Rôle des algorithmes du Machine Learning (ML) dans l'analyse de la relation entre la structure chimique et son activité (SAR)

Dans le cadre du projet SCHISM*

Hajar KARINE (née REHIOUI)
hajar.rehioui-karine@unicaen.fr

06/07/2021

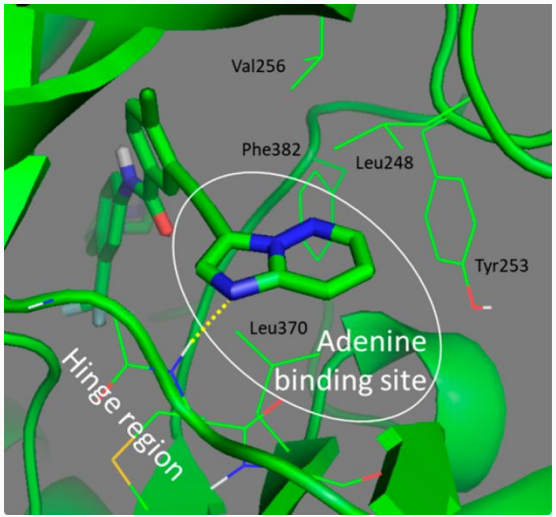
* Le projet SCHISM est financé par l'Union européenne dans le cadre du programme opérationnel FEDER/ FSE 2014-2020

Contexte général

Contexte général

Découverte de médicament (Drug Discovery)

Pour découvrir l'impact d'un médicament sur **une cible** (protein responsable d'une maladie), un certain nombre de **molécules** sont utilisées (ou testées) comme **inhibiteurs** de ces **cibles**. Si l'interaction entre la molécule est **la cible** se fait, on dit que la molécule est **active** sinon elle est **inactive**



1. METIVIER, Jean-Philippe, CUISSART, Bertrand, BUREAU, Ronan, *et al.* The pharmacophore network: a computational method for exploring structure–activity relationships from a large chemical data set. *Journal of medicinal chemistry*, 2018, vol. 61, no 8, p. 3551-3564.

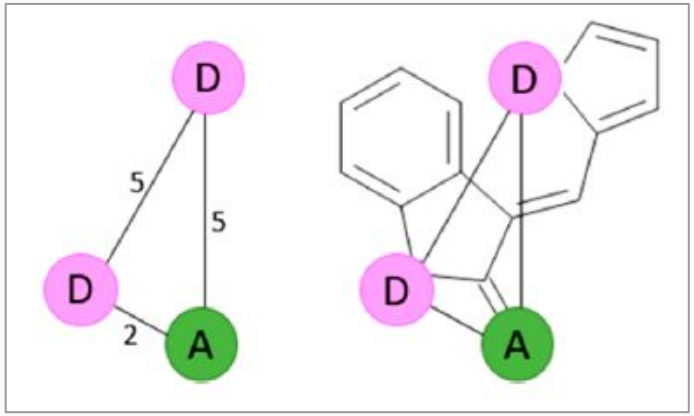
Contexte général

Pharmacophore

Un **Pharmacophore** correspond à un fragment d'une molécule responsable de son **activité biologique** (**active** ou **inactive**).

Les marqueurs pharmacophoriques sont identifiés afin de construire construire un graphe pharmacophorique

Une **molécule** pourra être décrite par un ensemble de **pharmacophores**

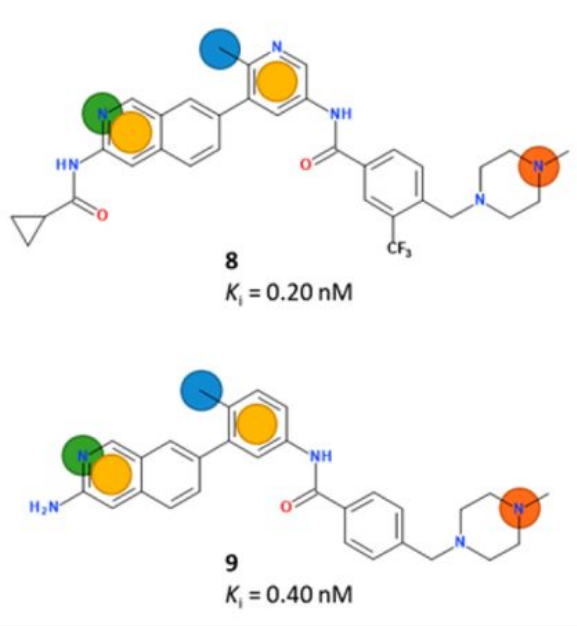


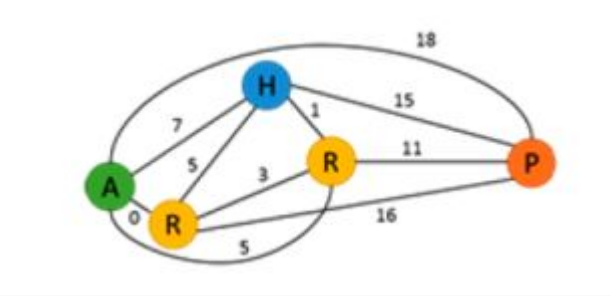
Exemple d'un pharmacophore et de son graphe

Structure Activity Relationship (SAR)

SAR

La relation entre la structure chimique et son activité part du principe que deux **structures chimiques similaires** auront **des activités similaires**





Machine Learning (ML) et analyse SAR

Pourquoi l'utilisation du ML[1]?

- la complexité des molécules étudiées
- la grande taille des données

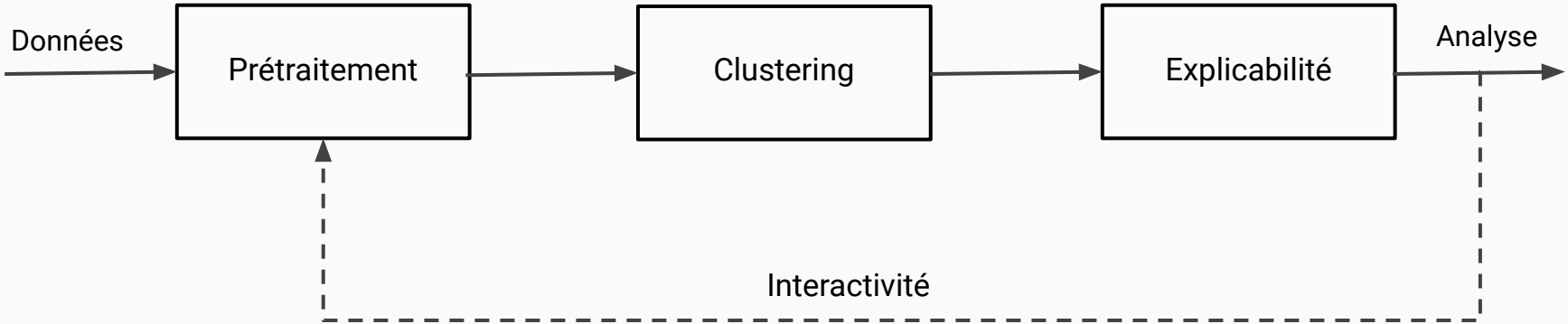
Les aspects du ML

- méthodes supervisées
- méthodes non supervisées
- méthodes semi-supervisées

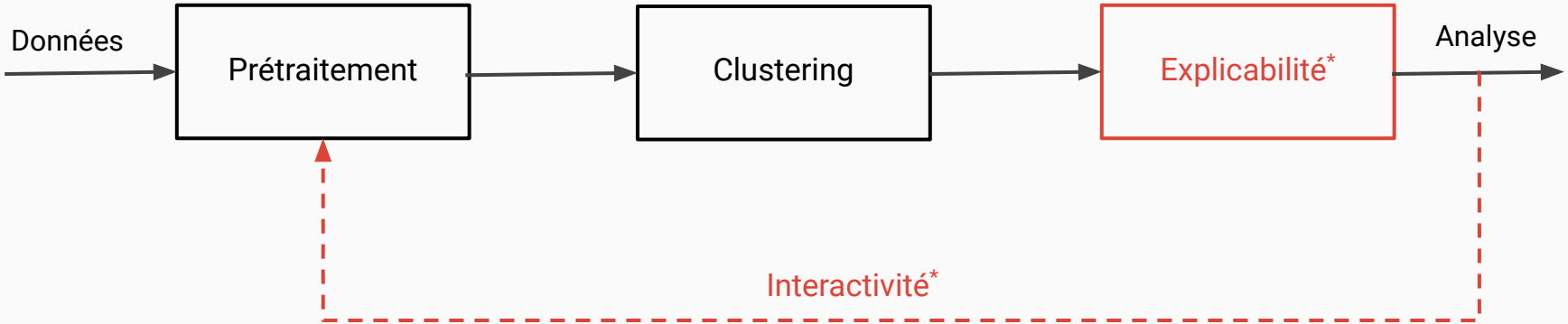
- prétraitement de la data: feature selection, feature reduction
- régression, classification, regroupement (clustering)
- analyse, interprétation, explication

1. VAMATHEVAN, Jessica, CLARK, Dominic, CZODROWSKI, Paul, *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 2019, vol. 18, no 6, p. 463-477.

Approches


Données
Prétraitement
Clustering
Explicabilité
Analyse
Interactivité

Approches



* perspectives

Prétraitement

Jeu de données étudié

Description

- L'objectif est l'étude d'une tyrosine kinase nommée BCR-ABL responsable de la leucémie.
- 1485 molécules qui ont pour cible BCR-ABL, et qui sont décrites par 112048 pharmacophores d'ordres 3 à 7.
- Jeu de données générés par le logiciel Norns[1] à base du Dataset CHEMBL[2]

Jeu de données	#instances	#attributs	Type de données
Molécules	1485	112048	binaires (0, 1)

Id_Mol	|A|A|A|-0	|A|A|A|-1	|A|A|A|-2	|A|A|A|-3
CHEMBL250213	1	0	0	0

1. METIVIER, Jean-Philippe, CUISSART, Bertrand, BUREAU, Ronan, *et al.* The pharmacophore network: a computational method for exploring structure–activity relationships from a large chemical data set. *Journal of medicinal chemistry*, 2018, vol. 61, no 8, p. 3551-3564.

2. GAULTON, Anna, HERSEY, Anne, NOWOTKA, Michał, *et al.* The ChEMBL database in 2017. *Nucleic acids research*, 2017, vol. 45, no D1, p. D945-D954.

Problèmes rencontrés et solutions

Problèmes

1) Problèmes de mémoire lors des exécutions

2) Données compliquées pour l'analyse clustering

Solutions:

Prétraitement de données:

1) Réduire le nombre d'attributs ⟶ Classes d'équivalence

2) Transformer les données pour obtenir de meilleurs résultats de clustering

Prétraitement : classes d’équivalence

Une classe d’équivalence (CE)

Une **CE** regroupe les **pharmacophores** qui **apparaissent exactement** dans les mêmes **molécule**s

Un des **pharmacophores** de chaque **CE** est choisi comme **représentant** de sa classe.

Cette méthode nous a permis de faire une **réduction de 80,25%**

Jeu de données	#instances	#attributs	Type de données
Molécules	1485	**22127**	binaires (0, 1)

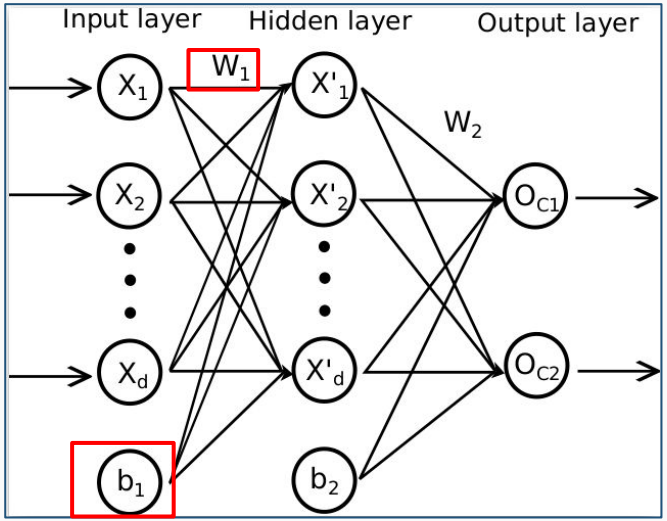
Prétraitement : transformation des vecteurs d'attributs

Transformation supervisée

Cette transformation a pour objectif de regrouper les molécules selon leur activité biologique.

Elle utilise un réseau de neurones *(NN) qui minimise la fonction d'optimisation "categorical_crossentropy"

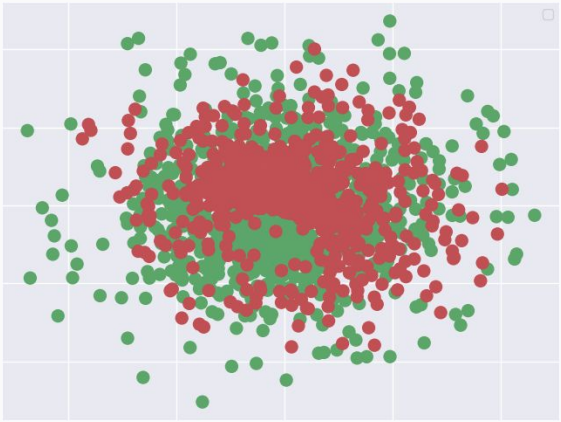
Les poids et biais qui ont été utilisés dans la phase d'entraînement sont utilisés pour faire une transformation linéaire



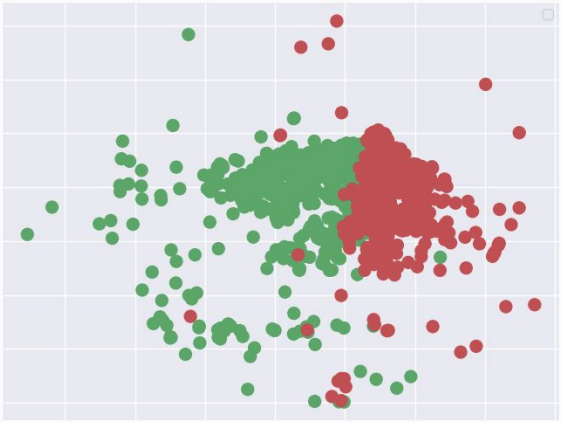
$$x'_{ij} = x_{ij} \times w_{jj} + b_j$$

* Implémentation faite à l'aide des bibliothèques Keras et Tensorflow

Prétraitement : transformation des vecteurs d’attributs



Original

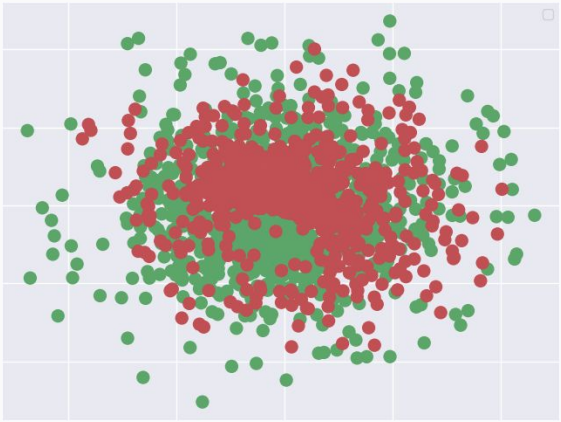


Transformé

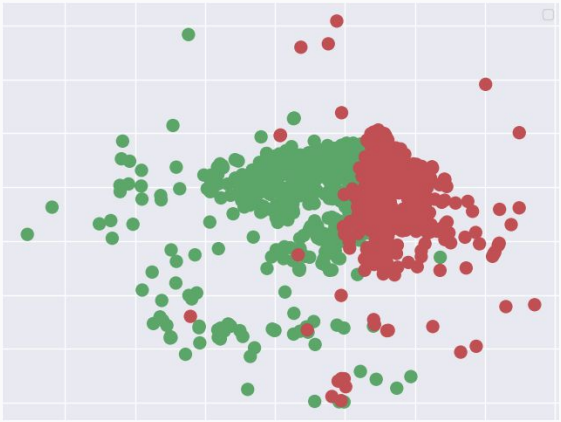
Inactive

Active

Prétraitement : transformation des vecteurs d'attributs



Original



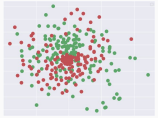
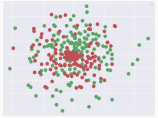
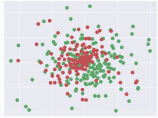
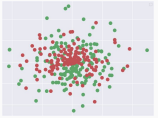
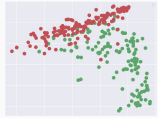
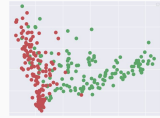
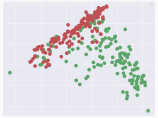
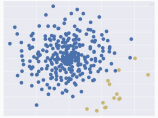
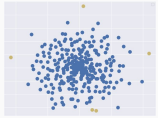
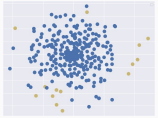
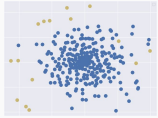
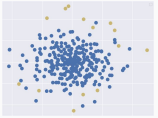
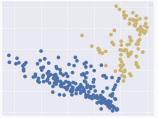
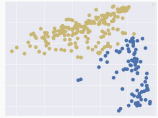
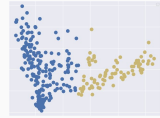
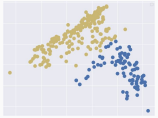
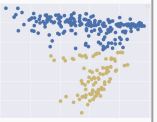
Transformé

Evaluation: NMI

NMI *(Normalized mutual information): est une mesure de qualité qui compare les clusters ou classes résultants avec la vérité terrain. Les résultats varient entre 0 (aucune information mutuelle) et 1 (corrélation parfaite)

* L'évaluation est faite à l'aide de la bibliothèque scikit-learn

Validation du modèle de la transformation (5-fold)

Jeu de données	Original					Transformé				
avant clustering										
après clustering										
NMI	0.088	0.045	0.054	0.024	0.088	0.480	0.395	0.354	0.396	0.427
NMI (5-folds)	0.060					0.410				

Légende: Inactive Active Cluster 0 Cluster 1

	Original		Transformé	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Inactive (711)	707 (99.44%)	4 (0.56%)	709 (99.72%)	2 (0.28%)
Active (774)	716 (92.51%)	58 (7.49%)	302 (39.02%)	472 (60.98%)

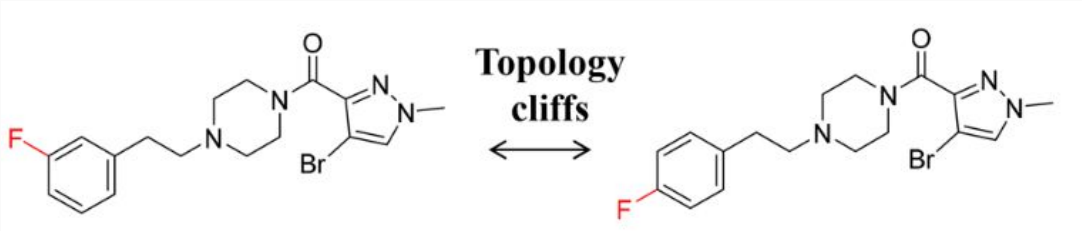
Clustering

Clustering et analyse SAR

Objectif

Le clustering permet de regrouper les structures similaires et de découvrir les sous-familles d'activités ainsi nous avons la possibilité d'analyser les clusters:

- homogènes → découvrir des structures d'activités similaires → Être apte à expliquer la cause de ce regroupement
- non homogènes → découvrir des activités cliffs → Être apte à expliquer la cause de l'activité cliff



Exemple[1] d'activité cliff

1. STUMPFE, Dagmar, HU, Huabin, et BAJORATH, Jürgen. Evolving concept of activity cliffs. ACS omega, 2019, vol. 4, no 11, p. 14360-14368.

Défis

Défis

Comment prédire la sous activité pour une nouvelle donnée ?

Solutions:

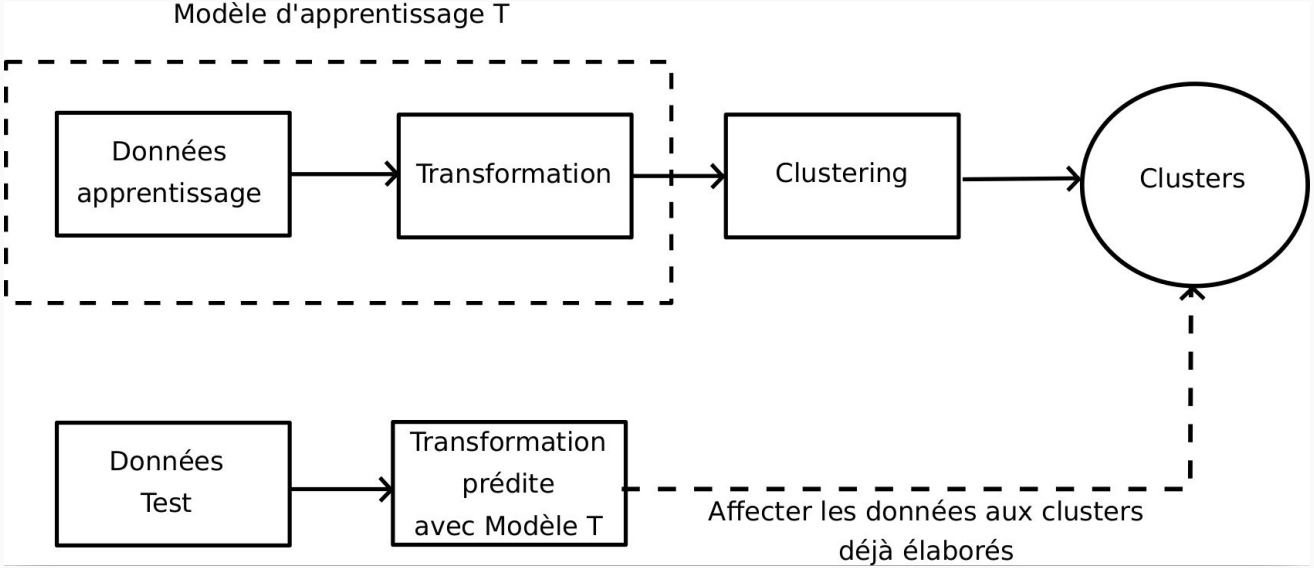
Introduire la notion du clustering prédictif

Clustering prédictif

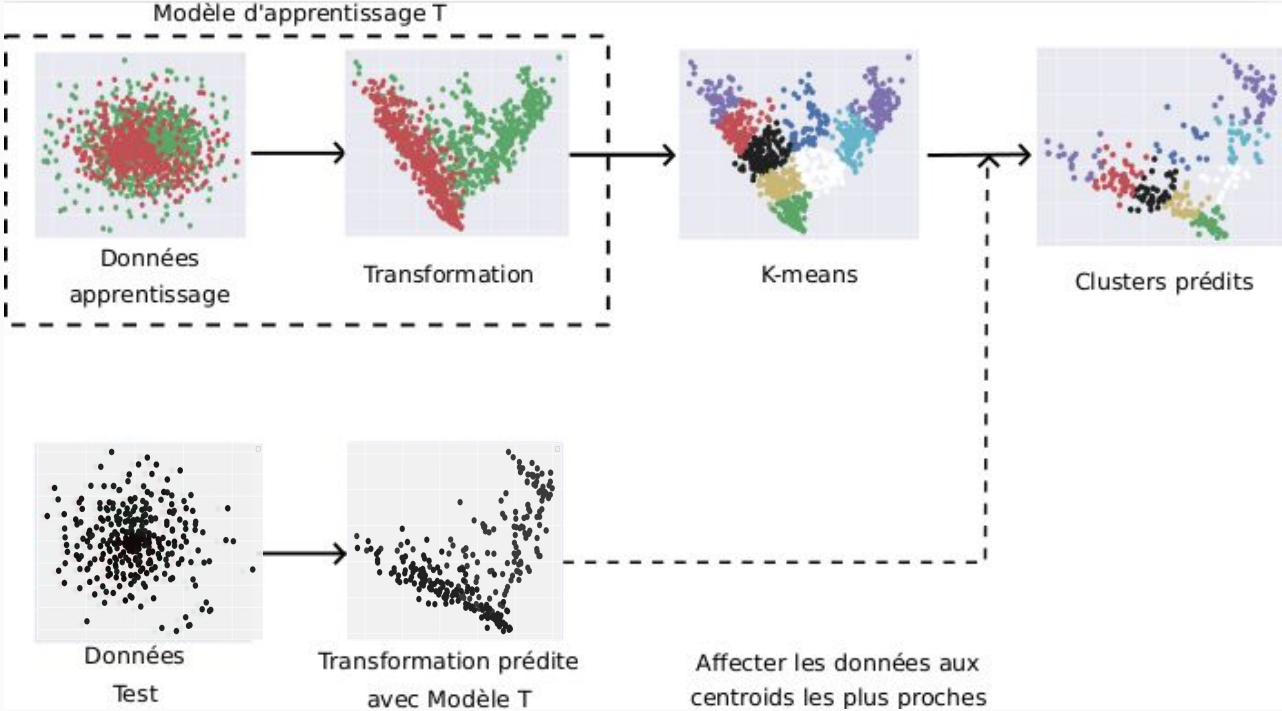
Principe

1) Appliquer l'une des méthodes de clustering sur l'ensemble d'apprentissage après sa transformation

2) Affecter à chaque cluster construit dans l'étape (1) les données qui lui correspondent (les plus similaires)



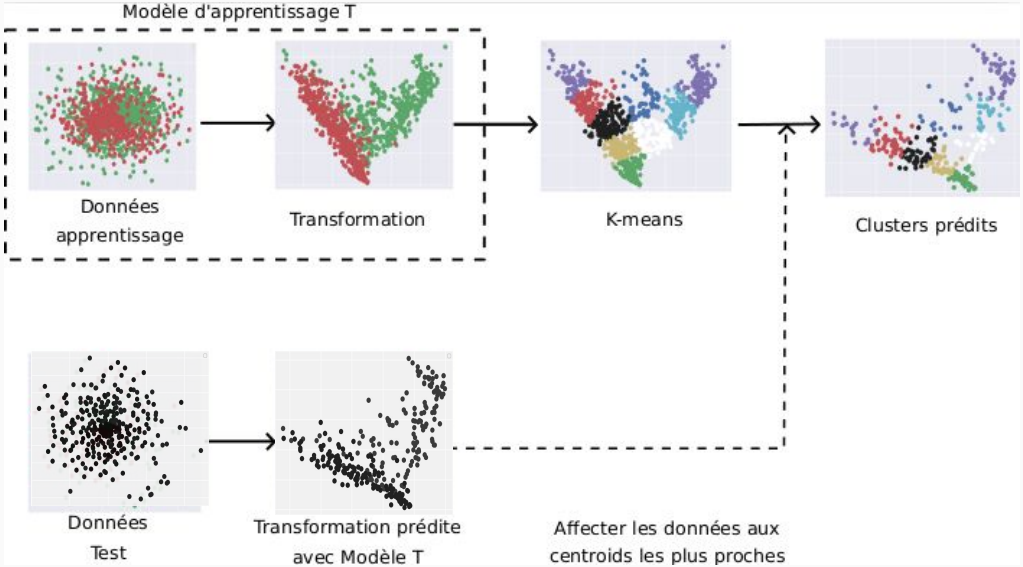
Clustering prédictif



Exemple clustering prédictif avec K-means*

* K-means et la prédiction de clusters est faite à l'aide de la bibliothèque scikit-learn

Clustering prédictif



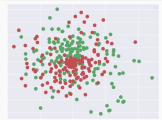
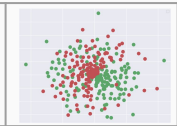
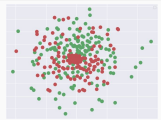
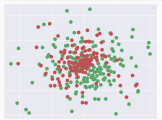
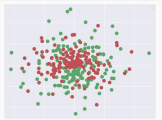
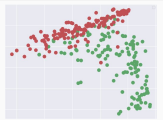
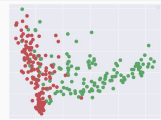
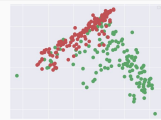
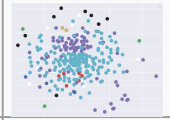
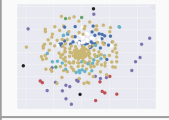
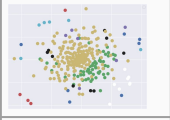
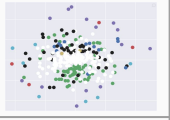
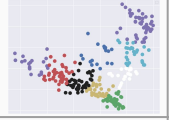
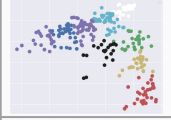
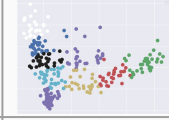
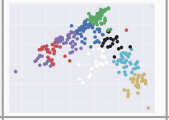
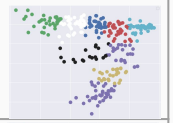
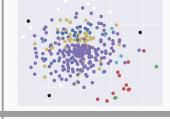
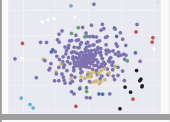
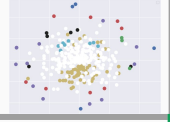
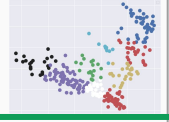
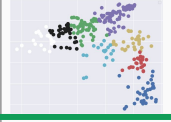
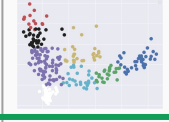
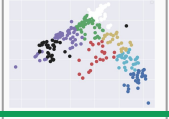
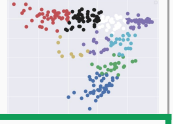
Exemple clustering prédictif avec K-means

Evaluation: Silhouette

Silhouette* est une mesure calculée en utilisant la distance moyenne intra-cluster et la distance moyenne du cluster le plus proche. La meilleure valeur est 1 et la moins bonne est -1. Les valeurs proches de 0 indiquent que les clusters se chevauchent. Les valeurs négatives indiquent généralement qu'un échantillon a été affecté au mauvais cluster, car un cluster différent est plus similaire.

* L'évaluation est faite à l'aide de la bibliothèque scikit-learn

Clustering prédictif (5-folds)

données	Original					Transformé				
avant clustering										
K-means										
Affinity-Propagation										
NMI K-means	0.252	0.263	**0.230**	**0.283**	0.258	**0.345**	0.288	0.300	**0.282**	0.278
silhouette K-means	0.130	**0.142**	**0.098**	**0.225**	-0.005	**0.472**	0.496	0.469	**0.489**	0.455
NMI Affinity	**0.301**	**0.302**	0.175	0.172	**0.273**	**0.345**	**0.291**	**0.306**	0.280	**0.299**
silhouette Affinity	**0.187**	0.075	0.089	0.196	**0.193**	0.471	0.479	**0.483**	0.466	**0.507**

Transformation (Matrice de confusion)

Meilleur fold (5)	Original	Transformé
Inactive (711)	0 0 0 0 3 4 4 128 3	0 1 3 35 0 0 38 27 38
Active (774)	4 63 9 8 0 4 2 63 1	45 8 18 11 29 15 19 7 2

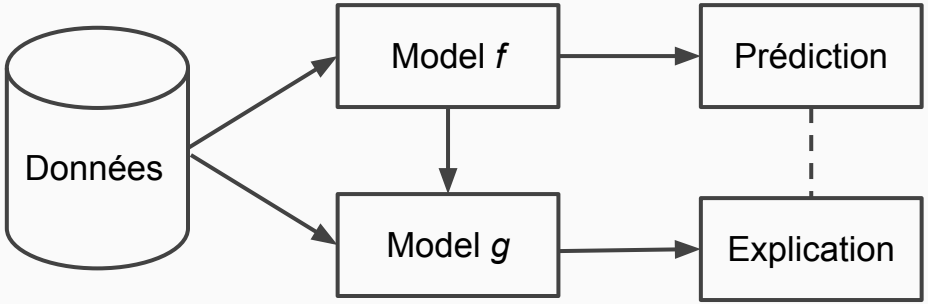
Moins bon fold(2)	Originale	Transformé
Inactive (711)	0 0 0 0 0 0 4 3 135	0 0 0 0 30 4 23 54 31
Active (774)	8 28 4 6 6 3 4 3 93	27 23 32 25 11 4 9 14 10

Analyse d'activité cliffs dans les régions fortement non homogènes

Perspectives
-Explicabilité-

Feature Importance [1]

Création d'un modèle parallèle *g* qui explique le modèle de prédiction *f*



Le modèle d'explication général se calcule comme suit :

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$$

L'idée de base c'est en faisant la somme des contributions de tous les features il faudra avoir la même prédiction (output) que le modèle orignal $f(x)$.

1. Lundberg, S. M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30; NIPS, 2017.

Feature Importance [1]

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$$

$x' \in \{0,1\}^M$, dite représentation interprétable, est un vecteur binaire indiquant la présence ou l'absence du feature. M est le nombre de features en entrée

ϕ_0 est dite valeur de base. il représente la valeur moyenne des prédiction du modèle *f*

ϕ_i est une valeur réelle désignant la contribution du feature

1. Lundberg, S. M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30; NIPS, 2017.

Algorithm SHAP [1,2]

Dans le cas de SHAP, la contribution de chaque feature i est calculée comme suit:

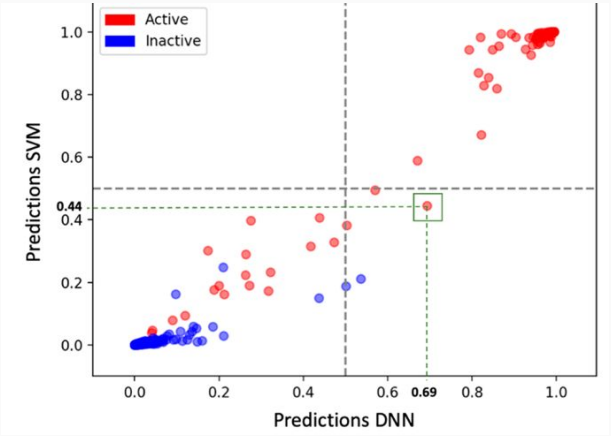
$$\phi_i = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![f(S \cup \{i\}) - f(S)]$$

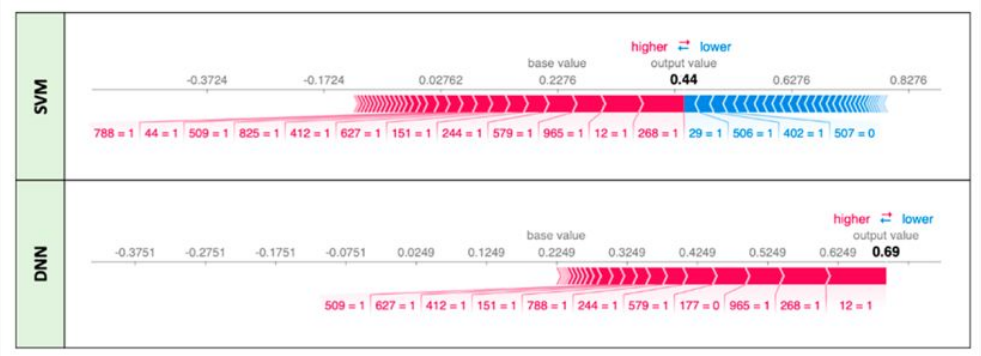
- N : l'ensemble de tous les features
- S : tous les sous ensembles de N moins le feature i
- $f(S \cup \{i\})$: prédiction d'un sous ensemble S en considérant le feature *i*
- $f(S)$: prédiction d'un sous ensemble S
- $f(S \cup \{i\}) - f(S)$: la contribution marginal du feature *i* dans chaque sous ensemble *S*
- $|S|!$: toutes les permutations former par S avant l'ajout de i.
- $(|N| - |S| - 1)!$: toutes les permutations former après l'ajout de i.

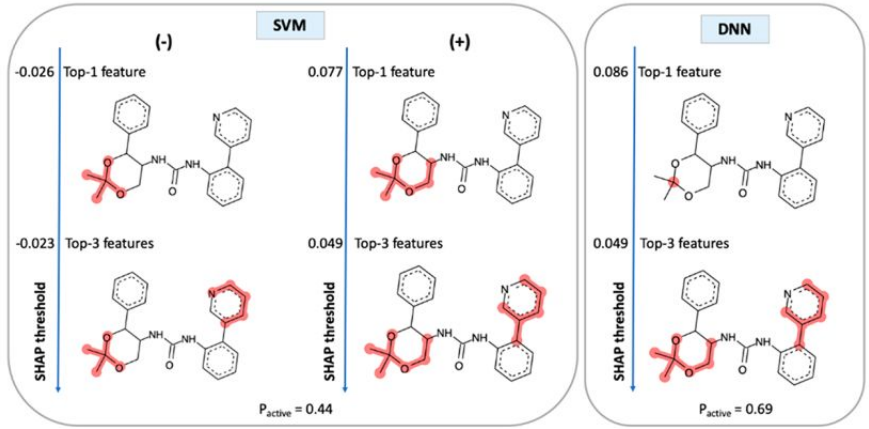
1. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J. Med. Chem. 2020, 63, 8761−8777.
2. exemple explicatif: https://towardsdatascience.com/explainable-ai-application-of-shapely-values-in-marketing-analytics-57b716fc9d1f

Algorithm SHAP in drug Discovery[1]

Les résultats sont obtenus en se basant sur le framework SHAP (python) développé par Scott Lundberg[2]



