มขึ้นของช่องทางการจัดจำหน่ายสินค้าได้ส่งผลกระทบอย่างมากต่อภาคธุรกิจ โดยธุรกิจต่างๆ ต้องเผชิญกับความท้าทายจากการแข่งขันที่เพิ่มขึ้นในตลาด รวมถึงการตอบสนองความต้องการของกลุ่มลูกค้าที่มีความหลากหลายมากขึ้น ไม่ว่าจะเป็นธุรกิจที่ดำเนินในรูปแบบออนไลน์หรือออฟไลน์ ต่างก็ต้องค้นหาและพัฒนาวิธีการใหม่ๆ เพื่อให้สามารถปรับตัวและสอดคล้องกับการเปลี่ยนแปลงที่เกิดขึ้น หนึ่งในกลยุทธ์ที่ได้รับความนิยมสำหรับดึงดูดลูกค้าและกระตุ้นยอดขาย คือ การจัดทำโปรโมชั่น

การจัดทำโปรโมชั่นเป็นกลยุทธ์ที่ธุรกิจนิยมใช้เพื่อดึงดูดลูกค้าให้เข้ามาเลือกซื้อสินค้าหรือใช้บริการอย่างต่อเนื่อง ปัจจุบัน ธุรกิจในช่องทางออนไลน์มีการจัดโปรโมชั่นที่หลากหลาย เช่น การให้ส่วนลดราคาสินค้า ส่วนลดเพิ่มเติมเมื่อชำระผ่านบัตรเครดิต ส่วนลดสำหรับยอดสั่งซื้อที่ตรงตามเงื่อนไข หรือการจัดส่งสินค้าฟรี สำหรับธุรกิจในช่องทางออฟไลน์ก็มีการใช้กลยุทธ์ส่งเสริมการขายเช่นกัน โดยมักมุ่งเน้นที่ลูกค้าที่เป็นสมาชิกของร้านค้า เช่น การให้ส่วนลดเมื่อซื้อสินค้าเฉพาะรายการ หรือส่วนลดท้ายใบเสร็จ ซึ่งรูปแบบการส่งเสริมการขายในทั้งสองช่องทางนี้มีความคล้ายคลึงกันในเรื่องของการใช้ส่วนลดเพื่อกระตุ้นให้ลูกค้ากลับมาซื้อซ้ำและสร้างความพึงพอใจ

การสมัครสมาชิกกับแหล่งจัดจำหน่ายหรือร้านค้าในอดีต อาจเป็นเรื่องที่ยุ่งยากและไม่สะดวกสบายสำหรับลูกค้า เนื่องจากเทคโนโลยียังไม่ได้รับการพัฒนาเหมือนดังในปัจจุบัน ระบบสมาชิกจึงมีการทำอย่างจำกัด โดยจะมีเฉพาะในธุรกิจที่มีความจำเป็น เช่น การให้เช่าสื่อบันเทิงหรือการให้ยืมหนังสือ ซึ่งกระบวนการสมัครสมาชิกมักต้องกรอกข้อมูลลงบนกระดาษและเก็บ

รักษาไว้ในรูปแบบเอกสาร อย่างไรก็ตาม ด้วยการพัฒนาเทคโนโลยีในปัจจุบัน การเก็บข้อมูลสมาชิกทำได้ง่ายและสะดวกขึ้น ทั้งกับธุรกิจและลูกค้า โดยลูกค้าสามารถกรอกข้อมูลผ่านแบบฟอร์มอิเล็กทรอนิกส์บนมือถือ หรือคอมพิวเตอร์ ซึ่งข้อมูลจะถูกจัดเก็บในระบบอัตโนมัติและอยู่ในรูปแบบที่ธุรกิจสามารถนำไปใช้ได้อย่างมีประสิทธิภาพ จึงไม่น่าแปลกใจที่ระบบสมาชิกกลายเป็นสิ่งที่ได้รับความนิยมในธุรกิจยุคนี้

การให้สิทธิพิเศษแก่ลูกค้าที่เป็นสมาชิกเป็นกลยุทธ์ที่ธุรกิจนิยมใช้เพื่อสร้างความพึงพอใจและกระตุ้นให้ลูกค้ากลับมาใช้บริการซ้ำ สิทธิพิเศษนี้ครอบคลุมทั้งลูกค้าใหม่ที่เพิ่งสมัครสมาชิกและลูกค้าเดิมที่เป็นสมาชิกอยู่แล้ว โดยธุรกิจมุ่งเน้นการส่งเสริมการขายผ่านสมาชิกเพื่อสร้างความสัมพันธ์ระยะยาวและจัดเก็บข้อมูลเกี่ยวกับลูกค้า การที่ธุรกิจเข้าถึงข้อมูลของลูกค้า เช่น ประวัติการซื้อสินค้า จะช่วยธุรกิจสามารถจัดทำโปรโมชั่นได้ตรงกับความต้องการของลูกค้าอย่างมีประสิทธิภาพ นอกจากนี้ การเพิ่มจำนวนสมาชิกยังเป็นประโยชน์ต่อธุรกิจ เนื่องจากข้อมูลธุรกรรมที่ได้จากการซื้อสินค้า จะทำให้ธุรกิจสามารถวิเคราะห์และปรับกลยุทธ์การตลาดได้ดีขึ้น ซึ่งส่งผลให้ธุรกิจเข้าใจความต้องการของลูกค้าและสามารถปรับปรุงบริการให้สอดคล้องกับความคาดหวังของลูกค้า

เทคโนโลยีในปัจจุบันไม่ได้เพียงแค่สร้างช่องทางการจัดจำหน่ายใหม่ๆ แต่ยังมีบทบาทสำคัญในด้านการศึกษา โดยเฉพาะการพัฒนาและประยุกต์ใช้เทคนิคการสร้างแบบจำลองที่หลากหลาย การเติบโตของเทคโนโลยีนี้ทำให้การวิเคราะห์ข้อมูลขนาดใหญ่ที่มนุษย์ไม่สามารถทำได้ด้วยตนเองกลายเป็นไปได้ด้วยแบบจำลองทางคอมพิวเตอร์ แบบจำลองเหล่านี้ไม่ได้ถูกจำกัดเพียงการใช้งานประเภทเดียว แต่สามารถนำมาปรับใช้ได้ในหลายด้าน ไม่ว่าจะเป็นการจำแนกประเภทข้อมูล การพยากรณ์แนวโน้ม การตรวจจับความผิดปกติในข้อมูล หรือแม้แต่งานอื่นๆ ที่ต้องการการวิเคราะห์อย่างลึกซึ้ง

การทำความเข้าใจลักษณะของกลุ่มลูกค้าเป็นสิ่งสำคัญยิ่งสำหรับธุรกิจในยุคที่มีการแข่งขันสูง การเข้าถึงและวิเคราะห์ข้อมูลในอดีตของธุรกิจ การคาดการณ์แนวโน้มในอนาคต และการนำข้อมูลเข้าสู่การวิเคราะห์ผ่านแบบจำลอง ช่วยให้ธุรกิจสามารถแบ่งกลุ่มลูกค้าตามพฤติกรรมหรือความต้องการที่เฉพาะเจาะจงได้ นอกจากนี้ ยังสามารถค้นหาความสัมพันธ์ของสินค้าภายในแต่ละกลุ่มลูกค้า ซึ่งข้อมูลที่ได้จากแบบจำลองเหล่านี้เป็นประโยชน์ในการตัดสินใจและวางแผนการส่งเสริมการขายให้ตรงกับกลุ่มเป้าหมายที่หลากหลาย ทำให้ธุรกิจสามารถออกแบบกลยุทธ์ที่มีประสิทธิภาพและสอดคล้องกับความต้องการเฉพาะของลูกค้าแต่ละกลุ่มได้อย่างเหมาะสม

ในความเป็นจริง ธุรกิจหลายแห่งยังไม่ได้นำข้อมูลที่มีอยู่ไปใช้ให้เกิดประโยชน์สูงสุด โดยส่วนใหญ่ข้อมูลถูกนำมาใช้เพียงเพื่อตรวจสอบจำนวนสินค้าคงคลัง คำนวณยอดขาย หรือดูสถิติการขายที่ผ่านมาเท่านั้น แต่กลับไม่ได้ลงลึกในการวิเคราะห์อย่างเต็มที่ อุปสรรคสำคัญอาจมาจากขนาดของข้อมูลที่ใหญ่เกินกว่าจะให้มนุษย์วิเคราะห์ได้อย่างมีประสิทธิภาพ หรือเกิดจากความซับซ้อนในการเลือกใช้ของแบบจำลองและเทคนิคต่างๆ ตลอดจนการอภิปรายผลลัพธ์ที่ได้จากการวิเคราะห์ผ่านแบบจำลอง การวิเคราะห์ข้อมูลที่มากเกินไปอาจทำให้การประมวลผลยากลำบาก รวมถึงการวิเคราะห์ที่มีขั้นตอนซับซ้อนก็อาจเป็นอุปสรรคในการอภิปรายผลลัพธ์และนำไปใช้งานจริง

ในยุคปัจจุบัน การใช้เทคโนโลยีในการวิเคราะห์ข้อมูลถือเป็นสิ่งสำคัญสำหรับธุรกิจที่ต้องการเข้าใจพฤติกรรมของลูกค้าและปรับกลยุทธ์ให้สอดคล้องกับความต้องการเฉพาะของลูกค้าได้อย่างแม่นยำ หนึ่งในวิธีการที่มีประสิทธิภาพคือการแบ่งกลุ่มลูกค้า (Customer Segmentation) ด้วยเทคนิคการแบ่งกลุ่มข้อมูล (Clustering) ซึ่งจะช่วยให้ธุรกิจสามารถจัดกลุ่มลูกค้าได้ตามลักษณะ หรือพฤติกรรมที่มีความคล้ายคลึงกัน การแบ่งกลุ่มนี้ จะทำให้ธุรกิจสามารถออกแบบกลยุทธ์การตลาดที่เหมาะสมกับลูกค้าแต่ละกลุ่มได้ ส่งผลให้การสื่อสารและการนำเสนอผลิตภัณฑ์นั้น ตรงกับความต้องการของลูกค้าในแต่ละกลุ่มมากขึ้น และสามารถสร้างประสบการณ์ที่ดีขึ้นให้กับลูกค้า

นอกจากนี้ การวิเคราะห์ความสัมพันธ์ระหว่างสินค้าที่ลูกค้ามักซื้อพร้อมกัน หรือที่เรียกว่า การวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis) ยังเป็นเทคนิคสำคัญที่ช่วยให้ธุรกิจเข้าใจพฤติกรรมการซื้อของลูกค้าอย่างลึกซึ้ง การวิเคราะห์นี้ช่วยให้ธุรกิจสามารถระบุสินค้าที่มีแนวโน้มถูกซื้อร่วมกันได้ ซึ่งสามารถนำไปสู่การออกแบบกลยุทธ์การส่งเสริมการขายหรือการจัดวางสินค้าในร้านค้าได้อย่างเหมาะสม เพื่อเพิ่มโอกาสในการซื้อสินค้าเพิ่มเติมและกระตุ้นยอดขาย

การวิเคราะห์ข้อมูลของลูกค้าด้วยเทคนิคการแบ่งกลุ่มข้อมูล ร่วมกับเทคนิคการวิเคราะห์ตระกร้าสินค้า จะช่วยให้ธุรกิจสามารถแบ่งกลุ่มลูกค้าอย่างแม่นยำ รวมถึงเข้าใจพฤติกรรมการเลือกซื้อสินค้าของลูกค้าแต่ละกลุ่มได้อย่างละเอียด ซึ่งจะสนับสนุนการวางแผนกลยุทธ์ทางการตลาด การบริหารสินค้าคงคลัง และการสร้างโปรโมชั่นที่ตรงใจลูกค้าในแต่ละกลุ่ม ผลลัพธ์ที่ได้ไม่เพียงแต่เพิ่มความพึงพอใจของลูกค้าเท่านั้น แต่ยังช่วยเสริมสร้างความสัมพันธ์ที่ยั่งยืนและความภักดีต่อแบรนด์

1.2 **ความมุ่งหมายของงานวิจัย**

การวิจัยในครั้งนี้ ผู้วิจัยมีจุดประสงค์ต่องานวิจัยที่ศึกษา ดังนี้

1. เพื่อศึกษาและวิเคราะห์แบบจำลองที่ใช้ในการแบ่งกลุ่มข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Learning) โดยแบบจำลองที่นำมาศึกษาประกอบด้วย 4 แบบจำลอง ได้แก่

- แบบจำลอง K-Means Clustering
- แบบจำลอง Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- แบบจำลอง Gaussian Mixture Model (GMM)
- แบบจำลอง Agglomerative Clustering

พร้อมทั้งเปรียบเทียบและประเมินประสิทธิภาพของแต่ละแบบจำลองที่ได้จากการทดลองปรับค่าพารามิเตอร์ของแบบจำลอง ด้วยเมทริกต์ Silhouette Score เพื่อให้สามารถระบุแบบจำลองที่เหมาะสมที่สุดสำหรับการแบ่งกลุ่มข้อมูลของชุดข้อมูลนี้

2. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ในการแบ่งกลุ่มข้อมูล ระหว่างกรณีที่มีการลดมิติของข้อมูล (Dimensionality Reduction) ด้วยเทคนิค Principal Component Analysis (PCA) กับกรณีที่ไม่ได้ลดมิติ เพื่อพิจารณาว่าการลดมิติส่งผลต่อประสิทธิภาพของแบบจำลองในการแบ่งกลุ่มข้อมูลอย่างไร

3. เพื่อศึกษาเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนในการวิเคราะห์ตระกร้าสินค้า ด้วยการใช้แบบจำลอง Frequent Pattern Growth (FP-Growth) เพื่อค้นหาความสัมพันธ์ระหว่างสินค้าและกลุ่มลูกค้า โดยทำการวิเคราะห์แยกออกเป็นตามกลุ่มของลูกค้าที่ได้จากการแบ่งกลุ่มข้อมูลในการศึกษาก่อนหน้านี้ รวมทั้งเปรียบเทียบผลลัพธ์ที่ได้จากการกำหนดค่า Support, Confidence, และ Lift ที่แตกต่างกัน

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. เพื่อให้สามารถแบ่งกลุ่มลูกค้าได้อย่างชัดเจน และทราบถึงลักษณะเฉพาะของลูกค้าในแต่ละกลุ่ม

2. เพื่อให้ทราบและเข้าใจเทคนิคที่ช่วยให้แบบจำลองในการแบ่งกลุ่มข้อมูลมีประสิทธิภาพในการวิเคราะห์ที่ดียิ่งขึ้น

3. เพื่อให้ทราบถึงความสัมพันธ์ของกลุ่มสินค้าที่มีแนวโน้มถูกซื้อพร้อมกันในแต่ละกลุ่มลูกค้า

4. เพื่อให้สามารถนำผลลัพธ์จากการศึกษาที่ได้นี้ ไปใช้ในการออกแบบกลยุทธ์ส่งเสริมการขายให้เหมาะสมกับกลุ่มลูกค้าแต่ละกลุ่มในอนาคต

5. เพื่อเป็นแนวทางในงานวิจัยที่ศึกษาเกี่ยวกับการแบ่งกลุ่มลูกค้า ร่วมกับการวิเคราะห์ตะกร้าสินค้าด้วยเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนในประเทศไทย

1.4 ความสำคัญของงานวิจัย

งานวิจัยนี้ศึกษาวิธีการแบ่งกลุ่มลูกค้าจากชุดข้อมูลร้านค้าปลีก โดยนำข้อมูลที่ได้จากการวิเคราะห์ผ่านเทคนิคการวิเคราะห์ RFM (RFM Analysis) มาใช้เป็นคุณลักษณะ (Features) ในแบบจำลองที่ใช้ในการแบ่งกลุ่ม ด้วยเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน ซึ่งจะทำการวิเคราะห์และเปรียบเทียบประสิทธิภาพของแบบจำลอง 4 แบบจำลอง ได้แก่ แบบจำลอง K-Means Clustering, แบบจำลอง DBSCAN, แบบจำลอง GMM, และแบบจำลอง Agglomerative Clustering รวมถึงการทดลองลดมิติของข้อมูล เพื่อให้ได้แบบจำลองในการแบ่งกลุ่มที่มีประสิทธิภาพที่ดีที่สุด นอกจากนี้ งานวิจัยยังนำผลลัพธ์จากแบบจำลองการแบ่งกลุ่มที่ได้ มาศึกษาการวิเคราะห์ตะกร้าของแต่ละกลุ่มลูกค้า ด้วยการใช้เทคนิค FP-Growth ในการค้นหาความสัมพันธ์ของสินค้าที่ลูกค้าแต่ละกลุ่มมักซื้อร่วมกัน ภายใต้เงื่อนของค่าพารามิเตอร์ที่ได้กำหนด

1.5 ขอบเขตของงานวิจัย

1.5.1 กลุ่มตัวอย่างที่ใช้ในการวิจัย

ชุดข้อมูลที่ใช้ในการวิจัยในครั้งนี้ เป็นชุดข้อมูลที่ชื่อว่า Retail Transaction Dataset จากแหล่งข้อมูลสาธารณะบนเว็บไซต์ Kaggle.com ซึ่งข้อมูลทั้งหมดถูกสร้างขึ้นมาจากไลบรารี่ (Library) ที่ชื่อว่า Python Faker โดยจัดเก็บข้อมูลธุรกรรมของร้านค้าปลีกตั้งแต่วันที่ 1 มกราคม 2020 ถึง 18 พฤษภาคม 2024 ประกอบไปด้วย 13 คุณลักษณะ และข้อมูลจำนวน 1,000,000 ธุรกรรม

1.5.2 ตัวแปรที่ศึกษา

- Transaction_ID คือ เลขที่ที่ใช้ระบุการทำธุรกรรมแต่ละรายการ
- Date คือ วันที่และเวลาที่ทำธุรกรรมแต่ละครั้ง
- Customer_Id คือ รหัสลูกค้าที่ทำการซื้อสินค้า
- Product คือ รายการสินค้าที่ถูกซื้อในแต่ละธุรกรรม
- Total_Items คือ ปริมาณของสินค้าที่ถูกซื้อในธุรกรรม
- Total_Cost คือ มูลค่ารวมของการซื้อของธุรกรรม
- Payment_Method คือ วิธีการชำระเงินในการทำธุรกรรม

- City คือ เมืองที่เกิดการทำธุรกรรม
- Store_Type คือ ประเภทของร้านค้า
- Discount_Applied คือ มีการใช้ส่วนลดกับธุรกรรมนั้นหรือไม่ (True: ใช้/False: ไม่ใช้)
- Customer_Category คือ ประเภทที่แสดงถึงภูมิหลังหรือกลุ่มอายุของลูกค้า
- Season คือ ฤดูกาลที่เกิดการทำธุรกรรม
- Promotion คือ ประเภทของโปรโมชั่นที่ใช้กับธุรกรรม

1.6 กรอบแนวคิดในงานวิจัย

งานวิจัยนี้ศึกษาแบบจำลองที่ใช้ในการแบ่งกลุ่มลูกค้า และการวิเคราะห์ตระกร้าสินค้าของลูกค้าแต่ละกลุ่มในธุรกิจร้านค้าปลีก ด้วยการใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน โดยใช้ชุดข้อมูลธุรกรรมในร้านค้าปลีก จากแหล่งข้อมูลสาธารณะบนเว็บไซต์ Kaggle.com ประกอบด้วย 13 คุณลักษณะ และจำนวนข้อมูลทั้งหมด 1,000,000 ธุรกรรม ซึ่งชุดข้อมูลนี้ได้ถูกบันทึกเป็นไฟล์นามสกุล CSV และจัดเก็บข้อมูลอยู่ในรูปแบบของตาราง

ในการศึกษานี้ ผู้วิจัยได้เลือกใช้ Google Colab เป็นเครื่องมือที่ใช้ในการประมวลผลข้อมูล และเขียนโปรแกรมด้วยภาษา Python โดยขั้นตอนในการศึกษาแบบจำลอง ประกอบด้วย 2 ขั้นตอนหลัก คือ 1. การสร้างแบบจำลองเพื่อใช้สำหรับแบ่งกลุ่มลูกค้า โดยทำการสร้างแบบจำลองทั้งหมด 4 แบบจำลอง ได้แก่ แบบจำลอง K-Means Clustering, แบบจำลอง DBSCAN, แบบจำลอง GMM, และแบบจำลอง Agglomerative Clustering ซึ่งในการสร้างแบบจำลองนี้ ใช้ข้อมูลที่ได้จากการวิเคราะห์ด้วย RFM Analysis มาใช้เป็นคุณลักษณะในแบบจำลองการแบ่งกลุ่ม รวมทั้งจะทำการทดลองปรับค่าพารามิเตอร์ของแต่ละแบบจำลอง รวมถึงการทดลองลดจำนวนมิติของข้อมูล ด้วยเทคนิค PCA และทำการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองและแต่ละรูปแบบของข้อมูล ด้วยการใช้เมทริกซ์ Silhouette Score ในการประเมินประสิทธิภาพ 2. การวิเคราะห์ตระกร้าสินค้าของแต่ละกลุ่มลูกค้าที่ได้จากแบบจำลองที่มีประสิทธิภาพที่ดีที่สุด 2 อันดับแรกในขั้นตอนก่อนหน้า มาทำการวิเคราะห์หากฎความสัมพันธ์ของสินค้า ซึ่งทดลองปรับค่าพารามิเตอร์ของแบบจำลอง เพื่อให้ได้ค่าที่เหมาะสมที่สุดในการนำมาใช้เป็นเงื่อนไขในการวิเคราะห์ตระกร้าสินค้าของลูกค้าแต่ละกลุ่ม

1.7 สมมุติฐานของการวิจัย

1. แบบจำลองที่ใช้ในการแบ่งกลุ่มจากเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนในแต่ละรูปแบบ สามารถแบ่งแยกกลุ่มลูกค้าแต่ละกลุ่มได้อย่างชัดเจนด้วยค่า Silhouette Score ที่มากกว่า 0.60

2. การลดมิติของข้อมูลที่ใช้ในการแบ่งกลุ่มลูกค้าของแบบจำลองทำให้ค่า Silhouette Score ที่ได้มีค่าสูงกว่าแบบจำลองที่ไม่ได้มีการลดมิติของข้อมูล เมื่อเปรียบเทียบกับแบบจำลองเดียวกัน

3. การวิเคราะห์ตระกร้าสินค้า พบกฎความสัมพันธ์ของสินค้าของแต่ละกลุ่มลูกค้าอย่างน้อย 2 กฎความสัมพันธ์ เมื่อกำหนดค่า Min Support เท่ากับ 50%, Min Confidence เท่ากับ 50% และ Lift ที่มีค่ามากกว่า 1

บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารทฤษฎีและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

2.1. ทฤษฎีที่เกี่ยวข้อง

2.1.1 ทฤษฎีเกี่ยวกับการเตรียมข้อมูล (Data Preprocessing)

2.1.2 ทฤษฎีเกี่ยวกับการลดจำนวนมิติของข้อมูล (Dimensionality Reduction)

2.1.3 ทฤษฎีเกี่ยวกับแบบจำลอง Recency, Frequency, Monetary (RFM Analysis)

2.1.4 ทฤษฎีเกี่ยวกับอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูล (Clustering)

2.1.5 ทฤษฎีเกี่ยวกับการเลือกจำนวนกลุ่มที่เหมาะสมด้วยเทคนิค Elbow Method

2.1.6 ทฤษฎีเกี่ยวกับอัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis)

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ทฤษฎีเกี่ยวกับการเตรียมข้อมูล (Data Preprocessing)

Han และคณะ (2011) อธิบายว่า การปรับช่วงขอบเขตของข้อมูล (Data Scaling) เป็นหนึ่งในกระบวนการที่สำคัญในการเตรียมข้อมูล (Data Preprocessing) ซึ่งจำเป็นอย่างยิ่งสำหรับอัลกอริทึมที่พึ่งพาการคำนวณระยะทาง เช่น K-Means Clustering หรือ Support Vector Machine (SVM) เนื่องจากค่าที่แตกต่างกันอย่างมากระหว่างคุณลักษณะ (Features) อาจทำให้อัลกอริทึมให้ความสำคัญกับบางคุณลักษณะมากเกินไป ส่งผลให้เกิดการบิดเบือนในการเรียนรู้ของแบบจำลอง ตัวอย่างเช่น ข้อมูลที่มีช่วงค่ากว้าง เช่น ยอดขายในหน่วยพันบาท อาจมีอิทธิพลมากกว่าข้อมูลที่มีช่วงค่าจำกัด เช่น จำนวนครั้งการซื้อ (1–50 ครั้ง) ซึ่งอาจนำไปสู่การวิเคราะห์ที่ไม่สะท้อนความสัมพันธ์ที่แท้จริงของข้อมูล

การปรับช่วงขอบเขตของข้อมูลยังมีบทบาทสำคัญในการเพิ่มประสิทธิภาพในการคำนวณ โดยจะช่วยลดปัญหาความไม่เสถียรทางตัวเลข ที่อาจจะเกิดขึ้นในอัลกอริทึมที่ต้องจัดการกับค่าตัวเลขขนาดใหญ่หรือเล็กมาก นอกจากนี้ การปรับช่วงขอบเขตของข้อมูลยังช่วยให้

อัลกอริทึมสามารถเรียนรู้และคำนวณได้เร็วขึ้น ช่วยลดระยะเวลาที่ใช้ในการประมวลผล และทำให้ผลลัพธ์ที่ได้มีความแม่นยำสูงขึ้น

เทคนิคที่ใช้ในการปรับช่วงขอบเขตของข้อมูลมีหลากหลาย ขึ้นอยู่กับลักษณะของข้อมูลและอัลกอริทึมที่จะนำไปใช้ในการวิเคราะห์ โดยหนึ่งในเทคนิคที่ได้รับนิยม คือ Standardization หรือ Z-Score Normalization ซึ่งเป็นการปรับค่าของข้อมูลให้มีค่าเฉลี่ย (μ) เท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐาน (σ) เท่ากับ 1 ดังแสดงในสมการ 2.1

Pedregosa และคณะ (2011) อธิบายว่า การทำ Standardization โดยใช้ StandardScaler เป็นแนวทางมาตรฐานในการเตรียมข้อมูลก่อนนำเข้าสู่อัลกอริทึมที่อิงกับการวัดระยะทาง ซึ่งเทคนิคนี้ช่วยลดผลกระทบจากความแตกต่างของช่วงขอบเขตข้อมูลในแต่ละคุณลักษณะ และช่วยเพิ่มประสิทธิภาพในการเรียนรู้ของแบบจำลอง

$$z = \frac{x-\mu}{\sigma} \tag{2.1}$$

โดยที่

z คือ ค่าที่ถูกปรับสเกลหลังการทำ Standardization

x คือ ค่าข้อมูลดั้งเดิมที่ต้องการปรับสเกล

μ คือ ค่าเฉลี่ยของข้อมูลในแต่ละมิติ (Feature)

σ คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูลในแต่ละมิติ (Feature)

เทคนิค Standardization เหมาะสำหรับข้อมูลที่มีการกระจายตัวแบบปกติ (Normal Distribution) โดยช่วยให้คุณลักษณะทั้งหมดมีการกระจายตัวอย่างสมมาตร และลดอิทธิพลจากหน่วยหรือขนาดของข้อมูลที่แตกต่างกันภายในชุดข้อมูล เช่น การวิเคราะห์ข้อมูลประชากรที่ประกอบด้วยข้อมูลรายได้ (หน่วยบาท) และอายุ (หน่วยปี) ซึ่งช่วยให้สามารถนำคุณลักษณะทั้งสองมาวิเคราะห์ร่วมกันได้โดยไม่เกิดความเอนเอียงจากความแตกต่างของหน่วยหรือขนาดของข้อมูล อย่างไรก็ตาม Zheng และ Casari (2018) อธิบายว่า แม้ Standardization จะมีประโยชน์ในด้านดังกล่าว แต่ก็มีข้อจำกัดที่สำคัญ โดยเฉพาะความอ่อนไหวต่อค่าความผิดปกติ (Outliers) ซึ่งอาจบิดเบือนค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ส่งผลให้ค่าที่ปรับสเกลคลาดเคลื่อนจากความเป็นจริงของข้อมูล นอกจากนี้ หากข้อมูลมีการกระจายตัวแบบเบ้ (Skewed Distribution) การใช้ Standardization อาจไม่สามารถแก้ไขปัญหาความไม่สมมาตรได้อย่างมีประสิทธิภาพเท่ากับเทคนิคอื่น

2.1.2 **ทฤษฎีเกี่ยวกับการลดจำนวนมิติของข้อมูล** (Dimensionality Reduction)

การลดจำนวนมิติของข้อมูลเป็นกระบวนการที่มีบทบาทสำคัญในงานวิเคราะห์ข้อมูล โดยมุ่งเน้นการลดจำนวนคุณลักษณะหรือจำนวนมิติในชุดข้อมูล เพื่อแก้ไขปัญหาความซับซ้อนของข้อมูลที่มีมิติจำนวนมาก และเพิ่มประสิทธิภาพของอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูล ปัญหาหลักที่เกิดจากข้อมูลที่มีมิติสูง เรียกว่า Curse of Dimensionality ซึ่งอาจทำให้โครงสร้างสำคัญในข้อมูลถูกบดบัง และระยะห่างระหว่างจุดข้อมูลเพิ่มขึ้น จนทำให้ผลลัพธ์ที่ได้มีความน่าเชื่อถือลดลง

นอกจากการลดความซับซ้อนของข้อมูลแล้ว การลดจำนวนมิติยังช่วยเพิ่มความเร็วในการคำนวณ ลดปัญหา Overfitting และช่วยให้ผู้วิเคราะห์สามารถสร้างแผนภาพ (Visualization) ของข้อมูลได้ดีขึ้น โดยกระบวนการลดจำนวนมิติ สามารถแบ่งออกเป็น 2 ประเภท ได้แก่ การเลือกคุณลักษณะ (Feature Selection) และ การสกัดคุณลักษณะ (Feature Extraction)

การเลือกคุณลักษณะ คือ กระบวนการเลือกคุณลักษณะที่สำคัญที่สุดในชุดข้อมูล และตัดคุณลักษณะที่ไม่เกี่ยวข้องหรือมีความสำคัญต่ำออก เทคนิคนี้จะช่วยลดความซับซ้อนของแบบจำลอง เพิ่มความเร็วในการประมวลผล และลดโอกาสเกิด Overfitting โดยมีเทคนิค ดังนี้

1. Filter Methods

ใช้ตัวชี้วัดทางสถิติ เช่น ค่า Chi-square, ค่า Correlation หรือค่า Mutual Information เพื่อวัดความสัมพันธ์ระหว่างคุณลักษณะและตัวแปรเป้าหมาย (Target Variable)

2. Wrapper Methods

ใช้แบบจำลองการเรียนรู้ด้วยเครื่อง เพื่อประเมินความสำคัญของคุณลักษณะ เช่น Recursive Feature Elimination (RFE)

3. Embedded Methods

ใช้อัลกอริทึมที่มีการประเมินความสำคัญของคุณลักษณะ (Feature Importance) ในตัว เช่น แบบจำลอง Decision Trees หรือแบบจำลอง Random Forest

การสกัดคุณลักษณะ คือ กระบวนการแปลงคุณลักษณะเดิมให้กลายเป็นคุณลักษณะใหม่ที่จัดเก็บข้อมูลรวมกัน และทำการลดจำนวนมิติลง โดยวิธีนี้จะสร้าง Latent Features ที่สามารถอธิบายความแปรปรวนหรือความสำคัญในข้อมูลได้ ตัวอย่างอัลกอริทึมที่เด่นชัด คือ PCA และ t-SNE

Principal Component Analysis หรือ PCA เป็นอัลกอริทึมที่นิยมใช้ในการลดจำนวนมิติที่อาศัยการแปลงเชิงเส้น (Linear Transformation) เพื่อค้นหาแกนใหม่ที่เรียกว่า Principal

Components โดยแกนเหล่านี้ถูกสร้างขึ้นเพื่ออธิบายความแปรปรวนในข้อมูลเดิมให้ได้มากที่สุด โดยแกนแต่ละแกนจะมีลักษณะเป็นอิสระต่อกัน แนวคิดพื้นฐานของ PCA นี้ถูกเสนอเป็นครั้งแรกโดย Pearson (1901) ซึ่งได้อธิบายวิธีการหาทิศทางของเส้นตรงหรือระนาบที่เหมาะสมที่สุดในเชิงเรขาคณิต สำหรับการฉายข้อมูลที่อยู่ในมิติสูงให้มีมิติน้อยลง โดยยังคงโครงสร้างสำคัญของข้อมูลไว้ให้มากที่สุด

Jolliffe และ Cadima (2016) อธิบายว่า PCA เป็นเทคนิคพื้นฐานที่มีประสิทธิภาพในการลดจำนวนมิติของข้อมูล โดยสามารถรักษาความแปรปรวนหลักของข้อมูลไว้ได้อย่างดี อีกทั้งยังสามารถประยุกต์ใช้ในสถานการณ์ที่ซับซ้อนมากขึ้น เช่น Sparse PCA ซึ่งช่วยให้ผลลัพธ์ตีความได้ง่ายขึ้นด้วยการลดจำนวนตัวแปรที่เกี่ยวข้องในแต่ละองค์ประกอบหลัก รวมถึงการนำไปใช้กับข้อมูลที่ไม่ใช่เชิงตัวเลข เช่น ข้อมูลประเภทกลุ่ม (Categorical Data) ผ่านการแปลงที่เหมาะสม PCA จึงได้รับความนิยมอย่างแพร่หลายในงานที่มีข้อมูลมิติสูง เช่น การวิเคราะห์ภาพหรือข้อมูลพันธุกรรม เนื่องจากสามารถลดขนาดข้อมูลได้อย่างมีประสิทธิภาพโดยไม่สูญเสียโครงสร้างสำคัญของข้อมูลเดิม

ขั้นตอนการลดจำนวนมิติของข้อมูล

1. การลบค่าเฉลี่ยออกจากข้อมูล (Mean Centering)

ปรับข้อมูลให้อยู่ในรูปศูนย์กลาง ด้วยการลบค่าเฉลี่ยของแต่ละคุณลักษณะกับข้อมูล ดังแสดงในสมการ 2.2 ซึ่งจะช่วยปรับข้อมูลให้มีค่าเฉลี่ยเป็นศูนย์ เป็นขั้นตอนพื้นฐานก่อนการคำนวณเมทริกซ์ความแปรปรวนร่วม

$$x' = x - \mu \tag{2.2}$$

โดยที่

x' คือ ค่าข้อมูลหลังจากลบค่าเฉลี่ยออก

x คือ ค่าข้อมูลต้นฉบับในแต่ละมิติ

μ คือ ค่าเฉลี่ยของข้อมูลในแต่ละมิตินั้น

2. การคำนวณเมทริกซ์ความแปรปรวนร่วม (Covariance Matrix)

คำนวณหาความสัมพันธ์ระหว่างคุณลักษณะ ดังแสดงในสมการ 2.3 โดยเมทริกซ์ C แสดงถึง ความแปรปรวนและความสัมพันธ์ระหว่างคุณลักษณะ ที่ช่วยระบุทิศทางที่มีความแปรปรวนสูงสุดในข้อมูล

$$C = \frac{1}{n-1} - X^T X \quad (2.3)$$

โดยที่

C คือ เมทริกซ์ความแปรปรวนร่วม (Covariance Matrix)

n คือ จำนวนตัวอย่าง (Sample Size)

X คือ เมทริกซ์ข้อมูลหลังจากลบค่าเฉลี่ยแล้ว

X^T คือ เมทริกซ์ X ที่ถูกสลับแถวกับคอลัมน์ (Transpose)

3. การคำนวณ Eigenvalues และ Eigenvectors

คำนวณ Eigenvalues (λ) และ Eigenvectors (v) โดยใช้สมการ 2.4 ซึ่ง Eigenvalues จะช่วยบ่งบอกว่า ความแปรปรวนมีมากน้อยเพียงใดในแต่ละแกนใหม่ (Principal Component) ในขณะที่ Eigenvectors จะระบุทิศทางของแกนเหล่านั้น

$$Cv = \lambda v \quad (2.4)$$

โดยที่

Cv คือ เมทริกซ์ความแปรปรวนร่วม (Covariance Matrix)

λ คือ Eigenvalue

v คือ Eigenvector

4. การเลือก Principal Components

จัดลำดับ Eigenvalues จากมากไปน้อย และเลือก Eigenvectors ที่มีความสัมพันธ์กับ Eigenvalues สูงที่สุด

5. การแปลงค่าข้อมูล

คำนวณค่าข้อมูลใหม่ (Z) ด้วยการคูณข้อมูลเดิมกับ Eigenvectors ที่ได้ทำการเลือก ดังแสดงในสมการ 2.5 โดยข้อมูลที่ได้ในมิติใหม่จะมีจำนวนมิติที่น้อยลง แต่ยังคงไว้ซึ่งความแปรปรวนสำคัญของข้อมูลเดิม

$$Z = XV \quad (2.5)$$

โดยที่

Z คือ ข้อมูลใหม่ที่ลดมิติแล้ว

X คือ ข้อมูลต้นฉบับที่ถูกลบค่าเฉลี่ยออกแล้ว

V คือ เมทริกซ์ที่ประกอบด้วย Eigenvectors ที่ถูกเลือกไว้

PCA ช่วยลดมิติของข้อมูลได้อย่างมีประสิทธิภาพในกรณีที่ข้อมูลมีลักษณะความสัมพันธ์แบบเชิงเส้น (Linear Relationships) โดยจะทำการรักษาความแปรปรวนสำคัญในข้อมูลเดิมเอาไว้ อัลกอริทึมนี้ยังช่วยปรับปรุงความเร็วและลดการใช้ทรัพยากรในกระบวนการคำนวณ อีกทั้ง ยังมีประโยชน์อย่างมากในการแสดงผลข้อมูล เมื่อจำนวนมิติถูกลดลงเหลือ 2 มิติ หรือ 3 มิติ ซึ่งช่วยให้ผู้วิเคราะห์เข้าใจข้อมูลได้ง่ายขึ้น

อย่างไรก็ตาม อัลกอริทึม PCA ไม่เหมาะสำหรับข้อมูลที่มีความสัมพันธ์แบบไม่เชิงเส้น (Non-linear Relationships) นอกจากนี้ การตีความแกนใหม่หรือ Principal Components อาจทำได้ยากในเชิงธุรกิจ เพราะแกนเหล่านี้เป็นการรวมกันเชิงเส้นของคุณลักษณะเดิม และไม่ได้สะท้อนถึงความหมายที่ชัดเจนในบริบทของปัญหาที่ศึกษา

2.1.3 ทฤษฎีเกี่ยวกับแบบจำลอง Recency, Frequency, Monetary (RFM)

แบบจำลอง Recency, Frequency, Monetary (RFM) เป็นหนึ่งในวิธีการที่ได้รับความนิยมในการวิเคราะห์พฤติกรรมลูกค้า โดยมุ่งเน้นการระบุและแบ่งกลุ่มลูกค้าที่มีมูลค่าต่อองค์กรตามข้อมูลพฤติกรรมการซื้อในอดีต แบบจำลองนี้พัฒนาขึ้นมาเพื่อตอบโจทย์การตลาดยุคใหม่ที่ต้องการความเข้าใจเชิงลึกเกี่ยวกับลูกค้า โดยอาศัยการวัดค่าผ่านสามคุณลักษณะที่สำคัญ ได้แก่ Recency (R), Frequency (F) และ Monetary (M) ซึ่งแสดงถึงระยะเวลานับจากการซื้อครั้งล่าสุด ความถี่ในการซื้อ และมูลค่ารวมของการซื้อสินค้า ตามลำดับ

แนวคิดแบบจำลอง RFM ได้รับการเสนอครั้งแรกโดย Hughes (1994) ซึ่งอธิบายว่าการจัดอันดับลูกค้าจากข้อมูลเชิงธุรกรรมใน 3 มิตินี้ สามารถช่วยให้นักการตลาดวางแผนแคมเปญได้แม่นยำและคุ้มค่า โดยไม่จำเป็นต้องอาศัยข้อมูลประชากรศาสตร์เพิ่มเติม ซึ่งทำให้ RFM เป็นเครื่องมือที่เรียบง่ายแต่มีประสิทธิภาพสูงในการระบุลูกค้าสำคัญ

แบบจำลอง RFM มีความสำคัญในการช่วยให้องค์กรสามารถระบุกลุ่มลูกค้าที่มีศักยภาพสูงเพื่อสร้างความสัมพันธ์เชิงบวกในระยะยาวและเพิ่มมูลค่าของลูกค้า (Customer Lifetime Value: CLV) การนำแบบจำลองนี้มาใช้จะช่วยให้องค์กรสามารถออกแบบกลยุทธ์ทางการตลาดได้อย่างมีประสิทธิภาพ รวมทั้งปรับแต่งให้เหมาะสมกับลูกค้าแต่ละกลุ่ม ทั้งนี้ ยังช่วยลด

ต้นทุนและเพิ่มผลตอบแทนจากการลงทุน (Return on Investment, ROI) ในการทำการตลาด ทั้งยัง มีบทบาทสำคัญในการช่วยองค์กรทำความเข้าใจพฤติกรรมของลูกค้าในระดับบุคคลและกลุ่มมากยิ่งขึ้น

Linoff และ Berry (2011) อธิบายว่า RFM เป็นเครื่องมือหลักในกระบวนการทำ Data-Driven Marketing โดยเฉพาะอย่างยิ่งเมื่อต้องการระบุกลุ่มลูกค้าที่มีแนวโน้มตอบสนองต่อแคมเปญโปรโมชั่น หรือกลุ่มที่มีแนวโน้มเลิกซื้อสินค้า ซึ่งสามารถนำไปสู่การพัฒนาแผน CRM ที่ตรงเป้าหมาย ตัวอย่างเช่น การวิเคราะห์ RFM สามารถช่วยระบุลูกค้าที่มีความภักดีต่อองค์กรสูง ซึ่งมักจะมีค่า Frequency และ Monetary สูง รวมถึงลูกค้าที่มีแนวโน้มลดลงในความสัมพันธ์กับองค์กรผ่านค่า Recency ที่สูงขึ้น

ขั้นตอนการวิเคราะห์แบบจำลอง RFM

กระบวนการของการวิเคราะห์แบบจำลอง RFM เริ่มต้นจากการรวบรวมข้อมูลของลูกค้า เช่น วันที่ของการทำธุรกรรม จำนวนครั้งของการซื้อ และมูลค่าการซื้อสินค้า จากนั้นข้อมูลดังกล่าวจะถูกแปลงให้เป็นตัวเลขเชิงเปรียบเทียบผ่านการคำนวณค่า Recency, ค่า Frequency, และค่า Monetary และจัดอันดับหรือแบ่งกลุ่มตามเกณฑ์ที่กำหนด เพื่อระบุความสำคัญของลูกค้าในแต่ละมิติ ซึ่งแสดงเป็น 3 ขั้นตอน ดังนี้

ขั้นตอนที่ 1: การรวบรวมข้อมูลที่เกี่ยวข้องกับการทำธุรกรรมของลูกค้า ได้แก่ วันที่ของการซื้อครั้งล่าสุดเพื่อใช้คำนวณค่า Recency จำนวนการซื้อสินค้าตลอดระยะเวลาที่กำหนดเพื่อใช้คำนวณค่า Frequency และมูลค่ารวมของการซื้อสินค้าตลอดเวลาเพื่อใช้คำนวณค่า Monetary

ขั้นตอนที่ 2: การคำนวณค่าของทั้ง 3 คุณลักษณะ โดยมีวิธีการคำนวณ ดังนี้

- Recency (R): วัดระยะเวลาจากการซื้อครั้งล่าสุดกับวันที่ที่กำหนด ดังในสมการ 2.6

$$Recency = Current\ Date - Last\ Purchase\ Date \quad (2.6)$$

- Frequency (F): วัดจำนวนครั้งที่ลูกค้าซื้อสินค้า ดังในสมการ 2.7

$$Frequency = Total\ Number\ of\ Purchases \quad (2.7)$$

- Monetary (M): คำนวณมูลค่ารวมของการซื้อสินค้า ดังในสมการ 2.8

$$Monetary = \sum_{i=1}^{n} Purchase\ Value_i \quad (2.8)$$

โดยข้อมูลแต่ละคุณลักษณะจะถูกแปลงให้เป็นช่วงคะแนน เช่น 1 ถึง 5 โดยอาจใช้วิธีการแบ่งข้อมูลตามค่าควอร์ไทล์ (Quartile) หรือค่าเปอร์เซ็นไทล์ (Percentile) ซึ่งคะแนนสูงสุด แสดงถึง ลูกค้าที่มีค่าที่ดีที่สุดในแต่ละคุณลักษณะ

ขั้นตอนที่ 3: หลังจากให้คะแนนในแต่ละคุณลักษณะแล้ว คะแนนเหล่านี้จะถูกรวมเพื่อสร้างคะแนน RFM ที่แสดงถึงความสำคัญของลูกค้าโดยรวม ตัวอย่างเช่น ลูกค้าที่ได้คะแนน R=5, F=5, และ M=5 ถือเป็นลูกค้าที่สำคัญที่สุดในแง่ของความสัมพันธ์กับองค์กร และจะทำการแบ่งกลุ่มลูกค้าจากคะแนน RFM ที่ได้

แบบจำลอง RFM มีข้อได้เปรียบอย่างมากในกระบวนการวิเคราะห์และทำความเข้าใจลูกค้า โดยเฉพาะในแง่ของการประหยัดเวลาและทรัพยากร เนื่องจากเป็นกระบวนการวิเคราะห์ที่เรียบง่ายแต่ทรงพลัง องค์กรสามารถนำแบบจำลอง RFM ไปใช้ระบุลูกค้าที่มีมูลค่าสูงได้อย่างรวดเร็ว เพียงแค่มีข้อมูลพื้นฐานเกี่ยวกับการทำธุรกรรมของลูกค้า เช่น วันที่ของการซื้อ จำนวนครั้งของการซื้อ และมูลค่าการซื้อสินค้า นอกจากนี้ แบบจำลอง RFM ยังช่วยสนับสนุนการตัดสินใจเชิงกลยุทธ์ทางการตลาดให้กับกลุ่มลูกค้าที่สำคัญที่สุด ซึ่งสามารถเพิ่มโอกาสในการรักษาความสัมพันธ์กับลูกค้ากลุ่มเป้าหมาย

อีกหนึ่งข้อได้เปรียบสำคัญ คือ แบบจำลอง RFM มีความยืดหยุ่นสูงและสามารถปรับใช้ได้ในหลากหลายอุตสาหกรรม ไม่ว่าจะเป็นธุรกิจค้าปลีก ธุรกิจบริการ หรือธุรกิจออนไลน์ การแบ่งกลุ่มลูกค้าด้วยแบบจำลอง RFM จะช่วยให้องค์กรสามารถปรับแต่งกลยุทธ์การตลาดให้เหมาะสมกับพฤติกรรมของลูกค้าแต่ละกลุ่มได้ง่ายขึ้น นอกจากนี้ แบบจำลอง RFM ยังมีประสิทธิภาพสูงในการช่วยองค์กรระบุโอกาสทางการตลาดใหม่ๆ โดยสามารถวิเคราะห์ลูกค้าที่มีความภักดีสูง หรือค้นพบกลุ่มลูกค้าที่มีโอกาสขยายมูลค่าผ่านการส่งเสริมการขาย

