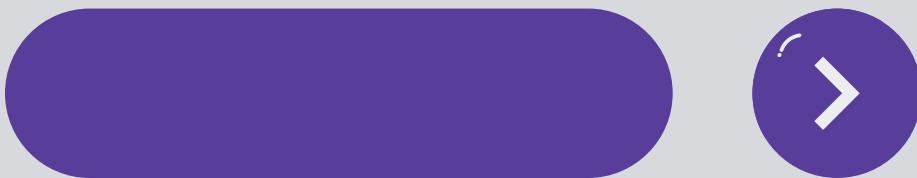


Stroke Prediction

การคำนายน้ำหนึ่งที่จะเป็นโรคหลอดเลือดในสมอง



สมาชิก

653380190-1 นายจักรกัตร วงศ์ศรีวรรณ Section 2

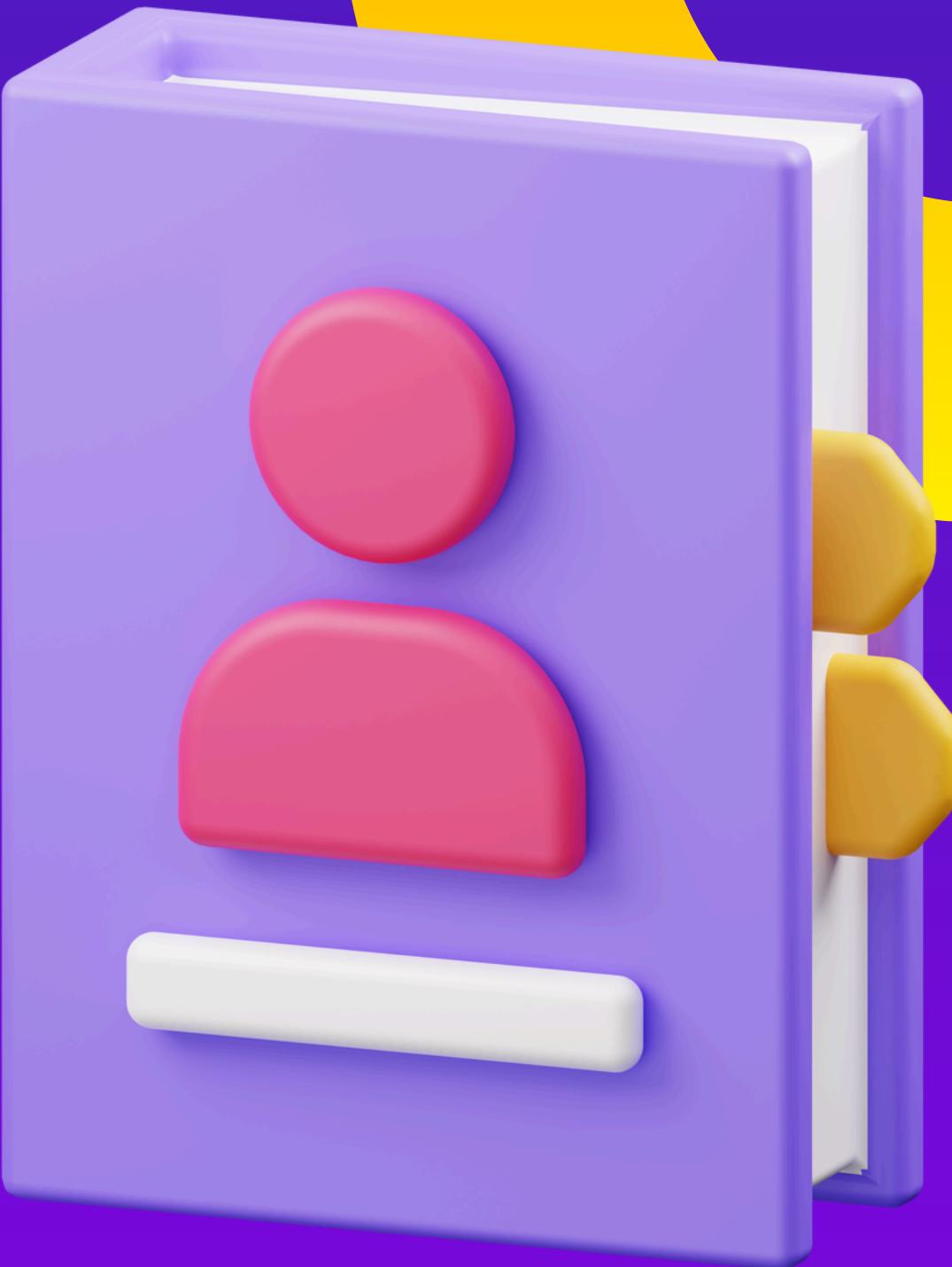
หน้าที่ เก็บข้อมูล วิเคราะห์ข้อมูล สร้างโมเดล

