

Third Group (Class DS2)

Analyzing Stroke Risk Factors Using Logistic Regression

12 DECEMBER, 2024

OUR TEAMS



**Nayla Poetry Salafyna
Mumtaz**

11220940000045



Putri Maesarah

11220940000046



Hamida

11210940000059



Miranita Anisa Rohmah

11220940000055



Awalia Damayanti

11220940000063



DATA UNDERSTANDING

The dataset contains information on demographic, health, and lifestyle factors related to stroke. Here's an overview of the data:

Dependent Variable = Stroke (0 = No , 1 = Yes)

Categoric Variable Predictors

1. Gender
2. Hypertension
3. Heart Disease
4. Ever Married
5. Work Type
6. Residence Type
7. Smoking Status

Continous Variable Predictors

1. Age
2. Average Glucose Level
3. BMI (Body Mass Index)



PURPOSES

The purposes of this analysis are

1. Identify which factors associated with the likelihood of having a stroke.
2. Build a logistic regression model to predict the likelihood of a person having a stroke based on relevant factors.
3. Evaluate the impact of these factors on stroke likelihood to provide guidance on prevention.



Pre Processing

In this process,

- We changed the type of a variable.
- We used the mean to impute missing values on BMI variable
- We remove one row with the value 'Other' on gender variable.

```
tibble [5,110 × 11] (S3: tbl_df/tbl/data.frame)
 $ gender      : chr [1:5110] "Male" "Female" "Male" "Female" ...
 $ age         : chr [1:5110] "67" "61" "80" "49" ...
 $ hypertension: num [1:5110] 0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease: num [1:5110] 1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married: chr [1:5110] "Yes" "Yes" "Yes" "Yes" ...
 $ work_type   : chr [1:5110] "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type: chr [1:5110] "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: chr [1:5110] "228.69" "202.21" "105.92" "171.23" ...
 $ bmi         : chr [1:5110] "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status: chr [1:5110] "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke      : num [1:5110] 1 1 1 1 1 1 1 1 1 1 ...
```

gender	age	hypertension	heart_disease	ever_married
Female:2994	Min. : 0.08	0:4612	0:4834	No :1757
Male :2115	1st Qu.:25.00	1: 498	1: 276	Yes:3353
Other : 1	Median :45.00			
NOISE	Mean :43.23			
	3rd Qu.:61.00			
	Max. :82.00			
work_type	Residence_type	avg_glucose_level	bmi	
children : 687	Rural:2514	Min. : 55.12	Min. :10.30	
Govt_job : 657	Urban:2596	1st Qu.: 77.25	1st Qu.:23.50	
Never_worked : 22		Median : 91.89	Median :28.10	
Private :2925		Mean :106.15	Mean :28.89	
Self-employed: 819		3rd Qu.:114.09	3rd Qu.:33.10	
		Max. :271.74	Max. :97.60	
			NA's :201	
			MISSING VALUE	
smoking_status	stroke			
formerly smoked: 885	Min. :0.00000			
never smoked :1892	1st Qu.:0.00000			
smokes : 789	Median :0.00000			
Unknown :1544	Mean :0.04873			
	3rd Qu.:0.00000			
	Max. :1.00000			



