

Statistical Learning Project

Filippo Santin
Gurjeet Singh
Francesca Zen

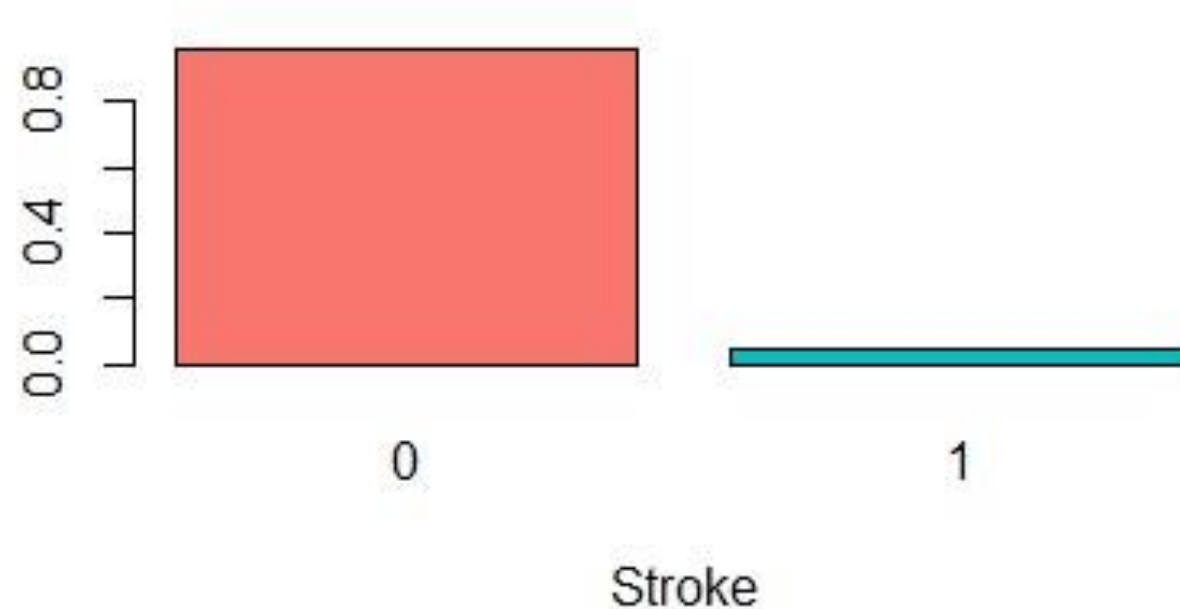
The background of the slide features a close-up of several blue and white capsules scattered on a surface. In the background, a book is visible with the title 'Brain Stroke' printed on its cover. The text is partially obscured by the capsules and a dark overlay.

Data analysis on stroke detection

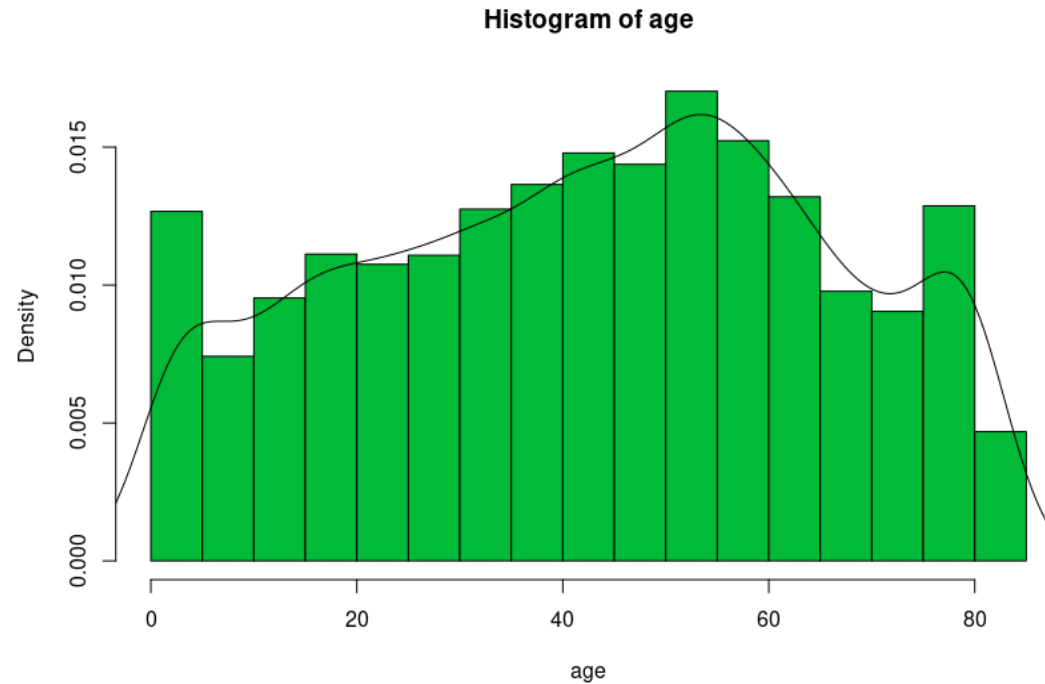
- “Stroke” is the medical term for damage to brain tissue or the death of a portion of it, due to insufficient blood supply to an area of the brain
- It is responsible for approximately 11% of total deaths
- Normal values of glucose: 60-110 mg/dl
 - > 126 diabetes
- Normal BMI range: 18.5 - 24.9
 - > 30.0 obesity

Dataset

- **Data variables:** id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke
- **Missing values**
- **Unbalanced data:**
4,26% of the people get a stroke



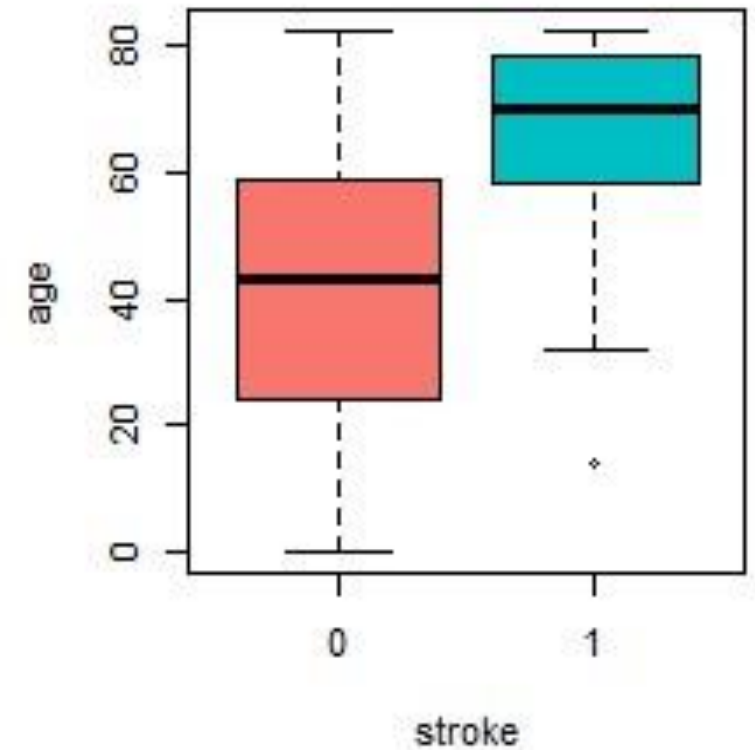
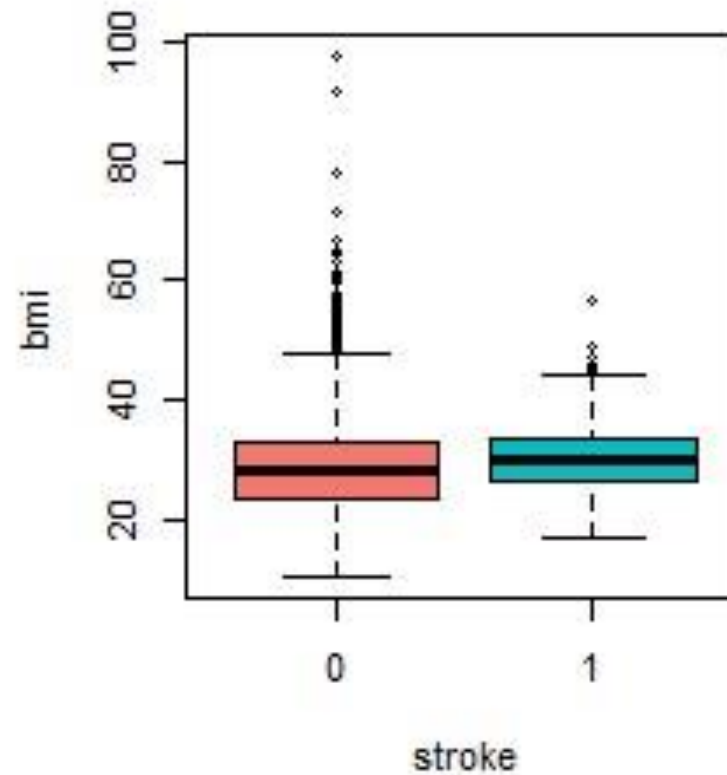
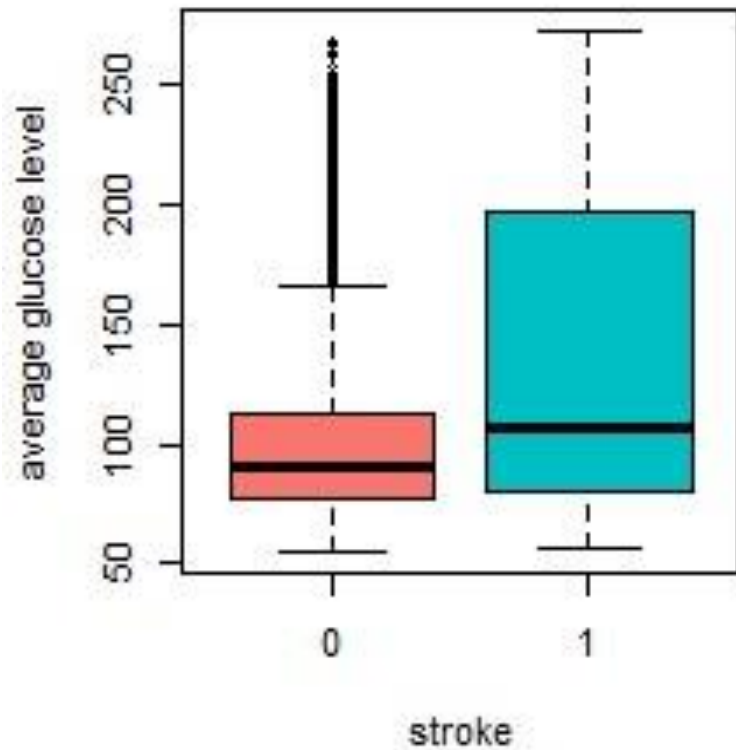
Dataset



