Exploratory Data Analysis: Which genes and their expression levels may be associated with developing familial alzheimer's disease?

Orfeas Gkourlias

2022-10-06

Introduction

This document aims to explore, analyse and explain the data set being used in answering the following research question: "Given 10 attributes, how do they compare in predicting the chances of a person's risk of a stroke?". As the res earch question implies, the data set consists of 10 attributes. This project has the goal of comparing those attributes, so that the most likely predictors for a stroke may be deduced. Some attributes affect each other, while others may not. Analysis of these correlations can help in finding the rankings of the attr ibutes.

To get a feel for what the scope and attributes of the data set consists of, it will be loaded and the first 10 results will be displayed.

```r
main <- read.csv("../data/stroke-data.csv")
head(main)
```

```
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            0             1          Yes       Private
## 2 51676 Female  61            0             0          Yes Self-employed
## 3 31112   Male  80            0             1          Yes       Private
## 4 60182 Female  49            0             0          Yes       Private
## 5  1665 Female  79            1             0          Yes Self-employed
## 6 56669   Male  81            0             0          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21  N/A    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
## 4          Urban            171.23 34.4          smokes      1
## 5          Rural            174.12   24    never smoked      1
## 6          Urban            186.21   29 formerly smoked      1
```

```r
nrow(main)
```

```
## [1] 5110
```

There are 12 attributes, 10 of which will be used in the analysis: Gender, age hypertension, heart_disease, ever_married, work_type, residence_type, avg_gluco se_level, bmi and smoking_status. The last column indicates whether the person has already experienced a prior stroke. This can be used to the train the machine learning model which will be utilized to answer the research question.

There are 5110 entries in this data set. This is also why the row numbers will not be replaced with the id's, because there is no order in the id numbers. They exceed the number 5110.

The attributes and their units can be seen in the code book on the next page.

Codebook

```
knitr::kable(codebook)
```

Column	Unit	Description
ID	Number	Unique patient identifier
Gender	Text	“Male”, “Female” or “Other”
Age	Number	Age of patient
Hypertension	Boolean	Whether patient has hypertension
Heart_disease	Boolean	Whether patient has a heart disease
Ever_married	Boolean	Whether patient has ever been married
Work_type	Text	Occupation status of patient
Residence_type	Text	Patient living enviroment
Avg_glucose_level	Number	Average glucose level in blood
BMI	Number	Body mass index of patient
Smoking_status	Boolean	Whether patient smokes or not
Stroke	Boolean	Whether patient has ever experienced a stroke

Initial Data and Attributes

In this section, the attributes will be examined individually. What these attri butes could mean for the research question will be discussed. Correlations will be observed in a later section. Any preprocessing or cleanup required will also be performed in this section.

ID

This column is neither noteworthy for analysis or data structure. This column will therefore be dropped, because the dataframe used already has row numbers and this makes the ID redundant.

```
main <- main[2:12]
```

Age

The age of the patient. At first sight, it might look redundant for this data to be stored as a float, since most of the data consists of a rounded age number. Some of the entries contain very young patients. The younger a patient is, the more important the specifity of the age is, since the age difference is still significant at that point. It is for that reason that any patient under the age of 2 will contain a float number, with two decimal numbers. A couple of those instances will be shown in vector format below:

```
head(c(main[main$age < 2, 2]))
```

```
## [1] 1.32 0.64 0.88 1.80 0.32 1.08
```