อย่างไรก็ตาม แม้ว่าแบบจำลอง RFM จะเป็นเครื่องมือที่มีประสิทธิภาพ แต่ก็มีข้อจำกัดบางประการที่ควรพิจารณา หนึ่งในข้อจำกัดสำคัญ คือ แบบจำลอง RFM ต้องการข้อมูลธุรกรรมที่ชัดเจนเพื่อใช้ในการวิเคราะห์ หากองค์กรไม่มีข้อมูลดังกล่าว หรือข้อมูลธุรกรรมที่เก็บไว้อยู่ในรูปแบบที่ไม่สามารถนำมาใช้งานได้ การวิเคราะห์นี้จะไม่สามารถดำเนินการได้อย่างมีประสิทธิภาพ

อีกทั้ง แบบจำลอง RFM ยังมุ่งเน้นเพียงสามมิติ ได้แก่ Recency, Frequency และ Monetary ซึ่งแม้ว่าจะครอบคลุมพฤติกรรมการซื้อในระดับหนึ่ง แต่ก็อาจไม่เพียงพอในการอธิบายความซับซ้อนของพฤติกรรมลูกค้าทั้งหมด เช่น ความพึงพอใจของลูกค้า การปฏิสัมพันธ์กับแบรนด์ หรือพฤติกรรมเชิงคุณภาพอื่นๆ นอกจากนี้ การให้คะแนนในแต่ละมิติ เช่น การแบ่งคะแนน RFM เป็นช่วงคะแนน 1 ถึง 5 อาจต้องการการปรับแต่งอย่างเหมาะสม เพื่อให้สอดคล้องกับบริบทเฉพาะของธุรกิจ หากการกำหนดคะแนนไม่ได้สะท้อนถึงลักษณะของลูกค้าอย่างถูกต้อง ผลลัพธ์จากการวิเคราะห์อาจนำไปสู่การตัดสินใจที่ไม่ถูกต้องได้

และสุดท้าย แบบจำลอง RFM ยังมีข้อจำกัดในแง่ของการใช้งานในกลุ่มลูกค้าที่ไม่ได้มีพฤติกรรมการซื้อที่ชัดเจน เช่น ลูกค้าที่ใช้บริการในลักษณะสมัครสมาชิก หรือกลุ่มลูกค้าที่มีพฤติกรรมบริโภคสินค้าอย่างต่อเนื่อง แต่ไม่ได้มีการทำธุรกรรมในรูปแบบที่สามารถวัดค่าได้ง่าย นอกจากนี้ การนำข้อมูลที่ได้จากแบบจำลอง RFM ไปใช้ในองค์กรที่มีจำนวนลูกค้ามาก อาจต้องการเทคนิคการวิเคราะห์ขั้นสูงเพิ่มเติม เช่น การใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสนับสนุนการวิเคราะห์ข้อมูลอย่างแม่นยำมากขึ้น ซึ่งงานวิจัยโดย Lee และ Jiang (2021) ได้แสดงให้เห็นถึงศักยภาพของการผสานแบบจำลอง RFM เข้ากับอัลกอริทึมการเรียนรู้ของเครื่องในการจำแนกลูกค้าอย่างแม่นยำมากยิ่งขึ้นในบริบทของธุรกิจออนไลน์

2.1.4 **ทฤษฎีเกี่ยวกับอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูล** (Clustering)

การแบ่งกลุ่มข้อมูลเป็นหนึ่งในเทคนิคสำคัญในเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน โดยมุ่งเน้นการค้นหาความสัมพันธ์หรือรูปแบบในข้อมูลโดยไม่มีตัวแปร การแบ่งกลุ่มข้อมูลจะช่วยให้ผู้วิเคราะห์สามารถค้นพบโครงสร้างที่ซ่อนอยู่ในข้อมูลขนาดใหญ่ โดยเป้าหมายหลัก คือ การจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน และแยกกลุ่มข้อมูลที่มีลักษณะแตกต่างออกจากกัน

ในกระบวนการนี้ ข้อมูลที่อยู่ในกลุ่มเดียวกันหรือคลัสเตอร์ (Cluster) จะมีความคล้ายคลึงกันของจุดข้อมูลภายในกลุ่ม (Intra-cluster Similarity) สูง และความแตกต่างระหว่างกลุ่ม (Inter-cluster dissimilarity) สูงที่สุด การแบ่งกลุ่มข้อมูลมีบทบาทสำคัญในหลากหลายสาขา เช่น การวิเคราะห์พฤติกรรมลูกค้า การจัดการระบบชีวสารสนเทศ การทำเหมืองข้อมูล และระบบแนะนำสินค้า

ประเภทของการแบ่งกลุ่มข้อมูล

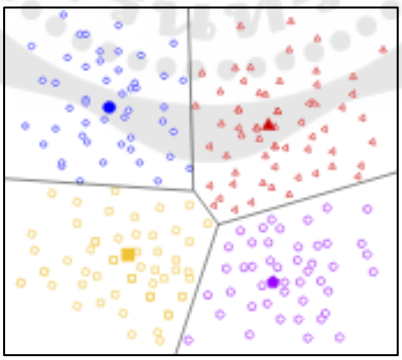
การแบ่งกลุ่มข้อมูลเป็นกระบวนการสำคัญในงานวิเคราะห์ข้อมูล โดยมุ่งเน้นการจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน โดยที่ข้อมูลภายในกลุ่มจะมีลักษณะ

ใกล้เคียงกันมากที่สุด ในขณะที่ข้อมูลระหว่างกลุ่มจะมีความแตกต่างกันอย่างชัดเจน โดยการแบ่งกลุ่มสามารถแบ่งออกเป็น 4 ประเภทหลัก ตามลักษณะการทำงานและวิธีการประมวลผล ดังนี้

1. **การแบ่งกลุ่มข้อมูลเกิดจากจุดข้อมูล** (Centroid-based Clustering)

เป็นการแบ่งกลุ่มข้อมูลโดยใช้จุดศูนย์กลาง (Centroid) เป็นตัวแทนของกลุ่ม ซึ่งจุดศูนย์กลางนี้จะถูกคำนวณมาจากค่าเฉลี่ยของจุดข้อมูลทั้งหมดภายในกลุ่ม การแบ่งกลุ่มประเภทนี้จะจัดกลุ่มข้อมูลเป็นแบบไม่ซ้อนทับ (Non-Hierarchical Clusters) โดยเน้นการลดความแปรปรวนภายในกลุ่มให้ต่ำที่สุด ทำให้ข้อมูลในแต่ละกลุ่มมีความใกล้เคียงกันมากที่สุด ซึ่ง Xu และ Tian (2015) อธิบายว่า การใช้จุดศูนย์กลางในการแบ่งกลุ่มข้อมูลช่วยลดความแปรปรวนภายในกลุ่มได้อย่างมีประสิทธิภาพ โดยการจัดกลุ่มตามค่าเฉลี่ยของจุดข้อมูลนั้นจะทำให้ข้อมูลในกลุ่มมีความใกล้เคียงกันมากที่สุด ดังแสดงในภาพประกอบ 2

ตัวอย่างอัลกอริทึมที่เด่นชัดของการแบ่งกลุ่มข้อมูลเกิดจากจุดข้อมูล คือ K-Means Clustering ซึ่งเป็นอัลกอริทึมที่นิยมใช้กันอย่างแพร่หลาย โดยผู้ใช้ต้องกำหนดกลุ่มล่วงหน้า จากนั้นอัลกอริทึมจะทำการสุ่มจุดศูนย์กลางเริ่มต้น และปรับปรุงตำแหน่งของจุดศูนย์กลางซ้ำๆ จนกว่ากลุ่มจะคงที่ จุดเด่นของวิธีนี้คือมีความรวดเร็วและเหมาะกับข้อมูลที่มีโครงสร้างชัดเจนและขนาดกลุ่มใกล้เคียงกัน อย่างไรก็ตาม การแบ่งกลุ่มข้อมูลจากจุดข้อมูลมีข้อจำกัด คือ อ่อนไหวต่อค่าเริ่มต้น และค่าข้อมูลที่มีความผิดปกติหรือ Outlier ซึ่งอาจทำให้ผลลัพธ์ที่ได้ไม่สมเหตุสมผล

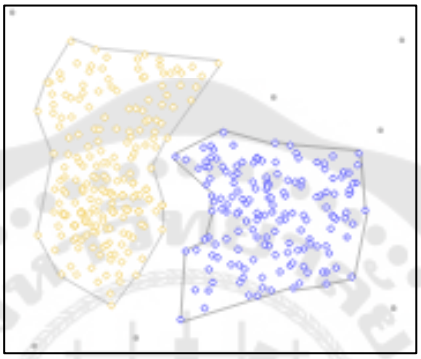


ภาพประกอบ 1 แสดงภาพตัวอย่างของการแบ่งกลุ่มข้อมูลเกิดจากจุดข้อมูล

ที่มา: (https://developers.google.com/machine-learning/clustering/clustering-algorithms)

2. การแบ่งกลุ่มของข้อมูลเกิดจากการกระจุกตัวของจุดข้อมูล (Density-based Clustering)

Xu และ Tian (2015) อธิบายว่า การแบ่งกลุ่มข้อมูลที่เกิดจากการกระจุกตัวของจุดข้อมูลจะคำนึงถึงพื้นที่ที่มีความหนาแน่นสูงเป็นหลัก โดยจะแบ่งกลุ่มข้อมูลที่อยู่ในพื้นที่หนาแน่นเดียวกันเข้าด้วยกันและกำหนดข้อมูลที่อยู่นอกพื้นที่หนาแน่นเป็นค่าข้อมูลที่มีความผิดปกติ (Outlier) ดังแสดงในภาพประกอบ 3



ภาพประกอบ 2 แสดงภาพตัวอย่างของการแบ่งกลุ่มของข้อมูลเกิดจากการกระจุกตัวของจุดข้อมูล

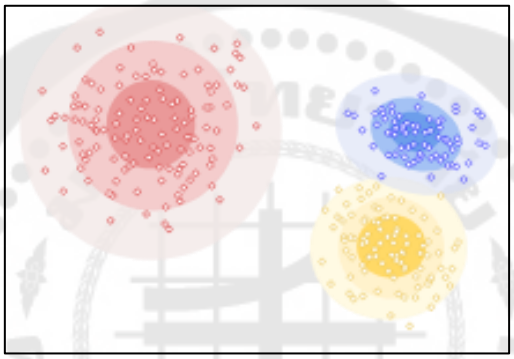
ที่มา: (https://developers.google.com/machine-learning/clustering/clustering-algorithms)

อัลกอริทึมที่โดดเด่นสำหรับการแบ่งกลุ่มของข้อมูลเกิดจากการกระจุกตัวของจุดข้อมูล ได้แก่ Density-Based Spatial Clustering of Applications with Noise (DBSCAN) ซึ่งทำงานโดยใช้ค่าพารามิเตอร์รัศมีของความหนาแน่นหรือ Epsilon และจำนวนจุดขั้นต่ำที่กำหนดว่าพื้นที่นี้มีความหนาแน่นเพียงพอหรือ MinPts อัลกอริทึม DBSCAN สามารถจัดการกับข้อมูลที่มีรูปทรงซับซ้อนและมี Noise ได้ดี อย่างไรก็ตาม อัลกอริทึมนี้มีข้อจำกัดในกรณีที่ข้อมูลมีความหนาแน่นไม่สม่ำเสมอหรือมีจำนวนมิติสูง เพราะการกำหนดค่าพารามิเตอร์ที่เหมาะสมทำได้ยาก

3. การแบ่งกลุ่มตามการกระจายตัวของข้อมูล (Distribution-based Clustering)

การแบ่งกลุ่มตามการกระจายตัวของข้อมูลใช้หลักการทางสถิติ โดยมีสมมติฐานว่า ข้อมูลในแต่ละกลุ่มสามารถอธิบายด้วยฟังก์ชันความหนาแน่น เช่น Gaussian Distribution โดยแต่ละกลุ่มจะมีศูนย์กลางและการกระจายตัวที่กำหนดไว้ โดยความน่าจะเป็นของข้อมูลแต่ละจุดจะลดลงตามระยะห่างจากศูนย์กลาง ดังแสดงในภาพประกอบ 4

Xu และ Tian (2015) อธิบายว่า อัลกอริทึมที่มีการแบ่งกลุ่มตามการกระจายตัวของข้อมูล เช่น Gaussian Mixture Model ใช้สมมติฐานของการกระจายตัวของข้อมูลแบบ Gaussian เพื่อระบุพารามิเตอร์ที่สามารถอธิบายข้อมูลที่กระจายตัวในแต่ละกลุ่มได้อย่างแม่นยำ ซึ่งใช้อัลกอริทึม Expectation-Maximization (EM) ในการประมาณพารามิเตอร์ของแต่ละคลัสเตอร์ เช่น ค่าความน่าจะเป็นที่จุดข้อมูลจะอยู่ในคลัสเตอร์ใดๆ วิธีนี้มีความยืดหยุ่นสูงและสามารถอธิบายข้อมูลที่มีการกระจายตัวหลากหลายได้ดี อย่างไรก็ตาม Distribution-based Clustering มีข้อจำกัดตรงที่ต้องกำหนดจำนวนคลัสเตอร์ล่วงหน้า และสมมติฐานการแจกแจงข้อมูลต้องมีความถูกต้อง ซึ่งในบางครั้งอาจไม่สามารถนำไปใช้กับข้อมูลที่มีการแจกแจงไม่ชัดเจน



ภาพประกอบ 3 แสดงภาพตัวอย่างของการแบ่งกลุ่มตามการกระจายตัวของข้อมูล

ที่มา: (https://developers.google.com/machine-learning/clustering/clustering-algorithms)

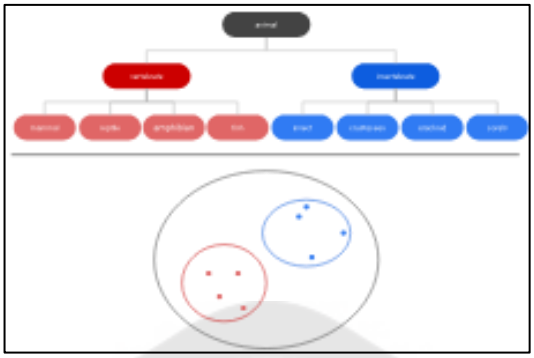
4. การแบ่งกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

การแบ่งกลุ่มแบบลำดับชั้นเป็นวิธีการแบ่งกลุ่มข้อมูลในลักษณะลำดับชั้น (Hierarchy) โดยแสดงผลลัพธ์ในรูปแบบโครงสร้างต้นไม้ หรือที่เรียกว่า Dendrogram ซึ่งแสดงถึงลำดับการรวมกลุ่มหรือแยกกลุ่มของข้อมูล ดังแสดงในภาพประกอบ 5

การแบ่งกลุ่มแบบลำดับชั้น สามารถแบ่งออกเป็น 2 ประเภทหลัก ได้แก่

1. Agglomerative Clustering (Bottom-Up): เริ่มต้นจากแต่ละจุดเป็นกลุ่มเดี่ยว จากนั้นจะทำการรวมกลุ่มที่ใกล้กันที่สุดไปเรื่อยๆ จนเหลือเพียงกลุ่มดียว

2. Divisive Clustering (Top-Down): เริ่มจากกลุ่มเดียวที่รวมข้อมูลทั้งหมด แล้วทำการแยกออกเป็นกลุ่มย่อยๆ ตามความแตกต่างของข้อมูล



ภาพประกอบ 4 แสดงภาพตัวอย่างของการจัดกลุ่มแบบลำดับชั้น

ที่มา: (https://developers.google.com/machine-learning/clustering/clustering-algorithms)

Xu และ Tian (2015) อธิบายว่า การแบ่งกลุ่มแบบลำดับชั้นสามารถจัดกลุ่มข้อมูลในลักษณะของต้นไม้ (Dendrogram) ซึ่งช่วยให้เข้าใจความสัมพันธ์ของข้อมูลในระดับต่างๆ ได้อย่างชัดเจน และไม่จำเป็นต้องกำหนดจำนวนกลุ่มล่วงหน้า ซึ่งเป็นข้อดีของการจัดกลุ่มแบบลำดับชั้น อย่างไรก็ตาม การแบ่งกลุ่มประเภทนี้ ก็มีข้อจำกัดในด้านความซับซ้อนเชิงเวลา เนื่องจากต้องคำนวณระยะห่างระหว่างข้อมูลทุกจุด ซึ่งทำให้ประมวลผลช้าในข้อมูลขนาดใหญ่

อัลกอริทึมในการแบ่งกลุ่มข้อมูล

1. K-Means Clustering

หลักการพื้นฐานของอัลกอริทึม K-Means Clustering คือ การแบ่งข้อมูล n จุดให้อยู่ใน k กลุ่ม หรือคลัสเตอร์ที่กำหนดล่วงหน้า โดยการค้นหาจุดศูนย์กลางที่เหมาะสมสำหรับแต่ละคลัสเตอร์ อัลกอริทึมนี้จะพยายามลดความแปรปรวนภายในกลุ่ม (Within-Cluster Variance) ให้น้อยที่สุด กระบวนการนี้จะอาศัยการคำนวณระยะห่างระหว่างจุดข้อมูลกับจุดศูนย์กลาง และปรับตำแหน่งของจุดศูนย์กลางใหม่ในแต่ละรอบจนกว่าการจัดกลุ่มไม่เปลี่ยนแปลง (Convergence)

MacQueen (1967) ซึ่งเป็นผู้ริเริ่มอัลกอริทึม K-Means ได้อธิบายว่า อัลกอริทึมนี้สามารถใช้ในการจัดกลุ่มข้อมูลหลายมิติ โดยการหาจุดศูนย์กลางที่เหมาะสมที่สุดเพื่อแบ่งข้อมูลออกเป็นกลุ่มที่มีลักษณะคล้ายคลึงกันมากที่สุด ผ่านการคำนวณระยะห่างระหว่างจุด

ข้อมูลแต่ละจุดกับจุดศูนย์กลางของแต่ละคลัสเตอร์ ซึ่งจะทำให้กลุ่มข้อมูลในแต่ละคลัสเตอร์มีความแตกต่างน้อยที่สุดจากจุดศูนย์กลางของกลุ่ม

อัลกอริทึม K-Means Clustering ทำงานได้ดีเมื่อข้อมูลมีโครงสร้างชัดเจนและสามารถแบ่งกลุ่มได้ตามระยะห่างทางเรขาคณิต เช่น กลุ่มข้อมูลที่กระจายตัวรอบจุดศูนย์กลางในรูปวงกลมหรือทรงกลม อย่างไรก็ตาม อัลกอริทึมนี้ไม่สามารถจับความสัมพันธ์เชิงซับซ้อนหรือรูปทรงที่ไม่เป็นทรงเรขาคณิตได้ เช่น รูปทรงโค้ง หรือข้อมูลที่มีลักษณะทับซ้อนกัน (Xu และ Tian, 2015)

ขั้นตอนการแบ่งกลุ่มข้อมูล

1. การกำหนดจำนวนคลัสเตอร์ที่ต้องการ: เลือกจำนวนคลัสเตอร์ที่ต้องการแบ่งข้อมูลออกเป็นจำนวน k กลุ่ม โดยค่าที่กำหนด k ควรพิจารณาจากวิธีต่างๆ เช่น Elbow Method หรือ Silhouette Score เพื่อกำหนดจำนวนคลัสเตอร์ที่เหมาะสม

2. การเลือกจุดศูนย์กลางเริ่มต้นแบบสุ่ม: เลือกจุดศูนย์กลางเริ่มต้นแบบสุ่ม k จุดจากข้อมูล หรือใช้วิธีอย่าง K-Means++ เพื่อกระจายจุดศูนย์กลางอย่างเหมาะสม

3. การจัดกลุ่มข้อมูลแต่ละจุดเข้ากับคลัสเตอร์ที่ใกล้กับจุดศูนย์กลางที่สุด: คำนวณระยะทางแบบยูคลิด (Euclidean Distance) จากจุดข้อมูล $\boldsymbol{x_i}$ ไปยังจุดศูนย์กลาง $\boldsymbol{c_k}$ โดยสามารถคำนวณได้จากสมการ 2.10

$$d(x_i, c_k) = \| x_i - c_k \| \tag{2.10}$$

โดยที่

x_i คือ จุดข้อมูลตัวที่ i ในข้อมูล

c_k คือ จุด Centroid ของคลัสเตอร์ k

4. ปรับปรุงจุดศูนย์กลางของแต่ละคลัสเตอร์ใหม่: คำนวณจุดศูนย์กลางใหม่ ($\boldsymbol{c_k}$) สำหรับคลัสเตอร์ $\boldsymbol{C_k}$ โดยใช้ค่าเฉลี่ยของจุดข้อมูลทั้งหมดในคลัสเตอร์นั้น ดังแสดงในสมการ 2.11

$$c_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{2.11}$$

โดยที่

c_k คือ จุดศูนย์กลางของคลัสเตอร์ k หลังการอัปเดต

$|c_k|$ คือ จำนวนจุดข้อมูลในคลัสเตอร์ C_k

x_i คือ จุดข้อมูลที่อยู่ในคลัสเตอร์ C_k

5. ดำเนินการตามขั้นตอนที่ 3 และ 4 อีกครั้ง จนกว่าจุดศูนย์กลางของกลุ่มจะไม่เปลี่ยนแปลง (Convergence)

อัลกอริทึม K-Means Clustering มีความเรียบง่ายและรวดเร็ว เหมาะสำหรับข้อมูลจำนวนมากที่มีโครงสร้างที่ชัดเจน สามารถนำไปใช้งานในหลากหลายบริบท เช่น การวิเคราะห์ลูกค้า การจัดการเอกสาร หรือการทำเหมืองข้อมูล แต่มีข้อจำกัด เนื่องจากต้องกำหนดจำนวนกลุ่มล่วงหน้า และมีความอ่อนไหวต่อจุดเริ่มต้น หากจุดเริ่มต้นไม่ดี อาจทำให้ได้ผลลัพธ์ที่ไม่เหมาะสม นอกจากนี้ อัลกอริทึม K-Means Clustering ยังไม่เหมาะสำหรับข้อมูลที่มีรูปทรงไม่ชัดเจนหรือมีค่า Noise และค่า Outliers เนื่องจากอาจทำให้การคำนวณจุดศูนย์กลางเบี่ยงเบน

2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN เป็นอัลกอริทึมที่ใช้ความหนาแน่นของข้อมูลในพื้นที่ใกล้เคียงเป็นเกณฑ์ในการสร้างกลุ่ม Ester และคณะ (1996) ซึ่งเป็นผู้ริเริ่มอัลกอริทึม DBSCAN อธิบายว่าอัลกอริทึมนี้สามารถค้นหากลุ่มข้อมูลที่มีความหนาแน่นสูงในพื้นที่ที่ไม่จำเป็นต้องมีการกำหนดจำนวนคลัสเตอร์ล่วงหน้า โดยการแยกจุดข้อมูลที่ไม่เข้ากับกลุ่มออกจากการวิเคราะห์ ซึ่งต่างจากอัลกอริทึมประเภทอื่นๆ เช่น K-Means Clustering ที่ต้องกำหนดจำนวนกลุ่มก่อน โดย Ester และคณะ (1996) ได้แสดงให้เห็นว่า DBSCAN ใช้หลักการของการกำหนด Core Points คือ จุดข้อมูลที่มีจำนวนจุดอื่นๆ อยู่ภายในระยะ Epsilon อย่างน้อย MinPts เพื่อทำการขยายกลุ่มจากจุดเหล่านี้ โดยอัลกอริทึม DBSCAN จะเน้นการค้นหากลุ่มข้อมูลที่มีความหนาแน่นสูง และแยกข้อมูลที่ไม่เป็นส่วนหนึ่งของกลุ่มใดหรือค่า Noise ออกไป ซึ่งช่วยให้อัลกอริทึมสามารถจัดการกับข้อมูลที่มีลักษณะซับซ้อน เช่น รูปทรงไม่สม่ำเสมอ หรือมีจุดที่เป็นค่า Outliers ได้

หลักการสำคัญของอัลกอริทึม DBSCAN คือ การนิยาม Core Point ซึ่งเป็นจุดข้อมูลที่มีจำนวนจุดอื่นๆ อยู่ในระยะทาง Epsilon อย่างน้อย MinPts และขยายคลัสเตอร์จากจุดเหล่านี้ กระบวนการนี้ทำให้ DBSCAN เหมาะกับข้อมูลที่มีความหลากหลายทางโครงสร้าง เช่น ข้อมูลทางชีววิทยา ภาพถ่ายดาวเทียม หรือข้อมูลเชิงพื้นที่

นิยามของจุดข้อมูล

1. Core Point คือ จุดข้อมูล p ที่มีจุดอื่นๆ อยู่ภายในระยะ Epsilon อย่างน้อย MinPts จุด

2. Border Point คือ จุดข้อมูล p ที่อยู่ในระยะ Epsilon ของ Core Point แต่มีจำนวนจุดในพื้นที่ Epsilon น้อยกว่า MinPts

3. Noise หรือ Outlier คือ จุดข้อมูลที่ไม่เป็นทั้ง Core Point และ Border Point

การคำนวณระยะห่างระหว่างจุดข้อมูล

จะทำการคำนวณระยะห่างระหว่างจุดข้อมูลเพื่อระบุว่า จุดข้อมูลใดเป็นเพื่อนบ้าน (Neighbor) ของจุดข้อมูล p ภายใต้ค่า Epsilon ที่กำหนด ซึ่งคำนวณได้จากสมการ 2.12

$$N_{\varepsilon}(p) = \{q \in D | \left|\left|p - q\right|\right| \leq \varepsilon\} \quad (2.12)$$

โดยที่

$N_{\varepsilon}(p)$ คือ เซตของจุดข้อมูล q ที่อยู่ในพื้นที่ใกล้เคียงของจุด p

q คือ จุดข้อมูลใดๆ ในชุดข้อมูล D

D คือ ชุดของจุดข้อมูลทั้งหมด

ε คือ ค่า Epsilon เป็นพารามิเตอร์ที่กำหนดรัศมีรอบจุด p เพื่อค้นหาจุดข้อมูลที่อยู่ในพื้นที่ใกล้เคียง

$\left|\left|p - q\right|\right|$ คือ ระยะห่างระหว่างจุด p และจุด q มักคำนวณระยะทางแบบยูคลิด (Euclidean Distance) ซึ่งสามารถคำนวณได้จากสมการที่ 2.13

$$\left|\left|p - q\right|\right| = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \quad (2.13)$$

โดยที่

x_p คือ ค่าพิกัดในแกน x ของจุด p

y_p คือ ค่าพิกัดในแกน y ของจุด p

x_q คือ ค่าพิกัดในแกน x ของจุด q

y_q คือ ค่าพิกัดในแกน y ของจุด q

ขั้นตอนการแบ่งกลุ่มข้อมูล

1. กำหนดค่าพารามิเตอร์ Epsilon และ MinPts

2. เลือกจุดเริ่มต้นจากข้อมูลใดๆ ที่ยังไม่ถูกตรวจสอบ และทำการตรวจสอบว่าจุดนั้นเป็น Core Point หรือไม่ โดยคำนวณจำนวนจุดในระยะ Epsilon

3. หากจุดเริ่มต้นเป็น Core Point ให้สร้างกลุ่มใหม่ โดยเพิ่มจุดข้อมูลทั้งหมดที่อยู่ในระยะ Epsilon ของ Core Point ลงในกลุ่ม หากพบจุดข้อมูลใหม่ที่เป็น Core Point ให้เพิ่มจุดข้อมูลในระยะ Epsilon ของจุดนั้นลงในกลุ่ม

4. ทำซ้ำจนกว่าข้อมูลทั้งหมดจะถูกตรวจสอบ หากไม่สามารถขยายกลุ่มได้อีก ให้กำหนดจุดข้อมูลที่เหลือเป็น Noise หรือเริ่มกลุ่มใหม่

อัลกอริทึม DBSCAN สามารถจัดการข้อมูลที่มีรูปทรงซับซ้อนหรือมีค่า Noise ได้ดี รวมถึงไม่จำเป็นต้องกำหนดจำนวนคลัสเตอร์ล่วงหน้า ซึ่งการตั้งค่าพารามิเตอร์ Epsilon และ MinPts มีผลอย่างมากต่อผลลัพธ์ หากข้อมูลมีความหนาแน่นแตกต่างกัน DBSCAN อาจไม่สามารถจัดกลุ่มได้ดี แต่หากการคำนวณระยะห่างระหว่างจุดข้อมูลมีจำนวนมาก อาจใช้ทรัพยากรมากเมื่อข้อมูลมีขนาดใหญ่

3. Gaussian Mixture Model (GMM)

อัลกอริทึม GMM ใช้สมมติฐานว่า ข้อมูลในแต่ละกลุ่มถูกสร้างขึ้นจากการแจกแจงแบบปกติ (Gaussian Distribution) หลายตัว โดยข้อมูลในคลัสเตอร์หนึ่ง จะมีการกระจายที่กำหนดด้วยค่าเฉลี่ยและความแปรปรวน Dempster และคณะ (1977) อธิบายว่า การใช้กระบวนการ EM (Expectation-Maximization) ทำให้สามารถประมาณค่าพารามิเตอร์ของการแจกแจง Gaussian แต่ละตัวจากข้อมูลที่ขาดหายไปหรือไม่สมบูรณ์ได้ ซึ่งการปรับค่าพารามิเตอร์จะทำอย่างต่อเนื่องในแต่ละรอบ จนกระทั่งค่า Log-Likelihood ของข้อมูลไม่เปลี่ยนแปลงหรือลดลงต่ำกว่าค่าที่กำหนด

ความพิเศษของอัลกอริทึม GMM คือ ความสามารถในการจัดการกับข้อมูลที่ซ้อนทับกัน โดยใช้ความน่าจะเป็นในการกำหนดคลัสเตอร์ที่ข้อมูลควรอยู่ โดย Xu และ Tian (2015) อธิบายว่าอัลกอริทึม GMM สามารถจัดการกับข้อมูลที่มีโครงสร้างซับซ้อน และสามารถจำแนกรูปร่างคลัสเตอร์ที่ไม่เป็นแบบกลม (Non-Spherical) ได้ดีกว่าอัลกอริทึม K-Means ซึ่งจำเป็นต้องใช้การจำแนกรูปร่างกลมที่สมบูรณ์

ขั้นตอนการแบ่งกลุ่มข้อมูล

1. กำหนดค่าพารามิเตอร์เริ่มต้น: กำหนดจำนวนคลัสเตอร์และค่าพารามิเตอร์สำหรับเริ่มต้น

- μ_k คือ เวกเตอร์ค่าเฉลี่ยของแต่ละคลัสเตอร์ (ตำแหน่งศูนย์กลาง)
- $\sum_k$ คือ เมทริกซ์ความแปรปรวนร่วมของแต่ละคลัสเตอร์ (รูปร่างและขนาด)
- w_k คือ น้ำหนักของคลัสเตอร์แต่ละคลัสเตอร์ ซึ่งแสดงถึงสัดส่วนของจุดข้อมูลในแต่ละคลัสเตอร์

2. Expectation-Step (E-Step):

- คำนวณความน่าจะเป็นที่จุดข้อมูล x_i มาจากคลัสเตอร์ k ดังแสดงในสมการ 2.14

$$P(x_i|\mu_k,\sum_k) = \frac{1}{(2\pi^{d/2}|\sum_k|^{1/2}} exp(-\frac{1}{2}(x_i-\mu_i)^T\sum_k{}^{-1}(x_i-\mu_i)) \quad (2.14)$$

โดยที่

x_i คือ จุดข้อมูลตัวที่ i ในข้อมูล

μ_k คือ เวกเตอร์ค่าเฉลี่ยของคลัสเตอร์ k

d คือ จำนวนมิติของข้อมูล

$(x_i - \mu_i)$ คือ เวกเตอร์ผลต่างระหว่างจุดข้อมูล x_i และค่าเฉลี่ย μ_i

- คำนวณ Responsibility (r_{ik}): เป็นความน่าจะเป็นแบบมีเงื่อนไขที่จุด x_i เป็นสมาชิกของคลัสเตอร์ k ดังแสดงในสมการ 2.15

$$r_{ik} = \frac{w_k P(x_i|\mu_k,\sum_k)}{\sum_{j=1}^{K} w_j P(x_i|\mu_k,\sum_j)} \quad (2.15)$$

โดยที่

r_{ik} คือ ค่าความน่าจะเป็นแบบมีเงื่อนไขที่แสดงว่า x_i มาจากคลัสเตอร์ k มีช่วงระหว่าง [0,1]

w_k คือ น้ำหนักของคลัสเตอร์ k

$P(x_i|\mu_k,\sum_k)$ คือ ความน่าจะเป็นของ x_i ภายใต้การแจกแจงแบบปกติ (*Gaussian Distribution)* ของคลัสเตอร์ k

μ_k คือ ค่าเฉลี่ยของคลัสเตอร์ k

x_i คือ จุดข้อมูลตัวที่ i ในข้อมูล

K คือ จำนวนคลัสเตอร์ทั้งหมด

3. Maximization-Step (M-Step): ปรับปรุงค่าพารามิเตอร์ของแต่ละคลัสเตอร์โดยใช้ค่า r_{ik}:

1. ปรับปรุงค่าเฉลี่ย (μ_k)

เพื่อคำนวณตำแหน่งศูนย์กลางของคลัสเตอร์ใหม่ ดังแสดงในสมการ 2.16

$$\mu_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}} \tag{2.16}$$

2. ปรับปรุงเมทริกซ์ความแปรปรวนร่วม ($\sum_k$)

เพื่อคำนวณความแปรปรวนและรูปร่างใหม่ของคลัสเตอร์ ดังแสดงในสมการ 2.17

$$\sum_k = \frac{\sum_{i=1}^{N} r_{ik}(x_i-\mu_i)(x_i-\mu_i)^T}{\sum_{i=1}^{N} r_{ik}} \tag{2.17}$$

3. ปรับปรุงน้ำหนัก (w_k)

เพื่อคำนวณสัดส่วนของจุดข้อมูลในแต่ละคลัสเตอร์ ดังแสดงในสมการ 2.18

$$w_k = \frac{\sum_{i=1}^{N} r_{ik}}{N} \tag{2.18}$$

4. ตรวจสอบเงื่อนไขการหยุด (Convergence): คำนวณค่า Log-Likelihood ของแบบจำลอง โดยจะหยุดกระบวนการหากค่า Log-Likelihood เปลี่ยนแปลงน้อยกว่าเกณฑ์ที่กำหนดหรือจำนวน Iteration ถึงค่าที่ตั้งไว้ล่วงหน้า ดังแสดงในสมการ 2.19

$$L = \sum_{i=1}^{N} log\ (\sum_{k=1}^{K} w_k P(x_i|\mu_{k,}\sum_k)) \quad (2.19)$$

โดยที่

L คือ ค่า Log-Likelihood ของข้อมูลทั้งหมดในอัลกอริทึม

w_k คือ น้ำหนักของคลัสเตอร์หรือสัดส่วนของคลัสเตอร์ k ในข้อมูลทั้งหมด.

N คือ จำนวนจุดข้อมูลทั้งหมดในชุดข้อมูล

K คือ จำนวนคลัสเตอร์ในแบบจำลอง

$\sum_{k=1}^{K} w_k P(x_i|\mu_{k,}\sum_k)$ คือ ผลรวมของค่าความน่าจะเป็นสำหรับทุกคลัสเตอร์ k ของจุดข้อมูล x_i.

อัลกอริทึม GMM รองรับข้อมูลที่มีการแจกแจงที่ซับซ้อน และสามารถจัดการข้อมูลที่มีรูปทรงหรือขอบเขตของกลุ่มที่ซ้อนทับกันได้ แต่ก็มีข้อจำกัด เนื่องจากต้องกำหนดจำนวนกลุ่มล่วงหน้า และอ่อนไหวต่อค่าเริ่มต้น อาจเกิดการ Overfitting หากข้อมูลมีค่า Noise สูง หรือไม่ได้กระจายตัวตาม Gaussian Distribution อาจทำให้แบบจำลองซับซ้อนเกินความจำเป็น

4. Agglomerative Clustering

Agglomerative Clustering เป็นอัลกอริทึมที่ใช้แนวคิดแบบลำดับชั้น โดยเริ่มต้นจากการพิจารณาจุดข้อมูลแต่ละจุดเป็นคลัสเตอร์เดี่ยว จากนั้นจึงทำการรวมกลุ่มจุดข้อมูลที่ใกล้เคียงกันที่สุดในแต่ละขั้นตอนจนกว่าจะได้จำนวนคลัสเตอร์ที่ต้องการ Johnson (1967) อธิบายว่า Agglomerative Clustering เป็นกระบวนการที่เริ่มจากการคำนวณระยะห่างระหว่างจุดข้อมูลทั้งหมดในชุดข้อมูล จากนั้นทำการรวมกลุ่มจุดข้อมูลที่ใกล้เคียงที่สุดในแต่ละขั้นตอนจนกว่าจะได้จำนวนคลัสเตอร์ที่ต้องการ อัลกอริทึมนี้ให้ความสำคัญกับโครงสร้างลำดับชั้นของข้อมูล ทำให้ผู้วิเคราะห์สามารถเข้าใจความสัมพันธ์เชิงลำดับชั้นได้ลึกซึ้งยิ่งขึ้น

อัลกอริทึม Agglomerative Clustering เหมาะสำหรับข้อมูลที่มีโครงสร้างซ้อนกันหลายระดับ เช่น ต้นไม้สปีชีส์ของสิ่งมีชีวิต หรือโครงสร้างองค์กร อย่างไรก็ตาม วิธีนี้มีความซับซ้อนทางคำนวณค่อนข้างสูงเมื่อจำนวนข้อมูลเพิ่มขึ้น เนื่องจากต้องคำนวณระยะทางระหว่างจุดข้อมูลหลายคู่ในทุกขั้นตอน

การคำนวณระยะห่างระหว่างกลุ่ม

- Single Linkage คือ ระยะห่างระหว่างคลัสเตอร์ C_i และ C_j ซึ่งจะใช้ระยะห่างที่สั้นที่สุดระหว่างจุดข้อมูลใน C_i และ C_j ดังแสดงในสมการ 2.20

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} ||x - y|| \quad (2.20)$$

- Complete Linkage คือ ระยะห่างระหว่างคลัสเตอร์ C_i และ C_j ซึ่งจะใช้ระยะห่างที่ยาวที่สุดระหว่างจุดข้อมูลใน C_i และ C_j ดังแสดงในสมการ 2.21

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} ||x - y|| \quad (2.21)$$

- Average Linkage ระยะห่างระหว่างคลัสเตอร์ C_i และ C_j ซึ่งจะใช้ค่าเฉลี่ยของระยะห่างระหว่างจุดข้อมูลใน C_i และ C_j ดังแสดงในสมการ 2.22

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} ||x - y|| \quad (2.22)$$

ขั้นตอนการแบ่งกลุ่มข้อมูล

1. เริ่มต้นด้วยจุดข้อมูลทั้งหมดในคลัสเตอร์แยกกัน (แต่ละจุดเป็นคลัสเตอร์ของตัวเอง)

2. คำนวณระยะห่างระหว่างคลัสเตอร์

3. ปรับปรุงคลัสเตอร์: รวมคลัสเตอร์ C_i และ C_j ตามประเภทการคำนวณระยะห่างระหว่างคลัสเตอร์ และปรับปรุงระยะห่างระหว่างคลัสเตอร์ทั้งหมด.

4. ดำเนินการซ้ำจนกว่าคลัสเตอร์จะลดเหลือจำนวนที่ต้องการ: ดำเนินการซ้ำในขั้นตอนที่ 2 และ 3 จนกว่าคลัสเตอร์ทั้งหมดจะเหลือเท่ากับจำนวนที่กำหนด

ข้อได้เปรียบของอัลกอริทึม Agglomerative Clustering คือ การที่ไม่ต้องกำหนดจำนวนคลัสเตอร์ล่วงหน้า เนื่องจากสามารถเลือกจุดหยุดที่เหมาะสมจาก Dendrogram ได้ เหมาะสำหรับข้อมูลที่มีโครงสร้างแบบลำดับชั้น เช่น ต้นไม้ตระกูล หรือความสัมพันธ์ระหว่างกลุ่ม และให้ผลลัพธ์ที่แสดงภาพรวมของข้อมูลในทุกระดับของการรวมกลุ่ม ทำให้ง่ายต่อการตัดสินใจเลือกจำนวนคลัสเตอร์ที่เหมาะสม แต่ก็มีข้อจำกัด คือ มีการใช้ทรัพยากรสูงเมื่อจำนวน

ข้อมูลเพิ่มขึ้น โดยเฉพาะหน่วยความจำและเวลาในการคำนวณ รวมทั้งอ่อนไหวต่อการกำหนดระยะทางระหว่างคลัสเตอร์ ซึ่งอาจส่งผลต่อคุณภาพของการจัดกลุ่ม

เมทริกซ์ในการวัดประสิทธิภาพอัลกอริทึม

Silhouette Score

เมทริกซ์ Silhouette Score ใช้สำหรับประเมินความเหมาะสมของการแบ่งกลุ่ม โดยจะทำการเปรียบเทียบระยะห่างเฉลี่ยระหว่างจุดข้อมูลภายในกลุ่มเดียวกัน (Intra-cluster Distance) กับระยะห่างเฉลี่ยระหว่างจุดข้อมูลกับกลุ่มที่ใกล้ที่สุด (Inter-cluster Distance) โดยจะมีค่าอยู่ระหว่าง -1 และ 1 (Rousseeuw, 1987) ซึ่งจะมีสูตรการคำนวณดังสมการ 2.23

$$S = \frac{b-a}{max(a,b)} \tag{2.23}$$

โดยที่

S คือ ค่า Silhouette Score สำหรับจุดข้อมูล i

a คือ ระยะเฉลี่ยระหว่างจุดข้อมูล i และจุดข้อมูลอื่นๆ ภายในกลุ่มเดียวกัน ซึ่งคำนวณได้จากสมการ 2.24

$$a = \frac{1}{|C|-1}\sum_{x \in C, x \neq i} ||x - i|| \tag{2.24}$$

โดยที่

$|C|$ คือ จำนวนจุดข้อมูลในกลุ่ม C

x คือ จุดข้อมูลที่อยู่ในกลุ่มเดียวกัน

i คือ จุดข้อมูลที่กำลังคำนวณ Silhouette Score อยู่

b คือ ระยะเฉลี่ยระหว่างจุดข้อมูล i และจุดข้อมูลในกลุ่มที่ใกล้ที่สุด (Nearest Neighbor Cluster) ซึ่งคำนวณได้จากสมการ 2.25

$$b = \min_{C' \neq C} \frac{1}{|C'|}\sum_{y \in C'} ||y - i|| \tag{2.25}$$

โดยที่

$|C'|$ คือ จำนวนจุดข้อมูลในกลุ่ม C'

y คือ จุดข้อมูลที่อยู่ในกลุ่มอื่น

i คือ จุดข้อมูลที่กำลังคำนวณ Silhouette Score อยู่

$max(a, b)$ คือ ค่าที่มากที่สุดระหว่าง a และ b

การวัดประสิทธิภาพของอัลกอริทึม

หลังจากได้ค่า Silhouette Score จากการคำนวณมาแล้ว จะทำการประเมินประสิทธิภาพของอัลกอริทึม โดยจะสามารถอธิบายได้ ดังนี้

- ค่า Silhouette Score ใกล้ 1 หมายถึง การแบ่งกลุ่มได้ดี
- ค่า Silhouette Score ใกล้ 0 หมายถึง ข้อมูลอาจอยู่ใกล้เส้นขอบระหว่างกลุ่ม
- ค่า Silhouette Score ติดลบ หมายถึง ข้อมูลอาจถูกจัดกลุ่มผิดหรือไม่เหมาะสม

สรุปข้อมูลของอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูล

จากได้ทำการศึกษาเกี่ยวกับอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูลทั้ง 4 อัลกอริทึม ซึ่งเป็นอัลกอริทึมที่มีความนิยมในการแบ่งกลุ่มแต่ละประเภท โดยประกอบด้วยอัลกอริทึม K-Means Clustering, อัลกอริทึม DBSCAN, อัลกอริทึม GMM, และอัลกอริทึม Agglomerative Clustering ซึ่งทำให้มีความเข้าใจในแต่ละอัลกอริทึมมากขึ้น ทั้งในเรื่องของวิธีการแบ่งกลุ่มข้อมูล ค่าพารามิเตอร์ จุดเด่น ตลอดจนข้อจำกัดของอัลกอริทึม ผู้วิจัยจึงได้ทำการสรุปข้อมูลของแต่ละอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูลที่ได้ทำการศึกษา และแสดงข้อมูลที่ได้ในตาราง 1

ตาราง 1 สรุปข้อมูลอัลกอริทึมที่ใช้ในการแบ่งกลุ่มข้อมูล (Clustering)

อัลกอริทึม	วิธีการแบ่งกลุ่ม	พารามิเตอร์	จุดเด่น	ข้อจำกัด
K-Means Clustering	ลดระยะห่างรวมของข้อมูลในกลุ่มจากจุดศูนย์กลางของคลัสเตอร์	- k: จำนวนคลัสเตอร์	- ใช้งานง่ายและประมวลผลเร็ว - เหมาะสำหรับข้อมูลที่มีโครงสร้างชัดเจน - ใช้ได้ดีกับข้อมูลที่มีลักษณะทรงกลม	- อ่อนไหวต่อจุดเริ่มต้น - ต้องกำหนดจำนวนคลัสเตอร์ล่วงหน้า - ไม่เหมาะกับข้อมูลที่มีค่า noise หรือรูปทรงไม่ชัดเจน

ตารางที่ 1 (ต่อ)

