

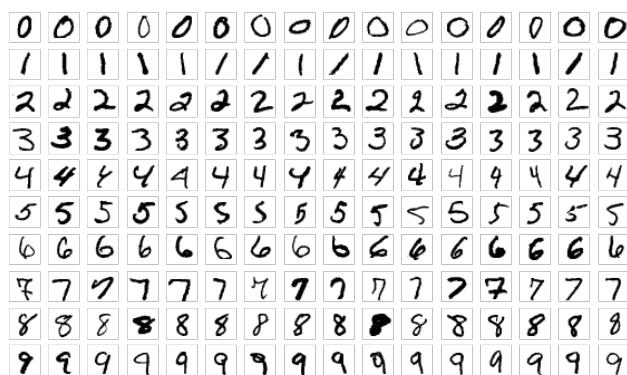
Document 1 : algorithme des k plus proches voisins (KNN)

L'algorithme des k plus proches voisins est un algorithme d'apprentissage machine **supervisé**. Il vise à classifier une donnée, à trouver l'étiquette d'une donnée.

L'algorithme est dit supervisé car il nécessite une intervention extérieure pour réaliser sa tâche. Un ensemble de données doit en effet être étiqueté au préalable pour permettre à l'algorithme d'opérer.

Le principe de l'algorithme est de calculer d'une façon ou d'une autre la distance entre la nouvelle donnée et chacune des données d'apprentissage, il classe ensuite ces distances en ordre croissant et enfin, il attribue l'étiquette la plus donnée (le mode) parmi les k données les plus proches (d'où son nom).

Exemple d'utilisation : on donne à l'algorithme un ensemble d'images de chiffres manuscrits et on lui indique le chiffre réellement représenté (l'étiquette) sur chaque image. Charge ensuite à l'algorithme de déterminer le chiffre écrit sur une nouvelle image manuscrite fournie.



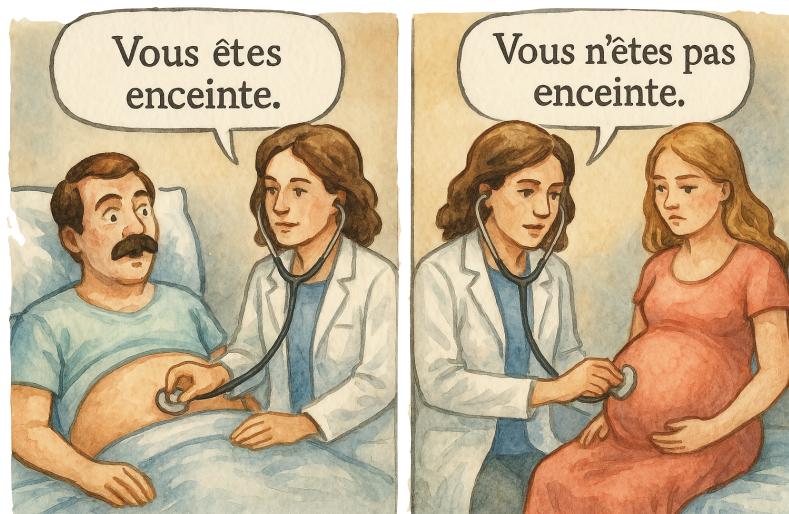
Document 2 : précision, sensibilité et exactitude

Supposons que l'on spécialise notre algorithme à identifier les 3. Il se borne alors juste à déterminer si l'image soumise est un 3 ou non. Quatre cas de figure peuvent se présenter :

- vrai positif (**VP**) : un 3 est reconnu comme un 3 par l'algo ;
- faux positif (**FP**) : un autre chiffre est reconnu comme un 3 ;
- vrai négatif (**VN**) : un autre chiffre n'est pas reconnu comme un 3 ;
- faux négatif (**FN**) : un 3 n'est pas reconnu comme un 3.

On définit alors trois grandeurs permettant d'évaluer la qualité de la prédiction :

- La **précision** : proportion de données bien prédites parmi les prédictions positives.
- La **sensibilité** : proportion de données bien prédites parmi les données positives (les vrais 3).
- L'**exactitude** : proportion de données bien prédites parmi l'ensemble des données.



- Quel cas de figure est illustré par chacune des images ?

Pour mettre de l'ordre, on a intérêt à utiliser un **tableau de contingence** (ou matrice de confusion).

		Réalité	
		il s'agit d'un 3	il ne s'agit pas d'un 3
Prédiction	un 3 est prédit		
	un 3 n'est pas prédict		

Désignons par **VP** l'effectif des vrais positifs, **FP** celui des faux positifs, **VN** celui des vrais négatifs et **FN** celui des faux négatifs.

- Placer **VP**, **FP**, **VN** et **FN** dans le tableau.
- Construire la grandeur précision à partir de ces effectifs.
Sur quelle partie du tableau se concentre-t-on alors ?
- Construire la grandeur sensibilité à partir de ces effectifs.
Sur quelle partie du tableau se concentre-t-on alors ?
- Construire la grandeur exactitude à partir de ces effectifs.

Un chef de projet vous demande d'obtenir la meilleure sensibilité possible sans plus de détail.

- Comment pourriez-vous lui fournir un algorithme très simple qui réponde à son cahier des charges trop limité ?
- Pourquoi augmenter la précision se fait souvent au détriment de la sensibilité ?