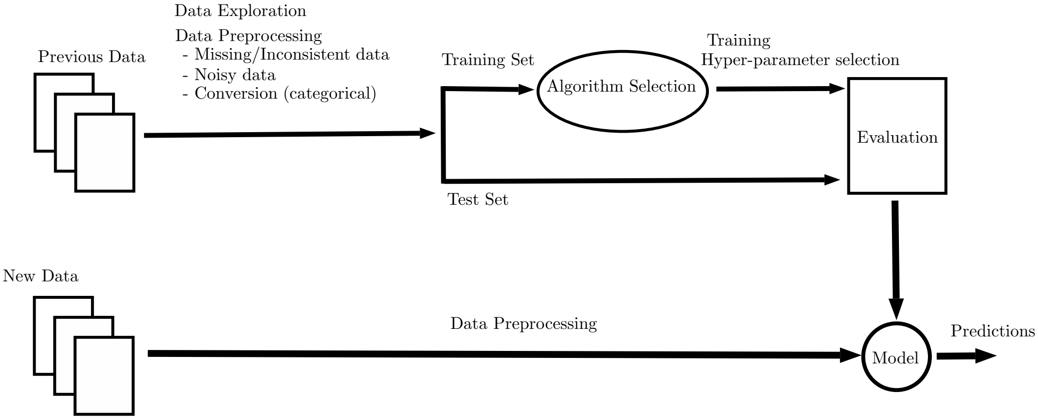
POLIMI GRADUATE SCHOOL OF MANAGEMENT

DATA PREPARATION

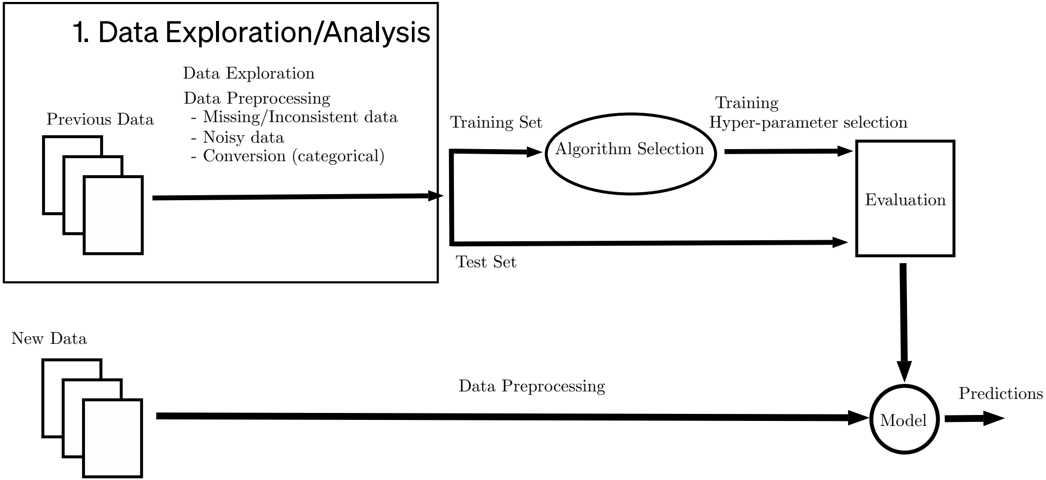
PERCORSO EXECUTIVE DATA SCIENCE AND BUSINESS ANALYTICS

Mauricio Soto - mauricioabel.soto@polimi.it

WORKFLOW

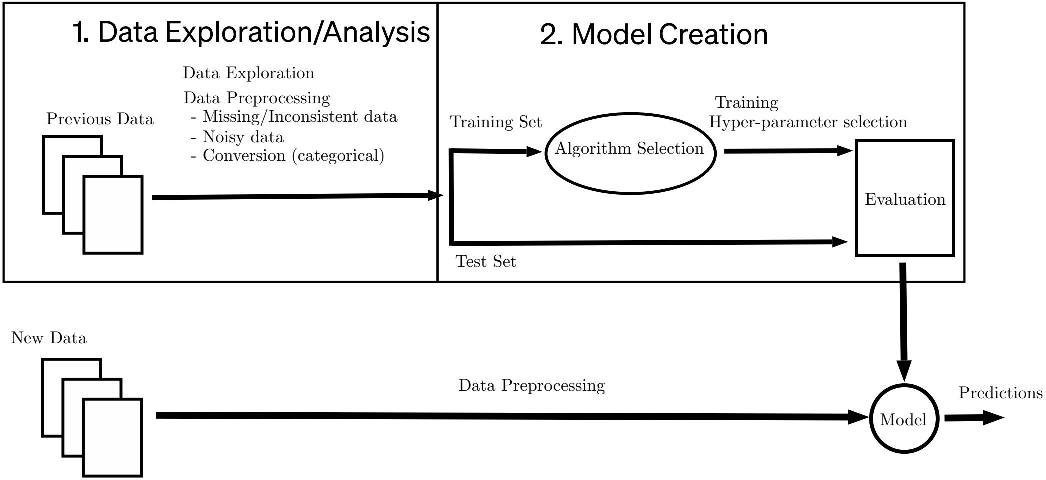

Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

WORKFLOW



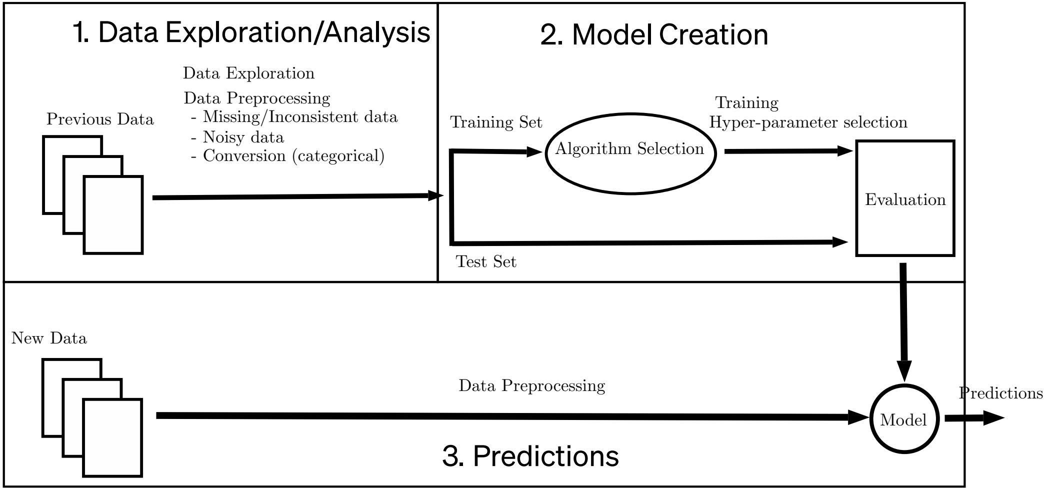
1. Data Exploration/Analysis
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

WORKFLOW



1. Data Exploration/Analysis
2. Model Creation
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

WORKFLOW


1. Data Exploration/Analysis
2. Model Creation
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions
3. Predictions

INCOMPLETE DATA

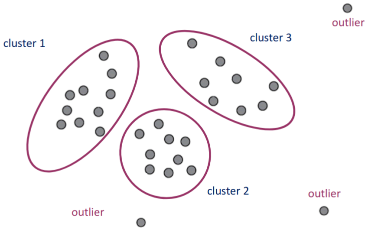
- Inspection
- Elimination
- Identification
- Replacement
  - mean value of numerical attributes
  - mean value of the target class
  - value estimated sing statistical inference

WHAT IS AN OUTLIER AND HOW TO RECOGNIZE IT

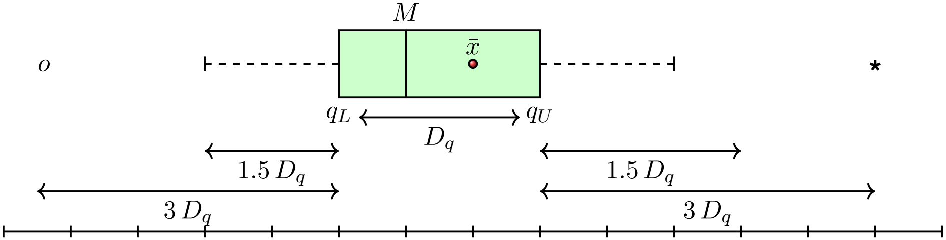
https://pollev.com/mauriciosoto

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 $[x'_{\text{min},j}, x'_{\text{max},j}]$

