1.1. Metodología

Desde el punto de vista metodológico, se ha desarrollado una metodología de trabajo común para todos los casos de uso, dado que todos ellos se ajustan a un desarrollo que puede enmarcarse bajo el paraguas de lo que de manera estandarizada se ha venido a denominar como la disciplina de la Ciencia de Datos o en su término anglosajón, más extendido, el Data Science, que está íntimamente ligado con la Estadística y la Inteligencia Artificial, y sus técnicas más conocidas, como las de Machine Learning, y dentro de estas las Redes Neuronales Artificiales, y la subclase de estas más popular, los modelos de Deep Learning, entre otras.

En este sentido, se ha definido un flujo de trabajo que cubre los siguientes dominios y procesos (se utiliza una terminología inglesa dado que muchos términos no se traducen directamente y es común trabajar en esta disciplina con esta terminología).

Data Science Workflow

I – Business Problem Domain

1. Domain Knowledge
2. Data-Driven Approach
3. Data Science Approach
4. Analytics Approach

II – Data Processing Domain

5. Data Collection
6. Data Adequacy
7. Data Sampling
8. Data Split (Train, Validation, Test)
9. Data Cleansing (Cleaning or Scrubbing)
10. Data Balancing Analysis
11. Exploratory Causal Analysis (ECA)
12. Exploratory Data Analysis (EDA)
13. Data Visualization

III – Features Engineering Domain

14. Feature Data Transform
15. Feature Importance
16. Feature Selection
17. Feature Extraction (Dimensional Reduction)
18. Feature Construction
19. Feature Transforms (Polynomial Features Transform)
20. Feature Learning

IV – Model Development Domain

21. Model Spot Checking
22. Model Evaluation
23. Model Selection
24. Model Tuning
25. Model Combination
26. Model Calibration
27. Model Uncertainty Analysis
28. Model Bias-Variance Trade-off Analysis
29. Model Interpretation
30. Model Prediction
31. Model Finalization
32. Model Saving

V – Model Operation Domain

33. Model Deployment
34. Model Execution
35. Model Analysis
36. Model Updating
37. Model Explainability

Domains Description

I – Problem

Business Problem Domain is the set of processes that allow us to understand and categorize the problem in an adequate way, in order to be able to use and select the following domains with their processes in a logical and correct way.

This domain will allow us to define the problem (knowledge representation), understand it in terms of a Data-Driven and the required approach (data modeling or data discovery), its viability, the types of analysis required (descriptive, predictive, ...), the typology of Data Sciences that describe the problem (classification, regression, ...), and so on. All this will allow to select in the following domains the set of processes, techniques, and algorithms, suitable for solving the problem.

II – Data Process

Data Process Domain is the set of processes that allows to obtain the necessary, adequate, and pertinent data set for the design and development of subsequent models.

The processes of this domain allow the obtaining, generation (if required), cleaning, adaptation, balancing, analysis, understanding, and appropriate division (train, test and validation) of the data.

III – Feat Enginee

Feature Engineering Domain is the set of processes that allows to obtain the set of characteristics (variables) that best represent the problem, and that allow obtaining the best model.

This domain should be used in order to analyze all the original characteristics available in order to obtain information about the most appropriate set of them, and then using transforming, selecting, building or extracting procedures, obtain the final set of characteristics.

IV – Model Dev

Model Development Domain is the set of processes that allows to obtain the best model (or set of models) to be implemented in order to solve the problem defined, from data and characteristics obtained in the previous processes.

V – Model Operat

Model Operation Domain is the set of processes that allow to deploy and execute models defined.

This domain covers the processes related to the analysis of the results of the models at runtime, especially the analysis related to the underspecification problem, as well as the necessary adaptations to mitigate said problem.

Processes Description

I – Problem

Knowledge

Domain Knowledge is the process to acquire and represent the knowledge about the environment in which the data is processed to reveal secrets of the data.

Data-Driven

Data-Driven is the process to approximate the problem in terms of data, in order to explain how to solve it from a data analysis.

Data Science

Data Science Approach Process refers about what types of approaches are required to approximate the problem in terms of data science techniques.

Analytics

Analytics Approach Process refers about what kind of analysis is required: Statistical, Descriptive, Diagnostic, Predictive, Prescriptive, etcetera.

II – Data Process

Data Collection

Data Collection is the process to obtain and generate (if required) necessary data to model the problem.

