Final Poject FDS: Stroke Prediction

Dipartimento di Ingegneria Informatica, Informatica e Statistica

SAPIENZA
UNIVERSITÀ DI ROMA

Andrea Potì, Amedeo Ranaldi,
Onur Ergun, Giulio D'Erasmo

Master Degree in Data Science

A. A. 2021-2022

Table of contents

1. Task and Motivation
2. Models, tools
3. Analysis

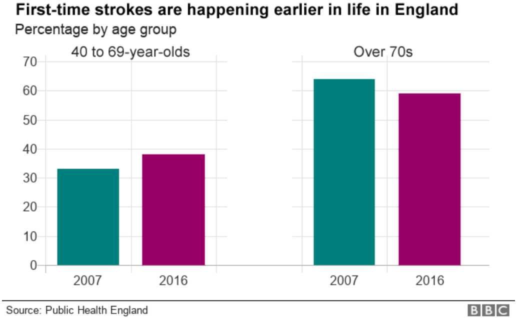
Task and Motivation

Task

Predicting the probability of a person having a stroke in the future.

Motivation

Each year over 13 million people will have a stroke and 5.5 million people dies as a result. Our motivation is to save lives by predicting if an individual will have a stroke or not.



First-time strokes are happening earlier in life in England

Percentage by age group

Source: Public Health England

Task and Motivation

Task

Predicting the probability of a person having a stroke in the future.

Motivation

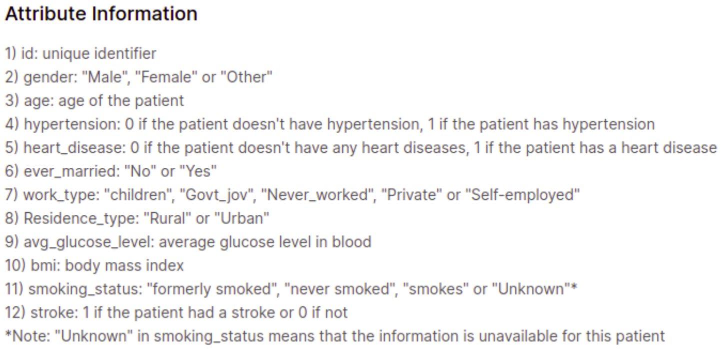
Each year over 13 million people will have a stroke and 5.5 million people dies as a result. Our motivation is to save lives by predicting if an individual will have a stroke or not.

Related work

Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults (Journal of the American Medical Informatics Association, 28(8), 2021, 1719–1727, 9 May 2021)

Tentative material and methods

- The Stroke Prediction dataset that we used has been taken from the following link : https://www.kaggle.com/fedesoriano/stroke-prediction-dataset



Attribute Information

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

Tentative material and methods

The models

For the implemention we choose to use and compare:

- Naive Bayes
- Logistic Regression
- Gaussian Discriminant Analysis (GDA)

Benchmark

Possible problem of the dataset:

- bmi column: some nan values to manage
- age column: the time is float... should we convert it to int?
- smoking-status column: some values are unkown.