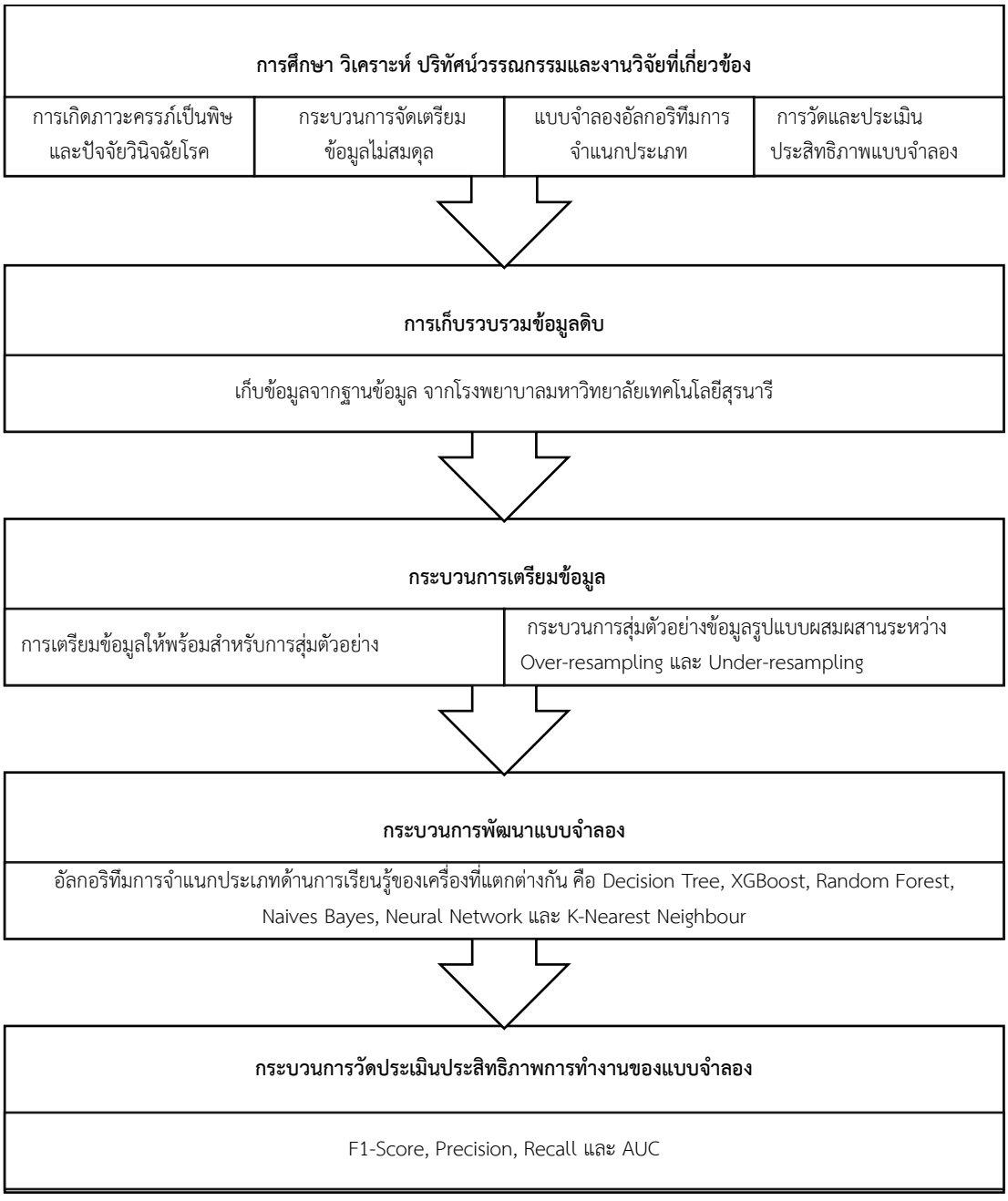
บทที่ 3
วิธีดำเนินการวิจัย

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%

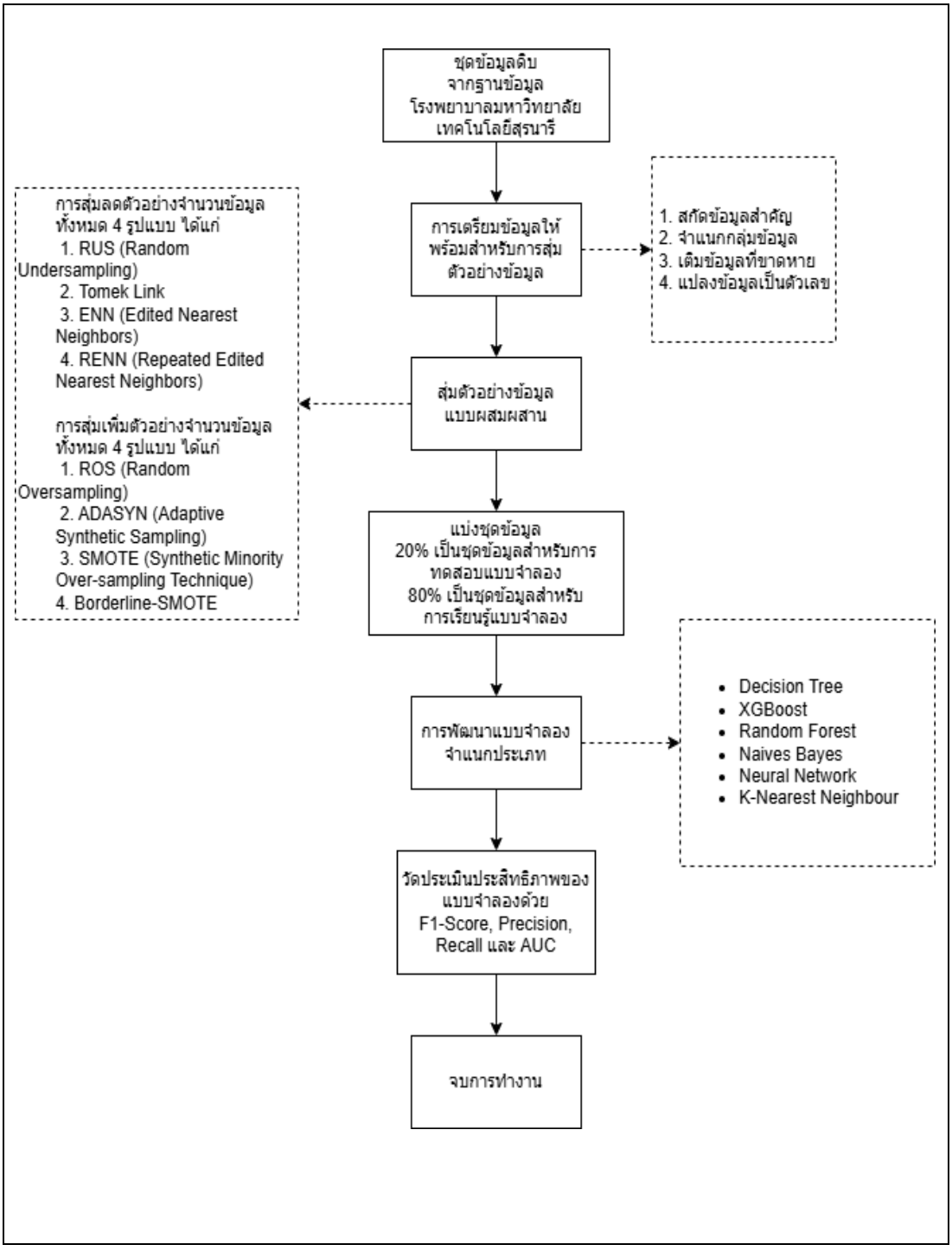
ชื่อคอลัมน์ที่เหลือ	**เปอร์เซ็นต์ข้อมูลที่หายไป (%)**
น้ำหนักตัว	20.0928
ค่าความดันโลหิต	20.0206
ภาวะอ้วน	19.5361
ดัชนีมวลกาย (BMI)	19.5361
อุณหภูมิร่างกาย	19.3608
ชีพจร	18.1443
ส่วนสูง	12.9381

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

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