การศึกษาเทคนิคการจัดกลุ่มโดย Machine Learning ในหุ้น SET100 ของตลาดหลักทรัพย์แห่งประเทศไทย

Study of Clustering Techniques Using Machine Learning on SET100 Stocks of the Stock Exchange of Thailand

พัชร นพคุณมงคลชัย, ปรารถนา มินเสน* และ ปิยะชาติ เวียงนาค

Patchara Noppakunmongkolchai, Prathana Minsan* and Piyachart Wiangnak

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงใหม่

Faculty of Science and Technology, Chiang Mai Rajabhat University

E-mail : patcharaa47@gmail.com, pradthana_min@g.cmru.ac.th* and piyachart@g.cmru.ac.th

*Corresponding author

(Received: 11 January 2025, Revised: 26 February 2025, Accepted: 18 March 2025)

https://doi.org/10.57260/stc.2025.1042

บทคัดย่อ

การซื้อขายหลักทรัพย์เป็นหนึ่งในช่องทางในการลงทุน ที่นำมาซึ่งผลตอบแทนมหาศาลแต่ผลตอบแทนดังกล่าวก็มาพร้อมกับความเสี่ยงที่สูงมากเช่นเดียวกัน ดังนั้นภารกิจหลักของนักลงทุนที่สำคัญอีกสิ่งหนึ่งก็คือการกระจายความเสี่ยงในการลงทุนโดยมีเป้าหมายเพื่อลดความเสี่ยง บทความนี้นำเสนอวิธีการจัดกลุ่มโดย Machine Learning ในหุ้น SET 100 ของตลาดหลักทรัพย์แห่งประเทศไทย เป็นเทคนิคแบบ Unsupervised Learning ประกอบด้วยเทคนิค K-means, Hierarchical, DBSCAN และ Fuzzy C-means โดยเทคนิคเหล่านี้สามารถนำมาประยุกต์ใช้ในการจัดกลุ่มหลักทรัพย์ที่มีการเปลี่ยนแปลงคล้ายกันได้ ในการศึกษาครั้งนี้ได้ทำการเก็บรวบรวมข้อมูลย้อนหลังหุ้น SET 100 ของตลาดหลักทรัพย์แห่งประเทศไทย ซึ่งเป็นหุ้นสามัญ 100 บริษัทที่มีมูลค่าตลาดสูงและการซื้อขายมีสภาพคล่องสูงอย่างสม่ำเสมอถือว่าเป็นหุ้นที่มีปัจจัยพื้นฐานดี มีอัตราการเติบโตของบริษัทอย่างต่อเนื่อง เป็นที่สนใจของนักลงทุนทั่วไป ในการจัดกลุ่มได้นำเอา อัตราส่วนเงินปันผลตอบแทน (Dividend Yield) และ ค่าความเสี่ยงของหุ้น (BETA) ระยะเวลา 3 ปี 274 ตัว ผลการจัดกลุ่มทั้ง 4 วิธี พบว่า K-means สามารถจัดออกมาได้ 3 กลุ่มได้แก่กลุ่มที่ หุ้นปลอดภัยแต่ให้ปันผลต่ำ 192 ตัว หุ้นปันผลสูง ความเสี่ยงต่ำ 4 ตัว และ หุ้นเติบโตและมีความเสี่ยงสูง 78 ตัวส่วนเทคนิค Hierarchical สามารถจัดกลุ่มได้ 3 กลุ่ม ได้แก่กลุ่มที่ หุ้นปลอดภัยแต่ให้ปันผลต่ำ 163 ตัว หุ้นปันผลสูง ความเสี่ยงต่ำ 4 ตัว และ หุ้นเติบโตและมีความเสี่ยงสูง 107 ตัวส่วนเทคนิค DBSCAN แบ่งเป็น 2 กลุ่มได้แก่กลุ่มหลัก 266 ตัว นอกกลุ่ม 8 ตัวส่วนเทคนิค Fuzzy C-means แบ่งเป็น 3 กลุ่ม ได้แก่หุ้นปันผลสูงและความเสี่ยงต่ำ 89 ตัว หุ้นที่มีความสมดุลระหว่างความเสี่ยงและผลตอบแทน 138 ตัว และหุ้นที่มีความผันผวนสูง 47 ตัว

คำสำคัญ: เปรียบเทียบการแบ่งกลุ่ม K means Hierarchical DBSCAN Fuzzy C-Means ข้อมูล SET100

Abstract

Securities trading is one of the investment channels that can generate substantial returns, but these returns also come with high risks. Therefore, an important responsibility of investors is to diversify risk to reduce potential losses. This article presents a method for clustering stocks in the SET100 of the Stock Exchange of Thailand using Machine Learning. It applies Unsupervised Learning techniques, including K-means, Hierarchical Clustering, DBSCAN, and Fuzzy C-means, which can group securities with similar price movements. The study collected historical data of SET 100 stocks, which consist of 100 common stocks with high market capitalization and consistent liquidity. These stocks are fundamentally strong, with continuous business growth, making them attractive to investors. For the clustering process, the study used two key factors: Dividend Yield and Stock Risk (BETA) over a three-year period, covering 274 stocks. The results from the four clustering methods showed that K-means divided the stocks into three groups: low-risk stocks with low dividends (192 stocks), high-dividend low-risk stocks (4 stocks), and growth stocks with high risk (78 stocks). Hierarchical Clustering also resulted in three groups: low-risk stocks with low dividends (163 stocks), high-dividend low-risk stocks (4 stocks), and growth stocks with high risk (107 stocks). DBSCAN identified two groups: the main cluster (266 stocks) and outlier stocks (8 stocks). Fuzzy C-means classified the stocks into three groups: high-dividend low-risk stocks (89 stocks), stocks with balanced risk and return (138 stocks), and highly volatile stocks (47 stocks).

