การเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนในตัวแปรตามสำหรับตัวแบบการจำแนก

Comparison of the efficiency of noise handling methods in dependent variable for classification model

กฤษฎี ศิริเรือง[1] และ ประภาศิริ รัชชประภาพรกุล[2*]
Kritsadee Siriruang[1] and Prapasiri Ratchaprapapornkul[2*]

Received: 23 April 2024 ; **Revised**: 28 May 2024 ; **Accepted**: 12 June 2024

บทคัดย่อ

ข้อมูลรบกวนเป็นปัญหาหลักที่พบในชุดข้อมูล ซึ่งหากเกิดกับตัวแปรตามจะนำไปสู่การเรียนรู้กลุ่มที่ผิดพลาด จึงต้องจัดการกับข้อมูลรบกวนก่อนนำไปวิเคราะห์จำแนกกลุ่มข้อมูล โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบการจำแนกระหว่างวิธีการขจัดข้อมูลรบกวนด้วยตัวกรองข้อมูล 4 วิธี ได้แก่ วิธี Condensed Nearest Neighbor (CNN), วิธี Edited Nearest Neighbors (ENN), วิธี Cross-Validated Committees Filter (CVCF) และ วิธี Iterative Partitioning Filter (IPF) และวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองข้อมูลรบกวนจากวิธีการขจัดข้อมูลรบกวนร่วมกับการประมาณค่าทดแทนพหุ 3 วิธี ได้แก่ วิธี polytomous regression (polyreg), วิธีป่าสุ่ม (random forest: rf) และ วิธี multiple imputation through XGBoost (mixgb) โดยศึกษาผ่านการจำลองข้อมูลแบบมอนติคาร์โล ภายใต้สถานการณ์ของขนาดตัวอย่างเท่ากับ 100, 500 และ 1,000 หน่วย ปริมาณข้อมูลรบกวนเท่ากับ 10%, 20%, 30% และ 40% โดยประสิทธิภาพของวิธีการจัดการข้อมูลรบกวน พิจารณาจากค่า F1 ของตัวแบบการจำแนก 4 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้ที่สุด, วิธีป่าสุ่ม, วิธีนาอีฟเบย์ และวิธีซัพพอร์ตเวกเตอร์แมชชีน และเปรียบเทียบประสิทธิภาพด้วยการวิเคราะห์ความแปรปรวนหลายทาง ผลการศึกษาพบว่าวิธีการจัดการข้อมูลรบกวนมีอิทธิพลปฏิสัมพันธ์กับทุกปัจจัย ได้แก่ ขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนกที่ส่งผลให้ค่าประสิทธิภาพ F1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยหากตัวอย่างขนาดเล็ก (n = 100) การปรับค่าตัวแปรตามจะมีแนวโน้มดีกว่าการขจัดข้อมูลรบกวน แต่เมื่อขนาดตัวอย่างขนาดใหญ่ขึ้นทั้ง 2 วิธีจะมีประสิทธิภาพใกล้เคียงกัน เมื่อปริมาณข้อมูลรบกวนเพิ่มขึ้นค่าประสิทธิภาพ F1 มีแนวโน้มลดลง และประสิทธิภาพของวิธีการจัดการด้วยการปรับค่าตัวแปรตามมีแนวโน้มดีกว่าในทุกตัวแบบ ในภาพรวมประสิทธิภาพของวิธีการปรับค่าตัวแปรตามด้วยตัวกรอง ENN ร่วมกับการประมาณค่าทดแทนพหุด้วยวิธี polyreg มีแนวโน้มให้ค่าประสิทธิภาพ F1 สูงสุดเกือบทุกกรณี ยกเว้นกรณีที่ขนาดตัวอย่างเป็น 1,000 หน่วย ที่วิธีการ ENN จะมีประสิทธิภาพสูงสุด ผลการวิจัยนี้ทำให้ได้ข้อค้นพบถึงวิธีการจัดการข้อมูลรบกวนที่เหมาะสมกับแต่ละสถานการณ์การจำแนกกลุ่มข้อมูล

คำสำคัญ: ข้อมูลรบกวน, การจำแนก, การประมาณค่าทดแทนพหุ, การเรียนรู้ของเครื่อง, ตัวกรองข้อมูลรบกวน

---

[1] นิสิตระดับมหาบัณฑิต สาขาวิชาวิธีวิทยาการพัฒนานวัตกรรมทางการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
[2] อาจารย์ประจำสาขาวิชาวิธีวิทยาการพัฒนานวัตกรรมทางการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
[1] Master degree student, Division of Methodology for Innovation Development in Education, Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University
[2] Lecturer, Division of Methodology for Innovation Development in Education, Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University
* Corresponding author E-mail: prapasiri.r@chula.ac.th

Abstract

Noisy data is a major problem often encountered in datasets. When noise affects the dependent variable, it can lead to incorrect group in classification. Therefore, it is essential to handle noise before analyzing and classifying the data. This research aims to compare the effectiveness of class noise handling method in classification model between noise removal methods using four noise filter: Condensed Nearest Neighbor (CNN), Edited Nearest Neighbors (ENN), Cross-Validated Committees Filter (CVCF), and Iterative Partitioning Filter (IPF), and relabel methods that utilize noise filter from noise removal methods along with multiple imputation from three methods: polytomous regression (polyreg), random forest (rf), and multiple imputation through XGBoost (mixgb). The study is conducted through Monte Carlo simulation under the scenario with sample sizes of 100, 500, and 1,000 units, and noise level of 10%, 20%, 30%, and 40%. The performance of noise handling methods is evaluated based on the F1 score of four data classification models: k-NN, Random Forest, Naïve Bayes, and Support Vector Machine. Comparison is done through N-Way analysis of variance (N-Way ANOVA). The study found that the class noise handling methods have an interaction effect with all factors, including sample size, noise level, and classification model, affecting the F1 score significantly at the .05 level of significance. For small sample sizes (n = 100), relabel method tended to perform better than remove method. However, as sample size increased, both methods showed similar performance. Overall, the combination of ENN noise filter with polytomous regression imputation tended to yield the highest F1 score in most cases, except for sample sizes of 1,000 units where ENN alone showed the highest performance. The findings of this research provide insights into appropriate class noise handling methods for different data classification scenarios.

Keywords: Noisy data, classification, multiple imputation, machine learning, noise filter

บทนำ

คุณภาพของข้อมูลมีความสำคัญอย่างยิ่งต่อการวิเคราะห์ข้อมูล โดยเฉพาะการเรียนรู้แบบมีผู้สอน (supervised learning) ในการเรียนรู้ของเครื่อง (machine learning) ที่ใช้สำหรับการจำแนกกลุ่มข้อมูล (classification) โดยชุดข้อมูลเรียนรู้ (training data) จะต้องมีคุณภาพเพียงพอ เพื่อให้ผลการทำนายในชุดข้อมูลใหม่มีความถูกต้องแม่นยำ (Moura *et al*., 2022) ข้อมูลรบกวน (noisy data) เป็นปัญหาหลักที่พบในชุดข้อมูล (Nguyen-Van, 2020; Zhu & Wu, 2004) โดยเป็นความผิดพลาดที่เกิดขึ้นอย่างสุ่ม ซึ่งสามารถเกิดได้ทั้งในกระบวนการจำแนก (classification) ที่มีตัวแปรตามเป็นตัวแปรจัดประเภท หรือกระบวนการถดถอย (regression) ที่มีตัวแปรตามเป็นตัวแปรต่อเนื่อง ซึ่งปัญหาข้อมูลรบกวนที่เกิดขึ้นนี้จะส่งผลต่อประสิทธิภาพของการวิเคราะห์ข้อมูล เพิ่มเวลา และความซับซ้อนในการวิเคราะห์

ข้อมูลรบกวนสามารถเกิดได้กับตัวแปรทั้ง 2 ประเภทในตัวแบบ ทั้งตัวแปรอิสระหรือตัวแปรคุณลักษณะ (attribute variable) และตัวแปรตาม (class variable) โดยข้อมูลรบกวนที่เกิดกับตัวแปรอิสระ เป็นข้อมูลที่มีค่าผิดพลาด สูญหาย ไม่สมบูรณ์ หรือข้อมูลไม่ตรงตามช่วงหรือลักษณะข้อมูลที่ควรจะเป็น ซึ่งจะเรียกข้อมูลรบกวนนี้ว่า “attribute noise” ส่วนข้อมูลรบกวนที่เกิดกับตัวแปรตามจะเรียกว่า class noise หรือ label noise เป็นการระบุกลุ่มของข้อมูลผิดพลาดที่อาจจะเกิดจากตัวบุคคลที่ทำหน้าที่ระบุกลุ่มผิดพลาด หรือข้อมูลที่นำมาใช้ตัดสินกลุ่มไม่มีคุณภาพเพียงพอ หรือเครื่องมือที่ใช้ในการเก็บข้อมูลบกพร่อง และมีปัญหาอื่นๆ เกิดขึ้นระหว่างการดำเนินการระบุกลุ่มข้อมูล (Sáez *et al*., 2013; García *et al*., 2019) โดยข้อมูลรบกวนที่เกิดกับตัวแปรตามนี้จะนำไปสู่การเรียนรู้กลุ่มที่ผิดพลาดเมื่อนำไปวิเคราะห์จำแนกกลุ่มข้อมูล

