การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง

STROKE PREDICTION USING MACHINE LEARNING

สากล พัชรปัญญวัฒน์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2564

การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง

สากล พัชรปัญญวัฒน์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2564

STROKE PREDICTION USING MACHINE LEARNING

SAKOL PATCHARAPANYAWAT

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

2021

สารนิพนธ์

เรื่อง

การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง

ของ

สากล พัชรปัญญวัฒน์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... ประธาน

(อาจารย์ ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(อาจารย์ ดร.ศุภร คนธภักดี)

ชื่อเรื่อง	การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง
ผู้วิจัย	สากล พัชรปัญญวัฒน์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2564
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. จันตรี ผลประเสริฐ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาการทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองในวัยผู้ใหญ่โดยใช้การเรียนรู้ของเครื่อง การศึกษานี้เราต้องการตรวจสอบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องในสามโมเดลซึ่งประกอบไปด้วยโมเดล Logistic Regression (LR), Random Forest (RF) และ Support vector machine (SVM) เราใช้ชุดข้อมูลการดูแลสุขภาพที่มีอยู่ในชุดข้อมูลของ Kaggle dataset ซึ่งมีข้อมูลผู้ป่วย 5,110 คนและเราเลือกผู้ป่วยเหลือเพียง 4,254 คน ที่เป็นผู้ป่วยวัยผู้ใหญ่ที่มีอายุ 18 ปีขึ้นไป เมทริกซ์ความสับสนใช้สำหรับการสรุปประสิทธิภาพของโมเดลการจำแนกประเภทประกอบไปด้วยค่าความแม่นยำ ค่าความเที่ยงตรง ค่าความไว ค่าความจำเพาะ ค่าประสิทธิภาพโดยรวม (f1-score) และ พื้นที่ใต้กราฟ AUC (Area Under The Curve) จากการทดลองครั้งนี้ RF เป็นโมเดลที่มีประสิทธิภาพที่ดีที่สุดด้วย ค่าความแม่นยำเท่ากับ 0.94 ค่าความเที่ยงตรงเท่ากับ 0.93 ค่าความไวเท่ากับ 0.95 ความจำเพาะเท่ากับ 0.93 ค่า ค่าประสิทธิภาพโดยรวม (f1-score) เท่ากับ 0.94 และค่าพื้นที่ใต้กราฟเท่ากับ 0.94 และสามอันดับสูงสุดของความสำคัญของฟีเจอร์ของโมเดล RF ที่มีลำดับตามความสำคัญจากมากไปน้อยคือตัวแปร อายุ มีค่า 0.38 ค่าเฉลี่ยของระดับน้ำตาลในเลือด มีค่า 0.20 และ ค่าดัชนีมวลกาย มีค่าเท่ากับ 0.15 ตามลำดับ

คำสำคัญ : โรคหลอดเลือดสมอง, การเรียนรู้ของเครื่อง, การประเมินความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง

Title	STROKE PREDICTION USING MACHINE LEARNING
Author	SAKOL PATCHARAPANYAWAT
Degree	MASTER OF SCIENCE
Academic Year	2021
Thesis Advisor	Assistant Professor Dr. Chantri Polprasert

In this study, we developed a machine learning (ML)-based approach for the prediction of stroke risk. To be specific, healthcare datasets containing 5,110 cases that are available in the Kaggle dataset were employed and then only 4,254 cases were selected, all adults, aged 18 years of age or older. In addition, the performance of three popular ML algorithms was compared and investigated, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). A confusion matrix was used to summarize the performance of a classification model with accuracy, precision, recall, f1-score, specificity, and the AUC (Area Under The Curve) represented the degree of separability. In the experiment, RF achieved the best performance with an accuracy of 0.94, a precision of 0.93, a recall of 0.95, an f1-score of 0.94, a specificity of 0.93, and an AUC of 0.94. The top three features of importance of the RF model included age at 0.39, average glucose level of 0.20, and body mass index at 0.15, respectively.

Keyword : Stroke, Machine learning, Stroke risk assessment

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จได้ด้วยการได้รับความอนุเคราะห์จาก ผศ. ดร. จันตรี ผลประเสริฐ อาจารย์ที่ปรึกษาที่ให้คำปรึกษาแนะนำในการทำสารนิพนธ์ตลอดจนสนับสนุนข้อมูลทางวิชาการ และขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ที่ได้ให้คำแนะนำ แนะแนวทางในการปรับปรุงสารนิพนธ์ให้มีความสมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณบัณฑิตวิทยาลัยมหาวิทยาลัยศรีนครินทรวิโรฒสำหรับการสนับสนุน โดยให้แนวทางในการนำเสนอผลงานวิจัยของบัณฑิตศึกษา รวมทั้งสนับสนุนทุนการศึกษาการเข้าร่วมประชุมและนำเสนอผลงานของนิสิต ทำให้ได้ประสบการณ์ในการจัดทำรูปแบบการรายงานและการนำเสนอสารนิพนธ์ให้ได้มาตราฐานอันเป็นการสนับสนุนให้งานวิจัยนี้มีคุณค่าและเป็นแนวทางให้เกิดการแลกเปลี่ยนความรู้ในด้านเทคโนโลยี คณิตศาสตร์ คอมพิวเตอร์ และการใช้เทคนิคการเรียนรู้ของเครื่องในการพัฒนาทางด้านการแพทย์และเป็นแนวทางในการนำไปพัฒนาประเทศชาติต่อไป

สากล พัชรปัญญวัฒน์

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ญ

สารบัญรูปภาพ ..... ฎ

บทที่ 1 บทนำ 1

ภูมิหลัง 1

ความมุ่งหมายของงานวิจัย 4

ความสำคัญของการวิจัย 4

ขอบเขตของการวิจัย 4

กรอบแนวคิดในงานวิจัย 5

บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง 7

1. โรคหลอดเลือดสมอง 7

2. การเรียนรู้ของเครื่อง 14

3. อัลกอริทึมการเรียนรู้ของเครื่องสำหรับการแบ่งประเภทของข้อมูล มีดังนี้ 19

4. การวิเคราะห์ข้อมูล 25

5. ประเมินประสิทธิภาพของแต่ละโมเดล 25

6. การทบทวนวรรณกรรม 32

6.1 Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. (Chun et al., 2021) 32

6.2 Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis. (Site, Nurmi, & Lohan, 2021) 33

6.3 Machine Learning for Brain Stroke: A Review. (Sirsat, Ferme, & Camara, 2020) 34

6.4 Using a Multiclass Machine Learning Model to Predict the Outcome of Acute Ischemic Stroke Requiring Reperfusion Therapy. (Chiu, Zeng, Cheng, Chen, & Lin, 2021) 36

6.5 Assessing stroke severity using electronic health record data: a machine learning approach. (Kogan et al., 2020) 37

6.6 Machine learning to predict mortality after rehabilitation among patients with severe stroke. (Scrutinio et al., 2020) 38

6.7 Value-Based Healthcare in Ischemic Stroke Care: Case-Mix Adjustment Models for Clinical and Patient-Reported Outcomes. (Tsevat & Moriates, 2018) 38

6.8 The Probability of Ischemic Stroke Prediction with a Multi-Neural-Network Model. (Liu, Yin, & Cong, 2020) 39

6.9 Machine Learning Approach to Identify Stroke Within 4.5 Hours. (Lee et al., 2020) 40

6.10 Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study.(Chilamkurthy et al., 2018) 41

6.11 Early Stroke Prediction Using Machine Learning (Sharma, Sharma, Kumar, & Sodhi, 2022) 41

บทที่ 3 วิธีการดำเนินการวิจัย 42

1. การกำหนดกลุ่มประชากร 42

2. ขั้นตอนการดำเนินการวิจัย 43

3. โมเดลที่ใช้ในการวิจัย 45

4. รวบรวมข้อมูล 45
5. จัดเตรียมข้อมูลเพื่อพร้อมสร้างโมเดล 48
6. การประเมินประสิทธิภาพของแต่ละโมเดล 70
บทที่ 4 ผลการดำเนินการวิจัย 71
1. ผลลัพธ์ของการศึกษาและวิเคราะห์ 71
1.1 การสร้างโมเดลกับข้อมูลที่เป็นข้อมูลที่ไม่สมดุล 71
1.1.1 การใช้อัลกอริทึม SVM ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล 71
1.1.2 การใช้อัลกอริทึม LR ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล 74
1.1.3 การใช้อัลกอริทึมป่าสุ่ม ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล 76
1.2 การสร้างโมเดลจากข้อมูลที่เป็นข้อมูลที่สมดุล 78
1.2.1 การใช้อัลกอริทึม SVM ในการสร้างโมเดลจากข้อมูล ที่สมดุล 79
1.2.2 การใช้อัลกอริทึม LR ในการสร้างโมเดลจากข้อมูลที่สมดุล 81
1.2.3 การใช้อัลกอริทึมป่าสุ่ม ในการสร้างโมเดลจากข้อมูลที่สมดุล 83
2. ผลการทดสอบสมมติฐานการวิจัย 90
บทที่5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ 94
สรุปผลการวิจัย 94
อภิปรายผล 95
ข้อเสนอแนะ 98
บรรณานุกรม 100
ประวัติผู้เขียน 103

สารบัญตาราง

หน้า

ตาราง 1 การคำนวณหาค่า Macro-average f1-score 31

ตาราง 2 การคำนวณหาค่า Weighted-average f1-score 31

ตาราง 3 การหาค่า average f1-scores ด้วยวิธี Micro-average method 32

ตาราง 4 แสดง Dataset ถูกแบ่งออกเป็นสองส่วนคือ train set 70%, test set 30 % 44

ตาราง 5 แสดงตัวแปร ประเภทของตัวแปร และความหมายของตัวแปรที่ใช้ในการศึกษา 45

ตาราง 6 แสดงการเปรียบเทียบประสิทธิภาพของโมเดลที่ได้จากข้อมูลที่ไม่สมดุล 78

ตาราง 7 การเปรียบเทียบประสิทธิภาพของโมเดลป่าสุ่ม, LR, SVM ที่ได้จากการใช้ข้อมูลที่สมดุล 85

ตาราง 8 แสดง Best score จากการ fine turning ด้วย GridSearchCV(CV = 10) ตัวพารามิเตอร์ต่างๆของโมเดลป่าสุ่ม, LR, และ SVM 85

ตาราง 9 แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม หลังการทำ hyperparameter tuning 88

ตาราง 10 แสดงการเปรียบเทียบประสิทธิภาพของแต่ละโมเดลที่ได้จากข้อมูลที่สมดุลและข้อมูลที่ไม่สมดุล 90

ตาราง 11 แสดงการสรุปเปรียบเทียบค่าประสิทธิภาพของแต่ละโมเดลที่ได้จาก balanced dataset และ หลังจากการทำ Hyperparameters Tuning 94

สารบัญรูปภาพ

หน้า

ภาพประกอบ 1 แสดงประเภทของโรคหลอดเลือดสมอง 2 ประเภท คือ หลอดเลือดแดงสมองอุดตัน (Ischemic stroke) และหลอดเลือดแดงสมองแตก (Hemorrhagic stroke) 10

ภาพประกอบ 2 แสดงความสัมพันธ์ของการเรียนรู้ของเครื่องเป็นประเภทย่อยของปัญญาประดิษฐ์ และการเรียนรู้เชิงลึกเป็นส่วนย่อยที่มีซับซ้อนของการเรียนรู้ของเครื่อง ตามลำดับ 15

ภาพประกอบ 3 แสดงการฟิตเส้นตรงสำหรับการทำ Linear Regression นั้นวัดค่าความแม่นยำจากผลรวมของระยะห่างระหว่างข้อมูลจริงกับค่าที่ทำนายจากโมเดล 20

ภาพประกอบ 4 แสดง Sigmoid เป็นฟังก์ชัน activation function ซึ่งอยู่ในเส้นโค้งรูปร่าง S..... 21

ภาพประกอบ 5 แสดงในขั้นตอนการทำงานของป่าสุ่ม จะทำการการจำแนกต้นไม้หลาย ๆ ต้น ซึ่งในต้นไม้ แต่ละต้นมีการแบ่งเป็นคลาส โดยที่ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจแต่ละต้นถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด (Majority Voting)..... 22

ภาพประกอบ 6 อัลกอริทึม SVM คือการหาเส้นตรงที่มีมาร์จินที่โตที่สุด (Maximum Margin) ที่สามารถแบ่งข้อมูลออกเป็น 2 คลาส 23

ภาพประกอบ 7 แสดงการจำแนกเชิงเส้นด้วยมาร์จินที่ (margin width) ใหญ่ที่สุด 24

ภาพประกอบ 8 แสดงการคำนวณ การหาค่า Margin 24

ภาพประกอบ 9 แสดงตารางตารางเมทริกซ์ความสับสน (Confusion Matrix)..... 26

ภาพประกอบ 10 แสดงการแบ่งข้อมูลออกเป็น 5-fold cross-validation โดย แบ่งข้อมูล ออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน..... 28

ภาพประกอบ 11 Confusion matrix แสดงตัวอย่างค่าการวัดประสิทธิภาพของโมเดลในการทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง 29

ภาพประกอบ 12 ตัวอย่าง classification_report ที่ได้จากการวัดประสิทธิภาพของโมเดล 30

ภาพประกอบ 13 ภาพแสดงอัลกอริทึมของปัญญาประดิษฐ์ที่โมเดลที่ใช้ ให้ประสิทธิภาพที่ดีในการทำนายผู้ป่วยโรคหลอดเลือดสมอง 34

ภาพประกอบ 14 Pie chart แสดงอัลกอริทึมของการเรียนรู้ของเครื่องที่ใช้ในขบวนการรักษาโรคหลอดเลือดสมองในช่วง การป้องกัน การวินิจฉัยโรค การรักษา และการติดตามผลการรักษา ... 35

ภาพประกอบ 15 แสดงขั้นตอนการดำเนินการวิจัย 43

ภาพประกอบ 16 แสดงจำนวนตัวแปรทั้งหมด 11 ตัวแปร แสดงค่า non-null ของแต่ละตัวแปร และแสดงชนิดของข้อมูล (dtypes) ของแต่ละตัวแปร 47

ภาพประกอบ 17 แสดงค่าต่างๆ ทางสถิติ ของตัวแปรต่างๆที่ใช้ในการศึกษาครั้งนี้ 48

ภาพประกอบ 18 แสดงการทดสอบหา Missing value ของตัวแปร พบว่าตัวแปร ‘bmi’ มีค่า missing value 181 records 49

ภาพประกอบ 19 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ age ในกลุ่มที่มีโอกาสไม่เป็น stroke กับกลุ่มที่มีโอกาสเป็น stroke 50

ภาพประกอบ 20 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ avg_glucose_level ในกลุ่มที่มีโอกาสไม่เน stroke กับกลุ่มที่มีโอกาสเป็น stroke 51

ภาพประกอบ 21 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ bmi ในกลุ่มที่มีโอกาสไม่เป็น stroke กับกลุ่มที่มีโอกาสเป็น stroke 52

ภาพประกอบ 22 แสดงความสัมพันธ์ของตัวแปรตัวที่เป็นตัวเลขกับการเกิดโรคหลอดเลือดสมอง 53

ภาพประกอบ 23 กราฟแสดงความสัมพันธ์ระหว่าง gender และ stroke 54

ภาพประกอบ 24 กราฟแสดงความสัมพันธ์ระหว่าง hypertension และ stroke 55

ภาพประกอบ 25 กราฟแสดงความสัมพันธ์ระหว่าง heart_disease และ stroke 56

ภาพประกอบ 26 กราฟแสดงความสัมพันธ์ระหว่าง ever_married และ stroke 57

ภาพประกอบ 27 กราฟแสดงความสัมพันธ์ระหว่าง work_type และ stroke 58

ภาพประกอบ 28 กราฟแสดงความสัมพันธ์ระหว่าง Residence_type และ stroke 59

ภาพประกอบ *29* กราฟแสดงความสัมพันธ์ระหว่าง smoking_status และ stroke 60

ภาพประกอบ 30 เรากำหนดให้ ‘’stroke” เป็น target attribute 61

ภาพประกอบ 31 กราฟแสดงความไม่สมดุลของข้อมูลของ stroke 61

ภาพประกอบ 32 แสดงการกระจายตัวของข้อมูลและแสดงถึงความสัมพันธ์ของข้อมูลระหว่างสองตัวแปรกับโอกาสการเกิดโรคหลอดเลือดสมอง 62

ภาพประกอบ 33 กราฟแท่งแสดงค่า Correlation Coefficient หรือ ค่าสหสัมพันธ์ระหว่างตัวแปร 2 ตัวที่มีอิทธิพลต่อโอกาสการเกิดโรคหลอดเลือดสมอง 65

ภาพประกอบ 34 แสดงค่า Correlation ที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง 66

ภาพประกอบ 35 แสดงการจัดการกับ missing values ของตัวแปร bmi ด้วยการลบแถวที่มีค่า missing value 67

ภาพประกอบ 36 การใช้ LabelEncoder function จัดการกับ categorical data 68

ภาพประกอบ 37 การทำ standardization หรือ normalization 68

ภาพประกอบ 38 การจัดการ imbalanced data ด้วยเทคนิค 69

ภาพประกอบ 39 กราฟแสดงจำนวนของข้อมูลที่มีความสมดุลของ target attribute (stroke) หลังจากใช้เทคนิค Synthetic Minority Oversampling Technique: SMOTE 70

ภาพประกอบ 40 Classification report ของโมเดล SVM ที่สร้างมาจากข้อมูลที่ไม่สมดุล 72

ภาพประกอบ 41 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล SVM ที่สร้างมาจากข้อมูลที่ไม่สมดุล 73

ภาพประกอบ 42 แสดงประสิทธิภาพของโมเดล LR ที่ได้จากการใช้ข้อมูลจากข้อมูลที่ไม่สมดุล . 74

ภาพประกอบ 43 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่ไม่สมดุล 75

ภาพประกอบ 44 แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากการใช้ข้อมูลข้อมูลที่ไม่สมดุล 76

ภาพประกอบ 45 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลข้อมูลที่ไม่สมดุล 77

ภาพประกอบ 46 Classification report ของโมเดล SVM ที่สร้างมาจากข้อมูลที่สมดุล 79

ภาพประกอบ 47 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล SVM ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล 79

ภาพประกอบ 48 กราฟแสดงประสิทธิภาพของโมเดล SVM ที่ได้จากข้อมูลข้อมูลที่สมดุล..... 80

ภาพประกอบ 49 Classification report ของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล..... 81

ภาพประกอบ 50 Confusion Matrix แสดงประสิทธิภาพของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล 81

ภาพประกอบ 51 กราฟแสดงประสิทธิภาพของโมเดล LR ที่ได้จากข้อมูลข้อมูลที่สมดุล..... 82

ภาพประกอบ 52 Classification report ของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลที่สมดุล 83

ภาพประกอบ 53 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลข้อมูลข้อมูลที่สมดุล..... 83

ภาพประกอบ 54 กราฟแสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากข้อมูลข้อมูลที่สมดุล โดยเปรียบเทียบการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation 84

ภาพประกอบ 55 Classification report หลังจาก fine turning ด้วย GridSearchCV ของโมเดลป่าสุ่ม ที่สร้างมาจาก balanced dataset 86

ภาพประกอบ 56 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม หลังจาก fine turning ด้วย GridSearchCV 87

ภาพประกอบ 57 กราฟค่าพื้นที่ใต้กราฟ AUC (area under curve) ของโมเดลป่าสุ่ม หลังจาก fine turning ด้วย GridSearchCV..... 88

ภาพประกอบ 58 แสดง Feature importance ของโมเดลป่าสุ่ม 89

ภาพประกอบ 59 แสดงการเปรียบเทียบประสิทธิภาพของโมเดล ระหว่างโมเดลที่ได้จากข้อมูลที่เป็น Raw dataset, Normalization dataset และ Standardization dataset (แสดงประสิทธิภาพของโมเดล LR, SVM และ โมเดลป่าสุ่ม) 91

ภาพประกอบ 60 Classification report แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ทดสอบกับ Harvard dataset..... 92

ภาพประกอบ 61 Confusion matrix แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จาก Harvard dataverse..... 93

ภาพประกอบ 62 แสดงความสัมพันธ์ของความสำคัญของฟีเจอร์ของโมเดลป่าสุ่ม ของตัวแปร อายุ และค่าเฉลี่ยของระดับน้ำตาลในเลือด กับโอกาสการเกิดโรคหลอดเลือดสมอง 96

บทที่ 1
บทนำ

ภูมิหลัง

โรคหลอดเลือดสมอง (cerebrovascular disease หรือ stroke) หรือโรคอัมพฤกษ์อัมพาต เป็นปัญหาด้านสุขภาพที่สำคัญที่เกิดขึ้นไปทั่วโลกในปัจจุบัน เนื่องจากเป็นโรคที่พบบ่อยและมีอัตราการเสียชีวิตและความพิการสูง จากข้อมูลจากองค์การอนามัยโลก (World Health Organization : WHO) พบว่า ปี *2563* มีผู้ป่วยเป็นโรคหลอดเลือดสมองกว่า 80 ล้านคน มีผู้เสียชีวิตประมาณ 5.5 ล้านคนและผู้ที่รอดชีวิตมักจะมีความพิการหลงเหลืออยู่ นอกจากนี้ยังพบผู้ป่วยรายใหม่เพิ่มขึ้นถึง 14.5 ล้านคนต่อปี โดยใน 25% เป็นผู้ป่วยที่มีอายุ 25 ปีขึ้นไป (Hfocus., Thursday, 29 October 2020)

สำหรับประเทศไทย จากรายงานข้อมูลย้อนหลัง 5 ปี ของกองยุทธศาสตร์และแผนงาน กระทรวงสาธารณสุข พบว่าจำนวนผู้ป่วยโรคหลอดเลือดสมอง ตั้งแต่ปี 2556-2560 จำนวนผู้ป่วยโรคหลอดเลือดสมองมีแนวโน้มเพิ่มสูงขึ้นทุกปี โดยในปี 2560 พบผู้ป่วยรายใหม่จำนวน 304,807 ราย และมีผู้เสียชีวิตจากโรคนี้ปีละไม่ต่ำกว่า 30,000 ราย (Hfocus., Thursday, 29 October 2020) และผู้ป่วยที่รอดชีวิตต้องเผชิญกับ ความพิการของร่างกาย ปัญหาการสื่อสาร ความรู้สึกนึกคิดที่เปลี่ยนไป ปัญหาเรื่องค่าใช้จ่าย การตกงาน และเรื่องสถานะและการใช้ชีวิตในสังคมที่เปลี่ยนไป ดังนั้นการป้องกันการเกิดโรคหลอดเลือดสมองจึงมีความสำคัญมากที่จะช่วยลดโอกาสการเกิดโรคหลอดเลือดสมองและการให้การรักษาที่รวดเร็วทันท่วงทีก็จะช่วยลดอัตราการตายและระดับความรุนแรงของความพิการลงได้ ดังนั้นจึงมีความจำเป็นอย่างมากที่จะต้องมีการพัฒนาระบบการดูแลรักษาผู้ป่วยโรคหลอดเลือดสมองให้มีมาตรฐานและครบวงจรในทุกระยะของการรักษา คือ

1. ระยะการป้องกัน คือการป้องกันและควบคุมปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคหลอดเลือดสมอง

2. ระยะการให้การรักษา คือการให้การรักษาผู้ป่วยโรคหลอดเลือดสมองตั้งแต่ในระยะเฉียบพลันที่รวดเร็วและได้มาตรฐาน ที่จะสามารถช่วยลดอัตราการตาย ลดภาวะแทรกซ้อนและลดความพิการของผู้ป่วย

3. ระยะการติดตามการรักษา คือการป้องกันการเกิดซ้ำด้วยการดูแลรักษาอย่างต่อเนื่องและติดตามดูแลควบคุมปัจจัยเสี่ยงต่าง ๆ ที่จะทำให้เกิดโรคหลอดเลือดสมอง

ซึ่งขั้นตอนเหล่านี้หากเป็นการดูแลรักษาที่มีมาตรฐานและมีประสิทธิภาพสูงสุดจะสามารถช่วยให้ผู้ป่วยฟื้นตัวได้ดีและช่วยลดค่าใช้จ่ายในการรักษาได้อย่างมาก อันจะส่งผลให้ลดความสูญเสียทางเศรษฐกิจของครอบครัว สังคม และประเทศได้ ซึ่งขบวนการดูแลรักษาในปัจจุบันนี้ต้องอาศัยแพทย์และบุคลากรทางการแพทย์ที่มีความเชี่ยวชาญ ในการรักษาและนำ เอาเทคโนโลยีสมัยใหม่ต่าง ๆ มาช่วยสนับสนุนการตัดสินใจในขบวนการการรักษาเพื่อให้มีการให้บริการที่มีมาตรฐานและผลการรักษาที่มีประสิทธิภาพสูงสุด

ดังนั้นการรักษาโรคหลอดเลือดสมองในปัจจุบันมีการนำ เอาเทคโนโลยีสมัยใหม่ต่าง ๆ เข้ามาใช้ในระบบการให้บริการทางการแพทย์อย่างมากมาย ทำให้เกิดข้อมูลที่มีจำนวนมหาศาลและหลากหลายทั้งข้อมูลการบันทึกประวัติผู้ป่วย บันทึกของแพทย์ บันทึกข้อมูลยา บันทึกอาการ บันทึกผลการตรวจจากห้องทดสอบ ภาพถ่ายรังสี ผลวินิจฉัย ข้อมูลการเบิกค่ารักษา ข้อมูลเศรษฐกิจเชิงสังคมของผู้ป่วย ไปจนถึงข้อมูลใหม่ ๆจากอุปกรณ์สวมใส่ติดตามอาการของผู้ป่วยและข้อมูลอื่น ๆ ที่ได้จาก Social media ต่าง ๆ โดยข้อมูลเหล่านี้มีความหลากหลายอย่างมาก ทั้งในเชิงขนาดของข้อมูล รูปแบบและความเร็วในการผลิตข้อมูล

ด้วยเหตุดังกล่าวการใช้ ปัญญาประดิษฐ์ (Artificial Intelligence : AI) ในการจัดการกับข้อมูลจำนวนมหาศาล (Big data) เพื่อสนับสนุนการนำข้อมูลมาใช้ประโยชน์ในขบวนการรักษา เช่น การใช้การวิเคราะห์ข้อมูลด้วยอัลกอริทึมของปัญญาประดิษฐ์ช่วยในการตัดสินใจในการวางแผนให้การรักษาโรคหลอดเลือดสมองในแต่ละระยะ จึงมีความสำคัญมากเพราะการวิเคราะห์หรือการตัดสินใจของมนุษย์หรือแพทย์ผู้เชี่ยวชาญที่ให้การวิเคราะห์วินิจฉัยโรคกับข้อมูลจำนวนมากมักจะเป็นข้อจำกัดที่อาจทำให้เกิดความผิดพลาดได้เมื่อต้องทำการตัดสินใจ ในขณะที่อัลกอริทึมปัญญาประดิษฐ์ไม่ได้อยู่ภายใต้ข้อจำกัดดังกล่าว ความจำเป็นในการตรวจหาความเสี่ยงของการเกิดโรคหลอดเลือดสมองตั้งแต่เนิ่น ๆ การวินิจฉัยที่แม่นยำ และการตัดสินใจให้การรักษาอย่างทันท่วงทีได้ส่งเสริมการใช้ ปัญญาประดิษฐ์ ในการดูแลโรคหลอดเลือดสมองเพิ่มมากขึ้น ตัวอย่าง เช่น Chilamkurthy และคณะ (Chilamkurthy et al., 2018)ใช้อัลกอริทึมการเรียนรู้เชิงลึกเพื่อระบุความผิดปกติโดยอัตโนมัติในการสแกนด้วยเครื่องเอกซเรย์คอมพิวเตอร์ที่ศีรษะและสมอง ทำงานได้ดีในการตรวจจับการตกเลือดในกะโหลกศีรษะ โดยโมเดลให้*ค่า* Area under the curve (AUC) เท่ากับ 0.94 ในการศึกษาอื่น Titano และคณะ (Titano et al., 2018) ได้แสดงให้เห็นถึงประสิทธิภาพของโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network :

CNN) ผ่านการทดลองแบบสุ่มที่มีกลุ่มควบคุม (blinded randomized controlled trial) แบบ double-blinded ซึ่งแสดงให้เห็นว่าระบบที่ใช้การเรียนรู้เชิงลึกสามารถตรวจจับเหตุการณ์ทางระบบประสาทเฉียบพลันในการถ่ายภาพกะโหลกได้เร็วกว่านักรังสีวิทยา

ดังนั้นการใช้อัลกอริทึมการเรียนรู้เชิงลึก (Deep learning) กับภาพทางการแพทย์สำหรับการวินิจฉัยโดยใช้คอมพิวเตอร์ช่วย ได้แสดงผลในเชิงบวกอย่างชัดเจนต่อประสิทธิภาพและคุณภาพของขั้นตอนการทำงานสำหรับการดูแลโรคหลอดเลือดสมอง ระบบปัญญาประดิษฐ์ ยังสามารถช่วยให้นักประสาทวิทยาโรคหลอดเลือดสมองระบุผู้ป่วยโรคหลอดเลือดสมองตีบเฉียบพลันได้เนื่องจากการอุดตันของหลอดเลือดขนาดใหญ่ โดยระบบแพลตฟอร์มการอ่านผลภาพจากเครื่องเอกซเรย์คอมพิวเตอร์ที่ศีรษะและสมอง ได้อย่างรวดเร็ว ได้รับการพิสูจน์แล้วว่าเป็นเครื่องมือที่มีประสิทธิภาพในการระบุผู้ป่วยที่มีการอุดตันของหลอดเลือดที่สามารถได้รับประโยชน์จากการรักษาเพื่อเปิดหลอดเลือดอีกครั้ง หรือการให้ยาละลายลิ่มเลือด ในอนาคตอันใกล้การผสมผสานระหว่าง ปัญญาประดิษฐ์ และการแพทย์ทางไกลสามารถมีบทบาทสำคัญในการประเมินอย่างรวดเร็วของการวินิจฉัยโรคหลอดเลือดสมองที่แม่นยำ และช่วยให้การพิจารณาให้การรักษาเพื่อเปิดหลอดเลือดอีกครั้ง หรือการให้ยาละลายลิ่มเลือดเป็นไปได้อย่างรวดเร็วทันเวลา

การทำงานร่วมกันระหว่างมนุษย์กับระบบปัญญาประดิษฐ์ สามารถเป็นพันธมิตรที่มีคุณค่าสำหรับการแพทย์ในหลาย ๆ ด้าน โดยความร่วมมือระหว่างแพทย์และเทคโนโลยีปัญญาประดิษฐ์ จะเป็นกุญแจสำคัญในการปรับปรุงคุณภาพและประสิทธิภาพของบริการด้านสุขภาพ ซึ่งสอดคล้องกับแนวทางของการให้การบริการทางการแพทย์ในระดับสากลคือหลักการการดูแลสุขภาพโดยเน้นคุณค่า (Value Based-Healthcare) โดย ศาสตราจารย์ Michael E. Porter แห่งมหาวิทยาลัยฮาร์วาด ได้นำแนวคิดทางธุรกิจที่มุ่งเน้น คุณค่าต่อลูกค้ามาประยุกต์ใช้ในระบบสุขภาพ ทั้งนี้ไม่ได้มุ่งเน้นที่การลดค่าใช้จ่ายโดยตรง แต่ให้ความสำคัญที่การเพิ่มผลลัพธ์สุขภาพ ของผู้ป่วย โดยแนวคิดนี้มีอิทธิพลต่อการปรับเปลี่ยนระบบการบริหารจัดการบริการสุขภาพในหลายประเทศ และสอดคล้องกับนโยบายของรัฐบาลของไทยในยุทธศาสตร์ Thailand 4.0 ที่สนับสนุนให้ใช้เทคโนโลยีในการพัฒนาการแพทย์ไทย

สำหรับการศึกษานี้ เราจะใช้ขบวนการเรียนรู้ของเครื่อง (Machine learning : ML) ซึ่งเป็นอัลกอริทึมอันหนึ่งที่สำคัญของระบบปัญญาประดิษฐ์ ในการทำนายโอกาสการเกิดโรคหลอดเลือดสมองในผู้ป่วยที่มีปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง เพื่อช่วยในการวางแผนการรักษาและลดความเสี่ยงลง

ความมุ่งหมายของงานวิจัย

ในการวิจัยครั้งนี้ได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อหาโมเดลการเรียนรู้ของเครื่องที่ช่วยสนับสนุนการตัดสินใจ ในขบวนการการวางแผนการรักษาของแพทย์ในการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง

2. เพื่อศึกษาตัวแปรที่มีความสัมพันธ์กับการเกิดโรคหลอดเลือดสมองโดยใช้หลักการการเรียนรู้ของเครื่อง เพื่อนำตัวแปรเหล่านั้นไปใช้ในการสร้างโมเดลที่ดีที่สุด ในการทำนายโอกาสของการเกิดโรคหลอดเลือดสมอง

3. เพื่อเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมของการเรียนรู้ของเครื่อง ในการทำการจัดหมวดหมู่ (classification) ในการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง โดยใช้ confusion matrix, accuracy, sensitivity, specificity, f1-score

ความสำคัญของการวิจัย

การวิจัยนี้เป็นการศึกษาเพื่อแบ่งกลุ่มของผู้ป่วยออกเป็นสองกลุ่ม คือ กลุ่มผู้ที่มีโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองและกลุ่มผู้ที่ไม่มีโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง ซึ่งแบบจำลองของการทำนายนี้จะเป็นประโยชน์มากในการช่วยสนับสนุนการตัดสินใจทางการแพทย์ในการวางแผนการป้องกันและรักษาผู้ที่มีปัจจัยของความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง โดยการสร้างโมเดลในการทำนายโรคหลอดเลือดสมองเราใช้ ขบวนการเรียนรู้ของเครื่อง เป็นเครื่องมือในการสร้างโมเดล โดยข้อมูลที่ใช้ในการสร้างโมเดลเป็นข้อมูลผู้ป่วยที่ได้จาก Kaggle dataset มีขนาด 317 KB ในรูปแบบไฟล์ CSV ข้อมูลของเราประกอบไปด้วย 11 features และ 5,110 records

ขอบเขตของการวิจัย

ประชากรที่ใช้ในการวิจัย

เป็นข้อมูลที่ได้มาจาก Kaggle dataset มีขนาด 317 KB ในรูปแบบไฟล์ CSV ข้อมูลของเราประกอบไปด้วย 11 features และ 5,110 records

กลุ่มตัวอย่างที่ใช้ในการวิจัย

ข้อมูลผู้ป่วยที่เป็นวัยผู้ใหญ่ที่มีอายุ 18 ปี ขึ้นไป ที่ได้จาก Kaggle dataset โดยแบ่งเป็น 2 กลุ่ม คือ

ผู้ป่วยที่มีปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง และผู้ป่วยที่เป็นโรคหลอดเลือดสมอง ตัวแปรที่ศึกษามีดังนี้

ตัวแปรใช้ในการวิจัย

ตัวแปรต้น ประกอบด้วย 10 ตัวแปรคือ gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status, avg_glucose_level, bmi, age

ตัวแปรตาม 1 ตัวแปร คือ stroke

กรอบแนวคิดในงานวิจัย

การวิจัยนี้เป็นการศึกษาเพื่อแบ่งกลุ่มของผู้ป่วยออกเป็นสองกลุ่ม (binary classification) คือ กลุ่มผู้ที่มีโอกาสจะเกิดโรคหลอดเลือดสมองและผู้ป่วยที่ไม่มีโอกาสเกิดโรคหลอดเลือดสมอง โดยใช้ การเรียนรู้ของเครื่อง เป็นเครื่องมือสำหรับสร้างแบบจำลองในการทำนายการเกิดโรคหลอดเลือดสมอง ของข้อมูลผู้ป่วยที่เก็บไว้ ในรูปแบบไฟล์ CSV ที่ประกอบไปด้วย 11 features และ 5,110 records โดย

1. ข้อมูลจะถูกเก็บในรูปแบบตาราง จากนั้นจึงใช้ การเรียนรู้ของเครื่อง มาเป็นเครื่องมือช่วยสร้าง โมเดลในการทำนายโอกาสการเกิดโรคหลอดเลือดสมองซึ่งการวิเคราะห์ข้อมูลและการสร้างโมเดลการทำนายถูกสร้างโดยใช้ภาษาโปรแกรม Python และใช้

2. ใช้ขบวนการตรวจสอบสำรวจข้อมูล (Exploratory data analysis *หรือ* EDA) ในการจัดการกับค่า missing value ของข้อมูล

3. การใช้ Synthetic Minority Oversampling Technique (SMOTE technique) จัดการกับ imbalanced dataset

4. การใช้ Feature selection techniques ก่อนการนำข้อมูลมาใช้ในการสร้างโมเดล

5. การสร้างแบบจำลองโดยใช้ Logistic regression, Support Vector Machine and Random Forest

6. วัดผลและเปรียบเทียบ performance โดยใช้ confusion matrix, accuracy, sensitivity, specificity, f1-score

สมมติฐานในการวิจัย

1. การใช้เทคนิค Synthetic Minority Oversampling (SMOTE) ในการจัดการกับ imbalanced dataset จะช่วยเพิ่มประสิทธิภาพของโมเดลในการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง

2. การใช้ Feature Engineering ก่อนการนำข้อมูลมาใช้ในการสร้างโมเดล จะช่วยเพิ่มประสิทธิภาพที่ได้จากแบบจำลองจากการเรียนรู้ของเครื่อง

ข้อจำกัดในการวิจัย

การศึกษานี้เป็นการศึกษาที่ใช้ข้อมูลจากแหล่งเดียว (single-center study) จึงควรมีการตรวจสอบ validation ด้วยข้อมูลจากแหล่งอื่นๆ (multi-center validation)

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องและได้นำเสนอตาม หัวข้อต่อไปนี้

