



แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตัน
ในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง
AUTOMATIC DIAGNOSIS MODEL FOR RISK OF SYMPTOMATIC VENOUS
THROMBOEMBOLISM BASED ON MACHINE LEARNING

อัจฉราภรณ์ สุขเพิ่ม

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดอุดตัน
ในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง



สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

AUTOMATIC DIAGNOSIS MODEL FOR RISK OF SYMPTOMATIC VENOUS
THROMBOEMBOLISM BASED ON MACHINE LEARNING



A Master's Project Submitted in Partial Fulfillment of the Requirements
for the Degree of MASTER OF SCIENCE
(Data Science)

Faculty of Science, Srinakharinwirot University

2020

Copyright of Srinakharinwirot University

สารนิพนธ์

เรื่อง

แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดอุดตัน

ในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง

ของ

อัจฉราภรณ์ สุขเพิ่ม

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

ที่ปรึกษาหลัก

(อาจารย์ ดร. วีระ สอึ้ง)

ประธาน

(รองศาสตราจารย์ ดร. สุพัฒน์ เอื้อทวีเกียรติ)

กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. นุวิทย์ วิวัฒน์วัฒนา)

ชื่อเรื่อง	แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดอุดตัน
	ในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง
ผู้วิจัย	อัจฉราภรณ์ สุขเพิ่ม
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	อาจารย์ ดร. วีระ สอึ้ง

ภาวะลิ้มเลือดอุดตันในหลอดเลือดดำเป็นโรคหนึ่งที่สำคัญที่มีผู้ป่วยเพิ่มมากขึ้นในประเทศไทยซึ่งเกิดจากการปิดกั้นการไหลเวียนของเลือดในหลอดเลือดดำ นอกจากนี้แบบจำลองการประเมินความเสี่ยงของภาวะลิ้มเลือดอุดตันในหลอดเลือดดำที่มีประสิทธิภาพจึงเป็นสิ่งสำคัญที่สุดเพื่อช่วยแพทย์วินิจฉัย งานวิจัยนี้ได้สร้างแบบจำลองการวินิจฉัยความเสี่ยงต่อการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำ โดยอาศัยหลักการเรียนรู้ของเครื่อง จากการเก็บรวบรวมข้อมูลผู้ป่วยในหอผู้ป่วยอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย งานวิจัยนี้ได้เตรียมข้อมูลทั้งหมด 1290 แถว 65 คอลัมน์ และตรวจสอบค่าที่ขาดหายไปพร้อมทั้งแปลงข้อมูลให้พร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไป จากนั้นแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบด้วยอัตราส่วน 70:30 ผลการทดลองของงานวิจัยได้เปรียบเทียบประสิทธิภาพของแต่ละ 3 อัลกอริทึม ประกอบด้วย Decision Tree, Logistic Regression และ Neural Network จากผลการทดลองแบบจำลอง Decision Tree โดยการปรับสมดุลของข้อมูลด้วยวิธี Class Weight ให้ค่า Sensitivity 25 เปอร์เซ็นต์ ค่า Precision 22 เปอร์เซ็นต์ และค่า F1-score 24 เปอร์เซ็นต์

คำสำคัญ : ภาวะลิ้มเลือดอุดตันในหลอดเลือดดำ, การเรียนรู้ของเครื่อง, การวินิจฉัยอัตโนมัติ

Title	AUTOMATIC DIAGNOSIS MODEL FOR RISK OF SYMPTOMATIC VENOUS THROMBOEMBOLISM BASED ON MACHINE LEARNING
Author	AUTCHARAPORN SUKPERM
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Dr. Vera Sa-ing

Venous thromboembolism (VTE) is an important disease in terms of increasing the number of patients because of a lack of awareness in Thailand on blocking blood flow in the veins. In addition, an effective assessment model of VTE risk was the most important factor for doctors in making a diagnosis. So, this research represents an automatic diagnosis model by using effective machine learning to predict the important risk factors of VTE by collecting patient data from the medical ward at King Chulalongkorn Memorial Hospital. This research prepared 1290 rows, 65 columns and investigated the missing values for transforming the data to be imported into each model and then separated the adjusted data for training and testing in the ratio of 70:30. The experimental results were compared to the effectiveness of three machine learning algorithms that consisted of the decision tree, logistic regression, and neural network. From the experimental result of the decision tree, this model represented a sensitivity of 25 %, precision of 22 % and F1-score of 24 % by adjusting the balance data with the class weight method for assisting diagnoses in the medical doctor.

Keyword : Venous Thromboembolism, Machine Learning, Automatic Diagnosis

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยด้วยความเมตตากรุณาช่วยเหลือ และความเอาใจใส่เป็นอย่างดี ยิ่งตลอดจนการให้คำแนะนำ และข้อคิดเห็นที่เป็นประโยชน์อย่างยิ่งสำหรับการปรับแก้ไขข้อบกพร่อง จากคณะกรรมการผู้ควบคุมสารนิพนธ์ ผู้วิจัยขอกราบขอบพระคุณ

อาจารย์ ดร.วีระ สอิ่ง ที่ได้ให้ความเมตตากรุณาเป็นที่ปรึกษาและให้ความช่วยเหลือชี้แนะแนวทางในสิ่งที่เป็นประโยชน์ต่อการศึกษาและการทำวิจัยครั้งนี้

ขอกราบขอบพระคุณ ศาสตราจารย์ ดร.นายแพทย์พลภัทร โรจน์นครินทร์ ผู้เชี่ยวชาญทางด้านการแพทย์ และนางเบญจพร อัศววัฒน์ นักเทคนิคการแพทย์ ชำนาญการพิเศษ หน่วยโลหิตวิทยา โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย ที่สนับสนุนให้ข้อมูลและคำแนะนำที่เป็นประโยชน์ในการทำวิจัยครั้งนี้

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.นุรีย์ วิวัฒน์วัฒนา พร้อมทั้งคณาจารย์ทุกท่านและกรรมการบริหารหลักสูตร สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้กรุณาให้ทั้งความรู้ในห้องเรียน ตลอดจนสั่งสอนให้มีคุณธรรม จริยธรรม ในทุกๆ ด้าน ในการดำเนินชีวิต

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.สุพัฒน์นา เอื้อทวีเกียรติ ประธานผู้ทรงคุณวุฒิในการสอบปากเปล่าสารนิพนธ์ ที่กรุณาให้คำแนะนำและข้อเสนอแนะเกี่ยวกับงานวิจัย เพื่อนำความรู้ที่ได้ไปปรับใช้ในอนาคตได้เป็นอย่างดีและเหมาะสม

ขอขอบคุณทุนอุดหนุนงานวิจัยจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ทำให้งานวิจัยสำเร็จลุล่วงไปด้วยดี

ขอขอบคุณเจ้าหน้าที่ ที่บัณฑิตวิทยาลัย สำหรับข้อมูลและการดำเนินงานเอกสารต่างๆ เกี่ยวกับการทำงานวิจัยครั้งนี้ ให้สำเร็จลุล่วงไปด้วยดี

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้ให้ความช่วยเหลือและเป็นกำลังใจให้กับผู้วิจัยตลอดมา

สุดท้ายนี้ ผู้วิจัยขอโน้มรำลึกถึงคุณบิดามารดาและครูอาจารย์ ที่อบรมสั่งสอนให้ความรู้ และเป็นกำลังใจ สนับสนุนผู้วิจัยตลอดมา

อัจฉราภรณ์ สุขเพิ่ม

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญ	ช
สารบัญตาราง.....	ฅ
สารบัญรูปภาพ	ญ
บทที่ 1 บทนำ.....	1
1.1 ความสำคัญและความเป็นมาของงานวิจัย.....	1
1.2 วัตถุประสงค์ของงานวิจัย.....	2
1.3 เป้าหมายและขอบเขตของงานวิจัย	3
1.4 ขั้นตอนการทำงานวิจัย.....	3
1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย.....	4
บทที่ 2 วรรณกรรม และงานวิจัยที่เกี่ยวข้อง.....	5
2.1 ภาวะล้มเหลือดุดตันในหลอดเลือดดำ.....	5
2.2 ข้อมูลที่สำคัญต่อการวิเคราะห์.....	10
2.3 การวิเคราะห์ด้วยหลักการของคอมพิวเตอร์.....	12
2.4 งานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ความเสี่ยงของการเกิดล้มเหลือดุดตัน	13
2.5 งานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่	24
บทที่ 3 วิธีการดำเนินการวิจัย.....	28
3.1 กระบวนการทำงานของแบบจำลอง	29
3.2 การเก็บรวบรวมข้อมูล (Data Acquisition).....	30

3.3 การคัดกรองข้อมูล (Data Filtering)	37
3.4 การสำรวจข้อมูล (Exploratory Data Analysis : EDA)	38
3.5 การเตรียมข้อมูล (Data Preparation)	44
3.6 หลักการที่ใช้ในการสร้างแบบจำลอง (Modeling).....	49
3.7 การประเมินผล (Evaluation)	51
บทที่ 4 ผลการดำเนินงานวิจัย.....	57
4.1 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Oversampling	58
4.2 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Undersampling.....	60
4.3 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE	62
4.4 ผลลัพธ์ของการสร้างแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class weight	64
4.5 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier	66
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ.....	71
5.1 สรุปผลการวิจัย.....	71
5.2 อภิปรายผลการวิจัย.....	72
5.3 ข้อเสนอแนะ.....	73
บรรณานุกรม	74
ประวัติผู้เขียน	77

สารบัญตาราง

หน้า

ตาราง 1 แสดงการเปรียบเทียบค่าความถูกต้องของ 3 อัลกอริทึม	26
ตาราง 2 แสดงคำอธิบายข้อมูลของตัวแปรในแต่ละคอลัมน์.....	30
ตาราง 3 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ Oversampling	47
ตาราง 4 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ Undersampling	48
ตาราง 5 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ SMOTE	49
ตาราง 6 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย วิธีการสมดุลข้อมูลแบบ Oversampling	58
ตาราง 7 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย วิธีการสมดุลข้อมูลแบบ Undersampling	60
ตาราง 8 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย วิธีการสมดุลข้อมูลแบบ SMOTE	62
ตาราง 9 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 2 อัลกอริทึมด้วย วิธีการสมดุลข้อมูลแบบ Class Weight.....	64
ตาราง 10 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย วิธีการสมดุลข้อมูลแบบ Ensemble sampling โดยให้ Balanced Bagging Classifier	66
ตาราง 11 แสดงผลลัพธ์ของแบบจำลองการทำนาย 3 อัลกอริทึมพร้อมเปรียบเทียบวิธีปรับความ สมดุลของข้อมูลทั้ง 5 วิธี	69

สารบัญรูปภาพ

หน้า

ภาพประกอบ 1 แสดงการไหลเวียนของเลือดในหลอดเลือด.....	1
ภาพประกอบ 2 แสดงภาวะหลอดเลือดดำชั้นลึกอุดตันที่ขา (Deep Vein Thrombosis, DVT).....	6
ภาพประกอบ 3 แสดงภาวะลิ่มเลือดอุดตันที่ปอด (Pulmonary Embolism, PE).....	6
ภาพประกอบ 4 แสดงแผนภาพแสดงการคัดกรองผู้ป่วยเพื่อวิจัหาปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตัน.....	8
ภาพประกอบ 5 แสดงปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ่มเลือดอุดตัน	9
ภาพประกอบ 6 โครงสร้างของ MLP ที่ใช้ประเมินความเสี่ยงของการเกิดลิ่มเลือดอุดตัน	10
ภาพประกอบ 7 แสดงการเปรียบเทียบความแตกต่างระหว่าง 2 โครงข่ายประสาทเทียม	14
ภาพประกอบ 8 แสดงตัวแปรที่ใช้ input เพื่อฝึกฝนและทดสอบข้อมูล	15
ภาพประกอบ 9 แสดงการเปรียบเทียบประสิทธิภาพแบบจำลอง ANNs กับ Logistic Regression	15
ภาพประกอบ 10 แสดงกราฟ ROC ของ ANNs และ Logistic Regression.....	16
ภาพประกอบ 11 แสดงกลุ่มของคุณลักษณะทางคลินิก	17
ภาพประกอบ 12 แสดงผลลัพธ์ของ Basic Predictors และ Predictors Based on Machine Learning with RO.....	18
ภาพประกอบ 13 แสดงโครงสร้างของแบบจำลอง Radial Basis Function (RBF) ANNs	19
ภาพประกอบ 14 เปรียบเทียบแบบจำลอง RBF ANNs และ Logistic regression สำหรับทำนายการเกิดลิ่มเลือดอุดตันของผู้ป่วยตับอ่อนอักเสบเฉียบพลัน	20
ภาพประกอบ 15 แสดงการเปรียบเทียบ ROC curves ของ Super learner model (ML) และ Reduced model (rML) เทียบกับ IMPROVE score.....	21
ภาพประกอบ 16 แสดงการกระจายของความเสี่ยงที่ทำนายด้วย rML.....	21
ภาพประกอบ 17 แสดงการเปรียบเทียบ ROC curves ของแต่ละโมเดล.....	22

ภาพประกอบ 18 เปรียบเทียบประสิทธิภาพของ Seely model กับ Machine learning models .23	23
ภาพประกอบ 19 แสดงประสิทธิภาพเทคนิค Khorana Score (KS) และเทคนิค ML-RO24	24
ภาพประกอบ 20 แสดงกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM.....25	25
ภาพประกอบ 21 แสดงแผนภาพกระบวนการทำงานของแบบจำลอง.....29	29
ภาพประกอบ 22 แสดงจำนวนผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตัน VTE positive(1) และไม่เป็นโรคลิ้มเลือดอุดตัน VTE negative(0)38	38
ภาพประกอบ 23 แสดงความสัมพันธ์ของข้อมูลที่น่ามาพิจารณาปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำจากข้อมูลทั้งหมด.....39	39
ภาพประกอบ 24 แสดงความสัมพันธ์ของปัจจัยเสี่ยงที่สัมพันธ์กันมากที่สุด.....41	41
ภาพประกอบ 25 แสดงข้อมูลของผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตันในหลอดเลือดดำ 27 ราย43	43
ภาพประกอบ 26 แสดงความกระจายตัวของผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตัน (เขียว) และไม่เป็นโรคลิ้มเลือดอุดตัน (ฟ้า)44	44
ภาพประกอบ 27 แสดงผลลัพธ์การแปลงข้อมูลให้เป็นตัวเลข ด้วย Label encoder45	45
ภาพประกอบ 28 แสดงข้อมูลทั้งสองคลาสที่ไม่สมดุลกัน (Imbalance)46	46
ภาพประกอบ 29 แสดงข้อมูลที่สมดุลกันด้วยวิธี Oversampling46	46
ภาพประกอบ 30 แสดงข้อมูลที่สมดุลกันด้วยวิธี Undersampling47	47
ภาพประกอบ 31 แสดงข้อมูลที่สมดุลกันด้วยวิธี SMOTE.....48	48
ภาพประกอบ 32 แสดงโมเดล Neural Network ที่มี 3 layers51	51
ภาพประกอบ 33 แสดงตาราง Confusion matrix.....51	51
ภาพประกอบ 34 ผลการทำนายการเกิดโรคลิ้มเลือดอุดตัน โดยใช้อัลกอริทึม Decision Tree ด้วยการสมดุลข้อมูลวิธี Class weight.....52	52
ภาพประกอบ 35 แสดงผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Oversampling58	58

ภาพประกอบ 36 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสุ่มลดข้อมูลแบบ Undersampling	60
ภาพประกอบ 37 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสุ่มลดข้อมูลแบบ SMOTE.....	62
ภาพประกอบ 38 ผลลัพธ์ของการสร้างแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสุ่มลดข้อมูลแบบ Class Weight.....	64
ภาพประกอบ 39 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสุ่มลดข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier.....	66

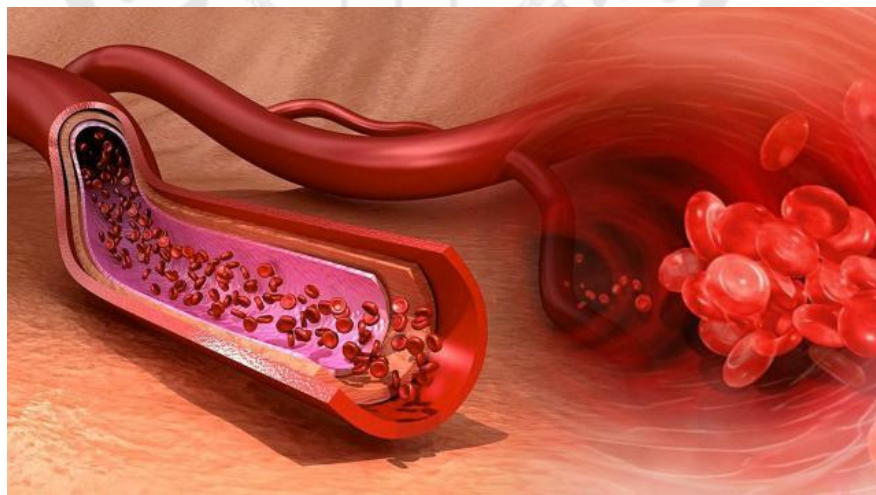


บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของงานวิจัย

ร่างกายของมนุษย์นั้นมีระบบต่าง ๆ ของอวัยวะที่ล้วนมีความสำคัญแตกต่างกันออกไป แต่ถ้าหากส่วนใดส่วนหนึ่งของระบบร่างกายเกิดความผิดปกติก็จะเกิดอันตรายกับร่างกายได้ ซึ่งในระบบไหลเวียนโลหิตของหลอดเลือดดำ (Vein) มีหน้าที่หลักคือขนส่งเลือดกลับเข้าสู่ระบบการทำงานของหัวใจ (Cardiovascular System) หรือมีหน้าที่จ่ายเลือดจากช่วงล่างของร่างกายไปยังหัวใจเพื่อฟอกเป็นเลือดแดงกลับมาหล่อเลี้ยงอวัยวะต่างๆของร่างกาย (Diawkee, 2018) หากเกิดมีลิ่มเลือดอุดตันขึ้นในหลอดเลือดดำ (Venous Thromboembolism, VTE) ซึ่งจะมีการจับตัวเป็นก้อนเป็นลิ่มทำให้เลือดไม่สามารถไหลเวียนได้ สามารถเกิดขึ้นได้ทุกบริเวณในร่างกายมนุษย์ โดยส่วนใหญ่มักจะพบภาวะการเกิดลิ่มเลือดอุดตันที่หลอดเลือดดำบริเวณขาของผู้ป่วย หรือเรียกในทางการแพทย์ว่า Deep Vein Thrombosis (DVT) ซึ่งจะมีอาการบวมแดงของขาข้างนั้นๆ มากไปกว่านั้นยังอาจจะทำให้ลิ่มเลือดเกิดไปอุดตันที่บริเวณของปอดได้ซึ่งเป็นอวัยวะที่สำคัญต่อการดำรงชีวิตของมนุษย์ ถ้าหากมีลิ่มเลือดไปอุดตันที่ปอดและไม่ได้ได้รับการรักษาอย่างทันท่วงทีก็อาจทำให้เสียชีวิตได้



ภาพประกอบ 1 แสดงการไหลเวียนของเลือดในหลอดเลือด

ที่มา : (THAIRATH, 2561)

ภาวะลิ้มเลือดอุดตันในหลอดเลือดดำเกิดจากปัจจัยเสี่ยงหลายๆ ปัจจัยด้วยกัน ตัวอย่างเช่น อายุ เพศ เชื้อชาติ พันธุกรรม การผ่าตัดและการได้รับบาดเจ็บจากอุบัติเหตุ เป็นต้น นอกจากนี้เมื่อเกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำขึ้นในร่างกายแล้ว อาจจะทำให้ต้องใช้ระยะเวลาในการรักษาที่โรงพยาบาลและยังต้องใช้ค่าใช้จ่ายในการรักษาพยาบาลที่มากอีกด้วย ถึงแม้ว่าโรคหลอดเลือดดำอุดตันเป็นโรคที่พบได้ไม่บ่อย แต่อาการที่เกิดขึ้นจากโรคนี้มีความรุนแรง และอาจนำไปสู่การสูญเสียชีวิตได้อย่างรวดเร็ว ทำให้ควรดูแลร่างกายให้แข็งแรงอยู่เสมอ โดยหลีกเลี่ยงปัจจัยเสี่ยงต่างๆ ที่ทำให้เกิดโรค มากไปกว่านั้นถ้าหากพบว่ามีความผิดปกติที่เกิดขึ้นกับร่างกายควรรีบพบแพทย์ทันที ดังนั้นภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำจึงเป็นโรคที่ควรศึกษาและทำความเข้าใจปัจจัยต่างๆ ที่จะนำไปสู่การเกิดโรคให้มากขึ้น มากไปกว่านั้นแพทย์และบุคลากรทางการแพทย์ที่เกี่ยวข้องในการรักษา ต้องตระหนักในการวินิจฉัยและการใช้เครื่องมือสำหรับประเมินการเกิดโรคที่เหมาะสม รวมทั้งพิจารณาอาการเพื่อให้เกิดการป้องกันและการรักษาที่เหมาะสมแก่ผู้ป่วยแต่ละราย เพื่อลดภาวะแทรกซ้อนต่างๆ และลดอัตราการเสียชีวิตของผู้ป่วยได้

ดังนั้นงานวิจัยนี้ผู้วิจัยต้องการประยุกต์เพื่อใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ เพื่อสร้างแบบจำลองของการทำนาย (Prediction Model) ถึงปัจจัยเสี่ยงที่อาจจะทำให้เกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำโดยได้ข้อมูลจากแพทย์ผู้เชี่ยวชาญที่เก็บข้อมูลของผู้ป่วย ที่เข้ารับการรักษาในหอผู้ป่วยอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย เพื่อใช้ในการช่วยให้แพทย์ตัดสินใจในการวินิจฉัยและรักษาได้แม่นยำมากขึ้น

1.2 วัตถุประสงค์ของงานวิจัย

ในการวิจัยครั้งนี้ได้ตั้งความมุ่งหมายไว้ดังนี้

1.2.1 เพื่อรวบรวมข้อมูลต่างๆ ที่สำคัญ เพื่อใช้ในการพยากรณ์การเกิดโรคลิ้มเลือดอุดตันในหลอดเลือดดำ

1.2.2 เพื่อศึกษาปัจจัยสำคัญ ที่สามารถนำไปสู่ความเสี่ยงของการเกิดโรคภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำได้

1.2.3 เพื่อสร้างแบบจำลองที่มีประสิทธิภาพ ในการทำนายความเสี่ยงของการเกิดโรคลิ้มเลือดอุดตันในหลอดเลือดดำได้

1.2.4 เพื่อทำการทดสอบประสิทธิภาพของแบบจำลองที่สร้างขึ้น ให้สามารถทำนายความเสี่ยงของการเกิดโรคภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำได้อย่างมีประสิทธิภาพ

1.2.5 เพื่อช่วยเพิ่มข้อมูลที่ใช้ในการพยากรณ์การเกิดโรคลิ้มเลือดอุดตันในหลอดเลือดดำเบื้องต้นให้กับแพทย์หรือผู้เชี่ยวชาญได้

1.3 เป้าหมายและขอบเขตของงานวิจัย

งานวิจัยนี้จะนำข้อมูลจากผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยแผนกอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย ที่เก็บรวบรวมภายในปี 2009 ที่ได้รับการอนุมัติจาก คณะกรรมการจริยธรรมของคณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยแล้ว เพื่อทำการสร้าง รูปแบบการเรียนรู้ของเครื่องในการทำนายความเสี่ยงที่จะนำไปสู่การเกิดโรคลิ่มเลือดอุดตันขึ้นใน หลอดเลือดดำ

1.4 ขั้นตอนการทำงานวิจัย

งานวิจัยนี้ได้พัฒนาระบบเพื่อช่วยในการวินิจฉัยภาวะการเกิดลิ่มเลือดอุดตันโดยใช้ เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยวิเคราะห์จากปัจจัยเสี่ยงของการเกิดโรค ดังกล่าว โดยแบ่งขั้นตอนออกเป็นดังนี้

