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.

#### Specifications

* Establish if the program will run on different OS platforms – It was not needed to validate on Mac OS. We did validate on Linux and Windows OS.
* Determine the software quality criteria: uptime, privacy – we did not receive any criteria.
* Determine the output that needs to be delivered: web-based or window interface – we have been asked to make a console based interface.

### R&D – learning goals, next to building the application

* Determine the key criteria upon which a choice can be made which tool(s) for data analyses to use, for example Pandas and MatPlotLib.
* Work with KATA’s – we had daily Kata/project updates with the teacher.
* We managed tasks and (sub)activities through Trello.
* Understand how the outcome of the model will be used: The doctor can enter the necessary data, based upon the patients information – see choices for columns in DATA TRANSFORMATION and ANALYSIS, entered via a console application, and see the results: BMI and an indication for “Lifespan”.
* We 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 application as a collection tool, for data stored on an external server, because it allows to have a temporary stored dataset.

### 02. DATA CLEANING

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

Through using SQLite we are able for getting a quick overview of the content of the tables. Next to this we exported the data to a CSV file and imported the data in Pandas. 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. The different types of bad data are:  
• Empty cells  
• Data in wrong format  
• Wrong data  
• Duplicates *(no duplicates found)*

We identified “errors” in 12 rows.

#### 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. Looking at the fact that this is a very small % against the total number of records so we decided to delete these rows.

***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 following 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 corelation 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 therefore could be considered as a 'trend'.

Add language about decision to keep al colums in.

#### 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.

#### 05. INTERFACE

We used below added table and a connection to the data as help to define the buttons in the interface.

Afbeelding met tafel

Automatisch gegenereerde beschrijving

***06. Trello task and time management***

