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