Data Mining and Machine Learning
Machine learning algorithms for environmental analysis.
An introduction.

Elena Grimaccia

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Aim of this lesson

- Knowledge of key machine learning models: Supervised learning approaches (generalised lineal models such as linear and not linear regression, logistic regression, predictions, qualitative predictors), Unsupervised learning techniques (Principal Component Analysis, Cluster Analysis) relevant to environmental data, Training- versus Test-Set Performance.
- Apply AI in Environmental Contexts: Utilize AI tools and techniques for environmental applications, including biodiversity monitoring, climate modelling, remote sensing, and big data analysis related to air, water, and soil quality.

Data mining: what is it? Everything related to data

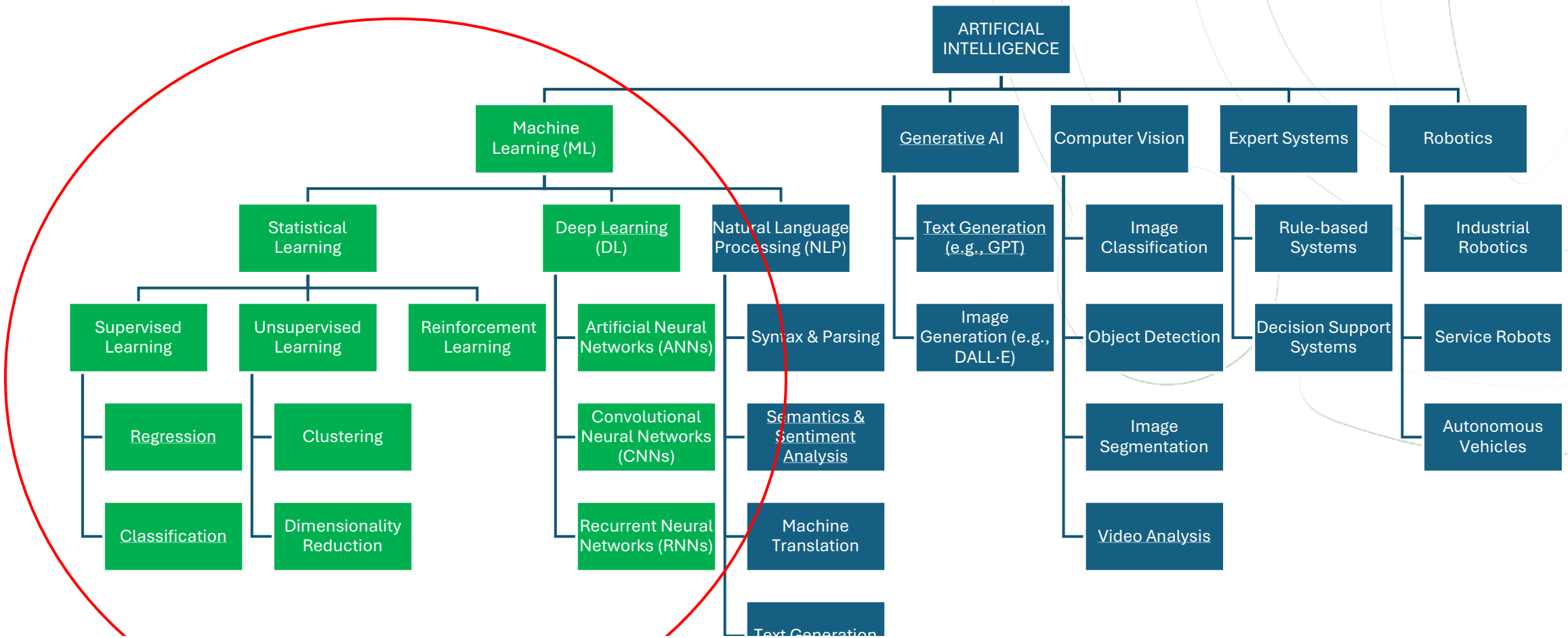
- The term "data mining" is a misnomer because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself.
- It also is a buzzword (OKAIRP conference 2005) and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support systems, including artificial intelligence (e.g., machine learning) and business intelligence. Often the more general terms (*large scale*) *data analysis* and *analytics*—or, when referring to actual methods, *artificial intelligence* and *machine learning*—are more appropriate.
- Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long
- https://mitmecsept.wordpress.com/wp-content/uploads/2017/04/data-mining-concepts-and-techniques-2nd-edition-impressao.pdf

Data mining: what is it? Everything related to data

- Data mining is the process of extracting and finding patterns in massive data sets involving methods at the intersection of machine learning, statistics, and database systems.
- Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use.
- Data mining is the analysis step of the "knowledge discovery in databases" process. It also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Machine Learning

Statistical methodologies (GLMs; classification; PCA; Clustering,....)



The approach to machine learning, typically aimed at prediction or optimization, pays little to no attention to the nature of the underlying data—something that is instead fundamental for a statistician.

Statistics is at the foundation of ML.

Machine learning methodologies have long been rooted in statistics:

- Michie, D., Spiegelhalter, D.J., Taylor, C.C. and Campbell, J. eds., 1995. Machine learning, neural and statistical classification. Ellis Horwood.
- Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression Trees (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470
- Golden, R. (2020). Statistical Machine Learning: A Unified Framework. United States: CRC Press

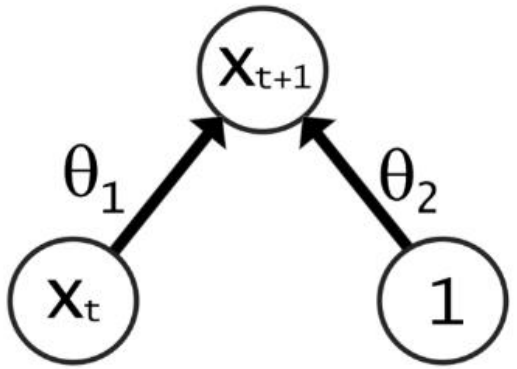


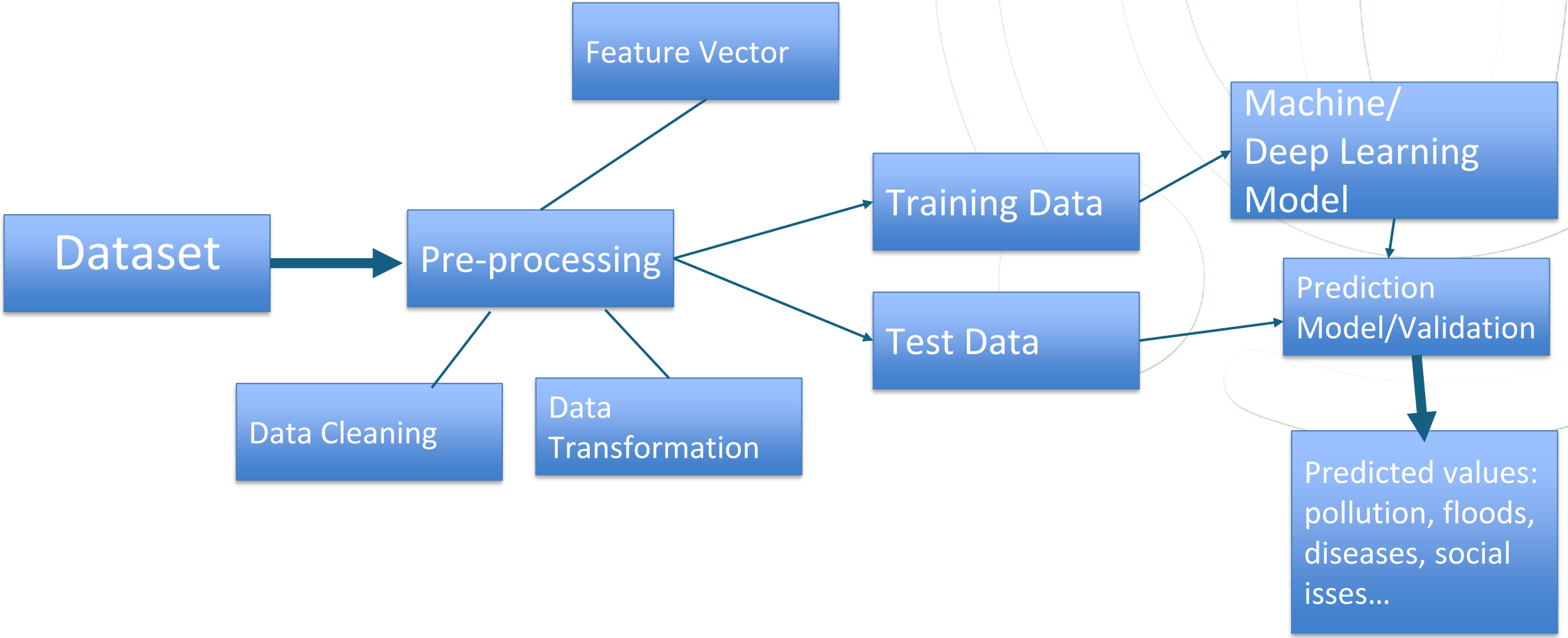
FIGURE 1.3
Artificial neural network node representation of stock market linear regression predictive model. An artificial neural network node representation of a linear regression model which predicts tomorrow's closing price x_{t+1} given today's closing price x_t using the formula $\hat{x}_{t+1} = \theta_1 x_t + \theta_2$.

