การพัฒนาระบบวิเคราะห์ภาวะเสี่ยงต่อการเกิดครรภ์เป็นพิษด้วยปัญญาประดิษฐ์

นางสาวปอรรัชม์ ปานใจนาม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมโทรคมนาคมและคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2567

AN ARTIFICIAL INTELLIGENCE SYSTEM FOR THE RISK ANALYSIS OF PRE-ECLAMPSIA

PAONRAT PANJAINAM

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Telecommunication and Computer Engineering
Suranaree University of Technology
Academic Year 2024

การพัฒนาระบบวิเคราะห์ภาวะเสี่ยงต่อการเกิดครรภ์เป็นพิษด้วยปัญญาประดิษฐ์

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

.....

(ผศ. ดร.นันทวุฒิ คะอังกุ)

ประธานกรรมการ

.....

(ผศ. ดร.ศรัญญา กาญจนวัฒนา)

กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

.....

(อ. ดร.รติพร จันทร์กลั่น)

กรรมการ

.....

(รศ. ดร.ยุพาพร รักสกุลพิวัฒน์)

รองอธิการบดีฝ่ายวิชาการและประกันคุณภาพ

.....

(รศ. ดร.พรศิริ จงกล)

คณบดีสำนักวิชาวิศวกรรมศาสตร์

ปอรรัชม์ ปานใจนาม: การพัฒนาระบบวิเคราะห์ภาวะเสี่ยงต่อการเกิดครรภ์เป็นพิษด้วยปัญญาประดิษฐ์ (AN ARTIFICIAL INTELLIGENCE SYSTEM FOR THE RISK ANALYSIS OF PRE-ECLAMPSIA)
อาจารย์ที่ปรึกษา : ผศ. ดร. ศรัญญา กาญจนวัฒนา, 53 หน้า.

คำสำคัญ: ภาวะครรภ์เป็นพิษ/ข้อมูลไม่สมดุล/การสุ่มตัวอย่างข้อมูล/การเรียนรู้ของเครื่อง

งานวิจัยนี้ศึกษาภาวะครรภ์เป็นพิษ ที่เป็นภาวะแทรกซ้อนที่สำคัญในหญิงตั้งครรภ์ซึ่งมีความสัมพันธ์กับความดันโลหิตสูง ส่งผลกระทบต่อสุขภาพของมารดาและทารกทั่วโลก การวินิจฉัยและการคัดกรองกลุ่มเสี่ยงยังมีข้อจำกัด โดยเฉพาะข้อมูลทางคลินิกที่มีความไม่สมดุลและซับซ้อน งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยเสี่ยงของภาวะครรภ์เป็นพิษ พัฒนาเทคนิคจัดการข้อมูลไม่สมดุลด้วยการสุ่มเพิ่มลดตัวอย่างข้อมูลแบบผสมผสาน และสร้างแบบจำลองปัญญาประดิษฐ์สำหรับทำนายความเสี่ยงของภาวะดังกล่าว โดยใช้ชุดข้อมูลเวชระเบียนจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี ข้อมูลถูกจัดกลุ่มเป็นปกติ เสี่ยง และครรภ์เป็นพิษ มีการปรับสมดุลข้อมูลด้วยเทคนิคการสุ่มลดตัวอย่างข้อมูลด้วย 4 วิธี คือ RUS Tomek Links ENN และ RENN ร่วมกับการสุ่มตัวอย่างข้อมูลด้วย 4 วิธี คือ ROS ADASYN SMOTE และ Borderline-SMOTE ทำให้เทคนิคการสุ่มเพิ่มลดตัวอย่างข้อมูลมี 16 วิธี จากนั้นพัฒนาแบบจำลองด้วยอัลกอริธึม ได้แก่ Decision Tree XGBoost Random Forest Naive Bayes Neural Network และ K-Nearest Neighbors ผลการทดลองพบว่าแบบจำลอง XGBoost ร่วมกับเทคนิค Tomek Links และ ROS มีประสิทธิภาพสูงสุด โดยมีค่า F1-Score เท่ากับ 0.9864 และ AUC เท่ากับ 1.0000

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา.....ปอรรัชม์.....

ปีการศึกษา 2567 ลายมือชื่ออาจารย์ที่ปรึกษา.....

PAONRAT PANJAIMNAM: AN ARTIFICIAL INTELLIGENCE SYSTEM FOR THE RISK ANALYSIS OF PRE-ECLAMPSIA

THESIS ADVISOR: ASSISTANT PROFESSOR SARUNYA KANJANAWATTANA , Ph.D. 53 PP.

Keywords: Preeclampsia/Imbalanced Data/Data Resampling/Machine Learning

This study investigates preeclampsia, a significant pregnancy complication associated with hypertension that adversely affects the health of both mothers and infants worldwide. Current diagnosis and risk screening face limitations, especially due to imbalanced and complex clinical data. The study aims to analyze risk factors for preeclampsia, develop hybrid resampling techniques to manage data imbalance, and build an artificial intelligence model to predict the risk of this condition. Medical record data from Suranaree University of Technology Hospital were categorized into normal, at-risk, and preeclampsia groups. Data balancing was achieved by combining four undersampling methods such as RUS, Tomek Links, ENN, and RENN with four oversampling methods such as ROS, ADASYN, SMOTE, and Borderline-SMOTE, resulting in 16 hybrid resampling techniques. Subsequently, predictive models were developed using algorithms including Decision Tree, XGBoost, Random Forest, Naive Bayes, Neural Network, and K-Nearest Neighbors. Experimental results showed that the XGBoost model combined with Tomek Links and ROS achieved the highest performance, with an F1-Score of 0.9864 and an AUC of 1.0000.

School of Computer Engineering Student's Signature.....

Academic year 2024 Advisor's Signature.....

กิตติกรรมประกาศ

วิทยานิพนธนี้สำเร็จลุลวงดวยดีเนื่องจากไดรับความชวยเหลืออยางดียิ่ง ทั้งดานวิชาการและดานการดำเนินงานวิจัยจากบุคคลและกลุมบุคคลตาง ๆ ผู้เขียนขอแสดงความขอบคุณเป็นอย่างสูงสำหรับคำแนะนำ การสนับสนุน และการชี้แนะจาก ผศ. ดร.ศรัญญา กาญจนวัฒนา ตลอดระยะเวลาการดำเนินงานวิจัยนี้ ขอขอบคุณเป็นพิเศษสำหรับ รองศาสตราจารย์ ผศ. ดร.นันทวุฒิ คะอังกุ และ ดร.รติพร จันทร์กลั่นที่ได้เข้าร่วมเป็นกรรมการสอบวิทยานิพนธ์ฉบับนี้

ขอขอบคุณเป็นพิเศษไปยังอาสาสมัครทุกท่านที่เข้าร่วมในการทดลองนี้ ความทุ่มเทและการปฏิบัติงานของพวกเขามีความสำคัญอย่างยิ่งในการสนับสนุนงานวิจัยนี้

ยิ่งไปกว่านั้น ผู้เขียนขอขอบคุณอย่างยิ่งสำหรับการสนับสนุนที่ได้รับจากทุน OROG Scholarships (External Grants and Scholarships for Graduate Students) ซึ่งเป็นทุนที่สำคัญอย่างยิ่งต่อการสำเร็จการศึกษาครั้งนี้

ปอรรัชม์ ปานใจนาม

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหางานวิจัย

ในปัจจุบันภาวะครรภ์เป็นพิษ (Preeclampsia) ถือเป็นโรคเฉพาะชนิดหนึ่งที่เกิดขึ้นในระหว่างการตั้งครรภ์และมีความเกี่ยวข้องการเกิดความดันโลหิตสูง (Hypertension) ขณะตั้งครรภ์ส่งผลให้เกิดความอันตรายถึงแก่ชีวิตแก่สตรีที่ตั้งครรภ์และทารกในครรภ์ อาการเริ่มต้นของภาวะนี้ประกอบไปด้วยการมีความดันโลหิตสูงร่วมกับการมีโปรตีนรั่วในปัสสาวะ ซึ่งถือเป็นสัญญาณเตือนที่ต้องได้รับการดูแลอย่างใกล้ชิด กรณีที่เข้าสู่ภาวะรุนแรงอาจจะนำไปสู่การชัก ซึ่งมีความสัมพันธ์โดยตรงกับระดับความดันโลหิตสูงในขณะตั้งครรภ์

ภาวะครรภ์เป็นพิษส่งผลกระทบต่อสุขภาพของหญิงตั้งครรภ์และทารกในครรภ์อย่างมาก โดยจะเกิดขึ้นหลังจากการตั้งครรภ์ครบ 20 สัปดาห์ พร้อมทั้งยังไม่เคยมีประวัติความดันโลหิตสูงมาก่อน ภาวะนี้ถูกบ่งชี้โดยการตรวจพบความดันโลหิตสูงทั้งค่าตัวบน (Systolic Blood Pressure: SBP) และตัวล่าง (Diastolic Blood Pressure: DBP) ความดันโลหิตตัวบนมีค่ามากกว่า140 มิลลิเมตรปรอทหรือพบว่ามีความดันโลหิตตัวล่างมากกว่า 90 มิลลิเมตรปรอท และพบโปรตีนที่พบในปัสสาวะ ข้อบ่งชี้เหล่านี้สามารถมีข้อใดข้อหนึ่งที่กล่าวมาก็สามารถบ่งชี้ได้ว่ามีภาวะครรภ์เป็นพิษขณะตั้งครรภ์ (รศ.พญ.เกษมศรี ศรีสุพรรณดิฐ, ม.ป.ป.) ภาวะนี้อาจเกิดขึ้นช้าถึง 4-6 สัปดาห์หลังคลอด (Lim, 2022) และยังไม่ทราบสาเหตุที่แน่ชัด ทั่วโลกมีหญิงตั้งครรภ์จำนวนมากที่เสี่ยงต่อภาวะครรภ์เป็นพิษ องค์การอนามัยโลกได้มีประกาศออกมาชัดเจนว่าผู้หญิงที่เสียชีวิตจากภาวะครรภ์เป็นพิษมีผลกระทบต่อการตั้งครรภ์ 2-8 % ทั่วโลก มีผู้หญิงประมาณ 76,000 ราย และทารก 500,000 ราย (World Health Organization, 2025) เสียชีวิตทุกปีเนื่องจากภาวะครรภ์เป็นพิษ ภาวะครรภ์เป็นพิษเป็นสาเหตุของการเสียชีวิตของมารดาประมาณร้อยละ 10 ในเอเชียและแอฟริกา ร้อยละ 25 ในละตินอเมริกา มีผลกระทบประมาณ 8-10 % ของการตั้งครรภ์ทั้งหมดในทวีปแอฟริกา (Vata et al., 2015) ข้อมูลการเสียชีวิตของมารดาในไทย 10 ปี ย้อนหลังจากการตายทั้งหมด 1,125 ราย มีการตายด้วยสาเหตุภาวะความดันโลหิตสูงในขณะตั้งครรภ์ การคลอด และระยะหลังคลอด 103 ราย คิดเป็น ร้อยละ 9 จากสาเหตุการเสียชีวิตทั้งหมด (กรมอนามัย กระทรวงสาธารณสุข, 2025) นอกจากนี้ในการศึกษาภาวะครรภ์เป็นพิษชนิดรุนแรงในสตรีตั้งครรภ์ในโรงพยาบาลมหาราชนครราชสีมามีมากถึงร้อยละ 7 ของการตั้งครรภ์ทั้งหมดในกลุ่มประชากรที่ศึกษา (สิรยา กิติโยดม, 2014)

วิธีสำหรับการป้องกันและลดความรุนแรงของภาวะครรภ์เป็นพิษมีหลากหลายวิธี โดยจำแนกกลุ่มเสี่ยงที่ควรได้รับการติดตามอย่างใกล้ชิดระหว่างตั้งครรภ์เป็นวิธีการหนึ่งที่มีประสิทธิภาพสูงเพื่อป้องกันการเกิดภาวะนี้ได้ทันท่วงที ในกลุ่มผู้ป่วยที่ถูกว่าเป็นกลุ่มเสี่ยงจะพบว่ามีอาการดังต่อไปนี้ ความดันโลหิตสูงและพบโปรตีนในปัสสาวะ อาการจุกแน่นลิ้นปี่ อาการตาพร่ามัว อาการปวดศรีษะ และมีอาการตัวบวม ซึ่งเกิดในกลุ่มสตรีผู้ที่เป็นการตั้งครรภ์ครั้งแรก ผู้มีอายุมากกว่า 40 ปี หรือผู้ที่เคยมีประวัติครรภ์เป็นพิษ หรือมีโรคประจำตัว เช่น โรคความดันโลหิตสูง โรคไต โรคเบาหวาน โรคทางภูมิคุ้มกัน โรคอ้วน น้ำหนักตัวเพิ่มผิดปกติจากเกณฑ์การเพิ่มน้ำหนักของสตรีตั้งครรภ์ อาการอ่อนเพลีย เมื่อพบอาการใดอาการหนึ่งจากกลุ่มครรภ์ปกติจะถูกจัดเป็นกลุ่มเสี่ยงทันที

วิธีการฝากครรภ์ตามแนวทางปกติเริ่มจากการตรวจร่างกาย สอบถามประวัติสุขภาพ และตรวจทางห้องปฏิบัติการในช่วงอายุครรภ์ก่อน 12 สัปดาห์ พร้อมติดตามอาการในแต่ละระยะผ่านการนัดตรวจและบันทึกลงในสมุดสีชมพู แม้ระบบนี้จะช่วยคัดกรองความเสี่ยงเบื้องต้นได้ แต่ยังมีข้อจำกัดในด้านความรวดเร็ว โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีความซับซ้อนหรืออาการยังไม่ชัดเจน ซึ่งอาจทำให้การวินิจฉัยกลุ่มเสี่ยงล่าช้าและเพิ่มโอกาสเกิดภาวะแทรกซ้อน เทคโนโลยีปัญญาประดิษฐ์จึงเข้ามามีบทบาทสำคัญในการเร่งการประมวลผลข้อมูลจำนวนมาก พร้อมให้ผลวิเคราะห์ความเสี่ยงแบบเรียลไทม์ภายในไม่กี่วินาที ช่วยให้บุคลากรทางการแพทย์สามารถตัดสินใจได้อย่างรวดเร็วและแม่นยำยิ่งขึ้น ลดข้อจำกัดของระบบคัดกรองแบบเดิมที่ต้องอาศัยเวลาและทรัพยากรจำนวนมาก

ในปัจจุบันมีการนำเทคโนโลยีทางปัญญาประดิษฐ์มาปรับใช้ในการป้องกันโรคต่าง ๆ เพิ่มขึ้น ในปัจจุบัน มีการนำมาใช้เพื่อลดการเกิดความรุนแรงของภาวะครรภ์เป็นพิษ แนวคิดในการป้องกันภาวะครรภ์เป็นพิษในหญิงตั้งครรภ์จึงเป็นสิ่งจำเป็น วิธีการป้องกันดังกล่าวจะต้องอาศัยข้อมูลที่หลากหลายจากสตรีที่ตั้งครรภ์เพื่อทำการวิเคราะห์ ซึ่งการค้นหาปัจจัยที่ส่งผลต่อการเกิดภาวะดังกล่าวสามารถนำไปสู่การจำแนกกลุ่มเสี่ยงและกำหนดวิธีการป้องกันและเฝ้าระวังที่เหมาะสม (De Kat et al., 2019) นอกจากนี้ ในปัจจุบันเทคโนโลยีด้านปัญญาประดิษฐ์ยังมีบทบาทสำคัญในการแก้ปัญหาต่าง ๆ ในด้านของวิทยาศาสตร์การแพทย์มากขึ้น สนับสนุนการตัดสินใจและการวางแผนการรักษาอย่างมีประสิทธิภาพ (Briceño-Pérez et al., 2009)

การพัฒนาปัญญาประดิษฐ์ให้ประสิทธิภาพที่ดีได้นั้นต้องอาศัยข้อมูลเพื่อใช้ในการเรียนรู้ที่เหมาะสม ความแม่นยำและความถูกต้องในการจำแนกกลุ่มข้อมูลส่วนน้อยในงานด้านการเรียนรู้ของเครื่อง (Machine Learning) การกระจายตัวของกลุ่มข้อมูลที่ไม่สมดุลทำให้เกิดความท้าทายในการแก้ปัญหาข้อมูลไม่สมดุล (Fotouhi et al., 2019) การกระจายตัวของข้อมูลที่ไม่สมดุลนี้เป็นความท้าทายอย่างมากในด้านการเรียนรู้ของเครื่อง เนื่องจากแบบจำลองเรียนรู้จะมีคุณภาพได้จำเป็นต้องมีการเรียนรู้กับข้อมูลจำนวนมากเพื่อเข้าใจรูปแบบของข้อมูล แต่สำหรับชุดข้อมูลส่วนน้อย

แบบจำลองอาจจะไม่มีความสามารถที่เพียงพอต่อการเรียนรู้และจำแนกข้อมูล ซึ่งมีผลทำให้แบบจำลองมีประสิทธิภาพการทำนายไปทิศทางเดียวกับกลุ่มข้อมูลส่วนมากที่มีจำนวนข้อมูลที่มากกว่า (Li et al., 2014) ข้อมูลในโลกความเป็นจริงมักพบปัญหาความไม่สมดุลของข้อมูลเสมอ ในชุดข้อมูลที่ไม่สมดุลนั้นมักเจอกับปัญหาที่ว่าข้อมูลกลุ่มที่สนใจมักมีจำนวนประชากรของข้อมูลน้อยมาก เมื่อเปรียบเทียบกับประชากรข้อมูลกลุ่มอื่น ๆ และอัตราส่วนระหว่างข้อมูลในแต่ละคลาสมีความแตกต่างกันมากเกิดปัญหาช่องว่างระหว่างกลุ่มของข้อมูล ปัญหาเกี่ยวกับข้อมูลไม่สมดุลยังเป็นความท้าทายสำหรับการทำนายชุดข้อมูลส่วนน้อย ทางด้านการแพทย์มักพบปัญหาข้อมูลไม่สมดุล เนื่องจากชุดข้อมูลในงานวิจัยมักจะพิจารณาถึงความผิดปกติของกลุ่มประชากรที่มีปริมาณข้อมูลที่น้อยกว่ากลุ่มประชากรปกติในสัดส่วนข้อมูลที่แตกต่างกันมาก ด้วยปัญหานี้ทำให้การจำแนกประเภทข้อมูลพบกับอุปสรรคในการจำแนกกลุ่มประชากรที่มีความผิดปกติ เพราะข้อมูลส่วนใหญ่เป็นข้อมูลของกลุ่มที่ปกติ (Guo et al., 2008)

ในงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษาและพัฒนาแบบจำลองทางด้านการเรียนรู้ของเครื่อง เพื่อทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษ โดยชุดข้อมูลที่นำมาศึกษาในครั้งนี้เป็นชุดข้อมูลเกี่ยวกับภาวะครรภ์เป็นพิษที่ได้รับจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี ผู้วิจัยได้ทำการศึกษาข้อมูลและมุ่งเน้นหาวิธีการจัดการข้อมูลให้อยู่ในรูปแบบสมดุล โดยเลือกวิธีการที่เหมาะสมกับชุดข้อมูลมากที่สุดและให้ผลลัพธ์ที่ดีที่สุด การวิจัยครั้งนี้ได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ของ โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี เลขที่โครงการ EC-65-121 เมื่อวันที่ 17 พฤศจิกายน พ.ศ.2565 โดยข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลเวชระเบียนที่ไม่สามารถระบุตัวบุคคลได้ (de-identified data) และไม่มีการเปิดเผยข้อมูลส่วนบุคคลของผู้ป่วย ข้อมูลทั้งหมดได้รับการเก็บรักษาอย่างปลอดภัย และนำมาใช้เพื่อวัตถุประสงค์ทางวิชาการเท่านั้น คณะกรรมการจริยธรรมฯ ได้พิจารณาและอนุญาตให้ไม่ต้องขอความยินยอมจากผู้ให้ข้อมูล เนื่องจากเป็นการใช้ข้อมูลทุติยภูมิที่ไม่มีผลกระทบต่อสิทธิ์หรือความเป็นส่วนตัวของผู้ป่วย ข้อมูลนี้มีข้อมูลคลินิกที่ได้รับคำสั่งให้ทำเป็นข้อมูลที่ไม่ระบุชื่อจากผู้เข้าร่วมทดลอง ชุดข้อมูลครรภ์เป็นพิษในงานวิจัยนี้ประกอบด้วยข้อมูลจำแนกออกเป็น 3 กลุ่ม ได้แก่ กลุ่มปกติ กลุ่มเสี่ยง และกลุ่มภาวะครรภ์เป็นพิษ โดยมีจำนวนตัวอย่างในแต่ละกลุ่มคือ 9,341 ตัวอย่างในกลุ่มปกติ 123 ตัวอย่างในกลุ่มเสี่ยง และ 9 ตัวอย่างในกลุ่มภาวะครรภ์เป็นพิษ ซึ่งสะท้อนถึงความไม่สมดุลของข้อมูลอย่างชัดเจน โดยมีอัตราส่วนระหว่างกลุ่มปกติ กับกลุ่มเสี่ยงเท่ากับประมาณ 76:1 และอัตราส่วนระหว่างกลุ่มเสี่ยงกับกลุ่มภาวะครรภ์เป็นพิษเท่ากับประมาณ 14:1 จากอัตราส่วนดังกล่าวจึงสามารถสรุปได้ว่าข้อมูลชุดนี้มีความไม่สมดุลอย่างมีนัยสำคัญ (Khushi et al., 2021)

1.2 วัตถุประสงค์ของการวิจัย

ผู้วิจัยได้ตั้งวัตถุประสงค์ในการวิจัยดังนี้

1.2.1 เพื่อวิเคราะห์ข้อมูลประวัติของสตรีตั้งครรภ์ในการหาความสัมพันธ์ของปัจจัยที่ทำให้การเกิดภาวะครรภ์เป็นพิษได้

1.2.2 เพื่อศึกษาหาแนวทางที่เหมาะสมในการจัดการข้อมูลที่ไม่สมดุลด้วยวิธี Hybrid Resampling เพื่อเตรียมความพร้อมให้กับข้อมูลก่อนสร้างแบบจำลอง

1.2.3 เพื่อพัฒนาแบบจำลองทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษของหญิงมีครรภ์ด้วยปัญญาประดิษฐ์

