

Machine Learning

Introduction part 2



Parcours Progis

Etudes, Medias, communication, Marketing

19.10.2026

Bahareh Afshinpour

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.

StatQuest — The Main Ideas of Fitting a Line to Data https://www.youtube.com/watch?v=PaFPbb66DxQ

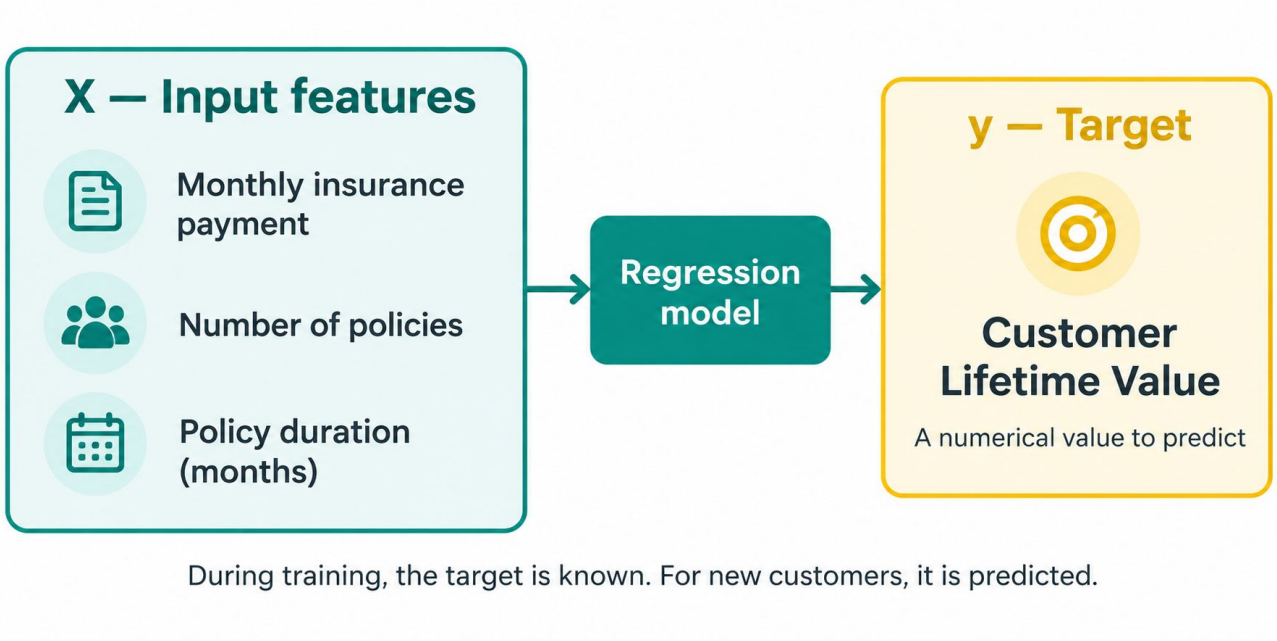
Recall — What is supervised learning?

Supervised learning learns from examples with known answers.(Labeled data)

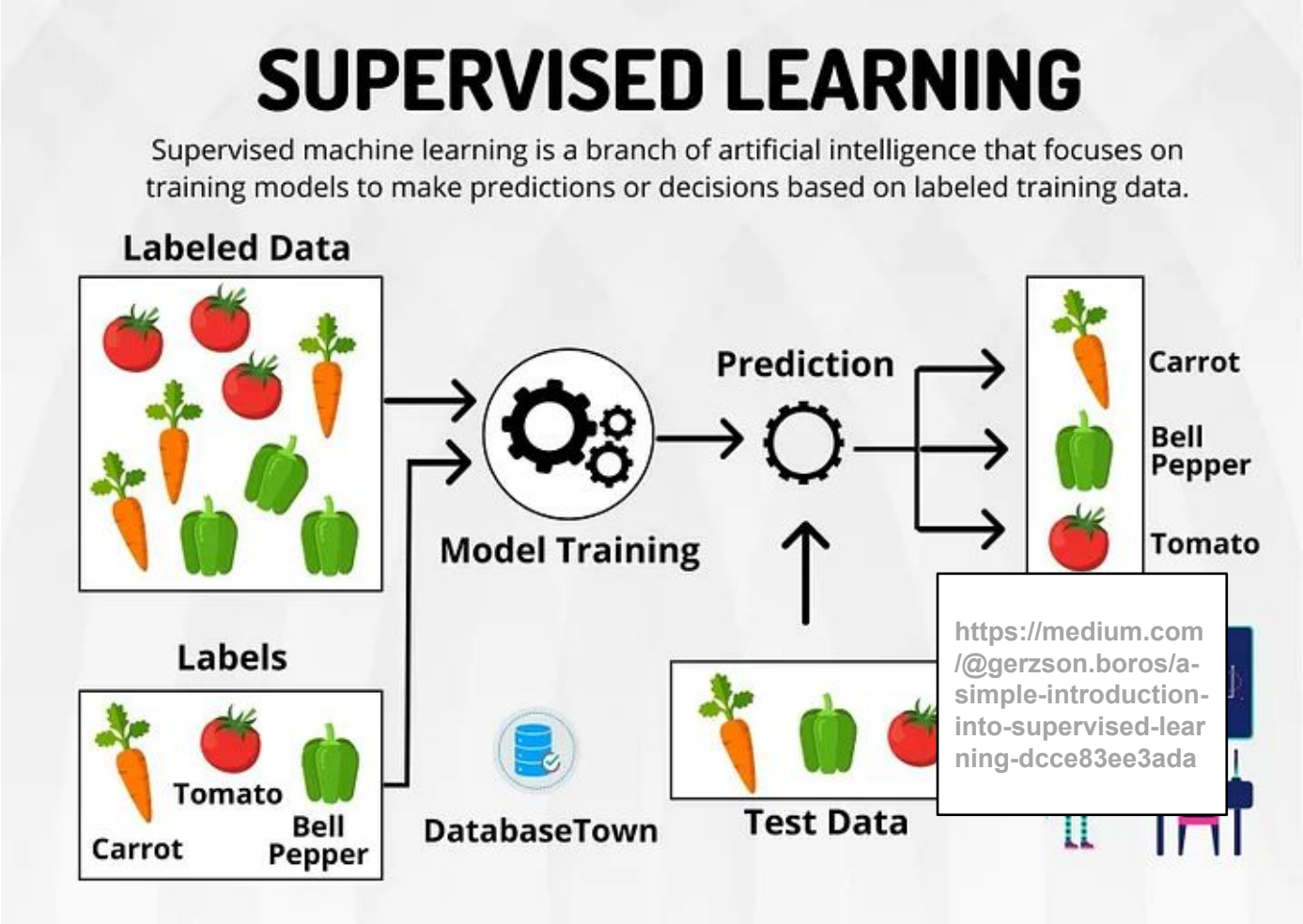
Each training example contains:

- **Input features (X):** information used to make a prediction.
- **Target (y):** the known outcome we want to predict.

The goal is to predict the target for **new observations not used during training**.



Recall — What is supervised learning?



Implementing a Supervised Learning Model

Define the prediction task — choose the target and suitable inputs.

Set aside test data — keep it separate from model development.

Explore and prepare the training data — inspect variables and learn any preprocessing rules.

Train the model — learn relationships between inputs and the target.

Evaluate on the test data — compare predictions with known answers.

Predict for new customers — apply the same preprocessing and trained model.

Implementing a Supervised Learning Model



What does each row represent? Identify the input features X and the target y.

Separate training and test data. Handle missing values and encode categories when needed.

Learn from the training data. Evaluate predictions on the test data. Predict outcomes for new customers.

Our example: use customer characteristics to predict Customer Lifetime Value.

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.

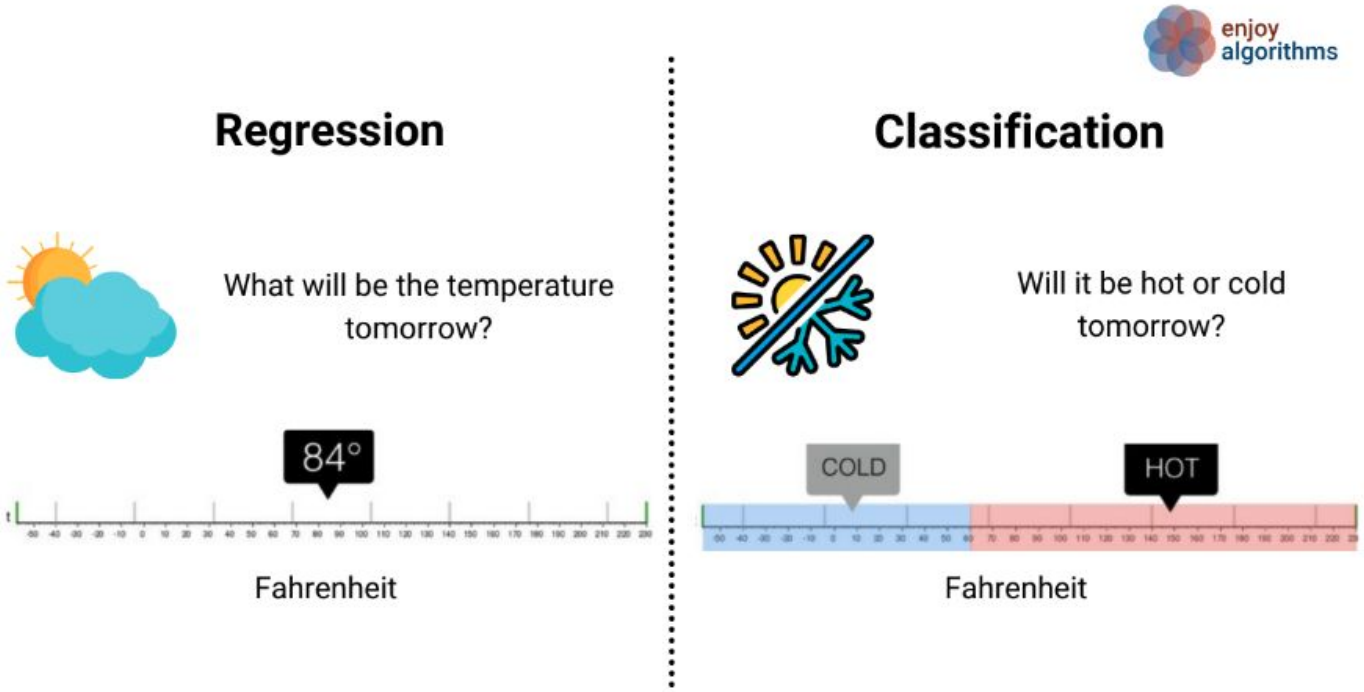
classification

-K-nearest Nieghbors
-Support vector machine
-Naive bayes classifier
-MLP
….

Regression

-Linear regression
-Logistic regression
….

Supervised Learning



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

Supervised Learning

Regression	Classification
Predict a numerical quantity	Predict a category
What is this customer's recorded lifetime value?	Will this customer respond to the campaign?
Target: **Customer Lifetime Value**	Target: **Response**
Example prediction: 6,500	Example prediction: Yes
Example method: linear regression	Example method: logistic regression

The type of target determines the task.



If we encode Response as 0 and 1, does it become a regression problem?

Unsupervised Learning Problem

Recall: What is unsupervised learning?

Unsupervised learning finds patterns in data without using a known target variable.

- We provide input features X, without target answers y.
- Clustering groups observations with similar characteristics.
- Dimensionality reduction represents data using fewer dimensions.

Example: identify groups of customers with similar insurance profiles.

Customer features → Clustering algorithm → Customer groups

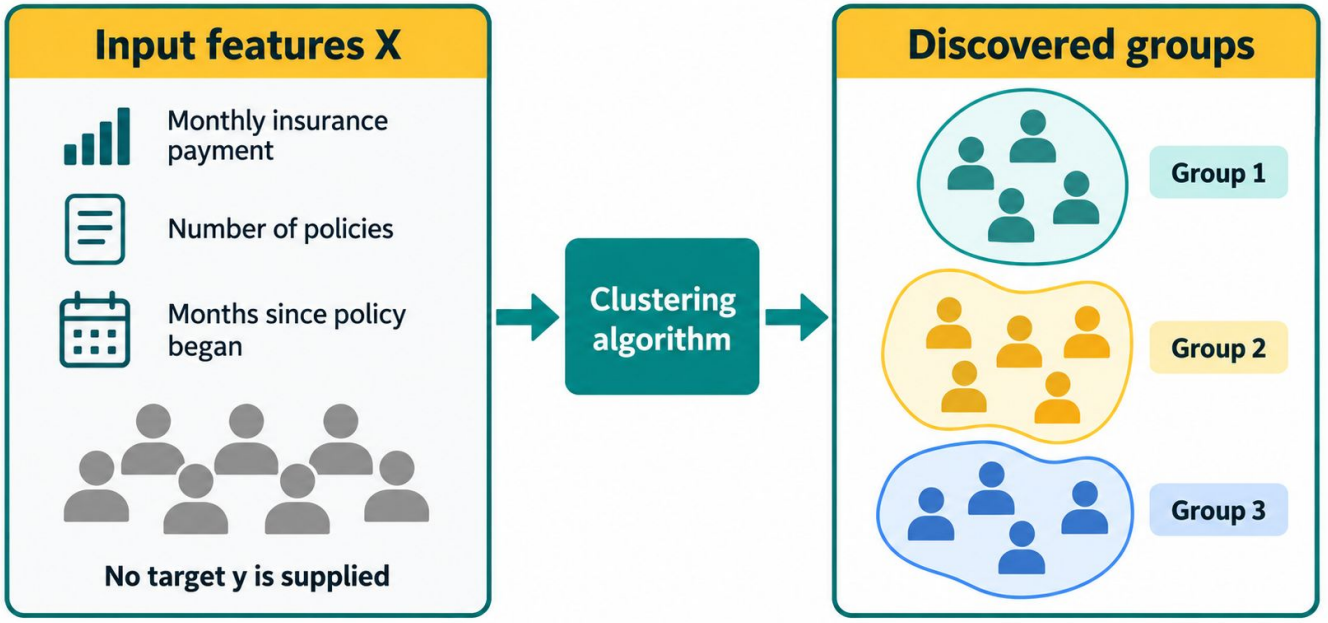
Example: Grouping similar customers

Question: Can we identify groups of customers with similar insurance profiles?

Input features X:

- Monthly insurance payment
- Number of policies
- Months since the policy began

Output: a cluster assigned to each customer.



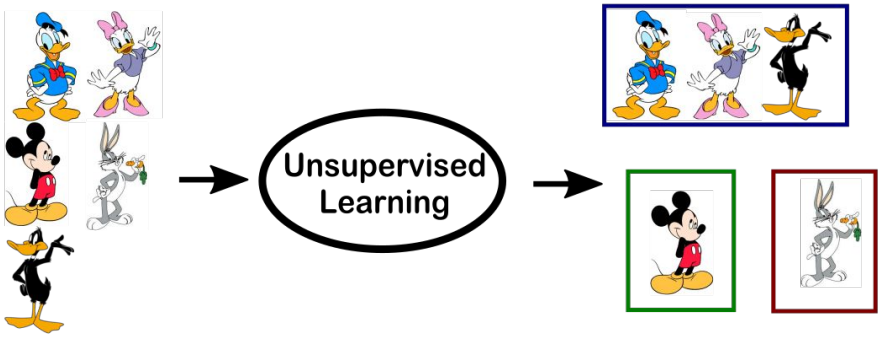
Groups are discovered from similarities in the selected features.

