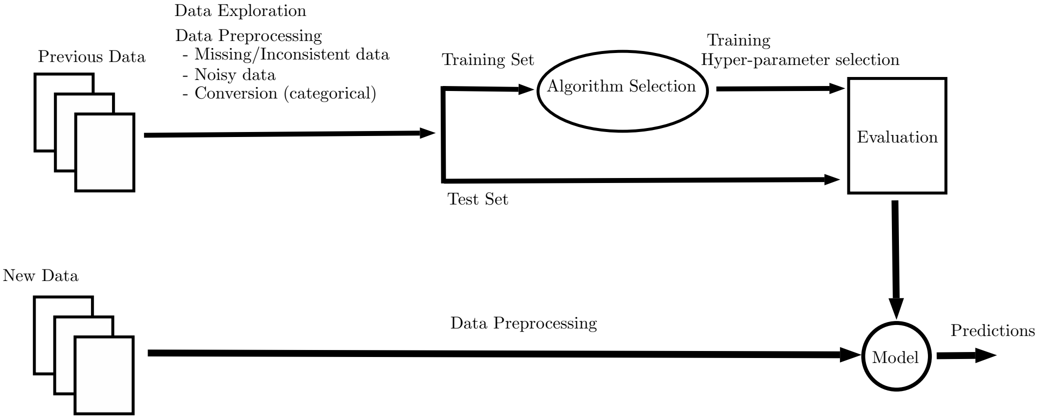
POLIMI GRADUATE SCHOOL OF MANAGEMENT

DATA PREPARATION

PERCORSO EXECUTIVE DATA SCIENCE AND BUSINESS ANALYTICS

Mauricio Soto - mauricioabel.soto@polimi.it

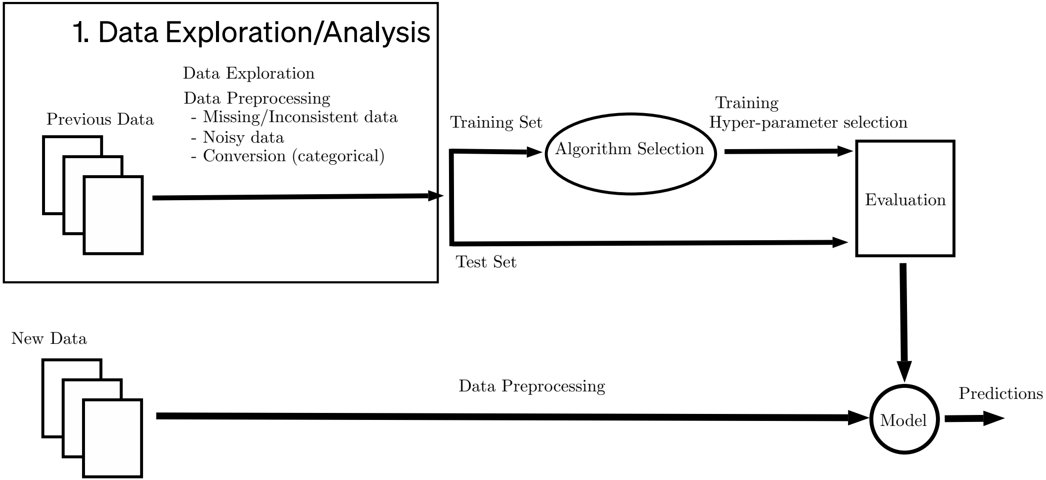
WORKFLOW



Data Exploration

Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)

WORKFLOW



1. Data Exploration/Analysis

Previous Data

Data Exploration

Data Preprocessing
 - Missing/Inconsistent data
 - Noisy data
 - Conversion (categorical)

Training Set

Algorithm Selection

Training
Hyper-parameter selection

Evaluation

Test Set

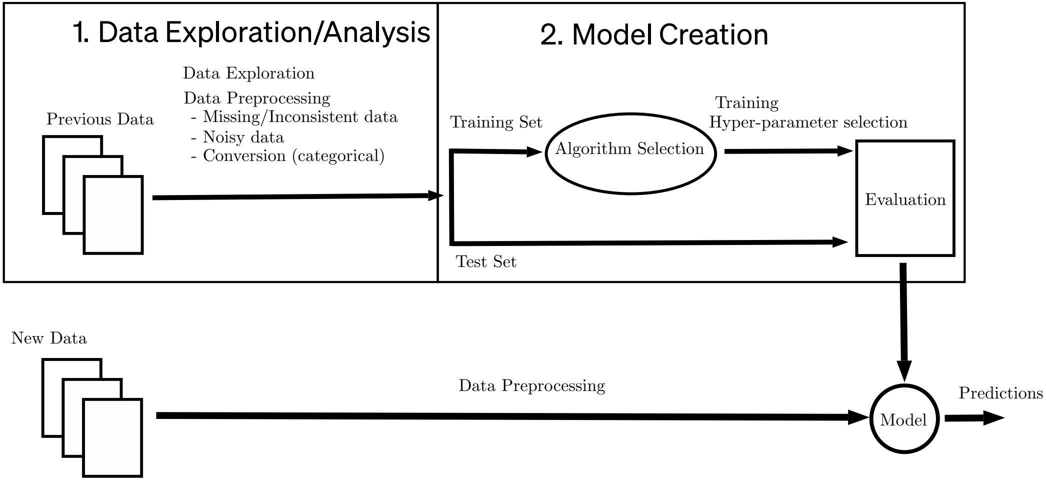
New Data

Data Preprocessing

Model

Predictions

WORKFLOW



1. Data Exploration/Analysis

Previous Data

Data Exploration

Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)

2. Model Creation

Training Set

Algorithm Selection

Training

Hyper-parameter selection

Evaluation

Test Set

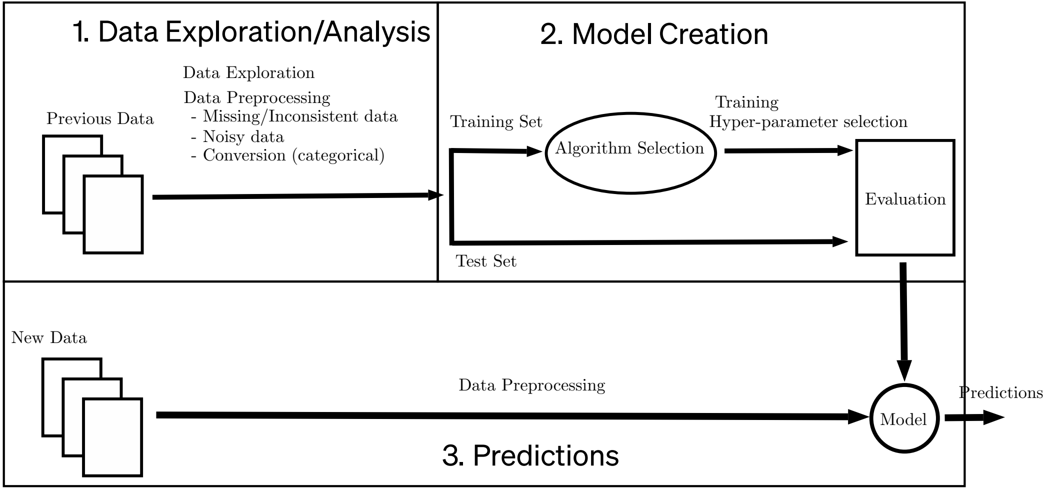
New Data

Data Preprocessing

Model

Predictions

WORKFLOW



POLIMI GRADUATE SCHOOL OF MANAGEMENT

INCOMPLETE DATA

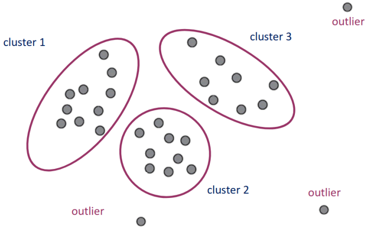
- Inspection
- Elimination
- Identification
- Replacement
  - mean value of numerical attributes
  - mean value of the target class
  - value estimated sing statistical inference

WHAT IS AN OUTLIER AND HOW TO RECOGNIZE IT

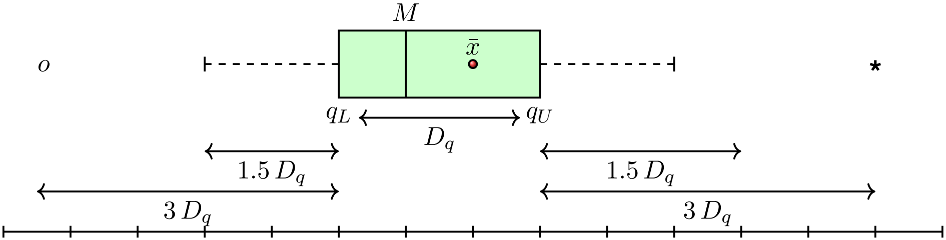
https://pollev.com/mauriciosoto

NOISY DATA

- Univariate
  - Normal-like distribution
    $$[\bar{\mu} - 2\bar{\sigma}, \bar{\mu} + 2\bar{\sigma}]$$
    contains about 96% of the data
  - In the general case, **Tchebysheff Theorem** states taht for $\gamma > 1$
    $$[\bar{\mu} - \gamma\bar{\sigma}, \bar{\mu} + \gamma\bar{\sigma}]$$
    contains $1 - 1/\gamma^2$ proportion of the observations
- Multi variate
  - Clustering techniques



BOX-PLOT



- $D_q = q_U - q_L = q_{0.75} - q_{0.25}$
- internal lower edge= $q_L - 1.5\,D_q$
- external lower edge= $q_L - 3\,D_q$

DATA TRANSFORMATION

- **Decimal Scaling**

$$x'_{ij} = \frac{x_{ij}}{10^k}$$

- **Min-Max** in the interval $[x'_{\text{min},j}, x'_{\text{max},j}]$

$$x'_{ij} = \frac{x_{ij} - x_{\text{min},j}}{x_{\text{max},j} - x_{\text{min},j}}(x'_{\text{max},j} - x'_{\text{min},j}) + x'_{\text{min},j}$$

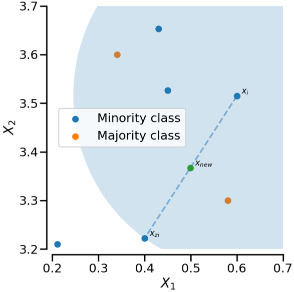
- **z-index**

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$$

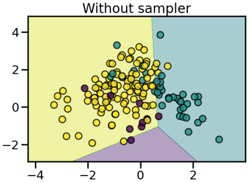
DATA REDUCTION

- **Sampling**
  - Simple sampling
  - Stratified sampling
- **Selection**
  - Filter methods
  - Wrapper methods
  - Embedded methods
- **Discretization, Aggregation**
- **Projection** (ex. PCA)

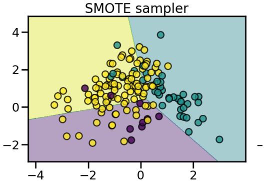
DATA UNBALANCE - SMOTE



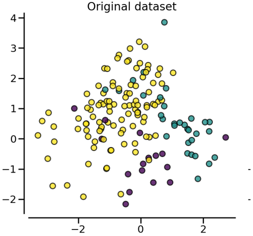
Without sampler



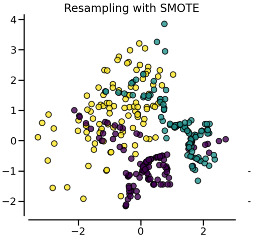
SMOTE sampler



Original dataset



Resampling with SMOTE



https://imbalanced-learn.org/

THANK YOU