Data Science Bootcamp

Project document

•Goal:

•the purpose of this project is to build a classification model <u>that helps to predict whether a patient is likely to get stroke</u> *based on the input parameters* like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Tools

Data Processing

Pandas, Numpy

Modelling

scikit-learn , Imbalanced-learn

Visualization

Matplotlib, Seaborn, Plotly, cufflinks

Stroke Prediction Dataset
11 clinical features for predicting stroke events

Data source "Kaggle":

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Target

```
df.head()
```

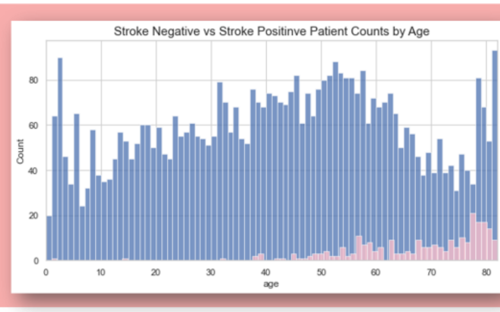
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1

Data has **5,110** instances and **11** attributes. "id is excluded"

Methods and techniques used to clean the dataset and prepare it….

1. Check for the null values (missing values).
2. Check for duplications.
3. Drop unnecessary columns (id).
4. Drop 'Other' gendered individuals to simplify the mathematical computations "since there is only one patient that has 'other' value as a gender"
5. 'bmi' column has 201 nan values so I decided to impute them with mean.
6. Round 'age' column and convert data type to integer.

Visualizing age and stroke correlation.



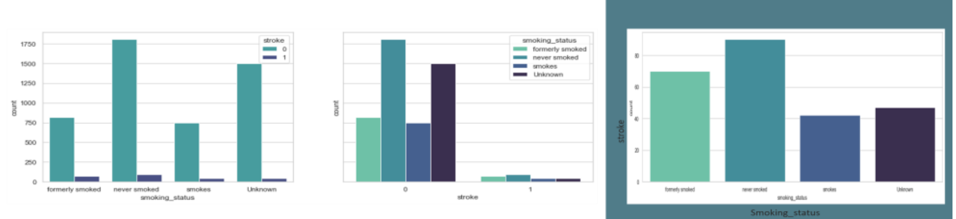
As we can see, we have <u>a balanced age distribution in the stroke negative group</u>.
However, <u>stroke positive patients are stacked to the right</u> side (older people).
<u>we can clearly see that age plays a huge role in predicting stroke.</u>

	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke
162	Female	1	0	0	No	children	Urban	70.37	28.9	Unknown	1
245	Female	14	0	0	No	children	Rural	57.93	30.9	Unknown	1

Data Preparation for Modelling:

1. All categorical data transformed into numerical by using dummies and converting to Boolean.
2. Encoding by "Gitting dummies" for 'work_type', 'smoking_status' columns.
3. Convert 'gender' column to Boolean 1 male - 0 female.
4. Convert 'ever_married' column to Boolean 1 married - 0 not married.
5. Convert 'residence_type' column to Boolean 1 urban - 0 rural.
6. scale the variance to make the data closer to normal distribution ('age', 'avg_glucose_level', 'bmi').
7. Splitting data into (testing & training) and then I make resampling by using SMOTE

Smoking status distribution in stroke patients



The highest stroke group is non-smokers!! Strange but true. This may be due to other factors and bias.

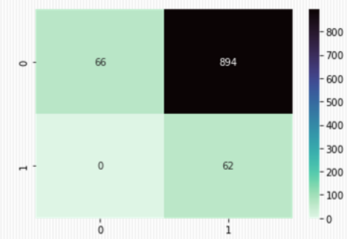
Stroke Relation with Smoking

Starting to build the Models

Logistic Regression Model

	Precision	Recall	f1-score	support
0	1.00	0.07	0.13	960
1	0.06	1.00	0.12	62

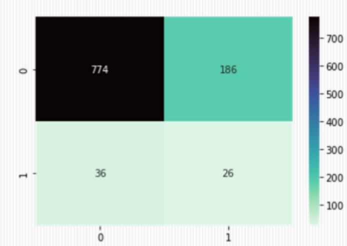
Accuracy			0.13	1022
Macro avg	0.53	0.53	0.13	1022
Weighted avg	0.94	0.13	0.13	1022



Random Forest Classifier Model

	Precision	Recall	f1-score	support
0	0.96	0.81	0.87	960
1	0.12	0.42	0.19	62

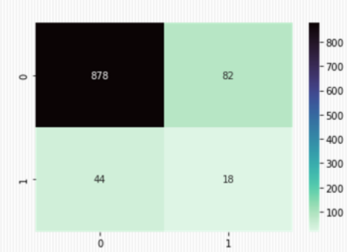
Accuracy			0.78	1022
Macro avg	0.54	0.61	0.53	1022
Weighted avg	0.91	0.78	0.83	1022



XGBOOST Model

	precision	recall	f1-score	support
0	0.95	0.91	0.93	960
1	0.18	0.29	0.22	62

Accuracy			0.88	1022
Macro avg	0.57	0.60	0.58	1022
Weighted avg	0.91	0.88	0.89	1022



XGBoost gives the best results with 0.88 accuracy and 0.91 recall

Done by: Eman Ahmed Alzhrani

supervised by: Ms. Mariam Elmasry