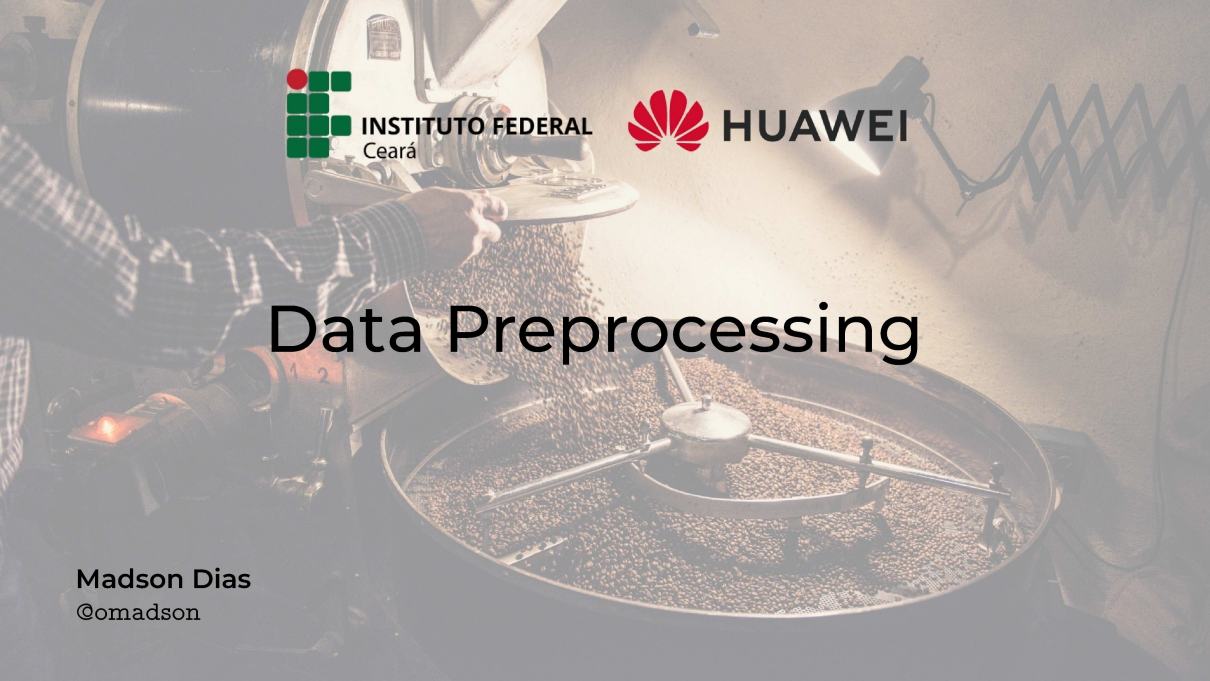


Data Preprocessing

Madson Dias
@omadson

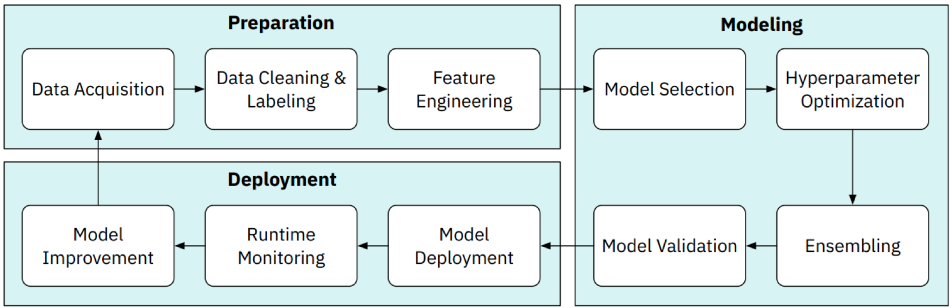
1. Data Science Workflow

2. Data Preprocessing
   - 2.1 Data sets, Features and Data Quality
   - 2.2 Feature encoding
   - 2.3 Missing Data
   - 2.4 Outliers
   - 2.5 Duplicate values
   - 2.6 Normalization