อัลกอริทึม	วิธีการแบ่งกลุ่ม	พารามิเตอร์	จุดเด่น	ข้อจำกัด
DBSCAN	จัดกลุ่มข้อมูลตามความหนาแน่นโดยแยกจุดข้อมูลเป็น Core Points, Border Points และ Noise/Outlier Points	- Epsilon: ระยะห่างที่กำหนดความหนาแน่น - MinPts: จำนวนจุดขั้นต่ำในบริเวณ Epsilon	- จัดการกับลัสเตอร์ที่มีรูปร่างซับซ้อนได้ดี - ทนทานต่อ Noise และ Outliers - ไม่จำเป็นต้องกำหนดจำนวนคลัสเตอร์	- อ่อนไหวต่อการตั้งค่าพารามิเตอร์ - ใช้งานยากเมื่อข้อมูลมีความหนาแน่นแตกต่างกัน
GMM	แบ่งกลุ่มโดยใช้การแจกแจงแบบ Gaussian หลายตัวเพื่อประเมินความน่าจะเป็นของจุดข้อมูล	- n_component: จำนวนคลัสเตอร์	- รองรับคลัสเตอร์ที่มีรูปร่างซับซ้อน - จัดการกับข้อมูลที่คลัสเตอร์ซ้อนทับกันได้ - ใช้เกณฑ์ความน่าจะเป็นในการตัดสินใจ	- ต้องกำหนดจำนวนคลัสเตอร์ล่วงหน้า - อ่อนไหวต่อค่าเริ่มต้นและค่า Noise - มีความซับซ้อนในการประมวลผล
Agglomerative Clustering	แบ่งกลุ่มแบบลำดับชั้น เริ่มจากจุดข้อมูลเดี่ยว รวมกลุ่มทีละคู่จนได้จำนวนคลัสเตอร์ที่ต้องการ	- ระยะห่างระหว่างกลุ่ม: Single Linkage/ Complete Linkage/ Average Linkage	- ไม่ต้องกำหนด k ล่วงหน้า - แสดงโครงสร้างลำดับชั้นชัดเจน - เหมาะสำหรับข้อมูลลำดับชั้น	- ใช้ทรัพยากรมากเมื่อจำนวนข้อมูลเพิ่มขึ้น - อ่อนไหวต่อการกำหนดระยะทางระหว่างคลัสเตอร์

2.1.5 **ทฤษฎีเกี่ยวกับการเลือกจำนวนกลุ่มที่เหมาะสมด้วยเทคนิค** Elbow Method

เทคนิค Elbow Method เป็นหนึ่งในเทคนิคที่นิยมใช้สำหรับการกำหนดจำนวนกลุ่มที่เหมาะสมในกระบวนการแบ่งกลุ่มข้อมูล โดยเฉพาะอย่างยิ่งสำหรับอัลกอริทึม K-Means Clustering เทคนิคนี้จะช่วยให้ผู้วิเคราะห์สามารถตัดสินใจได้อย่างมีหลักการเกี่ยวกับจำนวนกลุ่มที่ควรเลือกใช้งาน เป้าหมายของเทคนิค Elbow Method คือ การค้นหาจุดที่ความแปรปรวนภายในกลุ่ม (Within-Cluster Sum of Squares, WCSS) ลดลงในอัตราที่ชะลอตัว ซึ่งจุดนี้เรียกว่า Elbow บนกราฟ Thorndike (1953) เป็นผู้บุกเบิกแนวคิดเกี่ยวกับการเลือกจำนวนกลุ่มโดยสังเกตจากจุดเปลี่ยนที่ชัดเจนของตัวชี้วัด เช่น ความแปรปรวนภายในกลุ่ม ซึ่งในเวลาต่อมากลายเป็นแนวทางพื้นฐานของ Elbow Method ได้อธิบายว่า การแยกกลุ่มที่ดีควรส่งผลให้ความคล้ายคลึงกันภายในกลุ่มเพิ่มขึ้น ขณะที่ความแตกต่างระหว่างกลุ่มสูงขึ้นอย่างมีนัยสำคัญ

การแบ่งกลุ่มข้อมูลเป็นกระบวนการสำคัญในเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน ซึ่งจะไม่มีข้อมูลเป้าหมายที่ชัดเจน การเลือกจำนวนกลุ่มที่เหมาะสมเป็นหนึ่งในปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพของแบบจำลองที่ใช้ในการแบ่งกลุ่ม เทคนิค Elbow Method ถูกออกแบบมาเพื่อลดปัญหาจากการกำหนดค่าจำนวนกลุ่มแบบสุ่ม ซึ่งอาจไม่เหมาะสมกับลักษณะของข้อมูลเสมอไป เทคนิคนี้จึงช่วยเพิ่มความน่าเชื่อถือในการตัดสินใจ และเป็นเครื่องมือสำคัญในงานวิจัยเชิงข้อมูล

หลักการพื้นฐานของเทคนิค Elbow Method คือ มุ่งเน้นไปที่การประเมินความเหมาะสมของจำนวนกลุ่ม โดยพิจารณาจากค่า WCSS ซึ่งเป็นค่าความแปรปรวนภายในกลุ่มทั้งหมดจากจุดข้อมูลในกลุ่มไปยังจุดศูนย์กลางของกลุ่มนั้น โดยแสดงการคำนวณดังสมการ 2.9

$$WCSS = \sum_{k=1}^{K} \sum_{x \epsilon C_k} ||x_i - c_k||^2 \quad (2.9)$$

โดยที่:

K คือ จำนวนคลัสเตอร์ทั้งหมด

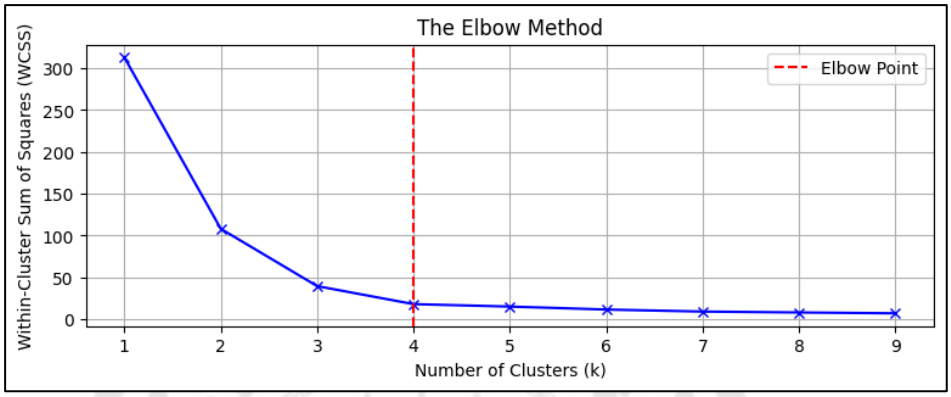
C_k คือ คลัสเตอร์ที่ k

x_i คือ ข้อมูลจุดที่ i

c_k คือ จุดศูนย์กลางของคลัสเตอร์ k

เมื่อจำนวนกลุ่มเพิ่มขึ้น ค่า WCSS จะลดลง เนื่องจากกลุ่มมีขนาดเล็กลงและจุดข้อมูลในแต่ละกลุ่มจะใกล้กับจุดศูนย์กลางมากขึ้น อย่างไรก็ตาม การลดลงของ WCSS จะชะลอ

ตัวเมื่อถึงจำนวนกลุ่มที่เหมาะสม ดังแสดงในภาพประกอบ 1 เนื่องจากการเพิ่มจำนวนกลุ่มที่มากเกินไป อาจไม่ได้ช่วยลดความแปรปรวนในข้อมูลได้อย่างมีนัยสำคัญ จุด Elbow ที่ปรากฏบนกราฟระหว่างค่า k และค่า WCSS คือ จุดที่ค่า WCSS เริ่มลดลงช้าลง ซึ่งสะท้อนถึงจำนวนกลุ่มที่เหมาะสมที่สุด



ภาพประกอบ 5 แสดงการเลือกจำนวนกลุ่มข้อมูลที่เหมาะของเทคนิค Elbow Method

ที่มา (https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans)

กระบวนการของเทคนิค Elbow Method เริ่มต้นจากการคำนวณค่า WCSS สำหรับแต่ละค่า k ที่กำหนดในช่วงที่สนใจ จากนั้นจะสร้างกราฟที่แสดงความสัมพันธ์ระหว่างจำนวนกลุ่ม k และค่า WCSS โดยแกน x แทนจำนวนคลัสเตอร์ และแกน y แทนค่า WCSS การเลือกจำนวนกลุ่มที่เหมาะสมสามารถทำได้โดยมองหาจุด "Elbow" บนกราฟ ซึ่งแสดงถึงจุดที่การลดลงของค่า WCSS มีแนวโน้มชะลอตัว

ขั้นตอนการเลือกจำนวนกลุ่มที่เหมาะสมด้วยเทคนิค Elbow Method

1. เลือกช่วงของจำนวนกลุ่มที่ต้องการวิเคราะห์ โดยกำหนดช่วงที่ต้องการวิเคราะห์ เช่น 1 ถึง 10 สำหรับค่า k

2. ประมวลผลอัลกอริทึม K-Means Clustering สำหรับแต่ละค่า k และคำนวณค่า WCSS

3. สร้างกราฟระหว่างค่า k และค่า WCSS โดยกำหนดแกน x แสดงจำนวนกลุ่มหรือค่า k และแกน y แสดงค่า WCSS

4. ค้นหาจุด Elbow โดยหาจุดที่กราฟมีการลดลงอย่างเด่นชัด ซึ่งสะท้อนถึงจำนวนกลุ่มที่เหมาะสมที่สุด

เทคนิค Elbow Method เป็นเครื่องมือที่มีความเรียบง่ายแต่ทรงพลังในการช่วยผู้วิเคราะห์ตัดสินใจเกี่ยวกับจำนวนกลุ่มที่เหมาะสม เทคนิคนี้ไม่ต้องพึ่งพาข้อมูลอ้างอิง (Ground Truth) และสามารถใช้งานได้อย่างมีประสิทธิภาพกับข้อมูลหลายประเภท ความสามารถในการแสดงผลผ่านกราฟช่วยให้ผู้ใช้งานเข้าใจและเปรียบเทียบผลลัพธ์ได้ง่าย นอกจากนี้ เทคนิค Elbow Method ยังช่วยลดความเสี่ยงของการแบ่งกลุ่มที่ไม่เหมาะสมจากการกำหนดจำนวนคลัสเตอร์แบบสุ่มหรือความรู้สึกส่วนตัวได้

อีกข้อได้เปรียบของ Elbow Method คือ ความยืดหยุ่นในการประยุกต์ใช้กับอัลกอริทึมการแบ่งกลุ่มอื่นที่ใช้แนวคิดคล้าย WCSS และสามารถใช้ร่วมกับเมทริกซ์ประเมินผลภายใน เช่น Silhouette Score เพื่อช่วยสนับสนุนการตัดสินใจเกี่ยวกับจำนวนกลุ่มที่เหมาะสม Bishop และ Nasrabadi (2006) อธิบายว่า Elbow Method เป็นเครื่องมือที่มีความเรียบง่ายแต่มีประสิทธิภาพในการระบุจำนวนกลุ่มที่เหมาะสมที่สุด โดยสามารถนำไปประยุกต์ใช้ได้ในหลากหลายสถานการณ์ ไม่จำกัดเพียงเฉพาะอัลกอริทึม K-Means แต่ยังสามารถใช้ประกอบกับอัลกอริทึมอื่นๆ ได้ นอกจากนี้ Han และคณะ (2011) ยังอธิบายว่า Elbow Method เป็นเครื่องมือที่มีความเรียบง่ายแต่สามารถให้ข้อมูลที่มีประโยชน์สูงสุดเมื่อนำมาใช้ในบริบทต่าง ๆ ที่เกี่ยวข้องกับการทำเหมืองข้อมูลและการประมวลผลข้อมูลขนาดใหญ่

อย่างไรก็ตาม เทคนิค Elbow Method มีข้อจำกัดหลายประการที่ควรพิจารณา ประการแรก คือ จุด Elbow บนกราฟอาจไม่ชัดเจนในบางกรณี โดยเฉพาะเมื่อข้อมูลไม่มีโครงสร้างที่สามารถแบ่งกลุ่มได้ง่าย หรือในกรณีที่การลดลงของ WCSS มีลักษณะต่อเนื่องและไม่มีจุดเปลี่ยนที่เด่นชัด ปัญหานี้อาจทำให้ผู้วิเคราะห์ต้องใช้ดุลยพินิจในการเลือกจำนวนคลัสเตอร์ซึ่งอาจนำไปสู่ความไม่แน่นอนในผลลัพธ์

นอกจากนี้ เทคนิค Elbow Method ยังอ่อนไหวต่อคุณภาพของข้อมูล หากข้อมูลมีลักษณะซับซ้อน มีค่า Noise จำนวนมาก หรือมีการกระจายตัวที่ไม่สมมาตร ค่า WCSS อาจไม่สะท้อนถึงโครงสร้างที่แท้จริงของข้อมูล ส่งผลให้จุด Elbow ที่พบไม่สามารถบ่งชี้จำนวนกลุ่มที่เหมาะสมได้อย่างถูกต้อง อีกทั้ง เทคนิค Elbow Method ยังมุ่งเน้นไปที่ความแปรปรวนภายในกลุ่มเพียงมิติเดียว โดยไม่พิจารณาเมทริกซ์ภายนอก หรือโครงสร้างเชิงคุณภาพอื่นๆ ของข้อมูล

2.1.6 ทฤษฎีเกี่ยวกับอัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis)

การวิเคราะห์ตะกร้าสินค้าเป็นเทคนิคสำคัญในการทำเหมืองข้อมูล (Data Mining) ที่ช่วยค้นหาความสัมพันธ์ระหว่างรายการสินค้าที่มักถูกซื้อร่วมกันในธุรกรรมเดียวกัน การค้นพบ

ความสัมพันธ์เหล่านี้ช่วยสนับสนุนการตัดสินใจทางธุรกิจในหลากหลายมิติ เช่น การเพิ่มยอดขาย การจัดโปรโมชั่น การปรับปรุงการจัดวางสินค้า และการจัดการสินค้าคงคลัง แนวคิดพื้นฐานของการวิเคราะห์นี้คือการใช้กฎความสัมพันธ์ (Association Rules) เพื่อระบุรูปแบบที่เกิดขึ้นบ่อยครั้งในข้อมูล เช่น กฎ “หากลูกค้าซื้อ A แล้ว พวกเขามีแนวโน้มที่จะซื้อ B ด้วย”

กฎความสัมพันธ์ถูกสร้างขึ้นจากกระบวนการที่เริ่มต้นด้วยการค้นหาชุดรายการสินค้าที่เกิดร่วมกันบ่อย (Frequent Itemsets) ในข้อมูลธุรกรรม โดยใช้ค่าตัวชี้วัดสำคัญ เช่น Support, Confidence, และ Lift เพื่อประเมินคุณภาพของกฎ ความสัมพันธ์ที่พบจากการวิเคราะห์นี้สามารถนำไปประยุกต์ใช้ได้ในหลายบริบท ตั้งแต่การแนะนำสินค้าผ่านระบบแนะนำ (Recommender Systems) ไปจนถึงการเพิ่มโอกาสในการ Cross-Selling และ Up-Selling

หลักการทำงานของการวิเคราะห์ตระกร้าสินค้า

การทำงานของการวิเคราะห์ตระกร้าสินค้าจะเริ่มต้นด้วยการการวิเคราะห์ข้อมูลจากชุดรายการสินค้าที่ลูกค้าซื้อ โดยขั้นตอนที่ใช้ในการวิเคราะห์ ดังนี้

1. การระบุชุดสินค้าที่เกิดร่วมกันบ่อย: ขั้นแรกของการวิเคราะห์ คือ การค้นหาชุดสินค้าที่ถูกซื้อร่วมกันบ่อย โดยใช้ค่า Support เป็นเกณฑ์ในการคัดเลือกชุดสินค้าที่พบในหลายๆ ธุรกรรม การเลือกชุดสินค้าเหล่านี้จะช่วยให้เราเข้าใจว่า ลูกค้าจะเลือกซื้อสินค้าชิ้นไหนคู่กับสินค้าชิ้นไหนบ่อยครั้ง

2. การสร้างกฎความสัมพันธ์: หลังจากที่เราได้ชุดสินค้าที่เกิดร่วมกันบ่อยๆ แล้ว ขั้นต่อไปคือนำข้อมูลเหล่านี้มาสร้างกฎความสัมพันธ์ เช่น หากลูกค้าซื้อสินค้า A แล้ว พวกเขามีแนวโน้มที่จะซื้อสินค้า B ด้วย โดยกฎเหล่านี้จะช่วยให้เราเข้าใจความสัมพันธ์ระหว่างสินค้าต่างๆ และสามารถนำไปใช้ในการแนะนำสินค้าให้ลูกค้าซื้อพร้อมกันได้

3. การประเมินคุณภาพของกฎความสัมพันธ์: เมื่อเราได้กฎความสัมพันธ์แล้วจะต้องใช้เมทริกซ์ต่างๆ เพื่อประเมินประสิทธิภาพของกฎเหล่านั้น โดยจะมีการใช้ค่า Support, Confidence, และ Lift เป็นเครื่องมือในการประเมินผล ซึ่งช่วยให้เราทราบว่า กฎไหนที่มีความน่าเชื่อถือและสามารถนำไปใช้ในการตัดสินใจทางธุรกิจ

อัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้า

ในการใช้อัลกอริทึมต่างๆ เพื่อค้นหาชุดสินค้าที่เกิดร่วมกันบ่อย และสร้างกฎความสัมพันธ์ถือเป็นขั้นตอนที่สำคัญในการทำงานของการวิเคราะห์ตระกร้าสินค้า ซึ่งแต่ละอัลกอริทึมจะมีวิธีการทำงานที่แตกต่างกัน ดังนี้

1. อัลกอริทึม Apriori

Agrawal และ Srikant (1994) ได้พัฒนาอัลกอริทึม Apriori ซึ่งมีพื้นฐานอยู่บนแนวคิด Downward Closure Property หรือที่เรียกว่า Anti-Monotonicity โดยอธิบายว่า หากชุดสินค้าหนึ่ง (Itemset) ไม่เกิดร่วมกันบ่อย (Frequent Itemset) ชุดสินค้าที่ใหญ่กว่านั้นซึ่งมีรายการสินค้าเพิ่มเติม ก็ไม่สามารถเกิดร่วมกันบ่อยได้เช่นกัน หลักการนี้ช่วยลดจำนวน Itemsets ที่ต้องพิจารณาและช่วยปรับปรุงประสิทธิภาพของการวิเคราะห์

การทำงานของ Apriori มักเริ่มต้นด้วยการคำนวณค่า Support ของชุดสินค้าขนาดเล็ก และจากนั้นค่อยๆ ขยายไปสู่ชุดสินค้าที่มีขนาดใหญ่ขึ้นโดยใช้กฎ Downward Closure Property ซึ่งหมายความว่า หากชุดสินค้าขนาดใหญ่มีค่า Support ต่ำ ชุดสินค้าขนาดเล็กที่ประกอบด้วยสินค้าจากชุดใหญ่ก็จะมีค่า Support ต่ำเช่นกัน

2. อัลกอริทึม Frequent Pattern Growth (FP-Growth)

Han และคณะ (2000) ได้พัฒนาอัลกอริทึม FP-Growth ขึ้นเพื่อแก้ปัญหาความซับซ้อนในการประมวลผลของอัลกอริทึม Apriori โดยหลีกเลี่ยงการสร้างชุดรายการสินค้าที่เป็นตัวเลือก (Candidate Itemsets) ซึ่งมักทำให้ใช้ทรัพยากรในการคำนวณสูง FP-Growth ด้วยการใช้โครงสร้างข้อมูลแบบ Frequent Pattern Tree (FP-Tree) เพื่อจัดเก็บรายการสินค้าที่เกิดร่วมกันในธุรกรรมต่าง ๆ โดย FP-Tree จะบีบอัดข้อมูลธุรกรรมให้มีขนาดเล็กลงผ่านการจัดเก็บรายการสินค้าที่เหมือนกันไว้ในโหนดเดียวกัน ซึ่งช่วยลดทั้งพื้นที่จัดเก็บและเวลาในการประมวลผล

กระบวนการค้นหา Frequent Itemsets ของ FP-Growth เริ่มจากการสร้าง Conditional Pattern Base ซึ่งเป็นชุดข้อมูลย่อยที่เกี่ยวข้องกับแต่ละรายการสินค้า แล้วจึงใช้กระบวนการค้นหารูปแบบที่เกิดบ่อยในลักษณะของ Divide-and-Conquer ซึ่งไม่จำเป็นต้องทดสอบทุกชุดที่เป็นไปได้เหมือนใน Apriori

FP-Growth จึงสามารถค้นหารูปแบบความสัมพันธ์ในข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพมากกว่า โดยเฉพาะในกรณีที่ข้อมูลมีรายการสินค้าจำนวนมาก อย่างไรก็ตาม การสร้างและจัดการ FP-Tree อาจมีความซับซ้อนในเชิงโครงสร้าง จึงเหมาะสำหรับนักวิเคราะห์ที่มีความเชี่ยวชาญด้านการจัดการข้อมูล

3. อัลกอริทึม Equivalence Class Transformation (Eclat)

Zaki และคณะ (1997) ได้นำเสนออัลกอริทึม Eclat ซึ่งมีแนวทางแตกต่างจากอัลกอริทึม Apriori และ FP-Growth โดยใช้แนวคิดการจัดเก็บข้อมูลในรูปแบบแนวตั้ง (Vertical Data Format) กล่าวคือ รายการสินค้าทุกตัวจะเชื่อมโยงกับ หมายเลขธุรกรรม (Transaction IDs) ที่สินค้านั้นปรากฏอยู่ หลักการสำคัญของอัลกอริทึมนี้ คือ การคำนวณ

Frequent Itemsets โดยอาศัยการหาจุดตัดของเซต (Set Intersection) ระหว่างรายการ หมายเลขธุรกรรมของสินค้าต่างๆ ซึ่งช่วยลดจำนวนรอบในการสแกนข้อมูลซ้ำ และเพิ่มประสิทธิภาพในการค้นหารูปแบบที่เกิดบ่อย

ข้อได้เปรียบของ Eclat คือความรวดเร็วในการค้นหา Frequent Itemsets โดยไม่จำเป็นต้องสร้าง Candidate Itemsets ซ้ำ ๆ เช่นเดียวกับใน Apriori อย่างไรก็ตาม Eclat อาจไม่เหมาะกับข้อมูลที่มีรายการสินค้าจำนวนมากหรือมีโครงสร้างซับซ้อน เนื่องจากอาจเกิด Overhead จากการจัดการชุดข้อมูลที่มีความสัมพันธ์หลากหลาย

เมทริกซ์ที่ใช้ในการประเมินผล

ในการวิเคราะห์ตระกร้าสินค้า การใช้เมทริกซ์ที่เหมาะสมในการประเมินกฎความสัมพันธ์เป็นสิ่งสำคัญในการตัดสินใจว่าเราควรจะใช้กฎไหนในการแนะนำสินค้าหรือการทำโปรโมชั่น เพื่อเพิ่มโอกาสในการขาย โดยเมทริกซ์ที่ใช้ จะประกอบด้วย

1. Support: เป็นตัววัดความถี่ของการเกิดร่วมกันของ Itemsets ในชุดข้อมูลทั้งหมด ซึ่งช่วยให้ทราบว่า ชุดสินค้าที่มีการซื้อร่วมกันบ่อยครั้งนั้นมีความสำคัญหรือไม่ในบริบทของข้อมูล การคำนวณ Support ช่วยให้เรารู้ว่า ชุดสินค้านั้นๆ เกิดขึ้นบ่อยในข้อมูลหรือไม่ โดยสามารถคำนวณได้จากสมการ 2.26

$$Support(A) = \frac{\text{จำนวนรายการที่มี } A}{\text{จำนวนทั้งหมดของรายการ}} \quad (2.26)$$

2. Confidence: ใช้ในการวัดความน่าจะเป็นที่สินค้าหนึ่งจะถูกซื้อเมื่อสินค้าหนึ่งถูกซื้อก่อน โดยคำนวณจากการหาร Support ของชุดสินค้ารวมกันด้วย Support ของชุดสินค้าหมายเหตุ ซึ่ง Confidence ช่วยให้ทราบว่าเมื่อผู้ซื้อสินค้าชิ้น A แล้ว พวกเขามีโอกาสสูงที่จะซื้อสินค้าชิ้น B ด้วยหรือไม่ โดยสามารถคำนวณได้จากสมการ 2.27

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \quad (2.27)$$

3. Lift: เป็นตัวชี้วัดที่บ่งชี้ถึงความแข็งแกร่งของการเชื่อมโยงระหว่างชุดสินค้า A และ B โดยการเปรียบเทียบกับการเกิดของสินค้าทั่วไป ซึ่งจะช่วยในการวัดความแข็งแกร่งของการเกิดร่วมกันระหว่างสินค้า A และ B โดยไม่ขึ้นกับความถี่ของการเกิดในตลาด โดยสามารถคำนวณได้จากสมการ 2.28

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} \quad (2.27)$$

สรุปข้อมูลของอัลกอริทึมที่ที่ใช้ในการวิเคราะห์ตระกร้าสินค้า

จากได้ทำการศึกษาเกี่ยวกับอัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้าแต่ละอัลกอริทึม ประกอบด้วยอัลกอริทึม Apriori, อัลกอริทึม FP-Growth, และอัลกอริทึม Eclat ซึ่งทำให้มีความเข้าใจในแต่ละอัลกอริทึมมากขึ้น ทั้งในเรื่องของวิธีการทำงาน การประมวลผล ความซับซ้อนในการคำนวณ จุดเด่น ข้อจำกัด ตลอดจนลักษณะของข้อมูลที่เหมาะสมกับอัลกอริทึม ผู้วิจัยจึงได้ทำการสรุปข้อมูลของแต่ละอัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้าที่ได้ทำการศึกษา และแสดงข้อมูลที่ได้ในตาราง 2

ตาราง 2 สรุปข้อมูลของอัลกอริทึมที่ใช้ในการวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis)

อัลกอริทึม	การทำงานของอัลกอริทึม	การประมวลผล	ความซับซ้อนในการคำนวณ	จุดเด่น	ข้อจำกัด	ลักษณะข้อมูลที่เหมาะสม
Apriori	ใช้ Downward Closure Property เพื่อลดจำนวน Candidate Itemsets	ใช้การสแกนข้อมูลหลายรอบ (Iterative Process)	มีความซับซ้อนสูงเมื่อข้อมูลมีธุรกรรมและรายการสินค้าจำนวนมาก	- เรียบง่าย เข้าใจง่าย - ใช้งานแพร่หลาย - มีประสิทธิภาพกับข้อมูลขนาดเล็กถึงปานกลาง	- การสร้าง Candidate Itemsets ใช้ทรัพยากรสูง - ไม่เหมาะกับข้อมูลขนาดใหญ่	- ข้อมูลขนาดเล็กถึงปานกลาง - ธุรกรรมที่ไม่ซับซ้อนมาก
FP-Growth	ใช้โครงสร้าง FP-Tree บีบอัดข้อมูล และค้นหา Frequent Itemsets ผ่าน Conditional Pattern Base	สแกนข้อมูลเพียง 2 รอบ และใช้โครงสร้างต้นไม้เพื่อบีบอัดข้อมูล	ความซับซ้อนลดลงเมื่อเทียบกับ Apriori โดยเฉพาะในข้อมูลขนาดใหญ่	- ประหยัดเวลาและพื้นที่ - เหมาะสำหรับข้อมูลขนาดใหญ่ - ประมวลผลเร็วกว่า Apriori	- ต้องสร้าง FP-Tree ซึ่งมีความซับซ้อนเชิงโครงสร้าง - ไม่เหมาะสำหรับผู้ใช้ที่ไม่ชำนาญ	- ข้อมูลขนาดใหญ่ - รายการสินค้าจำนวนมาก

ตาราง 2 (ต่อ)

อัลกอริทึม	การทำงานของอัลกอริทึม	การประมวลผล	ความซับซ้อนในการคำนวณ	จุดเด่น	ข้อจำกัด	ลักษณะข้อมูลที่เหมาะสม
Eclat	ใช้ Vertical Data Format และ Set Intersection ในการหาความสัมพันธ์	ใช้การคำนวณ Set Intersection ทำให้ลดการสแกนข้อมูล	มีความเร็วและเหมาะสำหรับข้อมูลที่มีความหนาแน่นสูง แต่ไม่เหมาะกับข้อมูลที่มีรายการสินค้าจำนวนมาก	- ใช้ทรัพยากรต่ำในการค้นหา Frequent Itemsets - มีประสิทธิภาพสำหรับข้อมูลที่มีความหนาแน่นสูง	- การจัดการข้อมูลแนวตั้งอาจใช้ทรัพยากรสูงในข้อมูลที่ซับซ้อน - ไม่เหมาะกับข้อมูลที่มีรายการสินค้าจำนวนมาก	- ข้อมูลที่มีความหนาแน่นสูง - ความสัมพันธ์ที่ชัดเจนระหว่างรายการ

2.2 งานวิจัยที่เกี่ยวข้อง

1. **บทความวิจัยเรื่อง** A Comprehensive Framework for Superstore Business with Employing Effective Clustering Techniques (Mahfuza et al., 2021)

บทความนี้มุ่งเน้นการพัฒนาโครงสร้างการวิเคราะห์ลูกค้าสำหรับห้างสรรพสินค้า โดยเน้นการใช้เทคนิคการจัดกลุ่มลูกค้าเพื่อเพิ่มประสิทธิภาพการวิเคราะห์มูลค่าของลูกค้าและผลกำไรของธุรกิจ แบบจำลอง RFM และ LRFM ถูกนำมาใช้เพื่อวิเคราะห์พฤติกรรมการซื้อของลูกค้า โดยพิจารณาปัจจัยสำคัญ ได้แก่ ความถี่ในการซื้อ, ความใกล้เคียงของการซื้อครั้งล่าสุด, ระยะเวลาความสัมพันธ์กับธุรกิจ, และมูลค่ารวมของการใช้จ่าย เพื่อแบ่งลูกค้าออกเป็นกลุ่มตามลักษณะพฤติกรรม

ข้อมูลที่ใช้ในการศึกษาได้มาจาก Global Superstore 2018 ซึ่งเป็นชุดข้อมูลตัวอย่างที่เผยแพร่โดย Tableau โดยประกอบด้วยคำสั่งซื้อจำนวน 51,290 รายการ และ 24 คุณลักษณะ ข้อมูลดังกล่าวถูกนำมาปรับแต่งให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ เช่น การคำนวณ

ค่าตัวแปรตามแบบจำลอง LRFM และการลดมิติข้อมูลด้วย PCA เพื่อลดความซับซ้อนของข้อมูล และเตรียมความพร้อมสำหรับกระบวนการจัดกลุ่ม

ในกระบวนการจัดกลุ่ม ผู้วิจัยได้ใช้ 3 อัลกอริทึม ได้แก่ K-Means Clustering, Agglomerative Clustering, และ Fuzzy C-Means Clustering เพื่อเปรียบเทียบประสิทธิภาพในการจัดกลุ่มลูกค้า ผลการวิเคราะห์แสดงให้เห็นว่า แบบจำลอง LRFM ที่ใช้งานร่วมกับอัลกอริทึม K-Means Clustering มีประสิทธิภาพสูงสุด โดยสามารถระบุได้ทั้งหมด 7 กลุ่ม และพบว่า กลุ่มที่ 3 เป็นกลุ่มที่สร้างกำไรสูงสุดสำหรับธุรกิจ ซึ่งลูกค้าในกลุ่มนี้มีลักษณะเป็นผู้บริโภคที่มีความภักดีสูงและมีศักยภาพในการสร้างกำไร ในขณะเดียวกัน กลุ่มลูกค้าที่มีลักษณะ Demanding เช่น กลุ่มที่ 2 และ 5 สร้างกำไรต่ำและมีต้นทุนในการดูแลสูง ซึ่งผู้ประกอบการควรหลีกเลี่ยงหรือปรับกลยุทธ์เพื่อเพิ่มประสิทธิภาพในการให้บริการ

2. **บทความวิจัยเรื่อง** An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market (John et al., 2023)

บทความนี้มุ่งเน้นการศึกษาการแบ่งกลุ่มลูกค้าในตลาดค้าปลีกของสหราชอาณาจักร โดยใช้เทคนิคการจัดกลุ่ม ซึ่งเป็นเทคนิคในการเรียนรู้ของเครื่องแบบไม่มีผู้สอน เพื่อวิเคราะห์พฤติกรรมการซื้อของลูกค้าและสนับสนุนการตัดสินใจเชิงกลยุทธ์ในธุรกิจค้าปลีก งานวิจัยนี้ได้นำแบบจำลอง RFM มาประยุกต์ใช้ในการประเมินคุณค่าลูกค้า โดยพิจารณาความใกล้เคียงของการซื้อครั้งล่าสุด, ความถี่ในการซื้อ, และมูลค่ารวมของการซื้อสินค้า เพื่อระบุประเภทของลูกค้า แบบจำลอง RFM ถูกนำมาใช้ร่วมกับอัลกอริทึมการจัดกลุ่ม 5 ประเภท ได้แก่ K-Means Clustering, GMM, DBSCAN, Agglomerative Clustering, และ Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) เพื่อเปรียบเทียบผลลัพธ์และค้นหาอัลกอริทึมที่มีประสิทธิภาพสูงสุดในการแบ่งกลุ่มลูกค้า

ข้อมูลที่ใช้ในงานวิจัยนี้นำมาจาก UCI Machine Learning Repository: Online Retail Dataset ซึ่งประกอบด้วยข้อมูลลูกค้าจำนวน 541,909 รายการ และ 8 คุณลักษณะ ข้อมูลดังกล่าว ถูกจัดเตรียมในกระบวนการการเตรียมข้อมูล เพื่อกำจัดค่าที่ขาดหาย (Missing Values) และข้อมูลที่ไม่ถูกต้อง หลังจากนั้นจึงทำการลดจำนวนมิติของข้อมูลด้วยอัลกอริทึม PCA ซึ่งถูกนำไปใช้กับอัลกอริทึม GMM เพื่อเพิ่มความแม่นยำและลดความซับซ้อน โดยจะลดจำนวนมิติข้อมูลให้อยู่ใน 2 มิติ เพื่อทำให้การวิเคราะห์และการแสดงผลข้อมูลง่ายขึ้น

จากการวิเคราะห์ข้อมูล ลูกค้าถูกจัดกลุ่มตามค่า RFM Score ซึ่งแบ่งออกเป็น 5 กลุ่มหลัก ได้แก่ Top Customers (8%), High-Value Customers (10%), Medium-Value Customers

(21%), Low-Value Customers (30%), และ Lost Customers (31%) โดยพบว่า Top Customers และ High-Value Customers เป็นกลุ่มที่สร้างกำไรสูงสุดและควรได้รับการดูแลอย่างใกล้ชิด ในขณะที่ Lost Customers และ Low-Value Customers ต้องการการปรับกลยุทธ์เพื่อนำกลับมาสร้างมูลค่าให้กับธุรกิจ ในส่วนของการจัดกลุ่มโดยใช้อัลกอริทึมต่างๆ พบว่า อัลกอริทึม GMM มีประสิทธิภาพสูงสุดด้วยค่า Silhouette Score ที่ 0.80 ซึ่งเหนือกว่า K-Means Clustering (0.64), Agglomerative Clustering (0.64), BIRCH (0.64) และ DBSCAN (0.62) โดยอัลกอริทึม GMM สามารถระบุรูปแบบข้อมูลที่ซับซ้อนและสร้างกลุ่มลูกค้าที่ชัดเจนได้ดีที่สุด การใช้งานร่วมกับ PCA ช่วยเพิ่มความแม่นยำของ GMM อย่างมาก ในขณะที่อัลกอริทึมอื่นไม่ได้ใช้ PCA ในการวิเคราะห์

3. **บทความวิจัยเรื่อง** **Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm (Sugiharto et al., 2023)**

บทความนี้มีเป้าหมายเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมการจัดกลุ่มลูกค้า ได้แก่ GMM, K-Means Clustering, และ BIRCH โดยมุ่งเน้นการศึกษาความเหมาะสมในการแบ่งกลุ่มลูกค้าตามข้อมูลส่วนบุคคล เช่น เพศ อายุ รายได้ต่อปี และคะแนนการใช้จ่าย ตัวชี้วัดที่ใช้ในการประเมินผลคือ Davies-Bouldin Score (DB Score) ซึ่งช่วยวัดคุณภาพของกลุ่มที่ได้จากการวิเคราะห์ งานวิจัยนี้มีเป้าหมายเพื่อช่วยให้ธุรกิจในภาคการตลาดและการค้าปลีกใฟ้สามารถเข้าใจข้อมูลของลูกค้าได้ดีขึ้น และนำข้อมูลไปพัฒนากลยุทธ์การแข่งขันในเชิงธุรกิจ

ข้อมูลที่ใช้ในงานวิจัยมาจากชุดข้อมูล "Shop Customer Data" ซึ่งเป็นชุดข้อมูลดัดแปลงจาก "Mall Customer Segmentation Data" โดยมีข้อมูลของลูกค้า 2,000 รายการ ประกอบด้วยคุณลักษณะสำคัญ ได้แก่ Customer ID, Gender, Age, Annual Income, และ Spending Score หลังจากนำเข้าข้อมูล ได้มีการเตรียมข้อมูล เช่น การลบตัวแปรที่ไม่เกี่ยวข้อง (เช่น อาชีพและขนาดครอบครัว) และแปลงข้อมูลเพศให้อยู่ในรูปแบบตัวเลข (Encoding) เพื่อเตรียมความพร้อมสำหรับการประมวลผล

ในงานวิจัยนี้ ใช้เทคนิค Elbow Method เป็นเครื่องมือสำคัญในการกำหนดจำนวนกลุ่มที่เหมาะสมที่สุดสำหรับการจัดกลุ่มลูกค้า ผลการวิเคราะห์ระบุว่าจำนวนกลุ่มที่เหมาะสมคือ 6 กลุ่ม ซึ่งถูกนำไปใช้กับทั้ง 3 อัลกอริทึม ได้แก่ GMM, K-Means Clustering, และ BIRCH เพื่อเปรียบเทียบประสิทธิภาพของแต่ละวิธี

Davies-Bouldin Score ถูกใช้เป็นตัวชี้วัดคุณภาพของการจัดกลุ่ม โดยคะแนนที่ต่ำกว่าหมายถึงกลุ่มที่มีความแตกต่างกันอย่างชัดเจนและมีความกระชับภายในกลุ่ม โดยได้ผลลัพธ์ของแต่ผลอัลกอริทึม ดังนี้

- K-Means Clustering มีค่า DB Score เท่ากับ 0.47941
- BIRCH มีค่า DB Score เท่ากับ 0.48156
- GMM มีค่า DB Score เท่ากับ 0.48821

จากผลลัพธ์ที่ได้ K-Means Clustering แสดงให้เห็นถึงความสามารถที่เหมาะสมที่สุดสำหรับชุดข้อมูลนี้ เนื่องจากความเรียบง่ายและความเหมาะสมของข้อมูลกับอัลกอริทึม อย่างไรก็ตาม BIRCH และ GMM ยังคงแสดงผลลัพธ์ที่ดีและเหมาะสำหรับการใช้งานในบริบทข้อมูลที่แตกต่างกัน

งานวิจัยยังระบุว่าคุณภาพของอัลกอริทึมการจัดกลุ่มขึ้นอยู่กับความเหมาะสมของชุดข้อมูลกับอัลกอริทึมนั้นๆ ในกรณีนี้ K-Means Clustering มีความเหมาะสมที่สุดสำหรับข้อมูลเชิงตัวเลขที่เรียบง่าย ในขณะที่ BIRCH หรือ GMM อาจให้ผลลัพธ์ที่ดีกว่าในบริบทของข้อมูลที่มีความซับซ้อนสูง เช่น ข้อมูลเชิงเวลา

4. **บทความวิจัยเรื่อง** Customer Segmentation and Personalized Marketing Using K-Means and APRIORI Algorithm (Gayathri & Arunodhaya, 2021)

บทความนี้นำเสนอการแบ่งกลุ่มลูกค้าและการตลาดเฉพาะบุคคล (Personalized Marketing) โดยใช้เทคนิคการจัดกลุ่มลูกค้าและการวิเคราะห์ตะกร้า เพื่อเพิ่มยอดขายและความพึงพอใจของลูกค้า งานวิจัยใช้แบบจำลอง RFM เป็นเครื่องมือในการวิเคราะห์และจำแนกลูกค้าโดยพิจารณาความถี่และมูลค่าการซื้อ เพื่อระบุประเภทลูกค้า พร้อมใช้อัลกอริทึม K-Means Clustering ในการจัดกลุ่มลูกค้าตามพฤติกรรมการซื้อ และอัลกอริทึม Apriori รวมถึงอัลกอริทึม ECLAT ในการแนะนำสินค้าแบบคอมโบ (Combo Offer Recommendation) เพื่อช่วยเพิ่มประสิทธิภาพในแผนกลยุทธ์การตลาด

ข้อมูลที่ใช้ในการวิจัยได้จาก Kaggle Online Retail Dataset ซึ่งรวบรวมข้อมูลการสั่งซื้อสินค้าจากลูกค้าหลายประเทศทั่วโลก คุณลักษณะสำคัญที่ใช้ ได้แก่ InvoiceNo, StockCode, Quantity, UnitPrice, และ CustomerID โดยข้อมูลถูกปรับแต่งผ่านกระบวนการเตรียมข้อมูล เช่น การลบค่าที่ขาดหาย และค่าที่ไม่สมเหตุสมผล เพื่อให้เหมาะสมสำหรับการวิเคราะห์

จากการวิเคราะห์ข้อมูลด้วย RFM Analysis ลูกค้าถูกแบ่งออกเป็น 3 กลุ่มหลัก โดยอัลกอริทึม K-Means Clustering ถูกนำมาใช้ โดยใช้ร่วมกับเทคนิค Elbow Method เพื่อกำหนดจำนวนกลุ่มที่เหมาะสม พบว่า จำนวนที่เหมาะสมคือ 3 กลุ่ม ซึ่งแบ่งตามลักษณะของความถี่และ

มูลค่าการซื้อเป็น High, Medium และ Low การวิเคราะห์เพิ่มเติมด้วยการวิเคราะห์ตระกร้าสินค้าสำหรับการแนะนำสินค้าแบบคอมโบ ซึ่งทำให้ได้ผลลัพธ์จากการวิเคราะห์ ดังนี้

- ECLAT Algorithm: แนะนำสินค้าได้ 72 รายการ เมื่อกำหนดค่า Min Support = 0.03

- Apriori Algorithm: แนะนำสินค้าได้ 31 รายการ เมื่อกำหนดค่า Min Support = 0.03 ค่า Confidence = 0.5 และค่า Lift >= 1.2

จากการวัดประสิทธิภาพของอัลกอริทึม K-Means Clustering พบว่า มีค่า Silhouette Score = 0.65 ซึ่งสะท้อนถึงความชัดเจนและความกระชับของกลุ่มลูกค้า ช่วยให้การวางแผนกลยุทธ์การตลาดในแต่ละกลุ่มมีประสิทธิภาพมากขึ้น

งานวิจัยสรุปว่า อัลกอริทึม Apriori มีความเหมาะสมกับการใช้งานในข้อมูลขนาดใหญ่ ในขณะที่อัลกอริทึม ECLAT ให้ผลลัพธ์ที่รวดเร็วกว่าในชุดข้อมูลขนาดเล็ก

5. **บทความวิจัยเรื่อง** K-Means Clustering-Based Market Basket Analysis: U.K. Online E-Commerce Retailer (Lim, 2021)

บทความนี้นำเสนอการประยุกต์ใช้อัลกอริทึม K-Means Clustering และการวิเคราะห์ตะกร้าสินค้า เพื่อทำความเข้าใจพฤติกรรมของลูกค้าในธุรกิจอีคอมเมิร์ซออนไลน์ของสหราชอาณาจักร โดยมุ่งเน้นการสนับสนุนการตัดสินใจเชิงกลยุทธ์และการเพิ่มยอดขาย งานวิจัยมีวัตถุประสงค์ในการวิเคราะห์รูปแบบการซื้อสินค้าของลูกค้า เพื่อแนะนำสินค้าที่เหมาะสมและเพิ่มประสิทธิภาพในด้านการตลาด

อัลกอริทึม K-Means Clustering ถูกใช้ในการจัดกลุ่มลูกค้าตามพฤติกรรมการซื้อ โดยการใช้เทคนิค Elbow Method ช่วยกำหนดจำนวนกลุ่มที่เหมาะสมที่สุดได้ที่ 4 กลุ่ม

การวิเคราะห์ตระกร้าสินค้าถูกนำมาใช้เพื่อวิเคราะห์สินค้าที่มักถูกซื้อร่วมกัน โดยใช้ค่า Support และค่า Confidence ในการสร้างคำแนะนำสินค้าแบบคอมโบ ผลลัพธ์แสดงว่าสินค้าที่มีค่า Support = 0.04 หมายความว่า สินค้าเหล่านี้ถูกซื้อร่วมกันใน 4% ของธุรกรรมทั้งหมด ในขณะที่ค่า Confidence = 0.65 แสดงถึง โอกาส 65% ที่ลูกค้าจะซื้อสินค้า B ร่วมกับสินค้า A หากได้ซื้อสินค้า A ไปแล้ว ค่าความสัมพันธ์นี้ถูกใช้เพื่อแนะนำสินค้าเพิ่มเติมให้ตรงตามพฤติกรรมการซื้อของลูกค้า

จากค่า Silhouette Score = 0.72 ชี้ให้เห็นว่ากลุ่มลูกค้าที่จัดกลุ่มด้วย K-Means Clustering มีความชัดเจนในแต่ละกลุ่มและมีความกระชับภายในกลุ่ม ในส่วนของการวิเคราะห์

ตะกร้าสินค้า ใช้ค่า Min Support = 0.03 เพื่อกรองสินค้าที่ถูกซื้อร่วมกันบ่อย และค่า Confidence = 0.6 เพื่อแนะนำสินค้าที่มีความน่าเชื่อถือสูงในการซื้อร่วมกัน

งานวิจัยสรุปว่า อัลกอริทึม K-Means Clustering เป็นเครื่องมือที่มีประสิทธิภาพในการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อ ขณะที่การวิเคราะห์ตะกร้าสินค้าก็ช่วยเพิ่มความสามารถในการแนะนำสินค้าแบบคอมโบอย่างแม่นยำ โดยมีค่า Support และค่า Confidence เป็นตัวชี้วัดที่สำคัญ

6. **บทความวิจัยเรื่อง** Market Basket Analysis of a Health Food Store in Thailand: A Case Study (Singha et al., 2024)

บทความนี้นำเสนอการวิเคราะห์ตะกร้าสินค้าของร้านอาหารเพื่อสุขภาพในประเทศไทย โดยมุ่งเน้นการศึกษาความสัมพันธ์ระหว่างสินค้าต่างๆ ในชุดข้อมูลการทำธุรกรรมของลูกค้า และนำผลลัพธ์ไปใช้ในการออกแบบกลยุทธ์การตลาดที่มีประสิทธิภาพ งานวิจัยนี้ใช้อัลกอริทึม FP-Growth มาพิจารณาเมทริกซ์ที่สำคัญ เช่น Support และ ps (การคำนวณความถี่ที่สังเกตได้เทียบกับความถี่ที่คาดการณ์) นอกจากนี้ยังมีการเพิ่มการวิเคราะห์ในเชิงมูลค่าทางการเงินของความสัมพันธ์ เพื่อช่วยในการออกแบบแคมเปญการตลาดที่เหมาะสม