ปัญหาข้อมูลรบกวนสามารถเกิดขึ้นได้ในหลายบริบท เช่น ในทางการแพทย์ พบว่า สามารถเกิดข้อมูลรบกวนจากการระบุกลุ่มโรคผิดพลาด โดยข้อผิดพลาดนี้มาจากตัวบุคคลผู้ระบุกลุ่มเองที่อาจจะทำงานภายใต้ความซับซ้อน กดดัน และเหนื่อยล้า (Sáez *et al*., 2016) ในทางวิทยาศาสตร์อาจเกิดข้อมูลรบกวนจากเครื่องมือที่ใช้วัดเกิดความบกพร่อง ทำให้ได้ข้อมูลที่คลาดเคลื่อนจากความเป็นจริง โดยเมื่อพิจารณาถึงในบริบททางการศึกษา พบว่า ในปัจจุบันได้มีการใช้สื่อเทคโนโลยี ปัญญาประดิษฐ์ (Artificial Intelligence: AI) ในการจัดการเรียนการสอน รวมทั้งมีการนำหลักการเรียนรู้ของเครื่อง และการวิเคราะห์เหมืองข้อมูลทางการศึกษามาวิเคราะห์การเรียนรู้ และพัฒนาตัวแบบเชิงทำนายเพื่อพัฒนาและปรับปรุงคุณภาพของผู้เรียน ข้อมูลที่เข้ามาจะมีขนาดใหญ่และมาจากหลายๆ แหล่ง ทำให้พบปัญหาข้อมูลรบกวนที่เป็นปัญหาสำคัญที่ต้องจัดการก่อนนำไปวิเคราะห์ (Hassan *et al*., 2020) ตัวอย่างของปัญหาข้อมูลรบกวนในทางการศึกษา

เช่น การจัดการเรียนการสอนรูปแบบใหม่ในลักษณะของห้องเรียนอัจฉริยะ (smart classroom) ที่มีการใช้ทั้งระบบออนไลน์ หรือออฟไลน์ และมีการใช้ระบบปัญญาประดิษฐ์ร่วมกับการเรียนรู้ของเครื่องในการจัดกลุ่มผู้เรียนจากพฤติกรรมและอารมณ์ความรู้สึก การที่ต้องพิจารณาตัดสินกลุ่มจากคุณลักษณะหลายๆ อย่างที่มีการนำเข้าข้อมูลจากเครื่องมือหรืออุปกรณ์หลายแหล่งพร้อมกัน อาจทำให้เกิดความคลาดเคลื่อนในการทำนายกลุ่มผู้เรียน ซึ่งทำให้เกิดข้อมูลรบกวนในตัวแปรตาม นอกจากนี้ข้อมูลรบกวนอาจเกิดได้ในกรณีที่ตัวบุคคลผู้ให้คะแนนเองทั้งเพื่อนร่วมชั้นและครูเกิดความคลาดเคลื่อน หรืออคติในการให้คะแนนเพื่อระบุกลุ่มของผู้เรียนในการวัดและประเมินผลการเรียนรู้ เมื่อต้องประเมินผลจากการทำกิจกรรมที่มีการลงมือปฏิบัติจริง พฤติกรรมในการเรียนรู้ และการนำเสนอที่มีหลายตัวชี้วัดพร้อมกัน (Kim *et al*., 2018; Miller & Soh, 2013) จากตัวอย่างดังกล่าวจะเห็นได้ว่าการเกิดข้อมูลรบกวนที่ตัวแปรตามสามารถเกิดได้หลายสถานการณ์ตามบริบทจริง ซึ่งในสถานการณ์จริงการระบุว่ากลุ่มที่ระบุเป็นข้อมูลรบกวนหรือเป็นกลุ่มที่ถูกต้อง สามารถกระทำได้หลายวิธี เช่น การสำรวจข้อมูลโดยใช้สถิติบรรยาย เพื่อดูลักษณะข้อมูล และความสัมพันธ์ของตัวแปรต่างๆ ก่อนนำข้อมูลไปวิเคราะห์ หรือใช้การแบ่งกลุ่มข้อมูล (clustering) ในการจัดกลุ่มเพื่อดูความแตกต่างของกลุ่มที่จัดขึ้นกับกลุ่มที่ระบุ นอกจากนี้ยังสามารถใช้การจำแนก (classification) ด้วยตัวแบบต่างๆ ในการระบุกลุ่มโดยหากกลุ่มที่ได้จากการทำนายไม่ตรงกับกลุ่มที่เก็บข้อมูลมา หรือไม่ตรงกับกลุ่มของข้อมูลอื่นที่มีคุณลักษณะใกล้เคียงกันก็จะสามารถบอกได้ว่าอาจจะเป็นข้อมูลรบกวน (Frénay & Verleysen, 2013)

นอกจากนี้การเกิดข้อมูลรบกวนในตัวแปรตามสามารถเกิดได้อย่างมีรูปแบบ โดย Schafer and Graham (2002) ได้ศึกษาไว้จำนวน 3 รูปแบบที่แตกต่างกัน ได้แก่ 1) NCAR (Noise Completely at Random) เป็นข้อมูลรบกวนที่เกิดขึ้นอย่างอิสระโดยไม่ขึ้นกับตัวแปรอิสระและตัวแปรตาม 2) NAR (Noise at Random) เป็นข้อมูลรบกวนที่เกิดขึ้นโดยขึ้นอยู่กับค่าของตัวแปรตาม และ 3) NNAR (Noise not at Random) เป็นข้อมูลรบกวนที่เกิดขึ้นโดยขึ้นอยู่กับค่าของตัวแปรอิสระและตัวแปรตาม ซึ่งโดยธรรมชาติของข้อมูลจริง รูปแบบของข้อมูลรบกวนที่เกิดขึ้นโดยทั่วไปมักจะเป็นแบบ NNAR เนื่องจากการเกิดข้อมูลรบกวนในตัวแปรตาม อาจเกิดมาจากการระบุกลุ่มผิดพลาดอันเนื่องมาจากข้อมูลจากตัวแปรอิสระที่เกิดความคลาดเคลื่อนจากการบันทึกข้อมูลโดยผู้วิจัยเอง หรือผู้ให้ข้อมูลมีการให้ข้อมูลไม่ครบถ้วนหรือปกปิดข้อมูลบางส่วนไว้ โดยปัจจุบันยังมีงานวิจัยจำนวนน้อยที่ศึกษาการจัดการกับข้อมูลรบกวนรูปแบบนี้ งานวิจัยนี้จึงมุ่งศึกษาเฉพาะปัญหาข้อมูลรบกวนที่เกิดกับตัวแปรตามที่มีลักษณะการเกิดแบบ NNAR

จากการศึกษางานวิจัยที่เกี่ยวข้องในเรื่องการจัดการกับปัญหาข้อมูลรบกวนที่ตัวแปรตาม พบว่าสามารถกระทำได้ 2 กระบวนการ ได้แก่ 1) กระบวนการระดับอัลกอริทึม เป็นการใช้ตัวแบบที่มีความทนต่อข้อมูลรบกวน โดยไม่ต้องมีการจัดการกับข้อมูลรบกวนก่อนนำไปวิเคราะห์ (Miao *et al*., 2015) และ 2) กระบวนการระดับข้อมูล เป็นการจัดการกับข้อมูลรบกวนโดยการขจัดข้อมูลรบกวน (remove) หรือปรับค่าของข้อมูลใหม่ (relabel) นิยมใช้ในอัลกอริทึมที่ไม่ทนต่อข้อมูลรบกวน หรือต้องการเพิ่มประสิทธิภาพของอัลกอริทึม (Brodley & Friedl, 1999) สำหรับวิธีการขจัดข้อมูลรบกวนจะใช้ ตัวกรองข้อมูลรบกวน (noise filter) ซึ่งแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่ม similarity ที่มีวิธีการพิจารณาข้อมูลรบกวนโดยใช้เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors: k-NN) และเทคนิคอื่นๆ ที่พิจารณาระยะทางระหว่างข้อมูล ตัวอย่างได้แก่ All-k Edited Nearest Neighbors (AENN), Blame Based Noise Reduction (BBNR), Condensed Nearest Neighbor (CNN) และ Edited Nearest Neighbors (ENN) และกลุ่ม ensemble ที่มีวิธีการพิจารณาข้อมูลรบกวนโดยใช้ตัวแบบในการจำแนกหลายๆ ครั้ง เพื่อตรวจสอบข้อมูลรบกวน ตัวอย่างได้แก่ Cross-Validated Committees Filter (CVCF), Ensemble Filter (EF), High Agreement Random Forest (HARF) และ Iterative Partitioning Filter (IPF) ซึ่งการใช้ตัวกรองข้อมูลรบกวนจะดำเนินการกับข้อมูลหลังตรวจพบด้วยการขจัดทั้งแถวข้อมูลออกจากชุดข้อมูล ซึ่งจะเกิดเป็นข้อมูลสูญหาย ที่ทำให้สูญเสียข้อมูลบางส่วนในการนำไปเรียนรู้

