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](http://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](https://en.wikipedia.org/wiki/Data_set) involving methods at the intersection of [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [statistics](https://en.wikipedia.org/wiki/Statistics) and [database systems](https://en.wikipedia.org/wiki/Database_system)
* **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](https://en.wikipedia.org/wiki/Model_validation) techniques for assessing how the results of a [statistical](https://en.wikipedia.org/wiki/Statistics) analysis will [generalize](https://en.wikipedia.org/wiki/Generalization_error) 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.