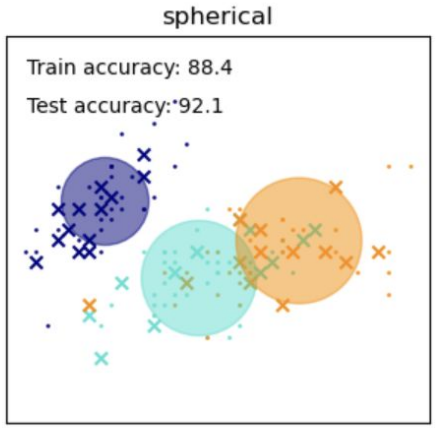
Illustrative example

Example: Grouping similar customers

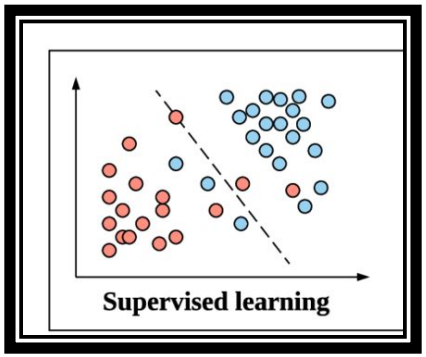
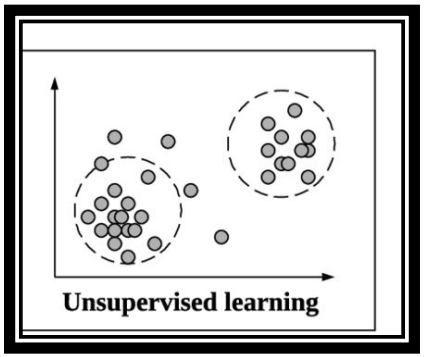
- Here we talk about grouping and there is no prediction aim.

- The lack of the target variable defines an unsupervised problem.





Supervised vs. Unsupervised

Supervised Learning	Unsupervised Learning
	

	Supervised learning	Unsupervised learning
Goal	Predict a specified target	Discover structure in the features.
During learning	Features X and known target y	Features X, without a target
Our example	Predict Customer Lifetime Value or Response	Group customers with similar insurance profiles
Result	A predicted value or category	Groups or a simplified representation

From Data to Predictions: Training and Testing

Why does the dataset matter?

A model learns relationships from the examples we provide.

Useful training data should:

- Contain information relevant to the prediction task.
- Have reliable feature values and target values.
- Reflect the customers for whom we want to make predictions.

More data alone does not guarantee a better model.

Reading our dataset: observations, features and target

- 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 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 terms variable, column, attributes and fields.

Size of house	Number of bedroom	Number of floor	Price

- Instance: is a synonymous with the terms 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		

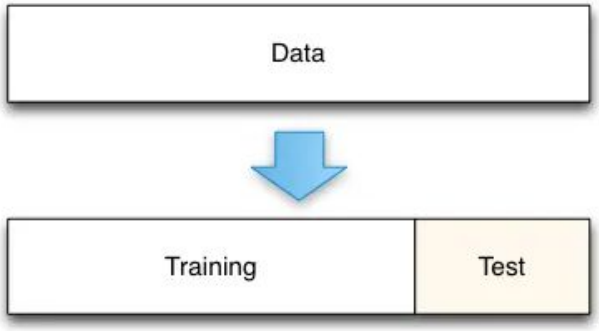
buying	Feature	Categorical	buying price
maint	Feature	Categorical	price of the maintenance
doors	Feature	Categorical	number of doors
persons	Feature	Categorical	capacity in terms of persons to carry
lug_boot	Feature	Categorical	the size of luggage boot
safety	Feature	Categorical	estimated safety of the 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 populatio
- Testing set
  - How the model actually behaves on unobserved data.



https://medium.com/analytics-vidhya/only-train-and-test-data-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)

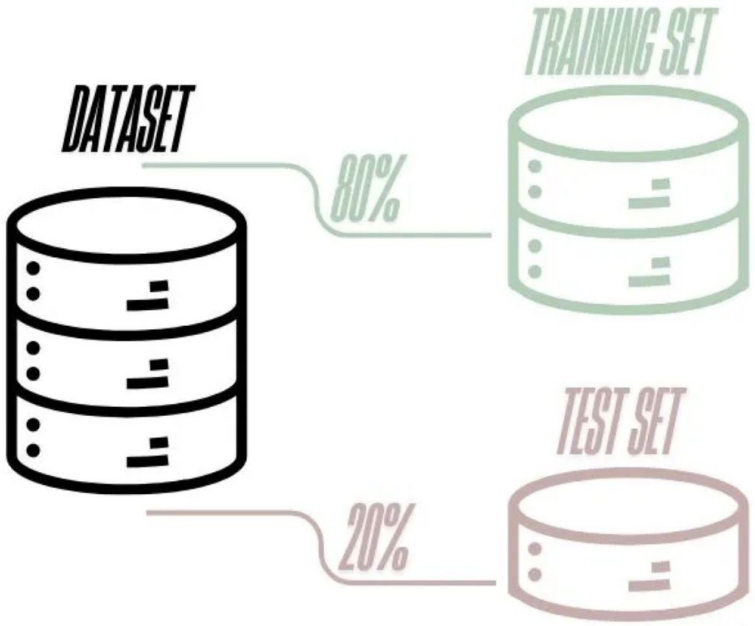
Enough Data: 80% is usually enough to train a good model

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

Training and testing

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

Why separate training and test data?

