**Data Science Project Protocol**

*Author(s):*

*Eti Mayer*

# Introduction

Here you have to give some known facts about the field you will work on.

Try to focus on the problems that are most common and then state the goals of the project.

You can try to answer to the following question:

* Which questions do we want to answer?
* What is known about the problem?
* How we define the outcome(s)?
* What is known to influence the outcome?
* Does we have any possible new knowledge that has not been in use before?

This part must be between half to one and half page.

F

acing the great challenge of climate changing becomes an essential to humankind, as the effects on climate increasingly visible. Floods, storms, fires and droughts become stronger and frequent. Glaciers leave their natural habitat and melt. Global ecosystems are changing rapidly. The growth of CO2 emissions is a major contributing factor to those changes. In order to reduce CO2 emissions, states must play an active role in the reducing CO2 emissions by setting targets to reduce the emissions by 20% compared to 1990 levels, increasing renewable energy (ex: solar, wind) market share by 20%, and a 20% increase in energy efficiency, policy known as 20/20/20 targets[[1]](#footnote-1).

Renewable energy is critical component in achieving those goals, unfortunately the high deployment and installation costs inhibitive a significant improvement in this factor. Over USD 1.7 trillion would be needed by 2030 to implement renewable energy targets contained in Nationally Determined Contributions (NDC)s worldwide[[2]](#footnote-2)

In this paper I want to predict the Δ between current CO2 emissions (CO2-0) to the CO2 emissions targets in 2030 (CO2-1), depending on historical data from 1990 until 2015 from the World Bank Open Data website[[3]](#footnote-3). I want to focus on renewable energy and its optimal level of using by using predictive tools from machine learning algorithms.

For this I have data collected by the World Bank including country name, country code, and various indictors between the years 1960-2015 by country including: GDP, CO2 emissions (kt), renewable energy usage etc.

# Methodology (Project design)

## Data

Here you have to describe how do you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss the limitations of the model, when it can be used and when not.

1. <https://www.recs.org/glossary/european-20-20-20-targets> [↑](#footnote-ref-1)
2. <https://www.irena.org/publications/2017/Nov/Untapped-potential-for-climate-action-NDC> [↑](#footnote-ref-2)
3. <https://data.worldbank.org/> [↑](#footnote-ref-3)