

Projet fil rouge Data Scientest

**Projet : Prévisions météo en Australie**

Réalisé par : Anne-claire OGIERAIKHI, Joseph CHARLES, Olivier Amable. et Geneviève Steele

Encadré par : Laurène BOUSKILA

Table des matières

[1. Présentation du sujet 3](#_Toc116526958)

[1.1. Les données 3](#_Toc116526959)

[1.2. Data visualisation 4](#_Toc116526960)

[1.2.1. Distribution de la variable cible 4](#_Toc116526961)

[1.2.2. Les variables qualitatives 4](#_Toc116526962)

[1.2.3. Les variables quantitatives 6](#_Toc116526963)

[1.2.4. Les matrices de corrélation 7](#_Toc116526964)

[1.2.5. Moyenne annuelle des précipitations 8](#_Toc116526965)

[1.2.6. Représentation cartographique 9](#_Toc116526966)

[2. Sélection de variables 10](#_Toc116526967)

[2.1. La régression de Lasso 10](#_Toc116526968)

[2.2. Sélection de variables avec SelectKBest 11](#_Toc116526969)

[3. Modélisation 11](#_Toc116526970)

# Présentation du sujet

Cet ensemble de données contient environ 10 ans d'observations météorologiques quotidiennes provenant différentes villes en Australie.

Le premier objectif est de prédire la variable cible : RainTomorrow. Elle signifie : a-t-il plu le jour suivant, oui ou non ? Cette colonne est Oui si la pluie pour ce jour était de 1mm ou plus. De même pour des prédictions de vent ou température.

Dans un second temps, on pourra effectuer des prédictions à long terme, en utilisant des techniques mathématiques d’analyse de séries temporelles, et/ou des réseaux de neurones récurrents.

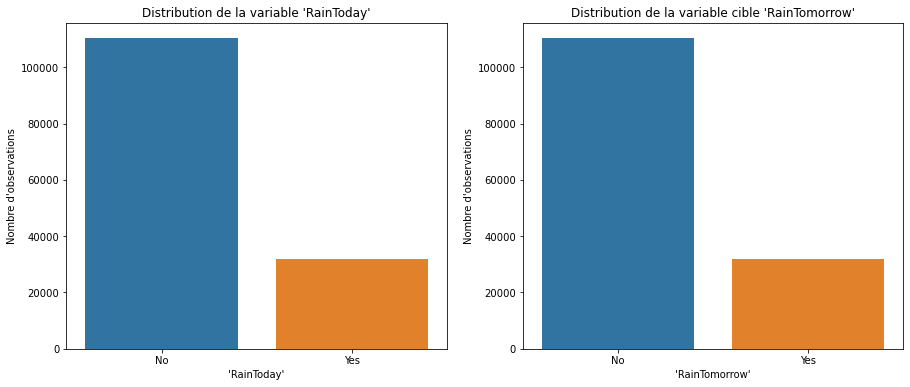
## Les données

Le jeu de données se compose de 23 variables :

