

Pipeline de l'apprentissage automatique

Des données au modèle

Aziza Merzouki

Université de Genève

28 Novembre, 2025

Outline

- 1 Etapes générales du Machine Learning**
- 2 Collecte des données et préparation**
- 3 Feature Engineering**
- 4 Entraînement du modèle**
- 5 Évaluation du modèle**
- 6 Conclusion**

Etapes générales du Machine Learning

- 1 Définition de l'objectif
- 2 Collecte des données
- 3 Préparation des données et Feature Engineering
- 4 Choix de l'algorithme
- 5 Entrainement et optimisation du modèle
- 6 Validation et ajustement du modèle (hyper-paramètres)
- 7 Test et déploiement du modèle

Définition de l'objectif

- Identifier clairement **l'objectif** du projet de machine learning
- Comprendre les **problèmes** à résoudre ou les **questions** posées
- Définir les **métriques** d'évaluation pour mesurer le **succès** du modèle



Specify and
Understand Business
problem



Obtain subject-
matter expertise



Define scope of
analysis and results



Consider success and
risks



Aim towards
stakeholders

Collecte des données

- Évaluer l'utilité des données par rapport à l'objectif fixé
- Identifier les **sources** de données disponibles
- Déterminer si les données proviennent de sources **internes** ou de fournisseurs **tiers**
 - Considérer les différentes méthodes **d'accès** aux données, ex. bases de données SQL, web-scraping
- Définir la **quantité** de données nécessaire pour l'entraînement du modèle, ex., 3 mois, 1 an, etc.

1

Volume

The size or amount of data

2

Velocity

The speed at which data is being generated

3

Variety

Diversity or Formats of data

4

Value

Insights gained from data is useful to the organization

5

Veracity

Verifying and validating the data

Collecte des données

- Spécifier le type des données qui seront collectées :
 - **Données structurées** : Types bien définis (nombres, chaînes de caractères) stockés dans des bases de données faciles à interroger.
 - **Données non structurées** : Structure non définie, telles que des e-mails, des fichiers textes.
 - **Données de séries temporelles** : Séquence de nombres collectés à intervalles réguliers sur une certaine période.

Préparation des données

- Explorer et analyser les données pour comprendre leur **nature** et leur **qualité**.



Préparation des données

- Identifier et traiter les **valeurs aberrantes** (outliers) : points de données nettement différents du reste, ex. clipping.
- Gérer les **valeurs manquantes**, ex. supprimer les observations ou les features, imputer les données.

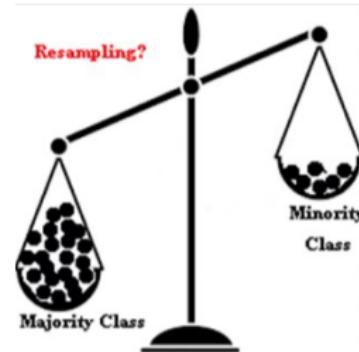
Inputs, X

Output, target, label, y

Country	Age	Salary	Purchased
France	44	72000	No
Spain	37	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	830000000	No
France	37	67000	Yes

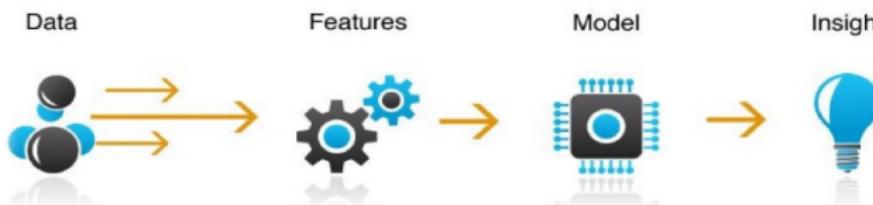
Préparation des données

- Gérer les données **déséquilibrées** : proportions différentes entre les classes.
 - Ex. dans la détection de fraude, il peut y avoir beaucoup plus de transactions normales que de transactions frauduleuses.
 - Méthodes: sur-échantillonnage aléatoire (Random Over-Sampling), sous-échantillonnage aléatoire (Random Under-Sampling).



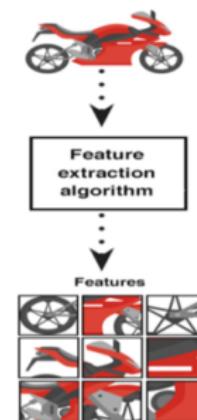
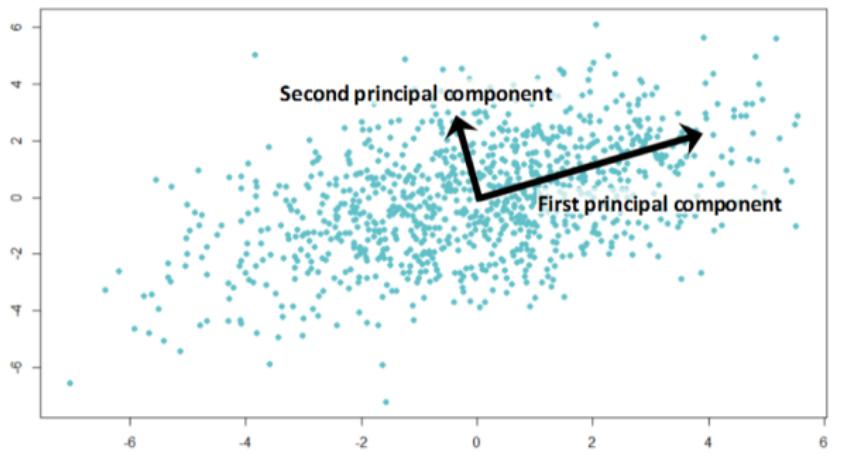
Feature Engineering - Définition

- Extraire des caractéristiques (features), généralement sous la forme de **colonnes structurées**.
- Une caractéristique est un **attribut disponible** dans les données.
- Une meilleure ingénierie des caractéristiques augmente les chances d'obtenir de **meilleures performances** du modèle.



Feature Engineering - Extraction

- **Extraction des caractéristiques** : Convertir les données brutes en caractéristiques numériques **uniques** permettant de représenter de manière précise et complète l'ensemble de données d'origine.



Feature Engineering - Transformation

- **Transformation des caractéristiques :** Appliquer aux caractéristiques des techniques telles que la **normalisation**, l'**encodage** des variables catégorielles, la **discrétisation** et l'application de fonctions **logarithmiques**.

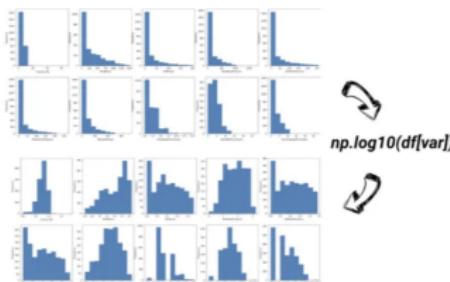
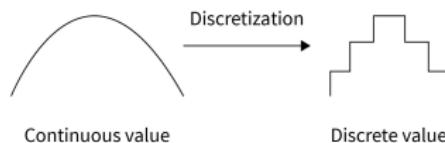
MinMaxScaler()

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

StandardScaler()

$$x' = \frac{x - \text{mean}}{\text{standard deviation}}$$

Classe	Code	Couleur	Rouge	Vert	Bleu
Eco	0	Rouge	1	0	0
Business	1	Vert	0	1	0
Super	2	Bleu	0	0	1

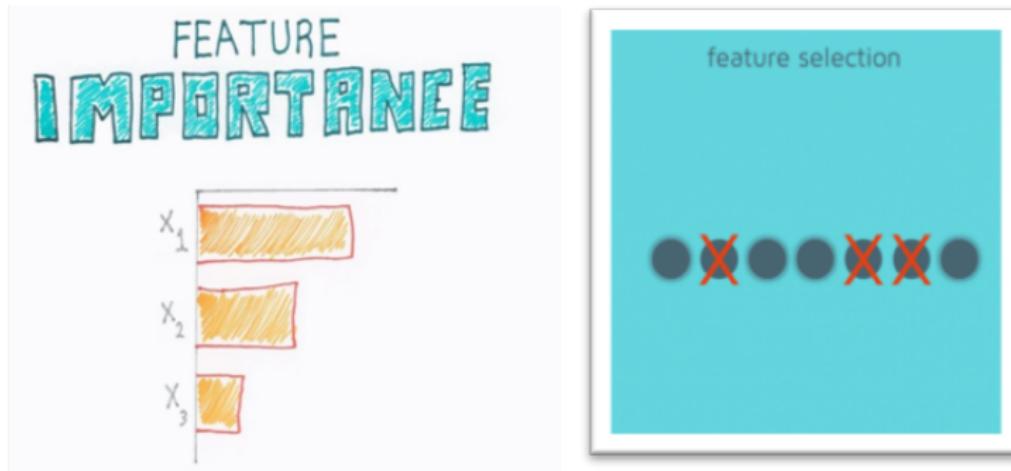


[Label Encoding in Python Explained.](#)

[Log Transformation: Purpose and Interpretation.](#)

Feature Engineering - Sélection

- **Sélection des caractéristiques** : Utiliser des méthodes d'évaluation d'importance et de sélection des caractéristiques pour obtenir les caractéristiques requises par le modèle.



Beginner's guide for feature selection.

Feature Selection with Embedded Methods.

Quizz

Analyse Exploratoire des Données

Maggie est une data scientist dans une entreprise qui développe un système de messagerie. Ils veulent construire un ensemble de modèles de machine learning pour traiter le spam et comprendre les comportements de leurs clients. La première étape du processus pour Maggie est de réaliser une Analyse Exploratoire des Données (EDA). Elle souhaite faire une présentation à son équipe sur tout ce qu'ils accompliront pendant cette phase.

Lesquelles des étapes suivantes Maggie réalise-t-elle pendant ce processus ?

- Pendant la phase d'Analyse Exploratoire des Données, Maggie évaluera la performance de son modèle.
- Pendant la phase d'Analyse Exploratoire des Données, Maggie surveillera la performance de son modèle.
- Pendant la phase d'Analyse Exploratoire des Données, Maggie examinera la distribution des caractéristiques de son ensemble de données pour identifier des problèmes potentiels ou des schémas qu'elle pourra utiliser pour améliorer son modèle.
- Pendant la phase d'Analyse Exploratoire des Données, Maggie évaluera la qualité de son ensemble de données, y compris en vérifiant les valeurs manquantes ou corrompues, pour éviter une mauvaise performance du modèle.

Quizz

Sélection des caractéristiques

Ben travaille avec une entreprise immobilière pour classifier les propriétés dans la ville. Ils souhaitent organiser chaque propriété par sa valeur. Il veut commencer par organiser les données et préparer les caractéristiques pour aider à construire un modèle de machine learning pour prédire la valeur des propriétés potentielles.

Lesquelles des colonnes du jeu de données de Ben sont des caractéristiques numériques pertinentes pour le modèle ?

- Le numéro de téléphone de la propriété.
- La superficie de la propriété.
- Le nombre de chambres.
- Le nombre de salles de bains.

Quizz

Encodage

Kali travaille dans une entreprise qui gère une plateforme éducative. Elle travaille sur un projet de classification des données de performance des élèves. Son équipe doit gérer quelques caractéristiques catégorielles dans le jeu de données. Un membre de l'équipe a suggéré d'utiliser l'encodage par étiquettes, tandis qu'un autre a préconisé l'encodage One-Hot. Kali doit décider, alors elle a demandé à ses collègues d'écrire un résumé des deux approches.

Voici quelques points ressortis dans le résumé, mais ils sont contradictoires.

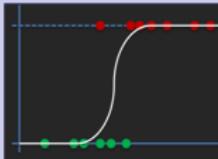
Veuillez sélectionner ceux que vous pensez corrects.

- L'encodage One-Hot remplace chaque étiquette de la caractéristique catégorielle par un entier unique basé sur l'ordre alphabétique.
- L'encodage One-Hot crée des caractéristiques supplémentaires en fonction du nombre de valeurs uniques dans la caractéristique catégorielle.
- L'encodage par étiquettes remplace chaque étiquette de la caractéristique catégorielle par un entier unique basé sur l'ordre alphabétique.
- L'encodage par étiquettes crée des caractéristiques supplémentaires en fonction du nombre de valeurs uniques dans la caractéristique catégorielle.

Choix de l'algorithme

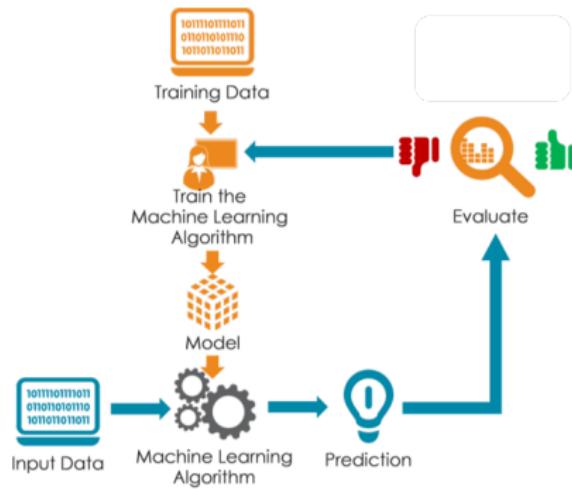
Choisir l'algorithme approprié en fonction du:

- Type de **problème**
- Taille et type des **données**
 - Petites vs grandes données
 - Données numériques vs catégoriques vs textuelles vs données d'images
- **Précision** du modèle

Binary (Logistic Regression)	Classification (SVM)	Classification(LDA)
 <ul style="list-style-type: none">• Credit risk• Medical conditions• Person will perform action or not	 <ul style="list-style-type: none">• Customer Classification• Mails Classification	 <ul style="list-style-type: none">• Topic Discovery• Sentiment Analysis• Automated Document Tagging

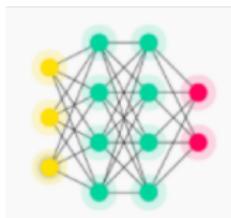
Entraînement du Modèle

- Fournir à un algorithme ML des **données d'entraînement** à partir desquelles il peut **apprendre**.
- Utiliser ce modèle ML pour obtenir des **prédictions** sur de nouvelles données.

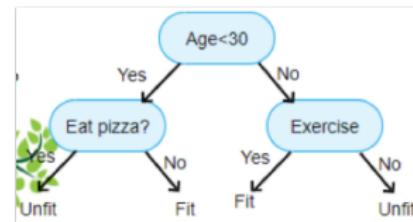


Modèle de machine learning

- Le terme "modèle de machine learning" fait référence à la **représentation** de ce qui a été appris lors du processus d'entraînement.
- La représentation du modèle peut prendre différentes **formes** selon le type d'algorithme de machine learning et le problème résolu, ex. un ensemble de **poids** et de **biais** appris pour un réseau neuronal, ou la **structure** d'un arbre de décision avec les **critères de division** appris à chaque nœud.



$$y = ax + b$$

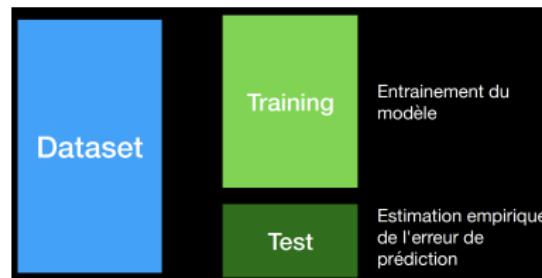


Prérequis pour l'entraînement du modèle

- Séparation des données en ensembles **d'entraînement**, de **validation** (facultatif) et de **test** pour estimer la **performance** du modèle sur des données non vues.
Pas de pourcentage de séparation optimal. Peut être de 80%-20%, 75%-25% ou autres.
 - Ensemble de **validation** : Utilisé pour ajuster les hyperparamètres de l'algorithme.



Validation croisée (Cross-validation)



(a) Train-Test Split



(b) Validation croisée k-Fold

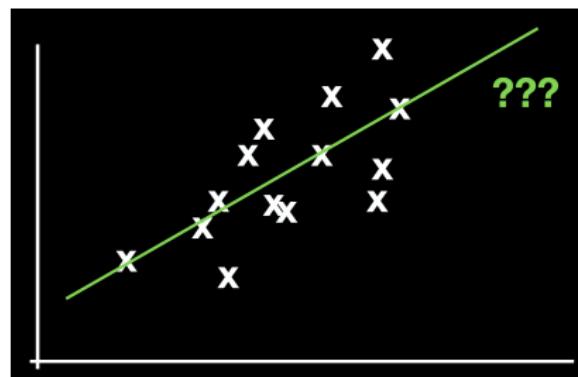
Entraînement du Modèle - Fonction de Coût

L'optimisation des paramètres du modèle vise à **minimiser la fonction de coût**.

- La fonction de coût est une mesure de **l'écart** entre les **prédictions** du modèle et les **valeurs réelles**.
- Elle quantifie à quel point le modèle est **performant**.
- En **minimisant la fonction** de coût, on **ajuste les paramètres** du modèle pour qu'il soit plus précis.

Exemple: Régression Linéaire

- Etablir la relation linéaire entre une variable cible y et une ou plusieurs variables d'entrée x .
- Données: $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$.
- Modèle: $h_w(x)$ est une équation linéaire de la forme $y = w_1 + w_2x$, où w_1, w_2 sont les coefficients du modèle.



Exemple: Régression Linéaire - Fonction de Coût

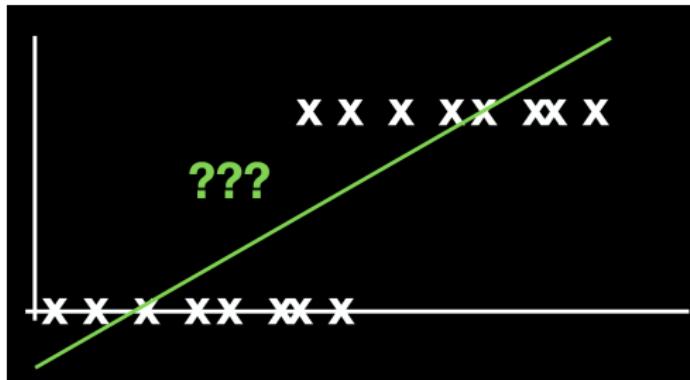
La fonction de coût utilisée pour l'algorithme de **régression linéaire** est **l'erreur quadratique moyenne** (Mean Squared Error, MSE) :

$$MSE = J(w_1, w_2, \dots) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

où y_i sont les valeurs réelles, \hat{y}_i sont les prédictions du modèle $h_w(x_i)$, et N est le nombre d'échantillons.

Exemple: Régression Logistique

- Utilisée pour la classification binaire, et prédire des probabilités de classes.
- Données: $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, où $y_i = 1$ si i appartient à la classe positive C, 0 sinon

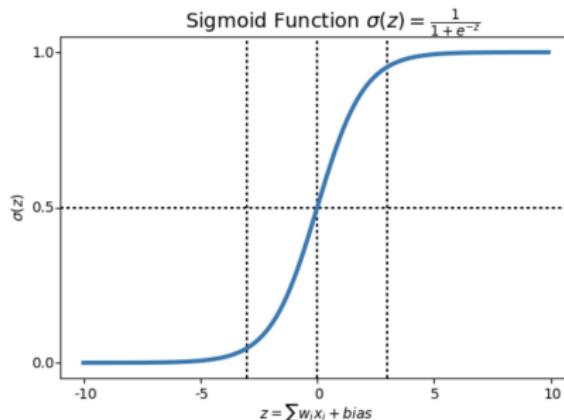


Exemple: Régression Logistique

- Modèle de la probabilité d'appartenir à la classe C:

$$h_w(x) = P(y = 1) = \sigma(w_0 + w_1 x)$$

, où w_0, w_1 sont les poids du modèle et (σ) la fonction sigmoid.



Exemple: Régression Logistique - Fonction de Coût

La fonction de coût utilisée pour l'algorithme de **régression logistique** est la **cross-entropy**:

- La cross-entropy (entropie croisée) est une mesure de l'écart entre les probabilités prédites \hat{y} et les vraies étiquettes y .
- Pour un seul exemple de donnée i , la formule de cross-entropy est :

$$J(y_i, \hat{y}_i) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

où y_i est la vraie étiquette (0 ou 1) et \hat{y}_i est la probabilité prédite.

- L'objectif de l'entraînement est de minimiser la moyenne de toutes les valeurs $J(y_i, \hat{y}_i)$ pour tous les N exemples d'entraînement.

Exemple: Classification Multiclasse - Fonction de Coût

L'entropie croisée est couramment utilisée comme fonction de coût pour évaluer un modèle de classification multiclasse.

Pour K classes, l'entropie croisée est définie comme suit :

$$J(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_i^{(k)} \log(\hat{y}_i^{(k)})$$

Où :

$y_i^{(k)}$: La véritable étiquette (1 si l'exemple i est de la classe k , sinon 0)

$\hat{y}_i^{(k)}$: La probabilité que l'exemple i soit de la classe k selon le modèle

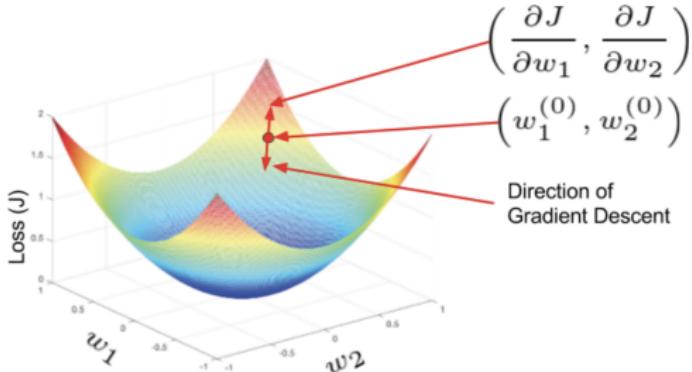
Méthodes d'optimisation

- **L'optimisation** est une étape cruciale dans l'entraînement du modèle, visant à **ajuster les paramètres** pour **minimiser la fonction de coût**.
- Différentes méthodes d'optimisation sont utilisées pour atteindre cet objectif.

Méthode de Gradient Descent

- Principe : ajuster les paramètres dans la **direction opposée du gradient** (pente) de la fonction de coût.
- Processus itératif : à chaque étape, les paramètres sont mis à jour proportionnellement au gradient et à un **taux d'apprentissage** λ .

$$w^{(s+1)} = w^{(s)} - \lambda \nabla J$$



Gradient Descent Algorithm: How Does it Work in Machine Learning?.

Méthode de Stochastic Gradient Descent (SGD)

- Variante du Gradient Descent.
- Au lieu d'utiliser l'ensemble complet de données pour calculer le gradient, SGD utilise un seul échantillon (**mini-batch**) à chaque étape.
- **Accélère** le processus d'entraînement et rend la convergence plus rapide.

[Gradient Descent Algorithm: How Does it Work in Machine Learning?](#)

Méthode de Stochastic Gradient Descent avec Momentum

- Amélioration du SGD standard.
- Il introduit un concept de "**momentum**" pour accélérer la convergence et atténuer les **oscillations** et éviter les minima locaux dans la descente du gradient.
- Le momentum ajoute une composante proportionnelle à la **direction précédente du gradient** lors de la mise à jour des paramètres.

Méthode d'Adam (Adaptive Moment Estimation)

- Méthode d'optimisation adaptative qui combine les avantages de SGD et d'autres techniques.
- Il ajuste les taux d'apprentissage pour chaque paramètre en fonction des estimations du premier moment (moyenne) et du second moment (variance non centrée) des gradients.
- Cela permet **d'adapter automatiquement le taux d'apprentissage** à chaque paramètre et à chaque itération.

[Gentle Introduction to the Adam Optimization Algorithm for Deep Learning.](#)

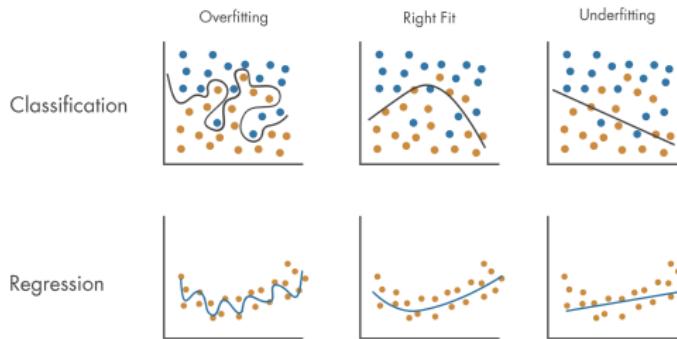
Généralisation du Modèle

- Un modèle entraîné doit être évalué et testé pour s'assurer qu'il peut **généraliser correctement** sur de **nouvelles données**.
- Le modèle doit être adapté aux données d'entraînement tout en évitant le **sur-apprentissage** (overfitting) et le **sous-apprentissage** (underfitting).

[Overfitting and Underfitting with Learning Curves.](#)

Ajustement du Modèle (Model Fitting)

- Un modèle **bien ajusté** produit des résultats précis sur les données d'entraînement et les nouvelles données.
- Un modèle **surajusté** (overfitted) correspond trop étroitement aux données d'entraînement, et ne généralise pas efficacement.
- Un modèle **sous-ajusté** (underfitted) ne correspond pas suffisamment aux données d'entraînement ni aux nouvelles données.



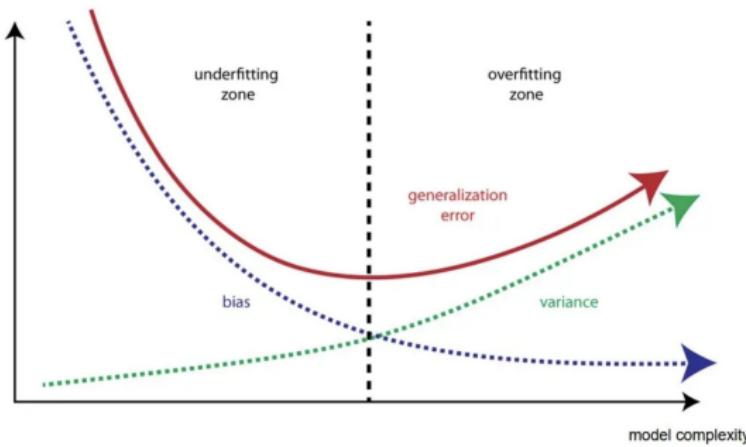
Dilemme Biais-Variance

- Les concepts de **biais** et de **variance** sont cruciaux pour comprendre la performance et la généralisation du modèle.
- **Biais** : Erreur due à des hypothèses simplificatrices du modèle, conduisant à des prédictions erronées sur les données d'entraînement et de test.
- **Variance** : Sensibilité excessive aux variations dans les données d'entraînement, entraînant des fluctuations importantes des prédictions entre différents ensembles de données.
- Le biais et la variance sont en compétition : réduire l'un augmente l'autre.

[Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning.](#)

Lien entre Biais, Variance et Généralisation

- Un modèle **surajusté** a une **variance élevée** mais un **faible biais**. Il apprend le bruit des données et ne généralise pas bien.
- Un modèle **sous-ajusté** a un **biais élevé** mais une **faible variance**. Il manque de complexité pour capturer les modèles sous-jacents.



Quizz

Jeux de Données - Division

Raegan est ingénierie en logiciel dans une entreprise qui développe une application de chat. Récemment, elle a commencé à suivre un cours en ligne sur le machine learning et a découvert le concept de division des jeux de données avant l'entraînement d'un modèle. Bien que le cours ait discuté de l'importance de diviser les données en ensembles d'entraînement, de validation et de test, il n'a pas expliqué pourquoi nous faisons cela.

Lesquelles des affirmations suivantes expliquent pourquoi nous divisons un jeu de données avant d'entraîner un modèle ?

- Nous divisons le jeu de données pour empêcher le modèle de surapprendre.
- Nous divisons le jeu de données pour réduire la mémoire nécessaire pour entraîner le modèle.
- Nous divisons le jeu de données pour accélérer le processus d'entraînement.
- Nous divisons le jeu de données pour évaluer précisément la performance du modèle.

Quizz

Jeux de Données - Mélange

Tessa est une data scientist travaillant dans le domaine de la vision par ordinateur. Elle a été chargée d'un projet impliquant la construction d'un modèle pour prédire des résultats spécifiques basés sur des données visuelles. Avant d'entraîner son modèle, Tessa divise son jeu de données en un ensemble d'entraînement et un ensemble de test. Elle comprend que mélanger le jeu de données avant de le diviser est crucial pour assurer des résultats précis.

Quelle est la raison principale de mélanger le jeu de données avant de le diviser en un ensemble d'entraînement et un ensemble de test ?

- Pour s'assurer que l'ensemble d'entraînement contient plus de données que l'ensemble de test.
- Pour s'assurer que les caractéristiques dans les ensembles d'entraînement et de test sont les mêmes.
- Pour s'assurer que les étiquettes de classe sont réparties de manière égale entre les ensembles d'entraînement et de test.
- Pour s'assurer que les ensembles d'entraînement et de test sont également difficiles pour que le modèle apprenne et évalue.

Quizz

Jeux de Données - Utilisation

Charlee est une data scientist dans le département financier d'une startup. Elle travaille sur un projet de construction d'un modèle de machine learning pour prédire les prix des actions de l'entreprise. Charlee sait que le succès de son modèle aura un impact significatif sur la performance financière de l'entreprise. Avant d'entraîner son modèle, Charlee doit diviser son jeu de données en trois ensembles : un ensemble d'entraînement, un ensemble de validation et un ensemble de test. Elle sait qu'il est essentiel d'utiliser chaque ensemble à des fins spécifiques pour assurer des résultats précis.

Lesquelles des affirmations suivantes sont vraies concernant les jeux de données d'entraînement et de validation pendant le processus de développement ?

- Le jeu de données d'entraînement ne doit être utilisé qu'une seule fois avant de tester le modèle avec l'ensemble de test.
- Le jeu de données d'entraînement peut être utilisé plusieurs fois tout au long du processus de développement du modèle.
- L'ensemble de validation ne doit être utilisé qu'une seule fois avant de tester le modèle avec l'ensemble de test.
- L'ensemble de validation peut être utilisé plusieurs fois tout au long du processus de développement du modèle.

Quizz

Modèles à haute variance

Thea assistait à une conférence sur le machine learning, désireuse d'élargir ses connaissances et de se connecter avec des professionnels du domaine. Lors d'une table ronde, l'un des intervenants a abordé le sujet des modèles à haute variance et de leur sensibilité aux données d'entraînement. Thea a écouté attentivement, car elle savait que ce serait un concept essentiel à comprendre pour son futur travail.

Lesquels des algorithmes suivants peuvent être considérés comme des modèles à haute variance ?

- Régression linéaire
- Régression logistique
- Arbres de décision
- k-Plus Proches Voisins (k-NN)

Évaluation du modèle

- Evaluer le modèle (entraîné) sur **l'ensemble de test** pour mesurer sa **performance**.
- Comprendre à quel point les prédictions du modèle correspondent aux valeurs réelles.
- Utiliser différentes **métriques** pour évaluer différents **types de modèles** (Définies à l'étape 1).



Métriques pour la Régression

- Erreur quadratique moyenne (MSE):

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

- Racine de l'erreur quadratique moyenne (RMSE):

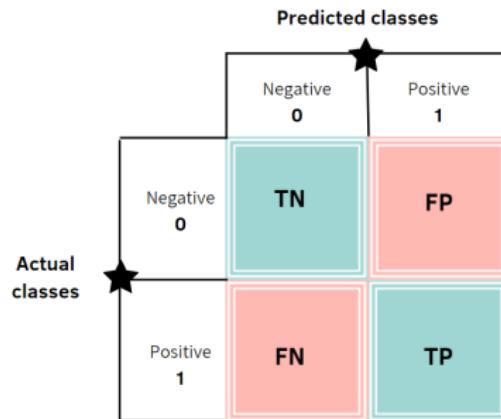
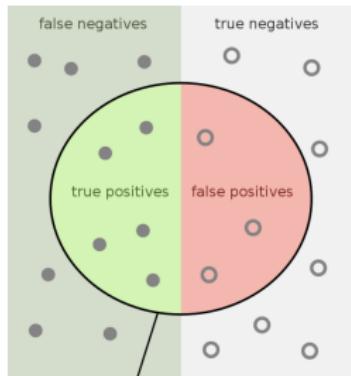
$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

- Erreur absolue moyenne (MAE):

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

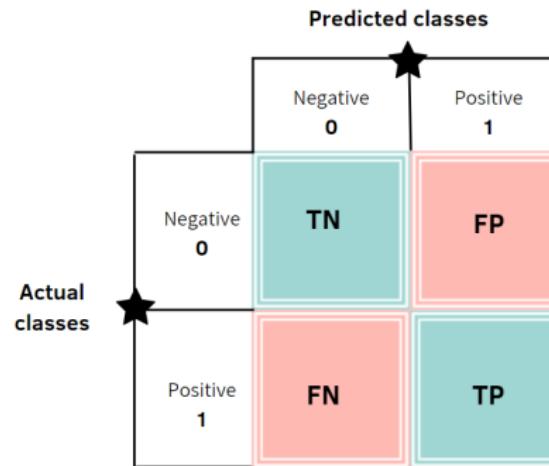
Métriques pour la Classification Binaire

- Evaluer la capacité du modèle à prédire des étiquettes binaires (par exemple, Oui/Non, Spam/Pas Spam).



Accuracy

■ **Accuracy** : $\frac{TP+TN}{TP+TN+FP+FN}$



Precision et Recall

- **Precision** : $\frac{TP}{TP+FP}$
 - Parmi tous les "vrais" prédicts, combien sont corrects?
- **Recall** : $\frac{TP}{TP+FN}$
 - Parmi tous les "vrais" réels, combien sont capturés?



F1-score

- **F1-score** : $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
 - Moyenne harmonique de la précision et du rappel (recall).

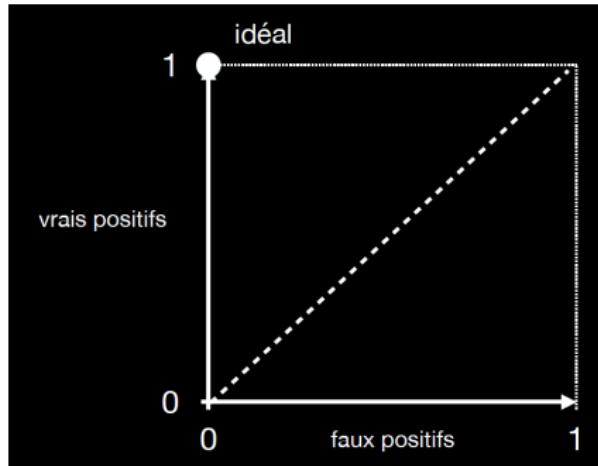
Sensitivité et Spécificité

- **Sensitivité = Recall :** $\frac{TP}{TP+FN}$
 - Parmi tous les "vrais" réels, combien sont capturés?
- **Spécificité :** $\frac{TN}{TN+FP}$
 - Parmi tous les "faux" réels, combien sont capturés?



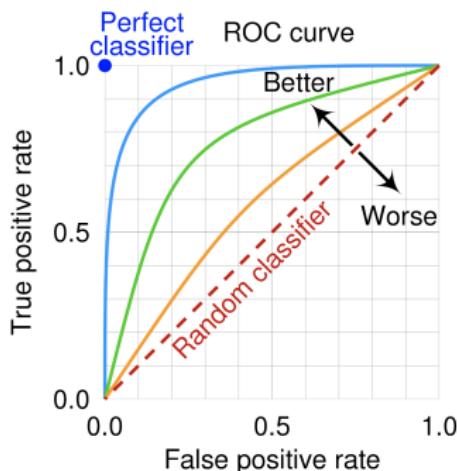
Courbe ROC

- Vue du **compromis** entre le **taux de vrais positifs** (TPR) et le **taux de faux positifs** (FPR) d'un modèle à différents **seuils de classification**.
 - Taux de vrais positifs (TPR) = Sensibilité : $\frac{TP}{TP+FN}$
 - Taux de faux positifs (FPR) = 1- Spécificité : $\frac{FP}{TN+FP}$



Courbe ROC

- Vue du **compromis** entre le **taux de vrais positifs** (TPR) et le **taux de faux positifs** (FPR) d'un modèle à différents **seuils de classification**.
 - Taux de vrais positifs (TPR) = Sensibilité : $\frac{TP}{TP+FN}$
 - Taux de faux positifs (FPR) = 1- Spécificité : $\frac{FP}{TN+FP}$



Métriques pour la Classification Multi-classes

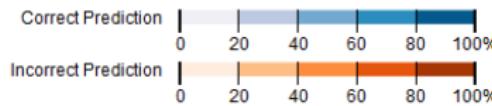
- Evaluer la capacité du modèle à prédire trois étiquettes ou plus (par exemple, Cloudy/Rain/Shine/Sunrise, Romance/Adventure/Thriller).

		Reality			
Confusion matrix		Cloudy	Rain	Shine	Sunrise
Prediction	Cloudy	39	6	13	1
	Rain	3	23	0	0
	Shine	6	0	30	1
	Sunrise	12	10	13	68

Score F1 Moyenné par Macro

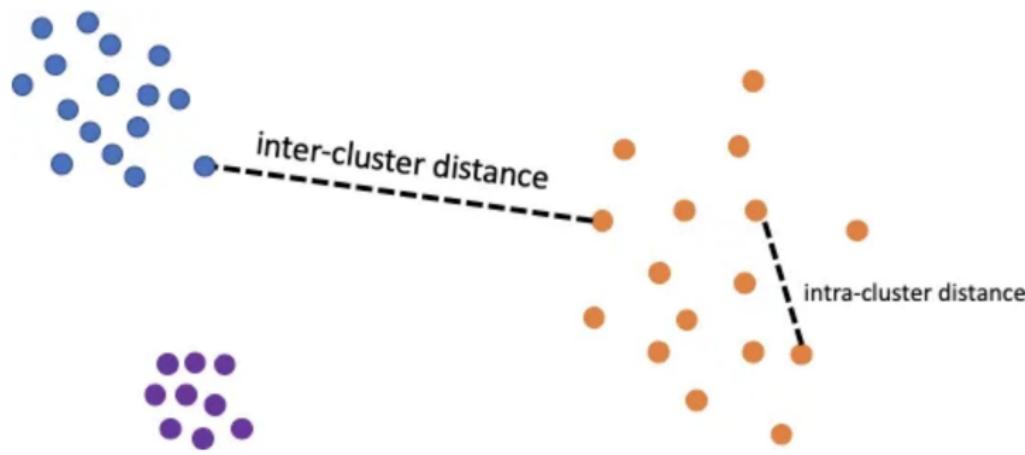
Macro average F1 score = $\frac{1}{K} \sum_{k=1}^K F1 \text{ score for class } k$

		Predicted Values			
		Romance	Thriller	Adventure	Total
True Values	Romance	57.92% (49.1k)			0.78
	Thriller	21.23% (18.0k)			0.33
True Values	Adventure	20.85% (17.7k)			0.32
	Total	77.56% (65.8k)	9.33% (7910)	13.12% (11.1k)	100.00% (84.8k)



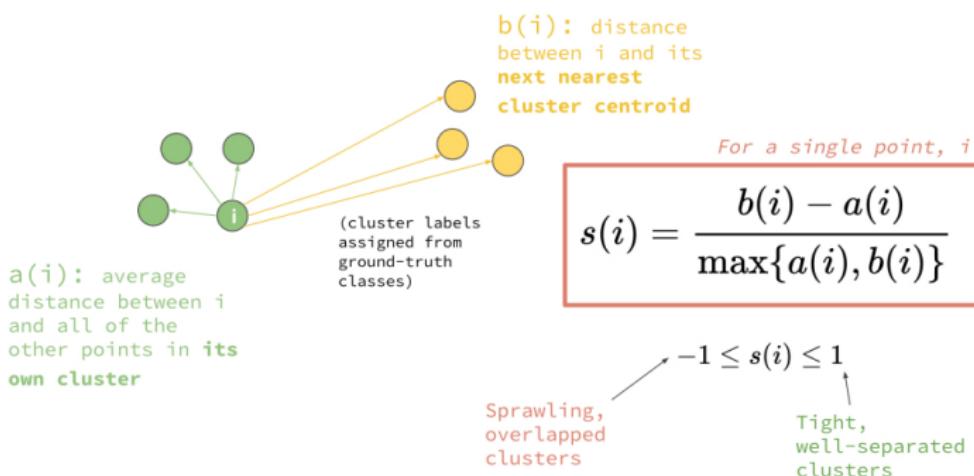
Métriques pour le Clustering

- Evaluer la qualité d'un clustering, i.e., petite variance intra-cluster (les points dans un groupe sont proches les uns des autres) et une grande variance inter-cluster (les clusters sont distants les uns des autres).



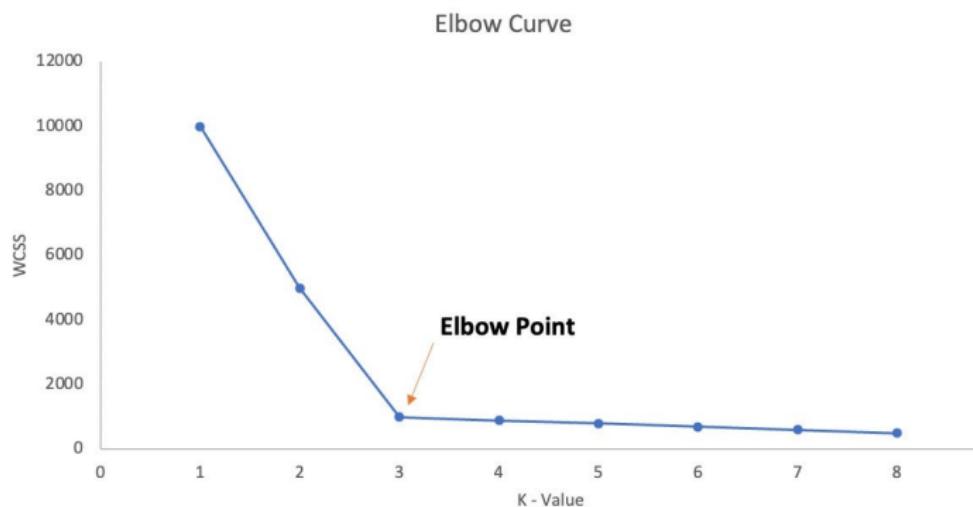
Silhouette

- Un coefficient de silhouette plus élevé indique des groupes mieux définis.



Méthode du coude WCSS

- WCSS (Within Cluster Sum of Squares) est la somme des carrés des distances entre les points de données et leur centre de cluster (inertie intra-classe).



Using the elbow method to determine the optimal number of clusters for k-means clustering.

Choix des Métriques

- Le choix des métriques dépend du **type de problème** et des **objectifs**.
- Par exemple, dans la détection de fraude, le rappel peut être plus important que la précision.
- Il est important d'utiliser **plusieurs métriques** pour obtenir une **vue complète** de la performance du modèle.

Déploiement du modèle

- Déployer le modèle dans un environnement de **production** pour une utilisation en temps réel.
- Surveiller les **performances** du modèle en production et effectuer des mises à jour si nécessaire.
- Intégrer le modèle dans des applications ou des **systèmes existants** pour prendre des décisions basées sur les prédictions du modèle.

[A Gentle Introduction to Concept Drift in Machine Learning.](#)

[Why You Should Care About Data and Concept Drift.](#)

Quizz

Fonctions de Perte

Mariana veut définir la fonction de perte correcte pour son modèle. Elle a lu qu'elle devrait se concentrer sur la différence positive moyenne entre les prédictions de son modèle et les valeurs cibles. Mais c'est ce que dit le livre. Maintenant, c'est à Mariana de traduire cela en termes techniques.

Laquelle des mesures d'erreur suivantes est celle dont Mariana a besoin ?

- Erreur Quadratique Moyenne (MSE)
- Erreur Absolue Moyenne (MAE)
- Erreur Positive Moyenne (MPE)
- Erreur Quadratique Moyenne Racine (RMSE)

Quizz

Méthode du Coude

En travaillant sur un projet d'analyse de données, Blair a utilisé l'algorithme K-Means. C'était sa première expérience avec un algorithme de clustering. K-Means était simple et rapide. Blair a écrit tout le code en quelques heures, mais il manquait quelque chose. La documentation mentionnait la "méthode du coude", mais Blair n'était pas sûre de son utilité.

Parmi les affirmations suivantes, laquelle est vraie à propos de la méthode du coude ?

- La méthode du coude détermine le nombre optimal de clusters.
- La méthode du coude détecte les outliers présents dans un jeu de données.
- La méthode du coude identifie les biais au sein d'un jeu de données.
- La méthode du coude détermine quelles caractéristiques expliquent le mieux les schémas observés dans un jeu de données.

Quizz

Recherche ML vs Production

Hallie a toujours été fascinée par l'IA et est impatiente de commencer sa carrière dans ce domaine. Cependant, elle ne sait pas si elle doit se concentrer sur la recherche ou se lancer dans un rôle plus orienté vers l'industrie. Pour l'aider à prendre une décision éclairée, nous allons souligner quelques différences clés entre le ML en environnement de recherche et ML en production.

Sélectionnez toutes les affirmations correctes.

- La priorité dans un environnement de recherche est généralement axée sur une performance plus élevée, soit la précision la plus élevée ou d'autres métriques pertinentes. Les environnements de production mettent davantage l'accent sur les coûts, la scalabilité et l'explicabilité.
- Dans un environnement de recherche, la plupart des travaux sont centrés sur l'entraînement initial et la validation du modèle. En production, il y a un focus important sur la surveillance et la maintenance des modèles.
- Les données utilisées dans un environnement de recherche sont généralement statiques, tandis que les données utilisées dans un cadre de production sont dynamiques et en constante évolution.
- Les environnements de recherche et de production se préoccupent de l'équité, mais les implications de l'équité en environnement de production sont souvent plus critiques en raison des conséquences réelles.

Conclusion

- Pipeline de l'apprentissage automatique : Des données au modèle
- Focus sur :
 - la préparation des données
 - le feature engineering
 - l'entraînement du modèle
 - l'évaluation du modèle

Travaux Pratiques

- Vous allez vous exercer à implémenter ces différentes étapes sur plusieurs problèmes ML!