Data Science Workflow

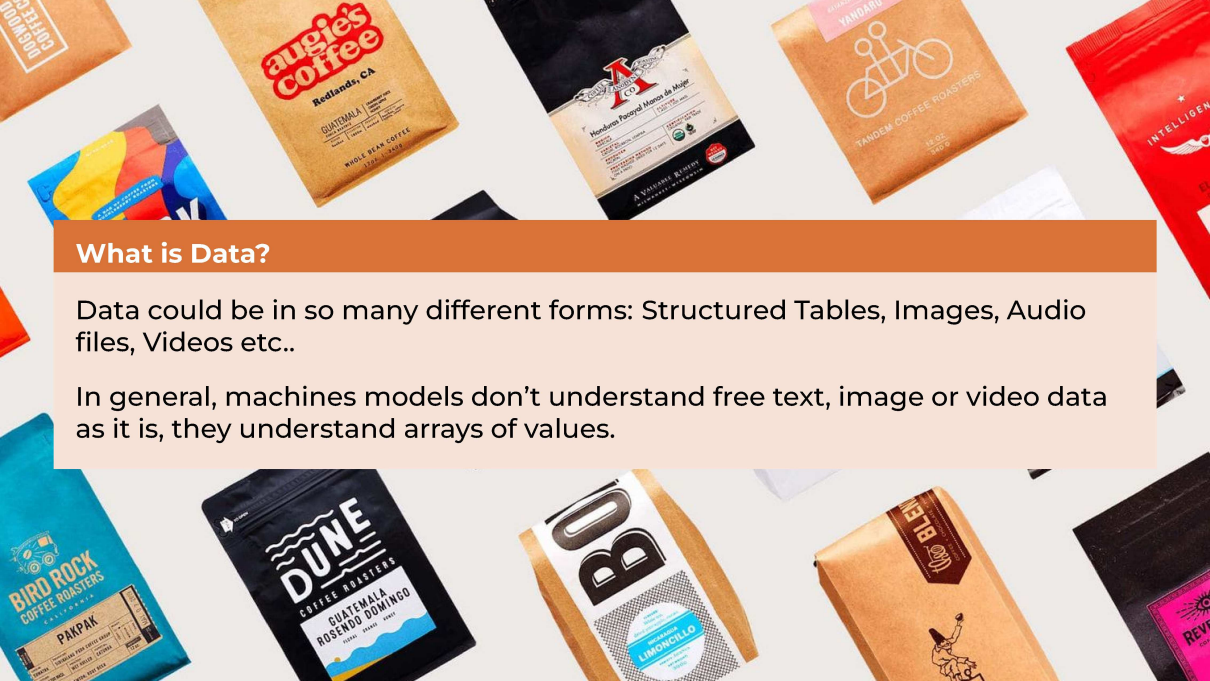
Data Science Workflow

A data science workflow defines the phases (or steps) in a data science project.


Preparation
Data Acquisition
Data Cleaning & Labeling
Feature Engineering
Modeling
Model Selection
Hyperparameter Optimization
Deployment
Model Improvement
Runtime Monitoring
Model Deployment
Model Validation
Ensembling


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... *MIGHT* NOT BE.
I GUESS THAT'S SOMETH-
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

Data Preprocessing



What is Data?

Data could be in so many different forms: Structured Tables, Images, Audio files, Videos etc..

In general, machines models don't understand free text, image or video data as it is, they understand arrays of values.

Data objects and Data sets

A data set can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities.

Supervised Learning: $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$

Unsupervised Learning: $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$

Data objects are described by a number of features, that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc.

Features are often called as variables, characteristics, fields, attributes, or dimensions.

Features

Categorical: Features whose values are taken from a defined set of values. For instance, days in a week: (*Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*) is a category because its value is always taken from this set. Another example could be the Boolean set: (*True, False*).

Numerical: Features whose values are continuous or integer-valued. They are represented by numbers and possess most of the properties of numbers. For instance, number of steps you walk in a day, or the speed at which you are driving your car.

