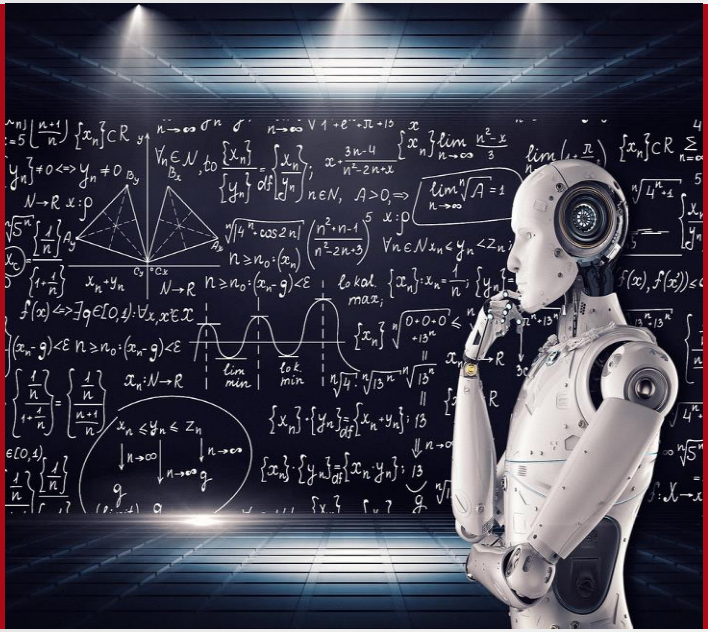
esprit
Se former autrement
HONORIS UNITED UNIVERSITIES



Chapter 1 : Data preparation

Unit: Machine Learning

United Nations Educational, Scientific and Cultural Organization
UNESCO Chair "Project-based learning" ESPRIT School of engineering, Tunisia

EUR-ACE® Délivrée par la Commission des Titres d'Ingénieur

CONFÉRENCE DES GRANDES ÉCOLES

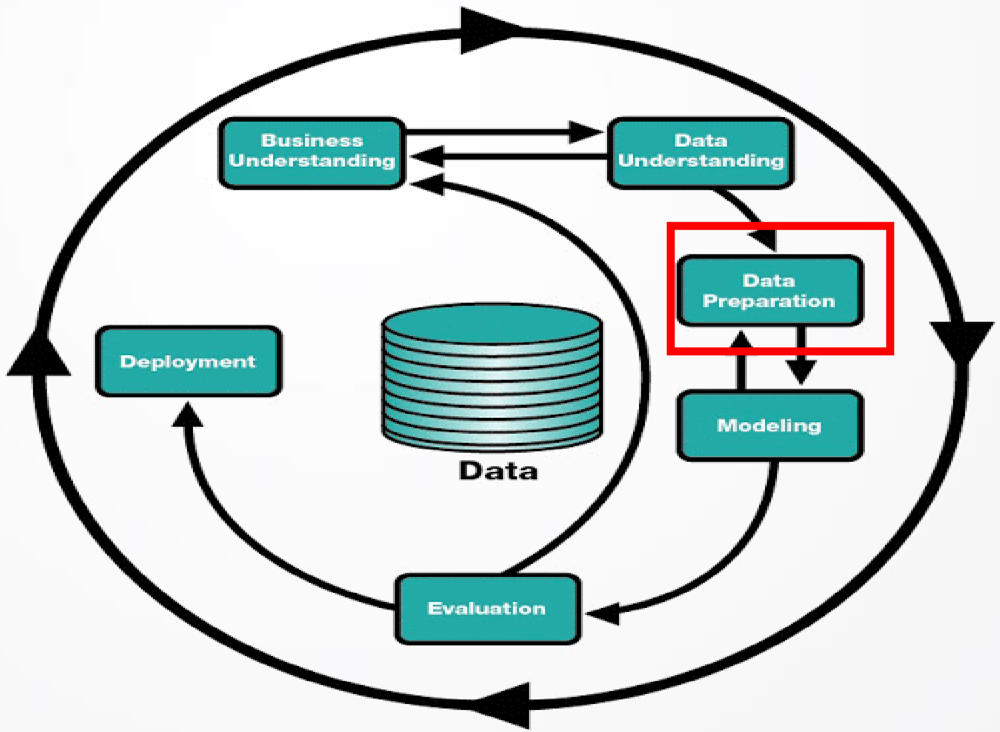
cdio CONCEIVE DESIGN IMPLEMENT OPERATE

Introduction

- Predictive modeling projects involve learning from data.
- The data refers to examples or cases from the field that characterizes the problem to be solved.
- On a Machine learning project, raw data typically cannot be used directly.
- This is because of reasons such as:
  - Machine learning algorithms require data to be numbers.
  - Some machine learning algorithms impose requirements on the data.
  - Statistical noise and errors in the data may need to be corrected…

Interest

- The raw data must be pre-processed prior to being used to fit and evaluate a machine learning model.
- This step in a predictive modeling project is referred to as “**data preparation**“



Standard tasks of data preparation

Data Cleaning
- Identifying and correcting mistakes or errors in the data.

Data Transforms
- Changing the scale or distribution of variables.

Feature Engineering
- Deriving new variables from available data.

Dimensionality Reduction

Feature Selection
- Identifying those input variables that are most relevant to the task.

Principal component Analysis
- Creating compact projections of the data.

Standard tasks of data preparation

Data Cleaning
- Identifying and correcting mistakes or errors in the data.

Data Transforms
- Changing the scale or distribution of variables.

Feature Engineering
- Deriving new variables from available data.

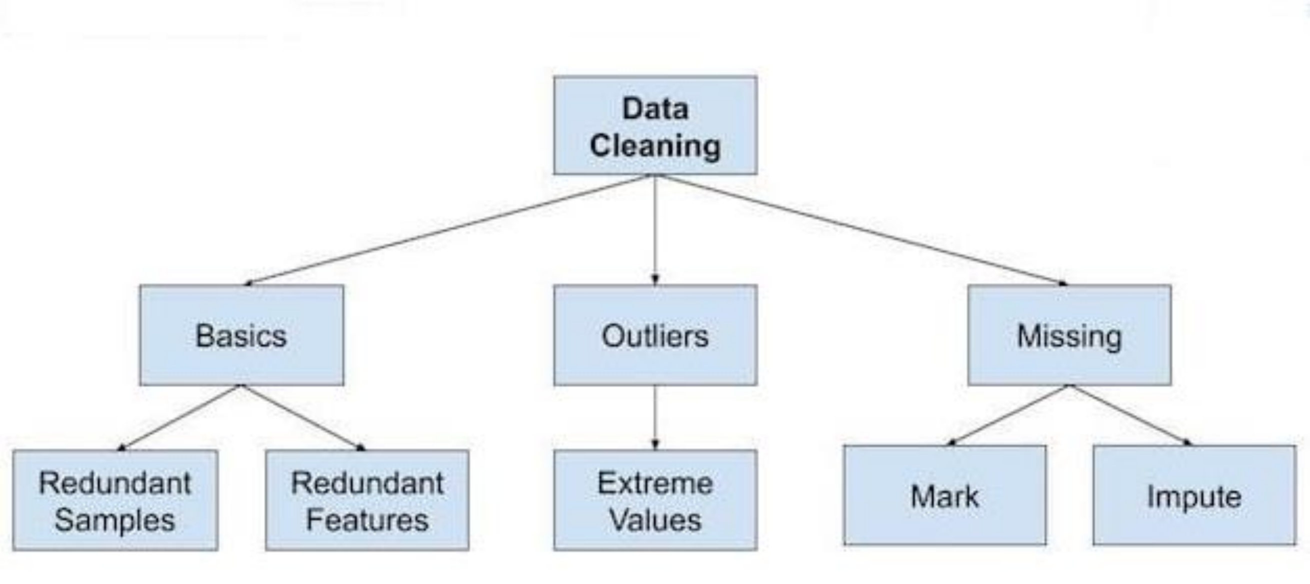
Dimensionality Reduction

Feature Selection
- Identifying those input variables that are most relevant to the task.

Principal component Analysis PCA
- Creating compact projections of the data.

Standard tasks

- Data cleaning includes simple tasks such as:
  - Define normal data.
  - Removing duplicate rows, redundant and irrelevant columns.
  - Identify outliers.
  - Dealing with missing values.



Data Cleaning: Redundant samples

pandas.DataFrame.duplicated

`DataFrame.duplicated(subset=None, keep='first')`

Return boolean Series denoting duplicate rows.

Considering certain columns is optional.

Parameters: **subset** : ***column label or sequence of labels, optional***

Only consider certain columns for identifying duplicates, by default use all of the columns.

keep : ***{'first', 'last', False}, default 'first'***

Determines which duplicates (if any) to mark.

- `first` : Mark duplicates as `True` except for the first occurrence.
- `last` : Mark duplicates as `True` except for the last occurrence.
- False : Mark all duplicates as `True`.

Returns: **Series**

Boolean series for each duplicated rows.

pandas.DataFrame.drop_duplicates

`DataFrame.drop_duplicates(subset=None, keep='first', inplace=False, ignore_index=False)`

Return DataFrame with duplicate rows removed.

Considering certain columns is optional. Indexes, including time indexes are ignored.

Parameters: **subset** : ***column label or sequence of labels, optional***

Only consider certain columns for identifying duplicates, by default use all of the columns.

keep : ***{'first', 'last', False}, default 'first'***

Determines which duplicates (if any) to keep. - `first` : Drop duplicates except for the first occurrence. - `last` : Drop duplicates except for the last occurrence. - False : Drop all duplicates.

inplace : ***bool, default False***

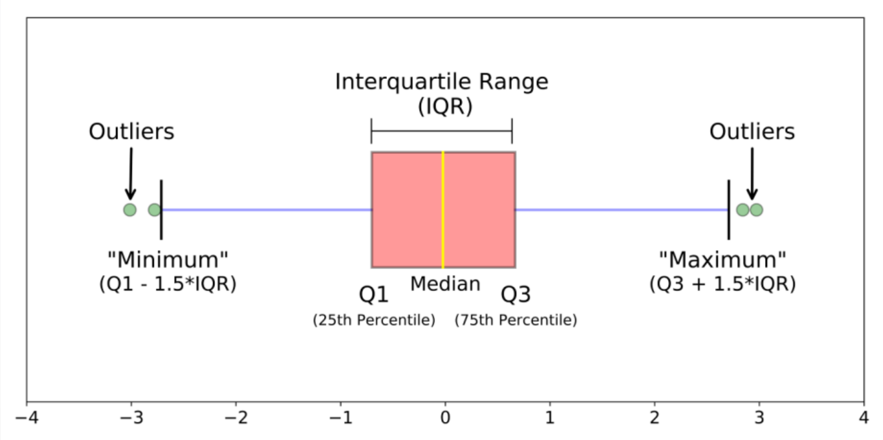
Whether to drop duplicates in place or to return a copy.

ignore_index : ***bool, default False***

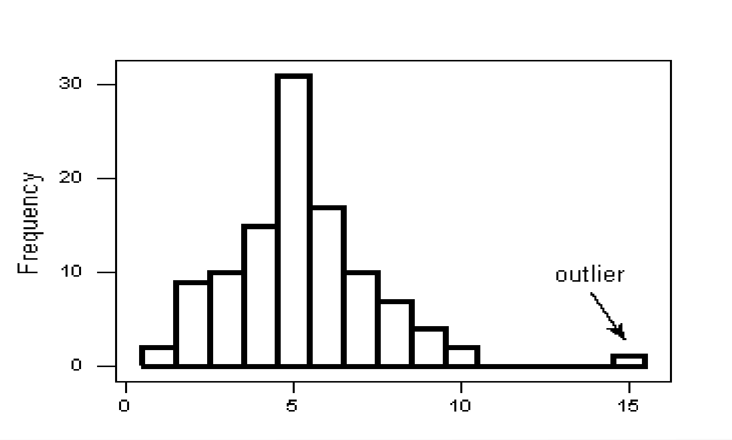
If True, the resulting axis will be labeled 0, 1, …, n - 1.

Data Cleaning: Outliers

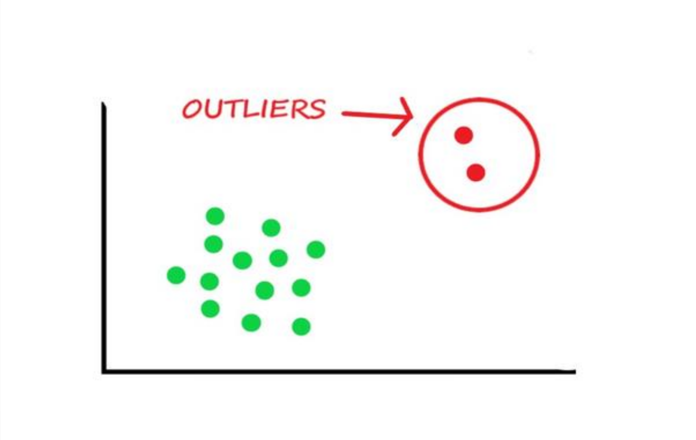
- Some methods to detect outliers:



Boxplot



Histogram



Scatter plot

Data Cleaning: Outliers

- Some methods to handle outliers:
  - Drop the outlier records.
  - Cap your outliers' data (min, max).
  - Assign a new value…

Data cleaning: Missing value

What is missing data?

- Missing data are defined as not available values, and that would be meaningful if observed.
- Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc.

- Filling in missing values with data is called data imputation.

Data cleaning: Missing value

Some data imputation methods:

- Delete individuals with missing data
- Replace missing data with a fixed value
- Replace missing data with a decision tree
- Replace missing data with nearest values
- Replace missing data with dedicated algorithms…

Data cleaning: Missing value

A popular approach for data imputation is to calculate a statistical value for each column (such as a mean) and replace all missing values for that column with the statistic.

Strategy	« mean »	« median »	« most_frequent»	« constant»
Definition	Replace missing values using the mean along each column	Replace missing values using the median along each column	Replace the missing value using the most frequent value along each column	Replace missing values with fill_value. • fill_value=0 when imputing numeric data • fill_value=missing_value” for strings or object data types
Data type	numeric data	numeric data	numeric data/Strings	numeric data/Strings
Attention: Missing values must be marked with NaN (Not-a-Number)				

Standard tasks of data preparation

Data Cleaning
- Identifying and correcting mistakes or errors in the data.

Data Transforms
- Changing the scale or distribution of variables.

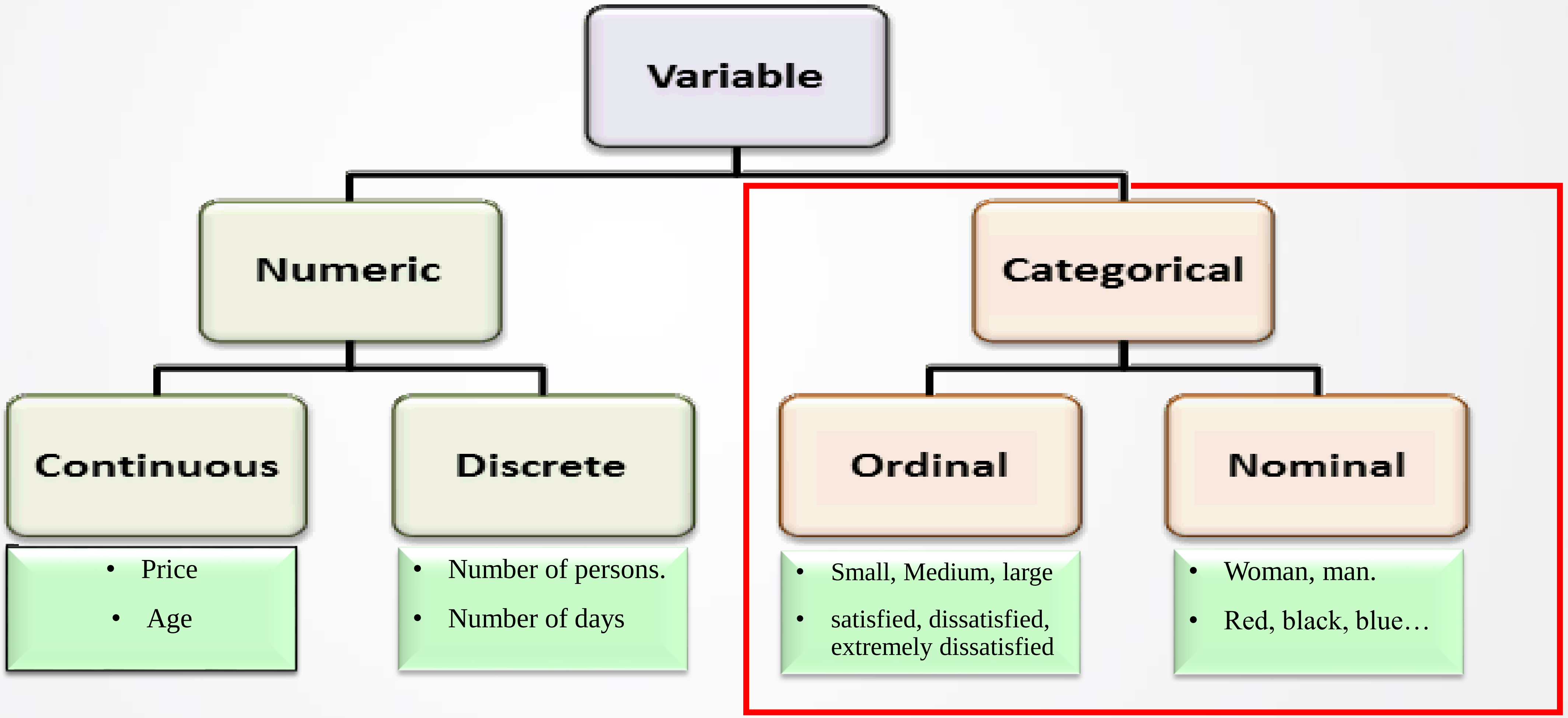
Feature Engineering
- Deriving new variables from available data.

Dimensionality Reduction

Feature Selection
- Identifying those input variables that are most relevant to the task.

Principal component Analysis
- Creating compact projections of the data.

Data Transformation: Categorical features (1/3)



Data Transformation: Categorical features (2/3)

- Transforming categorical data is an essential step during data preprocessing. sklearn's machine learning library require the input dataset to always have numeric values it does not support categorical data.
- It is necessary to **<u>convert</u>** categorical features to a numerical representation.
- Before you start transforming your data, it is important to figure out if the feature you're working on is ordinal (as opposed to nominal). An ordinal feature is best described as a feature with ordered categories.

Data Transformation: Categorical features (3/3)

- Once you know what type of categorical data you're working on, you can pick a suiting transformation tool. In **sklearn** that will be:
  - A OrdinalEncoder or LabelEncoder for ordinal data,

Ordinal Encoding

workclass
State-gov
Self-emp-not-inc
Private
Private
Private

workclass
0
1
2
2
2

original dataset

x_1	x_2	y
5	8	calabar
9	3	uyo
8	6	owerri
0	5	uyo
2	3	calabar
0	8	calabar
1	8	owerri

LabelEncoder

{
"calabar" ---> 0
"owerri" ---> 1
"uyo" ---> 2
}

dataset with encoded labels

x_1	x_2	y
5	8	0
9	3	2
8	6	1
0	5	2
2	3	0
0	8	0
1	8	1

  - A OneHotEncoder for nominal data.

OneHot Encoding

workclass
State-gov
Self-emp-not-inc
Private
Private
Private

State-gov	Self-emp-not-inc	Private
1	0	0
0	1	0
0	0	1
0	0	1
0	0	1

Data transformation: Feature scaling

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Data transformation: Feature scaling
Standardization

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Data Transformation: Feature scaling
Normalization

- Some normalization methods are :
  - Maximum Absolute Scaling

$$x_{scaled} = \frac{x}{max(|x|)}$$

  - Min-max normalization

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - Decimal scaling
  - Z-score normalization…

Standard tasks of data preparation

- **Data Cleaning**
  - Identifying and correcting mistakes or errors in the data.
- **Data Transforms**
  - Changing the scale or distribution of variables.
- **Feature Engineering**
  - Deriving new variables from available data.

Dimensionality Reduction

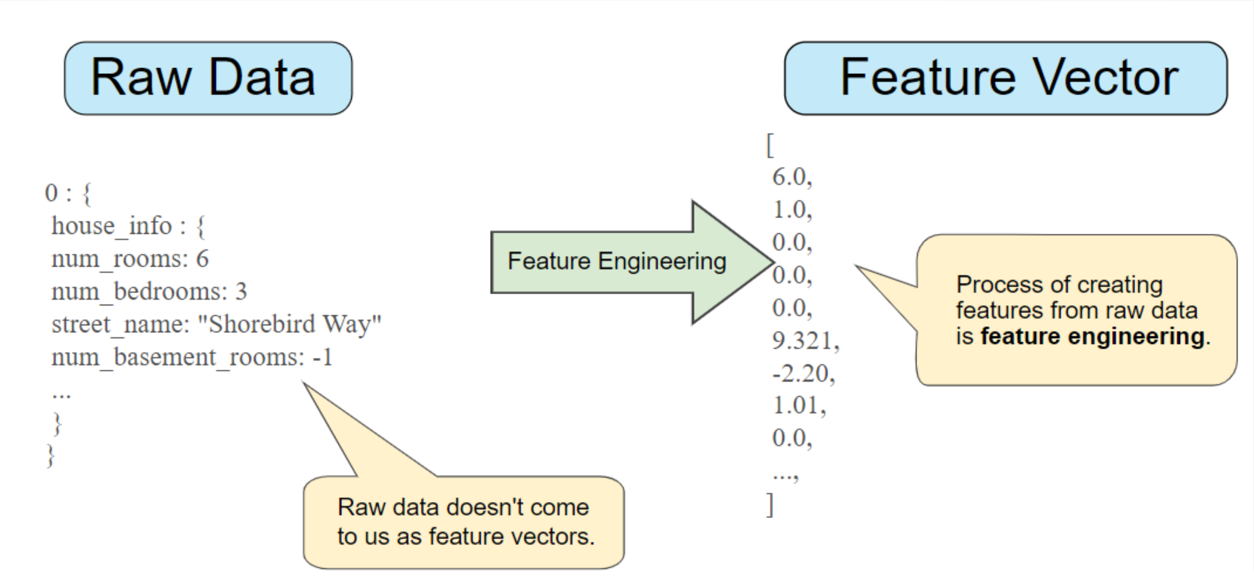
- **Feature Selection**
  - Identifying those input variables that are most relevant to the task.
- **Principal component Analysis**
  - Creating compact projections of the data.

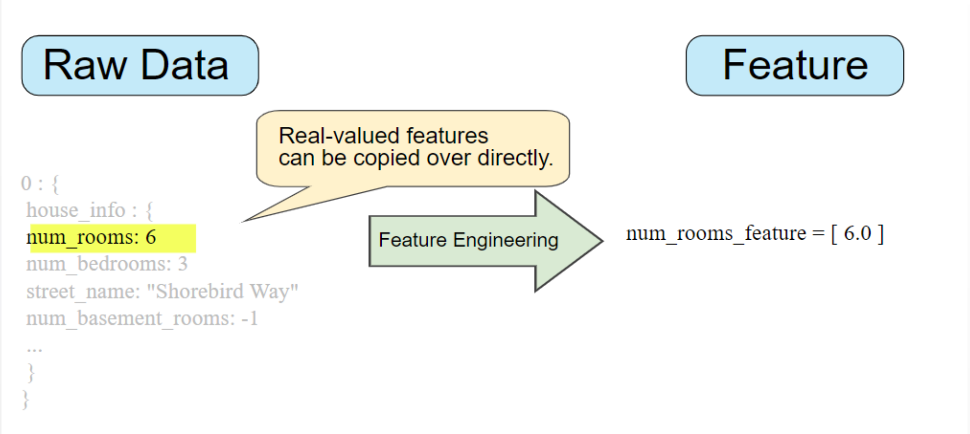
Feature engineering (1/2)

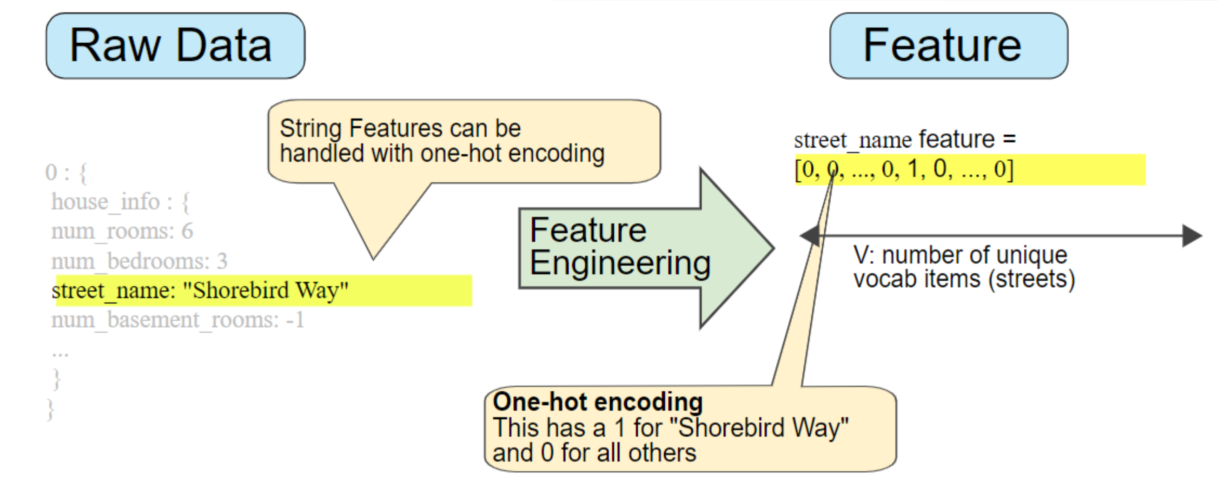
- Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data

- Example of Feature Engineering techniques:
  - **Creation of Features:** Sum, subtraction, average, min, max, product, quotient of the group of features.
  - **Extracting Features from a text**
  - **Topic extraction:** extract main topics from a text…
  - **Extracting features from an image**

Feature engineering (2/2)







Standard tasks of data preparation

Data Cleaning
- Identifying and correcting mistakes or errors in the data.

Data Transforms
- Changing the scale or distribution of variables.

Feature Engineering
- Deriving new variables from available data.

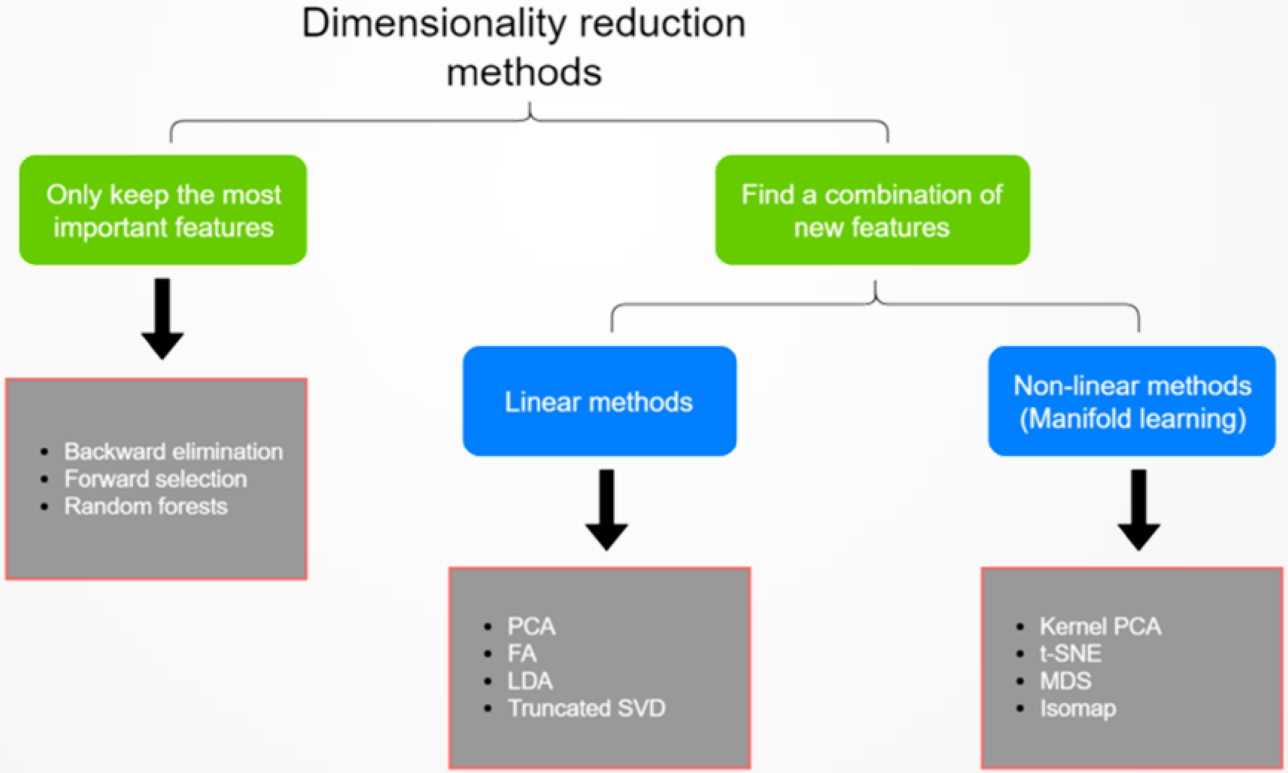
Dimensionality Reduction

Feature Selection
- Identifying those input variables that are most relevant to the task.

Principal component Analysis
- Creating compact projections of the data.

Dimensionality Reduction

- **Dimensionality reduction**, or **dimension reduction**, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data.



Standard tasks of data preparation

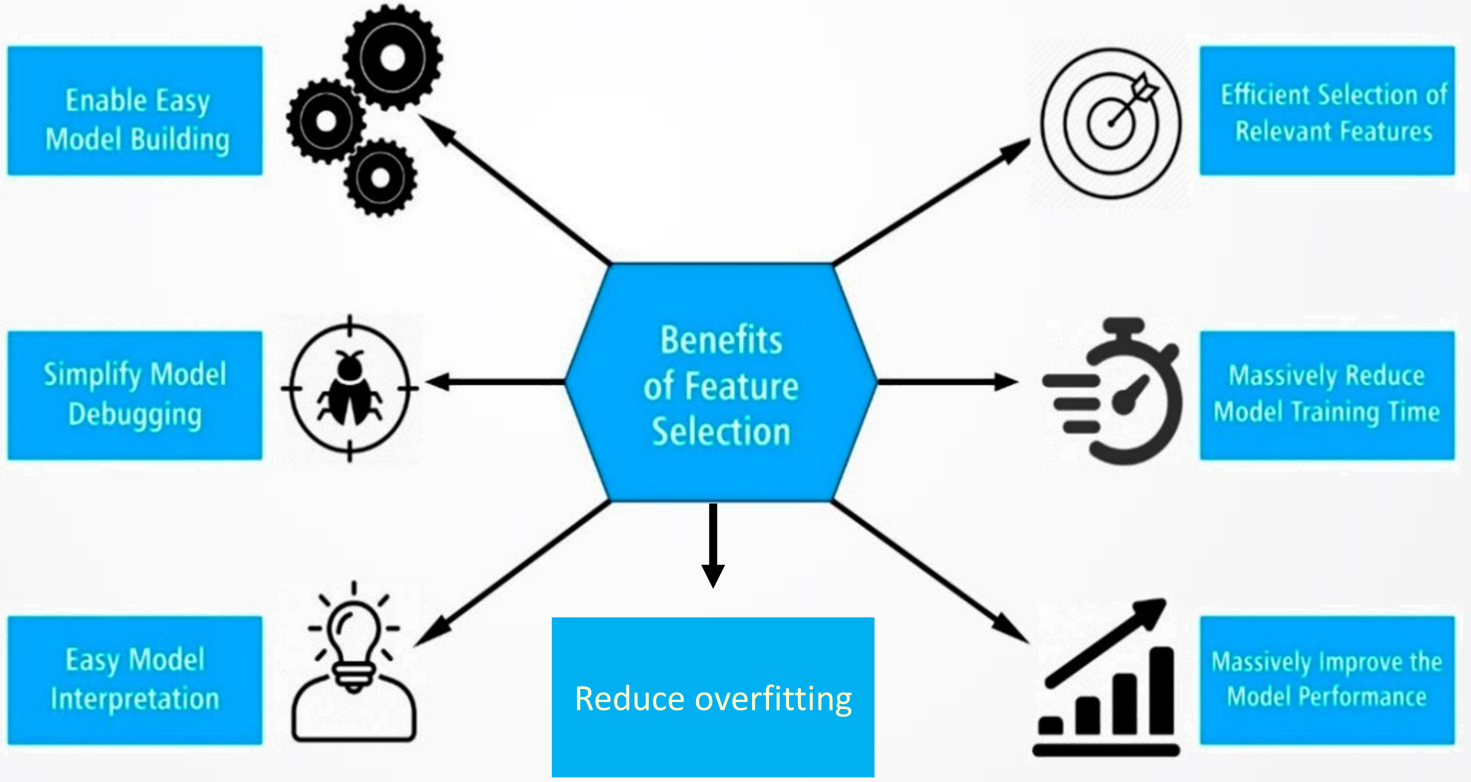
- **Data Cleaning**
  - Identifying and correcting mistakes or errors in the data.
- **Data Transforms**
  - Changing the scale or distribution of variables.
- **Feature Engineering**
  - Deriving new variables from available data.

Dimensionality Reduction

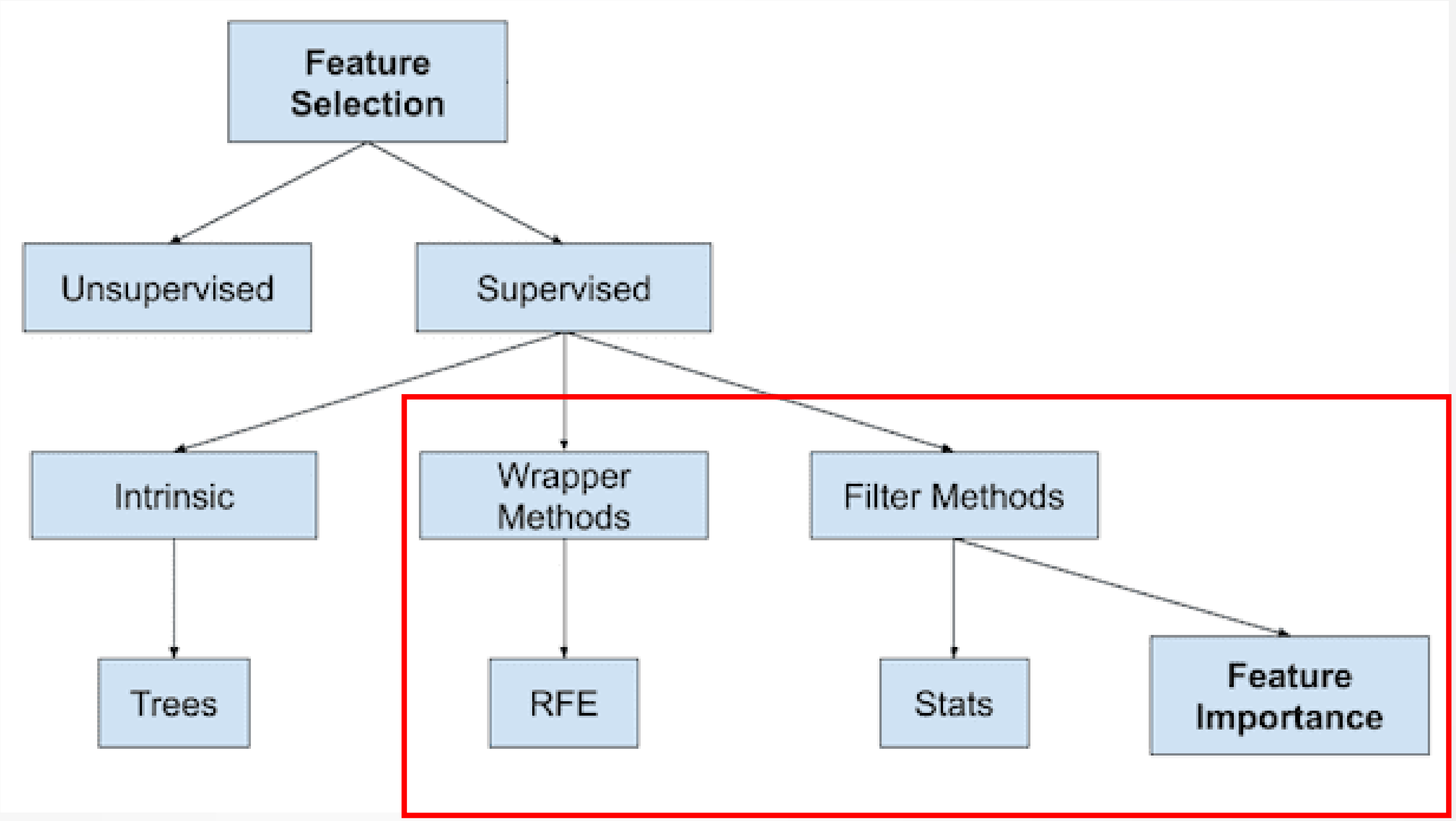
- **Feature Selection**
  - Identifying those input variables that are most relevant to the task.
- **Principal component Analysis**
  - Creating compact projections of the data.

Feature Selection

- Feature Selection: is the process of selecting the most important features to input in machine learning algorithms

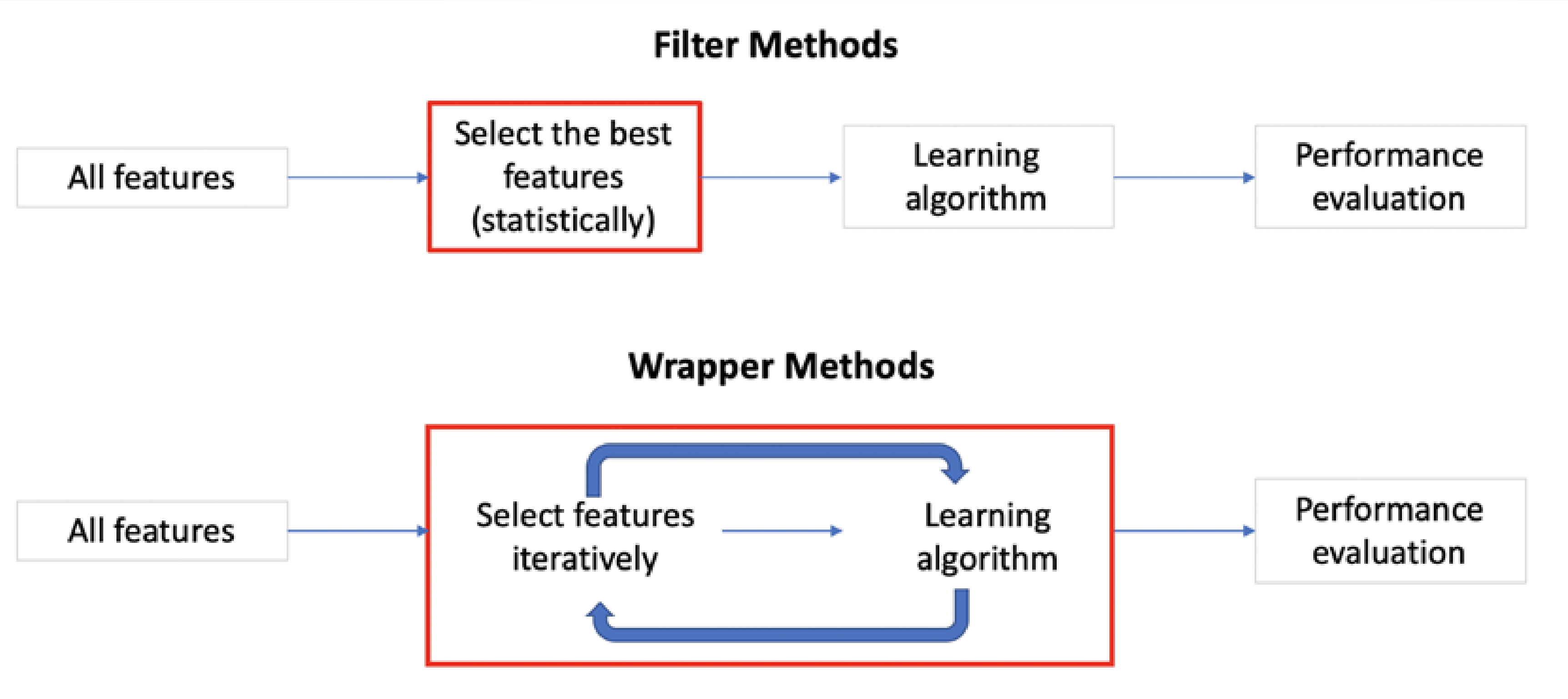


Feature Selection : Methods



Feature Selection : Methods

- **« Filter » vs. « Wrapper »:**



- Wrapper: choose in an iterative way the characteristics that give the best performing model.
- Filter: Assign a score to each input feature to select the best performing features.

Standard tasks of data preparation

Data Cleaning
- Identifying and correcting mistakes or errors in the data.

Data Transforms
- Changing the scale or distribution of variables.

Feature Engineering
- Deriving new variables from available data.

Dimensionality Reduction

Feature Selection
- Identifying those input variables that are most relevant to the task.

Principal component Analysis
- Creating compact projections of the data.