WELCOME

Stroke Prediction

Mentor: Văn Lương & Khánh Huyền

Member: Huy Anh

Introduction

1. Context & Objetive
2. Data Information
3. Exploratory data analysis
4. Model & Result
5. Evaluate Model
6. Reference

1. Context & Objetive

Context

According to the WHO stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths

Objetive

Predicts whether a patient is likely to get stroke based on various parameters

2. Data information

	Data	Note
	ID	Unique Identifier
Numerical	Age	Age of the patient
	AVG_Glucose_Level	AVG glucose in blood
	BMI	Body Mass index
Categorical	Gender	“Male”, “Female” or “Other”
	Hypertension	0 if patient doesn’t have hypertension 1 if patient have hypertension
	Heart_Disease	0 if patient doesn’t have heart disease 1 if patient have heart disease
	Ever_Married	“No” or “ Yes”
	Work_Type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed
	Residence_Type	"Rural" or "Urban"
	smoking_Status	"formerly smoked", "never smoked", "smokes" or "Unknown"
	Stroke	0 if patient doesn’t have Stroke 1 if patient have stroke

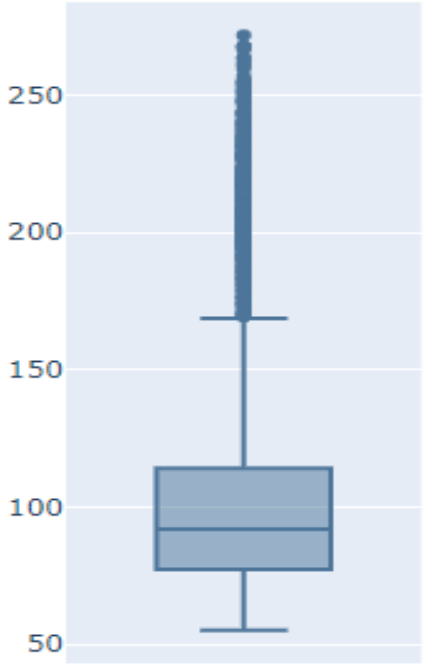
3. Exploratory Data Analysis

```
 #   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
```

About Data

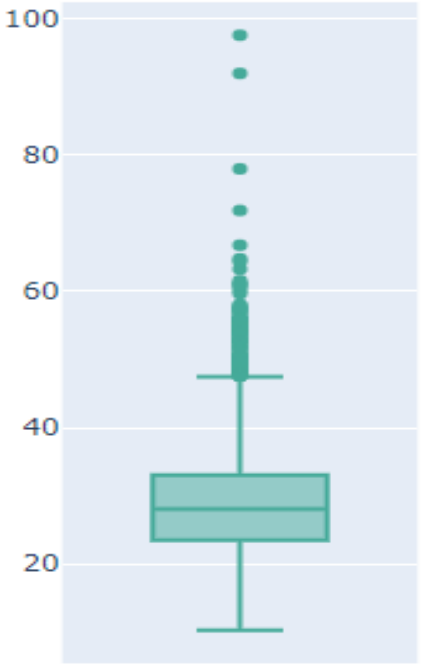
- Bảng có **5110 dòng** và **12 cột**
- **Cột BMI có dữ liệu null**
- Dữ liệu của cột **“Age”** đang là **dạng float**