ข้อมูลที่ใช้ประกอบด้วยธุรกรรมการซื้อสินค้าของลูกค้าจากร้านอาหารเพื่อสุขภาพในประเทศไทย โดยมีจำนวนธุรกรรมทั้งหมด 112,991 รายการ และสินค้าทั้งหมด 1,518 รายการ ซึ่งได้คัดเลือกเฉพาะสินค้า 238 รายการ ที่มียอดขายประจำปีไม่น้อยกว่า 20,000 บาท

ผลการวิเคราะห์แสดงให้เห็นว่า ความสัมพันธ์ระหว่างหมวดหมู่สินค้า ระหว่างผักสดและผลไม้สด มีค่า Support = 8.46% หมายความว่าในทุกๆ 100 รายการซื้อ จะมีการซื้อสินค้าจากทั้งสองหมวดหมู่ร่วมกันประมาณ 8 รายการ ค่าความสัมพันธ์ที่คาดการณ์ (ps) สูงถึง 319.69 ซึ่งสะท้อนถึงความสัมพันธ์ที่เกิดขึ้นบ่อยกว่าโดยคาดการณ์ จากผลลัพธ์นี้ยอดขายรวมจากความสัมพันธ์ระหว่างผักสดและผลไม้สดเท่ากับ 563,095 บาท คิดเป็น 3.76% ของยอดขายรวมประจำปี

ในส่วนของ ความสัมพันธ์ระหว่างสินค้ารายการเดียวกัน ตัวอย่างที่เห็นได้ชัดคือการซื้อ Iceberg Lettuce และ Cabbage ซึ่งพบว่ามีค่า Support = 3.27% และ ps = 518.30 โดยยอดขายจากความสัมพันธ์นี้มีมูลค่า 114,000 บาท คิดเป็นส่วนหนึ่งของยอดขายรวม

สำหรับความสัมพันธ์ระหว่างหมวดหมู่สินค้าและช่วงเวลาที่ซื้อ จากผลลัพธ์ พบว่าหมวดหมู่ผักสด และ ช่วงเวลาสาย มีค่า Support = 15.11% ซึ่งสูงกว่าค่าที่คาดการณ์ถึง 2,197 เท่า และสร้างยอดขายรวมได้ 1,271,857 บาท หรือ 8.49% ของยอดขายประจำปี ซึ่งบ่งชี้ถึง

ความสัมพันธ์ที่แข็งแกร่งและสามารถใช้ในการพัฒนากลยุทธ์การตลาดที่เน้นเวลาในการซื้อได้อย่างมีประสิทธิภาพ

งานวิจัยนี้แสดงให้เห็นว่าการใช้อัลกอริทึม FP-Growth ในการวิเคราะห์ตะกร้าสินค้าช่วยสร้างกลยุทธ์การตลาดที่มีประสิทธิภาพ เช่น การขายแบบ Cross-Selling และการส่งเสริมการขายแบบเจาะจงช่วงเวลาในวันหรือฤดูกาล นอกจากนี้ การใช้มูลค่าทางการเงินเป็นตัวชี้วัดร่วมกับ Support และ ps ช่วยให้การวิเคราะห์มีความแม่นยำและสามารถนำไปใช้ได้จริงในเชิงธุรกิจ

7. **บทความวิจัยเรื่อง** Market Basket Analysis Using FP-Growth Algorithm On Retail Sales Data (Pradana et al., 2022)

บทความนี้มุ่งเน้นการใช้อัลกอริทึม FP-Growth ในการวิเคราะห์ตะกร้าสินค้า โดยมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างสินค้าต่างๆ ในข้อมูลการทำธุรกรรมการขายของร้านค้าปลีก และนำผลลัพธ์ไปช่วยออกแบบกลยุทธ์การตลาดที่มีประสิทธิภาพ

ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลการทำธุรกรรมการขายในช่วงระยะเวลา 4 เดือน (มกราคม 2022 ถึง เมษายน 2022) โดยใช้ข้อมูลธุรกรรมทั้งหมด 571 รายการ ที่เกิดขึ้นในวันเสาร์และวันอาทิตย์ ซึ่งเป็นช่วงที่มียอดขายสูงสุด

ผลการวิเคราะห์ได้ผลลัพธ์ใน 4 กฎการซื้อสินค้าที่มีค่า Lift สูงกว่า 1 ซึ่งสะท้อนถึงความสัมพันธ์ที่มีความน่าเชื่อถือ เช่น:

- กฎที่ 1: หากลูกค้าซื้อ Nike Airforce One จะซื้อ Nike Socks ด้วย ความสัมพันธ์นี้มีค่า Support = 5.84%, Confidence = 89.29%, และ Lift = 11.58
- กฎที่ 2: หากลูกค้าซื้อ Nike Socks จะซื้อ Nike Airforce 1 ด้วย ความสัมพันธ์นี้มีค่า Support = 5.84%, Confidence = 75.76%, และ Lift = 11.58
- กฎที่ 3: หากลูกค้าซื้อ Champion Shorts จะซื้อ Champion Tees ด้วย ความสัมพันธ์นี้มีค่า Support = 4.44%, Confidence = 57.58%, และ Lift = 5.24
- กฎที่ 4: หากลูกค้าซื้อ Vans Old Skool จะซื้อ Vans Socks ด้วย ความสัมพันธ์นี้มีค่า Support = 5.37%, Confidence = 53.49%, และ Lift = 4.02

จากผลการศึกษา แสดงให้เห็นว่า อัลกอริทึม FP-Growth สามารถให้ผลลัพธ์ที่มีประสิทธิภาพในการค้นหาความสัมพันธ์ระหว่างสินค้าที่มักจะถูกซื้อร่วมกัน ซึ่งนำไปสู่การออกแบบกลยุทธ์การตลาด เช่น การขายแบบ Cross-Selling โดยสามารถจับคู่สินค้าสำหรับการส่งเสริมการขายได้ เช่น การจับคู่ Nike Airforce 1 และ Nike Socks หรือ Vans Old Skool และ Vans Socks เพื่อเพิ่มยอดขาย

การใช้ FP-Growth ในการวิเคราะห์ตลาดและพฤติกรรมการซื้อของลูกค้าช่วยให้ร้านค้าสามารถพัฒนากลยุทธ์การตลาดที่มีความแม่นยำและตรงกับความต้องการของลูกค้าได้ดียิ่งขึ้น นอกจากนี้การคำนวณ Support, Confidence, และ Lift ช่วยให้การวิเคราะห์มีความชัดเจนและสามารถนำไปใช้ได้จริงในเชิงธุรกิจ

8. **บทความวิจัยเรื่อง** **Market Basket Analysis Using Apriori and FP-Growth Algorithm (Hossain et al., 2019)**

บทความนี้มุ่งเน้นการใช้อัลกอริทึม Apriori และอัลกอริทึม FP-Growth ในการวิเคราะห์ตะกร้าสินค้า เพื่อศึกษาความสัมพันธ์ระหว่างสินค้าต่างๆ ในชุดข้อมูลการทำธุรกรรมของลูกค้า โดยมีวัตถุประสงค์ในการช่วยออกแบบกลยุทธ์การตลาดที่มีประสิทธิภาพ

ข้อมูลที่ใช้ในงานวิจัยนี้ประกอบด้วยธุรกรรมการซื้อสินค้าจากร้านค้าปลีก โดยมีการเลือกสินค้า 30%, 40%, 50%, และ 55% ของสินค้าที่ขายดีที่สุดเพื่อลดขนาดของข้อมูลและเพิ่มประสิทธิภาพในการประมวลผล ข้อมูลประกอบด้วย 3 ชุดข้อมูลหลัก ที่ใช้ในการทดสอบ ได้แก่ French Retail Dataset, Bakery Shop Dataset, และ Online Retail Dataset

ผลการวิเคราะห์ แสดงให้เห็นว่า เมื่อใช้การลดขนาดข้อมูลโดยเลือกสินค้าที่ขายดีที่สุด จะสามารถทำการวิเคราะห์ได้ในเวลาที่สั้นลง โดยไม่ส่งผลกระทบต่อคุณภาพของกฎการซื้อสินค้า ตัวอย่างเช่น การใช้ 50% ของสินค้าที่ขายดีที่สุด จะทำให้ผลลัพธ์ของ Frequent Itemsets และ Association Rules ที่ได้มีความคล้ายคลึงกับผลลัพธ์จากการใช้ข้อมูลทั้งหมด โดยไม่ต้องคำนวณข้อมูลทุกชิ้น ซึ่งได้ผลลัพธ์จากการทำลอง ดังนี้

- Apriori Algorithm ใช้เวลาในการประมวลผลมากกว่าหลังจากการลดขนาดข้อมูลด้วย 55% ของสินค้าที่ขายดีที่สุด ในขณะที่ FP-Growth Algorithm ใช้เวลาในการประมวลผลน้อยกว่า และทำงานได้เร็วขึ้นเมื่อใช้กับข้อมูลที่ลดขนาดแล้ว

- Support และ Confidence ถูกใช้ในการประเมินความสัมพันธ์ระหว่างสินค้า หมายความว่า Min Support = 0.03 และ Min Confidence = 0.5 เพื่อแนะนำสินค้าที่มีโอกาสถูกซื้อร่วมกันสูง

งานวิจัยนี้สรุปว่า อัลกอริทึม FP-Growth เหมาะสมกว่าการใช้อัลอริทึม Apriori ในการวิเคราะห์ตลาดขนาดใหญ่ เนื่องจากใช้เวลาน้อยกว่าและไม่ต้องการการคำนวณที่ซับซ้อนมากนัก นอกจากนี้ การลดขนาดข้อมูลด้วยการเลือก สินค้าที่ขายดีที่สุด ช่วยเพิ่มประสิทธิภาพในการวิเคราะห์โดยไม่สูญเสียความแม่นยำจากการค้นหาความสัมพันธ์ระหว่างสินค้า

9. **บทความวิจัยเรื่อง** Product Recommendations Using Market Basket Analysis with FP-Growth and Clustering Techniques (AL et al., 2022)

บทความนี้มุ่งเน้นการใช้การวิเคราะห์ตระกร้าสินค้า เพื่อศึกษาความสัมพันธ์ระหว่างสินค้าต่างๆ โดยใช้อัลกอริทึม FP-Growth ร่วมกับการแบ่งกลุ่มข้อมูลผ่านอัลกอริทึม K-Means Clustering เพื่อออกแบบกลยุทธ์การตลาดที่มีประสิทธิภาพสำหรับธุรกิจ Sukku Coffee & Space ซึ่งเป็นธุรกิจในอุตสาหกรรมอาหารและเครื่องดื่ม งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลการทำธุรกรรมของลูกค้าและแนะนำสินค้าที่เหมาะสมเพื่อเพิ่มยอดขาย โดยใช้ข้อมูลการทำธุรกรรมจริงระหว่างปี 2020-2022 จำนวน 34,745 รายการ

ข้อมูลถูกนำมาผ่านการวิเคราะห์ด้วยเทคนิค RFM Analysis ซึ่งช่วยในการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อ โดยการพิจารณาระยะเวลาการซื้อครั้งล่าสุด ความถี่ในการซื้อ และมูลค่าการใช้จ่ายทั้งหมด หลังจากการปรับช่วงขอบเขตของข้อมูลด้วย Z-Transformation แล้ว ข้อมูลถูกจัดกลุ่มโดยใช้อัลกอริทึม K-Means Clustering ซึ่งได้จำนวนกลุ่มที่เหมาะสมที่สุดคือ 3 กลุ่ม ได้แก่ กลุ่มลูกค้าที่ซื้อสินค้าบ่อยที่สุด กลุ่มลูกค้าที่มีความถี่ปานกลาง และกลุ่มลูกค้าที่ซื้อน้อยครั้งแต่มีมูลค่าการซื้อที่สูง

ในขั้นตอนการวิเคราะห์ความสัมพันธ์ของสินค้า อัลกอริทึม FP-Growth ถูกนำมาใช้เพื่อค้นหาความสัมพันธ์ของสินค้าที่มักถูกซื้อพร้อมกัน โดยใช้ Support และ Confidence เป็นเกณฑ์วัดความสำคัญของความสัมพันธ์ ตัวอย่างผลลัพธ์ที่ได้คือ ความสัมพันธ์ระหว่าง Sukku Aren และ Choco Berry ซึ่งมีค่า Support = 0.156 และ Confidence = 0.991 แสดงถึงโอกาสสูงที่ลูกค้าจะซื้อ Choco Berry เมื่อซื้อ Sukku Aren ผลการวิเคราะห์ยังพบความสัมพันธ์อื่นๆ ที่ช่วยในการสร้างกฎสำหรับแนะนำสินค้าและวางแผนโปรโมชั่นได้อย่างมีประสิทธิภาพ

นอกจากนี้ คุณภาพของการจัดกลุ่มข้อมูลยังได้รับการประเมินผ่านเมทริกซ์ เช่น ค่า Average within centroid distance ซึ่งมีค่าน้อยที่สุดในกลุ่ม Cluster 0 แสดงถึงการกระจายตัวที่ต่ำภายในกลุ่ม และค่า Davies-Bouldin Index ที่เท่ากับ -0.881 สะท้อนให้เห็นว่าการจัดกลุ่มข้อมูลด้วย K-Means Clustering นั้นมีความแม่นยำและชัดเจน ผลการศึกษาสรุปได้ว่า การผสานเทคนิค RFM Analysis, K-Means Clustering และ FP-Growth ช่วยให้สามารถวิเคราะห์พฤติกรรมลูกค้าและค้นหาความสัมพันธ์ของสินค้าที่มีความสำคัญ ซึ่งสามารถนำไปใช้ในการแนะนำสินค้าหรือวางกลยุทธ์การขายไขว้ (Cross-Selling) ได้อย่างแม่นยำ

10. **บทความวิจัยเรื่อง** Unveiling Patterns in the Night Market: A Machine Learning Approach to Customer Segmentation and Market Basket Analysis (Phyo & Uttama, 2023)

บทความนี้มีวัตถุประสงค์ในการศึกษาพฤติกรรมการซื้อของลูกค้าในตลาดกลางคืน โดยใช้เทคนิคการเรียนด้วยเครื่อง โดยเน้นไปที่การแบ่งกลุ่มลูกค้าและการวิเคราะห์ตะกร้าสินค้า เพื่อนำเสนอข้อมูลเชิงลึกเกี่ยวกับความชอบและแนวโน้มการซื้อสินค้า งานวิจัยนี้เก็บรวบรวมข้อมูลจากตลาดกลางคืนของมหาวิทยาลัยแม่ฟ้าหลวง จังหวัดเชียงราย โดยใช้อัลกอริทึม K-Means Clustering ในการจัดกลุ่มลูกค้า และใช้อัลกอริทึม Apriori สำหรับวิเคราะห์ความสัมพันธ์ระหว่างสินค้าที่ถูกซื้อพร้อมกัน

ในการศึกษานี้ ข้อมูลถูกเก็บรวบรวมจากสองแหล่ง ได้แก่ แบบสอบถาม Google Form จำนวน 43 ชุด และการสัมภาษณ์ลูกค้าในตลาดกลางคืนโดยตรง หลังจากกระบวนการทำความสะอาดข้อมูล (Data Cleaning) และการแปลงข้อมูลให้เป็นมาตรฐานด้วย Z-Score Normalization ข้อมูลถูกนำไปใช้ในการวิเคราะห์

ขั้นตอนแรกในการศึกษานี้คือการจัดกลุ่มลูกค้าด้วย K-Means Clustering โดยใช้เทคนิค Elbow Method ในการกำหนดจำนวนคลัสเตอร์ที่เหมาะสมที่สุด ซึ่งได้จำนวนกลุ่มเป็น 5 กลุ่ม ประสิทธิภาพของการจัดกลุ่มวัดโดยใช้เมทริกซ์ดังต่อไปนี้:

- Silhouette Score = 0.56
- Davies-Bouldin Index = 0.67
- Calinski-Harabasz Index = 321.5

ผลลัพธ์แสดงให้เห็นว่ากลุ่มลูกค้าที่ 3 และ 4 มีค่าเฉลี่ยการใช้จ่ายสูงสุด ในขณะที่กลุ่มที่ 0 มีกลุ่มลูกค้าหญิงเป็นสัดส่วนที่มากที่สุด ส่วนกลุ่มที่ 1 และ 2 มีการใช้จ่ายเฉลี่ยต่ำที่สุด โดยงานวิจัยแนะนำให้ตลาดกลางคืนควรให้ความสำคัญกับลูกค้าในกลุ่มที่ 3 และ 4 ซึ่งมีศักยภาพสูงในการสร้างรายได้

สำหรับการวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) ใช้อัลกอริทึม Apriori ในการสร้างกฎความสัมพันธ์ระหว่างสินค้า โดยใช้เกณฑ์ Minimum Support = 0.01, Minimum Confidence = 0.5 และ Minimum Lift = 1.3 ตัวอย่างกฎที่ค้นพบ ได้แก่:

- ความสัมพันธ์ระหว่าง "Street food and snacks" และ "Beverages" ซึ่งมีค่า Support = 0.017, Confidence = 1.000 และ Lift = 2.185
- ความสัมพันธ์ระหว่าง "Plants and flowers" และ "Fruits" ซึ่งมีค่า Support = 0.017, Confidence = 1.000 และ Lift = 3.471

- ความสัมพันธ์ระหว่าง "Desserts" และ "Beverages" ซึ่งมีค่า Support = 0.068, Confidence = 0.667 และ Lift = 1.457

เมื่อทดสอบกฎเหล่านี้กับกลุ่มลูกค้าใหม่จำนวน 20 คน พบว่า กฎที่มีประสิทธิภาพมากที่สุด คือ ความสัมพันธ์ระหว่าง "Street food and snacks" และ "Beverages" ซึ่งสามารถทำนายได้ถูกต้องถึง 10 คนจากทั้งหมด 20 คน ในขณะที่กฎอื่นๆ มีความแม่นยำน้อยกว่า

งานวิจัยนี้สามารถแบ่งกลุ่มลูกค้าด้วยอัลกอริทึม K-Means Clustering ได้อย่างมีประสิทธิภาพ โดยมีการประเมินประสิทธิภาพการจัดกลุ่มผ่าน Silhouette Score, Davies-Bouldin Index และ Calinski-Harabasz Index ในขณะเดียวกัน การวิเคราะห์ความสัมพันธ์ด้วยอัลกอริทึม Apriori ช่วยค้นหาความสัมพันธ์ของสินค้าที่มักถูกซื้อร่วมกัน โดยเมทริกซ์ Support, Confidence และ Lift ช่วยในการระบุความสัมพันธ์ที่มีนัยสำคัญ ข้อมูลที่ได้สามารถนำไปใช้ในการจัดวางร้านค้า การทำโปรโมชั่นร่วม และวางกลยุทธ์การตลาดเพื่อเพิ่มยอดขายได้อย่างมีประสิทธิภาพ

สรุปข้อมูลของงานวิจัยที่เกี่ยวข้อง

จากได้ทำการศึกษาเกี่ยวกับงานวิจัยที่เกี่ยวข้อง ซึ่งทำให้มีความเข้าใจในอัลกอริทึมแต่ละประเภทมากขึ้น รวมถึงได้ทราบถึง วิธีการจัดเตรียมข้อมูลก่อนที่จะวิเคราะห์ในแบบจำลอง การวิเคราะห์ข้อมูลในแบบจำลองต่างๆ ตลอดจนการวัดประสิทธิภาพของแบบจำลอง ผู้วิจัยจึงได้ทำการสรุปข้อมูลของแต่ละงานวิจัยที่ได้ทำการศึกษา และแสดงข้อมูลที่ได้ในตาราง 3

ตาราง 3 สรุปผลการทดลองของงานวิจัยที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
1	A Comprehensive Framework for Superstore Business with Employing Effective Clustering Techniques	Rezwana Mahfuza, Rafsan Shartaj Uddin, Yeaminur Rahman, and Md. Abdul Hai	2021	- RFM Analysis - LRFM - K-Means Clustering - Agglomerative Clustering - Fuzzy C-Means	ใช้แบบจำลอง RFM และ LRFM ร่วมกับ K-Means Clustering เพื่อแบ่งลูกค้าออกเป็น 7 กลุ่ม โดยกลุ่มที่ 3 สร้างกำไรสูงสุด และพบว่า การจัดกลุ่มด้วย LRFM ช่วยเพิ่มความแม่นยำในการระบุลูกค้าสำคัญที่สร้างกำไรให้ธุรกิจ

ตาราง 3 (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
2	An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market	Jeen Mary John, Olamilekan Shobayo, and Bayode Ogunleye	2023	- RFM Analysis - PCA - K-Means Clustering - GMM - DBSCAN - Agglomerative Clustering - BIRCH	ศึกษาและเปรียบเทียบอัลกอริทึมการจัดกลุ่มลูกค้าในตลาดค้าปลีก UK โดยใช้ RFM และ PCA พบว่า GMM มี Silhouette Score สูงสุดที่ 0.80 และเหมาะสมที่สุดในการแบ่งกลุ่มลูกค้าเมื่อเทียบกับ K-Means Clustering, Agglomerative และ DBSCAN
3	Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm	Nathanael Deciano Sugiharto, Darian Elbert, Juan Arnold, Ivan S. Edbert, and Derwin Suhartono	2023	- GMM - K-Means Clustering - BIRCH	เปรียบเทียบประสิทธิภาพของ K-Means Clustering, GMM และ BIRCH ในการจัดกลุ่มลูกค้าจากข้อมูลประชากร โดย K-Means Clustering มีค่า Davies-Bouldin Index ต่ำสุดที่ 0.47941 แสดงถึง คุณภาพการจัดกลุ่มที่ดีที่สุดสำหรับข้อมูลชุดนี้

ตาราง 3 (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
4	Customer Segmentation and Personalized Marketing Using K-Means and APRIORI Algorithm	Gayathri, and Arunodhaya	2021	- RFM Analysis - K-Means Clustering - Apriori - ECLAT	บทความนี้ใช้ RFM วิเคราะห์ลูกค้าและแบ่งกลุ่มด้วย K-Means Clustering ออกเป็น 3 กลุ่ม พร้อมใช้ Apriori และ ECLAT สร้างคำแนะนำสินค้าจากการวิเคราะห์ตะกร้าสินค้า พบว่า การใช้ Apriori สามารถสร้างกฎที่มี Support ≥ 0.03 และ Confidence ≥ 0.5 ช่วยเพิ่มยอดขายได้
5	K-Means Clustering-Based Market Basket Analysis: U.K. Online E-Commerce Retailer	Tristan Lim	2021	- K-Means Clustering - Elbow Method - Market Basket Analysis	นำ K-Means Clustering มาวิเคราะห์และจัดกลุ่มลูกค้าในธุรกิจอีคอมเมิร์ซออนไลน์ UK ออกเป็น 4 กลุ่ม และใช้ Market Basket Analysis เพื่อหาความสัมพันธ์ของสินค้า พบว่า สินค้าที่ถูกซื้อร่วมกันมี Support 0.04 และ Confidence 0.65

ตาราง 3 (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
6	Market Basket Analysis of a Health Food Store in Thailand: A Case Study	Kanokwan Singha, Parthana Parthanadee, Ajchara Kessuvan, and Jirachai Buddhakulsomsiri	2023	- FP-Growth	ศึกษาการวิเคราะห์ตะกร้าสินค้าของร้านอาหารเพื่อสุขภาพในไทยด้วย FP-Growth พบความสัมพันธ์สำคัญระหว่างหมวดหมู่ผักสดและผลไม้สดที่ Support 8.46% และ ps 319.69 พร้อมแนะนำกลยุทธ์ Cross-Selling เพื่อเพิ่มยอดขาย
7	Market Basket Analysis Using FP-Growth Algorithm On Retail Sales Data	Muhammad Raihan Pradana, Mohammad Syafrullah, Hendri Irawan, Irawan, Joko Christian Chandra, and Achmad Solichin	2022	- FP-Growth	บทความนี้ใช้ FP-Growth วิเคราะห์ข้อมูลการขายของร้านค้าปลีก โดยพบ 4 กฎความสัมพันธ์ที่มี Lift > 1 และค่า Support ≥ 4%, Confidence ≥ 50% ช่วยให้ร้านค้าสามารถออกแบบกลยุทธ์ Cross-Selling ได้แม่นยำยิ่งขึ้น

ตาราง 3 (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
8	Market Basket Analysis Using Apriori and FP-Growth Algorithm	Maliha Hossain, A H M Sarowar Sattar, and Mahit Kumar Paul	2019	- Apriori - FP-Growth	เปรียบเทียบ Apriori และ FP-Growth ในการวิเคราะห์ตะกร้าสินค้า พบว่า FP-Growth ใช้เวลาประมวลผลน้อยกว่าและมีประสิทธิภาพสูงกว่า โดยการลดข้อมูลเป็นสินค้าขายดี 55% ช่วยลดเวลาโดยไม่กระทบต่อคุณภาพของผลลัพธ์
9	Product Recommendations Using Market Basket Analysis with FP-Growth and Clustering Techniques	Muhammad Rizky AL, Meilita Tryana Sembiring, and Richo Giwana Resdy Maulana	2022	- RFM Analysis - K-Means Clustering - FP-Growth	ผสาน RFM และ K-Means Clustering แบ่งลูกค้าออกเป็น 3 กลุ่ม และใช้ FP-Growth วิเคราะห์ตะกร้าสินค้า พบความสัมพันธ์สินค้าระหว่าง Sukku Aren และ Choco Berry ที่มี Support 0.156 และ Confidence 0.991 ช่วยวางแผนโปรโมชั่นอย่างมีประสิทธิภาพ

ตาราง 3 (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	เทคนิคที่ใช้	ผลการทดลอง
10	Unveiling Patterns in the Night Market: A Machine Learning Approach to Customer Segmentation and Market Basket Analysis	Thandar Phyo, and Surapong Uttama	2023	- K-Means Clustering - Apriori	วิเคราะห์พฤติกรรมลูกค้าในตลาดกลางคืนโดยใช้ K-Means Clustering แบ่งลูกค้าออกเป็น 5 กลุ่ม พร้อมใช้ Apriori ค้นหาความสัมพันธ์สินค้า พบว่าความสัมพันธ์ระหว่าง Street Food และ Beverages สูงที่สุดด้วย Support 0.017 และ Confidence 1.0

บทที่ 3
วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการวิจัยตามขั้นตอนดังต่อไปนี้

1. กระบวนการสร้างแบบจำลอง
2. การเก็บรวบรวมข้อมูล (Data Collection)
3. การจัดระเบียบข้อมูล (Data Wrangling)
4. การวิเคราะห์ข้อมูลเชิงการตรวจสอบเบื้องต้น (Exploratory Data Analysis, EDA)
5. การเตรียมข้อมูล (Data Preprocessing)
6. การแบ่งกลุ่มข้อมูล (Clustering) ด้วยเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน (Unsupervised Learning)
7. การเปรียบเทียบประสิทธิภาพภายในแบบจำลองด้วยค่าพารามิเตอร์ที่แตกต่างกัน
8. การเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง
9. การวิเคราะห์ลักษณะลูกค้าของแต่ละแบบจำลอง
10. การวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis) ของแบบจำลองที่ดีที่สุด
11. สรุปผลการดำเนินงาน

3.1 กระบวนการสร้างแบบจำลอง

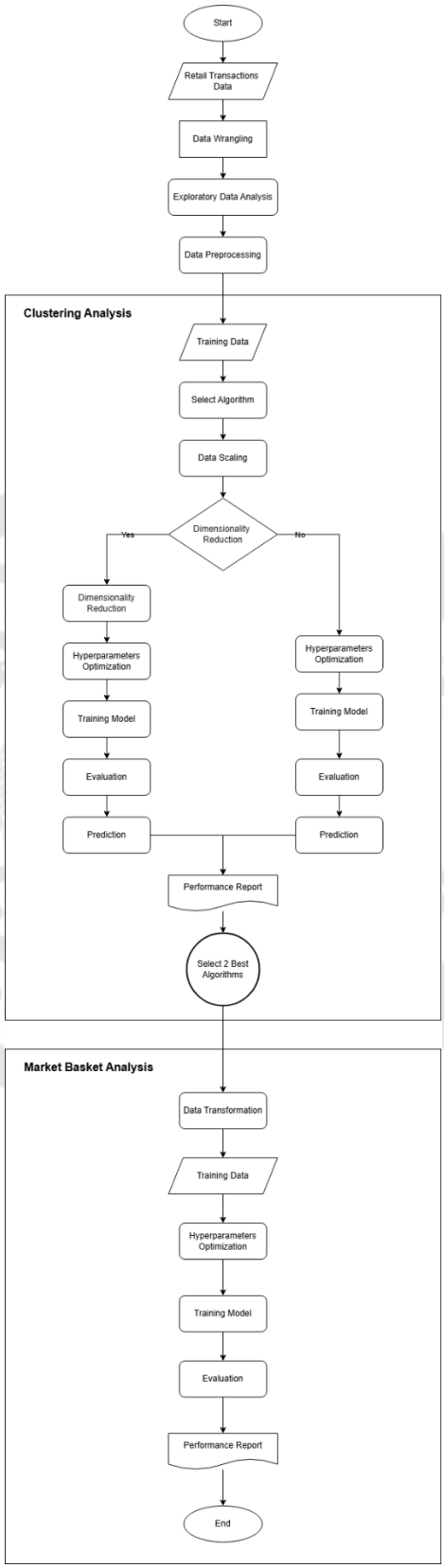
จากภาพประกอบ 6 แสดงถึงกระบวนการทำงานของแบบจำลองในงานวิจัย เริ่มต้นจากการนำเข้าข้อมูลธุรกรรมการซื้อสินค้าภายในร้านค้าปลีก จากนั้นทำการจัดระเบียบของข้อมูลด้วยการทำความสะอาดข้อมูล และเปลี่ยนแปลงประเภทและรูปแบบของข้อมูล และทำการวิเคราะห์ข้อมูลเชิงการตรวจสอบเบื้องต้นหรือ Exploratory Data Analysis ในลำดับถัดไป จากนั้นทำการเตรียมข้อมูลหรือ Data Preprocessing ด้วยการวิเคราะห์ข้อมูลผ่านแบบจำลอง RFM และนำผลลัพธ์ที่ได้จากการวิเคราะห์ไปทำการสุ่มข้อมูลตัวอย่างด้วยเทคนิค Random Sampling เพื่อให้ทรัพยากรที่ใช้สามารถประมวลผลข้อมูลจนแล้วเสร็จการวิจัยได้ เนื่องจากข้อมูลที่ใช้มีขนาดใหญ่เกินกว่าที่จะทำการวิจัยในทุกเทคนิคและแบบจำลอง

ขั้นตอนถัดไป คือ ส่วนของกระบวนการการแบ่งกลุ่มข้อมูล เริ่มต้นจากการเลือกใช้แบบจำลองสำหรับการแบ่งกลุ่มข้อมูล โดยในงานวิจัยนี้ได้เลือกใช้แบบจำลองทั้งหมด 4 แบบจำลอง ได้แก่ แบบจำลอง K-Means Clustering, แบบจำลอง DBSCAN, แบบจำลอง GMM, และแบบจำลอง Agglomerative Clustering เพื่อให้การแบ่งกลุ่มข้อมูลของแบบจำลองมี

ประสิทธิภาพที่ดียิ่งขึ้น จึงทำการทดลองปรับจูนค่าพารามิเตอร์ในค่าที่แตกต่างกัน การปรับช่วงของข้อมูล และการลดจำนวนมิติของข้อมูล โดยสามารถแบ่งข้อมูลที่ใช้ในการวิเคราะห์ในแบบจำลองการแบ่งกลุ่มข้อมูลได้ 2 รูปแบบ คือ ข้อมูลที่มีการปรับช่วงของข้อมูลร่วมกับการลดจำนวนมิติของข้อมูล และข้อมูลที่มีการปรับช่วงของข้อมูล แต่ไม่มีการลดจำนวนมิติของข้อมูล

เมื่อทำฝึกสอนแบบจำลองเรียบร้อยแล้ว ทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Silhouette Score และทำนายกลุ่มข้อมูลของแต่ละข้อมูล โดยนำข้อมูลการวัดประสิทธิภาพของทุกแบบจำลองในแต่ละรูปแบบข้อมูล มาทำการเปรียบเทียบประสิทธิภาพภายในแบบจำลองด้วยค่าพารามิเตอร์ที่แตกต่างกัน เพื่อคัดเลือกแบบจำลองที่ให้ประสิทธิภาพสูงสุดของแต่ละรูปแบบของข้อมูล คือ ข้อมูลที่มีการลดจำนวนมิติของข้อมูล และข้อมูลไม่มีการลดจำนวนมิติของข้อมูล สุดท้ายเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง โดยพิจารณาจากผลลัพธ์ของแบบจำลองทั้งหมด และคัดเลือกให้เหลือเพียง 2 แบบจำลอง จากทั้งหมด 8 แบบจำลอง เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพดีที่สุดในกระบวนการแบ่งกลุ่มข้อมูล

ขั้นตอนสุดท้าย คือ ส่วนของกระบวนการการวิเคราะห์ตระกร้าสินค้า เริ่มต้นจากการนำข้อมูลจาก 2 แบบจำลองที่ได้รับการคัดเลือก มาทำการเปลี่ยนรูปแบบของข้อมูลหรือ Data Transformation เพื่อให้เหมาะสมกับการนำมาใช้ในการฝึกสอนแบบจำลอง โดยทดลองปรับค่าพารามิเตอร์ ด้วยการใช้ค่า Support, Confidence, และ Lift เป็นเมทริกซ์ที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง เมื่อทำการฝึกสอนแบบจำลองแล้ว ได้ทำการประเมินผลลัพธ์ภายใต้ค่าพารามิเตอร์ที่กำหนด และสรุปผลการวิเคราะห์ทั้งหมดในตอนท้ายของกระบวนการวิจัย



ภาพประกอบ 6 แสดงกระบวนการสร้างแบบจำลอง

3.2 **การเก็บรวบรวมข้อมูล** (Data Collection)

การศึกษาในครั้งนี้ ผู้วิจัยได้นำข้อมูล Retail Transaction Dataset จากแหล่งข้อมูลสาธารณะบนเว็บไซต์ Kaggle.com มาใช้ในการทำวิจัย ซึ่งข้อมูลทั้งหมดถูกสร้างขึ้นมาจากไลบรารี่ (Library) ที่ชื่อว่า Python Faker โดยจัดเก็บข้อมูลธุรกรรมของร้านค้าปลีกตั้งแต่วันที่ 1 มกราคม 2020 ถึง 18 พฤษภาคม 2024 มีข้อมูลจำนวน 1,000,000 ธุรกรรม และ 13 คุณลักษณะ ดังแสดงในตาราง 4 ซึ่งข้อมูลถูกจัดเก็บอยู่ในรูปแบบตารางและบันทึกเป็นไฟล์นามสกุล CSV

ตาราง 4 แสดงคำอธิบายคุณลักษณะในชุดข้อมูล

คุณลักษณะ	**คำอธิบาย**
1. Transaction_ID	เลขที่ที่ใช้ระบุการทำธุรกรรมแต่ละรายการ
2. Date	วันที่และเวลาที่ทำธุรกรรมแต่ละครั้ง
3. Customer_Id	รหัสลูกค้าที่ทำการซื้อสินค้า
4. Product	รายการสินค้าที่ถูกซื้อในแต่ละธุรกรรม
5. Total_Items	ปริมาณของสินค้าที่ถูกซื้อในธุรกรรม
6. Total_Cost	มูลค่ารวมของการซื้อของธุรกรรม
7. Payment_Method	วิธีการชำระเงินในการทำธุรกรรม
8. City	เมืองที่เกิดการทำธุรกรรม
9. Store_Type	ประเภทของร้านค้า
10. Discount_Applied	มีการใช้ส่วนลดกับธุรกรรมนั้นหรือไม่
11. Customer_Category	ประเภทที่แสดงถึงภูมิหลังหรือกลุ่มอายุของลูกค้า
12. Season	ฤดูกาลที่เกิดการทำธุรกรรม
13. Promotion	ประเภทของโปรโมชั่นที่ใช้กับธุรกรรม

3.3 **การจัดระเบียบข้อมูล** (Data Wrangling)

3.3.1 **การทำความสะอาดข้อมูล** (Data Cleansing)

ผู้วิจัยได้ทำการสำรวจข้อมูลภายในชุดข้อมูลพบว่า มีค่าว่าง (Missing Value) ถูกจัดเก็บอยู่ในคุณลักษณะ Promotion ซึ่งเป็นคุณลักษณะที่เกี่ยวข้องกับประเภทของโปรโมชั่นที่ใช้ในแต่ละธุรกรรม เมื่อพิจารณาข้อมูลอื่นที่อยู่ในคุณลักษณะเดียวกัน ผู้วิจัยจึงตัดสินใจที่จะจัดการ

กับค่าว่างเหล่านี้ ด้วยการแทนที่ค่าว่างด้วยค่า None เนื่องจากการเติมคำว่า None เข้าไปในคุณลักษณะ Promotion จะสื่อถึงกรณีที่ไม่มีการใช้โปรโมชั่นในธุรกรรมดังกล่าว การเติมค่านี้เข้าไป จะช่วยให้ข้อมูลมีความสมบูรณ์ยิ่งขึ้นและสอดคล้องกับบริบทของคุณลักษณะนั้น

3.3.2 **การเปลี่ยนแปลงประเภทและรูปแบบของข้อมูล** (Data Transformation)

เมื่อข้อมูลถูกทำความสะอาดเรียบร้อยแล้ว จึงทำการสำรวจความถูกต้องของประเภทและรูปแบบของข้อมูล โดยพบว่า ประเภทข้อมูลของคุณลักษณะ Date นั้น เป็นประเภทแบบข้อความ (String) ซึ่งไม่ถูกต้อง ผู้วิจัยจึงได้ทำการเปลี่ยนแปลงประเภทข้อมูลของคุณลักษณะ Date ให้กลายเป็นประเภทแบบวันที่ (Date)

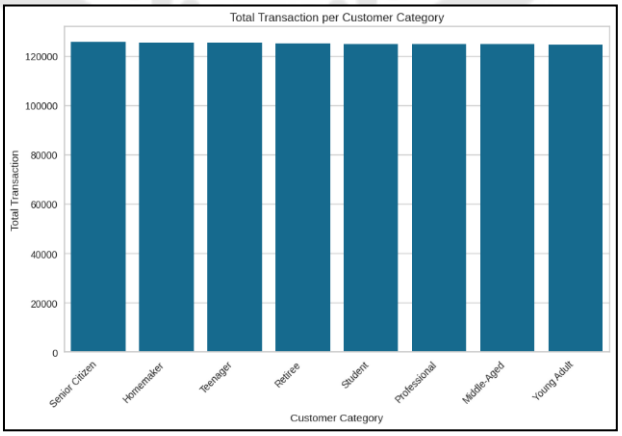
นอกจากนี้ ยังพบว่า ข้อมูลในคุณลักษณะ Product นั้น มีรูปแบบของข้อมูลที่ยากต่อการวิเคราะห์สินค้าในแต่ละรายการ ผู้วิจัยจึงตัดสินใจเปลี่ยนรูปแบบของข้อมูล จากเดิมที่ข้อมูลรายการสินค้าจะถูกจัดเก็บรวมกันอยู่ในลิสต์ (List) เปลี่ยนแปลงเป็นแยกข้อมูลรายการสินค้าของแต่ละธุรกรรมออกมาเป็นแถวละ 1 รายการสินค้า โดยที่ยังคงมีข้อมูลธุรกรรมเดิมประกอบอยู่ ซึ่งข้อมูลชุดนี้ ผู้วิจัยจะนำไปใช้ในการวิเคราะห์ข้อมูลเชิงการตรวจสอบเบื้องต้น (Exploratory Data Analysis) ของคุณลักษณะ Product

3.4 **การวิเคราะห์ข้อมูลเชิงการตรวจสอบเบื้องต้น** (Exploratory Data Analysis, EDA)

การวิเคราะห์ข้อมูลข้อมูลเชิงการตรวจสอบเบื้องต้นของแต่ละตัวแปรพบว่า ข้อมูลชุดนี้ประกอบไปด้วย 1,000,000 ธุรกรรม และ 13 คุณลักษณะ โดยจัดเก็บข้อมูลตั้งแต่วันที่ 1 มกราคม 2020 ถึง 18 พฤษภาคม 2024 เป็นระยะเวลา 3.5 ปี และเมื่อวิเคราะห์จำนวนของธุรกรรมที่เกิดขึ้นในแต่ละปีนั้น จะพบว่า จำนวนธุรกรรมที่เกิดขึ้นในแต่ละเดือน เมื่อเปรียบเทียบกับปีอื่นภายในชุดข้อมูล จะมีจำนวนธุรกรรมที่ใกล้เคียงกัน ดังแสดงในภาพประกอบ 7 ซึ่งในช่วงเดือนที่ 2 ของทุกปีจะเป็นเดือนที่มีจำนวนธุรกรรมน้อยที่สุด และมีจำนวนธุรกรรมมากที่สุดในเดือนที่ 3

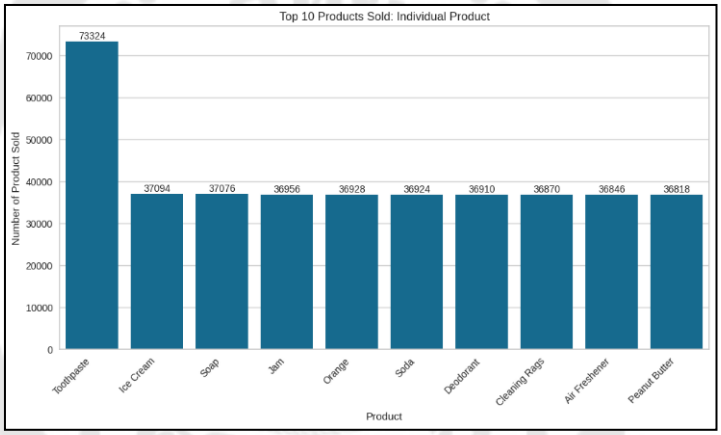
ภาพประกอบ 7 แสดงจำนวนธุรกรรมในแต่ละเดือน โดยเปรียบเทียบแต่ละปี

ข้อมูลลูกค้าที่มาใช้บริการในร้านค้าปลีกนั้น มีจำนวน 329,738 คน ซึ่งมีลูกค้าที่มีจำนวนธุรกรรมจำนวนมากที่สุดเป็นจำนวน 454 ธุรกรรม และจำนวนธุรกรรมน้อยที่สุดจำนวน 1 ธุรกรรม โดยเมื่อวิเคราะห์ข้อมูลกลุ่มลูกค้าที่เข้ามาใช้บริการ พบว่า ลูกค้ากลุ่มผู้สูงอายุ (Senior Citizen) เข้ามาใช้บริการซื้อสินค้ามากที่สุด ด้วยจำนวนธุรกรรม 125,485 ธุรกรรม คิดเป็น 12.55% รองลงมาคือ ลูกค้ากลุ่มพ่อบ้าน/แม่บ้าน (Homemaker) มีจำนวนธุรกรรม 125,418 ธุรกรรม คิดเป็น 12.54% ส่วนลูกค้ากลุ่มที่มาใช้บริการน้อยที่สุด เป็นลูกค้ากลุ่มคนหนุ่มสาว (Young Adult) ซึ่งมีจำนวนธุรกรรม 124,577 ธุรกรรม คิดเป็น 12.46% จากข้อมูลที่แสดงดังในภาพประกอบ 8 แสดงให้เห็นว่า สัดส่วนจำนวนธุรกรรมของลูกค้าแต่ละกลุ่มนั้น ไม่ได้มีจำนวนที่แตกต่างกันอย่างมีนัยสำคัญ



ภาพประกอบ 8 แสดงจำนวนธุรกรรมของลูกค้าแต่ละกลุ่ม

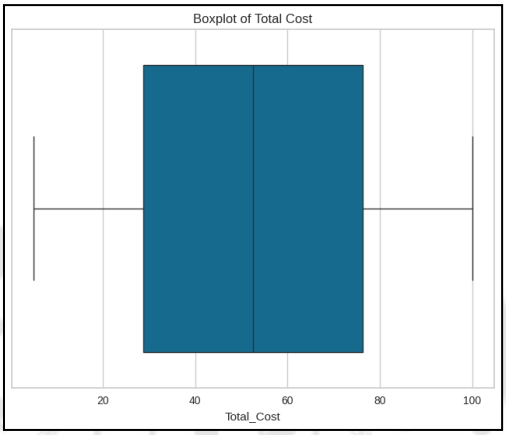
ด้านของข้อมูลสินค้า พบว่า ในชุดข้อมูลประกอบไปด้วยรายการสินค้าทั้งหมด 81 รายการ โดยสินค้าที่ได้รับความนิยมสูงที่สุด เมื่อพิจารณาจากจำนวนครั้งที่สินค้านั้นเกิดขึ้นในแต่ละธุรกรรม พบว่า แปรงสีฟัน (Toothpaste) เป็นสินค้าที่ถูกซื้อมากที่สุด มีจำนวนการสั่งซื้อถึง 73,324 ครั้งในธุรกรรมทั้งหมด ส่วนสินค้าที่มีจำนวนการสั่งซื้อน้อยที่สุดคือ ถุงขยะ (Trash Bags) ซึ่งมีจำนวนการซื้อ 36,187 ครั้ง แต่เมื่อวิเคราะห์สัดส่วนของจำนวนการสั่งซื้อในแต่ละสินค้า พบว่า สินค้าชนิดอื่นที่ไม่ใช่แปรงสีฟันนั้น มีจำนวนการซื้อที่ใกล้เคียงกัน ซึ่งเมื่อพิจารณาจากไอศครีม (Ice Cream) ที่เป็นสินค้าที่มีจำนวนการสั่งซื้อมากที่สุดเป็นอันดับ 2 โดยมีจำนวนธุรกรรม 37,094 ธุรกรรม เปรียบเทียบกับถุงขยะ (Trash Bags) พบว่า สัดส่วนของสินค้าทั้งสองชนิดมีสัดส่วนที่เท่ากับคือ 0.012 ของจำนวนสินค้าในธุรกรรมทั้งหมด ดังแสดงในภาพประกอบ 9



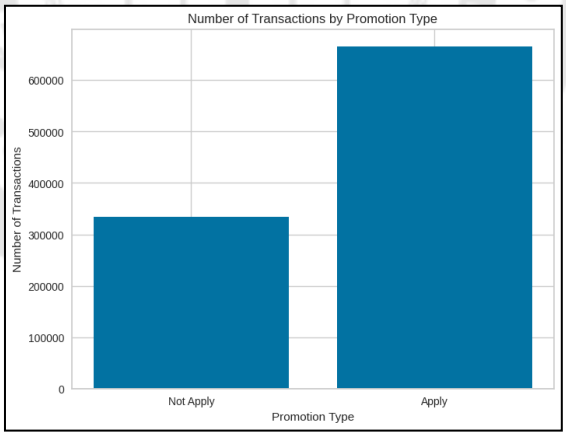
ภาพประกอบ 9 แสดงจำนวนสินค้าที่ขายดี 10 อันดับแรกของร้านค้าปลีก

