Master TRIED

Rapport : Etude de cas 1

Réalisé par :

Karim ASSAAD

Année universitaire :

2016/2017

Table des matières

Introduction :

Le formaldéhyde est l'un des principaux polluants intérieurs, Il est présent dans les environnements domestiques et professionnels puisqu'il entre dans la fabrication de plusieurs produits ménagers de ce fait son analyse en fonction des différentes variables de notre jeu de donnée seraient bénéfiques pour les études concernant la pollution de l'air.

Le jeu de donnée utilisée est une matrice de 126 variables et 567 individu dont 17 individus ayant des valeurs manquantes pour le formaldéhyde.

Tout au long de ce rapport je vais essayer d'étudier la variable explicative par le biais d'une analyse descriptive détaillé pour pouvoir par la suite construire un modèle prédictif adéquat pour la prédiction des valeurs manquantes.

Problématique :

L'objectif de cette étude est de prédire les valeurs manquantes du Formaldéhyde.

Pour ce faire on va commencer par effectuer une analyse descriptive des données, puis trouver les variables les plus corrélées et anti-corrélées avec le Formaldéhyde afin de les utiliser comme variables explicatives. Enfin on va créer plusieurs modèles de régression puis étudier leurs performances afin de choisir le meilleur modèle sur lequel on va se baser pour prédire les valeurs manquantes.

Présentation de la variable à expliquer (Formaldéhyde) :

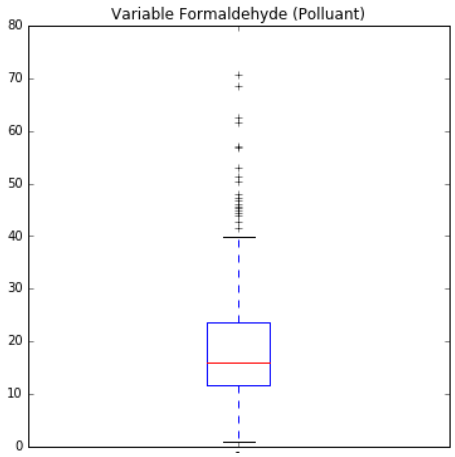


Figure 1 :Boxplot du Formaldéhyde

La variable explicative que j'ai utilisée est le Formaldéhyde, ce choix se justifie par l'importante présence de ce polluant dans notre vie quotidienne, en effet ce polluant est omniprésent dans l'air, on le trouve dans les produits basiques qu'on utilise tous les jours sans nous rendre compte des dangers qu'il peut causer de ce fait il est intéressant de traiter les données qui le prennent en compte afin de résoudre les problèmes liés aux effets de cette entité sur la composition de l'air.

D'après le Boxplot du Formaldéhyde on peut voir qu'il est possible d'avoir des valeurs aberrantes, mais pour l'instant on va pas les traiter.

Analyse descriptive :

Analyse unidimensionnel :

Tout d'abord j'ai commencé par analyser les variables quantitatives dont le nombre est 60 en appliquant différents tests tels que le test de Shapiro et de Chi2 pour voir si les données suivent les lois usuelles normale ou uniforme.

Les tests ont abouti à une distribution presque égale entre les variables, 17 variables suivent la loi normale et 16 suivent la loi uniforme.

Analyse bidimensionnel :

Avec des variables quantitatives

Dans un deuxième temps je me suis intéressé aux variables quantitatives en étudiant leurs dépendances avec le formaldéhyde par le biais du nuage des points. Selon les résultats obtenus j'ai conclu que les variables n'entretiennent aucune liaison avec le formaldéhyde que ce soit linaire, logarithmique ou autre.

