Data Science

Explore Series: 1 (18th April 2020)

Practical 1

The aim of this practical is to go through the data science life cycle with the objective of determining what variables are significant in determining whether a patient has severe malaria or not.

1. Data acquisition

There are 3 sets of data

1. **biodata.csv**

35 observations, 7 variables.

Biodata obtained from 35 patients diagnosed to have severe pneumonia.

	Variable	Description
1.	id	Patient unique ID
2.	date_adm	Date of admission
3.	hosp_id	Hospital ID
4.	sex	The patient's gender
5.	age_years	Documented age in years
6.	Age_mths2	Total age in months. Calculated from the age in years and age in months.
7.	village_name	Village of residence.

Note: Age in months in missing but this can be reverse calculated from age_years and Age_mths2

2. **clinical_examination.csv**

35 observations, 19 variables

Documentation of clinical symptoms of the 35 patients, the investigations ordered and the outcome.

	Variable	Description
1.	id	Patient unique identifier
2.	fever	Does the patient have fever?
3.	fever_dur	Duration of fever if the patient has fever
4.	temp	Body temperature of patient
5.	BS_or_RDT_ordered	Was a malaria blood slide or rapid diagnostic test ordered?
6.	BS_or_RDT_pos	Was the result of the blood slide or RDT positive?
7.	convulsions	Has the patient had any convulsion?
8.	convulsions_no	How many convulsions if they have had any?
9.	hb1_order	Was a haemoglobin test ordered
10.	hb1_result	Haemoglobin test results in g/dl

11.	trans_order	Was a blood transfusion ordered?
12.	pallor	How pale the patient is
13.	pallor_severe	Whether patient has severe pallor or not
14.	indrawing	Inward movement of lower chest when patient breathes in. (Sign of respiratory distress)
15.	acidotic_breathing	Deep and labored breathing (often due to metabolic acidosis)
16.	avpu	Patient’s level of consciousness. (Alert, Verbal responsive, Pain responsive, Unresponsive)
17.	Blantyre_score	Score of 0 to 5. Combines score of motor response, verbal response and eye movement in patient.
18.	days_adm	Duration of admission in days
19.	outcome	Was the patient Dead or Alive on discharge?

3. village_codes.csv

25 observations, 2 variables

Village names and their respective codes

	Variable	Description
1.	village	Village name
2.	code	Respective village code

Data Understanding and Preparation

Evaluate the different data sets by checking their dimensions and data types. Make any data type conversions necessary.

Aggregate the biodata and clinical_examination data and add the village codes for each record with respect to the village name

Check for duplicate if any.

Assess the data for these qualities; Validity, consistency, completeness and uniformity.

Exploratory Data Analysis

Conduct specific exploratory analysis for the numeric and the categorical variables.