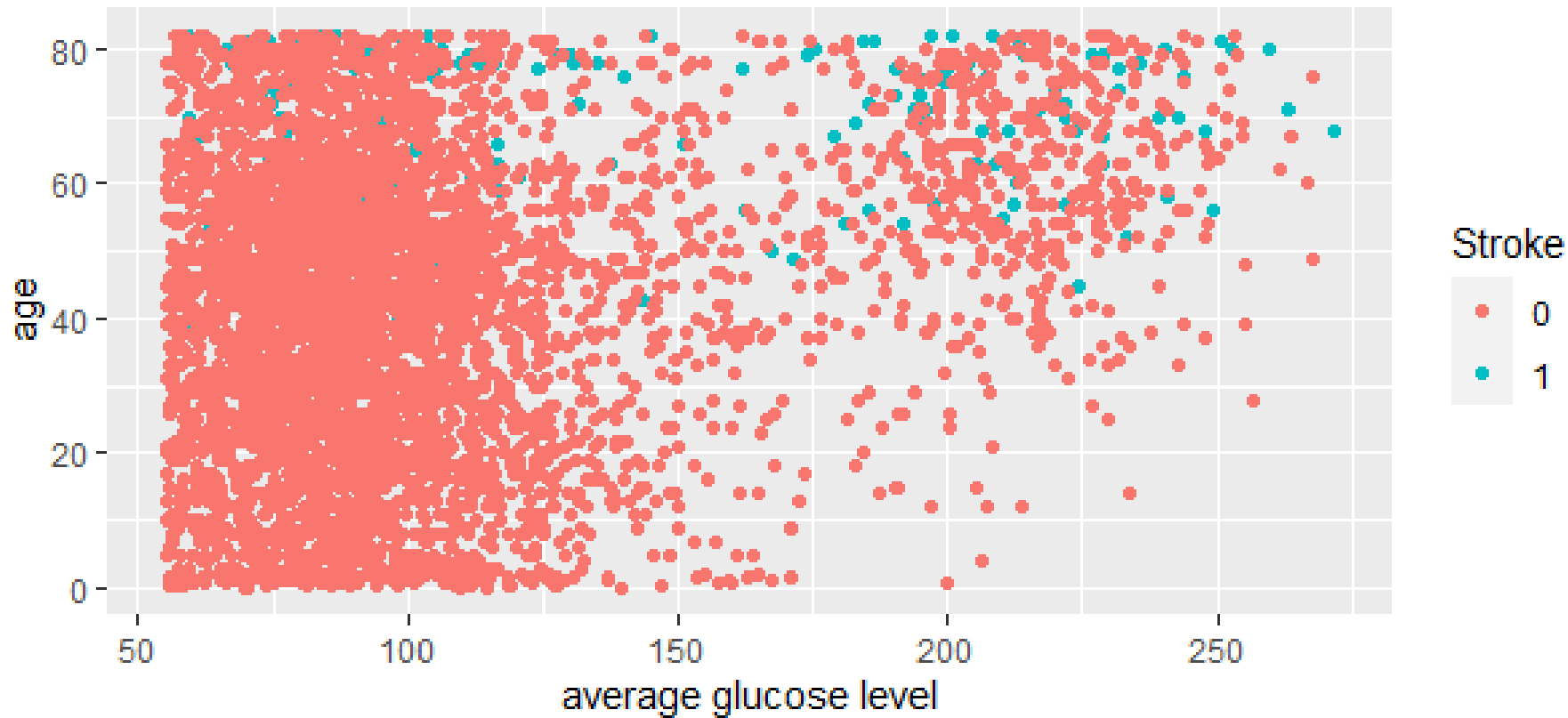
Data cover people of all ages from babies of 8 days to seniors of 82 years old

Explanatory Data Analysis (EDA)

- High glucose level and bmi do not imply directly a stroke
- Rare/interesting cases of stroke
- Strong relation between age and stroke

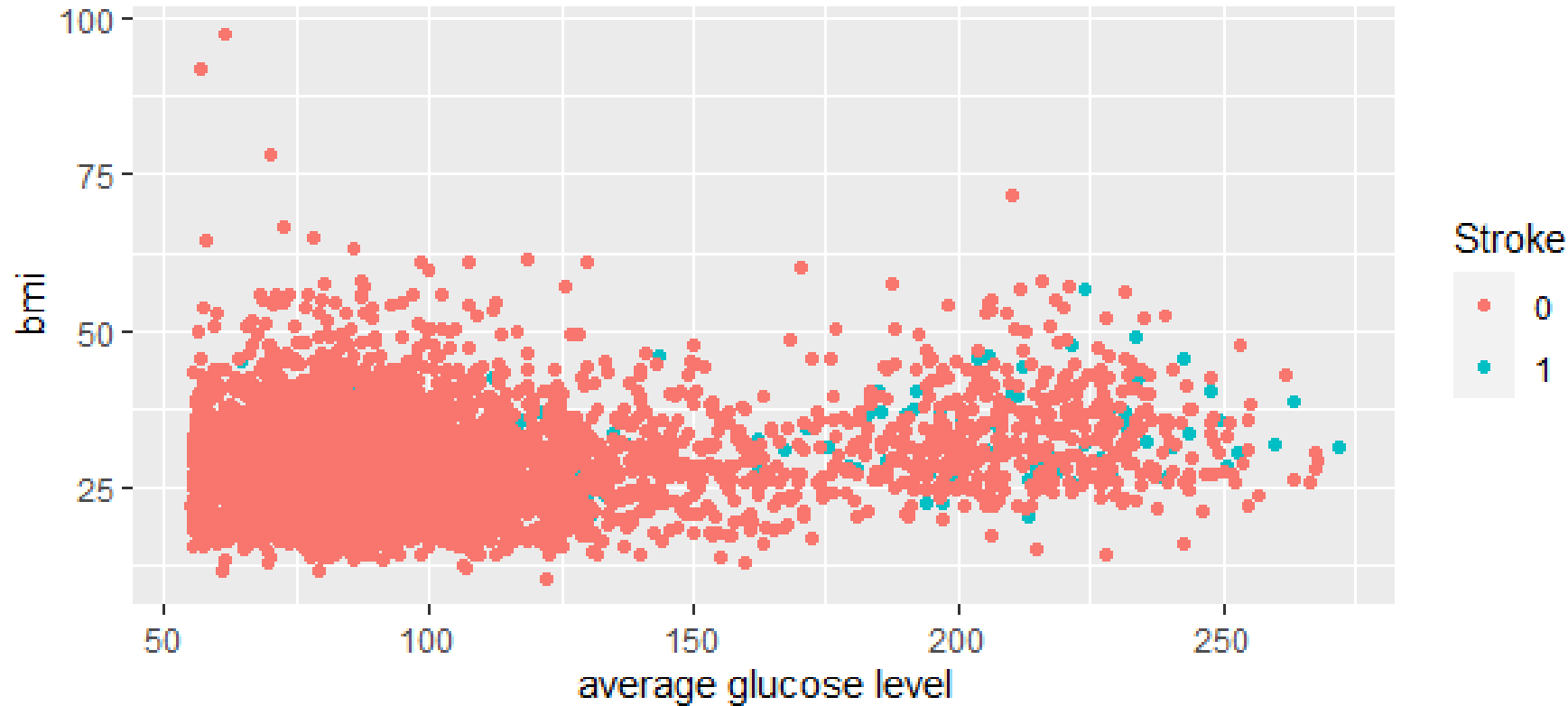


Not an easy problem



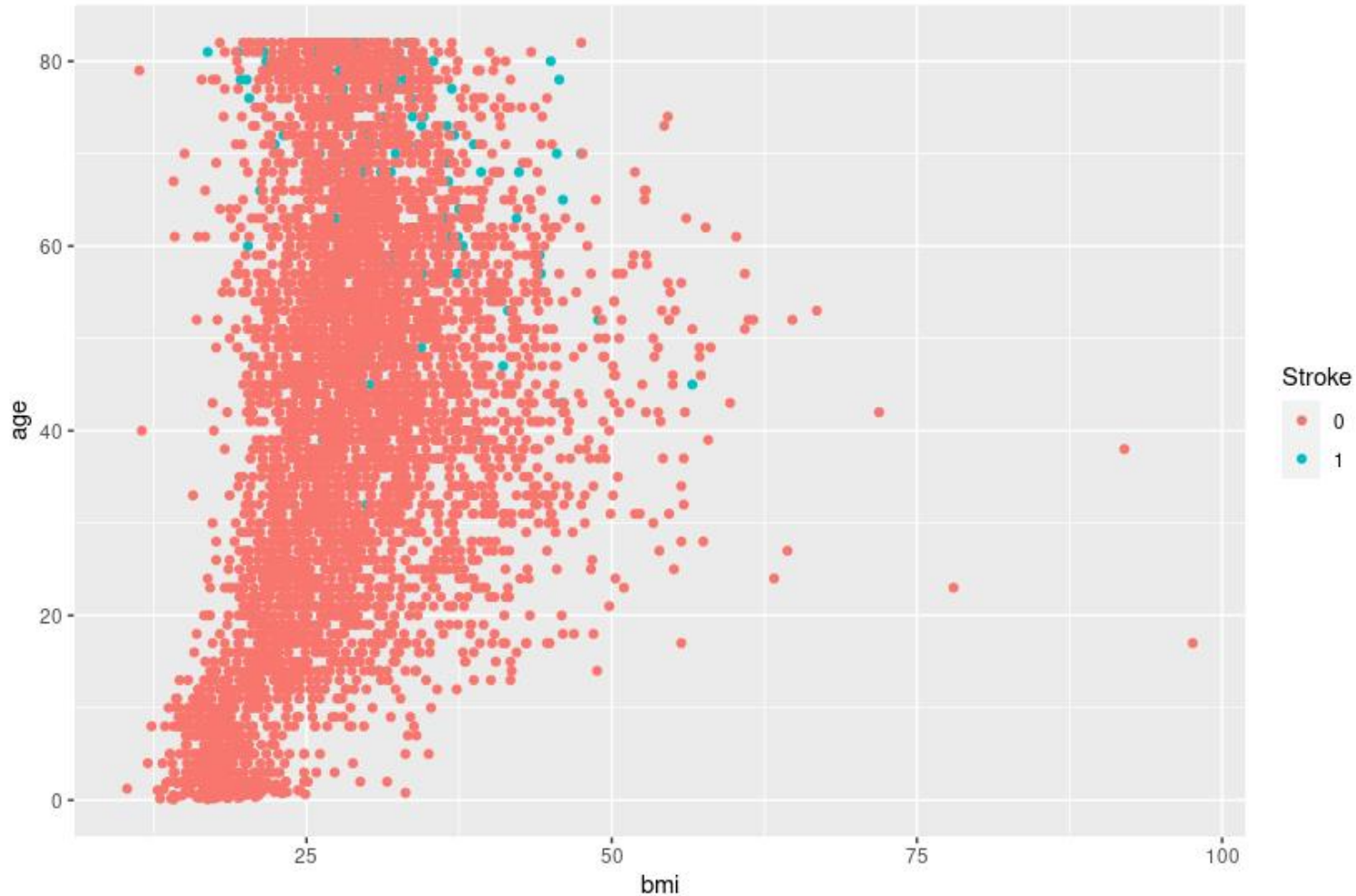
- Non-linear separable
- Not easy to identify a direct relationship with stroke diseases

Not an easy problem



Glucose levels and *Bmi* could not be so strictly related to the disease but maybe correlated to other illnesses linked (or not) to it.

Not an easy problem



- Strong correlation with *age*
- Weak correlation with *Bmi*

Correlation between features

- Presence of collinearity:
 - Age ~ Ever married : 0.68
 - Age ~ Work type : 0.54
 - Age ~ Smoking status : 0.39
 - Ever married ~ Bmi: 0.34
- Collinearity variables:
 - Stroke ~ Age: 0.23
 - Stroke ~ Hypertension: 0.14
 - Stroke ~ Avg. glucose level: 0.14



Relevant Questions:

- Which factors are the most related to the stroke disease?
- How strong are the relations between the features?
- Are the given variables enough to predict a good accuracy of some possible person affected by stroke?
- Is it possible to prevent the stroke?



Tested Models

- **Logistic Regression**
 - Full and Reduced Models
 - Interaction Models
 - Polynomial Models
- **Bayesian Models**
 - LDA Model
 - QDA Model



LOGISTIC REGRESSION

- A type of Generalized linear model (GLM)
- The dependent variable is binary

0 NO STROKE

1 STROKE

Model selection:

- Hypothesis
- p-value
- AIC



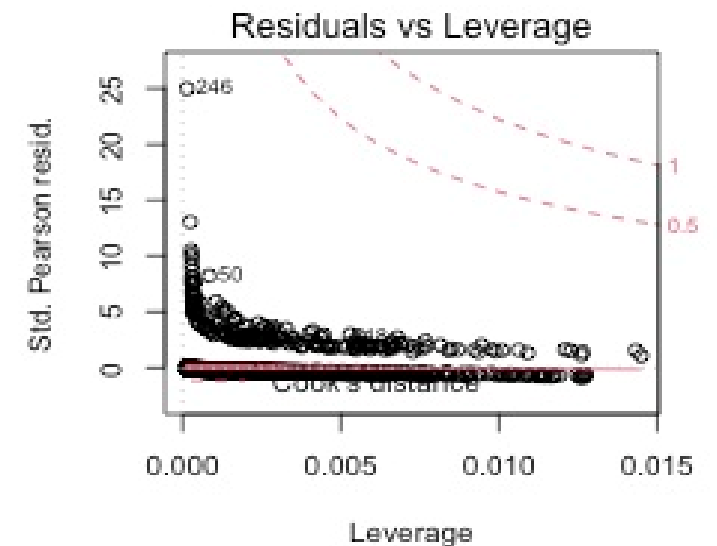
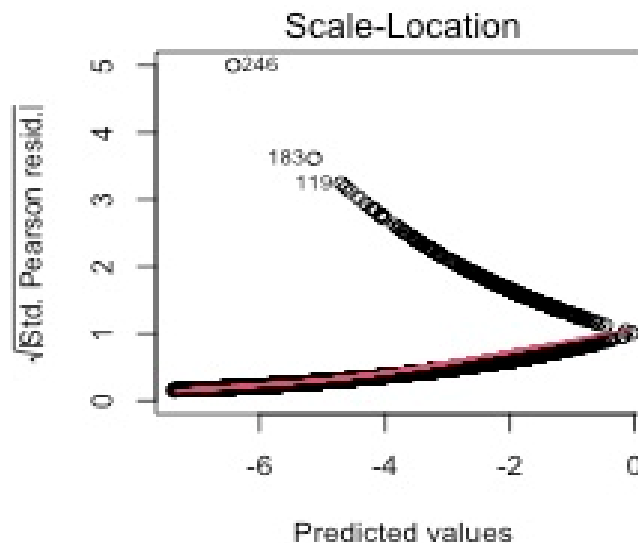
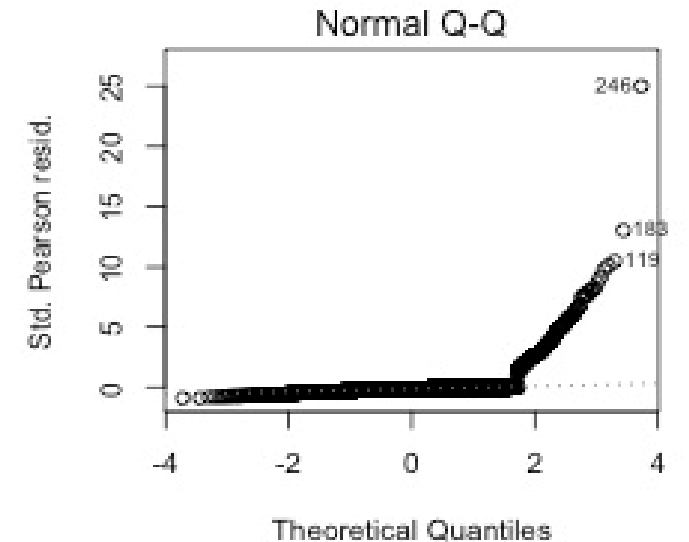
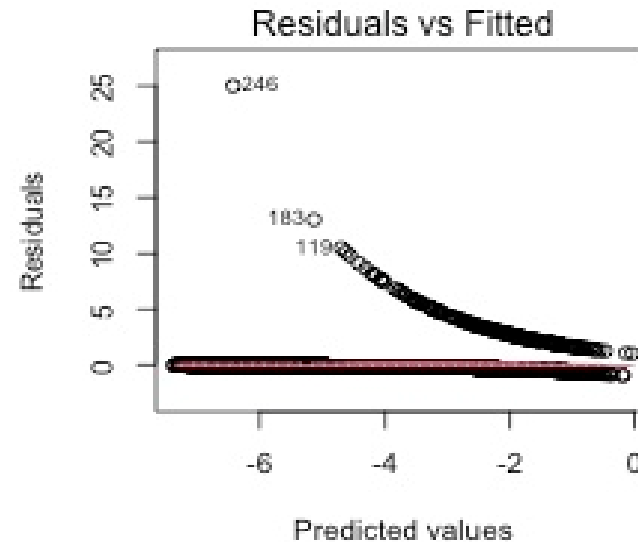
Reduced Model

Feature	Coef.	Level of significance
Age	0.067547	$< 2e-16$
Avg. Glucose level	0.004802	0.000129
Heart disease	0.404298	0.046895
Hypertension	0.539613	0.001820

AIC: 1384.6

Reduced Model Plots

- Non-linearity in dataset
- Residuals do not follow normal distribution
- Heteroscedasticity
- Leverage points
- Outliers



"Outliers"

Able to infer some particular stroke cases, *anomaly detection*.

	gender	age	hypert.	hd	ev_marr	work_type	res_type	glucose	bmi	smoking	stroke
119	Female	38	0	0	No	Self-employed	Urban	82.28	24.0	formerly smoked	1
183	Female	32	0	0	Yes	Private	Rural	76.13	29.9	smokes	1
246	Female	14	0	0	No	children	Rural	57.93	30.9	Unknown	1

Interaction between features

- `age ~ avg_glucose_level, heart_disease, bmi, hypertension`
- `avg_glucose_level ~ heart_disease, bmi, hypertension`
- `heart_disease ~ hypertension`
- `bmi ~ hypertension`



Best Interaction Model

Feature	Coef.	Level of significance
Age	0.070133	$< 2e-16$
Avg. Glucose level	0.004702	0.000176
Heart disease	2.765299	0.047694
Hypertension	0.536550	0.001880
Age:heart disease	-0.032872	0.091604

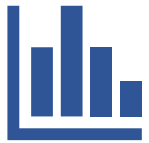
AIC: 1384

Best Model Selection

To choose the best model among the electives ones we used Training and Validation testing method.

Data splits should be done carefully cause of unblanced issue.

- 75% Training
- 25% Validation
- Both splits have 4% of stroke cases



Training set



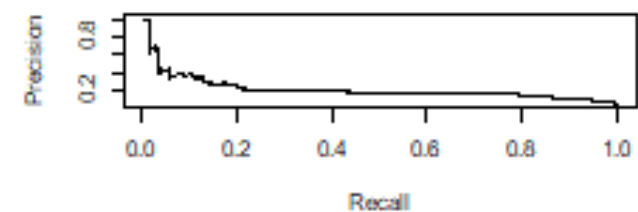
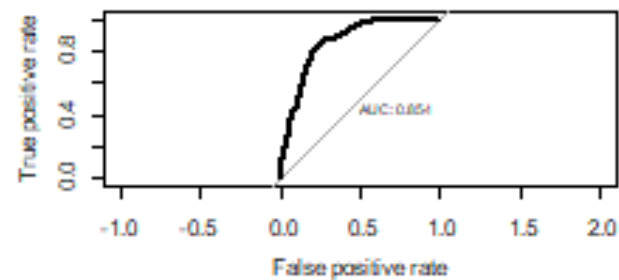
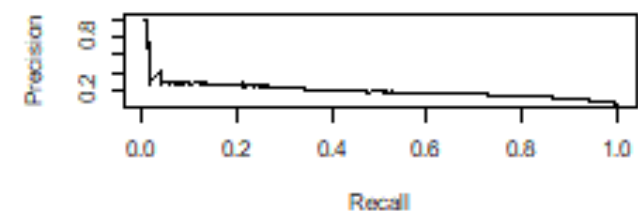
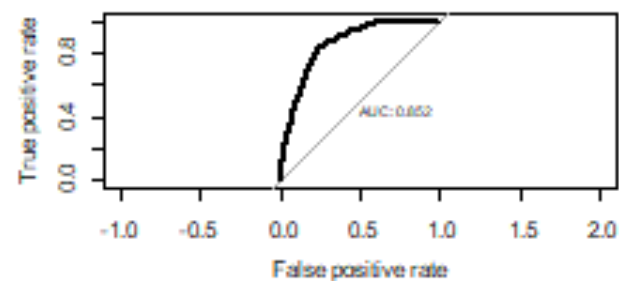
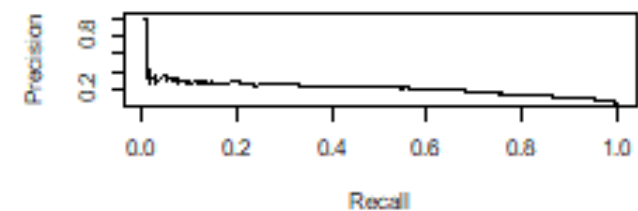
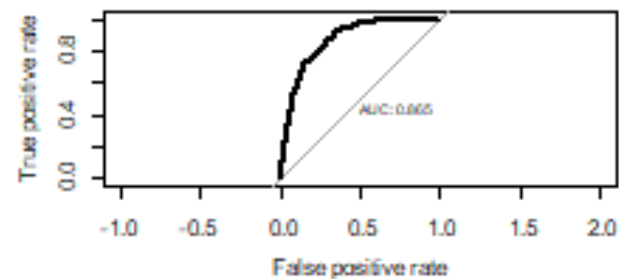
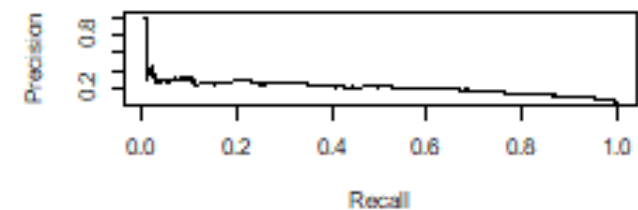
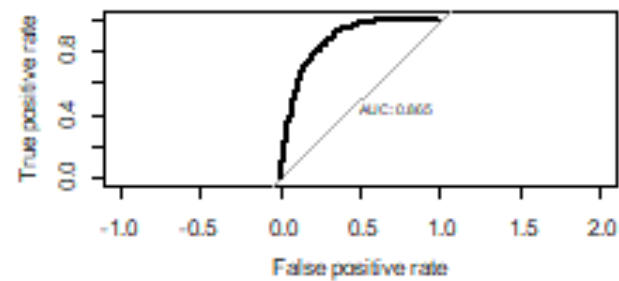
Validation set

Best Model Selection

- False Rates?
 - False negatives in medical cases
 - ROC or Precision-Recall curves?
 - Threshold



ROC VS Prec-Recall Curves



Best Model Selection

REDUCED MODEL			
		Predicted	
		0	1
Ground thruth	0	497	681
	1	1	48

INTERACTION MODEL			
		Predicted	
		0	1
Ground thruth	0	499	679
	1	1	48

LDA MODEL			
		Predicted	
		0	1
Ground thruth	0	490	688
	1	1	48

QDA MODEL			
		Predicted	
		0	1
Ground thruth	0	599	579
	1	4	45

Best Model Selection

- Error Rates -

REDUCED MODEL:

- Positive rates: 0.0658
- Negative rates: 0.002

INTERACTION MODEL:

- Positive rates: 0.066
- Negative rates: 0.002

LDA MODEL:

- Positive rates: 0.0652
- Negative rates: 0.002

QDA MODEL:

- Positive rates: 0.0721
- Negative rates: 0.0066



Conclusions

- Interaction model is the best
- Older people have higher probability to get a stroke
- Not easy to make secure predictions
- Increase the number of data
- Find more features related with stroke
- Find out the appropriate false rate