1. โรคหลอดเลือดสมอง
2. อัลกอริทึมการเรียนรู้ของเครื่องสำหรับการแบ่งประเภทของข้อมูล
3. การเลือกใช้โมเดล
4. การวิเคราะห์ข้อมูล
5. ประเมินประสิทธิภาพของแต่ละโมเดล
6. การทบทวนวรรณกรรม

1. โรคหลอดเลือดสมอง

โรคหลอดเลือดสมอง (Cerebrovascular disease or stroke) เกิดขึ้นเมื่อเลือดที่ไปเลี้ยงสมองบางส่วนถูกขัดขวางหรือมีปริมาณลดลง ทำให้เนื้อเยื่อสมองได้รับออกซิเจนและสารอาหารน้อยลงหรือไม่ได้รับออกซิเจนและสารอาหารเลย ทำให้เซลล์สมองเริ่มตายในเวลาไม่กี่นาที โรคหลอดเลือดสมองเป็นเหตุฉุกเฉินทางการแพทย์ และการรักษาอย่างทันท่วงทีเป็นสิ่งสำคัญ การดำเนินการในระยะแรกสามารถลดความเสียหายของสมองและภาวะแทรกซ้อนอื่นๆ ได้

1.1 อาการ (Symptoms)

เราสามารถสังเกตคนรอบข้างว่ามีอาการของโรคหลอดเลือดสมองที่เราสามารถสังเกต ได้แก่

1.1.1 ปัญหาในการพูดและการเข้าใจสิ่งที่คนอื่นพูด คุณอาจพบความสับสน พูดไม่ชัด หรือมีปัญหาในการทำความเข้าใจคำพูด

1.1.2 อัมพาตหรือชาที่ใบหน้า แขนหรือขา คุณอาจมีอาการชา อ่อนแรง หรืออัมพาตที่ใบหน้า แขนหรือขาอย่างกะทันหัน ซึ่งมักส่งผลกระทบเพียงด้านเดียวของร่างกาย พยายามยกแขนทั้งสองขึ้นเหนือศีรษะพร้อมกัน หากแขนข้างหนึ่งเริ่มตกลงมา คุณอาจเป็นโรคหลอดเลือดสมอง นอกจากนี้ ปากข้างหนึ่งของคุณอาจหย่อนยานเมื่อคุณพยายามยิ้ม

1.1.3 ปัญหาการมองเห็นในตาข้างเดียวหรือทั้งสองข้าง ทันใดนั้นคุณอาจมองเห็นภาพซ้อนเบลอหรือดำคล้ำในตาข้างเดียวหรือทั้งสองข้าง หรืออาจมองเห็นเป็นสองเท่า

1.1.4 ปวดศีรษะ. อาการปวดหัวอย่างกะทันหันและรุนแรง ซึ่งอาจมาพร้อมกับการอาเจียน เวียนศีรษะ หรือความรู้สึกผิดเพี้ยน อาจบ่งชี้ว่าคุณกำลังเป็นโรคหลอดเลือดสมอง

1.1.5 เดินลำบาก คุณอาจสะดุดหรือเสียสมดุล คุณอาจมีอาการวิงเวียนศีรษะอย่างกะทันหันหรือสูญเสียการประสานงาน

1.2 เมื่อไหร่จะไปหาหมอ (When to see a doctor)

ถ้าคุณพบสัญญาณใดๆหรืออาการของโรคหลอดเลือดสมอง Fast Stroke คือ อีกหนึ่งวิธีในการสังเกตตัวเองและคนใกล้ชิดว่ามีอาการของโรคหลอดเลือดสมองหรือไม่ โดยให้สังเกตอาการ ' F.A.S.T ' เพื่อการให้ความช่วยเหลือทางการแพทย์ทันที

1.2.1 Face ขอให้บุคคลนั้นยิ้ม ใบหน้าด้านใดด้านหนึ่งหย่อนยานหรือไม่

1.2.2 Arms ขอให้บุคคลนั้นยกแขนทั้งสองข้าง แขนข้างหนึ่งห้อยลงหรือไม่หรือแขนข้างหนึ่งไม่สามารถยกขึ้นได้

1.2.3 Speech ขอให้บุคคลนั้นพูดคำง่ายๆ คำพูดของเขาจะไม่ชัดเจนหรือแปลกๆหรือไม่

1.2.4 Time หากคุณสังเกตเห็นสัญญาณเหล่านี้ ให้รีบนำส่งโรงพยาบาลทันที

1.3 สาเหตุ (Causes)

มีสองสาเหตุหลักของโรคหลอดเลือดสมอง: หลอดเลือดแดงอุดตัน (โรคหลอดเลือดสมองตีบ) หรือการรั่วไหลหรือระเบิดของหลอดเลือด (จังหวะเลือดออก) บางคนอาจมีเพียงการหยุดชะงักชั่วคราวของการไหลเวียนของเลือดไปยังสมองที่เรียกว่าการโจมตีขาดเลือดชั่วคราว (TIA) ที่ไม่ก่อให้เกิดอาการถาวร

1.3.1 โรคหลอดเลือดสมองตีบ (Ischemic stroke)

นี่เป็นโรคหลอดเลือดสมองชนิดที่พบบ่อยที่สุด มันเกิดขึ้นเมื่อหลอดเลือดในสมองตีบหรืออุดตัน ทำให้เลือดไหลเวียน (ischemia) ลดลงอย่างรุนแรง หลอดเลือดตีบหรือตีบตันเกิดจากไขมันสะสมในหลอดเลือดหรือลิ่มเลือดหรือเศษเนื้อเยื่ออื่นๆ ที่เดินทางผ่านกระแสเลือด ส่วนใหญ่มักมาจากหัวใจ และติดอยู่ในหลอดเลือดในสมอง ในการวิจัยเบื้องต้นของงานวิจัยหลายชิ้นทำให้พบว่าการติดเชื้อ COVID-19 อาจเพิ่มความเสี่ยงต่อโรคหลอดเลือดสมองตีบ แต่ก็ยังจำเป็นต้องมีการศึกษาเพิ่มเติมอีก

1.3.2 โรคหลอดเลือดสมองแตก (Hemorrhagic stroke)

โรคหลอดเลือดสมองตีบเกิดขึ้นเมื่อหลอดเลือดในสมองรั่วหรือแตก ภาวะเลือดออกในสมองอาจเกิดจากภาวะต่างๆ ที่ส่งผลต่อหลอดเลือด ปัจจัยที่เกี่ยวข้องกับโรคหลอดเลือดสมองได้แก่:

1.3.2.1 ความดันโลหิตสูงที่ไม่สามารถควบคุมได้

1.3.2.2 การรักษามากเกินไปด้วยยาป้องกันการแข็งตัวของเลือดเช่นยาวาร์ฟาริน (anticoagulants)

1.3.2.3 เส้นเลือดโป่งพอง (aneurysms) ที่ถือว่าเป็นจุดอ่อนในผนังหลอดเลือด (โป่งพอง) การบาดเจ็บ เช่น อุบัติเหตุทางรถยนต์ (Trauma)

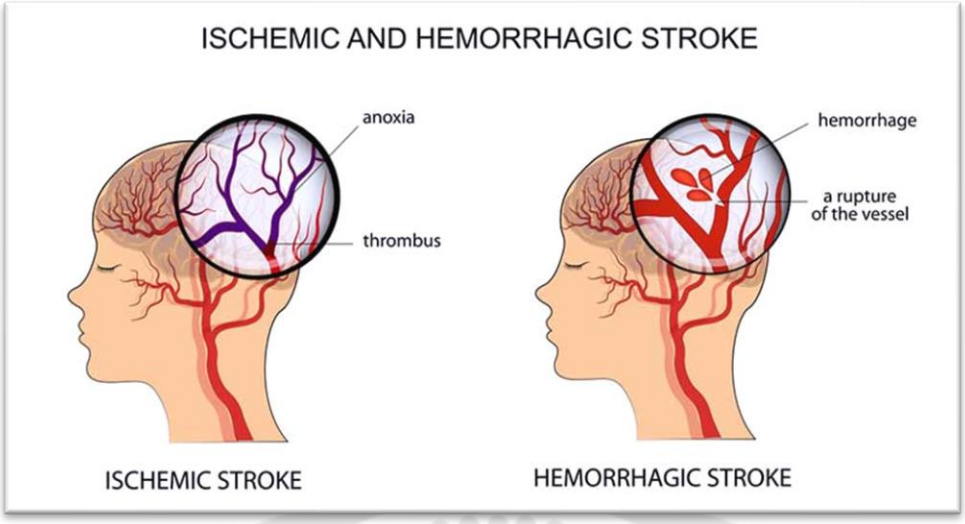
1.3.2.4 การสะสมของโปรตีนในผนังหลอดเลือดที่นำไปสู่ความอ่อนแอในผนังหลอดเลือด *(*cerebral amyloid angiopathy)

1.3.2.5 โรคหลอดเลือดแดงสมองตีบที่นำไปสู่การเกิดโรคหลอดเลือดแดงสมองแตก

อีกสาเหตุหนึ่งที่พบได้น้อยกว่าของการมีเลือดออกในสมองคือการแตกของหลอดเลือด AVM (Cerebral arteriovenous malformation) หรือโรคหลอดเลือดสมองเอวีเอ็ม

1.3.3 โรคสมองขาดเลือดชั่วคราว (Transient ischemic attack: TIA)

“MiniStroke” หรือ TIA (Transient Ischemic Attack) คือ อาการขาดเลือดชั่วคราว เป็นการหยุดชะงักของการไหลเวียนของเลือดไปยังส่วนหนึ่งของสมองไขสันหลังหรือจอประสาทตา ซึ่งอาจทำให้เกิดอาการคล้ายโรคหลอดเลือดสมอง แต่ไม่ทำลายเซลล์สมองหรือทำให้เกิดความพิการถาวร แต่ยังเป็นสิ่งจำเป็นที่เราจะต้องให้ความใส่ใจในการดูแลรักษาตั่งแต่ในช่วงแรกๆที่มีอาการ แม้ว่าจะเป็นอาการขาดเลือดชั่วคราวก็ตามเพราะถึงแม้ว่าอาการจะดีขึ้น หากมีอาการของอาการขาดเลือดชั่วคราว แสดงว่าเราอาจจะมีปัญหาของหลอดเลือดแดงอุดตันหรือตีบตันที่อาจจะไหลไปสู่หลอดเลือดในสมอง ดังนั้นการมีการขาดเลือดชั่วคราว จะทำให้เพิ่มความเสี่ยงที่จะเป็นโรคหลอดเลือดสมองอย่างเต็มตัวได้ในภายหลัง



ภาพประกอบ 1 แสดงประเภทของโรคหลอดเลือดสมอง 2 ประเภท คือ หลอดเลือดแดงสมองอุดตัน (Ischemic stroke) และหลอดเลือดแดงสมองแตก (Hemorrhagic stroke)

ที่มา : (Vectorstock, 2020 #60)

1.4 ปัจจัยเสี่ยง (Risk factors)

หลายปัจจัยสามารถเพิ่มความเสี่ยงของโรคหลอดเลือดสมองได้ ปัจจัยเสี่ยงที่อาจรักษาได้ ได้แก่

1.4.1 ปัจจัยเสี่ยงด้านไลฟ์สไตล์ (Lifestyle risk factors)

1.4.1.1 น้ำหนักเกินหรืออ้วน

1.4.1.2 การไม่ออกกำลังกาย

1.4.1.3 ดื่มหนักหรือเมามาย

1.4.1.4 การใช้ยาเสพติดที่ผิดกฎหมาย เช่น โคเคนและเมทแอมเฟตามีน

1.4.2 ปัจจัยเสี่ยงทางการแพทย์ (Medical risk factors)

1.4.2.1 ความดันโลหิตสูง

1.4.2.2 การสูบบุหรี่หรือการได้รับควันบุหรี่มือสอง

1.4.2.3 คอเลสเตอรอลสูง

1.4.2.4 โรคเบาหวาน

1.4.2.5 การหยุดหายใจขณะหลับ (sleep apnea)

1.4.2.6 โรคหัวใจและหลอดเลือด ได้แก่ ภาวะหัวใจล้มเหลว หัวใจบกพร่อง หัวใจติดเชื้อ หรือจังหวะการเต้นของหัวใจผิดปกติ เช่น ภาวะหัวใจห้องบนสั่นพลิ้ว

1.4.2.7 ประวัติส่วนตัวหรือครอบครัวของโรคหลอดเลือดสมอง หัวใจวาย หรืออาการขาดเลือดชั่วคราว

1.4.2.8 การติดเชื้อโควิด-19

1.4.3 ปัจจัยอื่นๆ ที่เกี่ยวข้องกับความเสี่ยงที่สูงขึ้นของโรคหลอดเลือดสมอง ได้แก่:

1.4.3.1 อายุ — ผู้ที่มีอายุ 55 ปีขึ้นไปมีความเสี่ยงต่อโรคหลอดเลือดสมองมากกว่าคนที่อายุน้อยกว่า

1.4.3.2 เชื้อชาติหรือชาติพันธุ์ — ชาวแอฟริกันอเมริกันและฮิสแปนิกมีความเสี่ยงที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่มาจากเชื้อชาติหรือชาติพันธุ์อื่น

1.4.3.3 เพศ — ผู้ชายมีความเสี่ยงต่อโรคหลอดเลือดสมองมากกว่าผู้หญิง ผู้หญิงมักจะแก่กว่าเมื่อมีโรคหลอดเลือดสมอง และมีโอกาสเสียชีวิตจากโรคหลอดเลือดสมองมากกว่าผู้ชาย

1.4.3.4 ฮอร์โมน — การใช้ยาคุมกำเนิดหรือการบำบัดด้วยฮอร์โมนที่มีเอสโตรเจนเพิ่มความเสี่ยง

1.5 ภาวะแทรกซ้อน (Complications)

โรคหลอดเลือดสมองบางครั้งอาจทำให้เกิดความทุพพลภาพชั่วคราวหรือถาวรได้ ขึ้นอยู่กับระยะเวลาที่สมองขาดการไหลเวียนของเลือดและส่วนใดได้รับผลกระทบ ภาวะแทรกซ้อนอาจรวมถึง

1.5.1 อัมพาตหรือสูญเสียการเคลื่อนไหวของกล้ามเนื้อ คุณอาจเป็นอัมพาตที่ด้านใดด้านหนึ่งของร่างกาย หรือสูญเสียการควบคุมกล้ามเนื้อบางส่วน เช่น กล้ามเนื้อข้างใดข้างหนึ่งหรือแขนข้างหนึ่ง

1.5.2 พูดหรือกลืนลำบาก โรคหลอดเลือดสมองอาจส่งผลต่อการควบคุมกล้ามเนื้อในปากและลำคอ ทำให้พูดไม่ชัดเจน กลืนหรือกินได้ยาก คุณอาจมีปัญหาด้านภาษา เช่น การพูดหรือทำความเข้าใจคำพูด การอ่าน หรือการเขียน

1.5.3 สูญเสียความทรงจำหรือมีปัญหาในการคิด หลายคนที่เป็นโรคหลอดเลือดสมองประสบกับการสูญเสียความทรงจำบางส่วน คนอื่นอาจมีปัญหาในการคิด การให้เหตุผล การตัดสิน และเข้าใจแนวคิด

1.5.4 ปัญหาทางอารมณ์ ผู้ที่เคยเป็นโรคหลอดเลือดสมองอาจควบคุมอารมณ์ได้ยากขึ้น หรืออาจมีอาการซึมเศร้าได้

1.5.5 ความเจ็บปวด. ความเจ็บปวด อาการชา หรือความรู้สึกผิดปกติอื่นๆ อาจเกิดขึ้นในส่วนต่างๆ ของร่างกายที่ได้รับผลกระทบจากโรคหลอดเลือดสมอง ตัวอย่างเช่น หากโรคหลอดเลือดสมองทำให้คุณเสียความรู้สึกที่แขนซ้าย คุณอาจจะรู้สึกเสียวซ่าที่แขนนั้นได้

1.5.6 การเปลี่ยนแปลงพฤติกรรมและความสามารถในการดูแลตนเอง ผู้ที่มีจังหวะอาจจะถอนตัวมากขึ้น พวกเขาอาจต้องการความช่วยเหลือเกี่ยวกับการดูแลและงานบ้านในแต่ละวัน

1.6 การป้องกัน (Prevention)

การทราบปัจจัยเสี่ยงโรคหลอดเลือดสมองของคุณ การทำตามคำแนะนำของผู้ให้บริการดูแลสุขภาพและการใช้ชีวิตอย่างมีสุขภาพดีเป็นขั้นตอนที่ดีที่สุดที่คุณสามารถทำได้ในการป้องกันโรคหลอดเลือดสมอง หากคุณเป็นโรคหลอดเลือดสมองหรือโรคสมองขาดเลือดชั่วคราว มาตรการเหล่านี้จะช่วยป้องกันการเกิดหรือเกิดซ้ำของโรคหลอดเลือดสมองอีก การดูแลติดตามผลที่คุณได้รับในโรงพยาบาลและหลังจากนั้นอาจมีบทบาทเช่นกัน กลยุทธ์การป้องกันโรคหลอดเลือดสมองหลายแบบเหมือนกับกลยุทธ์ในการป้องกันโรคหัวใจ โดยทั่วไป คำแนะนำในการดำเนินชีวิตอย่างมีสุขภาพ ได้แก่:

1.6.1 การควบคุมความดันโลหิตสูง นี่เป็นหนึ่งในสิ่งที่สำคัญที่สุดที่คุณสามารถทำได้เพื่อลดความเสี่ยงของโรคหลอดเลือดสมอง หากคุณเป็นโรคหลอดเลือดสมอง การลดความดันโลหิตสามารถช่วยป้องกัน TIA หรือโรคหลอดเลือดสมองได้ การเปลี่ยนแปลงวิถีชีวิตและการใช้ยาที่ดีต่อสุขภาพมักใช้รักษาความดันโลหิตสูง

1.6.2 ลดปริมาณคอเลสเตอรอลและไขมันอิ่มตัวในอาหารของคุณ การรับประทานคอเลสเตอรอลและไขมันให้น้อยลง โดยเฉพาะไขมันอิ่มตัวและไขมันทรานส์ อาจช่วยลดการสะสมในหลอดเลือดแดงได้ หากคุณไม่สามารถควบคุมคอเลสเตอรอลด้วยการเปลี่ยนแปลงอาหารเพียงอย่างเดียว แพทย์อาจสั่งยาลดคอเลสเตอรอล

1.6.3 การเลิกใช้ยาสูบ การสูบบุหรี่เพิ่มความเสี่ยงต่อโรคหลอดเลือดสมอง สำหรับผู้สูบบุหรี่และผู้ไม่สูบบุหรี่ที่ได้รับควันบุหรี่มือสอง การเลิกใช้ยาสูบช่วยลดความเสี่ยงของโรคหลอดเลือดสมอง

1.6.4 การควบคุมโรคเบาหวาน การรับประทานอาหาร การออกกำลังกาย และการลดน้ำหนักสามารถช่วยให้คุณรักษาระดับน้ำตาลในเลือดให้อยู่ในเกณฑ์ที่ดีได้ หากปัจจัยด้านไลฟ์สไตล์ไม่เพียงพอในการควบคุมโรคเบาหวาน แพทย์ของคุณอาจสั่งยารักษาโรคเบาหวาน

1.6.5 ควบคุมน้ำหนักเกิน การมีน้ำหนักเกินมีส่วนทำให้เกิดปัจจัยเสี่ยงอื่นๆ ของโรคหลอดเลือดสมอง เช่น ความดันโลหิตสูง โรคหัวใจและหลอดเลือด และโรคเบาหวาน

1.6.6 การรับประทานอาหารที่อุดมไปด้วยผักและผลไม้ การรับประทานอาหารที่มีผลไม้หรือผักอย่างน้อย 5 มื้อต่อวันอาจช่วยลดความเสี่ยงของโรคหลอดเลือดสมองได้ อาหารเมดิเตอร์เรเนียนที่เน้นน้ำมันมะกอก ผลไม้ ถั่ว ผัก และธัญพืชไม่ขัดสี อาจมีประโยชน์

1.6.7 ออกกำลังกายสม่ำเสมอ. การออกกำลังกายแบบแอโรบิกช่วยลดความเสี่ยงของโรคหลอดเลือดสมองได้หลายวิธี การออกกำลังกายสามารถลดความดันโลหิต เพิ่มระดับคอเลสเตอรอลชนิดดี และปรับปรุงสุขภาพโดยรวมของหลอดเลือดและหัวใจ ยังช่วยให้คุณลดน้ำหนัก ควบคุมเบาหวาน และลดความเครียด ค่อยๆ ออกกำลังกายในระดับปานกลางอย่างน้อย 30 นาที เช่น การเดิน วิ่งจ็อกกิ้ง ว่ายน้ำ หรือปั่นจักรยาน โดยออกกำลังกายอย่างน้อยสามถึงสี่วันต่อสัปดาห์

1.6.8 การดื่มแอลกอฮอล์ในปริมาณที่พอเหมาะ การดื่มแอลกอฮอล์ในปริมาณมากจะเพิ่มความเสี่ยงต่อความดันโลหิตสูง โรคหลอดเลือดสมองตีบ และโรคหลอดเลือดสมองตีบ แอลกอฮอล์อาจมีปฏิกิริยากับยาอื่นๆ ที่คุณกำลังใช้อยู่ อย่างไรก็ตาม การดื่มแอลกอฮอล์ในปริมาณเล็กน้อยถึงปานกลาง เช่น ดื่มวันละ 1 แก้ว อาจช่วยป้องกันโรคหลอดเลือดสมองตีบ และลดแนวโน้มการแข็งตัวของเลือดได้ พูดคุยกับแพทย์ของคุณเกี่ยวกับสิ่งที่เหมาะสมสำหรับคุณ

1.6.9 การรักษาภาวะหยุดหายใจในขณะหลับ (Obstructive Sleep Apnea: OSA) แพทย์ของคุณอาจแนะนำการศึกษาเรื่องการนอนหลับหากคุณมีอาการของ OSA ซึ่งเป็นความผิดปกติของการนอนหลับที่ทำให้คุณหยุดหายใจในช่วงเวลาสั้น ๆ ซ้ำ ๆ ระหว่างการนอนหลับ การรักษา OSA รวมถึงอุปกรณ์ที่ส่งแรงดันบวกของทางเดินหายใจผ่านหน้ากากเพื่อให้ทางเดินหายใจเปิดในขณะที่คุณหลับ

1.6.10 หลีกเลี่ยงยาเสพติดที่ผิดกฎหมาย ยาข้างถนนบางชนิด เช่น โคเคนและยาบ้า ถือเป็นปัจจัยเสี่ยงสำหรับ TIA หรือโรคหลอดเลือดสมอง

1.7 ยาป้องกัน (Preventive medications)

หากคุณเคยเป็นโรคหลอดเลือดสมองตีบหรือ TIA แพทย์ของคุณอาจแนะนำให้ใช้ยาเพื่อช่วยลดความเสี่ยงที่จะเป็นโรคหลอดเลือดสมองอีก ซึ่งรวมถึง:

1.7.1 ยาต้านเกล็ดเลือด (Anti-platelet drugs)

เกล็ดเลือดเป็นเซลล์ในเลือดที่เป็นก้อน ยาต้านเกล็ดเลือดทำให้เซลล์เหล่านี้มีความเหนียวน้อยลงและมีโอกาสเป็นก้อนน้อยลง ยาต้านเกล็ดเลือดที่ใช้กันมากที่สุดคือแอสไพริน แพทย์ของคุณสามารถช่วยคุณกำหนดขนาดยาแอสไพรินที่เหมาะสมสำหรับคุณได้

หลังจาก TIA หรือโรคหลอดเลือดสมองตีบเล็กน้อย แพทย์ของคุณอาจให้แอสไพรินและยาต้านเกล็ดเลือดเช่น clopidogrel (Plavix) แก่คุณเป็นระยะเวลาหนึ่งเพื่อลดความเสี่ยงของโรคหลอดเลือดสมองอื่น หากคุณไม่สามารถรับประทานแอสไพรินได้ แพทย์อาจสั่งจ่ายยาโคลพิโดเกรลเพียงอย่างเดียว

1.7.2 ยาต้านการแข็งตัวของเลือด (Anticoagulants)

ยาเหล่านี้ลดการแข็งตัวของเลือดมี ยาเฮปารินออกฤทธิ์เร็วและอาจใช้ในระยะสั้นในโรงพยาบาล ยาวาร์ฟารินที่ออกฤทธิ์ช้า (Jantoven) อาจใช้ในระยะยาว ยาวาร์ฟารินเป็นยาที่ทำให้เลือดบางลงอย่างมีประสิทธิภาพ ดังนั้น คุณจะต้องใช้ยาตรงตามที่กำหนดและคอยดูผลข้างเคียง คุณจะต้องตรวจเลือดเป็นประจำเพื่อติดตามผลของวาร์ฟาริน

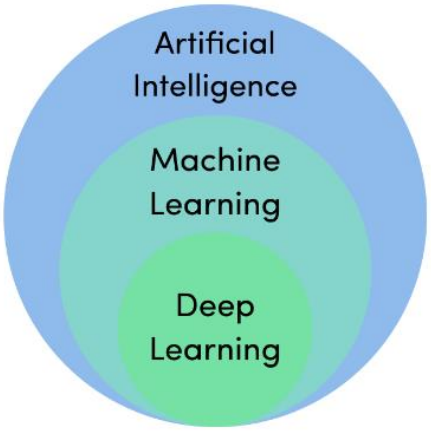
ยาต้านการแข็งตัวของเลือด (Anticoagulants) รุ่นใหม่หลายชนิดมีไว้เพื่อป้องกันโรคหลอดเลือดสมองในผู้ที่มีความเสี่ยงสูง ยาเหล่านี้รวมถึง dabigatran (Pradaxa), rivaroxaban (Xarelto), apixaban (Eliquis) และ edoxaban (Savaysa) พวกมันออกฤทธิ์สั้นกว่าวาร์ฟารินและมักจะไม่ต้องตรวจเลือดหรือตรวจจากแพทย์เป็นประจำ ยาเหล่านี้สัมพันธ์กับความเสี่ยงที่ลดลงของภาวะแทรกซ้อนจากเลือดออกเมื่อเปรียบเทียบกับวาร์ฟาริน

2. การเรียนรู้ของเครื่อง

ส่วนของการเรียนรู้ของเครื่อง ถูกใช้งานเสมือนเป็นสมองของปัญญาประดิษฐ์ เราอาจพูดได้ว่าปัญญาประดิษฐ์ ใช้การเรียนรู้ของเครื่อง ในการสร้างความฉลาด มักจะใช้เรียกโมเดลที่เกิดจากการเรียนรู้ของปัญญาประดิษฐ์ ไม่ได้เกิดจากการเขียนโดยใช้มนุษย์ มนุษย์มีหน้าที่เขียนโปรแกรมให้ปัญญาประดิษฐ์ เรียนรู้จากข้อมูลเท่านั้น ที่เหลือเครื่องจัดการเอง

การเรียนรู้ของเครื่อง เรียนรู้จากสิ่งที่เราส่งเข้าไปกระตุ้น แล้วจดจำเอาไว้เป็นมันสมอง ส่งผลลัพธ์ออกมาเป็นตัวเลข หรือ code ที่ส่งต่อไปแสดงผล หรือให้ตัวปัญญาประดิษฐ์ นำไปแสดงการกระทำ การเรียนรู้ของเครื่องเองสามารถเอาไปใช้งานได้หลายรูปแบบ ต้องอาศัยกลไกที่เป็นโปรแกรม หรือเรียกว่า อัลกอริทึม ที่มีหลากหลายแบบ โดยมีนักวิทยาการข้อมูล (Data Scientist) เป็นผู้ออกแบบ หนึ่งในอัลกอริทึม ที่ได้รับความนิยมสูงคือ การเรียนรู้เชิงลึกซึ่งถูกออกแบบมาให้ใช้

งานได้ง่าย และประยุกต์ใช้ได้หลายลักษณะงาน อย่างไรก็ตาม ในการทำงานจริงนักวิทยาการข้อมูล จำเป็นต้องออกแบบตัวแปรต่างๆ ทั้งในตัวของการเรียนรู้เชิงลึกเอง และต้องหา อัลกอริทึมอื่นๆ มาเป็นคู่เปรียบเทียบ เพื่อมองหาอัลกอริทึม ที่เหมาะสมที่สุดในการใช้งานจริง



ภาพประกอบ 2 แสดงความสัมพันธ์ของการเรียนรู้ของเครื่องเป็นประเภทย่อยของปัญญาประดิษฐ์ และการเรียนรู้เชิงลึกเป็นส่วนย่อยที่มีซับซ้อนของการเรียนรู้ของเครื่อง ตามลำดับ

จากภาพประกอบ 2 แสดงความสัมพันธ์ของปัญญาประดิษฐ์ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก โดยการเรียนรู้ของเครื่องเปรียบเสมือนหน่วยย่อยของปัญญาประดิษฐ์ แล้วการเรียนรู้เชิงลึกเองก็เป็นเสมือนวิธีการหนึ่งของการเรียนรู้ของเครื่องเช่นกัน เพียงแต่เป็นวิธีคิดของคอมพิวเตอร์ที่มีประสิทธิภาพสูงยิ่งกว่า โดยการทำงานจะช่วยเพิ่มประสิทธิภาพในการทำนายรวมไปถึงช่วยเพิ่มประสิทธิภาพของความถูกต้อง ข้อดีของการเรียนรู้เชิงลึก คือยิ่งมีข้อมูลเข้ามาให้ฝึกหัดมากเท่าไหร่ประสิทธิภาพในการคิดของคอมพิวเตอร์ก็ยิ่งสูงขึ้นเท่านั้น ต่างกับการเรียนรู้ของเครื่อง ที่เมื่อมีข้อมูลจำนวนมากประสิทธิภาพการทำงานจะไม่สามารถสูงขึ้นอีกจนกว่าจะได้รับการฝึกหัดเพิ่มเติม

ประเภทของปัญหาการเรียนรู้ของเครื่อง

ปัญหาการเรียนรู้ของเครื่อง สามารถจำแนกได้ 3 ประเภทใหญ่ๆ คือ

1.Supervised learning: คือปัญหาที่ในชุดข้อมูลนั้นมีคำตอบอยู่แล้ว โดยการแก้ปัญหานั้นอัลกอริทึมจำเป็นต้องใช้ ข้อมูลในส่วนสำหรับ train (training data) และส่วนที่รับกลับมาเพื่อปรับปรุง (feedback) จากมนุษย์เพื่อที่จะเรียนรู้ความสัมพันธ์ระหว่างข้อมูลที่ถูกป้อนเข้ามาสู่

ข้อมูลที่ออกไป ยกตัวอย่างเช่น การสร้างโมเดลทำนายภาพ หมากับแมว เราต้องรวมรวบข้อมูลรูปภาพของหมากับแมวจำนวนมากและให้คนมาดูว่าอันไหนเป็น หมาหรือแมว บ้าง แล้วนำมาสร้างโมเดลแยกแยะว่าอันไหนเป็นหมา อันไหนเป็นแมว จากข้อมูลเหล่านี้
เราสามารถใช้ supervised learning เมื่อผลลัพธ์ของข้อมูลเป็นสิ่งที่รู้อยู่แล้ว อัลกอริทึมนี้ก็จะทำนายข้อมูลใหม่ได้ โดยประเภทของ supervised learning มีอยู่ 2 ประเภทคือ

1.1 Regression: หาคำตอบที่เป็นตัวเลข เช่น ให้ข้อมูลการออกกำลังกายและการทานอาหารของเด็กคนหนึ่ง พยากรณ์ว่าเด็กคนนี้จะมีความสูงกี่เซนติเมตรในอีก 10 ปีข้างหน้า

1.2 Classification: หาคำตอบที่เป็นหมวดหมู่ เช่น ให้ภาพ mammogram เต้านม พยากรณ์ว่าคนไข้เป็นมะเร็งเต้านมหรือไม่ (คำตอบมีแต่ ใช่/ไม่ใช่ เรียกว่า Binary classification) หรือให้ข้อมูลเกี่ยวกับครอบครัวของเด็กคนหนึ่ง ทำนายว่าเด็กคนนี้จัดอยู่ในกลุ่มพัฒนาการดี / พัฒนาการปกติ / พัฒนาการช้า (คำตอบมีหลายกลุ่ม เรียกว่า Multiclass classification)

2. Unsupervised learning: คือปัญหาที่ยังไม่มีใครรู้ว่าคำตอบที่ถูกต้องคืออะไร โดยเป็นการเรียนรู้ที่ให้เครื่องจักรนั้นสามารถเรียนรู้ได้ด้วยตนเอง โดยไม่ต้องมีค่าเป้าหมายของแต่ละข้อมูล ซึ่งวิธีการคือมนุษย์จะเป็นผู้ใส่ข้อมูลต่าง ๆ และกำหนดสิ่งที่ต้องการจากข้อมูลเหล่านั้น โดยให้เครื่องจักรวิเคราะห์จากการจำแนกและสร้างแบบแผนจากข้อมูลที่ได้รับมา โดยตัวอย่างที่เห็นได้ชัดของ Machine Learning ในกลุ่ม Unsupervised Learning ที่ถูกนำมาประยุกต์ใช้งานในเชิงธุรกิจ คือ ระบบแนะนำผลิตภัณฑ์ ยกตัวอย่างเช่นการแนะนำคลิปวิดีโอใน YouTube ที่ทำการแบ่งหมวดหมู่ของคลิปวีดีโอต่าง ๆ เป็นต้น ดังนั้น Unsupervised Learning เป็นการสร้างโมเดลโดยใช้ข้อมูล input เพียงอย่างเดียว ไม่มี target ซึ่งการใช้งานหลักมี 2 อย่างคือ

2.1 Clustering: จัดกลุ่มข้อมูล เช่น มีข้อมูลผู้บริจาค ต้องการจัดกลุ่มผู้บริจาคเป็น 3 กลุ่ม เพื่อออกแบบกลยุทธ์การสื่อสารกับลูกค้าแต่ละกลุ่มที่ต่างกันออกไป

2.2 Non-clustering: ปัญหาอื่นๆ ที่ไม่ใช่การจัดกลุ่ม เช่น ตรวจหาข้อมูลชิ้นที่มีความผิดปกติจากพวก (Anomaly detection) แนะนำข้อมูลที่ผู้ใช้น่าจะสนใจ (Recommendation system) เป็นต้น

3. Reinforcement learning: คือการเรียนรู้ที่มีกลไกการเสริมแรงเพื่อให้คอมพิวเตอร์มีพฤติกรรมที่เราต้องการ

Reinforcement Learning เป็น Machine Learning Algorithm แบบหนึ่ง ซึ่งประกอบด้วยองค์ประกอบหลัก ดังต่อไปนี้

Agent – ผู้กระทำ Action

Action – การกระทำของ Agent ที่ส่งผลบางอย่างต่อ Environment

Environment – ระบบที่ Agent ต้องมีปฏิสัมพันธ์ด้วย

State – สถานการณ์ของ Environment ที่ทาง Agent สามารถรับรู้ได้

Policy – หลักการที่ Agent ใช้ในการตัดสินใจเลือก Action หลังจากประเมินสถานการณ์แล้ว

Reward – ตัวประเมินผลลัพธ์ที่เกิดจากการกระทำของ Agent เช่น คะแนน กำไรที่ได้รับหรือ ผลแพ้ชนะ เป็นต้น

โดยมีลักษณะหรือหลักการทำงานที่เหมือนกับตัดสินใจหรือการเรียนรู้ของมนุษย์ นั่นคือเป็นการเรียนรู้จากการลองผิดลองถูกและมีการเรียนรู้เกิดขึ้นระหว่างทางว่าการกระทำไหนดีหรือไม่ดี และพยายามเลือกเอาแนวทางที่ให้ผลผลลัพธ์ที่ดีที่สุดในการรับมือกับปัญหานั้นๆ ให้ดีที่สุด โดยการเรียนรู้เกิดมาจากการปฏิสัมพันธ์ (interaction) ระหว่างผู้เรียนรู้ (agent) กับสิ่งแวดล้อม (environment) โดย ผู้เรียนรู้ จะสามารถรับรู้สถานการณ์ของ สิ่งแวดล้อม รวมทั้งเรียนรู้ผ่านข้อผิดพลาดในอดีตที่เกิดขึ้นผ่าน State และเลือกการกระทำ Action ที่ส่งผลต่อ Environment โดยหวังว่าจะได้ผลลัพธ์ที่ให้ผลตอบแทน Reward ที่สูงที่สุด โดยแนวทางที่ดีที่สุดที่ผู้เรียนรู้ตัดสินใจในการเลือกทำนั้นๆก็คือ policy นั้นเอง ทตัวอย่างเช่น หลักการของ Reinforcement Learning ที่คล้ายกับการเรียนรู้ของมนุษย์เมื่อเปรียบเทียบกับกรณีที่นักศึกษาจะต้องอ่านหนังสือเตรียมตัวสอบ โดยจะแทนค่าผู้เรียนรู้คือ นักศึกษา (Agent) จะรับรู้ถึงสถานการณ์ (State) ของสิ่งรอบตัว (Environment) เช่น เด็กชายเอ ที่เป็นนักเรียนที่กำลังจะเตรียมตัวสอบ เขารับรู้ว่าสถานการณ์คะแนนเก็บของตนในปัจจุบันนั้นมีอยู่เท่าไหร่ โดยดูจากผลการสอบครั้งที่ผ่านๆมา แต่นักเด็กชายเอก็เลือกที่จะไม่อ่านหนังสือสอบ (Action) ทำให้ส่งผลถึงคะแนนการสอบคือเด็กชายเอได้คะแนนน้อย ซึ่งเป็นผลลัพธ์ที่ไม่ดี เพราะฉะนั้นทำให้เด็กชายเอเรียนรู้จากการกระทำจากครั้งก่อนของเขาว่า การไม่อ่านหนังสือก่อนการสอบจะส่งผลทำให้ผลลัพธ์ (Reward) ที่ไม่ดี เด็กชายเอจึงเกิดการเรียนรู้ว่าในการสอบครั้งต่อไป เขาต้องตั้งใจอ่านหนังสือเตรียมสอบเพื่อที่จะทำให้ได้ผลลัพธ์ที่ดีขึ้นเป็นต้น

