

Lecture_8

Sélection de caractéristiques(Feature

All Features



Feature Selection



Final Features



Selection)

Qu'est-ce que la sélection de caractéristiques ?

La sélection de caractéristiques est une méthode qui consiste à réduire le nombre de variables d'entrée d'un modèle en utilisant uniquement les données pertinentes et en éliminant le bruit présent dans les données. Il s'agit d'un processus automatisé de sélection des caractéristiques pertinentes pour votre modèle d'apprentissage automatique.

Nécessité de la sélection de caractéristiques

Students Dataset

Name	Marks	Address	Race	Religion	Attendance

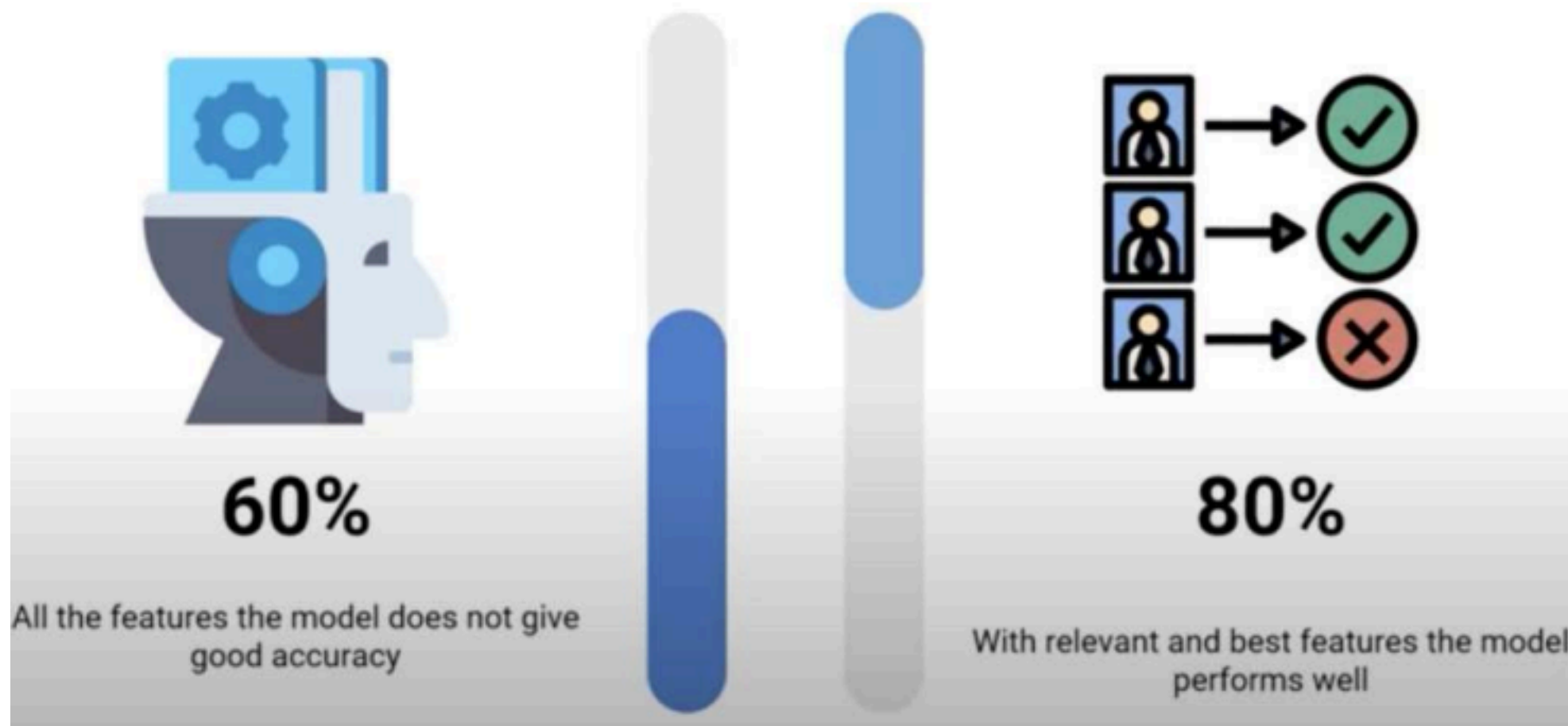
Nécessité de la sélection de caractéristiques

Students Dataset

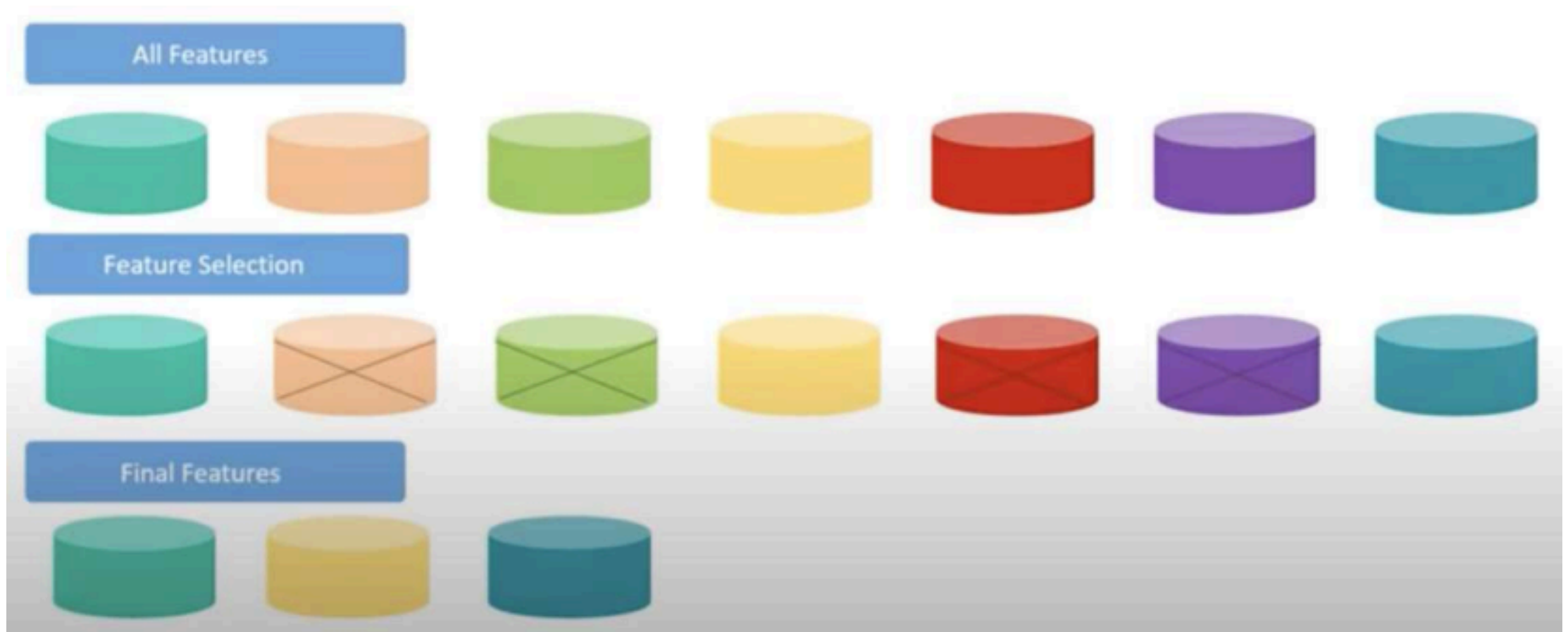
Name	Marks	Address	Race	Religion	Attendance

NB: On peut remarquer que la race et la religion sont des caractéristiques non pertinentes.

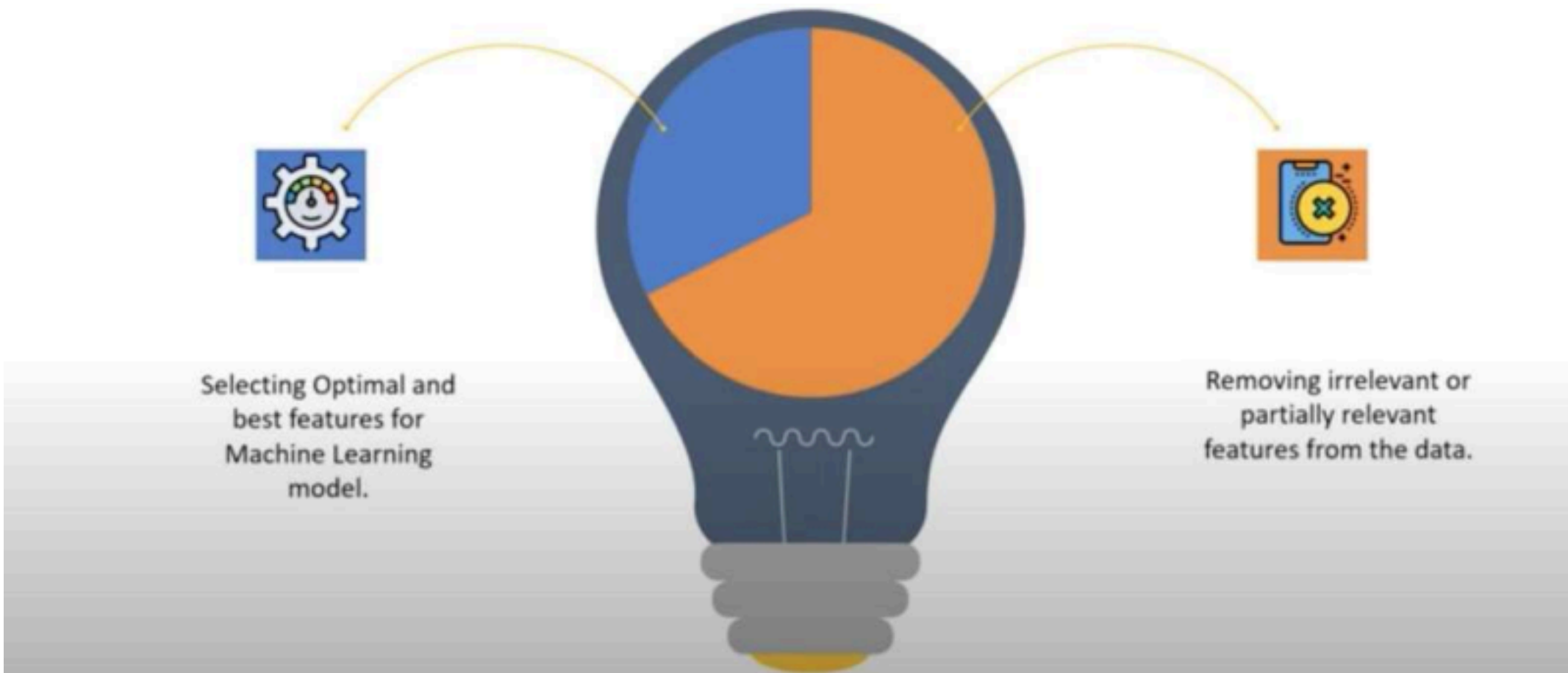
Remarque



Caractéristiques Pertinentes



What is feature Selction



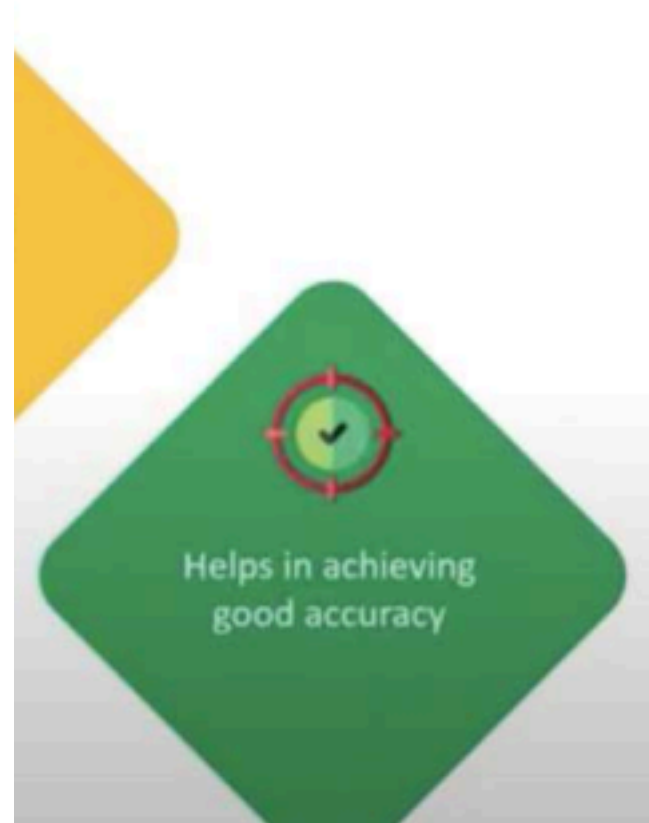
Pourquoi la sélection de caractéristiques est essentielle?

Simplifier le modèle
en réduisant le
nombre de variables

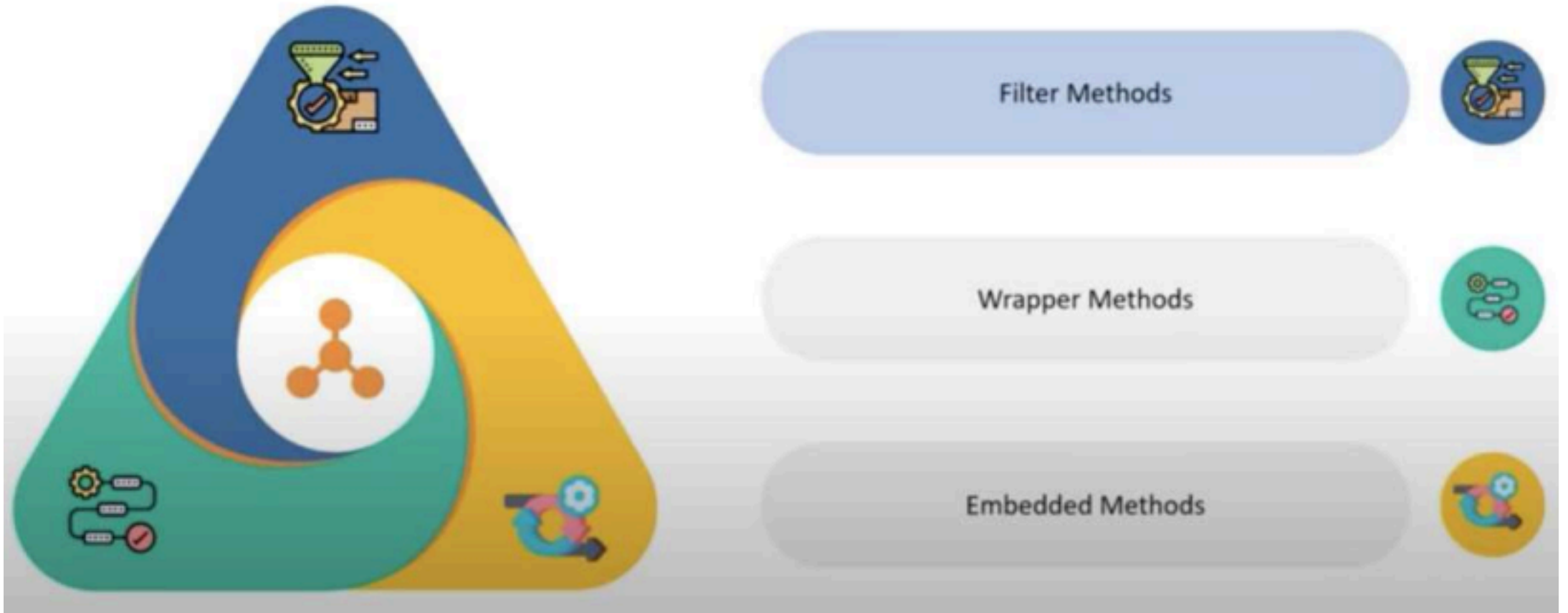
aide à
apprendre
le modèle

aide à obtenir
une bonne
précision

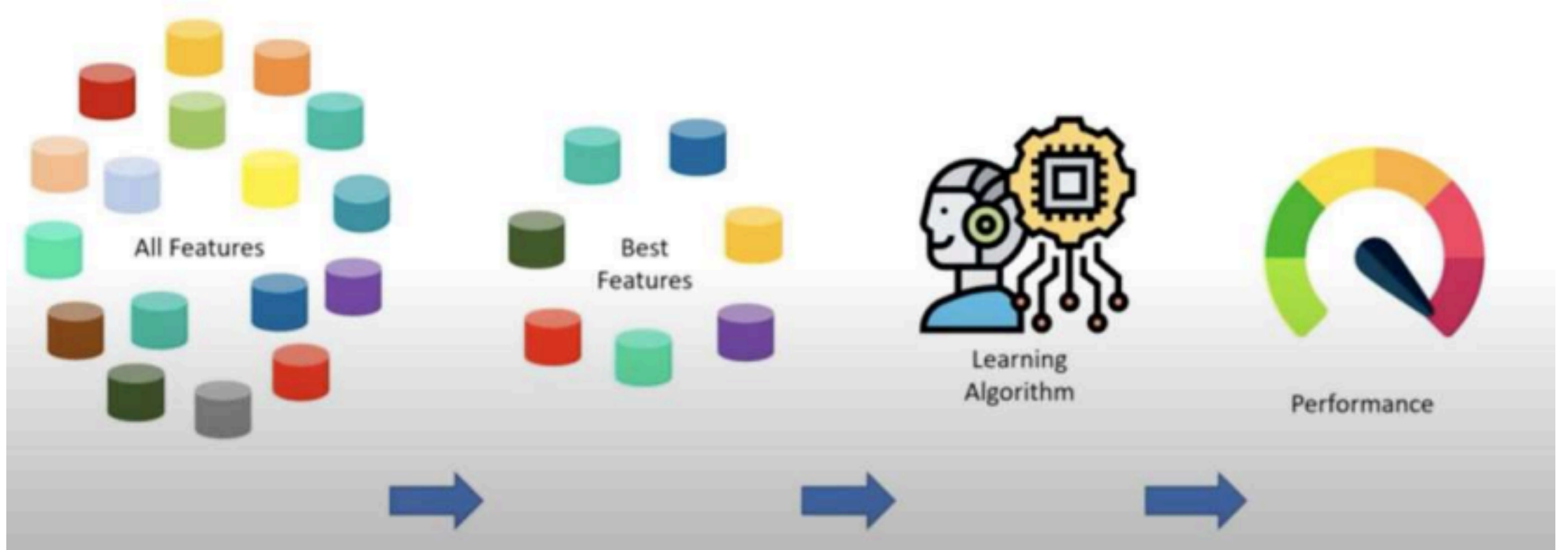
minimise le
coût de calcul



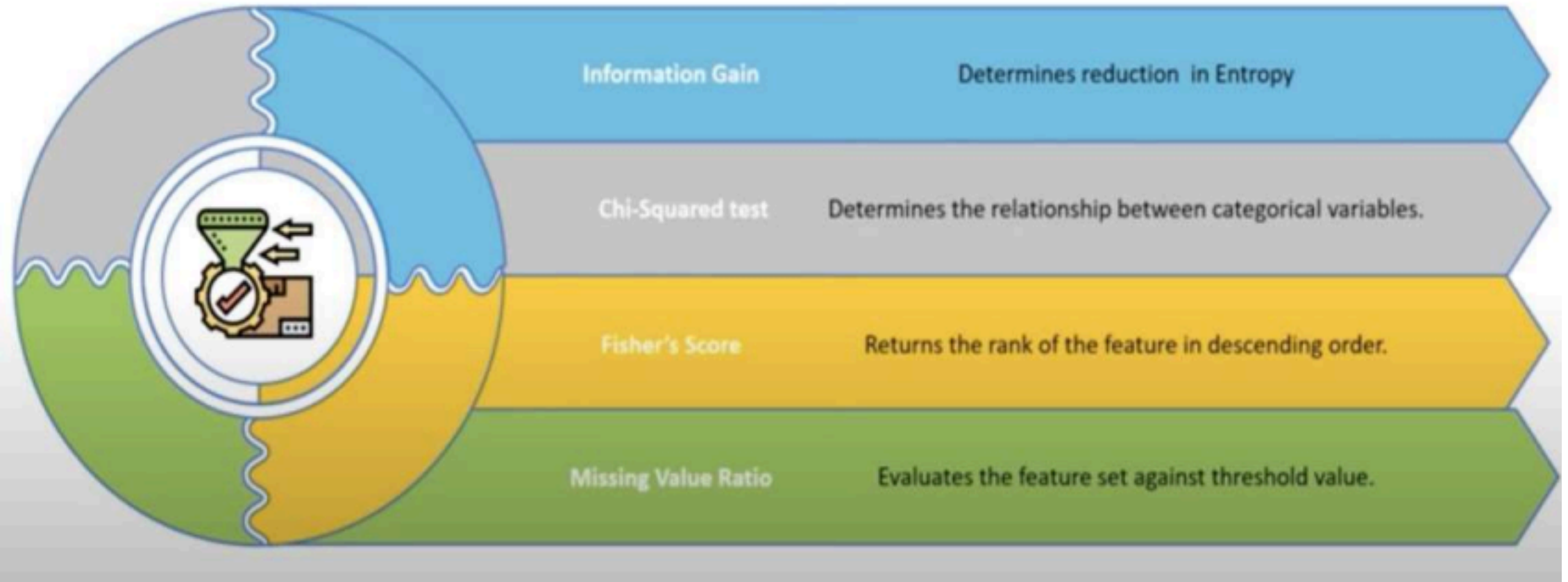
Feature Selection Techniques



Filter Methods Techniques



Filter Method Techniques



Example

1. Correlation Technique
2. Variance Threshold Technique

Lecture_9

Confusion matrix in Machine learning

Definition

Une matrice de confusion, également connue sous le nom de matrice d'erreur, est un outil utile utilisé dans l'apprentissage automatique pour évaluer les performances

d'un modèle de classification. Il s'agit essentiellement d'un tableau qui résume le nombre de prédictions correctes et incorrectes faites par le modèle pour un ensemble de données de test donné. Cela aide les chercheurs et les praticiens à comprendre les performances du modèle et à identifier les domaines à améliorer.

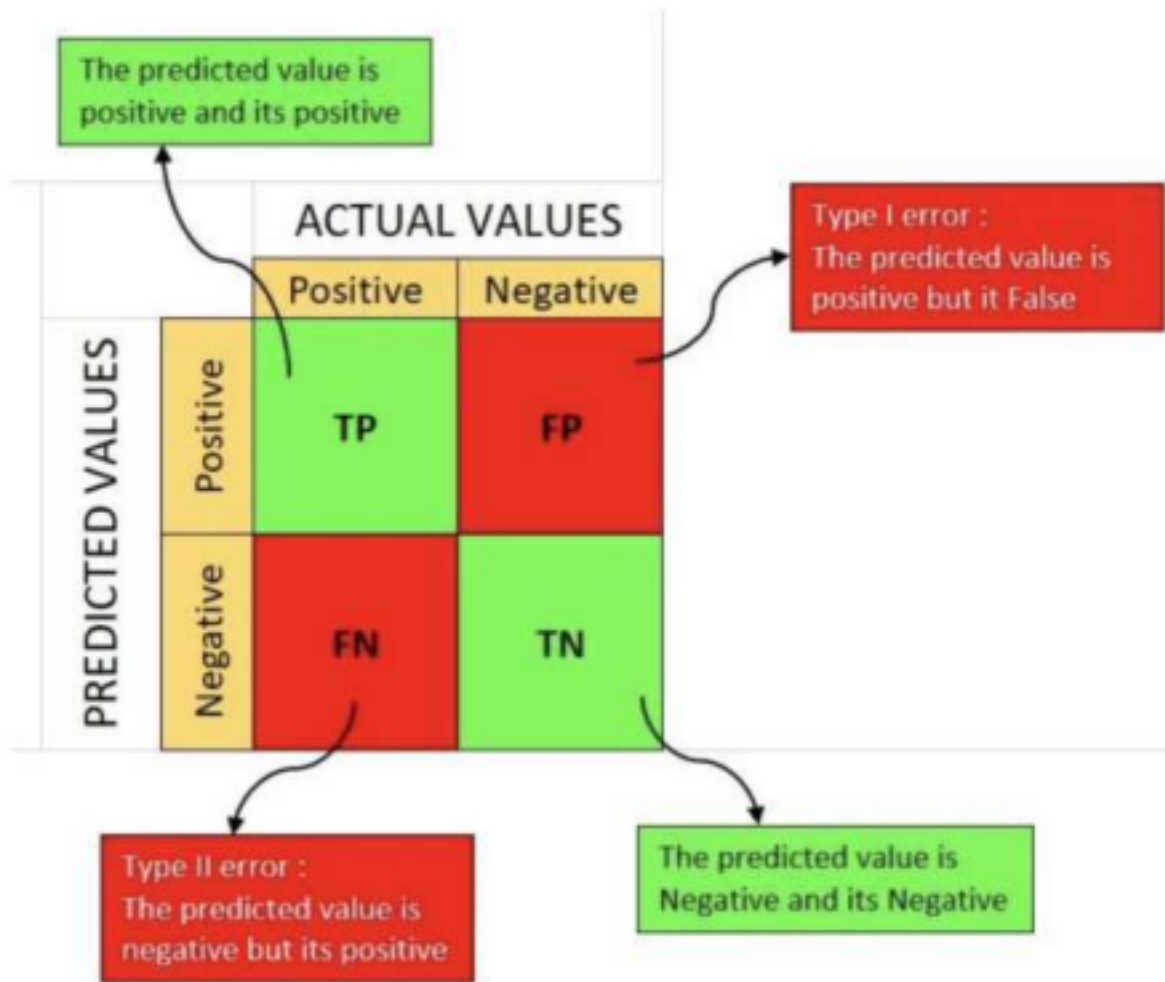
Remark

- **Vrais positifs (VP)** : Il s'agit des cas où le modèle a correctement prédit un résultat positif. •
- **Vrais négatifs (VN)** : Il s'agit des cas où le modèle a correctement prédit un résultat négatif. •
- **Faux positifs (FP)** : Il s'agit des cas où le modèle a prédit un résultat positif de manière incorrecte (également connu sous le nom d'erreur de type I).
- **Faux négatifs (FN)** : Il s'agit des cas où le modèle a prédit un résultat négatif de manière incorrecte (également connu sous le nom d'erreur de type II).
- **True positives (TP)**: These are instances where the model correctly predicted a positive outcome. •
- **True negatives (TN)**: These are instances where the model correctly predicted a negative outcome. •

False positives (FP): These are instances where the model incorrectly predicted a positive outcome (also known as Type I error).

- **False negatives (FN):** These are instances where the model incorrectly predicted a negative outcome (also known as Type II error).

Example



Une matrice de confusion est une matrice de taille $N \times N$ utilisée pour évaluer les performances d'un modèle de classification, où

	Predicted	
	Negative (N) -	Positive (P) +

N représente le nombre de classes cibles. Cette matrice compare les valeurs cibles réelles avec celles prédites par le modèle d'apprentissage automatique.

Attention:

1. Un bon modèle est celui qui a des taux de VP et de VN élevés, et des taux de FP et de FN faibles.





2. Si vous travaillez avec de données déséquilibré, il est toujours préférable d'utiliser la matrice de confusion comme critère d'évaluation de votre modèle d'apprentissage automatique.

Understanding Confusion Matrix in an easier way
Comprendre la matrice de confusion de manière plus simple

Nous avons un total de 20 chats et chiens et notre modèle prédit s'il s'agit d'un chat ou non.

Actual values = ['dog', 'cat', 'dog', 'cat', 'dog', 'dog', 'cat', 'dog', 'cat', 'dog', 'dog', 'dog', 'dog', 'cat', 'dog', 'dog',
'cat', 'dog', 'dog', 'cat']





Predicted values = ['dog', 'dog', 'dog', 'cat', 'dog', 'dog', 'cat', 'cat', 'cat', 'cat', 'dog', 'dog', 'dog', 'cat', 'dog', 'dog',

		PREDICTED VALUES	
		Positive (CAT)	Negative (DOG)
ACTUAL VALUES	Positive (CAT)	 TRUE POSITIVE 6 YOU ARE A CAT	 FALSE NEGATIVE 1 TYPE II ERROR YOU ARE A DOG
	Negative (DOG)	 FALSE POSITIVE 2 TYPE I ERROR YOU ARE A CAT	 TRUE NEGATIVE 11 YOU ARE NOT A CAT

'cat', 'dog', 'dog', 'cat']

Details

- **True Positive (TP) = 6**
- You predicted positive and it's true. You predicted that an animal is a cat and it actually is.
- **True Negative (TN) = 11**
- You predicted negative and it's true. You predicted that animal is not a cat and it actually is not (it's a dog).
- **False Positive (Type 1 Error) (FP) = 2**
- You predicted positive and it's false. You predicted that animal is a cat but it actually is not (it's a dog).
- **False Negative (Type 2 Error) (FN) = 1**

		PREDICTED VALUES	
		Positive (CAT)	Negative (DOG)
ACTUAL VALUES	Positive (CAT)	 <p>TRUE POSITIVE</p> <p>6</p> <p>YOU ARE A CAT</p>	 <p>FALSE NEGATIVE</p> <p>1</p> <p>TYPE II ERROR</p> <p>YOU ARE A DOG</p>
	Negative (DOG)	 <p>FALSE POSITIVE</p> <p>2</p> <p>TYPE I ERROR</p> <p>YOU ARE A CAT</p>	 <p>TRUE NEGATIVE</p> <p>11</p> <p>YOU ARE NOT A CAT</p>

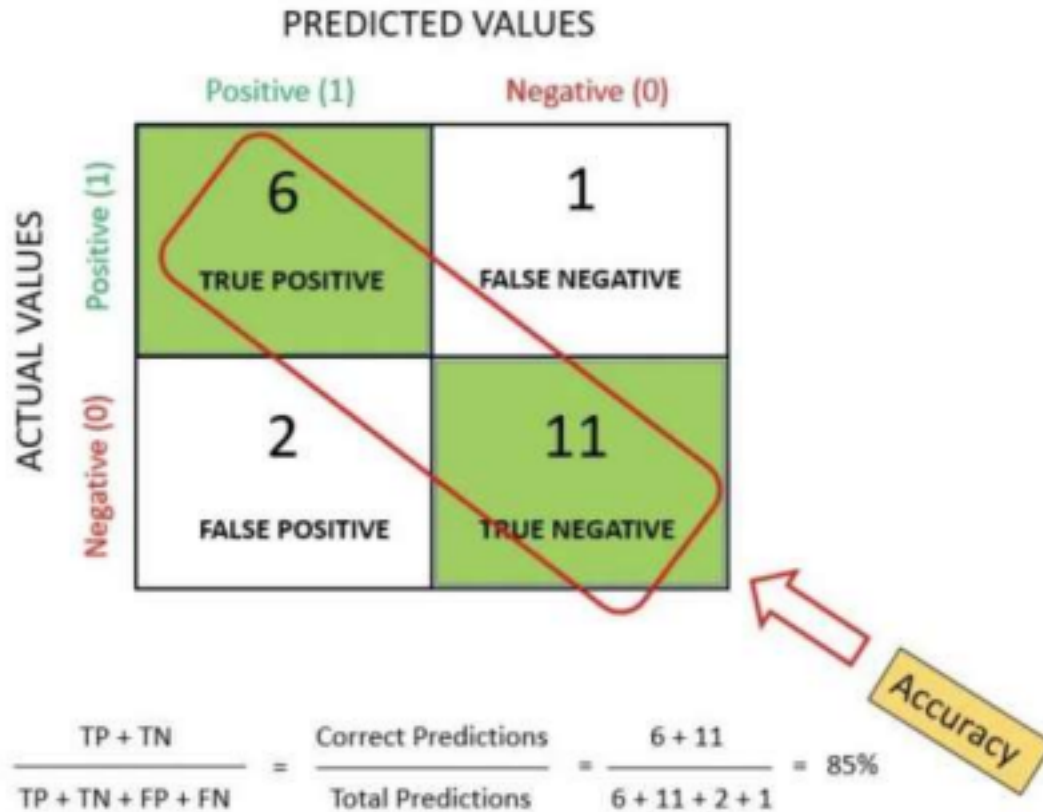
Classification Measure

Fondamentalement, il s'agit d'une version étendue de la matrice de confusion. Il existe des mesures autres que la matrice de confusion qui peuvent aider à mieux comprendre et analyser notre modèle et ses performances.

1. Accuracy
2. Precision
3. Recall (TPR, Sensitivity)
4. F1-Score

Accuracy

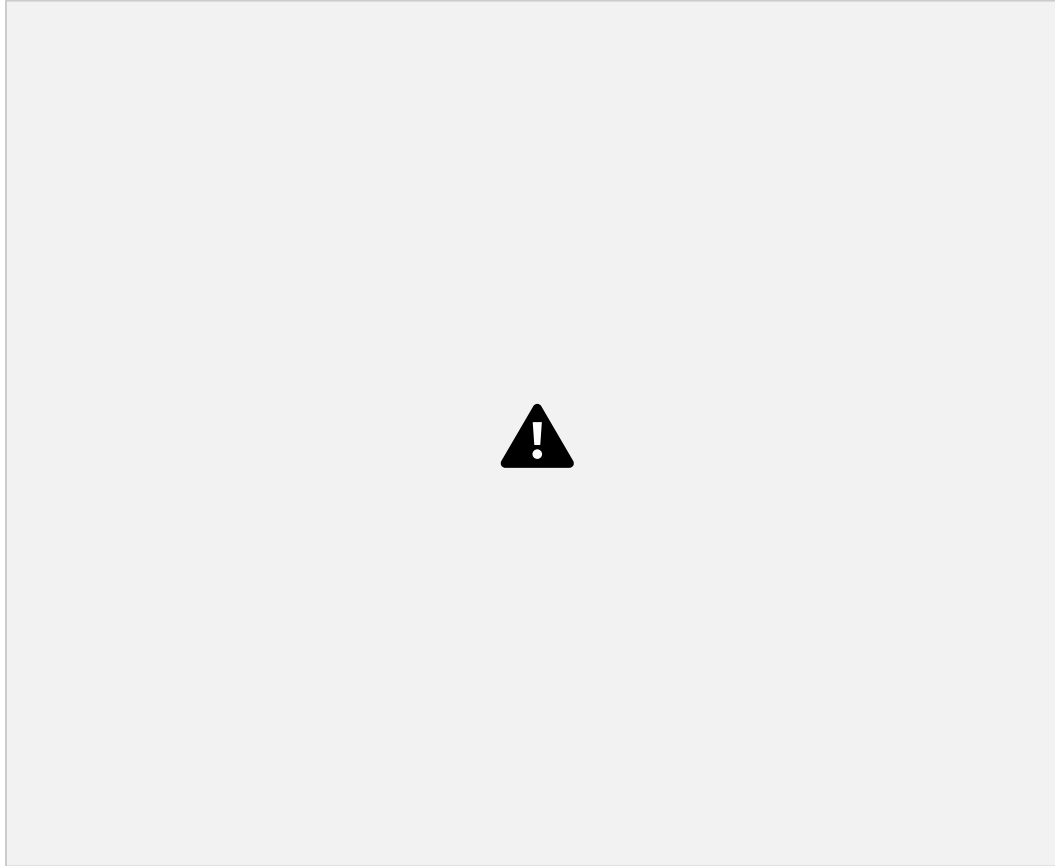
La précision mesure simplement la fréquence à laquelle le classificateur effectue la prédiction correcte. C'est le rapport entre le nombre de prédictions correctes et le nombre total de prédictions.



Precision

Il s'agit d'une mesure d'exactitude obtenue dans une véritable prédiction. En termes simples, cela nous indique combien de prédictions sont réellement positives sur

l'ensemble des prédictions positives totales.



Recall

Il s'agit d'une mesure des observations réelles qui sont prédites correctement, c'est-à-dire du nombre d'observations de classe positive qui sont réellement prédites comme positives. On l'appelle également sensibilité.

Exemple:

Le rappel est important dans les cas médicaux où peu importe que nous déclenchions une fausse alerte, mais les cas positifs réels ne doivent pas passer inaperçus !



F-measure / F1-Score

Le score F1 est un nombre compris entre 0 et 1 et constitue la moyenne harmonique de précision et de rappel. Nous utilisons la moyenne harmonique car elle n'est pas sensible aux valeurs extrêmement grandes, contrairement aux moyennes simples.



NB: En pratique, lorsque l'on cherche à augmenter la précision de notre modèle, le rappel diminue et vice versa. Le score F1 capture les deux tendances en une seule valeur.

Le score F1 est une moyenne harmonique de précision et de rappel. Par rapport à la moyenne arithmétique, la moyenne harmonique punit davantage les valeurs extrêmes. Le score F doit être élevé (idéalement 1).

Exemple