Advantages of Machine Learning

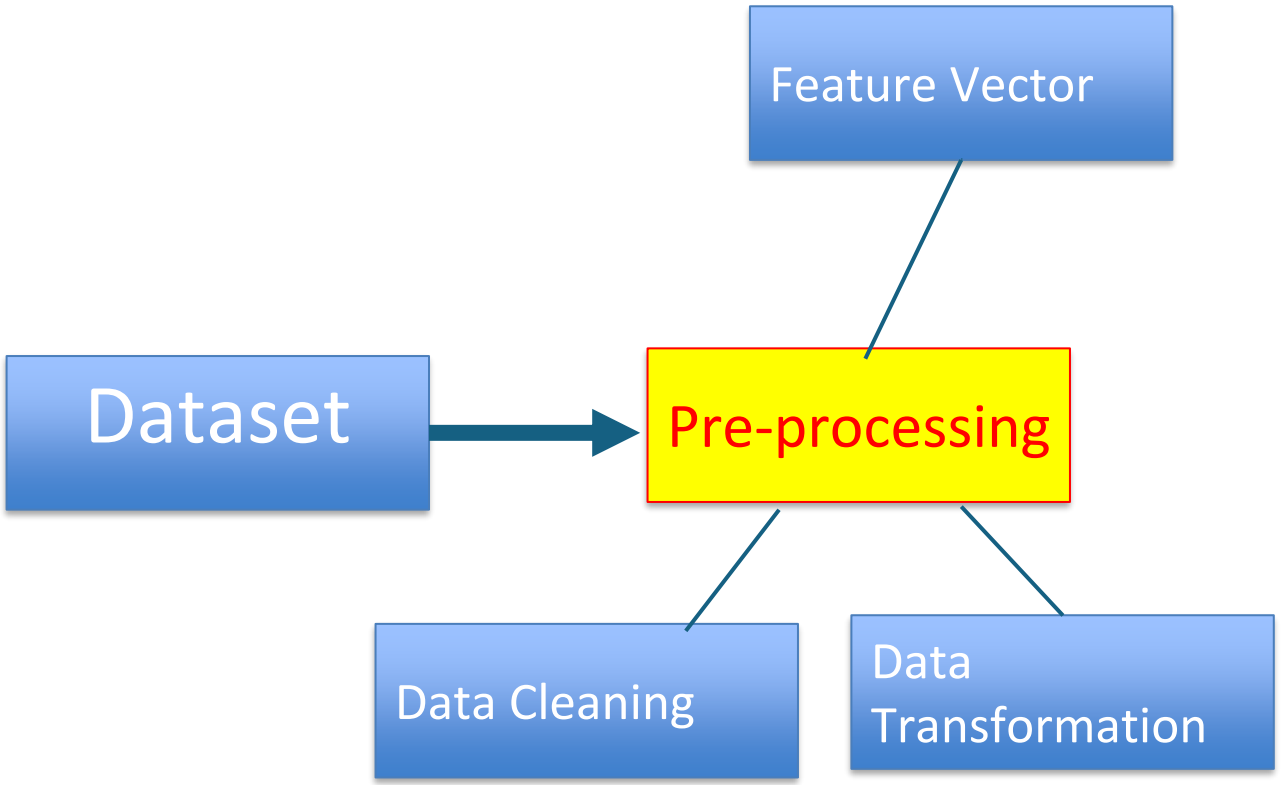
Machine learning has several advantages over traditional statistical methods.

- ML is a model-free philosophy that does not require prior statistical assumptions.
- The model can deal with non-linear and colinear data and handle sparse and high-dimensional data with high prediction ability.
- ML trains on part of the data (training dataset) and tests its performance on another part (testing dataset).
- This process makes out-of-sample prediction for ML algorithms high compared to traditional statistical methods, which are prone to overfitting.
- Dealing with non-structure data (medical images, words, multimedia) is another advantage of ML.

How do we actually perform a Machine Learning project?



1° step



It Improves Data Quality

How:

- It Handles Missing Data
- It Normalizes and Scales Data
- It Eliminates Duplicate Records
- It Handles Outliers
- It Helps in Enhancing Model Performance:

- developing new features
- modifying existing ones
- encoding category variables
- developing interaction terms
- retrieving pertinent data from text or timestamps.

https://lakefs.io/blog/data-preprocessing-in-machine-learning/

Missing Data

- Evaluate the data and look for missing values:
  - Missing values can break actual data trends and potentially result in additional data loss when entire rows and columns are deleted due to a few missing cells in the dataset.
  - “rubbish in, rubbish out”: is the concept that flawed, biased or poor quality (“rubbish") information or input produces a result or output of similar (“rubbish") quality, https://physicsworld.com/a/machine-learning-collaborations-accelerate-materials-discovery/

If you discover any, you can choose from two methods to deal with this issue:

- Remove the whole row with a missing value. However, eliminating the full row increases the likelihood of losing some critical data. This strategy is beneficial if the dataset is massive.
- Estimate the value.

Missing Data

- The Treatment of Missing Data
- Unit nonresponse: all responses relating to a unit are missing, for example because the interviewed person refused to answer the questionnaire or because they were not reachable.
- Item nonresponse: only partial data are available and data are missing only for some variables.
- Standard error, p-value and other measures of uncertainty calculated through the usual methods can be misleading if they do not take into account the uncertainty due to missing data .

Missing Data

MAR, MCAR, MNAR

- MAR (Missing At Random) allows the probabilities of observing missing data to depend on the observed data but not on the missing data. MAR means that the distribution of V is the same both among the cases (units) in which the variable is observed and among those in which it is missing.
- MCAR (Missing Completely At Random) occurs when the probability of observing missing data does not depend even on the observed values. For MCAR, missing data are a simple random sample of all possible values.
- When MAR holds, it is appropriate to base inferences on the parameters of the distribution of Y on the likelihood calculated on the observed data only.
- When the MAR condition does not hold, the data are called MNAR (Missing Not At Random).

Missing Data: what shall we do with them?

DELETION

- Case deletion consists of using only those units for which all considered variables have been observed.
- The only condition under which case deletion does not produce biases in estimates is that missing data are MCAR and only in some circumstances does it produce good inferences even under MAR.
- When missing data are not MCAR, the results obtained from case deletion can be biased, as complete data are not representative of the entire population.

Missing Data: what shall we do with them?

IMPUTATION

- Imputation methods consist of replacing missing data with plausible values in such a way as to produce a complete data set.
- Imputation methods are divided into deterministic and stochastic depending on whether the method considered always provides the same value for those units that have the same characteristics or not.
- Usually in the imputation process, a certain number of auxiliary variables are used that are statistically correlated with the variable on which non-responses occur through an imputation model.
- The main reason for using an imputation method is to reduce the bias due to non-responses.

Missing Data: what shall we do with them?

IMPUTATION METHODS

- Unconditional Mean: replacing missing data with the mean of the corresponding variable calculated for the observed values
- Unconditional Distribution: preserve the distribution of the variable. One of the most used methods is called hot deck. It consists of replacing each missing value with a random extraction of observed values.
- Conditional Mean: a regression model to predict Y from X. This method works particularly well for a limited class of problems but is, instead, discouraged when analyzing covariances and correlations because it inflates correlations between Y and X.
- Conditional Distribution: known as random regression imputation. Using a linear regression model as a starting point, one could add a random error to the estimate $\hat{Y}$. This method has the main advantage of keeping intact the distributions of variables and allowing the estimation of other quantities relative to the distribution.
- Nearest-Neighbour Imputation: based on donation, like hot deck, the donor is chosen through the minimization of a certain distance measure, function of auxiliary variables. The observed unit with the lowest distance relative to the non-responding unit is identified and its value is substituted for the missing unit for the variable of interest.

Exercise: Diagnosing and Handling Missing Data in a Realistic Scenario

Missing Data Analysis & Simple Imputation
Sample Dataset:

City	Temperature	PM2.5	Population
Milan	22	15	1350000
Rome	55	20	2870000
Naples	27		960000
Turin	21	18	
Toronto	10		
Lecce	25		
Florence	23	17	380000

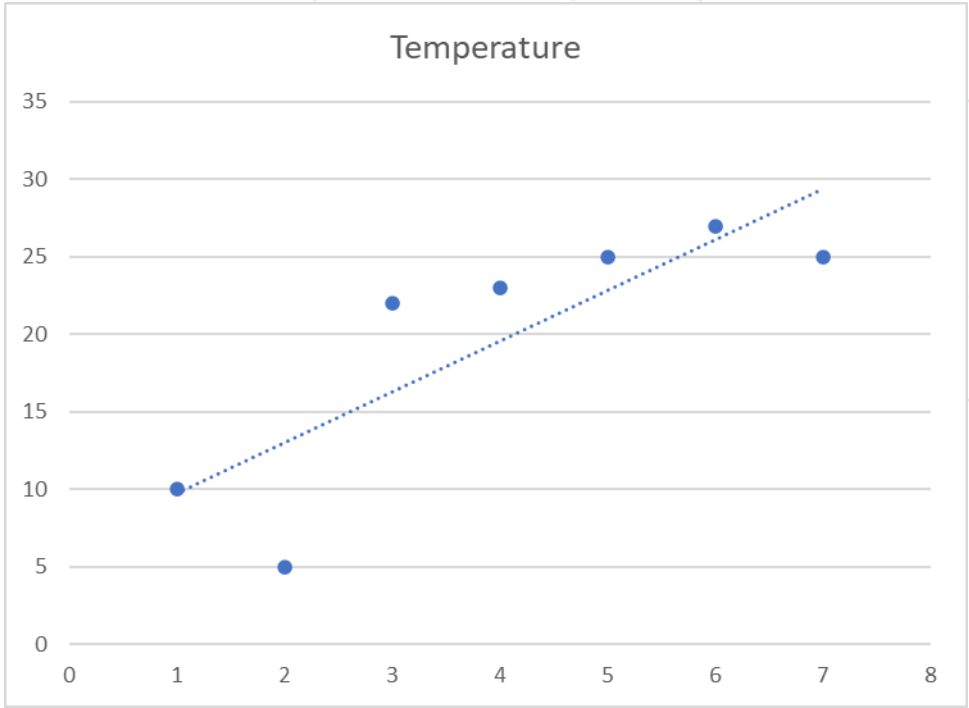
Task 1: Identify Missing Values and outliers
Task 2: evaluate randomness
Tak 3: impute

