การจัดกลุ่มลูกค้า E-COMMERCE ด้วย CLUSTERING ALGORITHMS

CUSTOMER SEGMENTATION IN E-COMMERCE USING CLUSTERING ALGORITHMS

คฑาวุธ โพธิ์ศรี

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

การจัดกลุ่มลูกค้า E-COMMERCE ด้วย CLUSTERING ALGORITHMS

คฑาวุธ โพธิ์ศรี

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

CUSTOMER SEGMENTATION IN E-COMMERCE USING CLUSTERING ALGORITHMS

KATAWUT PHOSRI

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

การจัดกลุ่มลูกค้า E-COMMERCE ด้วย CLUSTERING ALGORITHMS

ของ

คฑาวุธ โพธิ์ศรี

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

.....

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... กรรมการ

(ดร.ศุภร คนธภักดี)

ชื่อเรื่อง	การจัดกลุ่มลูกค้า E-COMMERCE ด้วย CLUSTERING ALGORITHMS
ผู้วิจัย	คฑาวุธ โพธิ์ศรี
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการจัดกลุ่มลูกค้าในธุรกิจ E-Commerce โดยใช้อัลกอริธึม Clustering ได้แก่ K-Means และ BIRCH เพื่อวิเคราะห์พฤติกรรมของลูกค้าจากข้อมูลธุรกรรมในชุดข้อมูล Online Retail II ที่รวบรวมการสั่งซื้อสินค้าออนไลน์จากประเทศสหราชอาณาจักรระหว่างปี 2009–2011 จำนวน 1,067,371 รายการ โดยดำเนินการเปรียบเทียบการจัดกลุ่มผ่านฟีเจอร์ 2 รูปแบบ ได้แก่ RFM (Recency, Frequency, Monetary) และ CAD (Country, AVG Basket Price, Day Type) ผลการทดลองแสดงให้เห็นว่าการใช้ข้อมูล CAD ผ่านอัลกอริธึม BIRCH ให้ประสิทธิภาพสูงสุดในการจัดกลุ่มลูกค้า โดยได้ค่าตัวชี้วัด Silhouette Score เท่ากับ 0.9666, Calinski-Harabasz Index เท่ากับ 123,831.82 และ Davies-Bouldin Index เท่ากับ 0.1923 ซึ่งแสดงถึงความชัดเจนและความแน่นของแต่ละกลุ่มอย่างมีนัยสำคัญ ในขณะที่การใช้ RFM ผ่าน K-Means ให้ผลลัพธ์ดีเช่นกันในด้านการตีความทางธุรกิจ แต่มีประสิทธิภาพรองลงมา จากผลการวิจัย พบว่า การเลือกฟีเจอร์และอัลกอริธึมที่เหมาะสมมีผลต่อคุณภาพของการจัดกลุ่มลูกค้าอย่างมาก โดยเฉพาะ BIRCH ที่สามารถแบ่งกลุ่มลูกค้าได้โดยไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า และเหมาะกับข้อมูลที่มีโครงสร้างชัดเจน งานวิจัยนี้จึงมีประโยชน์ต่อการกำหนดกลยุทธ์การตลาด การจัดการสินค้าคงคลัง และการให้บริการที่ตรงกับความต้องการของลูกค้าในธุรกิจ E-Commerce

คำสำคัญ : การแบ่งกลุ่มลูกค้า, อีคอมเมิร์ซ, K-Means, BIRCH, การวิเคราะห์ RFM

Title	CUSTOMER SEGMENTATION IN E-COMMERCE USING CLUSTERING ALGORITHMS
Author	KATAWUT PHOSRI
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

This study aims to explore customer segmentation in the E-Commerce industry using clustering algorithms, specifically K-Means and BIRCH, to analyze customer behavior based on transaction data from the Online Retail II dataset. The dataset consists of 1,067,371 online purchase records from the United Kingdom during the period 2009–2011. Two feature sets were employed for comparison: RFM (Recency, Frequency, Monetary) and CAD (Country, Average Basket Price, Day Type). Experimental results indicate that using the CAD feature set with the BIRCH algorithm yields the highest clustering performance, with a Silhouette Score of 0.9666, a Calinski-Harabasz Index of 123,831.82, and a Davies-Bouldin Index of 0.1923—demonstrating significant cohesion and separation among the clusters. Meanwhile, the RFM feature set combined with K-Means provides superior interpretability from a business perspective. The findings highlight that the choice of features and clustering algorithm significantly influences the quality of customer segmentation. Notably, BIRCH offers advantages in handling well-structured data without requiring a predefined number of clusters. This research contributes valuable insights for strategic marketing, inventory management, and delivering personalized services in the E-Commerce sector.

Keyword : Customer Segmentation, E-Commerce, K-Means Clustering, BIRCH, RFM

กิตติกรรมประกาศ

รายงานสารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาและความเอาใจใส่อย่างยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษา ซึ่งได้ให้คำแนะนำอย่างใกล้ชิดทั้งในด้านวิชาการและการดำเนินการวิจัย ตลอดจนให้ข้อเสนอแนะ รวมถึงการสนับสนุนในทุกขั้นตอนของการทำงานวิจัยในครั้งนี้ ข้าพเจ้าขอกราบขอบพระคุณท่านมา ณ ที่นี้ด้วยความเคารพอย่างสูง

ขอขอบคุณต่อคณาจารย์ทุกท่านในสาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้ถ่ายทอดความรู้ ประสบการณ์ และแรงบันดาลใจตลอดระยะเวลาของการศึกษา รวมถึงคณะกรรมการสอบปากเปล่าที่ได้ให้คำแนะนำและข้อเสนอแนะที่เป็นประโยชน์อย่างยิ่งต่อการพัฒนางานวิจัยฉบับนี้ให้มีความสมบูรณ์ยิ่งขึ้น

ข้าพเจ้าขอขอบคุณครอบครัว เพื่อนร่วมรุ่น และทุกท่านที่มีส่วนสนับสนุน ให้ความช่วยเหลือ และให้กำลังใจอย่างต่อเนื่องในการศึกษาและการทำวิจัย หวังเป็นอย่างยิ่งว่างานวิจัยฉบับนี้จะเป็นประโยชน์ต่อผู้ที่เกี่ยวข้องและผู้สนใจในด้านการจัดกลุ่มลูกค้า E-Commerce ไม่มากก็น้อย หากมีข้อผิดพลาดประการใด ข้าพเจ้าขอน้อมรับไว้ด้วยความเคารพ และขออภัยมา ณ โอกาสนี้ด้วย

คฑาวุธ โพธิ์ศรี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ธุรกิจยอดนิยมในปัจจุบัน ที่สามารถสร้างรายได้และยอดขายจำนวนมาก นั้นคือการซื้อขายสินค้าทางออนไลน์ จนทำให้เกิดการเติบโตเพิ่มขึ้นอย่างรวดเร็วในอุตสาหกรรม E-Commerce ส่งผลให้การบริหารจัดการรวมถึงการหาข้อมูลเชิงลึก มีความสำคัญและมีความท้าทายอย่างมาก เพื่อที่จะสามารถตอบสนองความต้องการของลูกค้า หรือสร้างประสบการณ์การใช้บริการให้ยิ่งขึ้น การแบ่งกลุ่มลูกค้าเป็นกลยุทธ์ทางการตลาดที่ได้รับการยอมรับอย่างมาก การได้รับข้อมูลเชิงลึกเกี่ยวกับพฤติกรรมของผู้บริโภค และกระบวนการในการตัดสินใจนั้นมีความสำคัญอย่างยิ่งในการกำหนดกลยุทธ์ หรือนโยบายที่มีประสิทธิภาพ เพื่อสร้างบริการลูกค้าต้องการ และเพิ่มผลกำไรให้กับภาคธุรกิจ

การนำ Clustering Algorithms มาใช้ในการจัดการข้อมูล E-Commerce จะสามารถเพิ่มประสิทธิภาพในหลายๆ ด้าน เช่น การทำการตลาดให้ตรงกับกลุ่มเป้าหมายมากยิ่งขึ้น , การคาดการณ์พฤติกรรมของลูกค้า และการจัดการคลังสินค้าเพื่อรองรับความต้องการที่เปลี่ยนแปลงไปของลูกค้า การนำ Clustering Algorithms จะทำให้สามารถจัดกลุ่มของลูกค้า ตามลักษณะหรือ พฤติกรรมที่คล้ายคลึงกัน ทำให้ธุรกิจสามารถนำข้อมูลเชิงลึกที่ได้ ไปใช้ในการสร้างกลยุทธ์และการตัดสินใจได้อย่างมีประสิทธิภาพ

จากที่กล่าวมา ทำให้ผู้วิจัยมีความสนใจศึกษาขั้นตอนและวิธีการต่างๆที่เกี่ยวข้องกับการแบ่งกลุ่มตามพฤติกรรมของผู้บริโภคในการซื้อสินค้าผ่านช่องทางออนไลน์ โดยใช้ข้อมูล Online Retail II จากเว็บไซต์ Kaggle โดยข้อมูลนี้เกี่ยวกับธุรกรรมที่เกิดขึ้นจากการขายสินค้าออนไลน์ในสหราชอาณาจักร โดยเก็บข้อมูลตั้งแต่ 1 ธันวาคม 2009 จนถึง 12 กันยายน 2011 มีข้อมูลทั้งหมด 1,067,371 รายการ และ 8 คุณลักษณะ เพื่อที่นำวิธีการนี้ในการหาข้อมูลเชิงลึก และนำไปใช้ประโยชน์ในการพัฒนากลยุทธ์ทางการตลาดหรือ วางแผนการดำเนินการธุรกิจได้

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเพิ่มประสิทธิภาพในการจัดการข้อมูลและการวางแผนในธุรกิจ E-Commerce โครงการนี้จึงมีวัตถุประสงค์ในการนำ Clustering Algorithm มาใช้ในการวิเคราะห์และจัดกลุ่มข้อมูลในด้านต่าง ๆ ที่เกี่ยวข้องกับลูกค้าและการดำเนินงาน โดยโครงการมีวัตถุประสงค์ดังนี้:

1. ศึกษาวิธีการแบ่งกลุ่มลูกค้า (Customer Segmentation) โดยใช้ Clustering Algorithms เพื่อวิเคราะห์พฤติกรรมการซื้อสินค้าของลูกค้าบนช่องทางออนไลน์

2. เพื่อศึกษาการประเมินประสิทธิภาพของ Clustering Algorithms ในการจัดกลุ่มลูกค้าและวิเคราะห์ลักษณะหรือพฤติกรรมที่คล้ายคลึงกันของลูกค้า

3. เปรียบเทียบผลลัพธ์ของ Clustering Algorithms ที่หลากหลาย เช่น K-Means, BIRCH เพื่อระบุอัลกอริทึมที่เหมาะสมที่สุดสำหรับการจัดกลุ่มลูกค้าในบริบทของข้อมูล E-Commerce

1.3 ความสำคัญของการวิจัย

ปัจจุบันธุรกิจ E-Commerce เติบโตขึ้นอย่างรวดเร็ว เนื่องจากการซื้อขายสินค้าทางออนไลน์เป็นที่นิยมและสร้างรายได้มหาศาล การแข่งขันในอุตสาหกรรมนี้จึงเพิ่มขึ้นตามไปด้วย ทำให้การบริหารจัดการข้อมูลเชิงลึกมีบทบาทสำคัญต่อการดำเนินธุรกิจ การทำความเข้าใจลูกค้าในมุมมองเชิงพฤติกรรม และการพัฒนาแผนกลยุทธ์ที่มีประสิทธิภาพเพื่อตอบสนองต่อความต้องการของลูกค้าได้อย่างเหมาะสมกลายเป็นปัจจัยสำคัญที่ส่งผลต่อความสำเร็จในธุรกิจ

การวิจัยครั้งนี้ให้ความสำคัญกับการศึกษานำ Clustering Algorithms มาประยุกต์ใช้ในข้อมูลที่เกี่ยวข้องกับ E-Commerce เพื่อจัดกลุ่มลูกค้าตามลักษณะหรือพฤติกรรมที่คล้ายคลึงกัน ซึ่งจะช่วยเพิ่มประสิทธิภาพในหลายๆด้าน เช่น การทำการตลาดที่ตรงเป้าหมาย การวิเคราะห์พฤติกรรมผู้บริโภค และการจัดการคลังสินค้า โดยข้อมูลที่ได้จากการจัดกลุ่มลูกค้าเหล่านี้ สามารถนำไปใช้พัฒนากลยุทธ์และสนับสนุนการตัดสินใจเชิงธุรกิจได้อย่างชัดเจน นอกจากนี้ การวิจัยยังมีความสำคัญในมิติของการสนับสนุนการตัดสินใจโดยใช้ข้อมูล (Data-Driven Decision Making) ซึ่งช่วยให้ธุรกิจสามารถวางแผนการดำเนินงานและการลงทุนได้อย่างมีประสิทธิภาพ ลดความเสี่ยงและความสูญเปล่าของทรัพยากร

ในขณะเดียวกันยังช่วยเพิ่มความสามารถในการแข่งขันในตลาดได้อย่างยั่งยืน อีกทั้งงานวิจัยนี้ยังมุ่งหวังที่จะสร้างองค์ความรู้ใหม่ที่สามารถนำไปพัฒนาต่อยอดในด้านการวิเคราะห์ข้อมูลสำหรับอุตสาหกรรม E-Commerce โดยเฉพาะการใช้ Clustering Algorithms เช่น K-Means, และ BIRCH เพื่อค้นหาวิธีการที่เหมาะสมที่สุดในการจัดกลุ่มลูกค้า งานวิจัยนี้จึงไม่เพียงแต่จะศึกษาเกี่ยวกับการใช้ Clustering Algorithm ในการแบ่งกลุ่มทางผู้วิจัยยังคาดหวังว่าจะสามารถเพิ่มมูลค่าให้กับธุรกิจ รวมทั้งยังเป็นแนวทางสำหรับการพัฒนาเทคโนโลยีและนวัตกรรมในอนาคต

1.4 ขอบเขตของงานวิจัย

งานวิจัยนี้มุ่งเน้นการใช้ชุดข้อมูล Online Retail II จากเว็บไซต์ Kaggle ซึ่งรวบรวมข้อมูลธุรกรรมการขายสินค้าทางออนไลน์ในสหราชอาณาจักรระหว่างวันที่ 1 ธันวาคม 2009 ถึง 12 กันยายน 2011 โดยมีข้อมูลทั้งหมด 1,067,371 รายการและ 8 คุณลักษณะ โดยจะดำเนินการปรับเปลี่ยนข้อมูลให้เหมาะสมสำหรับการวิเคราะห์และการแบ่งกลุ่ม โดยการใช้เทคนิค การสร้างคุณลักษณะ (Feature Engineering) ผ่านแบบจำลอง RFM กำหนดให้วิเคราะห์พฤติกรรมลูกค้าผ่านการประยุกต์ใช้ Clustering Algorithms ได้แก่ K-Means และ BIRCH เพื่อจัดกลุ่มลูกค้าตามลักษณะและพฤติกรรมที่คล้ายคลึงกัน และประเมินประสิทธิภาพของการจัดกลุ่มที่ได้ ว่ามีความเหมาะสมเพียงใด โดยใช้ตัวชี้วัด Silhouette Coefficient , Davies-Bouldin Index และ Calinski-Harabasz Index จากนั้นทำการวิเคราะห์ผลลัพธ์ที่ได้ พร้อมสรุปและอภิปรายผล

1.5 สมมติฐานของการวิจัย

1. การใช้ Clustering Algorithms (เช่น K-Means และ BIRCH) สามารถจัดกลุ่มลูกค้าตามลักษณะหรือพฤติกรรมที่คล้ายคลึงกันได้อย่างมีประสิทธิภาพ

2. ตัวชี้วัดประสิทธิภาพ (เช่น Silhouette Coefficient, Davies-Bouldin Index และ Calinski-Harabasz Index) สามารถแสดงความแตกต่างของคุณภาพการจัดกลุ่มระหว่างอัลกอริทึม K-Means และ BIRCH ได้อย่างชัดเจน

4. แบบจำลอง RFM (Recency, Frequency, Monetary) ช่วยเพิ่มความแม่นยำและความเข้าใจในพฤติกรรมของลูกค้าสำหรับการจัดกลุ่มโดย Clustering Algorithms

5. Clustering Algorithms แบบต่าง ๆ (K-Means, BIRCH) มีความเหมาะสมแตกต่างกันสำหรับการวิเคราะห์พฤติกรรมลูกค้าในบริบทของข้อมูล Online Retail II

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.เปรียบเทียบประสิทธิภาพของ Clustering Algorithms (K-Means, BIRCH) เพื่อระบุวิธีการที่เหมาะสมที่สุดสำหรับการจัดกลุ่มลูกค้า

2.สนับสนุนการตัดสินใจเชิงกลยุทธ์ เช่น การปรับปรุงการให้บริการ การกำหนดโปรโมชั่น หรือการจัดการคลังสินค้าให้สอดคล้องกับความต้องการของลูกค้า

3. ได้รับความรู้และความเข้าใจเพิ่มมากขึ้นเกี่ยวกับการใช้ Clustering Algorithm

บทที่ 2
วรรณกรรม และงานวิจัยที่เกี่ยวข้อง

บทนี้นำเสนอการทบทวนวรรณกรรมที่เน้นการศึกษาและวิเคราะห์งานวิจัยเกี่ยวกับการพัฒนาโมเดลการเรียนรู้ของเครื่องสำหรับการแบ่งกลุ่ม โดยผู้วิจัยได้ศึกษาข้อมูลจากแหล่งต่าง ๆ อย่างละเอียดและสังเคราะห์เนื้อหาตามประเด็นสำคัญดังนี้

1. แนวคิดเกี่ยวกับการจัดการและ ธุรกิจ อี คอมเมิร์ซ
2. RFM Model
3. เทคนิค Exploratory Data Analysis (EDA)
4. ทฤษฎีที่เกี่ยวข้องกับอัลกอริทึมในการแบ่งกลุ่ม (Clustering Algorithms)
   3.1 Partitioning Methods: K-Means
   3.2 Hierarchical Methods: BIRCH
5. การวัดประสิทธิภาพแบบจำลอง (Model Performance Metrics)
6. งานวิจัยที่เกี่ยวข้องการใช้ Clustering Algorithms ในการแบ่งกลุ่ม

การศึกษาวรรณกรรมในครั้งนี้มุ่งวิเคราะห์และรวบรวมข้อมูลจากหลากหลายแหล่ง เช่น บทความวิชาการ รายงานวิจัย และกรณีศึกษาที่เกี่ยวข้องกับอุตสาหกรรมอีคอมเมิร์ซ โดยเน้นไปที่งานวิจัยที่เผยแพร่ในช่วง 10 ปีที่ผ่านมา เพื่อให้ได้มุมมองที่ทันสมัยและเข้ากับบริบททางเทคโนโลยีในปัจจุบัน ผลลัพธ์จากการศึกษาเหล่านี้จะช่วยชี้ให้เห็นช่องว่างของความรู้ที่ยังไม่ได้รับการสำรวจ และนำไปสู่การพัฒนาโมเดล Clustering Algorithms ที่สามารถเพิ่มประสิทธิภาพการจัดการในอีคอมเมิร์ซได้อย่างมีประสิทธิภาพในอนาคต

2.1 แนวคิดเกี่ยวกับการจัดการและ ธุรกิจ อี คอมเมิร์ซ

e-Commerce (Electronic Commerce) อีคอมเมิร์ซเป็นกิจกรรมทางธุรกิจที่แลกเปลี่ยนสินค้าและบริการ ผ่านเทคโนโลยีเครือข่ายสารสนเทศ ในฐานะของแบบจำลองเศรษฐกิจดิจิทัลทั่วไป อีคอมเมิร์ซทำลายระยะห่างของเวลาและพื้นที่ โดยผู้ซื้อและผู้ขายสามารถดำเนินกิจกรรมทางการค้า การค้าขาย และการเงินทุกประเภทได้โดยไม่ต้องพบกัน (Jiang & Qin, 2024; Wei et al., 2025)

ในประเทศไทยจะใช้คำว่า พาณิชย์อิเล็กทรอนิกส์ หมายถึง การทำธุรกรรมซื้อขายสินค้าและบริการผ่านอินเทอร์เน็ต โดยใช้เว็บไซต์หรือแอปพลิเคชันเป็นสื่อกลาง ช่วยให้ผู้ใช้งานสามารถเข้าถึงสินค้าและบริการจากทุกที่ทั่วโลกตลอด 24 ชั่วโมง e-Commerce เป็นส่วนหนึ่งของ ธุรกรรม

ทางอิเล็กทรอนิกส์ (Electronic Transaction) ที่ครอบคลุมกิจกรรมต่าง ๆ เช่น การซื้อขายออนไลน์ การสมัครสมาชิก การทำสัญญา การโอนเงินอัตโนมัติ และการสื่อสารข้อมูลผ่านระบบออนไลน์ เพื่อวัตถุประสงค์ทางธุรกิจและการติดต่องานราชการ (อาภาภรณ์ วัธนกุล, 2012)

2.2 RFM Model

RFM (Recency, Frequency, Monetary) เป็นเครื่องมือที่แพร่หลายที่สุดสำหรับการสำรวจพฤติกรรม โดยโมเดลนี้จะจัดกลุ่มลูกค้าตามความใหม่ ความถี่ และมูลค่าทางการเงิน แต่จะไม่เน้นที่การดึงดูดลูกค้าใหม่ แต่จะใช้เหมาะสมที่สุดในการระบุลูกค้าที่จะทำการตลาดแบบกำหนดเป้าหมาย(Rungruang et al., 2024; Ryu et al., 2025) โดยพิจารณาจาก 3 ปัจจัยสำคัญ ได้แก่

