BEDOUCHA Gary
CONDORIS Dimitri
COHEN Ethan
MAGASSA Idrissa

Apprentissage Statistique

Sujet 4 :
Détection d'anomalies pour des débitmètres ultrasoniques

Résumé

Le sujet proposé nous invite à étudier un problème de classification et plus particulièrement la détection d'anomalies pour des débitmètres ultrasoniques. Le débitmètre à ultra-sons est un instrument utilisant les ultrasons pour mesurer la vitesse moyenne d'un fluide. Ainsi, l'objectif de ce projet est de pouvoir déterminer si ces des derniers sont défectueux ou non. Pour ce faire, nous disposons de diverses variables d'entrées ainsi que de plusieurs algorithmes d'apprentissage statistique, les algorithmes que nous utiliserons ici seront ceux des **K-Plus Proches Voisins**, de la **Régression Logistique** et finalement du **Machine à Vecteurs de Support**.

Sommaire

Introduction

Les données sur lesquelles se pose l'étude se constituent en quatre groupes, dans chacun de ces groupes nous disposons de plusieurs caractéristiques du débitmètre, cependant la variable à prédire varie en fonction du groupe, en particulier c'est le nombre de modalités qui va varier. Ici nous allons nous intéresser au groupe qui contient pour la variable à prédire seulement deux classes : en bon état ou défectueux.

Pour ce faire nous disposons d'un jeu de données constitué de plusieurs variables telles que la vitesse moyenne du son dans les huit conduits, le rapport de planéité, le flux croisé, la symétrie et d'autres encore.

Objectifs

Pour résoudre ce problème, qui est un problème de classification supervisée bimodale, nous allons utiliser plusieurs algorithmes classiques, les comparer entre eux et les critiquer. Notre objectif ici va donc être de maximiser un certain score que l'on précisera ensuite afin de créer une règle de classification efficace qui pourra servir à détecter des débitmètres défectueux afin de limiter au maximum les erreurs de mesures possibles lors d'études dans l'industrie pharmaceutique par exemple.

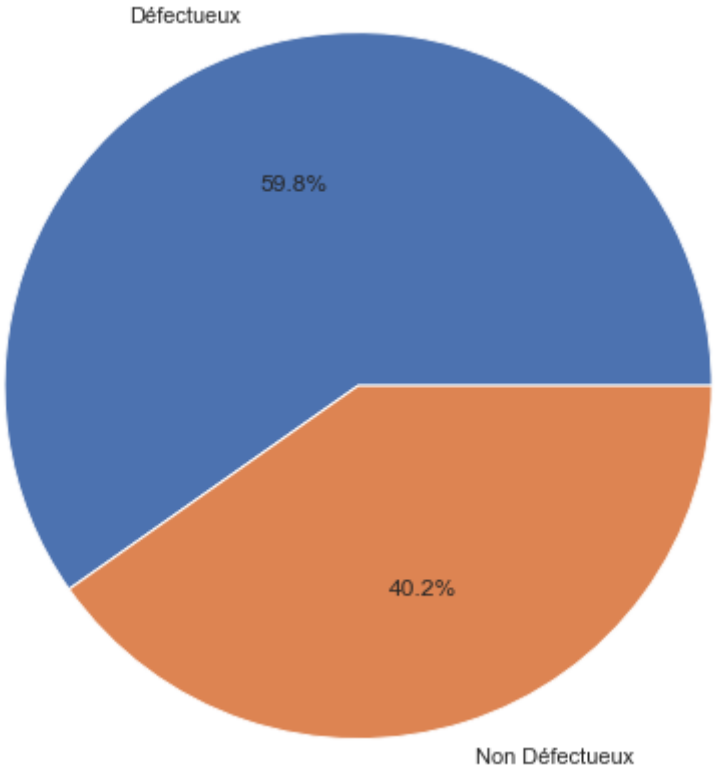
1 Explorations des données

1.1 Préparation de la base de données

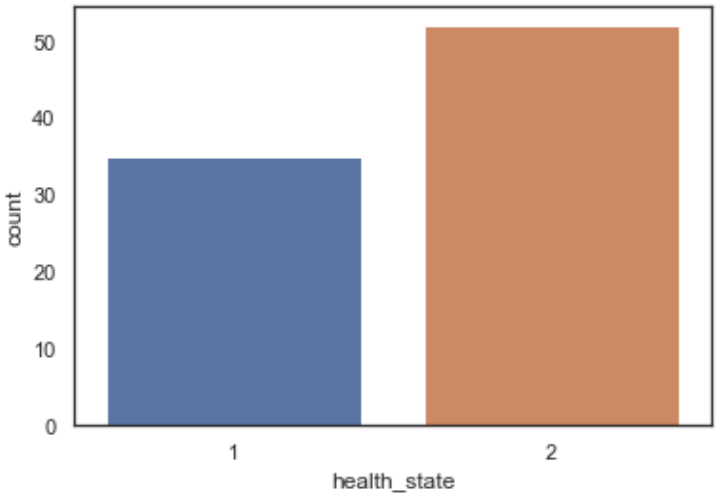
Nous disposons dans notre étude de 87 débitmètres et 36 de leurs caractéristiques constituées de 35 variables continues et de la variable d'intérêt, bimodale, à valeurs dans {1,2}, 1 représentant l'état non défectueux et 2 représentant l'état défectueux. Nous ne disposons d'aucune donnée manquante et, les variables explicatives étant toutes continues, d'aucune variable à recoder. Cependant on constate rapidement que l'on dispose de nombreuses covariables (quasiment autant que la moitié du nombre d'observations). Cela peut donc représenter un problème que l'on traitera plus tard.

1.2 Étude des variables

On regarde tout d'abord la variable d'intérêt, c'est-à-dire celle qui indique si le débitmètre est défectueux ou non. Nous allons regarder la répartition de celle-ci en fonction de ses deux modalités :



Répartition des modalités de la variable d'intérêt



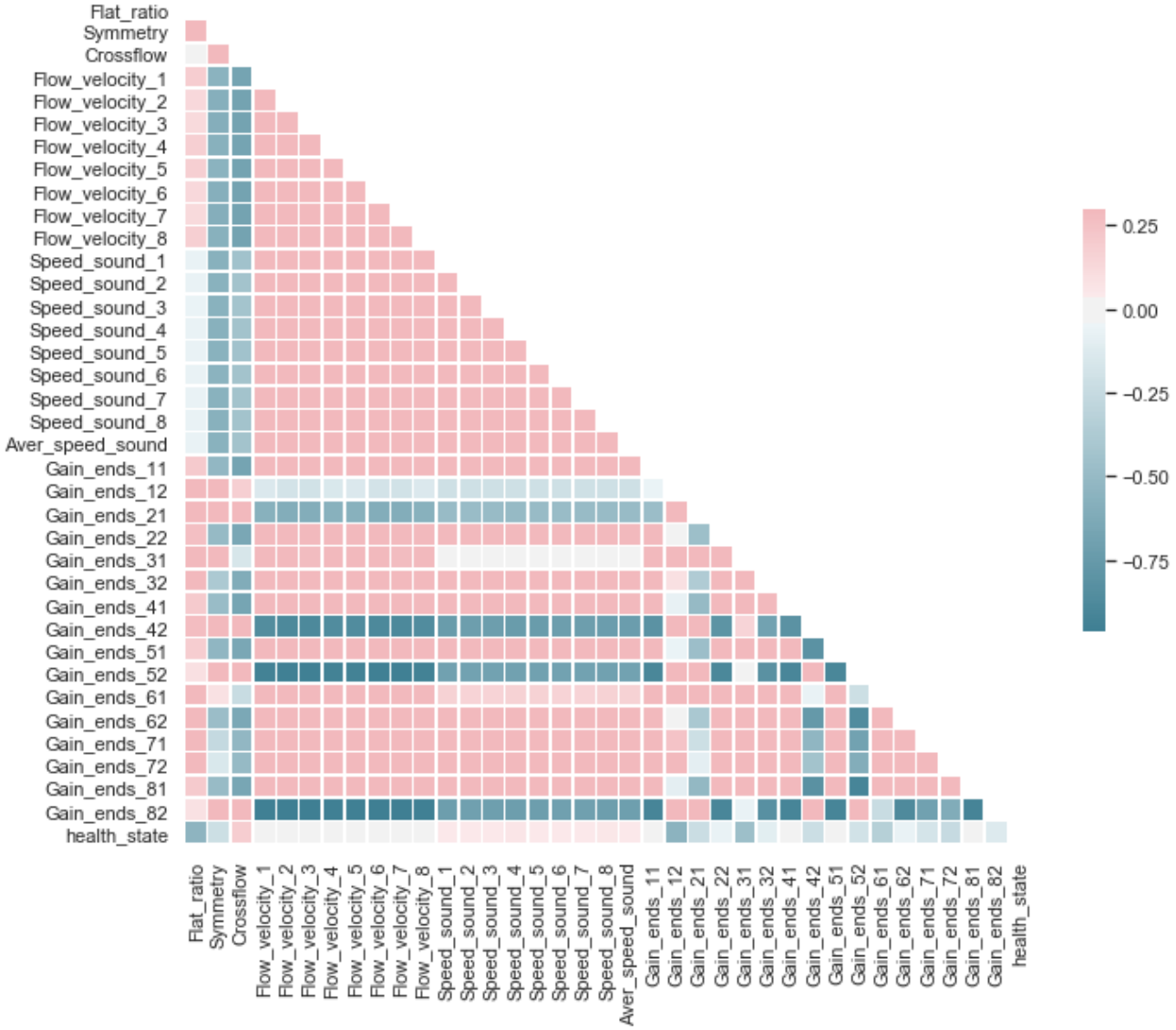
Nombre d'individus dans chaque classe

Nous pouvons constater que la répartition est plutôt homogène entre les deux classes, en effet, 40,2% des appareils sont en bon état contre 59.8% d'appareils défectueux, de plus comme nous ne disposons pas d'impératif quant à l'importance des faux positifs et faux négatifs, on pourra utiliser la précision comme indicatrice de la qualité d'un modèle. La précision est définie comme le nombre de prédictions correctes divisé par le nombre total de prédictions.

Pour les covariables nous allons nous intéresser à la matrice des corrélations définie par :

$$\forall\, i,j \leq p,\; M(i,j) = corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)}\sqrt{var(X_j)}}$$

Avec p le nombre de covariables.



Matrice des corrélations

On constate que plus ou moins toutes les covariables sont corrélées entre elles, cependant on remarque que certaines sont très fortement négativement corrélées entre elles (Gain_ends_82 ou encore Symmetry). Nous allons donc chercher à réduire la dimension pour voir si l'on ne peut pas représenter les individus dans un espace de dimension 3 voir 2.

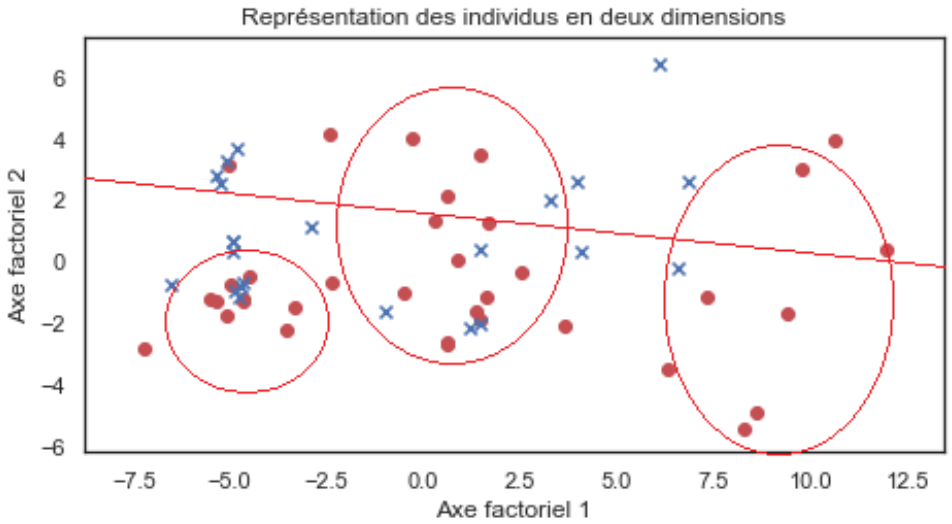
1.3 Réduction de dimension

Plusieurs variables semblent fortement corrélées entre elles, il peut donc y avoir redondance d'information dans notre modèle, pour vérifier cela nous allons lancer une ACP, ce qui consiste à trouver une base orthogonale dans l'espace des covariables. Pour pouvoir plus tard vérifier que nos résultats se confirment, c'est-à-dire que l'on peut réduire notre modèle à 3 variables, nous lancerons l'ACP sur une base d'apprentissage au préalable séparée d'une base de test. Il y a 29 observations dans la base de test et 58 dans la base d'apprentissage.

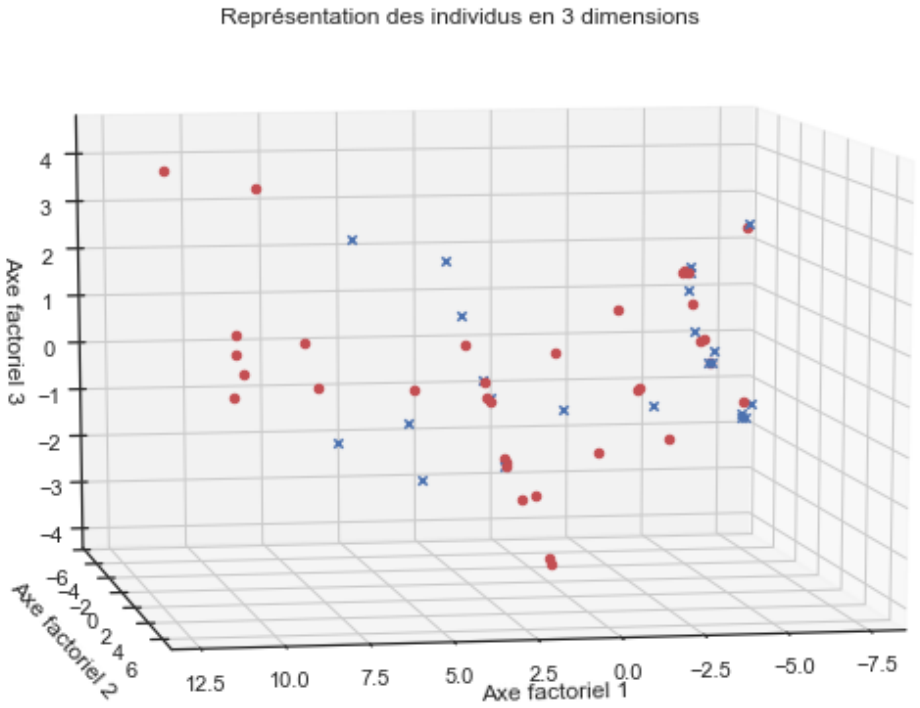
	Composante principale 1	Composante principale 2	Composante principale 2
Variance expliquée	71%	16%	8%

Tableau de la contribution des composantes principales à la variance

95% de la variation des covariables est expliquée par 3 composantes principales, on décide donc de représenter les individus dans cette base en 2 et 3 dimensions.



Représentation des individus dans la base des composantes principales, en rouge ceux défectueux



Représentation des individus dans la base des composantes principales, en rouge ceux défectueux

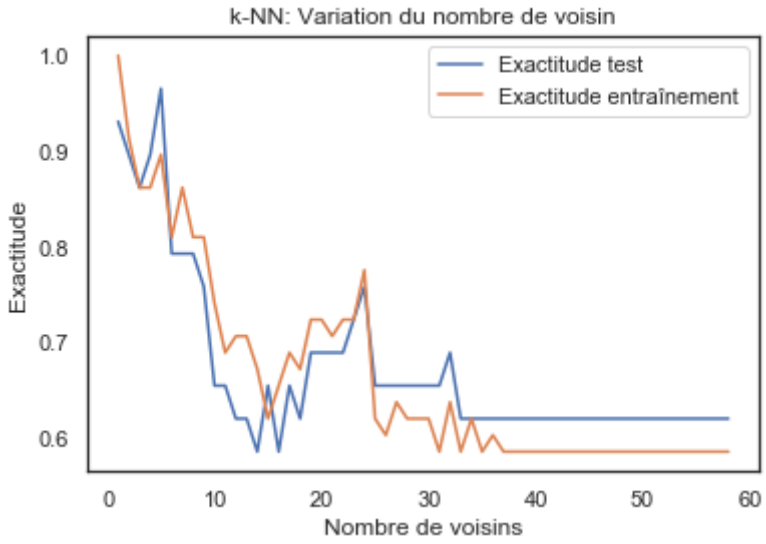
Nous pouvons constater que l'on pourrait tenter de séparer linéairement les deux groupes, cependant cette séparation n'étant pas triviale, on s'attend à obtenir un faible score si l'on applique un modèle de Machine à Supports de Vecteur avec un noyau linéaire, cependant on distingue bien 3 groupes de rouges sur la représentation en 2 dimensions avec quelques outliers bleus, il est donc raisonnable de penser qu'un noyau gaussien produirait de bien meilleurs résultats. Enfin les classes semblant regroupées entre elles, des méthodes de type Régression Logistique ou K-Plus Proches Voisins pourraient aussi s'avérer efficaces. (Même si pour la Régression Logistique les points ne sont pas disposés selon une sigmoïde, forme que l'on retrouve pour ce genre de modèle.)

2 Méthodes de classification

Comme nous l'avons vu sur la matrice de corrélations, certaines corrélations sont élevées, de plus en appliquant la méthode .describe() à notre jeu d'entraînement, on constate que les variances sont très différentes, ainsi pour éviter que certaines variables se fassent camoufler par d'autres plus grosses, on normalise nos données.

2.1 Méthode des K-Plus Proches Voisins

Il est nécessaire de bien comprendre que la réduction de dimension n'est pas vouée à sélectionner des variables, mais bien de représenter nos individus dans un espace de dimension inférieure. En effet le peu d'individus nous permettent de lancer n'importe lequel de nos modèles en un temps très limité (de l'ordre de la centième de seconde). Voilà pourquoi l'intégralité des variables sera utilisée lors de l'application de nos modèles.
Naïvement on commence par tester différentes valeurs du nombre de voisins afin de maximiser la précision de notre modèle sur la base test :



Précision en fonction du nombre de voisins

Si l'on choisit cette méthode, le nombre de voisins optimal serait 5, or en faisant cela, on sur-apprend possiblement la base test. Pour éviter de surapprendre la base test, on procède à une validation croisée et une recherche sur grille avec 10 échantillons. Avec cette méthode nous obtenons un nombre optimal de voisins de 3 avec une variance de 0.24 sur l'ensemble des

échantillons et un score de 90% sur la base de test. Nous obtenons aussi la matrice de confusion suivante :

Réel/Estimation	En bon état	Défectueux
En bon état	10	1
Défectueux	2	16

Matrice de confusion pour la méthode des K-PPV

On constate que l'on a autant de faux positif que de faux négatif, ce qui confirme le fait qu'il n'y a pas beaucoup d'inégalité de dispersion entre les classes. Cependant la variance de 24% indique que cette méthode est potentiellement en train de surapprendre notre base de données (ce qui serait induit par le nombre très faible d'individus). On obtiendrait donc peut-être un score beaucoup moins bon sur une base beaucoup plus grande. Nous allons donc essayer le modèle de la Régression Logistique afin de diminuer cette variance.

2.2 Régression logistique

La Régression Logistique est un modèle linéaire généralisé avec la variable d'intérêt distribué selon une loi binomiale (en l'occurrence il s'agit d'une loi de Bernoulli pour être plus précis). Elle est donc de moindre performance et donc diminue la possibilité de surapprentissage. On effectue donc de la même manière que pour la méthode des K –Plus Proches Voisins une validation croisée avec une grille de recherche comprenant, les pénalités Lasso/Ridge ainsi que la variation du coefficient devant ces pénalités. Nous trouvons avec cette stratégie une pénalité optimale L2 avec une constante qui vaut 0.7. Cette fois-ci la variance vaut 6% et la précision sur la base de test est de 80%, finalement nous obtenons la matrice de confusion suivante :

Réel/Estimation	En bon état	Défectueux
En bon état	10	1
Défectueux	5	13

Matrice de confusion pour la Régression Logistique

Pour cette méthode les conclusions sont contraires à celles d'avant, en effet la variance est faible, on a donc plus peur de sous-apprendre la base de données, de plus la précision a chuté de 10%, on peut donc encore améliorer notre modèle. Pour ce faire nous allons utiliser la méthode qui semblait le plus adapté à la vue de la représentation des individus en 2 dimensions, il s'agit de la méthode Machine à Vecteurs de Support.

2.3 Méthode Machine à Vecteurs de Support

Comme nous l'avons vu précédemment, les individus de notre base d'entraînement semblent tantôt séparables selon un noyau gaussien et légèrement linéairement aussi. Nous allons donc comme dans les précédents modèles, utiliser la faible taille de notre tableau pour pouvoir lancer une validation croisée avec une grille de recherche. Nous allons donc ici chercher plusieurs valeurs de la constante de pénalisation, du noyau (linéaire ou gaussien) ainsi que de la variance du noyau

gaussien. Nous obtenons comme paramètres optimaux pour cette méthode, un noyau gaussien, une constante de pénalisation valant 0.66 et finalement un gamma (égal à l'inverse de deux fois la variance) valant 1. Nous obtenons de plus une précision de 86% sur la base de test ainsi qu'une variance de 14% sur l'ensemble des échantillons de la validation croisée. Finalement nous obtenons la matrice de confusion suivante :

Réel/Estimation	En bon état	Défectueux
En bon état	8	3
Défectueux	1	17

Matrice de confusion pour la méthode SVM

Conclusion

Nous avons réussi à trouver un modèle qui paraît plutôt convenable, celui de la méthode SVM, en effet on a réussi à plus ou moins minimiser la variance et maximiser la précision en même temps, cependant on trouve une précision de 86% sur la base de test, vu le peu de donnée, il est logique de penser que peu importe notre minimisation de variance, les algorithmes que nous avons proposés ont tous surappris notre échantillon particulier et ne sont donc a priori pas généralisables. Il faudrait donc pour améliorer ceux-ci, beaucoup plus d'échantillons pour être certain de la fiabilité de leurs estimations.