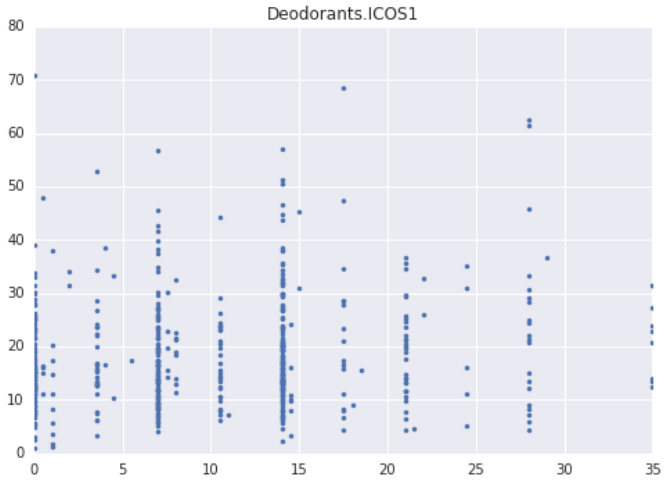


Figure 2 : Représentation du Formaldéhyde en fonction du Déodorant

L'exemple ci-dessus prend en compte le nuage des points pour la variable déodorant, on voit qu'il n'existe pas de liaison linéaire et c'est le cas pour la plupart des variables donc on ne peut en aucun cas les utiliser pour implémenter un modèle linéaire univarié.

Avec des variables qualitatives

En présentant les Boxplot des variables en fonction du formaldéhyde, on a obtenu des représentations similaires dont voici un exemple.

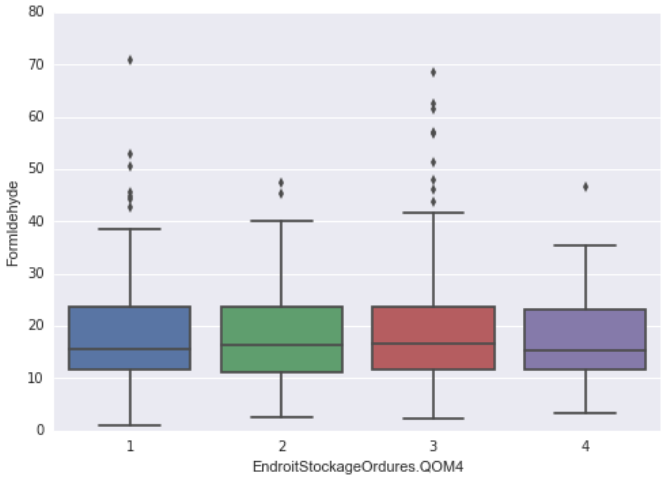


Figure 3 : Représentation du formaldéhyde en fonction de Endroit-Stockage-Ordures

Dans ce cas il s'agit d'une variable à quatre modalités dont on a présenté les box plots en fonction du formaldéhyde. Les moyennes, min et max des différentes modalités sont très proches l'une de l'autre donc il est quasi impossible de deviner à quelle classe appartiennent les valeurs des formaldéhydes. Ici par exemple, on est incapable d'attribuer la valeur 20 à une classe particulière donc les modalités de la variable en fonction du formaldéhyde sont linéairement inséparables.

Corrélation linéaire et de Spearman

Dans cette partie, J'ai calculé les corrélations linéaires et la corrélation de Spearman de toutes les données quantitatives une par une avec le formaldéhyde mais malheureusement tous les résultats étaient clairement mauvais ce qui me ramène à dire que les données utilisées ne sont pas adéquates pour pouvoir atteindre un bon résultat dans la création d'un modèle linéaire. Cette conclusion confirme donc toutes les informations obtenues précédemment.

Analyse multidimensionnelle :

Analyse par composante principale

Cette analyse servira à voir s'il existe une liaison linéaire entre les combinaisons des variables et le formaldéhyde. Pour Cela j'ai choisi d'établir dans un premier temps une ACP sur les différents groupements de variables. On a donc trois groupements, le premier représente les variables liées aux habitudes, le deuxième représente les variables liées aux ménages et le troisième représente les variables liées aux logements. Une fois les axes choisis et les corrélations établies, j'ai choisi de sélectionner les variables corrèles et anti corrèles avec le polluant qui est considéré comme une variable supplémentaire (Cela est fait en implémentant une fonction qui détecte les variables dont les coordonnes ont le même signe que celles de la variable supplémentaire sur les 2 axes principaux).