การซื้อสินค้าของลูกค้าในแต่ละครั้ง โดยเฉลี่ยลูกค้าซื้อสินค้าจำนวน 5-6 ชิ้น และมียอดใช้จ่ายเฉลี่ยอยู่ที่ประมาณ 52.45 ต่อธุรกรรม ดังแสดงข้อมูลในภาพประกอบ 10 ซึ่งวิธีการชำระเงินที่ลูกค้าเลือกใช้นั้น ไม่ได้มีความแตกต่างกันอย่างมีนัยสำคัญ คือ ลูกค้าเลือกชำระด้วยเงินสด (Cash) จำนวน 250,230 ธุรกรรม คิดเป็น 25.02%, ชำระด้วยบัตรเดบิต (Debit Card) จำนวน 250,074 ธุรกรรม คิดเป็น 25.01%, ชำระด้วยบัตรเครดิต (Credit Card) จำนวน 249,985 ธุรกรรม คิดเป็น 25.00%, และชำระด้วยการโอนชำระผ่านโทรศัพท์มือถือ (Mobile Payment) จำนวน 249,711 ธุรกรรม คิดเป็น 24.97% รวมถึงการใช้ส่วนลดในแต่ละธุรกรรมนั้น ก็มีจำนวนที่ใกล้เคียงกัน คือ ธุรกรรมที่ใช้ส่วนลดจำนวน 500,104 ธุรกรรม และธุรกรรมที่ไม่มีการใช้ส่วนลดจำนวน 499,896 ธุรกรรม และเมื่อวิเคราะห์ข้อมูลการใช้โปรโมชั่นในธุรกรรม พบว่า สัดส่วนธุรกรรมที่ไม่มี

การใช้โปรโมชั่น เมื่อเปรียบกับกับการใช้โปรโมชั่นในการลดราคาสินค้าที่กำหนด (Discount on Selected Items) หรือโปรโมชั่นซื้อ 1 แถม 1 (Buy One Get One) มีจำนวนที่น้อยกว่า โดยมี 333,943 ธุรกรรม คิดเป็น 33.39% ส่วนธุรกรรมที่มีการใช้โปรโมชั่นนั้น มีจำนวน 666,057 ธุรกรรม คิดเป็น 66.61% ดังแสดงในภาพประกอบ 11



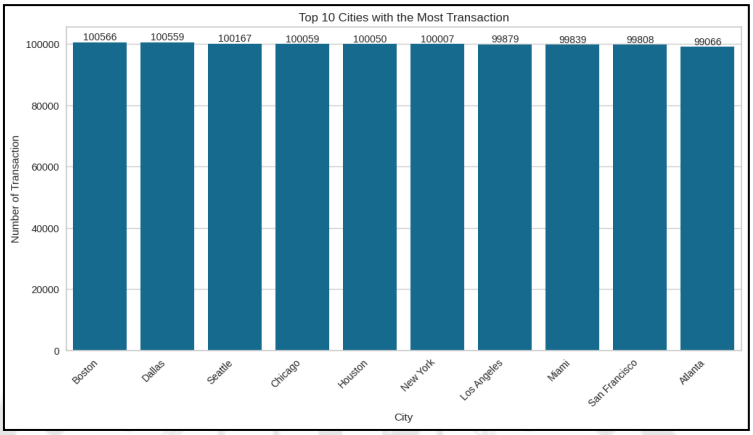
ภาพประกอบ 10 แสดงข้อมูลของคุณลักษณะมูลค่ารวมของการซื้อของธุรกรรม



ภาพประกอบ 11 แสดงจำนวนธุรกรรมที่มีการใช้โปรโมชั่น เปรียบเทียบกับไม่มีการใช้โปรโมชั่น

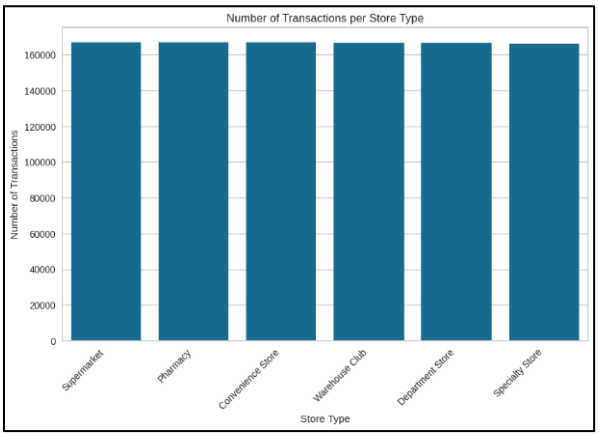
ข้อมูลร้านค้าปลีกในแต่ละเมือง แสดงให้เห็นว่า เมืองที่มีลูกค้าไปใช้บริการมากที่สุด ได้แก่ เมือง Boston ซึ่งมีลูกค้าไปใช้บริการมากถึงจำนวน 100,566 ธุรกรรม คิดเป็น 10.06% รองลงมาคือ เมือง Dallas มีจำนวนธุรกรรม 100,559 ธุรกรรม คิดเป็น 10.06% เช่นกัน ส่วนเมืองที่มีลูกค้าไปใช้บริการน้อยที่สุด ได้แก่ เมือง Atlanta ซึ่งมีจำนวนธุรกรรม 99,066 ธุรกรรม คิดเป็น 9.91% โดยหากเมื่อพิจารณาความแตกแต่งของจำนวนธุรกรรมที่เกิดขึ้นระหว่างเมือง Dallas ที่

เป็นเมืองที่มีจำนวนธุรกรรมมากที่สุด และเมือง Atlanta ที่มีจำนวนธุรกรรมน้อยที่สุด พบว่า จำนวนธุรกรรมที่แตกต่างของทั้งสองเมือง เท่ากับ 1,500 ธุรกรรม คิดเป็น 0.15% และเมื่อวิเคราะห์ข้อมูลจากกราฟที่แสดงจำนวนธุรกรรมของที่เกิดขึ้นในแต่ละเมือง ดังแสดงในภาพประกอบ 12 แสดงให้เห็นว่า ร้านค้าปลีกในแต่ละเมืองนั้น มีจำนวนธุรกรรมที่เกิดขึ้นใกล้เคียงกันมาก ซึ่งไม่ได้แสดงถึงความแตกต่างกันที่ชัดเจน



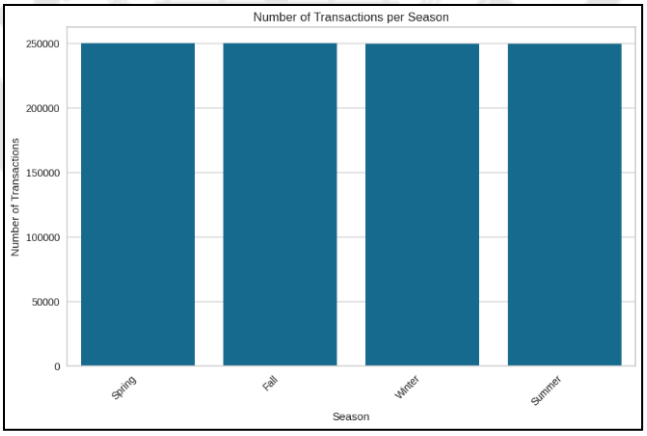
ภาพประกอบ 12 แสดงจำนวนธุรกรรมที่เกิดขึ้นในร้านค้าปลีกแต่ละเมือง

เมื่อทำการวิเคราะห์ข้อมูลประเภทของร้านค้าปลีก ก็พบว่า ร้านค้าปลีกที่เป็นซูเปอร์มาร์เก็ต (Supermarket) ได้รับความนิยมจากลูกค้ามากที่สุด เนื่องจากมีจำนวนธุรกรรมเท่ากับ 166,936 ธุรกรรม คิดเป็น 16.69% รองลงมาเป็นร้านปลีกที่เป็นร้านขายยา (Pharmacy) มีจำนวนธุรกรรม 166,915 ธุรกรรม คิดเป็น 16.69% ส่วนร้านค้าปลีกที่มีจำนวนธุรกรรมน้อยเป็นร้านค้าปลีกที่เป็นร้านค้าพิเศษ (Specialty Store) มีจำนวนธุรกรรม 166,101 ธุรกรรม คิดเป็น 16.61% ซึ่งเมื่อวิเคราะห์พบว่า จำนวนธุรกรรมของร้านค้าปลีกประเภทซูเปอร์มาร์เก็ต (Supermarket) และร้านค้าปลีกประเภทร้านค้าพิเศษ (Specialty Store) มีจำนวนธุรกรรมที่แตกต่างกันเพียง 835 ธุรกรรม หรือคิดเป็น 0.08% ซึ่งแสดงถึงใกล้เคียงกันอย่างมากของจำนวนธุรกรรมที่เกิดขึ้นในร้านค้าปลีกแต่ละประเภท ดังแสดงในภาพประกอบ 13



ภาพประกอบ 13 แสดงจำนวนธุรกรรมในร้านค้าปลีกแต่ละประเภท

ในด้านข้อมูลของฤดูกาล พบว่า ฤดูกาลที่ลูกค้ามาใช้บริการมากที่สุด เป็นฤดูใบไม้ผลิ (Spring) มีจำนวนธุรกรรม 250,368 คิดเป็น 25.04% ซึ่งรองลงมาจะเป็นฤดูใบไม้ร่วง (Fall) มีจำนวนธุรกรรม 250,248 ธุรกรรม คิดเป็น 25.03% ฤดูหนาว (Winter) มีจำนวนธุรกรรม 249,763 ธุรกรรม คิดเป็น 24.97% และฤดูร้อน (Summer) มีจำนวนธุรกรรม 249,621 ธุรกรรม คิดเป็น 24.96% ตามลำดับ จากภาพประกอบ 14 แสดงให้เห็นว่า จำนวนธุรกรรมที่เกิดขึ้นในแต่ละฤดูกาลนั้น มีจำนวนธุรกรรมที่เกิดขึ้นใกล้เคียงกัน



ภาพประกอบ 14 แสดงจำนวนธุรกรรมในแต่ละฤดูกาล

3.5 **การเตรียมข้อมูล** (Data Preprocessing)

3.5.1 **การวิเคราะห์ข้อมูลด้วยแบบจำลอง** RFM

ผู้วิจัยได้เริ่มต้นด้วยการคัดเลือกข้อมูลคุณลักษณะที่เกี่ยวข้องจากชุดข้อมูล โดยเลือกใช้ข้อมูลวันที่ทำธุรกรรม, เลขที่ของธุรกรรม, และมูลค่าของการซื้อสินค้า หลังจากนั้น จึงจัด

กลุ่มข้อมูลตามรหัสลูกค้าเพื่อนำไปวิเคราะห์ด้วยแบบจำลอง RFM โดยจะมีรายละเอียดในการคำนวณ ดังนี้

1. **การคำนวณ** Recency (R)

ผู้วิจัยคำนวณระยะเวลาที่ผ่านไปตั้งแต่การซื้อสินค้าครั้งล่าสุดของลูกค้า โดยนำวันที่ทำธุรกรรมล่าสุดของลูกค้าแต่ละรายมาลบกับวันที่ล่าสุดในชุดข้อมูล ซึ่งใช้เป็นวันที่ที่กำหนดเป็นจุดอ้างอิง เพื่อที่จะระบุว่าลูกค้ามาใช้บริการซื้อสินค้าล่าสุดเป็นระยะเวลาเท่าไหร่

2. **การคำนวณ** Frequency (F)

ผู้วิจัยนับจำนวนครั้งที่ลูกค้าทำธุรกรรมในช่วงเวลาที่กำหนด โดยนับจากจำนวนเลขที่ธุรกรรมที่เกิดขึ้นของแต่ละลูกค้า เพื่อระบุว่าลูกค้ามีความถี่ในการซื้อสินค้าอย่างไรในช่วงระยะเวลาที่วิเคราะห์

3. **การคำนวณ** Monetary (M)

ผู้วิจัยสรุปมูลค่าการซื้อทั้งหมดของลูกค้าแต่ละราย โดยคำนวณหายอดรวมของมูลค่าของการซื้อสินค้า เพื่อระบุว่าลูกค้ามีการใช้จ่ายเงินโดยรวมมากน้อยเพียงใด

หลังจากดำเนินการตามกระบวนการคำนวณ RFM แล้ว ทำให้ได้ผลลัพธ์เป็นข้อมูลจำนวน 329,728 รายการ ซึ่งแต่ละรายการเป็นข้อมูลของลูกค้าแต่ละรายในชุดข้อมูล โดยประกอบไปด้วยคุณลักษณะทั้งสาม คือ Recency, Frequency, และ Monetary ดังตัวอย่างที่แสดงในภาพประกอบ 15

Recency	Frequency	Monetary
249	4	279.85
836	3	215.32
33	12	772.36
393	3	118.52
56	5	236.76

ภาพประกอบ 15 แสดงตัวอย่างข้อมูลที่ได้หลังจากทำ RFM Analysis

ข้อมูลสถิติที่ได้จากการวิเคราะห์ข้อมูลที่ได้จากการทำ RFM ดังแสดงข้อมูลในตาราง 5 แสดงให้เห็นว่า พฤติกรรมของลูกค้าที่มาใช้บริการนั้น มีความหลากหลาย จากค่า Recency ที่มีค่าเฉลี่ยอยู่ที่ 600.50 วัน แสดงให้เห็นว่า ลูกค้าส่วนใหญ่ไม่ได้มาใช้บริการซื้อสินค้าเป็นระยะ

เวลานานเกือบ 2 ปี ซึ่งมีลูกค้าเพียง 25% ที่มาใช้บริการในช่วง 192 วันที่ผ่านมา และอาจมีลูกค้าบางรายที่กลายเป็นลูกค้าที่สูญเสียไปของร้านค้าปลีก โดยพิจารณาจากค่าสูงสุดเท่ากับ 1,599 วัน สำหรับค่า Frequency พบว่า ลูกค้าส่วนใหญ่มียอดการทำธุรกรรมเฉลี่ยเพียง 3 ครั้ง โดยมีลูกค้าส่วนมากที่ทำธุรกรรมเพียงครั้งเดียว แต่ยังมีลูกค้าส่วนน้อยที่ทำธุรกรรมสูงถึง 454 ครั้ง ซึ่งอาจเป็นกลุ่มลูกค้าที่มีความภักดีต่อแบรนด์ ในส่วนของค่า Monetary พบว่า ลูกค้ามียอดใช้จ่ายเฉลี่ย 159.08 โดย 75% ของลูกค้าใช้จ่ายต่ำกว่า 156.30 แต่ยังมีกลุ่มลูกค้าบางรายที่มียอดใช้จ่ายสูงถึง 23,768.53 แสดงให้เห็นถึงความแตกต่างในพฤติกรรมการใช้จ่ายของลูกค้าแต่ละราย

ตาราง 5 แสดงค่าทางสถิติของข้อมูลที่ได้จากการทำ RFM Analysis

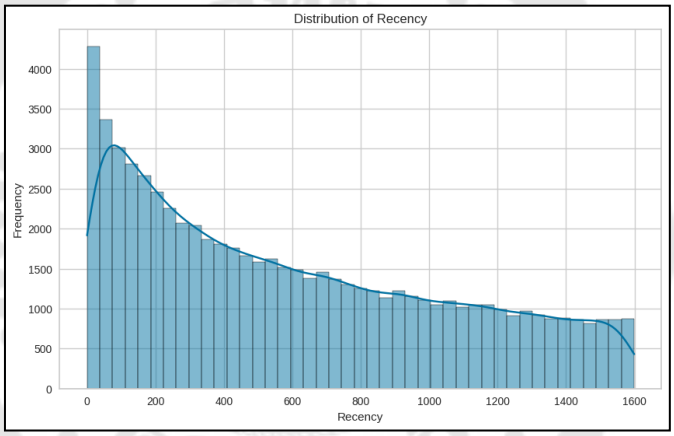
ค่าทางสถิติ	Recency	Frequency	Monetary
จำนวนข้อมูล (Number of Record)	329,738	329,738	329,738
ค่าเฉลี่ย (Mean)	600.50	3.03	159.08
ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)	460.07	6.38	338.88
ค่าน้อยที่สุด (Minimum)	0	1	5
ควอไทล์ที่ 25 (25^{th} Quartile)	192	1	46.59
ค่ามัธยฐาน (50^{th} Quartile)	503	1	83.05
ควอไทล์ที่ 75 (75^{th} Quartile)	957	3	156.30
ค่าสูงสุด (Maximum)	1,599	454	23,768.53

และเมื่อพิจารณาค่าความเบ้ (Skewness) และความโค้ง (Kurtosis) ของคุณลักษณะทั้งสาม ดังแสดงข้อมูลในตาราง 6 พบว่า มีลักษณะการกระจายข้อมูลที่แตกต่างกันอย่างชัดเจน คุณลักษณะ Frequency และคุณลักษณะ Monetary ดังแสดงในภาพประกอบ 16 และภาพประกอบ 17 ตามลำดับ มีค่าความเบ้บวกสูงหรือ Extremely Skewed Distribution คือ 14.680 และ 14.559 และค่าความโด่งที่สูงมากหรือ Leptokurtic Distribution คือ 425.799 และ 431.393 ซึ่งสะท้อนถึงการกระจายข้อมูลที่มีลักษณะปลายหางหนา (Heavy-Tails) และมีข้อมูลที่มีค่าผิดปกติ (Outlier) จำนวนมาก ส่วนคุณลักษณะ Recency ดังแสดงในภาพประกอบ 18 มีค่าความเบ้บวกต่ำกว่าหรือ Moderately Skewed Distribution ซึ่งมีค่าเท่ากับ 0.513 และค่าความ

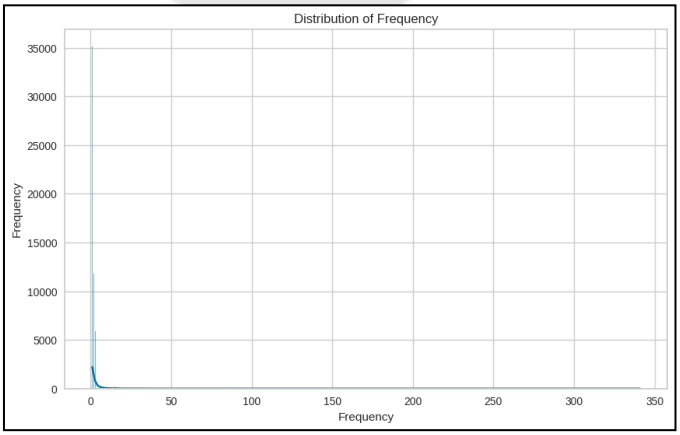
โด่งที่ต่ำกว่าปกติหรือ Platykurtic Distribution มีค่า -0.936 แสดงถึงลักษณะการกระจายข้อมูลที่สมดุลมากขึ้น และไม่แสดงความผิดปกติที่ชัดเจนเหมือนคุณลักณะอื่น

ตาราง 6 แสดงค่าความเบ้และค่าความโด่งของข้อมูลที่ได้จากการทำ RFM Analysis

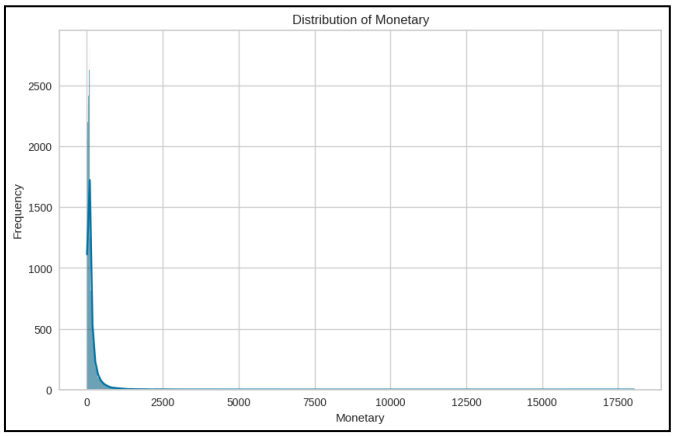
	Skewness	Kurtosis
Recency	0.512608	-0.936437
Frequency	14.680297	425.798537
Monetary	14.559395	431.393440



ภาพประกอบ 16 แสดงการกระจายตัวของข้อมูลคุณลักษณะ Recency



ภาพประกอบ 17 แสดงการกระจายตัวของข้อมูลคุณลักษณะ Frequency



ภาพประกอบ 18 การกระจายตัวของข้อมูลคุณลักษณะ Monetary

3.5.2 **การสุ่มข้อมูลตัวอย่างด้วยเทคนิค** Random Sampling

เนื่องจากข้อมูลที่มีอยู่ในตอนนี้นั้น มีขนาดใหญ่เกินกว่าที่ทรัพยากรที่ใช้ในการประมวลผลจะดำเนินการจนสำเร็จการวิจัยได้ ผู้วิจัยจึงทำการลดขนาดของชุดข้อมูลลง เพื่อให้สามารถประมวลผลจนแล้วเสร็จการวิจัยได้ โดยเลือกใช้วิธีการสุ่มตัวอย่างแบบสุ่ม (Random Sampling) แต่เนื่องจากการศึกษาในครั้งนี้ ล้วนใช้เทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน จึงทำให้ไม่สามารถกำหนดสัดส่วนของข้อมูลตามสัดส่วนของกลุ่มเป้าหมาย (Label) ได้

ผู้วิจัยจึงได้ทดลองทำการสุ่มตัวอย่าง โดยทดลองปรับลดสัดส่วนของข้อมูลลงครั้งละ 5% จนกว่าทรัพยากรที่ใช้ในการประมวลผลสามารถดำเนินการเสร็จสิ้น จนได้เป็นผลลัพธ์ของแต่ละแบบจำลองออกมาได้ ซึ่งสัดส่วนของข้อมูลที่เหมาะสมสุดที่จะสามารถดำเนินการได้ คือ สัดส่วน 20% ของจำนวนข้อมูลที่ได้จากการวิเคราะห์จากแบบจำลอง RFM ทั้งหมด โดยจะทำให้ข้อมูลที่นำไปใช้ในแบบจำลองนั้น มีจำนวน 65,947 รายการ และ 3 คุณลักษณะ

3.6 **การแบ่งกลุ่มข้อมูล** (Clustering) **ด้วยเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน** (Unsupervised Learning)

ข้อมูลที่ได้จากการวิเคราะห์ในแบบจำลอง RFM เป็นข้อมูลที่ถูกนำเข้าไปยังแต่ละแบบจำลอง โดยผู้วิจัยจะทำการปรับช่วงของข้อมูลให้มีค่าเฉลี่ย (Mean) เท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เท่ากับ 1 ด้วยการใช้เทคนิค Standardization เพื่อช่วยลดความต่างกันระหว่างข้อมูลในแต่ละคุณลักษณะ ซึ่งอาจส่งผลต่อการวิเคราะห์ของแบบจำลอง รวมถึงทำการทดลองสร้างแบบจำลองจากข้อมูล 2 รูปแบบ คือ 1.ข้อมูลที่ผ่านการลดจำนวนมิติ และ 2.ข้อมูลที่ไม่ได้ผ่านการลดจำนวนมิติ พร้อมทั้งทำการปรับค่าพารามิเตอร์ของแต่ละ

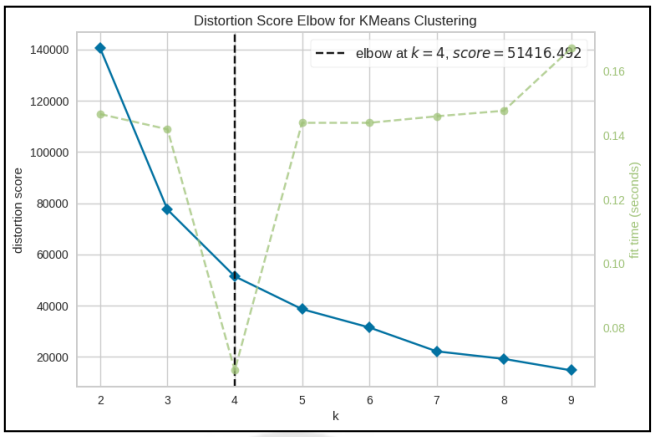
แบบจำลอง เพื่อค้นหาค่าที่ดีที่สุด ที่ทำให้แบบจำลองมีความสามารถแบ่งกลุ่มได้อย่างมีประสิทธิภาพมากที่สุด

การลดจำนวนมิติของข้อมูลในงานวิจัยนี้ ได้นำเทคนิค PCA มาใช้ในการดำเนินการ ซึ่งออกแบบให้ข้อมูลหลังจากการลดจำนวนมิตินั้น เหลือเพียง 2 มิติ เพื่อเพิ่มความสะดวกในการนำเสนอผลลัพธ์ในรูปแบบแผนภูมิ และช่วยให้การวิเคราะห์ข้อมูลและการสื่อสารผลลัพธ์มีความชัดเจนและเข้าใจง่ายยิ่งขึ้น การกำหนดจำนวนมิติที่ลดลงเป็น 2 มิตินี้ถูกนำไปใช้ในทุกการทดลองและแบบจำลองที่ศึกษา เพื่อให้เกิดความสม่ำเสมอและสามารถเปรียบเทียบผลลัพธ์ระหว่างแบบจำลองต่างๆ ได้อย่างมีประสิทธิภาพ

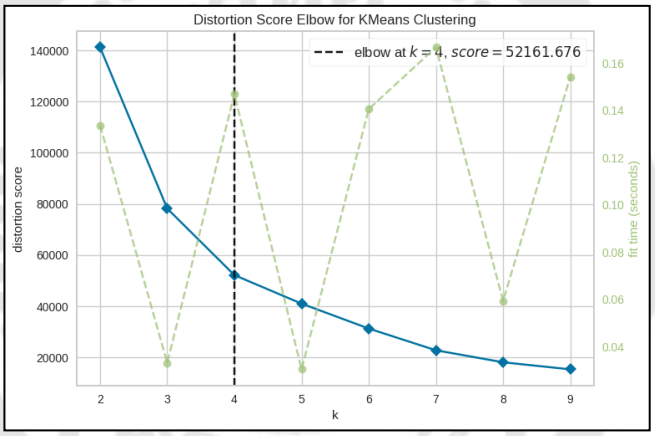
โดยแต่ละขั้นตอนที่ดำเนินการ ตั้งแต่ปรับช่วงของข้อมูล การลดจำนวนมิติของข้อมูล (กรณีที่มีทำ) ตลอดจนถึงการสร้างแบบจำลองที่ใช้ในการแบ่งกลุ่มข้อมูล ผู้วิจัยใช้เครื่องมือไปป์ไลน์ (Pipeline) มาช่วยในการดำเนินการและกำหนดขั้นตอนของแต่ละกระบวนการ เพื่อให้สะดวกต่อการสร้างแบบจำลองที่มีรูปแบบข้อมูลที่นำเข้าแตกต่างกัน รวมถึงการประเมินประสิทธิภาพของแต่ละแบบจำลอง

3.6.1 **แบบจำลอง** K-Means Clustering

ผู้วิจัยได้นำเทคนิค Elbow Method หนึ่งเทคนิคที่ได้รับความนิยมในการนำมาใช้เพื่อหาจำนวนคลัสเตอร์ที่เหมาะสมที่สุดที่จะนำใช้ในการกำหนดค่า K ของแบบจำลอง K-Means Clustering มาใช้ในการทดลองของงานวิจัยนี้ โดยเลือกใช้เมทริกซ์ Distortion Score ในการประเมินผลและค้นหาจำนวนคลัสเตอร์ที่เหมาะสม จากการทดลองนี้ แสดงให้เห็นว่า จำนวนคลัสเตอร์ที่เหมาะสมที่สุด เมื่อนำข้อมูลที่มีการลดจำนวนมิติของข้อมูลไปวิเคราะห์ผ่านแบบจำลอง คือ 4 คลัสเตอร์ ดังแสดงในภาพประกอบ 19 ส่วนด้านข้อมูลที่ไม่มีการลดจำนวนมิติของข้อมูล พบว่าจำนวนคลัสเตอร์ที่เหมาะสมที่สุด ที่นำไปวิเคราะห์ผ่านแบบจำลองนั้น มีจำนวนเท่ากับ 4 คลัสเตอร์ แสดงในภาพประกอบ 20



ภาพประกอบ 19 แสดงผลลัพธ์จากการทำ Elbow Method ของข้อมูลที่มีการลดจำนวนมิติ



ภาพประกอบ 20 แสดงผลลัพธ์จากการทำ Elbow Method ของข้อมูลที่ไม่มีการลดจำนวนมิติ

3.6.2 แบบจำลอง Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

ผู้วิจัยดำเนินการทดลองในการปรับค่าพารามิเตอร์ของแบบจำลอง Density-Based Spatial Clustering of Applications with Noise (DBSCAN) โดยปรับเฉพาะค่า Epsilon (Eps) ซึ่งเป็นค่าพารามิเตอร์ที่กำหนดระยะทางสูงสุดที่ใช้ในการพิจารณาว่าจุดข้อมูลใดอยู่ในบริเวณเดียวกัน (Neighborhood) และทำหน้าที่ควบคุมความหนาแน่นขั้นต่ำสำหรับการสร้างคลัสเตอร์ในแบบจำลอง โดยทำการทดลองปรับในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 0.2, 0.5 และ 1 นอกจากนี้ ผู้วิจัยยังมีการกำหนดค่า min_samples เท่ากับ 10 เพื่อให้คลัสเตอร์ที่ได้มีความหนาแน่นมากขึ้น ช่วยลดการเกิดคลัสเตอร์ขนาดเล็กที่อาจไม่มีนัยสำคัญต่อการวิเคราะห์ โดยทำ

การทดลองทั้งในแบบจำลองที่ใช้ข้อมูลที่ผ่านการลดจำนวนมิติ และแบบจำลองที่ใช้ข้อมูลที่ไม่ได้ลดจำนวนมิติ

3.6.3 **แบบจำลอง** Gaussian Mixture Model (GMM)

ผู้วิจัยได้ดำเนินการทดลองในการปรับค่าพารามิเตอร์ของแบบจำลอง Gaussian Mixture Model (GMM) โดยปรับเฉพาะค่า n_component ซึ่งเป็นค่าพารามิเตอร์ที่กำหนดจำนวนของ Gaussian Distribution หรือจำนวนคลัสเตอร์ที่ใช้ในการสร้างแบบจำลอง โดยทำการทดลองปรับในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 3, 4 และ 5 ทั้งในแบบจำลองที่ใช้ข้อมูลที่ผ่านการลดจำนวนมิติ และแบบจำลองที่ใช้ข้อมูลที่ไม่ได้ลดจำนวนมิติ

3.6.4 **แบบจำลอง** Agglomerative Clustering

ผู้วิจัยได้ดำเนินการทดลองในการปรับค่าพารามิเตอร์ของแบบจำลอง Agglomerative Clustering โดยปรับเฉพาะค่า n_clusters ซึ่งเป็นค่าพารามิเตอร์ที่กำหนดจำนวนคลัสเตอร์ในขั้นตอนสุดท้ายของแบบจำลอง โดยทำการทดลองปรับในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 2, 3 และ 4 ทั้งในแบบจำลองที่ใช้ข้อมูลที่ผ่านการลดจำนวนมิติ และแบบจำลองที่ใช้ข้อมูลที่ไม่ได้ลดจำนวนมิติ

3.7 การเปรียบเทียบประสิทธิภาพภายในแบบจำลองด้วยค่าพารามิเตอร์ที่แตกต่างกัน

ในการประเมินประสิทธิภาพของแบบจำลองในงานวิจัยนี้ ได้เลือกค่า Silhouette Score เป็นเมทริกซ์หลักในการประเมิน โดยหลังจากทำการสร้างแบบจำลองที่แบ่งกลุ่มข้อมูลลูกค้าแล้วตลอดจนวัดประสิทธิภาพของแบบจำลองและได้ค่า Silhouette Score ของแบบจำลองในแต่ละเงื่อนไขออกมา ผู้วิจัยทำการเปรียบเทียบผลลัพธ์ที่ได้ภายในแบบจำลอง ระหว่างแบบจำลองที่ใช้ข้อมูลที่มีการลดจำนวนมิติ และแบบจำลองที่ไม่ได้ใช้ข้อมูลที่มีการลดจำนวนมิติ โดยทำการเลือกแบบจำลองที่มีค่า Silhouette Score สูงที่สุด ซึ่งสะท้อนถึงประสิทธิภาพที่ดีที่สุดในการแบ่งกลุ่มและช่วยให้ได้แบบจำลองที่มีประสิทธิภาพสูงสุดสำหรับข้อมูลในแต่ละรูปแบบ ทำให้จำนวนของแบบจำลองที่มีในการทดลองทั้งหมดจะลดลงจาก 20 แบบจำลอง เหลือเพียง 8 แบบจำลอง

3.8 การเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง

เมื่อได้แบบจำลองที่มีประสิทธิภาพสูงที่สุดในแต่ละรูปแบบข้อมูลแล้ว พบว่า แต่ละแบบจำลองจะประกอบด้วยแบบจำลองที่ใช้ข้อมูลที่มีการลดจำนวนมิติ และแบบจำลองที่ไม่ใช้ข้อมูลที่มีการลดจำนวนมิติ ดังแสดงในภาพประกอบ 21 ในขั้นตอนนี้ ผู้วิจัยทำการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด เพื่อคัดเลือกแบบจำลองที่มีประสิทธิภาพสูงสุด โดยทำการ

พิจารณาจากค่า Silhouette Score อีกครั้ง และจะเลือกเพียง 2 แบบจำลองที่ดีที่สุดสำหรับการทำการวิเคราะห์ต่อไป

โดยจากการวิเคราะห์ผลลัพธ์ที่ได้ ผู้วิจัยได้ทำการเลือกแบบจำลอง DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ และแบบจำลอง DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ ซึ่งเป็นแบบจำลองที่มีประสิทธิภาพที่ดีที่สุด 2 อันดับแรก เมื่อเปรียบเทียบกับแบบจำลองอื่น

Model	Silhouette Score	Number of Clusters
DBSCAN without PCA	0.913339	2
DBSCAN with PCA	0.887072	3
Agglomerative Clustering without PCA	0.840405	2
Agglomerative Clustering with PCA	0.800971	2
K-Means with PCA	0.581552	4
K-Means without PCA	0.580628	4
GMM with PCA	0.147887	3
GMM without PCA	0.113903	3

ภาพประกอบ 21 แสดงประสิทธิภาพของแต่ละแบบจำลองในแต่ละรูปแบบของข้อมูล

3.9 การวิเคราะห์ลักษณะลูกค้าของแต่ละแบบจำลอง

ผู้วิจัยทำการนำข้อมูลที่ได้จากการแบ่งกลุ่มลูกค้าของแบบจำลองไปผสานกับข้อมูลที่ได้จากการทำ RFM Analysis ก่อนหน้า และทำการวิเคราะห์ลักษณะของลูกค้าแต่ละแบบจำลองจากการใช้ค่าทางสถิติ ได้แก่ ค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median) และค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ของแต่ละคลัสเตอร์ เพื่อศึกษาลักษณะเฉพาะของลูกค้าในแต่ละกลุ่มที่ได้จากการแบ่งกลุ่ม ว่ามีความแตกต่างหรือมีคุณสมบัติอย่างไร

3.10 การวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis) ของแบบจำลองที่ดีที่สุด

ผู้วิจัยดำเนินการผสานข้อมูลที่ได้จากการแบ่งกลุ่มลูกค้าของแต่ละแบบจำลองเข้ากับข้อมูลก่อนการทำ RFM Analysis จากนั้น ทำการปรับโครงสร้างข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) โดยนำอัลกอริทึม FP-Growth มาใช้ในการวิเคราะห์ และทดลองตั้งค่าพารามิเตอร์ในแต่ละแบบจำลอง ดังนี้

1. ค่า Min Support: ทดสอบค่าที่แตกต่างกัน 3 ระดับ ได้แก่ 0.01, 0.002, และ 0.001

2. ค่า Min Confidence: กำหนดค่าไว้ที่ 0.01

3. ค่า Lift: เลือกเฉพาะกฎที่มีค่า Lift มากกว่า 1

หลังจากการทดลองและค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับการแสดงกฎความสัมพันธ์ (Association Rules) ผู้วิจัยทำการเปรียบเทียบกฎความสัมพันธ์ที่ได้ในแต่ละกลุ่มลูกค้าและแต่ละแบบจำลอง จากนั้นวิเคราะห์ความสัมพันธ์ของสินค้าที่ปรากฏในแต่ละคลัสเตอร์ โดยอ้างอิงจากกฎความสัมพันธ์ที่ค้นพบ เมื่อได้ผลการวิเคราะห์ดังกล่าวแล้ว ผู้วิจัยนำเสนอคำแนะนำสำหรับลูกค้าในแต่ละกลุ่ม โดยพิจารณาจากความสัมพันธ์ระหว่างสินค้า

3.11 สรุปผลการดำเนินงาน

แบบจำลองการแบ่งกลุ่มข้อมูลด้วยเทคนิคการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน ได้ทำการแบ่งกลุ่มลูกค้าจากการข้อมูลธุรกรรมของร้านค้าปลีก ร่วมกับการใช้เทคนิคการปรับช่วงของข้อมูล รวมถึงเทคนิคการลดจำนวนมิติของข้อมูล ซึ่งทำให้ข้อมูลที่ใช้ในการวิเคราะห์ในแบบจำลองมีทั้งหมด 2 รูปแบบ คือ ข้อมูลที่มีการจำนวนมิติของข้อมูล และข้อมูลที่ไม่มีการลดจำนวนมิติของข้อมูล โดยจะใช้ค่า Silhouette Score เป็นตัววัดประสิทธิภาพของแบบจำลอง ซึ่งเลือกแบบจำลองที่มีค่าสูงที่สุด ในงานวิจัยได้ทำการเปรียบเทียบประสิทธิภาพภายในแบบจำลองด้วยค่าพารามิเตอร์ที่แตกต่างกัน และเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง ซึ่งทำการคัดเลือกแบบจำลองที่ประสิทธิภาพสูงที่สุดจำนวน 2 แบบจำลอง โดยพิจารณาจากผลลัพธ์ที่ได้จากการเปรียบเทียบ

ส่วนเทคนิคการวิเคราะห์ตระกร้าสินค้านั้น นำข้อมูลกลุ่มลูกค้าที่ได้จากการแบ่งกลุ่มร่วมกับข้อมูลธุรกรรมของลูกค้าแต่ละกลุ่มที่ได้จาก 2 แบบจำลองที่มีประสิทธิภาพสูงสุด เพื่อค้นหากฎความสัมพันธ์ของสินค้า นอกจากนี้ ได้ทำการทดลองปรับค่าพารามิเตอร์ในแต่ละแบบจำลองเพื่อเปรียบเทียบผลลัพธ์ที่ได้จากการทดลองแต่ละครั้ง

บทที่ 4
ผลการดำเนินงานวิจัย

งานวิจัยวิธีการแบบผสมผสานโดยใช้อัลกอริทึมการจัดกลุ่มและกฎความสัมพันธ์สำหรับการแนะนำสินค้าในธุรกิจค้าปลีก ซึ่งใช้ข้อมูลธุรกรรมการซื้อสินค้าภายในร้านค้าปลีกในช่วงระหว่างวันที่ 1 มกราคม 2020 ถึง 18 พฤษภาคม 2024 โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน ประกอบด้วย การจัดกลุ่มข้อมูล (Clustering) และการวิเคราะห์ตระกร้าสินค้า (Market Basket Analysis)

ผู้วิจัยได้ดำเนินการศึกษาอัลกอริทึมการจัดกลุ่ม ได้แก่ K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), และ Agglomerative Clustering โดยได้ทำการศึกษาตามกระบวนการของแต่ละอัลกอริทึม พร้อมทั้งศึกษาเทคนิคการลดจำนวนมิติของข้อมูล และค่าพารามิเตอร์ของอัลกอริทึมตลอดจนประเมินผลลัพธ์ที่ได้ของแต่ละอัลกอริทึม ตามรูปแบบข้อมูลที่ใช้ ซึ่งประกอบด้วยข้อมูลที่มีการลดจำนวนมิติ และข้อมูลที่ไม่มีการลดจำนวนมิติ รวมถึงค่าพารามิเตอร์ที่แตกต่างกัน

ทั้งนี้ ผู้วิจัยยังได้ดำเนินการศึกษาการวิเคราะห์ตระกร้าสินค้า โดยใช้ข้อมูลกลุ่มลูกค้าที่ได้จากอัลกอริทึมการจัดกลุ่มที่มีประสิทธิภาพที่ดีที่สุด 2 อันดับแรก มาวิเคราะห์ด้วยอัลกอริทึม FP-Growth โดยจะดำเนินการศึกษาและค้นหากฎความสัมพันธ์ระหว่างสินค้าของลูกค้าแต่ละกลุ่มภายใต้เงื่อนไขที่กำหนด เพื่อให้บรรลุวัตถุประสงค์ในการวิจัยที่ได้กำหนดไว้ ดังนี้

4.1. ผลลัพธ์ของอัลกอริทึมการจัดกลุ่มลูกค้า

4.1.1 อัลกอริทึม K-Means Clustering

4.1.2 อัลกอริทึม DBSCAN

4.1.3 อัลกอริทึม GMM

4.1.4 อัลกอริทึม Agglomerative Clustering

4.2. ผลลัพธ์ของกฎความสัมพันธ์จากข้อมูลการจัดกลุ่มลูกค้าภายในร้านค้าปลีก

4.2.1 ข้อมูลการจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

4.2.2 ข้อมูลการจัดกลุ่มของอัลกอริทึม Agglomerative Clustering ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

4.1 ผลลัพธ์ของอัลกอริทึมการจัดกลุ่มลูกค้า

อัลกอริทึมการจัดกลุ่มลูกค้า คือ อัลกอริทึมที่ใช้ข้อมูลคุณลักษณะของลูกค้าแต่ละรายในการวิเคราะห์และจำแนก โดยทำการจัดกลุ่มตามระดับความคล้ายคลึงกันของพฤติกรรมหรือคุณลักษณะที่มีร่วมกัน ผู้วิจัยได้คัดเลือกคุณลักษณะจากชุดข้อมูลเพื่อนำมาวิเคราะห์ด้วยแบบจำลอง RFM โดยใช้ข้อมูลวันที่ทำธุรกรรม (Date), หมายเลขธุรกรรม (Transaction_ID), และมูลค่าการซื้อสินค้า (Total_Cost) หลังจากการวิเคราะห์และจัดกลุ่มข้อมูลตามรหัสลูกค้า ได้ผลลัพธ์เป็น 3 คุณลักษณะ ได้แก่ Recency, Frequency, และ Monetary จากนั้น ชุดข้อมูลที่ได้ถูกนำมาสุ่มตัวอย่างด้วยเทคนิค Random Sampling โดยสุ่มเลือก 20% ของข้อมูลทั้งหมดที่ได้จากการวิเคราะห์ด้วยแบบจำลอง RFM ส่งผลให้มีจำนวนข้อมูลที่ใช้ในการวิเคราะห์ในลำดับถัดไปเท่ากับ 65,947 รายการ

ในงานวิจัยนี้ ผู้วิจัยได้เลือกใช้ 4 อัลกอริทึม สำหรับการจัดกลุ่มข้อมูล ได้แก่ K-Means Clustering, DBSCAN, GMM, และ Agglomerative Clustering ซึ่งแต่ละอัลกอริทึมมีหลักการทำงานและวิธีการประมวลผลที่แตกต่างกัน นอกจากนี้ ผู้วิจัยยังได้นำเทคนิค PCA ซึ่งเป็นเทคนิคที่ใช้การลดจำนวนมิติของข้อมูลมาใช้ร่วมกับอัลกอริทึมการจัดกลุ่ม เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการจัดกลุ่มข้อมูล โดยกำหนดให้จำนวนมิติใหม่ เท่ากับ 2 มิติ ซึ่งทำให้มีชุดข้อมูลที่ใช้ในการวิจัยสำหรับแต่ละอัลกอริทึม 2 ชุดข้อมูล คือ ชุดข้อมูลที่มีจำนวนมิติเท่าเดิม และชุดข้อมูลที่ผ่านการลดจำนวนมิติ

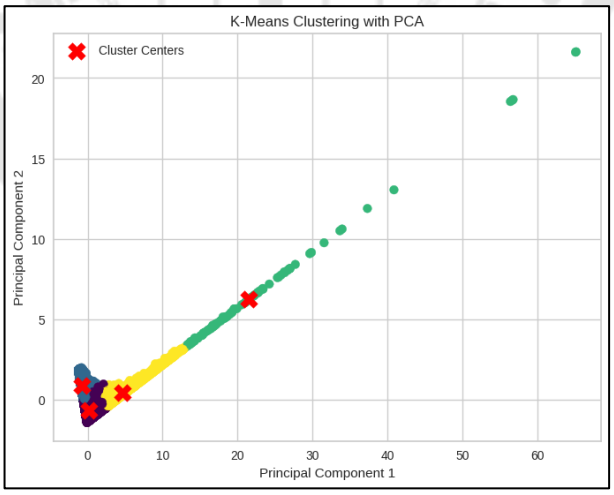
4.1.1 อัลกอริทึม K-Means Clustering

การจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม K-Means Clustering ได้นำเทคนิค Elbow Method มาใช้ในการหาจำนวนคลัสเตอร์ที่เหมาะสมที่สุดที่ใช้ในการกำหนดค่า K โดยพิจารณาจากเมทริกซ์ Distortion Score ในการประเมินผลและค้นหาจำนวนคลัสเตอร์ที่เหมาะสม จากการทดลอง พบว่า จำนวนคลัสเตอร์ที่เหมาะสมของทั้งชุดข้อมูลที่ไม่มีการลดจำนวนมิติ และชุดข้อมูลที่มีการลดจำนวนมิติ คือ 4 คลัสเตอร์ และเมื่อทำการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม K-Means Clustering ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ พบว่า คลัสเตอร์ที่ 0 มีจำนวนข้อมูลมากที่สุด คือ 38,266 และคลัสเตอร์ที่ 2 เป็นคลัสเตอร์ที่มีจำนวน 104 ซึ่งมีจำนวนน้อยที่สุด ดังแสดงข้อมูลในตารางที่ 7

ตาราง 7 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม K-Means Clustering โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	38,266
คลัสเตอร์ที่ 1	25,661
คลัสเตอร์ที่ 2	104
คลัสเตอร์ที่ 3	1,916

จากภาพประกอบที่ 22 แสดงให้เห็นว่า การจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม K-Means Clustering โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ มีการแบ่งขอบเขตคลัสเตอร์ของแต่ละคลัสเตอร์ได้อย่างชัดเจน ซึ่งจากการประเมินประสิทธิภาพของแบบจำลอง พบว่า ค่าที่ได้จากการทดสอบด้วยเมทริกซ์ Silhouette Score มีค่าเท่ากับ 0.5816 ซึ่งแสดงถึงประสิทธิภาพการจัดกลุ่มของแบบจำลองอยู่ในระดับปานกลางถึงดี มีความแตกต่างกันระหว่างกลุ่มในระดับหนึ่ง แต่อาจยังมีบางจุดที่อยู่ใกล้ขอบเขตของคลัสเตอร์อื่น



