การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคเค-มีนส์ และต้นไม้ตัดสินใจเพื่อแนะนำแผนการตลาดเฉพาะกลุ่ม

K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION TECHNIQUES FOR CREDIT CARD CUSTOMER SEGMENTATION AND PERSONALIZED MARKETING

ฐิติพร ฐิติพรธรรม

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคเค-มีนส์ และต้นไม้ตัดสินใจเพื่อแนะนำแผนการตลาดเฉพาะกลุ่ม

ฐิติพร ฐิติพรธรรม

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION TECHNIQUES FOR CREDIT CARD CUSTOMER SEGMENTATION AND PERSONALIZED MARKETING

THITIPORN THITIPORNDHARMA

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

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

Copyright of Srinakharinwirot University

สารนิพนธ์

เรื่อง

การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคเค-มีนส์ และต้นไม้ตัดสินใจเพื่อแนะนำแผนการตลาดเฉพาะกลุ่ม

ของ

ฐิติพร ฐิติพรธรรม

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)
คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก
(อาจารย์ ดร.ศุภร คนธภักดี)

..... ประธาน
(รองศาสตราจารย์ ดร.ดวงดาว วิชาดากุล)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

ชื่อเรื่อง	การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคเค-มีนส์ และต้นไม้ตัดสินใจเพื่อแนะนำแผนการตลาดเฉพาะกลุ่ม
ผู้วิจัย	ฐิติพร ฐิติพรธรรม
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2566
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศุภร คนธภักดี

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการจัดกลุ่มลูกค้าของบริษัทบัตรเครดิตแห่งหนึ่ง โดยการศึกษาแบ่งออกเป็น 2 ช่วง คือ ช่วงที่หนึ่งการจัดกลุ่มด้วยแบบจำลองเค-มีนส์ และช่วงที่สองการทำนายกลุ่มด้วยแบบจำลองต้นไม้ตัดสินใจ โดยในช่วงแรกใช้วิธีการหาจำนวนกลุ่ม (คลัสเตอร์) ที่เหมาะสมด้วย วิธีการ Elbow method, การวิเคราะห์ Silhouette, Davies-Bouldin index และCalinski-Harabasz index และในช่วงที่สองให้แบบจำลองต้นไม้ตัดสินใจทำนายกลุ่มลูกค้า เพื่อหาฟีเจอร์สำคัญที่ส่งผลให้แบบจำลองใช้เป็นกฎการตัดสินใจในการจัดกลุ่มลูกค้า และแสดงถึงลักษณะของลูกค้าในแต่ละกลุ่ม โดยการศึกษาในครั้งนี้ช่วยให้เข้าใจปัจจัยที่มีอิทธิพลต่อพฤติกรรมและความชอบของลูกค้าในแต่ละกลุ่มได้อย่างชัดเจนยิ่งขึ้น และในท้ายที่สุดงานวิจัยนี้มีการนำเสนอแนวทางสำหรับการออกแคมเปญทางการตลาด และโปรโมชั่นเพื่อตอบสนองความต้องการของลูกค้าในแต่ละกลุ่ม

คำสำคัญ : แบบจำลองเค-มีนส์, แบบจำลองต้นไม้ตัดสินใจ, การจัดกลุ่มลูกค้า, การวิเคราะห์ซิลูเอต, ตัวชี้วัดเดวี-บุลดีน, ตัวชี้วัดคาลินสกี-ฮาราบาซ

Title	K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION TECHNIQUES FOR CREDIT CARD CUSTOMER SEGMENTATION AND PERSONALIZED MARKETING
Author	THITIPORN THITIPORNDHARMA
Degree	MASTER OF SCIENCE
Academic Year	2023
Thesis Advisor	Subhorn Khonthapagdee , Ph.D.

This study aims to analyze customer segmentation of a credit card company. The analysis consists of two phases: K-Means clustering and Decision Tree classification. In the initial phase, various numbers of clusters were explored using the Elbow method, Silhouette analysis, the Davies-Bouldin index, and the Calinski-Harabasz index. The second phase focused on Decision Tree classification to identify key features that differentiate customer groups and capture characteristics of each cluster. Finally, guidelines for developing customized campaigns ore promotions were provided.

Keyword : K-means clustering, Decision Tree classification, Customer segmentation, Silhouette analysis, Davies-Bouldin index, Calinski-Harabasz index

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ดร.ศุภร คนธภักดี อาจารย์ที่ปรึกษา ที่ให้คำปรึกษา และคำแนะนำในการทำสารนิพนธ์ ตลอดจนสนับสนุนข้อมูลทางวิชาการ

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ที่ให้คำแนะนำ และข้อเสนอแนะที่เป็นประโยชน์ในการปรับปรุงสารนิพนธ์ให้ดียิ่งขึ้น

ขอกราบขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ สำหรับทุนสนับสนุนการนำเสนอผลงานวิจัยของนิสิตบัณฑิตศึกษาในงานประชุมวิชาการ ทำให้ได้รับประสบการณ์ที่ดี ในการเผยแพร่ผลงาน และแลกเปลี่ยนความรู้กับผู้นำเสนอท่านอื่น

สุดท้ายนี้ขอขอบพระคุณครอบครัวของผู้วิจัยที่เป็นกำลังใจจนสำเร็จการศึกษา และขอบคุณเพื่อนๆในสาขาวิชาที่คอยให้ความช่วยเหลือ และให้คำแนะนำด้วยดีเสมอมา

ฐิติพร ฐิติพรธรรม

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1

บทนำ

1.1 ภูมิหลัง

จากการเติบโตของเทคโนโลยีในปัจจุบัน การแข่งขันทางธุรกิจที่สูงขึ้น รวมถึงเป็นยุคที่ลูกค้าสามารถเข้าถึงข้อมูลของสินค้า และบริการของแบรนด์ต่างๆได้อย่างง่ายดายด้วยโทรศัพท์มือถือเพียงเครื่องเดียว ทำให้แต่ละธุรกิจต่างต้องแข่งขันกันอย่างหนัก เพื่อช่วงชิงลูกค้าและส่วนแบ่งการตลาด จากสาเหตุเหล่านี้ ทำให้แต่ละธุรกิจมีการวางแผนกลยุทธการตลาด ทั้งในด้านสินค้า บริการ ราคา และแคมเปญทางการตลาดในรูปแบบต่างๆ เพื่อนำเสนอสินค้า หรือบริการ ให้ตรงกับความต้องการของลูกค้า ทั้งนี้การส่งเสริมการขายที่ธุรกิจนิยมใช้กันบ่อยๆ คือการลดแลกแจกแถม ซึ่งเป็นแคมเปญการตลาดที่ส่งผลดีในระยะสั้น แต่ในระยะยาวไม่ส่งผลดีมากนัก เนื่องจากธุรกิจจะไม่สามารถรักษาลูกค้าที่ซื้อสินค้าเฉพาะช่วงลดราคาไว้ได้ และลูกค้าสามารถเปลี่ยนใจไปซื้อสินค้าจากคู่แข่งได้เสมอ หากสินค้าของคู่แข่งมีราคาที่ต่ำกว่า ดังนั้น ในการวางแผนกลยุทธทางการตลาด นอกจากการวางแผนกลยุทธในการหาลูกค้าใหม่แล้ว การรักษาลูกค้าให้อยู่กับธุรกิจก็มีความสำคัญเช่นกัน ซึ่งการที่ธุรกิจจะสามารถรักษาลูกค้าให้อยู่กับธุรกิจได้นั้น สิ่งที่สำคัญ คือ การรู้ข้อมูลเชิงลึกของลูกค้า เพื่อให้ธุรกิจสามารถนำเสนอสินค้า บริการ หรือ ข้อเสนอพิเศษที่ตรงกับความต้องการของลูกค้าของธุรกิจเอง อย่างถูกที่ ถูกเวลา และสร้างความแตกต่างจากคู่แข่ง

อนึ่ง การที่ธุรกิจรู้จักลูกค้า และสื่อสารหาลูกค้าทุกคนด้วยข้อเสนอแบบเดียวกัน อาจไม่ใช่วิธีการที่เหมาะสม ก่อให้เกิดต้นทุนสูง และไม่ได้ผลตอบรับที่ดี เนื่องจากลูกค้าแต่ละรายมีความต้องการที่แตกต่างกันออกไป ในขณะเดียวกันการสื่อสารหาลูกค้าด้วยรูปแบบ และข้อเสนอที่เฉพาะเจาะจงลงไปในแต่ละบุคคล ในทางปฏิบัติค่อนข้างทำได้ยาก โดยเฉพาะธุรกิจที่มีขนาดกลางถึงขนาดใหญ่ซึ่งมีปริมาณลูกค้าจำนวนมาก เพราะต้องใช้ต้นทุนทางด้านงบประมาณ แรงงาน รวมถึงต้นทุนเวลา ในการสื่อสารหาลูกค้าให้ครบทุกราย ในทางการตลาด ใช้เทคนิคการจัดกลุ่มลูกค้าออกเป็นกลุ่มย่อย (Customer Segmentation) โดยจัดกลุ่มลูกค้าที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน เพื่อให้ธุรกิจสามารถเข้าใจลักษณะเฉพาะที่มีร่วมกันของลูกค้า เข้าใจพฤติกรรมในการใช้จ่าย โดยประโยชน์ของการจัดกลุ่มลูกค้า คือ ช่วยให้ฝ่ายการตลาดสามารถวางแผนการขายสินค้า บริการ และข้อเสนอพิเศษที่เฉพาะเจาะจงลงไปในแต่ละกลุ่ม ช่วยปรับปรุงงานด้านลูกค้าสัมพันธ์ให้มีประสิทธิภาพ เป็นการสร้างความจงรักภักดีที่ลูกค้ามีต่อธุรกิจ

และทำให้ธุรกิจทราบถึงสัดส่วนรายได้ที่มาจากกลุ่มลูกค้า ทำให้ธุรกิจสามารถจัดสรรการใช้งบประมาณลงไปในแต่ละกลุ่มได้อย่างเหมาะสม นอกจากนี้ยังเพิ่มโอกาสทางธุรกิจในการวิจัยพัฒนาสินค้า รวมถึงการออกสินค้าใหม่

ในงานวิจัยนี้นำเสนอการจัดกลุ่มลูกค้าบัตรเครดิต โดยแบ่งเป็น 2 ช่วง คือ ช่วงที่ 1 การจัดกลุ่มลูกค้าบัตรเครดิตซึ่งเป็นชุดข้อมูลที่ไม่มีเลเบล ด้วยการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Machine Learning) โดยจัดกลุ่มจากลักษณะทางประชากรศาสตร์ และพฤติกรรม ด้วยแบบจำลองK-Means Clustering และช่วงที่ 2 ทำนายกลุ่มลูกค้าบัตรเครดิตด้วยด้วยการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning)ด้วยแบบจำลอง Decision Tree Classification เพื่อให้การจัดกลุ่มมีประสิทธิภาพ เพิ่มความน่าเชื่อถือให้แบบจำลอง ทราบถึงกฎการตัดสินใจ และฟีเจอร์ที่มีผลต่อการตัดสินใจของแบบจำลอง

1.2 ความมุ่งหมายของงานวิจัย

ในงานวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1.2.1 เพื่อจัดกลุ่มลูกค้าบัตรเครดิตที่มีลักษณะพฤติกรรมคล้ายคลึงกันออกเป็นกลุ่มย่อย โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Machine Learning) โดยการแบ่งกลุ่มด้วยแบบจำลอง K-Means Clustering

1.2.2 เพื่อทำนายกลุ่มของลูกค้าบัตรเครดิตโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning) โดยใช้แบบจำลอง Decision Tree Classification

1.2.3 เพื่อทราบถึงหลักการของแบบจำลอง K-Means Clustering ในการจัดกลุ่มและนำผลจากการจัดกลุ่มไปใช้ในการวางแผนงานด้านการตลาด และลูกค้าสัมพันธ์

1.2.4 เพื่อทราบถึงหลักการของแบบจำลอง Decision Tree Classification ในการทำนายกลุ่มของลูกค้าบัตรเครดิต และสามารถอธิบายผลของแบบจำลอง

1.2.5 เพื่อมีความรู้ และความเข้าใจในการแบ่งกลุ่มลูกค้าออกเป็นกลุ่มย่อย (Customer Segmentation)

1.3 ความสำคัญของงานวิจัย

งานวิจัยนี้ศึกษาการจัดกลุ่มลูกค้าบัตรเครดิตออกเป็นกลุ่มย่อย โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Machine Learning) ในการจัดกลุ่ม (Clustering) ด้วยแบบจำลอง K-Means และทำนายกลุ่มของลูกค้าโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning) โดยใช้แบบจำลอง Decision Tree Classification ทดลองกับ

ชุดข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com ประกอบไปด้วย 10 แอททริบิวต์ และมีจำนวนข้อมูลทั้งหมด 10,127 แถว

1.4 ขอบเขตของงานวิจัย

1.4.1 ประชากรที่ใช้ในการวิจัย

ชุดข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com ประกอบไปด้วย 10 แอททริบิวต์ และมีจำนวนข้อมูลทั้งหมด 10,127 แถว

1.4.2 กลุ่มตัวอย่างที่ใช้ในการวิจัย

ชุดข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com

1.4.3 ตัวแปรที่ศึกษา

- Age หรือ อายุของลูกค้า
- Gender หรือ เพศของลูกค้า (M แทนเพศชาย/F แทนเพศหญิง)
- Education_Level หรือ ระดับการศึกษา (High School/College/Graduate/Post-Graduate/Doctorate/Unknown/Uneducated)
- Marital_Status หรือ สถานภาพ (Single/Married/Divorced/Unknown)
- Months on book หรือ จำนวนเดือนที่ลูกค้าถือบัตรเครดิต
- Credit_Limit หรือ วงเงินบัตรเครดิตที่ลูกค้าได้รับ
- Total_Transaction_Amount หรือ ยอดเงินการใช้จ่ายทั้งหมด
- Total_Transaction_Count หรือ จำนวนครั้งในการใช้จ่าย
- Minimum_Income หรือ รายได้ขั้นต่ำที่ได้รับ
- Maximum_Income หรือ รายได้สูงสุดที่ได้รับ

1.5 กรอบแนวคิดในงานวิจัย

งานวิจัยนี้ศึกษาการจัดกลุ่มลูกค้าที่มีพฤติกรรมใกล้เคียงกันออกเป็นกลุ่มย่อยโดยใช้ Machine Learning เป็นเครื่องมือสำหรับสร้างแบบจำลองในการจัดกลุ่มลูกค้า และทำนายกลุ่มลูกค้า โดยใช้ชุดข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่ง ประกอบด้วย 10 แอททริบิวต์ และมีจำนวนข้อมูลทั้งหมด 10,127 แถว ข้อมูลถูกเก็บในรูปแบบตาราง จากนั้นจึงใช้ Machine Learning มาเป็นเครื่องมือในการสร้างแบบจำลองเพื่อจัดกลุ่มลูกค้า และทำนายกลุ่มลูกค้า โดย

เขียนด้วยโปรแกรมภาษา Python งานวิจัยนี้สร้างแบบจำลอง 2 ขั้นตอน คือ 1) การสร้างแบบจำลองเพื่อจัดกลุ่มลูกค้า ใช้เทคนิคการจัดการฟีเจอร์ (Feature Engineering) เพื่อปรับข้อมูลให้อยู่ในระดับเดียวกัน ก่อนเข้าสู่แบบจำลอง ใช้เทคนิคการจัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means Clustering ใช้วิธี Elbow method ในการหาจำนวนกลุ่มที่เหมาะสม และวัดผลประสิทธิภาพของแบบจำลองด้วย Silhouette Score, Davies-Bouldin Index และCalinski-Harabasz Index 2) การทำนายกลุ่มของลูกค้าด้วยแบบจำลอง Decision Tree Classification และวัดผลประสิทธิภาพของแบบจำลอง Classification ด้วยค่า Accuracy, Precision, Recall และF1-Score ขั้นตอนสุดท้าย อธิบายค่าความสำคัญของฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้ด้วย Feature Importance

1.6 สมมติฐานในการวิจัย

1.6.1 แบบจำลอง K-Means เป็นแบบจำลองที่ประสิทธิภาพในการจัดกลุ่มข้อมูล สามารถจัดการกับข้อมูลขนาดใหญ่ และใช้เวลาในการทำงาน (Fit data) น้อย

1.6.2 แบบจำลอง Decision Tree Classification เป็นแบบจำลองที่มีประสิทธิภาพในการทำนายกลุ่มลูกค้าที่มีมากกว่าสองคลาส และเป็นแบบจำลองที่สามารถแปลความได้ง่าย

1.6.3 แบบจำลอง Decision Tree Classification ทำให้ทราบถึงกฎการตัดสินใจของแบบจำลองในการทำนายข้อมูล ช่วยให้ผู้วิจัยทราบถึงฟีเจอร์ที่มีผลในการตัดสินใจของแบบจำลอง

1.6.4 วิธีการจัดกลุ่มลูกค้าโดยใช้เทคนิค Clustering และ Classification เป็นวิธีการจัดกลุ่มที่มีประสิทธิภาพ เนื่องจากทั้ง 2 เทคนิคส่งเสริมประสิทธิภาพกัน และการใช้ Clustering และ Classification ทำให้ทราบถึงข้อมูลเชิงลึกของพฤติกรรม และลักษณะของลูกค้าที่อยู่ในแต่ละกลุ่ม

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และนำเสนอตามหัวข้อต่อไปนี้

1. การแบ่งกลุ่มลูกค้า (Customer Segmentation)
2. เทคนิคในการจัดการข้อมูลดิบ (Feature Engineering)
3. การสร้างแบบจำลองในการจัดกลุ่มข้อมูล (Clustering Model)
4. การวัดผลประสิทธิภาพของแบบจำลองจัดกลุ่ม
5. การสร้างแบบจำลองในการทำนายกลุ่มลูกค้า
6. การวัดผลประสิทธิภาพของแบบจำลองทำนายกลุ่มลูกค้า
7. งานวิจัยที่เกี่ยวข้อง

2.1 การแบ่งกลุ่มลูกค้า (Customer Segmentation)

การแบ่งกลุ่มลูกค้าเป็นขั้นตอนหนึ่งของการจัดการการตลาด โดยมีทฤษฎี STP Theory STP Theory (Kotler & Keller, 2016)ประกอบไปด้วย Segmentation, Targeting และ Positioning ในงานวิจัยนี้นำเสนอ Segmentation หรือ การแบ่งกลุ่มลูกค้า

การแบ่งกลุ่มลูกค้า คือ กระบวนการในการแบ่งลูกค้าของธุรกิจออกเป็นกลุ่มย่อย โดยในแต่ละกลุ่มมีลักษณะที่เหมือน หรือ คล้ายคลึงกัน สามารถแบ่งออกได้ตามลักษณะ 4 ประเภท ได้แก่

1. แบ่งตามประชากรศาสตร์ (Demographic) คือ การแบ่งกลุ่มลูกค้าตาม เพศ อายุ ระดับการศึกษา สถานะ และอาชีพ เป็นต้น
2. แบ่งตามพื้นที่ (Geographic) คือ การแบ่งกลุ่มลูกค้าตามพื้นที่อยู่อาศัย ตำบล อำเภอ ประเทศ และความหนาแน่นของประชากรในพื้นที่ เป็นต้น
3. แบ่งตามหลักจิตวิทยา (Psychographic) คือ การแบ่งกลุ่มลูกค้าตามความสนใจ พฤติกรรม ความเชื่อ นิสัย ความชอบส่วนตัว เป็นต้น
4. แบ่งตามพฤติกรรม (Behavioral) คือ การแบ่งกลุ่มลูกค้าจากพฤติกรรมการซื้อสินค้า การใช้สินค้า ความถี่ในการซื้อสินค้า เวลาในการซื้อสินค้า ประเภทของสินค้าที่ซื้อ เป็นต้น

การแบ่งกลุ่มลูกค้า ทำให้ธุรกิจทราบข้อมูลเชิงลึกของลูกค้าในแต่ละกลุ่ม ช่วยให้ธุรกิจวางแผนการตลาด เพื่อนำเสนอผลิตภัณฑ์ บริการ ได้ตรงกับความต้องการของลูกค้า สร้างความสัมพันธ์กับลูกค้า มองเห็นโอกาสในการหาลูกค้าเพิ่มในอนาคต และสร้างรายได้เพิ่มให้กับธุรกิจ

2.2 **เทคนิคในการจัดการข้อมูลดิบ** (Feature Engineering)

Feature Engineering คือ กระบวนการในการแยกคุณลักษณะ หรือ ฟีเจอร์ (Feature) จากข้อมูลดิบที่ทำให้แบบจำลองสามารถทำงานได้ดีขึ้น โดยทั่วไปฟีเจอร์จะอยู่ในรูปแบบของโครงสร้างของคอลัมห์ หรือ แอททริบิวต์ โดยขั้นตอนในการทำ Feature Engineering ได้แก่

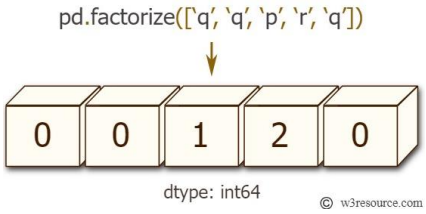
- การเลือกฟีเจอร์ (Feature selection)
- การจัดการค่าที่หายไป (Missing value)
- การจัดการข้อมูลที่ไม่สมดุลกัน (Imbalanced data)
- การจัดการค่าผิดปกติ (Outlier)
- การจัดกลุ่มค่าต่อเนื่อง (Binning)
- การแปลงค่าประเภทของข้อมูลให้เป็นตัวเลข (Encoding)
- การปรับช่วงค่าขอบเขตของฟีเจอร์ (Feature scaling)

เนื่องจากชุดข้อมูลที่นำมาใช้ในงานวิจัยเป็นชุดข้อมูลที่ผ่านการทำความสะอาดแล้ว ดังนั้น ในขั้นตอน Feature Engineering ในงานวิจัยนี้จึงเป็นขั้นตอนการแปลงค่าประเภทของข้อมูลให้เป็นตัวเลข (Encoding)และการปรับช่วงขอบเขตของฟีเจอร์ (Feature scaling) โดยมีรายละเอียดดังนี้

2.2.1 **การแปลงค่าประเภทของข้อมูลให้เป็นตัวเลข** (Encoding)

เนื่องจากแบบจำลองสามารถทำงานได้ดีเมื่อฟีเจอร์เป็นค่าประเภทตัวเลข และชุดข้อมูลที่ใช้ในงานวิจัยนี้ประกอบไปด้วยฟีเจอร์ประเภทตัวเลข และฟีเจอร์ประเภทหมวดหมู่ จึงใช้วิธีแปลงค่าฟีเจอร์ประเภทหมวดหมู่เป็นค่าตัวเลขด้วยฟังก์ชั่น Factorize ()

ฟังก์ชั่น Factorize () เป็นหนึ่งในฟังก์ชั่นของ Pandas ที่แปลงข้อมูลประเภทหมวดหมู่ หรือ ข้อมูลประเภทแจกแจงให้เป็นข้อมูลตัวเลข ทำให้ค่าฟีเจอร์ถูกแปลงเป็นตัวเลขในรูปแบบอาเรย์ (Array) ที่ตัวเลขระบุค่าความแตกต่างของข้อมูล ดังภาพประกอบที่ 1



ภาพประกอบ 1 แสดงวิธีแปลงค่าฟีเจอร์ด้วยฟังก์ชั่นFactorize()

ที่มา Pandas Series: factorize() function. Retrieved October 1, 2023, from https://www.w3resource.com/pandas/series/series-factorize.php

2.2.2 การปรับช่วงค่าขอบเขตของฟีเจอร์ (Feature scaling)

Feature scaling คือ การปรับช่วงค่าขอบเขตของฟีเจอร์ที่เป็นข้อมูลประเภทตัวเลขให้อยู่ในช่วงขอบเขตเดียวกัน เพื่อให้ข้อมูลเหมาะกับการที่แบบจำลองนำไปประมวลผล ในงานวิจัยนี้ทำ Feature scaling โดยการทำข้อมูลประเภทตัวเลขให้เป็นค่ามาตรฐาน (Normalization) ด้วยวิธี MinMaxScaler เนื่องจากชุดข้อมูลที่ใช้ในการวิจัยมีการกระจายตัวของข้อมูลที่ไม่ได้อยู่ในลักษณะปกติ (Normal Distribution)

MinMaxScaler เป็นการปรับค่าช่วงของข้อมูลให้อยู่ในช่วง [0,1] ซึ่งข้อมูลแต่ละตัวที่สนใจ (i) ขนาดของข้อมูลจะถูกลบด้วยค่าต่ำสุด และถูกหารด้วยส่วนต่างระหว่างค่าสูงสุดและค่าต่ำสุดของข้อมูล โดยมีสมการดังสมการที่ 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

2.3 การสร้างแบบจำลองในการจัดกลุ่มข้อมูล (Clustering Model)

Clustering model เป็น Machine learning ที่อยู่ในประเภท Unsupervised เป็นแบบจำลองที่ชุดข้อมูลไม่มีการกำหนดเลเบล หรือ ต้นแบบของผลลัพธ์ โดยการ Clustering เป็นการจัดกลุ่มข้อมูลที่อาศัยกระบวนการค้นหาโครงสร้างที่คล้ายคลึงกันในชุดข้อมูล ข้อมูลที่มีลักษณะคล้ายคลึงกันจะอยู่ในกลุ่มเดียวกัน เช่น การจัดกลุ่มลูกค้าจากพฤติกรรมการซื้อสินค้า ลูกค้าที่มีลักษณะคล้ายกันจะเป็นลูกค้าประเภทเดียวกัน ประเภทของClustering ที่ใช้ในการจัดกลุ่มมีหลายประเภท ได้แก่ 1) การจัดกลุ่มแบบมีลำดับชั้น (Hierarchical clustering) 2) การจัดกลุ่มข้อมูลจากจุดข้อมูล (Centroid-based clustering) 3) การจัดกลุ่มแบบการกระจาย

(Distribution-based clustering) 4) การจัดกลุ่มของข้อมูลจากการกระจุกตัวของจุดข้อมูล (Density-based clustering) 5) การจัดกลุ่มแบบคลุมเครือ (Fuzzy clustering) 6) การจัดกลุ่มของข้อมูลจากข้อจำกัด (Constraint-based clustering)

ในงานวิจัยนี้ ใช้แบบจำลองในการจัดกลุ่มข้อมูลจากจุดข้อมูล (Centroid-based clustering) เป็นเทคนิคในการจัดกลุ่มข้อมูลที่พิจารณาระยะห่างระหว่างจุดข้อมูลกับจุดศูนย์กลางของข้อมูล (Centroid) โดยแบบจำลองที่ใช้ ได้แก่ แบบจำลอง K-Means จึงขออธิบายเฉพาะแบบจำลองที่ใช้ในงานนี้

2.3.1 **แบบจำลอง** K-Means Clustering

เป็นแบบจำลองการจัดกลุ่มข้อมูลที่นิยมใช้ เนื่องจากมีวิธีการทำงานที่ไม่ซับซ้อน และเข้าใจง่ายโดย K-Means เป็นการตัดแบ่ง (Partition) ข้อมูลออกเป็น k กลุ่ม และแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (Centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดย K-Means มีขั้นตอนในการทำงานดังนี้

ขั้นตอนที่ 1 เริ่มจากการกำหนดค่า k หรือจำนวนกลุ่มที่ต้องการ k กลุ่ม

ขั้นตอนที่ 2 สุ่มวางตำแหน่งจุดศูนย์กลางของข้อมูล (Centroid)

ขั้นตอนที่ 3 จากจุดศูนย์กลาง (Centroid) แต่ละจุด ทำการคำนวณระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยระยะห่างคำนวณได้จากสมการ Euclidean distance โดยข้อมูลจะถูกจัดอยู่ในกลุ่ม Centroid ที่มีระยะทางใกล้ที่สุด สมการ Euclidean distance ดังสมการที่ 2

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

ขั้นตอนที่ 4 หลังจากทำการจัดกลุ่มข้อมูลใหม่แล้ว ทำการคำนวณค่าเฉลี่ยของสมาชิกในกลุ่ม เพื่อกำหนดเป็น Centroid ของกลุ่มข้อมูลใหม่ โดยในการคำนวณในแต่ละรอบค่าผลรวมของระยะห่างระหว่างสมาชิกในกลุ่มกับ Centroid (Within-Cluster-Sum-of-Squares หรือ WCSS) จะค่อยๆ ลดลง

ขั้นตอนที่ 5 ทำซ้ำในขั้นตอนที่ 3และ4 จนค่าของ Centroid ในกลุ่มข้อมูลใหม่ ไม่เปลี่ยนแปลง หรือได้ค่าไม่ต่างกับค่าของ Centroid ในรอบก่อนหน้า

2.4 การวัดผลประสิทธิภาพของแบบจำลองแบบจัดกลุ่ม

การวัดผลประสิทธิภาพของแบบจำลองแบบจัดกลุ่ม แบ่งการวัดประสิทธิภาพออกเป็น 3 ประเภท ได้แก่ 1) ตัวชี้วัดภายใน (Internal metric) 2) ตัวชี้วัดภายนอก (External metric) 3) เทคนิคการเลือกจำนวนClusterที่เหมาะสม (Validation method) ในงานวิจัยนี้การวัดผลประสิทธิภาพของแบบจำลองแบบจัดกลุ่ม ด้วยตัวชี้วัดภายใน (Internal metric) และเทคนิคการเลือกจำนวนClusterที่เหมาะสม (Validation method) จึงขออธิบายเฉพาะการวัดผลที่ใช้ในงานนี้

2.4.1 ตัวชี้วัดภายใน (Internal metric)

ตัวชี้วัดภายในจะขึ้นอยู่กับลักษณะของข้อมูลและCluster เช่น การเกาะกลุ่มของCluster การแยกตัวกันระหว่างCluster ความหนาแน่น และรูปทรงของCluster เป็นต้น ซึ่งตัวชี้วัดภายในที่ใช้ในงานวิจัยนี้ ได้แก่

2.4.1.1 Silhouette coefficient score วัดคุณภาพของกลุ่ม (Cluster) โดยพิจารณาจากการแยกตัวกันระหว่างกลุ่ม และความเหมือนของจุดข้อมูลในแต่ละกลุ่ม โดยคำนวณระยะห่างระหว่างกลุ่มเทียบกับระยะห่างระหว่างจุดข้อมูลในกลุ่มเดียวกัน โดยค่าจะอยู่ระหว่าง -1 ถึง 1

- ค่า Silhouette score เท่ากับ 1 แสดงถึงกลุ่มข้อมูลที่แยกตัวห่างจากกลุ่มข้อมูลอื่นอย่างชัดเจน และจุดข้อมูลในกลุ่มมีการเกาะกลุ่มกัน
- ค่า Silhouette score เท่ากับ 0 แสดงถึงขอบเขตของแต่ละกลุ่มข้อมูลมีการทับซ้อนกัน ไม่แยกตัวห่างจากกลุ่มข้อมูลอื่น
- ค่า Silhouette score เท่ากับ -1 แสดงถึงจุดข้อมูลอยู่ผิดกลุ่ม

สูตรการคำนวณ Silhouette coefficient score ดังสมการที่ 3

$$s = \frac{b - a}{\max(a, b)} \tag{3}$$

a แทนค่าเฉลี่ยของระยะห่างระหว่างตัวอย่างกับจุดข้อมูลอื่นๆทั้งหมดในกลุ่มเดียวกัน

b แทนค่าเฉลี่ยของระยะห่างระหว่างตัวอย่างกับจุดข้อมูลทั้งหมดในกลุ่มถัดไปที่อยู่ใกล้ที่สุด

2.4.1.2 Davies-Bouldin score คำนวณค่าเฉลี่ยอัตราส่วนระหว่างความแตกต่างภายในกลุ่มกับความแตกต่างระหว่างกลุ่มของแต่ละคู่กลุ่มข้อมูล โดยมีค่าตั้งแต่ 0 ขึ้นไป ซึ่งค่า Davies-Bouldin score เข้าใกล้ 0 แสดงถึงการจัดกลุ่มมีคุณภาพดี และแต่ละกลุ่มแยกตัวห่างกันอย่างชัดเจน

สูตรการคำนวณ Davies-Bouldin index ดังสมการที่ 4

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j} R_{ij} \tag{4}$$

K แทนกลุ่มข้อมูล

Rij แทนอัตราส่วนระหว่างกลุ่ม i และกลุ่ม j

Max แทนค่าอัตราส่วนภายในกลุ่มที่สูงที่สุด

2.4.1.3 Calinski-Harabasz index หรือที่รู้จักในชื่อ Variance Ratio Criterion วัดความแปรปรวนระหว่างกลุ่มเทียบกับความแปรปรวนภายในกลุ่ม หลักการคือการจัดกลุ่มที่ดีควรมีความแตกต่างระหว่างกลุ่มสูง และความแตกต่างภายในกลุ่มต่ำ ค่าดัชนีที่สูงแสดงถึงการจัดกลุ่มมีประสิทธิภาพดี โดยมีค่าตั้งแต่ 0 ขึ้นไป ซึ่งค่า Davies-Bouldin index สูงแสดงถึงถึงการเกาะกลุ่มกันของจุดข้อมูลในกลุ่ม และแต่ละกลุ่มแยกตัวห่างกันอย่างชัดเจน

สูตรการคำนวณ Calinski-Harabasz index ดังสมการที่ 5

$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1} \tag{5}$$

BGSS หรือ Between-Group Sum of Squares คำนวณระยะห่างระหว่างกลุ่มแต่ละกลุ่ม

WGSS หรือ Within-Group Sum of Squares คำนวนระยะห่างของจุดข้อมูลภายในกลุ่มเดียวกัน

N แทนจำนวนจุดข้อมูลทั้งหมด

K แทนจำนวนกลุ่มทั้งหมด

2.4.2 เทคนิคการเลือกจำนวนClusterที่เหมาะสม (Validation method)

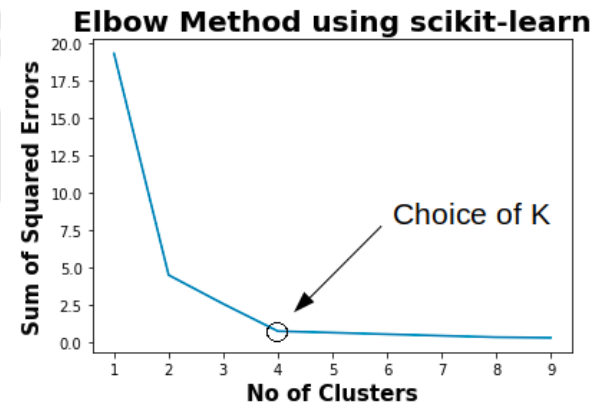
เนื่องจากแบบจำลองK-Means เป็นแบบจำลองที่ต้องมีการกำหนดจำนวนกลุ่ม (Cluster) ซึ่ง Elbow method เป็นเทคนิคในการเลือกจำนวน Clusterที่เหมาะสม โดยใช้วิธีการลดค่าความคลาดเคลื่อน (Error) ของผลรวมของระยะห่างระหว่างจุดข้อมูลและCentroid ของทุกCluster เรียกว่า Within-Cluster-Sum-of -Squares (WCSS)

สูตรการคำนวณ Within-Cluster-Sum-of -Squares ดังสมการที่ 6

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i in C_i}^{d_m} distance(d_i, C_k)^2 \right) \tag{6}$$

Where,
C is the cluster centroids and d is the data point in each Cluster.

เทคนิค Elbow method นิยมใช้ภาพประกอบในการพิจารณาจำนวนClusterที่เหมาะสม โดยจุดที่เหมาะสมในการเลือกจำนวนCluster คือ จุดที่มีการหักศอก ซึ่งที่เป็นจุดที่ค่า Within-Cluster-Sum-of -Squared ลดลงน้อยเมื่อจำนวนClusterเพิ่มขึ้น ดังภาพประกอบที่ 2



ภาพประกอบ 2 เทคนิคElbow MethodสำหรับแบบจำลองK-Means

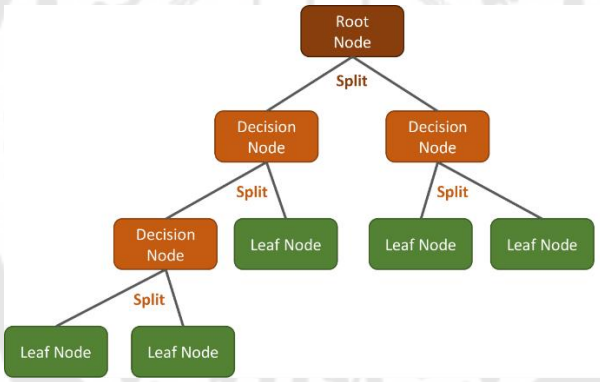
ที่ ม า Elbow method. Retrieved October 1, 2023, from https://dchandra.com/machine%20learning/2018/12/16/K-means-Clustering-Algorithm-using-scikit-learn.html

2.5 การสร้างแบบจำลองในการทำนายกลุ่มลูกค้า

2.5.1 แบบจำลอง Decision Tree Classification

แบบจำลอง Decision Tree Classification เป็นแบบจำลองประเภท Supervised learning แบบจำแนกหมวดหมู่ ที่ชุดข้อมูลมี Target หรือ ต้นแบบของผลลัพธ์ ไว้สำหรับให้แบบจำลองเรียนรู้ เพื่อให้แบบจำลองทำนายคำตอบออกมาตามTargetที่ได้กำหนดไว้ แบบจำลอง Decision Tree เป็น Rule-Based model คือ สร้างกฎ If-Else ขึ้นมาจากค่าของแต่ละฟีเจอร์ เพื่อแบ่งข้อมูลออกเป็นกลุ่มใหม่ที่สามารถอธิบาย Target ได้ดีที่สุด

วิธีการของแบบจำลอง Decision Tree คือ การค่อยๆแบ่งข้อมูลออกเป็น 2 ส่วน เริ่มจากส่วนที่อยู่ด้านล่างสุดเรียกว่า Root Node เป็นชุดข้อมูลกฎเกณฑ์ตั้งต้น เมื่อทำการวิเคราะห์ลงไปจนถึงกลุ่มสุดท้ายเรียกว่า Leaf Node ดังภาพประกอบที่ 3



ภาพประกอบ 3 วิธีการทำงานของแบบจำลองDecision Tree

ที่มา Decision Tree Method. Retrieved October 1, 2023 from https://blogs.fu-berlin.de/reseda/random-forest/

สิ่งที่สำคัญสำหรับแบบจำลอง Decision Tree Classification คือ การแตกสาขาของต้นไม้ในแต่ละครั้งจะต้องหาจุดที่ดีที่สุดในการแบ่งข้อมูลและลดค่าCost function ให้น้อยที่สุด โดย Cost function ที่ใช้ในการคำนวณ ได้แก่

2.5.1.1 Gini Impurity

เป็นการวัดค่าความไม่บริสุทธิ์ในการอธิบายTargetของกลุ่มที่ถูกแบ่งจากฟีเจอร์ ซึ่งหมายความว่ายิ่งค่า Impurity น้อยยิ่งแบ่งข้อมูลออกมาได้ดี

การคำนวณ Gini Impurity คือ การนำผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่สนใจคูณ (1 ลบ ค่าความน่าจะเป็นของเหตุการณ์ที่สนใจ) สูตรการคำนวณ Gini Impurity ดังสมการที่ 8

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{8}$$

2.5.1.2 Entropy

เป็นการวัดความไม่แน่นอนในการอธิบายTargetของกลุ่มที่ถูกแบ่งจากฟีเจอร์ ซึ่งหมายความว่ายิ่งค่า Entropyน้อยยิ่งแบ่งข้อมูลออกมาได้ดี

การคำนวณ Entropy คือ การนำผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่สนใจคูณ log ของความน่าจะเป็นของเหตุการณ์ที่สนใจ สูตรการคำนวณ Entropy ดังสมการที่ 9

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j) \tag{9}$$

2.6 **การวัดผลประสิทธิภาพของแบบจำลอง** Decision Tree Classification

ในงานวิจัยนี้สร้างแบบจำลอง Decision Tree ที่จัดการปัญหาแบบ Classification ซึ่งสิ่งที่ต้องการทำนายอยู่ในรูปหมวดหมู่ หรือ คลาส จึงใช้การวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall และF1-Score ซึ่งคำนวณได้จาก Confusion Matrix ดังตารางที่ 1

ตาราง 1แสดง Confusion Matrix

		Prediction	
Actual		TP	FN
		FP	TN

โดยที่

TP (True Positive)	คือ สิ่งที่แบบจำลองทำนายว่าจริงตรงกับสิ่งที่เกิดขึ้นว่าจริง
TN (True Negative)	คือ สิ่งที่แบบจำลองทำนายว่าไม่จริงตรงกับสิ่งที่เกิดขึ้นว่าไม่จริง
FP (False Positive)	คือ สิ่งที่แบบจำลองทำนายว่าจริงแต่สิ่งที่เกิดขึ้นคือไม่จริง
FN (False Negative)	คือ สิ่งที่แบบจำลองทำนายว่าไม่จริงแต่สิ่งที่เกิดขึ้นคือจริง

2.6.1 Accuracy

คือ ค่าความถูกต้องที่แบบจำลองทายถูก เป็นอัตราส่วนของการทำนายถูกกับจำนวนข้อมูลทั้งหมด แสดงได้ดังสมการที่ 10

$$Accuracy = \frac{TP + TN}{N} \tag{10}$$

2.6.2 Precision

คือ ค่าความแม่นยำของแบบจำลองในการทำนายว่าจริงตรงกับสิ่งที่เกิดขึ้นว่าจริงกับจำนวนข้อมูลทั้งหมดที่เป็นPositive แสดงได้ดังสมการที่ 11

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

2.6.3 Recall **หรือ** Sensitivity

คือ อัตราส่วนที่แบบจำลองทำนายว่าจริงกับจำนวนข้อมูลที่เป็นจริงทั้งหมด แสดงได้ดังสมการที่ 12

$$Recall = \frac{TP}{(TP + FN)} \tag{12}$$

2.6.4 F1-Score

คือ เป็นค่าเฉลี่ยแบบฮาร์โมนิกระหว่างPrecisionและRecall แสดงได้ดังสมการที่ 13

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (13)$$

2.7 งานวิจัยที่เกี่ยวข้อง

ในปัจจุบันมีงานวิจัยที่ศึกษาเทคนิคการจัดกลุ่มลูกค้าด้วยคุณลักษณะ และพฤติกรรมที่เหมือนหรือคล้ายคลึงกันโดยใช้เทคนิคการจัดกลุ่มในแบบจำลองประเภทต่างๆ ผู้วิจัยจึงศึกษาและพิจารณาวิธีการจากงานวิจัยเหล่านี้

2.7.1 **บทความเรื่อง** Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier **โดย** Shuzlina Abdul-Rahman, Nurin Faiqah Kamal Arifin, Mastura Hanafiah & Sofianita Mutalib(Abdul-Rahman et al., 2021)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าในธุรกิจประกัน เพื่อให้ฝ่ายการตลาดของธุรกิจประกันสามารถวางกลยุทธที่เหมาะสมกับลูกค้าแต่ละกลุ่ม ก่อให้เกิดประโยชน์สูงสุดแก่ลูกค้า และเพื่อให้ลูกค้าได้รับแผนประกันที่เหมาะกับความต้องการ และสร้างความพึงพอใจให้กับลูกค้า โดยใช้เทคนิค K-Modes ในการจัดกลุ่ม และใช้แบบจำลองDecision Tree Classifier ทำนายกลุ่มลูกค้าด้วย

จากการทดลองงานวิจัยนี้ใช้ชุดข้อมูลจากบริษัทประกันภัยแห่งหนึ่งในประเทศมาเลเซีย ที่เก็บในช่วงมกราคม 2561-ธันวาคม 2562 ประกอบไปด้วยข้อมูลจำนวน 37,181 แถว และเลือกแอททริบิวต์ที่เกี่ยวข้องจำนวน 12 แอททริบิวต์

ผลการวิจัยพบว่า การจัดกลุ่มลูกค้าด้วย K-Modes สามารถจัดกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ กลุ่มลูกค้าที่มีศักยภาพสูง (Potential High-Value Customer) กลุ่มลูกค้าที่มีศักยภาพต่ำ (Low Value Customer) และกลุ่มลูกค้าที่ไม่น่าสนใจ (Disinterested Customer) และการทำนายกลุ่มลูกค้าด้วยแบบจำลอง Decision Tree Classifier ได้ค่าAccuracy สูงสุดที่ 81.30% เมื่อปรับค่าพารามิเตอร์ด้วย Gini โดยตั้งค่า Leaf node สูงสุดไม่เกิน 50 และใช้เทคนิคการแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพแบบจำลองจำนวน 10 ส่วน (10-Fold Cross Validation)

2.7.2 **บทความเรื่อง** Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning **โดย** Musadig Aliyev, Elvin Ahmadov & Habil Gadirli(Aliyev et al., 2020)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าของธนาคาร เพื่อให้ธนาคารสามารถวางแผนการนำเสนอผลิตภัณฑ์ ช่วงเวลาที่เหมาะสมในการนำเสนอ และกลุ่มลูกค้าที่ควรนำเสนอเพื่อให้เหมาะกับลูกค้าแต่ละกลุ่ม ซึ่งการวางแผนนี้ทำให้ธนาคารได้รับผลตอบรับจากลูกค้าเป็นอย่างดี ทั้งผลตอบรับด้านการบริการใหม่ และการซื้อผลิตภัณฑ์ใหม่

งานวิจัยนี้ใช้ชุดข้อมูลจริงจากUnibank ซึ่งเป็นธนาคารในประเทศเอเซอร์ไบจัน โดยใช้เทคนิค RFM ร่วมกับแบบจำลองจัดกลุ่มลูกค้าแบบต่างๆ เทคนิค RFM ใช้การคำนวณมูลค่าของลูกค้าโดยคิดจาก Recency, Frequency และMonetary จากนั้นจึงนำมูลค่าของลูกค้าที่ได้จากการคำนวณ RFM มาแบ่งกลุ่มลูกค้าด้วยแบบจำลองจัดกลุ่ม ได้แก่ K-Means, DBSCAN และ Hierarchical

วิธีการที่ 1 นำมูลค่าของลูกค้าที่ได้จากการคำนวณ RFM มาจัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means และหาค่า K หรือ จำนวนCluster ที่เหมาะสมด้วยวิธี Elbow Method ได้ค่า K เท่ากับ 4 Cluster จากนั้นผู้เขียนได้ทำการแปลงข้อมูลเป็นภาพและสังเกตเห็นกลุ่มข้อมูลที่น่าสนใจ ซึ่งเป็นกลุ่มลูกค้าที่มีลักษณะพฤติกรรมไม่เหมือนกัน คือ ในกลุ่ม แยกย่อยออกเป็นกลุ่มจุดข้อมูลที่ค่าRecencyต่ำ และ กลุ่มจุดข้อมูลที่ค่าRecencyสูง จึงใช้แบบจำลอง K-Means ในการจัดกลุ่มลูกค้ากลุ่มนี้อีกขั้นตอนหนึ่ง ผลการวิจัยของวิธีการที่ 1 จัดกลุ่มลูกค้าออกเป็น 5 กลุ่ม ได้แก่ กลุ่มที่ 1 เป็นลูกค้าที่เคยใช้บริการสม่ำเสมอในอดีต และปัจจุบันไม่ใช้บริการ กลุ่มที่ 2 เป็นลูกค้าเคยใช้บริการน้อยครั้ง และปัจจุบันไม่ใช้บริการ กลุ่มที่ 3 เป็นลูกค้าที่ใช้บริการสม่ำเสมอ และปัจจุบันยังใช้บริการสม่ำเสมอ กลุ่มที่ 4 เป็นลูกค้าที่ใช้บริการอยู่ แต่ทำธุรกรรมน้อย และกลุ่มที่ 5 เป็นลูกค้าที่ใช้บริการอยู่ ทำธุรกรรมบ่อย และใช้จ่ายเงินสูง

วิธีการที่ 2 นำมูลค่าของลูกค้าที่ได้จากการคำนวณ RFM มาจัดกลุ่มลูกค้าด้วยแบบจำลอง DBSCAN ซึ่งแบบจำลองนี้แบ่งลูกค้าเป็น 2 กลุ่ม และมีกลุ่ม Outlier 1 กลุ่มซึ่งผู้เขียนสังเกตเห็นว่าในกลุ่มที่เป็น Outlier เป็นกลุ่มลูกค้าที่มีลักษณะพฤติกรรมใช้จ่ายเงินสูง และทำธุรกรรมบ่อย จึงใช้แบบจำลอง K-Means ในการจัดกลุ่มลูกค้ากลุ่มOutlier อีกขั้นตอนหนึ่ง ผลการวิจัยของวิธีการที่ 2 จัดกลุ่มลูกค้าออกเป็น 4 กลุ่ม ได้แก่ กลุ่มที่ 1 เป็นลูกค้าที่ใช้บริการอยู่ ทำธุรกรรมบ่อย และใช้จ่ายเงินสูง กลุ่มที่ 2 เป็นลูกค้าที่ในอดีตทำธุรกรรมบ่อย และใช้จ่ายเงินสูง ปัจจุบันไม่ใช้บริการแล้ว กลุ่มที่ 3 เป็นลูกค้าที่ใช้บริการอยู่ และกลุ่มที่ 4 เป็นลูกค้าที่ไม่ใช้บริการแล้ว

วิธีการที่ 3 นำมูลค่าของลูกค้าที่ได้จากการคำนวณ RFM มาจัดกลุ่มลูกค้าด้วยแบบจำลอง Hierarchical Clustering ซึ่งแบบจำลองนี้จัดกลุ่มลูกค้าออกเป็น 4 กลุ่ม เหมือนกับแบบจำลอง K-Means แต่เนื่องจากเกิดปัญหาความซับซ้อนในการทำงานของแบบจำลอง จึงตัดสินใจไม่ทดลองวิธีการนี้ต่อ

ในขั้นตอนสุดท้าย งานวิจัยนี้เปรียบเทียบประสิทธิภาพของทั้ง 3 วิธีพบว่า

วิธีที่ 1 ใช้เทคนิค RFM และแบบจำลอง K-Means 2 ขั้นตอน แบบจำลองนี้ทำงานได้เร็วที่สุด

วิธีที่ 2 ใช้เทคนิค RFM, แบบจำลอง DBSCAN และแบบจำลอง K-Means การใช้แบบจำลอง DBSCAN ช่วยตรวจจับข้อมูลที่ผิดปกติได้ดี ทำให้เจอจุดข้อมูลที่น่าสนใจที่เป็นกลุ่มลูกค้าที่มีมูลค่าสูง

วิธีที่ 3 ใช้เทคนิค RFM และแบบจำลองHierarchical Clustering เป็นวิธีการที่ไม่มีประสิทธิภาพ เนื่องจากแบบจำลองไม่สามารถจัดการกับชุดข้อมูลขนาดใหญ่ได้

2.7.3 **บทความเรื่อง** Incorporating K-Means, Hierarchical Clustering and PCA in Customer Segmentation **โดย** Azad Abdulhafedh(Abdulhafedh, 2021)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าบัตรเครดิต เพื่อให้ธุรกิจเข้าใจลูกค้า และสามารถวางแผนกลยุทธการตลาดที่เฉพาะเจาะจงลงไปในแต่ละกลุ่ม โดยใช้ชุดข้อมูลสาธารณะ Kaggle.com ซึ่งเป็นข้อมูลพฤติกรรมการใช้บัตรเครดิตของลูกค้าที่เก็บในปี 2018 ประกอบไปด้วยข้อมูลจำนวน 8,950 แถวและ 18 แอททริบิวต์ โดยการทดลองแบ่งออกเป็น 2 รูปแบบ คือ รูปแบบที่ 1 จัดกลุ่มข้อมูลด้วยแบบจำลอง K-Means และAgglomerative และวัดประสิทธิภาพของแบบจำลองด้วย Silhouette score, Davies-Bouldin Index และDunn index รูปแบบที่ 2 ลดมิติของข้อมูลด้วยเทคนิคPCA (Principle Component Analysis)ก่อน จากนั้นนำชุดข้อมูลที่ผ่านการลดมิติด้วยPCA มาจัดกลุ่มด้วยแบบจำลองจัดกลุ่มK-Means และวัดประสิทธิภาพของแบบจำลองด้วย Silhouette score, Davies-Bouldin Index และDunn index และนำผลลัพธ์ของการทดลองทั้ง 2 รูปแบบมาเปรียบเทียบกัน

จากทดลองรูปแบบที่ 1 แบบจำลองที่ 1 K-Means หาจำนวนCluster ที่เหมาะสมด้วยElbow method ได้ค่า K เท่ากับ 3 และวัดประสิทธิภาพของ K-Means ได้ค่าSilhouette score 0.31,Davies-Bouldin Index 2.02 และDunn index 0.64

จากทดลองรูปแบบที่ 1 แบบจำลองที่ 2 Agglomerative หลังจากการทำAverage Linkage ได้จำนวนCluster เท่ากับ 3 และวัดประสิทธิภาพของ Agglomerative ได้ค่าSilhouette score 0.29,Davies-Bouldin Index 1.83 และDunn index 0.57

จากการทดลองรูปแบบที่ 2 โดยใช้เทคนิคPCAในการลดมิติของข้อมูลก่อนการจัดกลุ่ม โดยPCA ลดมิติของข้อมูล โดยการลดจำนวนฟีเจอร์ลงจาก 18 ฟีเจอร์เป็น 4 ฟีเจอร์ และเมื่อแสดงภาพการจัดกลุ่มโดยใช้เทคนิคPCA พบว่าข้อมูลแบ่งออกเป็น 4 กลุ่ม จากนั้นผู้เขียนทำการจัดกลุ่มข้อมูลอีกครั้งด้วยแบบจำลองK-Means โดยเลือกค่า K เท่ากับ 4 และวัดประสิทธิภาพของของK-Means ได้ค่า Silhouette score 0.34,Davies-Bouldin Index 2.21 และDunn index 0.71

จากผลการวิจัยสรุปได้ว่าแบบจำลอง K-Means มีประสิทธิภาพในการจัดกลุ่มข้อมูลมากกว่าแบบจำลองAgglomerative เมื่อทดลองกับข้อมูลชุดนี้ และพบว่าการใช้เทคนิคPCA ร่วมกับแบบจำลองจัดกลุ่มK-Means ทำให้ค้นพบCluster หรือ กลุ่มข้อมูลเพิ่มขึ้น เมื่อเปรียบเทียบกับการจัดกลุ่มโดยใช้เทคนิคK-Meansเพียงอย่างเดียว ทำให้คะแนน Silhouette score และDunn index สูงขึ้นเมื่อใช้เทคนิคPCA ร่วมกับแบบจำลองจัดกลุ่มK-Means

2.7.4 **บทความเรื่อง** A Machine Learning approach to Segment the Customers of Online Sales Data for Better and Efficient Marketing Purposes **โดย** Mathes T, Sumathy G **และ**Maheshwari A(Mathes T., 2023)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าเป้าหมาย เพื่อให้ธุรกิจสามารถวางแผนการตลาด และนำเสนอแคมเปญให้ตรงกับความต้องการของลูกค้าในแต่ละกลุ่ม โดยมีเป้าหมายเพื่อปิดการขาย และสร้างรายได้ให้กับธุรกิจ การทดลองใช้ชุดข้อมูลที่ซื้อมาจากบริษัทShrine โดยข้อมูลเก็บมาจาก 3 แหล่ง ได้แก่ ข้อมูลของลูกค้าภายในบริษัท ข้อมูลที่เก็บจากช่องทางออนไลน์ และข้อมูลจากบริษัทจัดหางาน ประกอบไปด้วยข้อมูลจำนวน 9,108 แถวและ15 แอททริบิวต์ โดยผู้เขียนได้ทำการจัดการข้อมูล และตัดข้อมูลออกคงเหลือ 3,107 แถวและ 4 แอททริบิวต์

ในการทดลองนี้ทำการจัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means, Agglomerative, Mean-Shift และDBSCAN และวัดประสิทธิภาพของแบบจำลองด้วย Silhouette score

จากการทดลองพบว่า แบบจำลอง K-Means แบ่งข้อมูลออกเป็น 2 กลุ่ม แบบจำลอง Agglomerative จัดกลุ่มข้อมูลออกเป็น 3 กลุ่ม แบบจำลอง Mean-Shift จัดกลุ่มข้อมูลออกเป็น 14 กลุ่มและแบบจำลอง DBSCAN จัดกลุ่มข้อมูลออกเป็น 17 กลุ่ม ซึ่งผู้เขียนให้ความเห็นว่าแบบจำลอง K-Means และแบบจำลอง Agglomerative ให้ผลลัพธ์การจัดกลุ่มที่ไม่เพียงพอต่อการนำไปใช้ในการวางแผนการตลาด เนื่องจากแบ่งกลุ่มข้อมูลจากฟีเจอร์รายได้ และแบบจำลอง

Mean-Shift ให้ผลลัพธ์การจัดกลุ่มที่เพียงพอต่อการนำไปใช้ในการวางแผนการตลาด เนื่องจากจัดกลุ่มข้อมูลจากฟีเจอร์อายุ และรายได้ฟีเจอร์รายได้ และในแบบจำลอง DBSCAN มีการจัดกลุ่มข้อมูลที่เป็น Outlier แยกเป็นลูกค้าอีกกลุ่มหนึ่ง

เมื่อวัดประสิทธิภาพของแบบจำลองด้วย Silhouette score พบว่าแบบจำลองที่มีคะแนนสูง ได้แก่ แบบจำลอง Mean-Shift และแบบจำลอง Agglomerative โดยผู้เขียนให้ความเห็นในบทสรุปงานวิจัยว่า Silhouette score เป็นเมตริกเพื่อวัดประสิทธิภาพของแบบจำลอง แต่ในโลกความเป็นจริงการจัดกลุ่มลูกค้าออกเป็นกลุ่มย่อยหลายกลุ่มตอบโจทย์ต่อการวางแผนการตลาดแบบเฉพาะเจาะจง ดังนั้น แบบจำลองที่เหมาะกับการแบ่งกลุ่มลูกค้าในชุดข้อมูลนี้ คือ แบบจำลอง Mean-Shift และแบบจำลอง DBSCAN

2.7.5 **บทความเรื่อง** Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce **โดย** Ritu Punhani, V.P.S Arora, Sai Sabitha **และ**Vinod Kumar Shukla(Ritu Punhani, 2021)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้า ในเวปไซต์E-Commerce เพื่อให้ธุรกิจวางแผลกลยุทธการตลาดที่เฉพาะเจาะจงลงไปในแต่ละกลุ่ม ช่วยให้ธุรกิจเข้าใจลูกค้า เพื่อสร้างความสัมพันธ์กับลูกค้า สามารถวางแผนการสต็อคสินค้า และสร้างยอดขายให้กับธุรกิจE-Commerce

การทดลองใช้ชุดข้อมูลสาธารณะ Kaggle.com ซึ่งเป็นข้อมูลการซื้อสินค้าในเวปไซต์E-Commerce ประกอบไปด้วยข้อมูลจำนวน 25,000 แถวและ 7 แอททริบิวต์ ใช้โปรแกรม RapidMiner ในการจัดการข้อมูลโดยใช้เทคนิค K-Means ในการจัดกลุ่มลูกค้า และวัดประสิทธิภาพการจัดกลุ่มลูกค้าด้วย Davies-Bouldin index ซึ่งในงานวิจัยนี้เลือกจำนวนCluster เท่ากับ 4 กลุ่ม เนื่องจากมีคะแนนDavie-Bouldin index น้อยที่สุด แสดงถึงClusterแต่ละกลุ่มมีการแยกตัวห่างกันอย่างชัดเจน และเมื่อพิจารณาผลลัพธ์จากการแบ่งกลุ่ม พบข้อมูลเชิงลึกว่าสินค้าช่วง 503-505 เป็นสินค้าที่มียอดขายสูงสุด และสินค้าช่วง 512-513 เป็นสินค้าที่มียอดขายต่ำที่สุด นอกจากนี้ยังพบว่าลูกค้าส่วนใหญ่นิยมชำระค่าสินค้าด้วยบัตรเครดิตคิดเป็น 57.8% และรองลงมาเป็นชำระค่าสินค้าด้วยPayPalคิดเป็น 42.2%

2.7.6 บทความเรื่อง Comparative Analysis of The Application of Five Clustering Algorithms for Market Segmentation โดย Denys Teslenko, Sorokina Kyrylo Smelyakpv และOleksii Filipov(Denys Teslenko, 2023)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าโรงแรมโดยเปรียบเทียบอัลกอริธึม 5 แบบ ได้แก่ K-Means, BIRCH, Agglomerative, DBSCAN และOPTICS เพื่อประกอบการตัดสินใจว่าเทคนิคประเภทใดเหมาะสำหรับการจัดกลุ่มลูกค้า การทดลองใช้ชุดข้อมูลลูกค้าจากโรงแรม 4 ดาว เมืองลิสบอน ประเทศโปรตุเกส ประกอบไปด้วยข้อมูลจำนวน 83,905 แถวและ 31 แอททริบิวต์

การทดลองเริ่มต้นด้วยการเตรียมข้อมูล และเนื่องจากชุดข้อมูลนี้ประกอบไปด้วยฟีเจอร์ประเภทตัวเลข (Numerical feature) และฟีเจอร์ประเภทหมวดหมู่(Categorical feature) จึงทำการแปลงฟีเจอร์ประเภทหมวดหมู่เป็นฟีเจอร์ประเภทตัวเลขด้วยวิธี One-Hot encoding และเพื่อให้การคำนวณเร็วขึ้น งานวิจัยนี้ใช้เทคนิคRandom selection ในการลดข้อมูลจาก 83,950 แถวคงเหลือเป็น 2,000 แถว ในขั้นตอนการแบ่งกลุ่ม งานวิจัยนี้ใช้ DBSCANในการหาจำนวนCluster ที่เหมาะสม ได้จำนวนClusterที่เหมาะสมเท่ากับ 4 โดยในการทดลอง ให้แบบจำลองจัดกลุ่มที่ Cluster เท่ากับ 3,4 และ5 ด้วยอัลกอริธึม K-Means, BIRCH, Agglomerative, DBSCAN และOPTICS เมื่อวัดประสิทธิภาพของอัลกอริธึมด้วย Davies-Bouldin index พบว่า Agglomerative มีประสิทธิภาพดีที่สุด โดยมีค่า Davies-Bouldin index เมื่อจำนวนCluster เท่ากับ 3,4 และ5 ที่ 0.152, 0.367 และ0.537 ตามลำดับ และอัลกอริธึม DBSCAN และOPTICS มีประสิทธิภาพต่ำที่สุดเมื่อเปรียบเทียบกับอัลกอริธึมอื่น และเมื่อวัดประสิทธิภาพอัลกอริธึมด้วย Silhouette score พบว่า Agglomerative มีประสิทธิภาพดีที่สุด โดยมีค่า Silhouette score เมื่อจำนวนCluster เท่ากับ 3,4 และ5 ที่ 0.888, 0.705 และ0.593 ตามลำดับ และอัลกอริธึม DBSCAN และOPTICS มีประสิทธิภาพต่ำที่สุดเมื่อเปรียบเทียบกับอัลกอริธึมอื่นเช่นกัน

เพื่อเพิ่มประสิทธิภาพของอัลกอริธึม DBSCAN และOPTICS งานวิจัยนี้จึงทดลองใช้ Gower's distance แทน Euclidean distance จากการทดลองพบว่าอัลกอริธึม DBSCAN มีค่า Silhouette scoreสูงขึ้นเมื่อจำนวนCluster เท่ากับ 3,4 และ5 ที่ 0.145 (ค่าเดิม -0.457), 0.107 (ค่าเดิม -0.540) และ0.073 (ค่าเดิม -0.558) ตามลำดับ

สรุปผลงานวิจัยนี้เปรียบเทียบประสิทธิภาพของอัลกอริธึม 5 แบบกับชุดข้อมูลลูกค้าโรงแรม โดยอัลกอริธึมที่มีประสิทธิภาพมากที่สุด คือ อัลกอริธึม Agglomerative รองลงมาคือ BIRCH และK-Means และอัลกอริธึมที่มีประสิทธิภาพต่ำที่สุด คือ อัลกอริธึมประเภทDensity-based อย่างDBSCAN และOPTICS เพื่อปรับปรุงประสิทธิภาพของอัลกอริธึมประเภท Density-

based การใช้Gower's distance ช่วยให้อัลกอริธึม DBSCAN และOPTICS มีประสิทธิภาพดีขึ้น ซึ่งทั้งหมดนี้ขึ้นอยู่กับแต่ละธุรกิจจะพิจารณานำเทคนิคที่เหมาะสมไปประยุกต์ใช้

2.7.7 **บทความเรื่อง** Educational Data Mining Using Cluster Analysis and Decision Tree Technique: A Case Study **โดย** Snjezana Krizanic(Kri˘zanic, 2020)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลักษณะพฤติกรรมของนักเรียนในการเรียนผ่านระบบออนไลน์เพื่อค้นหารูปแบบของพฤติกรรมที่ส่งผลต่อคะแนนสอบของนักเรียน เพื่อปรับปรุงเนื้อหาวิชาที่สอน และเพื่อให้ครูผู้สอนปรับเทคนิคในการสอน โดยจัดกลุ่มข้อมูลด้วยโปรแกรม RapidMiner ใช้อัลกอริธึม K-Means ในการจัดกลุ่ม และอัลกอริธึม Decision Tree ในการวิเคราะห์กลุ่ม การทดลองใช้ชุดข้อมูลการเรียนผ่านระบบออนไลน์ของนักเรียนระดับอุดมศึกษาในประเทศโครเอเชีย เก็บข้อมูลในเดือนกุมภาพันธ์ถึงมิถุนายนในปี 2018 โดยชุดข้อมูลเป็นนักเรียนที่เข้าเรียนในระบบออนไลน์จำนวน 185 คนและ4 แอททริบิวต์ และมีการสอบกลางภาค 2 ครั้ง ซึ่งในแต่ละครั้งมีคะแนนเต็ม 40 คะแนน

จากการทดลองอัลกอริธึม K-Means แบ่งนักเรียนออกเป็น 3 กลุ่ม ได้แก่ กลุ่มที่ศูนย์จำนวน 84 คน กลุ่มที่หนึ่งจำนวน 82 คน และกลุ่มที่สองจำนวน 19 คน จากนั้นนำแต่ละกลุ่มเข้าอัลกอริธึม Decision Tree เพื่อสังเกตลักษณะพฤติกรรมของนักเรียนในแต่ละกลุ่ม โดย Decision Tree อธิบายลักษณะของนักเรียนแต่ละกลุ่มดังนี้

กลุ่มที่ศูนย์ นักเรียนในกลุ่มนี้ส่วนใหญ่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียนน้อย และได้คะแนนสอบกลางภาคน้อย และมีนักเรียนส่วนน้อยในกลุ่มนี้เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียน และกระดานพูดคุยบ่อย และนักเรียนที่ได้คะแนนสอบกลางภาคสูงในกลุ่มนี้เป็นนักเรียนที่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียน และแบบฝึกหัดในการเรียนแลปบ่อย

กลุ่มที่หนึ่ง นักเรียนในกลุ่มนี้ผสมผสานระหว่างนักเรียนที่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียน และแบบฝึกหัด นักเรียนที่ได้คะแนนสอบกลางภาคสูงในกลุ่มนี้เป็นนักเรียนที่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียน แบบฝึกหัด และแบบฝึกหัดในการเรียนแลปบ่อย และนักเรียนที่ได้คะแนนสอบกลางภาคน้อยในกลุ่มนี้เป็นนักเรียนที่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียนน้อย

กลุ่มที่สอง เป็นนักเรียนที่เข้าไปในระบบออนไลน์บ่อยและได้คะแนนสอบกลางภาคสูง

ผลการวิจัยสรุปว่า นักเรียนที่เข้าระบบเรียนออนไลน์ ที่เข้าไปในระบบเพื่อดาวน์โหลดเอกสารในการเรียน แบบฝึกหัด และแบบฝึกหัดในการเรียนแลปเป็นนักเรียนที่ได้คะแนนสอบกลางภาคสูง

2.7.8 **บทความเรื่อง** Customer Behavior Mining Framework (CBMF) Using Clustering and Classification Techniques **โดย** Farshid Abdi **และ**Shanghayegh Abolmakarem(Farshid Abdi, 2019)

งานวิจัยนี้นำเสนอวิธีการจัดกลุ่มลูกค้าในธุรกิจโทรคมนาคม เพื่อศึกษาพฤติกรรมของลูกค้าและนำข้อมูลมาพัฒนาระบบลูกค้าสัมพันธ์ เพื่อสร้างความพึงพอใจให้กับลูกค้า และรักษาลูกค้าให้อยู่กับธุรกิจ ในการทดลองใช้ชุดข้อมูลลูกค้าของธุรกิจโทรคมนาคมจำนวน 1,000 แถว 25 แอททริวต์ (แบ่งเป็น 24 แอททริบิวต์ และ1 เลเบล) โดยมีลูกค้าเลเบลChurned จำนวน 274 คนและลูกค้าเลเบลNon-churned จำนวน 726 คน

โดยงานวิจัยนี้แบ่งการทดลองออกเป็น 2 ช่วง คือ การทดลองช่วงที่หนึ่ง จัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means และวัดประสิทธิภาพของแบบจำลองด้วย Davies-Bouldin index และจัดลำดับความน่าสนใจของลูกค้าในแต่ละกลุ่ม และการทดลองช่วงที่สอง ทำนายลำดับความน่าสนใจของลูกค้า และทำนายการเลิกใช้บริการซึ่งแบ่งการเลิกใช้บริการออกเป็น 2 ช่วง คือ ช่วงที่หนึ่งทำนายการเลิกใช้บริการจากฟีเจอร์ลักษณะทางประชากรศาสตร์ และช่วงที่สองทำนายการเลิกใช้บริการจากฟีเจอร์พฤติกรรมของลูกค้า โดยทำนายหลังจากทำนายช่วงที่หนึ่งไปแล้ว 1 เดือน โดยใช้แบบจำลอง Neural Network และDecision Tree ในการทำนาย และวัดประสิทธิภาพของแบบจำลองด้วยค่าAccuracy, Precision, Recall และConfusion matrix

ผลการวิจัย ในการทดลองช่วงที่หนึ่ง จัดกลุ่มลูกค้าด้วยแบบจำลองK-Means แบ่งลูกค้าออกเป็น 6 กลุ่ม มีคะแนนDavies-Bouldin 0.88 โดยมีลูกค้ากลุ่มที่ศูนย์จำนวน 103 คน ลูกค้ากลุ่มที่หนึ่งจำนวน 144 คน ลูกค้ากลุ่มที่สองจำนวน 38 คน ลูกค้ากลุ่มที่สามจำนวน 58 คน ลูกค้ากลุ่มที่สี่จำนวน 251 คน และลูกค้ากลุ่มที่ห้าจำนวน 363 คน

โดยระดับความน่าสนใจของลูกค้าแบ่งออกเป็น 2 ระดับ คือ ระดับที่หนึ่ง ความน่าสนใจระดับปานกลาง คือ ลูกค้าที่มีค่าเฉลี่ยจำนวนชั่วโมงในการใช้งาน และจำนวนบริการที่ใช้ระดับปานกลาง และระดับที่สอง ความน่าสนใจระดับสูง คือ ลูกค้าที่มีค่าเฉลี่ยจำนวนชั่วโมงในการใช้งาน และจำนวนบริการที่ใช้ระดับสูง ซึ่งลูกค้าที่มีความน่าสนใจระดับปานกลาง ได้แก่ ลูกค้ากลุ่มที่ศูนย์ และลูกค้ากลุ่มที่ห้า และลูกค้าที่มีความน่าสนใจระดับสูง ได้แก่ ลูกค้ากลุ่มที่หนึ่ง ลูกค้ากลุ่มที่สอง ลูกค้ากลุ่มที่สาม และลูกค้ากลุ่มที่สี่

ผลการวิจัยในการทดลองช่วงที่สอง ช่วงทำนายระดับความน่าสนใจของลูกค้า พบว่าแบบจำลองNeural Network มีค่าAccuracy 98.61%, ค่าRecall 96.81% และค่าPrecision 97.87% และแบบจำลองDecision Tree มีค่าAccuracy 98.26%, ค่าRecall 97.63% และค่าPrecision 96.69%

ช่วงทำนายการเลิกใช้บริการครั้งที่หนึ่ง โดยพิจารณาฟีเจอร์ลักษณะทางประชากรศาสตร์ พบว่าแบบจำลองNeural Network มีค่าAccuracy 67.60%, ค่าRecall 80.19% และค่าPrecision 76.15% และแบบจำลองDecision Tree มีค่าAccuracy 69.00%, ค่าRecall 89.90% และค่าPrecision 73.20%

ช่วงทำนายการเลิกใช้บริการครั้งที่สอง โดยพิจารณาฟีเจอร์พฤติกรรมของลูกค้าหลังจากการทำนายครั้งที่หนึ่งไปแล้ว 1 เดือน พบว่าแบบจำลองNeural Network มีค่าAccuracy 75.61%, ค่าRecall 86.45% และค่าPrecision 81.86% และแบบจำลองDecision Tree มีค่าAccuracy 74.22%, ค่าRecall 88.89% และค่าPrecision 73.30% และในขั้นตอนสุดท้ายผู้เขียนแนะนำเทคนิคการจัดการลูกค้าสัมพันธ์ให้กับลูกค้าในกลุ่มน่าสนใจ และลูกค้าในกลุ่มที่มีพฤติกรรมจะเลิกใช้บริการ

บทที่ 3

การดำเนินการวิจัย

ในงานวิจัยนี้ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

3.1 กระบวนการทำงานของแบบจำลอง

3.2 การเก็บรวบรวมข้อมูล และตรวจสอบข้อมูล

3.3 การสำรวจข้อมูล (Exploratory Data Analysis)

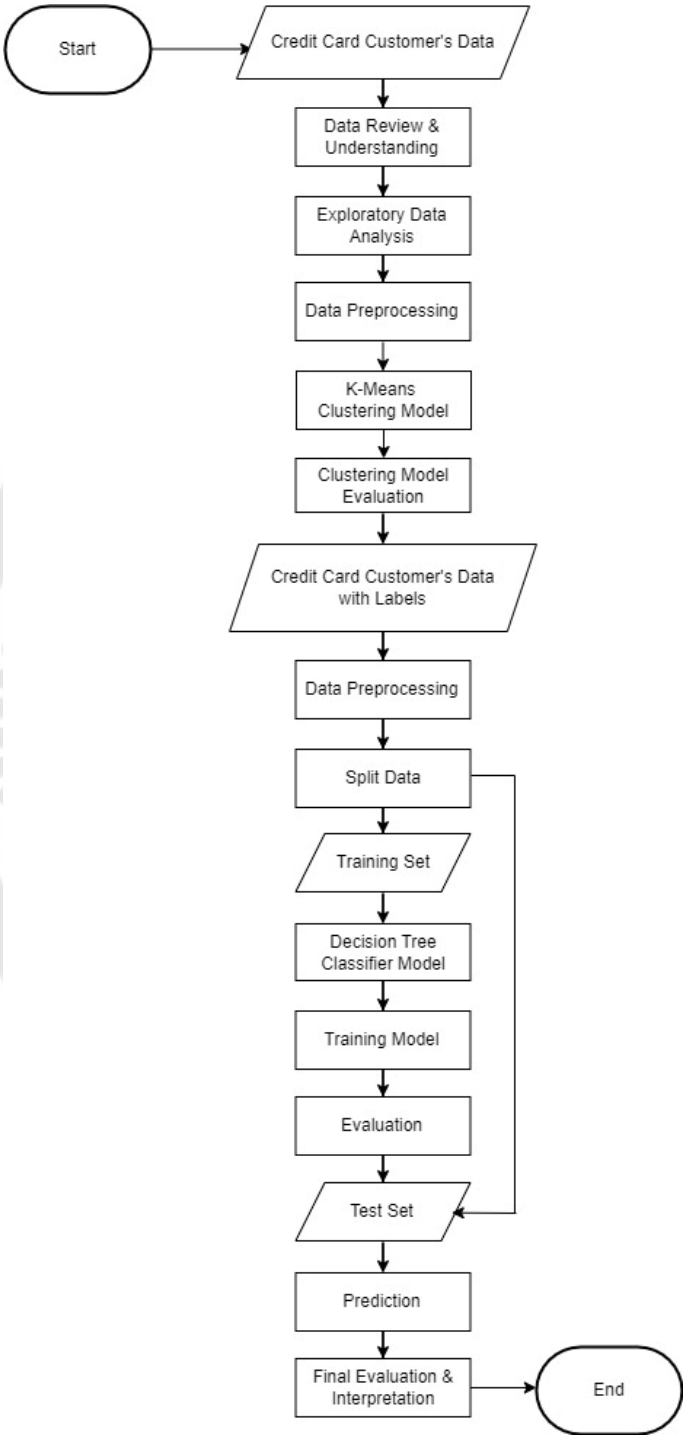
3.4 การเตรียมข้อมูล (Data Preprocessing)

3.5 การสร้างแบบจำลองจัดกลุ่มลูกค้า (Clustering Model)

3.6 การวัดผลลัพธ์จากแบบจำลองจัดกลุ่มลูกค้า

3.7 การสร้างแบบจำลองเพื่อทำนายกลุ่มลูกค้า (Classification Model)

3.1 กระบวนการทำงานของแบบจำลอง



ภาพประกอบ 4 แสดงกระบวนการทำงานของแบบจำลองในงานวิจัย

จากภาพประกอบที่ 4 แสดงถึงกระบวนการทำงานของแบบจำลอง เริ่มต้นจากนำเข้าชุดข้อมูลลูกค้าบัตรเครดิต จากนั้นทำเข้าใจข้อมูล และตรวจสอบข้อมูล Exploratory Data Analysis เพื่อศึกษาข้อมูลเชิงลึก จากนั้นทำการจัดเตรียมข้อมูล หรือ Data Preprocessing เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมต่อการทำงานของแบบจำลอง

ขั้นตอนถัดมา คือ การ ใช้แบบจำลอง K-Means ในการจัดกลุ่มข้อมูล จากนั้นวัดผลประสิทธิภาพของแบบจำลองการจัดกลุ่มข้อมูล และนำชุดข้อมูลที่มีเลเบลไปทำนายในแบบจำลองถัดไป

เข้าสู่กระบวนการของแบบจำลองทำนายกลุ่ม โดยเริ่มต้นจากการจัดเตรียมข้อมูลอีกครั้งเพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมต่อการทำงานของแบบจำลอง Decision Tree Classification จากนั้นทำการแบ่งข้อมูลออกเป็นสองชุดคือ ข้อมูลสำหรับการเรียนรู้ของแบบจำลองหรือ Training Set และข้อมูลสำหรับการทดสอบประสิทธิภาพของแบบจำลองหรือ Test Set เพื่อประสิทธิภาพของแบบจำลอง จึงใช้การปรับพารามิเตอร์ร่วมด้วย เมื่อแบบจำลองทำการเรียนรู้เสร็จเรียบร้อย จึงนำข้อมูลชุดทดสอบหรือ Test Set มาใช้วัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และConfusion Matrix นอกจากนี้ยังวิเคราะห์และตีความหมายของแบบจำลองด้วย Feature Importance

3.2 การเก็บรวบรวมข้อมูล และการตรวจสอบข้อมูล

3.2.1 การเก็บรวบรวมข้อมูล

ในงานวิจัยนี้ใช้ข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com ประกอบไปด้วย 10 แอททริบิวต์ และมีจำนวนข้อมูลทั้งหมด 10,127 แถว ดังภาพประกอบที่ 5 และแสดงแอททริบิวต์ของข้อมูลดังตารางที่ 2

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

    Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

[ ] df = pd.read_csv("/content/drive/MyDrive/Thitiporn_IS_Project/Dataset/creditcardusage.csv")

[ ] df.head()
```

	Age	Gender	Education_Level	Marital_Status	Months_on_book	Credit_Limit	Total_Trans_Amt	Total_Trans_Count	Minimum_income	Max_income
0	45	M	High School	Married	39	12691	1144	42	60000	80000
1	49	F	Graduate	Single	44	8256	1291	33	0	40000
2	51	M	Graduate	Married	36	3418	1887	20	80000	120000
3	40	F	High School	Unknown	34	3313	1171	20	0	40000
4	40	M	Uneducated	Married	21	4716	816	28	60000	80000

ภาพประกอบ 5 แสดงตัวอย่างข้อมูลห้าบรรทัดแรก

ตาราง 2 แสดงแอททริบิวต์ของข้อมูล

ลำดับ	ชื่อแอททริบิวต์	ข้อมูลภายในแอททริบิวต์	คำอธิบาย
1	Age	ปี	อายุของลูกค้า
2	Gender	M/F	เพศของลูกค้า
3	Education_Level	High School/Graduate/ College/Post-Graduate/ Doctorate/Uneducated/ Unknown	ระดับการศึกษาของลูกค้า
4	Marital_Status	Single/Married/Divorced/ Unknown	สถานภาพของลูกค้า
5	Months_on_book	เดือน	ระยะเวลาที่ลูกค้าถือบัตรเครดิต
6	Credit_Limit	จำนวนเงิน	วงเงินบัตรเครดิต
7	Total_Trans_Amount	จำนวนเงิน	ยอดเงินที่ใช้จ่ายบัตรเครดิต
8	Total_Trans_Count	จำนวนครั้ง	จำนวนครั้งของการใช้จ่ายบัตรเครดิต
9	Minimum_Income	จำนวนเงิน	รายได้ขั้นต่ำที่ลูกค้าได้รับ
10	Max_Income	จำนวนเงิน	รายได้สูงสุดที่ลูกค้าได้รับ

3.2.2 ตรวจสอบข้อมูลและพิจารณาข้อมูลมาใช้ในการวิเคราะห์

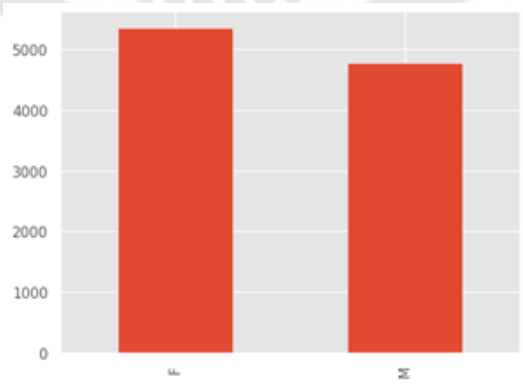
งานวิจัยนี้ใช้ภาษาไพธอน (Python) ในการวิเคราะห์ข้อมูล และสร้างแบบจำลอง เริ่มต้นด้วยการนำเข้าโมดูลสำหรับสร้างแบบจำลอง และนำไฟล์ข้อมูลที่ใช้สำหรับสร้างแบบจำลอง ดังภาพประกอบที่ 6

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.cluster import KMeans
    from kmodes.kprototypes import KPrototypes
    from sklearn.preprocessing import MinMaxScaler
    from sklearn.metrics import davies_bouldin_score
    from sklearn.metrics import silhouette_score
    from sklearn.metrics import calinski_harabasz_score

    # print the graphs in the notebook
    %matplotlib inline

    # set seaborn style to white
    sns.set_style("white")
    sns.set_theme()
    import plotly.express as px

    import warnings
    warnings.filterwarnings("ignore")
```

ภาพประกอบ 6 แสดงตัวอย่างโค้ดการนำเข้าโมดูลที่ใช้สำหรับการสร้างแบบจำลอง

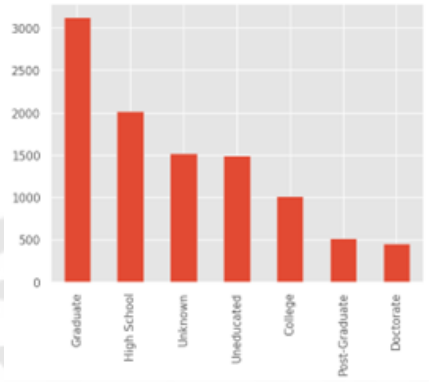
3.3 การสำรวจข้อมูล (Exploratory Data Analysis)

การสำรวจข้อมูลเป็นขั้นตอนที่สำคัญ เนื่องจากช่วยให้เข้าใจข้อมูลเชิงลึกของลูกค้าบัตรเครดิต โดยเริ่มต้นจากการสำรวจฟีเจอร์ Gender พบว่าจากลูกค้าทั้งหมด 10,127 คน แบ่งเป็นเพศหญิง 5,358 คน และเพศชาย 4,769 คน จะเห็นได้ว่าลูกค้าเพศชายมีจำนวนน้อยกว่าลูกค้าเพศหญิง ดังภาพประกอบที่ 7



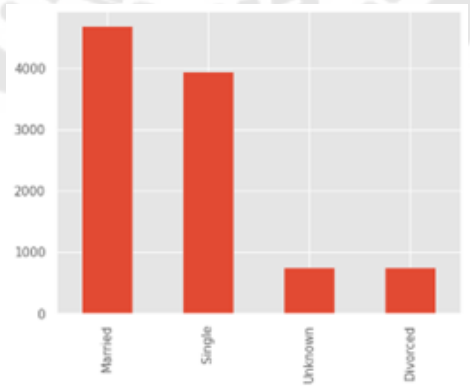
ภาพประกอบ 7 แสดงจำนวนลูกค้าเพศหญิงและเพศชาย

สำหรับฟีเจอร์ Education_Level แบ่งออกเป็น 7 ประเภท ได้แก่ High School 2,013 คน, Graduate 3128 คน, College 1,013 คน, Post-Graduate 516 คน, Doctorate 451 คน, Uneducated 1,487 คน และUnknown 1,519 คน จะเห็นได้ว่าลูกค้าส่วนใหญ่จบการศึกษาระดับปริญญาตรีและมัธยมศึกษาตอนปลาย ดังภาพประกอบที่ 8



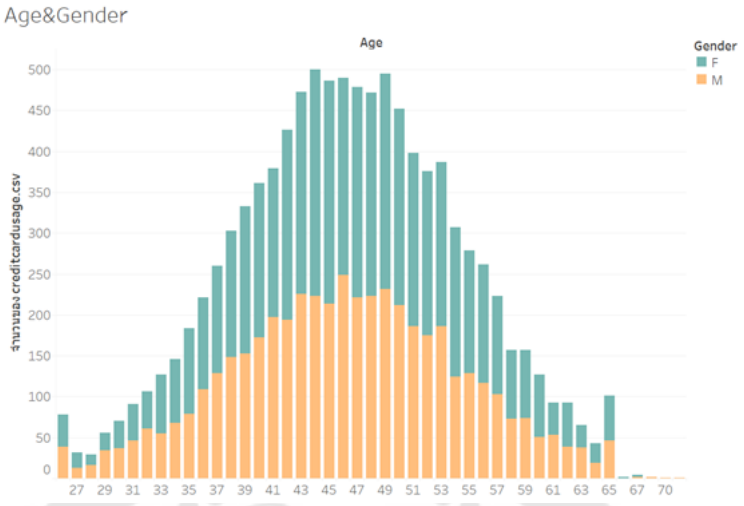
ภาพประกอบ 8 แสดงระดับการศึกษาของลูกค้า

สำหรับฟีเจอร์ Marital_Status แบ่งออกเป็น 4 ประเภท ได้แก่ Single 3,943 คน, Married 4,687 คน, Divorced 748 คน และUnknown 749 คน จะเห็นได้ว่าลูกค้าส่วนใหญ่อยู่ในสถานภาพโสดและแต่งงาน ดังภาพประกอบที่ 9



ภาพประกอบ 9 แสดงสถานภาพของลูกค้า

ในภาพประกอบที่ 10 แสดงฟีเจอร์ Age ที่ตรวจสอบจำนวนลูกค้าบัตรเครดิต พบว่ามีอายุตั้งแต่ 26-73 ปี ลูกค้าส่วนใหญ่มีอายุอยู่ในช่วง 38-54 ปี

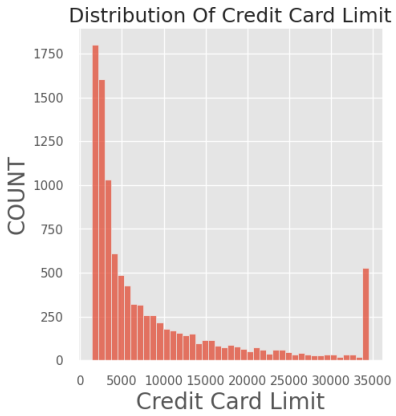


ภาพประกอบ 10 แสดงการกระจายตัวของอายุของลูกค้า

ในภาพประกอบที่ 11 แสดงฟีเจอร์ Months_on_book ที่ตรวจสอบจำนวนเดือนที่ลูกค้าถือบัตรเครดิต พบว่าลูกค้าถือบัตรเครดิตตั้งแต่ 13-56 เดือน และลูกค้าส่วนใหญ่ถือบัตรเครดิตเป็นระยะเวลา 36 เดือน

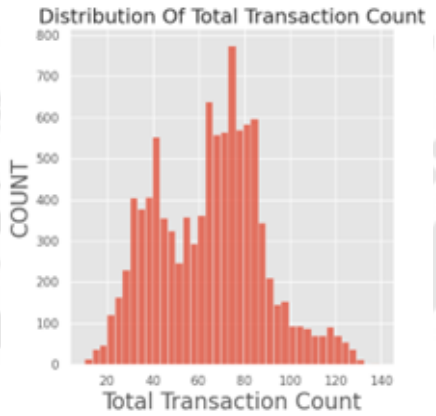
ภาพประกอบ 11 แสดงการกระจายตัวของระยะเวลาที่ลูกค้าถือบัตรเครดิต

ในภาพประกอบที่ 12 แสดงฟีเจอร์ Credit_Limit ที่ตรวจสอบวงเงินบัตรเครดิตของลูกค้า พบว่าลูกค้ามีวงเงินบัตรเครดิตตั้งแต่ 1,000-35,000 และลูกค้าส่วนใหญ่มีวงเงินบัตรเครดิต 1,000-3,500 และ34,500-35,000



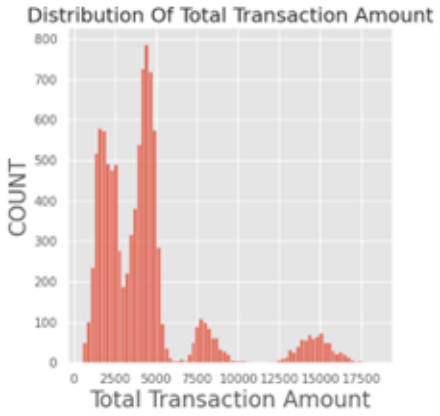
ภาพประกอบ 12 แสดงการกระจายตัวของวงเงินบัตรเครดิตที่ลูกค้าได้รับ

ในภาพประกอบที่ 13 แสดงฟีเจอร์ Total_Trans_Count ที่ตรวจสอบจำนวนการใช้บัตรเครดิต พบว่าลูกค้าใช้จ่ายบัตรเครดิตจำนวน 10-139 ครั้ง ซึ่งลูกค้าส่วนใหญ่ใช้บัตรเครดิตจำนวน 60-79 ครั้ง และลำดับรองลงมาจำนวน 40-59 ครั้ง



ภาพประกอบ 13 แสดงการกระจายตัวของจำนวนครั้งที่ลูกค้าใช้บัตรเครดิต

ในภาพประกอบที่ 14 แสดงฟีเจอร์ Total_Trans_Amt ที่ตรวจสอบยอดใช้จ่ายบัตรเครดิต พบว่าลูกค้าใช้จ่ายบัตรเครดิตตั้งแต่ 510-18,484 ซึ่งลูกค้าส่วนใหญ่ใช้จ่ายบัตรเครดิตในยอด 1,100-2,220 และลำดับรองลงมาในยอด 3,700-4,810



ภาพประกอบ 14 การกระจายตัวของยอดใช้จ่ายบัตรเครดิต

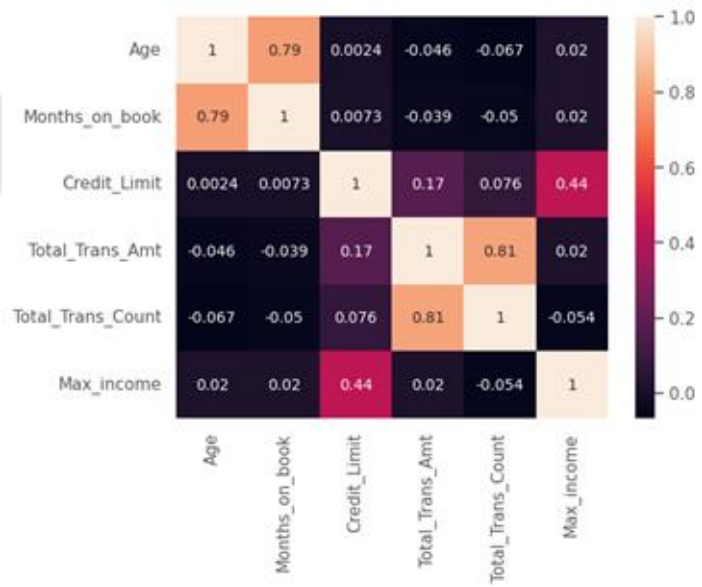
ในภาพประกอบที่ 15 แสดงฟีเจอร์ Minimum_Income ที่ตรวจสอบรายได้ขั้นต่ำของลูกค้า พบว่าลูกค้ามีรายได้ขั้นต่ำตั้งแต่ 0-120,000 ซึ่ง 0 ในชุดข้อมูลนี้แสดงถึงไม่ทราบข้อมูลรายได้ ซึ่งข้อมูลส่วนใหญ่ไม่มีข้อมูลรายได้ขั้นต่ำของลูกค้า ดังนั้น ผู้วิจัยจึงพิจารณาตัดฟีเจอร์นี้ออก เนื่องจากไม่แสดงถึงข้อมูลเชิงลึก

ภาพประกอบ 15 การกระจายตัวของรายได้ขั้นต่ำของลูกค้า

ในภาพประกอบที่ 16 แสดงฟีเจอร์ Max_Income ที่ตรวจสอบรายได้สุงสุดที่ลูกค้าได้รับ พบว่าลูกค้ามีรายได้สูงสุดตั้งแต่ 0-120,000 ซึ่ง 0 ในชุดข้อมูลนี้แสดงถึงไม่ทราบข้อมูลรายได้ ซึ่งลูกค้าส่วนใหญ่มีรายได้สูงสุด 40,000-120,000

ภาพประกอบ 16 การกระจายตัวของรายได้สูงสุดที่ลูกค้าได้รับ

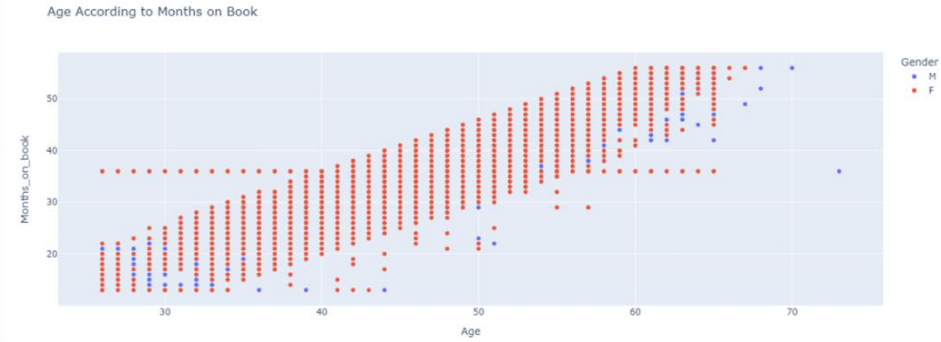
ในภาพประกอบที่ 17 แสดงการวิเคราะห์ความสัมพันธ์ระหว่างฟีเจอร์ พบว่ามีความสัมพันธ์ในทางบวกระหว่างฟีเจอร์ Ageและ Months_on_book 79%, ฟีเจอร์ Total_Trans_Amtและ Total_Trans_Count 81%และฟีเจอร์ Credit_Limitและ Max_Income 44%



ภาพประกอบ 17 แสดงการวิเคราะห์ความสัมพันธ์ของแต่ละฟีเจอร์

จากผลการวิเคราะห์ความสัมพันธ์ของฟีเจอร์ ผู้วิจัยทำการสำรวจข้อมูลเชิงลึกเพิ่ม จากภาพประกอบที่ 18 แสดงความสัมพันธ์ของฟีเจอร์ Ageและ Months_on_book พบว่า อายุมี

ความสัมพันธ์กับระยะเวลาในการถือบัตรเครดิต ยิ่งอายุมากขึ้น ยิ่งถือบัตรเครดิตเป็นระยะเวลานาน ซึ่งหมายความว่าลูกค้าบัตรเครดิตส่วนใหญ่มีความจงรักภักดีกับธนาคาร เนื่องจากยังคงถือบัตรเครดิตต่อเนื่อง และลูกค้าที่ถือบัตรเครดิตเป็นระยะเวลานานส่วนใหญ่เป็นเพศหญิง

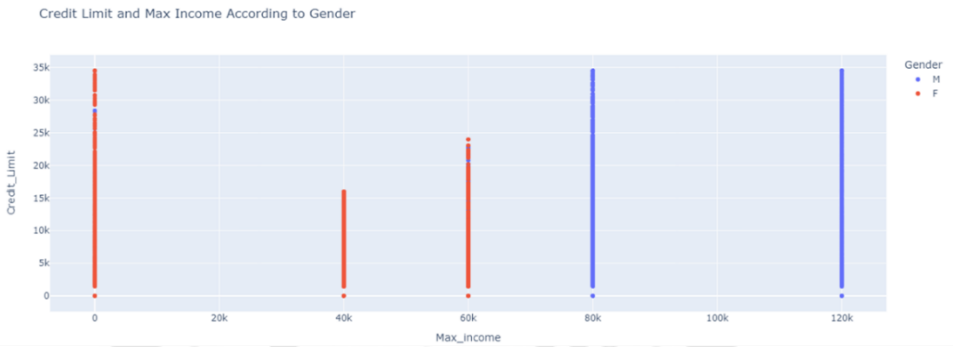


ภาพประกอบ 18 แสดงความสัมพันธ์ของฟีเจอร์ AgeและMonths_on_book

จากภาพประกอบที่ 19 แสดงความสัมพันธ์ของฟีเจอร์ Total_Trans_Amtและ Total_Trans_Count พบว่าจำนวนครั้งของการใช้บัตรเครดิตมีความสัมพันธ์กับยอดใช้จ่ายบัตรเครดิต จำนวนการใช้ที่มากขึ้นส่งผลถึงยอดใช้จ่ายที่เพิ่มขึ้น และยังพบว่ากลุ่มลูกค้าที่ใช้บัตรเครดิตจำนวนครั้งและยอดใช้จ่ายมากเป็นลูกค้าเพศชาย และลูกค้าเพศหญิงส่วนใหญ่มียอดใช้จ่ายบัตรเครดิตน้อย

ภาพประกอบ 19 แสดงการวิเคราะห์ความสัมพันธ์ของฟีเจอร์ Total_Trans_Amtและ Total_Trans_Count

จากภาพประกอบที่ 20 แสดงความสัมพันธ์ของฟีเจอร์ Credit_LimitและMax_Income พบว่าวงเงินบัตรเครดิตมีความสัมพันธ์กับรายได้สูงสุดที่ลูกค้าได้รับ กลุ่มลูกค้าที่มีรายได้สูง จะได้รับวงเงินบัตรเครดิตสูงส่วนใหญ่เป็นลูกค้าเพศชาย และเพศหญิงที่ไม่มีข้อมูลรายได้ได้รับวงเงินบัตรเครดิตสูงเช่นกัน



ภาพประกอบ 20 แสดงการวิเคราะห์ความสัมพันธ์ของฟีเจอร์ Credit_LimitและMax_Income

3.4 **การเตรียมข้อมูล** (Data Preprocessing)

เพื่อให้แบบจำลองสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ ในงานวิจัยนี้ทำการเตรียมข้อมูล 2 ขั้นตอน ได้แก่

3.4.1 **การแปลงค่าประเภทของข้อมูล** (Encoding)

เนื่องจากชุดข้อมูลมีฟีเจอร์ประเภทหมวดหมู่ ได้แก่ Gender, Education_Level และ Marital_Status จึงใช้วิธีแปลงค่าฟีเจอร์ประเภทหมวดหมู่เป็นค่าตัวเลขด้วยฟังก์ชั่น Factorize ดังภาพประกอบที่ 21

```
[ ] df['Gender'] = pd.factorize(df['Gender'])[0]
    df['Education_Level'] = pd.factorize(df['Education_Level'])[0]
    df['Marital_Status'] = pd.factorize(df['Marital_Status'])[0]

[ ] #Preview dataset again
    df.head()
```

	Age	Gender	Education_Level	Marital_Status	Months_on_book	Credit_Limit	Total_Trans_Amt	Total_Trans_Count	Minimum_income	Max_income
0	45	0	0	0	39	12691	1144	42	60000	80000
1	49	1	1	1	44	8256	1291	33	0	40000
2	51	0	1	0	36	3418	1887	20	80000	120000
3	40	1	0	2	34	3313	1171	20	0	40000
4	40	0	2	0	21	4716	816	28	60000	80000

ภาพประกอบ 21 แสดงการแปลงค่าฟีเจอร์ประเภทหมวดหมู่เป็นค่าตัวเลขด้วยฟังก์ชั่น Factorize

3.4.2 **การปรับช่วงค่าขอบเขตของฟีเจอร์** (Feature Scaling)

การปรับช่วงค่าขอบเขตของฟีเจอร์ที่เป็นข้อมูลประเภทตัวเลขให้อยู่ในช่วงขอบเขตเดียวกัน เพื่อให้ข้อมูลเหมาะกับการที่แบบจำลองนำไปประมวลผล ในงานวิจัยนี้ทำ Feature scaling โดยการทำข้อมูลประเภทตัวเลขให้เป็นค่ามาตรฐาน (Normalization) ด้วยวิธี MinMaxScaler เนื่องจากชุดข้อมูลที่ใช้ในการวิจัยมีการกระจายตัวของข้อมูลที่ไม่ได้อยู่ในลักษณะปกติ (Normal Distribution) ดังภาพประกอบที่ 22

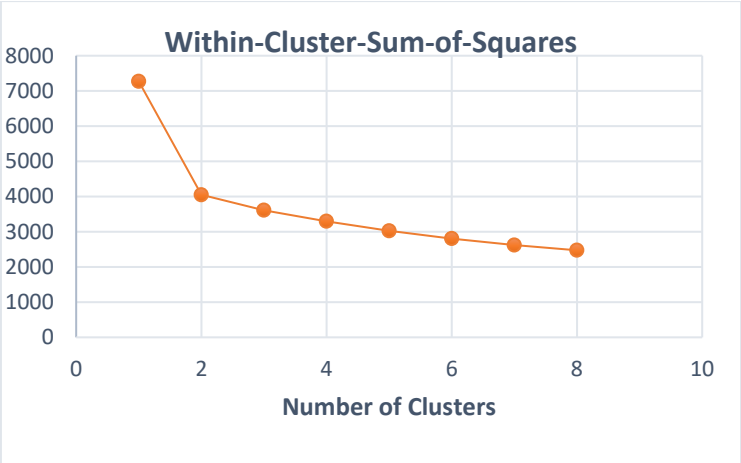
```
[ ] from sklearn.preprocessing import MinMaxScaler

    # Extract numerical column positions
    df_norm = ['Age', 'Gender', 'Education_Level','Marital_Status','Months_on_book', 'Credit_Limit', 'Total_Trans_Amt', 'Total_Trans_Count', 'Max_income']
    scaler = MinMaxScaler().fit(df[df_norm])

[ ] df[df_norm] = scaler.transform(df[df_norm])

[ ] df.head()
```

	Age	Gender	Education_Level	Marital_Status	Months_on_book	Credit_Limit	Total_Trans_Amt	Total_Trans_Count	Max_income
0	0.404255	0.0	0.000000	0.000000	0.604651	0.367685	0.035273	0.248062	0.666667
1	0.489362	1.0	0.166667	0.333333	0.720930	0.239193	0.043452	0.178295	0.333333
2	0.531915	0.0	0.166667	0.000000	0.534884	0.099027	0.076611	0.077519	1.000000
3	0.297872	1.0	0.000000	0.666667	0.488372	0.095984	0.036775	0.077519	0.333333
4	0.297872	0.0	0.333333	0.000000	0.186047	0.136632	0.017025	0.139535	0.666667

ภาพประกอบ 22 แสดงการแปลงช่วงค่าขอบเขตฟีเจอร์ด้วยวิธีMinMaxScaler

3.5 **การสร้างแบบจำลองจัดกลุ่มลูกค้า** (Clustering Model)

ในงานวิจัยนี้จัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means Clustering ซึ่งแบบจำลอง K-Means เป็นแบบจำลองที่ต้องกำหนดค่า K หรือ จำนวนกลุ่มที่ต้องการก่อนการจัดกลุ่ม จึงทำการหาค่า K ที่เหมาะสมด้วย Elbow Method โดยคำนวณค่าCost function ด้วยค่าผลรวมของระยะห่างระหว่างจุดข้อมูลและจุดศูนย์กลางของข้อมูล (Centroid) ของทุกCluster เรียกว่า Within-Cluster-Sum-of-Squares (WCSS) เทียบกับจำนวนCluster และใช้เทคนิค K-Means++ สุ่มวางตำแหน่งจุดศูนย์กลางของข้อมูล (Centroid) แรก



ภาพประกอบ 23 การหาจำนวนClusterของแบบจำลองK-Meansด้วยวิธีElbow Method

จากภาพประกอบที่ 23 การหาค่า K ที่เหมาะสมของแบบจำลองK-Means ด้วย Elbow Method มีจุดที่เป็น Elbow point (จุดหักศอก) ที่ค่า K เท่ากับ 2 มีค่า Within-Cluster-Sum-of-Squares เท่ากับ 4048.07, ค่า K เท่ากับ 3 มีค่า Within-Cluster-Sum-of-Squares เท่ากับ 3609.67 และ K เท่ากับ 4 มีค่า Within-Cluster-Sum-of-Squares เท่ากับ 3295.96

3.6 การวัดผลลัพธ์จากแบบจำลองจัดกลุ่มลูกค้า

ผู้วิจัยใช้วิธีวิเคราะห์ ประเมินผลลัพธ์ และเปรียบเทียบกลุ่มข้อมูลของแบบจำลองด้วยการวัดผลประสิทธิภาพของแบบจำลองด้วยตัวชี้วัดภายใน (Internal metric) ได้แก่ Silhouette Score, Davies-Bouldin index และ Calinski-Harabasz index

3.7 การสร้างแบบจำลองเพื่อทำนายกลุ่มลูกค้า (Classification Model)

งานวิจัยนี้ต้องการเพิ่มประสิทธิภาพและความน่าเชื่อถือของการจัดกลุ่มด้วยการทำนายกลุ่มของลูกค้า โดยนำชุดข้อมูลที่มีเลเบล เข้าสู่แบบจำลอง Decision Tree Classification และเนื่องจากชุดข้อมูลมีฟีเจอร์ประเภทหมวดหมู่ ได้แก่ Gender, Education_Level และ Marital_Status จึงใช้วิธีแปลงค่าฟีเจอร์ประเภทหมวดหมู่เป็นค่าตัวเลขด้วยฟังก์ชั่นFactorizeอีกครั้ง และเนื่องจากแบบจำลอง Decision Tree เป็นแบบจำลองประเภท Rule-Based ดังนั้น จึงไม่จำเป็นต้องทำการปรับช่วงค่าของฟีเจอร์

เมื่อเตรียมข้อมูลสำหรับการเรียนรู้เรียบร้อยแล้ว ขั้นตอนต่อไปคือการนำข้อมูลเข้าสู่แบบจำลอง โดยใช้ข้อมูลในการเรียนรู้ทั้งหมด 7,088 ข้อมูลและข้อมูลสำหรับการทดสอบทั้งหมด 3,039 ข้อมูล จากนั้นทำการทดลองด้วยแบบจำลอง Decision Tree ร่วมกับการปรับพารามิเตอร์

สำหรับแบบจำลอง Decision Tree Classification ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 6 พารามิเตอร์ ได้แก่

3.7.1 Criterion คือ สูตรที่ใช้วัดคุณภาพของการแบ่งพื้นที่ (Partition) ของต้นไม้ตัดสินใจ มีทั้งหมด 2 ค่า ได้แก่

- Gini คือ ค่าที่ใช้วัดความไม่บริสุทธิ์ของพาร์ทิชั่นที่ถูกแบ่งโดยใช้ฟีเจอร์หนึ่ง ดังนั้นค่า Gini น้อยแสดงถึงการแบ่งข้อมูลออกมาได้ดี

- Entropy คือ ค่าวัดความไม่แน่นอนของข้อมูล ซึ่งความไม่แน่นอนหมายถึงจำนวนข้อมูลที่ทายผิด ดังนั้นค่า Entropyน้อยแสดงถึงจำนวนข้อมูลที่ทายผิดน้อย

3.7.2 Max_depth คือ ค่าความลึกของต้นไม้ตัดสินใจ

3.7.3 Splitter คือ การเลือกการแยกฟีเจอร์ในแต่ละNode มีการแยก 2 ประเภท คือ

- Best (ค่าตั้งต้น) คือ อัลกอริธึมจะพิจารณาฟีเจอร์ทั้งหมด และเลือกการแยกNode ที่ดีที่สุด

- Random คือ อัลกอริธึมจะพิจารณาฟีเจอร์และเลือกการแยกNodeแบบสุ่มที่ดีที่สุด

3.7.4 Min_sample_leafs คือ จำนวนข้อมูลขั้นต่ำในLeaf Node ของต้นไม้ตัดสินใจ ถ้าจำนวนข้อมูลต่ำกว่าค่านี้ จะหยุดการแยกNode

3.7.5 Min_sample_split คือ จำนวนขั้นต่ำที่จำเป็นในNode เพื่อทำให้เกิดการแยก Node

3.7.6 Max_features คือ การกำหนดฟีเจอร์ที่ใช้ในการตัดสินใจ มีทั้งหมด 3 แบบคือ

- Auto (ค่าตั้งต้น) คือ การใช้ทุกฟีเจอร์ที่แบบจำลองมองว่าเหมาะสม

- Sqrt คือ การใช้ฟีเจอร์จำนวนรากที่สองของจำนวนฟีเจอร์ทั้งหมด

- Log2 คือ การใช้ฟีเจอร์จำนวนลอการิทึมฐานสองของจำนวนฟีเจอร์ทั้งหมด

เมื่อปรับพารามิเตอร์สำหรับแบบจำลอง Decision Tree Classification ทำให้ได้ผลออกมาดังตารางที่ 3

ตาราง 3 แสดงค่าพารามิเตอร์ในการปรับแบบจำลองDecision Tree

พารามิเตอร์	**แบบจำลองDecision Tree**
Criterion	Entropy
Max_depth	10
Splitter	Best
Min_sample_leaf	2
Min_sample_split	9
Max_features	Auto

หลังจากปรับจูนพารามิเตอร์ให้แบบจำลอง Decision Tree และให้แบบจำลองได้เรียนรู้ชุดข้อมูล หลังจากนั้นจึงทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และแสดงผลการทำนายของแบบจำลองด้วย Confusion Matrix

บทที่ 4

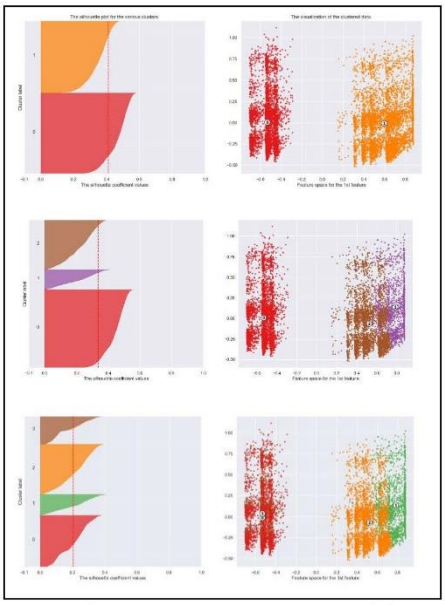
ผลการดำเนินการวิจัย

ในการวิจัยเพื่อการศึกษาวิธีการจัดกลุ่มข้อมูลลูกค้าบัตรเครดิตของธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนด้วยแบบจำลอง K-Means ในการจัดกลุ่ม และทำนายกลุ่มโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอนโดยใช้แบบจำลองDecision Tree ผู้วิจัยดำเนินการวิจัยโดยการศึกษาตามขบวนการ และขั้นตอนต่างๆ ตลอดจนวัดประสิทธิภาพ เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยที่กำหนดไว้ดังนี้

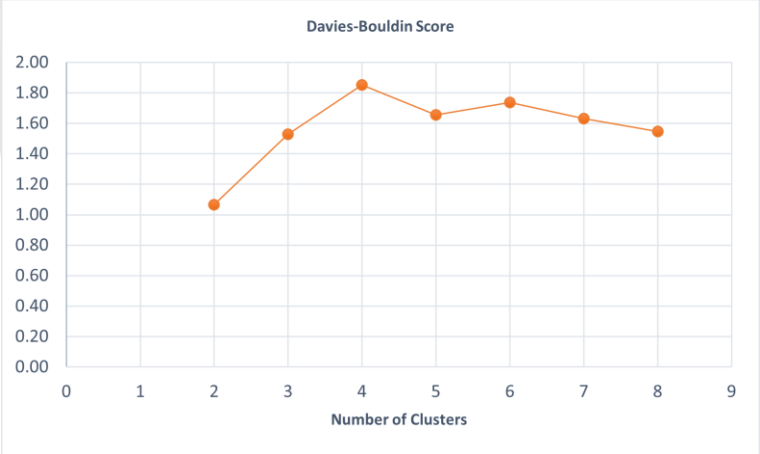
1. ผลลัพธ์ของการจัดกลุ่มด้วยแบบจำลองK-Means
2. ผลลัพธ์ของการทำนายกลุ่มด้วยแบบจำลองDecision Tree
3. วิเคราะห์ลักษณะของลูกค้าในแต่ละกลุ่ม (Cluster Analysis)

4.1 ผลลัพธ์ของการจัดกลุ่มด้วยแบบจำลองK-Means

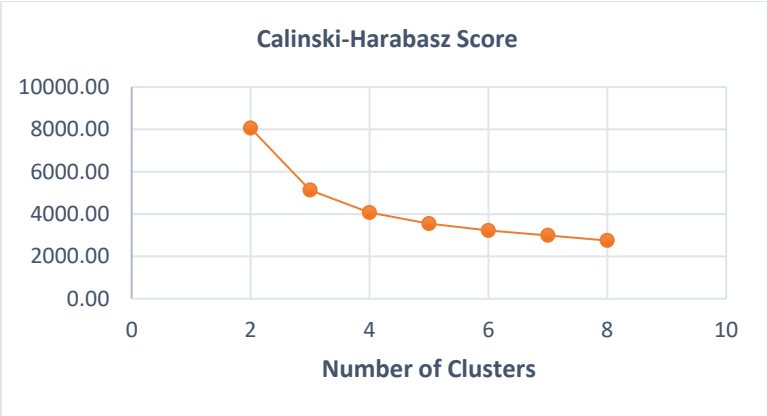
จากการสร้างแบบจำลอง K-Means ในการจัดกลุ่มข้อมูล โดยใช้การวัดผลด้วยSilhouette score จากภาพประกอบที่ 24 (ด้านซ้าย) แสดงให้เห็นว่า การจัดกลุ่มข้อมูลเมื่อค่า K เท่ากับ 2 มีค่าเฉลี่ยของSilhouette score สูงที่สุด คือ 0.41 รองลงมาคือการจัดกลุ่มเมื่อค่า K เท่ากับ 3 มีค่าเฉลี่ยเท่ากับ 0.33 และจากภาพประกอบที่ 24 (ด้านขวา) แสดงการกระจายตัวของข้อมูลสองมิติหลังจากการใช้เทคนิคPCA (Principle Component Analysis) ในการลดมิติข้อมูล กราฟแสดงการจัดกลุ่มข้อมูลอย่างชัดเจน โดยเฉพาะอย่างยิ่งเมื่อค่า K เท่ากับ 2 กลุ่มข้อมูลขนาดใหญ่แสดงด้วยสีแดง และกลุ่มข้อมูลขนาดเล็กแสดงด้วยสีส้ม และเมื่อค่า K เท่ากับ 3 กลุ่มข้อมูลขนาดเล็กกว่ามีการแบ่งแยกออกเป็นอีกกลุ่มหนึ่ง



ภาพประกอบ 24 (ด้านซ้าย) แสดงค่าSilhouette scoreและค่าเฉลี่ยเมื่อK=2,3 และ4 (ด้านขวา) แสดงจุดข้อมูลและCentroid (รูปวงกลม) ของแต่ละกลุ่ม



ภาพประกอบ 25 แสดงค่าDavies-Bouldin indexกับจำนวนCluster



ภาพประกอบ 26 แสดงค่าCalinski-Harabasz indexกับจำนวนCluster

ตาราง 4 แสดงค่าSilhouette score, Davies-Bouldin index และCalinski-Harabasz index

จำนวนคลัสเตอร์	Silhouette score	Davies-Bouldin index	Calinski-Harabasz index
2	0.41	1.06	8061.10
3	0.33	1.53	5134.41
4	0.25	1.85	4069.55
5	0.22	1.65	3550.40
6	0.22	1.74	3224.32

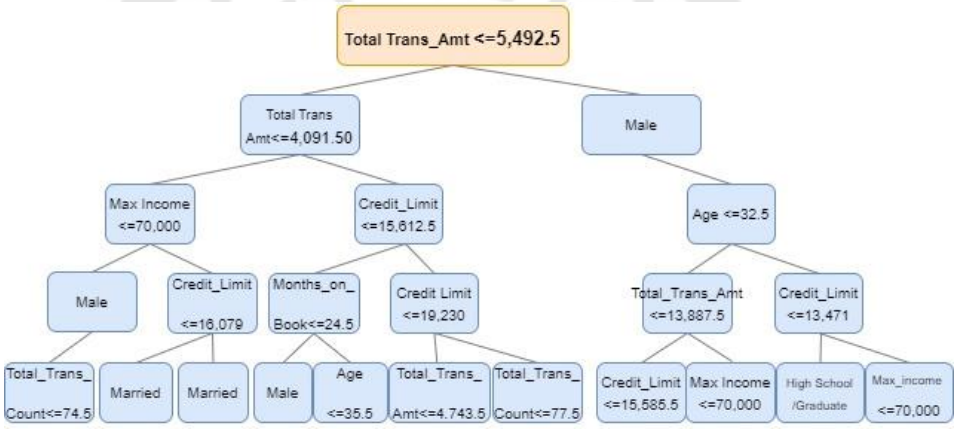
จากตารางที่ 4 จะเห็นว่าแบบจำลอง K-Means ที่ค่า K เท่ากับ 2 มีค่าSilhouette score เท่ากับ 0.41 และค่า K เท่ากับ 3 มีค่าSilhouette score เท่ากับ 0.33 แสดงถึงคุณภาพของCluster มีการแยกตัวออกห่างจากClusterอื่น และจุดข้อมูลในCluster เกาะกลุ่มกัน ค่าDavies-Bouldin ที่ค่า K เท่ากับ 2 มีค่าเท่ากับ 1.06 และK เท่ากับ 3 มีค่า Davies-Bouldin เท่ากับ 1.53 แสดงถึงClusterที่แยกตัวออกห่างจากClusterอื่นชัดเจน และค่าCalinski-Harabasz ที่ค่า Kเท่ากับ 2 เท่ากับ 8061.10 และK เท่ากับ3 มีค่า Calinski-Harabasz เท่ากับ 5134.41 แสดงถึงคุณภาพของ Cluster มีการแยกตัวออกห่างจากClusterอื่น และจุดข้อมูลในCluster เกาะกลุ่มกัน

จากผลลัพธ์ของตัวชี้วัด (Metric)ทั้งหมด จะเห็นว่าค่าที่ดีที่สุด คือ ค่า K หรือ จำนวน Clusterเท่ากับ 2 แต่เนื่องจากงานวิจัยนี้มีวัตถุประสงค์เพื่อนำไปวางแผนการตลาดเฉพาะกลุ่ม

ให้กับลูกค้า ดังนั้น การจัดลูกค้าออกเป็นกลุ่มย่อยหลายกลุ่มสามารถตอบโจทย์ธุรกิจมากกว่า ผู้วิจัยจึงพิจารณาเลือกจัดกลุ่มข้อมูลเป็น 3 กลุ่ม (Cluster) โดยแบบจำลองK-Means จัดกลุ่มข้อมูลออกมามีผลลัพธ์ดังนี้ Clusterที่ศูนย์ มีจำนวน 1,375 คน Clusterที่หนึ่ง มีจำนวน 5,358 คน และ Clusterที่สอง มีจำนวน 3,394 คน

4.2 ผลลัพธ์ของการทำนายกลุ่มด้วยแบบจำลองDecision Tree

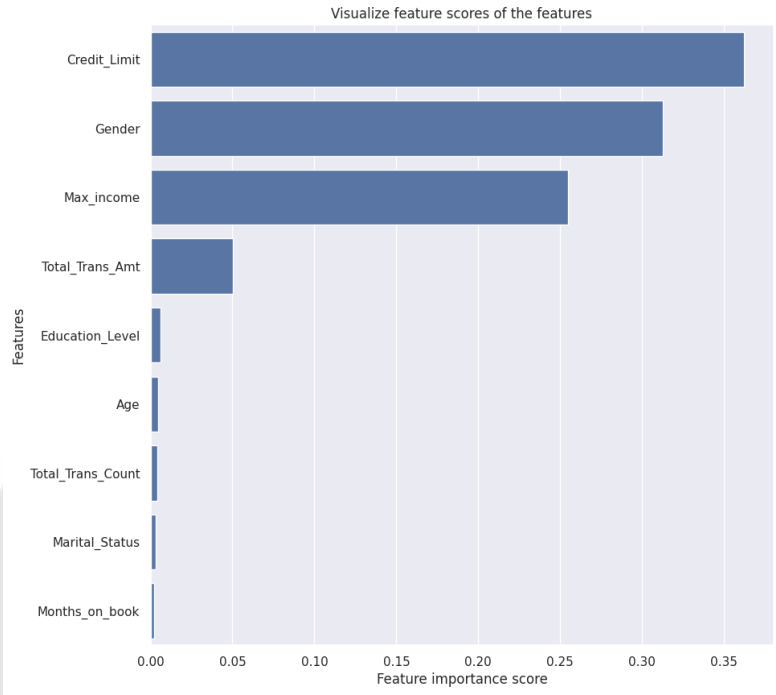
ในการวัดประสิทธิภาพของแบบจำลองDecision Tree ผู้วิจัยใช้ข้อมูลทดสอบทั้งหมด 3,039 แถว ประกอบไปด้วยลูกค้าใน Clusterที่ศูนย์ มีจำนวน 413 คน Clusterที่หนึ่งจำนวน 1,608 คน และ Clusterที่สองจำนวน 1,018 คน โดยแบบจำลองDecision Tree มีกฎการตัดสินใจในการแบ่งฟีเจอร์ดังภาพประกอบที่ 27



ภาพประกอบ 27 แสดงการแบ่งกฎการตัดสินใจของต้นไม้ 4 ชั้น

จากภาพประกอบ 27 แบบจำลองDecision Tree แบ่งกฎการตัดสินใจของต้นไม้ 4 ชั้น หลังจากคำนวณค่าEntropy โดยเริ่มต้นแบ่งที่ Root node ด้วยฟีเจอร์ Total_Trans_Amt ที่แบ่งกฎการตัดสินใจที่ยอดใช้จ่ายบัตรเครดิตน้อยกว่า 5,492.5 ในต้นไม้ชั้นที่หนึ่งแบ่งกฎการตัดสินใจด้วยฟีเจอร์ Total_Trans_Amt และGender ในต้นไม้ชั้นที่สองแบ่งกฎการตัดสินใจด้วยฟีเจอร์ Max_Income, Credit_Limit และAge ในต้นไม้ชั้นที่สามแบ่งกฎการตัดสินใจด้วยฟีเจอร์ Gender, Credit_Limit, Months_on_Book และ Total_Trans_Amt และในต้นไม้ชั้นที่สี่แบ่งกฎการตัดสินใจด้วยฟีเจอร์ Total_Trans_Count, Marital_Status, Gender, Age, Total_Trans_Amt, Credit_Limit, Max_Income และEducation_Level

ผู้วิจัยได้ทำการแสดงความสำคัญของฟีเจอร์ที่แบบจำลองDecision Tree ใช้ในการเรียนรู้ด้วย Feature importance เพื่อสำรวจว่าแต่ละCluster ถูกจัดด้วยฟีเจอร์ใด ดังภาพประกอบที่ 28



ภาพประกอบ 28 แสดงค่าฟีเจอร์สำคัญที่แบบจำลองใช้ในการเรียนรู้

จากภาพประกอบที่ 28 แสดงถึงฟีเจอร์สำคัญที่แบบจำลองDecision Tree ใช้ในการเรียนรู้มากที่สุด 4 ฟีเจอร์ ได้แก่ ฟีเจอร์Credit_Limit มีค่าFeature importance 0.362, ฟีเจอร์Gender มีค่าFeature importance 0.312, ฟีเจอร์Max_Income มีค่าFeature importance 0.255 และฟีเจอร์Total_Trans_Amt มีค่าFeature importance 0.050

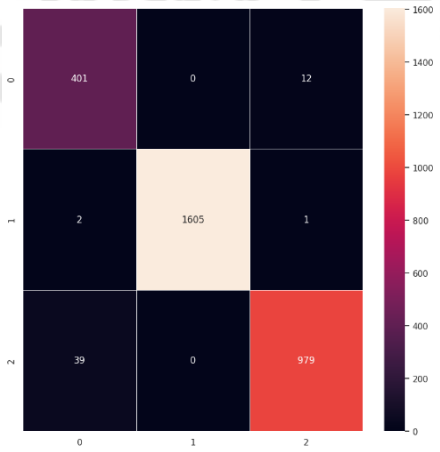
ในการทำนายกลุ่มลูกค้าด้วยแบบจำลองDecision Tree วัดประสิทธิภาพการทำงานของแบบจำลองDecision Tree ด้วยค่าAccuracy, Precision, Recall และF1-Score ดังตารางที่ 5

ตาราง 5 แสดงประสิทธิภาพของแบบจำลองDecision Tree

Model Performance	Accuracy	Precision	Recall	F1-Score
Cluster 0	0.97	0.91	0.97	0.94
Cluster 1	1.00	1.00	1.00	1.00
Cluster 2	0.96	0.99	0.96	0.97
Weighted Average	0.98	0.98	0.98	0.98

ในขั้นตอนของแบบจำลองทำนายกลุ่ม Decision Tree โดยให้แบบจำลองเรียนรู้ชุดข้อมูลที่มีเลเบล ซึ่งแบบจำลองประสบความสำเร็จในการทำนายกลุ่มข้อมูลทดสอบด้วยความแม่นยำ (Accuracy) ที่สูงถึง 0.98 นอกจากนี้ผลการประเมินประสิทธิภาพของแบบจำลองได้ค่าPrecision เท่ากับ 0.98 ค่าRecall เท่ากับ 0.98 และค่าF1-Score เท่ากับ 0.98 ซึ่งบ่งชี้ถึงแบบจำลองทำงานมีประสิทธิภาพดีเยี่ยม

จากตารางที่ 5 พบว่าแบบจำลองDecision Tree ทำนายกลุ่มลูกค้าในClusterที่หนึ่งถูกมากที่สุด และทำนายกลุ่มลูกค้าผิดพลาดในClusterที่ศูนย์ และClusterที่สอง เพื่อสังเกตจำนวนความถูกต้อง และความผิดพลาดในการทำนาย ผู้วิจัยได้ทำการแสดงผลของConfusion matrix ดังภาพประกอบ 29



ภาพประกอบ 29 แสดงผลของConfusion matrixของแบบจำลองDecision Tree

จากภาพประกอบ 29 พบว่ากลุ่มที่ทำนายถูกมากที่สุด คือ ลูกค้าในClusterที่หนึ่ง พบว่ามีการทำนายผิดพลาดจำนวน 3 ข้อมูล รองลงมาคือ ลูกค้าในClusterที่ศูนย์มีการทำนายผิดพลาดจำนวน 12 ข้อมูล และลูกค้าในClusterที่หนึ่งมีการทำนายผิดพลาดจำนวน 39 ข้อมูล

4.3 วิเคราะห์ลักษณะของลูกค้าในแต่ละกลุ่ม (Cluster Analysis)

จากการจัดกลุ่มลูกค้าบัตรเครดิตด้วยแบบจำลองK-Means ซึ่งแบ่งลูกค้าออกเป็น 3 กลุ่ม โดยแต่ละกลุ่มมีลักษณะของลูกค้าดังนี้

- Clusterที่ศูนย์ เป็นกลุ่ม “Selective Spender” ประกอบไปด้วยลูกค้าจำนวน 1,385 คน โดยลูกค้าใน Clusterที่ศูนย์มีลักษณะดังนี้

- เพศชาย
- ช่วงอายุ 26-65 ปี ประกอบไปด้วยลูกค้าอายุ 26-35 ปีจำนวน 70 คน อายุ 36-45 ปี จำนวน 536 คน อายุ 46-55 ปี จำนวน 647 คน และอายุ 56-65 ปีจำนวน 132 คน
- ระดับการศึกษา ลูกค้าจบการศึกษาระดับปริญญาตรีจำนวน 408 คน ลูกค้าที่จบการศึกษาระดับมัธยมศึกษาตอนปลายจำนวน 269 คน ลูกค้าที่ไม่ได้รับการศึกษาจำนวน 219 คน ลูกค้าที่ไม่มีข้อมูลการศึกษาจำนวน 209 คน ลูกค้าจบระดับวิทยาลัยจำนวน 148 คน ลูกค้าจบการศึกษาระดับปริญญาโทจำนวน 77 คน และลูกค้าจบการศึกษาระดับปริญญาเอกจำนวน 55 คน
- สถานภาพ ลูกค้าสถานภาพโสดจำนวน 582 คน ลูกค้าสถานภาพสมรสจำนวน 546 คน ลูกค้าสถานภาพหย่าร้างจำนวน 128 คน และลูกค้าที่ไม่มีข้อมูลสถานภาพจำนวน 122 คน
- ระยะเวลาที่ถือบัตรเครดิตอยู่ระหว่าง 13-56 เดือน ลูกค้าที่ถือบัตรเครดิต 13-24 เดือน จำนวน 76 คน ลูกค้าที่ถือบัตรเครดิต 25-36 เดือนจำนวน 761 คน ลูกค้าที่ถือบัตรเครดิต 37-48 เดือนจำนวน 485 คน และลูกค้าที่ถือบัตรเครดิต 49-56 เดือนจำนวน 63 คน
- วงเงินบัตรเครดิตที่ได้รับอยู่ระหว่าง 9,959-34,516 ลูกค้าที่ได้รับวงเงินน้อยกว่า 10,000 มีจำนวน 1 คน ลูกค้าที่ได้รับวงเงิน 10,000-20,000 มีจำนวน 281 คน ลูกค้าที่ได้รับวงเงิน 20,000-30,000 มีจำนวน 500 คน และลูกค้าที่ได้รับวงเงิน 30,000 ขึ้นไปมีจำนวน 603 คน

- จำนวนครั้งในการใช้บัตรเครดิตอยู่ระหว่าง 10-139 ครั้ง ลูกค้าจำนวน 365 คนใช้บัตรเครดิตน้อยกว่า 50 ครั้ง ลูกค้าจำนวน 814 คนใช้บัตรเครดิต 51-100 ครั้ง และลูกค้าจำนวน 206 คนใช้บัตรเครดิต 101 ครั้งขึ้นไป
- ยอดใช้จ่ายบัตรเครดิตอยู่ระหว่าง 597-15,399 ลูกค้าจำนวน 914 คนมียอดใช้จ่ายบัตรเครดิตน้อยกว่า 5,000 ลูกค้าจำนวน 238 คนมียอดใช้จ่ายบัตรเครดิต 5,000-10,000 ลูกค้าจำนวน 135 คนมียอดใช้จ่ายบัตรเครดิต 10,000-15,000 และลูกค้าจำนวน 98 คนมียอดใช้จ่ายบัตรเครดิต 15,000 ขึ้นไป
- รายได้สูงสุดที่ได้รับ ลูกค้าที่มีรายได้สูงสุด 120,000 จำนวน 1,100 คน ลูกค้าที่มีรายได้สูงสุด 80,000 จำนวน 268 คน ลูกค้าที่มีรายได้สูงสุด 60,000 จำนวน 13 คน และลูกค้าที่ไม่มีข้อมูลรายได้จำนวน 4 คน

- Clusterที่หนึ่ง เป็นกลุ่ม “Moderate Spender” ประกอบไปด้วยลูกค้าจำนวน 5,358 คน โดยลูกค้าใน Clusterที่หนึ่งมีลักษณะดังนี้

- เพศหญิง
- ช่วงอายุ 26-67 ปี ประกอบไปด้วยลูกค้าอายุ 26-35 ปีจำนวน 471 คน อายุ 36-45 ปีจำนวน 1,977 คน อายุ 46-55 ปี จำนวน 2,198 คน อายุ 56-65 ปีจำนวน 708 คน และอายุ 66 ปีขึ้นไปจำนวน 4 คน
- ระดับการศึกษา ลูกค้าจบการศึกษาระดับปริญญาตรีจำนวน 1,670 คน ลูกค้าที่จบการศึกษาระดับมัธยมศึกษาตอนปลายจำนวน 1,028 คน ลูกค้าที่ไม่มีข้อมูลการศึกษาจำนวน 812 คน ลูกค้าที่ไม่ได้รับการศึกษาจำนวน 796 คน คน ลูกค้าจบระดับวิทยาลัยจำนวน 532 คน ลูกค้าจบการศึกษาระดับปริญญาโทจำนวน 263 คน และลูกค้าจบการศึกษาระดับปริญญาเอกจำนวน 257 คน
- สถานภาพ ลูกค้าสถานภาพโสดจำนวน 2,125 คน ลูกค้าสถานภาพสมรสจำนวน 2,451 คน ลูกค้าสถานภาพหย่าร้างจำนวน 402 คน และลูกค้าที่ไม่มีข้อมูลสถานภาพจำนวน 380 คน
- ระยะเวลาที่ถือบัตรเครดิตอยู่ระหว่าง 13-56 เดือน ลูกค้าที่ถือบัตรเครดิต 13-24 เดือนจำนวน 432 คน ลูกค้าที่ถือบัตรเครดิต 25-36 เดือนจำนวน 2,886 คน ลูกค้าที่ถือบัตรเครดิต 37-48 เดือนจำนวน 1,702 คน และลูกค้าที่ถือบัตรเครดิต 49-56 เดือนจำนวน 338 คน

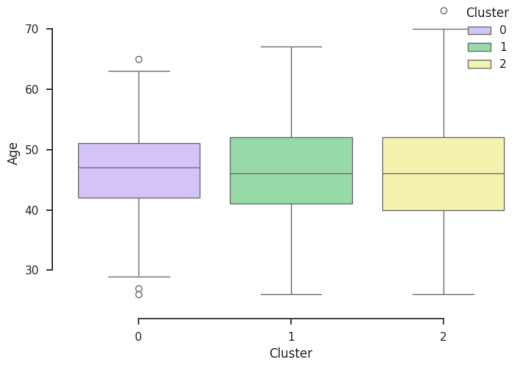
- วงเงินบัตรเครดิตที่ได้รับอยู่ระหว่าง 1,439-34,516 ลูกค้าที่ได้รับวงเงินน้อยกว่า 10,000 มีจำนวน 4,777 คน ลูกค้าที่ได้รับวงเงิน 10,000-20,000 มีจำนวน 439 คน ลูกค้าที่ได้รับวงเงิน 20,000-30,000 มีจำนวน 78 คน และลูกค้าที่ได้รับวงเงิน 30,000 ขึ้นไปมีจำนวน 64 คน
- จำนวนครั้งในการใช้บัตรเครดิตอยู่ระหว่าง 12-138 ครั้ง ลูกค้าจำนวน 1,473 คนใช้บัตรเครดิตน้อยกว่า 50 ครั้ง ลูกค้าจำนวน 3,626 คนใช้บัตรเครดิต 51-100 ครั้ง และลูกค้าจำนวน 259 คนใช้บัตรเครดิต 101 ครั้งขึ้นไป
- ยอดใช้จ่ายบัตรเครดิตอยู่ระหว่าง 510-17,437 ลูกค้าจำนวน 4,463 คนมียอดใช้จ่ายบัตรเครดิตน้อยกว่า 5,000 ลูกค้าจำนวน 596 คนมียอดใช้จ่ายบัตรเครดิต 5,000-10,000 ลูกค้าจำนวน 196 คนมียอดใช้จ่ายบัตรเครดิต 10,000-15,000 และลูกค้าจำนวน 103 คนมียอดใช้จ่ายบัตรเครดิต 15,000 ขึ้นไป
- รายได้สูงสุดที่ได้รับ ลูกค้าที่มีรายได้สูงสุด 60,000 จำนวน 1,014 คน ลูกค้าที่มีรายได้สูงสุด 40,000 จำนวน 3,284 คน และลูกค้าที่ไม่มีข้อมูลรายได้จำนวน 1,060 คน

- Clusterที่สอง เป็นกลุ่ม “Strategic Spender” ประกอบไปด้วยลูกค้าจำนวน 3,834 คน โดยลูกค้าใน Clusterที่สองมีลักษณะดังนี้

- เพศชาย
- ช่วงอายุ 26-73 ปี ประกอบไปด้วยลูกค้าอายุ 26-35 ปีจำนวน 378 คน อายุ 36-45 ปี จำนวน 1,229 คน อายุ 46-55 ปี จำนวน 1,290 คน อายุ 56-65 ปีจำนวน 481 คน และอายุ 66 ปีขึ้นไปจำนวน 6 คน
- ระดับการศึกษา ลูกค้าจบการศึกษาระดับปริญญาตรีจำนวน 1,050 คน ลูกค้าที่จบการศึกษาระดับมัธยมศึกษาตอนปลายจำนวน 716 คน ลูกค้าที่ไม่มีข้อมูลการศึกษาจำนวน 498 คน ลูกค้าที่ไม่ได้รับการศึกษาจำนวน 472 คน คน ลูกค้าจบระดับวิทยาลัยจำนวน 333 คน ลูกค้าจบการศึกษาระดับปริญญาโทจำนวน 176 คน และลูกค้าจบการศึกษาระดับปริญญาเอกจำนวน 139 คน
- สถานภาพ ลูกค้าสถานภาพโสดจำนวน 1,236 คน ลูกค้าสถานภาพสมรสจำนวน 1,690 คน ลูกค้าสถานภาพหย่าร้างจำนวน 218 คน และลูกค้าที่ไม่มีข้อมูลสถานภาพจำนวน 240 คน

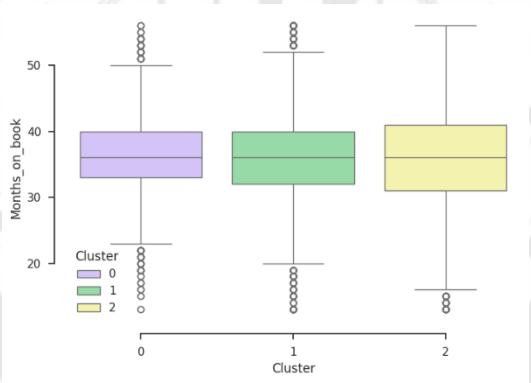
- ระยะเวลาที่ถือบัตรเครดิตอยู่ระหว่าง 13-56 เดือน ลูกค้าที่ถือบัตรเครดิต 13-24 เดือน จำนวน 339 คน ลูกค้าที่ถือบัตรเครดิต 25-36 เดือนจำนวน 1,771 คน ลูกค้าที่ถือบัตรเครดิต 37-48 เดือนจำนวน 1,020 คน และลูกค้าที่ถือบัตรเครดิต 49-56 เดือนจำนวน 254 คน
- วงเงินบัตรเครดิตที่ได้รับอยู่ระหว่าง 1,439-25,618 ลูกค้าที่ได้รับวงเงินน้อยกว่า 10,000 มีจำนวน 2,595 คน ลูกค้าที่ได้รับวงเงิน 10,000-20,000 มีจำนวน 770 คน และลูกค้าที่ได้รับวงเงิน 20,000-30,000 มีจำนวน 19 คน
- จำนวนครั้งในการใช้บัตรเครดิตอยู่ระหว่าง 10-131 ครั้ง ลูกค้าจำนวน 1,290 คนใช้บัตรเครดิตน้อยกว่า 50 ครั้ง ลูกค้าจำนวน 1,910 คนใช้บัตรเครดิต 51-100 ครั้ง และลูกค้าจำนวน 184 คนใช้บัตรเครดิต 101 ครั้งขึ้นไป
- ยอดใช้จ่ายบัตรเครดิตอยู่ระหว่าง 530-18,484 ลูกค้าจำนวน 2,863 คนมียอดใช้จ่ายบัตรเครดิตน้อยกว่า 5,000 ลูกค้าจำนวน 296 คนมียอดใช้จ่ายบัตรเครดิต 5,000-10,000 ลูกค้าจำนวน 141 คนมียอดใช้จ่ายบัตรเครดิต 10,000-15,000 และลูกค้าจำนวน 84 คนมียอดใช้จ่ายบัตรเครดิต 15,000 ขึ้นไป
- รายได้สูงสุดที่ได้รับ ลูกค้าที่มีรายได้สูงสุด 120,000 จำนวน 1,162 คน ลูกค้าที่มีรายได้สูงสุด 80,000 จำนวน 1,134 คน ลูกค้าที่มีรายได้สูงสุด 60,000 จำนวน 763 คนและลูกค้ามีรายสูงสุด 40,000 จำนวน 227 คน

ผู้วิจัยพิจารณาเปรียบเทียบความแตกต่างของกลุ่มลูกค้าบัตรเครดิตทั้ง 3 Cluster โดยพิจารณาฟีเจอร์ Age, Months_on_Book, Credit_Limit, Total_Trans_Count และ Total_Trans_Amt



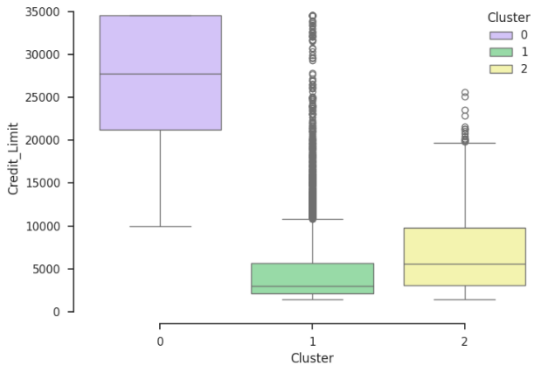
ภาพประกอบ 30 เปรียบเทียบอายุของลูกค้าในแต่ละCluster

จากภาพประกอบ 30 พบว่าช่วงอายุของลูกค้าในแต่ละกลุ่ม ส่วนใหญ่อยู่ในช่วง 40-52 ปี ซึ่งไม่มีความแตกต่างกันในแต่ละกลุ่ม



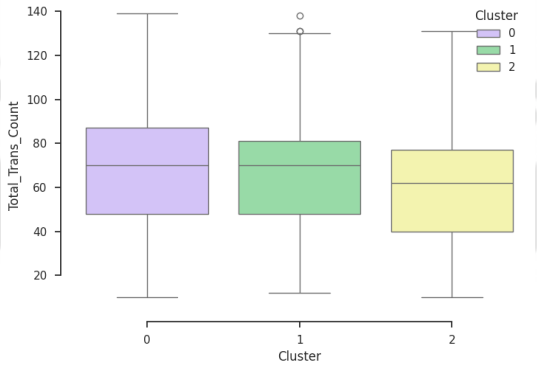
ภาพประกอบ 31 เปรียบเทียบระยะเวลาในการถือบัตรเครดิตของแต่ละCluster

จากภาพประกอบ 31 พบว่าระยะเวลาที่ลูกค้าถือบัตรเครดิตส่วนใหญ่ในแต่ละกลุ่มอยู่ในช่วง 31-41 เดือน ซึ่งไม่มีความแตกต่างกัน



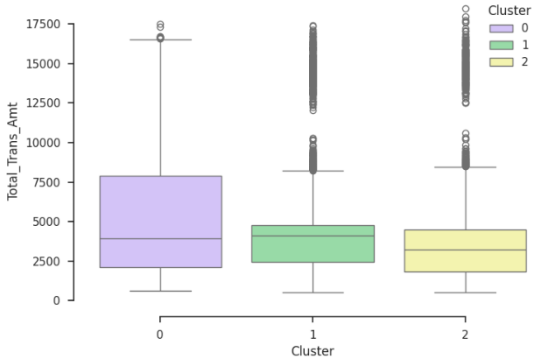
ภาพประกอบ 32 เปรียบเทียบวงเงินบัตรเครดิตของแต่ละCluster

จากภาพประกอบที่ 32 พบว่าวงเงินบัตรเครดิตที่ลูกค้าได้รับมีความแตกต่างกัน โดยลูกค้าในClusterที่ศูนย์ เป็นกลุ่มที่ได้รับวงเงินมากที่สุด ลำดับรองลงมา คือ ลูกค้าในClusterที่สอง และลูกค้าในClusterที่หนึ่งได้รับวงเงินบัตรเครดิตน้อยที่สุด



ภาพประกอบ 33 เปรียบเทียบจำนวนครั้งในการใช้บัตรเครดิตของแต่ละCluster

จากภาพประกอบที่ 33 พบว่าจำนวนในการใช้บัตรเครดิตของลูกค้ามีความแตกต่างกันเล็กน้อย โดยลูกค้าในClusterที่ศูนย์เป็นกลุ่มที่ใช้บัตรเครดิตจำนวน 48-87 ครั้ง ลำดับรองลงมาคือลูกค้าในClusterที่หนึ่งจำนวน 48-81 ครั้ง และลูกค้าในClusterที่สองจำนวน 40-77 ครั้ง



ภาพประกอบ 34 เปรียบเทียบยอดใช้จ่ายบัตรเครดิตของแต่ละCluster

จากภาพประกอบที่ 34 พบว่ายอดใช้จ่ายบัตรเครดิตมีความแตกต่างกัน โดยลูกค้าในClusterที่ศูนย์เป็นกลุ่มที่มียอดใช้จ่ายบัตรเครดิตมากที่สุด ในยอด 2,118-7,888 ลำดับรองลงมาคือ ลูกค้าในClusterที่หนึ่ง ในยอด 2,424-4,752 และลูกค้าในClusterที่สองในยอด 1,839-4,493

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยเพื่อศึกษาวิธีการจัดกลุ่มข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบจัดกลุ่มข้อมูล และทำนายกลุ่มข้อมูล ผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลองK-Means และDecision Tree เพื่อนำมาสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ศึกษาการจัดกลุ่มลูกค้าที่มีพฤติกรรมการใช้บัตรเครดิตออกเป็นกลุ่มย่อย และทำนายกลุ่มลูกค้า โดยใช้เทคนิคการเรียนรู้ของเครื่อง 2 วิธีดังนี้

5.1.1 การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน หรือ Unsupervised machine learning ด้วยแบบจำลองK-Means ใช้วิธีElbow method ในการเลือกจำนวนกลุ่ม (Cluster) ที่เหมาะสม เลือกค่า K เท่ากับ 3 กลุ่ม พบว่า Silhouette score มีค่า 0.33 ,Davies-Bouldin index มีค่า 1.53 และCalinski-Harabasz index มีค่า 5134.41

5.1.2 การทำนายกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน หรือ Supervised machine learning ด้วยแบบจำลองDecision Tree ร่วมกับการปรับพารามิเตอร์ด้วย GridsearchCV เท่ากับ 5 และวัดประสิทธิภาพการทำนายกลุ่มของแบบจำลองDecision Tree ให้ผลลัพธ์ที่ดีที่สุดที่ Accuracy 98%, Precision 98%, Recall 98%, F1-Score 98% และจากการพิจารณาConfusion matrix พบว่ากลุ่มที่ทำนายถูกมากที่สุด คือ กลุ่มClusterที่หนึ่ง และกลุ่มที่ทำนายผิดพลาดมากที่สุดเป็นกลุ่มClusterที่สอง ที่มักถูกทำนายผิดเป็นกลุ่มClusterที่ศูนย์ และในขั้นตอนสุดท้ายอธิบายค่าความสำคัญของฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้ด้วย Feature importance พบว่าฟีเจอร์ที่มีผลในการจัดกลุ่มลูกค้า ได้แก่ Credit_Limit มีค่า 0.36, Gender มีค่า 0.31, Max_Income มีค่า 0.25 และTotal_Trans_Amt มีค่า 0.05

5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ศึกษาวิธีการจัดกลุ่มข้อมูลการใช้บัตรเครดิตของลูกค้าธนาคารแห่งหนึ่ง โดยใช้ 2 เทคนิค คือ

5.2.1 การจัดกลุ่มข้อมูลด้วยแบบจำลองK-Means

จากการทดลอง และวัดประสิทธิภาพของแบบจำลองK-Means พบว่าSilhouette score, Davies-Bouldin index และCaliski-Harabasz index มีคะแนนดีที่สุดเมื่อพิจารณาจำนวนกลุ่มที่เหมาะสม หรือ ค่า K เท่ากับ 2 รองลงมา คือ ค่า K เท่ากับ 3 ดังตารางที่ 6 แต่เนื่องจากงานวิจัยนี้เป็นงานวิจัยที่มีความมุ่งหมายที่จะนำไปใช้ในงานทางด้านการตลาด การจัดกลุ่มลูกค้าออกเป็นกลุ่มย่อยหลายกลุ่ม ช่วยให้นักการตลาดทำความเข้าใจพฤติกรรมของลูกค้าได้อย่างเฉพาะเจาะจง และสามารถวางแผนกลยุทธการตลาดเฉพาะบุคคล (Personalized Marketing) เพื่อตอบสนองความต้องการของลูกค้าแต่ละกลุ่มได้อย่างมีประสิทธิภาพ จากเหตุผลเหล่านี้ผู้วิจัยจึงพิจารณาเลือกจำนวนกลุ่มที่ค่า K เท่ากับ 3 หรือ แบ่งลูกค้าออกเป็น 3 กลุ่ม

ตาราง 6 แสดงค่าประสิทธิภาพของแบบจำลองK-Means เมื่อจำนวนClusterตั้งแต่ 2-6

จำนวนคลัสเตอร์	Silhouette score	Davies-Bouldin score	Calinski-Harabasz score
2	0.41	1.06	8061.10
3	0.33	1.53	5134.41
4	0.25	1.85	4069.55
5	0.22	1.65	3550.40
6	0.22	1.74	3224.32

5.2.2 ทำนายกลุ่มข้อมูลด้วยแบบจำลองDecision Tree

จากการทดลอง และวัดประสิทธิภาพการทำนายกลุ่มลูกค้าด้วยแบบจำลอง Decision Tree เมื่อพิจารณาConfusion matrix พบว่าเกิดการทำนายทั้งถูก และผิดพลาดของแบบจำลอง ซึ่งพบว่าลูกค้าที่อยู่ในClusterที่หนึ่งเกิดการทำนายถูกมากที่สุด และเมื่อตรวจสอบข้อมูลในแต่ละกลุ่มพบว่าในClusterที่หนึ่ง ลูกค้าที่อยู่ในกลุ่มนี้เป็นลูกค้าเพศหญิงทั้งหมด ในขณะที่ลูกค้าในClusterที่ศูนย์ และClusterที่สองเป็นเพศชาย ฟีเจอร์Gender ที่ต่างกันอาจเป็นสาเหตุที่

ทำให้Clusterที่หนึ่งเกิดการทำนายถูกมากที่สุด ในขณะที่กลุ่มที่เกิดการทำนายผิด คือ Clusterที่ศูนย์ และClusterที่สอง ซึ่งเมื่อสำรวจข้อมูลพบว่าลูกค้าที่อยู่ใน Clusterที่ศูนย์ และClusterที่สองมีลักษณะบางอย่างที่คล้ายกัน เช่น เป็นลูกค้าเพศชาย ถือบัตรเครดิตอยู่ในช่วงระยะเวลา 2-4 ปี และมีพฤติกรรมการใช้จ่ายบัตรเครดิตน้อย โดยลูกค้าใน Clusterที่ศูนย์มีรายได้สูง และได้รับวงเงินบัตรเครดิตสูง ในขณะที่ลูกค้าClusterที่สองมีรายได้ปานกลางถึงสูง และได้รับวงเงินบัตรเครดิตน้อย เหล่านี้อาจเป็นสาเหตุที่ทำให้แบบจำลองทำนายลูกค้าในClusterที่สองเป็นลูกค้าใน Cluster ที่ศูนย์ และลูกค้าใน Clusterที่ศูนย์เป็นลูกค้าในClusterที่สอง

ผู้วิจัยได้ทำการแสดงค่าความสำคัญของแต่ละฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้ หรือ Feature importance เพื่อทำความเข้าใจกฎการตัดสินใจของแบบจำลองDecision Treeมากขึ้น ซึ่งฟีเจอร์ที่ส่งผลต่อการตัดสินใจของแบบจำลอง ได้แก่ Credit_Limit, Gender, Max_Income และTotal_Trans_Amt จะเห็นได้ว่าการจัดกลุ่มลูกค้า หรือ การทำ Customer Segmentation โดยใช้ 2 เทคนิค คือ การจัดกลุ่มลูกค้าด้วยแบบจำลองK-Means และทำนายกลุ่มลูกค้าด้วยแบบจำลองDecision Tree เป็น 2 เทคนิคที่ช่วยส่งเสริมประสิทธิภาพกัน โดยแบบจำลองK-Means จัดกลุ่มลูกค้าที่มีลักษณะเหมือนกันให้อยู่กลุ่มเดียวกัน ในขณะที่แบบจำลอง Decision Tree ทำให้เข้าใจการตัดสินใจของแบบจำลองในการเลือกใช้ฟีเจอร์ในการแบ่งกลุ่ม ซึ่งสำหรับงานด้านการตลาด หากธุรกิจมีลูกค้าใหม่เพิ่มเข้ามา สามารถนำ 2 เทคนิคนี้ไปช่วยงานด้านการจัดกลุ่มลูกค้า เพื่อส่งมอบสินค้า และบริการที่เหมาะสมกับลูกค้าได้แม่นยำยิ่งขึ้น

5.2 ข้อเสนอแนะ

5.2.1 แนะนำแผนการตลาด

ผลจากการวิเคราะห์การจัดกลุ่ม และลักษณะเฉพาะของกลุ่มลูกค้าที่ได้จากกฎการตัดสินใจของแบบจำลอง Decision Tree ผู้วิจัยขอเสนอแนวทางการตลาด เพื่อให้บริษัทบัตรเครดิตนำไปประยุกต์ใช้กับลูกค้าในแต่ละClusterดังนี้

1. ลูกค้าClusterที่ศูนย์ “Selective Spender” เป็นกลุ่มลูกค้าเพศชาย ส่วนใหญ่อยู่ในวัยกลางคนที่มีรายได้สูง ได้รับวงเงินบัตรเครดิตสูง มียอดใช้จ่ายบัตรเครดิตระดับน้อยถึงปานกลาง ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ระดับปานกลาง แคมเปญการตลาดสำหรับกลุ่มลูกค้า Selective Spender

- กระตุ้นการใช้จ่ายบัตรเครดิตด้วยโปรแกรมสะสมคะแนนแลกรับของรางวัล

- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะกับเพศชาย เพื่อกระตุ้นการใช้จ่ายผ่านบัตรเครดิต

2. ลูกค้าClusterที่หนึ่ง “Moderate Spender” เป็นกลุ่มลูกค้าเพศหญิงที่มีรายได้น้อยถึงปานกลาง ได้รับวงเงินบัตรเครดิตน้อย มียอดใช้จ่ายบัตรเครดิตระดับน้อยถึงปานกลาง ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ในระดับปานกลาง แคมเปญการตลาดสำหรับลูกค้ากลุ่ม Moderate Spender

- โปรแกรมผ่อนชำระ 0% สำหรับสินค้า หรือ บริการที่ราคาสูง
- กระตุ้นการใช้จ่ายบัตรเครดิตด้วยโปรแกรมสะสมคะแนนแลกรับของรางวัล
- เพิ่มวงเงินบัตรเครดิตให้กับลูกค้าที่มีรายได้ระดับปานกลางขึ้นไป ที่มีประวัติในการชำระหนี้ดี
- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะสำหรับเพศหญิง เพื่อกระตุ้นการใช้จ่ายผ่านบัตรเครดิต

3. ลูกค้าClusterที่สอง “Strategic Spender” เป็นกลุ่มลูกค้าเพศชาย ส่วนใหญ่อยู่ในวัยกลางคนถึงวัยเกษียณ มีรายได้ระดับปานกลางถึงสูง ได้รับวงเงินบัตรเครดิตน้อยจนถึงระดับปานกลาง มียอดใช้จ่ายบัตรเครดิตน้อยที่สุด ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ในระดับปานกลาง แคมเปญการตลาดสำหรับลูกค้ากลุ่มStrategic Spender

- แคมเปญลดอัตราดอกเบี้ยแบบขั้นบันไดเมื่อใช้จ่ายถึงยอดที่กำหนด
- กระตุ้นการใช้จ่ายด้วยโปรแกรมสะสมคะแนนแลกรับของรางวัล
- เพิ่มวงเงินบัตรเครดิตให้กับลูกค้าที่มีรายได้ระดับปานกลางขึ้นไปที่มีประวัติในการชำระหนี้ดี
- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะสำหรับวัยผู้ใหญ่จนถึงวัยเกษียณ

5.2.2 **ข้อเสนอแนะอื่น ๆ**

1. งานวิจัยนี้ใช้ชุดข้อมูลการใช้บัตรเครดิตของลูกค้าที่ประกอบไปด้วยฟีเจอร์เพศ อายุ สถานภาพ ระดับการศึกษา ระยะเวลาในการถือบัตรเครดิต จำนวนครั้งในการใช้บัตรเครดิต ยอดใช้จ่ายบัตรเครดิต และรายได้ของลูกค้าในการจัดกลุ่ม ซึ่งสำหรับงานทางด้านการตลาดหากมี

การเพิ่มฟีเจอร์อาชีพ ประเภทของสินค้า หรือ บริการที่ลูกค้าซื้อ ความถี่ในการใช้บัตรเครดิต อาจช่วยให้สามารถจัดกลุ่มลูกค้าได้หลากหลาย และมีลักษณะของลูกค้าที่เฉพาะเจาะจงมากยิ่งขึ้น ส่งผลให้ธุรกิจสามารถนำเสนอสินค้า หรือ บริการเฉพาะกลุ่ม ซึ่งเป็นข้อได้เปรียบทางธุรกิจหากทราบข้อมูลเชิงลึกของลูกค้า และนำมาวางแผนกลยุทธทางการตลาด

2. เนื่องจากงานวิจัยนี้ใช้แบบจำลอง K-Means ในการจัดกลุ่ม ดังนั้น อาจมีแบบจำลองประเภทอื่นที่มีความสามารถในการจัดกลุ่มได้อย่างมีประสิทธิภาพมากกว่า

บรรณานุกรม

Abdul-Rahman, S., Arifin, N. F. K., Hanafiah, M., & Mutalib, S. (2021). Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier. *International Journal of Advanced Computer Science and Applications(IJACSA)*, *12*(9). https://thesai.org/Publications/ViewPaper?Volume=12&Issue=9&Code=IJACSA&SerialNo=50

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, *3*. https://doi.org/10.12691/jcd-3-1-3 (Science and Education Publishing)

Aliyev, M., Ahmadov, E., & Gadirli, H. (2020). Segmentaing Bank Customers via RFM Model and Unsupervised Learning. https://arxiv.org/abs/2008.08662 (arXiv)

Denys Teslenko, A. S., Kyrylo Smelyakov, Oleksii Filipov. (2023). Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation. *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences*. https://doi.org/10.1109/ESTREAM59056.2023.10134796

Farshid Abdi, S. A. (2019). Customer Behavior Mining Framework (CBMF) using clustering and classification techniques. *Journal of Industrial Engineering International*. https://doi.org/10.1007/s40092-018-0285-3

Kotler, P., & Keller, K. L. (2016). *Marketing Management*. Pearson Education, Inc.

Kri˘zanic, S. z. (2020). Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*. https://doi.org/10.1177/1847979020908675

Mathes T., S. G., and Maheshwari A. (2023). A Machine Learning approach to segment the customers of online sales data for better and efficient marketing purposes. *International Conference on Artificial Intellegence and Knowledge Discovery in Concurrent Engineering*. https://doi.org/10.1109/ICECONF57129.2023.10084339 (IEEE Xplore)

Ritu Punhani, V. P. S. A., Sai Sabitha, Vinod Kumar Shukla. (2021). Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce. *2021 International Conference on Computational Intellegence and Knowledge Economy*. https://doi.org/10.1109/ICCIKE51210.2021.9410713

ภาคผนวก

ประวัติผู้เขียน