

# INTRODUCTION AU COURS D'ÉCONOMÉTRIE

**IG5**

2022/2023

## Économétrie : quelques définitions

### Définition 1 :

L'**économétrie** consiste en l'application des techniques **mathématiques** et **statistiques** à l'analyse de phénomènes économiques.

### Définition 2 :

L'**économétrie** est une branche de la science économique qui a pour objectif d'estimer et de tester les modèles économiques, à partir de données issues de l'observation du fonctionnement réel de l'économie ou provenant d'expériences contrôlées.

- ▶ Application des **statistiques** à l'analyse des **phénomènes économiques**.
- ▶ **Domaine de l'économie** visant à estimer et tester des **modèles économiques** basés sur des données du système économique réel ou d'expériences contrôlées.
- ▶ **Objectif** : vérifier la validité d'un **modèle économique** avec des jeux de données réels.

## Exemples

- Effet de l'expérience et de l'éducation sur le salaire ?
- Effet du prix du tabac sur la consommation ?
- Effet de la R&D sur les innovations des entreprises françaises ?

## C'est quoi un Modèle ?

Un modèle est une **représentation simplifiée de la réalité**, que l'on peut utiliser pour prédire ce qui se passerait dans certaines conditions. Ça peut être un dessin, une équation physique, une fonction mathématique, une courbe. . .

- Par exemple, si je lâche une pomme de 100 g depuis une hauteur de 4 mètres, en combien de temps tombera t'elle sur la tête de ce bon vieux Newton ? On peut prédire cela avec les équations de Newton.
- ▶ Pour prédire le prix d'un appartement en fonction de toutes ses caractéristiques, quelle est l'équation mathématique à entrer dans la machine ? Et quelle est l'équation pour reconnaître un chat sur une photo ?

### ☺ La solution ?

Laisser la machine trouver le modèle qui correspond le mieux à votre Dataset : c'est l'apprentissage supervisé.

### C'est quoi l'apprentissage supervisé ?



- ✓ Acheter un livre de traduction chinois français.
- ✓ Trouver un professeur de chinois.
- ▶ Le rôle du professeur ou du livre de traduction sera de **superviser votre apprentissage** en vous fournissant des **exemples** de traductions français chinois que vous devrez mémoriser.

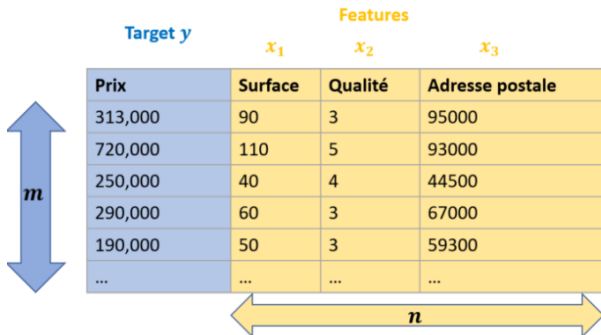
Pour maîtriser l'apprentissage supervisé, il faut absolument comprendre et connaître les 4 notions suivantes :

- Le **Dataset** (Importer un Dataset  $(x, y)$  qui contient nos exemples)
- Le **Modèle et ses paramètres** (Développer un Modèle aux paramètres aléatoires)
- La **Fonction Coût** (Développer une Fonction Coût qui mesure les erreurs entre le modèle et le Dataset)
- **Machine Learning** ( Développer un Algorithme d'apprentissage pour trouver les paramètres du modèle qui minimisent la Fonction Coût)

Dans le cadre de l'apprentissage supervisé, il existe deux sous-catégories : la **régression** et la **classification**.

## Régression :

*Exemple de Dataset sur des appartements*



The diagram shows a table with 6 rows and 4 columns. The first column is labeled 'Target y' and contains values: 313,000, 720,000, 250,000, 290,000, 190,000, and ... . The next three columns are labeled 'Features' and contain values: Surface (90, 110, 40, 60, 50, ...), Qualité (3, 5, 4, 3, 3, ...), and Adresse postale (95000, 93000, 44500, 67000, 59300, ...). A blue double-headed vertical arrow on the left is labeled 'm', indicating the number of rows. A yellow double-headed horizontal arrow at the bottom is labeled 'n', indicating the number of columns.

Target $y$	Features		
	$x_1$	$x_2$	$x_3$
Prix	Surface	Qualité	Adresse postale
313,000	90	3	95000
720,000	110	5	93000
250,000	40	4	44500
290,000	60	3	67000
190,000	50	3	59300
...	...	...	...

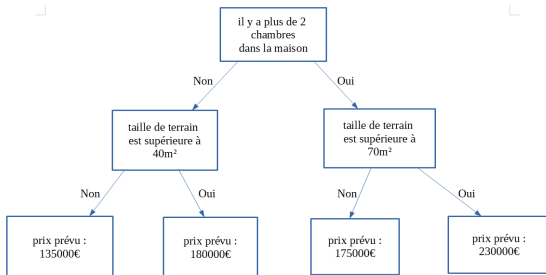
**Par convention:**

**$m$ :** nombre d'exemples

**$n$ :** nombre de features

- ▶ Régression simple,
- ▶ Régression polynomiale.
- ▶ Régression multiple.

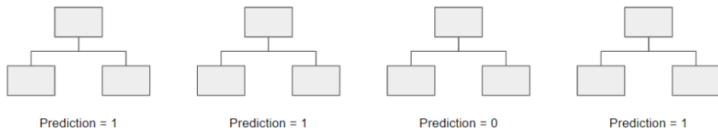
## Decision Tree (Arbre de décision) :



- ▶ Les arbres de décision sont un modèle populaire, utilisé dans la recherche opérationnelle, la planification stratégique et le Machine Learning.
- ▶ Chaque rectangle ci-dessus est appelé un **nœud**.



