CrowdHEALTH

Collective Wisdom Driving Public Health Policies

Del. no. – D5.18 Multimodal Forecasting and Causal Techniques: Software Prototype v1

Project Deliverable

This project has received funding from the European Union's Horizon 2020 Programme (H2020-SC1-2016-CNECT) under Grant Agreement No. 727560

D5.18 Multimodal Forecasting and Causal Techniques: Software Prototype v1

Work Package:		WP5
Due Date:		31/12/2017
Submission Date:		17/01/2018
Start Date of Project:		01/03/2017
Duration of Project:		36 Months
Partner Responsible of Deliverable:		Karolinska Institutet
Version:		1.4
Status:	☒ Final ☐ Draft ☐ Ready for internal Review ☐ Task Leader Accepted ☐ WP leader accepted ☒ Project Coordinator accepted	
Author name(s):	Sokratis Nifakos (KI) Anton Gradisek (JSI) Mitja Lustrek (JSI) Thanos Kosmidis (CRA) Christos Panagopoulos (BIO)	
Reviewer(s):	Serge Autexier (DFKI)	Andreas Menyctas (BIO)
Nature:	☐ R – Report ☒ D – Demonstrator	
Dissemination level:	☒ PU – Public ☐ CO – Confidential ☐ RE – Restricted	

REVISION HISTORY			
Version	Date	Author(s)	Changes made
0.1	06/04/2017	Organization name	TOC to share with the participants.
0.2	25/11/2017	KI/JSI	Tables with prototype components
0.3	12/12/2017	KI/JSI	Main components/Interfaces
0.4	18/12/2017	KI/JSI/ULJ	Executive Summary, User Interface
0.5	21/12/2017	KI/JSI	Revision
0.6	22/12/2017	JSI	Figures added
0.7	03/01/2018	KI	Formatting
0.8	04/01/2018	JSI	Formatting
0.9	06/01/2018	JSI	Addressed comments from per review
1.0	07/01/2018	JSI	Fixing the prototype overview
1.1	09/01/2018	KI/JSI/ULJ	Addressed comments from per review
1.2	11/01/2018	KI	Formatting
1.3	12/01/2018	KI	Final version
1.4	17/01/2018	ATOS	Quality review. Submission to EC

List of acronyms

BMI	Body Mass Index
CSV	Comma-separated values
JSI	Jozef Stefan Institute
REST- API	Representational state transfer - Application programming interface
ULJ	University of Ljubljana

Contents

Table of figures / tables

1. Executive Summary

This document aims at describing the Multimodal Forecasting and Causal Techniques module prototype which will form a part of the health analytics layer and will be used to perform several analytical procedures that will aid the policy making process. The initial prototype reflects the work that has been done in the scope of task T5.6 (Multimodal Forecasting and Causal Techniques) between M04-M11 and describes the main components of the prototype as well as the basic technologies used in its development.

Due to limited availability of data from the UCs, the current prototype is based exclusively on the SLOfit use case (ULJ and JSI). The Multimodal Forecasting and Causal Techniques module will eventually consist of two main building blocks: i) Causal analysis, ii) Forecasting. However, at this time only the forecasting module is included in the prototype. As a result, this document not only describes the currently available functionalities, but also gives a high level overview of the future components to be implemented in the next version of the prototype. Both causal analysis and forecasting of individual parameters will represent progress beyond the current state-of-the-art.

As the work to be done in the scope of T5.6 is not yet finalised, updated versions of this document are planned to be delivered during the process of the project (M22, M34) which will include the additional components and functionalities that will be implemented.

2. Prototype overview

2.1. Main components of the prototype

The causal analysis and forecasting module has been developed and tested based on the available data from the SLOfit project provided from University of Ljubljana. One of the main forecasting functionalities in the current version is the forecast of weight and the overall fitness status at the age of 18.

The main components of the prototype v.1 are as follows:

- Data cleaning app: Responsible for tidying the data (fixing typos in names, merging cases where names and surnames are mixed, fixing date of birth). This prototype is used in pre-processing.
- Missing value fixer: Populates the missing values with extrapolated values in order for the models to work (as the models often rely on continuous data and cannot handle missing data well). This prototype is used in pre-processing.
- Forecasting prototypes (3 prototypes): Forecast the weight, individual health parameter, and the overall fitness at the age of 18, based on the previous data.

2.2. Interfaces

At the current stage of the project we have one external interface of the prototype:

- Interface with data store CrowdHEALTH repository: offline interface which is based on CSV files.
- In the next versions, the prototype will interface with the Data Store via the JDBC driver for LeanXcale. It will also provide an REST API for other CrowdHEALTH modules to submit prediction tasks, the results of which will probably be written in the Data Store.

2.3. Baseline technologies and tools

The prototypes or their components are currently written in the following programming languages:

- Java (general-purpose programming language)
- R (programming language and software environment for statistical computing)
- Python (interpreted high-level programming language for general-purpose programming)

3. Source code

3.1. Availability and exploitation

In this part we present the current status and the different functionalities of the prototype, the software used and the execution process. Two of the main components are available on the CrowdHEALTH repository while the following components will be developed and implemented in the next versions of the prototype.

1. Component: **Data cleaning app.**

Functionality: Tidying data.

Software: R

Execution: Run on the data, likely before importing to the CrowdHEALTH platform.

Availability: CrowdHEALTH repository

Current Status: Functional

Details: The app identifies the individuals that are the same person but listed as two and merges them. Fills in the missing data, fixes typos, etc.

2. Component: **Missing Value Fixer**

Functionality: Populate the missing values with extrapolated values in order for the models to work.

Software: Python

Execution: Run on the data, likely before importing to the CrowdHEALTH platform.

Availability: CrowdHEALTH repository

Current Status: Functional

Details: As the models use continuous data while some measurements are occasionally missing, this component uses the rest of the data to populate the missing values with the values, extrapolated from the rest of the data. This is considered the reasonable within our models.

3. Component: **Weight Forecasting**

Functionality: Forecast the weight at the age of 18 based on the previous data.

Software: Python, will either be rewritten in Java or a Java wrapper will be developed

Execution: Select an individual from the database, return the forecast value (one can choose from various methods, such as linear regression or percentile method)

Availability: CrowdHEALTH platform, when implemented

Current Status: Functional

Details: The weight forecasting module uses various approaches to forecast the weight at a chosen year. The input parameters are the available weight and other data up to the current year. The percentile method looks at the general population, it works with the assumption that the individual development follows general trends. For example, if an individual at a chosen age is between 60 and 65 % weight percentile, it is likely that they will be in that interval at the age of 18, thus we use the data for 18-year olds as the basis for forecasting. Linear regression method uses a regression model trained on general population and uses the weights from previous years. This method can be improved when training the model on specific subsets of similar individuals, such as based on the growth spurt. A further improvement considers other types of available data as well. These improved methods are beyond the current state-of-the-art.

3.2. User Interface

The user interface is still in the concept stage as the prototypes only exist at the code level.

The user chooses the SLOfit parameter from a menu, for example weight or BMI. Next, the user chooses the available data for an individual, for example from 8-14 year, and the model that is to be used for forecasting. The output is the forecast weight or BMI at the age of 18.

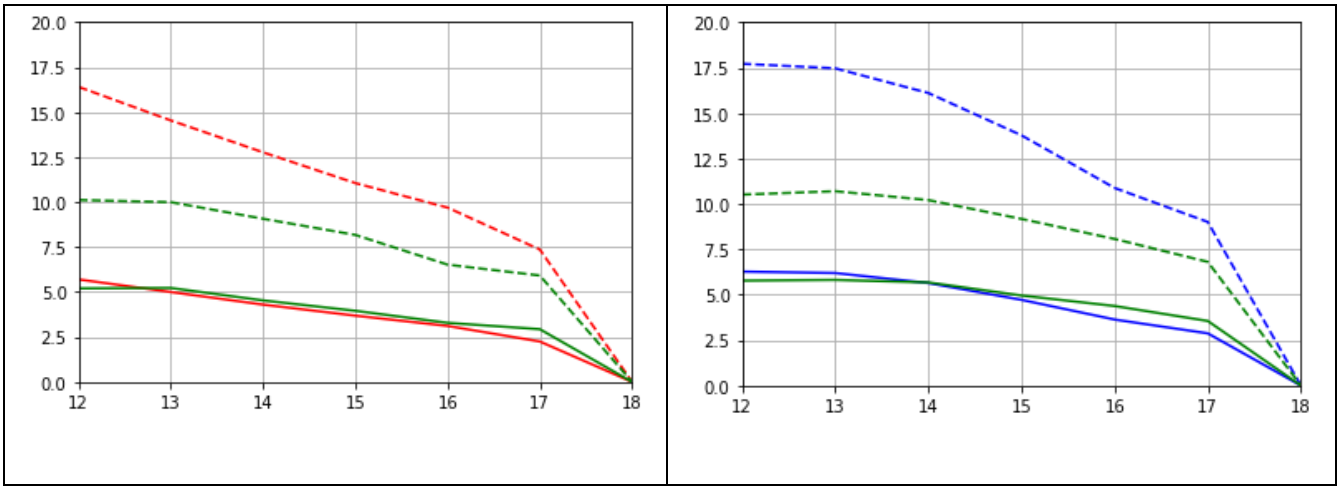

Figure 1 Weight forecast interface.

At the age of 18 for girls (left) and for boys (right), based on the data up to the year marked on the x-axis. The y-axis represents the error of the prediction. Solid green lines represent the mean error linear regression model while solid red or blue correspond to the percentile method. The dashed lines correspond to the standard deviation, in the same manner.