Keywords: Clustering, K-Means clustering, Hierarchical clustering, DBSCAN, Fuzzy C-Means, SET100 Data

บทนำ

ตลาดหลักทรัพย์แห่งประเทศไทย เป็นตลาดทุนที่ทำหน้าที่รวมการลงทุนของบริษัทหลากหลายกลุ่มอุตสาหกรรมเพื่อให้นักลงทุนได้พิจารณาเลือกลงทุนตามความต้องการ เพราะฉะนั้นทางเลือกในการหาผลตอบแทนจึงเป็นสิ่งหนึ่งของนักลงทุนที่จะตัดสินใจว่าจะลงทุนในรูปแบบใด แหล่งใด หรือลักษณะใด ด้วยความสำคัญของผลตอบแทนที่ต้องการจึงทำให้นักลงทุนให้ความสำคัญกับอัตราเงินปันผลตอบแทน มากกว่าที่จะลงทุนในลักษณะของการฝากเงินในสถาบันการเงิน นภสินธุ์ สานติวัตร และ สรศาสตร์ สุขเจริญสิน (2564) ได้กล่าวว่าผลตอบแทนที่นักลงทุนจะได้รับจากการลงทุนในการหลักทรัพย์ประกอบด้วย 2 ส่วน ได้แก่ การเพิ่มขึ้นของมูลค่าหลักทรัพย์และการได้รับเงินปันผล โดยเงินปันผลที่ได้รับนั้นเกิดจากผลการดำเนินงานที่มีกำไรของธุรกิจ และทำการจ่ายส่วนแบ่งออกมาในรูปของเงินปันผล โดยในบางบริษัทมีนโยบายจ่าย 2 ครั้ง 4 ครั้ง หรือ 1 ครั้งต่อปี และเมื่อผลการดำเนินงานดีมีการจ่ายเงินปันผลเพิ่มขึ้นก็จะทำให้อัตราเงินปันผลตอบแทนเพิ่มขึ้นตาม จึงเป็นส่วนประกอบที่ทำให้นักลงทุนเลือกลงทุนในธุรกิจนั้น ๆ นอกจากนี้ Zarah

(2019) ได้กล่าวว่าทฤษฎีสัญญาณ เป็นสิ่งสำคัญที่ทำให้เกิดการลงทุนของนักลงทุน เมื่อใดก็ตามที่มีการประกาศจ่ายเงินปันผลก็จะส่งผลทำให้เกิดการตัดสินใจลงทุน และส่งผลต่ออัตราเงินปันผลตอบแทนที่เพิ่มขึ้นหรือลดลงตามผลการดำเนินงานที่ธุรกิจได้ประกาศ อีกทั้ง Zarah (2018) พบว่า การประกาศจ่ายเงินปันผลมีอิทธิพลต่อการประเมินมูลค่าตลาดอีกด้วย (จักรกฤษณ์ มะโหฬาร และคณะ,)

เมื่อการลงทุนในตลาดหลักทรัพย์เป็นที่ต้องการของนักลงทุน เหตุด้วยผลตอบแทนที่สูงที่ออกมาในรูปของเงินปันผล และกำไรของการขายหลักทรัพย์ แต่ในขณะเดียวกันการลงทุนย่อมต้องมีการคัดเลือกดัชนีราคาที่มีผลการดำเนินงานอย่างต่อเนื่องและนิยม นั่นคือ บริษัทที่อยู่ในดัชนีราคาหุ้น SET100 ซึ่งเป็นดัชนีราคาที่แสดงระดับการเคลื่อนไหวของราคาหุ้นสามัญ 100 ตัว ที่มีมูลค่าตามราคาตลาดสูง มีการซื้อขายที่มีสภาพคล่องสูงอย่างสม่ำเสมอ มีสัดส่วนของผู้ถือหุ้นที่ผ่านเกณฑ์ตามที่คณะกรรมการกำกับดูแลตลาดหลักทรัพย์กำหนดไว้ เมื่อมีการลงทุนนักลงทุนย่อมต้องหาเครื่องมือมาใช้ในการพิจารณาการลงทุนเพื่อให้ได้ผลลัพธ์ที่สูง โดยมีการนำอัตราส่วนทางการเงินมาใช้ในการวิเคราะห์เช่น อัตราส่วนหนี้สินต่อส่วนของผู้ถือหุ้น อัตราผลตอบแทนจากสินทรัพย์รวม กำไรจากการดำเนินงาน กำไรสุทธิ งานวิจัยของวิวัฒน์วงศ์ บุญหนุน (2565) พบว่า อัตราส่วนหนี้สินต่อส่วนผู้ถือหุ้นมีอิทธิพลต่ออัตราเงินปันผลตอบแทน ชลวิช สุธัญญารักษ์ (2560) การลงทุนในหลักทรัพย์ที่มีผลการดำเนินงานในรูปของประสิทธิภาพกำไรที่ดีจะส่งผลทำให้เกิดความเสี่ยงต่ำและได้รับผลตอบแทนที่สูง นอกจากนี้สัดส่วนการถือหุ้นของผู้ถือหุ้นรายใหญ่ รายย่อย จำนวนคณะกรรมการบริษัทและอายุของกิจการ เป็นส่วนประกอบที่สำคัญต่อการบริหารธุรกิจ เมื่อกิจการมีโครงสร้างการถือหุ้น สัดส่วนการถือหุ้นที่ดีจะส่งผลทำให้ผลการดำเนินงานดีขึ้น

ดังนั้นในการวิจัยครั้งนี้จึงเป็นการศึกษาการใช้ Machine Learning เพื่อมาจัดกลุ่มในหุ้น SET100 (SET, 2024) ของตลาดหลักทรัพย์แห่งประเทศไทย โดยเก็บข้อมูลรายเดือนย้อนหลัง 3 ปี ตั้งแต่วันที่ 1 เดือนมกราคม 2022 ถึงวันที่ 1 เดือนธันวาคม 2024 เพื่อให้เห็นแนวโน้มของกลุ่มราคาของดัชนีหุ้นแต่ละตัว และผลตอบแทนในรูปของอัตราเงินปันผลที่คุ้มค่ากับการลงทุน เพื่อให้สามารถนำไปวิเคราะห์ก่อนที่จะพิจารณาลงทุนในบริษัทที่อยู่ในตลาดหลักทรัพย์

วัตถุประสงค์

เพื่อศึกษาการใช้ Machine Learning ด้วยวิธี K-means , Hierarchical , DBSCAN , Fuzzy C-means ในการจัดกลุ่มหุ้น SET100 ของตลาดหลักทรัพย์แห่งประเทศไทย

