PROJECT 1 GEZONDHEIDSZORG

#### INTRODUCTION

**AI ONDERZOEKERS**

7 MARCH 2023

Hossein, Hans, Kajal and Peter

#### Project information given by Hogeschool Rotterdam:

From a fictive site at Medical Center Randstad we can find data belonging to fictive patients regarding life expectancy and lifestyle. Our job is to 'scrape' that DataFrame.

#### The first goal

Develop a qualitative model that describes the relationship between lifestyle and age. If new data becomes available, the model must be easy to update. So an automatic data pipeline has to be built.

A modular application is built around this model that is suitable enough to be used in a doctor's practice or hospital. The application must therefore be able to run independently on a PC or laptop.

The doctor can enter the necessary data via a console application, and next see the results.   
The results are: the life expectancy that is predicted based on the behavior of certain variables.

***These variables/data are:***

• Genetic is the genetic age in years (not affected by lifestyle)

• Length is the height in centimeters

• Mass is the mass in kilograms

• Exercise is the amount of exercise in hours per day

• Smoking describes the number of cigarettes per day

• Alcohol describes the number of glasses of alcoholic drink per day

• Sugar describes the amount of sugar in lumps per day

Lifespan, age depends on the influence of lifestyle. To give the patient insight into a change of lifestyle, the doctor can adjust the style, for example adjusting the number of cigarettes per day. The patient is then shown how much influence this has on life expectancy. In addition, the patient is shown what the "predicted" influence will be on the insurance premium. After a session, the doctor can use the escape key to return to the starting point.

#### Our preliminary thoughts and specifications

• Establish if the program will run on different OS platforms

• Determine the software quality criteria: uptime, privacy

• Determine the output that needs to be delivered: web-based or window interface.

### R&D

• Determine the key criteria upon which a choice can be made which tool(s) for data analyses to use, i.e. Pandas. and MatPlotLib

• What are the different parts of the code in KATA: analysis, AI modelling, interface.

• Understand how the outcome of the model will be used: enduser input, adding patient records?

• Establish suitable naming convention for different files in the different phases.

### 00. HIGH LEVEL ARCHITECTURE – DATA PIPELINE

Afbeelding met diagram

Automatisch gegenereerde beschrijving

### 01. PIPELINE & DATA COLLECTION

Downloading and Cleaning DataBase through DB browser SQLite.  
  
We used the SQLite

### 02. DATA CLEANING

#### 2a. Viewing the Data INFO and Cleaning DataFrame

One of the most used method for getting a quick overview of the DataFrame, is the head() method. The head() method returns the headers and a specified number of rows, starting from the top.

DataBase *(before cleaning)* contains 4096 rows × 8 columns with different DataType values, however no 'non-null' values.

#### 2b. Cleaning DataFrame

#### Cleaning Data: Apply different techniques to clean the dataset

Data cleaning means fixing bad data in your data set. Bad data could be:  
• Empty cells  
• Data in wrong format  
• Wrong data  
• Duplicates *(no duplicates found)*

#### 2c. Info About the Data after cleaning

The DataFrames object has a method called info(), that gives you more information about the data set. Result in short: We have left 4084 rows each with their Unique ID containing float64 values and 9 columns and Column Headers (including Unique ID). 12 original rows were deleted.

***NOTE:*** *We found no outliers that require te be altered or to be deleted including its index (row).*  
*We left 'outliers' in because we believe the outliers are not unrealistic.*

#### 03. DATA TRANSFORMATION and ANALYSIS

After adding a new column 'BMI' our dataset results in 4084 rows × 10 columns.

Further analysis of Histograms, Heatmap, Scatter plots, pair-plots and Dendrograms gave us the follwing SUMMARY and ASSUMPTIONS:

We believe this DataFrame is 'flawed'. However discussing with our teachers we learned to see that this DataFrame is created with educational purposes.

**Pearson Correlation:**

There is one clear correlation between 'lifespan' and 'genetic'. In addition there is some correlation between 'exercise' and 'lifespan' and between 'alocohol' and 'lifespan'. Nevertheless their correlation-values in the Heatmap above appear to be under '0.1' and therefor could be considered as a 'trend'.

#### 04. SKLearn REGRESSION

Using SKLearn we were able to calculate our Regression Model and store Regression Coefficients into Lists 'regCoef' and 'regInter' and store these online using Pickle Dump.

When connecting again with our stored models and values (using Pickle again) we were able to connect all data with our interface.

Afbeelding met tafel

Automatisch gegenereerde beschrijving