cegos
Beyond knowledge

Analyser vos données grâce à l'intelligence artificielle

Pour accéder au Learning Hub à votre parcours et votre documentation

https://fr.mylearninghub.cegos.com/

Identifiant : Adresse email professionnelle

Première connexion : Cliquez sur "Créer ou réinitialiser votre mot de passe".

Tél : 01 55 00 93 07

Email : elearning@cegos.fr

pour toutes questions techniques

Sommaire interactif

Objectifs de la formation

- Comprendre les principes clés du traitement des données.
- Comprendre les principes fondamentaux de l'Intelligence Artificielle.
- Être capable de mettre en œuvre un projet d'analyse et de visualisation de données avec les techniques d’IA.
- Comprendre les risques liés à l'utilisation de l'IA pour le traitement des données de son organisation.

Introduction

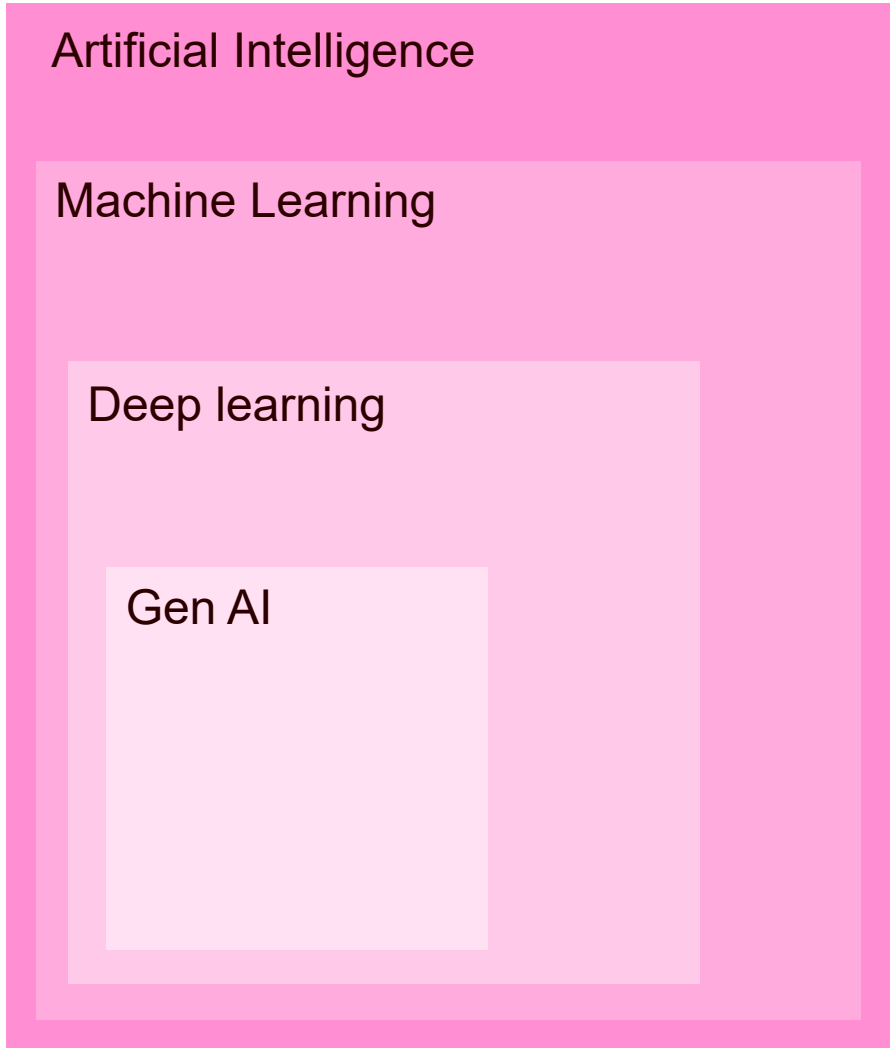
Définitions

- L’Intelligence artificielle (IA)
  - est un domaine de l'informatique qui vise à créer des systèmes capables d'effectuer des tâches qui nécessitent normalement l'intelligence humaine. Elle englobe diverses techniques permettant aux machines d'apprendre, de raisonner et de résoudre des problèmes. Contrairement à l'informatique traditionnelle, où chaque action est précisément programmée, l'IA permet aux systèmes de s'adapter et d'améliorer leurs performances avec l'expérience.
- Le Machine Learning (apprentissage automatique)
  - est une branche centrale de l'IA. Il s'agit d'une approche qui permet aux systèmes informatiques d'apprendre et de s'améliorer à partir de l'expérience, sans être explicitement programmés pour chaque tâche. En utilisant des algorithmes et des modèles statistiques, les systèmes de Machine Learning peuvent identifier des motifs dans les données, faire des prédictions et prendre des décisions avec un minimum d'intervention humaine. Cette approche est particulièrement utile pour traiter des problèmes complexes où les règles traditionnelles de programmation seraient difficiles à définir.

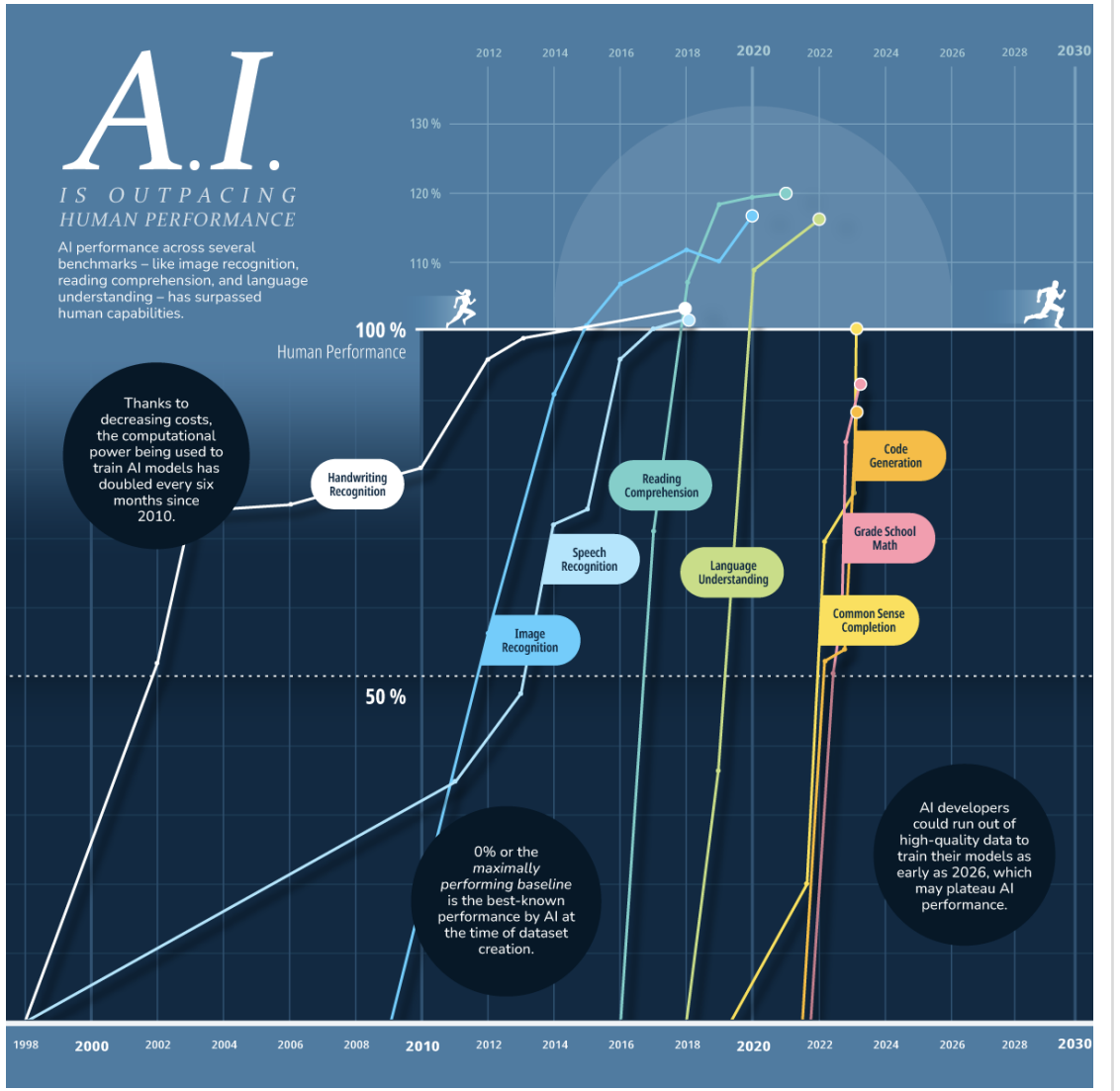
Définitions

- Le Deep Learning (apprentissage profond)
  - est une technique avancée de Machine Learning utilisant des réseaux de neurones artificiels à plusieurs couches. Ces réseaux traitent de grandes quantités de données pour apprendre et reconnaître des motifs complexes. Particulièrement efficace pour des tâches comme la reconnaissance d'images ou le traitement du langage naturel, le Deep Learning permet aux machines d'apprendre des représentations de données de plus en plus abstraites et sophistiquées.
- L'Intelligence artificielle générative (gen AI)
  - est un domaine de l'IA qui se concentre sur la création de contenu nouveau et original. En utilisant des modèles d'apprentissage avancés, elle peut produire du texte, des images, de la musique ou d'autres types de données qui semblent avoir été créés par des humains. Cette technologie ouvre de nouvelles possibilités en matière de création assistée par ordinateur, mais soulève également des questions éthiques sur l'authenticité et la propriété intellectuelle.

Définition



À partir de 2012, l'IA surpasse l'homme dans beaucoup des tâches cognitives



Source https://intelligence-artificielle.developpez.com

Raison de l’accélération de l’IA

Puissance des machines

- 1941 : Z3 (Konrad Zuse) - Environ 1 opération par seconde
- La machine Z3, créée par Konrad Zuse, était l'un des premiers ordinateurs électroniques programmables
- 2022 : 1 milliard de milliards d'opérations par seconde
- 2025 : 100 milliards de milliards d'opérations par seconde (projet d'Elon Musk)
- 2030 : 1000 milliards de milliards d'opérations par seconde (projet Dell Intel)

Algorithmes nb paramètres

- Les Algo d’IA des années 2000 avaient typiquement autour d'un million de paramètres
- AlexNet (2012) qui a relancé l'intérêt pour les réseaux de neurones profonds, avait 60 millions de paramètres
- VGG (2014) et ResNet (2015) ont poursuivi cette tendance avec respectivement 138 millions et 60 millions de paramètres
- La complexité a continué à augmenter avec des algos comme Inception (2015) et Transformer (2017), ce dernier étant la base des modèles GPT suivants
- GPT-1 (2018) a marqué un tournant avec 110 millions de paramètres, suivi par GPT-2 (2019) avec 1,5 milliard de paramètres
- GPT-3 (2020) a réalisé un bond quantique à 175 milliards de paramètres
- Finalement, GPT-4 (2023) repousse les limites avec un trillion de paramètres

Quantité de données apprentissage

- En 2014, environ 2 millions d'images étaient disponibles pour l'apprentissage
- En 2024, ce nombre a explosé pour atteindre 100 milliards d'images, auxquelles s'ajoutent 1000 milliards de mots, témoignant d'une croissance exponentielle des données disponibles pour entraîner des modèles d'intelligence artificielle

Chronologie de l'intelligence artificielle

1960 : Début de l'IA

1980 : IA : Systèmes Experts

2010 : IA predictive : machine learning, deep learning

2020 : IA par Renforcement : Alphafold

2022 : IA Générative : ChatGPT

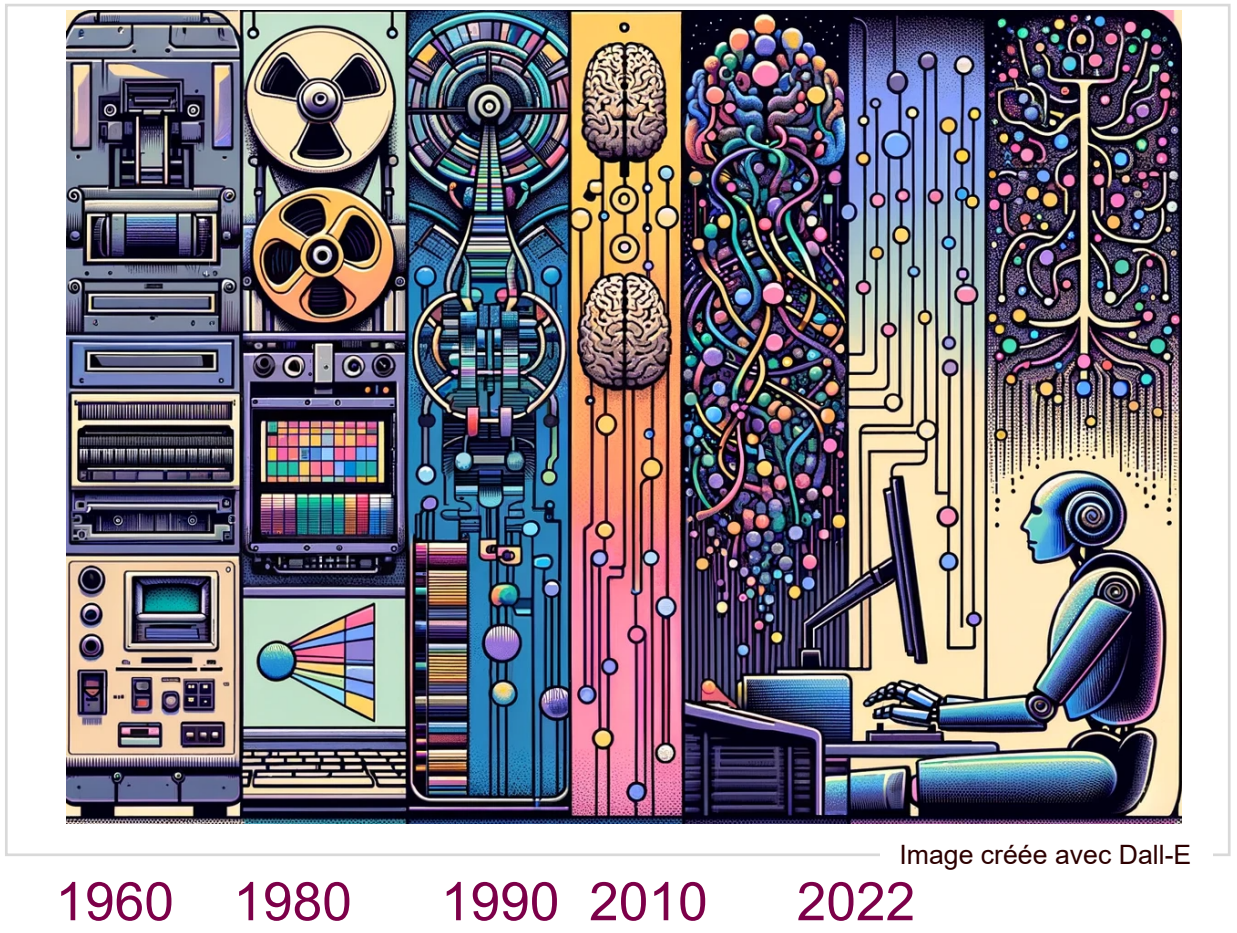


Image créée avec Dall-E

L'IA démystifiée

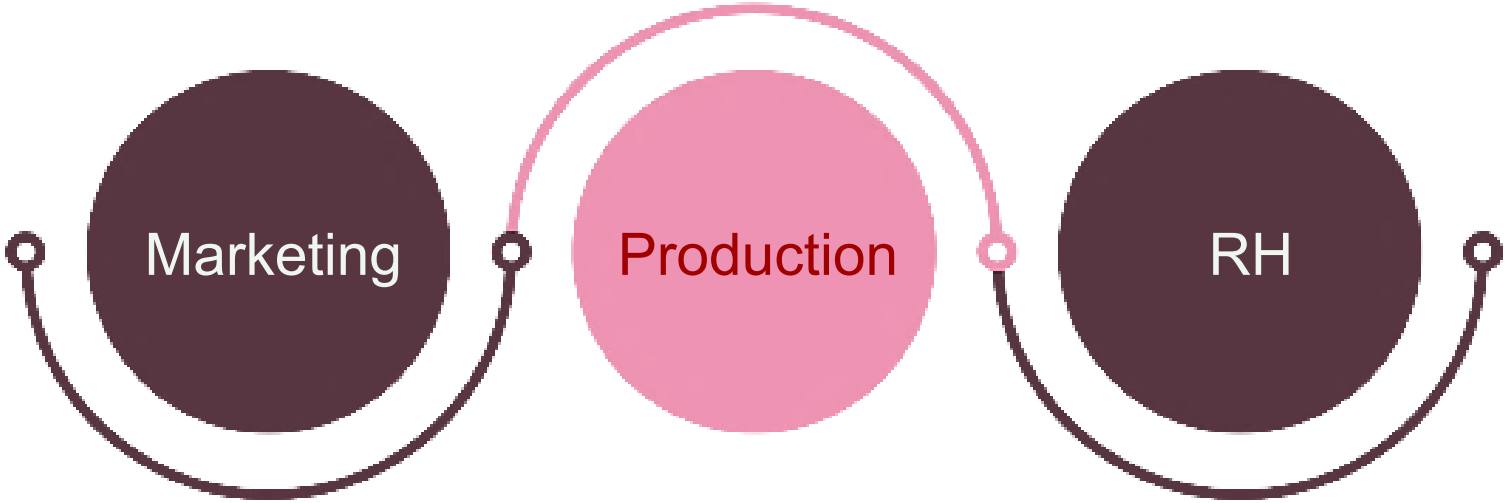
L'IA aujourd'hui c'est :

- Pas de conscience artificielle
- Pas de remplacement humain total
- Mais des outils puissants pour :
  - ▸Automatiser les tâches répétitives
  - ▸Analyser de grands volumes de données
  - ▸Prédire à partir d'historiques
  - ▸Générer du contenu
- Exemple concret : Gmail
  - ▸Tri automatique des spams (classification)
  - ▸Suggestions de réponses (génération)
  - ▸Détection des urgences (analyse)

Applications concrètes 2025

- Maintenance prédictive
- Contrôle qualité visuel
- Optimisation processus

Par fonction entreprise :



- Segmentation clients avancée
- Prédiction du churn
- Personnalisation temps réel

- Présélection CV
- Prédiction turnover
- Chatbot RH

Les types d'IA Panorama pratique 2025

L'IA discriminative

- Classifie et prédit
- Besoins : données labellisées historiques
- Exemples :
  - Scoring crédit (approuvé / refusé)
  - Détection fraude (normal / suspect)
  - Prévision ventes (chiffres)

L'IA générative

- Crée du nouveau contenu
- Besoins : grands volumes de données
- Exemples :
  - ChatGPT (texte)
  - Dall E (images)
  - Copilot (code)

Machine Learning vs Logiciel

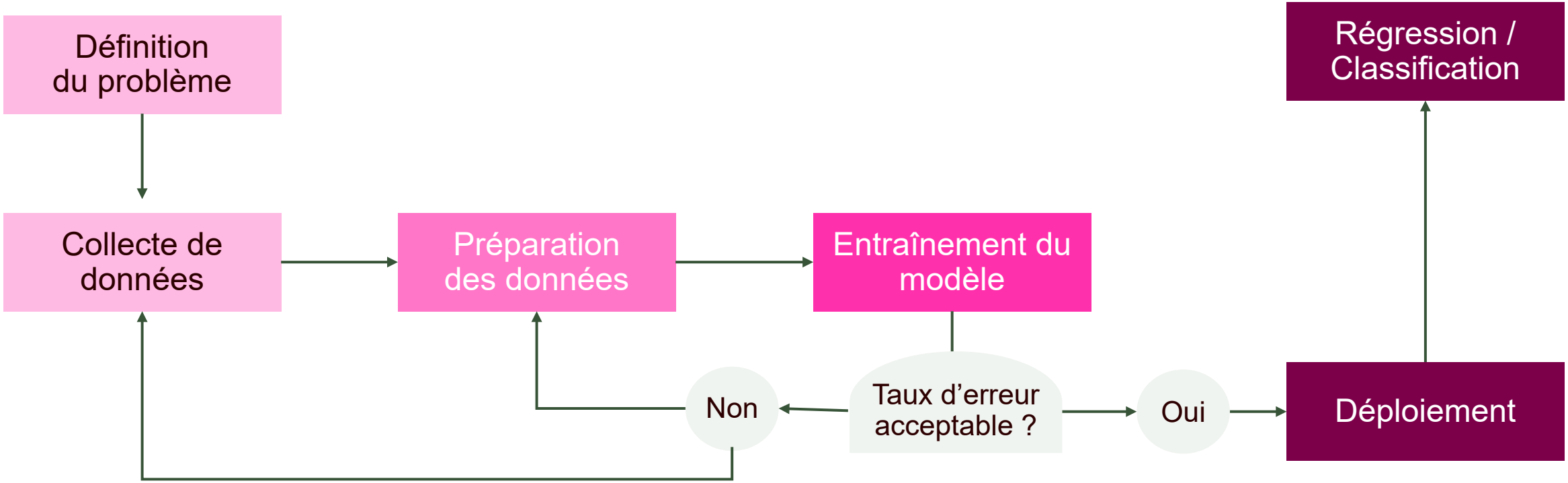
Le Machine Learning part des données d'entrée et des résultats attendus pour que l'ordinateur découvre lui-même les règles. L'algorithme « apprend » à partir des exemples et génère un programme capable de traiter de nouvelles données afin de produire de nouvelles sorties.

La programmation traditionnelle repose sur un schéma simple : on donne à l'ordinateur des données d'entrée et un programme (les règles écrites par le développeur). L'ordinateur exécute alors ces instructions et fournit un résultat. Ici, toutes les règles sont définies à l'avance par l'humain, et la sortie est toujours déterminée par le code.