ส่วนการจัดการข้อมูลรบกวนในรูปแบบของการปรับค่าข้อมูลตัวแปรตาม โดยไม่ขจัดแถวข้อมูลออกจากชุดข้อมูล มีงานวิจัยเสนอวิธีการไว้อย่างหลากหลาย เช่น การปรับค่าด้วยการเรียนรู้ค่าที่ถูกต้อง (Self- training correction) โดยใช้ตัวกรองข้อมูลรบกวนออกแล้วให้ตัวแบบเรียนรู้จากค่าที่ไม่ได้เป็นข้อมูลรบกวน และสร้างค่าใหม่ที่ถูกต้องขึ้นมาทดแทน (Nicholson *et al*., 2016) หรือใช้การจัดกลุ่มข้อมูลด้วย k-means เพื่อตรวจสอบข้อมูลรบกวน แล้วใช้ตัวแบบทำนายเพื่อปรับค่าใหม่ให้ถูกต้องตามกลุ่ม (Nematzadeh *et al*., 2020)

จากการศึกษาข้างต้น หากใช้ตัวกรองข้อมูลรบกวนจะทำให้เกิดข้อมูลสูญหาย ผู้วิจัยจึงศึกษาวิธีแก้ปัญหาข้อมูลสูญหาย พบว่า วิธีประมาณค่าการทดแทน (imputation) ที่มีการคำนวณค่ามาทดแทนข้อมูลที่สูญหาย ถูกนำมาใช้อย่างหลากหลายทั้งในทางสถิติและทางการเรียนรู้ด้วยเครื่อง (Sun *et al*., 2023) โดยสามารถแบ่งเป็น 1) การประมาณค่าการทดแทนด้วยค่าเดียว (single imputation) และ 2) การประมาณค่าทดแทนพหุ (multiple imputation) ซึ่งการประมาณค่าทดแทนพหุจะมีข้อดีกว่าการประมาณค่าด้วยค่าเดียว เนื่องจากจะคำนวณค่าทดแทนขึ้นมาหลายชุด แล้วใช้สถิติเพื่อประเมิน

หาค่าทดแทนที่ดีที่สุด ที่จะลดความลำเอียงจากการประมาณค่า (Van Buuren & Groothuis-Oudshoorn, 2011; Li *et al*., 2015) ทั้งนี้วิธีการประมาณค่าทดแทนพหุที่นิยมใช้ในปัจจุบัน คือ Multivariate Imputation by Chained Equations (MICE) (Sun *et al*., 2023) ซึ่งภายในวิธี MICE จะประมาณค่าโดยสามารถเลือกรูปแบบตัวแบบที่เหมาะสมกับข้อมูลได้ โดยตัวแบบที่เป็นค่าเริ่มต้นสำหรับประมาณค่าข้อมูลตัวแปรตามที่เป็นเชิงกลุ่มตั้งแต่ 2 กลุ่มขึ้นไป คือ Imputation of unordered data by polytomous regression (polyreg) หากพิจารณาถึงตัวแบบอื่นๆ ในงานวิจัยของ Buczak *et al*. (2023) ระบุว่าการใช้ตัวแบบป่าสุ่มในการประมาณค่าในอัลกอริทึมของ MICE มีประสิทธิภาพดีกว่าการใช้การประมาณค่าด้วยป่าสุ่ม (random forest: rf) ผ่านอัลกอริทึมอื่นในหลายสถานการณ์ของการเกิดข้อมูลสูญหาย นอกจากนี้ ในการประมาณค่าทดแทนที่กล่าวมา ถึงแม้จะมีประสิทธิภาพสูง แต่จะใช้เวลามากในประมวลผล จึงมีผู้พัฒนาวิธีการ Multiple Imputation Through XGBoost (mixgb) ที่มีการใช้อัลกอริทึม XGBoost ในการประมาณค่า มีการสุ่มตัวอย่างย่อย และใช้วิธีทำนายผ่านการสร้าง Boosting tree ที่สามารถทำงานได้เร็วกับข้อมูลขนาดใหญ่และมีความซับซ้อน (Deng & Lumley, 2023) จากที่กล่าวมา พบว่า วิธีการประมาณค่าทดแทนพหุสามารถนำไปทดแทนข้อมูลสูญหายได้อย่างมีประสิทธิภาพ ในงานวิจัยนี้จึงนำการประมาณค่าทดแทนดังกล่าวมาใช้ร่วมกับการใช้ตัวกรองข้อมูลรบกวน สำหรับใช้เป็นวิธีการปรับค่าตัวแปรตามที่เป็นข้อมูลรบกวนให้ถูกต้องก่อนการนำไปวิเคราะห์

จากประเด็นที่กล่าวมา งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบการจำแนกระหว่างวิธีการขจัดข้อมูลรบกวนด้วยตัวกรองข้อมูล 4 วิธี ได้แก่ วิธี CNN, วิธี ENN, วิธี CVCF และ วิธี IPF และวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองข้อมูลรบกวนจากวิธีการขจัดข้อมูลรบกวนร่วมกับการประมาณค่าทดแทนพหุ 3 วิธี ได้แก่ วิธี polyreg, วิธี rf และ วิธี mixgb เมื่อสถานการณ์ของขนาดตัวอย่างเท่ากับ 100, 500 และ 1,000 หน่วย ปริมาณข้อมูลรบกวนเท่ากับ 10%, 20%, 30% และ 40% และตัวแบบการจำแนกกลุ่มข้อมูลเป็น วิธีเพื่อนบ้านใกล้ที่สุด (k-Nearest Nearest Neighbors: k-NN), วิธีป่าสุ่ม (Random Forest: rf), วิธีนาอีฟเบย์ (Naïve Bayes: NB) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) ผ่านการจำลองข้อมูลแบบมอนติคาร์โล เพื่อให้ครอบคลุมกับสถานการณ์ที่แตกต่างกัน และให้ได้ข้อค้นพบถึงวิธีการจัดการข้อมูลรบกวนที่เหมาะสมกับแต่ละสถานการณ์

วิธีการวิจัย

งานวิจัยนี้เป็นการศึกษาผ่านการจำลองข้อมูลด้วยวิธีการมอนติคาร์โลผ่านโปรแกรม R เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบการจำแนกโดยวนซ้ำ 500 รอบ ซึ่งเป็นจำนวนรอบที่ให้ผลลัพธ์ที่มีความแม่นยำเพียงพอและใช้เวลาไม่มากเกินไปในการประมวลผล (Zhang & Cui, 2023) โดยมีขั้นตอนการวิจัย ดังนี้

1. การจำลองข้อมูลสมบูรณ์ กำหนดเงื่อนไขขนาดตัวอย่างขนาดตัวอย่างเป็น 3 ระดับ ได้แก่ 100, 500 และ 1,000 หน่วย กำหนดแปรอิสระจำนวน 10 ตัวแปร แบ่งเป็นตัวแปรจัดประเภท 5 ตัวแปร ที่มีการแจกแจงแบร์นูลลี โดยกำหนดพารามิเตอร์เป็น $x_i \sim Ber(0.5)$ โดยที่ i = 1, 2, … , 5 และตัวแปรต่อเนื่องมีการแจกแจงแบบปกติ จำนวน 5 ตัวแปร โดยกำหนดพารามิเตอร์เป็น $x_j \sim N(0,1)$ โดยที่ j = 6, 7, … , 10 การกำหนดให้มีตัวแปรแต่ละประเภทมีการแจกแจงเดียวกันเพื่อให้สอดคล้องกับการแปลงข้อมูลให้อยู่ในหน่วยเดียวกัน ในทางปฏิบัติ กำหนดให้ตัวแปรอิสระในแต่ละคู่ตัวแปรมีความสัมพันธ์กันขนาดต่ำ (r = 0.3) ผู้วิจัยจำลองตัวแปรตามเป็นตัวแปรจัดประเภทที่แบ่งออกเป็น 3 กลุ่ม โดยมีเงื่อนไขให้อัตราออด (odd ratio) เป็น 3 ระดับ โดยแต่ละระดับจะสุ่มจากช่วงดังต่อไปนี้ ระดับต่ำ [0,1), ระดับกลาง [1,2) และระดับสูง [2,3) และนำมาใช้ในการกำหนดค่าสัมประสิทธิ์การถดถอย (β) แต่ละตัวของตัวแบบ และจำลองตัวแปรตามผ่านตัวแบบการถดถอยลอจิสติกอเนกนาม (Multinomial Logistic Regression) ดังสมการ

$$\log(\pi_i/\pi_j) = \beta_{j0} + \beta_{j1}x_1 + \ldots + \beta_{jp}x_p$$

โดยที่ J แทน จำนวนกลุ่มของตัวแปรตามซึ่งแบ่งเป็น 3 กลุ่ม, p แทน จำนวนตัวแปรอิสระ (p = 1, 2, … , 10) และ j = 1, … , J -1