ภาพประกอบ 22 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม K-Means Clustering โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

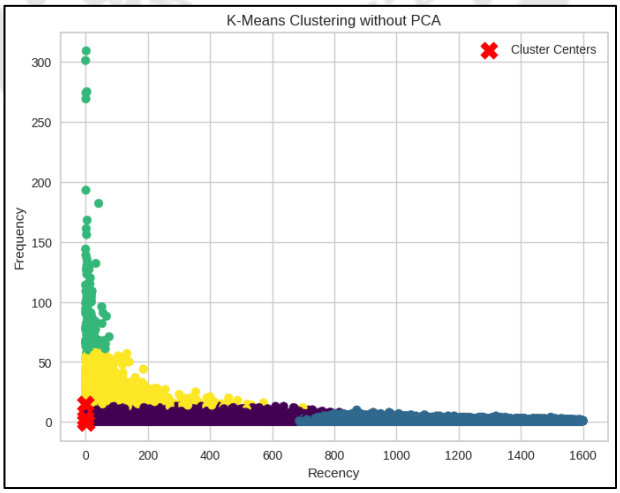
ส่วนผลลัพธ์ที่ได้จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม K-Means Clustering ด้วยชุดข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า คลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด

เป็นคลัสเตอร์ที่ 0 ซึ่งมีจำนวนข้อมูล 38,240 และคลัสเตอร์ที่ 2 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 119 ดังแสดงข้อมูลในตารางที่ 8

ตาราง 8 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม K-Means Clustering โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	38,240
คลัสเตอร์ที่ 1	25,602
คลัสเตอร์ที่ 2	119
คลัสเตอร์ที่ 3	1,986

และจากภาพประกอบที่ 23 แสดงให้เห็นว่า การแบ่งขอบเขตกลุ่มมีความชัดเจน แต่เนื่องจากการแสดงภาพของคลัสเตอร์อยู่ในรูปแบบ 2 มิติ โดยใช้ Frequency ในแกน Y และ Recency ในแกน X จึงทำให้จุดศูนย์กลางของคลัสเตอร์ที่แสดงในภาพดูมีระยะหรือตำแหน่งใกล้เคียงกัน โดยเมื่อพิจารณาค่าเมทริกซ์ Silhouette Score ของแบบจำลองนี้ ซึ่งมีค่าเท่ากับ 0.5806 ก็แสดงถึง ประสิทธิภาพในการจัดกลุ่มระดับปานกลางถึงดีของแบบจำลอง



ภาพประกอบ 23 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม K-Means Clustering โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม K-Means Clustering ด้วยชุดข้อมูล 2 รูปแบบ คือ ชุดข้อมูลที่มีการลดจำนวนมิติ และชุดข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า แบบจำลองมีการแบ่งขอบเขตของแต่ละคลัสเตอร์ได้อย่างชัดเจน และเมื่อพิจารณาค่า Silhouette Score ที่วัดได้ แสดงให้เห็นว่า ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลจากทั้ง 2 รูปแบบ มีประสิทธิภาพที่ใกล้เคียงกัน โดยแบบจำลองที่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ มีค่า Silhouette Score เท่ากับ 0.5816 ส่วนชุดข้อมูลที่มีการลดจำนวนมิติ มีค่า Silhouette Score เท่ากับ 0.5806 ดังแสดงในตารางที่ 9 ซึ่งแสดงถึง ความสามารถในการจัดกลุ่มในระดับปานกลางถึงดี แต่อาจยังมีข้อมูลบางจุดที่ยังทับซ้อนกันอยู่

ตาราง 9 แสดงข้อมูล Silhouette Score ของอัลกอริทึม K-Means Clustering

อัลกอริทึม	รูปแบบของชุดข้อมูล	จำนวนคลัสเตอร์	Silhouette Score
K-Means Clustering	ชุดข้อมูลที่มีการลดจำนวนมิติ	4	0.5816
	ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ	4	0.5806

4.1.2 อัลกอริทึม DBSCAN

การจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม DBSCAN มีการกำหนดค่า min_samples เท่ากับ 10 และทดลองปรับค่า Epsilon ในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 0.2, 0.5 และ 1 โดยทำการทดลองปรับค่าพารามิเตอร์ในระดับเดียวกันกับชุดข้อมูลทั้ง 2 รูปแบบ คือ ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ และชุดข้อมูลที่มีการลดจำนวนมิติ โดยได้ผลลัพธ์จากการทดลอง ดังนี้

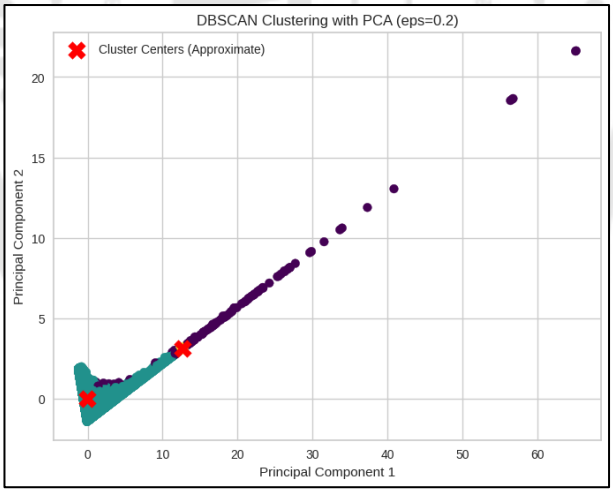
ผลลัพธ์ที่ได้จากการปรับค่า Epsilon = 0.2

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม DBSCAN ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ โดยเริ่มทดลองจากการกำหนดค่า Epsilon เท่ากับ 0.2 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 3 คลัสเตอร์ ซึ่งสามารถแบ่งออกได้เป็น 2 ส่วน คือ คลัสเตอร์ที่ 0 ถึง 1 คือ คลัสเตอร์ที่สามารถจัดกลุ่มได้ ส่วนคลัสเตอร์ที่ -1 เป็นข้อมูลที่ไม่สามารถรวมกลุ่มเป็นคลัสเตอร์ได้ ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูล จากตารางที่ 10 แสดงให้เห็นว่า คลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 65,790 และค่อนข้างกระจุกตัวอยู่ในกลุ่มเป็นส่วนใหญ่ และคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 17 ในส่วนของคลัสเตอร์ที่ -1 ซึ่งเป็นคลัสเตอร์ที่ไม่สามารถจัดกลุ่มได้นั้น มีจำนวนข้อมูลเท่ากับ 140

ตาราง 10 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.2 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	140
คลัสเตอร์ที่ 0	65,790
คลัสเตอร์ที่ 1	17

จากภาพประกอบที่ 24 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ มีรูปแบบการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.8011 ซึ่งแสดงให้เห็นว่า คลัสเตอร์ถูกแบ่งได้อย่างมีประสิทธิภาพสูง สามารถรวมกลุ่มข้อมูลที่มีความหนาแน่นสูงอยู่ภายในคลัสเตอร์เดียวกัน และแบ่งแยกระยะห่างระหว่างคลัสเตอร์อื่นที่ไม่สามารถจัดกลุ่มออกจากกันได้อย่างดี



ภาพประกอบ 24 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.2 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

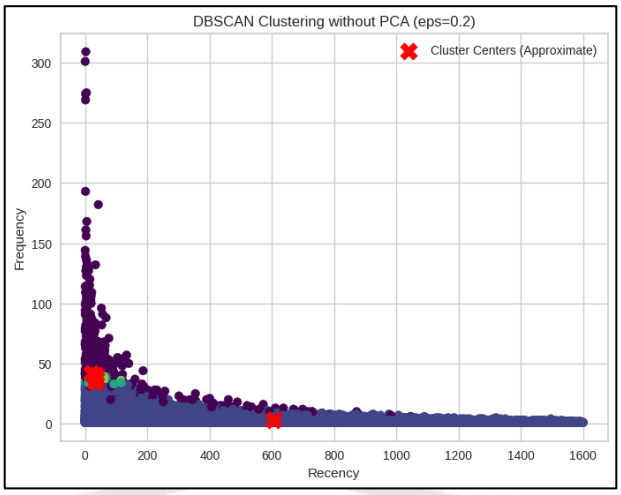
ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 6 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 65,419 ซึ่งยังคงมีการกระจุกของข้อมูลค่อนข้างสูงอยู่ ส่วนคลัส

เตอร์ที่ 4 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 9 ในส่วนของคลัสเตอร์ที่ -1 ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูลนั้น มีจำนวนข้อมูล เท่ากับ 417 ดังแสดงข้อมูลในตารางที่ 11

ตาราง 11 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.2 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	417
คลัสเตอร์ที่ 0	65,419
คลัสเตอร์ที่ 1	13
คลัสเตอร์ที่ 2	33
คลัสเตอร์ที่ 3	56
คลัสเตอร์ที่ 4	9

จากภาพประกอบที่ 25 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์ และตำแหน่งของจุดศูนย์กลางค่อนข้างดี แต่อาจมีบางจุดข้อมูลที่มีระยะใกล้หรือทับซ้อนกันระหว่างคลัสเตอร์ ประกอบกับค่า Silhouette Score เท่ากับ 0.6567 แสดงถึง ประสิทธิภาพในการจัดกลุ่มปานกลางถึงดี หากเปรียบเทียบกับการทดลองก่อนหน้า ที่มีการกำหนดค่า Epsilon เท่ากับ 0.2 เท่ากัน แต่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ พบว่า ประสิทธิภาพของอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิตินั้น มีประสิทธิภาพในการจัดกลุ่มที่สูงกว่าอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เท่ากับ 0.1444



ภาพประกอบ 25 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.2 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

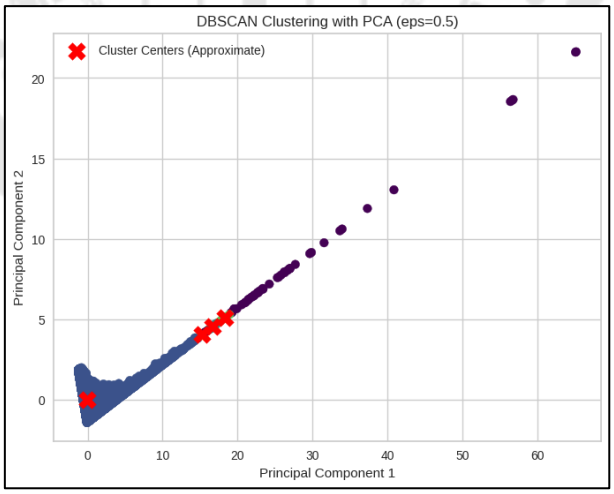
ผลลัพธ์ที่ได้จากการปรับค่า Epsilon = 0.5

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม DBSCAN ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า Epsilon เท่ากับ 0.5 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 5 คลัสเตอร์ ซึ่งสามารถแบ่งออกได้เป็น 2 ส่วน คือ คลัสเตอร์ที่ 0 ถึง 3 คือ คลัสเตอร์ที่สามารถจัดกลุ่มได้ ส่วนคลัสเตอร์ที่ -1 เป็นข้อมูลที่ไม่สามารถรวมกลุ่มเป็นคลัสเตอร์ได้ ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูล จากตารางที่ 12 แสดงให้เห็นว่าคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 65,865 และค่อนข้างกระจุกตัวอยู่ในกลุ่มเป็นส่วนใหญ่ และคลัสเตอร์ที่ 3 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 7 ในส่วนของคลัสเตอร์ที่ -1 ซึ่งเป็นคลัสเตอร์ที่ไม่สามารถจัดกลุ่มได้นั้น มีจำนวนข้อมูลเท่ากับ 48

ตาราง 12 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.5 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	48
คลัสเตอร์ที่ 0	65,865
คลัสเตอร์ที่ 1	15
คลัสเตอร์ที่ 2	12
คลัสเตอร์ที่ 3	7

จากภาพประกอบที่ 26 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ที่ 0 และคลัสเตอร์อื่น มีการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.8349 ซึ่งแสดงให้เห็นว่าอัลกอริทึมมีประสิทธิภาพในการจัดกลุ่มข้อมูลได้อย่างดีมาก มีความชัดเจนระหว่างคลัสเตอร์ และแทบจะไม่มีการทับซ้อนหรือระยะที่ใกล้ของจุดข้อมูลระหว่างคลัสเตอร์



ภาพประกอบ 26 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.5 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 3 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัส

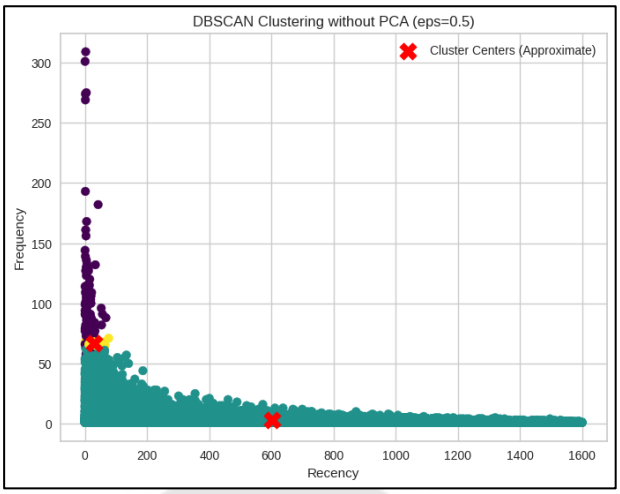
เตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 65,838 ซึ่งยังคงมีการกระจุกของข้อมูลค่อนข้างสูงอยู่ ส่วนคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 15 ในส่วนของคลัสเตอร์ที่ -1 ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูลนั้น มีจำนวนข้อมูล เท่ากับ 94 ดังแสดงข้อมูลในตารางที่ 13

ตาราง 13 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.5 โดยใช้ชุดข้อมูลที่มีการไม่มีลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	94
คลัสเตอร์ที่ 0	65,838
คลัสเตอร์ที่ 1	15

จากภาพประกอบที่ 27 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์มีการแบ่งกลุ่มระหว่างคลัสเตอร์ 0, 1, และ -1 รวมถึงตำแหน่งของจุดศูนย์กลางที่ห่างกันอย่างชัดเจน ประกอบกับค่า Silhouette Score เท่ากับ 0.8230 แสดงถึง ประสิทธิภาพที่สูงของอัลกอริทึมในการจัดกลุ่มที่สามารถรวมกลุ่มที่มีความหนาแน่นสูงอยู่ภายในคลัสเตอร์เดียวกัน และมีขอบเขตที่แบ่งแยกชัดเจนในแต่ละคลัสเตอร์

หากเมื่อเปรียบเทียบกันระหว่างผลลัพธ์ที่ได้จากการทดลองของข้อมูลทั้ง 2 รูปแบบ ที่มีการกำหนดค่า Epsilon เท่ากับ 0.5 ทั้งจากภาพประกอบและค่า Silhouette Score พบว่า ผลลัพธ์ของอัลกอริทึมที่ได้จากทั้ง 2 การทดลอง มีประสิทธิภาพในการจัดกลุ่มที่ดีมากข้อมูลภายในคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลาง มีการแบ่งขอบเขตกันได้อย่างชัดเจน แทบจะไม่มีจุดข้อมูลที่ทับซ้อนกัน โดยอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ จะมีค่า Silhouette Score ที่มากกว่าอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เท่ากับ 0.0119



ภาพประกอบ 27 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 0.5 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

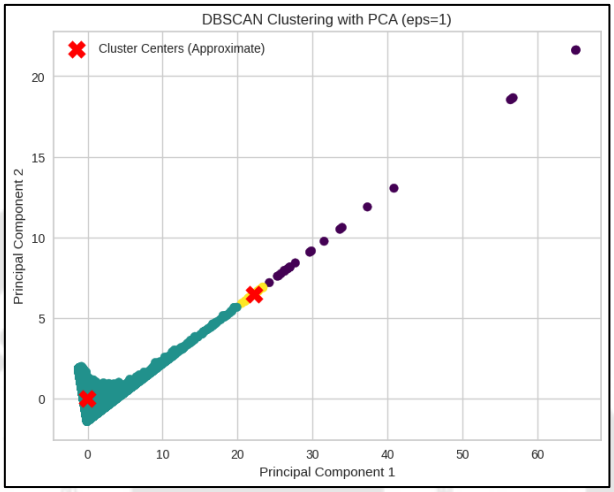
ผลลัพธ์ที่ได้จากการปรับค่า Epsilon = 1

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม DBSCAN ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า Epsilon เท่ากับ 1 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 3 คลัสเตอร์ ซึ่งสามารถแบ่งออกได้เป็น 2 ส่วน คือ คลัสเตอร์ที่ 0 ถึง 1 คือ คลัสเตอร์ที่สามารถจัดกลุ่มได้ ส่วนคลัสเตอร์ที่ -1 เป็นข้อมูลที่ไม่สามารถรวมกลุ่มเป็นคลัสเตอร์ได้ ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูล จากตารางที่ 14 แสดงให้เห็นว่า คลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 65,907 และค่อนข้างกระจุกตัวอยู่ในกลุ่มเป็นส่วนใหญ่ และคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 17 ในส่วนของคลัสเตอร์ที่ -1 ซึ่งเป็นคลัสเตอร์ที่ไม่สามารถจัดกลุ่มได้นั้น มีจำนวนข้อมูลเท่ากับ 23

ตาราง 14 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 1 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	23
คลัสเตอร์ที่ 0	65,907
คลัสเตอร์ที่ 1	17

จากภาพประกอบที่ 28 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ มีการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.8871 ซึ่งแสดงให้เห็นว่า อัลกอริทึมมีประสิทธิภาพในการจัดกลุ่มข้อมูลได้อย่างดีมาก มีความชัดเจนระหว่างคลัสเตอร์ และแทบจะไม่มีการทับซ้อนหรือระยะที่ใกล้ของจุดข้อมูลระหว่างคลัสเตอร์



ภาพประกอบ 28 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 1 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

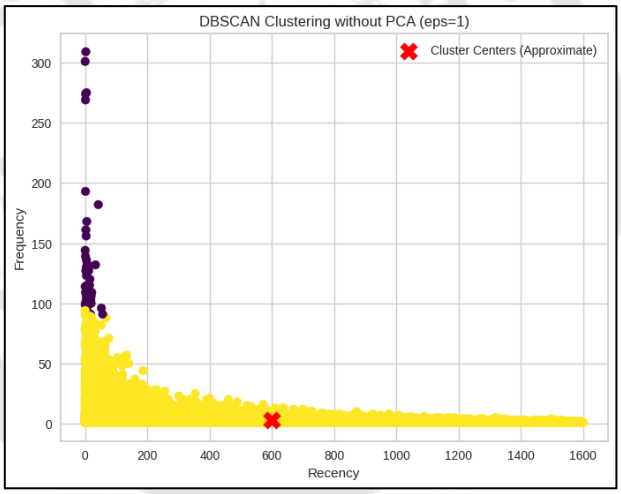
ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนของคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 2 คลัสเตอร์ ซึ่งแสดงว่า อัลกอริทึมสามารถจัดกลุ่มได้เพียงกลุ่มเดียว คือ คลัสเตอร์ที่ 0 ที่มีจำนวนข้อมูลเท่ากับ 65,901 ส่วนคลัสเตอร์ที่ -1 ซึ่งอาจเป็น Outlier/Noise ของชุดข้อมูลนั้น มีจำนวนข้อมูล เท่ากับ 46 ดังแสดงข้อมูลในตารางที่ 15

ตาราง 15 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 1 โดยใช้ชุดข้อมูลที่มีการไม่มีลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ -1 (Outlier)	46
คลัสเตอร์ที่ 0	65,901

จากภาพประกอบที่ 29 แสดงให้เห็นว่า คลัสเตอร์และตำแหน่งของจุดศูนย์กลางที่มีการแบ่งแยกกันชัดเจนมาก เนื่องจากในการทดลองนี้ ผลลัพธ์ที่ได้มีเพียงคลัสเตอร์เดียว สอดคล้องกับค่า Silhouette Score เท่ากับ 0.9133 ที่แสดงถึง ประสิทธิภาพที่สูงมากของอัลกอริทึมในการจัดกลุ่ม ที่สามารถรวมกลุ่มที่มีความหนาแน่นสูงอยู่ภายในคลัสเตอร์เดียวกัน และมีขอบเขตที่แบ่งแยกชัดเจนในแต่ละคลัสเตอร์

หากเมื่อเปรียบเทียบกันระหว่างผลลัพธ์ที่ได้จากการทดลองของข้อมูลทั้ง 2 รูปแบบ ที่มีการกำหนดค่า Epsilon เท่ากับ 1 ทั้งจากภาพประกอบและค่า Silhouette Score พบว่า ผลลัพธ์ของอัลกอริทึมที่ได้จากทั้ง 2 การทดลอง มีประสิทธิภาพในการจัดกลุ่มที่ดีมาก ข้อมูลภายในคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลาง มีการแบ่งขอบเขตกันได้อย่างชัดเจน แทบจะไม่มีจุดข้อมูลที่ทับซ้อนกัน โดยอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ จะมีค่า Silhouette Score ที่มากกว่าอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เท่ากับ 0.0262



ภาพประกอบ 29 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม DBSCAN เมื่อกำหนดค่า Epsilon เท่ากับ 1 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม DBSCAN ด้วยชุดข้อมูล 2 รูปแบบ คือ ชุดข้อมูลที่มีการลดจำนวนมิติ และชุดข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า แบบจำลองมีการแบ่งขอบเขตของแต่ละคลัสเตอร์ได้อย่างชัดเจน และเมื่อพิจารณาค่า Silhouette Score ที่วัดได้ แสดงให้เห็นว่า ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลจากทั้ง 2 รูปแบบ ด้วยอัลกอริทึมที่มีการกำหนดค่า Epsilon ที่แตกต่างกัน ส่วนใหญ่มีประสิทธิภาพในการจัดกลุ่มที่สูงมากและค่าที่ได้ค่อนข้างใกล้เคียงกัน โดยสันนิษฐานว่า ค่า Silhouette Score ที่สูงมากนั้น เป็นผลมาจากการที่

ข้อมูลส่วนใหญ่ถูกจัดกลุ่มเข้าด้วยกัน และข้อมูลส่วนที่เหลือถูกแบ่งออกไปเป็นคลัสเตอร์อื่นแทน ทำให้ข้อมูลมีการกระจุกตัวกันภายในคลัสเตอร์เดียว ส่งผลให้ผลลัพธ์จากการทดลองที่ได้แสดงขอบเขตการจัดกลุ่มที่ชัดเจนผ่านภาพประกอบ รวมถึงค่า Silhouette Score ที่วัดค่าได้

โดยแบบจำลองที่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ ที่มีค่า Silhouette Score สูงสุด เป็นแบบจำลองที่มีกำหนดค่า Epsilon เท่ากับ 1 โดยมีค่า Silhouette Score เท่ากับ 0.8871 ส่วนชุดข้อมูลที่มีการลดจำนวนมิติ เป็นแบบจำลองที่มีกำหนดค่า Epsilon เท่ากับ 1 เช่นเดียวกัน โดยมีค่า Silhouette Score เท่ากับ 0.9133 ดังแสดงในตารางที่ 16

ตาราง 16 แสดงข้อมูล Silhouette Score ของอัลกอริทึม DBSCAN

อัลกอริทึม	**รูปแบบของชุดข้อมูล**	Epsilon	**จำนวนคลัสเตอร์**	Silhouette Score
DBSCAN	ชุดข้อมูลที่มีการลดจำนวนมิติ	0.2	3	0.8011
		0.5	5	0.8349
		1	3	0.8871
	ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ	0.2	6	0.6567
		0.5	3	0.8230
		1	2	0.9133

4.1.3 **อัลกอริทึม** GMM

การจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม GMM ทำการทดลองปรับค่า n_component ซึ่งเป็นค่าพารามิเตอร์ที่กำหนดจำนวนของ Gaussian Distribution หรือจำนวนคลัสเตอร์ที่ใช้ในการสร้างแบบจำลอง โดยทำการทดลองปรับในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 3, 4, และ 5 ทั้งในแบบจำลองที่ใช้ข้อมูลที่ผ่านการลดจำนวนมิติ และแบบจำลองที่ใช้ข้อมูลที่ไม่ได้ลดจำนวนมิติ โดยได้ผลลัพธ์จากการทดลอง ดังนี้

ผลลัพธ์ที่ได้จากการปรับค่า n_component = 3

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม GMM ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ โดยเริ่มทดลองจากการกำหนดค่า n_component เท่ากับ 3 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 3 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่

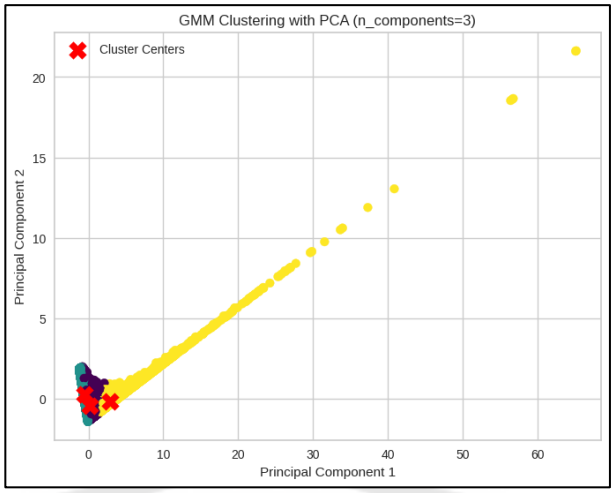
มีจำนวนข้อมูลมากที่สุด คือ 39,276 และคลัสเตอร์ที่ 2 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 4,538 ดังแสดงข้อมูลในตารางที่ 17

ตาราง 17 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 3 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	39,276
คลัสเตอร์ที่ 1	22,133
คลัสเตอร์ที่ 2	4,538

จากภาพประกอบที่ 30 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ไม่ค่อยชัดเจน มีการทับซ้อนกันของจุดข้อมูล และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.1479 แสดงให้เห็นว่า การจัดกลุ่มข้อมูลของอัลกอริทึมในการทดลองนี้มีประสิทธิภาพที่ต่ำมาก ซึ่งแสดงว่า ข้อมูลอาจไม่ได้ถูกจัดกลุ่มอย่างชัดเจน หรืออาจมีการซ้อนทับกันของคลัสเตอร์

ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 3 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 35,299 ส่วนคลัสเตอร์ที่ 2 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 5,908 ดังแสดงข้อมูลในตารางที่ 18

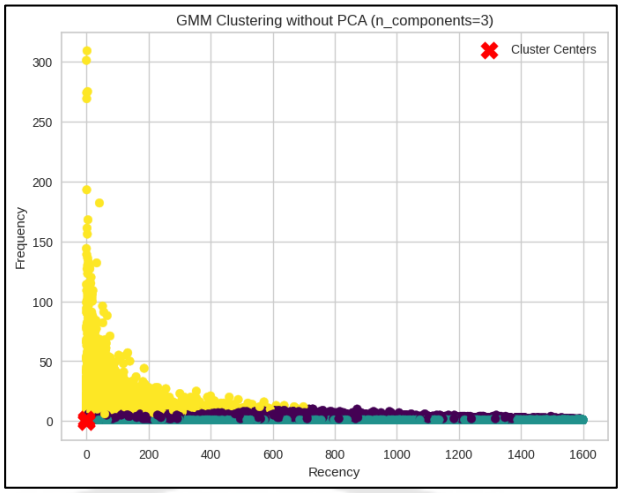


ภาพประกอบ 30 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 3 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

ตาราง 18 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 3 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	35,299
คลัสเตอร์ที่ 1	24,740
คลัสเตอร์ที่ 2	5,908

จากภาพประกอบที่ 31 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์ และตำแหน่งของจุดศูนย์กลางมีการทับซ้อนกันระหว่างคลัสเตอร์ และเมื่อพิจารณาประกอบกับค่า Silhouette Score เท่ากับ 0.1139 แสดงถึง ประสิทธิภาพในการจัดกลุ่มที่ต่ำมาก หากเปรียบเทียบกับการทดลองก่อนหน้า ที่มีการกำหนดค่า n_component เท่ากับ 3 เท่ากัน แต่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ พบว่า ประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มของทั้ง 2 การทดลอง มีประสิทธิภาพที่ต่ำมากทั้งคู่ แต่หากจะเปรียบเทียบกันนั้น ประสิทธิภาพของอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ จะมีประสิทธิภาพในการจัดกลุ่มที่ดีกว่าอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ ซึ่งมีค่าต่างกันเพียงเล็กน้อย คือ 0.034



ภาพประกอบ 31 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 3 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

ผลลัพธ์ที่ได้จากการปรับค่า n_component = 4

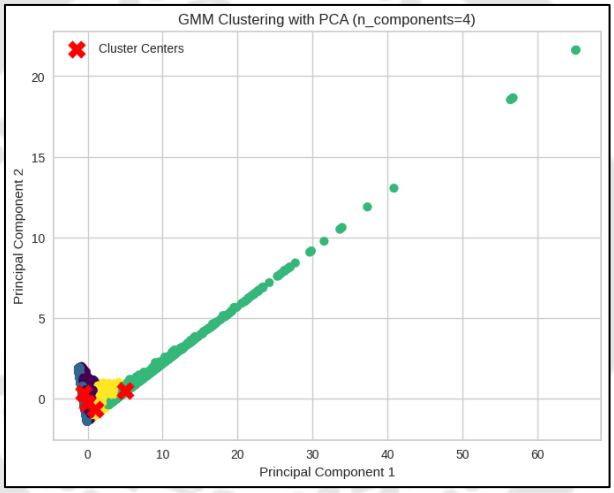
จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม GMM ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า n_component เท่ากับ 4 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 4 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 37,166 และคลัสเตอร์ที่ 3 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 1,636 ดังแสดงข้อมูลในตารางที่ 19

ตาราง 19 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 4 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	37,166
คลัสเตอร์ที่ 1	18,678
คลัสเตอร์ที่ 2	8,467
คลัสเตอร์ที่ 3	1,636

จากภาพประกอบที่ 32 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ไม่ค่อยชัดเจน มีการทับซ้อนกันของจุดข้อมูล และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.0457 แสดงให้เห็นว่า การจัดกลุ่มข้อมูลของอัลกอริทึมในการทดลองนี้มีประสิทธิภาพที่ต่ำมาก ซึ่งแสดงว่า ข้อมูลอาจไม่ได้ถูกจัดกลุ่มอย่างชัดเจน หรืออาจมีการซ้อนทับกันของคลัสเตอร์

ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 4 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 35,299 ส่วนคลัสเตอร์ที่ 2 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 2,086 ดังแสดงข้อมูลในตารางที่ 20

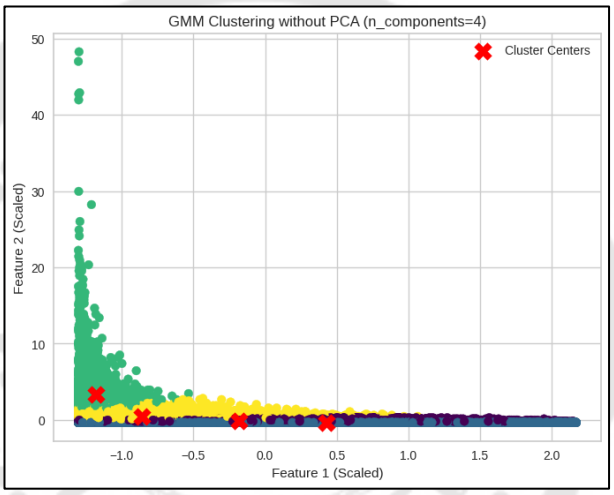


ภาพประกอบ 32 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 4 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

ตาราง 20 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 4 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	35,299
คลัสเตอร์ที่ 1	18,486
คลัสเตอร์ที่ 2	10,076
คลัสเตอร์ที่ 3	2,086

จากภาพประกอบที่ 33 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์ และตำแหน่งของจุดศูนย์กลางมีการทับซ้อนกันระหว่างคลัสเตอร์ และเมื่อพิจารณาประกอบกับค่า Silhouette Score เท่ากับ 0.0186 แสดงถึง ประสิทธิภาพในการจัดกลุ่มที่ต่ำมาก หากเปรียบเทียบกับการทดลองก่อนหน้า ที่มีการกำหนดค่า n_component เท่ากับ 4 เท่ากัน แต่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ พบว่า ประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มของทั้ง 2 การทดลอง มีประสิทธิภาพที่ต่ำมากทั้งคู่ แต่หากเปรียบเทียบกันนั้น ประสิทธิภาพของอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ จะมีประสิทธิภาพในการจัดกลุ่มที่ดีกว่าอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ ซึ่งมีค่าต่างกันเพียงเล็กน้อย คือ 0.0271



ภาพประกอบ 33 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM
เมื่อกำหนดค่า n_component เท่ากับ 4 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

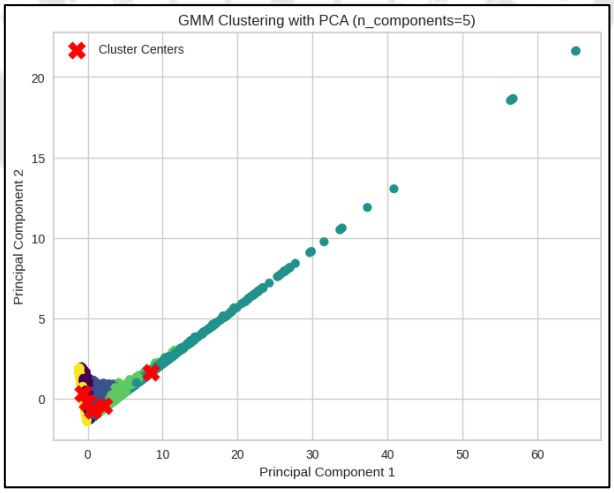
ผลลัพธ์ที่ได้จากการปรับค่า n_component = 5

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม GMM ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า n_component เท่ากับ 5 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 4 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 36,884 และคลัสเตอร์ที่ 4 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 529 ดังแสดงข้อมูลในตารางที่ 21

ตาราง 21 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 5 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	36,884
คลัสเตอร์ที่ 1	16,505
คลัสเตอร์ที่ 2	8,959
คลัสเตอร์ที่ 3	3,070
คลัสเตอร์ที่ 4	529

จากภาพประกอบที่ 34 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ไม่ค่อยชัดเจน มีการทับซ้อนกันของจุดข้อมูล และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.0118 แสดงให้เห็นว่า การจัดกลุ่มข้อมูลของอัลกอริทึมในการทดลองนี้มีประสิทธิภาพที่ต่ำมาก ซึ่งแสดงว่า ข้อมูลอาจไม่ได้ถูกจัดกลุ่มอย่างชัดเจน หรืออาจมีการซ้อนทับกันของคลัสเตอร์



ภาพประกอบ 34 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 5 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

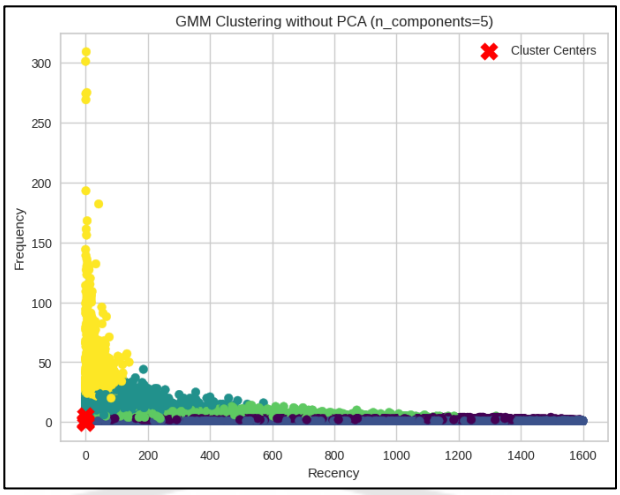
ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 5 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัส

เตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 35,299 ส่วนคลัสเตอร์ที่ 4 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 741 ดังแสดงข้อมูลในตารางที่ 22

ตาราง 22 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม GMM เมื่อกำหนดค่า n_component เท่ากับ 5 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	35,299
คลัสเตอร์ที่ 1	16,819
คลัสเตอร์ที่ 2	9,259
คลัสเตอร์ที่ 3	3,829
คลัสเตอร์ที่ 4	741

จากภาพประกอบที่ 35 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์ และตำแหน่งของจุดศูนย์กลางมีการทับซ้อนกันระหว่างคลัสเตอร์ จนไม่สามารถแยกคลัสเตอร์ออกจากกันได้ และเมื่อพิจารณาประกอบกับค่า Silhouette Score เท่ากับ -0.0095 แสดงถึง การจัดกลุ่มที่ไม่มีประสิทธิภาพ ซึ่งอาจมีสาเหตุมาจากการทับซ้อนกันของคลัสเตอร์จำนวนมาก หรือไม่มีโครงสร้างกลุ่มที่ชัดเจน หากเปรียบเทียบกับการทดลองก่อนหน้า ที่มีการกำหนดค่า n_component เท่ากับ 5 เท่ากัน แต่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ พบว่า ประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มของทั้ง 2 การทดลอง มีประสิทธิภาพที่ต่ำมากทั้งคู่ แต่หากเปรียบเทียบกันนั้น ประสิทธิภาพของอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ จะมีประสิทธิภาพในการจัดกลุ่มที่ดีกว่าอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ ซึ่งมีค่าต่างกันเพียงเล็กน้อย คือ 0.0213



ภาพประกอบ 35 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม GMM
เมื่อกำหนดค่า n_component เท่ากับ 5 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม GMM ด้วยชุดข้อมูล 2 รูปแบบ คือ ชุดข้อมูลที่มีการลดจำนวนมิติ และชุดข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า แบบจำลองมีการแบ่งขอบเขตของแต่ละคลัสเตอร์ได้อย่างชัดเจน และเมื่อพิจารณาค่า Silhouette Score ที่วัดได้ แสดงให้เห็นว่า ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลจากทั้ง 2 รูปแบบ ด้วยอัลกอริทึมที่มีการกำหนดค่า n_component ที่แตกต่างกัน ส่วนใหญ่มีประสิทธิภาพในการจัดกลุ่มที่ต่ำมากและค่าที่ได้ค่อนข้างใกล้เคียงกัน โดยสันนิษฐานว่า ค่า Silhouette Score ที่ต่ำมากนั้น เป็นผลมาจากการที่ข้อมูลที่นำมาใช้ทั้ง 3 คุณลักษณะ มีการแจกแจงแบบไม่ปกติ ซึ่งมีลักษณะเบ้ไปทางขวา โดยเฉพาะคุณลักษณะ Frequency และ Monetary ที่มีค่าความโด่งสูงมาก จึงอาจส่งผลให้ผลลัพธ์จากการจัดกลุ่มที่ได้นั้น ไม่มีประสิทธิภาพ

โดยแบบจำลองที่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ ที่มีค่า Silhouette Score สูงสุด จะเป็นแบบจำลองที่มีกำหนดค่า n_component เท่ากับ 3 โดยมีค่า Silhouette Score เท่ากับ 0.1479 ส่วนชุดข้อมูลที่มีการลดจำนวนมิติ จะเป็นแบบจำลองที่มีกำหนดค่า n_component เท่ากับ 3 เช่นเดียวกัน โดยมีค่า Silhouette Score เท่ากับ 0.1139 ดังแสดงในตารางที่ 23

ตาราง 23 แสดงข้อมูล Silhouette Score ของอัลกอริทึม GMM

อัลกอริทึม	**รูปแบบของชุดข้อมูล**	n_component	**จำนวนคลัสเตอร์**	Silhouette Score
GMM	ชุดข้อมูลที่มีการลดจำนวนมิติ	3	3	0.1479
		4	4	0.0457
		5	5	0.01183
	ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ	3	3	0.1139
		4	4	0.0186
		5	5	-0.0095

4.1.4 **อัลกอริทึม** Agglomerative Clustering

การจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม Agglomerative Clustering ทำการทดลองปรับค่า n_clusters ซึ่งเป็นค่าพารามิเตอร์ที่กำหนดจำนวนของคลัสเตอร์ที่จะใช้ในการสร้างแบบจำลอง โดยจะทำการทดลองปรับในระดับที่แตกต่างกันแบบสุ่ม ได้แก่ 2, 3, และ 4 ทั้งในแบบจำลองที่ใช้ข้อมูลที่ผ่านการลดจำนวนมิติ และแบบจำลองที่ใช้ข้อมูลที่ไม่ได้ลดจำนวนมิติ โดยได้ผลลัพธ์จากการทดลอง ดังนี้

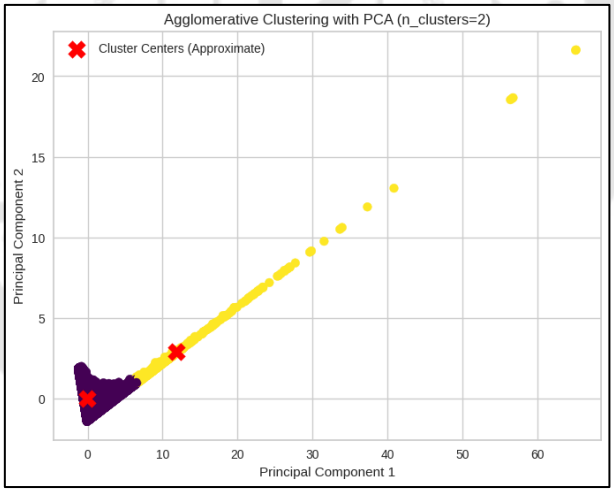
ผลลัพธ์ที่ได้จากการปรับค่า n_clusters = 2

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม Agglomerative Clustering ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า n_clusters เท่ากับ 2 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 2 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 65,516 และคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 431 ดังแสดงข้อมูลในตารางที่ 24

ตาราง 24 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 2 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	65,516
คลัสเตอร์ที่ 1	431

จากภาพประกอบที่ 36 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ มีรูปแบบการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.8001 ซึ่งแสดงให้เห็นว่า คลัสเตอร์ถูกแบ่งได้อย่างมีประสิทธิภาพสูง มีการแบ่งกลุ่มข้อมูลออกจากกันได้อย่างดี แทบจะไม่มีการทับซ้อนกันของข้อมูลระหว่างคลัสเตอร์



ภาพประกอบ 36 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 2 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

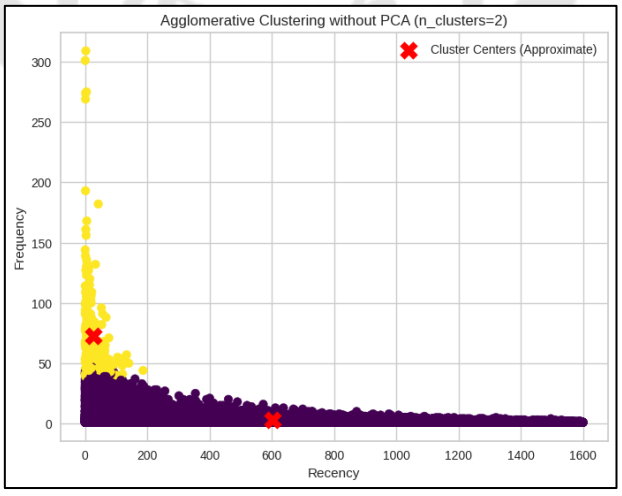
ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 2 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 65,699 ส่วนคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 248 ดังแสดงข้อมูลในตารางที่ 25

ตาราง 25 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 2 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	65,699
คลัสเตอร์ที่ 1	248

จากภาพประกอบที่ 37 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์มีการแบ่งกลุ่มระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ห่างกันอย่างชัดเจน ประกอบกับค่า Silhouette Score เท่ากับ 0.8404 แสดงถึง ประสิทธิภาพที่สูงของอัลกอริทึมในการจัดกลุ่ม ที่สามารถแบ่งแยกขอบเขตชัดเจนในแต่ละคลัสเตอร์

หากเมื่อเปรียบเทียบกันระหว่างผลลัพธ์ที่ได้จากการทดลองของข้อมูลทั้ง 2 รูปแบบ ที่มีการกำหนดค่า n_clusters เท่ากับ 2 ทั้งจากภาพประกอบและค่า Silhouette Score พบว่า ผลลัพธ์ของอัลกอริทึมที่ได้จากทั้ง 2 การทดลอง มีประสิทธิภาพในการจัดกลุ่มที่ดีมาก ข้อมูลภายในคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลาง มีการแบ่งขอบเขตกันได้อย่างชัดเจน แทบจะไม่มีจุดข้อมูลที่ทับซ้อนกัน โดยอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ มีค่า Silhouette Score ที่มากกว่าอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เท่ากับ 0.0403



ภาพประกอบ 37 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 2 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

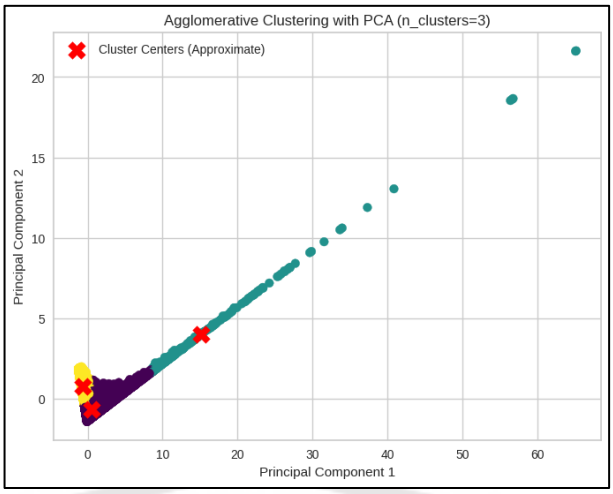
ผลลัพธ์ที่ได้จากการปรับค่า n_clusters = 3

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม Agglomerative Clustering ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า n_clusters เท่ากับ 3 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 3 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 35,754 และคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 248 ดังแสดงข้อมูลในตารางที่ 26

ตาราง 26 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 3 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	35,754
คลัสเตอร์ที่ 1	248
คลัสเตอร์ที่ 2	29,945

จากภาพประกอบที่ 38 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ มีรูปแบบการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.5457 ซึ่งแสดงให้เห็นว่า คลัสเตอร์ถูกแบ่งได้อย่างมีประสิทธิภาพในระดับปานกลาง มีการแบ่งกลุ่มข้อมูลออกจากกันได้อย่างดี แต่อาจมีการซ้อนกันของข้อมูลระหว่างคลัสเตอร์อยู่บ้าง



ภาพประกอบ 38 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 3 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

