# Choix de Modèles & Modèles Linéaires Généralisés

Haeji Yun

2023-06-29

Dans cette analyse, nous nous intéressons à prédire s'il va pleuvoir ou pas le lendemain à Bâle en nous basant sur différentes variables météorologiques qui composent notre jeu de données.

Notre jeu de données contient 1180 observations et 46 variables. Les observations correspondent aux différents jours entre 2010 et 2018 et les variables correspondent à différentes caractéristiques météorologiques.

## Chargement de données

Notre variable d'intérête est la variable qualitative *pluie.demain* qui est un boléen indiquant s'il a plu le lendemain ou pas. La valeur est *True* s'il a plu le lendemain et *False* dans le cas contraire.

Les autres 45 variables sont quantitatives qui vont nous servir à expliquer la variable cible. Elles sont composées de :

- Les variables de temps telles que l'année, le mois, le jour, l'heure et la minute
- La moyenne, la minimale et la maximale de différentes caractéristiques météorologiques telles que la température, l'humidité, la pression, la nébulosité en pourcentage, la nébulosité forte, la nébulosité moyenne, la nebulosité faible, la vitésse du vent à 10 m, la vitesse du vent à 80 m, la vitesse du vent à 900 m, la rafale
- La moyenne de la direction de vent à 10 m, la direction de vent à 80 m, et la direction de vent à 900 m
- Les valeurs totales de précipitations, de néige, d'ensoleilement, de rayonnement solaire

Dans nos variables, nous avons à la fois les mêmes variables mesurées sur de niveaux différentes telles que la nébulosité d'intensité différente et le vent à altitudes différentes, et les mêmes variables représentées par des mesures différentes telles que la moyenne, la minimale et la maximale. Nous pouvons déjà supposer qu'il y aura beaucoup de variables qui sont corrélées entre elles.

Le fichier csv qui contient notre jeu de données est téléchargé sous forme d'un dataframe. Nous allons nous contenter d'afficher la dimension, la liste et le type de variables de notre jeu de données car il est difficile d'avoir toutes les colonnes en dataframe sur une page à cause de nombre élevé de variables.

```
## 'data.frame':    1180 obs. of  46 variables:
##  $ Year                                       : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 20
##  $ Month                                      : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Day                                        : int  2 4 6 8 10 12 14 16 18 20 ...
##  $ Hour                                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minute                                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Temperature.daily.mean..2.m.above.gnd.     : num  15 17.3 21.6 20.2 22.6 ...
##  $ Relative.Humidity.daily.mean..2.m.above.gnd.: num  76.5 77.6 69.5 75.1 73.5 ...
```

```
##  $ Mean.Sea.Level.Pressure.daily.mean..MSL.    : num  1015 1017 1015 1008 1004 ...
##  $ Total.Precipitation.daily.sum..sfc.         : num  1 0 3.7 0.2 0 2.2 1.8 1.8 17.5 1.2 ...
##  $ Snowfall.amount.raw.daily.sum..sfc.         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Total.Cloud.Cover.daily.mean..sfc.          : num  79.8 4.7 42.1 67.5 56.3 ...
##  $ High.Cloud.Cover.daily.mean..high.cld.lay.  : num  3 0.67 21.21 54.71 50.25 ...
##  $ Medium.Cloud.Cover.daily.mean..mid.cld.lay. : num  31.6 0 25.9 65.8 55.3 ...
##  $ Low.Cloud.Cover.daily.mean..low.cld.lay.    : num  79.2 4.5 35.3 18.9 34.2 ...
##  $ Sunshine.Duration.daily.sum..sfc.           : num  287.2 821.4 441.3 41.9 473.2 ...
##  $ Shortwave.Radiation.daily.sum..sfc.         : num  6710 7974 4834 5390 7216 ...
##  $ Wind.Speed.daily.mean..10.m.above.gnd.      : num  11.64 6.34 8.4 5.4 9.16 ...
##  $ Wind.Direction.daily.mean..10.m.above.gnd.  : num  275 230 215 205 179 ...
##  $ Wind.Speed.daily.mean..80.m.above.gnd.      : num  14.99 8.92 10.38 6.53 11.91 ...
##  $ Wind.Direction.daily.mean..80.m.above.gnd.  : num  268 199 208 206 186 ...
##  $ Wind.Speed.daily.mean..900.mb.              : num  20.6 27.9 18.9 10.4 21.9 ...
##  $ Wind.Direction.daily.mean..900.mb.          : num  180.4 93.7 250.1 238.6 153 ...
##  $ Wind.Gust.daily.mean..sfc.                  : num  14.88 9.48 13.5 5.31 12.21 ...
##  $ Temperature.daily.max..2.m.above.gnd.       : num  18.5 25 26.2 24.2 30.7 ...
##  $ Temperature.daily.min..2.m.above.gnd.       : num  11.1 10.4 17.7 14.7 16.9 ...
##  $ Relative.Humidity.daily.max..2.m.above.gnd. : int  94 92 91 89 97 92 96 96 97 95 ...
##  $ Relative.Humidity.daily.min..2.m.above.gnd. : int  59 54 57 62 39 65 69 64 74 61 ...
##  $ Mean.Sea.Level.Pressure.daily.max..MSL.     : num  1017 1019 1016 1010 1006 ...
##  $ Mean.Sea.Level.Pressure.daily.min..MSL.     : num  1014 1016 1013 1006 1001 ...
##  $ Total.Cloud.Cover.daily.max..sfc.           : num  100 28 100 100 100 100 100 100 100 100 ...
##  $ Total.Cloud.Cover.daily.min..sfc.           : num  0 0 0 0 0 0 0 100 0 0 ...
##  $ High.Cloud.Cover.daily.max..high.cld.lay.   : int  16 11 100 100 100 28 100 100 100 24 ...
##  $ High.Cloud.Cover.daily.min..high.cld.lay.   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Medium.Cloud.Cover.daily.max..mid.cld.lay.  : int  100 0 100 100 100 100 100 100 100 41 ...
##  $ Medium.Cloud.Cover.daily.min..mid.cld.lay.  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Low.Cloud.Cover.daily.max..low.cld.lay.     : int  100 28 100 100 100 100 100 100 100 100 ...
##  $ Low.Cloud.Cover.daily.min..low.cld.lay.     : int  0 0 0 0 0 0 0 29 0 0 ...
##  $ Wind.Speed.daily.max..10.m.above.gnd.       : num  22 15.5 22.7 10.7 20.5 ...
##  $ Wind.Speed.daily.min..10.m.above.gnd.       : num  5.62 1.08 2.41 0 2.52 2.28 1.3 4.32 7.2 8.05 .
##  $ Wind.Speed.daily.max..80.m.above.gnd.       : num  23.8 18.7 32 10.2 23.4 ...
##  $ Wind.Speed.daily.min..80.m.above.gnd.       : num  8.65 0 0.51 1.44 2.97 ...
##  $ Wind.Speed.daily.max..900.mb.               : num  32.1 48.1 44 22.2 40.8 ...
##  $ Wind.Speed.daily.min..900.mb.               : num  12.25 6.62 5.48 4.69 4.68 ...
##  $ Wind.Gust.daily.max..sfc.                   : num  25.2 20.2 41.8 11.2 24.1 ...
##  $ Wind.Gust.daily.min..sfc.                   : num  6.48 2.16 1.08 0.36 1.44 ...
##  $ pluie.demain                                : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
```

