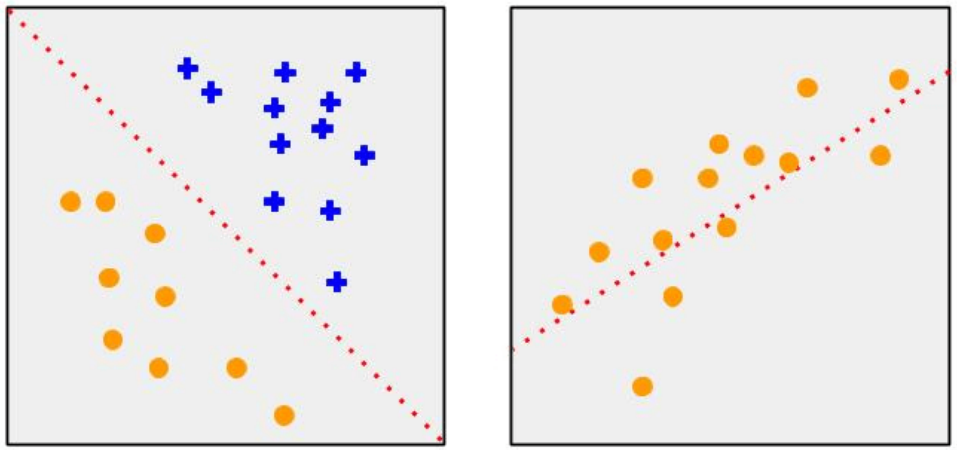
Apprentissage supervisé

Partie 3

Analyse de données et Classification 2
ENSEEIHT - 3ème année Sciences du Numérique

Contact :
Sandrine.Mouysset@irit.fr
sandrine.mouysset@toulouse-inp.fr

2 principaux types d'apprentissage supervisé :



Classification | Regression

Approches par modèles supervisés :

- Arbres de Décision
- Apprentissage d'ensemble : Forêts aléatoires
- Réseaux de neurones
- Ouverture sur le deep learning

Méthodes d'évaluation :

- Validation croisée
- Matrice de confusion
- Précision, Rappel, F-mesure
- Courbe ROC
- Ouverture sur l'explicabilité

Validation croisée : méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage.

- *Testset validation ou holdout method :* on divise l'échantillon de taille n en deux sous-échantillons, le premier dit d'*apprentissage* (communément supérieur à 60 % de l'échantillon) et le second dit de *test*. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'erreur est estimée en calculant un test, une mesure ou un score de performance du modèle sur l'échantillon de test;
- *k-fold cross-validation :*
- *Leave-one-out cross-validation (LOOCV) :*

Validation croisée : méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage.

- *Testset validation ou holdout method :*
- *k-fold cross-validation :* on divise l'échantillon original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation et les $k-1$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode le score de performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $k-1$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction ;
- *Leave-one-out cross-validation (LOOCV) :* cas particulier de la deuxième méthode où $k = n$, c'est-à-dire que l'on apprend sur $n-1$ observations puis on valide le modèle sur la énième observation et l'on répète cette opération n fois.

Matrice de confusion : consiste à compter le nombre de fois où des observations de la classe A ont été rangées dans la classe B.

- Chaque ligne de la matrice de confusion représente la *classe réelle* tandis que chaque colonne représente une *classe prédite*.
- Les éléments hors diagonaux représenteront les erreurs de classification. *Exemple : si on souhaite connaître le nombre de fois où le classifieur a pris des 5 pour des 3, on regardera l'élément hors diagonal (5,3) dans la matrice de confusion.*
- Les éléments diagonaux représenteront le nombre d'éléments bien classés. *Classification parfaite = matrice de confusion diagonale*

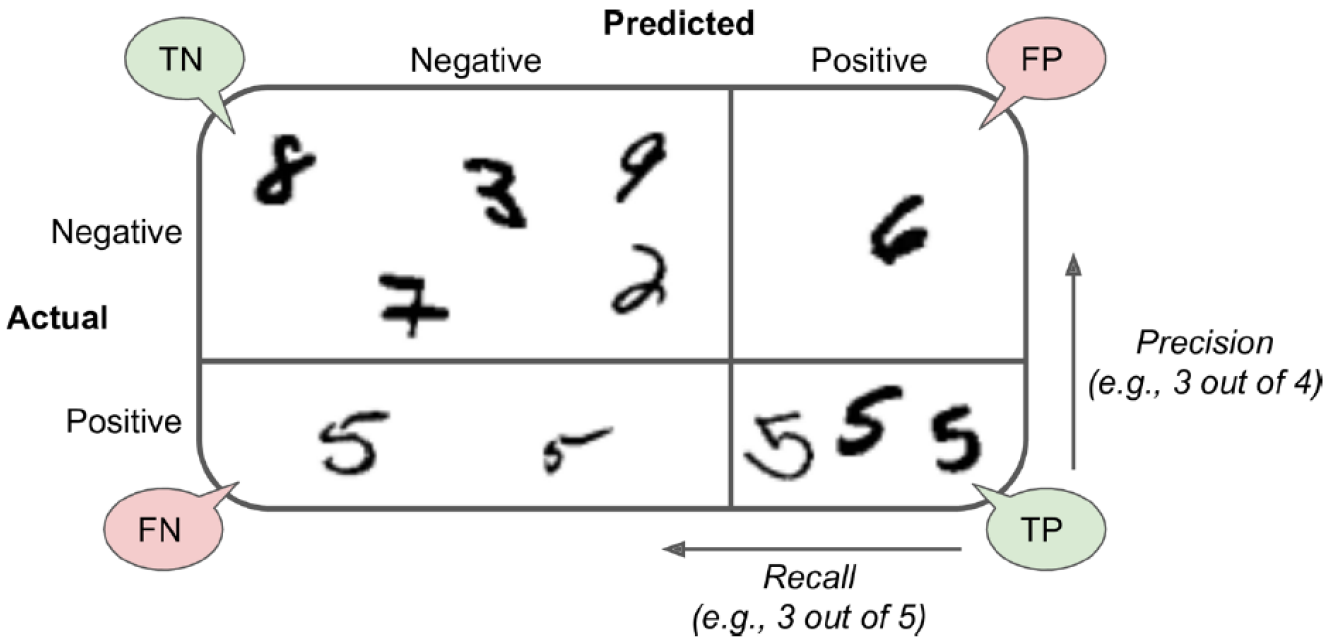
Précision, Rappel : à partir de la matrice de confusion, on peut évaluer d'autres métriques basées sur les notions de vrais positifs (TP), faux positifs (FP) et respectivement vrais et faux négatifs (TN, FN)

- **Précision :** évalue l'exactitude des prédictions positives

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

- **Rappel :** évalue le taux d'observations positives ayant été correctement détectées par le classifieur