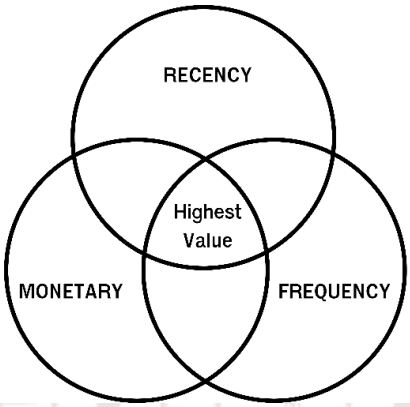
Recency: หมายถึง "ความใหม่" ของการทำธุรกรรมครั้งล่าสุดของลูกค้า โดยคำนวณจากจำนวนวันที่ผ่านมานับตั้งแต่วันที่ลูกค้าทำธุรกรรมครั้งล่าสุดจนถึงวันที่ปัจจุบัน ตัวชี้วัดนี้มีความสำคัญเพราะลูกค้าที่เพิ่งทำธุรกรรมเมื่อไม่นานมักแสดงถึงความสนใจและความสัมพันธ์ที่ยังคงมีกับธุรกิจ ยิ่งระยะเวลาน้อยเท่าใด ก็ยิ่งบ่งบอกถึงโอกาสที่ลูกค้าจะกลับมาซื้อซ้ำหรือใช้บริการอีกครั้ง ตัวอย่างเช่น หากลูกค้าทำการซื้อครั้งล่าสุดเมื่อ 7 วันที่ผ่านมา แสดงว่าลูกค้ามี Recency สูง ซึ่งต่างจากลูกค้าที่ไม่ได้ซื้อสินค้าหรือใช้บริการมาเป็นเวลาหลายเดือน

Frequency: หมายถึง "ความถี่" หรือจำนวนครั้งที่ลูกค้าทำธุรกรรมภายในช่วงเวลาที่กำหนด ตัวชี้วัดนี้ใช้วัดระดับความสัมพันธ์และความภักดีของลูกค้าที่มีต่อธุรกิจ ยิ่งลูกค้ามีความถี่ในการซื้อสินค้าหรือใช้บริการสูง ก็ยิ่งแสดงถึงความสัมพันธ์ที่ต่อเนื่องและมีแนวโน้มที่จะสร้างรายได้ให้ธุรกิจในระยะยาว ตัวอย่างเช่น หากลูกค้าทำการซื้อสินค้า 10 ครั้งในช่วง 6 เดือนที่ผ่านมา แสดงว่าลูกค้ารายนี้มี Frequency สูงและอาจเป็นกลุ่มลูกค้าที่ควรได้รับการดูแลเป็นพิเศษ ความถี่ในการทำธุรกรรม

Monetary: หมายถึง "มูลค่ารวม" ของการใช้จ่ายที่ลูกค้าได้ทำไว้กับธุรกิจในช่วงเวลาที่กำหนด ตัวชี้วัดนี้สะท้อนถึงความสำคัญของลูกค้าในแง่ของรายได้ที่ธุรกิจได้รับจากพวกเขา ลูกค้าที่มี Monetary สูงคือกลุ่มที่สร้างรายได้มาก และอาจเป็นกลุ่มที่ควรได้รับการดูแลหรือเสนอโอกาสพิเศษ ตัวอย่างเช่น หากลูกค้าใช้จ่ายรวม 50,000 บาทในช่วง 6 เดือนที่ผ่านมา แสดงว่าลูกค้ารายนี้มีมูลค่าสูงและมีความสำคัญต่อธุรกิจ

โมเดล RFM ช่วยให้บริษัทสามารถระบุกลุ่มลูกค้าที่แตกต่างกันและพัฒนากลยุทธ์ทางการตลาดเฉพาะกลุ่มได้อย่างมีประสิทธิภาพ ความเรียบง่ายและประสิทธิภาพของโมเดลนี้ช่วยให้

บริษัทสามารถจัดสรรทรัพยากรได้ดีขึ้น โดยมุ่งเน้นไปที่ลูกค้าที่มีมูลค่าสูง และส่งเสริมความภักดีของลูกค้า เพื่อเพิ่มความสามารถในการแบ่งกลุ่มและทำความเข้าใจลูกค้าให้ลึกซึ้งยิ่งขึ้น สามารถเสริมโมเดล RFM ด้วยมิติใหม่ ๆ หรือวิธีการอื่นที่เกี่ยวข้อง เช่น การใช้อัลกอริทึมการจัดกลุ่ม (Clustering Algorithms) เพื่อการวิเคราะห์ที่แม่นยำยิ่งขึ้น วิธีการนี้ได้รับความสนใจอย่างมากในด้านการสร้างกลยุทธ์ทางการตลาดเฉพาะเจาะจง (Anitha & Patil, 2022)



ภาพประกอบ 1 โมเดล RFM

2.3 เทคนิค Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) เป็นขั้นตอนสำคัญในกระบวนการ Data Science ที่ใช้เทคนิคการสร้างภาพข้อมูล (data visualization) และวิธีการเชิงปริมาณ (quantitative methods) เพื่อทำความเข้าใจชุดข้อมูล ก่อนนำไปวิเคราะห์หรือสร้างโมเดล(van Lingen et al., 2024) โดยมีจุดมุ่งหมายดังนี้

1.ตรวจสอบคุณภาพข้อมูล คือการค้นหาข้อผิดพลาด เช่น ค่าที่หายไป (missing values), ค่าผิดปกติ (outliers) และความไม่สมบูรณ์ของข้อมูล

2.เลือกแบบจำลองเบื้องต้น คือการใช้ข้อมูลจาก EDA เพื่อช่วยตัดสินใจเลือกโมเดลที่เหมาะสมกับลักษณะของข้อมูล

3.วิเคราะห์ความสัมพันธ์ระหว่างตัวแปร คือการตรวจสอบทิศทางและความสัมพันธ์เชิงลึกระหว่างตัวแปร เช่น การวิเคราะห์ Correlation

4.สรุปและทำความเข้าใจชุดข้อมูล คือการช่วยให้ได้มุมมองที่ชัดเจนเกี่ยวกับลักษณะสำคัญของข้อมูล เช่น การกระจายตัว (distribution), แนวโน้ม (trends) และการแบ่งกลุ่มข้อมูล (clustering)

2.4 ทฤษฎีที่เกี่ยวข้องกับอัลกอริทึมในการแบ่งกลุ่ม (Clustering Algorithms)

Data Clustering เป็นการจัดกลุ่มข้อมูลเป็นหนึ่งในเทคนิคการวิเคราะห์ข้อมูลที่สำคัญและได้รับความนิยมในงานด้านการทำเหมืองข้อมูล (data mining) และการเรียนรู้ของเครื่อง (machine learning) เนื่องจากการทำ labeling ข้อมูลด้วยมนุษย์มีค่าใช้จ่ายสูงและใช้เวลามาก การจัดกลุ่มข้อมูลจึงช่วยแก้ปัญหาโดยการจัดตัวอย่างข้อมูลที่ไม่มี label ให้อยู่ในกลุ่มเดียวกันตามความคล้ายคลึงกัน และแยกตัวอย่างที่แตกต่างกันออกเป็นกลุ่มอื่น

การจัดกลุ่มข้อมูลมีประโยชน์ทั้งในงานสำรวจข้อมูลเพื่อค้นหารูปแบบที่ซ่อนอยู่ และในขั้นตอนการเตรียมข้อมูลเพื่อสนับสนุนงานวิเคราะห์อื่น ๆ เช่น การแบ่งส่วนภาพ (image segmentation), การจดจำรูปแบบ (pattern recognition), และการวิเคราะห์เครือข่าย (network analysis)

วิธีการจัดกลุ่มข้อมูลสามารถแบ่งออกเป็น 3 ประเภทหลัก:

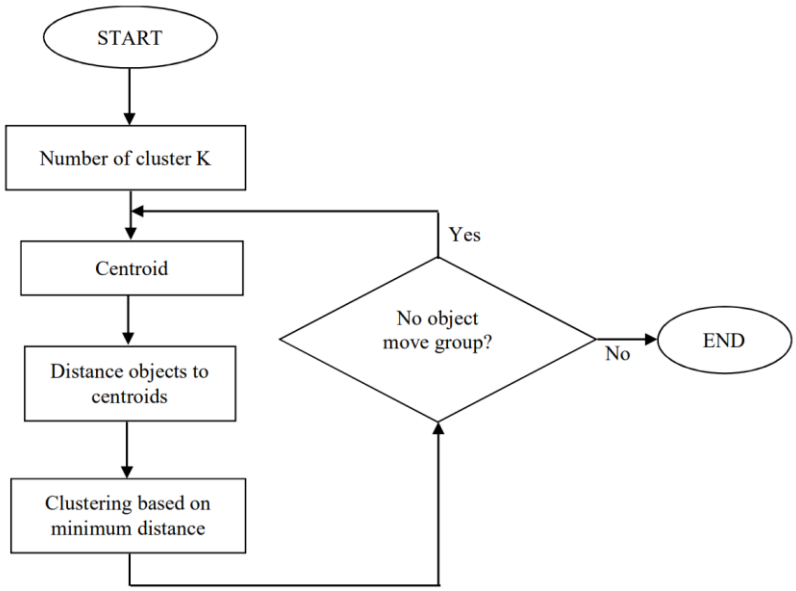
1. Partitioning Methods: เป็นเทคนิคที่ใช้แบ่งกลุ่มออกตามจำนวนกลุ่มที่กำหนดไว้ล่วงหน้า โดยอาศัยการกำหนดจุดศูนย์กลาง (Centroid) หรือจุดอ้างอิง (Medoid) เหมาะกับกลุ่มข้อมูลที่ชัดเจนไม่ทับซ้อนกันโดยมีอัลกอริทึมที่นิยมใช้คือ K-Means (Selim & Ismail, 1984)

K-Means Clustering

การจัดกลุ่มด้วย K-Means เป็นอัลกอริทึมที่ได้รับความนิยมมากที่สุด โดยจะเริ่มจากการกำหนดจำนวนกลุ่ม K กลุ่มและกำหนดจุดศูนย์กลาง (Centroid) จำนวน K จุดให้เป็น Centroid เริ่มต้น จากนั้นจะนำข้อมูลจัดเป็นกลุ่มเข้ากับจุดศูนย์กลางที่ใกล้ที่สุด โดยอ้างอิงจากจุดศูนย์กลางที่เลือก ของแต่ละกลุ่ม เมื่อสร้างกลุ่มแล้วจุดศูนย์กลางของแต่ละกลุ่มจะเปลี่ยนแปลง จากนั้นอัลกอริทึมจะทำขั้นตอนดังกล่าวซ้ำๆ จนกว่าจุดศูนย์กลางของแต่ละกลุ่มจะไม่เปลี่ยนแปลง(Selim & Ismail, 1984)

K-means clustering เป็น greedy algorithm ที่รับประกันว่าจะหาค่าผลลัพธ์ไปสู่จุดต่ำสุดในระดับเฉพาะที่ (local minimum) แต่การหาค่าต่ำสุดของ score function ของ K-means นั้นเป็นปัญหาที่ทราบกันว่ามีความซับซ้อนระดับ NP-Hard (Non-deterministic Polynomial-time Hard เป็นคำที่ใช้อธิบายความซับซ้อนของปัญหาทางคอมพิวเตอร์ในทางทฤษฎี (complexity theory) โดย NP-Hard สำหรับ K-Means Clustering คือการหาศูนย์กลางกลุ่มที่เหมาะสมที่สุดและจะไม่มีวิธีแก้ที่มีประสิทธิภาพสำหรับทุกกรณีในเวลา Polynomial-Time เมื่อจำนวนกลุ่มหรือจำนวนข้อมูลเพิ่มขึ้น) โดยปกติแล้วเงื่อนไขการลู่เข้าสู่จุดศูนย์กลางจะสามารถยอมรับได้ในทาง

ปฏิบัติคือ การทำซ้ำจะหยุดเมื่อข้อมูลเหล่านั้นเปลี่ยนแปลงกลุ่มไม่เกิน 1 % จากกลุ่มเดิม (Kodinariya & Makwana, 2013)



ภาพประกอบ 2 ผังการทำงานของอัลกอริธึม K-Means Clustering

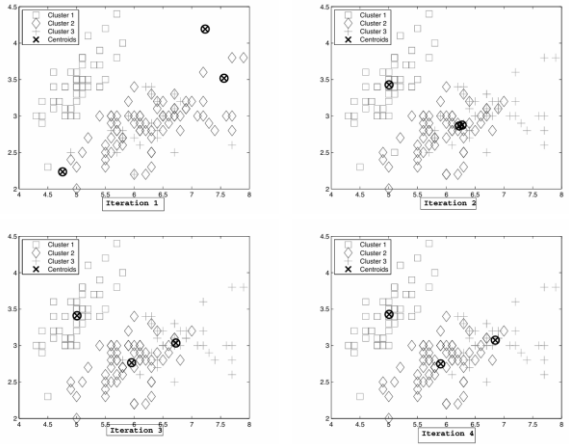
รูปแบบการวัดระยะห่างแบบต่างๆที่นำมาใช้ได้คือ ระยะแมนฮัตตัน (Manhattan distance หรือ L_1 Norm), ระยะยูคลิเดียน (Euclidean distance หรือ L_2 Norm), ความคล้ายคลึงของโคไซน์ (Cosine similarity) สำหรับการจัดกลุ่มแบบ K-means ระยะยูคลิเดียนเป็นตัวเลือกที่นิยมใช้มากที่สุด จากที่กล่าวมาข้างต้น แสดงให้เห็นว่า เราสามารถได้ผลลัพธ์ของการจัดกลุ่มที่แตกต่างกันตามค่า K และรูปแบบการวัดระยะที่แตกต่างกัน

ฟังก์ชันวัตถุประสงค์ที่ถูกใช้ในอัลกอริทึม K-means เรียกว่า ผลรวมของความคลาดเคลื่อนยกกำลังสอง (Sum of Squared Errors - SSE) หรือ ผลรวมของเศษเหลือยกกำลังสอง (Residual Sum of Squares - RSS) โดยมีสูตรทางคณิตศาสตร์ดังนี้

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - C_k\|^2 \qquad (1)$$

$$C_k = \frac{x_i \in C_k}{|C_k|} \qquad (2)$$

กำหนดให้ชุดข้อมูล $D = \{x_1, x_2, \ldots, x_n,\}$ ประกอบด้วย N จุด โดยกำหนดให้สัญลักษณ์ $C = \{C_1, C_2, \ldots, C_k, \ldots, C_K,\}$ แทนการจัดกลุ่มที่ได้หลังจากใช้ K-means clustering ค่าผลรวมของความคลาดเคลื่อนยกกำลังสอง SSE สำหรับการจัดกลุ่มนี้ จะนิยามไว้ในสมการ (1.1) โดยที่ c_k คือจุดศูนย์กลางของกลุ่ม c_k คือเป้าหมายการหาการดกลุ่มที่ลดค่า SSE ให้ต่ำที่สุด ขั้นตอนการกำหนดกลุ่มและการแก้ไขแบบวนซ้ำในอัลกอริทึม K-means มีเป้าหมายเพื่อลดค่า SSE ให้ต่ำที่สุดสำหรับจุดศูนย์กลางที่กำหนดไว้ (Lu et al., 2025)



ภาพประกอบ 3 การทำงานของ K-Means บนชุดข้อมูล Fisher Iris

ภาพประกอบ 3 แสดงขั้นตอนการทำงานของอัลกอริทึม K-means บนชุดข้อมูล Fisher Iris โดยจะเริ่มจากการสุ่มจุดศูนย์กลางสามจุดใน Iteration 1 ซึ่งข้อมูลแต่ละจุดจะถูกจัดให้อยู่ในกลุ่มที่ใกล้จุดศูนย์กลางที่สุด จากนั้นใน Iteration 2, 3 จะเห็นจุดศูนย์กลางใหม่ที่ถูกคำนวณจากค่าเฉลี่ยของข้อมูลในแต่ละกลุ่มจนไม่มีการเปลี่ยนแปลงใน Iteration 4 (Jain & K., 2013)

การกำหนดจำนวนกลุ่ม และการเลือกจุดเริ่มต้น

MacQueen ได้เสนอวิธีการกำหนดค่าเริ่มต้นแบบง่าย ๆ โดยการเลือกจุดเริ่มต้น K จุดแบบสุ่ม (MacQueen, 1967) วิธีนี้เป็นวิธีที่ง่ายที่สุดและได้รับการใช้งานอย่างแพร่หลาย ซึ่งทำให้เห็นว่าปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพของอัลกอริทึม K-Mean คือการเลือกจุดศูนย์กลางเริ่มต้น และ การกำหนดจำนวนกลุ่ม K จึงทำให้มีวิธีการในการกำหนดจำนวนกลุ่ม K เริ่มต้นสำหรับ K-Means ที่ได้รับความนิยมและ ประสบความสำเร็จในการปรับปรุงประสิทธิภาพของการจัดกลุ่ม คือการใช้ Elbow Methods, Silhouette Coefficient, K-Means++

Elbow Methods คือวิธีที่ใช้ในการกำหนดจำนวนกลุ่ม (K) ที่เหมาะสมโดยจะวัดความคลาดเคลื่อนยกกำลังสอง (Sum of Squared Errors - SSE) เทียบกับจำนวนกลุ่ม (K) และนำค่าที่ได้มากำหนดจุดบนกราฟ เมื่อได้ค่า SSE ของแต่ละจำนวนกลุ่ม (K) แล้วจะลากเส้นเชื่อมกันระหว่างจุด และจำนวนกลุ่มที่เหมาะสมนั้นคือ จุดที่จำนวนกลุ่ม (K) ในกราฟเกิดมุมที่ชัดเจนซึ่งมีลักษณะคล้าย ข้อศอก (Elbow Criterion)

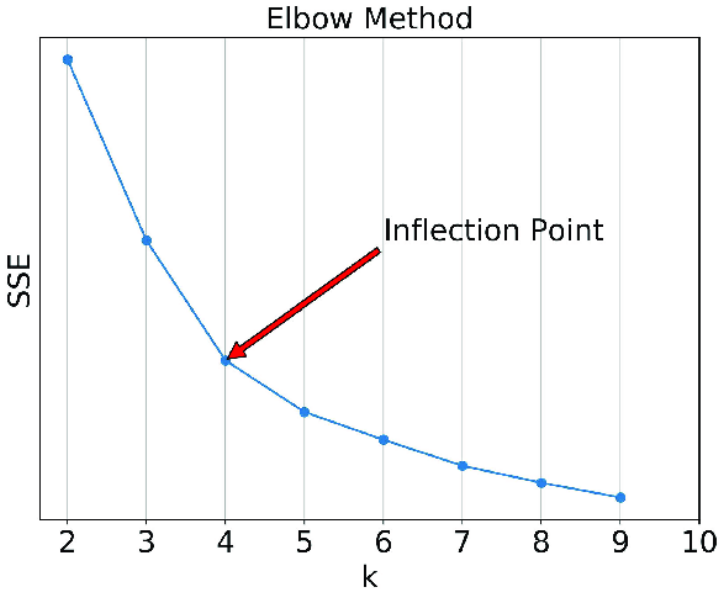
แนวคิดคือ จะเริ่มด้วย K=2 และจะเพิ่มค่าของ K ขึ้นทีละ 1 ในแต่ละขั้น และจะถูกนำไปคำนวนที่ค่าหนึ่งของ K ค่าที่คำนวณออกมาได้จะลดลงอย่างรวดเร็ว และจะเข้าสู่ช่วงที่คงที่เมื่อเราเพิ่มจำนวน K ต่อไปโดยจุดที่ K ลดลงมามากที่สุดนั้นคือ จำนวนกลุ่มที่ต้องการ เหตุผลก็เพราะว่าการเพิ่มจำนวนกลุ่ม K หลังจากนี้กลุ่ม ใหม่จะมีความใกล้เคียงกับกลุ่มเดิมที่มีอยู่แล้ว ดังภาพประกอบ 4 โดยมีสมการดังนี้

$$SSE = \sum_{i=1}^{n} (y_i - y')^2 \qquad (3)$$

N = จำนวนอินสแตนซ์

y_i = ค่าของอินสแตนซ์ที่ i^{th}

y' = ค่าเฉลี่ยของอินสแตนซ์



ภาพประกอบ 4 แสดงจุดที่ K ลดลงมามากที่สุดทำให้เกิดการหักศอกใน Elbow Method

2. Hierarchical Methods

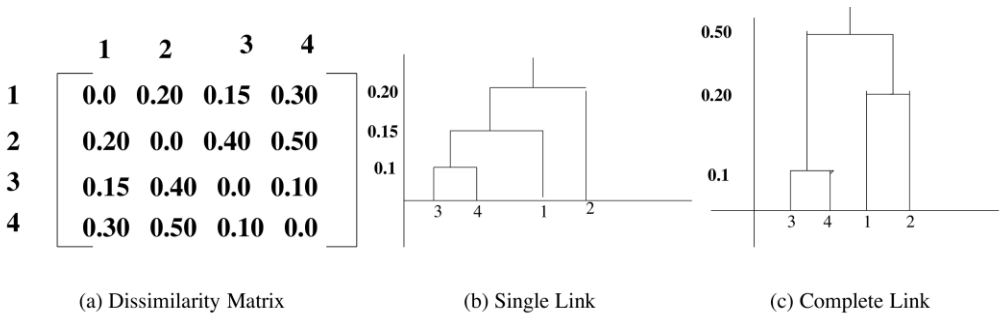
การจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering) (Jain et al., 1999) เป็นวิธีการที่ถูกพัฒนาขึ้นเพื่อแก้ปัญหาข้อจำกัดของวิธีการจัดกลุ่มแบบแบ่งส่วน (Partitional Clustering) ที่มักต้องกำหนดจำนวนกลุ่ม (K) ล่วงหน้าและมีลักษณะไม่แน่นอนในธรรมชาติ อัลกอริทึมแบบลำดับขั้นให้กลไกการจัดกลุ่มที่ยืดหยุ่นและมีความแน่นอนมากขึ้น โดยแบ่งออกเป็นสองประเภทหลัก ได้แก่