Missing Data

Exercise: Diagnosing and Handling Missing Data in a Realistic Scenario

Task 1: Identify Missing Values and outliers
Task 2: evaluate randomness
Tak 3: impute

City	Temperature	PM2.5	PM2.5	Population	
Toronto	10	2	13.0		1
Ottawa	5	2	13.0		1
Milan	22	15	15	1350000	0
Florence	23	17	17	380000	0
Lecce	25	17.3	13.0		1
Naples	27	17.3	13.0	960000	0
Rome	25	20	20	2870000	0
Number of missing values	0	0		3	
Unconditional Mean	19.6	13.0			
Conditional Mean|Italy	23.3	17.3			
Conditional Mean|Canada		13.0			



Application on air pollution data

- The goal of this work is to analyse urban pollution and its determinants.
- The final dataset consists of 61 variables and 230 total observations, referring to cities from various OECD countries.
- These countries include: Australia, Austria, Belgium, Bulgaria, Canada, Switzerland, Colombia, Germany, Denmark, Estonia, Spain, Finland, France, Great Britain, Croatia, Hungary, Ireland, Italy, Japan, Korea, Lithuania, Mexico, Netherlands, Norway, New Zealand, Poland, Portugal, Sweden, Turkey, USA.

Step 1: Data Collection

- Climate data was obtained from the “aqicn” website, an organization behind the World Air Quality Index (since 2007), gathering data from over 30,000 stations in 130 countries.
- https://aqicn.org/data-platform/covid19/
- Variables collected per city include:
  - Meteorological: precipitation, temperature, humidity, wind speed/gust, pressure
  - Pollutants: SO2, NO2, O3, CO, PM10, PM2.5
  - Only January data from 2018 to 2022 was used and averaged per city.
  - Socioeconomic data was downloaded from OECD databases, focused on 2018.

Step 2: Merging Datasets
Socioeconomic data was downloaded from OECD databases-2018

Cities appearing in both datasets were matched and merged into an initial “Original DB” file.

Variable	Label
Average population size of local governments	APS
Elderly dependency ratio (65+ over 15–64)	ED_R
Elderly population (65+)	EPG
Employment rate (15–64 over total 15–64)	EMP_R
GDP (Million USD, 2015 constant prices and PPP)	GDP
Metropolitan area GDP share of national GDP	GDP_MA
GDP per capita	GDP_PC
Labour force (15–64)	LF
Participation rate	PART_R
Population density (inhabitants per km^2)	PD
Metropolitan population share of national population	POP_MA
Total population	POP
Territorial fragmentation	TF
Unemployment (15+)	UNEMP
Unemployment rate (15–64)	UNEMP_R
Working age population (15–64)	WP
Youth dependency ratio (0–14 over 15–64)	YD_R
Youth population (0–14)	YP

Further steps

Step 3: Socioeconomic Imputation

- Missing socioeconomic data was filled using OECD and Eurostat values:
- Data from the same city and year was prioritized.
- If unavailable, data from the same region or nearest year (±2 years) was used.

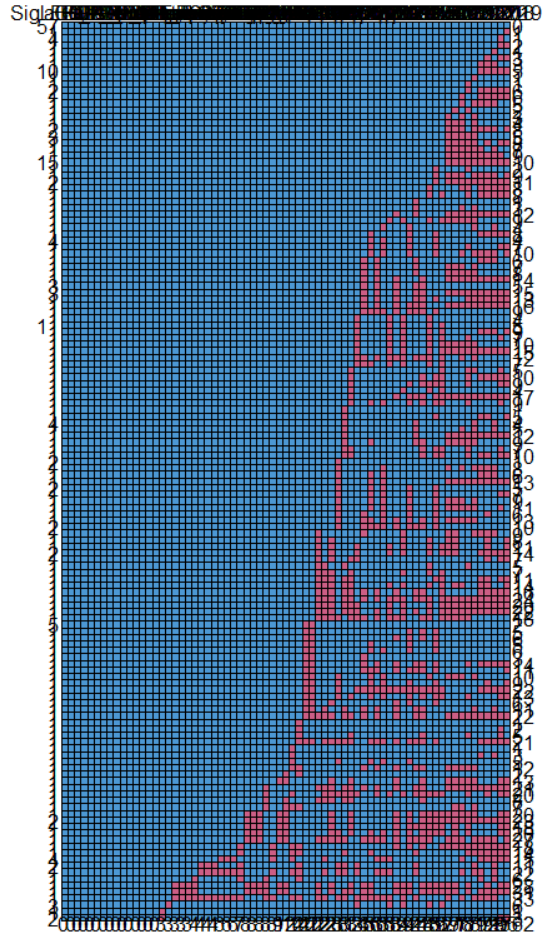
Step 4: Cleaning

- Variables with too many missing values were removed.
- The “Complete DB” was created, and a missing value analysis was conducted.

Visual and statistical analysis (R packages)

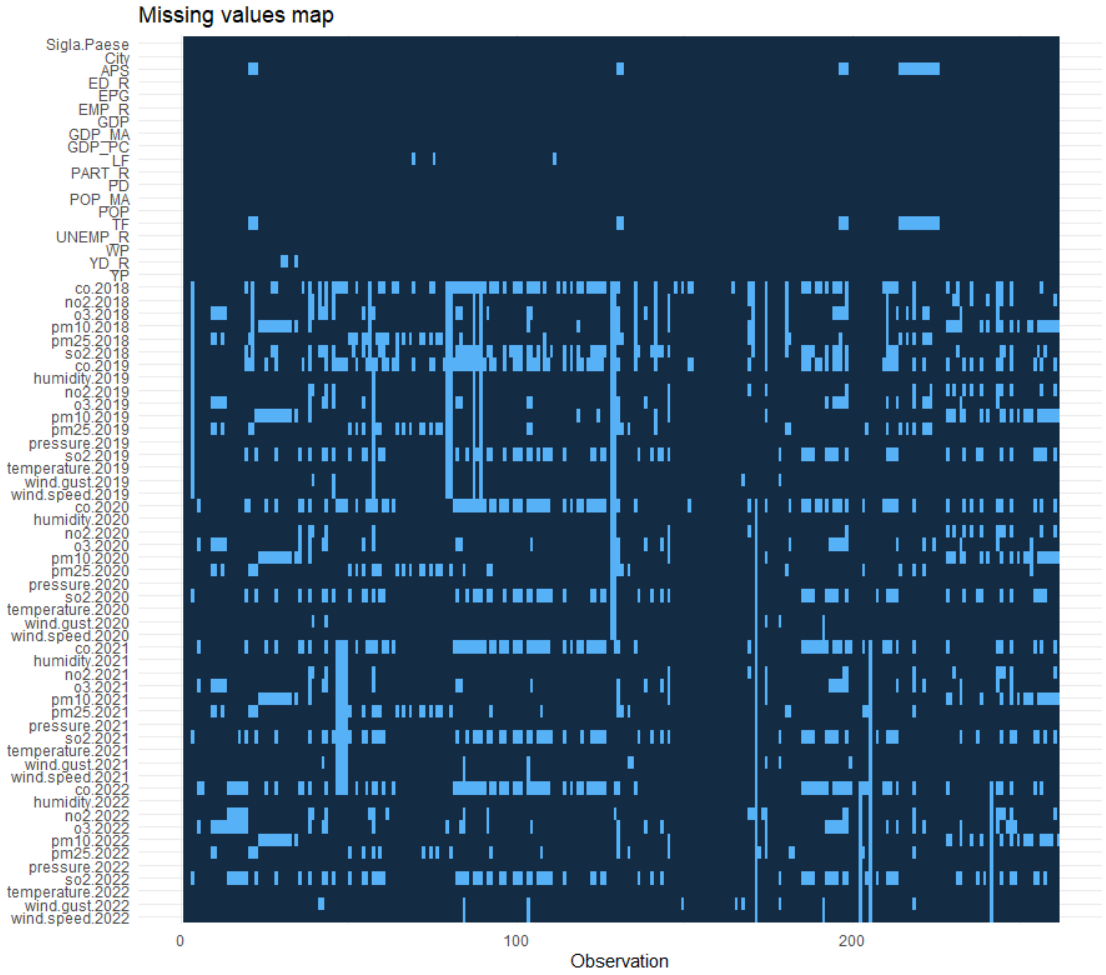
MICE, Finalfit, Amelia, Visdat, VIM, ggplot2) were used.
Visualizations suggested about 10% missing data, mostly MAR.

The MICE package (Figure) allows us to understand how many observations are missing 0, 1, 2, etc., data points, and in which variables these missing values occur.