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

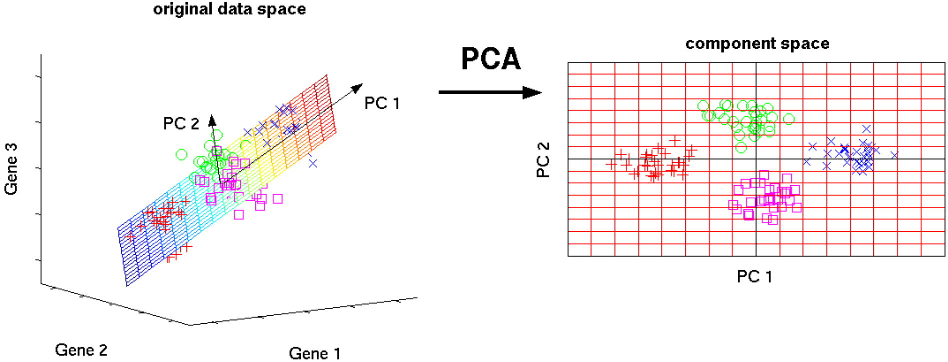
- **z-index**

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

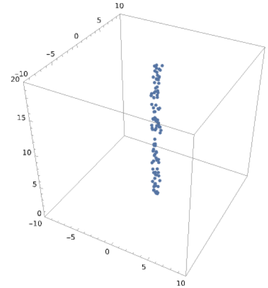
DATA REDUCTION

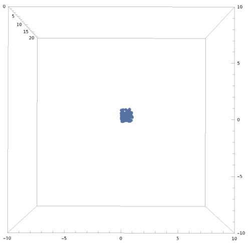
- **Sampling**
  - Simple sampling
  - Stratified sampling
- **Selection**
  - Filter methods
  - Wrapper methods
  - Embedded methods
- **Discretization, Aggregation**
- **Projection** (ex. PCA)

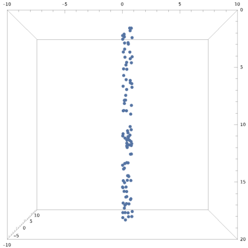
PCA: PRINCIPAL COMPONENT ANALYSIS



PCA-INTUITION







Compute projections that better capture the **variance**

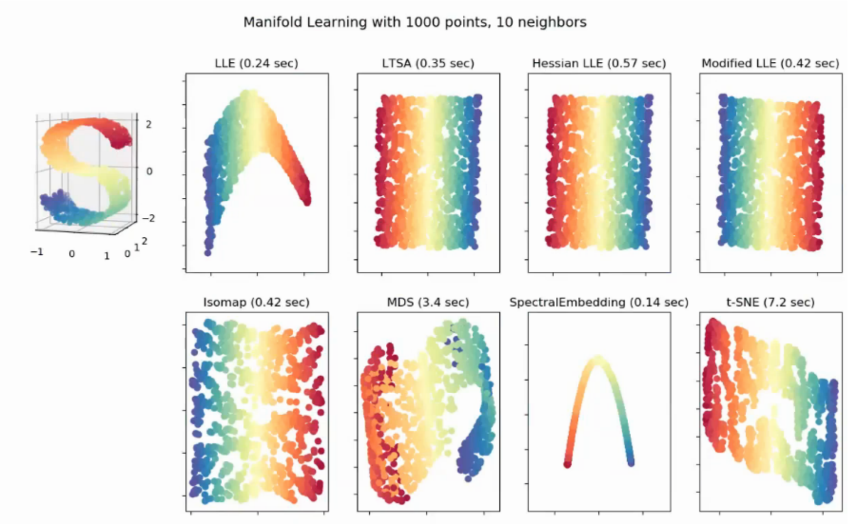
PCA: PRINCIPAL COMPONENT ANALYSIS

- Covariance data matrix $V = X'X$
- $\bar{x}_{ij} = x_{ij} - \bar{\mu}_j$
- New components p_j obtained as a linear transformation of original data $p_j = X\, w_j$
- Variance of $p_j = w_j'\, X'X\, w_j = w_j' V w_j$
- Maximizing the variance:

$$\max_{w_1} w_1' V w_1$$
$$\text{s.t. } w_1' w_1 = 1$$

- w_j is the j-th eigenvector of V, which explains a variance λ_j which is the j-th eigenvector

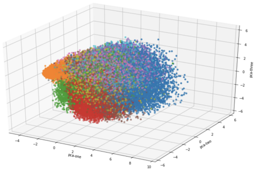
NONLINEAR REDUCTION



T-SNE: T-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING

We aim to project observations preserving observation distance.

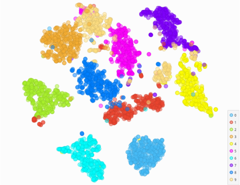
High-Dimensional Space



Distance as Normal Distribution

$$p_{ij} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq l} \exp\left(-\left\|x_k - x_l\right\|^2 / 2\sigma_i^2\right)}$$

Low-Dimensional Space



Distance as t-Students Distribution

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}}$$

We try to minimize the divergence between the distributions

$$KL(P\|Q) = \sum_i \sum_{j \neq i} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$ (Kullback-Leibler)

THANK YOU