**Study of HIV epidemiology indicators and treatments for children and adolescents.**

* TEAM COMPONENTS:

Kristjan Pekk, Pärt Alango and Yaiza Rubio Chavida

* TASK 1 (Setting up):

Project repository:<https://github.com/Yaiza0706/DataScienceProject>

* TASK 2 (Business understanding):
* Identifying your business goals
  + Background
  + Business goals
  + Business success criteria
* Assessing your situation
  + Inventory of resources
  + Requirements, assumptions, and constraints
  + Risks and contingencies
  + Terminology
  + Costs and benefits
* Defining your data-mining goals
  + Data-mining goals
  + Data-mining success criteria

**Goal 1:** Analysis of our first dataset to see how the number of children with HIV changes depending on different features like age, gender, continent, year...

**Goal 2:** Analysis of the treatments that were made in the last years depending also on different features.

**Goal 3:** Compare the values obtained before (number of positives in HIV and number of treatments) and make an analysis with the obtained results (using different methods covered in the course (we will look for the best one for each conclusion).

* TASK 3 (Data understanding):

After defining our goals, we looked for the data to start with gathering, first phase of Data understanding. For this first method,we entered the Unicef web (<https://data.unicef.org/resources/dataset/>) and we found many datasets available. After seeing all the different topics, we went into VIH section, one that was more similar to our goals (( <https://data.unicef.org/resources/dataset/hiv-aids-statistical-tables/>). In this web page, we analyze which datasets of the given would be useful for us . After choosing two of them, we verified that the data we wanted existed and we identified them as databases. We downloaded them and uploaded to Github (platform we are using to share between us our progress). We also discuss about the source to analyze our data and we finally decided Jupyter Notebook using Python 3 (method used in our different homework and studied during the Introduction to Data Science course). Moreover, our datasets looked compatible to it.

After this previous study, we looked at our data and started the second phase: describing it. The datasets we obtained were two:

**Dataset 1 (11290 KB):** It has information about HIV indicators from children and adolescents, depending on the continent, sex, age(0-19) and the year (1990- 2019).This dataset includes six key indicators for monitoring the HIV response for children and adolescents. Indicators are presented for the years 1990-2019, and disaggregated by age, sex and country where available. It also contains the datasource . It is formed by 240430 rows.

**Dataset 2 (177 KB):** It contains the coverage of antiretroviral treatment (ART) among children(0-14 years) with HIV. This dataset includes two key indicators for monitoring treatment programmes for children living with HIV. Indicators are presented for the years 2010-2019 and countries where available. It also contains the datasource . It is formed by 2118 rows.

After seeing what information we had, we passed to the third phase: exploring the data. After a deep analysis, we discovered that we would have to analyze the data according to the different indicators, some of them are number of children, another ones were represented by percentage… So we realized we couldn´t mix the values to get good results. We also saw that some columns were not useful for our analysis (like the data source), so we will eliminate them. Finally, we will reduce first dataset to study years from 2010 to 2019, because we have no data about previous years in the second dataset, so we can´t compare them.

Finally we did the last step of data understanding: verifying data quality. We transformed our datasets(xlsx) into csv files. We read them using pandas with ‘,’ separator. We had some problems because some of the numbers where giving with dot as a decimal and other ones with a comma. To solve this issue, we change separators from ‘,’ into ‘;’. After that, we got better results, but some rows where still read as a singe column . Then, we added different parameters to pd.read\_csv, but some mistakes appeared. We finally solved these problems using “ encoding='latin-1' ”. With this last change, we had our data prepared to start working on it.

* TASK 4 (Planning your project):

\*\**disclaimer: tasks’ time estimation may be extremely inaccurate (everything comes down to how well can students put their previously acquired knowledge to practice)*

***[Preparation]***

**Task 0:** get comfortable with different visualization methods (i.e plots) | *est. time: 3-4h*

**Task 0.1:** find best fields from both datasets to analyze       *| est.time: <1h*

***[First dataset]***

**Task 1:** find correlations/differences amongst the number of children with HIV and their age, gender       *| est.time: 3h*

**Task 2:** connect results from task 1 with continents (data origin)       *| est.time: 1-2h*

**Task 3:** visualize found data       *| est.time: 3h*

***[Second dataset]***

**Task 1:** connect an find correlations between the percentage of HIV-diagnosed children receiving treatment and their country/region       *| est.time: 3h*

**Task 2:** connect results from **task 1** with results from **first dataset**     *| est.time: 2h*

**Task 3:** visualize found data       *| est.time: 3-4h*

***[Finally]***

**Task 1:** find additional correlations from 1st dataset (i.e using ‘mother to child transmission rate) *| est.time: 1-9h* **Task 2:** optimize visualization       *| est.time: 2-3h* **Task 3:** optimize code       *| est.time: 3h* **Task 4:** conclusion/report       *| est.time: 7-8h*

***[Used methods]***

All work and analysis will be conducted using Jupyter Notebook and Excel. Languages used will be Python and R.

We cannot yet specify all methods that will be used, but used methods will be reported and their usage described appropriately.