Final Report: Stroke Prediction

BY ENJEL

Business Background

Background

I am a data scientist working in a hospital in New York, USA.

Stroke is a dangerous disease. To enable doctors to diagnose their stroke patients more accurately and more early on, we were assigned to build a stroke predictor based on a patient's condition.

Business Objectives

The objective of this project is to predict whether or not someone is likely to have a stroke event in the future, so that doctors can quickly warn their patients, give them better care and attention, and assign the right medications for them.

Expected Output

The output of this project is a model that can provide predictions of whether or not someone will have a stroke event or not, based on its given features, such as age, BMI, work type, etc.

Project Limitation

Due to limitations of time and data on this project, I decided to focus on using features from one dataset, which these contains features: ID, age, BMI, work type, heart disease, hypertension, smoke status, and marital status.

Analytical Approach

MACHINE LEARNING TECHNIQUE

Supervised Learning (Classification) to predict whether or not someone will have a stroke event.

PERFORMANCE MEASURES

Recall, precision, F1, and ROC-AUC score, to minimize the error of stroke prediction.

Data Understanding and Data Exploration

Data Info

Data Shape

This dataset of Stroke Prediction was obtained from Kaggle.

It has 5110 rows and 12 columns.

Features Overview

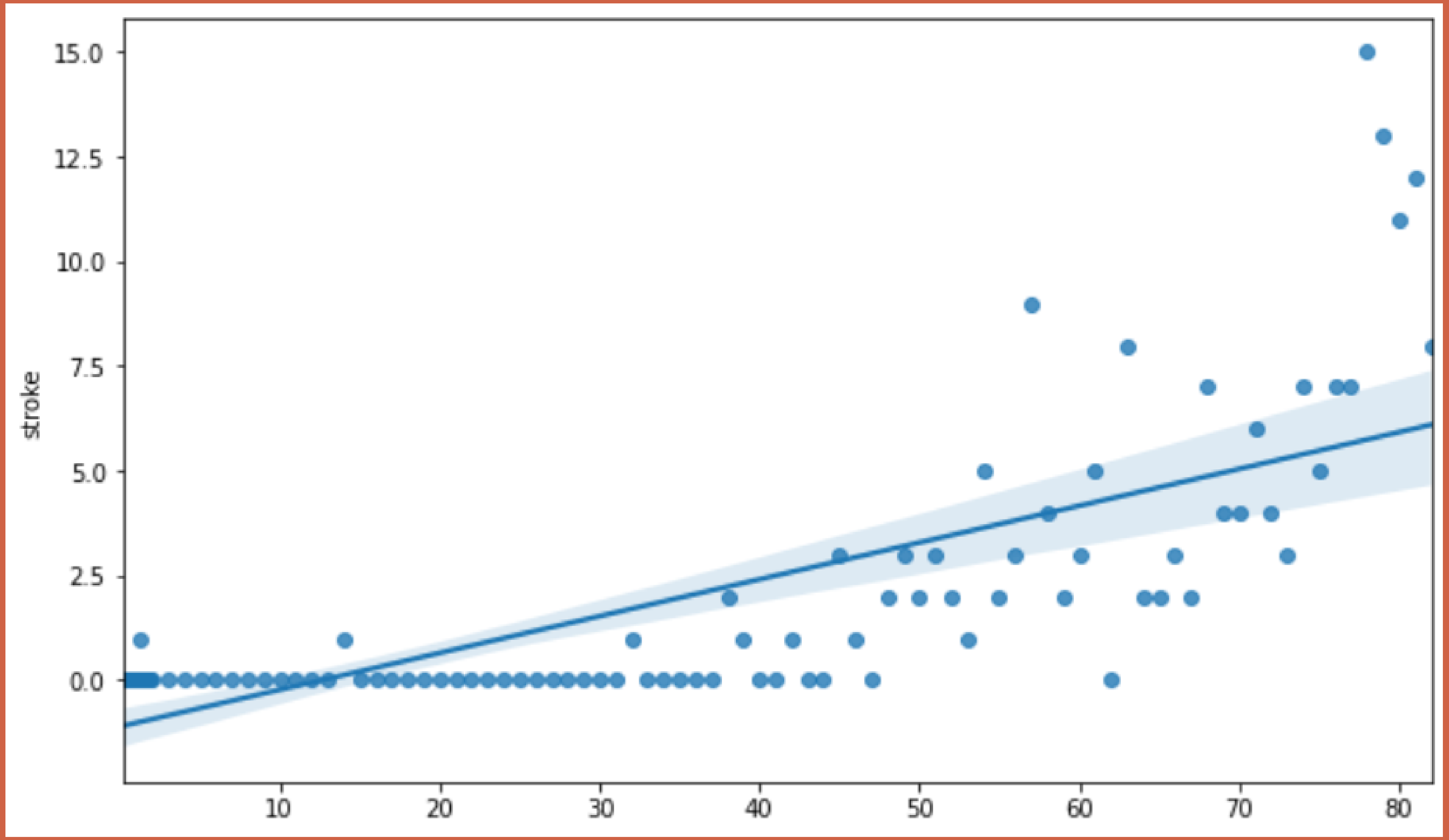
Categorical Features: gender, ever_married, work_type, Residence__type, smoking_status

