Investigating Stroke Risk Factors

From Health Data to Early Intervention

Alex Quao

Bernice Akoto

Faustina Asare

Patience Asabea Ansah

Sandra Adomako

Made with GAMMA

Table of Contents

- Problem Statement
- Objectives
- Dataset Overview
- Analysis
- Modelling Approach: Algorithms & Techniques
- Modelling Approach: Evaluation Results
- Key Takeaways
- Next Steps
- Conclusion

Made with GAMMA

The Problem: Why Stroke Prediction Matters

What is Stroke?

A stroke happens when part(s) of the brain does not get enough oxygen leading to the death of brain cells. (CDC, 2024)

Why does it matter?

Stroke is a major cause of death and disability in Ghana and the world. (BioMed Central) People are generally unaware of the stroke risk they carry.

How can data help?

About 80% of stroke cases are preventable. (Heart.Org, 2021) Timely intervention saves lives and improves outcomes.

Made with GAMMA

Objectives

Risk Assessment

Who is most at risk of stroke?

Lifestyle Patterns

What behaviors correlate with stroke risk?

Demographic Factors

How do demographics influence stroke likelihood?

Medical Factors

How do medical markers influence stroke occurrence?

Prevention

Improve early detection and prevention

Made with GAMMA

Dataset Overview

43K Records

Variables

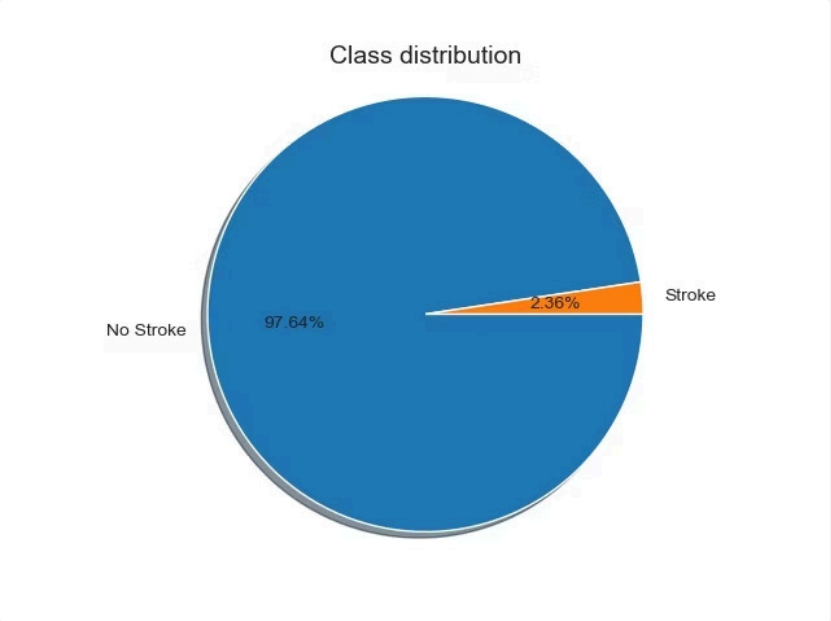
- Age
- BMI
- Smoking Status
- Gender
- Marital Status
- etc.

Imbalanced Data set

- Only 2.3% had stroke

Missing Values

- 30% of BMI values
- 3% of smoking status values



Made with GAMMA

Analysis

🧊 Age

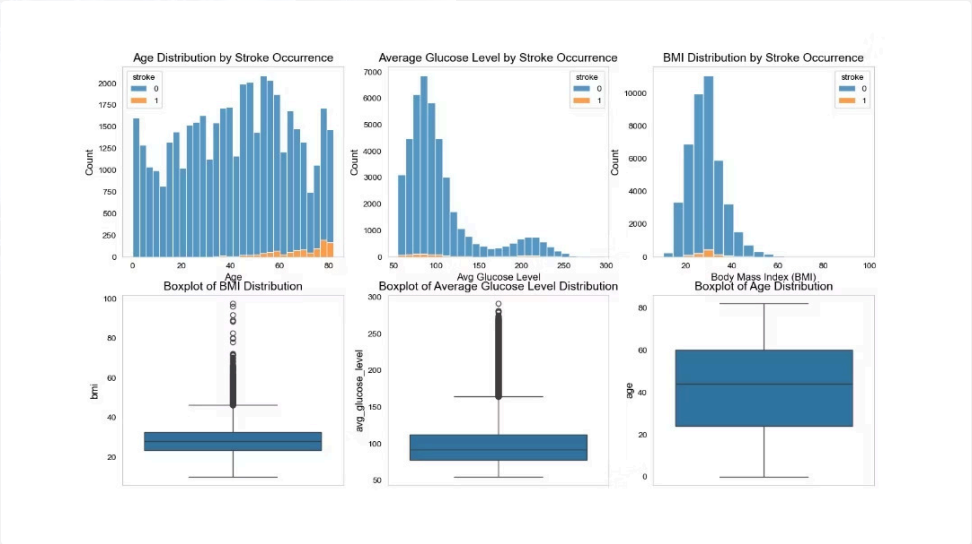
- Stroke occurrences start to significantly increase at age 40 and upwards.

🧊 Average Glucose Level

- Average glucose shows a bimodal distribution within both stroke and non-stroke populations with higher levels of stroke occurrence among High Glucose Level group.

🧊 BMI

- BMI distribution is normal.



Made with GAMMA

Analysis

👫 Gender

- Females had marginally more strokes by count alone. However, had proportionally lower stroke occurrence.

💍 Marital Status

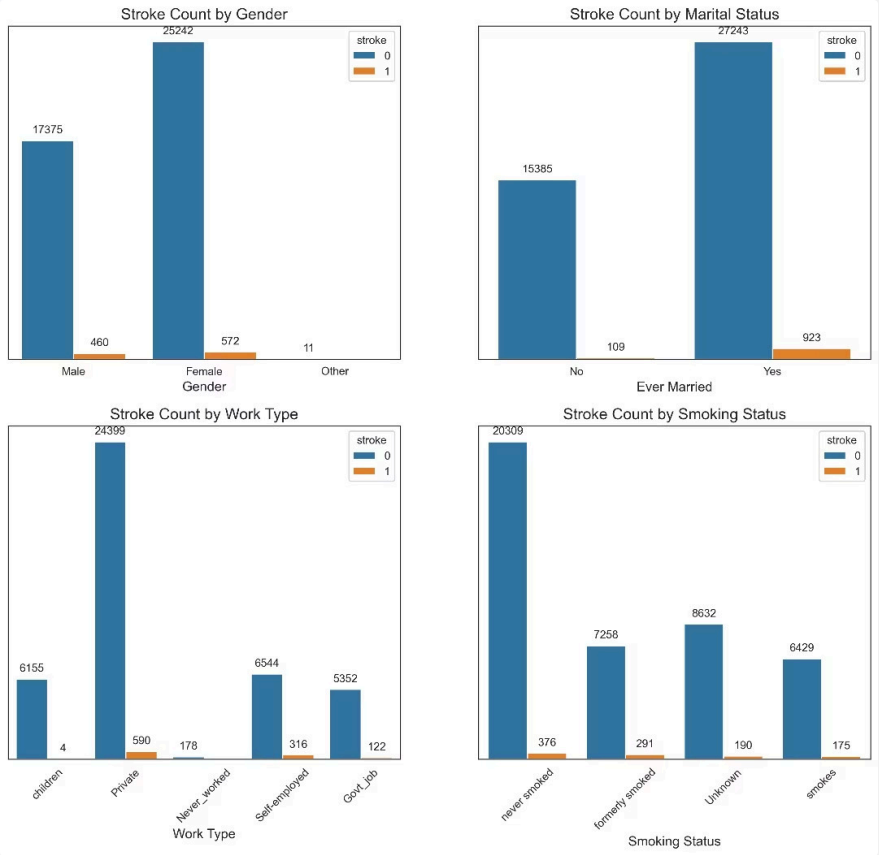
- Proportionally, people who have never been married have a lower incidence of stroke compared to those who had ever been married.

💼 Work Type

- Self-employed people proportionally have the highest stroke incidences.

🚬 Smoking Status

- Proportionally, people who have never smoked have a lower incidence of stroke compared to those who previously smoked or still smoke.

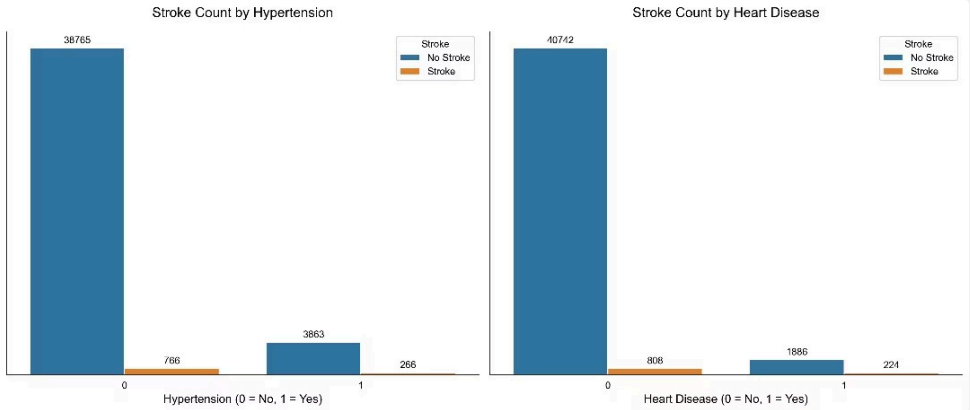


Made with GAMMA

Analysis

❤️‍🩹 Heart disease & Hypertension

- People with hypertension and heart disease experienced a higher incidence of stroke compared to those without these conditions.

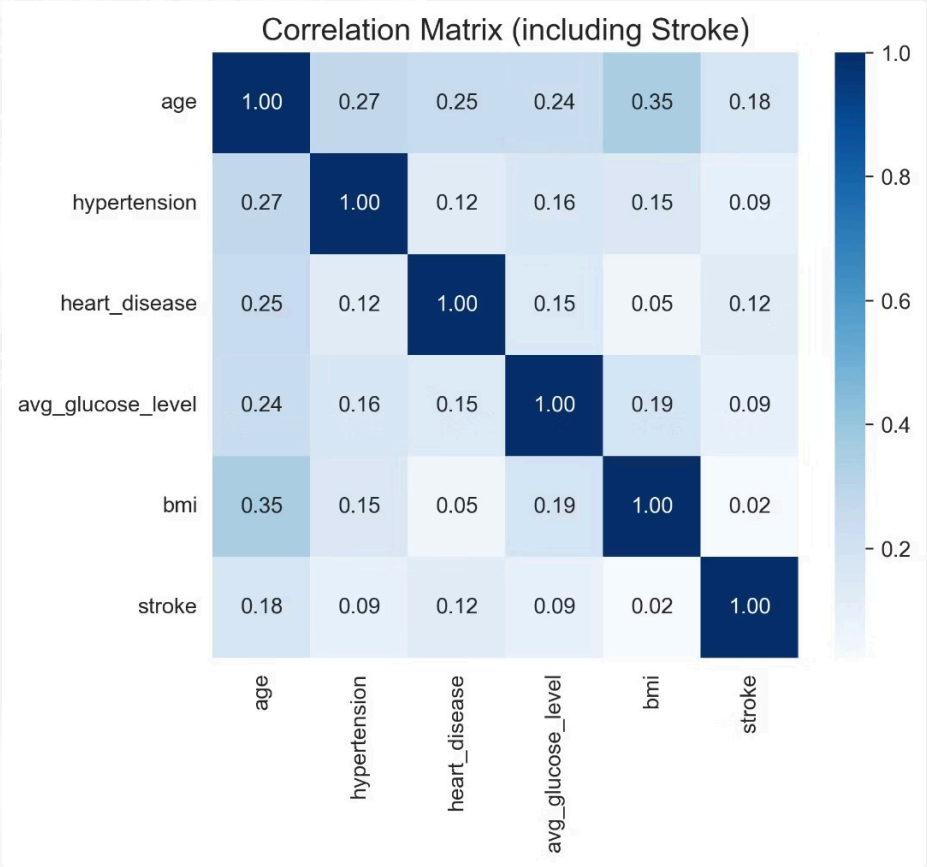


Made with GAMMA

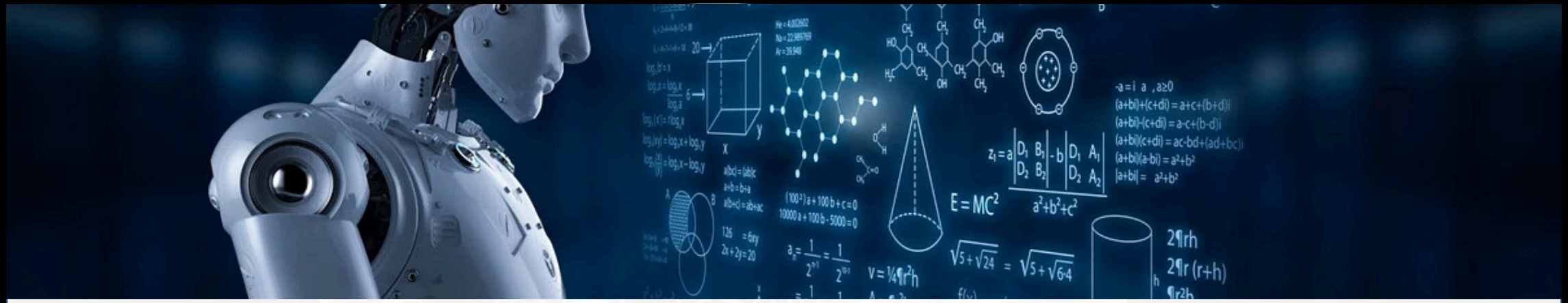
Analysis

Observations

- No attribute alone correlates strongly with stroke occurrence.
- Age has the strongest correlation (0.18) with stroke occurrence followed by heart disease (0.12) and glucose level (0.09).



Made with GAMMA



Machine Learning: Algorithms & Techniques

1

SMOTE

- To generate synthetic stroke cases to balance the data set.

2

Training Models

- Random Forest
- K-Nearest Neighbour
- Logistic Regression

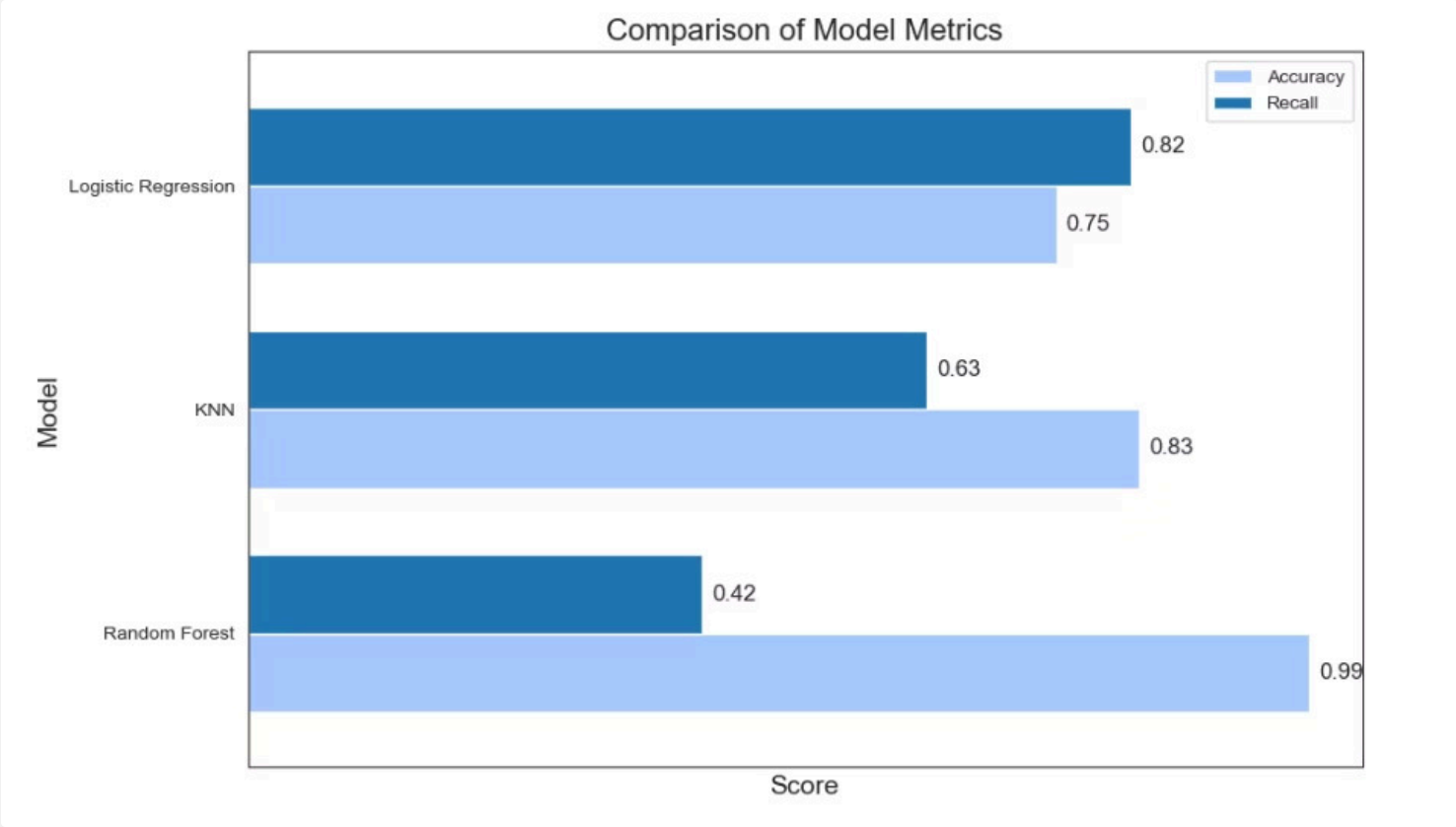
3

Model Evaluation

Test and compare the accuracy of the three models.

Made with GAMMA

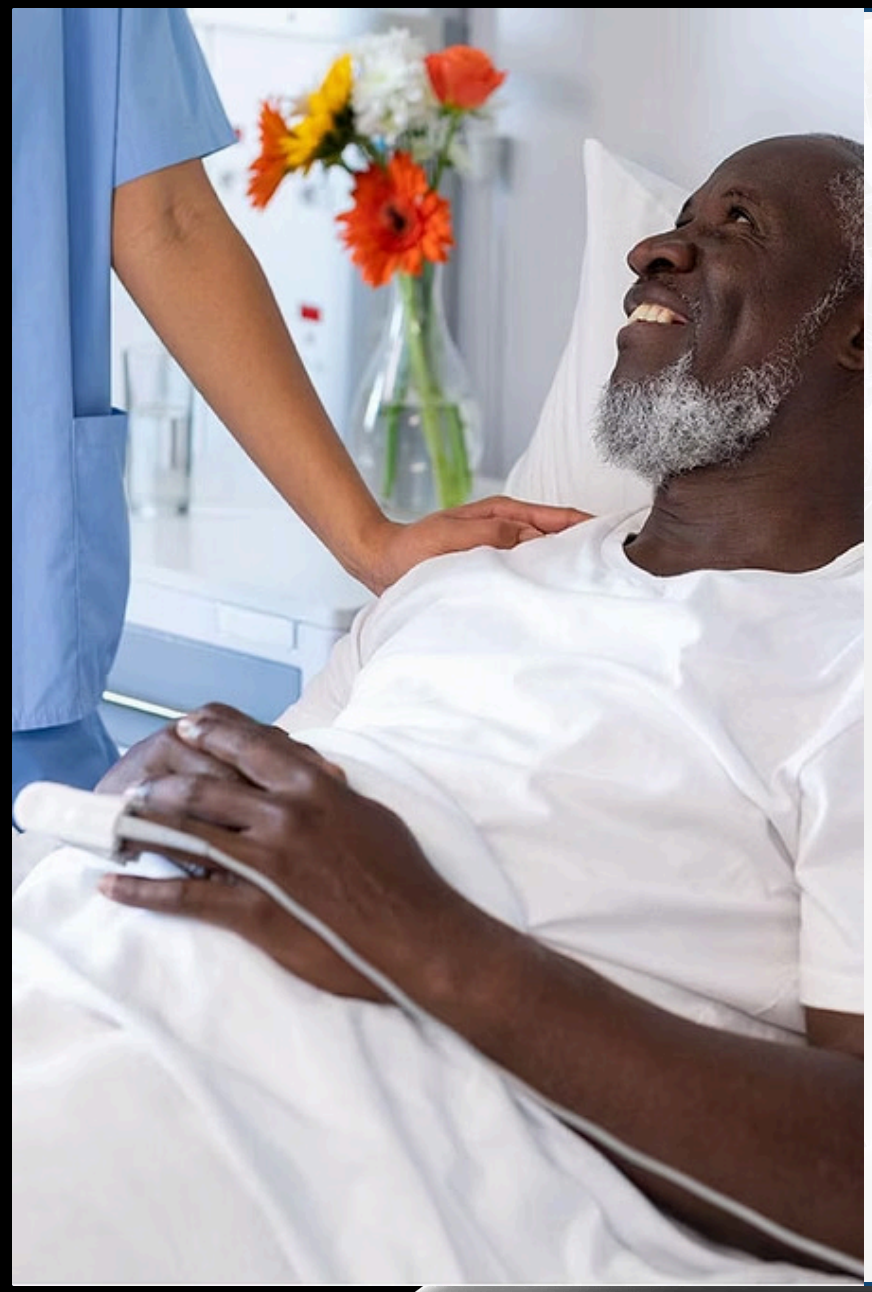
Machine Learning: Evaluation Results



Observations

- Random Forest has the best Accuracy Score
- Logistic Regression has the best Recall Score

Made with GAMMA



Key Takeaways

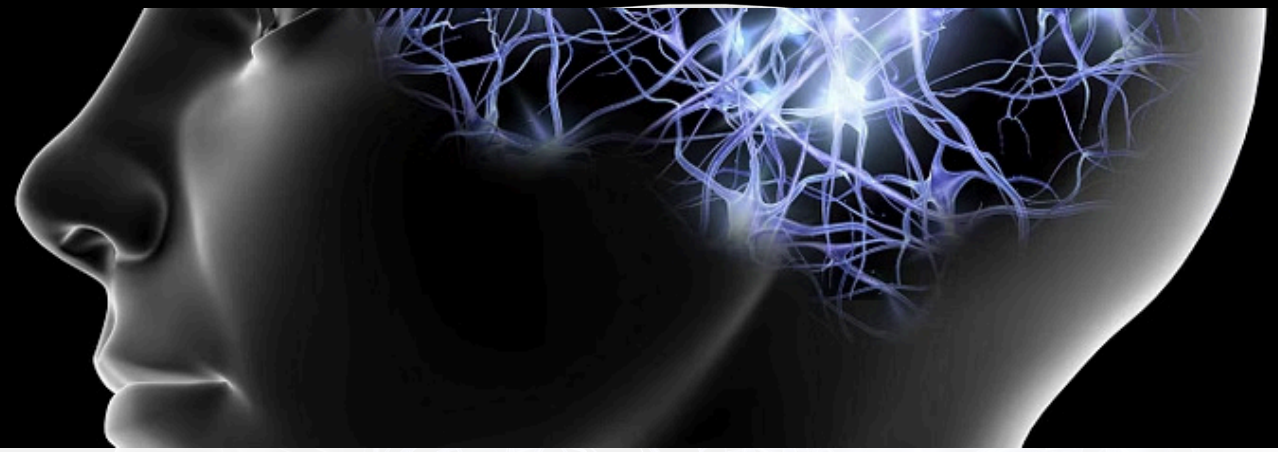
- Most important metrics to focus on for intervention programs:
  - Age : 40 and above
  - Heart Disease
  - High Glucose Level {170mg/dL to 250mg/dL}

Made with GAMMA

Next Steps

- Expand the dataset to include more real-world stroke occurrence metrics - This can help improve Accuracy and Recall Scores to make the machine learning models more suitable for clinical use.
- Address data entry omissions when collecting data for subsequent studies

Made with GAMMA



Conclusion: From Data to Prevention

Machine learning can play a role in saving lives through preventative health measures.

Thank You!!

Made with GAMMA