REPORT

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 our business goals

- Background

The main driving force behind our project is to analyze the number of people with HIV in relation to different features in the data available. The features might be age, gender, continent, year etc. In doing this, we hope to better our understanding of conducting a proper scientific research while abiding to the CRISP-DM process model.

- Business goal

The goal of our project is to contribute to the overall understanding and spread of this virus and the treatment efficiency in treatments conducted in the time period of 2010-2019.

- Business success criteria

The afore stated goals can be considered accomplished if project results indicate new, previously undiscovered correlations between features and a person either having HIV or not.

Assessing our situation

- Inventory of resources

Resources available for this project can be divided into 3 categories: people, technical resources and the actual data. Research will be conducted by Kristjan Pekk, Arean-Pärt Alango and Yaiza Rubio Chavida.

While somewhat inexperienced in the field of epidemiology, the researchers possess basic knowledge about working with data, including the use of necessary software, using methods for descriptive data analysis and carrying out statistical tests and interpreting the results.

In regards to technical resources, every group member has a laptop and the necessary software is already installed. The group uses GitHub as a central repository to store the project related code and Jupyter Notebook with Python 3 to explore the data, make visualizations and conduct the majority of our research with. The data that we have gathered so far from UNICEF's dataset archives will suffice to get started.

- Requirements, assumptions and constraints

The project deadline is December 14 2020, by which we will have finished research and drawn conclusions from the results. All data and software necessary to complete the project is open source.

- Risks and contingencies

The main risk regarding project completion is if one or more team members lose focus on what to do next, in which case the other team members will help them back on track.

- Terminology

Some domain specific vocabulary will be used in the documentation and presentation of this project, so to help all parties of interest have an equal understanding of our work, some of the more important terms will be defined.

  - Dataset – a collection of related information that is composed of separate elements but can be manipulated as a unit by a computer.
  - Feature – in machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed.
  - Data mining – the process of discovering patterns in large datasets involving methods of machine learning, statistics and different sources of data.
  - Machine learning – the study of computer algorithms that improve automatically through experience.
  - Data visualization – means of displaying information e.g. data in the form of charts, diagrams and pictures.
  - Correlation - a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)

Defining our data-mining goals

- Data-mining goals

To clean the data and find correlations between people's conditions and their features in the datasets and also to find out how many positive cases are getting treatment.
For example, these results could then be visualized by heat maps, where areas with more HIV cases are more highlighted.

- Data-mining success criteria

Map all the correlations using different features against the patients' conditions.
Determine the success rate of treatment among known treatment cases and the overall percentage of people with HIV who have turned to getting treatment.

- ## 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 programs for children living with HIV. Indicators are presented for the years 2010-2019 and countries where available. It also contains the data source. 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, and other 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 single 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)*

[Data 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 Modeling]
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 Modeling]
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*

[Evaluation and deployment]

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.