Stroke Prediction

6304062630016 กฤษญา โมรา[1], 6304062630091 ณัชชา วิเศษสุทธิ์[2] , 6304062630229 ปวีณอร สำลี[3]

บทคัดย่อ

ในปัจจุบันโรคหลอดเลือดสมอง (Stroke) เป็นสาเหตุการเสียชีวิตอันดับ 2 ของโลก และยังมีแนวโน้มที่จะเพิ่มมากขึ้นทุกปี โรคหลอดเลือดสมองหรือที่เรียกกันว่า อัมพฤกษ์ อัมพาต หรือทางการแพทย์เรียกว่า STROKE เป็นโรคที่มีความรุนแรงสูงถึงขั้นเสียชีวิต และแม้ว่าจะไม่เสียชีวิตแต่ทำให้เกิดความพิการระยะยาว ต้องอาศัยความช่วยเหลือจากผู้อื่นตลอดชีวิต เกิดความสูญเสียทางเศรษฐกิจและสังคม

การสังเกตลักษณะของอาการและการวินิจฉัยโรคจึงสำคัญมาก และเนื่องด้วยในปัจจุบันการรักษาสามารถเข้าถึงได้ง่ายกว่าสมัยก่อน แต่วงการแพทย์ยังขาดแคลนบุคลากร ทำให้แพทย์บางท่านต้องทำงานอย่างหนักเพื่อรักษาชีวิตผู้คนคงดีกว่าหากมีตัวช่วยแบ่งเบาการทำงานนั้นให้เบาบางลง จึงเกิดอัลกอริทึมที่ใช้โมเดลในการช่วยทำนายว่าผู้ป่วยท่านใดมีความเสี่ยงที่จะเป็นโรคหลอดเลือดสมอง(Stroke) บ้าง ก็จะช่วยในการวินิจฉัยโรคในเบื้องต้นเพื่อคัดกรองผู้ป่วย

คำหลัก: Stroke , Hypertension , Heart Disease , Age , BMI

1. ที่มาและความสำคัญ

ตามสถิติที่ world health organization(WHO) ได้ระบุไว้ โรคหลอดเลือดสมอง (Stroke) เป็นสาเหตุของการเสียชีวิตอันดับ 2 ของโลก มากถึง 11% ของการเสียชีวิตทั้งหมด และอันดับ 3 ของความพิการ ประมาณ 2 ใน 3 ของผู้ป่วยโรคนี้เกิดขึ้นในประเทศที่น้อยพัฒนาหรือกำลังพัฒนา ซึ่งปัจจุบันถือเป็นสาเหตุการเสียชีวิตอันดับ 1 ในเพศหญิง เนื้อสมองของผู้ป่วยจะถูกทำลาย สูญเสียการทำหน้าที่จนเกิดอาการของอัมพฤกษ์ อัมพาต หรือร้ายแรงถึงขั้นเสียชีวิตได้ อาการสมองขาดเลือดจะเกิดแบบเฉียบพลัน มีอาการชาที่ใบหน้า ปากเบี้ยว พูดไม่ชัด แขนหรือขาอ่อนแรงข้างใดข้างหนึ่งหรือทั้งสองข้าง เคลื่อนไหวไม่ได้หรือเคลื่อนไหวลำบาก เดินเซ ปวดศีรษะมาก ตามัวมองเห็นไม่ชัด โดยอาการเกิดขึ้นอย่างทันทีทันใด นอกจากความพิการทางกายแล้ว ยังมีผลต่อความคิด การวางแผน ความจำ ทำให้เกิดความจำเสื่อมในระยะต่อมา ซึ่งมักถูกมองข้ามไปในผู้ป่วยส่วนใหญ่

ดังนั้นจะเห็นว่าโรคหลอดเลือดสมองเป็นปัญหาสำคัญของประเทศ จำเป็นต้องมีแนวทางการรักษาหรือการป้องกันไม่ให้เกิดโรคหลอดเลือดสมอง (Stroke) กับประชนชน โดยอาจปรับเปลี่ยนพฤติกรรมการบริโภคหรือออกกำลังกาย หมั่นดูแลสุขภาพ และเข้ารับการตรวจสุขภาพประจำปี ซึ่งหากมีโมเดลมาช่วยในการวินิจฉัยโรค ก็จะทำให้การวินิจฉัยนั้นรวดเร็วและมีประสิทธิภาพมากขึ้น สามารถเข้ารักษาได้ทันท่วงที และเป็นการช่วยแบ่งเบาการทำงานของบุคลากรทางการแพทย์เพื่อให้สามารถเตรียมการรักษาได้มีอย่างประสิทธิภาพมาก และช่วยลดแนวโน้มของผู้ป่วยโรคหลอดเลือดสมอง (Stroke) ลงมา

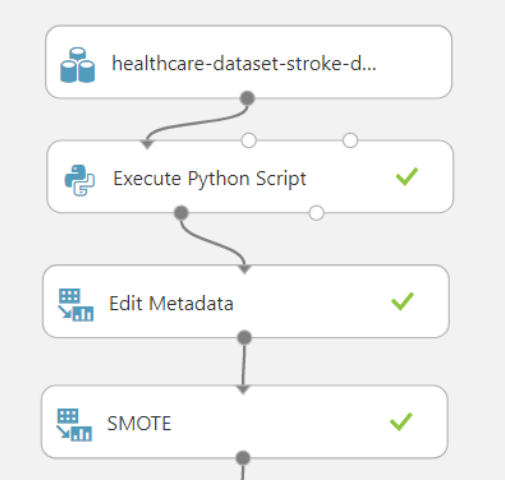
2. วิธีการ

2.1 โหลด Dataset มาจาก Kaggle แล้วนำเข้ามาใน Azure โดยมีชื่อไฟล์ว่า healthcare-dataset-stroke-data.csv

2.2 ทำการ Clean Data โดยใช้ Execute Python Script

2.3 แบ่งกลุ่มข้อมูลโดยใช้ Edit Metadata

2.3 ทำให้ข้อมูลสมดุลกัน โดยใช้ SMOTE



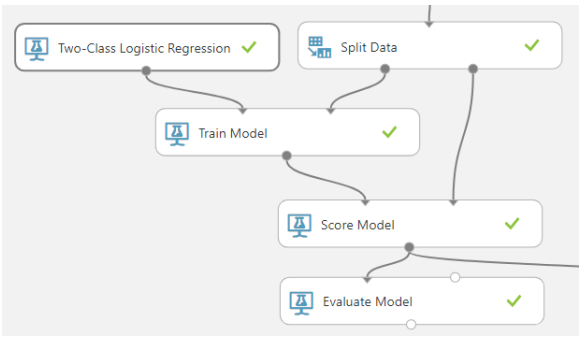
2.4 เลือกโมเดล Two-Class Logistic Regression เพื่อใช้ในการทำนายข้อมูล

2.5 Split Data แบ่งเป็น Train 80% และ Test 20%

2.6 Train Model

2.7 ดูว่าโมเดลทำนายเป็นอย่างไร โดยใช้ Score Model

2.8 Evaluate Model เพื่อวัดประสิทธิภาพของโมเดล



2.9 Convent to CSV เพื่อนำข้อมูลที่ผ่านการ Train แล้วไปใช้ประโยชน์ต่อไป

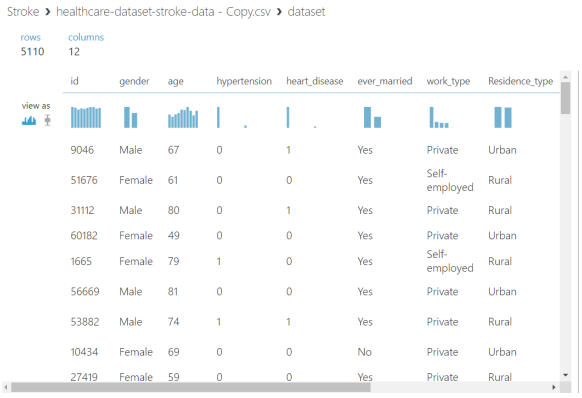
3. การออกแบบการทดลอง

3.1 ชุดข้อมูลที่ใช้

ใช้ Dataset จาก Kaggle โดยมีชื่อไฟล์ว่า healthcare-dataset-stroke-data.csv

โดยจะมีชื่อคอลัมน์ดังนี้

- id
- heart_disease
- ever_married
- stroke
- gender
- work_type
- age
- Residence_type
- hypertension
- avg_glucose_level
- bmi
- smoking_status



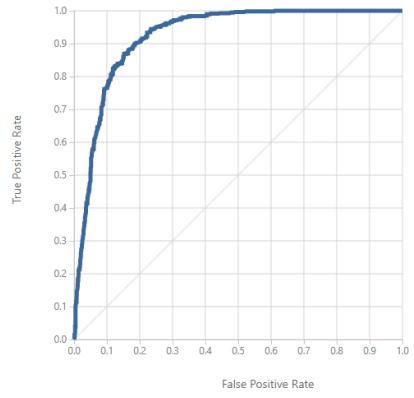
Stroke › healthcare-dataset-stroke-data - Copy.csv › dataset

rows 5110 columns 12

heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
1	Yes	Private	Rural	105.92	32.5	never smoked	1
0	Yes	Private	Urban	171.23	34.4	smokes	1
0	Yes	Self-employed	Rural	174.12	24	never smoked	1
0	Yes	Private	Urban	186.21	29	formerly smoked	1
1	Yes	Private	Rural	70.09	27.4	never smoked	1
0	No	Private	Urban	94.39	22.8	never smoked	1
0	Yes	Private	Rural	76.15	N/A	Unknown	1

3.2 วิธีการวัดความถูกต้อง

Stroke › Evaluate Model › Evaluation results



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
457	98	0.859	0.796	0.5	0.924

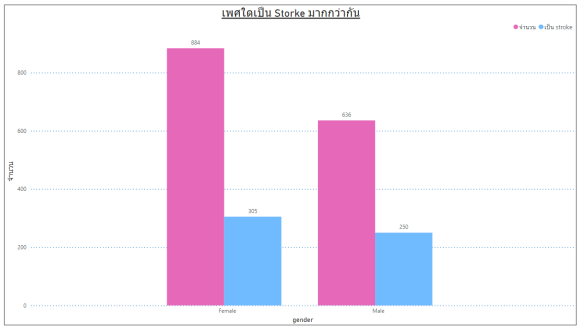
False Positive	True Negative	Recall	F1 Score
117	848	0.823	0.810

Positive Label	Negative Label
1	0

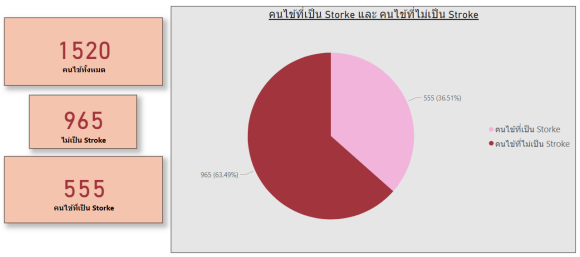
จะดูจาก ROC Curve ถ้ามีความโค้งที่เข้าใกล้แกน Y มากเท่าไหร่แสดงว่ามีความถูกต้องสูง ส่วน AUC , Accuracy , Precision Recall , F1 Score จะดูจากผลลัพธ์ที่ได้จากการคำนวณ

4. ผลการทดลอง

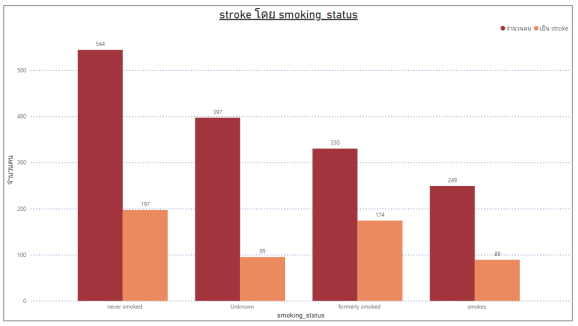
นำไฟล์ CSV ที่ได้จากการ Train โมเดลเพื่อนำไปวาดกราฟดูผลลัพธ์ที่ได้ เพื่อวิเคราะห์ความเกี่ยวข้อง , แนวโน้มและการเปรียบเทียบข้อมูล



เปรียบเทียบว่าเพศใดเป็นโรค Stroke มากกว่ากัน พบว่าโรค Stroke จะพบในผู้หญิงมากกว่าผู้ชาย



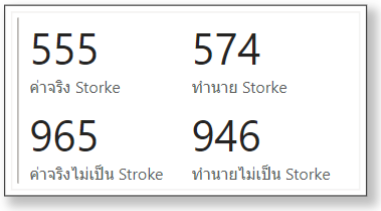
เปรียบเทียบว่าคนไข้ที่เป็น Stroke กับคนไข้ที่ไม่เป็น Stroke พบว่าคนไข้ที่ไม่เป็น Stroke มีจำนวนเยอะกว่า



เปรียบเทียบว่าคนไข้ที่เป็น Stroke มีสถานะการสูบบุหรี่แบบใด พบว่าคนส่วนใหญ่ที่เป็น Stroke มีสถานะสูบบุหรี่เป็น never smoked

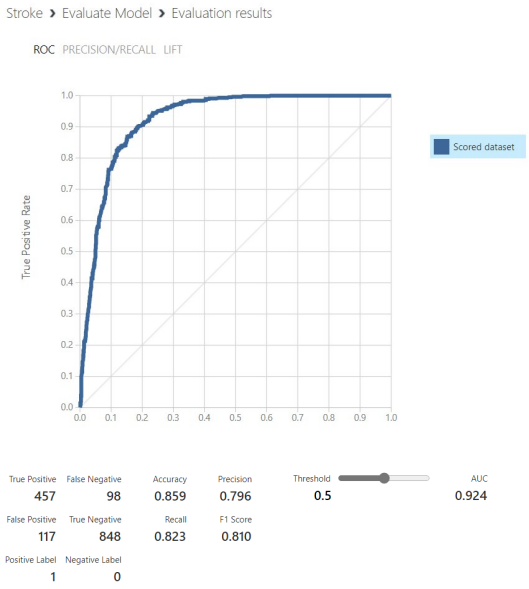
Stroke	คนไข้ทั้งหมด	hypertension	avg_bmi	heart_disease	avg_glucose_level
No Storke	965	88	28.00	38	104.35
Storke	555	63	25.15	11	129.94
ผลรวม	**1520**	**151**	**26.96**	**49**	**113.69**

เปรียบเทียบให้เห็นว่าผู้ป่วยที่เป็น Stroke เป็น Hypertension , Heart disease กี่คน และมีช่วงค่าเฉลี่ยของ BMI และ Glucose เท่าใด



เปรียบเทียบค่าจริงของ Stroke กับค่าที่โมเดลเราทำนายออกมา

5. สรุปผลการทดลอง



ROC Curve เป็นอีกตัวที่วัดประสิทธิภาพว่าโมเดลเราแยกกลุ่มของคนที่เป็น Stroke กับคนที่ไม่เป็น Stroke ได้ดีมากแค่ไหน ซึ่ง ROC Curve ที่ได้จากการ Train โมเดลมีพื้นที่ใต้กราฟที่เยอะและลักษณะการโค้งแบบนี้มันแปลว่าผลลัพธ์ที่ได้ค่อนข้างดี แสดงว่ามี True Positive Rate ยิ่งเข้าใกล้แกน Y แปลว่าโมเดลจำแนกได้ดี มีความแม่นยำสูง

AUC เป็นตัวบอกว่ามีพื้นที่ใต้กราฟเท่าไหร่ ซึ่งถ้ามีค่าเป็น 1 คือค่าในอุดมคติ ยิ่งมีค่าเข้าใกล้ 1 ก็แปลว่าโมเดลทำนายได้ถูกต้องเยอะมากเท่านั้น ซึ่งโมเดลนี้ทำ AUC ได้สูงถึง 0.924

Accuracy: คือค่าความแม่นยำรวมของทั้งโมเดล ซึ่งได้ 0.859

Precision: ค่าความแม่นยำ เกิดจากการนำค่า TP มาเทียบกับ FP ซึ่งได้ 0.796

Recall: ค่าความถูกต้อง เกิดจากการนำค่า TP มาเทียบกับ FN ซึ่งได้ 0.823

F1 Score: ค่าเฉลี่ยของ Precision และ Recall ซึ่งได้ 0.810

การที่ได้ผลลัพธ์เหล่านี้ในค่าที่สูงแปลว่าโมเดลมีความสามารถในการทำนายสูง อาจกล่าวได้ว่าเป็นโมเดลที่ดีและช่วยเพิ่มประสิทธิภาพให้กับการวินิจฉัยโรคได้ ลดการทำงานที่หนักเกินไปของบุคลากรทางการแพทย์และใช้โมเดลเข้ามามีบทบาทในการช่วยประเมินผู้ที่มีแนวโน้มการเป็นโรค Stroke ได้ในเบื้องต้น เพื่อการรักษาทีทันท่วงทีและลดแนวโน้มของผู้ป่วยที่เป็นโรค Stroke

6. เอกสารอ้างอิง

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

https://www.prd.go.th/th/content/category/detail/id/9/iid/130588