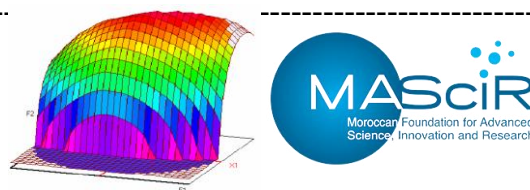


Année Universitaire : 2020-2021



Master Sciences et Techniques CAC Ageq

Chimiométrie et Analyse Chimique : Application à la gestion de la qualité

MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master Sciences et Techniques

Construction des modèles de prédiction de quantité d'azote et de carbone dans les feuilles des plantes de tomates

Présenté par:

Ayoub ZININI

Encadré par:

- Reda RABIE - MASciR
- El Mestafa EL HADRAMI - FST Fès

Soutenu Le 13 Juin 2021 devant le jury composé de :

- El Mestafa EL HADRAMI
- Bouchaib IHSSANE
- Mohammed EL ASRI

Stage effectué à : MAScIR

Remerciements :

Mes remerciements vont premièrement à **Dieu**, le tout puissant pour la volonté, la santé, la patience, et le courage qu'il m'a donné pour terminer ce travail.

Ce travail résulte de la contribution de plusieurs personnes, que ce soit d'un point professionnel m'apportant des connaissances et savoir-faire. Tout d'abord, je tiens à remercier la **société MAScIR** pour le rôle qu'elle joue dans l'intégration des stagiaires au sein de la société. Alors, j'aimerais remercier toutes personnes ayant contribué au bon déroulement de mon stage à cette société.

Mes vifs remerciements vont à mon encadrant Pr. **El Mestafa EL HADRAMI** de la FST Fès pour le suivi continué tout le long de la réalisation de ce projet de fin d'études et qui n'a pas cessé de me donner ses conseils et remarques.

Ma profonde et vive reconnaissance s'adresse aussi aux membres de jury Messieurs les professeurs **Bouchaib IHSSANE** et **Mohammed EL ASRI** d'avoir bien voulu accepter sans réserve de juger ce modeste travail.

Finalement, je remercie notre chef de filière Monsieur **Taoufiq SAFFAJ** pour son support et bonne gestion, et bien sur **tous les professeurs de département de chimie de la FST Fès** pour leurs formations, conseils et soutiens.

Sommaire des titres :

Introduction.....	1
Chapitre I : Brève présentation de la société et quelques rappels sur les outils chimiométriques	
I- Présentation de la société :	2
I.1- Organisme d'accueil :	2
I.2- MAScIR :	2
I.2.1- MAScIR MicroElectronics :	2
I.2.1.1- Missions :	3
I.2.1.2-Laboratoires :	4
I.2.2- MAScIR Biotechnologie :	4
I.2.3- Nanomatériaux et Nanotechnologies :	4
I.2.4- Partenaires de la fondation :	4
II – Définition d'apprentissage automatique :	5
II.1- Apprentissage supervisé :	6
II.2- Apprentissage non supervisé :	6
III – Modélisation :	6
III.1- Régression :	6
III.2- Classification :	6
III.3- Idée générale sur le projet Demandé :	6
IV- Notions sur la source des données et les étapes de préparation :.....	6
IV.1- Spectroscopie :	6
IV.1.1- Spectroscopie proche infrarouge - PIR :	7
IV.1.1.1- PIR-FTIR par transmission (absorption) :	9
IV.1.1.2- PIR-FTIR par réflexion :	10

IV.2- Prétraitements des données obtenues :	10
I.V2.1- Correction log :	11
IV.2.2- Normalisation de signal :	11
IV.2.3- Traitement de Savitsky-Golay :	11
IV.2.4- Detrend :	11
IV.2.5- Dérivé de spectre :	11
IV.2.6- SNV :	11
IV.2.7- SMC :	11
V- Notions de modélisation :	12
V.1- Algorithmes de modélisation :	12
V.2- Les algorithmes utilisés dans ce projet :	12
V.2.1- PLS : Régression à moindres carrées partiels :	12
V.2.2- PCR : Régression sur les composantes principales :	14
V.2.3- MLP : Réseaux de neurones multi couches :	14
V.2.4- KNN : K proches voisins :	16
V.2.1- SVM : Machine à vecteurs de support :	17
V.2.1- DT : Arbre de décision :	18
V.2.1- RFT : Forêt d'arbres décisionnels :	19
VI- Outils de validation de modèle :	20
VI.1- Détermination des relations entre les paramètres :	20
VI.2- Quantification d'erreur :	20
VI.3- Tests de choix de meilleur modèle :	20
Chapitre II : Recherche de modèles de détermination du taux de carbone et de d'azote dans les feuilles de tomates cerise par les méthodes chimiométriques.....	
I- A propos des données :	21

I.1- Présentation des données :	21
I.2- Analyses élémentaires des données :	23
I.2.1- Box plot:	23
I.2.2- Droite d'Henry :	24
I.2.3- Test de Shapiro-Wilk:	25
II- Examen des modèles sur les données prétraitées de Carbone 1 :	26
II.1- Prétraitements effectués :	26
II.2- Examen des performances :	26
II.3- Examen de résidu :	30
II.4- Conclusion :	31
III- Examen des modèles sur les données prétraitées de Carbone 2 :	31
III.1- Prétraitements effectués :	31
III.2- Examen des performances :	31
III.3- Examen de résidu :	35
III.4- Conclusion :	36
IV- Examen des modèles sur les données prétraitées de Carbone 3 :	36
IV.1- Prétraitements effectués :	36
IV.2- Examen des performances :	36
IV.3- Examen de résidu :	40
IV.4- Conclusion :	41
V- Examen des modèles sur les données prétraitées d'Azote 1 :	41
V.1- Type des prétraitements effectués :	41
V.2- Examen des performances :	41
V.3- Examen de résidu :	45

V.4- Conclusion :	46
VI- Examen des modèles sur les données prétraitées d'Azote 2 :	46
VI.1- Type des prétraitements effectués :	46
VI.2- Examen des performances :	46
VI.3- Examen de résidu :	50
VI.4- Conclusion :	51
VII- Conclusion générale :	52
Références :	53

Sommaire des Tableaux :

<u>Tableau 1 :</u> Une prévision des données de Carbone.....	22
<u>Tableau 2 :</u> Une prévision des données d’Azote.....	22
<u>Tableau 3 :</u> Résultats d’examen de performances effectué sur les données de Carbone prétraitées 1.....	26
<u>Tableau 4 :</u> Résultats d’examen de normalité de résidus effectué sur les données de Carbone prétraitées 1.....	30
<u>Tableau 5 :</u> Résultats d’examen de performances effectué sur les données de Carbone prétraitées 2.....	31
<u>Tableau 6 :</u> Résultats d’examen de normalité de résidus effectué sur les données de Carbone prétraitées 2.....	35
<u>Tableau 7 :</u> Résultats d’examen de performances effectué sur les données de Carbone prétraitées 3.....	36
<u>Tableau 8 :</u> Résultats d’examen de normalité de résidus effectué sur les données de Carbone prétraitées 3.....	40
<u>Tableau 9 :</u> Résultats d’examen de performances effectué sur les données d’Azote prétraitées 1.....	41
<u>Tableau 10 :</u> Résultats d’examen de normalité de résidus effectué sur les données d’Azote prétraitées 1.....	45
<u>Tableau 11 :</u> Résultats d’examen de performances effectué sur les données d’Azote prétraitées 2.....	46
<u>Tableau 12 :</u> Résultats d’examen de normalité de résidus effectué sur les données d’Azote prétraitées 2.....	50

Sommaire des figures :

<u>Figure 1 :</u> Les branches importantes d'apprentissage automatique des machines.....	5
<u>Figure 2 :</u> Les intervalles des longueurs d'ondes.....	7
<u>Figure 3 :</u> Montage de la méthode PIR-FTIR.....	8
<u>Figure 4 :</u> Mécanisme d'interféromètre de Michelson.....	8
<u>Figure 5 :</u> L'interférogramme à la sortie du détecteur.....	9
<u>Figure 6 :</u> Mécanisme de mesure de transmittance.....	9
<u>Figure 7 :</u> Phénomène de réflectance diffuse.....	10
<u>Figure 8 :</u> Les composantes d'un perceptron.....	15
<u>Figure 9 :</u> Exemples de fonctions d'activation.....	15
<u>Figure 10 :</u> Les composantes d'un réseau MLP	16
<u>Figure 11 :</u> Les étapes d'algorithme KNN.....	16
<u>Figure 12 :</u> Les composantes d'un algorithme SVM.....	17
<u>Figure 13 :</u> Présentation générale d'un arbre de décision.....	18
<u>Figure 14 :</u> Exemple d'arbre de décision.....	19
<u>Figure 15 :</u> Les étapes de Forêt d'arbres décisionnels.....	19
<u>Figure 16 :</u> Spectre de réflectance des données de Carbone.....	22
<u>Figure 17 :</u> Spectre de réflectance des données d'Azote.....	22
<u>Figure 16 :</u> Spectres de réflectance des données de Carbone avant et après un exemple de prétraitement - prétraitement effectué : MSC + Savitsky-Golay	22
<u>Figure 17 :</u> Spectres de réflectance des données d'Azote avant et après un exemple de prétraitement - prétraitement effectué : Moving average + GAP1 + MSC	23
<u>Figure 18 :</u> Box plot de réponse des données de Carbone.....	23
<u>Figure 19 :</u> Box plot de réponse des données d'Azote.....	24

<u>Figure 20</u> : Plot d'Henry des données de Carbone.....	25
<u>Figure 21</u> : Plot d'Henry des données d'Azote.....	25
<u>Figure 22</u> : Résultats de test Shapiro-Wilk sur les données de Carbone.....	26
<u>Figure 23</u> : Résultats de test Shapiro-Wilk sur les données d'Azote.....	26
<u>Figure 24</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle PLS.....	27
<u>Figure 25</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle PCR.....	27
<u>Figure 26</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle SVR.....	28
<u>Figure 27</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle RFT.....	28
<u>Figure 28</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle DT.....	29
<u>Figure 29</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle KNN.....	29
<u>Figure 30</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 1 - modèle MLP.....	30
<u>Figure 31</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle PLS.....	32
<u>Figure 32</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle PCR.....	32
<u>Figure 33</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle SVR.....	33
<u>Figure 34</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle RFT.....	33
<u>Figure 35</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle DT.....	34
<u>Figure 36</u> : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle KNN.....	34

<u>Figure 37 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de carbone 2 - modèle MLP.....	35
<u>Figure 38 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle PLS.....	37
<u>Figure 39 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle PCR.....	37
<u>Figure 40 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle SVR.....	38
<u>Figure 41 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle RFT.....	38
<u>Figure 42 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle DT.....	39
<u>Figure 43 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle KNN.....	39
<u>Figure 44 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone 3 - modèle MLP.....	40
<u>Figure 45 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle PLS.....	42
<u>Figure 46 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle PCR.....	42
<u>Figure 47 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle SVR.....	43
<u>Figure 48 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle RFT.....	43
<u>Figure 49 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle DT.....	44
<u>Figure 50 :</u> Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote 1 - modèle KNN.....	44

<u>Figure 51 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 1 - modèle MLP.....	45
<u>Figure 52 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle PLS.....	47
<u>Figure 53 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle PCR.....	47
<u>Figure 54 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle SVR.....	48
<u>Figure 55 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle RFT.....	48
<u>Figure 56 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle DT.....	49
<u>Figure 57 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle KNN.....	49
<u>Figure 58 :</u> Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote 2 - modèle MLP.....	50

Tableau des abréviations :

Abréviation	Nom complet
PLS	Partial least squared regression
PCR	Principal components regression
SVR	Support vector machine regressor
KNN	K nearest neighbors
MLP	Multi layers perceptron
DT	Decision tree
RFT	Random forest
MSE	Mean squared error
RMSE	Root mean squared error
ME	Max error
PIR-FTIR (ou : NIR-FTIR)	Near infrared with Fourier transform
SNV	Standard normal variation
MSC	Multiplicative scatter correction
CV	Cross validation
RPD	Ratio of performance to deviation
AIC	Akaike information criterion
BIC	Bayesian information criterion
Corr	Correlation

Introduction

L'expérience moderne cherche à réduire le temps, les équipements, le coût, les expériences... C'est une victoire du monde scientifique après le développement des méthodes de modélisation et prédiction des réponses désirées avec un gain de temps et de coût importants. L'apprentissage moderne des machines est basé sur des méthodes statistiques/mathématiques et informatiques développées pendant plusieurs années du dernier siècle, et sont maintenant de plus en plus applicables dans plusieurs domaines. En chimie, la modélisation et les méthodes d'apprentissage sont basées dans la majorité des cas sur les méthodes spectrales, ce qui fait gagner du temps par rapport à l'utilisation des méthodes de titrage et les méthodes gravimétriques, conductimétriques..., et minimise le coût aussi dans le cas de quelques méthodes analytiques en particulier les méthodes basées sur les rayonnements infrarouges.

L'objectif de ce projet est l'utilisation de ces outils chimiométriques lors de la détermination du taux de carbone et d'azote dans les feuilles de tomates cerise. Notre travail sera basé sur les méthodes de traitement des signaux infrarouge enregistrés à partir de plusieurs expériences appliquées sur des échantillons de feuilles de tomates cerise collectées de la région d'Agadir. Les données enregistrées sont : la réflectance des échantillons dans la zone spectrale comprise entre 350 et 2500 cm^{-1} . La méthode de référence utilisée pour déterminer les quantités de Carbone et Azote totales est la méthode Kjeldahl.

Ce manuscrit sera organisé comme suit :

- Chapitre 1 : présentation de la société où le stage s'est déroulé et rappel sur les méthodes chimiométriques utilisées
- Chapitre II : Recherche de modèles de détermination du taux de carbone et de d'azote dans les feuilles de tomates cerise par les méthodes chimiométriques

CHAPITRE 1 : BRÈVE
PRÉSENTATION DE LA
SOCIÉTÉ ET QUELQUES
RAPPELS SUR LES OUTILS
CHIMIOMÉTRIQUES

Chapitre I : Brève présentation de la société et quelques rappels

sur les outils chimiométriques

I- Présentation de la société :

I.1- Organisme d'accueil :

Avant de présenter et de développer notre sujet en profondeur, il paraît, dans un premier temps, opportun de commencer par une présentation générale de la société, de survoler rapidement son historique, afin de mieux développer le sujet et la problématique traitée lors de ce travail.

I.2- MAScIR :

MAScIR (Moroccan foundation for Advanced Science, Innovation and Research) est un organisme de recherche à caractère scientifique et technologique. Il est voué à la recherche en nanotechnologie, en biotechnologie, en technologie numérique, en microélectronique, en énergie et en environnement, la fondation se veut présenter là où les enjeux de la société l'exigent. Rassemblant d'éminents chercheurs des quatre coins du monde, MAScIR regroupe des équipes scientifiques œuvrant dans des domaines innovants et complémentaires et met à leur disposition des instruments scientifiques de pointe. Initialement fondée en 2007 par le Gouvernement Marocain en tant que fondation à but non lucratif.

- MAScIR MicroElectronics : créé vers la fin de l'année 2008, a pour objectif de devenir un centre de Recherche et Développement dans le domaine de la microélectronique.
- MAScIR Biotechnology : deuxième centre inscrit dans MAScIR œuvrant dans le domaine de la biotechnologie : recherche et développement des médicaments ou des biocides.
- Nano Technology : qui a pour mission de mener des recherches appliquées, innovantes et à la fine pointe de la technologie pour créer de la propriété intellectuelle et des prototypes dans le domaine des nanomatériaux et des nanotechnologies. Ces recherches sont menées par une équipe internationale de haut calibre travaillant dans un environnement unique et utilisant une infrastructure de pointe.

I.2.1- MAScIR MicroElectronics :

MAScIR MicroElectronics est un centre d'innovation et de développement des technologies dans le milieu microélectronique. Il se concentre sur le micro packaging, l'ingénierie, les tests

de simulation, le design, la qualification, le prototypage de produits micro-électroniques et les systèmes embarqués. Parmi les projets traités par MAScIR MicroElectronics, nous citons :

- Design et Micro packaging CSP et PILR,
- Tests de fiabilité sur les packages
- Circuits embarqués sur une application wafer level camera fabriquée au Maroc par Nemotek Technologie.

MAScIR MicroElectronics fait partie de la fondation à but non lucratif MAScIR dont les domaines de recherche sont la Nanotechnologie, la Microélectronique, les énergies, et l'environnement.

La mission de MAScIR est de promouvoir l'excellence dans la recherche et le développement des technologies au MAROC dans le but de générer de la valeur, des emplois à travers l'intégration dans le marché des technologies avancées dans le but de produire de la propriété intellectuelle et de créer des spin-offs.

MAScIR MicroElectronics offre ces services aux clients industriels et partenaires académiques dans bon nombre de domaines :

- Prototypage de Circuits Intégrés incluant le design de packages et la modélisation, l'assemblage et la qualification.
- Systèmes embarqués.
- Management de projets.

I.2.1.1- Missions :

La mission du MAScIR MicroElectronics est de devenir un centre d'excellence en Microélectronique et en Miniaturisation incluant l'électronique Digitale et Analogique, le Logiciel, l'Optique et les Technologies de packaging.

MAScIR MicroElectronics fournit des services pour des clients industriels, mais elle développe aussi son propre business dans les domaines suivants :

- L'intégration et la miniaturisation des systèmes microélectroniques.
- L'analyse de fiabilité et défaillance des produits.
- Modélisation des systèmes complexes
- Prototypage et industrialisation des produits innovants.
- Industrialisation des idées et résultats académiques.

I.2.1.2-Laboratoires :

MAScIR possède plusieurs laboratoires équipés de technologie avancée :

- Chambre blanche
- Laboratoire optique
- Laboratoire électronique

I.2.2- MAScIR Biotechnologie :

Le Centre de Biotechnologie de MAScIR combine une expertise scientifique de classe mondiale avec des laboratoires de recherche pour offrir des solutions efficaces innovantes et économiques afin de répondre à la demande du marché et aux besoins sociaux du Maroc et de sa région.

L'équipe de la biotechnologie médicale est principalement axée sur le développement de kits de diagnostics médicaux innovants ciblant les maladies les plus répandues au Maroc telles que le cancer et les maladies infectieuses. L'équipe s'intéresse également au co-développement des prototypes biogénétiques en collaboration avec des partenaires industriels.

L'équipe de la biotechnologie verte se concentre sur des projets dans le domaine de la recherche appliquée des plantes, y compris l'exploitation, l'amélioration et les nouvelles technologies. Les principaux domaines de recherche sont :

- Bioénergie à base de microalgues
- Crop Breeding
- Biotechnologie microbienne appliquée au développement agricole et alimentaire

I.2.3- Nanomatériaux et Nanotechnologies :

Des chercheurs de l'Institut de MAScIR des nanomatériaux et les nanotechnologies travaillent au niveau minuscule d'un nanomètre, où la matière montre des propriétés physiques, chimiques inhabituelles.

La mission du centre est de concevoir, d'inventer et d'encourager la recherche et le développement des applications de la nanotechnologie, favorisant ainsi le transfert de technologie à l'industrie afin de répondre à la demande du marché.

I.2.4- Partenaires de la fondation :

Le modèle de management de MAScIR ainsi que le savoir-faire et le professionnalisme de ses équipes scientifiques ont permis à la Fondation d'asseoir sa notoriété auprès des industriels et renforcer son éventail de partenaires industriels tant à l'échelle nationale qu'internationale. Les principaux partenaires de la fondation MAScIR Microélectroniques sont :

- Lear Corporation : L'un des principaux fournisseurs mondiaux de sièges automobiles et les systèmes de gestion de l'énergie électrique.
- Thales : figure parmi les leaders européens de la fabrication et de la commercialisation d'équipements et de systèmes électroniques destinés aux secteurs de l'aérospatial, du transport, de la défense et de la sécurité.
- OCP : est un acteur incontournable sur le marché des phosphates et de ses produits dérivés. Présent sur toute la chaîne de valeur, il est le premier exportateur de cette matière dans le monde.
- COSUMAR : est un groupe marocain, filiale de la Société nationale d'investissement, spécialisé dans l'extraction, le raffinage et le conditionnement du sucre sous différentes formes. Il est devenu l'unique opérateur sucrier marocain après l'acquisition de SUTA, SUCRAFOR, SUNABEL et SURAC en 2005.

II– Définition d'apprentissage automatique :

L'apprentissage automatique est un moyen pour rendre les machines apprennent se qu'il faut faire sans les programmer d'une manière classique. Cette technique est une combinaison entre la programmation classique, la recherche traditionnelle et les lois statistiques/mathématiques. La figure suivante montre les types d'apprentissage automatique. [1]

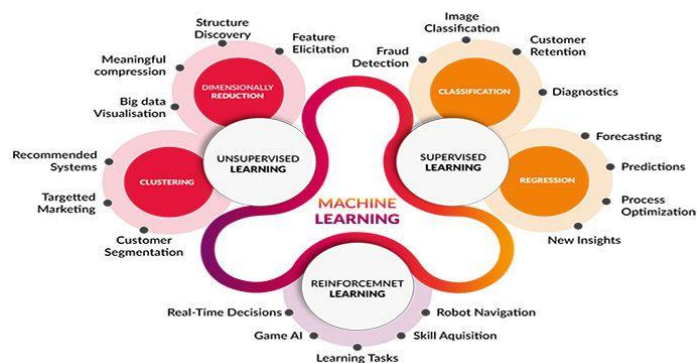


Figure 1 : Les branches importantes d'apprentissage automatique des machines.

