Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification

การปรับปรุงประสิทธิภาพการเรียนรู้ของเครื่องจักรในข้อมูลเรซูเม่ที่ไม่สมดุลโดยใช้ SMOTE สำหรับการจำแนกประเภทผู้สมัครงาน

Warissara Vasuarayasak[1], Natdanai Mangklang[2], Jirapat Somseang[3], and Chaisiri Sanitphonklang[4,*]
วริสรา วสุอารยะศักดิ์[1], ณัฐดนัย มังคลัง[2], จิรภัทร สมเสียง[3], และ ชัยศิริ สนิทพลกลาง[4,*]

Received: 15 March 2025;
Revised: 7 April 2025;
Accepted: 13 April 2025;
Published: 20 April 2025;

Abstract

The problem of data imbalance in machine learning processes is a significant limitation affecting model performance, particularly when the minority class has considerably fewer samples than the majority class. This imbalance causes the model to be biased and less accurate in classification. A widely adopted solution to this issue is using SMOTE, which generates new synthetic samples for the minority class by calculating the distance between existing data points. This research presents the application of SMOTE to improve the performance of machine learning models, specifically Decision Trees, Random Forests, and K-NN, on imbalanced datasets. Results indicate that the accuracy of the models improved from 0.83, 0.86, and 0.80 to 0.84, 0.88, and 0.82, respectively; there is statistical significance after applying SMOTE. This demonstrates the effectiveness of the technique in addressing data imbalance issues. The findings confirm that SMOTE enhances the ability to classify minority class data, resulting in models that are more accurate and suitable for applications in contexts where data imbalance is a concern. By helping reduce the time spent on personnel selection for the recruitment department.

Keywords: Class Imbalance, SMOTE, Machine Learning, Resume Classification

[1] Student, Program in Computer Science and Artificial Intelligence, Faculty of Science, Chadrakasem Rajabhat University, Bangkok 10900, Thailand; นักศึกษา สาขาวิชาวิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม จังหวัดกรุงเทพมหานคร 10900 ประเทศไทย; Email: 6511500461@chandra.ac.th

[2] Student, Program in Computer Science and Artificial Intelligence, Faculty of Science, Chadrakasem Rajabhat University, Bangkok 10900, Thailand; นักศึกษา สาขาวิชาวิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม จังหวัดกรุงเทพมหานคร 10900 ประเทศไทย; Email: 6511518166@chandra.ac.th

[3] Student, Program in Computer Science and Artificial Intelligence, Faculty of Science, Chadrakasem Rajabhat University, Bangkok 10900, Thailand; นักศึกษา สาขาวิชาวิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม จังหวัดกรุงเทพมหานคร 10900 ประเทศไทย; Email: 6511518513@chandra.ac.th

[4,*] Assistant Professor, Dr., Program in Computer Science and Artificial Intelligence, Faculty of Science, Chadrakasem Rajabhat University, Bangkok 10900, Thailand; ผู้ช่วยศาสตราจารย์ ดร., สาขาวิชาวิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม จังหวัดกรุงเทพมหานคร 10900 ประเทศไทย; Email: chaisiri.s@chandra.ac.th

*Corresponding authors: Chaisiri Sanitphonklang (chaisiri.s@chandra.ac.th)

Citation: Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

บทคัดย่อ

ปัญหาความไม่สมดุลของข้อมูลในกระบวนการเรียนรู้ของเครื่องเป็นข้อจำกัดสำคัญที่ส่งผลต่อประสิทธิภาพของโมเดล โดยเฉพาะในกรณีที่กลุ่มข้อมูลกลุ่มน้อยมีจำนวนน้อยกว่ากลุ่มข้อมูลกลุ่มใหญ่ ทำให้โมเดลเรียนรู้มีความลำเอียงและจำแนกข้อมูลได้ไม่แม่นยำ วิธีการแก้ไขปัญหานี้ในปัจจุบันที่ได้รับความนิยมคือการเพิ่มข้อมูลตัวอย่างด้วยวิธี SMOTE ซึ่งสร้างตัวอย่างข้อมูลใหม่ในกลุ่มชนกลุ่มน้อยโดยอาศัยการคำนวณระยะห่างระหว่างข้อมูลที่มีอยู่ งานวิจัยนี้นำเสนอการประยุกต์ใช้ SMOTE ในการปรับปรุงประสิทธิภาพการเรียนรู้ของเครื่องด้วยโมเดล Decision Tree, Random Forest, และ K-NN จากชุดข้อมูลที่ไม่สมดุล พบว่าค่าความถูกต้องของโมเดลเพิ่มขึ้นจาก 0.83, 0.86, และ 0.80 เป็น 0.84, 0.88, และ 0.82 ตามลำดับโดยมีนัยสำคัญทางสถิติหลังจากใช้ SMOTE ซึ่งแสดงให้เห็นถึงประสิทธิภาพของวิธีการดังกล่าวในการแก้ไขปัญหาความไม่สมดุลของข้อมูล ผลการวิจัยยืนยันว่า SMOTE ช่วยปรับปรุงความสามารถในการจำแนกข้อมูลชนกลุ่มน้อยทำให้โมเดลมีความแม่นยำและเหมาะสมต่อการใช้งานในบริบทที่ข้อมูลมีความไม่สมดุล โดยเข้ามาช่วยลดเวลาในการคัดเลือกบุคคลของฝ่ายสรรหาบุคลากรได้อีกด้วย

คำสำคัญ: ความไม่สมดุลของคลาส, SMOTE, การเรียนรู้ของเครื่อง, การจำแนกเรซูเม่

1. บทนำ (Introduction)

ในยุคปัจจุบัน กระบวนการสรรหาและคัดเลือกบุคลากรถือเป็นหนึ่งในองค์ประกอบสำคัญที่ส่งผลต่อความสำเร็จขององค์กร (Akkharatwiwatthanathorn, 2020) อย่างไรก็ตาม การพิจารณาคัดเลือกผู้สมัครจากเอกสารสมัครงาน (Resume) ซึ่งมักประกอบด้วยข้อมูลที่หลากหลายและซับซ้อน เป็นกระบวนการที่ต้องใช้เวลาและอาศัยความแม่นยำจากผู้คัดเลือกอย่างมาก ในบริบทนี้ การนำเทคโนโลยีปัญญาประดิษฐ์และการเรียนรู้ของเครื่องเข้ามาช่วยในการประมวลผลและแนะนำผู้สมัครที่เหมาะสมกับตำแหน่งงานจึงเป็นแนวทางที่มีศักยภาพในการเพิ่มประสิทธิภาพและลดความลำเอียงในการคัดเลือกบุคคล

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลการเรียนรู้แบบมีผู้สอนที่สามารถพยากรณ์ตำแหน่งงานที่เหมาะสมสำหรับผู้สมัครโดยพิจารณาข้อมูลที่ระบุใน Resume ของผู้สมัคร พร้อมทั้งใช้โมเดลสำหรับเรียนรู้แบบมีผู้สอน 3 อัลกอริทึมด้วยกันและทำการเปรียบเทียบประสิทธิภาพในการทำงานของทั้งสามอัลกอริทึม (Simarmata et.al., 2024; Valen-Dacanay & Palaoag, 2023; Maurya et.al., 2022) ซึ่งได้แก่ Decision Tree, Rainforest และ K-NN อย่างไรก็ตาม ความท้าทายที่สำคัญคือปริมาณข้อมูลที่มีจำกัด ซึ่งอาจส่งผลต่อประสิทธิภาพของการเรียนรู้ของโมเดล ในการแก้ไขปัญหาดังกล่าว การวิจัยนี้ได้นำวิธีการ SMOTE (Synthetic Minority Over-sampling Technique) เข้ามาช่วยในการสร้างข้อมูลตัวอย่างเพิ่ม เพื่อปรับสมดุลของข้อมูลและเพิ่มประสิทธิภาพของโมเดลในการเรียนรู้จากข้อมูลที่มีจำกัด (Sun et.al., 2023)

