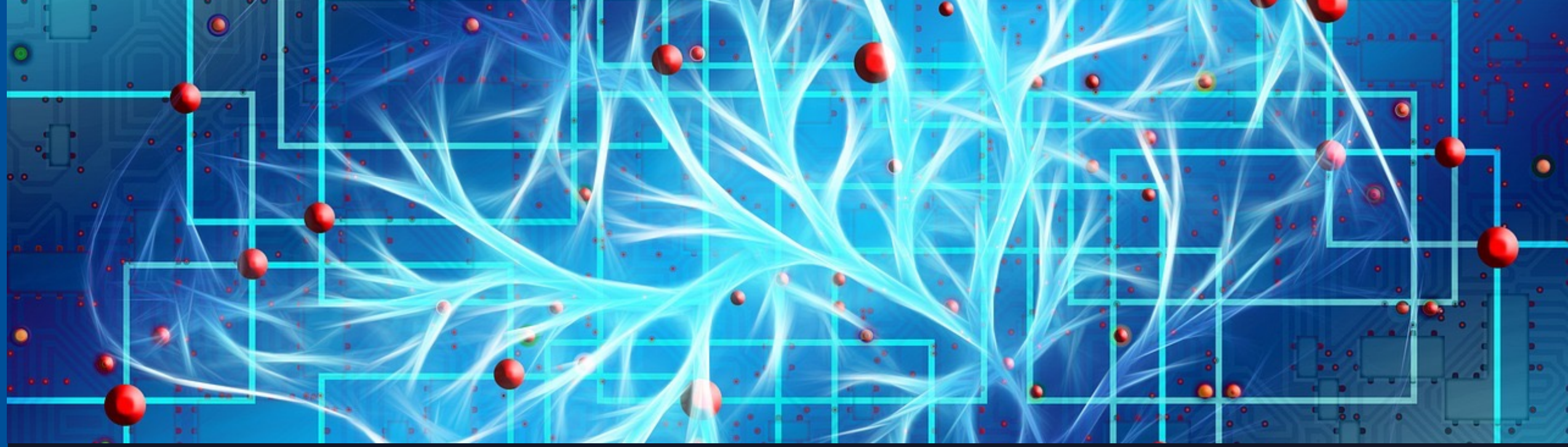


# Stroke Classification

Jordana Tepper, Troy Hendrickson, Viktoria Szontagh

# Agenda

1 | BUSINESS UNDERSTANDING

2 | DATA

3 | BEST MODEL AND RESULTS

4 | RECOMMENDATIONS

5 | LIMITATIONS AND NEXT STEPS

# Business Understanding

- **STAKEHOLDERS**

  The Mount Sinai Hospital in New York

- **THE PROBLEM**

  According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Additionally, in the US, someone has a stroke every 40 seconds.

  On top of that, strokes are a known complication of surgery.

- **THE PROJECT**

  Develop a model that acts as a preliminary assessment to determine whether a person is likely to have a stroke or not during surgery using available data. The results will determine if further screening is needed.

- **THE GOAL**

  Introduce the model that arbitrates the best results to identify patients in need of extra screening before surgery

# The Data

- **FEATURES**

  BMI, Smoking Status, Glucose level, Age, Gender, Ever Married, Residence Type, Working Status, Heart Disease, Hypertension

- **MISSING DATA**

  BMI (Body Mass Index)

- **IMBALANCE**

  95% - No stroke

  5% - Stroke

- **KAGGLE 'STROKE PREDICTION DATA SET'**

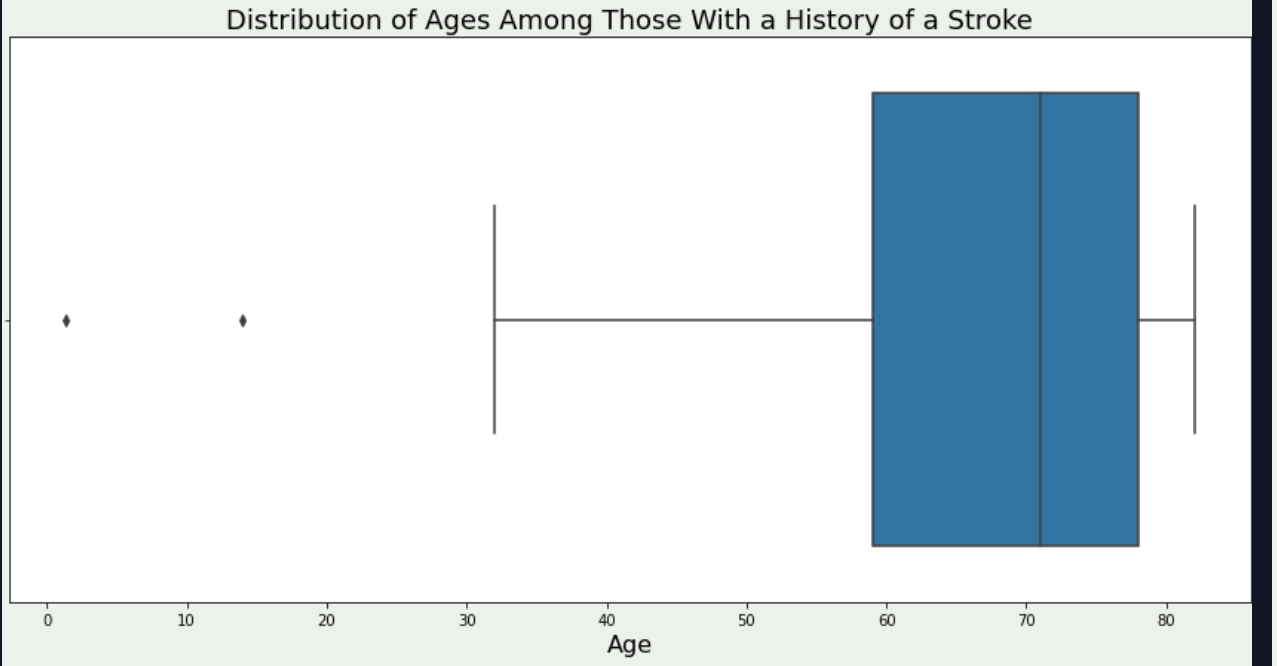
5k+

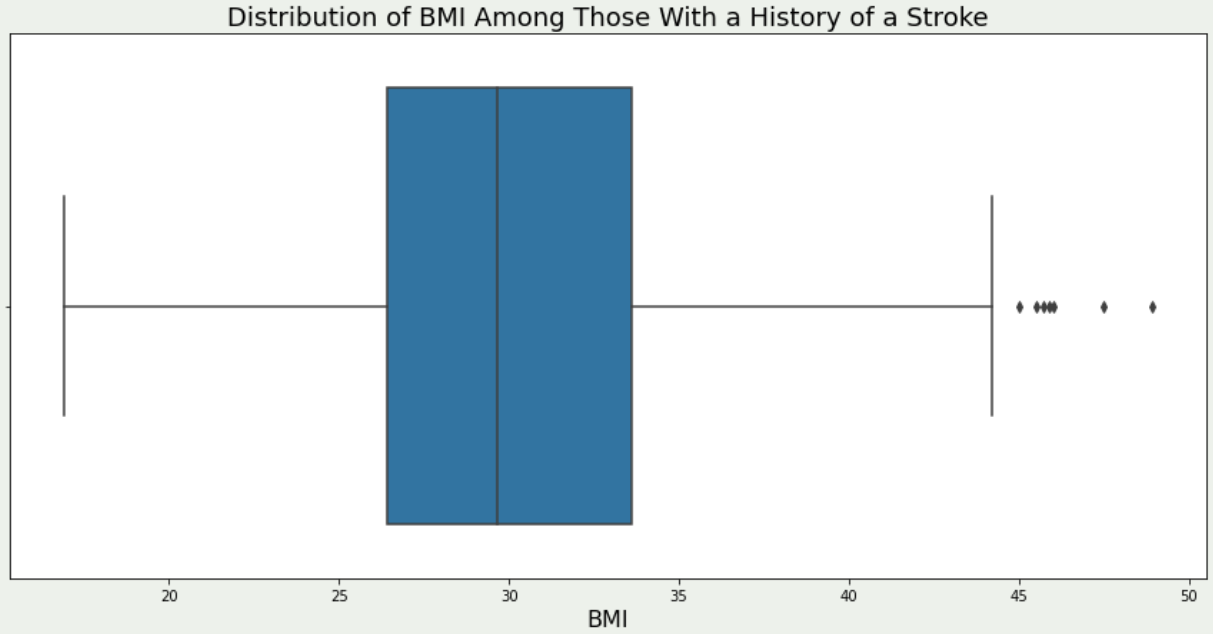
ROWS

11

COLUMNS

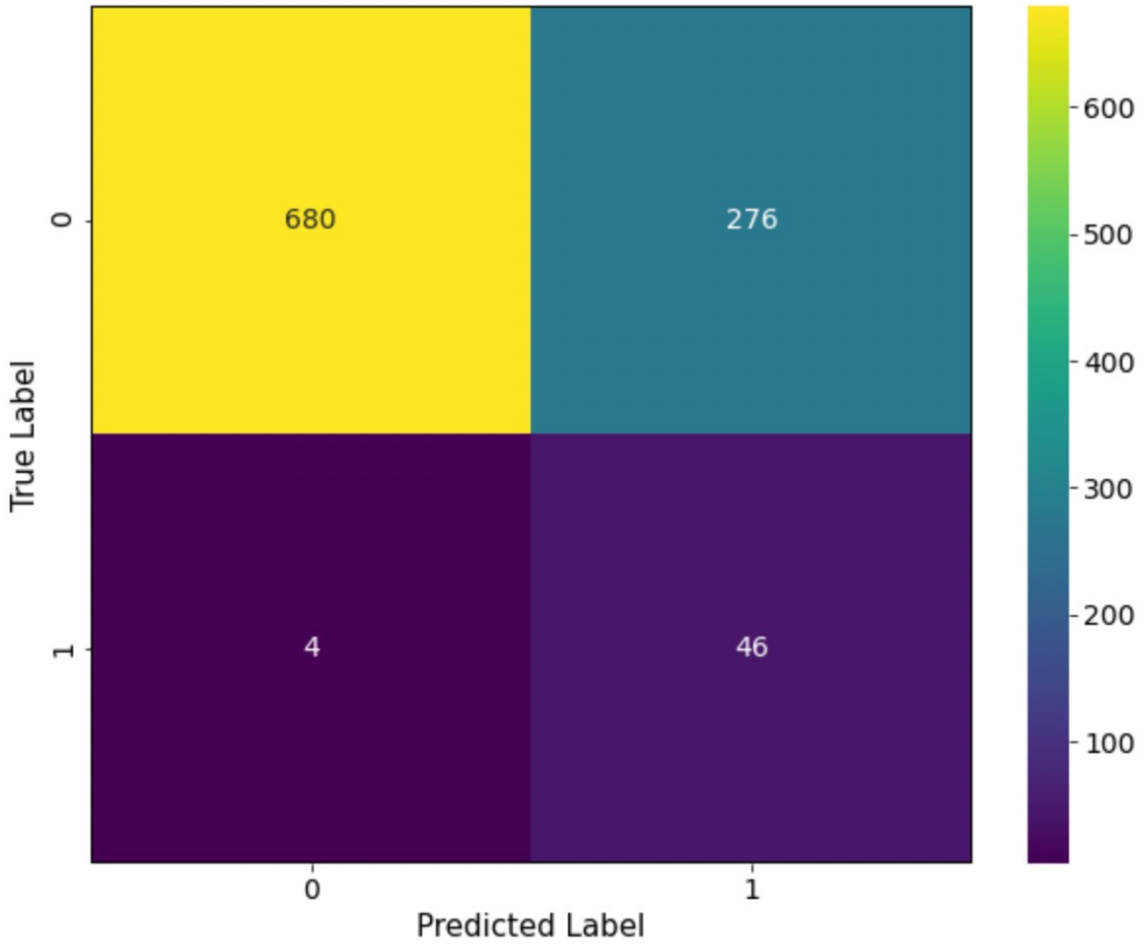
# Data Processing

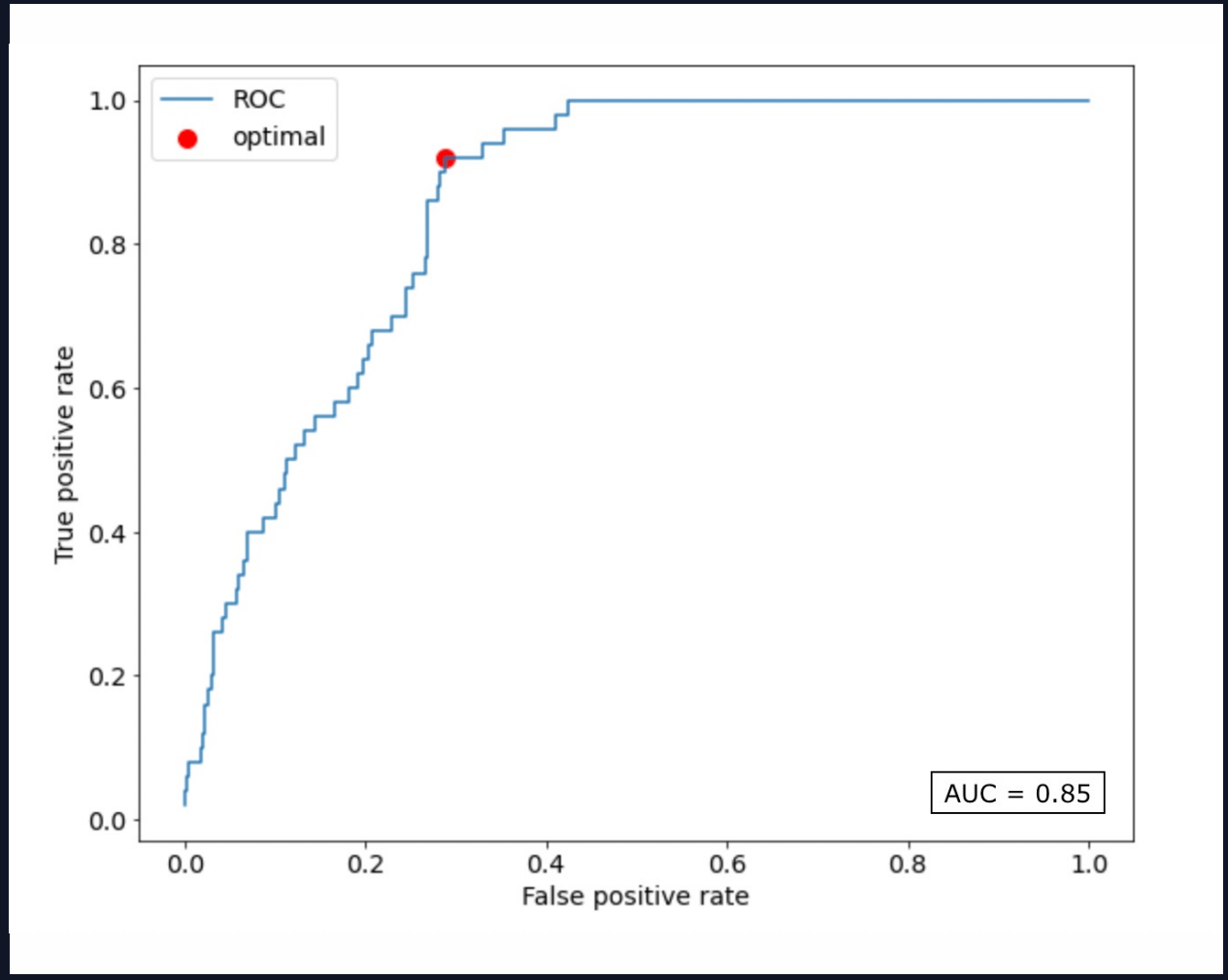


# Best Model

Logistic Regression - Optimal Threshold (0.446)



## CLASSIFICATION MODEL: LOGISTIC REGRESSION

Included optimal threshold for best results

## RECALL SCORE: 92%

The main metric used to determine the accuracy of our model

A false negative is more costly than a false positive.

## BETTER THAN BAYES?

Had best recall was our Gaussian Naive Bayes model with a recall score of 94% but a false positive rate of 0.60.

# Recommendations

## MODEL TYPE

Based on our project, we propose that logistic regression has the best classification of stroke risk and most effectively minimizes both the false negatives and false positives.

## OPTIMIZE SCREENING COSTS

By decreasing the false positives, it gives less room to the insurance company to reject claims - benefiting both the patients and the hospital.

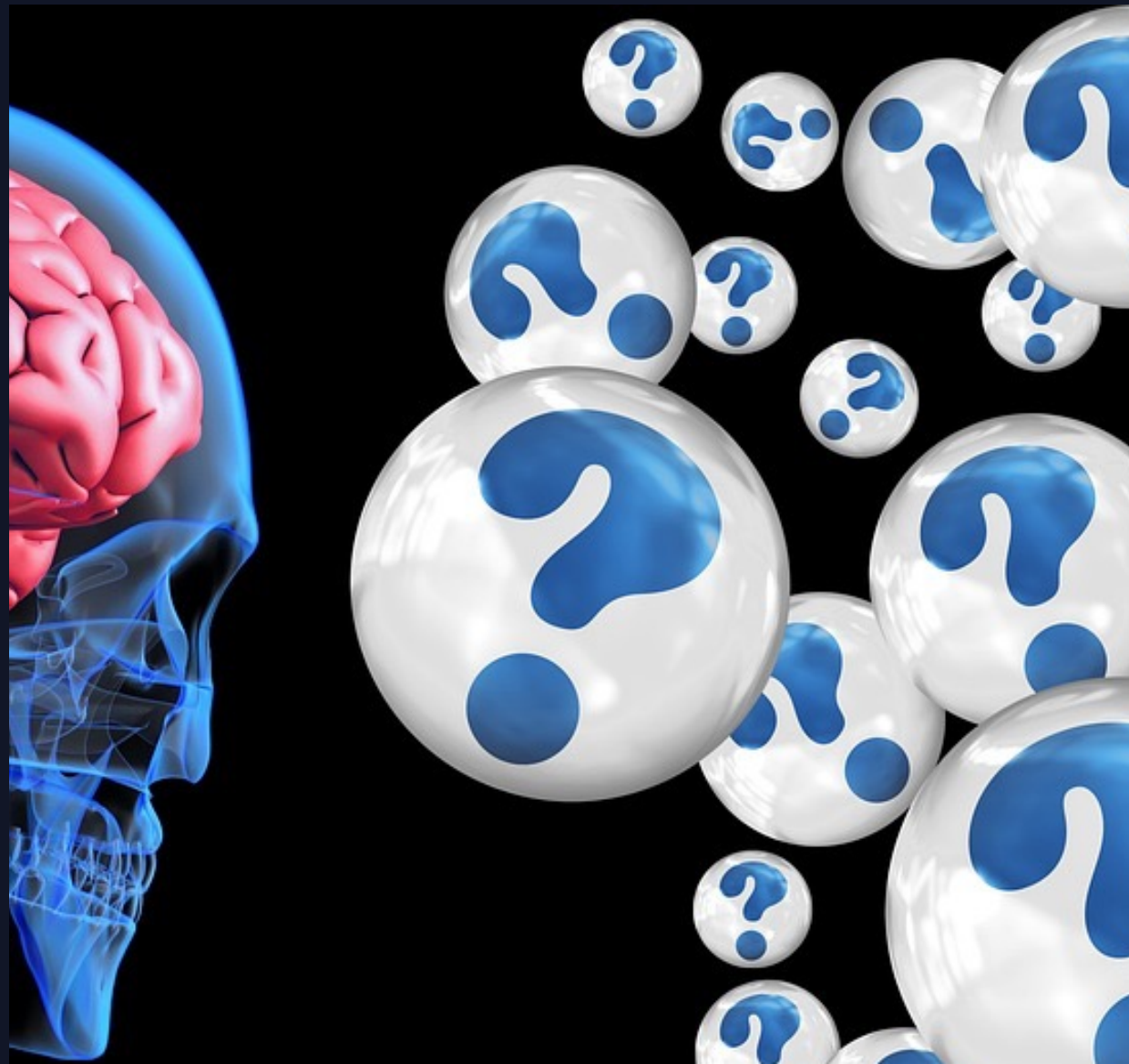
# Limitations and Next Steps

**Limitations**

- Medication is not taken into account
- Missing Data
- Unknown origin of the dataset

**Next Steps**

- Cholesterol
- Family history of stroke
- Number of strokes
- Race (i.e., the likelihood of stroke among different races)



# Contact Us


## TROY HENDRICKSON

troyhendrickson@gmail.com
gitHub.com/tkhendrix22
linkedin.com/in/troy-hendrickson

## JORDANA TEPPER

jtepper724@gmail.com
gitHub.com/jordanate
linkedin.com/in/jordana-tepper

## VIKTORIA SZONTAGH

vikkiszontagh@gmail.com
gitHub.com/vszontagh
linkedin.com/in/viktoriaszontagh