Data Adequacy

Data Adequacy is the process to adapt basic and fundamental aspects of the raw data (File Format, Data Separator, Feature Names, Tidy Data, etcetera).

Data Sampling

Data Sampling is the process to obtain a subset with the appropriate size of data (examples/rows, not features/columns) for analysis, maintaining the same statistical significance as raw dataset.

Data Split

Data Split is the process of selecting the appropriate division of the data set into train, test and validation set.

Data Cleansing

Data Cleaning refers to identifying and correcting (or removing) errors in the dataset that may negatively impact a predictive model, replacing, modifying, or deleting the dirty or coarse data.

Data Balancing

Data Balancing is the process to obtain an adequate data balance if is required, in order to have the adequate amount of data that reflects the intrinsic structure of the problem to be solved.

ECA

Exploratory or Root Cause Analysis (ECA) is the process of discovering the root causes of problems in order to identify appropriate solutions.

EDA

Exploratory Data Analysis (EDA) is the process to analyzing data sets to summarize their main characteristics.

Data Visualization

Data Visualization is the process of performing a statistical graphical analysis of the data.

III – Feat Enginee

- **Feat Data Transf** — Feature Data Transform is the process that allows change (if is required) the type and/or distribution of data features (e.g. scaling, normalizing o standardizing data features).
- **Feat Importance** — Feature Importance is the process that assigns scores to the input characteristics to a model, which indicate the relative importance of each characteristic, in order, for example, to be able to select the most important ones.
- **Feat Selection** — Feature Selection is the process where you automatically or manually select the most relevant features which contribute most to the correct output of the model.
- **Feat Extraction** — Feature Extraction is the process related to dimensionality reduction (or dimension reduction) that creates a projection of the data (high-dimensional space) resulting in entirely new input features (low-dimensional space), so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.
- **Feat Construction** — Feature Construction is the process related to create new features from your existing ones to improve model performance.
- **Feat Transform** — Feature Transform (Polynomial Features Transform) is the process to create new features by raising existing features to an exponent, in order to see if they improve model performance, when the input features interact in unexpected and often nonlinear ways.
- **Feat Learning** — Feature learning or representation learning is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data.