งานวิจัยนี้คาดหวังที่จะช่วยสนับสนุนฝ่ายทรัพยากรบุคคล (Human Resources: HR) ในการตัดสินใจที่มีความแม่นยำและเป็นธรรมมากขึ้น โดยมุ่งเน้นการลดระยะเวลาและความซับซ้อนในการพิจารณาคัดเลือกบุคคลผ่านการใช้ระบบที่พัฒนาขึ้น ทั้งนี้ยังเป็นการเปิดโอกาสให้องค์กรสามารถประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์เพื่อเพิ่มประสิทธิภาพในกระบวนการสรรหาและคัดเลือกบุคลากรในอนาคต

2. วัตถุประสงค์งานวิจัย (Research Objectives)

1. เพื่อพัฒนาโมเดลการเรียนรู้แบบมีผู้สอนที่สามารถพยากรณ์ตำแหน่งงานที่เหมาะสมสำหรับผู้สมัครโดยพิจารณาข้อมูลที่ระบุใน Resume ของผู้สมัครที่มีข้อมูลไม่สมดุล

2. เพื่อเปรียบเทียบประสิทธิภาพโมเดลการเรียนรู้แบบมีผู้สอนของเครื่องเรียนรู้ด้วย 3 โมเดล ได้แก่ Decision Tree, Rainforest และ K-NN ในแง่ค่าความถูกต้อง (Accuracy) สำหรับข้อมูลก่อนและหลังสมดุลข้อมูล

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48.
https://doi.org/10.14456/jcct.2025.4.

3. กรอบแนวคิดงานวิจัย (Conceptual Framework)

Input ประกอบด้วยข้อมูลสรุปของผู้สมัครงานและข้อมูลทักษะของตำแหน่งงานใน 3 ด้านหลัก ซึ่งเป็นข้อมูลพื้นฐานที่ใช้ในการวิเคราะห์

Process เน้นการวิเคราะห์ข้อมูลเชิงลึกในหลายมิติ เช่น การศึกษาความต้องการของตำแหน่งงาน การประเมินทักษะ การวิเคราะห์โครงสร้างองค์กร และการประเมินคุณภาพข้อมูล นอกจากนี้ ยังมีการจัดการปัญหาข้อมูลไม่สมดุล (Data Imbalance) ซึ่งเกิดจากจำนวนข้อมูลในบางกลุ่มที่มีน้อยกว่ากลุ่มอื่น ส่งผลต่อความแม่นยำของโมเดล เพื่อแก้ไขปัญหานี้ได้มีการใช้เทคนิค SMOTE (Synthetic Minority Oversampling Technique) ในการสร้างข้อมูลตัวอย่างเสริมจากกลุ่มที่มีข้อมูลน้อยเพื่อเพิ่มความสมดุลของข้อมูลก่อนการสร้างโมเดลการจับคู่

Output คือ โมเดลการจับคู่ผู้สมัครงานกับตำแหน่งงานที่เหมาะสมบนพื้นฐานของข้อมูลที่ได้รับการปรับสมดุล

Outcome คือ การนำโมเดลไปใช้ในองค์กรเพื่อเพิ่มประสิทธิภาพในการคัดเลือกบุคลากร ลดอัตราการลาออก และเพิ่มความแม่นยำในการตัดสินใจ และขอนำเสนอกรอบแนวคิดการวิจัยครั้งนี้ ดัง Figure 1.

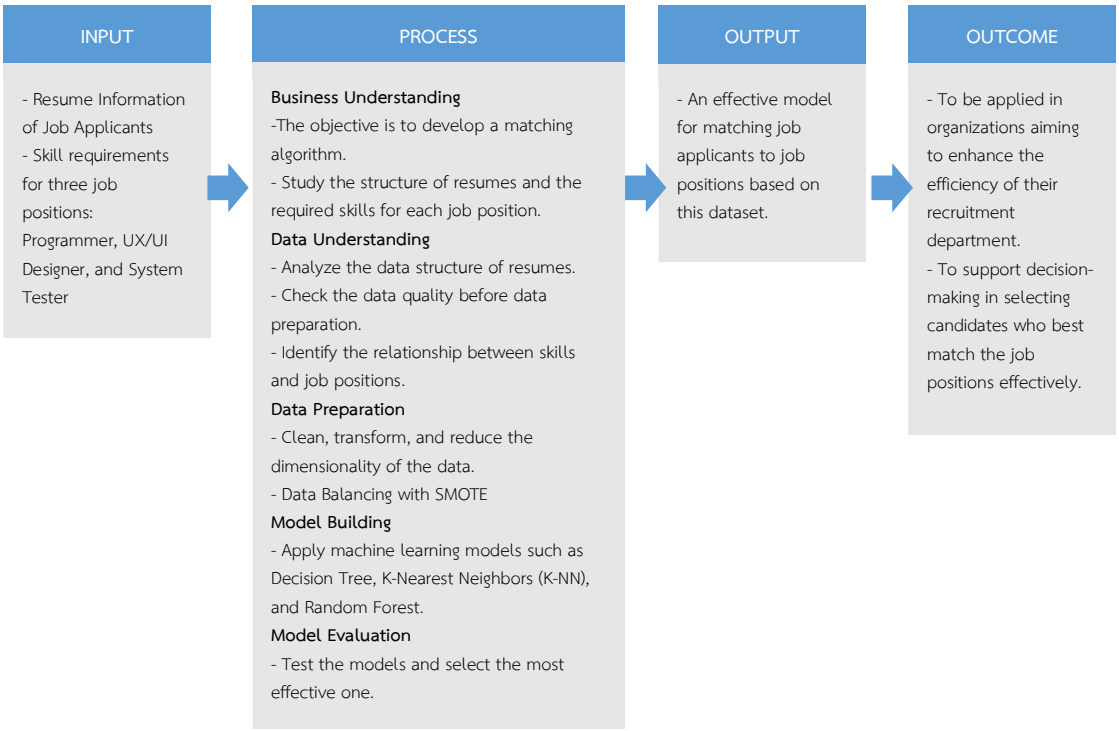


Figure 1. Conceptual Framework

จาก Figure 1. ได้ดำเนินการภายใต้กระบวนการเครื่องเรียนรู้ (Machine Learning Process) 6 ขั้นตอน ได้แก่ การรวบรวมข้อมูล (Data Collection) การจัดเตรียมข้อมูล (Data Preparation) การสมดุลข้อมูล (Imbalanced Data) การแบ่งข้อมูล (Split Dataset) การสร้างโมเดล (Modeling) และการนำไปใช้งาน ซึ่งได้แสดงลำดับดำเนินการดัง Figure 2.

4. การทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้อง (Literature Review)

สำหรับการวิจัยครั้งนี้ได้เล็งเห็นความสำคัญในการทบทวนวรรณกรรมที่เกี่ยวข้อง เพื่อนำมาสู่การปรับปรุงหรือต่อยอดในงานเดิมซึ่งจะสร้างความน่าเชื่อถือให้กับบทความนี้ โดยครั้งนี้ได้มีการทบทวนวรรณกรรม ดังนี้

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

4.1 การคัดเลือกบุคคลผ่านทางเอกสารสมัครงาน (Resume)

การคัดเลือกผู้สมัครงานผ่านเอกสารสมัครงาน (Resume) เป็นขั้นตอนสำคัญในกระบวนการสรรหาบุคคล โดยในปัจจุบันองค์กรส่วนใหญ่อาศัยการพิจารณาโดยเจ้าหน้าที่ทรัพยากรบุคคล (HR) ซึ่งใช้เกณฑ์ที่หลากหลาย เช่น การศึกษา ประสบการณ์ และทักษะ อย่างไรก็ตาม กระบวนการดังกล่าวเผชิญกับความท้าทาย เช่น ความลำเอียงในการตัดสินใจ ความไม่สอดคล้องของข้อมูลที่กรอกมา และปริมาณผู้สมัครที่มากเกินไป ส่งผลให้การคัดเลือกอาจขาดความแม่นยำและประสิทธิภาพ

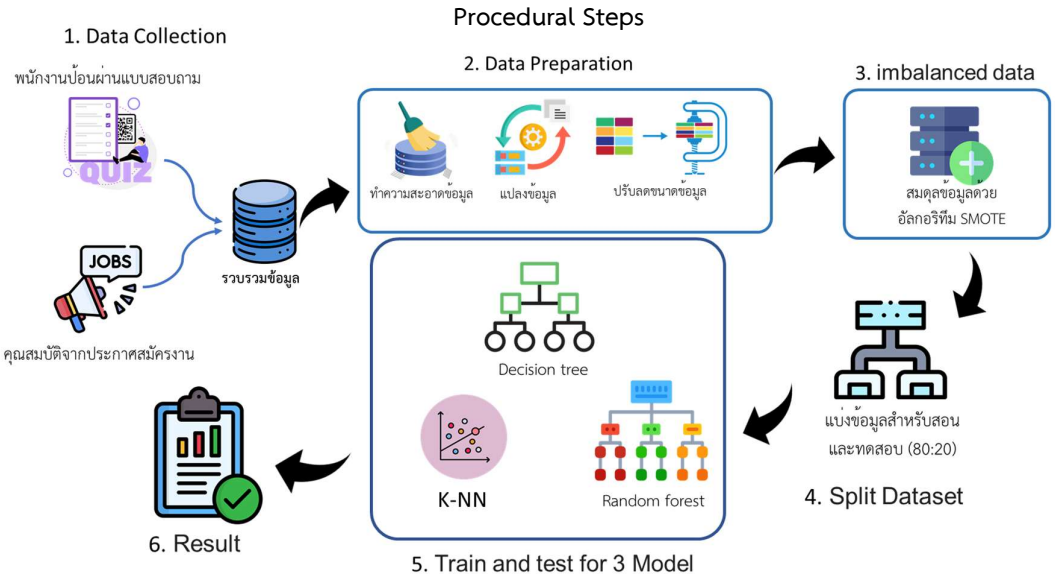


Figure 2. Procedural steps within the Machine Learning process.

แนวทางที่ถูกนำมาใช้แก้ไขปัญหาในปัจจุบันคือการประยุกต์ใช้ระบบปัญญาประดิษฐ์ (AI) และการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งช่วยประมวลผลข้อมูลจาก Resume อย่างรวดเร็วและอัตโนมัติ ตัวอย่างวิธีการ เช่น การวิเคราะห์ข้อความด้วยเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การสร้างโมเดลการเรียนรู้แบบมีผู้สอนเพื่อจัดกลุ่มผู้สมัคร และการใช้เทคนิค SMOTE เพื่อแก้ปัญหาข้อมูลไม่สมดุล วิธีการเหล่านี้ช่วยเพิ่มความแม่นยำ ลดอคติ และปรับปรุงกระบวนการคัดเลือกบุคคลอย่างมีประสิทธิภาพมากขึ้น ซึ่งปัจจุบันนี้ได้มีการนำเอาเทคโนโลยีปัญญาประดิษฐ์เข้ามาช่วยแก้ไขปัญหาการคัดเลือกผู้สมัครจากประวัติย่อ (Patil, 2023; Ali et al., 2022; Modak et al., 2024) ซึ่งได้ประสิทธิภาพที่ดีขึ้น แต่ยังเจอปัญหาความครบถ้วนของข้อมูลผู้สมัครที่ไม่เพียงพอเนื่องจากการจำกัดปริมาณข้อมูลประวัติอย่างย่อที่น้องไป

4.2 การจำแนกข้อมูล (Classification)

การจำแนกข้อมูล เป็นเทคนิคสำคัญในด้านวิทยาศาสตร์ข้อมูล (Data Science) และการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งมีบทบาทสำคัญในการวิเคราะห์และพยากรณ์ข้อมูล โมเดลแบบจำแนกถูกออกแบบมาเพื่อแบ่งชุดข้อมูลออกเป็นกลุ่มหรือคลาส โดยใช้คุณสมบัติต่าง ๆ ของข้อมูลเป็นปัจจัยในการตัดสินใจ ตัวอย่างการใช้งานเช่น การวินิจฉัยโรค การประเมินเครดิตทางการเงิน และการคัดกรองอีเมลสแปม

การประยุกต์ใช้โมเดลจำแนกข้อมูลช่วยเพิ่มประสิทธิภาพในงานที่เกี่ยวข้องกับการตัดสินใจ เช่น การลดเวลาที่ใช้ในการวิเคราะห์ข้อมูล การลดความผิดพลาดจากมนุษย์ และการเพิ่มความแม่นยำในการคัดเลือกข้อมูล อย่างไรก็ตาม ความท้าทายของการใช้โมเดลจำแนกข้อมูล ได้แก่ การเลือกคุณสมบัติที่เหมาะสม (Feature Selection) การแก้ไขปัญหาข้อมูลไม่สมดุล (Imbalanced Data) และการจัดการกับข้อมูลที่มีมิติสูง (High Dimensional Data) เทคนิคต่าง ๆ เช่น การใช้

Citation: Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

วิธีการ Oversampling อย่าง SMOTE (Synthetic Minority Over-sampling Technique) และการลดมิติข้อมูลด้วย PCA (Principal Component Analysis) ถูกนำมาใช้เพื่อแก้ปัญหาเหล่านี้

โมเดลแบบจำแนกข้อมูลมีหลักการทำงานโดยพิจารณาจากตัวแปรต้น (Input Features) และตัวแปรผลลัพธ์ (Output Class) จากชุดข้อมูลที่มีการกำหนดคลาสไว้แล้ว (Supervised Learning) เพื่อสร้างแบบจำลองที่สามารถทำนายคลาสของข้อมูลใหม่ได้ คุณลักษณะสำคัญของโมเดลเหล่านี้คือความสามารถในการจัดการข้อมูลที่มีความซับซ้อน และการให้ผลลัพธ์ที่แม่นยำและสอดคล้องกับบริบท สามารถจำแนกโมเดล ดังนี้

1. Decision Tree เป็นโมเดลที่อาศัยโครงสร้างต้นไม้ในการแบ่งข้อมูลออกเป็นกลุ่มย่อยตามเกณฑ์ที่กำหนด (Patel & Rana, 2014) โหนดแต่ละโหนดในต้นไม้แทนตัวแปรหรือลักษณะเฉพาะของข้อมูลที่ถูกเลือกให้เป็นเกณฑ์ตัดสินใจ โดยการสร้างต้นไม้จะเลือกตัวแปรที่ช่วยลดความไม่แน่นอน (Uncertainty) ของข้อมูล เช่น การวัดโดยค่า Gini Index หรือ Entropy หลักการทำงานของ Decision Tree ทำให้โมเดลสามารถแปลผลได้ง่ายและเข้าใจได้ชัดเจน ซึ่งเหมาะสำหรับงานที่ต้องการความโปร่งใสในการตัดสินใจ ข้อดีของ Decision Tree คือความเรียบง่ายและความสามารถในการจัดการข้อมูลประเภทต่าง ๆ แต่ข้อเสียคือความเสี่ยงที่จะเกิด Overfitting เมื่อข้อมูลมีความซับซ้อน