On ne sait pas au préalable ce que cette méthode va retourner néanmoins, elle servira à confirmer ou rejeter les conclusions vues précédemment concernant la linéarité des variables et mettra en évidence les variables liées aux facteurs causant le formaldéhyde.

Les résultats de l'ACP sont visualisés par le tableau ci-dessous :

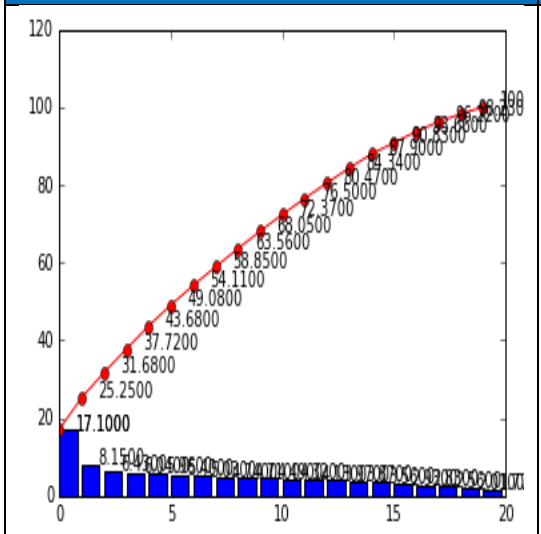
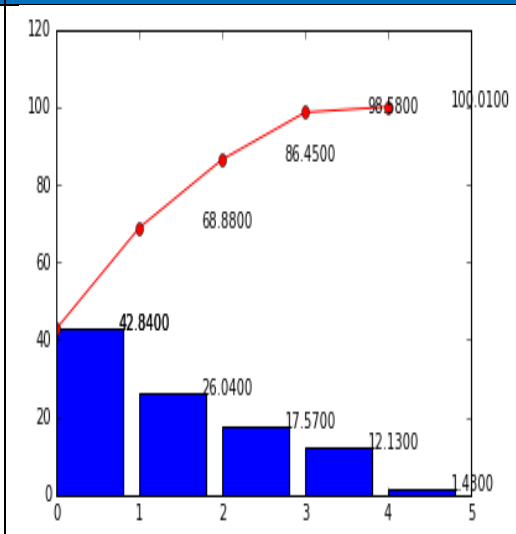
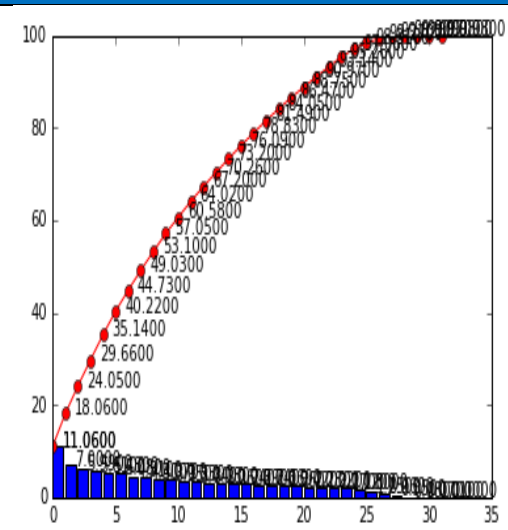
Habitude	Menage	Logement
		
On choisit 10 composantes	On choisit 3 composantes	On choisit 15 composantes

Figure 4 : Représentation des inertie et des inertie cumuler de chaque topologie.

En prenant en considération les graphes de la variance cumulées, j'ai choisi de prendre 10 composantes principales des variables liées aux habitudes car ils fournissent plus de 70% de l'information, 3 composantes pour les variables de ménages puisqu'ils présentent plus de 75% de l'information ainsi que 15 composantes pour les variables de logement car leurs inerties présentent plus de 70% de l'inertie totale.