On va se concentrer sur les deux branches les plus présentes dans le domaine chimique : Apprentissage supervisé et non supervisé.

II.1- Apprentissage supervisé :

C'est un type d'apprentissage contenant des données à réponses avec des labels c'est-à-dire une réponse d'entrée étalonnée. Ce type d'apprentissage est utile dans les problèmes de classification comme le traitement des images, prédictions liées aux sujets d'environnement, économie..., et quelques problèmes de régression. [1]

II.2- Apprentissage non supervisé :

Au contraire d'apprentissage supervisé, l'apprentissage non supervisé est appliqué sur des données d'entrée non étalonnée et réponses sans labels. Cette technique est utile aux domaines physiques/chimiques comme les traitements des signaux spectrales, analyse vocales..., dont la majorité des problèmes sont des régressions. [1]

III – Modélisation :

La modélisation est une étape majeure dans le processus d'apprentissage qui crée un lien direct et principal entre les données et la décision de sortie d'appareil. Cette modélisation est basée sur la création d'une relation binaire entre les données d'entrée et de sortie selon des algorithmes de régression ou de classification selon le type de problème et des données collectées.

III.1- Régression :

Dans ce type de modélisation : la réponse est une valeur simple calculée à partir d'un ensemble des données uni ou multi-dimensionnels, soit avec utilisation des coefficients (régression paramétrique) ou sans utilisation des coefficients (régression non paramétrique). [2]

III.2- Classification :

Cette modélisation est basée sur le même enchaînement des données d'entrée, mais la sortie est à forme de classes. Cette réponse est déterminée à partir des régressions non paramétriques.

IV- Notions sur la source des données et les étapes de préparation :

IV.1- Spectroscopie :

La spectroscopie est une technique d'analyse de type interactions matière-lumière basée sur les mesures de grandeurs physiques excitées par les ondes électromagnétiques pour quantifier la composition d'analyte. Cette technique est divisée en plusieurs types selon les intervalles des fréquences ou des longueurs d'ondes définies :

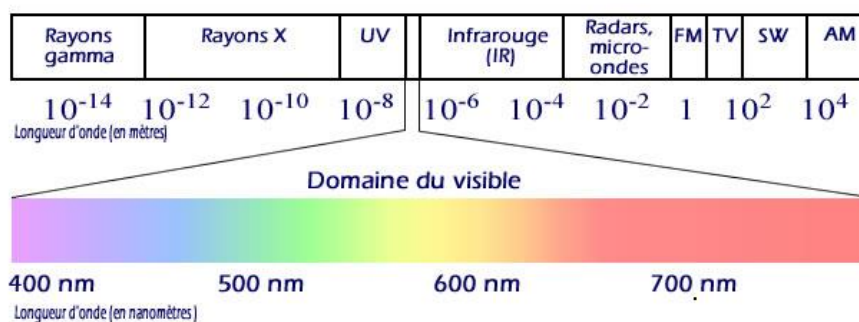


Figure 2 : Les intervalles des longueurs d'ondes

La méthode utilisée dans ce projet est incluse dans le domaine infrarouge, s'appelle PIRFTIR (Proche infrarouge : transformation de Fourier infrarouge).

IV.1.1- Spectroscopie proche infrarouge - PIR :

La spectroscopie proche infrarouge est une technique concentrée sur l'intervalle optique qui commence à **13000** et fini à **4000 cm^{-1}** .

C'est la première plage en relation avec les vibrations moléculaires. Elle présente un grand intérêt pratique pour les applications analytiques.

Dans ce projet, on a utilisé le type PIR-FTIR :

Le spectromètre à transformée de Fourier (PIR-FTIR) [3] de type non dispersif a été développé dans les années 1960. Il comporte les éléments suivants : une source de radiations IR, l'interféromètre de Michelson, un compartiment d'échantillon, un détecteur, un amplificateur, le convertisseur analogique numérique qui interroge le détecteur à des intervalles réguliers et transforme le signal analogique en un signal numérique manipulable par le système informatique et un ordinateur. La radiation PIR passe à travers l'échantillon après l'interféromètre et atteint le détecteur. Puis le signal est amplifié et converti en un signal digital par l'amplificateur et le convertisseur analogique numérique, respectivement. Finalement, le signal est transféré à un ordinateur où la transformée de Fourier est calculée.

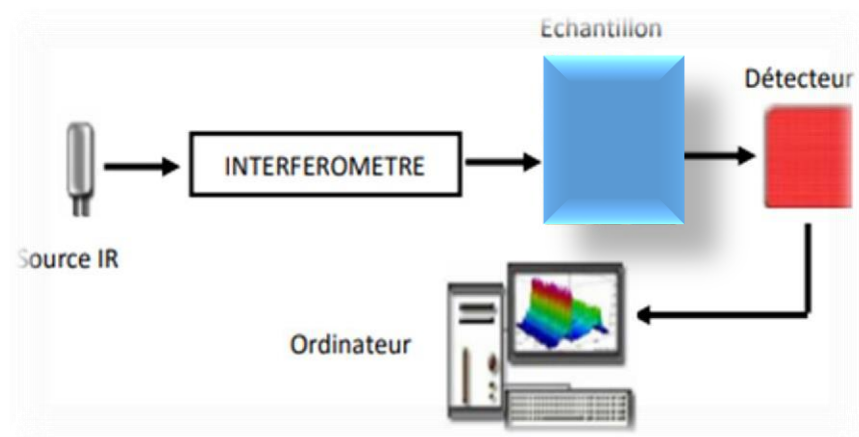


Figure 3 : Montage de la méthode PIR-FTIR

L'interféromètre de Michelson, qui est une pièce essentielle du spectromètre FT-IR, comporte deux miroirs perpendiculaires et un séparateur de faisceau. Un de ces miroirs est fixe (stationnaire) et l'autre est mobile. Le séparateur de faisceau est désigné pour transmettre la moitié de l'intensité lumineuse et réfléchir l'autre moitié. Par la suite, la lumière transportée et la lumière réfléchie frappent le miroir stationnaire et le miroir mobile, respectivement. Quand ils sont réfléchis par les deux miroirs, les deux faisceaux de lumière se recombinent ensemble au niveau du séparateur de faisceaux de manière cohérente. Ces faisceaux combinés sont envoyés vers l'échantillon. Le signal sortant de l'interféromètre créé par ces faisceaux combinés est enregistré par le détecteur pour constituer un interférogramme.

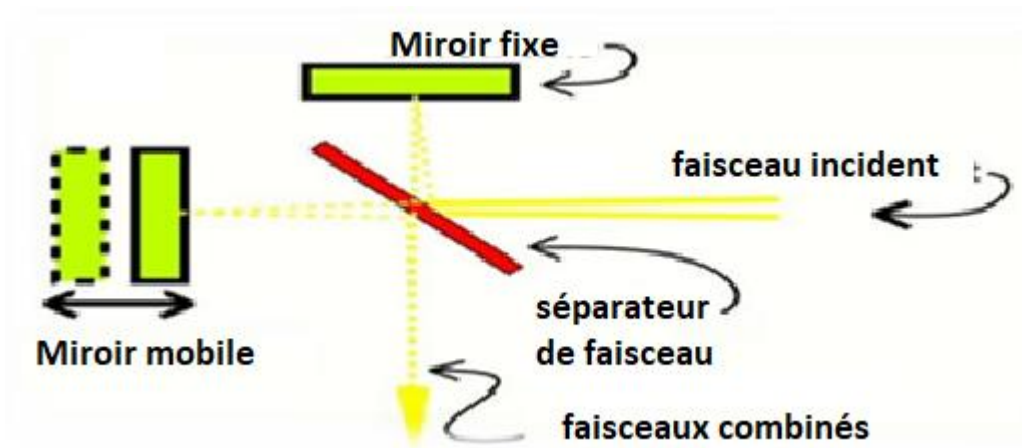


Figure 4 : Mécanisme d'interféromètre de Michelson

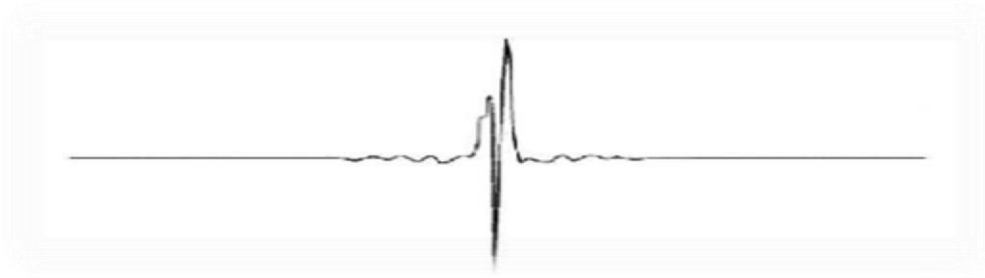


Figure 5 : L'interférogramme à la sortie du détecteur

La transformée de Fourier est une procédure mathématique qui permet de décomposer un interférogramme en une somme de sinusoides représentant chacune une onde donnée. La fréquence et l'amplitude de ces ondes sont calculées à partir des données de l'interférogramme recueillies par le spectromètre. Cette technique permet de mesurer un spectre entier de l'échantillon

Techniques d'acquisition des spectres FTIR deux techniques d'acquisition peuvent être utilisées :

- Transmission,
- Réflexion diffuse

IV.1.1.1- PIR-FTIR par transmission (absorption) :

Le procédé de mesure par transmission infrarouge est le plus utilisé [3, 4, 5] en raison de sa simplicité de mise en œuvre. Son principe de mesure est représenté sur la Figure ci-dessous. Pour chaque longueur d'onde, la transmittance T est définie par : $T = I/I_0$

Où I et I_0 représentent respectivement les intensités transmises de l'échantillon et d'une référence. La transmittance est souvent remplacée par son pourcentage (%T) ou par l'absorption A définie par : $A = \log (I_0/I) = \log (1/T)$

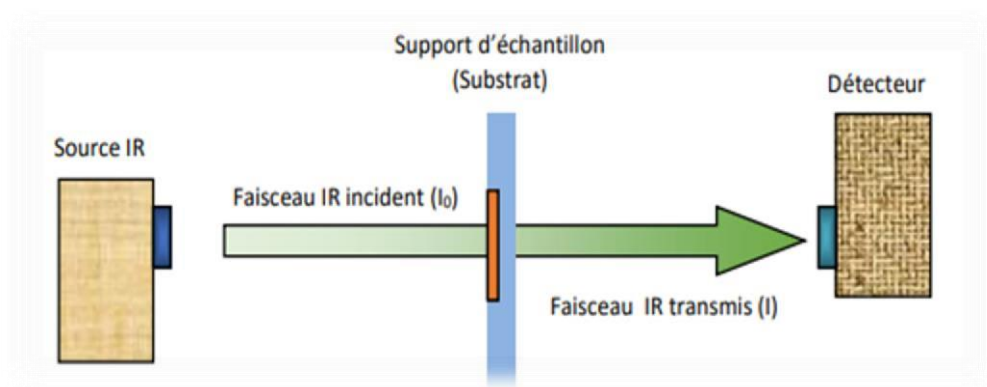


Figure 6 : Mécanisme de mesure de transmittance

IV.1.1.2- PIR-FTIR par réflexion :

La réflexion diffuse est la deuxième technique la plus courante d'enregistrement des spectres infrarouges. En mode réflexion diffuse, le faisceau infrarouge incident arrive à la surface de l'échantillon avec un certain angle. Les molécules vont absorber à certaines longueurs d'onde et le signal va repartir vers le détecteur.

L'inconvénient principal de ce mode est la perte d'une partie du rayonnement incident par diffusion au travers de l'échantillon.

Cette technique de réflexion diffuse peut être utilisée pour des échantillons sous forme de poudres ou de solides ayant une surface rugueuse comme du papier, du tissu. Dans la technique de réflectance diffuse, la taille des particules et l'homogénéité des échantillons en poudre joue un rôle important sur la qualité du spectre. Afin d'obtenir un spectre de bonne qualité, l'échantillon doit être broyé extrêmement fin. De plus, ce mode d'analyse est totalement non destructif.

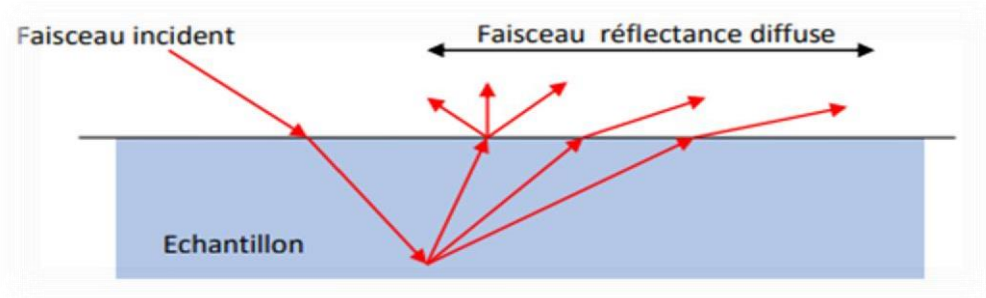


Figure 7 : Phénomène de réflectance diffuse

IV.2- Prétraitements des données obtenues :

Les théorèmes fondamentaux de modélisation insistent la normalité des données d'apprentissage, l'absence des valeurs aberrantes et des effets spectroscopiques dues à l'appareil et les conditions externes.

Ce qui implique l'importance de la préparation des données pour éviter les problèmes d'erreurs d'apprentissage.

Les effets spectroscopiques qui provoquent des déformations de spectre par affectation d'absorbance ou une autre grandeur optique mesurée sont :

- Effet additif : dû à la diffusion des photons $A_f(\lambda)$ (source physique) et le bruit de fond $A_b(\lambda)$ (origine liée au équipements).

- Effet multiplicatif : on peut le représenter sous forme d'un paramètre d'équation : $K.\varepsilon(\lambda).L.C$, le facteur K représente cet effet.
- On peut sommer les effets dans un terme unique : $A(\lambda) = K.\varepsilon(\lambda).L.C + Af(\lambda) + Ab(\lambda)$

Parmi les prétraitements qu'on peut utiliser on trouve :

IV.2.1- Correction log :

- Appliqué sur l'effet multiplicatif
- Transforme la constante K d'effet multiplicatif en constante $\log(k)$ qui est une constante additive qui sera éliminée après.

IV.2.2- Normalisation de signal :

- Correction concentrée sur l'effet multiplicatif
- Un algorithme qui élimine la ligne de base

IV.2.3- Traitement de Savitsky-Golay :

- Cette correction affecte l'effet additif
- Un algorithme de lissage qui élimine la ligne de base et les bruit de fond.

IV.2.4- Detrend :

- Un rôle important dans l'élimination d'effet additif
- Un algorithme qui élimine la ligne de base

IV.2.5- Dérivé de spectre :

- Appliqué sur l'effet additif
- Un processus mathématique qui élimine la ligne de base

IV.2.6- SNV :

- C'est une méthode combinée
- Enlever la moyenne d'un spectre à chacun de ses points, puis les diviser par la norme de spectre.

IV.2.7- SMC :

- C'est une méthode combinée

- Identifier des effets additifs et multiplicatifs d'un spectre par rapport à un spectre de référence, cette méthode est basée sur des calculs matriciels de régression.

V- Notions de modélisation :

V.1- Algorithmes de modélisation :

Les algorithmes de modélisation déterminent la façon d'obtention des entrées et de calcul des sorties ; Ces algorithmes se différencient au niveau de structure des données entrantes et sortantes, paramètres de réglage de modèle et procédure interne, ce qui produit la différence des performances et le temps d'exécution entre les modèles.

V.2- Les algorithmes utilisés dans ce projet :

V.2.1- PLS : Régression à moindres carrées partiels :

- Aperçu sur la PLS :

La **régression PLS** est une technique récente qui généralise et combine les colonnes obtenues par **analyse en composantes principales** et **de la régression multiple**. Son objectif est de prédire ou d'analyser un ensemble de **variables dépendantes** à partir d'un ensemble de variables indépendantes ou de prédicteurs. Cette prédiction est réalisée en extrayant des prédicteurs un ensemble de facteurs orthogonaux appelés **variables latentes** qui ont la **meilleure puissance prédictive**.

La **régression PLS** est particulièrement utile lorsque nous devons prédire un ensemble de variables dépendantes d'un grand ensemble de **variables indépendantes** (prédicteurs). Elle est née dans les **sciences sociales**, spécifiquement en économie (Herman Wold 1966) mais est devenue populaire d'abord en **Chimie** en partie au fils d'Herman (Svante 2001) et en **évaluation sensorielle**. Mais la **régression PLS** devient aussi un outil de choix en **sciences sociales** comme **technique multivariée** pour les données non expérimentales et expérimentales. Elle a d'abord été présentée comme un algorithme proche de la méthode de puissance (utilisé pour le calcul des vecteurs propres) mais a été rapidement interprétée dans un cadre statistique. [4]

- Notions et notations préalables :

Les observations I décrites par K variables dépendantes sont stockées dans une matrice $I \times K$ notée Y , les valeurs de J prédicteurs collectées sur ces observations I sont rassemblées dans la matrice $I \times J$ X . [4]

- **Objectif de la régression PLS : Prédire Y à partir de X :**

Le but de la régression PLS est de prédire Y à partir de X et de décrire leur structure commune. Lorsque Y est un vecteur et X est un rang complet, cet objectif pourrait être atteint en utilisant la régression multiple ordinaire. Lorsque le nombre de prédicteurs est important par rapport au nombre d'observations, X est susceptible d'être singulier et l'approche de régression n'est plus faisable (c'est-à-dire en raison de la multi-colinéarité). Plusieurs approches ont été développées pour faire face à ce problème. Une approche consiste à éliminer certains prédicteurs (par exemple, en utilisant des méthodes par étapes) une autre, appelée régression sur composantes principales, consiste à effectuer une analyse en composantes principales (ACP) de la matrice X et ensuite utiliser les composantes principales (c'est-à-dire les vecteurs propres) de X comme régresseurs sur Y. Techniquement en ACP, X est décomposé en utilisant sa décomposition en valeurs singulières comme : $X = S\Delta V^T$

Avec : $S^T S = V^T V = I$, (Ce sont les matrices des vecteurs singuliers gauche et droit), et Δ étant une matrice diagonale avec les valeurs singulières comme éléments diagonaux. Les vecteurs singuliers sont ordonnés en fonction de leurs valeurs singulières correspondantes qui correspondent à la racine carrée de la variance de X expliquée par chaque vecteur singulier. Les vecteurs singuliers de gauche (c'est-à-dire les colonnes de S) sont ensuite utilisés pour prédire Y en utilisant la régression standard car l'orthogonalité du singulier vecteurs élimine le problème de la multi-colinéarité. Mais, le problème de choisir un sous-ensemble optimal de prédicteurs demeure. Un possible la stratégie consiste à ne conserver que quelques-uns des premiers composants. Mais ceux-ci les composantes sont choisies pour expliquer X plutôt que Y, et donc, rien ne garantit que les composantes principales, qui expliquent X, sont pertinents pour Y.

En revanche, la régression PLS trouve des composants de X qui sont également pertinent pour Y. Plus précisément, la régression PLS recherche un ensemble de composantes (appelées vecteurs latents) qui effectue une décomposition simultanée de X et Y avec la contrainte que ces composantes expliquent autant que possible la covariance entre X et Y. Cette étape généralise l'ACP. Il est suivi d'une étape de régression où la décomposition de X est utilisée pour prédire Y. [4]

- **Régression PLS avec NIPALS :**

C'est la méthode la plus utile parmi les approches de régression PLS [5] car elle minimise l'erreur des informations perdu pendant les étapes de recherche de combinaison entrées-sorties.

Les équations importantes de cette régression sont :

- $w = X'Y$
- $W = w / \|w\|$
- $t = XW$
- $p = X' t (t't)^{-1}$
- $\hat{c} = t' Y (t't)^{-1}$
- $Y_{\text{prédites}} = t \hat{c}$
- $X_{\text{prédites}} = t p'$
- $X_{\text{résiduelles}} = X - t p'$
- $Y_{\text{résiduelles}} = Y - t \hat{c}$

Tels que :

- p : matrice des loadings
- t : matrice des scores
- w : poids de pondération non normés.
- W : poids de pondération normés.
- $\hat{c}$: matrice des coefficients obtenus par moindres carrées.