2. Random Forest เป็นการขยายตัวของ (Gehrke et al., 2000) Decision Tree โดยการรวม Decision Trees หลายต้นเข้าด้วยกันผ่านกระบวนการ Bagging หรือ Bootstrap Aggregating วิธีการนี้ช่วยลดความเอนเอียงของโมเดล (Bias) และเพิ่มความแม่นยำในผลลัพธ์ Random Forest ใช้การสุ่มตัวแปรที่นำมาใช้ในการสร้างต้นไม้แต่ละต้น ทำให้โมเดลมีความหลากหลายและมีความทนทานต่อข้อมูลที่มี Noise จุดเด่นของ Random Forest คือความสามารถในการจัดการข้อมูลขนาดใหญ่และซับซ้อน แต่ข้อเสียคือการตีความผลลัพธ์ของโมเดลทำได้ยากกว่าการใช้ Decision Tree

3. K-Nearest Neighbors (K-NN) เป็นโมเดลที่ทำงานโดยเปรียบเทียบข้อมูลใหม่กับข้อมูลตัวอย่างที่อยู่ใกล้เคียงในพื้นที่มิติข้อมูล (Feature Space) โมเดลนี้ใช้ระยะทาง (Cheng et al., 2014) เช่น ระยะทางแบบ Euclidean หรือ Manhattan ในการวัดความใกล้เคียง หลักการทำงานของ K-NN คือการนับคลาสที่พบมากที่สุดในกลุ่มข้อมูลตัวอย่างใกล้เคียงจำนวน k ตัว ข้อดีของ K-NN คือความง่ายในการใช้งานและความสามารถในการรองรับข้อมูลที่มีรูปแบบต่าง ๆ อย่างไรก็ตาม ข้อเสียคือความต้องการทรัพยากรสูงเมื่อข้อมูลมีขนาดใหญ่ และประสิทธิภาพที่ลดลงเมื่อข้อมูลมีมิติมาก

โมเดล Decision Tree และ Random Forest มีจุดเด่นในการแปลผลและการจัดการข้อมูลที่มีโครงสร้างชัดเจน ในขณะที่ K-NN เหมาะกับการประมวลผลข้อมูลแบบเรียลไทม์ แต่มีข้อจำกัดเมื่อเผชิญกับข้อมูลขนาดใหญ่ Decision Tree เหมาะสำหรับงานที่ต้องการความเข้าใจง่าย ส่วน Random Forest เหมาะกับงานที่ต้องการความแม่นยำสูง และ K-NN เหมาะสำหรับงานที่ต้องการการทำนายข้อมูลที่ไม่มีความซับซ้อนมาก

การเลือกโมเดลจำแนกข้อมูลควรพิจารณาจากลักษณะของชุดข้อมูลและเป้าหมายของการวิเคราะห์ ทั้งนี้ การประยุกต์ใช้ Decision Tree, Random Forest และ K-NN ในงานวิจัยและการประยุกต์ใช้จริงได้พิสูจน์ให้เห็นถึงประสิทธิภาพของโมเดลจำแนกข้อมูลในการเพิ่มความแม่นยำและประสิทธิภาพของระบบ

4.3. วิธีการ SMOTE

Synthetic Minority Oversampling Technique : SMOTE (Chawla et al., 2012; Rezki et al, 2024; Aubaidan et al., 2024; Boonchob, 2022) เป็นวิธีการแก้ปัญหาข้อมูลไม่สมดุล โดยเน้นสร้างข้อมูลสังเคราะห์ในคลาสส่วนน้อยที่มีจำนวนสองคลาส เพื่อให้โมเดลเรียนรู้ลักษณะข้อมูลได้ดียิ่งขึ้น ซึ่งถือว่ายังเป็นช่องว่างที่เกิดขึ้นเพราะแท้จริงแล้วข้อมูลจำนวนมากมีหลายคลาส ขั้นตอนเริ่มจากการเลือกตัวอย่างในคลาสส่วนน้อยแบบสุ่ม หาเพื่อนบ้านใกล้เคียงด้วยระยะทาง Euclidean และสร้างตัวอย่างใหม่ด้วยการคำนวณค่ากลางเชิงเส้นระหว่างตัวอย่างต้นแบบและเพื่อนบ้าน วิธีนี้ลดปัญหา Overfitting จากการคัดลอกข้อมูลเดิมแบบซ้ำ

SMOTE ถูกใช้อย่างกว้างขวางในงานวิเคราะห์ข้อมูล เช่น การจำแนกประเภทหลายหมวดหมู่ด้วยการเรียนรู้ของเครื่องเพื่อพยากรณ์ระดับน้ำตาลสะสมในเลือดของผู้ป่วยเบาหวาน (Posri, 2024) ช่วยเพิ่มความแม่นยำในการทำนายข้อมูลกลุ่มส่วนน้อย อย่างไรก็ตาม ข้อจำกัดอยู่ที่การสร้างข้อมูลในกรณีที่ข้อมูลมี Noise สูง หรือความไม่สมดุลที่รุนแรง ซึ่งอาจทำให้ข้อมูลสังเคราะห์ไม่สะท้อนความเป็นจริงได้อย่างเหมาะสม ขั้นตอนของวิธีการ SMOTE ดังนี้

1. การเลือกข้อมูลต้นแบบ (Random Selection)

เลือกตัวอย่างในคลาสส่วนน้อยแบบสุ่มหนึ่งตัวอย่างจากชุดข้อมูลต้นฉบับ

2. การเลือกเพื่อนบ้านใกล้เคียง (K-Nearest Neighbors)

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

คำนวณระยะห่างระหว่างตัวอย่างต้นแบบกับตัวอย่างอื่นในคลาสเดียวกันโดยใช้ระยะทาง Euclidean และเลือกเพื่อนบ้านใกล้เคียงจำนวน K ตัว

3. การสร้างตัวอย่างสังเคราะห์ เลือกตัวอย่างหนึ่งในเพื่อนบ้านใกล้เคียง จากนั้นสร้างตัวอย่างใหม่โดยการคำนวณค่ากลางเชิงเส้น (Linear Interpolation) ระหว่างตัวอย่างต้นแบบและตัวอย่างที่เลือก โดยสูตรดังแสดงไว้ในสมการที่ 1

$$X_{syntheti} = X_{original} + \lambda\left(X_{neighbor} - X_{original}\right) \quad \text{... (1)}$$

โดยที่ λ เป็นค่าสุ่มในช่วง [0,1]

ข้อมูลไม่สมดุล ด้วยการสร้างตัวอย่างสังเคราะห์ที่ช่วยเสริมความหลากหลายของข้อมูล เทคนิคนี้ได้แสดงให้เห็นถึงความสำคัญในหลากหลายบริบทของการแก้ปัญหาด้านข้อมูล อย่างไรก็ตาม ควรพิจารณาข้อจำกัดและปรับใช้ให้เหมาะสมกับลักษณะของข้อมูลและบริบทของปัญหา

5. วิธีดำเนินงานวิจัย (Research Methodology)

การวิจัยนี้มุ่งพัฒนาโมเดลจำแนกบุคคลให้เหมาะสมกับตำแหน่งงานโดยอิงข้อมูลจากเอกสารประวัติผู้สมัคร (Resume) ผ่านกระบวนการเครื่องเรียนรู้แบบมีผู้สอน (Supervised Learning) เน้นการจัดการข้อมูลที่ไม่สมดุลและคัดเลือกคุณลักษณะสำคัญ เพื่อเพิ่มความแม่นยำของโมเดลในการสนับสนุนการตัดสินใจของฝ่ายสรรหาบุคคลในองค์กร ภายใต้ 6 ขั้นตอน ดังนี้

5.1 การรวบรวมข้อมูล

การรวบรวมข้อมูลสำหรับการวิจัยครั้งนี้ดำเนินการผ่านแบบสอบถามบน Google Form โดยสอบถามความคิดเห็นเกี่ยวกับคุณลักษณะและทักษะที่จำเป็น ทั้ง Soft Skills เช่น การสื่อสาร การแก้ปัญหา และ Hard Skills เช่น การเขียนโปรแกรม การออกแบบ UX/UI และการทดสอบระบบ จากผู้ตอบแบบสอบถามที่มีประสบการณ์ในสายงานที่เกี่ยวข้อง นอกจากนี้ ข้อมูลยังถูกสกัดจากเอกสารประกาศรับสมัครงานในตำแหน่งโปรแกรมเมอร์ นักออกแบบ UX/UI และนักทดสอบระบบ ข้อมูลที่ได้จะถูกนำมาใช้สร้างชุดข้อมูลสำหรับฝึกสอนโมเดลจำแนกบุคคลให้เหมาะสมกับตำแหน่งงาน เพื่อเพิ่มความแม่นยำในการพิจารณาคัดเลือกบุคคล ซึ่งรวบรวมได้ทั้งสิ้นจำนวน 200 รายการ

5.2 การทำความเข้าใจกับข้อมูล

การทำความเข้าใจกับข้อมูลที่รวบรวมมาเป็นขั้นตอนสำคัญในกระบวนการเรียนรู้ของเครื่อง (Machine Learning) โดยข้อมูลที่ได้จากแบบสอบถามและเอกสารประกาศรับสมัครงานจะถูกตรวจสอบความสมบูรณ์ ความถูกต้อง และความสอดคล้องกับเป้าหมายการวิจัย ข้อมูลจะถูกจำแนกเป็นสองประเภท ได้แก่ Soft Skills และ Hard Skills ซึ่งข้อมูลแต่ละชุดจะถูกวิเคราะห์เพื่อระบุคุณลักษณะที่มีผลต่อการจำแนกตำแหน่งงาน เช่น โปรแกรมเมอร์ นักออกแบบ UX/UI และนักทดสอบระบบ

กระบวนการนี้รวมถึงการแปลงข้อมูลข้อความให้อยู่ในรูปแบบเชิงตัวเลขที่เหมาะสม เช่น การใช้เทคนิคการเข้ารหัส (Encoding) และการสร้างเวกเตอร์คุณลักษณะ (Feature Vector) เพื่อให้โมเดลสามารถนำไปวิเคราะห์ได้อย่างมีประสิทธิภาพ นอกจากนี้ ยังทำการจัดการกับข้อมูลที่ขาดสมดุล เช่น การใช้เทคนิค SMOTE เพื่อเพิ่มจำนวนตัวอย่างในกลุ่มที่มีจำนวนน้อย โดยกระบวนการทั้งหมดนี้มีเป้าหมายเพื่อเตรียมข้อมูลที่เหมาะสมและมีคุณภาพสำหรับการเรียนรู้ของโมเดล

5.3 การเตรียมข้อมูล

การจัดเตรียมข้อมูลสำหรับการวิจัยนี้ได้ดำเนินการผสานข้อมูลจากสองแหล่งสำคัญ ได้แก่ข้อมูลจากแบบสอบถามบน Google Form ซึ่งรวบรวมข้อมูลเกี่ยวกับทักษะและคุณลักษณะที่จำเป็นของพนักงาน และข้อมูลจากประกาศรับสมัครงานในรูปแบบออนไลน์ที่เกี่ยวข้องกับตำแหน่งโปรแกรมเมอร์ นักออกแบบ UX/UI และนักทดสอบระบบ โดยข้อมูลจากทั้งสองแหล่งได้ผ่านกระบวนการรวมกันและวิเคราะห์ให้อยู่ในรูปแบบเดียวกัน เพื่อให้พร้อมสำหรับการนำไปใช้งานในกระบวนการเรียนรู้ของเครื่อง

ข้อมูลทั้งหมดถูกแบ่งออกเป็นสองประเภทหลัก ได้แก่ Soft Skills เช่น ความคิดสร้างสรรค์ การทำงานเป็นทีม และการแก้ปัญหา และ Hard Skills เช่น การเขียนโปรแกรม การออกแบบอินเทอร์เฟซ และการทดสอบซอฟต์แวร์ โดยข้อมูล

Citation: Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

ถูกจัดกลุ่มตามตำแหน่งงานเพื่อระบุทักษะที่สำคัญสำหรับแต่ละตำแหน่ง นอกจากนี้ ได้มีการทำความสะอาดข้อมูล (Data Cleaning) เช่น การจัดการข้อมูลที่สูญหายหรือไม่สมบูรณ์ โดยการใช้วิธีการเติมข้อมูลด้วยค่าที่เหมาะสม

ข้อมูลที่ผ่านการทำความสะอาดถูกแปลงให้อยู่ในรูปแบบที่เหมาะสมสำหรับโมเดลการเรียนรู้ เช่น การแปลงข้อมูลเชิงหมวดหมู่ (Categorical Data) เป็นค่าตัวเลขผ่านเทคนิค Encoding และการปรับขนาดข้อมูล (Scaling) เพื่อลดความแปรปรวนของข้อมูลและเพิ่มประสิทธิภาพของการเรียนรู้ กระบวนการนี้ช่วยให้ข้อมูลมีคุณภาพและพร้อมสำหรับขั้นตอนการสร้างโมเดลจำแนกตำแหน่งงานอย่างมีประสิทธิผล โดยครั้งนี้ได้จัดแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลฝึก (Training Set) กับ ข้อมูลทดสอบ (Test Set) โดยมีอัตราส่วน 70:30

5.4. การเพิ่มจำนวนข้อมูลตัวอย่างด้วย SMOTE

การเพิ่มจำนวนข้อมูลตัวอย่างด้วยวิธี SMOTE (Synthetic Minority Over-sampling Technique) เป็นเทคนิคที่ได้รับความนิยมในการแก้ไขปัญหาความไม่สมดุลของข้อมูล (Data Imbalance) โดยเฉพาะอย่างยิ่งในกรณีที่จำนวนตัวอย่างของกลุ่มข้อมูลชนกลุ่มน้อย (Minority Class) มีไม่เพียงพอ ซึ่งอาจส่งผลให้โมเดลการเรียนรู้ของเครื่องขาดความแม่นยำในการจำแนกกลุ่มดังกล่าว อาทิเช่นการจำแนกโรคราก่อนเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึมต้นไม้ตัดสินใจพบว่าเมื่อมีการสร้างความสมดุลของข้อมูลด้วย SMOTE แล้วประสิทธิภาพการจำแนกมีความแม่นยำเพิ่มขึ้นจากเดิม 97.33% เป็น 99.33% (Sanitphonklang, 2023)

สำหรับกรณีนี้ มีข้อมูลเริ่มต้นเพียง 200 รายการ ประกอบด้วย 3 คลาส ได้แก่ คลาสโปรแกรมเมอร์ (P) จำนวน 113 รายการ คลาสนักทดสอบระบบ (T) จำนวน 50 รายการ และ คลาสนักออกแบบ UX/UI (U) จำนวน 37 รายการ เมื่อพิจารณาแล้วเห็นมีสองคลาสที่มีจำนวนรายการน้อยมาก คือ คลาสโปรแกรมเมอร์กับคลาสนักออกแบบ UX/UI ในวิจัยนี้จึงได้ใช้เทคนิค SMOTE โดยขั้นตอนแรกของการใช้ SMOTE คือการสร้างข้อมูลตัวอย่างใหม่ในกลุ่มข้อมูลชนกลุ่มน้อยผ่านการสังเคราะห์ข้อมูล โดยพิจารณาคุณลักษณะ (Features) ที่มีอยู่ในข้อมูลเดิม ขั้นตอนสำคัญมีดังนี้

1. การเลือกข้อมูลต้นแบบ ข้อมูลต้นแบบ (Sample) จากกลุ่มข้อมูลชนกลุ่มน้อยจะถูกเลือกแบบสุ่ม
2. การเลือกข้อมูลเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors) สำหรับข้อมูลต้นแบบที่เลือก จะค้นหาข้อมูลเพื่อนบ้านที่ใกล้ที่สุด K ตัวอย่าง (โดยทั่วไป K มีค่าประมาณ 5) โดยใช้ระยะห่างเชิง Euclidean ระหว่างข้อมูลในเชิงคุณลักษณะ
3. การสังเคราะห์ข้อมูลใหม่ สร้างข้อมูลตัวอย่างใหม่โดยการคำนวณค่าถ่วงน้ำหนักระหว่างข้อมูลต้นแบบกับเพื่อนบ้านที่ใกล้ที่สุดที่เลือกมา

ในกรณีนี้ ข้อมูลใหม่จำนวน 101 รายการถูกสร้างเพิ่มขึ้นให้กับคลาสที่มีจำนวนตัวอย่างน้อย 2 คลาส คือ คลาสนักทดสอบระบบกับนักออกแบบ UX/UI ได้เป็น นักทดสอบระบบจากเดิมมีจำนวนตัวอย่าง 50 รายการ เพิ่มเป็น 94 รายการ เพิ่มขึ้น 44 รายการ สำหรับนักออกแบบระบบจากเดิมมีจำนวนตัวอย่าง 37 รายการ เพิ่มเป็น 94 รายการ เพิ่มขึ้น 57 รายการ ดังแสดงไว้ใน Table 1.

Table 1. Comparison of sample sizes before and after applying SMOTE.

Class	Before	Increased	After
Programmer (P)	113	-	113
System Tester (T)	50	44	94
UX/UI Designer (U)	37	57	94
Total	**200**	**101**	**301**

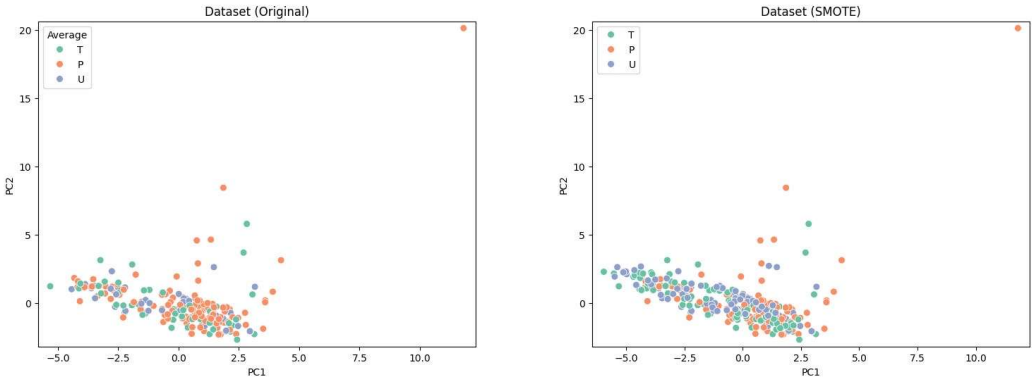
จาก Table 1. จะเห็นว่ามีการเพิ่มข้อมูลทั้งหมดเป็น 301 รายการ การใช้ SMOTE ช่วยเพิ่มความหลากหลายของข้อมูลตัวอย่างในกลุ่มข้อมูลส่วนน้อย ซึ่งช่วยแก้ปัญหา Overfitting ที่โมเดลทำการเรียนรู้ข้อมูลกลุ่มมากมากเกินไปทำให้เกิดความแม่นยำกับข้อมูลเหล่านั้นมากเกินไป ซึ่งอาจเกิดความลำเอียงในการทำนายข้อมูลในอนาคตและยังเป็นการเพิ่มความสามารถของโมเดลการเรียนรู้ในการจำแนกข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น โดยเฉพาะในกรณีที่มีจำนวนข้อมูลจำกัด

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

SMOTE จึงเป็นวิธีที่มีประสิทธิภาพ ซึ่งเหมาะสมสำหรับการเพิ่มคุณภาพของข้อมูลสำหรับการเรียนรู้ของเครื่องในกรณีที่ข้อมูลมีปริมาณน้อยและขาดความสมดุล แสดงผลการเปรียบเทียบข้อมูล ดัง Figure 3.

5.5 สร้างโมเดลการจำแนก

สำหรับขั้นตอนนี้ได้นำข้อมูลที่ผ่านขั้นตอนการจัดเตรียมข้อมูลเข้าไปให้เครื่องเรียนรู้แล้วทำการเปรียบเทียบประสิทธิภาพโมเดลสำหรับเรียนรู้ข้อมูล ซึ่งได้แก่ Decision Tree, Random Forest และ K-NN โดยทำการเรียนรู้ในข้อมูลก่อนและหลังเพิ่มข้อมูลตัวอย่างด้วย SMOTE



(a) (b)

Figure 3. Comparing dataset used for Machine Learning before (a) and after (b) sample augmentation.

จาก Figure 3. พบว่า Figure 3(a) เป็นข้อมูลก่อนเพิ่มข้อมูลตัวอย่าง และ Figure 3(b) เป็นข้อมูลหลังเพิ่มข้อมูลตัวอย่าง

6. ผลการวิจัย (Results)

6.1 ผลการพัฒนาโมเดลการเรียนรู้แบบมีผู้สอนที่สามารถพยากรณ์ตำแหน่งงานที่เหมาะสมสำหรับผู้สมัครโดยพิจารณาข้อมูลที่ระบุใน Resume ของผู้สมัครที่มีข้อมูลไม่สมดุล

ผลการวิจัย ดังนี้ อัลกอริทึม Decision Tree ซึ่งกำหนด Parameter ที่สำคัญได้แก่ ฟังก์ชันสำหรับวัดคุณภาพการแยก (Criterion= ‘gini’) ค่าความลึกของต้นไม้ (max_depth = None) โดยมีค่าความถูกต้อง เท่ากับ 0.83 อัลกอริทึม Random Forest โดยกำหนดค่า default parameter ของ Random Forest โดยมีจำนวนต้นไม้เท่ากับ 10 ซึ่งมีค่าความถูกต้อง เท่ากับ 0.86 และ อัลกอริทึม K-NN ซึ่งกำหนดจำนวนเพื่อนบ้าน (ค่า K) เท่ากับ 3 โดยมีค่าความถูกต้อง เท่ากับ 0.80 ตามลำดับ โดยวัดประสิทธิภาพมุมมองค่าความถูกต้อง (Accuracy) ด้วย Confusion Matrix ดัง Table 2.

Table 2. Results of learning from data before sample augmentation.

Algorithms	Accuracy	Precision	Recall	F1 score
Decision Tree	0.83	0.83	0.86	0.88
Random Forest	0.86	0.86	0.90	0.86
K-NN	0.80	0.86	0.86	0.83

จากนั้นได้นำค่าความถูกต้องสำหรับทั้ง 3 อัลกอริทึม โดยมีการเพิ่มข้อมูลตัวอย่างด้วยวิธี SMOTE ได้ผลลัพธ์และกำหนดค่า parameter ของแต่ละอัลกอริทึมมตามการทดลองก่อนมีการเพิ่มข้อมูลตัวอย่างด้วย SMOTE ดังนี้ อัลกอริทึม Decision Tree มีค่าความถูกต้อง เท่ากับ 0.84 อัลกอริทึม Random Forest มีค่าความถูกต้อง เท่ากับ 0.88 และอัลกอริทึม K-NN มีค่าความถูกต้อง เท่ากับ 0.82 สำหรับ Table 3.

Citation: Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

Table 3. Results of learning from data after sample augmentation.

Algorithms	Accuracy	Precision	Recall	F1 score
Decision Tree	0.84	0.84	0.83	0.85
Random Forest	0.88	0.88	0.92	0.87
K-NN	0.82	0.82	0.84	0.84

นอกจากนั้นในการวิจัยครั้งนี้ยังได้ทำการทดลองสร้างโมเดลการจำแนกกับชุดข้อมูลที่มีการเพิ่มด้วย SMOTE ออกแบบ 2 แบบ คือ แบบที่ 1 ใช้ข้อมูลที่เพิ่มด้วย SMOTE ทั้งหมดอยู่ในข้อมูลฝึก (Training Set) แบบที่ 2 ใช้ข้อมูลที่เพิ่มด้วย SMOTE ทั้งบางส่วน (ร้อยละ 50) อยู่ในข้อมูลฝึก (Training Set) และบางส่วน ร้อยละ 50) ใช้เป็นข้อมูลทดสอบ (Test Set) แล้วนำไปเรียนรู้ด้วย Random Forest ซึ่งเป็นอัลกอริทึมที่มีประสิทธิภาพที่สุดจากทั้ง 3 อัลกอริทึมที่ได้ใช้เป็นการทดลองในวิจัยนี้ ผลที่ได้พบว่า แบบที่ 1 ค่าความถูกต้องในการจำแนกเป็น 0.87 และแบบที่ 2 ความถูกต้องในการจำแนกเป็น 0.86 โดยมีประสิทธิภาพแตกต่างกันเพียงเล็กน้อย

6.2 ผลการเปรียบเทียบประสิทธิภาพโมเดลการเรียนรู้แบบมีผู้สอนของเครื่องเรียนรู้ด้วย 3 โมเดล ได้แก่ Decision Tree, Rainforest และ K-NN ในแง่ค่าความถูกต้อง (Accuracy) สำหรับข้อมูลก่อนและหลังสมดุลข้อมูล

การวิจัยนี้ได้ทำการประเมินประสิทธิภาพของโมเดล Decision Tree, Random Forest และ K-NN โดยเปรียบเทียบค่าความถูกต้อง (Accuracy) ของโมเดลก่อนและหลังการเพิ่มข้อมูลตัวอย่างด้วยเทคนิค SMOTE ผลการทดลองพบว่าโมเดล Random Forest มีค่าความถูกต้องสูงสุดทั้งก่อนและหลังการเพิ่มข้อมูลตัวอย่าง โดยในชุดข้อมูลเดิม ค่าความถูกต้องของ Decision Tree, Random Forest และ K-NN เท่ากับ 0.83, 0.86 และ 0.80 ตามลำดับ (แสดงใน Table 2.) ส่วนในชุดข้อมูลที่ผ่านการเพิ่มข้อมูลตัวอย่างด้วย SMOTE ค่าความถูกต้องเพิ่มขึ้นเป็น 0.84, 0.88 และ 0.82 ตามลำดับ (แสดงใน Table 3.)

ผลการทดลองออกแบบการใช้ข้อมูลที่เพิ่มด้วย SMOTE ในสองรูปแบบ โดยใช้โมเดล Random Forest ซึ่งมีประสิทธิภาพดีที่สุดในการทดลองครั้งนี้ แบบที่ 1 ใช้ข้อมูลที่เพิ่มด้วย SMOTE ทั้งหมดในชุดข้อมูลฝึก (Training Set) ให้ค่าความถูกต้องเท่ากับ 0.87 ขณะที่แบบที่ 2 ใช้ข้อมูลที่เพิ่มด้วย SMOTE ร้อยละ 50 ในชุดข้อมูลฝึกและร้อยละ 50 ในชุดข้อมูลทดสอบ (Test Set) ให้ค่าความถูกต้องเท่ากับ 0.86 ผลลัพธ์ดังกล่าวแสดงให้เห็นว่าการเพิ่มข้อมูลตัวอย่างด้วย SMOTE ช่วยเพิ่มประสิทธิภาพของโมเดลการเรียนรู้ โดยเฉพาะอย่างยิ่งในบริบทของข้อมูลที่ไม่สมดุล

7. สรุปและอภิปรายผลการวิจัย (Conclusion and Discussion)

การวิจัยนี้ได้ทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึม Decision Tree, Random Forest และ K-NN บนชุดข้อมูลที่ไม่สมดุลทั้งก่อนและหลังการเพิ่มข้อมูลตัวอย่างด้วยเทคนิค SMOTE ผลการทดลองแสดงให้เห็นว่า การเพิ่มข้อมูลตัวอย่างด้วย SMOTE ช่วยเพิ่มค่าความถูกต้องของโมเดลทั้งสาม โดยเฉพาะโมเดล Random Forest ที่มีค่าความถูกต้องสูงสุดในทั้งสองกรณี (0.86 ก่อนเพิ่มข้อมูลตัวอย่าง และ 0.88 หลังเพิ่มข้อมูลตัวอย่าง) สะท้อนถึงศักยภาพของ SMOTE ในการแก้ปัญหาความไม่สมดุลของข้อมูล และเป็นงานวิจัยแรกที่ทดสอบ SMOTE กับข้อมูลเรซูเม่ภาษาไทย

นอกจากนี้ การทดลองออกแบบการใช้ข้อมูลที่เพิ่มด้วย SMOTE ในสองรูปแบบ โดยเน้นที่ Random Forest แสดงให้เห็นว่าการใช้ข้อมูลที่เพิ่มด้วย SMOTE ทั้งหมดในชุดฝึก (Training Set) ให้ค่าความถูกต้องสูงกว่าเล็กน้อย (0.87) เมื่อเปรียบเทียบกับการใช้ข้อมูลที่เพิ่มบางส่วนในชุดฝึกและทดสอบ (0.86) ความแตกต่างดังกล่าวอาจเกิดจากผลกระทบของการกระจายข้อมูลในชุดทดสอบที่ไม่สมดุล อย่างไรก็ตาม ทั้งสองวิธีมีประสิทธิภาพใกล้เคียงกันและเหมาะสมสำหรับบริบทที่ต้องการปรับปรุงความแม่นยำของโมเดลในโมเดลข้อมูลที่ไม่สมบูรณ์ นอกจากนี้พบแนวทางการการเพิ่มประสิทธิภาพ ดังนี้ 1) การทดลองกับชุดข้อมูลหลากหลาย ควรทดสอบเทคนิค SMOTE กับชุดข้อมูลประเภทอื่นที่มีลักษณะและระดับความไม่สมดุลที่แตกต่างกัน เพื่อประเมินความทั่วไปของผลลัพธ์ 2) เปรียบเทียบกับเทคนิคอื่น ควรนำเทคนิคเพิ่มข้อมูลตัวอย่างอื่น เช่น ADASYN หรือ BORDERLINE-SMOTE มาทดลองเปรียบเทียบกับ SMOTE เพื่อค้นหาวิธีการที่เหมาะสมที่สุดสำหรับบริบทที่แตกต่างกัน 3) วิเคราะห์ผลกระทบของพารามิเตอร์ SMOTE การวิเคราะห์ผลกระทบของจำนวนตัวอย่างที่เพิ่มและระยะห่างของข้อมูลในการคำนวณอาจช่วยปรับปรุงประสิทธิภาพของโมเดลเพิ่มเติม 4) การใช้วิธีการอื่นร่วมกัน ควรพิจารณาการใช้เทคนิคการเลือกคุณลักษณะหรือการสร้างคุณลักษณะร่วมกับ SMOTE เพื่อเพิ่มความแม่นยำของโมเดลมาก

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

ยิ่งขึ้น 5) ข้อมูลเพียง 200 ตัวอย่าง อาจไม่เพียงพอสำหรับการสรุปทั่วไป และ 6) การศึกษานี้ใช้ข้อมูลเรซูเม่เฉพาะ 3 ตำแหน่งงาน อาจไม่ครอบคลุมอุตสาหกรรมอื่น

