Prévision météo en Australie

Exploitation, visualisation et preprocessing de données

Table des matières

[1 Source des données 3](#_Toc144795777)

[2 Compréhension des variables 4](#_Toc144795778)

[2.1 Variable cible 5](#_Toc144795779)

[2.2 Variables catégorielles 5](#_Toc144795780)

[2.3 Variables numériques 7](#_Toc144795781)

[2.4 Relation et corrélation entre les variables 9](#_Toc144795782)

[3 Missing data 10](#_Toc144795783)

[3.1 Détermination des valeurs manquantes. 10](#_Toc144795784)

[3.2 Traitement des données manquantes 15](#_Toc144795785)

[4 Outliers 15](#_Toc144795786)

# Source des données

Nous avons utilisé un ensemble de données contenant 10 ans d’observations météorologiques quotidiennes provenant de plusieurs stations météorologiques australiennes. Ces observations sont des observations météorologiques quotidiennes réalisées à 9h et 15h sur une période de 10 ans, du 01/12/2008 au 25/06/2017.

<https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

Le Tableau 1 représente tous les paramètres de données collectés

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nom de colonne** | **Unité** | **Explication** |
| 1 | Date |  | Date d’observation |
| 2 | Location |  | Location de la station météo |
| 3 | MinTemp | Degrés Celsius | Température minimum en 24 heures jusqu’à 9am |
| 4 | MaxTemp | Degrés Celsius | Température maximum en 24 heures jusqu’à 9am |
| 5 | Rainfall | Millimètres | Précipitation en 24 heures jusqu’à 9am |
| 6 | Evaporation | Millimètres | Évaporation en 24 heures jusqu’à 9am |
| 7 | Sunshine | Heure | Soleil radieux en 24 heures jusqu’à minuit |
| 8 | WindGustDir | 16 compass points | Direction de la rafale de vent la plus forte en 24 heures jusqu’à minuit |
| 9 | WindGustSpeed | Kilomètres par heure | Vitesse de la rafale de vent la plus forte en 24 heures jusqu’à minuit |
| 10 | WindDir9am | 16 compass points | Direction de vent à 9am |
| 11 | WindDir3pm | 16 compass points | Direction de vent à 3pm |
| 12 | WindSpeed9am | Kilomètres par heure | Vitesse de vent à 9am |
| 13 | WindSpeed3pm | Kilomètres par heure | Vitesse de vent à 3pm |
| 14 | Humidity9am | Pourcentage | Humidité relative à 9am |
| 15 | Humidity3pm | Pourcentage | Humidité relative à 3pm |
| 16 | Pressure9am | Hectopascals | Pression atmosphérique réduite au niveau moyen de la mer à 9am |
| 17 | Pressure3pm | Hectopascals | Pression atmosphérique réduite au niveau moyen de la mer à 3pm |
| 18 | Cloud9am | Huitièmes | Fraction de ciel obscurcie par les nuages à 9am |
| 19 | Cloud3pm | Huitièmes | Fraction de ciel obscurcie par les nuages à 3pm |
| 20 | Temp9am | Degrés Celsius | Température à 9am |
| 21 | Temp3pm | Degrés Celsius | Température à 3pm |
| 22 | RainToday | Binaire (Yes, No) | La journée en cours a-t-elle reçu des précipitations supérieures à 1 mm en 24 heures jusqu’à 9h ? |
| 23 | RainTomorrow | Binaire (Yes, No) | Le lendemain a-t-il reçu des précipitations dépassant 1 mm en 24 heures jusqu’à 9am ? |

Tableau  : Les paramètres de données collectées

On enlève la colonne « Date » lorsqu’on utilise des modèles de prédiction comme des modèles de classification.

Le Tableau 2 représente un Overview (généré par la librairie *ydata\_profiling*) du dataframe de 22 colonnes. L’ensemble de données contient 145 460 d’observations dont il y a 21 observations qui sont redondantes.

