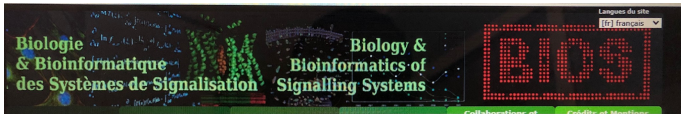
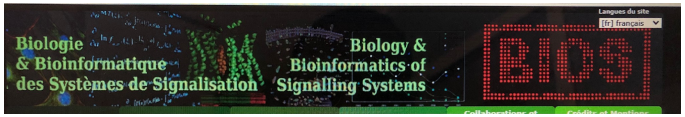


Analyse de l'expression différentielle par les auto encodeurs

Alpha Oumar Ba

Université de Paris Cité

Février -Aout 2022



Rapport de stage Master M2 IMB

Tutrice entreprise : Misbah Razzaq
misbah.razzaq@inrae.fr

Tuteur académique :Vittorio Perduca
vittorio.perduca@gmail.com

Équipe :BIOS

1. Analyse Exploratoire de notre Jeux de données
   - Détection des valeurs aberrantes
2. Réduction de dimensions
   - Méthodes Analyse des composantes principales
     - Visualisation
   - Méthode t-distributed Stochastic Neighbor Embedding
3. Auto encodeur
   - Description de notre modèle
4. Visualisation t-SNE encodeur
   - TSNE
5. Interprétation Modèle
   - Avantage de SHAP

[3] [1] [2]

Pierre-Marc Jodoin.
Ift 603/ift 712–techniques d'apprentissage.
2021.

Marc Parizeau.
Réseaux de neurones.
GIF-21140 et GIF-64326, 124, 2004.

Fernanda et Johnson Ian Wattenberg, Martin et Viégas.
Comment utiliser efficacement t-sne.
Distiller, 1.

Présentation entreprise

- Siège INRAE [a]
- Année de création 1983
- Forme juridique : Établissement public national à caractère administrative
- Activité : recherche développement en autres sciences physique et naturelles

a. Institut de Recherche pour l'agriculture et l'environnement

Définition et exemple

- Analyse et Exploitation des données : Les données transcriptomics permettent principalement de mesurer l'abondance des acides ribo-nucléiques messagers (ARNm) pour un grand nombre de gènes de manière simultanée. Dans le cas d'une analyse différentielle, cette technique permet en particulier de comparer l'expression des gènes dans différente condition.

	C01_FC_NS	C02_FC_S	C03_FB1_NS	C04_FB1_S	C05_FB2_NS	C06_FB2_S	C19_FC_NS	C20_FC_S	C21_FB1_NS	C22_FB1_S	...	C42_FB2_S	B1a
0	34.81	37.97	35.30	37.23	37.59	38.26	36.92	40.29	36.11	39.09	...	27.20	37.91
1	36.73	38.98	41.54	35.71	36.76	35.10	31.63	33.15	36.44	32.62	...	40.63	36.21
2	34.18	35.63	33.06	37.59	36.00	36.25	33.19	32.72	35.03	37.03	...	39.07	37.55
3	38.40	37.28	35.47	38.75	33.93	39.65	38.19	35.25	38.83	41.51	...	37.02	37.46
4	37.48	41.67	42.32	38.22	39.72	36.94	34.61	32.98	39.62	38.24	...	41.51	36.83

Figure – Exemple Jeux de données

- Dimension : 20330 variables et 27 observation

Objectif

Analyser ⇒ **Appliquer** ⇒ **Ameliorer** ⇒ **Réseau de neurones artificielles**

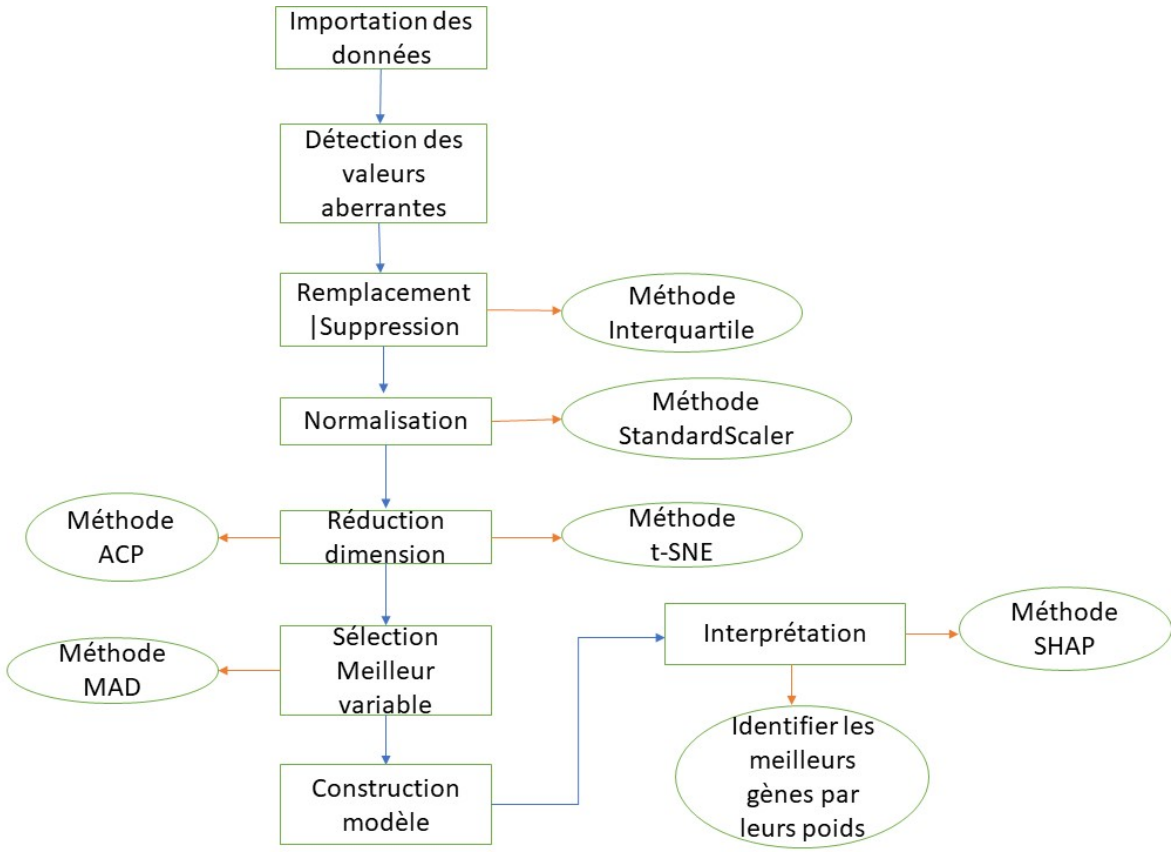


Figure – Description travail

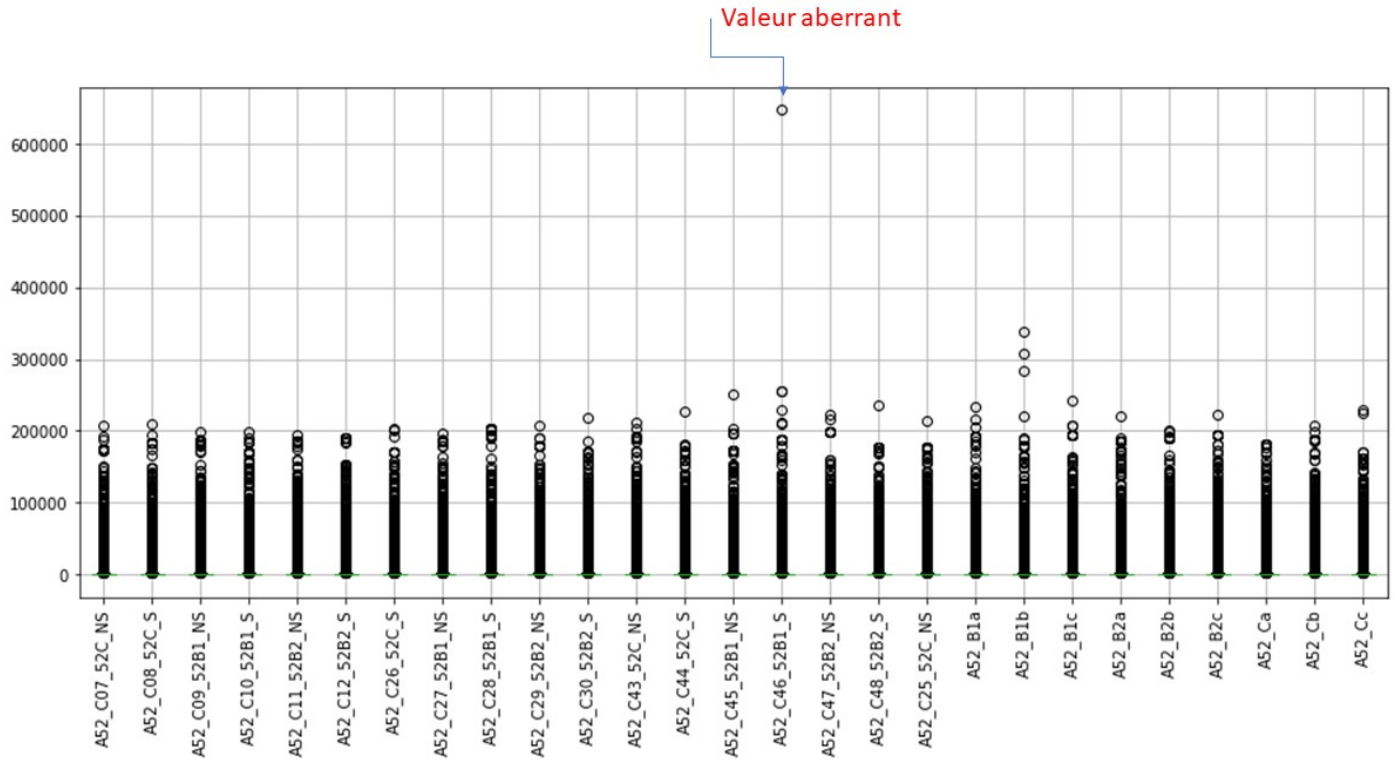


Figure – Détection des valeurs aberrantes

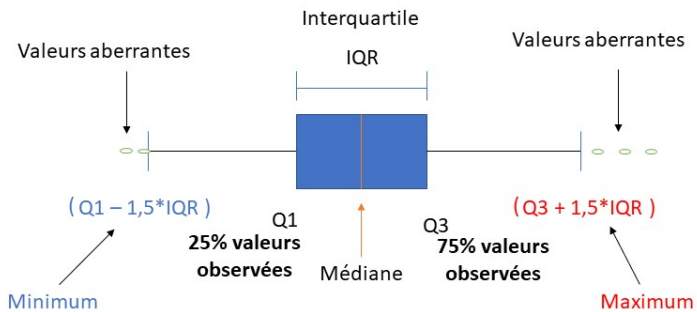


Figure – Boxplot avec les valeurs aberrantes

Remplacement des valeurs aberrantes

Méthode inter-quartile

C'est une méthode qui consiste à remplacer tous les valeurs qui sont plus **grand** que le Maximum par upper et tous les valeurs sont plus **grand** que Minimum par lower.

remarque

$$iqr = Q_3 - Q_1 \Longrightarrow \text{interquartile} \quad (1)$$

$$lower = Q_1 - 1.5iqr \quad (2)$$

$$upper = Q_3 + 1.5iqr \quad (3)$$

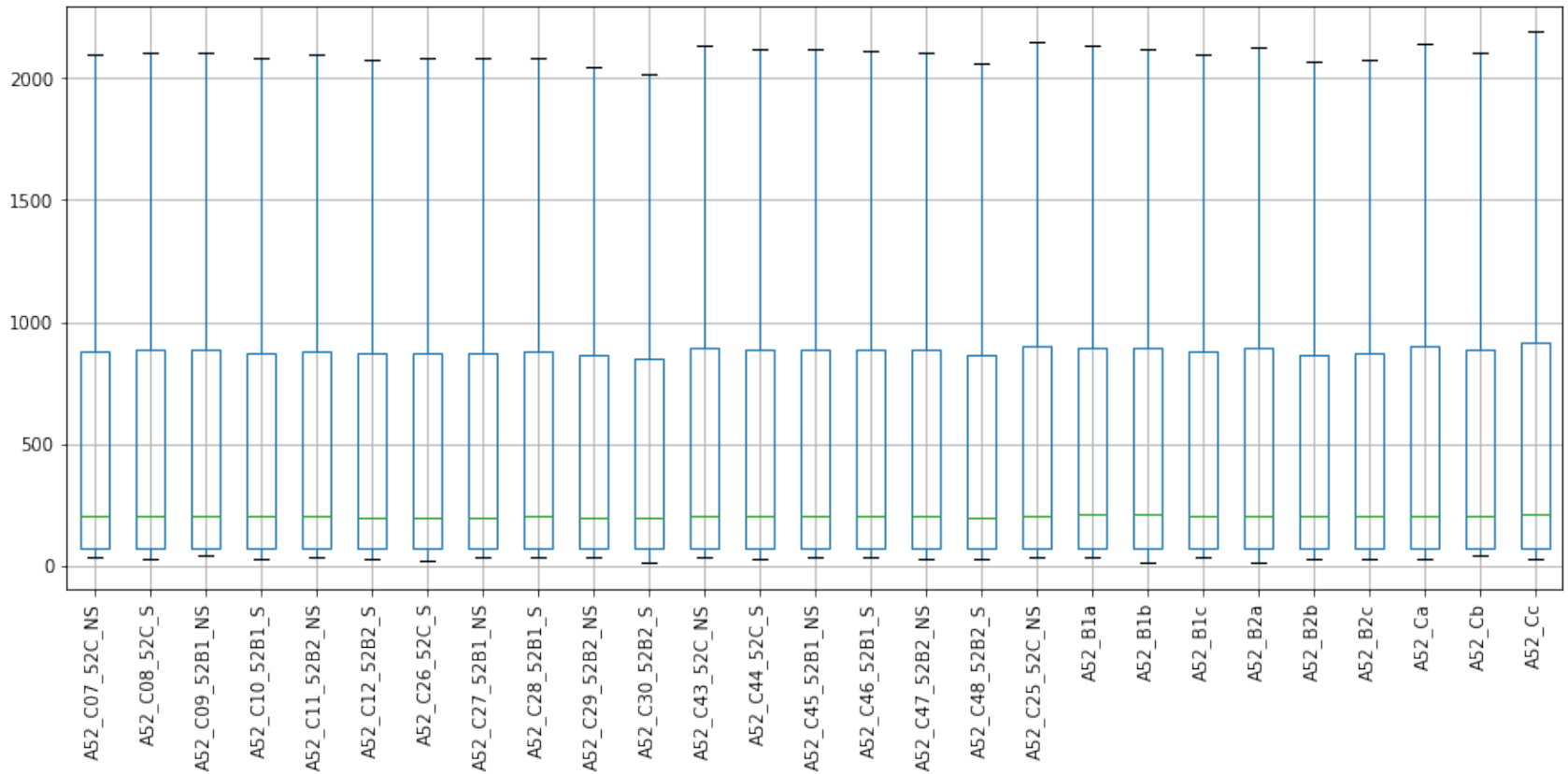


Figure – Boxplot de substitution

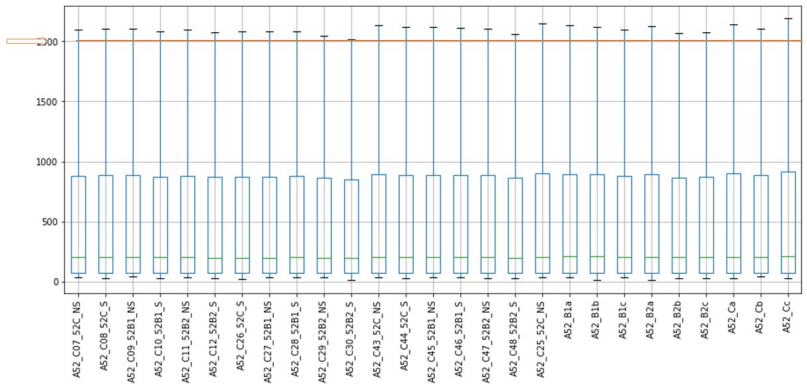


Figure – Boxplot de substitution

Attention

On voit que pour certains observations les niveaux d'expressions varient beaucoup. Les plages de valeurs sont différentes d'une observation à l'autre. Cette situation peut impacter fortement l'analyse des composantes principal (ACP) qui est principalement basée sur la variance. Les observations avec une forte variance absolue auront plus de poids dans l'analyse que ceux avec une faible variance.

Normalisation Méthode StandardScaler

La standardisation donne à une variable une moyenne nulle et un écart-type de 1, similaire à une Loi normale centrée et réduite.

Calcul :

$$x_{stand} = \frac{x-\mu}{\sigma}$$ avec x : les observations , μ : la moyenne et σ : l'écart type

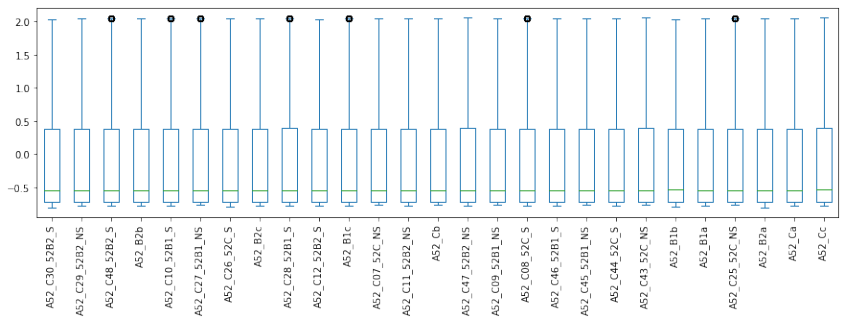


Figure – Boxplot de substitutions normalisé

Apprentissage non supervisée

L'apprentissage non supervisé est une branche des apprentissages automatiques caractérisée par l'analyse et le regroupement de données non étiquetées.