8. ข้อเสนอแนะงานวิจัย (Recommendation)

จากผลการวิจัยพบว่าเทคนิค SMOTE ช่วยเพิ่มประสิทธิภาพของโมเดลการเรียนรู้ โดยเฉพาะในกรณีของข้อมูลไม่สมดุล ซึ่งโมเดล Random Forest ให้ค่าความถูกต้องสูงสุดทั้งก่อนและหลังการเพิ่มข้อมูลตัวอย่าง ดังนั้นข้อเสนอแนะสำหรับการวิจัยในอนาคต ดังนี้

1. การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) ควรศึกษาการปรับแต่งพารามิเตอร์ของโมเดล Random Forest เพิ่มเติม เพื่อเพิ่มประสิทธิภาพและความแม่นยำของโมเดลให้ดียิ่งขึ้น

2. การทดลองกับชุดข้อมูลที่หลากหลาย ควรนำเทคนิค SMOTE ไปประยุกต์ใช้กับชุดข้อมูลที่มีความไม่สมดุลในบริบทอื่น เช่น การตรวจจับการทุจริตทางการเงินหรือการวินิจฉัยโรค

3. การประเมินโมเดลด้วย Metric อื่น นอกเหนือจากค่าความถูกต้อง (Accuracy) ควรพิจารณาค่า F1-Score, Precision และ Recall เพื่อตรวจสอบประสิทธิภาพในเชิงลึก

4. การเพิ่มประสิทธิภาพของการแบ่งข้อมูล การใช้ SMOTE ในสัดส่วนที่เหมาะสมระหว่างชุดข้อมูลฝึกและชุดข้อมูลทดสอบ เช่น การเพิ่มข้อมูลใน Training Set ในสัดส่วนที่เหมาะสมเพื่อหลีกเลี่ยงปัญหาการ Overfitting

5. เสนอแนวทางปรับใช้ SMOTE ร่วมกับโมเดล Random Forest เพื่อเพิ่มประสิทธิภาพระบบคัดเลือกบุคคล ซึ่งจะทำให้การวิจัยในอนาคตจะสามารถพัฒนาระบบการเรียนรู้ที่แม่นยำและเสถียรยิ่งขึ้นในบริบทของข้อมูลไม่สมดุล

9. กิตติกรรมประกาศ

ขอขอบพระคุณและขอแสดงความนับถือสำหรับการสนับสนุนทุนวิจัยจาก คณะวิทยาศาสตร์ สาขาวิชาวิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ รวมทั้งสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏจันทรเกษม (CRURDI) ที่ให้การสนับสนุน สถานที่ เครื่องมือ และความอนุเคราะห์ในทุกด้านจนส่งให้งานวิจัยชิ้นนี้สำเร็จด้วยดีตลอดมา

10. เอกสารอ้างอิง (References)

Akkharatwiwatthanathorn S. (2020*). Artificial Intelligence Innovation with Recruitment*. [Master's dissertation, Mahidol University]. CMMU Digital Archive. https://archive.cm.mahidol.ac.th/handle/123456789/3777. (In Thai)

Ali, I., Mughal, N., Khan, Z. H., Ahmed, J., & Mujtaba, G. (2022). Resume Classification System using Natural Language Processing and Machine Learning Techniques. *Mehran University Research Journal of Engineering and Technology, 41*(1), 65–79. https://doi.org/10.22581/muet1982.2201.07.

Aubaidan, B. H., Kadir, R. A., & Ijab, M. T. (2024). A Comparative Analysis of Smote and CSSF Techniques for Diabetes Classification Using Imbalanced Data. *Journal of Computer Science, 20*(9), 1146–1165. https://doi.org/10.3844/jcssp.2024.1146.1165

Boonchob, T. (2022). *Job-candidate Classifying and Ranking System with Machine Learning Method*. [Master's dissertation, Chulalongkorn University]. Chula Digital Collections. https://digital.car.chula.ac.th/chulaetd/5817. (In Thai)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953.

Cheng, D., Zhang, S., Deng, Z., Zhu, Y., & Zong, M. (2014). kNN Algorithm with Data-Driven k Value. In Luo, X., Yu, J. X., & Li, Z. (Eds), Advanced Data Mining and Application (8933, 499–512). Springer International Publishing. https://doi.org/10.1007/978-3-319-14717-8_39.

Citation: Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.

Gehrke, J., Ramakrishnan, R., & Ganti, V. (2000). RainForest-A Framework for Fast Decision Tree Construction of Large Datasets. *Data Mining and Knowledge Discovery*, *4*(2/3), 127–162. https://doi.org/10.1023/A:1009839829793.

Maurya, L. S., Hussain, M. S., & Singh, S. (2022). Machine Learning Classification Models for Student Placement Prediction Based on Skills. International Journal of Artificial Intelligence and Soft Computing, 7(3), 194–207. https://doi.org/10.1504/IJAISC.2022.10051214.

Modak, S., Shinde, P., Tiwari, A., & Nalamwar, S. (2024). A Review of Resume Analysis and Job Description Matching Using Machine Learning. *International Journal on Recent and Innovation Trends in Computing and Communication*, *12*(2), 247–250.

Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm for Classification. *International Journal of Engineering Development and Research, 2*(1), 1–5.

Patil, P. R. (2023). Resume Classification-based on Personality using Machine Learning Algorithm. *International Journal of Scientific and Research Publications, 13*(2), 335–341. https://doi.org/10.29322/IJSRP.13.02.2023.p13440.

Posri, N. (2024). Machine Learning-Based Multiclass Classification for Predicting the Cumulative Blood Sugar Levels in Type 2 Diabetes Patient [Master's dissertation, Sukhothai Thammathirat Open University]. STOUIR at Sukhothai Thammathirat Open University. https://ir.stou.ac.th/handle/123456789/13029. (In Thai)

Rezki, M. K., Mazdadi, M. I., Indriani, F., Muliadi, M., Saragih, T. H., & Athavale, V. A. (2024). Application Of SMOTE To Address Class Imbalance in Diabetes Disease Classification Utilizing C5.0, Random Forest, And SVM. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(4), 343–354. https://doi.org/10.35882/jeeemi.v6i4.434.

Sanitphonklang, C. (2023). The Identification of Root and Stem Rot Disease of Durian Trees by Algorithm Decision Tree. *Journal of Science Ladkrabang, 32*(1), 72–87. (In Thai)

Simarmata, J. E., Weber, G.-W., & Chrisinta, D. (2024). Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines. *Jurnal Matematika, Statistika Dan Komputasi*, *20*(3), 623–638. https://doi.org/10.20956/j.v20i3.32970.

Sun, S., Zhou, X., Wei, J., Xiao, Y., & Wang, J. (2023, November 17-19). An Optimization of SMOTE for Anomaly Detection Based on High Contribution Sample Screening. *2023 China Automation Congress (CAC)*, 2010–2014. https://doi.org/10.1109/CAC59555.2023.10451412.

Valen-Dacanay, J. G., & Palaoag, T. D. (2023, March 18-20). Exploring The Learning Analytics of Skill-Based Course Using Machine Learning Classification Models. *2023 11th International Conference on Information and Education Technology (ICIET)*, 411–415. https://doi.org/10.1109/ICIET56899.2023.10111210.

Citation:
Vasuarayasak, W., Mangklang, N., Somseang, J., & Sanitphonklang, C. (2025). Enhancing Machine Learning Performance on Imbalanced Resume Data Using SMOTE for Job Candidate Classification. *Journal of Computer and Creative Technology, 3*(1), 38-48. https://doi.org/10.14456/jcct.2025.4.