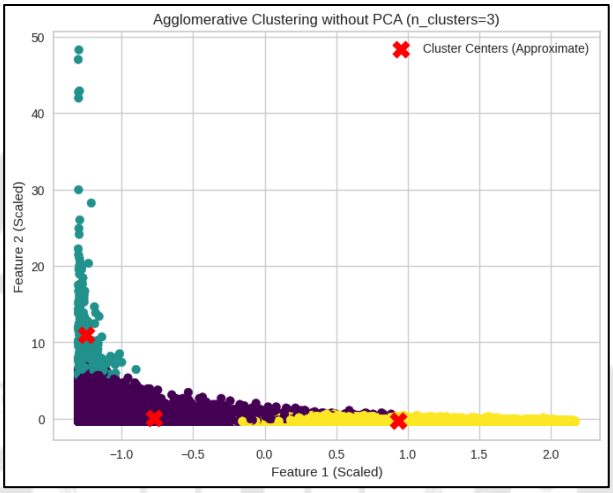
ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 3 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 0 มีจำนวนข้อมูลเท่ากับ 35,754 ส่วนคลัสเตอร์ที่ 1 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 248 ดังแสดงข้อมูลในตารางที่ 27

ตาราง 27 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 3 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	35,754
คลัสเตอร์ที่ 1	248
คลัสเตอร์ที่ 2	29,945

จากภาพประกอบที่ 39 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์มีการแบ่งกลุ่มระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ห่างกันอย่างชัดเจน ประกอบกับค่า Silhouette Score เท่ากับ 0.5457 แสดงถึง ประสิทธิภาพระดับปานกลางของอัลกอริทึมในการจัดกลุ่ม ที่สามารถแบ่งแยกขอบเขตแต่ละคลัสเตอร์ค่อนข้างดี แต่อาจยังมีการทับซ้อนกันของข้อมูล

หากเมื่อเปรียบเทียบกันระหว่างผลลัพธ์ที่ได้จากการทดลองของข้อมูลทั้ง 2 รูปแบบ ที่มีการกำหนดค่า n_clusters เท่ากับ 3 ทั้งจากภาพประกอบและค่า Silhouette Score พบว่า ผลลัพธ์ของอัลกอริทึมที่ได้จากทั้ง 2 การทดลอง มีประสิทธิภาพในการจัดกลุ่มในระดับปานกลาง ซึ่งข้อมูลภายในคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลาง มีการแบ่งขอบเขตกันได้ค่อนข้าง โดยทั้งอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ และที่ใช้ข้อมูลที่มีการลดจำนวนมิติ มีค่า Silhouette Score เท่ากัน คือ 0.5457



ภาพประกอบ 39 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 3 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

ผลลัพธ์ที่ได้จากการปรับค่า n_clusters = 4

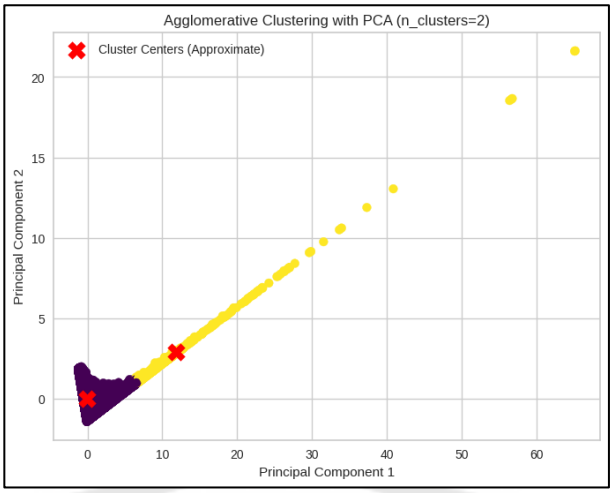
จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม Agglomerative Clustering ด้วยชุดข้อมูลที่มีการลดจำนวนมิติ เมื่อทดลองกำหนดค่า n_clusters เท่ากับ 4 พบว่า จำนวนคลัสเตอร์ที่สามารถจัดกลุ่มได้จากการใช้แบบจำลองนี้ มีจำนวน 4 คลัสเตอร์ ซึ่งคลัสเตอร์ที่ 3 เป็นคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ 34,103 และคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 431 ดังแสดงข้อมูลในตารางที่ 28

ตาราง 28 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 4 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	431
คลัสเตอร์ที่ 1	4,131
คลัสเตอร์ที่ 2	27,282
คลัสเตอร์ที่ 3	34,103

จากภาพประกอบที่ 40 แสดงถึง การแบ่งกลุ่มข้อมูลระหว่างคลัสเตอร์ มีรูปแบบการแบ่งขอบเขตของคลัสเตอร์ที่ชัดเจน รวมถึงตำแหน่งของจุดศูนย์กลางที่แตกต่างกัน และเมื่อพิจารณาถึงค่า Silhouette Score เท่ากับ 0.5257 ซึ่งแสดงให้เห็นว่า คลัสเตอร์ถูกแบ่งได้อย่างมีประสิทธิภาพในระดับปานกลาง มีการแบ่งกลุ่มข้อมูลออกจากกันได้อย่างดี แต่อาจมีบางจุดที่ข้อมูลทับซ้อนกันระหว่างคลัสเตอร์

ผลลัพธ์ที่ได้จากการทดลองกับข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่า จำนวนคลัสเตอร์ที่ได้จากการจัดกลุ่ม มีจำนวน 4 คลัสเตอร์ ซึ่งคลัสเตอร์ที่มีจำนวนข้อมูลมากที่สุด คือ คลัสเตอร์ที่ 1 มีจำนวนข้อมูลเท่ากับ 33,744 ส่วนคลัสเตอร์ที่ 0 เป็นคลัสเตอร์ที่มีจำนวนน้อยที่สุด คือ 248 ดังแสดงข้อมูลในตารางที่ 29



ภาพประกอบ 40 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 4 โดยใช้ชุดข้อมูลที่มีการลดจำนวนมิติ

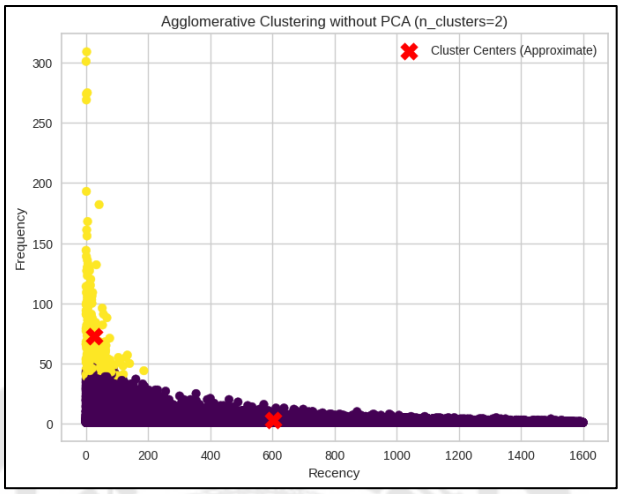
ตาราง 29 แสดงจำนวนข้อมูลในแต่ละคลัสเตอร์ที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 4 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

คลัสเตอร์	**จำนวนข้อมูล**
คลัสเตอร์ที่ 0	248
คลัสเตอร์ที่ 1	33,744
คลัสเตอร์ที่ 2	29,945
คลัสเตอร์ที่ 3	2,010

จากภาพประกอบที่ 41 แสดงให้เห็นว่า การจัดกลุ่มของคลัสเตอร์มีการแบ่งกลุ่มระหว่างคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลางที่ห่างกันอย่างชัดเจน ประกอบกับค่า Silhouette Score เท่ากับ 0.5457 แสดงถึง ประสิทธิภาพที่สูงของอัลกอริทึมในการจัดกลุ่ม ที่สามารถแบ่งแยกขอบเขตชัดเจนในแต่ละคลัสเตอร์

หากเมื่อเปรียบเทียบกันระหว่างผลลัพธ์ที่ได้จากการทดลองของข้อมูลทั้ง 2 รูปแบบ ที่มีการกำหนดค่า n_clusters เท่ากับ 4 ทั้งจากภาพประกอบและค่า Silhouette Score พบว่า ผลลัพธ์ของอัลกอริทึมที่ได้จากทั้ง 2 การทดลอง มีประสิทธิภาพในการจัดกลุ่มที่ดีมาก ข้อมูลภายในคลัสเตอร์ รวมถึงตำแหน่งของจุดศูนย์กลาง มีการแบ่งขอบเขตกันได้อย่างชัดเจน

แทบจะไม่มีจุดข้อมูลที่ทับซ้อนกัน โดยอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ มีค่า Silhouette Score ที่มากกว่าอัลกอริทึมที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เท่ากับ 0.02



ภาพประกอบ 41 แสดงคลัสเตอร์และจุดศูนย์กลางที่ได้จากอัลกอริทึม Agglomerative Clustering เมื่อกำหนดค่า n_clusters เท่ากับ 4 โดยใช้ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ

จากการทดลองจัดกลุ่มลูกค้าโดยใช้อัลกอริทึม Agglomerative Clustering ด้วยชุดข้อมูล 2 รูปแบบ คือ ชุดข้อมูลที่มีการลดจำนวนมิติ และชุดข้อมูลที่ไม่มีการลดจำนวนมิติ พบว่าแบบจำลองมีการแบ่งขอบเขตของแต่ละคลัสเตอร์ได้อย่างชัดเจน และเมื่อพิจารณาค่า Silhouette Score ที่วัดได้ แสดงให้เห็นว่า ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลจากทั้ง 2 รูปแบบ ด้วยอัลกอริทึมที่มีการกำหนดค่า n_clusters ที่แตกต่างกัน ส่วนใหญ่มีประสิทธิภาพในการจัดกลุ่มในระดับปานกลางถึงสูง รวมถึงมีการกระจายข้อมูลไปตามจำนวนคลัสเตอร์ที่กำหนด ทำให้จำนวนข้อมูลที่ได้ในแต่ละคลัสเตอร์มีการกระจุกตัวน้อยกว่าอัลกอริทึม DBSCAN

โดยแบบจำลองที่ใช้ชุดข้อมูลที่มีการลดจำนวนมิติ ที่มีค่า Silhouette Score สูงสุด เป็นแบบจำลองที่มีกำหนดค่า n_clusters เท่ากับ 2 โดยมีค่า Silhouette Score เท่ากับ 0.8001 ส่วนชุดข้อมูลที่มีการลดจำนวนมิติ เป็นแบบจำลองที่มีกำหนดค่า n_clusters เท่ากับ 2 เช่นเดียวกัน โดยมีค่า Silhouette Score เท่ากับ 0.8404 ดังแสดงในตารางที่ 30

ตาราง 30 แสดงข้อมูล Silhouette Score ของอัลกอริทึม Agglomerative Clustering

อัลกอริทึม	รูปแบบของชุดข้อมูล	n_clusters	จำนวนคลัสเตอร์	Silhouette Score
Agglomerative Clustering	ชุดข้อมูลที่มีการลดจำนวนมิติ	2	2	0.8001
		3	3	0.5457
		4	4	0.5257
	ชุดข้อมูลที่ไม่มีการลดจำนวนมิติ	2	2	0.8404
		3	3	0.5457
		4	4	0.5457

ผลลัพธ์ที่ได้จากการทดลองจัดกลุ่มข้อมูลลูกค้า โดยใช้อัลกอริทึมทั้ง 4 อัลกอริทึม คือ K-Means Clustering, DBSCAN, GMM, และ Agglomerative Clustering รวมถึงการทดลองปรับค่าพารามิเตอร์ จากภาพประกอบที่ 42 พบว่า อัลกอริทึม DBSCAN เป็นอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลมากที่สุด จากค่า Silhouette Score ที่สูงถึง 0.9133 สำหรับข้อมูลที่ไม่มีการลดจำนวนมิติ และ 0.8871 สำหรับข้อมูลที่มีการลดจำนวนมิติ โดยเมื่อพิจารณาถึงจำนวนคลัสเตอร์ที่อัลกอริทึมแบ่งออกมาได้ พบว่า สาเหตุที่อัลกอริทึมมีค่า Silhouette Score ที่สูงมากนั้น เป็นเพราะข้อมูลที่ถูกจัดกลุ่ม มีการกระจุกตัวรวมอยู่ในคลัสเตอร์ใดคลัสเตอร์หนึ่ง ประกอบกับมีจำนวนของคลัสเตอร์ที่สามารถจัดกลุ่มได้จำนวนไม่มาก ส่งผลให้มีความหนาแน่นภายในคลัสเตอร์สูง และมีระยะห่างระหว่างคลัสเตอร์ที่ชัดเจน โดยคาดว่า เกิดจากการที่จุดข้อมูลมีระยะห่างกันระหว่างกันไม่มาก และเมื่อปรับค่าพารามิเตอร์ Epsilon เท่ากับ 1 จึงทำให้จุดข้อมูลที่อยู่ใกล้กันเหล่านั้น ถูกจัดมาอยู่รวมในคลัสเตอร์เดียวกัน

ลำดับถัดมา คือ อัลกอริทึม Agglomerative Clustering ซึ่งมีประสิทธิภาพในการจัดกลุ่มข้อมูลที่สูงใกล้เคียงกับอัลกอริทึม DBSCAN เนื่องจากมีการกำหนดจำนวนของคลัสเตอร์ไว้ก่อนดำเนินการทดลอง จึงทำให้อัลกอริทึมจำเป็นต้องจะกลุ่มภายใต้จำนวนคลัสเตอร์ที่กำหนด ซึ่งส่งผลให้จำนวนข้อมูลในแต่ละคลัสเตอร์ มีการกระจายตัวได้ดีกว่าอัลกอริทึม DBSCAN ถึงแม้ว่าข้อมูลยังคงมีการกระจุกตัวอยู่ที่คลัสเตอร์ใดคลัสเตอร์หนึ่งอยู่บ้าง แต่ก็ลดน้อยลงและทำการกระจายข้อมูลไปยังแต่ละคลัสเตอร์มากกว่า

สำหรับอัลกอริทึม K-Means Clustering นั้น ได้นำเทคนิค Elbow Method มาใช้ในการค้นหาจำนวนคลัสเตอร์ที่เหมาะสม ซึ่งจากการทดลอง ก็พบว่า ประสิทธิภาพในการจัดกลุ่มของอัลกอริทึมมีค่าอยู่ในระดับปานกลาง

และสุดท้าย อัลกอริทึม GMM เป็นอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มที่ต่ำที่สุด ถึงแม้ว่าจะมีการทดลองปรับค่าพารามิเตอร์ คือ n_component ที่เป็นตัวกำหนดจำนวนคลัสเตอร์ของอัลกอริทึม ในระดับที่แตกต่างกัน แต่จากการทดลองนั้น แสดงให้เห็นว่า ค่า Silhouette Score ที่ได้รับ ยังคงอยู่ในระดับที่ต่ำจนถึงไม่มีประสิทธิภาพ ซึ่งสันนิษฐานว่า ประสิทธิภาพที่ต่ำของอัลกอริทึม อาจมีผลมาจากรูปแบบการกระจายตัวของข้อมูลที่นำมาใช้ในการจัดกลุ่มของอัลกอริทึม ซึ่งจากการที่ได้สำรวจข้อมูลก่อนนำมาทดลอง พบว่า ข้อมูลทั้ง 3 คุณลักษณะ ได้แก่ Recency, Frequency, และ Monetary นั้น มีลักษณะการแจกแจงแบบไม่ปกติ คือ เบ้ไปทางขวา รวมทั้งคุณลักษณะ Frequency และ Monetary ยังมีค่าความโด่งที่สูงมากเช่นกัน ด้วยเหตุนี้ จึงเป็นสาเหตุให้อัลกอริทึม GMM ไม่สามารถจัดกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ

	Model	Silhouette Score	Number of Clusters
0	DBSCAN without PCA	0.913339	2
1	DBSCAN with PCA	0.887072	3
2	Agglomerative Clustering without PCA	0.840405	2
3	Agglomerative Clustering with PCA	0.800971	2
4	K-Means with PCA	0.581552	4
5	K-Means without PCA	0.580628	4
6	GMM with PCA	0.147887	3
7	GMM without PCA	0.113903	3

ภาพประกอบ 42 แสดงผลลัพธ์ Silhouette Score ของแต่ละอัลกอริทึมและรูปแบบของข้อมูล

หลังจากได้ผลลัพธ์จากการจัดกลุ่มข้อมูลลูกค้าของแต่ละอัลกอริทึมและรูปแบบของข้อมูลแล้ว ผู้วิจัยได้ทำการเลือก 2 อัลกอริทึมที่มีประสิทธิภาพดีที่สุด ได้แก่ อัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ และอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ มาทำการศึกษาลักษณะลูกค้าของแต่ละคลัสเตอร์ภายในแต่ละอัลกอริทึม ซึ่งทำให้ได้ผลลัพธ์จากการศึกษา ดังนี้

ลักษณะลูกค้าที่ได้จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

จากข้อมูลในตารางที่ 31 แสดงให้เห็นว่า ลูกค้าในคลัสเตอร์ที่ -1 มีแนวโน้มเป็นลูกค้าประจำที่มีความภักดีร้านค้าปลีกสูง เนื่องจากคุณลักษณะ Recency มีค่า Mean เท่ากับ 10.83 วัน และ Median เท่ากับ 5.5 วัน บ่งบอกว่า ลูกค้ามาทำการซื้อสินค้าล่าสุดในช่วงเวลาไม่นานมาก ซึ่งหมายความว่า ลูกค้ากลุ่มนี้ มีการกลับมาซื้อซ้ำบ่อยครั้ง และมาซื้ออย่างสม่ำเสมอ

ด้านความถี่ของการซื้อ คุณลักษณะ Frequency ที่มีค่า Mean เท่ากับ 137.59 ครั้ง และ Median เท่ากับ 115 ครั้ง แสดงให้เห็นว่า ลูกค้ากลุ่มนี้มักทำธุรกรรมบ่อย และโดยเฉลี่ยแล้วจำนวนครั้งที่ซื้อนั้นสูงกว่ากลุ่มอื่นมาก ซึ่งอาจเป็นผลมาจากความเชื่อมั่นในแบรนด์หรือความพึงพอใจต่อสินค้าและบริการ และจากค่า Standard Deviation ที่มีค่า 57.77 ชี้ให้เห็นถึง ความแตกต่างของพฤติกรรมในลูกค้ากลุ่มนี้ โดยลูกค้าบางส่วนอาจซื้อสินค้าบ่อยมากกว่ากลุ่มอื่น แสดงถึง ความหลากหลายในการซื้อสินค้า และการเลือกสินค้าที่ตรงกับความต้องการ

และเมื่อพิจารณาด้านมูลค่าการใช้จ่ายจากคุณลักษณะ Monetary ที่มีค่า Mean เท่ากับ 7,218.63 และ Median เท่ากับ 6,135.65 บ่งบอกว่า ลูกค้ากลุ่มนี้มียอดใช้จ่ายต่อคำสั่งซื้อค่อนข้างสูง โดยอาจเป็นกลุ่มลูกค้าที่ซื้อสินค้าปริมาณมาก หรือเลือกซื้อสินค้าที่มีราคาสูง ส่งผลให้ลูกค้ากลุ่มนี้ เป็นลูกค้ากลุ่มสำคัญที่สร้างรายได้ให้กับธุรกิจ รวมถึงมีค่า Standard Deviation เท่ากับ 3,109.62 ที่ค่อนข้างกว้าง แสดงว่า มีลูกค้าบางกลุ่มที่มีการใช้จ่ายมากเป็นพิเศษ

จากการวิเคราะห์ข้อมูลข้างต้น พบว่า ลูกค้าในคลัสเตอร์ที่ -1 มีแนวโน้มเป็นลูกค้าที่มีศักยภาพสูง และควรได้รับการดูแลเป็นพิเศษ ธุรกิจอาจใช้กลยุทธ์รักษาความสัมพันธ์กับลูกค้าผ่านโปรแกรมสะสมแต้ม ข้อเสนอพิเศษ หรือโปรโมชั่นเฉพาะบุคคล เพื่อกระตุ้นให้พวกเขากลับมาซื้อซ้ำและเพิ่มมูลค่าต่อคำสั่งซื้อในอนาคต

ส่วนลูกค้าในคลัสเตอร์ที่ 0 นั้น มีแนวโน้มที่เป็นลูกค้าใหม่ หรือลูกค้าขาจร ที่ไม่ได้มีการซื้อสินค้าอย่างต่อเนื่อง สังเกตได้จากคุณลักษณะ Recency ที่มีค่า Mean เท่ากับ 600.34 วัน และ Median เท่ากับ 503 วัน สะท้อนให้เห็นว่าพวกเขาไม่ได้ทำธุรกรรมมานานมากแล้ว หรืออาจเคยซื้อสินค้าเพียงไม่กี่ครั้งในอดีต แต่ไม่ได้กลับมาซื้อต่อในช่วงที่ผ่านมา

ด้านความถี่ของการซื้อ คุณลักษณะ Frequency ที่มีค่า Mean เท่ากับ 2.94 ครั้ง และ Median เท่ากับ 1 ครั้ง ชี้ให้เห็นว่า ลูกค้ากลุ่มนี้ส่วนใหญ่มักซื้อสินค้าเพียงหนึ่งหรือสองครั้งเท่านั้น ค่า Median ที่ต่ำเมื่อเทียบกับ Mean บ่งบอกว่า มีลูกค้าส่วนน้อยที่ซื้อซ้ำหลายครั้ง แต่โดยรวมแล้วส่วนใหญ่เป็นลูกค้าที่ซื้อเพียงครั้งเดียวแล้วไม่ได้กลับมาอีก

เมื่อพิจารณาด้านมูลค่าการใช้จ่ายจากคุณลักษณะ Monetary ที่มีค่า Mean เท่ากับ 154.15 และ Median เท่ากับ 83.27 แสดงว่า ลูกค้ากลุ่มนี้มียอดใช้จ่ายที่ค่อนข้างต่ำ และไม่ได้เป็นกลุ่มที่สร้างรายได้หลักให้กับธุรกิจ

จากการวิเคราะห์ข้อมูลข้างต้น แสดงให้เห็นว่า ลูกค้าในคลัสเตอร์ที่ 0 มีแนวโน้มที่มีโอกาสในการพัฒนา ซึ่งธุรกิจสามารถใช้กลยุทธ์กระตุ้นให้เกิดการซื้อซ้ำ เช่น การส่งอีเมลแจ้งเตือนข้อเสนอพิเศษ โปรโมชั่นส่วนลดสำหรับการซื้อครั้งถัดไป หรือการให้สิทธิพิเศษเฉพาะลูกค้าใหม่ เพื่อเพิ่มโอกาสให้พวกเขากลายเป็นลูกค้าประจำในอนาคต

ตาราง 31 แสดงค่าทางสถิติของแต่ละคุณลักษณะจากผลลัพธ์การจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

คุณลักษณะ	**ค่าทางสถิติ**	Cluster -1	Cluster 0
Recency	Mean	10.83	600.34
	Median	5.5	503
	Standard Deviation	12.91	460.32
Frequency	Mean	137.59	2.94
	Median	115.0	1.0
	Standard Deviation	57.77	5.03
Monetary	Mean	7,218.63	154.15
	Median	6,135.65	83.27
	Standard Deviation	3,109.62	268.23

ลักษณะลูกค้าที่ได้จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ

จากข้อมูลในตารางที่ 32 แสดงให้เห็นว่า ลูกค้าในคลัสเตอร์ที่ -1 มีแนวโน้มเป็นลูกค้าประจำที่มีความภักดีสูง และสร้างรายได้หลักให้กับธุรกิจ พิจารณาจากคุณลักษณะ Recency ที่มีค่า Mean เท่ากับ 7.09 วัน และ Median เท่ากับ 4.0 วัน สะท้อนว่า ลูกค้ามีการซื้อสินค้าล่าสุดในช่วงเวลาไม่นานมาก และกลับมาซื้อซ้ำอย่างต่อเนื่อง

ในด้านความถี่ของการซื้อ คุณลักษณะ Frequency ที่มีค่า Mean เท่ากับ 172.74 ครั้ง และ Median เท่ากับ 139.0 ครั้ง แสดงให้เห็นว่า ลูกค้ากลุ่มนี้มักทำธุรกรรมบ่อยมากและมี

ความภักดีในระดับสูง ค่า Standard Deviation เท่ากับ 64.25 ชี้ให้เห็นถึงความแตกต่างของพฤติกรรมในกลุ่มนี้ โดยลูกค้าบางรายอาจซื้อบ่อยมากกว่ารายอื่น ซึ่งแสดงถึงความหลากหลายของกลุ่ม

ในด้านมูลค่าการใช้จ่าย คุณลักษณะ Monetary มีค่า Mean เท่ากับ 9,044.98 บาท และ Median เท่ากับ 7,521.87 บาท สะท้อนว่า ลูกค้ากลุ่มนี้มียอดใช้จ่ายเฉลี่ยต่อคำสั่งซื้อที่สูงมาก ซึ่งอาจมาจากการซื้อสินค้าปริมาณมาก หรือเลือกซื้อสินค้าที่มีราคาสูง โดยมีค่า Standard Deviation เท่ากับ 3,541.99 แสดงให้เห็นว่า มีลูกค้าบางกลุ่มที่มีการใช้จ่ายมากเป็นพิเศษ

จากการวิเคราะห์ข้อมูลข้างต้น พบว่า ลูกค้าในคลัสเตอร์ที่ -1 เป็นกลุ่มที่มีศักยภาพสูง ควรได้รับการดูแลเป็นพิเศษ ธุรกิจอาจใช้กลยุทธ์ในการรักษาความสัมพันธ์ เช่น การมอบข้อเสนอพิเศษแบบเฉพาะบุคคล โปรแกรมสะสมแต้ม หรือสิทธิพิเศษสำหรับลูกค้าระดับ VIP เพื่อกระตุ้นให้ลูกค้ากลุ่มนี้ยังคงภักดี และเพิ่มมูลค่าการซื้อในระยะยาว

ส่วนลูกค้าในคลัสเตอร์ที่ 0 มีแนวโน้มเป็นลูกค้าใหม่หรือลูกค้าขาจร ที่ไม่ได้ซื้อสินค้าอย่างต่อเนื่อง ซึ่งสามารถพิจารณาจากคุณลักษณะ Recency ที่มีค่า Mean เท่ากับ 600.29 วัน และ Median เท่ากับ 503.0 วัน สะท้อนให้เห็นว่า ลูกค้าไม่ได้ทำธุรกรรมมาเป็นเวลานาน หรืออาจเคยซื้อสินค้าเพียงไม่กี่ครั้งในอดีต แต่ไม่ได้กลับมาซื้อต่อ

ด้านความถี่ของการซื้อ คุณลักษณะ Frequency ที่มีค่า Mean เท่ากับ 2.95 ครั้ง และ Median เท่ากับ 1 ครั้ง ชี้ให้เห็นว่า ลูกค้ากลุ่มนี้ส่วนใหญ่มักซื้อสินค้าเพียงหนึ่งหรือสองครั้งเท่านั้น ค่า Median ที่ต่ำกว่า Mean บ่งบอกว่า มีลูกค้าส่วนน้อยที่ซื้อซ้ำหลายครั้ง แต่โดยรวมแล้ว ส่วนใหญ่เป็นลูกค้าที่ซื้อเพียงครั้งเดียวแล้วไม่ได้กลับมาอีก

เมื่อพิจารณาด้านมูลค่าการใช้จ่ายจากคุณลักษณะ Monetary ที่มีค่า Mean เท่ากับ 154.57 บาท และ Median เท่ากับ 83.27 บาท แสดงว่า ลูกค้ากลุ่มนี้มียอดใช้จ่ายที่ค่อนข้างต่ำ และไม่ได้เป็นกลุ่มที่สร้างรายได้หลักให้กับธุรกิจ

จากการวิเคราะห์ข้อมูลข้างต้น พบว่า ลูกค้าในคลัสเตอร์ที่ 0 เป็นลูกค้ากลุ่มที่มีโอกาสในการพัฒนา ซึ่งธุรกิจสามารถใช้กลยุทธ์กระตุ้นให้เกิดการซื้อซ้ำ เช่น การส่งอีเมลแจ้งเตือนข้อเสนอพิเศษ โปรโมชั่นส่วนลดสำหรับการซื้อครั้งถัดไป หรือการให้สิทธิพิเศษเฉพาะลูกค้าใหม่ เพื่อเพิ่มโอกาสให้พวกเขากลายเป็นลูกค้าประจำในอนาคต

และสำหรับลูกค้าในคลัสเตอร์ที่ 1 นั้น มีแนวโน้มที่เป็นลูกค้าประจำที่มีมูลค่าการซื้อสูง และมีการซื้อสินค้าต่อเนื่อง พิจารณาจากคุณลักษณะ Recency ที่มีค่า Mean เท่ากับ 11.00

วัน และ Median เท่ากับ 14.0 วัน แสดงว่า ลูกค้ามีการทำธุรกรรมอย่างสม่ำเสมอ และยังคงซื้อสินค้าภายในช่วงเวลาที่ไม่นานมาก

ในแง่ของความถี่ของการซื้อสินค้า คุณลักษณะ Frequency ที่มีค่า Mean เท่ากับ 106.88 ครั้ง และ Median เท่ากับ 106.0 ครั้ง แสดงให้เห็นว่า ลูกค้ากลุ่มนี้ทำธุรกรรมบ่อยมากและมีความภักดีต่อแบรนด์สูง ลูกค้ากลุ่มนี้มีการซื้อสินค้าซ้ำอย่างสม่ำเสมอ โดยมีค่า Standard Deviation เพียง 5.12 ซึ่งต่ำ แสดงถึงพฤติกรรมการซื้อที่ค่อนข้างคล้ายกันในกลุ่มนี้

ด้านของมูลค่าการใช้จ่าย คุณลักษณะ Monetary มีค่า Mean เท่ากับ 5,608.68 บาท และ Median เท่ากับ 5,609.46 บาท สะท้อนว่า ลูกค้ากลุ่มนี้มีมูลค่าการซื้อที่ค่อนข้างสูง โดยมีค่า Standard Deviation เท่ากับ 331.64 แสดงให้เห็นว่า ลูกค้าในกลุ่มนี้มีลักษณะการใช้จ่ายที่ใกล้เคียงกัน

จากการวิเคราะห์ข้อมูลข้างต้น แสดงให้เห็นว่า ลูกค้าในคลัสเตอร์ที่ 1 เป็น ลูกค้าประจำที่มีมูลค่าการซื้อสูง ธุรกิจควรเน้นการรักษาความสัมพันธ์ และเพิ่มมูลค่าการซื้อในอนาคต โดยการนำเสนอ Personalized Offers ที่ตรงกับความสนใจของลูกค้า และการสร้าง Loyalty Program เพื่อกระตุ้นให้ลูกค้าใช้จ่ายมากขึ้น รวมถึงการเสนอสินค้าใหม่ หรือข้อเสนอพิเศษสำหรับลูกค้าที่มีการใช้จ่ายสูง จะช่วยรักษาความภักดีและเพิ่มโอกาสในการทำธุรกรรมในอนาคต

ตาราง 32 แสดงค่าทางสถิติของแต่ละคุณลักษณะจากผลลัพธ์การจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ

คุณลักษณะ	**ค่าทางสถิติ**	Cluster -1	Cluster 0	Cluster 1
Recency	Mean	7.09	600.29	11
	Median	4	503	14
	Standard Deviation	10.37	460.33	7.66
Frequency	Mean	172.74	2.95	106.88
	Median	139	1	106
	Standard Deviation	64.25	5.10	5.12
Monetary	Mean	9,044.98	154.57	5,608.68
	Median	7,521.87	83.27	5,609.46
	Standard Deviation	3,541.99	271.84	331.64

4.2 ผลลัพธ์ของกฎความสัมพันธ์จากข้อมูลการจัดกลุ่มลูกค้าภายในร้านค้าปลีก

การวิเคราะห์ตระกร้าสินค้า คือ เทคนิคการวิเคราะห์ข้อมูลที่ใช้ในการตรวจสอบการซื้อสินค้าที่มักเกิดขึ้นร่วมกันในครั้งเดียว โดยจะใช้กฎความสัมพันธ์ในการตรวจหาความเชื่อมโยงระหว่างสินค้าที่ลูกค้ามักจะซื้อพร้อมกัน

ในงานวิจัยนี้ ผู้วิจัยได้นำอัลกอริทึม FP-Growth มาใช้ในการวิเคราะห์หากฎความสัมพันธ์ของข้อมูลการจัดกลุ่มที่ได้จาก 2 อัลกอริทึมที่มีประสิทธิภาพดีที่สุด ได้แก่ อัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ และอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ โดยจะทำการศึกษาผลลัพธ์ที่ได้จากการทดลองปรับค่าพารามิเตอร์ภายใต้เงื่อนไขที่กำหนด คือ 1. ทดลองปรับค่า Min Support ที่แตกต่างกัน 3 ระดับ ได้แก่ 0.01, 0.002, และ 0.001, 2. กำหนดค่า Min Confidence เท่ากับ 0.01, และ 3. เลือกเฉพาะกฎที่มีค่า Lift มากกว่า 1

4.2.1 ข้อมูลการจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

การจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ ทำให้สามารถจัดกลุ่มข้อมูลได้เป็น 2 คลัสเตอร์ แบ่งเป็นคลัสเตอร์ที่ -1 มีจำนวนข้อมูล เท่ากับ 46 และคลัสเตอร์ที่ 0 มีจำนวนข้อมูล เท่ากับ 65,901 ซึ่งทำการศึกษาด้วยการทดลองปรับค่าพารามิเตอร์ในระดับที่แตกต่างกัน โดยทำให้ได้ผลลัพธ์จากการทดลอง ดังนี้

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.01

จากการทดลองวิเคราะห์หากฎความสัมพันธ์ด้วยการกำหนดค่าพารามิเตอร์ตามเงื่อนไข พบว่า ทั้งสองคลัสเตอร์ไม่มีข้อมูลสินค้าที่มีกฎความสัมพันธ์จากค่าที่กำหนด ซึ่งอาจเกิดจากการที่ค่าพารามิเตอร์ที่กำหนดไว้นั้น มีค่าสูงเกินไป เนื่องจากชุดข้อมูลที่ใช้เป็นข้อมูลธุรกรรมภายในร้านค้าปลีก ซึ่งมักมีจำนวนธุรกรรมที่มาก ทำให้การกำหนดค่า Support, Confidence, หรือ Lift ที่สูงเกินไป อาจทำให้ไม่สามารถค้นพบกฎความสัมพันธ์ที่มีความหมายได้ เพราะกรองข้อมูลบางส่วนออกไป จนเหลือเฉพาะกฎที่หายากหรือไม่เกิดขึ้นบ่อย

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.001

เมื่อทำการทดลองปรับลดค่า min_support ลง เหลือเท่ากับ 0.001 พบว่า คลัสเตอร์ที่ -1 มีจำนวนกฎความสัมพันธ์ เท่ากับ 2,328 ซึ่งจากภาพประกอบที่ 43 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Banana ⟶ Toothpaste ซึ่งมีค่า Support, Confidence, และ Lift มากที่สุด รองลงมา คือ Ketchup ⟶ Toothpaste และ Toothpaste ⟶ Banana ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
1534	(Banana)	(Toothpaste)	0.034919	0.072997	0.003476	0.099548	1.363715
1718	(Ketchup)	(Toothpaste)	0.036341	0.072997	0.003476	0.095652	1.310352
1535	(Toothpaste)	(Banana)	0.072997	0.034919	0.003476	0.047619	1.363715
1719	(Toothpaste)	(Ketchup)	0.072997	0.036341	0.003476	0.047619	1.310352
1376	(Potatoes)	(Toothpaste)	0.034445	0.072997	0.003318	0.096330	1.319641

ภาพประกอบ 43 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ -1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.001

ส่วนผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 0 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 252 และจากภาพประกอบที่ 44 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Ketchup ⟶ Toothpaste รองลงมา คือ Toothpaste ⟶ Ketchup และ Hair Gel ⟶ Toothpaste ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
992	(Ketchup)	(Toothpaste)	0.036347	0.071116	0.002606	0.071692	1.008107
993	(Toothpaste)	(Ketchup)	0.071116	0.036347	0.002606	0.036642	1.008107
318	(Hair Gel)	(Toothpaste)	0.035831	0.071116	0.002565	0.071573	1.006425
319	(Toothpaste)	(Hair Gel)	0.071116	0.035831	0.002565	0.036062	1.006425
2798	(Peanut Butter)	(Orange)	0.036523	0.036631	0.001455	0.039842	1.087648

ภาพประกอบ 44 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 0 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.001

จากผลลัพธ์ที่ได้ของทั้งสองคลัสเตอร์ พบว่า กฎความสัมพันธ์ที่ได้จากข้อมูลในคลัสเตอร์ที่ -1 ที่มีค่าสูงสุด คือ Banana ⟶ Toothpaste มีค่า Support เท่ากับ 0.0035 ค่า Confidence เท่ากับ 0.0995 และค่า Lift เท่ากับ1.3637 โดยมีค่ามากกว่าค่าสูงสุดของกฎความสัมพันธ์ที่ได้จากข้อมูลในคลัสเตอร์ที่ 0 คือ Ketchup ⟶ Toothpaste มีค่า Support เท่ากับ 0.0026 ค่า Confidence เท่ากับ 0.0717 และค่า Lift เท่ากับ 1.0081

โดยสามารถอธิบายได้ว่า โอกาสที่ลูกค้าในคลัสเตอร์ที่ -1 ที่ซื้อ Banana และจะซื้อ Toothpaste ร่วมด้วย มีมากกว่าโอกาสที่ลูกค้าในคลัสเตอร์ที่ 0 ที่ซื้อ Ketchup แล้วจะซื้อ Toothpaste และเมื่อพิจารณาถึงค่า Lift ก็พบว่า สินค้า Banana และ Toothpaste ที่ได้จากคลัส

เตอร์ -1 มีแนวโน้มถูกซื้อร่วมกันมากกว่าที่จะเกิดขึ้นจากการซื้อแบบสุ่ม ในขณะที่ ค่า Lift ของคลัสเตอร์ 0 มีค่าใกล้เคียง 1 แสดงว่า กฎ Ketchup ⟶ Toothpaste อาจไม่มีความสัมพันธ์เชิงบวกที่ชัดเจน

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.002

จากการทดลองปรับค่า min_support เท่ากับ 0.002 พบว่า คลัสเตอร์ที่ -1 มีจำนวนกฎความสัมพันธ์ เท่ากับ 268 ซึ่งจากภาพประกอบที่ 45 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Banana ⟶ Toothpaste ซึ่งมีค่า Support, Confidence, และ Lift มากที่สุด รองลงมา คือ Ketchup ⟶ Toothpaste และ Toothpaste ⟶ Banana ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
144	(Banana)	(Toothpaste)	0.034919	0.072997	0.003476	0.099548	1.363715
152	(Ketchup)	(Toothpaste)	0.036341	0.072997	0.003476	0.095652	1.310352
145	(Toothpaste)	(Banana)	0.072997	0.034919	0.003476	0.047619	1.363715
153	(Toothpaste)	(Ketchup)	0.072997	0.036341	0.003476	0.047619	1.310352
136	(Potatoes)	(Toothpaste)	0.034445	0.072997	0.003318	0.096330	1.319641

ภาพประกอบ 45 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ -1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.002

ส่วนผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 0 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 4 และจากภาพประกอบที่ 46 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Ketchup ⟶ Toothpaste รองลงมา คือ Toothpaste ⟶ Ketchup และ Hair Gel ⟶ Toothpaste ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
22	(Ketchup)	(Toothpaste)	0.036347	0.071116	0.002606	0.071692	1.008107
23	(Toothpaste)	(Ketchup)	0.071116	0.036347	0.002606	0.036642	1.008107
6	(Hair Gel)	(Toothpaste)	0.035831	0.071116	0.002565	0.071573	1.006425
7	(Toothpaste)	(Hair Gel)	0.071116	0.035831	0.002565	0.036062	1.006425

ภาพประกอบ 46 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 0 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.002

จากผลลัพธ์ที่ได้ของทั้งสองคลัสเตอร์ พบว่า กฎความสัมพันธ์ที่ได้จากข้อมูลในคลัสเตอร์ที่ -1 ที่มีค่าสูงสุด คือ Banana ⟶ Toothpaste มีค่า Support เท่ากับ 0.0035 ค่า Confidence เท่ากับ 0.0995 และค่า Lift เท่ากับ1.3637 ส่วนกฎความสัมพันธ์ที่ได้จากข้อมูลในคลัสเตอร์ที่ 0 ที่มีค่าสูงสุด คือ Ketchup ⟶ Toothpaste มีค่า Support เท่ากับ 0.0026 ค่า Confidence เท่ากับ 0.0717 และค่า Lift เท่ากับ 1.0081 ซึ่งผลลัพธ์ที่มาได้นั้น เป็นกฎความสัมพันธ์เดียวกันที่ได้จากการทดลองปรับค่า min_support = 0.001

4.2.2 ข้อมูลการจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ

การจัดกลุ่มของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ ทำให้สามารถจัดกลุ่มข้อมูลได้เป็น 3 คลัสเตอร์ แบ่งเป็นคลัสเตอร์ที่ -1 มีจำนวนข้อมูล เท่ากับ 23 คลัสเตอร์ที่ 0 มีจำนวนข้อมูล เท่ากับ 65,907 และคลัสเตอร์ที่ 1 มีจำนวนข้อมูล เท่ากับ 17 ซึ่งทำการศึกษาด้วยการทดลองปรับค่าพารามิเตอร์ในระดับที่แตกต่างกัน โดยทำให้ได้ผลลัพธ์จากการทดลอง ดังนี้

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.01

จากการทดลองวิเคราะห์หากฎความสัมพันธ์ด้วยการกำหนดค่าพารามิเตอร์ตามเงื่อนไข พบว่า ทั้งสองคลัสเตอร์ไม่มีข้อมูลสินค้าที่มีกฎความสัมพันธ์จากค่าที่กำหนด ซึ่งอาจเกิดจากการที่ค่าพารามิเตอร์ที่กำหนดไว้นั้น มีค่าสูงเกินไป เนื่องจากชุดข้อมูลที่ใช้เป็นข้อมูลธุรกรรมภายในร้านค้าปลีก ซึ่งมักมีจำนวนธุรกรรมที่มาก ทำให้การกำหนดค่า Support, Confidence, หรือ Lift ที่สูงเกินไป อาจทำให้ไม่สามารถค้นพบกฎความสัมพันธ์ที่มีความหมายได้ เพราะกรองข้อมูลบางส่วนออกไป จนเหลือเฉพาะกฎที่หายากหรือไม่เกิดขึ้นบ่อย

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.001

เมื่อทำการทดลองปรับลดค่า min_support ลง เหลือเท่ากับ 0.001 พบว่า คลัสเตอร์ที่ -1 มีจำนวนกฎความสัมพันธ์ เท่ากับ 2,382 ซึ่งจากภาพประกอบที่ 47 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Banana ⟶ Toothpaste รองลงมา คือ Potatoes ⟶ Toothpaste และ Toothpaste ⟶ Banana ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
1558	(Banana)	(Toothpaste)	0.036748	0.072993	0.004027	0.109589	1.501370
1476	(Potatoes)	(Toothpaste)	0.037000	0.072993	0.004027	0.108844	1.491156
1559	(Toothpaste)	(Banana)	0.072993	0.036748	0.004027	0.055172	1.501370
1477	(Toothpaste)	(Potatoes)	0.072993	0.037000	0.004027	0.055172	1.491156
3896	(Ketchup)	(Toothpaste)	0.036748	0.072993	0.003775	0.102740	1.407534

ภาพประกอบ 47 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ -1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.001

ส่วนผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 0 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 256 และจากภาพประกอบที่ 48 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Ketchup ⟶ Toothpaste รองลงมา คือ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
980	(Ketchup)	(Toothpaste)	0.036365	0.071109	0.002619	0.072025	1.012883
981	(Toothpaste)	(Ketchup)	0.071109	0.036365	0.002619	0.036833	1.012883
316	(Hair Gel)	(Toothpaste)	0.035871	0.071109	0.002563	0.071439	1.004641
317	(Toothpaste)	(Hair Gel)	0.071109	0.035871	0.002563	0.036037	1.004641
2782	(Peanut Butter)	(Orange)	0.036519	0.036612	0.001456	0.039876	1.089156

Toothpaste ⟶ Ketchup และ Hair Gel ⟶ Toothpaste ตามลำดับ

ภาพประกอบ 48 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 0 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.001

และด้านผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 1 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 3,566 และจากภาพประกอบที่ 49 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Laundry Detergent ⟶ Trash Cans รองลงมา คือ Trash Cans ⟶ Laundry Detergent และ Cheese ⟶ Toothpaste ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
4668	(Laundry Detergent)	(Trash Cans)	0.039075	0.041277	0.006604	0.169014	4.094648
4669	(Trash Cans)	(Laundry Detergent)	0.041277	0.039075	0.006604	0.160000	4.094648
4520	(Cheese)	(Toothpaste)	0.036874	0.074298	0.005504	0.149254	2.008845
4521	(Toothpaste)	(Cheese)	0.074298	0.036874	0.005504	0.074074	2.008845
1238	(Shower Gel)	(Toothpaste)	0.035773	0.074298	0.004953	0.138462	1.863590

ภาพประกอบ 49 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.001

จากผลลัพธ์ที่ได้จากคลัสเตอร์ที่ -1 พบว่า ลูกค้าส่วนใหญ่มีแนวโน้มซื้อสินค้าหลายประเภทในครั้งเดียว โดยเฉพาะสินค้าที่เกี่ยวข้องกับการใช้ชีวิตประจำวัน เช่น Banana ⟶ Toothpaste และ Potatoes ⟶ Toothpaste ซึ่งทั้งสองกฎมีค่า Lift ที่สูงกว่า 1 แสดงถึงความสัมพันธ์ที่มีความน่าจะเป็นสูงกว่าการซื้อแบบสุ่ม บ่งบอกว่า ลูกค้ากลุ่มนี้อาจมีพฤติกรรมการซื้อสินค้าหลายรายการในคำสั่งซื้อเดียวกัน เช่น การซื้อของสดและสินค้าที่ใช้ในชีวิตประจำวันพร้อมกัน ซึ่งอาจเป็นลักษณะของลูกค้าที่ซื้อสินค้าประจำอย่างมีแบบแผน

ส่วนลูกค้าในคลัสเตอร์ที่ 0 มีพฤติกรรมการซื้อสินค้าที่ไม่ค่อยมีความสัมพันธ์ของสินค้า เช่น Ketchup ⟶ Toothpaste ซึ่งมีค่า Lift ใกล้เคียง 1 แสดงว่า ไม่มีความสัมพันธ์เชิงบวกที่ชัดเจนระหว่างสินค้าทั้งสอง ซึ่งสะท้อนให้เห็นว่า ลูกค้าในกลุ่มนี้อาจเป็นกลุ่มลูกค้าขาจรที่ซื้อสินค้าผ่านโปรโมชั่นหรือซื้อแบบไม่ได้ตั้งใจ โดยสินค้าในกลุ่มนี้อาจไม่ค่อยถูกซื้อร่วมกันเป็นประจำ