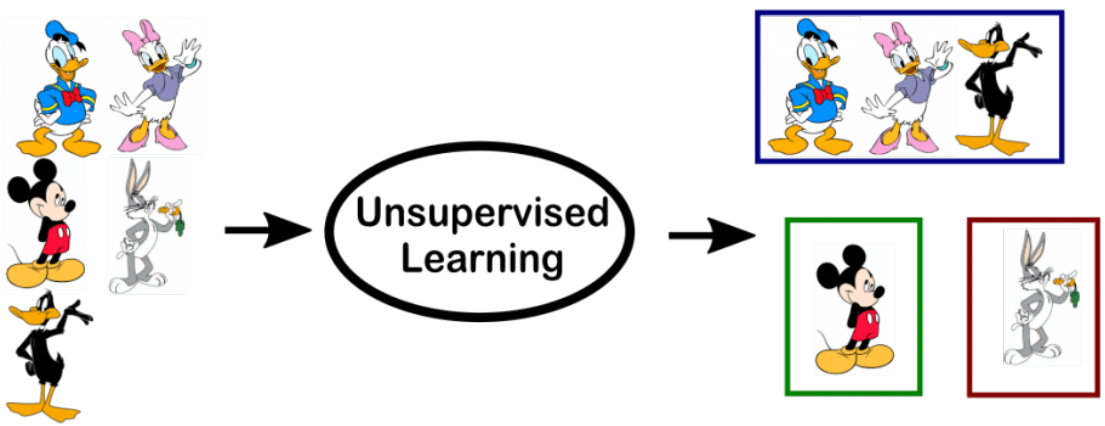


Figure – Exemple d'apprentissage non supervisées

Exemple méthode non supervisée

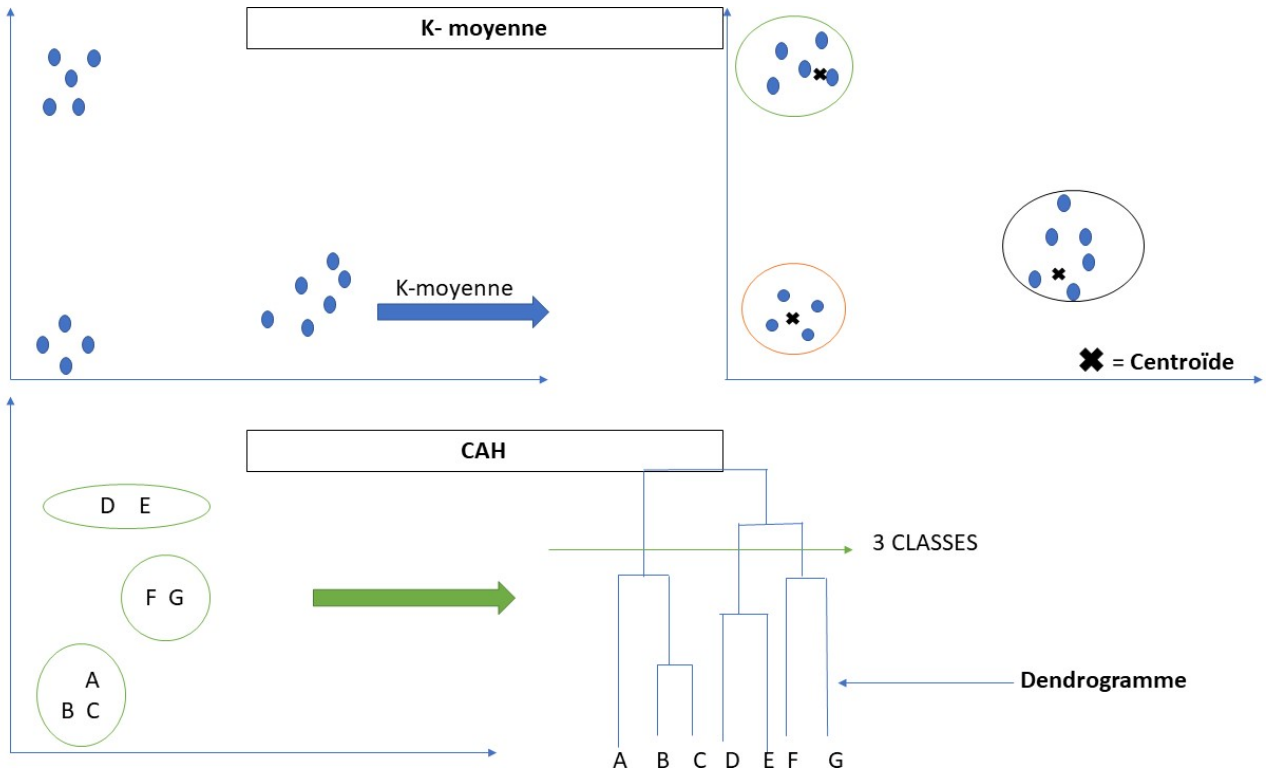


Figure – méthode d'apprentissage non supervisée

Méthode Analyse de composantes principales (ACP)

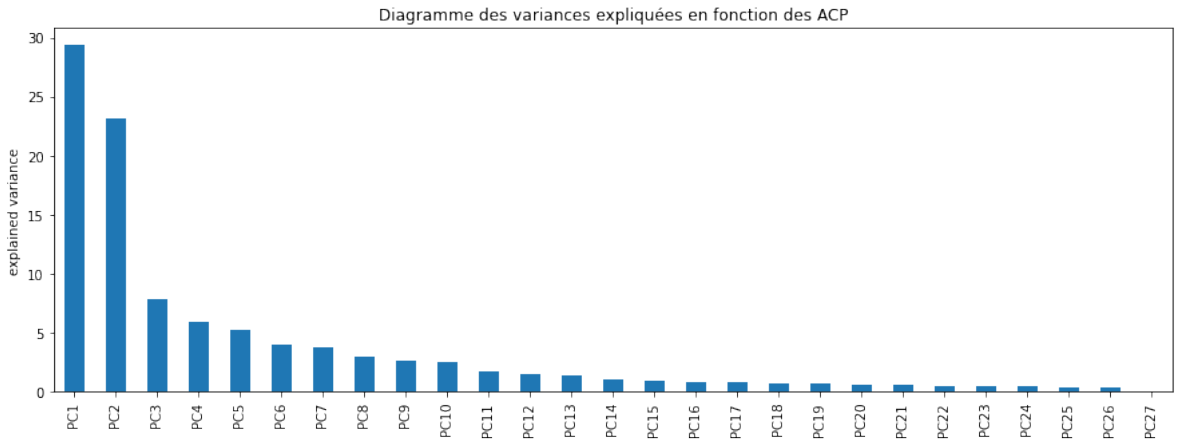


Figure – Diagramme des variances expliquées

Après calcul des variances expliquées on a trouvé : PC1 : 29.34 et PC2 :23.09 en addition celle-ci on se retrouve à 52.44.

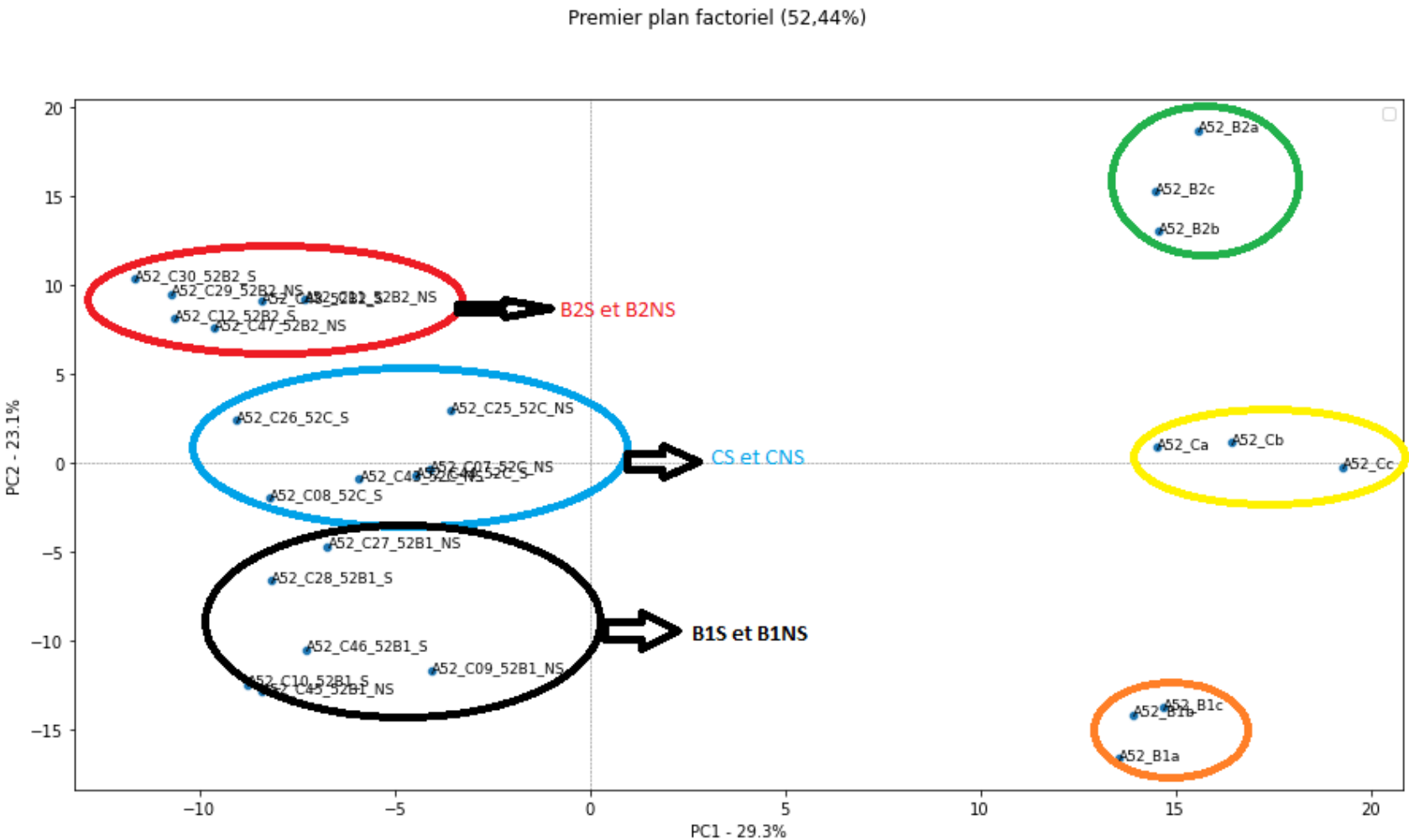


Figure – PCA de substitution

Méthode t-distributed Stochastic Neighbor Embedding

Définition

t-SNE est un algorithme d'apprentissage non supervisée connu notamment pour sa capacité à faciliter la visualisation des **données non linéaires** ayant beaucoup de variables.

Algorithme t-SNE

Etape1 : Grande dimension

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

Avec σ_i est la variance de la gaussienne centrée sur le point de données x_i

Etape 2: Faible dimension

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

Logiquement les probabilités conditionnelles $P_{j|i}$ et $q_j|\mathrm{i}$ doivent être égales pour une représentation parfaite de la similarité des points de données dans les différences espaces dimensionnels.