On peut voir clairement que les combinaisons linéaires des variables créent par l'ACP ne fournissent pas une quantité suffisante d'informations.

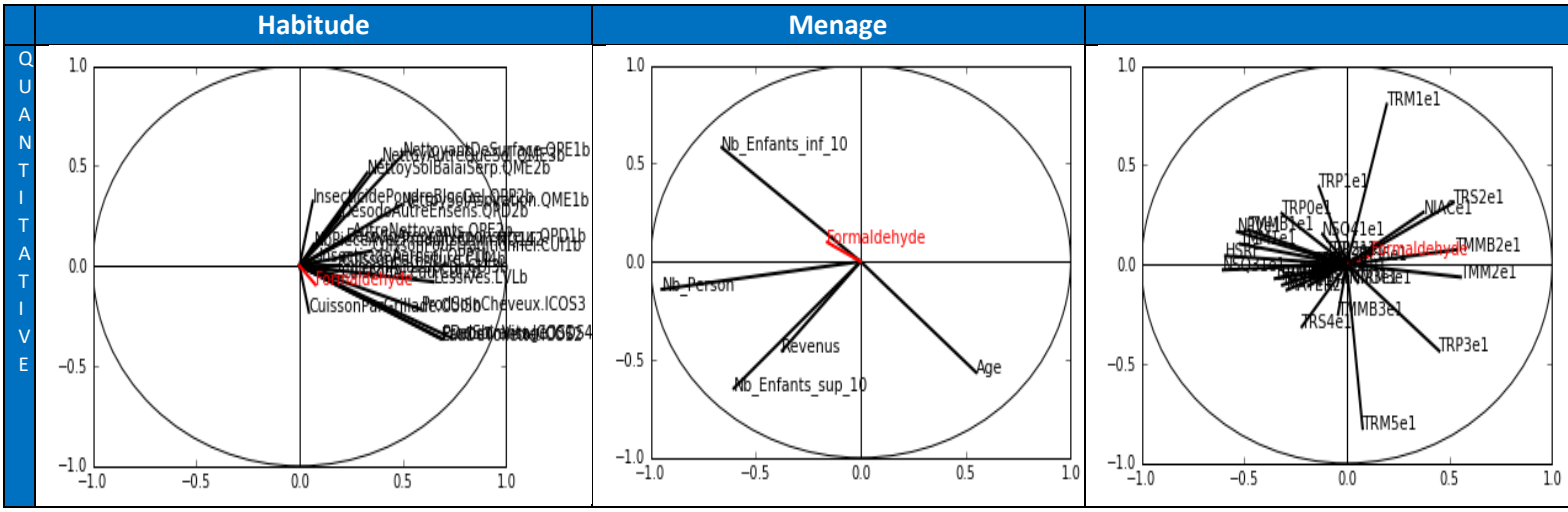


Figure 5 :Représentation des cercles de corrélation pour chaque topologie avec le Formaldéhyde en variable supplémentaire

On a obtenu 33 variables corrèles et 45 non corrèles avec le formaldéhyde.

Selon la visualisation des cercles de corrélations et la projection de la variable supplémentaire on voit que le polluant est faiblement présenté sur les axes principaux. Ici aussi on peut voir clairement que nos conclusions vues dans les partie analyse unidimensionnelle et bidimensionnelle par rapport à non linéarité des variables correspondent au résultat fournie par l'ACP.

Tout cela m'amène à utiliser un modèle non linéaire et rejeter les méthodes de sélection stepwise (AIC,BIC) et les méthodes de régularisation.

Sélection des variables :

Sélection par l'importance des variables via les arbres de décisions :

Pour la sélection des variables, j'ai eu recours à une modélisation en utilisant les arbres décisions car elles fournissent les variables les plus importantes mais cette méthode a retourné une erreur très grande qui sera donnée ultérieurement.

Sélection par variance pour les variables quantitatives et par distribution de fréquence des variables qualitative :

Puisque les méthodes classiques ne retournaient pas un résultat satisfaisant, j'ai choisi une méthode simple et claire pour faire face à la mauvaise distribution des données. Cette méthode consiste à choisir les variables quantitatives qui ont une grande variance et les variables qualitatives dont les modalités sont presque équi-distribués. Par suite les variables obtenues et vont être utilisées pour la modélisation sont :

'MATER2', 'NIACe1','NSQ31e1','TMM1e1','TMM2e1','TMMB1e1','TMMB2e1', 'TMMB3e1','TRM1e1','TRM5e1','TRP0e1', 'TRP3e1','TRS1e1','TRS2e1','TRS3e1', 'TRS4e1','Nb_Enfants_inf_10', 'Nb_Enfants_sup_10', 'Lessives.LVLb', 'Deodorants.ICOS1', 'NettoyantDeSurface.QPE1b', 'AutreNettoyants.QPE2b', 'DesodoAutreEnsens.QPD2b','InsecticidePoudreBlocGel.QPP2b','NettoySolAspiration.QME1b', 'NettoySolBalaiSerp.QME2b', 'NettoyAutreQueSol.QME3b','CuissonFourTraditionnel.CUI1b', 'CuissonALeau.CUI2b', 'CuissonALaVapeur.CUI3b', 'CuissonParGrillade.CUI5b', 'HCU31', 'KCC34', 'KCC5'

Modélisation :

Les modèles :

Les données sont non linaire donc dans cette partie j'ai choisi des méthodes qui correspondent à ce type de données. J'ai effectué de divers modèles sur les données en commençant par les arbres de régression et les Radom Forest en passant par le SVM pour aboutir enfin au PMC. Chacun des modèles était construit à partir des variables sélectionnes précédemment et ont donné des résultats différents.

SVM et Arbre de régression

Tout d'abord j'ai commencé à établir les modèles d'arbres de régression et de SVM de noyau radial mais leurs résultats n'étaient pas satisfaisants car la marge d'erreur était assez importante.

Random Forest

Après avoir effectué l'algorithme 3 fois en prenant n_estimators qui correspondent respectivement à 10, 30 et 100 arbres j'ai choisis un nombre d'arbres égale à 100 puisque cela m'a permis d'avoir les meilleures performances. Comme il s'agit d'un problème de régression on va fixer le paramètre max_features aux nombre de variables explicatives qui est égale à 78.

Random Forest et les arbres de régression sont de types « Ensemble » et donc leurs fonctionnements sont basés sur l'importance des variables. Par la suite je vais utiliser une librairie python pour sélectionner les meilleures variables des modèles et ré-exécuter les deux algorithmes en prenant en compte les variables pertinentes ainsi j'ai pu établir une sur sélection et avoir 41 variables significatives pour ce modèle.

Perceptron Multicouche

J'ai commencé par normaliser les données pour pouvoir établir le modèle.

En prenant seulement 5 neurones cachés je n'ai pas obtenue un résultat concret, j'ai donc augmenté le nombre de neurones cachés.

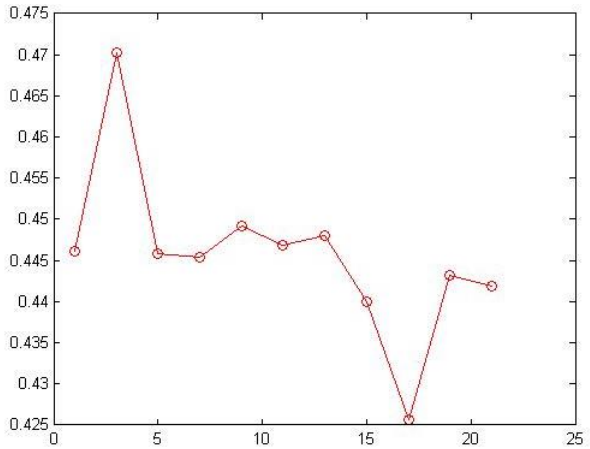


Figure 6 : Représentation des erreurs d'apprentissage des modèles en fonction des neurones cacher

Cette figure démontre qu'il faut prendre 17 neurones pour minimiser l'erreur du modèle, en effet ce nombre correspond à 0.42 qui est la valeur initiale de l'erreur.

Validation :

Pour valider les modèles et obtenir des erreurs de généralisation, j'ai utilisé la validation croisée à trois piliers au lieu d'utiliser une simple division en un ensemble d'apprentissage et de validation car le nombre des individus n'est pas très important aux alentours de 500.

Les erreurs et les performances :

	Arbre de Décision	Random Forest	Arbre de Décision Améliorer	Random Forest Améliorer	SVM (Radial)	PMC
Erreur app	0	0.19	0	0.2	0.52	0.42
Erreur gen	0.83	0.79	0.86	0.78	0.84	0.45
Performance	0.22	0.10	0.31	0.13	0.07	0.47

Tableau 1 : Représentation des erreur d'apprentissage, des erreur général et des performance des diffèrent modèles

Selon le tableau l'erreur d'apprentissage pour les Random forest et les arbres de régression est très négligeable (0 et 0.19) mais cela est dû à un sur apprentissage car l'erreur de généralisation est très importante, pour cela je les ai régularisé mais sans avoir un effet remarquable donc ces modèles sont à rejeter ainsi que le SVM dont la performance est presque nulle.

En me basant sur l'erreur de généralisation et la performance, j'ai conclu que le meilleur modèle est le PMC puisqu'il atteint une performance de plus de 40% mais même avec ces résultats le modèle reste assez moyen et cela est peut-être dû à la nature des données et aux critères du formaldéhyde dont la présence est importante dans l'air.

Clustering :

Dans cette partie, je vais chercher à étudier, s'elle existe, la sous linéarité des variables, pour ce fait j'ai procédé comme suit :

J'ai utilisé le clustering Kmeans pour mettre les données dans différentes classes que je vais étudier par la suite une par une. Le meilleur clustering est obtenue avec 8 classes. J'ai donc répartie mes données en 8 classes et j'ai obtenue une distribution à peu prés égales entre les différentes classes, comme indiqué dans le tableau ci-dessous.

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Nombre	51	85	73	49	62	**95**	71	44

Tableau 2 : Le nombre des individus dans chaque classe

Cet exemple prend en compte la variable déodorant normalisé en fonction du formaldéhyde pour chacune des classes

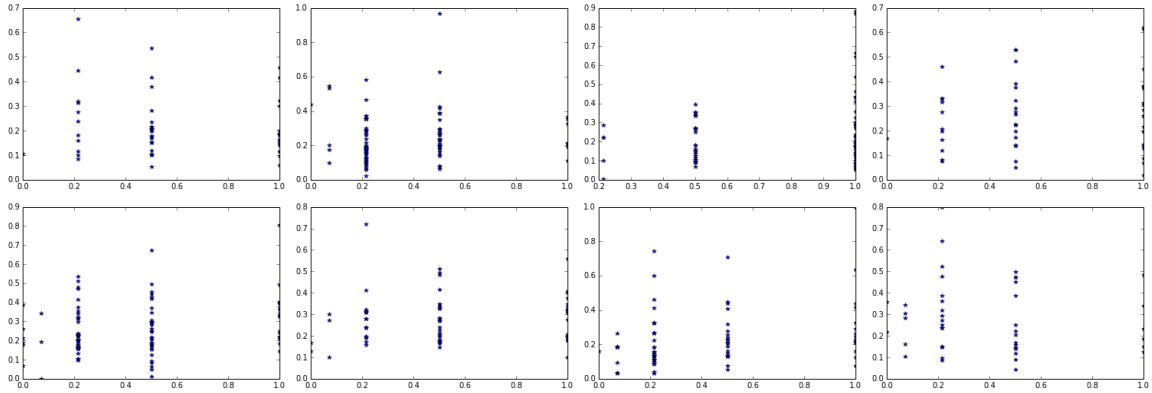


Figure 7 : Représentation du Formaldéhyde en fonction de la variables Déodorant pour chaque classe

On peut voir que la répartition en classe n'a pas améliorer la dépendance entre le formaldéhyde et le déodorant, c'est le cas aussi pour le reste des variables.