3. Exploratory Data Analysis



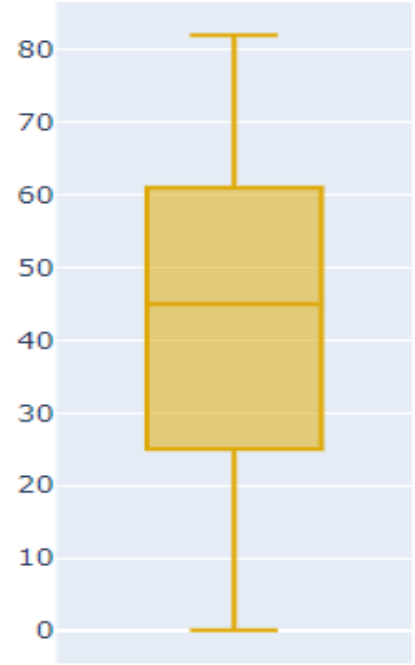
AVG_Glucose

- Max: 372.74
- Min : 55.12
- Median: 91.88



BMI

- Max: 97.6
- Median : 28.1
- Median: 10.3

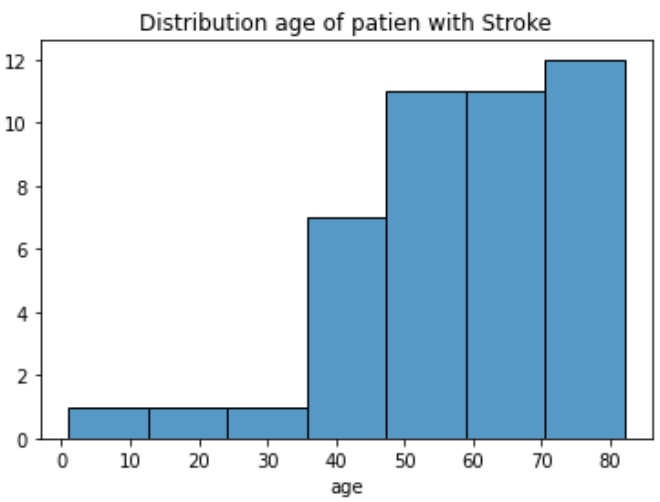


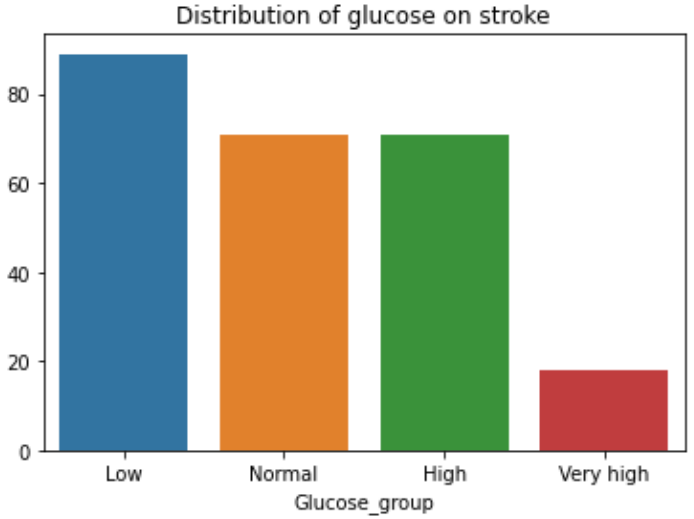
Age

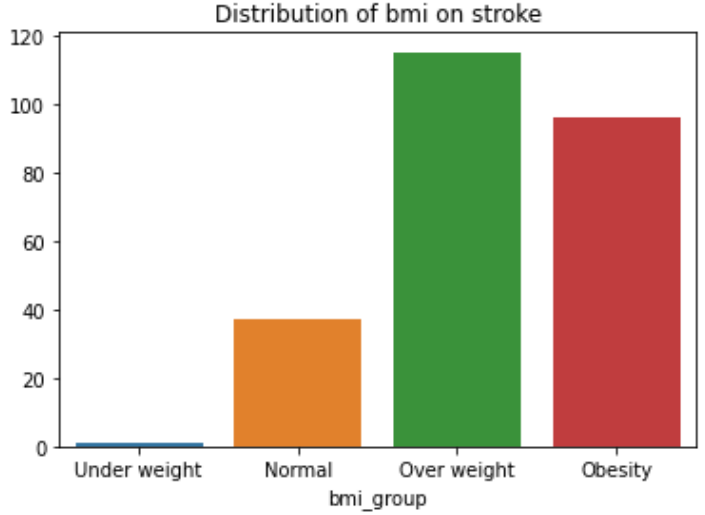
- Max: 82
- Min : 0
- Median: 45

- Dữ liệu của AVG_Glucose và BMI có outlier, nhưng AVG_Glucose BMI có thể là các chỉ số nguyên nhân dẫn đến stroke
- **Chia giá trị dữ liệu thành các nhóm nhỏ:** ở các cột “AVG_glucose”, “Bmi” và “Age”
- **BMI:** Do có giá trị null, nên đã thay thế giá trị null bằng một giá trị cụ thể rồi chia thành các nhóm nhỏ