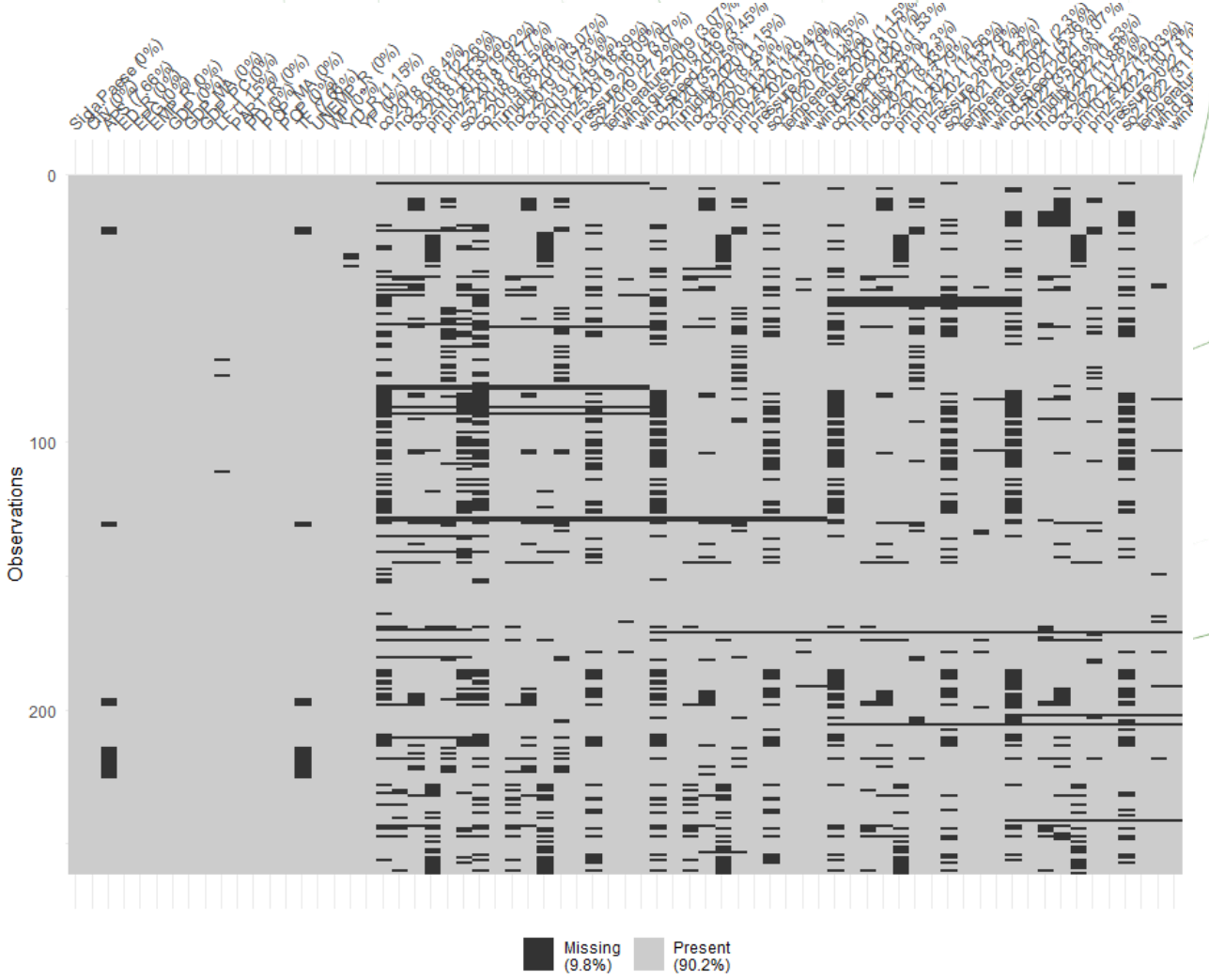
FinalFit Missing Visualisation

The Finalfit package (Figure) shows where each variable has missing information.
So, while the previous graph displayed the distribution of data by frequency, here we have a simple graphical visualization of the table imported into R.



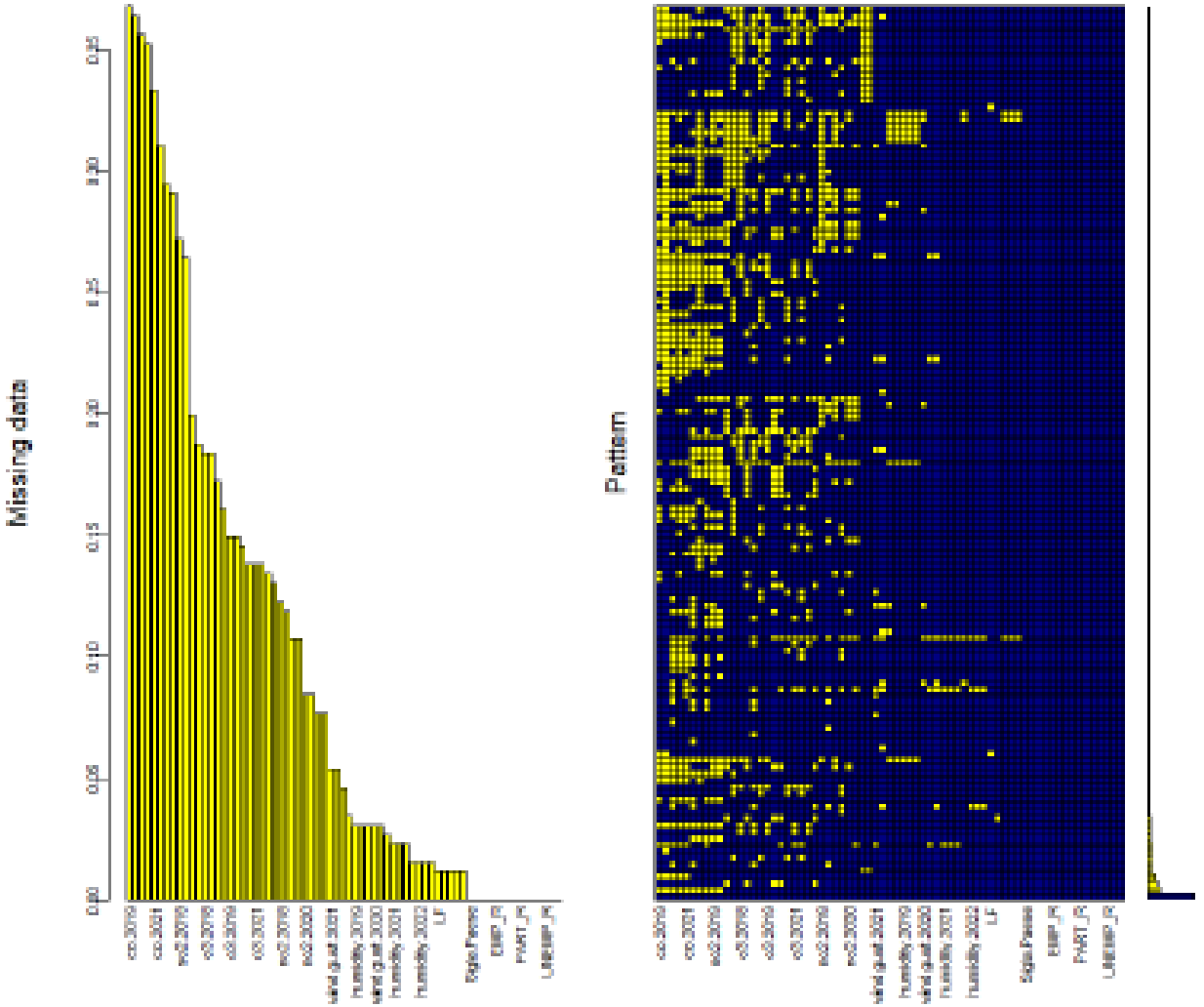
Visdat Missing Visualisation

The Visdat package (Figure) provides a simple graphical visualization of the table imported into R, but in addition to the previous one, it also shows the percentage of missing data and observed data.



