Automatic Stroke Detecting System

Capstone Project

DSA_0395

- **Data Source:** https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- **Data attributes/columns :** ID , Gender , Age , Hypertension , Heart disease , Ever married , Work type , Average glucose level , BMI (Body Mass Index) , Smoking status , Stroke

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
13	8213	Male	78.0	0	1	Yes	Private	Urban	219.84	NaN	Unknown	1
19	25226	Male	57.0	0	1	No	Govt_job	Urban	217.08	NaN	Unknown	1
27	61843	Male	58.0	0	0	Yes	Private	Rural	189.84	NaN	Unknown	1

Image 01: Data table

- **Data pre-processing steps:**

1. Identify duplicates
2. Identify null values
3. Remove null values
4. Check unique values
5. Check memory usage
6. Check data type
7. Change data types into Category
8. Assign numbers into Objects
9. Rearrange column order
10. Data visualization
11. Check correlation/Heat map
12. Train data set split

- **Output for :**

data.describe(include='all').transpose()

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	5110.0	NaN	NaN	NaN	36517.829354	21161.721625	67.0	17741.25	36932.0	54682.0	72940.0
gender	5110	3	Female	2994	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	5110.0	NaN	NaN	NaN	43.226614	22.612647	0.08	25.0	45.0	61.0	82.0
hypertension	5110.0	NaN	NaN	NaN	0.097456	0.296607	0.0	0.0	0.0	0.0	1.0
heart_disease	5110.0	NaN	NaN	NaN	0.054012	0.226063	0.0	0.0	0.0	0.0	1.0
ever_married	5110	2	Yes	3353	NaN	NaN	NaN	NaN	NaN	NaN	NaN
work_type	5110	5	Private	2925	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Residence_type	5110	2	Urban	2596	NaN	NaN	NaN	NaN	NaN	NaN	NaN
avg_glucose_level	5110.0	NaN	NaN	NaN	106.147677	45.28356	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	NaN	NaN	NaN	28.893237	7.854067	10.3	23.5	28.1	33.1	97.6
smoking_status	5110	4	never smoked	1892	NaN	NaN	NaN	NaN	NaN	NaN	NaN
stroke	5110.0	NaN	NaN	NaN	0.048728	0.21532	0.0	0.0	0.0	0.0	1.0

Image 02

- **Correlation Matrix :**

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1.000000	0.274329	0.256999	0.235725	0.333738	0.232221
hypertension	0.274329	1.000000	0.115991	0.180543	0.167811	0.142515
heart_disease	0.256999	0.115991	1.000000	0.154525	0.041357	0.137938
avg_glucose_level	0.235725	0.180543	0.154525	1.000000	0.175502	0.138936
bmi	0.333738	0.167811	0.041357	0.175502	1.000000	0.042374
stroke	0.232221	0.142515	0.137938	0.138936	0.042374	1.000000

Image 03: Correlation matrix

- **Used ML model : Logistic Regression**

```
model = LogisticRegression()
```

- Heat map,

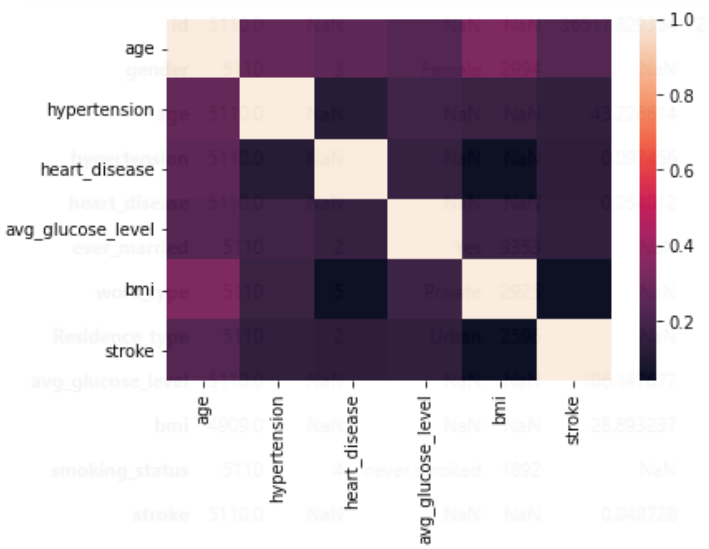

Image 04: Heat map

- **One hot encode prediction table,**

	y_act	y_pred	y_pred_prob_0	y_pred_prob_1	y_pred_0
4336	0	0	0.747968	0.252032	1
3709	0	0	0.984575	0.015425	1
964	0	0	0.925583	0.074417	1
2647	0	0	0.995012	0.004988	1
3262	0	0	0.981664	0.018336	1

Image 05

- **Confusion matrix,**

y_pred	0	All
y_act		
0	1401	1401
1	72	72
All	1473	1473

Image 06

- **Accuracy:** 0.9511201629327902 (95%)