We want a model that works on customers it did not learn from.

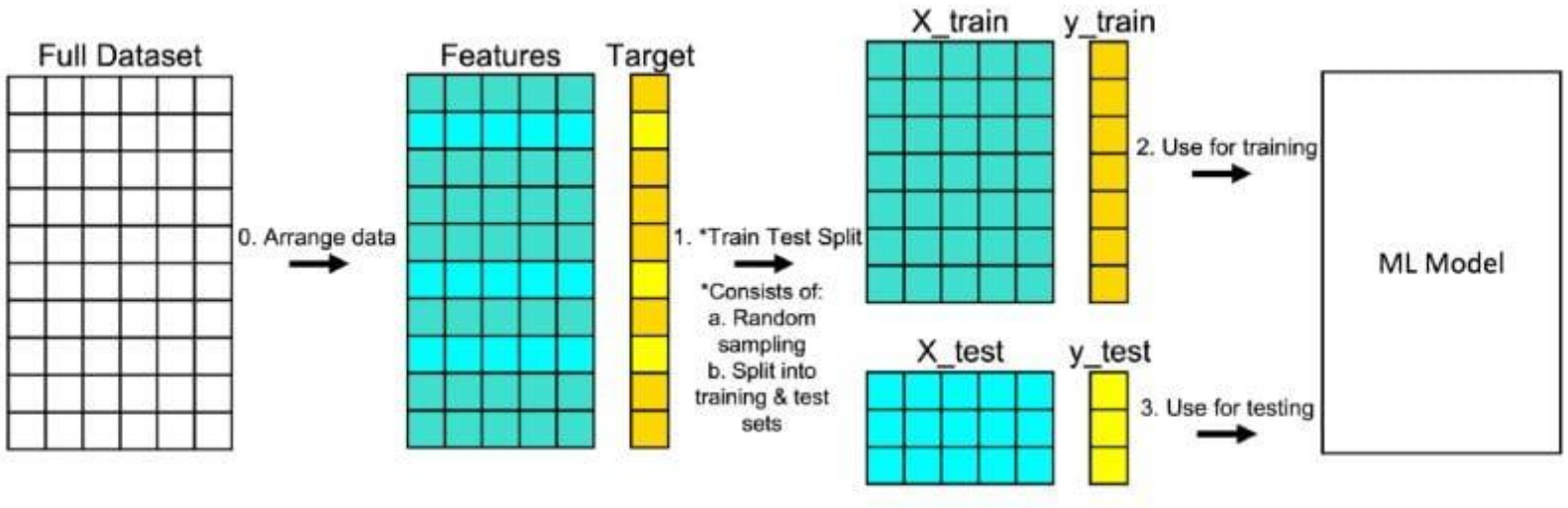
Training set: used to learn relationships between features and the target.

Test set: kept separate to evaluate predictions on unseen examples.

Generalisation means performing well on new observations.

Training = practice questions. Testing = new exam questions.

Train, predict, evaluate



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

Test targets are known to us, but are not given to the model when it makes predictions.

Step	What happens?
1. Train	Learn from `X_train` and `y_train`.
2. Predict	Use `X_test` to produce predictions: `y_pred`.
3. Evaluate	Compare `y_pred` with the actual values: `y_test`.

At which step do we use `y_test`?

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

```
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
```

Understanding and Preparing the Data.

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 of variables

Type	Meaning	Example
Categorical	Describes a category or group	Sales Channel: Agent, Branch, Call Center, Web
Numerical — discrete	Takes separate, countable values	Number of Policies: 1, 2, 3…
Numerical — continuous	Represents a quantity measured on a scale	Customer Lifetime Value

Numbers can also represent categories: Response encoded as 0/1 remains categorical.

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)**

Types of variables

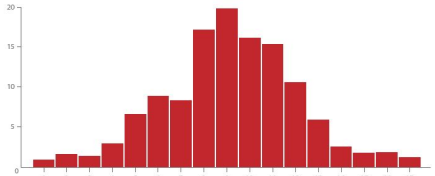
Question	Summary or chart
How many customers responded?	Counts, percentages and a **bar chart**
How are customer values distributed?	Mean, median and a **histogram**
Are there unusually high customer values?	A **box plot**, followed by investigation
Do monthly payments relate to customer value?	A **scatter plot**

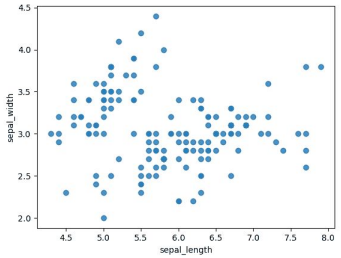


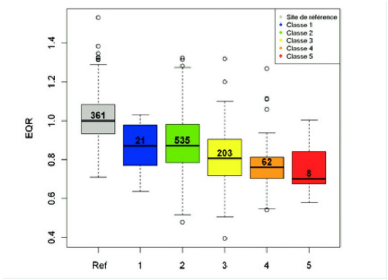
Would you use a histogram or a bar chart to show Yes and No responses?"

Choose the right summary and chart

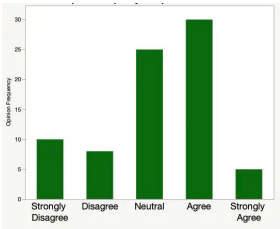
- **Graphiques sur données numériques**
  - **Histogramme, Nuage de points(Scatter plot), Boites a moustaches (box plot)**

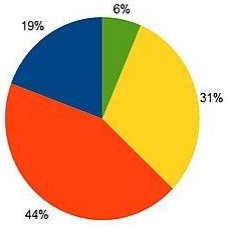






- **Graphiques sur données catégorielles**
  - **Les diagrammes à barres, les diagrammes circulaires (aussi appelés « en camembert »)**





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')
```

Why are predictions imperfect?

- Noise is any unwanted anomaly in the data and due to noise, the class may be more difficult to learn
- Customers with similar recorded features can still have different outcomes.

Possible reasons include:

- **Measurement or recording errors:** a value was entered incorrectly.
- **Incorrect target labels:** a response was recorded as No instead of Yes.
- **Unobserved factors:** relevant information is missing from our features.
- .....

Label these as **possible explanations**, not problems confirmed in your dataset.

Check data quality before modelling

- Data cleaning is the process of finding and correcting or deleting irrelevant, missing, duplicate, or otherwise useless data from a datset.
- Data cleaning means identifying data-quality problems and deciding how to handle them.

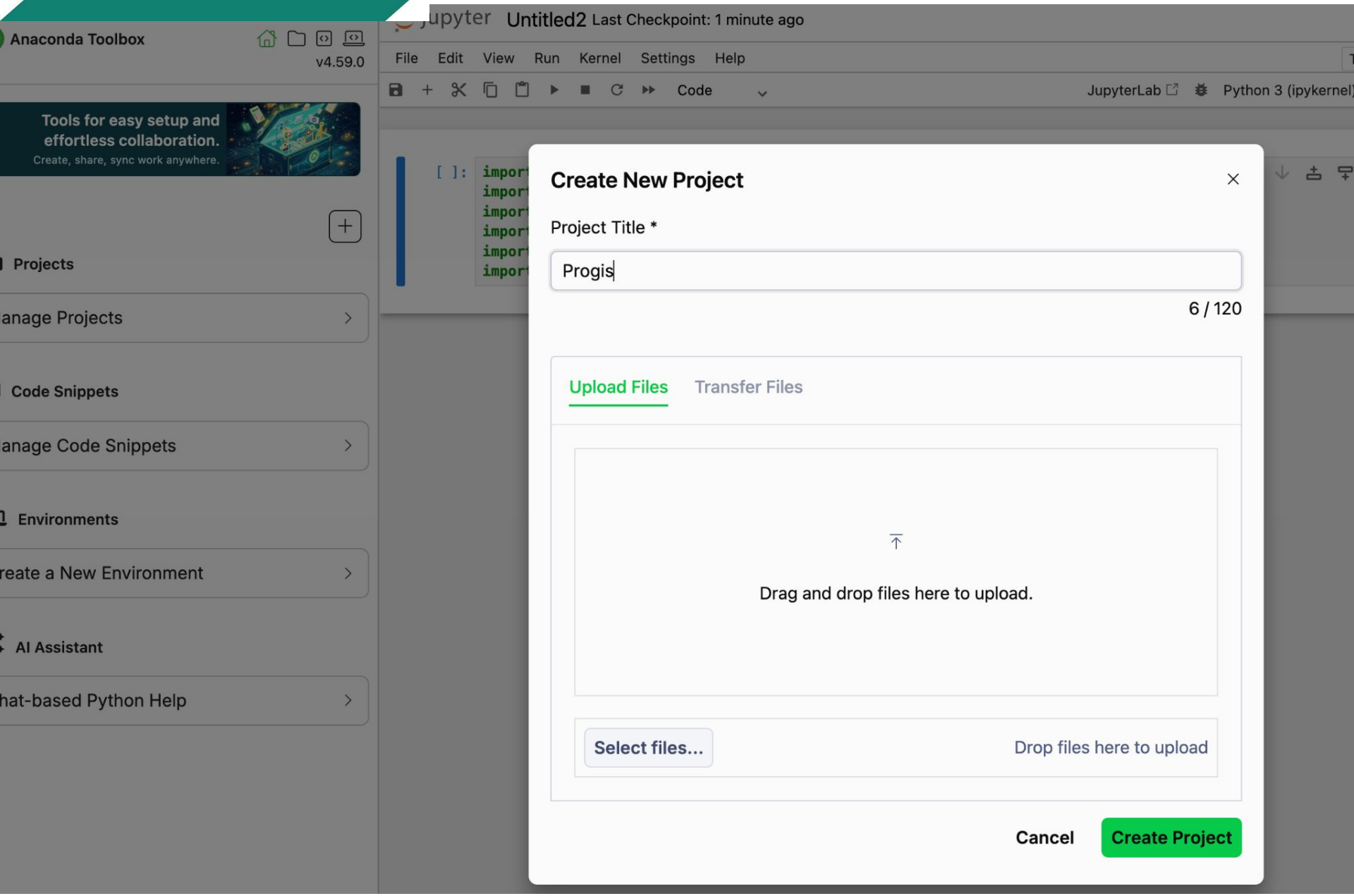
What to check	Example	Possible action
Missing values	An empty monthly-payment field	Investigate; choose an appropriate treatment
Duplicate records	The same customer record imported twice	Verify before removing
Unusual values	A very high customer value	Investigate; keep it if valid

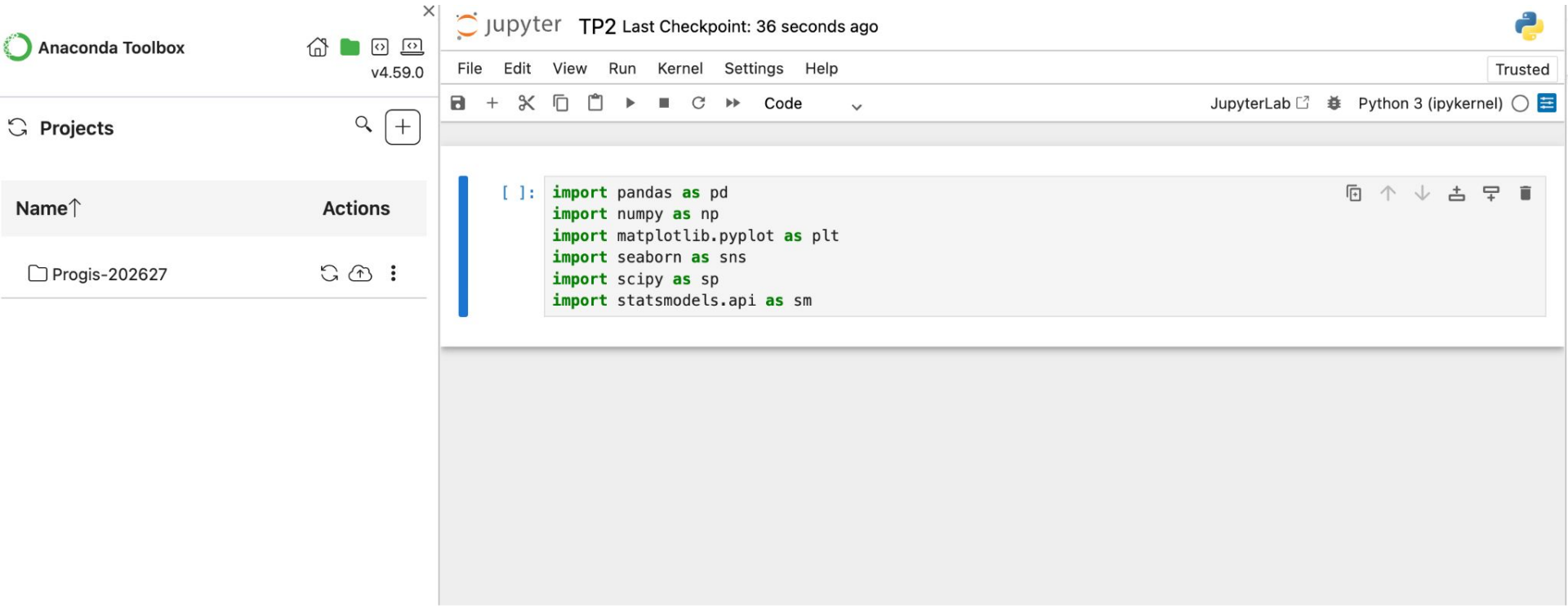
*To perform data cleansing, all incorrect, incomplete, and irrelevant data must be found, then either **replaced**, **removed**, or **modified.***



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

TP


Anaconda Toolbox
v4.59.0
Tools for easy setup and effortless collaboration.
Create, share, sync work anywhere.
Projects
anage Projects
Code Snippets
anage Code Snippets
Environments
reate a New Environment
AI Assistant
hat-based Python Help
upyter Untitled2 Last Checkpoint: 1 minute ago
File Edit View Run Kernel Settings Help
Code
JupyterLab
Python 3 (ipykernel)
Create New Project
Project Title *
Progis
6 / 120
Upload Files
Transfer Files
Drag and drop files here to upload.
Select files...
Drop files here to upload
Cancel
Create Project


Anaconda Toolbox
v4.59.0
Projects
Name
Actions
Progis-202627
jupyter
TP2 Last Checkpoint: 36 seconds ago
File
Edit
View
Run
Kernel
Settings
Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
[ ]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy as sp
import statsmodels.api as sm

Understanding customer data

Our case study

An insurance company has a dataset describing its customers, their policies, and their recorded campaign responses.

Today, we will learn to:

- Identify observations, features, and targets.
- Load and explore a dataset with pandas.
- Summarise data and create simple charts.
- Explain what the results tell us and what remains uncertain.

ANACONDA

jupyter Notebook 7.5.7



Learning from insurance customer data

An insurance company has customers who pay for car insurance.

Each customer has a history with the company:

how long their policy has existed,

how many policies they hold, and

whether they have unresolved complaints…..

What we know about customers
Their monthly insurance payment, policy history, coverage, sales channel, and unresolved complaints.

Task	Question
Regression	Can we estimate the customer value recorded in the dataset?
Classification	Can we predict the recorded campaign response: Yes or No?

Our goal: learn how to prepare data, train a model, and evaluate its predictions.

Our customer dataset

Each row represents one customer.

Our dataset contains **9,134 customers** and **9 columns**, describing:

- **Their insurance:** monthly payment, coverage, and number of policies.
- **Their relationship with the company:** months since the policy began and unresolved complaints.
- **Their sales channel:** agent, branch, call centre, or web.
- **Their recorded customer value and campaign response.**

Customer ID identifies each person; we will not use it to make predictions.

Customer	Monthly Premium Auto	Months Since Policy Inception	Number of Policies	Number of Open Complaints	Sales Channel	Coverage	Customer Lifetime Value	Response
BU79786	69	5	1	0	Agent	Basic	2763.519279	No
QZ44356	94	42	8	0	Agent	Extended	6979.535903	No
AI49188	108	38	2	0	Agent	Premium	12887.43165	No
WW63253	106	65	7	0	Call Center	Basic	7645.861827	No
HB64268	73	44	1	0	Agent	Basic	2813.692575	No
OC83172	69	94	2	0	Web	Basic	8256.2978	Yes
XZ87318	67	13	9	0	Agent	Basic	5380.898636	Yes
CF85061	101	68	4	0	Agent	Premium	7216.100311	No
DY87989	71	3	2	0	Agent	Basic	24127.50402	Yes

Our customer dataset

Column	Plain-language meaning
Customer	A code identifying the customer, like a student number
Monthly Premium Auto	The monthly amount the customer pays for car insurance. *Premium* means the price of insurance
Months Since Policy Inception	How many months have passed since the insurance policy began
Number of Policies	How many insurance contracts the customer holds
Number of Open Complaints	How many customer complaints remain unresolved
Sales Channel	How the insurance was sold: Agent, Branch, Call Center, or Web
Coverage	The coverage category: Basic, Extended, or Premium. Here, *Premium* is a category name
Customer Lifetime Value	The customer value number supplied in the dataset
Response	Whether the dataset records a campaign response: Yes or No

Customer Lifetime Value estimates a customer's financial value over their relationship with a company. In our dataset, CLV is already provided, but its exact calculation is not documented.

Observations, features, and targets

Observation: one customer
Each row describes a customer.

Features (X): information used to make a prediction
Examples: monthly insurance payment, number of policies, and months since the policy began.

Target (y): the value we want to predict
The target depends on our question:

Our question	Target	Task
What is the customer's recorded value?	Customer Lifetime Value	**Regression:** predict a number
Did the customer respond to the campaign?	Response: Yes / No	**Classification:** predict a category

The target must not appear among the input features.

? **If we predict campaign response, which column must we remove from X?**

Observations, features, and targets

Observation: one customer
Each row describes a customer.

Features (X): information used to make a prediction
Examples: monthly insurance payment, number of policies, and months since the policy began.

Target (y): the value we want to predict
The target depends on our question:

Our question	Target	Task
What is the customer's recorded value?	Customer Lifetime Value	**Regression:** predict a number
Did the customer respond to the campaign?	Response: Yes / No	**Classification:** predict a category

The target must not appear among the input features.



If we predict campaign response, which column must we remove from X?

Response

Numerical and categorical variables

- Columns contain different types of information.
- Their type helps us choose how to summarize, visualize, and prepare them for a model.

Type	Meaning	Examples from our dataset
Numerical	Quantities we can measure or count	Monthly Premium Auto, Number of Policies, Customer Lifetime Value
Categorical	Groups or labels	Sales Channel, Coverage, Response



Is the number `of Open Complaints` numerical or categorical? Why?

Working with data in pandas

Pandas is a Python library for working with tables.

A **DataFrame** is a table with rows and named columns.

We will use it to explore our customer dataset.

```
import pandas as pd

df = pd.read_csv("Progis_Customer_Data.csv")

df.head()
```

A CSV file stores tabular data as text. Keep it in the same folder as your notebook for this example.



"Does `df.head()` load only five customers?"

A first look at our dataset

Before building a model, we check the table's size, column names, and data types.

Python command	What it tells us
`df.head()`	Shows the first five rows
`df.shape`	Gives the number of rows and columns
`df.columns`	Lists the column names
`df.info()`	Summarises column types and non-missing counts

Our dataset contains **9,134 rows** and **9 columns**.

? Does having 9 columns mean we have 9 input features for our model?

End

EMAIL: bahareh.afshinpour@iepg.fr