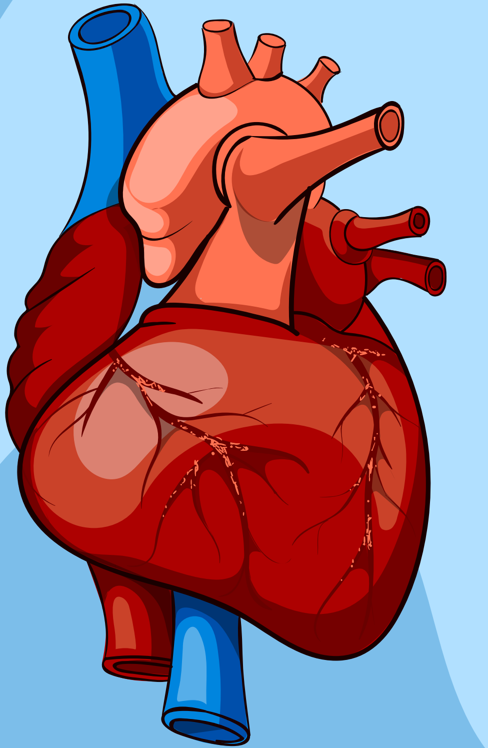


# How Likely are You to Have a Stroke?

Helena Lemay, Henry Guzman, Ethan Lui, Terry Tu

# TABLE OF CONTENTS

01

## Introduction

Description of variables and scenario

02

## Causes

What are our most correlated variables?

03

## Key Forecast

What are we trying to predict?

04

## Descriptive Models

What does our data tell us?

05

## Forecasting Models

What did we find and what sort of scenarios would this apply in?

06

## Summary

What have we learned?

# INTRODUCTION

The Impact and Importance of Possible Findings

# CAUSES OF STROKE

**130 mm**

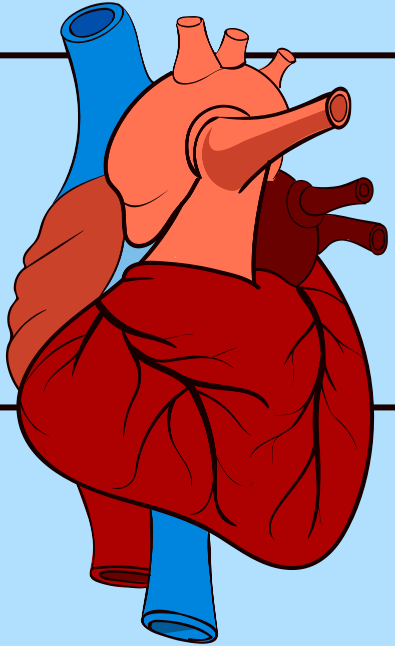
## High Blood Pressure

On inside walls of arteries

**11.3%**

## Diabetes

Glucose is too high



**1 in 5**

## Smoking

Plaque buildup in arteries

**18.2 mil**

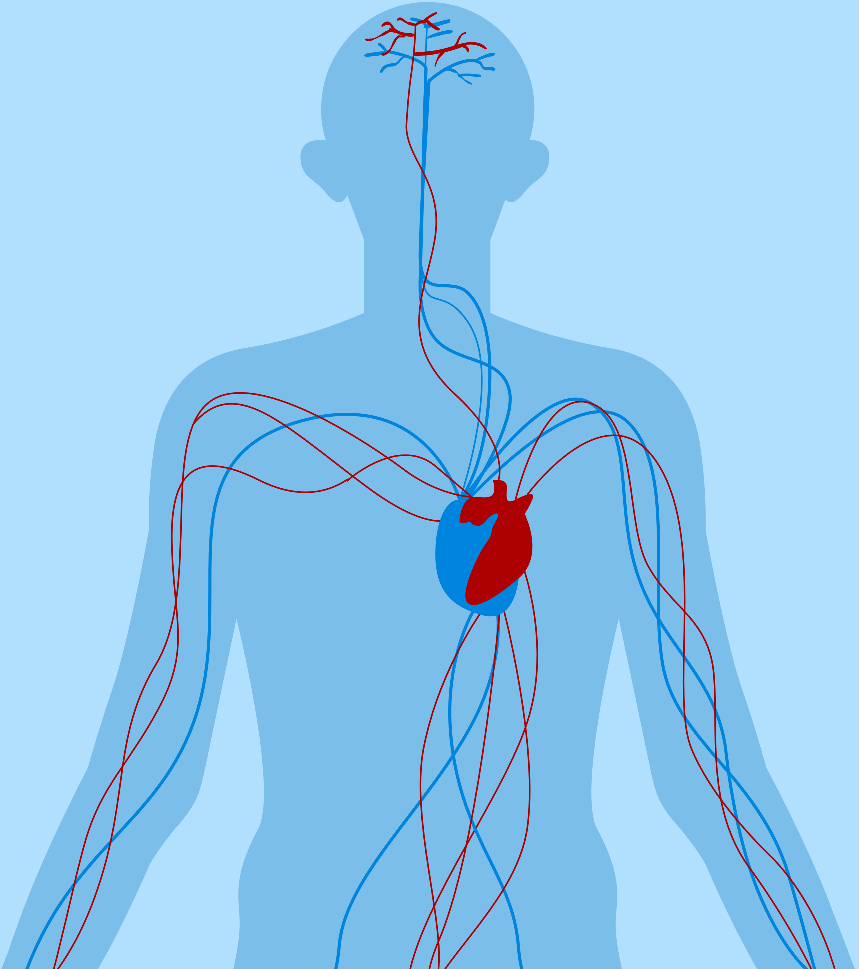
## Coronary Heart Disease

High levels of LDL cholesterol

02

# What are the key variables for predicting the occurrence of a stroke?

And why do they matter?