V.2.2- PCR : Régression sur les composantes principales :

Est une relation linéaire entre les entrées obtenues par analyse de composantes principales soit par la méthode de décomposition en valeurs singulières ou la méthode NIPALS et les sorties enregistrées dans le tableau des données. [6]

V.2.3- MLP : Réseaux de neurones multi couches :

Dans cet algorithme de modélisation [7], l'unité la plus importante est le perceptron, ce dernier est une unité numérique qui a des entrées, des sorties et des poids de pondération ; Les entrées sont pondérées par les poids, puis elles entrent dans le perceptron qui est composé d'une fonction d'activation qui calcule la sortie selon les paramètres réglés par l'utilisateur sachant que les performances de perceptron seront maximales, l'architecture de modèle est expliquée dans les figures suivantes :

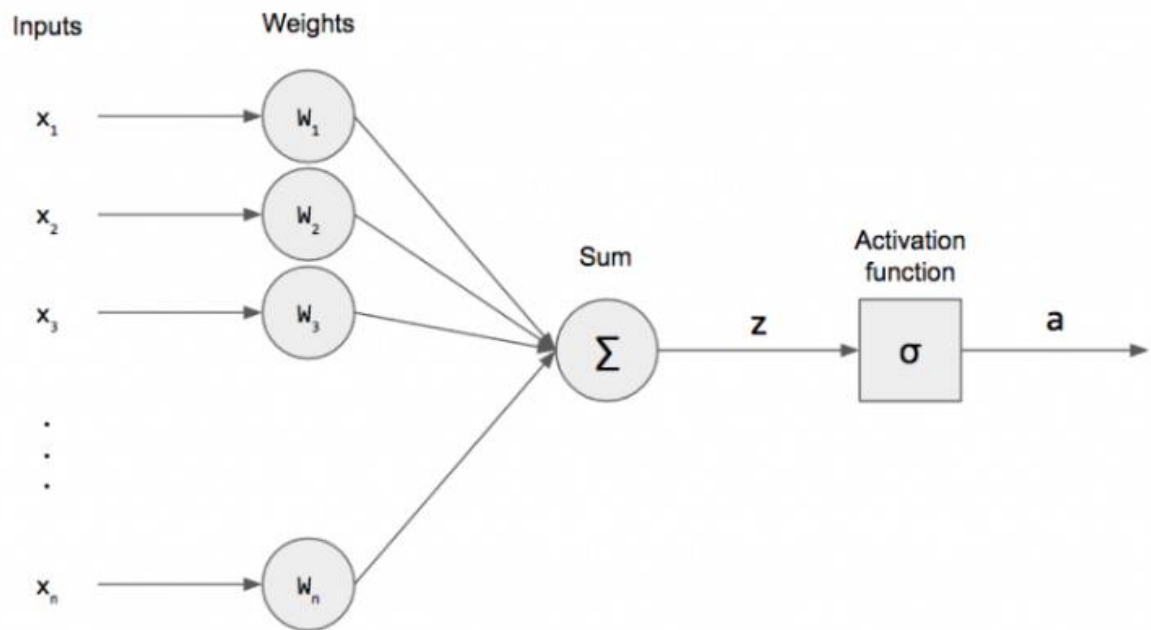


Figure 8 : Les composantes d'un perceptron

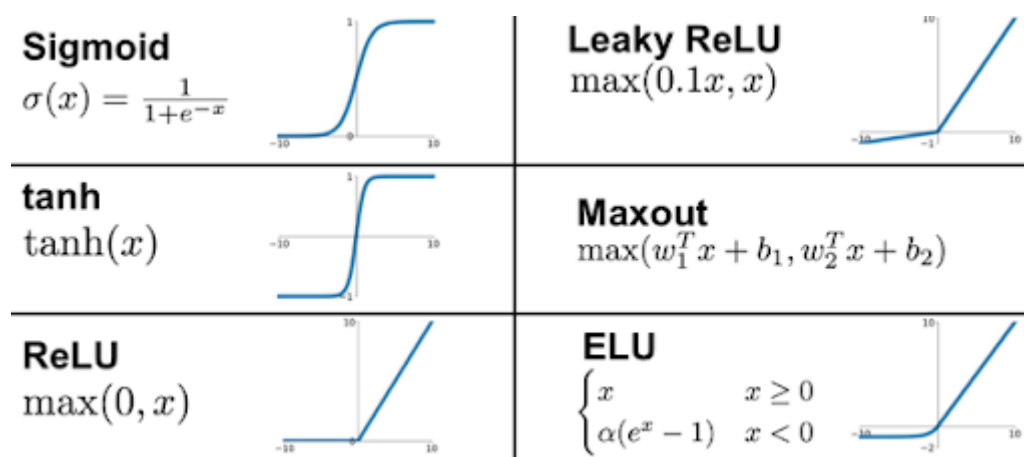


Figure 9 : Exemples de fonctions d'activation

Cet algorithme est basé sur l'union de plusieurs couches de perceptrons dans un seul système de modélisation des données, on choisit les réglages de ce système tels que les performances seront maximales pour avoir une régression bien corrélée, ajustée et donne des résultats proches de la réalité.

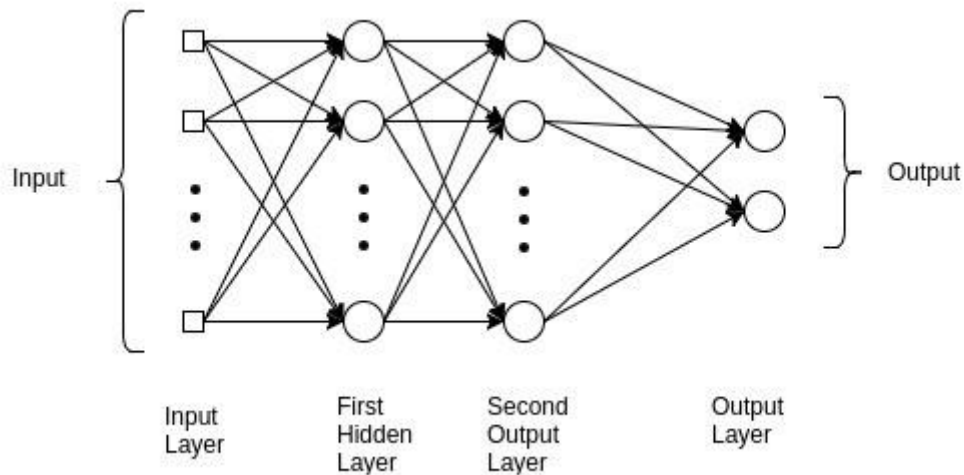


Figure 10 : Les composantes d'un réseau MLP

V.2.4- KNN : K proches voisins :

Cette méthode [7] est basée sur les étapes suivantes, soit en problèmes de classification ou de régression, les étapes sont détaillées sous forme écrite et figure ci-dessous :

- Calcul des distances entre le point cible et tous les autres points
- Détermination de valeur de K utilisant un facteur de performance déjà choisi, il faut choisir un nombre impair à la fin
- Détermination de K voisins les plus proches
- Vote sur la réponse entre les proches voisins
- En cas de problèmes de régression, on prend la moyenne des votes comme valeur prédite

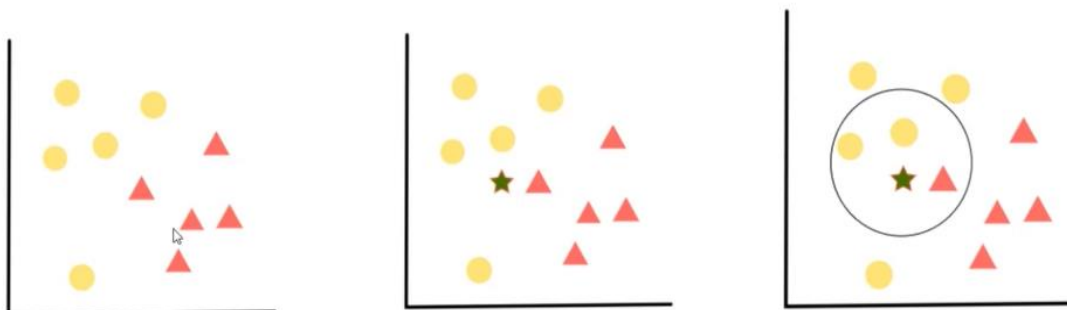


Figure 11 : Les étapes d'algorithme KNN

V.2.1- SVM : Machine à vecteurs de support :

C'est un algorithme [8] qui cherche un hyperplan d'équation $Y = w X + b$ dont w et b sont les paramètres obtenus par le modèle ; Pour éviter les erreurs d'écart, cette recherche est limitée par des bandes de décision tels que : $-a < Y - (w X + b) < a$ dont a est la constante de bande de décision ; en synthèse :

- Hyperplan : il s'agit essentiellement d'une ligne de séparation entre deux classes de données dans SVM. Mais dans l'algorithme de régression, c'est la ligne qui sera utilisée pour prédire la sortie.
- Bande de décision : cette bande est considérée comme une ligne de démarcation d'un côté de laquelle se trouvent des exemples positifs et de l'autre côté des exemples négatifs. Sur cette ligne même, les exemples peuvent être classés comme positifs ou négatifs. Ce même concept de SVM sera également appliqué dans l'algorithme de régression.
- L'équation d'hyperplan peut être considérée comme la fonction d'activation des perceptrons.

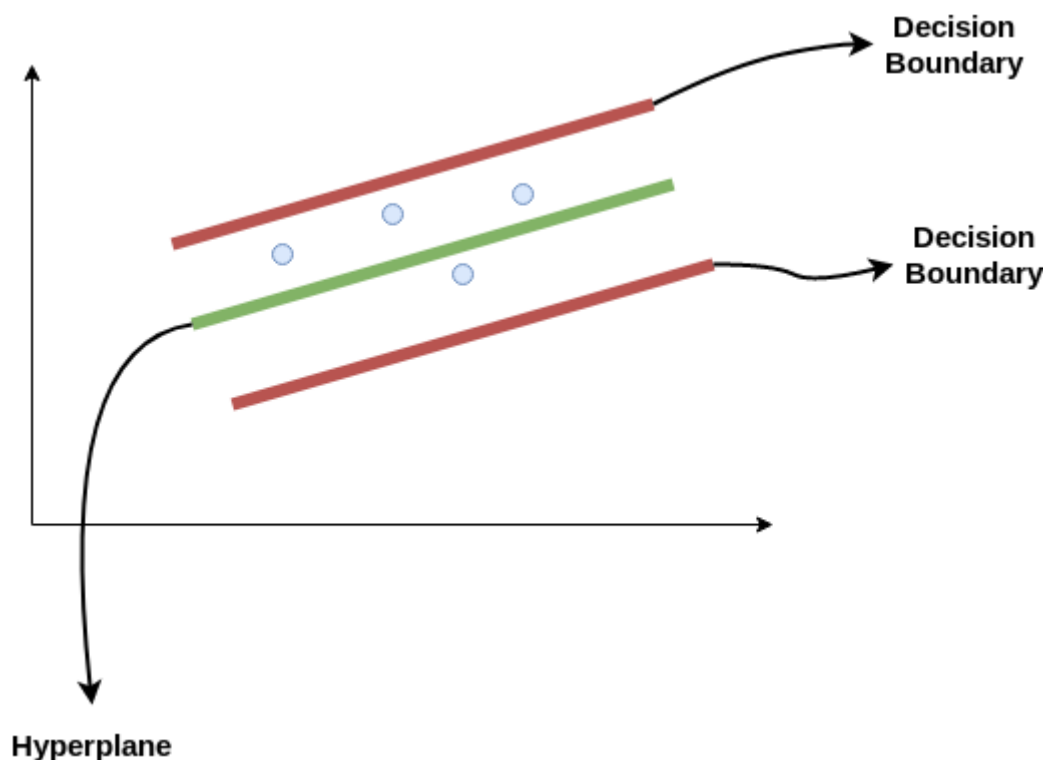


Figure 12 : Les composantes d'un algorithme SVM

V.2.1- DT : Arbre de décision :

Un arbre de décision [7] [9] est un algorithme d'apprentissage automatique supervisé, il est utilisé à la fois dans les algorithmes de classification et de régression ; L'arbre de décision est comme un arbre avec des nœuds ; Les branches dépendent d'un certain nombre de facteurs, on divise les données en branches comme dans la figure 14 jusqu'à ce qu'on atteigne une valeur seuil, un arbre de décision se compose des nœuds racine, des nœuds enfants et des nœuds feuilles distribués de la manière suivante :

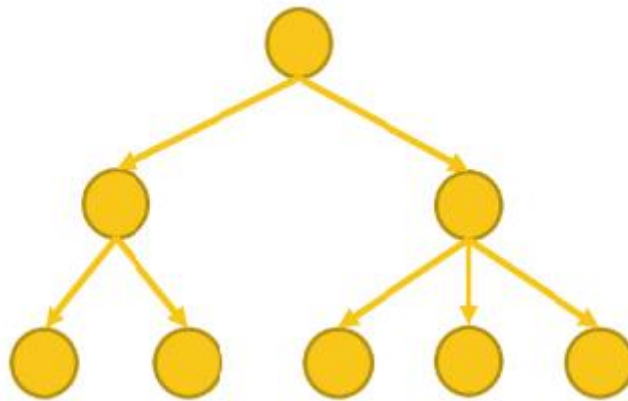


Figure 13 : Présentation générale d'un arbre de décision

Chaque arbre de décision contient les éléments suivants :

- Nœuds : C'est le point où l'arbre se divise en fonction de la valeur d'un attribut / caractéristique de l'ensemble de données
- Bords : Il dirige le résultat d'une scission vers le nœud suivant que nous pouvons voir dans la figure ci-dessus. Il existe un avantage pour chaque valeur potentielle de chacun de ces attributs / fonctionnalités.
- Racine : C'est le nœud où la première scission a lieu
- Feuilles : ce sont les nœuds terminaux qui prédisent le résultat de l'arbre de décision

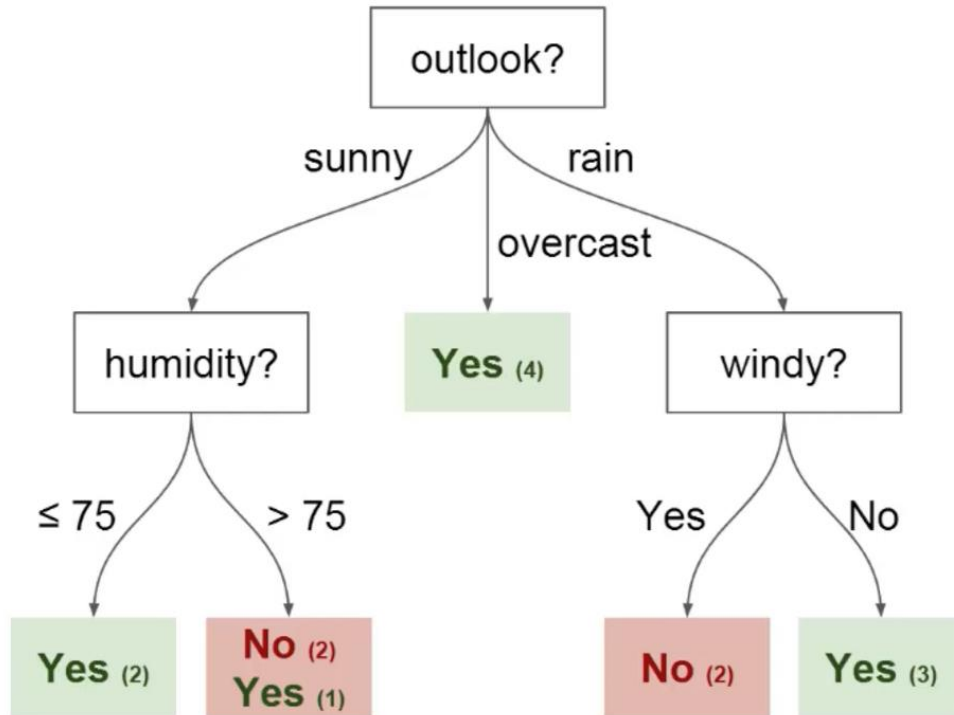


Figure 14 : Exemple d'arbre de décision

V.2.1- RFT : Forêt d'arbres décisionnels :

Cet algorithme [7] se base sur l'union de plusieurs arbres décisionnels dans un seul outil de régression tels que le mécanisme est :

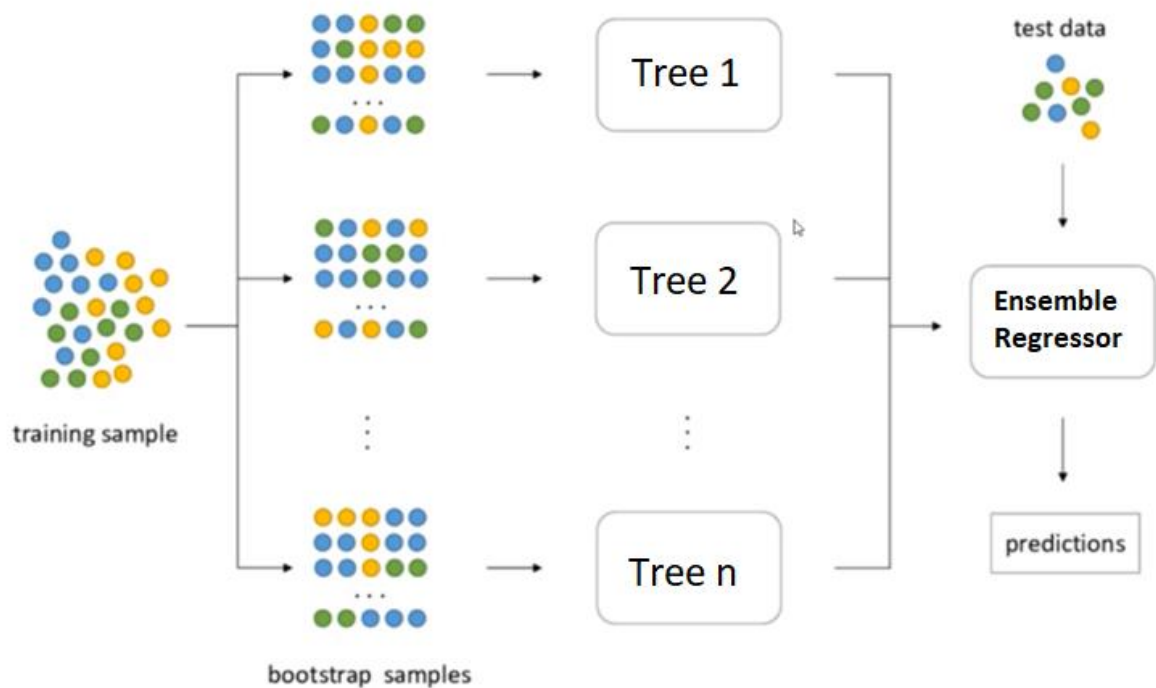


Figure 15 : Les étapes de Forêt d'arbres décisionnels

VI- Outils de validation de modèle :

L'examen d'un modèle est effectué sur plusieurs niveaux qui sont :

VI.1- Détermination des relations entre les paramètres :

- Coefficient de détermination R^2 : sert à quantifier la corrélation entre les entrées et les sorties de modèle. On peut écrire sa valeur en forme de pourcentage. La valeur doit être proche de 100%.
- Coefficient de détermination ajusté $R^2_{\text{ajusté}}$: quantifie simultanément la corrélation entrées-sorties et l'ajustement de cette combinaison linéaire. Ce coefficient doit converger vers 100% en forme de pourcentage.
- Coefficient de corrélation $\text{Corr}(A, B)$: utilisé pour détecter les variables les plus liées entre elles. Si la valeur de corrélation est au voisinage de 1 : on dit qu'il y a une corrélation positive forte ; Si la valeur est au voisinage de -1 : on dit que la corrélation est forte négativement.
- Facteur de colinéarité VIF : utilisé pour détecter les facteurs colinéaires entre eux si la valeur VIF est supérieure ou égale à 1.

VI.2- Quantification d'erreur de prédiction :

- MSE, mean squared error: $\frac{1}{n} \sum_1^n (y_i - y_{\text{pred } i})^2$; tels que : $y_{\text{pred } i}$: élément de vecteur des valeurs prédites et y_i : élément de vecteur de réponse réelle.
- RMSE, Root mean squared error: $\sqrt{\text{MSE}}$
- ME, max error : $\max(Y - \hat{Y})$; tels que : Y : vecteur de réponse réelle et $\hat{Y}$: vecteur des valeurs prédites.

VI.3- Tests de choix de meilleur modèle :

- RPD, est une grandeur utile pour les modèles entraînés par des données spectrales tels que la relation est : $\text{Ecart_type}(Y)/\text{RMSE}(Y, \hat{Y})$. [11]
- AIC, est un critère utile pour criblage des modèles tels que la relation est : $n * \log(\text{mse}) + 2 * \text{num_param}$. [10]
- BIC, est un critère utile pour criblage des modèles tels que la relation est : $n * \log(\text{mse}) + \text{num_params} * \log(n)$. [10]

CHAPITRE 2 : PARTIE
EXPÉRIMENTALE -
RECHERCHE DE MODÈLES
DE DÉTERMINATION DU
TAUX DE CARBONE ET DE
D'AZOTE DANS LES
FEUILLES DE TOMATES
CERISE PAR LES
MÉTHODES
CHIMIOMÉTRIQUES

Chapitre II : Partie expérimentale - Recherche de modèles de détermination du taux de carbone et de d'azote dans les feuilles de tomates cerise par les méthodes chimiométriques

Dans cette partie nous comptons développer un analyseur de mesure de la quantité totale de carbone et azote en utilisant des technologies à base spectroscopique infrarouge portative, et en incluant l'élaboration d'un modèle mathématique de prédiction locale propre à chaque élément cible.

Ma propre mission dans ce projet est l'étude des meilleurs paramètres possibles pour plusieurs algorithmes de modélisation et le choix du meilleur modèle puis le valider, en utilisant les données rassemblées à partir de l'analyse PIR-FTIR. Toutes mes analyses sont effectuées par le langage de programmation Python 3.7 utilisant les bibliothèques : Sklearn, Matplotlib et Seaborn.

