BRAIN STROKE MODEL PREDICTION

Author: Annbellah Mbungu

CONTENTS

- **Overview**
  - Business Objective
- **Data Understanding**
  - Stakeholder
  - The data
- **Feature Importance**
  - Age
  - Avg_glucose_levels
  - Bmi
- **Modeling**
  - Winning Model
  - Evaluation Metric
- **Recommendations**

BUSINESS PROBLEM

The primary goal of this project is to develop a predictive model that can accurately identify individuals at high risk of having a stroke based on various health and demographic factors. By leveraging this model, healthcare providers can proactively manage and mitigate stroke risks, ultimately improving patient outcomes and reducing healthcare costs associated with stroke-related treatments and complications.

DATA UNDERSTANDING

Stakeholder: Healthcare Providers

The Data

The title of this dataset is called "Brain Stroke dataset" from kaggle.com

1) gender: "Male", "Female" or "Other"

2) age: age of the patient

3) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

4) heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 5) Ever-married: "No" or "Yes"

6) work type: "children", "Govtjov", "Never worked", "Private" or "Self-employed"

7) Residencetype: "Rural" or "Urban"

8) avg glucose level: average glucose level in blood

9) BMI: body mass index

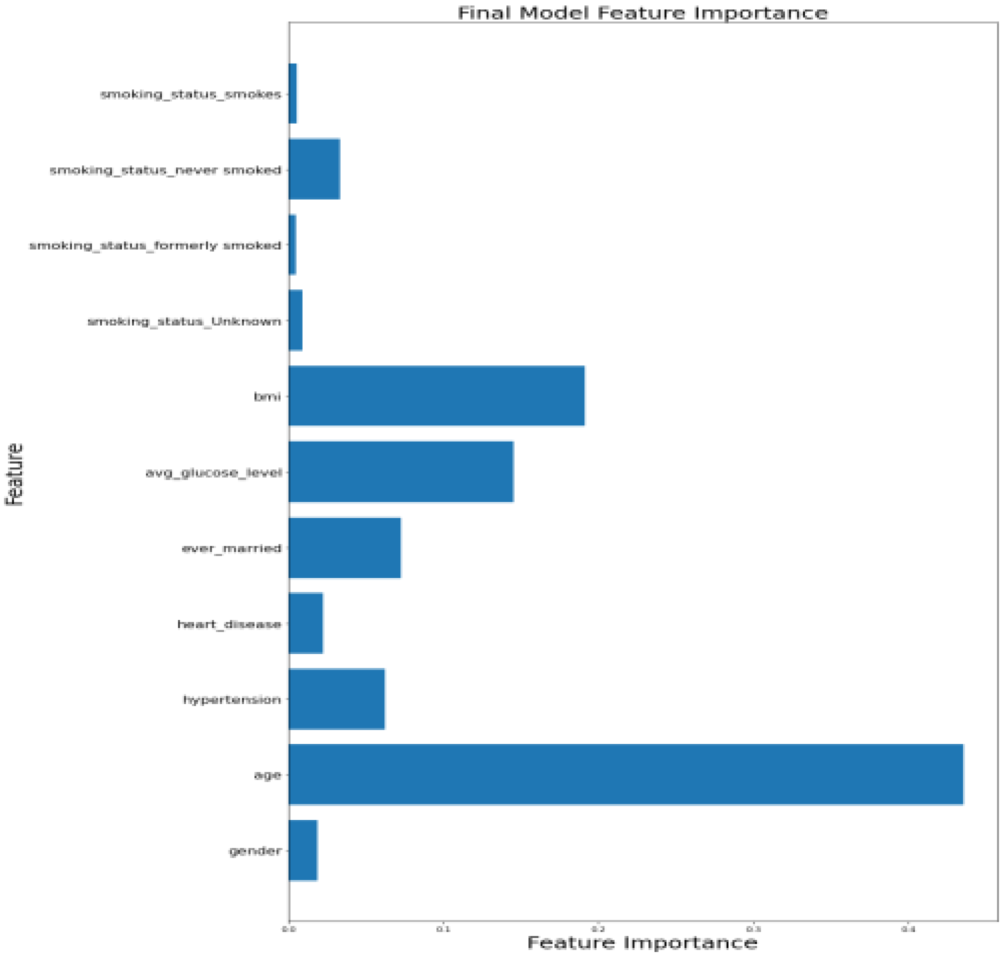
10) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

11) stroke: 1 if the patient had a stroke or 0 if not

FEATURE IMPORTANCE

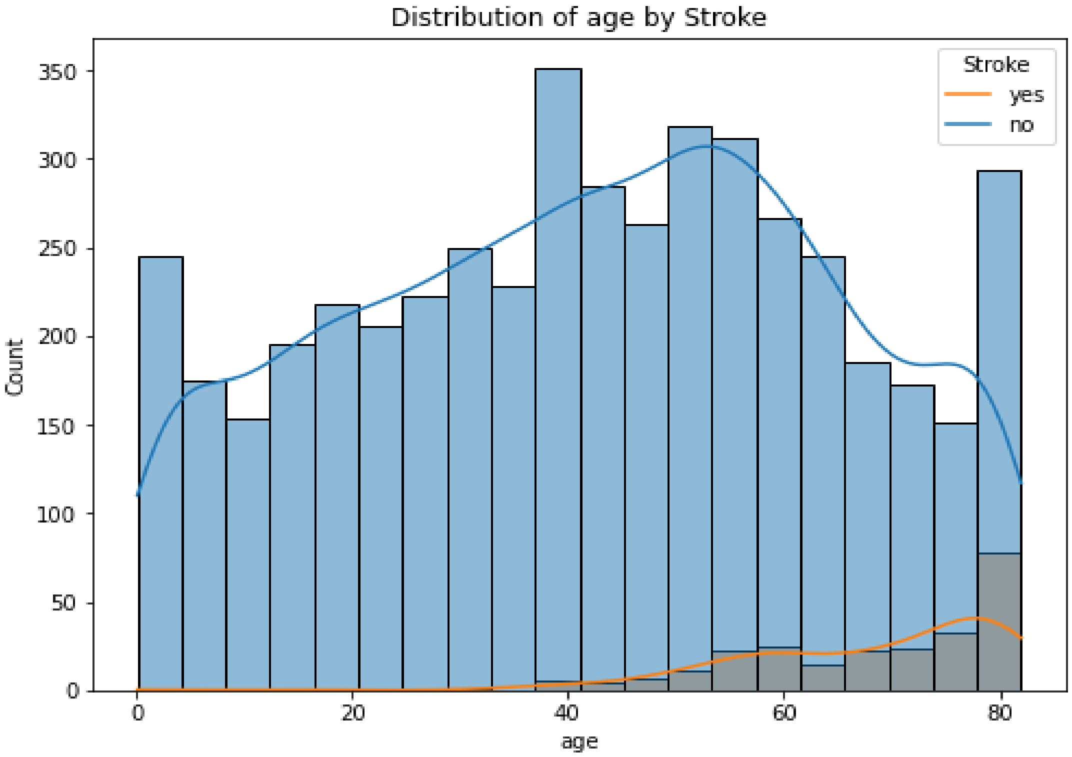
Weighted Features

1. Age
2. BMI
3. Avg_glucose_level



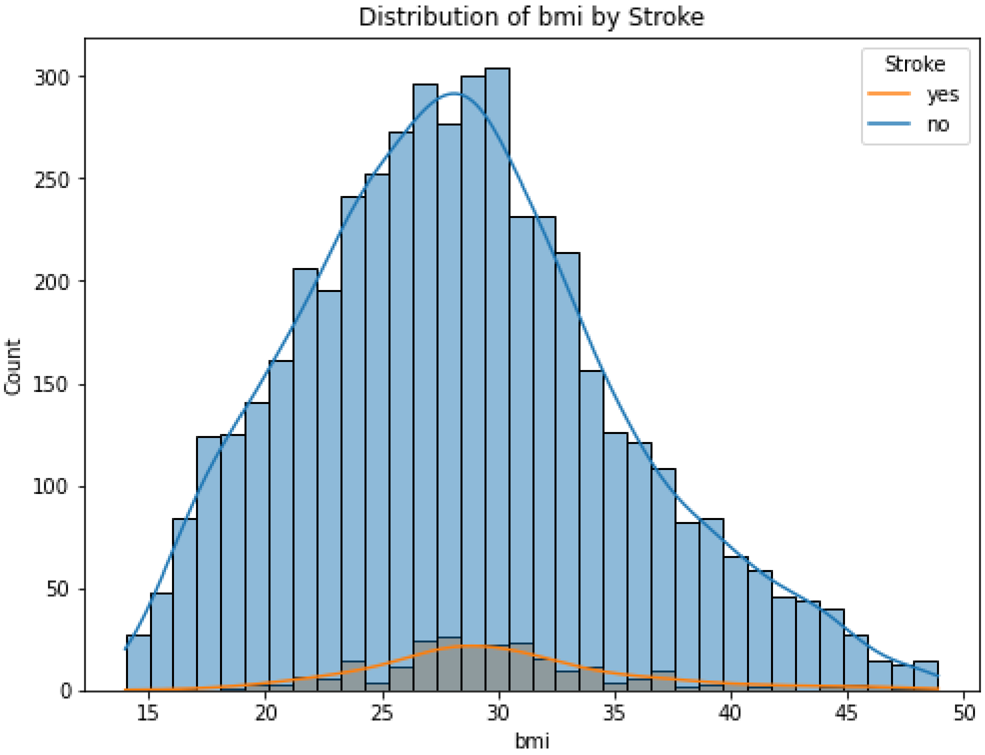
AGE

- People between age 60-80 are the most likely to have a stroke



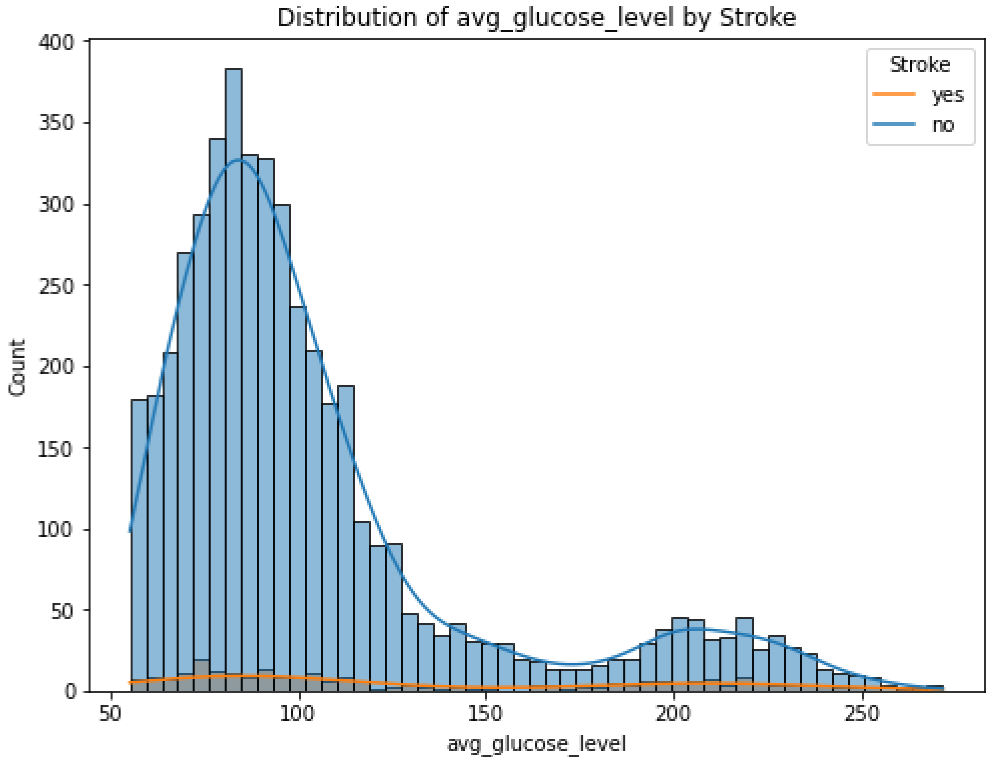
BMI

People with BMIs between 26-32 are most likely to have a stroke.



AVG_GLUCOSE_LEVELS

Most people with glucose levels between 60 – 100 are most likely to have a stroke.



THE MODELS

Logistic Regression
Recall Score (Train): 80%
Recall Score (Test): 82%
Decision Tree
Recall Score (Train): 81%
Recall Score (Test):74%
KNN
Recall Score (Train):87%
Recall Score (Test):78%
KNN With GridSearchCV
Recall Score (Train):85%
Recall Score (Test):84%
Random Forest
Recall Score (Train):90%
Recall Score (Test):84%
Random Forest with GridSearchCV:
Recall Score (Train):90%
Recall Score (Test):82%

WINNING MODEL

Considering both the recall on the training and test sets, the KNN with GridSearchCV and Random Forest (without GridSearchCV) models are the top contenders:

- KNN with GridSearchCV: 85% recall on training, 84% recall on test.
- Random Forest: 90% recall on training, 84% recall on test.

Both models perform equally well on the test set. However, the KNN with GridSearchCV has a smaller gap between training and test recall, suggesting it might generalize slightly better without as much over fitting compared to the Random Forest.

RECOMMENDATIONS

- Deploy the KNN with GridSearchCV model due to its balanced and high recall scores.

- Continuously monitor and update the model with new data to maintain performance.

- Analyze feature importance and collect additional data to enhance model accuracy.

- Use resampling techniques and cost-sensitive learning to handle imbalanced data.

- Develop a user-friendly interface and provide training for healthcare providers.

- Regularly evaluate the model for biases and ensure data privacy and security.