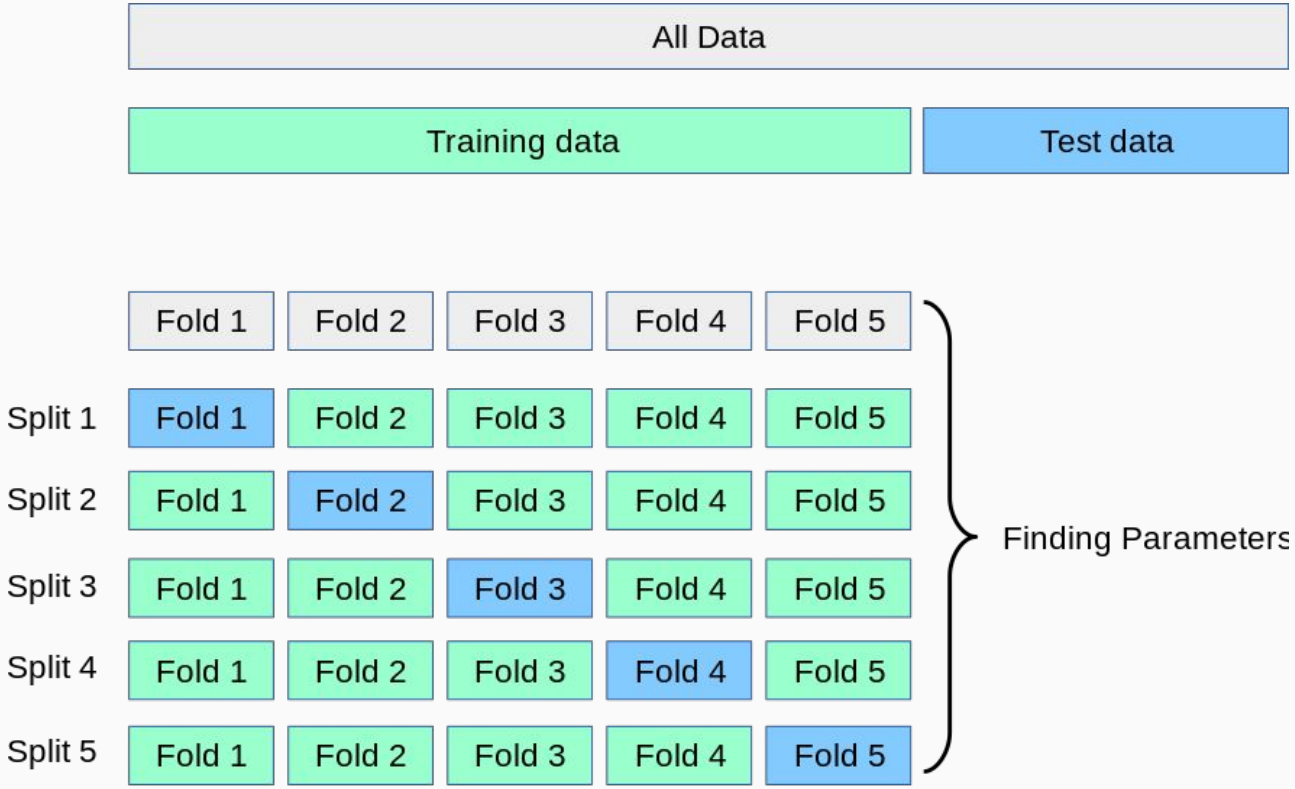




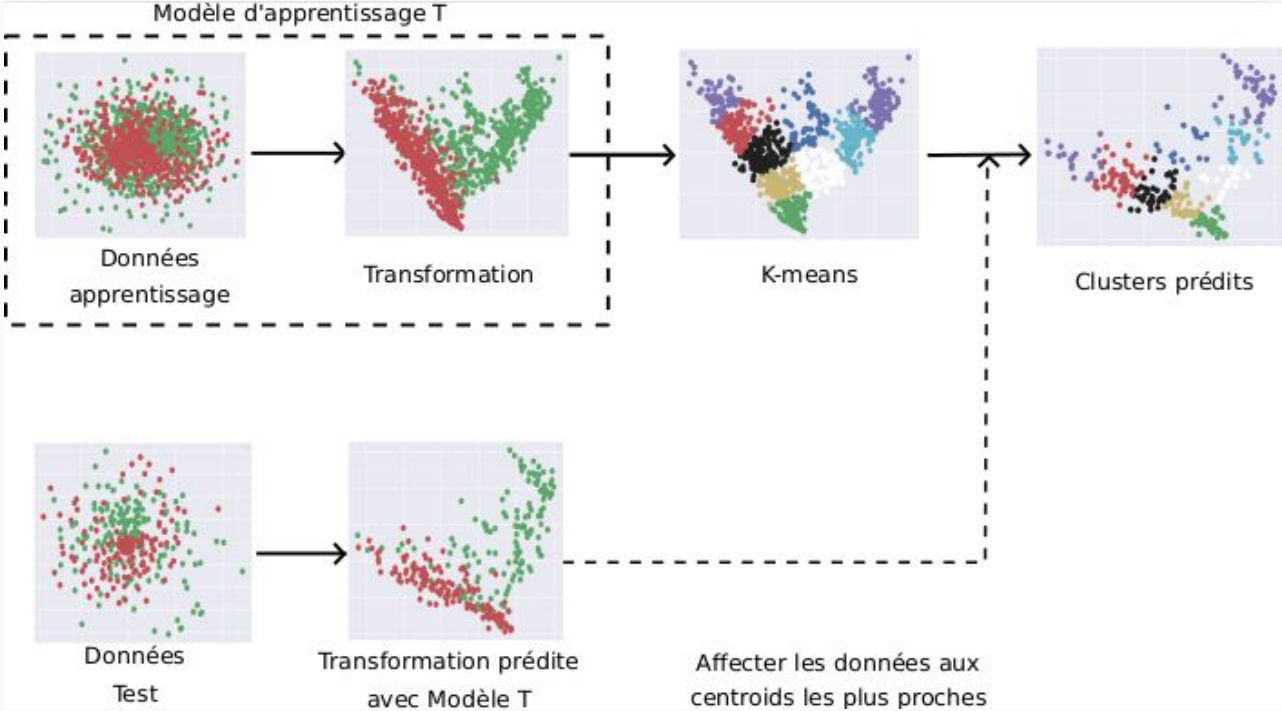
1. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J. Med. Chem. 2020, 63, 8761−8777.
2. https://github.com/slundberg/shap

Merci pour votre attention

Annexe


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters

Clustering prédictif



Exemple clustering prédictif avec K-means*

* K-means et la prédiction de clusters est faite à l'aide de la bibliothèque scikit-learn