* *Date* : Date d'observation.
* *Location* : Ville où se situe la station météorologique.
* *MinTemp* : Température minimale en degrés Celsius.
* *MaxTemp* : Température maximale en degrés Celsius.
* *Rainfall* : Quantité de pluie enregistrée pour la journée en millimètres.
* *Evaporation* : Niveau des bacs d'évaporation de classe A (mm) dans les 24 heures jusqu'à 9h.
* *Sunshine* : Nombre d'heures d'ensoleillement dans la journée.
* *WindGustDir* : Direction des plus fortes rafales de vent dans les 24 heures jusqu'à minuit.
* *WindGustSpeed* : Vitesse en km/h des plus fortes rafales de vent dans les 24 heures jusqu'à minuit.
* *WindDir9am* : Direction du vent à 9h du matin.
* *WindDir3pm* : Direction du vent à 3h de l'après-midi.
* *WindSpeed9am* : Vitesse du vent en km/h moyennée sur 10 minutes avant 9h du matin.
* *WindSpeed3pm* : Vitesse du vent en km/h moyennée sur 10 minutes avant 3h de l'après-midi.
* *Humidity9am* : Humidité en pourcentage à 9h du matin.
* *Humidity3pm* : Humidité en pourcentage à 3h de l'après-midi.
* *Pressure9am* : Pression atmosphérique (hpa) au niveau de la mer à 9h.
* *Pressure3pm* : Pression atmosphérique (hpa) au niveau de la mer à 15h.
* *Cloud9am* : Opacité du ciel obscurci par les nuages à 9h mesuré en "oktas (0 signifie que le ciel est complètement éclairci tandis que 8 indique que le ciel est couvert.
* *Cloud3pm* : Opacité du ciel obscurci par les nuages à 15h mesuré en "oktas
* *Temp9am* : Température en degrés Celsius à 9h du matin.
* *Temp3pm* : Température en degrés Celsius à 3h de l'après-midi.
* *RainToday* : vaut 1 si les précipitations (en mm) dans les 24 heures avant 9h sont supérieur à 1mm, 0 sinon.
* *RainTomorrow* : vaut 1 si les précipitations (en mm) du lendemain sont supérieur à 1mm, 0 sinon

## Data visualisation

### Distribution de la variable cible



Les variables "RainToday" et "RainTomorrow" présentent une forte disparité entre leurs classes, avec :

* 22.4% des observations qui correspondent à la classe minoritaire "Yes",
* 77.6% des observations qui correspondent à la classe majoritaire "No".

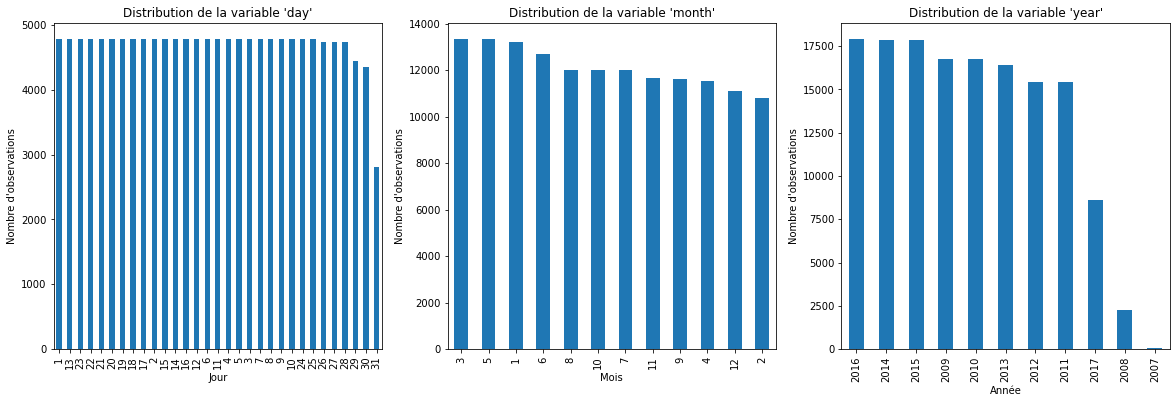
Il s'agit d'un problème de classification binaire sur jeu de données déséquilibré. Il est possible que ces deux variables soient fortement corrélées.

La mesure du niveau de corrélation entre ses deux variables qualitatives par le V de Cramer nous donne 0.3079. La p\_value est nulle et le V\_Cramer n'est pas proche de 1, donc les variables 'RainToday' et 'RainTomorrow' ne sont pas très corrélées.

### Les variables qualitatives

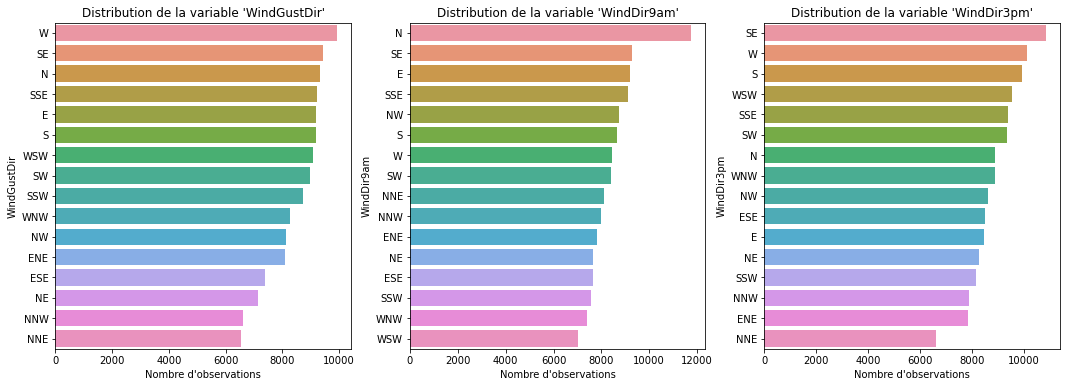
D’après le graphique ci-dessous, les derniers jours du mois possèdent moins d'observations, ce qui est cohérent sur une période de 10 années. Les observations sont inégalement réparties sur les mois de l'année. L’année 2007 ne possède quasiment pas d'observations. Les années 2008 et 2017 possèdent beaucoup moins d'observations que les autres années.

**Mettre en avant la suppression de la variable Date**



Les observations sont inégalement réparties selon les villes où se situe la station météorologique. La majorité des villes possèdent entre 3000 et 3500 observations.





Les variables "WindGustDir", "WindDir9am" et "WindDir3pm" présentent 16 modalités chacune et la distribution des observations assez uniforme entre les différentes classes.

Pour la suite du projet, nous décidons d’encoder les variables qualitatives.

### Les variables quantitatives

Une image contenant table

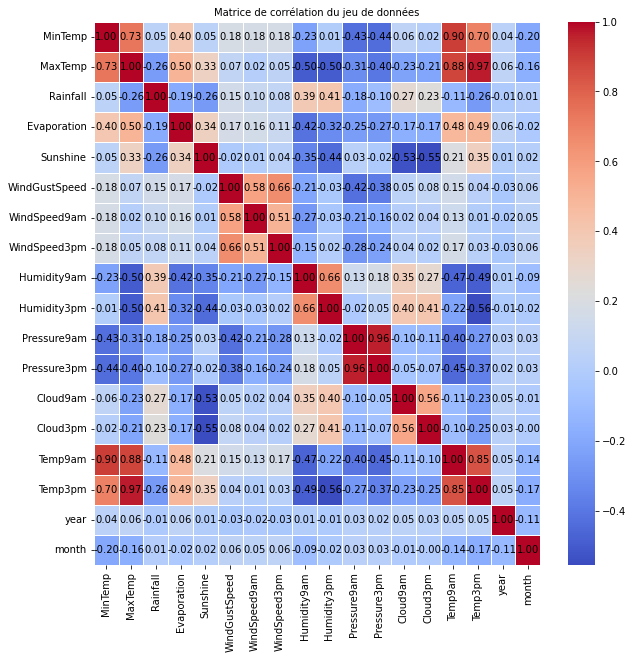
Description générée automatiquement

Une image contenant table

Description générée automatiquement Une image contenant table

Description générée automatiquement

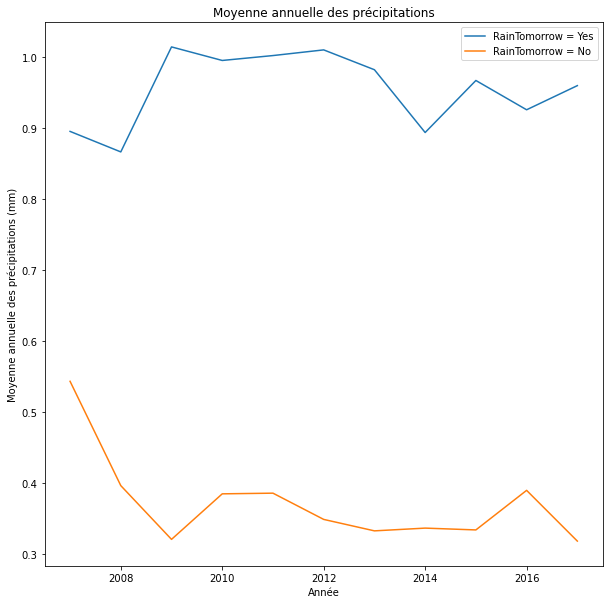
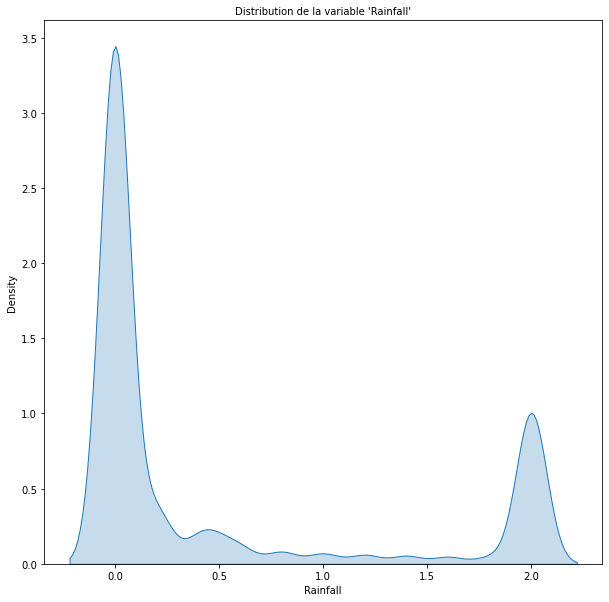
### Les matrices de corrélation



**Mettre matrice de corrélation mise à jour sans les variables corrélées entre elles**

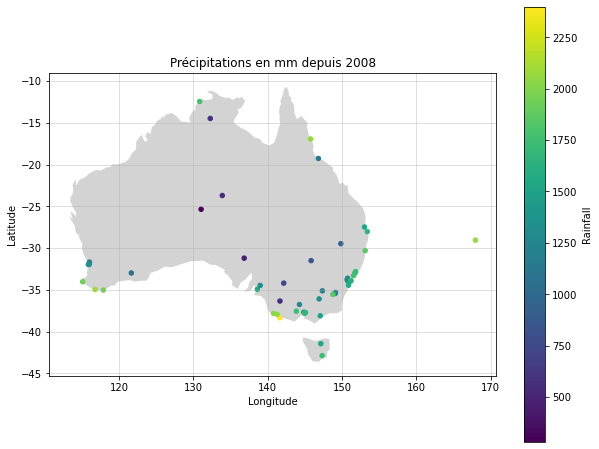
* Les variables 'MinTemp' et 'MaxTemp' sont fortement positivement corrélées (coefficient de corrélation = 0.73)
* Les variables 'MinTemp' et 'Temp9am' sont fortement positivement corrélées (coefficient de corrélation = 0.90)
* Les variables 'MinTemp' et 'Temp3pm' sont fortement positivement corrélées (coefficient de corrélation = 0.70)
* Les variables 'MaxTemp' et 'Temp9am' sont fortement positivement corrélées (coefficient de corrélation = 0.88)
* Les variables 'MaxTemp' et 'Temp3pm' sont fortement positivement corrélées (coefficient de corrélation = 0.97)
* Les variables 'WindGustSpeed' et 'WindSpeed3pm' sont fortement positivement corrélées (coefficient de corrélation = 0.66)
* Les variables 'Pressure9am' et 'Pressure3pm' sont fortement positivement corrélées (coefficient de corrélation = 0.96)
* Les variables 'Temp9am' et 'Temp3pm' sont fortement positivement corrélées (coefficient de corrélation = 0.85)
* Les variables 'Cloud9am' ou 'Cloud3pm' et 'Sunshine' sont fortement negativement corrélées (coefficient de corrélation = -0.70)

### Moyenne annuelle des précipitations



On observe que la distribution de la variable "Rainfall" est fortement concentrée autour des valeurs 0 mm et 2 mm. Sur le deuxième graphique, lorsque "RainTomorrow" est égale à "Yes", la moyenne annuelle des précipitations est supérieure à 0.8 mm mais lorsque "RainTomorrow" est égale à "No", la moyenne annuelle des précipitations est inférieure à 0.6 mm.

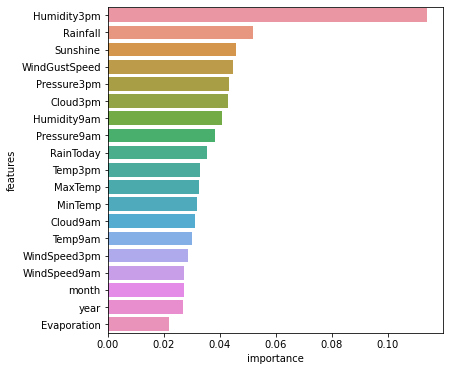
### Représentation cartographique



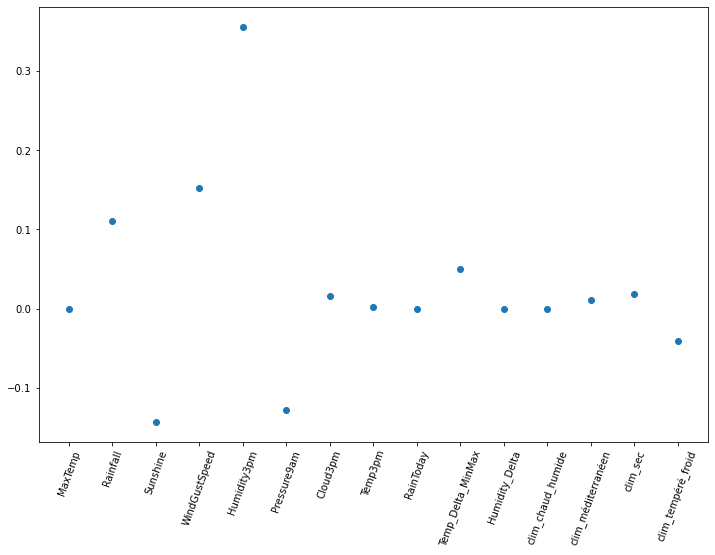
Les villes australiennes se trouvant près des côtes ont tendance à recevoir plus de précipitations que les villes se trouvant à l’intérieur du pays. En effet, ses villes côtières ont depuis 2008 reçu plus de 1750 mm de précipitations.

# Sélection de variables

À la suite d’une classification par forêt aléatoire, nous avons fait ressortir les variables par ordre d’importance :



### La régression de Lasso



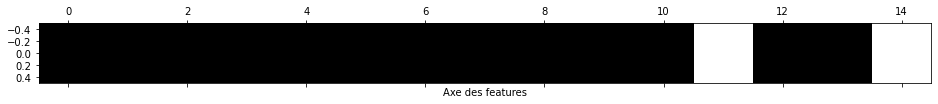
Score du modèle sur l'ensemble d’apprentissage : 0.2997084093093404

Score du modèle sur l'ensemble de test : 0.30626955235235587

10 variables sur 12 ont été sélectionnées par le modèle, avec le paramètre "alpha = 0.01". Les coefficients des 2 autres variables sont estimées à 0.

Le meilleur paramètre retenu par la Grid search est {'alpha': 0.001} avec pour sélection de variable : ['MaxTemp', 'Rainfall', 'Sunshine', 'WindGustSpeed', 'Humidity3pm','Pressure9am', 'Cloud3pm', ‘Temp3pm', 'Temp\_Delta\_MinMax', 'Humidity\_Delta', 'clim\_méditerranéen','clim\_sec', 'clim\_tempéré\_froid']

### Sélection de variables avec SelectKBest



Les features retenus sont ['MaxTemp', 'Rainfall', 'Sunshine', 'WindGustSpeed', 'Humidity3pm', 'Pressure9am', 'Cloud3pm', 'Temp3pm', 'RainToday', 'Temp\_Delta\_MinMax','Humidity\_Delta', 'clim\_méditerranéen', 'clim\_sec']

# Modélisation