Etape 3

Pour optimiser cette distribution, t-SNE utilise la divergence de Kullback-Leibler entre entre les probabilités conditionnelles de $P_{j|i}$ et $q_j|\mathrm{i}$

$$c = \sum_{x \in \chi} P_{(x)} \log\left(\frac{P(x)}{Q(x)}\right)$$

La perplexité : $Perp(P_i) = 2^{H(P_i)},$ Avec H l'entropie $H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}.$

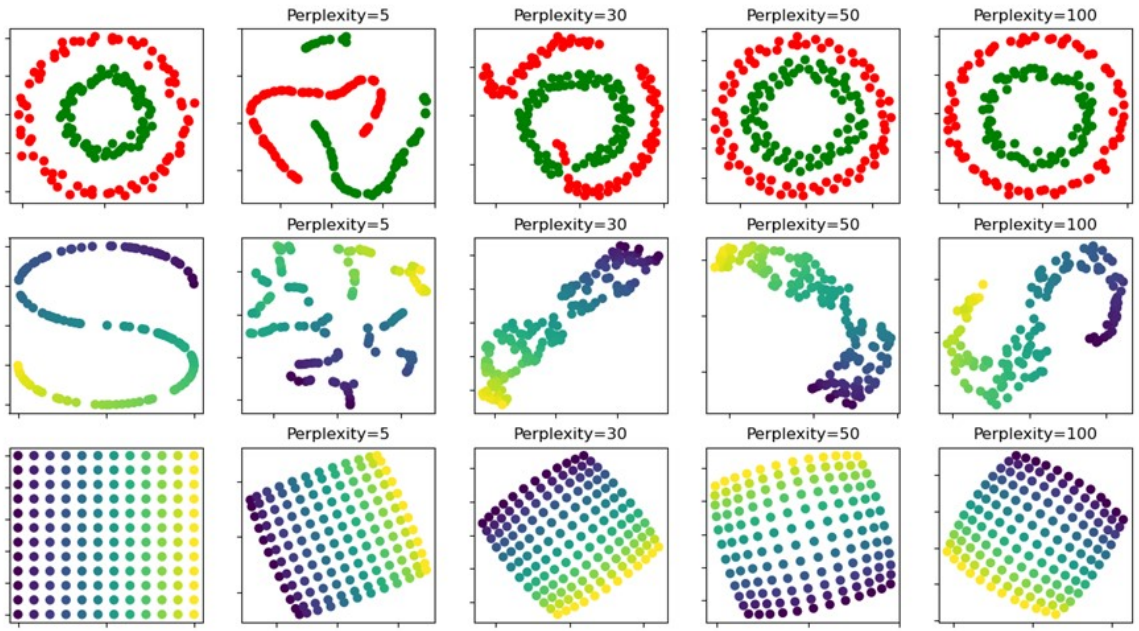


Figure – l'effet de diverses valeurs de perplexité

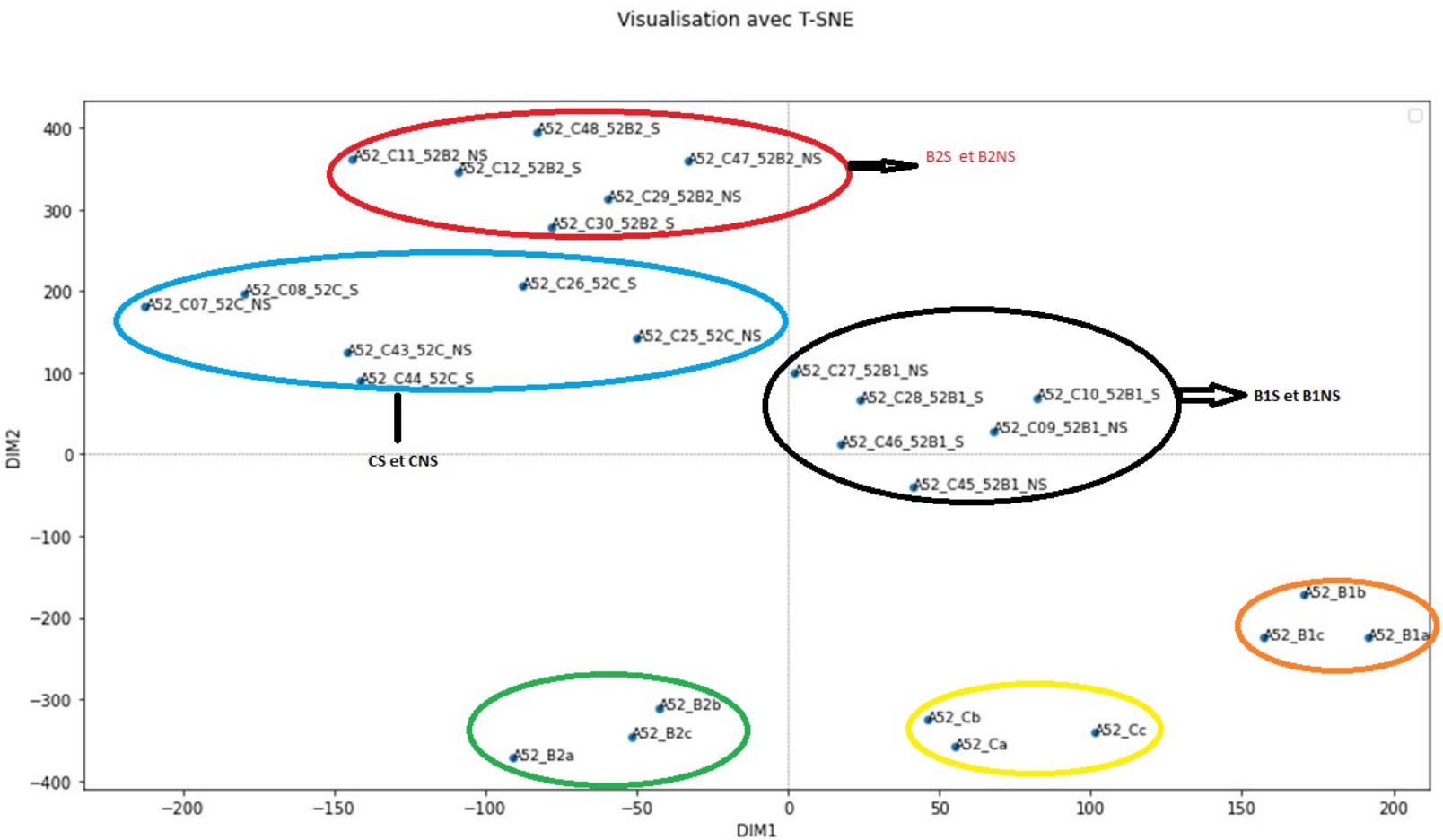


Figure – t-SNE de substitution

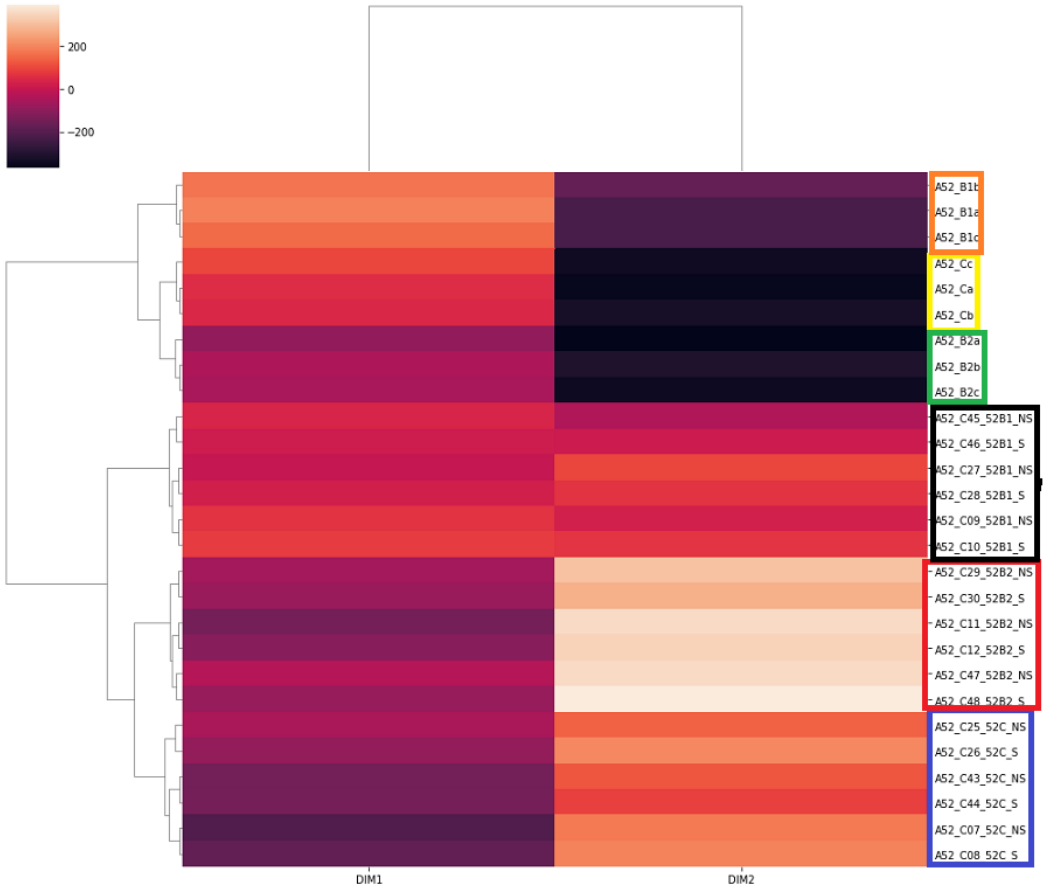


Figure – t-SNE heatmap

☐ Suppressions des valeurs aberrantes

Méthode suppression des valeurs aberrantes :

C'est une méthode comme son nom l'indique supprime les valeurs aberrantes par le biais de la méthode boxplot, qui consiste à supprimer les valeurs qui sont au dessus de *upper* et en dessous *lower*. Après suppression des valeurs aberrantes les variables sont passés 20330 $\longrightarrow$ 4955

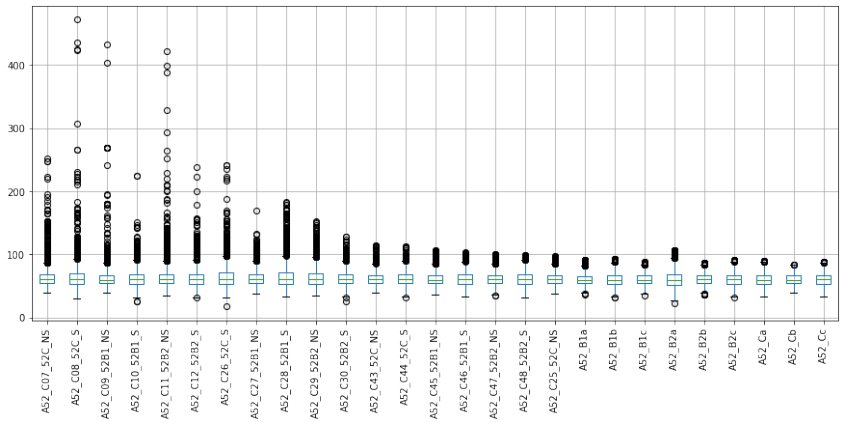


Figure – Boxplot de suppression

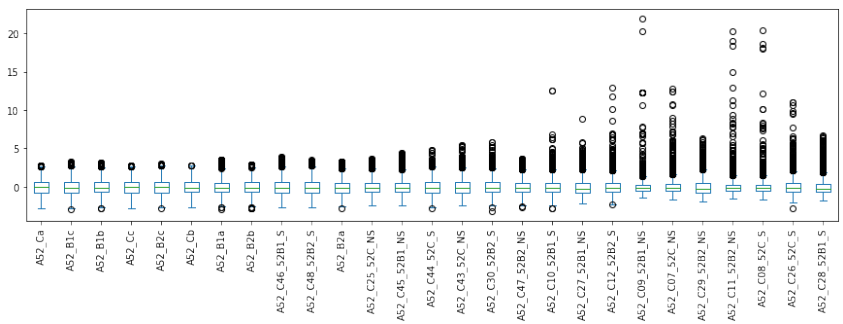


Figure – Boxplot de suppression Normalisé

ACP suppression

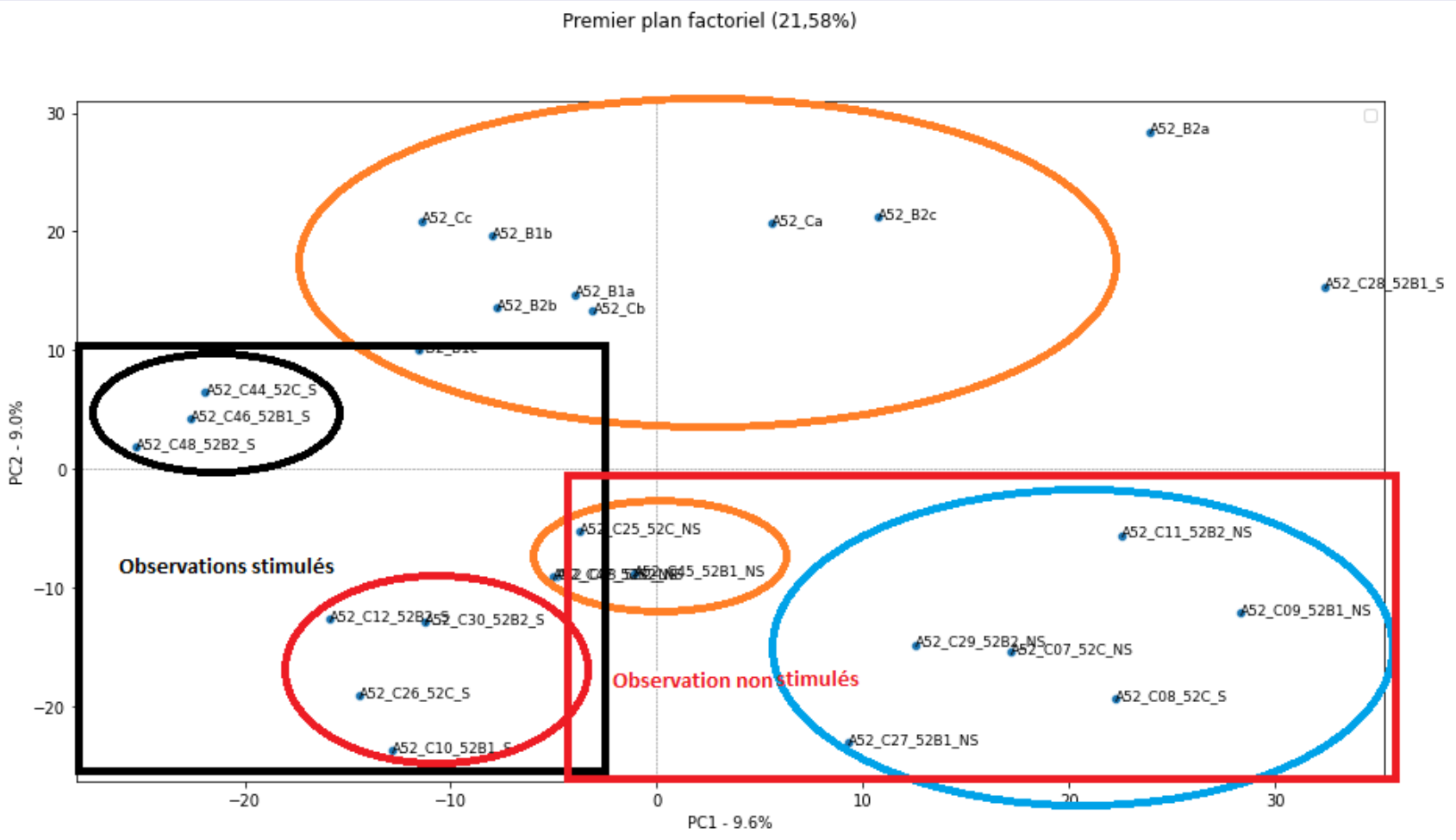


Figure – ACP suppression A52-NormDataAnnot

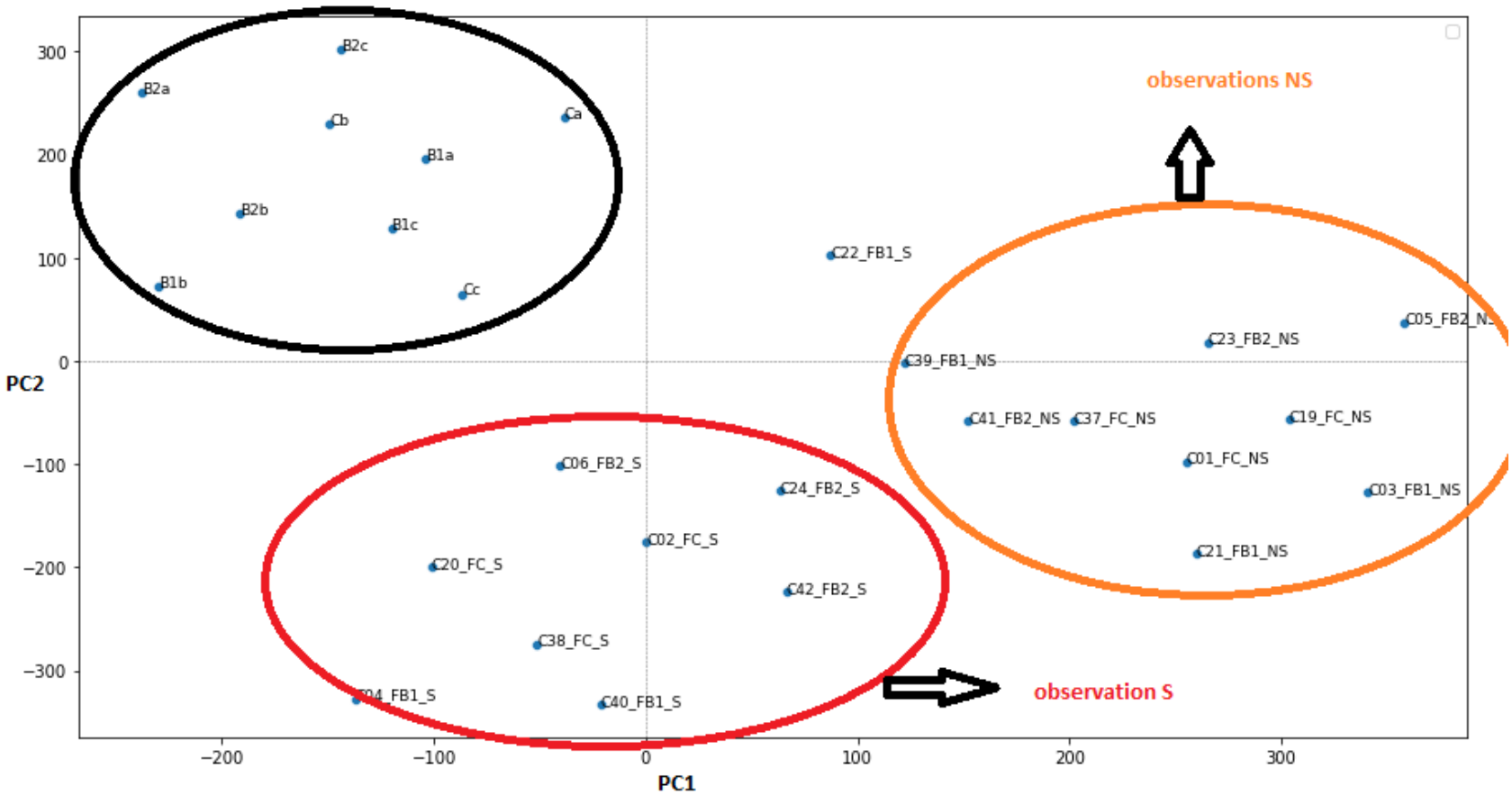


Figure – ACP suppression du jeu de données AF-NormDataAnnot

On voit que dans l'ACP de suppression, les observations sont regroupées en trois (3) clusters.

- les observations stimulées
- les observations non stimulées
- B1a, B1b,B1c, B2a, B2b, B2c,Ca ,Cb et Cc

Apprentissage Profond

Dans l'apprentissage profond, rien n'est programmé explicitement. Il s'agit d'une classe d'apprentissage automatique qui utilise de nombreuses unités de traitement non linéaire pour effectuer l'extraction de caractéristiques ainsi que la transformation.

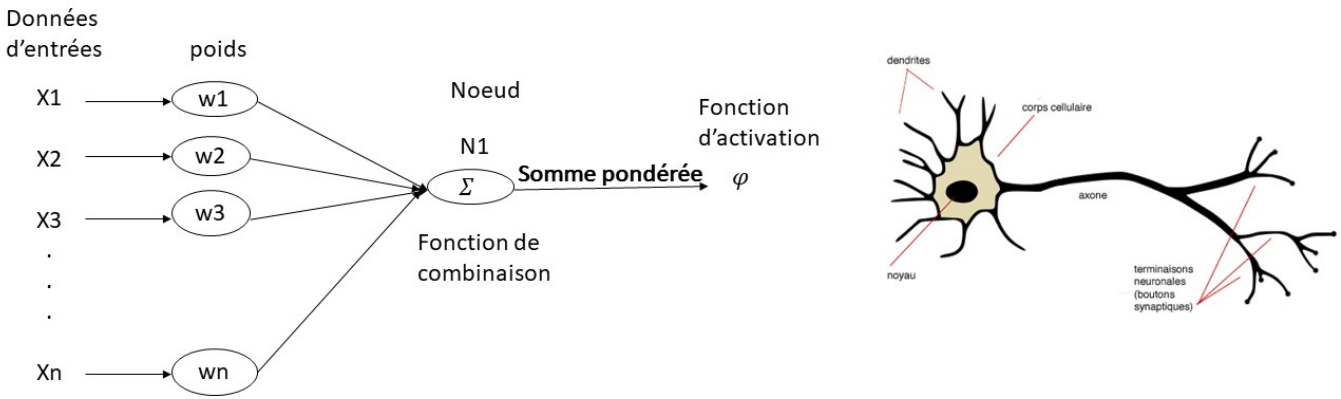


Figure – Neurone artificiel | biologique

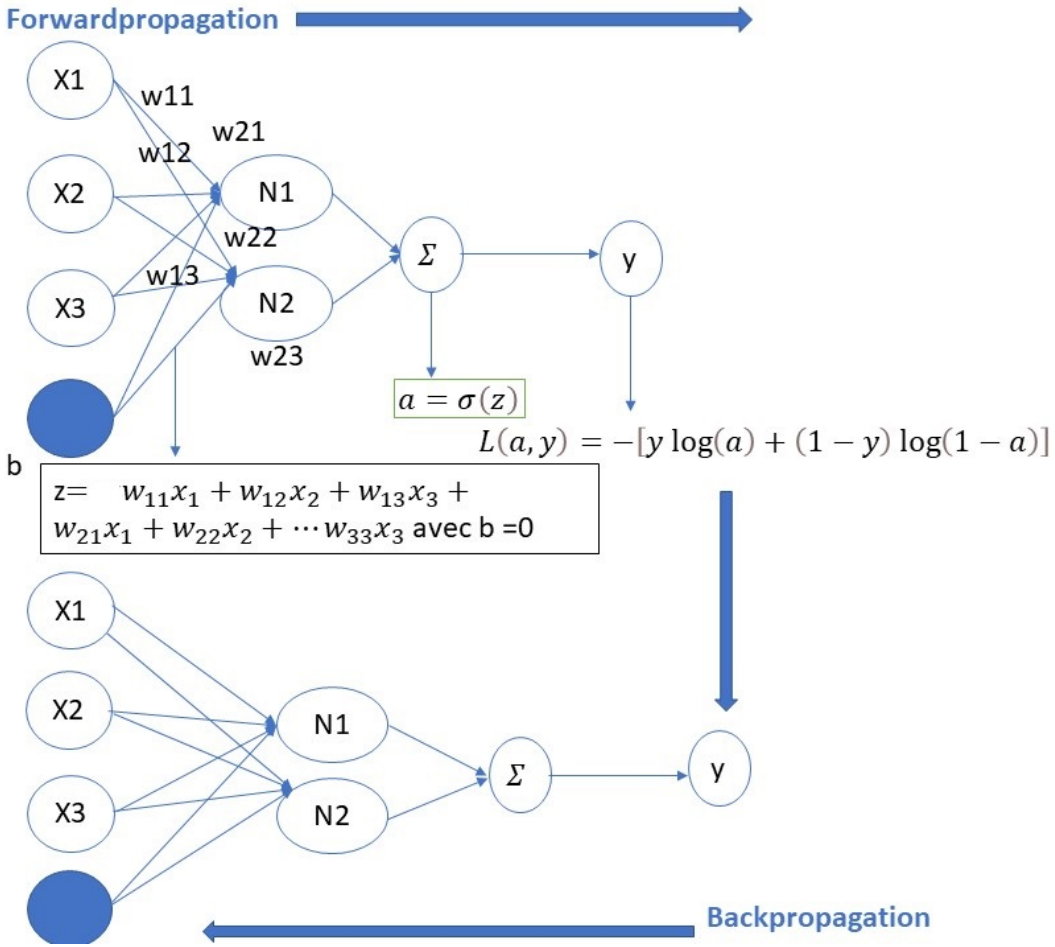
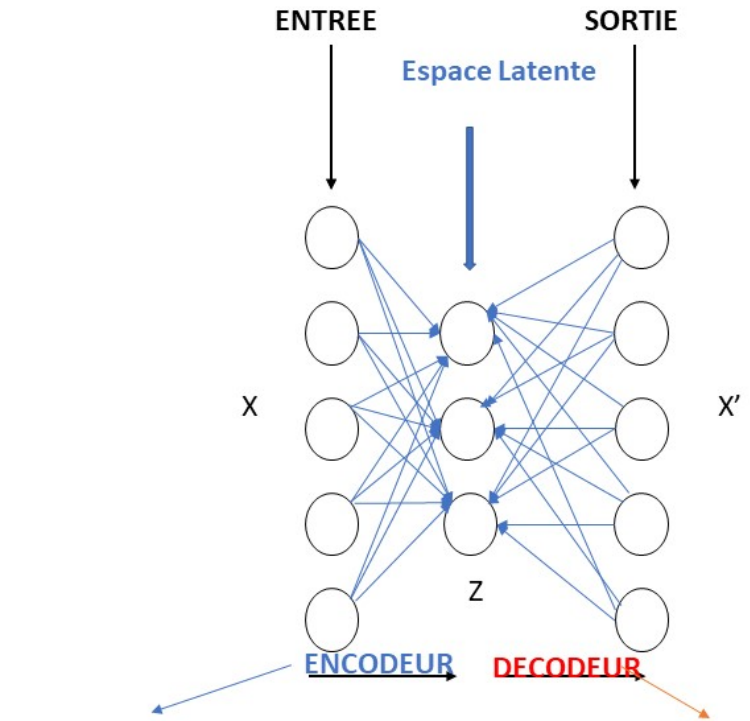


Figure – rétro propagation

Auto encodeur

Définition

Auto encoder (AE) est un réseau de neurones artificiels qui apprend à coder efficacement des représentations pour des données non-étiquetées, puis à utiliser les représentations apprises pour reconstruire l'entrée d'origine aussi proche que possible. Il réduit la dimensionna-lité en entrainant le réseau à extraire des informations utiles et à ignorer le bruit.



$\phi: x \cdot\rightarrow F$

$\psi: F \rightarrow x$

$\phi, \psi = argmin\|x - (\psi - \Phi)x\|$

ou $x = \mathbb{R}^d$ est l'espace ou sont X et X'

$F = \mathbb{R}^P$ est l'espace ou se trouve le code Z

$z = \sigma(wx + b)$

σ est la fonction d'activation
W une matrice de poids
b le biais

$x' = \sigma'(w'z + b')$

$\sigma', w' et\ b'$ du decodeur peuvent différer ou non des σ ,w et b de l'encodeur

Figure – Architecture Autoencodeur

AE se compose de deux parties :

- Encodeur : cette partie comprime l'entrée dans une représentation spatiale qui peut être représentée par la fonction d'encodage

$$h = f(x) \tag{4}$$

- Décodeur : cette partie reconstruite l'entrée à partir de la représentation spatiale potentielle et peut être représentée par la fonction

$$r = g(h) \tag{5}$$

Ainsi l'ensemble de l'AE peut-être décrit par une fonction

$$g(f(x)) = r \tag{6}$$

ou r est la sortie similaire à l'entrée d'origine X.