2. การจำลองข้อมูลรบกวน งานวิจัยนี้ศึกษาข้อมูลรบกวนที่เกิดอย่างไม่เป็นอิสระโดยขึ้นอยู่กับทั้งตัวแปรอิสระและตัวแปรตามเอง โดยมีกลไกการเกิดข้อมูลรบกวนแบบ NNAR (Noise Not At Random) โดยจะเกิดเมื่อข้อมูลมีความใกล้เคียงกับชุดข้อมูลในอีกกลุ่ม จึงอยู่บริเวณขอบของกลุ่ม หากจะจำลองข้อมูลรบกวนรูปแบบนี้ สามารถจำลองได้โดยการแทนที่ค่าของตัวแปรตามที่อยู่บริเวณขอบกลุ่มข้อมูลที่เชื่อมต่อกับชุดข้อมูลอีกกลุ่มหรือบริเวณที่ข้อมูลในกลุ่มนั้นมีความหนา

แน่นน้อยเป็นค่าของอีกกลุ่ม (Moura *et al*., 2022; Garcia *et al*., 2019) ซึ่ง Sáez (2022) ได้ระบุว่าข้อมูลรบกวนแบบ NNAR มีลักษณะเป็น Neighborwise borderline label noise ผู้วิจัยจึงจำลองข้อมูลรบกวนในงานวิจัยนี้ให้เป็นลักษณะ Neighborwise borderline label noise โดยทำการจำลองผ่านแพ็คเกจ “noisemodel” ของโปรแกรม R ซึ่งจะกำหนดค่าพารามิเตอร์ level แสดงเงื่อนไขปริมาณข้อมูลรบกวนเป็น 10%, 20%, 30% และ 40% ตามลำดับ โดยจำลองข้อมูลรบกวนดังกล่าวในชุดข้อมูลเรียนรู้ (training data)

3. การจัดการข้อมูลรบกวน วิธีการจัดการข้อมูลรบกวนที่ศึกษาในงานวิจัยนี้แบ่งออกเป็น 2 วิธี ได้แก่

3.1 วิธีการขจัดข้อมูลรบกวน (remove) เป็นการใช้ตัวกรองข้อมูลรบกวน (noise filter) เพื่อขจัดแถวข้อมูลที่ถูกตรวจจับว่าเป็นข้อมูลรบกวนผ่านอัลกอริทึมที่แตกต่างกันของแต่ละตัวกรองออกจากชุดข้อมูลเรียนรู้ (training data) ผลที่ได้จากการขจัดทั้งแถวที่มีตัวแปรตามเป็นข้อมูลรบกวนจะทำให้ลดจำนวนชุดข้อมูลในเรียนรู้ที่จะนำไปเป็นตัวนำเข้าในกระบวนการจำแนกกลุ่มข้อมูล (Prati *et al*., 2019) โดยตัวกรองข้อมูลรบกวนสามารถแบ่งได้เป็น 2 กลุ่ม ได้แก่ กลุ่ม similarity ที่ใช้หลักการ k-NN และระยะห่างระหว่างข้อมูลในการคัดกรอง และกลุ่ม ensemble ที่ใช้หลักการทำนายจากตัวแบบจำแนกหลายๆ ครั้ง ในการศึกษาครั้งนี้ผู้วิจัยเลือกตัวกรองข้อมูลรบกวนมาศึกษากลุ่มละ 2 ตัวกรอง รวมทั้งหมด 4 ตัวกรอง ได้แก่ ในกลุ่ม similarity เลือก 2 วิธี ได้แก่ 1) CNN มีหลักการ คือ ใช้ 1-NN เพื่อหาเพื่อนบ้านที่ใกล้ที่สุด หากกลุ่มไม่ตรงกันจะขจัดออก และ 2) ENN มีหลักการ คือ หาเพื่อนบ้านที่ใกล้ที่สุดด้วย k-NN และจะขจัดออกถ้าข้อมูลตัวนั้นมีกลุ่มแตกต่างจากเพื่อนบ้าน และในกลุ่ม ensemble เลือกวิธี 2 วิธี ได้แก่ 1) CVCF มีหลักการคือ แบ่งข้อมูลออกเป็น n folds แล้วสร้าง ต้นไม้ประเภท C4.5 ด้วย n-1 folds เพื่อทดสอบข้อมูลที่เป็นข้อมูลรบกวน และ 2) IPF มีหลักการคือ ใช้ต้นไม้ประเภท C4.5 เพื่อกำจัดชุดข้อมูลที่มีข้อมูลรบกวน โดยจะทำวนซ้ำจนกระทั่งไม่มีข้อมูลถูกกำจัดอีกต่อไป การใช้ตัวกรองข้อมูลทั้งหมดจะใช้แพ็คเกจ“NoiseFiltersR” ในโปรแกรม R

3.2 วิธีการปรับค่าตัวแปรตาม (relabel) เป็นการปรับค่าตัวแปรตามที่ถูกตรวจสอบว่าเป็นข้อมูลรบกวน ให้เป็นกลุ่มที่ถูกต้อง โดยในการศึกษาครั้งนี้มีการใช้ตัวกรองข้อมูลรบกวนจากวิธีการขจัดข้อมูลรบกวนมาเป็นวิธีในการตรวจสอบข้อมูลรบกวนหลังจากนั้นจะใช้เก็บวิธีค่าดัชนีที่ระบุการเป็นข้อมูลรบกวนไว้ แล้วลบค่าเฉพาะตัวแปรตามที่แสดงว่าเป็นข้อมูลรบกวน โดยไม่ได้ลบทั้งแถวข้อมูลเหมือนวิธีขจัดข้อมูลรบกวน หลังจากนั้นจะใช้การประมาณค่าทดแทนพหุ (multiple imputation) ในการทำนายค่าตัวแปรตามโดยครั้งนี้เลือกวิธีการประมาณค่าทดแทนมาใช้ร่วมกับตัวกรองข้อมูลจำนวน 3 วิธี ได้แก่ 1) polyreg วิธีนี้ใช้ประมาณค่าข้อมูลที่เป็นตัวแปรจัดประเภทที่มากกว่า 2 กลุ่มที่ไม่มีอันดับโดยใช้ตัวแบบถดถอยพหุกลุ่มในการประมาณค่าที่เหมาะสม สำหรับงานวิจัยนี้ใช้งานผ่านแพ็คเกจ “mice” ในโปรแกรม R 2) rf เป็นวิธีที่มีการประมาณค่าโดยใช้ตัวแบบ random forest ในการทำนายข้อมูลสูญหาย โดยใช้งานผ่านแพ็คเกจ “mice” ในโปรแกรม R และ 3) mixgb เป็นวิธีที่มีการประมาณค่าโดยใช้ตัวแบบ XGBoost ในการทำนายข้อมูลสูญหาย ใช้งานผ่านแพ็คเกจ “mixgb” ในโปรแกรม R สำหรับวิธีการปรับค่าตัวแปรตามมีการใช้ตัวกรองข้อมูลจากวิธีขจัดข้อมูลรบกวน 4 วิธี ร่วมกับการประมาณค่าทดแทนพหุ จำนวน 3 วิธี จึงมีวิธีย่อยสำหรับวิธีการปรับค่าตัวแปรตามทั้งหมด 12 วิธี แสดงดัง Table 1

Table 1 Combination between noise filter and multiple imputation for relabel method.

Noise filter	Multiple imputation method		
	polyreg	**rf**	**mixgb**
CNN	CNN + polyreg	CNN + rf	CNN + mixgb
ENN	ENN + polyreg	ENN + rf	ENN + mixgb
CVCF	CVCF + polyreg	CVCF + rf	CVCF + mixgb
IPF	IPF + polyreg	IPF + rf	IPF + mixgb

4. การสร้างตัวแบบการจำแนก งานวิจัยนี้ต้องการเปรียบเทียบประสิทธิภาพการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบ มีการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเรียนรู้ (training data) 70% และชุดข้อมูลทดสอบ (testing data) 30% สำหรับชุดข้อมูลเรียนรู้จะมีการจำลองข้อมูลรบกวนตามวิธีการในข้อ 2 ส่วนชุดข้อมูลทดสอบจะเป็นชุดข้อมูลที่ไม่ได้มีการจำลองข้อมูลรบกวน สำหรับตัวแบบจำแนกกลุ่มข้อมูลที่ใช้ในงานวิจัยนี้ทั้งหมด 4 ตัวแบบ ได้แก่ 1) k-NN เป็นอัลกอริทึมที่พิจารณาเพื่อนบ้านที่อยู่รอบๆ สำหรับการตัดสินใจว่าจะข้อมูลที่กำหนดจะอยู่ในกลุ่มใดตามเพื่อนบ้านที่อยู่ใกล้ที่สุด โดยจะพิจารณาจากระยะห่างระหว่างจุดข้อมูลจากนั้นจึงทำการเลือกกลุ่มโดยยึดเสียงข้างมาก 2) Random Forest (RF) เป็นอัลกอริทึมที่ใช้ต้นไม้การตัดสินใจหลายๆ ครั้ง แล้วพิจารณาผลลัพธ์ การแบ่งกลุ่มโดยยึดเสียงข้างมาก 3) Naïve Bayes (NB) เป็นอัลกอริทึมที่ใช้หลักการของความน่าจะเป็นจากพื้นฐานทฤษฎีของเบย์ (Bayes theorem) ซึ่งจะใช้การคำนวณความน่าจะเป็นแบบมีเงื่อนไข และ 4) Support Vector Machine (SVM) เป็นอัลกอริทึมที่แบ่ง Class ของ

ข้อมูลออกจากกัน โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการเรียนรู้ โดยเน้นการหาเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด สำหรับค่าพารามิเตอร์ในแต่ละตัวแบบ กำหนดให้ใช้เป็นค่าเริ่มต้นของตัวแบบ

5. การประเมินประสิทธิภาพของวิธีการจัดการข้อมูลรบกวน จะประเมินโดยใช้ค่า F1 ของตัวแบบการจำแนกที่มีการใช้ค่าจากตาราง confusion matrix งานวิจัยนี้มีเป้าหมายเพื่อการเปรียบเทียบวิธีการจัดการข้อมูลรบกวน จึงอธิบายความหมายของค่าต่างๆ ใน confusion matrix ได้ดังนี้ 1) True Positive (TP) คือ จำนวนข้อมูลที่เป็นข้อมูลรบกวน และถูกระบุว่าเป็นข้อมูลรบกวน 2) False Negative (FN) คือ จำนวนข้อมูลที่เป็นข้อมูลรบกวน แต่ถูกระบุว่าไม่เป็นข้อมูลรบกวน 3) False Positive (FP) คือ จำนวนข้อมูลที่ไม่เป็นข้อมูลรบกวน แต่ถูกระบุว่าเป็นข้อมูลรบกวน และ 4) True Negative (TN) คือ จำนวนข้อมูลที่ไม่เป็นข้อมูลรบกวน และถูกระบุว่าไม่เป็นข้อมูลรบกวน (Li & Mao, 2023; Nematzadeh *et al*., 2020) โดยค่าประสิทธิภาพ F1 จะคำนวณมาจากค่าเฉลี่ยฮาร์โมนิกของ precision และ recall ดังสมการ

$$\text{preciasion} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FP}$$

$$F1 = \frac{2 \text{ x precision x recall}}{\text{precision + recall}}$$

6. การวิเคราะห์ประสิทธิภาพของวิธีการจัดการข้อมูลรบกวน ทำโดยวิเคราะห์ความแปรปรวนหลายทาง (N-Way ANOVA) ซึ่งเป็นการเปรียบเทียบความแตกต่างของค่าเฉลี่ย F1 ที่มีปัจจัยหลายปัจจัยพร้อมๆ กัน ได้แก่ วิธีการจัดการข้อมูลรบกวน ขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนก พร้อมทั้งดูปฏิสัมพันธ์ระหว่างวิธีการจัดการข้อมูลรบกวน กับปัจจัยอื่นๆ ในสถานการณ์ที่จำลองข้อมูลขึ้น เพื่อวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบจำแนกกลุ่มข้อมูล เมื่อสถานการณ์ของขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนกกลุ่มข้อมูลที่แตกต่างกัน

ผลการวิจัย

1. ประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนในภาพรวม

จากการวิเคราะห์ความแปรปรวนหลายทาง พบว่า วิธีการจัดการข้อมูลรบกวนมีอิทธิพลปฏิสัมพันธ์กับทุกปัจจัย ได้แก่ ขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนกที่ส่งผลให้ค่าประสิทธิภาพ F1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยมีขนาดอิทธิพลที่พิจารณาจากค่า Partial eta squared () แสดงดัง Table 2

จากผลการวิเคราะห์ข้อมูลตาม Table 2 พบว่า วิธีการจัดการข้อมูลรบกวนมีอิทธิพลขนาดใหญ่ (= 0.63) ต่อค่าประสิทธิภาพ F1 กล่าวคือ วิธีการจัดการข้อมูลรบกวนที่แตกต่างกันส่งผลให้ค่าประสิทธิภาพ F1 แตกต่างกันมาก โดยในภาพรวมเมื่อจัดการข้อมูลรบกวนโดยวิธีการปรับค่าตัวแปรตามมีแนวโน้มที่มีค่าประสิทธิภาพ F1 สูงกว่าวิธีขจัดข้อมูลรบกวนเป็นส่วนใหญ่ โดยประสิทธิภาพของแต่ละวิธีแตกต่างกันตาม Figure1 สำหรับอิทธิพลปฏิสัมพันธ์จะนำเสนอแยกรายประเด็นในหัวข้อถัดไป

Table 2 Comparison of the effect size from Partial eta squared for each factor influencing the F1 score.

Effect	ηp^2	p - value
Model	0.03	< .001
Sample size	0.01	< .001
Noise level	0.31	< .001
Noise handling	0.63	< .001
Noise handling * Model	0.01	< .001
Noise handling * Sample size	0.12	< .001
Noise handling * Noise level	0.01	< .001

เมื่อพิจารณาตาม Figure1 พบว่า วิธีย่อยของวิธีการปรับค่าตัวแปรตาม พบว่า วิธี ENN + polyreg จะมีค่ามัธยฐานของค่าเฉลี่ย F1 สูงที่สุด นอกจากนี้ยังพบว่าวิธีปรับค่าตัวแปรตามที่มีการใช้ตัวกรองกลุ่ม ENN จะมีค่าประสิทธิภาพโดยเฉลี่ยสูงกว่าวิธีการที่ใช้ตัวกรองกลุ่มอื่น เมื่อพิจารณาวิธีย่อยของวิธีการขจัดข้อมูลรบกวน พบว่า วิธี IPF มีค่าเฉลี่ย F1 สูงที่สุด รองลงมาคือ ENN, CVCF และ CNN ตามลำดับ

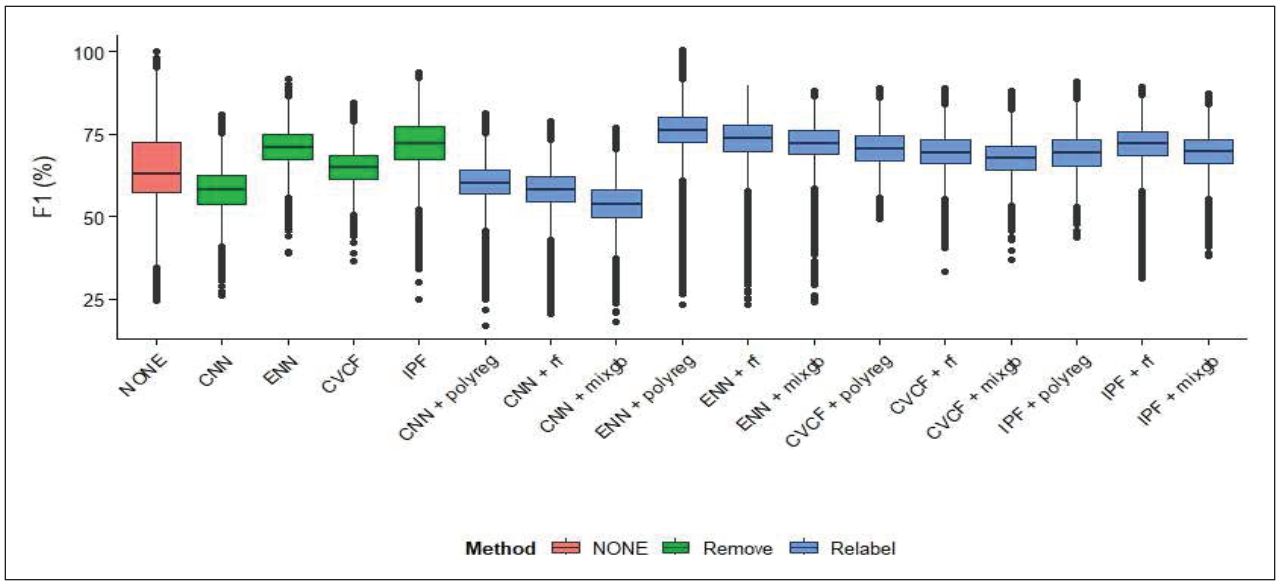

Figure 1 The average F1 score of each noise handling methods overall.

2. ผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนเมื่อพิจารณาร่วมกับขนาดตัวอย่าง

จากผลการวิเคราะห์ข้อมูลตาม Table 2 พบว่า ปฏิสัมพันธ์ของวิธีการจัดการข้อมูลรบกวนกับขนาดตัวอย่าง มีอิทธิพลขนาดกลาง (= 0.12) ต่อค่าประสิทธิภาพ F1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($p < .001$) โดยค่าประสิทธิภาพ F1 เมื่อพิจารณาเฉพาะวิธีการจัดการข้อมูลรบกวนร่วมกับขนาดตัวอย่างทั้ง 3 ขนาด ได้แก่ 100, 500 และ 1,000 หน่วย โดยใช้ปริมาณข้อมูลรบกวนทุกระดับ และทุกตัวแบบการจำแนก แสดงดัง Table 3

กรณีขนาดตัวอย่าง 100 หน่วย พบว่า วิธีการปรับค่าตัวแปรตามส่วนใหญ่มีแนวโน้มที่มีค่าเฉลี่ย F1 สูงกว่าวิธีการขจัดข้อมูลรบกวน ยกเว้นวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองกลุ่ม CNN ที่มีค่า F1 ต่ำกว่าการไม่จัดการข้อมูลรบกวน โดยวิธีการจัดการข้อมูลรบกวนที่มีประสิทธิภาพสูงสุดเมื่อขนาดตัวอย่าง 100 หน่วย ได้แก่ ENN + polyreg รองลงมา คือ ENN + rf และ ENN + mixgb ตามลำดับ

กรณีขนาดตัวอย่าง 500 หน่วย พบว่า วิธีการปรับค่าตัวแปรตามบางวิธีและวิธีการขจัดข้อมูลรบกวนบางวิธี มีแนวโน้มที่จะมีค่าประสิทธิภาพ F1 โดยเฉลี่ยใกล้เคียงกัน เมื่อพิจารณาวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองกลุ่ม CNN พบว่ามีประสิทธิภาพต่ำสุด เช่นเดียวกับกรณีขนาดตัวอย่าง 100 หน่วย โดยวิธีการจัดการข้อมูลรบกวนที่มีประสิทธิภาพสูงสุดเมื่อขนาดตัวอย่าง 500 หน่วย ได้แก่ ENN + polyreg รองลงมา คือ IPF และ IPF + rf ตามลำดับ

กรณีขนาดตัวอย่าง 1,000 หน่วย พบว่า วิธีการขจัดข้อมูลรบกวน มีแนวโน้มที่จะมีค่าเฉลี่ย F1 สูงกว่าวิธีการปรับค่าตัวแปรตามบางวิธี โดยเมื่อพิจารณาวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองกลุ่ม CNN มีประสิทธิภาพต่ำสุด เช่นเดียวกับกรณีขนาดตัวอย่าง 100 และ 500 หน่วย โดยวิธีการจัดการข้อมูลรบกวนที่มีประสิทธิภาพสูงสุดเมื่อขนาดตัวอย่าง 1,000 หน่วย ได้แก่ ENN รองลงมา คือ IPF และ ENN + polyreg ตามลำดับ

Table 3 The average F1 score of each noise handling method when considering with sample size.

method		Sample size		
		n = 100	n = 500	n = 1000
None	NONE	64.61	64.75	64.80
Remove	CNN	52.57	59.18	61.53
	ENN	67.91	71.08	74.19
	CVCF	62.48	65.18	66.79
	IPF	66.22	76.15	73.96
Relabel	CNN + polyreg	60.79	60.26	59.53
	CNN + rf	58.13	58.48	57.18
	CNN + mixgb	54.19	54.39	52.60
	ENN + polyreg	79.02	76.53	72.92
	ENN + rf	76.17	72.35	71.93
	ENN + mixgb	72.52	72.87	72.07
	CVCF + polyreg	71.03	70.50	71.04
	CVCF + rf	69.48	69.96	69.39
	CVCF + mixgb	67.60	68.04	67.96
	IPF + polyreg	70.85	68.07	69.81
	IPF + rf	71.57	73.16	71.92
	IPF + mixgb	70.10	70.15	69.24

3. ผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนเมื่อพิจารณาร่วมกับปริมาณข้อมูลรบกวน

จากผลการวิเคราะห์ข้อมูลตาม Table 2 พบว่า ปฏิสัมพันธ์ของวิธีการจัดการข้อมูลรบกวนกับปริมาณข้อมูลรบกวนมีอิทธิพลขนาดเล็ก (= 0.01) ต่อค่าประสิทธิภาพ F1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($p < .001$) โดยค่าประสิทธิภาพ F1 เมื่อพิจารณาวิธีการจัดการข้อมูลรบกวนร่วมกับปริมาณข้อมูลรบกวนทั้ง 4 ระดับ ได้แก่ 10%, 20%, 30% และ 40% โดยใช้ขนาดตัวอย่างทุกระดับ และทุกตัวแบบการจำแนก แสดงดัง Table 4

ค่าประสิทธิภาพ F1 เมื่อใช้วิธีการจัดการข้อมูลรบกวนในทุกวิธีมีแนวโน้มลดลงตามปริมาณข้อมูลรบกวนที่เพิ่มขึ้น โดยการจัดการข้อมูลรบกวนโดยวิธีการปรับค่าตัวแปรตามมีแนวโน้มที่จะมีค่าประสิทธิภาพ F1 สูงกว่าวิธีขจัดข้อมูลรบกวน ยกเว้นวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรอง CNN จะมีค่า F1 โดยเฉลี่ยต่ำกว่าการไม่จัดการข้อมูลรบกวน ซึ่งเป็นไปในทิศทางเดียวกันตามปริมาณรบกวนทั้ง 4 ระดับ

กรณีที่ปริมาณข้อมูลรบกวนเป็น 10% และ 20% พบว่า วิธีการที่มีค่าประสิทธิภาพ F1 สูงที่สุด ได้แก่ ENN + polyreg, ENN + rf และ ENN + mixgb ตามลำดับ ซึ่งเป็นวิธีการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองข้อมูล ENN ทั้งหมด แต่เมื่อพิจารณาวิธีการขจัดข้อมูลรบกวนโดยไม่มีการปรับค่าตัวแปรตาม พบว่าประสิทธิภาพของการใช้ตัวกรอง IPF จะมีแนวโน้มสูงกว่าการใช้ตัวกรอง ENN เล็กน้อย

กรณีที่ปริมาณข้อมูลรบกวนเป็น 30% และ 40% พบว่า วิธีการที่มีค่าประสิทธิภาพ F1 สูงที่สุด ได้แก่ ENN + polyreg, ENN + rf และ IPF ตามลำดับ ซึ่ง 2 วิธีแรกเป็นวิธีการปรับค่าตัวแปรตาม ส่วนวิธี IPF เป็นวิธีในกลุ่มของการขจัดข้อมูลรบกวน และมีค่าประสิทธิภาพใกล้เคียงกับวิธีการปรับค่าตัวแปรตามตัวอื่นๆ บางวิธี ซึ่งวิธีการทั้งปรับค่าตัวแปรตามและวิธีการขจัดข้อมูลรบกวน ในกรณีที่ประมาณรบกวนเป็น 30% และ 40% ส่วนใหญ่มีแนวโน้มที่มีค่าเฉลี่ย F1 ไม่ต่างกันมากนัก ยกเว้นวิธีที่มีการขจัดข้อมูลรบกวน และปรับค่าตัวแปรตามด้วยตัวกรอง CNN

Table 4 The average F1 score of each noise handling methods when considering with noise levels.

method		Noise level			
		10%	20%	30%	40%
None	NONE	67.55	66.08	64.11	61.13
Remove	CNN	61.50	58.70	56.15	54.70
	ENN	75.38	71.99	69.35	67.51
	CVCF	69.26	65.76	63.07	61.18
	IPF	75.81	72.68	70.67	69.27
Relabel	CNN + polyreg	65.40	60.62	58.96	55.79
	CNN + rf	63.88	57.21	55.80	54.84
	CNN + mixgb	59.65	53.45	51.83	49.99
	ENN + polyreg	81.51	76.71	75.13	71.28
	ENN + rf	79.39	73.12	71.85	69.57
	ENN + mixgb	78.18	72.38	70.57	68.82
Relabel	CVCF + polyreg	76.58	70.56	69.25	67.05
	CVCF + rf	75.06	69.45	67.96	65.97
	CVCF + mixgb	73.31	67.67	66.29	64.20
	IPF + polyreg	75.16	69.40	68.01	65.73
	IPF + rf	77.76	72.10	70.64	68.37
	IPF + mixgb	75.25	69.90	68.13	66.05

4. ผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนเมื่อพิจารณาร่วมกับตัวแบบที่ใช้ในการจำแนกกลุ่มข้อมูล

จากผลการวิเคราะห์ข้อมูลตาม Table 2 พบว่า ปฏิสัมพันธ์ของวิธีการจัดการข้อมูลรบกวนกับตัวแบบที่ใช้ในการจำแนกกลุ่มข้อมูลมีอิทธิพลขนาดเล็ก (η_p^2 = 0.01) ต่อค่าประสิทธิภาพ F1 อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ($p < .001$) โดยค่าประสิทธิภาพ F1 เมื่อพิจารณาวิธีการจัดการข้อมูลรบกวนร่วมกับตัวแบบที่ใช้ในการจำแนกกลุ่มข้อมูล ได้แก่ k-NN, Random Forest, Naïve Bayes และ Support Vector Machine โดยใช้ปริมาณข้อมูลรบกวนทุกระดับ และขนาดตัวอย่างทุกระดับ แสดงดัง Table 5

ผลการศึกษาพบว่า ในภาพรวมวิธีการจัดการข้อมูลรบกวนโดยการปรับค่าตัวแปรตาม จะมีแนวโน้มที่ให้ค่าประสิทธิภาพ F1 สูงกว่าวิธีการขจัดข้อมูลรบกวน เมื่อพิจารณาวิธีการปรับค่าตัวแปรตามและวิธีการขจัดข้อมูลรบกวนที่มีการใช้ตัวกรอง CNN จะมีค่าประสิทธิภาพ F1 ต่ำในทุกๆ ตัวแบบ

กรณีตัวแบบ k-NN และ Random Forest พบว่า วิธีการที่มีค่าประสิทธิภาพ F1 สูงที่สุดเป็นวิธีการเหมือนกัน ได้แก่ ENN + polyreg, ENN + rf และ IPF ตามลำดับ

กรณีตัวแบบ Naïve Bayes และ Support Vector Machine พบว่า วิธีการที่มีค่าประสิทธิภาพ F1 สูงที่สุดเป็นวิธีการเหมือนกัน ได้แก่ ENN + polyreg, ENN + rf และ ENN + mixgb ตามลำดับ ซึ่งเป็นวิธีการปรับค่าตัวแปรตามทั้งหมด นอกจากนี้ยังพบว่าส่วนใหญ่วิธีการปรับค่าตัวแปรตามจะมีประสิทธิภาพดีกว่าวิธีขจัดข้อมูลรบกวน

Table 5 The average F1 score of each noise handling methods when considering with classification models.

Method		Model			
		k-NN	RF	NB	SVM
None	NONE	65.98	64.26	63.57	65.06
Remove	CNN	57.84	58.68	56.50	58.03
	ENN	72.46	73.12	69.37	69.28
	CVCF	65.16	65.80	63.48	64.84
	IPF	73.45	74.11	70.48	70.39
Relabel	CNN + polyreg	60.73	60.90	59.08	60.07
	CNN + rf	58.69	58.99	56.43	57.61
	CNN + mixgb	53.65	55.03	52.56	53.67
	ENN + polyreg	76.29	77.48	74.94	75.93
	ENN + rf	73.96	74.45	72.08	73.43
	ENN + mixgb	73.02	73.48	71.24	72.22
	CVCF + polyreg	71.23	71.83	69.59	70.79
	CVCF + rf	70.00	70.62	68.31	69.50
	CVCF + mixgb	68.39	68.79	66.48	67.81
	IPF + polyreg	69.91	70.71	68.23	69.46
	IPF + rf	72.66	73.16	70.82	72.22
	IPF + mixgb	70.37	70.79	68.49	69.68

สรุปและวิจารณ์ผลการวิจัย

จากผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนที่เกิดขึ้นในตัวแปรตามของตัวแบบจำแนกกลุ่มข้อมูลเมื่อมีสถานการณ์ของขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนกกลุ่มข้อมูลที่แตกต่างกัน พบว่า วิธีการจัดการข้อมูลรบกวนมีอิทธิพลปฏิสัมพันธ์กับทุกปัจจัย ได้แก่ ขนาดตัวอย่าง ปริมาณข้อมูลรบกวน และตัวแบบการจำแนกที่ส่งผลให้ค่าประสิทธิภาพ F1 แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 โดยสามารถสรุปวิธีการจัดการข้อมูลรบกวนที่มีประสิทธิภาพสูงสุดในแต่ละสถานการณ์ได้ดัง Table 6

Table 6 Summary of the most effective class noise handling methods for each situation.

Situation		Method
Sample size	100	ENN + polyreg
	500	ENN + polyreg
	1,000	ENN
Noise level	10%	ENN + polyreg
	20%	ENN + polyreg
	30%	ENN + polyreg
	40%	ENN + polyreg
Model	k-NN	ENN + polyreg
	RF	ENN + polyreg
	NB	ENN + polyreg
	SVM	ENN + polyreg

เมื่อพิจารณาประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนร่วมกับขนาดตัวอย่าง พบว่าหากกลุ่มตัวอย่างมีขนาดเล็ก (n=100) วิธีการปรับค่าตัวแปรตามมีแนวโน้มที่จะมีประสิทธิภาพสูงกว่าวิธีขจัดข้อมูลรบกวน หากกลุ่มตัวอย่างขนาดกลาง (n=500) ประสิทธิภาพของวิธีปรับค่าตัวแปรตามส่วนใหญ่มีแนวโน้มใกล้เคียงกับวิธีการขจัดข้อมูลรบกวน และเมื่อกลุ่มตัวอย่างขนาดใหญ่ (n=1,000) การขจัดข้อมูลรบกวนมีแนวโน้มที่จะให้ประสิทธิภาพสูงกว่า โดยวิธีการที่มีแนวโน้มมีค่าประสิทธิภาพสูงสุดสำหรับตัวอย่างขนาดเล็กและกลาง คือ ENN + polyreg และสำหรับตัวอย่างขนาดใหญ่ คือ ENN จะเห็นว่าหากกลุ่มตัวอย่างขนาดเล็กการขจัดข้อมูลรบกวนจะทำให้สูญเสียข้อมูลที่จะนำไปเรียนรู้และชุดข้อมูลขนาดเล็กอาจจะทำให้เกิดปัญหา overfitting ได้ นอกจากนี้ในงานวิจัย Rajput *et al*. (2023) ยังพบว่าตัวแบบบางตัวแบบมีประสิทธิภาพไม่ดีหากกลุ่มตัวอย่างขนาดเล็ก ดังนั้น วิธีการปรับค่าตัวแปรตามจึงเป็นวิธีที่เหมาะสมกว่าสำหรับกลุ่มตัวอย่างขนาดเล็ก

เมื่อพิจารณาประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนร่วมกับปริมาณข้อมูลรบกวน พบว่า ประสิทธิภาพ F1 เมื่อใช้วิธีการจัดการข้อมูลรบกวนในทุกวิธีมีแนวโน้มลดลงตามปริมาณข้อมูลรบกวนที่เพิ่มขึ้น และวิธีการจัดการที่ให้ประสิทธิภาพสูงสุดในทุกระดับปริมาณข้อมูลรบกวน คือวิธีการปรับค่าตัวแปรตาม ENN + polyreg แต่หากพิจารณาวิธีการขจัดข้อมูลรบกวนเพียงอย่างเดียว ตัวกรอง IPF จะมีประสิทธิภาพดีกว่าตัวกรอง ENN แสดงให้เห็นว่าควรพิจารณาประสิทธิภาพของตัวกรองร่วมกับวิธีการประมาณค่าทดแทนพหุด้วย ซึ่งจะเป็นข้อเสนอแนะในการศึกษาครั้งต่อไป สอดคล้องกับงานวิจัยของ Luengo *et al*. (2021) ที่พบว่า หากมีการใช้ตัวกรองแบบผสมการเลือกตัวกรองหลักที่มีประสิทธิภาพดีไม่ได้เป็นการบอกว่าประสิทธิภาพของการจัดการกับข้อมูลรบกวนครั้งนั้นจะดีเสมอไปควรพิจารณาถึงประสิทธิภาพของวิธีการอื่นๆ ที่นำมาใช้ร่วมกันด้วย