# Etude Exploratoire

## Valeurs manquantes et valeurs uniques

Nous n'avons pas de données manquantes dans notre jeu de données.

```
sum(is.na(meteo_train))
```

```
## [1] 0
```

Néanmoins, nous observons que la variables *Hour* et *Minute* ont une valeur unique *0* pour toutes les observations. Ces deux variables ne sont donc pas informatives et ne donnent aucune explication sur notre variable cible. Nous pouvons déjà supprimer les deux.

```
##       Hour       Minute
##  Min.   :0   Min.   :0
##  1st Qu.:0   1st Qu.:0
##  Median :0   Median :0
##  Mean   :0   Mean   :0
##  3rd Qu.:0   3rd Qu.:0
##  Max.   :0   Max.   :0
```

Nous retrouvons avec un jeu de données avec 1180 observations et 44 variables au lieu de 46.

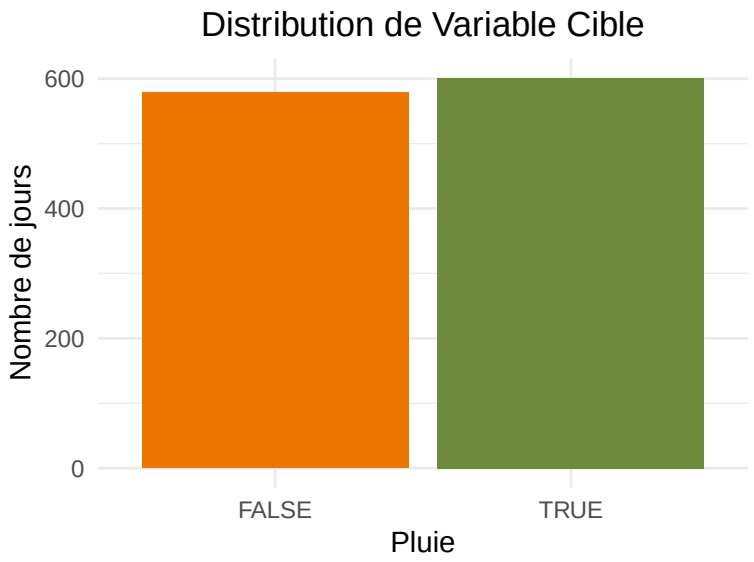
```
## [1] 1180   44
```

## Variable cible

Notre variable d'intérêt *pluie.demain* a 49% de valeur *True* et 51% de valeur *False*.

Nous avons une quantité équivalente des deux classes. Cela nous évitera de réduire la performance de prédiction en apprenant mieux une classe avec plus de données disponibles par rapport à l'autre classe.

Nous pourrons également diviser les données d'entraînement et de test de façon aléatoire par la suite.