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

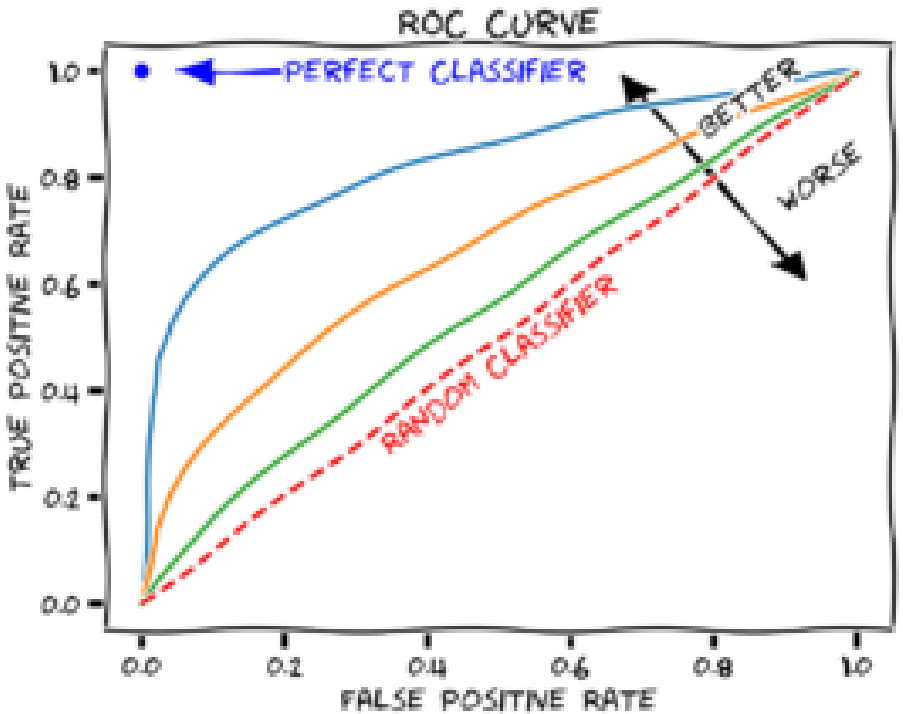


F-mesure : moyenne *harmonique* de la précision et du rappel donnant davantage de poids aux faibles valeurs.

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Rappel}} = 2\frac{Precision \times Rappel}{Precision \ + \ Rappel}$$

Interprétation : le classifieur n'obtiendra un bon score F_1 que si son rappel et sa précision sont élevés. Le score F_1 favorise donc les classifieurs ayant une précision et un rappel similaires (ce qui n'est pas forcément souhaitable suivant l'étude que vous menez).

Courbe ROC : courbe d'efficacité du récepteur (Receiver Operating Characteristic ou ROC) très semblable à la courbe précision/rappel mais elle croise les taux de vrais positifs (rappel) avec le taux de faux positifs (c-à-d le pourcentage d'observations négatives qui sont incorrectement classées comme positives (FP=1-TP)).

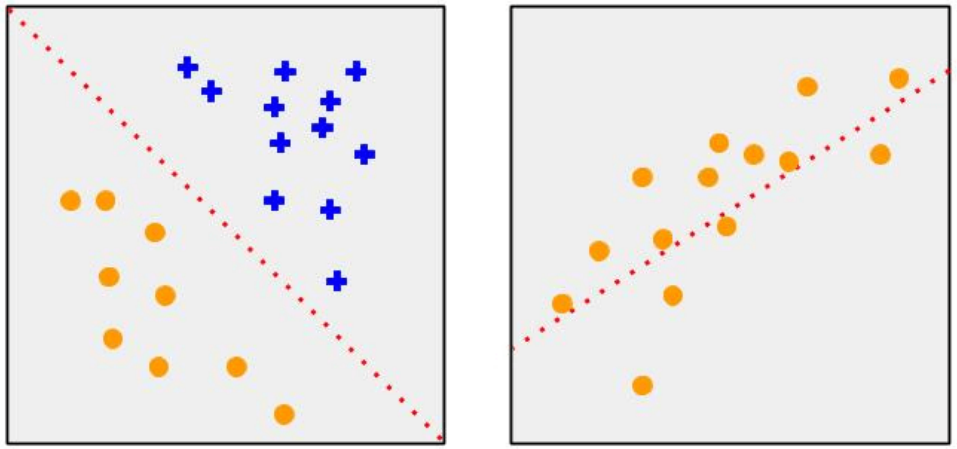


Courbe ROC : courbe d'efficacité du récepteur (Receiver Operating Characteristic ou ROC) très semblable à la courbe précision/rappel mais elle croise les taux de vrais positifs (rappel) avec le taux de faux positifs (c-à-d le pourcentage d'observations négatives qui sont incorrectement classées comme positives (FP=1-TP)).

Différences entre courbes Précision/Rappel et ROC :

- préférer une courbe Précision/Rappel si la classe positive est rare ou si on attache est plus importance aux faux positifs qu'aux faux négatifs;
- préférer la courbe ROC ou le score ROC AUC représentant l'aire sous la courbe ROC (égal à 1 si classifieur parfait) dans les autres cas.

2 principaux types d'apprentissage supervisé :



Approches par modèles supervisés :

- Arbres de Décision
- Apprentissage d'ensemble : Forêts aléatoires
- Réseaux de neurones
- Ouverture sur le Deep Learing

Méthodes d'évaluation :

- Validation croisée
- Matrice de confusion
- Precision, Rappel, F-mesure
- Courbe ROC
- Ouverture sur l'explicabilité

Deep Learning, boosting... sont souvent considérées comme des **boîtes noires**.

Comment évaluer réellement un modèle de Machine Learning s'il n'a aucun moyen d'apprécier le processus logique qui a conduit à générer tel ou tel résultat?

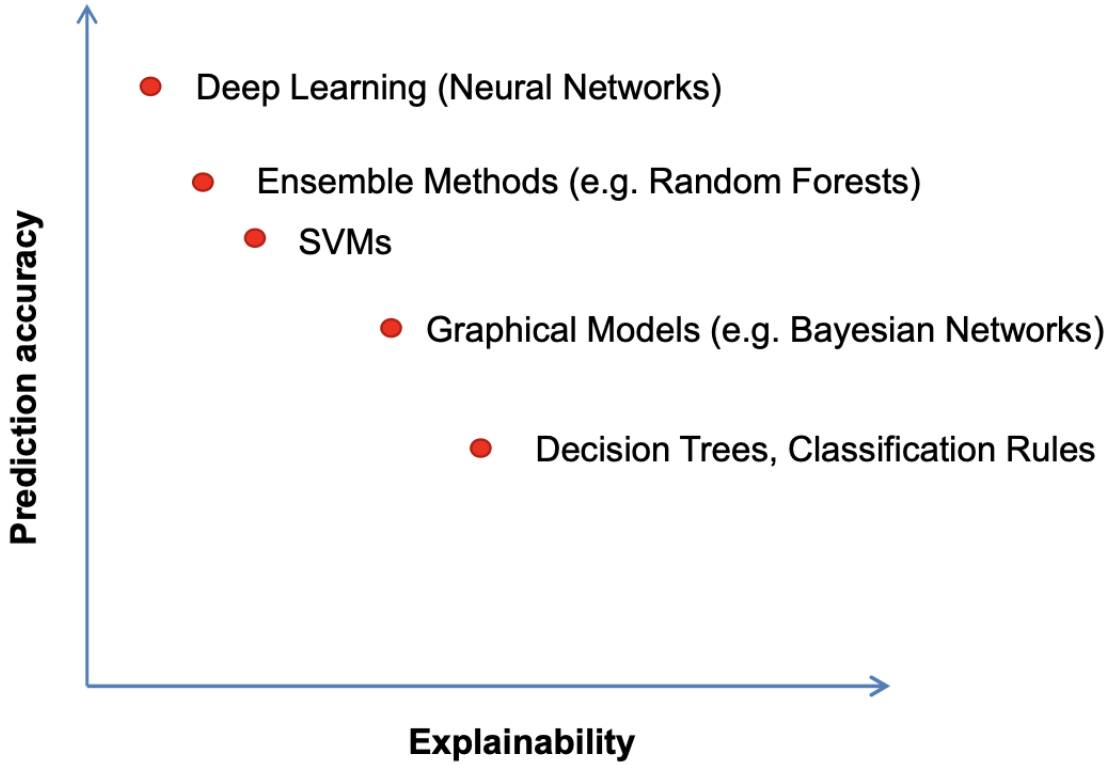


Figure: Caractère explicable des modèles de Machine Learning

$\Rightarrow$ interprétabilité et explicabilité

L'intelligibilité vient après l'étape de modélisation (celle où l'on construit le modèle prédictif).

- **Explicabilité** (ou *intelligibilité locale*) s'intéresse aux variables qui ont été déterminantes pour une décision particulière : LIME, Shapley Value
- **Interprétabilité** (ou *intelligibilité globale*) propose une évaluation globale d'un processus de décision : Shapley Value, feature importance...

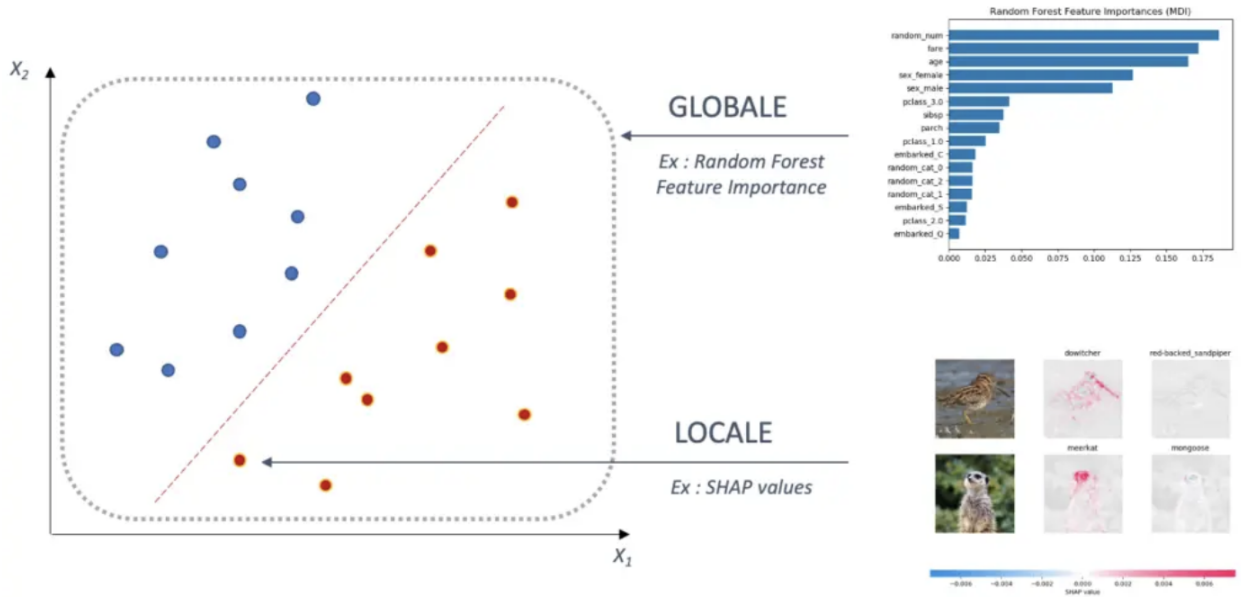


Figure: Exemple d'intelligibilité locale et globale

LIME (Local Interpretable Model-agnostic Explanations) : un modèle local qui cherche à expliquer la prédiction d'un individu par analyse de son voisinage.

- *Interprétable.* Il fournit une compréhension qualitative entre les variables d'entrée et la réponse. Les relations entrées-sortie sont faciles à comprendre.
- *Simple localement.* Le modèle est globalement complexe, il faut alors chercher des réponses localement plus simples.
- *Agnostique.* Il est capable d'expliquer n'importe quel modèle de machine learning.

Fonctionnement de LIME

- *1ère étape :* l'algorithme LIME génère des nouvelles données, dans un voisinage proche de l'individu à expliquer.

- *2ème étape :* LIME entraîne un modèle transparent sur les prédictions du modèle « boîte noire » complexe qu'on cherche à interpréter. Il apprend ainsi à l'aide d'un modèle simple et donc interprétable (par exemple, une régression linéaire ou un arbre de décision).

Le modèle transparent joue donc le rôle de modèle de substitut pour interpréter les résultats du modèle complexe d'origine.

SHAP (SHapley Additive exPlanations) utilise les valeurs de Shapley comme base, indiquant dans quelle mesure une caractéristique contribue à une prédiction

- agnostique
- **Interprétation locale :** estime l'effet de la caractéristique unique sur la variable cible

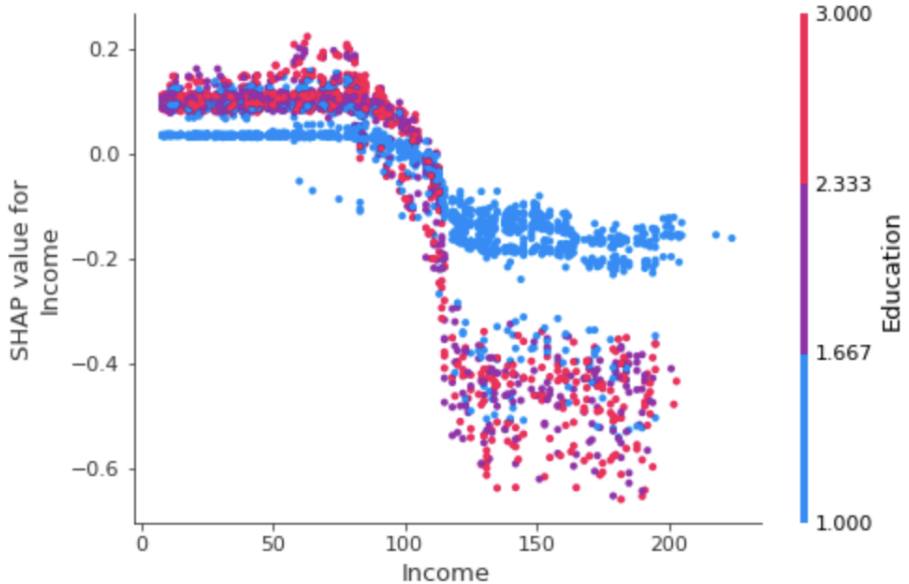
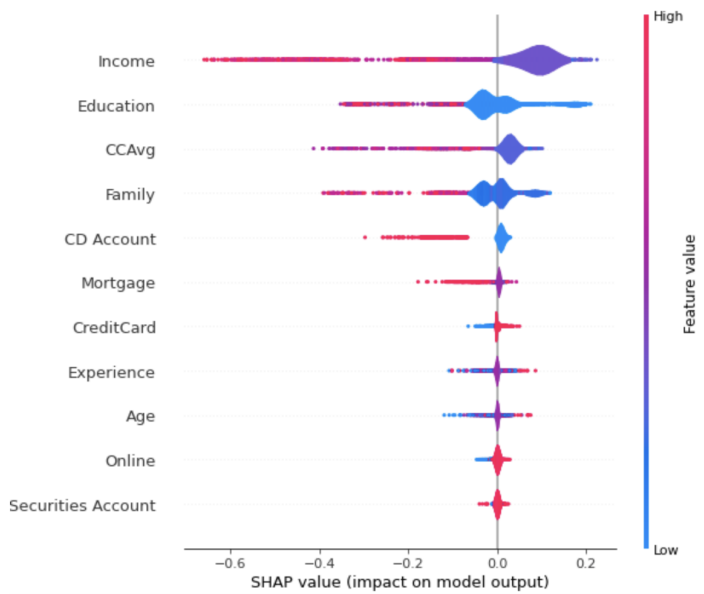


Figure: Exemple : relation négative entre le revenu et le prêt personnel, c'est-à-dire qu'un revenu élevé diminue les chances de l'échantillon de rejeter le prêt personnel.

SHAP (SHapley Additive exPlanations) utilise les valeurs de Shapley comme base, indiquant dans quelle mesure une caractéristique contribue à une prédiction

- agnostique
- **Interprétation locale :** estime l'effet de la caractéristique unique sur la variable cible
- **Interprétation globale :** explication globale de l'ensemble du modèle



- Une valeur de revenu élevée fait baisser la prédiction.
- Une valeur moyenne ou élevée de la variable "éducation" diminue également les chances de la classe actuelle.
- Des caractéristiques telles que "online" et " Securities Account" ont peu ou pas d'effet sur la prédiction, car leurs valeurs SHAP sont proches de 0.

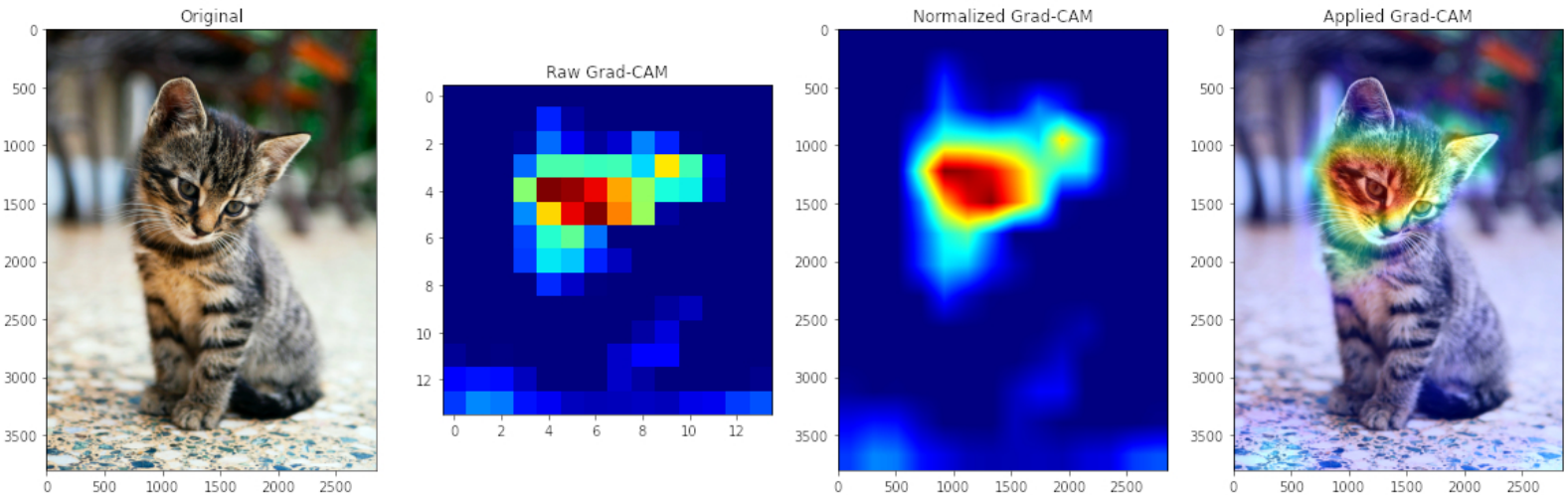
Fonctionnement Shapley value

- *1ère étape :* calcul des valeurs de Shapley pour un individu en particulier : simuler différentes combinaisons de valeurs pour les variables d'entrée
- *2ème étape :* Pour chaque combinaison, calculer la différence entre la valeur prédite et la moyenne des prédictions. La valeur de Shapley d'une variable correspond alors à la moyenne de la contribution de sa valeur en fonction des différentes combinaisons.

Figure: Exemple : SHAP permet de traduire la prédiction d'un individu en explication sous forme de somme de contributions de chacune des variables.

Interprétabilité pour les architectures convolutives CNN :

- **Grad-CAM** consiste à chercher quelles parties de l'image ont conduit un réseau neuronal convolutif à sa décision finale. Cette méthode consiste à produire des cartes thermiques représentant les classes d'activation sur les images reçues en entrée. Une classe d'activation est associée à une classe de sortie spécifique



Interprétabilité pour les architectures convolutives CNN :

- **Occlusion Map :** calcule l'importance de chaque parcelle en observant le changement de la sortie du modèle lorsque la parcelle est supprimée. Les résultats individuels peuvent être assemblés en une carte d'attribution.