# Machine Learning

## Introduction part 2- Regression

**Parcours Progis**

**Etudes, Medias, communication, Marketing**

03.10.2025

Université Grenoble Alpes

Sciences Po Grenoble UGA

# References


- **Book: AI for Marketing and Product Innovation,** Dr. A. K., Pradeep, Andrew Appel, and Stan Sthanunathan.

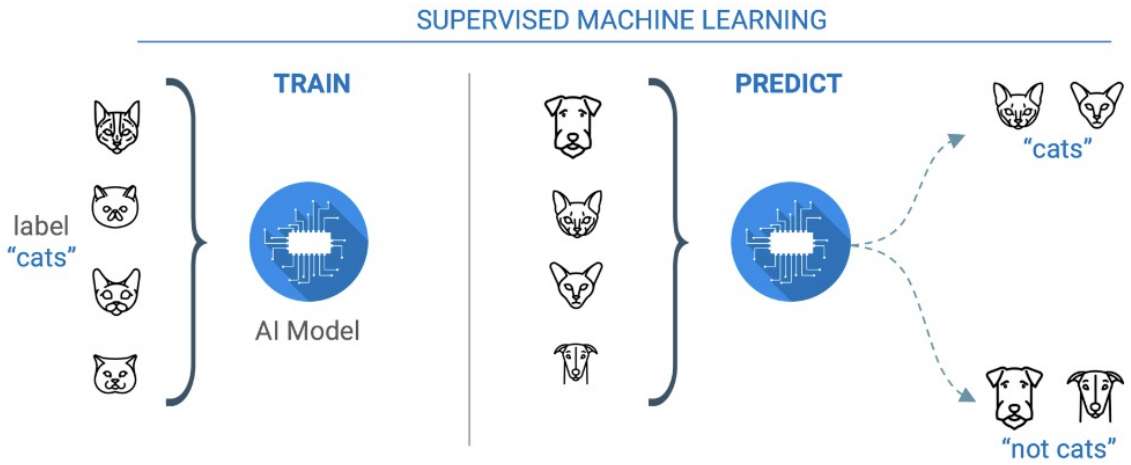
- **Book**: Machine Learning Implémentation en Python avec Scikit-learn (2e édition)- Virginie MATHIVET

- **Book:** Hwang, Yoon. 2019. *Hands-On Data Science for Marketing*. Packt Publishing, Ltd. Chapter 3, 128–156.

# Supervised learning

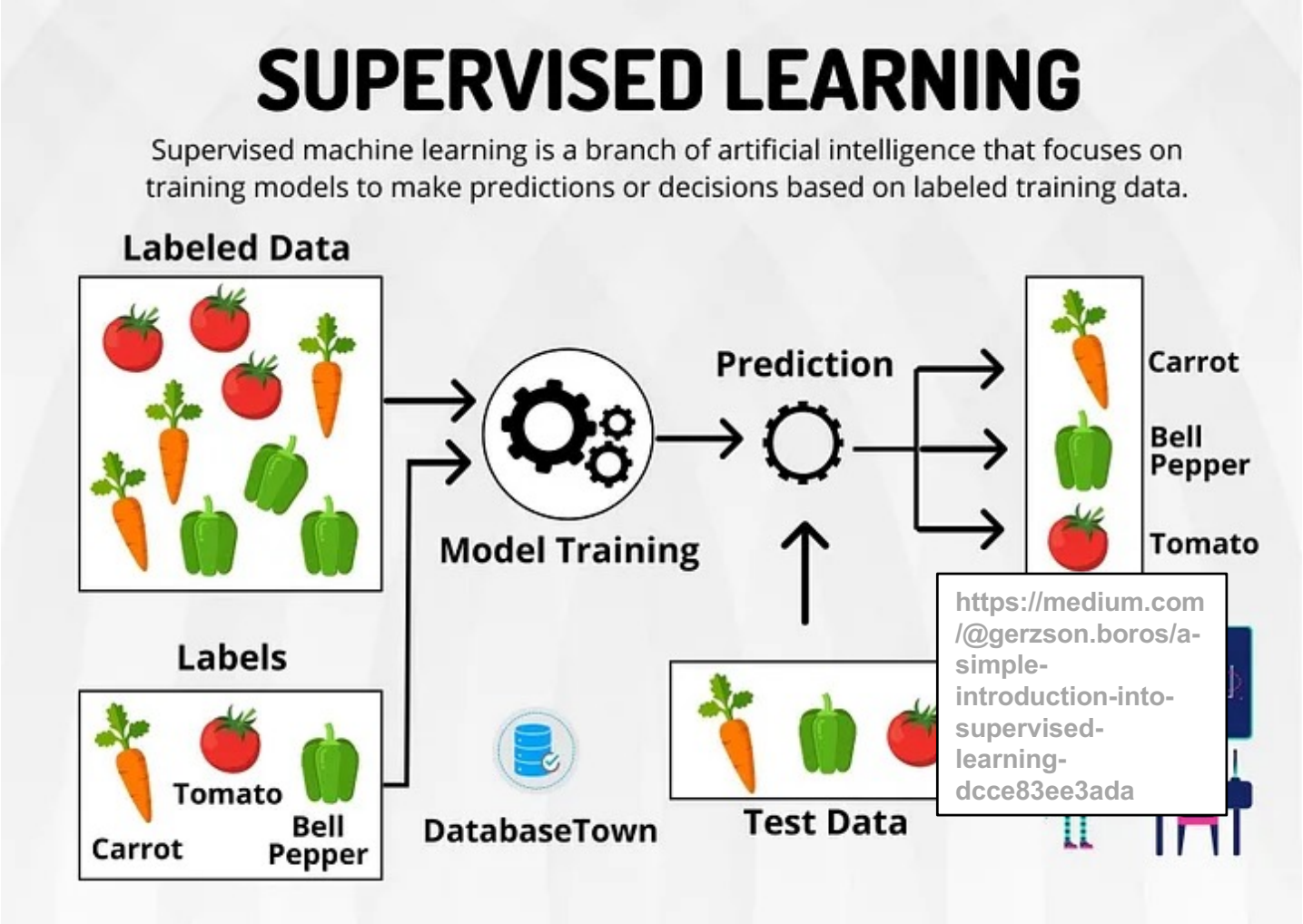
Supervised learning is a fundamental branch of machine learning where algorithms, learned patterns, and relationships in **labeled** training data

to make **predictions** or **classify** new **unseen** data points.



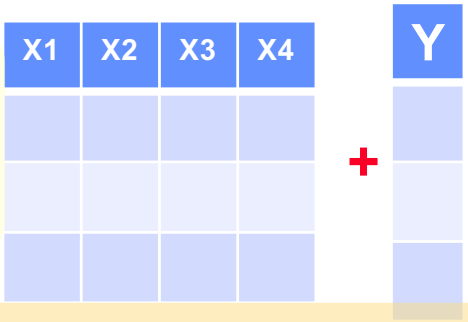

https://abeyon.com/how-do-machines-learn/

# Supervised learning



# Implementing Supervised Learning Model

1. Extract Data
2. Join data with Target




3. Data cleaning

4. Univariate and Bi-Variate Analysis

5. Select the Bests features

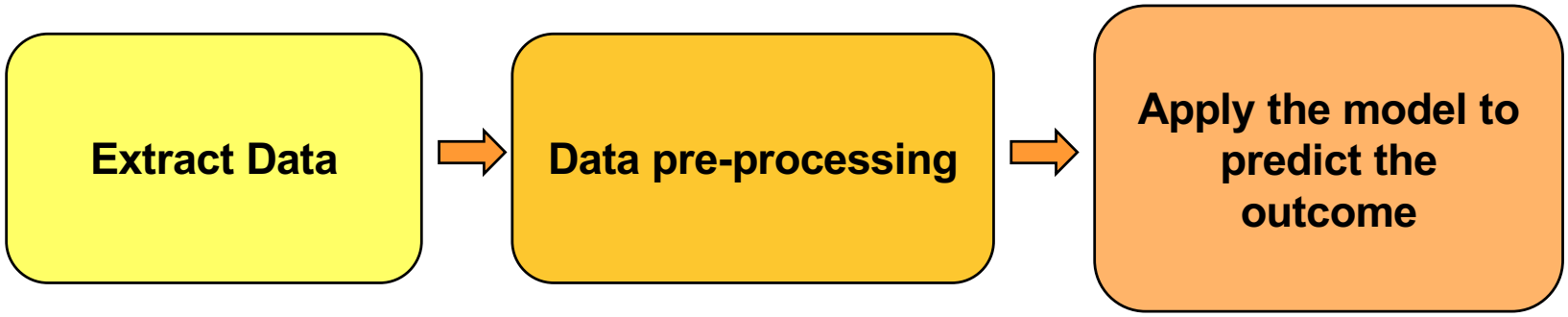
6. Data Split (Building training and testing Samples)

7. Apply appropriate Supervised Algorithm

8. Validate model results

# Implementing Supervised Learning Model

- **So in general:**



# What we learned in Supervised learning

- In classification problems the task is to assign new inputs to one of a number of discrete classes or categories.
- However, there are many other tasks, which we shall refer to as *regression* problems, in which the outputs represent the values of continuous variables.

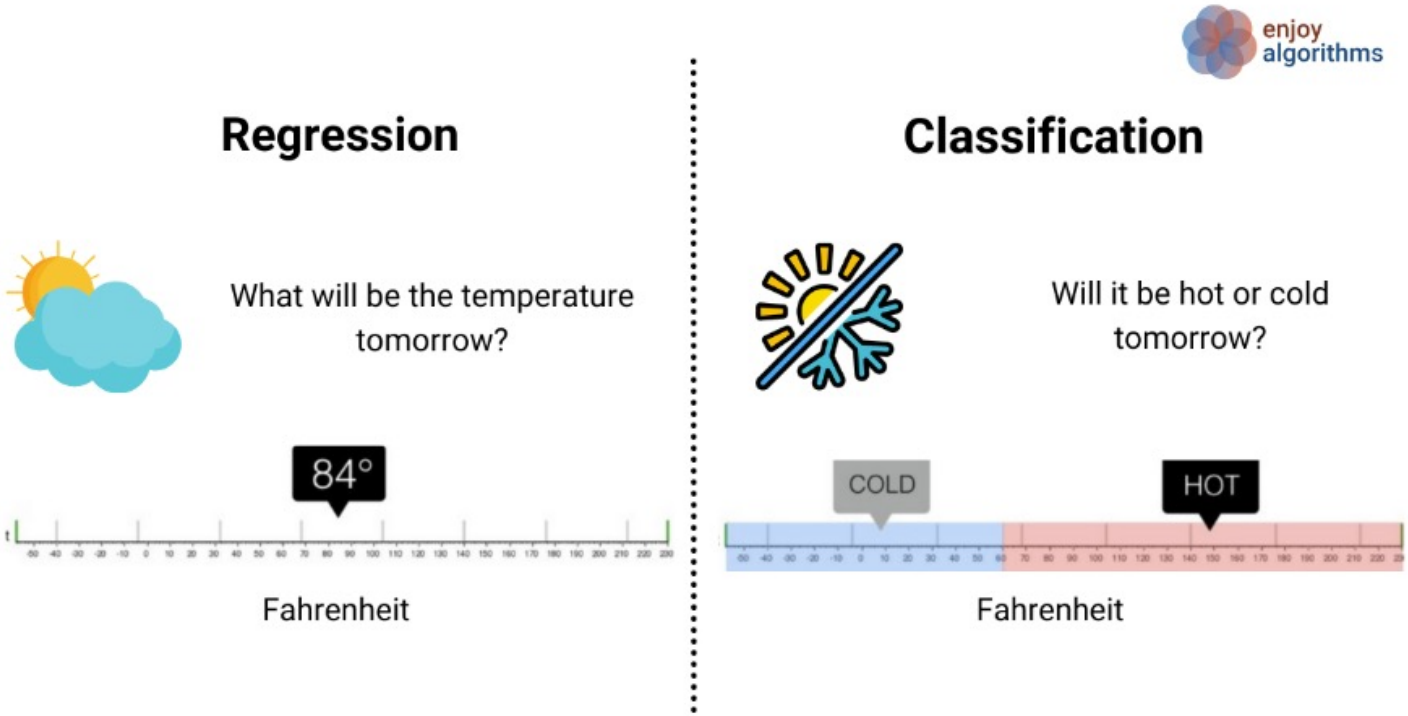
**classification**

-K-nearest Nieghbors
-Support vector machine
-Naive bayes classifier
-MLP
-Random forest

**Regression**

-Linear regression
-Decision tree regression
-Logistic regression

# Supervised Learning

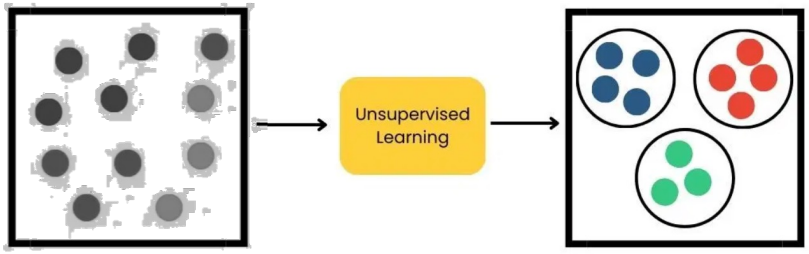



https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

# Unsupervised Learning Problem

# Unsupervised Learning

- Here, we let the model work on its own to discover information that may not be visible to the human eye.
- It means, the unsupervised algorithm trains on the dataset, and draws conclusions on unlabeled data.
- Unsupervised learning has more difficult algorithms than supervised learning since we know little to no information about the data.
- Dimension reduction and clustering are the most widely used unsupervised machine learning techniques.

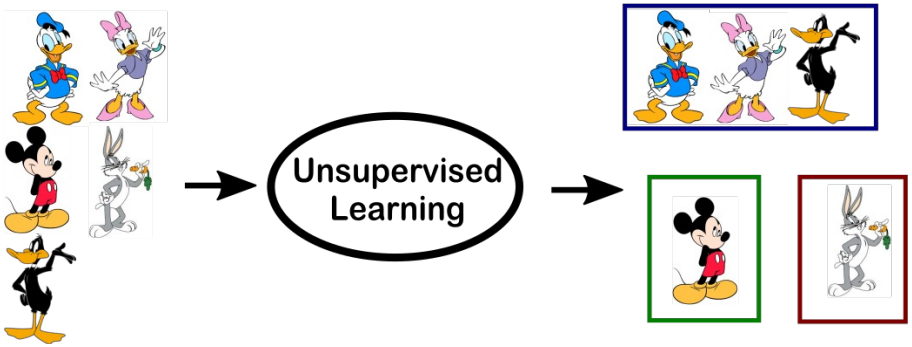


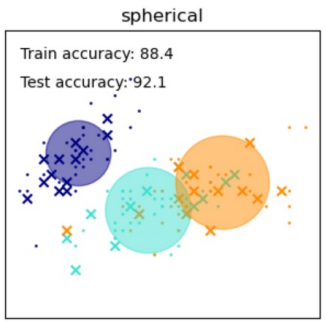

https://www.bombaysoftwares.com/blog/introduction-to-unsupervised-learning

# Unsupervised Learning

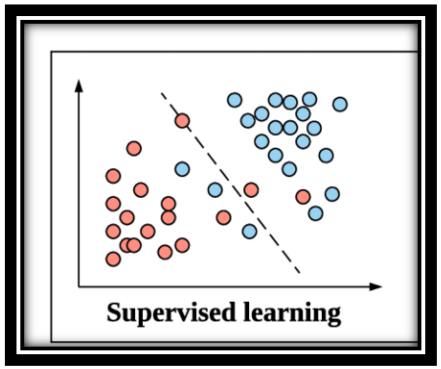
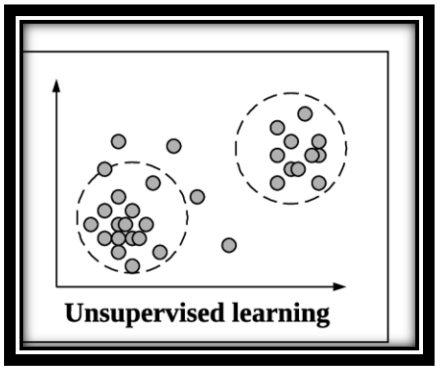
- Here we only talk about grouping and there is no prediction aim.
- Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points.
- The lack of the target variable defines an unsupervised problem.

| X1 | X2 | X3 | X4 | X5 | Cluster ID |
|---|---|---|---|---|---|
| 7 | 1 | 3 | 1 | 2 | 1 |
| 2 | 4 | 7 | 12 | 9 | 1 |
| 5 | 9 | 1 | 10 | 6 | 2 |
| 8 | 11 | 2 | 3 | 4 | 3 |
| 1 | 5 | 5 | 7 | 1 | 3 |



# Supervised Vs Unsupervised

| Supervised Learning | Unsupervised Learning |
| --- | --- |
|   |   |
| Our model should effectively classify observations into either success or failure classes based on the features that we have. | Good clusters should capture similar observations within and capture different observations across clusters. |

# Why do you need datasets?

- Machine learning algorithms learn from data. A machine learning algorithm identifies trends, and relationships, and makes predictions based on large volume of data given to training the models.

One of the major challenges faced by marketers today is the limited access **to customer data**. This is due to privacy regulations, increasing consumer concerns about data privacy, and the complexity of gathering and analyzing large volume of data.

# What is a Supervised Learning Dataset?

- One of the fundamental steps in any machine learning project is importing the dataset.
- What is a DATASET?
  - A dataset in machine learning is a collection of instances (instance refers to a single row of data) that all share some common features and attributes.
  - Two kinds of datasets are required

    -**Training Dataset** — The data that is fed into the machine learning algorithm for training.

    -**Test Dataset** or Validation Dataset — The data that is used to evaluate and test that the machine learning model is interpreting accurately.

# What is feature, instance and target variable?

- Feature: synonymous with the term variable, column, attributes and fields.

| Size of house | Number of bedroom | Number of floor | Price |
|---|---|---|---|
| | | | |

- Instance: is a synonymous with the term row, datapoint and case.

| Size of house | Number of bedroom | Number of floor | Price |
|---|---|---|---|
| 98 | 2 | 1 | 440000 |
| 130 | 3 | 2 | 800000 |
| 78 | 1 | 1 | 250000 |

First instance

Target variable

# Dataset example (supervised)

| buying | maint | doors | persons | lug_boot | safety | class |
|---|---|---|---|---|---|---|
| vhigh | med | 2 | more | med | high | acc |
| vhigh | med | 2 | more | big | low | unacc |
| vhigh | med | 2 | more | big | med | acc |
| vhigh | med | 2 | more | big | high | acc |
| vhigh | med | 3 | 2 | small | low | unacc |
| vhigh | med | 3 | 2 | small | med | unacc |
| vhigh | med | 3 | 2 | small | high | unacc |
| vhigh | med | 3 | 2 | med | low | unacc |
| vhigh | med | 3 | 2 | med | med | unacc |
| vhigh | med | 3 | 2 | med | high | unacc |
| vhigh | med | 3 | 2 | big | | |
| vhigh | med | 3 | 2 | big | | |
| vhigh | med | 3 | 2 | big | | |
| vhigh | med | 3 | 4 | small | | |

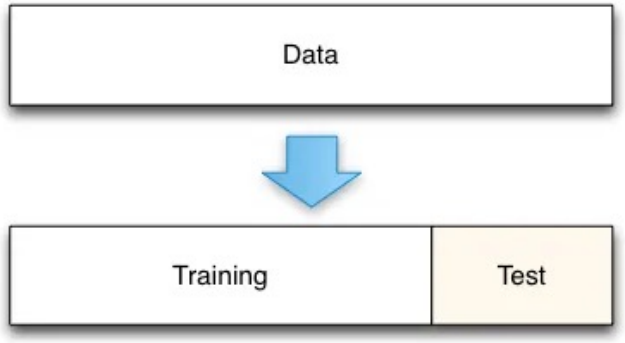
| | | | |
|---|---|---|---|
| buying | Feature | Categorical | buying price |
| maint | Feature | Categorical | price of the maintenance |
| doors | Feature | Categorical | number of doors |
| persons | Feature | Categorical | capacity in terms of persons to carry |
| lug_boot | Feature | Categorical | the size of luggage boot |
| safety | Feature | Categorical | estimated safety of the car |
| class | Target | Categorical | evaluation level (unacceptable, acceptable, good, very good) |

**Target variable: Synonymous with the terms "predictant," "response," or "dependent variable."**

https://www.kaggle.com/datasets/stealthtechnologies/car-evaluation-classification

# Training and Testing

- Universal set population
- Training set
  - Build a model only on a portion of the population.
- Testing set
  - How the model actually behaves on unobserved data.




https://medium.com/analytics-vidhya/only-train-and-test-set-is-not-enough-for-generalizing-ml-model-significance-of-validation-set-cf68bb26881a

Three parts to machine learning
First is get the data,
second is build your model on the training set and
third is to validate the performance of the model on the testing set.

# Why Split the Data?

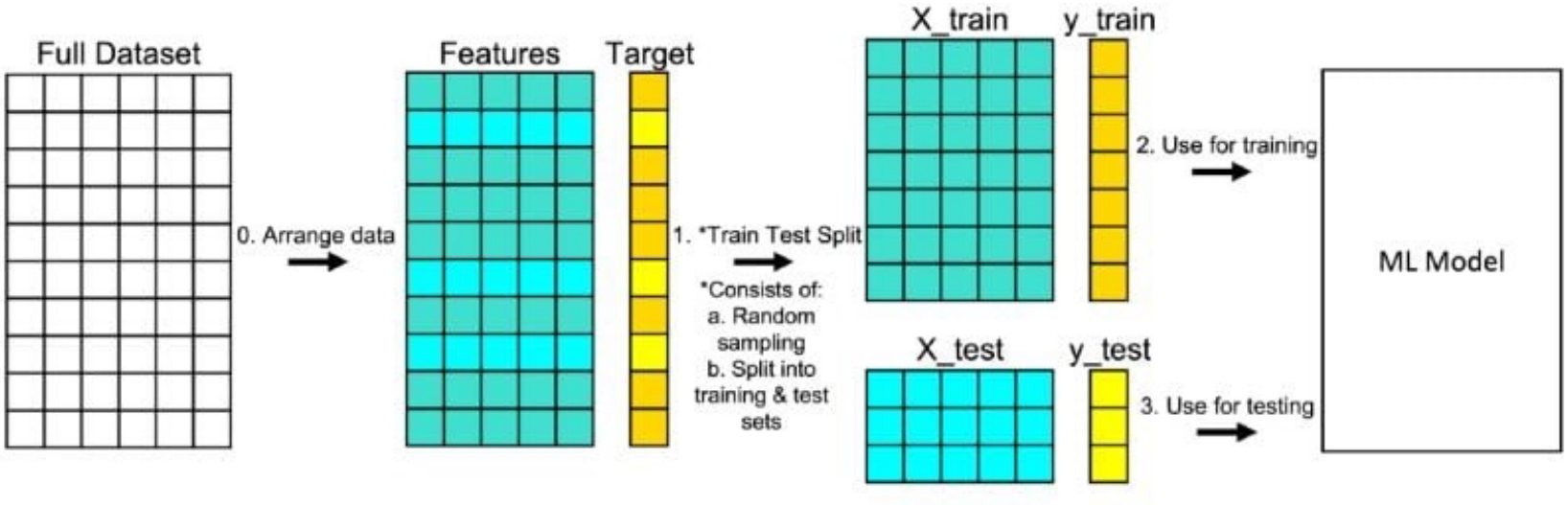
Build a model that works well on new, unseen data

- Train Dataset (80%):
  - Used to teach the model (like studying for an exam)
- Test Dataset (20%):
  - Used to evaluate the model (like taking the exam)

Enough Data: 80% is usually enough to train a good model

Reliable Test: 20% is enough to evaluate performance without being too small.

# Training and testing



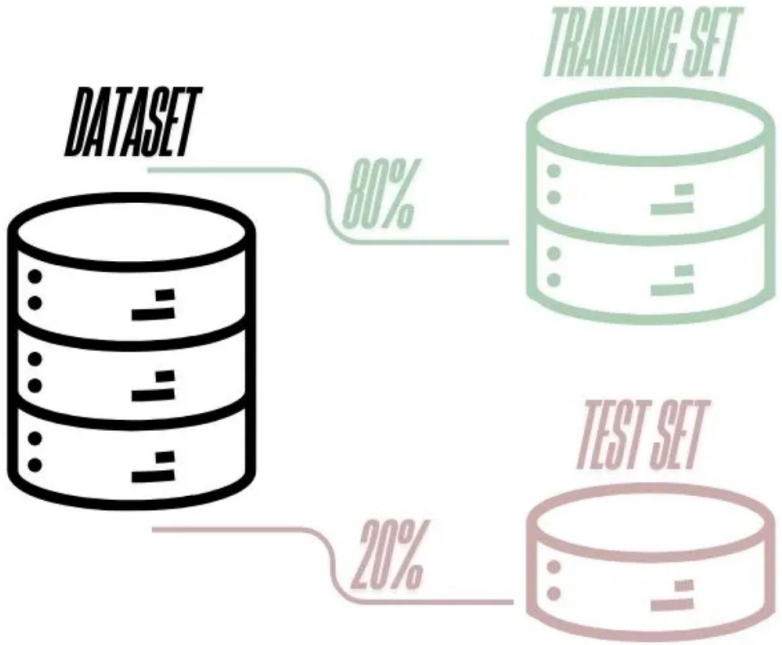

https://builtin.com/data-science/train-test-split

1. **Train**: Fit the model on **X_train**, **y_train**
2. **Predict**: Use the model to predict **y_test** from **X_test**
3. **Evaluate**: Compare predictions to **actual y_test** values.

# Training and testing

**Training Set**: Used to train the model and help it learn patterns.

**Test Set**: Used to evaluate how well the model generalizes to new data.




https://python.plainenglish.io/train-test-split-in-python-a-step-by-step-guide-with-example-for-accurate-model-evaluation-53741204ff7d

# How can we access the datasets?

https://blog.learnamp.com/why-asking-questions-is-the-key-to-business-success

# Importing the Data Set in Machine Learning Using Sklearn datasets

- **The scikit-learn library offers several datasets for machine learning. These datasets are available through sklearn.datasets and fall into different categories**

- `load_iris` : Iris flower dataset (classification)
- `load_digits` : Handwritten digits dataset (classification)
- `load_wine` : Wine dataset (classification)
- `load_breast_cancer` : Breast cancer dataset (classification)
- `load_diabetes` : Diabetes dataset (regression)
- `load_linnerud` : Linnerud dataset (multivariate regression)


**-Importing Scikit-Learn**

make sure you have Scikit-Learn installed. If not, you can install it using:

```
pip install scikit-learn
```

**-Loading Sample Datasets**

To load a sample dataset, you can use the following code:

```
from sklearn import datasets
# Load the iris dataset
iris = datasets.load_iris()
 # Access the features and target variable
X = iris.data # Features
y = iris.target # Target variable
```

# Importing the Data Set in Machine Learning Using Sklearn datasets

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [12, 4]
from sklearn import datasets

# Load the diabetes dataset
diabetes = datasets.load_diabetes(as_frame=True)
print(diabetes.DESCR)
```

```
**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one ye

  :Attribute Information:
      - age     age in years
      - sex
      - bmi     body mass index
      - bp      average blood pressure
      - s1      tc, T-Cells (a type of white blood cells)
      - s2      ldl, low-density lipoproteins
      - s3      hdl, high-density lipoproteins
      - s4      tch, thyroid stimulating hormone
      - s5      ltg, lamotrigine
      - s6      glu, blood sugar level
```

# Different forms of data

Grouping by color is one way to organize categorical data.

- **Categorical data:**
  - **A categorical characteristic** is an attribute that can take one of a limited and usually fixed number of possible values based on a qualitative property. **For example, hair color (black, orange, blond)**
  - Example in marketing: understanding customer segments through categorical data allows for targeted advertising and personalized marketing campaigns.

- **Numerical data:**
  - If an entity represents a characteristic measured **in numbers, it is called a numerical entity.**
  - **Discrete data: countable items. For example** Shoe sizes.
  - **Continuous data: can take any value within a range, such as temperatures or distances.**

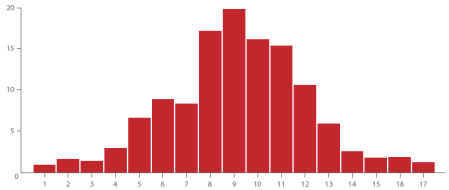
# Types de données

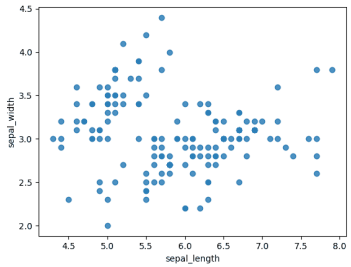
- **Pour chaque variable du jeu de données, deux informations devront être produites :**
  - **une analyse statistique, sous forme textuelle ou tabulaire.**
  - **une représentation graphique des données (visualisation des données)**

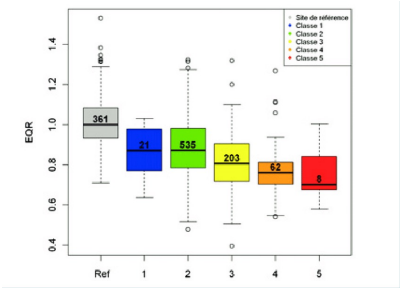
- **Les outils à utiliser sont cependant différents selon le type de la variable (catégorielle ou numérique)**

# Une représentation graphique des données

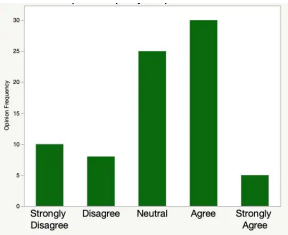
- **Graphiques sur données numériques**
  - **Histogramme, Nuage de points(Scatter plot), Boites a moustaches (box plot)**

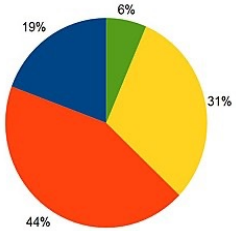





- **Graphiques sur données catégorielles**
  - **Les diagrammes à barres, les diagrammes circulaires (aussi appelés « en camembert »)**



# Importing the Data Set in Machine Learning Using Scikit-Learn from CSV file

- **Loading CSV Datasets**

For datasets in CSV format, you can use libraries like Pandas to load the data and then convert it to NumPy arrays for further processing with Scikit-Learn:

```
data = pd.read_csv('./WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv', encoding='latin1')
data.columns
```

```
Index(['Customer', 'State', 'Customer Lifetime Value', 'Response', 'Coverage',
       'Education', 'Effective To Date', 'EmploymentStatus', 'Gender',
       'Income', 'Location Code', 'Marital Status', 'Monthly Premium Auto',
       'Months Since Last Claim', 'Months Since Policy Inception',
       'Number of Open Complaints', 'Number of Policies', 'Policy Type',
       'Policy', 'Renew Offer Type', 'Sales Channel', 'Total Claim Amount',
       'Vehicle Class', 'Vehicle Size'],
      dtype='object')
```

# Noise

- Noise is any unwanted anomaly in the data and due to noise, the class may be more difficult to learn
- There are several interpretations of noise:
  - There may be imprecision in recording the input attributes, which may shift the data points in the input space.
  - There may be errors in labeling the data points, which may relabel positive instances as negative and vice versa. (called *teacher noise)*
  - There may be additional attributes, which we have not taken into account, that affect the label of an instance. (It may be *hidden* or *latent)*

# Data cleansing (data scrubbing)

- Data cleansing is the process of finding and correcting or deleting irrelevant**,** missing, duplicate, or otherwise useless data from a datset.
- This is a necessary step designed to purify data so that algorithms can work faster and make more accurate predictions.

-Reasons for the corruption of data are :
user error, dummy data, and workarounds.
-Data cleansing functions may include the enhancement, harmonization, and standardization of data.

*To perform data cleansing, all incorrect, incomplete, and irrelevant data must be found, then either **replaced**, **removed**, or **modified.***

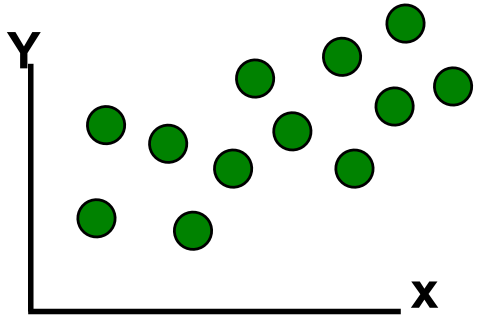

https://www.dataversity.net/what-is-data-cleansing/

# Regression

- It can help to invistigate the relation between variables.
- We commonly use three major types of regression:
  - Linear regression
  - Polynomial regression
  - And Logistic regression

# Regression

- Imagine we collect the price of a package of pumpkins on different days of the year.
- the x value of each point is the day of the year the package is sold.
- And the Y value is the price of the package.

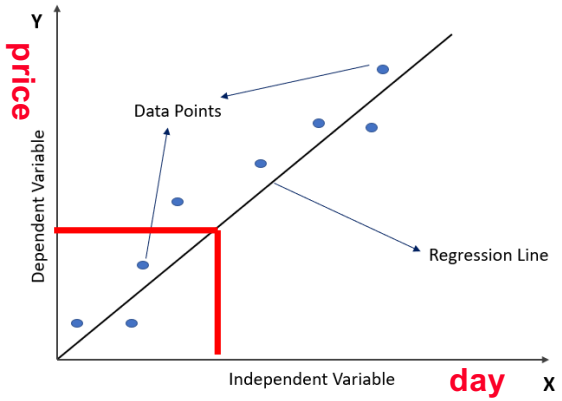



- You can use regression to find a mathematical formula that represents the general trend of the data.
- This formula encodes the relationship between our x and y variables
- And enables us to predict y for any given value of x.

# Linear Regression

- It uses a straight line to approximate the trend of our data points.
- y = mx + b, where m is the slope of the line, b is the y-intercept, and x is the independent variable.
- A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

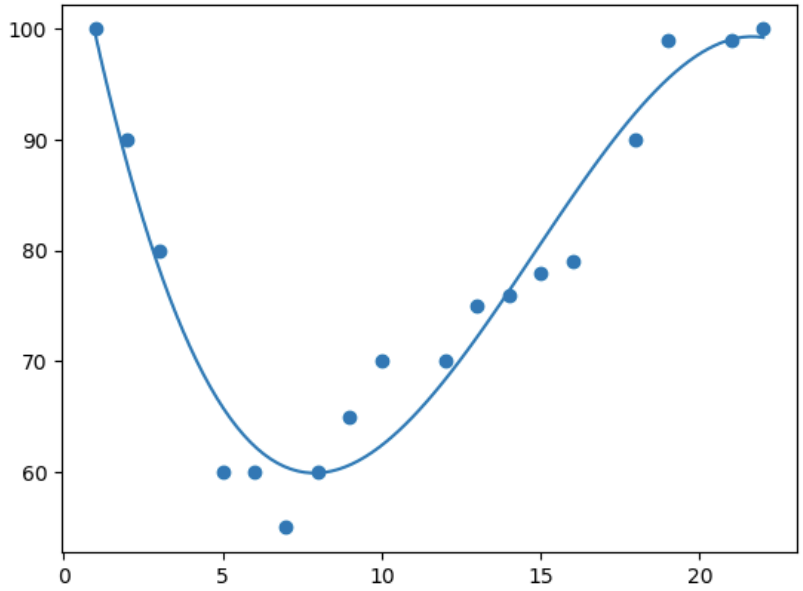
In pumpkin dataset, we want to **predict** the price of a package of pumpkins for the Particular day of the year




https://wisdomml.in/linear-regression-in-machine-learning-a-comprehensive-guide/

# Polynomial regression

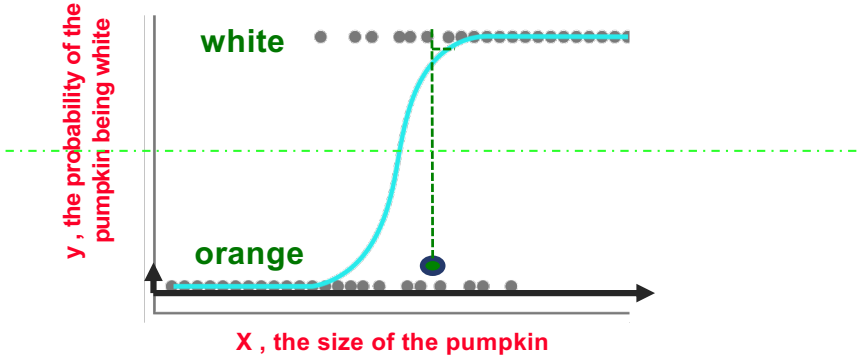
- In some situations, we can consider polynomial regression which is really just an extension of linear regression that uses a curve to represent a relationship between variables.
- We introduce a new term in our formula that include the square of x variable.



# Logistic regression

- We want to predict whether the color of a pumpkin is orange or white based on its size.
- The value we want to predict is a category either orange or white.

**When we want to predict a category logistic regression is a great method**



# Exemple de cas pratique

- **La regression est utilisée dans de nombreux cas potentiels :**
  - **Prédire le prix d'un produit ou service en fonction de ses caractéristiques, comme le prix d'un appartement en fonction de sa taille, le nombre de chambres , etc.**
  - **Évaluer les risques d'un évènement en fonction d'autres évènements ou informations.**
  - **Estimez la qualité d'une production en fonction des données physiques du processus (température, pression, humidité…).**
  - **….**

# Drivers Behind Marketing Engagement

- **Regression analysis allows marketers to explore the relationships between independent variables, such as advertising spend, social media presence, customer demographics, and the dependent variable of marketing engagement.**

- By analyzing the statistical significance and coefficients of these variables, marketers can identify the key factors that influence engagement and gain insights into their relative impact.

**Book:** Hwang, Yoon. 2019. *Hands-On Data Science for Marketing*. Packt Publishing, Ltd. Chapter 3, 128–156.

# End