Binary Numerical Features: hypertension, heart_disease, stroke

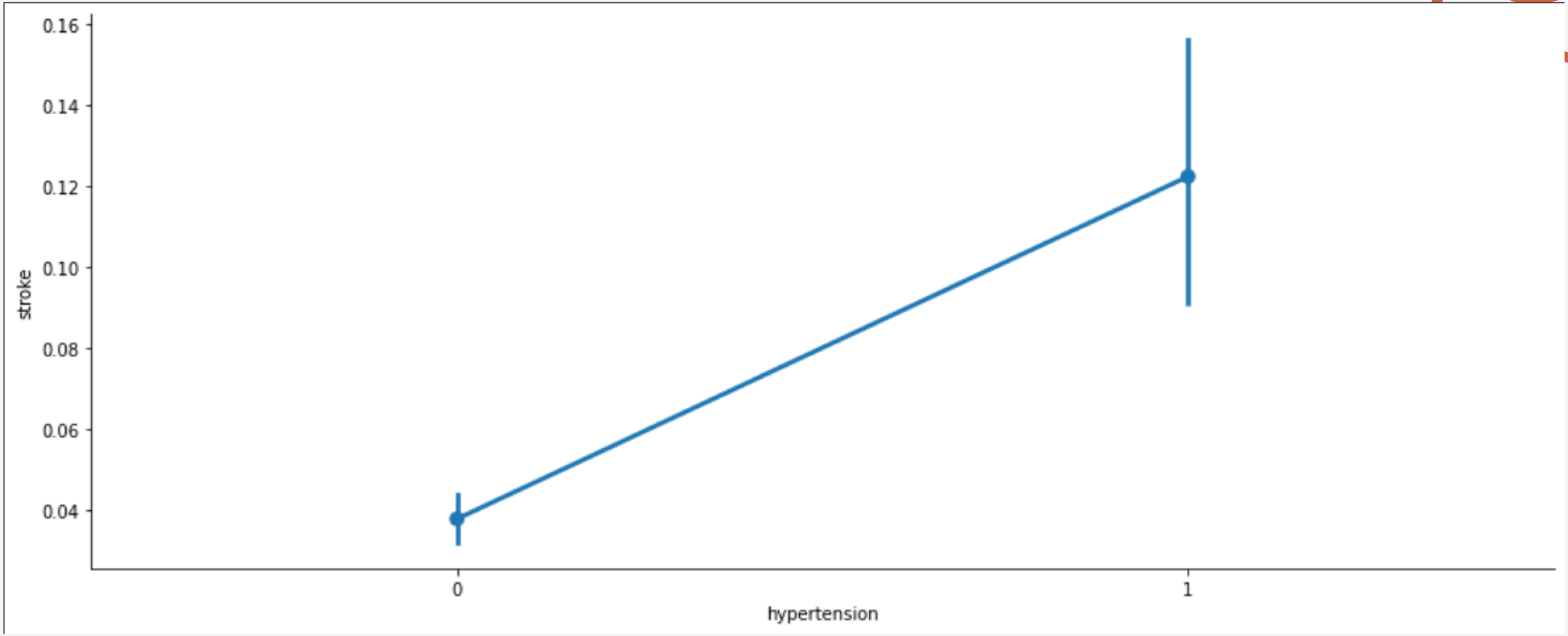
Continuous Numerical Features: age, avg_glucose_level, bmi

Issues to note

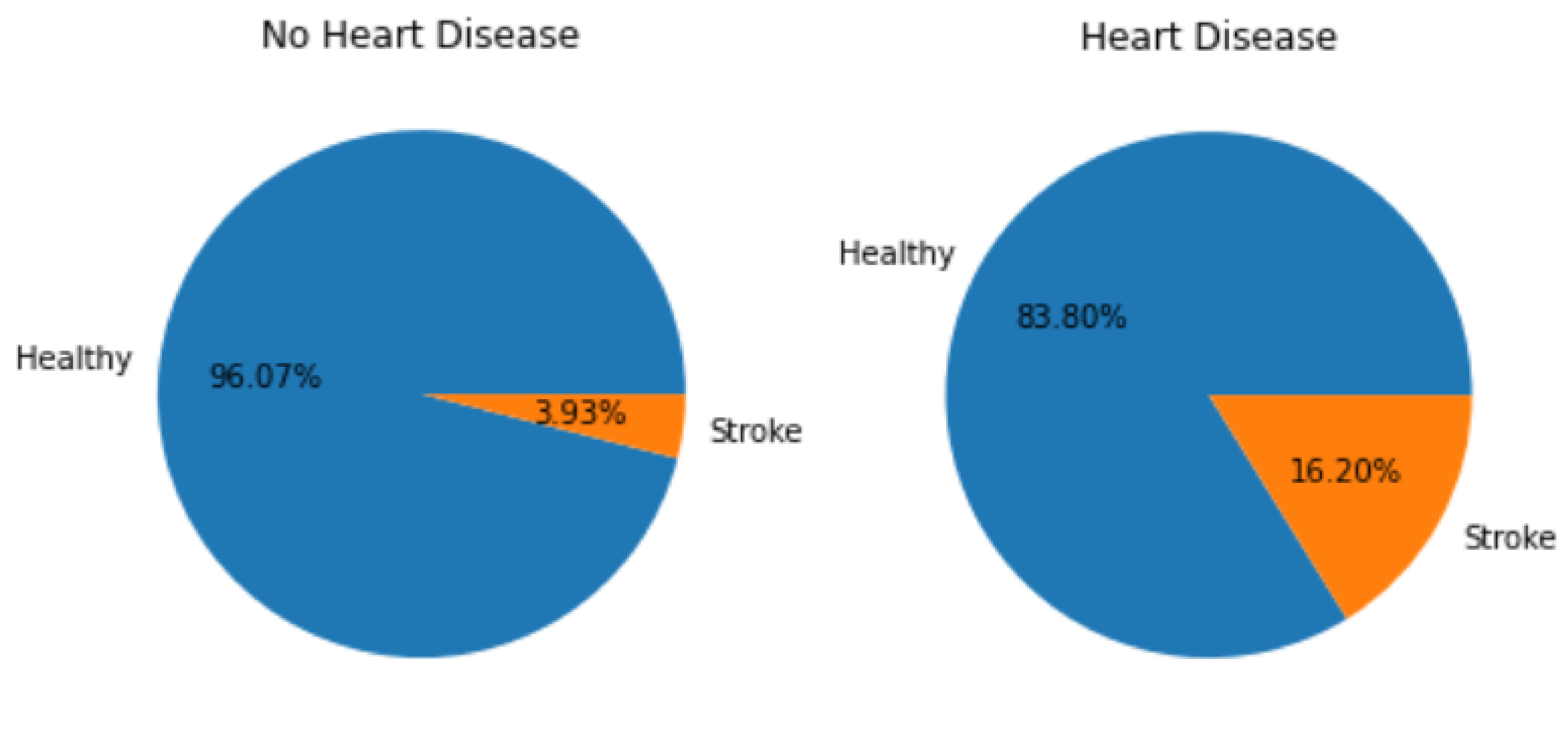
There are a few issues to note:
The target column is very imbalanced (5: 95).
There are null values in the bmi column.
There is an unusual value in the gender column ("other").



As we suspected, the age column is a relatively good predictor of stroke. The older one gets, the more likely he/she is going to have a stroke.

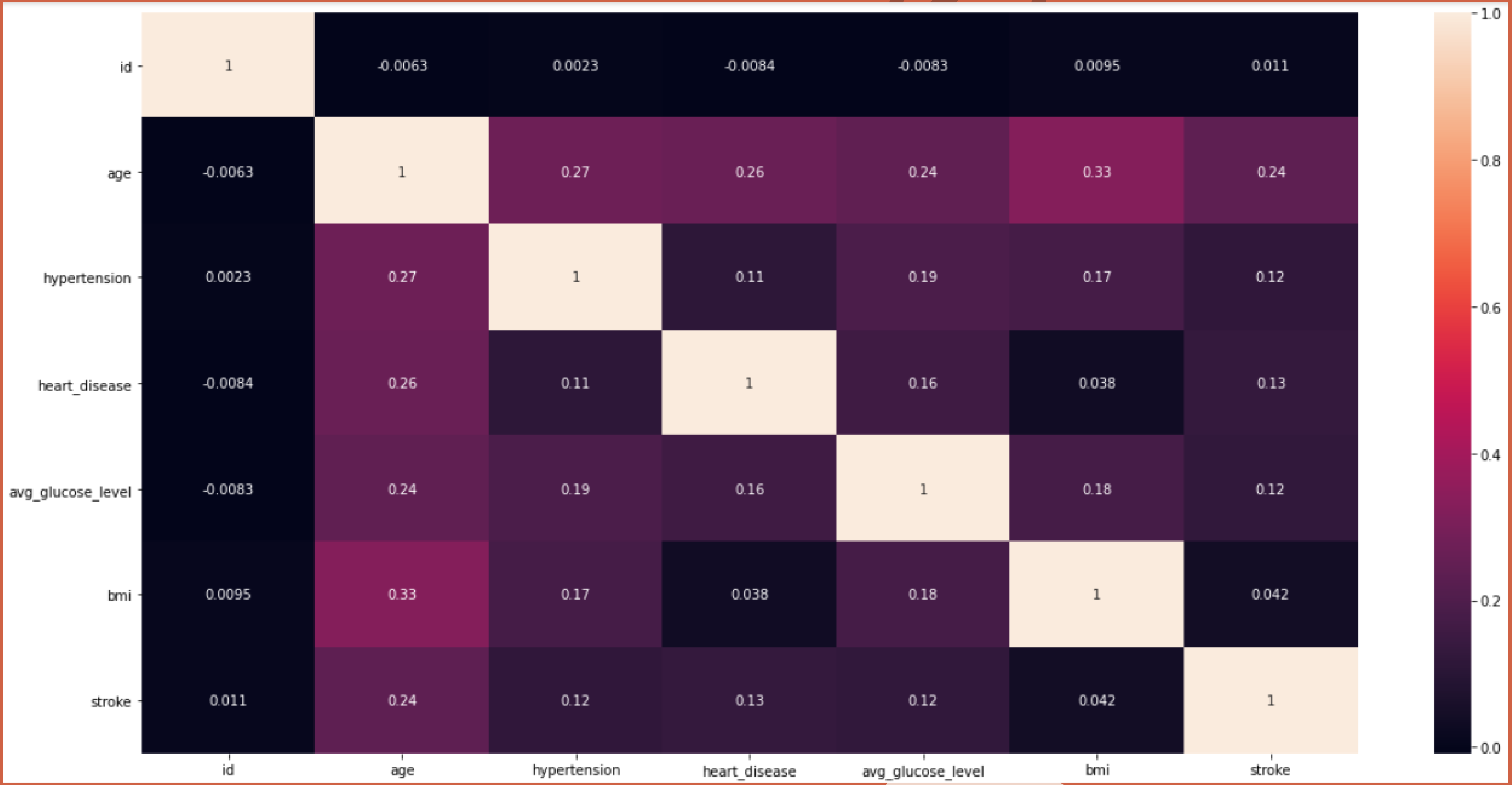


From this plot, we can also see that hypertension is a good predictor of strokek. If someone has hypertension, he/she will most probably have a stroke event.



Having a heart disease significantly increases the chance of someone having a stroke by 4 times. Therefore, heart disease is a good predictor of a stroke event.

Heatmap of numerical values


id
age
hypertension
heart_disease
avg_glucose_level
bmi
stroke
id
age
hypertension
heart_disease
avg_glucose_level
bmi
stroke