ในขณะที่ลูกค้าในคลัสเตอร์ที่ 1 มีพฤติกรรมการซื้อสินค้าจำเป็นที่มีความสัมพันธ์ชัดเจน เช่น Laundry Detergent ⟶ Trash Cans ซึ่งมีค่า Lift ที่สูงมาก แสดงให้เห็นว่า ลูกค้าในกลุ่มนี้มักซื้อสินค้าที่เกี่ยวข้องกันอย่างสม่ำเสมอ และมีแนวโน้มซื้อสินค้าประเภทของใช้ภายในบ้านในปริมาณที่มากขึ้น เช่น การซื้อผงซักฟอกและถังขยะพร้อมกัน ซึ่งบ่งบอกถึงพฤติกรรมการซื้อที่มีแบบแผนและมักจะซื้อในปริมาณที่มากตามความจำเป็น

ผลลัพธ์ที่ได้จากการปรับค่า min_support = 0.002

จากการทดลองปรับค่า min_support เท่ากับ 0.002 พบว่า คลัสเตอร์ที่ -1 มีจำนวนกฎความสัมพันธ์ เท่ากับ 710 ซึ่งจากภาพประกอบที่ 50 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Banana ⟶ Toothpaste รองลงมา คือ Potatoes ⟶ Toothpaste และ Toothpaste ⟶ Banana ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
324	(Banana)	(Toothpaste)	0.036748	0.072993	0.004027	0.109589	1.501370
308	(Potatoes)	(Toothpaste)	0.037000	0.072993	0.004027	0.108844	1.491156
325	(Toothpaste)	(Banana)	0.072993	0.036748	0.004027	0.055172	1.501370
309	(Toothpaste)	(Potatoes)	0.072993	0.037000	0.004027	0.055172	1.491156
672	(Ketchup)	(Toothpaste)	0.036748	0.072993	0.003775	0.102740	1.407534

ภาพประกอบ 50 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ -1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.002

ส่วนผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 0 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 4 และจากภาพประกอบที่ 51 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Ketchup ⟶ Toothpaste รองลงมา คือ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
22	(Ketchup)	(Toothpaste)	0.036365	0.071109	0.002619	0.072025	1.012883
23	(Toothpaste)	(Ketchup)	0.071109	0.036365	0.002619	0.036833	1.012883
6	(Hair Gel)	(Toothpaste)	0.035871	0.071109	0.002563	0.071439	1.004641
7	(Toothpaste)	(Hair Gel)	0.071109	0.035871	0.002563	0.036037	1.004641

Toothpaste ⟶ Ketchup และ Hair Gel ⟶ Toothpaste ตามลำดับ

ภาพประกอบ 51 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 0 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.002

และด้านผลลัพธ์จากการทดลองที่ได้จากคลัสเตอร์ที่ 1 พบว่า มีจำนวนกฎความสัมพันธ์ เท่ากับ 1,032 และจากภาพประกอบที่ 52 แสดงให้เห็นว่า กฎความสัมพันธ์ที่มีค่า Support, Confidence, และ Lift มากที่สุด คือ Laundry Detergent ⟶ Trash รองลงมา คือ Trash Cans ⟶ Laundry Detergent และ Cheese ⟶ Toothpaste ตามลำดับ

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
990	(Laundry Detergent)	(Trash Cans)	0.039075	0.041277	0.006604	0.169014	4.094648
991	(Trash Cans)	(Laundry Detergent)	0.041277	0.039075	0.006604	0.160000	4.094648
972	(Cheese)	(Toothpaste)	0.036874	0.074298	0.005504	0.149254	2.008845
973	(Toothpaste)	(Cheese)	0.074298	0.036874	0.005504	0.074074	2.008845
244	(Shower Gel)	(Toothpaste)	0.035773	0.074298	0.004953	0.138462	1.863590

ภาพประกอบ 52 แสดงกฎความสัมพันธ์ 5 อันดับแรกของคลัสเตอร์ที่ 1 จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_support = 0.002

จากผลลัพธ์ที่ได้จากคลัสเตอร์ที่ -1 พบว่า กฎความสัมพันธ์ที่สำคัญ คือ Banana ⟶ Toothpaste มีค่า Lift เท่ากับ 1.5014 ซึ่งสูงกว่ากฎความสัมพันธ์อื่นในคลัสเตอร์นี้ เช่น Potatoes ⟶ Toothpaste และ Ketchup ⟶ Toothpaste กฎความสัมพันธ์เหล่านี้มีค่า Lift สูงกว่า 1 แสดงถึง ความสัมพันธ์ที่มีนัยสำคัญระหว่างสินค้าที่ถูกซื้อร่วมกันในระดับที่มากกว่าการซื้อแบบสุ่ม ซึ่งอาจบ่งชี้ได้ว่า ลูกค้าในคลัสเตอร์ที่ -1 มีแนวโน้มที่ซื้อสินค้าที่ไม่เกี่ยวข้องกัน เช่น กล้วยและยาสีฟัน หรือมันฝรั่งกับยาสีฟัน ในคำสั่งซื้อเดียวกัน

ส่วนในคลัสเตอร์ที่ 0 กฎความสัมพันธ์ที่สำคัญ คือ Ketchup ⟶ Toothpaste ซึ่งมีค่า Lift เท่ากับ 1.0129 แสดงถึง ความสัมพันธ์ที่ไม่ชัดเจนระหว่างสินค้าทั้งสอง เนื่องจากค่า Lift มีค่าใกล้เคียง 1 โดยลูกค้าในกลุ่มนี้อาจซื้อสินค้าที่ไม่เกี่ยวข้องกัน เช่น Ketchup และ Toothpaste ในกรณีที่มีโปรโมชั่นพิเศษหรือลูกค้าขาจรที่ซื้อสินค้าตามความสะดวกและโอกาส

ในขณะที่คลัสเตอร์ที่ 1 มีกฎความสัมพันธ์ที่สำคัญ เช่น Laundry Detergent ⟶ Trash Cans ซึ่งมีค่า Lift ที่สูงมาก เท่ากับ 4.0946 แสดงถึง ความสัมพันธ์ที่สูงระหว่างสินค้าประเภทของใช้ในบ้าน ลูกค้าในคลัสเตอร์นี้มักซื้อสินค้าที่จำเป็นใช้ในชีวิตประจำวัน เช่น ผงซักฟอกและถังขยะพร้อมกัน โดยมีการซื้อสินค้าเหล่านี้ในปริมาณที่มากขึ้น บ่งบอกถึง พฤติกรรมการซื้อที่มีแบบแผน และมักจะซื้อในปริมาณที่มากอย่างต่อเนื่อง เช่นเดียวกับ กฎ Cheese ⟶ Toothpaste ที่มีค่า Lift เท่ากับ 2.0088 ซึ่งแสดงถึงความสัมพันธ์เชิงบวกที่ชัดเจนมากในการซื้อสินค้าในตะกร้าสินค้าเดียวกัน

จากการทดลองตลอดจนได้ผลลัพธ์กฎความสัมพันธ์ของข้อมูลที่ได้มาจากแต่ละอัลกอริทึมการจัดกลุ่ม โดยจากกฎความสัมพันธ์ที่ได้จากข้อมูลของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ ดังแสดงในตารางที่ 33 พบว่า คลัสเตอร์ที่ -1 มีกฎความสัมพันธ์ที่

สำคัญ ได้แก่ Banana ⟶ Toothpaste และ Ketchup ⟶ Toothpaste ซึ่งมีค่า Support เท่ากับ 0.003476 และ Confidence ที่ค่อนข้างสูง เมื่อเทียบกับคลัสเตอร์อื่น โดยเฉพาะกฎ Banana ⟶ Toothpaste ที่มีค่า Confidence เท่ากับ 0.099548 และ Lift เท่ากับ 1.363715 แสดงให้เห็นว่าลูกค้าที่ซื้อกล้วยมักจะซื้อยาสีฟันร่วมด้วยในอัตราสูงกว่าการซื้อแบบสุ่ม

ด้านคลัสเตอร์ที่ 0 มีกฎความสัมพันธ์ที่น่าสนใจ ได้แก่ Ketchup ⟶ Toothpaste และ Hair Gel ⟶ Toothpaste ซึ่งมีค่า Support และค่า Lift ที่ต่ำกว่า เมื่อเทียบกับคลัสเตอร์ที่ -1 ซึ่งหมายความว่า ความสัมพันธ์ระหว่างสินค้าของคลัสเตอร์นี้ จะไม่สูงเท่ากับคลัสเตอร์ที่ -1

หากทำการเปรียบเทียบกันระหว่าง 2 คลัสเตอร์ จะพบว่า คลัสเตอร์ที่ -1 จะมีความสัมพันธ์ระหว่างสินค้าที่สูงกว่า โดยเฉพาะกฎ Banana ⟶ Toothpaste ที่มีค่า Lift สูงกว่า 1.3 ซึ่งแสดงให้เห็นถึง แนวโน้มการซื้อสินค้าร่วมกันที่น่าสนใจ ส่วนคลัสเตอร์ที่ 0 นั้น จะมีความสัมพันธ์ระหว่างสินค้าต่ำกว่าเล็กน้อย และค่า Lift ที่ใกล้เคียง 1 บ่งบอกว่า สินค้าเหล่านี้ไม่ได้มีความสัมพันธ์กันมากนัก

การนำไปประยุกต์ใช้กับธุรกิจ สามารถนำข้อมูลกฎความสัมพันธ์ที่ได้จากคลัสเตอร์ที่ -1 ไปใช้ในการออกโปรโมชั่นสำหรับลูกค้าที่อยู่ในกลุ่ม เช่น การจัดชุดสินค้า หรือการจัดโปรโมชั่นในการซื้อคู่กันสำหรับ Banana และ Toothpaste หรือ Ketchup และ Toothpaste เพื่อกระตุ้นยอดขาย ส่วนในคลัสเตอร์ที่ 0 อาจต้องหาสินค้าอื่นมาช่วยเพิ่มความสัมพันธ์ เช่น การจัดโปรโมชันร่วมกับสินค้าที่เกี่ยวข้องกับ Hair Gel เป็นต้น

ตาราง 33 แสดงข้อมูลกฎความสัมพันธ์ 3 อันดับแรกที่ได้จากการทดลองปรับค่า min_support ของอัลกอริทึม DSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ

Cluster	min_support	Antecedent	Consequent	Support	Confidence	Lift
Cluster = -1	0.001	Banana	Toothpaste	0.003476	0.099548	1.363715
		Ketchup	Toothpaste	0.003476	0.095652	1.310352
		Toothpaste	Banana	0.003476	0.047619	1.363715
	0.002	Banana	Toothpaste	0.003476	0.099548	1.363715
		Ketchup	Toothpaste	0.003476	0.095652	1.310352
		Toothpaste	Banana	0.003476	0.047619	1.363715
Cluster = 0	0.001	Ketchup	Toothpaste	0.002606	0.071692	1.008107
		Toothpaste	Ketchup	0.002606	0.036642	1.008107
		Hair Gel	Toothpaste	0.002565	0.071573	1.006425
	0.002	Ketchup	Toothpaste	0.002606	0.071692	1.008107
		Toothpaste	Ketchup	0.002606	0.036642	1.008107
		Hair Gel	Toothpaste	0.002565	0.071573	1.006425

ทางด้านกฎความสัมพันธ์ที่ได้จากข้อมูลของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ ดังแสดงในตารางที่ 34 พบว่า คลัสเตอร์ที่ -1 มีกฎความสัมพันธ์ที่สำคัญ ได้แก่ Banana ⟶ Toothpaste และ Potatoes ⟶ Toothpaste โดยมีค่า Support เท่ากับ 0.004027 และ Confidence ที่ค่อนข้างสูง เมื่อเทียบกับคลัสเตอร์อื่น โดยเฉพาะกฎ Banana ⟶ Toothpaste ที่มีค่า Confidence เท่ากับ 0.1096 และ Lift เท่ากับ 1.5014 แสดงให้เห็นว่า ลูกค้าที่ซื้อกล้วยมักจะซื้อยาสีฟันร่วมด้วยในอัตราที่สูงกว่าการซื้อแบบสุ่ม นอกจากนี้ ยังพบว่า กฎ Potatoes ⟶ Toothpaste มีค่า Lift ที่สูงถึง 1.4912 ซึ่งบ่งบอกถึงความสัมพันธ์ที่ค่อนข้างเด่นชัดระหว่างสินค้าทั้งสองชนิด

ด้านคลัสเตอร์ที่ 0 พบว่า มีกฎความสัมพันธ์ที่น่าสนใจ ได้แก่ Ketchup ⟶ Toothpaste และ Hair Gel ⟶ Toothpaste ซึ่งมีค่า Support และ Lift ที่ต่ำกว่าคลัสเตอร์ที่ -1 แสดงให้เห็นถึง ความสัมพันธ์ที่ไม่สูงเท่ากับคลัสเตอร์ที่ -1 โดยเฉพาะกฎ Ketchup ⟶ Toothpaste ที่มีค่า Lift เท่ากับ 1.0129 และกฎ Hair Gel ⟶ Toothpaste ที่มีค่า Lift เท่ากับ 1.0046 แสดงให้เห็นถึง ความสัมพันธ์ระหว่างสินค้าที่ต่ำ และมีแนวโน้มที่เป็นการซื้อแบบสุ่ม

ส่วนด้านคลัสเตอร์ที่ 1 พบว่า กฎ Laundry Detergent ⟶ Trash Cans และ Trash Cans ⟶ Laundry Detergent มีค่า Lift สูงถึง 4.0946 ขณะที่กฎ Cheese ⟶ Toothpaste มีค่า Lift อยู่ที่ 2.0088 แสดงให้เห็นถึงความสัมพันธ์ที่สูงมากระหว่างสินค้าทั้งสองชนิด โดยเฉพาะ Laundry Detergent ⟶ Trash Cans ซึ่งมีค่า Confidence เท่ากับ 0.1690 และ Lift เท่ากับ 4.0946 สะท้อนให้เห็นว่า ผงซักฟอกมักถูกซื้อร่วมกับถังขยะในอัตราที่สูงมาก

การนำไปประยุกต์ใช้กับธุรกิจ สามารถนำข้อมูลกฎความสัมพันธ์ที่ได้จากคลัสเตอร์ที่ -1 ไปใช้ในการออกโปรโมชั่น เช่น การจัดชุดสินค้า หรือการทำโปรโมชั่นซื้อคู่กันระหว่าง Banana และ Toothpaste หรือ Potatoes และ Toothpaste เพื่อกระตุ้นยอดขาย ในขณะที่คลัสเตอร์ที่ 0 สามารถใช้ข้อมูลเหล่านี้เพื่อหาโอกาสในการทำโปรโมชันร่วมกันระหว่าง Ketchup และ Toothpaste หรือ Hair Gel และ Toothpaste แต่เนื่องจากความสัมพันธ์ที่ค่อนข้างต่ำอาจต้องมีการเพิ่มสินค้าอื่น เพื่อเสริมความสัมพันธ์ของสินค้าเหล่านี้ให้สูงขึ้น ส่วนในคลัสเตอร์ที่ 1 สามารถใช้กฎ Laundry Detergent ⟶ Trash Cans ในการออกโปรโมชั่นสำหรับลูกค้าที่ซื้อสินค้าในหมวดของใช้ในบ้าน เช่น การจัดโปรโมชันที่มี Laundry Detergent และ Trash Cans หรือ Cheese และ Toothpaste เพื่อดึงดูดให้ลูกค้าซื้อสินค้าทั้งสองประเภทพร้อมกัน ซึ่งจะช่วยกระตุ้นยอดขายในกลุ่มสินค้าดังกล่าว

ตาราง 34 แสดงข้อมูลกฎความสัมพันธ์ 3 อันดับแรกที่ได้จากการทดลองปรับค่า min_support ของอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ

Cluster	min_support	Antecedent	Consequent	Support	Confidence	Lift
Cluster = -1	0.001	Banana	Toothpaste	0.004027	0.109589	1.501370
		Potatoes	Toothpaste	0.004027	0.108844	1.491156
		Toothpaste	Banana	0.004027	0.055172	1.501370
	0.002	Banana	Toothpaste	0.004027	0.109589	1.501370
		Potatoes	Toothpaste	0.004027	0.108844	1.491156
		Toothpaste	Banana	0.004027	0.055172	1.501370
Cluster = 0	0.001	Ketchup	Toothpaste	0.002619	0.072025	1.012883
		Toothpaste	Ketchup	0.002619	0.036833	1.012883
		Hair Gel	Toothpaste	0.002563	0.071439	1.004641
	0.002	Ketchup	Toothpaste	0.002619	0.072025	1.012883
		Toothpaste	Ketchup	0.002619	0.036833	1.012883
		Hair Gel	Toothpaste	0.002563	0.071439	1.004641
Cluster = 1	0.001	Laundry Detergent	Trash Cans	0.006604	0.169014	4.094648
		Trash Cans	Laundry Detergent	0.006604	0.160000	4.094648
		Cheese	Toothpaste	0.005504	0.149254	2.008845
	0.002	Laundry Detergent	Trash Cans	0.006604	0.169014	4.094648
		Trash Cans	Laundry Detergent	0.006604	0.160000	4.094648
		Cheese	Toothpaste	0.005504	0.149254	2.008845

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

เนื่องจากการแข่งขันในอุตสาหกรรมร้านค้าปลีกที่มีความเข้มข้นมากขึ้น และการเพิ่มขึ้นของช่องทางการจัดจำหน่ายใหม่ๆ รวมถึงพฤติกรรมของลูกค้าที่มีความหลากหลายและมุ่งเน้นเฉพาะกลุ่มมากขึ้นในปัจจุบัน ทำให้ผู้วิจัยตระหนักว่า การทำความเข้าใจลักษณะของลูกค้าเป็นสิ่งสำคัญอย่างยิ่ง ซึ่งสามารถช่วยธุรกิจในการตัดสินใจที่สำคัญ ไม่ว่าจะเป็นการออกแบบแคมเปญการตลาดที่เหมาะสมกับแต่ละกลุ่มลูกค้า หรือการเลือกสินค้าที่จะส่งเสริมการขายร่วมกันได้อย่างมีประสิทธิภาพ

ผู้วิจัยจึงจัดทำงานวิจัยนี้ขึ้นมา เพื่อดำเนินการศึกษาอัลกอริทึมการจัดกลุ่มแต่ละประเภท และการวิเคราะห์กฎความสัมพันธ์ภายในกลุ่มลูกค้า โดยใช้ข้อมูลธุรกรรมจากร้านค้าปลีก ผู้วิจัยมีความคาดหวังว่า ผลการศึกษาที่ได้จากงานวิจัยนี้ จะช่วยให้ธุรกิจสามารถเลือกอัลกอริทึมที่เหมาะสมกับข้อมูลในการสร้างแบบจำลองการจัดกลุ่มลูกค้า และสามารถนำข้อมูลจากการจัดกลุ่มที่ได้ไปทำการวิเคราะห์ตระกร้าสินค้าเพื่อหากฎความสัมพันธ์ที่สำคัญภายในแต่ละกลุ่มได้ รวมถึงสามารถประยุกต์ใช้แบบจำลองนี้ ในการจัดกลุ่มลูกค้าในธุรกิจต่างๆ หรือข้อมูลลูกค้าใหม่ในอนาคตได้

ผู้วิจัยได้ทำการฝึกสอนแบบจำลองที่สร้างด้วยอัลกอริทึมประเภทต่างๆ ตลอดจนวัดประสิทธิภาพของแบบจำลอง รวมถึงวิเคราะห์ตระกร้าสินค้า เพื่อค้นหากฎความสัมพันธ์ที่มีภายในกลุ่มนั้นๆ ภายใต้ค่าพารามิเตอร์ที่กำหนด โดยจะทำการสรุปผลแบ่งตามหัวข้อ ดังนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

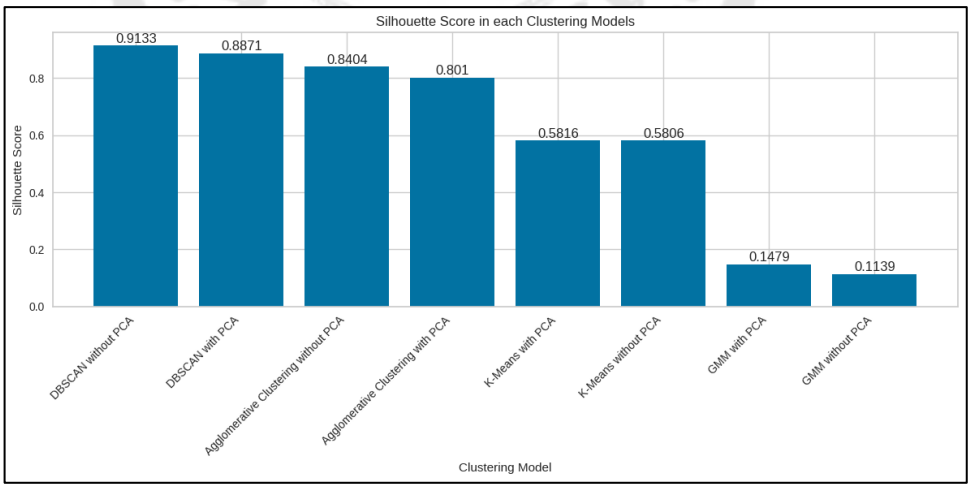
งานวิจัยนี้ศึกษาอัลกอริทึมการจัดกลุ่มข้อมูล และการวิเคราะห์ตระกร้าสินค้าเพื่อวิเคราะห์หากฎความสัมพันธ์ในลูกค้าแต่ละกลุ่ม โดยใช้ข้อมูลธุรกรรมร้านค้าปลีก และเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนหรือ Unsupervised Learning และทำการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมการจัดกลุ่มทั้งหมด 4 อัลกอริทึม ได้แก่ K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), และ Agglomerative Clustering ซึ่งข้อมูลที่จะนำไปศึกษาในแบบจำลองที่สร้าง

จากอัลกอริทึมการจัดกลุ่มแต่ละประเภท ผู้วิจัยได้ทำการจัดเตรียมข้อมูล โดยเริ่มจากการนำข้อมูลไปวิเคราะห์ผ่านเทคนิคการวิเคราะห์ RFM และนำผลลัพธ์ที่มาใช้เป็นคุณลักษณะในแบบจำลอง รวมถึงทำการแบ่งข้อมูลเป็น 2 รูปแบบ คือ ข้อมูลที่มีการลดจำนวนมิติด้วยเทคนิค Principal Component Analysis (PCA) และข้อมูลที่ไม่มีการลดจำนวนมิติ และทำการปรับค่าพารามิเตอร์ของแต่ละอัลกอริทึมในค่าที่แตกต่างกัน โดยจะนำข้อมูลที่ได้จากอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มที่ดีที่สุด 2 อันดับแรก ไปทำการศึกษาการวิเคราะห์ตระกร้าสินค้า โดยใช้อัลกอริทึม Frequent Pattern Growth (FP-Growth) ในการวิเคราะห์หากฎความสัมพันธ์ของลูกค้าแต่กลุ่มภายใต้ค่าพารามิเตอร์ที่ถูกกำหนดแตกต่างกัน

เมื่อทำการวัดประสิทธิภาพของอัลกอริทึมการจัดกลุ่มด้วยเมทริกซ์ Silhouette Score พบว่า อัลกอริทึม DBSCAN เป็นอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มสูงที่สุด ซึ่งมีค่า Silhouette Score เท่ากับ 0.9133 สำหรับข้อมูลที่ไม่มีการลดจำนวนมิติ และ 0.8871 สำหรับข้อมูลที่มีการลดจำนวนมิติ รองลงมา คือ อัลกอริทึม Agglomerative Clustering ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ มีค่าเท่ากับ 0.8404 และอัลกอริทึม Agglomerative Clustering ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ 0.8010 ถัดมา คือ อัลกอริทึม K-Means Clustering ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ มีค่าเท่ากับ 0.5816 และอัลกอริทึม K-Means Clustering ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ 0.5806 และสุดท้าย คือ อัลกอริทึม GMM ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ มีค่าเท่ากับ 0.1479 และอัลกอริทึม GMM ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ 0.1140 ดังแสดงข้อมูลในตารางที่ 35 และรูปภาพประกอบที่ 53

ตาราง 35 แสดงประสิทธิภาพของแต่ละอัลกอริทึมการจัดกลุ่ม

อัลกอริทึม	รูปแบบของข้อมูล	ค่าพารามิเตอร์	Silhouette Score	จำนวนคลัสเตอร์
DBSCAN	ไม่มีการลดจำนวนมิติ	min_samples = 10 Epsilon = 1	0.9133	2
DBSCAN	มีการลดจำนวนมิติ	min_samples = 10 Epsilon = 1	0.8871	3
Agglomerative Clustering	ไม่มีการลดจำนวนมิติ	n_clusters = 2	0.8404	2
Agglomerative Clustering	มีการลดจำนวนมิติ	n_clusters = 2	0.8010	2
K-Means Clustering	ไม่มีการลดจำนวนมิติ	k = 4	0.5806	4
K-Means Clustering	มีการลดจำนวนมิติ	K = 4	0.5816	4
GMM	ไม่มีการลดจำนวนมิติ	n_components = 3	0.1140	3
GMM	มีการลดจำนวนมิติ	n_components = 3	0.1479	3



ภาพประกอบ 53 แสดงประสิทธิภาพของแต่ละอัลกอริทึมการจัดกลุ่ม
โดยเรียงลำดับจากค่ามากไปหาค่าน้อย

จากการเปรียบเทียบประสิทธิภาพของแบบจำลอง พบว่า แบบจำลองที่สร้างจากอัลกอริทึม DBSCAN ด้วยข้อมูลที่ไม่มีการลดจำนวนมิติ และอัลกอริทึม DBSCAN ด้วยข้อมูลที่มีการลดจำนวนมิติ เป็นอัลกอริทึมที่มีประสิทธิภาพที่ดีที่สุด 2 อันดับแรก ซึ่งเป็นอัลกอริทึมที่ถูกเลือกเพื่อนำข้อมูลไปวิเคราะห์หากฎความสัมพันธ์ในลำดับถัดไป

โดยเมื่อวิเคราะห์กฎความสัมพันธ์ของลูกค้าแต่ละกลุ่มที่ได้จากอัลกอริทึม DBSCAN ที่ใช้ข้อมูลทั้งสองรูปแบบ ดังแสดงข้อมูลในตาราง 36 พบว่า จำนวนกฎความสัมพันธ์ที่ค้นพบมีความแตกต่างกันตามค่าพารามิเตอร์ min_support ที่กำหนด และเมื่อกำหนดค่า min_support = 0.01 พบว่า ไม่สามารถค้นพบกฎความสัมพันธ์ระหว่างสินค้าในกลุ่มลูกค้าที่ได้จากอัลกอริทึม DBSCAN ซึ่งใช้ข้อมูลทั้งสองรูปแบบได้ ซึ่งอาจเกิดจากความถี่ของการซื้อสินค้าร่วมกันในกลุ่มลูกค้าเหล่านี้อาจยังไม่เพียงพอต่อการสร้างกฎความสัมพันธ์ภายใต้เงื่อนไขดังกล่าว

สำหรับอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่ไม่มีการลดจำนวนมิติ คลัสเตอร์ -1 มีจำนวนกฎความสัมพันธ์สูงที่สุด โดยเมื่อกำหนด min_support = 0.001 พบกฎมากถึง 2,328 กฎ และลดลงเหลือ 268 กฎ เมื่อเพิ่ม min_support เป็น 0.002 สะท้อนให้เห็นว่า ถึงแม้ข้อมูลจุดนอกอาจไม่ถูกจัดกลุ่มอย่างชัดเจน แต่ยังคงมีความสัมพันธ์ของสินค้าที่ลูกค้าซื้อร่วมกัน

ในขณะที่ คลัสเตอร์ที่ 0 มีจำนวนกฎความสัมพันธ์น้อยลงอย่างมาก โดยพบ 252 กฎ ที่ min_support = 0.001 และลดลงเหลือเพียง 4 กฎ ที่ min_support = 0.002 แสดงให้เห็นว่าความสัมพันธ์ของสินค้าภายในกลุ่มนี้มีแนวโน้มกระจัดกระจายมากขึ้น เมื่อเพิ่มค่าความถี่ขั้นต่ำ

ทางด้านอัลกอริทึม DBSCAN ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ พบว่า คลัสเตอร์ที่ -1 มีจำนวนกฎความสัมพันธ์มากที่สุด โดยเมื่อกำหนด min_support = 0.001 พบกฎความสัมพันธ์มากถึง 2,382 กฎ และลดลงเหลือ 710 กฎ เมื่อเพิ่ม min_support เป็น 0.002 สะท้อนให้เห็นว่าแม้ข้อมูลบางส่วนอาจไม่ถูกจัดกลุ่มอย่างชัดเจน หรือไม่ได้อยู่ในคลัสเตอร์ แต่ยังคงมีความสัมพันธ์ระหว่างสินค้าที่ลูกค้าซื้อร่วมกันในระดับที่สูงพอสมควร

ในขณะที่ คลัสเตอร์ที่ 0 พบจำนวนกฎความสัมพันธ์ที่น้อยลงอย่างมาก โดยพบ 256 กฎ เมื่อกำหนด min_support = 0.001 และลดลงเหลือเพียง 4 กฎ ที่ min_support = 0.002 แสดงให้เห็นว่า ความสัมพันธ์ระหว่างสินค้าภายในกลุ่มนี้มีแนวโน้มกระจัดกระจายหรือไม่ชัดเจนมากขึ้นเมื่อมีการเพิ่มค่าของ min_support

สำหรับคลัสเตอร์ที่ 1 พบว่า มีจำนวนกฎความสัมพันธ์มากที่สุดเมื่อกำหนด min_support = 0.001 โดยมี 3,566 กฎ และลดลงเหลือ 1,032 กฎ เมื่อเพิ่ม min_support เป็น

0.002 ซึ่งแสดงให้เห็นว่า ความสัมพันธ์ระหว่างสินค้ามีค่าสูงในกลุ่มนี้ แม้จะมีการปรับค่า min_support เพิ่มขึ้น แต่ยังคงพบกฎความสัมพันธ์จำนวนมากและมีความสำคัญ

ตาราง 36 แสดงจำนวนกฎความสัมพันธ์ที่ได้จากการปรับค่า min_support

อัลกอริทึม	คลัสเตอร์	min_support	จำนวนกฎความสัมพันธ์
DBSCAN without PCA	-1	0.001	2,328
		0.002	268
	0	0.001	252
		0.002	4
DBSCAN with PCA	-1	0.001	2,382
		0.002	710
	0	0.001	256
		0.002	4
	1	0.001	3,566
		0.002	1,032

5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ศึกษาอัลกอริทึมการจัดกลุ่มข้อมูลทั้งหมด 4 อัลกอริทึม ได้แก่ K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), และ Agglomerative Clustering โดยใช้ข้อมูลธุรกรรมร้านค้าปลีก และเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Learning) จากนั้น นำผลลัพธ์ที่ได้จากการจัดกลุ่มมาวิเคราะห์ตะกร้าสินค้าเพื่อค้นหากฎความสัมพันธ์ที่น่าสนใจในแต่ละกลุ่มด้วยอัลกอริทึม FP-Growth

เริ่มต้นจากการนำข้อมูลมาวิเคราะห์ผ่านแบบจำลอง RFM (Recency, Frequency, Monetary) และจัดเตรียมข้อมูลใน 2 รูปแบบ คือ ข้อมูลที่มีการลดจำนวนมิติด้วยเทคนิค Principal Component Analysis (PCA) และข้อมูลที่ไม่มีการลดจำนวนมิติ เพื่อศึกษาผลกระทบของจำนวนมิติต่อคุณภาพการจัดกลุ่ม ผู้วิจัยทำการทดลองปรับค่าพารามิเตอร์หลักของอัลกอริทึมแต่ละตัว

ในช่วงค่าที่หลากหลาย เพื่อค้นหาค่าพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุด โดยประเมินประสิทธิภาพของการจัดกลุ่มด้วย Silhouette Score เมื่อคัดเลือกอัลกอริทึมการจัดกลุ่ม 2 อันดับแรกที่มีประสิทธิภาพสูงสุดแล้ว จึงนำผลลัพธ์กลุ่มลูกค้าเหล่านั้น ไปวิเคราะห์ด้วยอัลกอริทึม FP-Growth เพื่อค้นหากฎความสัมพันธ์ภายในแต่ละกลุ่ม ภายใต้เงื่อนไขของ min_support ที่แตกต่างกัน 3 ระดับ ได้แก่ 0.01, 0.002, และ 0.001, min_confidence เท่ากับ 0.01, และ min_lift มากกว่าหรือเท่ากับ 1

จากผลการศึกษาอัลกอริทึมการจัดกลุ่ม พบว่า อัลกอริทึม DBSCAN ให้ผลลัพธ์ที่ดีที่สุด และรองลงมา คือ อัลกอริทึม Agglomerative Clustering โดยพิจารณาจากค่า Silhouette Score ซึ่งสะท้อนถึงประสิทธิภาพในการจัดกลุ่มของอัลกอริทึม ขณะที่อัลกอริทึม K-Means Clustering และ GMM ให้ค่าที่ต่ำกว่า

อัลกอริทึม DBSCAN เป็นอัลกอริทึมที่มีประสิทธิภาพสูงสุดในการจัดกลุ่มข้อมูลลูกค้า โดยสามารถนำไปใช้ในการวิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้าได้อย่างมีประสิทธิภาพ อัลกอริทึม DBSCAN เหมาะกับการระบุความสัมพันธ์ของสินค้าระหว่างลูกค้าที่มีพฤติกรรมการซื้อแบบหลากหลายและไม่เป็นกลุ่มชัดเจน เนื่องจากสามารถระบุจุดที่เป็น Noise/Outlier ซึ่งจะสามารถจับกลุ่มลูกค้าที่มีลักษณะการซื้อที่กระจัดกระจายหรือมีความหลากหลายได้ดี

และจากผลการศึกษาจากการวิเคราะห์ตระกร้าสินค้าด้วยอัลกอริทึม FP-Growth โดยการใช้ข้อมูลการจัดกลุ่มของอัลกอริทึม DBSCAN ด้วยข้อมูลทั้งสองรูปแบบ พบว่า แม้ข้อมูลในคลัสเตอร์ที่เป็น Noise หรือ Outlier ซึ่งไม่ได้ถูกรวมอยู่ในกลุ่มหลัก แต่ยังคงสามารถค้นพบกฎความสัมพันธ์ที่สำคัญได้อย่างมากมาย ซึ่งแสดงให้เห็นถึง ความสามารถของอัลกอริทึม DBSCAN ในการค้นหาความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลที่ไม่ชัดเจนหรือกระจายตัว การค้นพบกฎความสัมพันธ์ในคลัสเตอร์ที่เป็น Noise/Outlier นี้ แสดงให้เห็นถึง ประสิทธิภาพของ DBSCAN ในการจัดกลุ่มข้อมูลที่มีพฤติกรรมการซื้อที่กระจัดกระจาย และไม่มีรูปแบบที่ชัดเจน ซึ่งอาจถูกมองข้ามในอัลกอริทึมการจัดกลุ่มอื่น ที่เน้นการจัดกลุ่มข้อมูลที่มีโครงสร้างหรือรูปแบบของข้อมูลที่ชัดเจน

ขณะที่อัลกอริทึม Agglomerative Clustering มีประสิทธิภาพในการจัดกลุ่มรองลงมาจากอัลกอริทึม DBSCAN โดยเหมาะสำหรับข้อมูลลูกค้าที่มีพฤติกรรมการซื้อที่แน่นอนและสามารถคาดการณ์ได้ ซึ่งสามารถสร้างกลุ่มที่มีความสัมพันธ์ภายในสูงและคาดการณ์พฤติกรรมได้แม่นยำ นอกจากนี้ยังเหมาะกับข้อมูลที่กระจายตัวอย่างต่อเนื่องหรือเป็นลำดับขั้น ซึ่งช่วยให้การจำแนกกลุ่มลูกค้าได้ดีในกรณีที่ข้อมูลมีโครงสร้างชัดเจน

ในส่วนของอัลกอริทึม K-Means Clustering และ GMM ถึงแม้จะให้ผลลัพธ์ที่มีค่า Silhouette Score ต่ำกว่าอัลกอริทึม DBSCAN และ Agglomerative Clustering แต่ทั้งสองอัลกอริทึมยังคงมีประโยชน์ในกรณีที่ข้อมูลมีโครงสร้างที่เป็นกลุ่มทรงกลมชัดเจน หรือสามารถจำแนกกลุ่มได้อย่างง่ายดาย โดยเฉพาะในกรณีที่ต้องการการคำนวณที่มีประสิทธิภาพสูงสำหรับชุดข้อมูลขนาดใหญ่ อัลกอริทึม K-Means Clustering เหมาะกับข้อมูลที่มีความแตกต่างชัดเจนในลักษณะการซื้อสินค้า เนื่องจากใช้การคำนวณตามระยะทางจากจุดศูนย์กลาง ขณะที่อัลกอริทึม GMM เหมาะกับข้อมูลที่มีการแจกแจงเป็นแบบปกติ (Gaussian Distribution) ซึ่งช่วยให้การจำแนกกลุ่มมีความยืดหยุ่นมากขึ้นในการจัดการกับข้อมูลที่มีการกระจายตัวที่ไม่สมมาตรหรือซับซ้อน

อย่างไรก็ตาม การประเมินประสิทธิภาพอัลกอริทึมจัดกลุ่มด้วย Silhouette Score เพียงตัวชี้วัดเดียว อาจยังไม่เพียงพอที่จะสะท้อนคุณภาพของอัลกอริทึมได้อย่างครบถ้วน ตามผลลัพธ์ที่ได้จากอัลกอริทึม DBSCAN ทั้งในกรณีที่ใช้ข้อมูลที่ไม่มีการลดมิติ และข้อมูลที่มีการลดมิติ พบว่าจำนวนคลัสเตอร์ที่ได้ยังไม่เหมาะสม โดยอัลกอริทึมที่ใช้ข้อมูลที่ไม่มีการลดมิติให้ผลลัพธ์เพียงคลัสเตอร์เดียว หากไม่รวม Outliers ขณะที่อัลกอริทึมที่ใช้ข้อมูลลดมิติให้ 2 คลัสเตอร์เท่านั้น ประกอบกับจุดข้อมูลที่เป็น Cluster Point ส่วนใหญ่กระจุกอยู่ในคลัสเตอร์ใดคลัสเตอร์หนึ่งอย่างหนาแน่น แสดงให้เห็นว่า อัลกอริทึมจัดกลุ่มไม่มีประสิทธิภาพอย่างแท้จริง จึงควรนำตัวชี้วัดอื่นมาประเมินร่วมด้วย รวมถึงการพิจารณาจำนวนคลัสเตอร์ และจำนวนจุดข้อมูลของแต่ละคลัสเตอร์ควบคู่กันไปในการวิเคราะห์ เพื่อให้ได้อัลกอริทึมการจัดกลุ่มที่สมบูรณ์ยิ่งขึ้น

สำหรับผลลัพธ์จากการลดจำนวนมิติของข้อมูล พบว่า ไม่มีความแตกต่างอย่างมีนัยสำคัญ เนื่องจากจำนวนคุณลักษณะก่อนและหลังการลดมิติมีความใกล้เคียงกัน ส่งผลให้ค่า Silhouette Score ที่ได้จากอัลกอริทึมในทั้งสองรูปแบบของข้อมูลมีค่าที่ไม่แตกต่างกันอย่างเด่นชัด

นอกจากนี้ การกำหนดค่า min_support ที่สูง และมุ่งเน้นการพิจารณารายการสินค้าที่มีค่าสูง ทำให้ผลลัพธ์ส่วนใหญ่ที่ได้ มีรายการสินค้าใน Antecedents และ Consequents เพียง 1 รายการ ซึ่งจากการศึกษา พบว่า มีเพียงคลัสเตอร์ที่ 1 ที่ใช้ข้อมูลที่มีการลดจำนวนมิติ เมื่อกำหนดค่า min_suport = 0.001 มีข้อมูลที่มีรายการสินค้ามากกว่า 1 จำนวน 1,068 จากจำนวนกฎความสัมพันธ์ที่มี 3,566 แสดงให้เห็นว่า การกำหนดค่าเงื่อนไขที่สูง อาจทำให้พลาดกฎที่มีประโยชน์เชิงธุรกิจในเชิงการจับคู่สินค้าหรือกลุ่มสินค้าที่ควรนำมาใช้ต่อยอดได้มากกว่านี้

5.3 ข้อเสนอแนะ

1. ในงานวิจัยนี้ มีการใช้เมทริกซ์ที่วัดประสิทธิภาพของอัลกอริทึมการจัดกลุ่มค่อนข้างจำกัด ซึ่งอาจทำให้การประเมินประสิทธิภาพของอัลกอริทึมขาดความหลากหลายในการวิเคราะห์ หากมีการใช้เมทริกซ์เพิ่มเติมเพื่อประเมินประสิทธิภาพของอัลกอริทึมร่วมด้วย อาจช่วยให้สามารถวิเคราะห์และระบุสาเหตุที่ทำให้อัลกอริทึมมีประสิทธิภาพที่ดีหรือแย่ได้อย่างชัดเจนยิ่งขึ้น

2. เนื่องจากข้อมูลที่ใช้ในการจัดกลุ่มของงานวิจัยมีรูปแบบการแจกแจงที่ไม่ปกติ คือ มีลักษณะเบ้ไปทางขวา จึงขอแนะนำให้ทดลองนำชุดข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) มาทำการจัดกลุ่มข้อมูล โดยใช้อัลกอริทึมที่นำเสนอในงานวิจัยนี้ พร้อมทั้งทำการปรับค่าพารามิเตอร์ของแต่ละอัลกอริทึมตามงานวิจัย พร้อมทั้งสังเกตผลลัพธ์ที่ได้จากการทดลองเพื่อประเมินว่า ประสิทธิภาพของอัลกอริทึมยังคงคงที่ตามที่คาดหวัง หรือมีการเปลี่ยนแปลงไปอย่างไร ซึ่งผลลัพธ์ที่ได้จะช่วยให้การเลือกใช้อัลกอริทึมในงานวิจัยหรือการประยุกต์ใช้งานจริงมีความแม่นยำและมีประสิทธิภาพมากยิ่งขึ้น

3. ในงานวิจัยนี้ได้เลือกใช้อัลกอริทึม FP-Growth สำหรับการวิเคราะห์ตะกร้าสินค้าและหากฎความสัมพันธ์เพียงอัลกอริทึมเดียว ซึ่งอาจจำกัดความหลากหลายในการวิเคราะห์และผลลัพธ์ที่ได้รับ หากมีการต่อยอดงานวิจัยด้วยการทดลองใช้อัลกอริทึมอื่นๆ เช่น Apriori หรือ Eclat เพื่อเปรียบเทียบผลลัพธ์ที่ได้ ทั้งในแง่ของคุณภาพของกฎความสัมพันธ์ที่ค้นพบ รวมถึงระยะเวลาที่ใช้ในการประมวลผล อาจช่วยให้เห็นความแตกต่างในประสิทธิภาพของแต่ละอัลกอริทึมได้ชัดเจนยิ่งขึ้น

4. เนื่องจากงานวิจัยนี้ดำเนินการวิเคราะห์ตะกร้าสินค้า โดยการจัดเรียงลำดับกฎความสัมพันธ์ตามค่า Support, Confidence และ Lift จึงส่งผลให้กฎความสัมพันธ์ที่ได้ในช่วงค่าที่สูง มักมีรายการสินค้าเพียงรายการเดียวทั้งในส่วนของ Antecedents และ Consequents ซึ่งหากมีการต่อยอดงานวิจัยโดยแยกการวิเคราะห์ตามจำนวนรายการสินค้าที่ปรากฏใน Antecedents และ Consequents อาจช่วยให้สามารถค้นพบกฎความสัมพันธ์ที่มีความน่าสนใจและมีความหลากหลายมากยิ่งขึ้น

5. การต่อยอดงานวิจัยโดยการใช้ Generative AI ในการประเมินอัลกอริทึมที่เหมาะสมกับลักษณะข้อมูลแต่ละประเภท จะช่วยเพิ่มความแม่นยำในผลลัพธ์ และการแนะนำสินค้าจากกฎความสัมพันธ์ที่ได้จากการจัดกลุ่มลูกค้า จะช่วยปรับปรุง Recommendation System ให้สามารถตอบสนองความต้องการของลูกค้าได้ตรงกลุ่มมากยิ่งขึ้น ซึ่งจะเสริมสร้างความพึงพอใจของลูกค้าและเพิ่มความสามารถในการแข่งขันของธุรกิจ

บรรณานุกรม

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules in Large Databases* Proceedings of the 20th International Conference on Very Large Data Bases,

AL, M. R., Sembiring, M. T., & Maulana, R. G. R. (2022). Product Recommendations Using Market Basket Analysis with FP-Growth and Clustering Techniques. *Proceedings of the First Australian International Conference on Industrial Engineering and Operations Management, Sydney, Australia, December 20-21, 2022*.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering*, *40*(1), 16-28.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. kdd,

Gayathri, K., & Arunodhaya, R. (2021). Customer segmentation and personalized marketing using k-means and apriori algorithm. Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India,

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann. https://books.google.co.th/books?id=pQws07tdpjoC

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, *29*(2), 1-12.

Hossain, M., Sattar, A. S., & Paul, M. K. (2019). Market basket analysis using apriori and FP growth algorithm. 2019 22nd international conference on computer and information technology (ICCIT),

Huges, A. (1994). Strategic Database Marketing: The Master Plan for Starting and Managing a Profitable, Customer-Based Marketing program. *Probus, Pub Co*.

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics*, *2*(4), 809-823.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241-254.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Lee, H. F., & Jiang, M. (2021). A hybrid machine learning approach for customer loyalty prediction. Neural Computing for Advanced Applications: Second International Conference, NCAA 2021, Guangzhou, China, August 27-30, 2021, Proceedings 2,

Lim, T. (2021). K-Means Clustering-Based Market Basket Analysis: UK Online E-Commerce Retailer. ICIT,

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics,

Mahfuza, R., Uddin, R. S., Rahman, Y., & Hai, M. A. (2021). A Comprehensive Framework for Superstore Business with Employing Effective Clustering Techniques. 2021 24th International Conference on Computer and Information Technology (ICCIT),

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. *Service Business*, *3*, 117-130.

Patil, P. (n.d.). *Retail transactions dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Phyo, T., & Uttama, S. (2023). Unveiling Patterns in the Night Market: A Machine Learning Approach to Customer Segmentation and Market Basket Analysis. 2023 7th International Conference on Information Technology (InCIT),

Pradana, M. R., Syafrullah, M., Irawan, H., Chandra, J. C., & Solichin, A. (2022). Market basket analysis using FP-growth algorithm on retail sales data. 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI),

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Singha, K., Parthanadee, P., Kessuvan, A., & Buddhakulsomsiri, J. (2024). Market Basket Analysis of a Health Food Store in Thailand: A Case Study. *International Journal of Knowledge and Systems Science (IJKSS)*, *15*(1), 1-14.

Sugiharto, N. D., Elbert, D., Arnold, J., Edbert, I. S., & Suhartono, D. (2023). Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm. 2023 6th International Conference on Information and Communications Technology (ICOIACT),

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267-276.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of data science*, *2*(2), 165-193.

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. KDD,

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

ภาคผนวก

ประวัติผู้เขียน