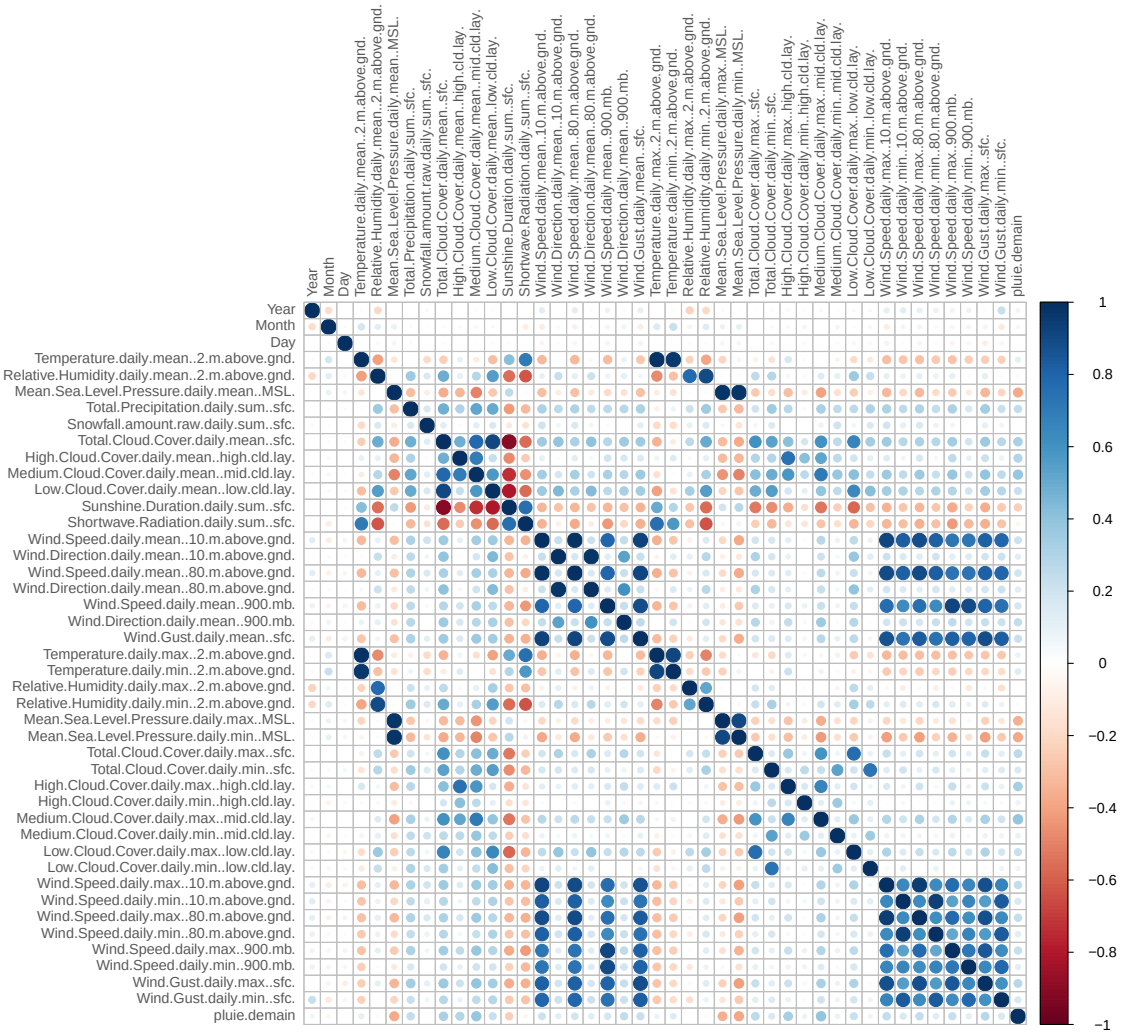



## Variables explicatives

Nous pouvons étudier la corrélation de variables avec une matrice de corrélation.

Comme attendue, nous observons des corrélations sur des blocs de variables :

- Entre les différents niveaux de même variable : la nébulosité d'intensité différente et le vent à altitudes différents
- Entre les différentes mesures de même variable : la moyenne, la minimale et la maximale de la température, de l'humidité et de la pression
- Entre les différents variables : la vitesse de vent, la direction de vent et la rafale.




## Sélection de variables

Comme notre étude consiste à prédire s'il va pleuvoir le lendemain ou pas, c'est une variable discrète binaire *vrai* ou *faux* que nous voudrons obtenir. Il s'agit d'une classification.

Nous allons donc utiliser la régression logistique qui estime pour chaque observation la probabilité qu'un événement se produise, que nous pouvons utiliser pour classifier selon le seuil de probabilité que nous fixons.

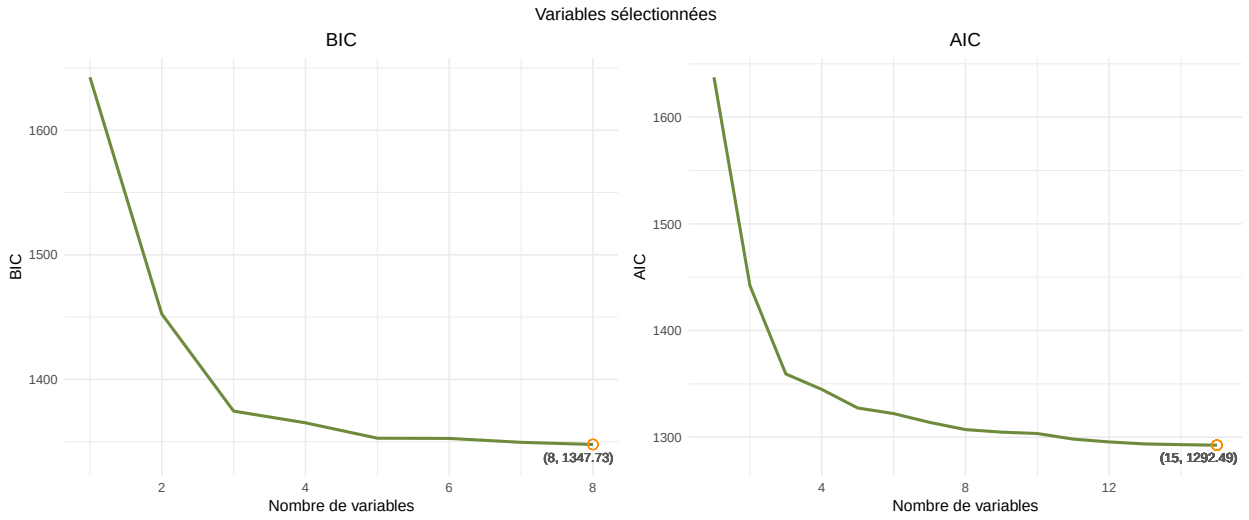
Parmi les critères de sélection de modèles, nous ne pourrons pas utiliser R2 ajusté et Cp de Mallows car ils utilisent le résidu de modèle linéaire. Nous allons utiliser les critères *aic* et *bic* qui utilisent la vraisemblance.

Les steps *both*, *forward*, et *backward* de *aic* et de *bic* ont donné à peu près les mêmes résultats. Pour cela, nous avons opté de garder le step *forward* pour faciliter la compréhension de sélection effectuée.

Avec le step *forward*, l'initialisation est faite avec le modèle qui ne continet pas de variablie explicative. Nous calculons *aic* et *bic* du modèle. Puis la variable qui minimise *aic* ou *bic* est ajouté. Cet ajout de variable continue une par une jusqu'à ce que le modèle obtenu n'a plus de *aic* ou *bic* inférieur au modèle précédant.

Le meilleur modèle en terme de *bic* a gardé 8 variables et celui de *aic* a gardé 15 variables.

Les modèles minimisent chacun son *bic* et son *aic*. Le *bic* est égale à 1347,73 avec 8 variables et le *aic* est égale à 1292,49 avec 15 variables.




Les 15 variables sélectionnées par *aic* sont affichées ci-dessous. Nous observons les variables ayant des corrélations detectées précedemment entre:

- les nébulosités d'intensité différente
- la vitesse de vent, la direction de vent et la rafale
- la température minimale et la température maximale

```
##  [1] "(Intercept)"
##  [2] "Medium.Cloud.Cover.daily.max..mid.cld.lay."
##  [3] "Mean.Sea.Level.Pressure.daily.min..MSL."
##  [4] "Temperature.daily.min..2.m.above.gnd."
##  [5] "Wind.Gust.daily.max..sfc."
##  [6] "Total.Cloud.Cover.daily.mean..sfc."
##  [7] "Temperature.daily.max..2.m.above.gnd."
##  [8] "Wind.Direction.daily.mean..900.mb."
##  [9] "Year"
## [10] "Wind.Speed.daily.mean..80.m.above.gnd."
## [11] "Wind.Speed.daily.min..10.m.above.gnd."
## [12] "Wind.Speed.daily.max..10.m.above.gnd."
## [13] "Wind.Direction.daily.mean..80.m.above.gnd."
## [14] "Snowfall.amount.raw.daily.sum..sfc."
## [15] "Total.Cloud.Cover.daily.min..sfc."
```

Le *bic* a retenu beaucoup moins de variables par rarpport à *aic*. Nous observons quand même les variables ayant des corrélations détectées précedemment entre :

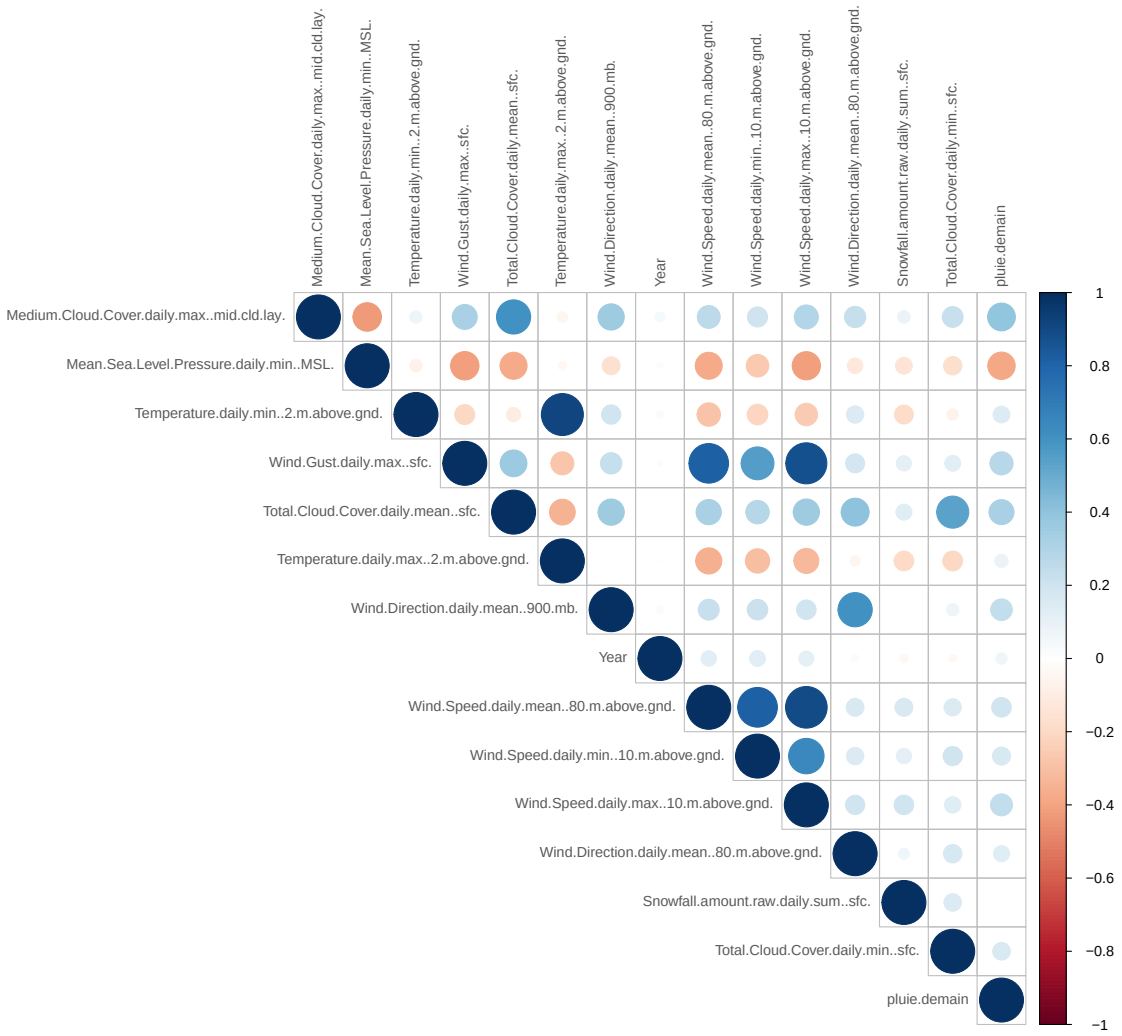
- la température minimale et la température maximale
- la nébulosité moyenne et la nébulosité maximale
- La direction de vente et la rafale

```
## [1] "(Intercept)"
## [2] "Medium.Cloud.Cover.daily.max..mid.cld.lay."
## [3] "Mean.Sea.Level.Pressure.daily.min..MSL."
```

```
## [4] "Temperature.daily.min..2.m.above.gnd."
## [5] "Wind.Gust.daily.max..sfc."
## [6] "Total.Cloud.Cover.daily.mean..sfc."
## [7] "Temperature.daily.max..2.m.above.gnd."
## [8] "Wind.Direction.daily.mean..900.mb."
```

Nous allons afficher la corrélation entre les variables retnues par *aic* et éliminer les variables corrélées en privilégiant les variables retenues par *bic* :

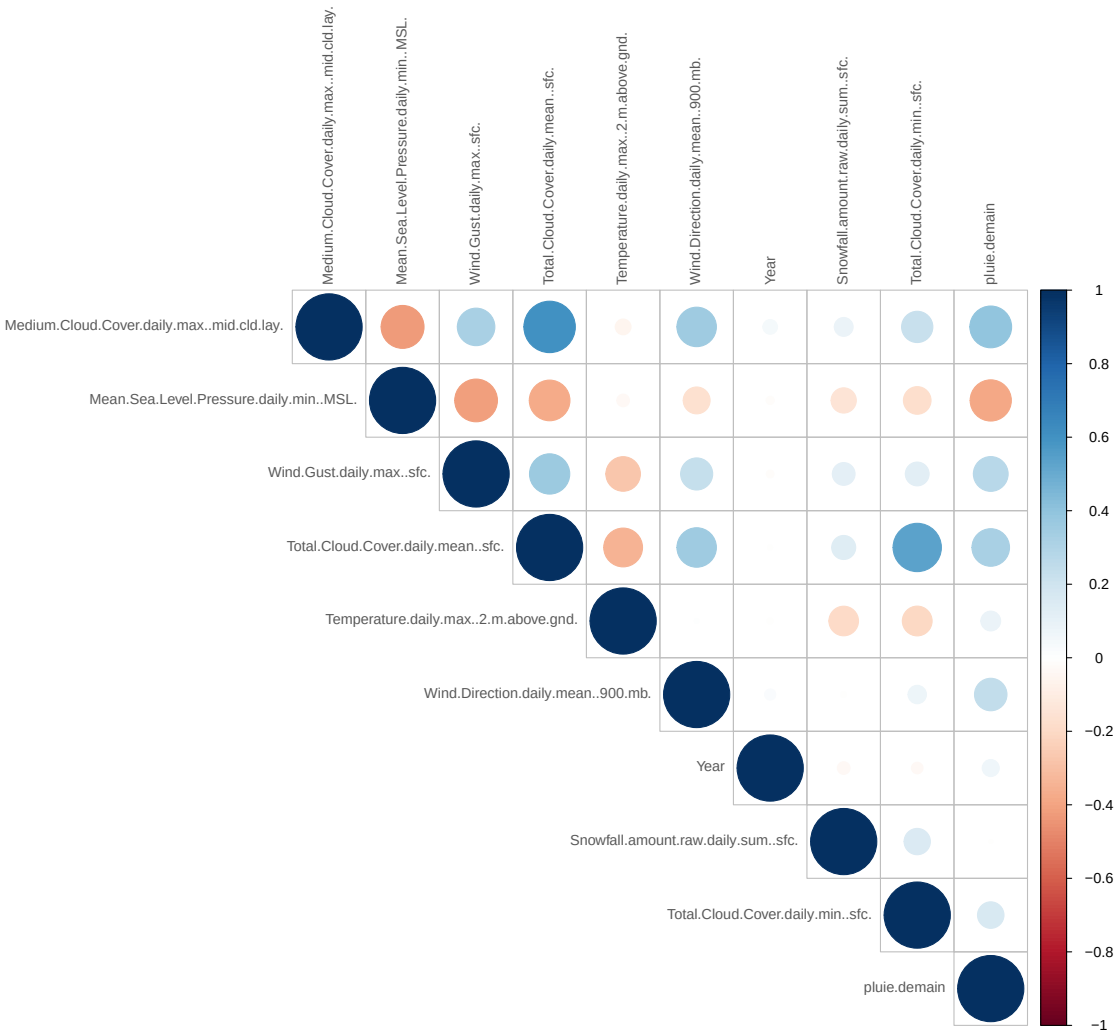
- Il y a une forte corrélation entre la rafale maximale avec la vitesse minimale à 10 m d'altitude, la vitesse maximale à 10 m d'altitude, et la vitesse moyenne à 80 m d'altitude. Aucune des 4 variables a été retenue par *bic*. Nous allons garder seulement la rafale maximale. Comme elle est la plus fortement corrélée avec les trois autres variables, elle expliquerait bien les variabilités expliquées par ces trois variables après leur élimination.
- Une autre forte corrélation est présente entre la direction moyenne de vent à 80 m d'altitude et la direction moyenne de vent à 900 m d'altitude. Nous allons garder celle à l'altitude 900 m car elle est retenue par *bic* alors que celle à l'altitude 80 m n'est pas rentenue par *bic*.
- Nous avons également une forte corrélation etnre la température minimale et la température maximale. Néanmoins les deux variables sont retenues par *bic*. Nous allons garder la température maximale car l'entraînement par la régression logistique considère qu'elle seule est significative.



En éliminant les 5 variables corrélées, nous n'avons plus de problème de corrélations. Nous nous retrouvons avec 10 variables qui sont :

- La nébulosité maximale
- La nébulosité minimale
- La nébulosité moyenne
- La pression minimale
- La rafale
- La température maximale
- La direction moyenne de vent à l'altitude de 900 m
- L'année
- La quantité totale de negie
- Le booléen indiquant s'il a plu le lendemain ou pas

Ce sera notre jeu de données final pour trouver un modèle de prédiction.



# Régression Logistique

Nous pouvons effectuer la régression logistique avec les 10 variables que nous avons rentenu.

## Séparation de données

Pour cela, nous allons séparer le jeu de données en set d'entraînement et set de test afin de pouvoir évaluer et valider la performance du modèle.

Le set d'entrainement sert à entraîner le modèle. Le modèle apprends du set en estimant les coefficients qui définissent la relation entre la variable cible et les variables explicatives.

Le set de test sert à évaluer la performance du modèle. Seules avec les variables explicatives, nous utilisons le modèle pour prédire la variable d'intérêt et comparons avec les vraies valeurs.

L'entraînement sur le set d'entraînement et l'évaluation de performance sur le set de test nous permettrait d'estimer à quel point le modèle est performant sur les données non observées.

Les deux classes de notre variable cible sont bien équilibrées. Nous pouvons donc séparer les données de façon aléatoire. 80% de nos données seront gardées pour l'entraînement et 20% pour le test.

```
set.seed(1)
split = sample(2, nrow(selection_final), replace = T, prob = c(0.8, 0.2))
train = selection_final[split == 1, ]
test = selection_final[split == 2, ]
```

## Modèle 1

Dans un premier temp, nous pouvons entraîner en gardant tous les 10 covariables du set d'entraînement. Il y a quatre variables, *la nébulosité maximale*, *la pression minimale*, *la rafale*, et *la température maximale* qui sont detectés significatives.

```
##
## Call:
## glm(formula = pluie.demain ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2918  -0.8793   0.3620   0.8585   2.6821
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                           -25.472068  65.935698  -0.386
## Medium.Cloud.Cover.daily.max..mid.cld.lay.   0.010960   0.002360   4.645
## Mean.Sea.Level.Pressure.daily.min..MSL.     -0.069984   0.011651  -6.007
## Wind.Gust.daily.max..sfc.                    0.023394   0.006425   3.641
## Total.Cloud.Cover.daily.mean..sfc.           0.005145   0.003942   1.305
## Temperature.daily.max..2.m.above.gnd.        0.057114   0.011986   4.765
## Wind.Direction.daily.mean..900.mb.           0.001996   0.001128   1.769
## Year                                         0.046331   0.032105   1.443
## Snowfall.amount.raw.daily.sum..sfc.          0.191863   0.424334   0.452
## Total.Cloud.Cover.daily.min..sfc.            0.007517   0.004294   1.751
##                                          Pr(>|z|)
```

```
## (Intercept)                                  0.699262
## Medium.Cloud.Cover.daily.max..mid.cld.lay. 3.40e-06 ***
## Mean.Sea.Level.Pressure.daily.min..MSL.    1.89e-09 ***
## Wind.Gust.daily.max..sfc.                  0.000272 ***
## Total.Cloud.Cover.daily.mean..sfc.         0.191856
## Temperature.daily.max..2.m.above.gnd.      1.89e-06 ***
## Wind.Direction.daily.mean..900.mb.         0.076833 .
## Year                                       0.148991
## Snowfall.amount.raw.daily.sum..sfc.        0.651160
## Total.Cloud.Cover.daily.min..sfc.          0.080021 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1314.0  on 947  degrees of freedom
## Residual deviance: 1032.9  on 938  degrees of freedom
## AIC: 1052.9
##
## Number of Fisher Scoring iterations: 4
```

## Modèle 2

Nous pouvons créer un deuxième modèle en gardant les 4 covariables detectées comme significatives dans le modèle 1. Ce modèle detecte tous les covariables comme significatives. Nous pouvons sélectionner ce modèle.

Nous avons des coefficients positifs pour la nébulosité maximale, la température maximale et la direction moyenne de vent. La probabilité de pleuvoir le lendemain augmente avec l'augmentation de l'un des ces trois variables.

Nous avons un coefficient négatif pour la pression minimale. La probabilité de pleuvoir le lendemain diminue avec l'augmentation de pression.

```
##
## Call:
## glm(formula = pluie.demain ~ . - Total.Cloud.Cover.daily.mean..sfc. -
##     Year - Wind.Gust.daily.max..sfc. - Snowfall.amount.raw.daily.sum..sfc. -
##     Total.Cloud.Cover.daily.min..sfc., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0637  -0.9140   0.4922   0.8737   2.6751
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                85.310993  11.210480   7.610
## Medium.Cloud.Cover.daily.max..mid.cld.lay.  0.014446   0.002035   7.099
## Mean.Sea.Level.Pressure.daily.min..MSL.    -0.086283   0.011014  -7.834
## Temperature.daily.max..2.m.above.gnd.       0.031800   0.010049   3.164
## Wind.Direction.daily.mean..900.mb.          0.003112   0.001069   2.910
##                                            Pr(>|z|)
## (Intercept)                                2.74e-14 ***
## Medium.Cloud.Cover.daily.max..mid.cld.lay. 1.25e-12 ***
## Mean.Sea.Level.Pressure.daily.min..MSL.    4.72e-15 ***
```

```
## Temperature.daily.max..2.m.above.gnd.       0.00155 **
## Wind.Direction.daily.mean..900.mb.          0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1314.0  on 947  degrees of freedom
## Residual deviance: 1058.4  on 943  degrees of freedom
## AIC: 1068.4
##
## Number of Fisher Scoring iterations: 4
```

## Modèle 3

Nous pouvons également créer un troisième modèle en appliquant la fonction *step* qui nous donne un modèle avec une combinaison de covariables qui minimise *aic*

Le modèle sélectionné contient 6 variables qui sont toutes significatives.

Nous avons des coefficients positifs pour la nébulosité moyenne maximale, la rafale maximale, la température maximale, la directin moyenne de vent, et la nébulosité totale minimale. La probabilité de pleuvoir le lendemain augmente avec l'augmentation de l'un des ces trois variables.

Nous avons un coefficient négatif pour la pression minimale. La probabilité de pleuvroir le lendemain diminue avec l'augmentation de pression.

```
##
## Call:
## glm(formula = pluie.demain ~ Medium.Cloud.Cover.daily.max..mid.cld.lay. +
##     Mean.Sea.Level.Pressure.daily.min..MSL. + Wind.Gust.daily.max..sfc. +
##     Temperature.daily.max..2.m.above.gnd. + Wind.Direction.daily.mean..900.mb. +
##     Total.Cloud.Cover.daily.min..sfc., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2632  -0.8766   0.3619   0.8682   2.6572
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             70.227629  11.802626   5.950
## Medium.Cloud.Cover.daily.max..mid.cld.lay.  0.012580   0.002086   6.032
## Mean.Sea.Level.Pressure.daily.min..MSL.    -0.072200   0.011549  -6.252
## Wind.Gust.daily.max..sfc.                   0.023814   0.006388   3.728
## Temperature.daily.max..2.m.above.gnd.       0.050511   0.010958   4.609
## Wind.Direction.daily.mean..900.mb.          0.002386   0.001093   2.183
## Total.Cloud.Cover.daily.min..sfc.           0.010268   0.003762   2.730
##                                          Pr(>|z|)
## (Intercept)                             2.68e-09 ***
## Medium.Cloud.Cover.daily.max..mid.cld.lay. 1.62e-09 ***
## Mean.Sea.Level.Pressure.daily.min..MSL.    4.06e-10 ***
## Wind.Gust.daily.max..sfc.                  0.000193 ***
## Temperature.daily.max..2.m.above.gnd.      4.04e-06 ***
## Wind.Direction.daily.mean..900.mb.         0.029060 *
## Total.Cloud.Cover.daily.min..sfc.          0.006340 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1314.0  on 947  degrees of freedom
## Residual deviance: 1036.7  on 941  degrees of freedom
## AIC: 1050.7
##
## Number of Fisher Scoring iterations: 4
```

# Évaluation de modèle

Nous allons maintenant évaluer le modèle 2 et le modèle 3 qui ont gardé que des covariables significatives en faisant la prédiction sur le set de test.

Pour commencer, nous allons fixé le seuil à 50%. Tous les observations ayant une probabilité de pleuvoir le lendemain supérieur à 0,5 auront la valeur *True* et le reste la valeur *False*.

## Précision

La précision est une métrique basée sur la matrice de confusion pour évaluer la performance de modèle de classification. Elle correspond à la proportion de bonne prédiction sur la prédiction totale.

Le modèle 2 a une précision de 73%.

```
## [1] 0.7284483
```

```
##
## pred_2  FALSE TRUE
##   FALSE    72   23
##   TRUE     40   97
```

Le modèle 3 a également une précision de 73%.

```
## [1] 0.7284483
```

```
##
## pred_3  FALSE TRUE
##   FALSE    72   23
##   TRUE     40   97
```

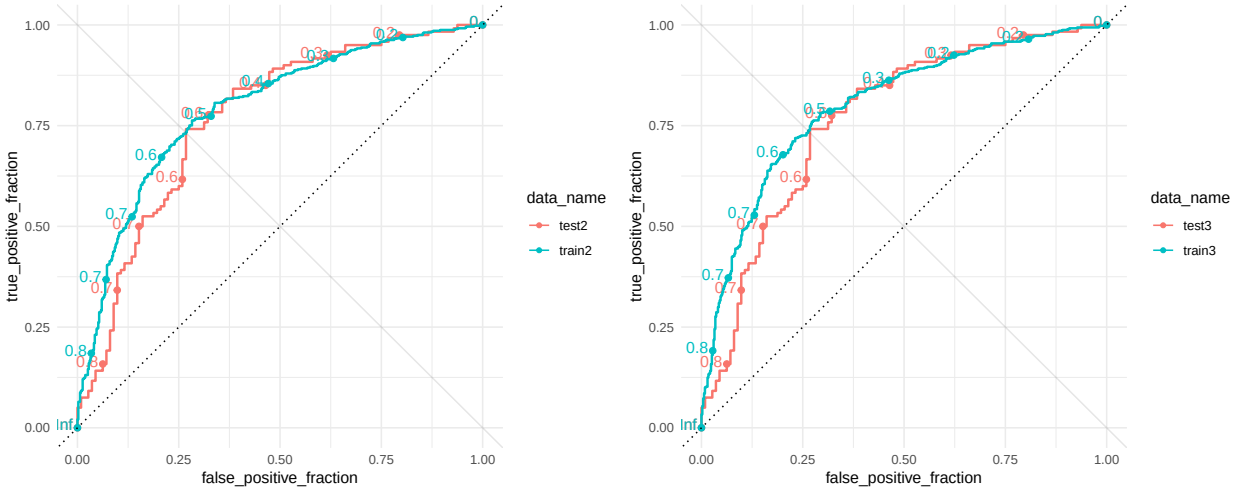
Dans notre cas, la valeur positive semble mieux predire correctement que la valeur négative mais la différence entre les deux classes n'est pas alertant.

## ROC

Une optimisation du taux de vrai positif et du taux de faux positif à l'aide de la courbe ROC nous permettra d'avoir un modèle performant. La courbe ROC représente le taux de vrai positif en fonction du taux de faux positif.

Nous cherchons un seuil qui maximise le taux de vrai positif et minimise le taux de faux positif. Notre courbe ROC est symétrique avant et après le seuil 0,5.

0,5 est le bon seuil pour notre modèle.




## AUC

Nous pouvons également regarder l'AUC, laire sous la courbe ROC. Elle représente le degré de séparabilité. Elle nous dit à quel point le modèle distingue entre les différentes classes.

L'AUC est en général compris entre 0,5 et 1. où 0,5 est égale à un modèle aléatoire et 1 est égale au modèle parfait qui prédit parfaitement toutes les observations

Pour le modèle 2, nous avons une AUC de 0,77 pour le set test et 0,79 pour le set train.

```
calc_auc(rocr2)
```

```
##   PANEL group data_name       AUC
## 1     1     1     test2 0.7679315
## 2     1     2    train2 0.7926919
```

Pour le modèle 3, nous avons une AUC de 0,77 pour le set test et 0,80 pour le set train.

```
calc_auc(rocr3)
```

```
##   PANEL group data_name       AUC
## 1     1     1     test3 0.7679315
## 2     1     2    train3 0.8018805
```

Les deux modèles ont des performances très correctes. Le modèle 3 ajuste légèrement mieux mais tous les deux modèles généralisent aussi bien sur le set test. Nous allons donc garder le modèle 3.

## Prédiction

Avec notre modèle retenu, nous pouvons faire la prédiction avec un seuil de 0,5 sur le jeu de données à prédire fourni.

Au final, 55% de jours est prédit pluvieux le lendemain et 45% de jours le contraire parmi les 290 jours.