การเรียนรู้เชิงลึก

การเรียนรู้เชิงลึกคือ การเรียนรู้เชิงลึก เป็นส่วนหนึ่งของวิธีการการเรียนรู้ของเครื่องบนพื้นฐานของโครงข่ายปราสาทเทียมและการเรียนเชิงคุณลักษณะ การเรียนรู้สามารถเป็นได้ทั้งแบบการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอน คำว่า "ลึก"ในความหมายมาจากการที่มีชั้นของโครงข่ายหลายชั้น ที่มีประสิทธิภาพมากขึ้น การเรียนที่สะดวกขึ้น และการเข้าใจในโครงสร้างที่ชัดเจนขึ้น

พื้นฐานของการเรียนรู้เชิงลึกคือ อัลกอริทึมที่พยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบไปด้วยโครงสร้างย่อย ๆ หลายอัน และแต่ละอันนั้นได้มาจากการแปลงที่ไม่เป็นเชิงเส้น การเรียนรู้เชิงลึกอาจมองได้ว่าเป็นวิธีการหนึ่งของการเรียนรู้ของเครื่องที่พยายามเรียนรู้วิธีการแทนข้อมูลอย่างมีประสิทธิภาพ ตัวอย่างเช่น รูปภาพภาพหนึ่ง สามารถแทนได้เป็นเวกเตอร์ของความสว่างต่อจุดพิกเซล หรือมองในระดับสูงขึ้นเป็นเซ็ตของขอบของวัตถุต่างๆ หรือมองว่าเป็นพื้นที่ของรูปร่างใด ๆ ก็ได้ การแทนความหมายดังกล่าวจะทำให้การเรียนรู้ที่จะทำงานต่าง ๆ ทำได้ง่ายขึ้น ไม่ว่าจะเป็นการรู้จำใบหน้าหรือการรู้จำการแสดงออกทางสีหน้า การเรียนรู้เชิงลึกถือว่าเป็นวิธีการที่มีศักยภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้แบบไม่มีผู้สอนหรือการเรียนรู้แบบกึ่งมีผู้สอน

นักวิจัยในสาขานี้พยายามจะหาวิธีการที่ดีขึ้นในการแทนข้อมูลแล้วสร้างแบบจำลองเพื่อเรียนรู้จากตัวแทนของข้อมูลเหล่านี้ในระดับใหญ่ บางวิธีการก็ได้แรงบันดาลใจมาจากสาขาประสาทวิทยาขั้นสูง โดยเฉพาะเรื่องกระบวนการตีความหมายในกระบวนการประมวลผลข้อมูลในสมอง ตัวอย่างของกระบวนการที่การเรียนรู้เชิงลึกนำไปใช้ได้แก่ การเข้ารหัสประสาท อันเป็นกระบวนการหาความสัมพันธ์ระหว่างตัวกระตุ้นกับการตอบสนองของเซลล์ประสาทในสมอง นักวิจัยด้านการเรียนรู้ของเครื่องได้เสนอสถาปัตยกรรมการเรียนรู้หลายแบบบนหลักการของการเรียนรู้เชิงลึกนี้ ได้แก่ โครงข่ายประสาทเทียมแบบลึก (Deep Artificial Neural Networks) โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks) โครงข่ายความเชื่อแบบลึก (Deep Belief Networks) และโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) ซึ่งมีการนำมาใช้งานอย่างแพร่หลายในทางคอมพิวเตอร์วิทัศน์ การรู้จำเสียงพูด การประมวลผลภาษาธรรมชาติ การรู้จำเสียง และชีวสารสนเทศศาสตร์

อัลกอริทึมการเรียนรู้ของเครื่องสำหรับการศึกษานี้

จากการทบทวนวรรณกรรมเราพบว่ามีหลายอัลกอริทึมของการเรียนรู้ของเครื่องที่ใช้ในการสร้างโมเดลประเภทต่างๆ ในผู้ป่วยโรคหลอดเลือดสมอง และเราพบว่ามีหลายอัลกอริทึม ที่ให้

ประสิทธิภาพที่ดีในการทำนายในโรคหลอดเลือดสมองและงานทางการแพทย์อื่นๆ ดังแสดงในการทบทวนวรรณกรรมข้อ 6.2 และ 6.3 ดังนั้นข้อมูลจากการทบทวนวรรณกรรมเหล่านี้ เราจึงใช้เป็นข้อมูลในการตัดสินใจเลือกใช้ 3 อัลกอริทึมคือ Logistic regression, Support Vector Machine , Random forest ในการสร้างโมเดลในการทำนายเพราะเป็นอัลกอริทึมที่มีประสิทธิภาพดีในการทำนายในทางการแพทย์

3. อัลกอริทึมการเรียนรู้ของเครื่องสำหรับการแบ่งประเภทของข้อมูล มีดังนี้

อัลกอริทึมที่ใช้ในการแบ่งประเภทข้อมูลของผู้ป่วยที่เสี่ยงต่อการเกิดโรคหลอดเลือดสมอง

3.1 **การวิเคราะห์การถดถอยโลจิสติค** (Peeradon Samasiri, Scientist, & (GBDi), 2021)

$$P(y = 1|x, \beta) = \frac{e^{f(x,\beta)}}{1 + e^{f(x,\beta)}} \quad (1)$$

สมการคำนวณค่าความน่าจะเป็นที่จะเป็น 1

เมื่อ x คือเซ็ตของตัวแปรที่ส่งผลต่อการทำนาย,

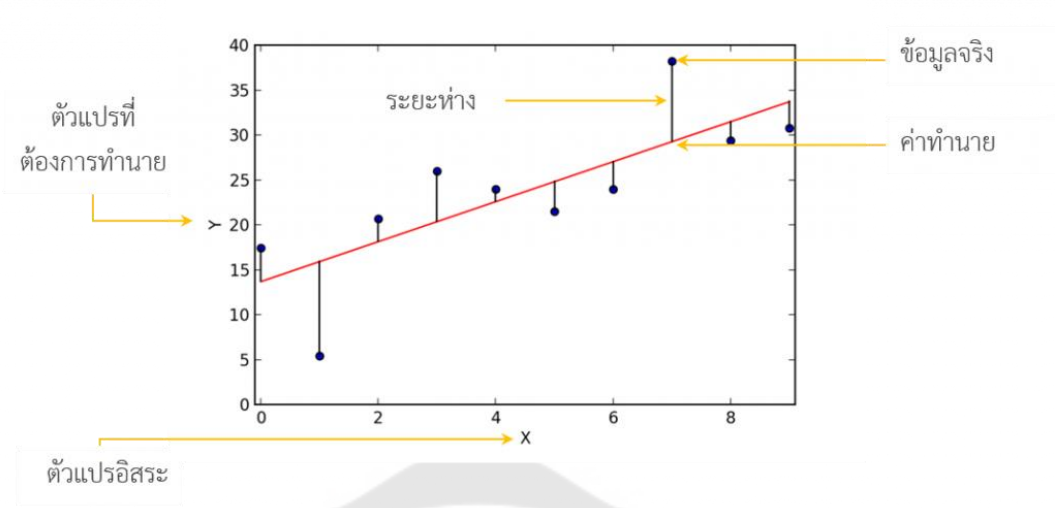
y คือแปรตาม(ผลการทำนาย)

β คือเซ็ตของพารามิเตอร์ของโมเดลทำนาย

$f(x)$ เป็นฟังก์ชันเชิงเส้นในรูปของ $f(x) = \beta 0 + \beta 1X1 + \beta 1X2 + \cdots \beta nXn$

และ $P(y = 1|x, \beta)$ คือค่าความน่าจะเป็นที่ค่าตัวแปรนั้นจะถูกทำนายให้เป็น 1

ในขณะที่ผลจากการทำนายด้วย Linear Regression สามารถเขียนแทนได้ด้วยเส้นตรง (สำหรับกรณีทำนายผลจากหนึ่งตัวแปร) ผลรวมระยะห่างระหว่างค่าทำนายกับค่าของข้อมูลจริงบ่งชี้ความแม่นยำของการทำนายดังกล่าว เพื่อหาเส้นตรงที่เหมาะสม เราอาจสร้างเส้นตรงหลาย ๆ เส้นก่อนจะเฟ้นหาเส้นตรงที่ให้ผลรวมระยะห่างที่ต่ำที่สุด ซึ่งจะถือเป็นโมเดลที่ดีที่สุดที่ใช้อธิบายข้อมูลชุดนั้น ๆ ดังแสดงในภาพประกอบ 3



ภาพประกอบ 3 แสดงการฟิตเส้นตรงสำหรับการทำ Linear Regression นั้นวัดค่าความแม่นยำจากผลรวมของระยะห่างระหว่างข้อมูลจริงกับค่าที่ทำนายจากโมเดล

ที่มา : (Samasiri, 2021 #61)

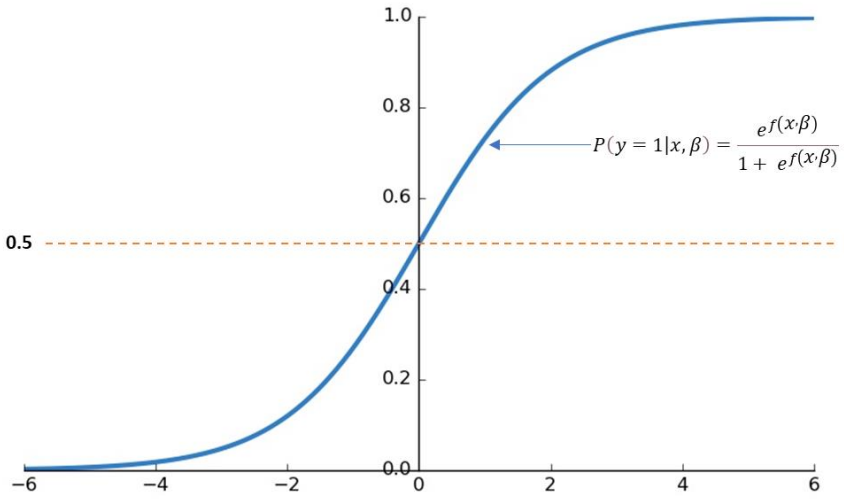
อย่างไรก็ดีเมื่อใช้หลักการดังกล่าวกับการทำนายด้วย Logistic Regression พบว่าค่าผลรวมที่ได้จากการเทียบระยะห่างระหว่างเส้นโค้งซิกมอยด์กับข้อมูลจริงสามารถหาจุดที่ต่ำที่สุดได้ยาก ในทางปฏิบัติเราจะคำนวณหาค่าผลคูณความน่าจะเป็น ซึ่งสะท้อนความศักยภาพของโมเดลในการฟิตเข้ากับตัวข้อมูล (Likelihood) แทนผลรวมระยะห่าง โมเดลเส้นโค้งซิกมอยด์ที่ให้ค่าความเป็นไปได้สูงสุด (Maximum Likelihood) จะถูกเลือกให้เป็นโมเดลที่ดีที่สุดที่ใช้อธิบายข้อมูลชุดนั้น ในทางปฏิบัติเรานิยมคำนวณค่าลอการิทึมของผลคูณความน่าจะเป็น (Log Likelihood) โดยสมการผลคูณความน่าจะเป็นจะถูกเปลี่ยนให้อยู่ในรูปของผลรวม ซึ่งสามารถคำนวณหาค่าสูงสุดผ่านการหาอนุพันธ์ (Differentiation) ได้ง่ายกว่าการหาอนุพันธ์ของผลคูณ ผลบวกดังกล่าวอยู่ในรูปของ

$$LL = log(Likelihood) = \sum_i \begin{cases} log\left(\frac{e^{f(x_i)}}{1+e^{f(x)}}\right), & y_i = 1 \\ log\left(1 - \frac{e^{f(x_i)}}{1+e^{f(x_i)}}\right), & y_i = 0 \end{cases} \quad (2)$$

สมการคำนวณ Log Likelihood ของโมเดล

แทน โดยพจน์แรกว่าด้วยความเป็นไปได้ที่จุด i ใด ๆ ที่ทราบมาก่อน (labelled) ว่าเป็น 1 นั้น จะถูกโมเดลทำนายว่าเป็น 1 ในขณะที่พจน์ที่สองว่าด้วยความเป็นไปได้ที่จุด i ใดๆ ที่ทราบมาก่อน

ว่าเป็น 0 นั้นจะถูกโมเดลทำนายว่าเป็น 0 ด้วยเหตุนี้ผลรวมของสองพจน์จึงเป็นผลรวมความเป็นไปได้สำหรับกรณีที่โมเดลทายถูก (1 หรือ 0) ทั้งหมด โมเดลที่มีค่า β ที่สร้างความเป็นไปได้ที่สูงที่สุด (Maximum Likelihood) จึงถูกเลือกให้เป็นโมเดลที่เป็นตัวแทนของข้อมูลชุดนั้น



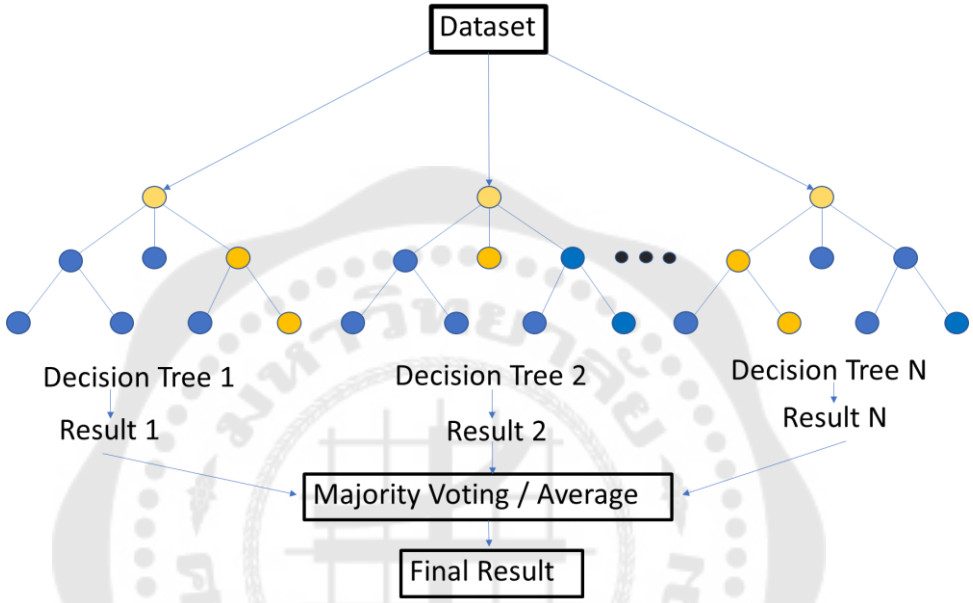
ภาพประกอบ 4 แสดง Sigmoid เป็นฟังก์ชัน activation function ซึ่งอยู่ในเส้นโค้งรูปร่าง S

จากภาพประกอบ 4 แสดง Sigmoid เป็นฟังก์ชัน activation function ซึ่งอยู่ในเส้นโค้งรูปร่าง S โดยเส้นโมเดล (เส้นสีน้ำเงิน) ที่จะใช้แยกแยะเปรียบเทียบกับข้อมูลจริงที่ได้จากการใช้ฟังก์ชัน $f(x,\beta)$ ที่แปลงค่าตัวแปรอิสระหลายตัวให้สามารถนำมาทำนายได้บนแกนนอน โดยแกนตั้งแสดงค่าความน่าจะเป็นที่ $P(y=1|x,\beta)$ โมเดลจะคำนวณค่า $P(y=1|x,\beta)$ จากค่าตัวแปรอิสระจากข้อมูลหนึ่งๆ ที่มีค่าอยู่ระหว่าง 0 ถึง 1

3.2 ป่าสุ่ม

Random forest: RF เกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ ซึ่งค่าความคลาดเคลื่อนโดยรวมของป่าไม้จะถูกเปลี่ยนให้เป็นค่าลิมิต ทำให้จำนวนของต้นไม้ในป่าเพิ่มขึ้น ค่าความคลาดเคลื่อน โดยรวมจะขึ้นกับความมั่นคง (Strength) ของต้นไม้แต่ละต้น รวมถึงความสัมพันธ์กันระหว่างต้นไม้เหล่านั้น โดยจะใช้วิธีการสุ่มเลือกคุณสมบัติเพื่อการแบ่งแยกโหนดทำให้ค่าความผิดพลาดลดลง ขั้นตอนวิธีนี้ จะมีประสิทธิภาพมากเมื่อนำไปใช้วิเคราะห์เกี่ยวกับการประมาณการขนาดใหญ่ เราสามารถสร้างแบบจำลองที่ใช้ต้นไม้หลาย ๆ ต้นในการตัดสินใจเพื่อนำมาประมวลผล ซึ่งมีความแม่นยำสูง สามารถจัดการข้อมูลได้มากและเหมาะสำหรับข้อมูลที่

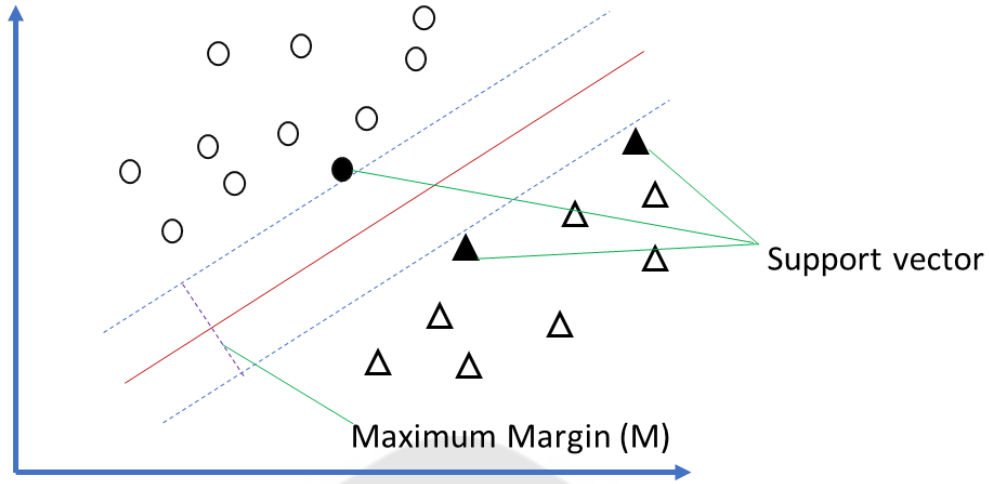
มีความสำคัญ ซึ่งโมเดลป่าสุ่มนี้เป็นหนึ่งในกลุ่มของโมเดลที่เรียกว่า Ensemble learning ที่มีหลักการคือการเทรนโมเดลที่เหมือนกันหลายๆ ครั้ง (หลาย Instance) บนข้อมูลชุดเดียวกัน โดยแต่ละครั้งของการเทรนจะเลือกส่วนของข้อมูลที่เทรนไม่เหมือนกัน แล้วเอาการตัดสินใจของโมเดลเหล่านั้นมาโหวตกันว่า Class ไหนถูกเลือกมากที่สุด (Majority Voting) ดังแสดงในภาพประกอบ 5



ภาพประกอบ 5 แสดงในขั้นตอนการทำงานของป่าสุ่ม จะทำการการจำแนกต้นไม้หลาย ๆ ต้น ซึ่งในต้นไม้ แต่ละต้นมีการแบ่งเป็นคลาส โดยที่ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจแต่ละต้นถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด (Majority Voting)

3.3 ซัพพอร์ตเวกเตอร์แมชชีน

Support Vector Machine: SVM เป็นตัวจำแนกเชิงเส้น (Linear Classifier) แบบ 2 คลาส ซึ่งเป็นที่ยอมรับถึงประสิทธิภาพของการจำแนกที่เหนือว่าวิธีการจำแนกอื่น ๆ ข้อได้เปรียบของ SVM คือมีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมากได้ นอกจากนี้การใช้ฟังก์ชัน คอร์เนล (Kernel Function) เพื่อแปลงข้อมูลไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ (Feature Space) สามารถจำแนกข้อมูลที่มีความคลุมเครือได้อย่างมีประสิทธิภาพ หลักการของ SVM คือการหาเส้นตรงที่มีมาร์จินที่โตที่สุด (Maximum Margin) ที่สามารถแบ่งข้อมูลออกเป็น 2 คลาส ดังตัวอย่างในภาพที่ 9 เป็นข้อมูลขนาด 2 มิติ โดนถูกจำแนกออกเป็น 2 คลาส ได้แก่ + (O) และคลาส – (Δ) โดยเส้นตรงที่ใช้แบ่งข้อมูลมีมาร์จินเท่ากับ M=2w ซึ่ง เป็นความกว้างระหว่างเส้นตรงกับซัพพอร์ตเวกเตอร์ (Support vector) ของข้อมูลทั้ง 2 คลาส (● และ ▲)



ภาพประกอบ 6 อัลกอริทึม SVM คือการหาเส้นตรงที่มีมาร์จินที่โตที่สุด (Maximum Margin) ที่สามารถแบ่งข้อมูลออกเป็น 2 คลาส

ภาพประกอบ 6 การใช้เส้นตรงสำหรับแบ่งข้อมูลเป็น 2 กลุ่มด้วยมาร์จินที่ใหญ่ที่สุด (Maximum Margin) เป็นวิธีที่การันตีได้ว่าจะสามารถแยกข้อมูลได้โดยมีความผิดพลาดน้อยที่สุด โดยมี support vector เป็นตัวกำหนดขนาดของ Margin ดังนั้นถ้าข้อมูลมีการเปลี่ยนแปลงใด ๆ เส้นตรงจำแนกก็ยังขึ้นอยู่กับ support vector ซึ่งจะยังเป็น Maximum Margin อยู่

ในการหา Maximum Margin ในเชิงคณิตศาสตร์ จะเห็นได้ว่าข้อมูล x จะถูกแบ่งเป็นระนาบบวกและระนาบลบ โดยมีสมการคือ $w \cdot x+b >=1$ สำหรับคลาส + และ $w \cdot x+b <=-1$ สำหรับคลาส - ดังนั้นจะสามารถจำแนกข้อมูลได้โดย

$$+1 \text{ ถ้า } w \cdot x + b >= +1 \quad (3)$$

$$-1 \text{ ถ้า } w \cdot x + b <= -1 \quad (4)$$

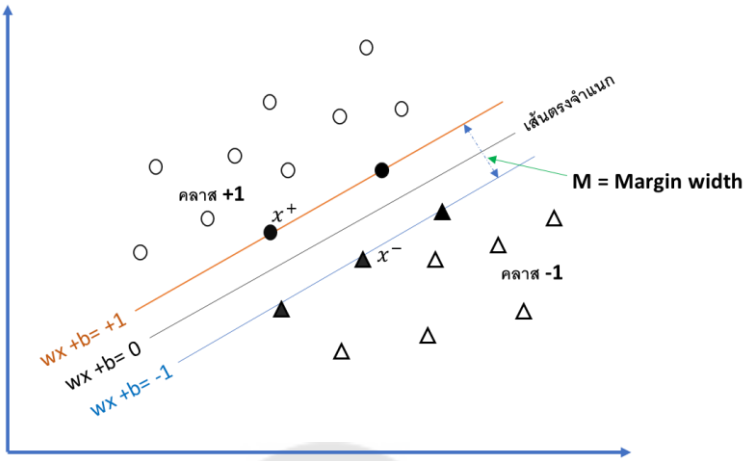
$$\text{ถ้า } -1 < w \cdot x + b < +1 \quad (5)$$

x คือ ตัวแปรต้น (predictor variable)

y คือ ตัวแปรตาม (dependent variable)

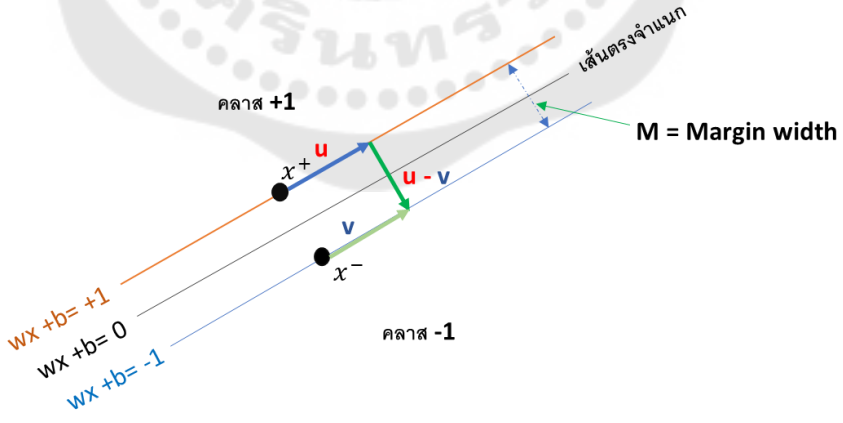
w คือ ความชัน (slope)

b คือ จุดตัดแกน y (y-intercept)



ภาพประกอบ 7 แสดงการจำแนกเชิงเส้นด้วยมาร์จินที่ (margin width) ใหญ่ที่สุด

จากภาพประกอบ 7 เป็นการหาความกว้างมาร์จิน (margin width) กำหนดให้ x- เป็นจุดอยู่บนระนาบลบ และให้ x+ เป็นจุดอยู่บนระนาบบวก และทั้งสองจุดใกล้กันมากที่สุด ดังภาพประกอบ 8 จะมีเวกเตอร์ u บนระนาบบวกและทำนองเดียวกันเวกเตอร์ v บนระนาบลบ และมีเวกเตอร์ w ที่เป็นเวกเตอร์น้ำหนักของเส้นจำแนก ดังนั้นจะเกิดเวกเตอร์ตั้งฉากกันจะได้ (u-v)·w = 0 การคำนวณหาความกว้างของมาร์จินในเทอมของ w และ b สามารถหาได้จากระบบสมการต่อไปนี้



ภาพประกอบ 8 แสดงการคำนวณ การหาค่า Margin

จากภาพประกอบ 8 การคำนวณ การหาค่า Margin มีสมการทางคณิตศาตร์ดังต่อไปนี้

$$w \cdot x+ + b = +1 \quad (6)$$

$$w \cdot x– + b = -1 \quad (7)$$

$$x+ = x– + \delta w \quad (8)$$

$$x+ – x–| = M \quad (9)$$

เมื่อ δ เป็นขนาดของเวกเตอร์ w จะได้ว่า จาก สมการ (8) และ (9)

$$w \cdot (x– + \delta w) + b = +1 \quad (10)$$

$$w \cdot x– + b + \delta w \cdot w = +1 \quad (11)$$

$$-1 + \delta w \cdot w = 1 \quad (12)$$

$$\text{ดังนั้น } \delta = 2/(w.w) \quad (13)$$

จากสมการที่ (12) และ (13) จะได้

$$M = |x+ – x–| = |\delta w| = \delta \text{ sqrt}(w \cdot w) \quad (14)$$

จากสมการที่ (13) และ (14) จะได้

$$M = (2/(w.w))\ \text{sqrt}(w \cdot w) = 2/(\text{sqrt}(w \cdot w)) \quad (15)$$

เนื่องจากเราสามารถทราบค่าของเวกเตอร์ w และ ค่าคงที่ b จะทำให้สามารถคำนวณค่า margin M ได้จากสมการที่ 15 จากนั้นใช้วิธีการค้นหาค่า M ที่มากที่สุดด้วยวิธีต่าง ๆ เช่น Gradient Descent, Simulated Annealing, Newton method, EM เป็นต้น เพื่อให้ได้ค่าคำตอบที่เหมาะสมต่อไป

4. การวิเคราะห์ข้อมูล

ในการศึกษานี้เป็นการวิเคราะห์ข้อมูลแบบพยากรณ์ (Predictive analytics) เป็นการวิเคราะห์เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูล ที่ได้เกิดขึ้นแล้วกับแบบจำลองทางสถิติ หรือ เทคโนโลยีปัญญาประดิษฐ์ต่างๆ เพื่อทำนายโอกาสการเกิดโรคหลอดเลือดสมอง โดยการใช้อัลกอริทึมของการเรียนรู้ของเครื่อง

5. ประเมินประสิทธิภาพของแต่ละโมเดล

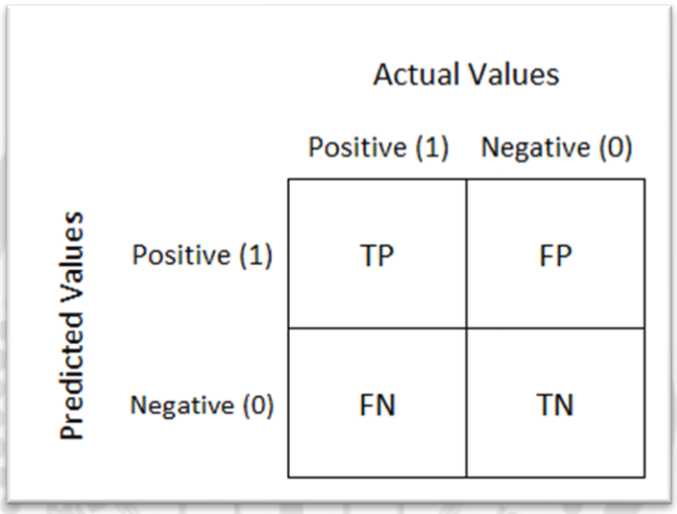
ประเมินประสิทธิภาพของแต่ละโมเดล

การตรวจประเมินประสิทธิภาพของโมเดลเราใช้ Confusion Metrix และการทำ cross validation ดังรายละเอียดต่อไปนี้

5.1 ตารางเมทริกซ์ความสับสน (Confusion Matrix) และการคำนวณหาค่าประสิทธิภาพของโมเดล

ตารางเมทริกซ์ความสับสน คือตารางสำคัญในการวัดความสามารถของ machine learning ในการแก้ปัญหา classification

โดยตาราง ตารางเมทริกซ์ความสับสน คือการประเมินผลลัพธ์การทำนาย (หรือผลลัพธ์จากโปรแกรม) เปรียบเทียบกับผลลัพธ์จริงๆ ที่หาโดยคน ดังแสดงรายละเอียดในภาพประกอบ 9



ภาพประกอบ 9 แสดงตารางตารางเมทริกซ์ความสับสน (Confusion Matrix)

จากภาพประกอบ 9 เป็นตารางเมทริกซ์ความสับสน ขนาด 2x2 มีชื่อเรียก และความหมายของแต่ละช่อง

True Positive (TP) คือ ทายว่าใช่ แล้วใช่จริง ๆ (Hit)

True Negative (TN) คือ ทายว่าไม่ใช่ แล้วไม่ใช่จริง ๆ (Correct Rejection)

False Positive (FP) คือ ทายว่าใช่ แต่จริง ๆ แล้วไม่ใช่ (False Alarm, Type I error)

False Negative (FN) คือ ทายว่าไม่ใช่ แต่จริง ๆ แล้วมันใช่ (Miss, Type II error)

Condition Positive (P) คือ จำนวนของที่ใช่ทั้งหมด ที่อยู่ในข้อมูล = TP + FN

Condition Negative (N) คือ จำนวนของที่ไม่ใช่ทั้งหมด ที่อยู่ในข้อมูล = FP + TN

หมายเหตุ True = ทายถูก, False = ทายผิด, Positive = ทายว่าใช่, Negative = ทายว่าไม่ใช่

ค่าต่างๆที่ใช้ในการประเมินประสิทธิภาพของโมเดล คือ

5.1.1 ค่าความแม่นยำ (Accuracy)

จำนวนครั้งที่ทายถูก หารด้วยจำนวนครั้งที่ทายทั้งหมด หมายความว่าทายแม่นแค่ไหน แบบรวม ๆ

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

5.1.2 ค่าความเที่ยงตรง (Precision)

หมายควายว่าในบรรดาคนที่ตรวจได้ผลบวกทั้งหมด (TP+FN) มีกี่คนที่เป็นโรคจริง (TP)

จำนวนครั้งที่ทายว่า Positive แล้วถูก หารด้วยจำนวนครั้งที่ทายว่า Positive ทั้งหมด

Precision มีอีกชื่อว่า positive predictive value (PPV)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

5.1.3 การทายว่า Positive มีความสำคัญ ตัวอย่างเช่น ผลตรวจผลเลือด เป็น Positive อาจจะแปลว่าเป็นโรค ดังนั้นจึงไม่ควร Positive พร่ำเพรื่อ ถ้าไม่แน่ใจจริง ๆ แต่ทั้งนี้ขึ้นอยู่กับงานด้วย

5.1.4 ค่าระลึกคืน (Recall/Sensitivity)

หมายควายว่าในบรรดาคนที่เป็นโรคจริงๆทั้งหมด (TP+FN) มีกี่คนที่ตรวจได้ผลบวก (TP)

Recall มีอีกหลายชื่อ เช่น sensitivity, recall, hit rate, หรือ true positive rate (TPR)

$$\text{Recall} = \frac{TP}{TP+FN} \quad (18)$$

5.1.5 ค่าความจำเพาะ (Specificity)

หมายควายว่าในบรรดาคนที่ไม่เป็นโรคทั้งหมด (TN+FP) มีกี่คนที่ตรวจได้ผลลบ (TN)

Specificity มีอีกหลายชื่อ เช่น selectivity หรือ true negative rate (TNR)

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (19)$$

5.1.6 ค่าประสิทธิภาพโดยรวม (f1-Score)

f1-Score คือ Harmonic mean ของ Precision และ Recall

$$\text{f1-score} = 2 * \left(\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}\right) \quad (20)$$

ตัววัดประสิทธิภาพของโมเดลการจำแนก ประเภทข้อมูล

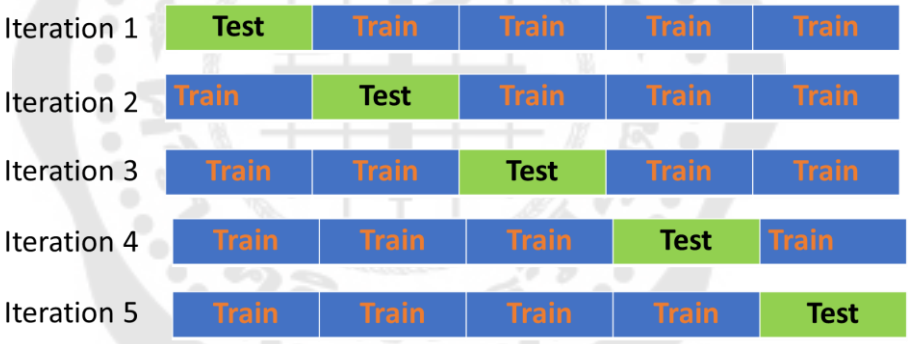
การนำโมเดลไปใช้งานจริงได้นั้น จำเป็นจะต้อง ทราบประสิทธิภาพของโมเดลก่อน โดยทั่วไปแล้วจะมี % ตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่าง ๆ อยู่ 5 ค่า คือ

1. ค่าความแม่นยำ (Precision) คือค่าที่ดูสิ่งที่ ทำนายออกมาแล้วทายถูกได้กี่เปอร์เซ็นต์
2. ค่าความระลึก (Recall) คือจำนวนที่ทำนาย ถูกกี่ตัว เป็นการวัดความถูกต้องของโมเดล

3. ค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) คือค่าเฉลี่ย ของค่าความแม่นยำและค่าความระลึก

4. ค่าความถูกต้อง (Accuracy) คือจำนวน ข้อมูลที่ทำนายถูกทุกคลาส เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส

5. การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพ (Cross validation Test) ของโมเดลการจำแนกประเภทข้อมูลด้วยวิธี Cross validation Test เป็นวิธีที่นิยมใช้ในการทดสอบประสิทธิภาพของ โมเดล เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัด ประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการ แบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือทำการแบ่งข้อมูล ออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือการแบ่งข้อมูล ออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูล เท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบ ประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวน ที่แบ่งไว้

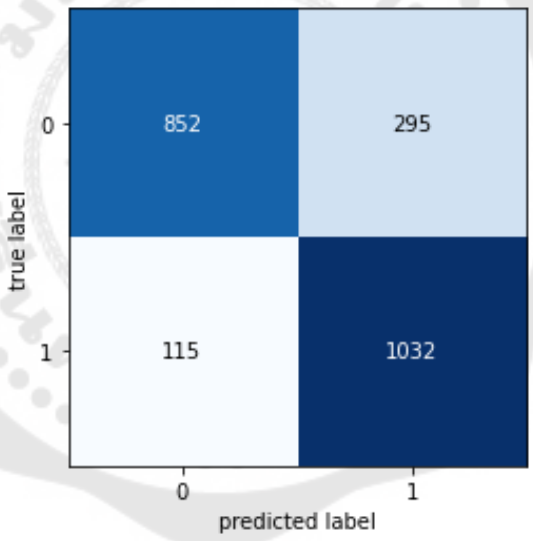


ภาพประกอบ 10 แสดงการแบ่งข้อมูลออกเป็น 5-fold cross-validation โดย แบ่งข้อมูล ออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน

จากภาพประกอบ 10 เป็นการทำ 5-fold cross-validation คือทำการแบ่งข้อมูล ออกเป็น 5 ส่วนแล้วแบ่ง 1 ส่วนเป็น Test set และอีก 4 ส่วนเป็น Train set โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัว Test ประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวน

5.2 การหาค่า Micro, Macro & Weighted Averages of F1 Score

อย่างไรก็ตามในการทดสอบประสิทธิภาพของโมเดลทั้งใน binary classification หรือ multiclassification จะมีค่าที่เราสามารถแสดงประสิทธิภาพของโมเดลได้คือค่า Accuracy, Precision, Recall, และ F1-score ซึ่งค่าเหล่านี้จะมีค่าเฉพาะสำหรับแต่ละคลาสและมีหลายค่าตามจำนวนของคลาสที่มี ดังนั้นจะเป็นสิ่งที่ดีกว่าหากเราสามารถแสดงค่าเหล่านี้เป็นค่าเฉลี่ยโดยรวมเป็นตัวเลขเพียงตัวเดียวเพื่ออธิบายประสิทธิภาพโดยรวมของโมเดล โดยวิธีการหาค่าเฉลี่ยของค่าrecision, Recall, F1-score มีวิธีการคำนวณสามแบบคือ Micro average , Macro average และ Weight average ด้วย Function classification_report ดังตัวอย่างต่อไปนี้ จะแสดงวิธีคำนวณค่า Average f1-score ด้วยวิธีการ Micro average method , Macro average method และ Weight average method