เมื่อพิจารณาประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนร่วมกับตัวแบบที่ใช้ในการจำแนกกลุ่มข้อมูล พบว่า ในภาพรวมวิธีการจัดการข้อมูลรบกวนโดยการปรับค่าตัวแปรตามจะมีแนวโน้มที่ให้ค่าประสิทธิภาพ F1 สูงกว่าวิธีการขจัดข้อมูลรบกวนเล็กน้อย และวิธีการจัดการที่ให้ประสิทธิภาพสูงสุดในทุกตัวแบบที่ใช้ในการจำแนกกลุ่มข้อมูล คือวิธีการปรับค่าตัวแปรตาม ENN + polyreg

จากผลการศึกษาได้ข้อสรุปถึงประสิทธิภาพของวิธีการ ENN + polyreg ที่มีแนวโน้มให้ค่าประสิทธิภาพ F1 สูงสุดในเกือบทุกสถานการณ์ ยกเว้นกรณีที่กลุ่มตัวอย่างเป็น 1,000 หน่วย โดยวิธีการนี้เป็นการปรับค่าตัวแปรตามโดยใช้ตัวกรอง ENN ในการตรวจสอบข้อมูลรบกวน ที่มีหลักการคือ ตรวจจับข้อมูลที่มีค่าเพื่อนบ้านที่ใกล้ที่สุดหากค่าตัวแปรตามไม่ตรงกันจะดำเนินการขจัดออก ผลของการใช้ตัวกรอง ENN แสดงให้เห็นว่ามีแนวโน้มที่มีประสิทธิภาพดีเมื่อใช้ร่วมกับวิธีประมาณค่าทดแทนพหุในสถานการณ์ของงานวิจัยนี้ที่มีข้อมูลรบกวนแบบ NNAR ซึ่งสอดคล้องกับงานวิจัยของ (Sáez *et al*., 2013) ที่พบว่าการใช้ตัวกรอง ENN และตัวกรองอื่นที่พัฒนามาจาก ENN จะมีประสิทธิภาพดีกว่าตัวกรองกลุ่มอื่นๆ ส่วนวิธีการที่มีแนวโน้มให้ค่าประสิทธิภาพต่ำสุด คือ การใช้ตัวกรอง CNN ทั้งในการขจัดข้อมูลรบกวนและวิธีการปรับค่าตัวแปรตามที่ถึงแม้จะมีหลักการพื้นฐานที่ใช้หลักการ k-NN และระยะห่างระหว่างข้อมูล แต่อัลกอริทึมนี้จะลดขนาดของชุดข้อมูลลงให้เล็กที่สุดโดยใช้ 1-NN และมักจะใช้สำหรับการลดข้อมูลที่ไม่ให้สารสนเทศ คงเหลือเฉพาะข้อมูลที่สำคัญ (Sutton, 2012) ซึ่งไม่เหมาะสมที่จะนำมาใช้เป็นตัวกรองข้อมูลรบกวนรูปแบบ NNAR นี้ที่ข้อมูลในแต่ละกลุ่มจะมีความใกล้เคียงกัน

สำหรับข้อเสนอแนะในการศึกษาครั้งต่อไป สามารถขยายขอบเขตการศึกษาในการเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลรบกวนด้วยการปรับค่าตัวแปรตามที่มีการใช้ตัวกรองข้อมูลรบกวนร่วมกับวิธีการประมาณค่าทดแทนพหุวิธีอื่นๆ ที่ยังไม่ได้ทำการศึกษาในครั้งนี้พร้อมทั้งศึกษาปฏิสัมพันธ์ระหว่างตัวกรองข้อมูลรบกวนและวิธีประมาณค่าทดแทนพหุ และควรเพิ่มเติมถึงวิธีการจัดการข้อมูลรบกวนในข้อมูลรบกวนแบบ NCAR และ NAR เพื่อขยายองค์ความรู้ในการจัดการข้อมูลรบกวนนอกเหนือจากที่ได้ศึกษาไป

เอกสารอ้างอิง

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, *11*, 131–167. https://doi.org/10.1613/jair.606

Buczak, P., Chen, J. J., & Pauly, M. (2023). Analyzing the effect of imputation on classification performance under MCAR and MAR missing mechanisms. *Entropy*, *25*(3), Article 521. https://doi.org/10.3390/e25030521

Deng, Y., & Lumley, T. (2024). Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, *33*(2), 531–549. https://doi.org/10.1080/10618600.2023.2252501

Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Garcia, L. P., Lehmann, J., de Carvalho, A. C., & Lorena, A. C. (2019). New label noise injection methods for the evaluation of noise filters. *Knowledge-Based Systems*, *163*, 693–704. https://doi.org/10.1016/j.knosys.2018.09.031

Hassan, H., Ahmad, N. B., & Anuar, S. (2020, May). Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. *Journal of Physics: Conference Series*, *1529*(5), Article 052041. IOP Publishing. https://doi.org/10.1088/1742-6596/1529/5/052041

Kim, Y., Soyata, T., & Behnagh, R. F. (2018). Towards emotionally aware AI smart classroom: Current issues and directions for engineering and education. *IEEE Access*, *6*, 5308–5331. https://doi.org/10.1109/ACCESS.2018.2791861

Li, C., & Mao, Z. (2023). A label noise filtering method for regression based on adaptive threshold and noise score. *Expert Systems with Applications*, *228*, Article 120422. https://doi.org/10.1016/j.eswa.2023.120422

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *JAMA*, *314*(18), 1966–1967. https://doi.org/10.1001/jama.2015.15281

Luengo, J., Sanchez-Tarrago, D., Prati, R. C., & Herrera, F. (2021). Multiple instance classification: Bag noise filtering for negative instance noise cleaning. *Information Sciences*, *579*, 388–400. https://doi.org/10.1016/j.ins.2021.07.076

Miao, Q., Cao, Y., Xia, G., Gong, M., Liu, J., & Song, J. (2015). RBoost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically sTable base learners. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(11), 2216–2228. https://doi.org/10.1109/TNNLS.2015.2475750

Miller, D., & Soh, L.-K. (2013). Meta-reasoning algorithm for improving analysis of student interactions with learning objects using supervised learning. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 129–135). International Educational Data Mining Society. https://educationaldatamining.org/files/conferences/EDM2013/papers/paper_6.pdf

Moura, K. G., Prudêncio, R. B., & Cavalcanti, G. D. (2022). Label noise detection under the noise at random model with ensemble filters. *Intelligent Data Analysis*, *26*(5), 1119–1138. https://doi.org/10.3233/IDA-215980

Nematzadeh, Z., Ibrahim, R., & Selamat, A. (2020). Improving class noise detection and classification performance: A new two-filter CNDC model. *Applied Soft Computing*, *94*, Article 106428. https://doi.org/10.1016/j.asoc.2020.106428

Nguyen-Van, T. (2020). A discrete-time state estimation for nonlinear systems with noises. *IEEE Access*, *8*, 147089–147096. https://doi.org/10.1109/ACCESS.2020.3014377

Nicholson, B., Sheng, V. S., & Zhang, J. (2016). Label noise correction and application in crowdsourcing. *Expert Systems with Applications*, *66*, 149–162. https://doi.org/10.1016/j.eswa.2016.09.003

Prati, R. C., Luengo, J., & Herrera, F. (2019). Emerging topics and challenges of learning from noisy data in nonstandard classification: A survey beyond binary class noise. *Knowledge and Information Systems*, *60*(1), 63–97. https://doi.org/10.1007/s10115-018-1244-4

Rajput, D., Wang, W. J., & Chen, C. C. (2023). Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, *24*(1), Article 48. https://doi.org/10.1186/s12859-023-05156-9

Sáez, J. A. (2022). Noise models in classification: Unified nomenclature, extended taxonomy and pragmatic categorization. *Mathematics*, *10*(20), Article 3736. https://doi.org/10.3390/math10203736

Sáez, J. A., Galar, M., Luengo, J., & Herrera, F. (2013). Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness. *Information Sciences*, *247*, 1–20. https://doi.org/10.1016/j.ins.2013.06.002

Sáez, J. A., Krawczyk, B., & Wo□niak, M. (2016). On the influence of class noise in medical data classification: Treatment using noise filtering methods. *Applied Artificial Intelligence*, *30*(6), 590–609. https://doi.org/10.1080/08839514.2016.1193719

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, *227*, Article 120201. https://doi.org/10.1016/j.eswa.2023.120201

Sutton, O. (2012). *Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction* [Lecture notes]. Department of Mathematics, University of Leicester.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Zhang, J., & Cui, S. (2023). Investigating the number of Monte Carlo simulations for statistically stationary model outputs. *Axioms*, *12*(5), Article 481. https://doi.org/10.3390/axioms12050481

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, *22*(3–4), 177–210. https://doi.org/10.1007/s10462-004-0751-8