## Random Forest (Forêt d'arbres) :



- Quel est l'intérêt de cette méthode ?  
En s'appuyant sur un modèle de prévalence de la majorité (c'est-à-dire sur lequel la majorité l'emporte), il réduit le risque d'erreur d'un arbre individuel.

# Classification

## Exemple :

Classer un email en tant que 'spam' ou 'non spam'.

Dans ce genre de problème, on aura un Dataset contenant une variable target  $y$  pouvant prendre 2 valeurs seulement, par exemple 0 ou 1

- si  $y = 0$ , alors l'email n'est pas un spam
- si  $y = 1$ , alors l'email est un spam

On dit également que l'on a 2 classes, c'est une **classification binaire**.

▶ Régression logistique, Support Vector Machine (SVM),...

## Alors c'est quoi un modèle économique ?

Le modèle économique, ou **business model**, décrit la manière dont **l'entreprise** va créer de la valeur pour l'ensemble des parties prenantes et gagner de l'argent.

Plusieurs éléments sont très importants pour élaborer un modèle économique : la mission de l'entreprise, la vision des créateurs d'entreprises, les **objectifs**, les **ressources** et les **compétences**. Il traduit la mise en œuvre de la stratégie des entrepreneurs.

Le modèle économique se construit en se focalisant sur les **clients**, car ceux sont eux qui doivent bénéficier de la valeur apportée par le produit ou le service de l'entreprise, et ceux sont eux qui procurent les revenus à l'entreprise.

## Techniques statistiques sous-jacentes ?

- Régression linéaire simple
- Régression linéaire multiple
- Généralisations comme la régression logistique (économétrie sur variables qualitatives : **Classification**)
- Séries temporelles

## Exemples.

- Effet du prix du tabac sur la consommation,
- Effet de la recherche et du développement sur l'innovation de l'industrie,
- Effet des politiques gouvernementales sur la richesse de la population.

## Mots clés : notions à revoir ?

- Variables aléatoires
- Échantillon
- Techniques d'estimation (méthode des moments, méthode du max de vraisemblance, MCO...)
- Tests (test de student, test de Fisher...)

## Et le scoring c'est quoi ?

Application de la technique de classement **prédictif** à certaines **problématiques de l'entreprise**.

La construction d'un **score** fait appel à la **modélisation prédictive** et on parle de score quand la variable à prédire a 2 modalités possibles (oui/non).

**Origine** : "**credit scoring**" consiste à prédire la probabilité d'être un bon ou un mauvais payeur dans l'octroi d'un prêt.

## Exemple :

- Dans l'assurance de risque, des produits obligatoires (automobile, habitation)
  - Soit prendre un client à un concurrent
  - Soit faire monter en gamme un client que l'on détient déjà.
- D'où les sujets dominants
  - Attrition
  - Ventes croisées (cross-selling)
  - Montées en gamme (up-selling)
- Et les exemples de scores
  - Score de risque  
Prédire les impayés ou la fraude
  - Score d'appétence (ou de propension)  
Prédire l'achat d'un produit ou service
  - Score d'attrition  
Prédire le départ d'un client vers un concurrent
  - Mais aussi en médecine : diagnostic (oui/non), analyse des courriels : spam (oui/non)

- Définition de la variable à expliquer
  - En médecine, souvent naturelle  
*Un patient a, ou non, une tumeur (et encore faut-il distinguer les différents stades de la tumeur).*
  - Dans la banque : qu'est-ce qu'un client non risqué ?  
*aucun impayé ? 1 impayé ? ... ?*
  
- Biais de sélection
  - En risque, certaines demandes sont refusées et on ne peut donc pas mesurer la variable à expliquer.
  - En appétence : certaines populations n'ont jamais été ciblées. Le produit ne leur a donc pas été proposé.

## **Techniques (statistiques) sous-jacentes ?**

- Techniques de régression
- Régression logistique
- Analyse discriminante
- Arbre de décision
- Bayes naif

## **Notions à définir**

- Validation des modèles
- Matrice de confusion, sensibilité, spécificité, courbes ROC



## Mais avant, préparation des données - Analyse exploratoire des données

- Les données sont-elles fiables ? (valeurs aberrantes, valeurs manquantes ?) et que faire si non ?
- Etudier la distribution des données
- Faut-il recoder les variables ? regrouper des modalités ?
- Faut-il transformer les variables continues ?
- Faut-il discrétiser les variables continues ?
- Faut-il se contruire d'autres variables (indicateurs pertinents, composantes principales) ?

## Complément valeurs manquantes

- toujours vérifier que les valeurs manquantes ne viennent pas d'un pb technique ou d'individus qui ne devraient pas se trouver dans la base.
- des solutions ? à choisir selon les cas et toujours avec prudence !!!
  - supprimer les obs
  - ne pas utiliser la variable concernée
  - traiter la valeur manquante comme une valeur à part entière
  - imputation (plusieurs techniques et jamais neutre...)
  - la remplacer grâce à une source externe

## Complément : Pourquoi discrétiser les variables continues ?

- appréhender des liaisons non linéaires voire non monotones entre les variables continues et la variable à expliquer
- neutraliser les valeurs extrêmes
- gérer les valeurs manquantes ou imprécises
- traiter simultanément des données quantitatives et qualitatives