La démarche du projet commence par les réglages des paramètres de modélisation, puis on choisit le meilleur modèle, on le valide et finalement on fait un test qui présente ces performances dans la vraie vie.

I- A propos des données :

I.1- Présentation des données :

En exemple, une partie des données brutes est présentée sous forme de deux tableaux : données de Carbone et données d'Azote. La première colonne de chaque tableau présente les concentrations (en % par 100 g de la matière sèche) de Carbone (mesurée par perte de masse) et Azote (mesurée par Kjeldahl) respectivement, et les autres colonnes présentent les réflectances enregistrées par la méthode spectrale PIR-FTIR (les titres de ces colonnes sont des fréquences en cm^{-1}); les lignes de ces tableaux présentent les essais sur chaque échantillon ; la base des données de Carbone contient 361 lignes x 2152 colonnes et la base des données d'Azote contient 470 lignes x 2152 colonnes.

	C(%) f(cm^{-1})	350	351	352	353	354	355	...	2500
1	33,20572	0,685202	0,703378	0,707244	0,700635	0,710577	0,71256	...	0,654053
2	48,40181	1,211069	1,220758	1,17734	1,153564	1,207712	1,223311	...	0,636301
3	46,2405	0,806229	0,812663	0,821668	0,829485	0,832806	0,830015	...	0,611575

4	45,6938	0,690813	0,696243	0,70462	0,718818	0,736409	0,750347	...	0,458071
5	53,12453	1,141793	1,118914	1,150225	1,198876	1,146297	1,200844	...	0,969213
...	...	...	...	...	...	...	...	...	...
361	44,39214	0,662118	0,668036	0,686908	0,704108	0,693279	0,704106	...	0,543783

Tableau 1 : Une prévision des données de Carbone

	N(%) f(Cm ⁻¹)	350	351	352	353	354	355	...	2500
1	0,33555	0,685202	0,703378	0,707244	0,700635	0,710577	0,71256	...	0,654053
2	1,390083	1,211069	1,220758	1,17734	1,153564	1,207712	1,223311	...	0,636301
3	4,606212	1,141793	1,118914	1,150225	1,198876	1,146297	1,200844	...	0,969213
4	2,614611	1,088999	1,105354	1,099355	1,091881	1,109881	1,113219	...	0,817567
5	1,167868	0,816191	0,822812	0,830625	0,834704	0,83348	0,838764	...	0,380638
...	...	...	...	...	...	...	...	...	...
470	2,715643	0,892809	0,89273	0,883102	0,875398	0,905165	0,918359	...	0,552016

Tableau 2 : Une prévision des données d'Azote

Les figures ci-dessous représentent les spectres de réflectance des données de carbone et d'azote brutes (à gauche de la figure) avec des exemples des prétraitement effectués parmi plusieurs expériences (à droite de la figure) :

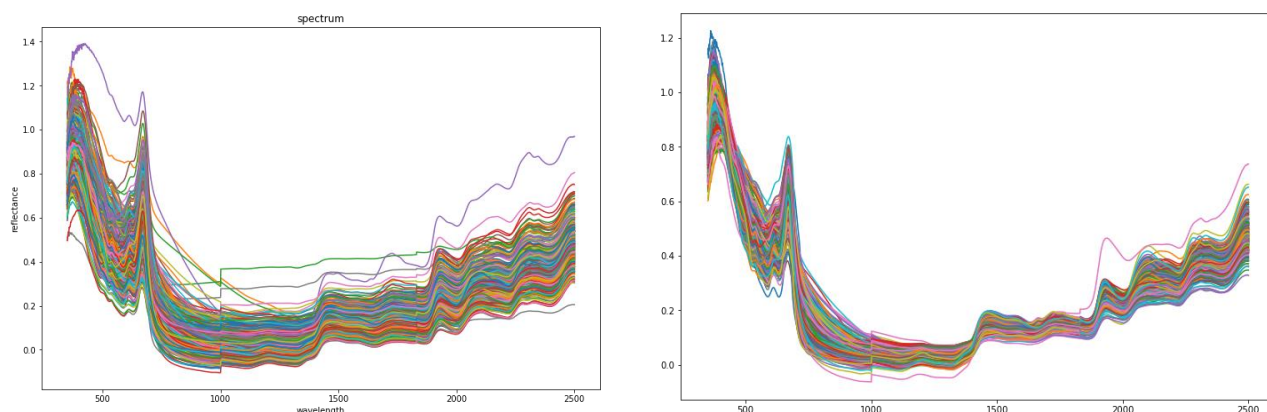


Figure 16 : Spectres de réflectance des données de Carbone avant et après un exemple de prétraitement - prétraitement effectué : MSC + Savitsky-Golay

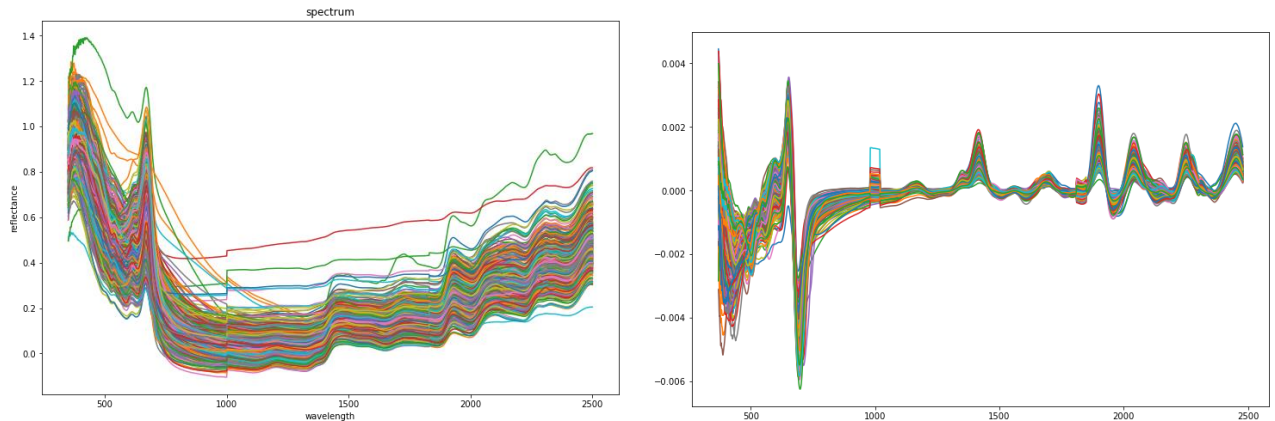


Figure 17 : Spectres de réflectance des données d'Azote avant et après un exemple de prétraitement - prétraitement effectué : Moving average + GAP1 + MSC

On remarque la présence des effets aditif, multiplicatif... et les spectres aberrants, d'où l'importance des prétraitements.

A propos des prétraitements effectués, on a fait plusieurs expériences des prétraitements et dans la partie pratique on va tester les données prétraitées, puis on va choisir le meilleur choix. Il faut noter bien qu'on ne va pas travailler avec toute la bande spectrale, il y a une sélection des bandes utiles pour la modélisation.

I.2- Analyses élémentaires des données :

I.2.1- Box plot :

Pour avoir une idée sur les valeurs aberrantes dans la réponse on fait un box plot :

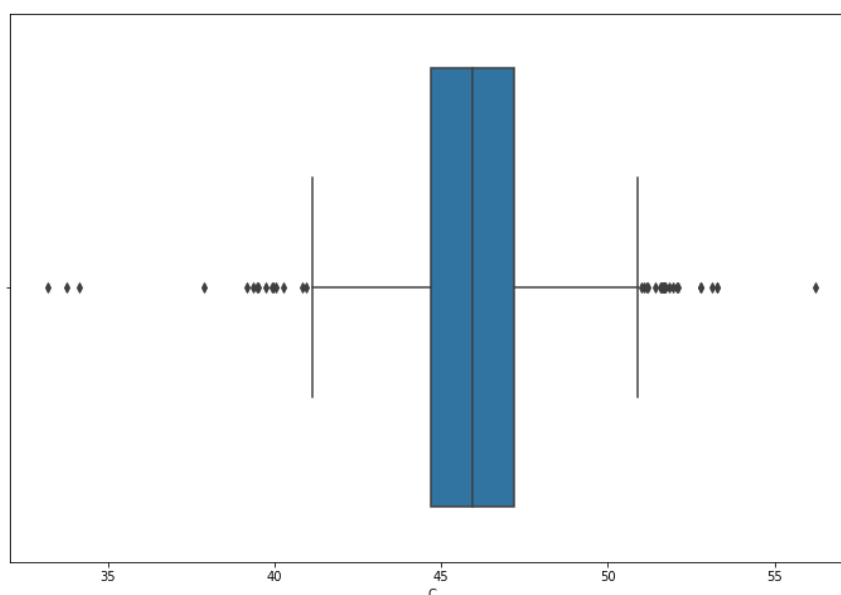


Figure 18 : Box plot de réponse des données de Carbone

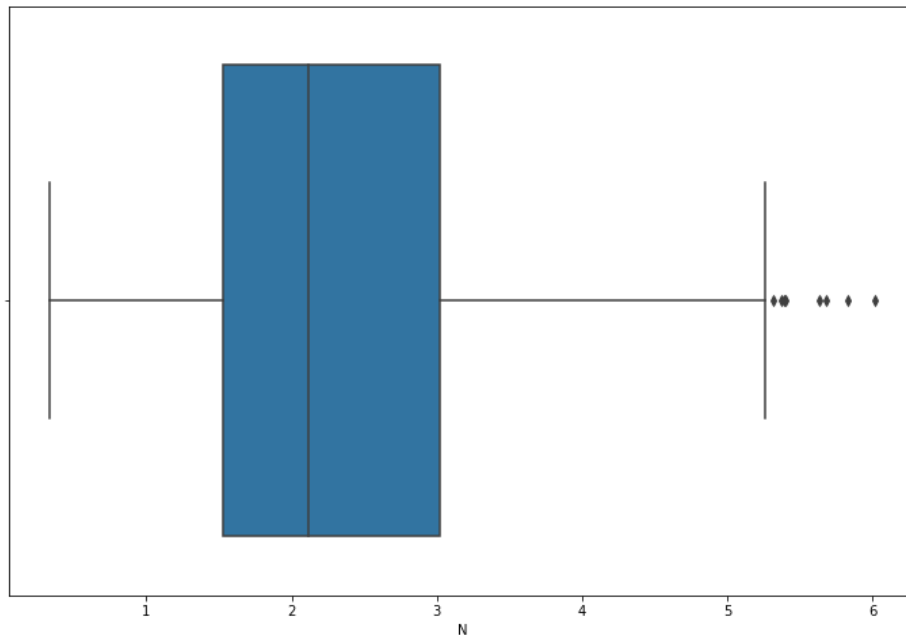


Figure 19 : Box plot de réponse des données d’Azote

Ce test montre qu’il y a des valeurs aberrantes, maintenant il faut voir l’effet de ces valeurs sur la normalité.

I.2.2- Droite d’Henry :

Le test d’Henry effectué sur les vecteurs de réponse nous donne :

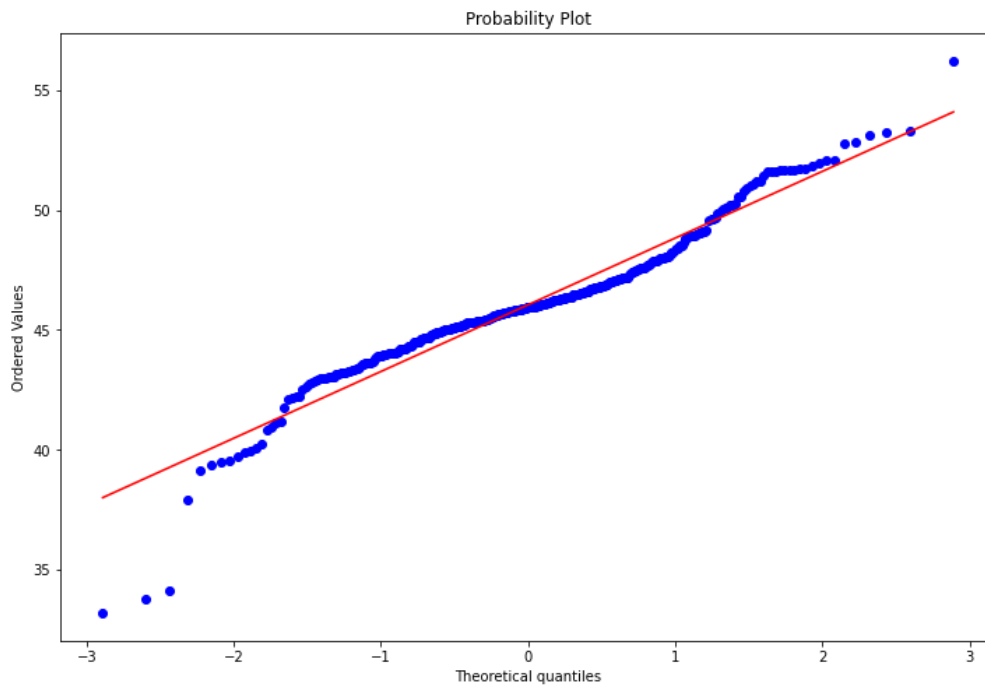


Figure 20 : Plot d’Henry des données de Carbone

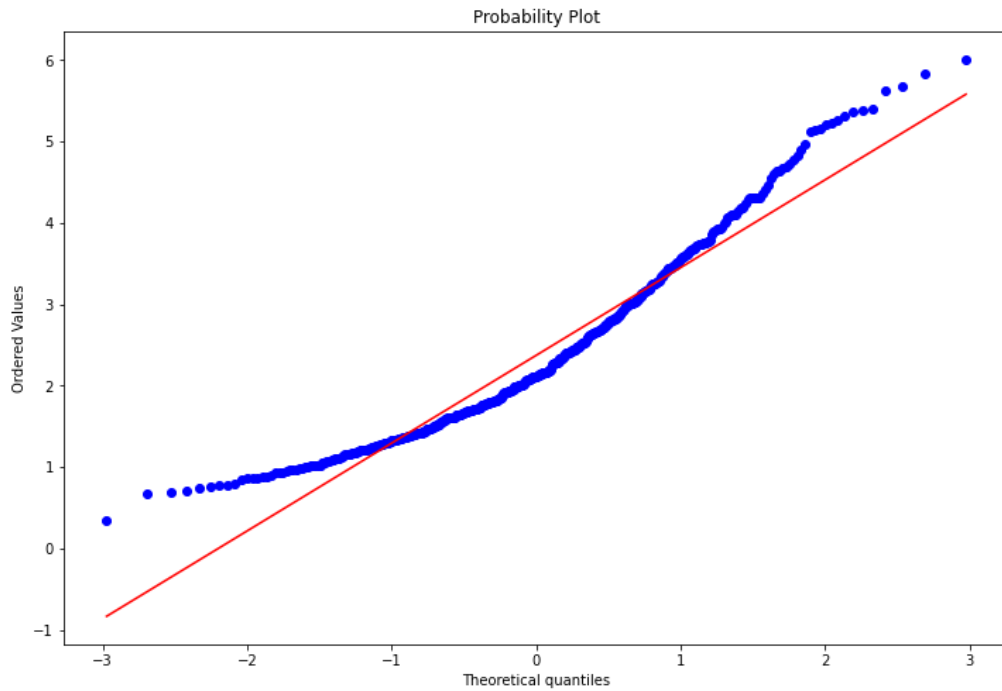


Figure 21 : Plot d'Henry des données d'Azote

On constate que la présence des valeurs aberrantes provoque un écart à la normalité.

I.2.3- Test de Shapiro-Wilk:

Ce test quantifie l'écart à la normalité des vecteurs de réponse en termes de quantile et probabilité.

```
Quantile shapiro : 0.9441915154457092
propability shapiro : 2.0193163807746828e-10
desicion : Not Normal
```

Figure 22 : Résultats de test Shapiro-Wilk sur les données de Carbone

```
Quantile shapiro : 0.9397115707397461
propability shapiro : 7.154277712785595e-13
desicion : Not Normal
```

Figure 23 : Résultats de test Shapiro-Wilk sur les données d'Azote

Ces résultats montrent que l'écart à la normalité est très grand.

II- Examen des modèles sur les données prétraitées de Carbone – prétraitement 1 :

II.1- Prétraitements effectués :

Les prétraitements effectués sur les données de carbone sont :

- MSC
- Lissage Savitsky-Golay

II.2- Examen des performances :

Maintenant, il faut valider le choix des paramètres des modèles et choisir le meilleur modèle.

N.B : La signification des couleurs dans tous les tableaux est :

- Rouge : rejeté
- Jaune : incertain
- Vert : accepté

Modèles	R ² train %	R ² CV %	R ² test %	RMSE train	RMSE CV	RMSE test
PLS	89.853817	80.918207	77.440639	0.906132	1.169510	1.163661
PCR	84.303495	80.214124	71.248917	1.127045	1.209116	1.313682
SVR	97.693127	42.290064	38.926634	0.432068	2.129837	1.914648
RF	90.805312	19.379742	45.505277	0.862598	2.503612	1.808591
DT	60.854194	-33.94847	-16.14645	1.779846	3.115360	2.640378
KNN	100.000000	6.912335	44.804552	0.000000	2.694251	1.820182
MLP	80.683654	74.191180	66.253086	1.250266	1.375814	1.423246

Tableau 3 : Résultats d'examen de performances effectué sur les données de Carbone prétraitées – prétraitement 1

Pour examiner l'ajustement des modèles, on fait le plot : predicted VS target.

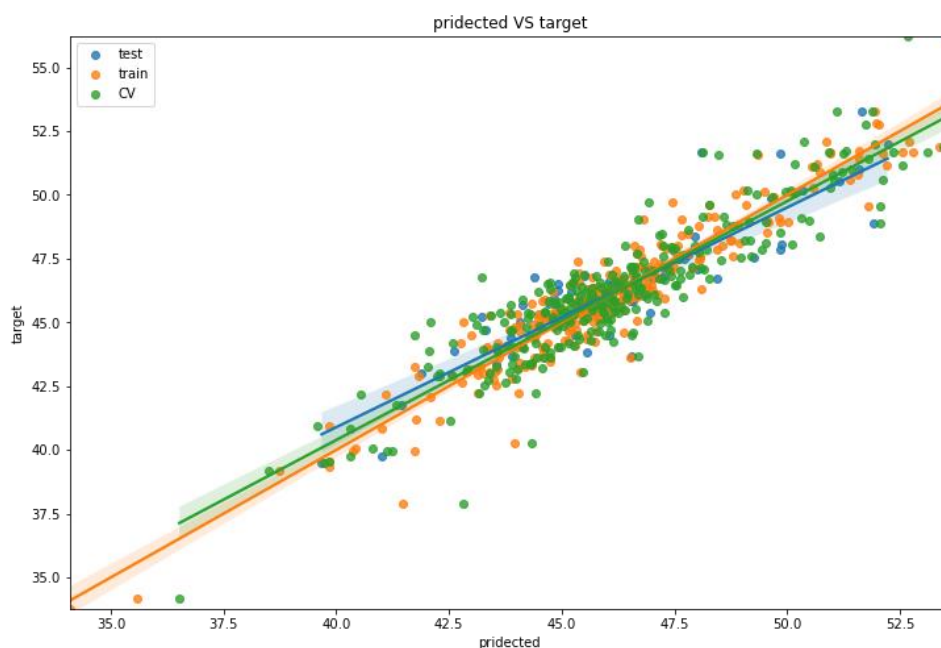


Figure 24 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle PLS

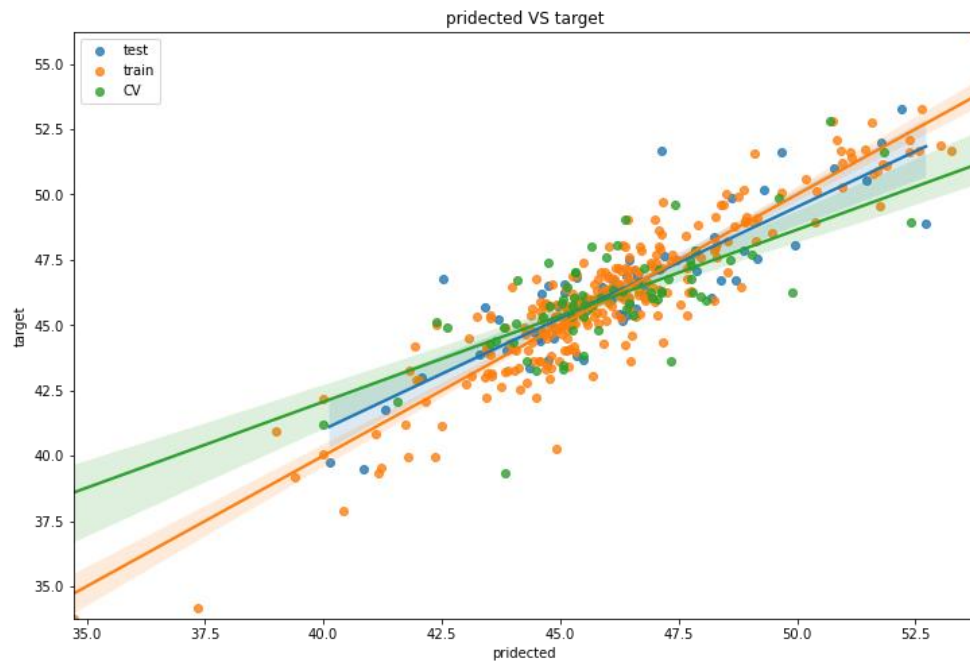


Figure 25 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle PCR

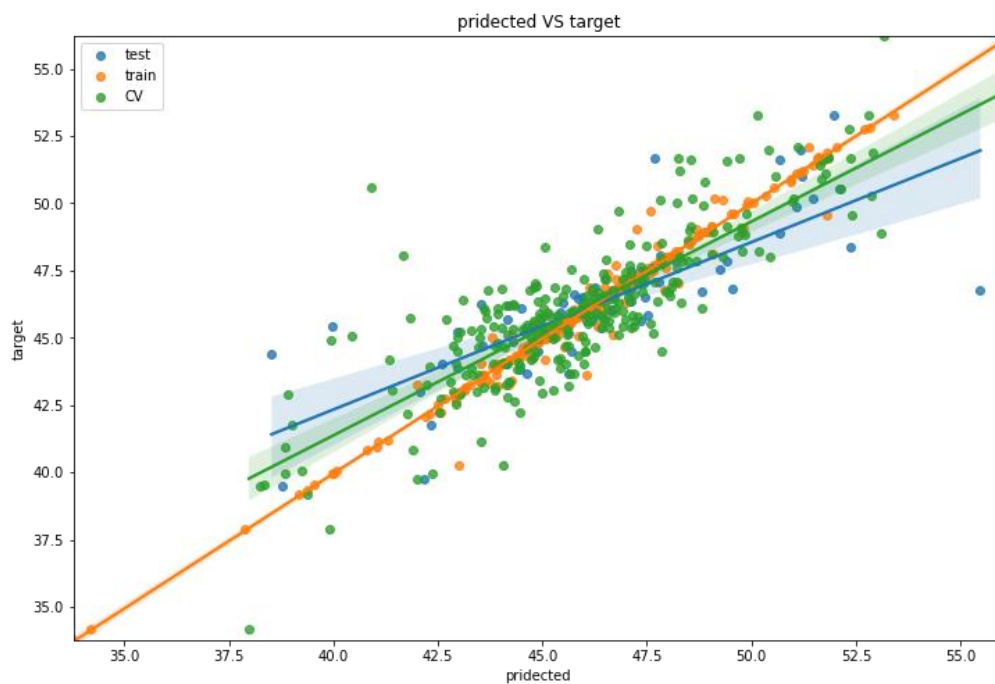


Figure 26 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle SVR

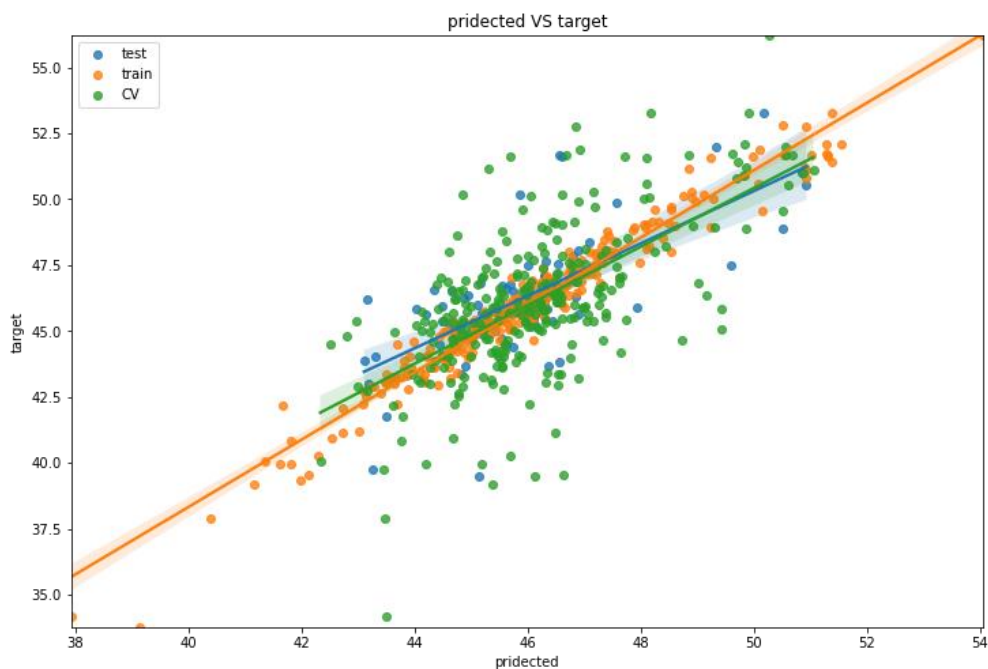


Figure 27 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle RFT

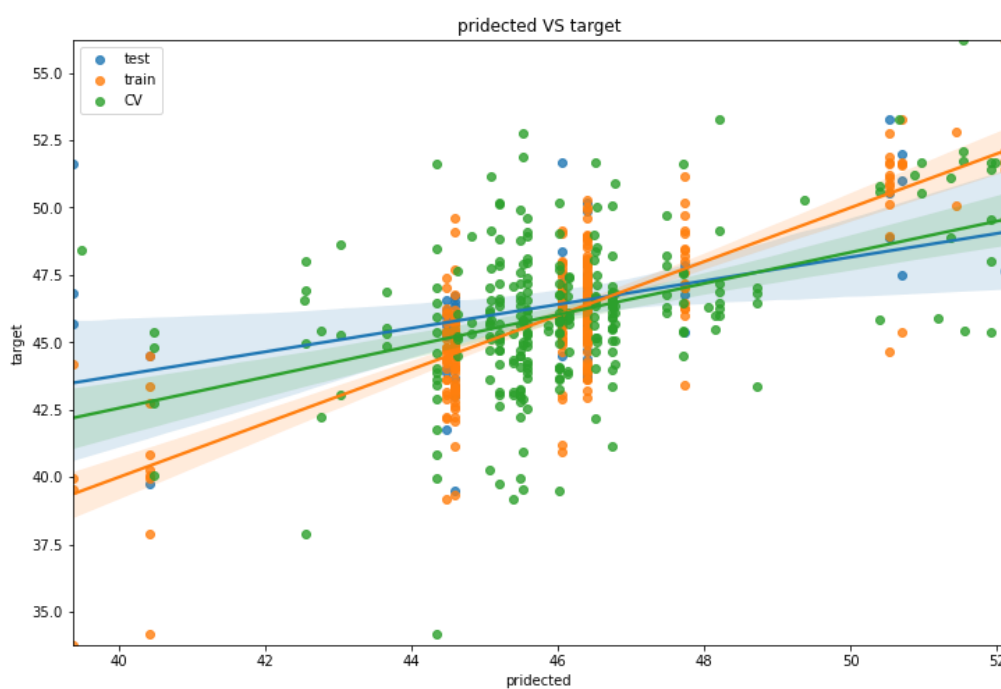


Figure 28 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle DT

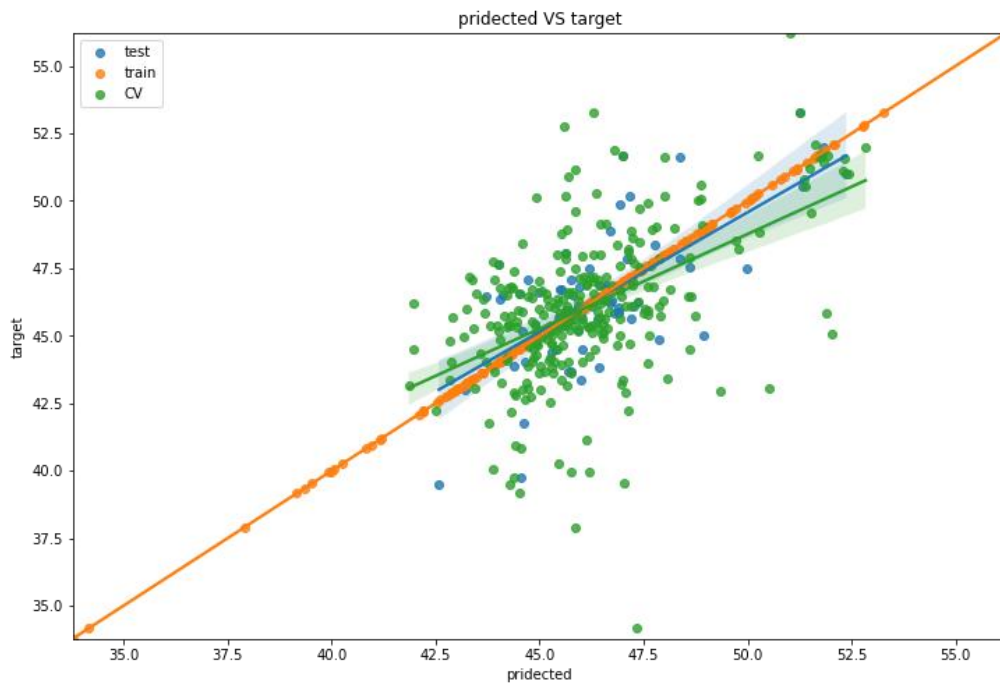


Figure 29 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle KNN

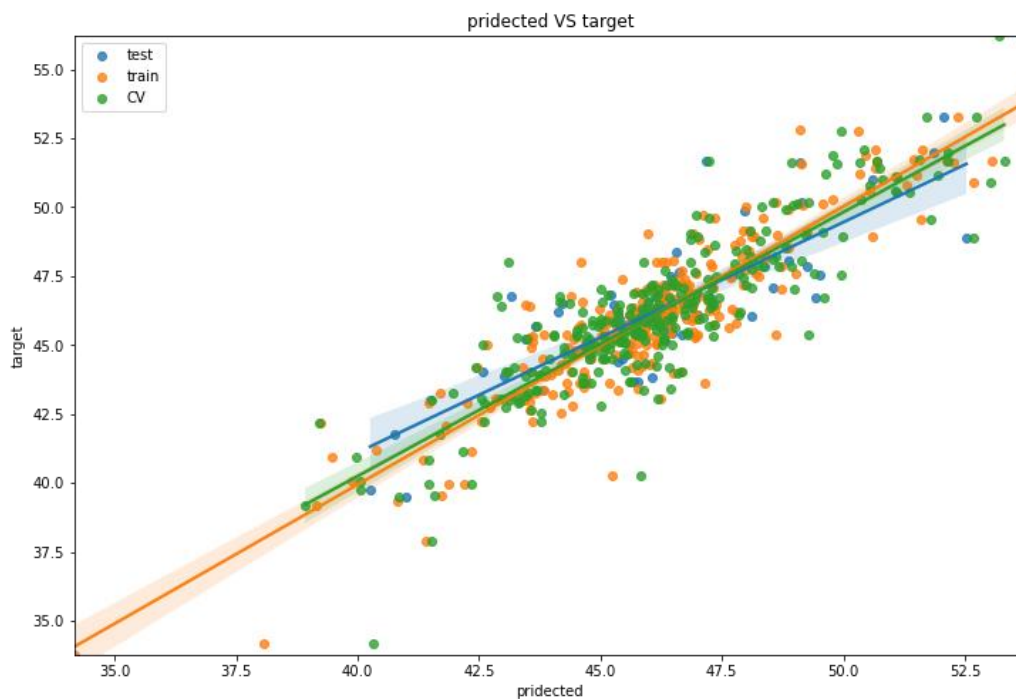


Figure 30 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 1 - modèle MLP

II.3- Examen de résidu :

En utilisant le test de Shapiro-Wilk, ~~voici~~ les résultats d’examen de résidus sont regroupés dans le tableau ci-dessous :

Modèle	Shapiro quantile	Probabilité	Décision
PLS	0.975365	0.182667	Normal
PCR	0.948532	0.00607862	Not Normal
SVR	0.87567	4.81103e-06	Not Normal
RFT	0.87567	4.81103e-06	Not Normal
DT	0.888413	1.39259e-05	Not Normal
KNN	0.888413	1.39259e-05	Not Normal
MLP	0.948532	0.00607862	Not Normal

Tableau 4 : Résultats d'examen de normalité de résidus effectué sur les données de Carbone prétraitées – prétraitement 1

II.4- Conclusion :

Les modèles acceptés sont PLS et PCR, mais on va sélectionner le modèle PLS à cause de ces performances test.

III- Examen des modèles sur les données prétraitées de Carbone – prétraitement 2 :

III.1- Prétraitements effectués :

Les prétraitements effectués sur les données de carbone sont :

- SNV
- Lissage Savitky-Golay

III.2- Examen des performances :

Modèles	R ² train %	R ² CV %	R ² test %	RMSE train	RMSE CV	RMSE test
PLS	87.440933	83.271319	76.412813	1.008135	1.126883	1.189874
PCR	85.064236	78.410816	74.950842	1.099395	1.257448	1.226195
SVR	85.512236	68.189522	72.816781	1.082781	1.543379	1.277361
RF	90.864347	28.439447	38.879849	0.859824	2.359630	1.915381

DT	64.112894	14.501535	-12.01933	1.704155	2.572445	2.593042
KNN	100.000000	11.906107	33.357776	0.000000	2.607821	2.000036
MLP	72.317243	69.709848	51.231933	1.496734	1.491008	1.710925

Tableau 5 : Résultats d'examen de performances effectué sur les données de Carbone prétraitées – prétraitement 2

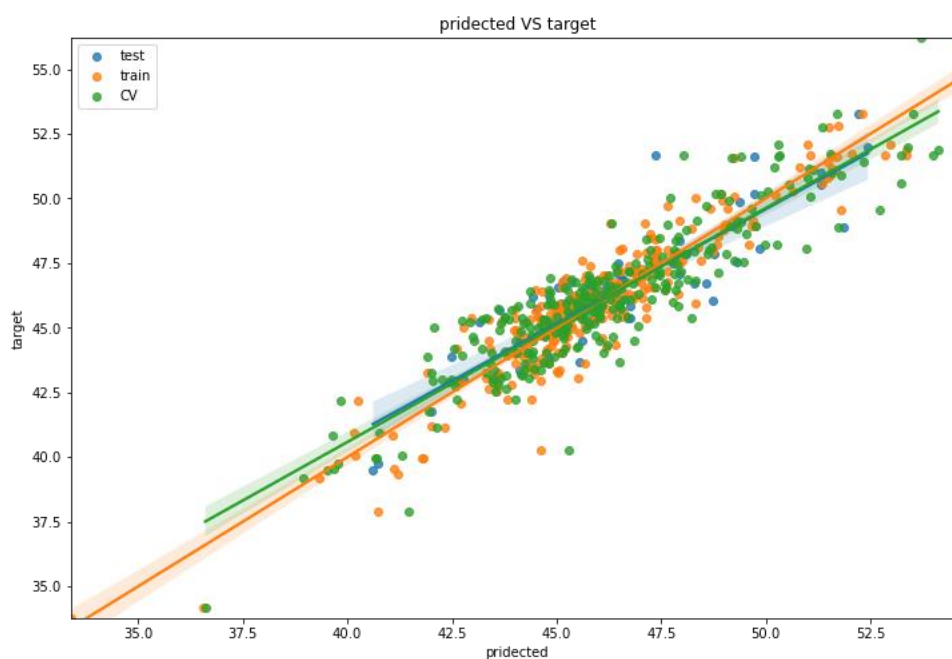


Figure 31 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle PLS

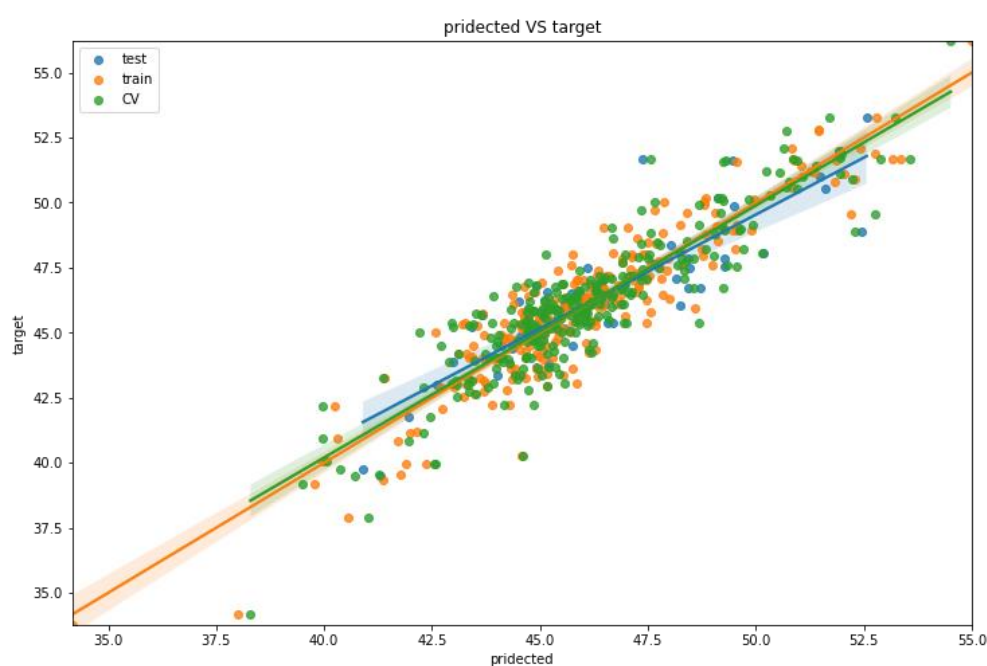


Figure 32 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle PCR

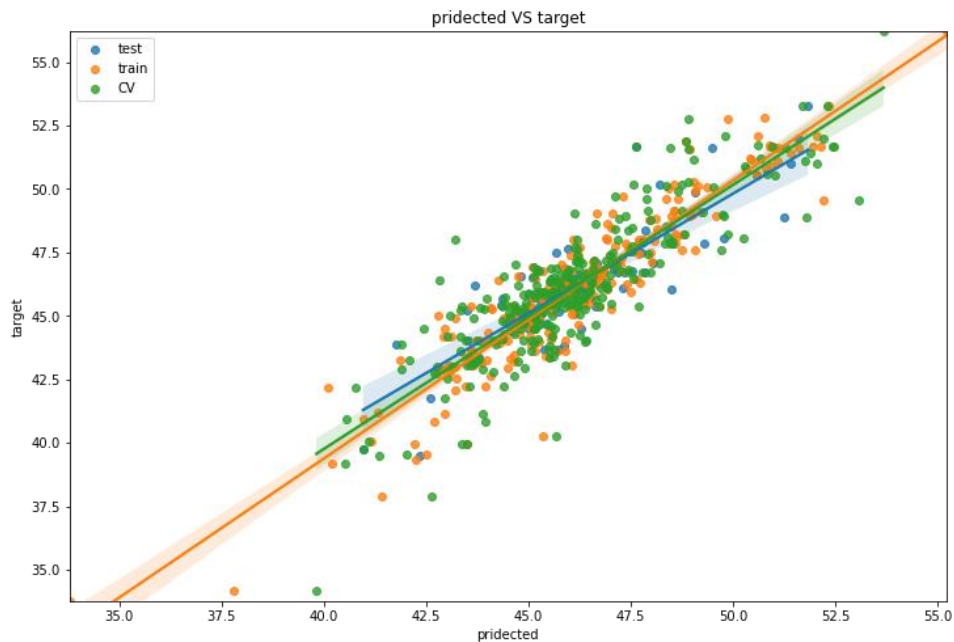


Figure 33 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle SVR

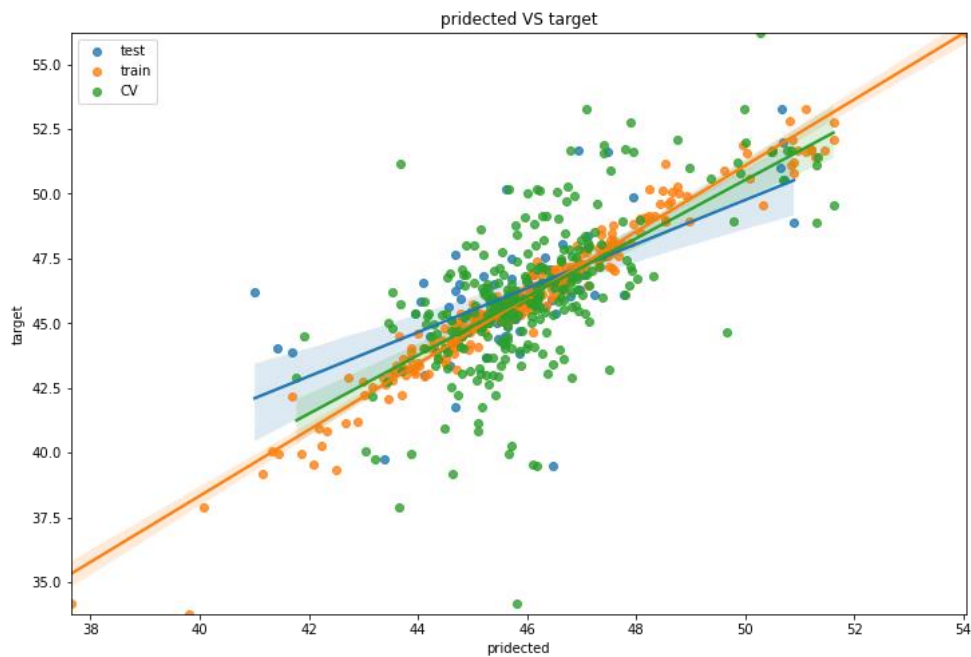


Figure 34 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle RFT

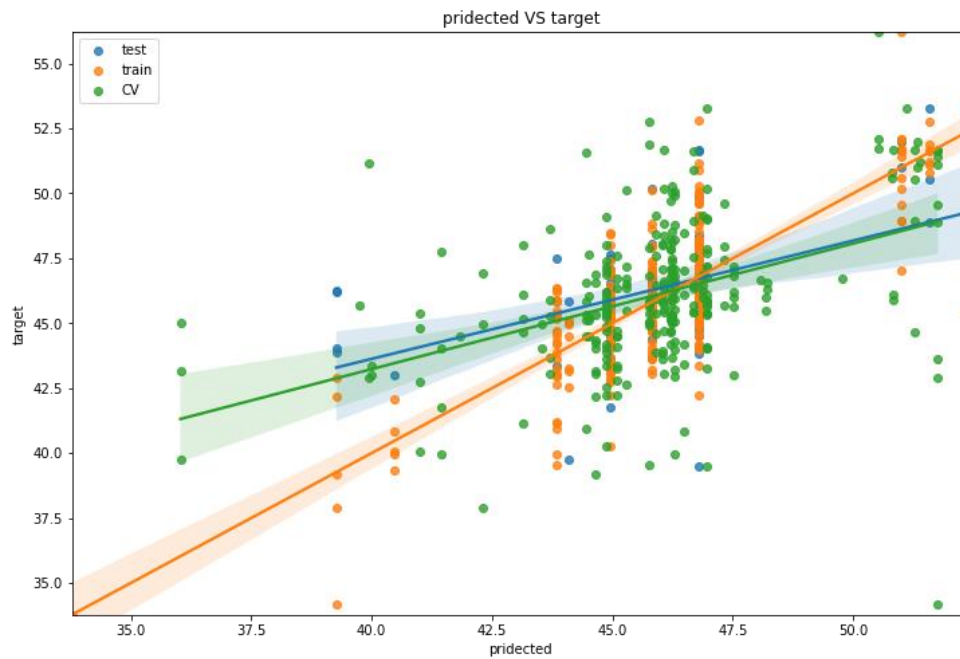


Figure 35 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle DT

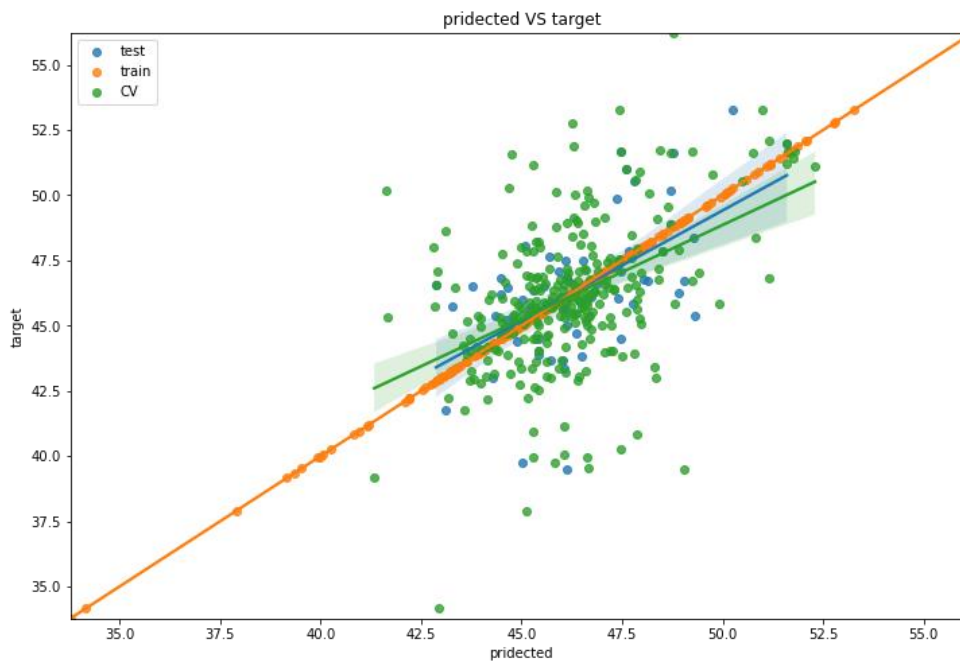


Figure 36 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle KNN

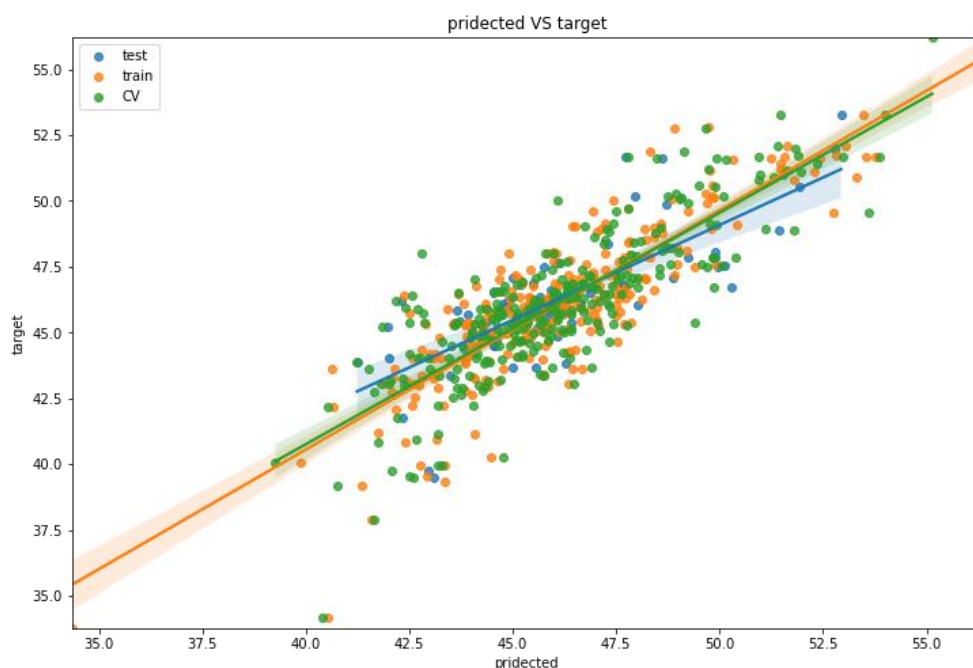


Figure 37 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de carbone – prétraitement 2 - modèle MLP

III.3- Examen de résidu :

Modèle	Shapiro quantile	Probabilité	Décision
PLS	0.975174	0.178233	Normal
PCR	0.970662	0.0993596	Normal
SVR	0.412517	3.03502e-15	Not Normal
RFT	0.913702	0.000140899	Not Normal
DT	0.917336	0.000201651	Not Normal
KNN	0.958939	0.0218844	Not Normal
MLP	0.841949	3.76133e-07	Not Normal

Tableau 6 : Résultats d’examen de normalité de résidus effectué sur les données de Carbone prétraitées – prétraitement 2

III.4- Conclusion :

Tous les modèles sont refusés sauf PLS et PCR, mais on va sélectionner le modèle PLS à cause de ces performances test.

IV- Examen des modèles sur les données prétraitées de Carbone – prétraitement 3 :

IV.1- Prétraitements effectués :

Les prétraitements effectués sur les données de carbone sont :

- MSC
- Detrend 1

IV.2- Examen des performances :

Modèles	R ² train %	R ² CV %	R ² test %	RMSE train	RMSE CV	RMSE test
PLS	89.264564	81.953756	78.250914	0.932073	1.152594	1.142572
PCR	82.816876	80.307342	66.344030	1.179210	1.204096	1.421327
SVR	99.825657	81.826014	-12673.05569	0.118780	1.159700	27.689205
RF	91.162072	25.327745	39.408932	0.845698	2.420639	1.907073
DT	63.241105	-0.749747	-4.646361	1.724730	2.783435	2.506254
KNN	100.000000	38.291704	41.880002	0.000000	2.180600	1.867781
MLP	88.876196	82.048555	77.454162	0.948782	1.145031	1.163312

Tableau 7 : Résultats d'examen de performances effectué sur les données de Carbone prétraitées – prétraitement 3

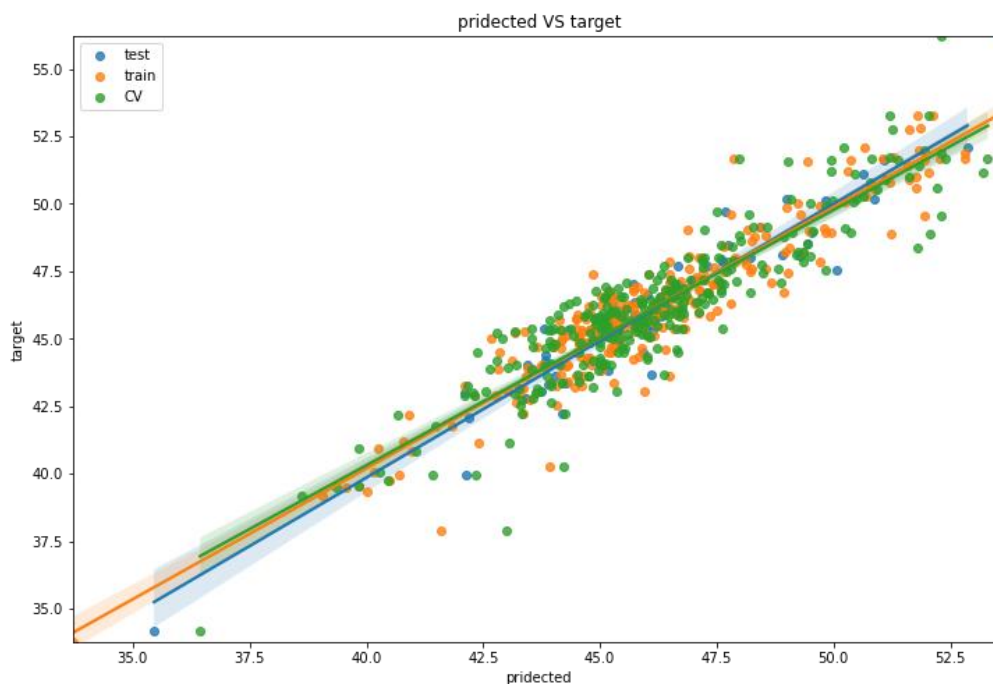


Figure 38 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle PLS

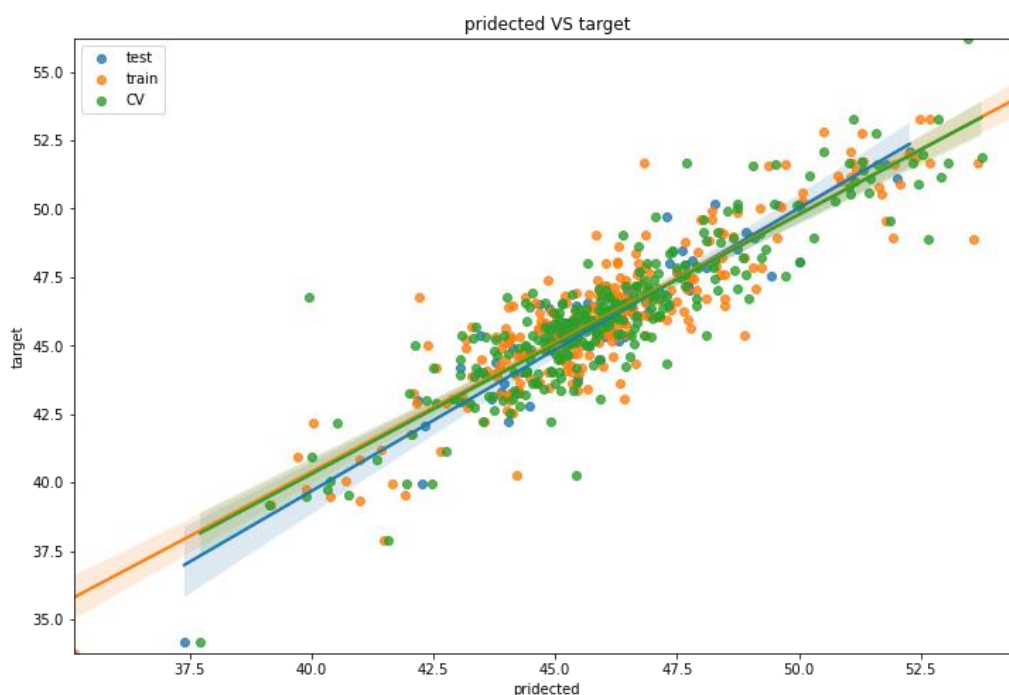


Figure 39 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle PCR

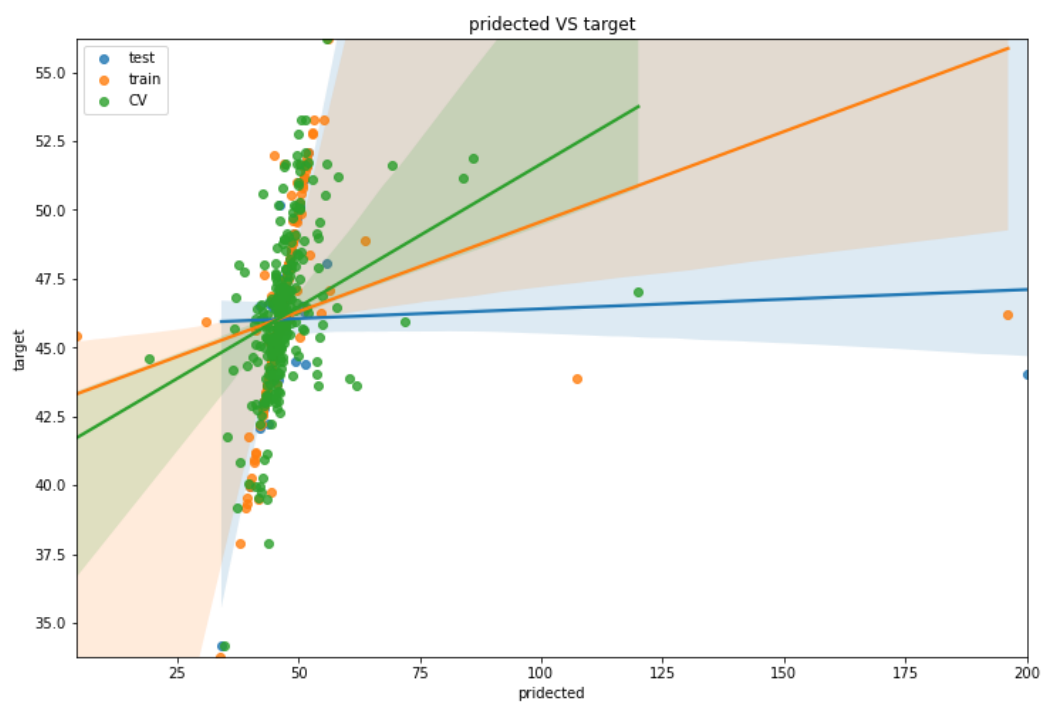


Figure 40 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle SVR

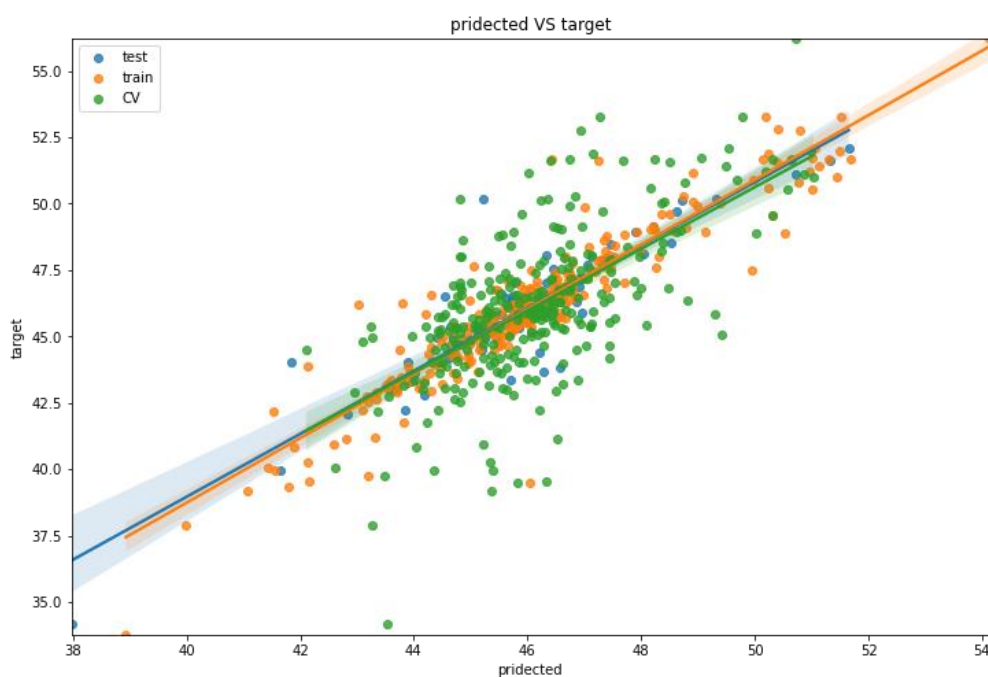


Figure 41 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle RFT

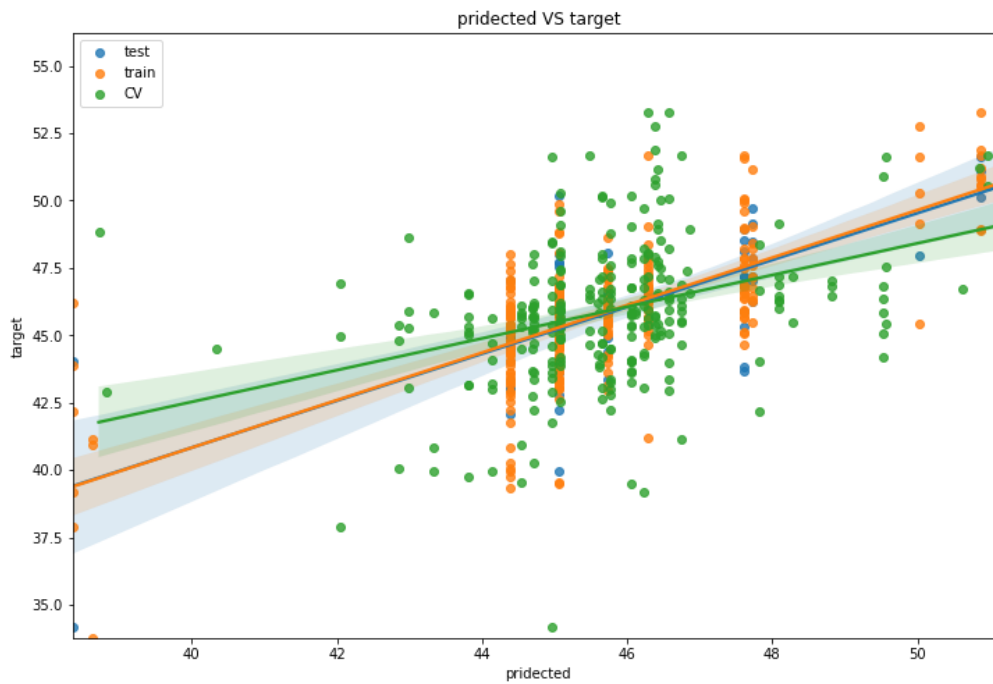


Figure 42 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle DT

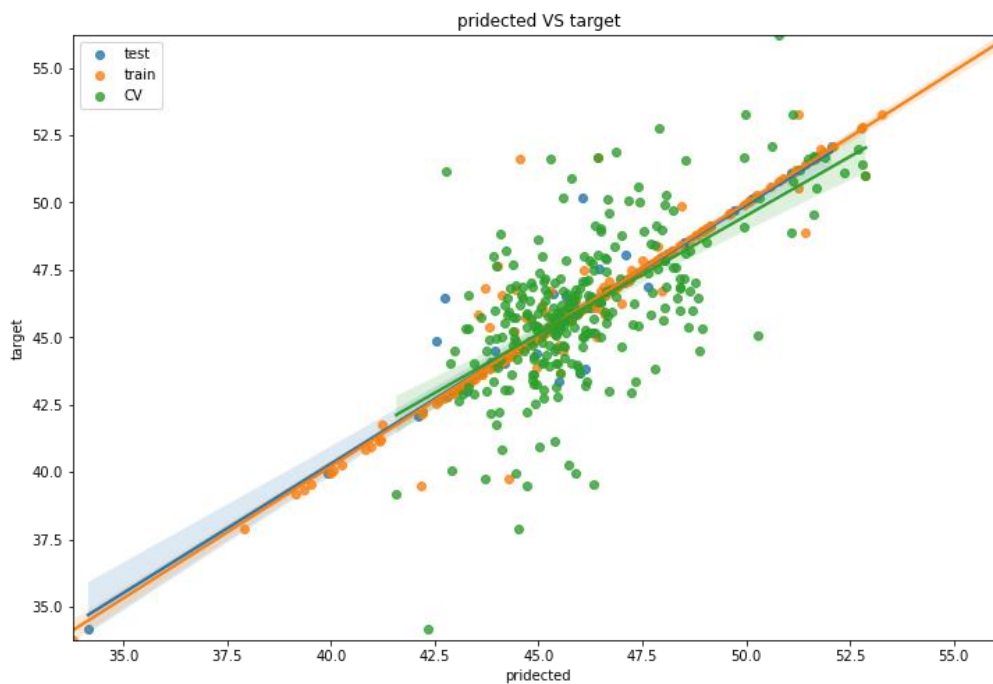


Figure 43 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle KNN

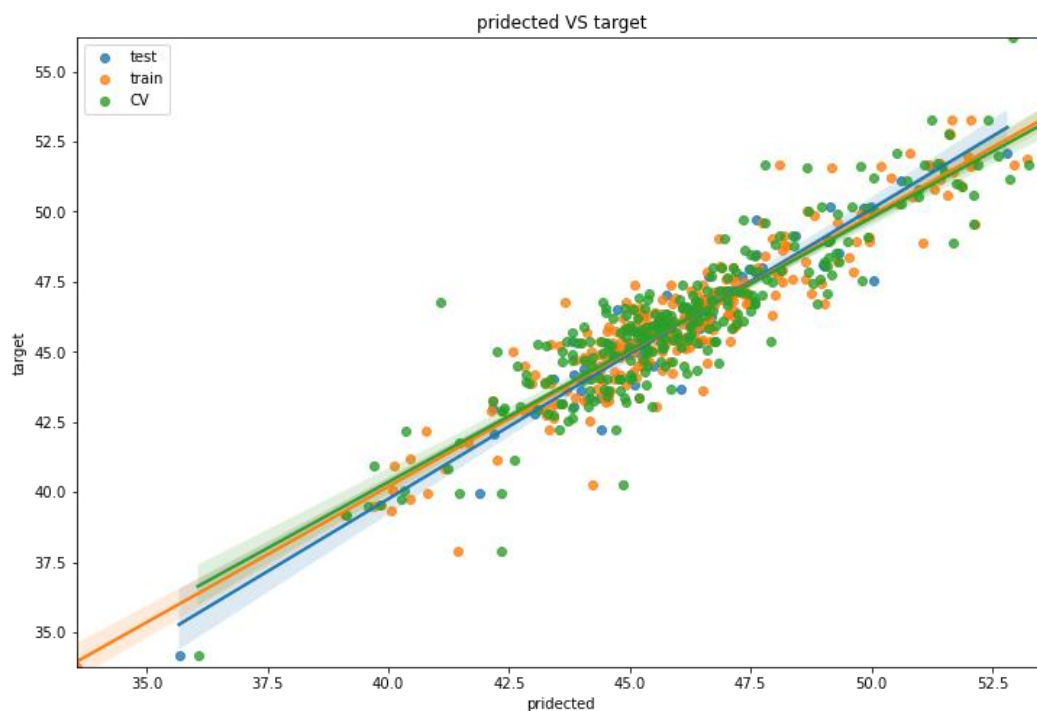


Figure 44 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées de Carbone – prétraitement 3 - modèle MLP

IV.3- Examen de résidu :

Modèle	Shapiro quantile	Probabilité	Décision
PLS	0.967543	0.0661697	Normal
PCR	0.974324	0.159795	Normal
SVR	0.412517	3.03502e-15	Not Normal
RFT	0.913702	0.000140899	Not Normal
DT	0.917336	0.000201651	Not Normal
KNN	0.958939	0.0218844	Not Normal
MLP	0.841949	3.76133e-07	Not Normal

Tableau 8 : Résultats d’examen de normalité de résidus effectué sur les données de Carbone prétraitées – prétraitement 3

IV.4- Conclusion :

Ces examens montrent que le meilleur modèle est : PLS ; le modèle PCR est aussi acceptable ; les autres modèles ont des problèmes de surapprentissage, performances test faibles, manque d'ajustement et/ou une sortie écartée à la normalité.

V- Examen des modèles sur les données prétraitées d'Azote – prétraitement 1 :

V.1- Type des prétraitements effectués :

Les prétraitements effectués sur ces données sont :

- Moving average
- GAP 1
- MSC

V.2- Examen des performances :

Modèles	R ² train %	R ² CV %	R ² test %	RMSE train	RMSE CV	RMSE test
PLS	96.022713	92.740618	95.291439	0.204742	0.266394	0.254422
PCR	94.720422	93.285529	94.522784	0.235892	0.257405	0.274405
SVR	97.195201	92.456015	95.476329	0.171935	0.273964	0.249377
RFT	97.207821	77.700086	81.071604	0.171548	0.470464	0.510115
DT	77.975511	54.688459	50.434044	0.481799	0.666462	0.825473
KNN	100.000000	50.501271	58.714749	0.000000	0.701746	0.753371
MLP	91.822698	89.519498	93.318739	0.293574	0.321169	0.303068

Tableau 9 : Résultats d'examen de performances effectué sur les données d'Azote prétraitées
– prétraitement 1

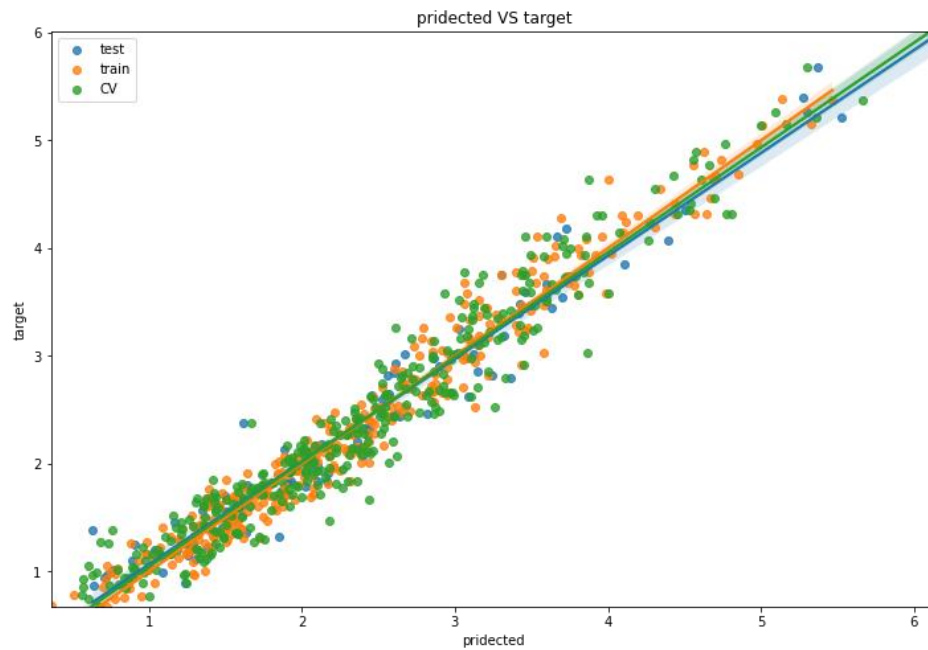


Figure 45 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 1 - modèle PLS

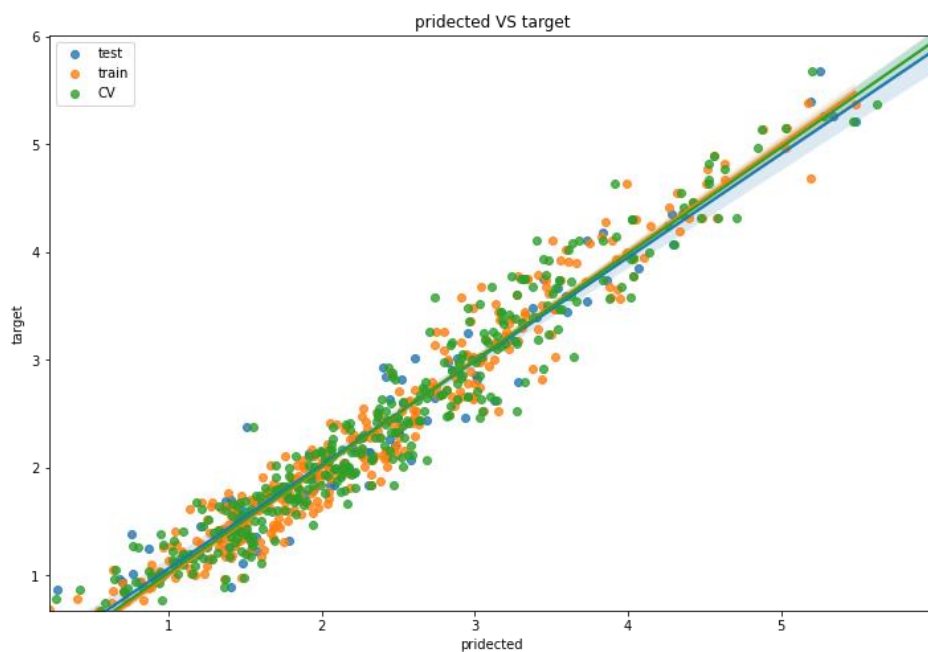


Figure 46 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 1 - modèle PCR

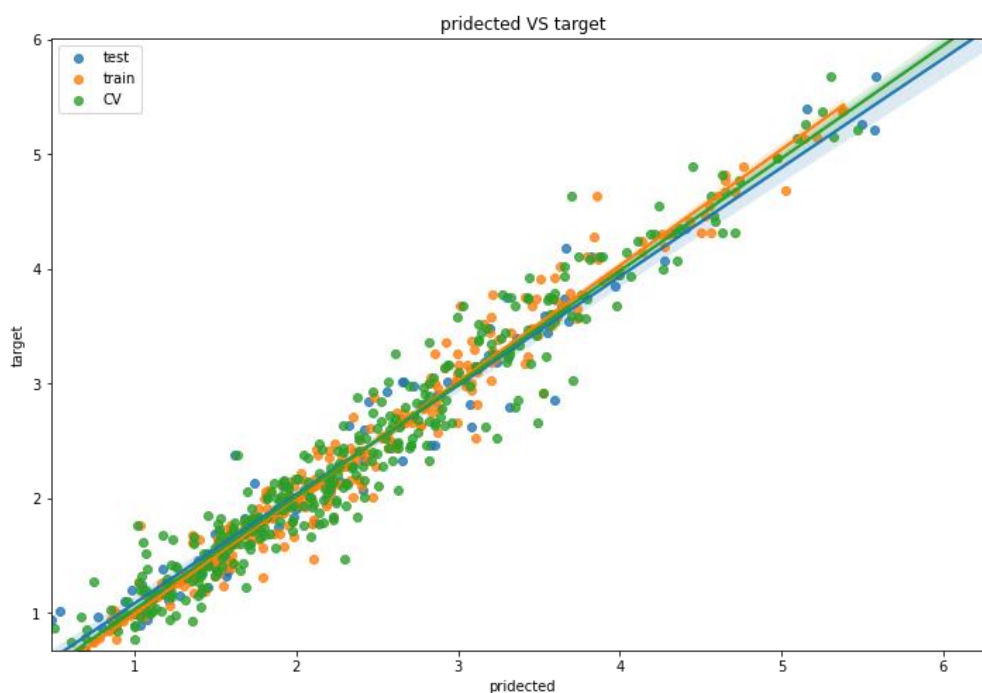


Figure 47 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote – prétraitement 1 - modèle SVR

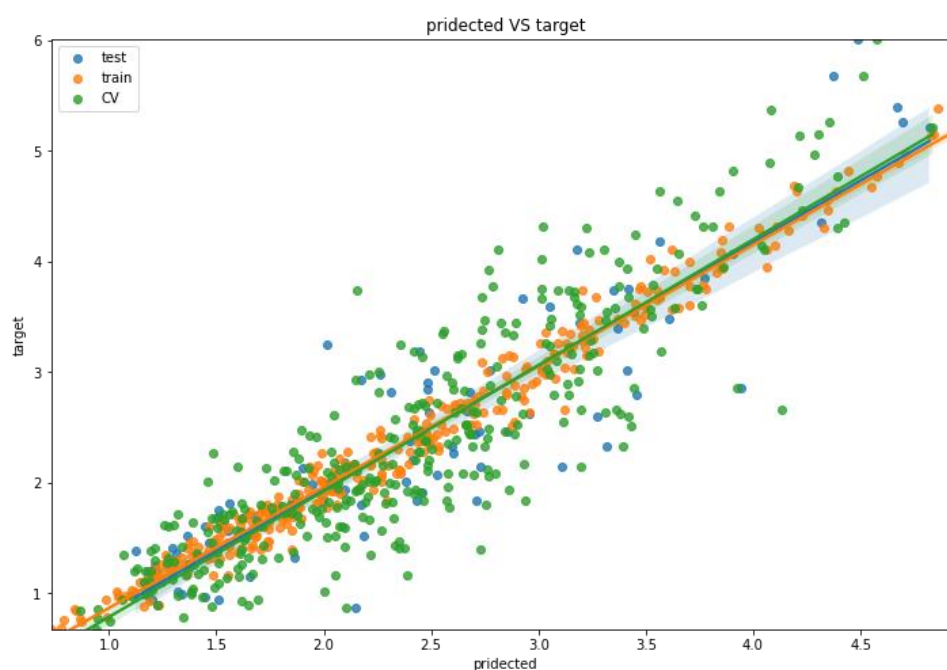


Figure 48 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées d'Azote – prétraitement 1 - modèle RFT

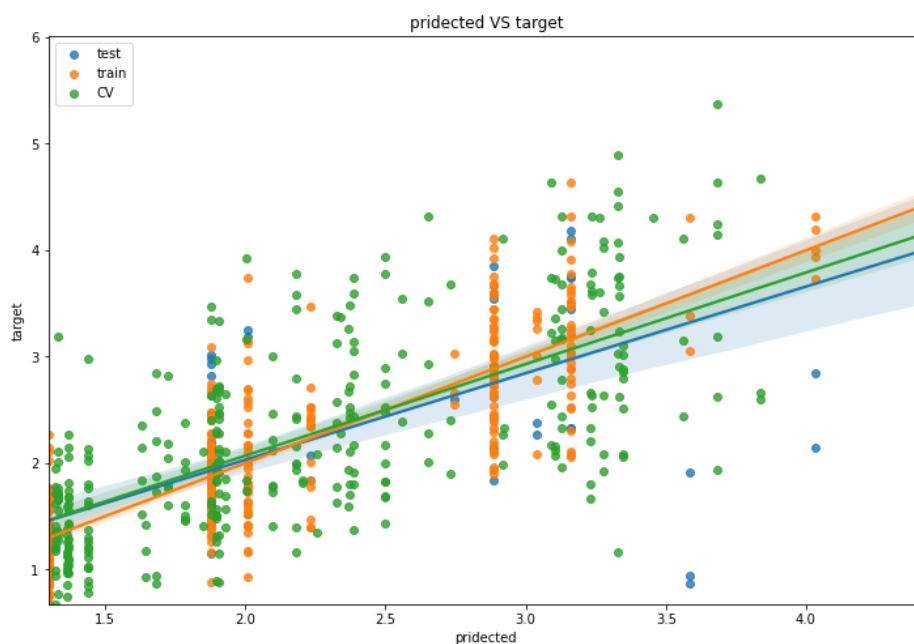


Figure 49 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 1 - modèle DT

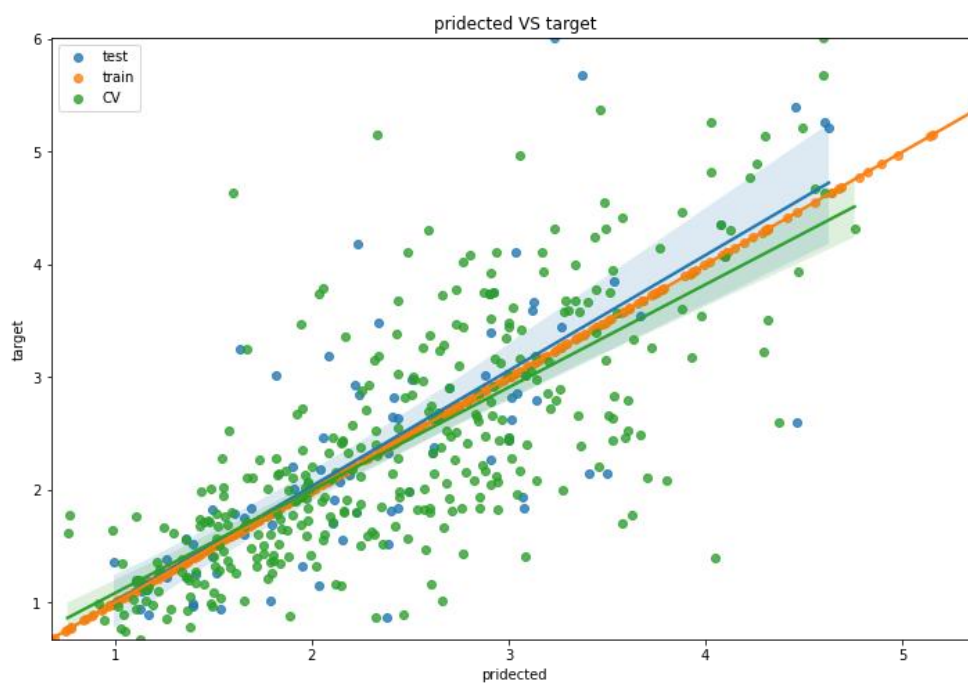


Figure 50 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 1 - modèle KNN

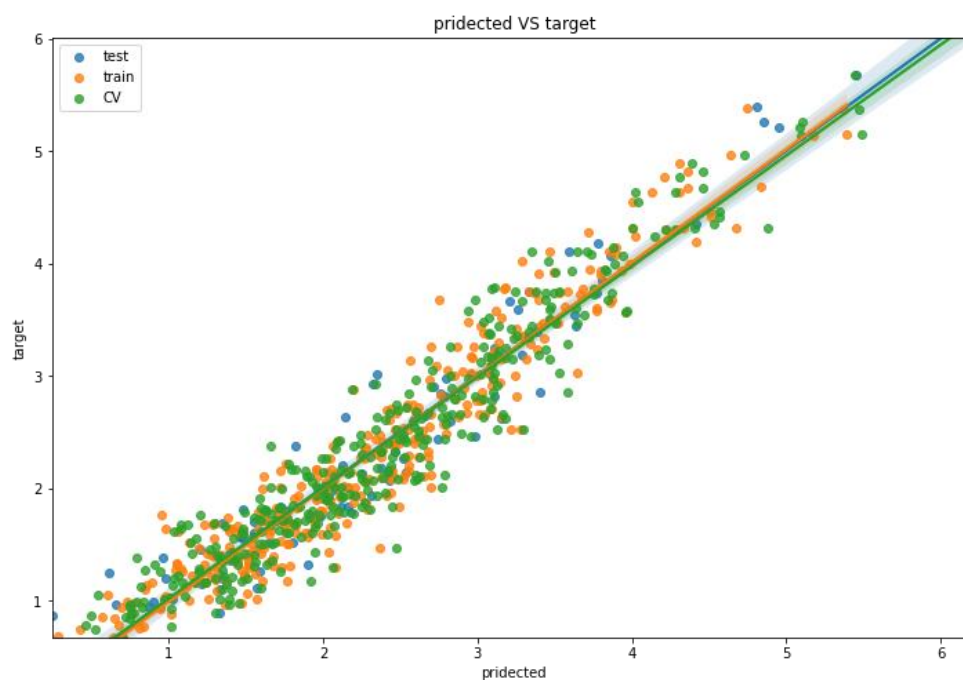


Figure 51 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées
d’Azote – prétraitement 1 - modèle MLP

V.3- Examen de résidu :

Modèle	Shapiro quantile	Probabilité	Décision
PLS	0,985022	0,41566	Normal
PCR	0,972071	0,056522	Normal
SVR	0.989119	0.688083	Normal
RFT	0,95676	0,005434	Not Normal
DT	0,945476	0,001117	Not Normal
KNN	0,939239	0,000491	Not Normal
MLP	0,989441	0,711101	Normal

Tableau 10 : Résultats d’examen de normalité de résidus effectué sur les données d’Azote
prétraitées – prétraitement 1

Et pour le critère RPD :

PLS	4.05443
PCR	3.75918
SVR	3.75918
RFT	2.02217
DT	1.24963
KNN	1.36923
MLP	3.40365

V.4- Conclusion :

Les modèles RFT, DT et KNN sont éliminés car ils donnent des sorties écartées de la normalité, subissent un surapprentissage et les performances de test ne sont pas satisfaisantes.

Pour le test d'ajustement, les modèles RFT, DT et KNN sont éliminés car ils ne sont pas ajustés.

Le critère RDP montre que le modèle PLS est le meilleur choix.

