

|  |
| --- |
| Machine Learning |
| Rendu TD 3 : Feature Select |
|  |

|  |
| --- |
| ANIS MOURI Christophe MOYEN    15/12/2016 |

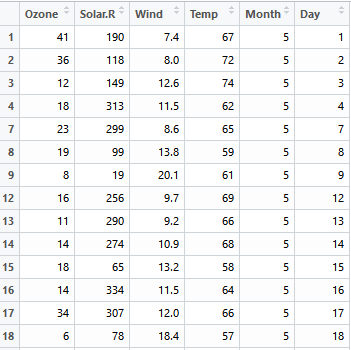
# Présentation du TP

## Objectif

Dans ce TP seront tester plusieurs méthodes de sélections de variables et étudier leurs importances pour le modèle. On étudiera la fiabilité de chacune des méthodes et leurs dépendances aux échantillons choisis.

## Analyse apriori du jeu de données :

Exemple des données étudiées :



Ce jeu de donnée contient 11 observations et 6 variables

[,1] Ozone numeric Ozone (ppb)

[,2] Solar.R numeric Solar R (lang)

[,3] Wind numeric Wind (mph)

[,4] Temp numeric Temperature (degrees Farenheit)

[,5] Month numeric Month (1-12)

[,6] Day numeric Day of month (1-31)

Notre étude portera sur les variables significatives dans la détermination du taux de pollution d’Ozone.

## Exécution du script et commentaires :

Avant l’application des tests on transforme les données quantitatives du taux de concentration de l’ozone vers une donnée binaire en prenant comme seuil de pollution 180 ppb. Les nouvelles données seront donc « Y » pour pollué et « N » pour non pollué.

### Test de fischer :

On a décidé d’appliquer le test de fischer sur toutes les variables pour voir les différences des résultats.

# FDS score between Ozone and each other variable

FishT = FDS(DS$Temp,DS$Ozone)

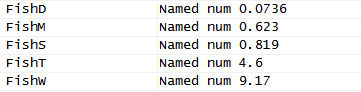
FishS = FDS(DS$Solar.R,DS$Ozone)

FishW =FDS(DS$Wind,DS$Ozone)

FishM =FDS(DS$Month,DS$Ozone)

FishD =FDS(DS$Day,DS$Ozone)

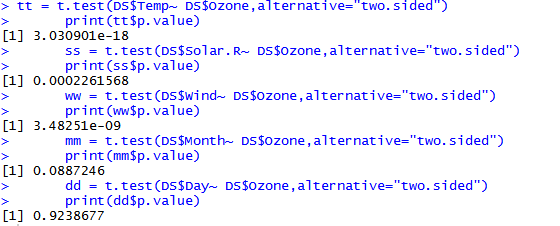
On a observé les résultats ci-dessous :



On voit bien que les deux variables les plus significatives sont la température et le vent (elles ont les indices les plus élevés).

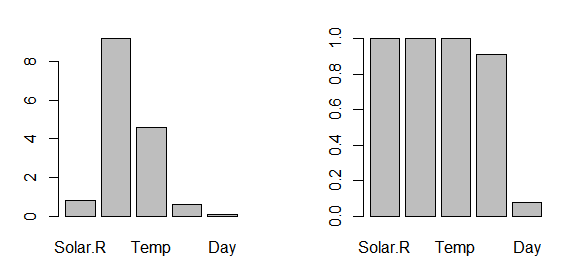
### Test de Student :

En analogie avec le test de Fischer, on a dupliqué ce test sur toutes les variables pour voir les différences entre ces dernières. On voit bien que les deux variables les plus significatives encore une fois sont la température et le vent.

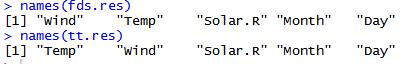


La p-value représente le seuil d’acceptabilité d’une variable selon une fonction gaussienne.

## Comparaison entre les deux tests : (fds.res et tt.res)



Tri des variables par ordre décroissant sur la signification des variables (de la plus importante à la moins importante)



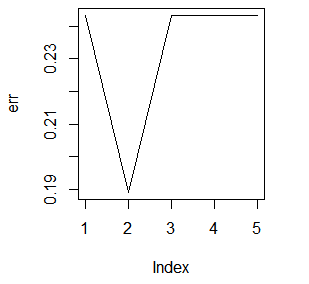
La p-value est mois de 5% pour les variables Temp, Wind et Solar.R.

Avec l’utilisation de CART :

En choisissant le tri var.fds :



On aura le résultat suivant :



> combien

[1] 2

> optvars = lesvars[1:combien]

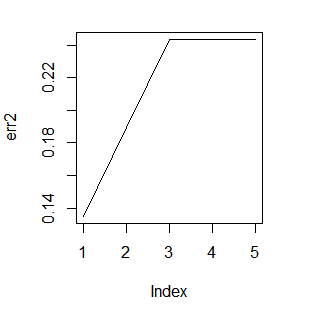
> optvars

[1] "Wind" "Temp"

CART, nous montre que c’est lorsqu’il y a 2 variables (combien) que le résultat est le plus probant. Ces deux variables sont : « Wind » et « Temp ». C’est ce que l’on observe sur la courbe, c’est en ajoutant les deux premières variables que l’erreur est la plus faible.

Test avec un tri différent des variables (var.tt) : 

On aura les résultats suivants :



> combien2

[1] 1

> optvars2

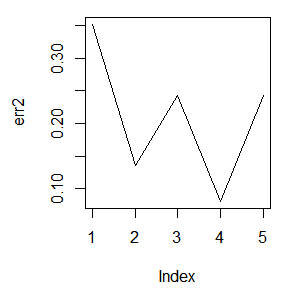
[1] "Temp"

Donc ici, la seule variable « Temp » est significative.

Pour voir l’importance de l’ordre des variables en entrée, on introduit un arrangement aléatoire :

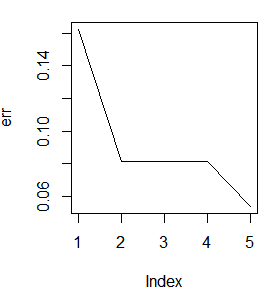


On obtient les résultats suivants :



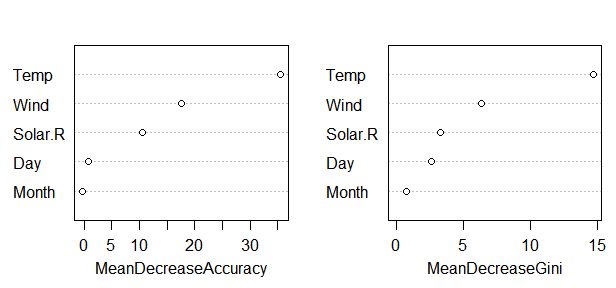
Le nombre de variables significatives prisent en compte ici est 4. En prenant comme hypothèse le minimum des erreurs dues au CART.

On remarque aussi qu’en renouvelant l’opération plusieurs fois, et donc des échantillons différents, on voit que le résultat fluctue beaucoup :



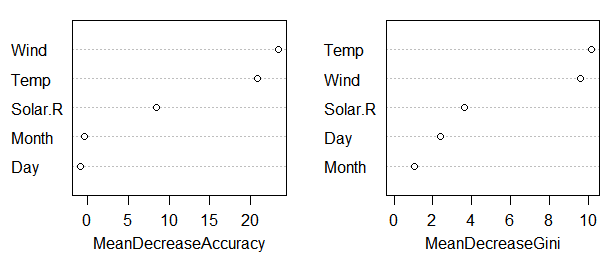
Dans cet exemple la sélection s’est faite sur 5 variables significatives.

## Importance des variables avec Random Forest



Le "Mean Decrease Accuracy" évalue l’importance des variables dans un contexte prédictif quant au "Mean Decrease Gini" correspond à l’apport d’une variable sur l’homogénéité d’un noeud ne sera pas présenté. Il s’inscrit plutôt dans un contexte explicatif.

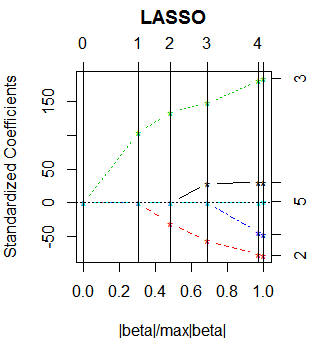
On observe donc que la variable la plus significative est « Temp ».



Mais selon l’échantillon de test, la variable « Wind » peut être la plus significative.

Donc la valeur de la variable la plus significative est dépendante de l’échantillon en « Random Forest ».

# LASSO



Le premier indice est la variable 3 « temp », le second est la variable 2 « wind ».

Plus la variable température augmente, plus la pollution est importante.

Plus la variable wind est importante, mois la pollution est importante.