Link to github: https://github.com/kare22/IDSCI2020EMISSION

TASK 2. BUSINESS UNDERSTANDING

Backround

CO2 emission proves to be a rising problem in worlds economy and earths welfare. It is believed that global average temperatures have increased by more than 1℃ since pre-industrial times. Industrialisation and human activity are the main reasons for CO2 emission which has lead to the climate change.
Due time many regions are prone to excessive CO2 emissions and our aim is to find out if/how this is connected to regional poverty, education and world development in general.

Business goals

Our business goal is to make a model based on the datasets, which we can use to interpret the CO2 emissions all around the world, therefore help areas in need of improving education, access to food, medicine, first aid etc. With our analysis we can determine which areas are beneficial to renewable energy production and which are not.

Business success criteria

If we can gain insight on how the emission works over the years regarding economical factors we can find common patterns and therefore build prediction models. Since our data is based on countries all over the world and the time scale fits the timeframe when the major climate changes(CO2 emission) are happening, we believe our Business success criteria is quite high.

Assessing the situation

Inventory of resources

Our data is collected from a website called databank.worldbank.org. From there we are using four datasets on poverty, education, world development and CO2 emissions. The timescale is between 1960 - 2019 and we will feature 264 countries.

Requirements, assumptions, and constraints

As a group our end goal is to present a poster with different analytical ways in processing the given data. For each team member there is a given topic where we gather the data and put it together to analyse with different ways learned from this course. If there is missing data from a certain year

or from a specific country we will narrow our research. Overall we assume that CO2 emission is more severe in developed countries.

Risks and contingencies

Our work strategy is based on a work plan. We know the deadline, so each week until the presentation, we set goals to finish (gathering data, analysing it etc.) the tasks at hand.

Terminology

* **Data pre-processing** - consists of data preparation and data characterisation. First means that we make the data presentable to are use (includes standardising data formats, enriching source data, and/or removing outliers) and then we summarise the general characteristics or features of a target class of data
* **Ordinal data** - Ordinal data is a kind of categorical data with a set order or scale to it
* **Continuous data**- Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.
* **Normalising the features -** to scale a variable to have a values between 0 and 1
* **Data Mining** - is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems
* **Pattern Evolution** - is defined as as identifying strictly increasing patterns representing knowledge based on given measures
* **Cross-validation -** is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set
* **Missing data** - Data values can be missing because they were not measured, not answered, were unknown or were lost

Costs and benefits

Due to the fact our goal is to find the optimal solution for reducing CO2 emissions, we don't yet know, what it is, therefore we can't give an estimate on its cost. Although we can clearly state the benefits of decreasing CO2 emissions… As damaging lives and nature are usually irreversible, it is adamant that we don't pollute our environment. Also if the CO2 level in our atmosphere rises, it will cause climate change through the greenhouse effect, which will have a grate impact on our lives and nature in general.

Further more when dealing with CAB, it is not very efficient for environmental topics, because we can't put a price tag on lives or welfare. Thus the real benefits may not seems that important now, but their cost in the long run when kept unfixed is usually greater.

All-in-all we believe that this is one of the important topics of our lifetime and needs to be addressed, thus the research topic.

Defining your data-mining goals

Data-mining goals

Our goal is to make a combined dataset that has country/year as row values and different series values in the columns. By doing so we will use different data minig methods to analyse the data and find common patterns.

Data-mining success criteria

Our Data-mining success depends on data minig methods accuracy. We aim to find common patterns with pattern tracking on our data, and how much missing data do we have in our dataset. Overall we try do use as many learned techniques (Neural Network, Anomaly Analysis, Clustering Analysis etc.) to find best resaults for our goals.

Task 3. Data understanding

- Gathering data

  For our project ’Factors that effect CO2 emissions’ our team collects data from World Bank website databank.worldbank.org.

  - Outline data requirements

    We prefer that the data we collect is in csv-format and our source satsifies that condition. We try to collect data about 264 countries and the time range is from 1960 till 2019.

  - Verify data availability

    If data turn out to be incomplete for certain countries or for a certain periood of time, we will narrow down the country choices and the periood.

  - Define selection criteria

    We will collect data from the website databank.worldbank.org. Our goal is to find features that correlate best with CO2 emissions on diferent countries through years and find the best models within the set of features to predict CO2 emissions.

- Describing data

We collect data about CO2 emissions and three main topics: world develepment indicators, education statistics, poverty and equity. For each topic we search data from about 5-7 series.

For example about World Develepment Indicators we suppose that CO2 emission is correlated with 1) GDP per capita 2) households final expenditure 3) population 4) industry (% of GDP) 5) electricity production from oil, gas and coal sources 6) goods export.

Our data has three dimensions: country, year and values from diferent series. We put values to the columns (approximately 15-21 + 1 columns) and country/year together to the rows (maximum ca 15500 rows).

Most of the data will be numerical - continuos data. We will have categorical - ordinal data, also (for example data about education). We will use different methods to preprocess data. For example different education levels we will put on the scale: 0 means no education, 5 means high education and 5 is better than other values till 0. Numeric continiuos data we will categorize. We will normalize all features so that we place values between 0 and 1 separately for each feature.

- Exploring data

  Having read the data more closely we conclude that the main probleem is missing data. Many countries has missing data for following reasons: countries has a short or fragmentary history – for example countries of the former Yugoslavia or Soviet Union; countries are or have been at war for some years. If country has no data for longer periood we will exclude the country. If there will be missing data for some years only, we will replace gaps with averages.

- Verifying data quality

  It seems that data quality from World Bank is quite good – missing data is marked in the same way only (no zeros, Nan’s and spaces together). It’s possible to layout the data in the necessary way (choose columns and rows) on the webpage, already. Exporting data in csv-format is also possible.

Plan

1. Collecting test data – each member will select 5-7 features from one set (there are 3 in total) and create a subset
    - **a.** Education Statistics – Karel **(2h)**
    - **b.** World Development Indicators – Andres **(2h)**
    - **c.** Poverty and Equity – Merilin **(2h)**
2. Generating the test set
    a. Join all features with the label set **(2h) TBD**
        i. Countries as rows and year x feature as columns
    b. Choose which periods should be left out or filled with the arithmetic mean **(2h) TBD**
    c. Splitting the set into train (70% of data) and test (30% of data), for this we randomly choose 3 years per decade, because the rise of emissions is linear, we don't want same indexed years per decade. **(2h) TBD**
3. Finding the best methods for estimating the growth of CO2 emissions in a country.
    - **a.** Using machine learning models and natural language processing (afterwards testing two or more most effective methods together to try and produce an even better result) **21h for everyone**
4. Testing the results with a custom use case (created by hand) to be sure of the results.
    - **a.** Create the use case **(3h) TBD**
    b. Test all satisfactory methods with the use case **(5h) TBD**
5. Finding the logical assumptions from the results and concluding ways to improve a countries CO2 emissions**. (7h) TBD**
6. Creating the poster **2h for everyone**