Australia Weather Report

AYADI Ben-Mekki

BICAKCI Alain

LEVÊQUE Jonas

SIMON Samuel

Contents

[1. Rapport d’exploration, de data visualisation et de pre-processing des données 4](#_Toc136979495)

[1.1. Contexte 4](#_Toc136979496)

[1.2. Objectifs 7](#_Toc136979497)

[1.3. Compréhension et manipulation des données 7](#_Toc136979498)

[1.3.1. Cadre 7](#_Toc136979499)

[1.3.2. Pertinence 9](#_Toc136979500)

[1.3.2.1. Données manquantes 9](#_Toc136979501)

[1.3.2.2. Détection des outliers 10](#_Toc136979502)

[1.3.2.3. Matrice de corrélation 11](#_Toc136979503)

[1.3.2.4. Feature Engineering 12](#_Toc136979504)

[1.3.2.5. Principales caractéristiques de variables explicatives 13](#_Toc136979505)

[1.3.2.6. Principal Component Analysis 16](#_Toc136979506)

[1.3.2.7. Visualisations des données avec la variable cible 18](#_Toc136979507)

[2. Data pre-processing 22](#_Toc136979508)

[2.1. Advanced Feature engineering 22](#_Toc136979509)

[2.2. Suppression des variables inutiles 22](#_Toc136979510)

[2.3. Missing values 22](#_Toc136979511)

[2.4. Outlier treatment 23](#_Toc136979512)

[2.5. Encodage des variables catégorielles 23](#_Toc136979513)

[2.6. Vérification de l’utilité des variables sélectionnées 23](#_Toc136979514)

[3. Simple Modeling Techniques 23](#_Toc136979515)

[3.1. Régression Logistique 23](#_Toc136979516)

[3.2. Comparaison avec des modèles de machine learning plus complexes 26](#_Toc136979517)

[3.3. Optimisation des modèles 27](#_Toc136979518)

[3.4. Focale sur le Random Forest 30](#_Toc136979519)

[3.5. Optimisation Bayésienne 32](#_Toc136979520)

[3.6. Lazy Classifier 32](#_Toc136979521)

[3.7. Conclusion de la section optimisation des modèles 33](#_Toc136979522)

[4. Imbalanced data 34](#_Toc136979523)

[4.1. Traitement des données 34](#_Toc136979524)

[4.2. Analyse des données en suréchantillonnage 35](#_Toc136979525)

[4.3. Analyse des données en sous-échantillonnage 35](#_Toc136979526)

[4.4. Intégration de seuils de classification 36](#_Toc136979527)

[4.5. Balanced Random Forest Classifier 37](#_Toc136979528)

[4.6. Conclusion de la section imbalanced data 37](#_Toc136979529)

[5. Advanced Modeling Techniques XGBoost and Keras 38](#_Toc136979530)

[5.1. Modélisation de réseau de neurones avec Keras 38](#_Toc136979531)

[5.2. Modélisation XGBoost 39](#_Toc136979532)

[5.3. Conclusion des modèles deep learning 42](#_Toc136979533)

[6. Modélisation régionale 43](#_Toc136979534)

[6.1. Préparation des jeux de données 43](#_Toc136979535)

[6.2. Modélisation par Random Forest 44](#_Toc136979536)

[6.3. Courbes d’apprentissage 44](#_Toc136979537)

[6.4. Optimisation Bayésienne du Random Forest 45](#_Toc136979538)

[6.5. Conclusion des modèles régionaux 46](#_Toc136979539)

[7. Conclusions 46](#_Toc136979540)

[8. Références 46](#_Toc136979541)

# Rapport d’exploration, de data visualisation et de pre-processing des données

## Contexte

D’un point de vue métier, ce projet constitue un exemple de jeu de données que nous pouvons rencontrer ou qui à terme pourrait le devenir. En effet, chacun d’entre nous est confronté à des données de type big data, liées de manière plus ou moins éloignée à l’environnement. En ce sens, il était logique que nous sélectionnions ce projet en tant qu’exemple de modélisation de jeu de données en data science.

D’un point de vue technique et scientifique, ce projet propose des séries de données temporelles sur le climat en Australie. Le projet permet de manipuler des données de type séries temporelles qui à terme sont utilisées pour prédire le climat. Cependant, cela peut également être utilisé pour décrire un climat futur en se basant sur des modèles de prédictions du Groupe Intergouvernemental sur l’Evolution du Climat (GIEC/IPCC).

Ainsi, d’un point de vue scientifique, les données à exploiter rentrent directement dans un des challenges les plus importants de ce siècle : le changement climatique. Cela impacte également les systèmes économiques et notre relation aux écosystèmes. Les rapports successifs récents du GIEC indiquent notamment une augmentation des phénomènes climatiques extrêmes pour la région Australie-Asie. Cela se traduit par avec des inondations, des sécheresses, des feux et tempêtes de grande ampleur à des fréquences plus élevées que par le passé[[1]](#footnote-1).

Dans le cas de l’Australie, les feux de brousses dus à des sécheresses importantes ont été dévastateurs, avec des implications sur la couche d’ozone et sur l’augmentation de la température au niveau de la stratosphère[[2]](#footnote-2), ces deux facteurs étant impliqués dans le climat. Les conséquences économiques pour l’Australie sont importantes avec une baisse du PIB comprise entre 0.2 et 0.5 points pour la saison 2019-2020[[3]](#footnote-3). En termes de valeur, les dommages causés par ces incendies sont estimés à 5 milliards de dollars (AUS).

En conséquence, pouvoir prédire la pluie en Australie est un facteur majeur dans la propension au risque d’incendies.

Il existe plusieurs types de climats en Australie, impliquant différents types de pluie pouvant se former. En effet, la forme la plus importante et la plus dominante est **la pluie convective**. Celle-ci se retrouve dans les régions de hautes latitudes, avec des climats tropicaux et subtropicaux. Cette forme de pluie est généralement accompagnée d’orages.

**La pluie conventionnelle** est un type de pluie qui est affecté par les montagnes et les régions montagneuses ; car c'est la forme de pluie la plus dominante et elle dépend de la latitude (Collier, 2003). La formation des précipitations conventionnelles se produit lorsque l'air à la surface de la terre devient intense par la chaleur du soleil. L'air chaud est plus léger que l'air frais, il s'évapore donc de la surface de la terre et forme des nuages dans l'atmosphère. Plus les vapeurs d'eau s'élèvent, plus ces vapeurs se déplacent progressivement vers le haut vers la zone d'air convergent et forment des nuages épais et lourds. Les nuages lourds et instables montent encore et l'instabilité de ces nuages les oblige alors à retomber à la surface de la terre sous forme de pluie conventionnelle (Selase, Agyimpomaa, Selasi, & Hakii, Precipitation and Rainfall Types with Their Characteristic Features, 2015).

**La pluie orographique** est la forme de pluie formée par l'air humide que l'on peut généralement observer au-dessus des montagnes. L'air humide au-dessus des montagnes est évaporé ou soulevé vers le haut. Lorsque l'air humide est soulevé et atteint un certain niveau, il se refroidit ; les nuages orographiques se forment puis se condensent et forment la précipitation. La pluviométrie orographique est formée par les terres de latitude moyenne comme celle à grandes montagnes (Gray & Seed, 2006). La pluie orographique a de minuscules gouttes d'eau qui se condensent. Ces petites gouttes d'eau des nuages, puis ces petits nuages se rassemblent pour former de plus gros nuages. Ces nuages se transforment également en neige au bout d'un certain temps (Jr., 2012). Les précipitations orographiques sont observées sur les montagnes des latitudes moyennes avec un axe perpendiculaire à la direction du vent dominant. Ces directions provoquent des transitions abruptes des précipitations et pourraient être mieux observées avec deux chaînes adjacentes de la montagne pour faire circuler davantage l'air humide. Les plus stables et ceux-ci sont expérimentés principalement dans l'après-midi des étés avec des orages dynamiques. La formation discrète de précipitations orographiques est parfois également observée sur les petites montagnes (Roe, 2005). Les précipitations orographiques sont dues au soulèvement de masses d'air par le vent (Smith & Evans, 2007).

**La pluie cyclonique ou frontale** est le dernier et troisième type de pluie. Le cyclonique par son nom représente la tempête et se produit lorsque les masses d'air aux caractéristiques distinctes entrent en collision les unes avec les autres. La collision de l'air léger qui est chaud et de l'air froid qui est lourd se produit ; l'air froid favorise l'air chaud car il est plus léger à monter. L'air ascendant se refroidit en formant des vapeurs d'eau. Le processus de condensation initie et forme les nuages (Thatcher, Takayabu, Yokoyama, & Pu, 2012). La formation de nuages devient lourde lorsqu'ils rencontrent d'autres nuages et ces nuages lourds deviennent instables et retombent sur la terre sous forme de précipitations cycloniques. Les précipitations cycloniques sont courantes dans les zones tropicales avec 23% de degrés de latitude nord et sud de l'équateur avec la latitude de la zone tempérée de 66% degrés nord et sud. C'est la raison pour laquelle elle est également connue sous le nom de pluie frontale (Thatcher, Takayabu, Yokoyama et Pu, 2012). Les précipitations frontales/cycloniques ont une période et un moment spécifiques où elles sont plus dominantes et les précipitations sont plus rapides et concentrées. La pluie cyclonique/frontale peut avoir des précipitations prolongées qui pourraient être prolongées et garder le temps humide pendant des journées entières. C'est une distribution de type lognormal et sa distribution dépend de la zone de précipitation (Cheng & Qi, 2001).

**La répartition des formes de précipitations en Australie** :

* La pluie convective est la forme la plus courante, en particulier dans les régions tropicales et subtropicales du nord et de l'est du pays. Les orages convectifs sont fréquents pendant la saison des pluies (de novembre à avril) et peuvent être très intenses, avec de fortes pluies, des éclairs et des vents violents.
* La pluie conventionnelle est également assez courante, en particulier dans les régions de haute altitude et de montagne. Cette forme de précipitation se produit lorsque l'air chaud à la surface de la terre est soulevé par le soleil et forme des nuages dans l'atmosphère. Lorsque ces nuages deviennent assez lourds et instables, ils se déplacent vers le bas et tombent sous forme de pluie conventionnelle
* La pluie orographique est principalement observée dans les régions de montagne et de haute altitude, en particulier dans les États de Victoria et de Nouvelle-Galles du Sud. Cette forme de pluie se produit lorsque l'air humide est soulevé par les montagnes et se refroidit, formant des nuages qui se condensent et tombent sous forme de pluie.
* La pluie cyclonique ou frontale est principalement observée dans les régions côtières, en particulier dans le nord et l'est du pays. Cette forme de pluie se produit lorsque deux masses d'air de caractéristiques différentes entrent en collision, soulevant l'air chaud et formant des nuages qui se condensent et tombent sous forme de pluie. Les cyclones tropicaux sont également une source fréquente de pluie cyclonique en Australie, en particulier dans le nord du pays.

En résumé, la proportion de pluie reçue en Australie dépend de la région et de la période de l'année, mais la pluie convective est généralement la forme la plus courante de précipitation dans les régions tropicales et subtropicales, tandis que la pluie conventionnelle et orographique sont plus courantes dans les régions de montagne et de haute altitude, et la pluie cyclonique est principalement observée dans les régions côtières.

En Australie, le régime des précipitations est fortement lié aux saisons. Comparée aux autres masses continentales, l'Australie est l'une des plus arides. **Plus de 80 % de sa surface connaît une pluviométrie annuelle inférieure à 600 millimètres** En passant à un autre extrême, certaines côtes nord du Queensland ont des moyennes annuelles de plus de 4 000 mm, le record australien de 12 461 mm ayant été atteint en 2000 au sommet du Mount Bellenden Ker, qui possède aussi depuis 1986 le record mondial de précipitations en 8 jours avec 3 847 mm.

En Australie, les précipitations varient considérablement d'une région à l'autre en raison de facteurs tels que la distance de la côte, l'altitude et la proximité de zones montagneuses.

Cependant, en général, l'Australie est un pays à précipitations relativement faibles, avec une moyenne annuelle de pluie inférieure à 600 mm pour la plupart des régions.

**En général, les précipitations sont plus élevées dans les régions côtières et les zones montagneuses et plus faibles dans les régions intérieures et désertiques.**

La moyenne de la quantité de pluie varie en fonction de la saison, avec une moyenne plus élevée en été et en automne et une moyenne plus faible au printemps et en hiver. La pluie peut varier considérablement d'un mois à l'autre, même au sein de la même saison.

## Objectifs

Les principaux objectifs que nous souhaitons atteindre lors de ce projet sont les suivants :

* **Obtenir un jeu de données exploitables, permettant la création d’un ou plusieurs modèles de prédiction de pluie en Australie** ;
* **Utiliser ces données et ces modèles afin de prédire les précipitations sur les 10 prochaines années (à préciser au fur et à mesure du projet).**

Par rapport à la problématique proposée, voici l’expertise de chaque membre du groupe :

* AYADI Ben-Mekki : travaille à l’Institut de Radioprotection et de Sûreté Nucléaire, notamment dans le bureau de modélisation de dispersion de rejets accidentels de radionucléides dans l’atmosphère en lien avec les prévisions météorologiques fournies par Météo France (modèle ARPEGE).
* BICAKCI Alain : Ingénieur production à BPCE IT, j’occupe principalement le rôle de data analyst, pas d’activité liée à la météo. J’ai des connaissances de base sur les phénomènes météorologiques courants, prévision du temps, les fonds froids et chauds, les précipitations et les vents. Familier également sur les tendances à long terme, telles que les changements et les catastrophes climatiques.
* LEVÊQUE Jonas : Utilisation de données géolocalisées avec ArcGIS, sur la thématique de la qualité de l’eau aux US. Interactions avec des amis/collègues sur l’évaporation et les réserves d’eau dans le monde en 2014-2019, mais pas dans le cadre de ce projet.
* SIMON Samuel : travaille à l’ingénierie de maintenance RATP, en particulier, sur la déformation des voies pendant les périodes de canicule. Contacts fréquents avec la société Metigate qui a développé un modèle prédictif de la température du rail en fonction des conditions météo.

## Compréhension et manipulation des données

### Cadre

Le jeu de données que nous avons acquis provient de la plateforme gratuite KAGGLE[[4]](#footnote-4) contenant de multiples jeux de données. Le jeu de données sélectionné est WeatherAUS.csv. Il contient 145 460 lignes pour 23 colonnes (variables). Au total, 45 villes australiennes (points de relevés météorologiques) sont présentes dans le dataset. Plusieurs variables sont proposées afin de prédire l’occurrence de la pluie, sur une période de Janvier 2007 à Juin 2017 (voir figure suivante).



Figure 1 : Localisation des villes et types de climat

Les types de climat sont répertoriés à partir de la carte suivante :

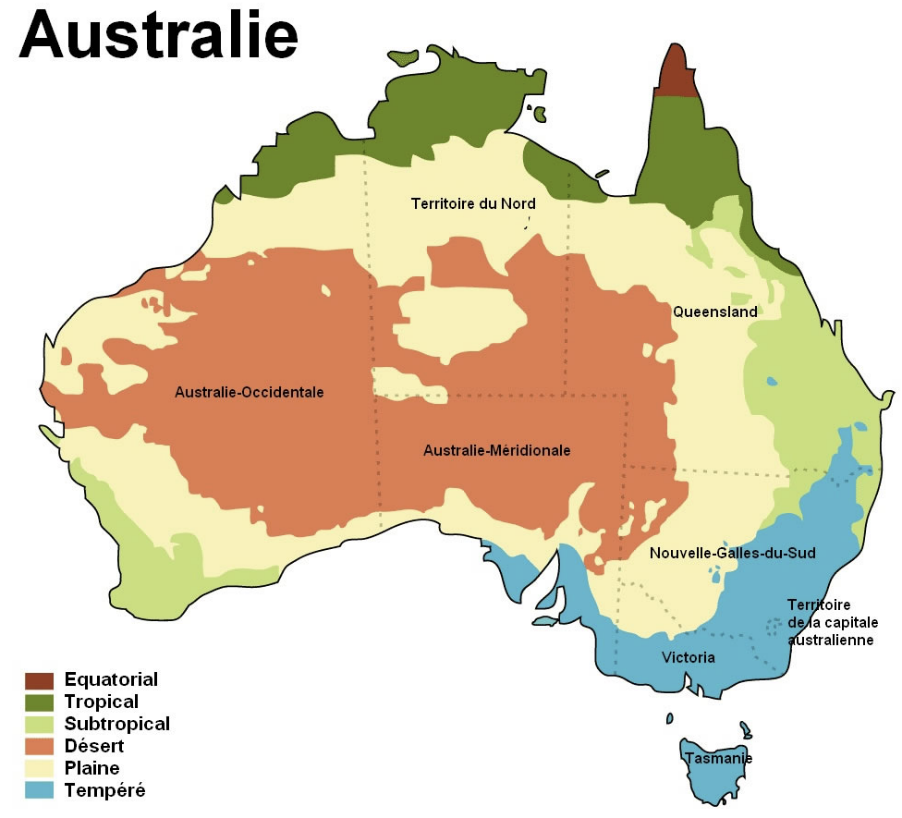


Figure 2 : Différents climats en Australie

Il est également possible d’utiliser la classification selon Koppen-Geiger sur les différents types de climats.

### Pertinence

Les variables qui semblent les plus pertinentes au regard des objectifs fixés sont les variables relatives à la pression, la température, la direction et la force du vent, la couverture nuageuse et l’ensoleillement. La variable cible est la présence de pluie le jour suivant les relevés météorologiques d’une journée.

La météorologie se base sur plusieurs paramètres essentiels de l’atmosphère (Dickinson *et al.*, 1996). Les prédictions des températures, de l’ensoleillement ou de la présence de pluie s’appuient sur ces paramètres. L’un des premiers facteurs de pluie est la pression atmosphérique. En effet, les masses d’air chaud et froid créent des différentiels de pression : plus la pression augmente, plus la température augmente. L’inverse est également vrai. Par ailleurs, du fait des marées, des différences de pression et de la terre en rotation, le vent est généré. Celui-ci est une résultante de ces interactions de masse d’air. Il est également une résultante de l’évaporation de l’eau au niveau de la mer. Le rayonnement du soleil, et le rayonnement infrarouge de la terre participent tous deux également à ces phénomènes et aux gradients de températures. La couverture nuageuse qui est la résultante de la vapeur d’eau stockée sous forme de gaz impliquant directement la présence de pluie lors de sa condensation. Cette couverture nuageuse est l’une des dimensions les plus importantes pour estimer l’apparition de pluie (Dickinson *et al.*, 1996).

Les différentes variables sont reprises dans le graphique suivant :

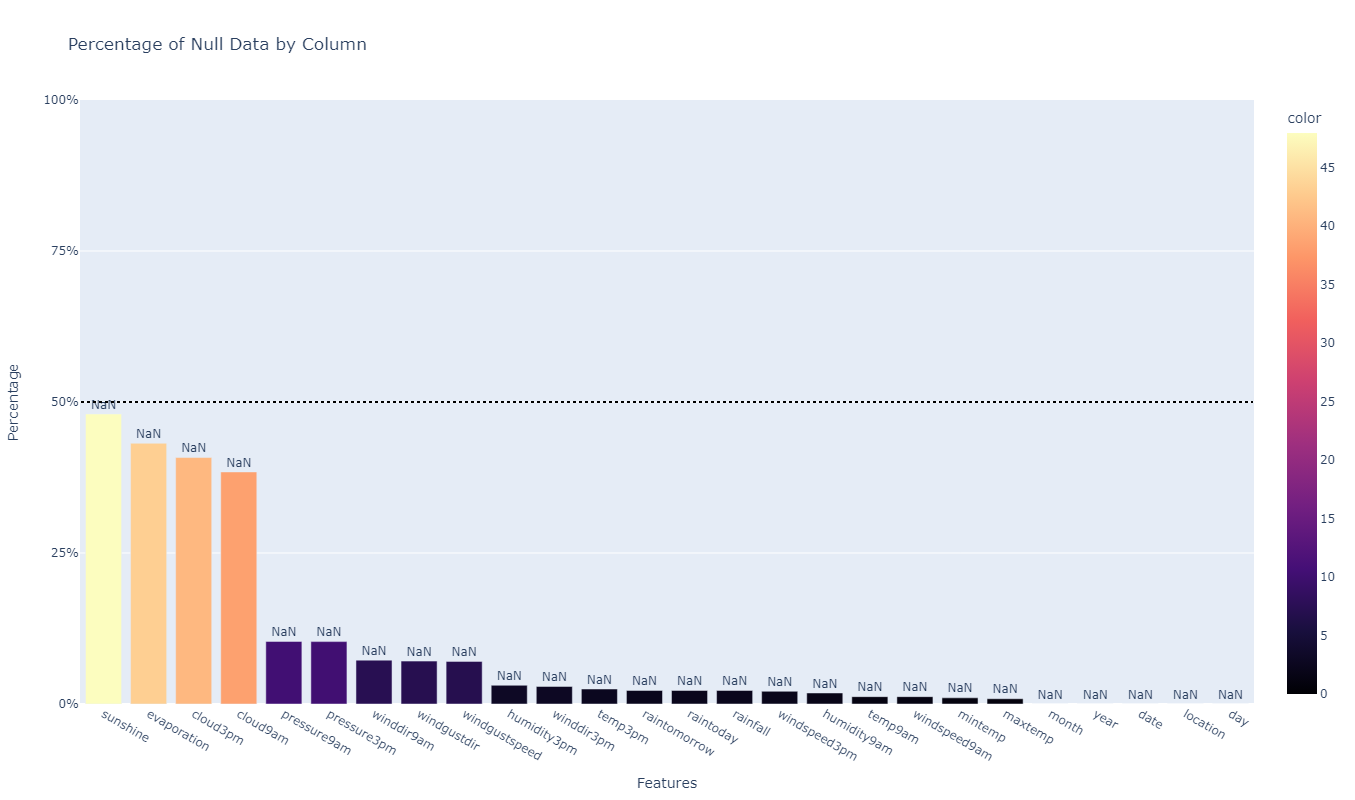


Figure 3 : Variables relatives au jeu de données et données manquantes

#### Données manquantes

Comme il est possible de l’observer dans le tableau précédent, le paramètre le plus problématique actuellement est le traitement des données manquantes. En effet, certaines variables possèdent de nombreuses données manquantes. Cela affecte les modèles que nous pouvons essayer de créer.

Solutions :

Afin de traiter le problème des données manquantes, nous avons étudié plusieurs choix possibles :

* La suppression des lignes pour lesquelles les données sont manquantes. Cela implique une baisse d’information de la base de données. Par exemple, certaines villes ont beaucoup de valeurs manquantes, ce qui peut générer une baisse de la qualité du modèle généré plus tard, et de mauvaises prédictions pour ces villes. En réalisant cette hypothèse, nous sommes arrivés à 56 420 observations retenues contre 145 460 observations avec des données manquantes, soit une réduction de 62 % des données du dataset ;
* Supprimer les variables explicatives qui ont plus de 40% de données manquantes. A ce titre, Evaporation, Sunshine, Cloud9am et Cloud3pm sont des variables que nous pourrions retirer du dataset. Nous obtenons 112 925 observations avec cette méthode. D’après la littérature, ces variables demeurent malgré tout importantes dans la prédiction de la pluie. Nous n’avons donc pas retenu cette méthode ;
* Remplacement des valeurs manquantes par la médiane et le mode. Cette méthode permet de conserver toutes les observations du dataset. Cependant, la médiane et le mode restent une information générale, là où des données peuvent être plus extrêmes ;
* Remplacement des valeurs par une valeur aléatoire. Plutôt que d’avoir une tendance, cette méthode a le mérité d’être objective. Malgré tout, cela génère des biais qui peuvent avoir des impacts importants dans la qualité du modèle final.
* Utiliser des valeurs proches pour remplacer les valeurs. En utilisant les villes localisées à 100 kilomètres les unes des autres, il peut être possible d’orienter le remplacement de valeurs manquantes.

Au vu de ces réflexions et des tests réalisés, nous avons souhaité mettre en place la dernière solution. L’objectif étant de pouvoir conserver un maximum de données, sans créer de biais trop important. Ainsi, nous avons cherché à éliminer les observations qui comportaient trop de valeurs manquantes sur plusieurs variables. Une étape de feature engineering a été préalablement réalisée (l’affectation de chaque ville à sa géolocalisation et à son type de climat (voir section sur le feature engineering)). Pour des valeurs manquantes sur une ville, nous avons ainsi remplacé une valeur par celle d’une observation connue pour le même jour sur une ville adjacente (inférieur à 100 kilomètres). Si nous n’avons pas pu remplacer une valeur, nous avons supprimé l’observation. Avec cette méthode, nous obtenons 114 474 observations, soit 78,7% des observations du jeu de données initial.

Le jeu de données que vous avons sur la variable cible est déséquilibré. En effet, pour 77% du temps, il ne pleut pas le lendemain. Cela signifie que les tests réalisés doivent être suffisamment robustes pour prendre en compte cette différence. En effet, cette différence peut notamment affecter la variance.

#### Détection des outliers

Après la résolution des données manquantes, nous avons traité la détection des outliers. En effet, la figure suivante montre les outliers des premières variables disponibles dans le dataset :

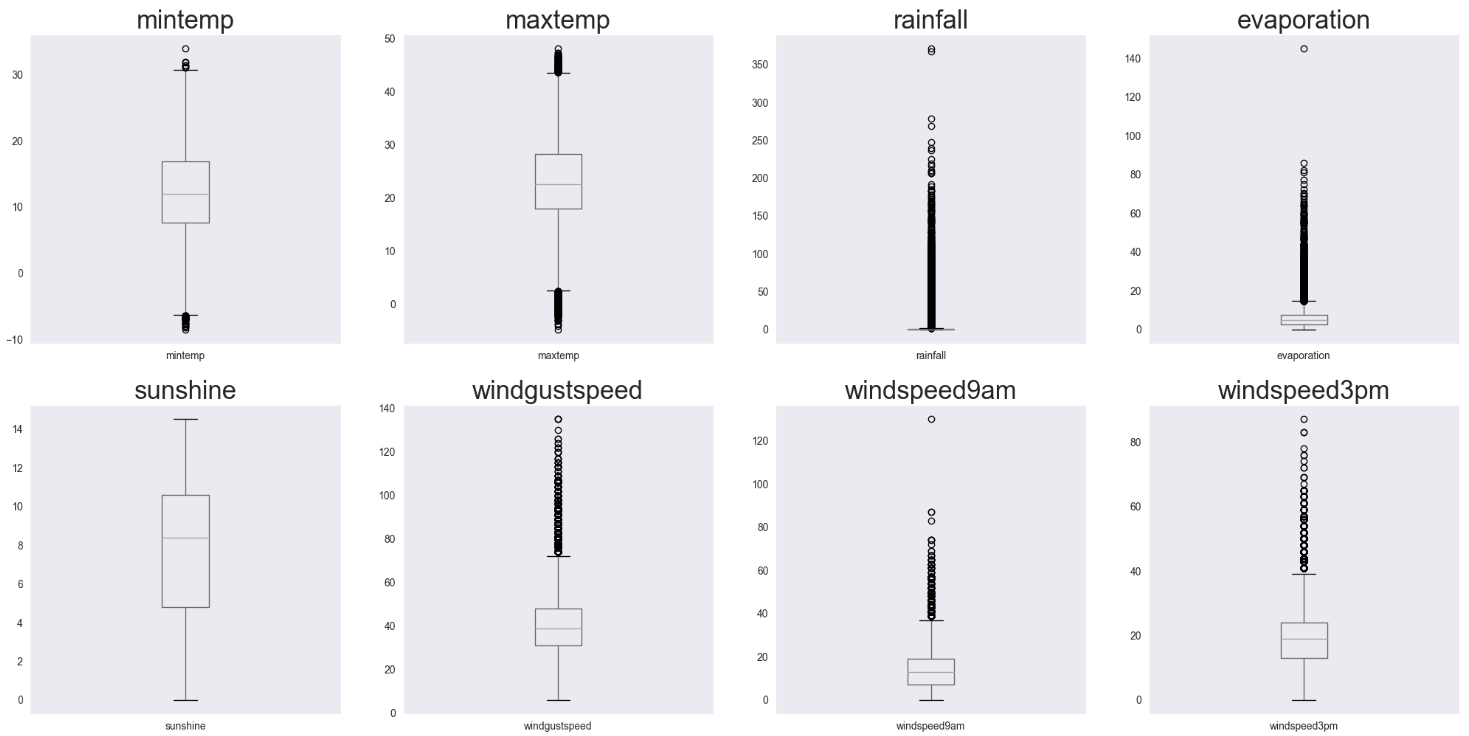


Figure 4 : Détection des outliers par boxplots

Bien que ces boxplots soient intéressants, les outliers présents reflètent notamment les différences entre les régions et les types de climats qui existent en Australie. En effet, les mesures d’évaporation ou de températures peuvent grandement fluctuer en fonction de la saison et de la localisation du point de mesure. Par ailleurs, Aguinis *et al.* définissent des recommandations pour identifier et manipuler les outliers (Aguinis *et al.*, 2013), puisque ceux-ci apparaissent à différents niveaux. Ces auteurs indiquent qu’il n’est pas toujours nécessaire de modifier les outliers et que ceux-ci sont parfois des informations réelles, à prendre en compte. Dans notre cas, en utilisant les percentiles à 99%, nous obtenons des valeurs normales. Les outliers sur rainfall par exemple proviennent d’une observation prise dans un lieu où la pluie tombe de manière importante (des records y sont atteints).

A partir de ces réflexions, nous faisons le choix d’inclure ces outliers dans le jeu de données.

#### Matrice de corrélation

Une première approche des données a été de comprendre les interactions entre les données et les divers liens qui existent entre elles. Pour cela, nous avons réalisé une matrice de corrélation, en utilisant le coefficient R de Pearson. Nous pouvons noter que certaines variables sont fortement corrélées tandis que d’autres non. La comparaison de la littérature par rapport à une matrice de corrélation est également importante. En effet, cela permet de considérer les variables explicatives qui se rapprochent. Cela est notamment essentiel dans de futures modélisations : il faut éviter la tautologie (redondance) des variables explicatives. D’après la littérature, la valeur absolue d’un coefficient de corrélation supérieur à 0.70 indique une corrélation importante (à confirmer avec des p-values sur des datasets avec un nombre d’observations faible, ce qui n’est pas notre cas ici). La figure suivante présente la matrice de corrélation entre toutes les variables :

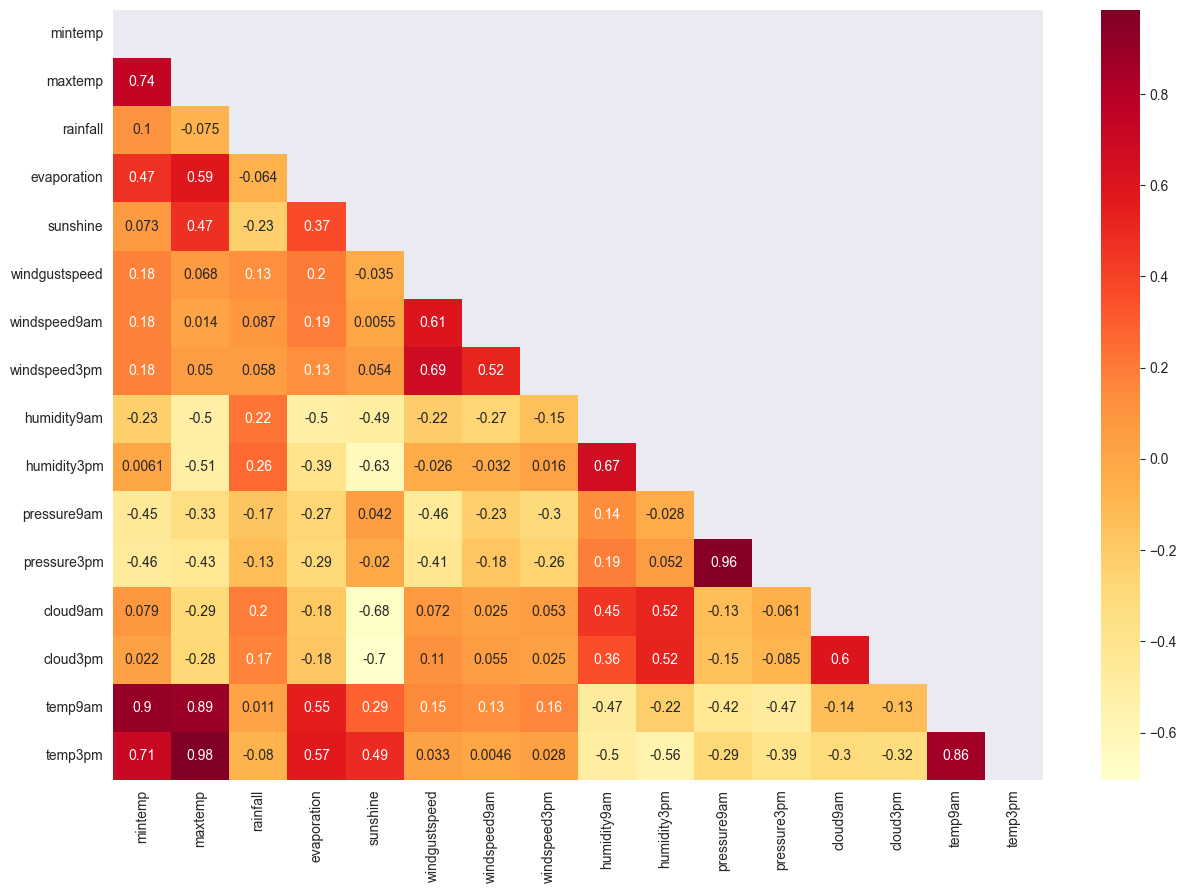


Figure 5 : Matrice de corrélation

Cette figure permet de visualiser les relations avec une heatmap : plus les cellules sont foncées, plus une corrélation positive entre deux variables est forte. Les cellules claires représentent une corrélation négative entre deux variables. Nous pouvons noter que les températures, l’ensoleillement et l’évaporation sont relativement positivement corrélés. La couverture nuageuse est positivement corrélée à l’humidité, et négativement corrélée à l’ensoleillement. La pression est négativement corrélée aux températures, à la vitesse du vent et à l’évaporation. L’évaporation est fortement positivement corrélée aux températures.

Ces interactions seront importantes à tenir compte lors de la modélisation.

#### Feature Engineering

Cette section est importante puisqu’elle traduit les différentes créations de variables secondaires à partir du jeu de données.

Ainsi, voici les variables que nous avons créé afin de faciliter la création d’un modèle et qui permettent d’enrichir un modèle :

* La temporalité. Nous avons créé des variables de temps par mois et par année (la donnée initiale étant par jour).
* La saisonnalité. Cette variable nous semble importante car des différences sont expliquées par la saison. L’Australie étant située dans l’Hémisphère Sud, les saisons sont inversées :
  + Eté : Décembre, Janvier, Février
  + Automne : Mars, Avril, Mai
  + Hiver : Juin, Juillet, Août
  + Printemps : Septembre, Octobre, Novembre
* Le climat. Nous avons utilisé le climat des localisations des villes/points de relevés afin de les regrouper. Nous avons défini quatre climats différents :
  + Tempéré
  + Subtropical
  + Tropical
  + Désertique
* La géographie du lieu. Les lieux précisent notamment l’altitude. Nous avons défini cinq géographies différentes :
  + Côtes
  + Montagnes
  + Désert
  + Plaines
  + Ile
* Différentiel de pression. La pression étant importante dans la littérature, il semblait adéquat de créer une variable de différence entre la pression maximale et minimale de la journée.

#### Principales caractéristiques de variables explicatives

Les figures présentées dans cette section permettent de visualiser les données en fonction des variables au cours du temps ou des saisons.

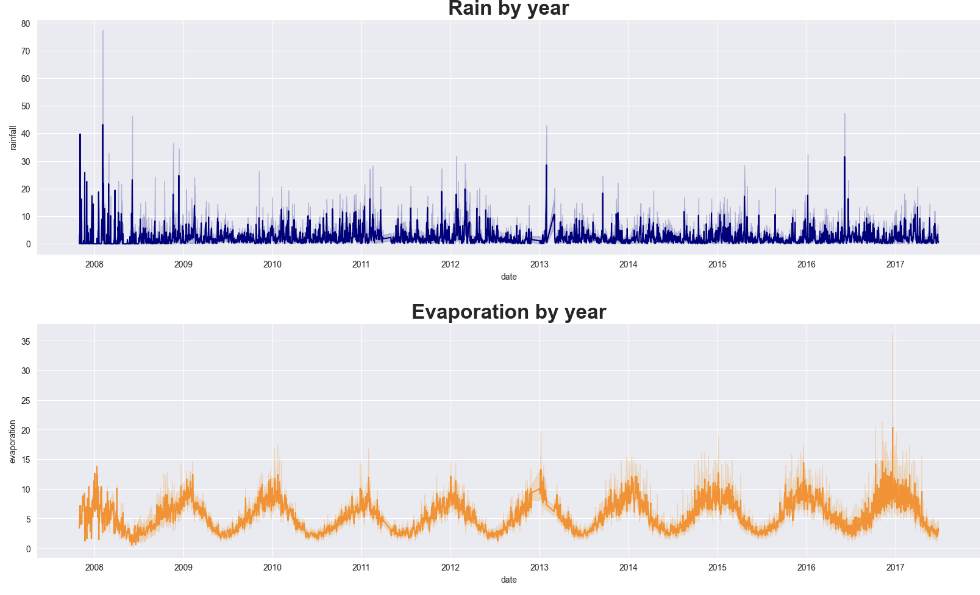


Figure 6 : Evolution de la pluie et de l’évaporation au cours du temps

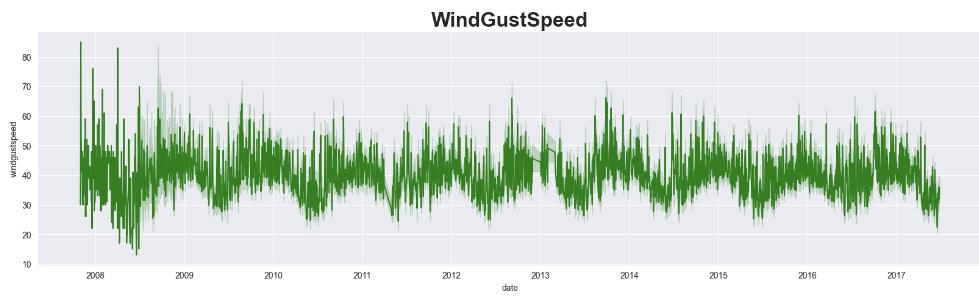


Figure 7 : Evolution du vent au cours du temps

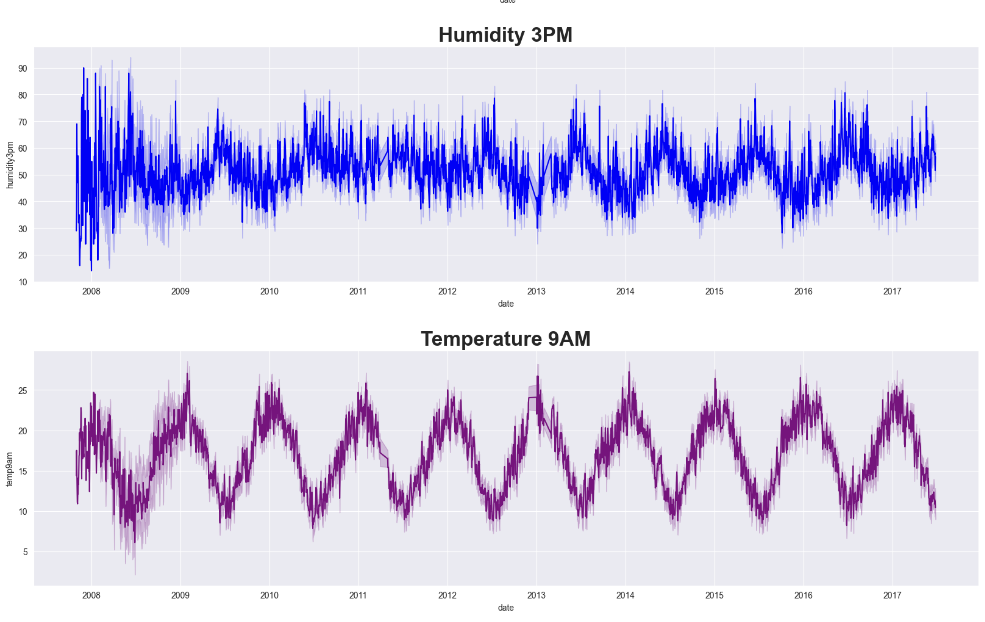


Figure 8 : Evolution de l’humidité (après-midi) et de la température (le matin) au cours du temps

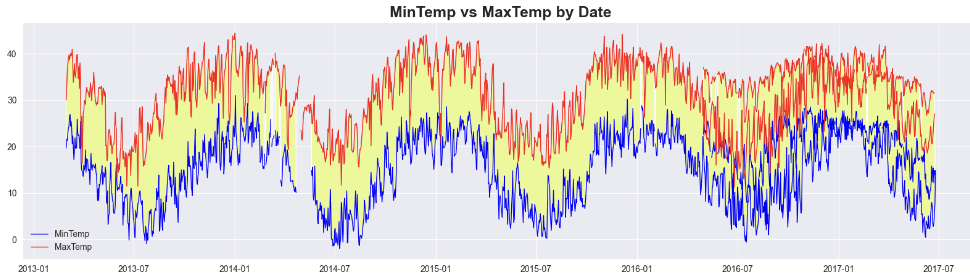


Figure 9 : Evolution de la différence de température entre les maximas et les minimas au cours du temps

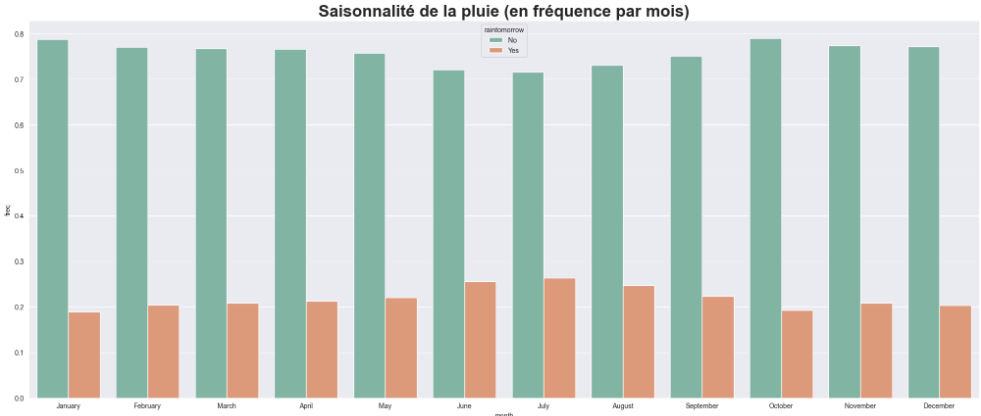


Figure 10 : Evolution de la pluie en fonction du mois

Cette figure montre notamment la fréquence de pluie globalement plus importante sur les mois de Mai-Août, qui correspondent à l’hiver Australien.

#### Principal Component Analysis

Afin de mieux comprendre la variance en fonction des variables du jeu de données. La figure suivante décrit les dimensions qui permettent d’expliquer la variance. Cette analyse ne permet pas d’identifier une variable dont la variance serait négligeable et ne participerait pas à expliquer une partie significative de la variance totale du système.

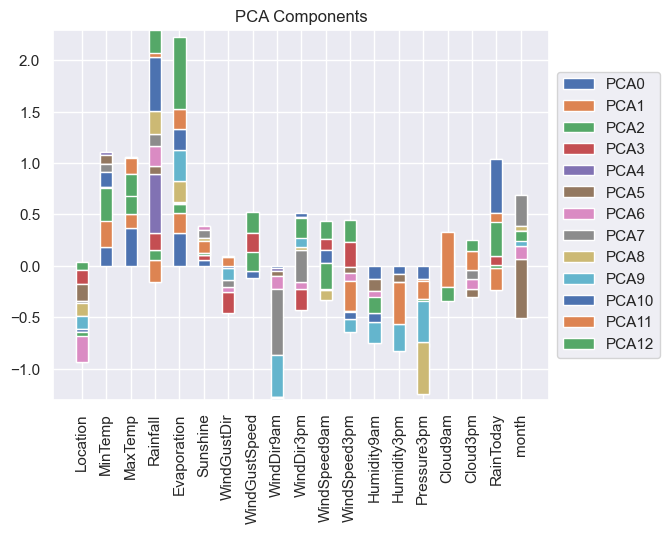


Figure 11 : PCA sur l’ensemble de variables du dataset

La figure suivante représente la répartition de la variable cible “RainTomorrow” selon les 2 composantes principales de l’analyse ACP (environ 40% de la variance expliquée). La répartition de la variable cible en 2 sous-domaines relativement distincts permet d’anticiper que le jeu de données est adapté à un problème de classification et devrait permettre d'entraîner efficacement les modèles.

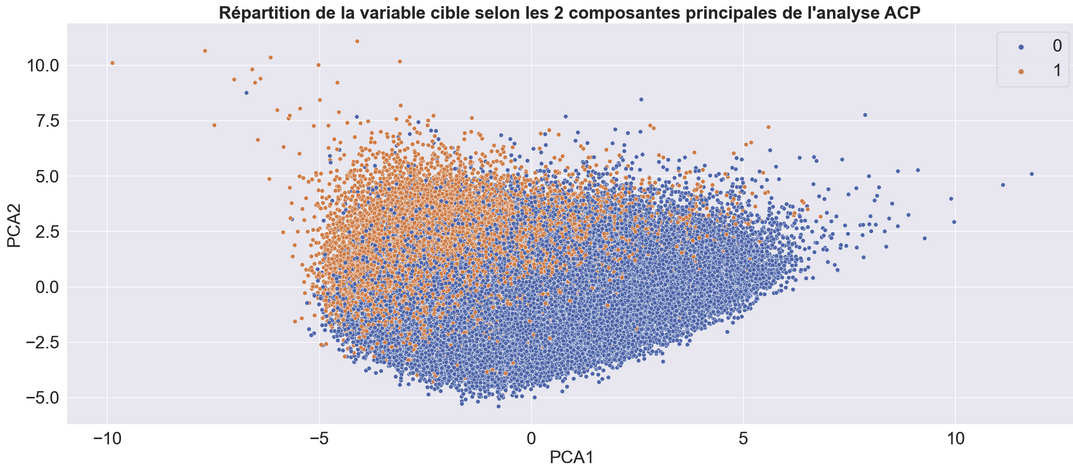


Figure 12 : PCA sur l’ensemble des variables du dataset

#### Visualisations des données avec la variable cible

Afin de mieux comprendre le jeu de données, nous avons réalisé des visualisations des variables explicatives en fonction de la variable cible.

Par exemple, la figure suivante propose la comparaison entre les minima et maxima de températures par rapport à la prévision de pluie.

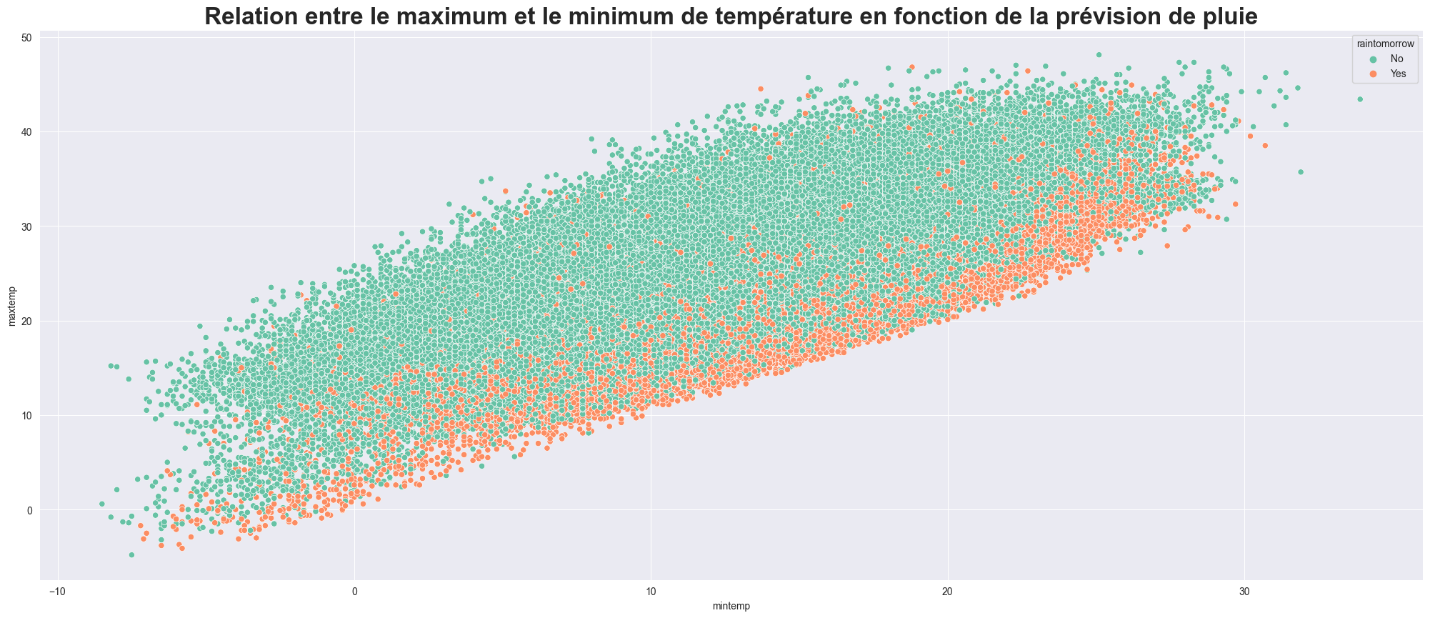


Figure 13 : Nuage de points entre les températures et la prévision de pluie

Il semble que la relation soit linéaire. On peut noter que les températures plus élevées ont tendance à avoir moins de pluie que les températures faibles.



Figure 14 : Evolution de la pluie en fonction de l’humidité

Plus l’humidité est importante la veille, plus la tendance de pluie le jour suivant est importante.

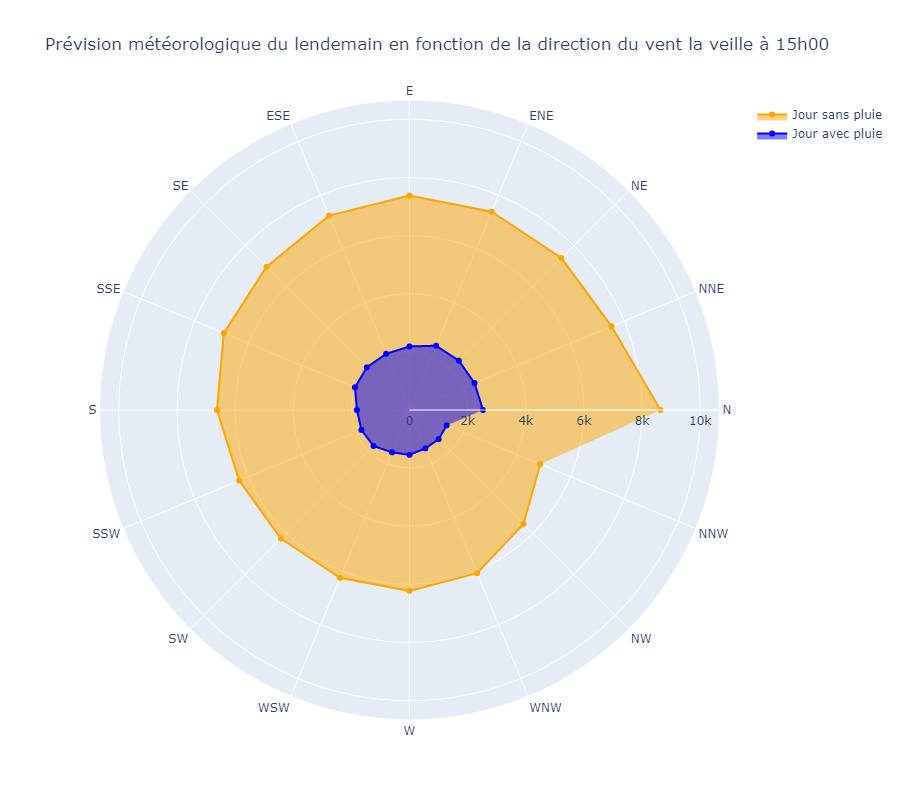


Figure 15 : Présence des jours avec pluie ou sans pluie en fonction de la direction du vent la veille à 15h

Les directions du vent provenant du Nord-Ouest apportent le moins de précipitations. (En même temps, moins de villes sont localisées sur la côte Nord-Ouest de l’Australie).

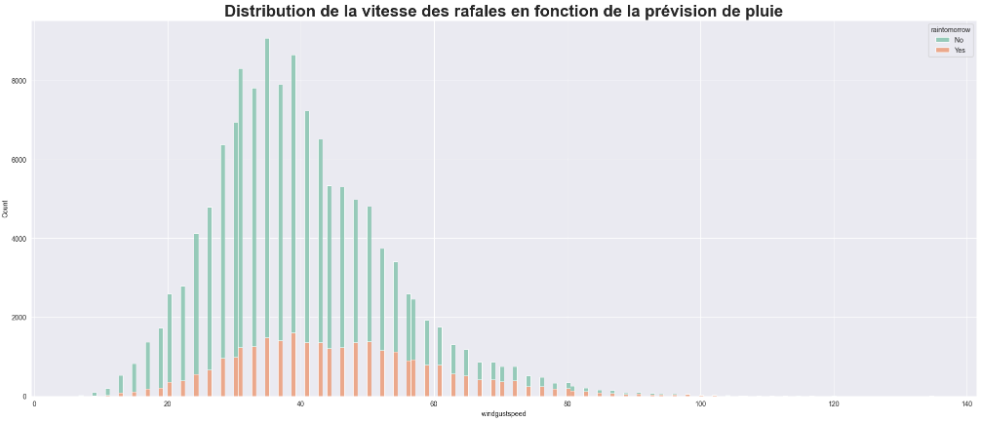


Figure 16 : Présence de pluie en fonction des rafales de vent

Il ne semble pas y avoir de lien entre la vitesse du vent et la présence de pluie.

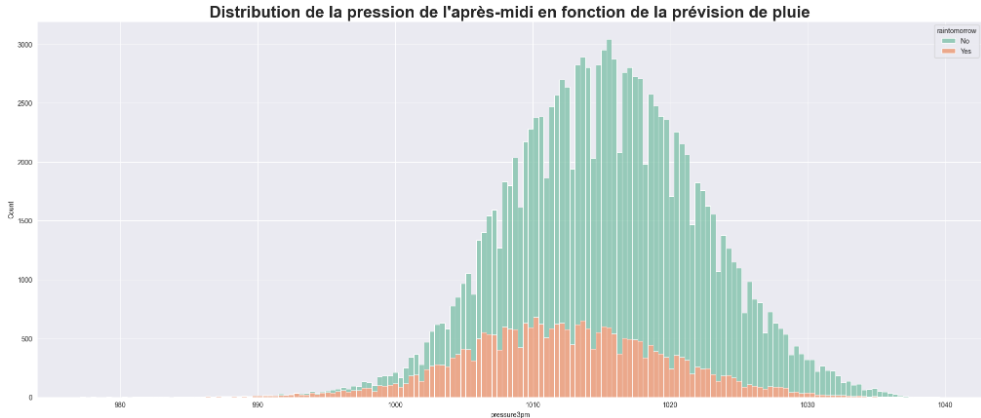


Figure 17 : Présence de pluie en fonction de la pression de l’après-midi de la veille

Cette figure indique que de faibles pressions ont tendance à apporter de la pluie.

# Data pre-processing

Après avoir nettoyé et exploré le jeu de données existant, nous avons réalisé les premières modélisations de données à partir de l’analyse exploratoire. Rapidement, nous avons été confrontés à des questionnements des valeurs manquantes et la possibilité d’améliorer certaines variables.

### Advanced Feature engineering

En réalisant nos premières modélisations, nous avons trouvé que les données climatiques exprimées en quatre classes différentes ne permettaient pas d’expliquer une variance intéressante. En cherchant dans la littérature, nous avons déterminé que l’usage de la classification du climat selon la méthode de Köppen était la plus approprié. Dans notre cas, nous avons déterminé quatre classes de climat :

* Chaud et humide ;
* Tempéré et froid ;
* Sec ;
* Méditerranéen.

Cette variable n’a finalement pas contribué à améliorer l’explication de la variance et a été retirée des variables explicatives.

Nous avons également décidé d’utiliser les points cardinaux, Nord, Sud, Est et Ouest pour déterminer la position géographique d’un point de données par rapport à la carte de l’Australie.

### Suppression des variables inutiles

Une première conséquence de l’analyse exploratoire est la multicolinéarité présente parmi les variables explicatives. Nous avons donc retiré les variables qui apportaient une information déjà renseignée par une autre afin d’avoir des variables explicatives expliquant une part individuelle de la variance. Les variables retenues ont été conservées dans un fichier csv « data\_features ».

Parmi les variables non retenues pour nos analyses, nous pouvons noter : 'date', 'location', 'maxtemp', 'mintemp', 'climat\_koppen', 'temp9am', 'temp3pm', 'pressure9am'. Celles-ci présentent des corrélations trop élevées.

### Missing values

Par ailleurs, nous avons fait le choix de traiter les valeurs manquantes des variables sunshine, evaporation, cloud3pm et cloud9am. Les villes étant proches les unes des autres, nous avons estimé qu’il était possible de pouvoir retrouver le climat d’une ville par proximité, dans un rayon de 500 km. Si une ville possède une donnée manquante, nous remplaçons cette valeur par celle de la ville la plus proche. Bien que cette méthode puisse intégrer un biais dans nos analyses, elle semble moins biaisée qu’un simple remplacement par le mode, la moyenne ou la médiane.

### Outlier treatment

Avec ce nouveau jeu de données, nous avons réalisé un traitement des valeurs aberrantes à l’aide du test de Tukey. L’objectif est de repérer les valeurs qui sortent des valeurs de référence selon l’interquartile range. L’option a été de retirer les valeurs aberrantes du jeu de données. Nous obtenons un jeu de données avec 72 707 entrées.

### Encodage des variables catégorielles

Afin de réaliser des modélisations adéquates, nous avons encodé les différentes variables catégorielles. Celles en « Yes », « No » remplacées par « 1 » et « 0 » respectivement. Les autres variables étant encodées selon une méthode de label encoding.

### Vérification de l’utilité des variables sélectionnées

Dans l’objectif de déterminer le choix des variables retenues, une simple régression logistique est lancée sur l’ensemble du jeu de données. Les variables proposent des p-values inférieures à .05, ce qui montre une influence significative sur la variance de la variable cible. Elles sont donc toutes importantes dans l’explication de la variance. Cette régression simple est réalisée avec le package statsmodel. Une régression utilisant le model OLS (Ordinary Least Square) est utilisé, pour indiquer une différence de catégorie entre la valeur 0 qui est un jour sans pluie et 1 qui représente un jour avec pluie (une quantité physique).

# Simple Modeling Techniques

L’objectif de ces premières modélisations était d’obtenir des résultats préliminaires avec des modèles simples. En effet, nous devions connaître les performances de modèles simples. Un modèle de type régression est facilement interprétable comparé à des modèles plus complexes (réseaux de neurones), mais perd en performance.

La variable cible ‘raintomorrow’ a été isolée et les autres variables conservées dans un jeu de données csv en tant que data-features.

Pour débuter, nous avons transformé les données en un jeu d’entrainement et un jeu de test. 80% du jeu de test compose le jeu d’entraînement contre 20% pour le jeu de test. Ce choix est motivé par le fait que ce soit une méthode standard en data science. Par ailleurs, une autre proportion devrait être calculée selon le modèle choisi et donc possiblement recalculée pour chaque test.

Par ailleurs, nous avons créé un jeu de données features dit PCA, qui sont basés sur l’ACP réalisée précédemment.

### Régression Logistique

Le premier modèle utilisé ici est une régression logistique (*Logit Regression*). Nous avons utilisé le jeu de données classique (sans PCA). La matrice de confusion est présentée dans la figure suivante.

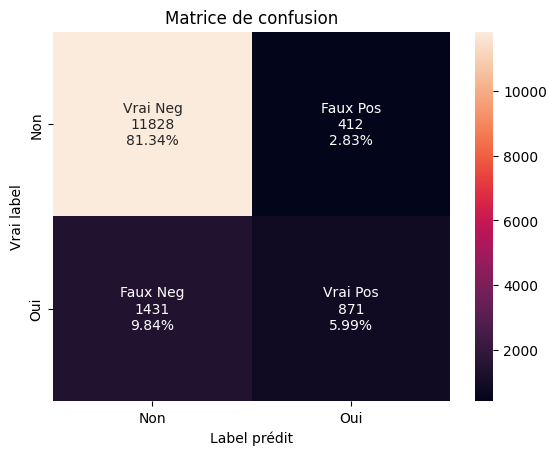


Figure 18 : Matrice de confusion de la Régression Logistique

D’après cette figure, il est possible de remarquer que la classe présentant les « vrais négatifs » est majoritaire. C’est-à-dire que le modèle arrive à prédire correctement dans 81.34% des cas qu’il ne va pas pleuvoir. A l’inverse, la classe des « faux négatifs » contient 9.84% des observations. Cela signifie que dans 9.84% des cas, le modèle prédit qu’il ne va pas pleuvoir, alors qu’il a plu en réalité. En agriculture, cela peut affecter une décision de récolte et un taux d’humidité trop important. Cette classe est relativement importante par rapport aux deux autres restantes alors que cette classe, et les faux positifs doivent avoir les scores les plus faibles. En effet, nous cherchons à connaître la classe des « vrais positifs » : que le modèle arrive à correctement prédire dans seulement 5.99% des cas. Le modèle réalise encore des erreurs en classifiant les « faux positifs » à hauteur de 2.83%. Le modèle prédit qu’il va pleuvoir alors qu’il n’a pas plu en réalité. Cela peut être impactant pour la gestion de l’irrigation en Australie, avec la ressource en eau.

Ce modèle « simple » de régression logistique indique que dans 87% des cas, le modèle classifie correctement les jours sans pluie (81.34%) et les jours avec pluie (5.99%).

La classe 0 (pas de pluie le lendemain) obtient un F1-score de 0.93, tandis que la classe 1 (pluie le lendemain) obtient un score de 0.49. Nous avons cherché à optimiser le score F1, ici avec l’attribut *weighted*, pour prendre en compte la différence importante entre les deux classes (la classe 0 étant majoritaire par rapport à la classe 1). Le score du F1-Weighted (score F1 pondéré) maximise donc la classe 0 malgré cet attribut intégré.

La précision indique le nombre de vrais positifs prédits par le modèle, divisé par le nombre total de positifs (vrais positifs et faux positifs). Ce score permet de répondre à la question : « quelle proportion d’identification positive était réellement correcte ? ».

Le rappel (recall) indique le nombre de vrais positifs divisé par les nombres de vrais positifs et de faux négatifs. Ce score permet de répondre à la question suivante : « quelle proportion de positifs ont été correctement identifiés ? ».

La métrique F1 est une moyenne arithmétique de la précision et du rappel.

En termes de résultats statistiques, le modèle précédent peut être traduit selon le tableau suivant qui représente les odd-ratios de chaque variable indicatrice :

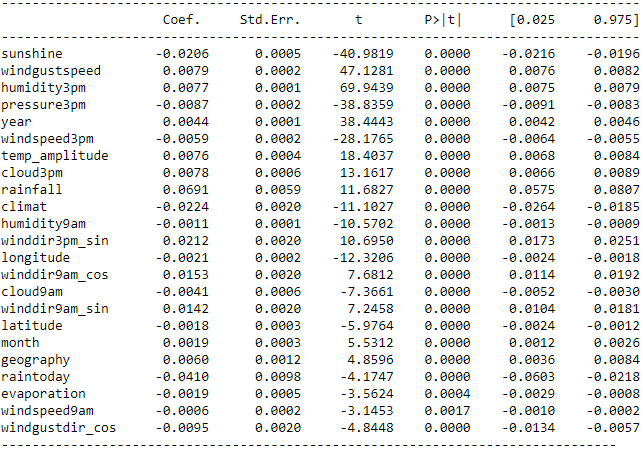


Figure 19 : Variables explicatives du modèle

Prises ensemble, les 23 variables prédisent la présence de pluie ou non le lendemain (c2 = -18 444, *df* = 23, *p* < 0.001). Les coefficients peuvent être interprétés de la manière suivante : pour une variable continue, si un coefficient est négatif, cela signifie qu’une augmentation d’une unité de la variable concernée diminue la probabilité de pluie pour le lendemain. En revanche, un coefficient positif indique que l’augmentation d’une unité de variable entraine l’augmentation de la probabilité de pluie le lendemain.

Par exemple, la probabilité (odd-ratio) qu’il pleuve diminue de 0.0206 lorsque la variable sunshine est augmentée d’une unité. A l’inverse, avec une augmentation d’une unité de windgustspeed, la probabilité (odd-ratio) qu’il pleuve augmente de 0.0079. La statistique « t » indique la force de la variable dans la détermination de la probabilité. On voit donc que humidity3pm représente la variable avec le plus de poids dans le modèle, tout comme sunshine et windgustspeed. Pour les variables catégorielles, le même schéma s’applique, en passant d’une unité à une autre.

### Comparaison avec des modèles de machine learning plus complexes

Pour espérer obtenir un meilleur score de prédiction de la classe 1, nous avons utilisé d’autres modèles de classification : le machine à support de vecteur (SVM), l’arbre de décision (decision tree), le modèle de forêt aléatoire (random forest), et le classificateur de plus proches voisins K (Knn).

A ce stade, nous avons également utilisé le jeu de données optimisé sur la PCA, pour comparer.

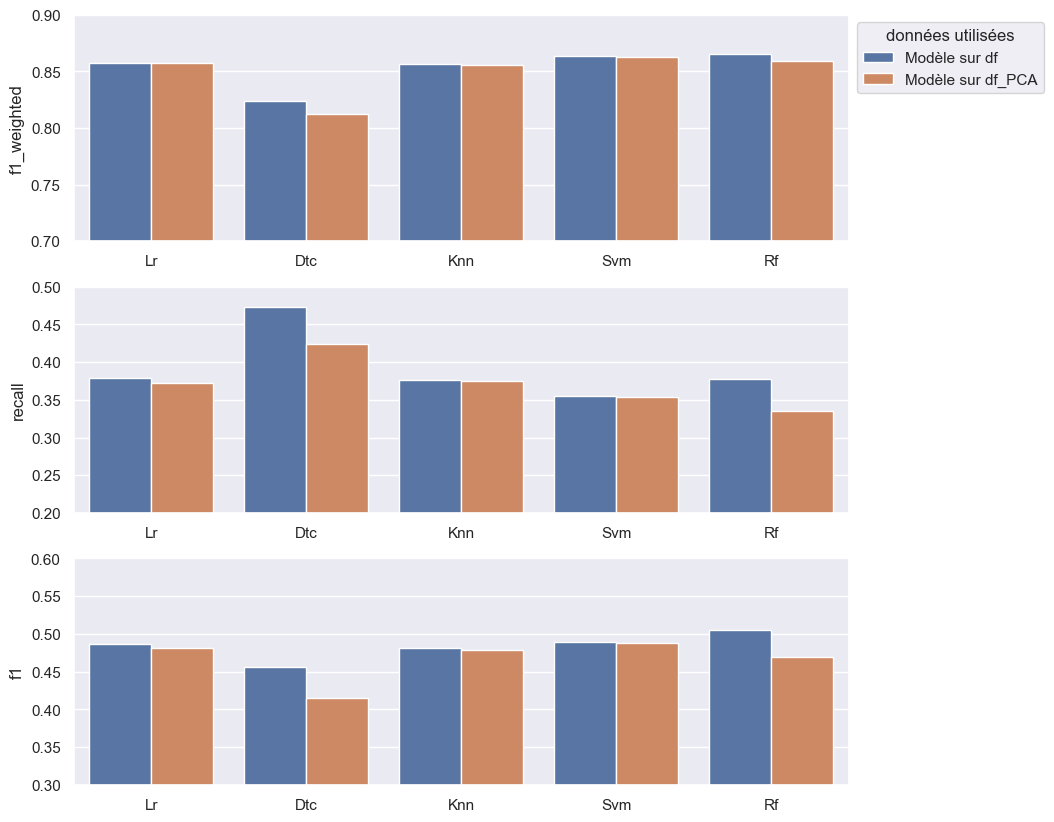


Figure 20 : Comparaison des différents modèles et des scores obtenus

La figure précédente intègre la comparaison des scores F1, F1-weighted et du rappel, comme métriques pour déterminer les modèles plus intéressants que la régression logistique.

Dans un premier temps, il est important de noter que l’optimisation en PCA ne permet pas d’obtenir des scores plus intéressants. Nous souhaitons tout de même explorer la modélisation avec ceux-ci.

En termes de comparaison de modèles, il semble que l’arbre de décision obtient un score limité sur le score F1 et F1-weighted. Il performe moins bien que le modèle de régression logistique. Le modèle Knn est également légèrement en-deçà des scores offerts par le modèle initial. Le modèle SVM est légèrement supérieur, mais le modèle de forêts aléatoires (RF) est le modèle qui ressort avec les meilleures métriques. Cela signifie que ce modèle classifie le mieux les données fournies.

### Optimisation des modèles

Afin d’optimiser les modèles, nous avons choisi de réaliser un réaliser un Grid Search avec la maximisation de la métrique « F1-weighted » qui nous semble la plus pertinente pour définir les modèles les plus optimaux dans la prédiction de la pluie le lendemain.

Chaque modèle cité précédemment a été optimisé selon les meilleurs paramètres, et pour les deux jeux de données : PCA et le jeu de données « classique ».

Tableau 1 : Comparaison des meilleurs scores et paramètres des modèles avec Grid Search

|  |  |  |  |
| --- | --- | --- | --- |
| Modèle | Caractéristiques | Df | Df\_PCA |
| LR | F1\_weighted (test) | 85.78% | 85.70% |
|  | Rappel | 37.83% | 37.27% |
|  | ROC score | 67.23% | 66.99% |
|  | Matrice de confusion | [[11828 ; 412]  [1431 ; 871]] | [[11837 ; 403]  [1444 ; 858]] |
|  | Best parameters | {'C': 100, 'class\_weight': None, 'penalty': 'l1', 'solver': 'liblinear'} | {'C': 100, 'class\_weight': None, 'penalty': 'l1', 'solver': 'saga'} |
| Decision Tree | F1\_weighted (test) | 84.48% | 83.13% |
|  | Rappel | 32.62% | 26.02% |
|  | ROC score | 64.55% | 61.44% |
|  | Matrice de confusion | [[11809 ; 431]  [1551 ; 751]] | [[11856 ; 384]  [1703 ; 599]] |
|  | Best parameters | {'class\_weight': None, 'criterion': 'entropy', 'max\_depth': 40, 'max\_features': None, 'max\_leaf\_nodes': 10, 'random\_state': 4, 'splitter': 'best'} | {'class\_weight': None, 'criterion': 'gini', 'max\_depth': 40, 'max\_features': None, 'max\_leaf\_nodes': 10, 'random\_state': 4, 'splitter': 'best'} |
| KNN | F1\_weighted (test) | 85.80% | 85.83% |
|  | Rappel | 35.58% | 36.05% |
|  | ROC score | 66.49% | 66.68% |
|  | Matrice de confusion | [[11922 ; 318]  [1483 ; 819]] | [[11909 ; 331]  [1472 ; 830]] |
|  | Best parameters | {'metric': 'euclidean', 'n\_neighbors': 10, 'weights': 'distance'} | {'metric': 'euclidean', 'n\_neighbors': 10, 'weights': 'distance'} |
| SVM (SVC) | F1\_weighted (test) | 86.99% | 86.85% |
|  | Rappel | 42.09% | 69.29% |
|  | ROC score | 69.55% | 41.61% |
|  | Matrice de confusion | [[11874 ; 366]  [1333 ; 969]] | [[11869 ; 371]  [1344 ; 958]] |
|  | Best parameters | {'C': 10.0, 'class\_weight': None, 'degree': 3, 'kernel': 'rbf'} | {'C': 10.0, 'class\_weight': None, 'degree': 3, 'kernel': 'rbf'} |
| Random Forest | F1\_weighted (test) | 86.66% | 85.85% |
|  | Rappel | 37.79% | 33.84% |
|  | ROC score | 67.84% | 65.95% |
|  | Matrice de confusion | [[11982 ; 258]  [1432 ; 870]] | [[12002 ; 238]  [ 1523 ; 779]] |
|  | Best parameters | {'class\_weight': None, 'criterion': 'entropy', 'max\_features': 'auto', 'n\_estimators': 2000, 'random\_state': 42} | {'class\_weight': None, 'criterion': 'gini', 'max\_features': 'auto', 'n\_estimators': 2000, 'random\_state': 42} |

Finalement, les modélisations supplémentaires avec le jeu de données en PCA ne nous permettent pas d’obtenir de différence notable sur les performances. Nous avons donc décidé de ne pas poursuivre avec ce jeu de données.

En utilisant le Grid Search et les meilleurs paramètres trouvés, nous avons créé les modèles optimisés. La figure suivante indique les comparaisons entre les scores des modèles optimisés et non optimisés (de base).

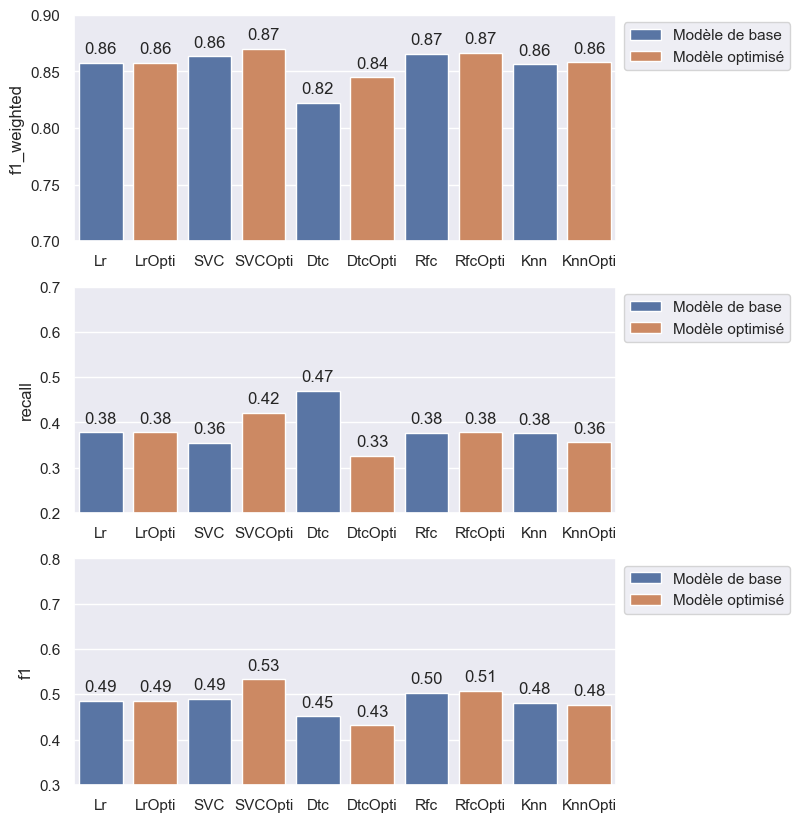


Figure 21 : Comparaison des performances des modèles optimisés ou non

Ce tableau confirme que les modèles de Random Forest et de SVM demeurent les plus précis dans la détermination de la pluie le lendemain, avec un score qui est plus intéressant. Toutefois, le modèle SVM conduit à des temps de calculs très longs (plusieurs heures). En ce sens, le Random Forest est plus court en temps de calcul et nous préférons conserver celui-ci.

### Focale sur le Random Forest

La figure suivante montre le détail des décisions qui sont prises lors de la création du modèle de Random Forest et de toutes les ramifications qui permettent de prédire la classe.

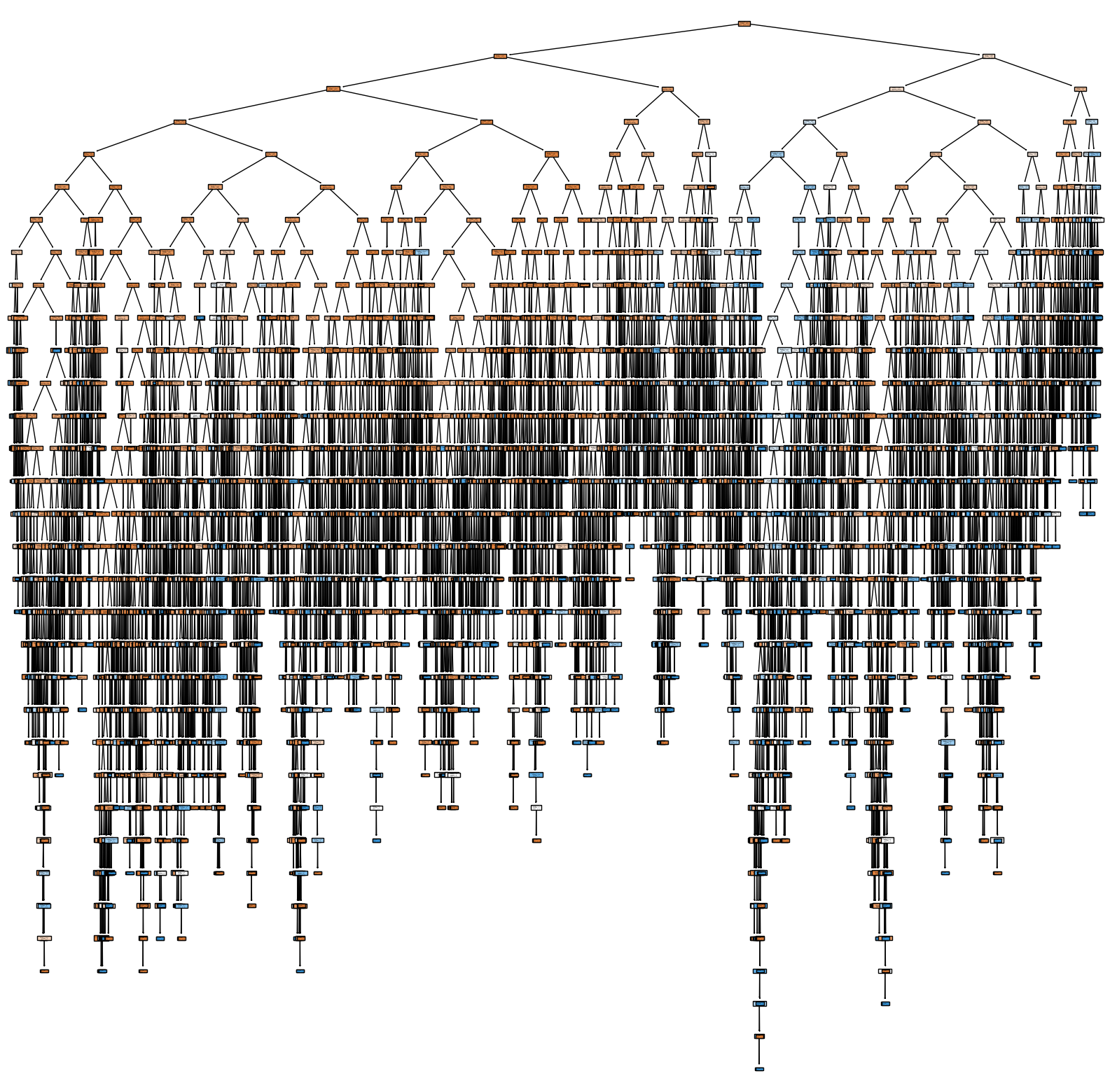


Figure 22 : Arbre de décision du Random Forest Classifier

En utilisant un modèle avec moins de dimensions, on peut obtenir la figure suivante :

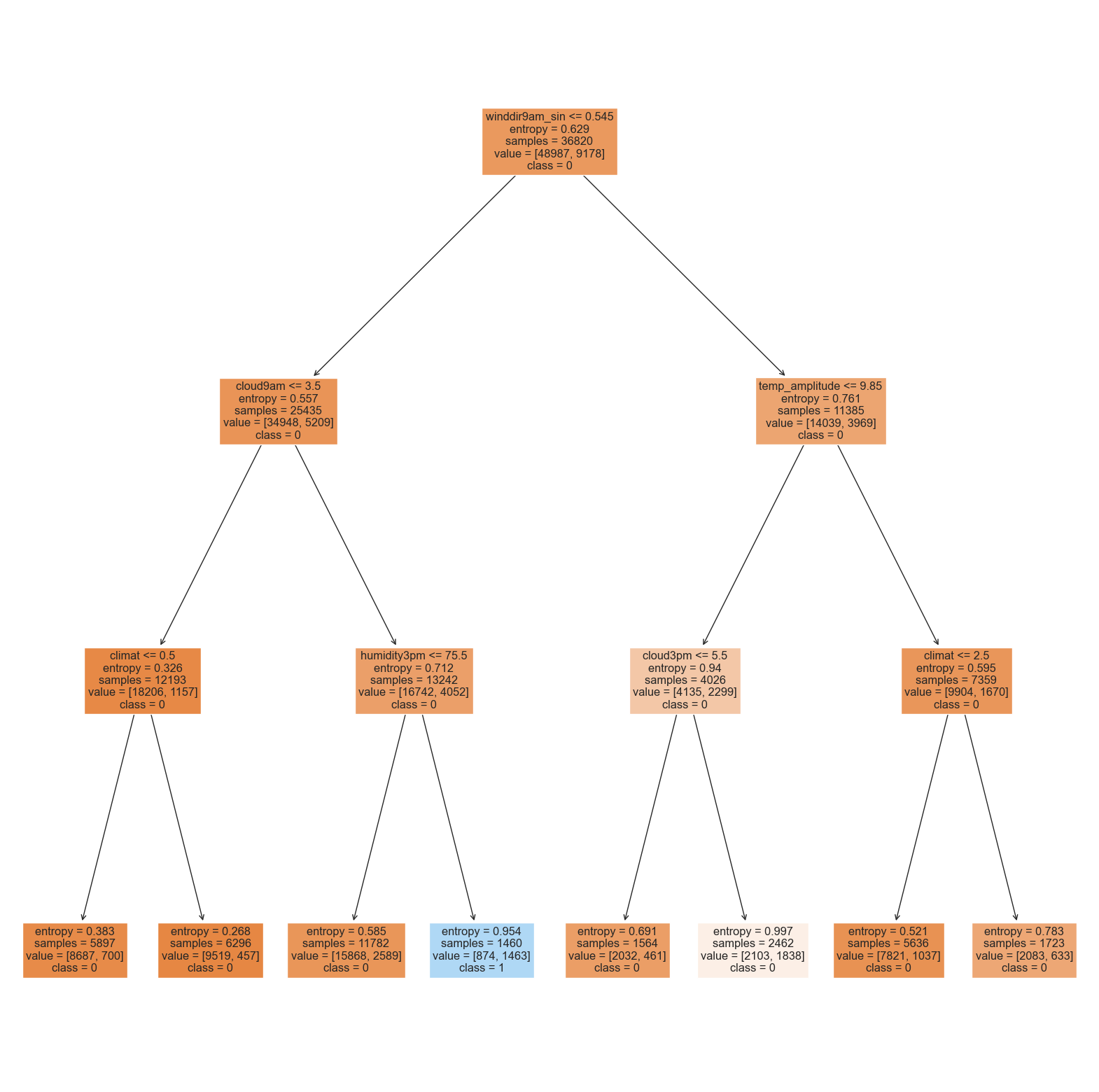


Figure 23 : Arbre de décision pour un modèle (réduit) de Random Forest

Il est possible d’interpréter un tel arbre en indiquant que la variable la plus importante du modèle est le sinus de la direction du vent à 9h, qui va déterminer en premier lieu la classification. En-dessous d’un score à 0.54, le modèle classifie la donnée pour vérifier la couverture nuageuse à 9h. Au fur et à mesure, le modèle choisit et dirige vers d’autres variables pour obtenir une classe finale. C’est bien la limite de la dimensionalité, car plus il y a de dimensions, plus il y a de possibilités et de ramifications. C’est arbre correspond au choix de la première itération : la première classification.

### Optimisation Bayésienne

Afin de comparer les optimisations, nous avons réalisé une optimisation bayésienne sur ce modèle. L'optimisation bayésienne est une méthode d'optimisation globale séquentielle qui cherche à minimiser une fonction d'objectif en construisant un modèle probabiliste de cette fonction. Le concept de base de l'optimisation bayésienne est d'utiliser la règle de Bayes pour construire un modèle probabiliste. Ce modèle est ensuite utilisé pour déterminer le prochain ensemble de paramètres à évaluer pour la fonction d'objectif. L'optimisation bayésienne est souvent utilisée pour l'optimisation des hyperparamètres, où la fonction d'objectif est la performance d'un modèle.

Par rapport au nombre d’itérations, c’est moins long qu’un Grid Search classique. C’est une recherche intelligente, qui permet de laisser l’algorithme définir lui-même les hyperparamètres les plus intéressants. C’est une étude de sensibilité de chaque paramètre.

Ici, l’optimisation bayésienne renvoie les hyperparamètres suivants pour le modèle Random Forest :

* n\_estimators : Le nombre d'arbres dans la forêt = 774 ;
* max\_depth : La profondeur maximale des arbres = 15 ;
* min\_samples\_split : Le nombre minimal d'échantillons requis pour scinder un nœud interne = 1 ;
* min\_samples\_leaf : Le nombre minimal d'échantillons requis pour être à un nœud feuille = 4 ;
* max\_features : Le nombre de caractéristiques à considérer lors de la recherche du meilleur split, exprimé en pourcentage = 79.82%.

Le score de l’optimisation atteint 87%. Le meilleur score (F1-weighted) obtenu avec les hyperparamètres définis ici est de 88%.

Pour un niveau de performance final équivalent à celui des hyperparamètres déterminés par Gridsearch, l'optimisation à travers une recherche bayesienne offre un confort intéressant pour le datascientist :

* La possibilité de tester un grand intervalle de valeurs pour chacun des hyperparamètres, ce qui évite de devoir trop préjuger du résultat (notamment de quels seront les hyperparamètres les plus influents).
* Un temps de calcul très raisonnable.

### Lazy Classifier

Nous avons également utilisé un Lazyclassifier afin de comparer les scores des différents modèles possibles utilisés en machine learning :

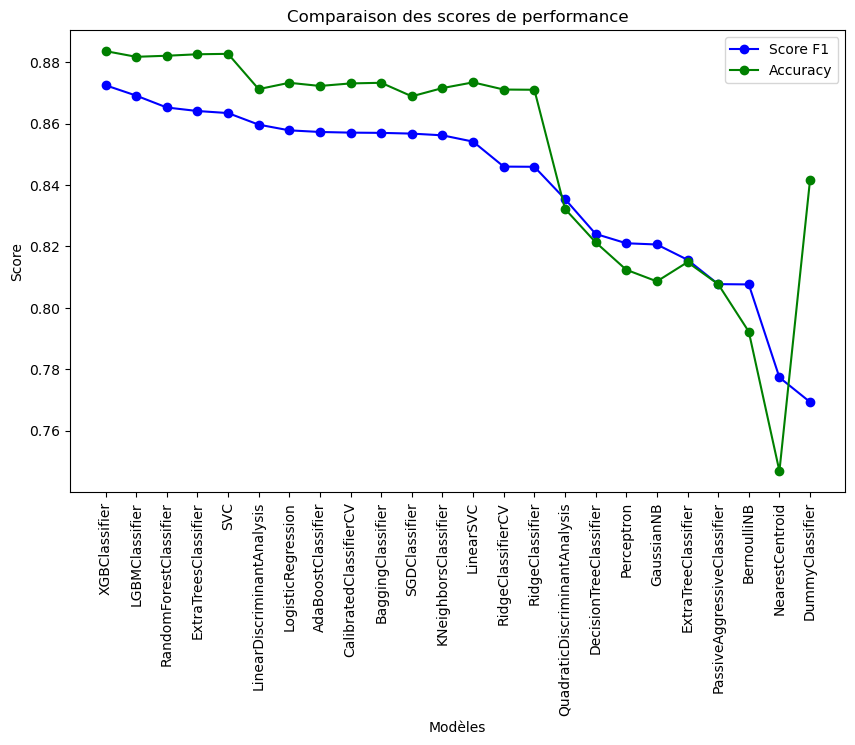


Figure 24 : Comparaison des scores (accuracy et F1) selon les différents modèles

En utilisant cette technique, il est possible d’observer que le XGBoost Classifieret le Random Forest Classifier sont les modèles avec les meilleurs scores (accuracy et F1). Nous conservons ces modèles pour la suite de notre exploration de modélisation.

### Conclusion de la section optimisation des modèles

Les modèles spécifiés ici indiquent des scores relativement importants dans la définition de la classe 0 : les jours sans pluie. Nous notons toutefois une redondance dans le manque de fiabilité sur la prédiction de la classe 1 : les jours avec pluie. Il faut rappeler que ces modèles sont basés sur l’ensemble de l’Australie et créent un modèle de généralisation des données. Il existe donc des variabilités importantes et notamment dans le cas de la prédiction des jours de pluie. L’utilisation de modèles et techniques prenant en compte le déséquilibre de données et des méthodes algorithmiques plus poussées vont être explorées dans les sections suivantes.

# Imbalanced data

L’utilisation d’un jeu de données déséquilibré créé un problème de taille pour la modélisation dans le sens où la prédiction de la classe minoritaire (ici, la classe 1 = pluie le lendemain) est difficilement prédite par les modèles. Plusieurs options de traitement des données permettent de passer outre et de rééquilibrer le jeu de données. Dans cette section, nous avons cherché à traiter les données de manière égale.

Il est à noter que dans ce cas, nous n’avons pas utilisé la métrique F1-weighted pour tenir compte de l’égalité des classes.

### Traitement des données

Pour réaliser ce traitement, nous avons utilisé les méthodes d’oversampling et d’undersampling. La première technique consiste à créer un jeu de données en simulant des données de la classe minoritaire jusqu’à atteindre un nombre égal d’observations entre la classe majoritaire et minoritaire (figure suivante).

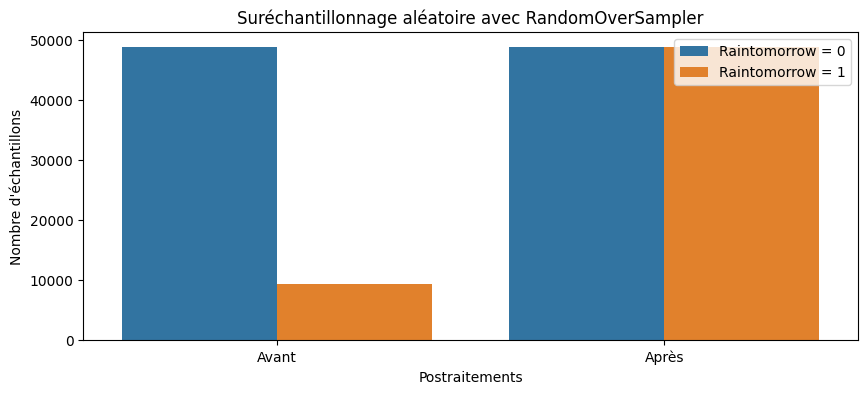


Figure 25 : Suréchantillonnage de la classe minoritaire

La deuxième technique consiste à l’inverse. C’est-à-dire à une sélection aléatoire des données de la classe majoritaire pour faire correspondre le nombre d’observations à la classe minoritaire (figure suivante).

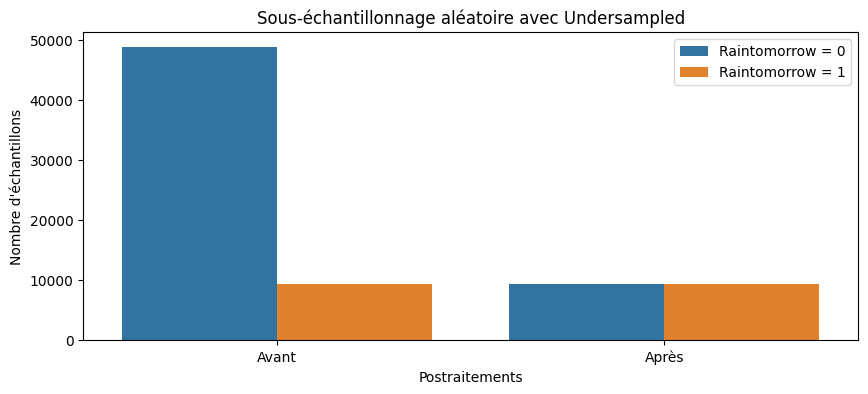


Figure 26 : Sous-échantillonnage de la classe majoritaire

### Analyse des données en suréchantillonnage

Pour tester les deux hypothèses, nous avons fait appel à une méthode de SVM.

En suréchantillonnage, en analysant le rapport de classification, nous pouvons conclure que le modèle SVM entraîné sur les données d'entraînement déséquilibrées a une performance moyenne sur les données de test déséquilibrées. La métrique de rappel (recall) est relativement élevée pour la classe minoritaire (classe 1), ce qui indique que le modèle est relativement bon pour identifier les vrais positifs. Cependant, la précision (precision) est relativement faible pour la même classe, ce qui signifie que le modèle peut également produire un grand nombre de faux positifs. La métrique F1-score est une mesure harmonique entre la précision et le rappel et est donc une mesure plus équilibrée de la performance globale du modèle. Le score F1 est d'environ 0,5 pour la classe 1 minoritaire, cela indique que le modèle peut être amélioré pour mieux l'identifier.

### Analyse des données en sous-échantillonnage

Dans le cas du sous-échantillonnage, le modèle SVM a des performances relativement bonnes pour prédire la classe "0" (pas de pluie demain), mais des performances plus faibles pour prédire la classe "1" (pluie demain). Le modèle a correctement prédit la classe "0" pour la grande majorité des échantillons (9230 sur 12240, soit une précision de 0.93), mais il a eu plus de difficulté à prédire la classe "1" (seulement 1647 prédictions correctes sur 2302, soit une précision de 0.35). Le score F1 est d'environ 0,77 pour la classe 1 minoritaire, cela indique que le modèle est plus performant avec un sous-échantillonnage des données.

Cependant, ce modèle implique la perte d’observations et donc de données réelles par rapport à la classe majoritaire. En ce sens, nous avons cherché à obtenir de meilleurs résultats avec d’autres méthodes d’optimisation.

### Intégration de seuils de classification

L’une des optimisations possibles est d’intégrer des seuils permettant de faciliter la classification. Ainsi, en reprenant un modèle SVM, il est possible d’indiquer un seuil au-delà duquel la classe 1 est identifiée. Pour un seuil établi à 0.40, nous obtenons les métriques suivantes :

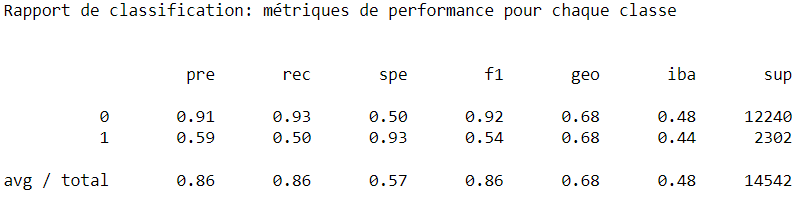


Figure 27 : Différentes métriques obtenues avec un seuil de .40

Le rapport de classification montre que le modèle a une précision élevée pour la classe 0 (pas de pluie demain), mais une précision beaucoup plus faible pour la classe 1 (pluie demain). Le rappel avec un rappel plus important pour la classe 0 que la classe 1. La spécificité est à l’inverse meilleure pour la classe 1. Le score F1 est plus élevé pour la classe 0 que pour la classe 1. La moyenne géométrique et l'exactitude équilibrée sont similaires pour les deux classes et indiquent une performance globale moyenne du modèle. L’ensemble de ces résultats suggère que le modèle a plus de difficulté à prédire les jours avec pluie que les jours sans pluie.

Nous avons également cherché à connaitre l’évolution des métriques en fonction de la valeur du seuil :

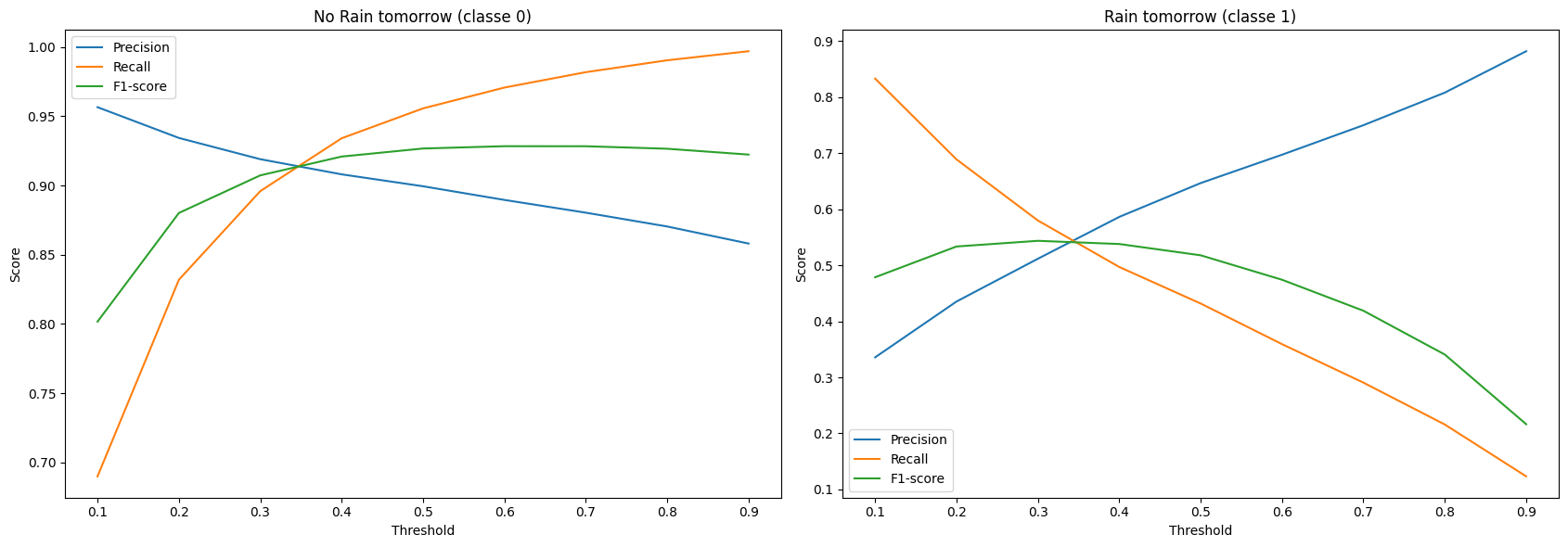


Figure 28 : Effet de la valeur du seuil sur les métriques pour la classe 0 (gauche) et la classe 1 (droite)

En ce qui concerne les performances maximales atteintes pour la classe 1, la précision maximale est de 0,81, le rappel maximal est de 0,50 et le F1-score maximal est de 0,62. Ces valeurs montrent que le modèle a des difficultés à prédire la classe 1 (pluie le lendemain).

En analysant les résultats pour la classe 1, on peut noter que la précision, le rappel et le F1-score augmentent tous avec la valeur du seuil jusqu'à un certain point, après quoi ils commencent à diminuer. Cela suggère qu'il existe un seuil optimal pour la classe 1 au-delà duquel le modèle commence à perdre en performances. Le rappel pour la classe 1 est assez faible, ce qui signifie que le modèle manque de nombreuses prédictions positives (la prédiction indique de la pluie et il pleut effectivement le lendemain).

Cette conclusion nous a conduit à rechercher et à tester un autre modèle.

### Balanced Random Forest Classifier

L’objectif du Balanced Random Forest Classifier est de pouvoir intégrer la différence de poids entre les classes dans l’algorithme. En performant ce modèle, nous obtenons les résultats suivants :

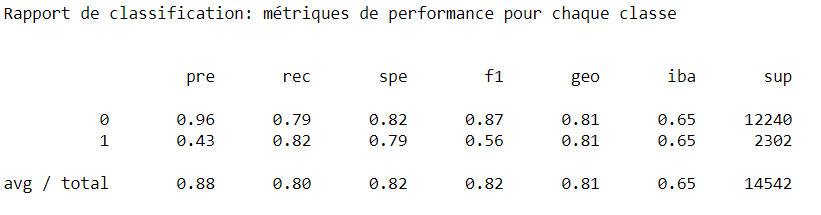


Figure 29 : Différentes métriques en utilisant un Balanced Random Forest Classifier

En ce qui concerne les performances maximales atteintes pour la classe 1, la précision maximale est de 0,43, le rappel maximal est de 0,82 et le F1-score maximal est de 0,56. Ces valeurs montrent que le modèle a des difficultés à prédire la pluie le lendemain (classe 1 = raintomorrow). D'après la matrice de confusion, le modèle a prédit correctement 9710 observations de la classe 0 (pas de pluie demain) et 1891 observations de la classe 1 (pluie demain), tandis que 2530 observations de la classe 0 ont été mal classées en classe 1 et 411 observations de la classe 1 ont été mal classées en classe 0. On voit donc ici des faux positifs qui sont créés par le modèle.

Le rapport de classification déséquilibré montre que le modèle a un rappel est élevé pour la classe 1 (80%), ce qui indique que le modèle a identifié la plupart des observations de la classe 1, mais le rappel est plus faible pour la classe 0 (80%), ce qui indique que le modèle a manqué certaines observations de la classe 0. Le score F1 est de 0,55 pour la classe 1, ce qui signifie que le modèle a un compromis entre la précision et le rappel pour cette classe.

En conclusion, le modèle Balanced Random Forest Classifier semble mieux performer que le modèle SVM dans la prédiction de la classe déséquilibrée. Cependant, le modèle a encore besoin d'être amélioré pour être plus performant pour la classe 1.

### Conclusion de la section imbalanced data

En analysant les différentes approches, nous pouvons conclure les points suivants :

En utilisant un jeu de données déséquilibré avec un modèle SVM, nous avons obtenu des performances moyennes sur les données de test déséquilibrées. Bien que le rappel pour la classe minoritaire fût relativement élevé, la précision était relativement faible, indiquant que le modèle pouvait produire un grand nombre de faux positifs. Cela montre l'importance d'utiliser des techniques de sur-échantillonnage et de sous-échantillonnage pour équilibrer les classes du jeu de données.

En utilisant la méthode de sur-échantillonnage aléatoire avec RandomOverSampler(), nous avons équilibré les classes et entraîné un modèle SVM avec des performances un peu améliorées. Le modèle a réussi à prédire la classe minoritaire avec une précision et un rappel plus raisonnables, avec un score F1 amélioré.

Cependant, en utilisant la méthode de sous-échantillonnage aléatoire avec RandomUnderSampler(), le modèle SVM a montré des performances relativement similaires à celles du modèle déséquilibré. Bien que le rappel pour la classe minoritaire ait été relativement élevé, la précision était relativement faible, indiquant que le modèle pouvait produire un grand nombre de faux positifs.

On rappelle, comme indiqué précédemment que le modèle BalancedRandomForestClassifier semble mieux performer que le modèle SVM dans la prédiction de la classe déséquilibrée.

Il existe donc plusieurs méthodes pour améliorer la performance de prédiction de la classe minoritaire tout en sacrifiant potentiellement la performance globale du modèle. Cette approche peut être appliquée à divers cas d'utilisation de prévisions météo. Cela pourrait être utile pour planifier des activités en plein air ou pour prendre des décisions liées aux opérations agricoles ou industrielles, par exemple.

Cependant, les modèles utilisés ici ont encore besoin d'être améliorés pour être plus performants dans la prédiction de la classe 1. C'est ce que nous proposons dans les sections suivantes.

# Advanced Modeling Techniques XGBoost and Keras

Les conclusions des sections précédentes nous invitent à explorer des modèles plus complexes pour prédire la classe 1 : les jours avec pluie le lendemain.

### Modélisation de réseau de neurones avec Keras

Dans un premier modèle, nous utilisons un modèle de réseau de neurones à l’aide de Keras. Le modèle est de type séquentiel et contient plusieurs couches denses (Fully Connected) avec des fonctions d'activation ReLU et des couches Dropout pour la régularisation. La couche de sortie utilise une activation sigmoïde pour la classification binaire. La compilation du modèle définit l'optimiseur, la fonction de perte et les métriques d'évaluation. L'optimiseur Adam est utilisé avec un taux d'apprentissage de 0.001. La fonction de perte est "binary\_crossentropy", adaptée à la classification binaire, et la métrique d'évaluation est l'accuracy (précision).

Nous avons choisi 100 epochs. Nous obtenons un score d’accuracy de 84%, ce qui n’est pas meilleur que les autres modèles plus simples. La fonction de perte est de 0.44. La validation de l’accuracy suit la même courbe que celle de l’accuracy, ce qui indique que notre modèle ne fait pas de surapprentissage. Ce modèle est en réalité décevant dans le sens où malgré la complexité du modèle, il ne parvient pas à obtenir un score tangiblement supérieur à celui du modèle simple de régression logistique.

### Modélisation XGBoost

XGBoost est particulièrement adapté aux problèmes de classification et offre de meilleures performances et une précision accrue par rapport aux autres algorithmes de Gradient Boosting.¶

XGBoost utilise une parallélisation pour entraîner les arbres, inclut des termes de régularisation dans sa fonction de coût pour éviter le surapprentissage (overfitting) et est capable de gérer des données manquantes en trouvant automatiquement la meilleure direction pour les observations manquantes.

En optimisant le model, nous obtenons la matrice de confusion suivante :

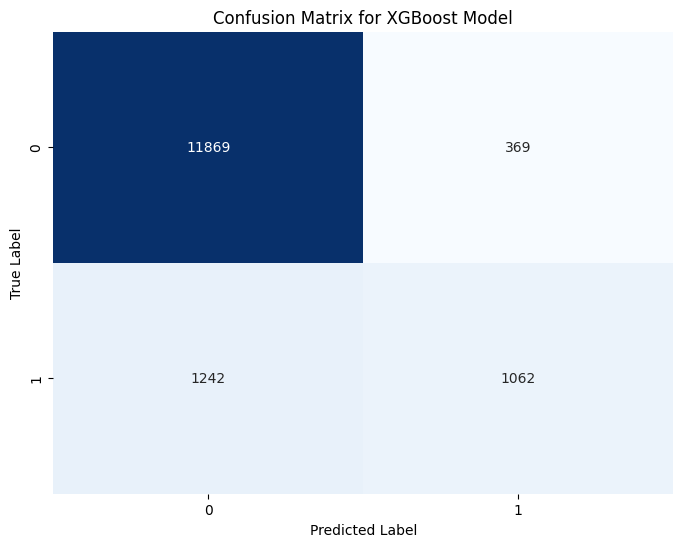


Figure 31 : Matrice de confusion pour le modèle XGBOOST

Nous obtenons ici un accuracy de 88.92%, ce qui est meilleur que les modèles précédents. Le F1-score pour la classe 1 est ici de 56.87%, ce qui est mieux que les autres modèles. Le rappel est de 46% pour la classe 1.

A l’aide des librairies shap et lime, nous pouvons faire les interprétations suivantes :

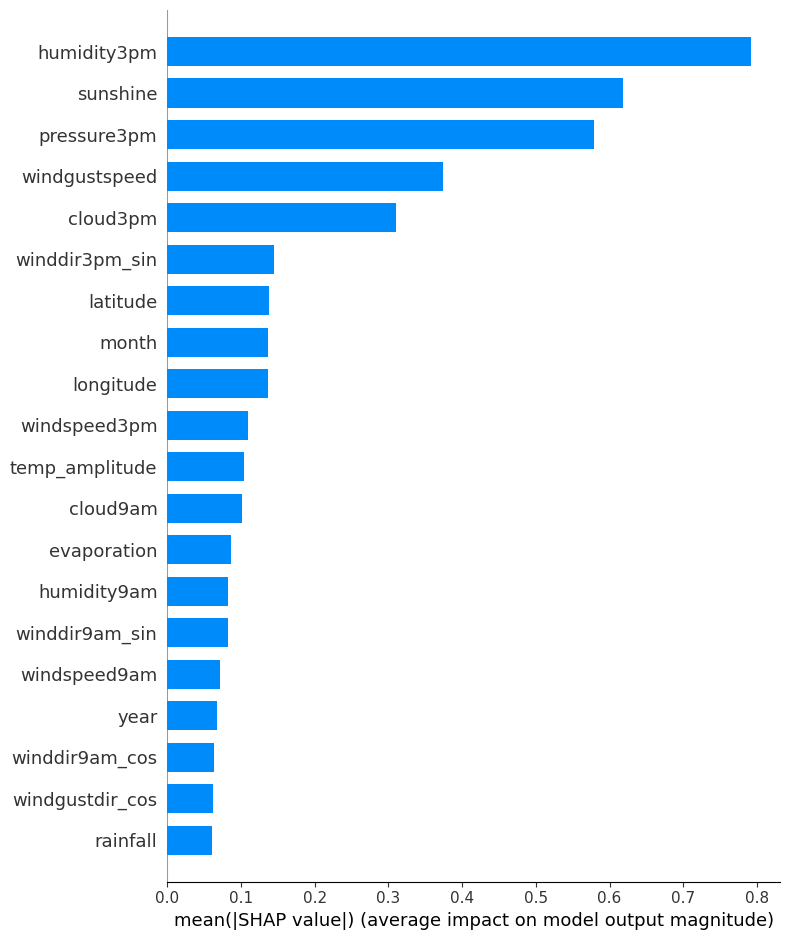


Figure 32 : Degré d’importance des variables dans l’explication du modèle

Cette figure nous indique que l’humidité à 15h est la variable la plus importante pour classifier les données. Les variables sunshine (degré d’ensoleillement), pressure 3pm (pression atmosphérique à 15h), windgustspeed (vitesse de pointe du vent) et cloud 3pm (couverture nuageuse à 15h) sont également importantes. La figure suivante montre également le sens de cette variation :

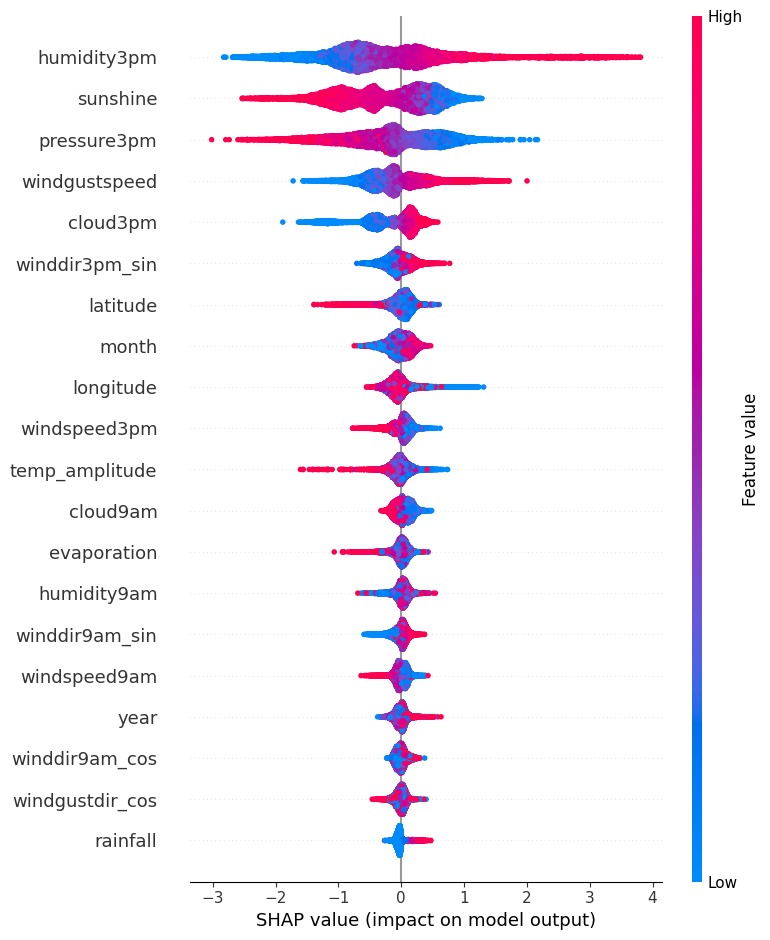


Figure 33 : Impact des variables explicatives sur le modèle de prédiction

Ici, une couleur rouge indique la propension à classifier une donnée dans la classe 1 : jour de pluie, tandis que la couleur rouge indique la classe 1. Par exemple, une valeur élevée pour la variable « humidité à 15h » nous indique que l’humidité est élevée, et que le modèle aura tendance à classer cette observation dans la classe 1. A l’inverse, un score faible dans cette variable (bleu) nous indique une humidité faible et que le modèle risque de classer notre observation dans classe 0. Nous pouvons interpréter chaque variable ainsi.

La librairie Lime permet aussi de comparer des observations uniques. Par exemple, ici sur les deux figures suivantes, il est possible de remarquer que l’observation 1 avait 95% de chance d’être classée dans la classe 0, et 74% pour l’observation 23. La libraire permet d’indiquer les valeurs critiques de chaque variable que le modèle a spécifié pour déterminer l’appartenance à une classe ou à une autre. La couleur orange spécifie les valeurs obtenues par les observations qui dépassent le seuil de validité d’appartenance à la classe 1. Les valeurs de chaque observation unique sont affichées dans la colonne de droite.

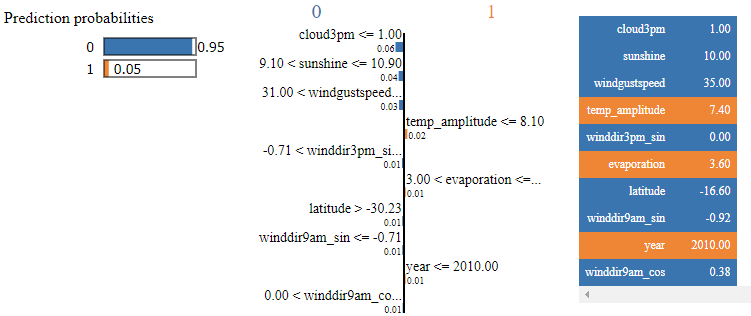


Figure 34 : Interprétation de l’observation 1

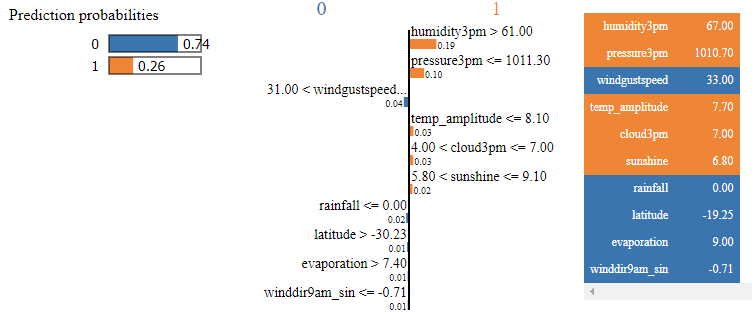


Figure 35 : Interprétation de l’observation 23

### Conclusion des modèles deep learning

Bien que les modèles soient plus complexes, ils sont censés pouvoir mieux interpréter les données à leur disposition afin de prédire des jours de pluie. Il est à noter que le modèle de réseau de neurones ne parvient pas à obtenir des scores meilleurs que les modèles plus simples. En revanche, le modèle XGBOOST obtient de bons scores prédictifs. Il est le seul à pouvoir prédire 1000 jours de pluie dans le jeu de données, et atteint un score d’accuracy de presque 89%.

Dans un second temps, nous avons cherché à utiliser des modèles régionaux : c’est-à-dire utiliser la géographie ou le type de climat pour déterminer au mieux un modèle de prédiction de pluie le lendemain.

# Modélisation régionale

Comme indiqué précédemment, l’objectif de cette section est de pouvoir tester l’hypothèse de modèles régionaux pour prédire la présence de pluie le lendemain.

### Préparation des jeux de données

Afin de réaliser ce type de modèles, nous avons divisé le jeu de données utilisé précédemment en accordance avec le type de climat de chaque localisation du jeu de données. Nous avons utilisé le climat selon la classification Koppen-Geiger. Cinq jeux de données ont été ainsi créés :

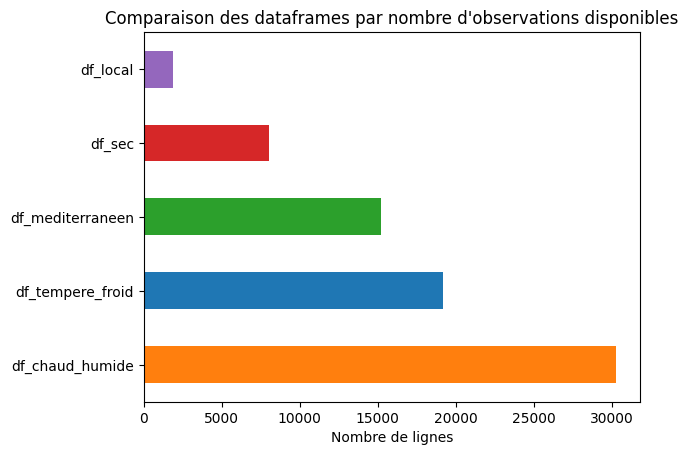


Figure 36 : Type de jeu de données et nombre d’observations associées

Le Df\_local contient les données de Norfolk Island qui est une île au milieu du Pacifique. C’est pour cela que nous avons isolé ce dataframe.

A partir des résultats que nous avions précédemment, nous avons sélectionné le Random Forest Classifier et le XGBoost pour modéliser les différents jeux de données.

### Modélisation par Random Forest

Le tableau suivant récapitule les scores pour les différents jeux de données.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Métrique | Chaud humide | Tempéré froid | Méditerranéen | Sec | Local |
| Précision classe 0 | 0.89 | 0.89 | 0.90 | 0.95 | 0.81 |
| Rappel classe 0 | 0.98 | 0.97 | 0.97 | 0.99 | 0.92 |
| Précision classe 1 | 0.76 | 0.73 | 0.79 | 0.86 | 0.65 |
| Rappel classe 1 | 0.36 | 0.37 | 0.52 | 0.45 | 0.40 |
| F1  classe 1 | 0.48 | 0.49 | 0.63 | 0.59 | 0.49 |
| F1 pondéré (weighted) | 0.86 | 0.86 | 0.88 | 0.93 | 0.77 |
| Précision globale | 0.88 | 0.87 | 0.89 | 0.94 | 0.78 |

Figure 37 : Tableau récapitulatif des scores du Random Forest selon les différents modèles

En résumé, les modèles ont des performances variables en fonction de l'ensemble de données et de la classe prédite. Le modèle df\_chaud\_humide, df\_sec et le modèle df\_mediterraneen semblent avoir des performances supérieures par rapport aux autres modèles en termes de précision, rappel et F1-score pondéré.

### Courbes d’apprentissage

Pour chaque ensemble de données, le code génère un graphique de courbes d'apprentissage avec le nombre d'exemples d'entraînement sur l'axe des x et le score F1 Weighted sur l'axe des y. Le graphique contient deux courbes, une pour le score d'entraînement et une pour le score de validation croisée. Les zones ombrées autour de chaque courbe représentent l'écart type des scores.

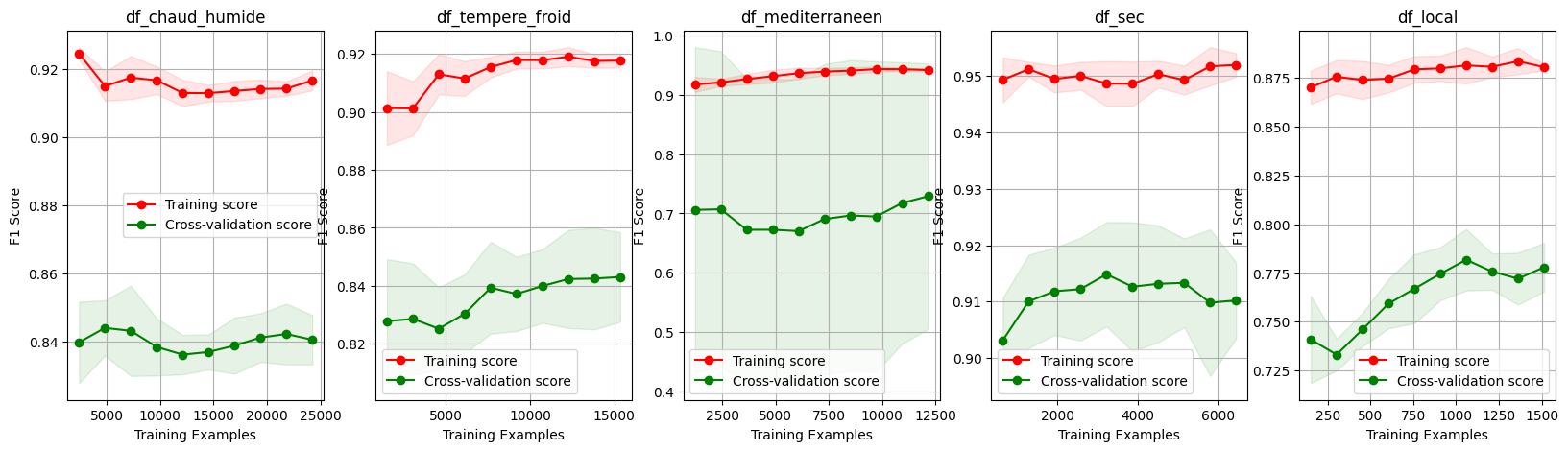


Figure 38 : Courbes d’apprentissage selon les jeux de données

Le but des courbes d'apprentissage est d'évaluer les performances du classificateur sur l'ensemble d'entraînement et l'ensemble de validation à mesure que le nombre d'exemples d'entraînement augmente. Si le score d'entraînement est beaucoup plus élevé que le score de validation, cela indique un surapprentissage, ce qui signifie que le modèle est trop complexe et s'ajuste trop étroitement aux données d'entraînement, ce qui entraîne une mauvaise généralisation aux nouvelles données.

Le paramètre train\_sizes spécifie les tailles des sous-ensembles d'entraînement, qui sont uniformément réparties entre 10% et 100% de la taille totale de l'ensemble d'entraînement. Le paramètre cv spécifie le nombre de plis dans la validation croisée. Le paramètre n\_jobs spécifie le nombre de cœurs de CPU à utiliser pour le calcul parallèle.

Cependant, la largeur de la zone ombrée autour des courbes varie entre les ensembles de données, avec df\_mediterraneen ayant la zone la plus large, ce qui indique une plus grande variabilité dans les scores de performances.

Les scores d'entraînement et de test augmentent à mesure que la taille de l'ensemble d'entraînement augmente. Cela indique que le modèle apprend et améliore ses performances avec plus de données.

Les scores d'entraînement plus élevés que les scores de test indiquent un surapprentissage du modèle sur les données d'entraînement, car le modèle parvient à mieux s'ajuster à ces données spécifiques. Cependant, le fait que les scores de test augmentent également avec la taille de l'ensemble d'entraînement suggère que le modèle parvient toujours à généraliser correctement sur de nouvelles données. La valeur moyenne sur le score de cross validation reste élevée (courbe verte) et stable quelle que soit la taille des données d'entrainement. De plus, l'écart-type (ombre verte) reste réduite.

Il peut y avoir une légère variation des scores d'entraînement et de test en fonction de la taille de l'ensemble d'entraînement. Cependant, ces variations restent généralement assez stables, ce qui indique une certaine cohérence dans les performances du modèle.

Cependant, la largeur de la zone ombrée autour des courbes varie entre les ensembles de données, avec df\_mediterraneen ayant la zone la plus large, ce qui indique une plus grande variabilité dans les scores de performances. Cela peut être dû à une plus grande variabilité dans les données de départs (features).

### Optimisation Bayésienne du Random Forest

Dans un second temps, nous avons cherché à optimiser les modèles par une optimisation Bayésienne :

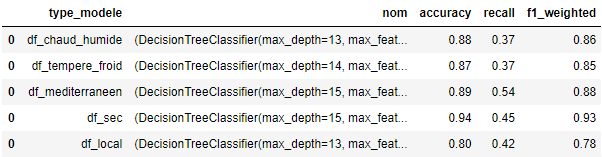


Figure 39 : Résultats de l’optimisation

En comparant les résultats avant et après l'optimisation bayésienne, on peut voir une amélioration significative des performances des modèles. Avant l'optimisation, les précisions variaient entre 0,79 et 0,94, tandis que les rappels variaient entre 0,35 et 0,52. Après l'optimisation, les précisions sont restées stables, mais les rappels ont augmenté, allant de 0,37 à 0,54. Les scores F1 pondérés ont également augmenté dans certains cas. L'optimisation bayésienne a permis de trouver des hyperparamètres plus adaptés à chaque modèle et a amélioré les performances des modèles.

Les courbes d'apprentissage montrent que les performances du classificateur de forêt aléatoire s'améliorent à mesure que le nombre d'exemples d'entraînement augmente (sauf pour le modèle temps sec). Les courbes des scores d'entraînement et de validation croisée sont proches les unes des autres, ce qui indique qu'il n'y a pas de surapprentissage significatif. Cependant, la largeur de la zone ombrée autour des courbes varie entre les ensembles de données, avec df\_mediterraneen ayant la zone la plus large, ce qui indique une plus grande variabilité dans les scores de performances

### Conclusion des modèles régionaux

L’utilisation de modèles régionaux pour prédire la pluie le lendemain semble être mieux adaptée pour prédire la pluie le lendemain. Les scores des métriques utilisées sont relativement plus élevés que sur les modèles simples. Les temps de mise en place de ces modèles peuvent être importants.

# Conclusions

L’étude de ce jeu de données nous a amené à réaliser diverses modélisations, de modèles simples (régression logistique) vers des modèles plus complexes (réseau de neurones).

Nous avons appris que le Random Forest Classifier permettait d’obtenir les scores les plus intéressants en termes de métriques. Par ailleurs, l’explication de ce modèle peut être réalisée grâce à un arbre de décision. Cependant, celui-ci peut être complexe à interpréter du fait du nombre important de dimensions. Le XGBoost parvient tout de même à être facilement interprétable. L’interprétation des modèles est souvent un problème pour les modèles plus complexes de machine learning (la fameuse black box). De même, les temps de process peuvent être longs sur des modèles de type SVM ou Random Forest, tandis qu’une régression logistique est plus rapide.

La question se pose donc dans le cas d’un jeu de données comme le nôtre : quelle est la meilleure façon de traiter les données ? Nous avons réussi à obtenir de meilleurs scores avec des modèles Random Forest et XGBoost, mais pour quel gain final en termes d’explication ?

En revanche, l’utilisation de différents modèles selon les régions climatiques permet de mieux spécifier les données et de pouvoir atteindre des scores plus pertinents. Ici encore, la mise en place de tels modèles requiert du temps de process qui peut être long.

Finalement, nous sommes à la croisée des chemins en termes de décision : faut-il maximiser l’efficience et utiliser un modèle plus simple pour décrire les données, ou alors permettre d’avoir des modèles plus précis au détriment du temps de process ?

Par ailleurs, en fonction de l’utilité de la précision, il sera préférable de conserver tel ou tel modèle, en fonction de la métrique l’on souhaite maximiser, et si l’on préfère maximiser des faux négatifs ou des faux positifs.

# Références

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301.

Dickinson, R. D., Meleshko, V., Randall, D., Sarachik, E., Silva-Dias, P., & Slingo, A. (1996). Climate processes. In *Climate Change 1995: the science of climate change. Contribution of WG1 to the Second Assessment Report of the IPCC* (pp. 193-227). Cambridge University Press.

1. https://www.ipcc.ch/report/ar6/wg2/about/factsheets/ [↑](#footnote-ref-1)
2. https://www.nature.com/articles/d41586-022-02782-w [↑](#footnote-ref-2)
3. https://www.dtf.vic.gov.au/victorias-economic-bulletin-volume-5/economic-impacts-2019-20-bushfires-victoria [↑](#footnote-ref-3)
4. <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package> [↑](#footnote-ref-4)