XAI - Final Project

Professor: Carlos Andres Peña
Assistant: Shabnam Ataee

June 2021

Data Set

Data Set

- Description. The goal of this project is to verify the impact of pesticides and soil care products on the health and quality of vineyard soils in Switzerland.
- Data Set. The following vineyard soil samples are available:
  - *104 soil samples in 2015*
  - *73 soil samples in 2016*
- For each soil sample, there exists
  - *2683 features as input variables*
    - Available in CSV files *'ProtistAmpliconSequenceVariants_ASV_2015.csv' & 'ProtistAmpliconSequenceVariants_ASV_2016.csv'*
  - *'Cu_mg_kg' as target (output variable)*
    - Available as a column in CSV files *'Env_2015.csv' & 'Env_2016.csv'*
  - The first column in all CSV files is considered as *index*.

- Note. *The provided data is confidential.*

Data Set (Cont.)

- Train Set. We consider data set in 2015 as train set.
- Test Set. We consider data set in 2016 as test set.

- The values of target variable *'Cu_mg_kg'* are float numbers. This is a *regression problem by default*. In this project, we ask you to convert it into a *classification problem with 4 different classes* in such a way that the values of *'Cu_mg_kg'* in the train set are *equally distributed* between these four classes.

Problem

Overall Problem

Question 1. What are *the most relevant features* (Feature Selection) in this classification problem?

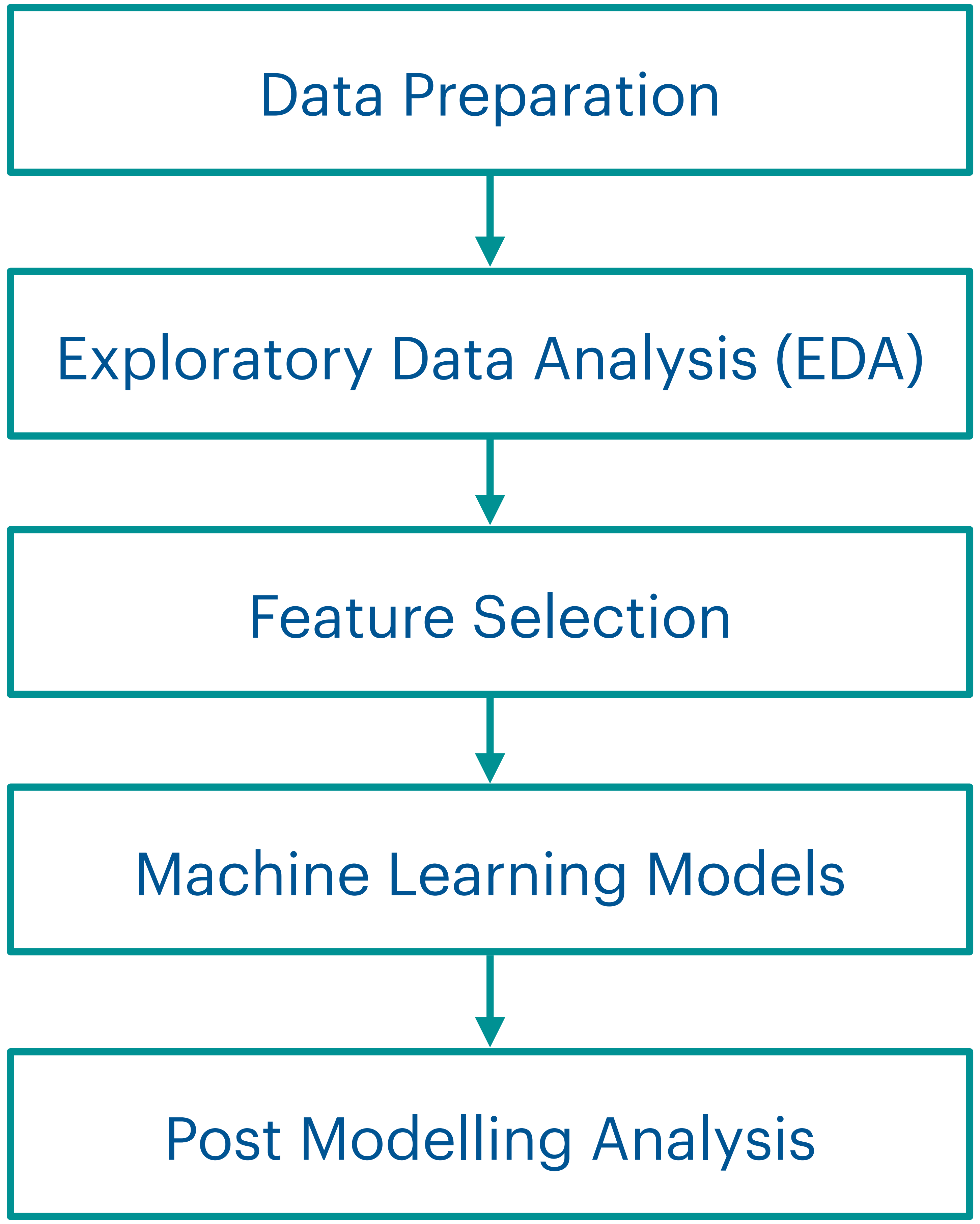
Question 2. How could we predict *class of 'Cu_mg_kg'* in *vineyard soil samples* in 2016 using *interpretability-oriented machine learning (ML)* models?