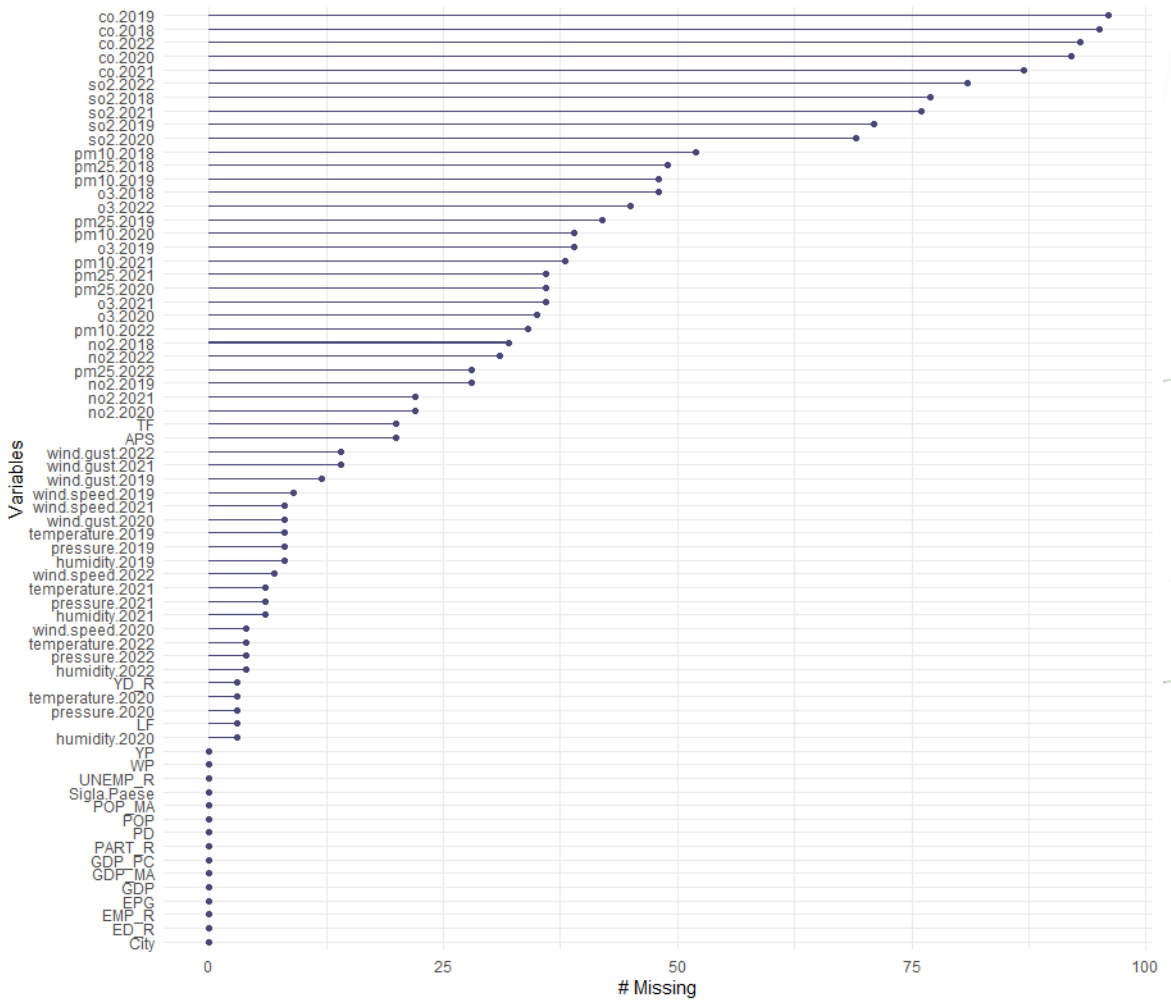
Visualization with the VIM package

- the visualization with the VIM package is slightly different (Figure), which outputs two graphs:
- one on the left, where the variables are ordered in descending order based on the number of missing observations;
- the second on the right, which shows the general pattern of the table—following the same principle as Finalfit—although the variables are placed on the x-axis this time and possibly in a different order.



Visualisation ggplot

- Number of missing values by variable



Algorithm

- Cities or variables with low data coverage were filtered out.
- Cities with ≥21% missing data were removed.
- Final dataset: 61 variables, 230 cities, 6% missing data.
- Final test confirmed MAR nature of missing data using Little's MCAR test.
- Then, we imputed our data with the following:

Algorithm

1. Load the dataset with 6% missing data
2. Impute the socioeconomic variables using the mice function with the **"mean"** method
3. Impute the other variables using the mice function with the **"lasso.norm"** method
4. Display the new complete dataset and verify the absence of missing values
5. Export the complete and imputed dataset: **DB_final**

MICE (Multiple Imputation by Chained Equations)

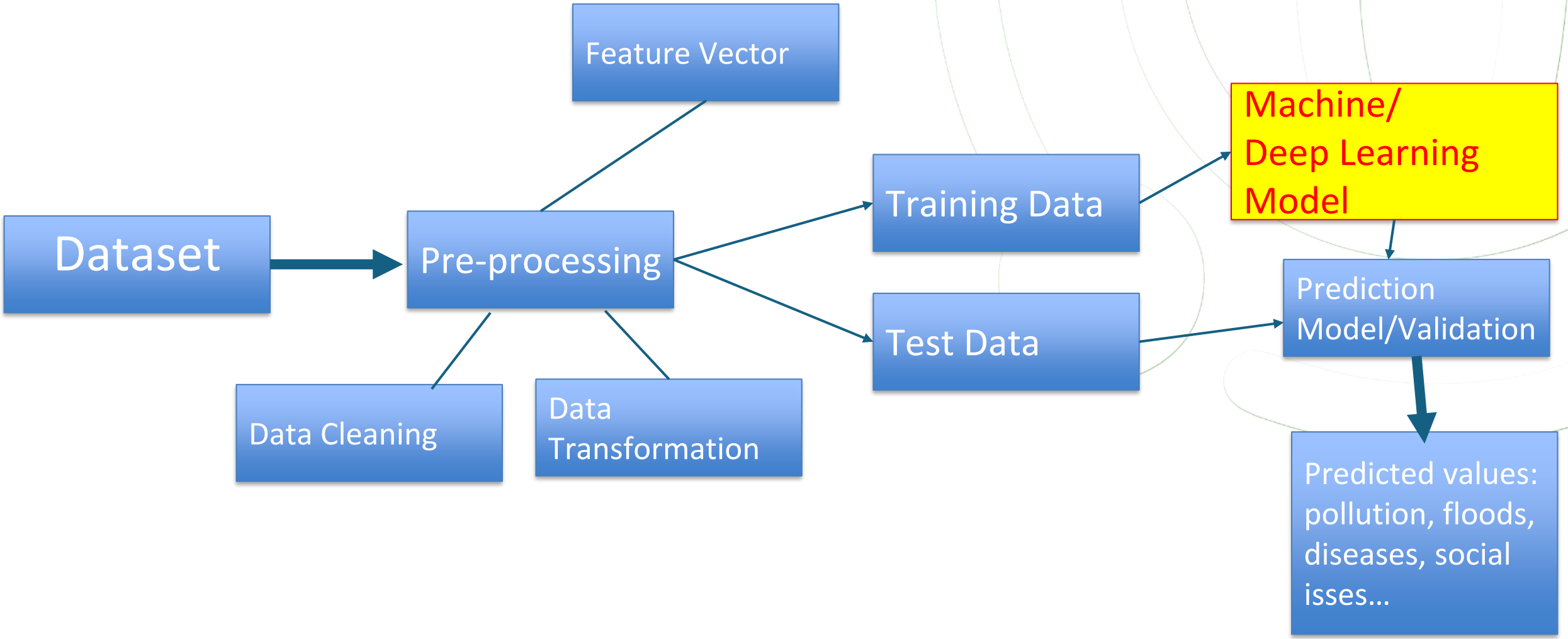
- MICE operates in 5 steps:
- Temporarily replace missing values using means.
- Reinsert missing values in one variable.
- Predict missing values via regression using other variables.Replace missing values.
- Repeat steps 2–4 for each variable, over m cycles (usually m = 5).
- This study used lasso.norm regression, ideal for handling multicollinearity in socioeconomic variables (e.g., population and population density).
- LASSO Regression = Least Absolute Shrinkage and Selection Operator
  - A linear regression method that shrinks some coefficients to 0, improving model performance by eliminating irrelevant predictors. Effective in the presence of multicollinearity.

Possible R code

```
# Load required libraries
library(naniar)     # for missing data summaries
library(VIM)        # for visualization
library(mice)       # for multiple imputation
library(ggplot2)    # for visualizations
library(dplyr)      # for data manipulation

# Step 1: Load a sample dataset with missing data (you can replace with your own)
data("airquality")   # Sample dataset with missing values
df <- airquality

# Add some artificial categorical data for the sake of the example
df$CityGroup <- sample(c("Low income", "High income", NA), size = nrow(df), replace = TRUE, prob = c(0.4, 0.5, 0.1))

# Step 2: Explore missingness
summary(df)
vis_miss(df)                    # General pattern
gg_miss_var(df)                 # Missing % per variable
aggr(df, col=c('skyblue','orange'),     # VIM plot
     numbers=TRUE, sortVars=TRUE,
     labels=names(df), cex.axis=.7, gap=3)

# Step 3: Run MCAR test (Little's test via mice)
mcar_test_result <- mice::mcar(df)
print(mcar_test_result)

# Step 4: Define methods for imputation
# We'll use mean for Wind (as an example), and lasso.norm for others

init <- mice(df, maxit = 0)
methods <- init$method

# Customize methods: use mean for one, lasso.norm for others (only numerical variables allowed here)
methods["Wind"] <- "mean"
methods["Ozone"] <- "lasso.norm"
methods["Solar.R"] <- "lasso.norm"
methods["Temp"] <- "lasso.norm"
methods["CityGroup"] <- "polyreg"   # multinomial logistic regression for categorical

# Step 5: Run MICE imputation with 5 iterations
imputed_data <- mice(df, method = methods, m = 5, seed = 123)

# Step 6: Check convergence
plot(imputed_data)

# Step 7: Extract completed dataset (first of 5)
complete_data <- complete(imputed_data, 1)

# Step 8: Export or inspect
View(complete_data)
write.csv(complete_data, "imputed_dataset.csv", row.names = FALSE)
```

mice::mcar() performs a test of Missing Completely at Random (MCAR). lasso.norm is useful when multicollinearity is suspected in numeric predictors. polyreg is used for categorical variables with more than 2 levels .The number of imputations (m = 5) follows Rubin's rules for balance between performance and stability.

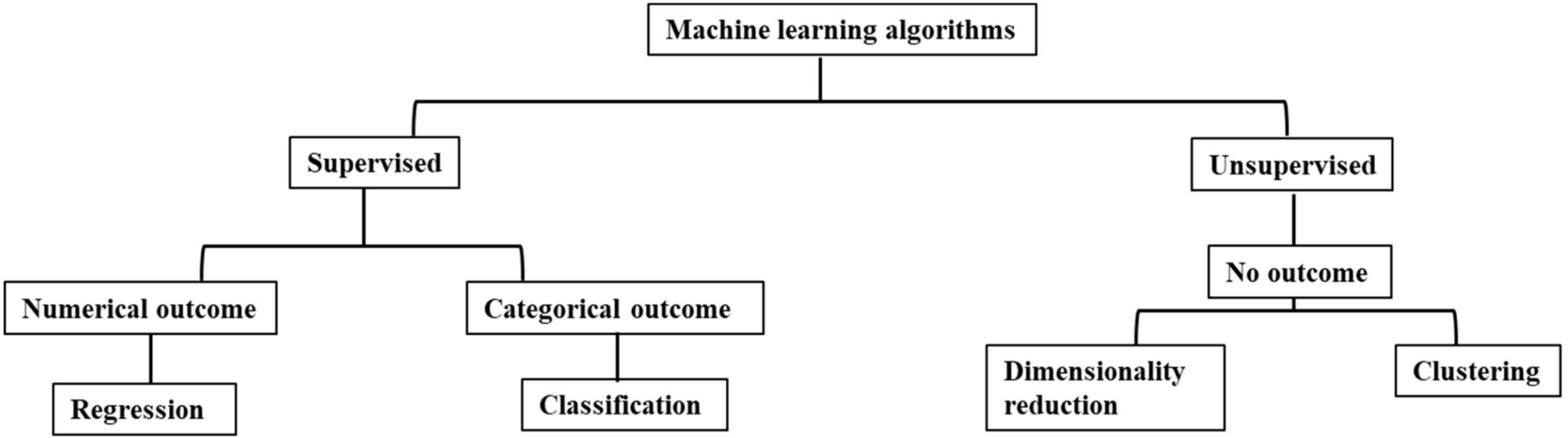
How do we actually perform a Machine Learning project?



Machine Learning

ML algorithms have three main subcategories (supervised, unsupervised, and reinforcement learning)

- Unsupervised ML involves analyzing and clustering unlabeled datasets to uncover hidden patterns or groupings. Its ability to identify similarities and differences in data makes it particularly useful for exploratory data analysis. Unsupervised ML models are employed for clustering and dimensionality reduction tasks.
- Supervised ML uses a training set containing input–output pairs to teach the algorithm to learn a function that maps inputs to outputs.



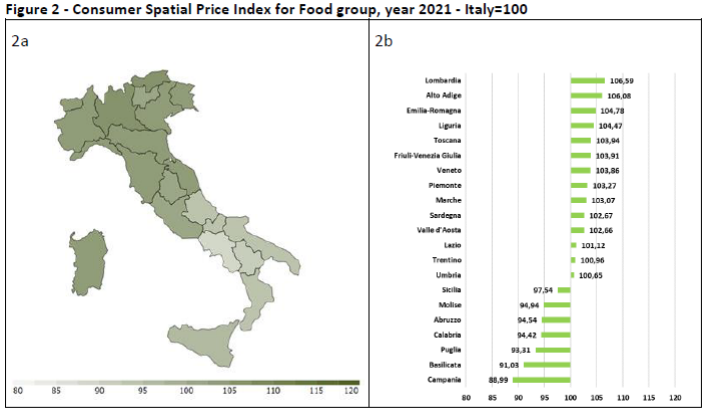
Real world applications: official statistics

Consumer Spatial Price Indices

The sources used for compiling these new indicators are mainly those of the consumer price survey complemented by surveys carried out for the specific purpose of calculating the regional spatial indices. In summary, the three data sources are:

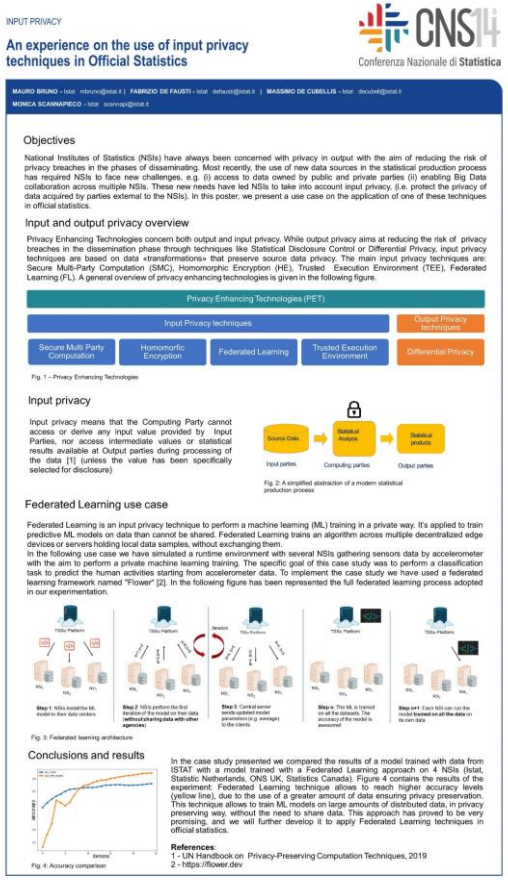
- **Scanner data**. A unique identifier (bar code) characterizes each product, therefore the comparability in space is guaranteed. Information on turnover and quantity allows to calculate the unit value (average price) for each bar code and to weigh it, fully guaranteeing the principle of representativeness.
- **CPI (Consumer Price Index) data**. For some product categories (fresh fish, fruit, and vegetables), the definitions in the traditional CPI data collection are detailed enough to allow the use of these data in compliance with comparability. The products of these categories included in the consumer price basket are widely distributed throughout the country.

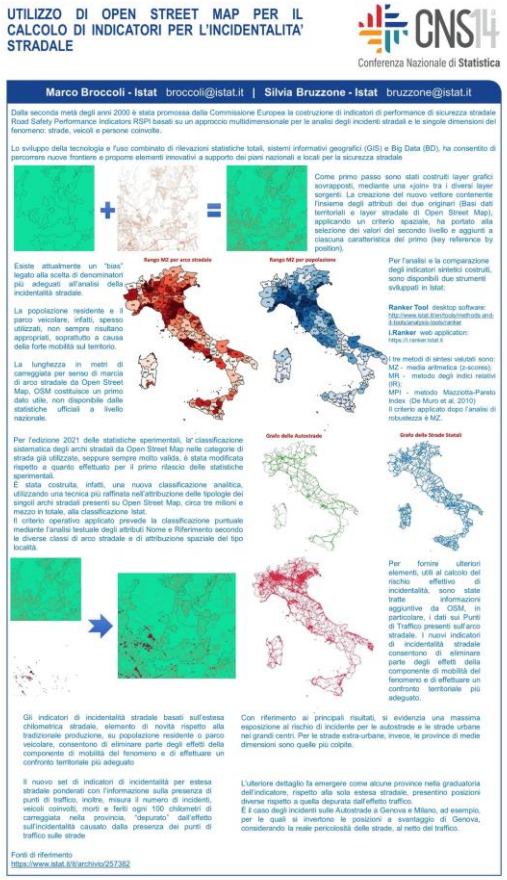
Figure 2 - Consumer Spatial Price Index for Food group, year 2021 - Italy=100

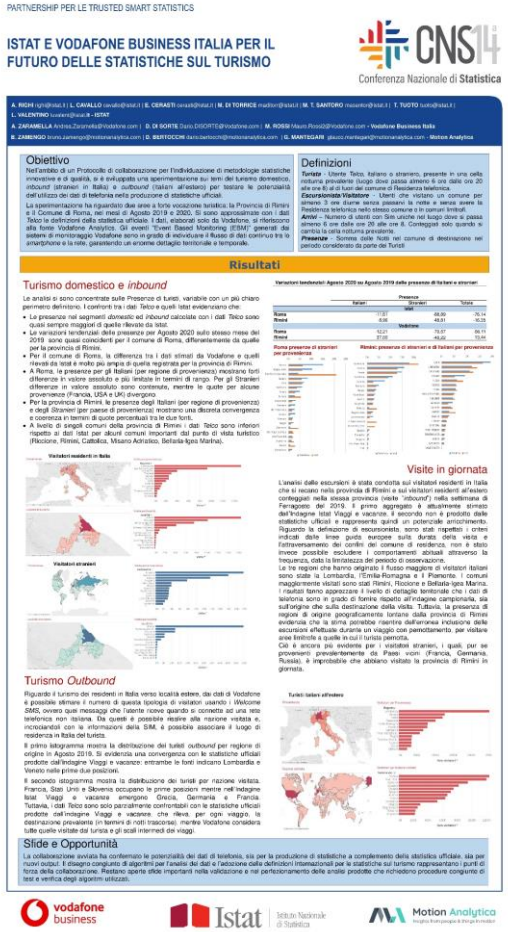


Source: Istat

Real world applications: official statistics







Real world applications: official statistics: social mood on economy Index

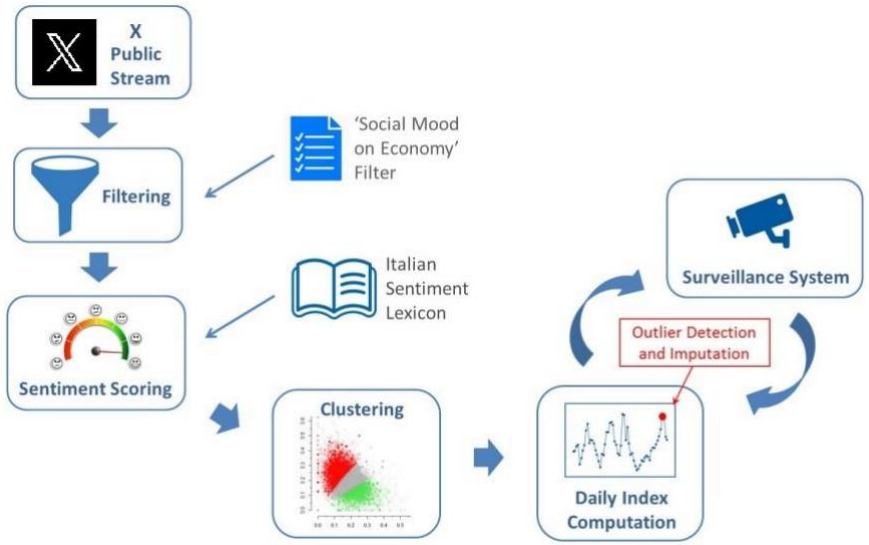
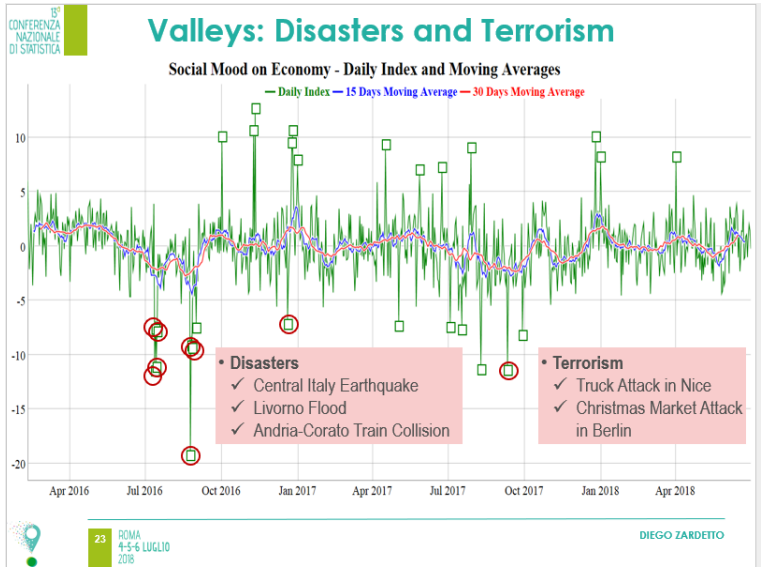
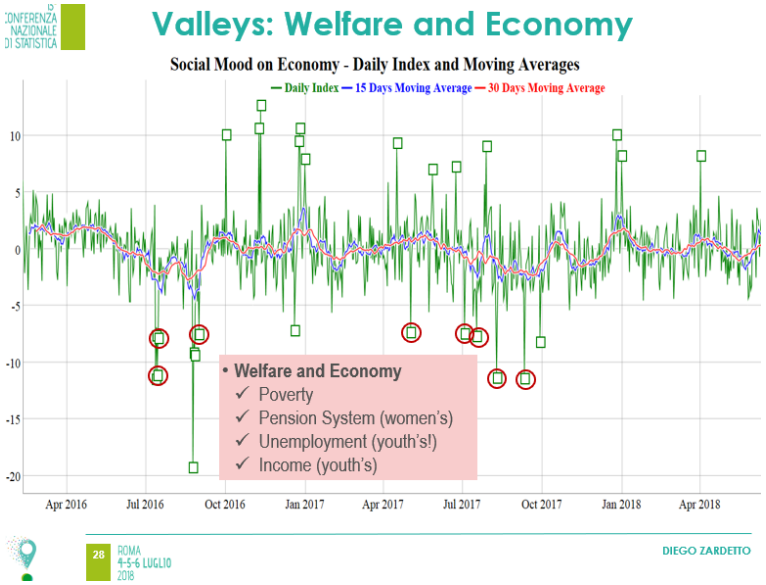


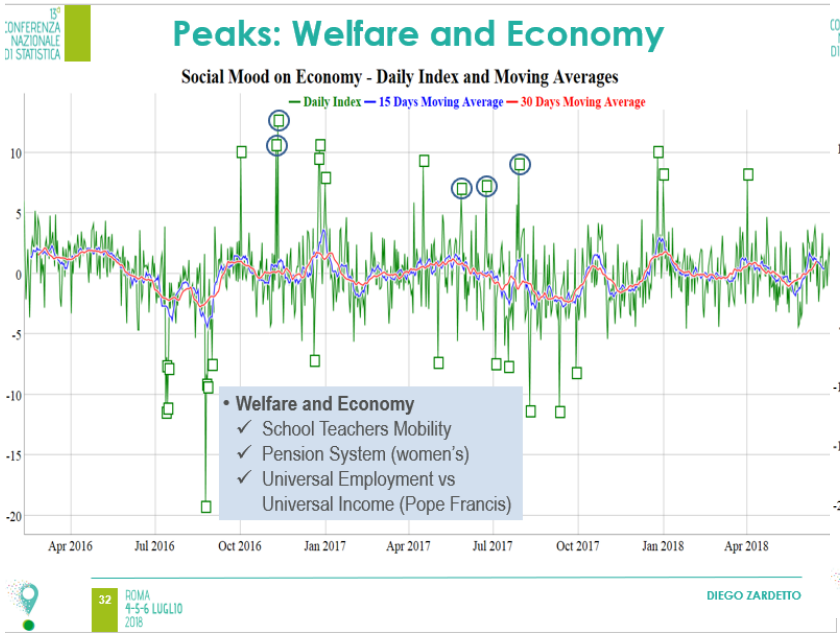
Figure 1: A schematic representation of the processing pipeline of the Social Mood on Economy Index

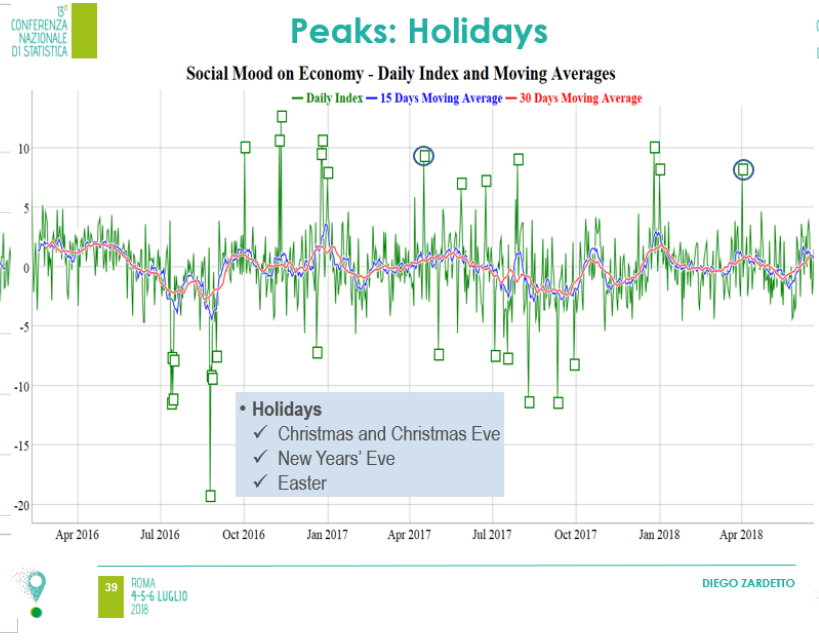
https://www.istat.it/wp-content/uploads/2024/05/Methodological_Note.pdf

