TASK II - Business Understanding

With the global standard of living increasing every year and things such as high speed transportation and internet becoming more and more available peoples every day lives will no doubt have changed in the passed 50 years. But just how different are peoples lives in different countries ?

The goal of this project is to collect and present population statistics of countries in order to find out, if an individuals country of inhabitants can be predicted using machine learning. The main question the project will try to answer is, to what degree of accuracy can an individuals country of inhabitants be predicted using machine learning?

The first criteria for success is a dataset that can be displayed in a human readable form. To get an overview over the collected data and to make sure that data gathered from national statistical offices are equal. Each nation has their own systems for education, healthcare ,etc - thus a common set of variables needs to be established. The second criteria is that the dataset can be used for machine learning. When in the first criteria it was to find similar characteristics between each country, then here the goal is prepare the data for machine learning. And finally in order to succeed the project must provide an answer to the research question.

The project will be carried out by 2 people, who will use their personal computers to collect, clean and analyse the data. Git version control software and GitHub online repository will be used for collaboration between team members. The main method for communication between teammates will be Slack. Google Colab will be used for the coding and data visualization part of the project. The project will have access to various public online sources of data as well as information that can be found through academic search engines available to the University of Tartu students.

The project needs to be compleated by the 16. of december where it will be present as a poster. Also the project plan has to be in place and presented by the 2. of december. In order to carry out the project it is assumed that the required data remains available.

The progress of the project can be hindered by a number of factors. If a team member should lose access to the internet or their personal computer has a hardware issue, they will be able to complete their work at the university. If the repository at GitHub should become unavailable the project will be recovered from a team members local machine or Google Drive and a new repository will be created at Bitbucket or GitLab. In case of communication problems teammates can communicate via email and if problems should occur with Google Colab, then the coding part of the project can be done in jupyter notebook.

The main term used throughout the project will be population characteristics which in this case refers to characteristics that describe the population living in one country i.e. median age, percentage of population working in the industrial sector, most popular sport, average hours spent watching television and so on. When speaking about the characteristics the term standardised means that the individual ways countries define levels of education and things of that nature have been changed to a common set of variables.

The benefits of the project are purely academic. No monetary value will be generated through this study, however the results of the data gathered could help businesses design and market products that have the largest international appeal. The project is also carried out by students as part of introductory course to data science in the University of Tartu, so all the costs associated with the project, i.e. hardware costs and access to the internet, will be covered by the team, meaning the costs and the benefits of the project are balanced.

The goal of the data mining portion of the project is to create one standardised dataset that contains the population characteristics of nations, which can be visualized on a poster and used as teaching data for a machine learning algorithm. The secondary goal is to create smaller datasets that display interesting statistics that may not be as useful for the main goal.

The data mining portion can be considered successful if a main dataset can be created that includes at least 50 countries and 20 different characteristics that have been standardised.

TASK III

In order to meet the data mining goals the project requires population data about a large number of countries. What exact characteristics will be used in the final version of the project will be largely determined by what is available for all countries. Data that can be accessed as a CSV file is preferable, however manual copying of data may be the only option in some cases.

In cases of a countries population statistics being unavailable, outdated or a linguistic barrier preventing the use of the data, the data can be left out because the goal of the project is not to have 100 coverage of every country on Earth, but to have enough data so that some conclusions can be made.

The main criteria for the data are that it can be found in english, it is not older than 5 years and the methodology for the collection of data can be verified. In the case of sources with conflicting information, data gathered by international organizations will be preferred over data gather by local organizations. Datasets that can not be read or that have over 10 % of the data missing will be discarded. The scope of the project may narrow down to only countries in the European Union, if data on the minimal amount of common characteristics can not be found for nations around the globe.

Since our goal is to mix and match different datasets it is difficult to give a simple overview. Generally the data we are looking at are in CSV format and they contain a lot of information that will be used in our project but some columns of values be extracted and added to our larger dataset of characteristics. The field used for the project will contain percentage values or numeric values i.e. average amount of hours spent by a member of the population doing something.

Data gathered from international statistical databases (UN, EU) are already clean, balanced and related to each other. For example, while comparing married men and women then the categorization to age group has already done for us. But usually these databases give data for general and simplified answers, but to get much more detailed custom characteristics, we must look for countries national statistical agencies. While collecting our data piece by piece, then some custom categorization work is needed.

Our work accuracy hugely depends on the amount of features we can have for each country and also by the correctness of the data. Some of the data is gathered with census or survey and this can give invalid outputs since one is not as accurate or reliable than other. Similar problem arises when comparing data taken from national statistical offices, where one countries questionnaires are more detailed and accurate than others. Final problem that have occurred with quality is when collecting data from various indexes that are useful to us. Some of them do not show in detail what and from where their data is from. This kind of data is not trustworthy and therefore we do not use them.

Project Plan

Tasks

1. Gather national characteristics data from different sources. (Georg Reintam 8h, Taavet Simo 8h)

   The main tool to find the data will be google and once the data is found we will use the pandas library for python to store the data in a way it is easy to work with. The part that makes this task time consuming is finding the characteristics for which many countries have reliable data. When determining a new datasets compatibility with our dataset the first step is to make sure that the country names match with our dataset. The second step is to add the new characteristics to our dataset. Finally we check if the new characteristics columns contain values for each country we have listed in the first column, if there are many empty slots the characteristic column will be removed, if only a few countries are missing then we determine if those countries can be removed from the dataset.

2. Cleaning and evaluating the data. Some additional data gathering may also be required. (Georg Reintam 7h ,Taavet Simo 7h)

   In this step we evaluate the created dataset as a whole using pandas and numpy. Values that appear out of place will be double checked and possibly removed. If the dataset becomes too thin, additional data will be gathered.

3. Analysing and visualizing the cleaned data (Georg Reintam 5h, Taavet Simo 5h)

   We use pandas library and possibly other visualization tools to present the information gathered in our dataset. We will also try to high light interesting information i.e. what percéntage of a nations population watches football or what is the big mac index of each country.

4. Implementing machine learning to find an answer to the research question (Georg Reintam 6h, Taavet Simo 6h)

   In this part we use sklearn to implement a machine learning algorithm that can make a prediction on where someone might live based on our dataset and characteristic values given as test data. We also had the idea of using tableau to make the output more visually pleasing
5. Making conclusions and designing the poster (Georg Reintam 4h, Taavet Simo 4h)

   In this task we will design the poster. How we do this is yet to be decided.