Le Machine Learning

- Le Machine Learning fonctionne comme l'apprentissage humain : il apprend à partir d'exemples. Imaginons un enfant qui apprend à reconnaître les chats :
  - Il voit de nombreux chats (données d'entraînement)
  - Il identifie des caractéristiques communes (features)
  - Il apprend à reconnaître un chat (modèle)
  - Il teste sa capacité sur de nouveaux chats (validation)

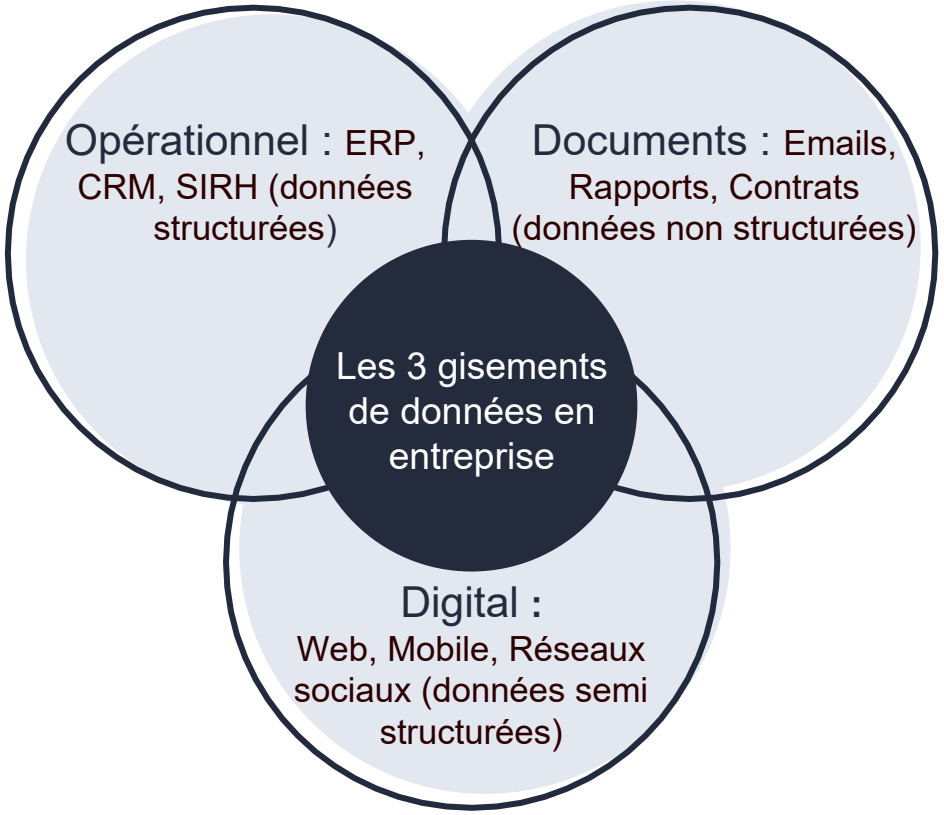
Cycle de vie d'un projet de machine learning



Collecte des données

Sources de données

Vue d'ensemble



- Exemple concret :
  - Pour analyser la satisfaction client
    - ERP : historique achats
    - CRM : interactions service client
    - Digital : avis en ligne

Exploration rapide

Explorer 100 lignes manuellement

Analyser 10 000 lignes avec Excel

Traiter 1M+ lignes avec outils spécialisés

La méthode 1 10 100

Pour 100 lignes, chercher :

- Valeurs extrêmes
- Formats inhabituels
- Incohérences évidentes

Collecte moderne Architecture type

De la source au stockage :

1. Captation :
   - API, Connecteurs, ETL
2. Stockage :
   - Data Lake (brut) → Data Warehouse (propre)
3. Accès :
   - Requêtage SQL, Export fichiers, API

Exemple : Collecte données ventes

1. Temps réel :
   - API Shopify → Lake
2. Quotidien :
   - Export SAP → Warehouse
3. Mensuel :
   - Rapports consolidés

Les 3 piliers de la collecte de données

- Pertinence
  - Alignement avec les objectifs business
  - Focus sur les données à forte valeur ajoutée
  - Éviter la collecte excessive
- Qualité
  - Définition claire des standards de qualité
  - Validation à la source
  - Procédures de contrôle
- Efficacité
  - Automatisation des processus de collecte
  - Optimisation des ressources
  - Réduction des coûts de traitement

Les méthodes de collecte

- Collecte active → Contrôle maximal mais coût élevé
  - Formulaires et questionnaires
  - Saisie manuelle
  - Capteurs et IoT
- Collecte passive → Volume important mais qualité variable
  - Logs systèmes
  - Trackers web
  - Données transactionnelles
- Collecte hybride
  - Combine les approches
  - Validation croisée
  - Enrichissement mutuel

La stratégie d’échantillonnage (sampling)

- Pourquoi échantillonner ?
  - Réduire les coûts de collecte
  - Accélérer le traitement
  - Tester rapidement des hypothèses
- Les Méthodes
  - Aléatoire simple : Chaque élément a la même probabilité
  - Stratifié : Respecte les proportions des sous-groupes
  - Systématique : Sélection à intervalle régulier
- Points d'Attention
  - Taille d'échantillon minimale
  - Représentativité
  - Biais potentiels

Gouvernance de la collecte

- Cadre Réglementaire
  - Conformité RGPD
  - Bases légales de collecte
  - Durées de conservation
- Processus qualité
  - Validation des sources
  - Contrôles automatisés
  - Gestion des anomalies
- Documentation
  - Dictionnaire des données
  - Procédures de collecte
  - Traçabilité des modifications

Rôles et responsabilités

- Data Owner (Responsable Métier) Exemple Directeur Marketing pour les données clients
  - Définit les règles métier et les priorités
  - Valide les usages et droits d'accès
  - Fixe les objectifs de qualité
- Data Steward (Garant Technique) Exemple : Architecte Data pour les schémas de données
  - Assure la qualité technique et la conformité
  - Gère les métadonnées et la documentation
  - Implémente les contrôles automatisés
- Data Collector (Opérationnel) Exemple Commercial pour la saisie CRM
  - Exécute la collecte selon les procédures
  - Applique les contrôles qualité
  - Remonte les anomalies

Types de données et pièges associés

Dates :

- Formats multiples (FR, US, ISO)
- Fuseaux horaires mélangés
- Années sur 2 ou 4 chiffres

Texte :

- Accents et caractères spéciaux
- Casse incohérente
- Abréviations variables

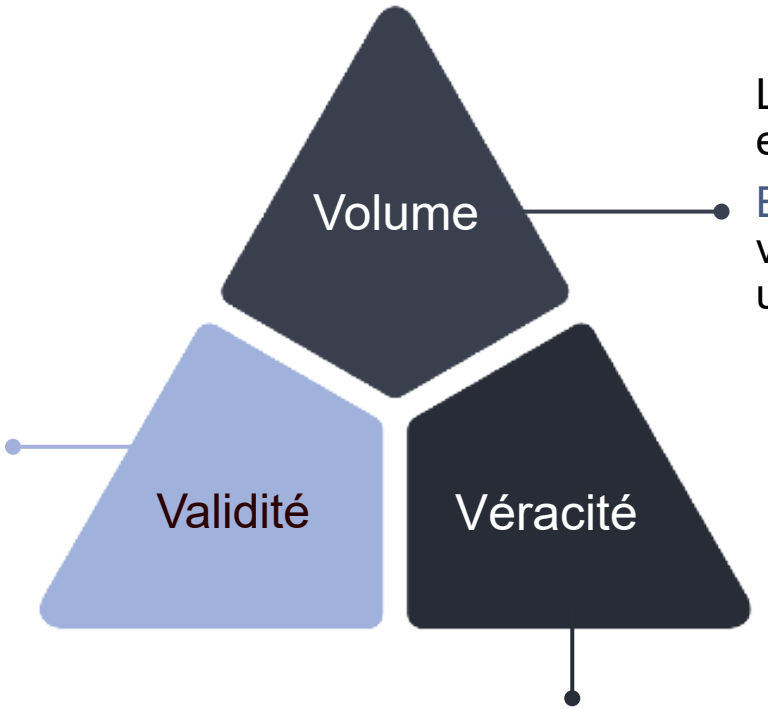
Numériques :

- Séparateurs décimaux (,/.)
- Unités mélangées (€/$)
- Arrondis différents

Validation des données

Approche pratique

Règle des 3V



Les données sont-elles complètes ?

Exemple : 90 % des ventes doivent avoir un code produit

Les données sont-elles cohérentes ?

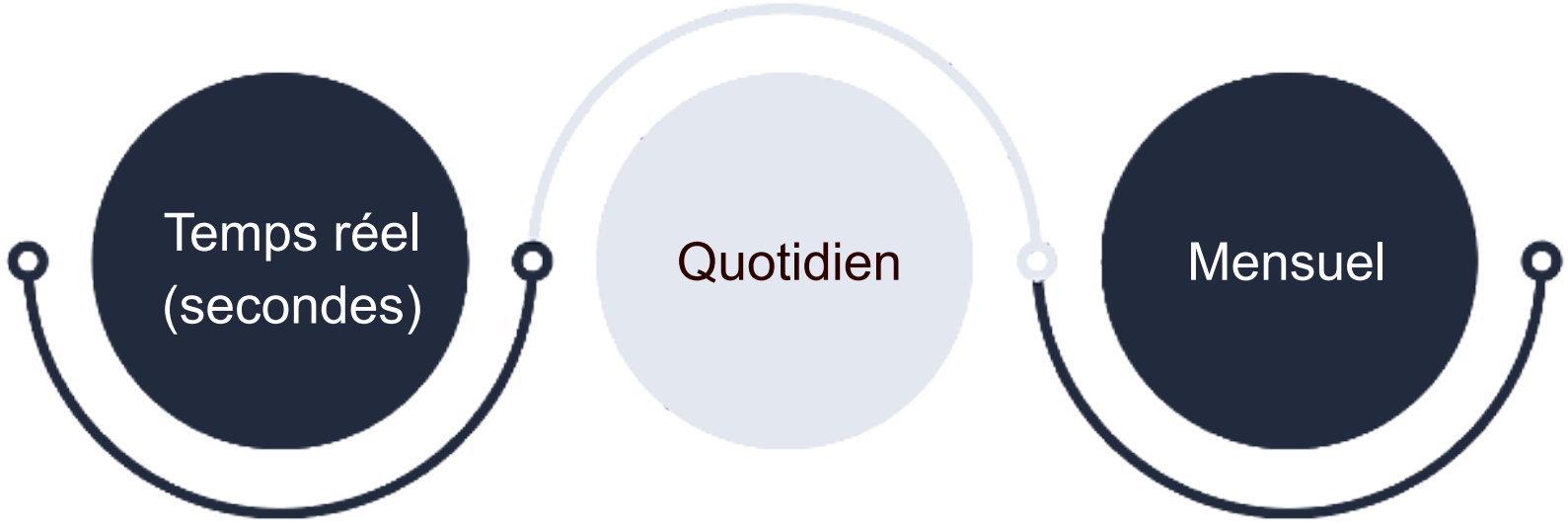
Exemple : Pas de date de livraison avant date commande

Les données sont-elles exactes ?

Exemple : Prix unitaire × Quantité = Montant total

Fréquence de collecte

Impact business



- Fraude bancaire
- Monitoring production
- Trading algorithmique

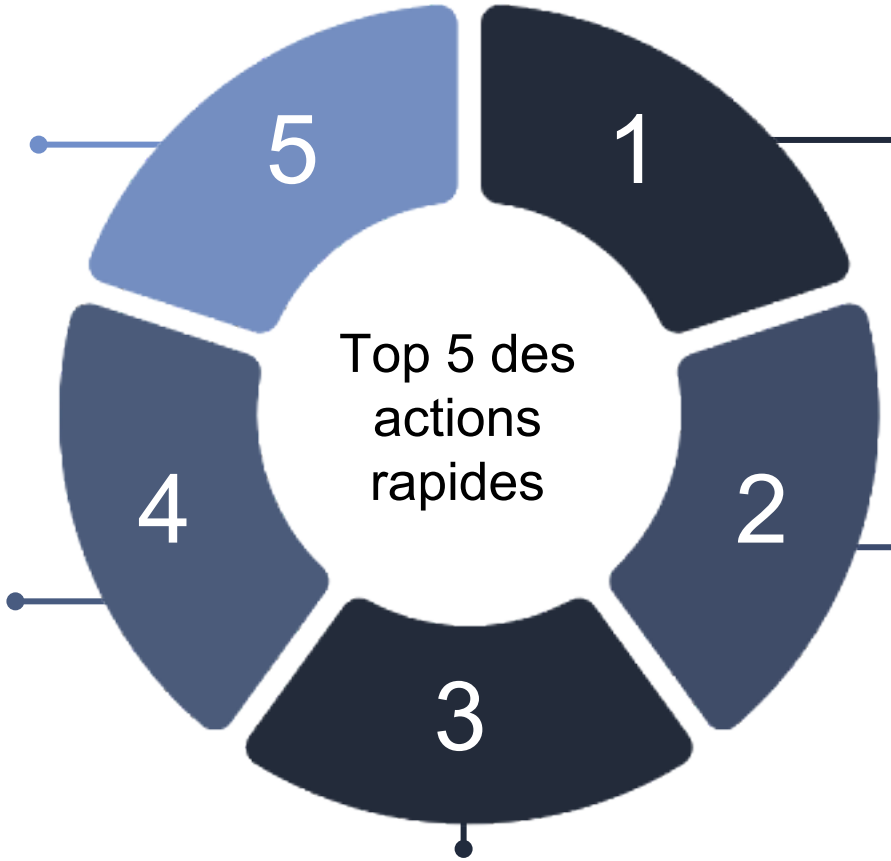
- Ventes magasins
- Stock produits
- KPIs opérationnels

- Reporting financier
- Analyse performance
- Tableaux de bord stratégiques

Coûts cachés de la collecte

- Équation simple : Coût total =
  - Coût technique (stockage, traitement)
  - Coût humain (collecte, maintenance)
  - Coût opportunité (données manquantes)
  - Valeur générée (décisions, optimisations)
- Exemple : Collecte données clients
  - Technique : 10K € / an
  - Humain : 0.5 ETP = 30K € / an
  - Opportunité : 50K € ventes manquées
  - Valeur : 200K € upsell → ROI positif malgré coûts importants

Quick wins collecte de données



Former les utilisateurs aux bonnes pratiques

Standardiser la saisie avec des listes déroulantes

Créer des tableaux de bord de qualité basiques

Mettre en place des contrôles de cohérence simples

Documenter la signification de chaque champ

Gains typiques :

- 60 % d'erreurs de saisie
- 40 % de temps de nettoyage

De la Collecte à la Préparation

Les meilleures données du monde sont inutiles si elles ne sont pas correctement préparées pour l'analyse

- Ce que nous avons vu
  - Les sources de données
  - Les méthodes de collecte
  - La gouvernance et les rôles
- Ce qui nous attend
  - Le nettoyage des données
  - L'enrichissement
  - La validation qualité

cegos
Beyond knowledge

Préparation des données

- L'analyse des données est le nouveau moteur de la performance entreprise

Réalité du terrain

- Temps consacré à la préparation des données : en moyenne, 45 % du temps des data scientists est dédié à la préparation des données (source : Anaconda).
- Coût des erreurs de données : les grandes entreprises peuvent perdre jusqu'à 15 millions d'euros par an en raison d'erreurs de données (source : Gartner).
- Taux d'échec des projets data : environ 76 % des projets data échouent en raison de données mal préparées (source : McKinsey).
- Amélioration de la performance commerciale : un nettoyage efficace des données peut entraîner une augmentation de 23 % de la performance commerciale (source : Harvard Business Review).

Erreurs courantes concrètes

- Exemple 1 : Fusion échouée AVIS/AVISE dans base client = doublon
- Exemple 2 : Date au format US vs EU = commandes mal datées
- Exemple 3 : Codes postaux sur 4 digits = géolocalisation impossible. Solution pour chaque cas présenté avec impact business réel.

Qualité des données
Mesures concrètes

- 3 Métriques essentielles à suivre :

1. Complétude :
   - % de champs remplis (minimum 95 %)
2. Exactitude :
   - % de valeurs correctes (objectif 99 %)
3. Cohérence :
   - % de règles métier respectées (objectif 100 %)

- Exemple : Base clients Telecom

1. Avant nettoyage :
   - 82 % 90 % 95 %
2. Après nettoyage :
   - 98 % 99 % 100 %
3. Impact :
   - +15 % de ventes croisées

Méthode CARD : Correction, Agrégation, Restructuration, Dédoublonnage

- Correction : « Straßbourg » → « Strasbourg », « 75O00 » → « 75000 »
- Agrégation : Regrouper les achats par client/mois plutôt que par ticket
- Restructuration : Séparer « NOM Prénom » en deux colonnes
- Dédoublonnage : Identifier les clients uniques malgré les variantes

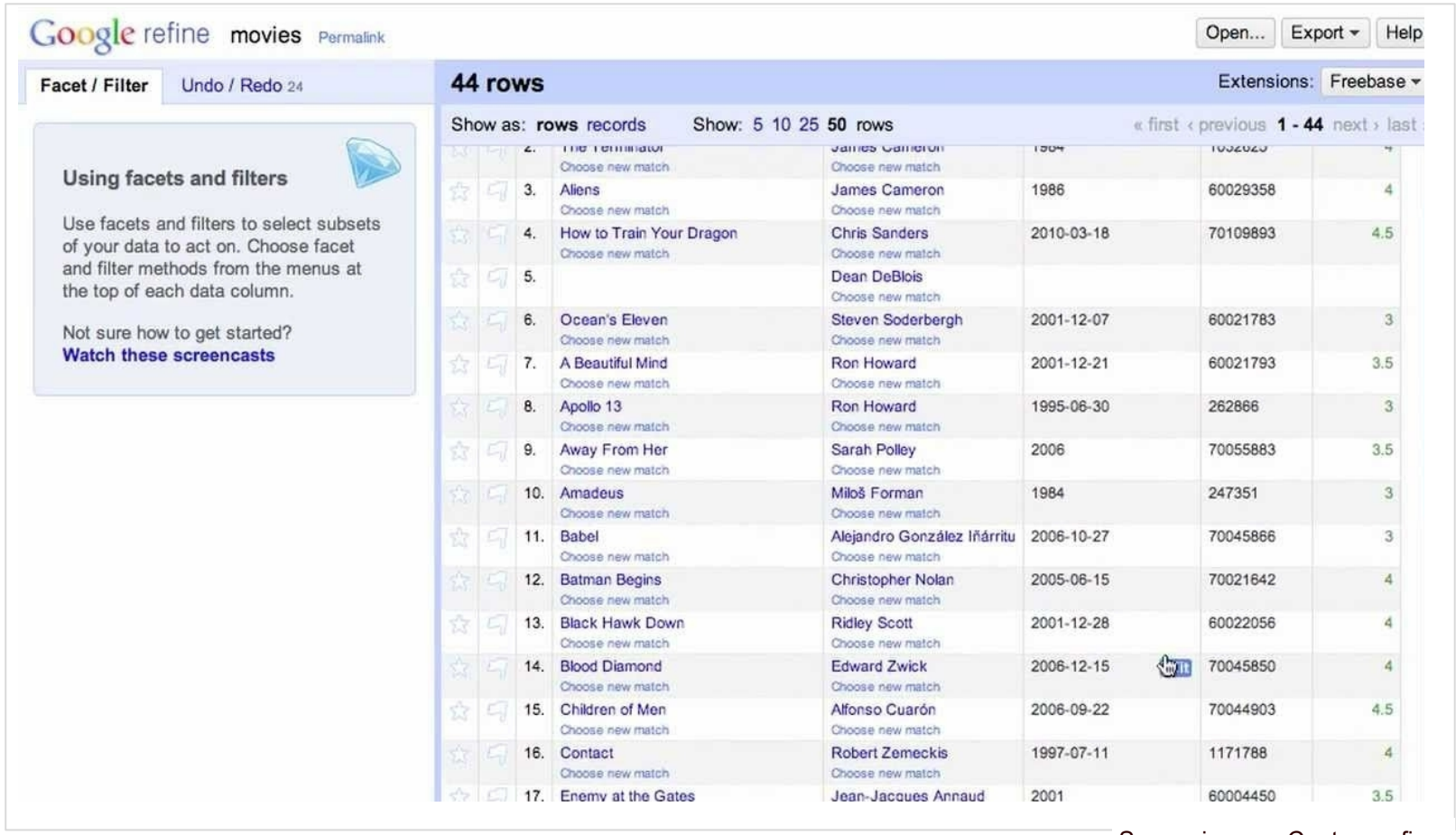
Techniques d'enrichissement

- Internal matching :
  - Fusionner les données CRM/ERP/SIRH Exemple : Ajouter historique SAV aux données commerciales
- Données externes :
  - APIs gratuites utiles Exemple : Code postal → données démographiques INSEE
- Web scraping :
  - Enrichir avec données concurrentielles Exemple : Prix du marché pour positionnement produit

Tests qualité minimum

- Checklist à appliquer systématiquement :
  - Aucune valeur aberrante (ex. : âge > 150 ans)
  - Formats cohérents (dates, montants, codes)
  - Totaux cohérents (CA mensuel = Σ CA quotidien)
  - Règles métier (remise max = 50 %)

OpenRefine : pour éditer ou traiter des données en masse



Source image : Capture refine

https://youtu.be/B70J_H_zAWM

Installer OpenRefine

- https://openrefine.org
- **Ouvrir** OpenRefine → **Create Project** → importer clients_ventes_final.csv
- Vérifier l'aperçu → Create Project (en haut à droite).
- Repères : barre bleue = nom du projet, panneau gauche = Facet/Filter, tableau au centre.

– Cartographie rapide des erreurs

- Sur les colonnes **Nom**, **Ville**, **Produit**, **Email** :
- Flèche ▾ (en-tête) → **Facet** → **Text facet**.
- Observer : casse (PARIS/Paris), variantes (plan c/Plan C), possibles doublons d’email.
- **But** : repérer avant d’agir.

OpenRefine

Activité
30 min

– Normaliser les Villes

- **Ville** → flèche ▾ → **Facet** → **Text facet**.
- Dans la liste à gauche :
- Cliquer **edit** en face des valeurs (ex. PARIS → Paris, Straßbourg → Strasbourg).
- Option : flèche ▾ **Edit cells** → **Cluster and edit…**
- Méthodes : *fingerprint*, *ngram-fingerprint* → **Merge selected & re-cluster**.
- **But** : homogénéiser rapidement.

– Corriger les Produits

- ▶**Produit → Facet → Text facet.**
- ▶Unifier : Edit cells → Common transforms → To lowercase
- ▶Corriger : **Produit → Edit cells → Transform…**
  - value.replace("plan","formule")

Préparer la colonne Nom

- **Edit column → Split into several columns…** (séparateur = espace, limit = 2) → Nom1 (prénom) / Nom2 (nom).
- **Nom1** puis **Nom2** → **Edit cells** → **Common transforms** → **To lowercase**.
- Option : **Cluster & edit** pour corriger trémas/accents proches (Müller/Muller).

Dates : conversion directe

- **Date_Commande → Edit cells → Common transforms → To date**.
- Vérifier via **Facet → Customized facets → Facet by type =** *date* uniquement.
- Option d'affichage : **Edit cells → Transform… →** value.toString("yyyy-MM-dd").
- **But** : dates ISO propres.

– Montants : 3 clics no-code

- **1. Supprimer espaces** : **Edit cells** → **Replace…** → **Replace**
- Find: \s (cocher *regex*), Replace: *(vide)* → **OK**
- **2. Virgules → points** : **Edit cells** → **Transform…** → **Replace**
- Find: ,, Replace: . → **OK**
- **3. Nombre** : **Common transforms** → **To number**

- Vérifier : **Facet by type** = *number*.

– Codes postaux & Emails

- **Code_Postal** :
  - Facet → customized facets →Text length facet → repérer les 4 chiffres
  - Filtrer sur “4”
  - Edit cells → Transform → value + 0
  - Vérifier longueur = 5
- **Email : Facet → Custom text Facet:**
- if(value.contains("@"),
  if(value.contains("."), "Valide", "Invalide"),
  "Invalide")
- **But** : préparer le dédoublonnage fiable par email.

Dédoublonnage par Email

- **Email** → **Facet** → **Duplicates facet** (si dispo) → cliquer **true**.
- **Sélection** : en-tête de la colonne de sélection (à gauche) → **Select all**.
- Barre au-dessus du tableau → **All** → **Remove duplicate rows**.
- **Effet** : garde la 1re occurrence de chaque email, supprime les copies exactes visibles.

– Contrôles finaux & Export

- **Facets rapides** :
- **Facet by type** : *date* sur Date_Commande, *number* sur Montant.
- **Text facet** : Ville, Produit → plus de variantes surprises.
- **Email** : aucun doublon restant.
- **Undo/Redo** : remonter si besoin, garder un instantané.
- **Export** : **Export** → **Comma-Separated Value (.csv)** → *dataset propre*.

Data Mining

- Science d'extraction de connaissances à partir de grands volumes de données par l'analyse automatique des patterns, tendances et relations cachées.
- En pratique :
  - ►Exploration : découverte de structures dans les données
  - ►Prédiction : identification des relations pour anticiper des comportements
  - ►Segmentation : regroupement d'observations similaires
  - ►Détection : identification d'anomalies et événements rares
- Finalité : Transformer les données brutes en insights actionnables pour la prise de décision.

Les outils qu'il vous faut

- Pour commencer :
  - Excel suffit jusqu’à 1 M lignes
- Pour collaborer :
  - Google Sheets permet le travail en équipe
- Pour industrialiser :
  - Orange Data Mining offre puissance sans code
- Pour scaler :
  - Dataiku / Databricks / Snowflake / DataRobot gèrent les grands volumes

Installer Orange Data Mining

- https://orangedatamining.com/download/
- Télécharger : https://www.kaggle.com/competitions/titanic/data?select=train.csv

Data Dictionary

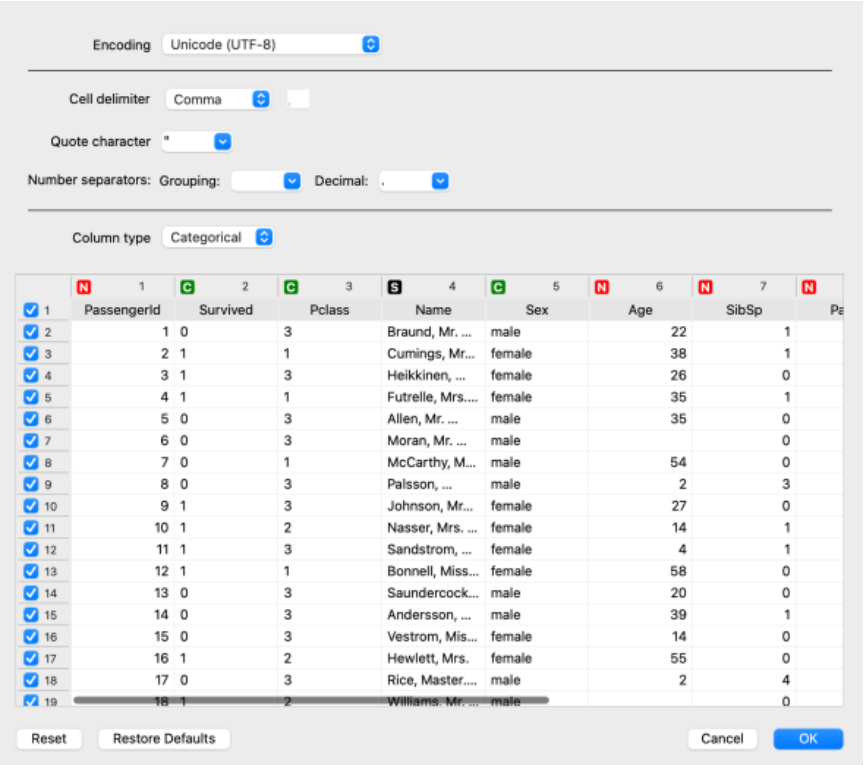
Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Étape 1 : Chargement des données

1. Ouvrir Orange Data Mining
2. Ajouter widget « CSV Import »
3. Depuis Titanic charger train.csv
4. Typer les données
    1. Quelles sont les discrètes ? => categorical
    2. Quelles sont les données numériques ? => numeric
    3. Quelles sont les données textuelles ? => text

– Étape 1 : Typer les données

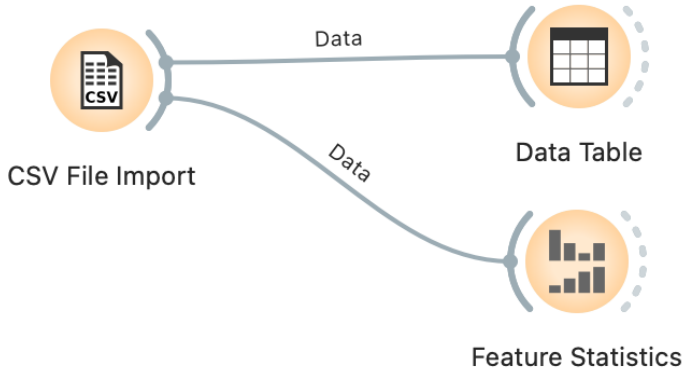
- **Categorical :** Survived, Sex, Embarked, Pclass
- **Numeric** : PassengerId, Age, SibSp, Parch , Fare,
- **Text** : Name, Cabin, Ticket



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1							
2	1	0	3	Braund, Mr. ...	male	22	1
3	2	1	1	Cumings, Mr...	female	38	1
4	3	1	3	Heikkinen, ...	female	26	0
5	4	1	1	Futrelle, Mrs....	female	35	1
6	5	0	3	Allen, Mr. ...	male	35	0
7	6	0	3	Moran, Mr. ...	male		0
8	7	0	1	McCarthy, M...	male	54	0
9	8	0	3	Palsson, ...	male	2	3
10	9	1	3	Johnson, Mr...	female	27	0
11	10	1	2	Nasser, Mrs. ...	female	14	1
12	11	1	3	Sandstrom, ...	female	4	1
13	12	1	1	Bonnell, Miss...	female	58	0
14	13	0	3	Saundercock...	male	20	0
15	14	0	3	Andersson, ...	male	39	1
16	15	0	3	Vestrom, Mis...	female	14	0
17	16	1	2	Hewlett, Mrs.	female	55	0
18	17	0	3	Rice, Master....	male	2	4
19	18	1	2	Williams, Mr....	male		0

Étape 2 : Visualiser les données

1. Connecter à « Data Table » et « Feature Statistics »
2. Observer les colonnes :
    1. Quelles sont les données incomplètes ?
    2. Quelle sont les données utiles ?
    3. Quelle est la cible (target) ?
    4. Astuce : mettre temporairement les données « text » en « categorical » pour compter les valeurs manquantes dans Feature Statistics (CSV File Import > Import Options)



– Étape 2 : Visualiser les données

Feature Statistics

	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
C	Cabin					4.9			687 (77 %)
N	Age		29.6991		28	0.48878	0.42	80	177 (20 %)
C	Embarked			S		0.76			2 (0 %)
N	PassengerId		446	1	446	0.576703	1	891	0 (0 %)
C	Survived			0		0.666			0 (0 %)

Color: None

Send Automatically

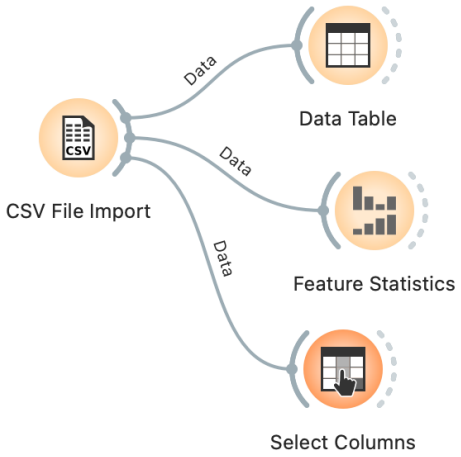
891 - | 12

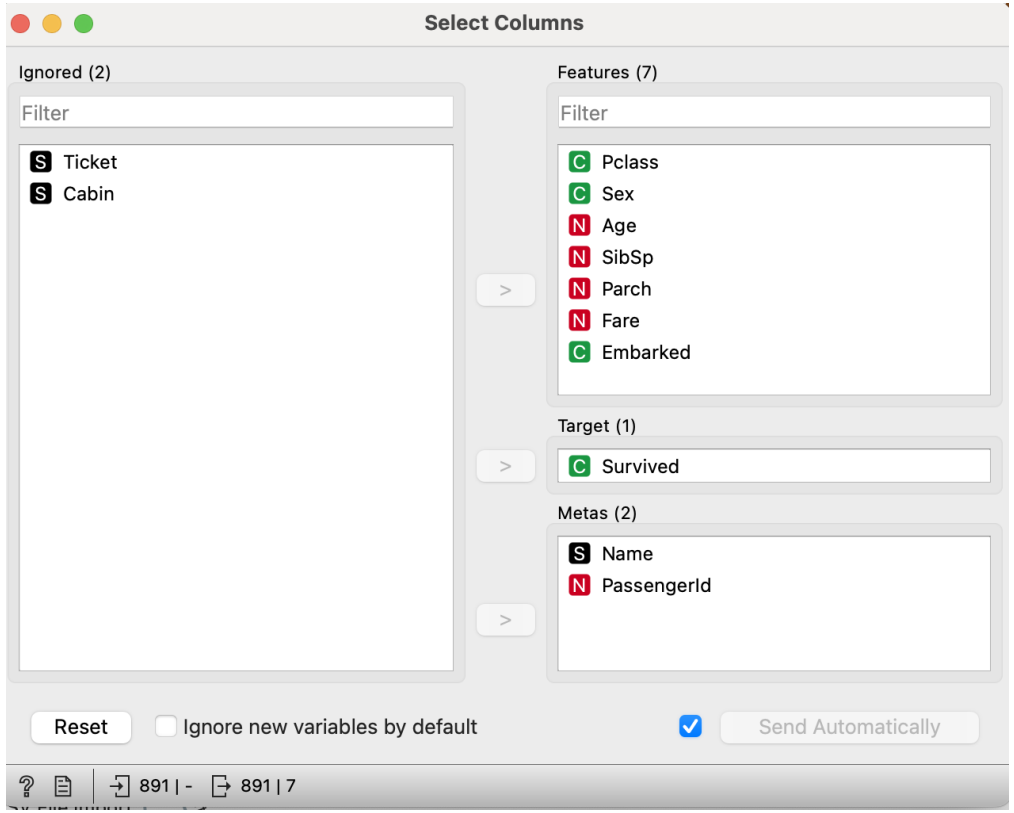
Étape 2 : Visualiser les données

Colonne	Ne pas garder ?	Pourquoi ?
ticket	✅ **Skip**	Trop de valeurs uniques, sans signification catégorielle. Exemple : `A/5  21171` n'a pas de structure exploitable sans traitement spécifique. Trop bruité pour la visualisation ou le ML.
cabin	✅ **Skip**	Données très **manquantes** (~80%) + grande **variabilité** (ex: C85, E46). Pas utilisable sans un gros travail de nettoyage/encodage.
name	❌ comme Feature, ✅ **Meta**	Trop **spécifique à chaque passager**, mais utile pour lecture humaine. Ne doit pas être utilisé dans les modèles ou visualisations.
passengerId	❌ comme Feature, ✅ **Meta**	Simple identifiant (1, 2, 3...) → aucune valeur explicative. Sert juste à retrouver un passager.

– Étape 2 : Définir la target et les features

1. Connecter à « Select Columns » à CSV File Import





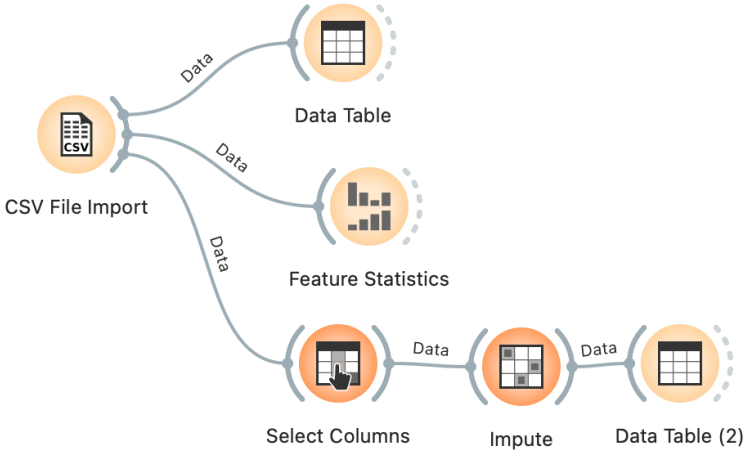
Étape 3 : Preprocess

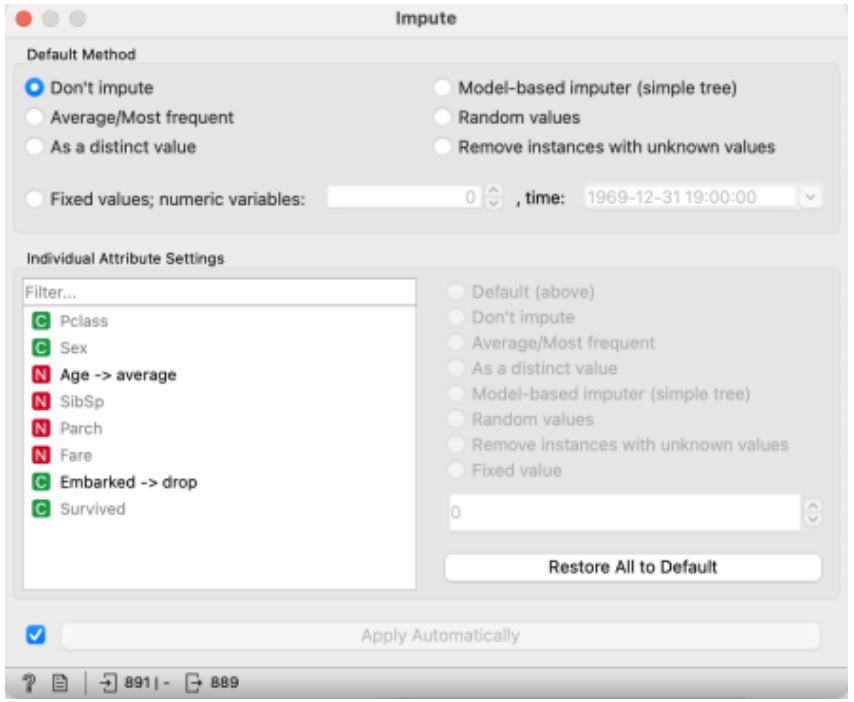
- Certaines colonnes (notamment Age) ont des **valeurs manquantes**, ce qui peut gêner :
  - Les visualisations (Box Plot, Scatter)
  - Les algorithmes de machine learning
- Connecter « Impute » à « Select Columns »
  - Age → moyenne ou médiane
  - Embarked → drop
- Connecter « Data Table » et vérifier
  - Quelle valeur a remplacé les valeurs « Age » manquantes ?
  - Combien de row par rapport au départ ?

Orange Data Mining

Activité
30 min

– Étape 3 : Preprocess





Conclusion et partage

- Discuter des découvertes principales
- Quels facteurs semblent les plus importants ?
- Quelles autres questions cela soulève ?

De la préparation à la visualisation

- Ce que nous avons vu
  - Des données propres
  - Des formats standardisés
  - Des contrôles qualité
- Ce qui nous attend
  - Donner du sens aux données
  - Créer des visualisations impactantes
  - Communiquer efficacement les insights

cegos
Beyond knowledge

Visualisation des données

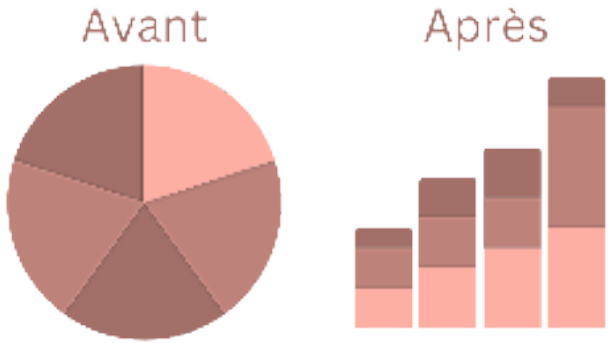
Pourquoi visualiser ?

- Les 3 objectifs :
  - Explorer → Comprendre les données
  - Expliquer → Communiquer les insights
  - Engager → Déclencher des actions
- Impact :
  - Gain temps analyse
  - Aide à la mémorisation
  - Augmentation adhésion décisions

Les erreurs classiques

Top 5 des pièges à éviter :

1. Camemberts 3D illisibles



2. Axes tronqués trompeurs
3. Couleurs sans logique
4. Trop d'informations
5. Pas de contexte

Choisir le bon graphique

- Règles simples :
  - Comparaison → Barres
  - Évolution → Ligne
  - Composition → Aires
  - Distribution → Histogramme
  - Relations → Scatter plot

Test rapide :
Si vous devez expliquer le graph plus de 30s, changez de format

Orange Data Mining

Activité

45 min

Étape 4 : Distribution

- Connecter « Distributions » à « Impute »
  - Combien de passagers ont survécu ? Est-ce une minorité ou une majorité ?
  - Y a-t-il plus d'hommes ou de femmes ?
  - Quelle classe est la plus représentée ?
  - Quelle est la tranche d'âge la plus représentée ? Y a-t-il des enfants ?
  - Est-ce que la majorité des passagers ont payé cher ou peu cher ?

Étape 4 : Distribution

- Connecter « Distributions » à « Impute »
- En activant “Cumulative Distribution” et “Split by: Survived” dans le widget Distributions :
  - La plupart des enfants ont-ils survécu ?
  - Les passagers ayant payé un tarif plus élevé avaient-ils plus de chances de survivre ?
  - Quelle classe a le plus contribué aux survivants ?
  - Quelle proportion des femmes a survécu comparée aux hommes ?

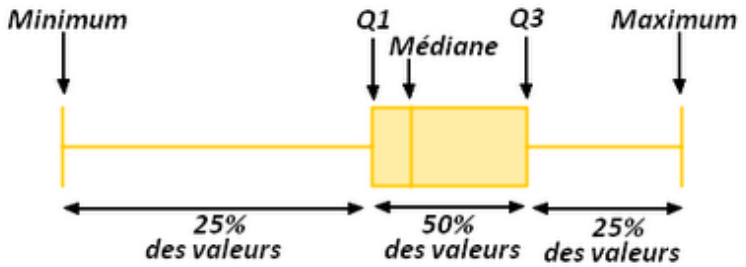
Orange Data Mining

Activité

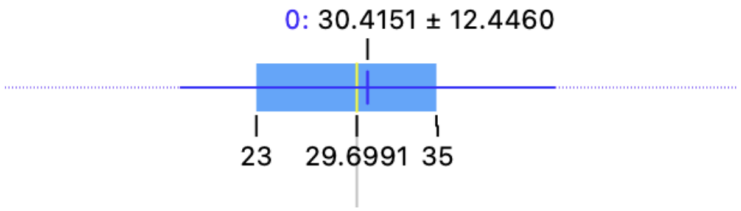
45 min

– Étape 5 : BoxPlot (Boîtes à moustaches)

- Connecter « BoxPlot » à « Impute »
  - Les survivants étaient-ils plus jeunes ?
  - Les passagers de 1ère classe ont-ils payé plus cher ?
  - Les survivants ont-ils payé un tarif plus élevé ?
  - Y a-t-il une différence d'âge selon le sexe ?



Moyenne + écart-type



Médiane

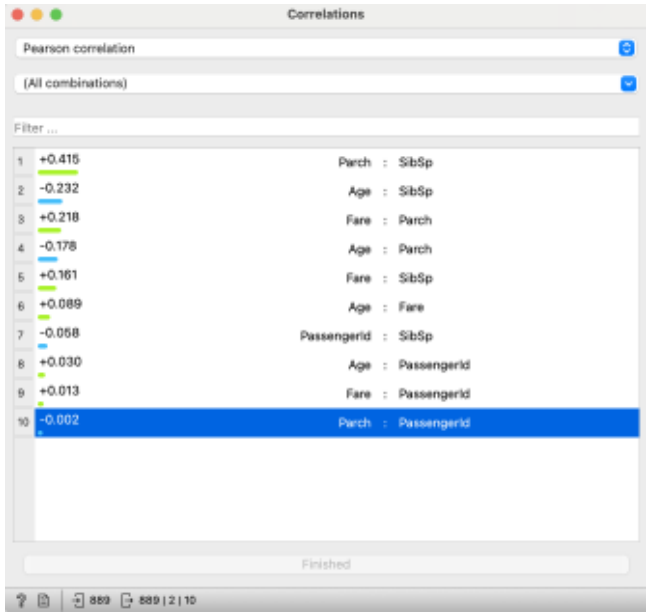
Orange Data Mining

Activité

45 min

– Étape 6 : Correlations

- Connecter « Correlations » à « Impute »



📖 Lecture des résultats

Corrélation (r)	Variables comparées	Interprétation pédagogique
+0.415	`Parch` / `SibSp`	📌 **Corrélation modérée** : voyager en famille implique frères/sœurs/parents
-0.232	`Age` / `SibSp`	🔍 Plus l'âge augmente, moins on voyage avec frères/sœurs
+0.218	`Fare` / `Parch`	🟢 Légère corrélation : familles paient souvent un peu plus
-0.178	`Age` / `Parch`	🔍 Adultes voyagent moins avec enfants
+0.161	`Fare` / `SibSp`	🟢 Familles avec frères/sœurs paient un peu plus
< ±0.10	`Age` / `Fare`, `Pclass`, etc.	Corrélations très faibles, peu interprétables

Étape 7 : Scatter Plot

- Connecter « Scatter Plot » à « Impute » :
  - Activer **Jittering**
  - Color : Survived
  - Shape : Survived

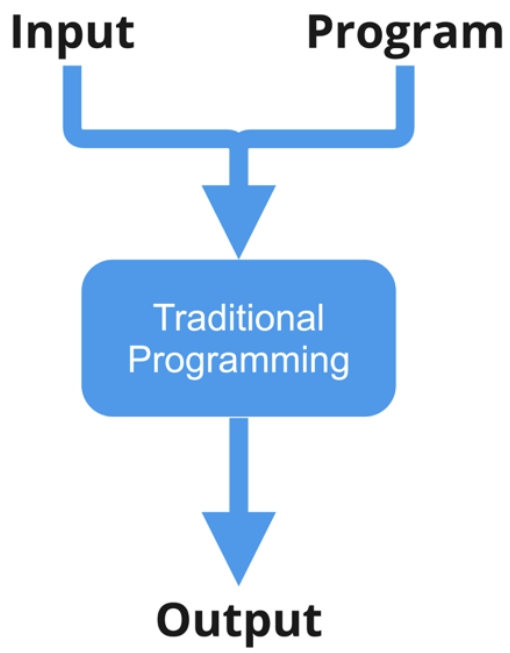
- Est-ce que voyager en famille améliore les chances de survie ? (SibSp vs Parch)
- Est-ce que le prix du billet reflète la classe et la survie ? (Fare vs Pclass)

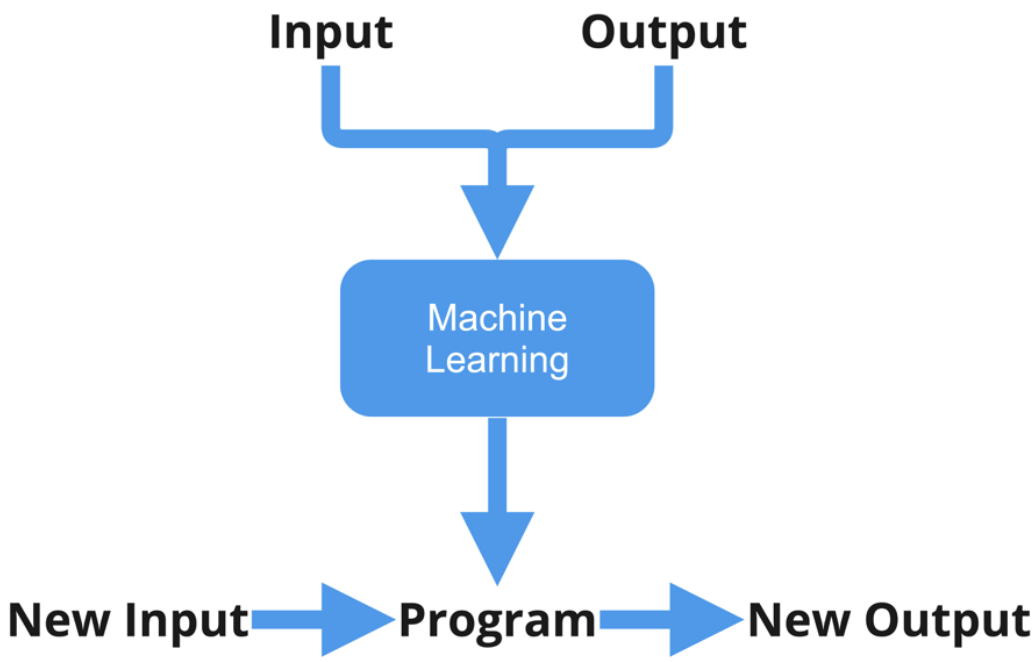
cegos
Beyond knowledge

Le Machine Learning

Machine Learning (Rappel)

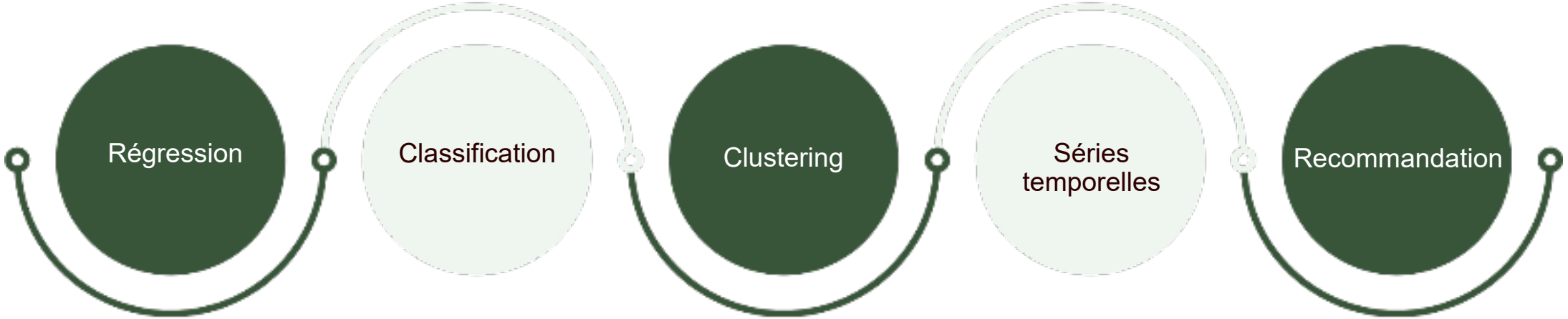
General programming vs. Machine Learning





Les algorithmes essentiels

- Top 5 des algorithmes business :



- Attribue une catégorie
- Exemple : Segmentation client
- Métrique : % précision

- Analyse de tendances
- Exemple : Prévision stock
- Métrique : Erreur prévision

- Prédit une valeur
- Exemple : Prévision CA mensuel
- Métrique : Erreur moyenne (RMSE)

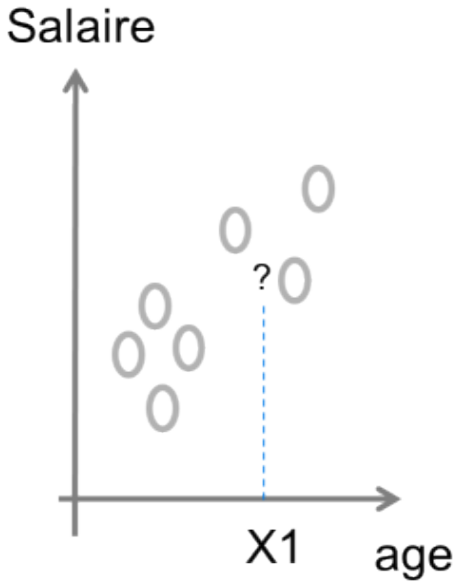
- Regroupe les similaires
- Exemple : Comportements d'achat
- Métrique : Cohésion groupes

- Suggère des options
- Exemple : Produits associés
- Métrique : Taux de clic

Méthodes supervisées / non supervisées

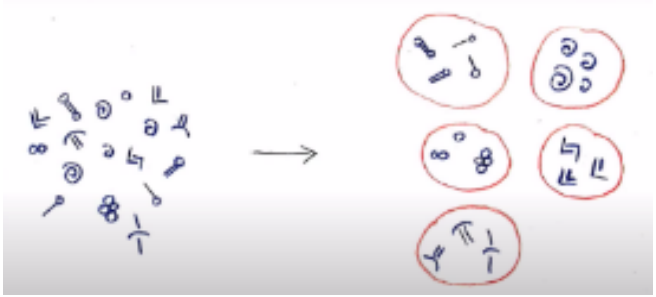
- Exemple

Méthodes supervisées



L'apprentissage supervisé utilise des données étiquetées (avec réponses connues) pour entraîner des modèles qui prédisent des résultats.

Méthodes non supervisées



L'apprentissage non supervisé travaille avec des données non étiquetées pour découvrir des structures ou des patterns intrinsèques.

Deux grandes familles de méthodes

Méthodes supervisées

Classification

- Reconnaitre des chiffres, images, classification binaire

Régression

- Prédire : Régression y = f(x), ventes, churn

Méthodes non supervisées

Clustering

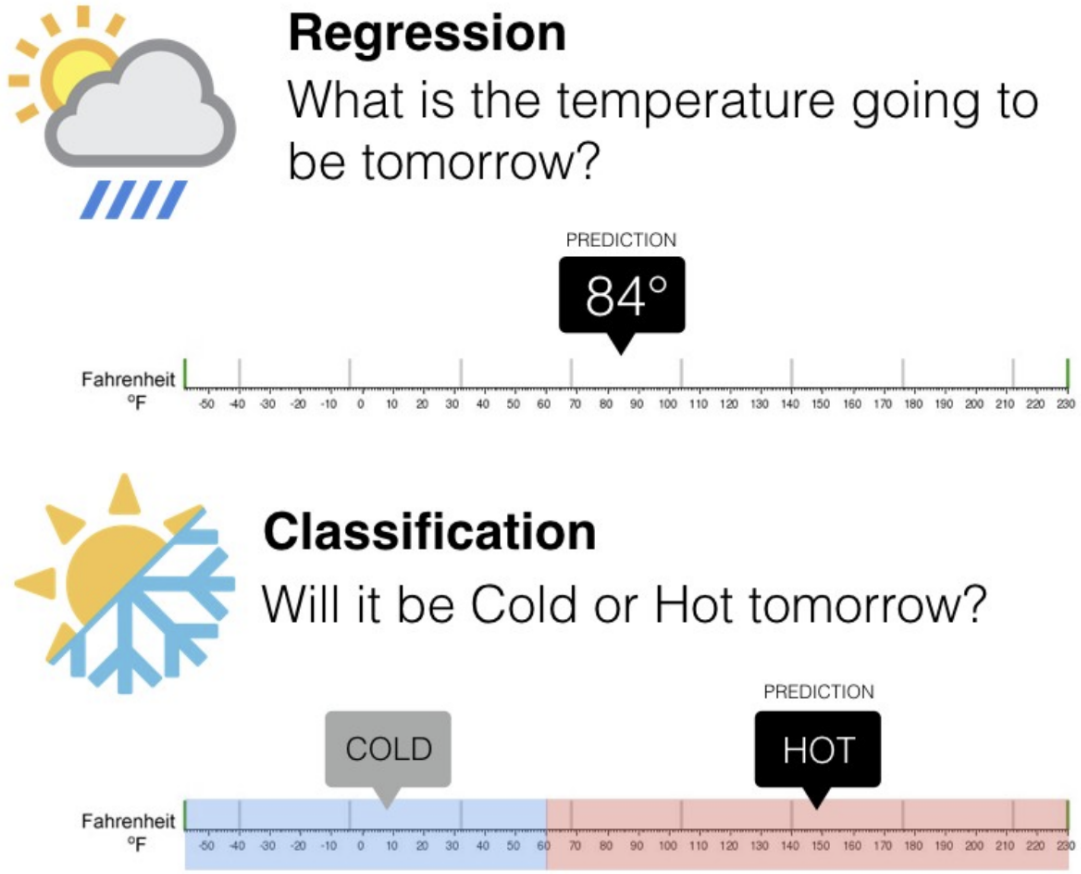
- Identification segment de marché
- Découverte de pattern

Réduction de dimension

Détection des anomalies

Machine Learning

Classification vs Regression



Machine Learning : supervisé

Apprentissage supervisé (Supervised Learning)

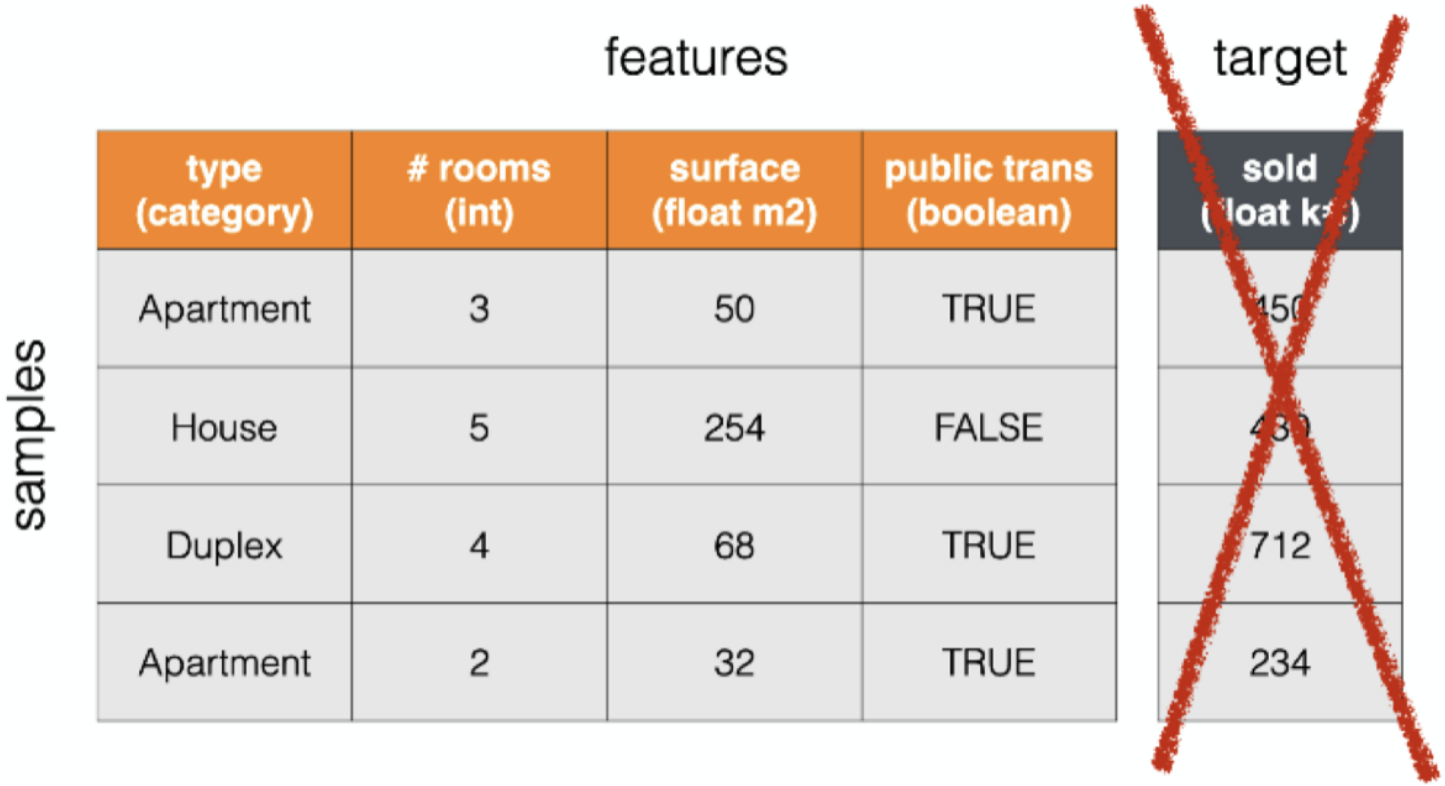
Développer un modèle prédictif basé à la fois sur les données d'entrée et de sortie.

	features			target
samples	type (category)	# rooms (int)	surface (float m2)	public trans (boolean)	sold (float k€)
	Apartment	3	50	TRUE	450
	House	5	254	FALSE	430
	Duplex	4	68	TRUE	712
	Apartment	2	32	TRUE	234

Machine Learning : non supervisé

Apprentissage non supervisé (Unsupervised Learning)

But : Regrouper et interpréter les données uniquement à partir des données d'entrée (*input data*).



Machine Learning: Jargon

Les **features** (caractéristiques) peuvent aussi être appelées **input**, **X**, **variables**.

La **target** (cible) peut aussi être appelée **output**, **y**, **label**, **classe**.

Les **samples** (échantillons) peuvent aussi être appelés **rows** (lignes) ou **observations**.

	features			target
samples	type (category)	# rooms (int)	surface (float m2)	public trans (boolean)	sold (float k€)
	Apartment	3	50	TRUE	450
	House	5	254	FALSE	430
	Duplex	4	68	TRUE	712
	Apartment	2	32	TRUE	234

Définitions

- En équipe proposez une définition courte des algorithmes de Machine Learning suivants :
  - Régression linéaire
  - Régression logistique
  - Arbre de décision
  - Naïve Bayes
  - Réseaux de neurones
  - K Nearest Neighbors (KNN)
  - K Means – Nuées dynamiques
- Vous pouvez utiliser les sources que vous souhaitez

Les algorithmes du Machine Learning les plus utilisés

- Régression linéaire :
  - Modèle qui établit une relation linéaire entre variables indépendantes et une variable dépendante continue. Prédit une valeur numérique en trouvant la meilleure droite (ou hyperplan) minimisant l'erreur entre prédictions et valeurs réelles.
- Régression logistique :
  - Modèle de classification qui utilise une fonction logistique pour estimer la probabilité qu'une instance appartienne à une classe. Idéale pour les problèmes binaires, mais extensible au multiclasse.
- Arbres de décision :
  - Structure hiérarchique qui divise les données selon des règles de décision (si-alors) basées sur les caractéristiques. Facile à interpréter visuellement, modélise des relations non linéaires et gère naturellement différents types de données.
- Naïve Bayes :
  - Algorithme probabiliste basé sur le théorème de Bayes qui suppose l'indépendance entre les caractéristiques. Rapide, efficace pour les grands ensembles de données et particulièrement adapté au traitement de texte et à la classification d'emails.

REF-Gr_00055019,3.0,CURRENT, Bon pour publi

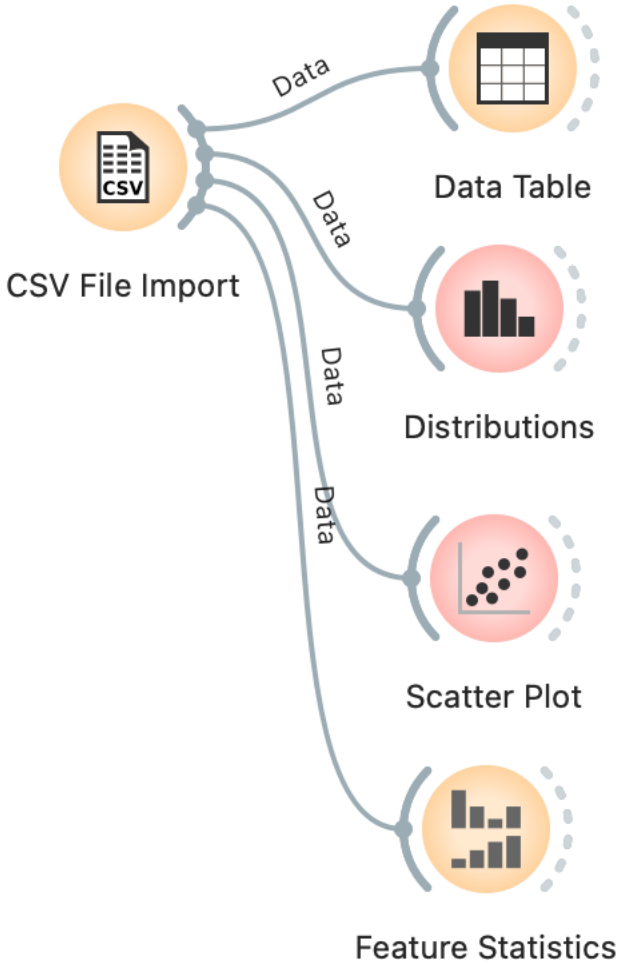
Les algorithmes du Machine Learning les plus utilisés

- Réseaux de neurones :
  - Modèles inspirés du cerveau humain composés de couches de neurones interconnectés. Capables d'apprendre des représentations complexes à partir des données pour diverses tâches comme la classification d'images ou le traitement du langage naturel.
- K Nearest Neighbors (KNN) :
  - Classifie un point en fonction des classes des K points les plus proches. Mémorise les exemples d'entraînement et prédit selon le "vote majoritaire" du voisinage. Simple mais puissant pour les relations complexes avec suffisamment de données.
- K Means – Nuées dynamiques :
  - Algorithme de clustering qui partitionne les données en K groupes en minimisant la distance entre les points et le centre de leur cluster. Rapide et efficace pour de grands ensembles de données, mais nécessite de spécifier le nombre de clusters à l'avance.

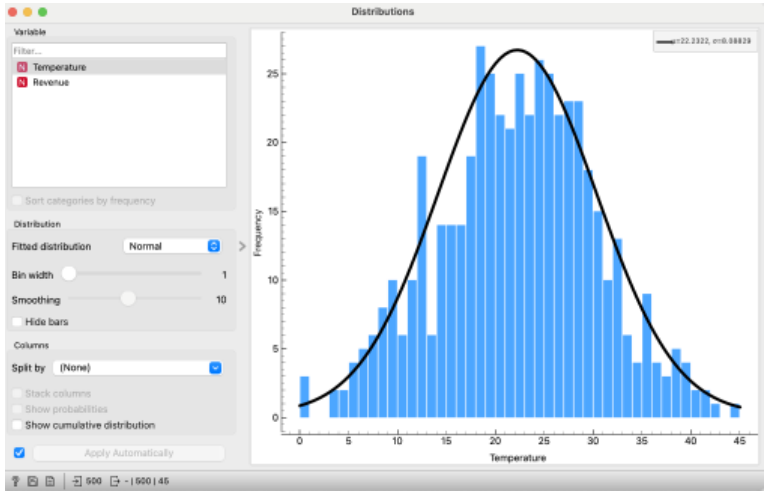
Les algorithmes du Machine Learning les plus utilisés

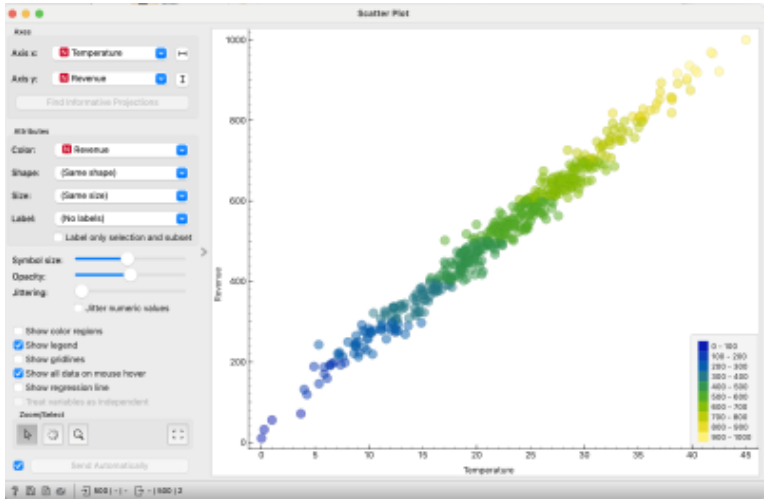
- Régression linéaire : https://youtu.be/PaFPbb66DxQ
- Régression logistique : https://youtu.be/yIYKR4sgzI8
- Arbres de décision : https://youtu.be/_L39rN6gz7Y
- Naïve Bayes : https://youtu.be/O2L2Uv9pdDA
- Réseaux de neurones : https://youtu.be/CqOfi41LfDw
- K Nearest Neighbors (KNN) : https://youtu.be/HVXime0nQeI
- K Means – Nuées dynamiques : https://youtu.be/4b5d3muPQmA

Étape 1 : Visualiser les données IceCreamData.csv

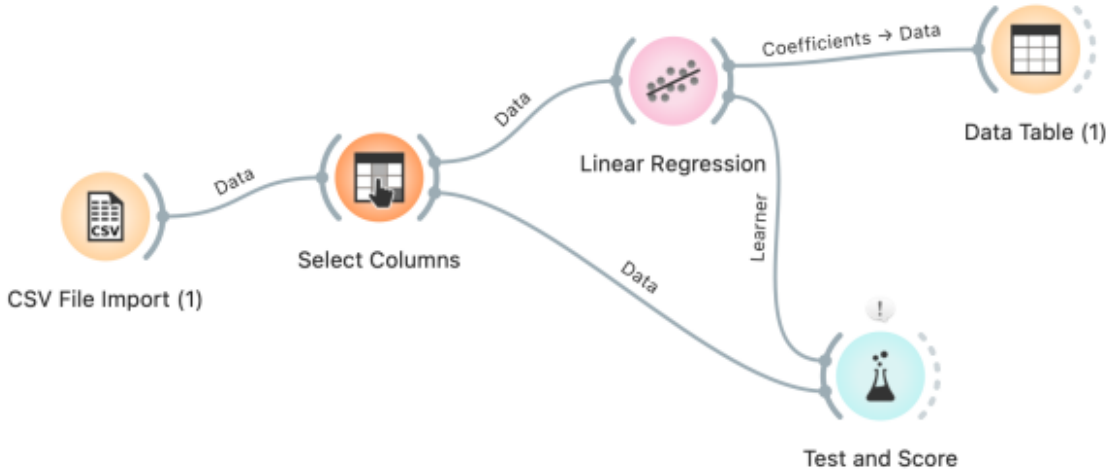


– Étape 1 : Visualiser les données IceCreamData.csv

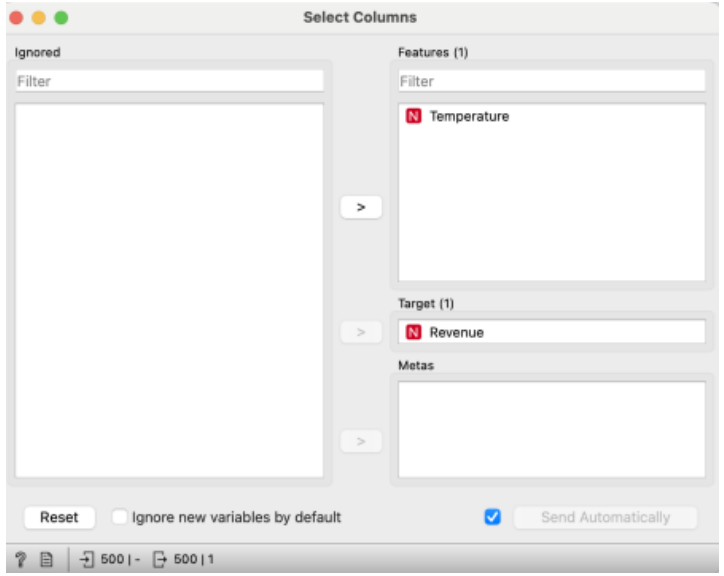




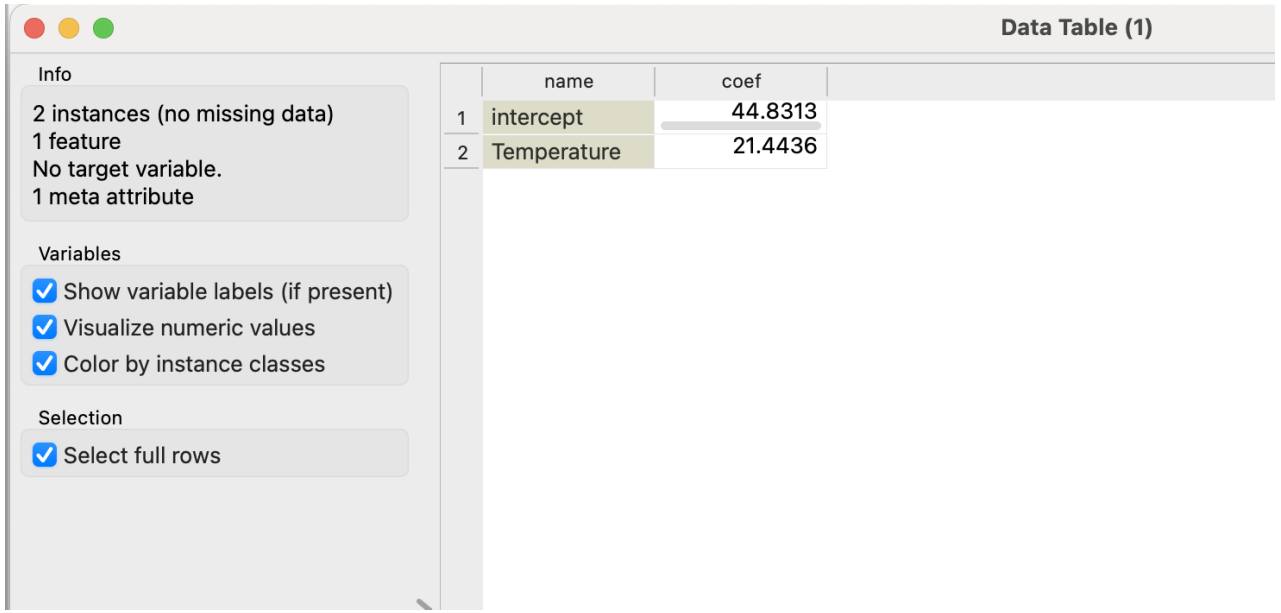
– Étape 2 : Régression Linéaire



f(x) = ax + b
Trouver le coefficient directeur (a) et l'ordonnée (b)



– Étape 2 : Régression Linéaire



Évaluer un modèle

- Tout comme un élève passe des examens pour évaluer ses connaissances, un modèle de ML doit être évalué avec des métriques précises
- Ces métriques nous permettent de :
  - Mesurer objectivement la qualité des prédictions
  - Comparer différents modèles entre eux
  - Détecter les problèmes (biais, sur apprentissage)
  - Suivre l'évolution des performances dans le temps

Métriques pour la régression

Métriques pour la régression

Régression linéaire

MSE (Mean Squared Error - Erreur quadratique moyenne)
Définition : Moyenne des carrés des écarts entre prédictions et valeurs réelles
Valeur cible : Aussi proche de 0 que possible (pas de valeur absolue standard car dépend de l'échelle des données)
Interprétation : Plus la valeur est faible, meilleur est le modèle

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Métriques pour la régression

Régression linéaire

RMSE (Root Mean Squared Error - Racine de l'erreur quadratique moyenne)
Définition : Racine carrée du MSE, dans la même unité que la variable cible
Valeur cible : Aussi proche de 0 que possible
Interprétation : Représente l'écart-type des résidus du modèle

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

Métriques pour la régression

Régression linéaire

MAE (Mean Absolute Error - Erreur absolue moyenne)
Définition : Moyenne des valeurs absolues des écarts
Valeur cible : Aussi proche de 0 que possible
Interprétation : Moins sensible aux valeurs aberrantes que le MSE/RMSE

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Métriques pour la régression

Régression linéaire

R^2 (Coefficient de détermination)
Définition : Proportion de la variance expliquée par le modèle
Valeur cible : Entre 0 et 1, avec > 0.7 généralement considéré comme bon
Interprétation : 0.8 signifie que 80% de la variance est expliquée par le modèle

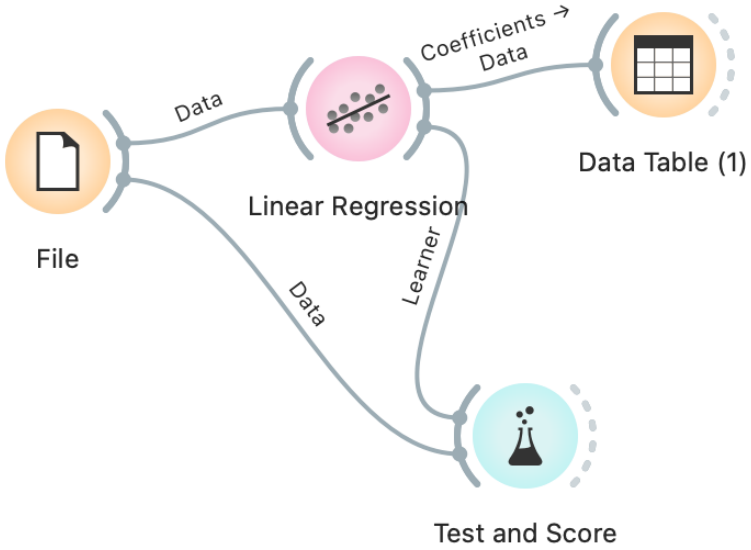
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

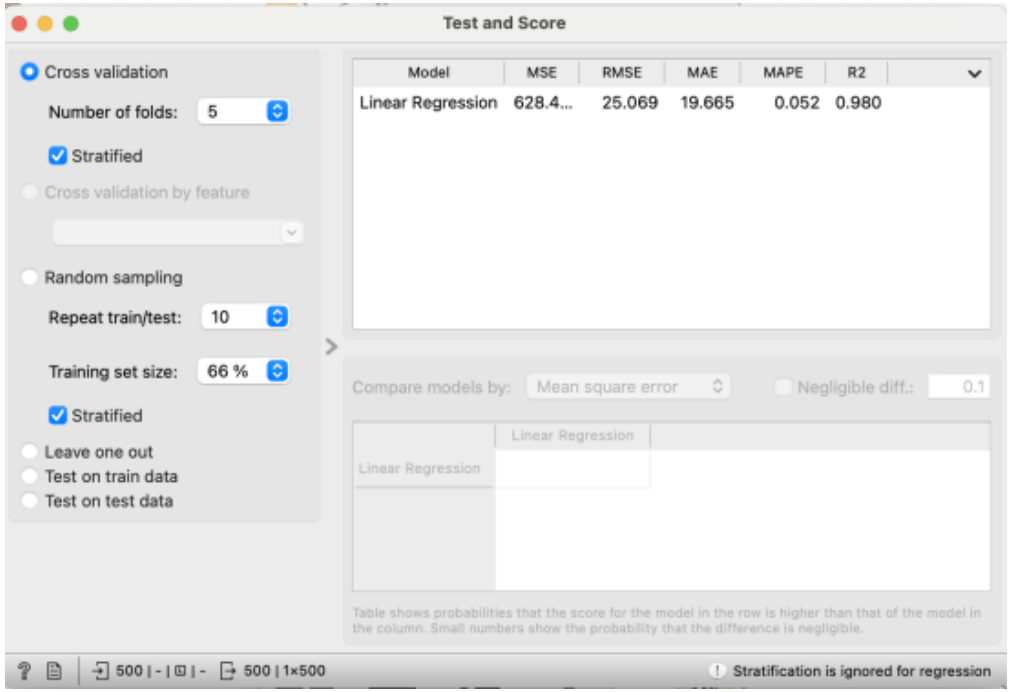
Orange Data Mining

Activité

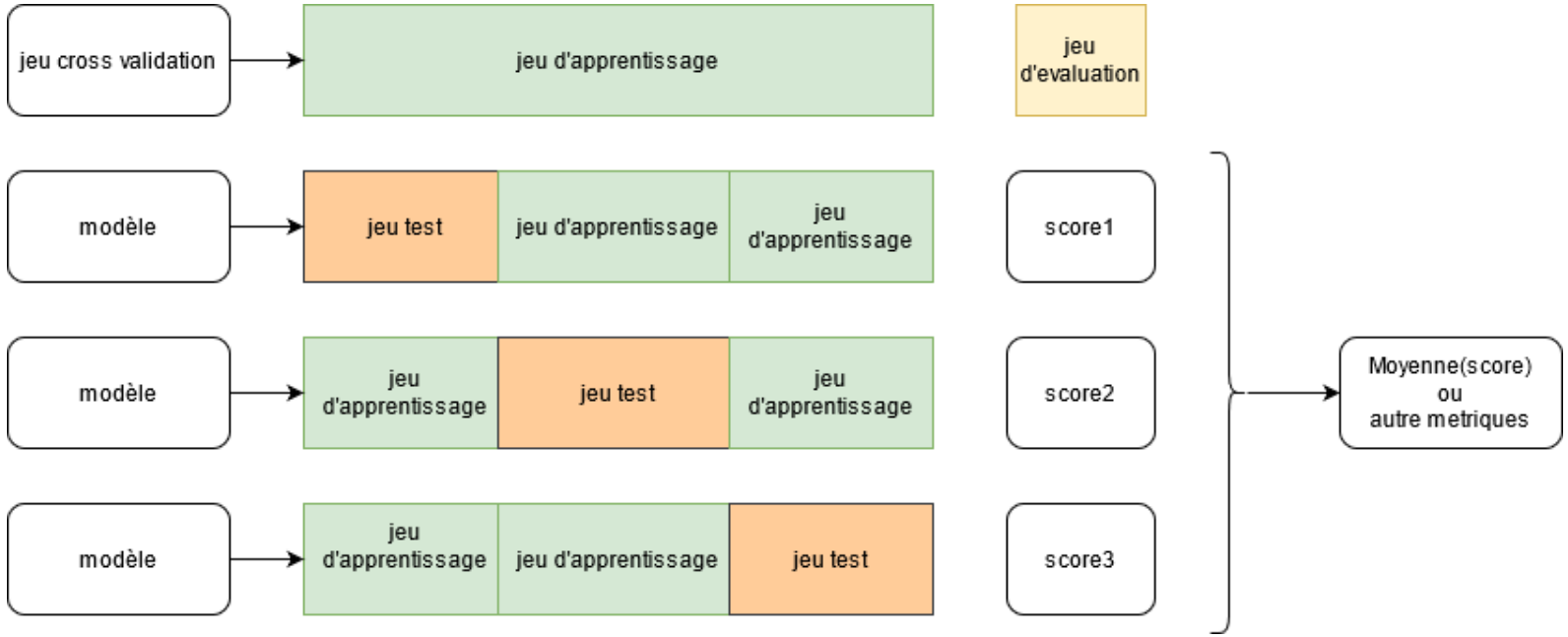
30 min

– Étape 3 : Test





Étape 3 : Test



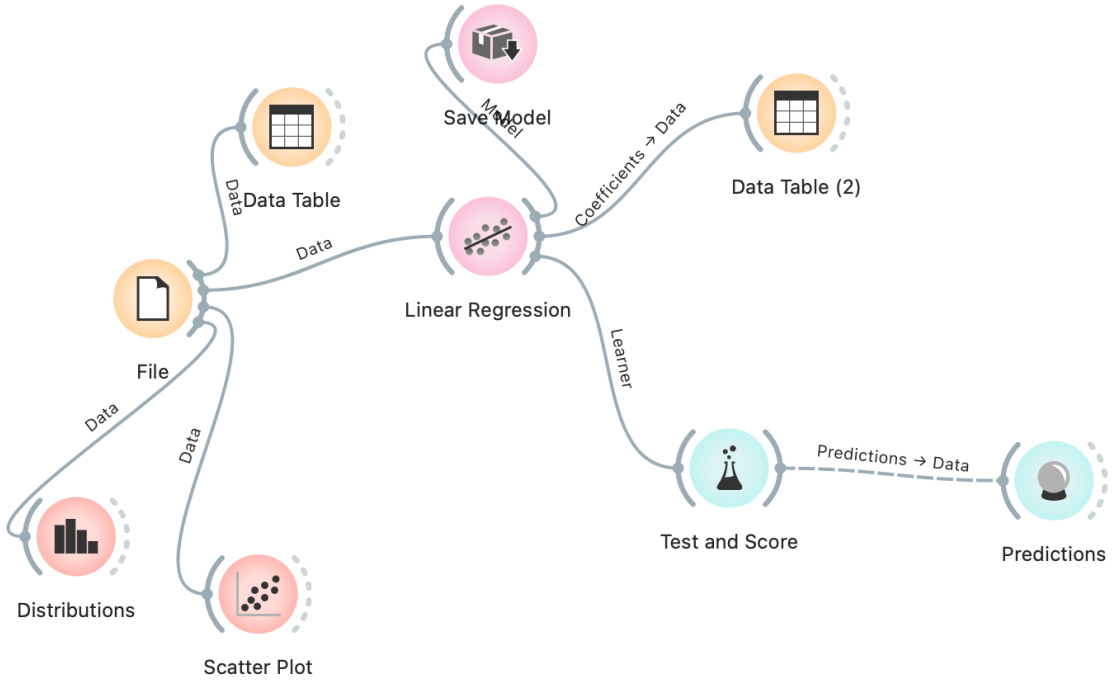
Source : https://collonvillethomas.wordpress.com

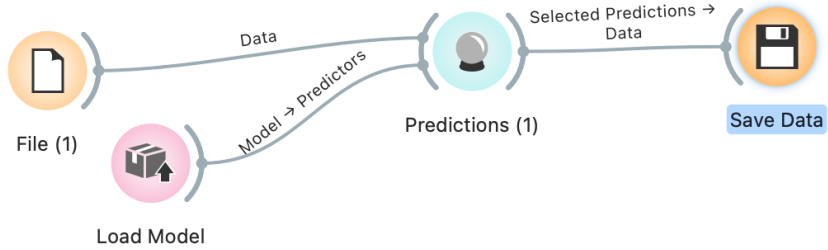
Orange Data Mining

Activité

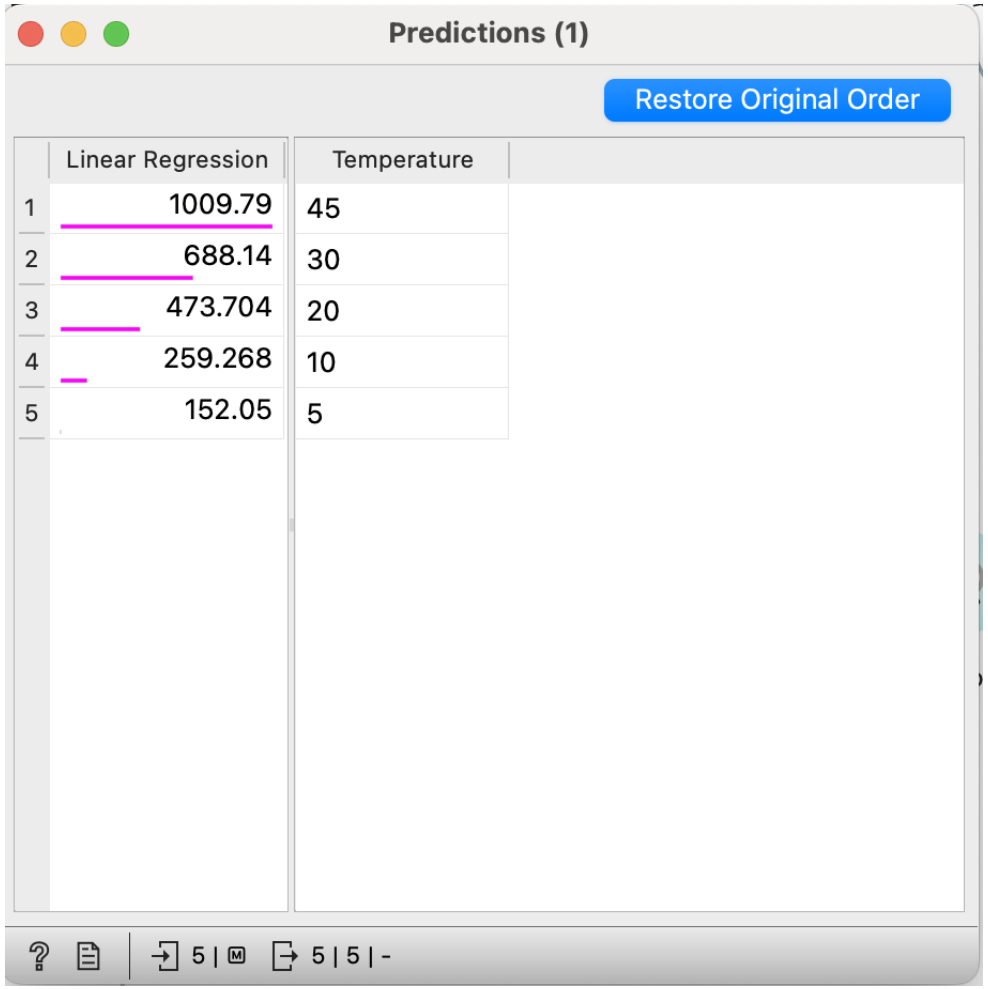
30 min

– Étape 4 : Prédiction





– Étape 4 : Prédiction

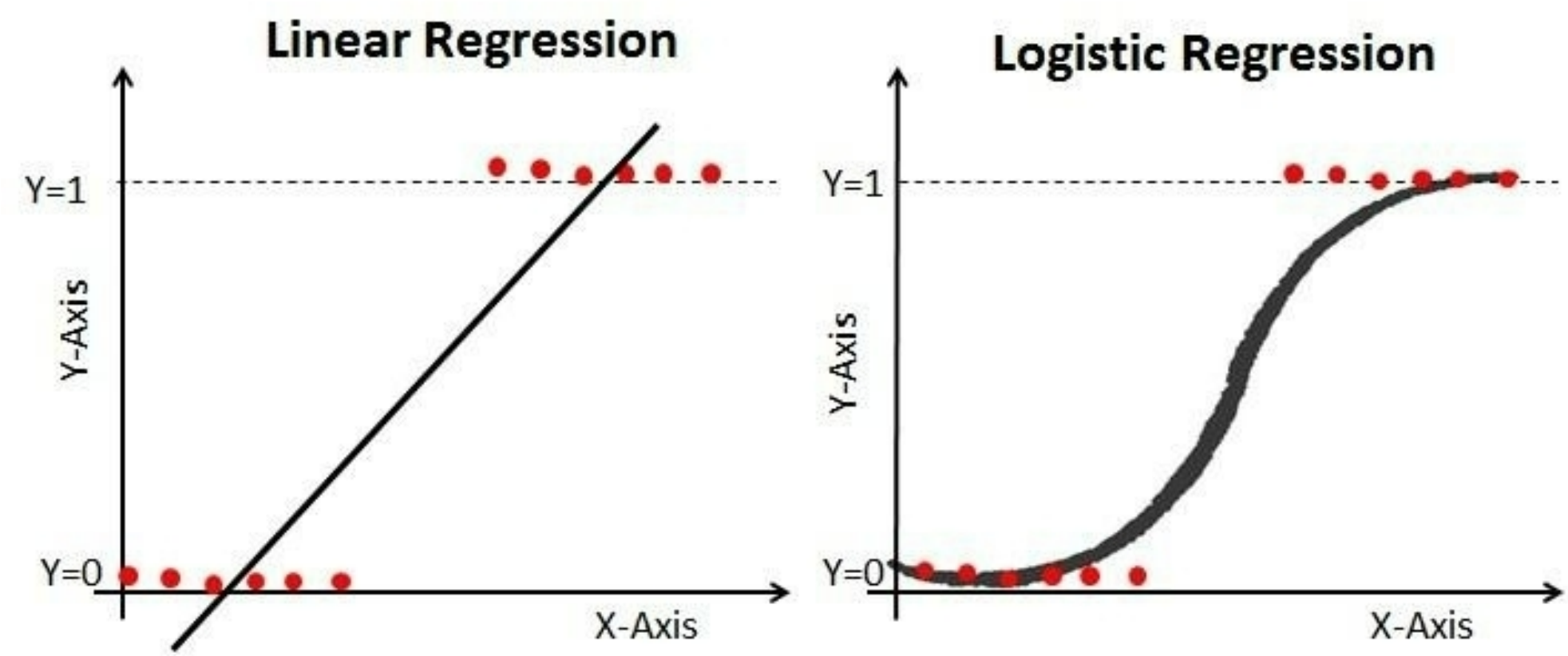


Régression Linéaire vs Régression Logistique

La régression linéaire est utilisée pour prédire une valeur numérique continue. Elle établit une relation entre des variables explicatives (features) et une cible (target), en ajustant une droite (ou un plan en multivarié) qui minimise l'erreur entre prédictions et valeurs réelles.

La régression logistique poursuit un objectif différent : résoudre des problèmes de classification, en prédisant une probabilité d'appartenance à une classe. Elle transforme la combinaison linéaire des variables par une fonction sigmoïde qui ramène le résultat entre 0 et 1.

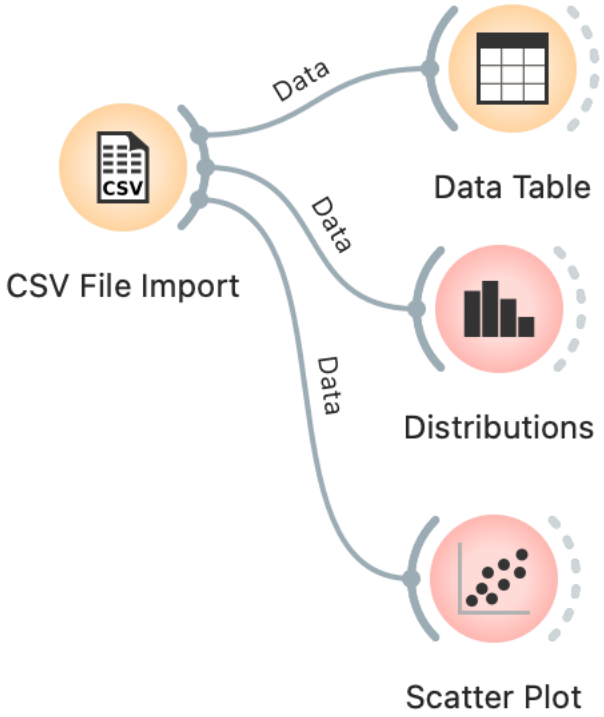
Régression Linéaire vs Régression Logistique



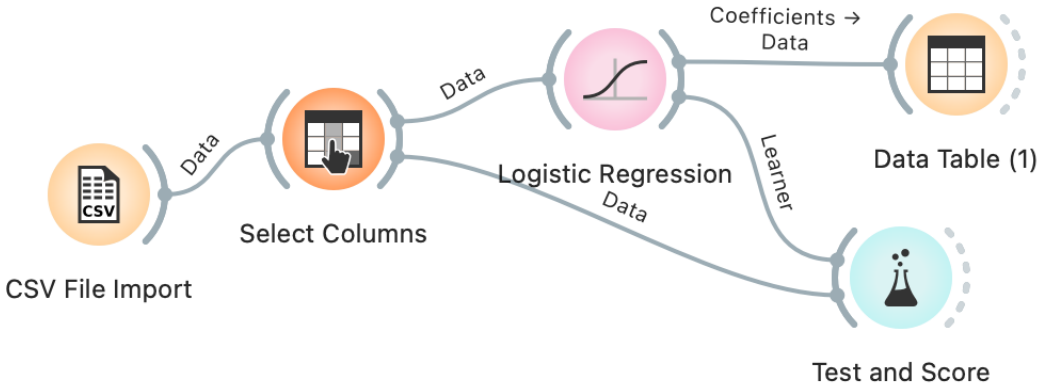
Régression Logistique

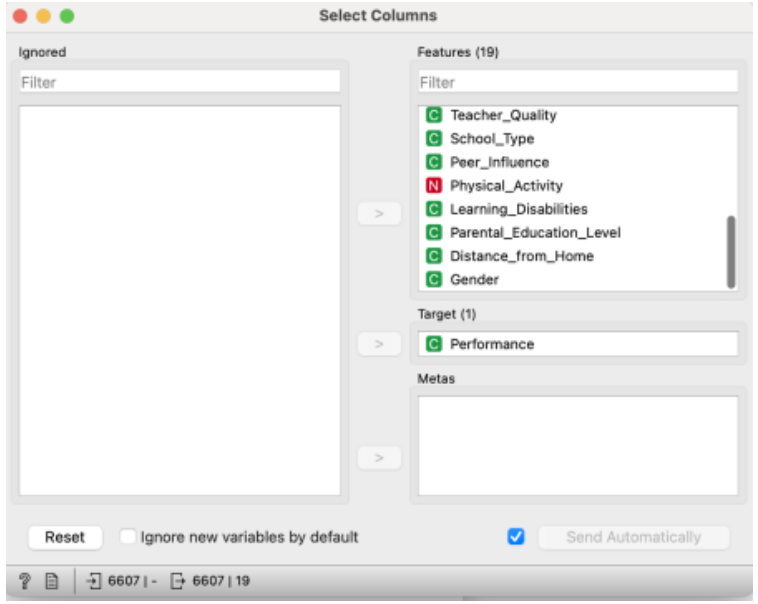
- **Télécharger et analyser**: student_performance.csv

Étape 1 : Visualiser les données



Étape 2 : Définir les colonnes



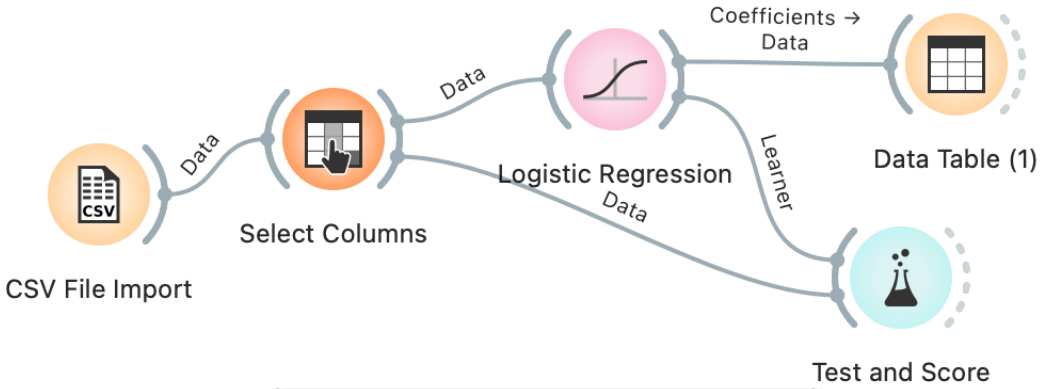


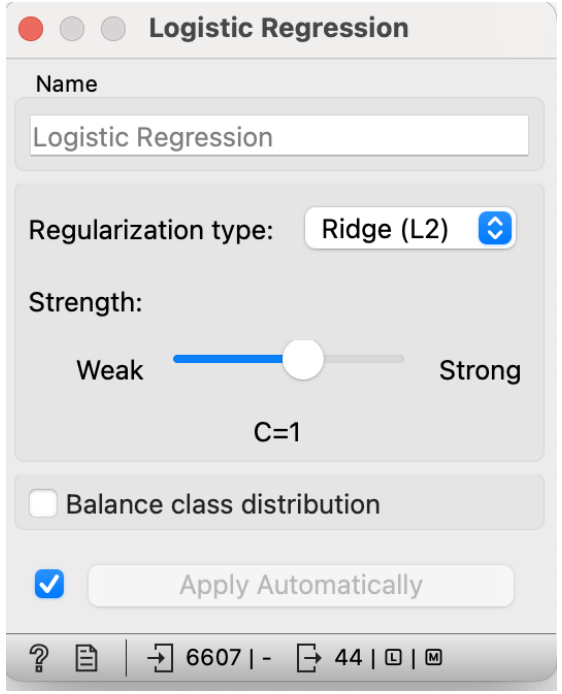
Orange Data Mining

Activité

30 min

– Étape 2 : Régression Logistique





– Étape 2 : Régression Logistique

Data Table (1)

Info
44 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes

Selection
Select full rows

Restore Original Order

	name	Pass
4	Parental_Involvement=High	1.77197
7	Access_to_Resources=High	1.6846
33	Learning_Disabilities=No	1.29423
23	Teacher_Quality=High	1.29074
37	Parental_Education_Level=Postgraduate	1.24587
14	Motivation_Level=High	1.15969
31	Peer_Influence=Positive	1.04742
41	Distance_from_Home=Near	1.0159
19	Tutoring_Sessions	0.98459
20	Family_Income=High	0.923685
18	Internet_Access=Yes	0.918004
2	Hours_Studied	0.695794
11	Extracurricular_Activities=Yes	0.564466
32	Physical_Activity	0.558218
3	Attendance	0.487374
40	Distance_from_Home=Moderate	0.425584
6	Parental_Involvement=Medium	0.386547
16	Motivation_Level=Medium	0.319
9	Access_to_Resources=Medium	0.268125
42	Distance_from_Home=Unknown	0.206291
13	Previous_Scores	0.115379
44	Gender=Male	0.07443
22	Family_Income=Medium	0.0700737
28	School_Type=Public	0.0669341
12	Sleep_Hours	0.0522684
27	School_Type=Private	0.051215
43	Gender=Female	0.0437191
35	Parental_Education_Level=College	-0.0151117
38	Parental_Education_Level=Unknown	-0.0305046
25	Teacher_Quality=Medium	-0.0523784
30	Peer_Influence=Neutral	-0.145237
26	Teacher_Quality=Unknown	-0.20893
10	Extracurricular_Activities=No	-0.446317
29	Peer_Influence=Negative	-0.784033
17	Internet_Access=No	-0.799855
21	Family_Income=Low	-0.87561
24	Teacher_Quality=Low	-0.911286
36	Parental_Education_Level=High School	-1.0821
34	Learning_Disabilities=Yes	-1.17608
15	Motivation_Level=Low	-1.36054
39	Distance_from_Home=Far	-1.52962
8	Access_to_Resources=Low	-1.83458
5	Parental_Involvement=Low	-2.04036
1	intercept	-50.1008

Accuracy

- La précision globale mesure le pourcentage de prédictions correctes parmi toutes les prédictions.
- Formule : (Vrais Positifs + Vrais Négatifs) / Nombre total de prédictions
- Exemple concret : un modèle de détection de fraude qui identifie correctement 95 transactions sur 100 a une accuracy de 95 %
- Attention : l'accuracy peut être trompeuse avec des données déséquilibrées

Precision

- La précision mesure la proportion de vrais positifs parmi tous les cas prédits comme positifs
- Formule : Vrais Positifs / (Vrais Positifs + Faux Positifs)
- Exemple concret : Si un modèle identifie 10 emails comme spam mais que seulement 8 sont réellement du spam, sa précision est de 8/10 = 80 %
- Cas d'usage : Crucial quand les faux positifs sont coûteux (ex: blocage de transactions légitimes)

Recall

- Le recall mesure la proportion de vrais positifs correctement identifiés parmi tous les cas réellement positifs.
- Formule : Vrais Positifs / (Vrais Positifs + Faux Négatifs)
- Exemple concret : Si parmi 100 fraudes réelles, un modèle en détecte 90, son recall est de 90 %
- Cas d'usage : Important quand manquer un cas positif est grave (ex. : détection de maladies)

F1 Score

- Le F1 Score est la moyenne harmonique entre la precision et le recall, offrant un équilibre entre les deux métriques.
- Formule : 2 × (Precision × Recall) / (Precision + Recall)
- Exemple concret : Un modèle avec une precision de 80 % et un recall de 70 % aura un F1 Score de 74.7 %.
- Utilisation : Particulièrement utile quand on cherche un compromis entre precision et recall.

Métriques pour la classification

Régression logistique

Accuracy (Précision globale)
Définition : Proportion de prédictions correctes
Valeur cible : > 0.7 généralement, mais dépend fortement du contexte et de l'équilibre des classes
Interprétation : Problématique pour les datasets déséquilibrés

Precision (Précision)
Définition : Proportion des prédictions positives qui sont correctes
Valeur cible : > 0.8 pour la plupart des applications
Interprétation : Crucial quand les faux positifs sont coûteux

Recall (Rappel, Sensibilité)
Définition : Proportion des cas positifs réels correctement identifiés
Valeur cible : > 0.8 idéalement
Interprétation : Crucial quand les faux négatifs sont coûteux (ex: détection de maladies)

F1-Score
Définition : Moyenne harmonique de Precision et Recall
Valeur cible : > 0.75 généralement considéré comme bon
Interprétation : Bon équilibre entre precision et recall, utile pour les datasets déséquilibrés

AUC-ROC (Area Under Receiver Operating Characteristic Curve)
Définition : Mesure la capacité du modèle à distinguer les classes
Valeur cible :
0.5 = aléatoire (pas mieux qu'un tirage à pile ou face)
0.7-0.8 = acceptable
0.8-0.9 = excellent
0.9 = exceptionnel
Interprétation : Représente la probabilité que le modèle classe correctement un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire

MCC (Matthews Correlation Coefficient)
Définition : Coefficient de corrélation entre les classifications observées et prédites
Valeur cible : Entre -1 et 1, où 1 représente une prédiction parfaite
Interprétation : Considéré comme une mesure équilibrée même avec des classes déséquilibrées

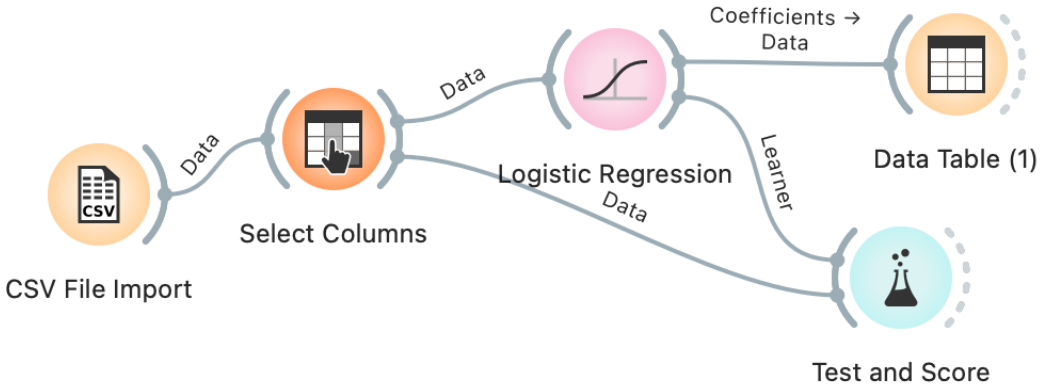
REF-Gr_00055019,3.0,CURRENT, Bon pour publi

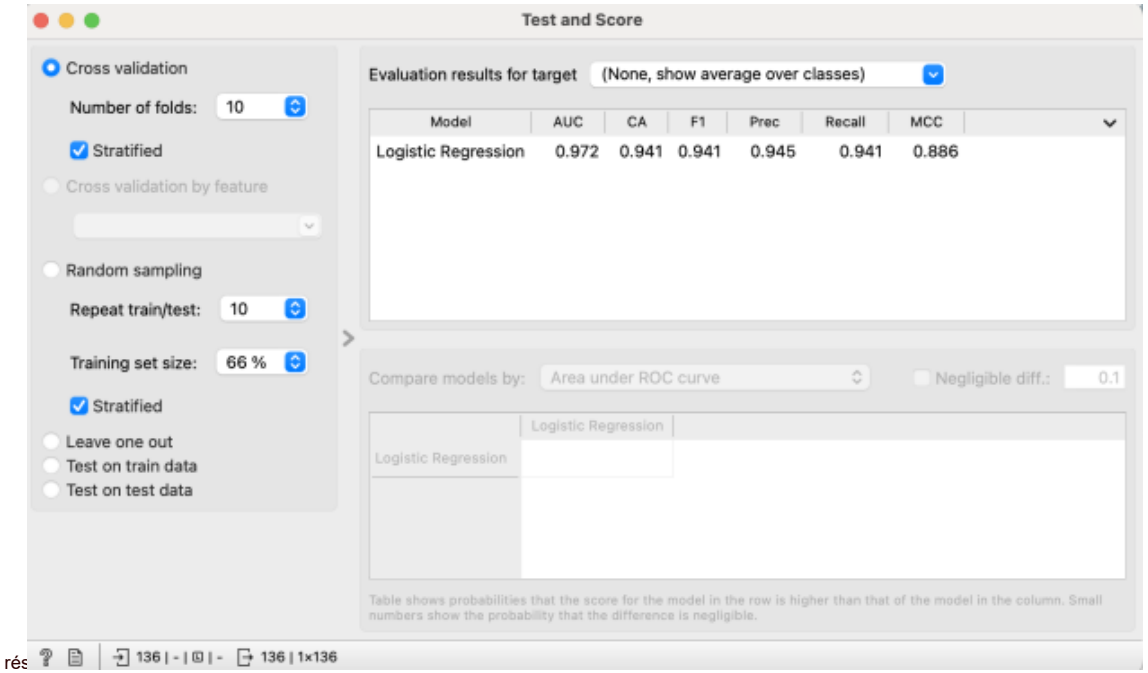
Orange Data Mining

Activité

30 min

– Étape 4 : Test and score

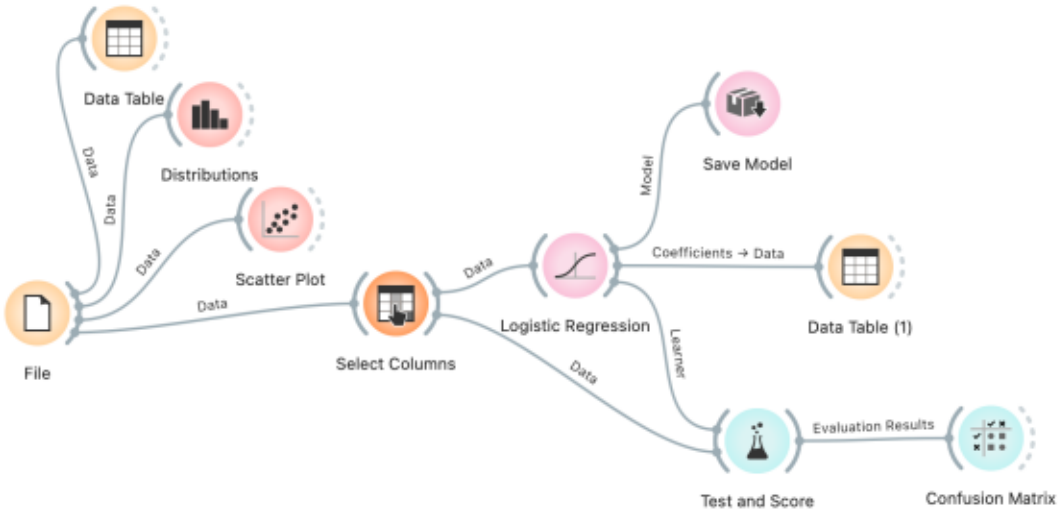


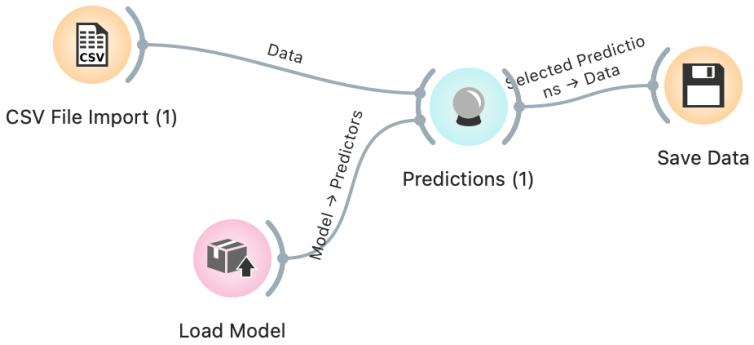


Orange Data Mining

Activité
30 min

– Étape 5 : Prédiction





Orange Data Mining

Activité

30 min

– Étape 5: Prédiction

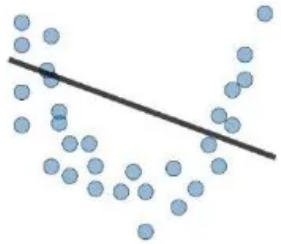
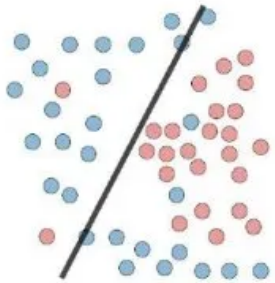
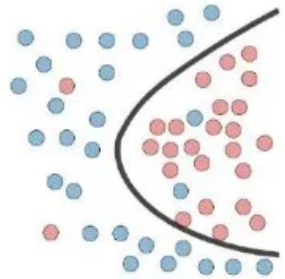
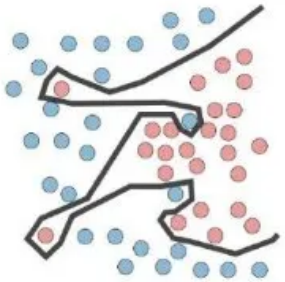
Predictions (1)

Show probabilities for: Classes known to the model — Restore Original Order

	Logistic Regression	Hours_Studied	Attendance	ırental_Involveme	:cess_to_Resourc	acurricular_Activi	Sleep_Hours	Previous_Scores
1	1.00 : 0.00 → Fail	5	55	Low	Low	No	6	45
2	0.00 : 1.00 → Pass	15	80	Medium	Medium	Yes	7	65
3	0.00 : 1.00 → Pass	25	90	High	High	Yes	8	85
4	0.55 : 0.45 → Fail	12	70	Low	Medium	No	6	60
5	0.00 : 1.00 → Pass	20	95	High	High	Yes	8	90

5 | M 1 | 5 | -

Courbe d'apprentissage

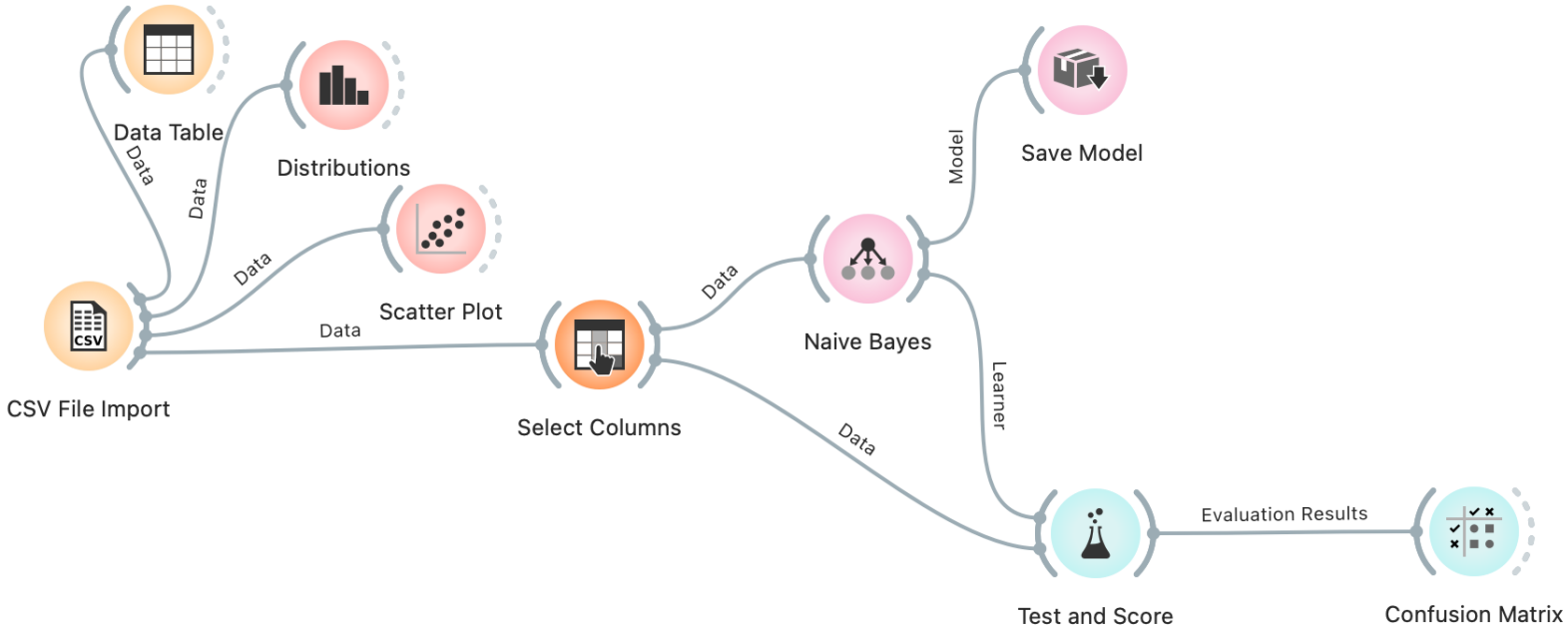
	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

Orange Data Mining

Activité

30 min

– Reprenons Titanic

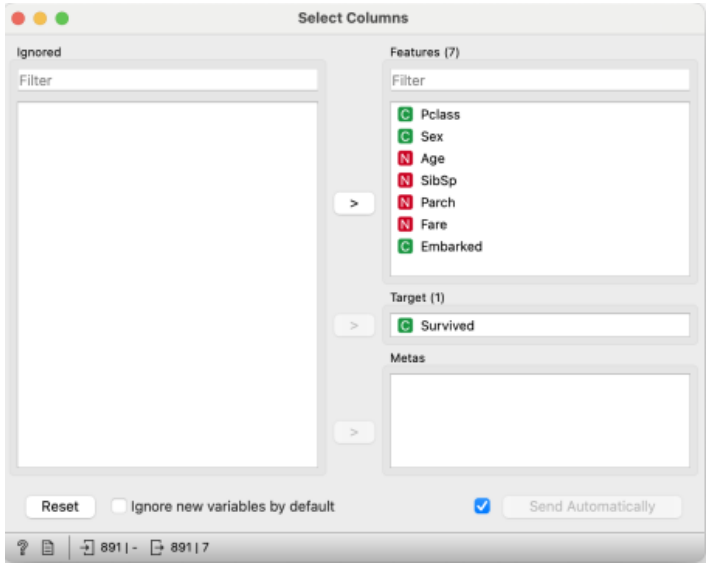


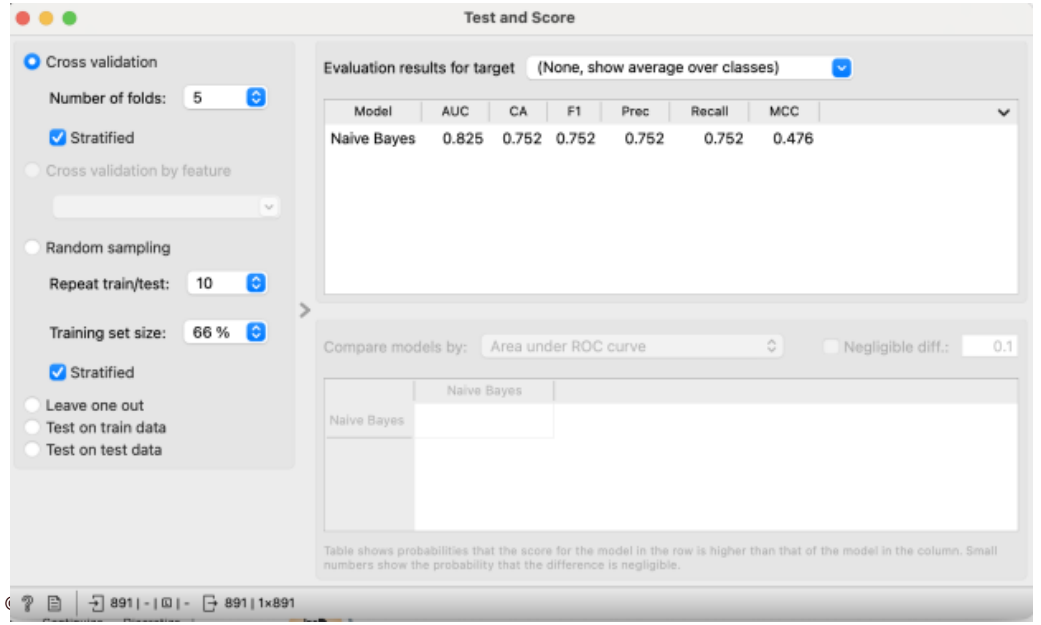
Orange Data Mining

Activité

30 min

– Reprenons Titanic





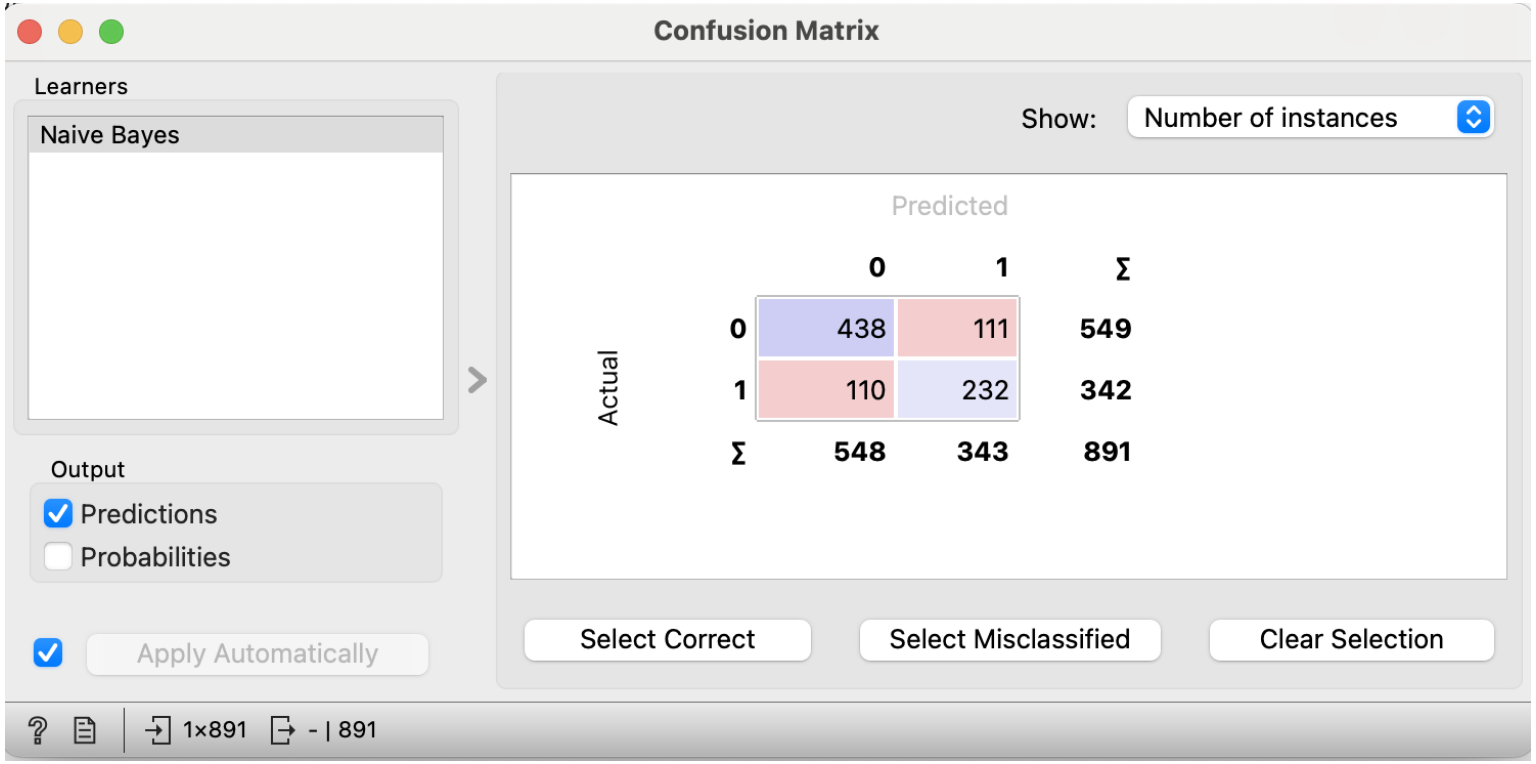
Orange Data Mining

Activité

30 min

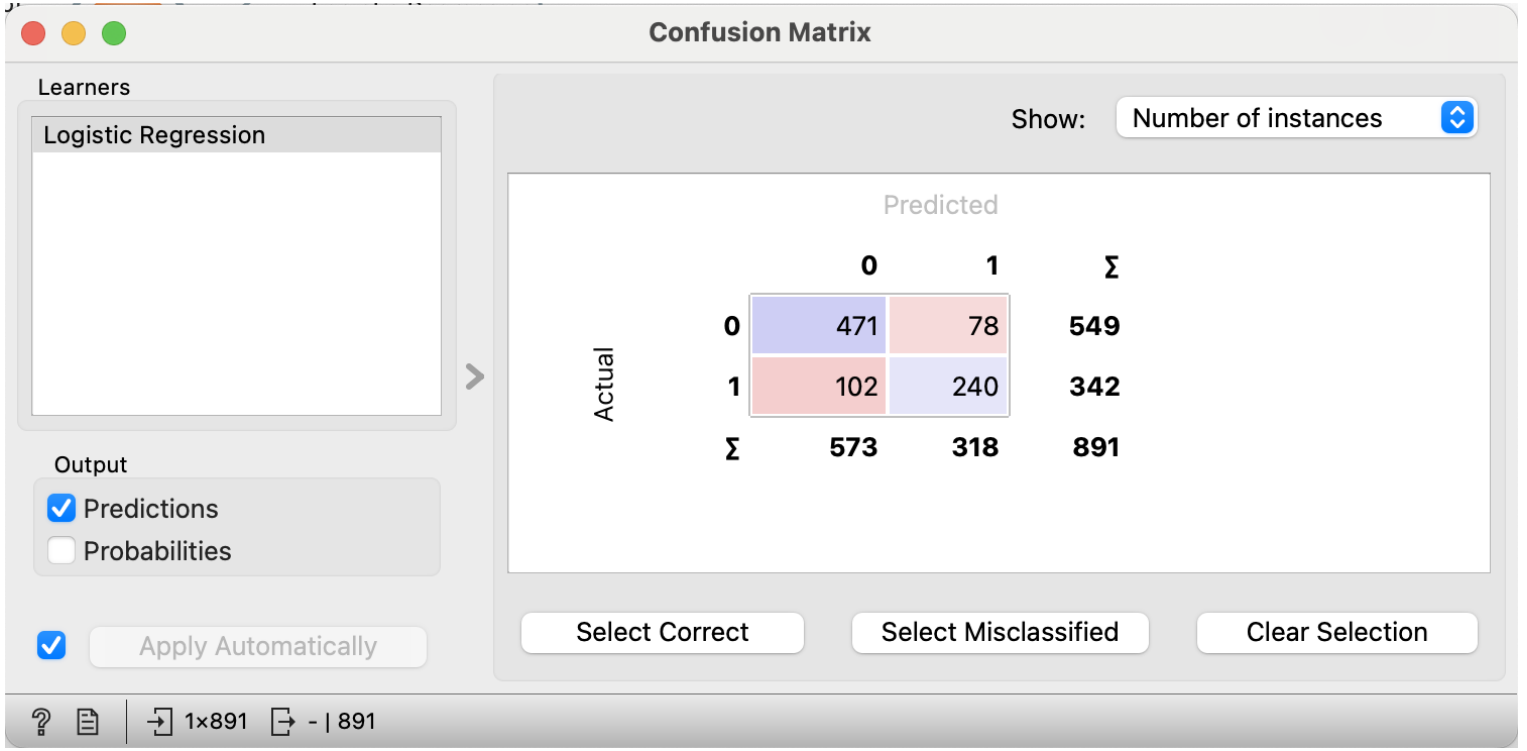
– Reprenons Titanic

- ▸Comparer avec la régression logistique



Reprenons Titanic

- Comparer avec la régression logistique

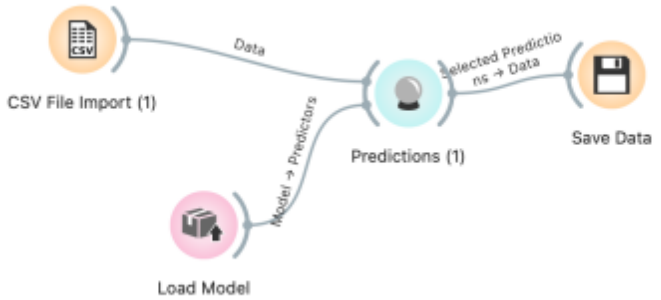


Orange Data Mining

Activité

30 min

– Reprenons Titanic



Predictions (1)

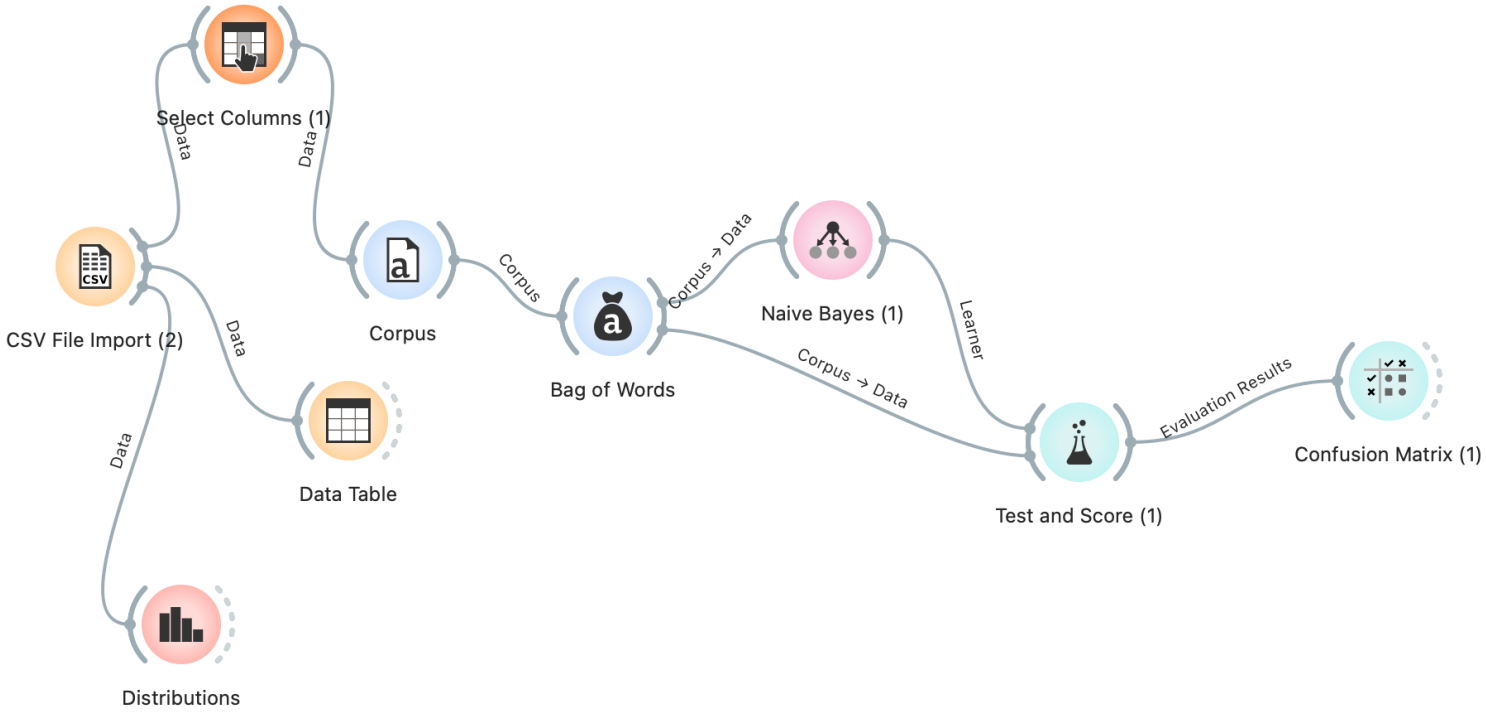
Show probabilities for: Classes known to the model

	Naive Bayes	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0.96 : 0.04 → 0	3	male	34.5	0	0	7.8292	Q
2	0.58 : 0.42 → 0	3	female	47	1	0	7	S
3	0.86 : 0.14 → 0	2	male	62	0	0	9.6875	Q
4	0.96 : 0.04 → 0	3	male	27	0	0	8.6625	S
5	0.26 : 0.74 → 1	3	female	22	1	1	12.2875	S
6	0.94 : 0.06 → 0	3	male	14	0	0	9.225	S
7	0.65 : 0.35 → 0	3	female	30	0	0	7.6292	Q
8	0.45 : 0.55 → 1	2	male	26	1	1	29	S
9	0.47 : 0.53 → 1	3	female	18	0	0	7.2292	C
10	0.91 : 0.09 → 0	3	male	21	2	0	24.15	S
11	0.97 : 0.03 → 0	3	male	?	0	0	7.8958	S
12	0.68 : 0.32 → 0	1	male	46	0	0	26	S
13	0.05 : 0.95 → 1	1	female	23	1	0	82.2667	S
14	0.64 : 0.36 → 0	2	male	63	1	0	26	S
15	0.05 : 0.95 → 1	1	female	47	1	0	61.175	S
16	0.06 : 0.94 → 1	2	female	24	1	0	27.7208	C
17	0.82 : 0.18 → 0	2	male	35	0	0	12.35	Q
18	0.94 : 0.06 → 0	3	male	21	0	0	7.225	C
19	0.45 : 0.55 → 1	3	female	27	1	0	7.925	S
20	0.56 : 0.44 → 0	3	female	45	0	0	7.225	C
21	0.19 : 0.81 → 1	1	male	55	1	0	59.4	C
22	0.92 : 0.08 → 0	3	male	9	0	1	3.1708	S
23	0.09 : 0.91 → 1	1	female	?	0	0	31.6833	S
24	0.19 : 0.81 → 1	1	male	21	0	1	61.3792	C
25	0.02 : 0.98 → 1	1	female	48	1	3	262.375	C
26	0.83 : 0.17 → 0	3	male	50	1	0	14.5	S
27	0.02 : 0.98 → 1	1	female	22	0	1	61.9792	C
28	0.94 : 0.06 → 0	3	male	22.5	0	0	7.225	C
29	0.68 : 0.32 → 0	1	male	41	0	0	30.5	S
30	0.78 : 0.22 → 0	3	male	?	2	0	21.6792	C

418 | 8 | 1 | 418 | -

Spams avec Naive Bayes

- Installer via: Options -> add-ons -> Text
- Charger le fichier : sms_spam.csv



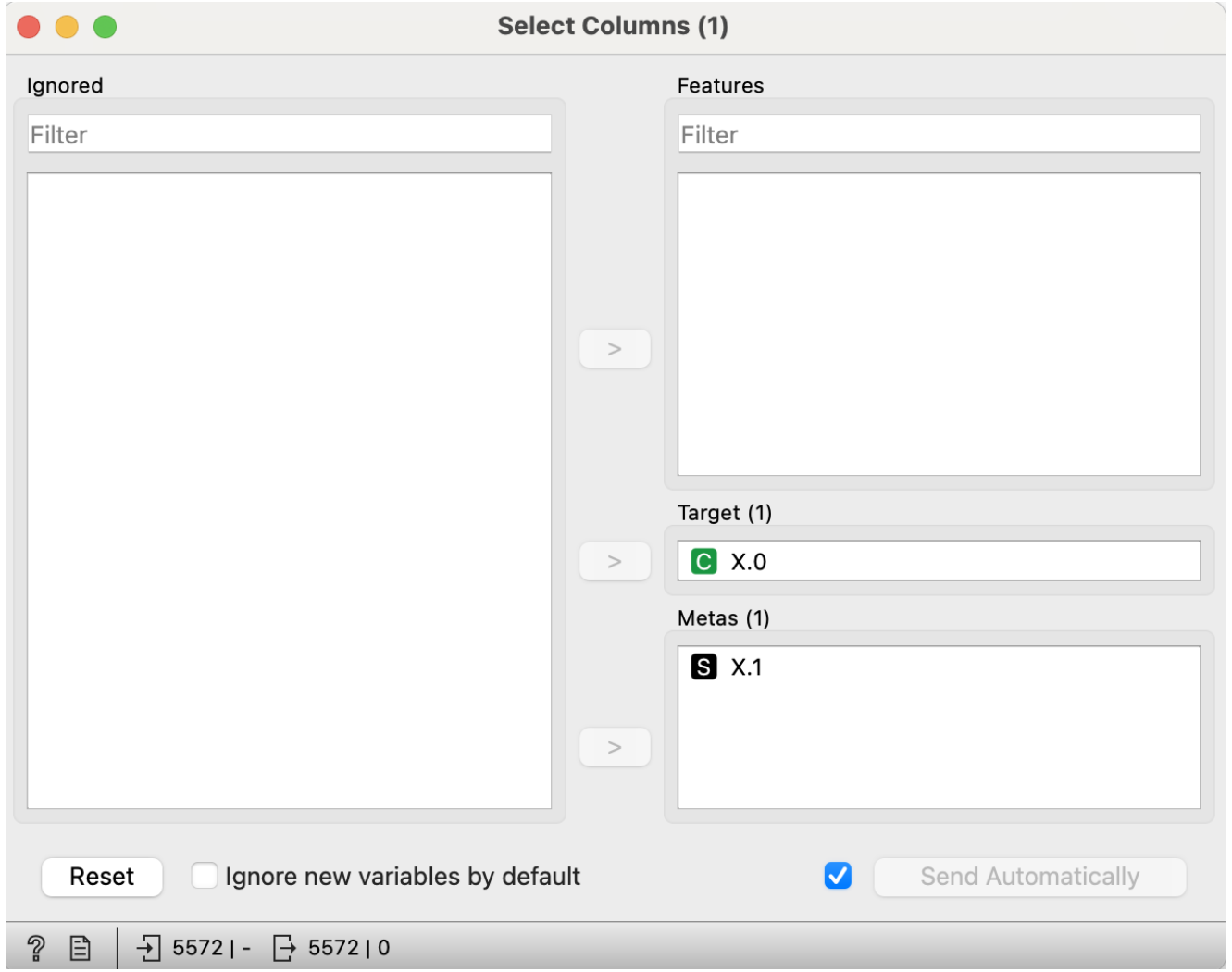
– Spams avec Naive Bayes

Column type

	1	2	3	4	5
1	v1	v2			
2	ham	Go until juro...			
3	ham	Ok lar... Joki...			
4	spam	Free entry in ...			
5	ham	U dun say so...			
6	ham	Nah I don't ...			
7	spam	FreeMsg He...			
8	ham	Even my ...			
9	ham	As per your ...			
10	spam	WINNER!! As...			
11	spam	Had your ...			
12	ham	I'm gonna be...			
13	spam	SIX chances ...			
14	spam	URGENT! Yo...			
15	ham	I've been ...			
16	ham	I HAVE A DA...			
17	spam	XXXMobileM...			
18	ham	Oh k...i'm ...			
19	ham	Eh u ...			

Reset | Restore Defaults | Cancel | OK

– Spams avec Naive Bayes

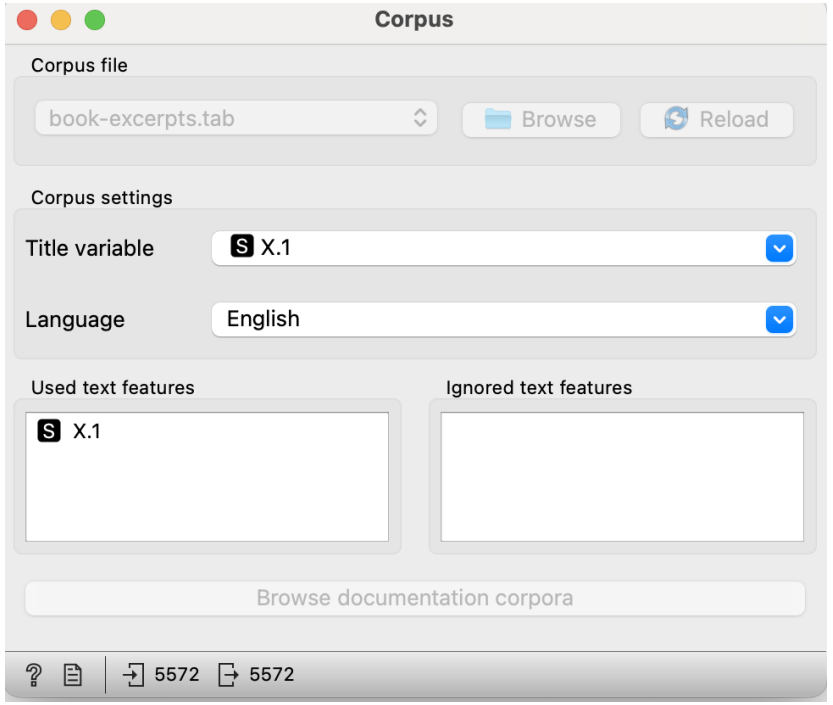


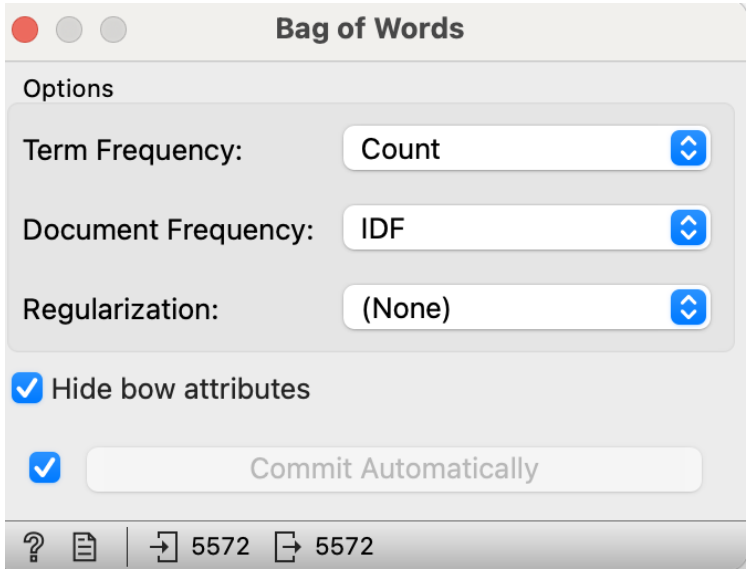
Orange Data Mining

Activité

30 min

– Spams avec Naive Bayes



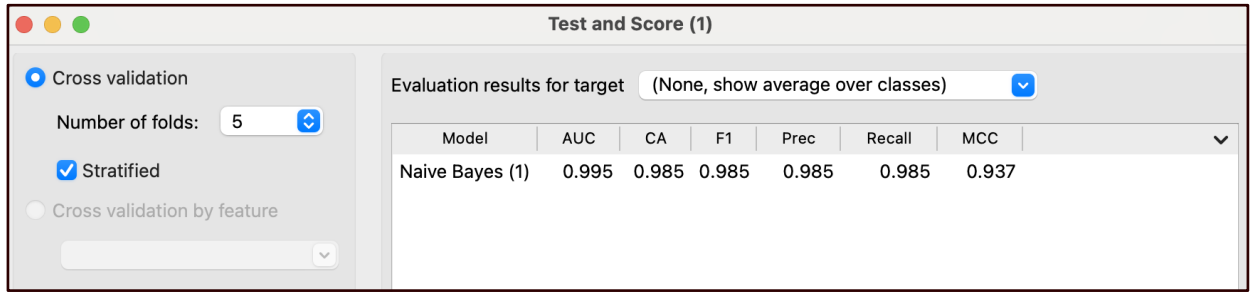


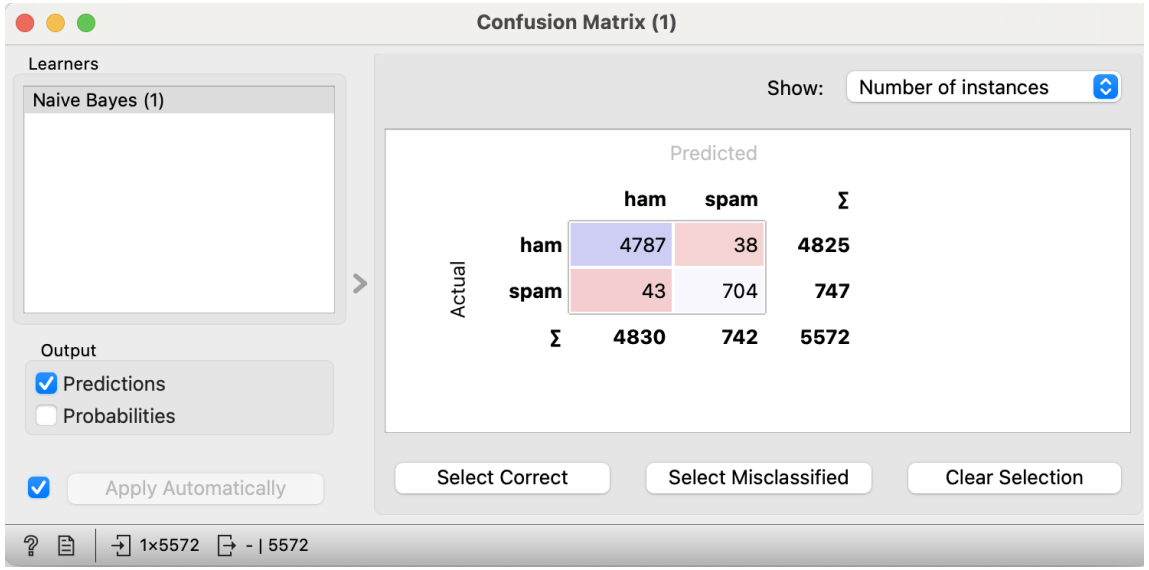
Orange Data Mining

Activité

30 min

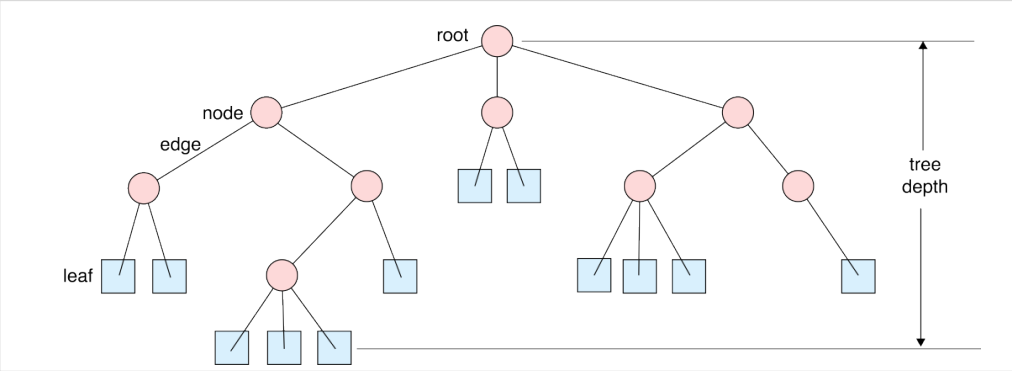
– Spams avec Naive Bayes





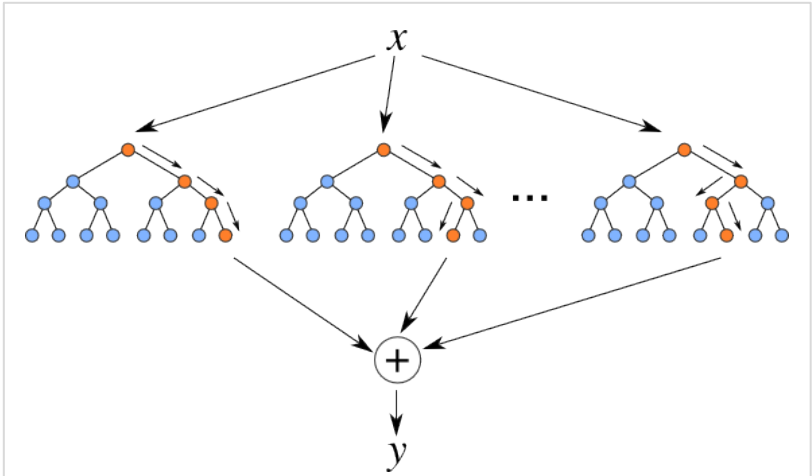
Arbre de décision

- Comment poser une série de questions, prendre une décision finale pour classifier un objet ?
- En apprentissage automatique, ce processus est appelé un arbre de décision
- Un nœud pose une question afin d'aider à classer les données
- Vous commencez avec un nœud qui se branche ensuite vers un autre nœud, en répétant ce processus jusqu'à ce que vous atteigniez une feuille. Une branche représente les différentes possibilités auxquelles ce nœud pourrait conduire. Une feuille est la fin d'un arbre de décision, ou un nœud qui n'a plus de branches



Random forest (forêt aléatoire)

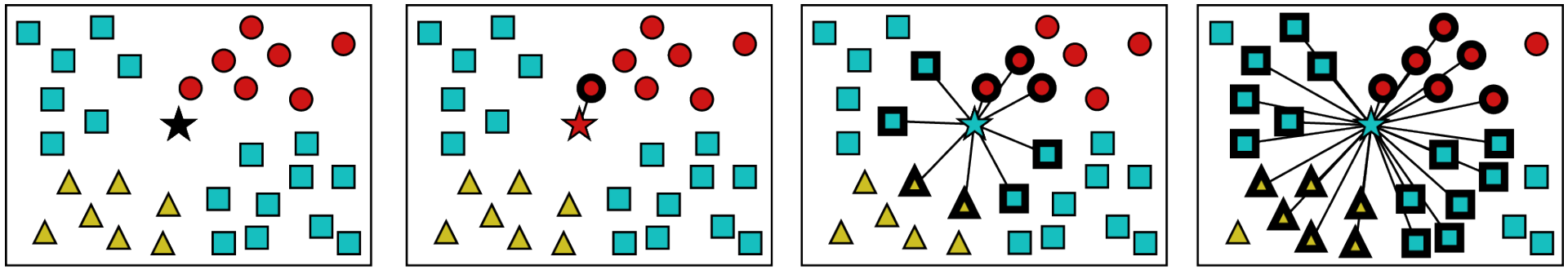
- L'algorithme de forêt aléatoire est composé de différents arbres de décision, chacun avec les mêmes nœuds, mais utilisant des données différentes qui conduisent à des feuilles différentes. Il fusionne les décisions de plusieurs arbres de décision afin de trouver une réponse, qui représente la moyenne de tous ces arbres de décision
- L'algorithme de forêt aléatoire est un modèle d'apprentissage supervisé ; il utilise des données étiquetées pour «apprendre » comment classer les données non étiquetées
- Avantages :
  - Empêche le surajustement (overfit) des données
  - Rapide à entraîner avec les données de test
- Les inconvénients :
  - Lent à créer des prédictions une fois le modèle créé
  - Sensibles aux valeurs aberrantes et aux trous dans les données



Source image : Logiciel d'arbre de décision

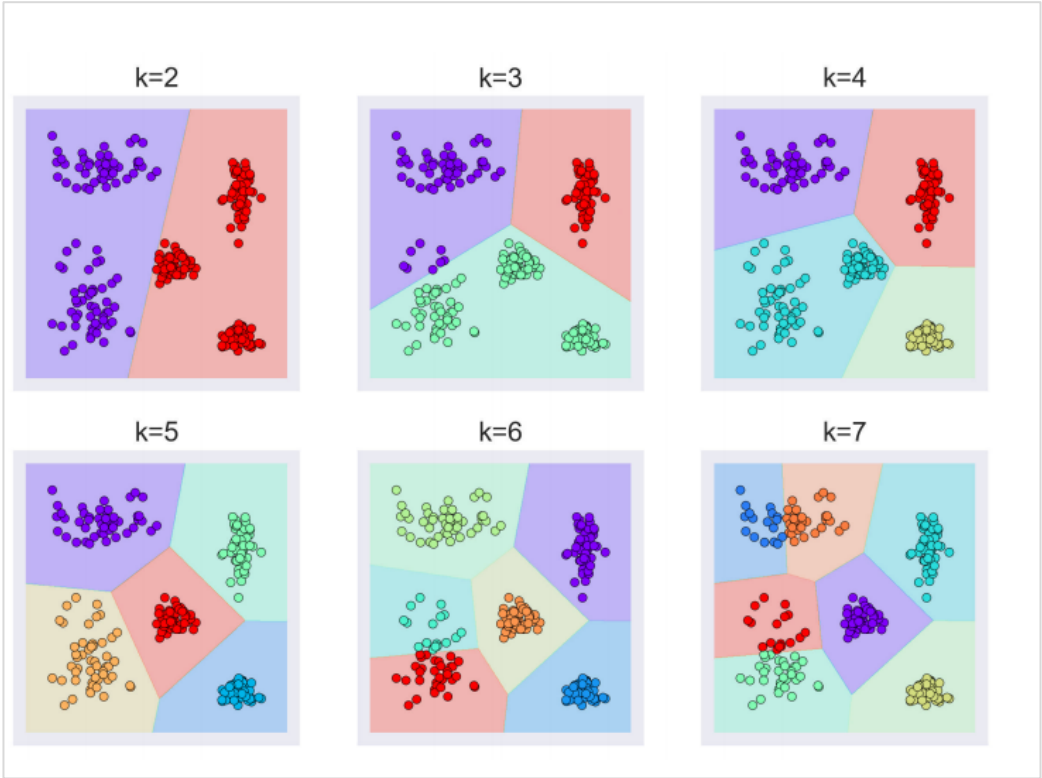
k-Nearest Neighbors (KNN)

- L'algorithme cherche les k plus proches voisins
  - K = 1 (b)
  - k = 9 (c)
  - K = 25 (d)
- Puis par mécanisme de vote attribue une catégorie



Non supervisé K-means

- K-means attribue des points de données à des catégories ou des clusters en trouvant la distance moyenne entre les points de données. Il répète ensuite cette technique afin d'effectuer des classifications plus précises au fil du temps
- Avantages : Rapide et efficace. Fonctionne sur des données numériques non étiquetées. Technique itérative
- Les inconvénients : Vous devez choisir votre propre valeur k. Beaucoup de répétition. Ne fonctionne pas bien lorsque des valeurs aberrantes sont présentes



Les limites actuelles de l'IA

Ce qui marche

- Tâches répétitives bien définies
- Patterns dans les données historiques
- Génération de contenu « inspiré »

Ce qui ne marche pas

- Raisonnement causal
- Adaptation à l'imprévu
- Créativité pure
- Jugement éthique

Exemple :
Recrutement ✓
Tri CV sur critères objectifs X
Évaluation soft skills

Points clés pour réussir avec l'IA

Activité

– Les 5 règles d'or :

Commencer petit, scalable

- Projet pilote < 3 mois
- ROI mesurable
- Périmètre limité mais extensible

Qualité des données d'abord

- Nettoyage rigoureux
- Sources fiables
- Mise à jour régulière

Focus sur la valeur business

- Objectifs SMART
- KPIs définis avant
- Impact mesurable

Validation humaine systématique

- Contrôle des résultats
- Feedback continu
- Amélioration itérative

Formation des équipes

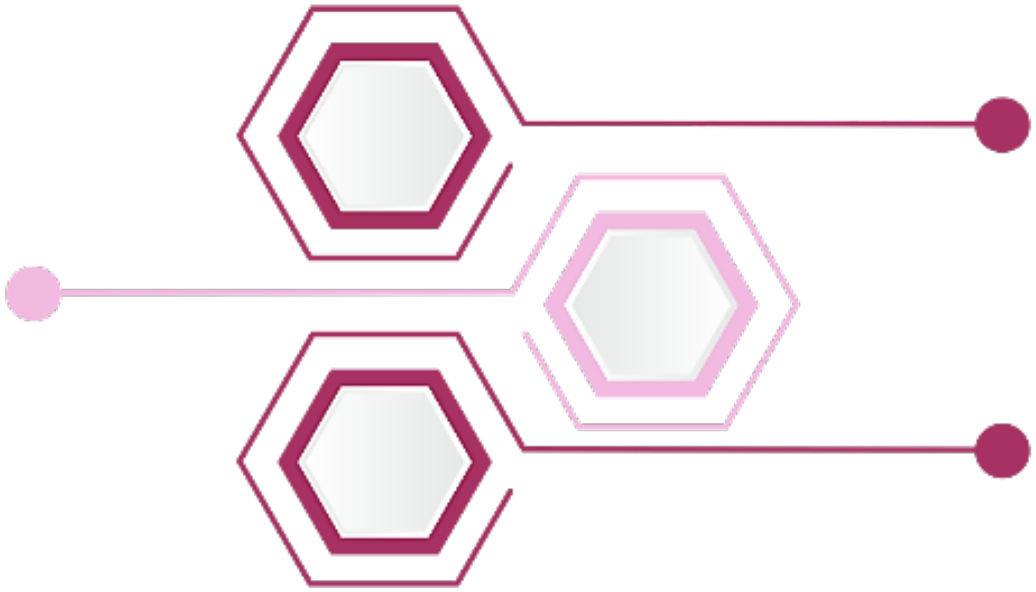
- Montée en compétences
- Adhésion utilisateurs
- Documentation claire

Enjeux Éthiques et Réglementation de l'IA

Contexte Réglementaire en 2025

- Le contexte réglementaire en 2025 est marqué par plusieurs textes clés :

IA Act : réglementation spécifique aux systèmes d'intelligence artificielle.



RGPD (Règlement Général sur la Protection des Données) : cadre légal pour la protection des données personnelles.

Directive NIS2 : renforcement de la cybersécurité pour les infrastructures critiques.

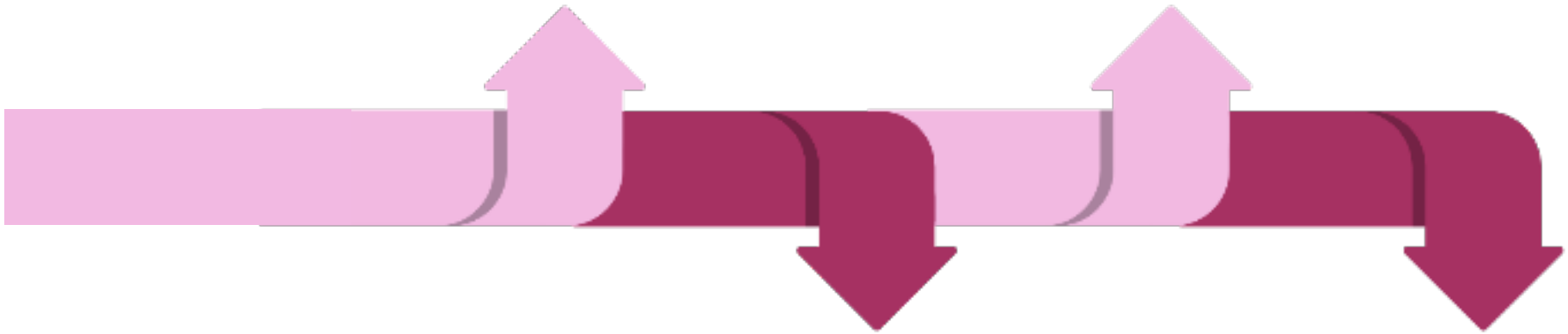
- Les amendes pour non-conformité peuvent atteindre 7 % du chiffre d'affaires mondial d'une entreprise.

Classification des Systèmes IA selon l'IA Act

- L'IA Act classe les systèmes IA en quatre catégories de risque :

Risque inacceptable :
systèmes interdits
(ex. : notation sociale par l'État).

Risque limité :
exigences de transparence
(ex. : chatbots).



Risque élevé :
systèmes soumis
à des obligations strictes
(ex. : IA pour le recrutement).

Risque minimal :
pas de réglementation spécifique
(ex. : filtres anti-spam).

Exigences pour les Systèmes à Risque Élevé

- Gestion des risques : mise en place de processus pour identifier, analyser et atténuer les risques liés à l'IA.
- Documentation technique : fournir une documentation détaillée sur le fonctionnement du système.
- Qualité des données : s'assurer que les données utilisées sont pertinentes, représentatives et exemptes de biais.
- Transparence et information : informer clairement les utilisateurs sur le fonctionnement du système.
- Surveillance humaine : maintenir un contrôle humain pour superviser les décisions prises par l'IA.

Gestion des Biais et de l'Équité

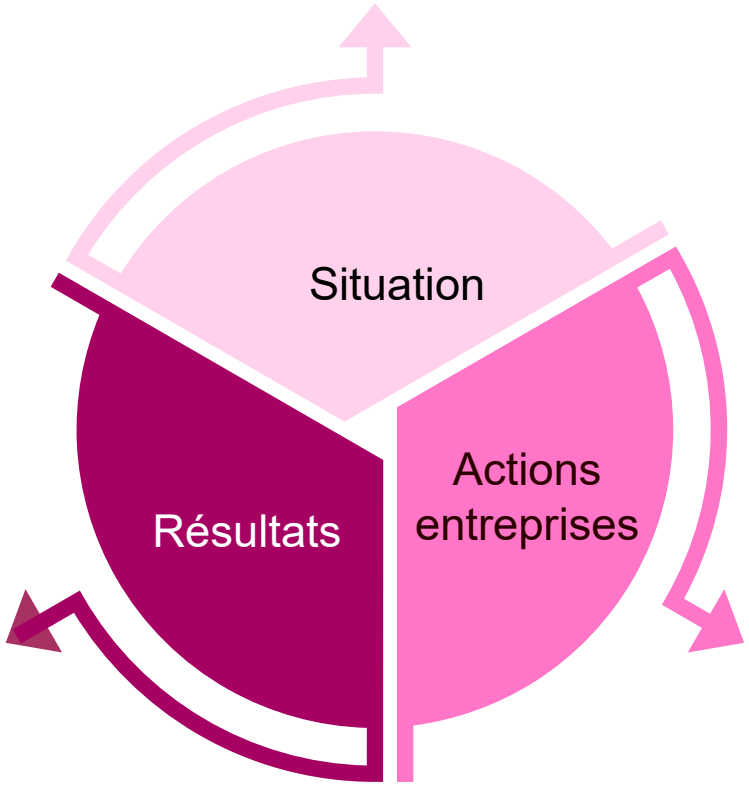
- Les biais dans les systèmes IA peuvent provenir de :
  - Données historiques biaisées : reflétant des discriminations passées.
  - Biais d'échantillonnage : échantillon non représentatif de la population cible.
  - Biais algorithmiques : introduits par les méthodes de modélisation

- Pour gérer ces biais :
  - Analyse exploratoire des données pour détecter les déséquilibres.
  - Techniques de rééchantillonnage pour équilibrer les données.
  - Métriques d'équité pour évaluer la performance du modèle sur différents groupes.
  - Audits réguliers pour surveiller et corriger les dérives.

Cas Pratique : Biais dans un Modèle de Recrutement

Une entreprise utilise un modèle d'IA pour filtrer les candidatures.
Il est constaté que le modèle favorise systématiquement les candidats masculins au détriment des candidates féminines.

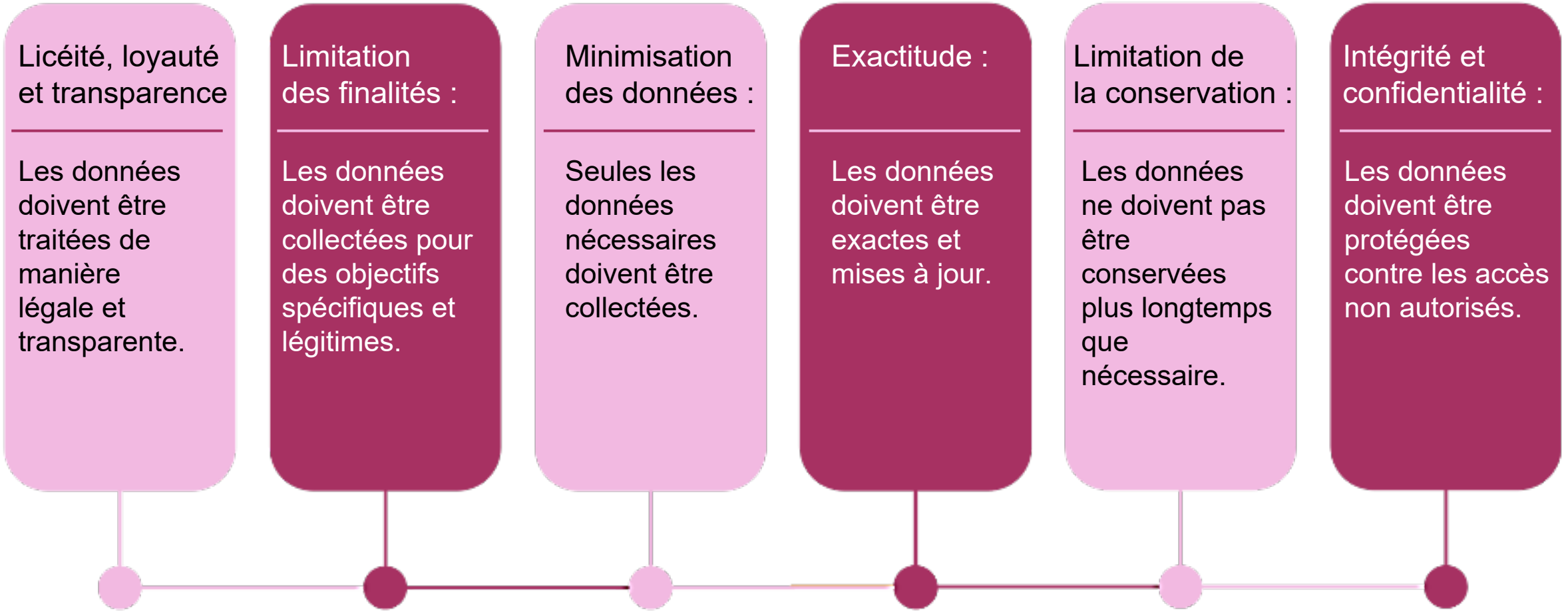
- Amélioration de l'équité dans le processus de recrutement.
- Conformité avec les réglementations anti-discrimination.
- Renforcement de la réputation de l'entreprise en matière de diversité.



- Audit des données : identification d'un déséquilibre dans les données d'entraînement, reflétant un historique de recrutement biaisé
- Rééquilibrage des données : ajustement pour inclure une représentation équitable des genres.
- Modification du modèle : application de contraintes d'équité pour assurer une sélection non discriminatoire

RGPD et IA : Principes Clés

Les principes du RGPD applicables à l'IA incluent :



Cas Pratique : Non-Conformité RGPD

- Mise en place d'un système de consentement sur les obligations légales
- Formation du personnel sur les obligations légales
- Révision des processus internes de collecte et de traitement des données.

Une entreprise de marketing utilise un modèle d'IA pour cibler des publicités personnalisées. Elle collecte des données de navigation sans consentement explicite des utilisateurs.

- Plainte déposée auprès de la CNIL.
- Amende de 500 000 € pour non-respect du RGPD
- Perte de confiance des clients et impact négatif sur la réputation.

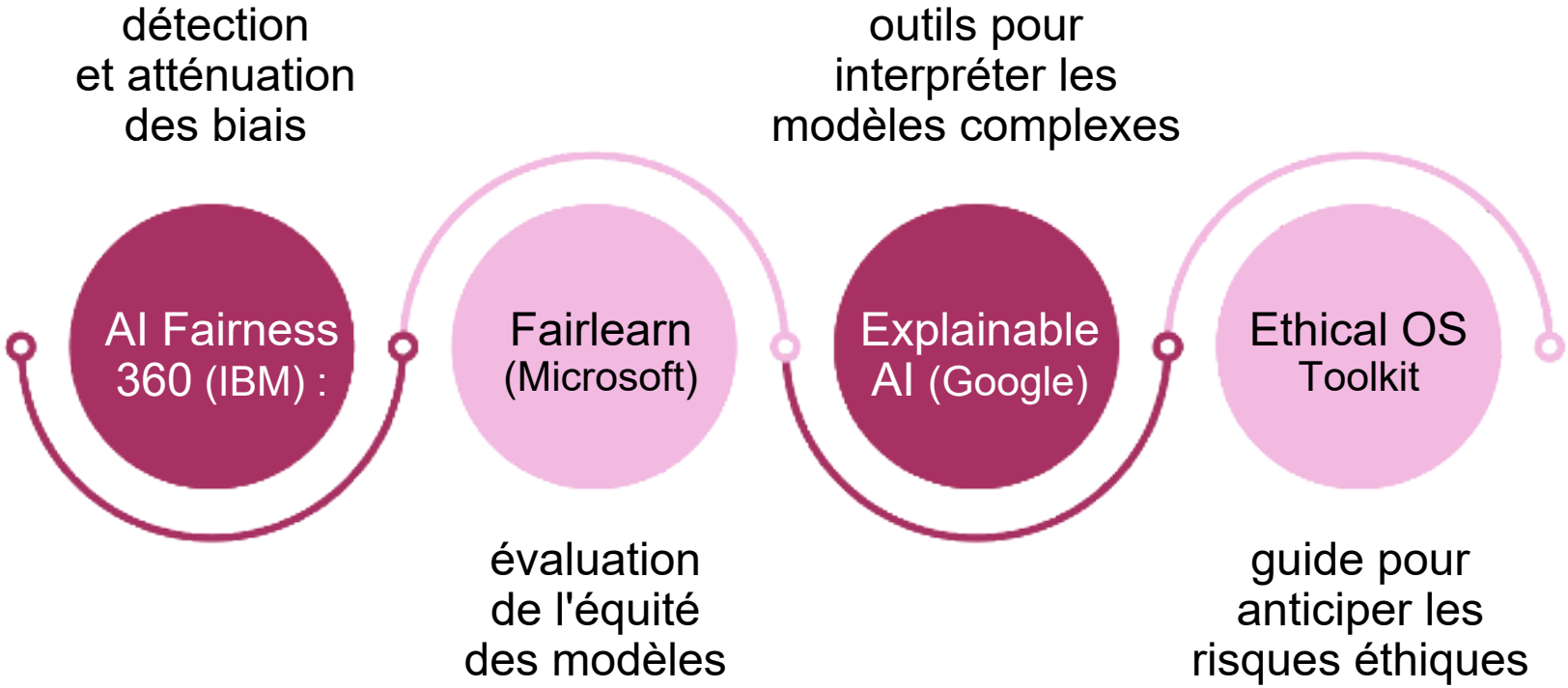
Guide pour une IA Responsable

Le framework E.T.H.I.C propose cinq principes pour une IA responsable :

1. Explicable : les décisions de l'IA doivent être compréhensibles.
2. Transparent : le fonctionnement et le développement de l'IA doivent être transparents.
3. Humain : maintenir une supervision humaine sur les systèmes IA.
4. Intègre : assurer l'éthique du système par des tests et validations réguliers.
5. Contrôlé : réaliser des audits externes pour vérifier la conformité.

Outils pour l'Éthique en IA

- Des outils pour intégrer l'éthique dans vos projets IA :



Documentation Légale Requise

- Pour assurer la conformité et faciliter les audits, il est important de maintenir une documentation complète :
  - Registre des traitements : détails sur les données collectées, les finalités, les bases légales, etc.
  - Analyse d'Impact sur la Protection des Données (AIPD) : évaluation des risques pour les droits et libertés des personnes.
  - Politique de confidentialité : informations claires pour les utilisateurs sur la gestion de leurs données.
  - Contrats avec les sous-traitants : clauses spécifiques sur la protection des données.

Rapport du gouvernement

Source : https://www.info-socialrh.fr/le-rapport-sur-lia-met-en-exergue-limportance-du-dialogue-social-extraits-895597.php

– Explication des différentes méthodes de scoring

Information Gain (Gain d'Information)
Valeurs : petal length (1.086), petal width (1.059), sepal length (0.624), sepal width (0.361)
Explication : Mesure la réduction d'entropie obtenue en divisant les données selon un attribut. Plus la valeur est élevée, plus l'attribut est informatif pour la classification. Les caractéristiques des pétales (longueur et largeur) sont nettement plus informatives que celles des sépales.

Gain Ratio (Ratio de Gain d'Information)
Valeurs : petal length (0.544), petal width (0.532), sepal length (0.313), sepal width (0.183)
Explication : Version normalisée du gain d'information qui pénalise les attributs ayant de nombreuses valeurs distinctes. Permet une comparaison plus équitable entre attributs. Le classement relatif des attributs reste similaire à celui du gain d'information.

Gini (Diminution de l'indice Gini)
Valeurs : petal length (0.423), petal width (0.407), sepal length (0.247), sepal width (0.154)
Explication : Mesure de la réduction de l'impureté (hétérogénéité) dans les sous-ensembles après division. Plus utilisée dans les algorithmes comme CART. Confirme également l'importance des caractéristiques des pétales.

ANOVA (Analyse de Variance)
Valeurs : petal length (1179.034), petal width (959.324), sepal length (119.265), sepal width (47.364)
Explication : Test statistique qui évalue si les moyennes de différents groupes sont significativement différentes. Valeurs élevées indiquent une forte relation entre l'attribut et la variable cible. Les écarts importants entre les valeurs soulignent la différence de pouvoir discriminant entre les caractéristiques des pétales et des sépales.

χ^2 (Chi-carré)
Valeurs : petal length (98.946), petal width (94.162), sepal length (79.243), sepal width (50.082)
Explication : Mesure l'indépendance entre variables. Une valeur élevée indique une forte dépendance entre l'attribut et la classe cible. Le classement des caractéristiques est cohérent avec les autres méthodes.

ReliefF
Valeurs : petal width (0.366), petal length (0.360), sepal length (0.136), sepal width (0.126)
Explication : Évalue la capacité d'un attribut à distinguer entre instances proches. Légèrement différent des autres métriques, ici la largeur des pétales est classée juste au-dessus de la longueur, mais avec une différence minime.

FCBF (Fast Correlation-Based Filter)
Valeurs : petal length (1.542), petal width (1.451), sepal width (0.255), sepal length (0.000)
Explication : Méthode qui identifie les attributs pertinents tout en éliminant les redondances. La valeur nulle pour sepal length suggère qu'elle est considérée comme redondante une fois les autres caractéristiques prises en compte.

cegos
Beyond knowledge

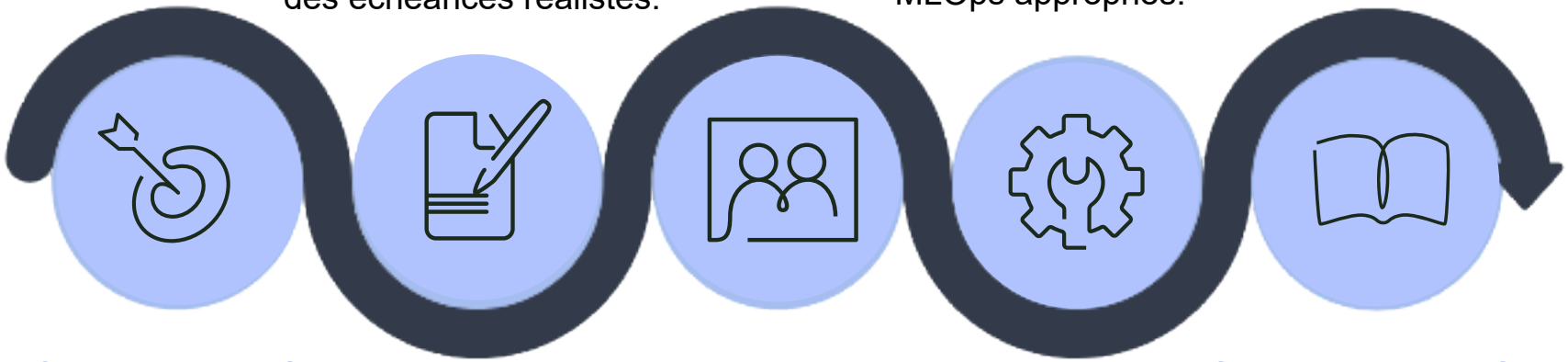
Conclusion

Plan d'Action Personnel

- Identifiez les actions que vous allez entreprendre dans les prochains jours pour appliquer ce que vous avez appris.

Planifier les étapes : établir un plan d'action avec des échéances réalistes.

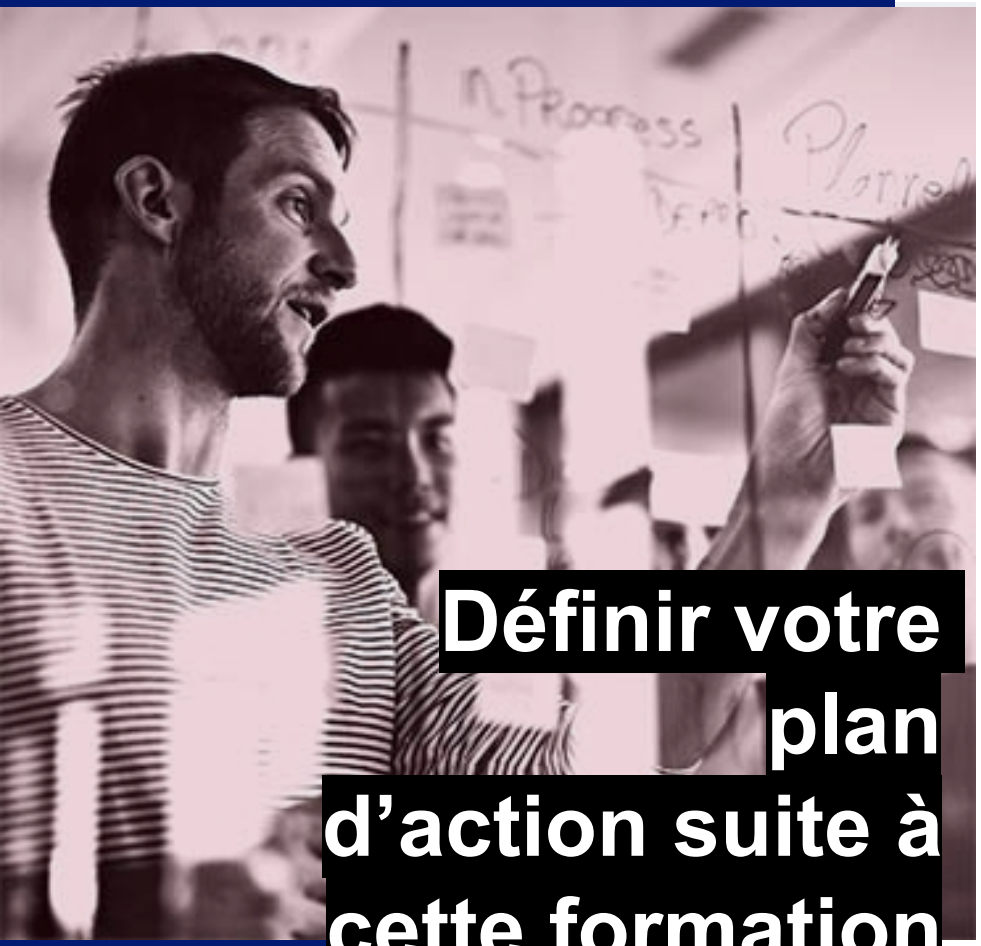
Mettre en place des outils : adopter les outils de gestion de projet, d'AutoML, de MLOps appropriés.



Définir des objectifs concrets : choisir un projet sur lequel appliquer ces compétences.

Partager avec votre équipe : présenter les apprentissages et proposer des améliorations aux processus existants.

Continuer à vous former : identifier des ressources pour approfondir certains sujets.



Définir votre plan d'action suite à cette formation

- Votre feedback sur la formation
- Quelles sont les 3 actions prioritaires que vous allez lancer pour utiliser l'IA en analyse de données ?
  - 1)
  - 2)
  - 3)

N’hésitez pas à nous contacter

Adresse
19 Rue René Jacques,
92130 Issy-les-Moulineaux

Email
espace-client@cegos.fr

Réseaux sociaux

Téléphone
01 55 00 90 90