- Note. This problem is a *classification problem*.

Pipeline

Pipeline

- Follow the proposed pipeline to solve this classification problem.



Data Preparation

Exploratory Data Analysis (EDA)

Feature Selection

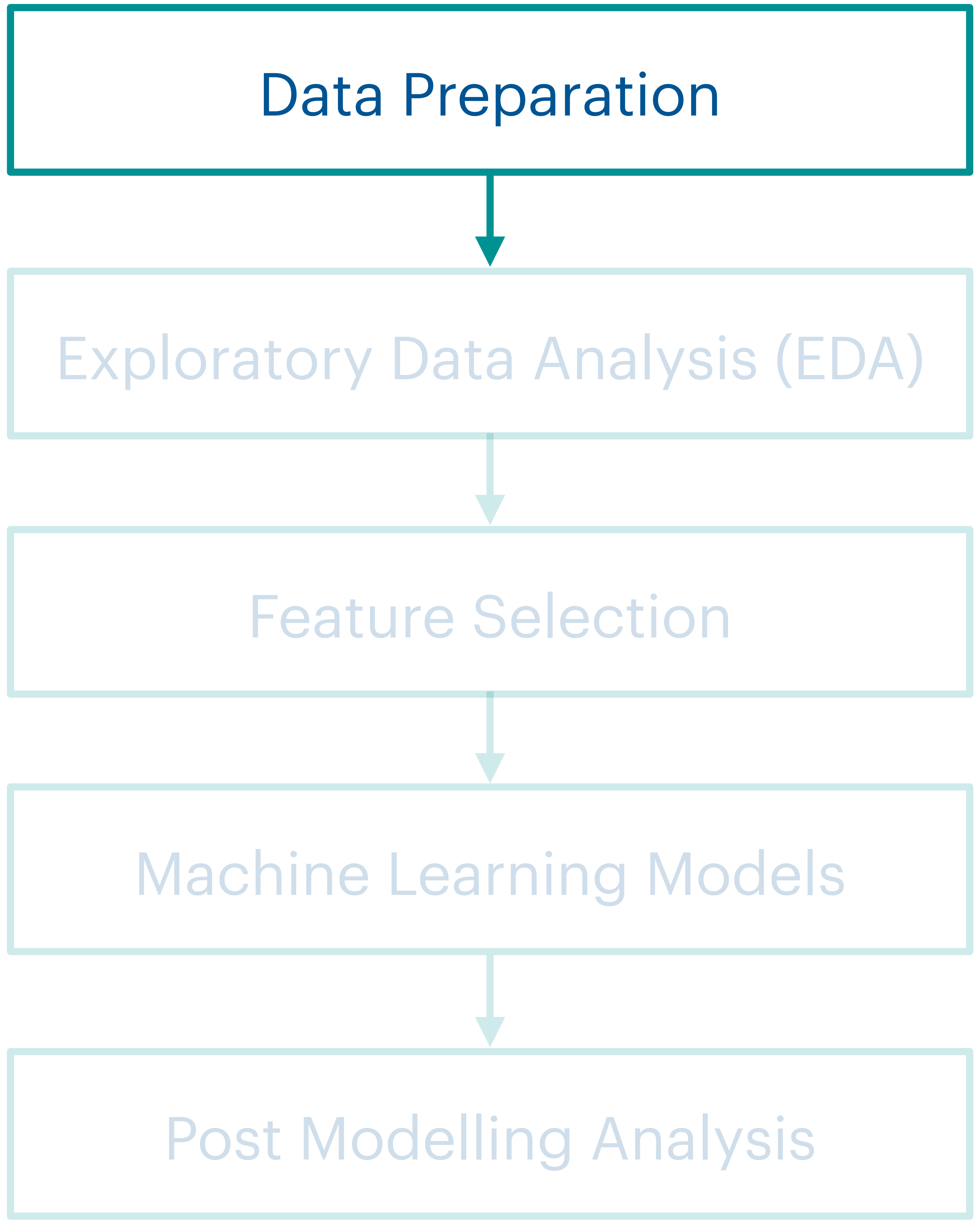
Machine Learning Models

Post Modelling Analysis

Data Preparation

Data Preparation

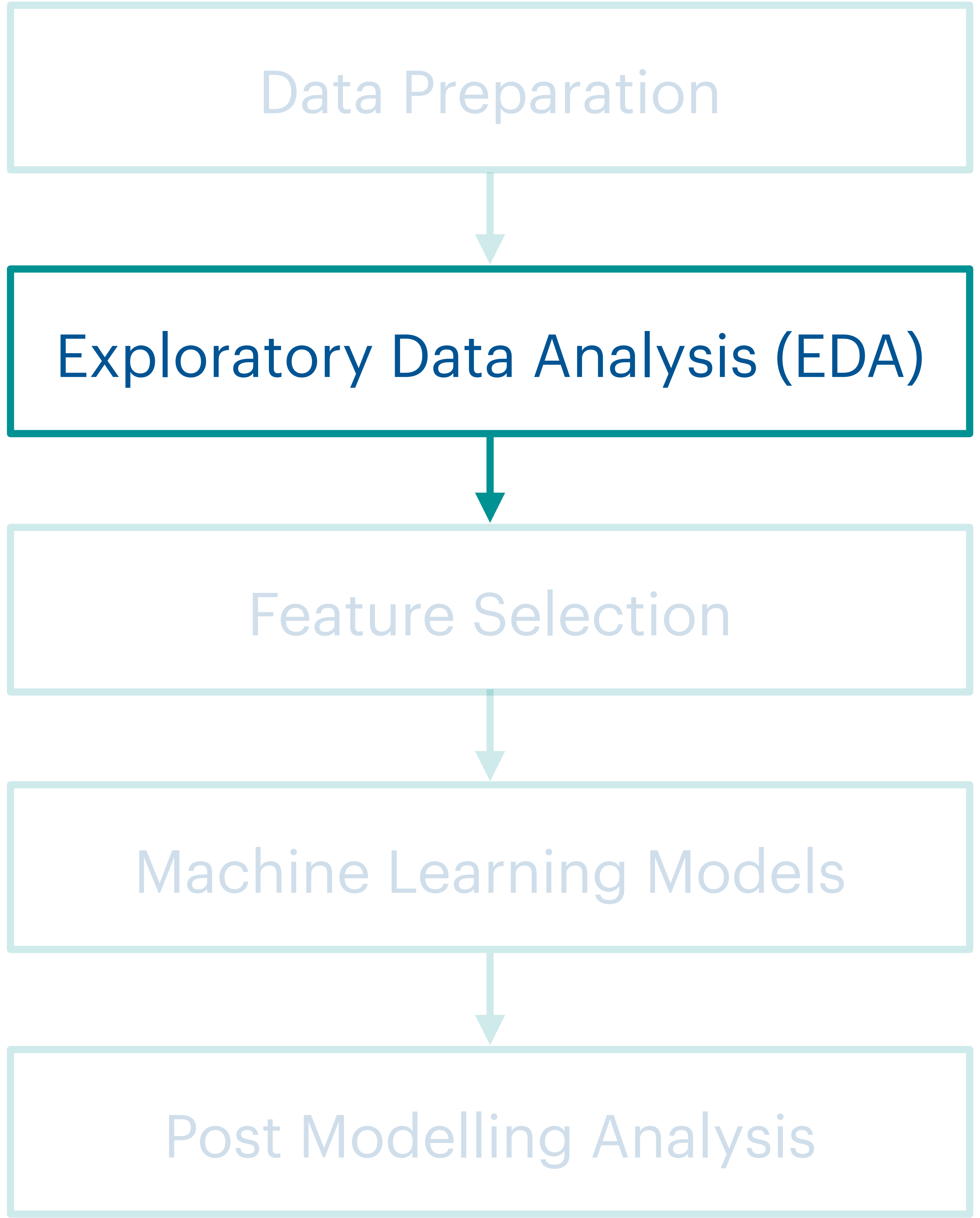
- Data Preparation is composed of *formatting* and also *cleaning* data from *missing values* and *outliers*. Prepare data for the next steps.
- The values of *'Cu_mg_kg'* are float numbers. Convert this regression problem into a *classification problem with 4 different classes* in such a way that the values of *'Cu_mg_kg'* in the train set are *equally distributed* between these four classes.



Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

- Do some kind of Exploratory Data Analysis (EDA) to become more familiar with the given data sets.



Data Preparation

Exploratory Data Analysis (EDA)