1.4.1 ศึกษางานวิจัยเกี่ยวกับภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ จาก บทความวิจัยเรื่อง Risk factors for symptomatic venous thromboembolism in Thai hospitalised medical patients (Rojnuckarin et al., 2011) ซึ่งได้กล่าวไว้ว่าการป้องกันการเกิด ลิ่มเลือดอุดตันในหลอดเลือดดำไม่สามารถลดอัตราการเสียชีวิตของผู้ป่วยในโรงพยาบาลได้ จึงมี วัตถุประสงค์ในการหาปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำเพื่อเป็น เป้าหมายหลักในการป้องกันการเกิดโรคให้กับผู้ป่วย ในโรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย โดยข้อมูลที่น่ามาวิเคราะห์แพทย์ได้ทำการเก็บข้อมูลของผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วย อายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ ในปี 2009

1.4.2 งานวิจัยนี้จะทำการวิจัยต่อยอดจากงานวิจัยเรื่อง Risk factors for symptomatic venous thromboembolism in Thai hospitalised medical patients โดยจะใช้ เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาทำการวิเคราะห์ข้อมูลดังกล่าวเพื่อ ประเมินประสิทธิภาพของโมเดลในการทำนายการเกิดโรคลิ่มเลือดอุดตันในหลอดเลือดดำ และ สามารถให้ข้อมูลกับแพทย์เพื่อเป็นแนวทางในการรักษาผู้ป่วยได้

1.4.3 ขั้นตอนการเตรียมข้อมูล แปลงข้อมูล และจัดการข้อมูลที่มีปัญหา หรือไม่ เหมาะสมต่อการนำมาวิเคราะห์

1.4.4 ทำการวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis : EDA) จาก ข้อมูลปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตันหลาย ๆ ปัจจัยที่ได้จากการรวบรวมข้อมูลจากแพทย์ โดยได้รับการอนุมัติจากคณะกรรมการจริยธรรมของคณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย แล้ว ซึ่งประกอบด้วย 1,290 แถว และ 65 คอลัมน์

1.4.5 ทดสอบแบบจำลองที่เลือกใช้ในการทำนายความความเสี่ยงของการเกิดโรคภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำ ด้วยหลักการวัดทางคณิตศาสตร์ที่มีประสิทธิภาพ

1.4.6 สรุปผลการทดลอง เพื่อนำเสนอแบบจำลองที่มีประสิทธิภาพในการทำนายความความเสี่ยงของการเกิดโรคภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำ

1.4.7 วิเคราะห์ และอภิปรายผลการทดลอง แบบจำลองที่เลือกใช้ในการทำนายความความเสี่ยงของการเกิดโรคภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำ เพื่อแนะนำการปรับปรุงหรือเสนอแนะข้อจำกัดของแบบจำลอง

1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1.5.1 สามารถศึกษาปัจจัยต่าง ๆ ของการทำนายการเกิดโรคลิ้มเลือดอุดตันขึ้นในหลอดเลือดดำได้

1.5.2 สามารถพัฒนาแบบจำลอง ที่สามารถช่วยในการทำนายของแพทย์ในการรักษาผู้ป่วยได้ง่ายขึ้นได้

1.5.3 สามารถเพิ่มความถูกต้องในการทำนายให้กับแพทย์ด้วยผลลัพธ์ของการทำนายจากแบบจำลองการทำนายที่พัฒนาขึ้นได้

1.5.4 แบบจำลองที่สร้างขึ้น สามารถช่วยลดเวลาในการวินิจฉัยให้กับแพทย์ได้

บทที่ 2

วรรณกรรม และงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลองในการทำนายภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ และได้นำเสนอข้อค้นพบต่อไปนี้

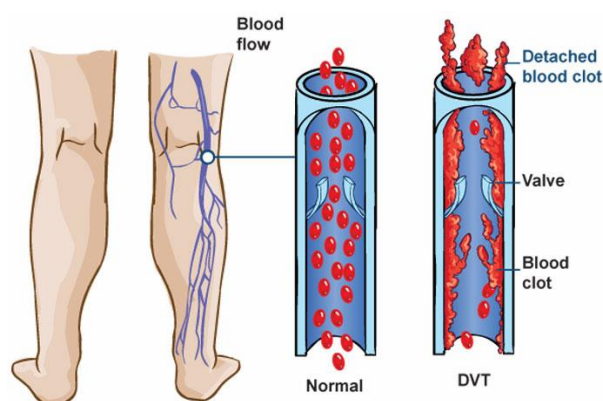
- 2.1 ทฤษฎีเกี่ยวกับภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ
- 2.2 ข้อมูลที่สำคัญต่อการวิเคราะห์ภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ
- 2.3 ทฤษฎีของการวิเคราะห์ข้อมูลด้วยหลักการทางคอมพิวเตอร์
- 2.4 งานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ความเสี่ยงของภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ
- 2.5 งานวิจัยที่ศึกษาเกี่ยวกับข้อมูลขนาดใหญ่

2.1 ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ

ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ หรือ Venous Thromboembolism (VTE) เกิดจากการที่มีลิ่มเลือดไปขัดขวางการไหลเวียนของเลือดในหลอดเลือดดำ ซึ่งการจับตัวกันเป็นลิ่มเลือดจะขัดขวางการไหลเวียนของเลือดอย่างช้าๆ ประกอบด้วยภาวะหลอดเลือดดำชั้นลึกอุดตันที่ขา (Deep Vein Thrombosis, DVT) ซึ่งส่วนใหญ่เกิดที่หลอดเลือดดำชั้นลึกที่ขา และอาจเกิดลิ่มเลือดหลุดไปอุดหลอดเลือดแดงที่ปอดทำให้เกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดแดงที่ปอด (Pulmonary Embolism, PE) (Uaprasert, Chanswangphuwana, & al., 2020) เป็นสาเหตุให้เกิดการเสียชีวิตอย่างเฉียบพลันได้

2.1.1 ภาวะหลอดเลือดดำชั้นลึกอุดตันที่ขา (Deep Vein Thrombosis, DVT)

การเกิดลิ่มเลือดอุดตันที่ขานั้น เริ่มจากการที่มีการไหลเวียนของเลือดน้อย ได้แก่ หลอดเลือดดำ Soleal ของน่อง และบริเวณ Valve pocket ซึ่งมักเป็นตำแหน่งเริ่มต้นของการเกิดลิ่มเลือดอุดตันในชั้นลึกของขา แสดงดังภาพประกอบ 2 ส่วนใหญ่เกิดจากผู้ป่วยที่ไม่ได้เคลื่อนไหวเป็นเวลานาน ๆ ผู้ป่วยที่มีภาวะหัวใจวาย เป็นต้น ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ ไม่ได้เกิดขึ้นแต่บริเวณขาเพียงที่เดียว อาจเกิดได้ทุกที่ในร่างกาย แต่บริเวณขานั้นเกิดขึ้นได้ง่ายกว่าบริเวณอื่น ๆ

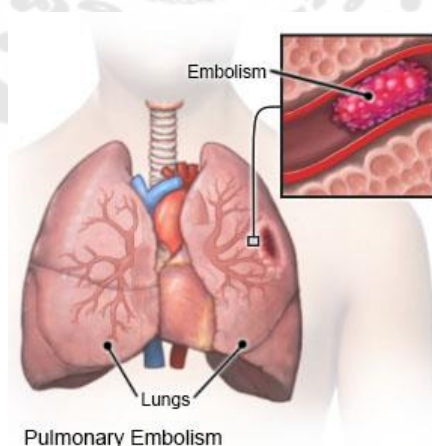


ภาพประกอบ 2 แสดงภาวะหลอดเลือดดำชั้นลึกอุดตันที่ขา (Deep Vein Thrombosis, DVT)

ที่มา : (TSUPAMAN, 2556)

2.1.2 ภาวะลิ่มเลือดอุดตันในหลอดเลือดแดงที่ปอด (Pulmonary Embolism, PE)

อาการของลิ่มเลือดไปอุดตันที่ปอดเริ่มจากการเกิดลิ่มเลือดอุดตันที่บริเวณขาเป็นส่วนใหญ่ ซึ่งเป็นหลอดเลือดที่เชื่อมต่อกับหัวใจข้างขวาและปอด ทำให้ลิ่มเลือดนั้นสามารถหลุดเข้าไปอุดตันที่ปอดได้ แสดงดังภาพประกอบ 3 หรืออาจจะเกิดลิ่มเลือดที่บริเวณอื่น ๆ ก่อนก็ได้ เช่น บริเวณแขน เป็นต้น ในผู้ป่วยที่มีลิ่มเลือดไปอุดตันที่ปอด จะมีอาการเหนื่อยและเจ็บหน้าอก ร่วมกับหรือมีความรู้สึกเจ็บในขณะหายใจ (ตั้งสุจริตวิจิตร, 2018)



ภาพประกอบ 3 แสดงภาวะลิ่มเลือดอุดตันที่ปอด (Pulmonary Embolism, PE)

ที่มา : (Drugs, 2564)

2.1.3 อาการที่บ่งชี้ของการเกิดโรค

อาการของภาวะลิ่มเลือดอุดตันในหลอดเลือดดำมักขึ้นอยู่กับตำแหน่งของอวัยวะที่มีลิ่มเลือดอุดตัน ผู้ป่วยภาวะหลอดเลือดดำชั้นลึกอุดตัน บางรายอาจไม่แสดงอาการผิดปกติให้เห็น บางรายอาจมีอาการปวด บวมแดง ส่วนผู้ป่วยที่มีภาวะลิ่มเลือดไปอุดตันที่ปอด จะมีอาการเจ็บหน้าอก หายใจถี่ หายใจไม่อิ่ม ไอหรือไอมีเลือดปน หมดสติและหัวใจเต้นผิดปกติ (Uaprasert et al., 2020)

จากการค้นคว้าศึกษางานวิจัยเรื่องปัจจัยเสี่ยงของภาวะหลอดเลือดดำอุดตันในผู้ป่วยที่เข้ารับการรักษาในโรงพยาบาลในประเทศไทย ซึ่งเป็นการคาดการณ์การความเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำในชาวเอเชีย พบว่าปัจจัยเสี่ยงของอาการที่ทำให้เกิดภาวะลิ่มเลือดอุดตันในผู้ป่วยที่มีความเสี่ยงสูง 2.1% ซึ่งใกล้เคียงกับความเสี่ยงของผู้ป่วยในแถบทางตะวันตกที่มีความเสี่ยงประมาณ 2% ปัจจัยเสี่ยงทั่วไปที่ใช้ในการศึกษาของงานวิจัยนี้จำแนกผู้ป่วยในกลุ่มที่เป็น DVT (0.09%) ตรวจพบในผู้ป่วย 1,160 รายที่มีความเสี่ยงต่ำ แม้ว่าจะมีกลุ่มที่ไม่ได้ติดตามหลังจาก discharge ส่วนใหญ่ (75%) ของผู้ป่วย VTE เกิดขึ้นระหว่างนอนรักษาอยู่ที่โรงพยาบาล พบว่าผู้ป่วยที่เป็นโรคแพภูมิตัวเองและเป็นเนื้องอกจะมีความเสี่ยงสูงต่อการเกิดภาวะลิ่มเลือดอุดตัน (Rojnuckarin et al., 2011)

2.1.4 การรักษา

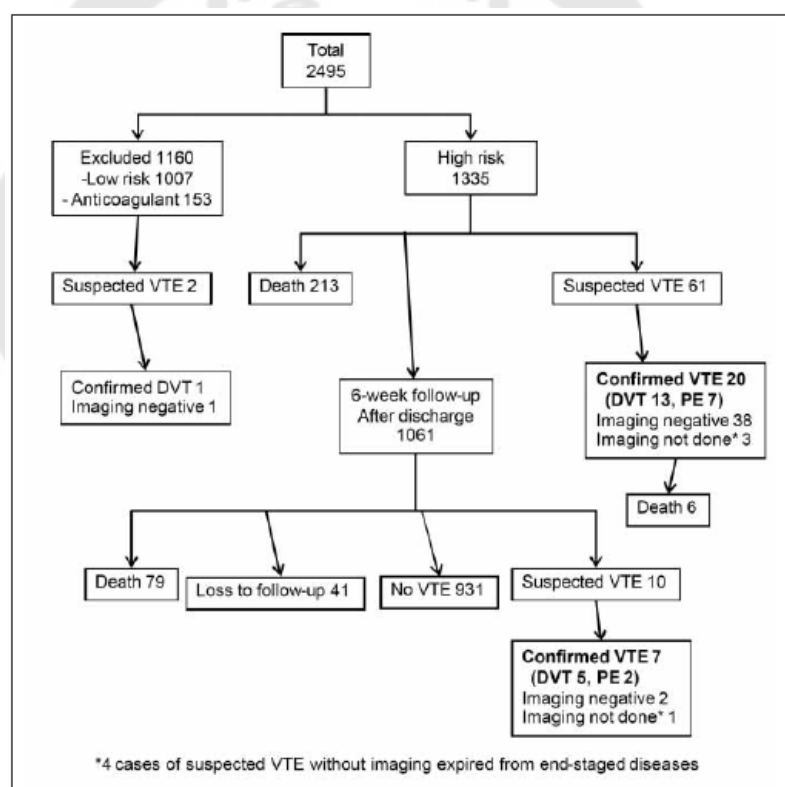
การรักษาภาวะลิ่มเลือดอุดตันในหลอดเลือดดำในปัจจุบันคือ การให้ยาต้านการแข็งตัวของเลือด เพื่อป้องกันการแข็งตัวเพิ่มขึ้นของลิ่มเลือด และป้องกันการกลับเป็นซ้ำทั้งในระยะแรกและระยะยาว โดยลิ่มเลือดที่เกิดขึ้นจะค่อยๆสลายไป โดยการทำงานของระบบละลายลิ่มเลือด (Uaprasert et al., 2020) การให้ยาต้านการแข็งตัวของเลือดจึงถือเป็นการรักษาหลักในภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ

2.1.5 การวินิจฉัย

การวินิจฉัยภาวะหลอดเลือดดำอุดตันอาศัยทั้งลักษณะทางคลินิก การเจาะเลือดตรวจ Biochemical Marker และการตรวจวินิจฉัยทางรังสีวิทยา ซึ่งจะพิจารณาการส่งตรวจอย่างเป็นระบบและเหมาะสม เพื่อให้ได้การวินิจฉัยที่ถูกต้องและคุ้มค่ามากที่สุด ที่สำคัญภาวะลิ่มเลือดอุดตันในหลอดเลือดดำหากได้รับการวินิจฉัยตั้งแต่เริ่มแรกและมีการรักษาให้เหมาะสมในช่วง 3-6 เดือนแรก จะสามารถรักษาให้หายขาดได้ (Uaprasert et al., 2020)

การวินิจฉัยปัจจัยเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำโรงพยาบาลในประเทศไทย โดยใช้ข้อมูลของผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ ในปี 2009 เป็นเวลา 3 วันหรือนานกว่านั้น ได้รับการประเมินจากปัจจัยเสี่ยงที่รู้จักกัน

ดี ได้แก่ ภาวะหัวใจล้มเหลว, โรคทางเดินหายใจเรื้อรังที่มีอาการกำเริบเฉียบพลัน, อัมพฤกษ์หรืออัมพาต, เป็นทั้งมะเร็งเม็ดเลือดและไม่ใช่มะเร็งเม็ดเลือด, โรคภูมิแพ้ตัวเอง, หายใจผิดปกติต้องใช้เครื่องช่วยหายใจหรือการช่วยหายใจแบบไม่ต้องผ่าตัด, กระดูกหัก, โรคไขข้ออักเสบที่ขา, ประวัติก่อนเป็น VTE, ประวัติครอบครัวของการเป็น VTE, ประวัติก่อนเป็นมะเร็ง, เส้นเลือดขอด, การใช้ Oestrogen, ประวัติการอุดตัน, platelet counts $> 600 \times 10^9/l$, body mass index $\geq 30 \text{ kg/m}^2$ และกลุ่มอาการของโรคไต แสดงการคัดกรองผู้ป่วยดังภาพประกอบ 4 ฟอรัมการประเมินความเสี่ยงถูกแนบมาในแฟ้มประวัติที่อยู่หอผู้ป่วย แต่ละปัจจัยเสี่ยงที่เกิดขึ้นถูกประเมินขั้นต้นโดยแพทย์และตรวจสอบในภายหลัง หลังจาก Discharge กรณีที่ไม่มีปัจจัยเสี่ยงข้างต้นหรือได้รับยาละลายลิ่มเลือดระหว่างอยู่หอผู้ป่วยจะได้รับการยกเว้นนำมาวิเคราะห์ (Rojnuckarin et al., 2011)



ภาพประกอบ 4 แสดงแผนภาพแสดงการคัดกรองผู้ป่วยเพื่อวิจัหาปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตัน

ที่มา : (Rojnuckarin et al., 2011)

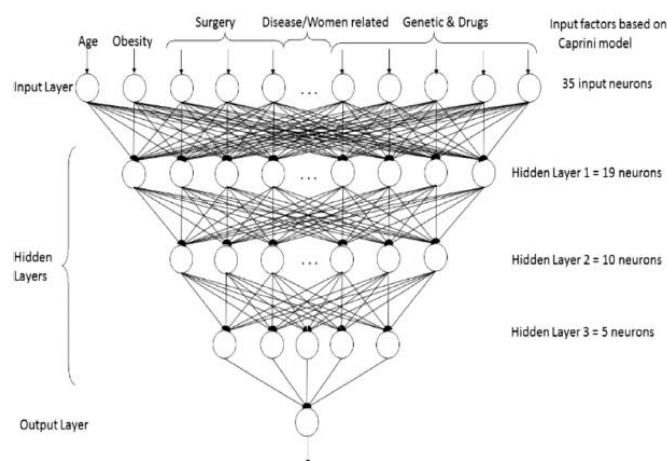
ปัจจัยเสี่ยงของ VTE แสดงเป็นความเสี่ยงสัมพัทธ์ 95% ของช่วงความมั่นใจ (CI) และ P-values จากการทดสอบ Chi Square ปัจจัยที่สำคัญนำไปวิเคราะห์ในแบบหลายตัวแปร สถิติทั้งหมดถูกคำนวณโดยโปรแกรม SPSS 16.0 for Windows ผลลัพธ์ที่ได้จากการวิเคราะห์พบว่า ภาวะลิ่มเลือดอุดตันในหลอดเลือดดำพบได้ทั่วไปในโรงพยาบาลในเอเชีย จากผลการวิจัยพบว่าผู้ป่วยที่เป็นโรคมะเร็งตัวเองและผู้ป่วยที่เป็นมะเร็งจะมีความเสี่ยงสูงต่อการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ จึงต้องมีกลยุทธ์ในการป้องกันโรคในกลุ่มเหล่านี้ (Rojnuckarin et al., 2011) แสดงดังภาพประกอบ 5

Risk factors	Relative risk (95% CI)	P-value
Autoimmune disease	11.83 (3.89–35.97)	<0.001
Solid tumours	4.66 (1.84–11.79)	0.001
Family history of VTE	120.28 (6.89–2101)	0.001
Varicose vein	40.09 (3.75–429.27)	0.002
Oestrogen	17.08 (1.18–248.20)	0.038
95% CI, 95% confidence interval.		

ภาพประกอบ 5 แสดงปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ่มเลือดอุดตัน

ที่มา : (Rojnuckarin et al., 2011)

มีอีกงานวิจัยที่น่าสนใจที่ผู้วิจัยได้ศึกษาเป็นงานวิจัยเรื่องระบบช่วยสนับสนุนการตัดสินใจทางคลินิกของการจำแนกความเสี่ยงของการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ Zelal Qatawneh และคณะ ได้นำหลักการทางคอมพิวเตอร์เข้ามามีบทบาทช่วยในการวินิจฉัยปัจจัยเสี่ยงของการเกิดลิ่มเลือดอุดตันโดยใช้ระบบโครงข่ายประสาทเทียม Multilayer Perceptron (MLP) โดยใช้ Resilient Backpropagation algorithm (Rprop) ใช้ในการฝึกอบรมข้อมูลที่ Input ปัจจัยเสี่ยงของการเกิดโรค แล้วระบบนำไปวิเคราะห์ในชั้น Hidden Layer ผลลัพธ์ Output ที่ได้ออกมาจะเป็นระดับคะแนนความเสี่ยงของการเกิดโรค โดยแบ่งเป็น 5 ระดับความเสี่ยง ได้แก่ Low, Lower-Mild, Higher-Mild, Moderate และ High แสดงดังภาพประกอบ 6 เพื่อช่วยแพทย์ประกอบการตัดสินใจในการวินิจฉัยผู้ป่วย ซึ่งผลการประเมินประสิทธิภาพความถูกต้องของระบบดังกล่าวอยู่ที่ 81% (Qatawneh, Alshraideh, Almasri, Tahat, & Awidi, 2019) ซึ่งระบบในงานวิจัยนี้แสดงให้เห็นดังภาพประกอบ



ภาพประกอบ 6 โครงสร้างของ MLP ที่ใช้ประเมินความเสี่ยงของการเกิดลิ่มเลือดอุดตัน

ที่มา : (Qatawneh et al., 2019)

2.2 ข้อมูลที่สำคัญต่อการวิเคราะห์

ภาวะหลอดเลือดดำอุดตัน เป็นสาเหตุสำคัญที่ทำให้เกิดการเสียชีวิตและทุพพลภาพมากที่สุดในโลก จากผลการศึกษาขององค์การอนามัยโลกพบว่าสาเหตุการเสียชีวิตและทุพพลภาพเกิดจากการที่ผู้ป่วยมานอนรับการรักษาตัวที่โรงพยาบาล โดยนอนอยู่กับเตียงเป็นเวลานาน ขยับขาไม่ได้ ลูกเดินไม่ได้ แพทย์จึงแนะนำให้บริหารขา เพื่อให้เลือดไหลเวียนได้เร็วขึ้น ถึงแม้ว่าภาวะหลอดเลือดอุดตันจะสามารถป้องกันได้ แต่ก็มักจะพบว่าเป็นปัจจัยเสี่ยงให้เกิดภาวะหัวใจล้มเหลว ภาวะหลอดเลือดสมองตีบ ซึ่งเป็นกลุ่มของโรคหัวใจและหลอดเลือดที่เป็นสาเหตุสำคัญที่ทำให้ผู้ป่วยเสียชีวิต (สภากาชาดไทย, 2018)

จากงานวิจัยต่างได้มีการศึกษาปัจจัยเสี่ยงมากมายที่เป็นสาเหตุของการเกิดภาวะหลอดเลือดอุดตัน ทั้งปัจจัยภายใน เช่น อายุ เชื้อชาติ พันธุกรรม มะเร็ง โรคหัวใจ เป็นต้น และปัจจัยภายนอก เช่น อาหาร สิ่งแวดล้อม การเดินทาง เป็นต้น โดยผู้ที่มีความเสี่ยงหลายปัจจัยร่วมกันเพิ่มโอกาสเสี่ยงต่อการเกิดภาวะหลอดเลือดดำอุดตันมากขึ้น

อายุ อุบัติการณ์การเกิดหลอดเลือดดำอุดตันสูงขึ้นชัดเจนในผู้สูงอายุ โดยเฉพาะในอายุที่มากกว่า 50 ปี (Uaprasert et al., 2020)

เพศ ผลของเพศต่ออุบัติการณ์การเกิดภาวะหลอดเลือดดำอุดตันมีความแตกต่างกันในหลายการศึกษาและขึ้นอยู่กัช่วงอายุของผู้ป่วย บางการศึกษาพบว่าอุบัติการณ์การเกิดภาวะ

หลอดเลือดอุดตันระหว่างเพศชายและเพศหญิงไม่แตกต่างกัน โดยการศึกษาที่ขัดแย้งกันในกลุ่มประชากรที่ต่างกันอาจมีผลจากปัจจัยอื่นร่วมด้วย (Uaprasert et al., 2020)