1.3 ขอบเขตการวิจัย

ในงานวิจัยนี้เป็นการศึกษาและพัฒนาแบบจำลองทำนายความเสี่ยงของการเกิดครรภ์เป็นพิษของหญิงมีครรภ์ด้วยปัญญาประดิษฐ์ เพื่อให้สามารถหาปัจจัยเสี่ยงการเกิดภาวะครรภ์เป็นพิษและจำแนกประเภทความเสี่ยงของผู้ป่วยได้ โดยกำหนดขอบเขตของงานวิจัยดังนี้

1.3.1 ข้อมูลที่นำมาใช้ในการวิจัยนั้นมาจากการเก็บรวบรวมข้อมูลจากผู้ป่วยด้วยข้อมูลจริงจาก โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี ระหว่างปี 2565 ถึง 2566

1.3.2 งานวิจัยนี้ศึกษาจากข้อมูลผู้ป่วยที่เคยได้รับการวินิจฉัยว่าตั้งครรภ์เท่านั้น

1.4 ประโยชน์ที่ได้รับ

ประโยชน์ที่ได้รับหลังจากการศึกษาและพัฒนาวิจัยนี้ ได้แก่

1.4.1 สามารถวิเคราะห์ของข้อมูลของสตรีตั้งครรภ์และหาความสัมพันธ์ของปัจจัยที่นำไปสู่การเกิดภาวะครรภ์เป็นพิษได้

1.4.2 การพัฒนาแบบทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษของหญิงมีครรภ์ด้วยปัญญาประดิษฐ์ โดยสามารถวิเคราะห์ข้อมูลหาปัจจัยสำคัญที่เพิ่มความเสี่ยงของการเกิด ภาวะครรภ์เป็นพิษได้

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

บทนี้กล่าวถึงการทบทวนวรรณกรรมที่เกี่ยวข้องกับงานวิจัยนี้ โดยมีรายละเอียดเกี่ยวกับการเตรียมข้อมูลในรูปแบบเชิงตัวเลขเพื่อนำไปใช้สำหรับการวิเคราะห์และประมวลผล รวมถึงการใช้เทคนิคการสุ่มตัวอย่างรูปแบบผสมผสาน ซึ่งประกอบไปด้วยการสุ่มลดจำนวนตัวอย่างในกลุ่มข้อมูลส่วนใหญ่และการสุ่มเพิ่มจำนวนกลุ่มตัวอย่างในกลุ่มข้อมูลส่วนน้อย เพื่อปรับให้กลุ่มข้อมูลมีความสมดุลกัน นอกจากนี้ยังมีการกล่าวถึงกระบวนการที่ใช้ในการจัดการข้อมูล อัลกอริทึมที่ใช้ในการจำแนกประเภท ตลอดจนวิธีการวัดผลและประเมินประสิทธิภาพของแบบจำลอง โดยงานวิจัยนี้มีการอ้างอิงถึงงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.1 การจัดการเตรียมข้อมูล

ในปัจจุบันความแม่นยำและความถูกต้องในการจำแนกกลุ่มข้อมูลชนกลุ่มน้อยในงานด้านการเรียนรู้ของเครื่อง เป็นประเด็นที่ได้รับความสนใจ เนื่องจากการกระจายตัวของข้อมูลที่ไม่สมดุล ทำให้เกิดความท้าทายในการแก้ปัญหานี้ (Fotouhi et al., 2019) การกระจายตัวที่ไม่สมดุลนี้เป็นอุปสรรคสำคัญในงานด้านการเรียนรู้ของเครื่องอย่างมาก เนื่องจากแบบจำลองมักเรียนรู้จากข้อมูลส่วนใหญ่ที่มีจำนวนมากกว่า ส่งผลให้ขาดความสามารถเพียงพอในการจำแนกข้อมูลกลุ่มน้อย ทั้งนี้ โมเดลจึงมักทำนายผลลัพธ์ไปในทิศทางเดียวกับข้อมูลกลุ่มใหญ่ (Li et al., 2014)

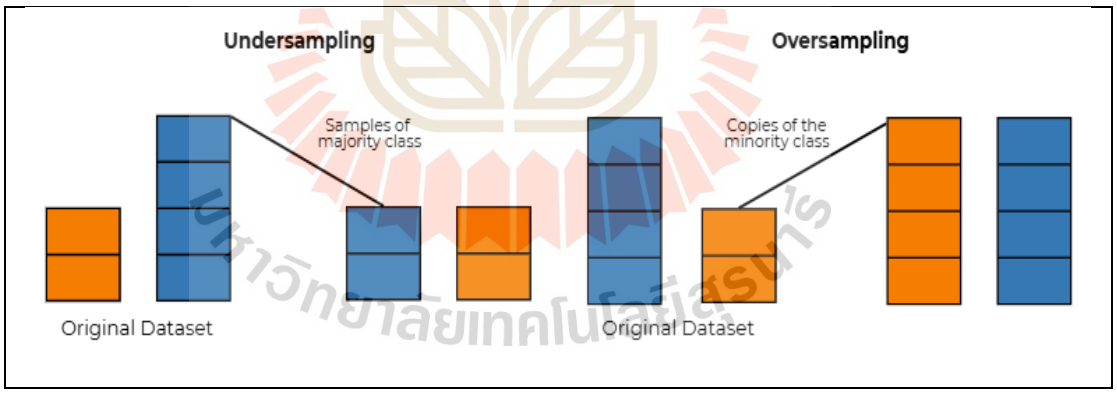
จากการศึกษาพบว่าเพื่อให้สามารถใช้วิธีการสุ่มตัวอย่างได้อย่างมีประสิทธิภาพ จำเป็นต้องแปลงข้อมูลให้เป็นข้อมูลเชิงตัวเลข (Khushi et al., 2021) ก่อนที่จะดำเนินการแก้ปัญหาข้อมูลที่ไม่สมดุล ในงานวิจัยนี้จะใช้วิธีการจัดการข้อมูลในรูปแบบผสมผสาน รูปแบบนี้คือการนำวิธีการจัดการข้อมูลที่ไม่สมดุลมาประยุกต์ใช้ โดยเลือกใช้วิธีการสุ่มลดตัวอย่างกลุ่มข้อมูลร่วมกับการสุ่มเพิ่มตัวอย่างข้อมูล ปกติแล้ววิธีการสุ่มตัวอย่างจะใช้กับชุดข้อมูลที่มีเพียงสองกลุ่ม แต่ชุดข้อมูลในงานวิจัยนี้ประกอบด้วยสามกลุ่ม คือ กลุ่มผู้ป่วยปกติ จำนวน 9,341 ตัวอย่าง จากข้อมูลทั้งหมด กลุ่มผู้ป่วยมีความเสี่ยง จำนวน 123 ตัวอย่าง จากข้อมูลทั้งหมด และกลุ่มผู้ป่วยที่มีภาวะครรภ์เป็นพิษ จำนวน 9 ตัวอย่าง จากข้อมูลทั้งหมด

ดังนั้นจึงได้นำวิธีการสุ่มเพื่อลดจำนวนกลุ่มตัวอย่างมาใช้กับกลุ่มผู้ป่วยปกติ เพื่อให้จำนวนข้อมูลสมดุลกับกลุ่มผู้ป่วยที่มีความเสี่ยง ต่อมาจึงนำวิธีการสุ่มเพื่อเพิ่มจำนวนตัวอย่างกับกลุ่มข้อมูลผู้ป่วยที่มีภาวะครรภ์เป็นพิษ เพื่อให้จำนวนข้อมูลสมดุลกับกลุ่มผู้ป่วยที่มีความเสี่ยง วิธีการนี้ส่งผลให้ทั้งสามกลุ่มข้อมูลมีปริมาณข้อมูลที่สมดุลกันมากขึ้น

ก่อนที่จะทำการสุ่มตัวอย่างข้อมูล จะมีการประเมินความไม่สมดุลของข้อมูลผ่านการคำนวณอัตราส่วนระหว่างกลุ่มข้อมูลที่มีส่วนมากและกลุ่มที่มีส่วนน้อย ซึ่งชุดข้อมูลนี้มีอัตราส่วนความไม่สมดุลอย่างชัดเจน ดังนั้นจึงจำเป็นต้องใช้วิธีการสุ่มตัวอย่างแบบผสมผสานเพื่อลดจำนวนข้อมูลในกลุ่มปกติและเพิ่มข้อมูลในกลุ่มครรภ์เป็นพิษ เพื่อให้ค่าสัดส่วนของชุดข้อมูลเข้าใกล้ 1 ซึ่งแสดงถึงความสมดุลที่ดีขึ้น (Noorhalim et al., 2019) การเลือกใช้วิธีการเหล่านี้ยังอ้างอิงจากงานวิจัยที่ได้นำเทคนิคต่าง ๆ มาปรับใช้ในการจำแนกประเภทข้อมูล (Khushi et al., 2021)

ในด้านวิทยาศาสตร์การแพทย์ เครื่องมือที่ใช้ในการจัดการกับความไม่สมดุลของกลุ่มตัวอย่างข้อมูลได้รับการนำมาใช้ในงานวิจัยหลายชิ้น และมีการจัดการข้อมูลที่ไม่สมดุลในหลากหลายบริบทเพื่อการวินิจฉัย เช่น การวินิจฉัยโรคมะเร็งปอด (Khushi et al., 2021), ผลข้างเคียงทางการแพทย์ (Santiso et al., 2019), และการระบุสารพันธุกรรม (Herndon et al., 2016)

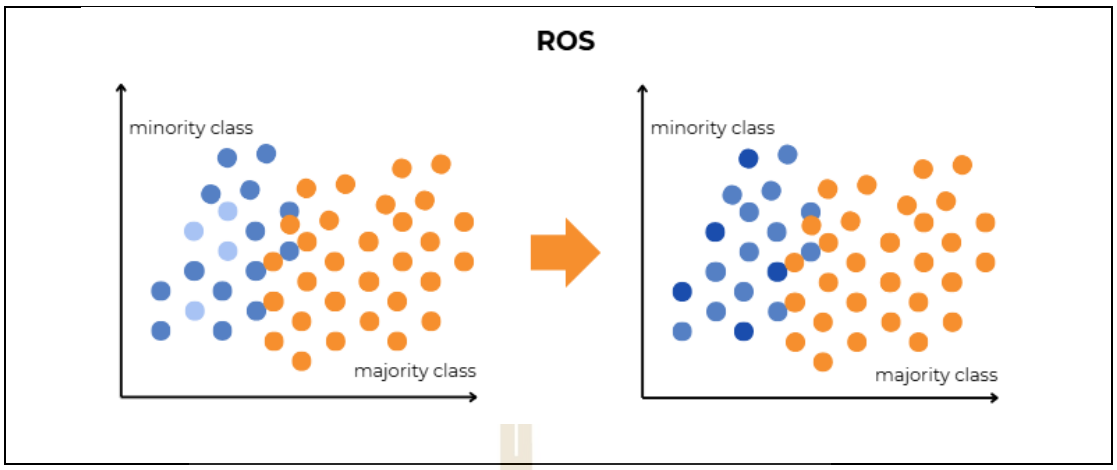
วิธีการระดับข้อมูลในปัจจุบันจะถูกใช้ในขั้นตอนการเตรียมข้อมูล โดยการสุ่มตัวอย่างข้อมูลและจัดการการกระจายตัวของกลุ่มประชากรในข้อมูลใหม่ เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล (Longadge and Dongre, 2013) โดยการเพิ่มจำนวนข้อมูลในกลุ่มข้อมูลชนกลุ่มน้อย และลดจำนวนข้อมูลในกลุ่มข้อมูลชนกลุ่มมาก ในวิธีการระดับข้อมูลส่วนถัดไปจะอธิบายวิธีการต่างๆ แบบกระชับสั้น ๆ เกี่ยวกับเทคนิคในระดับข้อมูล



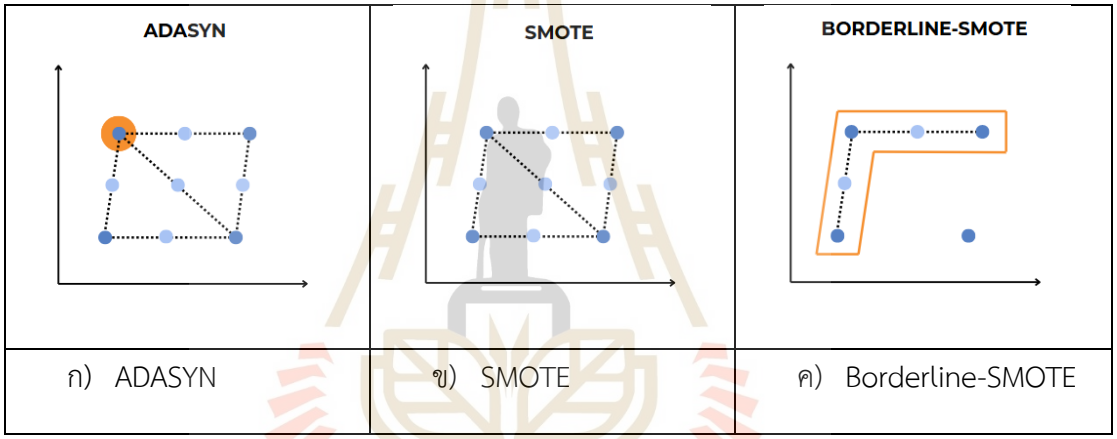
รูปที่ 2.1 หลักการทำงานการสุ่มเพิ่มและลดตัวอย่างข้อมูล

โดยจากการศึกษาการสุ่มเพิ่มตัวอย่างของข้อมูลนั้นจะเป็นการเพิ่มกลุ่มข้อมูลกลุ่มน้อยขึ้นมาและการสุ่มตัวอย่างลดข้อมูลนั้นจะเป็นการลดกลุ่มข้อมูลส่วนใหญ่ลงมานั่นเอง เป็นไปดังรูปที่ 2.1

วิธีการสุ่มเพิ่มตัวอย่างมีดังนี้: Random Oversampling (ROS), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE



รูปที่ 2.2 หลักการทำงานการสุ่มเพิ่มข้อมูลด้วยวิธี ROS



ก) ADASYN | ข) SMOTE | ค) Borderline-SMOTE

รูปที่ 2.3 หลักการทำงานการสุ่มเพิ่มข้อมูลด้วยวิธี ADASYN (ก) SMOTE (ข) และ Borderline-SMOTE (ค)

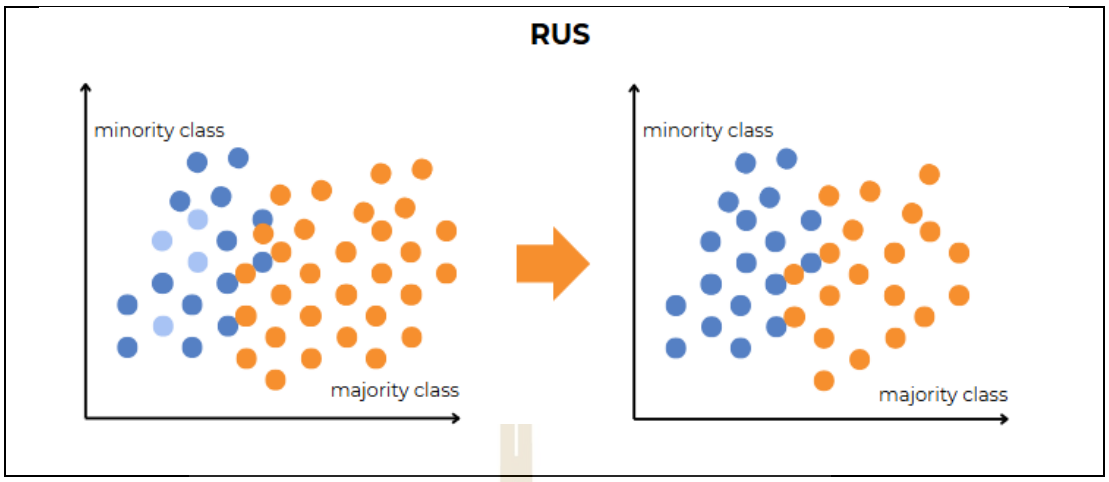
1) ROS เป็นเทคนิคการวิเคราะห์ข้อมูลในขั้นต้นที่ถูกพัฒนาขึ้นเพื่อใช้ในการสุ่มเพิ่มข้อมูลตัวอย่างในคลาสชนกลุ่มน้อย (Batista et al., 2004) ดังรูปที่ 2.2

2) ADASYN ใช้การกระจายตัวถ่วงน้ำหนักในกลุ่มตัวอย่างของข้อมูลคลาสชนกลุ่มน้อยตามระดับความยากในการเรียนรู้ โดยจะสร้างกลุ่มตัวอย่างมากขึ้นในส่วนที่มีความยากในการเรียนรู้สูงกว่า (He et al., 2008) ดังรูปที่ 2.3

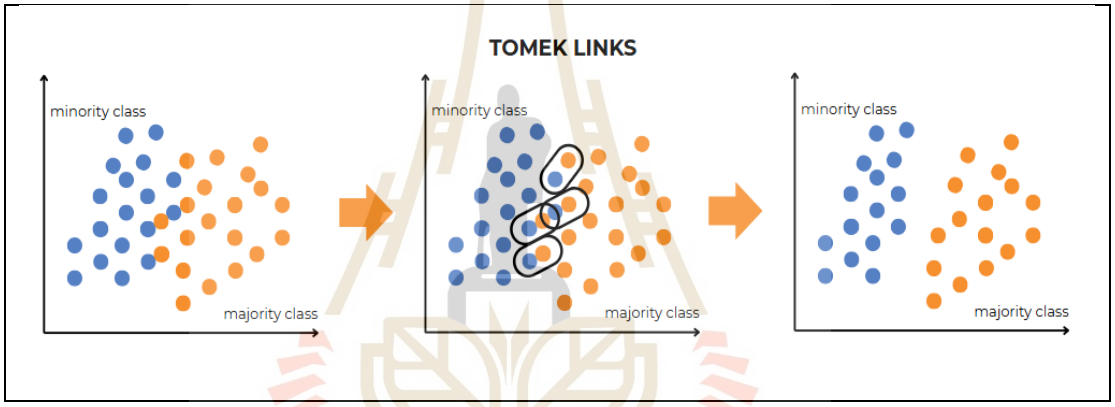
3) SMOTE ใช้ k-Nearest Neighbors (k-NN) เพื่อประมาณค่า ก่อนที่จะทำการสังเคราะห์ข้อมูลใหม่ในคลาสชนกลุ่มน้อยระหว่างข้อมูลเดิม (Chawla et al., 2002) ดังรูปที่ 2.3

4) Borderline-SMOTE เป็นการใช้ SMOTE กับข้อมูลที่อยู่ในเขตชายขอบ ซึ่งมักจะถูกจำแนกผิดประเภท และทำการสังเคราะห์ข้อมูลบริเวณนั้นเพิ่ม (Han et al., 2005) ดังรูปที่ 2.3

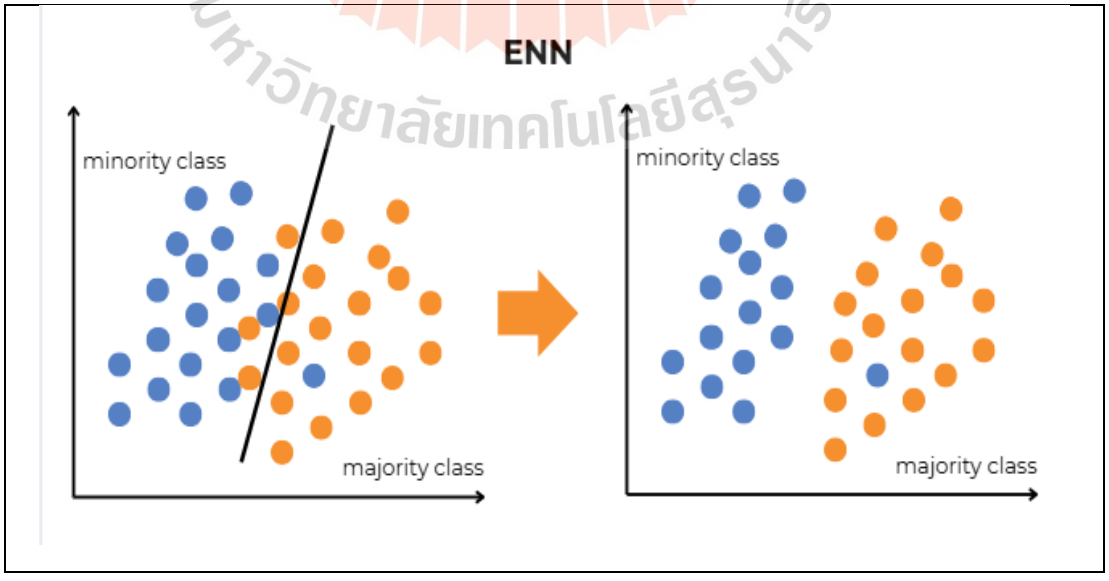
วิธีการสุ่มลดตัวอย่างมีดังนี้: Random Undersampling (RUS), Tomek Links, Edited Nearest Neighbors (ENN), Repeated Edited Nearest Neighbors (RENN)



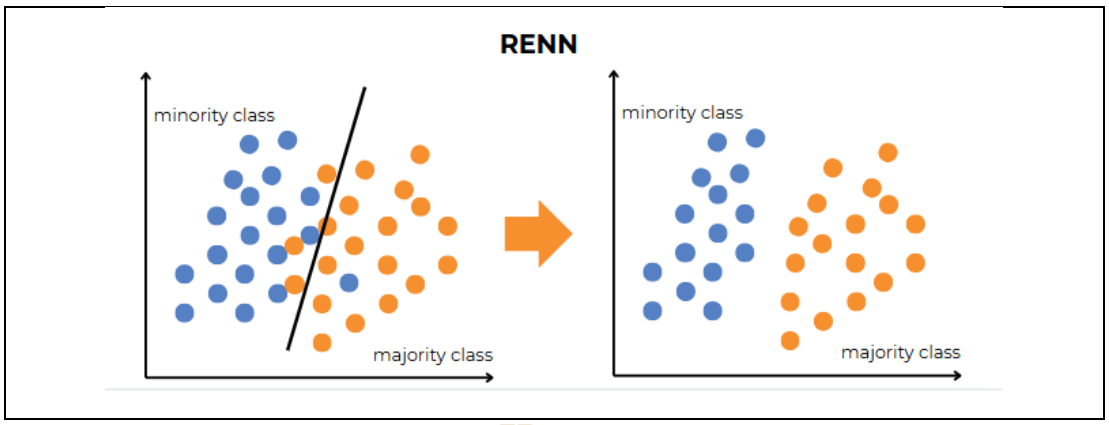
รูปที่ 2.4 หลักการทำงานการสุ่มลดข้อมูลด้วยวิธี RUS



