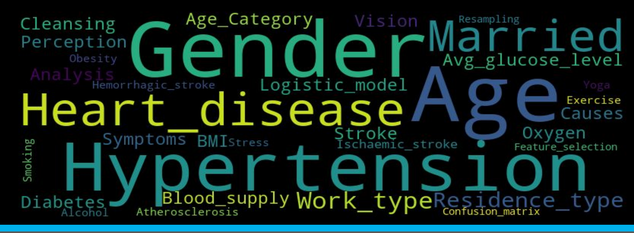
Stroke Prediction

By: Sarah El Shatby



Problem Statement

Analysing “healthcare-dataset-stroke-data.csv” dataset to identify patients who are more likely to have a stroke.

Data Source

The dataset downloaded from Kaggle contains 5110 records with 1 attributes.

Analysis

I- Assessing
II- Cleaning

Variables

Categorical variables

1. Gender
2. Hypertension
3. Heart disease
4. Ever married
5. Work type
6. Residence type
7. Stroke

Continuous variables

1. Age
2. BMI
3. Average glucose level

Feature Selection

- Using chi squared statistical test to select the most related features that will be used to build the model.

Most related features:

Age	**3662.95**
Average glucose level	**1700.8**

Least related features:

Gender	**0.224**
Residence type	**0.639**

Prediction model (Logistic Regression)

- A simple binary logistic model to predict strokes using age & average glucose levels.
- The model has an accuracy of **77%** and was able to predict a stroke for a new dataset of patients.

-New dataset

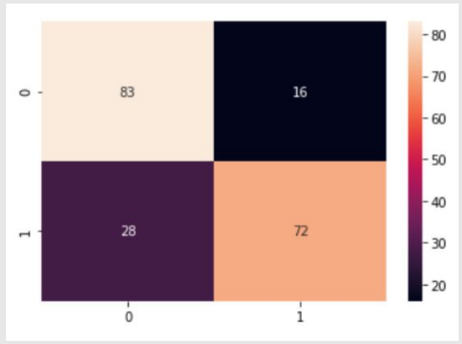
	age	avg_glucose_level
0	20	120
1	35	140
2	70	160
3	80	200
4	90	170
5	100	150

- Prediction for the new dataset

```
   age  avg_glucose_level
0   20                120
1   35                140
2   70                160
3   80                200
4   90                170
5  100                150
[0 0 0 0 0 1]
```

Model Evaluation

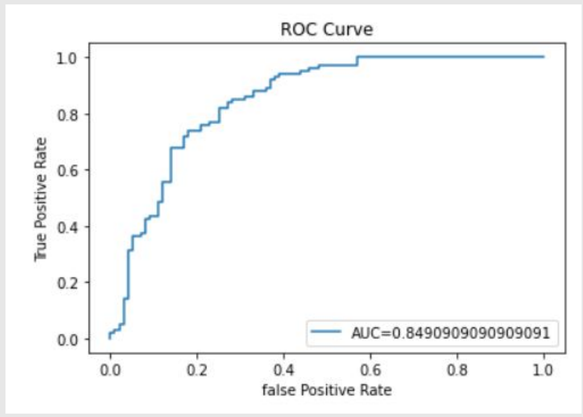
- A confusion matrix was used to evaluate the model’s performance.
- A heat map was used to visualize the matrix.



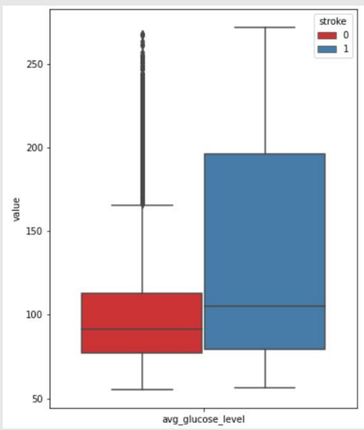
Model report

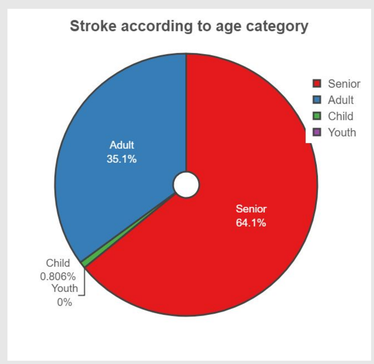
```
Accuracy:  0.7788944723618091
Recall:  0.8383838383838383
Precision: 0.7477477477477478
CL Report:               precision    recall  f1-score   support

           0       0.82      0.72      0.77       100
           1       0.75      0.84      0.79        99

    accuracy                           0.78       199
   macro avg       0.78      0.78      0.78       199
weighted avg       0.78      0.78      0.78       199
```

ROC Curve



Visualizations





Conclusions

- Seniors have a higher risk of strokes.
- The level of average glucose also shows a significant correlation to strokes.
- Hypertension is the 4th most related feature to strokes & heart disease is the 3rd most impacting one.

References

Stroke Project References