ภาพประกอบ 11 Confusion matrix แสดงตัวอย่างค่าการวัดประสิทธิภาพของโมเดลในการทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง

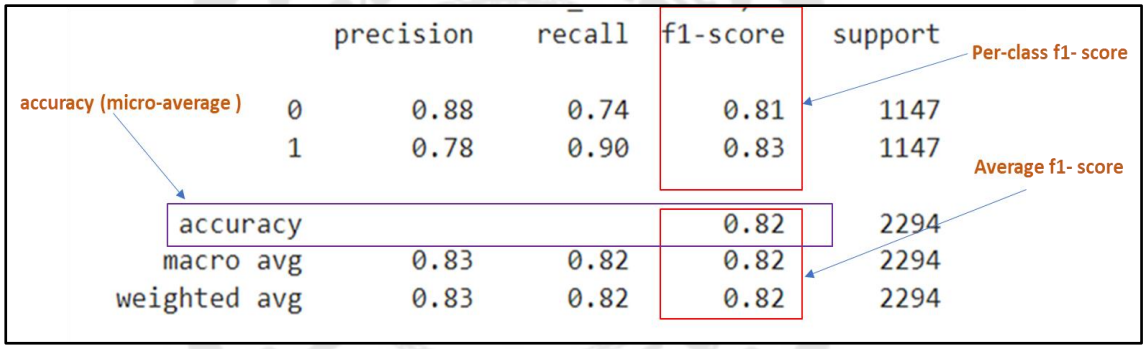
จากภาพประกอบ 11 Confusion matrix แสดงตัวอย่างค่าประสิทธิภาพของโมเดลที่ทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง โดยแสดงค่าประสิทธิภาพต่างๆ ของโมเดลดังนี้ TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 1032

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 852

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 295

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง(stroke) โดยมีค่าเท่ากับ155

โดยค่า TP,TN,FP,FN ในตาราง Confusion matrix จะแทนด้วยค่าความถี่ของจำนวนข้อมูลซึ่งเราสามารถใช้ค่าใน Confusion Matrix มาคำนวน ค่าการประเมินประสิทธิภาพของการทำนายด้วย Model ของเรา ในรูปแบบค่าต่างๆ ได้หลายค่าตามภาพประกอบ 12



ภาพประกอบ 12 ตัวอย่าง classification_report ที่ได้จากการวัดประสิทธิภาพของโมเดล

จากภาพประกอบ12 นอกจาก Precision, Recall, f1-score แล้ว classification_report ยังแสดงจำนวนข้อมูลในแต่ละ Class ด้วยค่า Support และในกรอบสีแดงได้แสดงค่าของ f1-score ทั้ง Per-class f1-score และ Average f1-score และในกรอบสีม่วงได้แสดงค่าของ accuracy (หรือค่า micro-average score) ซึ่งจะได้แสดงให้ทราบถึงรายละเอียดต่อไป

จากข้อมูล Classification_report และ Confusion matrix เราสามารถนำมาหาค่าเฉลี่ยแบบต่างๆ ได้ ไม่ว่าจะเป็นค่าเฉลี่ยของ Precision, Recall และ f1-score แต่สำหรับตัวอย่างนี้เราจะคำนวณค่าเฉลี่ยของ f1- score (average f1-scores) โดยวิธีการหาค่าเฉลี่ยมีสามวิธีที่แตกต่างกันในการคำนวณคือ Macro average method , Weight average method,และ Micro average method ดังตัวอย่างต่อไปนี้

5.2.1 **การหาค่า** average f1-scores **ด้วยวิธี** Macro-average method

ตาราง 1 การคำนวณหาค่า Macro-average f1-score

Label	Per-class f1-score	Macro-Average f1-score
0	0.81	$\frac{0.81+0.83}{2} = 0.82$
1	0.83	

จากตาราง 1 การหาค่าเฉลี่ยมาโคร (Macro average) อาจเป็นวิธีที่ตรงไปตรงมาที่สุดในบรรดาวิธีการหาค่าเฉลี่ยที่มีอยู่มากมาย โดยคะแนน macro-averaged f1-score (หรือคะแนน Macro f1-score) คำนวณโดยใช้ค่าเฉลี่ยเลขคณิต (หรือค่าเฉลี่ยไม่ถ่วงน้ำหนัก / unweighted mean) ของคะแนน f1-score ต่อคลาสทั้งหมด สำหรับวิธีการนี้เป็นการปฏิบัติกับทุกคลาสอย่างเท่าเทียมกันโดยไม่คำนึงถึงค่าจำนวนตัวอย่างของแต่ละคลาส (ไม่สนค่า support values ของแต่ละคลาส) จากการคำนวณค่า 0.82 ที่เราคำนวณข้างต้นตรงกับค่า macro-averaged f1-score ในรายงานการจัดหมวดหมู่ (classification report) ที่แสดงในภาพประกอบ 12

5.2.2 **การหาค่า** average f1-scores **ด้วยวิธี** Weighted-average method

ตาราง 2 การคำนวณหาค่า Weighted-average f1-score

Label	Per-class f1-score	Support	Support proportion	Weighed-Average f1-score
0	0.81	1147	1147 / 2294 = 0.5	$(0.81 * 0.5) + (0.83 * 0.5) = 0.82$
1	0.83	1147	1147 / 2294 = 0.5	

จากตาราง 2 เป็นการคำนวณหาค่า Weighted-average f1-score โดย 'weight' หมายถึงสัดส่วนของแต่ละคลาสที่สัมพันธ์กับผลรวมของค่า support ทั้งหมด และ Support หมายถึงจำนวนที่เกิดขึ้นจริงของแต่ละคลาสในชุดข้อมูล ด้วยการถัวเฉลี่ยถ่วงน้ำหนัก weighted average ค่าเฉลี่ยของผลลัพธ์ที่ได้จะนำมาพิจารณาถึงการมีส่วนร่วมของแต่ละคลาสโดยถ่วงน้ำหนักด้วยจำนวนตัวอย่างของคลาสที่ระบุ แล้วค่าที่เราคำนวณได้เท่ากับ 0.82 นับเป็นคะแนน weighted-averaged ของ f1-score ในรายงาน classification_report การจัดหมวดหมู่นั้นเอง

5.2.3 การหาค่า average f1-scores ด้วยวิธี Micro-average method

ตาราง 3 การหาค่า average f1-scores ด้วยวิธี Micro-average method

Label	True Positives(**TP**)	False Negatives(**FN**),	False Positives (**FP**)	Micro-Average f1-score
0 and 1	1032	115	295	$\frac{\mathrm{TP}}{\mathrm{TP}+0.5(\mathrm{FP}+\mathrm{FN})} = \frac{1032}{1032+0.5(295+115)} = 0.82$

จากตาราง 3 เราจะพบว่าใน classification report คะแนนของ micro f1-score เป็น 0.82 และจะถูกแสดงเป็นค่า 'accuracy' (ดังที่แสดงในภาพประกอบ 12) แล้วเหตุใดจึงไม่มีแถวใน report ที่ระบุว่าเป็น 'micro avg' เหตุผลก็คือค่า micro-average จะคำนวณจากสัดส่วนของการสังเกตค่าที่ทำนายได้ถูกต้อง (proportion of correctly classified) ต่อจำนวนของการสังเกตทั้งหมด หากเราคิดถึงสิ่งนี้ คำจำกัดความนี้คือสิ่งที่เราใช้คำนวณความถูกต้องโดยรวมของระบบนั้นก็คือค่า accuracy นั้นเอง

จากตัวอย่างข้างบนค่า Macro avg, Micro avg และ Weight avg ของ f1-score มีค่าเท่ากัน เนื่องจากจำนวนข้อมูลทั้ง 2 Classes เท่ากัน จัดเป็นข้อมูลแบบ balanced Classes ซึ่งเมื่อข้อมูลมีลักษณะ balanced เราจะดูค่าเฉลี่ยได้ทั้งแบบ Macro avg, Micro avg และ Weight avg แต่ถ้าเราพบว่าค่าของ Macro avg, Macro avg และ Weight avg ของ f1-score มีค่าต่างกัน เนื่องจากจำนวนข้อมูลทั้ง 2 Classes ไม่เท่ากัน (Imbalanced Classes) ซึ่งเมื่อข้อมูลมีลักษณะ imbalanced เราก็จะดูค่าเฉลี่ยแบบ Weight avg เป็นหลัก

6. การทบทวนวรรณกรรม

6.1 Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. (Chun et al., 2021)

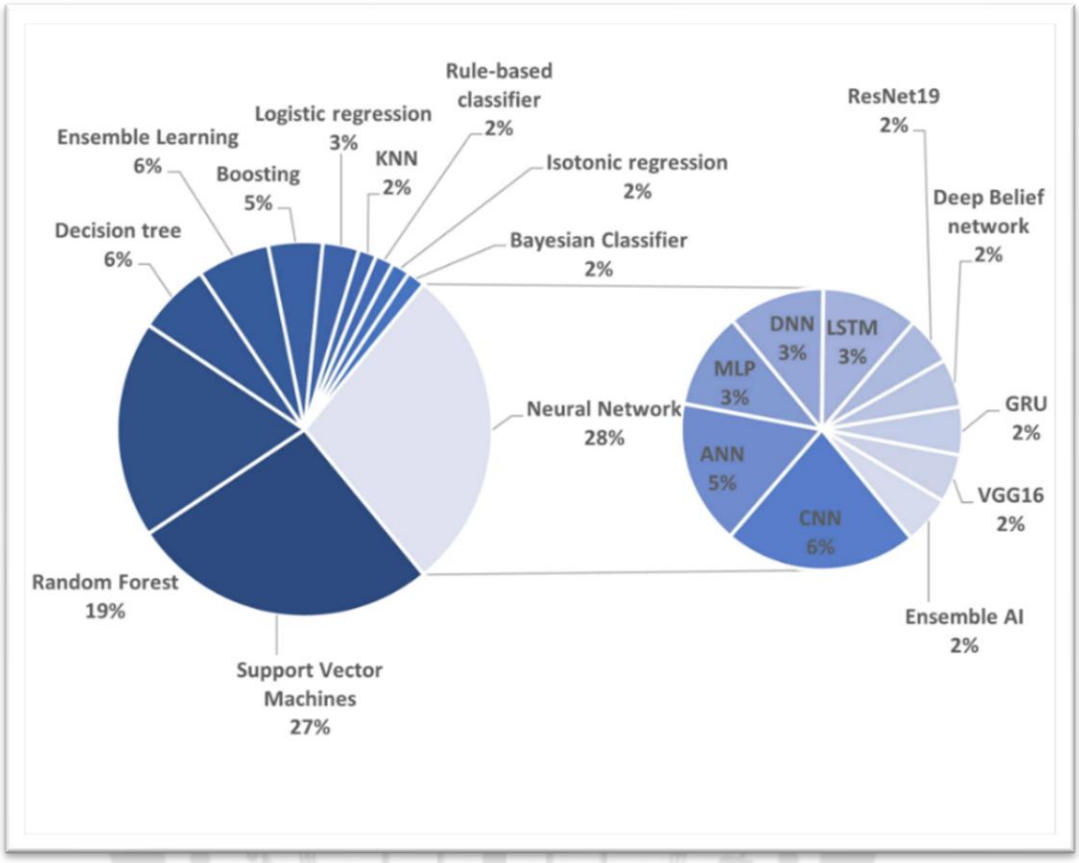
ทำการศึกษาแบบ prospective cohort study คือ เป็นการศึกษาตามรุ่น (cohort study) ที่ติดตามกลุ่มบุคคลที่มีลักษณะคล้าย ๆ กัน (cohort) โดยทำการศึกษากับวัยผู้ใหญ่ชาวจีนจำนวน

503,842 คน ใน 10 พื้นที่จีน ที่ไม่มีประวัติการเกิดโรคหลอดเลือดสมองมาก่อน สำหรับการทำนายความเสี่ยงต่อโรคหลอดเลือดสมองในช่วงเวลา 9 ปี ผลพบว่า GBT(gradient boosted trees) ให้ผลของการแยกแยะการเกิดโรคที่ดีที่สุด (AUROC: 0.833 ในผู้ชายและ 0.836 ในผู้หญิง) และการสอบเทียบ โดยให้ผลลัพธ์ที่สอดคล้องกันในแต่ละช่วงเวลาของการติดตาม ensemble approachให้ค่า accuracy สูงขึ้นแบบค่อยเป็นค่อยไป (ผู้ชาย: 76% ผู้หญิง: 80%) ค่าความจำเพาะ (ผู้ชาย: 76% ผู้หญิง: 81%) และค่า positive predictive (ผู้ชาย: 26% ผู้หญิง: 24%) เมื่อเทียบกับวิธีอื่น ๆของแต่ละอัลกอริทึม

6.2 Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis. (Site, Nurmi, & Lohan, 2021)

การศึกษานี้เป็นการศึกษาที่มีการการทบทวนงานวิจัยอย่างเป็นระบบในเชิงลึกโดยเลือกงานวิจัยตามข้อกำหนดของ PRISMA (Preferred reporting items for systematic reviews and meta-analyses) คือ เป็นระเบียบปฏิบัติมาตรฐาน เพื่อการรายงานที่โปร่งใสและสมบูรณ์แบบในงานปริทัศน์เป็นระบบ และปัจจุบันเป็นมาตรฐานบังคับของวารสารแพทย์กว่า 170 วารสารทั่วโลก โดยงานวิจัยที่ได้รับเลือกได้รับคะแนนการทบทวนงานวิจัยมากกว่า 50%

การทบทวนนี้เน้นที่โรคต่างๆ ต่อไปนี้เพื่อรับข้อมูล eHealth: เบาหวานชนิดที่ 1 และชนิดที่ 2 ความดันโลหิตสูงและความดันเลือดต่ำ ภาวะหัวใจห้องบนเต้นผิดปกติ เช่น หัวใจเต้นเร็ว หัวใจเต้นช้า และโรคที่เกี่ยวข้องกับไข้ ข้อมูลสำหรับการทบทวนวรรณกรรมอย่างเป็นระบบรวบรวมจากฐานข้อมูลสี่แห่ง ได้แก่ Medline, ProQuest, Scopus และ Web of Science เราเลือกการศึกษา 67 ชิ้นสำหรับการทบทวนเชิงลึกขั้นสุดท้ายจากเอกสารที่เลือกไว้ล่วงหน้าจำนวน 1530 ฉบับ การศึกษาของเราระบุว่าข้อมูล eHealth ส่วนใหญ่ได้มาจากเซ็นเซอร์ เช่น มาตรความเร่งไจโรสโคป คลื่นไฟฟ้าหัวใจ (Electrocardiogram : ECG) จอภาพคลื่นไฟฟ้าสมอง (Electroencephalogram : EEG) และเซ็นเซอร์ระดับน้ำตาลในเลือด การศึกษานี้ยังตรวจสอบประเภทคุณลักษณะต่างๆ วิธีการแยกคุณลักษณะ และอัลกอริทึมการเรียนรู้ของเครื่อง ที่ใช้สำหรับการวิเคราะห์ข้อมูล eHealth การตรวจสอบของเรายังแสดงให้เห็นว่าอัลกอริทึมโครงข่ายประสาทเทียมและอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชิน ได้แสดงให้เห็นประสิทธิภาพที่ดีที่สุดในการวิเคราะห์ข้อมูลด้านการดูแลสุขภาพจากอัลกอริทึมการเรียนรู้ของเครื่อง อื่นๆ ที่ศึกษาในเอกสาร ดังแสดงในภาพประกอบ 13



ภาพประกอบ 13 ภาพแสดงอัลกอริทึมของปัญญาประดิษฐ์ที่โมเดลที่ใช้ ให้ประสิทธิภาพที่ดีในการทำนายผู้ป่วยโรคหลอดเลือดสมอง

6.3 Machine Learning for Brain Stroke: A Review. (Sirsat, Ferme, & Camara, 2020)

การศึกษานี้ได้ทบทวนงานวิจัยทั้งหมด 39 ชิ้นจากผลลัพธ์ของฐานข้อมูลทางวิทยาศาสตร์บนเว็บ ScienceDirect เกี่ยวกับการเรียนรู้ของเครื่อง ที่ใช้วิเคราะห์ข้อมูล สำหรับโรคหลอดเลือดสมองตั้งแต่ปี 2550 ถึง 2562 โดยการศึกษางานวิจัยนี้ทั้งหมดถูกจัดกลุ่มเป็น 4 หมวดหมู่ดังต่อไปนี้:

(ก) การป้องกันโรคหลอดเลือดสมอง

(ข) การวินิจฉัยโรคหลอดเลือดสมอง

(ค) การรักษาโรคหลอดเลือดสมอง และ

(ง) การพยากรณ์โรคหลอดเลือดสมองการทำนายผลผลลัพธ์หลังจากการคัดกรองการศึกษาทั้งหมด เราพบว่าการศึกษา

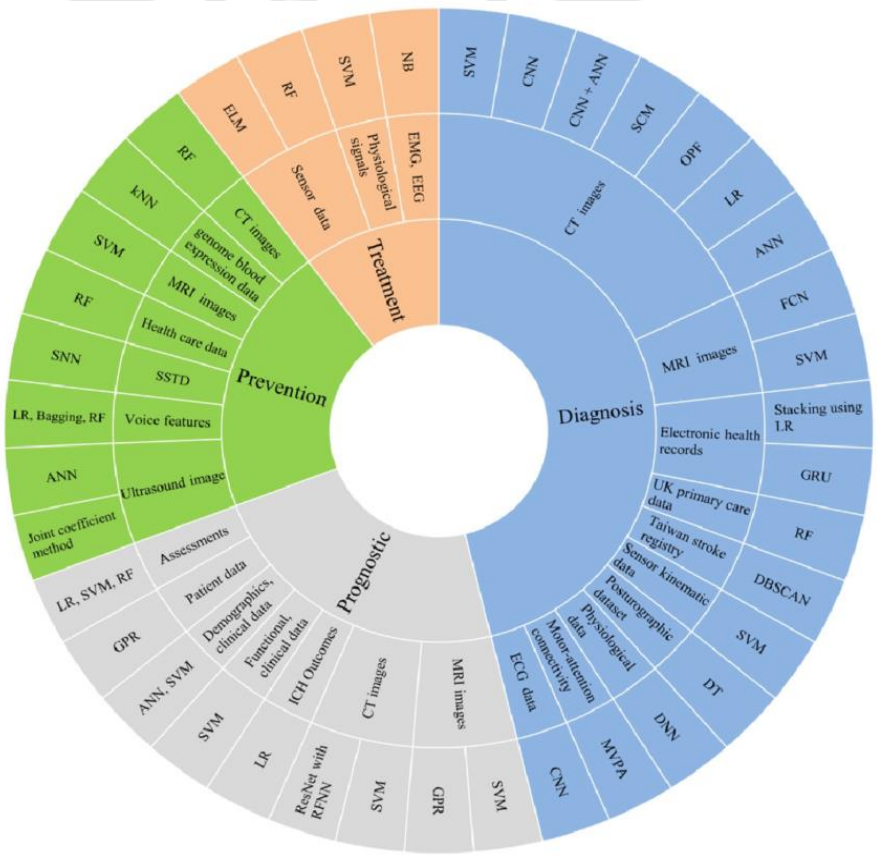
8 เรื่องเกี่ยวกับการป้องกันโรคหลอดเลือดสมอง

18 การศึกษาเกี่ยวกับการวินิจฉัยโรคหลอดเลือดสมอง

4 การศึกษาเกี่ยวกับการรักษาโรคหลอดเลือดสมอง

และ 9 การศึกษาเกี่ยวกับการพยากรณ์โรคหลอดเลือดสมอง

ผลลัพธ์ที่ได้ โดยพบว่าโมเดลของ ซัพพอร์ตเวกเตอร์แมชชีน และ ป่าสุ่ม เป็นโมเดลที่เหมาะสมที่สุดและเป็นเทคนิคที่มีประสิทธิภาพมากสุดซึ่งใช้ในแต่ละหมวดหมู่ ที่ใช้ในการศึกษาในโรคหลอดเลือดสมอง ดังแสดงในภาพประกอบ 14



ภาพประกอบ 14 Pie chart แสดงอัลกอริทึมของการเรียนรู้ของเครื่องที่ใช้ในขบวนการรักษาโรคหลอดเลือดสมองในช่วง การป้องกัน การวินิจฉัยโรค การรักษา และการติดตามผลการรักษา

6.4 Using a Multiclass Machine Learning Model to Predict the Outcome of Acute Ischemic Stroke Requiring Reperfusion Therapy. (Chiu, Zeng, Cheng, Chen, & Lin, 2021)

การทำนาย functional outcome ในผู้ป่วยโรคหลอดเลือดสมองตีบเป็นประโยชน์สำหรับการตัดสินใจทางคลินิก โดยการศึกษาอื่น ๆก่อนหน้านี้ได้กำหนดผลลัพธ์ในการทำนายออกได้เป็น การคาดคะเนผลลัพธ์ที่ดี (favorable outcomes) และการคาดคะเนผลผลลัพธ์ที่ไม่ดี (Miserable outcomes) โดยใช้หลักเกณฑ์จากเครื่องมือ modified Rankin Scale (mRS) ก่อนที่เราจะพิจารณาการให้การรักษา invasive intervention

ส่วนการเรียนรู้ด้วยเครื่องครั้งนี้ เรามุ่งที่จะพัฒนาแบบจำลองการจำแนกแบบหลายคลาส(multiclass classification model)สำหรับการทำนายผลลัพธ์ในผู้ป่วยโรคหลอดเลือดสมองตีบเฉียบพลันที่ต้องใช้การรักษาแบบการเปิดหลอดเลือดที่อุดตันอีกครั้ง (reperfusion therapy) ซึ่งเป็นการศึกษาย้อนหลังที่ศูนย์การแพทย์โรคหลอดเลือดสมองในไต้หวัน รวมผู้ป่วยโรคหลอดเลือดสมองตีบเฉียบพลันที่เข้ารับการรักษาระหว่างเดือนมกราคม 2559 ถึงธันวาคม 2562 และผู้ที่ได้รับการบำบัดด้วยการ reperfusion therapy ผลลัพธ์ทางคลินิกแบ่งออกเป็น 3 ระดับคือ

- ผลลัพธ์ที่ดี (favorable outcome)
- ผลลัพธ์ขั้นปานกลาง (intermediate outcome)
- และผลลัพธ์ที่ไม่ดี (miserable outcome)

เราพัฒนาโมเดลการเรียนรู้ของเครื่องหลายคลาสที่แตกต่างกันสี่แบบ (Logistic Regression, Supportive Vector Machine, Random Forest และ Extreme Gradient Boosting) เพื่อทำนายผลลัพธ์ทางคลินิก(clinical outcomes)และเปรียบเทียบประสิทธิภาพกับคะแนนของเครื่องมือการวัด DRAGON โดยมีผู้ป่วย 590 รายถูกรวมในการศึกษานี้ โดยพบว่าในจำนวนนี้ 180 (30.5%) รายมีผลดีและ 152 (25.8%) รายมีผลลัพธ์ที่ไม่ดี โมเดลการเรียนรู้ของเครื่องที่เลือกทั้งหมดให้ประสิทธิภาพในการทำนายเหนือกว่า DRAGON score ในด้านความแม่นยำของการทำนายผลลัพธ์ (Logistic Regression: 0.70, Supportive Vector Machine: 0.67, Random Forest: 0.69 และ Extreme Gradient Boosting: 0.67 เทียบกับ DRAGON: 0.51, $p < 0.001$) ในบรรดาแบบจำลองที่เลือกทั้งหมด Logistic Regression ยังมีประสิทธิภาพที่ดีกว่าคะแนน DRAGON score ในด้านค่าพยากรณ์เชิงบวก ความไว และความจำเพาะ เมื่อเทียบกับ DRAGON score วิธีการเรียนรู้ของเครื่องแบบหลายคลาสแสดงให้เห็นประสิทธิภาพที่ดีขึ้นในการคาดคะเน

functional outcome ใน 3 เดือนของผู้ป่วยโรคหลอดเลือดสมองตีบเฉียบพลันที่จำเป็นต้องได้รับการรักษาแบบการเปิดหลอดเลือดที่อุดตันอีกครั้ง

6.5 Assessing stroke severity using electronic health record data: a machine learning approach. (Kogan et al., 2020)

การศึกษานี้ “การประเมินความรุนแรงของโรคหลอดเลือดสมองโดยใช้ข้อมูลบันทึกสุขภาพอิเล็กทรอนิกส์: แนวทางการเรียนรู้ของเครื่อง” ความรุนแรงของโรคหลอดเลือดสมอง (Stroke severity) เป็นตัวทำนายที่สำคัญของผลลัพธ์ของการรักษาทางคลินิก (clinical outcome) ของผู้ป่วย และมักวัดด้วยคะแนน National Institutes of Health Stroke Scale (NIHSS) และเนื่องจากคะแนนเหล่านี้มักถูกบันทึกเป็นข้อความในรายงานของแพทย์ ดังนั้น จึงไม่ค่อยมีการนำเอาข้อมูลออกมาแปรเป็นระดับความความรุนแรงของโรคหลอดเลือดสมอง จุดมุ่งหมายของการศึกษานี้คือการใช้โมเดลการเรียนรู้ของเครื่องเพื่อกำหนดคะแนน NIHSS สำหรับผู้ป่วยทุกรายที่มีโรคหลอดเลือดสมองที่เพิ่งได้รับการวินิจฉัยใหม่จากข้อมูลบันทึกสุขภาพอิเล็กทรอนิกส์จากหลายสถาบัน multi-institution electronic health record (EHR) วิธีการ: คะแนน NIHSS ที่มีอยู่ในชุดข้อมูล Optum© de-identified Integrated Claims-Clinical ถูกดึงออกมาจากบันทึกของแพทย์โดยใช้วิธีการประมวลผลภาษาธรรมชาติ (natural language processing : NLP) กลุ่มที่วิเคราะห์ในการศึกษานี้ประกอบด้วยผู้ป่วย 7,149 รายที่เป็นผู้ป่วยในหรือผู้ป่วยห้องฉุกเฉินที่ถูกวินิจฉัยว่าเป็นเป็นโรคหลอดเลือดสมองตีบ โรคหลอดเลือดสมองแตก หรือโรคที่สมองขาดเลือดชั่วคราว และสอดคล้องกับคะแนน NIHSS ที่สกัดด้วย NLP และที่ถูกจัดออกมาเป็นกลุ่มย่อย (holdout set) ของผู้ป่วยเหล่านี้ (n = 1033, 14%) ถูกจัดไว้เพื่อตรวจสอบความถูกต้องของประสิทธิภาพของแบบจำลองและผู้ป่วยที่เหลือ (n = 6116, 86%) ถูกใช้สำหรับการฝึกโมเดล มีการประเมินโมเดลการเรียนรู้ของเครื่องหลายแบบ และพารามิเตอร์ปรับให้เหมาะสมโดยใช้การตรวจสอบ cross-validation จากtraining set โดยโมเดลที่มีประสิทธิภาพสูงสุดเป็น Random forest model ได้รับการประเมินในท้ายที่สุดใน holdout set

ผลลัพธ์ : ใช้ประโยชน์จากจากการเรียนรู้ของเครื่อง เราระบุปัจจัยหลักในข้อมูลบันทึกสุขภาพทางอิเล็กทรอนิกส์สำหรับการประเมินความรุนแรงของโรคหลอดเลือดสมอง รวมถึงการเสียชีวิตภายในเดือนเดียวกับการเกิดโรคหลอดเลือดสมอง ระยะเวลาพักรักษาตัวในโรงพยาบาลหลังเกิดโรคหลอดเลือดสมอง การวินิจฉัยภาวะสมองเสื่อม/ภาวะกลืนลำบาก การวินิจฉัยโรคอัมพาตครึ่งซีก การอนุญาตให้กลับบ้านหรือดูแลตนเอง การเปรียบเทียบคะแนน NIHSS ที่กำหนด

กับคะแนน NIHSS ที่สกัดด้วย NLP ในชุดข้อมูลholdout data set ให้ผล R2 (สัมประสิทธิ์การกำหนด) ที่ 0.57, R (ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน) ที่ 0.76 และ(root-mean-squared error) ข้อผิดพลาดค่าเฉลี่ยรากที่สองที่ 4.5

6.6 Machine learning to predict mortality after rehabilitation among patients with severe stroke. (Scrutinio et al., 2020)

โรคหลอดเลือดสมองเป็นหนึ่งในสาเหตุสำคัญของการเสียชีวิตและความพิการทั่วโลก ผู้รอดชีวิตจากโรคหลอดเลือดสมองประมาณ 20-25% มีความทุพพลภาพขั้นรุนแรง ซึ่งสัมพันธ์กับความเสี่ยงต่อการเสียชีวิตที่เพิ่มขึ้น การพยากรณ์โรคเป็นสิ่งที่สำคัญในกระบวนการตัดสินใจทางคลินิก วิธีการเรียนรู้ด้วยเครื่อง ได้รับความนิยมเพิ่มขึ้นในการวางแผนการรักษาทางการแพทย์ จุดมุ่งหมายของการศึกษาครั้งนี้แบ่งออกเป็น 2 ประเด็น (twofold) คือ : การประเมินประสิทธิภาพด้วยอัลกอริทึม ML tree-based สำหรับโมเดลการทำนายการเสียชีวิตใน 3 ปีในผู้ป่วยโรคหลอดเลือดสมอง 1,070 คนที่มีความทุพพลภาพขั้นรุนแรงซึ่งเสร็จสิ้นการฟื้นฟูแล้ว และเปรียบเทียบกับประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่อง กับการถดถอยโลจิสติกแบบมาตรฐาน และพบว่าโมเดลการถดถอยโลจิสติกแบบมาตรฐาน (standard logistic regression) ให้พื้นที่ภายใต้เส้นกราฟ (AUC) ที่ 0.745 และได้รับการปรับเทียบอย่างดี ที่เกณฑ์ความเสี่ยงที่เหมาะสมที่สุด (optimal risk threshold) แบบจำลองมีความแม่นยำ 75.7% ค่าพยากรณ์เชิงบวก (PPV) 33.9% และค่าพยากรณ์เชิงลบ (NPV) ที่ 91.0% อัลกอริทึม อัลกอริทึมการเรียนรู้ของเครื่องมีประสิทธิภาพเหนือกว่าโมเดลการถดถอยลอจิสติกส์ เมื่อใช้เทคนิคSMOTEและRandom Forests ได้ค่า AUC ที่ 0.928 และความแม่นยำ 86.3% PPV คือ 84.6% และ NPV 87.5%

6.7 Value-Based Healthcare in Ischemic Stroke Care: Case-Mix Adjustment Models for Clinical and Patient-Reported Outcomes. (Tsevat & Moriates, 2018)

การรายงานผลผลลัพธ์ของการรักษาโดยผู้ป่วย (Patient-Reported Outcome Measures : PROMs) ถูกเสนอเพื่อใช้เป็นตัวเปรียบเทียบคุณภาพการดูแลสุขภาพในโรงพยาบาลต่างๆ โดยใช้วิธี case-mix adjustment

จุดมุ่งหมายของการศึกษานี้คือการพัฒนาและเปรียบเทียบแบบจำลอง case-mix models สำหรับการตาย ผลลัพธ์การทำงานของร่างกายจากการรักษา (functional outcome) และการวัดผลลัพธ์การรักษาที่รายงานโดยผู้ป่วย (PROM) ในการดูแลโรคหลอดเลือดสมองตีบ

โดยข้อมูลจากผู้ป่วยโรคหลอดเลือดสมองตีบที่เข้ารับการรักษาในศูนย์โรคหลอดเลือดสมองสี่แห่งในประเทศเนเธอร์แลนด์ระหว่างปี 2557 และ 2559 พร้อมข้อมูลผลลัพธ์ของการรักษาที่มีอยู่ (N = 1,022) ได้ถูกนำมาทำการวิเคราะห์โดยมีแบบจำลองการปรับแบบผสม (Case-mix adjustment models)ได้รับการพัฒนาสำหรับอัตราการตาย โดยมี modified Rankin Scale (mRS) scores และ EQ-5D index scores ใช้วิเคราะห์ร่วมกับ binary logistic, proportional odds and linear regression models ใช้วิเคราะห์ร่วมกับ stepwise backward selection และความสามารถในการทำนายของแบบจำลองเหล่านี้จะถูกกำหนดด้วยสถิติ R-squared (R2) และ Area-under-the-receiver-operating-characteristic-curve (AUC)

ผลการศึกษาพบว่า อายุ NIHSS score ในช่วงแรกรับเข้ารับการรักษา และภาวะหัวใจล้มเหลวเป็นตัวพยากรณ์ทั่วไปใน case-mix adjustment modelsทั้ง 3 โมเดล ตัวทำนายเฉพาะสำหรับ EQ-5D ได้แก่ เพศ (β = 0.041) สถานะทางเศรษฐกิจและสังคม (β = − 0.019) และสัญชาติ (β = − 0.074) ค่า R2 สำหรับโมเดลการถดถอยสำหรับการตาย (5 ตัวทำนาย), คะแนน mRS (ตัวทำนาย 9 ตัว) และคะแนน EQ-5D utility score (ตัวทำนาย 12 ตัว) คือR2 = 0.44, R2 = 0.42 และ R2 = 0.37ตามลำดับ

6.8 The Probability of Ischemic Stroke Prediction with a Multi-Neural-Network Model. (Liu, Yin, & Cong, 2020)

ดังที่ทราบกันดีว่าโรคหลอดเลือดสมองได้กลายเป็นโรคที่สำคัญที่เป็นอันตรายต่อสุขภาพของผู้คน โรคหลอดเลือดสมองชนิดตีบ คิดเป็นประมาณ 85% ของโรคหลอดเลือดสมอง จากการวิจัยพบว่า การพยากรณ์และการป้องกันในระยะเริ่มต้นสามารถลดอัตราการเกิดโรคได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม เป็นการยากที่จะทำนายโรคหลอดเลือดสมองตีบเนื่องจากข้อมูลที่เกี่ยวข้องกับโรคนี้มีหลายรูปแบบ เพื่อให้ได้ความแม่นยำสูงในการทำนายและรวมตัวทำนาย (The stroke risk predictors) ความเสี่ยงโรคหลอดเลือดสมองที่ได้รับจากนักวิจัยคนก่อน ๆ ได้มีการเสนอวิธีการในการทำนายความน่าจะเป็นของการเกิดโรคหลอดเลือดสมองตามโครงสร้างโครงข่ายประสาทเทียมแบบหลายรุ่น ด้วยวิธีนี้ ความแม่นยำของการทำนายโรคหลอดเลือดสมองตีบจะดีขึ้นโดยการประมวลผลข้อมูลหลายโมดอลผ่านเครือข่ายประสาทเทียมแบบ end-to-end หลายเครือข่าย ในวิธีนี้ การดึงข้อมูลคุณลักษณะของข้อมูลที่มีโครงสร้าง (อายุ เพศ ประวัติความดันโลหิตสูง ฯลฯ) และการสตรีมข้อมูล (อัตราการเต้นของหัวใจ ความดันโลหิต ฯลฯ) โดยอิงจากโครงข่ายประสาทเทียมจะรับรู้ก่อน โมเดลโครงข่ายประสาทเทียมสำหรับการผสานคุณลักษณะจะ

ถูกสร้างขึ้นเพื่อให้เกิดการรวมคุณลักษณะของข้อมูลที่มีโครงสร้างและข้อมูลการสตรีม ในที่สุด โมเดลการทำนายสำหรับการทำนายความน่าจะเป็นของโรคหลอดเลือดสมองได้มาจากการฝึกอบรม ดังแสดงในผลการทดลอง ความแม่นยำของการทำนายโรคหลอดเลือดสมองตีบถึง 98.53% ความแม่นยำในการทำนายที่สูงเช่นนี้จะเป็นประโยชน์ในการป้องกันการเกิดโรคหลอดเลือดสมอง

6.9 Machine Learning Approach to Identify Stroke Within 4.5 Hours. (Lee et al., 2020)

การศึกษานี้มีวัตถุประสงค์เพื่อตรวจสอบความสามารถของเทคนิคการเรียนรู้ของเครื่อง (ML) ที่วิเคราะห์การถ่ายภาพด้วยคลื่นแม่เหล็กไฟฟ้า DWI(diffusion-weighted imaging) และการถ่ายภาพด้วยคลื่นสนามแม่เหล็กไฟฟ้า FLAIR (fluid-attenuated inversion recovery) เพื่อระบุผู้ป่วยว่าอยู่ภายในช่วงเวลาที่กำหนดที่เหมาะสำหรับการสลายลิ่มเลือด

วิธีการ: มีการวิเคราะห์ภาพ DWI และ FLAIR ของผู้ป่วยโรคหลอดเลือดสมองตีบเฉียบพลันต่อเนื่องกันภายใน 24 ชั่วโมงหลังจากเริ่มมีอาการที่ชัดเจนโดยใช้วิธีการประมวลผลภาพอัตโนมัติ (automatic image processing approaches) โดยกระบวนการเหล่านี้ประกอบไปด้วย การแบ่งส่วน infarct (infarct segmentation), การกำหนดขอบเขตของรอยโรคให้ชัดเจนด้วย DWI, and FLAIR imaging registration และการแยกคุณลักษณะของภาพ (image feature extraction) คุณลักษณะเวกเตอร์ (vector features) ทั้งหมด 89 รายการจากลำดับภาพแต่ละภาพถูก captured และนำมาใช้ใน ML โดยแบบจำลอง ML โมเดลสามแบบได้ถูกนำมาใช้ในการคลาดเดาประมาณระยะเวลาของการเกิดของโรคหลอดเลือดสมอง (estimate stroke onset time) สำหรับการจำแนกประเภทไบนารี (≤4.5 ชั่วโมง)โดยมีโมเดลที่ใช้ดังนี้ : การถดถอยโลจิสติก (support vector machine, and random forest) เครื่องเวกเตอร์สนับสนุน (support vector machine) และป่าสุ่ม (random forest)