VI- Examen des modèles sur les données prétraitées d'Azote – prétraitement 2 :

VI.1- Type des prétraitements effectués :

Les prétraitements effectués sur ces données sont :

- Moving average
- Detrend1
- MSC

VI.2- Examen des performances :

Modèles	R ² train %	R ² CV %	R ² test %	RMSE train	RMSE CV	RMSE test
PLS	95.224387	92.487788	93.596178	0.235779	0.287697	0.258599
PCR	92.769572	90.003385	92.410823	0.290117	0.331932	0.281516
SVR	99.959802	-6.740467	90.340057	0.021632	1.100824	0.317610
RFT	98.335348	86.670602	92.250511	0.139204	0.383500	0.284474
DT	87.420431	75.515414	80.991817	0.382670	0.517324	0.445530
KNN	100.000000	39.782895	51.875883	0.000000	0.820815	0.708905
MLP	87.883117	82.065391	87.482990	0.375566	0.445429	0.361540

Tableau 11 : Résultats d'examen de performances effectué sur les données d'Azote prétraitées – prétraitement 2

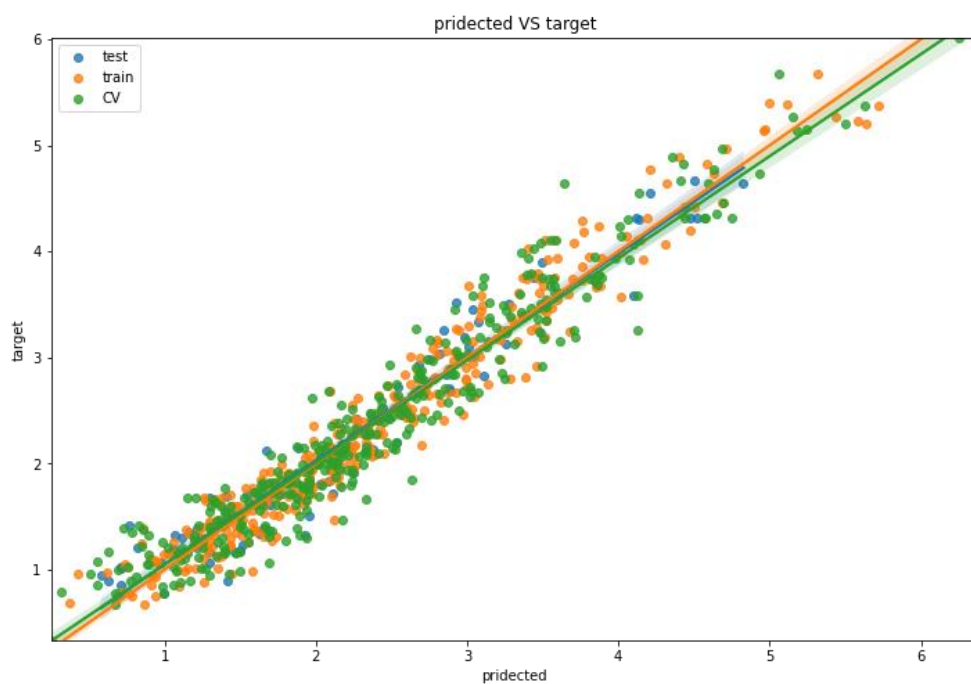


Figure 52 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle PLS

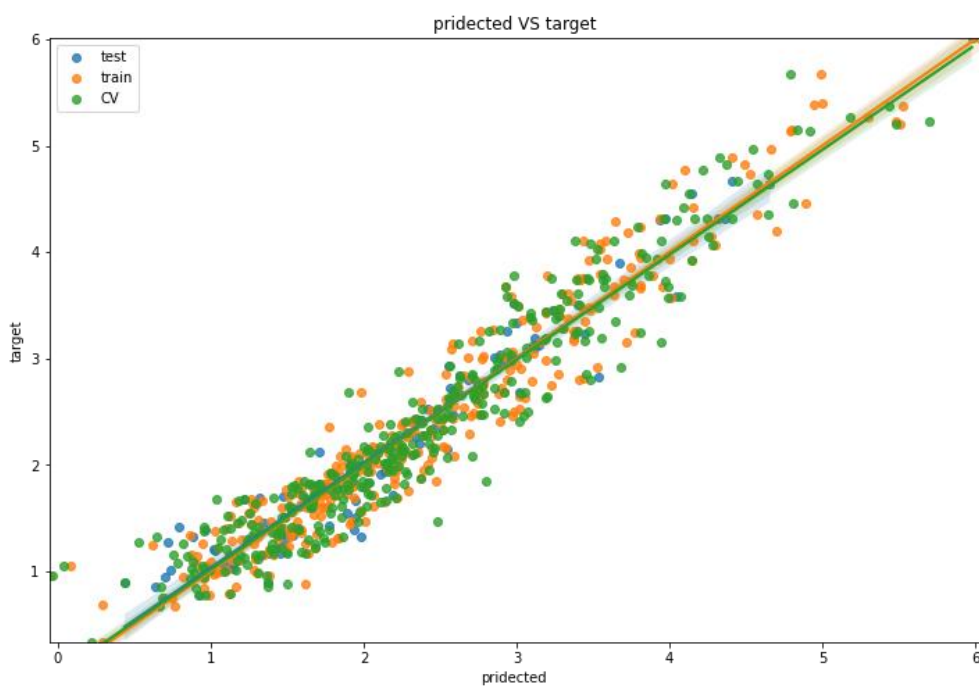


Figure 53 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle PCR

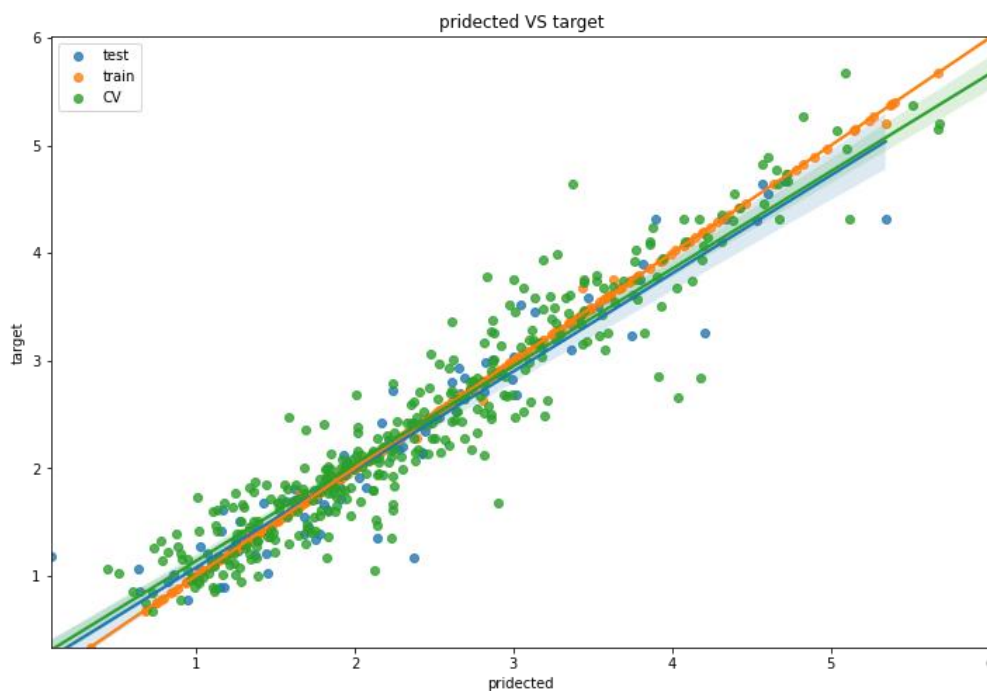


Figure 54 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle SVR

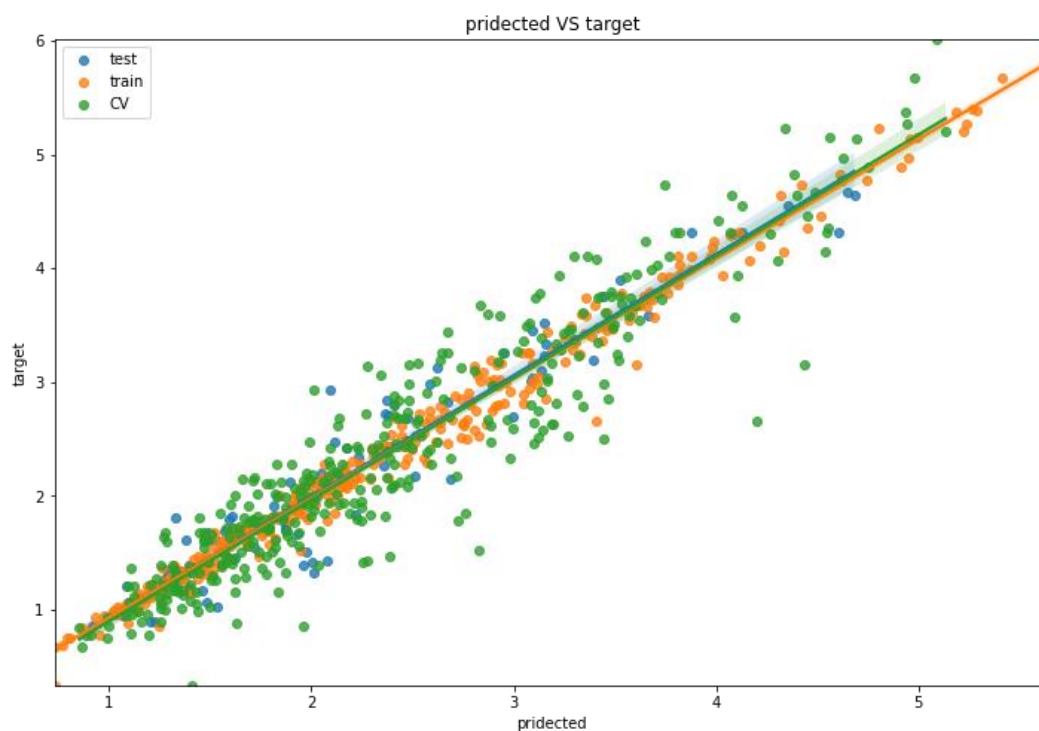


Figure 55 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle RFT

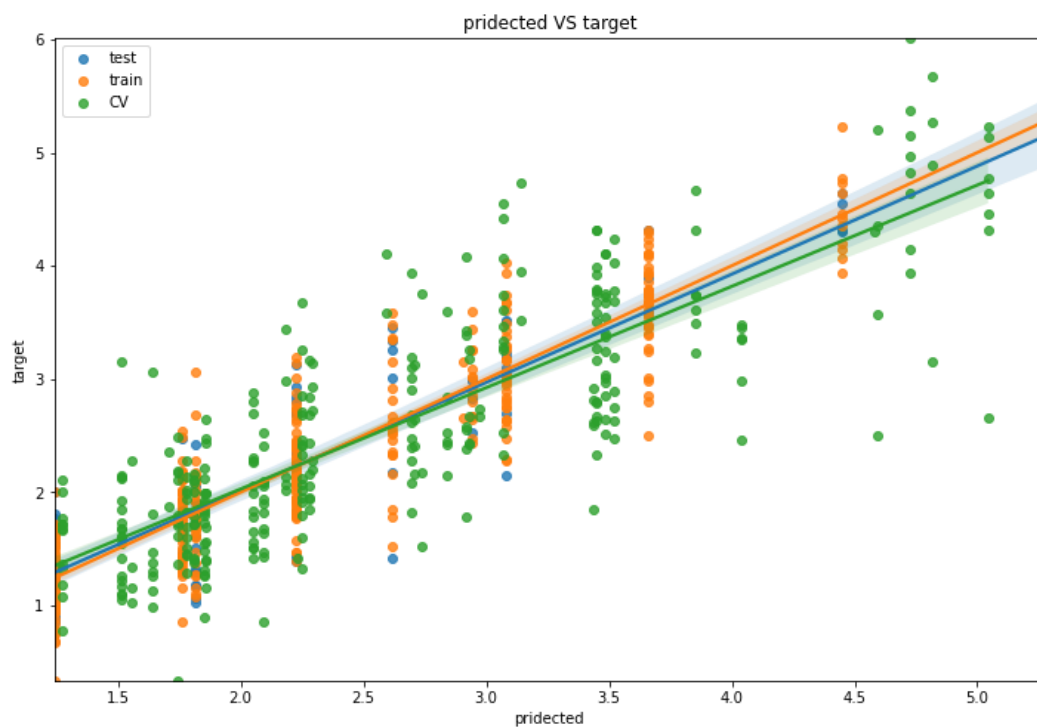


Figure 56 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle DT

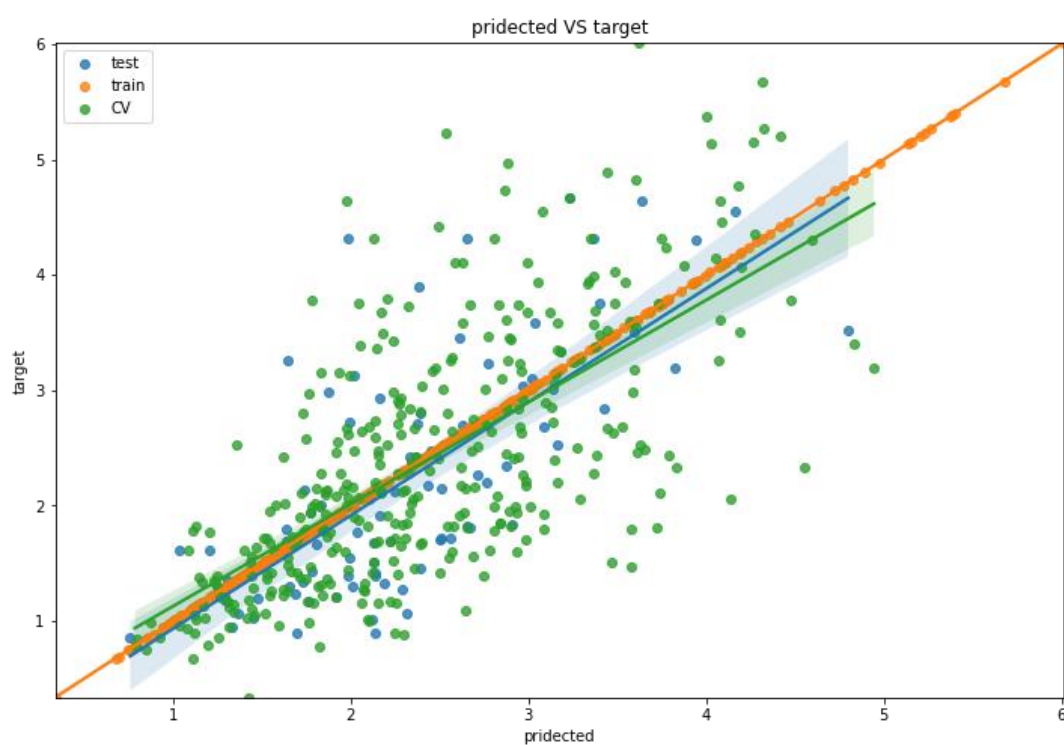


Figure 57 : Examen d'ajustement entre les réponses prédites et réelles - données prétraitées
d'Azote – prétraitement 2 - modèle KNN

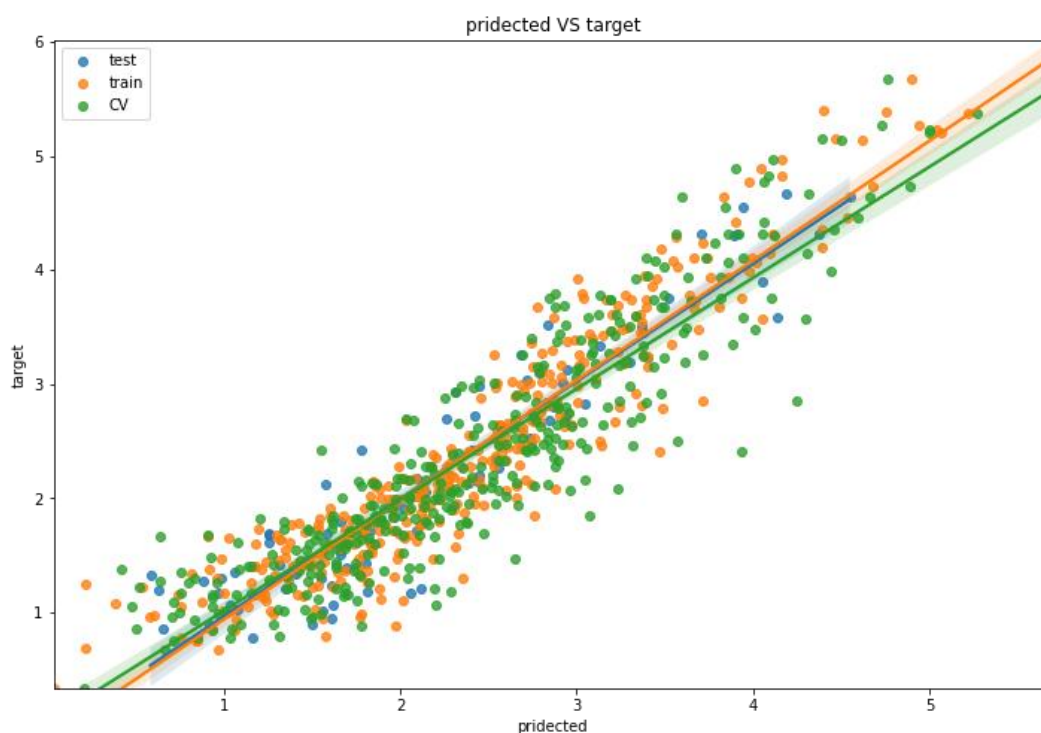


Figure 58 : Examen d’ajustement entre les réponses prédites et réelles - données prétraitées d’Azote – prétraitement 2 - modèle MLP

VI.3- Examen de résidu :

Modèle	Shapiro quantile	Probabilité	Décision
PLS	0.989845	0.726343	Normal
PCR	0.992328	0.888732	Normal
SVR	0.90932	1.20068e-05	Not Normal
RFT	0.507382	9.01461e-16	Not Normal
DT	0.373778	1.09604e-17	Not Normal
KNN	0.947798	0.00132061	Not Normal
MLP	0.983805	0.334737	Normal

Tableau 12 : Résultats d’examen de normalité de résidus effectué sur les données d’Azote prétraitées – prétraitement 2

Et pour le critère RPD :

PLS	3.974056
PCR	3.650533
SVR	3.650533
RFT	3.612577
DT	2.306659
KNN	1.449680
MLP	2.842519

VI.4- Conclusion :

Les modèles RFT, DT et KNN sont éliminés car ils donnent des sorties écartées à la normalité, subissent un surapprentissage et les performances de test ne sont pas satisfaisantes.

Pour le test d'ajustement, les modèles RFT, DT et KNN sont éliminés car ils ne sont pas ajustés.

Le critère RDP montre que le modèle PLS est le meilleur choix.

Conclusion générale :

Dans ce sujet, on a essayé de construire, valider et tester un modèle qui prédit les quantités d'Azote et Carbone totales dans les plantes de tomate cerises en pourcentage par 100g de la matière sèche sans écarts provoqués par effet de matrice, matériels utilisés ou grandeurs incontrôlables (lumière naturelle, climat...) et avec un coût minimal.

Basant sur la méthode PIR-FTIR et exactement par le procédé de réflectance avec une méthode de référence Kjeldahl pour l'Azote et perte de masse pour le Carbone, on a réussi à délivrer deux modèles de prédiction :

- Un modèle PLS d'algorithme NIPALS, basé sur 17 variables latentes et entraîné avec les données prétraitées par Moving average+Gap1+MSC qui prédit la quantité totale d'Azote.
- Un modèle PLS d'algorithme NIPALS, basé sur 20 variables latentes entraîné avec les données prétraitées par MSC+Detrend1 qui prédit la quantité totale de Carbone.

En termes de performances, ces modèles sont acceptables, mais ils ont besoin de plus d'amélioration. Cette étape d'amélioration a besoin de plus de matériel et d'expériences. Nous pensons que si ce problème de manque de matériel est surmonté, on peut améliorer les performances des modèles faibles ou bien essayer d'autres algorithmes d'apprentissage qui peuvent donner des résultats plus performants.

Pour la livraison des modèles de prédiction, on peut livrer le modèle de prédiction de quantité d'Azote avec un risque très faible, et son homologue de prédiction de Carbone avec un risque un peu plus élevé que le premier modèle, ce qui fait que dernier modèle sera valable uniquement pour la description des échantillons.

Références :

- [1]: Tarek Hamdi, [Machine Learning: Types of Problems and Tasks](#) , 19/06/2017.
- [2]: Jill C. Stoltzfus PhD, [Logistic Regression: A Brief Primer](#) , 13/10/2013.
- [3]: David Aymonin, [Projet de recherche : Utilisation des semi-conducteurs dans la purification de l'eau à la présence de lumière - école nationale des bibliothécaires](#) , promotion 1988-1989.
- [4]: Neil Salkind, [Partial least square Regression, Hervé Abdi](#) , 2007
- [5]: Rustam Guliev, [Understanding NIPALS algorithm for PLS](#) , 25/03/2017
- [6]: Statistics4U Courses and Tutorials, [Principal components regression](#)
- [7]: Lucas Eduardo de Oliveira Aparecido, Glauco de Souza Rolim, Jose Reinaldo da Silva Cabral De Moraes, Cicero Teixeira Silva Costa et Paulo Sergio de Souza, [Machine learning algorithms for forecasting the incidence of Coffee](#) , 08/12/2019.
- [8]: Alakh Sethi, [Support Vector Regression Tutorial for Machine Learning](#) , 27/03/2020.
- [9]: Pranshu Sharma, [Decision Tree Classification | Guide to Decision Tree Classification](#) , 29/04/2021.
- [10]: Jason Brownley, [Probabilistic Model Selection with AIC, BIC and MDL](#) , 30/10/2019.
- [11]: Silvia Sans, Joan Ferré, Ricard Boqué, José Sabaté, Joan Casals and Joan Simó, [Estimating Sensory Properties with Near-Infrared Spectroscopy: A Tool for Quality Control and Breeding of 'Calçots' \(Allium cepa L.\)](#) , 11/06/2020.