การจัดกลุ่มแบบรวบรวม (Agglomerative Clustering)

วิธีการนี้เริ่มจากการสร้างกลุ่มเล็กที่สุด (Singleton Clusters) ที่ประกอบด้วยข้อมูลเพียงตัวเดียว จากนั้นจะรวมกลุ่มสองกลุ่มเข้าด้วยกันครั้งละคู่จนเกิดโครงสร้างแบบลำดับขั้นจากล่างขึ้นบน (Bottom-Up Hierarchy) ขั้นตอนพื้นฐานที่เกี่ยวข้องในอัลกอริทึมการจัดกลุ่มแบบลำดับขั้นเชิงรวบรวมมีดังนี้ ก่อนอื่น ต้องใช้ตัววัดความใกล้เคียงเฉพาะ (Proximity Measure) เพื่อสร้างเมทริกซ์ความไม่คล้ายคลึง (Dissimilarity Matrix) และแสดงจุดข้อมูลทั้งหมดในระดับล่างสุดของเดนโดรแกรม (Dendrogram) จากนั้น กลุ่มที่ใกล้เคียงกันที่สุดจะถูกรวมกันในแต่ละระดับ และอัปเดตเมทริกซ์ความไม่คล้ายคลึงให้สอดคล้องกัน กระบวนการรวมกลุ่มเชิงรวบรวมนี้ดำเนินต่อไปจนกว่าจะได้กลุ่มใหญ่ที่สุดที่รวมข้อมูลทั้งหมดไว้ในกลุ่มเดียว ซึ่งจะเป็นจุดยอดของเดนโดรแกรม และถือเป็นการสิ้นสุดกระบวนการรวมกลุ่ม

วิธีการจัดกลุ่มแบบรวบรวมที่ได้รับความนิยมมากที่สุดคือการจัดกลุ่มแบบ Single Link และ Complete Link ในการจัดกลุ่มแบบ Single Link ความคล้ายคลึงระหว่างสองกลุ่มจะถูกกำหนดโดยความคล้ายคลึงระหว่างสมาชิกที่ใกล้กันที่สุด (Nearest Neighbor) ของแต่ละกลุ่ม วิธีการนี้ให้ความสำคัญกับบริเวณที่กลุ่มมีความใกล้ชิดกันมากที่สุด โดยละเลยโครงสร้างโดยรวมของกลุ่ม ดังนั้น วิธีนี้จึงถูกจัดให้อยู่ในประเภทของวิธีการจัดกลุ่มที่อิงตามความคล้ายคลึงในระดับท้องถิ่น (Local Similarity-Based Clustering Method) Single Link สามารถจัดกลุ่มข้อมูลที่มีรูปร่างยาวหรือไม่เป็นวงรีได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม ข้อเสียสำคัญของวิธีนี้คือมีความไวต่อสัญญาณรบกวน (Noise) และจุดข้อมูลที่อยู่นอกกลุ่ม (Outliers) ในข้อมูล วิธีการนี้ยังมีความไวต่อจุดข้อมูลที่อยู่นอกกลุ่ม (Outliers) ทั้ง Single Link และ Complete Link มีการตีความในลักษณะกราฟเชิงทฤษฎี (Graph-Theoretic Interpretations) โดยกลุ่มที่ได้จากการจัดกลุ่มแบบ Single Link จะสอดคล้องกับส่วนประกอบที่เชื่อมต่อกัน (Connected Components) ของกราฟ

และกลุ่มที่ได้จากการจัดกลุ่มแบบ Complete Link จะสอดคล้องกับคลิกสูงสุด (Maximal Cliques) ของกราฟ



(a) Dissimilarity Matrix (b) Single Link (c) Complete Link

ภาพประกอบ 5 ภาพประกอบของ Agglomerative Clustering

ภาพประกอบที่ 5 แสดง การแสดงภาพการจัดกลุ่มแบบลำดับขั้น (Agglomerative Clustering)

(a) เมทริกซ์ความไม่คล้ายคลึงที่คำนวณจากจุดข้อมูลสี่จุดแบบสุ่ม

(b) เดนโดรแกรมที่ได้จากวิธีการจัดกลุ่มแบบลำดับขั้น Single Link

(c) เดนโดรแกรมที่ได้จากวิธีการจัดกลุ่มแบบลำดับขั้น Complete Link

แสดงเมทริกซ์ความไม่คล้ายคลึงและเดนโดรแกรมสองแบบที่ได้จากการใช้อัลกอริทึม Single Link และ Complete Link กับชุดข้อมูลตัวอย่าง ในเดนโดรแกรม แกน X แสดงถึงวัตถุข้อมูล และแกน Y แสดงถึงค่าความไม่คล้ายคลึง (ระยะทาง) ที่ข้อมูลถูกรวมกัน ความแตกต่างในการรวมกลุ่มระหว่างเดนโดรแกรมทั้งสองเกิดจากเกณฑ์ที่แตกต่างกันระหว่างอัลกอริทึม Single Link และ Complete Link ใน Single Link จุดข้อมูล 3 และ 4 ถูกรวมกันที่ระยะทาง 0.1 ตามที่แสดงใน (b) จากนั้นตามการคำนวณในสมการในภาพประกอบที่ (6) กลุ่ม (3, 4) จะถูกรวมกับจุดข้อมูล 1 ในระดับถัดไป และในระดับสุดท้ายกลุ่ม (3, 4, 1) จะถูกรวมกับจุดข้อมูล 2 ใน Complete Link การรวมกลุ่มสำหรับ (3, 4) จะถูกตรวจสอบกับจุดข้อมูล 1 และ 2 และเนื่องจาก d(1,2)=0.20 d(1,2)=0.20 จุดข้อมูล 1 และ 2 จะถูกรวมกันในระดับถัดไป สุดท้าย กลุ่ม (3, 4) และ (1, 2) จะถูกรวมกันในระดับสุดท้าย สิ่งนี้อธิบายถึงความแตกต่างในกระบวนการจัดกลุ่มของทั้งสองกรณี (Jain, 2010)

$$
\begin{aligned}
d_{min}((3,4),1) &= min(d(3,1),d(4,1)) = 0.15 \\
d_{min}((3,4,1),2) &= min(d(3,2),d(4,2),d(1,2)) = 0.20 \\
d_{max}((3,4),1) &= max(d(3,1),d(4,1)) = 0.30 \\
d_{max}((3,4),2) &= max(d(3,2),d(4,2)) = 0.50 \\
d_{max}((3,4),(1,2)) &= max(d(3,1),d(3,2),d(4,1),d(4,2)) = 0.50
\end{aligned}
$$

ภาพประกอบ 6 สมการการวัดระยะทางระหว่างกลุ่มข้อมูล (clusters)

การจัดกลุ่มแบบแบ่งแยก (Divisive Clustering)

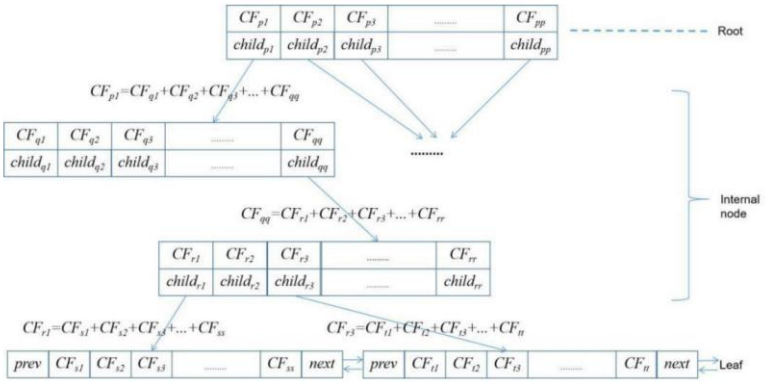
เป็นกระบวนการแบบจากบนลงล่าง (Top-Down Approach) โดยเริ่มต้นที่ราก (Root) ซึ่งมีจุดข้อมูลทั้งหมด และทำการแบ่งกลุ่มซ้ำๆ เพื่อสร้างเดนโดรแกรม วิธีนี้มีข้อได้เปรียบด้านประสิทธิภาพมากกว่าการจัดกลุ่มแบบรวบรวม (Agglomerative Clustering) โดยเฉพาะในกรณีที่ไม่จำเป็นต้องสร้างลำดับขั้นทั้งหมดจนถึงระดับจุดข้อมูลเดี่ยว (Individual Leaves) นอกจากนี้ยังถือเป็นกระบวนการแบบองค์รวม (Global Approach) เนื่องจากมีข้อมูลทั้งหมดครบถ้วนก่อนที่จะทำการแบ่งกลุ่ม

ประเด็นในกระบวนการจัดกลุ่มแบบแบ่งแยก (Issues in Divisive Clustering)มีปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพหลายประการ เริ่มต้นด้วย เกณฑ์การแบ่งกลุ่ม (Splitting Criterion) ซึ่งมักใช้เกณฑ์ค่า SSE (Sum of Squares Error) ของ Ward's K-means เพื่อพิจารณาความเหมาะสมของการแบ่ง โดยการลดค่า SSE ที่มากจะแสดงถึงคุณภาพของการแบ่งกลุ่ม อย่างไรก็ตาม เนื่องจาก SSE ใช้ได้กับข้อมูลเชิงตัวเลขเท่านั้น จึงสามารถใช้ดัชนี Gini เพื่อจัดการข้อมูลประเภทตัวแปรเชิงระบุได้ วิธีการแบ่งกลุ่ม (Splitting Method) ก็มีบทบาทสำคัญ วิธี Bisecting K-means (กำหนด K=2) มักใช้ในการแบ่งกลุ่ม เนื่องจากลดเวลาในการคำนวณเกณฑ์ของ Ward ได้และช่วยให้ได้ผลลัพธ์ที่มีคุณภาพสูงขึ้น สำหรับ การเลือกกลุ่มที่จะแบ่งต่อ (Choosing the Cluster to Split) แม้จะไม่สำคัญเท่ากับสองปัจจัยแรก แต่การเลือกกลุ่มที่เหมาะสมสำหรับการแบ่งต่อสามารถช่วยสร้างเดนโดรแกรมที่กะทัดรัดได้ วิธีที่ง่ายที่สุดคือการตรวจสอบค่า Square Errors ของกลุ่มและเลือกกลุ่มที่มีค่าสูงสุด สุดท้าย การจัดการกับสัญญาณรบกวน (Handling Noise) เป็นสิ่งสำคัญ เนื่องจากจุดข้อมูลที่มีสัญญาณรบกวนอาจสร้างกลุ่มที่ผิดปกติได้ การใช้เกณฑ์หยุด (Threshold) สามารถช่วยระบุว่าเมื่อใดควรหยุดการแบ่งกลุ่มเพิ่มเติม เพื่อป้องกันผลกระทบที่เกิดจากสัญญาณรบกวนในข้อมูล.

ลำดับขั้นของกลุ่มสามารถมองได้ในลักษณะต้นไม้แบบไบนารี (Binary Tree) โดยจุดเริ่มต้นที่ราก (Root) แทนข้อมูลทั้งหมดที่ต้องการจัดกลุ่ม ซึ่งอยู่ที่ระดับบนสุดของลำดับขั้น (Level 0) ในแต่ละระดับ กลุ่มย่อยหรือโหนดลูกจะเป็นชุดย่อยของข้อมูล ขณะที่ฐานของลำดับขั้นจะเป็นจุดข้อมูลเดี่ยว (Singleton Points) ซึ่งเป็นใบไม้ (Leaves) ของต้นไม้ โครงสร้างลำดับขั้นนี้เรียกว่า Dendrogram ข้อดีที่สำคัญคือความสามารถในการ "ตัด" โครงสร้างลำดับขั้นในระดับที่ต้องการ เพื่อให้ได้จำนวนกลุ่มที่เหมาะสมโดยไม่ต้องกำหนดจำนวน K ล่วงหน้า การจัดกลุ่มแบบลำดับขั้นเหมาะสำหรับการวิเคราะห์ข้อมูลที่ต้องการความยืดหยุ่นและการแสดงผลในลักษณะโครงสร้างลำดับขั้น โดยสามารถปรับการตัดสินใจตามระดับที่สนใจได้อย่างง่ายดาย ทำให้เป็นเครื่องมือที่ทรงพลังสำหรับการวิเคราะห์ข้อมูลที่ซับซ้อน

BIRCH Clustering

Balanced Iterative Reducing and Clustering Using Hierarchies (Zhang et al., 1996) เป็นอัลกอริทึมที่มีลักษณะเฉพาะที่ผสมผสานแนวคิดของทั้ง Agglomerative และ Divisive Clustering โดยจะเริ่มจากการรวมข้อมูลเข้าด้วยกันเป็นกลุ่มเล็ก ๆ ก่อน (เหมือน agglomerative) แต่ถ้ากลุ่มไหนใหญ่เกินไปหรือเกินเงื่อนไขที่ตั้งไว้ มันก็จะแตกกลุ่มออก (เหมือน divisive) การวางแผนและการจัดกลุ่มแบบสมดุลที่ดำเนินการโดยวิธีเชิงลำดับขั้น วัตถุประสงค์ของการพัฒนาอัลกอริทึมนี้คือเพื่อทำการจัดกลุ่มข้อมูลได้อย่างมีประสิทธิภาพและแม่นยำ โดยใช้การสแกนชุดข้อมูลเพียงครั้งเดียว โครงสร้างข้อมูลของ BIRCH มีความคล้ายคลึงกับโครงสร้างต้นไม้แบบ B+ tree โดยใช้ต้นไม้ที่เรียกว่า Clustering Feature Tree (CF Tree) แต่ละโหนดในต้นไม้จะประกอบด้วยคุณลักษณะการจัดกลุ่ม (Clustering Features) หลายรายการ และแบ่งเป็นชั้น ๆ จนถึงโหนดใบ (Leaf Nodes) โหนดใบเหล่านี้ยังประกอบด้วย CF หลายรายการเช่นกัน สำหรับโหนดที่ไม่ใช่ใบ (Non-leaf Nodes) จะมีตัวชี้ (Pointer) ไปยังโหนดในชั้นลำดับถัดไป และโหนดใบทั้งหมดจะเชื่อมต่อกันผ่านโครงสร้างข้อมูลแบบ Double Linked List ดังที่แสดงในภาพประกอบ 7



ภาพประกอบ 7 โครงสร้างของต้นไม้คุณลักษณะการจัดกลุ่ม (Clustering Feature Tree)

ตามคำจำกัดความของลักษณะการจัดกลุ่ม (Clustering features) และคุณลักษณะของต้นไม้คุณลักษณะ (feature tree) แนวคิดหลักของ BIRCH คือการสร้างต้นไม้ CF (CF tree) สำหรับตัวอย่างข้อมูลทั้งหมดในชุดข้อมูลของกิจกรรมการเรียนรู้เชิงปฏิสัมพันธ์ ผลลัพธ์ที่ได้หลังจากการจัดกลุ่ม BIRCH คือจำนวนโหนด CF (CF nodes) โดยตัวอย่างของแต่ละโหนดคือกลุ่ม (clusters) และการจัดกลุ่ม BIRCH ของกิจกรรมการเรียนรู้เชิงปฏิสัมพันธ์คือกระบวนการสร้างต้นไม้ CF กระบวนการหลักจะแสดงในอัลกอริทึมการจัดกลุ่ม BIRCH โดยในระหว่างการออกแบบอัลกอริทึมการจัดกลุ่ม BIRCH จำเป็นต้องพิจารณาข้อยกเว้นบางประการ ซึ่งได้แก่

เนื้อหาของหลักสูตรและช่วงเวลาคือสองเงื่อนไขดัชนีที่ใช้ในการแยกชุดตัวอย่างของกิจกรรมการเรียนรู้เชิงปฏิสัมพันธ์ ซึ่งสามารถทำการวิเคราะห์การจัดกลุ่มแบบขนานได้ ในช่วงเริ่มต้นของการสร้างต้นไม้ CF ของกิจกรรมการเรียนรู้เชิงปฏิสัมพันธ์ จำเป็นต้องกรองและลบโหนด CF ที่ผิดปกติบางส่วนออก โดยทั่วไปโหนดเหล่านี้มีจุดตัวอย่างน้อยมาก สำหรับชุดข้อมูลที่อยู่ใกล้กับ CF อย่างมาก ควรดำเนินการรวมชุดข้อมูล

เริ่มต้นโดยใช้การเดินแบบสุ่ม (random walk) เพื่อค้นหาเส้นทางสำคัญของกิจกรรมการเรียนรู้เชิงปฏิสัมพันธ์และดำเนินการเตรียมข้อมูลล่วงหน้าด้วยเบาะแส จากนั้นจึงจัดกลุ่มชุดข้อมูล CF ทั้งหมด เพื่อให้ได้ต้นไม้ CF ที่ดียิ่งขึ้น ซึ่งสามารถกำจัดโครงสร้างต้นไม้ที่ไม่สมเหตุสมผลซึ่งเกิดจากขนาดตัวอย่างที่ใหญ่ และหลีกเลี่ยงผลกระทบอย่างร้ายแรงจากข้อจำกัดจำนวนที่มีต่อโครงสร้างของต้นไม้ CF ซึ่งอาจทำให้โครงสร้างต้นไม้แบ่งย่อยผิดพลาด ผ่านอัลกอริทึมเพื่อสร้างจุดศูนย์กลาง (center points) ของโหนด CF ทั้งหมด จากนั้นจึงจัดประเภทและจัดกลุ่มตัวอย่างทั้งหมดตามค่าของระยะห่างระหว่างจุดศูนย์กลางกับค่าความต่าง (threshold) เพื่อช่วยลดผลลัพธ์การจัดกลุ่มที่ไม่สมเหตุสมผลซึ่งเกิดจากข้อจำกัดบางประการในระหว่างการสร้างต้นไม้ CF

ก่อนการจัดกลุ่ม BIRCH เส้นทางสำคัญของชุดข้อมูลจะถูกเดินทางผ่านด้วยวิธีการเดินแบบสุ่ม (random walk) เพื่อปรับปรุงความเหมาะสมของตัวอย่างที่ใช้ในการจับคู่และเพิ่มความเร็วในการจัดกลุ่ม นอกจากนี้ โหนดของต้นไม้ CF สามารถเพิ่ม ลบ และรวมได้อย่างรวดเร็ว ซึ่งช่วยให้ชุดข้อมูลเริ่มต้นสามารถกรองและดึงข้อมูลล่วงหน้าได้อย่างมีประสิทธิภาพ รวมถึงการระบุข้อมูลที่ผิดปกติและข้อมูลที่รบกวนได้ตั้งแต่เนิ่น ๆ เพื่อปรับปรุงประสิทธิภาพและคุณภาพของการวิเคราะห์ข้อมูล การเริ่มต้น ตามข้อกำหนดการวิเคราะห์ที่แตกต่างกัน ข้อมูลของกิจกรรมการเรียนรู้แบบโต้ตอบจะถูกแบ่งออกเป็นชุดข้อมูลย่อยที่แตกต่างกัน

ขั้นตอนที่ 1: เรียกใช้อัลกอริทึม RW เพื่อสร้างเส้นทางสำคัญของชุดข้อมูลย่อย เส้นทางสำคัญนี้จะถูกแมปกับเงื่อนไขการคัดกรองรองของกิจกรรมการเรียนรู้แบบโต้ตอบ และผลลัพธ์การคัดกรองคือชุดข้อมูลที่จะทำการจัดกลุ่ม

ขั้นตอนที่ 2: อ่านจุดตัวอย่างแรกจากชุดข้อมูลและเติมข้อมูลใน CF (Cluster Feature) ใหม่ให้เป็นโหนดราก

ขั้นตอนที่ 3: หากรัศมีของทรงกลมที่ใช้ CF เป็นจุดศูนย์กลางน้อยกว่าค่าขีดจำกัด T ลังจากเพิ่มตัวอย่างใหม่แล้ว ให้ปรับปรุงข้อมูล CF ทั้งหมดบนเส้นทาง หากไม่เป็นไปตามเงื่อนไข ให้ไปที่ขั้นตอนที่ 4

ขั้นตอนที่ 4: หากจำนวน CF ของโหนดใบปัจจุบันน้อยกว่าค่าขีดจำกัด L ให้สร้าง CF ใหม่และใส่ตัวอย่างใหม่ จากนั้นนำ CF ใหม่นี้เป็นโหนดใบ และปรับปรุงข้อมูล CF ทั้งหมดบนเส้นทาง หากไม่เป็นไปตามเงื่อนไข ให้ไปที่ขั้นตอนที่ 5

ขั้นตอนที่ 5: แยกโหนดใบปัจจุบันออกเป็นสองโหนดใบใหม่ และเลือกข้อมูล CF ทั้งหมดในโหนดใบเดิมที่สอดคล้องกับข้อมูล CF สองชุดที่อยู่ไกลที่สุดจากทรงกลมเป็นโหนด CF แรกของโหนดใบใหม่

ขั้นตอนที่ 6: ตามหลักการของค่าขีดจำกัดความแตกต่างสัมบูรณ์และข้อมูล CF อื่น ๆ ตัวอย่างใหม่จะถูกส่งกลับไปยังโหนดใบที่เกี่ยวข้อง ทำการตรวจสอบโหนดภายในซ้ำ หากต้องการแยก ให้ทำกระบวนการแยกของขั้นตอนที่ 4 และ 5 ซ้ำ

2.5 การวัดประสิทธิภาพแบบจำลอง (Model Performance Metrics)

2.5.1 Silhouette Coefficient (SC)

Silhouette Coefficient (Kaufman & Rousseeuw, 1990) มาตรวัดที่ใช้ประเมินคุณภาพการจัดกลุ่มข้อมูล (Clustering) โดยไม่ต้องอาศัยข้อมูลภายนอกหรือป้ายกำกับ (Label) มาตรวัดนี้ช่วยประเมินว่าแต่ละจุดข้อมูลอยู่ในกลุ่มที่เหมาะสมหรือไม่ โดยพิจารณาจาก 2 ปัจจัยหลัก

1.ความใกล้ชิดภายในกลุ่ม $a(x_i)$: ค่าเฉลี่ยระยะทางระหว่างจุดข้อมูล xi กับจุดข้อมูลอื่นๆ ในกลุ่มเดียวกัน

ค่า $a(x_i)$ ต่ำ แสดงว่าจุดข้อมูลนั้นอยู่ใกล้กับจุดอื่นในกลุ่มเดียวกัน แปลว่าการจัดกลุ่มนั้นเหมาะสม

ค่า $a(x_i)$ สูง แสดงว่าจุดข้อมูลนั้นอยู่ห่างจากจุดอื่นในกลุ่มเดียวกัน แปลว่าการจัดกลุ่มอาจไม่เหมาะสม

2.ความห่างจากกลุ่มอื่น $b(x_i)$: ค่าเฉลี่ยระยะทางระหว่างจุดข้อมูล xi กับจุดข้อมูลในกลุ่มอื่นที่ใกล้ที่สุด

ค่า $b(x_i)$ สูง แสดงว่าจุดข้อมูลนั้นอยู่ห่างจากกลุ่มอื่น แปลว่าการจัดกลุ่มแยกกลุ่มได้ดี

ค่า $b(x_i)$ ต่ำ แสดงว่าจุดข้อมูลนั้นอยู่ใกล้กับกลุ่มอื่น แปลว่าการจัดกลุ่มอาจไม่ชัดเจน

จากนั้นคำนวณ Silhouette Score ของแต่ละจุดข้อมูล xi โดยใช้สูตร

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \qquad (4)$$

โดยค่า Silhouette Score จะอยู่ในช่วง -1 ถึง 1:

ค่าใกล้ 1: แสดงว่าจุดข้อมูลนั้นอยู่ในกลุ่มที่เหมาะสมและแยกจากกลุ่มอื่นได้ดี

ค่าใกล้ 0: แสดงว่าจุดข้อมูลนั้นอาจอยู่ใกล้ขอบเขตระหว่างกลุ่ม

ค่าใกล้ -1: แสดงว่าจุดข้อมูลนั้นควรอยู่ในกลุ่มอื่น

2.5.2 Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) (Frédéric et al., 2023)

เป็นมาตรวัดที่ใช้ประเมินคุณภาพของการจัดกลุ่ม (Clustering) โดยไม่ต้องอาศัยข้อมูลภายนอก (Unsupervised Evaluation) เช่นเดียวกับ Dunn's Index (DI) แต่มีแนวคิดที่แตกต่างกัน โดย DBI จะวัดว่าแต่ละกลุ่มมีความคล้ายคลึงกันแค่ไหน โดยDavies-Bouldin Index ถูกกำหนดโดยพิจารณาจาก 2 ปัจจัยหลัก ดังนี้

1. การกระจายตัวภายในกลุ่ม (Cluster Scatter, S_i)

เป็นการวัดการกระจายของจุดข้อมูลภายในแต่ละกลุ่ม โดยคำนวณจากค่าเฉลี่ยระยะทางของทุกจุดข้อมูลในกลุ่มหนึ่งไปยังเซนทรอยด์ของกลุ่มนั้น หาก S_i มีค่าต่ำ แสดงว่าคลัสเตอร์มีความกระชับดี

2. ระยะห่างระหว่างกลุ่ม (Cluster Separation $d(C_i, C_J)$)

เป็นการวัดระยะห่างระหว่างเซนทรอยด์ของสองกลุ่มที่แตกต่างกัน โดยถ้าระยะห่าง $d(C_i, C_J)$ สูง หมายความว่ากลุ่มถูกแยกออกจากกันได้ดี

สูตรการคำนวณ Davies-Bouldin Index

Davies-Bouldin Index ถูกคำนวณจากค่า R_{ij} ซึ่งเป็นอัตราส่วนระหว่าง การกระจายตัวกลุ่ม และ ระยะห่างระหว่างกลุ่ม

$$R_{ij} = \frac{S_i + S_j}{d(C_i, C_J)} \tag{5}$$

ค่าที่ใช้สำหรับการคำนวณ DBI คือค่าเฉลี่ยของค่า R_{ij} ที่มากที่สุดของแต่ละกลุ่ม

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} R_{ij} \tag{6}$$

โดยที่ S_i = ค่าเฉลี่ยของระยะห่างระหว่างจุดข้อมูลในกลุ่มที่ i กับเซนทรอยด์ของกลุ่ม

$d(C_i, C_J)$= ระยะห่างระหว่างเซนทรอยด์ของกลุ่มที่ i และ j

K = จำนวนกลุ่มทั้งหมด

การแสดงผลค่า Davies-Bouldin Index

หากค่า DBI ต่ำ หมายความว่าการจัดกลุ่มมีคุณภาพดี เพราะ กลุ่มมีการกระจายตัวภายในที่ต่ำ และ มีการแยกกันได้ชัดเจน

หากค่า DBI สูง หมายความว่าการจัดกลุ่มมีคุณภาพไม่ดี เนื่องจาก กลุ่มมีความคล้ายคลึงกันมากเกินไป หรือ เกิดการทับซ้อนกัน

โดยทั่วไป ค่า DBI ใกล้ 0 เป็นค่าที่ดีที่สุด เนื่องจากแสดงว่ากลุ่มมีความแตกต่างกันชัดเจนและมีความกระชับที่ดี

2.5.3 Calinski-Harabasz Index (CHI)

Calinski-Harabasz Index (CHI) หรือที่เรียกว่า Variance Ratio Criterion (VRC) เป็นอีกหนึ่งมาตรวัดที่ใช้ประเมินคุณภาพของการจัดกลุ่ม (Clustering) โดยพิจารณาความกระชับของจุดข้อมูลภายในกลุ่มและการแยกจากกันของกลุ่ม ซึ่งแนวคิดหลักของ Calinski-Harabasz Index (CHI) คำนวณโดยใช้ อัตราส่วนระหว่างการกระจายตัวระหว่างกลุ่ม (Between-Cluster Dispersion) และ การกระจายตัวภายในกลุ่ม (Within-Cluster Dispersion)

1. การกระจายตัวระหว่างกลุ่ม (Between-Cluster Dispersion , B_k)

คือวัดระยะห่างระหว่างเซนทรอยด์ของแต่ละกลุ่มกับเซนทรอยด์ของข้อมูลทั้งหมด ถ้าค่ามากแสดงว่ากลุ่มมีการแยกกันชัดเจน

2. การกระจายตัวภายในกลุ่ม (Within-Cluster Dispersion, W_k)

เป็นการวัดว่าจุดข้อมูลในแต่ละกลุ่มกระจายตัวมากน้อยเพียงใด ถ้าค่าต่ำ แสดงว่าจุดข้อมูลภายในกลุ่มมีความกระชับกันดี โดยสูตรการคำนวณ Calinski-Harabasz Index มีดังนี้

$$CHI = \frac{\frac{B_k}{k-1}}{\frac{W_k}{(n-k)}} \qquad (7)$$

B_k = ผลรวมของความแปรปรวนระหว่างกลุ่ม

W_k = ผลรวมของความแปรปรวนภายในกลุ่ม

k = จำนวนคกลุ่ม

n = จำนวนข้อมูลทั้งหมด

การแสดงผลผลค่า Calinski-Harabasz Index

หากค่า CHI สูง แสดงว่าการจัดกลุ่มมีคุณภาพดี เพราะ กลุ่มแยกจากกันชัดเจนและแต่ละกลุ่มมีความกระชับ

หากค่า CHI ต่ำ หมายความว่าการจัดกลุ่มอาจไม่ดี เนื่องจากกลุ่มอาจทับซ้อนกันหรือจุดข้อมูลภายในกลุ่มกระจายตัวมากเกินไป

2.6 งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยฉบับนี้ได้ศึกษาและค้นคว้างานวิจัยที่เกี่ยวข้องต่างๆ โดยมีรายละเอียดดังนี้

1. บทความวิจัยเรื่อง SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

โดย J. MacQueen (1967) ได้นำเสนอ K-Means Clustering Algorithm ซึ่งเป็นกระบวนการที่ใช้ในการแบ่งกลุ่มข้อมูลแบบ Partition-Based Clustering โดยใช้หลักการลด Within-Cluster Variance เพื่อทำให้ข้อมูลภายในแต่ละกลุ่มมีความคล้ายคลึงกันมากที่สุด MacQueen อธิบายว่า K-Means จะทำงานโดยเริ่มจากการสุ่มเลือกจุดศูนย์กลางของแต่ละกลุ่ม จากนั้นกำหนดให้แต่ละจุดข้อมูลอยู่ในกลุ่มที่มีศูนย์กลางใกล้ที่สุด และอัปเดตศูนย์กลางของแต่ละกลุ่มซ้ำ ๆ จนกว่าศูนย์กลางจะไม่เปลี่ยนแปลง กระบวนการนี้สามารถใช้ได้กับ N-dimensional population และเป็นวิธีที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลขนาดใหญ่บนคอมพิวเตอร์ นอกจากนี้ MacQueen ยังได้ศึกษา คุณสมบัติของ K-Means ในเชิงทฤษฎี โดยเชื่อมโยงกับแนวคิดของกฎจำนวนมาก (Law of Large Numbers) เพื่ออธิบายพฤติกรรมของอัลกอริธึมเมื่อจำนวนข้อมูลเพิ่มขึ้น และรายงานผลการทดลองจากการประมวลผลข้อมูลจริงที่แสดงให้เห็นว่า K-Means มีแนวโน้มที่จะลด Within-Cluster Variance ได้ดี อย่างไรก็ตาม งานวิจัยยังชี้ให้เห็นว่าอัลกอริธึมนี้ อาจไม่สามารถหาผลลัพธ์ที่เป็น Optimal Partition เสมอไป เนื่องจากขึ้นอยู่กับจุดเริ่มต้นและอาจติดอยู่ในค่าที่ไม่ใช่ค่าที่ดีที่สุด (Local Optima) โดยไม่มีวิธีที่สามารถรับประกันได้ว่าจะได้คำตอบที่ดีที่สุดเสมอ นอกจากนี้ งานวิจัยยังกล่าวถึง การนำ K-Means ไปใช้ในงานด้านการจัดกลุ่มข้อมูล (Similarity Grouping), การพยากรณ์แบบไม่เชิงเส้น (Nonlinear Prediction), และการทดสอบความเป็นอิสระของตัวแปร (Nonparametric Tests for Independence) ทำให้ K-Means Clustering เป็นเทคนิคที่มีศักยภาพสูงสำหรับการวิเคราะห์ข้อมูลแบบ Multivariate (MacQueen, 1967)

2. บทความวิจัยเรื่อง Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering

งานวิจัยนี้นำเสนอแนวทางใหม่ในการแบ่งกลุ่มลูกค้า (Customer Segmentation) โดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Machine Learning) และการรวมกลุ่ม (Clustering Ensemble) เพื่อเพิ่มความแม่นยำและประสิทธิภาพในการแบ่งกลุ่ม โดยใช้ 4 อัลกอริธึมพื้นฐาน ได้แก่ DBSCAN, K-means, Mini Batch K-means และ Mean Shift เพื่อสร้างผลลัพธ์ที่หลากหลาย และนำผลลัพธ์ที่ได้มาปรับปรุงด้วย Spectral Clustering เพื่อรวมกลุ่มข้อมูลให้แม่นยำยิ่งขึ้น การประเมินผลใช้ดัชนี ARI, NMI, SC และ DI โดยพบว่าโมเดลที่นำเสนอให้ผลลัพธ์ที่ดีที่สุดเมื่อเทียบกับอัลกอริธึมพื้นฐาน นอกจากนี้ งานวิจัยยังได้ทดลองใช้โมเดลกับข้อมูลของประชาชนโมร็อกโกในช่วงวันที่ 03/06/2022 ถึง 19/08/2022 และพบว่าสามารถแบ่งกลุ่มลูกค้าได้อย่างมีประสิทธิภาพ ผลลัพธ์ชี้ให้เห็นถึงศักยภาพของการใช้ Clustering Ensemble และ Spectral Clustering ในการแบ่งกลุ่มลูกค้า ซึ่งสามารถนำไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนากลยุทธ์การตลาดได้อย่างมีประสิทธิภาพ อีกทั้งยังสามารถขยายขอบเขตการศึกษาด้วยการทดสอบโมเดลกับชุดข้อมูลอื่น ๆ และเปรียบเทียบกับอัลกอริธึมขั้นสูง เช่น Deep Learning เพื่อเพิ่มความแม่นยำในการแบ่งกลุ่มลูกค้าต่อไป (Hicham & Karim, 2022)

3. บทความวิจัยเรื่อง Customer Segmentation using RFM Model and K-Means Clustering

งานวิจัยนี้มีเป้าหมายในการแบ่งกลุ่มลูกค้าโดยใช้ RFM Model (Recency, Frequency, Monetary) และ K-Means Clustering เพื่อวิเคราะห์พฤติกรรมการซื้อของลูกค้าและช่วยให้ธุรกิจสามารถกำหนดกลยุทธ์ทางการตลาดได้อย่างแม่นยำ โดยใช้ชุดข้อมูล UK's E-commerce dataset จาก UCI Machine Learning Repository ซึ่งครอบคลุมธุรกรรมการซื้อระหว่างปี 2010-2011 งานวิจัยนี้ดำเนินการผ่าน 6 ขั้นตอน ได้แก่ การทำความเข้าใจธุรกิจ, การวิเคราะห์ข้อมูล, การเตรียมข้อมูล, การพัฒนาโมเดล, การประเมินผล และการพัฒนาเว็บแอปพลิเคชัน สำหรับการวิเคราะห์ข้อมูลลูกค้า โดยใช้ Elbow Method กำหนดจำนวนกลุ่มที่เหมาะสม และวัดประสิทธิภาพด้วย Silhouette Index (ค่าเฉลี่ย 0.442) ผลลัพธ์ที่ได้แบ่งลูกค้าออกเป็น 4 กลุ่ม ได้แก่ Class A (ลูกค้ารายได้สูงสุด), Class B (ลูกค้าปานกลาง), Class C (ลูกค้ากำลังลดการซื้อ), และ Class D (ลูกค้าที่ซื้อน้อยที่สุด) การวิจัยสรุปว่าโมเดลนี้สามารถช่วยให้ธุรกิจเข้าใจพฤติกรรมของลูกค้าได้ดีขึ้นและสามารถพัฒนากลยุทธ์ทางการตลาดเพื่อลดอัตราการสูญเสียลูกค้าได้อย่างมีประสิทธิภาพ

โดยสามารถนำไปปรับปรุงเพิ่มเติมได้ด้วยวิธีการ Clustering อื่น เช่น Hierarchical Clustering หรือ DBSCAN เพื่อเพิ่มความแม่นยำในการจัดกลุ่มลูกค้า (Rahul et al., 2021)

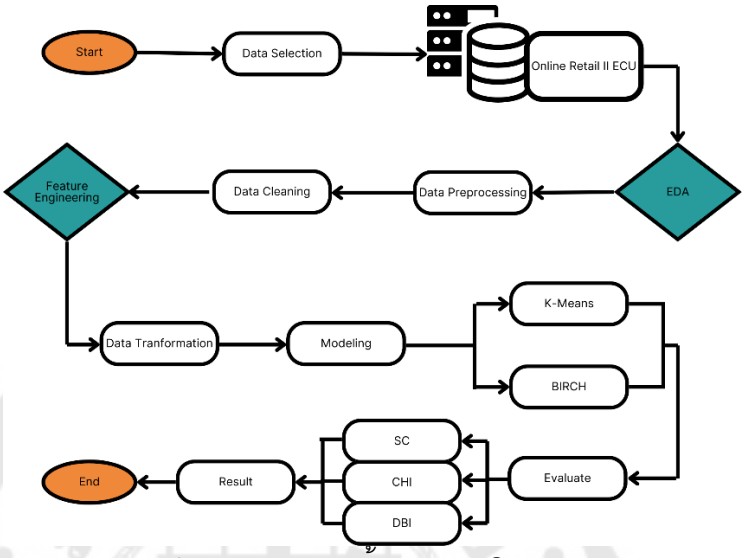
4. บทความวิจัยเรื่อง Variations on the Clustering Algorithm BIRCH

งานวิจัย "Variations on the Clustering Algorithm BIRCH" มุ่งเน้นการพัฒนา BIRCH ให้ทำงานได้ดีขึ้นโดยไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า ซึ่งเป็นปัญหาหลักของอัลกอริธึมการจัดกลุ่มหลายตัว ทีมวิจัยได้เสนอ A-BIRCH ที่ช่วยให้ BIRCH คำนวณค่าพารามิเตอร์ threshold ได้อัตโนมัติ ทำให้สามารถจัดกลุ่มข้อมูลได้แม่นยำขึ้นโดยไม่ต้องพึ่งการตั้งค่าจากผู้ใช้ และลดขั้นตอนการจัดกลุ่มสุดท้ายที่มักใช้ทรัพยากรสูง นอกจากนี้ยังพัฒนา MBD-BIRCH (Multiple Branch Descent BIRCH) ซึ่งช่วยลดข้อผิดพลาดจากการที่กลุ่มถูกแบ่งผิดพลาด (supercluster splitting) ที่มักเกิดในโครงสร้างต้นไม้ลึก ผลการทดลองแสดงให้เห็นว่าการปรับปรุงนี้ช่วยให้ BIRCH ทำงานได้เร็วขึ้นและจัดกลุ่มข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพโดยไม่ต้องกำหนดค่าพารามิเตอร์ที่ยุ่งยาก (Boris et al., 2018)

5. บทความวิจัยเรื่อง Comparison of Topic Modelling Approaches in the Banking Context งานวิจัยนี้เปรียบเทียบเทคนิค Hierarchical Clustering และ Partition-Based Clustering สำหรับการจำแนกหัวข้อในบริบทของอุตสาหกรรมธนาคาร โดยใช้ข้อมูลจาก Twitter ของลูกค้าธนาคารในไนจีเรีย ซึ่งพบว่า Hierarchical Dirichlet Process (HDP) แม้จะสามารถปรับจำนวนกลุ่มได้แบบไดนามิกโดยไม่ต้องกำหนดล่วงหน้า แต่กลับมีปัญหาในการจัดการกับข้อมูลสั้นและกระจัดกระจาย เช่น ทวีตของลูกค้า ทำให้ได้ค่า coherence score ต่ำกว่าเทคนิคอื่น ๆ ในทางตรงกันข้าม Partition-Based Clustering โดยเฉพาะ K-Means ร่วมกับ Kernel Principal Component Analysis (KernelPCA) ใน BERTopic สามารถลดมิติของข้อมูลและเพิ่มประสิทธิภาพการจัดกลุ่มได้ดีกว่า ส่งผลให้ได้ค่า coherence score สูงสุดที่ 0.8463 ซึ่งเหนือกว่าการใช้ LDA, LSI หรือ HDP งานวิจัยนี้สรุปว่า แม้ Hierarchical Clustering จะมีข้อดีด้านความยืดหยุ่น แต่สำหรับข้อมูลสั้นและซับซ้อนเช่น ข้อความจากโซเชียลมีเดีย การใช้ Partition-Based Clustering ร่วมกับเทคนิคลดมิติ เช่น KernelPCA มีประสิทธิภาพสูงกว่าในการจำแนกหัวข้อและวิเคราะห์ความคิดเห็นของลูกค้าในภาคธนาคาร (Ogunleye et al., 2023)

งานวิจัยที่เกี่ยวข้องจะมีการประยุกต์ใช้อัลกอริธึมการจัดกลุ่มและการวิเคราะห์พฤติกรรมลูกค้าอย่างหลากหลาย แต่ยังมีข้อจำกัดบางประการที่แตกต่างจากบริบทของงานวิจัยฉบับนี้ ซึ่งรายละเอียดของความแตกต่างดังกล่าวจะนำเสนออย่างชัดเจนในบทที่ 3

บทที่3
ขั้นตอนการดำเนินการวิจัย



ภาพประกอบ 8 แสดงขั้นตอนการวิจัย

กระบวนการวิจัยนี้ครอบคลุมการดำเนินงานที่เกี่ยวข้องกับการเลือกข้อมูล การเตรียมข้อมูล การปรับเปลี่ยนข้อมูล และการวิศวกรรมคุณลักษณะเพื่อสร้างโมเดลการจัดกลุ่มลูกค้าที่มีประสิทธิภาพ โดยมีขั้นตอนดังนี้

3.1 Data Selection (**การเลือกข้อมูล**):

ข้อมูลที่ใช้ในการวิจัยนี้มาจาก Online Retail II UCI บนแพลตฟอร์ม Kaggle ซึ่งประกอบด้วยข้อมูลการสั่งซื้อสินค้าออนไลน์ โดยรวมถึงข้อมูลที่สำคัญหลายประเภท เช่น ข้อมูลลูกค้า พฤติกรรมการซื้อ สินค้า คำสั่งซื้อ เพื่อให้ได้ข้อมูลที่เหมาะสมในการวิเคราะห์และการสร้างโมเดลการจัดกลุ่ม ข้อมูลที่คัดเลือกจะถูกตรวจสอบและกำหนดคุณสมบัติที่สำคัญในการวิจัยนี้ เช่น ลักษณะภูมิศาสตร์ พฤติกรรมการซื้อ ความถี่ในการสั่งซื้อ และความต้องการการจัดส่งในพื้นที่ต่าง ๆ

ภาพประกอบ 9 แสดงหน้าเว็บไซต์ของข้อมูลที่นำมาใช้

3.2 Data understanding (**การทำความเข้าใจข้อมูล**)

สืบค้นและรวบรวมข้อมูลที่น่าเชื่อถือ พร้อมประเมินคุณภาพข้อมูลเพื่อให้เข้าใจลักษณะเฉพาะ และเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์อย่างเหมาะสม จากข้อมูลที่เก็บตั้งแต่ปี 2009 - 2011 มีข้อมูลทั้งหมด 1,067,371 รายการ มี 8 คุณลักษณะ รายละเอียดตามตารางที่ 1 โดยที่ Customer ID มี Missing value 243,007 รายการ และมีแถวที่มีข้อมูลซ้ำกันจำนวน 34,335 รายการ InvoiceDate มี Data Type เป็น object รายละเอียดตามรูปภาพที่ 10

ตาราง 1 แสดงรายละเอียดคุณลักษณะของชุดข้อมูล

Column Name	Description
Invoice	เลขใบแจ้งหนี้
StockCode	รหัสสินค้า
Description	รายละเอียดสินค้า
Quantity	จำนวนสินค้า
InvoiceDate	วันที่ออกใบแจ้งหนี้
Price	ราคา
Customer ID	รหัสลูกค้า
Country	ประเทศ

```
print(dt.shape)
```

```
 Dimension of Data is :
(1067371, 8)
```

```
# สร้าง DataFrame สำหรับการแสดงผลลัพธ์
results = pd.DataFrame({
    'Column': dt.columns,
    'Data Type': dt.dtypes.values,
    'Duplicated Rows': [dt.duplicated().sum()] * len(dt.columns),
    'Missing Values': dt.isnull().sum().values
})

# แสดงผล
print(results)
```

```
         Column Data Type  Duplicated Rows  Missing Values
0       Invoice    object            34335               0
1     StockCode    object            34335               0
2   Description    object            34335            4382
3      Quantity     int64            34335               0
4   InvoiceDate    object            34335               0
5         Price   float64            34335               0
6   Customer ID   float64            34335          243007
7       Country    object            34335               0
```

ภาพประกอบ 10 แสดงรายละเอียดของข้อมูลที่นำมาใช้

3.3 Data Preparation (**การเตรียมข้อมูล**)

1. Data Cleaning (**การทำความสะอาดข้อมูล**)

เป็น กระบวนการจัดการและปรับปรุงข้อมูลดิบ (Raw Data) ให้มีคุณภาพและความถูกต้องเพื่อให้อยู่ในรูปแบบที่พร้อมสำหรับการวิเคราะห์หรือใช้งานในกระบวนการต่าง ๆ เช่น การสร้างแบบจำลอง หรือการแสดงผลข้อมูล กระบวนการนี้เป็นขั้นตอนสำคัญใน Data Preprocessing เพื่อให้ได้ข้อมูลที่เชื่อถือได้และปราศจากความผิดพลาด โดย ขั้นตอนที่พบใน Data Cleaning มีรายละเอียดตามตาราง 2

ตาราง 2 อธิบายการทำ Data Cleaning

ขั้นตอน	รายละเอียด
การจัดการค่าข้อมูลที่ขาดหาย (Missing Data)	เติมค่าที่ขาดหายด้วยค่าเฉลี่ย ค่ากลาง หรือค่าที่เหมาะสม หรืออาจลบแถว/คอลัมน์ที่มีข้อมูลขาดหายมากเกินไป
การลบข้อมูลซ้ำ (Duplicate Data)	ตรวจสอบและลบแถวข้อมูลที่ซ้ำกันเพื่อลดความซับซ้อนและป้องกันข้อมูลผิดพลาด
การจัดการข้อมูลผิดพลาด (Erroneous Data)	แก้ไขข้อมูลที่ไม่สมเหตุสมผล เช่น วันที่ผิดช่วง หรือค่าที่อยู่นอกขอบเขต
การแปลงค่าข้อมูล (Data Transformation)	ปรับเปลี่ยนรูปแบบข้อมูล เช่น การแปลงข้อความให้เป็นตัวพิมพ์เล็ก/ใหญ่ การเปลี่ยนหน่วย หรือการเข้ารหัสข้อมูล
การจัดการข้อมูลนอกกรอบ (Outliers)	ตรวจสอบค่าที่แปลกหรือสุดขั้ว และตัดสินใจว่าจะลบหรือปรับปรุง
การจัดโครงสร้างข้อมูล	ปรับข้อมูลให้อยู่ในรูปแบบที่เป็นระเบียบ เช่น การแบ่งคอลัมน์ หรือการรวมข้อมูลจากหลายแหล่ง

Data Cleaning จึงมีบทบาทสำคัญในการเพิ่มความถูกต้องและความน่าเชื่อถือของผลการวิเคราะห์ข้อมูล ช่วยป้องกันข้อผิดพลาดที่อาจเกิดขึ้นในโมเดลหรือกระบวนการวิเคราะห์อีกทั้งยังช่วยลดความซับซ้อนของข้อมูล ทำให้การดำเนินงานในขั้นตอนถัดไปง่ายและมีประสิทธิภาพมากขึ้น โดยสรุป การทำความสะอาดข้อมูลคือกระบวนการที่จำเป็นสำหรับการปรับปรุงคุณภาพของข้อมูลและลดความเสี่ยงจากข้อผิดพลาดในกระบวนการจัดการข้อมูลทั้งหมด โดยมีการทำความสะอาดข้อมูลตามภาพประกอบด้านล่าง

```
# แปลงประเภทข้อมูล
ddt['InvoiceDate'] = pd.to_datetime(ddt['InvoiceDate'])
ddt.dtypes
```

	0
Invoice	object
StockCode	object
Description	object
Quantity	int64
InvoiceDate	datetime64[ns]
Price	float64
Customer ID	float64
Country	object

dtype: object

ภาพประกอบ 11แสดงชนิดของข้อมูลที่ถูกแก้ไข

```
#ลบข้อมูลที่ค่า CustomerID ว่างเปล่า (Blank)
ddt['Customer ID'].replace('', np.nan, inplace=True)
ddt.dropna(subset=['Customer ID'], inplace=True)
print('\n Dimension of Data is : ')
print(ddt.shape)
```

```
 Dimension of Data is :
(824364, 8)
```

ภาพประกอบ 12 แสดง Dimension หลังจากลบค่า Blank

```
[ ] # ลบข้อมูลที่บันทึกซ้ำ
    dddt=ddt.drop_duplicates(keep='first')
    print(dddt.shape)
```

```
(797885, 8)
```

ภาพประกอบ 13 แสดง Dimension หลังจากลบค่า Duplicates

```
[16] # ตรวจสอบค่า Price และ Quantity ที่มีค่า <= 0
     price_zero_count = (dddt[['Price']] <= 0).sum()
     quantity_negative_count = (dddt[['Quantity']] <= 0).sum()

     # สร้าง DataFrame เพื่อแสดงผล
     summary_table = pd.DataFrame({
         'Condition': ['Price <= 0', 'Quantity <= 0'],
         'Count': [price_zero_count.values[0], quantity_negative_count.values[0]]
     })

     # แสดงผลลัพธ์
     print(summary_table)
```

```
        Condition  Count
0     Price <= 0     70
1  Quantity <= 0  18390
```

```
[ ]  # คำนวณค่าเฉลี่ยของ Price โดยไม่รวมค่าที่เป็น 0
     average_price = dddt.loc[dddt['Price'] != 0, 'Price'].mean()
     # แทนค่าที่เป็น 0 ด้วยค่าเฉลี่ย
     dddt['Price'] = dddt['Price'].replace(0, average_price)
```

```
[18] # คำนวณค่าเฉลี่ยของ Quantity โดยไม่รวมค่าที่เป็น 0 และค่าติดลบ
     average_quantity = dddt.loc[dddt['Quantity'] > 0, 'Quantity'].mean()
     # แทนค่าที่เป็น 0 และค่าติดลบด้วยค่าเฉลี่ย
     dddt.loc[dddt['Quantity'] <= 0, 'Quantity'] = average_quantity
```

```
[19] # ตรวจสอบค่า Price และ Quantity ที่มีค่า <= 0
     price_zero_count = (dddt[['Price']] <= 0).sum()
     quantity_negative_count = (dddt[['Quantity']] <= 0).sum()

     # สร้าง DataFrame เพื่อแสดงผล
     summary_table = pd.DataFrame({
         'Condition': ['Price <= 0', 'Quantity <= 0'],
         'Count': [price_zero_count.values[0], quantity_negative_count.values[0]]
     })

     # แสดงผลลัพธ์
     print(summary_table)
```

```
        Condition  Count
0     Price <= 0      0
1  Quantity <= 0      0
```

ภาพประกอบ 14 การแทนค่าเฉลี่ยลงใน Price และ Quantity มีค่าน้อยกว่าหรือเท่ากับ 0

2. Future Engineering (**การสร้างตัวแปรใหม่จากข้อมูลที่มีอยู่**)

คือกระบวนการปรับปรุงหรือสร้างตัวแปรใหม่ (features) เพื่อเพิ่มประสิทธิภาพของโมเดล Machine Learning โดยเน้นการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและเป็นประโยชน์ที่สุด ตัวอย่างกระบวนการที่สำคัญในงานวิจัยนี้คือการนำข้อมูลที่มีอยู่มาสร้างตัวแปร RFM เพิ่ม เนื่องจากข้อมูลเดิมนั้นเป็นแค่การเก็บข้อมูลเบื้องต้นจากวันที่ทำรายการ จำนวน และราคาของสินค้าที่ถูกสั่งซื้อโดยมีการพิจารณาตัวแปรดังนี้

1. Recency (R): ความใหม่ของการซื้อครั้งล่าสุด

วัดจากระยะเวลาตั้งแต่การทำธุรกรรมครั้งล่าสุดจนถึงปัจจุบัน ลูกค้าที่ซื้อสินค้าหรือใช้บริการล่าสุดมักแสดงถึงความสนใจหรือความสัมพันธ์ที่ดีต่อธุรกิจ เช่น ลูกค้าที่ซื้อสินค้าซ้ำภายใน 7 วันที่ผ่านมาอาจถือว่ามีความ "ใหม่" สูง ตามภาพประกอบที่ โดยมีสมการทางคณิตศาสตร์ตามสมการที่ดังนี้

$$Recency = (Recent_{date} - Last\ invoiceDate) \qquad (8)$$

```
[22] #Calculating Recency
     data_recency = data.groupby(by='Customer ID',
                                 as_index=False)['InvoiceDate'].max()
     data_recency.columns = ['Customer ID', 'LastInvoiceDate']
     recent_date = data_recency['LastInvoiceDate'].max()
     data_recency['Recency'] = data_recency['LastInvoiceDate'].apply(
         lambda x: (recent_date - x).days)
     print(data_recency.head(5))
```

```
   Customer ID     LastInvoiceDate  Recency
0      12346.0 2011-01-18 10:17:00      325
1      12347.0 2011-12-07 15:52:00        1
2      12348.0 2011-09-25 13:13:00       74
3      12349.0 2011-11-21 09:51:00       18
4      12350.0 2011-02-02 16:01:00      309
```

ภาพประกอบ 15 การสร้างตัวแปร Recency

2. Frequency (F): ความถี่ในการซื้อสินค้า/ใช้บริการ

วัดจำนวนครั้งที่ลูกค้าทำธุรกรรมภายในช่วงเวลาที่กำหนด ลูกค้าที่มีความถี่ในการซื้อสินค้าสูงแสดงถึงความสัมพันธ์ที่ต่อเนื่องและความภักดีต่อธุรกิจ ตามภาพประกอบที่ โดยมีสมการทางคณิตศาสตร์ตามสมการด้านล่าง

$$Frequency = Count(InvoiceDate) \quad (9)$$

```
#Calculating Frequency
frequency_data = data.drop_duplicates().groupby(
    by=['Customer ID'], as_index=False)['InvoiceDate'].count()
frequency_data.columns = ['Customer ID', 'Frequency']
frequency_data.head(5)
```

	Customer ID	Frequency
0	12346.0	46
1	12347.0	222
2	12348.0	51
3	12349.0	180
4	12350.0	17

ภาพประกอบ 16 การสร้างตัวแปร Frequency

3. Monetary (M): มูลค่ารวมของการซื้อสินค้าหรือการใช้บริการ

วัดมูลค่าทั้งหมดของเงินที่ลูกค้าใช้จ่ายกับธุรกิจ ลูกค้าที่มีมูลค่าการใช้จ่ายที่สูง ถือเป็นกลุ่มลูกค้าที่สำคัญ เพราะสร้างรายได้ให้กับธุรกิจมาก

$$Total_k = Price_{\ k} \times Quantity_k \qquad (10)$$

$Total_k =$ มูลค่ารวมของรายการที่ k
$Price_{\ k} =$ ราคาต่อหน่วยที่ซื้อสินค้าของรายการที่ k
$Quantity_k =$ จำนวนสินค้าที่ซื้อของรายการที่ k

$$Monetary_i = \sum_{k \in Transaction_i} Total_k \qquad (11)$$

$Total_k =$ มูลค่ารวมของแต่ละรายการธุรกรรมที่ลูกค้า $\boldsymbol{i}$ ทำไว้
$Transactions_{\ i} =$ เซตรายการธุรกรรมทั้งหมดของลูกค้า i
$Monetary_i =$ มูลค่ารวมทั้งหมดที่ลูกค้า i ใช้จ่าย

```
#Calculating Monetary Value
data['Total'] = data['Price']*data['Quantity']
monetary_data = data.groupby(by='Customer ID', as_index=False)['Total'].sum()
monetary_data.columns = ['Customer ID', 'Monetary']
monetary_data.head(5)
```

	Customer ID	Monetary
0	12346.0	82420.361206
1	12347.0	4921.530000
2	12348.0	2019.400000
3	12349.0	4754.886096
4	12350.0	334.400000

ภาพประกอบ 17 การสร้างตัวแปร Monetary

4. Country (C): ประเทศที่ลูกค้าใช้บริการ

มีประโยชน์ในการวิเคราะห์ข้อมูลเชิงธุรกิจและการสร้างโมเดล Machine Learning โดยช่วยระบุแหล่งที่มาของลูกค้า ทำให้สามารถวิเคราะห์พฤติกรรมการซื้อในแต่ละประเทศได้ เช่น ความแตกต่างของยอดขาย การตั้งราคาสินค้า หรือแนวโน้มการซื้อเฉพาะพื้นที่ ในเชิงการตลาดสามารถใช้เพื่อปรับแต่งกลยุทธ์ให้เหมาะกับแต่ละภูมิภาค เช่น การทำแคมเปญส่งเสริมการขายหรือการบริหารสินค้าคงคลัง นอกจากนี้ ในงาน Machine Learning การเข้ารหัสค่าประเทศ (Country Encoding) ช่วยให้โมเดลสามารถใช้ข้อมูลประเทศเป็นปัจจัยในการทำนายผลลัพธ์ได้อย่างมีประสิทธิภาพมากขึ้น

โดยโค้ดนี้ใช้ Label Encoding เพื่อแปลงค่าหมวดหมู่ในคอลัมน์ Country ให้เป็นตัวเลขสำหรับใช้ในโมเดล Machine Learning โดยเริ่มจากการตรวจสอบจำนวนประเทศที่ไม่ซ้ำกันใน data2['Country'] ด้วย .nunique() จากนั้นใช้ LabelEncoder() จาก sklearn.preprocessing เพื่อเข้ารหัสชื่อประเทศเป็นตัวเลข และสร้างคอลัมน์ใหม่ชื่อ Country_encoded ซึ่งเก็บค่าที่เข้ารหัสแล้ว สุดท้ายแสดงผลลัพธ์โดยลบค่าที่ซ้ำกันออกและเรียงลำดับตามรหัสของประเทศ ซึ่งช่วยให้สามารถนำข้อมูลประเทศไปใช้ในโมเดลได้ง่ายขึ้นแทนการใช้ข้อมูลตัวอักษร

```python
# Country
import pandas as pd
from sklearn.preprocessing import LabelEncoder

# ตรวจสอบค่าประเทศที่มีในคอลัมน์ 'Country'
print("Unique Countries:", data2['Country'].nunique())

# ใช้ Label Encoding แปลงชื่อประเทศเป็นตัวเลข
le = LabelEncoder()
data2['Country_encoded'] = le.fit_transform(data2['Country'])

# ตรวจสอบผลลัพธ์
print(data2[['Country', 'Country_encoded']].drop_duplicates().sort_values('Country_encoded'))
```

```
Unique Countries: 41
                  Country  Country_encoded
178             Australia                0
17287             Austria                1
206105            Bahrain                2
173               Belgium                3
348243             Brazil                4
416551             Canada                5
9628       Channel Islands                6
12263               Cyprus                7
629059      Czech Republic                8
4793               Denmark                9
440                   EIRE               10
```

ภาพประกอบ 18 การสร้างตัวแปร Country

5. AVG Basket Price: ราคาเฉลี่ยต่อคำสั่งซื้อ

คำนวณจาก ยอดซื้อรวมของลูกค้าแต่ละราย หารด้วยจำนวนครั้งที่ลูกค้าคนนั้นซื้อสินค้า ช่วยให้ธุรกิจเข้าใจพฤติกรรมการใช้จ่ายของลูกค้าได้ดีขึ้น และนำไปใช้วางแผนกลยุทธ์เพิ่มยอดขาย เช่น การทำโปรโมชั่น การขายสินค้าเป็นชุด (Bundle) หรือการแนะนำสินค้าเพิ่มเติม เพื่อให้ลูกค้าซื้อเพิ่ม โดยมีสมการดังนี้

$$\text{ราคาเฉลี่ยต่อคำสั่งซื้อ} = \frac{\text{ยอดซื้อรวมของลูกค้าแต่ละราย}}{\text{จำนวนครั้งที่ลูกค้าซื้อสินค้า}} \qquad (12)$$

```python
# ✅ ตรวจสอบว่ามี 'TotalPrice' และ 'Customer ID' หรือไม่
required_columns = ['Customer ID', 'TotalPrice']
missing_columns = [col for col in required_columns if col not in data2.columns]

if missing_columns:
    raise KeyError(f"Missing columns: {missing_columns}. Please check your dataset.")

# ✅ คำนวณค่าเฉลี่ย Total Basket Price ของแต่ละลูกค้า (Customer ID)
avg_basket_price = data2.groupby('Customer ID')['TotalPrice'].mean().reset_index()

# ✅ เปลี่ยนชื่อคอลัมน์ให้เป็น 'AVG_Basket_Price'
avg_basket_price.rename(columns={'TotalPrice': 'AVG_Basket_Price'}, inplace=True)

# ✅ แสดงตัวอย่างผลลัพธ์ 10 แถวแรก
display(avg_basket_price.head(10))
```

	Customer ID	AVG_Basket_Price
0	12346.0	1719.182319
1	12347.0	22.266087
2	12348.0	39.596078
3	12349.0	26.389278
4	12350.0	19.670588
5	12351.0	14.330000
6	12352.0	126.144918
7	12353.0	16.948333
8	12354.0	18.610345
9	12355.0	27.074571

ภาพประกอบ 19 การสร้างตัวแปร AVG Basket Price

6. Day Type : ประเภทของวัน

ใช้ในการวิเคราะห์ข้อมูลพฤติกรรมลูกค้า โดยแบ่งวันออกเป็นหมวดหมู่ต่างๆ เช่น วันธรรมดา (Weekday) vs. วันหยุด (Weekend), วันทำงาน vs. วันหยุดนักขัตฤกษ์, หรืออาจรวมถึงฤดูกาล หรือ ช่วงเทศกาลพิเศษ เช่น Black Friday หรือวันปีใหม่ การจำแนกประเภทของวันช่วยให้ธุรกิจเข้าใจแนวโน้มการซื้อของลูกค้าในแต่ละช่วงเวลา และนำไปปรับกลยุทธ์การขาย โปรโมชั่น หรือการบริหารสินค้าคงคลังให้เหมาะสมกับพฤติกรรมผู้บริโภค

โดยโค้ดนี้ทำหน้าที่สร้างคอลัมน์ Day type เพื่อระบุว่าแต่ละวันที่อยู่ใน Invoice Date เป็นวันธรรมดา (Weekday) หรือ วันหยุดสุดสัปดาห์ (Weekend) โดยเริ่มจากการแปลง Invoice Date ให้เป็นรูปแบบ datetime จากนั้นใช้ฟังก์ชัน get_day_type() เพื่อตรวจสอบว่าวันที่นั้นเป็นวันจันทร์-ศุกร์ (คืนค่า "Weekday") หรือวันเสาร์-อาทิตย์ (คืนค่า "Weekend") แล้วนำฟังก์ชันนี้ไปใช้กับทุกค่าของ Invoice Date ผ่าน .apply() เพื่อสร้างคอลัมน์ใหม่ สุดท้ายแสดงผลลัพธ์โดยลบค่าที่ซ้ำกันและแสดงตัวอย่างข้อมูล ซึ่งช่วยให้สามารถวิเคราะห์พฤติกรรมการซื้อของลูกค้าในวันธรรมดาและวันหยุดได้ง่ายขึ้น

```python
# Daytype

# แปลง InvoiceDate เป็น datetime (ถ้ายังไม่ได้แปลง)
data2['InvoiceDate'] = pd.to_datetime(data2['InvoiceDate'])

# ฟังก์ชันกำหนดประเภทของวัน
def get_day_type(date):
    if date.weekday() < 5:  # วันจันทร์ (0) ถึงศุกร์ (4) -> Weekday
        return 'Weekday'
    else:  # วันเสาร์ (5) และอาทิตย์ (6) -> Weekend
        return 'Weekend'

# สร้างคอลัมน์ 'Daytype'
data2['Daytype'] = data2['InvoiceDate'].apply(get_day_type)

# ตรวจสอบผลลัพธ์
print(data2[['InvoiceDate', 'Daytype']].drop_duplicates().head())
```

```
           InvoiceDate  Daytype
0  2009-12-01 07:45:00  Weekday
8  2009-12-01 07:46:00  Weekday
12 2009-12-01 09:06:00  Weekday
31 2009-12-01 09:08:00  Weekday
54 2009-12-01 09:24:00  Weekday
```

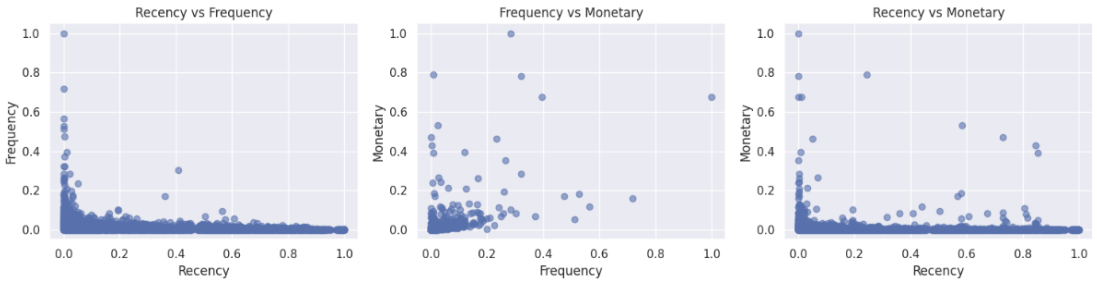
ภาพประกอบ 20 การสร้างตัวแปร Day Type

3.4 Modeling (การสร้างแบบจำลอง)

การสร้างแบบจำลองในการวิจัยนี้ใช้กระบวนการวิเคราะห์ข้อมูลด้วย Clustering Algorithms ได้แก่ K-Means, BIRCH โดยเริ่มจากการปรับข้อมูลด้วย Feature Engineering ผ่านโมเดล RFM (Recency, Frequency, Monetary) และ CAD (Country , AVG Basket Price , Day Type) เพื่อสร้างคุณลักษณะที่สะท้อนพฤติกรรมการซื้อของลูกค้า จากนั้นนำข้อมูลมา Normalize ก่อนที่จะเข้าสู่กระบวนการจัดกลุ่ม (Clustering) เพื่อค้นหาลักษณะหรือพฤติกรรมที่คล้ายคลึงกันระหว่างกลุ่มลูกค้า

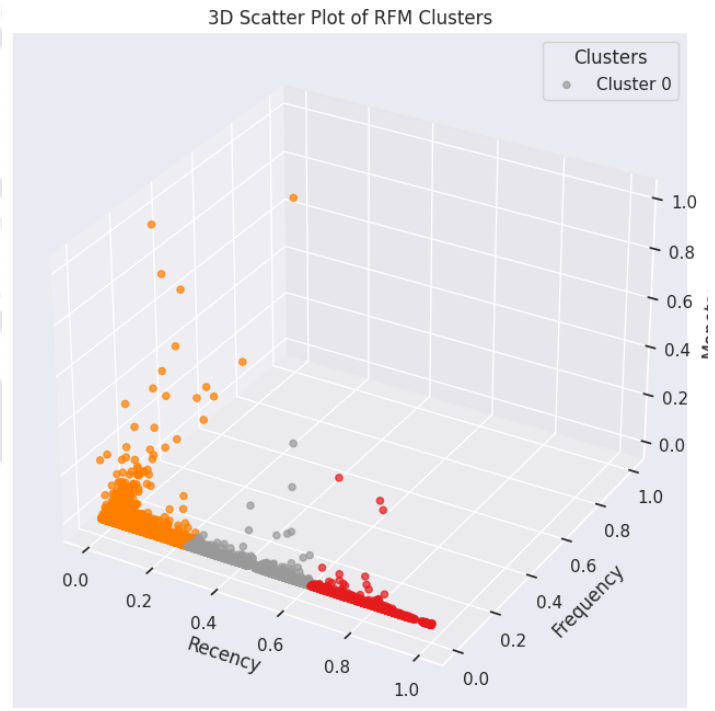
ตาราง 3 เปรียบเทียบ Clustering Algorithm 3 ประเภท (K-Means , BIRCH)

คุณลักษณะ	K-Means	BIRCH
วิธีการทำงาน	แบ่งกลุ่มโดยพิจารณาจากจุดศูนย์กลางของแต่ละกลุ่มและลดระยะห่างภายในกลุ่ม	ใช้โครงสร้างต้นไม้ (Tree Structure) เพื่อจัดการข้อมูลแบบลำดับชั้น
ข้อมูลที่ต้องการ	ต้องกำหนดจำนวนกลุ่มล่วงหน้า (k)	ไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า เหมาะกับข้อมูลขนาดใหญ่
การจัดการ Outliers	อ่อนไหวต่อ Outliers	มีความสามารถจัดการ Outliers ได้ดี
ข้อดี	เข้าใจง่ายและคำนวณเร็ว	มีประสิทธิภาพสูงกับข้อมูลขนาดใหญ่
	ใช้งานได้ดีกับข้อมูลขนาดเล็กถึงกลางที่เป็นทรงกลม	ใช้หน่วยความจำน้อย รองรับข้อมูลที่มีการอัปเดต
ข้อจำกัด	อ่อนไหวต่อ Outliers และค่าเริ่มต้นของ Centroids	มีความซับซ้อนในโครงสร้างและการปรับแต่งพารามิเตอร์
	ใช้งานได้ไม่ดีกับกลุ่มที่มีรูปทรงซับซ้อน	มีข้อจำกัดเมื่อข้อมูลไม่ได้เรียงลำดับหรือมีโครงสร้างที่ชัดเจน
ประสิทธิภาพ (Scalability)	ทำงานได้ดีในข้อมูลขนาดเล็กถึงกลาง	ทำงานได้ดีมากในข้อมูลขนาดใหญ่
การใช้งานที่เหมาะสม	ข้อมูลที่เป็นทรงกลมและมีจำนวนกลุ่มชัดเจน การแบ่งกลุ่มลูกค้า	ข้อมูลขนาดใหญ่ ข้อมูลที่มีโครงสร้างลำดับชั้น



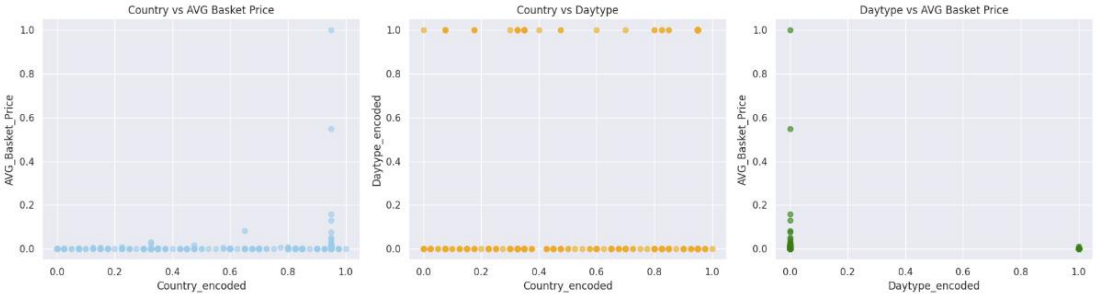
ภาพประกอบ 21 การกระจายตัวของข้อมูล RFM (Recency, Frequency, Monetary)

ก่อนนำข้อมูลไปจัดกลุ่มด้วยอัลกอริธึม Clustering ได้มีการสำรวจลักษณะการกระจายตัวของตัวแปร RFM (Recency, Frequency, Monetary) ผ่านกราฟแบบ 2 มิติ โดยข้อมูลถูกปรับให้อยู่ในช่วงเดียวกันด้วยวิธี Normalize



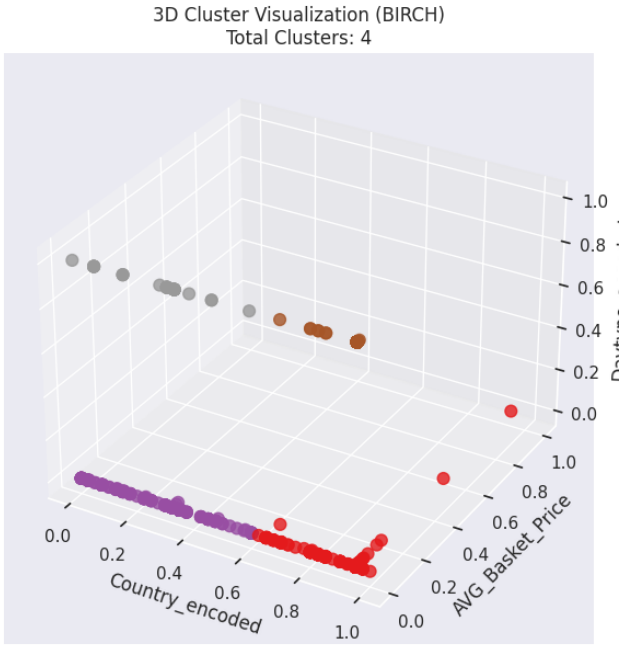
ภาพประกอบ 22 การจัดกลุ่มลูกค้าแบบสามมิติด้วย K-Means (RFM)

จากภาพแสดงการจัดกลุ่มลูกค้าโดยใช้ K-Means Clustering กับข้อมูล RFM (Recency, Frequency, Monetary) ซึ่งผ่านการทำ Min-Max Scaling แล้ว โดยแสดงผลแบบสามมิติ โดยแต่ละจุดแทนลูกค้า 1 ราย และแต่ละสีแสดงถึงกลุ่มที่แตกต่างกัน ผลลัพธ์แสดงให้เห็นว่าลูกค้าที่มีพฤติกรรมคล้ายกัน (เช่น ซื้อบ่อย ใช้จ่ายสูง) ถูกจัดอยู่ในกลุ่มเดียวกันอย่างชัดเจน ซึ่งช่วยให้สามารถนำไปใช้วางกลยุทธ์การตลาดแบบเฉพาะกลุ่มได้



ภาพประกอบ 23 การกระจายตัวของข้อมูล CAD (Country, AVG Basket Price, Daytype)

จากการสำรวจลักษณะการกระจายตัวของข้อมูลฟีเจอร์ Country, AVG Basket Price และ Day type ผ่านการเข้ารหัสและการทำ Normalize พบว่าข้อมูลมีแนวโน้มกระจุกตัวในช่วงค่าต่ำของแต่ละตัวแปร และมีความหนาแน่นของข้อมูลสูงในบางจุด ขณะที่จุดกระจายออก (outliers) มีจำนวนน้อย



ภาพประกอบ 24 การจัดกลุ่มลูกค้าแบบสามมิติด้วย BIRCH (CAD)

จากภาพแสดงการจัดกลุ่มลูกค้าในรูปแบบสามมิติด้วยอัลกอริธึม BIRCH โดยใช้ฟีเจอร์ Country (หลังการเข้ารหัส), ค่าเฉลี่ยต่อคำสั่งซื้อ (AVG_Basket_Price) และประเภทวัน (Daytype) ซึ่งผ่านการทำ Min-Max Scaling แล้ว ผลลัพธ์แสดงให้เห็นการกระจายตัวของลูกค้าใน 4 กลุ่ม โดยลูกค้าในแต่ละกลุ่มมีลักษณะพฤติกรรมการซื้อที่คล้ายกัน เช่น กลุ่มที่มีการใช้จ่าย

เฉลี่ยสูงหรือซื้อสินค้าเฉพาะในวันหยุด ซึ่งสามารถนำไปใช้ในการกำหนดกลยุทธ์ทางการตลาดแบบเฉพาะกลุ่มได้อย่างมีประสิทธิภาพ

3.5 Model Evaluation (การประเมินประสิทธิภาพ)

การประเมินผลโมเดลในงานวิจัยนี้จะใช้ Clustering Metrics หลายประเภทเพื่อตรวจสอบประสิทธิภาพและความแม่นยำของการจัดกลุ่มลูกค้า โดยใช้เกณฑ์ดังนี้:

Silhouette Coefficient (SC): วัดประสิทธิภาพการจัดกลุ่มโดยพิจารณาความหนาแน่นของข้อมูลในแต่ละกลุ่มและความห่างระหว่างกลุ่ม ค่าสูงบ่งชี้ว่ากลุ่มมีความแตกต่างกันชัดเจน ซึ่งช่วยให้สามารถประเมินได้ว่าการจัดกลุ่มที่เกิดขึ้นสามารถแยกข้อมูลแต่ละกลุ่มได้ดีเพียงใด

Davies-Bouldin Index (DBI): จะวัดว่าแต่ละกลุ่มมีความคล้ายคลึงกันมากน้อยเพียงใด หาก DBI มีค่าต่ำหมายความว่าการจัดกลุ่มมีคุณภาพดี เพราะ กลุ่มมีการกระจายตัวภายในที่ต่ำและ มีการแยกกันได้ชัดเจน

Calinski-Harabasz Index (CHI) ประเมินคุณภาพของการจัดกลุ่ม (Clustering) โดยพิจารณาความกระชับของจุดข้อมูลภายในกลุ่มและการแยกจากกันของกลุ่ม ค่า CHI สูง แสดงว่าการจัดกลุ่มมีคุณภาพดี เพราะ กลุ่มแยกจากกันชัดเจนและแต่ละกลุ่มมีความกระชับ

ด้วยการใช้เกณฑ์ดังกล่าวนี้ในการวิจัยจะประเมินคุณภาพของการจัดกลุ่มเพื่อเลือกโมเดลที่มีประสิทธิภาพสูงสุดในการวิเคราะห์ข้อมูลลูกค้า E-Commerce

3.6 การวิเคราะห์ความสัมพันธ์ระหว่างผลการจัดกลุ่ม

ในการศึกษานี้ได้ทำการจัดกลุ่มลูกค้าโดยใช้เทคนิคการจัดกลุ่มที่แตกต่างกัน ได้แก่ K-Means และ BIRCH โดยใช้ชุดข้อมูลคุณลักษณะที่แตกต่างกันสองชุด เพื่อวิเคราะห์และทำความเข้าใจพฤติกรรมของลูกค้าในมุมมองที่หลากหลายด้วยค่า Adjusted Rand Index (ARI)

Adjusted Rand Index (ARI) เป็นตัวชี้วัดที่ใช้ในการประเมินความเหมือนหรือความสอดคล้องกันระหว่างผลการจัดกลุ่ม (Clustering) สองชุด โดยพิจารณาว่าคู่ของตัวอย่าง (pair of data points) ในชุดข้อมูลถูกจัดให้อยู่ในกลุ่มเดียวกันหรือไม่ในแต่ละการจัดกลุ่ม ซึ่ง ARI ได้รับการพัฒนามาจาก Rand Index (RI) โดยเพิ่มการปรับค่าตามโอกาสที่อาจเกิดขึ้นโดยบังเอิญ

บทที่ 4
ผลการศึกษา

การแบ่งกลุ่มลูกค้าใน E-Commerce ด้วย Clustering Algorithms ได้ดำเนินการตามลำดับขั้นตอนต่างๆ เพื่อให้บรรลุวัตถุประสงค์ที่กำหนดไว้ดังนี้

1. ผลลัพธ์การจัดกลุ่มลูกค้าด้วย K-means Clustering โดยใช้ RFM
2. ผลลัพธ์การจัดกลุ่มลูกค้าด้วย BIRCH Clustering โดยใช้ RFM
3. ผลลัพธ์การจัดกลุ่มลูกค้าด้วย K-means Clustering โดยใช้ CAD
4. ผลลัพธ์การจัดกลุ่มลูกค้าด้วย BIRCH Clustering โดยใช้ CAD
5. การเปรียบเทียบผลการจัดกลุ่มระหว่าง RFM และ CAD ด้วย K-Means และ BIRCH

4.1 ผลลัพธ์การจัดกลุ่มลูกค้าด้วย K-means Algorithm โดยใช้ RFM

แบบจำลองการแบ่งกลุ่มด้วย RFM โดยใช้ข้อมูลลูกค้า ตามภาพประกอบที่ 21 และต่อไปจะเป็นการหาค่า K ที่ดีที่สุดจากการทำ Elbow Method ตามภาพประกอบที่ 22 และจะนำมาทำการแบ่งกลุ่มโดยใช้ K-means Clustering ตามภาพประกอบที่ 23 จากนั้นจะมาประเมินผลด้วย Silhouette , Calinski-Harabasz Index , Davies-Bouldin Index ตามภาพประกอบที่ 24

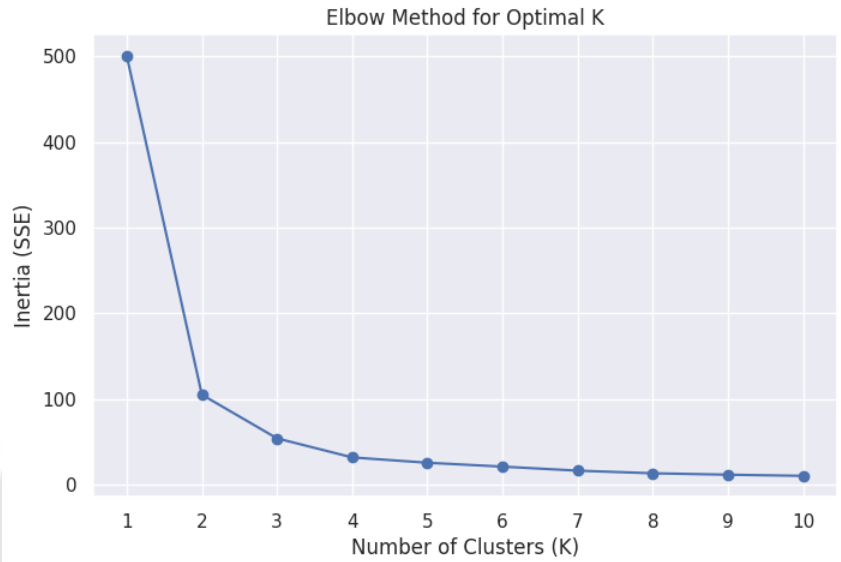
	Customer ID	Recency	Frequency	Monetary
0	12346.0	0.440379	0.031558	0.116584
1	12347.0	0.001355	0.013807	0.007955
2	12348.0	0.100271	0.007890	0.002849
3	12349.0	0.024390	0.007890	0.006707
4	12350.0	0.418699	0.000000	0.000468
...	...	...	...	...
5937	18283.0	0.004065	0.041420	0.003862
5938	18284.0	0.581301	0.001972	0.001118
5939	18285.0	0.894309	0.000000	0.000599
5940	18286.0	0.644986	0.003945	0.002208
5941	18287.0	0.056911	0.013807	0.005954

5942 rows × 4 columns

ภาพประกอบ 25 แสดงข้อมูล RFM ที่นำมาใช้ในการในการวิเคราะห์

1. ผลวิเคราะห์จำนวนกลุ่มที่เหมาะสมด้วย Elbow Method และ K-means Algorithm

จากการวิเคราะห์ Elbow Method พบว่าจำนวนกลุ่มที่เหมาะสมคือ 3 กลุ่ม เนื่องจากเป็นจุดที่เกิดการหักศอก ตามภาพประกอบที่ 22 และภาพประกอบที่ 23 แสดงผลลัพธ์ของการจัดกลุ่มโดยการใช้ K-Means Algorithm



ภาพประกอบ 26 แสดงจำนวนกลุ่ม RFM ที่เหมาะสมด้วย Elbow Method

RFM K-Means Clustering Full Data (Normalized):

	Customer ID	Recency	Frequency	Monetary	Cluster
0	12346.0	0.440379	0.031558	0.116584	2
1	12347.0	0.001355	0.013807	0.007955	1
2	12348.0	0.100271	0.007890	0.002849	1
3	12349.0	0.024390	0.007890	0.006707	1
4	12350.0	0.418699	0.000000	0.000468	2
...	...	...	...	...	...
5937	18283.0	0.004065	0.041420	0.003862	1
5938	18284.0	0.581301	0.001972	0.001118	2
5939	18285.0	0.894309	0.000000	0.000599	0
5940	18286.0	0.644986	0.003945	0.002208	0
5941	18287.0	0.056911	0.013807	0.005954	1

5942 rows × 5 columns

ภาพประกอบ 27 แสดงผลลัพธ์ RFM ที่ผ่าน K-means Algorithm

2. การประเมินประสิทธิภาพ RFM โดยใช้ K-means Algorithm

ผลลัพธ์ในการประเมินประสิทธิภาพ คะแนน Silhouette Score อยู่ที่ 0.679 คะแนน Calinski-Harabasz Index อยู่ที่ 24535.95 คะแนน Davies-Bouldin Index อยู่ที่ 0.49 ตามภาพประกอบที่ 24

	Metric	Value
0	Silhouette Score	0.678998
1	Calinski-Harabasz Index	24535.945273
2	Davies-Bouldin Index	0.490270

ภาพประกอบ 28 แสดงการประเมินประสิทธิภาพ RFM โดยใช้ K-means Algorithm

4.2 ผลลัพธ์การจัดกลุ่มลูกค้าด้วย BIRCH Clustering โดยใช้ RFM

แบบจำลองการแบ่งกลุ่มด้วย RFM โดยใช้ข้อมูลลูกค้า ตามภาพประกอบที่ 21 หลังจากนั้นจะนำข้อมูลผ่าน BIRCH Algorithm ซึ่งแบ่งกลุ่มได้ 4 กลุ่ม ตามภาพประกอบที่ 25 และจะประเมินประสิทธิภาพโดยผลลัพธ์ในการประเมิน คะแนน Silhouette Score อยู่ที่ 0.677 คะแนน Calinski-Harabasz Index อยู่ที่ 24520.84 คะแนน Davies-Bouldin Index อยู่ที่ 0.50 ตามภาพประกอบที่ 26

📌 BIRCH Clustering พบทั้งหมด 4 กลุ่ม
RFM BIRCH Clustering Full Data:

	Recency	Frequency	Monetary	Cluster
0	0.440379	0.031558	0.116584	0
1	0.001355	0.013807	0.007955	1
2	0.100271	0.007890	0.002849	1
3	0.024390	0.007890	0.006707	1
4	0.418699	0.000000	0.000468	0
...	...	...	...	...
5937	0.004065	0.041420	0.003862	1
5938	0.581301	0.001972	0.001118	0
5939	0.894309	0.000000	0.000599	2
5940	0.644986	0.003945	0.002208	2
5941	0.056911	0.013807	0.005954	1

5942 rows × 4 columns

ภาพประกอบ 29 แสดงผลลัพธ์ RFM ที่ผ่าน BIRCH Algorithm

	Metric	Value
0	Silhouette Score	0.677281
1	Calinski-Harabasz Index	24520.843409
2	Davies-Bouldin Index	0.495538

ภาพประกอบ 30 แสดงการประเมินประสิทธิภาพ RFM โดยใช้ BIRCH Algorithm

4.3 ผลลัพธ์การจัดกลุ่มลูกค้าด้วย K-means Clustering โดยใช้ CAD

แบบจำลองการแบ่งกลุ่มด้วย RFM โดยใช้ข้อมูลลูกค้า ตามภาพประกอบที่ 27 และต่อไปจะเป็นการหาค่า K ที่ดีที่สุดจากการทำ Elbow Method ตามภาพประกอบที่ 28 และจะนำมาทำการแบ่งกลุ่มโดยใช้ K-means Clustering ตามภาพประกอบที่ 29 จากนั้นจะมาประเมินผลด้วย Silhouette , Calinski-Harabasz Index , Davies-Bouldin Index ตามภาพประกอบที่ 30

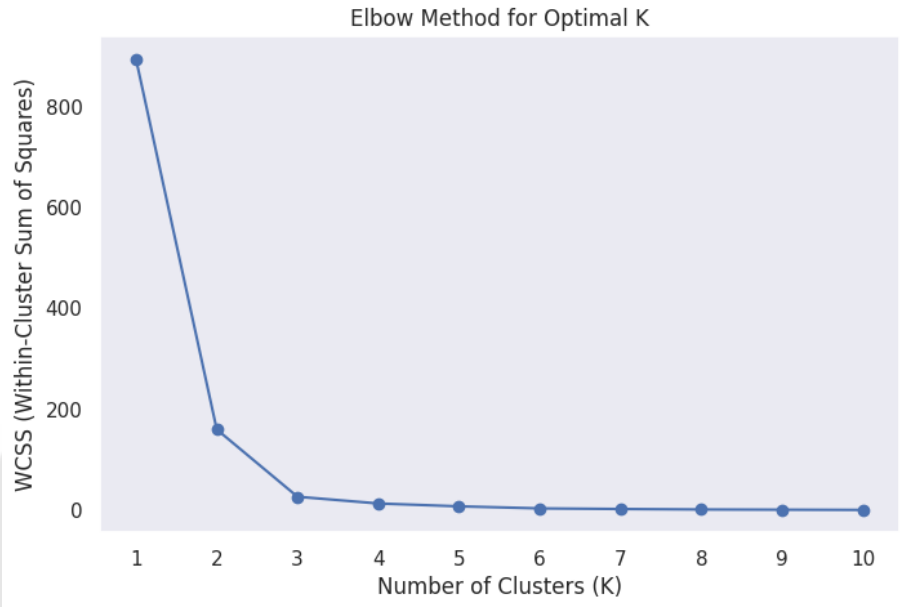
	Customer ID	Country_encoded	AVG_Basket_Price	Daytype
0	13085.0	0.950	0.000440	Weekday
2	13078.0	0.950	0.000421	Weekday
3	15362.0	0.950	0.000166	Weekday
4	18102.0	0.950	0.007634	Weekday
5	12682.0	0.325	0.000273	Weekday
...	...	...	...	...
36760	15195.0	0.950	0.045834	Weekday
36835	13436.0	0.950	0.000176	Weekday
36838	15520.0	0.950	0.000409	Weekday
36872	13298.0	0.950	0.001066	Weekday
36965	12713.0	0.350	0.000233	Weekday

5881 rows × 4 columns

ภาพประกอบ 31แสดงข้อมูล CAD ที่นำมาใช้ในการในการวิเคราะห์

1. ผลวิเคราะห์จำนวนกลุ่มที่เหมาะสมด้วย Elbow Method และ K-means Algorithm

จากการวิเคราะห์ Elbow Method พบว่าจำนวนกลุ่มที่เหมาะสมคือ 3 กลุ่ม เนื่องจากเป็นจุดที่เกิดการหักศอก ตามภาพประกอบที่ 28 และภาพประกอบที่ 29 แสดงผลลัพธ์ของการจัดกลุ่มโดยการใช้ K-Means Algorithm



ภาพประกอบ 32แสดงจำนวนกลุ่ม CAD ที่เหมาะสมด้วย Elbow Method

📌 K-Means Clustering พบทั้งหมด 3 กลุ่ม
CAD K-Means Clustering Full Data:

🔷 ข้อมูลหลังจัดกลุ่ม:

	Customer ID	Cluster	Country_encoded	AVG_Basket_Price	Daytype_encoded
0	13085.0	0	0.950	0.000440	0
2	13078.0	0	0.950	0.000421	0
3	15362.0	0	0.950	0.000166	0
4	18102.0	0	0.950	0.007634	0
5	12682.0	2	0.325	0.000273	0
...	...	...	...	...	...
36760	15195.0	0	0.950	0.045834	0
36835	13436.0	0	0.950	0.000176	0
36838	15520.0	0	0.950	0.000409	0
36872	13298.0	0	0.950	0.001066	0
36965	12713.0	2	0.350	0.000233	0

5881 rows × 5 columns

ภาพประกอบ 33 แสดงผลลัพธ์ CAD ที่ผ่าน K-means Algorithm

2. การประเมินประสิทธิภาพ RFM โดยใช้ K-means Algorithm

ผลลัพธ์ในการประเมินประสิทธิภาพ คะแนน Silhouette Score อยู่ที่ 0.961 คะแนน Calinski-Harabasz Index อยู่ที่ 93293.74 คะแนน Davies-Bouldin Index อยู่ที่ 0.164 ตามภาพประกอบที่ 30

	Metric	Value
0	Silhouette Score	0.961330
1	Calinski-Harabasz Index	93293.743824
2	Davies-Bouldin Index	0.163960

ภาพประกอบ 34 แสดงการประเมินประสิทธิภาพ CAD โดยใช้ K-means Algorithm

4.4 ผลลัพธ์การจัดกลุ่มลูกค้าด้วย BIRCH Clustering โดยใช้ CAD

แบบจำลองการแบ่งกลุ่มด้วย CAD โดยใช้ข้อมูลลูกค้า ตามภาพประกอบที่ 27 หลังจากนั้นจะนำข้อมูลผ่าน BIRCH Algorithm ซึ่งแบ่งกลุ่มได้ 4 กลุ่ม ตามภาพประกอบที่ 31 และจะประเมินประสิทธิภาพโดยผลลัพธ์ในการประเมิน คะแนน Silhouette Score อยู่ที่ 0.967 คะแนน Calinski-Harabasz Index อยู่ที่ 123831.82 คะแนน Davies-Bouldin Index อยู่ที่ 0.19 ตามภาพประกอบที่ 32

◆ จำนวนคลัสเตอร์ที่ได้จาก BIRCH: 4

◆ ข้อมูลหลังใช้ BIRCH Clustering (Full Data):

	Customer ID	Cluster_BIRCH	Country_encoded	AVG_Basket_Price	Daytype_encoded
0	13085.0	0	0.950	0.000440	0
2	13078.0	0	0.950	0.000421	0
3	15362.0	0	0.950	0.000166	0
4	18102.0	0	0.950	0.007634	0
5	12682.0	1	0.325	0.000273	0
...	...	...	...	...	...
36760	15195.0	0	0.950	0.045834	0
36835	13436.0	0	0.950	0.000176	0
36838	15520.0	0	0.950	0.000409	0
36872	13298.0	0	0.950	0.001066	0
36965	12713.0	1	0.350	0.000233	0

5881 rows × 5 columns

ภาพประกอบ 35 แสดงผลลัพธ์ CAD ที่ผ่าน BIRCH Algorithm

	Metric	Value
0	Silhouette Score	0.966645
1	Calinski-Harabasz Index	123831.822030
2	Davies-Bouldin Index	0.192343

ภาพประกอบ 36 แสดงการประเมินประสิทธิภาพ CAD โดยใช้ BIRCH Algorithm

4.5 การเปรียบเทียบผลการจัดกลุ่มระหว่าง RFM และ CAD ด้วย K-Means และ BIRCH

Method Metric	Eva_RFM_KMeans	Eva_RFM_Birch	Eva_CAD_KMeans	Eva_CAD_Birch
Silhouette Score	0.678998	0.677281	0.961330	0.966645
Calinski-Harabasz Index	24535.945273	24520.843409	93293.743824	123831.822030
Davies-Bouldin Index	0.490270	0.495538	0.163960	0.192343

ภาพประกอบ 37 ผลลัพธ์การประเมินการจัดกลุ่ม RFM และ CAD ด้วย K-Means และ BIRCH

จากภาพประกอบที่ 33 แสดงให้เห็นว่าการใช้ข้อมูล CAD ผ่าน BIRCH Algorithm มีประสิทธิภาพสูงสุดในการแบ่งกลุ่มข้อมูลเมื่อพิจารณาจากตัวชี้วัดทั้งหมด โดยมี Calinski-Harabasz Index สูงที่สุดที่ 123,831.82 แสดงถึงโครงสร้างกลุ่มที่กะทัดรัดและแยกจากกันได้ดี นอกจากนี้ Davies-Bouldin Index ที่ต่ำสุดที่ 0.1923 บ่งบอกว่ากลุ่มมีความแตกต่างกันมากและมีความแน่นภายในกลุ่มสูง อีกทั้ง Silhouette Score อยู่ที่ 0.9666 ซึ่งเป็นค่าที่สูงที่สุด แสดงให้เห็นว่าจุดข้อมูลแต่ละจุดถูกจัดกลุ่มได้อย่างเหมาะสมและมีความแตกต่างจากกลุ่มอื่นอย่างชัดเจน

การใช้ข้อมูล CAD ผ่าน K-Means ให้ผลลัพธ์ที่ดีมากในการจัดกลุ่มข้อมูล โดยมี Silhouette Score เท่ากับ 0.9613 ซึ่งแสดงให้เห็นว่าข้อมูลแต่ละจุดถูกจัดอยู่ในคลัสเตอร์ที่เหมาะสมและมีการแยกตัวที่ชัดเจน Calinski-Harabasz Index มีค่า 93,293.74 ซึ่งสูงกว่ากลุ่มที่ใช้ RFM อย่างมีนัยสำคัญ บ่งบอกว่าคลัสเตอร์มีความกะทัดรัดและแยกออกจากกันได้ดี นอกจากนี้ Davies-Bouldin Index มีค่าต่ำที่สุดที่ 0.1639 แสดงว่าคลัสเตอร์มีความแตกต่างกันอย่างชัดเจนมากกว่าวิธีอื่น ๆ โดยรวมแล้ว การใช้ข้อมูล CAD ผ่าน K-Means มีประสิทธิภาพสูงในการจัดกลุ่มข้อมูลลูกค้า และให้ผลลัพธ์ที่ดีมากเมื่อใช้ข้อมูล CAD แม้ว่า การใช้ข้อมูล CAD ผ่าน BIRCH จะให้ผลลัพธ์ที่ดีกว่าเล็กน้อย แต่ K-Means ยังคงเป็นทางเลือกที่มีประสิทธิภาพและสามารถใช้งานได้ดี

การใช้ข้อมูล RFM ผ่าน BIRCH ให้ผลลัพธ์ในการจัดกลุ่มที่ใกล้เคียงกับ การใช้ข้อมูล RFM ผ่าน K-Means แต่มีประสิทธิภาพต่ำกว่าการใช้ CAD โดยมี Silhouette Score เท่ากับ 0.6773 ซึ่งบ่งชี้ว่าคลัสเตอร์มีการแยกตัวในระดับปานกลาง Calinski-Harabasz Index อยู่ที่ 24,520.84 ซึ่งต่ำกว่ากลุ่มที่ใช้ CAD อย่างมาก แสดงว่าคลัสเตอร์มีความกะทัดรัดและสามารถแยกออกจากกันได้ในระดับหนึ่ง แต่ไม่เด่นชัดเท่า CAD นอกจากนี้ Davies-Bouldin Index มีค่า 0.4955 ซึ่งสูงกว่าวิธีอื่น ๆ ทั้งหมด แสดงว่าคลัสเตอร์มีความคล้ายคลึงกันมากกว่าหมวดหมู่ CAD และอาจมีความซ้อนทับของข้อมูลสูงกว่า โดยรวมแล้ว การใช้ข้อมูล RFM ผ่าน BIRCH เป็นตัวเลือกที่พอใช้ได้ในการจัดกลุ่มข้อมูล RFM แต่มีประสิทธิภาพน้อยกว่าทางเลือกอื่น โดยเฉพาะเมื่อเปรียบเทียบกับการใช้ CAD

จากผลการประเมินประสิทธิภาพการจัดกลุ่ม พบว่า CAD feature ที่ประกอบด้วย Country, AVG Basket Price และ Day type เมื่อใช้ร่วมกับ BIRCH Clustering ให้ค่าการประเมินที่ดีที่สุดในทุกตัวชี้วัด โดยเฉพาะ Silhouette Score และ Calinski-Harabasz Index ที่มีค่าสูงกว่าวิธีอื่นอย่างชัดเจน สะท้อนให้เห็นว่า BIRCH สามารถจับกลุ่มลูกค้าตามลักษณะเฉพาะของฟีเจอร์ CAD ได้อย่างมีประสิทธิภาพ ทั้งนี้อาจเนื่องจากลักษณะของข้อมูลที่มีความเป็นกลุ่มอยู่ชัดเจน โดยเฉพาะฟีเจอร์เชิงหมวดหมู่ที่มีค่าเพียง 0 และ 1 ซึ่งเหมาะสมกับการแบ่งกลุ่มแบบลำดับชั้นตามหลักการของ BIRCH

ในขณะที่การใช้ข้อมูล RFM โดยผ่าน K-Means Algorithm นั้นให้ผลลัพธ์ที่ดีกว่าวิธี BIRCH เล็กน้อย โดยมี Silhouette Score เท่ากับ 0.6789 ซึ่งแสดงให้เห็นว่าข้อมูลแต่ละจุดอยู่ในกลุ่มที่เหมาะสมในระดับปานกลาง ขณะที่ Calinski-Harabasz Index มีค่า 24,535.94 บ่งบอกว่ากลุ่มมีการกระจายตัวที่ค่อนข้างดีเมื่อเทียบกับ BIRCH แต่ยังต่ำกว่ากลุ่มที่ใช้ข้อมูล CAD อย่างมาก นอกจากนี้ Davies-Bouldin Index อยู่ที่ 0.4902 ซึ่งดีกว่า BIRCH เล็กน้อย หมายความว่ากลุ่มมีความแตกต่างกันในระดับหนึ่ง แต่ยังไม่ดีเท่าการใช้ CAD สรุปได้ว่าการใช้ข้อมูล RFM โดยผ่าน K-Means Algorithm เหมาะกับการแบ่งกลุ่มข้อมูลลูกค้าโดยใช้ข้อมูล RFM แต่ยังไม่ใช่วิธีที่ดีที่สุดเมื่อเปรียบเทียบกับการใช้ข้อมูล CAD ผ่าน BIRCH Algorithm

บทที่ 5
สรุปผลการวิจัย อภิปรายผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการวิเคราะห์การแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้าจากการขายสินค้าออนไลน์ในสหราชอาณาจักร โดยเก็บข้อมูลตั้งแต่ 1 ธันวาคม 2009 จนถึง 12 กันยายน 2011 มีข้อมูลทั้งหมด 1,067,371 ด้วยแบบจำลอง RFM (Recency, Frequency, Monetary) กับ แบบจำลอง CAD (Country , AVG Basket Price , Day Type) โดยใช้เทคนิค K-Means และ BIRCH ในการทำ Clustering Algorithm โดยใช้ Elbow Method ในการกำหนดจำนวนกลุ่มใน K-means และให้ BIRCH กำหนดจำนวนกลุ่มเองได้ผลลัพธ์ดังต่อไปนี้

จากผลการประเมิน พบว่าเทคนิค K-Means ให้ผลลัพธ์ที่ดีที่สุดเมื่อใช้กับข้อมูล RFM โดยมีค่า Silhouette Score เท่ากับ 0.67และ Calinski-Harabasz Index สูงถึง 24,535.95 ซึ่งบ่งชี้ว่าการกระจายตัวของกลุ่มมีความเหมาะสม และกลุ่มที่ได้มีความชัดเจน ในขณะที่ ค่า Davies-Bouldin Index อยู่ที่ 0.49 แสดงถึงการทับซ้อนระหว่างกลุ่มในระดับต่ำ ส่งผลให้สามารถแยกแยะกลุ่มลูกค้าได้อย่างมีประสิทธิภาพ

สำหรับการจัดกลุ่มโดยใช้ BIRCH ให้ผลลัพธ์ที่ดีที่สุดเมื่อใช้กับข้อมูล CAD พบว่าผลลัพธ์อยู่ในระดับที่ดี โดยมีค่า Silhouette Score เท่ากับ 0.966 และค่า Davies-Bouldin Index ต่ำเพียง 0.192 แสดงถึงความชัดเจนของกลุ่มและการแยกแยะกลุ่มที่มีประสิทธิภาพ กลุ่มลูกค้ามีลักษณะแตกต่างกันชัดเจนตามฟีเจอร์ประเทศ ประเภทวัน และราคาต่อคำสั่งซื้อเฉลี่ย เทคนิคการจัดกลุ่ม BIRCH ให้ผลลัพธ์ที่ดีที่สุดเมื่อใช้กับ CAD เนื่องจากสามารถสร้างกลุ่มที่มีความกะทัดรัดและชัดเจนกว่า แม้ว่า K-Means จะมีค่า Davies-Bouldin Index ที่ต่ำกว่าเล็กน้อย แต่โดยรวมแล้ว BIRCH ทำให้กลุ่มมีความคงตัวและมีการแยกกลุ่มที่ดีกว่า

หากต้องการแบ่งกลุ่มลูกค้าโดยอ้างอิงจากข้อมูลเชิงลึกของลูกค้า การใช้ข้อมูล CAD ผ่าน BIRCH เป็นตัวเลือกที่ดีที่สุดในแง่ของคุณภาพการจัดกลุ่ม แต่หากจะวิเคราะห์พฤติกรรมลูกค้าโดยใช้ข้อมูล RFM การใช้ K-Means Clustering จะให้ผลลัพธ์ในการแบ่งกลุ่มดีกว่า BIRCH

1. ข้อมูล CAD (Country, AVG Basket Size, Day Type) ผ่านการใช้วิธี BIRCH Clustering มีผลลัพธ์โดย BIRCH เลือกจำนวนกลุ่มได้ 4 กลุ่ม คือ

ตาราง 4 สรุปลักษณะกลุ่มลูกค้าที่ได้จาก BIRCH Clustering (ฟีเจอร์ CAD)

กลุ่มที่	จำนวนลูกค้า	AVG Basket Price	Day Type	Country (Top3)
1	4674	76.75	Weekday	1. United Kingdom (4,533)
				2. Spain (37)
				3. Switzerland (19)
2	349	62.24	Weekday	1. France (86)
				2. Italy (17)
				3. Australia (13)
3	826	23.92	Weekend	1. United Kingdom (819)
				2. Sweden (2)
				3. Switzerland (2)
4	32	35.49	Weekend	1. France (7)
				2. Japan (2)
				3. Australia (1)

กลุ่มที่ 1 มีลูกค้าจำนวน 4674 คน มี AVG Basket Price อยู่ที่ 76.75 และลูกค้าทั้งหมดทำการซื้อสินค้าในวันธรรมดา (Weekday) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ UK จำนวน 4533 คน อันดับ 2 คือ Spain จำนวน 37 คน และ Switzerland จำนวน 19 คน

กลุ่มที่ 2 มีลูกค้าจำนวน 349 คน มี AVG Basket Price อยู่ที่ 62.24 และลูกค้าทั้งหมดทำการซื้อสินค้าในวันธรรมดา (Weekday) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ France จำนวน 86 คน Italy เป็นอันดับ 2 ที่ 17 คน และอันที่ 3 คือ Australia จำนวน 13 คน

กลุ่มที่ 3 มีลูกค้าจำนวน 826 คน มี AVG Basket Price อยู่ที่ 23.92 และลูกค้าทั้งหมดทำการซื้อสินค้าในวันหยุด (Weekend) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ United Kingdom จำนวน 819 คน อันดับ 2 คือ Sweden จำนวน 2 คน และ Switzerland จำนวน 2 คน

กลุ่มที่ 4 มีลูกค้าจำนวน 32 คน มี AVG Basket Price อยู่ที่ 35.49 และลูกค้าทั้งหมดทำการซื้อสินค้าในวันหยุด (Weekend) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ France จำนวน 7 คน อันดับ 2 คือ Japan จำนวน 2 คน และ Australia จำนวน 1 คน

2. ข้อมูล RFM (Recency, Frequency, Monetary) ผ่านการใช้วิธี K-Means Clustering มีโดยแบ่งลูกค้าออกเป็น 3 กลุ่มดังต่อไปนี้

ตาราง 5 สรุปลักษณะกลุ่มลูกค้าที่ได้จาก K-Means Clustering (ฟีเจอร์ RFM)

กลุ่มที่	จำนวนลูกค้า	Recency (AVG)	Frequency (AVG)	Monetary (AVG)
1	808	601.26	1.88	2386
2	3621	51.57	10.45	5301.28
3	1513	348.43	3.48	1741.83

กลุ่มที่ 1 มีจำนวนลูกค้า 808 คน ค่า Recency เฉลี่ยคือ 601.26 ค่า Frequency เฉลี่ยคือ 1.88 Monetary เฉลี่ยเท่ากับ 2386.00

กลุ่มที่ 2 มีจำนวนลูกค้า 3621 คน ค่า Recency เฉลี่ยคือ 51.57 ค่า Frequency เฉลี่ยคือ 10.45 Monetary เฉลี่ยเท่ากับ 5301.28

กลุ่มที่ 3 มีจำนวนลูกค้า 1513 คน ค่า Recency เฉลี่ยคือ 348.43 ค่า Frequency เฉลี่ยคือ 3.48 Monetary เฉลี่ยเท่ากับ 1741.83

5.2 สรุปและอภิปรายผล

งานวิจัยนี้มุ่งเน้นการวิเคราะห์และเปรียบเทียบประสิทธิภาพของการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้าใน การขายสินค้าออนไลน์ในสหราชอาณาจักร โดยใช้ข้อมูลตั้งแต่ 1 ธันวาคม 2009 – 12 กันยายน 2011 รวม 1,067,371 รายการ การทดลองแบ่งเป็นสองแบบจำลอง ได้แก่ RFM Model และ CAD Model (Country, AVG Basket Price, Day Type) และใช้เทคนิค K-Means และ BIRCH ในการทำ Clustering

การกำหนดจำนวนกลุ่มใน K-Means ใช้วิธี Elbow Method เพื่อหาจุดที่เหมาะสมที่สุด ในขณะที่ BIRCH สามารถกำหนดจำนวนกลุ่มได้เองโดยอัตโนมัติ จากผลการทดลองพบว่า CAD Model โดยใช้ BIRCH ให้ผลลัพธ์ดีกว่า K-means และ RFM Model โดยการใช้ K-means ให้

ผลลัพธ์ที่ดีกว่า BIRCH ในแง่ของคุณภาพการจัดกลุ่ม ซึ่งพิจารณาจาก Silhouette Score และ Davies-Bouldin Index

ซึ่งการใช้ข้อมูล CAD ผ่าน BIRCH Clustering ได้ทำการแบ่งกลุ่มลูกค้าออกเป็น 4 กลุ่ม โดยมี Calinski-Harabasz Index สูงที่สุดที่ 123,831.82 แสดงถึงโครงสร้างกลุ่มที่กะทัดรัดและแยกจากกันได้ดี นอกจากนี้ Davies-Bouldin Index ที่ต่ำสุดที่ 0.1923 บ่งบอกว่ากลุ่มมีความแตกต่างกันมากและมีความแน่นภายในกลุ่มสูง"อีกทั้ง Silhouette Score อยู่ที่ 0.9666 ซึ่งเป็นค่าที่สูงที่สุด แสดงให้เห็นว่าจุดข้อมูลแต่ละจุดถูกจัดกลุ่มได้อย่างเหมาะสมและมีความแตกต่างจากกลุ่มอื่นอย่างชัดเจนโดยแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 มีลูกค้าจำนวน 4674 คน เป็นกลุ่มที่มีลูกค้าเยอะที่สุด มี AVG Basket Price อยู่ที่ 76.75 ซึ่งเป็นกลุ่มที่มีค่าเฉลี่ยการใช้จ่ายต่อคครั้งสูงที่สุด และลูกค้าทั้งหมดทำการซื้อสินค้าในวันธรรมดา (Weekday) ซึ่งแสดงให้เห็นว่าลูกค้าส่วนมากจะใช้จ่ายในวันธรรมดา อันดับ 1 สั่งซื้อสินค้าจากประเทศ United Kingdom จำนวน 4533 คน โดยสรุปเป็นกลุ่ม ลูกค้าในประเทศเยอะสุด ใช้จ่ายเยอะ ในวันธรรมดา

กลุ่มที่ 2 มีลูกค้าจำนวน 349 คน มีจำนวนลูกค้าเป็นอันดับ 3 มี AVG Basket Price อยู่ที่ 62.24 ซึ่งยังถือว่ามีการใช้จ่ายที่ค่อนข้างสูง และลูกค้าทั้งหมดทำการซื้อสินค้าในวันธรรมดา (Weekday) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ France จำนวน 86 คน Italy เป็นอันดับ 2 ที่ 17 คน และอันที่ 3 คือ Australia จำนวน 13 คน ซึ่งประเทศที่ลูกค้าอยู่นั้นคือต่างประเทศ โดยสรุปเป็นกลุ่ม ลูกค้าต่างประเทศ จ่ายค่อนข้างเยอะ ในวันธรรมดา

กลุ่มที่ 3 มีลูกค้าจำนวน 826 คน นับว่าเป็นอันดับที่ 2 โดย มี AVG Basket Price อยู่ที่ 23.92 ซึ่งใช้จ่ายน้อยที่สุด และลูกค้าทั้งหมดทำการซื้อสินค้าในวันหยุด (Weekend) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ United Kingdom จำนวน 819 คน อันดับ 2 คือ Sweden จำนวน 2 คน และ Switzerland จำนวน 2 คน โดยสรุปแล้วเป็นกลุ่ม ลูกค้าภายในประเทศที่ทำการซื้อสินค้าซื้อสินค้าในวันหยุด และใช้จ่ายต่อครั้งน้อยมาก

กลุ่มที่ 4 มีลูกค้าจำนวน 32 คน เป็นกลุ่มที่มีลูกค้าน้อยที่สุด แต่มี AVG Basket Price อยู่ที่ 35.49 เป็นอันดับที่ 3 มีการใช้จ่ายมากกว่ากลุ่มที่ 3 และลูกค้าทั้งหมดทำการซื้อสินค้าในวันหยุด (Weekend) โดยลูกค้า 3 อันดับแรกสั่งซื้อสินค้าจากประเทศ Franceจำนวน 7 คน อันดับ 2 คือ Japan จำนวน 2 คน และ Australia จำนวน 1 คน โดยสรุปแล้วเป็นกลุ่ม ลูกค้าต่างประเทศที่ใช้จ่ายปานกลางในวันหยุด

การใช้ข้อมูล RFM ผ่าน K-Means Clustering ให้ผลลัพธ์ที่ดีสุดโดยหาจำนวนกลุ่มผ่าน Elbow Method ได้จำนวนกลุ่มเท่ากับ 3 ได้คะแนนประเมินประสิทธิภาพกลุ่ม Calinski-Harabasz Index มีค่า 24,535.94 บ่งบอกว่ากลุ่มมีการกระจายตัวที่ค่อนข้างดีเมื่อเทียบกับ แต่ยังต่ำกว่ากลุ่มที่ใช้ข้อมูล CAD อย่างมาก นอกจากนี้ Davies-Bouldin Index อยู่ที่ 0.4902 หมายความว่ากลุ่มมีความแตกต่างกันในระดับหนึ่ง แต่ยังไม่ดีเท่าการใช้ CAD โดยมีรายละเอียดแต่ละกลุ่มดังนี้

กลุ่มที่ 1 มีจำนวนลูกค้า 808 คน ค่า Recency เฉลี่ยคือ 601.26 นับว่าเป็นระยะห่างที่ค่อนข้างมากที่ลูกค้ากลับมาซื้อสินค้าอีกครั้ง หรือไม่ได้กลับมาซื้อสินค้าเลย ค่า Frequency เฉลี่ยคือ 1.88 แสดงให้เห็นถึงแนวโน้มในการซื้อสินค้าซ้ำที่น้อยมากเมื่อเทียบกับกลุ่มอื่น ส่วนค่า Monetary เฉลี่ยเท่ากับ 2386.00 ถือว่ายังเป็นกลุ่มลูกค้าที่ใช้จ่ายปานกลาง โดยอาจจะเรียกกลุ่มนี้ว่า เป็นกลุ่มที่มีการใช้จ่ายปานกลาง มีโอกาสที่จะกลับมาซื้อน้อย ควรหากลยุทธ์ในการดึงดูดให้กลับมาใช้ซ้ำ

กลุ่มที่ 2 มีจำนวนลูกค้า 3621 คน ค่า Recency เฉลี่ยคือ 51.57 ซึ่งแสดงให้เห็นว่าลูกค้ากลุ่มนี้มักจะกลับมาซื้อสินค้าอีกครั้งในช่วงเวลาที่ผ่านมา ค่า Frequency เฉลี่ยคือ 10.45 หมายถึงลูกค้ากลับมาซื้อสินค้าที่บ่อยมากกว่ากลุ่มอื่น Monetary เฉลี่ยเท่ากับ 5301.28 เป็นกลุ่มที่มีการใช้จ่ายเฉลี่ยเยอะที่สุด โดยสามารถเรียกกลุ่มได้ว่าเป็นกลุ่มที่ มีกำลังทรัพย์ และกลับมาใช้ซ้ำบ่อย เป็นกลุ่มลูกค้าที่ควรรักษาไว้ให้ได้เนื่องจากมีจำนวนมากที่สุดด้วย

กลุ่มที่ 3 มีจำนวนลูกค้า 1513 คน ค่า Recency เฉลี่ยคือ 348.43 แสดงให้เห็นว่าลูกค้าทำการซื้อสินค้าที่น้อยลงแต่ยังคงมีกลับมาซื้อบ้าง ค่า Frequency เฉลี่ยคือ 3.48 แต่ยังเป็นกลุ่มลูกค้าที่มีความถี่ในการซื้อปานกลาง Monetary เฉลี่ยเท่ากับ 1741.83 เป็นกลุ่มลูกค้าที่ใช้จ่ายน้อยที่สุด โดยอาจเรียกกลุ่มนี้ได้ว่าเป็นกลุ่มที่ควรสนใจน้อยที่สุด เพราะว่า มูลค่าในการใช้จ่ายมีน้อยและความถี่ในการซื้อก็ไม่ได้มาก

เพื่อตรวจสอบว่าผลการจัดกลุ่มจากทั้งสองวิธีมีความสอดคล้องกันหรือไม่ ได้มีการประเมินโดยใช้ Adjusted Rand Index (ARI) ซึ่งเป็นตัวชี้วัดที่นิยมใช้ในการวัดความใกล้เคียงระหว่างสองชุดของกลุ่ม (Clustering) โดยไม่คำนึงถึงลำดับหรือชื่อของกลุ่ม จากการคำนวณ ARI ระหว่างผลการจัดกลุ่มจาก K-Means และ BIRCH พบว่า ค่า ARI เท่ากับ -0.0081 ซึ่งใกล้เคียงกับศูนย์ และเป็นค่าที่ติดลบเพียงเล็กน้อย แสดงให้เห็นว่า ผลการจัดกลุ่มของทั้งสองวิธีไม่มีความสัมพันธ์กันอย่างมีนัยสำคัญ หรืออาจกล่าวได้ว่า ทั้งสองชุดข้อมูลได้สะท้อนคุณลักษณะของลูกค้าที่แตกต่างกันโดยสิ้นเชิง

ค่าดัชนี ARI ที่ต่ำมากบ่งชี้ว่า กลุ่มลูกค้าที่ถูกแบ่งโดยอิงจากพฤติกรรมการซื้อ (RFM) และกลุ่มที่แบ่งโดยบริบทของการซื้อ (CAD) มีโครงสร้างของกลุ่มที่แตกต่างกันมาก ซึ่งไม่สามารถใช้แทนกันได้โดยตรง อย่างไรก็ตาม ความแตกต่างนี้เป็นประโยชน์ในเชิงกลยุทธ์ เพราะแสดงให้เห็นว่าธุรกิจสามารถนำมุมมองทั้งสองมาร่วมกันใช้เพื่อออกแบบการตลาดที่แม่นยำยิ่งขึ้น เช่น

1. ใช้กลุ่ม RFM เพื่อออกแบบแคมเปญ Loyalty Program หรือการ Retarget ลูกค้าที่มีศักยภาพสูง
2. ใช้กลุ่ม CAD เพื่อปรับกลยุทธ์การกำหนดราคาหรือเวลาการตลาดให้เหมาะสมกับแต่ละประเทศหรือช่วงเวลา

ผลลัพธ์นี้สนับสนุนแนวคิดในการใช้ Multidimensional Segmentation หรือการจัดกลุ่มลูกค้าจากหลายมุมมอง เพื่อเพิ่มความลึกในการวิเคราะห์และสร้างคุณค่าเชิงกลยุทธ์สูงสุดให้กับธุรกิจ

จากการวิเคราะห์ดังกล่าว ผลการประเมิณระหว่าง K-Means และ BIRCH ชี้ให้เห็นว่าการใช้ข้อมูล CAD ผ่าน BIRCH Clustering มีความสามารถในการปรับจำนวนกลุ่มได้ดีกว่า เนื่องจากมีคะแนนประเมินประสิทธิภาพที่ดี และการใช้ข้อมูล RFM ผ่าน K-Means Clustering ก็ยังคงจัดกลุ่มให้มีความเสถียรและสามารถตีความเชิงธุรกิจได้ง่ายกว่า การวิเคราะห์นี้จึงแสดงให้เห็นว่า CAD Model ผ่าน BIRCH Clustering เหมาะสำหรับการ แบ่งกลุ่มที่มีความแตกต่างชัดเจน แต่ RFM Model ผ่าน K-Means ยังคงมีจุดแข็งในด้านการตีความทางธุรกิจ

โดยการพิจารณาเลือกใช้งานชุดข้อมูล และ Clustering Algorithm สามารถพิจารณาได้ดังต่อไปนี้

1. เลือกใช้ RFM Model ร่วมกับ K-Means Clustering เมื่อต้องการการแบ่งกลุ่มที่สามารถนำไปวิเคราะห์หา ลูกค้าประจำ ลูกค้าที่ใช้จ่ายน้อย หรือมีแนวโนมเลิกใช้บริการ ซึ่งจะใช้ได้ดีเมื่อมีข้อมูล พฤติกรรมการซื้อในอดีต และต้องการนำไปใช้กับแคมเปญการตลาด

2. เลือกใช้ CAD Model ร่วมกับ BIRCH Clustering เมื่อต้องการการแบ่งกลุ่มลูกค้าที่มีความแม่นยำ และสะท้อนถึงพฤติกรรมในปัจจุบัน การดูพฤติกรรมลูกค้าตามช่วงเวลา หรือแบ่งตามประเทศ

3. เลือกใช้ K-means Clustering เมื่อมีโครงสร้างกลุ่มชัดเจน ต้องการการแบ่งกลุ่มที่รวดเร็ว และปรับเปลี่ยนได้ง่าย

4. เลือกใช้ BIRCH Clustering ถ้าต้องการการแบ่งกลุ่มที่สามารถปรับเปลี่ยนได้เองตามโครงสร้างข้อมูล และต้องการความแม่นยำในการจัดกลุ่มลูกค้า

5.3 ข้อเสนอแนะ

การนำฟีเจอร์ CAD (Country, AVG Basket Price, Day type) มาใช้ในการจัดกลุ่มลูกค้าด้วย BIRCH มีความสมเหตุสมผลและให้ผลลัพธ์ที่ดีทั้งในเชิงสถิติและการประยุกต์ใช้งานจริงในทางธุรกิจ โดย BIRCH สามารถแยกกลุ่มลูกค้าที่มีลักษณะพฤติกรรมแตกต่างกันได้อย่างชัดเจน เช่น ลูกค้าต่างประเทศที่มีกำลังซื้อสูงในช่วงวันหยุด หรือกลุ่มที่มียอดสั่งซื้อต่ำในวันธรรมดา สามารถนำไปใช้ในการออกแบบกลยุทธ์การตลาดแบบเจาะจงได้อย่างมีประสิทธิภาพ

อย่างไรก็ตาม การใช้ฟีเจอร์ที่มีค่าจำกัดเพียง 0 และ 1 อย่าง Country และ Day type อาจมีข้อจำกัดในด้านความละเอียดของการจำแนกลูกค้า เนื่องจากค่าดังกล่าวไม่สามารถสะท้อนความแตกต่างเชิงลึกได้เท่ากับตัวแปรเชิงต่อเนื่อง เช่น Recency หรือ Monetary จาก RFM นอกจากนี้ BIRCH แม้จะจัดกลุ่มได้ดีในชุดข้อมูลที่มีโครงสร้างชัดเจน แต่ก็อาจไวต่อการตั้งค่าพารามิเตอร์ เช่น threshold และ branching factor ซึ่งหากตั้งค่าไม่เหมาะสมอาจส่งผลต่อคุณภาพของการจัดกลุ่มได้เช่นกัน

สำหรับแนวทางในการพัฒนางานวิจัยในอนาคตโดยการใช้ Clustering อื่นๆเช่น DBSCAN , Gaussian Mixture Model (GMM) ซึ่งอาจให้ผลลัพธ์ที่ดีขึ้นหรือเหมาะกับข้อมูลบางประเภท การเปรียบเทียบอัลกอริธึมเหล่านี้ จะช่วยให้เข้าใจจุดแข็งและข้อจำกัดของแต่ละเทคนิคได้ชัดเจนมากยิ่งขึ้น หรืออาจจะหาปัจจัยภายนอกอื่นๆที่ส่งผลต่อพฤติกรรมการซื้อมาวิเคราะห์ เช่น ฤดูกาล,สภาพเศรษฐกิจ หรือสถานการณ์รอบโลก โดยการนำข้อมูลจาก Google Trends หรือจาก Social Media มาวิเคราะห์ร่วมกันก็อาจจะให้ผลลัพธ์ที่ดีมากขึ้น

บรรณานุกรม

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785-1792. Retrieved 2022/05/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S1319157819309802

Boris, L., Ana, K., Bersant, D., Dženan, S., Peter, R., & Axel, K. (2018). Variations on the Clustering Algorithm BIRCH. *Big Data Research*, *11*, 44–53. https://doi.org/10.1016/j.bdr.2017.09.002

Frédéric, R., Rabia, R., & Serge, G. (2023). PDBI: A Partitioning Davies-Bouldin Index for Clustering Evaluation. *Neurocomputing*, *523*, 249–270. https://doi.org/10.1016/j.neucom.2022.10.061

Hicham, N., & Karim, S. (2022). Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *13*(10), 122–130. Retrieved formhttps://thesai.org/Downloads/Volume13No10/Paper_14-Analysis_of_Unsupervised_Machine_Learning.pdf

Jain, & K., A. (2013). Data Clustering: 50 Years Beyond K-Means. In C. C. Aggarwal & C. K. Reddy, *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC. Retrieved formhttps://people.cs.vt.edu/~reddy/papers/DCBOOK.pdf

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666. Retrieved 2010/06/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0167865509002323

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323. Retrieved formhttps://dl.acm.org/doi/pdf/10.1145/331499.331504

Jiang, P.-c., & Qin, S. (2024). E-commerce empowers urban entrepreneurial activity—Empirical evidence from the construction of national e-commerce demonstration cities. *Cities*, *150*, 105092. Retrieved 2024/07/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0264275124003068

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Online Library. Retrieved formhttps://www.researchgate.net/publication/220695963_Finding_Groups_in_Data_An_Introduction_To_Cluster_Analysis

Kodinariya, T. M., & Makwana, D. P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*. Retrieved formhttps://www.researchgate.net/publication/313554124

Lu, J., Luo, T., & Li, K. (2025). A forward k-means algorithm for regression clustering. *Information Sciences*, *711*, 122105. Retrieved 2025/09/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0020025525002373

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

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context.

Applied Sciences, *13*(2), 797. https://doi.org/10.3390/app13020797

Rahul, S., Laxmiputra, S., & Saraswati, J. (2021). Customer Segmentation using RFM Model and K-Means Clustering. *International Journal of Scientific Research in Science and Technology*, *8*(3), 591–597. https://doi.org/10.32628/IJSRST2183118

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449. Retrieved 2024/03/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0957417423019516

Ryu, D. Y., Ko, Y. K., & Ko, Y. D. (2025). RFM analysis for profiling profitable customers based on characteristics of the hotel industry. *International Journal of Hospitality Management*, *129*, 104176. Retrieved 2025/08/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0278431925000994

Selim, S. Z., & Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(1), 81–87. Retrieved formhttps://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4767478

van Lingen, H. J., Suarez-Diez, M., & Saccenti, E. (2024). Normalization of gene counts affects principal components-based exploratory analysis of RNA-sequencing data. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1867*(4), 195058. Retrieved 2024/12/01/, from Retrieved

formhttps://www.sciencedirect.com/science/article/pii/S1874939924000543

Wei, B., Zhao, C., Cai, W., & Luo, M. (2025). Online markets, offline homes: E-commerce development and rural migrant workers' urban settlement intentions in China. *Cities*, *161*, 105867. Retrieved 2025/06/01/, from Retrieved formhttps://www.sciencedirect.com/science/article/pii/S0264275125001672

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. SIGMOD Conference,

อาภาภรณ์ วัธนกุล. (2012). ปัจจัยที่มีความสัมพันธ์ต่อพฤติกรรมการซื้อสินค้าของผู้บริโภคผ่านทางเว็บไซต์พาณิชย์อิเล็กทรอนิกส์ยอดนิมยมของประเทศไทย มหาวิทยาลัยศรีนครินทรวิโรฒ]. กรุงเทพฯ.

ประวัติผู้เขียน