653380341-6 นางสาววารณี อุบลสุเรนทร์ Section 2

หน้าที่ วิเคราะห์ข้อมูล จัดทำรายงาน

653380348-2 นางสาวสิริยากร อาจยางคำ Section 2

หน้าที่ วิเคราะห์ข้อมูล จัดทำพรีเซนเทชั่น

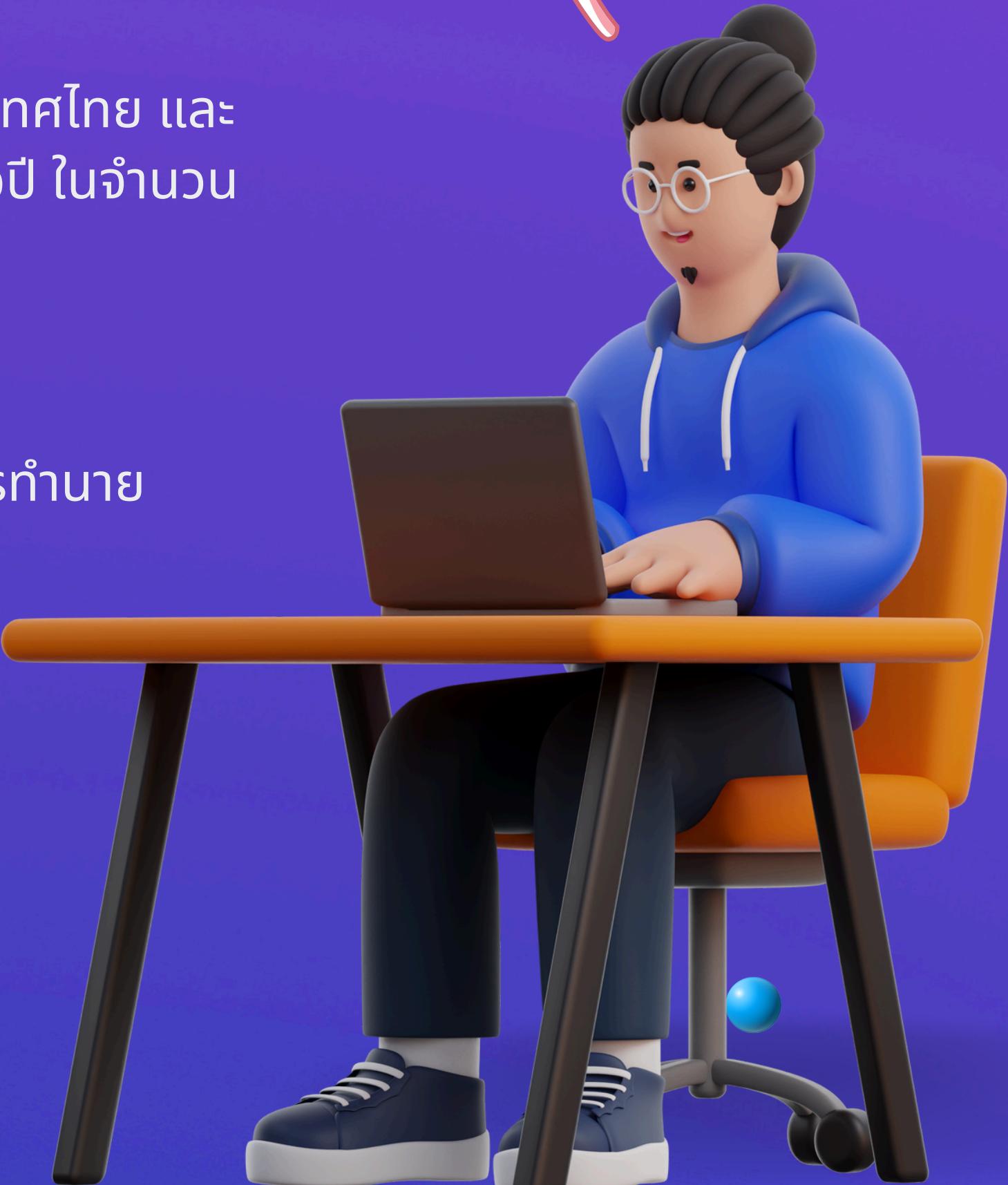


ความเป็นมาและความสำคัญของปัญหา



โรคหลอดเลือดสมอง เป็นโรคที่พบบ่อย เป็นปัญหาสาธารณสุขของประเทศไทย และ กั่งโลก ปัจจุบันมี ผู้ป่วยโรคหลอดเลือดสมองเกิดใหม่ราว 15 ล้านรายต่อปี ในจำนวนนี้เสียชีวิตประมาณ 5 ล้านราย ที่เหลือ พิการเป็นส่วนใหญ่

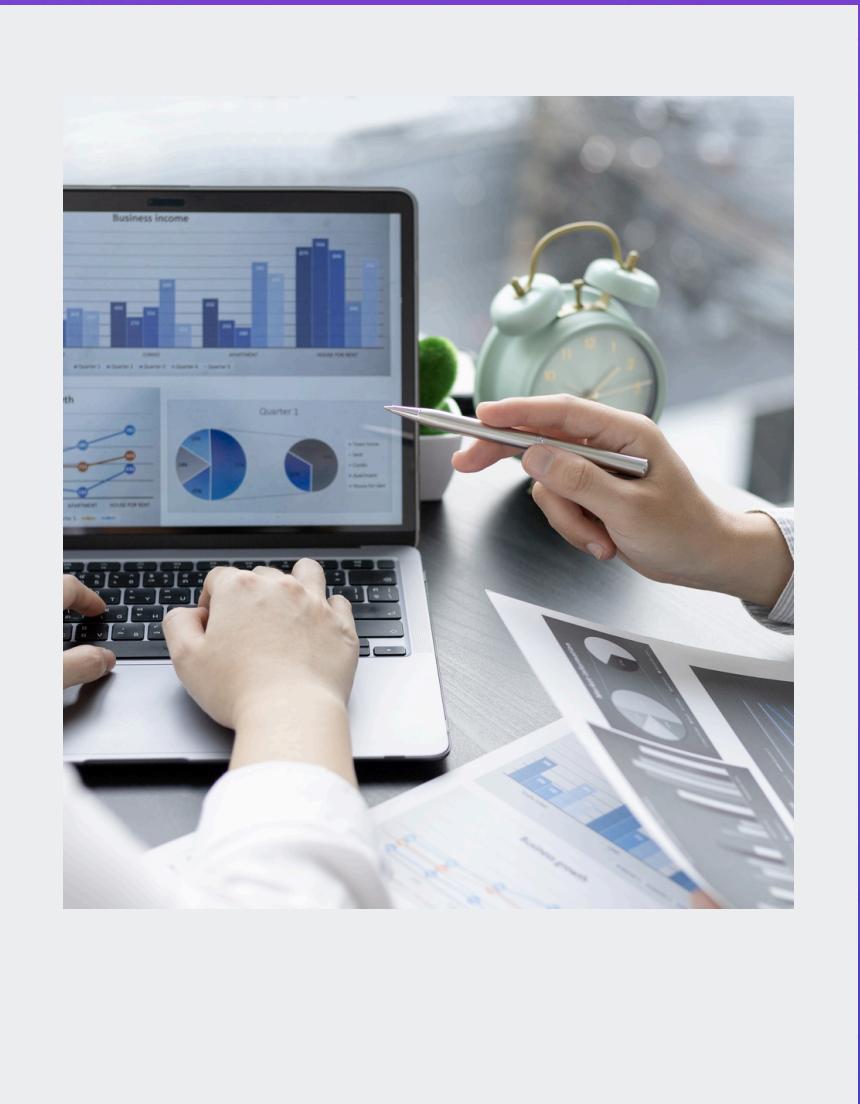
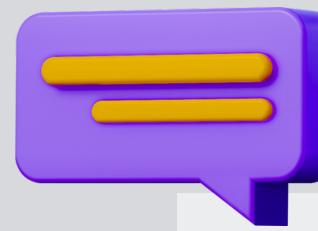
ด้วยสาเหตุดังกล่าวข้างต้นผู้จัดทำจึงให้ความสนใจในสร้างตัวแบบในการกำนยา แนวโน้มการเกิดโรคหลอดเลือดสมองของผู้ป่วย