3. Exploratory Data Analysis

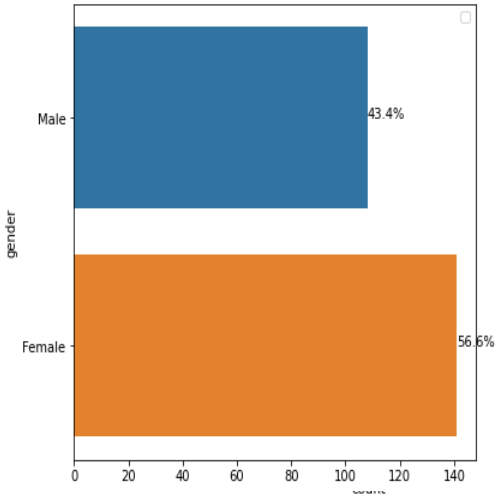


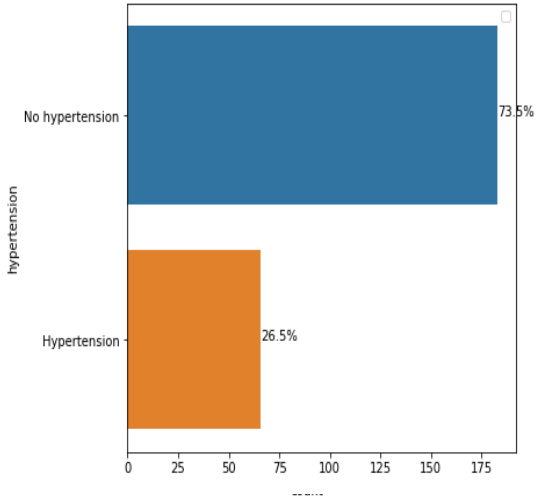


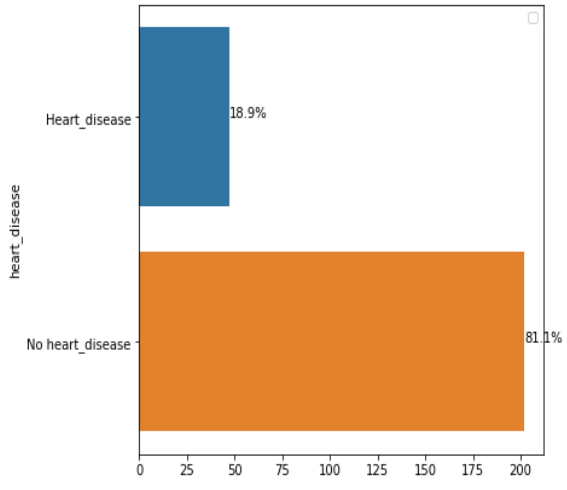


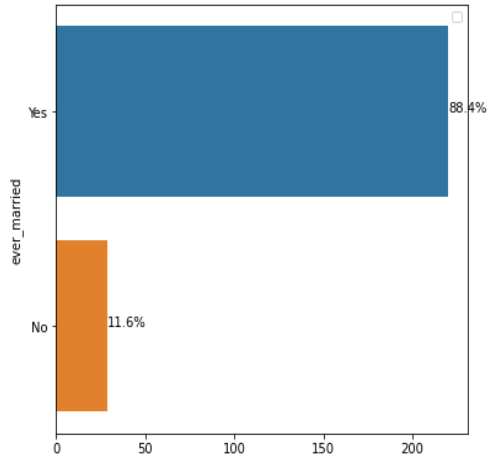
- Những người **trên 40 tuổi** có xu hướng bị “Stroke” nhiều hơn
- “Glucose” **Không ảnh hướng** đến việc bị “Stroke”
- **“BMI” cao** cũng là nguyên nhân dẫn đến “stroke”

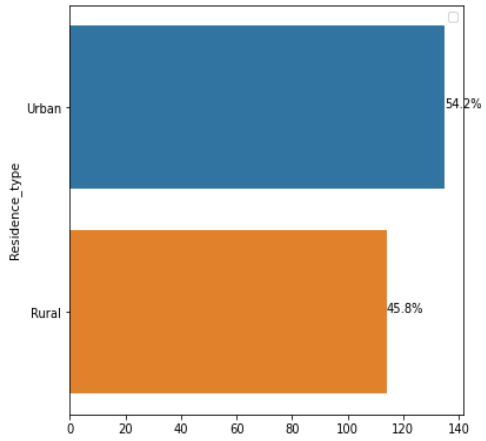
3. Exploratory Data Analysis

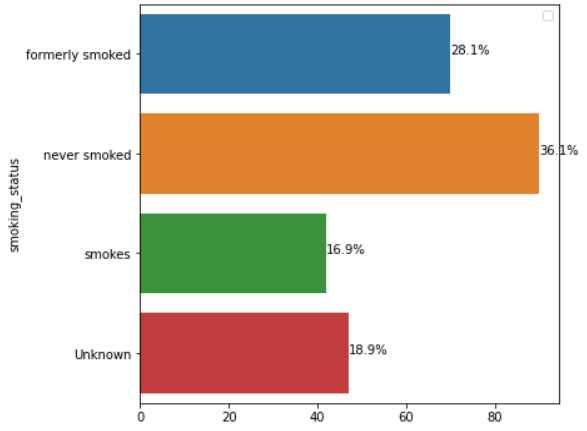






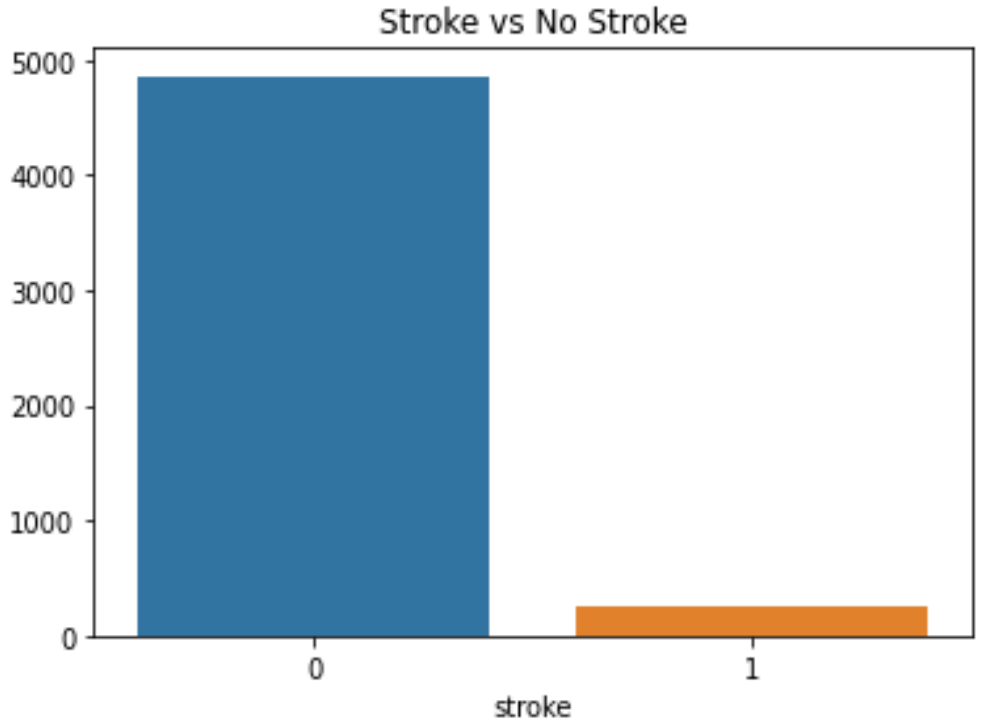






- Trong những người bị stroke tỉ lệ **nữ giới** chiếm 56%, **nam giới** chiếm 43.4%
- **Thành thị** có tỉ lệ bị stroke **cao hơn** là ở **nông thôn**
- Những người **có gia đình** có tỉ lệ stroke **cao hơn** là **chưa lập gia đình**
- Hypertension, Heart Disease và Smoking thì không liên quan đến Stroke

4. Modeling



- Sau khi xử lý và làm sạch dữ liệu, Tiến hành **Enconding** các Categorical data
- Dữ liệu đang bị **mất cân bằng** giữa “Stroke” và “No Stroke” => Thực hiện cân bằng data để giúp cho mô hình hoạt động được tốt và chính xác hơn

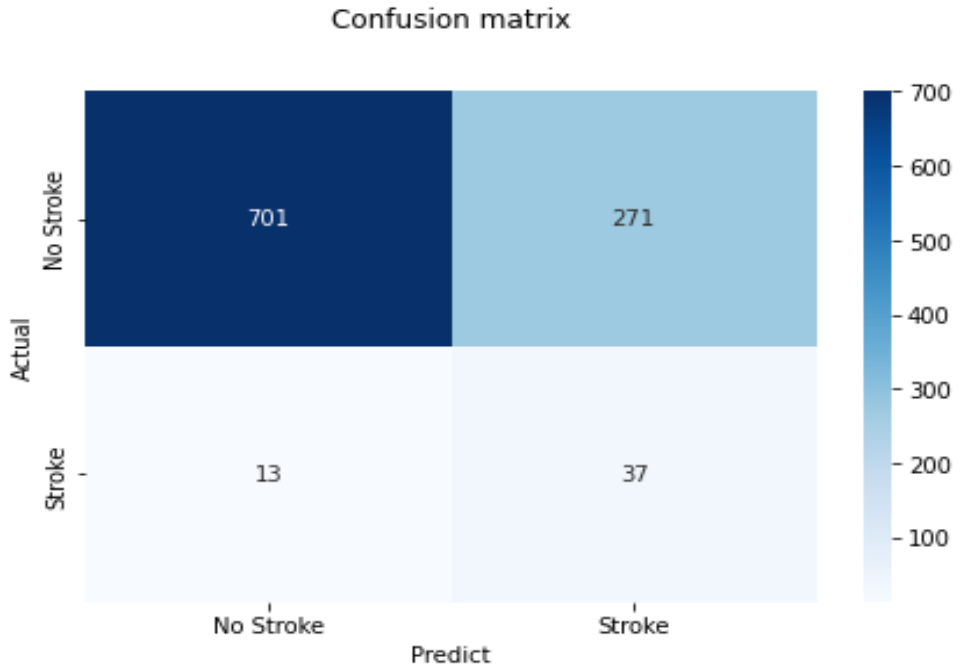
4. Modeling

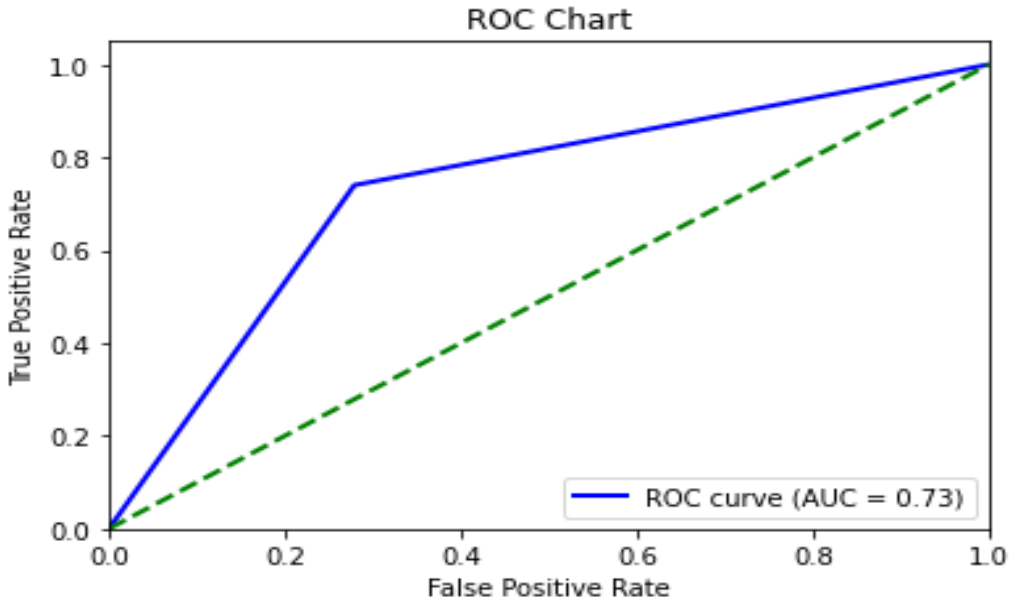
Model	Accuracy	Precision	Recall	F1_Score
Navie Bayes	72%	12%	74%	21%
Mnb	70%	9%	58%	16%
Logistic Regresion	70%	8%	48%	14%
Decision Tree	72%	10%	62%	18%
Random Forest	79%	11%	48%	18%
SVM	66%	9%	62%	15%
KNN	84%	11%	32%	16%
Adaboost	81%	6%	20%	9%
Gradien	79%	7%	26%	11%
Xgb	79%	7%	28%	12%
Light GBM	77%	8%	34%	13%

- Sau khi áp dụng các model để tìm dự đoán thì KNN có accuracy là 84% ,Adaboost 81% và random forest là 79%
- Tuy nhiên đây là bài toán phân loại Stroke nên ta xét thêm các chỉ số f1, Precision và Recall thì Gnb lại cho kết quả cao nhất

-> Chọn mô hình Navie Bayes

5. Predict & Evaluate





- **Precision = 12%** -> trong tổng 308 người dự đoán là bị stroke , thì có 37 người thực sự bị stroke.
- **Recall = 74%** -> Trong 50 người thực sự bị stroke thì mô hình đoán đc 37 người
- **F1 = 21%**

- **FNR = FN/(TP + FN)** = 26% -> Tỉ lệ dự báo sai của mô hình là 26%
- **FPR = FP/(FP+ TN)** = 27% -> Tỉ lệ dự báo nhầm của mô hình là 27%
- **AUC** = 0.73

6. Reference

- **Linkdataset:** https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- **Link Bmi level**: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
- **Link glucose level:** https://www.credihealth.com/blog/normal-blood-glucose-levels-in-adults/

THANKS