เชื้อชาติ จากการศึกษาอุบัติการณ์ของภาวะหลอดเลือดอุดตันในแต่ละเชื้อชาติซึ่งอาศัยอยู่ในรัฐ หรือภูมิภาคเดียวกันในสหรัฐอเมริกาพบว่า มีอุบัติการณ์การเกิดสูงสุดในชาวแอฟริกันอเมริกัน ตามด้วยชาวคอเคเซียน ฮิสพานิกส์และพบอุบัติการณ์ต่ำสุดในชาวเอเชียซึ่งต่ำประมาณ 4-5 เท่า (Uaprasert et al., 2020) แสดงให้เห็นว่าประชากรที่อาศัยอยู่ในสิ่งแวดล้อมใกล้เคียงกัน แต่เชื้อชาติแตกต่างกันมีอุบัติการณ์ของการเกิดภาวะหลอดเลือดอุดตันแตกต่างกันมาก ดังนั้น ความแตกต่างทางเชื้อชาติซึ่งมีความสัมพันธ์กับปัจจัยทางพันธุกรรมมีส่วนสำคัญต่อการเกิดภาวะหลอดเลือดอุดตัน

พันธุกรรม มีการศึกษาจำนวนมากในชาวตะวันตกซึ่งแสดงให้เห็นถึงปัจจัยทางพันธุกรรมที่มีผลต่อการเพิ่มความเสี่ยงต่อการเกิดภาวะหลอดเลือดอุดตันที่เกิดครั้งแรกและเกิดซ้ำ (Uaprasert et al., 2020) โดยปัจจัยทางพันธุกรรมนี้ส่งผลให้ผู้ป่วยมีภาวะเลือดแข็งตัวง่ายผิดปกติ ซึ่งแบ่งได้ 2 กลุ่มคือ

กลุ่มที่ 1 การขาดสารต้านการแข็งตัวของเลือด พบบ่อยในชาวเอเชีย รวมทั้งคนไทย แต่พบน้อยในชาวตะวันตก

กลุ่มที่ 2 การเพิ่มขึ้นของปริมาณหรือการทำงานของ Coagulation Factor เป็นความผิดปกติที่พบบ่อยที่สุดในชาวตะวันตกแต่พบน้อยมากในคนเอเชียรวมทั้งคนไทย ซึ่งความผิดปกติในกลุ่มนี้มีความเสี่ยงต่อการเกิดภาวะหลอดเลือดดำอุดตันน้อยกว่ากลุ่มแรก

ผู้ป่วยเป็นมะเร็ง พบอุบัติการณ์การเกิดภาวะหลอดเลือดดำอุดตันในผู้ป่วยมะเร็งสูงกว่าประชากรปกติมาก การศึกษาของ Blom และคณะ พบว่ามะเร็งเพิ่มความเสี่ยงถึง 7 เท่า (Uaprasert et al., 2020) โดยมะเร็งที่มีความเสี่ยงต่อการเกิดหลอดเลือดดำอุดตันสูง ได้แก่ มะเร็งกระดุก มะเร็งสมอง มะเร็งตับอ่อน มะเร็งรังไข่ มะเร็งโลหิต มะเร็งปอด มะเร็งทางเดินอาหาร และมะเร็งตับ สำหรับปัจจัยที่เพิ่มความเสี่ยงต่อการเกิดหลอดเลือดดำอุดตันในผู้ป่วยมะเร็ง ได้แก่ การแพร่กระจาย ยาเคมีบำบัด ฮอร์โมนบำบัดในผู้ป่วยมะเร็งเต้านมแต่ไม่เพิ่มขึ้นในมะเร็งต่อมลูกหมาก ส่วนการผ่าตัดและฉายแสงไม่เพิ่มความเสี่ยงในผู้ป่วยมะเร็ง

ภาวะเจ็บป่วยทางอายุรกรรม พบว่ามีความเสี่ยงต่อการเกิดภาวะหลอดเลือดดำอุดตันเพิ่มขึ้น โดยเฉพาะการรักษาในโรงพยาบาลพบว่าผู้ป่วยที่เกิดภาวะหลอดเลือดดำอุดตันขณะรักษาตัวในโรงพยาบาลส่วนใหญ่เป็นผู้ป่วยในหอผู้ป่วยอายุรกรรมทั่วไปโดยไม่ได้รับการป้องกันด้วยยาต้านการแข็งตัวของเลือดย่อยละ 5-45 (Uaprasert et al., 2020) ภาวะที่เพิ่มความเสี่ยง ได้แก่

มะเร็ง อัมพาต ภาวะหัวใจล้มเหลว โรคหัวใจขาดเลือด เบาหวาน ความอ้วน ไขมันในเลือดสูง การอยู่เนืองนิตย์ โรคหัวใจขาดเลือด เป็นต้น

การผ่าตัดและการได้รับบาดเจ็บ การผ่าตัดเป็นปัจจัยเสี่ยงสูงสุดโดยมีอัตราเสี่ยงสูงถึง 21.7 เท่า และการได้รับบาดเจ็บมีอัตราเสี่ยงสูงรองลงมา 12.7 เท่า โดยขึ้นอยู่กับชนิดของการผ่าตัดและการบาดเจ็บ (Uprasert et al., 2020)

2.3 การวิเคราะห์ด้วยหลักการของคอมพิวเตอร์

2.3.1 การจำแนกประเภทข้อมูล (Classification)

เป็นกระบวนการสร้างแบบจำลองเพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (Training Data) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลสอนเป็นข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะนำมาเปรียบเทียบกับกลุ่มที่หาได้จากแบบจำลอง เพื่อทดสอบประสิทธิภาพว่ามีความถูกต้องมากน้อยแค่ไหน และปรับปรุงแบบจำลองจนกว่าจะได้ค่าความถูกต้องหรือค่าอื่น ๆ ที่ผู้วิจัยสนใจ เป็นระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามาจะนำข้อมูลมาผ่านแบบจำลองโดยแบบจำลองจะสามารถทำนายกลุ่มของข้อมูลนี้ได้ (อัมพวัน)

2.3.2 การทำความเข้าใจกับข้อมูล (Data understanding)

ในขั้นตอนนี้เป็นขั้นตอนที่ต้องทำความเข้าใจว่าข้อมูลแต่ละตัวแปรที่นำมาใช้นั้นแต่ละตัวแปรคืออะไรบ้าง ทำการตรวจสอบข้อมูลที่ได้อามาเพื่อดูความถูกต้องของข้อมูล แล้วทำการพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์หรือไม่

2.3.3 การเตรียมข้อมูล (Data Preparation)

เป็นขั้นตอนที่ทำการเตรียมข้อมูลเพื่อนำไปวิเคราะห์ แบ่งออกเป็น 3 ขั้นตอนย่อย ๆ (พัชรวงศ์ศักดิ์, 2020) ได้แก่

2.3.3.1 Data Selection

เป็นขั้นตอนของการเลือกตัวแปรที่จะนำมาใช้งาน เลือกตัวแปรที่มีความสัมพันธ์กันส่งผลต่อกันและเป็นประโยชน์ต่อการนำไปทำนาย

2.3.3.2 Data Cleansing

เป็นขั้นตอนการหาว่าข้อมูลมีสิ่งผิดปกติอะไรบ้าง เช่น ข้อมูลมีค่าว่างที่ขาดหายไปหรือไม่ ถ้ามีจะต้องแก้ไขอย่างไรบ้าง อาจจะเพิ่มข้อมูลที่ขาดหายไปด้วยค่าทางสถิติ

2.3.3.3 Data Transformation

เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการวิเคราะห์

2.3.4 การสร้างโมเดล (Modeling)

เป็นขั้นตอนการวิเคราะห์ด้วยเทคนิค การจำแนกประเภทข้อมูล (Classification) ซึ่งเป็นกระบวนการสร้างโมเดลเพื่อจัดแบ่งกลุ่มข้อมูล โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ หรือเรียกว่า Training data และนำข้อมูลส่วนที่เหลือจากข้อมูลสอนมาเป็นข้อมูลเพื่อใช้ในการทดสอบ หรือเรียกว่า Testing data แล้วทำการทดสอบความถูกต้อง โดยการปรับจูนโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ

2.3.5 การประเมินผล (Evaluation)

ในขั้นตอนนี้เป็นการประเมินประสิทธิภาพของผลลัพธ์จากโมเดลวิเคราะห์ข้อมูล อาจใช้การเปรียบเทียบจากผลลัพธ์ที่ได้จากแต่ละเทคนิคที่ใช้งานว่าโมเดลไหนดีกว่ากัน เมื่อเลือกโมเดลที่ให้ความถูกต้องดีแล้วก็สามารถนำไปใช้งานต่อได้

2.4 งานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ความเสี่ยงของการเกิดลิ่มเลือดอุดตัน

งานวิจัยนี้ได้ทำการทบทวนวรรณกรรมและศึกษาค้นคว้างานวิจัยอย่างมากมายที่เกี่ยวข้องกับการวินิจฉัยและวิเคราะห์ความเสี่ยงของการเกิดโรคภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ โดยเลือกงานวิจัยบางส่วนที่มีรายละเอียดสำคัญกับงานวิจัยดังกล่าวการต่อไปนี้

2.4.1 บทความวิจัยเรื่อง The Prediction of the Risk Level of Pulmonary Embolism and Deep Vein Thrombosis through Artificial Neural Network (Agharezaei et al., 2016)

บทความนี้ได้ทำนายระดับความเสี่ยงของการเกิดลิ่มเลือดอุดตันที่ปอดของผู้ป่วย โดยใช้โครงข่ายประสาทเทียม 2 ชนิด ได้แก่ Feed-Forward Back Propagation และ Elman Back Propagation ใน Software MATLAB ในการวิเคราะห์ข้อมูล โดยโครงสร้างประกอบด้วย Input Layer 31 neurons และ Output Layer 1 neuron ที่แบ่งเป็นระดับความเสี่ยงของการเกิดโรค 4 ระดับ ได้แก่ Low Risk, Moderate Risk, High Risk และ Highest Risk เพื่อช่วยให้แพทย์วินิจฉัยผู้ป่วยที่มีความเสี่ยงสูงได้แม่นยำมากขึ้น

ผลลัพธ์ของการเปรียบเทียบโมเดลระหว่าง Feed-Forward Back Propagation และ Elman Back Propagation ดังภาพประกอบ 7 พบว่า Feed-Forward Back Propagation เมื่อทำการ Optimized ระบบแล้วได้ค่า Correlation Coefficient 97% ซึ่งพบว่ามีประสิทธิภาพดีกว่า Elman Back Propagation ที่ได้ Correlation Coefficient 95% เนื่องจาก Feed-Forward Back Propagation นั้นมีความเร็วในการส่งข้อมูลสูงและใช้จำนวนเซลล์ประสาทที่น้อยในชั้น Hidden Layer

Model Number	Network type	Learning Method	Transfer Function	Number of neurons in hidden layer	MAPE (%)	Incorrectly Classified Instances
1	Feed-forward backprop	Train LM	TanSig	10	4.35	4
2	Elman backprop	Train LM	TanSig	20	4.43	4

ภาพประกอบ 7 แสดงการเปรียบเทียบความแตกต่างระหว่าง 2 โครงข่ายประสาทเทียม

ที่มา : (Agharezaei et al., 2016)

ในอนาคตงานวิจัยนี้จะพัฒนาโดยเพิ่มขนาดตัวอย่างที่นำมาทำนายให้มากขึ้น และปรับโครงสร้างของโครงข่ายประสาทเทียมด้วยวิธีที่ชาญฉลาดขึ้น เช่น Genetic Algorithm และ Fuzzy algorithm

2.4.2 บทความวิจัยเรื่อง Artificial Neural Networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis (Fei, Hu, Li, Wang, & Zong, 2017)

บทความนี้ได้ใช้โครงข่ายประสาทเทียม (ANNs) ในการทำนายการเกิดลิ้มเลือดอุดตันของหลอดเลือดดำ (Portosplenomesenteric Venous Thrombosis, PSMVT) โดยศึกษากับผู้ป่วย 72 ราย ที่เป็นโรคตับอ่อนอักเสบเฉียบพลัน สุ่มแบ่งข้อมูลเพื่อฝึกฝน 48 รายและใช้ในการทดสอบ 24 ราย บ้อนตัวแปร Input 11 ตัวแปร และ Output 1 ตัวแปร แบ่งเป็น Positive result คือ 1 และ Negative result คือ 0 ดังภาพประกอบ 8 แล้ววิเคราะห์ข้อมูลทางสถิติเพื่อดูประสิทธิภาพการทำงานของระบบโดยใช้ SPSS17.0 เปรียบเทียบความสามารถในการทำนายของ ANN กับ Logistic Regression พบว่า ANN มี Specificity สูงกว่า Sensitivity ดังภาพประกอบ 9 ทำให้จำแนกผู้ป่วยที่มี PSMVT เป็น positive ในกลุ่มผู้ป่วยที่มี PSMVT เป็น negative พบความแตกต่างระหว่าง False-Positive Rate (FPR, 14.3%) และ False-Negative Rate (FNR, 20.0%) มีความแตกต่างกันเล็กน้อย ซึ่งหมายความว่าแบบจำลอง ANN สามารถปรับสมดุลในเรื่องของ

Sensitivity (SEN) และ Specificity (SPE) และบอกว่า ANNs ให้ความแม่นยำมากกว่า Logistic Regression ในการทำนายการเกิดลิ้มเลือดอุดตัน (PSMVT) หลังจากมีตับอ่อนอักเสบเฉียบพลัน

No.	Variable code	Variable description	Mean (SD or %)	
			Training set (48)	Validation set (24)
1	Age (year)	Age	42.8 (6.6)	42.3 (5.9)
2	Sex	Sex, male, <i>n</i> (%)	26 (54.2%)	13 (54.1%)
3	Hct	Red blood cell specific volume (impacting whole blood viscosity)	0.382 (0.029)	0.392 (0.017)
4	PT (second)	Prothrombin time (indicating exogenous coagulation function)	14.10 (3.76)	14.28 (3.96)
5	FBG (mmol L ⁻¹)	Fasting blood glucose (indicating pancreatic endocrine function)	8.42 (3.17)	7.92 (3.01)
6	D-dimer (mg L ⁻¹)	D-dimer (indicating fibrinolysis function)	0.44 (0.11)	0.45 (0.08)
7	[Ca ²⁺] (mmol L ⁻¹)	Concentration of serum calcium (indicating the severity of acute pancreatitis)	2.31 (0.43)	2.26 (0.37)
8	TG (mmol L ⁻¹)	Triglyceride (indicating lipid metabolism)	2.03 (0.27)	2.17 (0.31)
9	AMY (Unit)	Serum amylase, somogyi method (high diagnostic sensitivity of pancreatitis)	349.1 (76.5)	371.4 (67.9)
10	APACHEII score	Acute physiology and chronic health evaluation II score (indicating the severity of acute pancreatitis, less affected by treatment)	5.02 (1.93)	4.91 (1.76)
11	Ranson score	Ranson score (indicating the severity of acute pancreatitis)	2.16 (0.47)	2.09 (0.29)
12	Output	PSMVT (positive, %)	19 (39.6%)	9 (37.5%)

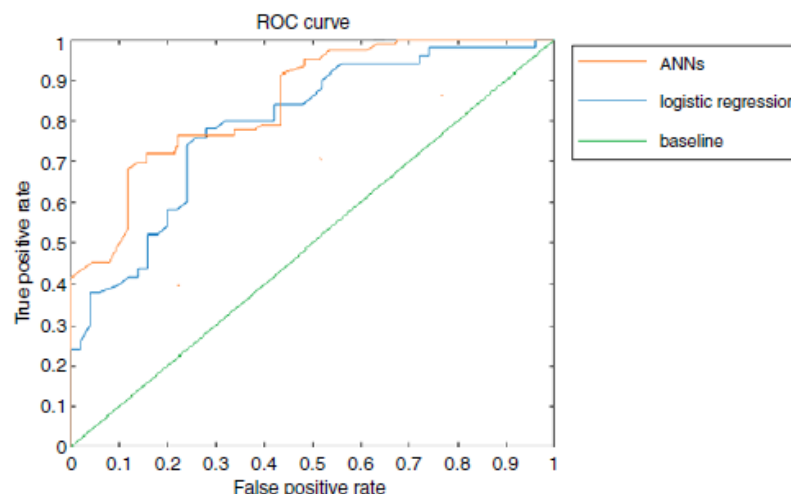
ภาพประกอบ 8 แสดงตัวแปรที่ใช้ input เพื่อฝึกฝนและทดสอบข้อมูล

ที่มา : (Fei, Hu, Li, et al., 2017)

Variable	ANNs model	Logistic regression model	Difference between models (95% CI)	<i>P</i> -value
Sensitivity	80.0%	70.0%	10.0% (−14.3–34.3%)	0.039
Specificity	85.7%	71.4%	14.3% (−8.6–37.2%)	0.027
PPV	77.6%	61.9%	15.7% (−9.9–41.3%)	0.011
NPV	90.7%	78.9%	11.8% (−8.2–31.8%)	0.026
Accuracy	83.3%	60.7%	22.6% (−1.9–47.1%)	0.004
AUC	0.849	0.716	0.298 (0.086–0.397)	0.041

ภาพประกอบ 9 แสดงการเปรียบเทียบประสิทธิภาพแบบจำลอง ANNs กับ Logistic Regression

ที่มา : (Fei, Hu, Li, et al., 2017)



ภาพประกอบ 10 แสดงกราฟ ROC ของ ANNs และ Logistic Regression

ที่มา : (Fei, Hu, Li, et al., 2017)

ในอนาคตงานวิจัยนี้จะปรับปรุงประสิทธิภาพของแบบจำลองโดยอาจจะเพิ่มจำนวนผู้ป่วยให้มากขึ้นเพื่อให้แบบจำลองดีขึ้นและใช้แบบจำลองอื่น เช่น Support Vector Machines, Genetic Algorithm, Clustering Algorithm, Decision Tree และอื่นๆ ที่สามารถ Integrate กับ ANNs เพื่อให้โมเดลทำนายได้มีประสิทธิภาพมากขึ้น

2.4.3 บทความวิจัยเรื่อง Risk Assessment for venous thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patient: A Machine Learning Approach (Ferroni, Zanzotto, Scarpato, Riondino, Nanni, et al., 2017)

บทความนี้ได้ใช้โมเดลการเรียนรู้แบบ Multiple kernel learning (MKL) บนพื้นฐานของ Support Vector Machines (SVM) และ Random optimization (RO) เพื่อทำนายความเสี่ยงของการเกิดลิ่มเลือดอุดตัน ที่ให้ประสิทธิภาพการจำแนกที่ดีที่สุดในช่วงฝึกอบรวม 3-fold cross-validation และชุดการทดสอบ โดยมีข้อมูล $n=1179$ และจัดกลุ่มคนไข้ออกเป็น 9 กลุ่ม ดังภาพประกอบ 11 แล้ววิเคราะห์โมเดลด้วย ML-RO

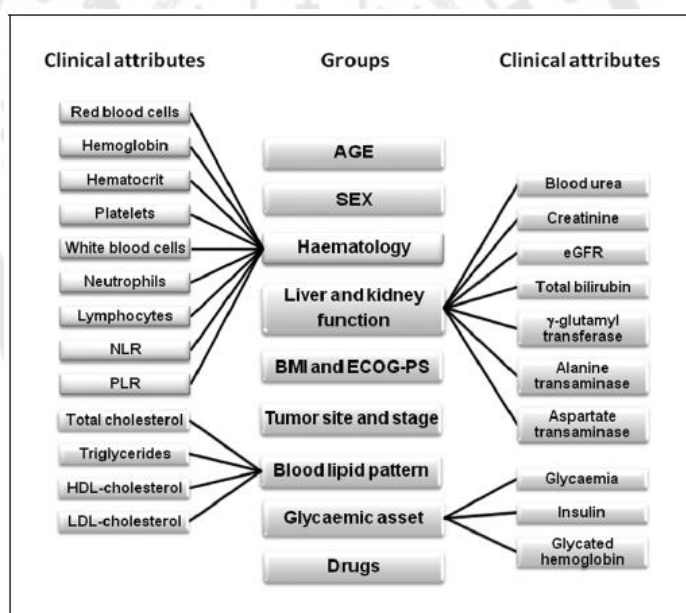
งานวิจัยนี้ได้เน้นถึงข้อดีของการเพิ่มประสิทธิภาพความสัมพันธ์ในกลุ่มของคุณลักษณะทางคลินิกในการเลือกตัวทำนายความเสี่ยงของการเกิดลิ่มเลือดอุดตัน การวิเคราะห์ด้วยการเรียนรู้ Multiple Kernel Learning โดยการทำงานร่วมกันของ Support Vector Machines

(SVM) และ Random optimization (RO) สร้างตัวทำนายความเสี่ยงในประชากรของผู้ป่วยมะเร็งที่เป็นตัวแทนของกลุ่มประชากรทั่วไป เพื่อกำหนดความสำคัญของกลุ่มคุณลักษณะทางคลินิกที่แตกต่างกันของแอททริบิวต์การทำนายขั้นสุดท้ายด้วยฟังก์ชันดัง (2-1) ต่อไปนี้

$$f(x) = \text{sgn}\left(\sum_{i=1}^N \alpha_i \langle \vec{w}_i, \vec{g}_i \rangle + b\right) \quad (2-1)$$

ตัวทำนายปัจจัยเสี่ยงของการเกิดลิ้มเลือดอุดตันถูกเรียนรู้ด้วย n-fold cross-validation บนชุดข้อมูลฝึกฝน และวิเคราะห์หา F-measure จากสมการ (2-2) ดังนี้

$$F - \text{measure} = \frac{2PR}{P+R} \quad (2-2)$$



ภาพประกอบ 11 แสดงกลุ่มของคุณลักษณะทางคลินิก

ที่มา : (Ferroni, Zanzotto, Scarpato, Riondino, Nanni, et al., 2017)

โดยแบ่งข้อมูลเพื่อเรียนรู้และทดสอบออกเป็น 70:30 และวิเคราะห์หา F-measure, Precision (PPV = Positive Predictive Value), Recall (Sensitivity), -LR = Negative Likelihood Ratio และ +LR = Positive Likelihood Ratio ผลลัพธ์ที่ได้พบว่า Support Vector Machines

(SVM) เป็นวิธีการที่เหมาะสมที่สุดในการทำนายความเสี่ยงของการเกิดลิ่มเลือดอุดตัน แสดงดังภาพประกอบ 12

Method	Precision (PPV)	Recall (Sensitivity)	F-Measure	+LR (95% CI)	-LR (95% CI)
Khorana ($k \geq 3$) ^a	0.136	0.111	0.122	1.90 (0.46–6.05)	0.94 (0.76–1.04)
Khorana-ML	0.136	0.111	0.122	1.91 (0.46–6.08)	0.94 (0.76–1.04)
Basic-ML-K	0.099	0.852	0.177	1.33 (1.01–1.50)	0.41 (0.13–0.99)
ML-RO-1-K	0.105	0.741	0.184	1.43 (1.01–1.73)	0.54 (0.24–0.99)
ML-RO-2-K	0.100	0.778	0.177	1.35 (0.98–1.60)	0.53 (0.22–1.03)
ML-RO-3-K	0.112	0.704	0.193	1.52 (1.05–1.90)	0.55 (0.27–0.95)
ML-RO-4-K	0.100	0.667	0.174	1.35 (0.91–1.71)	0.66 (0.34–1.09)
ML-RO-5-K	0.096	0.778	0.171	1.29 (0.94–1.53)	0.56 (0.23–1.10)
Basic-ML	0.078	0.593	0.137	1.02 (0.66–1.35)	0.97 (0.44–1.50)
ML-RO-1	0.082	0.556	0.143	1.08 (0.68–1.47)	0.91 (0.53–1.36)
ML-RO-2	0.122	0.889	0.214	1.68 (1.29–1.86)	0.24 (0.06–0.65)
ML-RO-3	0.119	0.815	0.208	1.64 (1.21–1.90)	0.37 (0.14–0.78)
ML-RO-4	0.092	0.593	0.159	1.23 (0.79–1.63)	0.79 (0.44–1.21)
ML-RO-5	0.108	0.741	0.188	1.46 (1.03–1.77)	0.53 (0.24–0.96)

Note. ML-RO models are ranked according to F-measure on the training set (see Table 1). PPV = positive predictive value; -LR = negative likelihood ratio; CI = confidence interval; +LR = positive likelihood ratio; ML = machine learning; RO = random optimization.
a. Patients with brain cancer ($n = 2$) were excluded from the analysis; Khorana score not applicable.

ภาพประกอบ 12 แสดงผลลัพธ์ของ Basic Predictors และ Predictors Based on Machine Learning with RO

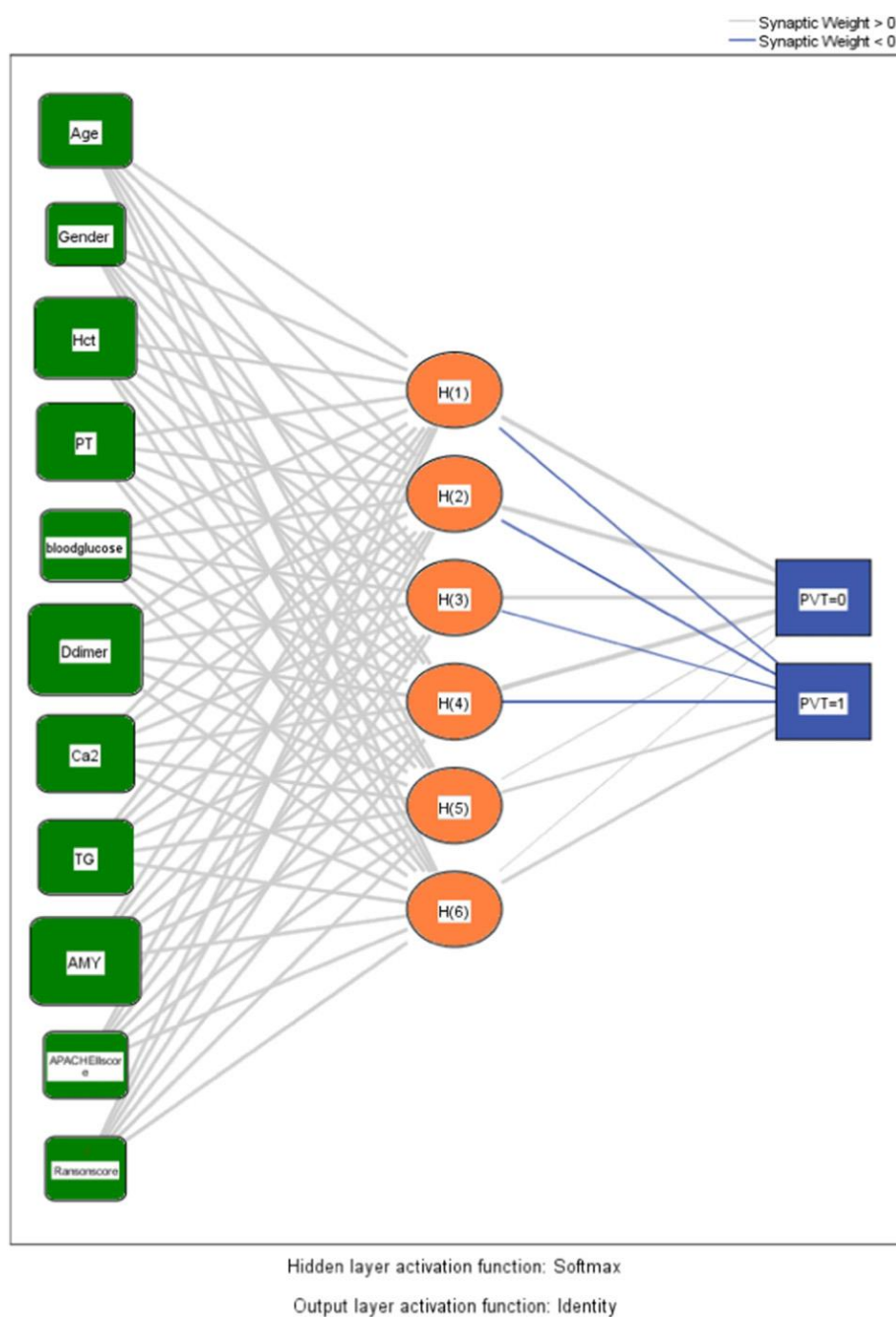
ที่มา : (Ferroni, Zanzotto, Scarpato, Riondino, Nanni, et al., 2017)

ในส่วนของการจำกัดของงานวิจัยนี้ได้ออกแบบและตรวจสอบความถูกต้องบนชุดข้อมูลที่ไม่ได้สกัดจาก HER ของผู้ป่วยรายเดียวเพราะเป็นข้อจำกัดด้านความเป็นส่วนตัวในเรื่องของการระบุตัวตน และวิจัยนี้ยังต้องปรับแต่งมากในเรื่องของโครงสร้างและขอบเขตของข้อมูลที่ไม่ใช่โครงสร้าง ได้แก่ ข้อมูลประชากรทางการแพทย์ประวัติครอบครัว สัญญาณชีพร ยา และการวินิจฉัยอื่นๆ

2.4.4 บทความวิจัยเรื่อง Predicting risk for portal vein thrombosis in acute pancreatitis patients : A comparison of radical basis function artificial neural network and logistic regression models (Fei, Hu, Gao, et al., 2017)

บทความนี้ได้สร้างแบบจำลอง Radical Basis Function (RBF) ANNs model และ Logistic regression model เพื่อทำนายอุบัติการณ์ของการเกิดตับอ่อนอักเสบเฉียบพลัน (Acute Pancreatitis, AP) ที่ทำให้เกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามมา โดยได้ input ปัจจัยเสี่ยงทั้งหมด 11 ปัจจัย และ Output ออกมาเป็น Binary Classification ดังภาพประกอบ 13 แล้วทำ

การเปรียบเทียบโมเดล RBF ANNs กับ Logistic Regression พบว่า RBF ANNs ให้ประสิทธิภาพของ SEN, SPE, PPV และ Accuracy ดีกว่า Logistic Regression แสดงดังภาพประกอบ 14



ภาพประกอบ 13 แสดงโครงสร้างของแบบจำลอง Radial Basis Function (RBF) ANNs

ที่มา : (Fei, Hu, Gao, et al., 2017)

Variable	RBF ANNs model	Logistic regression model	Difference between models (95%CI)	P-value
Sensitivity	73.3%	63.9%	9.4% (− 0.2%–17.0%)	0.032
Specificity	91.4%	83.1%	8.3% (− 4.6%–21.7%)	0.044
PPV	68.8%	52.0%	16.8% (5.8%–25.2%)	0.006
NPV	93.0%	89.9%	3.1% (− 7.9%–19.6%)	0.107
Accuracy	87.7%	78.5%	9.2% (0.3%–21.6%)	0.016
AUC	0.904	0.736	0.168 (0.003–0.262)	0.029

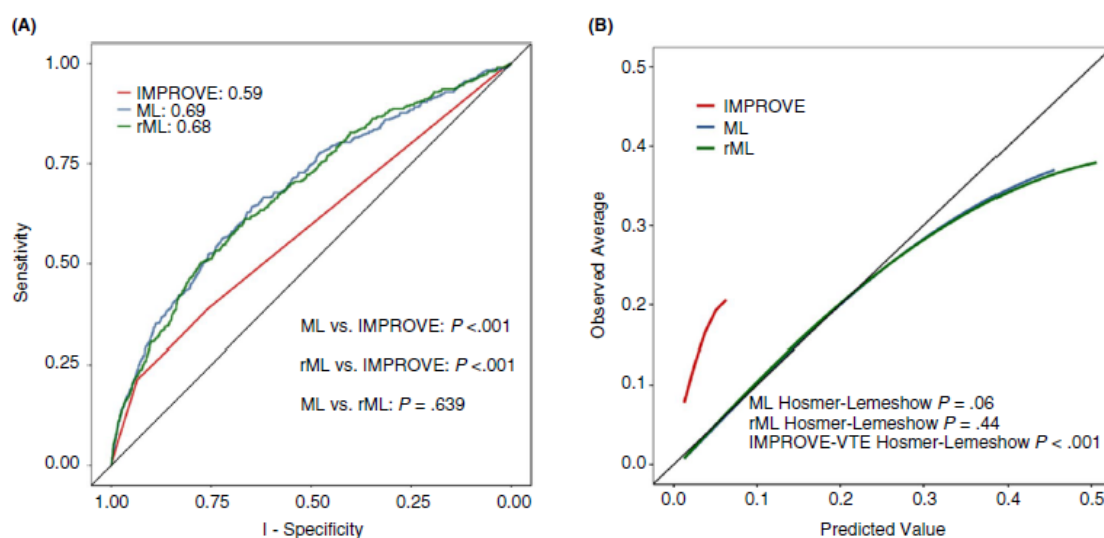
ภาพประกอบ 14 เปรียบเทียบแบบจำลอง RBF ANNs และ Logistic regression สำหรับทำนายการเกิดลิ่มเลือดอุดตันของผู้ป่วยตับอ่อนอักเสบเฉียบพลัน

ที่มา : (Fei, Hu, Gao, et al., 2017)

จากภาพประกอบ 14 แบบจำลอง RBF ANNs มีแนวโน้มที่สามารถทำนายการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำในผู้ป่วยที่เป็นตับอ่อนอักเสบเฉียบพลัน (Acute Pancreatitis, AP) ได้มีประสิทธิภาพมากกว่า Logistic Regression

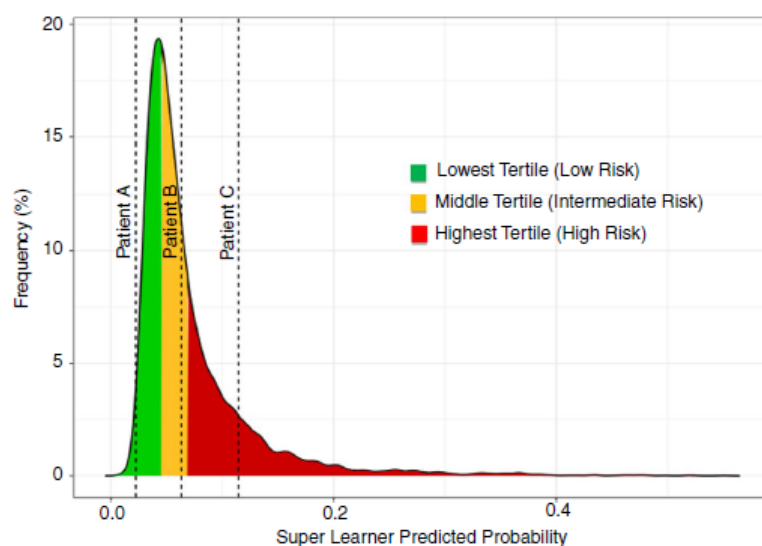
2.4.5 บทความวิจัยเรื่อง Machine learning to predict venous thrombosis in acutely ill medical patients (Nafee et al., 2020)

บทความนี้ได้ใช้ Super Learner Model (ML) และ Reduced Model (rML) เทียบกับ IMPROVE Score (International Medical Prevention Registry on Venous Thromboembolism) ในการทำนายการเกิดลิ่มเลือดอุดตัน ทำให้ทราบว่าผู้ป่วยที่เสี่ยงเกิดภาวะลิ่มเลือดอุดตันมีโอกาสเพิ่มขึ้นเป็น 5 เท่าของภาวะการเกิดลิ่มเลือดอุดตันที่ Tertile ต่ำสุด และพบว่าวิธี Super Learner Model (ML) ได้ค่า C-Statistic สูงสุดสำหรับการทำนายการเกิดลิ่มเลือดอุดตัน แสดงดังภาพประกอบ 15 เมื่อเทียบกับ IMPROVE Score และประเมินความเสี่ยงที่ได้รับ การสอบเทียบกับผลลัพธ์ที่เกิดจากการสังเกตการณ์ แสดงผลลัพธ์ดังภาพประกอบ 16



ภาพประกอบ 15 แสดงการเปรียบเทียบ ROC curves ของ Super learner model (ML) และ Reduced model (rML) เทียบกับ IMPROVE score

ที่มา : (Nafee et al., 2020)



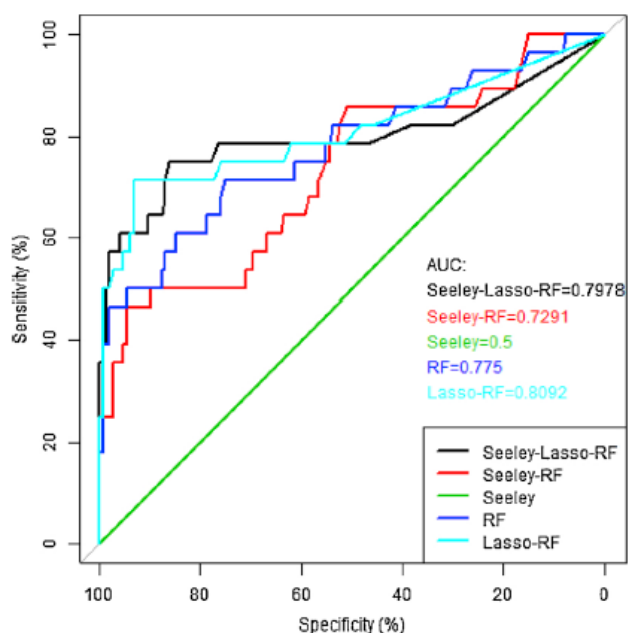
ภาพประกอบ 16 แสดงการกระจายของความถี่ที่ทำนายด้วย rML

ที่มา : (Nafee et al., 2020)

อนาคตงานวิจัยนี้พัฒนา Super Learner Model (ML) ให้ใช้งานบนแพลตฟอร์มบนมือถือ ซึ่งจะเพิ่มประสิทธิภาพทั้งความแม่นยำและความสะดวก โดยจะเพิ่มข้อมูลในการทำนายให้มากขึ้น

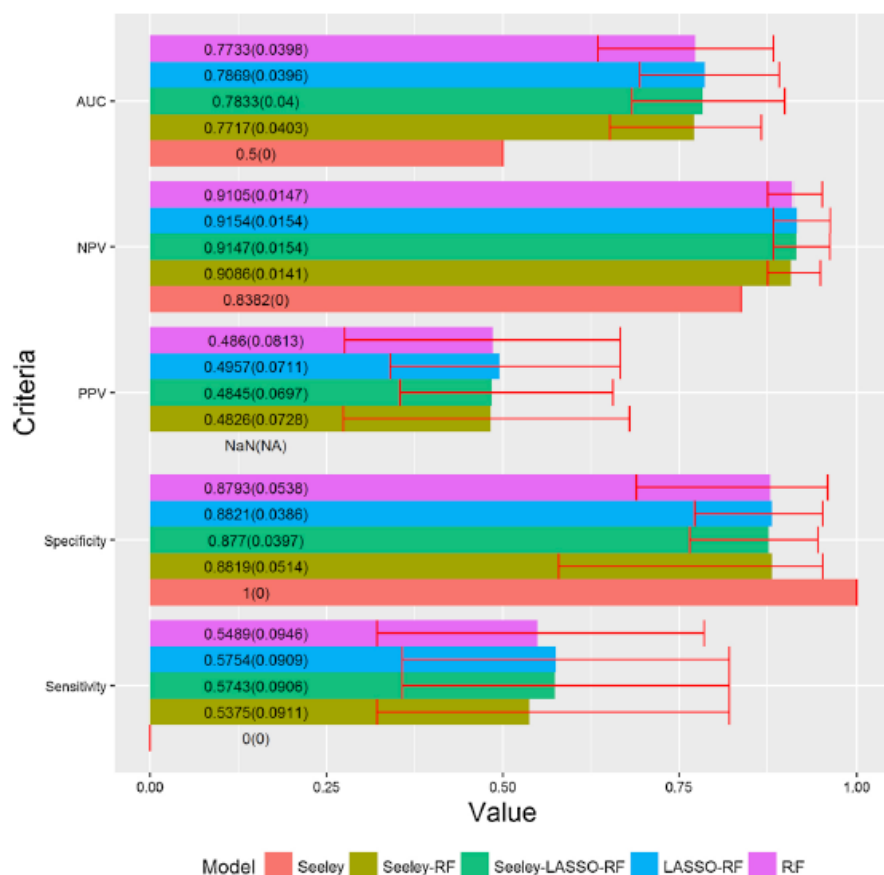
2.4.6 บทความวิจัยเรื่อง Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer (Liu et al., 2019)

บทความนี้ได้ใช้เทคนิค Machine Learning ประเมินประสิทธิภาพที่ระบุความสามารถในการทำนายความเสี่ยงของการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ใส่สายสวน PICC ได้ โดยใช้ 5 โมเดล ได้แก่ Seely, Seely-RF, Seely-LASSO-RF, RF และ LASSO-RF เพื่อเปรียบเทียบผลลัพธ์การประเมินประสิทธิภาพของโมเดล แสดงดังภาพประกอบ 17 และเปรียบเทียบกับผลลัพธ์ของโมเดลที่ใช้อยู่ในปัจจุบัน (Seely) กับ machine learning พบว่า machine learning ให้ประสิทธิภาพที่ดีกว่า แสดงดังภาพประกอบ 18 เพื่อช่วยในการตัดสินใจในการป้องกันโรคและลดอัตราการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ใส่สายสวน PICC ได้อย่างมีประสิทธิภาพ



ภาพประกอบ 17 แสดงการเปรียบเทียบ ROC curves ของแต่ละโมเดล

ที่มา (Liu et al., 2019)



ภาพประกอบ 18 เปรียบเทียบประสิทธิภาพของ Seely model กับ Machine learning models

ที่มา : (Liu et al., 2019)

ในอนาคตงานวิจัยนี้จะนำเรื่องของอนุกรมเวลาและตัวแปรทางการแพทย์อื่นๆที่เกี่ยวข้องมาวิเคราะห์ให้ละเอียดมากขึ้น

2.4.7 บทความวิจัยเรื่อง Validation of a Machine Learning Approach for Venous Thromboembolism Risk Prediction in Oncology (Ferroni, Zanzotto, Scarpato, Riondino, Guadagni, et al., 2017)

บทความนี้ได้ใช้เทคนิค Machine Learning (ML) และ Random Optimization (RO) พัฒนาในเรื่องของการทำนายความเสี่ยงของการเกิดหลอดเลือดอุดตัน เพื่อที่จะเป็นแนวทางในการออกแบบเว็บไซต์เพื่อแยกความเสี่ยงของการเกิดลิ่มเลือดอุดตันในผู้ป่วยมะเร็งที่ได้รับ

เคมีบำบัด และเปรียบเทียบประสิทธิภาพของทั้ง 2 เทคนิคดังกล่าวกับเทคนิคที่ใช้ในปัจจุบันซึ่งก็คือ Khorana Score (KS)

โดยจะใช้เทคนิค Machine Learning (ML) และ Random Optimization (RO) กำหนดความสำคัญของกลุ่มต่างๆทางคลินิก ซึ่งมีปัจจัยต่างๆ เช่น อายุ เพศ ขนาดของก้อนเนื้องอก ลักษณะทางโลหิตวิทยา ไขมันในเลือด การทำงานของตับและไต ค่าดัชนีมวลกาย สถานภาพ และยาต้านมะเร็งที่ผู้ป่วยได้รับ จากข้อมูลทั้งหมด $n = 1433$ ปัจจัยดังกล่าวได้ถูกป้อนเข้าไปวิเคราะห์ในรูปแบบตัวเลข

Receiver operating characteristics	KS*	ML predictor
Sample size	605	608
Area under the ROC curve**	0.589	0.716
Standard error	0.042	0.036
Positive likelihood ratio	1.58 (0.48–4.30)	2.30 (1.70–2.82)
Negative likelihood ratio	0.96 (0.83–1.04)	0.46 (0.28–0.69)

ภาพประกอบ 19 แสดงประสิทธิภาพเทคนิค Khorana Score (KS) และเทคนิค ML-RO

ที่มา : (Ferroni, Zanzotto, Scarpato, Riondino, Guadagni, et al., 2017)

จากภาพประกอบ 19 ผลลัพธ์ของงานวิจัยนี้พบว่าวิธี ML-RO มีประสิทธิภาพกว่าวิธี Khorana Score (KS) จึงเหมาะและเป็นประโยชน์ในการออกแบบ Web Service ที่สามารถอินเตอร์เฟซแก่แพทย์เพื่อช่วยในขั้นตอนการตัดสินใจรักษาผู้ป่วย

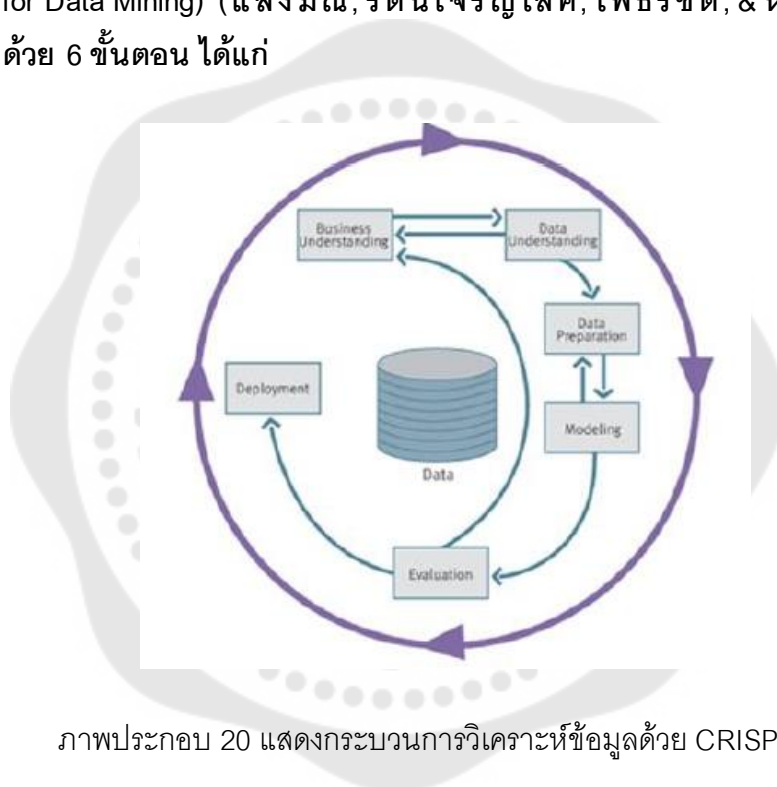
2.5 งานวิจัยที่ศึกษาเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่

งานวิจัยนี้ได้ศึกษาเพิ่มเติมเกี่ยวกับเทคนิคการวิเคราะห์ข้อมูลขนาดใหญ่ ซึ่งมีรายละเอียดดังต่อไปนี้

ศึกษางานวิจัยเรื่อง การสร้างโมเดลการทำนายการกลับมารักษาซ้ำของผู้ป่วยโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล (Building the Predicting Model of the Chance of Repeating the Treatment of Diabetes Patient Using Data Mining) (หนูสวัสดิ์, 2017) เป็น

งานวิจัยที่ประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อทำนายการกลับมารักษารักษาซ้ำของผู้ป่วยโรคเบาหวาน โดยวิเคราะห์จากปัจจัยเสี่ยงที่จะกลับมารักษารักษาซ้ำ โดยทำการสร้างโมเดลการทำนายจากเทคนิค Decision Tree, Naïve Bays และ K-Nearest Neighbors แล้วเปรียบเทียบประสิทธิภาพของแต่ละโมเดล เพื่อนำไปประยุกต์ใช้ในการสร้างระบบสนับสนุนการตัดสินใจแนวทางการรักษาของแพทย์ ซึ่งมีทฤษฎีและแนวคิดในการทำวิจัยดังนี้

กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM (Cross-Industry Standard Process for Data Mining) (แสงมณี, รัตนเจริญเลิศ, โพธิ์รัชต์, & หนูสวัสดิ์, 2017) ประกอบด้วย 6 ขั้นตอน ได้แก่



ภาพประกอบ 20 แสดงกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM

ที่มา (แสงมณี et al., 2017)

1.การทำความเข้าใจในปัญหา : Business Understanding

เป็นขั้นตอนแรกสุดในกระบวนการ CRISP-DM เป็นการทำความเข้าใจระบุปัญหาระบุ output ที่ต้องการได้จากการวิเคราะห์ข้อมูลด้วย Data mining โดยงานวิจัยเรื่องนี้ได้อธิบายถึงความสำคัญ สาเหตุ ความร้ายแรงของการเกิดโรคเบาหวาน

2.การทำความเข้าใจกับข้อมูล : Data Understanding

เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมา

ใช้ในการวิเคราะห์งานวิจัยดังกล่าวได้ใช้ชุดข้อมูลของผู้ป่วยโรคเบาหวานหรือเสี่ยงจะเป็นโรคเบาหวาน โรงพยาบาลของสหรัฐอเมริกา ตั้งแต่ปี 2542-2551 (หนูสวัสดิ์, 2017)

3.การเตรียมข้อมูล : Data Preparation

ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (Raw Data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (Data Cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (Scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น ซึ่งงานวิจัยดังกล่าวได้ทำการตัดแอททริบิวต์ที่ไม่ส่งผลต่อค่าคำตอบ โดยจะเลือกเฉพาะแอททริบิวต์หรือปัจจัยที่ส่งผลต่อการทำนายโอกาสการกลับมารักษาตัวของโรคเบาหวาน ยังได้ทำการลดความซ้ำซ้อนของข้อมูลโดยเลือกทำการแปลงข้อมูลอายุของผู้ป่วยออกเป็นช่วงๆ

4.การสร้างโมเดล : Modeling

เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิค Data mining เช่น การจำแนกประเภทของข้อมูล (Classification) โดยงานวิจัยดังกล่าวได้ใช้เทคนิค Decision Tree, Naïve Bays และ K-Nearest Neighbors และใช้โปรแกรม Rapid miner studio 7.2 และแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพแบบ วิธี Cross-validation test โดยแบ่งข้อมูล ค่า K=10 กลุ่ม, ค่า K=15 กลุ่ม และ ค่า K=20 กลุ่ม (หนูสวัสดิ์, 2017) เพื่อหาความถูกต้องของโมเดลที่มากที่สุด

5.การประเมินผล : Evaluation

ในขั้นตอนนี้เป็นการประเมินประสิทธิภาพของผลลัพธ์จากโมเดลวิเคราะห์ข้อมูลว่าครอบคลุมและสามารถตอบโจทย์หรือไม่ ผลลัพธ์ของการประเมินผลของงานวิจัยดังกล่าวพบว่าเทคนิค Decision tree ให้ค่าความถูกต้องของโมเดลดีที่สุด แสดงผลดังตาราง 1

ตาราง 1 แสดงการเปรียบเทียบค่าความถูกต้องของ 3 อัลกอริทึม

	Decision Tree	Naïve Bayes	K-Nearest Neighbor
Accuracy (%)	85.50	84.62	76.72

6.การนำไปใช้งาน : Deployment

การนำผลลัพธ์หรือองค์ความรู้ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining นำไปใช้งานจริง ซึ่งงานวิจัยดังกล่าวนำไปประยุกต์ใช้ในทางการแพทย์ เพื่อช่วยสนับสนุนการตัดสินใจ ลดระยะเวลาในการทำการรักษาของแพทย์ได้



บทที่ 3

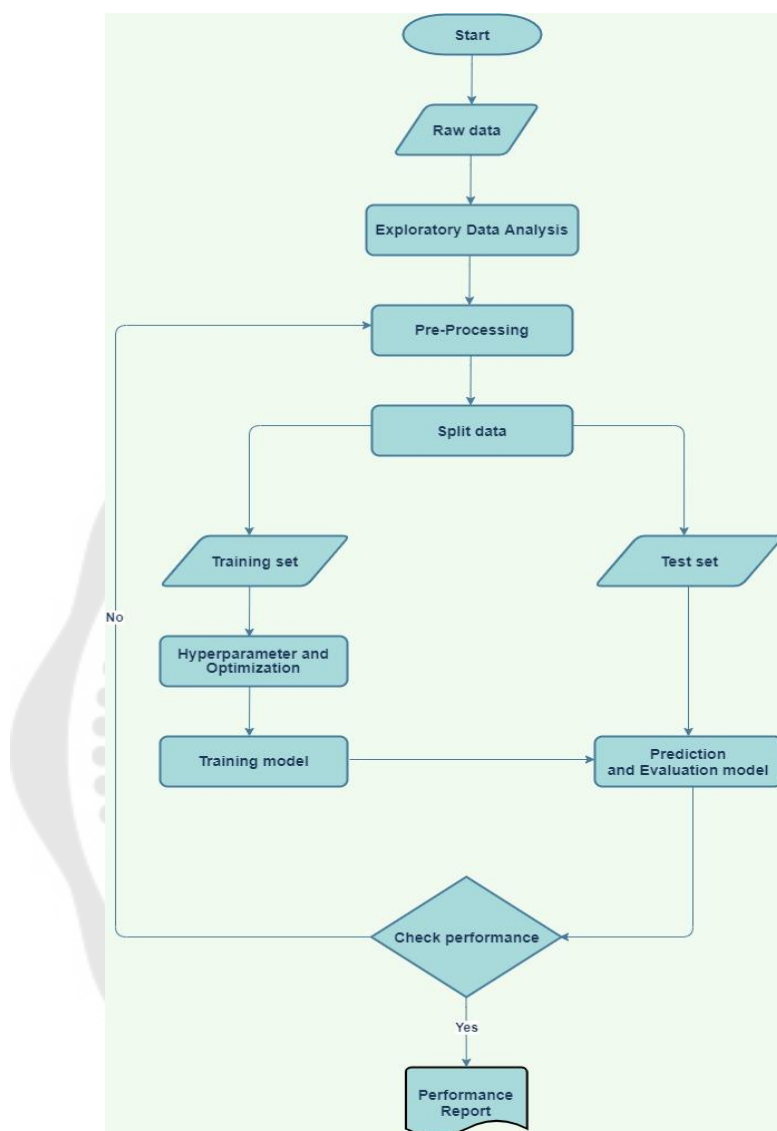
วิธีการดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

- 3.1 กระบวนการทำงานของแบบจำลอง
- 3.2 การรวบรวมข้อมูล (Data Acquisition)
- 3.3 การคัดกรองข้อมูล (Data Filtering)
- 3.4 การสำรวจข้อมูล (Exploratory Data Analysis : EDA)
- 3.5 การเตรียมข้อมูล (Data Preparation)
- 3.6 หลักการที่ใช้ในการสร้างแบบจำลอง (Modeling)
- 3.7 การประเมินผล (Evaluation)



3.1 กระบวนการทำงานของแบบจำลอง



ภาพประกอบ 21 แสดงแผนภาพกระบวนการทำงานของแบบจำลอง

จากแผนภาพกระบวนการทำงานของแบบจำลองที่แสดงในภาพประกอบ 21 อธิบายถึงการทำงานของแบบจำลองเพื่อเปรียบเทียบสมรรถนะของแต่ละอัลกอริทึม โดยเริ่มจากการนำเข้าข้อมูลดิบ ทำการสำรวจข้อมูล (Exploratory Data Analysis : EDA) โดยทำการวิเคราะห์หาความสัมพันธ์ระหว่างตัวแปรต่างๆ หรือปัจจัยเสี่ยงของการเกิดล้มเลือดอุดตัน จากนั้นทำการเตรียมข้อมูล ตรวจสอบค่า Missing Values แปลงข้อมูลให้เหมาะสมเพื่อพร้อมที่จะนำเข้าแบบจำลองในการวิเคราะห์ต่อไป และแบ่งข้อมูลสำหรับการฝึกฝน (Training Data) และข้อมูล

สำหรับการทดสอบ (Testing Data) ด้วยอัตราส่วน 70:30 จากนั้นปรับจูนหาไฮเปอร์พารามิเตอร์ที่เหมาะสม และดีที่สุดที่สามารถทำให้แบบจำลองทำนายได้มีประสิทธิภาพมากที่สุดแล้ว สร้างแบบจำลองโดยทดลองใช้อัลกอริทึมการเรียนรู้ของเครื่อง 3 อัลกอริทึม ประกอบด้วย Decision Tree, Logistic Regression และ Neural Network สุดท้ายทำการประเมินประสิทธิภาพของแบบจำลองว่ามีประสิทธิภาพมากน้อยเพียงใด ถ้ายังมีประสิทธิภาพมากก็ทำการตรวจสอบในส่วนของการทำ Pre-Processing อีกครั้ง

3.2 การเก็บรวบรวมข้อมูล (Data Acquisition)

งานวิจัยนี้นำข้อมูลที่ถูกระบุรวบรวมโดยแพทย์ผู้ทำวิจัยหาปัจจัยเสี่ยงของการเกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำ โดยเก็บข้อมูลจากผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยแผนกอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ ภายในปี 2009 ที่ได้รับการอนุมัติจากคณะกรรมการจริยธรรมของคณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยแล้ว ซึ่งข้อมูลประกอบด้วยทั้งหมด 1,290 แถว คือจำนวนผู้ป่วยที่ถูกสำรวจ และมี 65 คอลัมน์ คือปัจจัยต่างๆที่นำมาประกอบการพิจารณาหาความเสี่ยงของการเกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำ ซึ่งได้มาในลักษณะเป็นไฟล์ .csv แสดงคำอธิบายข้อมูลของแต่ละตัวแปรดังต่อไปนี้

ตาราง 2 แสดงคำอธิบายข้อมูลของตัวแปรในแต่ละคอลัมน์

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
1	No	เลขที่
2	Number	เลขที่
3	Ward	หอผู้ป่วย

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
4	GPrivateICU	General ward สามัญ Private ward พิเศษ ICU ward
5	Age	อายุ
6	Date	วันที่เก็บข้อมูล
7	Sex	เพศ
8	Weight	น้ำหนัก
9	Height	ส่วนสูง
10	BMI	ดัชนีมวลกาย
11	Race	สัญชาติ
12	Immobilized	การไม่ขยับเคลื่อนไหว

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
13	CHF	โรคหัวใจวาย
14	Paraparesis	ขาอ่อนแรง 2 ข้าง
15	Cancer	โรคมะเร็ง
16	CAType	ชนิดมะเร็ง
17	Lymphoma	มะเร็งต่อมน้ำเหลือง
18	Hemato	มะเร็งโลหิตวิทยา
19	Stage	ระยะของมะเร็ง
20	Respirator	ใช้เครื่องช่วยหายใจ
21	COPD	โรคปอดอุดกั้นเรื้อรัง
22	CompressionFx	กระดูกสันหลังหักแบบยุบตัว

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
23	Infection	การติดเชื้อ
24	Hemiparesis	การอ่อนแรงครึ่งซีก
25	BedRidden	คนไข้ติดเตียง
26	SLE	โรค Systemic Lupus Erythematosus
27	Arthritis	ข้ออักเสบ
28	Surgery	หลังผ่าตัด
29	SType	ชนิดการผ่าตัด
30	HxVTE	ประวัติของ VTE ในอดีต
31	FamilyVTE	ประวัติครอบครัวมี VTE
32	Hxcancer	ประวัติมะเร็งในอดีต

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
33	Varicose	เส้นเลือดขอดที่ขา
34	Estrogen	ได้รับ Estrogen เช่น ยาคุมกำเนิด
35	Thrombophilia	ภาวะเลือดแข็งตัวง่าย
36	Ttype	ชนิดของภาวะเลือดแข็งตัวง่าย
37	Platelet	ระดับเกล็ดเลือด
38	Fat	อ้วน
39	Nephrotic	โรคไตที่มีโปรตีนรั่วออกมา
40	@4Risk	มีปัจจัยเสี่ยงมากกว่า 4 ข้อ
41	Prophylaxis	การได้รับยาป้องกัน VTE
42	OtherIndi	การได้รับยาด้านการแข็งตัวของเลือด

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
43	IndiType	ข้อบ่งชี้ในการได้ยาต้านการแข็งตัวของเลือด
44	LMWH	Low Molecular Weight Heparin
45	UFH	Unfractionated Heparin
46	Warfarin	ยา Warfarin
47	Autoimmune	โรคแพ้ภูมิตัวเอง
48	FinalDx	การวินิจฉัยโรคสุดท้าย
49	Comorbid	โรคที่เป็นร่วม
50	Complications	ภาวะแทรกซ้อน
51	Procedure	หัตถการที่ทำ
52	Death	Death

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
53	DeadDate	วันที่เสียชีวิต
54	CauseofDeath	สาเหตุการเสียชีวิต
55	CauseOfDeath1	สาเหตุการเสียชีวิต
56	Autopsy	ได้รับการตรวจศพ
57	DischargeDate	วันที่ออกจากโรงพยาบาล
58	LOS	Length of stay ใน โรงพยาบาล
59	DateEval1	วันที่ตรวจหลัง Discharge
60	DateEval2	วันที่ตรวจหลัง Discharge
61	OPDdeath	เสียชีวิตหลังกลับบ้าน
62	OPDdeadDate	วันที่เสียชีวิตหลังกลับบ้าน

ตาราง 2 (ต่อ)

ลำดับ	ตัวแปรข้อมูล	คำอธิบายข้อมูล
63	OPDCOD	สาเหตุการตาย
64	totalDeath	ผู้ป่วยที่เสียชีวิตทั้งหมด
65	VTEpositive	ภาวะลิ่มเลือดอุดตัน

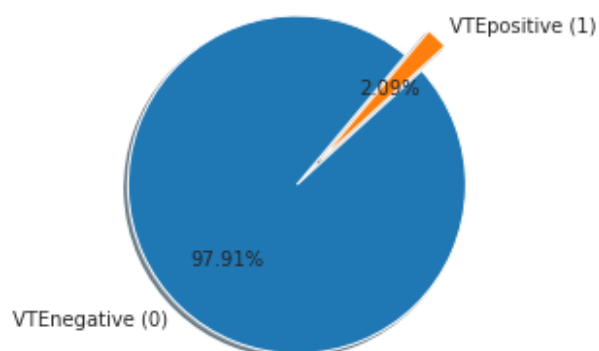
3.3 การคัดกรองข้อมูล (Data Filtering)

งานวิจัยนี้นำข้อมูลที่ถูกรวบรวมโดยแพทย์ผู้ทำวิจัยหาปัจจัยเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ โดยเก็บข้อมูลจากผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยแผนกอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ ภายในปี 2009 ที่ได้รับการอนุมัติจากคณะกรรมการจริยธรรมของคณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยแล้ว ซึ่งข้อมูลประกอบด้วยทั้งหมด 1,290 แถว คือจำนวนผู้ป่วยที่ถูกสำรวจ และมี 65 คอลัมน์ คือปัจจัยต่างๆที่นำมาประกอบการพิจารณาหาความเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำซึ่งจะรวมถึงคอลัมน์ที่เป็นคำตอบ (Class Label) ด้วย

นำข้อมูลทั้งหมดวิเคราะห์หาปัจจัยเสี่ยงของภาวะการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำ วิเคราะห์ออกมาในรูปแบบ Pie Chart พบว่าผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ VTE positive มีอยู่ 2.09 เปอร์เซ็นต์ และผู้ป่วยที่ไม่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ VTE negative มีอยู่ 97.91 เปอร์เซ็นต์ ของข้อมูลทั้งหมด แสดงดังภาพประกอบ 22 ซึ่งผู้วิจัยได้จัดให้เป็นในส่วนของ Class Label โดยกำหนดให้

0 คือ ผู้ป่วยที่ไม่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ (VTE negative)

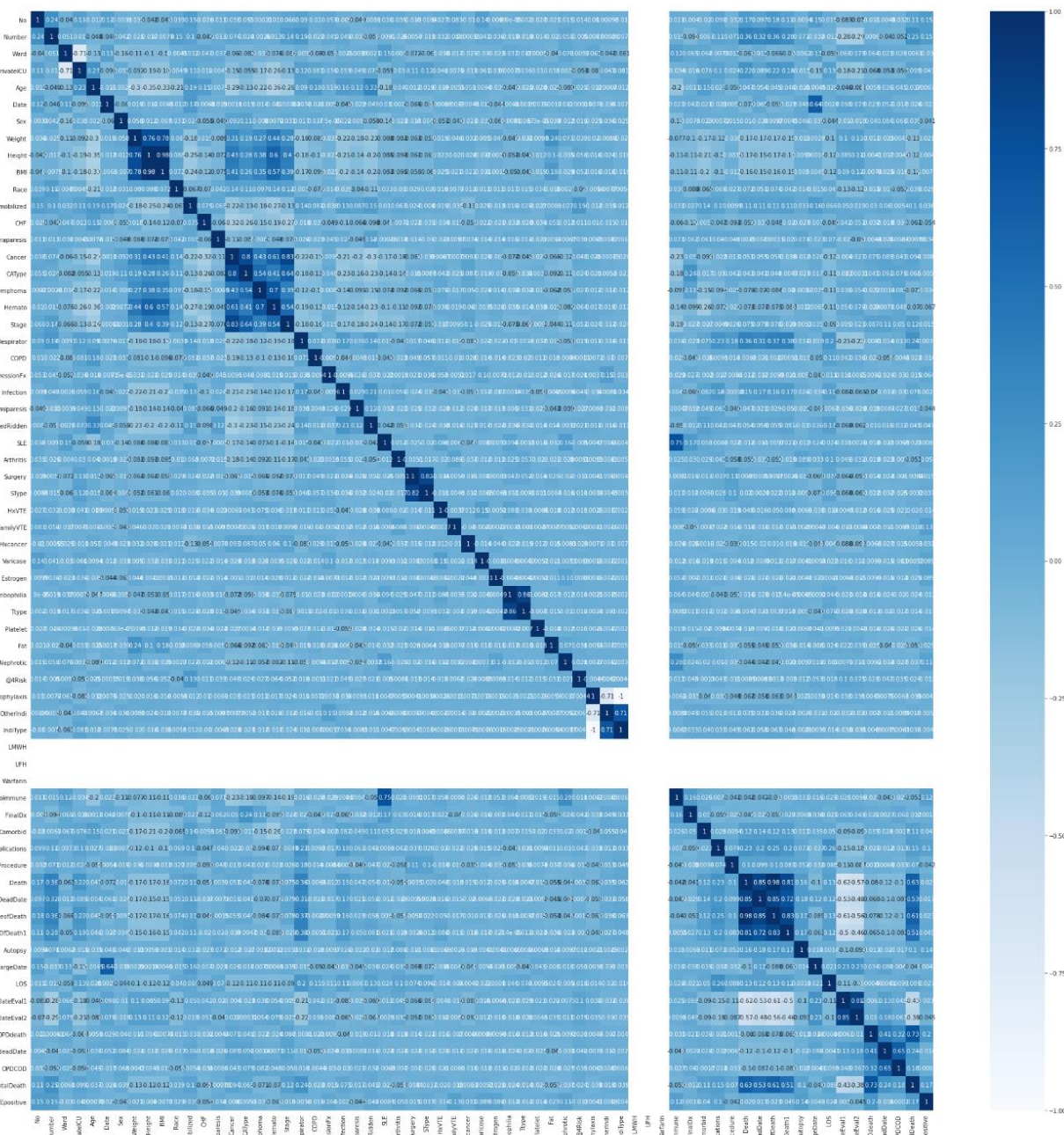
1 คือ ผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตันในหลอดเลือดดำ (VTE positive)



ภาพประกอบ 22 แสดงจำนวนผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตัน VTE positive(1) และไม่เป็นโรคลิ่มเลือดอุดตัน VTE negative(0)

3.4 การสำรวจข้อมูล (Exploratory Data Analysis : EDA)

วิเคราะห์ความสัมพันธ์ระหว่างข้อมูลหรือตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไปว่ามีความสัมพันธ์กันในระดับใด และมีความสัมพันธ์กันในทิศทางใด เช่น มีความสัมพันธ์กันมากหรือน้อยและสัมพันธ์ในทิศทางเดียวกันหรือตรงกันข้ามกันของข้อมูลปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ่มเลือดอุดตันทั้งหมด โดยใช้ Heatmap ได้ผลดังภาพประกอบ 23



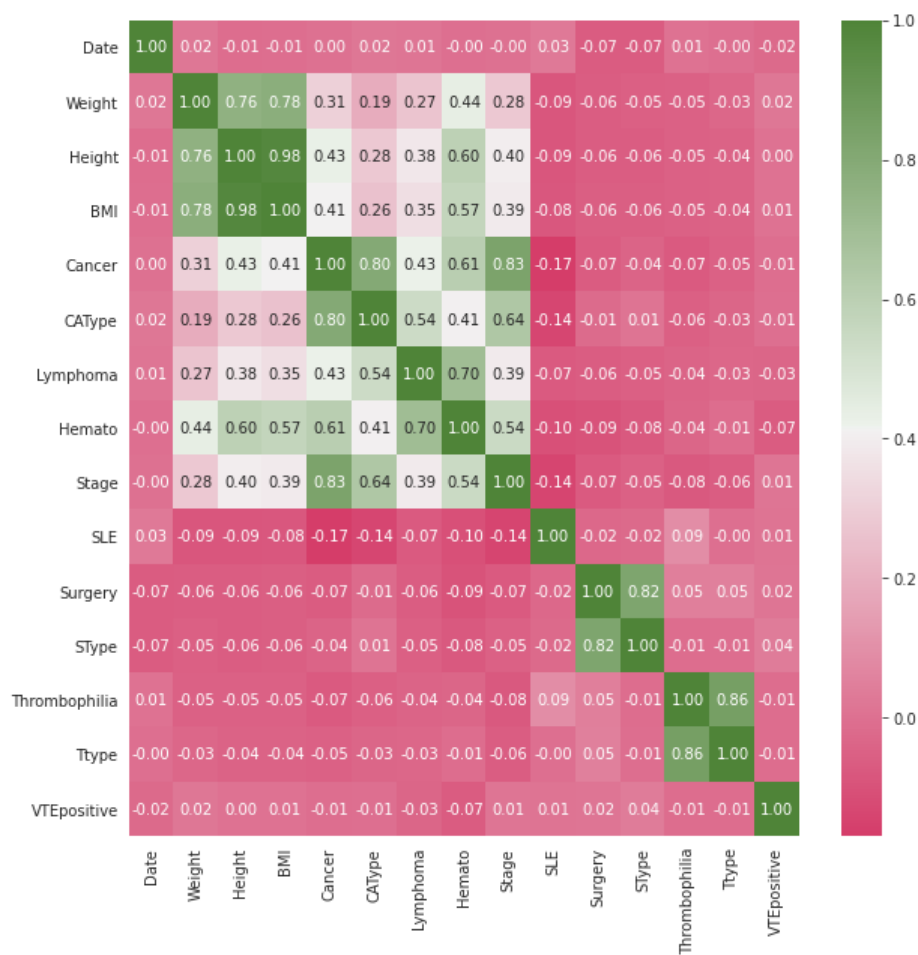
ภาพประกอบ 23 แสดงความสัมพันธ์ของข้อมูลที่น่ามาพิจารณาปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ้ม

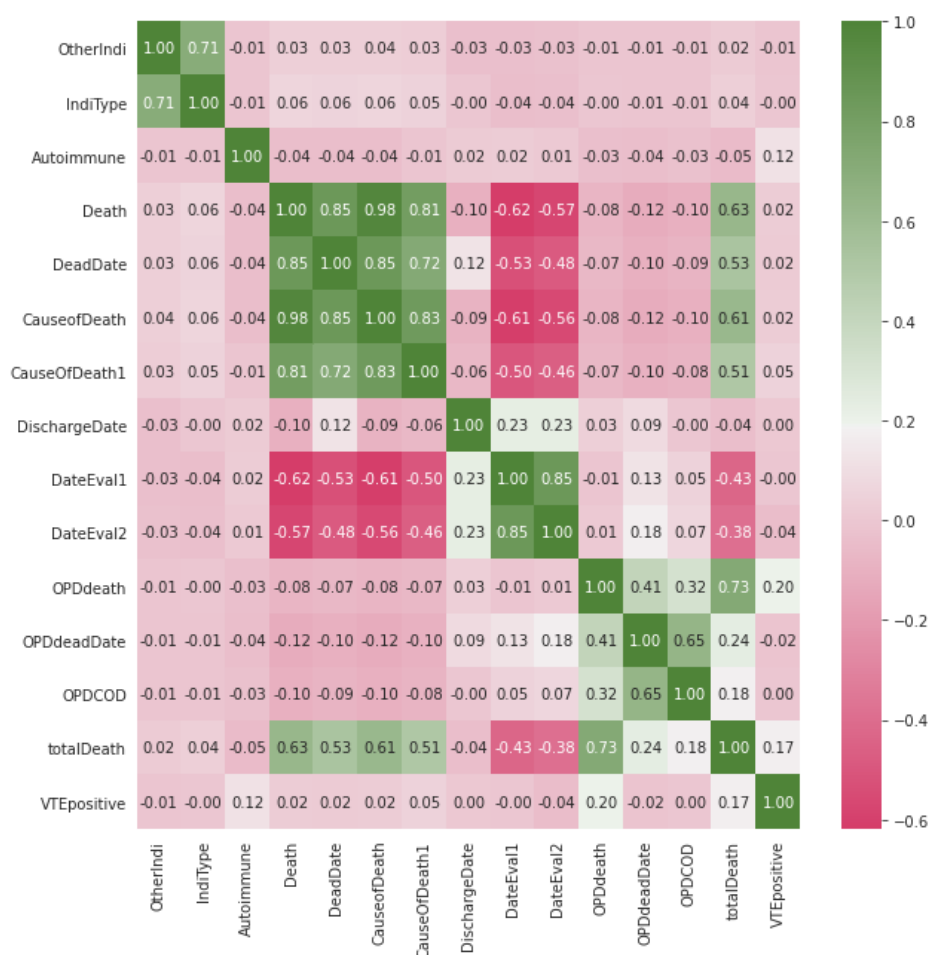
เลือดอุดตันในหลอดเลือดดำจากข้อมูลทั้งหมด

จากภาพประกอบ 23 พบว่าตัวแปรแต่ละคอลลัมน์มีความสัมพันธ์กันน้อยมากเมื่อเทียบกับ Class Label เนื่องจากข้อมูลทั้งหมดที่แพทย์ผู้เชี่ยวชาญได้เก็บข้อมูลไว้นั้นเป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Variable) เช่น เทียบกลุ่มอายุ มากกว่า 40 ปี หรือ น้อยกว่า 40 ปี ไม่ได้ให้อายุที่เป็นปัจจุรย์กับการเป็น VTE positive และ VTE negative จึงทำให้ไม่มีความสัมพันธ์กันระหว่าง

ตัวแปรในแต่ละคอลัมน์กับ Class Label จากข้อมูลทั้งหมด 1,290 แถว 65 คอลัมน์ ผู้วิจัยได้เลือก
พลอตกราฟเฉพาะคอลัมน์ที่มีความสัมพันธ์กันมากที่สุดจากข้อมูลทั้งหมดแสดงดังภาพประกอบ

24





ภาพประกอบ 24 แสดงความสัมพันธ์ของปัจจัยเสี่ยงที่สัมพันธ์กันมากที่สุด

จากภาพประกอบ 24 ความสัมพันธ์ของแต่ละตัวแปรปัจจัยเสี่ยงที่ทำให้เกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำนั้น ซึ่งมีตัวแปรจำนวนมากผู้วิจัยได้เลือกเฉพาะตัวแปรที่มีความสัมพันธ์กันสูงๆ มีค่าสูงๆ ดังต่อไปนี้

1. ความสัมพันธ์ของ Height (ส่วนสูง) กับ BMI (ดัชนีมวลกาย) มีค่าเท่ากับ 0.98
2. ความสัมพันธ์ของ Death (ผู้ป่วยเสียชีวิตในโรงพยาบาล) กับ CauseofDeath (สาเหตุของการเสียชีวิต) มีค่ามากที่สุดเท่ากับ 0.98
3. ความสัมพันธ์ของ DeadDate (วันที่เสียชีวิต) กับ CauseofDeath (สาเหตุของการเสียชีวิต) มีค่ามากที่สุดเท่ากับ 0.85
4. ความสัมพันธ์ของ Death (ผู้ป่วยเสียชีวิตในโรงพยาบาล) กับ DeadDate (วันที่เสียชีวิต) มีค่ามากที่สุดเท่ากับ 0.85

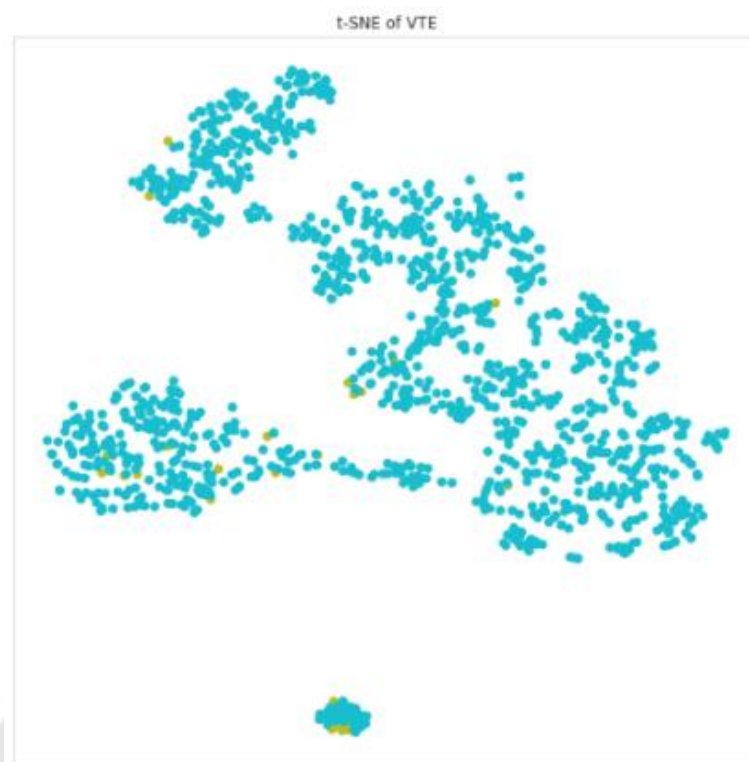
5. ความสัมพันธ์ของ Cancer (โรคมะเร็ง) กับ Stage (ระยะของมะเร็ง) มีค่าเท่ากับ 0.83
6. ความสัมพันธ์ของ Death (ผู้ป่วยเสียชีวิตในโรงพยาบาล) กับ CauseofDeath1 (สาเหตุการเสียชีวิต) มีค่ามากที่สุดเท่ากับ 0.81
7. ความสัมพันธ์ของ Cancer (โรคมะเร็ง) กับ CAType (ชนิดมะเร็ง) มีค่าเท่ากับ 0.80
8. ความสัมพันธ์ของ BMI (ดัชนีมวลกาย) กับ Weight (น้ำหนัก) มีค่าเท่ากับ 0.78
9. ความสัมพันธ์ของ Weight (น้ำหนัก) กับ Height (ความสูง) มีค่าเท่ากับ 0.76
10. ความสัมพันธ์ของ OPDdeath (เสียชีวิตหลังกลับบ้าน) กับ totalDeath (ผู้ป่วยที่เสียชีวิตทั้งหมด) มีค่าเท่ากับ 0.73
11. ความสัมพันธ์ของ Lymphoma (มะเร็งต่อมน้ำเหลือง) กับ Hemato (มะเร็งโลหิตวิทยา) มีค่าเท่ากับ 0.70
12. ความสัมพันธ์ของ OPDCOD (สาเหตุการตาย) กับ totalDeath (ผู้ป่วยที่เสียชีวิตทั้งหมด) มีค่าเท่ากับ 0.65
13. ความสัมพันธ์ของ CAType (ชนิดมะเร็ง) กับ Stage (ระยะของมะเร็ง) มีค่าเท่ากับ 0.64
14. ความสัมพันธ์ของ Death (ผู้ป่วยเสียชีวิตในโรงพยาบาล) กับ totalDeath (ผู้ป่วยที่เสียชีวิตทั้งหมด) มีค่าเท่ากับ 0.63
15. ความสัมพันธ์ของ Hemato (มะเร็งโลหิตวิทยา) กับ Cancer (โรคมะเร็ง) มีค่าเท่ากับ 0.61
16. ความสัมพันธ์ของ CauseofDeath (สาเหตุของการเสียชีวิต) กับ totalDeath (ผู้ป่วยที่เสียชีวิตทั้งหมด) มีค่าเท่ากับ 0.61
17. ความสัมพันธ์ของ Lymphoma (มะเร็งต่อมน้ำเหลือง) กับ CAType (ชนิดมะเร็ง) มีค่าเท่ากับ 0.54
18. ความสัมพันธ์ของ Hemato (มะเร็งโลหิตวิทยา) กับ Stage (ระยะของมะเร็ง) มีค่าเท่ากับ 0.54
19. ความสัมพันธ์ของ DeadDate (วันที่เสียชีวิต) กับ totalDeath (ผู้ป่วยที่เสียชีวิตทั้งหมด) มีค่าเท่ากับ 0.53

ผู้วิจัยได้ทำการสำรวจความกระจายของข้อมูลเฉพาะผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตันในหลอดเลือดดำ ซึ่งมี 27 ราย ของข้อมูลทั้งหมด โดยแสดงผลข้อมูลด้วยวิธี t-distributed Stochastic Neighbor Embedding : t-SNE ซึ่งเป็นวิธีการลดมิติของข้อมูลให้สามารถวิเคราะห์ออกมาง่ายขึ้น ดังภาพประกอบ 25



ภาพประกอบ 25 แสดงข้อมูลของผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตันในหลอดเลือดดำ 27 ราย

จากนั้นดูความสัมพันธ์ของข้อมูลคนที่เป็โรคลิ้มเลือดอุดตันกับและไม่เป็นโรคลิ้มเลือดอุดตันทั้งหมด พบว่ามีความสัมพันธ์ของปัจจัยเสี่ยงที่หลากหลายมา ซึ่งคนที่จะเป็นโรคนี้อาจเกิดจากหลายปัจจัยไม่ได้มีปัจจัยที่เฉพาะเจาะจง เช่น จากการพลอตกราฟความสัมพันธ์ Heatmap พบว่า Cancer (โรคมะเร็ง) กับ Stage (ระยะของมะเร็ง) มีความสัมพันธ์กันเอง แต่ไม่ได้มีความสัมพันธ์กับ Class VTE positive เลย จึงทำให้ความสัมพันธ์นี้ไม่ได้เป็นปัจจัยที่ทำให้เกิดโรคลิ้มเลือดอุดตันได้ แสดงผลข้อมูลด้วยวิธี t-distributed Stochastic Neighbor Embedding : t-SNE ซึ่งมีความกระจายตัวกันมาก ดังภาพประกอบ 26



ภาพประกอบ 26 แสดงความกระจายตัวของผู้ป่วยที่เป็นโรคลิ่มเลือดอุดตัน (เขียว) และไม่เป็นโรคลิ่มเลือดอุดตัน (ฟ้า)

3.5 การเตรียมข้อมูล (Data Preparation)

ในการเตรียมข้อมูลสำหรับการฝึกฝนแบบจำลองของข้อมูลที่ได้จากปัจจัยเสี่ยงของการเกิดภาวะลิ่มเลือดอุดตันในหลอดเลือดดำ โดยได้ข้อมูลจากแพทย์ผู้เชี่ยวชาญที่รวบรวมข้อมูลจากผู้ป่วยที่เข้ารับการรักษาในหอผู้ป่วยแผนกอายุรศาสตร์ ของโรงพยาบาลจุฬาลงกรณ์ เมื่อปี 2009 ได้จัดการกับข้อมูลเพื่อนำไปสร้างแบบจำลองดังต่อไปนี้

3.5.1 จัดการค่าว่าง (Missing Values)

จัดการกับข้อมูลที่ขาดหายไปหรือค่าว่าง (Missing Values) พบข้อมูลที่ขาดหายไปมีจำนวนมาก ได้จัดการกับข้อที่ขาดหายไปด้วยหลักการเติมศูนย์ (Fill missing values with zero) เพราะข้อมูลคนไข้แต่ละคนไม่สามารถแทนที่กันได้จึงใช้หลักการเติมค่าว่างด้วยตัวเลขศูนย์

3.5.2 การแปลงข้อมูล

เนื่องจากข้อมูลที่ทำกรสำรวจออกมานั้นมีบาง Feature ที่ชนิดของข้อมูลเป็นตัวอักษร จะนำไปวิเคราะห์ได้ข้อมูลต้องเป็นตัวเลข ผู้วิจัยจึงได้จัดการกับชนิดข้อมูลดังกล่าวด้วยวิธีการแปลงข้อมูลแบบ Label Encoder ซึ่งได้ผลลัพธ์ดังภาพประกอบ 27

FinalDx	Comorbid	Procedure	Death	CauseofDeath	Autopsy	DischargeDate
SLE	Myelopathy	Insertion of endotracheal tube, Mechanical ven...	0	0	0	4/8/2009
Guillain-Barre syndrome	Delirium, not superimposed on dementia, so des...	Insertion of endotracheal tube, Mechanical ven...	1	1	1	1/6/2009
Other sstreptococcal septicaemia	Nephrotic syndrome, focal and segmental glomer...	Computerized axial tomography of abdomen	0	0	0	5/28/2009
Cellulitis	DM, Bullous pemphigoid	Computerized axial tomography of abdomen and h...	0	0	0	7/14/2009
Secondary malignant neoplasm of bone and bone ...	Radiotherapy session, Malignant neoplasm, Obst...	Computerized axial tomography of abdomen and h...	0	0	0	11/22/2008

➔

FinalDx	Comorbid	Procedure	Death	CauseofDeath	Autopsy	DischargeDate
528	668	383	0	0	0	255
239	316	383	1	1	1	28
468	738	132	0	0	0	276
140	304	146	0	0	0	322
536	861	146	0	0	0	105

ภาพประกอบ 27 แสดงผลลัพธ์การแปลงข้อมูลให้เป็นตัวเลข ด้วย Label encoder

3.5.3 แบ่งข้อมูลสำหรับการฝึกฝน และสำหรับทดสอบ

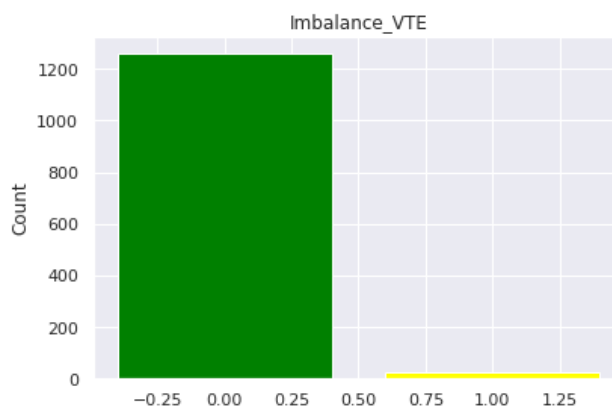
เมื่อทำการจัดการกับข้อมูลให้พร้อมที่จะนำไปสร้างแบบจำลอง ผู้วิจัยได้ทำการแบ่งข้อมูลสำหรับการฝึกฝน (Training data) และข้อมูลสำหรับทดสอบ (Testing data) ด้วยอัตราส่วน 70:30 จากข้อมูลทั้งหมด 1,290 แถว จะได้อัตราส่วนเป็น 903:387 เพื่อนำไปสร้างแบบจำลองในลำดับต่อไป

3.5.4 จัดการกับข้อมูลที่ไม่สมดุลกัน (Imbalance Data)

เนื่องจากข้อมูลของวิจยนี้เป็นปัญหาแบบ Binary Classification เพื่อทำการจำแนกโรคล้มเลือดอุดตัน ซึ่งแบ่งออกเป็น 2 ประเภท คือ 0 และ 1 โดยที่ได้จัดกลุ่มเป็นดังนี้

0 คือ ผู้ป่วยที่ไม่เป็นโรคล้มเลือดอุดตัน

1 คือ ผู้ป่วยที่เป็นโรคล้มเลือดอุดตัน

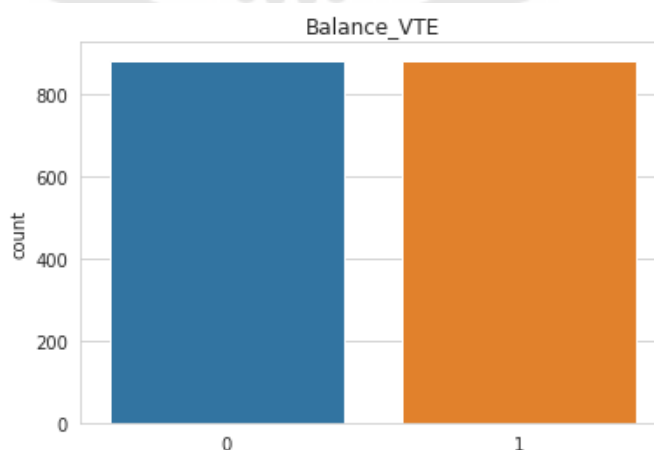


ภาพประกอบ 28 แสดงข้อมูลทั้งสองคลาสที่ไม่สมดุลกัน (Imbalance)

จากการสำรวจข้อมูลแสดงดังภาพประกอบ 28 พบว่าผู้ป่วยที่ไม่เป็นโรคลิ้มเลือดอุดตัน หรือ Class 0 มีจำนวน 1,263 ราย หรือคิดเป็น 97.91 เปอร์เซ็นต์ และผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตัน Class 1 มีจำนวน 27 ราย หรือคิดเป็น 2.09 เปอร์เซ็นต์ ซึ่งเป็นผลการสำรวจที่เป็นปัญหาของข้อมูลที่ไม่สมดุลกัน (Imbalance Data) คือผู้ป่วยที่ไม่เป็นโรคลิ้มเลือดอุดตันมีเยอะกว่าผู้ป่วยที่เป็นโรคลิ้มเลือดอุดตัน จะเกิดการ Sensitive กับข้อมูลคลาสที่เยอะกว่า ทำให้เกิด Bias ขึ้นกับข้อมูลเพราะอัลกอริทึมจะไม่สามารถตรวจพบข้อมูลในคลาสที่น้อยกว่า

ดังนั้นผู้วิจัยจึงจัดการกับปัญหากับข้อมูลที่ไม่สมดุลโดยมี 5 วิธี ดังต่อไปนี้

1.วิธีสุ่มเกิน หรือ Oversampling เป็นวิธีเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนเท่ากับจำนวนที่อยู่ในกลุ่มส่วนมาก



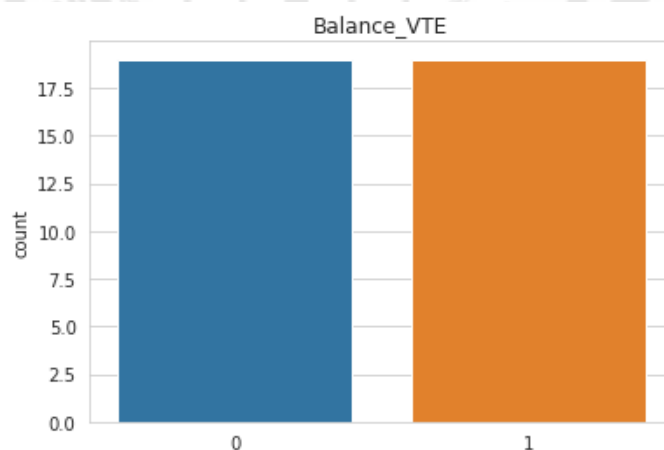
ภาพประกอบ 29 แสดงข้อมูลที่สมดุลกันด้วยวิธี Oversampling

จากภาพประกอบ 29 ได้ทำการสมดุลข้อมูลด้วยวิธี Oversampling จากข้อมูลการฝึกฝนทั้งหมด 903 แสดงจำนวนก่อนและหลังการสมดุลข้อมูลดังตาราง 3

ตาราง 3 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ Oversampling

Oversampling method	VTE negative (0)	VTE positive (1)
ก่อนทำ Oversampling	884	19
หลังทำ Oversampling	884	884

2.วิธีสุ่มลด หรือ Undersampling เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนเท่ากับจำนวนที่อยู่ในกลุ่มส่วนน้อย



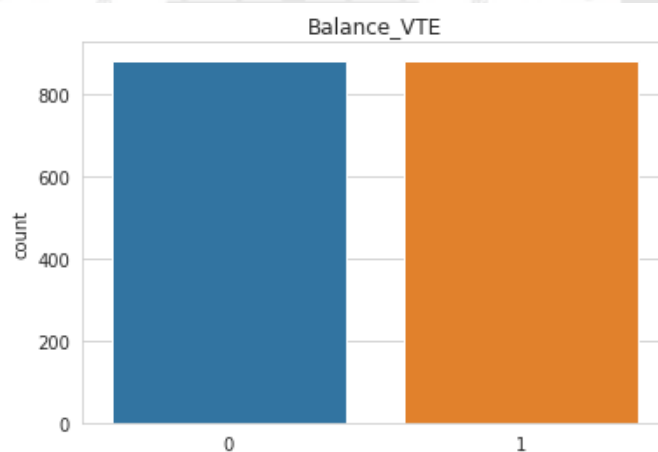
ภาพประกอบ 30 แสดงข้อมูลที่สมดุลกันด้วยวิธี Undersampling

จากภาพประกอบ 30 ได้ทำการสมดุลข้อมูลด้วยวิธี Undersampling จากข้อมูลการฝึกฝนทั้งหมด 903 แสดงจำนวนก่อนและหลังการสมดุลข้อมูลดังตาราง 4

ตาราง 4 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ Undersampling

Undersampling method	VTE negative (0)	VTE positive (1)
ก่อนทำ Undersampling	884	19
หลังทำ Undersampling	19	19

3.วิธีสังเคราะห์ข้อมูลเพิ่ม หรือ Synthetic Minority Oversampling TEchnique : SMOTE เป็นเทคนิคการสุ่มตัวอย่างแบบสุ่มเพิ่มแทนที่จะสุ่มเพิ่มจากข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิม หลักการคือการสุ่มจากข้อมูลเพื่อนบ้านที่อยู่ใกล้ที่สุด



ภาพประกอบ 31 แสดงข้อมูลที่สมดุลกันด้วยวิธี SMOTE

จากภาพประกอบ 31 ได้ทำการสมดุลข้อมูลด้วยวิธี SMOTE จากข้อมูลการฝึกฝนทั้งหมด 903 แสดงจำนวนก่อนและหลังการสมดุลข้อมูลดังตาราง 5

ตาราง 5 แสดงจำนวนการสมดุลข้อมูลก่อนและหลังทำ SMOTE

SMOTE method	VTE negative (0)	VTE positive (1)
ก่อนทำ SMOTE	884	19
หลังทำ SMOTE	884	884

4.วิธี Class weight คือ การให้น้ำหนักในแต่ละ Class ซึ่งจะให้ความสำคัญกับข้อมูลกลุ่มส่วนน้อยมากกว่า เพื่อให้ผลลัพธ์ที่ได้นั้นมาจากการเรียนรู้จากทุกกลุ่มที่เท่าเทียมกัน

5.วิธี Ensemble sampling โดยใช้ Balanced Bagging Classifier คือ ความเป็นของการสุ่มตัวอย่างของ Training data ที่จะถูกทำให้กระจายไปตามกลุ่มที่อยู่ใกล้เคียงกันทำให้ข้อมูลสมดุลขึ้น

3.6 หลักการที่ใช้ในการสร้างแบบจำลอง (Modeling)

หลังจากที่เตรียมข้อมูลที่เหมาะสมและพร้อมที่จะนำไปสร้างแบบจำลอง ทำการปรับจูนหาพารามิเตอร์ที่เหมาะสมและดีที่สุดของแต่ละอัลกอริทึมแล้วนำข้อมูลเข้าแบบจำลอง โดยแต่ละอัลกอริทึมมีหลักการดังต่อไปนี้

3.6.1 อัลกอริทึม Decision Tree

อัลกอริทึม Decision Tree เป็นเทคนิคต้นไม้ตัดสินใจแบบต้นไม้เดียวโดดๆ ที่เรียนรู้จากคุณลักษณะของข้อมูลแล้วสร้างผังการตัดสินใจคล้ายๆ กับต้นไม้ (สระอุบล, 2020) ซึ่งจะพยายามแบ่งข้อมูลออกมาเป็นคลาสให้ชัดเจนได้มากที่สุด

3.6.2 อัลกอริทึม Logistic Regression

อัลกอริทึม Logistic Regression เป็น Extended Version ของ Linear Regression นิยมใช้กับปัญหา Binary Classification ทำนายคลาสที่มี 2 คลาส เช่น 0 กับ 1 เป็นต้น เนื่องจาก Linear Regression ไม่เหมาะสมกับปัญหา Binary Classification เพราะผลของสมการ Linear

Regression เป็นแบบเส้นตรงมีโอกาสที่จะต่ำกว่า 0 หรือสูงกว่า 1 ได้ จึงไม่เหมาะสม ทำให้มีการแก้ปัญหาดังกล่าว โดยมี Sigmoid function เพื่อมา Normalize ให้ได้ค่าอยู่ระหว่าง 0-1 ซึ่งจะมี Exponential unction (e) ในการเปลี่ยนสมการ linear เป็น Non-Linear ดังสมการ (3-1) (SATANGMONGKOL, 2019)

$$\text{Sigmoid (Z)} = \frac{e^Z}{1+e^Z} \quad (3-1)$$

ผลลัพธ์ที่ได้จาก Sigmoid Function คือความน่าจะเป็นที่สามารถทำนายคลาสได้ เช่น 1= positive และ 0 = negative เป็นต้น

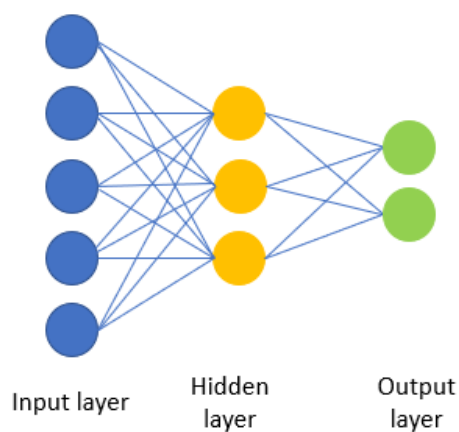
3.6.3 เทคนิค Neural Network

เทคนิค Neural Network หรือเรียกว่าโครงข่ายประสาทเทียมเป็นเทคนิคที่จำลองการทำงานของสมองมนุษย์ ซึ่งประกอบด้วยเซลล์ประสาทที่เชื่อมโยงกันและส่งกระแสไฟฟ้าเพื่อให้เซลล์ประสาทถัดไปทำงาน แสดงดังภาพประกอบ 32 ในโครงข่ายประสาทเทียมก็จะแบ่งออกเป็น 3 layers ดังนี้

1. Input Layer เป็น Layer แรกของโมเดลที่มีจำนวน Input เท่ากับตัวแปรที่ใช้ใน Training data ซึ่งตัวแปรนั้นต้องเป็นตัวเลขเท่านั้น

2. Hidden Layer เป็น Layer ตรงกลางของโมเดลที่อาจจะมีได้มากกว่า 1 Hidden Layer ก็ได้ จะมีเส้นเชื่อมโยงจาก Input Layer มาที่ Hidden Layer เชื่อมกันทุกโหนด เส้นที่เชื่อมกันนั้นจะมีค่าน้ำหนักที่ถูกปรับจนได้ค่าที่เหมาะสมจากการ Train Model ซึ่งจำนวน Hidden Layer ที่เหมาะสมโดยทั่วไปใช้แค่ 1 Layer เพื่อไม่ให้โมเดลมีความซับซ้อนมากเกินไปจนเกิดปัญหา Overfitting จนไม่สามารถนำไปใช้กับข้อมูลใหม่ได้ (พัชรวงศ์ศักดิ์, 2020)

3. Output Layer เป็น Layer ที่แสดงผลลัพธ์ของโมเดล ซึ่งจะมี Output Node เท่ากับค่าคำตอบ เช่น แสดงผลเป็น Yes และ No



ภาพประกอบ 32 แสดงโมเดล Neural Network ที่มี 3 layers

3.7 การประเมินผล (Evaluation)

งานวิจัยนี้เป็นการทำนายผลลัพธ์ที่เป็น Classification ในทางการแพทย์ จึงจำเป็นต้องมีการประเมินผลลัพธ์ให้แม่นยำที่สุด โดยใช้ตาราง Confusion Matrix ในประเมินผลของแบบจำลองดังต่อไปนี้

		ทำนาย (Prediction)	
		Negative (0)	Positive (1)
ค่าจริง (Actual)	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

ภาพประกอบ 33 แสดงตาราง Confusion matrix

จากภาพประกอบ 33 อธิบายรายละเอียดได้ดังต่อไปนี้

True Negative (TN) คือ ทำนายถูกต้อง สำหรับผลในทางลบ

เช่น ทางกายภาพตรวจแล้วพบว่าไม่เป็นโรคลิ่มเลือดอุดตัน (VTE negative) ทาง Machine Learning ถือว่าให้ผลทำนายถูกในทางลบ หรือไม่เป็นโรคลิ่มเลือดอุดตันกับอกได้ ถูกว่าไม่เป็นโรคลิ่มเลือดอุดตัน

True Positive (TP) คือ ทำนายถูกต้อง สำหรับผลในทางบวก

เช่น ทางกายภาพตรวจแล้วพบว่า เป็นโรคลิ่มเลือดอุดตัน (VTE positive) ทาง Machine Learning ถือว่าให้ผลทำนายถูกในทางบวก หรือเป็นโรคลิ่มเลือดอุดตันกับอกได้ ถูกว่า เป็นโรคลิ่มเลือดอุดตัน

False Negative (FN) คือ ทำนายผิด โดยให้ผลเป็นลบ

เช่น ทำนายแล้วพบว่าไม่เป็นโรคลิ่มเลือดอุดตัน (VTE negative) แต่จริงๆ แล้ว เป็นโรคลิ่มเลือดอุดตัน (VTE positive) สรุปว่าให้ผลทำนายผิด

False Positive (FP) คือ ทำนายผิด โดยให้ผลเป็นบวก

เช่น ทำนายแล้วพบว่า เป็นโรคลิ่มเลือดอุดตัน (VTE positive) แต่จริงๆ แล้ว ไม่ได้ เป็นโรคลิ่มเลือดอุดตัน (VTE negative) สรุปว่าให้ผลทำนายผิด

ยกตัวอย่างผลการทำนายการเกิดโรคลิ่มเลือดอุดตัน โดยใช้อัลกอริทึม Decision Tree ด้วยการสมดุลข้อมูลวิธี Class weight ดังภาพประกอบ 34

	precision	recall	f1-score	support
0	0.98	0.98	0.98	379
1	0.22	0.25	0.24	8
accuracy			0.97	387
macro avg	0.60	0.62	0.61	387
weighted avg	0.97	0.97	0.97	387
0.9664082687338501				
[[372 7]				
[6 2]]				
DecisionTree_Class weight				

ภาพประกอบ 34 ผลการทำนายการเกิดโรคลิ่มเลือดอุดตัน โดยใช้อัลกอริทึม Decision Tree ด้วยการสมดุลข้อมูลวิธี Class weight

จากภาพประกอบ 34 อัลกอริทึม Decision Tree สามารถทำนายคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน (0) ทั้งหมด 379 คน และทำนายคนที่เป็นโรคลิ้มเลือดอุดตัน (1) ทั้งหมด 8 คน อธิบายการอ่านค่าจาก Confusion Matrix ดังนี้

True Negative (TN) คือ ทำนายถูกว่าไม่เป็นโรคลิ้มเลือดอุดตันทั้งหมด 372 คน

False Positive (FP) คือ ทำนายว่าเป็นโรคลิ้มเลือดอุดตันทั้งหมด 7 คน แต่จริงๆ แล้วไม่ได้เป็นโรคลิ้มเลือดอุดตัน

True Positive (TP) คือ ทำนายถูกว่าเป็นโรคลิ้มเลือดอุดตันทั้งหมด 2 คน

False Negative (FN) คือ ทำนายว่าไม่ได้เป็นโรคลิ้มเลือดอุดตันทั้งหมด 6 คน แต่จริงๆ แล้วเป็นโรคลิ้มเลือดอุดตัน

จากผลลัพธ์ของ Classification report ในภาพประกอบ 34 ผู้วิจัยได้ประเมินผลประสิทธิภาพของแบบจำลองโดยจะเน้นดูค่าที่แบบจำลองทำนายว่าเป็นโรคลิ้มเลือดอุดตัน ซึ่งอธิบายค่าต่างๆ ได้ดังต่อไปนี้

3.7.1 Accuracy

Accuracy คือเปอร์เซ็นต์ความถูกต้อง คือจำนวนที่ทำนายถูกหารด้วยจำนวนทั้งหมด ดังสมการ (3-2)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3-2)$$

จากภาพประกอบ 34 สามารถคำนวณเปอร์เซ็นต์ความถูกต้องได้ดังนี้

$$\text{Accuracy} = \frac{2 + 372}{2 + 372 + 7 + 6} = \frac{374}{387} = 0.97$$

3.7.2 Recall

Recall คือความแม่นยำที่สนใจผลลัพธ์เทียบกับที่เป็นของจริง (Actual) ดังสมการ (3-3)

$$\text{Recall(Positive)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3-3)$$

จากภาพประกอบ 34 สามารถคำนวณค่า Recall Positive ได้ดังนี้

$$\text{Recall(VTEpositive(1))} = \frac{2}{2 + 6} = 0.25$$

Recall Positive หรือ True Positive Rate หรือ Sensitivity จากผลการคำนวณ มีคนที่เป็นโรคลิ่มเลือดอุดตันของจริงทั้งหมด 8 คน ทำนายถูก 2 คน แบบจำลองมีความแม่นยำในการทำนาย 0.25 หรือ 25 เปอร์เซ็นต์

ในการดูผล Sensitivity มีประโยชน์ในการป้องกันผลลบปลอม (False Negative) เพราะว่าการทดสอบยิ่งไวมากเท่าไรโอกาสการได้ผลลบ เช่น การพบว่าไม่เป็นโรคลิ่มเลือดอุดตัน แต่จริงๆ เป็นโรคลิ่มเลือดอุดตันก็จะน้อยลง ถ้าความไวอยู่ที่ 100 เปอร์เซ็นต์ โอกาสได้ผลลบปลอมก็จะอยู่ที่ 0 เปอร์เซ็นต์ จากตัวอย่างการคำนวณ Recall Positive หรือ Sensitivity ของภาพประกอบ 34 ให้ค่า Recall Positive อยู่ที่ 25 เปอร์เซ็นต์ หมายความว่าโอกาสได้ผลลบปลอมนั้นจะมีค่ามากถึง 75 เปอร์เซ็นต์ คือทำนายหาคนเป็นโรคลิ่มเลือดอุดตันไม่ถูกต้องเป็นจำนวนมาก

$$\text{Recall(Negative)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3-4)$$

จากภาพประกอบ 34 สามารถคำนวณค่า Recall Negative ได้ดังนี้

$$\text{Recall(VTEnegative(0))} = \frac{372}{372 + 7} = 0.98$$

Recall Negative หรือ True Negative Rate หรือ Specificity (ความจำเพาะ) จากผลการคำนวณ มีคนที่เป็นไม่โรคลิ่มเลือดอุดตันของจริงทั้งหมด 379 คน ทำนายถูก 372 คน แบบจำลองมีความแม่นยำในการทำนาย 0.98 หรือ 98 เปอร์เซ็นต์

ในการดูผล Specificity มีประโยชน์ในการป้องกันผลบวกปลอม (False Positive) เพราะว่าการทดสอบยิ่งจำเพาะมากเท่าไรโอกาสการได้ผลบวก เช่น การพบว่า เป็นโรคลิ่มเลือดอุดตัน แต่จริงๆ แล้ว ไม่ได้เป็นโรคลิ่มเลือดอุดตัน ก็น้อยลงเท่านั้น ถ้าความจำเพาะอยู่ที่ 100 เปอร์เซ็นต์ โอกาสได้ผลบวกปลอมก็จะอยู่ที่ 0 เปอร์เซ็นต์ จากตัวอย่างการคำนวณ Recall Negative หรือ Specificity ของภาพประกอบ 34 ให้ค่า Recall Negative อยู่ที่ 98 เปอร์เซ็นต์ หมายความว่าโอกาส

ได้ผลบวกปลอมนั้นจะมีค่า 2 เปอร์เซ็น คือทำนายหาคนไม่เป็นโรคล้มเลือดอุดตันถูกต้องเป็นจำนวนมาก

3.7.3 Precision

Precision คือความแม่นยำผลทำนาย แสดงดังสมการ (3-5) และ (3-6)

$$\text{Precision(Positive)} = \frac{TP}{TP+FP} \quad (3-5)$$

จากภาพประกอบ 35 สามารถคำนวณค่า Precision Positive ได้ดังนี้

$$\text{Precision(Positive(1))} = \frac{2}{2+7} = 0.22$$

Precision Positive จากผลการคำนวณระบบทำนายว่ามีคนเป็นโรคล้มเลือดอุดตันทั้งหมด 9 คน ทำนายถูก 2 คน ทำนายผิด 7 คน แบบจำลองมีความแม่นยำในการทำนาย 0.22 หรือ 22 เปอร์เซ็น

$$\text{Precision(Negative)} = \frac{TN}{TN+FN} \quad (3-6)$$

จากภาพประกอบ 34 สามารถคำนวณค่า Precision Negative ได้ดังนี้

$$\text{Precision(Negative(0))} = \frac{372}{372+6} = 0.98$$

Precision Negative จากผลการคำนวณระบบทำนายว่ามีคนไม่เป็นโรคล้มเลือดอุดตันทั้งหมด 378 คน ทำนายถูก 372 คน ทำนายผิด 6 คน แบบจำลองมีความแม่นยำในการทำนาย 0.98 หรือ 98 เปอร์เซ็น

3.7.4 F1-score

F1-score เป็นค่าที่แสดงประสิทธิภาพ โดยการนำ Precision และ Recall มาคำนวณหาค่าเฉลี่ย ซึ่งค่าสูงๆ ถือว่าแบบจำลองมีประสิทธิภาพดี คำนวณดังสมการ (3-7)

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3-7)$$

จากภาพประกอบ 34 สามารถคำนวณค่า F1-score ได้ดังนี้

$$F1 - score(Positive(1)) = \frac{2 * 0.22 * 0.25}{0.22 + 0.25} = 0.24$$

$$F1 - score(Negative(0)) = \frac{2 * 0.98 * 0.98}{0.98 + 0.98} = 0.98$$

จากการอ่านตาราง Confusion Matrix ดังภาพประกอบ 34 ให้ข้อสรุปว่าถ้า True Negative (TN) และ True Positive (TP) ยิ่งค่าสูงยิ่งดี เพราะทำนายถูกต้อง ส่วน False Negative (FN) และ False Positive (FP) คือความผิดพลาดของการทำนาย ค่าต่ำ ๆ จึงจะถือว่าดี ดังนั้นจากผลการคำนวณดังกล่าวให้ผลได้ไม่ตื้นนัก การวัดผลในทางการแพทย์ค่า Recall Positive หรือ Sensitivity (ความไว) หรือ True Positive ในทางปฏิบัติควรเลือกใช้การตรวจที่มีค่า Sensitivity สูง ๆ ในการตรวจคัดกรองผู้ป่วยเพื่อทำการรักษาในลำดับถัดไป

บทที่ 4

ผลการดำเนินงานวิจัย

จากการวิจัยเพื่อสร้างแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษากระบวนการต่างๆ สืบหาข้อมูล เตรียมข้อมูลเพื่อให้พร้อมสำหรับสร้างแบบจำลอง แบ่งข้อมูลสำหรับการฝึกฝน (Training data) และข้อมูลสำหรับทดสอบ (Testing data) ด้วยอัตราส่วน 70:30 จากข้อมูลทั้งหมด 1,290 แถว จะได้อัตราส่วนของข้อมูลที่แบ่งเป็น 903:387 จากนั้นสร้างข้อมูลให้สมดุลของ Class Label ทั้งหมด 5 วิธี ได้แก่ Oversampling, Undersampling, SMOTE, Class weight และ Ensemble sampling โดยใช้ Balanced Bagging Classifier ตลอดจนการวัดประสิทธิภาพเพื่อเปรียบเทียบแบบจำลอง 3 อัลกอริทึม ได้แก่ Decision Tree, Logistic Regression และ Neural Network และประเมินประสิทธิภาพของแบบจำลองโดยใช้ตาราง Confusion matrix เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยที่ได้กำหนดไว้ ผลลัพธ์ที่ได้อธิบายดังต่อไปนี้

4.1 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Oversampling

4.2 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Undersampling

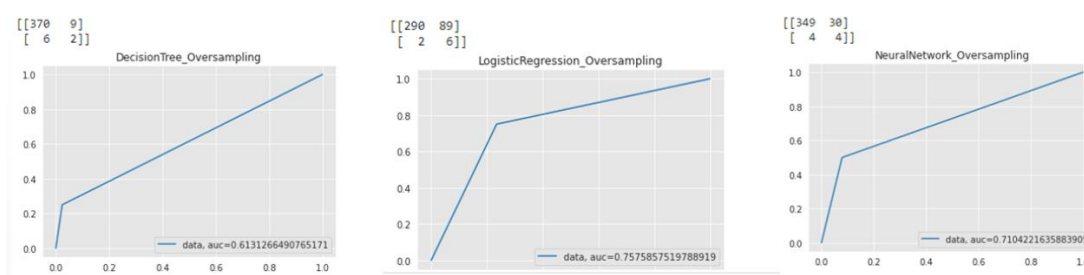
4.3 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE

4.4 ผลลัพธ์ของการสร้างแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class weight

4.5 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier

4.1 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลงข้อมูลแบบ

Oversampling



ภาพประกอบ 35 แสดงผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลงข้อมูลแบบ

Oversampling

จากภาพประกอบ 35 ผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลงข้อมูลแบบ Oversampling แสดงดังตาราง 6

ตาราง 6 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลงข้อมูลแบบ Oversampling

Oversampling	Decision Tree	Logistic Regression	Neural Network
Accuracy	0.96	0.76	0.91
Recall (VTEpositive(1))	0.25	0.75	0.50
Recall (VTEnegative(0))	0.98	0.77	0.92
Precision (VTEpositive(1))	0.18	0.06	0.12

ตาราง 6 (ต่อ)

Oversampling	Decision Tree	Logistic Regression	Neural Network
Precision (VTEnegative(0))	0.98	0.99	0.99
F1-score(VTEpositive(1))	0.21	0.12	0.95
F1-score(VTEnegative(0))	0.98	0.86	0.19

จากภาพประกอบ 35 เปรียบเทียบประสิทธิภาพของ 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Oversampling สังเกตได้ว่า ทั้ง 3 อัลกอริทึมทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ได้น้อยมาก จากข้อมูลสำหรับการทดสอบทั้งหมด 387 คน

4.1.1 แบบจำลอง Decision Tree

4.1.1.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 370 คน และทำนายผิด 9 คน

4.1.1.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 2 คน และทำนายผิด 6 คน

4.1.2 แบบจำลอง Logistic Regression

4.1.2.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 290 คน และทำนายผิด 89 คน

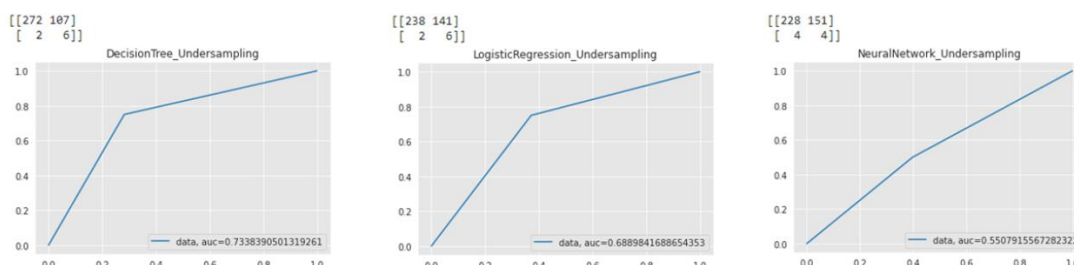
4.1.2.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 2 คน และทำนายผิด 6 คน

4.1.3 แบบจำลอง Neural Network

4.1.3.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 349 คน และทำนายผิด 30 คน

4.1.3.2 สามารถทำนายคนที่เป็นโรคล้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน
ทำนายถูกทั้งหมด 4 คน และทำนายผิด 4 คน

4.2 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Undersampling



ภาพประกอบ 36 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ
Undersampling

จากภาพประกอบ 36 ผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย
วิธีการสมดุลข้อมูลแบบ Undersampling แสดงดังตาราง 7

ตาราง 7 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วย
วิธีการสมดุลข้อมูลแบบ Undersampling

Undersampling	Decision Tree	Logistic Regression	Neural Network
Accuracy	0.72	0.63	0.60
Recall (VTEpositive(1))	0.75	0.75	0.50
Recall (VTEnegative(0))	0.72	0.63	0.60

ตาราง 7 (ต่อ)

Undersampling	Decision Tree	Logistic Regression	Neural Network
Precision (VTEpositive(1))	0.05	0.04	0.03
Precision (VTEnegative(0))	0.99	0.99	0.98
F1-score(VTEpositive(1))	0.10	0.08	0.05
F1-score(VTEnegative(0))	0.83	0.77	0.75

จากภาพประกอบ 36 เปรียบเทียบประสิทธิภาพของ 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Undersampling สังเกตได้ว่า ทั้ง 3 อัลกอริทึมทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ได้น้อยมาก จากข้อมูลสำหรับการทดสอบทั้งหมด 387 คน

4.2.1 แบบจำลอง Decision Tree

4.2.1.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 272 คน และทำนายผิด 107 คน

4.2.1.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 2 คน และทำนายผิด 6 คน

4.2.2 แบบจำลอง Logistic Regression

4.2.2.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 238 คน และทำนายผิด 141 คน

4.2.2.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 2 คน และทำนายผิด 6 คน

4.2.3 แบบจำลอง Neural Network

4.2.3.1 สามารถทำนายคนที่ไม่เป็นโรคล้มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 228 คน และทำนายผิด 151 คน

4.2.3.2 สามารถทำนายคนที่เป็นโรคล้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 4 คน และทำนายผิด 4 คน

4.3 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE



ภาพประกอบ 37 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE

จากภาพประกอบ 37 ผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE แสดงดังตาราง 8

ตาราง 8 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE

SMOTE	Decision Tree	Logistic Regression	Neural Network
Accuracy	0.96	0.80	0.87
Recall (VTEpositive(1))	0.12	0.62	0.50
Recall (VTEnegative(0))	0.97	0.81	0.88

ตาราง 8 (ต่อ)

SMOTE	Decision Tree	Logistic Regression	Neural Network
Precision (VTEpositive(1))	0.09	0.06	0.08
Precision (VTEnegative(0))	0.98	0.99	0.99
F1-score(VTEpositive(1))	0.11	0.12	0.14
F1-score(VTEnegative(0))	0.98	0.89	0.93

จากภาพประกอบ 37 เปรียบเทียบประสิทธิภาพของ 3 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ SMOTE สังเกตได้ว่า ทั้ง 3 อัลกอริทึมทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ได้น้อยมากจากข้อมูลสำหรับการทดสอบทั้งหมด 387 คน

4.3.1 แบบจำลอง Decision Tree

4.3.1.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 369 คน และทำนายผิด 10 คน

4.3.1.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 1 คน และทำนายผิด 7 คน

4.3.2 แบบจำลอง Logistic Regression

4.3.2.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 306 คน และทำนายผิด 73 คน

4.3.2.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 5 คน และทำนายผิด 3 คน

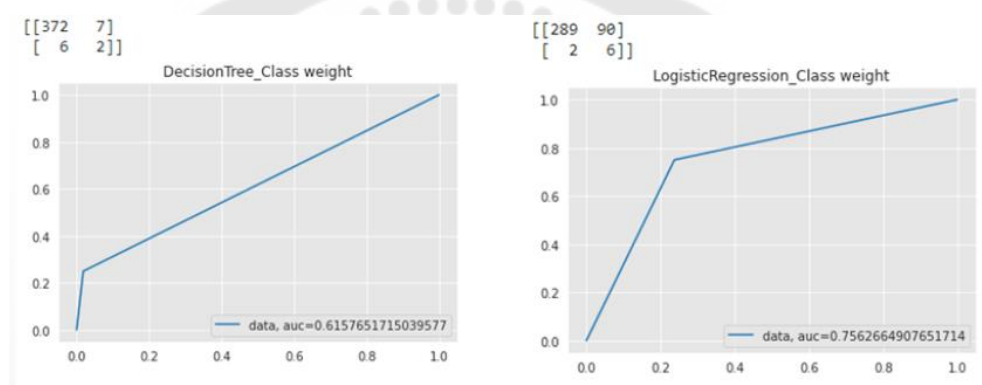
4.3.3 แบบจำลอง Neural Network

4.3.3.1 สามารถทำนายคนที่ไม่เป็นโรคล้มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 334 คน และทำนายผิด 45 คน

4.3.3.2 สามารถทำนายคนที่เป็นโรคล้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 4 คน และทำนายผิด 4 คน

4.4 ผลลัพธ์ของการสร้างแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class weight

เนื่องจากการทำสมดุลข้อมูลด้วยวิธี Class Weight ไม่สามารถใช้ได้กับ Neural Network ได้ ผู้วิจัยจึงได้เปรียบเทียบ 2 อัลกอริทึม คือ Decision Tree กับ Logistic Regression



ภาพประกอบ 38 ผลลัพธ์ของการสร้างแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class Weight

จากภาพประกอบ 38 ผลการประเมินประสิทธิภาพของแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class Weight แสดงดังตาราง 9

ตาราง 9 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class Weight

Class weight	Decision Tree	Logistic Regression
Accuracy	0.97	0.76

ตาราง 9 (ต่อ)

Class weight	Decision Tree	Logistic Regression
Recall (VTEpositive(1))	0.25	0.75
Recall (VTEnegative(0))	0.98	0.76
Precision (VTEpositive(1))	0.22	0.06
Precision(VTEnegative(0))	0.98	0.99
F1-score(VTEpositive(1))	0.24	0.12
F1-score(VTEnegative(0))	0.98	0.86

จากภาพประกอบ 38 เปรียบเทียบประสิทธิภาพของ 2 อัลกอริทึมด้วยวิธีการสมดุลข้อมูลแบบ Class weight สังเกตได้ว่า ทั้ง 2 อัลกอริทึมทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ได้น้อยมาก จากข้อมูลสำหรับการทดสอบทั้งหมด 387 คน

4.4.1 แบบจำลอง Decision Tree

4.4.1.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 372 คน และทำนายผิด 7 คน

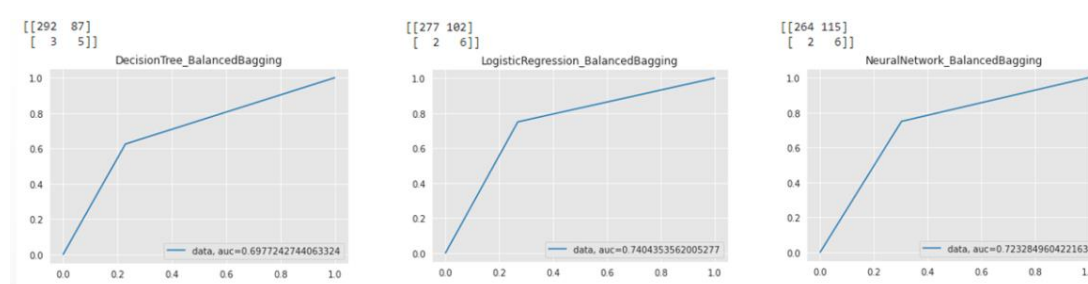
4.4.1.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 2 คน และทำนายผิด 6 คน

4.4.2 แบบจำลอง Logistic Regression

4.4.2.1 สามารถทำนายคนที่ไม่เป็นโรคล้มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 289 คน และทำนายผิด 90 คน

4.4.2.2 สามารถทำนายคนที่เป็นโรคล้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 6 คน และทำนายผิด 2 คน

4.5 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมมูลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier



ภาพประกอบ 39 ผลลัพธ์ของการสร้างแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมมูลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier

จากภาพประกอบ 39 ผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมมูลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier แสดงดังตาราง 10

ตาราง 10 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง 3 อัลกอริทึมด้วยวิธีการสมมูลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier

Balanced Bagging Classifier	Decision Tree	Logistic Regression	Neural Network
Accuracy	0.77	0.73	0.70

ตาราง 10 (ต่อ)

Balanced Bagging Classifier	Decision Tree	Logistic Regression	Neural Network
Recall (VTEpositive(1))	0.62	0.75	0.75
Recall (VTEnegative(0))	0.77	0.73	0.70
Precision (VTEpositive(1))	0.05	0.06	0.05
Precision (VTEnegative(0))	0.99	0.99	0.99
F1-score(VTEpositive(1))	0.10	0.10	0.09
F1-score(VTEnegative(0))	0.87	0.84	0.82

จากภาพประกอบ 39 เปรียบเทียบประสิทธิภาพของ 3 อัลกอริทึมด้วยวิธีการสมมูลข้อมูลแบบ Ensemble sampling โดยใช้ Balanced Bagging Classifier สังเกตได้ว่า ทั้ง 3 อัลกอริทึมทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ได้น้อยมาก จากข้อมูลสำหรับการทดสอบทั้งหมด 387 คน

4.5.1 แบบจำลอง Decision Tree

4.5.1.1 สามารถทำนายคนที่ไม่เป็นโรคลิ่มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 292 คน และทำนายผิด 87 คน

4.5.1.2 สามารถทำนายคนที่เป็นโรคลิ่มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 5 คน และทำนายผิด 3 คน

4.5.2 แบบจำลอง Logistic Regression

4.5.2.1 สามารถทำนายคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 277 คน และทำนายผิด 102 คน

4.5.2.2 สามารถทำนายคนที่เป็นโรคลิ้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 6 คน และทำนายผิด 2 คน

4.5.3 แบบจำลอง Neural Network

4.5.3.1 สามารถทำนายคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน (0) ของจริงทั้งหมด 379 คน ทำนายถูกทั้งหมด 264 คน และทำนายผิด 115 คน

4.5.3.2 สามารถทำนายคนที่เป็นโรคลิ้มเลือดอุดตัน (1) ของจริงทั้งหมด 8 คน ทำนายถูกทั้งหมด 6 คน และทำนายผิด 2 คน

เนื่องจากงานวิจัยนี้เป็นงานวิจัยที่ใช้ข้อมูลทางการแพทย์มาสร้างแบบจำลองการทำนายในความเป็นจริงแล้วทางการแพทย์เมื่อจะพยากรณ์โรคให้กับคนหนึ่งคนที่มีอาการนั้น จะพิจารณาสนใจคนที่แพทย์พยากรณ์แล้วว่ามีอาการผิดปกติทำให้เกิดโรคมามากกว่าคนที่ไม่พบอาการที่จะบ่งบอกว่าเป็นโรคได้ ซึ่งก็คือเป็นคนปกติ ถ้าหากพบอาการผิดปกติจะสำคัญกว่าคนที่ปกติเพราะจะต้องทำการรักษาให้กับคนที่มีอาการผิดปกติในลำดับถัดไป ดังนั้นจากผลลัพธ์ภาพประกอบ 35-39 ผู้วิจัยเลือกสนใจผลลัพธ์ที่แบบจำลองสามารถทำนายคนที่เป็นโรคลิ้มเลือดอุดตันได้ โดยจะพิจารณา ค่า Sensitivity, Precision และ F1-score เป็นหลัก เพราะค่าเหล่านี้จะเป็นที่ยอมรับในทางการแพทย์ในการพยากรณ์โรค

จากภาพประกอบที่ 35-39 แบบจำลองทั้ง 3 อัลกอริทึมสามารถทำนายคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน (0) ได้ดีแต่ประสิทธิภาพในการทำนายคนที่เป็นโรคลิ้มเลือดอุดตันนั้นน้อยมาก อาจเป็นเพราะข้อมูลในการวิเคราะห์มีน้อยเกินไปแม้ได้ทำวิธีการสมดุลข้อมูลไปแล้ว นำผลการทดลองทั้งหมดสรุปได้ดังตารางที่ 11

ตาราง 11 แสดงผลลัพธ์ของแบบจำลองการทำนาย 3 อัลกอริทึมพร้อมเปรียบเทียบวิธีปรับความสมดุลของข้อมูลทั้ง 5 วิธี

Balanced Method	Evaluate Model	Decision Tree	Logistic Regression	Neural Network
Oversampling	Sensitivity	0.25	0.75	0.50
	Precision	0.18	0.06	0.12
	F1-score	0.21	0.12	0.19
Undersampling	Sensitivity	0.75	0.75	0.50
	Precision	0.05	0.04	0.03
	F1-score	0.10	0.08	0.05
SMOTE	Sensitivity	0.12	0.62	0.50
	Precision	0.09	0.06	0.08
	F1-score	0.11	0.12	0.14

ตาราง 11 (ต่อ)

Balanced Method	Evaluate Model	Decision Tree	Logistic Regression	Neural Network
Class Weight	Sensitivity	0.25	0.75	-
	Precision	0.22	0.06	-
	F1-score	0.24	0.12	-
Ensemble sampling	Sensitivity	0.62	0.75	0.75
	Precision	0.05	0.06	0.05
	F1-score	0.10	0.10	0.09

เนื่องจากการวัดผลในทางการแพทย์ค่า Recall หรือ Sensitivity (ความไว) ค่า Precision และค่า F1-score ในทางปฏิบัติทางการแพทย์ได้เลือกใช้การตรวจที่มีค่าดังกล่าวสูงๆ ในการตรวจคัดกรองผู้ป่วยที่เป็นโรคซึมเศร้า เพราะเป็นการทำนายที่ถูกต้อง

จากผลการทดลองตาราง 11 พบว่าแบบจำลอง Decision Tree ด้วยการปรับความสมดุลของข้อมูลด้วยวิธี Class Weight ให้ค่าสูงที่สุด แต่ให้ผลลัพธ์ที่ไม่มีประสิทธิภาพ

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการทำวิจัยเรื่องแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองแต่ละเทคนิคเพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ภาวะลิ้มเลือดอุดตันในหลอดเลือดดำเกิดจากปัจจัยเสี่ยงหลาย ๆ ปัจจัยด้วยกัน เมื่อเกิดภาวะลิ้มเลือดอุดตันในหลอดเลือดดำขึ้นในร่างกายแล้ว อาจจะทำให้ต้องใช้เวลาในการรักษาที่โรงพยาบาล ถึงแม้ว่าโรคหลอดเลือดดำอุดตันเป็นโรคที่พบได้ไม่บ่อย แต่อาการที่เกิดขึ้นจากโรคนี้มีความรุนแรงและอาจนำไปสู่การสูญเสียชีวิตได้อย่างรวดเร็ว ทำให้ควรดูแลร่างกายให้แข็งแรงอยู่เสมอ ดังนั้นภาวะการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำจึงเป็นโรคที่ควรศึกษาและทำความเข้าใจปัจจัยต่าง ๆ ที่จะนำไปสู่การเกิดโรคให้มากขึ้น มากไปกว่านั้นแพทย์และบุคลากรทางการแพทย์ที่เกี่ยวข้องในการรักษา ต้องตระหนักในการวินิจฉัยและการใช้เครื่องมือสำหรับประเมินการเกิดโรคที่เหมาะสม

งานวิจัยนี้ได้สร้างแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ้มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง จากการเก็บรวบรวมข้อมูลโดยแพทย์ผู้เชี่ยวชาญได้เก็บข้อมูลผู้ป่วยในหอผู้ป่วยแผนกอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย ซึ่งมีข้อมูลทั้งหมด 1,290 แถว 65 คอลัมน์ จากนั้นผู้วิจัยได้ทำการเตรียมข้อมูลเพื่อวิเคราะห์ ตรวจสอบค่าที่ขาดหายไปของข้อมูล พร้อมทั้งแปลงข้อมูลที่เป็นตัวอักษรให้เป็นตัวเลขพร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไป จากนั้นแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบด้วยอัตราส่วน 70:30 และเนื่องจากข้อมูลงานวิจัยนี้มีปัญหาข้อมูลที่ไม่สมดุล เพราะมีจำนวนคนที่ไม่เป็นโรคลิ้มเลือดอุดตันทั้งหมด 1,263 คน และมีจำนวนคนที่เป็โรคลิ้มเลือดอุดตันทั้งหมด 27 คน จะเห็นได้ว่ากลุ่มของคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน นั้นมีจำนวนน้อยมากเมื่อเทียบกับกลุ่มคนที่ไม่เป็นโรคลิ้มเลือดอุดตัน ซึ่งจะเป็นกระทบต่อการสร้าง

แบบจำลองการทำนาย ผู้วิจัยจึงจัดการกับปัญหาข้อมูลที่ไม่สมดุลทั้งหมด 5 วิธี ได้แก่ Oversampling, Undersampling, Synthetic Minority Oversampling TEchnique : SMOTE, Class Weight และ Ensemble Sampling จากนั้นทำการจูนไฮเปอร์พารามิเตอร์เพื่อหาพารามิเตอร์ที่ดีที่สุดของแต่ละอัลกอริทึมให้สามารถทำนายได้มีประสิทธิภาพที่สุด โดยจะทำการประเมินประสิทธิภาพของแบบจำลองด้วยตาราง Confusion Matrix ซึ่งแบบจำลองการทำนาย 3 อัลกอริทึมที่เลือกใช้ ได้แก่ Decision Tree, Logistic Regression และ Neural Network

เนื่องจากการวัดผลในทางการแพทย์ค่า Recall หรือ Sensitivity, ค่า Precision และค่า F1-score ในทางปฏิบัติทางการแพทย์ได้เลือกใช้การตรวจที่มีค่าดังกล่าวสูง ๆ ในการตรวจคัดกรองผู้ป่วยที่เป็นโรคคลุ้มเลือดตัน เพราะเป็นการทำนายที่ถูกต้อง จากผลการทดลองพบว่าแบบจำลอง Decision Tree ด้วยการปรับความสมดุลของข้อมูลด้วยวิธี Class Weight ให้ค่า Sensitivity เท่ากับ 0.25 หรือ 25 เปอร์เซ็นต์ ค่า Precision เท่ากับ 0.22 หรือ 22 เปอร์เซ็นต์ และ ค่า F1-score เท่ากับ 0.24 หรือ 24 เปอร์เซ็นต์

5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ได้สร้างแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดคลุ้มเลือดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง เป็นการจำแนกประเภทและทำนายข้อมูลแบบ Classification ปัจจัยที่อาจมีผลต่อการทดลองของงานวิจัย คือ จากข้อมูลที่น่ามาวิเคราะห์นั้นเป็นข้อมูลทางการแพทย์ที่มีจำนวนน้อยมากเพียง 1,290 แถว พบข้อมูลที่มีความหลากหลายของแต่ละ Feature และยังเป็นปัญหาข้อมูลที่ไม่สมดุลกัน ซึ่งพบผู้ป่วยที่ไม่เป็นโรคคลุ้มเลือดตันมีจำนวนมากกว่าผู้ป่วยที่เป็นโรคคลุ้มเลือดตัน จะเกิดการ Sensitive กับข้อมูลคลาสที่มากกว่า ทำให้เกิด Bias ขึ้นกับข้อมูล เพราะอัลกอริทึมจะไม่สามารถตรวจพบข้อมูลในคลาสที่น้อยกว่าได้ การจัดการกับปัญหาข้อมูลที่ไม่สมดุลนั้น เป็นวิธีการเปลี่ยนแปลงโครงสร้างและคุณสมบัติของข้อมูลเดิม โดยการเพิ่มเข้าหรือเอาออกอาจทำให้เกิดการสูญเสียคุณสมบัติเดิมของข้อมูลเดิมก็เป็นไปได้ ด้วยความจำเป็นของงานวิจัยนี้จึงต้องจัดการและแก้ปัญหามุมข้อมูลให้สมดุลขึ้น เพื่อที่จะเพิ่มประสิทธิภาพการทำนายของแบบจำลองให้มีประสิทธิภาพที่ดีขึ้น และมุ่งเน้นในการประเมินผลการทดลอง เพื่อให้แบบจำลองสามารถทำนายคนที่เป็นโรคคลุ้มเลือดตันให้ถูกต้องที่สุด โดยเน้นดูผลลัพธ์ที่ค่า Sensitivity, Precision และ F1-score จากผลการทดลองแบบจำลองสามารถทำนายคนที่ไม่เป็นโรคคลุ้มเลือดตันได้ดีกว่า จึงให้ผล Accuracy ที่สูง ซึ่งเกิด Bias ขึ้น เพราะปัญหาข้อมูลที่ไม่สมดุล แบบจำลองได้ไปทำนายคลาสที่ที่จำนวนมากได้ดีกว่า หากจะนำค่า Accuracy มาพิจารณาด้วยนั้นจึงไม่เหมาะสม

ถึงแม้ว่างานวิจัยนี้ได้จัดการกับปัญหาข้อมูลที่ไม่สมดุลแล้วก็ยังให้ประสิทธิภาพของแบบจำลองที่ไม่ดีนัก อาจเกิดจากปัญหาของข้อมูลที่นำมาวิเคราะห์นั้นยังไม่เพียงพอ มีจำนวนข้อมูลที่น้อยมาก ไม่เหมาะสมที่จะนำมาสร้างแบบจำลองให้มีประสิทธิภาพได้

5.3 ข้อเสนอแนะ

5.3.1 ในงานวิจัยเป็นงานวิจัยที่สร้างแบบจำลองการทำนายทางด้านการแพทย์ ซึ่งมีข้อมูลที่นำมาวิเคราะห์จำนวนน้อย อนาคตอาจจะต้องมีข้อมูลสำหรับการวิเคราะห์ที่มากขึ้นกว่านี้ เพื่อผลการทำนายของแบบจำลองที่ดีขึ้น สามารถที่จะนำไปพัฒนาแบบจำลองเพื่อช่วยให้แพทย์วินิจฉัยการเกิดโรคล้มเลือดอุดตันได้มีประสิทธิภาพและแม่นยำมากขึ้น

5.3.2 เนื่องจากข้อมูลที่นำมาวิเคราะห์ในงานวิจัยนี้ ตัวแปรแต่ละคอลัมน์มีความสัมพันธ์กันน้อยมากเมื่อเทียบกับ Class Label ซึ่งข้อมูลทั้งหมดที่แพทย์ผู้เชี่ยวชาญได้เก็บข้อมูลแล้ววิเคราะห์ไว้ในลักษณะเป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Variable) เช่น เทียบกลุ่มอายุมากกว่า 40 ปี หรือ น้อยกว่า 40 ปี ไม่ได้ให้อายุที่เป็นปีจริงๆ กับการเป็น VTE positive และ VTE negative จึงทำให้ไม่มีความสัมพันธ์กันระหว่างตัวแปรในแต่ละคอลัมน์กับ Class Label อาจเป็นปัจจัยหนึ่งที่ทำให้การทำนายของแบบจำลองไม่มีประสิทธิภาพ

5.3.3 อนาคตในการเก็บรวบรวมข้อมูลอาจจะต้องตรวจสอบชนิดการเก็บข้อมูล เพื่อมาวิเคราะห์ให้ละเอียดยิ่งขึ้นและเรียนรู้การวิเคราะห์ตัวเลขในทางการแพทย์ให้มากขึ้น

5.3.4 ลดความหลากหลายของข้อมูล จัดกลุ่มของข้อมูลให้เหมาะสม ทำ Feature selection ลด Dimension ของข้อมูล เพื่อการสร้างแบบจำลองที่มีประสิทธิภาพมากขึ้น

5.3.5 สร้างแบบจำลองด้วยอัลกอริทึมอื่นๆ ที่หลากหลายขึ้น

บรรณานุกรม

- Agharezaei, L., Agharezaei, Z., Nemati, A., Bahaadinbeigy, K., Keynia, F., Baneshi, M., . . . Agharezaei, a. (2016). The Prediction of the Risk Level of Pulmonary Embolism and Deep Vein Thrombosis through Artificial Neural Network. *Acta Informatica Medica*, 24(5).
- Diawkee, T. (2018). หลอดเลือดดำอุดตัน อันตรายที่ต้องระวัง. <https://bit.ly/33tdufF>
- Fei, Y., Hu, J., Gao, K., Tu, J., Li, W. Q., & Wang, W. (2017). Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. *J Crit Care*, 39, 115-123.
- Fei, Y., Hu, J., Li, W. Q., Wang, W., & Zong, G. Q. (2017). Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. *J Thromb Haemost*, 15(3), 439-445.
- Ferroni, P., Zanzotto, F. M., Scarpato, N., Rioldino, S., Guadagni, F., & Roselli, M. (2017). Validation of a Machine Learning Approach for Venous Thromboembolism Risk Prediction in Oncology. *Dis Markers*, 2017, 8781379.
- Ferroni, P., Zanzotto, F. M., Scarpato, N., Rioldino, S., Nanni, U., Roselli, M., & Guadagni, F. (2017). Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients. *Med Decis Making*, 37(2), 234-242.
- Liu, S., Zhang, F., Xie, L., Wang, Y., Xiang, Q., Yue, Z., . . . Yu, C. (2019). Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer. *Int J Med Inform*, 129, 175-183.
- Nafee, T., Gibson, C. M., Travis, R., Yee, M. K., Kerneis, M., Chi, G., . . . Goldhaber, S. Z. (2020). Machine learning to predict venous thrombosis in acutely ill medical patients. *Res Pract Thromb Haemost*, 4(2), 230-237.
- Qatawneh, Z., Alshraideh, M., Almasri, N., Tahat, L., & Awidi, A. (2019). Clinical decision support system for venous thromboembolism risk classification. *Applied Computing and Informatics*, 15(1), 12-18.

- Rojnuckarin, P., Uaprasert, N., Vajragupta, L., Numkarunrunrote, N., Tanpowpong, N., & Sutcharitchan, P. (2011). Risk factors for symptomatic venous thromboembolism in Thai hospitalised medical patients. *Thromb Haemost*, 106(6), 1103-1108.
- SATANGMONGKOL, K. (2019). อธิบาย Logistic Regression พร้อมโค้ดตัวอย่างใน R. <https://datarockie.com/2019/05/12/logistic-regression-r/>
- Uaprasert, N., Chanswangphuwana, C., & al., e. (2020). *Essential Clinical Hematology*.
- แสงมณี, ว., รัตนเจริญเลิศ, ว., โพธิ์รัตน์, ณ., & หนูสวัสดิ์, เ. (2017). การสร้างโมเดลทำนายโอกาสการกลับมารักษาทัวซ้ำของผู้ป่วยโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล.
- ตั้งสุจริตวิจิตร, ว. (2018). รู้เท่ารู้ทันภาวะเสี่ยง "ลิ่มเลือดอุดตันในปอด". <https://bit.ly/324ULXb>
- พัชรวงศ์ศักดิ์, เ. (2020). *A little book of Big Data and Machine Learning*.
- สมาคมฯ ไทย, ส. ค. (2018). "ลิ่มเลือดในหลอดเลือดดำ" โรคอันตรายถึงชีวิตที่น้อยคนจะรู้จัก". <https://bit.ly/2Smqzb8>
- สระอุบล, ก. (2020). เรียนรู้ *Data Science* และ *AI:Machine Learning* ด้วย *Python*.
- หนูสวัสดิ์, ว. แ. ว. ร. ณ. โ. แ. เ. (2017). การสร้างโมเดลทำนายโอกาสการกลับมารักษาทัวซ้ำของผู้ป่วยโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล.
- อัมพวัน, โ.). *Data mining—การทำเหมืองข้อมูล*.

ประวัติผู้เขียน

ชื่อ-สกุล	อัจนราภรณ์ สุขเพิ่ม
วัน เดือน ปี เกิด	23 พฤศจิกายน 2533
สถานที่เกิด	อุบลราชธานี
วุฒิการศึกษา	พ.ศ.2556 วิทยาศาสตรบัณฑิต ฟิสิกส์ จาก มหาวิทยาลัยขอนแก่น
ที่อยู่ปัจจุบัน	23 หมู่ 2 ตำบลไผ่พลย์ อำเภอน้ำขุ่น จังหวัดอุบลราชธานี 34260

