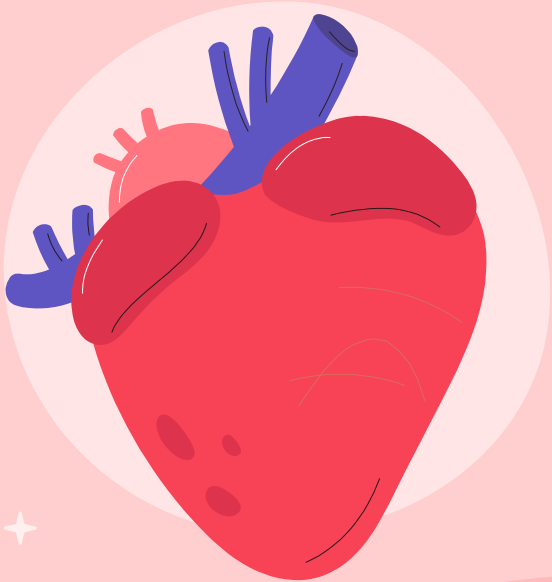
STROKE RISK PREDICTION



Team Member -
Harsh Sanjay Shah
Vrushabh Kumar Shrimali
Jenith Mayurbhai Suvagia

INTRODUCTION

- Stroke prediction is a crucial field of medical research.
- Identifying stroke risk factors is essential for early prevention.
- Timely interventions can save lives and reduce the burden of stroke-related disabilities.

Importance:
- Stroke is a leading cause of death and disability worldwide.
- Early prevention and intervention are key to reducing its impact.
- In 2023, stroke caused 610,000 deaths and 3.3 million disabilities in US. Strokes is the 5th leading cause of death an leading cause of adult disabilities.

Objective

Stroke prediction analysis involves using data to predict the likelihood of an individual having a stroke. It includes data collection, preprocessing, and the development of models to identify factors that influence stroke risk, with the goal of providing early intervention and personalized medical advice based on an individual's characteristics and medical history. This process typically encompasses data cleaning, exploratory data analysis, feature selection, model building, training, evaluation, and deployment to assist healthcare professionals in assessing stroke risk and taking preventive measures.

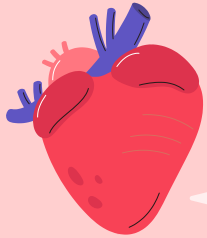
Team Member

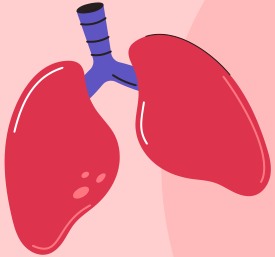
Harsh Sanjay Shah

Vrushabh Shrimali

Jenith M Suvagia

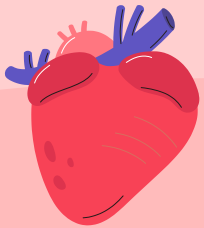
TABLE OF CONTENTS





1

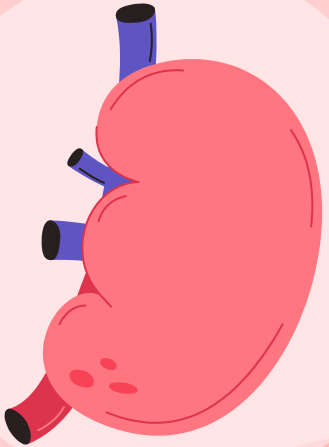
Project Aim

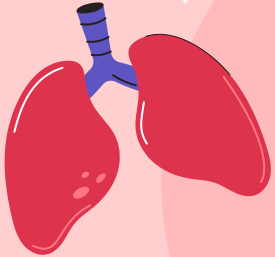


- The primary aim of this project is to predict an individual's risk of experiencing a stroke.

THE IMPORTANCE OF EARLY DETECTION

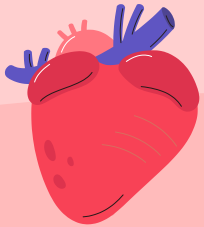
- Early detection can significantly improve stroke outcomes.
- Allows for timely intervention and preventive measures.
- Enhances the quality of life for individuals at risk.





2

Data Exploration



Information of the Dataset

Overview Of the Data

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Information about Data Types and Missing Values

```
Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

```
missing_percentage = (Data.isnull().mean() * 100).round(2)
```

```
missing_percentage
```

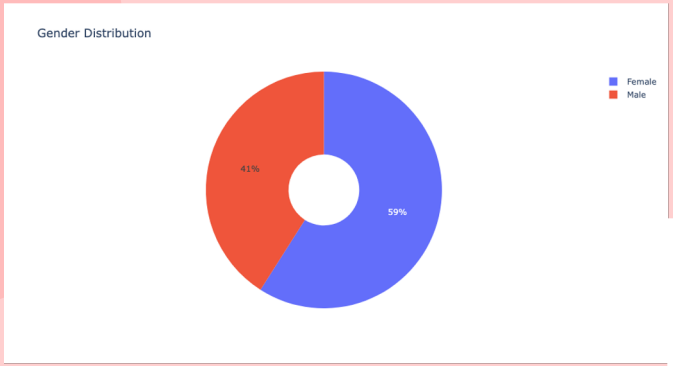
```
id                   0.00
gender               0.00
age                  0.00
hypertension         0.00
heart_disease        0.00
ever_married         0.00
work_type            0.00
Residence_type       0.00
avg_glucose_level    0.00
bmi                  3.93
smoking_status       0.00
stroke               0.00
dtype: float64
```

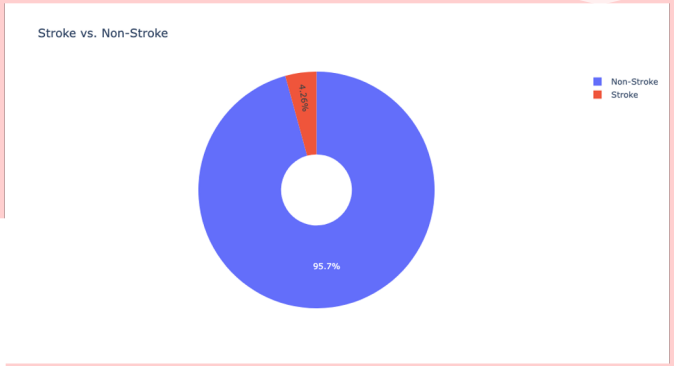
summary statistics

```
Data.describe()
```

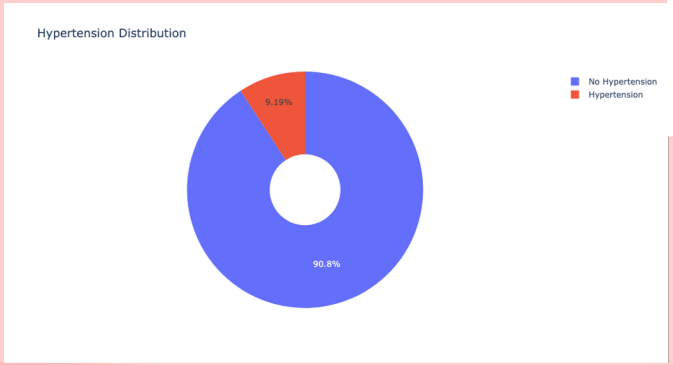
	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	4908.000000	4908.000000	4908.000000	4908.000000	4908.00000	4908.000000
mean	42.857579	0.091891	0.049511	105.297402	28.89456	0.042584
std	22.577004	0.288901	0.216954	44.425550	7.85432	0.201937
min	0.000000	0.000000	0.000000	55.120000	10.30000	0.000000
25%	25.000000	0.000000	0.000000	77.067500	23.50000	0.000000
50%	44.000000	0.000000	0.000000	91.680000	28.10000	0.000000
75%	60.000000	0.000000	0.000000	113.495000	33.10000	0.000000
max	82.000000	1.000000	1.000000	271.740000	97.60000	1.000000

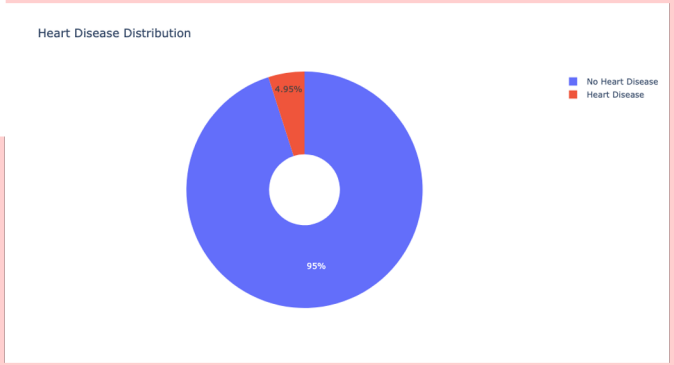
Pie Chart



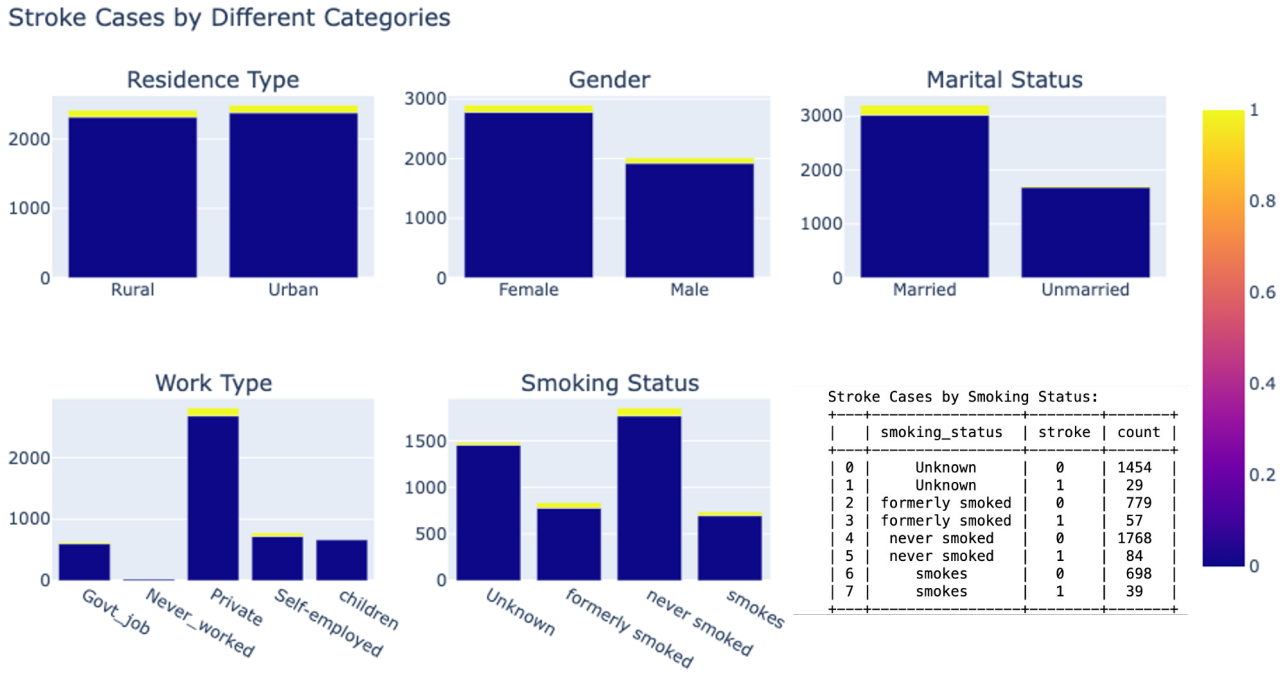



```
Female    2897
Male      2011
Name: gender, dtype: int64
0    4699
1     209
Name: stroke, dtype: int64
0    4457
1     451
Name: hypertension, dtype: int64
0    4665
1     243
Name: heart_disease, dtype: int64
```





Bar Plot



Stroke Cases by Smoking Status:

	smoking_status	stroke	count
0	Unknown	0	1454
1	Unknown	1	29
2	formerly smoked	0	779
3	formerly smoked	1	57
4	never smoked	0	1768
5	never smoked	1	84
6	smokes	0	698
7	smokes	1	39

Stroke Cases by Residence Type:

	Residence_type	stroke	count
0	Rural	0	2318
1	Rural	1	100
2	Urban	0	2381
3	Urban	1	109

Stroke Cases by Gender:

	gender	stroke	count
0	Female	0	2777
1	Female	1	120
2	Male	0	1922
3	Male	1	89

Stroke Cases by Marital Status:

	ever_married	stroke	count
0	Married	0	3018
1	Married	1	186
2	Unmarried	0	1681
3	Unmarried	1	23

Stroke Cases by Work Type:

	work_type	stroke	count
0	Govt_job	0	602
1	Govt_job	1	28
2	Never_worked	0	22
3	Private	0	2683
4	Private	1	127
5	Self-employed	0	722
6	Self-employed	1	53
7	children	0	670
8	children	1	1

Wilcoxon Test

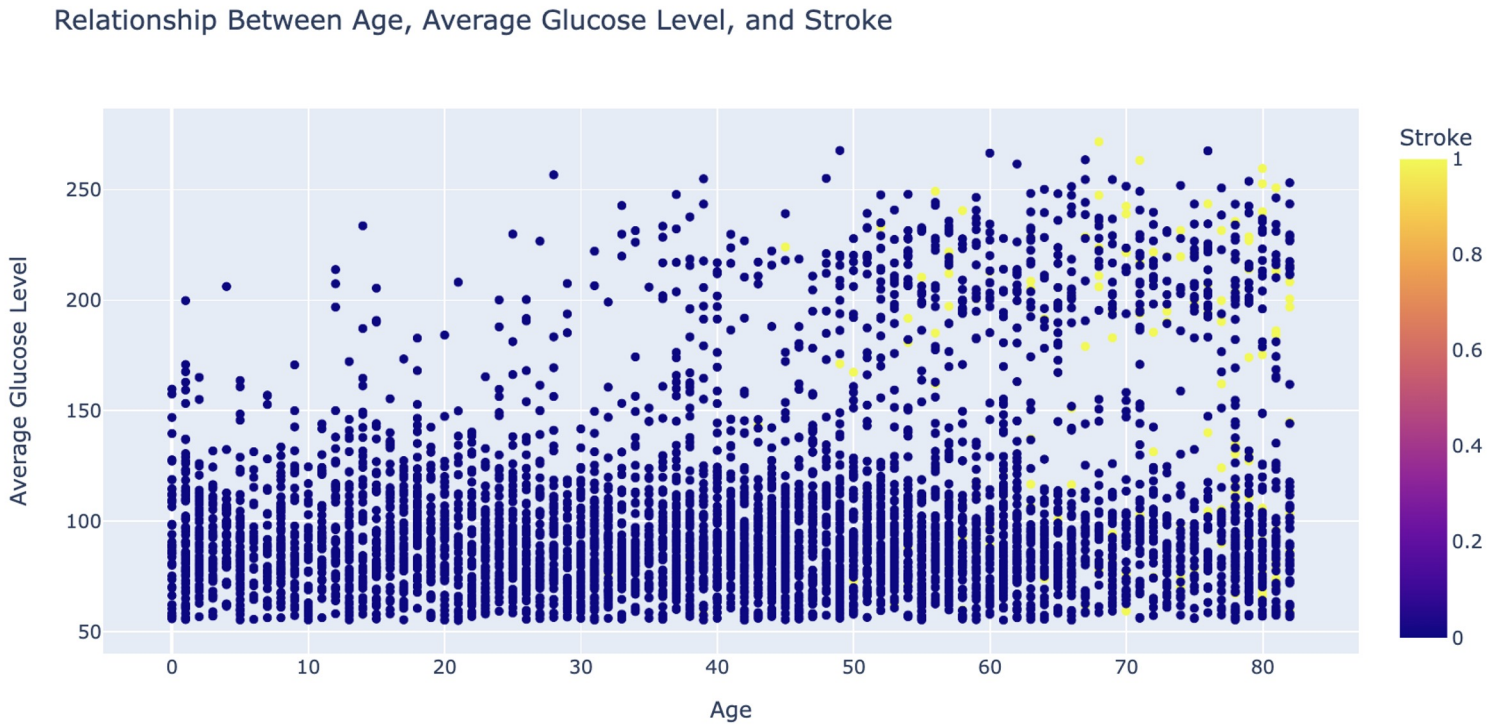
In [72]:
```
from scipy.stats import wilcoxon

# Perform the Wilcoxon test
wilcoxon_test_statistic, wilcoxon_test_p_value = wilcoxon(stroke)

# Print the results
print('Wilcoxon test statistic:', wilcoxon_test_statistic)
print('Wilcoxon test p-value:', wilcoxon_test_p_value)
```

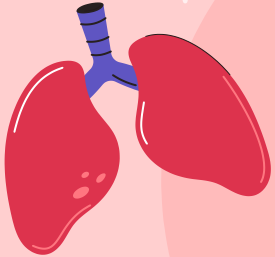
```
Wilcoxon test statistic: 0.0
Wilcoxon test p-value: 2.2700733178214844e-47
```

In [53]:
```
# Define the variables of interest
variables = ['age', 'hypertension', 'heart_disease', 'avg_glucose_level', 'bmi']

# Group the data by "stroke"
grouped_data = Data.groupby('stroke')

for variable in variables:
    result = stats.mannwhitneyu(grouped_data.get_group(1)[variable], grouped_data.get_group(0)[variable])
    print(f"Wilcoxon Test for '{variable}' p-value:", result.pvalue)
```

```
Wilcoxon Test for 'age' p-value: 6.377307674435073e-61
Wilcoxon Test for 'hypertension' p-value: 1.8221841128541993e-23
Wilcoxon Test for 'heart_disease' p-value: 4.377497602090701e-22
Wilcoxon Test for 'avg_glucose_level' p-value: 8.038249301233431e-10
Wilcoxon Test for 'bmi' p-value: 0.00010418772123808268
```

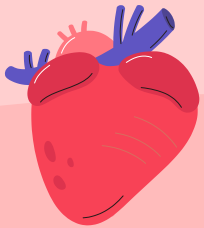
Visualization





3

STATISTICAL ANALYSIS



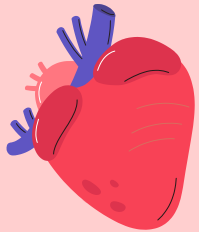
T Tests, Chi-Squared, Wilcoxon test, ANOVA

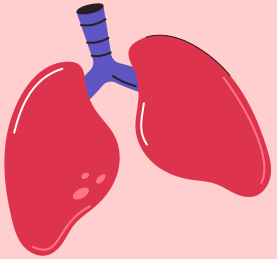
Chi-Squared Test

Column name	Chi-Square Test p-value:
Work Type	1.7076444710201084e-08
Ever married	3.2092342416320797e-13
Heart Disease	2.0990170036138994e-21
Hypertension	6.143875464115642e-2
Gender	0.6805108914997836
Residence Type	0.7272126134406378

ANOVA Test

Column name	p-value:	F-statistic:
Age	4.386488271550836e-61	279.59
Avg Glucose Level	1.3476353968158962e-22	96.63
BMI	0.0030083559555237 09	8.81

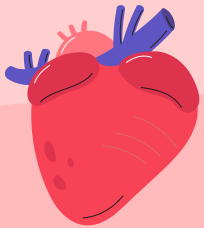




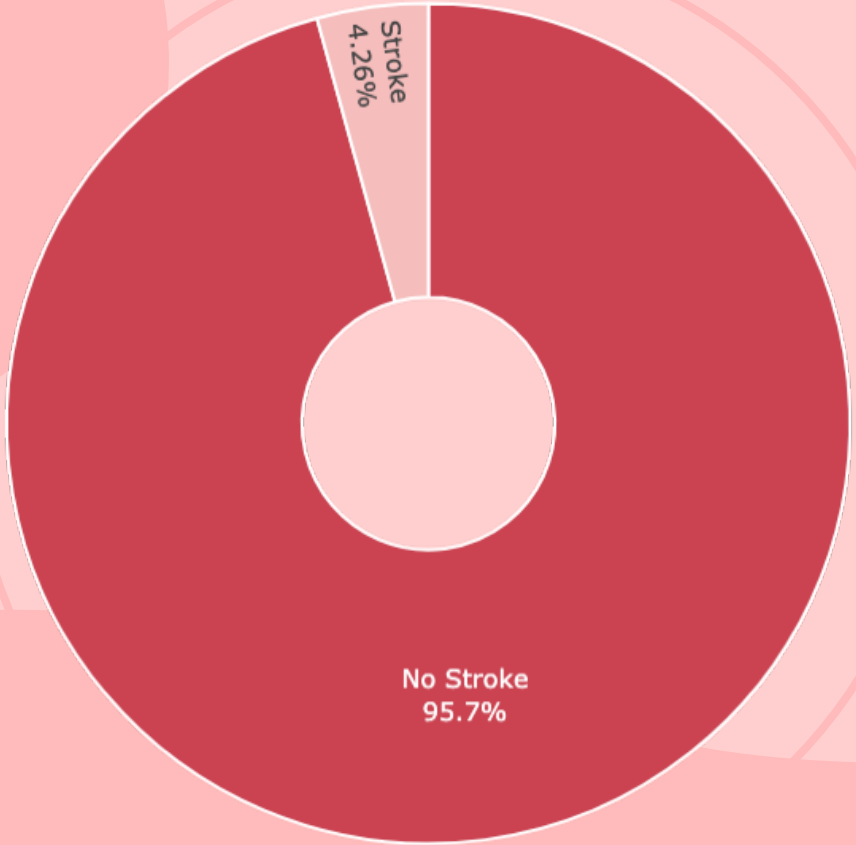
4

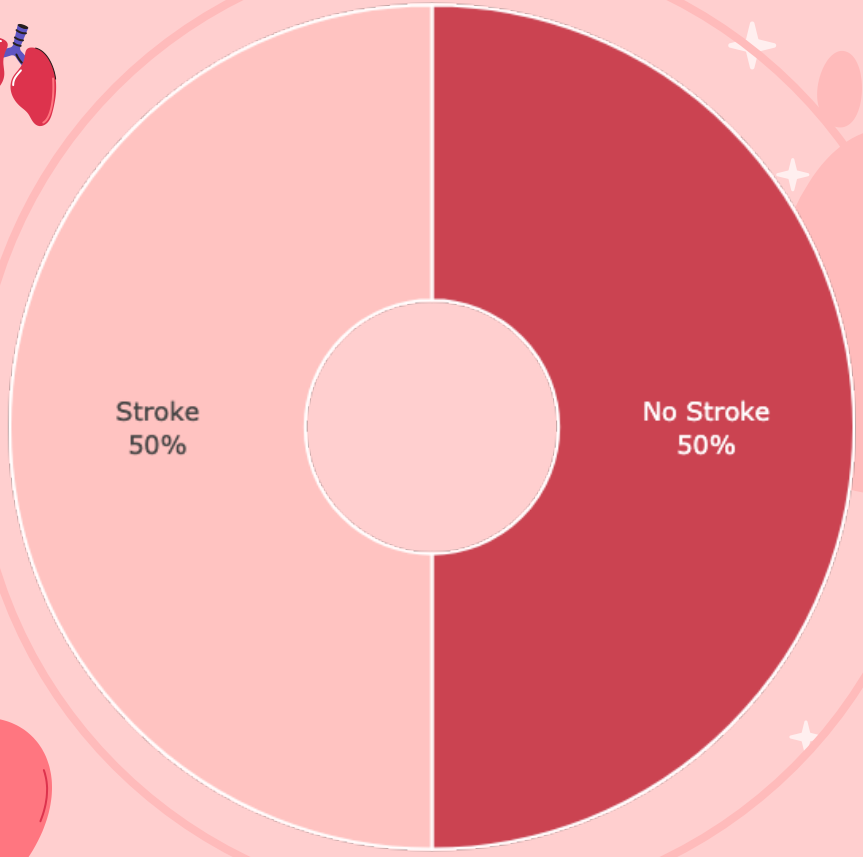
ML MODEL

Random Forest



Stroke Proportion Before SMOTE





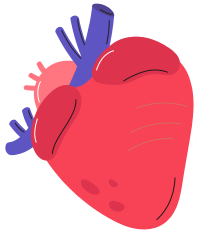
Stroke Proportion After SMOTE

Cross Validation

Random Forest Performance Metrics

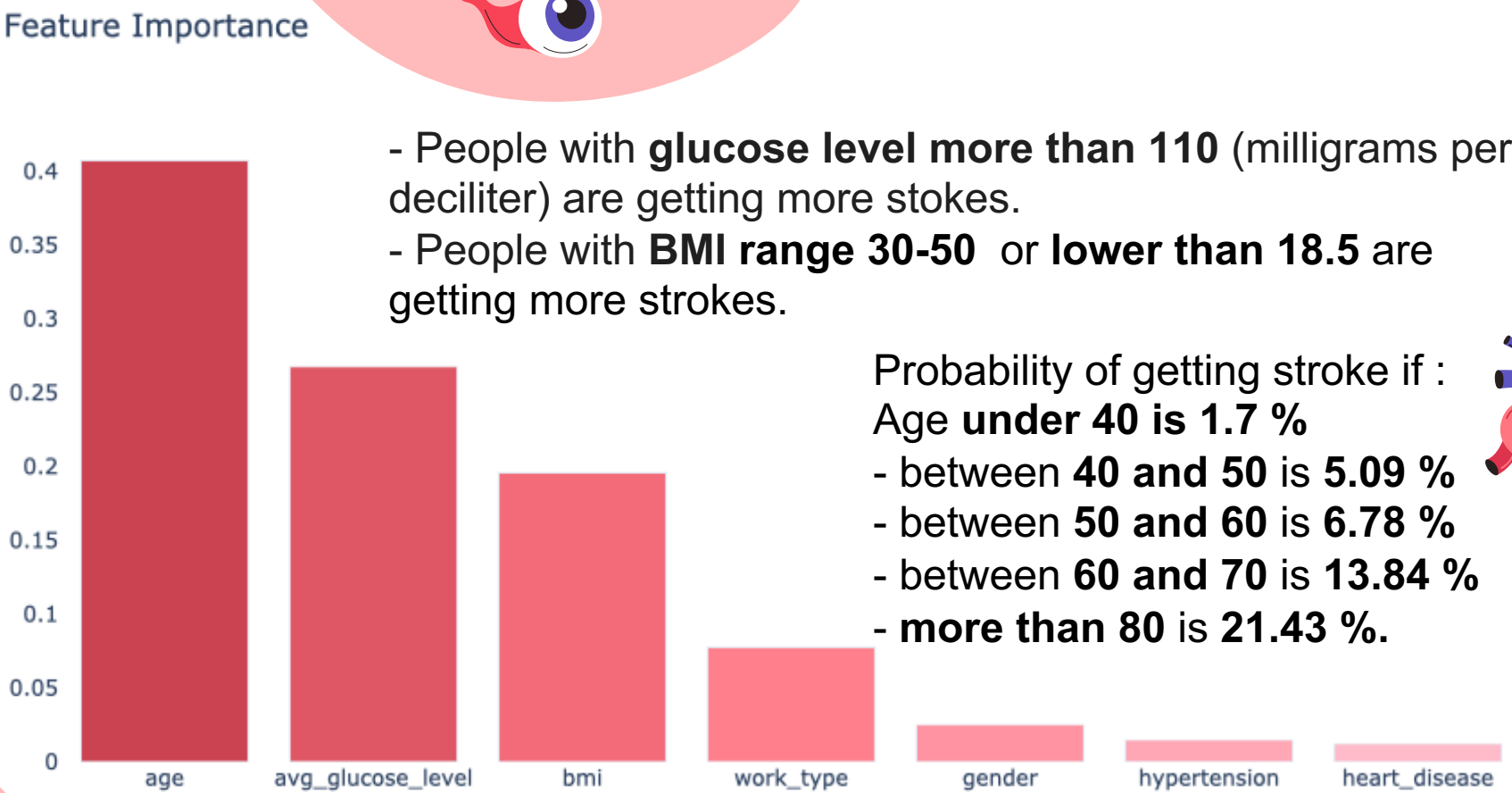
```
Accuracy: 0.9350563286944996
Confusion Matrix:
 [[689  73]
 [ 25 722]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.90      0.93       762
           1       0.91      0.97      0.94       747

    accuracy                           0.94      1509
   macro avg       0.94      0.94      0.94      1509
weighted avg       0.94      0.94      0.94      1509
```

- The model achieved a strong **93.51% accuracy**. It demonstrated high precision and recall for both classes, with **F1-scores** of **0.93 and 0.94.**



- **722 true positive** predictions and **689 true negative** predictions. There were **73 false positive** predictions and **25 false negative** predictions.

summary



- People with **glucose level more than 110** (milligrams per deciliter) are getting more stokes.
- People with **BMI range 30-50** or **lower than 18.5** are getting more strokes.

Probability of getting stroke if :
Age **under 40 is 1.7 %**
- between **40 and 50** is **5.09 %**
- between **50 and 60** is **6.78 %**
- between **60 and 70** is **13.84 %**
- **more than 80** is **21.43 %.**

THANK YOU