รูปที่ 2.5 หลักการทำงานการสุ่มลดข้อมูลด้วยวิธี Tomek Links



รูปที่ 2.6 หลักการทำงานการสุ่มลดข้อมูลด้วยวิธี ENN



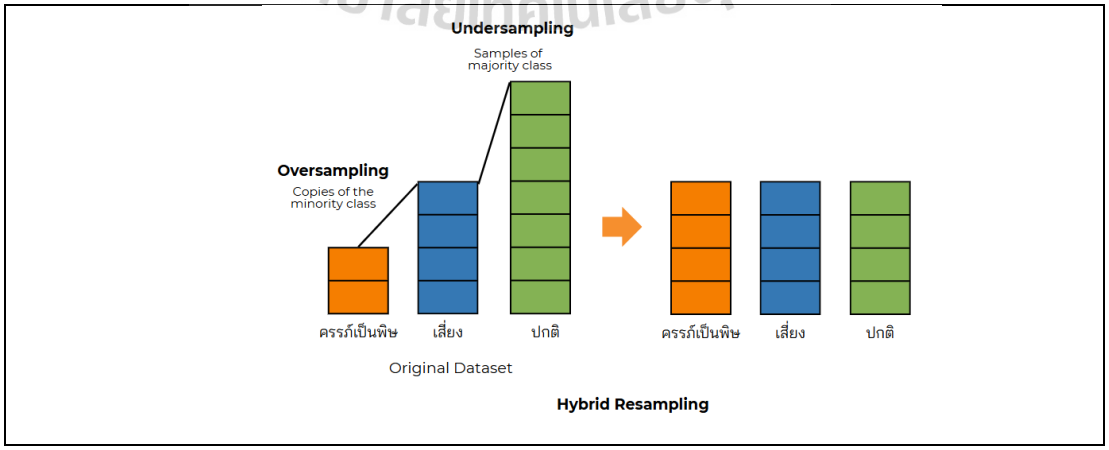
รูปที่ 2.7 หลักการทำงานการสุ่มลดข้อมูลด้วยวิธี RENN

1) RUS เป็นเทคนิคการสุ่มตัวอย่างที่เก่าแก่และได้รับการพัฒนามาตั้งแต่แรก โดยจะลดจำนวนตัวอย่างจากกลุ่มข้อมูลส่วนมาก (Prusa et al., 2015) ดังรูปที่ 2.4

2) Tomek Links สองตัวอย่างข้อมูลจะถูกพิจารณาว่าเป็น Tomek Links หากทั้งสองตัวอย่างอยู่ในกลุ่มข้อมูลที่แตกต่างกัน และเป็นเพื่อนบ้านที่ใกล้ที่สุดในกลุ่มข้อมูล ทั้งนี้สามารถหมายถึงตัวอย่างที่มีข้อผิดพลาดที่อาจจะรบกวน และเป็นตัวอย่างจากกลุ่มข้อมูลส่วนมากจะถูกลบออก (Tomek, 1976) ดังรูปที่ 2.5

3) ENN ในวิธีนี้ กลุ่มตัวอย่างแต่ละกลุ่มจะได้รับการทดสอบโดยใช้ k Nearest Neighbors (k-NN) กับกลุ่มข้อมูลที่เหลือ ตัวอย่างที่ถูกจำแนกไม่ถูกต้องจะถูกตัดทิ้ง และตัวอย่างที่เหลือจะกลายเป็นชุดข้อมูลที่ได้รับการแก้ไข (Broder et al., 1985) ดังรูปที่ 2.6

4) RENN เป็นการทำ ENN ซ้ำ จนกว่าจะสามารถแก้ไขชุดข้อมูลที่ไม่สามารถใช้ในการฝึกสอนได้ โดยจะไม่ถูกกระทบจากการกำจัดตัวอย่างอีกต่อไป (Wilson, 1972) ดังรูปที่ 2.7



รูปที่ 2.8 หลักการทำงานการสุ่มเพิ่มลดตัวอย่างข้อมูลแบบผสมผสาน

วิธีการสุ่มตัวอย่างแบบผสมผสานในงานวิจัยนี้ได้มีการเพิ่มจำนวนประชากรข้อมูลกลุ่มส่วนน้อย และลดจำนวนประชากรของข้อมูลส่วนมาก ที่ผสมผสานจับคู่ระหว่าง การสุ่มลดตัวอย่างกลุ่มข้อมูลชนกลุ่มมาก และ การสุ่มเพิ่มตัวอย่างข้อมูลชนกลุ่มน้อย ดังรูปที่ 2.8

2.2 การออกแบบแบบจำลอง

ในงานวิจัยนี้มีการใช้จะใช้ 6 อัลกอริทึมการจำแนกประเภทด้านการเรียนรู้ของเครื่อง คือ Decision Tree, Extreme Gradient Boosting (XGBoost), Random Forest, Naive Bayes, Neural Network, และ K-Nearest Neighbors (Kaur et al., 2019; Papacharalampous et al., 2021) ซึ่งเป็นตัวที่ใช้ในงานทางการสร้างแบบจำลองเพื่อทำนายข้อมูลที่ไม่สมดุล ปัญหาด้านข้อมูลทางการแพทย์

ตารางที่ 2.1 รายการสรุปการออกแบบแบบจำลอง

อัลกอริทึม	การใช้งาน	การอ้างอิง
Decision Tree	ใช้จำแนกประเภทข้อมูลที่ไม่สมดุล โดยปรับปรุงความแม่นยำผ่านการเลือกอัลกอริทึมที่เหมาะสมและการทดลองในชุดข้อมูลหลากหลาย	(Krawczyk et al., 2014), (Sanz et al., 2015), (Park and Ghosh, 2014)
XGBoost	เปรียบเทียบประสิทธิภาพกับอัลกอริทึมอื่นในงานวิเคราะห์ข้อมูล เช่น ปริมาณน้ำฝนและการวินิจฉัยโรคไตเรื้อรัง พบว่าให้ผลลัพธ์ที่ดีที่สุด	(Papacharalampous et al., 2021), (Ogunleye et al., 2020)
Random Forest	ใช้ในการจำแนกข้อมูลไม่สมดุล เช่น ข้อมูลข้อความและโรคมะเร็งปอด โดยให้ผลลัพธ์ดีที่สุดเมื่อเทียบกับอัลกอริทึมอื่น	(Wu et al., 2014), (Khushi et al., 2021), (Khabsa et al., 2016)
Naive Bayes	ใช้จัดการข้อมูลไม่สมดุล เช่น การจัดกลุ่มนักเรียนและในด้านวิทยาศาสตร์การแพทย์ โดยเน้นการจัดการข้อมูลที่มีมิติสูง	(Yap et al., 2014), (Bhandari and Patel, 2015), (Kaur et al., 2019)
Neural Network	ใช้ในหลากหลายงาน เช่น การประมวลผลข้อมูลลิ้นอิเล็กทรอนิกส์ การตรวจจับการทุจริตบัตรเครดิต และการปรับปรุงการกระจายข้อมูลในชุดข้อมูลที่ไม่สมดุล	(Liu et al., 2013), (Fu et al., 2016), (Jeatrakul et al., 2010)
K-Nearest Neighbors (KNN)	ใช้จำแนกประเภทข้อมูลไม่สมดุล เช่น การพยากรณ์โรคมะเร็ง โดยมีการเพิ่มตัวอย่างในคลาสที่มีจำนวนน้อย และเสนอ Hybrid Weighted Nearest Neighbors เพื่อเพิ่มความแม่นยำในข้อมูลไม่สมดุล	(Patel and Thakur, 2016), (Majid et al., 2014)

2.2.1 Decision Tree เป็นเครื่องมือทางอัลกอริทึมที่สามารถใช้ในการจำแนกประเภทได้ (Krawczyk et al., 2014) โดยการใช้ Decision Tree ในการจำแนกประเภทข้อมูลที่มีลักษณะไม่สมดุล ซึ่งเป็นลักษณะเด่นของชุดข้อมูลจริง ที่มักพบว่ามีกลุ่มข้อมูลที่ไม่สมดุล การเลือกใช้อัลกอริทึมในการจำแนกประเภทจึงมีความสำคัญ นอกจากนี้ยังได้มีการประยุกต์ใช้ในการคำนวณที่มีความแม่นยำดี โดยการนำไปใช้กับชุดข้อมูล 11 ชุด (Sanz et al., 2015) และงานวิจัยนี้ยังได้ใช้ในการแก้ปัญหาข้อมูลที่ไม่สมดุล ซึ่งมีประเภทของกลุ่มข้อมูลหลากหลาย และให้ผลลัพธ์ที่มีประสิทธิภาพในผลการทดลอง (Park and Ghosh, 2014)

2.2.2 XGBoost XGBoost นำไปใช้ในการเปรียบเทียบประสิทธิภาพกับอัลกอริทึมต่าง ๆ ในชุดข้อมูลปริมาณน้ำฝน ซึ่งผลลัพธ์แสดงให้เห็นว่า XGBoost มีประสิทธิภาพดีที่สุดในงานนี้ (Papacharalampous et al., 2021) นอกจากนี้ในงานวิจัยของ (Ogunleye et al., 2020) ได้เสนอให้ใช้ XGBoost ในการวินิจฉัยโรคไตเรื้อรัง

2.2.3 Random ถูกนำเสนอโดย (Wu et al., 2014) เป็นเครื่องมืออัลกอริทึมการเรียนรู้ของเครื่อง ซึ่งนำไปใช้ในการจำแนกประเภทข้อความที่มีข้อมูลไม่สมดุล ในงานการวินิจฉัยโรคมะเร็งปอดวิธีการนี้ได้รับการพิสูจน์ว่าให้ผลลัพธ์ดีที่สุดเมื่อเปรียบเทียบกับวิธีการอื่นๆ (Khushi et al., 2021) และในงานของ (Khabsa et al., 2016) ได้มีการนำเสนอการใช้ Random Forest เป็นกลยุทธ์ที่มีความแม่นยำในการศึกษาที่เกี่ยวข้อง โดยใช้ในการทบทวนระบบที่จำลองมาจากปัญหาการจำแนกประเภทที่ใช้กับข้อมูลไม่สมดุล

2.2.4 Naive Bayes ในงานที่มีการเปรียบเทียบวิธีการที่เหมาะสมในการจัดการกับข้อมูลที่ไม่สมดุล มีการนำเสนอ Naive Bayes เพื่อใช้ในการจำแนกข้อมูลที่มีความไม่สมดุล (Yap et al., 2014) ในงานของ (Bhandari and Patel, 2015) ได้นำเสนอการใช้ Naive Bayes กับชุดข้อมูลที่มีมิติสูงและข้อมูลที่ไม่สมดุลในการจัดกลุ่มนักเรียน และในการทบทวนระบบเกี่ยวกับความท้าทายในการจัดการข้อมูลที่ไม่สมดุลในการเรียนรู้ของเครื่องมือ ยังมีการเสนอวิธีนี้ในงานวิจัยด้านวิทยาศาสตร์การแพทย์ (Kaur et al., 2019)

2.2.5 Neural Network โครงข่ายประสาทเทียมมีการใช้งานที่หลากหลายในอัลกอริทึมการเรียนรู้ของเครื่อง (machine learning) ซึ่ง (Liu et al., 2013) ใช้โครงข่ายประสาทเทียมในการประมวลผลข้อมูลลิ้นอิเล็กทรอนิกส์ เพื่อตรวจจับตัวอย่างเครื่องดื่มส้มและน้ำส้มสายชูจีน และเปรียบเทียบกับอัลกอริทึมอื่นๆ โดยพบว่าโครงข่ายประสาทเทียมให้ผลลัพธ์ที่แม่นยำกว่า (Fu et al., 2016) ใช้โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network) ในการ

ตรวจจับการทุจริตบัตรเครดิตจากธุรกรรมของธนาคารพาณิชย์ โดยผลลัพธ์แสดงว่า ใช้โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network) มีประสิทธิภาพดีกว่าวิธีดั้งเดิมในการตรวจจับการทุจริต (Jeatrakul et al., 2010) เสนอการใช้เทคนิคการสุ่มตัวอย่างแบบลดตัวอย่าง และเพิ่มตัวอย่าง โดยใช้โครงข่ายประสาทเทียมที่เสริมสมรรถนะและเทคนิค SMOTE เพื่อแก้ปัญหาความไม่สมดุลของข้อมูล ซึ่งช่วยปรับปรุงการกระจายข้อมูลในชุดข้อมูลและเพิ่มประสิทธิภาพในการฝึกโมเดล

2.2.6 K-Nearest Neighbors เป็นอัลกอริทึมการจำแนกประเภทที่ได้รับความนิยมในด้านการเรียนรู้ของเครื่อง เนื่องจากใช้งานง่ายและไม่ซับซ้อน ในการจัดการข้อมูลที่ไม่สมดุล (Patel and Thakur, 2016) ได้เสนออัลกอริทึมแบบไฮบริดที่เรียกว่า Hybrid Weighted Nearest Neighbors ซึ่งกำหนดน้ำหนักและจำนวน k ตามขนาดของแต่ละคลาส เพื่อเพิ่มความแม่นยำในการจำแนกประเภท (Majid et al., 2014) ศึกษาการพยากรณ์โรคมะเร็งเต้านมและมะเร็งลำไส้ใหญ่ โดยปรับสมดุลข้อมูลผ่านการเพิ่มตัวอย่างในคลาสที่มีจำนวนน้อย จากนั้นเปรียบเทียบผลระหว่าง SVM และ KNN ซึ่งพบว่า SVM ให้ผลลัพธ์ที่ดีกว่า KNN

2.3 งานวิจัยที่เกี่ยวข้องกับภาวะครรภ์เป็นพิษ

ภาวะครรภ์เป็นพิษเป็นภาวะสุขภาพที่ส่งผลต่อหญิงตั้งครรภ์หลังอายุครรภ์ 20 สัปดาห์ โดยมีลักษณะสำคัญคือความดันโลหิตสูงร่วมกับการพบโปรตีนในปัสสาวะ ภาวะนี้สามารถกระทบต่อการทำงานของหลายระบบในร่างกาย และเพิ่มความเสี่ยงต่อภาวะแทรกซ้อนที่รุนแรงทั้งต่อมารดาและทารก เช่น การคลอดก่อนกำหนด หรือความล้มเหลวของไตและตับ การรักษาที่มีประสิทธิภาพที่สุดในปัจจุบันคือการให้คลอดเมื่อถึงระยะเวลาที่เหมาะสม (Vata et al., 2015; Briceño-Pérez et al., 2009)

ตลอดช่วงหลายปีที่ผ่านมา ได้มีความพยายามจากหลากหลายงานวิจัยในการพัฒนาวิธีการคัดกรองหรือทำนายความเสี่ยงของภาวะครรภ์เป็นพิษให้ได้ตั้งแต่ระยะเริ่มต้น โดยใช้ข้อมูลจากประวัติสุขภาพของหญิงตั้งครรภ์ เช่น ดัชนีมวลกาย ระดับความดันโลหิต สารชีวภาพในเลือด หรือการตรวจด้วยอัลตราซาวด์ อย่างไรก็ตาม แนวทางและตัวแปรที่ใช้ในแต่ละงานวิจัยยังแตกต่างกัน ทำให้ยังไม่สามารถพัฒนาแบบจำลองที่ใช้ได้อย่างแพร่หลายในสถานพยาบาลทั่วไปได้ (De Kat et al., 2019)

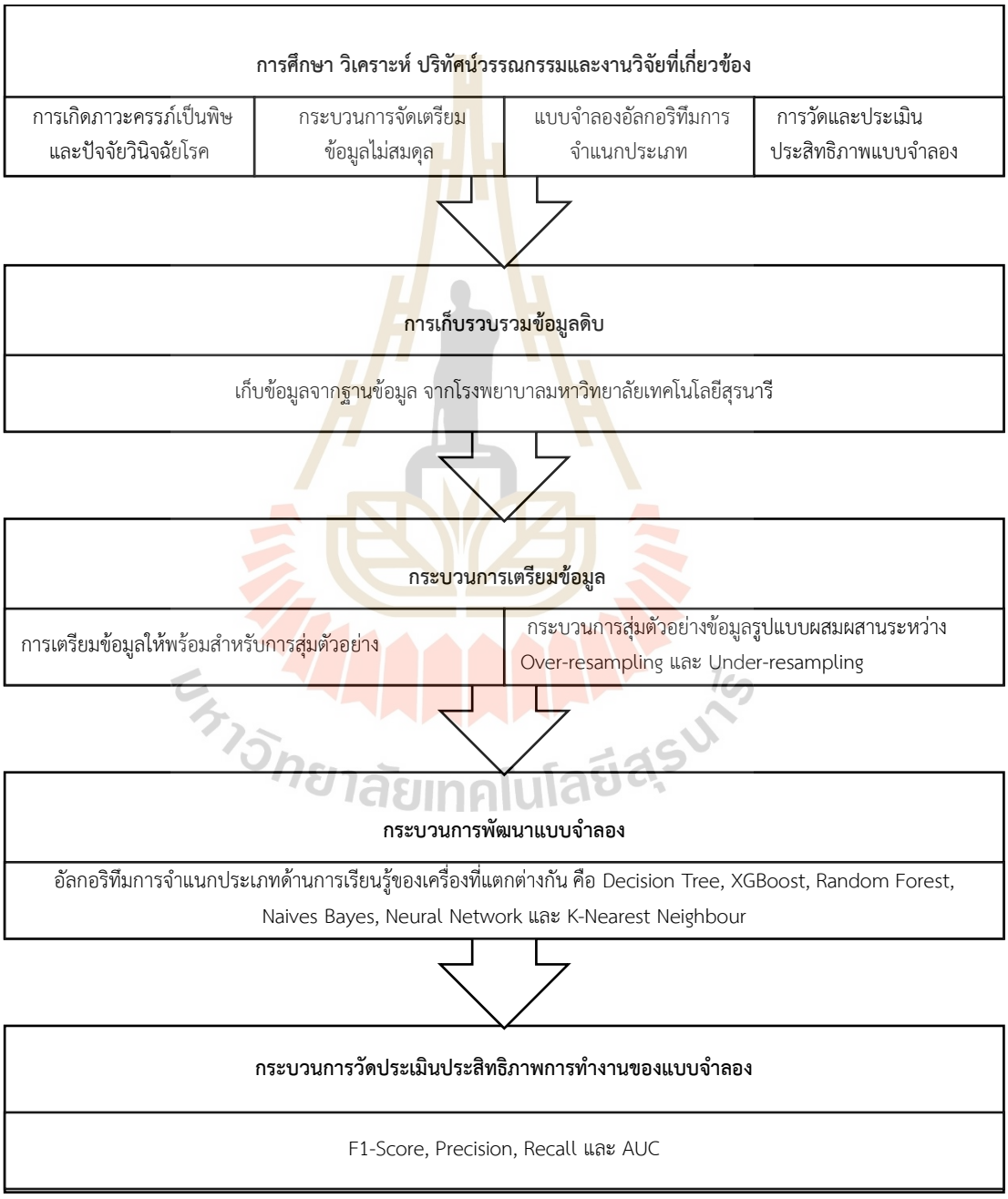
นอกจากนี้ อัตราการเกิดภาวะครรภ์เป็นพิษยังแตกต่างกันในแต่ละภูมิภาค โดยเฉพาะในประเทศที่มีทรัพยากรจำกัด เช่น ประเทศในแถบแอฟริกา ซึ่งพบอัตราการเกิดสูงกว่าค่าเฉลี่ยโลกอย่างชัดเจน ปัจจัยเสี่ยงที่เกี่ยวข้องอาจรวมถึงภาวะโภชนาการ ความรู้เกี่ยวกับโรค และการเข้าถึงบริการฝากครรภ์ที่เหมาะสม ในบางพื้นที่พบว่าหญิงตั้งครรภ์กว่า 70% ไม่เคยได้รับข้อมูลพื้นฐานเกี่ยวกับภาวะนี้เลย (Vata et al., 2015)

แม้จะมีความพยายามในการพัฒนาแนวทางการป้องกันและการทำนายความเสี่ยงของภาวะครรภ์เป็นพิษ แต่ปัจจุบันองค์ความรู้เกี่ยวกับสาเหตุที่แท้จริงของโรคยังมีจำกัด อีกทั้งยังต้องการงานวิจัยเพิ่มเติม โดยเฉพาะในบริบทของประเทศที่มีข้อจำกัดด้านระบบสาธารณสุข เช่น เอธิโอเปีย ซึ่งจำเป็นต้องมีแนวทางที่เหมาะสมกับบริบทจริง เพื่อช่วยลดความเสี่ยงและผลกระทบที่อาจเกิดขึ้น

บทที่ 3

วิธีดำเนินการวิจัย

3.1 ภาพรวมการทำวิทยานิพนธ์



รูปที่ 3.1 ผังแสดงภาพรวมการทำวิทยานิพนธ์

3.2 ชุดข้อมูล

ตารางที่ 3.1 รายการสรุปข้อมูลที่ขอจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี

ข้อมูลก่อนการตั้งครรภ์	- น้ำหนัก - ส่วนสูง - ประวัติความดันโลหิต - ประวัติโปรตีนในปัสสาวะ - ประวัติโรคประจำตัว - วันเกิด - BMI
ข้อมูลระหว่างการตั้งครรภ์	- BMI (ขณะตั้งครรภ์) - ประวัติการเป็นครรภ์เป็นพิษ - เป็นความดันโลหิตสูงหรือไม่ - เป็นเบาหวานหรือไม่ ประเภทใด - ประวัติน้ำหนักของแม่ - ประวัติความดันโลหิต (ขณะตั้งครรภ์) - ประวัติโปรตีนในปัสสาวะ - ประวัติโรคประจำตัว - อาการป่วยที่เกิดขึ้นระหว่างตั้งครรภ์ เช่น อาการปวดหัว จุกแน่นใต้ลิ้นปี่ ตาพร่ามัว หรือน้ำหนักเพิ่มขึ้นเร็วผิดปกติ การติดเชื้อในกระแสเลือด เป็นไข้ - อายุครรภ์ของการตั้งครรภ์ - วันที่พบแพทย์ - วันที่มาฝากครรภ์ - ความเข้มของเลือด - เกล็ดเลือด - การแข็งตัวของเลือด - เคยมีอาการตัวบวมหรือไม่ - เป็นท้องแฝดหรือไม่ - เป็นการตั้งครรภ์ครั้งแรกหรือไม่ - มีอาการครรภ์เป็นพิษหรือไม่ ระดับใด

งานวิจัยฉบับนี้ได้ดำเนินการรวบรวมข้อมูลผู้ป่วยจากฐานข้อมูลของโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี จำนวนทั้งสิ้น 1,491 ราย โดยอยู่ภายใต้การปรึกษาหารือกับแพทย์ผู้เชี่ยวชาญด้านสูตินรีเวช และเจ้าหน้าที่จากแผนกเทคโนโลยีสารสนเทศ เพื่อรับรองความถูกต้องและความน่าเชื่อถือของข้อมูลที่นำมาใช้ในการศึกษา จากนั้นได้มีการกำหนดประเภทของข้อมูลที่มีความจำเป็นต่อการดำเนินการวิจัย และจัดเก็บข้อมูลดังกล่าวจากเวชระเบียนหรือสื่อบันทึกข้อมูลอื่น ๆ โดยดำเนินการภายใต้หลักการคุ้มครองข้อมูลส่วนบุคคลอย่างเคร่งครัด โดยไม่มีการเปิดเผยข้อมูลที่สามารถระบุตัวตนของผู้ป่วยได้ ข้อมูลที่ใช้ในการศึกษาครอบคลุมช่วงเวลาการฝากครรภ์ระหว่างปี พ.ศ. 2558 ถึง พ.ศ. 2566 โดยการวิจัยนี้ได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีสุรนารี เลขที่ EC-65-121

ตารางที่ 3.2 รายการสรุปโปรไฟล์ของชุดข้อมูลก่อนทำการสุ่มตัวอย่าง

ชื่อคอลัมน์	**ชนิดของข้อมูล**	**คำอธิบายลักษณะข้อมูล**
เป้าหมายคลาส	ตัวเลข	0 = ปกติ (normal), 1 = ภาวะครรภ์เป็นพิษ (PCS), 2 = เสี่ยง (risk)
ปวดหัว	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 มีอาการปวดหัว
จุกแน่นลิ้นปี่	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 มีอาการจุกแน่นลิ้นปี่
มองเห็นไม่ชัด	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 มีอาการมองเห็นไม่ชัด
บวมทั่วร่างกาย	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 มีอาการบวมทั่วร่างกาย
รหัสผู้ป่วย	ตัวเลข	ข้อมูลเชิงปริมาณที่ใช้เป็นตัวระบุ (ID) สำหรับผู้ป่วยแต่ละคน
อายุ	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับอายุของผู้ป่วย
อายุมากกว่า 40 ปี	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 ผู้ป่วยมีอายุเกิน 40 ปี
ส่วนสูง	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับส่วนสูงของผู้ป่วย
ชีพจร	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับอัตราการเต้นของชีพจรของผู้ป่วย
อุณหภูมิร่างกาย	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับอุณหภูมิของร่างกาย
ดัชนีมวลกาย (BMI)	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับดัชนีมวลกาย (BMI) ของผู้ป่วย
ภาวะอ้วน	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 ผู้ป่วยมีภาวะอ้วน
ค่าความดันโลหิต	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับระดับความดันโลหิตของผู้ป่วย
น้ำหนักตัว	ตัวเลข	ข้อมูลเชิงปริมาณเกี่ยวกับน้ำหนักตัวของผู้ป่วย
รหัสโรคต่าง ๆ	ตัวเลข	ถ้ามีอาการมีค่าเป็น 0 ไม่มีอาการ, ถ้ามีอาการมีค่าเป็น 1 รหัสโรคที่เกี่ยวข้องกับผู้ป่วย

ตารางที่ 3.3 รายการคอลัมน์ที่เหลือหลังจากการตัดข้อมูลที่ขาดหายไปมากกว่า 60%

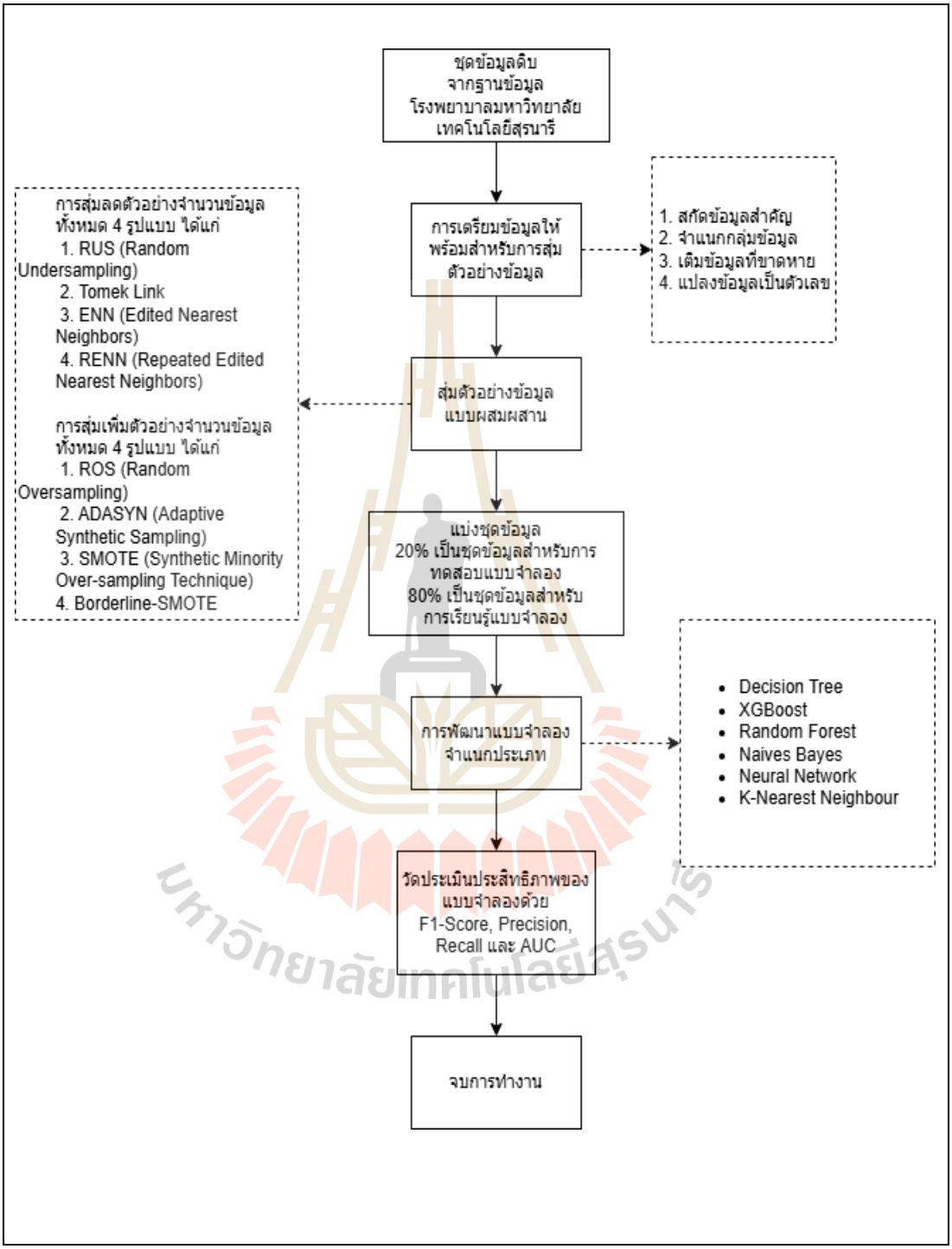
ชื่อคอลัมน์ที่เหลือ	**เปอร์เซ็นต์ข้อมูลที่หายไป (%)**
น้ำหนักตัว	20.0928
ค่าความดันโลหิต	20.0206
ภาวะอ้วน	19.5361
ดัชนีมวลกาย (BMI)	19.5361
อุณหภูมิร่างกาย	19.3608
ชีพจร	18.1443
ส่วนสูง	12.9381

โปรไฟล์ข้อมูล (Data Profiling) สำหรับชุดข้อมูลที่ได้จากโรงพยาบาลเทคโนโลยีสุรนารี เกี่ยวกับผู้ที่มีประวัติฝากครรภ์ที่ผ่านการเตรียมข้อมูลเรียบร้อยก่อนทำการสุ่มตัวอย่างข้อมูลก่อนนำไปสร้างแบบจำลอง มีรายละเอียดดังตารางที่ 3.2 เดิมทีข้อมูลจากโรงพยาบาลเทคโนโลยีสุรนารีได้ถูกคัดเลือกจากผู้ที่มีประวัติการฝากครรภ์ โดยเริ่มจากชุดข้อมูลดิบที่มีจำนวน 9,700 แถว และ 731 คอลัมน์ ซึ่งหลังจากการลบข้อมูลที่ไม่เกี่ยวข้องหรือไม่ใช้งาน และตรวจสอบข้อมูลที่ขาดหาย พบว่า ข้อมูลบางคอลัมน์ เช่น ค่าโปรตีนในปัสสาวะ, อายุครรภ์, และภาวะท้องเป็นครรภ์แรก และอีกหลายคอลัมน์มีการขาดหายมากกว่า 60% ซึ่งทำให้ไม่สามารถกรอกข้อมูลได้จากข้อมูลก่อนหน้า จึงได้ตัดคอลัมน์เหล่านั้นออกจากการพิจารณา

ในส่วนที่ข้อมูลยังขาดหายอยู่ตามที่แสดงในตารางที่ 3.3 ได้มีการเติมข้อมูลที่ขาดหายด้วยวิธีการที่เหมาะสม โดยเติมข้อมูลโดยใช้ข้อมูลในอดีตรายบุคคลได้ โดยการเติมข้อมูลส่วนนี้จะใช้ข้อมูลก่อนหน้าที่ขาดไปมาเติมส่วนที่ขาดหาย แต่ถ้าหากไม่มีข้อมูลก่อนหน้าจะใช้ค่าเฉลี่ยของแต่ละคนในการเติมค่า และข้อมูลที่ใช้เป็นข้อมูลเฉพาะรายบุคคล

ชุดข้อมูลนี้แสดงถึงความไม่สมดุลของข้อมูลที่ชัดเจน โดยในกลุ่มเป้าหมายคลาสปกติ มีจำนวนตัวอย่างมากถึง 9,341 ราย ในขณะที่กลุ่มมีความเสี่ยง มีเพียง 123 ราย และกลุ่มภาวะครรภ์เป็นพิษ มีเพียง 9 ราย การกระจายข้อมูลที่ไม่สมดุลนี้อาจทำให้โมเดลการเรียนรู้ของเครื่อง ถูกโน้มน้าวไปในทิศทางของกลุ่มที่มีจำนวนตัวอย่างมาก (ปกติ) ซึ่งอาจส่งผลต่อประสิทธิภาพในการจำแนกกลุ่มที่มีจำนวนตัวอย่างน้อย (เช่น กลุ่มมีความเสี่ยงและภาวะครรภ์เป็นพิษ) การจัดการกับข้อมูลไม่สมดุลนี้จึงเป็นสิ่งสำคัญในการเพิ่มความแม่นยำและประสิทธิภาพในการวิเคราะห์และทำนายผล

3.3 กรอบแนวคิดการวิจัยและการเลือกรูปแบบแบบจำลอง



รูปที่ 3.2 ผังแสดงแนวคิดการวิจัย

3.4 วิธีการเตรียมข้อมูล

การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ ขั้นตอนแรกในงานวิจัยคือการเตรียมข้อมูลให้พร้อมสำหรับการนำไปประมวลผล รวมถึงการนำข้อมูลไปวิเคราะห์ในขั้นตอนถัดไป นอกจากนี้ยังมีการปรับลักษณะข้อมูลให้เข้ากับอัลกอริทึมสำหรับการสร้างโมเดล รวมทั้งการออกแบบการทดลองในขั้นตอนการจัดการกับความไม่สมดุลของข้อมูล และการพัฒนาแบบจำลอง

ขั้นตอนการดำเนินงานมีดังนี้

3.4.1 การสกัดข้อมูลที่สำคัญจากข้อความประวัติของผู้ป่วยปัจจุบันในกลุ่มประชากร ซึ่งประกอบด้วยข้อมูลเกี่ยวกับอาการ รหัสโรค และอาการ รวมถึงข้อมูลที่สกัดออกมา เช่น อายุการตั้งครรภ์ ข้อมูลเกี่ยวกับการตั้งครรภ์ครั้งแรกหรือไม่ อาการจุกแน่นบริเวณลิ้นปี่ อาการตาพร่ามัว อาการปวดศีรษะ อาการตัวบวม และอาการอ่อนเพลีย โดยข้อมูลเหล่านี้จะถูกนำมาใช้ในขั้นตอนการจำแนกประเภทกลุ่มผู้ป่วย เพื่อเพิ่มประสิทธิภาพในขั้นตอนถัดไปของการทำงาน

3.4.2 การจำแนกกลุ่มของข้อมูล โดยข้อมูลเริ่มต้นจะระบุกลุ่มที่เป็นครรภ์ปกติและกลุ่มครรภ์เป็นพิษ จากนั้นจะจำแนกกลุ่มเสี่ยงเพิ่มเติมจากกลุ่มครรภ์ปกติ เพื่อให้สามารถใช้จำแนกกลุ่มเสี่ยงของข้อมูลได้อย่างมีประสิทธิภาพ กลุ่มเสี่ยงจะประกอบด้วยอาการดังนี้ ความดันโลหิตสูงและพบโปรตีนในปัสสาวะ อาการจุกแน่นที่ลิ้นปี่ อาการตาพร่ามัว อาการปวดศีรษะ อาการตัวบวม การตั้งครรภ์ครั้งแรก ผู้ตั้งครรภ์อายุเกิน 40 ปี ประวัติครรภ์เป็นพิษ โรคประจำตัว เช่น โรคความดันโลหิตสูง โรคไต โรคเบาหวาน โรคภูมิคุ้มกันบกพร่อง โรคอ้วน น้ำหนักตัวเพิ่มผิดปกติตามเกณฑ์การเพิ่มน้ำหนักของหญิงตั้งครรภ์, และอาการอ่อนเพลีย เมื่อพบอาการใดอาการหนึ่งจากกลุ่มครรภ์ปกติ จะถูกจัดให้อยู่ในกลุ่มเสี่ยงทันที (Pinitlertsakun and Wayuhuerd., 2021; พญ. ปวีณา บุตรดีวงศ์, ม.ป.ป.)

3.4.3 หลังจากได้กลุ่มเสี่ยงแล้วก็ทำการเติมข้อมูลด้วยข้อมูลรายบุคคล

3.4.4 การแปลงข้อมูลที่ไม่ใช่ข้อมูลตัวเลขให้เป็นข้อมูลตัวเลข เนื่องจากต้องนำข้อมูลไปใช้กับเทคนิคการสุ่มตัวอย่างใช้ได้เฉพาะกับข้อมูลเชิงตัวเลขเท่านั้น ดังนั้นจึงมีการแปลงข้อมูลที่ไม่ใช่ตัวเลขให้เป็นข้อมูลตัวเลข

3.4.5 เมื่อข้อมูลกลายเป็นตัวเลขแล้ว พบว่าอีกหลายคอลัมน์มีการขาดหายมากกว่า 60% ซึ่งทำให้ไม่สามารถกรอกข้อมูลได้จากข้อมูลก่อนหน้า จึงได้ตัดคอลัมน์เหล่านั้นออกจากการพิจารณา ในส่วนที่ข้อมูลยังขาดหายอยู่ตามที่แสดงในตารางที่ 3.3 ได้มีการเติมข้อมูลที่ขาดหายด้วยวิธีการที่เหมาะสม โดยเติมข้อมูลโดยใช้ข้อมูลในอดีตรายบุคคลได้ โดยการเติมข้อมูลส่วนนี้จะใช้ข้อมูลก่อนหน้าที่ขาดไปมาเติมส่วนที่ขาดหาย แต่ถ้าหากไม่มีข้อมูลก่อนหน้าจะใช้ค่าเฉลี่ยของแต่ละคนในการเติมค่า และข้อมูลที่ใช้เป็นข้อมูลเฉพาะรายบุคคล

3.4.6 ชุดข้อมูลนี้แสดงถึงความไม่สมดุลของข้อมูลที่ชัดเจน โดยในกลุ่มเป้าหมายคลาสปกติ มีจำนวนตัวอย่างมากถึง 9,341 ราย ในขณะที่กลุ่มมีความเสี่ยง มีเพียง 123 ราย และกลุ่มภาวะครรภ์เป็นพิษ มีเพียง 9 ราย การกระจายข้อมูลที่ไม่สมดุลนี้อาจทำให้โมเดลการเรียนรู้ของเครื่อง ถูกโน้มน้าวไปในทิศทางของกลุ่มที่มีจำนวนตัวอย่างมาก (ปกติ) ซึ่งอาจส่งผลต่อประสิทธิภาพในการจำแนกกลุ่มที่มีจำนวนตัวอย่างน้อย (เช่น กลุ่มมีความเสี่ยงและภาวะครรภ์เป็นพิษ) การจัดการกับข้อมูลไม่สมดุลนี้จึงเป็นสิ่งสำคัญในการเพิ่มความแม่นยำและประสิทธิภาพในการวิเคราะห์และทำนายผล

3.4.7 การสุ่มตัวอย่างเพื่อให้ข้อมูลในแต่ละกลุ่มมีความสมดุล โดยการผสมผสานวิธีการสุ่มลดตัวอย่างในกลุ่มข้อมูลส่วนมากลง และการสุ่มเพิ่มตัวอย่างในกลุ่มข้อมูลส่วนน้อยขึ้น เพื่อให้ข้อมูลในแต่ละกลุ่มมีความสมดุลในระดับเดียวกัน ในงานวิจัยนี้ได้ใช้รูปแบบการสุ่มลดตัวอย่างจำนวนข้อมูลทั้งหมด 4 รูปแบบ ได้แก่ RUS, Tomek Links, ENN และ RENN ได้ใช้รูปแบบการสุ่มเพิ่มตัวอย่างจำนวนข้อมูลทั้งหมด 4 รูปแบบ ได้แก่ ROS, ADASYN, SMOTE และ Borderline-SMOTE

3.5 การเลือกแบบจำลอง

งานวิจัยนี้ศึกษาอัลกอริทึมการจำแนกประเภทด้านการเรียนรู้ของเครื่อง เพื่อสร้างแบบจำลองการจำแนก 6 รูปแบบที่แตกต่างกัน ได้แก่ Decision Tree, XGBoost, Random Forest, Naive Bayes, Neural Network และ K-Nearest Neighbors โดยมีวิธีการดำเนินการสร้างแบบจำลองแต่ละประเภทดังนี้

3.5.1 การสร้างแบบจำลองด้วย Decision Tree แบบจำลองการเรียนรู้ของเครื่องแบบมีผู้สอนที่ใช้โครงสร้างต้นไม้ในการจำแนกข้อมูล โดยเริ่มจากโหนดรากแล้วแตกกิ่งตามเงื่อนไขที่ใช้คุณลักษณะต่าง ๆ ซึ่งเลือกโดยเกณฑ์ ในการแยกข้อมูลที่เหมาะสมที่สุดในแต่ละขั้น ข้อมูลคุณลักษณะที่ผ่านการเตรียมจะถูกป้อนเข้าไป เพื่อให้แบบจำลองจำแนกประเภทผู้ป่วยตามเป้าหมาย พร้อมค่าความน่าจะเป็น โดยไม่ปรับค่าพารามิเตอร์ใด ๆ เพื่อเน้นการเปรียบเทียบประสิทธิภาพของวิธีการสุ่มตัวอย่างเป็นหลักและวัดประสิทธิภาพของแบบจำลองก่อนนำไปเปรียบเทียบกับวิธีอื่น

3.5.2 การสร้างแบบจำลองด้วย XGBoost การสร้างแบบจำลองด้วย XGBoost ในงานวิจัยนี้ใช้การตั้งค่าพารามิเตอร์แบบเริ่มต้น โดยไม่มีการปรับแต่งเพิ่มเติม เนื่องจากมุ่งเน้นการเปรียบเทียบประสิทธิภาพของวิธีการสุ่มตัวอย่างข้อมูล XGBoost เป็นอัลกอริทึมที่พัฒนาจาก Gradient Boosting ซึ่งใช้แนวคิดของการเรียนรู้แบบผสม (Ensemble Learning)โดยสร้างชุดของต้นไม้

ตัดสินใจที่ทำงานร่วมกันเพื่อลดข้อผิดพลาดในแต่ละรอบการเรียนรู้ โดยผลลัพธ์จะถูกนำมาวิเคราะห์เพื่อประเมินประสิทธิภาพของแต่ละเทคนิคการสุ่มตัวอย่าง

3.5.3 การสร้างแบบจำลองด้วย Random Forest เป็นเทคนิคการเรียนรู้แบบผสม โดยอาศัยการสร้างต้นไม้ตัดสินใจจำนวนมากจากข้อมูลย่อยที่ได้จากการสุ่มแบบมีการแทนที่ (Bootstrap Sampling) และเลือกคุณลักษณะแบบสุ่มในแต่ละโหนด โดยผลลัพธ์สุดท้ายจะได้จากการโหวตเสียงข้างมากของต้นไม้ทั้งหมด แบบจำลองมีความสามารถในการจัดการกับข้อมูลที่มีมิติสูงและทนต่อค่าผิดปกติได้ สำหรับการศึกษานี้ใช้การตั้งค่าพารามิเตอร์แบบค่าเริ่มต้น โดยไม่ทำการปรับแต่งเพิ่มเติม เพื่อเน้นการเปรียบเทียบประสิทธิภาพของวิธีการสุ่มตัวอย่าง (Resampling) โดยเฉพาะในกรณีของข้อมูลที่ไม่สมดุล ทั้งนี้ การประเมินแบบจำลองจะพิจารณาจากความสามารถในการจำแนกกลุ่มเป้าหมายและค่าความน่าจะเป็นของการจัดกลุ่มที่ได้จากข้อมูลคุณลักษณะของผู้ป่วยที่ผ่านการเตรียมไว้ล่วงหน้า

3.5.4 การสร้างแบบจำลองด้วย Naive Bayes เป็นอัลกอริทึมที่ใช้ทฤษฎีความน่าจะเป็นแบบเบย์ในการคำนวณความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรผลลัพธ์เมื่อกำหนดค่าตัวแปรต้น โดยมีสมมติฐานว่าตัวแปรต้นทุกตัวเป็นอิสระต่อกัน หลังจากฝึกสอนแบบจำลองแล้ว จะทำการวัดประสิทธิภาพและบันทึกผลการดำเนินการ

3.5.5 การสร้างแบบจำลองด้วย Neural Network หรือโครงข่ายประสาทเทียมเป็นโมเดลที่ได้รับแรงบันดาลใจจากการทำงานของสมองมนุษย์ ซึ่งประกอบด้วยชั้นข้อมูลนำเข้า (Input layer) ชั้นซ่อน (Hidden layers) และชั้นข้อมูลส่งออก (Output layer) โดยแต่ละชั้นมีโหนดหรือนิวรอนที่เชื่อมต่อกัน การเรียนรู้ของเครือข่ายเกิดจากการปรับค่าน้ำหนักระหว่างการเชื่อมต่อ หลังจากฝึกสอนเสร็จสิ้น จะทำการทดสอบและประเมินประสิทธิภาพของแบบจำลองที่ได้

3.5.6 การสร้างแบบจำลองด้วย K-Nearest Neighbors (KNN) เป็นเทคนิคในการจำแนกประเภทข้อมูลที่ทำงานโดยการหาค่าคลาสที่พบบ่อยที่สุดจากกลุ่มข้อมูลที่ใกล้เคียงที่สุด K ตัว ซึ่งจะคำนวณจากระยะห่างระหว่างข้อมูลทดสอบกับข้อมูลในชุดฝึกสอน โดยไม่ต้องมีการฝึกสอนแบบเชิงลึกหรือปรับค่าน้ำหนักเหมือน Neural Network หลังจากฝึกสอนด้วยการจัดกลุ่มข้อมูลในลักษณะนี้แล้ว จะทำการทดสอบและประเมินผลการจำแนกเพื่อวัดประสิทธิภาพของแบบจำลอง

แบบจำลอง 6 ประเภทในการศึกษานี้มีพื้นฐานมาจากการทบทวนวรรณกรรมที่เกี่ยวข้องกับการจัดการข้อมูลไม่สมดุลในบริบททางการแพทย์ ผลการศึกษาแสดงให้เห็นว่าแบบจำลองประเภท tree-based models ได้แก่ Decision Tree, Random Forest และ XGBoost เป็นที่นิยมและมีการอ้างอิงอย่างแพร่หลายในวรรณกรรม เนื่องจากมีประสิทธิภาพสูงในการจัดการข้อมูลที่มีการกระจาย

ตัวไม่สมดุล นอกจากนี้ได้ทำการเพิ่มแบบจำลอง Naive Bayes, Neural Network และ K-Nearest Neighbors (Kaur et al., 2019) เพื่อประเมินประสิทธิภาพในการจัดการข้อมูลไม่สมดุลลงในการศึกษา การเปรียบเทียบแบบจำลองที่มีพื้นฐานการทำงานแตกต่างกันนี้จะช่วยให้เข้าใจคุณลักษณะสำคัญของแบบจำลองที่เหมาะสมกับการจัดการข้อมูลไม่สมดุลในบริบททางการแพทย์ และนำไปสู่การคัดเลือกแบบจำลองที่เหมาะสมที่สุดสำหรับการวิจัย

3.6 การประเมินประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลองการจำแนกประเภทมีความสำคัญอย่างยิ่ง โดยเฉพาะในกรณีที่ข้อมูลมีความไม่สมดุล การใช้เพียงค่าความถูกต้อง (Accuracy) อาจไม่เพียงพอและอาจทำให้เกิดการตีความผลที่คลาดเคลื่อน งานวิจัยนี้จึงใช้มาตรวัดประสิทธิภาพหลายรูปแบบที่เหมาะสมกับการประเมินแบบจำลองในบริบทของข้อมูลที่ไม่สมดุล ได้แก่ Precision, Recall, F1-Score และ Area under the curve (AUC)

ในการคำนวณค่าประสิทธิภาพต่างๆ โดยแสดงความสัมพันธ์ระหว่างค่าที่ทำนายได้และค่าจริง แบ่งเป็น 4 ส่วนดังนี้ True Positive (TP) จำนวนตัวอย่างที่ทำนายว่าถูกต้องในกลุ่มที่สนใจ, False Positive (FP) จำนวนตัวอย่างที่ทำนายว่าผลาดในกลุ่มที่สนใจ, False Negative (FN) จำนวนตัวอย่างที่ทำนายว่าเป็นผิดพลาดในกลุ่มรองลงมา และ True Negative (TN) จำนวนตัวอย่างที่ทำนายถูกต้องในกลุ่มรองลงมา

3.6.1 ความแม่นยำ (Precision) คือค่าที่ใช้วัดความแม่นยำในการทำนายเฉพาะแต่ละคลาส โดยค่าที่เท่ากับหนึ่ง แสดงว่าแบบจำลองสามารถทำนายกลุ่มเป้าหมายได้แม่นยำสูง มีการทำนายผิดน้อย ส่วนค่าที่ใกล้ศูนย์หมายถึงแบบจำลองมีการทำนายผิดว่าเป็นกลุ่มเป้าหมายจำนวนมาก เหมาะสำหรับกรณีที่ต้องการลดการวินิจฉัยผิด เช่น การคัดกรองโรคที่ไม่ต้องการให้บุคคลปกติถูกระบุว่าเป็นโรค

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

3.6.2 ความถูกต้อง (Recall) คือเป็นการวัดความถูกต้องของแบบจำลอง โดยพิจารณาแยกทีละคลาส ค่าสูงใกล้ 1 หมายถึงแบบจำลองสามารถตรวจจับกลุ่มที่สนใจได้เกือบทั้งหมด ค่าต่ำใกล้ 0 หมายถึงแบบจำลองพลาดการตรวจจับกลุ่มสนใจจำนวนมาก เหมาะสำหรับกรณีที่ต้องการลดความผิดพลาดการคัดกรองโรคร้ายแรงที่ไม่ควรพลาดการวินิจฉัย

$$Recall = \frac{TP}{TP + FN}$$

3.6.3 F1-Score คือ เป็นค่าเฉลี่ยแบบฮาร์โมนิค (Harmonic Mean) ระหว่าง Precision และ Recall ช่วยให้เห็นภาพรวมของแบบจำลองที่ต้องสมดุลระหว่างความแม่นยำและความถูกต้อง ค่าสูงใกล้ 1 หมายถึงแบบจำลองมีความแม่นยำและความถูกต้องสูง และตรงกันข้ามเมื่อค่าต่ำใกล้ 0 เหมาะสำหรับกรณีข้อมูลไม่สมดุลเนื่องจากให้ความสำคัญกับ ทำนายผิดพลาด F1-Score จะต่ำหากมีค่า Precision หรือ Recall ต่ำ แม้ว่าอีกค่าหนึ่งจะสูงก็ตาม

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.6.4 AUC คือค่าที่ใช้วัดประสิทธิภาพของแบบจำลอง โดยเฉพาะเมื่อข้อมูลมีความไม่สมดุล (imbalanced data) ซึ่งค่านี้ยิ่งสูงยิ่งดี หมายถึงแบบจำลองสามารถรักษาค่าความแม่นยำ (Precision) ไว้ได้ในระดับที่สูงในช่วงของค่า Recall ที่ต่างกัน ค่าสูงใกล้ 1 หมายถึงแบบจำลองมีประสิทธิภาพดีทั้งในด้านการดึงตัวอย่างที่มีภาวะครรภ์เป็นพิษออกมา (Recall) และทำนายตัวอย่างภาวะครรภ์เป็นพิษได้แม่นยำ (Precision) และตรงกันข้ามเมื่อค่าต่ำใกล้ 0 หมายความว่าไม่สามารถแยกแยะกลุ่มภาวะครรภ์เป็นพิษได้ เหมาะสำหรับกรณีข้อมูลไม่สมดุลเนื่องจากใช้ในการประเมินแบบจำลองที่ใช้กับข้อมูลทางการแพทย์ที่มักมีความไม่สมดุล โดย P ในสมการแทน ค่า Precision และ R ในสมการแทน Recall ที่จุด i

$$AUC \approx \sum_{i=1}^{n-1} (R_{i+1} - R_i) \cdot \frac{P_{i+1} + P_i}{2}$$

บทที่ 4
ผลการวิจัยและการอภิปรายผล

4.1 วิธีการทดลอง

การทดลองในงานวิจัยนี้มีวัตถุประสงค์สอดคล้องกับวัตถุประสงค์หลักของการวิจัยที่กำหนดไว้ โดยมุ่งดำเนินการตามแนวทางดังต่อไปนี้ ประการแรก เพื่อวิเคราะห์ความสัมพันธ์ระหว่างปัจจัยต่าง ๆ จากข้อมูลประวัติทางคลินิกของสตรีตั้งครรภ์กับการเกิดภาวะครรภ์เป็นพิษ โดยประยุกต์ใช้เทคนิคการวิเคราะห์ข้อมูลเชิงลึก ประการที่สอง เพื่อคัดเลือกวิธีการจัดการกับข้อมูลที่มีความไม่สมดุลให้เหมาะสมกับลักษณะของชุดข้อมูล โดยเฉพาะอย่างยิ่งการเปรียบเทียบประสิทธิภาพของเทคนิคการสุ่มแบบผสมผสานในกระบวนการเตรียมข้อมูลสำหรับการพัฒนาแบบจำลองทำนายภาวะครรภ์เป็นพิษ และประการสุดท้าย เพื่อพัฒนาแบบจำลองการทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษในสตรีตั้งครรภ์ได้

ในการดำเนินการทดลองดังกล่าว ได้ให้ความสำคัญกับการศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการจัดการข้อมูลที่มีความไม่สมดุลด้วยวิธีการต่าง ๆ เพื่อกำหนดแนวทางที่เหมาะสมที่สุดในการพัฒนาแบบจำลองสำหรับการทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษ

4.1.1 การแบ่งชุดข้อมูล การวิจัยครั้งนี้ได้ดำเนินการแบ่งชุดข้อมูลออกเป็นสองส่วนตามอัตราส่วน 80:20 โดยใช้วิธีการสุ่มแบบผสมผสาน เพื่อให้ได้ชุดข้อมูลที่เป็นตัวแทนที่สมดุลของประชากรข้อมูลทั้งหมด ประกอบด้วย ชุดข้อมูลสำหรับการฝึกสอน (Training Set) คิดเป็นร้อยละ 80 ของข้อมูลทั้งหมด สำหรับการพัฒนาและปรับปรุงแบบจำลองเชิงทำนาย และชุดข้อมูลสำหรับการทดสอบ (Testing Set) คิดเป็นร้อยละ 20 ของข้อมูลทั้งหมด สำหรับการประเมินประสิทธิภาพของแบบจำลองที่พัฒนาขึ้น

ทั้งนี้ กระบวนการแบ่งชุดข้อมูลดำเนินการภายหลังจากการประยุกต์ใช้เทคนิคการสุ่มแบบผสมผสาน เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล โดยข้อมูลที่ผ่านการปรับสมดุลด้วยแต่ละเทคนิคจะถูกแบ่งตามสัดส่วนที่กำหนดไว้เพื่อให้การเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาจากแต่ละเทคนิคการสุ่มแบบผสมผสานมีความน่าเชื่อถือ ได้ดำเนินการสุ่มแบ่งข้อมูลตามสัดส่วนที่กำหนดก่อนนำไปพัฒนาแบบจำลอง อย่างไรก็ตาม เนื่องด้วยข้อจำกัดของกระบวนการสุ่มตัวอย่างจึงไม่สามารถควบคุมให้แต่ละวิธีการสุ่มแบบผสมผสานมีชุดข้อมูลที่เหมือนกันทั้งหมดได้ในทุกกรณี แม้จะสามารถใช้ชุดข้อมูลเดียวกันได้เฉพาะที่มีวิธีการสุ่มแบบเดียวกันในการพัฒนาแบบจำลอง

4.1.2 เทคนิคการสุ่มแบบผสมผสาน การวิจัยนี้ประสบกับความท้าทายด้านข้อมูลไม่สมดุล โดยพบว่าจำนวนตัวอย่างของคลาสตรวจพบภาวะครรภ์เป็นพิษมีปริมาณน้อยอย่างมีนัยสำคัญ จึงได้ประยุกต์ใช้เทคนิคการสุ่มแบบผสมผสานเพื่อปรับสมดุลข้อมูลก่อนการพัฒนาแบบจำลองทำนาย โดยการผสมผสานระหว่างวิธีการสุ่มลดตัวอย่างและวิธีการสุ่มเพิ่มตัวอย่าง

วิธีการสุ่มเพิ่มตัวอย่าง คือ ROS, ADASYN, SMOTE และ Borderline SMOTE วิธีการสุ่มลดตัวอย่าง คือ RUS, Tomek Links, ENN และ RENN

การผสมผสานวิธีการสุ่มตัวอย่าง ในงานวิจัยนี้ ผู้วิจัยได้ทำการทดลองใช้การผสมผสานระหว่างวิธีการสุ่มลดตัวอย่างและวิธีการสุ่มเพิ่มตัวอย่างจำนวน 16 แบบ ดังนี้

1) ENN + ADASYN (enn_adasyn)
2) ENN + Borderline SMOTE (enn_borderline_smote)
3) ENN + Random Oversampling (enn_ros)
4) ENN + SMOTE (enn_smote)
5) RENN + ADASYN (renn_adasyn)
6) RENN + Borderline SMOTE (renn_borderline_smote)
7) RENN + Random Oversampling (renn_ros)
8) RENN + SMOTE (renn_smote)
9) RUS + ADASYN (rus_adasyn)
10) RUS + Borderline SMOTE (rus_borderline_smote)
11) RUS + Random Oversampling (rus_ros)
12) RUS + SMOTE (rus_smote)
13) Tomek Links + ADASYN (tomek_adasyn)
14) Tomek Links + Borderline SMOTE (tomek_borderline_smote)
15) Tomek Links + Random Oversampling (tomek_ros)
16) Tomek Links + SMOTE (tomek_smote)

กระบวนการทำงานของเทคนิคการสุ่มแบบผสมผสานเหล่านี้มีขั้นตอนสำคัญดังต่อไปนี้ ข้อมูลดิบจะผ่านกระบวนการทำความสะอาดและการเตรียมข้อมูลเบื้องต้น จากนั้นนำข้อมูลเข้าสู่กระบวนการสุ่มลดตัวอย่างเป็นลำดับแรก เพื่อขจัดข้อมูลที่อาจเป็นสัญญาณรบกวน หรือข้อมูลผิดปกติ ในคลาสส่วนใหญ่ หลังจากนั้นข้อมูลที่ผ่านการสุ่มลดตัวอย่างแล้วจะถูกนำไปผ่านกระบวนการสุ่มเพิ่มตัวอย่าง เพื่อเพิ่มจำนวนตัวอย่างในคลาสส่วนน้อยให้เกิดความสมดุลมากขึ้น ทั้งนี้ข้อมูลที่ผ่านการปรับสมดุลด้วยเทคนิคการสุ่มแบบผสมผสานแต่ละรูปแบบจะถูกนำไปใช้ในการพัฒนาและประเมินประสิทธิภาพของแบบจำลองทำนายต่อไป

4.2 ผลการวิจัย

ตารางที่ 4.1 ผลการวิจัยจากแบบจำลอง Decision Tree

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
Decision Tree	enn_adasyn	0.9688	0.9722	0.9683	0.9747
	enn_borderline_smote	0.9380	0.9420	0.9444	0.9522
	enn_ros	0.9535	0.9565	0.9565	0.9633
	enn_smote	0.9535	0.9565	0.9565	0.9633
	renn_adasyn	0.9688	0.9722	0.9683	0.9747
	renn_borderline_smote	0.9535	0.9565	0.9565	0.9633
	renn_ros	0.9223	0.9275	0.9333	0.9417
	renn_smote	0.9690	0.9710	0.9697	0.9749
	rus_adasyn	0.8565	0.8522	0.8629	0.8801
	rus_borderline_smote	0.7823	0.7851	0.7816	0.8171
	rus_ros	0.8473	0.8555	0.8488	0.8769
	rus_smote	0.8135	0.8185	0.8157	0.8464
	tomek_adasyn	0.9507	0.9507	0.9507	0.9597
	tomek_borderline_smote	0.9048	0.9070	0.9032	0.9208
	tomek_ros	0.9598	0.9677	0.9565	0.9689
	tomek_smote	0.9190	0.9237	0.9164	0.9335

การใช้แบบจำลอง Decision Tree ให้ผลลัพธ์ที่ดีที่สุดในค่า F1-Score ซึ่งมีค่าเท่ากับ 0.9690 จากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย RENN ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย SMOTE นอกจากนี้ยังให้ผลลัพธ์ที่ดีในค่า Recall และ AUC อีกด้วย ส่วนค่า Precision ได้ค่าที่สูงสุดเท่ากับ 0.9722 เมื่อใช้เทคนิคการสุ่มลดตัวอย่างด้วย ENN ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ADASYN และเมื่อใช้เทคนิคการสุ่มลดตัวอย่างด้วย RENN ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ADASYN

ตารางที่ 4.2 ผลการวิจัยจากแบบจำลอง XGBoost

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
XGBoost	enn_adasyn	0.9531	0.9522	0.9547	0.9779
	enn_borderline_smote	0.9535	0.9565	0.9565	0.9888
	enn_ros	0.9535	0.9565	0.9565	0.9888
	enn_smote	0.9535	0.9565	0.9565	0.9888
	renn_adasyn	0.9531	0.9522	0.9547	0.9779
	renn_borderline_smote	0.9535	0.9565	0.9565	0.9888
	renn_ros	0.9535	0.9565	0.9565	0.9888
	renn_smote	0.9535	0.9565	0.9565	0.9888
	rus_adasyn	0.8995	0.8999	0.9017	0.9530
	rus_borderline_smote	0.9038	0.9144	0.8995	0.9358
	rus_ros	0.8868	0.8902	0.8915	0.9683
	rus_smote	0.8934	0.8954	0.8947	0.9410
	tomek_adasyn	0.9751	0.9798	0.9722	0.9892
	tomek_borderline_smote	0.9720	0.9798	0.9667	0.9806
	tomek_ros	0.9864	0.9841	0.9892	1.0000
	tomek_smote	0.9720	0.9798	0.9667	0.9934

การใช้แบบจำลอง XGBoost ให้ผลลัพธ์ที่ดีที่สุดในด้าน F1-Score, Precision, Recall, และ AUC โดยมีค่าเท่ากับ 0.9864, 0.9841, 0.9892, และ 1.0000 ตามลำดับ ซึ่งได้มาจากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ROS

ตารางที่ 4.3 ผลการวิจัยจากแบบจำลอง Random Forest

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
Random Forest	enn_adasyn	0.9392	0.9420	0.9408	0.9809
	enn_borderline_smote	0.9556	0.9593	0.9565	0.9807
	enn_ros	0.9556	0.9593	0.9565	0.9823
	enn_smote	0.9397	0.9461	0.9420	0.9823
	renn_adasyn	0.9392	0.9420	0.9408	0.9788
	renn_borderline_smote	0.9556	0.9593	0.9565	0.9810
	renn_ros	0.9556	0.9593	0.9565	0.9804
	renn_smote	0.9556	0.9593	0.9565	0.9776
	rus_adasyn	0.9269	0.9292	0.9264	0.9671
	rus_borderline_smote	0.9005	0.9046	0.9001	0.9589
	rus_ros	0.8990	0.9077	0.9023	0.9649
	rus_smote	0.9143	0.9157	0.9146	0.9740
	tomek_adasyn	0.9624	0.9706	0.9583	0.9875
	tomek_borderline_smote	0.9592	0.9571	0.9618	0.9750
	tomek_ros	0.9601	0.9558	0.9677	0.9860
	tomek_smote	0.9603	0.9586	0.9640	0.9923

การใช้แบบจำลอง Random Forest ให้ผลลัพธ์ที่ดีที่สุดในค่า F1-Score โดยมีค่าเท่ากับ 0.9624 และค่า Precision เท่ากับ 0.9706 เมื่อใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ADASYN ในขณะที่ค่า Recall สูงสุดอยู่ที่ 0.9677 เมื่อใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ROS สำหรับค่า AUC ผลลัพธ์ที่ดีที่สุดคือ 0.9923 ซึ่งได้จากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย SMOTE

ตารางที่ 4.4 ผลการวิจัยจากแบบจำลอง Naive Bayes

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
Naive Bayes	enn_adasyn	0.6692	0.7183	0.6952	0.7399
	enn_borderline_smote	0.6587	0.6615	0.6616	0.7442
	enn_ros	0.6587	0.6615	0.6616	0.7442
	enn_smote	0.6587	0.6615	0.6616	0.7442
	renn_adasyn	0.6692	0.7183	0.6952	0.7399
	renn_borderline_smote	0.6587	0.6615	0.6616	0.7442
	renn_ros	0.6587	0.6615	0.6616	0.7442
	renn_smote	0.6587	0.6615	0.6616	0.7442
	rus_adasyn	0.4641	0.4803	0.4783	0.4907
	rus_borderline_smote	0.2922	0.2530	0.3484	0.5439
	rus_ros	0.3906	0.3982	0.3977	0.5715
	rus_smote	0.3930	0.4049	0.3880	0.4914
	tomek_adasyn	0.6970	0.7564	0.7325	0.7769
	tomek_borderline_smote	0.6727	0.6834	0.6839	0.7082
	tomek_ros	0.5211	0.4641	0.6667	0.7363
	tomek_smote	0.6686	0.7156	0.7075	0.7423

การใช้แบบจำลอง Naive Bayes ให้ผลลัพธ์ที่ดีที่สุดในด้าน F1-Score, Precision, Recall, และ AUC โดยมีค่าเท่ากับ 0.6970, 0.7564, 0.7325, และ 0.7769 ตามลำดับ ซึ่งได้มาจากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างด้วย ADASYN

