INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Soutenance de PFE : Pluie en Australie

Théophile THIERRY

25 février 2022

Bruno PORTIER

Sujet du projet

“Utilisation d’algorithmes de machines learning pour la résolution d’un challenge Kaggle.”

- Présentation de la base de données
  - Présentation des variables
  - Cartographie, étude des climats et périodicité
  - Complétion des données
  - Analyse des relations entre nos variables
- Prédiction
  - Mise en lumière du déséquilibre avec une ACP
  - Méthodes de rééchantillonnage et de modification de variables avec la régression logistique
  - Construction d’autres modèles de prédiction

Présentation de la base de données

Présentation des variables

Les 23 variables de la base de données “Rain in Australia”

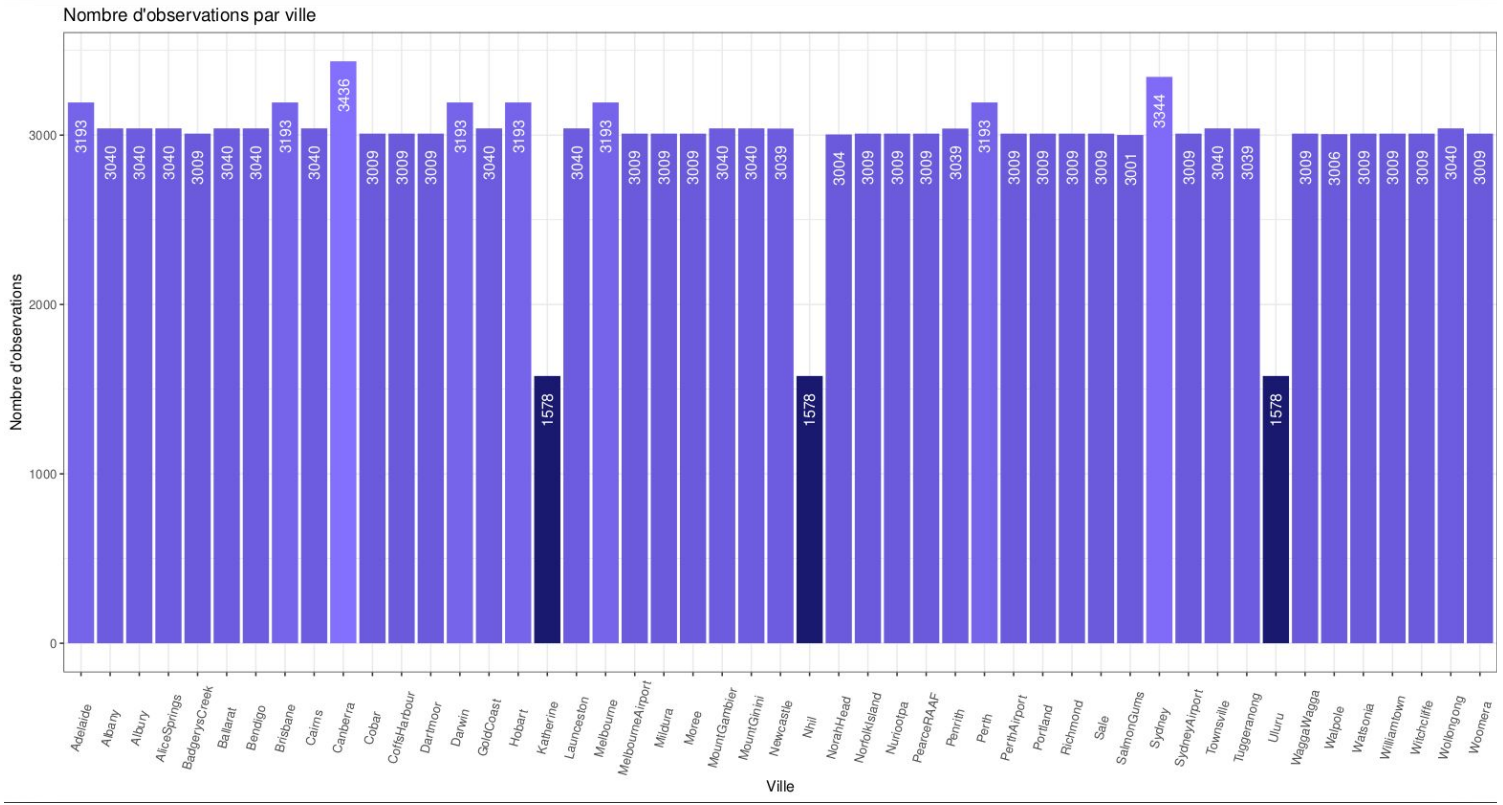
- Date
- Location
- MinTemp / MaxTemp / Temp9am / Temp3pm
- Rainfall / RainToday / RainTomorow / Evaporation
- Sunshine / Cloud9am / Cloud3pm
- WindGustDir / WindGustSpeed / WindDir9am / WindDir3pm / WindSpeed9am / WindSpeed3pm
- Humidity9am / Humidity3pm
- Pressure9am / Pressure3pm

Variables Numériques

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	Std.
MinTemp	-8.50	7.60	12.00	12.19	16.90	33.90	1485.00	6.40
MaxTemp	-4.80	17.90	22.60	23.22	28.20	48.10	1261.00	7.12
Rainfall	0.00	0.00	0.00	2.36	0.80	371.00	3261.00	8.48
Evaporation	0.00	2.60	4.80	5.47	7.40	145.00	62790.00	4.19
Sunshine	0.00	4.80	8.40	7.61	10.60	14.50	69835.00	3.79
WindGustSpeed	6.00	31.00	39.00	40.04	48.00	135.00	10263.00	13.61
WindSpeed9am	0.00	7.00	13.00	14.04	19.00	130.00	1767.00	8.92
WindSpeed3pm	0.00	13.00	19.00	18.66	24.00	87.00	3062.00	8.81
Humidity9am	0.00	57.00	70.00	68.88	83.00	100.00	2654.00	19.03
Humidity3pm	0.00	37.00	52.00	51.54	66.00	100.00	4507.00	20.80
Pressure9am	980.50	1012.90	1017.60	1017.65	1022.40	1041.00	15065.00	7.11
Pressure3pm	977.10	1010.40	1015.20	1015.26	1020.00	1039.60	15028.00	7.04
Cloud9am	0.00	1.00	5.00	4.45	7.00	9.00	55888.00	2.89
Cloud3pm	0.00	2.00	5.00	4.51	7.00	9.00	59358.00	2.72
Temp9am	-7.20	12.30	16.70	16.99	21.60	40.20	1767.00	6.49
Temp3pm	-5.40	16.60	21.10	21.68	26.40	46.70	3609.00	6.94

TABLE 1 – Résumé des variables numériques.

Variables Factorielles (Date & Location)



Variables Factorielles (Vent et Pluie)

	E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW
Degré	0.0	22.5	45.0	67.5	90.0	112.5	135.0	157.5	180.0	202.5	225.0	247.5	270.0	292.5	315.0	337.5

TABLE 5 – Les 16 points cardinaux en degrés.

	No	Yes	NA's
Compte	110319	31880	3261
%	75.84	21.92	2.24

TABLE 6 – Variable RainToday.

	No	Yes	NA's
Compte	110316	31877	3267
%	75.84	21.91	2.25

TABLE 7 – Variable RainTomorrow.

Cartographie

Dessiner une carte

Système Géodésique

Ellipsoïde

Celui du World Geodetic System 84

01

Système de coordonnées

Latitude et Longitude

02

03

Projection

Dite “géographique” :
Latitude = X et
Longitude = Y

Etude Des Climats

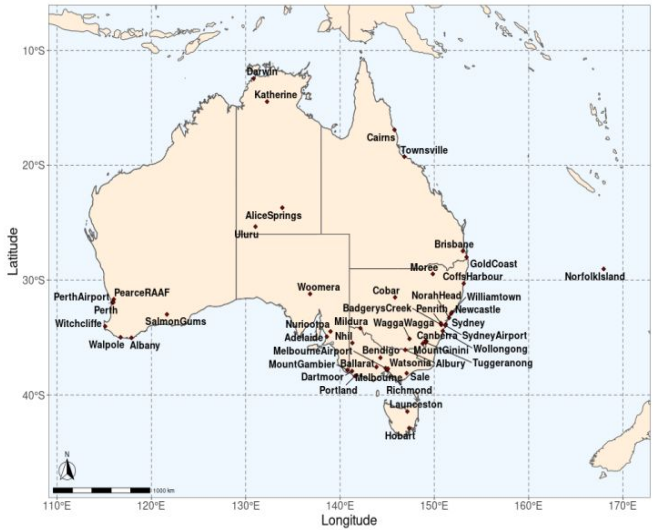


FIGURE 3 – Carte de l'Australie avec les lieux observés.

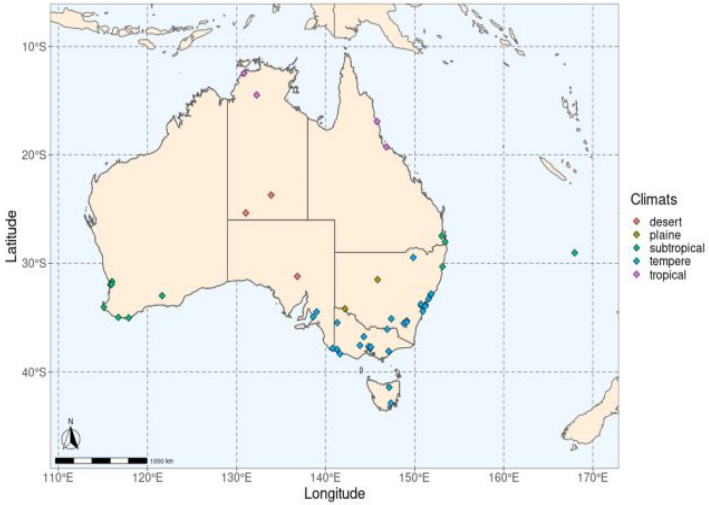
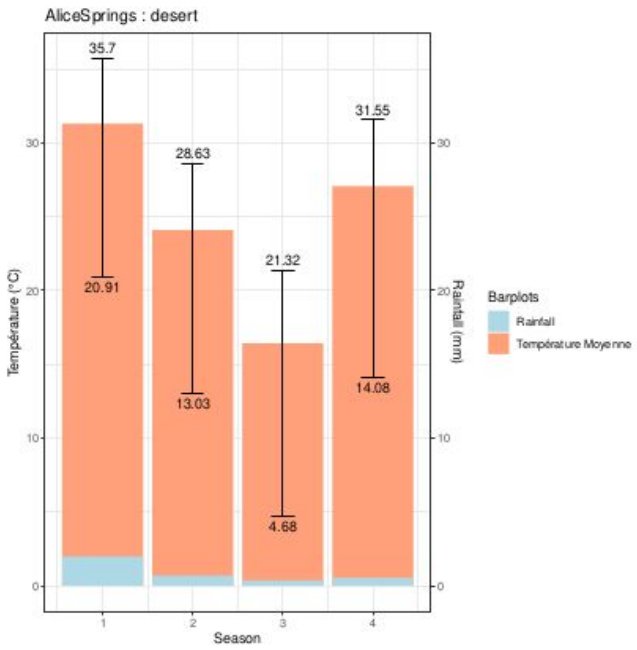
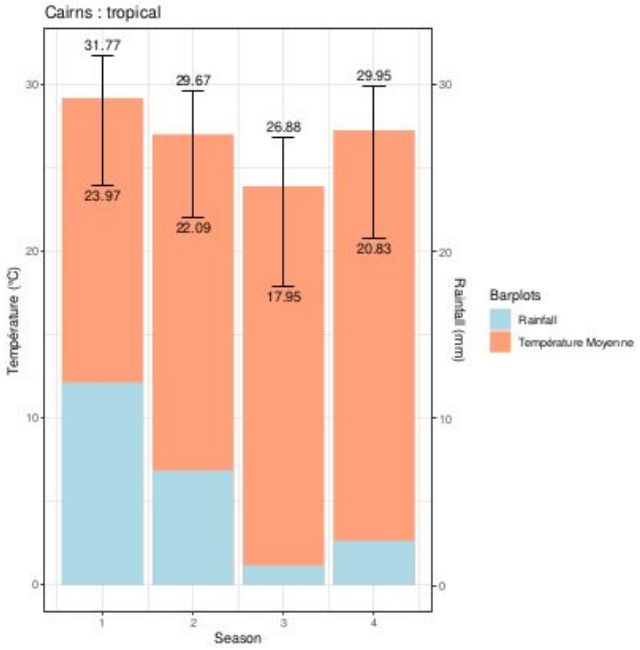


FIGURE 4 – Carte de l'Australie avec les climats des lieux observés.

Etude des climats (particularités & saisons)





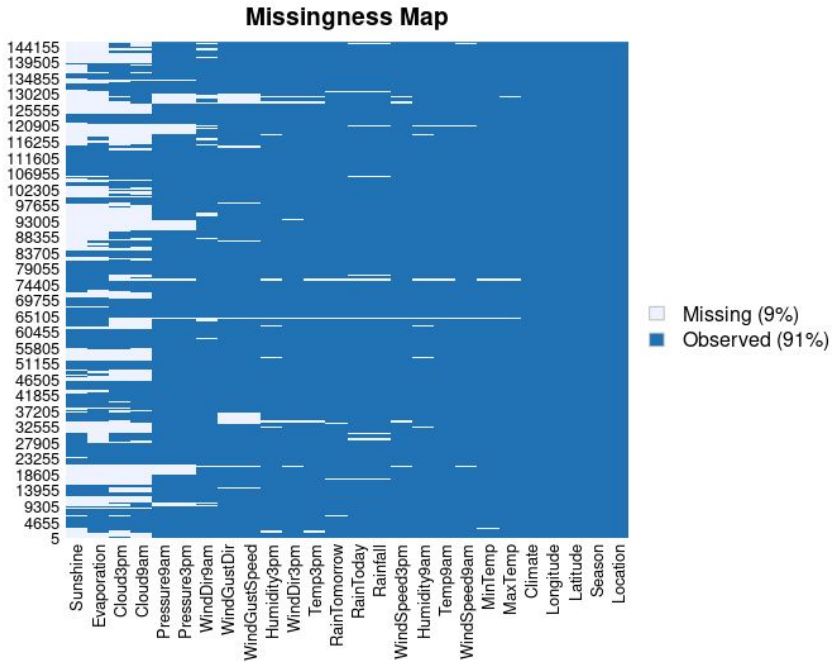
Les nouvelles variables

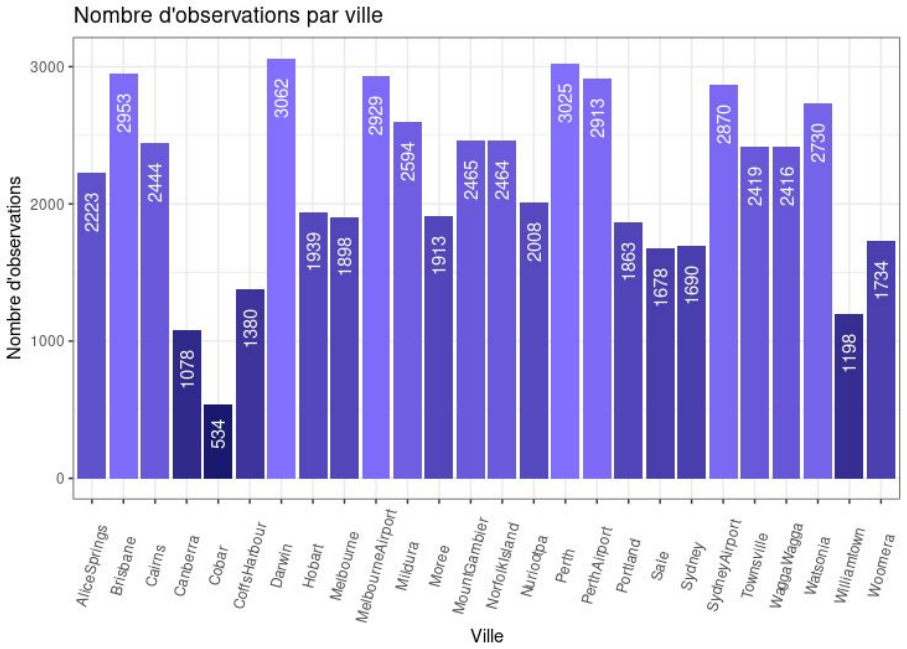
Que des numériques sauf Season (4 niveaux), Climate (5 niveaux) et les variables booléennes RainToday et RainTomorrow :

- Date > Season
- Location > Longitude & Latitude
- MinTemp / MaxTemp / Temp9am / Temp3pm
- Rainfall / RainToday / RainTomorow / Evaporation
- Sunshine / Cloud9am / Cloud3pm
- WindGustDir / WindGustSpeed / WindDir9am / WindDir3pm / WindSpeed9am / WindSpeed3pm
- Humidity9am / Humidity3pm
- Pressure9am / Pressure3pm
- + Climate

Complétion des données

Pourquoi Compléter ?





Comment Compléter ?

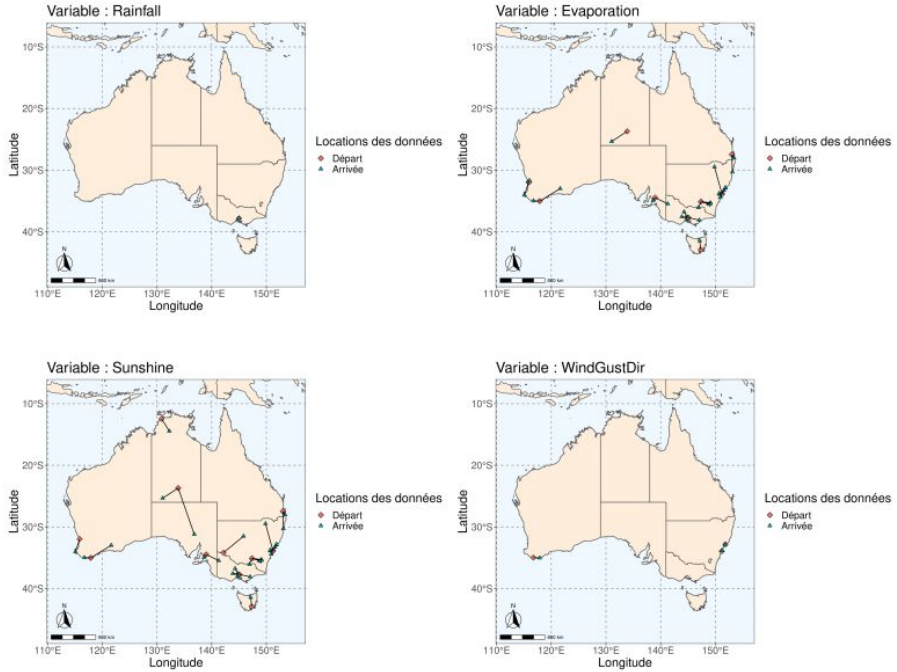
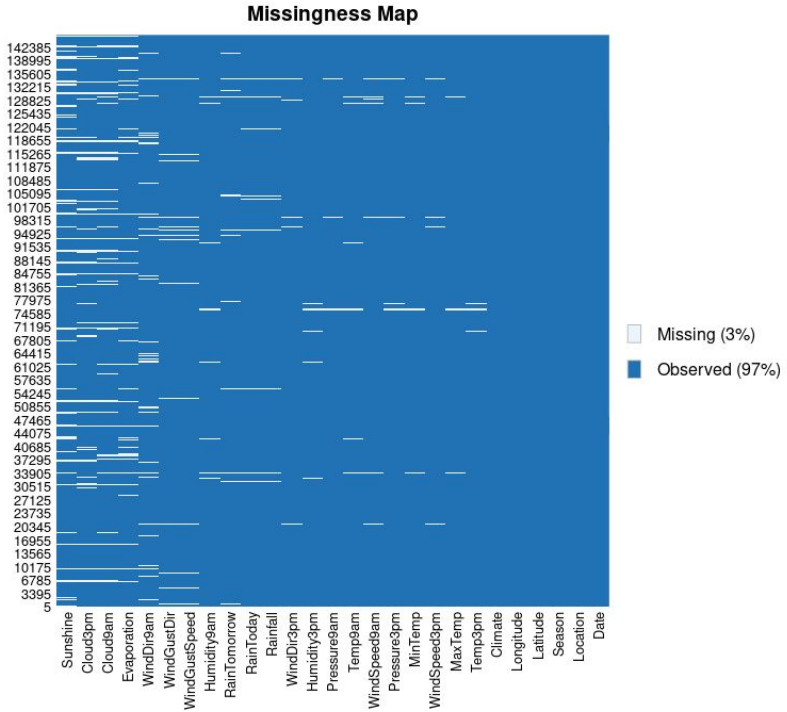
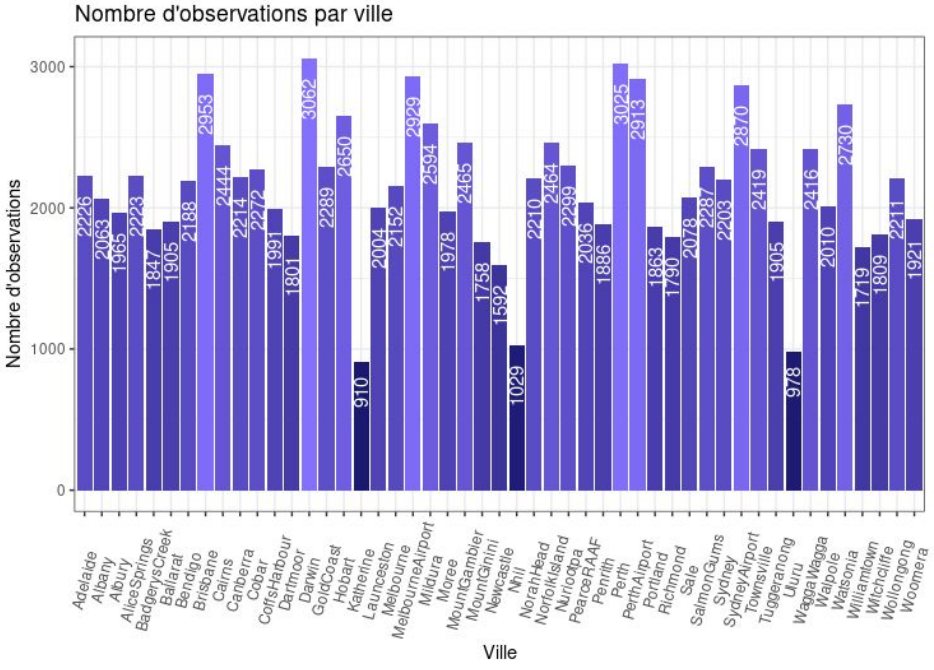


FIGURE 19 – Chemin des observations copiées (ville de départ et ville(s) d'arrivée(s)) pour certaines des variables complétées 1/4.

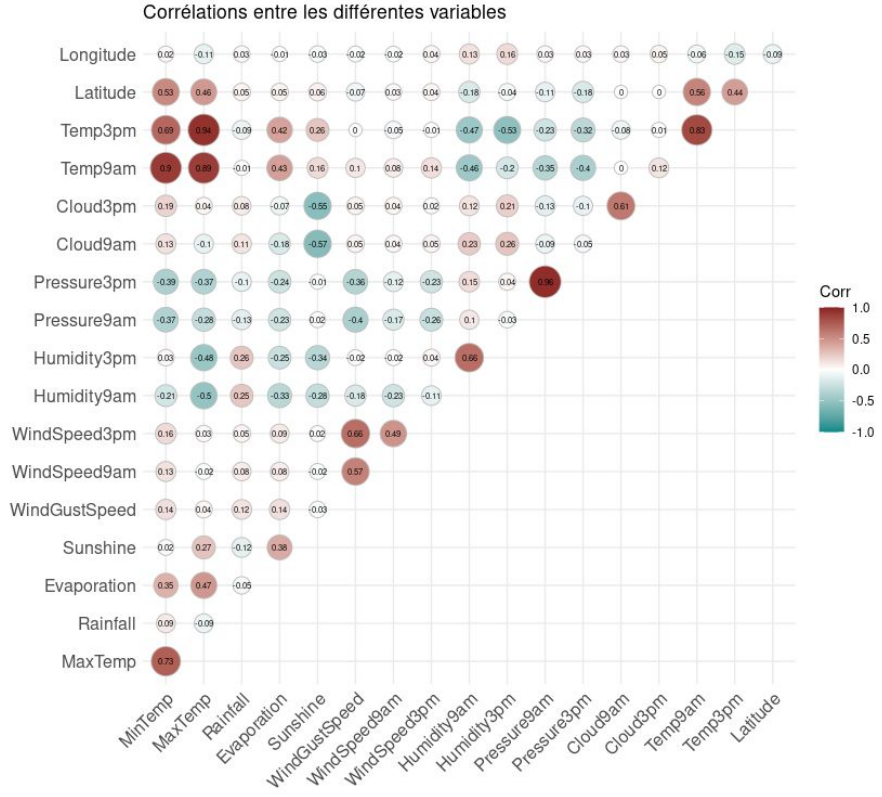
Après Complétion



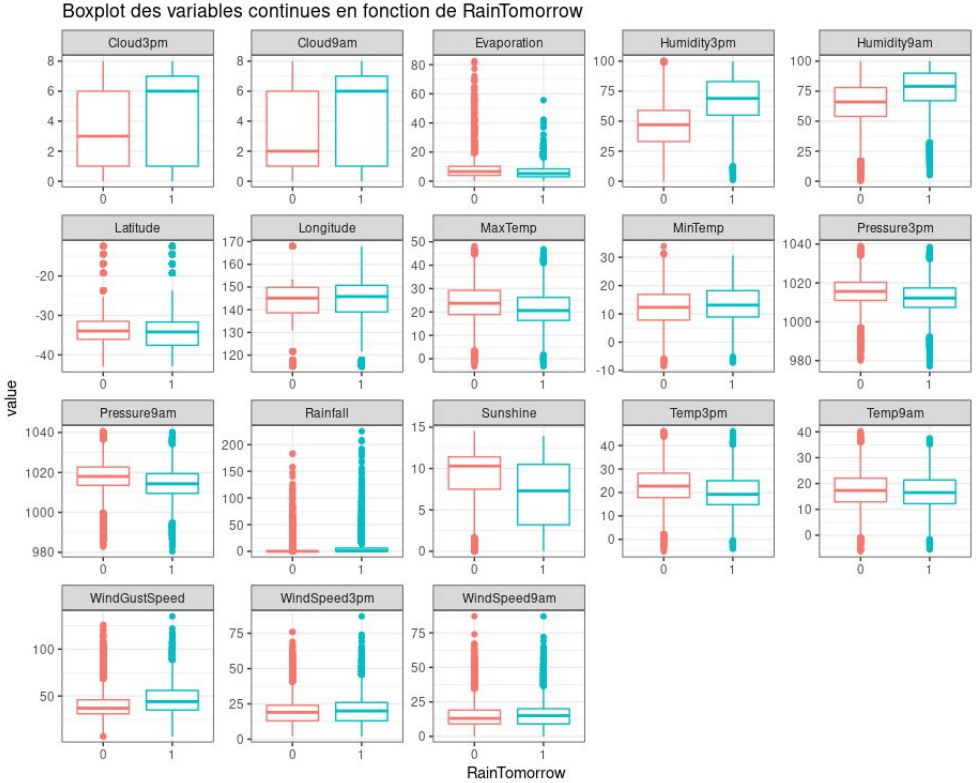


Relations entre les variables

Corrélations

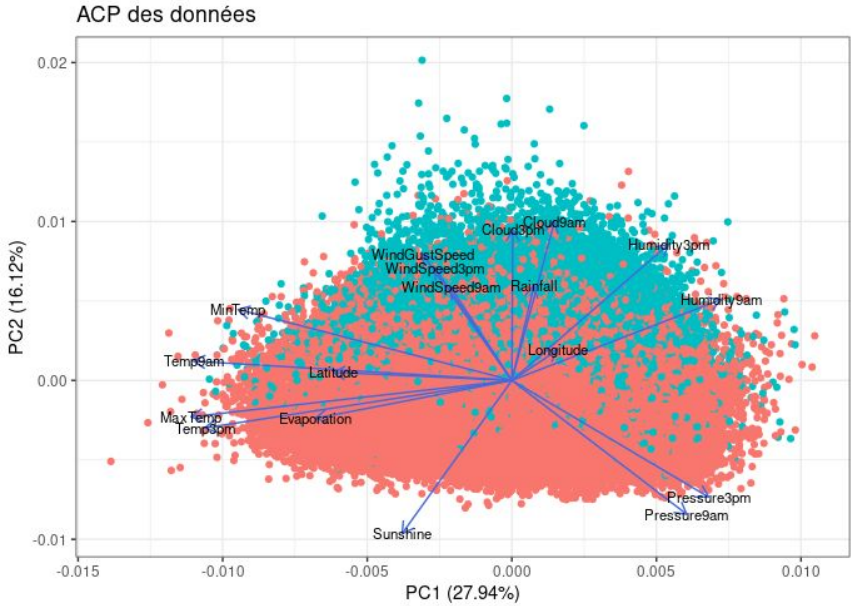


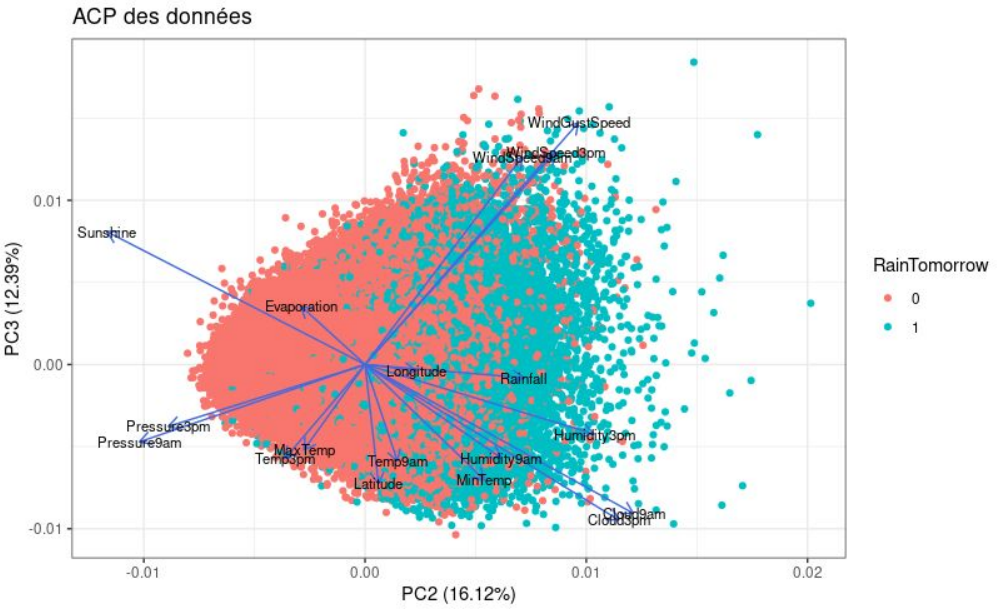
Boîtes à moustaches pour RainTomorrow



Prédictions

ACP


ACP des données
PC2 (16.12%)
PC1 (27.94%)
0.02
0.01
0.00
-0.01
-0.015
-0.010
-0.005
0.000
0.005
0.010
Cloud3pm
Cloud9am
Humidity3pm
WindGustSpeed
WindSpeed3pm
WindSpeed9am
Rainfall
Humidity9am
MinTemp
Longitude
Temp9am
Latitude
MaxTemp
Temp3pm
Evaporation
Pressure3pm
Pressure9am
Sunshine


ACP des données
PC3 (12.39%)
PC2 (16.12%)
0.01
0.00
-0.01
-0.01
0.00
0.01
0.02
WindGustSpeed
WindSpeed3pm
WindSpeed9am
Sunshine
Evaporation
Longitude
Rainfall
Pressure3pm
Pressure9am
Humidity3pm
MaxTemp
Temp3pm
Temp9am
Humidity9am
MinTemp
Latitude
Cloud9am
Cloud3pm
RainTomorrow
0
1

One-hot encoding pour Season et Climate

Season.1	Season.2	Season.3	Season.4	Climate.1	Climate.2	Climate.3	Climate.4	Climate.5
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00

TABLE 10 – Extrait du début des variables après *one-hot encoding* des variables Season et Climate.

Rééchantillonnage

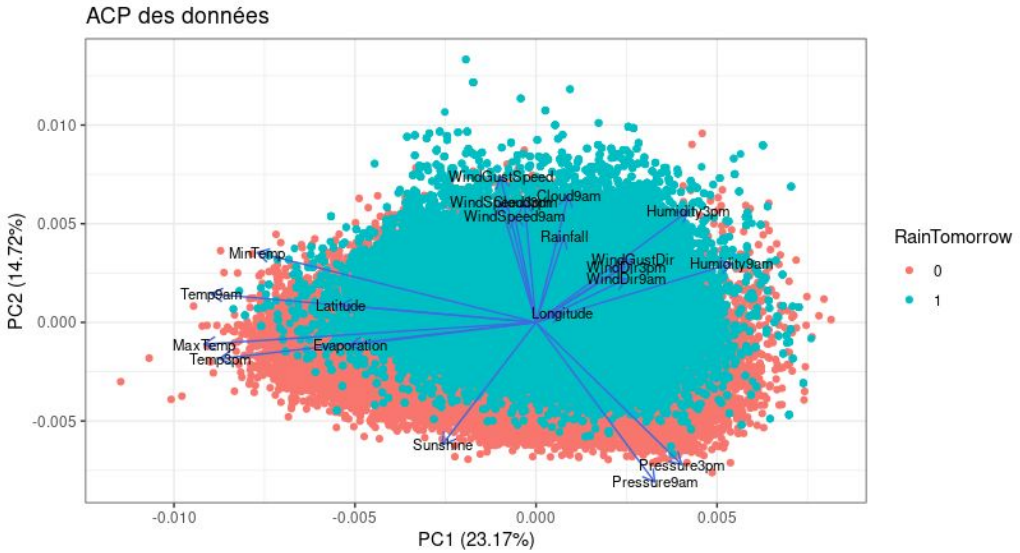
Down Sample : trop de pertes.

Up Sample : trop de redondances.

SMOTE : données synthétiques, mais seulement des variables numériques (one-hot encoding).

RainTomorrow	Compte	%
0	82367	54.22
1	69537	45.78

TABLE 11 – Distribution des valeurs de la variable RainTomorrow après SMOTE.



ACP après avoir fait upSample

Régression logistique

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.538e+01	1.790e+00	25.349	< 2e-16	***
MinTemp	1.567e-02	5.504e-03	2.847	0.00442	**
MaxTemp	-6.312e-02	5.926e-03	-10.652	< 2e-16	***
Rainfall	1.668e-02	1.584e-03	10.526	< 2e-16	***
Evaporation	2.418e-04	3.100e-03	0.078	0.93783	
Sunshine	-6.412e-02	3.735e-03	-17.168	< 2e-16	***
WindGustDir	1.754e-04	1.318e-04	1.331	0.18333	
WindGustSpeed	5.620e-02	1.162e-03	48.359	< 2e-16	***
WindDir9am	-3.727e-04	1.190e-04	-3.132	0.00174	**
WindDir3pm	-7.989e-05	1.329e-04	-0.601	0.54767	
WindSpeed9am	-1.354e-02	1.573e-03	-8.604	< 2e-16	***
WindSpeed3pm	-3.047e-02	1.639e-03	-18.597	< 2e-16	***
Humidity9am	9.532e-03	1.077e-03	8.852	< 2e-16	***
Humidity3pm	5.313e-02	1.023e-03	51.915	< 2e-16	***
Pressure9am	1.019e-01	5.925e-03	17.192	< 2e-16	***
Pressure3pm	-1.517e-01	5.860e-03	-25.893	< 2e-16	***
Cloud9am	2.585e-03	4.836e-03	0.534	0.59301	
Cloud3pm	4.043e-02	5.098e-03	7.931	2.17e-15	***
Temp9am	5.863e-02	8.532e-03	6.872	6.31e-12	***
Temp3pm	3.384e-02	4.235e-03	7.989	1.36e-15	***
RainToday1	5.710e-01	2.791e-02	20.457	< 2e-16	***
Season2	3.692e-01	3.288e-02	11.226	< 2e-16	***
Season3	6.623e-01	4.272e-02	15.505	< 2e-16	***
Season4	4.318e-01	3.324e-02	12.992	< 2e-16	***
Latitude	6.860e-03	4.780e-03	1.435	0.15121	
Longitude	-1.338e-02	1.096e-03	-12.205	< 2e-16	***
Climate2	-1.288e-01	3.931e-02	-3.276	0.00105	**
Climate3	-3.819e-01	8.957e-02	-4.264	2.01e-05	***
Climate4	-1.001e+00	1.026e-01	-9.757	< 2e-16	***
Climate5	1.398e-01	6.631e-02	2.108	0.03502	*

Performances sur dataTest en fonction de la “version” de la base de données :

- Sans rééchantillonnage : 84%, MAIS ! Balanced Accuracy de 71% car Spécificité de 47%. Évidence du déséquilibre.
- Up sampling : 77% et Balanced de 77%.
- SMOTE : 78% et Balanced de 77%. On privilégie cette version pour la suite.

Variables ignorées lors de la sélection (AIC) :

- Variables superflues de Climate et Season
- Cloud9am
- Latitude

Balanced Accuracy = (Sensibilité + Spécificité) / 2

Autres modèles et comparaison

	Sensitivity	Specificity	Balanced Accuracy
glm	0.82	0.74	0.78
glm.r	0.82	0.74	0.78
lda	0.82	0.72	0.77
qda	0.77	0.78	0.77
tmax	0.94	0.90	0.92
topt	0.88	0.76	0.82
rf (maxnodes = 1000)	0.90	0.81	0.86
rf (maxnodes = 2000)	0.91	0.83	0.87
glm test	0.82	0.74	0.78
glm.r test	0.82	0.74	0.78
lda test	0.83	0.73	0.78
qda test	0.78	0.78	0.78
tmax test	0.87	0.81	0.84
topt test	0.88	0.75	0.81
rf (maxnodes = 1000) test	0.91	0.83	0.87
rf (maxnodes = 2000) test	0.91	0.83	0.87

Conclusion