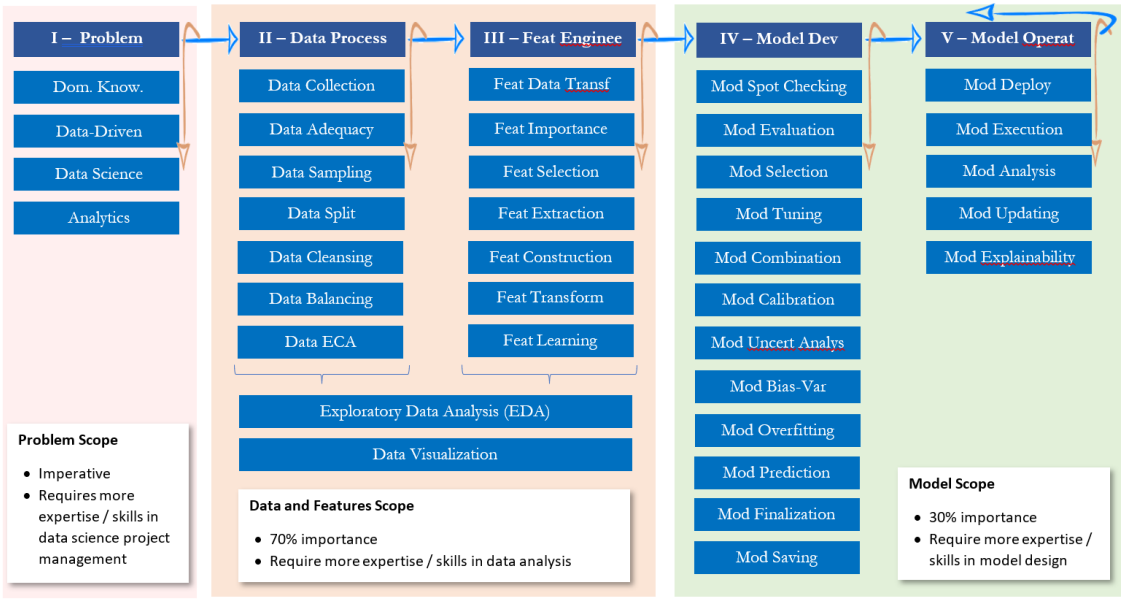
IV – Model Dev

- **Mod Spot Checking** — Spot Checking process is a set of techniques that allows to have a first approximation to know which algorithm will perform best on your data beforehand using a process of trial and error.
- **Mod Evaluation** — Model Evaluation is the process that allows us to evaluate the Models performance focused in the use of multiple Metrics and Resampling techniques.
- **Mod Selection** — Model Selection process is a set of techniques that allows to choose one from among a set of models, using Statistical Hypothesis Tests to address if the difference in skill between machine learning models is real, or due to a statistical chance; and the other hand, Probabilistic Statistical Measures to quantify both the model performance on the training dataset and the complexity of the model.
- **Mod Tuning** — Model Tuning process is a set of techniques that allows to optimizing hyperparameters of the model.
- **Mod Combination** — Model Combination process is a set of techniques that allows to merge or combine the outputs of the different models to obtain best results.
- **Mod Calibration** — Model Calibration process is a set of techniques that allows to calibrate the model when the probability estimate of a data point belonging to a class is very important. Calibration is comparison of the actual output and the expected output given by a system.
- **Mod Uncert Analys** — Model Uncertainty Analysis process is a set of techniques that allows to understanding why your model is uncertain and how to estimate there level of uncertainty.
- **Mod Bias-Var** — Model Bias-Variance Trade-off Analysis process is a set of techniques that allows to address the bias-variance dilemma or problem trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.
- **Mod Overfitting** — Model Overfitting Analysis process is a set of techniques that allows identify the overfitting as a possible cause of poor generalization performance of a predictive model.
- **Mod Prediction** — Model Prediction process allow get the output of the model for unseen data.
- **Mod Finalization** — Model Finalization process is a set of techniques that allows to finalize your machine learning model in order to make predictions on new data, training the final model selected on all available data.
- **Mod Saving** — Model Saving process allow takes a trained model and saves the entire transformation pipeline and trained model.