ตารางที่ 4.5 ผลการวิจัยจากแบบจำลอง Neural Network

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
Neural Network	enn_adasyn	0.8870	0.8925	0.8837	0.9531
	enn_borderline_smote	0.8928	0.8947	0.8953	0.9774
	enn_ros	0.8930	0.8914	0.8990	0.9798
	enn_smote	0.9200	0.9211	0.9205	0.9819
	renn_adasyn	0.8532	0.8666	0.8486	0.9517
	renn_borderline_smote	0.9323	0.9372	0.9313	0.9770
	renn_ros	0.9034	0.9051	0.9060	0.9771
	renn_smote	0.9064	0.9104	0.9060	0.9802
	rus_adasyn	0.8938	0.9052	0.8876	0.9015
	rus_borderline_smote	0.8546	0.8677	0.8566	0.8652
	rus_ros	0.9047	0.9180	0.9023	0.9443
	rus_smote	0.8546	0.8677	0.8566	0.8710
	tomek_adasyn	0.9080	0.9033	0.9216	0.9885
	tomek_borderline_smote	0.8630	0.8718	0.8700	0.8879
	tomek_ros	0.8831	0.8798	0.8883	0.9214
	tomek_smote	0.8791	0.8839	0.8845	0.9139

การใช้แบบจำลอง Neural Network ให้ผลลัพธ์ที่ดีที่สุดในด้าน F1-Score, Precision, และ Recall โดยมีค่าเท่ากับ 0.9323, 0.9372, และ 0.9313 ตามลำดับ ซึ่งได้มาจากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย RENN ร่วมกับการสุ่มเพิ่มตัวอย่างด้วย Borderline-SMOTE ส่วนค่า AUC มีค่าเท่ากับ 0.9885 ซึ่งได้จากการใช้เทคนิคการสุ่มลดตัวอย่างด้วย Tomek ร่วมกับการสุ่มเพิ่มตัวอย่างด้วย ADASYN

ตารางที่ 4.6 ผลการวิจัยจากแบบจำลอง K-Nearest Neighbors (KNN)

แบบจำลอง	เทคนิคการผสมผสาน วิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	Precision	Recall	AUC
K-Nearest Neighbors (KNN)	enn_adasyn	0.9531	0.9522	0.9547	0.9776
	enn_borderline_smote	0.9535	0.9565	0.9565	0.9749
	enn_ros	0.9535	0.9565	0.9565	0.9749
	enn_smote	0.9535	0.9565	0.9565	0.9749
	renn_adasyn	0.9531	0.9522	0.9547	0.9776
	renn_borderline_smote	0.9535	0.9565	0.9565	0.9749
	renn_ros	0.9535	0.9565	0.9565	0.9749
	renn_smote	0.9535	0.9565	0.9565	0.9749
	rus_adasyn	0.7862	0.7890	0.7893	0.8745
	rus_borderline_smote	0.7747	0.7762	0.7753	0.8172
	rus_ros	0.8896	0.8871	0.8990	0.9174
	rus_smote	0.7604	0.7692	0.7587	0.8314
	tomek_adasyn	0.9387	0.9378	0.9400	0.9785
	tomek_borderline_smote	0.9190	0.9164	0.9237	0.9749
	tomek_ros	0.9208	0.9231	0.9355	0.9786
	tomek_smote	0.9177	0.9177	0.9177	0.9711

จากการทดสอบแบบจำลอง K-Nearest Neighbors (KNN) โดยใช้เทคนิคการผสมผสานระหว่างวิธีการสุ่มลดตัวอย่างและวิธีการสุ่มเพิ่มตัวอย่างในแต่ละวิธี พบว่าแบบจำลองที่ใช้วิธีการสุ่มลดตัวอย่าง ENN ร่วมกับการสุ่มเพิ่มตัวอย่าง Borderline-SMOTE, ROS, และ SMOTE รวมถึงแบบจำลองที่ใช้วิธีการสุ่มลดตัวอย่าง RENN ร่วมกับการสุ่มเพิ่มตัวอย่าง Borderline-SMOTE, ROS, และ SMOTE ให้ผลลัพธ์ดีที่สุดในด้าน F1-Score, Precision, และ Recall โดยมีค่าเท่ากับ 0.9535, 0.9565, และ 0.9565 ตามลำดับ ส่วนค่า AUC สูงสุดที่ได้จากการใช้วิธีการสุ่มลดตัวอย่าง Tomek ร่วมกับการสุ่มเพิ่มตัวอย่าง Borderline-SMOTE คือ 0.9786

4.3 การอภิปรายผล

ตารางที่ 4.7 ตารางสรุปผลการทดลอง

แบบจำลอง	เทคนิคการผสมผสานวิธีการสุ่มเพิ่มลดตัวอย่าง	F1-Score	AUC
Decision Tree	renn_smote	0.9690	0.9749
XGBoost	tomek_ros	0.9864	1.0000
Random Forest	tomek_adasyn	0.9624	0.9875
	tomek_smote	0.9603	0.9923
Naive Bayes	tomek_adasyn	0.6970	0.7769
Neural Network	renn_borderline_smote	0.9323	0.9770
	tomek_adasyn	0.9080	0.9885
K-Nearest Neighbors (KNN)	enn_borderline_smote	0.9535	0.9749
	enn_ros	0.9535	0.9749
	enn_smote	0.9535	0.9749
	renn_borderline_smote	0.9535	0.9749
	renn_ros	0.9535	0.9749
	renn_smote	0.9535	0.9749
	tomek_ros	0.9208	0.9786

จากการวิเคราะห์ข้อมูลร่วมกับวัตถุประสงค์ของการวิจัยในครั้งนี้ มีเป้าหมายเพื่อศึกษาปัจจัยที่มีความสัมพันธ์กับการเกิดภาวะครรภ์เป็นพิษในสตรีมีครรภ์ และแสวงหาแนวทางที่เหมาะสมในการจัดการข้อมูลที่ไม่สมดุล เพื่อใช้ในการเตรียมข้อมูลก่อนการสร้างแบบจำลองปัญญาประดิษฐ์สำหรับการทำนายความเสี่ยงของภาวะครรภ์เป็นพิษ โดยใช้ชุดข้อมูลจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี ขั้นตอนการดำเนินงานประกอบด้วย การวิเคราะห์ข้อมูลเบื้องต้นเพื่อศึกษาความสัมพันธ์ระหว่างตัวแปร การเตรียมข้อมูลซึ่งครอบคลุมถึงการสกัดข้อมูลสำคัญ การจำแนกกลุ่มข้อมูล การเติมค่าข้อมูลที่ขาดหาย และการแปลงข้อมูลให้อยู่ในรูปแบบเชิงตัวเลข พร้อมทั้งการจัดการข้อมูลที่ไม่สมดุลผ่านกระบวนการสุ่มตัวอย่างทั้งในลักษณะของการลดและเพิ่มจำนวนข้อมูล จากนั้นแบ่งข้อมูลออกเป็น 80% สำหรับการฝึกแบบจำลอง และ 20% สำหรับการทดสอบแบบจำลอง ทั้งนี้ ได้ดำเนินการพัฒนาแบบจำลองด้วยอัลกอริธึมหลายประเภท ได้แก่ Decision Tree, XGBoost, Random Forest, Naive Bayes, Neural Network และ K-Nearest Neighbors พร้อมทั้งประเมินประสิทธิภาพของแบบจำลองโดยใช้เกณฑ์วัดผล ได้แก่ F1-Score, Precision, Recall และ AUC เพื่อเปรียบเทียบและเลือกวิธีที่เหมาะสมที่สุดสำหรับการนำไปใช้ในบริบทของข้อมูลทางการแพทย์ที่ไม่สมดุลชุดนี้

ข้อมูลทางการแพทย์มีลักษณะที่ซับซ้อนทำให้การใช้ระบบที่อาศัยกฎเกณฑ์แบบตายตัว (Rule-based System) มีข้อจำกัดสำคัญหลายประการ ความซับซ้อนของข้อมูลทางการแพทย์เกิดจากการที่ร่างกายมนุษย์เป็นระบบที่มีปฏิสัมพันธ์ซับซ้อนระหว่างปัจจัย ซึ่งแต่ละปัจจัยสามารถส่งผลกระทบต่อกัน เมื่อความซับซ้อนของข้อมูลเพิ่มขึ้น กฎเกณฑ์ที่ต้องใช้ในการตัดสินใจทางการแพทย์ก็จะมีความซับซ้อนเพิ่มขึ้นตามไปด้วย การสร้างกฎที่ครอบคลุมทุกสถานการณ์ที่เป็นไปได้จึงเป็นเรื่องที่ยาก ในกรณีของภาวะครรภ์เป็นพิษ หากต้องสร้างกฎที่ครอบคลุมปัจจัยเสี่ยงทั้งหมด เช่น อายุ น้ำหนัก ความดันโลหิต ระดับโปรตีนในปัสสาวะ ประวัติการเจ็บป่วย ประวัติครอบครัว และปัจจัยอื่น ๆ อีกมากมาย กฎที่ได้จะมีความซับซ้อนอย่างมหาศาล และที่สำคัญคือจะมีกฎที่ขัดแย้งกันเองในหลายสถานการณ์ เพราะปัจจัยต่าง ๆ เหล่านี้ไม่ได้มีผลกัดแย้งกันอย่างเป็นอิสระ แต่มีปฏิสัมพันธ์ซับซ้อนที่ไม่สามารถแยกออกจากกันได้ ในขณะที่ระบบปัญญาประดิษฐ์สามารถเรียนรู้และปรับตัวจากข้อมูลใหม่ได้อย่างอัตโนมัติ โดยไม่ต้องมีการเขียนกฎใหม่ทุกครั้ง ทำให้สามารถรองรับความซับซ้อนและความไม่แน่นอนของข้อมูลทางการแพทย์ได้ดีกว่า พร้อมทั้งสร้างการคาดการณ์ที่มีความแม่นยำสูงกว่าจากการวิเคราะห์รูปแบบที่ซ่อนอยู่ในข้อมูลจำนวนมาก

ตารางที่ 4.8 ตารางผลการทดลองแบบจำลอง XGBoost ร่วมกับเทคนิค tomek_ros

ชื่อคลาสข้อมูล	F1-Score	Precision	Recall	AUC
ปกติ	0.9810	0.9626	1.0000	0.9626
เสี่ยง	1.0000	1.0000	1.0000	1.0000
ครรภ์เป็นพิษ	0.9778	1.0000	0.9565	0.9565

ในส่วนนี้เป็นการนำเสนอผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องที่ใช้เทคนิคการสุ่มลดตัวอย่างแบบผสมผสาน โดยพิจารณาค่า F1-Score และ AUC เป็นตัวชี้วัดหลักในการประเมินประสิทธิภาพของแบบจำลอง ผลลัพธ์ซึ่งแสดงไว้ในตารางที่ 4.8 แสดงให้เห็นรูปแบบที่แตกต่างกันอย่างชัดเจนในลักษณะของแต่ละแบบจำลองต่อเทคนิคการสุ่มที่แตกต่างกัน ในตารางที่ 4.7 จะแสดงให้เห็นประสิทธิภาพการทำงานของโมเดลแยกคลาส จากผลในตางรางจะเห็นได้ว่าแบบจำลองสามารถทำนายได้ดีในทุกกลุ่มคลาสข้อมูล

แบบจำลอง XGBoost ที่ใช้ร่วมกับเทคนิค tomek_ros ให้ผลลัพธ์ที่ดีที่สุดโดยรวม โดยได้ค่า F1-Score เท่ากับ 0.9864 และ AUC เท่ากับ 1.0000 ซึ่งแสดงถึงความสมดุลระหว่างค่าความแม่นยำ (Precision) และค่าความสามารถในการดึงข้อมูลที่เกี่ยวข้อง (Recall) ได้เป็นอย่างดี รวมถึงสามารถแยกแยะข้อมูลระหว่างคลาสได้ สะท้อนให้เห็นถึงศักยภาพของ XGBoost เมื่อใช้ร่วมกับเทคนิคการสุ่มอย่างเหมาะสมในการจัดการกับชุดข้อมูลที่ไม่สมดุล โดยตัว XGBoost มีจุดเด่นด้านการสร้างขอบเขตการตัดสินใจที่ชัดเจนและสามารถเรียนรู้เพื่อแก้ไขข้อผิดพลาดได้อย่างต่อเนื่อง การใช้ Tomek Links ช่วยลดข้อมูลรบกวนที่อยู่ใกล้ขอบเขตของคลาส ทำให้โมเดลสามารถแยกคลาสได้ชัดเจนยิ่งขึ้น ในขณะที่ ROS เพิ่มจำนวนข้อมูลกลุ่มน้อยโดยการทำซ้ำข้อมูลเดิม จึงไม่ก่อให้เกิดข้อมูลรบกวนเพิ่มเติม ส่งผลให้การเรียนรู้มีความสมดุลและเสถียรมากขึ้น ส่งผลให้โมเดลสามารถจำแนกข้อมูลได้อย่างถูกต้องสมบูรณ์ ทั้งในด้าน F1-Score และ AUC ซึ่งประสิทธิภาของ XGBoost เป็นไปตามงานของ (Ogunleye et al., 2020)

แบบจำลอง Random Forest มีความแม่นยำโดยเมื่อใช้ tomek_adasyn ให้ค่า F1-Score เท่ากับ 0.9624 และ AUC เท่ากับ 0.9875 ในขณะที่ tomek_smote แม้จะได้ค่า F1-Score ลดลงเล็กน้อยที่ 0.9603 แต่กลับได้ค่า AUC สูงขึ้นที่ 0.9923 เป็นไปตามงานของ (Khushi et al., 2021) ที่ได้พิสูจน์ไว้ในด้านของค่า AUC แต่การทดลองของเราได้ผลลัพธ์ที่ดีกว่า ซึ่งสะท้อนถึงความสามารถในการจำแนกข้อมูลได้ดีในลักษณะที่แตกต่างกัน โดย Random Forest มีจุดเด่นในการฝึกจากชุดข้อมูลย่อยหลายชุด ทำให้สามารถทนต่อข้อมูลรบกวนภายในคลาสได้ดี แต่ยังต้องการขอบเขตการตัดสินใจที่ชัดเจน การใช้ Tomek Links จึงช่วยลดข้อมูลรบกวนที่อยู่ใกล้ขอบเขตคลาส ทำให้การจำแนกแม่นยำขึ้น สำหรับ ADASYN ซึ่งเน้นสร้างตัวอย่างเพิ่มเติมในบริเวณที่โมเดลเรียนรู้ได้ยาก จึงช่วยปรับปรุงค่า F1-Score ได้สูงกว่า ส่วน SMOTE ซึ่งสร้างตัวอย่างใหม่ระหว่างข้อมูลเดิม ทำให้ข้อมูลกระจายตัวอย่างสม่ำเสมอและเป็นกลุ่ม ส่งผลให้โมเดลสามารถแยกแยะคลาสได้ดียิ่งขึ้นในแง่ของ AUC

แบบจำลอง Decision Tree ที่ใช้เทคนิค renn_smote ก็มีประสิทธิภาพที่ดีเช่นกัน โดยได้ค่า F1-Score เท่ากับ 0.9690 และ AUC เท่ากับ 0.9749 สะท้อนถึงประสิทธิภาพในการจำแนกข้อมูลได้อย่างแม่นยำ Decision Tree เป็นโมเดลที่ตัดสินใจจากกฎระหว่างคุณสมบัติ ซึ่งมักจะทำงานได้ดีเมื่อข้อมูลไม่ซับซ้อนมากนัก แต่มีข้อจำกัดคือไม่ทนต่อข้อมูลรบกวนและข้อมูลซ้ำซ้อนในระดับสูง การใช้ RENN ซึ่งคัดกรองตัวอย่างที่ทำนายผิดซ้ำ ๆ ออกไป ช่วยลดรบกวนและลดการซ้อนทับระหว่างคลาสได้อย่างมีประสิทธิภาพ ในขณะที่ SMOTE ทำหน้าที่สร้างตัวอย่างใหม่จากข้อมูลเดิมในลักษณะที่ไม่เพิ่มข้อมูลรบกวนมากนักและไม่ก่อให้เกิดข้อมูลซ้ำซ้อน ส่งผลให้แบบจำลองสามารถเรียนรู้ได้อย่างเหมาะสมกับโครงสร้างการตัดสินใจของ Decision Tree

แบบจำลอง XGBoost นั้นจะมีประสิทธิภาพที่เหนือกว่า Decision Tree และ Random Forest เนื่องจากมีการเรียนรู้เป็นแบบลำดับที่ช่วยลดการทำนายที่ผิดผลาดก่อนหน้า ทำให้สามารถเรียนรู้ได้ดีกว่าเมื่อเทียบกับแบบจำลอง Random Forest ที่มีการเรียนรู้แบบสุ่ม ซึ่งเป็นไปตามงานวิจัยของ (Papacharalampous et al., 2021) และในแบบจำลอง Decision Tree ที่เป็นการเรียนรู้โครงสร้างแบบต้นไม้เพียงต้นเดียว ทำให้แบบจำลอง XGBoost เป็นทางเลือกที่เหมาะสมสำหรับชุดข้อมูลนี้

แบบจำลอง Naive Bayes มีผลการทดสอบที่ด้อยกว่าตัวอื่นอย่างชัดเจน โดยได้ค่า F1-Score เท่ากับ 0.6970 และ AUC เท่ากับ 0.7769 เมื่อใช้เทคนิค tomek_adasyn ซึ่งสะท้อนถึงข้อจำกัดของโมเดลนี้ที่สมมติว่าคุณสมบัติทั้งหมดเป็นอิสระต่อกันภายใต้เงื่อนไขของคลาสเดียวกัน ขณะที่ข้อมูลทางการแพทย์ส่วนใหญ่มีความสัมพันธ์ระหว่างคุณสมบัติ เช่น ความดันโลหิต ระดับน้ำตาล และอายุ ทำให้ Naive Bayes ไม่สามารถจับความสัมพันธ์เหล่านี้ได้อย่างเหมาะสม แม้ว่า tomek_adasyn จะช่วยลดข้อมูลที่ทำให้เกิดความยากในการทำนายและลดการทับซ้อนของข้อมูลบริเวณขอบเขตคลาส ทำให้ขอบเขตการจำแนกชัดเจนขึ้น แต่ก็ไม่เพียงพอที่จะชดเชยข้อจำกัดของโมเดล ส่งผลให้ประสิทธิภาพโดยรวมต่ำกว่าแบบจำลองอื่นอย่างเห็นได้ชัด

แบบจำลอง Neural Network แสดงประสิทธิภาพในการเรียนรู้รูปแบบที่ซับซ้อนระหว่างคุณลักษณะต่าง ๆ ภายในแต่ละคลาส โดยเมื่อใช้เทคนิค renn_borderline_smote ได้ค่า F1-Score เท่ากับ 0.9323 และ AUC เท่ากับ 0.9770 ซึ่งเทคนิคนี้เน้นปรับขอบเขตของข้อมูลให้ชัดเจนและเพิ่มความหลากหลายของตัวอย่าง ทำให้โมเดลมีความแม่นยำในการจำแนกสูงกว่า ขณะที่การใช้ tomek_adasyn ให้ค่า F1-Score ลดลงเล็กน้อยเป็น 0.9080 แต่ได้ค่า AUC สูงกว่าอยู่ที่ 0.9885 เนื่องจากการเพิ่มข้อมูลในบริเวณที่โมเดลทำนายยากช่วยให้การแยกแยะคลาสทำได้ดีขึ้น ทั้งนี้ RENN และ Tomek Links ทำหน้าที่ลดข้อมูลรบกวนและความทับซ้อน ช่วยให้โมเดลเรียนรู้จากข้อมูลที่มีความชัดเจนมากขึ้น ขณะที่ ADASYN และ Borderline-SMOTE สร้างตัวอย่างที่หลากหลายกว่าวิธีอื่น ส่งเสริมการเรียนรู้ที่หลากหลายและครอบคลุมมากขึ้น

แบบจำลอง K-Nearest Neighbour (KNN) ทำงานโดยทำนายคลาสจากเพื่อนบ้าน K ตัวที่ใกล้ที่สุดในชุดข้อมูลฝึกสอน โดยเมื่อใช้เทคนิค RENN และ ENN ร่วมกับ Borderline-SMOTE หรือ SMOTE สามารถทำได้ดีโดยมีค่า F1-Score อยู่ที่ 0.9535 และ AUC เท่ากับ 0.9749 เทคนิคเหล่านี้ช่วยลดความซ้ำซ้อนและความทับซ้อนของข้อมูลบริเวณชายขอบคลาส พร้อมทั้งเพิ่มความหลากหลายของข้อมูล ทำให้เกิดกลุ่มตัวแทนเพื่อนบ้านที่ชัดเจนและครอบคลุมมากขึ้น ส่งผลให้ F1-Score สูงขึ้น ขณะที่การใช้ tomek_ros แม้จะมีค่า F1-Score ลดลงเล็กน้อยเป็น 0.9208 แต่ได้ค่า AUC สูงกว่าอยู่ที่ 0.9749 เนื่องจากวิธีนี้ช่วยทำให้ขอบเขตข้อมูลชัดเจนมากขึ้นโดยการลดข้อมูลซ้ำซ้อนและความทับซ้อน แต่ ROSไม่ได้เพิ่มความหลากหลายของข้อมูลใหม่ จึงส่งผลให้อัตราการแยกคลาส AUC ดีขึ้นเล็กน้อยเมื่อเทียบกับวิธีอื่น ๆ

จากการทดลองนี้ พบว่าวิธีการที่ลดข้อมูลที่น่าสนใจคือ Tomek Links, ENN และ RENN ที่ทำงานร่วมกับการเพิ่มข้อมูลด้วย ADASYN, Borderline-SMOTE, ROS และ SMOTE โดยแต่ละเทคนิคมีจุดเด่นและข้อจำกัดที่แตกต่างกัน ซึ่งสามารถเลือกใช้ได้ตามลักษณะของชุดข้อมูลและวัตถุประสงค์ของการวิจัย

Tomek Links ในกรณีที่ต้องการลดข้อมูลที่มีการทับซ้อนระหว่างข้อมูลแต่ละคลาส ควรใช้ Tomek Links เนื่องจากเทคนิคนี้มุ่งเน้นการระบุและกำจัดคู่ข้อมูลที่อยู่ใกล้ชิดกันระหว่างคลาสที่แตกต่างกัน ทำให้สามารถลดความคลุมเครือในการแยกแยะข้อมูลระหว่างคลาส และช่วยให้ขอบเขตการตัดสินใจของแบบจำลองชัดเจนขึ้น

ENN และ RENN หากต้องการลดตัวอย่างโดยเน้นลดตัวอย่างที่มักทำนายผิด ควรเลือกใช้ ENN และ RENN โดย ENN จะกำจัดตัวอย่างที่ถูกจำแนกผิดโดยเพื่อนบ้านใกล้เคียง ส่วน RENN มีข้อดีคือมีการทำซ้ำหลายรอบ ทำให้เหมาะกับข้อมูลที่มีข้อมูลรบกวนปริมาณมาก เนื่องจากการทำซ้ำจะช่วยกำจัดข้อมูลรบกวนได้อย่างทั่วถึงและมีประสิทธิภาพมากขึ้น

ROS สำหรับการเพิ่มข้อมูลที่มีการกระจายตัวดีและต้องการศึกษาข้อมูลระดับพื้นฐาน ควรเลือกใช้ ROS เนื่องจากเป็นวิธีการที่เข้าใจง่ายและไม่ซับซ้อน อย่างไรก็ตาม เนื่องจากเป็นการนำข้อมูลเดิมมาทำซ้ำ จึงทำให้เกิดความเสี่ยงต่อการเรียนรู้ได้ดีเฉพาะกับชุดข้อมูลที่ทำการเรียนรู้ ซึ่งอาจส่งผลต่อประสิทธิภาพในการทำนายข้อมูลใหม่

ADASYN หากต้องการหลีกเลี่ยงการสร้างข้อมูลเดิมเพิ่มขึ้น ให้ใช้ เนื่องจากเทคนิคนี้จะสร้างข้อมูลสังเคราะห์ใหม่โดยมุ่งเน้นไปยังบริเวณที่มีความยากในการเรียนรู้ ทำให้แบบจำลองสามารถเรียนรู้จากข้อมูลที่หลากหลายและซับซ้อนมากขึ้น

Borderline-SMOTE หากต้องการเพิ่มตัวอย่างที่มักทำนายผิดพลาดบริเวณชายขอบ เนื่องจากเป็นการเพิ่มข้อมูลในบริเวณชายขอบของกลุ่มตัวอย่าง ซึ่งเป็นบริเวณที่มีความเสี่ยงสูงในการถูกจำแนกผิด การเพิ่มข้อมูลในบริเวณดังกล่าวจะช่วยเสริมสร้างความสามารถของแบบจำลองในการจัดการกับกรณีที่ท้าทาย

SMOTE สำหรับชุดข้อมูลที่ไม่มีการกระจุกตัวในแต่ละคลาส เพื่อให้ได้ข้อมูลใหม่ที่มีความกระจายตัวสม่ำเสมอ โดย SMOTE จะสร้างตัวอย่างสังเคราะห์ใหม่ในบริเวณระหว่างตัวอย่างที่มีอยู่เดิม ทำให้เกิดการกระจายข้อมูลที่สมดุลมากขึ้น

แม้ว่าการวิจัยครั้งนี้จะประสบความสำเร็จในการพัฒนาแบบจำลองเพื่อทำนายความเสี่ยงของภาวะครรภ์เป็นพิษจากชุดข้อมูลของโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี แต่ยังคงมีข้อจำกัดบางประการที่ควรได้รับการพิจารณาเพื่อการพัฒนาต่อไปในอนาคต ชุดข้อมูลที่ใช้ในการศึกษานี้มาจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารีในช่วงปีพุทธศักราช 2565-2566 เท่านั้น จึงมีปริมาณจำกัดและไม่ครอบคลุมกลุ่มประชากรที่หลากหลาย เนื่องจากข้อจำกัดด้านการเข้าถึงข้อมูลจากกฎระเบียบ PDPA ที่เข้มงวด รวมถึงขั้นตอนการขออนุมัติด้านจริยธรรมที่ซับซ้อนและใช้เวลานาน

ส่งผลต่อความสามารถในการทั่วไปของแบบจำลองเมื่อใช้งานกับกลุ่มประชากรอื่น การเพิ่มความหลากหลายของข้อมูลและการเก็บรวบรวมข้อมูลจากหลายแหล่งจะช่วยยกระดับประสิทธิภาพและความน่าเชื่อถือของแบบจำลองในบริบทที่กว้างขึ้น

สำหรับงานวิจัยในอนาคต ควรมุ่งเน้นการขยายชุดข้อมูลให้ครอบคลุมความหลากหลายของประชากรมากขึ้น รวมถึงพัฒนาแบบจำลองเฉพาะทางที่ใช้เทคนิคอัลกอริทึมหรือโมเดลที่สามารถจัดการกับข้อมูลที่ไม่สมดุลได้อย่างมีประสิทธิภาพ โดยไม่ต้องพึ่งพาการสังเคราะห์ข้อมูลใหม่ได้

บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์หลักสามประการ ประการแรกเพื่อวิเคราะห์ข้อมูลประวัติของสตรีตั้งครรภ์เพื่อหาความสัมพันธ์ของปัจจัยที่ก่อให้เกิดภาวะครรภ์เป็นพิษ ประการที่สองเพื่อศึกษาและหาแนวทางที่เหมาะสมในการจัดการข้อมูลที่ไม่สมดุล เพื่อเตรียมความพร้อมของข้อมูลก่อนการสร้างแบบจำลอง และประการสุดท้ายเพื่อพัฒนาแบบจำลองทำนายความเสี่ยงของการเกิดภาวะครรภ์เป็นพิษในหญิงมีครรภ์ด้วยปัญญาประดิษฐ์

การดำเนินการทดลองในงานวิจัยนี้เป็นไปอย่างสอดคล้องกับวัตถุประสงค์ที่ตั้งไว้ โดยเริ่มจากวัตถุประสงค์แรก จากการศึกษาวรรณกรรมที่เกี่ยวข้อง ร่วมกับการปรึกษาผู้เชี่ยวชาญและวิเคราะห์ข้อมูลจากกลุ่มเสี่ยง พบว่าปัจจัยทางคลินิกหลายประการ เช่น น้ำหนักตัว ความดันโลหิต ภาวะอ้วน ดัชนีมวลกาย (BMI) อุณหภูมิร่างกาย และชีพจร มีความสัมพันธ์อย่างมีนัยสำคัญกับการเกิดภาวะครรภ์เป็นพิษ สอดคล้องกับผลการศึกษาของ (De Kat et al., 2019) ซึ่งแสดงให้เห็นว่าสามารถใช้ข้อมูลประวัติของสตรีตั้งครรภ์วิเคราะห์หาความสัมพันธ์กับความเสี่ยงของภาวะครรภ์เป็นพิษได้อย่างเหมาะสม

เพื่อวัตถุประสงค์ประการที่สองการจัดการข้อมูลที่ไม่สมดุลด้วยวิธี Hybrid Resampling เพื่อเตรียมความพร้อมของข้อมูลก่อนการสร้างแบบจำลองทั้ง 16 รูปแบบ ผลการทดลองแสดงให้เห็นว่าเทคนิค Tomek Links ร่วมกับ Random Oversampling สามารถเพิ่มความสมดุลของข้อมูล ลดข้อมูลรบกวนที่ส่งผลต่อความแม่นยำของแบบจำลองได้อย่างมีประสิทธิภาพ ซึ่งช่วยให้แบบจำลองที่พัฒนาขึ้นมีประสิทธิผลดีขึ้นอย่างชัดเจน ดังนั้น การใช้เทคนิค Tomek Links ร่วมกับ ROS ถือเป็นวิธีที่เหมาะสมและมีประสิทธิภาพสำหรับการจัดการกับปัญหาข้อมูลไม่สมดุลก่อนการสร้างแบบจำลองได้

แบบจำลอง XGBoost ที่ประยุกต์ใช้ร่วมกับเทคนิค Tomek Links ร่วมกับ Random Oversampling สามารถให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า F1-Score เท่ากับ 0.9864 และค่า AUC เท่ากับ 1.0000 สะท้อนให้เห็นถึงความสามารถในการทำนายของแบบจำลองในการทำนายความเสี่ยงของภาวะครรภ์เป็นพิษในหญิงตั้งครรภ์ ด้วยการใช้ปัญญาประดิษฐ์ได้ซึ่งตอบวัตถุประสงค์ประการที่สาม

ชุดข้อมูลที่ใช้ในการวิจัยนี้มีข้อจำกัดในด้านจำนวนแล`ะความหลากหลาย เนื่องจากเก็บรวบรวมจากโรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารีในช่วงปี 2565-2566 เท่านั้น ทำให้ข้อมูลไม่ครอบคลุมกลุ่มประชากรที่หลากหลายทั้งในด้านภูมิศาสตร์และลักษณะประชากรศาสตร์ นอกจากนี้การเข้าถึงข้อมูลยังได้รับผลกระทบจากกฎหมายคุ้มครองข้อมูลส่วนบุคคล (PDPA) และกระบวนการขออนุญาตจริยธรรมที่ซับซ้อนและใช้เวลานาน ซึ่งเป็นอุปสรรคสำคัญในการรวบรวมข้อมูลจำนวนมากและหลากหลายสำหรับการวิจัย การพัฒนางานวิจัยในอนาคตจึงควรมุ่งเน้นการขยายฐานข้อมูลโดยการเก็บข้อมูลจากหลายโรงพยาบาลและในช่วงเวลาที่ยาวนานขึ้น เพื่อเพิ่มความครอบคลุมและความน่าเชื่อถือของแบบจำลอง พร้อมทั้งพัฒนาเทคนิคหรืออัลกอริทึมเฉพาะทางที่สามารถจัดการกับข้อมูลที่ไม่สมดุลได้อย่างมีประสิทธิภาพโดยไม่ต้องพึ่งพาการสร้างข้อมูลสังเคราะห์เพิ่มเติม เพื่อยกระดับความแม่นยำและความยั่งยืนของระบบทำนายในบริบทที่หลากหลายมากขึ้นในอนาคต

รายการอ้างอิง

กรมอนามัย กระทรวงสาธารณสุข. (2025). อัตราส่วนการตายมารดา. สืบค้นจาก https://datacatalog.anamai.moph.go.th/dataset/mchboard_mmr_01

ปวีณา บุตรดีวงศ์. (ม.ป.ป.). "ภาวะครรภ์เป็นพิษ" ความเสี่ยงต่อชีวิตของคุณแม่และทารกในครรภ์. สืบค้นจาก https://www.phyathai.com/th/article/2615-%E0%B8%A0%E0%B8%B2%E0%B8%A7%E0%B8%B0%E0%B8%84%E0%B8%A3%E0%B8%A3%E0%B8%A0%E0%B9%8C%E0%B9%80%E0%B8%9B%E0%B9%87%E0%B8%99%E0%B8%9E%E0%B8%B4%E0%B8%A9_%E0%B8%84%E0%B8%A7%E0%B8%B2

อรวรรณ พินิจเลิศสกุล และ ศุภาวดี วายุเหือด. (2021). ภาวะน้ำหนักเกินและอ้วนในสตรีตั้งครรภ์: บทบาทที่ท้าทายของพยาบาลผดุงครรภ์. วารสารการพยาบาลและการดูแลสุขภาพ, 39(1). สืบค้นจาก https://he01.tci-thaijo.org/index.php/jnat-ned/article/view/247475/168300

สิรยา กิติโยดม. (2018). การศึกษาภาวะชักในสตรีตั้งครรภ์ที่มีภาวะครรภ์เป็นพิษชนิดรุนแรง. วารสารการแพทย์โรงพยาบาลศรีสะเกษ สุรินทร์ บุรีรัมย์, 29(3), 129–138. สืบค้นจาก https://he02.tci-thaijo.org/index.php/MJSSBH/article/view/122515

เกษมศรี ศรีสุพรรณดิฐ. (ม.ป.ป.). ครรภ์เป็นพิษและโรคพิษแห่งครรภ์ระยะชัก: Preeclampsia with Severe Features and Eclampsia. สืบค้นจาก https://w1.med.cmu.ac.th/obgyn/lessons/33344/

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 20-29. https://doi.org/10.1145/1007730.1007735

Bhandari, S. M., & Patel, K. (2015). A review on using clustering and classification techniques to predict student failure with high dimensional and imbalanced data.

Briceño-Pérez, C., Briceño-Sanabria, L., & Vigil-De Gracia, P. (2009). Prediction and Prevention of Preeclampsia. Hypertension in Pregnancy, 138–155. https://doi.org/10.1080/10641950802022384

Broder, A. Z., Bruckstein, A. M., & Koplowitz, J. (1985). On the performance of edited nearest neighbor rules in high dimensions. IEEE Transactions on Systems, Man, and Cybernetics, 136-139. https://doi.org/10.1109/TSMC.1985.6313401

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

De Kat, A. C., Hirst, J., Woodward, M., Kennedy, S., & Peters, S. A. (2019). Prediction models for preeclampsia: A systematic review. Pregnancy Hypertension, 48–66. https://doi.org/10.1016/j.preghy.2019.03.005

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2018.12.003

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing (pp. 483–490). Springer.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. In IEEE 2008 Fourth International Conference on Natural Computation (pp. 192–201). https://doi.org/10.1109/ICNC.2008.871

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), Advances in Intelligent Computing. ICIC 2005 (Vol. 3644). https://doi.org/10.1007/11538059_91

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 1322-1328). https://doi.org/10.1109/IJCNN.2008.4633969

Herndon, N., & Caragea, D. (2016). A Study of Domain Adaptation Classifiers Derived From Logistic Regression for the Task of Splice Site Prediction. IEEE Transactions on NanoBioscience, 75-83. https://doi.org/10.1109/TNB.2016.2522400

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In Proceedings of the International Conference on Neural Information Processing (pp. 152–159). Springer.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys. https://doi.org/10.1145/3343440

Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., & Ouzzani, M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. Machine Learning, 465–482. https://doi.org/10.1007/s10994-015-5535-7

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access. https://doi.org/10.1109/access.2021.3102399

Krawczyk, B., Wo**ź**niak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing, 554-562. https://doi.org/10.1016/j.asoc.2013.08.014

Li, Q., & Mao, Y. (2014). A review of boosting methods for imbalanced data classification. Pattern Analysis and Applications, 679–693. https://doi.org/10.1007/s10044-014-0392-8

Lim, K.-H. (2022). Preeclampsia. สืบค้นจาก https://emedicine.medscape.com/article/1476919-overview?form=fpf

Liu, M., Wang, M., Wang, J., & Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical, 177, 970–980.

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. https://doi.org/10.48550/arXiv.1305.1707

Majid, A., Ali, S., Iqbal, M., & Kausar, N. (2014). Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Computer Methods and Programs in Biomedicine, 113(3), 792–808.

Noorhalim, N., Ali, A., & Shamsuddin, S. M. (2019). Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. In Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017) (pp. 19-43). https://doi.org/10.1007/978-981-13-7279-7

Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2131-2140. https://doi.org/10.1109/TCBB.2019.2911071

Papacharalampous, G., Tyralis, H., Doulamis, A., & Doulamis, N. (2021). Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. Hydrology. https://doi.org/10.3390/hydrology10020050

Park, Y., & Ghosh, J. (2014). Ensembles of (α)-Trees for Imbalanced Classification Problems. IEEE Transactions on Knowledge and Data Engineering, 131-143. https://doi.org/10.1109/TKDE.2012.255

Patel, H., & Thakur, G. S. (2016). A hybrid weighted nearest neighbor approach to mine imbalanced data. In Proceedings of the International Conference on Data Mining (DMIN'16) (p. 106). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In 2015 IEEE International Conference on Information Reuse and Integration (pp. 197-202). https://doi.org/10.1109/IRI.2015.39

Santiso, S., Casillas, A., & Pérez, A. (2019). The class imbalance problem detecting adverse drug reactions in electronic health records. Health Informatics Journal, 1768-1778. https://doi.org/10.1177/14604582187994

Sanz, J. A., Bernardo, D., Herrera, F., Bustince, H., & Hagras, H. (2015). A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications With Imbalanced Data. IEEE Transactions on Fuzzy Systems, 973-990. https://doi.org/10.1109/TFUZZ.2014.2336263

Tomek, I. (1976). Two modifications of CNN.

Vata, P. K., Chauhan, N., Nallathambi, A., & Hussein, F. E. R. (2015). Assessment of prevalence of preeclampsia from Dilla region of Ethiopia. https://doi.org/10.1186/s13104-015-1821-5

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics, 408-421. https://doi.org/10.1109/TSMC.1972.4309137

World Health Organization. (2025). Pre-eclampsia. สืบค้นจาก https://www-who-int.translate.goog/news-room/fact-sheets/detail/pre-eclampsia?_x_tr_sl=en&_x_tr_tl=th&_x_tr_hl=th&_x_tr_pto=tc

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S.-S. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems, 105–116. https://doi.org/10.1016/j.knosys.2014.06.004

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In Proceedings of the 1st International Conference on Advanced Data and Information Engineering (pp. 13–22). https://doi.org/10.1007/978-981-4585-18-7_2

ภาคผนวก ก

ผลงานที่ตีพิมพ์ในระหว่างการศึกษา

ผลงานที่ตีพิมพ์ในระหว่างการศึกษา

Panjainam, P., & Kanjanawattana, S. (2024). A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data. ICRSA 2024: 2024 7th International Conference on Robot Systems and Applications, Bangkok, Thailand. 12 – 14 September 2024, 5 Pages, https://doi.org/10.1145/3702468.3702477

A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data

Paonrat Panjainam
School of Computer Engineering Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
m6500979@g.sut.ac.th

Sarunya Kanjanawattana
School of Computer Engineering Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
sarunya.k@sut.ac.th

Abstract

Extremely severe preeclampsia is a disorder that emerges during pregnancy and is defined by the development of high blood pressure; both the mother and the fetus may perish as a result. However, the number of preeclampsia patients is substantially fewer in terms of statistics compared to that in a typical pregnancy. Thus, the uneven data offers a barrier to the building of highly effective machine learning models. This study attempted to overcome the problem of data imbalance in the preeclampsia dataset. Four unique undersampling algorithms were implemented: Random Undersampling (RUS), Tomek Link (Tomek), Edited Nearest Neighbors (ENN), and Repeated Edited Nearest Neighbors (RENN). Similarly, four other oversampling procedures were employed: Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE, Synthetic Minority Oversampling Technique (SMOTE), and Random Oversampling (ROS). In order to investigate the balanced data, Decision Tree, Naive Bayes, Random Forest, and XGBoost were deployed. According to the experimental findings, the XGBoost model, which employed the hybrid resampling technique Tomek and ROS, demonstrated the greatest degree of efficacy, as indicated by an AUC analysis.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Machine learning approaches*; *Bio-inspired approaches*; Generative and developmental approaches.

Keywords

Data imbalance, Data resampling, Undersampling, Preeclampsia data, Oversampling

ACM Reference Format:
Paonrat Panjainam and Sarunya Kanjanawattana. 2024. A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data. In *2024 7th International Conference on Robot Systems and Applications (ICRSA) (ICRSA 2024), September 12–14, 2024, Bangkok, Thailand.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3702468.3702477

ICRSA 2024, September 12–14, 2024, Bangkok, Thailand

ACM ISBN 979-8-4007-1703-1/24/09
https://doi.org/10.1145/3702468.3702477

1 Introduction

Currently, preeclampsia is a specific disease occurring during pregnancy that involves the development of high blood pressure during pregnancy, which is very severe, resulting in death for the pregnant woman and the fetus. The initial symptoms of this condition in pregnant women are protein leakage in the urine and elevated blood pressure. Additionally, close monitoring and attention should be given to this confluence of conditions. Epilepsy, which is directly associated with hypertension during pregnancy, may occur in severe cases of symptoms. A preeclampsia disorder affects both the mother and the fetus in terms of health. Additionally, mortality may be a possibility. It tends to occur after the 20th week of pregnancy, even in the absence of medical documentation indicating hypertension which, the systolic and diastolic blood pressures constantly exceed 140 mmHg and 90 mmHg, respectively. In addition, the presence of protein in the urine is a critical symptom. Hypertension is considered a notable indicator of preeclampsia. Globally, preeclampsia claims the lives of approximately 76,000 women and 500,000 infants across the globe. Eight to ten percent of pregnancies are impacted[25].

There are several approaches to mitigating the severity of preeclampsia, including the division of high-risk groups that necessitate constant monitoring to prevent the condition during pregnancy. The following symptoms have been identified as representing the high-risk category: Pregnant women who experience symptoms of epigastric tightness, blurred vision, pain, swelling, high blood pressure, protein in the urine, 40-year-old-over pregnant women, having medical history of preeclampsia, or have congenital disorders including to obesity, hypertension, renal dysfunction, diabetes, immunodeficiency, or abnormal weight gain.

To investigate the current state of research concerning the risk of preeclampsia, we examined artificial intelligence methods utilized to predict the occurrence of preeclampsia in expectant women, as well as prevention guidelines that account for potential preeclampsia initiating factors [5]. As anticipated, artificial intelligence technology has taken on a progressively significant role in addressing numerous challenges within the context of medical science by helping in decision-making and effective treatment planning [3].

The effectiveness of artificial intelligence is dependent on the quality of the data. In addressing the issue of unbalanced data, the unbalanced distribution of information poses an intriguing challenge because if the quantity and quality of the data are sufficient to discern patterns within the data, then a machine learning model may exhibit respectable performance[6]. Conversely, the model may have lacked sufficient intelligence to learn and classify data for only a small portion of the collection [15]. Typically, significant

imbalances occur in data in the actual world. Imbalanced datasets frequently exhibit a relatively small population size for the class of interest in comparison to the other classes, leading to substantial inconsistencies in the data and disparities in proportions between classes, especially for medical and health data [7].

In this study, our objective was to rectify the problem of imbalanced data pertaining to the risk of preeclampsia through the implementation of hybrid, undersampling, and oversampling techniques. Additionally, an assessment was conducted on the efficacy of specific machine learning models, including Extreme Gradient Boosting (XGBoost), Naive Bayes, Decision Tree, and Random Forest. As our data characteristic, there were three distinct groups in the datasets: normal, risk, and preeclampsia. To consider how data unbalancing, the data ratio between the majority data sample represented by the normal class and the minority data sample represented by the risk class was measured roughly 76. Moreover, a data ratio between the majority data samples (the risk group) and a minority of data (Preeclampsia group) is roughly 14. The calculated ratios were explicitly defined as imbalanced data [13].

2 Literature Reviews

2.1 Data Pre-processing

Data Pre-processing addressing the issue of unbalanced data during data distribution presents an intriguing challenge [6]. Too small portion of data directly impacts to the quality of machine learning models because the models learn the pattern from the majority of data [15]. To effectively implement the sampling method, the unbalanced data should be transformed into numerical data prior to addressing the problem [13].