Feature Selection

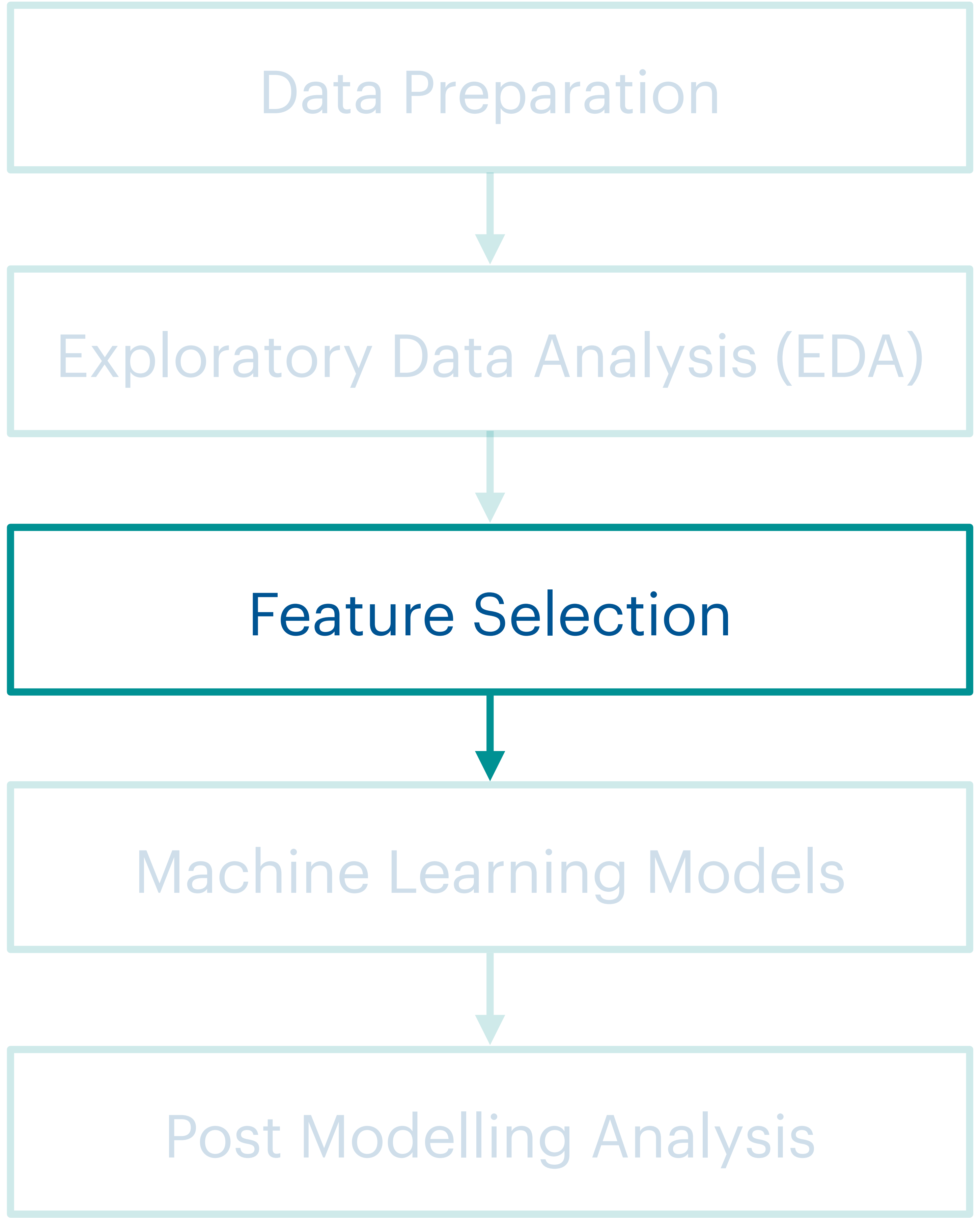
Machine Learning Models

Post Modelling Analysis

Feature Selection

Feature Selection

- Use a Feature Selection technique to select a subset of features
  - *Size(selected_features) should be between 100 and 200. (i.e., one order of magnitude reduction)*


Data Preparation

Exploratory Data Analysis (EDA)

Feature Selection

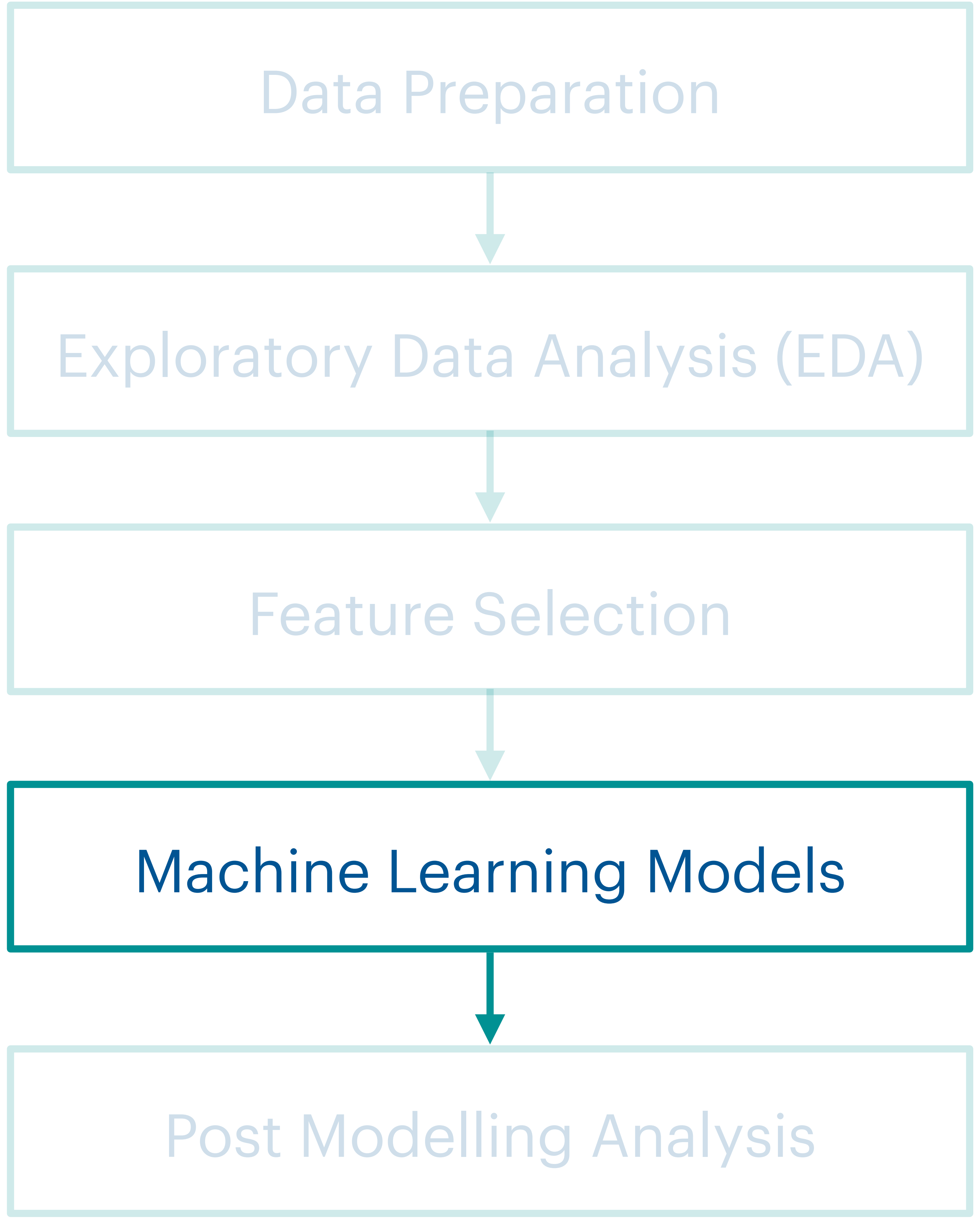
Machine Learning Models

Post Modelling Analysis

Machine Learning Models

Interpretability-Oriented Machine Learning Models

- Use *one baseline* and *two different interpretability-oriented machine learning models* to solve this regression problem.
  - *Baseline*
  - *Model A*
  - *Model B*
- At least in one of the proposed models, use *grid search* to tune hyper-parameters.



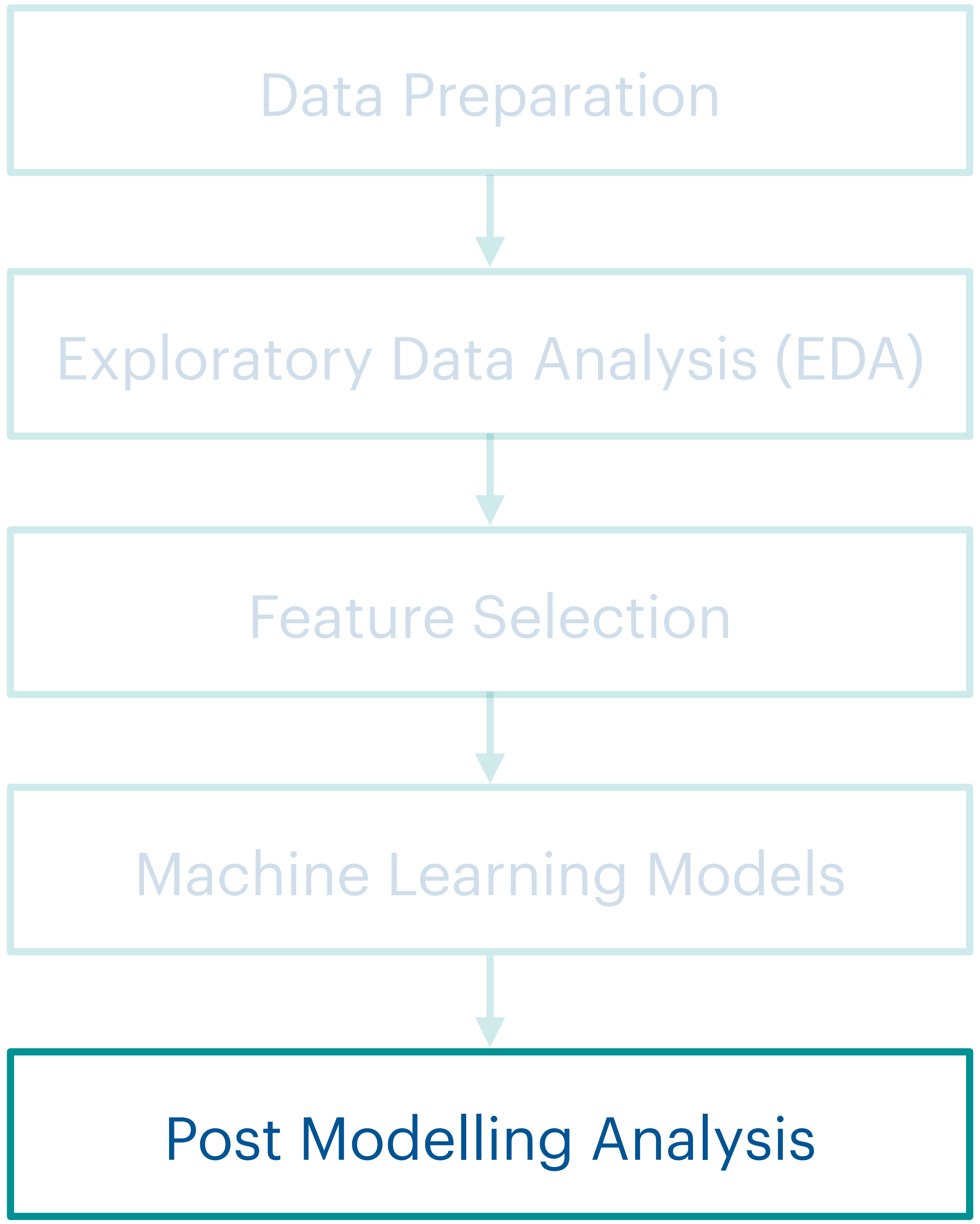
Evaluation Metrics

- For each model, calculate the following evaluation metrics on the test set:
  - Accuracy
  - Recall
  - Precision
  - F1-Score

Post Modelling Analysis

Post Modelling Analysis

- In Post Modelling Analysis,
  - Compare the different models in terms of *performance* and *interpretability*.
  - Analyse the *trade-off* between performance and interpretability.
  - Select one model that you consider as the best one and analyse it carefully.



Optional Section with Bonus

Optional Section

with Bonus

- Consider this problem as a regression problem and repeat the last two phases of the pipeline *(Machine Learning Models & Post-Modelling Analysis)* to solve this regression Problem.
- *Mean-Absolute-Error (MAE)* is a metric to evaluate the performance of regression models. It is the mean of the absolute values of each prediction error on all instances of the test set.
  - For each model, calculate *MAE* on test set.

* Bonus: Up to 0.5 additional points in your worst note (Labs average, Test 1, or Project)