วัตถุประสงค์

1) เพื่อสร้างตัวแบบหรือโมเดลที่ใช้สำหรับการคำนวณแนวโน้มที่จะเกิดโรคหลอดเลือดในสมอง (Stroke)

2) เพื่อศึกษาและเรียนรู้กระบวนการในการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูล



Questioning

ปัจจัยใดบ้างที่ทำให้เกิดโรคหลอดเลือดในสมอง ?



Get the data

FEDESORIANO · UPDATED 3 YEARS AGO

▲ 2758 New Notebook Download (69 kB)

Stroke Prediction Dataset

11 clinical features for predicting stroke events

Data Card Code (1030) Discussion (43)

About Dataset

Similar Datasets

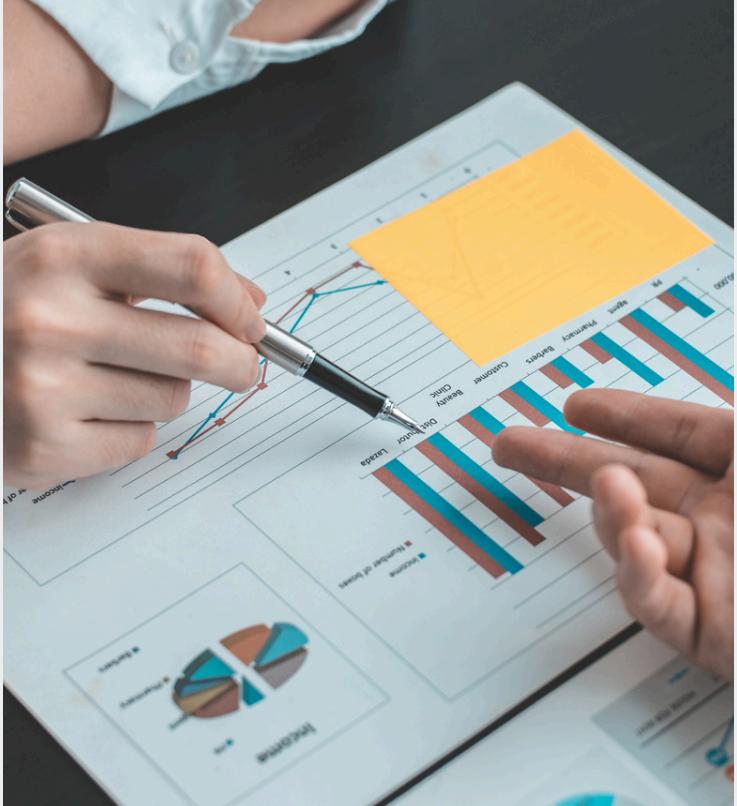
- [HIGHLIGHTED] CERN Electron Collision Data [LINK](#)
- Hepatitis C Dataset: [LINK](#)
- Body Fat Prediction Dataset: [LINK](#)
- Cirrhosis Prediction Dataset: [LINK](#)

Usability ⓘ
10.00

License
Data files © Original Authors

Expected update frequency
Never

Tags



เลือกจากปัจจัยที่จะส่งผลต่อการเกิดโรคมากที่สุด โดยมีดังนี้

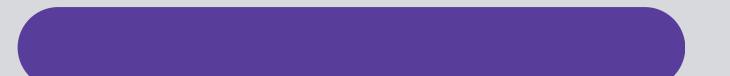
- 1) อายุ (Age)
- 2) เพศ (Gender)
- 3) ประวัติโรคความดัน (Hypertension)
- 4) ประวัติโรคหัวใจ (Heart disease)
- 5) ประวัติการแต่งงาน (Ever married)
- 6) อาชีพ (Work type)
- 7) ที่อยู่ โดยจำแนก ในเมืองและชนบท (Residence type)
- 8) ค่าเฉลี่ยกลูโคส (Avg glucose level)
- 9) ค่า BMI (BMI)
- 10) ประวัติการสูบบุหรี่ (Smoking status)

การเลือกกลุ่ม ประชากร

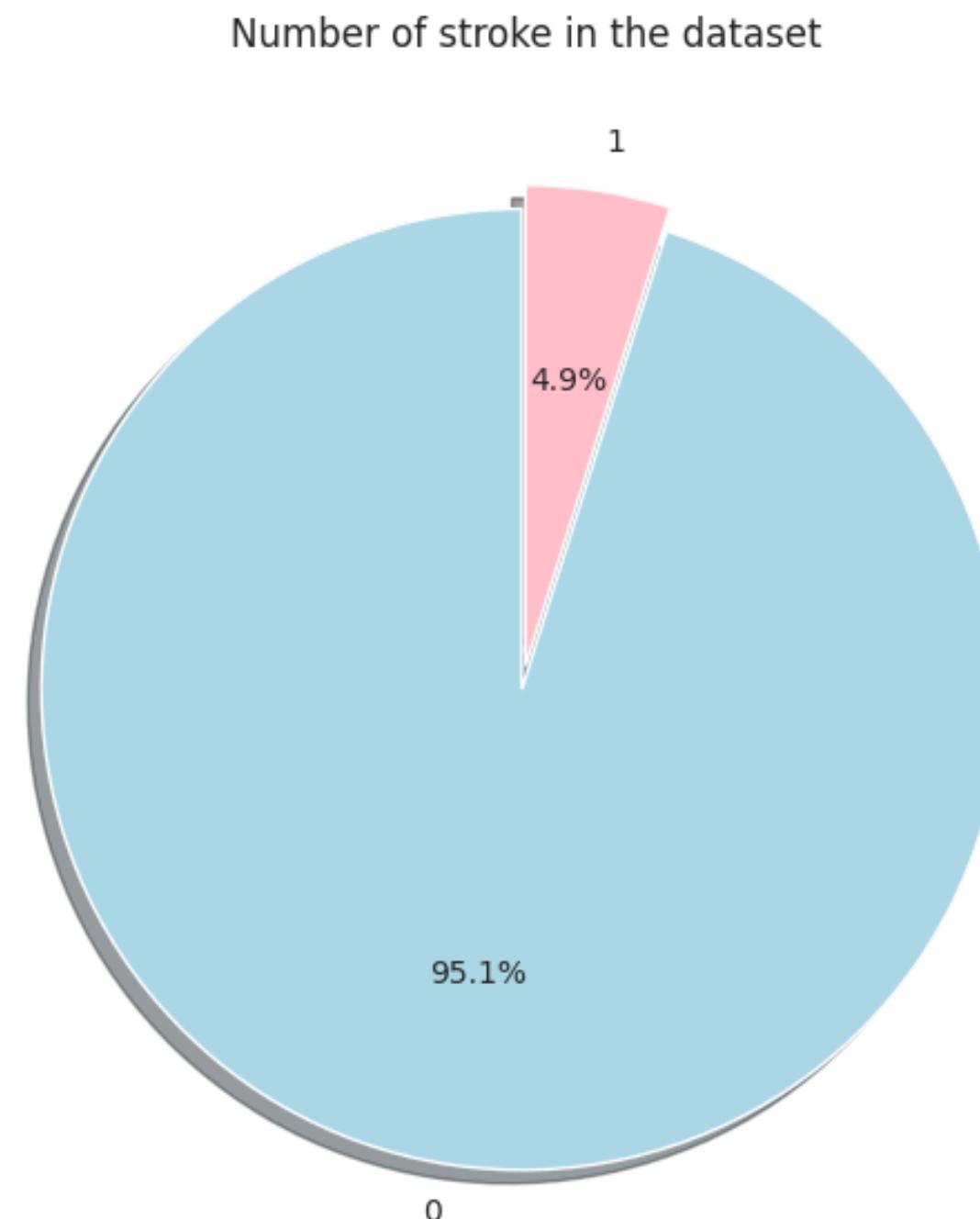


EDA

Exploratory Data
Analysis

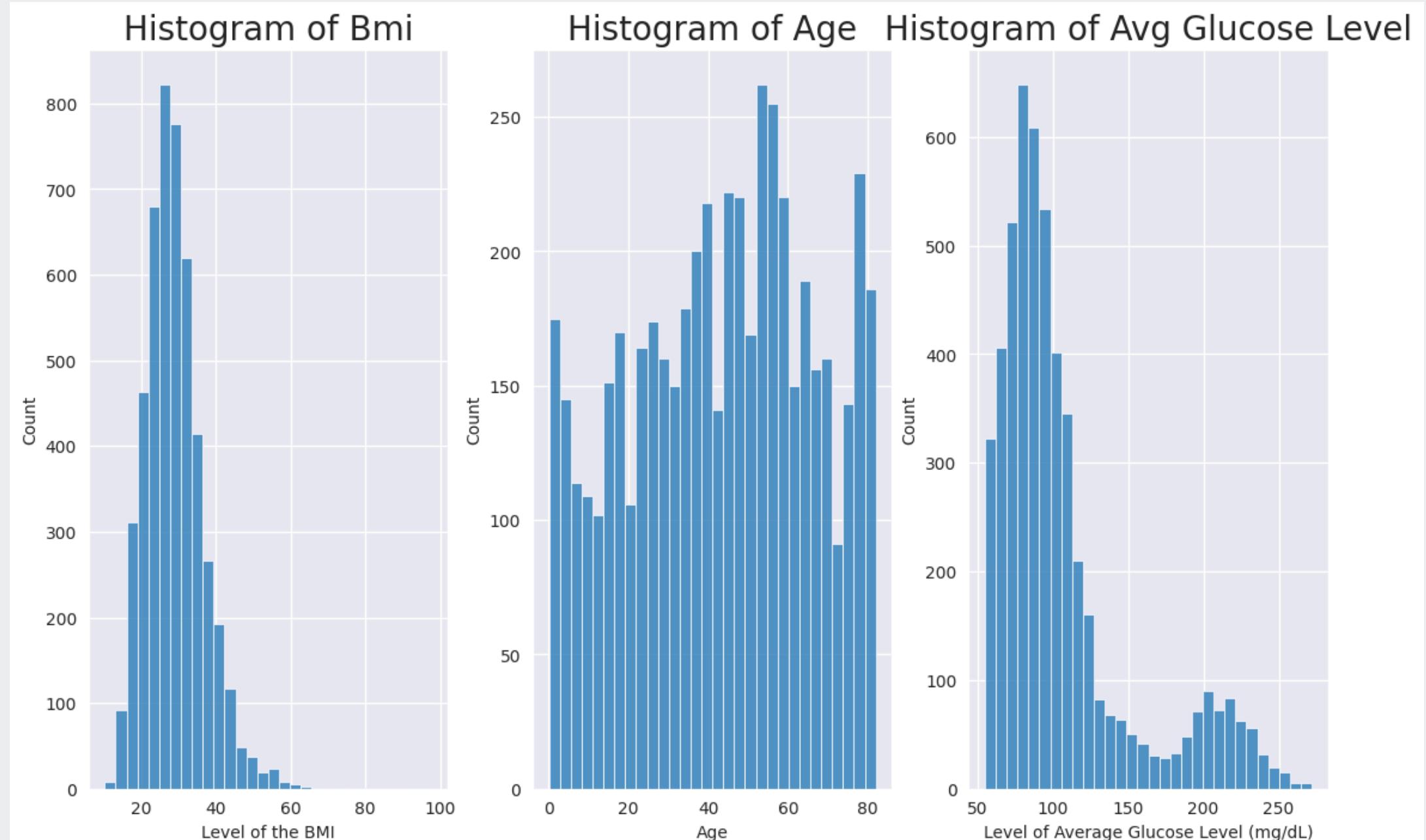


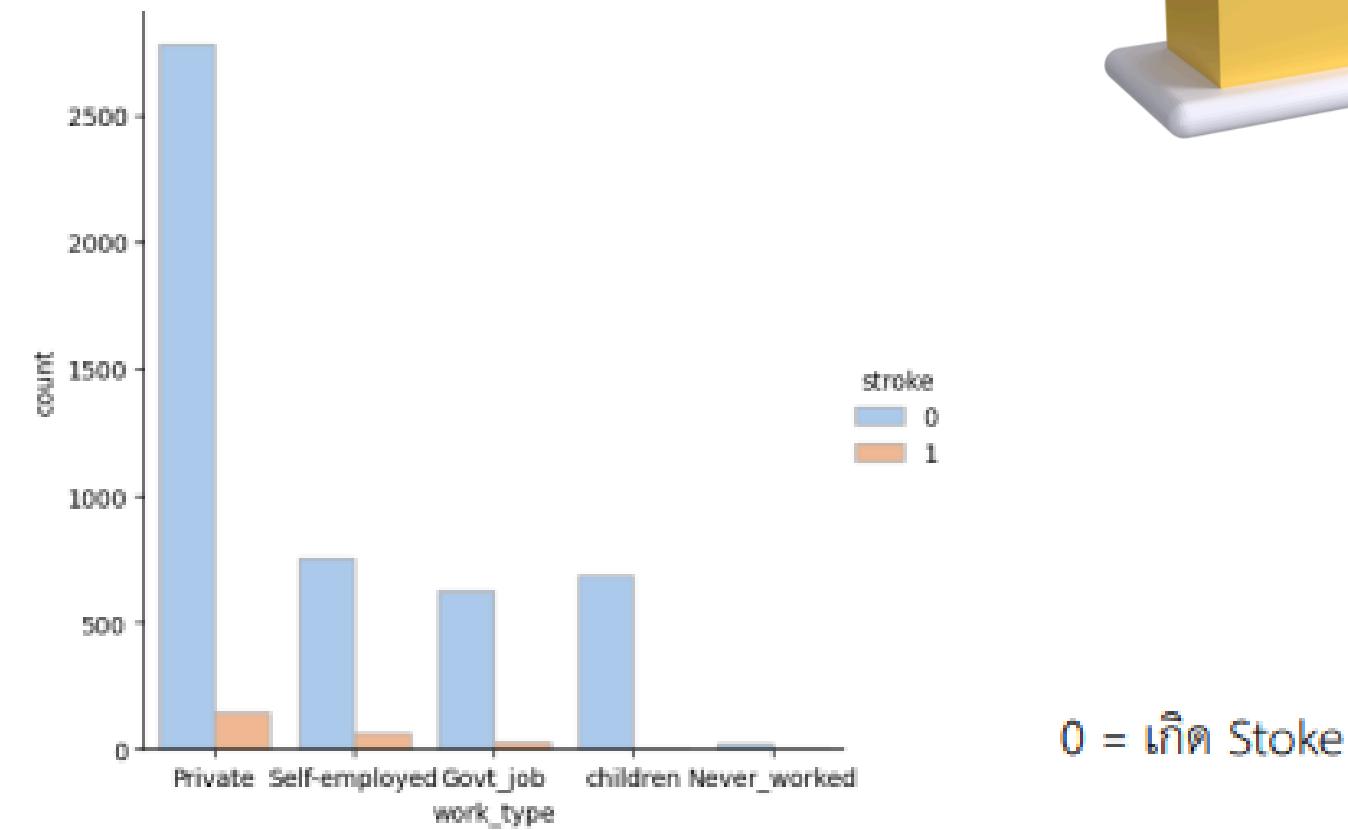
จำนวนคนเกิด Stroke



ทำการแสดงการกระจายค่า BMI Age และ Average Glucose Level จะได้ผลลัพธ์ดังนี้