Description

La fonction d'activation sert avant tout à modifier de manière non-linéaire les données.Dans notre modèle on a utilisé la fonction Rectified Linear Unit(ReLU) pour notre encoder et la fonction Sigmoid pour le décodeur.

$$\text{ReLU}(x) = \max(x, 0) \tag{7}$$

$$\text{sigmoid}(x) = \frac{1}{(1 + e^{(-x)})} \tag{8}$$

Notre modèle est une structure de réseau de neuronal simple à quatre couches : une couche d'entrée qui est notre « Input » , une couche cachée « Hidden layer » une dimension latente et une couche de sortie « Output ».
Avant la construction de notre modèle avec le jeu de données initial (27,20330), nous avons sélectionné 10000 variables les plus exprimés de manières variable par écart absolu médiane (MAD).
MAD = $median(|X_i - median(X)|)$ avec i = 1....n

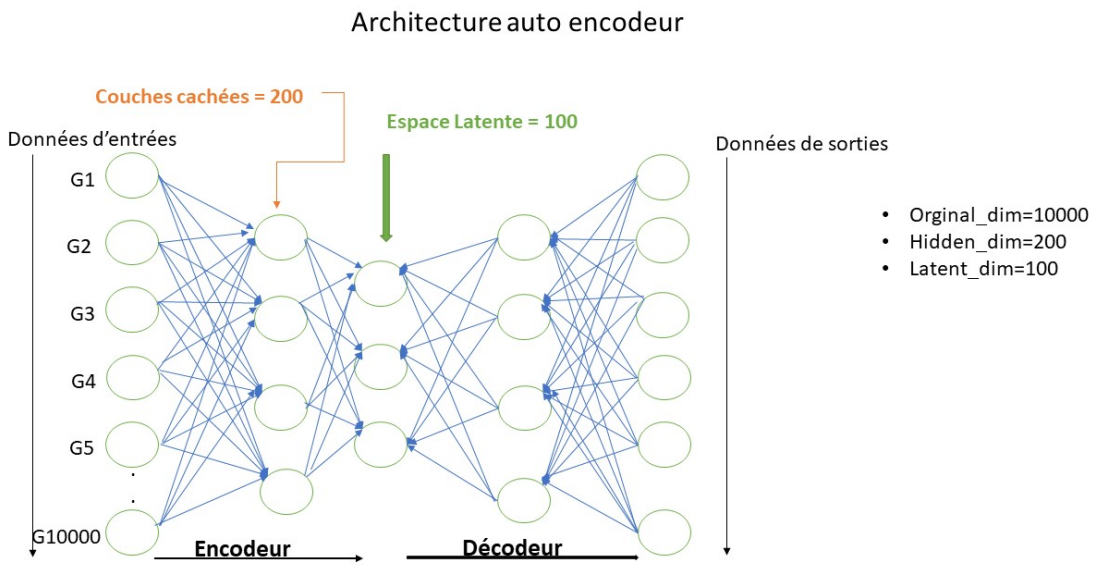


Figure – Architecture modèle

L'étape suivante consiste à compiler le modèle. Pour la compilation, nous avons besoin d'un optimiseur Adam, nous avons opté la fonction perte « mean absolute error ». Nous avons choisis mae car il est robuste aux valeurs aberrantes.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - xi|}{n} \tag{9}$$

avec y_i la prédiction et x_i la vraie valeur.

Nous pouvons représenter la décroissance de la fonction perte en utilisant l'objet 'historique' renvoyer par fit.

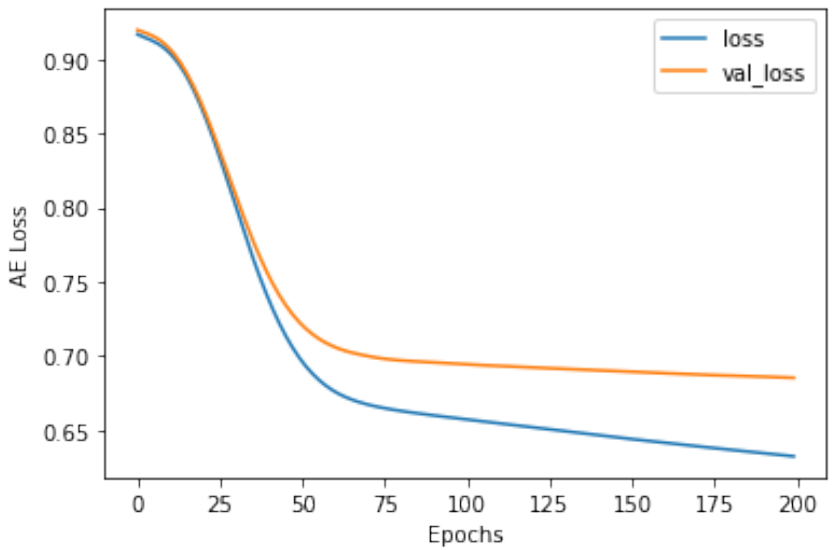


Figure – Modèle Auto-encoder

En effet pour avoir un meilleur modèle on applique des réglages sur les hyper-paramètres (epoch,batch size, latent dim et hidden dim) par la fonction RandomizedSearchCV. Nous avons sélectionné la recherche aléatoire car elle fonctionne plus rapidement qu'une recherche par grille.

Exemple Méthode d'optimisation

- Grid Search
- Random Search
- Hyperband
- Approches bayésiennes

Les meilleurs paramètres trouvés sont :

1. epoch : 200
2. batch size : 32
3. latent dim :70
4. hidden dim : 400

Nous pouvons représenter la décroissance de la fonction perte en utilisant l'objet 'historique' renvoyer par fit.

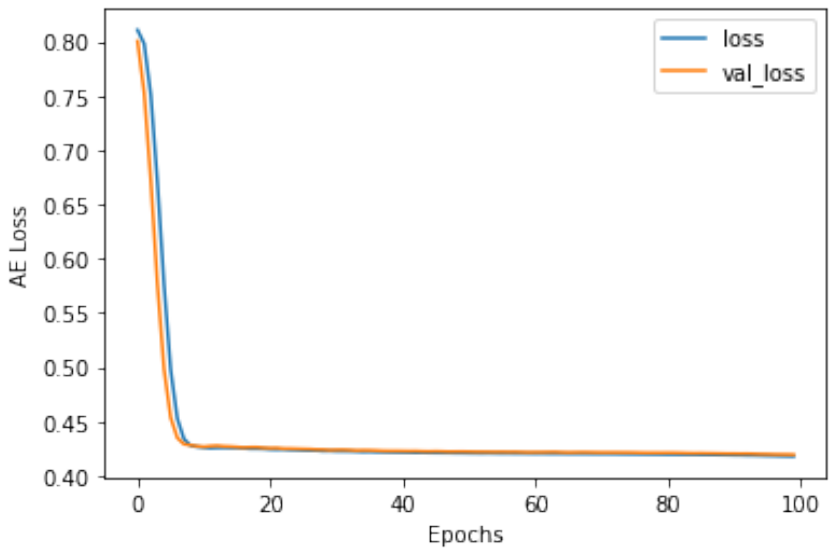


Figure – Modèle Auto-encoder

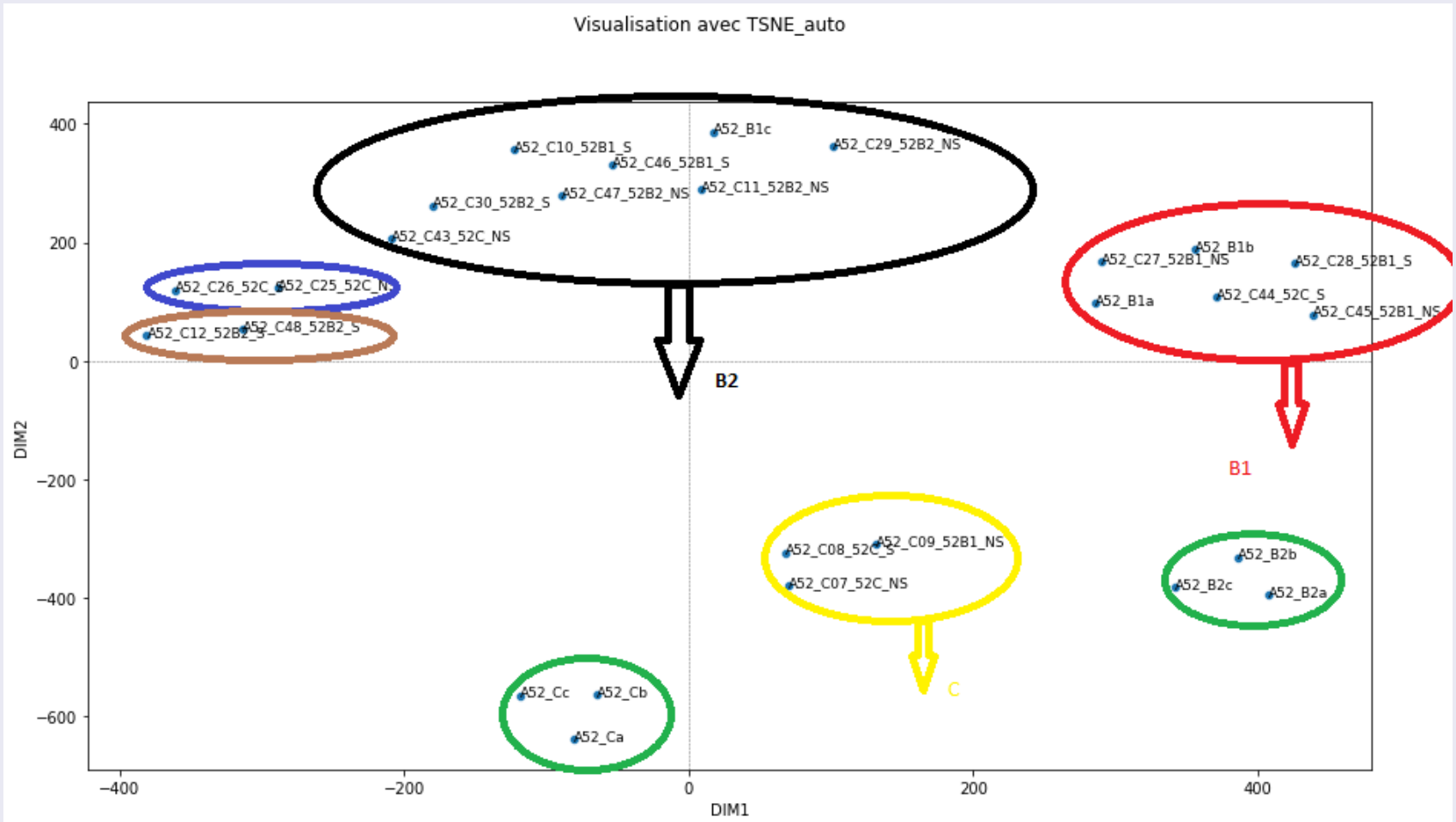
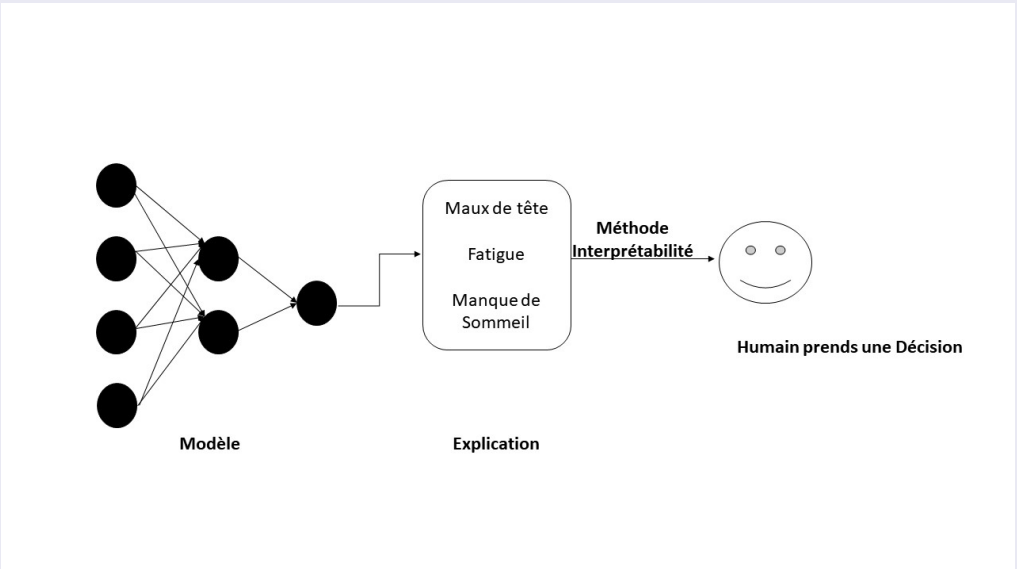


Figure – Visualisation t-SNE encodeur

Interprétation Modèle

Il est difficile de définir mathématiquement l'interprétabilité. L'interprétabilité est le degré auquel un humain peut comprendre la cause d'une décision. Donnons quelques exemples de méthodes d'interprétations.



- lime [1] Lime est capable d'expliquer n'importe qu'elle classificateur de boîte noire avec deux classes ou plus. La technique est de comprendre le modèle en perturbant l'entrée d'échantillon et comprenant comment les prédictions changent.

- Deeplift Fonctionnalité importantes pour l'apprentissage en profondeur. C'est une méthode pour décomposer la prédiction de sortie d'un réseau de neurones sur une entrée spécifique en retro propageant les contribution de tous les neurones du réseau.

- Méthode Identification des meilleurs poids sur notre modèle

- Méthode SHAP

Identification des meilleurs poids

Keras a implémenté certaines fonctions pour obtenir ou définir des poids pour chaque couche. Après on calcule la moyenne de ces poids ainsi on calcule l'écart-type pour sélectionner les meilleurs caractéristiques .

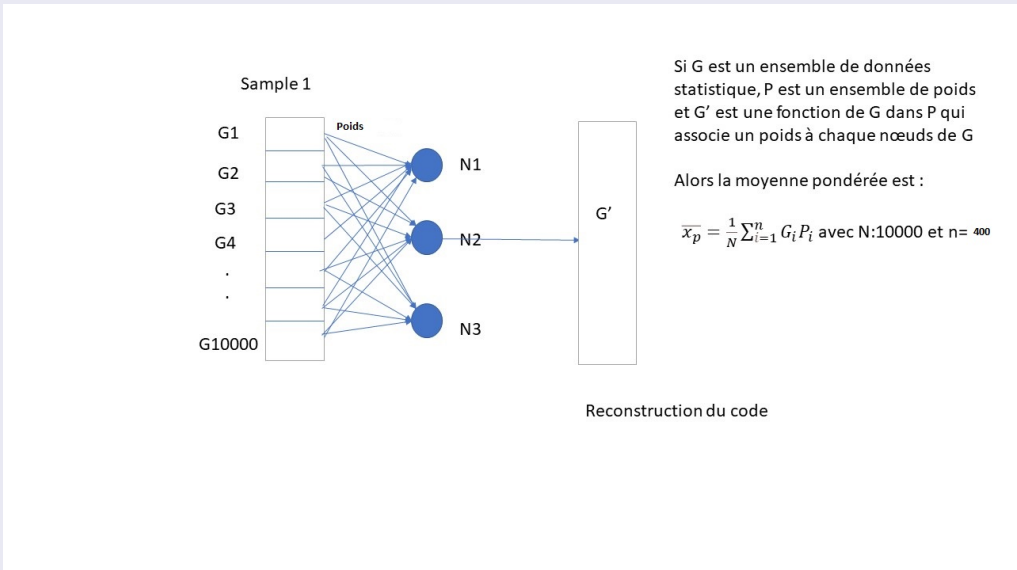


Figure – Identification des meilleurs poids

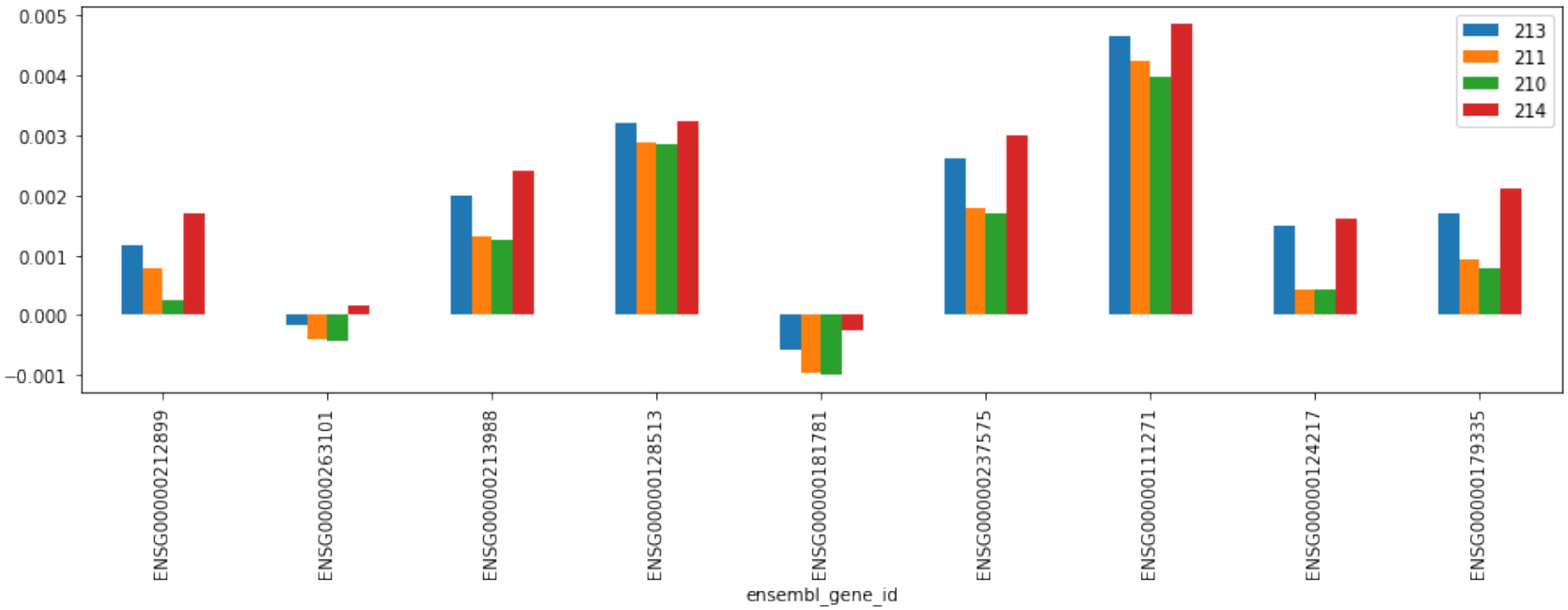


Figure – Keras get weight

Perspective

- Comparaison de l'ACP
- Interprétation de l'espace latente
- Regroupement des gènes
- Interprétation des résultats sur la base de la suppression des données
- Intégration éventuelle avec d'autres ensemble de données

Dans la suite de notre projet on va continuer à interpréter notre modèle par la méthode SHAP.

Méthode SHAP

SHAP (Shapley Additive explanation) exploite l'idée des valeurs de shapley pour la notation de l'influence des caractéristiques du modèle. En d'autres termes,les valeurs de shapley prennent en compte toutes les prédictions possibles pour une instance en utilisant toutes les combinaisons possibles d'entrée.

1. méthode globale : consiste à comprendre la structure globale de la façon dont un modèle prend une décision
2. méthode local : consiste à comprendre comment le modèle à pris des décisions pour une seule instance.


MERCI