การประเมินประสิทธิภาพ

ในการประเมินประสิทธิภาพของโมเดล ML และการอ่านผล DWI-FLAIR ของมนุษย์ ที่ไม่ตรงกัน ในการระบุเวลาของการเกิดโรคหลอดเลือดสมองภายใน 4.5 ชั่วโมง
โดยนำมาเปรียบเทียบกันใน ความไว (sensitivity) และความจำเพาะ (specificity)
ผลลัพธ์-

วิเคราะห์ข้อมูลจากผู้ป่วยทั้งหมด 355 ราย พบว่าการอ่านผลของ DWI-FLAIR ของโมเดล ML ไม่ตรงกันจากการอ่านของมนุษย์ระบุเวลาของการเกิดของโรคหลอดเลือดสมองของผู้ป่วยภายใน 4.5 ชั่วโมงหลังจากเริ่มมีอาการโดยมี ความไว 48.5% และความจำเพาะ 91.3% โดยพบว่าอัลกอริทึม ML มีความไวสูงกว่าเครื่องอ่านของมนุษย์อย่างมีนัยสำคัญ (75.8% สำหรับ logistic regression, P=0.020; 72.7% สำหรับ support vector machine, P=0.033; 75.8% สำหรับ random forest, P=0.013) ในการตรวจหาผู้ป่วยภายใน 4.5 ชั่วโมง ความจำเพาะเปรียบเทียบกันได้ (82.6% สำหรับการถดถอยโลจิสติก, P=0.157; 82.6% สำหรับเวกเตอร์เครื่องสนับสนุน, P=0.157; 82.6% สำหรับป่าสุ่ม, P=0.157)

6.10 Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study.(Chilamkurthy et al., 2018)

การศึกษาเป็นการเก็บรวบรวมชุดข้อมูลย้อนหลังของการทำสแกนด้วยเครื่องเอกซเรย์คอมพิวเตอร์ที่ศีรษะจำนวน 313,318 ภาพ พร้อมกับรายงานทางคลินิกจากศูนย์ประมาณ 20 แห่งในอินเดียระหว่างวันที่ 1 มกราคม 2011 ถึง 1 มิถุนายน 2017 และใช้อัลกอริทึมการเรียนรู้เชิงลึกเพื่อระบุความผิดปกติโดยอัตโนมัติในการสแกนด้วยเครื่องเอกซเรย์คอมพิวเตอร์ที่ศีรษะ พบว่าอัลกอริทึมการเรียนรู้เชิงลึกทำงานได้ดีในการตรวจจับการตกเลือดในกะโหลกศีรษะ (AUC, 0.94 [0.92–0.97]) และกระดูกไหปลาร้าหัก (AUC, 0.92 [0.91–0.94])

6.11 Early Stroke Prediction Using Machine Learning (Sharma, Sharma, Kumar, & Sodhi, 2022)

ในการทดลองนี้ใช้ข้อมูลจาก Kaggle dataset โดยได้ใช้อัลกอริทึมการจำแนกประเภทต่างๆ สำหรับการทำนายโรคหลอดเลือดสมองในระยะแรก โรคหลอดเลือดสมองเป็นหนึ่งในโรคที่ร้ายแรงมากในโลก และมีส่วนรับผิดชอบต่อการเสียชีวิตจำนวนมากทั้งทางตรงและทางอ้อม มีการใช้เทคนิคการทำเหมืองข้อมูลที่หลากหลายในอุตสาหกรรมการดูแลสุขภาพเพื่อช่วยในการวินิจฉัยและตรวจหาโรคในระยะเริ่มต้น องค์ประกอบหรือปัจจัยเสี่ยงหลายอย่างที่นำไปสู่โรคหลอดเลือดสมองได้รับการพิจารณาเป็นอันดับแรกๆ ในขบวนการการรักษา ดังนั้นเครื่องมือที่ใช้ในการมองหาผู้ที่มีโอกาสจะเป็นโรคหลอดเลือดสมองมากกว่าคนอื่นๆ จึงเป็นเรื่องสำคัญ และการศึกษานี้ได้อัลกอริทึมป่าสุ่ม ที่ให้ประสิทธิภาพดีที่สุดในการจำแนกผู้ที่มีโอกาสจะเป็นโรคหลอดเลือดสมอง ซึ่งทำให้ได้ความแม่นยำถึง 98.94 เปอร์เซ็นต์ และค่าความไวอยู่ที่ 98.90 เปอร์เซ็นต์

บทที่ 3
วิธีการดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

1. การกำหนดกลุ่มประชากร (patient classification)
2. ขั้นตอนการดำเนินการวิจัย (Research methodology)
3. โมเดลที่ใช้ในการวิจัย (Machine learning algorithms for Classification)
4. การเก็บรวบรวมข้อมูล (Data collection)
5. เตรียมข้อมูลให้พร้อมสร้างโมเดล (Prepare the Data)
6. การประเมินประสิทธิภาพของแต่ละโมเดล (Model evaluation)

1. การกำหนดกลุ่มประชากร

ประชากร

ในการศึกษานี้เป็นการศึกษาโอกาสการเกิดโรคหลอดเลือดสมอง โดยแบ่งผู้ป่วยออกเป็น 2 ประเภท คือ

1. กลุ่มผู้ป่วยที่มีโอกาสเกิดโรคหลอดเลือดสมอง (*Stroke)*

คือ ผู้ป่วยที่มีปัจจัยเสี่ยงที่มีโอกาสเกิดโรคหลอดเลือดสมองหรืออัมพฤกษ์อัมพาต

2. กลุ่มผู้ป่วยที่มีโอกาสไม่เกิดโรคหลอดเลือดสมอง (*No stroke)*

คือ ผู้ป่วยที่มีปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมองแต่มีโอกาสไม่เกิดโรคหลอดเลือดสมอง

การเลือกกลุ่มตัวอย่าง

ในการศึกษานี้เป็นการศึกษาการเกิดโรคหลอดเลือดสมองในวัยผู้ใหญ่ โดยเราศึกษาในกลุ่มผู้ป่วยที่มีอายุตั่งแต่ 18 ปีขึ้นไป โดยแบ่งผู้ป่วยออกเป็น 2 ประเภท คือ

กลุ่มผู้ป่วยที่เป็นโรคหลอดเลือดสมอง (stroke)

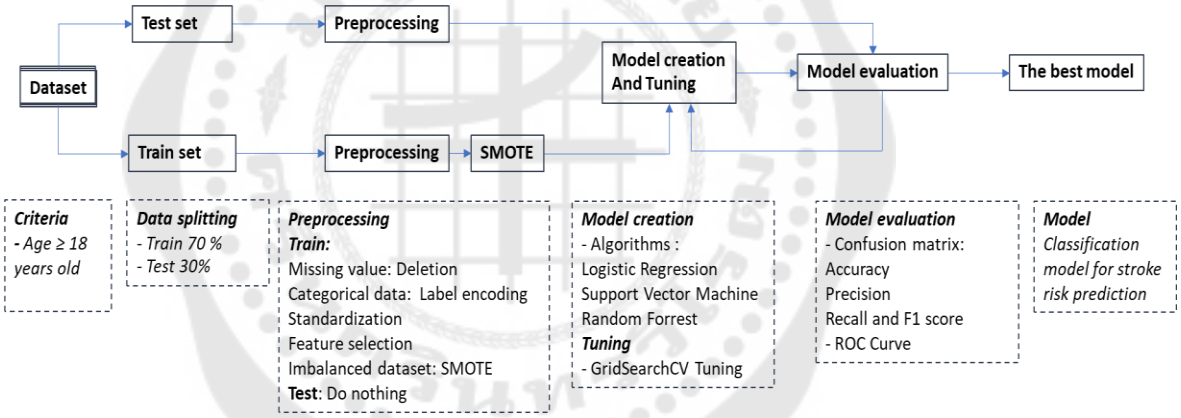
กลุ่มผู้ป่วยที่มีปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (Stroke Risk Factors)

กำหนดหัวข้อวิจัย (Define research topics)

จากการศึกษาพบว่าโรคหลอดเลือดสมองเป็นโรคที่พบบ่อยและส่งผลกระทบเป็นวงกว้างในหลาย ๆ ด้าน ตามที่ได้กล่าวแล้วในบทนำ ดังนั้นหัวข้อการวิจัยนี้ได้เล็งเห็นว่าขบวนการทางปัญญาประดิษฐ์สามารถที่จะช่วยเพิ่มคุณค่าของการศึกษาโดยช่วยสนับสนุนในขบวนการตัดสินใจในการดูแลรักษาโรคหลอดเลือดสมองให้มีประสิทธิภาสูง และสอดคล้องตามหลักการการดูแลสุขภาพโดยเน้นคุณค่า (Value Based-Healthcare) และสอดคล้องกับนโยบายของรัฐบาลของไทยในยุทธศาสตร์ Thailand 4.0 ที่สนับสนุนให้ใช้เทคโนโลยีในการพัฒนาการแพทย์ไทย

หัวข้อการวิจัย : การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง

2. ขั้นตอนการดำเนินการวิจัย



ภาพประกอบ 15 แสดงขั้นตอนการดำเนินการวิจัย

จากภาพประกอบ 15 เราสรุปขั้นตอนการดำเนินการวิจัยเป็นขั้นตอนต่างๆ ได้ดังนี้

2.1 Dataset เป็นชุดข้อมูลเกี่ยวกับสุขภาพที่มีอยู่ในชุดข้อมูลของ Kaggle dataset (Fedesoriano, 2021) ซึ่งมีข้อมูลผู้ป่วย 5,110 คนและเราเลือกผู้ป่วยเหลือเพียง 4,254 คน ที่เป็นผู้ป่วยวัยผู้ใหญ่ที่มีอายุ 18 ปีขึ้นไป

2.2 Data splinting แบ่งข้อมูลออกมาเป็น Training set 70% และ Test set 30% โดยนำข้อมูลส่วน Training set มาทำการสำรวจข้อมูลเพื่อหา insights ของข้อมูลโดยศึกษาแต่ละ attribute และคุณลักษณะของมัน เช่น ชื่อ,ประเภท % ของ missing value Outliers ของ

ข้อมูล การกระจายของข้อมูลเป็นรูปแบบใด กรณีนี้เป็น supervised learning ให้กำหนด target attribute Visualize ข้อมูล (plot กราฟ) ศึกษาความสัมพันธ์ (correlation) ระหว่าง attribute ต่างๆและ Test set โดยชุดทดสอบ (Test Set) เราจะวางไว้เฉยๆ โดยไม่ต้องไปยุ่งหรือแอบส่องมัน สำหรับการจัดการและการแบ่งชุดข้อมูลได้แสดงรายละเอียดดังตารางที่ 4

ตาราง 4 แสดง Dataset ถูกแบ่งออกเป็นสองส่วนคือ train set 70%, test set 30 %

Kaggle Dataset	The total number of each topic	No stroke	Stroke
Original dataset	5,110	4719	249
Missing value ('bmi')	201	162	39
Adult dataset (Age ≥ 18 years)	4073	3865	247
Child dataset (Age < 18 years)	856	854	2
Adult dataset after missing value deletion	4254	4007	247
Splitting (Train 70 % : Test 30 %)	70% Train = 2978: 30%Test = 1276	Train 2832: Test 1214	Train 146: Test 62

จากภาพประกอบ 4 ที่ใช้ในการศึกษานี้ มีข้อมูลของผู้ป่วยทั้งหมด 5,110 คน เราจัดการค่า missing values โดยการลบแถวนั้นออก และข้อมูลผู้ป่วยที่เราใช้ในการศึกษานี้เหลือเพียง 4,254 คน ที่เป็นผู้ป่วยวัยผู้ใหญ่ที่มีอายุ 18 ปีขึ้นไป ที่ประกอบไปด้วยผู้ป่วยที่เป็นไม่เป็น stroke จำนวน 4,007 คน และผู้ป่วยที่เป็น stroke จำนวน 247 คน

2.3 Preprocessing Data

เตรียมข้อมูลให้พร้อมสร้างโมเดล โดยทำ Data Cleaning: พิจารณาเกี่ยวกับข้อมูลมี missing values 181 records เราจัดการโดย drop แถวนี้ออก

ค่า avg_glucose_level มี scale ที่แตกต่างกันมากเราจัดการด้วยการทำ Feature Engineering โดยการทำ standardization technique ให้มีค่าที่อยู่ใน scale ที่ใกล้เคียงกัน

2.4 Model Selection and creation

ในการศึกษานี้เราเลือกใช้ algorithms ที่ให้ประสิทธิภาพในการทำ classification ที่ดีกับข้อมูลทางการแพทย์ 3 algorithms คือ SVM RF และ LR
แล้วปรับจูนโมเดลหาผลลัพธ์ที่ขึ้นและเลือกโมเดลที่มีประสิทธิภาพที่ดีที่สุด

2.5 Model evaluation

เป็นการวัดผลและเปรียบเทียบ performance ของแต่ละโมเดล โดยใช้ข้อมูลจาก Test set ที่เตรียมไว้และวัดผลและเปรียบเทียบ performance ของแต่ละโมเดลโดยใช้ confusion matrix, Accuracy, sensitivity, specificity, F1 score

3. โมเดลที่ใช้ในการวิจัย

การเลือกโมเดลการเรียนรู้ของเครื่อง เราจะเลือกโมเดลที่หลากหลายและเหมาะสมกับประเภทของข้อมูล สำหรับการศึกษานี้เราเลือกใช้ 3 algorithms ดังนี้ Logistic regression, Support Vector Machine และ Random forest

4. รวบรวมข้อมูล

ข้อมูลผู้ป่วยได้มาจาก Kaggle dataset มีขนาด 317 KB ในรูปแบบไฟล์ CSV ข้อมูลของเราประกอบไปด้วย 11 features และ 5,110 records

ตาราง 5 แสดงตัวแปร ประเภทของตัวแปร และความหมายของตัวแปรที่ใช้ในการศึกษา

No.	Attribute	Values	Definition
1	gender	String literal(Male, Female, Other)	**บอกเพศของผู้ป่วย**
2	age	Integer	**บอกอายุของผู้ป่วย**
3	hypertension	Integer (1, 0)	**บอกผู้ป่วยเป็นโรคความดันสูงหรือไม่**
4	heart_disease	Integer (1, 0)	**บอกผู้ป่วยเป็นโรคหัวใจหรือไม่**
5	ever_married	String literal (Yes, N0)	**บอกสถานะของผู้ป่วยว่าแต่งงานหรือไม่**
6	work_type	String literal (Child, Govt_worked, Private, Self-employed)	**บอกประเภทของงานที่ผู้ป่วยทำงานอยู่**

ตาราง 5 (ต่อ)

No.	Attribute	Values	Definition
7	Residence_type	String literal (Urban,Rural)	**บอกสถานที่อยู่อาศัยของผู้ป่วย**
8	avg_glucose_level	Floating point number	**ค่าเฉลี่ยของระดับน้ำตาลในเลือด**
9	bmi	Floating point number	**ค่าดัชนีมวลกายของผู้ป่วย**
10	Smoking_status	String literal (formerly smoked, never smoked, smoke, unknown)	**บอกสถานะการสูบบุหรี่ของผู้ป่วย**
11	stroke	Integer (1, 0)	**บอกสถานะการเป็นโรคหลอดเลือดสมองของผู้ป่วย**

จากตารางที่ 5 แสดงให้เห็นตัวแปร ประเภทของตัวแปร และความหมายของตัวแปรที่ใช้ในการศึกษาโดยจำนวนตัวแปร (Variables หรือ Features) ที่มีอยู่ใน Dataset ที่ใช้ในการศึกษานี้ พบว่ามีทั้งหมด 11 ตัวแปรประกอบไปด้วย gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke

4.1 จากการตรวจสอบข้อมูลเราสามารถแบ่งข้อมูลของตัวแปร ออกได้เป็น 2 ประเภท

4.1.1 ตัวแปรหรือข้อมูลที่เป็นตัวเลข(Numerical Features)

มี age, avg_glucose_level, bmi

4.1.2 ตัวแปรหรือข้อมูลที่เป็นข้อมูลเชิงคุณภาพ (Qualitative Feature, Categorical Feature)

มี gender, hypertension,heart_disease, ever_married, work_type, Residence_type, smoking_status, stroke

สำรวจข้อมูลและการวิเคราะห์ข้อมูล (Explore the Data and Data Analysis

การแสดงรายละเอียดของข้อมูล

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4254 entries, 0 to 5109
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   gender             4254 non-null   object 
 1   age                4254 non-null   float64
 2   hypertension       4254 non-null   int64  
 3   heart_disease      4254 non-null   int64  
 4   ever_married       4254 non-null   object 
 5   work_type          4254 non-null   object 
 6   Residence_type     4254 non-null   object 
 7   avg_glucose_level  4254 non-null   float64
 8   bmi                4073 non-null   float64
 9   smoking_status     4254 non-null   object 
 10  stroke             4254 non-null   int64  
dtypes: float64(3), int64(3), object(5)
memory usage: 398.8+ KB
```

ภาพประกอบ 16 แสดงจำนวนตัวแปรทั้งหมด 11 ตัวแปร แสดงค่า non-null ของแต่ละตัวแปร และแสดงชนิดของข้อมูล (dtypes) ของแต่ละตัวแปร

จากภาพประกอบ 16 แสดงรายละเอียดของข้อมูลที่เราจะศึกษาเกี่ยวกับการทำนายโอกาสการเกิดโรคหลอดเลือดสมองในวัยผู้ใหญ่ ที่เราเริ่มอายุตั่งแต่อายุ 18 ปีขึ้นไป

- โดยจำนวนตัวแปรที่มีอยู่ใน Dataset ที่ใช้ในการศึกษานี้ พบว่ามีทั้งหมด 11 ตัวแปรประกอบไปด้วย ‘gender’, ‘age’, ‘hypertension’, 'heart_disease’, ‘ever_married', 'work_type', ‘Residence_type', ‘'avg_glucose_level’, ‘bmi’, ‘smoking_status', ‘stroke’ แสดงค่าให้เห็นการหาค่า non-null ของแต่ละตัวแปร ซึ่งพบว่าทุกตัวแปรมีค่า non-null 4,254 ตัว ยกเว้นตัวแปร bmi มีค่าค่า non-null 4,073 ตัว (นั้นคือมี null เท่ากับ 181 records)
- แสดงชนิดของข้อมูลของแต่ละตัวแปร โดยมีข้อมูลชนิด float64 จำนวน 3 ตัวแปร ข้อมูลชนิด int64 จำนวน 3 ตัวแปร และข้อมูลชนิด objects จำนวน 5 ตัวแปร

การแสดงค่าทางสถิติของข้อมูล

```
df.describe()
```

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	4254.000000	4254.000000	4254.000000	4254.000000	4254.000000	4254.000000
mean	50.202163	0.116831	0.064645	108.514394	30.432752	0.058063
std	17.829035	0.321257	0.245927	47.769400	7.079512	0.233890
min	18.000000	0.000000	0.000000	55.120000	11.300000	0.000000
25%	36.000000	0.000000	0.000000	77.482500	25.600000	0.000000
50%	50.500000	0.000000	0.000000	92.465000	29.600000	0.000000
75%	64.000000	0.000000	0.000000	116.135000	33.800000	0.000000
max	82.000000	1.000000	1.000000	271.740000	92.000000	1.000000

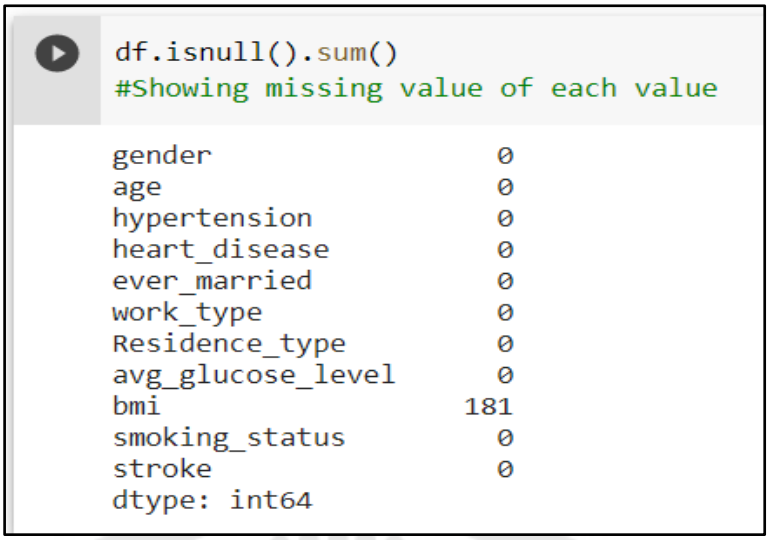
ภาพประกอบ 17 แสดงค่าต่างๆ ทางสถิติ ของตัวแปรต่างๆที่ใช้ในการศึกษาครั้งนี้

จากภาพประกอบ 17 แสดงให้เห็นความแตกต่างของค่าของข้อมูลที่มีช่วงของค่าของข้อมูลที่มีค่าแตกต่างกันมากในแต่ละตัวแปร ดังนั้นการจะนำข้อมูลไปใช้ในการสร้างโมเดลจะต้องมีการทำให้ข้อมูลอยู่ในช่วงเดียวกันก่อนโดยการทำ feature scaling เพราะการ Scale ค่ามีผลมากกับ Linear Model ถ้าเกิดข้อมูลของเราผ่านการ Scale มาแล้ว ค่าความแปรปรวนก็จะเท่าๆ กันในทุกๆ feature และอีกทั้งยังทำให้ค่า mean ของข้อมูลของเราเป็นกลางมากขึ้น ทำให้การสร้างโมเดล ได้ประสิทธิภาพที่ดีขึ้นด้วย

การทำ Feature Scaling คือ วิธีการปรับช่วงขอบเขตของข้อมูลชนิดตัวเลข Cardinal แต่ละ Feature (Field) ให้อยู่ในช่วงเดียวกัน ที่เหมาะกับการนำไปประมวลผลต่อ เข้าสูตรคำนวณได้ง่าย เช่น ช่วง [0, 1] หรือ [-1, 1] ได้ผลลัพธ์อยู่ในช่วงที่กำหนด เรียกว่า Data Normalization นิยมทำในขั้นตอน Preprocessing จัดเตรียมข้อมูล ก่อนป้อนให้โมเดลใช้เทรน

5. จัดเตรียมข้อมูลเพื่อพร้อมสร้างโมเดล

การตรวจสอบหา missing value ของข้อมูล



```
df.isnull().sum()
#Showing missing value of each value

gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  181
smoking_status         0
stroke                 0
dtype: int64
```

ภาพประกอบ 18 แสดงการทดสอบหา Missing value ของตัวแปร พบว่าตัวแปร ‘bmi’ มีค่า missing value 181 records

จากภาพประกอบที่ 18 แสดงการหาค่า null ในแต่ละ features ซึ่งพบว่าตัวแปร ‘bmi’ มีค่า null หรือ missing value เท่ากับ 181 records

การจัดการกับค่าปัญหา missing value มีสองวิธีหลักๆ คือ

1. **วิธีการลบข้อมูล** (Listwise Deletion or Complete Case Analysis)

กรณีที่ข้อมูลสูญหายเกิดขึ้นหลายตัวแปร แต่ปริมาณการสูญหายไม่เกิน 5% ของข้อมูลทั้งหมด เราก็จะตัดหรือลบข้อมูลส่วนที่สูญหายออกทั้ง record ซึ่งถือเป็นวิธีพื้นฐานที่นิยมใช้กัน ข้อดีคือง่ายและสามารถวิเคราะห์เชิงเปรียบเทียบระหว่างตัวแปรได้ เพราะแต่ละตัวแปรมีขนาดเท่ากัน แต่ข้อเสียที่เกิดขึ้นคือ ผลลัพธ์ที่ได้จากการวิเคราะห์เชื่อถือไม่ได้ 100% เนื่องจากมีข้อมูลบางส่วนถูกตัดออกไป ทำให้ข้อมูลไม่ครบถ้วนสมบูรณ์ ที่แย่ไปกว่านั้นคือ ถ้าข้อมูลที่สูญหายมีลักษณะกระจายตัว การตัดข้อมูลสูญหายทิ้งอาจจะทำให้ข้อมูลมีความเอนเอียง เบ้ซ้าย เบ้ขวา เนื่องจากข้อมูลของกลุ่มตัวอย่างแต่ละกลุ่มถูกตัดออกไม่เท่ากัน

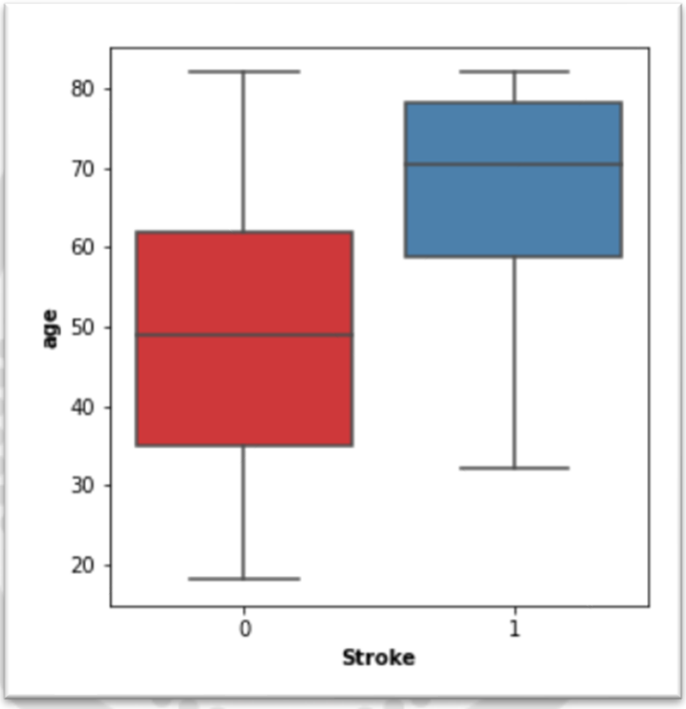
2. **วิธีการประมาณค่าข้อมูลสูญหาย** (Imputation Methods)

เป็นวิธีการประมาณค่าสูญหายโดยเอาหลักการทางคณิตศาสตร์ มาเติมเต็มค่าที่สูญหายไป ทำให้ผลลัพธ์สุดท้ายคล้ายกับว่าไม่เคยมีข้อมูลสูญหายเกิดขึ้นมาก่อนเลย ซึ่งมีหลากหลายวิธีมาก เช่น วิธีการประมาณค่าโดยวิธีเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) วิธีการประมาณค่าด้วยค่าเฉลี่ย (Mean Imputation: MI) เป็นต้น

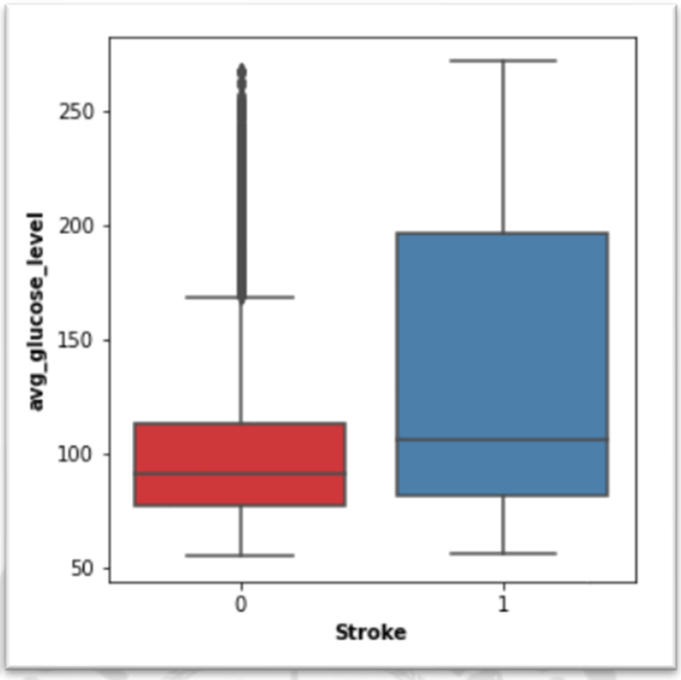
สำหรับการศึกษานี้เราเลือกใช้วิธีการจัดการกับ missing value ของตัวแปร bmi ด้วยวิธีการลบแถวที่มีค่า missing value ออก ดังแสดงในภาพประกอบ 35

การหาค่า outliers

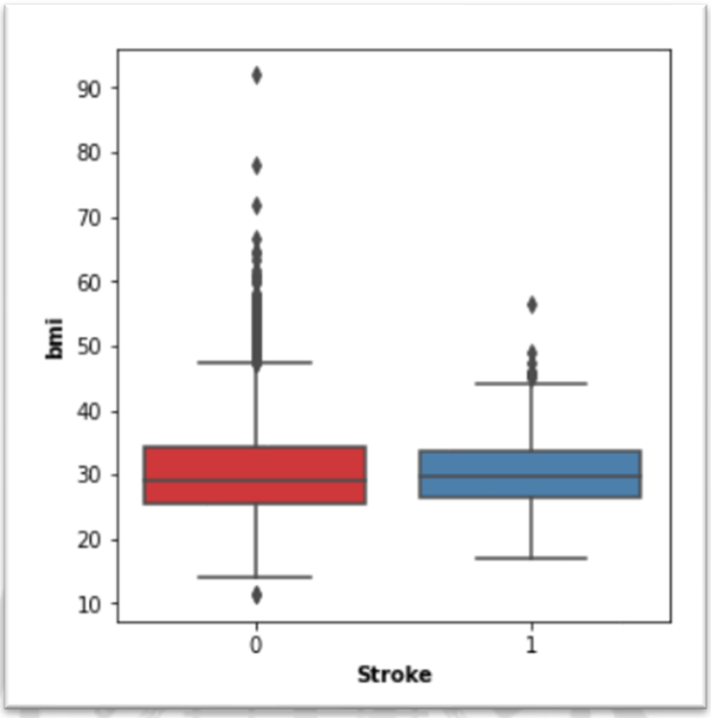
การหาค่า outliers ของตัวแปรที่เป็นตัวเลข (Numerical Features) ‘age’, ‘agv_glucose_level’ and ‘bmi’



ภาพประกอบ 19 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ age ในกลุ่มที่มีโอกาสไม่เป็น stroke กับกลุ่มที่มีโอกาสเป็น stroke



ภาพประกอบ 20 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ avg_glucose_level ในกลุ่มที่มีโอกาสไม่เน stroke กับกลุ่มที่มีโอกาสเป็น stroke

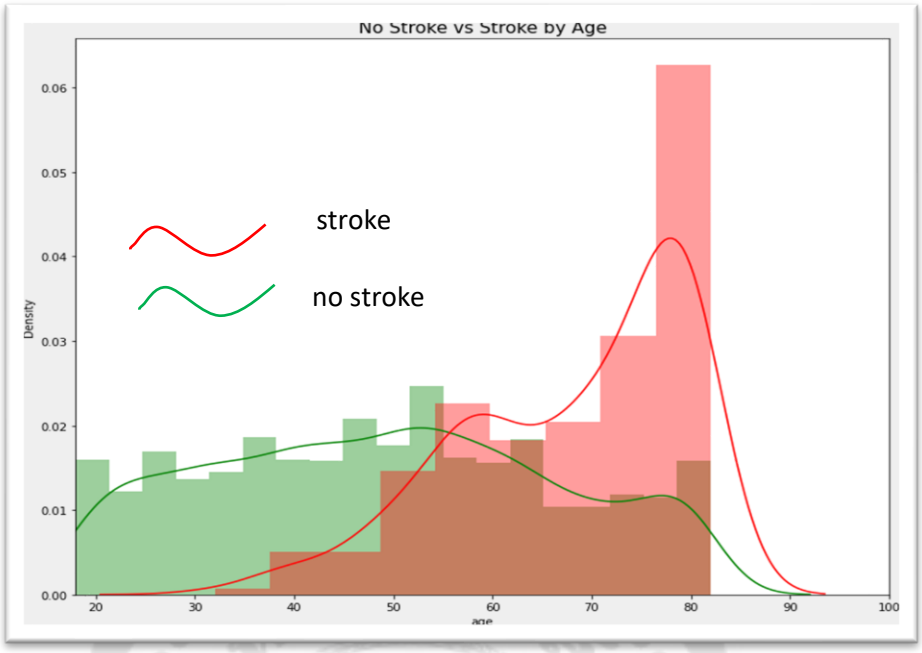


ภาพประกอบ 21 แสดง outliers ของตัวแปรที่เป็นตัวเลขของ bmi ในกลุ่มที่มีโอกาสไม่เป็น stroke กับกลุ่มที่มีโอกาสเป็น stroke

จากภาพประกอบที่ 19, 20, 21 แสดงบ็อกซ์พล็อตที่แสดงค่า outlier ในส่วนของตัวแปรของ 'avg_glucose_level' และ 'bmi' แต่เมื่อพิจารณาในทางการแพทย์แล้วพบว่าค่าของตัวแปรเหล่านี้ถือว่าเป็นค่าที่เป็นไปได้ในทางการแพทย์ ในผู้ป่วยที่มีปัจจัยเสี่ยง เราจึงไม่ตัดค่าเหล่านี้ออก

การกระจายของข้อมูลเป็นรูปแบบในกลุ่ม Numerical Feature

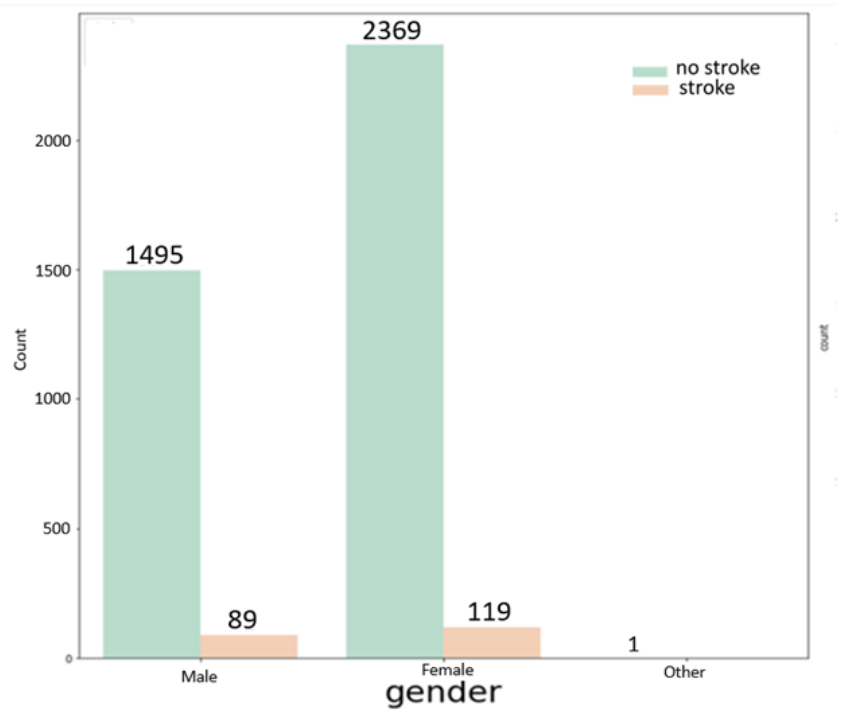
ตัวแปรที่เป็นตัวเลข (Numerical Features)



ภาพประกอบ 22 แสดงความสัมพันธ์ของตัวแปรตัวที่เป็นตัวเลขกับการเกิดโรคหลอดเลือดสมอง

จากภาพประกอบ 22 การกระจายตัวของข้อมูลของตัวแปรที่เป็นตัวเลขเราพบว่า ผู้ป่วยที่มีอายุมาก (>65 ปี) เป็นปัจจัยที่มีความสัมพันธ์กับการเกิดโรคหลอดเลือดสมอง

ตัวแปรตัวที่เป็นข้อมูลเชิงคุณภาพ (Categorical Features)



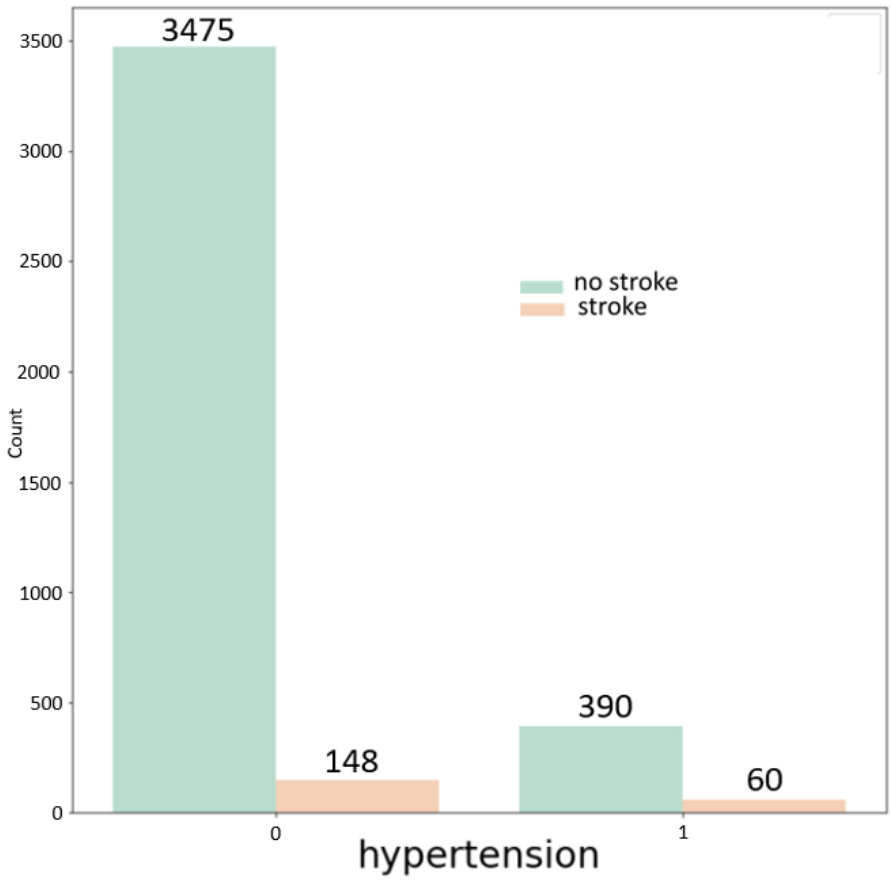
ภาพประกอบ 23 กราฟแสดงความสัมพันธ์ระหว่าง gender และ stroke

จากภาพประกอบ 23 ชุดข้อมูลประกอบด้วยเพศหญิง 61.10% และเพศชาย 38.90% และผู้ชายมีโอกาสเกิดเป็นโรคหลอดเลือดสมองมากกว่าผู้หญิงประมาณ 0.74 % โดย

ผู้หญิงมีโอกาสเป็นโรคหลอดเลือดสมอง 4.78%

ผู้ชายมีโอกาสเป็นโรคหลอดเลือดสมอง 5.62%

บุคคลจากเพศอื่นมีโอกาสเป็นโรคหลอดเลือดสมอง 0.0%

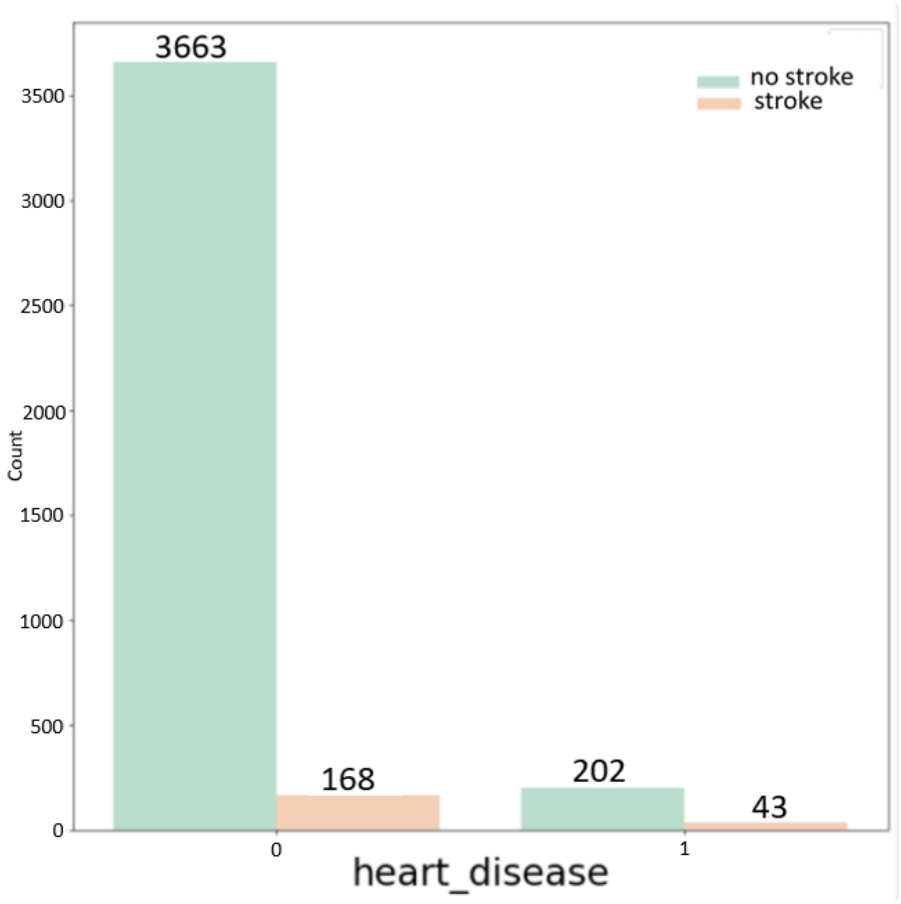


ภาพประกอบ 24 กราฟแสดงความสัมพันธ์ระหว่าง hypertension และ stroke

จากภาพประกอบ 24 พบว่าผู้ที่เป็นโรคความดันโลหิตสูงมีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าผู้ที่ไม่เป็นโรคความดันโลหิตสูง 9.24%

ผู้ที่เป็นโรคความดันโลหิตสูงมีโอกาสเป็นโรคหลอดเลือดสมองร้อยละ 13.33 %

คนไม่มีความดันโลหิตสูง มีโอกาสเป็นโรคหลอดเลือดสมอง 4.09 %

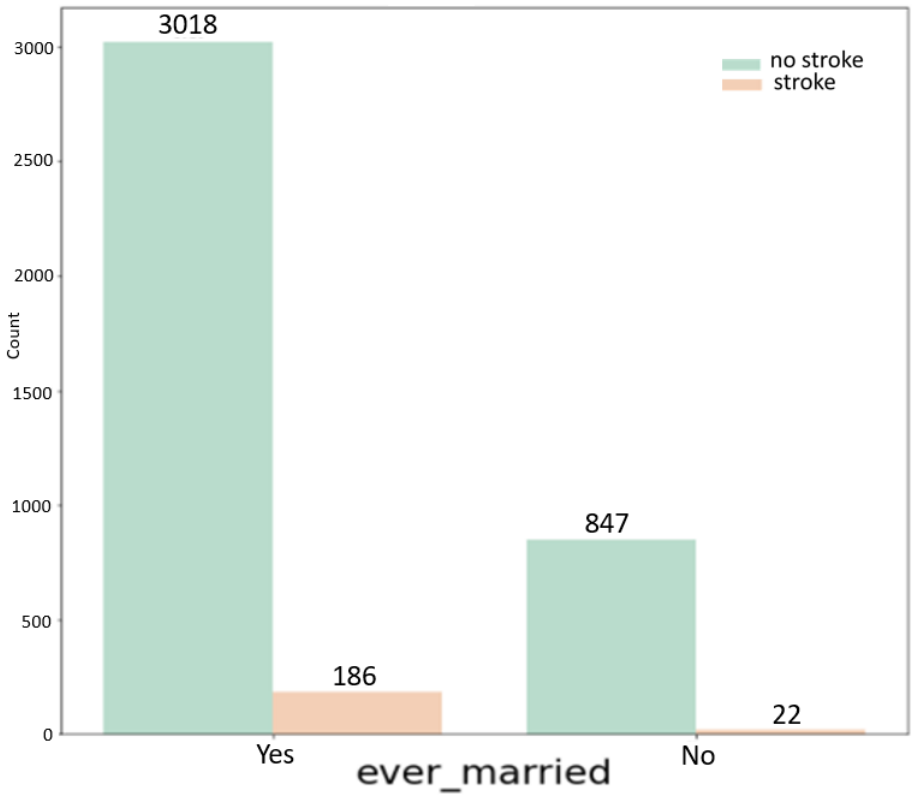


ภาพประกอบ 25 กราฟแสดงความสัมพันธ์ระหว่าง heart_disease และ stroke

จากภาพประกอบ 25 พบว่าผู้ที่เป็นโรคหัวใจมีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าผู้ที่ไม่เป็นโรคหัวใจ 12.14%

คนเป็นโรคหัวใจ มีโอกาสเป็นโรคหลอดเลือดสมอง 16.53%

คนไม่เป็นโรคหัวใจ มีโอกาสเป็นโรคหลอดเลือดสมอง 4.39%

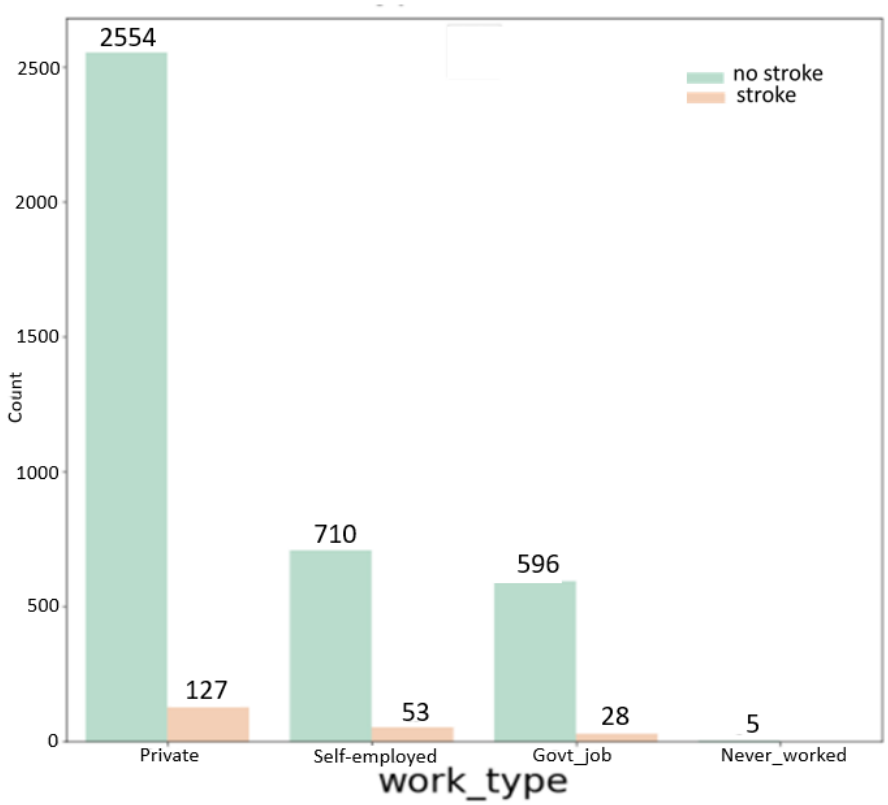


ภาพประกอบ 26 กราฟแสดงความสัมพันธ์ระหว่าง ever_married และ stroke

จากภาพประกอบ 26 พบว่าคนที่แต่งงานแล้วมีโอกาสเป็นโรคหลอดเลือดสมองสูงกว่าคนที่ยังไม่ได้แต่งงาน 3.28%

คนที่แต่งงานแล้ว (หรือเคยแต่งงานมาก่อน) มีโอกาสเป็น ที่จะเป็นโรคหลอดเลือดสมอง 5.81%

คนที่ไม่เคยแต่งงานมีโอกาสเป็นโรคหลอดเลือดสมอง 2.53%



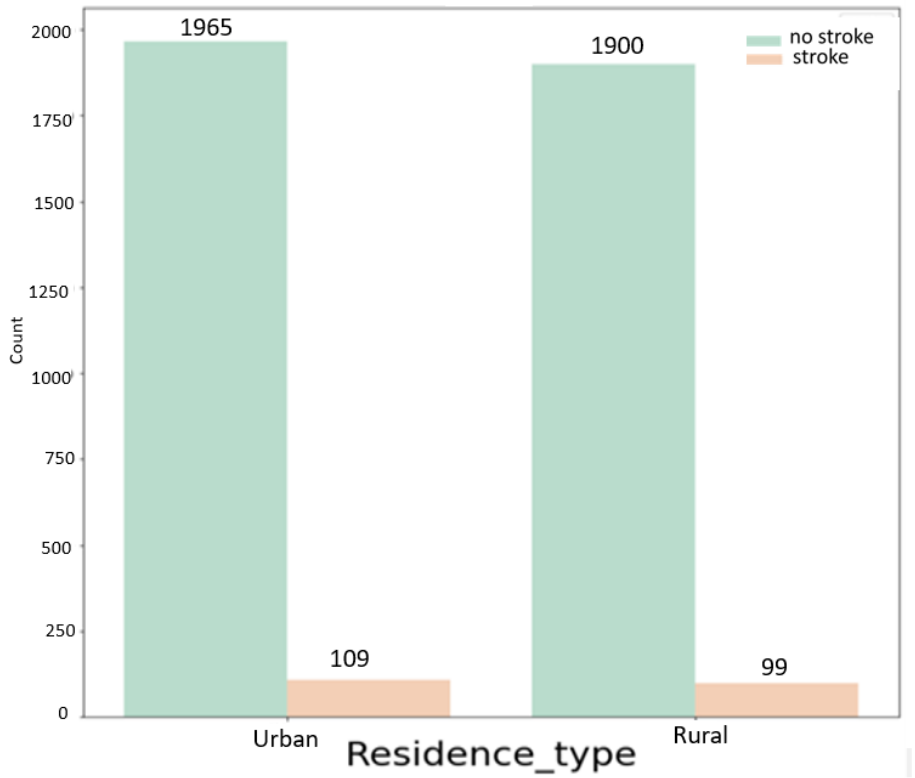
ภาพประกอบ 27 กราฟแสดงความสัมพันธ์ระหว่าง work_type และ stroke

จากภาพประกอบ 27 พบว่าคนที่ทำงานในราชการมีความน่าจะเป็นประมาณ 4.49% ที่จะเป็นโรคหลอดเลือดสมอง

ผู้ที่ทำงานในภาคเอกชนมีโอกาสเป็นโรคหลอดเลือดสมอง 4.74%

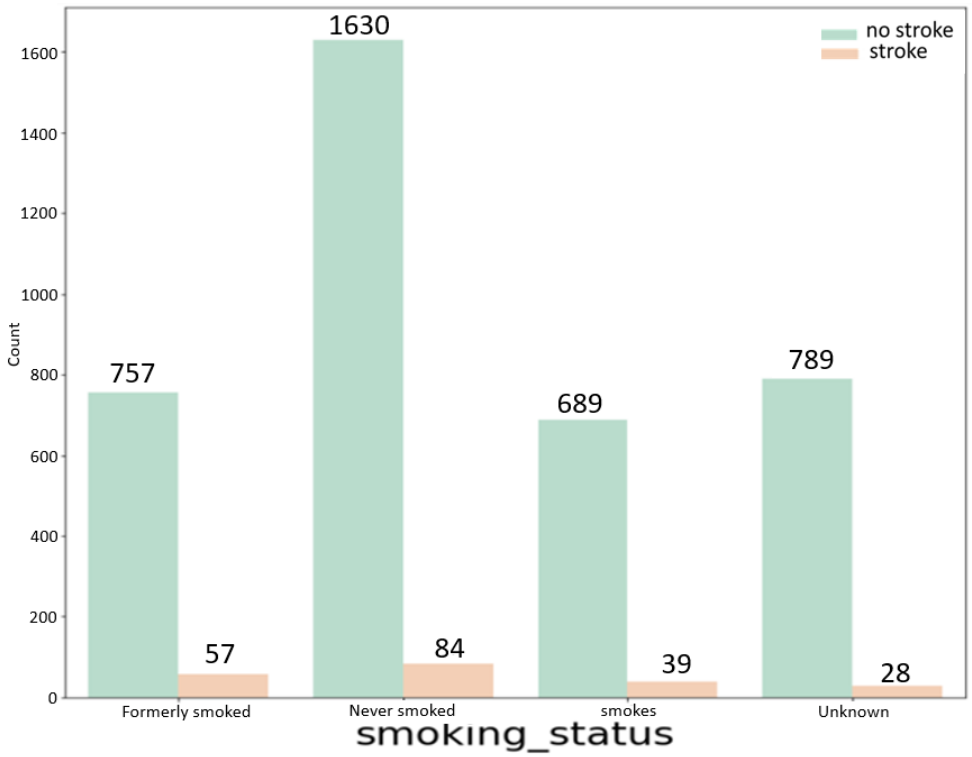
ผู้ที่ประกอบอาชีพอิสระมีโอกาสเป็นโรคหลอดเลือดสมอง 6.95%

ผู้ที่ไม่มีงานทำมีโอกาสเป็นโรคหลอดเลือดสมอง 0.00%



ภาพประกอบ 28 กราฟแสดงความสัมพันธ์ระหว่าง Residence_type และ stroke

จากภาพประกอบ 28 จากข้อมูลข้างต้นค่อนข้างชัดเจนว่าไม่มีความแตกต่างระหว่างประเภทที่อยู่อาศัย (residence type) กับความสัมพันธ์ที่มีต่อโอกาสเกิดโรคหลอดเลือดสมอง
คนที่อาศัยอยู่ในเขตเมืองมีโอกาสเป็นโรคหลอดเลือดสมอง 5.26%
คนที่อาศัยอยู่ในพื้นที่ชนบทมีโอกาสเป็นโรคหลอดเลือดสมองได้ 4.95 %



ภาพประกอบ *29* กราฟแสดงความสัมพันธ์ระหว่าง smoking_status และ stroke

จากภาพประกอบ 29 คนที่เคยสูบบุหรี่มาก่อนมีโอกาสที่จะเป็นโรคหลอดเลือดสมอง 7.0%

คนที่ไม่เคยสูบบุหรี่มีโอกาสที่จะเป็นโรคหลอดเลือดสมอง 4.9%

คนที่สูบบุหรี่มีโอกาสที่จะเป็นโรคหลอดเลือดสมอง 5.36%

ผู้ที่ไม่ทราบประวัติการสูบบุหรี่มีโอกาสเป็นโรคหลอดเลือดสมองร้อยละ 3.43

คนที่เคยสูบบุหรี่มาก่อนมีโอกาสที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่ไม่เคยสูบบุหรี่ 2.1%

การแบ่งข้อมูลออกเป็นตัวแปรต้นและตัวแปรตามในการสร้างโมเดล

ในการสร้างโมเดลการเรียนรู้ของเครื่องแบบ supervised learning เราได้กำหนด target attribute (attributeหรือตัวแปร ที่เป็นคำตอบ) โดยเรากำหนดให้ ‘’stroke” เป็น target attribute (y)

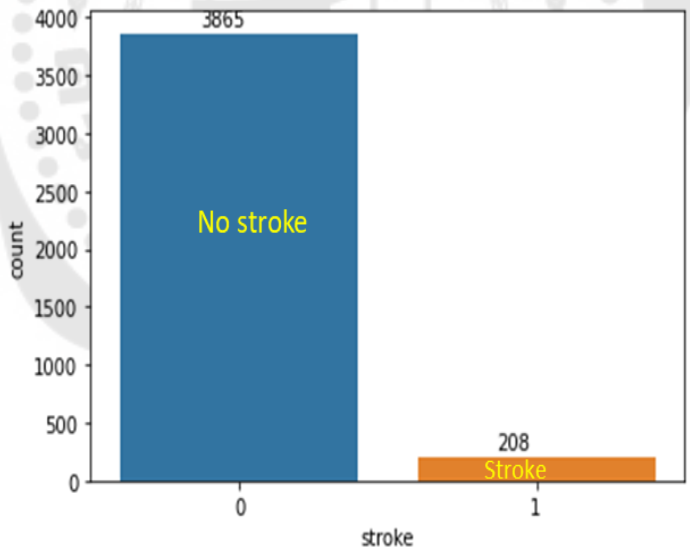
```
[ ]  X = df.iloc[:,0:-1].values
     y = df.iloc[:, -1].values
     # This will split the data into target and values column with arrays shape
```

ภาพประกอบ 30 เรากำหนดให้ ‘‘stroke” เป็น target attribute

จากภาพประกอบ 30 แสดงการกำหนดตัวแปรต้น (x) และกำหนดตัวแปรตามหรือตัวแปรที่เป็น target attribute (y)

ชุดข้อมูลที่ไม่สมดุล (Imbalanced Datasets)

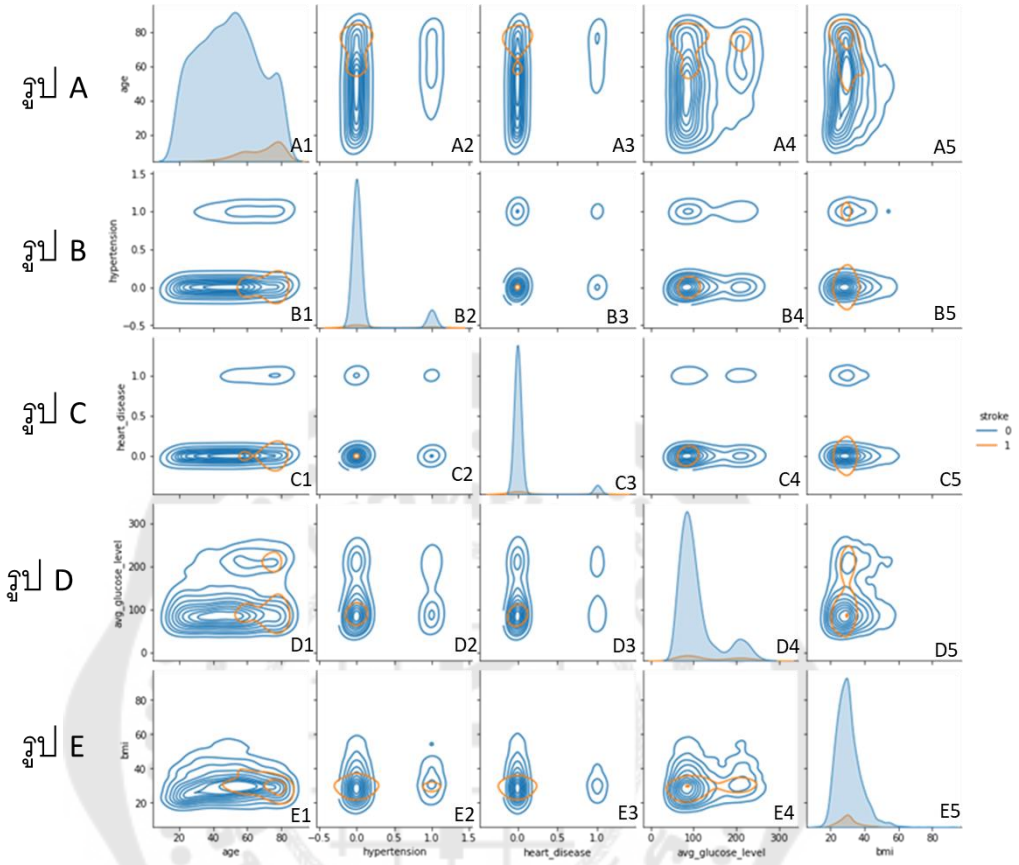
ซึ่งชุดข้อมูลที่ไม่สมดุล (Imbalanced Datasets) ทั้งสองชุดจะถูกนำเข้าสู่ขบวนการสร้างโมเดลพร้อมกันทั้งหมด ซึ่งจะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด กล่าวคือ ข้อมูลที่อยู่ในกลุ่มส่วนน้อยจะถูกจัดให้ไปอยู่ในกลุ่มส่วนมากทั้งหมด ซึ่งจะนำไปสู่ปัญหาที่เรียกว่า ปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (misclassification)



ภาพประกอบ 31 กราฟแสดงความไม่สมดุลของข้อมูลของ stroke

จากภาพประกอบ 31 จากกราฟแสดงตัวแปรที่เป็น target(stroke) ที่แสดงให้เห็นว่าข้อมูลที่ได้มาเป็น Imbalanced Dataset โดยข้อมูลมีผู้ที่ไม่เป็นโรคหลอดเลือดสมอง 4,007 คน และผู้ที่เป็นโรคหลอดเลือดสมอง 247 คน

Data Visualization (pair plot)



ภาพประกอบ 32 แสดงการกระจายตัวของข้อมูลและแสดงถึงความสัมพันธ์ของข้อมูลระหว่างสองตัวแปรกับโอกาสการเกิดโรคหลอดเลือดสมอง

จากภาพประกอบ 32 แสดงการกระจายตัวของข้อมูลและความสัมพันธ์ของข้อมูลระหว่างสองตัวแปรกับโอกาสการเกิดโรคหลอดเลือดสมอง โดยเมื่อวิเคราะห์ข้อมูลจากการทำ Visualization เราพบว่า

- รูป A ตัวแปร age พบว่า ผู้ที่มีอายุประมาณ 55 - 85 ปีมีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง (แสดงในภาพ A1)

- รูป A ตัวแปร age กับ รูป B ตัวแปร hypertension พบว่า ผู้ที่ไม่มีโรคความดันสูงและมีอายุอายุประมาณ 55 - 85 ปี มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มผู้ที่มีโรคความดันสูงในทุกช่วงอายุ พบว่าไม่มีความสัมพันธ์กับโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ B1)

- รูป A ตัวแปร age กับ รูป C ตัวแปร heart_disease พบว่า ผู้ที่ไม่มีโรคหัวใจและมีอายุประมาณ 55 - 85 ปีมีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มผู้ที่มีโรคหัวใจในทุกช่วงอายุ พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ C1)

- รูป A ตัวแปร age กับ รูป D ตัวแปร avg_glucose_level พบว่า ผู้ที่ไม่มีระดับน้ำตาลในเลือดสูงประมาณ 55-150 และมีอายุประมาณ 55- 85 ปี พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง และกลุ่มผู้ที่มีระดับน้ำตาลในเลือดสูงประมาณ 200 – 250 และมีอายุประมาณ 65-80 ปี พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ D1)

- รูป A ตัวแปร age กับ รูป E ตัวแปร bmi พบว่า ผู้ที่มีอายุประมาณ 55 – 85 ปี และมีดัชนีมวลกายประมาณ 20 - 40 มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง (แสดงในภาพ E1

- รูป B ตัวแปร hypertension พบว่า ผู้ที่ไม่มีโรคความดันสูงมีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มผู้ที่มีโรคความดันสูง พบว่าไม่มีความสัมพันธ์กับโอกาส (แสดงในภาพ B2)

- รูป B ตัวแปร hypertension กับรูป C ตัวแปร heart_disease พบว่า ผู้ที่ไม่มีโรคหัวใจและไม่มีโรคความดันสูง มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มความสัมพันธ์อื่นๆ ของตัวแปร hypertension กับ ตัวแปร heart_disease พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ B3)

- รูป B ตัวแปร hypertension กับรูป D ตัวแปร avg_glucose_level พบว่าผู้ที่ไม่มีโรคความดันสูงและผู้ที่มีระดับน้ำตาลในเลือดสูงประมาณ 50 – 150 พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มความสัมพันธ์อื่นๆ ของตัวแปร hypertension กับ ตัวแปร avg_glucose_level พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ B4)

- รูป B ตัวแปร hypertension กับรูป E ตัวแปร bmi พบว่าผู้ที่ไม่มีโรคความดันสูงและผู้ที่มีดัชนีมวลกายประมาณ 20 – 40 มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง และในผู้ที่มีโรคความดันสูงและผู้ที่มีดัชนีมวลกายประมาณ 30 – 35 มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง (แสดงในภาพ B5)

- รูป C ตัวแปร heart_disease พบว่า ผู้ที่ไม่มีโรคหัวใจ มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มผู้ที่มีโรคหัวใจ พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ C3)

- รูป C ตัวแปร heart_disease กับรูป D ตัวแปร avg_glucose_level พบว่า ผู้ที่ไม่เป็นโรคหัวใจและมีระดับน้ำตาลในเลือดสูงประมาณ 50 -150 พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มความสัมพันธ์อื่นๆ ของตัวแปร heart_disease กับ ตัวแปร avg_glucose_level พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ C4)

- รูป C ตัวแปร heart_disease กับรูป E ตัวแปร bmi พบว่าผู้ที่ไม่เป็นโรคหัวใจและผู้ที่มีดัชนีมวลกายประมาณ 20 – 40 มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง แต่ในกลุ่มความสัมพันธ์อื่นๆ ของตัวแปร heart_disease กับ ตัวแปร bmi พบว่าไม่มีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ C5)

- รูป D ตัวแปร avg_glucose_level พบว่า ผู้ที่ไม่มีระดับน้ำตาลในเลือดสูงประมาณ 50-250 พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ D4)

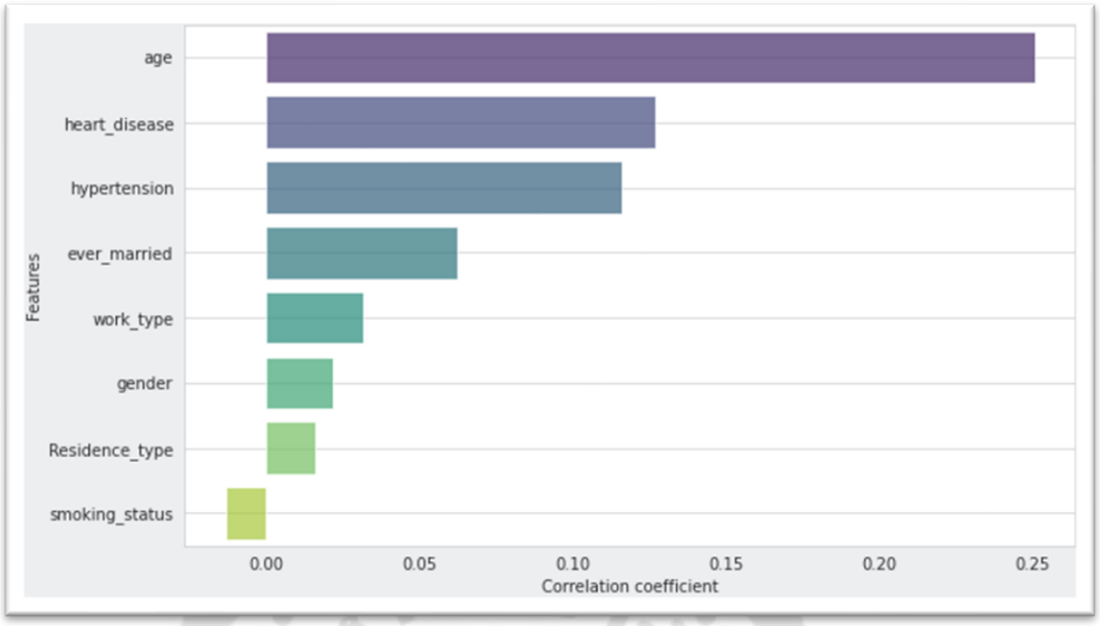
- รูป D ตัวแปร avg_glucose_level กับรูป E ตัวแปร bmi พบว่า ผู้ที่ไม่มีระดับน้ำตาลในเลือดสูงประมาณ 55 – 250 และมีดัชนีมวลกายประมาณ 25 - 45 พบว่ามีความสัมพันธ์กับการมีโอกาสเกิดโรคหลอดเลือดสมอง (แสดงในภาพ D5)

- รูป E ตัวแปร bmi พบว่า ผู้ที่มีดัชนีมวลกายประมาณ 20 - 40 มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง (แสดงในภาพ E5)

จากข้อมูลที่แสดงความสัมพันธ์ของสองตัวแปรต่างๆ สรุปโดยรวม พบว่าผู้ป่วยที่มีอายุประมาณ 55 - 85 ปี ไม่เป็นโรคเบาหวานและมีระดับน้ำตาลในเลือด 55 -150 ไม่เป็นโรคหัวใจ ไม่เป็นโรคความดันสูง และผู้มีดัชนีมวลกายประมาณ 20 - 40 พบว่ามีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง

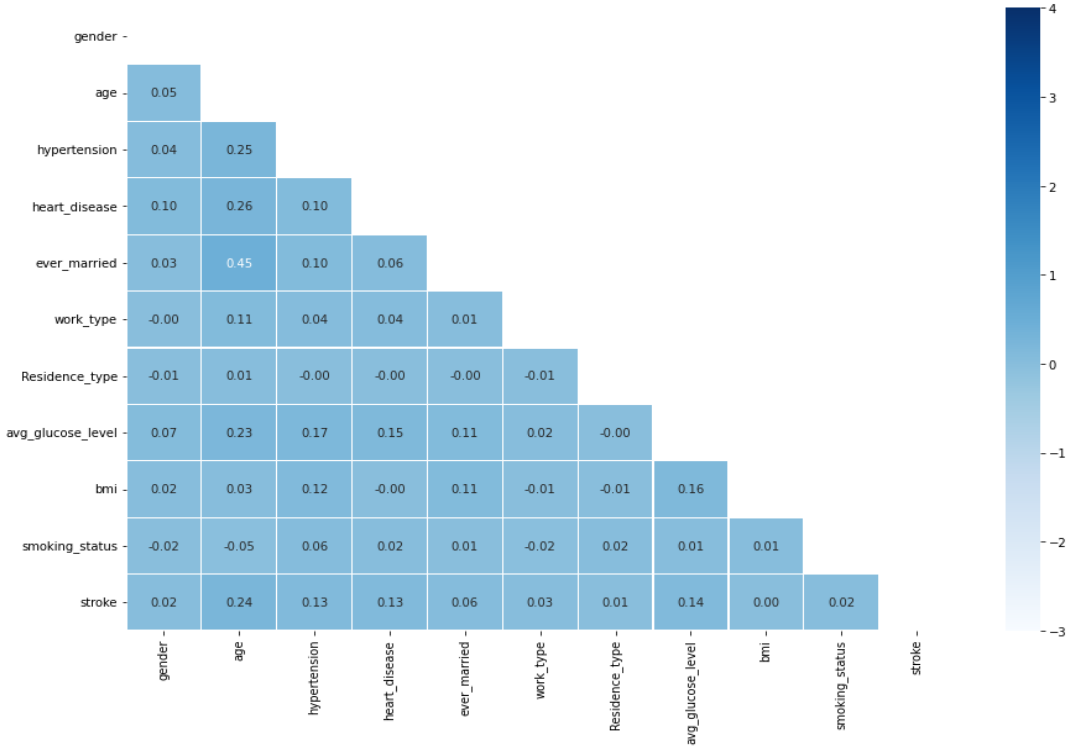
แต่ทั้งนี้ทั้งนั้นข้อมูลเหล่านี้ยังเป็นข้อมูลดิบที่ยังไม่ผ่านขบวนการจัดการกับ imbalanced dataset, data cleaning และ data standardization ซึ่งความสัมพันธ์ระหว่างตัวแปรอาจมีความคลาดเคลื่อนได้

ศึกษาความสัมพันธ์ (correlation) **ระหว่าง** attribute **ต่างๆ**



ภาพประกอบ 33 กราฟแท่งแสดงค่า Correlation Coefficient หรือ ค่าสหสัมพันธ์ระหว่างตัวแปร 2 ตัวที่มีอิทธิพลต่อโอกาสการเกิดโรคหลอดเลือดสมอง

จากภาพประกอบ 33 กราฟแท่งแสดงค่า Correlation Coefficient หรือ ค่าสหสัมพันธ์ เป็นการดูทิศทางความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่มีอิทธิพลต่อโอกาสการเกิดโรคหลอดเลือดสมอง โดยค่าความสัมพันธ์ระหว่างตัวแปรจะอยู่ในมาตราการวัดระดับ Interval หรือ Ratio Scale โดยปกติจะมีค่าอยู่ระหว่าง -1.00 ถึง 1.00



ภาพประกอบ 34 แสดงค่า Correlation ที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง

จากภาพประกอบ 34 ตารางแสดงค่า Correlation ที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมอง โดยค่าความสัมพันธ์ระหว่างตัวแปรจะอยู่ในมาตราการวัดระดับ Interval หรือ Ratio Scale โดยปกติจะมีค่าอยู่ระหว่าง -1.00 ถึง 1.00 จากภาพประกอบ 33 และ 34 แสดงค่า Correlation(Rabiablok et al., 2021) เป็นการแสดงให้เห็นเป็นการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่อยู่ในมาตราการวัดระดับ Interval หรือ Ratio Scale ค่าที่ได้เรียกว่า "สัมประสิทธิ์สหสัมพันธ์" โดยปกติจะมีค่าอยู่ระหว่าง -1.00 ถึง 1.00

- ถ้ามีค่าติดลบหมายความว่า ตัวแปร 2 ตัวมีความสัมพันธ์ในทิศทางตรงกันข้าม
- ถ้ามีค่าเป็นบวกหมายความว่า ตัวแปร 2 ตัวมีความสัมพันธ์ในทิศทางเดียวกัน
- ถ้ามีค่าเป็น 0 หมายความว่าตัวแปร 2 ตัวไม่มีความสัมพันธ์กัน

โดยจากภาพแสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่มีความสัมพันธ์กับโอกาสการเกิดโรคหลอดเลือดสมองเราพบว่าตัวแปรที่มีความสัมพันธ์ต่อการเกิดโรคหลอดเลือดสมอง เรียงลำดับจากตัวแปร ที่มีความสัมพันธ์ต่อโอกาสการเกิดโรคหลอดเลือดสมองจากมากไปน้อย คือ age, heart

disease, hypertension, ever_married, work_type, gender, resident_type, smoking ตามลำดับ

5. **เตรียมข้อมูลให้พร้อมสร้างโมเดล**

5.1 **การทำความสะอาดข้อมูล** (Data Cleaning)

การจัดการกับ Missing values

จากการศึกษานี้เป็นการศึกษาโอกาสการเกิดโรคหลอดเลือดสมองในวัยผู้ใหญ่ซึ่งมีจำนวน 4,073 records และมีค่า missing value ของตัวแปร bmi อยู่ 181 records หรือคิดเป็น 4.25 % ซึ่งกรณีที่ข้อมูลสูญหายเกิดขึ้นแต่ปริมาณการสูญหายไม่เกิน 5% ของข้อมูลทั้งหมด เราจึงตัดหรือลบข้อมูลส่วนที่สูญหายออกทั้ง records

```
#Removing missing values
df.dropna(inplace=True)
```

ภาพประกอบ 35 แสดงการจัดการกับ missing values ของตัวแปร bmi ด้วยการลบแถวที่มีค่า missing value

การใช้ LabelEncoder function **จัดการกับข้อมูลที่เป็นตัวแปรข้อความให้เป็นตัวเลข**

เป็นการแปลงจาก Categorical Feature ขนาด N เป็น Integer Feature ที่มี Range = N คือการแปลงข้อมูลที่เป็นตัวแปรกลุ่ม เช่น เพศ (ชาย หญิง) ให้อยู่ในรูปของตัวเลขในชุดคำสั่ง

Changing Category to Numerical Values

```
[ ] # create encoder for each categorical variable
    label_gender = LabelEncoder()
    label_married = LabelEncoder()
    label_work = LabelEncoder()
    label_residence = LabelEncoder()
    label_smoking = LabelEncoder()
```

```
[ ] clean_data['gender'] = label_gender.fit_transform(clean_data['gender'])
    clean_data['ever_married'] = label_married.fit_transform(clean_data['ever_married'])
    clean_data['work_type']= label_work.fit_transform(clean_data['work_type'])
    clean_data['Residence_type']= label_residence.fit_transform(clean_data['Residence_type'])
    clean_data['smoking_status']= label_smoking.fit_transform(clean_data['smoking_status'])
    with pd.option_context('expand_frame_repr', False):
        print(clean_data.head())
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	67.0	0	1	1	2	1	228.69	36.6	1	1
2	1	80.0	0	1	1	2	0	105.92	32.5	2	1
3	0	49.0	0	0	1	2	1	171.23	34.4	3	1
4	0	79.0	1	0	1	3	0	174.12	24.0	2	1
5	1	81.0	0	0	1	2	1	186.21	29.0	1	1

ภาพประกอบ 36 การใช้ LabelEncoder function จัดการกับ categorical data

จากภาพประกอบ 36 แสดงวิธีการแปลงข้อมูลประเภทข้อความ ‘gender’, ‘ever_married’, ‘work_type’, ‘Residence_type’, ‘smoking_status’ ให้เป็นตัวเลข 0, 1, 2, หรือ 3 ไปเลย โดยการใช้เทคนิคของ Label Encoder function

5.2 Feature Scaling

โดยการทำ standardize หรือ normalize เพื่อให้ข้อมูลแต่ละ feature อยู่ใน scale ที่ใกล้เคียงกัน ก่อนนำเข้าขบวนการ training สร้างโมเดล (Standardize the data before training)

```
[24] from sklearn.preprocessing import StandardScaler
     sc = StandardScaler()
```

```
     X_train.age=sc.fit_transform(X_train.age.values.reshape(-1,1))
     X_test.age=sc.transform(X_test.age.values.reshape(-1,1))
     X_train.bmi=sc.fit_transform(X_train.bmi.values.reshape(-1,1))
     X_test.bmi=sc.transform(X_test.bmi.values.reshape(-1,1))
     X_train.avg_glucose_level=sc.fit_transform(X_train.avg_glucose_level.values.reshape(-1,1))
     X_test.avg_glucose_level=sc.transform(X_test.avg_glucose_level.values.reshape(-1,1))
```

ภาพประกอบ 37 การทำ standardization หรือ normalization

จากภาพประกอบ 37 แสดงการทำ Feature Scaling โดยการทำ standardization เพื่อเป็นการปรับช่วงขอบเขตของข้อมูลชนิดตัวเลข Cardinal แต่ละ Feature (Field) ให้อยู่ในช่วง

เดียวกัน ที่เหมาะกับการนำไปประมวลผลต่อโดยทำให้เข้าสูตรคำนวณได้ง่ายขึ้นด้วยการปรับข้อมูลให้อยู่ในช่วง [-1, 1] ก่อนป้อนให้โมเดลใช้เทรน

5.3 **การจัดการกับปัญหาข้อมูลไม่สมดุลด้วยเทคนิค** Synthetic Minority Oversampling Technique (SMOTE)

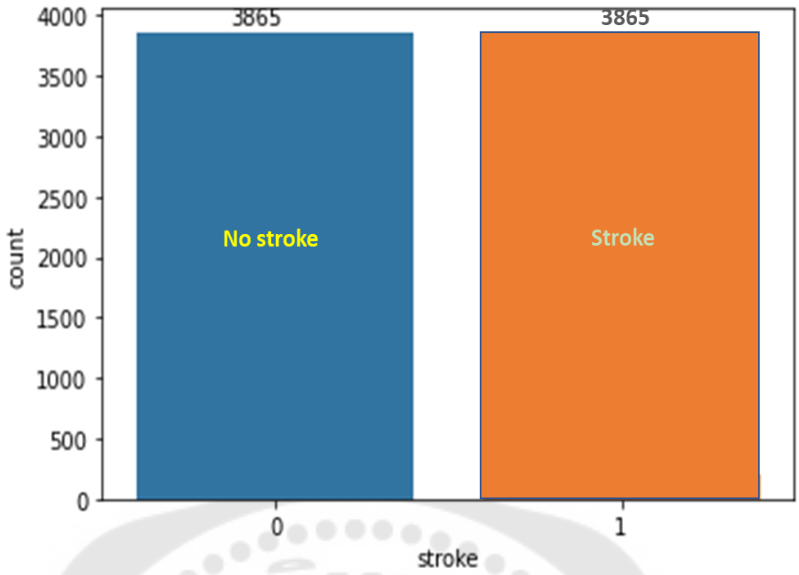
SMOTE เป็นหนึ่งในเทคนิคการสุ่มตัวอย่างที่ได้รับความนิยมมากที่สุดซึ่งพัฒนาโดย Chawla et al (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) ซึ่งต่างจากการสุ่มตัวอย่างเกินขนาดที่ทำซ้ำจากกลุ่มตัวอย่างข้อมูลส่วนน้อย (minority class) แต่เทคนิค SMOTE เป็นการสร้างตัวอย่างตามระยะห่างของแต่ละข้อมูล (โดยปกติใช้ระยะทางแบบ Euclidean distance และกลุ่มข้อมูลส่วนน้อยที่อยู่ใกล้ที่สุด ดังนั้นตัวอย่างที่สร้างขึ้นจึงแตกต่างจากกลุ่มชนกลุ่มน้อยดั้งเดิม เพื่อสร้างความสมดุลของข้อมูล

```
sm = SMOTE()
X_oversampled, y_oversampled = sm.fit_resample(X, y)

 y_oversampled.value_counts()

1    3865
0    3865
Name: stroke, dtype: int64
```

ภาพประกอบ 38 การจัดการ imbalanced data ด้วยเทคนิค

จากภาพประกอบ 38 แสดง Code ในการใช้ SMOTE technique เพื่อจัดการปัญหา imbalanced class ให้เป็น balanced class โดยหลังจากใช้ SMOTE technique ทำให้ Class 0 และ Class 1 มีจำนวนข้อมูลเป็น 3865 records เท่ากัน

Synthetic Minority Oversampling Technique : SMOTE



ภาพประกอบ 39 กราฟแสดงจำนวนของข้อมูลที่มีความสมดุลของ target attribute (stroke) หลังจากใช้เทคนิค Synthetic Minority Oversampling Technique: SMOTE

6. การประเมินประสิทธิภาพของแต่ละโมเดล

การวัดผลและเปรียบเทียบ performance ของโมเดลเราใช้ confusion matrix, accuracy, sensitivity, specificity, f1-score และ พื้นที่ใต้กราฟ AUC

บทที่ 4
ผลการดำเนินการวิจัย

การวิจัยนี้มีประโยชน์ที่เป็นเครื่องมือที่ช่วยสนับสนุนในขบวนการการตัดสินใจทางการแพทย์ โดยผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนต่างๆ จนได้โมเดลที่ใช้ในการช่วยประเมินความเสี่ยงของโอกาสการเกิดโรคหลอดเลือดสมองที่มีประสิทธิภาพตามวัตถุประสงค์ที่ได้กำหนดไว้ ได้ดังนี้

1. ผลลัพธ์ของการศึกษาและวิเคราะห์
2. ผลการทดสอบสมมติฐานการวิจัย

1. ผลลัพธ์ของการศึกษาและวิเคราะห์

จากการศึกษานี้เราเลือกใช้โมเดลป่าสุ่ม LR , และ SVM และเทคนิคต่างๆในการในการจัดการ กับข้อมูลเพื่อหาโมเดลที่ดีที่สุดในการประเมินความเสี่ยงของโอกาสการเกิดโรคหลอดเลือดสมอง โดยผลการศึกษามีดังนี้

1.1 การสร้างโมเดลกับข้อมูลที่เป็นข้อมูลที่ไม่สมดุล

จากการเลือกใช้โมเดลป่าสุ่ม LR , และ SVM กับข้อมูล imbalanced dataset เราได้ผลการศึกษาดังแสดงในภาพประกอบ 40, 42 และ 44

1.1.1 การใช้อัลกอริทึม SVM ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล

จาก Confusion matrix จากภาพประกอบ 41 ของโมเดล SVM เราสามารถคำนวนค่า Precision, Recall, F1-score รวมทั้งค่าเฉลี่ยของ Precision, Recall, F1-score ทั้งแบบ Macro avg และ Weight avg ด้วย Function classification_report ซึ่งค่า Macro avg และ Weight avg สามารถคำนวณได้ค่าดังต่างๆ ดังนี้

Support Vector Machine Model

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1160
1	0.00	0.00	0.00	62
accuracy			0.95	1222
macro avg	0.47	0.50	0.49	1222
weighted avg	0.90	0.95	0.92	1222

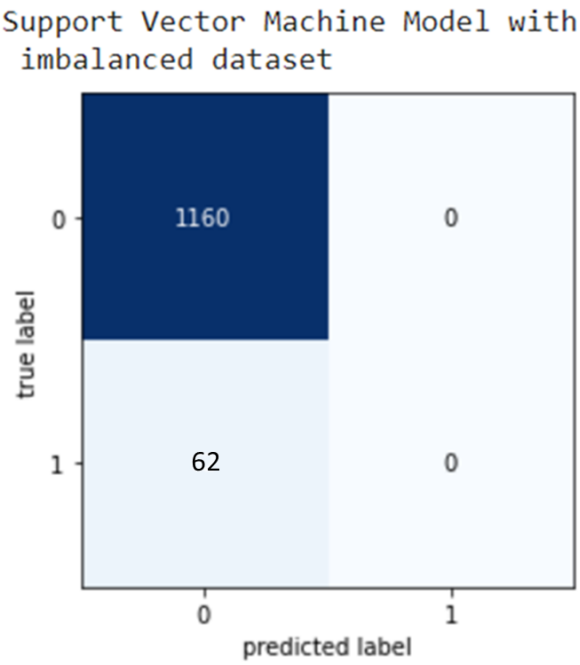
ภาพประกอบ 40 Classification report ของโมเดล SVM ที่สร้างมาจากข้อมูลที่ไม่สมดุล

จากภาพประกอบ 40 เราสามารถคำนวนค่าประสิทธิภาพของโมเดล SVM ได้ดังนี้ จากข้อมูลนี้เราพบว่าเป็นข้อมูลแบบ imbalanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.95 ให้ค่า macro avg f1-score เท่ากับ 0.49 และค่า weighted avg f1-score เท่ากับ 0.92 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดล LR โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้

ทำนายคลาส 0 (No stroke, Majority class) ได้ค่า recall 1.00 ค่า precision 0.95 และค่า f1-score 0.97

ทำนายคลาส 1(Stroke, Minority class) ได้ค่า recall 0.00 ค่า precision 0.00 และค่า f1-score 0.00

การที่คลาส 1 (Stroke) มีจำนวนของข้อมูลน้อยกว่า คลาส 0 มาก จึงทำให้โมเดลที่ได้มามีการทำนายข้อมูลโดยมีแนวโน้มที่จะลำเอียงไปทางคลาสที่มีจำนวนมากกว่า ดังนั้นโมเดลนี้จึงไม่เหมาะที่จะนำไปใช้ในการทำนายโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง



ภาพประกอบ 41 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล SVM ที่สร้างมาจากข้อมูลที่ไม่สมดุล

จากภาพประกอบ 41 Confusion matrix แสดงค่าประสิทธิภาพของโมเดล SVM โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 0

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 1160

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 0

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 62

1.1.2 การใช้อัลกอริทึม LR ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล

จาก Confusion matrix จากภาพประกอบ 43 ของโมเดล LR เราสามารถคำนวนค่า Precision, Recall, F1-score รวมทั้งค่าเฉลี่ยของ Precision, Recall, F1-score ทั้งแบบ Macro avg และ Weight avg ด้วย Function classification_report ซึ่งค่า Macro avg และ Weight avg สามารถคำนวณได้ค่าดังต่างๆ ดังนี้

Logistic Reg Model

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1160
1	0.00	0.00	0.00	62
accuracy			0.95	1222
macro avg	0.47	0.50	0.49	1222
weighted avg	0.90	0.95	0.92	1222

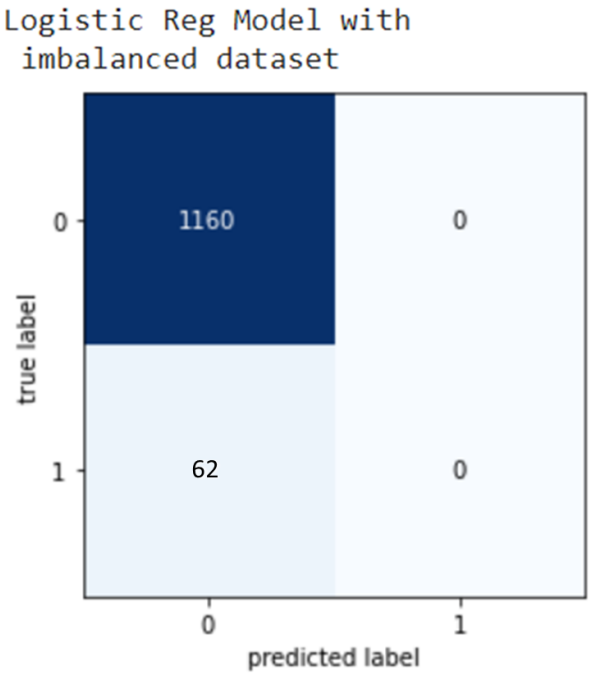
ภาพประกอบ 42 แสดงประสิทธิภาพของโมเดล LR ที่ได้จากการใช้ข้อมูลจากข้อมูลที่ไม่สมดุล

จากภาพประกอบ 42 เราสามารถคำนวนค่าประสิทธิภาพของโมเดล *LR* ได้ดังนี้ จากข้อมูลนี้เราพบว่าเป็นข้อมูลแบบ imbalanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.95 ให้ค่า macro avg f1-score เท่ากับ 0.49 และค่า weighted avg f1-score เท่ากับ 0.92 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดล LR โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้

ทำนายคลาส 0 (No stroke, Majority class) ได้ค่า recall 1.00 ค่า precision 0.95 และค่า f1-score 0.97

ทำนายคลาส 1(Stroke, Minority class) ได้ค่า recall 0.00 ค่า precision 0.00 และค่า f1-score 0.00

การที่คลาส 1(Stroke) มีจำนวนของข้อมูลน้อยกว่า คลาส 0 มาก จึงทำให้โมเดลที่ได้มามีการทำนายข้อมูลโดยมีแนวโน้มที่จะลำเอียงไปทางคลาสที่มีจำนวนมากกว่า ดังนั้นโมเดลนี้จึงไม่เหมาะที่จะนำไปใช้ในการทำนายโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง



ภาพประกอบ 43 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่ไม่สมดุล

จากภาพประกอบ 43 Confusion matrix แสดงค่าประสิทธิภาพของโมเดล LR โดยมีค่าของ

TP = สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่า จริง (stroke) และสิ่งที่เกิดขึ้น ก็คือ จริง (stroke) มีค่าเท่ากับ 0

TN = สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง (no stroke) และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง (no stroke) มีค่าเท่ากับ 1160

FP = สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง (stroke) แต่สิ่งที่เกิดขึ้น คือ ไม่จริง (no stroke) มีค่าเท่ากับ 0

FN = สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง (no stroke) แต่สิ่งที่เกิดขึ้น คือ จริง (stroke) มีค่าเท่ากับ 62

1.1.3 การใช้อัลกอริทึมป่าสุ่ม ในการสร้างโมเดลกับข้อมูลที่ไม่สมดุล

จาก Confusion matrix จากภาพประกอบ 45 ของโมเดลป่าสุ่ม เราสามารถคำนวนค่า Precision, Recall, F1-score รวมทั้งค่าเฉลี่ยของ Precision, Recall, F1-score ทั้งแบบ Macro avg และ Weight avg ด้วย Function classification_report ซึ่งค่า Macro avg และ Weight avg สามารถคำนวณได้ค่าดังต่างๆ ดังนี้

Random Forest Classifier Model

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1160
1	0.00	0.00	0.00	62
accuracy			0.95	1222
macro avg	0.47	0.50	0.49	1222
weighted avg	0.90	0.95	0.92	1222

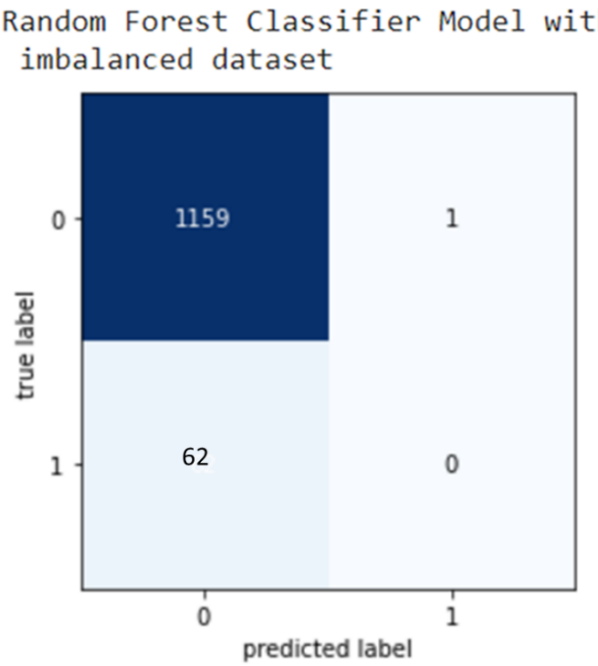
ภาพประกอบ 44 แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากการใช้ข้อมูลข้อมูลที่ไม่สมดุล

จากภาพประกอบ 44 เราสามารถคำนวนค่าประสิทธิภาพของโมเดลป่าสุ่ม ได้ดังนี้ จากข้อมูลนี้เราพบว่าเป็นข้อมูลแบบ imbalanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.95 ให้ค่า macro avg f1-score เท่ากับ 0.49 และค่า weighted avg f1-score เท่ากับ 0.92 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดลป่าสุ่ม โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้

ทำนายคลาส 0 (No stroke, Majority class) ได้ค่า recall 1.00 ค่า precision 0.95 และค่า f1-score 0.97

ทำนายคลาส 1(Stroke, Minority class) ได้ค่า recall 0.00 ค่า precision 0.00 และค่า f1-score 0.00

การที่คลาส 1(Stroke) มีจำนวนของข้อมูลน้อยกว่า คลาส 0 มาก จึงทำให้โมเดลที่ได้มามีการทำนายข้อมูลโดยมีแนวโน้มที่จะลำเอียงไปทางคลาสที่มีจำนวนมากกว่า ดังนั้นโมเดลนี้จึงไม่เหมาะที่จะนำไปใช้ในการทำนายโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง



ภาพประกอบ 45 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลข้อมูลที่ไม่สมดุล

จากภาพประกอบ 45 Confusion matrix แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 0

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 1151

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆ แล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 1

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆ แล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 62

ตาราง 6 แสดงการเปรียบเทียบประสิทธิภาพของโมเดลที่ได้จากข้อมูลที่ไม่สมดุล

Model performance with an imbalanced dataset			
Models	precision	recall	F1-score
RandomForest (Accuracy = 0.95)			
0	0.95	1.00	0.97
1	0.00	0.00	0.00
Logistic regression (Accuracy = 0.95)			
0	0.95	1.00	0.97
1	0.00	0.00	0.00
Support Vector Machine (Accuracy = 0.95)			
0	0.95	1.00	0.97
1	0.00	0.00	0.00

จากตาราง 6 แสดงประสิทธิภาพของโมเดลป่าสุ่ม, LR, SVM ที่ได้จากการใช้ข้อมูล imbalanced dataset โดยพบว่ามีประสิทธิภาพพอๆกันโดยมีค่า accuracy เท่ากับ 0.94 ค่า recall เท่ากับ 0.00 ค่า precision เท่ากับ 0.00 และค่า f1-score เท่ากับ 0.00

1.2 การสร้างโมเดลจากข้อมูลที่เป็นข้อมูลที่สมดุล

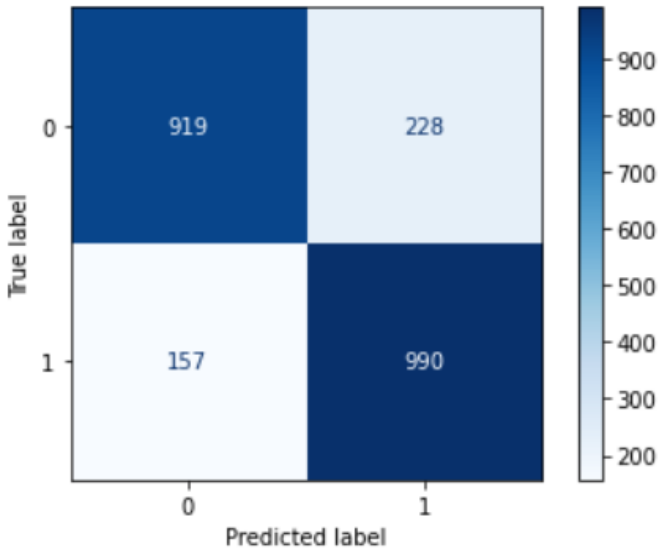
การเลือกใช้โมเดลป่าสุ่ม, LR, และ SVM กับข้อมูล balanced dataset ที่ได้จากการใช้เทคนิค SMOTE และการทำการประเมินประสิทธิภาพของโมเดลด้วยเทคนิค 10 Cross validation โดยเราได้ผลการศึกษาดังแสดงในภาพประกอบ 46, 49 และ 52

1.2.1 การใช้อัลกอริทึม SVM ในการสร้างโมเดลจากข้อมูล ที่สมดุล

SVC	precision	recall	f1-score	support
0	0.85	0.80	0.83	1147
1	0.81	0.86	0.84	1147
accuracy			0.83	2294
macro avg	0.83	0.83	0.83	2294
weighted avg	0.83	0.83	0.83	2294

ภาพประกอบ 46 Classification report ของโมเดล SVM ที่สร้างมาจากข้อมูลที่สมดุล

จากภาพประกอบ 46 แสดงประสิทธิภาพของโมเดล SVM ที่ได้จากการใช้ข้อมูล balanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.83 ให้ค่า macro avg f1-score เท่ากับ 0.83 และค่า weighted avg f1-score เท่ากับ 0.83 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดล SVM โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้ ทำนายคลาส 0 (No stroke) ได้ค่า recall 0.80 ค่า precision 0.85 และค่า f1-score 0.83 ทำนายคลาส 1(Stroke) ได้ค่า recall 0.86 ค่า precision 0.81 และค่า f1-score 0.84



ภาพประกอบ 47 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดล SVM ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล

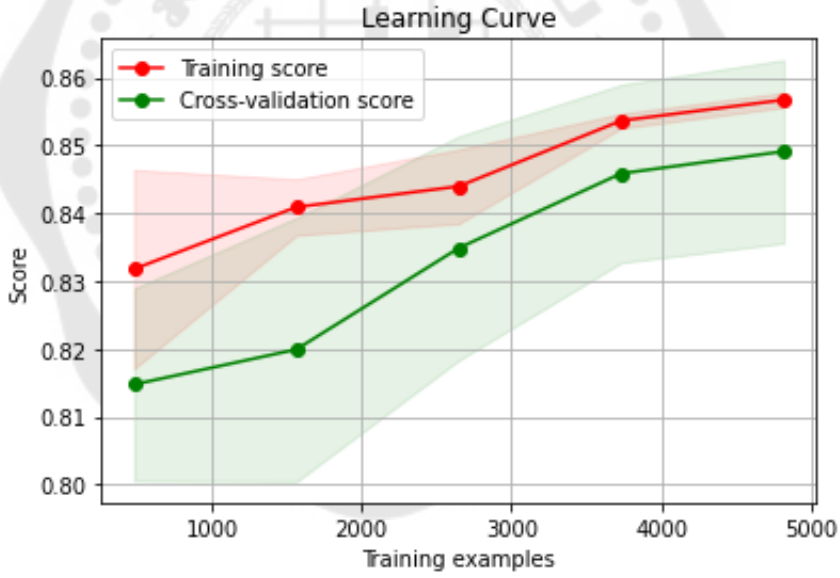
จากภาพประกอบ 47 Confusion matrix แสดงค่าประสิทธิภาพของโมเดล SVM โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 990

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 919

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 228

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 157



ภาพประกอบ 48 กราฟแสดงประสิทธิภาพของโมเดล SVM ที่ได้จากข้อมูลข้อมูลที่สมดุล

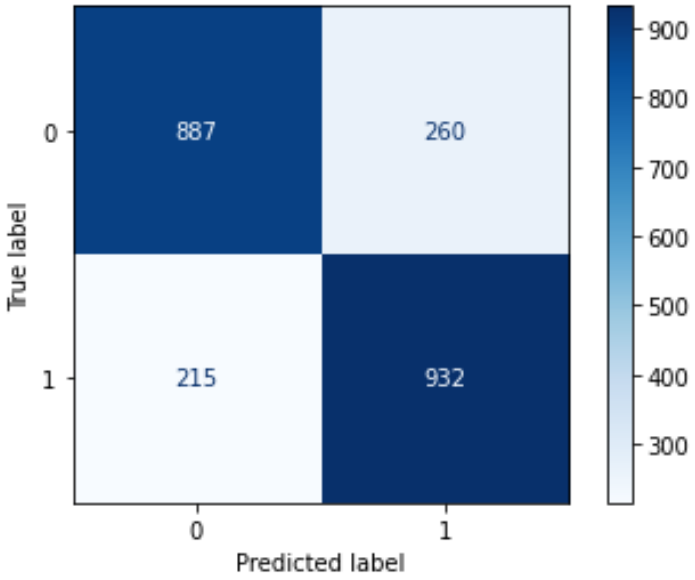
จากภาพประกอบ 48 กราฟแสดงประสิทธิภาพของโมเดล SVM ที่ได้จาก balanced dataset โดยใช้เทคนิค SMOTE และ Cross validation โดยเปรียบเทียบการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validationโดยพบว่าเมื่อจำนวน training samples มากขึ้นการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation มีแนวโน้มการเรียนรู้ได้ดีขึ้น

1.2.2 การใช้อัลกอริทึม LR ในการสร้างโมเดลจากข้อมูลที่สมดุล

Logistic Regression	precision	recall	f1-score	support
0	0.80	0.77	0.79	1147
1	0.78	0.81	0.80	1147
accuracy			0.79	2294
macro avg	0.79	0.79	0.79	2294
weighted avg	0.79	0.79	0.79	2294

ภาพประกอบ 49 Classification report ของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล

จากภาพประกอบ 49 แสดงประสิทธิภาพของโมเดล LR ที่ได้จากการใช้ข้อมูล balanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.79 ให้ค่า macro avg f1-score เท่ากับ 0.79 และค่า weighted avg f1-score เท่ากับ 0.79 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดล SVM โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้ ทำนายคลาส 0 (No stroke) ได้ค่า recall 0.77 ค่า precision 0.80 และค่า f1-score 0.79 ทำนายคลาส 1(Stroke) ได้ค่า recall 0.81 ค่า precision 0.78 และค่า f1-score 0.80



ภาพประกอบ 50 Confusion Matrix แสดงประสิทธิภาพของโมเดล LR ที่สร้างมาจากข้อมูลข้อมูลที่สมดุล

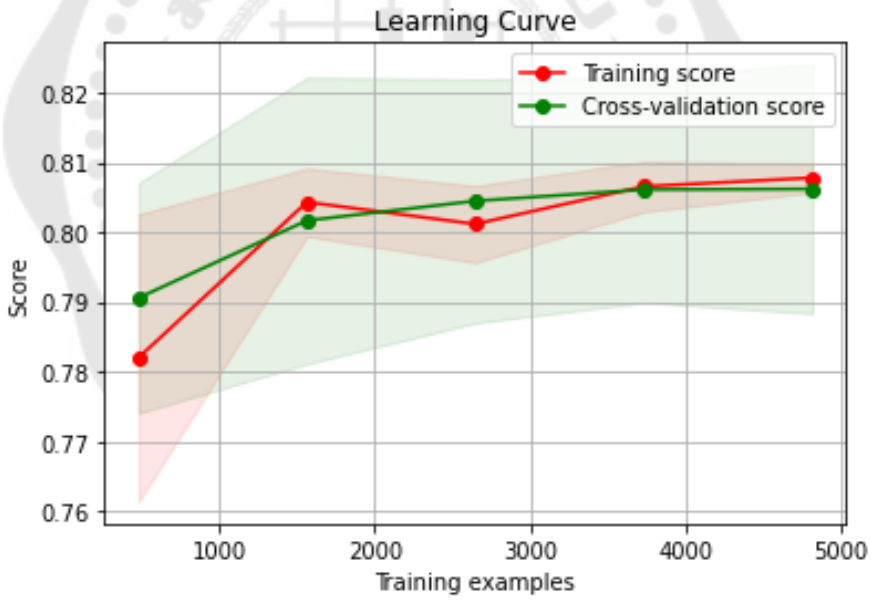
จากภาพประกอบ 50 Confusion matrix แสดงค่าประสิทธิภาพของโมเดล LR โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 932

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 887

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 260

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 215



ภาพประกอบ 51 กราฟแสดงประสิทธิภาพของโมเดล LR ที่ได้จากข้อมูลข้อมูลที่สมดุล

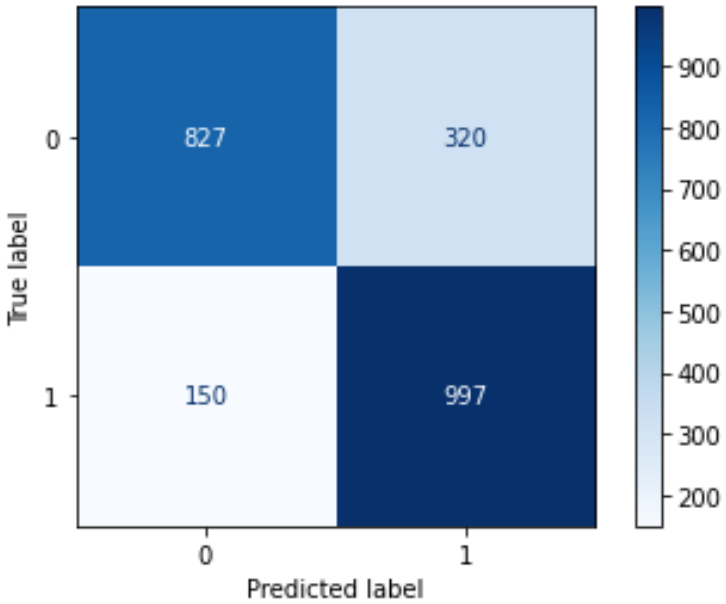
จากภาพประกอบ 51 กราฟแสดงประสิทธิภาพของโมเดล LR ที่ได้จาก balanced dataset โดยใช้เทคนิค SMOTE และ Cross validation โดยเปรียบเทียบการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation โดยพบว่าเมื่อจำนวน training samples มากขึ้นการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation มีแนวโน้มการเรียนรู้ได้ดีขึ้น

1.2.3 การใช้อัลกอริทึมป่าสุ่ม ในการสร้างโมเดลจากข้อมูลที่สมดุล

RandomForest Classifier	precision	recall	f1-score	support
0	0.85	0.72	0.78	1147
1	0.76	0.87	0.81	1147
accuracy			0.80	2294
macro avg	0.80	0.80	0.79	2294
weighted avg	0.80	0.80	0.79	2294

ภาพประกอบ 52 Classification report ของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลที่สมดุล

จากภาพประกอบ 52 แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากการใช้ข้อมูล balanced dataset ที่ได้ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.80 ให้ค่า macro avg f1-score เท่ากับ 0.79 และค่า weighted avg f1-score เท่ากับ 0.79 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดล SVM โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้ ทำนายคลาส 0 (No stroke) ได้ค่า recall 0.72 ค่า precision 0.85 และค่า f1-score 0.78 ทำนายคลาส 1(Stroke) ได้ค่า recall 0.87 ค่า precision 0.76 และค่า f1-score 0.81



ภาพประกอบ 53 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่สร้างมาจากข้อมูลข้อมูลข้อมูลที่สมดุล

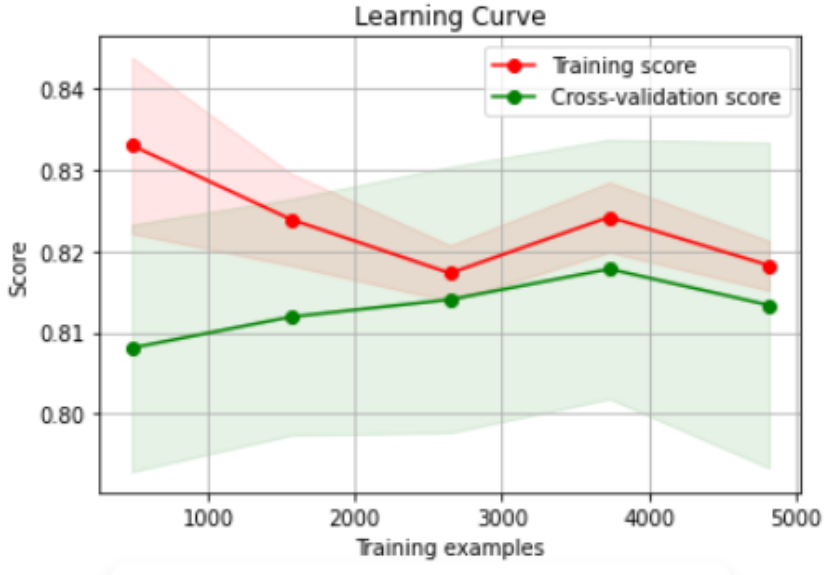
จากภาพประกอบ 53 Confusion matrix แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 997

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 827

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆ แล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 320

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆ แล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 150



ภาพประกอบ 54 กราฟแสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากข้อมูลข้อมูลที่สมดุล โดยเปรียบเทียบการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation

จากภาพประกอบ 54 กราฟแสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จาก balanced dataset โดยใช้เทคนิค SMOTE และ Cross validation โดยเปรียบเทียบการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation โดยพบว่าเมื่อจำนวน training samples มาก

ขึ้นการเรียนรู้ของโมเดลที่ได้จาก training dataset กับข้อมูล Cross validation มีแนวโน้มการเรียนรู้ได้ดีขึ้น

ตาราง 7 การเปรียบเทียบประสิทธิภาพของโมเดลป่าสุ่ม, LR, SVM ที่ได้จากการใช้ข้อมูลที่สมดุล

Model performance with a balanced dataset after using SMOTE			
Models	precision	recall	F1-score
RandomForest (Accuracy = 0.79)			
0	0.84	0.71	0.77
1	0.75	0.86	0.80
Logistic regression (Accuracy = 0.79)			
0	0.80	0.77	0.79
1	0.78	0.81	0.80
Support Vector Machine (Accuracy = 0.83)			
0	0.85	0.80	0.83
1	0.81	0.86	0.84

จากตาราง 7 แสดงประสิทธิภาพของโมเดลป่าสุ่ม, LR, SVM ที่ได้จากการใช้ข้อมูลที่สมดุล โดยพบว่ามีประสิทธิภาพพอๆ กันโดยมีค่า accuracy, precision และค่า f1-score ใกล้เคียงกัน

Hyperparameters Tuning (**โดยใช้** GridSearchCV)

ตาราง 8 แสดง Best score จากการ fine turning ด้วย GridSearchCV(CV = 10) ตัวพารามิเตอร์ต่างๆของโมเดลป่าสุ่ม, LR, และ SVM

Model	Parameters	Best parameters	Best score
RandomForest	{'n_estimators':[100,150,200,250], 'criterion':['gini','entropy'],}	{'criterion': 'gini', 'n_estimators': 250}	93.572
Logistic regression	{"penalty": ['l1', 'l2'], 'C': [0.001, 0.01, 0.025,0.05]}	{'C': 0.05, 'penalty': 'l2'}	79.901

ตาราง 8 (ต่อ)

Model	Parameters	Best parameters	Best score
Support Vector Machine	{'C':[0.5,0.75,1, 1.5], 'kernel':['linear', 'rbf']}	{'C': 1.5, 'kernel': 'rbf'}	84.343

จากตาราง 8 แสดงการ fine turning ตัวพารามิเตอร์ต่างๆของโมเดลป่าสุ่ม, LR และ SVM และผลของประสิทธิภาพของโมเดลโดยดูจากค่า Best score (accuracy) โดยเราได้โมเดลป่าสุ่ม เป็นโมเดลที่มีประสิทธิภาพดีที่สุดโดยมี พารามิเตอร์ที่ดีที่สุดที่ใช้คือ Criterion เป็น ฟังก์ชันที่ใช้ในการวัดประสิทธิภาพของการแยกโหนดของ Decision Tree ใช้ Entropy (Information Gain)

max_features คือ ค่าที่กำหนดจำนวนของฟีเจอร์ที่ Decision Tree แต่ละต้นจะสามารถใช้ในการสร้างโมเดล กำหนดแบบ *auto*

n_estimators คือ จำนวน Decision Tree ที่จะใช้ในโมเดลป่าสุ่ม คือ 200

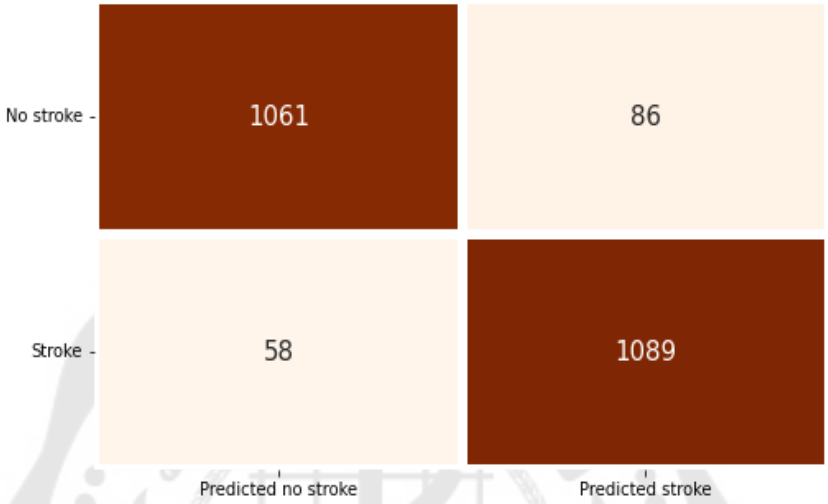
RandomForestClassifier(criterion='entropy', n_estimators=200) Model

	precision	recall	f1-score	support
0	0.95	0.93	0.94	1147
1	0.93	0.95	0.94	1147
accuracy			0.94	2294
macro avg	0.94	0.94	0.94	2294
weighted avg	0.94	0.94	0.94	2294

ภาพประกอบ 55 Classification report หลังจาก fine turning ด้วย GridSearchCV ของโมเดลป่าสุ่ม ที่สร้างมาจาก balanced dataset

จากภาพประกอบ 55 แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จากการใช้ข้อมูล balanced dataset และทำการ fine turning ด้วย GridSearchCV ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.94 ให้ค่า macro avg f1-score เท่ากับ 0.94 และค่า weighted avg f1-score เท่ากับ 0.94 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดลป่าสุ่ม โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้ ทำนายคลาส 0 (No stroke) ได้ค่า recall 0.93 ค่า precision 0.95 และค่า f1-score 0.94 ทำนายคลาส 1(Stroke) ได้ค่า recall 0.95 ค่า precision 0.93 และค่า f1-score 0.94

โมเดลป่าสุ่ม ที่ผ่านการทำ fine turning ด้วย GridSearchCV พบว่ามีประสิทธิภาพในการทำนายข้อมูลดีที่สุดเมื่อเทียบกับโมเดลอื่นๆ ดังนั้นโมเดลนี้จึงมีความเหมาะสมที่จะนำไปใช้ในการทำนายโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองได้มากที่สุด



ภาพประกอบ 56 Confusion Matrix ที่แสดงประสิทธิภาพของโมเดลป่าสุ่ม หลังจาก fine turning ด้วย GridSearchCV

จากภาพประกอบ 56 Confusion matrix แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม โดยมีค่าของ

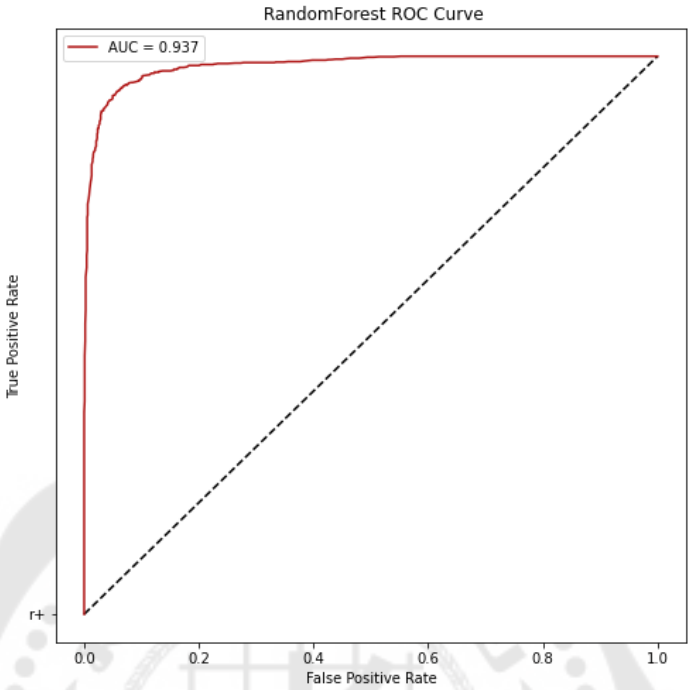
TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 1089

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 1061

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 86

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 58

พื้นที่ใต้กราฟของโมเดลป่าสุ่ม หลังจาก fine turning **ด้วย** GridSearchCV



ภาพประกอบ 57 กราฟค่าพื้นที่ใต้กราฟ AUC (area under curve) ของโมเดลป่าสุ่ม หลังจาก fine turning ด้วย GridSearchCV

จากภาพประกอบ 57 แสดงค่าพื้นที่ใต้กราฟ AUC (area under curve) ที่บอกถึงประสิทธิภาพของโมเดลป่าสุ่ม ที่สามารถแยกกลุ่มคนที่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองและกลุ่มคนที่ไม่มีความเสี่ยงต่อการต่อการเกิดโรคหลอดเลือดสมอง โดยมีจุด cut-off ของโมเดลเท่ากับ 0.937 เพื่อให้การแบ่งคนที่เป็นโรคกับไม่เป็นโรคของเรามีความถูกต้องมากที่สุด และผิดพลาดน้อยที่สุด

โมเดลป่าสุ่มกับกับไฮเปอร์พารามิเตอร์ที่ดีที่สุด

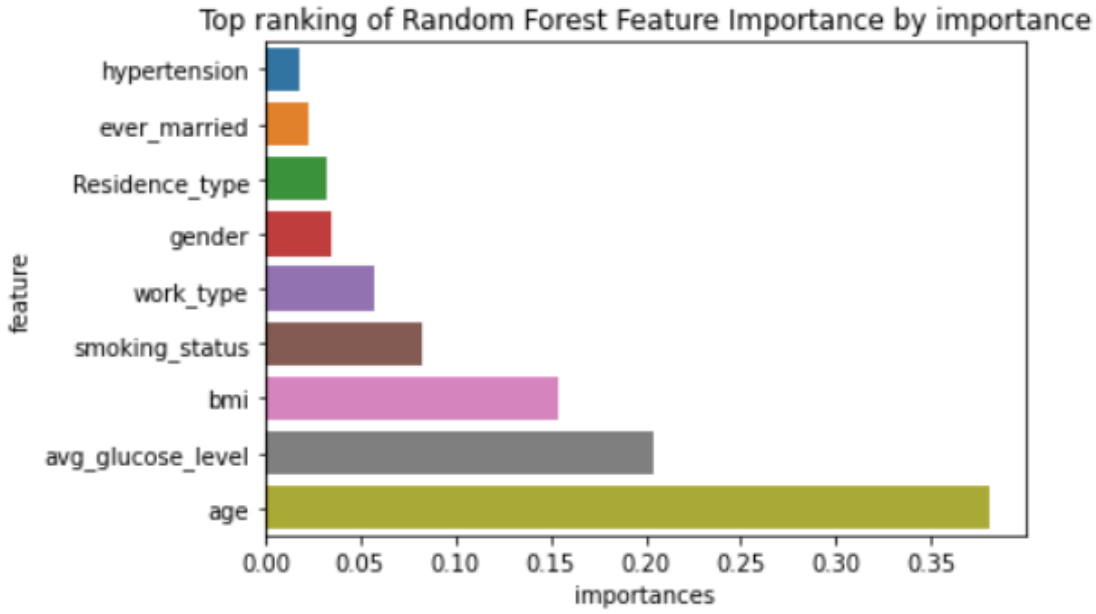
สรุปค่าaccuracy สูงสุดของโมเดลป่าสุ่มกับไฮเปอร์พารามิเตอร์ที่ดีที่สุด หลังจากการ fine turning โมเดลป่าสุ่ม, LR, SVM ด้วย GridSearchCV (CV = 10) แล้วพบว่าโมเดลของป่าสุ่ม ให้ประสิทธิภาพของโมเดลที่ดีที่สุดในการศึกษานี้ ดังที่แสดงข้อมูลในตาราง 9

ตาราง 9 แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม หลังการทำ hyperparameter tuning

Model RandomForest (Accuracy = 0.94, specificity : 0.93)	precision	recall	f1-score
0 (No stroke)	0.95	0.93	0.94
1 (Stroke)	0.93	0.95	0.94
The best parameters	criterion: entropy, n_estimators: 200		

จากตาราง 9 แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม หลังการทำ hyperparameter tuning แล้วโดยพบว่าเป็นโมเดลที่มีประสิทธิภาพดีที่สุดในการศึกษานี้ โดยให้ค่า ค่าความแม่นยำเท่ากับ 0.94 ค่าความเที่ยงตรงเท่ากับ 0.93 ค่าความไวเท่ากับ 0.95 ความจำเพาะเท่ากับ 0.93 ค่า f1-score เท่ากับ 0.94 และค่าพื้นที่ใต้กราฟเท่ากับ 0.94

ความสำคัญของ Feature ต่างๆของโมเดลป่าสุ่ม



ภาพประกอบ 58 แสดง Feature importance ของโมเดลป่าสุ่ม

จากภาพประกอบ 58 แสดง Feature importance ของโมเดลป่าสุ่ม โดยเรียงจากค่ามากไปค่าน้อยมี age, avg_glucose_level, bmi, smoking_status, work_type, gender, Residence_type, ever_married and hypertension โดยมีค่า importance value ดังที่แสดงต่อไปนี้

0.381, 0.204, 0.154, 0.083, 0.057, 0.035, 0.033, 0.023, 0.018, 0.014 ตามลำดับ

2. ผลการทดสอบสมมติฐานการวิจัย

ตาราง 10 แสดงการเปรียบเทียบประสิทธิภาพของแต่ละโมเดลที่ได้จากข้อมูลที่สมดุลและข้อมูลที่ไม่สมดุล

Model performance with an imbalanced dataset				Model performance with a balanced dataset after using SMOTE			
Models	precision	recall	F1-score	Models	precision	recall	F1-score
RandomForest (Accuracy = 0.95)				RandomForest (Accuracy = 0.79)			
0	0.95	1.00	0.97	0	0.84	0.71	0.77
1	0.00	0.00	0.00	1	0.75	0.86	0.80
Logistic regression (Accuracy = 0.95)				Logistic regression (Accuracy = 0.79)			
0	0.95	1.00	0.97	0	0.80	0.77	0.79
1	0.00	0.00	0.00	1	0.78	0.81	0.80
Support Vector Machine (Accuracy = 0.95)				Support Vector Machine (Accuracy = 0.83)			
0	0.95	1.00	0.97	0	0.85	0.80	0.83
1	0.00	0.00	0.00	1	0.81	0.86	0.84

จากตาราง 10 แสดงการเปรียบเทียบประสิทธิภาพของโมเดล ระหว่างโมเดลที่ได้จากข้อมูลที่เป็น imbalanced dataset และ balanced dataset (แสดงประสิทธิภาพของโมเดล LR, SVM และ โมเดลป่าสุ่ม)

1. **การใช้เทคนิค** SMOTE ในการจัดการกับข้อมูลที่ไม่สมดุล จะช่วยเพิ่มประสิทธิภาพ ของโมเดลในการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง โดยจากตาราง 10 พบว่า การทำนายโอกาสการเกิดโรคหลอดเลือดสมอง (minority class) ของโมเดลที่ได้จากข้อมูลที่ไม่สมดุล ไม่สามารถทำนายค่าได้ถูกเลยถึงแม้ว่าจะให้ค่า accuracy โดยรวมของโมเดลโดยรวมเท่ากับ 95% ก็ตาม ส่วนประสิทธิภาพของโมเดลที่ได้จากข้อมูลที่สมดุล ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 79% ถึงจะได้ค่าน้อยกว่าที่ได้จากข้อมูลที่ไม่สมดุล แต่กับได้ค่า recall ที่สูงมากกว่าและสามารถทำนายทั้งสองคลาสได้เป็นอย่างดี ดังนั้นโมเดลที่ได้จากข้อมูลที่สมดุล ให้ประสิทธิภาพที่ดีกว่าเพราะว่าสามารถฟิตเข้ากับข้อมูลอื่นๆ ดีกว่า เป็นไปตามสมมุติฐานที่ตั้งไว้

	Model Name	Feature Scaling	Accuracy	Recall	Precision	F1
0	SVC	Raw	0.737576	0.778553	0.719581	0.747906
1	SVC	Normalization	0.804708	0.807323	0.803122	0.805217
2	SVC	Standardization	0.832171	0.863121	0.812808	0.837209
3	Logistic Regression	Raw	0.792502	0.813426	0.780753	0.796755
4	Logistic Regression	Normalization	0.793374	0.814298	0.781590	0.797609
5	Logistic Regression	Standardization	0.792938	0.812554	0.781879	0.796922
6	RandomForest Classifier	Raw	0.789451	0.863993	0.751897	0.804057
7	RandomForest Classifier	Normalization	0.789451	0.863993	0.751897	0.804057
8	RandomForest Classifier	Standardization	0.789451	0.863993	0.751897	0.804057

ภาพประกอบ 59 แสดงการเปรียบเทียบประสิทธิภาพของโมเดล ระหว่างโมเดลที่ได้จากข้อมูลที่เป็น Raw dataset, Normalization dataset และ Standardization dataset (แสดงประสิทธิภาพของโมเดล LR, SVM และ โมเดลป่าสุ่ม)

2. **การใช้** Feature Engineering ก่อนการนำข้อมูลมาใช้ในการสร้างโมเดล จะช่วยเพิ่มประสิทธิภาพของโมเดลจากการเรียนรู้ของเครื่อง จากภาพประกอบ 59 เราพบว่าการทำ Feature Engineering โดยการทำ Normalization dataset หรือ Standardization ก่อนการสร้างโมเดล จะทำให้ช่วยเพิ่มประสิทธิภาพของโมเดลป่าสุ่ม, LR และ SVM แต่ในโมเดลป่าสุ่ม พบว่าการทำและไม่ทำ Feature Engineering จะไม่มีผลต่อประสิทธิภาพของโมเดล เพราะโมเดลป่าสุ่มเป็นโมเดลประเภท tree-based model ที่ต้องการทำ partitioning แต่การทำ Feature Engineering จะมีผลต่อโมเดลประเภท distance based เช่น โมเดล LR, SVM

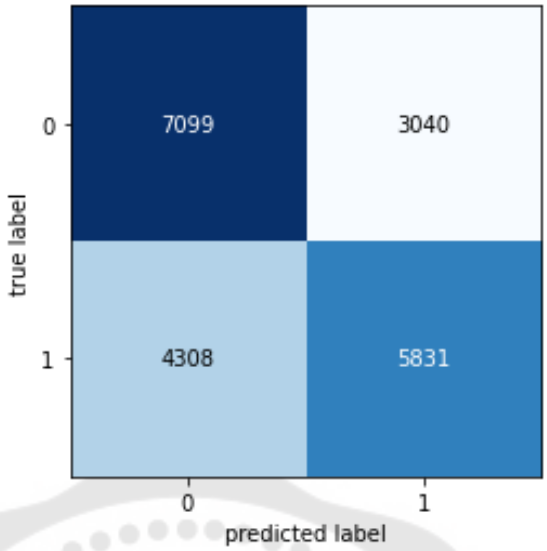
3. การทดสอบประสิทธิภาพของโมเดลป่าสุ่ม กับ dataset อื่นๆ

การทดสอบประสิทธิภาพของโมเดลป่าสุ่ม model กับ dataset จาก Harvard Dataverse (M, 2021)

```
The RF model performance with other datasets (Harvard dataset) for predicting stroke risk.
RandomForestClassifier(criterion='entropy', n_estimators=200) Model
              precision    recall  f1-score   support

           0       0.62      0.70      0.66     10139
           1       0.66      0.58      0.61     10139

    accuracy                           0.64     20278
   macro avg       0.64      0.64      0.64     20278
weighted avg       0.64      0.64      0.64     20278

[[7099 3040]
 [4308 5831]]
--------------------------------
```

ภาพประกอบ 60 Classification report แสดงประสิทธิภาพของโมเดลป่าสุ่ม ที่ทดสอบกับ Harvard dataset

จากภาพประกอบ 60 แสดงผลการทดสอบประสิทธิภาพของโมเดลป่าสุ่ม กับ dataset จากแหล่งอื่นๆ (multi-center validation) โดยในการทดสอบทดสอบประสิทธิภาพของโมเดลป่าสุ่ม ครั้งนี้เราใช้ dataset จาก Harvard dataset พบว่าโมเดลป่าสุ่ม ให้ค่าประสิทธิภาพของโมเดลดังนี้ ให้ค่า accuracy โดยรวมของโมเดลเท่ากับ 0.64 ให้ค่า macro avg f1-score เท่ากับ 0.64 และค่า weighted avg f1-score เท่ากับ 0.64 และนอกจากนี้ยังแสดงประสิทธิภาพของโมเดลป่าสุ่ม โดยการแสดงผลของค่าประสิทธิภาพค่าอื่นๆของการทำนายดังนี้ ทำนายคลาส 0 (No stroke) ได้ค่า recall 0.70 ค่า precision 0.62 และค่า f1-score 0.66 ทำนายคลาส 1(Stroke) ได้ค่า recall 0.58 ค่า precision 0.66 และค่า f1-score 0.61



ภาพประกอบ 61 Confusion matrix แสดงค่าประสิทธิภาพของโมเดลป่าสุ่ม ที่ได้จาก Harvard dataverse

จากภาพประกอบ 61 Confusion matrix แสดงค่าประสิทธิภาพของโมเดล RF ที่ได้จาก Harvard โดยมีค่าของ

TP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 5831

TN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) พบว่าตรงกับสิ่งที่เกิดขึ้นจริงคือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) โดยมีค่าเท่ากับ 7099

FP = จากการทำนายว่ามีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) แต่พบว่าจริงๆแล้วนั้น คือไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองเลย (no stroke) โดยมีค่าเท่ากับ 3040

FN = จากการทำนายว่าไม่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (no stroke) แต่พบว่าจริงๆแล้วนั้นมีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (stroke) โดยมีค่าเท่ากับ 4308

บทที่5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยนี้เป็นการศึกษาเพื่อหาโมเดลในการทำนายโอกาสการเกิดโรคหลอดเลือดสมองซึ่งใช้ ข้อมูลเกี่ยวกับผู้ที่เป็นผู้ที่มีปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมองและผู้ที่เป็นโรคหลอดเลือดสมองจากการ Kaggle dataset โดยใช้เทคนิคการเรียนรู้ของเครื่องในการการวิเคราะห์ข้อมูลและในการสร้างโมเดลการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง โดยผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลองแต่ละเทคนิค เพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้ 1. สรุปผลการวิจัย 2. อภิปรายผลการวิจัย 3. ข้อเสนอแนะ

สรุปผลการวิจัย

จากการทดลองครั้งนี้เราเลือกใช้โมเดลป่าสุ่ม LR, และ SVM เพื่อสร้างโมเดลและหาโมเดลที่เหมาะสมที่สุดในการทำนายโอกาสการเกิดโรคหลอดเลือดสมองโดยจากการทดลองเราได้วัดผลและเปรียบเทียบประสิทธิภาพของแต่ละโมเดลโดยใช้ confusion matrix วัดค่า accuracy, sensitivity, specificity, f1-score และได้แสดงผลตามตาราง 11

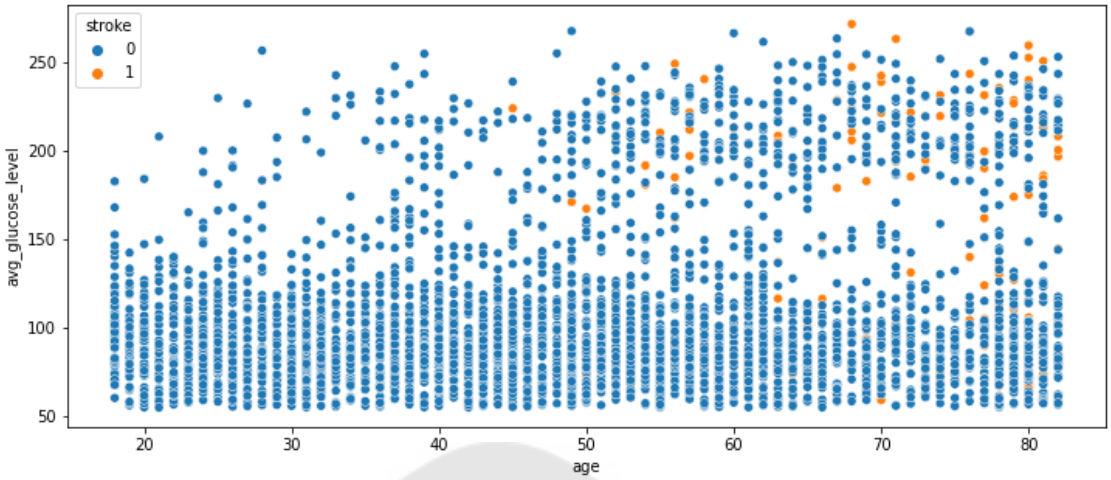
ตาราง 11 แสดงการสรุปเปรียบเทียบค่าประสิทธิภาพของแต่ละโมเดลที่ได้จาก balanced dataset และ หลังจากการทำ Hyperparameters Tuning

Model performance after tuning with GridSearchCV (CV: number of cross-validation = 10)			
Models	precision	recall	f1-score
RandomForest (Accuracy = 0.94 ,Specticity = 0.92)			
0	0.95	0.92	0.94
1	0.92	0.95	0.94
Logistic regression (Accuracy = 0.79 ,Specticity = 0.77)			
0	0.81	0.77	0.79
1	0.78	0.81	0.80
Support Vector Machine (Accuracy = 0.83 ,Specticity = 0.78)			
0	0.86	0.80	0.83
1	0.84	0.83	0.83

จากตาราง 11 เราพบว่าการทดลองครั้งนี้โมเดลป่าสุ่ม เป็นโมเดลที่มีประสิทธิภาพที่ดีที่สุดด้วยค่าความแม่นยำเท่ากับ 0.94 ค่าความเที่ยงตรงเท่ากับ 0.93 ค่าความไวเท่ากับ 0.95 ความจำเพาะเท่ากับ 0.93 ค่า f1-score เท่ากับ 0.94 และค่าพื้นที่ใต้กราฟเท่ากับ 0.94 และ Feature importance ของโมเดลป่าสุ่ม โดยเรียงจากค่ามากไปค่าน้อยมี age, avg_glucose_level, bmi, smoking_status, work_type, gender, Residence_type, ever_married and hypertension โดยมีค่า importance value เท่ากับ 0.381, 0.204, 0.154, 0.083, 0.057, 0.035, 0.033, 0.023, 0.018, 0.014 ตามลำดับ

อภิปรายผล

จากการทดลองครั้งนี้ โมเดลป่าสุ่ม เป็นโมเดลที่มีประสิทธิภาพที่ดีที่สุดด้วย ค่าความแม่นยำเท่ากับ 0.94 ค่าความเที่ยงตรงเท่ากับ 0.93 ค่าความไวเท่ากับ 0.95 ความจำเพาะเท่ากับ 0.93 ค่า f1-score เท่ากับ 0.94 และมีค่าพื้นที่ใต้กราฟ AUC ที่บอกถึงประสิทธิภาพของโมเดล RF ที่สามารถแยกกลุ่มคนที่มีความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองและกลุ่มคนที่ไม่มีความเสี่ยงต่อการต่อการเกิดโรคหลอดเลือดสมองออกจากกัน โดยมีจุด cut-off (จุดที่มีค่า sensitivity และ specificity สูงที่สุด) ของโมเดลเท่ากับ 0.94 ซึ่งถือว่าเป็นจุดที่โมเดลป่าสุ่ม ให้ประสิทธิภาพที่ดีเยี่ยม (excellent performance) และสามอันดับสูงสุดของความสำคัญของฟีเจอร์ของโมเดลป่าสุ่ม ที่มีลำดับตามความสำคัญจากมากไปน้อยคือตัวแปร อายุ มีค่า 0.383 ค่าเฉลี่ยของระดับน้ำตาลในเลือด มีค่า 0.203 และ ค่าดัชนีมวลกาย มีค่าเท่ากับ 0.153 ตามลำดับ



ภาพประกอบ 62 แสดงความสัมพันธ์ของความสำคัญของฟีเจอร์ของโมเดลป่าสุ่ม ของตัวแปร อายุ และค่าเฉลี่ยของระดับน้ำตาลในเลือด กับโอกาสการเกิดโรคหลอดเลือดสมอง

จากภาพประกอบ 62 แสดง scattering plot ของความสัมพันธ์ของความสำคัญของฟีเจอร์ของโมเดลป่าสุ่ม ของตัวแปร อายุ และค่าเฉลี่ยของระดับน้ำตาลในเลือดกับโอกาสความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง ซึ่งเราพบว่าเมื่อผู้ที่มีอายุมากกว่า 55 ปีและหรือมีค่าเฉลี่ยของระดับน้ำตาลในเลือดมากกว่า 100 มิลลิกรัม/เดซิลิตร มีความเสี่ยงหรือมีโอกาสต่อการเกิดโรคหลอดเลือดสมองมากขึ้น

การศึกษานี้เราพบว่าปัจจัยเสี่ยงที่ทำให้มีความเสี่ยงหรือมีโอกาสต่อการเกิดโรคหลอดเลือดสมองมีหลายปัจจัยเสี่ยงและเมื่อเราวิเคราะห์ปัจจัยเสี่ยงแต่ละประเภทกับโอกาสต่อการเกิดโรคหลอดเลือดสมอง เราพบว่ามีข้อมูลที่น่าสนใจดังต่อไปนี้

- อายุ เราพบว่าเมื่ออายุมากขึ้นจะทำให้มีโอกาสต่อการเกิดโรคหลอดเลือดสมองเพิ่มมากขึ้น

- เพศ ผู้ชายมีโอกาสเกิดโรคหลอดเลือดสมองมากกว่าผู้หญิงประมาณ 0.74 %

- สถานะภาพการแต่งงาน พบว่าคนที่แต่งงานแล้วมีโอกาสเป็นโรคหลอดเลือดสมองสูงกว่าคนที่ยังไม่ได้แต่งงาน 3.28%

- ผู้ที่เป็นโรคความดันโลหิตสูงพบว่ามีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าผู้ที่ไม่เป็นโรคความดันโลหิตสูง 9.24%

- ผู้ที่เป็นโรคหัวใจพบว่ามีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าผู้ที่ไม่เป็นโรคหัวใจ 12.14%

- ผู้ชายมีโอกาสเกิดโรคหลอดเลือดสมองมากกว่าผู้หญิงประมาณ 0.74 %

- คนที่เคยสูบบุหรี่มาก่อนมีโอกาสที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่ไม่เคยสูบบุหรี่ 2.1%

- ผู้ที่อาศัยในเมืองมีโอกาสที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่อาศัยอยู่ในเขตชนบท 0.31%

- ผู้ที่ทำธุรกิจส่วนตัวมีโอกาสที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่ทำงานกับบริษัทเอกชน 2.21%

- ผู้ที่ทำธุรกิจส่วนตัวมีโอกาสที่จะเป็นโรคหลอดเลือดสมองมากกว่าคนที่ทำงานรับราชการ 1.96%

โรคหลอดเลือดสมองเป็นโรคที่ในทางการแพทย์ถือว่าเป็นโรคที่มีผลกระทบที่คุกคามต่อชีวิตของมนุษย์ ซึ่งหากพบว่ามีอาการของโรคหลอดเลือดสมองเกิดขึ้นผู้ป่วยจะต้องได้รับการรักษาโดยเร็วที่สุด เพื่อหลีกเลี่ยงโอกาสการเสียชีวิตและช่วยลดภาวะแทรกซ้อนอื่น ๆ ที่จะเกิดขึ้นตามมา ถึงอย่างไรก็ตามในทางการแพทย์การป้องกันการเกิดโรคหลอดเลือดสมองนั้นเราถือเป็นสิ่งที่สำคัญที่สุด ดังนั้นการพัฒนาโมเดลการเรียนรู้ของเครื่องที่จะสามารถช่วยในการตรวจหาโอกาสการเกิดโรคหลอดเลือดสมองตั้งแต่เนิ่น ๆ จะช่วยให้แพทย์สามารถใช้เป็นเครื่องมือที่ใช้ในการสนับสนุนการตัดสินใจของแพทย์ในการวางแผนการป้องกันหรือรักษาโรคหลอดเลือดสมองให้มีประสิทธิภาพมากขึ้น ซึ่งจะช่วยลดโอกาสการเกิดโรคหลอดเลือดสมองหรือช่วยบรรเทาผลกระทบและความรุนแรงของโรคที่จะตามมาได้

การศึกษานี้มีการศึกษาประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องหลายอัลกอริทึมในการนำมาใช้สร้างโมเดลในการทำนายโอกาสการเกิดโรคหลอดเลือดสมองตามขั้นตอนของ data science โดยพิจารณาจากตัวแปรที่เป็นปัจจัยเสี่ยงต่อการเกิดโรคหลอดเลือดสมองและเราพบว่า การจำแนกประเภทของโมเดล มีประสิทธิภาพเหนือกว่ากระบวนการอื่นๆ เมื่อทดสอบประสิทธิภาพของโมเดล ด้วยความแม่นยำในการจำแนกประเภทที่ *94* เปอร์เซ็นต์ สำหรับขอบเขตในการศึกษาในอนาคต เราสามารถที่จะพัฒนาปรับปรุงให้โมเดลมีประสิทธิภาพที่ดีมากขึ้นได้ โดยการใช้ชุดข้อมูลที่มีขนาดใหญ่และทดลองใช้อัลกอริทึมการเรียนรู้ของเครื่องประเภทอื่นๆ เช่น AdaBoost, Bagging, Deep learning เป็นต้น

สถาปัตยกรรมการเรียนรู้ด้วยเครื่องยังสามารถช่วยประชาชนทั่วไปในการตรวจจับความน่าจะเป็นของโอกาสที่จะเป็นโรคหลอดเลือดสมองที่เกิดขึ้นในผู้ป่วยที่เป็นผู้ใหญ่ได้ ซึ่งจะทำให้

ผู้ป่วยได้รับการรักษาโรคหลอดเลือดสมองได้ตั้งแต่เนิ่น ๆ โดยการวางแผนการดูแลควบคุมปัจจัยเสี่ยงต่าง ๆ ที่ทำให้มีโอกาสที่จะเป็นโรคหลอดเลือดสมองได้อย่างมีประสิทธิภาพ และเมื่อเราทราบโอกาสของความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองแล้วเราสามารถวางแผนในการจัดการความเสี่ยงได้โดยการควบคุมปัจจัยเสี่ยงที่สามารถเปลี่ยนแปลงแก้ไขได้ ได้แก่ การควบคุมความดันโลหิตสูง ควบคุมระดับน้ำตาลในเลือดในเลือด ลดเลิกการสูบบุหรี่ ควบคุมภาวะไขมันในเลือดสูง ดูแลรักษาโรคหัวใจอย่างสม่ำเสมอ การออกกำลังกายอย่างสม่ำเสมอ ควบคุมน้ำหนัก งดเว้นการดื่มแอลกอฮอล์ สำหรับผู้ที่มีปัจจัยเสี่ยงดังกล่าวหลายตัวก็จะมีผลทำให้ความเสี่ยงสูงขึ้นด้วยแบบทวีคูณ ดังนั้น การลดปัจจัยเสี่ยงเหล่านี้โดยเร็วและสม่ำเสมอ จะเพิ่มโอกาสในการป้องกันการเกิดโรคหลอดเลือดสมองได้มากขึ้น ปัจจัยเสี่ยงที่กล่าวมานี้ส่วนใหญ่สามารถแก้ไขได้ด้วยการปรับเปลี่ยนพฤติกรรมและการใช้ยา

ข้อเสนอแนะ

การศึกษานี้เป็นการศึกษาที่ใช้ข้อมูลจากแหล่งเดียว (single-center study) ซึ่งหากมีการตรวจสอบ validation ด้วยข้อมูลจากแหล่งอื่นๆ (multi-center validation) ก็จะช่วยในการตรวจสอบประสิทธิภาพของโมเดลในการปรับตัวเข้ากับข้อมูลชุดอื่นๆได้ดีแค่ไหน
ในอนาคตเรายังสามารถนำแนวคิดและวิธีการการเรียนรู้ของเครื่องจากการศึกษานี้สามารถนำไปปรับใช้เพื่อสร้างแบบจำลองสำหรับการทำนายความเสี่ยงโรคหลอดเลือดสมองให้มีระดับความเสี่ยงที่มีระดับความเสี่ยงในหลายๆระดับมากขึ้น และเรายังสามารถนำขบวนการของการเรียนรู้ของเครื่องไปช่วยในการพัฒนาปรับปรุงประสิทธิภาพของเครื่องมือการทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมอง (CVD Risk) ที่มีใช้อยู่ในปัจจุบัน ซึ่งเป็นเครื่องมือที่สร้างมาจากพื้นฐานของหลักสถิติ (traditional statistics)

ข้อควรพิจารณาของระบบปัญญาประดิษฐ์กับการแพทย์

ระบบปัญญาประดิษฐ์ สามารถช่วยในการทำนายโรคและการตัดสินใจในขบวนการรักษา โดยระบบปัญญาประดิษฐ์ให้ประสิทธิภาพและความแม่นยำที่เทียบเท่ากับผู้เชี่ยวชาญ เราคิดว่าระบบปัญญาประดิษฐ์จะช่วยได้ แต่ยังไม่ถึงกับแทนที่แพทย์ เพราะเรายังไม่ทราบว่าระบบปัญญาประดิษฐ์สามารถทำงานได้ดีเพียงใดในสถานพยาบาล การใช้ระบบปัญญาประดิษฐ์ ในโรงพยาบาลอาจก่อให้เกิดคำถามด้านจริยธรรมและกฎหมายใหม่ ๆ แพทย์อาจต้องเผชิญกับปัญหาความรับผิดเมื่อมีการปฏิบัติตามคำแนะนำของระบบปัญญาประดิษฐ์ โดยเฉพาะอย่างยิ่ง

เมื่อคำแนะนำเหล่านี้ไม่สอดคล้องกับการตัดสินใจของแพทย์หรือตามความรู้ทางคลินิกและสัญชาตญาณการตัดสินใจ นอกจากนี้ระบบปัญญาประดิษฐ์ยังไม่สามารถอธิบายบริบททางสังคมและวัฒนธรรมที่มีอิทธิพลต่อการดูแลผู้ป่วยได้ ด้วยเหตุนี้ อุปสรรคและหลุมพรางเหล่านี้จึงควรได้รับการพิจารณาอย่างรอบคอบเมื่อใช้เทคโนโลยี ระบบปัญญาประดิษฐ์กับระบบการรักษาทางคลินิก

บรรณานุกรม

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR),* 16, 321-357.

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N., Venugopal, V., . . . Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet,* 392.

Chiu, I. M., Zeng, W. H., Cheng, C. Y., Chen, S. H., & Lin, C. R. (2021). Using a Multiclass Machine Learning Model to Predict the Outcome of Acute Ischemic Stroke Requiring Reperfusion Therapy. *Diagnostics (Basel),* 11(1).

Chun, M., Clarke, R., Cairns, B. J., Clifton, D., Bennett, D., Chen, Y., . . . China Kadoorie Biobank Collaborative, G. (2021). Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *J Am Med Inform Assoc,* 28(8), 1719-1727.

Fedesoriano. (2021). StrokePredictionDataset. version1. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata

Hfocus. (Thursday, 29 October 2020). Know quickly. *survive! Cerebrovascular disease is the leading cause of the elderly. handicapped-death.* https://www.hfocus.org/content/2020/10/20381

Hospital, F. o. M. S. How is cerebrovascular disease treated? https://www.si.mahidol.ac.th/center/sirirajstrokecenter/TH/StrokeContent/content/people/cure-stroke.aspx

Kogan, E., Twyman, K., Heap, J., Milentijevic, D., Lin, J. H., & Alberts, M. (2020). Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak,* 20(1), 8.

Lee, H., Lee, E. J., Ham, S., Lee, H. B., Lee, J. S., Kwon, S. U., . . . Kang, D. W. (2020). Machine Learning Approach to Identify Stroke Within 4.5 Hours. *Stroke,* 51(3), 860-866.

Liu, Y., Yin, B., & Cong, Y. (2020). The Probability of Ischaemic Stroke Prediction with a

Multi-Neural-Network Model. *Sensors (Basel),* 20(17).

M, M. (2021). *Replication Data for: Prediction of Cerebral Stroke*. Retrieved from: https://doi.org/10.7910/DVN/44RCPZ

Peeradon Samasiri, P., Scientist, D., & (GBDi), G. B. D. I. (2021). Logistic Regression By Microsoft Excel. https://bigdata.go.th/big-data-101/lr-excel/

Rabiablok, V., Manager, S. D. S. a. P., Government Big Data Institute (GBDi), Mayurasakhon, N., Scientist, D., & (GBDi), G. B. D. I. (2021, July 8, 2021). การวิเคราะห์ความสัมพันธ์กับข้อมูลขนาดใหญ่. https://bigdata.go.th/big-data-101/data-science/correlation-analysis-in-big-data/

Scrutinio, D., Ricciardi, C., Donisi, L., Losavio, E., Battista, P., Guida, P., . . . D'Addio, G. (2020). Machine learning to predict mortality after rehabilitation among patients with severe stroke. *Sci Rep,* 10(1), 20127.

Sharma, C., Sharma, S., Kumar, M., & Sodhi, A. (2022). Early Stroke Prediction Using Machine Learning.

Sirsat, M. S., Ferme, E., & Camara, J. (2020). Machine Learning for Brain Stroke: A Review. *J Stroke Cerebrovasc Dis,* 29(10), 105162.

Site, A., Nurmi, J., & Lohan, E. S. (2021). Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis. *IEEE Access,* 9, 112221-112235.

Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., . . . Oermann, E. K. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med,* 24(9), 1337-1341.

Tsevat, J., & Moriates, C. (2018). Value-Based Health Care Meets Cost-Effectiveness Analysis. *Ann Intern Med,* 169(5), 329-332.

ประวัติผู้เขียน

ชื่อ-สกุล	สากล พัชรปัญญวัฒน์
วัน เดือน ปี เกิด	19 กุมภาพันธ์ 2517
สถานที่เกิด	นครราชสีมา
วุฒิการศึกษา	พ.ศ.2540 วิทยาศาสตร์บัณฑิต (กายภาพบำบัด) จาก มหาวิทยาลัยศรีนครินทรวิโรฒ
ที่อยู่ปัจจุบัน	44/137 ถนน กาญจนาภิเษก 5/7 ท่าแร้ง บางเขน กรุงเทพ 10220