การศึกษาเปรียบเทียบวิธีการเติมค่าสูญหายสำหรับการจัดกลุ่มข้อมูลที่มีการสูญหาย

A comparative study on imputation methods for clustering with missing data

พินิจชญา โชติถนอม[1], นิวรรณ วัฒนกิจรุ่งโรจน์[1*] และ ศศิพร ทองแม้น[2]

Pinidchaya Chottanom[1], Niwan Wattanakitrungroj[1*] and Sasiporn Tongman[2]

Received: 8 December 2024 ; **Revised**: 13 January 2025 ; **Accepted**: 19 March 2025

บทคัดย่อ

การจัดการข้อมูลสูญหายด้วยการเติมข้อมูลมีความท้าทายต่อการวิเคราะห์ข้อมูล โดยเฉพาะข้อมูลที่มีน้อยและมีความสำคัญไม่สามารถตัดออกได้ โดยทั่วไปข้อมูลสูญหายอาจเติมโดยใช้ค่ากลางของข้อมูล เช่น ค่าเฉลี่ยหรือฐานนิยม (Mean/Mode) อย่างไรก็ตาม วิธีการดังกล่าวไม่เหมาะสมสำหรับข้อมูลที่มีการกระจายตัวที่หลากหลาย การประยุกต์ใช้การเรียนรู้ของเครื่องเพื่อทำนายค่าที่สูญหายไปจึงเป็นทางเลือกหนึ่ง งานวิจัยนี้ศึกษาผลกระทบของวิธีการเติมค่าข้อมูลสูญหายต่อประสิทธิภาพในการจัดกลุ่มโดยใช้ข้อมูลจำลองสองมิติและข้อมูลจริงเกี่ยวกับลูกค้าธนาคาร ซึ่งแต่ละชุดข้อมูลได้ทดลองให้มีอัตราการสูญหายสามระดับ ต่อมาเปรียบเทียบการเติมค่าข้อมูลสูญหาย 5 วิธี ได้แก่ วิธี Mean/Mode วิธี K-Nearest Neighbors (KNN) วิธี Multivariate Imputation by Chained Equations (MICE) KNN วิธี MICE Random Forest และวิธี MICE Bayesian Ridge จากนั้นจัดกลุ่มด้วยวิธี Hierarchical Clustering และวิธี Density-based Spatial Clustering of Applications with Noise (DBSCAN) Clustering โดยกำหนดค่าพารามิเตอร์หลากหลาย เพื่อเปรียบเทียบประสิทธิภาพ ผลการทดลองพบว่า การเติมข้อมูลสูญหายในข้อมูลที่มีอัตรการสูญหายทั้งสามระดับด้วยวิธี KNN วิธี MICE KNN และ วิธี MICE Random Forest จะให้ค่า Adjusted Rand Index (ARI) ค่า Normalized Mutual Information (NMI) ค่า Fowlkes-Mallows Index (FMI) และ ค่า Purity สูงสุดเป็นอันดับต้นเสมอ นอกจากนี้ แต่ละชุดข้อมูลจะพบวิธีการจัดกลุ่มที่เหมาะสมแตกต่างกันไป ดังนั้น การใช้วิธีการที่เหมาะสมทั้งวิธีการจัดกลุ่มและวิธีการเติมข้อมูลสูญหายแทนที่จะใช้วิธีการเติมด้วยค่ากลางแบบดั้งเดิมจะช่วยเพิ่มประสิทธิภาพการจัดกลุ่มได้มากขึ้น โดยได้ค่าประสิทธิภาพที่ดีของตัวชี้วัดเหล่านี้อยู่ในช่วงค่า 0.81 ถึงค่าใกล้ 1

คำสำคัญ: การจัดกลุ่มแบบลำดับชั้น, การจัดกลุ่ม DBSCAN, ข้อมูลสูญหาย, วิธีการเติมค่าสูญหาย

Abstract

Managing missing data by data imputation always challenges the subsequent data analysis process. Especially when a small amount of important data cannot be eliminated. In General, missing data can be imputed by filling its central value, such as the mean or mode value, although such methods might be unsuitable for every dataset with various distributions. Therefore, machine learning approaches to predict missing values are an interesting option. In this research, the impact of five missing data imputation methods was experimented with by clustering data with three missing rate levels. Two simulated 2D datasets and one real-world bank customer dataset were used in this work. Five missing data imputation methods, including Mean/Mode, K-Nearest Neighbors (KNN), Multivariate Imputation by Chained Equations (MICE), KNN, MICE Random Forest, and MICE Bayesian Ridge methods, were compared. After imputation, the data were clustered by both Hierarchical Clustering and Density-based Spatial Clustering of

[1] คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

[2] สาขาเทคโนโลยีชีวภาพ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

[1] School of Information Technology, King Mongkut's University of Technology Thonburi

[2] Department of Biotechnology, Faculty of Science and Technology, Thammasat University

* Corresponding author; e-mail address: watta.niwan@gmail.com

Applications with Noise (DBSCAN) Clustering approaches with various parameter setups for comparison. The experimental results show that KNN, MICE KNN, and MICE Random Forest imputation methods usually perform well, and consistently yield high values of Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index (FMI), and Purity regardless of data missing rate level. In addition, each clustering method may be suitable for each dataset differently. Therefore, using both an appropriate clustering approach and an imputation method to fill up data in this study, rather than traditional mean and mode imputation techniques, can expressively enhance clustering performance values ranging from 0.81 to nearly 1.

Keywords: Hierarchical clustering, DBSCAN clustering, missing data, imputation method

บทนำ

การดำเนินธุรกิจในปัจจุบันให้ความสำคัญการจัดการข้อมูลลูกค้าเป็นอย่างยิ่ง การทำความเข้าใจความต้องการและการวิเคราะห์กลุ่มลูกค้าปัจจุบันและอนาคตเป็นเรื่องที่กำลังท้าทายการดำเนินธุรกิจ โดยเฉพาะธุรกิจการเงิน การบริการ อสังหาริมทรัพย์ การศึกษา และการซื้อสินค้า เป็นต้น การจัดกลุ่มลูกค้า (Customer clustering) คือ การจัดให้กลุ่มลูกค้าที่มีพฤติกรรมใกล้เคียงกันหรือเหมือนกันไว้ในกลุ่มเดียวกัน (Maryani, 2018) วิธีการจัดกลุ่มลูกค้ามีหลากหลายรูปแบบ ซึ่งการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เป็นหนึ่งในวิธีที่นิยมใช้ โดยสามารถจัดกลุ่มลูกค้าได้โดยไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า และสร้างโครงสร้างการจัดกลุ่มที่สามารถแสดงความสัมพันธ์ของข้อมูลในระดับต่าง ๆ ได้อย่างชัดเจน แต่ในการวิเคราะห์กลุ่มลูกค้าจากข้อมูลที่ได้จากการเก็บรวบรวมจริง มักประสบปัญหาข้อมูลสูญหาย (Missing Data) ซึ่งส่งผลให้การจัดกลุ่มลูกค้าอาจไม่แม่นยำ ดังนั้น การจัดการข้อมูลสูญหายจึงมีความสำคัญอย่างยิ่ง ดังนั้นจึงต้องทำการทำความสะอาดข้อมูล (Data Cleansing) ซึ่งเป็นกระบวนการในการตรวจสอบ แก้ไข ลบ แทนที่ และจัดรูปแบบของข้อมูลที่ไม่สมบูรณ์หรือข้อมูลที่ผิดพลาด เพื่อให้ข้อมูลมีความถูกต้องเป็นระเบียบพร้อมสำหรับการนำไปใช้งานอย่างมีประสิทธิภาพ นอกจากนี้ การจัดการข้อมูลสูญหายยังจำเป็นต้องดำเนินการผ่านกระบวนการนี้ด้วย เพื่อให้ได้ข้อมูลที่สมบูรณ์และเหมาะสมสำหรับการวิเคราะห์ต่อไป การจัดการข้อมูลสูญหายมีหลายวิธี เช่น การตัดข้อมูลที่มีการสูญหายทิ้งไป การประมาณค่าพารามิเตอร์ การเติมค่าข้อมูลสูญหายโดยค่าประมาณจากการคำนวณ (พัชนา สุวรรณแสน, 2562; Batista *et al.,* 2003; Sanjar *et al.,* 2020)

ในงานวิจัยนี้สนใจศึกษาอิทธิพลของวิธีการจัดการข้อมูลสูญหายในการจัดกลุ่มข้อมูลลูกค้า วิเคราะห์ความสัมพันธ์ของวิธีการจัดการข้อมูลสูญหายที่มีปริมาณข้อมูลสูญหายแตกต่างกันที่มีต่อประสิทธิภาพการจัดกลุ่ม

หลักการและงานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้กล่าวถึง ประเภทของข้อมูลสูญหาย วิธีการจัดการข้อมูลสูญหาย การจัดกลุ่มข้อมูล และงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

1. ประเภทของข้อมูลสูญหาย

ในการวิเคราะห์ข้อมูล สามารถแบ่งประเภทข้อมูลสูญหายได้เป็น 3 ประเภท ดังนี้

1) การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely At Random; MCAR) เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นแบบสุ่มจากค่าทั้งหมด ความน่าจะเป็นที่ข้อมูลสูญหายที่จุดใด ๆ ไม่มีความสัมพันธ์กับค่าของข้อมูลอื่นในตัวแปรที่มีข้อมูลสูญหายเองหรือตัวแปรอื่น

2) การสูญหายแบบสุ่ม (Missing At Random: MAR) เป็นลักษณะของข้อมูลสูญหายที่ขึ้นกับตัวแปรอื่นในชุดข้อมูล แต่ไม่ขึ้นกับค่าของตัวแปรที่สูญหายเอง

3) การสูญหายแบบไม่สุ่ม (Missing Not At Random; MNAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นแบบสุ่ม โดยค่าของข้อมูลสูญหายมีความสัมพันธ์กับค่าของข้อมูลที่ไม่สูญหายในตัวแปรเดียวกันรวมถึงตัวแปรตัวอื่น เช่น ลูกค้าที่รายได้ต่ำอาจไม่กรอกข้อมูล Income เพื่อหลีกเลี่ยงการเปิดเผย

2. วิธีการจัดการข้อมูลสูญหาย

การจัดการข้อมูลสูญหายมีได้หลายวิธี โดยทั่วไปเราสามารถแบ่งวิธีการจัดการข้อมูลสูญหายเป็น 2 แบบหลัก ได้แก่

1) การตัดข้อมูลสูญหาย (Ignoring and discarding data) ใช้ในกรณีที่ข้อมูลสูญหายเกิดขึ้นหลายตัวแปร แต่สัดส่วนการสูญหายไม่ควรเกิน 5% ของข้อมูลทั้งหมด มีวิธีลบข้อมูล 2 แบบ ได้แก่ การลบแบบ Listwise คือ การลบข้อมูลทั้งหมดของแถวที่มีค่าสูญหาย เหมาะสำหรับข้อมูลที่สูญหายน้อย และการลบแบบ Pairwise deletion มีการวิเคราะห์ความสัมพันธ์ระหว่างคู่ของตัวแปร โดยจะวิเคราะห์ข้อมูลจากข้อมูล

ส่วนที่ไม่มีค่าสูญหายทั้งสองตัวแปร วิธีนี้มีข้อดีในส่วนของการใช้งานข้อมูลอย่างเต็มที่และมีประสิทธิภาพแต่ซับซ้อนกว่าวิธี Listwise

2) การประมาณค่าพารามิเตอร์ (Parameter estimation) การประมาณค่าความเป็นไปได้มากที่สุด (Maximum likelihood estimation) ซึ่งคำนวณจากข้อมูลทั้งหมด โดยค่าความเป็นไปได้มากที่สุดจะถูกคำนวณโดยแยกคำนวณจากข้อมูลที่ไม่มีการสูญหายของบางตัวแปร และคำนวณจากข้อมูลในทุกตัวแปร จากนั้นค่าความเป็นไปได้ที่มากที่สุดจากการคำนวณทั้งสองจะถูกนำใช้ในการประมาณค่าข้อมูลสูญหาย วิธีการเติมค่าข้อมูลสูญหายด้วยค่าประมาณเป็นวิธีการประมาณค่าสูญหายโดยเอาหลักการทางคณิตศาสตร์มาเติมค่าที่สูญหายไป เช่น การแทนข้อมูลสูญหายของตัวแปรนั้นด้วยค่าเฉลี่ย ค่ามัธยฐาน หรือค่าฐานนิยมของตัวแปร การแทนค่าข้อมูลสูญหายด้วยค่าทำนายด้วยการวิเคราะห์ถดถอย (Regression Method) การแทนค่าด้วยวิธีเคเนียร์เรสเบอร์ (K-Nearest Neighbors; KNN) วิธีเกรย์เคเนียร์เรสเบอร์ (Grey KNN) และวิธีการแทนค่าแบบพหุด้วยสมการลูกโซ่ (Multivariate Imputation by Chained Equations; MICE) เป็นต้น (Pan *et al.,* 2015)

3. การจัดกลุ่มข้อมูล

ในงานนี้ศึกษาประสิทธิภาพของการเติมข้อมูลสูญหายที่มีผลต่อวิธีการจัดกลุ่มข้อมูล ซึ่งการวิเคราะห์กลุ่มข้อมูล (Cluster Analysis) เป็นเทคนิคจัดกลุ่มหน่วยข้อมูล หรือเป็นการแบ่ง คน สัตว์ สิ่งของ องค์กร ฯลฯ ออกเป็นกลุ่ม โดยที่ภายในกลุ่มแต่ละกลุ่มหน่วยตัวอย่างจะต้องมีลักษณะคล้ายคลึงกัน และระหว่างกลุ่มหน่วยตัวอย่างจะต้องมีลักษณะที่แตกต่างกัน การจัดกลุ่มเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นวิธีการจัดกลุ่มข้อมูลโดยไม่ต้องใช้ข้อมูลตัวอย่างที่มีการระบุป้ายกำกับ (Labels) ไว้ล่วงหน้า การจัดกลุ่มช่วยในการค้นหารูปแบบ หรือโครงสร้างที่ซ่อนอยู่ในข้อมูล การจัดกลุ่มมีหลายวิธี โดยในงานนี้ทำการทดลองกับวิธีการจัดกลุ่มที่นิยมสองวิธี ได้แก่ Density-based Clustering และ Hierarchical Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) เป็นวิธีการจัดกลุ่มข้อมูลโดยอาศัยความหนาแน่นของข้อมูลในบริเวณหนึ่ง เหมาะสำหรับการจัดการข้อมูลที่มีรูปทรงไม่แน่นอนและสามารถจัดการข้อมูลที่เป็นสัญญาณรบกวน (Noise) โดยใช้พารามิเตอร์สำคัญสองตัว ได้แก่ ค่า Eps สำหรับกำหนดรัศมีของพื้นที่ใกล้เคียง และ ค่า MinPts สำหรับกำหนดจำนวนจุดขั้นต่ำในพื้นที่นั้นที่ถือว่าเป็นกลุ่ม จุดที่มีความหนาแน่นเพียงพอจะถูกจัดเป็น “Core Point” และขยายกลุ่มโดยรวมจุดใกล้เคียงที่เชื่อมโยงกันผ่าน Core Point วิธีนี้มีประโยชน์ในกรณีที่จำนวนคลัสเตอร์ไม่ชัดเจนล่วงหน้า และมีข้อได้เปรียบในด้านการจัดการข้อมูลที่ซับซ้อนเมื่อเทียบกับวิธีการจัดกลุ่มแบบดั้งเดิม (Ester *et al.,* 1996)

Hierarchical Clustering คือ การจัดกลุ่มข้อมูลด้วยวิธีการรวมกลุ่มทีละขั้น (Agglomerative) หรือลดกลุ่มทีละขั้น (Divisive) โดยการสร้างโครงสร้างแบบลำดับชั้น (Hierarchical Structure) เป็นวิธีการจัดกลุ่มที่ไม่จำเป็นต้องทราบจำนวนจำนวนคลัสเตอร์มาก่อน และสามารถใช้กับข้อมูลผสมโดยใช้ร่วมกับมาตรวัดระยะห่างสำหรับข้อมูลผสม เช่น มาตรวัด Gower’s Distance เป็นต้น แตกต่างจาก K-Means Clustering ที่ต้องทราบจำนวนกลุ่มมาก่อน (กัลยา วานิชบัญชา, 2558)

4. งานวิจัยที่เกี่ยวข้อง

Wood *et al*. (2004) ได้สำรวจผลงานวิจัยที่มีการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet and New England Journal of Medicine จำนวน 71 เรื่อง พบว่า งานวิจัยส่วนใหญ่ (ร้อยละ 89) มีข้อมูลสูญหายและไม่มีการจัดการก่อนการวิเคราะห์ ซึ่งระดับความรุนแรงของผลกระทบนี้ขึ้นอยู่กับคุณลักษณะและความสำคัญของข้อมูลสูญหาย จึงควรมีการศึกษาจัดการข้อมูลสูญหายอย่างเหมาะสม

Batista *et al.* (2003) ได้ทำการทดลองทดสอบประสิทธิภาพของการเติมค่าข้อมูลสูญหายด้วย KNN กับ Mean/Mode พบว่าวิธี KNN มีประสิทธิภาพในการประมาณค่าสูญหายที่ดีกว่า นอกจากนั้น ผู้วิจัยยังได้พัฒนาวิธี KNN ให้มีประสิทธิภาพมากขึ้นโดยวิธีการถ่วงน้ำหนัก

พัชนา สุวรรณแสน (2562) ใช้วิธีการจัดการข้อมูลสูญหาย ด้วยวิธี KNN พบว่าสามารถเป็นแนวทางในการทำให้ได้ข้อมูลที่มีความสมบูรณ์เหมาะสมในการดำเนินงานในขั้นตอนต่อไปของการวิจัย

อุมาพร ยกกำพล และคณะ (2561) เปรียบ เทียบประสิทธิภาพวิธีการจัดกลุ่มข้อมูลแบบลำดับขั้น และการจัดกลุ่มแบบเคมีน สำหรับข้อมูลผสมเชิงหมวดหมู่ กับเชิงตัวเลข โดยแปลงตัวแปรเชิงหมวดหมู่ เป็นค่าทวิภาคและแปลงตัวแปรให้เป็นบรรทัดฐาน พบว่าวิธีการจัดกลุ่มแบบลำดับขั้นมีประสิทธิภาพดีกว่าวิธีการจัดกลุ่มแบบเคมีน แต่การแปลงตัวแปรให้เป็นบรรทัดฐานจะช่วยเพิ่มประสิทธิภาพการจัดกลุ่มแบบเคมีนได้มากขึ้น

ศรีรักษ์ ศรีทองชัย (2566) ศึกษาวิธีการเติมค่าข้อมูลสูญหายในการพยากรณ์ความเข้มข้นของ PM 2.5 ด้วยแบบจำลองโครงข่ายประสาทเทียม ด้วยวิธีค่าเฉลี่ย วิธีการประมาณค่าในช่วงแบบเชิงเส้น วิธี KNN และ MICE พบว่าวิธี KNN และ วิธี MICE ให้ค่าความผิดพลาดน้อยที่สุด ทั้งสองวิธีให้ผลลัพธ์ดีกว่าวิธีค่าเฉลี่ยและวิธีการประมาณค่าในช่วงแบบเชิงเส้น แต่อย่างไรก็ตามเงื่อนไขของแบบจำลองมีอิทธิพลต่อความแม่นยำของการใช้ค่าพยากรณ์จากวิธีดังกล่าว

การดำเนินงานวิจัย

ในหัวข้อนี้จะกล่าวถึงการดำเนินงานวิจัยประกอบด้วย ข้อมูลที่ใช้ การเติมค่าสูญหาย การจัดกลุ่ม การประเมินผล ดัง Figure 1

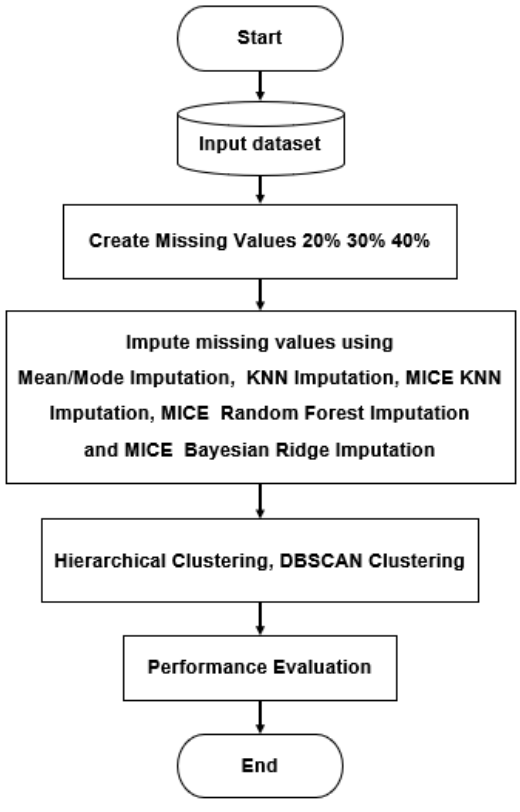


Figure 1 Overview of the experimental processes

1. ข้อมูลที่ใช้ในงานวิจัย

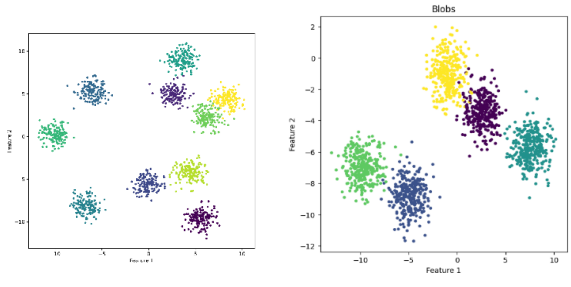
ในการศึกษาการเติมข้อมูลที่สูญหายเพื่อการจัดกลุ่ม ใช้ข้อมูล ดัง Table 1 ประกอบด้วยข้อมูลสังเคราะห์จำนวน 2 ชุด คือ DS1 และ DS2 ที่สร้างขึ้นโดยกำหนดให้ข้อมูลของสมาชิกในแต่ละกลุ่มมีการกระจายตัวตามการแจกแจงแบบปกติ (Normal Distribution) โดยข้อมูลของแต่ละกลุ่มถูกกำหนดให้มีลักษณะแตกต่างกันในเชิงตำแหน่ง (Mean) และการกระจายตัว (Variance) เพื่อให้เกิดความชัดเจนของกลุ่ม หรือในบางกรณีอาจมีการซ้อนทับกันเล็กน้อยระหว่างกลุ่ม ดัง Figure 2 สำหรับชุดข้อมูลจริง (Real Data) เป็นข้อมูลเกี่ยวกับข้อมูลลูกค้าผู้ถือบัญชีธนาคาร มีรายละเอียดข้อมูลดัง Table 2 และเพื่อให้เห็นความสัมพันธ์เชิงภาพของข้อมูลจริงจึงแสดงเป็นองค์ประกอบหลัก (PCA) 2 มิติ ดัง Figure 3

Table 1 Data used in the experiments

Data Set	Data Count	Description
DS1	1,500	2D Simulated Data 5 Clusters
DS2	1,500	2D Simulated Data 10 Clusters
DS3	1,000	Bank Customer Data, partially selected from a public data source (Singal, 2023), consisting of 7 features: Credit Score, Country, Age, Balance, Products Number, Credit Card, and Estimated Salary.

Table 2 Details of the DS3 dataset

Feature	Range of Values
Credit Score	300 - 850
Country	France, Germany, Spain
Age	24 - 83
Balance	2658 - 214347
Products Number	1 - 5
Credit Card	Binary: Yes (1), No (0)
Estimated Salary	600 - 5029354



(a) DS1 (b) DS2

Figure 2 DS1 and DS2 datasets

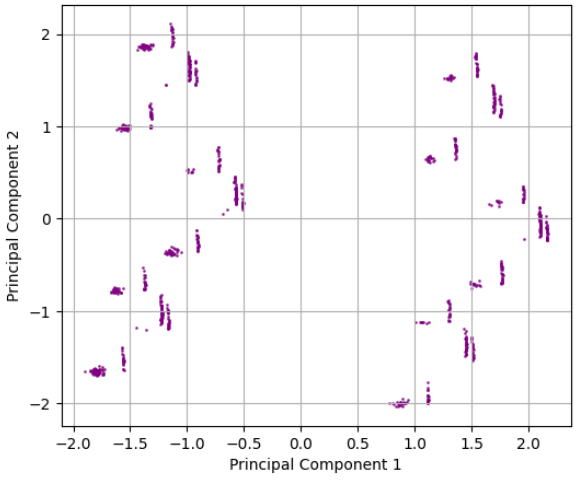


Figure 3 The first two principal components of the DS3 dataset. Denote that categorical data were transformed to binary features by one-hot encoding process

ในแต่ละชุดข้อมูล จะมีการจำลองการสูญหายของข้อมูลแบบ MNAR โดยมี Missing rate 20% 30% และ 40% เพื่อทำการทดลองเปรียบเทียบวิธีการเติมค่าสูญหายแล้วนำไปจัดกลุ่มต่อไป

2. วิธีประมาณการค่าสูญหาย

ในงานนี้ ใช้วิธีประมาณการค่าสูญหาย มี 5 วิธี ดังต่อไปนี้

2.1 Mean/Mode Imputation

การแทนค่าสูญหายด้วยค่าเฉลี่ยสำหรับตัวแปรสำหรับข้อมูลเชิงตัวเลข และแทนค่าสูญหายด้วยค่าฐานนิยมสำหรับตัวแปรสำเชิงหมวดหมู่ (Zhang, 2016)

2.2 KNN Imputation

การแทนค่าสูญหายด้วยวิธี KNN (Murti *et al.*, 2019) จะคำนวณระยะห่างระหว่างข้อมูลตัวที่มีค่าสูญหายที่ต้องการจะเติมกับข้อมูลอื่นทุกตัว หากค่าคุณลักษณะ (features) ที่ไม่มีค่าสูญหายมีความใกล้เคียงกัน จะถูกนำมาใช้ในการแทนค่าสูญหาย โดยพิจารณาข้อมูลที่ใกล้เคียงกันหรือเรียกว่าเป็นเพื่อนบ้านกัน k ตัว นำมาหาค่าเฉลี่ยแล้วแทนค่าให้กับข้อมูลที่สูญหายนั้น โดยค่าสูญหาย $\hat{x}$ คำนวณดังสมการ (3)

$$\hat{x} = \frac{\sum_{i=1}^{k} x_i}{k} \tag{3}$$

เมื่อ แทนค่าคุณลักษณะในคอลัมน์ที่ต้องการเติมค่าสูญขายของเพื่อนบ้านตัวที่

2.3 MICE KNN Imputation

การแทนค่าสูญหายด้วยวิธี MICE (Multivariate Imputation by Chained Equations) (Van Buuren & Groothuis-Oudshoorn, 2011) ใช้การประมาณค่าแบบหลายตัวแปร เติมค่าที่ขาดหายในแต่ละคอลัมน์แบบลูกโซ่ (chain) โดยใช้ K-Nearest Neighbors เพื่อคาดเดาค่าที่เหมาะสมเหมาะสำหรับการเติมค่าสูญหายซึ่งสัมพันธ์กับหลายตัวแปรในชุดข้อมูล

2.4 MICE Random Forest Imputation

การแทนค่าสูญหายด้วยวิธี MICE ใช้การประมาณค่าแบบหลายตัวแปร เติมค่าที่ขาดหายในแต่ละคอลัมน์แบบวนซ้ำ โดยใช้ Random Forest (Shah *et al.*, 2014) เป็นโมเดลสำหรับการคาดเดาค่าที่เหมาะสมสำหรับชุดข้อมูลที่ซับซ้อนซึ่งมีความสัมพันธ์แบบไม่เชิงเส้นระหว่างตัวแปร Random Forest เป็นโมเดลที่ใช้การรวมผล (ensemble) ของต้นไม้ตัดสินใจ (Decision Trees) โดยจะสร้างต้นไม้หลายต้นจากตัวอย่างสุ่มในชุดข้อมูล ในการแบ่งข้อมูลแต่ละครั้ง เพื่อลดการเกิด overfitting และเพิ่มความแม่นยำ สำหรับคอลัมน์ที่มีค่าหาย (missing values) ใช้ Random Forest สร้างโมเดลโดยพิจารณาตัวแปรที่เหลืออยู่ ค่าที่หายจะถูกประมาณจากค่าเฉลี่ยของผลลัพธ์จากต้นไม้หลายต้นใน Random Forest การคาดการณ์ค่าที่ขาดหาย $\hat{y}$ ใช้ค่าเฉลี่ยจากต้นไม้ดังสมการ (4)

$$\hat{y} = \frac{1}{N_t}\sum_{i=1}^{N_t} T_i(x) \tag{4}$$

โดยที่ $T_i(x)$ คือ ผลลัพธ์จากต้นไม้ i และ N_t คือ ต้นไม้ทั้งหมดใน Random Forest

2.5 MICE Bayesian Ridge Imputation

การแทนค่าสูญหายด้วยวิธี ใช้การประมาณค่าแบบหลายตัวแปร เติมค่าที่ขาดหายในแต่ละคอลัมน์แบบวนซ้ำ โดยใช้ Bayesian Ridge Regression (Mostafa *et al.*, 2020) เป็นโมเดลสำหรับคาดเดาค่าที่เหมาะสม Bayesian Ridge Regression เป็นการปรับปรุงโมเดล Linear Regression โดยเพิ่มการแจกแจงความน่าจะเป็นแบบ Bayesian เพื่อประเมินพารามิเตอร์ ของโมเดล สมการพื้นฐานของ Ridge Regression เป็นดังสมการ (5)

$$\hat{y} = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{5}$$

สำหรับคอลัมน์ที่มีค่าหาย ใช้ Bayesian Ridge Regression สร้างโมเดลโดยพิจารณาตัวแปรอื่น ๆ ที่เหลืออยู่ ค่าที่ขาดหาย

จะถูกแทนด้วยค่าเฉลี่ยจากการแจกแจง posterior ของโมเดล Bayesian ค่าประมาณ β คำนวณจาก posterior distribution เพื่อประมาณค่าพารามิเตอร์ที่ดีที่สุด

3. จัดกลุ่มข้อมูล

ในงานนี้ ศึกษาอิทธิพลของวิธีการเติมข้อมูลสูญหายที่มีต่อการจัดกลุ่มข้อมูล มีวิธีจัดกลุ่มข้อมูล 2 วิธี ได้แก่ วิธี Hierarchical Clustering และ วิธี DBSCAN Clustering ใช้ร่วมกับ Gower's distance ซึ่งเป็นการคำนวณระยะห่างระหว่างคู่ของข้อมูลที่ประกอบด้วยคุณลักษณะที่เป็นข้อมูลเชิงตัวเลขและข้อมูลเชิงหมวดหมู่ โดยจะคำนวณระยะห่างจากลักษณะต่าง ๆ ของข้อมูลแต่ละประเภทและนำมารวมกันเป็นค่าระยะห่างสุดท้าย

สำหรับ Hierarchical Clustering จะศึกษาสามวิธี ได้แก่ วิธี Single Linkage วิธี Complete Linkage และวิธี Ward Linkage

สำหรับ DBSCAN Clustering จะศึกษาการกำหนดพารามิเตอร์ MinPts สามค่า ได้แก่ MinPts =3 MinPts =5 และ MinPts =10 จากนั้น กำหนดช่วงค่า Eps เพื่อคำนวณหาค่า Eps ที่เหมาะสมสำหรับ MinPts แต่ละค่า

4. การประเมินประสิทธิภาพ

เนื่องจากในงานนี้ทำการศึกษาวิธีการเติมข้อมูลที่มีการสูญหายที่มีผลต่อการจัดกลุ่ม จึงดำเนินการวัดประสิทธิภาพจากผลการจัดกลุ่มที่ได้ (Cluster Labels) เมื่อเทียบกับการจัดกลุ่มจริง (Ground Truth Labels) โดยใช้ตัวชี้วัดดังต่อไปนี้

4.1 Adjusted Rand Index (ARI)

ARI วัดว่าการจับคู่ระหว่างคู่ของข้อมูลในผลการจัดกลุ่มที่ได้ และ การจัดกลุ่มจริงนั้นตรงกันมากน้อยเพียงใด โดยคำนึงถึงผลลัพธ์ที่อาจเกิดขึ้นแบบสุ่ม มีสูตรการคำนวนดังสมการ (6)

$$ARI = \frac{RI - expecte\ \ RI}{\max(RI) - expected\ RI} \tag{6}$$

เมื่อ $RI = \frac{TP+TN}{TP+FP+FN+TN}$ ซึ่งอยู่ในช่วง [0,1]

TP: จำนวนคู่ข้อมูลที่ถูกจัดให้อยู่ในกลุ่มเดียวกันทั้งใน Cluster Labels และ Ground Truth Labels

TN: จำนวนคู่ข้อมูลที่ถูกจัดให้อยู่ในกลุ่มต่างกันทั้งใน Cluster Labels และ Ground Truth Labels

FP: จำนวนคู่ข้อมูลที่ถูกจัดให้อยู่ในกลุ่มเดียวกันใน Cluster Labels แต่ต่างกลุ่มใน Ground Truth Labels

FN: จำนวนคู่ข้อมูลที่ถูกจัดให้อยู่ในกลุ่มต่างกันใน Cluster Labels แต่กลุ่มเดียวกันใน Ground Truth Labels

4.2 Fowlkes-Mallows Index (FMI)

FMI เป็นตัวชี้วัดที่ใช้ในการประเมินคุณภาพของการจัดกลุ่ม โดยการเปรียบเทียบการจัดกลุ่มที่ได้จากโมเดลกับการจัดกลุ่มจริงโดยเฉพาะความแม่นยำและความสมบูรณ์ของการจัดกลุ่ม มีสูตรการคำนวณ ดังสมการ (7)

$$FMI = \sqrt{\frac{TP}{(TP+FP)} \cdot \frac{TP}{(TP+FN)}} \tag{7}$$

4.3 Normalized Mutual Information (NMI)

NMI เป็นตัวชี้วัดที่ใช้ในการประเมินคุณภาพของการจัดกลุ่ม โดยเปรียบเทียบ Cluster Labels ที่ได้จากการจัดกลุ่มกับ Ground Truth ซึ่งแสดงถึงระดับ ความสัมพันธ์ระหว่างการจัดกลุ่มและคำตอบที่แท้จริง มีสูตรการคำนวณ ดังสมการ (8)

$$NMI = \frac{2 \cdot I(X;Y)}{H(X)+H(Y)} \tag{8}$$

เมื่อ X แทน Cluster Labels และ Y แทน Ground Truth

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

$$H(Y) = -\sum_{y \in Y} p(y) \log p(y)$$

4.4 Purity

Purity เป็นตัวชี้วัดที่ใช้ในการประเมินคุณภาพของการจัดกลุ่มโดยการวัดว่าแต่ละกลุ่มที่ได้จากการจัดมีการจับข้อมูลที่มีลักษณะเหมือนกันมากน้อยแค่ไหนโดยการเปรียบเทียบกลุ่มที่ได้กับ Ground Truth Labels มีสูตรการคำนวณ ดังสมการ (9)

$$Purity = \frac{1}{N} \sum_{K=1}^{K} max\ |C_k \cap L_j| \tag{9}$$

เมื่อ N แทนจำนวนข้อมูลทั้งหมด

K แทนจำนวนกลุ่มที่ได้จากการจัดกลุ่ม

C_k แทนข้อมูลทั้งหมดในกลุ่มที่ k จากการจัดกลุ่ม

L_j แทนข้อมูลทั้งหมดใน Ground Truth Label j

สำหรับข้อมูลสังเคราะห์ DS1 และ DS2 กำหนด Ground Truth Labels จากกลุ่มที่สังเคราะห์ขึ้นโดยตรง แต่สำหรับ DS3 ซึ่งเป็นข้อมูลจริง จะกำหนด Ground Truth Labels จากการจัดกลุ่มข้อมูลที่ไม่มีการสูญหาย โดยใช้วิธี KneeLocator Method [12] ในการหาจำนวนกลุ่มดีที่สุดสำหรับวิธี Hierarchical Clustering ส่วน DBSCAN Clustering จะใช้วิธีเพื่อกำหนดค่า Eps ที่เหมาะสมสำหรับ MinPts แต่ละค่า

ผลการทดลอง

ผลการเปรียบเทียบการเติมข้อมูลสูญหาย 5 วิธี ได้แก่ วิธี Mean/Mode วิธี KNN วิธี MICE KNN วิธี MICE Random Forest และวิธี MICE Bayesian Ridge ที่ส่งผลต่อประสิทธิภาพการจัดกลุ่มบนชุดข้อมูล DS1 DS2 และ DS3 ซึ่งแต่ละชุดข้อมูลมีการจำลองสถานการณ์การสูญหาย 3 ระดับ ได้แก่ สูญหาย 20% 30% และ 40% โดยใช้วิธีการจัดกลุ่ม 2 วิธี คือ Hierarchical Clustering และ DBSCAN Clustering แต่ละวิธีมีการตั้งค่าพารามิเตอร์ต่างกัน 3 แบบ จากนั้นวิเคราะห์ประสิทธิภาพการจัดกลุ่ม ดังต่อไปนี้

การจัดกลุ่มด้วย Hierarchical Clustering

Tables 5-7 แสดงผลการทดลองบนชุดข้อมูล DS1 DS2 และ DS3 ตามลำดับ ซึ่งเป็นการเปรียบเทียบค่าประสิทธิภาพ 4 ค่า คือ ARI NMI FMI และ Purity ซึ่งเป็นตัวชี้วัดประสิทธิภาพการจัดกลุ่มบนชุดข้อมูล DS1 DS2 และ DS3 หลังจากการเติมข้อมูลสูญหาย 5 วิธี โดยวิธีการจัดกลุ่มคือ Hierarchical Clustering (HC) ร่วมกับการใช้วิธีการรวมสองกลุ่มย่อยแบบ Single Linkage แบบ Complete Linkage หรือ แบบ Ward Linkage

เนื่องจากชุดข้อมูล DS1 และ DS2 ได้มาจากการสังเคราะห์ข้อมูลขึ้นมาด้วยวิธีเดียวกัน แต่มีจำนวนกลุ่ม 5 และ 10 กลุ่ม ตามลำดับ ผลการทดลองบนข้อมูล DS1 และ DS2 จึงใกล้เคียงกัน ดัง Tables 5-6 พบว่าประสิทธิภาพการจัดกลุ่มโดยวิธี Hierarchical Clustering (HC) แบบ Ward linkage ให้ค่าประสิทธิภาพสูงในอันดับต้นสามอันดับแรกในทุกตัวชี้วัด ได้แก่ ARI NMI FMI และ Purity เมื่อแทนค่าข้อมูลสูญหายด้วยวิธี KNN และ วิธี MICE KNN กล่าวคือ มีผลค่าในช่วง 0.93 ถึงค่าใกล้ 1 ในทั้ง 3 ระดับการสูญหาย 20% 30% และ 40% ที่ได้ทำการทดลอง ส่วนวิธีที่มีประสิทธิภาพสูงรองลงมาคือ วิธี MICE Random Forest กับการแทนที่ข้อมูลสูญหาย 20% และ 30% ซึ่งยังคงมีค่าประสิทธิภาพสูง 0.97 ขึ้นไป ส่วนการแทนค่าข้อมูลสูญหาย 40% บนชุดข้อมูล DS1 พบว่า วิธี MICE Random Forest ก็ยังแสดงผลประสิทธิภาพสูงในอันดับต้นติดอยู่สามอันดับแรก ในขณะที่การแทนค่าข้อมูลสูญหาย 40% บนชุดข้อมูล DS2 พบว่า ผลจากวิธี MICE Random Forest ในบางตัวชี้วัดไม่ได้ติดอยู่ในสามอันดับแรก แต่ก็ยังแสดง ค่าประสิทธิภาพสูง 0.97 ขึ้นไปในทุกตัวชี้วัด นอกจากนี้ การใช้ Complete linkage ยังให้ผลที่ดีในการทดลองกับชุดข้อมูล DS2 ในบางกรณี แต่อย่างไรก็ตาม การใช้ Ward linkage มักจะให้ผลที่ดีโดยทั่วไปในทั้งสองชุดข้อมูล ต่อมาเมื่อพิจารณารูปร่างคลัสเตอร์ของชุดข้อมูล DS1 และ DS2 ใน Figure 2 จะเห็นว่า ข้อมูลมีลักษณะกลุ่มเป็นรูปร่างกลมวงรีแยกจากกันค่อนข้างชัดเจน ดังนั้น ผลการทดลองที่ได้นี้อาจจะสามารถประยุกต์ใช้เติมข้อมูลสูญหายที่มีระดับมากถึง 40% แทนที่วิธีการที่มักใช้กันอย่างแพร่หลายอย่างวิธี Mean/Mode สำหรับชุดข้อมูล DS3 เป็นข้อมูลจริงเกี่ยวกับลูกค้าธนาคารที่มีจำนวน 7 คุณสมบัติ ซึ่งจะกำหนด Ground truth label ดังกล่าวไว้ในหัวข้อที่ 3 ของการดำเนินงานวิจัย ผลการเปรียบเทียบการเติมข้อมูลสูญหาย 5 วิธี ดัง Table 7 ที่มีข้อมูลสูญหาย 20% 30% และ 40% ปรากฏว่า การแทนค่าข้อมูลสูญหายด้วยวิธี MICE Random Forest ผนวกกับการจัดกลุ่มโดยวิธี Hierarchical Clustering แบบ Ward linkage ให้ค่าประสิทธิภาพสูงในสองอันดับแรกในทุกตัวชี้วัด ได้แก่ ARI NMI FMI และ Purity โดยมีผลอยู่ในช่วง 0.67 ถึง 0.88 ส่วนวิธีที่มีประสิทธิภาพรองลงมาสำหรับกรณีการแทนค่าข้อมูลสูญหาย 20% และ 30% คือ วิธี KNN และวิธี MICE KNN เมื่อผนวกกับการจัดกลุ่มโดยวิธี HC แบบ Ward linkage แต่ในกรณีการแทนที่ข้อมูลสูญหาย 40% การเติมข้อมูลสูญหายด้วยวิธี MICE Random Forest และวิธี MICE KNN ผนวกกับ HC แบบ Complete linkage ก็ให้ผลประสิทธิภาพสูงในอันดับต้นสามอันดับแรกในทุกตัวชี้วัด

กล่าวโดยภาพรวม จากผลการทดลองกับทั้ง 3 ชุดข้อมูลข้างต้น และข้อมูลมีอัตราการสูญหายไม่เกิน 40% การแทนค่าข้อมูล สูญหายด้วย MICE Random Forest กับการจัดกลุ่มด้วยวิธี Hierarchical Clustering (HC) แบบ Ward linkage มักจะให้ค่า ประสิทธิภาพที่สูงในทุกตัวชี้วัด

Table 5 Performance of Hierarchical Clustering on DS1 with three levels missing data imputed by five methods

HC	Imputation Methods	DS1 with 20% missing				DS1 with 30% missing				DS1 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
Single Linkage	Mean/Mode	0.2317	0.4237	0.5260	0.4020	0.0000	0.0079	0.4448	0.2040	0.2089	0.4719	0.5399	0.4013
	KNN	0.3738	0.5870	0.6180	0.4027	0.3738	0.5870	0.6180	0.4027	0.2095	0.4727	0.5405	0.4013
	MICE KNN	0.3738	0.5870	0.6180	0.4027	0.3734	0.5864	0.6175	0.4033	0.2089	0.4718	0.5396	0.4020
	MICE Random Forest	0.3734	0.3734	0.6175	0.4033	0.3738	0.5870	0.6180	0.4027	0.2089	0.4718	0.5396	0.4020
	MICE Bayesian Ridge	0.0000	0.0079	0.4448	0.2040	0.0000	0.0079	0.4448	0.2040	0.4814	0.7396	0.6723	0.6013
Complete Linkage	Mean/Mode	0.4570	0.6167	0.5691	0.6087	0.4720	0.6005	0.6277	0.5567	0.3714	0.6134	0.5521	0.5973
	KNN	0.7411	0.8124	0.7946	0.8760	0.5833	0.7337	0.6982	0.6313	0.3801	0.6723	0.5801	0.6000
	MICE KNN	0.6538	0.7535	0.7285	0.8200	0.5843	0.7353	0.7004	0.626	0.3571	0.6458	0.5659	0.5907
	MICE Random Forest	0.6212	0.7358	0.7067	0.7840	0.6336	0.7597	0.7181	0.7373	0.3680	0.6691	0.5700	0.5993
	MICE Bayesian Ridge	0.3768	0.5401	0.5452	0.5467	0.1936	0.3798	0.4686	0.4473	**0.7178**$^{(2)}$	0.8497	**0.7901**$^{(2)}$	0.7933
Ward Linkage	Mean/Mode	0.6582	0.7544	0.7299	0.7820	0.6938	0.7786	0.7629	0.7873	0.5627	0.7618	0.6792	0.7380
	KNN	**0.9802**$^{(2)}$	**0.9729**$^{(3)}$	**0.9841**$^{(3)}$	**0.9920**$^{(2)}$	**0.9867**$^{(2)}$	**0.9817**$^{(2)}$	**0.9894**$^{(2)}$	**0.9947**$^{(2)}$	**0.9301**$^{(1)}$	**0.9396**$^{(1)}$	**0.9441**$^{(1)}$	**0.9700**$^{(1)}$
	MICE KNN	**0.9802**$^{(2)}$	**0.9740**$^{(2)}$	**0.9842**$^{(2)}$	**0.9920**$^{(2)}$	**0.9901**$^{(1)}$	**0.9867**$^{(1)}$	**0.9921**$^{(1)}$	**0.9960**$^{(1)}$	**0.9301**$^{(1)}$	**0.9396**$^{(1)}$	**0.9441**$^{(1)}$	**0.9700**$^{(1)}$
	MICE Random Forest	**0.9851**$^{(1)}$	**0.9797**$^{(1)}$	**0.9881**$^{(1)}$	**0.9940**$^{(1)}$	**0.9852**$^{(3)}$	**0.9807**$^{(3)}$	**0.9881**$^{(3)}$	**0.9940**$^{(3)}$	0.7144	**0.8530**$^{(2)}$	0.7866	**0.7960**$^{(2)}$
	MICE Bayesian Ridge	0.6010	0.7130	0.6862	0.7387	0.4354	0.5633	0.5775	0.5580	0.5754	0.7623	0.6873	0.7433

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

Table 6 Performance of Hierarchical Clustering on DS2 with three levels missing data imputed by five methods

HC	Imputation Methods	DS2 with 20% missing				DS2 with 30% missing				DS2 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
Single Linkage	Mean/Mode	0.389	0.6474	0.4958	0.5407	0.2273	0.5565	0.3944	0.45	0.1204	0.456	0.3554	0.378
	KNN	0.7333	0.8965	0.787	0.7007	0.7338	0.8977	0.7874	0.7013	0.734	0.8983	0.7877	0.7007
	MICE KNN	0.7326	0.8947	0.7863	0.7007	0.7317	0.8949	0.7856	0.7856	0.8063	0.9293	0.8403	0.8
	MICE Random Forest	0.7326	0.8958	0.7863	0.7007	0.7331	0.896	0.7866	0.7013	0.797	0.9203	0.8321	0.7967
	MICE Bayesian Ridge	0.4643	0.7665	0.5748	0.6107	0.3816	0.6886	0.4919	0.576	0.117	0.4482	0.3523	0.3587
Complete Linkage	Mean/Mode	0.4969	0.6891	0.5525	0.6513	0.3067	0.5550	0.4046	0.5173	0.2286	0.5162	0.3862	0.4213
	KNN	0.9661	0.9754	0.9695	0.9840	**0.9956**$^{(1)}$	**0.9948**$^{(1)}$	**0.9960**$^{(1)}$	**0.9980**$^{(1)}$	0.9882	0.9881	0.9894	0.9947
	MICE KNN	**0.9754**$^{(2)}$	**0.9790**$^{(2)}$	**0.9778**$^{(2)}$	**0.9887**$^{(2)}$	0.9475	0.9691	0.9528	0.9733	**0.9926**$^{(2)}$	**0.9917**$^{(2)}$	**0.9933**$^{(2)}$	**0.9967**$^{(2)}$
	MICE Random Forest	0.8863	0.9469	0.8991	0.9227	0.9770	0.9791	0.9792	0.9893	0.9839	0.9841	0.9855	0.9927
	MICE Bayesian Ridge	0.5427	0.7296	0.5926	0.6853	0.4697	0.6664	0.5296	0.5873	0.3483	0.5907	0.4622	0.4973
Ward Linkage	Mean/Mode	0.5129	0.6990	0.5668	0.7000	0.4034	0.6285	0.4780	0.6093	0.4845	0.6758	0.5507	0.6387
	KNN	**0.9867**$^{(1)}$	**0.9861**$^{(1)}$	**0.9880**$^{(1)}$	**0.9940**$^{(1)}$	**0.9882**$^{(2)}$	**0.9881**$^{(2)}$	**0.9894**$^{(2)}$	**0.9947**$^{(2)}$	**0.9941**$^{(1)}$	**0.9934**$^{(1)}$	**0.9947**$^{(1)}$	**0.9973**$^{(1)}$
	MICE KNN	**0.9753**$^{(3)}$	**0.9789**$^{(3)}$	**0.9778**$^{(2)}$	**0.9887**$^{(2)}$	**0.9882**$^{(2)}$	**0.9877**$^{(3)}$	**0.9894**$^{(2)}$	**0.9947**$^{(2)}$	**0.9911**$^{(3)}$	**0.9904**$^{(3)}$	**0.9920**$^{(3)}$	**0.9960**$^{(3)}$
	MICE Random Forest	0.9724	0.9771	0.9752	0.9873	0.9755	0.9773	0.9779	0.9887	0.9779	0.9802	0.9801	0.9900
	MICE Bayesian Ridge	0.5528	0.7593	0.6033	0.7080	0.5182	0.7189	0.5706	0.6553	0.6114	0.7544	0.6522	0.7047

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

Table 7 Performance of Hierarchical Clustering on DS3 with three levels missing data imputed by five methods

HC	Imputation Methods	DS3 with 20% missing				DS3 with 30% missing				DS3 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
Single Linkage	Mean/Mode	0.4165	0.5840	0.5261	0.7000	0.3902	0.5708	0.5065	0.7000	0.3999	0.5796	0.5137	0.7000
	KNN	0.4087	0.5675	0.5175	0.7000	0.4032	0.5632	0.5136	0.7000	0.4055	0.5669	0.5153	0.7000
	MICE KNN	0.4087	0.5675	0.5175	0.7000	0.4032	0.5632	0.5136	0.7000	0.4055	0.5669	0.5153	0.7000
	MICE Random Forest	0.4087	0.5675	0.5175	0.7000	0.4032	0.5632	0.5136	0.7000	0.4055	0.5669	0.5153	0.7000
	MICE Bayesian Ridge	0.4087	0.5675	0.5175	0.7000	0.4032	0.5632	0.5136	0.7000	0.4055	0.5669	0.5153	0.7000
Complete Linkage	Mean/Mode	0.2450	0.3590	0.5149	0.5820	0.1340	0.2255	0.4510	0.5190	0.2224	0.3160	0.4837	0.5440
	KNN	0.5039	0.5701	0.6499	0.7790	0.5095	0.5797	0.6518	0.7860	0.3989	0.4663	0.5768	0.7290
	MICE KNN	0.5139	0.5916	0.6553	0.7880	0.4579	0.4749	0.6237	0.7240	**0.4860**$^{(3)}$	**0.5420**$^{(3)}$	**0.6379**$^{(3)}$	**0.7690**$^{(3)}$
	MICE Random Forest	0.5322	0.6175	0.6689	0.7950	0.4129	0.5125	0.5905	0.7220	**0.7630**$^{(1)}$	**0.8644**$^{(1)}$	**0.8012**$^{(1)}$	**0.8700**$^{(1)}$
	MICE Bayesian Ridge	0.1158	0.1956	0.3954	0.4960	0.1388	0.2420	0.4315	0.5100	0.6419	0.7654	0.6979	0.7800
Ward Linkage	Mean/Mode	0.1670	0.2313	0.4823	0.5870	0.1869	0.2359	0.4860	0.6060	0.2194	0.2545	0.4986	0.6300
	KNN	**0.6798**$^{(1)}$	**0.6898**$^{(2)}$	**0.7917**$^{(1)}$	**0.8740**$^{(1)}$	**0.6492**$^{(3)}$	**0.6484**$^{(2)}$	**0.7720**$^{(3)}$	**0.8630**$^{(2)}$	0.4373	0.5288	0.6313	0.7420
	MICE KNN	**0.6717**$^{(3)}$	**0.6658**$^{(3)}$	**0.7869**$^{(3)}$	**0.8730**$^{(2)}$	**0.6497**$^{(2)}$	**0.6480**$^{(3)}$	**0.7723**$^{(2)}$	**0.8630**$^{(2)}$	0.4345	0.5248	0.6295	0.7410
	MICE Random Forest	**0.6792**$^{(2)}$	**0.6909**$^{(1)}$	**0.7913**$^{(2)}$	**0.8740**$^{(1)}$	**0.6794**$^{(1)}$	**0.6819**$^{(1)}$	**0.7913**$^{(1)}$	**0.8750**$^{(1)}$	**0.6685**$^{(2)}$	**0.6707**$^{(2)}$	**0.7844**$^{(2)}$	**0.8700**$^{(1)}$
	MICE Bayesian Ridge	0.1322	0.1893	0.4526	0.5810	0.0441	0.0470	0.3890	0.5220	0.0426	0.0473	0.3894	0.5200

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

การจัดกลุ่มด้วย DBSCAN Clustering

Tables 8-10 แสดงผลการทดลองบนชุดข้อมูล DS1 DS2 และ DS3 ตามลำดับ ซึ่งเป็นการเปรียบเทียบค่าประสิทธิภาพ 4 ค่า คือ ARI NMI FMI และ Purity ซึ่งเป็นตัวชี้วัดประสิทธิภาพการจัดกลุ่มบนชุดข้อมูล DS1 DS2 และ DS3 หลังจากการเติมข้อมูลสูญหาย 5 วิธี ดังกล่าวข้างต้น โดยวิธีการจัดกลุ่ม คือ วิธี DBSCAN Clustering ซึ่งได้ทดลองตั้งค่า MinPts 3 ค่า คือ ค่า MinPts เท่ากับ 3 ค่า MinPts เท่ากับ 5 และ ค่า MinPts เท่ากับ 10 แต่ละค่าหาค่า Eps ที่เหมาะสมด้วย Elbow Method

จากผลการทดลองข้างต้นของชุดข้อมูล DS1 DS2 และ DS3 เมื่อพิจารณาค่าประสิทธิภาพที่ดีสามอันดับแรกในแต่ละตัวชี้วัดที่ระดับข้อมูลสูญหายต่างกัน จะพบว่า แต่ละชุดข้อมูลจะมีค่า MinPts ได้เหมาะสมมากกว่า 1 ค่า

สำหรับชุดข้อมูล DS1 (Table 8) วิธีการเติมข้อมูลสูญหายที่ให้ประสิทธิภาพที่ดีติดสามอันดับต้นในตัวชี้วัดส่วนใหญ่ในการทดลองทั้ง 3 ระดับข้อมูลสูญหาย คือ วิธี MICE KNN และรองลงมา คือ วิธี MICE Random Forest เมื่อ DBSCAN Clustering กำหนด MinPts เท่ากับ 3 หรือ 5

ส่วนผลการจัดกลุ่มในชุดข้อมูล DS2 (Table 9) วิธีการเติมข้อมูลสูญหายที่ให้ประสิทธิภาพที่ดีติดสามอันดับต้น ได้แก่ วิธี KNN วิธี MICE KNN และ วิธี MICE Random Forest พบได้กระจายในการทดลองกำหนดค่า MinPts ทั้ง 3 ค่า ต่อมาหากเลือกวิธีการเติมข้อมูลเพียงวิธีเดียวที่ให้ประสิทธิภาพที่ดีในทั้ง 3 ระดับข้อมูลสูญหาย เมื่อกำหนดค่า MinPts เป็น 3 ค่านี้ ปรากฏว่า วิธี MICE Random Forest จะให้ผลค่าประสิทธิภาพที่ดีในทุกตัวชี้วัด ในช่วงค่า 0.80-0.98

ในการทดลองจัดกลุ่มด้วย DBSCAN Clustering ในทั้ง 3 ระดับข้อมูลสูญหายบนชุดข้อมูลจริง DS3 (Table 10) ปรากฏว่า ที่การกำหนดค่า MinPts เท่ากับ 5 วิธีการเติมข้อมูลสูญหายที่ให้ประสิทธิภาพที่ดีติดสามอันดับต้นในทุกตัวชี้วัด ได้แก่ วิธี KNN วิธี MICE KNN และ วิธี MICE Random Forest

ประสิทธิภาพการเติมข้อมูลสูญหายในชุดข้อมูล DS1 DS2 และ DS3

จะเห็นได้ว่า วิธีการจัดกลุ่มพร้อมทั้งพารามิเตอร์ และวิธีการเติมค่าสูญหายมีผลต่อประสิทธิภาพการจัดกลุ่ม Figures 4-6 แสดงการเปรียบเทียบวิธีการเติมค่าสูญหายบนชุดข้อมูล DS1 DS2 และ DS3 ตามลำดับ ทั้ง 3 ระดับข้อมูลสูญหาย โดยเลือกผลการจัดกลุ่มที่ดีสุดมาเปรียบเทียบกัน จากผลการเปรียบเทียบ พบว่า Hierarchical Clustering ให้ผลค่าตัวชี้วัดประสิทธิภาพที่ดีกว่าวิธี DBSCAN Clustering ในชุดข้อมูล DS1 และ DS 2 (Figure 4-5) โดยส่วนใหญ่จะเป็นผลการเติมค่าสูญหายด้วยวิธี KNN และวิธี MICE KNN รองลงมาคือวิธี MICE Random Forest แต่สำหรับชุดข้อมูล DS3

ซึ่งเป็นข้อมูลจริง (Figure 6) พบว่า DBSCAN Clustering ให้ผลค่าตัวชี้วัดประสิทธิภาพที่ดีกว่าวิธี Hierarchical Clustering โดยจะเป็นผลการเติมค่าสูญหายด้วยวิธี KNN วิธี MICE KNN และวิธี MICE Random Forest ที่ให้ค่าประสิทธิภาพดีใกล้เคียงกัน

จากผลการทดลองข้างต้นโดยรวมพบว่า การเติมข้อมูลสูญหายด้วยวิธีการที่มีพื้นฐานมาจากแมชีนเลิร์นนิ่ง จะได้ประสิทธิภาพที่ดีกว่าวิธีพื้นฐานอย่างการเติมด้วยค่ากลาง นอกจากนี้จะเห็นได้ว่า วิธีการเติมข้อมูลสูญหายวิธีหนึ่ง หากใช้เติมข้อมูลสูญหายที่ชุดข้อมูลแตกต่างกัน แม้จะใช้วิธีการจัดกลุ่มแบบเดียวกัน จะได้ประสิทธิภาพการจัดกลุ่มที่แตกต่างกัน เมื่อดูที่ค่าตัววัดเดียวกัน หากมองในอีกมุม ชุดข้อมูลเดียวกันที่มีอัตราการสูญหายเท่ากัน หากใช้วิธีการเติมข้อมูลสูญหายแตกต่างกัน แม้จะใช้วิธีการจัดกลุ่มและพารามิเตอร์เหมือนกัน ประสิทธิภาพในการจัดกลุ่มมีความแตกต่างกันในค่าตัววัดเดียวกันเช่นกัน

ดังนั้น ปัจจัยที่เกี่ยวข้องกับประสิทธิภาพการจัดกลุ่ม ได้แก่ ชุดข้อมูล วิธีการเติมข้อมูลสูญหาย อัตราการสูญหายของข้อมูล อัลกอริทึมการจัดกลุ่มและการกำหนดพารามิเตอร์ ซึ่งการจัดการข้อมูลสูญหายอย่างมีประสิทธิภาพสำหรับการจัดกลุ่ม หากเข้าใจการกระจายของตัวข้อมูลรวมถึงอัตราการสูญหาย จะทำให้มีผลดีเกี่ยวข้องกับการเลือกใช้วิธีการและพารามิเตอร์ที่เหมาะสมได้

Table 8 Performance of DBSCAN Clustering on DS1 with three levels missing data imputed by five methods

DB SCAN	Imputation Methods	DS1 with 20% missing				DS1 with 30% missing				DS1 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
MinPts = 3	Mean/Mode	0.2197	0.4024	0.5092	0.4100	0.2073	0.4047	0.5145	0.4067	0.4605	0.6759	0.6353	0.6233
	KNN	0.3651	0.5544	0.6012	0.4087	0.7450	0.8374	0.8110	0.7907	0.4770	0.7265	0.6670	0.5993
	MICE KNN	**0.5919**$^{(3)}$	0.7343	**0.7190**$^{(3)}$	**0.6040**$^{(1)}$	**0.9451**$^{(1)}$	**0.9226**$^{(1)}$	**0.9561**$^{(1)}$	**0.9780**$^{(1)}$	**0.7386**$^{(2)}$	**0.8315**$^{(2)}$	**0.8015**$^{(2)}$	**0.8073**$^{(1)}$
	MICE Random Forest	0.5915	**0.7353**$^{(3)}$	0.7188	**0.6040**$^{(1)}$	**0.8510**$^{(2)}$	**0.8584**$^{(2)}$	**0.8815**$^{(2)}$	**0.9740**$^{(2)}$	0.7248	0.8037	**0.7936**$^{(3)}$	0.7880
	MICE Bayesian Ridge	0.0001	0.0195	0.4392	0.2113	0.0001	0.0161	0.4396	0.2107	0.4605	0.7045	0.6492	0.6080
MinPts = 5	Mean/Mode	0.2574	0.5103	0.5332	0.4880	0.2271	0.4825	0.5227	0.4580	0.7158	0.7960	0.7814	0.7893
	KNN	0.3687	0.5615	0.6076	0.4007	0.3694	0.5639	0.6099	0.3993	0.4785	0.7258	0.6681	0.5980
	MICE KNN	**0.5986**$^{(1)}$	**0.7513**$^{(1)}$	**0.7271**$^{(1)}$	0.5960	**0.7503**$^{(3)}$	**0.8442**$^{(3)}$	**0.8154**$^{(3)}$	**0.7840**$^{(3)}$	**0.7268**$^{(3)}$	**0.8201**$^{(3)}$	0.7912	**0.7940**$^{(2)}$
	MICE Random Forest	**0.5973**$^{(2)}$	**0.7462**$^{(2)}$	**0.7252**$^{(2)}$	**0.5967**$^{(3)}$	0.6019	0.7279	0.7044	0.7100	0.7208	0.7982	0.7883	0.7787
	MICE Bayesian Ridge	0.0001	0.0134	0.4393	0.2087	0.0000	0.0027	0.4437	0.2027	0.4604	0.7025	0.6492	0.6073
MinPts = 10	Mean/Mode	0.0000	0.0013	0.4463	0.2007	0.2271	0.4825	0.5227	0.4580	0.2087	0.4636	0.5332	0.4080
	KNN	0.3703	0.5670	0.6113	0.3993	0.3713	0.5711	0.6130	0.4007	**0.7645**$^{(1)}$	**0.8682**$^{(1)}$	**0.8279**$^{(1)}$	**0.7920**$^{(3)}$
	MICE KNN	0.3712	0.5695	0.6123	0.4007	0.3707	0.5724	0.6127	0.4007	0.4642	0.6916	0.6520	0.6007
	MICE Random Forest	0.3715	0.5736	0.6137	0.4013	0.3690	0.5561	0.6052	0.4100	0.2133	0.4769	0.5310	0.4267
	MICE Bayesian Ridge	0.0000	0.0036	0.4393	0.2053	0.0047	0.0682	0.4219	0.2473	0.2070	0.4340	0.5278	0.4093

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

Table 9 Performance of DBSCAN Clustering on DS2 with three levels missing data imputed by five methods

DB SCAN	Imputation Methods	DS2 with 20% missing				DS2 with 30% missing				DS2 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
MinPts = 3	Mean/Mode	0.5022	0.7268	0.5676	0.7120	0.3934	0.6793	0.5019	0.6327	0.6947	0.5921	0.4334	0.4307
	KNN	0.8031	**0.9221**$^{(3)}$	0.8367	0.7947	0.8004	0.9170	0.8335	0.7933	0.8889	0.9544	0.9040	0.8940
	MICE KNN	0.8036	**0.9246**$^{(2)}$	0.8374	0.7960	0.8796	**0.9417**$^{(2)}$	0.8951	0.8873	0.9417	**0.9568**$^{(2)}$	0.9484	**0.9893**$^{(1)}$
	MICE Random Forest	0.8002	0.9175	0.8314	0.8107	0.8667	0.9313	0.8831	0.8873	**0.9470**	0.9513	**0.9525**$^{(3)}$	**0.9820**$^{(2)}$
	MICE Bayesian Ridge	0.5216	0.7285	0.5827	0.7227	0.4063	0.7048	0.5104	0.6727	0.3390	0.6663	0.4872	0.5653
MinPts = 5	Mean/Mode	0.2903	0.6234	0.4552	0.5147	0.5294	0.7201	0.5869	0.6947	0.3390	0.6679	0.6679	0.5713
	KNN	0.5178	0.8182	0.6421	0.6000	**0.9001**$^{(2)}$	0.9110	**0.9108**$^{(2)}$	**0.9460**$^{(3)}$	**0.9644**$^{(1)}$	**0.9603**$^{(1)}$	**0.9681**$^{(1)}$	0.9767
	MICE KNN	**0.8174**$^{(2)}$	0.8847	0.8370	**0.8540**$^{(3)}$	**0.8878**$^{(3)}$	0.9102	**0.9005**$^{(3)}$	**0.9547**$^{(2)}$	0.8597	0.9013	0.8771	0.9620
	MICE Random Forest	**0.8107**$^{(3)}$	0.9204	**0.8774**$^{(1)}$	**0.8967**$^{(1)}$	0.8594	0.9238	0.8763	0.8813	0.8805	0.9017	0.8943	0.9587
	MICE Bayesian Ridge	0.2182	0.6194	0.4471	0.4567	0.5444	0.7407	0.5992	0.7253	0.5506	0.7627	0.6062	0.7380
MinPts = 10	Mean/Mode	0.5313	0.7393	0.5968	0.7167	0.2768	0.6133	0.4303	0.5453	0.4682	0.6938	0.5483	0.6307
	KNN	**0.8581**$^{(1)}$	0.9181	**0.8746**$^{(2)}$	**0.8733**$^{(2)}$	0.8609	0.9229	0.8775	0.8753	**0.9580**$^{(2)}$	**0.9550**$^{(3)}$	**0.9624**$^{(2)}$	0.9733
	MICE KNN	0.5182	0.8194	0.6426	0.6000	**0.9636**$^{(1)}$	**0.9584**$^{(1)}$	**0.9674**$^{(1)}$	**0.9780**$^{(1)}$	0.9287	0.9444	0.9370	**0.9807**$^{(3)}$
	MICE Random Forest	0.8050	**0.9287**$^{(1)}$	**0.8392**$^{(3)}$	0.7973	0.8726	**0.9375**$^{(3)}$	0.8890	0.8873	0.9459	0.9479	0.9515	0.9767
	MICE Bayesian Ridge	0.5160	0.7197	0.5776	0.7187	0.2847	0.6323	0.4383	0.5573	0.5702	0.7716	0.6264	0.7407

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

Table 10 Performance of DBSCAN Clustering on DS3 with three levels missing data imputed by five methods

DB SCAN	Imputation Methods	DS3 with 20% missing				DS3 with 30% missing				DS3 with 40% missing			
		ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity	ARI	NMI	FMI	Purity
MinPts = 3	Mean/Mode	0.6746	0.8036	0.7373	0.7490	0.1395	0.5264	0.2659	0.2090	0.1517	0.5509	0.2723	0.2370
	KNN	0.8578	0.8638	0.8723	0.8670	0.2143	0.6037	0.3088	0.2850	0.2252	0.6119	0.3168	0.2990
	MICE KNN	0.8343	0.8465	0.8501	0.8730	0.2174	0.6071	0.3101	0.2880	0.2224	0.6091	0.3144	0.2980
	MICE Random Forest	0.8489	0.8588	0.8632	0.8800	0.2223	0.6147	0.3142	0.2930	0.2247	0.6129	0.3165	0.2950
	MICE Bayesian Ridge	0.8724	0.8585	0.8846	0.8760	0.1854	0.5781	0.2868	0.2600	0.1926	0.5817	0.2932	0.2690
MinPts = 5	Mean/Mode	0.6955	0.8120	0.7522	0.7600	0.2185	0.6056	0.3076	0.2960	0.5758	0.7273	0.6630	0.6610
	KNN	**0.8820**$^{(3)}$	**0.8890**$^{(3)}$	**0.8940**$^{(3)}$	**0.8900**$^{(3)}$	**0.7962**$^{(2)}$	**0.8276**$^{(2)}$	**0.8190**$^{(2)}$	**0.8110**$^{(2)}$	**0.8124**$^{(2)}$	**0.8372**$^{(2)}$	**0.8328**$^{(2)}$	**0.8240**$^{(2)}$
	MICE KNN	0.8707	0.8772	0.8840	0.8830	**0.7867**$^{(3)}$	**0.8200**$^{(3)}$	**0.8111**$^{(3)}$	**0.8030**$^{(3)}$	**0.8098**$^{(3)}$	**0.8357**$^{(3)}$	**0.8306**$^{(3)}$	**0.8230**$^{(3)}$
	MICE Random Forest	**0.8843**$^{(1)}$	**0.8915**$^{(2)}$	**0.8960**$^{(1)}$	**0.8920**$^{(1)}$	**0.8051**$^{(1)}$	**0.8360**$^{(1)}$	**0.8263**$^{(1)}$	**0.8160**$^{(1)}$	**0.8165**$^{(1)}$	**0.8410**$^{(1)}$	**0.8363**$^{(1)}$	**0.8260**$^{(1)}$
	MICE Bayesian Ridge	0.8636	0.8493	0.8771	0.8720	0.6002	0.7300	0.6719	0.6960	0.5872	0.7203	0.6647	0.6690
MinPts = 10	Mean/Mode	0.6802	0.8141	0.7428	0.7490	0.5516	0.7056	0.6381	0.6600	0.0027	0.0268	0.3038	0.1350
	KNN	0.8790	0.8876	0.8914	0.8870	0.7534	0.7951	0.7804	0.7980	0.6597	0.7542	0.7088	0.7160
	MICE KNN	0.8717	0.8821	0.8848	0.8850	0.7539	0.7938	0.7806	0.7980	0.5418	0.6972	0.6331	0.6470
	MICE Random Forest	**0.8825**$^{(2)}$	**0.8920**$^{(1)}$	**0.8944**$^{(2)}$	**0.8910**$^{(2)}$	0.5352	0.6681	0.6205	0.6440	0.7731	0.8071	0.7971	0.8150
	MICE Bayesian Ridge	0.7121	0.8173	0.7573	0.7910	0.5913	0.7113	0.6577	0.7010	0.5595	0.6840	0.6374	0.6580

Note: Bold values indicating the top three values in each metric (or each column) with a superscript number in each bracket denoting the ranking.

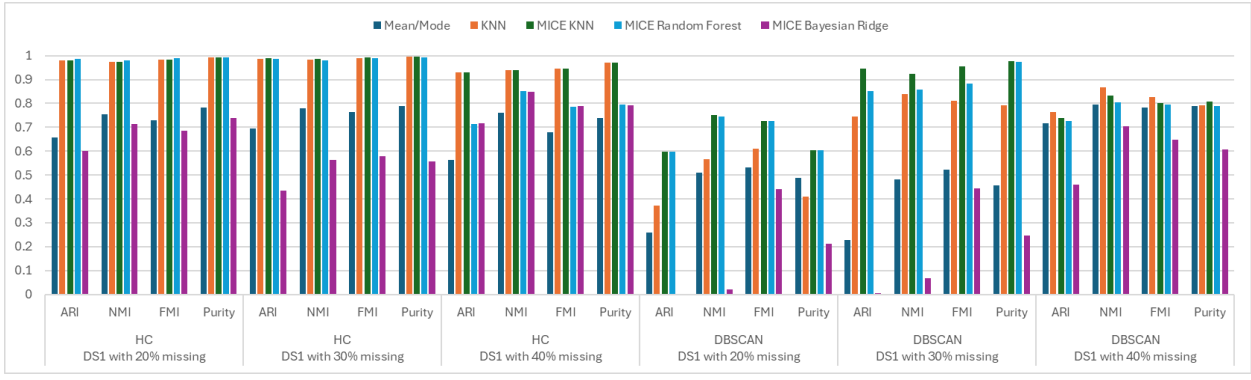


Figure 4 Comparison of DS1 best performance yielded by five imputation methods

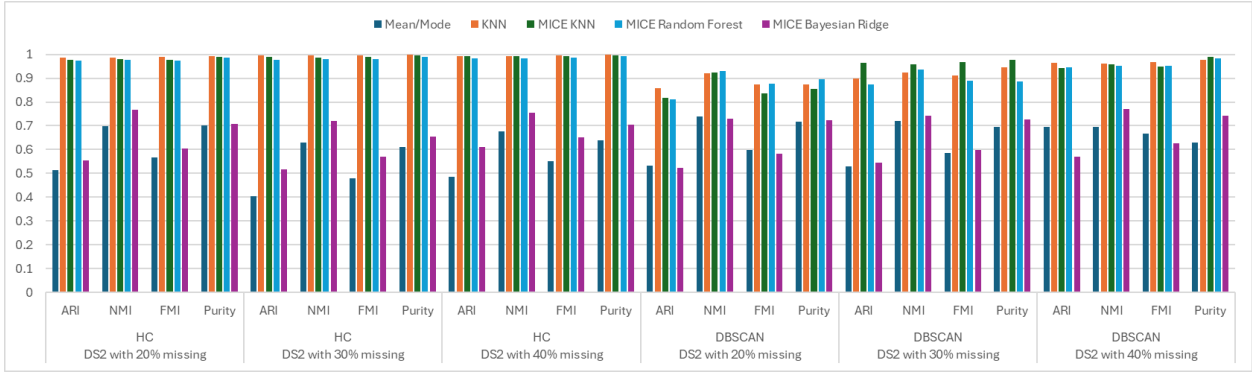


Figure 5 Comparison of DS2 best performance yielded by five imputation methods

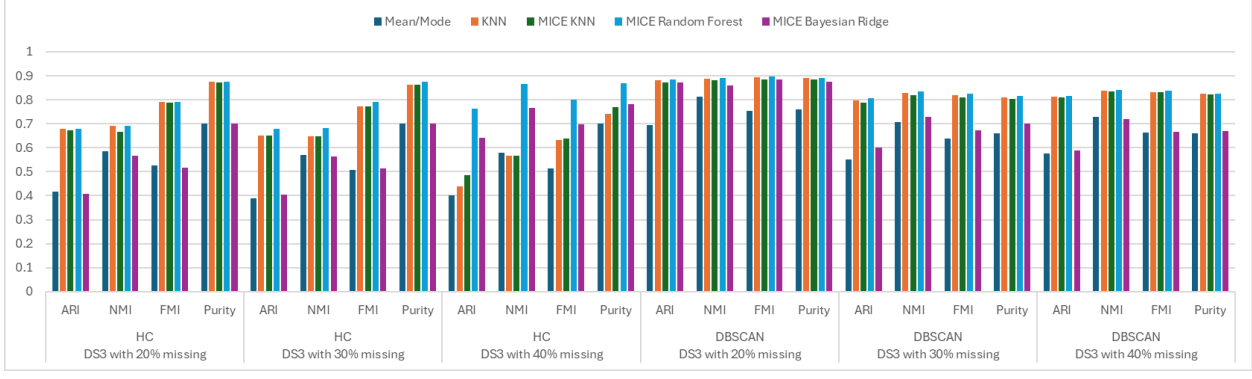


Figure 6 Comparison of DS3 best performance yielded by five imputation methods

สรุปผลการทดลองและข้อเสนอแนะ

ในงานวิจัยนี้ศึกษาอิทธิพลของวิธีการประมาณค่าข้อมูลสูญหายที่มีต่อประสิทธิภาพการจัดกลุ่ม โดยใช้ข้อมูลจำลองและข้อมูลจริงที่เกี่ยวกับลูกค้าธนาคาร โดยใช้วิธีจัดกลุ่ม Hierarchical Clustering และ DBSCAN Clustering จากการทดลองพบว่า เมื่อกำหนดค่าพารา มิเตอร์ที่เหมาะสม วิธี KNN วิธี MICE KNN และวิธี MICE Random Forest ถือเป็นวิธีที่มีประสิทธิภาพในการจัดการข้อมูลที่สูญหายได้ดี ที่ให้ผลค่า ARI NMI FMI และ Purity สูงกว่าวิธีอื่น ๆ ในทุกชุดข้อมูลที่ทำการทดลอง แม้ว่าจะมีการสูญหายของข้อมูลมากถึง 40% ก็ตาม อย่างไรก็ดี หากทำการทดสอบวิธีการจัดกลุ่มในรูปแบบอื่น เช่น K-Means หรือ Gaussian mixture models (GMM) เป็นต้น อาจช่วยให้ได้ผลลัพธ์ที่หลากหลายและเหมาะสมกับลักษณะของข้อมูลแบบอื่น นอกจากนี้ ประสิทธิภาพการใช้พื้นที่และเวลาในการทำนายเพื่อเติมค่าข้อมูลสูญหายเป็นปัจจัยที่สำคัญสำหรับแนวทางในการวิจัยในอนาคตที่จะเลือกวิธีการเติมค่าสูญหาที่เหมาะสมมากขึ้น

เอกสารอ้างอิง

กัลยา วานิชบัญชา. (2558). *การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows* (พิมพ์ครั้งที่ 10). ศูนย์หนังสือจุฬาลงกรณ์มหาวิทยาลัย.

พัชนา สุวรรณแสน. (2562). การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์. *วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา, 11*(2), 45–57.

ศรีรักษ์ ศรีทองชัย. (2566). การเปรียบเทียบวิธีการประมาณค่าสูญหายในการพยากรณ์ความเข้มข้นของ PM2.5 ด้วยโครงข่ายประสาทเทียม LSTM. *วารสารนวัตกรรมการเรียนรู้และเทคโนโลยี, 3*(1), 59–67.

อุมาพร ยกกำพล, อัชฌาณัท รัตนเลิศนุสรณ์ และอุไรวรรณ เจริญกีรติกุล. (2561). การเปรียบเทียบประสิทธิภาพของการจัดกลุ่มข้อมูลวิธีการแบบลำดับขั้นและวิธีการเคมีนสำหรับข้อมูลผสมเชิงหมวดหมู่กับเชิงตัวเลข. ใน *การประชุมวิชาการสถิติประยุกต์และเทคโนโลยีสารสนเทศระดับชาติ ประจำปี 2561: Digital Transformation and Thailand 4.0* (น. 1–6). คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

Batista, G. E. A. P. A., & Monard, M. C. (2003). *Experimental comparison of k-Nearest Neighbour and Mean or Mode Imputation Methods with the internal strategies used by C4.5 and CN2 to treat missing data*. University of São Paulo at São Carlos.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). AAAI Press.

Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, & Pratama, E. A. (2018). Customer segmentation based on RFM model and clustering techniques with K-Means algorithm. *2018 Third International Conference on Informatics and Computing (ICIC)*, 1–6.

Mostafa, S. M., Eladimy, A. S., Hamad, S., & Amano, H. (2020). CBRG: A novel algorithm for handling missing data using Bayesian ridge regression and feature selection based on gain ratio. *IEEE Access, 8*, 216969–216985.

Murti, D. M. P., Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019). K-Nearest Neighbor (k-NN) based missing data imputation. *2019 5th International Conference on Science in Information Technology (ICSITech)*, 83–88.

Pan, R., Yang, T., & Cao, J. (2015). Missing data imputation by k nearest neighbours based on grey relational structure and mutual information. *Applied Intelligence, 43*(3), 614–632.

Sanjar, K., Bekhzod, O., Kim, J., & Paul, A. (2020). Missing data imputation for geolocation-based price prediction using KNN–MCF method. *ISPRS International Journal of Geo-Information, 9*(6), Article 347.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology, 179*(6), 764–774.

Singal, A. (2023). *Bank term deposit predictions* [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/bank-term-deposit-prediction

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1*(4), 368–376.

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine, 4*(1), Article 19.