About

This document will give a short insight about the project we have been given and how we plan to realise the outcome of the project.

Air Quality

Green Team



Viona Bodlaender Heather van Diemen Luuk Bilsen Kirill Smirnov

# Preface

Contents

[Preface 1](#_Toc127185973)

[1 Phase 1: Proposal 3](#_Toc127185974)

[1.1 Stakeholders and domain understanding 3](#_Toc127185975)

[1.2 Possible Restrictions 3](#_Toc127185976)

[1.2.1 What limitations do we have in our project? 3](#_Toc127185977)

[1.3 Data Sourcing 3](#_Toc127185978)

[1.4 Analytic Approach 3](#_Toc127185979)

[2 Phase 2: Provisioning 4](#_Toc127185980)

[2.1 Data Requirements 4](#_Toc127185981)

[2.1.1 Proposal, Domain Understanding. 4](#_Toc127185982)

[2.1.2 Find the Stakeholders. 4](#_Toc127185983)

[2.1.3 Identify required data elements. 4](#_Toc127185984)

[2.1.4 Identify candidate data sources. 4](#_Toc127185985)

[2.2 Data Collection 5](#_Toc127185986)

[2.2.1 Determine what information you want to collect. 5](#_Toc127185987)

[2.2.2 Where to store your retrieved data? 5](#_Toc127185988)

[2.2.3 Use a (traceable) version and naming system. 5](#_Toc127185989)

[2.2.4 Determine how and how often you want to retrieve the data/ 5](#_Toc127185990)

[3 Phase 3: Predictions 6](#_Toc127185991)

[3.1 Preprocessing 6](#_Toc127185992)

[3.1.1 Feature Selection 6](#_Toc127185993)

[3.1.2 Scaling 6](#_Toc127185994)

[3.1.3 Bias/variance trade-off 6](#_Toc127185995)

[3.1.4 Hyperparameter tuning 6](#_Toc127185996)

[3.2 Modelling 7](#_Toc127185997)

[3.3 Evaluation 7](#_Toc127185998)

[4 Phase 4: Delivery 8](#_Toc127185999)

# Phase 1: Proposal

The proposal covers 4 components:

* Domain Understanding (can be found in appendix A and B)
* Possible restrictions
* Data Sourcing
* Analytic Approach

## Stakeholders and domain understanding

*To improve the daily practices of domain experts, you need to know as much as possible about the specific domain. Thus, a thorough exploration and understanding of the domain in which you work, gained through different types of research, is always the starting point of any AI project.*

## Possible Restrictions

### What limitations do we have in our project?

In this project we have some restrictions that we must consider before we start working on it. Here are a few we know of:

## Data Sourcing

*This section describes what kind of data we have and how we got this data.*

## Analytic Approach

*The next step in writing a proposal is to define a specific goal and approach for the project.*

# Phase 2: Provisioning

Provisioning covers 4 components:

* Data Requirements
* Data Collection
* Data Understanding
* Data Preparation

## Data Requirements

The essence of Data Requirements is to understand that a stakeholder is storing this data for a specific reason or purpose

### Proposal, Domain Understanding.

*Make a list of the domains and describe the relation between them.*

### Find the Stakeholders.

*From whom is the storage of data beneficial in the domain you are researching? Try to approach this creatively. Don’t look only inside your research domains but also in adjacent ones. There is a possibility you’ll find something unexpected.*

### Identify required data elements.

*Figure out the facts and dimensions you require to answer your research. Multiple facts can only be joined through shared dimensions, there are specific procedures you must obey to make these joins work!*

### Identify candidate data sources.

*Combine all the facts and then document each potential data source.*

## Data Collection

Data collection is the process of gathering information from an established system. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posted.

### Determine what information you want to collect.

*Set a time frame you want to look for. Try to be as specific as you can. This is a premise you can use for looking at the right location*

### Where to store your retrieved data?

*Where are we going to store out data. Think about 24/7 availability of the data if you work in a group.*

### Use a (traceable) version and naming system.

*Because you need to refer to the dataset frequently, you must think of an easy identifiable naming system.*

### Determine how and how often you want to retrieve the data/

*This depends highly on the quantity of data or load and update strategy of the source.*

# Phase 3: Predictions

Predictions covers 3 components:

* Preprocessing
* Modelling
* Evaluation

## Preprocessing

At this stage, you have a clean dataset, and you also have a rough idea of what machine learning techniques you're going to apply. The next step is then, to determine which algorithms you are going to use and check their documentation for whatever preprocessing you need to apply.

### Feature Selection

*Feature selection is a step that selects only the useful features (as in, those that do explain the variance in the target variable) and eliminates all other features.*

### Scaling

*Scaling is a process that transforms the values of your features against the same scale, as to make sure that no feature weighs more simply because the values that it uses are higher in respect to the values of another feature.*

### Bias/variance trade-off

*The Bias-variance tradeoff is a property of a model that we have to measure during model evaluation process.*

### Hyperparameter tuning

*A machine learning model has two types of parameters, (internal) parameters that are tuned during training and hyperparameters that can be specified (manually) for better performance when we create an instance from a model.*

## Modelling

The next step is to start building a model (or models) that is based on the delivery of phase 1 (your project proposal) and the outcome of phase 2 (a cleaned dataset). Ultimately you most likely must select and apply various modelling techniques, but in the first iteration you are probably best off choosing a quick algorithm, like nearest neighbors, to get some early impressions.

## Evaluation

At this step you are going to evaluate the built model(s) in the Modelling step of this phase. You will be using various model evaluation (quality and performance) techniques. At the end of this step you will either be confident with the evaluated model(s) that will be used in the Deployment step of the Delivery phase or you will have to improve the quality and performance in the Modelling step based on the evaluation metrics results.

# Phase 4: Delivery

**Appendix A: Stakeholders**

**Appendix B: Domain Understanding**