Features

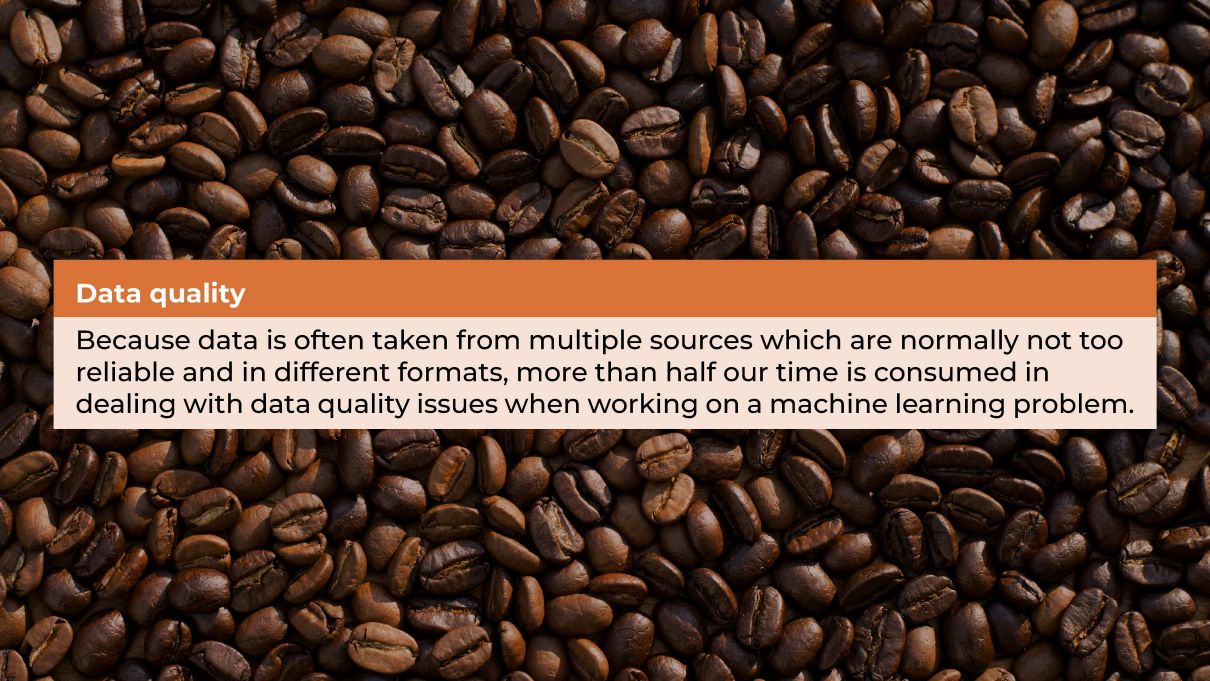
Useless: unique, discrete data without relationship with the outcome variable

Nominal: discrete values without relationship between the different categories

Ordinal: discrete integers that can be ranked or sorted

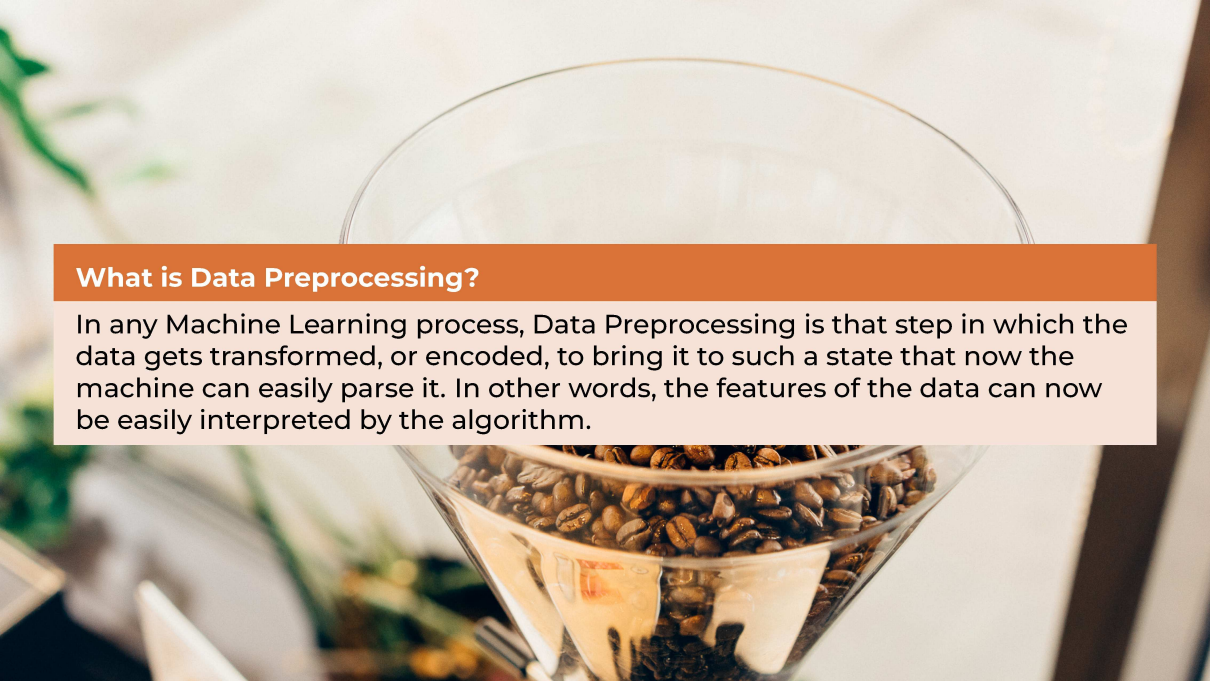
Binary: discrete data that can be in only one of two categories

Time: cyclical, repeating continuous form of data



Data quality

Because data is often taken from multiple sources which are normally not too reliable and in different formats, more than half our time is consumed in dealing with data quality issues when working on a machine learning problem.



What is Data Preprocessing?

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

Feature encoding

Feature encoding is basically performing transformations on the data such that it can be easily accepted as input for machine learning algorithms while still retaining its original meaning.

They are generally applied to categorical data

- Nominal: One-hot-encoding, dummy encoding
- Ordinal: Put to integers
- Binary: dummy encoding

Missing Data

Some column attributes just don’t exist

Eliminate rows with missing data

Eliminate columns with missing data

Imputation

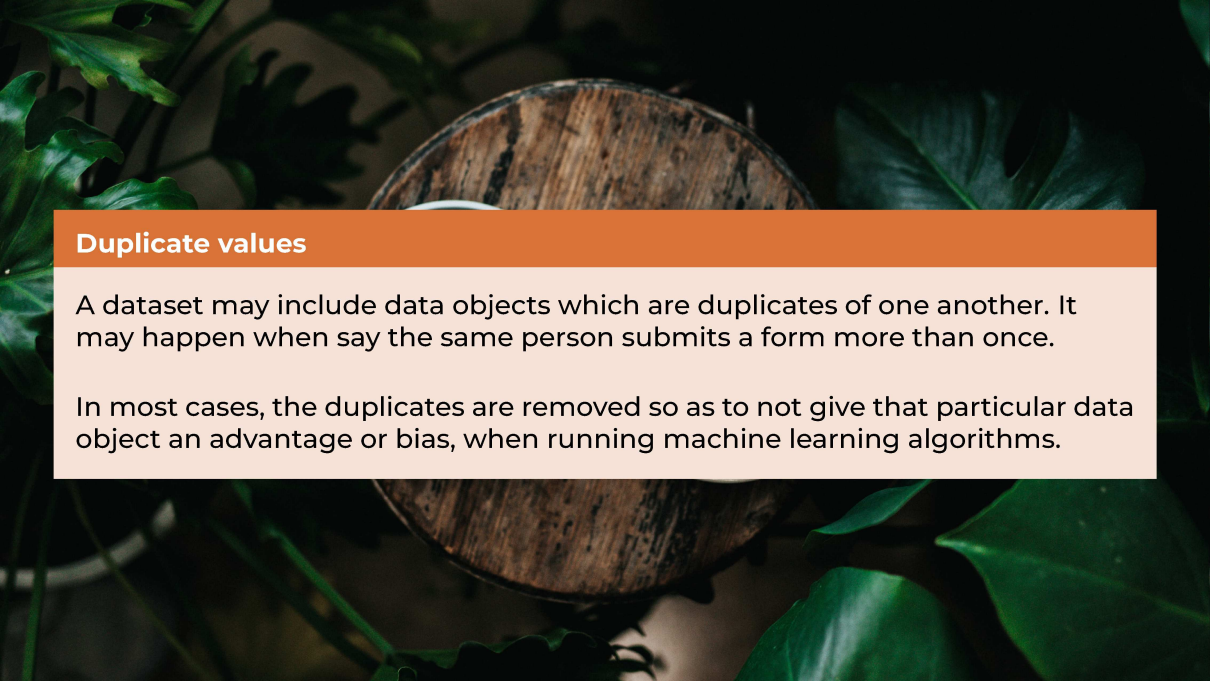
- mean, median, mode
- simple interpolation
- classification/regression models

Outliers

Extreme values that are outside the range of what is expected and unlike the other data

Methods to detect

- Standard Deviation Method
- Interquartile Range Method
- Automatic Outlier Detection



Duplicate values

A dataset may include data objects which are duplicates of one another. It may happen when say the same person submits a form more than once.

In most cases, the duplicates are removed so as to not give that particular data object an advantage or bias, when running machine learning algorithms.

Feature selection

Select a subset of input features from the data set.

Unsupervised: Do not use the target variable.

Supervised: Use the target variable.

- **Wrapper**: Search for well-performing subsets of features.
- **Filter**: Select subsets of features based on their relationship with the target.
- **Intrinsic**: Algorithms that perform automatic feature selection during training.

Dimensionality Reduction: Project input data into a lower-dimensional feature space.

Normalization

Normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging.

z-score

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

Min-Max

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{2}$$

omadson.codes