Our unbalanced dataset consisted of three groups: 9,341 records for the normal group, 123 records for the high-risk group, and 9 records for the preeclampsia group. To guarantee an enhancement in the quality of our dataset, it was therefore essential to implement the potential data unbalancing technique. Our approach involved boosting the proportion of the small data set, such as the preeclampsia group, while decreasing the size of the large data set, such as the normal group [16]. Subsequently, the quantity of data in each class should be approximately equivalent to that of the high-risk group. Eventually, each cohort comprised a substantial quantity of data.

As an indicator of data imbalance, the data imbalance ratio compares the proportions of the majority and minority data groups [13]. The ratio is ranged between zero to one. The nearly one ratio means data more balanced and vice versa [17]. As previously stated, medical science datasets were more susceptible to the problem of data imbalance than those of academic fields. A review of several prior studies led us to the conclusion that they had addressed the issue at hand, such as the diagnosis of lung cancer [13], genetic material identification [10], medical adverse effects [22], and so forth.

The oversampling methods were as follows: Random Oversampling (ROS), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Over-sampling Technique (SMOTE), and Borderline-SMOTE.

(1) ROS is an early data analysis technique that involves the random addition of minority class samples to the data [1].
(2) ADASYN divides minority class data samples according to learning difficulty and employs a weighted distribution. A smaller subset of the data in a group serves to generate a majority sample, which is a more challenging set to grasp [9].
(3) SMOTE uses KNN to pre-synthesize estimates on the minority data [4].
(4) Borderline-SMOTE applies SMOTE to data on the borderline of the data samples frequently misclassified as demographic clusters [8].

The undersampling methods are as follows: Random Undersampling (RUS), Tomek Link (Tomek), Edited Nearest Neighbors (ENN), and Repeated Edited Nearest Neighbors (RENN).

(1) RUS is the earliest technique since its inception for undersampling. By doing so, the random sample is diminished to encompass the majority group [21].
(2) Tomek evaluates two data samples as potentially related if they are adjacent and of distinct data types. Error-containing (Error data) or noisy (Noise data) data, as well as samples from the majority of data categories, are eliminated [24].
(3) ENN was applied to every sample that was evaluated by KNN on the remaining sample data. Erroneously classified segments of sample data are discarded. Additionally, the corrected data set comprises the remaining samples [26].
(4) RENN is the iterative application of ENN until the elimination of the untrainable data set no longer has an impact [26].

Hybrid resampling methods, which combine and pair undersampling of majority data and oversampling of minority data, were utilized to increase the population of minority groups while decreasing the population of majority data.

2.2 Classifiers

The study applied four machine learning classification methods, namely Decision Tree, Naive Bayes, Random Forest, and XGBoost, to develop models for predicting unbalanced medical data. The performance of the models was evaluated using the Area Under the Curve (AUC) measure, as defined by [11, 19].

The Decision Tree algorithm is a powerful tool for tackling classification problems. Krawczyk et al. [14] applied Decision Trees to detect data with imbalanced characteristics, a prominent trait in actual datasets. In addition, Sanz et al. [23] deployed Decision Trees on 11 datasets to successfully solve the issue of imbalanced data across several categories, yielding promising experimental outcomes [20].

Naive Bayes, as studied by Yap et al. [28], did a comparative investigation of machine learning algorithms to determine the most effective technique for managing imbalanced datasets. In their study, Bhandari et al. [2] applied the Naive Bayes algorithm to evaluate datasets with a high number of dimensions and imbalanced data. They demonstrated the usefulness of Naive Bayes in reliably recognizing different populations within these datasets. Moreover, Naive Bayes is widely seen in the literature while dealing with the classification of medical research data [11].

The Random Forest technique, introduced by Wu et al. [27], has proven to be a flexible and successful solution for text classification

tasks that deal with imbalanced data. Khushi et al. [13] confirmed its higher performance in a lung cancer diagnostic trial compared to competing approaches. In their research work, Khabsa et al. [12] suggested adopting Random Forest as a good strategy for categorizing imbalanced data.

XGBoost, as noted by Papacharalampous et al. [19], demonstrated great performance as compared to other tree-based classification algorithms when applied to rainfall datasets. Ogunleye et al. [18] suggested the application of XGBoost for the diagnosis of chronic renal sickness, offering additional proof of its usefulness in dealing with diverse types of datasets.

3 Methodology

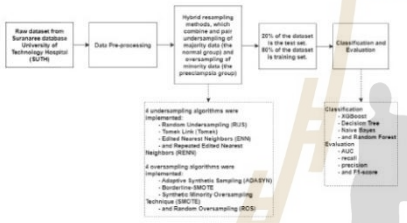


Figure 1: The framework of methodology

Figure 1 presents the methods and steps of the methodology. The purpose of this research was to find the appropriate resampling method to deal with imbalanced data sets related to preeclampsia in pregnancy. These datasets from the database of Suranaree University of Technology Hospital (SUTH) have three classes: the normal group, the risk group, and the preeclampsia group. As indicated by the data ratio, the distribution of the data in this set was unbalanced. Hence, the risk of preeclampsia, as assessed by the four chosen classifiers, can be predicted through the implementation of data imbalance methods.

3.1 Data Pre-processing

Data preparation required various actions, including converting the data into a suitable format for analysis, arranging the data for inquiry, modifying data attributes, and addressing missing data. The goal of the tests was to create models and determine the most effective approach for efficiently addressing data imbalances. Significant data items were acquired from patient histories, such as pregnancy age, number of pregnancies, and symptoms including headache, impaired vision, swollen tongue, and general edema. The information was separated into three independent groups: normal, at-risk, and preeclampsia. Notably, to determine the danger group, we carefully distinguished those normal pregnancy groups that incorporate damaging material. A case of a mother enduring hypertension without being diagnosed with preeclampsia is one example. Her disease necessitates more regular monitoring by doctors compared to a normal pregnancy, which supports her identification as a high-risk person.

The at-risk group comprised patients who possessed significant attributes, including raised blood pressure and protein in the urine, headache, swollen tongue, blurred vision, pre-existing pregnancies characterized by such symptoms, patients aged 40 and above, a prior diagnosis of preeclampsia, chronic illnesses including hypertension, kidney disease, diabetes, autoimmune disorders, obesity, abnormal weight gain in comparison to pregnancy weight gain standards, and fatigue. The patient became immediately classed into the risk group in the event that any symptoms showing a sign of the normal pregnancy group were detected.

Then, the data transformation was conducted. We changed categorical data to numerical data to prepare the process of data balance. This study utilized four separate undersampling techniques: RUS, Tomek, ENN, RENN, as well as four distinct oversampling techniques: Borderline-SMOTE, ROS, ADASYN, SMOTE.

3.2 Classification and Evaluation

To assess the efficacy of our resampling approaches, we choose four classification techniques: Decision Tree, Naive Bayes, Random Forest, and XGBoost. We deployed each model using its default parameter settings to preserve a uniform basis for comparison. The study employed these algorithms to validate which method of data unbalancing handling was suitable for imbalanced medical data.

This investigation implements the Area Under the Curve (AUC) as a metric to assess the performance of the classifier. A relationship between the True Positive Rate (TPR) and False Positive Rate, which ranges from 0 to 1, is expressed by the area under the curve. Effectiveness of the model increases as the value approaches 1. AUC is applicable to model performance evaluation in unbalanced data sets [11, 19]. Furthermore, the effectiveness of the models is evaluated through the utilization of recall, precision, and F1-score metrics.

4 Results

Naïve Bayes, the tomek_adasyn had the best performance. These are shown in Table 1, which produced the following superior values for AUC, F1-score, Recall, and Precision: 0.7769, 0.6970, 0.7325, and 0.7564, respectively.

Decision Tree, the renn_smote had the best performance in AUC and Precision, enn_adasyn. The renn_adasyn had the best performance in F1-score and Recall. These are shown in Table 2, which resulted in the following more effective values for AUC, F1-score, Recall, and Precision: 0.9749, 0.9688, 0.9722, and 0.9697, respectively.

Random Forest, the tomek_smote had the most significant performance in AUC, F1-score, and Recall. The tomek_adasyn had the best result in Precision. These are shown in Table 3, which produced the following superior values for AUC, F1-score, Recall, and Precision: 0.9929, 0.9730, 0.9785, and 0.9706, respectively.

XGBoost, the tomek_ros had the best performance. These are shown in Table 4, which produced the following superior values for AUC, F1-score, Recall, and Precision: 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

Table 1: Results for Naïve Bayes on test set

Model	Technique	AUC	F1-score	Recall	Precision
Naive Bayes	enn_adasyn	0.7399	0.6692	0.6952	0.7183
	enn_borderline_smote	0.7442	0.6587	0.6616	0.6615
	enn_ros	0.7442	0.6587	0.6616	0.6615
	enn_smote	0.7442	0.6587	0.6616	0.6615
	renn_adasyn	0.7399	0.6692	0.6952	0.7183
	renn_borderline_smote	0.7442	0.6587	0.6616	0.6615
	renn_ros	0.7442	0.6587	0.6616	0.6615
	renn_smote	0.7442	0.6587	0.6616	0.6615
	ras_adasyn	0.4907	0.4641	0.4783	0.4803
	ras_borderline_smote	0.5439	0.2922	0.3484	0.2530
	ras_ros	0.5715	0.3906	0.3977	0.3982
	ras_smote	0.4914	0.3930	0.3880	0.4049
	tomek_adasyn	**0.7769**	**0.6970**	**0.7325**	**0.7564**
	tomek_borderline_smote	0.7082	0.6727	0.6839	0.6834
	tomek_ros	0.7363	0.5211	0.6667	0.4641
	tomek_smote	0.7423	0.6686	0.7075	0.7156

Table 2: Results for Decision Tree on test set

Model	Technique	AUC	F1-score	Recall	Precision
Decision Tree	enn_adasyn	0.9747	0.9688	**0.9722**	0.9683
	enn_borderline_smote	0.9522	0.9380	0.9420	0.9444
	enn_ros	0.9633	0.9535	0.9565	0.9565
	enn_smote	0.9633	0.9535	0.9565	0.9565
	renn_adasyn	0.9747	0.9688	**0.9722**	0.9683
	renn_borderline_smote	0.9633	0.9535	0.9565	0.9565
	renn_ros	0.9417	0.9223	0.9275	0.9333
	renn_smote	**0.9749**	**0.9690**	0.9710	**0.9697**
	ras_adasyn	0.8801	0.8565	0.8522	0.8629
	ras_borderline_smote	0.8171	0.7823	0.7851	0.7816
	ras_ros	0.8769	0.8473	0.8555	0.8488
	ras_smote	0.8464	0.8135	0.8185	0.8157
	tomek_adasyn	0.959	0.950	0.950	0.950
	tomek_borderline_smote	0.9208	0.9048	0.9070	0.9032
	tomek_ros	0.9689	0.9598	0.9677	0.9565
	tomek_smote	0.9335	0.9190	0.9237	0.9164

Table 3: Results for Random Forest on test set

Model	Technique	AUC	F1-score	Recall	Precision
Random Forest	enn_adasyn	0.9876	0.9383	0.9371	0.9400
	enn_borderline_smote	0.9806	0.9556	0.9565	0.9593
	enn_ros	0.9795	0.9556	0.9565	0.9593
	enn_smote	0.9818	0.9556	0.9565	0.9593
	renn_adasyn	0.9900	0.9546	0.9547	0.9562
	renn_borderline_smote	0.9793	0.9556	0.9565	0.9593
	renn_ros	0.9780	0.9431	0.9458	0.9431
	renn_smote	0.9805	0.9431	0.9458	0.9431
	ras_adasyn	0.9609	0.9120	0.9088	0.9171
	ras_borderline_smote	0.9534	0.9005	0.9001	0.9046
	ras_ros	0.9636	0.9157	0.9168	0.8902
	ras_smote	0.9683	0.9157	0.9168	0.9219
	tomek_adasyn	0.9892	0.9624	0.9583	**0.9706**
	tomek_borderline_smote	0.9889	0.9592	0.9618	0.9571
	tomek_ros	0.9859	0.9603	0.9640	0.9586
	tomek_smote	**0.9929**	**0.9730**	**0.9785**	0.9697

Table 4: Results for XGBoost on test set

Model	Technique	AUC	F1-score	Recall	Precision
XGBoost	enn_adasyn	0.9779	0.9531	0.9547	0.9522
	enn_borderline_smote	0.9888	0.9535	0.9565	0.9565
	enn_ros	0.9888	0.9535	0.9565	0.9565
	enn_smote	0.9888	0.9535	0.9565	0.9565
	renn_adasyn	0.9779	0.9531	0.9547	0.9522
	renn_borderline_smote	0.9888	0.9535	0.9565	0.9565
	renn_ros	0.9888	0.9535	0.9565	0.9565
	renn_smote	0.9888	0.9535	0.9565	0.9565
	ras_adasyn	0.9530	0.8995	0.9017	0.8999
	ras_borderline_smote	0.9358	0.9038	0.8995	0.9144
	ras_ros	0.9683	0.8868	0.8915	0.9219
	ras_smote	0.9410	0.8934	0.8947	0.8954
	tomek_adasyn	0.9953	0.9751	0.9722	0.9798
	tomek_borderline_smote	0.9806	0.9720	0.9667	0.9798
	tomek_ros	**1.0000**	**0.9864**	**0.9892**	**0.9841**
	tomek_smote	0.9934	0.9720	0.9667	0.9798

5 Discussion

Based on the actual data developed by employing the Naive Bayes classifier, it was concluded that the tomek_adasyn model offered higher values for AUC, F1-score, Recall, and Precision: 0.7769, 0.6970, 0.7325, and 0.7564, respectively. In the instance of the decision tree, the renn_adasyn and enn_adasyn processes yielded the largest feasible F1-score and Recall values, which were 0.9688 and 0.9722, respectively. The optimum values for both AUC and Precision have been determined using the renn_smote data preparation method, resulting in values of 0.9697 and 0.9749, respectively.

When processing data in the tomek_smote format, the best results for Random Forest AUC, F1-score, and Recall were 0.9929, 0.9730, and 0.9785, respectively. Using the tomek_adasyn format, the highest Precision value of 0.9706 was reached. The most suitable values for AUC, F1-score, Recall, and Precision for XGBoost were obtained using the tomek_ros data preparation approach, with values of 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

From this data, it can be claimed that the Naïve Bayes classifier offers pretty average results when compared with other models. Naïve Bayes performs well on data with independent features, which is rare in medical datasets where qualities are highly correlated. Conversely, it performs well with Decision Tree, Random Forest, and XGBoost models.

Decision Trees typically overfit owing to the over-memorization of training data, rendering them unable to accurately forecast incoming input, especially when dealing with complex datasets or uncontrolled tree depth. Random Forest, based on a bagging ensemble technique, reduces overfitting and increases model robustness by training each tree on unique subsets of features and a random sample of data. XGBoost, a boosting ensemble technique, exceeds both Decision Trees and Random Forest thanks to its superior optimization approaches, regularization, and handling of challenging datasets. Iteratively improving model predictions, XGBoost generates decision trees sequentially, with each tree focused on repairing the faults committed by the preceding ones. This extensive development makes it extremely accurate and robust.

6 Conclusion

This study tries to uncover the correct technique to handle the unequal clinical dataset related with preeclampsia. The study used four independent undersampling strategies (RUS, Tomek, ENN, RENN) and four distinct oversampling approaches (ADASYN, SMOTE, Borderline-SMOTE). An assessment of the efficacy of four prediction models was conducted: XGBoost, Decision Tree, Naive Bayes, and Random Forest. Each model was assessed using AUC, F1-score, Recall, and Precision. The best AUC, F1-score, Recall, and Precision

values for the dataset attained with tomek_ros and the XGBoost model were 1.0000, 0.9864, 0.9892, and 0.9841, respectively.

However, some of the limitations of the study include; the use of one dataset and therefore the study can only be applied narrowly. As for the future studies, the authors suggest including different dataset clarifying performance with cross-validation testing and using demographics or medical situations different from the current study. It is necessary to derive new treatments on the algorithm level like cost-sensitive learning or ensemble learning to evade biases and constraints of data level approaches. Thus, comparative analysis of the data level and algorithm level might allow in the future to make a concerted system in the form of a combined approach. Thus, future studies must broaden the variety of methods and data sets, which will enhance the robustness and practicality of resampling approaches in medical data analysis.

Acknowledgments

This work was supported by (i) Suranaree University of Technology (SUT), (ii) Thailand Science Research and Innovation (TSRI), (iii) National Science, and Research and Innovation Fund (NSRF) (NRIIS Project Number 179264). This work has been improved grammatically and organizationally through the use of QuillBot, ChatGPT, and Grammarly. The author expresses gratitude for the help of these tools, which contribute to improving the overall quality of this work.

References

[1] G. E. Batista, R. C. Prati, and M. C. Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* (2004), 20–29. https://doi.org/10.1145/1007730.1007735

[2] Surya Mukesh Bhandari and Kunal Patel. 2015. A review on using clustering and classification techniques to predict student failure with high dimensional and imbalanced data. (2015).

[3] Carlos Briceño-Pérez, Liliana Briceño-Sanabria, and Paulino Vigil-De Gracia. 2009. Prediction and Prevention of Preeclampsia. *Hypertension in Pregnancy* (2009), 138–155. https://doi.org/10.1080/10641950802022384

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* (2002), 321–357. https://doi.org/10.1613/jair.953

[5] A. C. De Kat, J. Hirst, M. Woodward, S. Kennedy, and S. A. Peters. 2019. Prediction models for preeclampsia: A systematic review. *Pregnancy Hypertension* (2019), 48–66. https://doi.org/10.1016/j.preghy.2019.03.005

[6] S. Fotouhi, S. Asadi, and M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics* (2019). https://doi.org/10.1016/j.jbi.2018.12.003 Advance online publication.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. 2008. On the Class Imbalance Problem. In *IEEE 2008 Fourth International Conference on Natural Computation*. 192–201. https://doi.org/10.1109/ICNC.2008.871

[8] H. Han, W. Y. Wang, and B. H. Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing. ICIC 2005*, D. S. Huang, X. P. Zhang, and G. B. Huang (Eds.). 3644. https://doi.org/10.1007/11538059_91

[9] H. He, Y. Bai, E. A. Garcia, and S. Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[10] N. Herndon and D. Caragea. 2016. A Study of Domain Adaptation Classifiers Derived From Logistic Regression for the Task of Splice Site Prediction. *IEEE Transactions on NanoBioscience* (2016), 75–83. https://doi.org/10.1109/TNB.2016.2522400

[11] Harpreet Kaur, Himanshu S. Pannu, and Amandeep K. Malhi. 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *Comput. Surveys* (2019). https://doi.org/10.1145/3343440

[12] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning* (2016), 465–482. https://doi.org/10.1007/s10994-015-5535-7

[13] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes. 2021. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* (2021). https://doi.org/10.1109/access.2021.3102399

[14] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing* (2014), 554–562. https://doi.org/10.1016/j.asoc.2013.08.014

[15] Q. Li and Y. Mao. 2014. A review of boosting methods for imbalanced data classification. *Pattern Analysis and Applications* (2014), 679–693. https://doi.org/10.1007/s10044-014-0392-8

[16] R. Longadge and S. Dongre. 2013. Class imbalance problem in data mining review. (2013). https://doi.org/10.48550/arXiv.1305.1707

[17] N. Noorhalim, A. Ali, and S. M. Shamsuddin. 2019. Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. In *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. 19–43. https://doi.org/10.1007/978-981-13-7279-7

[18] Ayokunle Ogunleye and Qian-Gen Wang. 2020. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

[19] Georgios Papacharalampous, Hristos Tyralis, Anastasios Doulamis, and Nikolaos Doulamis. 2021. Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. *Hydrology* (2021). https://doi.org/10.3390/hydrology10020050

[20] Yoonsik Park and Joydeep Ghosh. 2014. Ensembles of (α)-Trees for Imbalanced Classification Problems. *IEEE Transactions on Knowledge and Data Engineering* (2014), 131–143. https://doi.org/10.1109/TKDE.2012.255

[21] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano. 2015. Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In *2015 IEEE International Conference on Information Reuse and Integration*. 197–202. https://doi.org/10.1109/IRI.2015.39

[22] S. Santiso, A. Casillas, and A. Pérez. 2019. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Informatics Journal* (2019), 1768–1778. https://doi.org/10.1177/1460458218799470

[23] Javier Alfonso Sanz, Daniel Bernardo, Francisco Herrera, Humberto Bustince, and Hani Hagras. 2015. A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications With Imbalanced Data. *IEEE Transactions on Fuzzy Systems* (2015), 973–990. https://doi.org/10.1109/TFUZZ.2014.2336263

[24] I. Tomek, A. Z. Broder, A. M. Bruckstein, and J. Koplowitz. 1985. Two modifications of CNN. *On the performance of edited nearest neighbor rules in high dimensions* (1985), 136–139. https://doi.org/10.1109/TSMC.1985.6313401

[25] P. K. Vata, N. Chauhan, A. Nallathambi, and F. E. R. Hussein. 2015. Assessment of prevalence of preeclampsia from Dilla region of Ethiopia. (2015). https://doi.org/10.1186/s13104-015-1821-5

[26] D. L. Wilson. 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics* (1972), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

[27] Qinglong Wu, Yuhang Ye, Huan Zhang, Michael K. Ng, and Su-En Ho. 2014. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems* (2014), 105–116. https://doi.org/10.1016/j.knosys.2014.06.004

[28] Bee Wah Yap, Khairul Azmi Abu Rani, Hani Al-Faris Abd Rahman, Simon Fong, Zuraini Khairudin, and Nor Aniza Abdullah. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the 1st International Conference on Advanced Data and Information Engineering*. 13–22. https://doi.org/10.1007/978-981-4585-18-7_2

ประวัติผู้เขียน

นางสาวปอรรัชม์ ปานใจนาม เกิดเมื่อวันที่ 22 มกราคม พ.ศ. 2543 ที่จังหวัดบุรีรัมย์ประเทศไทย เขาสำเร็จการศึกษาระดับประถมศึกษาจากโรงเรียนประโคนชัยวิทยาและสำเร็จการศึกษาระดับมัธยมศึกษาตอนปลายจากโรงเรียนบุรีรัมย์พิทยาคม เขาได้รับปริญญาตรีวิศวกรรมศาสตร์บัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ จากมหาวิทยาลัยเทคโนโลยีสุรนารี จากนั้นในปี พ.ศ. 2565 เขาได้ศึกษาต่อในระดับปริญญาโท สาขาวิชาวิศวกรรมคอมพิวเตอร์ ณ สำนักวิชาวิศวกรรมคอมพิวเตอร์ สถาบันวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา ประเทศไทย