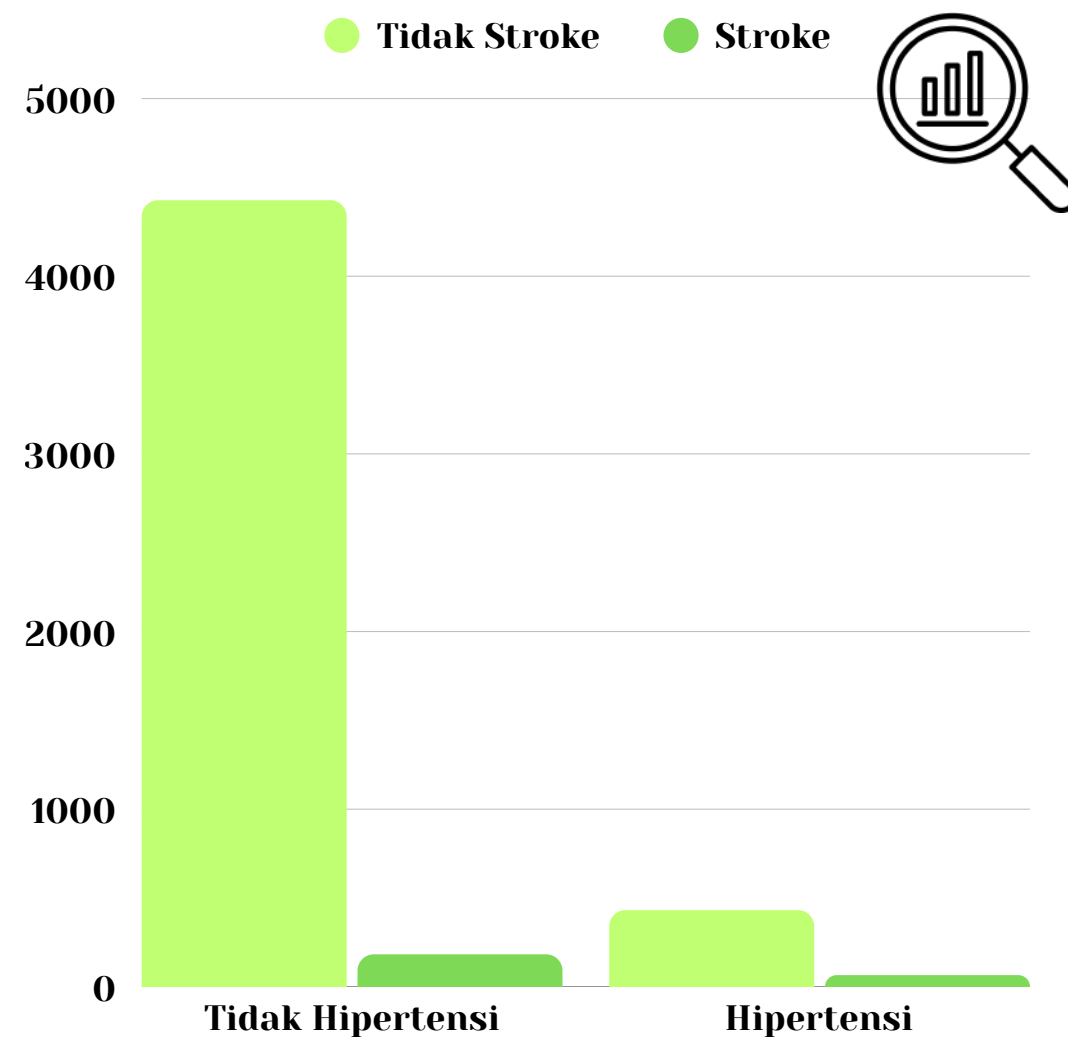
MODEL BUILDING



EXPLORATORY DATA ANALYSIS (EDA)

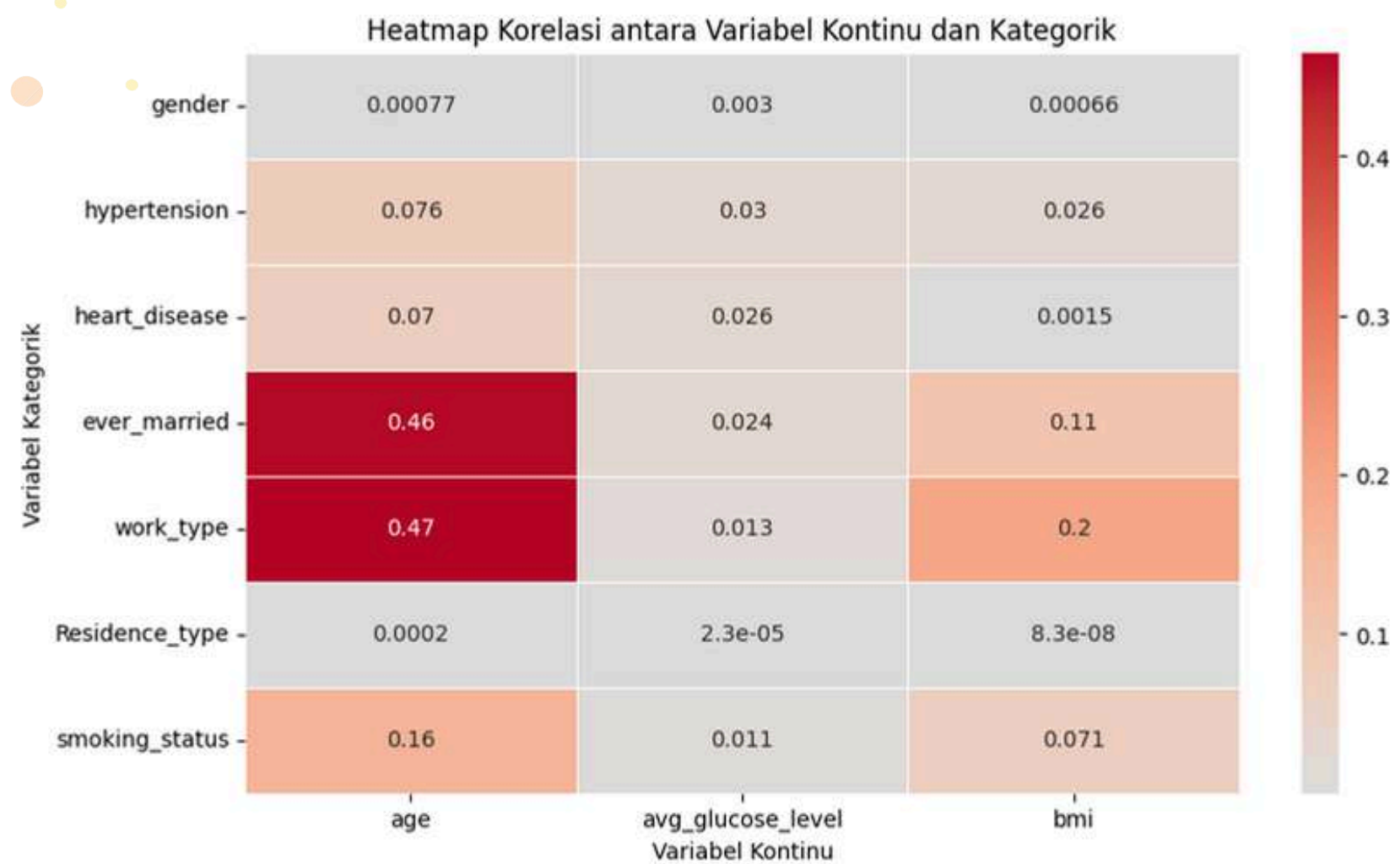


Proportion of Stroke Cases by Hypertension Status



the proportion of stroke cases is relatively higher among individuals with hypertension compared to those without.

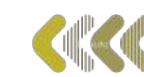
Correlation Between Independent variable



Higher correlation represented with darker colors, indicating stronger relationships.



MODEL ASSUMPTION



MODEL ASSUMPTION

```
summary(glm.logit)
```

```
Call:
glm(formula = stroke ~ avg_glucose_level + Residence_type + work_type +
    ever_married + bmi + heart_disease * hypertension + age *
    heart_disease + age * hypertension + smoking_status * gender,
    family = binomial(link = "logit"), data = data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-6.815132	0.793585	-8.588	< 2e-16
avg_glucose_level	0.003885	0.001202	3.233	0.00122
Residence_typeUrban	0.070406	0.138235	0.509	0.61053
work_typeGovt_job	-1.166791	0.852573	-1.369	0.17114
work_typeNever_worked	-10.349112	309.501676	-0.033	0.97333
work_typePrivate	-1.031307	0.836513	-1.233	0.21763
work_typeSelf-employed	-1.410105	0.855842	-1.648	0.09943
ever_marriedYes	-0.203064	0.225201	-0.902	0.36721
bmi	0.001710	0.011413	0.150	0.88088
heart_disease1	2.682193	1.309459	2.048	0.04053
hypertension1	1.322502	1.011679	1.307	0.19113
age	0.079217	0.006563	12.071	< 2e-16
smoking_statusnever smoked	-0.150851	0.232383	-0.649	0.51624
smoking_statussmokes	-0.063148	0.310224	-0.204	0.83870
smoking_statusUnknown	-0.167639	0.286369	-0.585	0.55828
genderMale	-0.039452	0.264232	-0.149	0.88131
heart_disease1:hypertension1	-0.325321	0.415329	-0.783	0.43346
heart_disease1:age	-0.032203	0.018322	-1.758	0.07881
hypertension1:age	-0.012384	0.014592	-0.849	0.39604
smoking_statusnever smoked:genderMale	-0.190103	0.361767	-0.525	0.59925
smoking_statussmokes:genderMale	0.333174	0.426240	0.782	0.43442
smoking_statusUnknown:genderMale	0.224330	0.410354	0.547	0.58460

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1990.3 on 5108 degrees of freedom
Residual deviance: 1574.4 on 5087 degrees of freedom
AIC: 1618.4

Number of Fisher Scoring iterations: 14

is there multicollinearity?

NO

The picture shows that several variables are significant (with p-values less than 5%), and the difference in deviance indicates that the model has at least one significant variable.

```
> p_value <- 1 - pchisq( 415.9 , 21)
> p_value
[1] 0
```



MULTICOLLINEARITY

See that the value of VIF (third column) for all variabel is less than 10. That's indicates there is no multicollinearity on this model assumption.

	GVIF	Df	$GVIF^{(1/(2*Df))}$
avg_glucose_level	1.121441	1	1.058981
Residence_type	1.009382	1	1.004680
work_type	1.498642	4	1.051870
ever_married	1.106981	1	1.052132
bmi	1.128648	1	1.062379
heart_disease	52.827435	1	7.268248
hypertension	40.862915	1	6.392411
age	1.769327	1	1.330161
smoking_status	7.075194	3	1.385553
gender	3.637534	1	1.907232
heart_disease:hypertension	1.650024	1	1.284533
heart_disease:age	53.649611	1	7.324589
hypertension:age	42.135591	1	6.491193
smoking_status:gender	14.389092	3	1.559572



MODEL SELECTION

Backward, Forward, and Stepwise Elimination

```
Call: glm(formula = stroke ~ age + avg_glucose_level + hypertension +  
heart_disease + age:heart_disease, family = binomial(link = "logit"),  
data = data)
```

Coefficients:

(Intercept)	age	avg_glucose_level	hypertension1
-7.646978	0.071512	0.004033	0.378696
heart_disease1	age:heart_disease1		
2.745935	-0.033691		

```
Degrees of Freedom: 5108 Total (i.e. Null); 5103 Residual  
Null Deviance: 1990  
Residual Deviance: 1588 AIC: 1600
```

Backward, forward, and stepwise selection methods all lead to the same result: age, hypertension, heart disease, average glucose levels, and interaction between age & heart disease are associated with an increased likelihood of having a stroke.

$$\text{log odds of having stroke} = \beta_0 + \beta_1(\text{avg_glucose_level}) + \beta_2(\text{hypertension}) + \beta_3(\text{age}) + \beta_4(\text{heart_disease}) + \beta_5(\text{age} * \text{heart_disease})$$



Evaluation and Model Checking

Hosmer and Lemeshow goodness of fit (GOF) test

H0 : The model fits the data well

H1 : The model doesn't fits the data well

ResourceSelection 0.3-6

2023-06-27

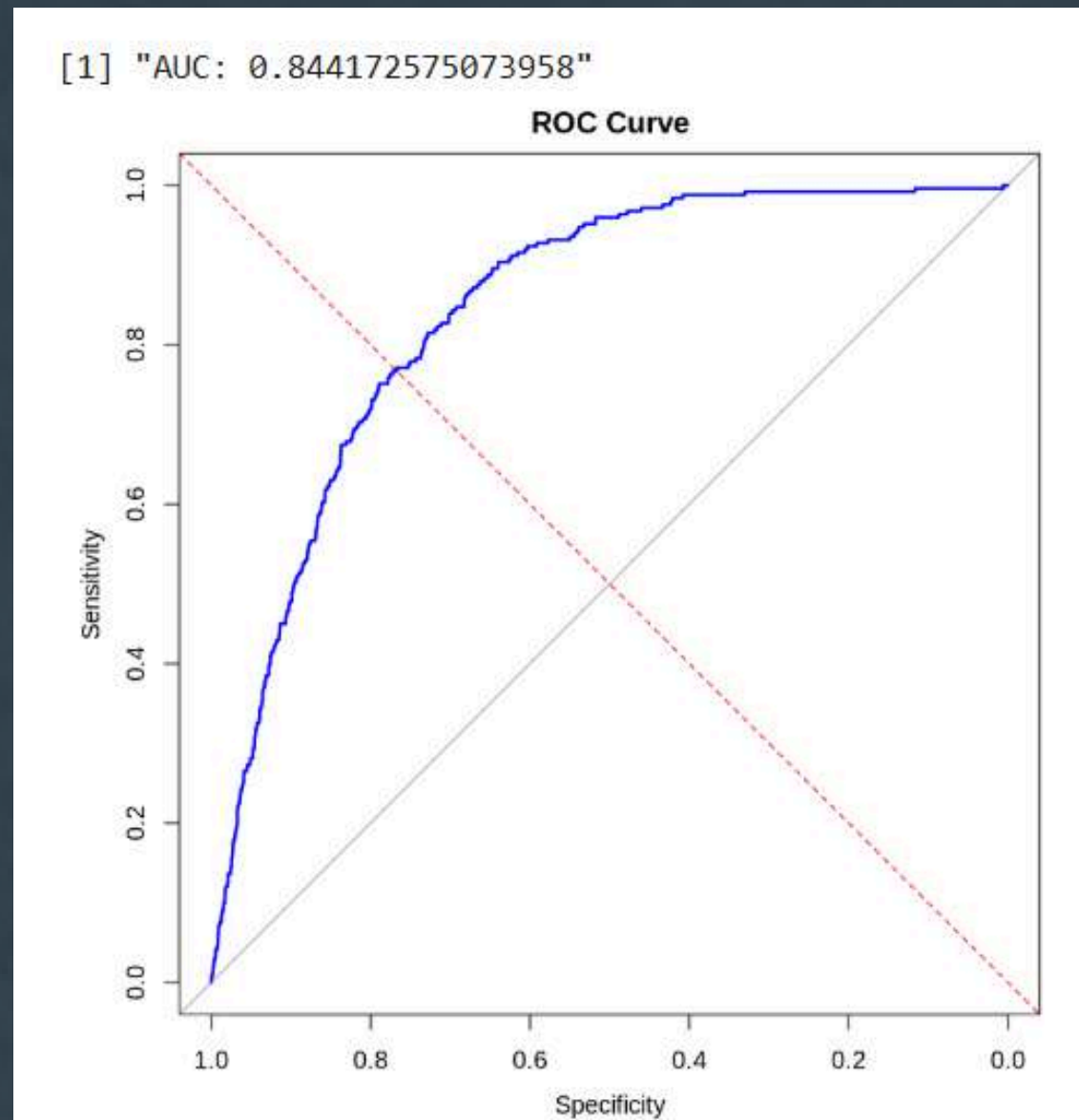
Hosmer and Lemeshow goodness of fit (GOF) test

```
data: data$stroke, predicted_probabilities  
X-squared = 5.3998, df = 8, p-value = 0.7141
```



Since the p-value is 0.7141, which is greater than 5%, so we accept the null hypothesis and conclude that the logistic regression model fits the data well.

ROC CURVE



The AUC = 0.844 indicates the model has excellent classification ability, so the model can distinguish between people who had a stroke and those who did not have a stroke.



Wald Test for β

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

```
Call:
glm(formula = stroke ~ avg_glucose_level + hypertension + age *
     heart_disease, family = binomial(link = "logit"), data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.646978   0.376319 -20.320  < 2e-16 ***
avg_glucose_level  0.004033   0.001161   3.474 0.000513 ***
hypertension1    0.378696   0.162217   2.335 0.019569 *
age             0.071512   0.005430  13.170  < 2e-16 ***
heart_disease1   2.745935   1.293066   2.124 0.033705 *
age:heart_disease1 -0.033691   0.018075  -1.864 0.062329 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1990.3  on 5108  degrees of freedom
Residual deviance: 1588.3  on 5103  degrees of freedom
AIC: 1600.3
```



For β_1 , β_2 , β_3 , and β_4 the p-value is less than 5%, so we reject the null hypothesis and conclude that β_1 , β_2 , β_3 , and β_4 is not equal to 0. And for β_5 the p-value is greater than 5%, so we accept the null hypothesis and conclude that β_5 equal to 0.

CLASSIFICATION TABLE

	Actual	
Predicted	0	1
0	1985	4
1	2875	245

Accuracy: 0.4364846
Error: 0.5635154
Sensitivity: 0.9839357
Specificity: 0.4084362

Although the model's accuracy is low (43.64%), its high sensitivity (98.39%) shows it is very effective at detecting people who have had a stroke. This is crucial in stroke detection, where identifying positive cases is more important than avoiding false positives. However, with low specificity (40.84%), the model struggles to correctly classify those without stroke.

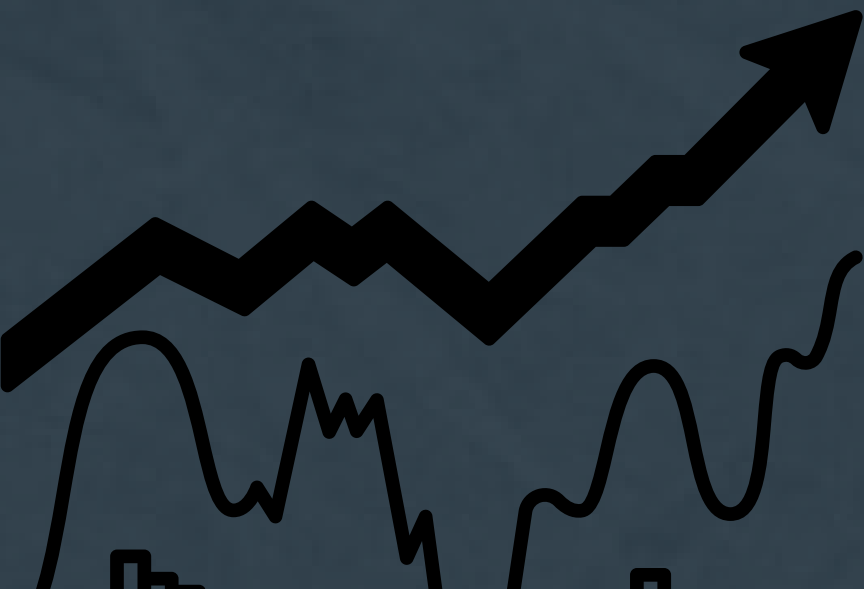
Interpretation and Conclusion

Model Estimate

So, based on Wald Test for β we have estimated model, such as :

- log odds of having stroke = $-7.65 + 0.004(\text{avg_glucose_level}) + 0.379(\text{hypertension}) + 0.071(\text{age}) + 2.75(\text{heart_disease}) + 0(\text{age} * \text{heart_disease})$

log odds of having stroke = $-7.65 + 0.004(\text{avg_glucose_level}) + 0.379(\text{hypertension}) + 0.071(\text{age}) + 2.75(\text{heart_disease})$



INTERPRETATION

log odds of having stroke = $-7.65 + 0.004(\text{avg_glucose_level}) + 0.379(\text{hypertension}) + 0.071(\text{age}) + 2.75(\text{heart_disease})$

1. **avg_glucose_level**: For 1 unit increase in glucose level the odds of stroke increases by $e^{0.004}$.
2. **hypertension**: The presence of hypertension increases the odds of stroke by $e^{0.379}$.
3. **age**: For 1-year increase in age the odds of stroke increases by $e^{0.071}$.
4. **heart_disease** : The presence of heart disease increases the odds of stroke by $e^{2.75}$.



CONCLUSION

Which factors associated with the likelihood of having a stroke ?

- Age
- Hypertension
- Heart Disease
- Average Glucose Level

How to predict the likelihood of having a stroke ?

Just by inputting variable values into the estimate model. However, the model needs to be updated regularly with the latest data to ensure optimal performance

Recommendations for Stroke Prevention

- Regularly check-up especially for older individuals to manage stroke risk factors
- Manage healthy blood glucose, blood pressure with medication and a healthy lifestyle.
- Do proper treatment for heart disease and cholesterol management to prevent blood clot formation.





WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION