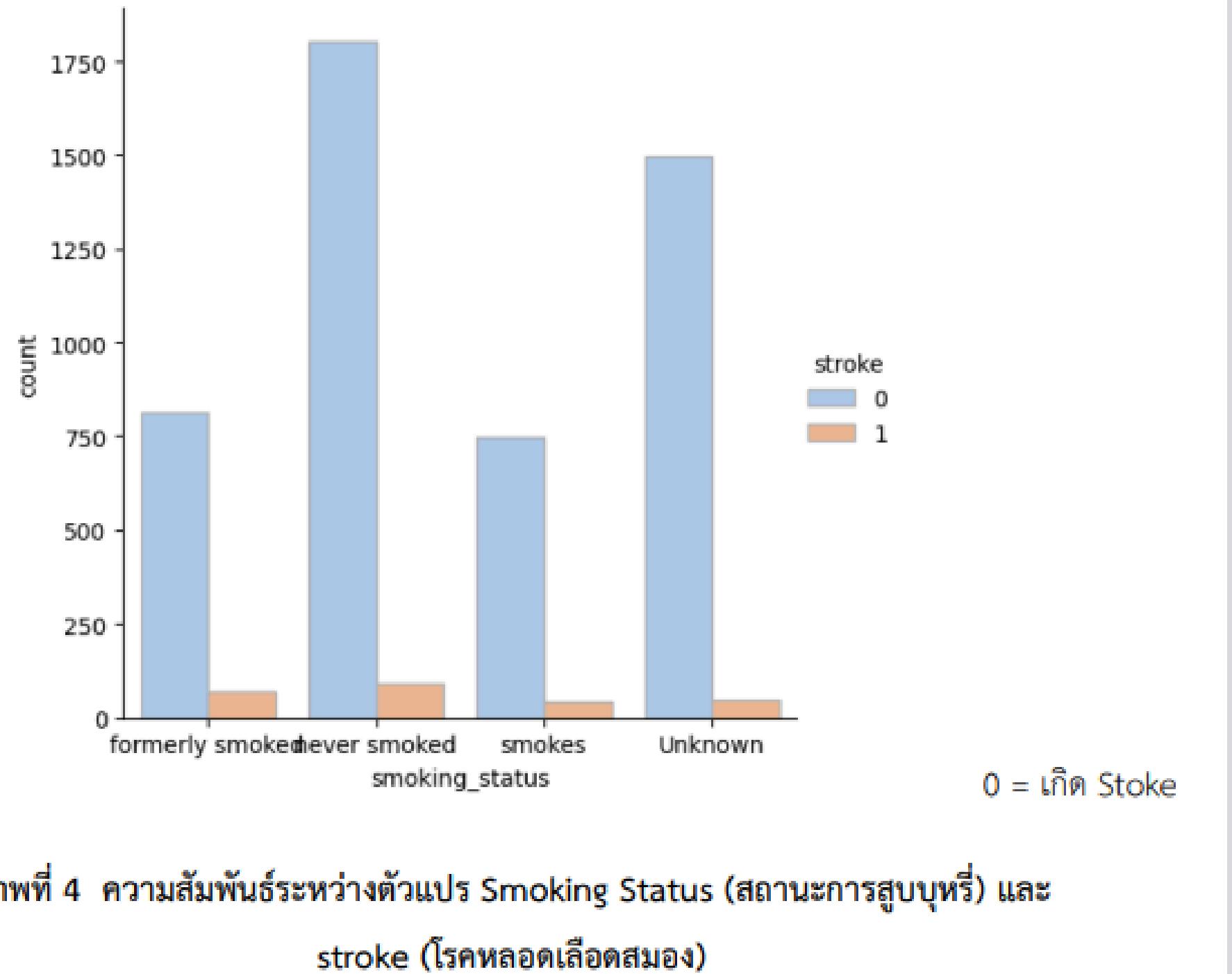
- กราฟ Histogram ของ BMI (คอลัมน์ค่าดัชนีมวลกาย BMI) จะแสดงการกระจายของค่า BMI จึงเห็นว่ากราฟนี้แสดงความถี่ของค่า BMI ที่ต่างกัน
- กราฟ Histogram ของ Age (คอลัมน์อายุ) จะแสดงการกระจายของอายุ
- กราฟ Histogram ของ Average Glucose Level (คอลัมน์ระดับน้ำตาลเลือดเฉลี่ย) จะแสดงการกระจายของระดับน้ำตาลเลือดเฉลี่ยในหน่วย mg/dL





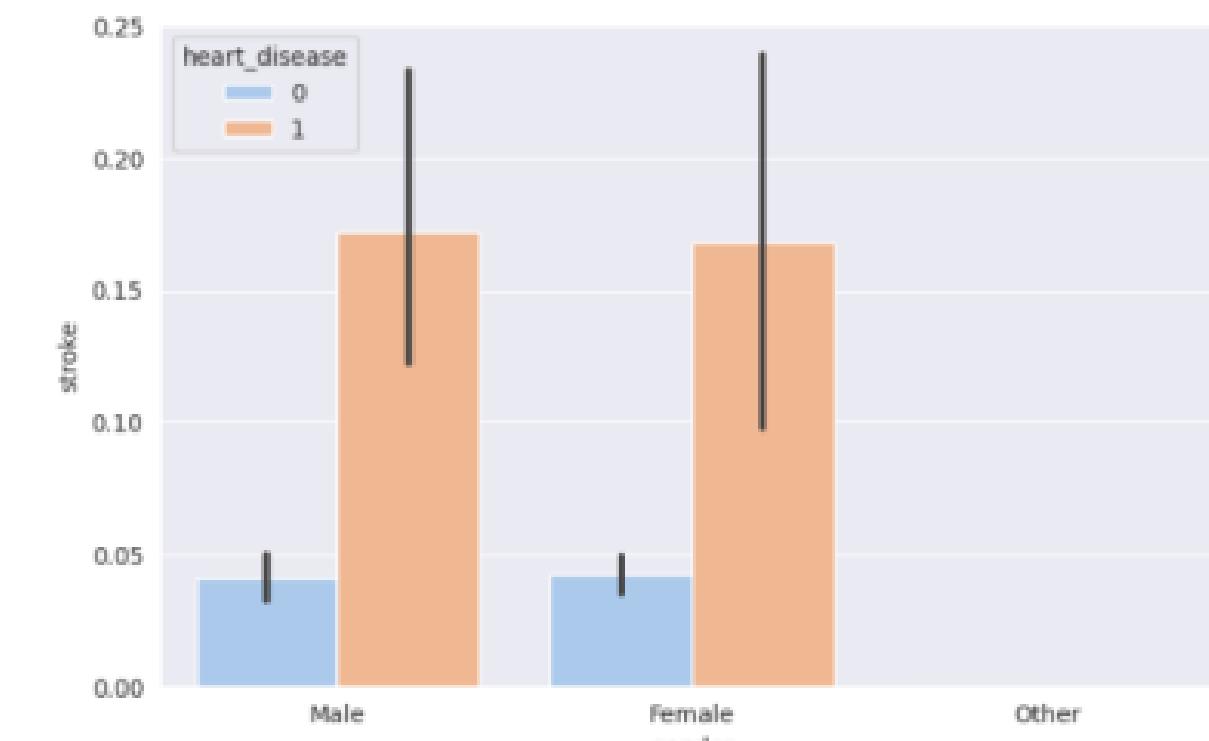
ภาพที่ 3 ความสัมพันธ์ระหว่างอาชีพ (work_type) และการเป็นโรคหลอดเลือดสมอง (stroke)

Categorical plot ของ Work type ออกแบบมาเพื่อดูว่า อาชีพไหนมีการเกิด Stroke 多 กว่า กัน จะเห็นได้ว่า work type เมื่อรวมการเกิด Stroke ก็จะ คือ 249 คน สามารถเห็นได้ว่า กลุ่ม 'Private' มีจำนวนผู้ป่วยโรคหลอดเลือดสมองมากที่สุด (149 คน) และ 'Never worked' ไม่มีผู้ป่วยโรคหลอดเลือดสมองเลย (0 คน)



Categorical plot ของ Smoking Status ออกแบบมาเพื่อดูว่า การสูบบุหรี่มีผลต่อการเกิด Stroke หรือไม่
Encoding
 0 = unknown
 1 = formerly smoked (เคยสูบบุหรี่)
 2 = never smoked (ไม่เคยสูบบุหรี่)
 3 = smokes (สูบเป็นกิจวัตร)
 จะเห็นได้ว่าบุหรี่แบบจะไม่ส่งผลให้เป็น Stroke มากรัก
 เนื่องจากว่ามีอัตราการเกิด Stroke เท่า ๆ กัน

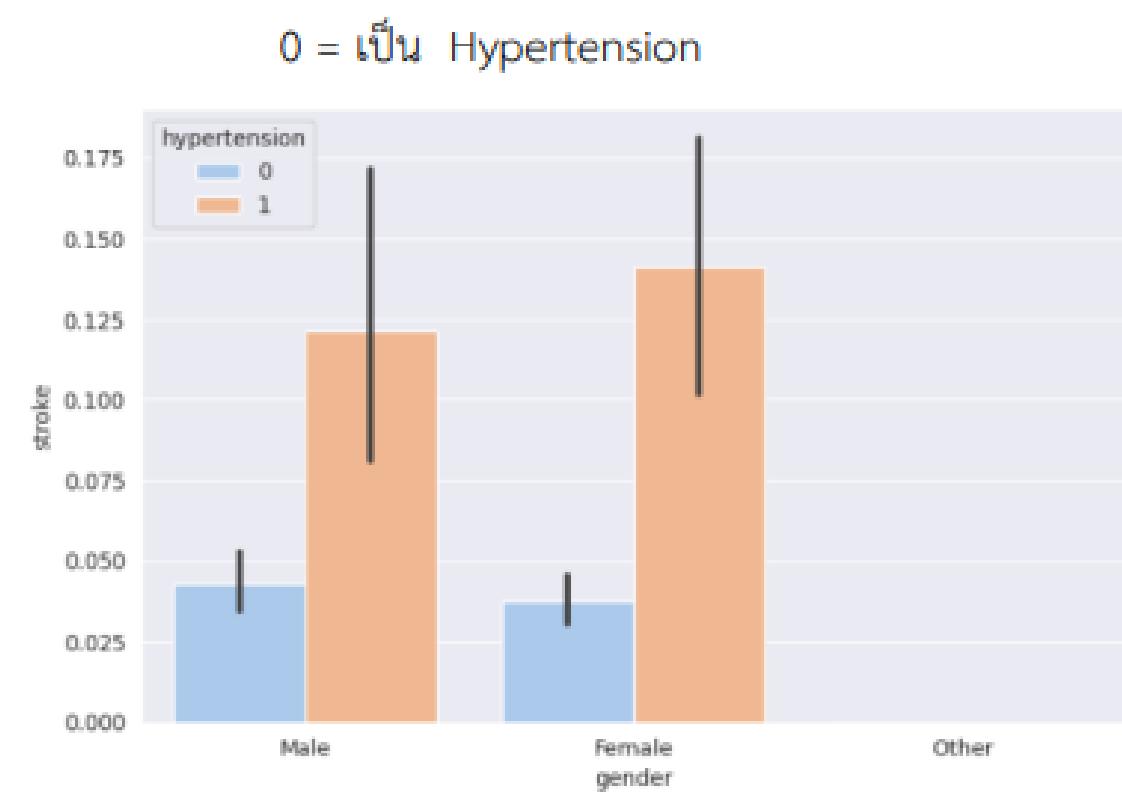
0 = เป็น Heart disease



Categorical plot เพื่อแสดงค่าของ gender(เพศ) และ Stroke เมื่อเปรียบเทียบกับ heart disease(โรคหัวใจ), residence type(ภูมิลำเนาที่อยู่) และ hypertension(โรคความดัน) คือจะสังเกตเห็นได้ว่า

- เพศ ไม่บ่งบอกว่าเกี่ยวกับ Stroke เพศชายและหญิงมีอัตราการเกิดโรคหัวใจใกล้เคียงกัน
- Hypertension(โรคความดันสูง) โรคความดันสูงมีแนวโน้มส่งเสริมให้เกิด Stroke
- Heart disease (โรคหัวใจ) โรคหัวใจมีแนวโน้มส่งเสริมให้เกิด Stroke อาจพิจารณาได้ว่า โรคความดันสูงและโรคหัวใจเป็นสิ่งที่ส่งเสริมให้เกิด Stroke

ภาพที่ 5 แสดงความสัมพันธ์ระหว่างเพศ โรคหลอดเลือดในสมอง และโรคหัวใจ



ภาพที่ 6 ความสัมพันธ์ระหว่างเพศ โรคหลอดเลือดสมอง และโรคความดันโลหิตสูง

Balancing Value จาก SMOTE จะเป็นการสร้างข้อมูลเทียม เนื่องจากมีข้อมูลที่มันไม่ค่อย balance กันสังเกตุจากต่อไปนี้ stroke = 1 = 249 คน แต่หลังจาก SMOTE จะเป็นการสร้างข้อมูลเทียมขึ้นมาเพื่อให้ตัว model มันมีตัวอย่างในการ train ขึ้นด้วย'

ก่อน OverSampling จำนวน Stroke ที่เป็น '1' คือ 249

ก่อน OverSampling จำนวน Stroke ที่เป็น '0' คือ 4861

หลัง OverSampling จำนวน features คือ (7788, 8)

หลังจาก OverSampling, จำนวนคนที่เป็น Stroke คือ 3894

หลังจาก OverSampling, จำนวนคนที่ไม่เป็น Stroke คือ 3894

Building Models



Logistic Regression

Logistic Regression

Validation Accuracy: 0.9461839530332681

Training Accuracy: 0.9525440313111546

KNearest Neighbours

KNearest Neighbor

Validation Accuracy: 0.9422700587084148

Training Accuracy: 0.9554794520547946

การประเมินผล
ตัวแบบ



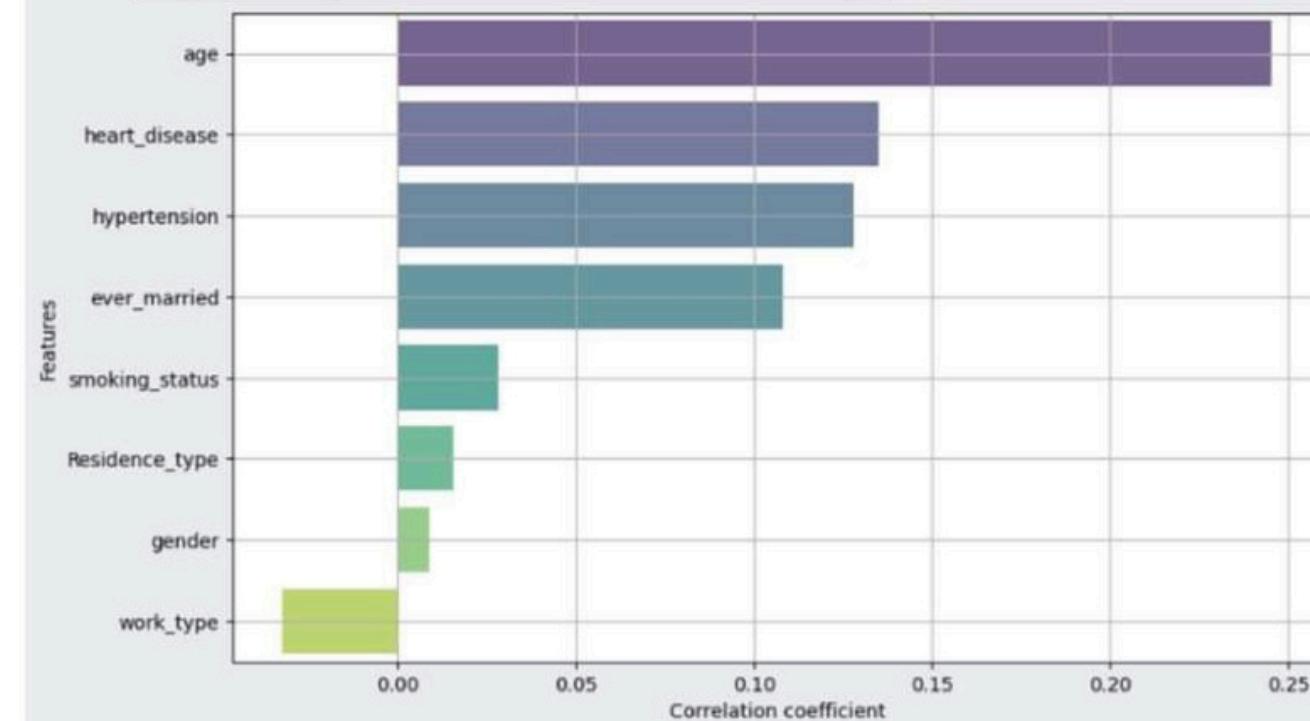
สรุปผลการดำเนินการ



จากข้อมูลที่ได้ทำการวิเคราะห์ ปัจจัยที่ส่งผลต่อการเป็นโรคหลอดเลือดในสมองจะมี อายุ โรคความดันโลหิต โรคหัวใจ ส่วนปัจจัยอื่นๆมีส่วนน้อยมากหรือไม่มีเลย

Correlation Feature w/ Stroke

Comparison of categorical features with numerical that correlated to target.



ภาพที่ 8 กราฟแท่งแสดงความสัมพันธ์ (correlation) ของตัวแปรต่างๆ และโรคหลอดเลือดในสมอง (Stroke)

Thank You Everyone