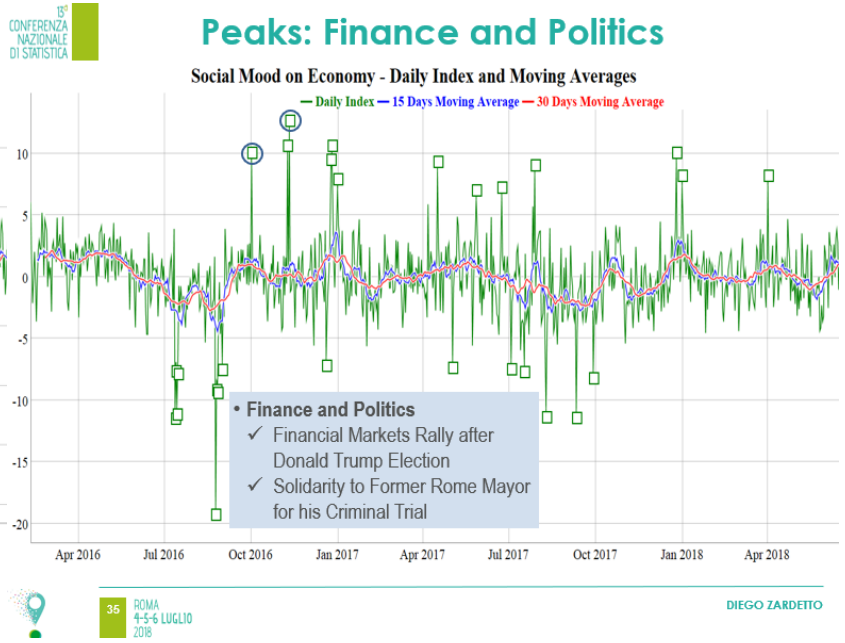




Real world applications: official statistics: social mood on economy Index

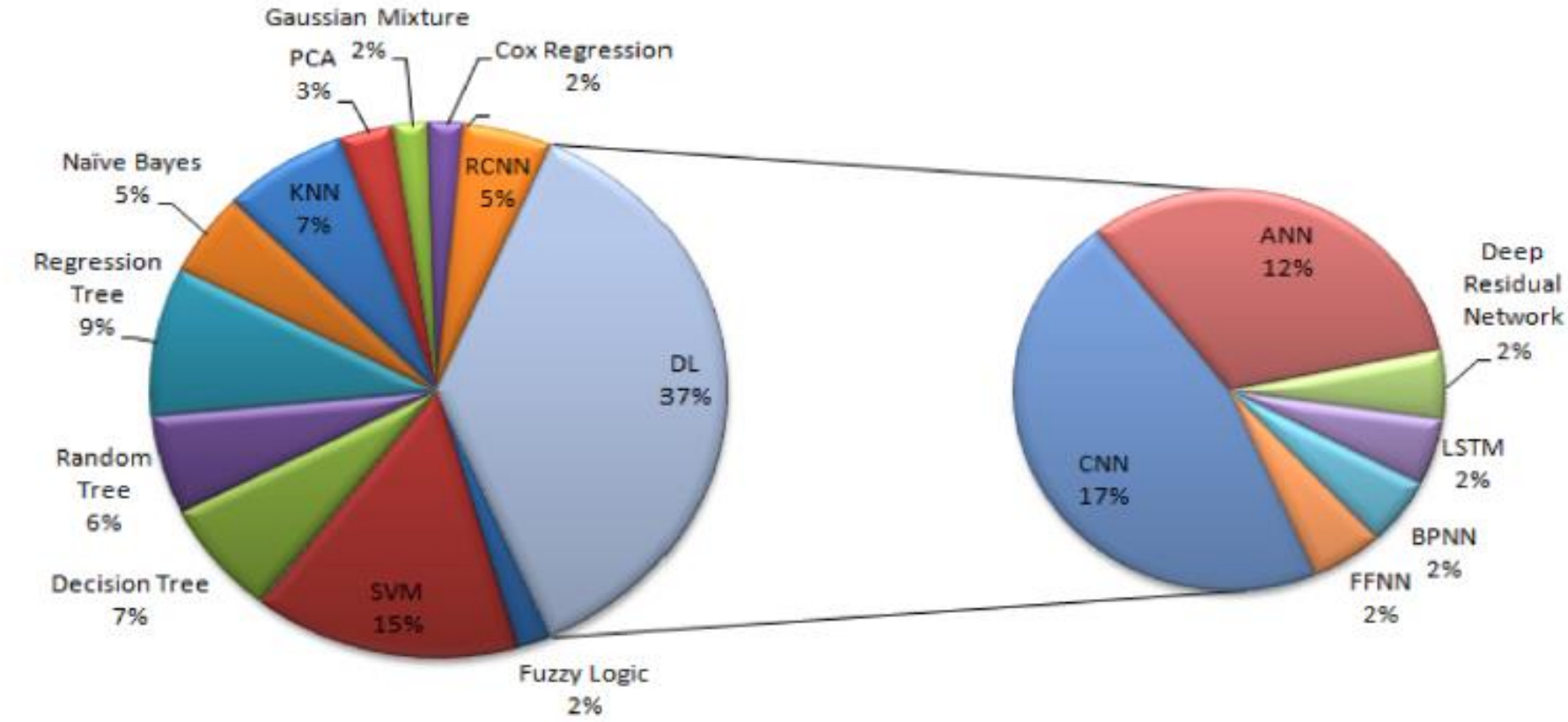






https://www.istat.it/wp-content/uploads/2024/05/Methodological_Note.pdf

Most employed ML methods



Kumar, Y., Koul, A., Singla, R. *et al.* Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Human Comput* **14**, 8459–8486 (2023). https://doi.org/10.1007/s12652-021-03612-zc

MACHINE LEARNING TECHNIQUES APPLICATIONS IN ENVIRONMENTAL SCIENCE

- Supervised learning
- Unsupervised leaarning
- Reinforcement learning

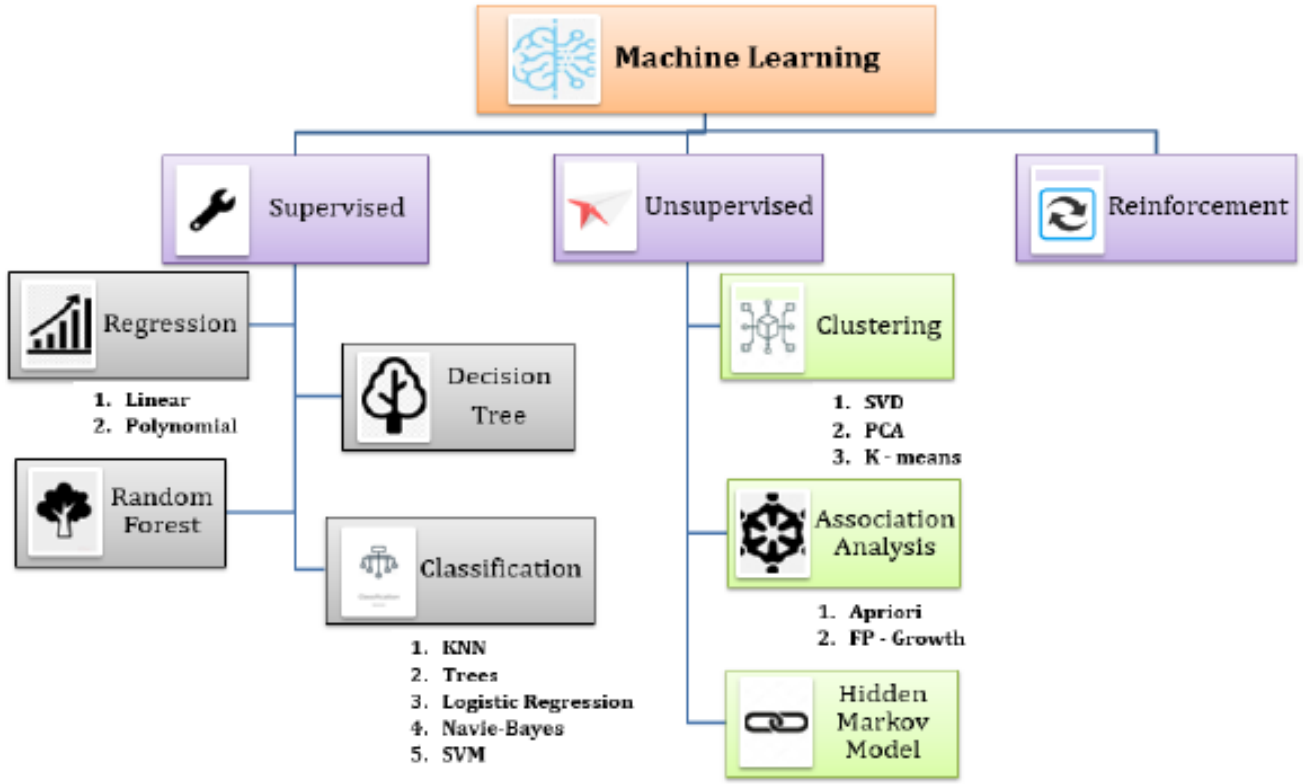


Figure 1. Categories of Machine Learning algorithms

More in detail:

- Let’s see what famous references tell us……

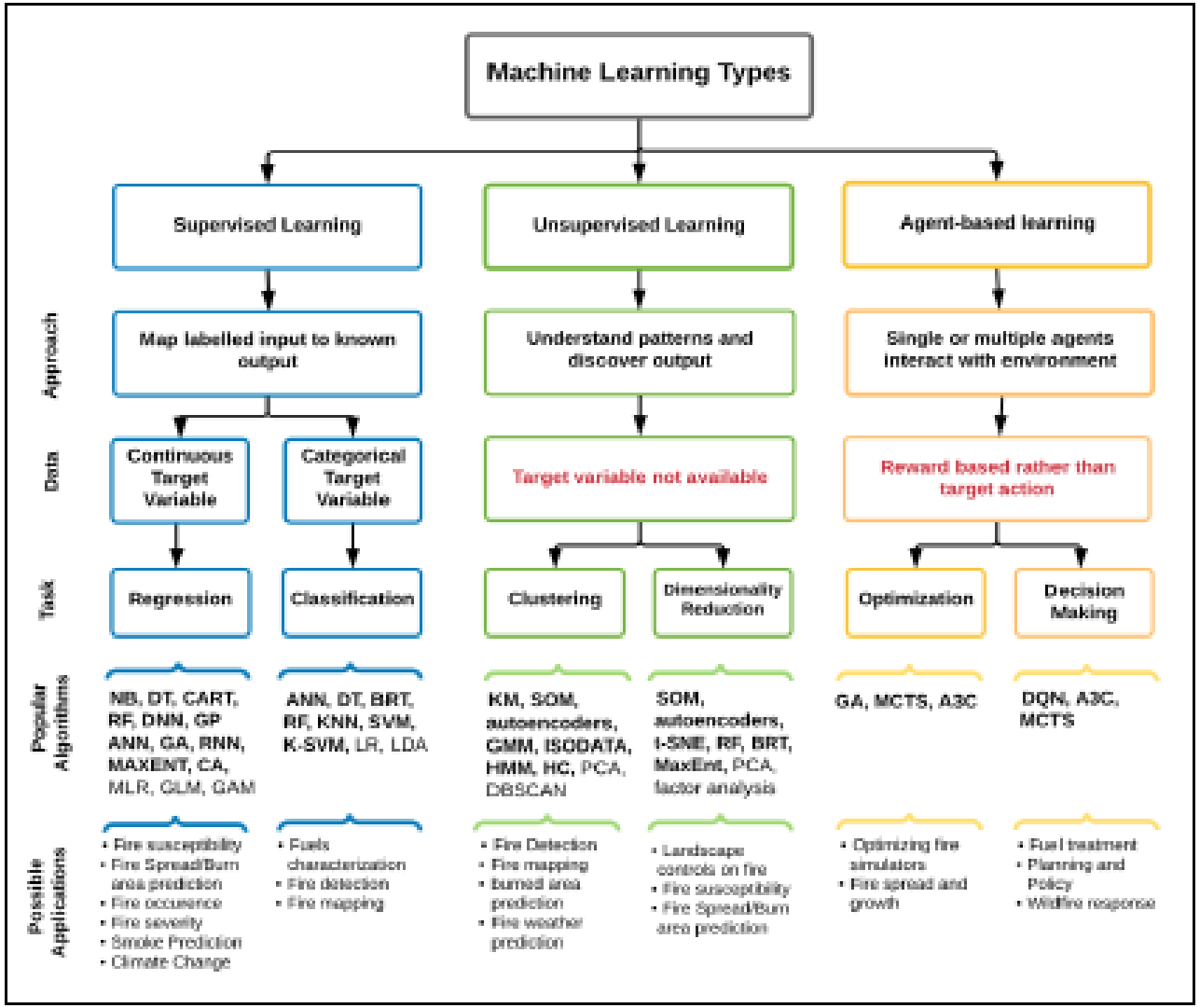


Fig 5: *diagram depicting the primary types of machine learning, data types, and modeling tasks, highlighting their associations with widely used algorithms and applications in wildfire science and management. Algorithms in bold indicate core ML methods, whereas non-bolded algorithms are generally not classified as core ML (Piyush et al., 2020).*

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani