Machine Learning

Anis Cherfi

Classification

Classification

Définition:

La classification permet de prédire si un élément est membre d'un groupe ou d'une catégorie donnée.



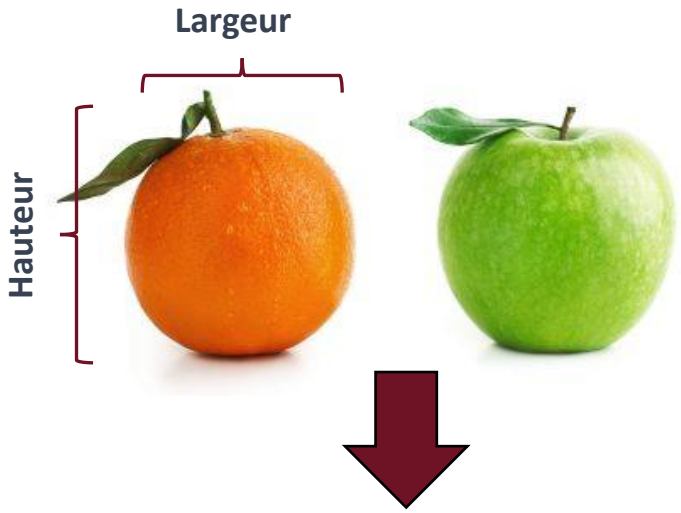
Classes:

- Identification de groupes avec des profils particuliers.
- Possibilité de décider de l'appartenance d'une entité à une classe.

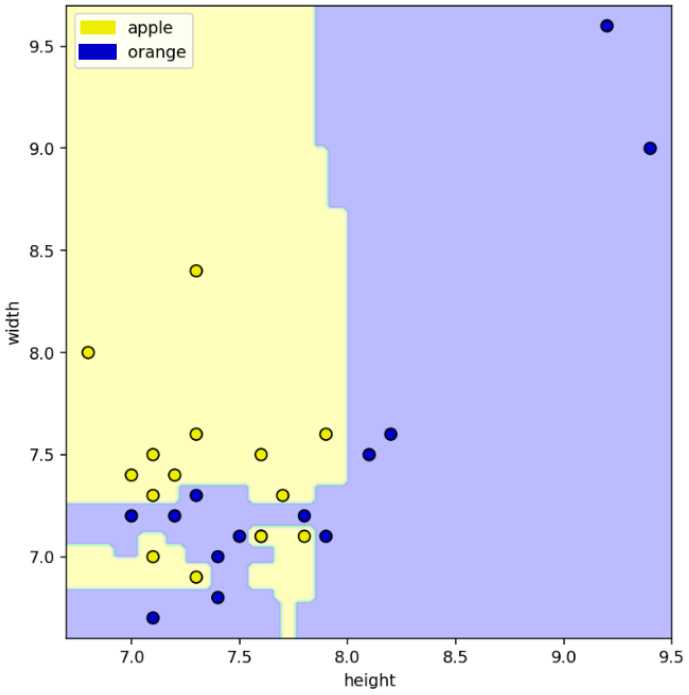
Caractéristiques de classification:

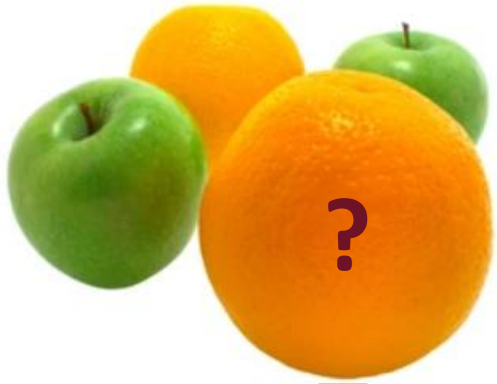
- Apprentissage supervisé: classes connues à l'avance.
- Qualité de la classification (taux d'erreur)

Exemple: Classification



fruit_name	fruit_subtype	mass	width	height	color_score
apple	granny_smith	192	8.4	7.3	0.55
apple	granny_smith	180	8.0	6.8	0.59
apple	granny_smith	176	7.4	7.2	0.60
apple	braeburn	178	7.1	7.8	0.92
apple	braeburn	172	7.4	7.0	0.89
orange	spanish_jumbo	342	9.0	9.4	0.75
orange	spanish_jumbo	356	9.2	9.2	0.75
orange	spanish_jumbo	362	9.6	9.2	0.74







MODEL

orange

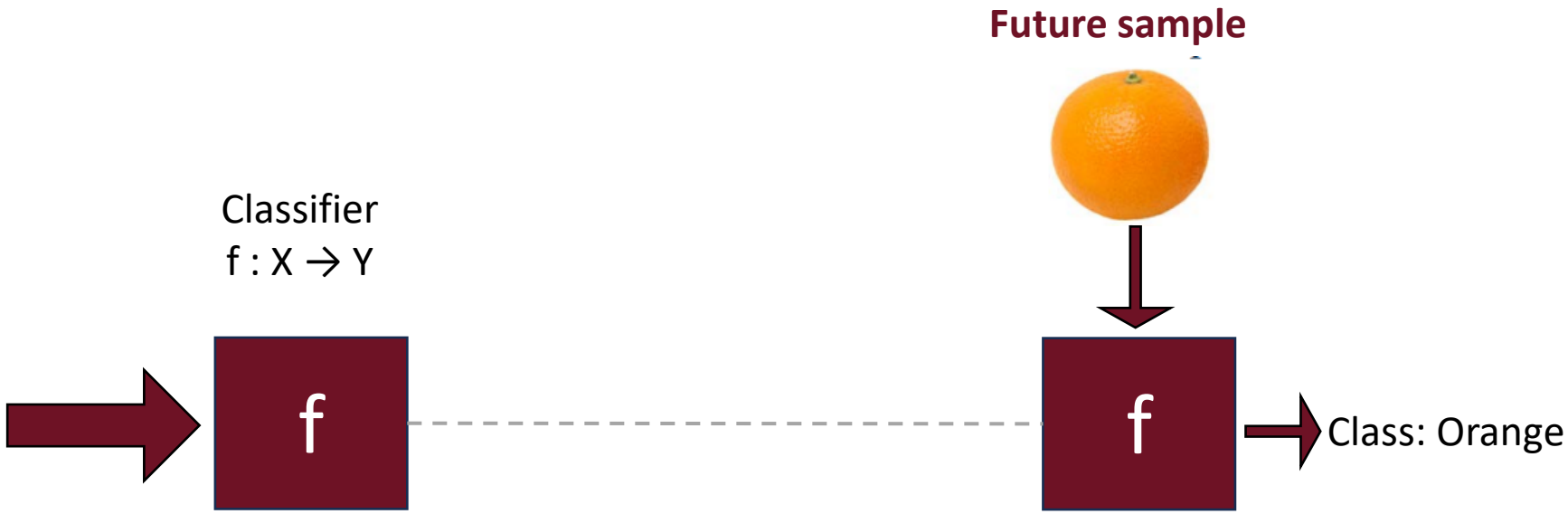
Processus de classification

Etape 1:

- Construction du modèle à partir de l'ensemble d'apprentissage (training set).

Training set

Instance X	Y (class, label)
X1 	y1: Apple
X2 	y2: Apple
X3 	Y3: Orange
X4 	Y4: Orange



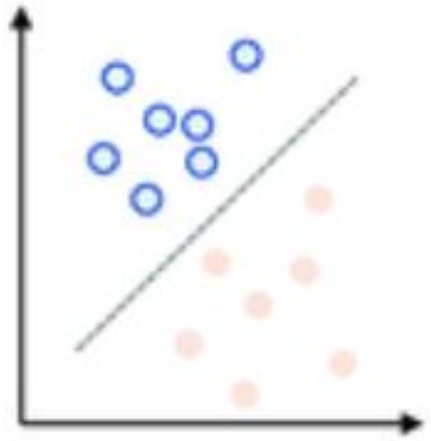
Etape 2:

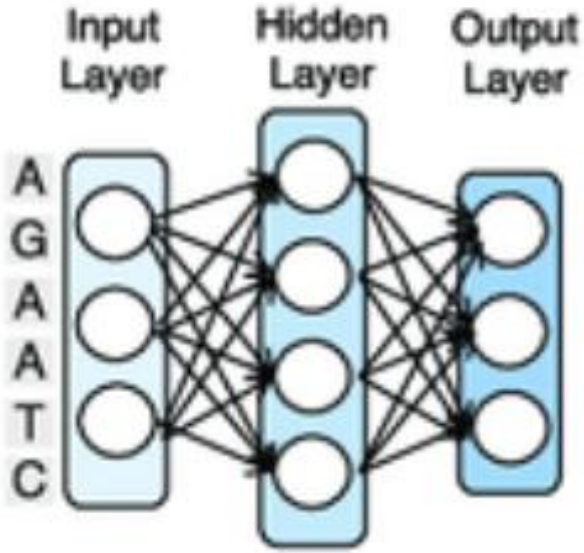
- Utilisation du modèle: tester la précision du modèle (test set) et l'utiliser dans la classification de futur donnée(nouvelles données)

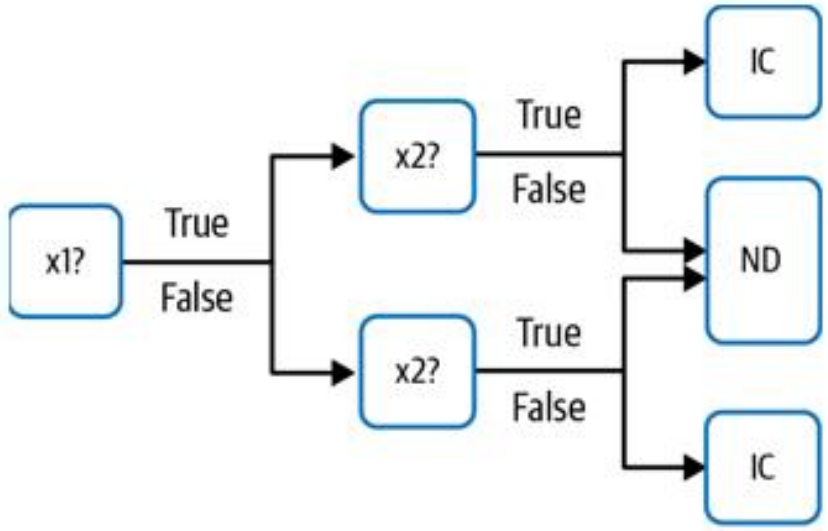
Etape 1: Construction de modèle

Chaque instance est affectée à une classe selon ces valeurs.

1. La classe d'une instance est déterminée par l'attribut classe.
2. L'ensemble des instances d'apprentissage (train set) est utilisé dans la construction du modèle (entrainement).
3. Le modèle est représenté par des règles de classification (Algorithme d'apprentissage)





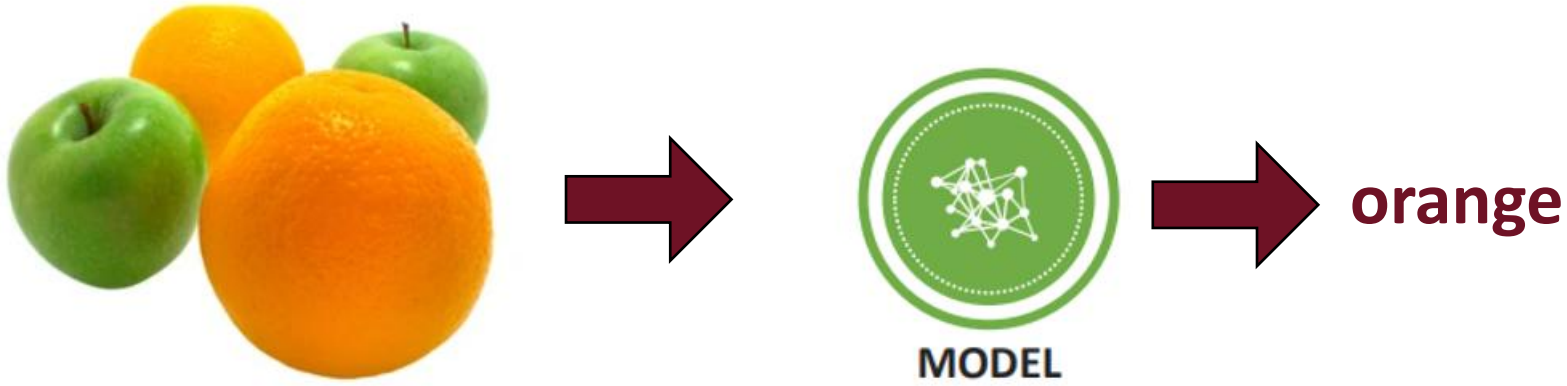


Etape 2: Utilisation du modèle

Classification de nouvelles instances ou données inconnues

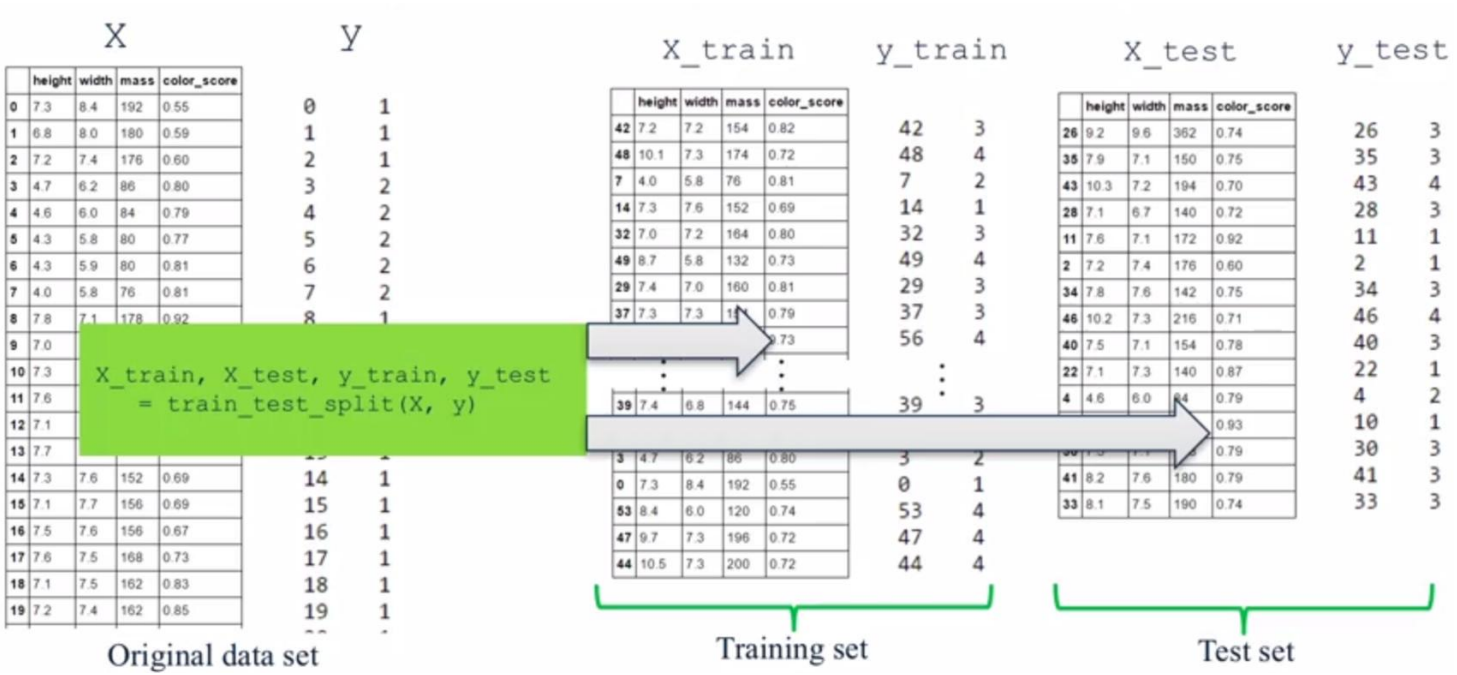
Estimer le taux d'erreur du modèle

- La classe connue d'une instance de test est comparée avec le résultat du modèle.
- Taux d'erreur = pourcentage de tests incorrectement classés par le modèle.



Méthodes d'évaluation

En apprentissage automatique, nous divisons généralement nos données en deux sous-ensembles : les données d'entraînement et les données de test. Nous adaptons notre modèle aux données d'entraînement, pour ensuite faire les prévisions sur les données de test.

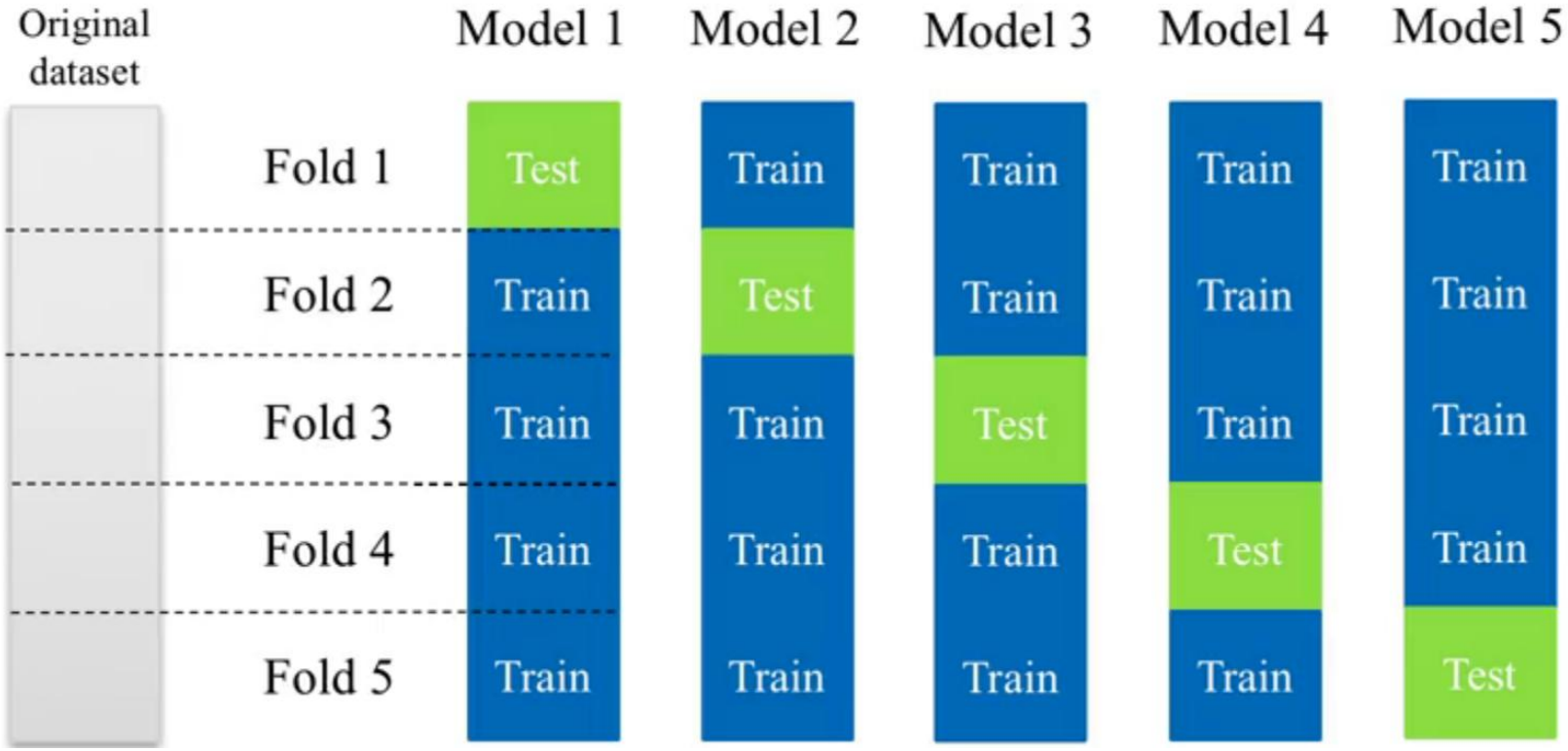


Mais cette méthode n'est pas très fiable et peut causer un risque dans le cas où une classe de nos données sera exclu ou qu'une classe soit beaucoup mieux représentée que les autres, il en résultera un sur-ajustement des données.

Le modèle sera très performant sur les données d'apprentissage (train set), mais probablement très inexact sur de nouvelles instances.

Méthodes d'évaluation: Cross validation

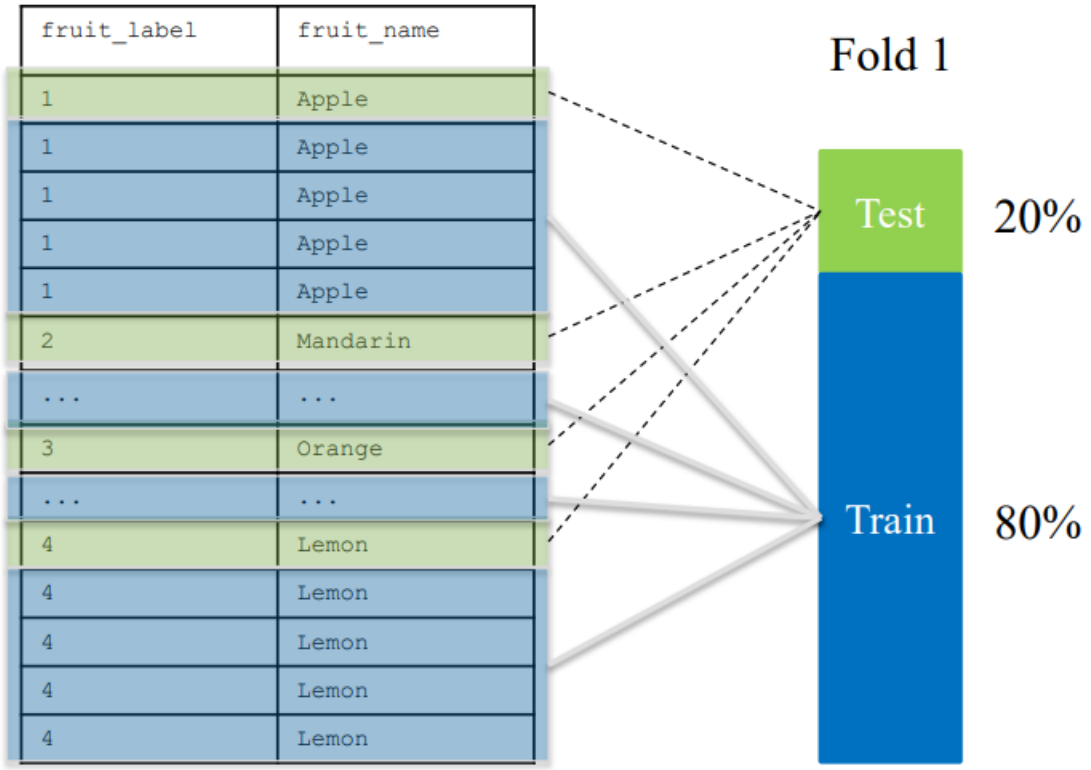
La validation croisée est une bonne technique d 'évaluation de la performance en généralisation d'un classificateur. L'idée derrière cette technique est de tester le modèle du classificateur obtenu sur des données qui n'ont pas participé à l'apprentissage, afin de permettre de prédire le comportement du classificateur face aux nouvelles données.



sklearn.model_selection.Kfold

Méthodes d'évaluation: Stratified Cross-validation

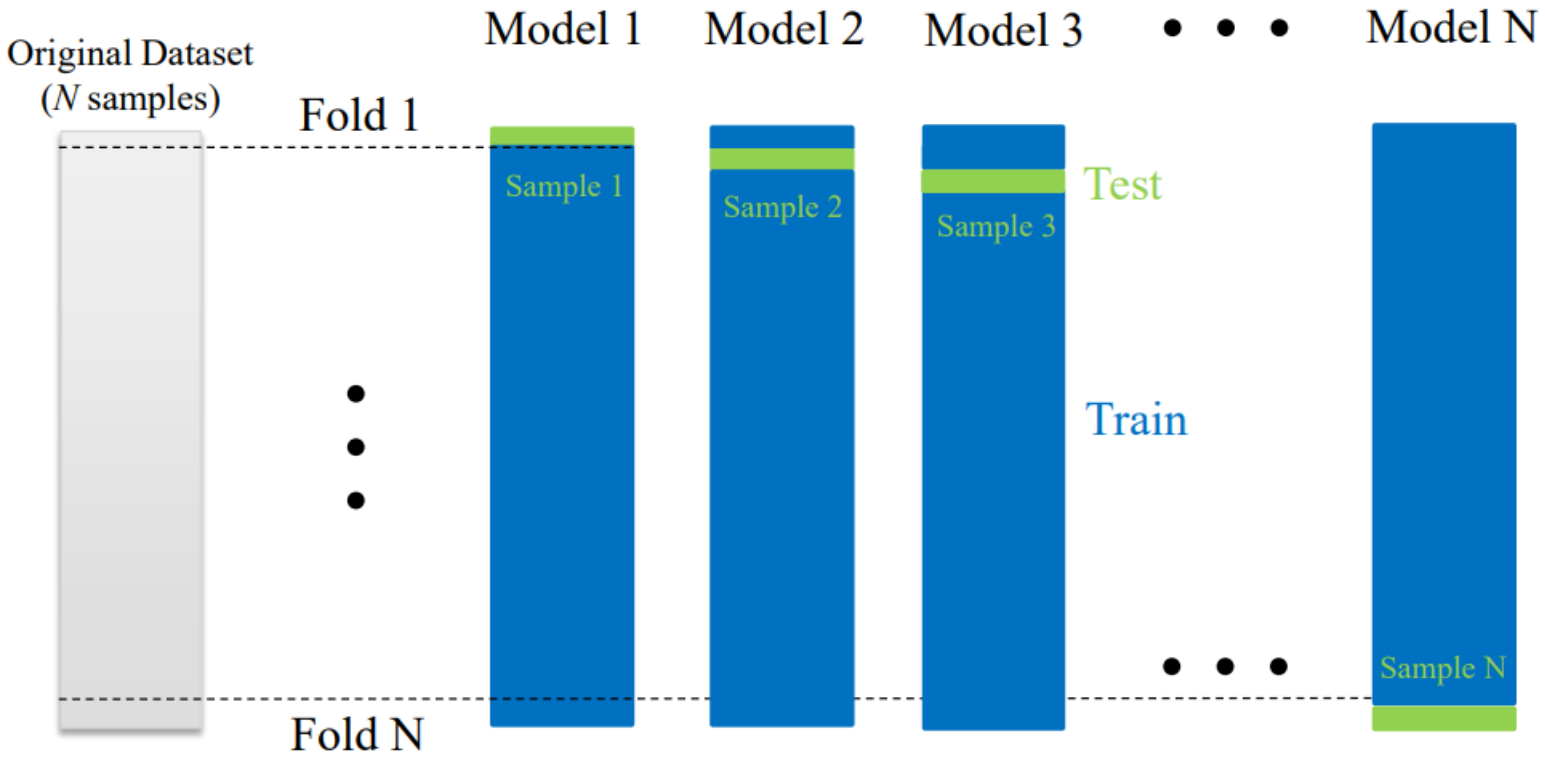
La validation croisée est une bonne technique d 'évaluation de la performance en généralisation d'un classificateur. L'idée derrière cette technique est de tester le modèle du classificateur obtenu sur des données qui n'ont pas participé à l'apprentissage, afin de permettre de prédire le comportement du classificateur face aux nouvelles données.



sklearn.model_selection.Kfold

Méthodes d'évaluation: Leave-one-out cross-validation

La validation croisée est une bonne technique d 'évaluation de la performance en généralisation d'un classificateur. L'idée derrière cette technique est de tester le modèle du classificateur obtenu sur des données qui n'ont pas participé à l'apprentissage, afin de permettre de prédire le comportement du classificateur face aux nouvelles données.



sklearn.model_selection.Kfold

Métriques d'évaluation

Evaluer les performances d'un modèle de classification est primordial:

- Pour savoir si le modèle est globalement significatif: Mon modèle traduit-il vraiment une causalité ?
- Pour se donner une idée des performances en déploiement : Quelle sera la fiabilité (les couts associés) lorsque j'utiliserai mon modèle ?
- Pour comparer plusieurs modèles candidats: Lequel parmi plusieurs modèles sera le plus performant compte tenu de mes objectifs ?

Remarque:

La mesure et l'évaluation de la performance d'un modèle de classification se fait toujours sur l'échantillon de test: Il faut tester la performance de modèle sur des données qui n'ont pas été utilisées pour construire le modèle de classification..

Métriques d'évaluation

Plusieurs indicateurs permettent de mesurer la performance des modèles de classification.
Chaque indicateur a ses spécificités.
Il faut bien souvent en utiliser plusieurs pour avoir une vision complète de la performance de votre modèle.

Pour évaluer la performance d'un modèle de classification nous présentons quatre indicateurs qui sont calculés A partir de la matrice de confusion:

- L'accuracy.
- Le recall.
- La precision.
- F1 score.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.72      0.75      0.73       162
           1       0.57      0.52      0.54       100

    accuracy                           0.66       262
   macro avg       0.64      0.64      0.64       262
weighted avg       0.66      0.66      0.66       262
```

Exemple de classification

Nous avons une base de données client qui ont été abonnés à un service.

- Des clients qui sont encore abonnés.
- Des clients qui ont résilié le service.

Client	Text	Purchases	Data	Age	Churn?
1	72	0	33.6	50	Abonné
2	66	0	26.6	31	Abonné
3	64	0	23.3	32	Abonné
4	66	94	28.1	21	Abonné
5	0	0	35.3	29	Résilié
6	72	175	25.8	51	Résilié
7	0	0	30	32	Abonné
8	84	230	45.8	31	Abonné
9	110	240	45.4	54	Résilié
10	64	0	27.4	40	Abonné

Exemple de classification

Nous construisons un score de churn : pour chaque client, on prédit s'il va résilier ou conserver son abonnement le mois suivant.

Client	Text	Purchases	Data	Age	Churn?	Prédiction
1	72	0	33.6	50	Abonné	Résilié
2	66	0	26.6	31	Abonné	Résilié
3	64	0	23.3	32	Abonné	Résilié
4	66	94	28.1	21	Abonné	Abonné
5	0	0	35.3	29	Résilié	Abonné
6	72	175	25.8	51	Résilié	Résilié
7	0	0	30	32	Abonné	Abonné
8	84	230	45.8	31	Abonné	Abonné
9	110	240	45.4	54	Résilié	Résilié
10	64	0	27.4	40	Abonné	Abonné

Quelle est la performance de ce score ?
A quel point je peux lui faire confiance pour prédire les résiliations futures?

Exemple de classification

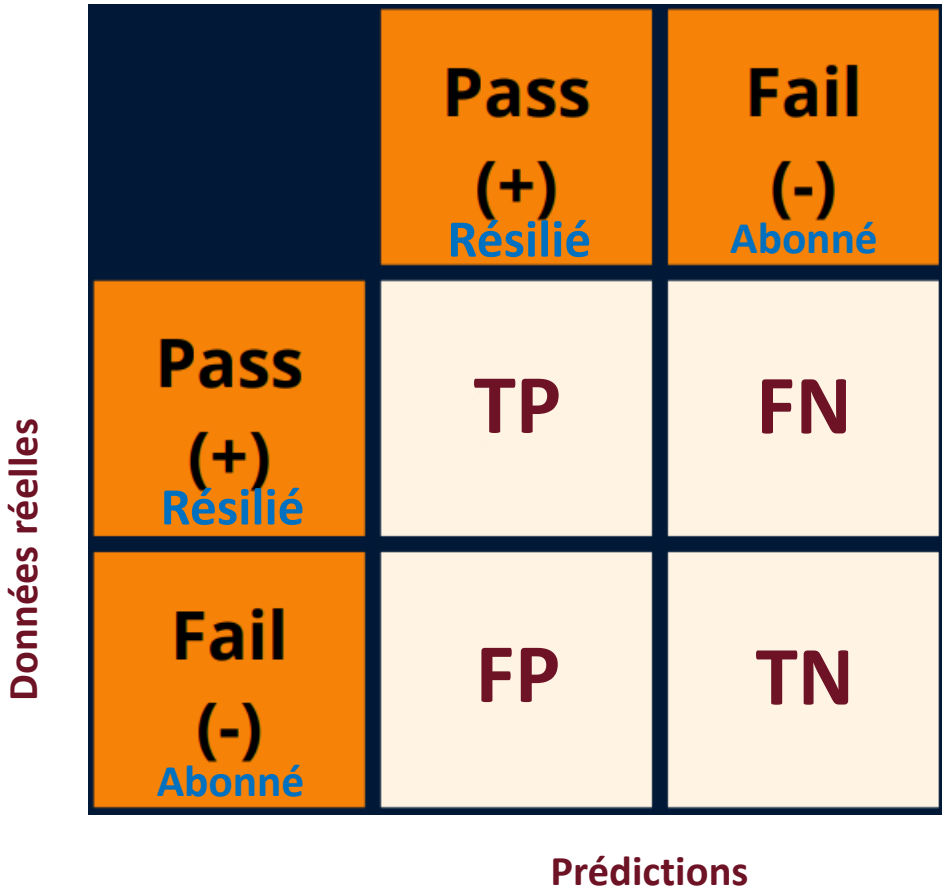
Nous construisons un score de churn : pour chaque client, on prédit s'il va résilier ou conserver son abonnement le mois suivant.

Client	Text	Purchases	Data	Age	Churn?	Prédiction
1	72	0	33.6	50	Abonné	Résilié
2	66	0	26.6	31	Abonné	Résilié
3	64	0	23.3	32	Abonné	Résilié
4	66	94	28.1	21	Abonné	Abonné
5	0	0	35.3	29	Résilié	Abonné
6	72	175	25.8	51	Résilié	Résilié
7	0	0	30	32	Abonné	Abonné
8	84	230	45.8	31	Abonné	Abonné
9	110	240	45.4	54	Résilié	Résilié
10	64	0	27.4	40	Abonné	Abonné

Quelle est la performance de ce score ?
A quel point je peux lui faire confiance pour prédire les résiliations futures?

Matrice de confusion

Une matrice de confusion sert à évaluer la qualité d'une classification. Elle est obtenue en comparant les données classées avec des données de référence (test set) qui doivent être différentes de celles ayant servi à réaliser la classification (train set).



Matrice de confusion

Les fausses prédictions:

- **Nombre de faux négatifs (FN):** les clients qui ont résilié mais pour lesquels le score a prédit à tort qu'ils allaient rester abonnés.
- **Nombre de faux positifs (FP):** les clients qui sont restés abonnés alors que le score a prédit à tort qu'ils allaient résilier.

Les bonnes prédictions:

- **Nombre de vrais positifs (VP):** les clients qui ont résilié pour lesquels le score a bien prédit qu'ils allaient résilier.
- **Nombre de vrais négatifs(VN):** les clients qui sont toujours abonnés et pour lesquels l'algorithme a bien prédit qu'ils resteraient abonnés.

Métriques d'évaluation

Définition:

Accuracy: mesure de la proportion d'exemples correctement étiquetés, soit le complémentaire à 1 du taux d'erreur.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Parfois l'accuracy ne suffit pas:
Considérons un problème de 2-classes:

Nombre de Classes 0 égal à 9990
Nombre de Classes 1 égal à 10.
La base de données n'est pas équilibrée.

Si le modèle prédit que tout est de classe 0, la précision est de 9990/10000 = 99, 9%.
La précision est trompeuse car le modèle ne détecte aucun exemple de classe 1.

Métriques d'évaluation

Définition:

Precision: On appelle précision, ou valeur positive prédictive(positive predictive value, PPV) la proportion de prédictions correctes parmi les prédictions positives.

$$Precision = \frac{TP}{TP + FP}$$

La précision permet de répondre à la question suivante:
Quelle proportion d'identifications positives était effectivement correcte ?

Un modèle de classification ne produisant aucun faux positif a une précision de 1, 0.

Métriques d'évaluation

Définition:

Recall: On appelle rappel (recall en anglais), ou sensibilité(sensitivity en anglais), le taux de vrais positifs, c'est-à-dire la proportion d'exemples positifs correctement identifiés

$$Recall = \frac{TP}{TP + FN}$$

Le recall (rappel) permet de répondre à la question suivante :
Quelle proportion de résultats positifs réels a été identifiée correctement ?

Un modèle ne produisant aucun faux négatif a un rappel de 1, 0.

Métriques d'évaluation

- Pour évaluer les performances d’un modèle de façon complète: Il faut analyser à la fois la précision et le rappel.
- La précision et rappel sont fréquemment en tension: l’amélioration de la précision se fait généralement au détriment du rappel et réciproquement.
- Si on veut comparer les performances de deux classificateurs et on a:
    - Supposons que le classificateur A a un recall plus élevé et le classificateur B a une précision plus élevée.

Alors on ne peut pas comparer les classificateurs A et B

Différents outils ont été créés pour évaluer simultanément la précision et le rappel. La F-score en fait partie.

Métriques d'évaluation

Le F1 score combine la précision et le recall d'un classificateur en une seule métrique en prenant leur moyenne harmonique.
Le F1 score est utilisé pour comparer les performances de deux classificateurs dans le cas suivant:

Supposons que le classificateur A a un recall plus élevé et le classificateur B a une précision plus élevée.

Dans ce cas, les F1 score des deux classificateurs peuvent être utilisés pour déterminer celui qui produit les meilleurs résultats.

Définition:

F1 score: On appelle F-mesure(F-score ou F1-score en anglais) la moyenne harmonique de la précision et du rappel. C'est la moyenne pondérée de la précision et du rappel. Par conséquent, ce score prend en compte à la fois les faux positifs et les faux négatifs

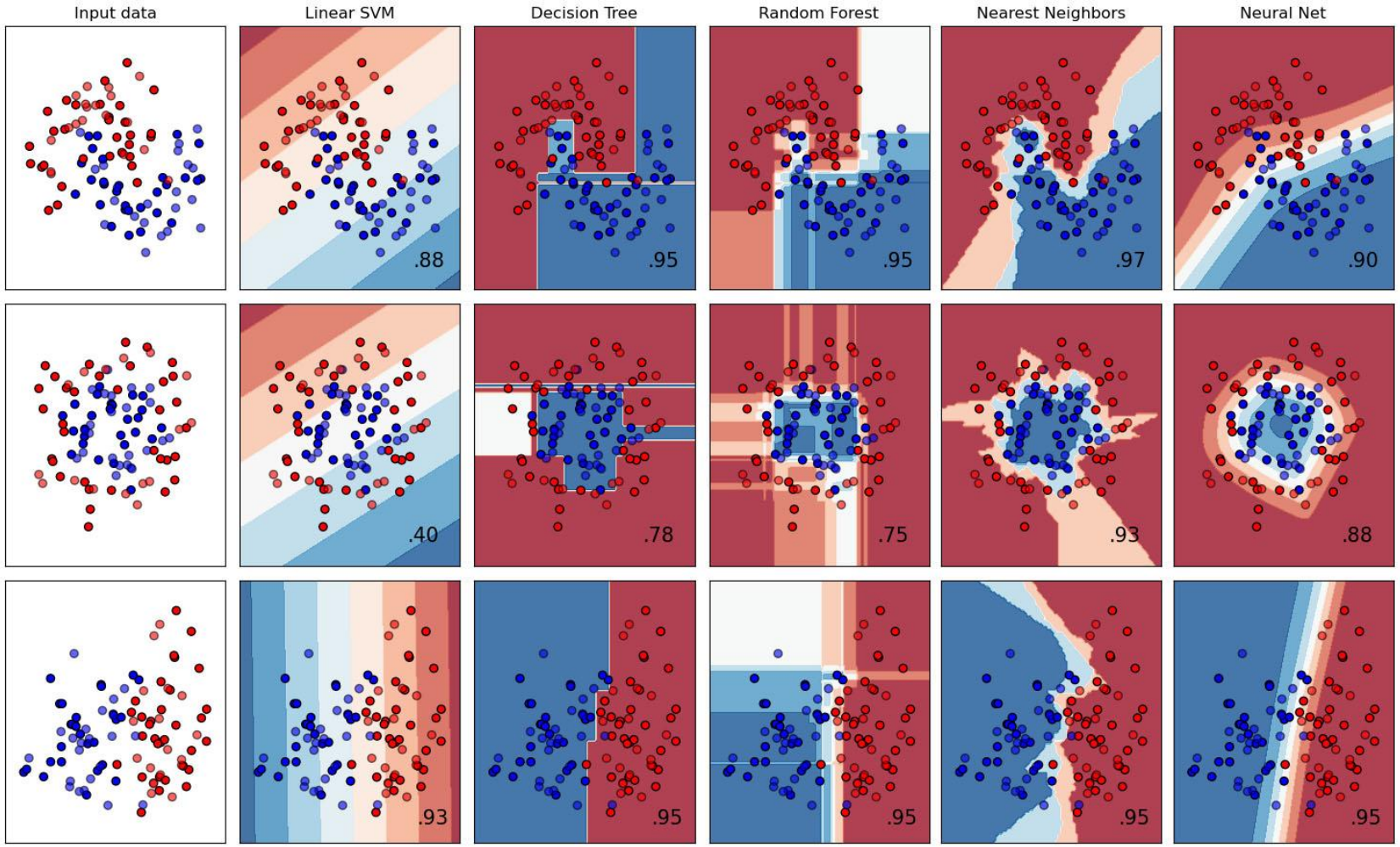
$$F = 2\frac{\text{Précision . Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Méthode de classification

Il existe des méthodes de classification:

- k-Nearest Neighbor (k-NN).
- Arbre de décision.
- Support vector Machine (SVM)
- Naive Bayes
- Régression logistique.

Méthode de classification



Modèles linéaires: séparation linéaire

Modèles d'arbre de décision: séparation en ensemble de rectangles

K plus proches voisins: séparation qui se base sur des mesures de distances

Réseaux de neurones: une séparation non linéaire

Méthode linéaire

	ALGORITHM	DESCRIPTION	APPLICATIONS	ADVANTAGES	DISADVANTAGES
Linear Models	**Linear Regression**	A simple algorithm that models a linear relationship between inputs and a continuous numerical output variable	USE CASES 1. Stock price prediction 2. Predicting housing prices 3. Predicting customer lifetime value	1. Explainable method 2. Interpretable results by its output coefficients 3. Faster to train than other machine learning models	1. Assumes linearity between inputs and output 2. Sensitive to outliers 3. Can underfit with small, high-dimensional data
	Logistic Regression	A simple algorithm that models a linear relationship between inputs and a categorical output (1 or 0)	USE CASES 1. Credit risk score prediction 2. Customer churn prediction	1. Interpretable and explainable 2. Less prone to overfitting when using regularization 3. Applicable for multi-class predictions	1. Assumes linearity between inputs and outputs 2. Can overfit with small, high-dimensional data
	Ridge Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients closer to zero. Can be used for classification or regression	USE CASES 1. Predictive maintenance for automobiles 2. Sales revenue prediction	1. Less prone to overfitting 2. Best suited where data suffer from multicollinearity 3. Explainable & interpretable	1. All the predictors are kept in the final model 2. Doesn't perform feature selection
	Lasso Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients to zero. Can be used for classification or regression	USE CASES 1. Predicting housing prices 2. Predicting clinical outcomes based on health data	1. Less prone to overfitting 2. Can handle high-dimensional data 3. No need for feature selection	1. Can lead to poor interpretability as it can keep highly correlated variables

Méthode de classification

	ALGORITHM	DESCRIPTION	APPLICATIONS	ADVANTAGES	DISADVANTAGES
Tree-Based Models	**Decision Tree**	Decision Tree models make decision rules on the features to produce predictions. It can be used for classification or regression	USE CASES 1. Customer churn prediction 2. Credit score modeling 3. Disease prediction	1. Explainable and interpretable 2. Can handle missing values	1. Prone to overfitting 2. Sensitive to outliers
	Random Forests	An ensemble learning method that combines the output of multiple decision trees	USE CASES 1. Credit score modeling 2. Predicting housing prices	1. Reduces overfitting 2. Higher accuracy compared to other models	1. Training complexity can be high 2. Not very interpretable
	Gradient Boosting Regression	Gradient Boosting Regression employs boosting to make predictive models from an ensemble of weak predictive learners	USE CASES 1. Predicting car emissions 2. Predicting ride hailing fare amount	1. Better accuracy compared to other regression models 2. It can handle multicollinearity 3. It can handle non-linear relationships	1. Sensitive to outliers and can therefore cause overfitting 2. Computationally expensive and has high complexity
	XGBoost	Gradient Boosting algorithm that is efficient & flexible. Can be used for both classification and regression tasks	USE CASES 1. Churn prediction 2. Claims processing in insurance	1. Provides accurate results 2. Captures non linear relationships	1. Hyperparameter tuning can be complex 2. Does not perform well on sparse datasets
	LightGBM Regressor	A gradient boosting framework that is designed to be more efficient than other implementations	USE CASES 1. Predicting flight time for airlines 2. Predicting cholesterol levels based on health data	1. Can handle large amounts of data 2. Computational efficient & fast training speed 3. Low memory usage	1. Can overfit due to leaf-wise splitting and high sensitivity 2. Hyperparameter tuning can be complex