สมมติฐานของงานวิจัย

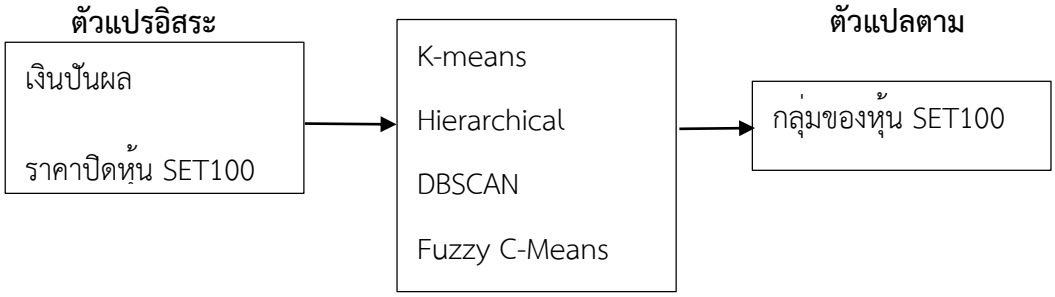
การใช้ Machine Learning สามารถนำมาจัดกลุ่มหุ้น SET100 ของตลาดหลักทรัพย์แห่งประเทศไทย

ระเบียบวิธีวิจัย

ขอบเขตงานวิจัย

ข้อมูลรายเดือนของ เงินปันผล ราคาปิดหุ้น SET100 จากตลาดหลักทรัพย์แห่งประเทศไทย ตั้งแต่วันที่ 1 เดือนมกราคม 2022 ถึง วันที่ 1 เดือนธันวาคม 2024

กรอบแนวคิดในการวิจัย



ภาพที่ 1 เทคนิคการจัดกลุ่มแบบ Machine Learning (ที่มา: คณะวิจัย, 2567)

วิธีการวิจัย

การศึกษาวิจัยครั้งนี้มีขั้นตอนวิธีการทดลอง ดังนี้

1. การจัดเตรียมข้อมูล

การวิจัยนี้มุ่งเน้นเพื่อการใช้ Machine Learning จัดกลุ่มหุ้น SET 100 ของตลาดหลักทรัพย์แห่งประเทศไทย ด้วยวิธี K-means, Hierarchical, DBSCAN, Fuzzy C-means โดยใช้ข้อมูล อัตราผลตอบแทนจากเงินปันผล และ ค่าความเสี่ยงของหุ้น จำแนกรายเดือน ตั้งแต่วันที่ 1 เดือนมกราคม 2022 ถึง วันที่ 1 เดือนธันวาคม 2024 เนื่องจากข้อมูลเป็น Miss Data ดังนั้นจึงมีข้อมูลจำนวน 274 ตัว ที่สามารถนำไปจัดกลุ่มข้อมูลได้

1.1 อัตราผลตอบแทนจากเงินปันผล (Dividend Yield)

อัตราเปรียบเทียบเงินปันผล จ่ายต่อหุ้นสามัญเทียบกับราคาตลาดของหุ้นสามัญ ใช้เพื่อดูว่าผลตอบแทนว่าหากลงทุนซื้อหุ้น ณ ระดับราคาตลาดปัจจุบัน จะมีโอกาสได้รับเงินปันผลคิดเป็นอัตราร้อยละของราคาหุ้น (SET investnow, 2568)

$$\text{อัตราผลตอบแทนจากเงินปันผล} = \frac{\text{เงินปันผลต่อหุ้น} \times 100}{\text{ราคาปิดตลาดของหุ้น}} \quad (1)$$

1.2 ค่าความเสี่ยงของหุ้น (BETA, β)

ค่าสถิติที่ใช้วัดความสัมพันธ์ระหว่าง ผลตอบแทนของหุ้น กับ ผลตอบแทนของตลาด โดยใช้หลักการของ ความแปรปรวนและความสัมพันธ์ของตัวแปร (Sharpe, 1964)

$$\beta = \frac{Cov(R_i, R_m)}{Var(R_m)} \quad (2)$$

โดย R_i ผลตอบแทนของหุ้นที่ต้องการวัด

R_m ผลตอบแทนของตลาด (เช่น ดัชนี SET100, ดัชนี SET100 TRI)

$Cov(R_i,R_m)$ ค่าสหสัมพันธ์ระหว่างผลตอบแทนของหุ้นกับตลาด

$Var(R_m)$ ค่าความแปรปรวนของผลตอบแทนตลาด

2. การจัดกลุ่ม(Clustering) แบบ Unsupervised Learning

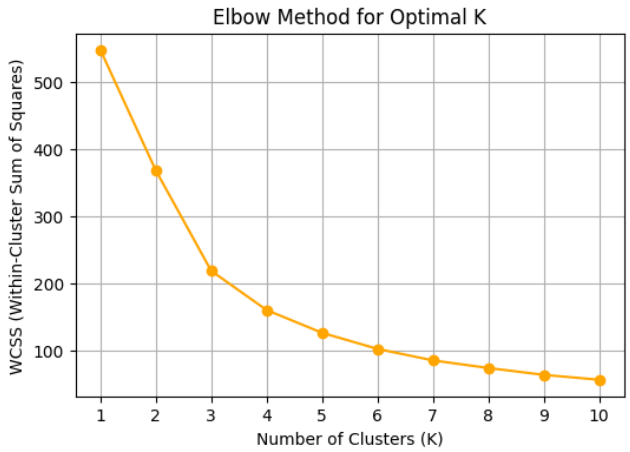
2.1 การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow Method

เทคนิคที่ใช้ในการหาค่าจำนวนกลุ่ม (k) ที่เหมาะสมสำหรับการทำการจัดกลุ่มด้วยวิธี K-means โดยวิธีนี้จะช่วยในการเลือกจำนวนกลุ่มที่ดีที่สุดที่สามารถอธิบายข้อมูลได้ดีที่สุด โดยจะทำการเปรียบเทียบ ค่า Within-Cluster Sum of Squares (WCSS) หรือ SSE (Sum of Squared Errors) ซึ่งเป็นค่าความผิดพลาดภายในกลุ่มที่เกิดขึ้นจากการจัดกลุ่ม

การใช้ Elbow Method จะทำการคำนวณค่า WCSS หรือ SSE สำหรับค่า k ต่าง ๆ แล้วทำการสร้างกราฟ โดยที่แกน x คือจำนวนกลุ่ม และแกน y คือค่า WCSS/SSE เมื่อเพิ่มจำนวนกลุ่มไปเรื่อย ๆ จะพบว่าค่า WCSS/SSE ลดลงจนถึงจุดหนึ่งที่ไม่ลดลงอย่างรวดเร็ว หรืออาจลดลงในอัตราที่ช้าลง จุดนี้จะเป็น "ข้อศอก" หรือ Elbow ซึ่งแสดงถึงจำนวนกลุ่มที่เหมาะสมที่สุดสำหรับข้อมูล

ขั้นตอนการใช้ Elbow Method:

1. เริ่มจากการกำหนดจำนวน k ตั้งแต่ 1 ถึง k ที่สูงสุด
2. คำนวณค่า WCSS หรือ SSE สำหรับแต่ละค่า k
3. สร้างกราฟระหว่าง k และ WCSS/SSE
4. ดูจุดที่ค่า WCSS/SSE เริ่มลดลงช้าลง ซึ่งจะเป็นค่าของ k ที่เหมาะสมที่สุด



ภาพที่ 2 กลุ่มที่เหมาะสม (ที่มา: คณะวิจัย, 2567)

จากภาพที่ 2 สามารถสังเกตเห็นว่า K = 3 เป็นจุดที่เหมาะสม เนื่องจาก

ค่าความคลาดเคลื่อน (WCSS) ลดลงอย่างรวดเร็ว จนถึง K = 3

หลังจาก K = 3 การลดลงของ WCSS เริ่มช้าลง ซึ่งหมายความว่าเพิ่มจำนวนกลุ่มมากกว่านี้ไม่สามารถทำให้ค่า WCSS ลดลงอย่างมากได้

2.2 การจัดกลุ่มด้วยวิธี K-means

K-means (Softnix, 2004) คือ วิธีการหนึ่งแบบ Unsupervised Learning คือการเรียนรู้แบบไม่ต้องสอน โดยหน้าที่หลักคือ การจัดกลุ่มมีตัวเลขประกอบอย่างน้อย 2 ตัวแปรขึ้นไป

วิธีการของ K-means มี 4 ขั้นตอนดังนี้

1. กำหนดจำนวนกลุ่มขึ้นมาก่อนเช่น 2 กลุ่ม (กำหนดเป็น C_1 และ C_2) และสุ่มตำแหน่งแกน x,y ให้กับ C_1 และ C_2 จะได้ $C_1\,(x_1,y_1)$ และ $C_2\,(x_2,y_2)$
2. พิจารณาตำแหน่งของสมาชิกแต่ละสมาชิกว่าอยู่ใกล้ค่าใดมากกว่ากันก็ให้ค่านั้นเป็นสมาชิกของ C นั้น จากตรงนี้จะรู้แล้วว่าสมาชิกแต่ละค่าอยู่ในกลุ่มใดระหว่าง C_1 และ C_2
3. ปรับ x,y ของ C_1 และ C_2ใหม่ให้อยู่ตรงกลางของกลุ่ม
4. ทำตามข้อ 2 และข้อ 3 อีกครั้งจนกว่า C_1 และ C_2 ตำแหน่งไม่เปลี่ยน

สูตรการคำนวน K-means Clustering โดยใช้ระยะทางยูคลิด (Euclidean Distance)

$$\mathrm{d}(\mathrm{x_i},\mathrm{c_j}) = \sqrt{\sum_{\mathrm{m=1}}^{\mathrm{n}} (\mathrm{x_{im}},\mathrm{c_{jm}})^2} \qquad (3)$$

โดย $d(x_i, c_j)$ คือระยะทางระหว่างจุดข้อมูล x_i และเซ็นทรอยด์ c_j

x_{im} คือพิกัดของจุดข้อมูล x_i ในมิติที่ m

c_{jm} คือพิกัดของเซ็นทรอยด์ c_j ในมิติที่ m

n คือจำนวนมิติของข้อมูล

2.3 การจัดกลุ่มด้วยวิธี Hierarchical

Hierarchical Clustering เป็นเทคนิคในการจัดกลุ่มข้อมูลที่ใช้วิธีการสร้างความสัมพันธ์ระหว่างข้อมูลจากการจัดเรียงตามระยะห่างระหว่างข้อมูล ซึ่งการจัดเรียงนี้จะมีการสร้างโครงสร้างเป็นต้นไม้ (Tree-like Structure) ที่เรียกว่า Dendrogram เพื่อแสดงความสัมพันธ์ระหว่างกลุ่ม

สูตร Hierarchical Clustering โดยใช้ระยะทางยูคลิด (Euclidean Distance)

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{x}_i - \mathrm{y}_i)^2} \qquad (4)$$

โดย x และ y คือเวกเตอร์ของจุดที่ต้องการคำนวณระยะทาง

n คือจำนวนมิติของเวกเตอร์

การรวมกลุ่มใช้ระยะทางระหว่างจุดที่ใกล้ที่สุดในกลุ่ม

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \quad (5)$$

โดย n_i และ n_j คือจำนวนในกุ่ม C_i และ C_j ตามลำดับ

$\bar{x}_i$ และ $\bar{x}_j$ คือค่าเฉลี่ยในกลุ่ม C_i และ C_j ตามลำดับ

2.4 การจัดกลุ่มด้วยวิธี DBSCAN

Density-Based Spatial Clustering of Applications with Noise: DBSCAN เป็นเทคนิคการจัดกลุ่ม ที่อิงตามความหนาแน่นของจุดข้อมูลในพื้นที่ ซึ่งช่วยในการค้นหากลุ่มของข้อมูลที่มีความหนาแน่นสูงในขณะที่ไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า

สูตรระยะทางยูคลิด (Euclidean Distance) (Weisstein, 2002)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (6)$$

โดย x และ y คือเวกเตอร์ของจุดข้อมูลสองจุดที่ต้องการคำนวณระยะทาง

n คือจำนวนมิติ (dimensions) ของข้อมูล

2.5 การจัดกลุ่มด้วยวิธี Fuzzy C-means

Fuzzy C-means Clustering (FCM) เป็นเทคนิคในการทำกลุ่มที่ยืดหยุ่นกว่า K-means เนื่องจากจุดข้อมูลแต่ละจุดสามารถเป็นสมาชิกของหลายกลุ่มได้ในเวลาเดียวกัน โดยมีระดับความเป็นสมาชิก (Membership Degree) ที่แตกต่างกัน ซึ่งช่วยในการจัดการกับข้อมูลที่มีความคลุมเครือและมีโครงสร้างที่ซับซ้อน

Fuzzy C-means เป็นวิธีการจัดกลุ่มข้อมูลที่กำหนดจำนวนกลุ่ม CCC ล่วงหน้า คล้ายกับ K-means โดยเริ่มต้นด้วยการสุ่มค่าศูนย์กลางของกลุ่ม (centroid) และค่าการเป็นสมาชิกของแต่ละจุดข้อมูล จากนั้นคำนวณ centroid ใหม่โดยอิงตามค่าการเป็นสมาชิก และอัปเดตค่าการเป็นสมาชิกของแต่ละจุดตามระยะห่างระหว่างจุดข้อมูลกับ centroid จุดที่อยู่ใกล้ centroid มากกว่าจะมีค่าการเป็นสมาชิกสูงกว่า กระบวนการนี้ทำซ้ำจนกว่าการเปลี่ยนแปลงของค่าการเป็นสมาชิกและ centroid จะน้อยมาก หรือครบจำนวนรอบที่กำหนด ค่าความเป็นสมาชิกใน Fuzzy C-means ถูกคำนวณตามสมการของ Bezdek (1981) ตามสมการ (7) และ (8)

การอัปเดตค่าการเป็นสมาชิก

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}} \quad (7)$$

โดย u_{ij} คือ ค่าการเป็นสมาชิกของจุดข้อมูล x_j ในกลุ่ม i

v_i คือ centroid ของกลุ่ม i

m คือ parameter ที่ควบคุมระดับความฟัซซี่ (โดยทั่วไป $m = 2$)

การคำนวณ centroid

$$v_i = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_j}{\sum_{j=1}^{N} u_{ij}^{m}} \quad (8)$$

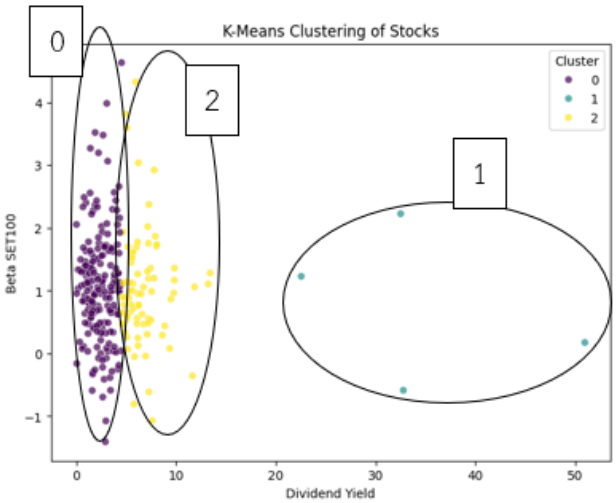
โดย v_i คือ centroid ของกลุ่ม i

x_j คือ จุดข้อมูล

u_{ij} คือ ค่าการเป็นสมาชิก

ผลการวิจัย

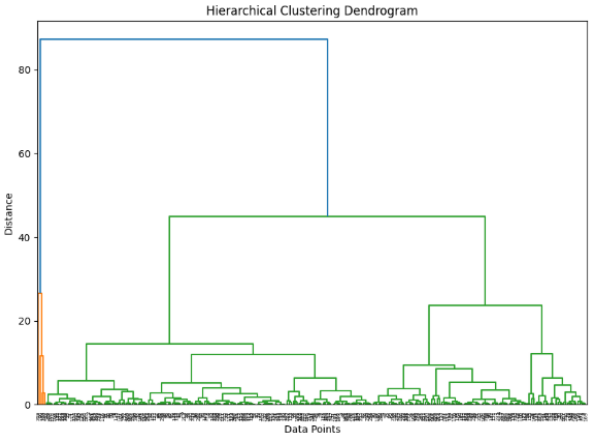
1. ผลการจัดกลุ่มด้วย K-means Clustering



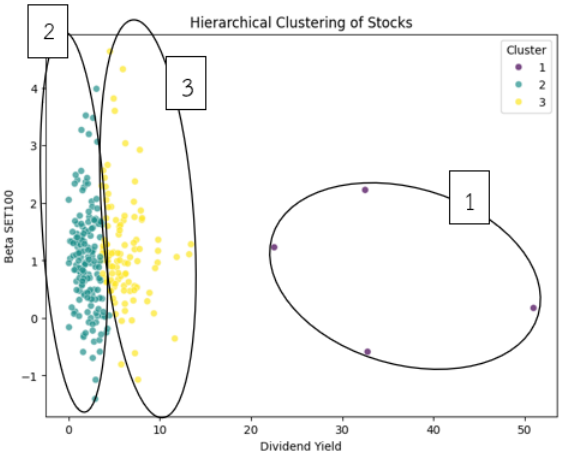
ภาพที่ 3 ผลการจัดกลุ่ม (Clustering) ของ K-means Clustering (ที่มา: คณะวิจัย, 2567)

K-means เป็นวิธีที่ใช้แบ่งข้อมูลออกเป็นกลุ่มโดยการหาจุดศูนย์กลาง (Centroid) ของแต่ละกลุ่ม จากภาพที่ 3 แสดงให้เห็นว่าหุ้นสามารถถูกแบ่งออกเป็น 3 กลุ่ม ได้แก่ Cluster 0, 1 และ 2 ข้อมูลส่วนใหญ่อยู่ในช่วงอัตราเงินปันผล ต่ำระหว่าง 0-10 และมีค่าความเสี่ยง อยู่ระหว่าง -1 ถึง 4 กลุ่มสีม่วง (Cluster 0) มีขนาดใหญ่ที่สุดและประกอบด้วยหุ้นที่มีอัตราเงินปันผลต่ำและกระจายตัวของค่าความเสี่ยงในช่วงกว้าง ส่วนกลุ่มสีเหลือง (Cluster 2) เป็นกลุ่มที่มีอัตราเงินปันผลสูงขึ้น ในขณะที่กลุ่มสีฟ้า (Cluster 1) แสดงหุ้นที่มีอัตราเงินปันผลสูงมาก ซึ่งบางตัวเกิน 50 และถือเป็น Outlier

2. ผลการจัดกลุ่มด้วย Hierarchical Clustering



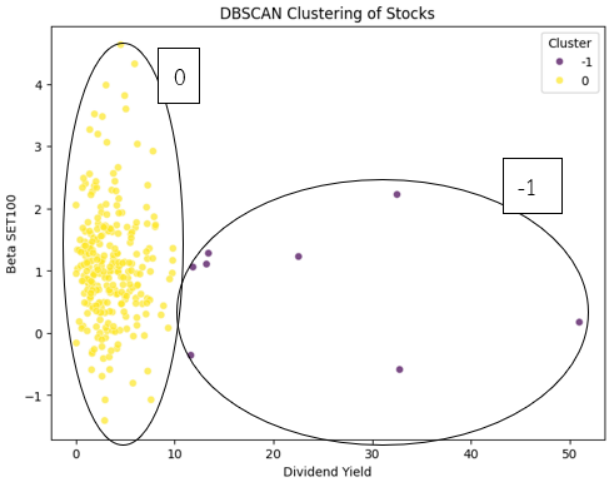
ภาพที่ 4 Dendrogram ของ Hierarchical
(ที่มา: คณะวิจัย, 2567)



ภาพที่ 5 การจัดกลุ่มของ Hierarchical
(ที่มา: คณะวิจัย, 2567)

การจัดกลุ่มแบบลำดับชั้น (Agglomerative Clustering) ถูกนำมาใช้และได้ผลลัพธ์เป็น 3 กลุ่ม ได้แก่ Cluster 1, 2 และ 3 จากภาพที่ 4 แสดงให้เห็นว่าการจัดกลุ่ม (Clustering) มีความแตกต่างจาก K-means โดยการกระจายตัวของข้อมูลแต่ละกลุ่มมีลักษณะที่ต่างกัน Cluster 1 (สีม่วง) ประกอบด้วยหุ้นที่มีอัตราเงินปันผลสูงมาก ซึ่งเป็น Outlier ส่วน Cluster 2 (สีฟ้า) มีความเสี่ยงที่กระจายตัวมากขึ้นและมีรูปแบบที่แตกต่างจาก K-means ในขณะที่ Cluster 3 (สีเหลือง) มีลักษณะการกระจุกตัวของข้อมูลที่คล้ายกับ K-means มากกว่า

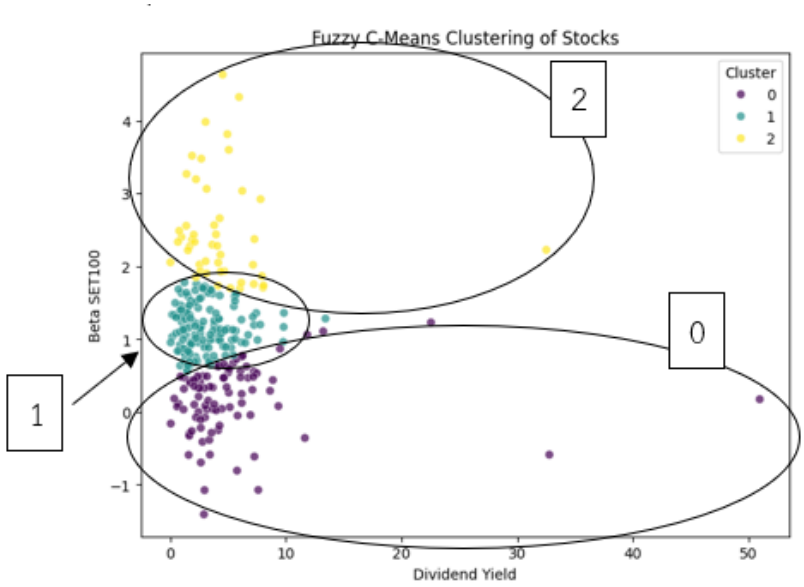
3. ผลการจัดกลุ่มด้วย DBSCAN



ภาพที่ 6 ผลการจัดกลุ่ม (Clustering) ของ DBSCAN (ที่มา: คณะวิจัย, 2567)

DBSCAN เป็นวิธีการจัดกลุ่ม (Clustering) ที่ใช้ความหนาแน่นของจุดข้อมูลในการกำหนดกลุ่ม ซึ่งได้ผลลัพธ์เป็น 2 กลุ่ม ได้แก่ Cluster 0 (สีเหลือง) และ -1 (Noise หรือ Outlier สีม่วง) เมื่อเปรียบเทียบกับ K-means และ Hierarchical แล้ว DBSCAN สามารถระบุ Outlier ได้ชัดเจนมากขึ้น โดยเฉพาะข้อมูลที่มีอัตราเงินปันผล สูงมาก เช่น มากกว่า 20 ซึ่งถูกจัดให้อยู่ใน Cluster -1 ในฐานะ Outlier ส่วน Cluster 0 เป็นกลุ่มหลักที่รวมข้อมูลที่มีอัตราเงินปันผลอยู่ในช่วง 0-10 และค่าความเสี่ยง ที่กระจายตัว

4. ผลการจัดกลุ่มด้วย Fuzzy C-means Clustering



ภาพที่ 7 ผลการจัดกลุ่ม (Clustering) ของ Fuzzy C-means (ที่มา: คณะวิจัย, 2567)

Fuzzy C-means เป็นวิธีที่ใช้แนวคิด Fuzzy Logic ซึ่งอนุญาตให้แต่ละจุดข้อมูลสามารถเป็นสมาชิกของหลายกลุ่มได้พร้อมกัน ผลลัพธ์ที่ได้แบ่งออกเป็น 3 กลุ่ม ได้แก่ Cluster 0, 1 และ 2 Cluster 0 (สีม่วง) ครอบคลุมจุดข้อมูลที่มีอัตราเงินปันผล ต่ำและมีค่าความเสี่ยง ที่กระจายตัว ในขณะที่ Cluster 1 (สีฟ้า) มีลักษณะคล้ายกับ Hierarchical Clustering โดยมีค่าความเสี่ยง สูงกว่า ส่วน Cluster 2 (สีเหลือง) เป็นกลุ่มที่แยกออกมาจากข้อมูลที่มีอัตราเงินปันผลสูงและมีความกระจายตัวของค่าความเสี่ยงมากขึ้น

ตารางที่ 1 ความแตกต่างของเทคนิคการจัดกลุ่มโดย Machine Learning ในแต่ละวิธี

Method	การกำหนดกลุ่มล่วงหน้า	จำนวนกลุ่ม	การจัดกลุ่ม	จุดเด่น	จุดด้อย
K-means	ใช่	3	กระจายตัวชัดเจน แต่ Outlier อาจไม่ได้ถูกจัดกลุ่มแยก	เร็ว ใช้งานง่าย	ไม่เหมาะกับข้อมูลที่มี Outlier มาก
Hierarchical	ไม่	3	จัดกลุ่มแบบลำดับชั้น อาจรวม Outlier ไว้ในกลุ่ม	ไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า	ใช้เวลาและหน่วยความจำมากขึ้น
DBSCAN	ไม่	2	สามารถระบุ Outlier ได้ดี	จัดการข้อมูลหนาแน่นได้ดี แยก Noise ออก	ไม่เหมาะกับข้อมูลที่กระจายตัวแบบไม่สม่ำเสมอ
Fuzzy C-means	ใช่	3	จุดข้อมูลอาจเป็นสมาชิกของหลายกลุ่มพร้อมกัน	หมาะกับข้อมูลที่มีการซ้อนทับกัน	คำนวณซับซ้อนและใช้เวลามากกว่าปกติ

การอภิปรายผล

ในการวิเคราะห์ข้อมูลหุ้นโดยการใช้ Machine Learning จัดกลุ่มหุ้น SET100 4 วิธี ได้แก่ K-means, Hierarchical Clustering, DBSCAN และ Fuzzy C-means พบว่าแต่ละวิธีให้ผลลัพธ์ที่แตกต่างกันไปตามลักษณะของข้อมูลและวิธีการที่ใช้ในการจัดกลุ่ม

K-means ใช้หลักการจัดกลุ่มโดยอาศัยจุดศูนย์กลาง (Centroid) ของแต่ละกลุ่ม ซึ่งสามารถแยกข้อมูลออกเป็น 3 กลุ่มได้อย่างชัดเจน โดยกลุ่มที่มี Dividend Yield ต่ำจะถูกจัดให้อยู่ใน Cluster 0 (สีม่วง) และ Cluster 2 (สีเหลือง) ที่มี Dividend Yield สูงขึ้น ส่วนกลุ่มสีฟ้า (Cluster 1) แสดงถึงหุ้นที่มี Dividend Yield สูงและเป็น Outlier ในกราฟ อย่างไรก็ตาม K-means อาจไม่สามารถจัดการกับ Outlier ได้ดีนัก เนื่องจากใช้ระยะห่างจาก Centroid เป็นหลักในการตัดสินใจ

Hierarchical Clustering เป็นการจัดกลุ่มแบบลำดับชั้น ซึ่งมีข้อดีคือไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า เวลาในการคำนวณมากกว่า ผลลัพธ์ที่ได้สามารถจัดกลุ่มได้ 3 กลุ่มเช่นเดียวกับ K-means แต่การจัดกลุ่ม ใน Hierarchical Clustering จะมีความกระจายของข้อมูลที่ต่างออกไป โดยเฉพาะ Cluster 1 (สีม่วง) ซึ่งประกอบไปด้วยหุ้นที่มี Dividend Yield สูงมาก ในขณะที่ Cluster 2 (สีฟ้า) มี Beta ที่กระจายกว้างขึ้น

DBSCAN เป็นวิธีที่เหมาะกับข้อมูลที่มี Outlier มาก เนื่องจากสามารถแยกข้อมูลที่เป็น Noise ออกมาได้อย่างชัดเจน ในกราฟพบว่า DBSCAN แบ่งข้อมูลออกเป็น 2 กลุ่มหลักคือ Cluster 0 (สีเหลือง) และ Noise (-1) ซึ่งเป็น Outlier (สีม่วง) ที่มี Dividend Yield สูงผิดปกติ ซึ่งเป็นลักษณะที่ต่างจาก K-means และ Hierarchical Clustering ที่ไม่สามารถแยก Outlier ได้ชัดเจน DBSCAN จะมีประโยชน์มากสำหรับข้อมูลที่มีลักษณะไม่สม่ำเสมอหรือกระจายตัวไม่เป็นกลุ่ม

Fuzzy C-means เป็นวิธีที่ใช้หลักการ Fuzzy Logic โดยอนุญาตให้ข้อมูลสามารถเป็นสมาชิกของหลายกลุ่มพร้อมกันได้ ซึ่งแตกต่างจากทั้ง K-means และ Hierarchical Clustering ที่จัดกลุ่ม ข้อมูลให้เป็นสมาชิกของกลุ่มเดียว ในผลลัพธ์ของ Fuzzy C-means พบว่า Cluster 0 (สีม่วง) ครอบคลุมข้อมูลที่มี Dividend Yield ต่ำและ Beta กระจายตัว ขณะที่ Cluster 1 (สีฟ้า) มีหุ้นที่มี Beta สูง และ Cluster 2 (สีเหลือง) จัดกลุ่มหุ้นที่มี Dividend Yield สูง

จากการวิเคราะห์นี้ การเลือกใช้ Machine Learning จัดกลุ่มหุ้น SET100 ขึ้นอยู่กับลักษณะของข้อมูลและวัตถุประสงค์ของการวิเคราะห์ หากต้องการแยก Outlier ออกจากข้อมูลหลัก DBSCAN จะเป็นวิธีที่เหมาะสมที่สุด แต่ถ้าต้องการจัดกลุ่มข้อมูลที่ชัดเจนและใช้งานง่าย K-means เป็นตัวเลือกที่ดี ส่วน Hierarchical Clustering จะเหมาะกับการศึกษาความสัมพันธ์และโครงสร้างของข้อมูล ในขณะที่ Fuzzy C-means จะเหมาะกับข้อมูลที่มีความซ้อนทับและต้องการให้ข้อมูลสามารถเป็นสมาชิกของหลายกลุ่มได้

การศึกษานี้สอดคล้องกับงานวิจัยการพยากรณ์ประสิทธิภาพของหุ้นในตลาดหลักทรัพย์แห่งประเทศไทย โดยใช้เทคนิค Multinomial Logistic Regression เพื่อจำแนกกลุ่มหุ้นได้เป็น 3 กลุ่ม (สุภาภรณ์ ทิมสำราญ และ ถวิล นิลใบ, 2021) ซึ่งวิจัยนี้ใช้ Machine Learning ได้เป็น 3 กลุ่มเหมือนกัน ทั้งสองงานวิจัยมีเป้าหมายเดียวกัน คือ จำแนกหุ้นออกเป็นกลุ่มต่างๆ ตามระดับความเสี่ยงและผลตอบแทน เพื่อช่วยให้นักลงทุนสามารถเลือกลงทุนได้อย่างเหมาะสม

บทสรุปและข้อเสนอแนะ

สรุป

วิจัยนี้ได้จัดทำเรื่องวิธีการจัดกลุ่มโดย Machine Learning ในหุ้น SET100 ของตลาดหลักทรัพย์แห่งประเทศไทย ด้วยเทคนิคแบบ Unsupervised Learning ด้วยเทคนิค K-means, Hierarchical, DBSCAN และ Fuzzy C-means ซึ่งสามารถนำมา ประยุกต์ใช้ในการจัดกลุ่ม หลักทรัพย์ที่มีการเปลี่ยนแปลง คล้ายกันได้ โดยทำการเก็บรวบรวมข้อมูลย้อนหลังของ SET 100 ของตลาดหลักทรัพย์แห่งประเทศไทย ซึ่งเป็นหุ้นสามัญ 100 ตัว โดยใช้ข้อมูลคือ เงินปันผล และ ราคาปิด ซึ่งลักษณะข้อมูลมีการจำแนกเป็นรายเดือน โดยใช้ข้อมูล ตั้งแต่วันที่ 1 เดือนมกราคม 2022 ถึง วันที่ 1 เดือนธันวาคม 2024 รวมข้อมูล 3 ปีมีทั้งหมด 300 ตัว แต่มีข้อมูลบางตัวไม่สมบูรณ์จึงตัดออกเป็น Missing Data ดังนั้นจึงเหลือข้อมูลเพียง 274 ตัว ซึ่งผู้วิจัยได้ทำการจัดกลุ่ม แล้วทั้ง 4 วิธีได้ผลดังนี้ K-means มีทั้งหมด 3 กลุ่มมีกลุ่มที่ หุ้นปลอดภัยแต่ให้ปันผลต่ำ 192 ตัว หุ้นปันผลสูง ความเสี่ยงต่ำ 4 ตัว และ หุ้นเติบโตและมีความเสี่ยงสูง 78 ตัว, Hierarchical แบ่งเป็น 3 กลุ่ม มี

กลุ่มที่หุ้นปลอดภัยแต่ให้ปันผลต่ำ 163 ตัว หุ้นปันผลสูง ความเสี่ยงต่ำ 4 ตัว และ หุ้นเติบโตและมีความเสี่ยงสูง 107 ตัว, DBSCAN แบ่งเป็น 2 กลุ่มคือ กลุ่มหลัก 266 ตัว นอกกลุ่ม 8 ตัว, Fuzzy C-means แบ่งเป็น 3 กลุ่ม หุ้นปันผลสูงและความเสี่ยงต่ำ 89 ตัว มีความสมดุลระหว่างความเสี่ยงและผลตอบแทน 138 ตัว มีความผันผวนสูง 47 ตัวจะเห็นได้ว่าแต่ละกลุ่มมีการจัดกลุ่มที่แตกต่างกัน

หากข้อมูลมี Outlier จำนวนมาก การใช้ DBSCAN อาจเหมาะสมที่สุด, หากต้องการเข้าใจความสัมพันธ์ของข้อมูลเชิงโครงสร้าง Hierarchical Clustering อาจให้ข้อมูลเชิงลึกมากขึ้น, ถ้าต้องการจัดกลุ่มที่ง่ายและรวดเร็ว K-means เป็นทางเลือกที่ดี, หากหุ้นมีพฤติกรรมที่เปลี่ยนแปลงและอาจเป็นสมาชิกของหลายกลุ่ม Fuzzy C-means จะช่วยให้การวิเคราะห์มีความยืดหยุ่นมากขึ้น

ข้อเสนอแนะ

1. การทำวิจัยนี้เป็นการจัดกลุ่มหุ้น SET100 สามารถนำไปใช้ในการวิเคราะห์กับหุ้น SET50, SET1000 ได้
2. ปัจจุบันความนิยมในการลงทุนหุ้นมีมากในกลุ่มวัยรุ่นหรือเริ่มต้นทำงาน ซึ่งคนกลุ่มนี้ยังขาดประสบการณ์หรือความเข้าใจในการลงทุน ดังนั้นการใช้ Machine Learning เพื่อในการจัดกลุ่มเพื่อการตัดสินใจในการลงทุนและประเมินความเสี่ยงในการลงทุนได้

เอกสารอ้างอิง

จักรกฤษณ์ มะโหฬาร, สุมาลี นามโชติ และ นฤพล อ่อนวิมล. (2567). ปัจจัยที่มีอิทธิพลต่ออัตราเงินปันผลตอบแทนของบริษัทที่จดทะเบียนในตลาดหลักทรัพย์แห่งประเทศไทยดัชนีSET100. *วารสารสุทธิปริทัศน์, 38*(1), 83-99. สืบค้นจาก https://so05.tci-thaijo.org/index.php/DPUSuthiparithatJournal/article/article/view /269045/181752

ชลวิช สุธัญญารักษ์. (2560). ความเสี่ยงและอัตราผลตอบแทนของหลักทรัพย์หมวดการท่องเที่ยวและสันทนาการ โดยใช้แบบจำลอง CAPM. *วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์, 11*(3), 13-23. สืบค้นจาก https://so02.tci-thaijo.org/index.php/JournalGradVRU/article/view/108212

นภสินธุ์ สานติวัตร และ สรศาสตร์ สุขเจริญสิน. (2564). การศึกษาความสัมพันธ์ระหว่างการจ่ายเงินปันผลและการเปลี่ยนแปลงของกำไรในอนาคต. *จุฬาลงกรณ์ธุรกิจปริทัศน์, 43*(169), 1-21. สืบค้นจาก https://so01.tci-thaijo.org/index.php/CBSReview/article/view/250228

สุภาภรณ์ ทิมสำราญ และ ถวิล นิลใบ. (2564). การพยากรณ์ประสิทธิภาพของหุ้น ในตลาดหลักทรัพย์แห่งประเทศไทย. *วารสารรามคำแหง ฉบับบัณฑิตวิทยาลัย, 4*(3), 75-87. สืบค้นจาก http://www.rujogs.ru.ac.th/Journals/ClickLink/149

Bezdek, J. C. (1981). Objective function clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 3*(3), 376–386. Retrieved from https://doi.org/10.1109/TPAMI.1981.4767076

SET. (2567). *ข้อมูล SET100 จากตลาดหลักทรัพย์แห่งประเทศไทย.* สืบค้นจาก https://shorturl .asia/4Shpr

SET investnow. (2568). *อัตราผลตอบแทนจากเงินปันผล.* สืบค้นจาก https://www.setinvestnow.com /th/glossary/dividend-yield

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442. Retrieved from https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Softnix. (2004). *K-means*. Retrieved from https://shorturl.asia/cdo3e

Weisstein, E. W. (2002). *Euclidean distance*. MathWorld – A Wolfram Web Resource. Retrieved from https://mathworld.wolfram.com/EuclideanDistance.html

Zarah, P. (2018). Ability of net income in predicting dividend yield operating cash flow as a moderating variable. *Archives of Business Research, 6*(1), 226-234. Retrieved from http://dx.doi.org/10.14738/abr.61.4128

Zarah, P. (2019). Empirical evidence of market reactions based on signaling theory in Indonesia stock exchange. *Investment Management and Financial Innovations, 16*(2), 66-77. http://dx.doi.org/10.21511/imfi.16(2).2019.0