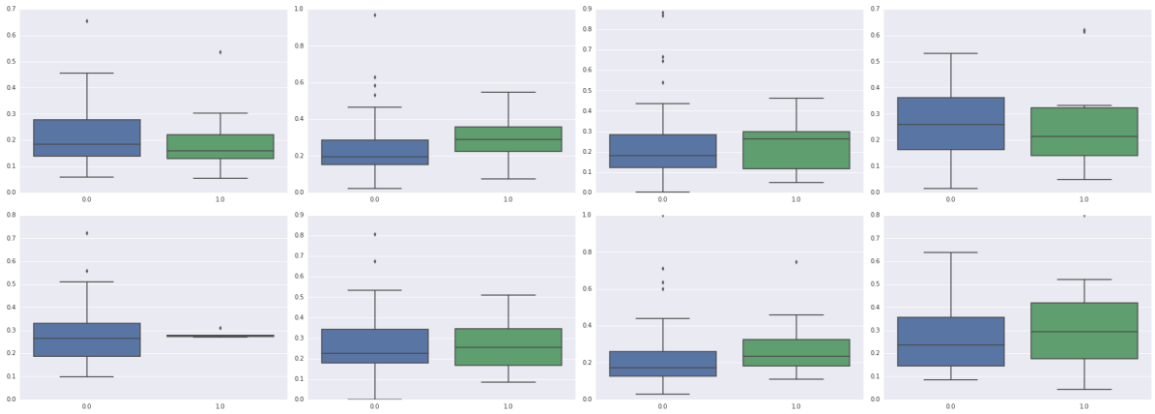


Figure 8 : Représentation du Formaldéhyde en fonction de la variables HCU31 pour chaque classe

En présentant les boxplots de la variable HCU31 en fonction du formaldéhyde pour chaque classe, on remarque qu'il y a une légère amélioration mais les données restent linéairement inséparables.

Modélisation sur les clusters :

Par cette méthode, j'espérais obtenir des classes avec des données plus linéaires avec le Formaldéhyde mais la linéarité reste très négligeable et les modèles linéaires ne semblent toujours pas marcher ce qui m'a amené à réutiliser le PMC qui était le modèle le plus performant dans le cas général. Pour ce fait, j'ai fait un modèle PMC avec une architecture basée sur 17 neurones cachés pour les classes 1, 4, 5, 7 et 8 en utilisant comme première fonction d'activation tangente tanh et une deuxième fonction d'activation linéaires. Pour le reste des classes j'ai choisi de faire un modèle de PMC avec une architecture de 23 neurones cachés par le biais d'une fonction d'activation sigmoïde et une fonction d'activation softmax.

Validation :

Pour la validation j'ai utilisé la méthode de Leave one out car le volume des données au sein de chaque classe est très faible et la cross validation ne sera pas efficace, mais cette méthode reste très couteuse en terme de calcul car il en faut établir un modèle pour chaque individu.

Résultats :

PMC	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Architecture	(Sig,Lin,17)	(Tan,Lin,17)	(Tan,Lin,13)	(Tahn,Lin,17)	(Tahn,Lin,17)	(Tahn,Lin,17)	(Tahn,Lin,17)	(Tahn,Lin,17)
Erreur gen	33%	35%	37%	35%	40%	36%	40%	45%
Performance	49%	64%	59%	72%	51%	69%	70%	50%

Tableau 3 : Représentation de l'Architecture, l'erreur de Généralisation et la performance pour chaque modèle

Le PMC à retourner des résultats plutôt satisfaisants, la performance atteint même 70% pour quelques classes ce qui n'était pas possible en prenant en compte un seul échantillon.

En effet cette classification non supervisée a été bénéfique d'un point de vue performance et minimisation de l'erreur de généralisation puisqu'elle a permit d'étudier chaque groupement de donnée comme étant un échantillon à part entière et d'attribuer à chaque classe le modèle qui lui correspond le plus. Vu la qualité des données, cette solution est la plus optimale car elle met de l'ordre à la distribution des données en les classant selon leurs critères communs et donc de tirer le maximum d'avantage de leurs distributions.

	PMC
Erreur gen	0.35
Performance	0.61

Tableau 4 : Représentation de l'erreur de Généralisation et la performance moyenne des modèles

En comparant l'erreur moyenne de généralisation des différentes classes avec l'erreur obtenue précédemment en utilisant une seule classe on remarque qu'elle a diminué de 0.45 à 0.37 ce qui est plutôt bon, on voit aussi qu'il y a une amélioration remarquable surtout que la performance a évolué de 0.4 à 0.6. On valide donc ce modèle et on passe à la prédiction des valeurs manquantes.

Prédiction des valeurs manquante :

Dans un premier temps, j'ai commencé par chercher les valeurs manquantes, j'ai trouvé qu'il existe 17 valeurs qui manquent que j'ai supprimés du jeu de données et qui n'étaient pas prises en compte pour la réalisation des différents modèles ainsi que pour la classification. Dans un second temps, il me faut connaitre à quelle classe correspond chaque valeur manquante pour pouvoir utiliser le modèle adéquat pour la prédire, pour ce fait j'ai construit un modèle en utilisant K Nears Neighbours en considérant les classes provenant du clustering kmeans comme variables à expliquer et le reste des variables comme variables explicatives (sauf pour le formaldéhyde).

L'avantage d'utiliser K Nears Neighbours est qu'elle est basée sur la distance euclidienne tout comme le clustering kmeans ce qui devrait donner plus de précision pour la prédiction des clusters pour les individus ayant des valeurs manquantes dans la variable formaldéhyde.

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Class8**
Nombre	2	3	2	3	2	**4**	0	1

Tableau 5 : Le nombre des individus contenant des valeurs manquantes dans chaque classe

Après avoir affecter chaque valeur manquante à sa classe, j'ai obtenue deux valeurs pour la première classe, trois pour la seconde, deux valeurs manquantes pour la troisième classe, trois valeurs pour la quatrième, deux pour la cinquième classe, quatre pour la sixième classe et une valeur qui manque pour la huitième classe.

J'ai donc prédit chaque variable en fonction du modèle correspond à sa classe et obtenue les résultats suivants : [7.49, 7.49, -3, -1.21, -1.29, 7.26, 7.26, 93.96, 93.96, 93.96, 61.57, 61.57, 71.93, 71.93, 71.93, 71.93, 94.94]

Conclusion :

La nature du jeu de donnée et leur mauvaise distribution à compliqué la tâche de sélection de variable et de modélisation, en effet il n'existait aucune forme de liaison linéaire, exponentielle ou autre entre les données ce qui m'a amené à chercher une sous distributions entre ces données afin de l'améliorer en utilisant le clustering kmeans. La répartition des données en classes n'a pas donné le résultat attendu et n'a pas réussi à mettre en évidence une quelle conque liaison linéaire entre les données, néanmoins cette étape était tranchante pour l'étape de modélisation car m'a permis d'utiliser le perceptron multicouche avec différentes architectures pour chaque classe.

Le résultat était plutôt bon comparé aux résultats obtenues précédemment et a permis d'améliorer la performance et de minimiser l'erreur de généralisation mais cette méthode a ses limites puisqu'elle utilise un temps de calcul énorme qui peut durer même 10 minute par classe.

Pour aboutir à des meilleurs résultats, il aurait peut-être fallu utiliser la carte topologique pour effectuer le clustering.