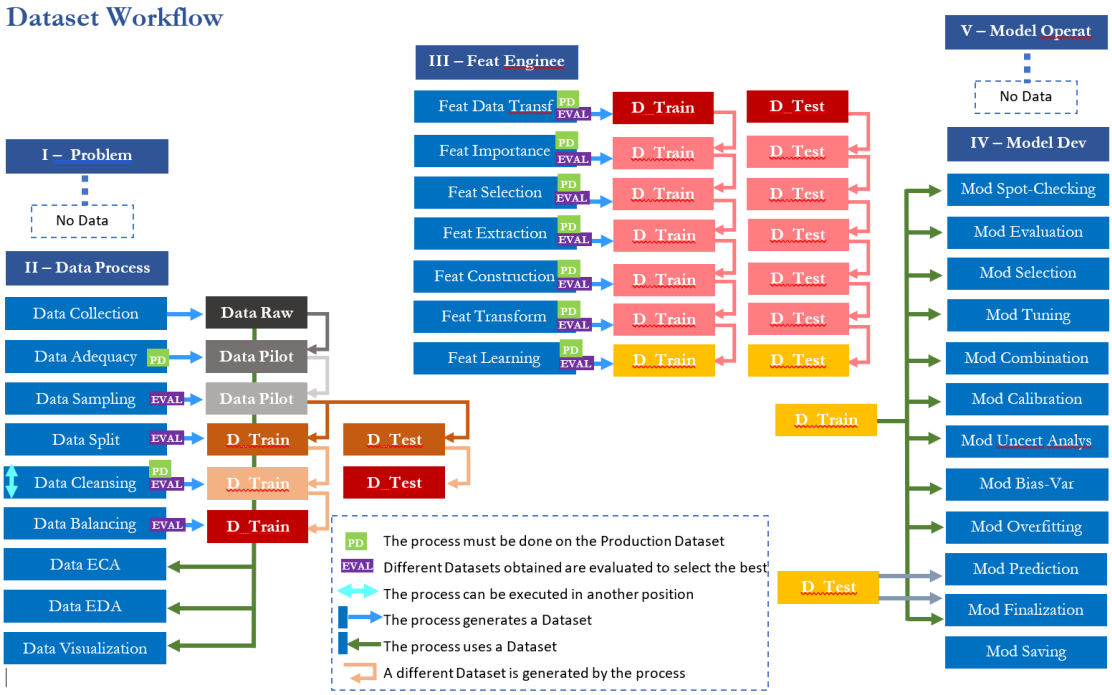
V – Model Operat

- **Mod Deploy** — Model Deploy process allows to integrate the model into an existing production environment to make practical business decisions based on data.
- **Mod Execution** — Model Execution process allows to execute the trained model with new unseen data in a production environment.
- **Mod Analysis** — Model Analysis process allows to perform an analysis in order to mitigate possible unexpectedly poor behavior when the models are deployed in real-world domains, because model or data drift, underspecification or uncertainty analysis.
- **Mod Updating** — Model Updating process allows to update the model in real time or production time, in order to adapt it dynamically to new data or changes in the problem domain.
- **Mod Explainability** — Model Explainability process seeks to understanding of why machine learning models make the decisions they do, and why it matters.

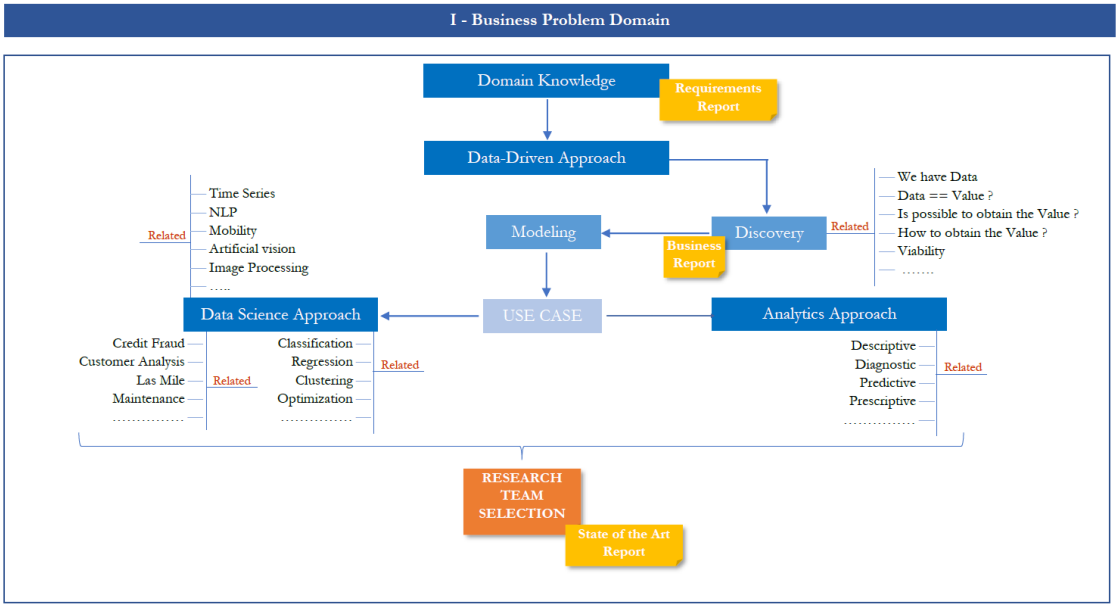
El flujo de trabajo a partir de estos dominios y procesos es el siguiente:

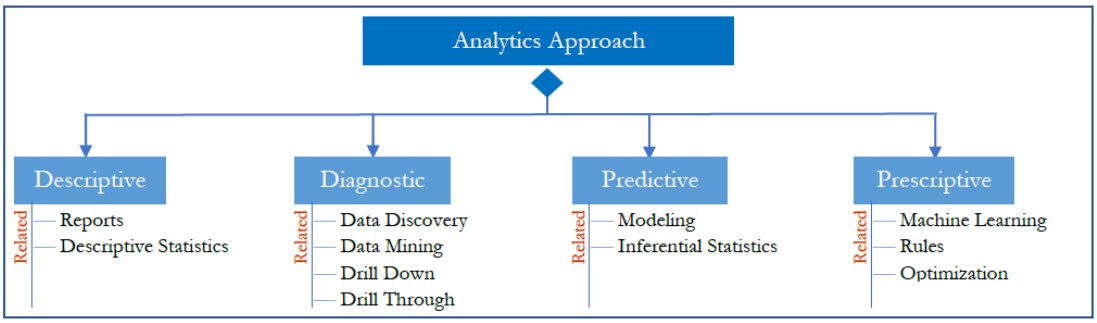


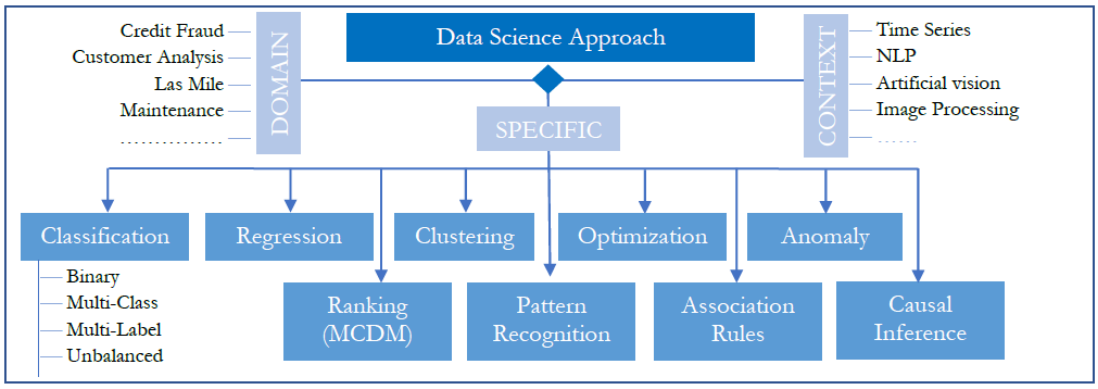
En referencia a los trabajos relacionados con los “datasets”, los flujos de trabajo generales se representan de la siguiente forma:

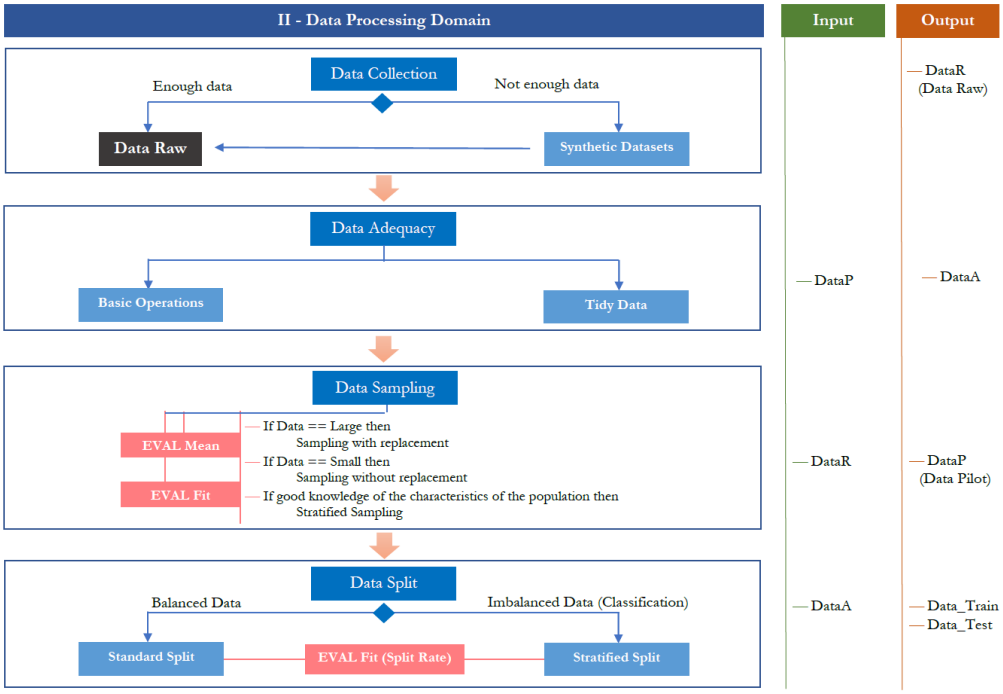


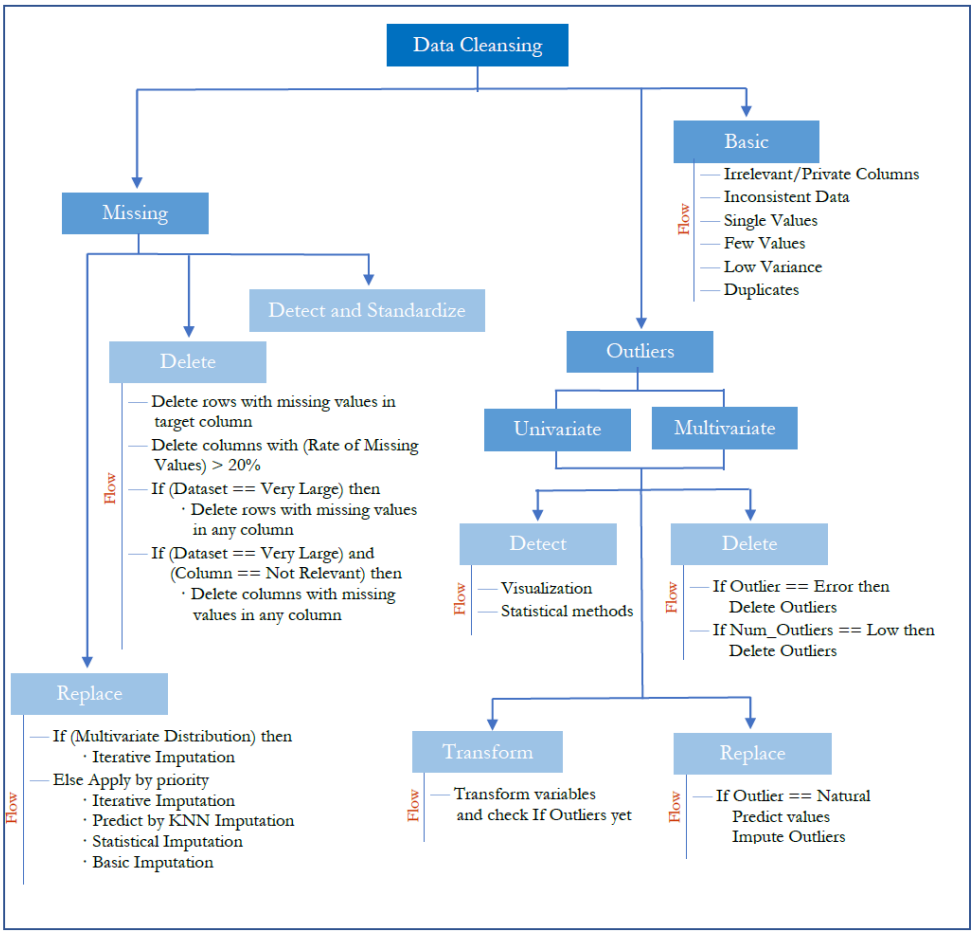
Algunos de los procesos, que requieren despliegue, se esquematizan a continuación:








II - Data Processing Domain
Input
Output
Data Collection
Enough data
Not enough data
Data Raw
Synthetic Datasets
DataR
(Data Raw)
Data Adequacy
Basic Operations
Tidy Data
DataP
DataA
Data Sampling
EVAL Mean
If Data == Large then
Sampling with replacement
If Data == Small then
Sampling without replacement
EVAL Fit
If good knowledge of the characteristics of the population then
Stratified Sampling
DataR
DataP
(Data Pilot)
Data Split
Balanced Data
Imbalanced Data (Classification)
Standard Split
EVAL Fit (Split Rate)
Stratified Split
DataA
Data_Train
Data_Test


Data Cleansing
Missing
Detect and Standardize
Delete
Flow
Delete rows with missing values in target column
Delete columns with (Rate of Missing Values) > 20%
If (Dataset == Very Large) then
· Delete rows with missing values in any column
If (Dataset == Very Large) and (Column == Not Relevant) then
· Delete columns with missing values in any column
Replace
Flow
If (Multivariate Distribution) then
· Iterative Imputation
Else Apply by priority
· Iterative Imputation
· Predict by KNN Imputation
· Statistical Imputation
· Basic Imputation
Basic
Flow
Irrelevant/Private Columns
Inconsistent Data
Single Values
Few Values
Low Variance
Duplicates
Outliers
Univariate
Multivariate
Detect
Flow
Visualization
Statistical methods
Delete
Flow
If Outlier == Error then
Delete Outliers
If Num_Outliers == Low then
Delete Outliers
Transform
Flow
Transform variables and check If Outliers yet
Replace
Flow
If Outlier == Natural
Predict values
Impute Outliers

Se definen en la metodología definida aspectos relativos a la mejora continua: