Healthcare-Stroke- Prediction

Abstract:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient

Design:

In order to determine the types of transaction statue, data was downloaded from Kaggle. Then, multiple models were implemented to get the best one to make a clear classification

Data Description:

The original source for this data is here, and we have taken from kaggel . This data set is name as
healthcare-dataset-stroke-data contain 12 column and 5111 rows.

Features:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not
*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Algorithms:

- Exploratory Data Analysis was done to the dataset.
- Building multiple models and finding out the well-suited one for this specific dataset.

Cleaning:

drop null values

Feature Engineering:

dummy variable

Model Building:

Around 5 models were tried and played with to get the best model that goes hand in hand with the dataset. After performing simple train and validation on the models one was chosen for further investigation. Models trained was:

- Logistic regression (Baseline)
- KNN
- Random forest
- XGB Classifier
- Decision trees

The Best Models: Logistic regression

Dealing with Class Imbalance by .

Evaluating Cross Validation and gridsearch for the best models.

Tools Description

The main technologies and libraries that will be used :

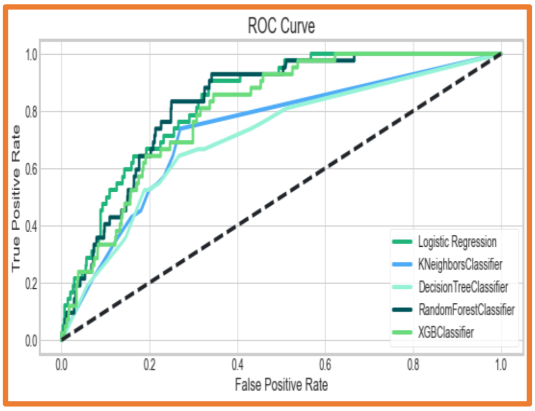
1. Python
2. Jupyter Notebook

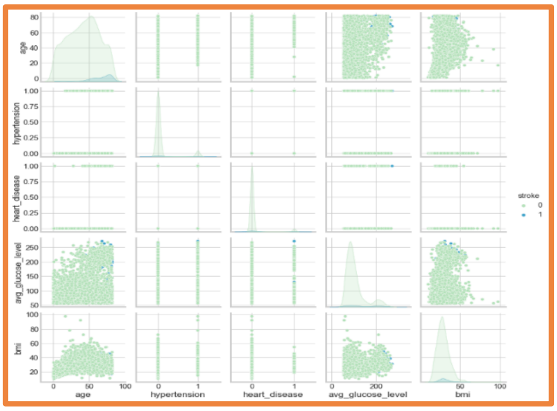
Libraries:

1. Pandas
2. Matplotlib
3. Seaborn
4. Numpy
5. Sklearn

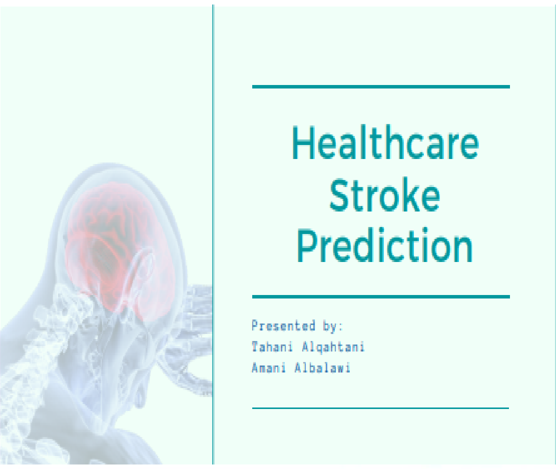
Communication:

Charts:



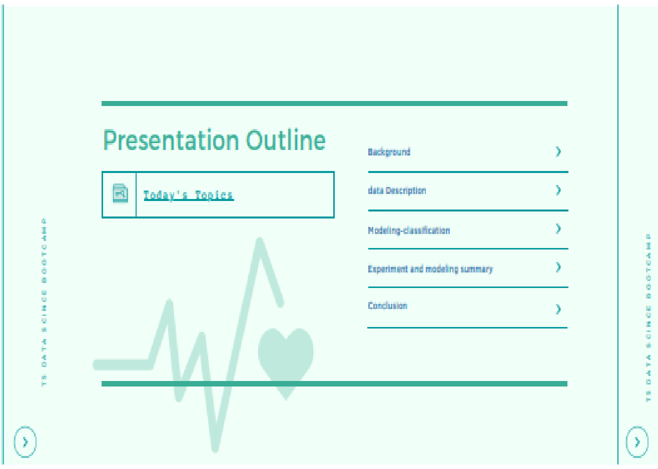


Presentation snips:



Healthcare Stroke Prediction

Presented by:
Tahani Alqahtani
Amani Albalawi



Presentation Outline

Today's Topics

- Background
- data Description
- Modeling-classification
- Experiment and modeling summary
- Conclusion

Background

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally

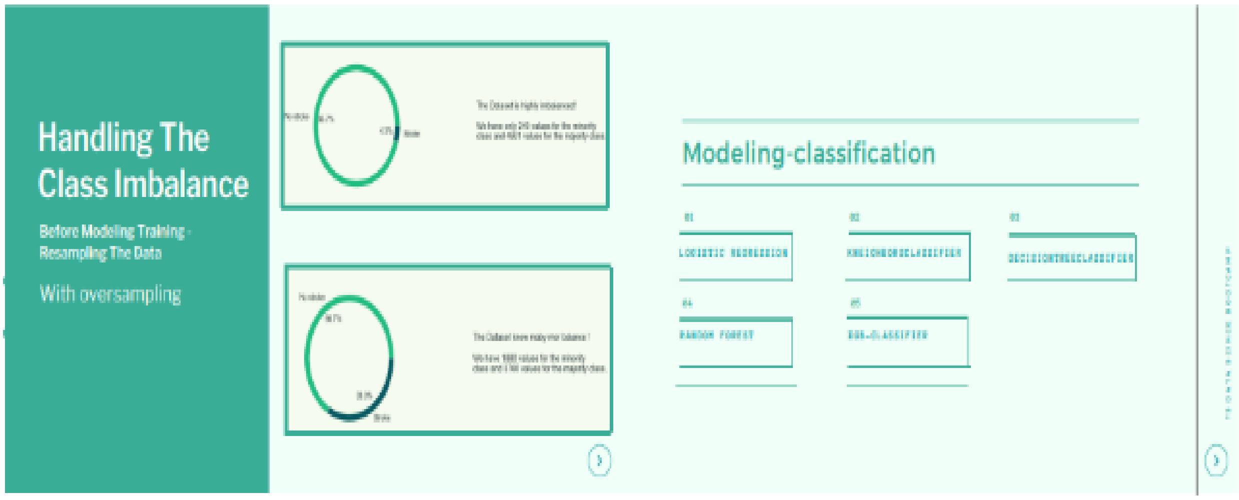
11%

total deaths of stroke

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Data description:

The original source for this data from kaggel . This data set is name as healthcare-dataset-stroke-data contain 12 column and 5111 rows.



Handling The Class Imbalance

Before Modeling Training - Resampling The Data

With oversampling

The Dataset is highly imbalanced!

We have only 249 values for the minority class and 4861 values for the majority class.

The Dataset is now mainly more balance !

We have 1888 values for the minority class and 3766 values for the majority class.

Modeling-classification

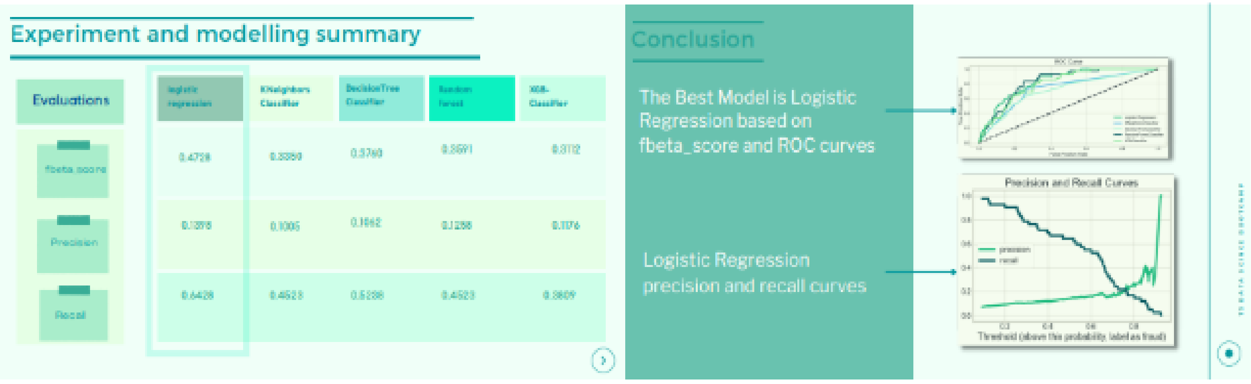
01 LOGISTIC REGRESSION

02 KNEIGHBORSCLASSIFIER

03 DECISIONTREECLASSIFIER

04 RANDOM FOREST

05 XGB-CLASSIFIER



Experiment and modelling summary

Evaluations	logistic regression	KNeighbors Classifier	DecisionTree Classifier	Random forest	XGB-Classifier
fbeta_score	0.4728	0.3350	0.3760	0.3591	0.3712
Precision	0.1378	0.1005	0.1662	0.1288	0.1176
Recall	0.6428	0.4523	0.5238	0.4523	0.3809

Conclusion

The Best Model is Logistic Regression based on fbeta_score and ROC curves

Logistic Regression precision and recall curves

ROC Curve
Precision and Recall Curves
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.2
0.4
0.6
0.8
Threshold (above this probability, label as fraud)

By: Amani Albalawi

&

Tahani Alqhtani