# Hypothesis and Background Research

# WHAT CAUSES STROKE?

01

## Expectations

- Age and Hypertension are positively correlated with stroke
- Individuals with a higher BMI are more likely to experience a stroke
- Smoking and alcohol consumption increase the chances of a stroke
- Individuals with diabetes have a higher risk of stroke
- Potential gender differences in stroke rates
- Heart conditions are correlated to a higher likelihood of a stroke

02

## Research Hypothesis

Certain risk factors, such as age, high blood pressure, smoking, diabetes, and BMI, are associated with an increased likelihood of stroke occurrence.

# Hypotheses

01

## Null hypothesis/ A priori claim

Previous studies have consistently demonstrated that age, high blood pressure, smoking, diabetes, and BMI are recognized risk factors for stroke

02

## Alternative Hypothesis

Commonly identified risk factors such as age, hypertension, smoking, high BMI, and diabetes are not attributed to an increased likelihood of a stroke.

# SECONDARY RESEARCH AND DATA

### A Priori Claim

Assumptions made based on logical conclusions for influence

### Experience

First hand experience
Slight exposure to medical field

### Research

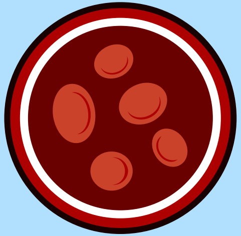
Cross referenced information through medical journals and databases

# Process and Descriptive Statistics

# 4,909 / 5,110

Observations

# DATASET VARIABLES



## Predictors

**Numeric:**
Id
Age
Average Glucose Level
BMI

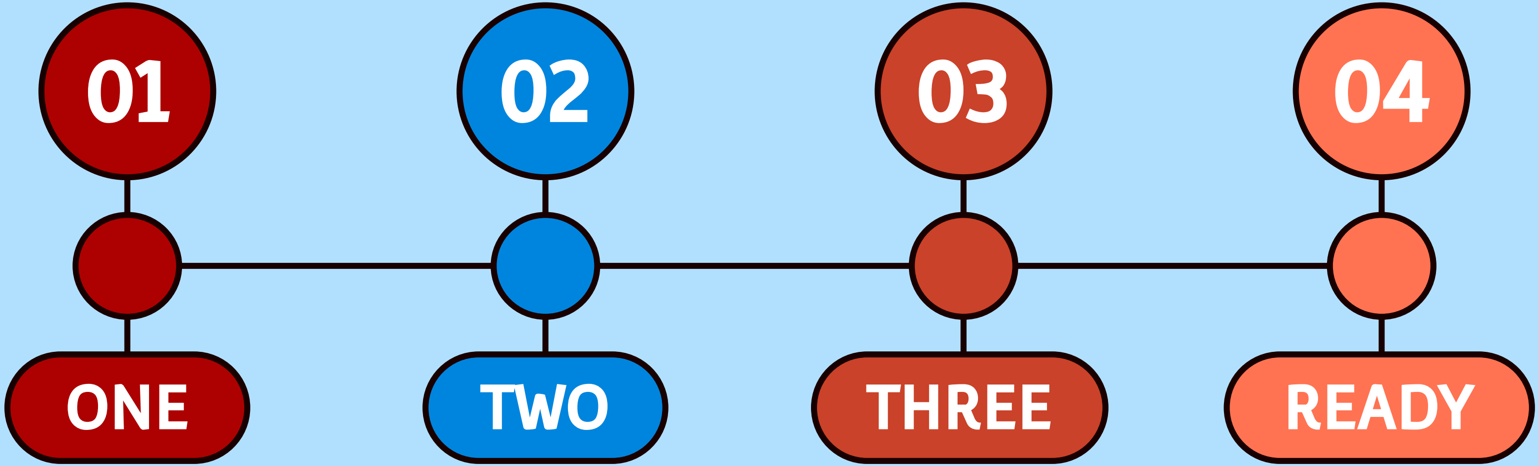
**Binary:**
Gender
Marital Status
Heart Disease
Residence Type

**Leveled:**
Work Type
Smoking Status

## Response

Stroke

# DATA CLEANING



01

ONE

**Chr→Num**

Gender
Ever_married
Residence_Type

02

TWO

**Factors**

Work_Type
Smoking_status

03

THREE

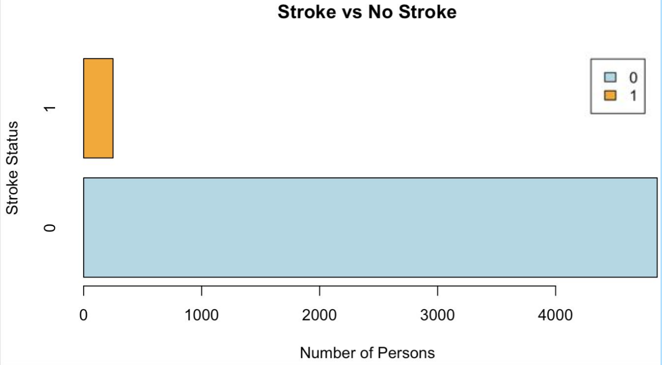
**Omit N/A**

BMI

04

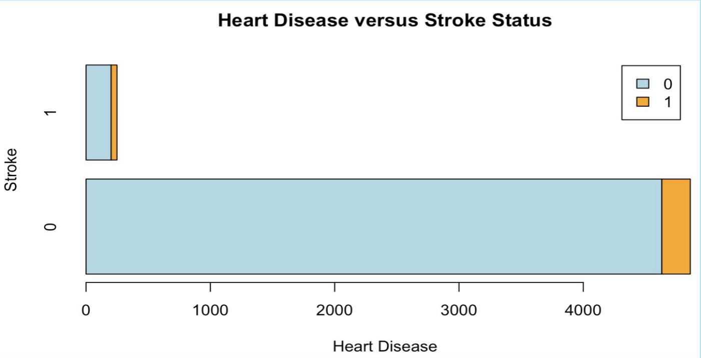
READY

# WHAT DOES THE DATA LOOK LIKE?



No Stroke

Stroke



No Heart Disease

Heart Disease

**4.9%**

**Stroke**

**95.1%**

**No Stroke**

**4.2%**

**No Heart Disease**

Out of people who <u>do not have heart disease</u>, 4.2% have had a stroke

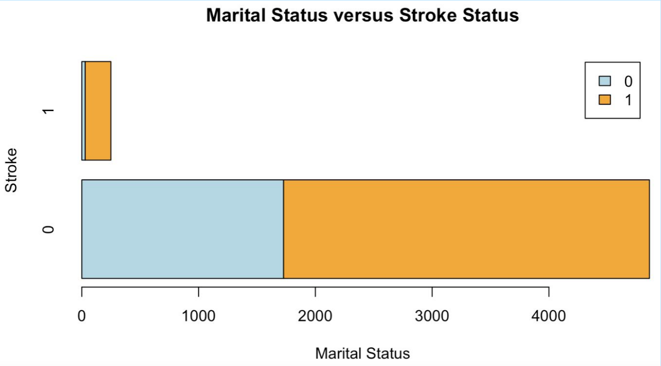
**17%**

**Heart Disease**

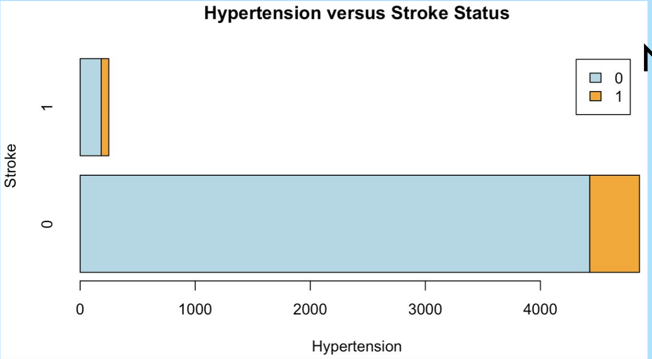
Out of people who <u>have heart disease</u>, 17% have had a stroke

0.2980024

# WHAT DOES THE DATA LOOK LIKE?



Not Married
Married



No Hypertension
Hypertension

**1.7%**

**Not Married**

Out of people who <u>are not married</u>, 1.7% have had a stroke

**6.6%**

**Married**

Out of people who <u>are married</u>, 6.6% have had a stroke

**4%**

**No Hypertension**

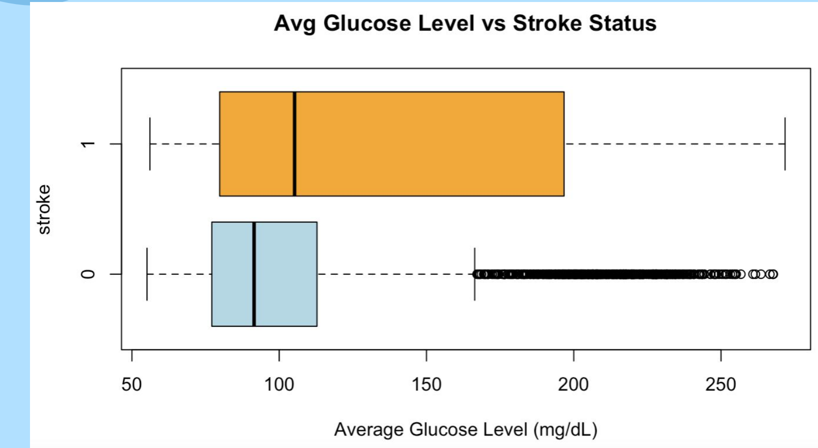
Out of people who <u>do not have hypertension</u>, 4% have had a stroke

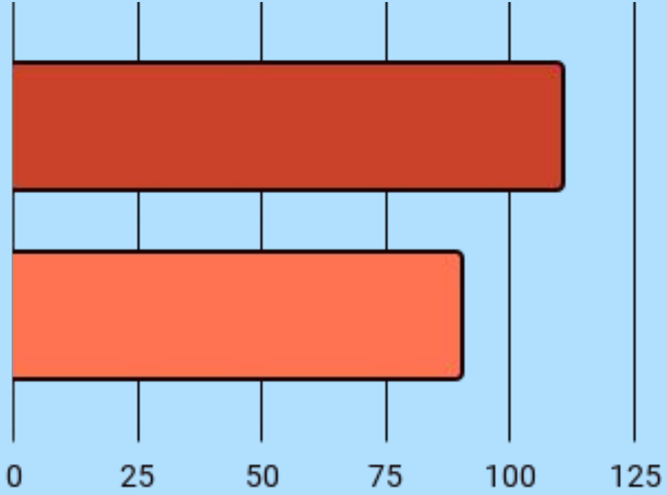
**13.3%**

**Hypertension**

Out of people who <u>have hypertension</u>, 13.3% have had a stroke

# WHAT DOES THE DATA LOOK LIKE?





Median Level of Glucose

**110** **STROKE**

Median level of glucose

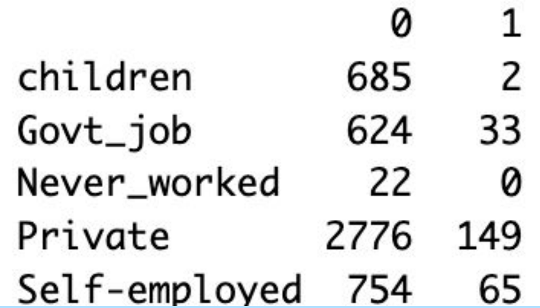
**90** **NO STROKE**

Median level of glucose

# WHAT DOES THE DATA LOOK LIKE?

No Stroke Stroke

Work Type



```
                 0    1
children       685    2
Govt_job       624   33
Never_worked    22    0
Private       2776  149
Self-employed  754   65
```

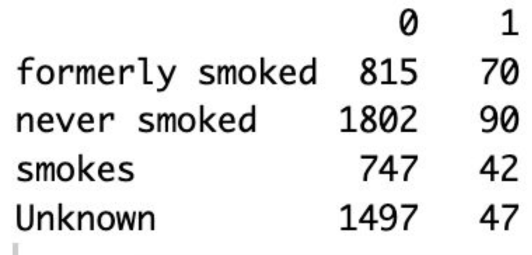
**7.9%** **92.1%**

**Stroke** **No Stroke**

Out of people who are self-employed, 7.9% have had a stroke

No Stroke Stroke

Smoke Status



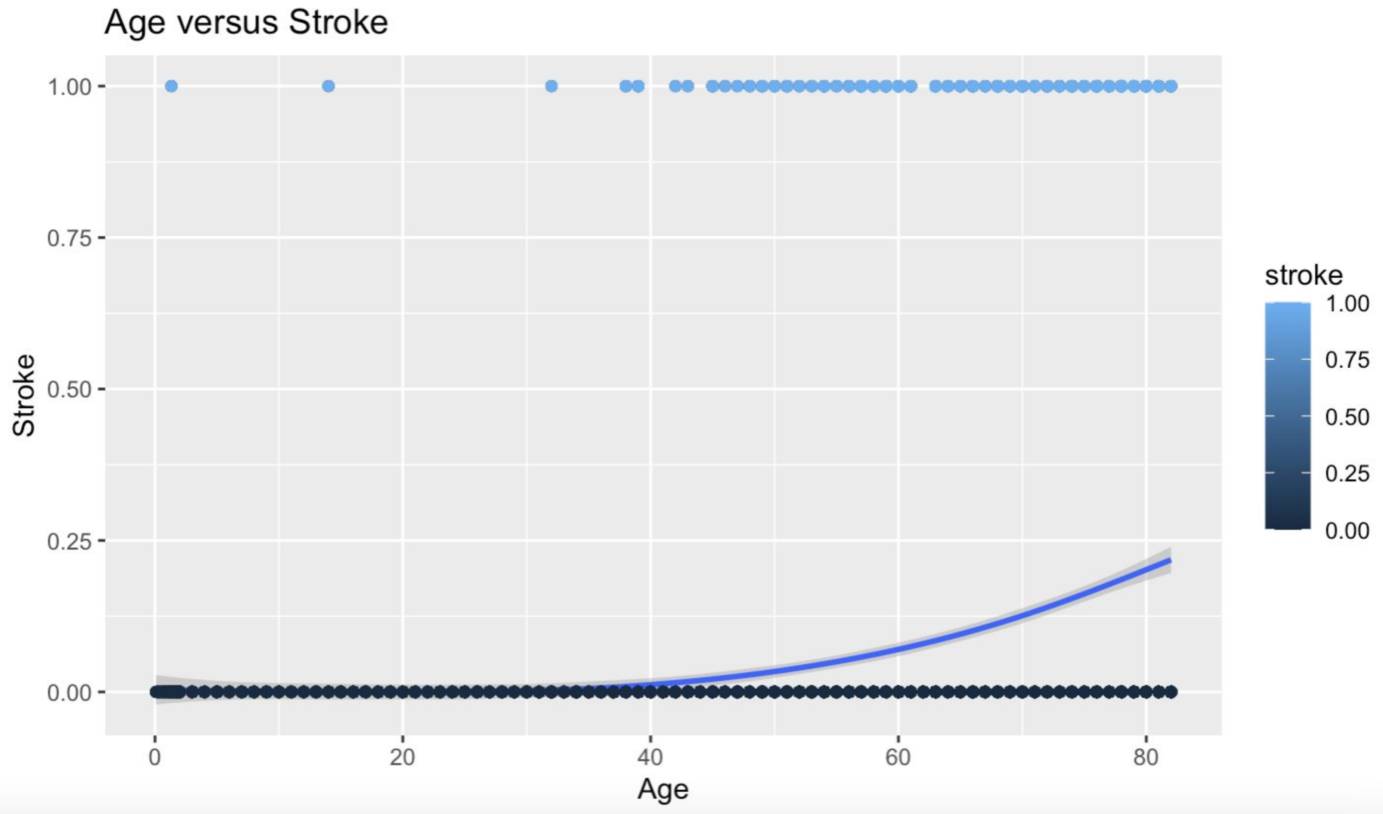
```
                    0   1
formerly smoked   815  70
never smoked     1802  90
smokes            747  42
Unknown          1497  47
```

**7.9%** **92.1%**

**Stroke** **No Stroke**

Out of people who have formerly smoked, 7.9% have had a stroke

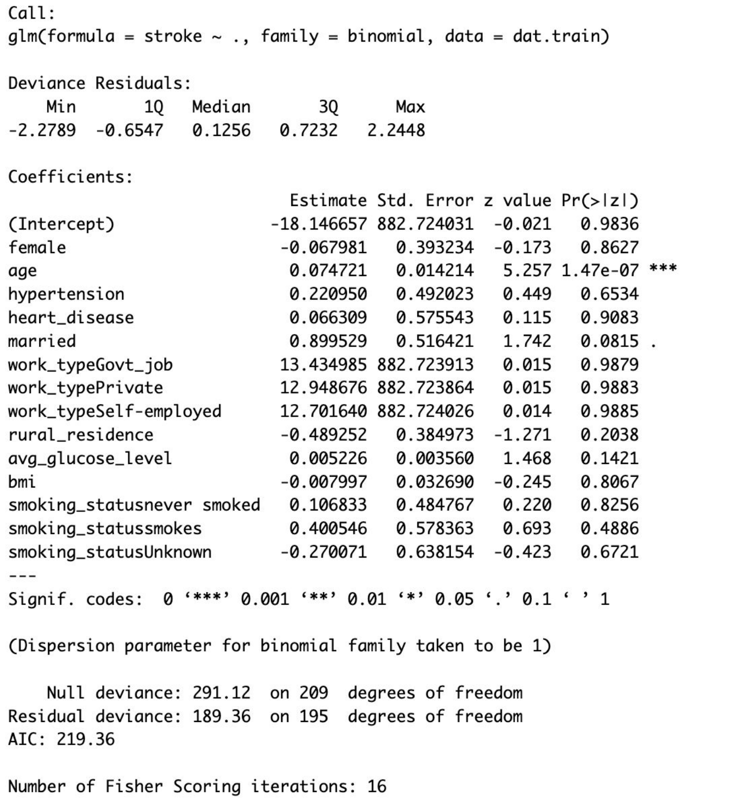
# WHAT DOES THE DATA LOOK LIKE?



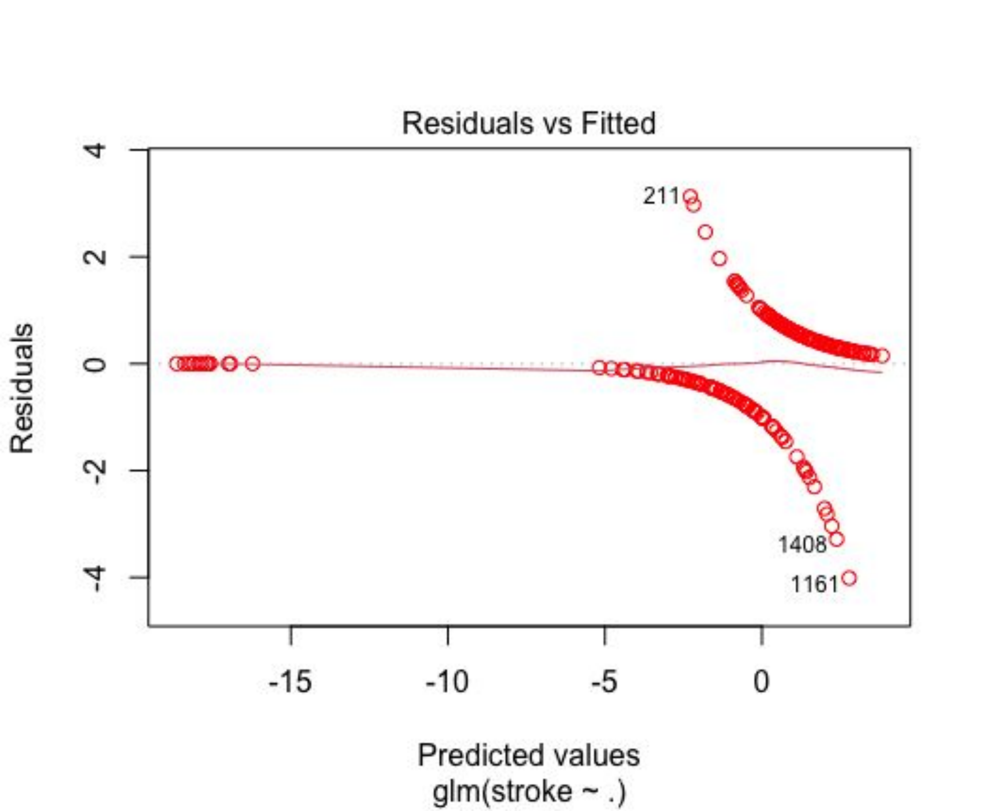
# Results from Analytical Analysis

# Logistic Regression

# How does our regression compare?

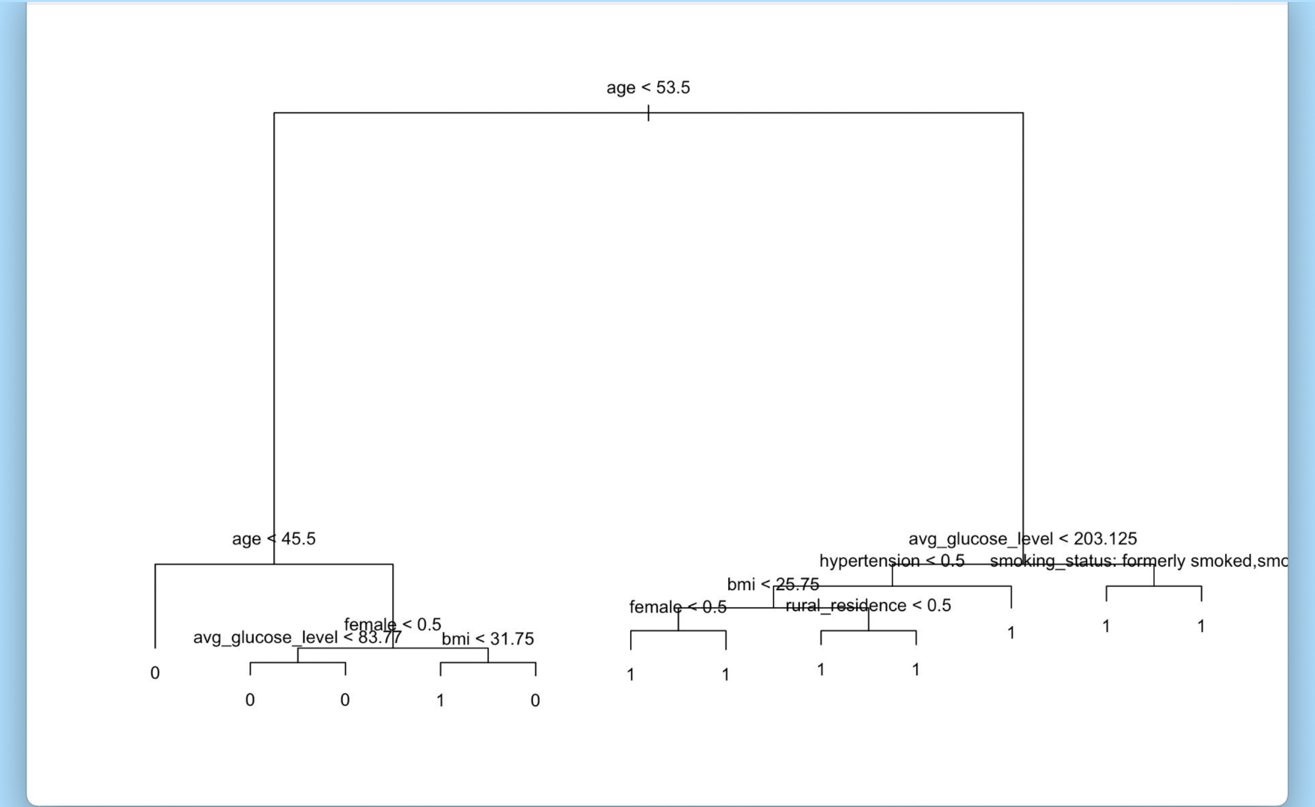


```
Call:
glm(formula = stroke ~ ., family = binomial, data = dat.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2789  -0.6547   0.1256   0.7232   2.2448

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -18.146657 882.724031  -0.021   0.9836
female                      -0.067981   0.393234  -0.173   0.8627
age                          0.074721   0.014214   5.257 1.47e-07 ***
hypertension                 0.220950   0.492023   0.449   0.6534
heart_disease                0.066309   0.575543   0.115   0.9083
married                      0.899529   0.516421   1.742   0.0815 .
work_typeGovt_job           13.434985 882.723913   0.015   0.9879
work_typePrivate            12.948676 882.723864   0.015   0.9883
work_typeSelf-employed      12.701640 882.724026   0.014   0.9885
rural_residence             -0.489252   0.384973  -1.271   0.2038
avg_glucose_level            0.005226   0.003560   1.468   0.1421
bmi                         -0.007997   0.032690  -0.245   0.8067
smoking_statusnever smoked   0.106833   0.484767   0.220   0.8256
smoking_statussmokes         0.400546   0.578363   0.693   0.4886
smoking_statusUnknown       -0.270071   0.638154  -0.423   0.6721
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 291.12  on 209  degrees of freedom
Residual deviance: 189.36  on 195  degrees of freedom
AIC: 219.36

Number of Fisher Scoring iterations: 16
```



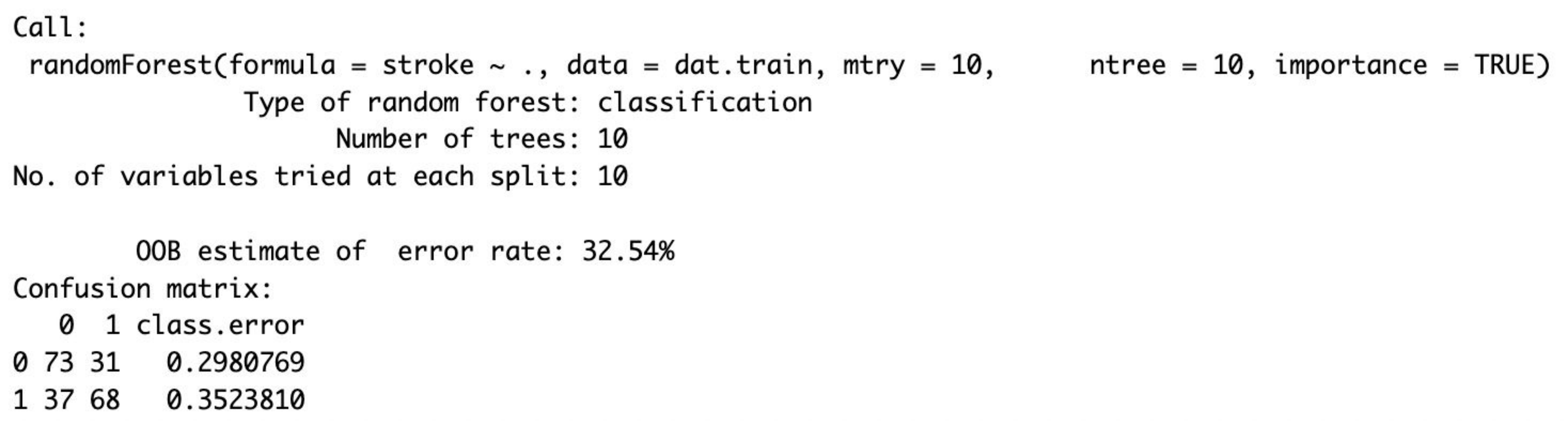
# Classification Tree

# HOW DO YOU CLASSIFY OBSERVATIONS?



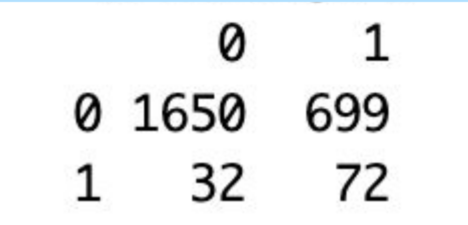
# Forecasts and Predictions

# MAKING PREDICTIONS



```
Call:
 randomForest(formula = stroke ~ ., data = dat.train, mtry = 10,      ntree = 10, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 10
No. of variables tried at each split: 10

        OOB estimate of  error rate: 32.54%
Confusion matrix:
   0  1 class.error
0 73 31   0.2980769
1 37 68   0.3523810
```

# Prediction – Classification Tree



```
     0    1
0 1650  699
1   32   72
```

# Prediction – Classification Tree

01

## Associated Forecasting Error

| | |
|---|---|
| **Prediction Error** | Prediction error of 0.2980769. |

02

## Model Performance Evaluation

| | |
|---|---|
| **Accuracy: 70.2%** | Moderate to low confidence in performance of the classification tree model. |

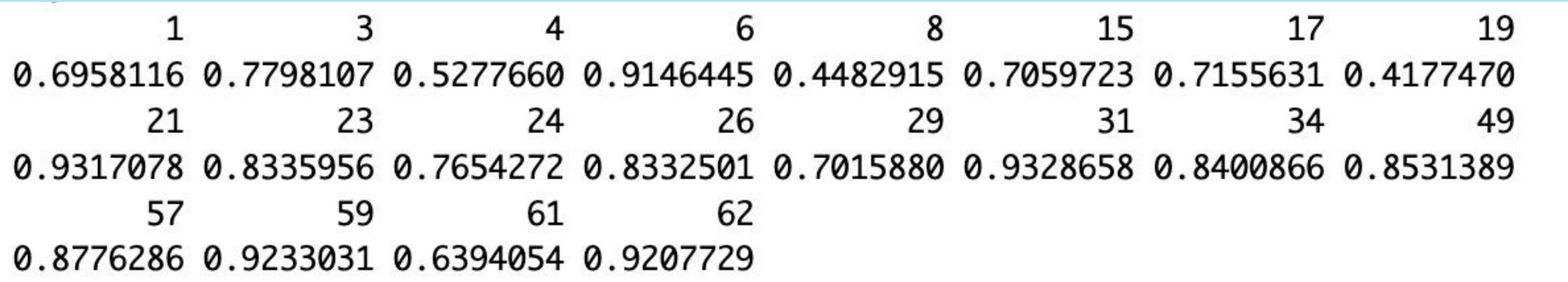
03

## Considerations

| | |
|---|---|
| **Factors** | Pruning reduced available predictors for consideration. |

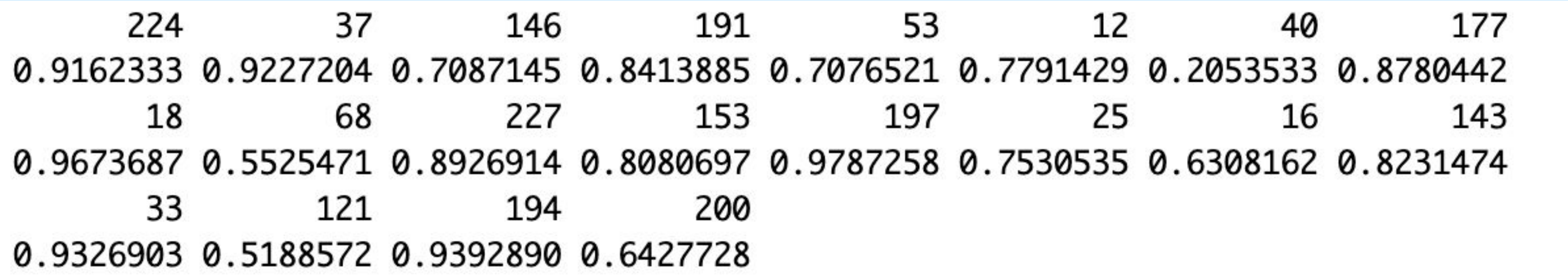
# Forecast – Logistic Regression

## Test Set:



```
        1         3         4         6         8        15        17        19 
0.6958116 0.7798107 0.5277660 0.9146445 0.4482915 0.7059723 0.7155631 0.4177470 
       21        23        24        26        29        31        34        49 
0.9317078 0.8335956 0.7654272 0.8332501 0.7015880 0.9328658 0.8400866 0.8531389 
       57        59        61        62 
0.8776286 0.9233031 0.6394054 0.9207729 
```

## Training Set/ Actual Results:



```
      224        37       146       191        53        12        40       177 
0.9162333 0.9227204 0.7087145 0.8413885 0.7076521 0.7791429 0.2053533 0.8780442 
       18        68       227       153       197        25        16       143 
0.9673687 0.5525471 0.8926914 0.8080697 0.9787258 0.7530535 0.6308162 0.8231474 
       33       121       194       200 
0.9326903 0.5188572 0.9392890 0.6427728 
```

# Forecast – Logistic Regression

## 01 Associated Forecasting Error

| | |
|---|---|
| **Forecast Error** | Forecasting error of 0.3048085. |

## 02 Model Performance Evaluation

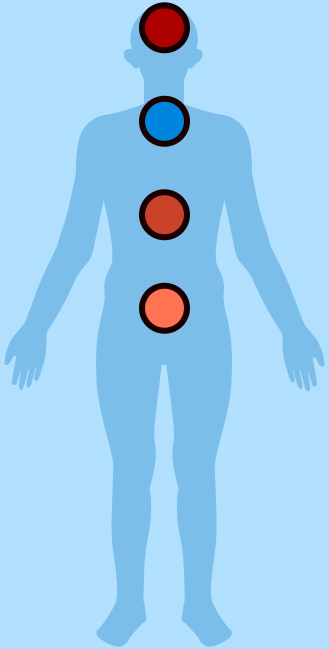
| | |
|---|---|
| **Accuracy: 69.5%** | Moderate to low confidence in performance of the logistic regression model. |

## 03 Considerations

| | |
|---|---|
| **Factors** | Limited sample size and exclusion of certain variables |

# Conclusion

# Our Purpose



01

## Prevention

Awareness around high probability variables

02

## Education

Teach others about strokes and actions around them

03

## Treatment

For those in the ‘red areas’ what can we focus on first?

04

## Forecasting

Meet conditions = get results!

## BMI

Excess fat can highly impact
the chance of having a stroke

## Age

With age, arteries get
narrower and harder

## Glucose

Clogged blood vessels due to
increased fatty deposits