Les 22 variables se divisent en 3 types dont

* 4 variables catégorielles : *Location, WindGustDir, WindDir9am et WindDir3pm*
* 2 variables booléennes : *RainToday, RainTomorrow*
* 16 variables numériques.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Tableau  : Overview du dataset

# Compréhension des variables

Dans la section précédente, nous décrivions l’ensemble de données dans son intégralité alors que nous essayons maintenant de décrire avec précision toutes les variables qui nous intéressent.

## Variable cible

Nous intéressons à prédire la variable « *RainTomorrow* » via les autres variables explicatives. Le Tableau 3 montre qu’il y a 2.2% des observations n’ayant pas d’information sur cette variable. On observe que le nombre de jours où il ne pleut pas est environ 4 fois plus grand que le nombre de jours où il pleut. Cela nous suggère que dans la phase de modélisation, nous pouvons peut-être mettre en œuvre des techniques d’équilibrage des données pour ne pas confondre notre modèle.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Tableau  : Variable cible

## Variables catégorielles

La Figure 1 représente la distribution de la variable *Location*. Les trois stations météorologiques *Uluru, Katherine* et *Nhil* contiennent les moins observations que les autres. Les deux stations *Sydney* et *Canberra* contiennent plus d’observations que les autres. Le nombre d’observations des autres stations reste à peu près homogène.

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

La Figure 2 représente la distribution de la direction du vent à différents moments de la journée.

* *WindDir9am*: à 9 heures du matin, la direction du vent est principalement du nord et au moins de l'ouest au sud-ouest.
* *WindDir3pm* : de même, à 15h, le vent souffle principalement du sud-est et au moins du nord-nord-est.
* *WindGustDir* : la direction des rafales de vent est principalement à l’ouest et au moins au nord-nord-est, comme le vent de 15h.

Figure : Distribution de la variable Location

Une image contenant texte, capture d’écran, Rectangle, ligne

Description générée automatiquement

Figure : Distribution des variables concernant la direction du vent

## Variables numériques

La Figure 3 représente la distribution de chaque variable numérique. Nous pouvons remarquer que seuls certaines d’entre elles sont distribuées presque normalement, comme *MinTemp*, *Humidity3pm*, *Pressure9am*, *Pressure3pm*, *Temp9am*, *Temp3pm*) tandis que d’autres sont soit asymétriques à droite, soit à gauche.

Une image contenant diagramme, origami, conception

Description générée automatiquement

Figure : Distribution des variables numériques

Le Figure 3 représente les boxplots des variables numériques (sauf pour la variable *RainToday*). On constate que les intervalles des valeurs sont bien différents d’une variable à l’autre. On observe aussi que certaines variables, comme *Rainfall*, *Evaporation* et les trois variables concernant la vitesse du vent, contiennent des valeurs aberrantes.

Une image contenant texte, diagramme, Rectangle, capture d’écran

Description générée automatiquement

Figure : Boxplot des variables numériques

## Relation et corrélation entre les variables

L’idée est maintenant de trouver des relations intéressantes qui montrent l’influence d’une variable sur l’autre, de préférence sur la variable cible.

Une image contenant diagramme, capture d’écran

Description générée automatiquement

Figure : Corrélation entre les variables numériques

# Missing data

## Détermination des valeurs manquantes.

Après avoir compris l’ensemble de données, nous devons rechercher les valeurs manquantes. Les valeurs manquantes dans un ensemble de données jouent un rôle très important dans un projet. Si elles ne sont pas traitées, les résultats risquent de ne pas être pertinents.

Les données manquantes peuvent être divisées en 3 catégories :

* **Missing Completely at Random (MCAR)**: ce sont des données qui manquent complètement au hasard. C’est-à-dire que les valeurs manquantes n’ont aucune corrélation avec d’autres valeurs de l’ensemble de données observées ou manquantes.
* **Missing at Random (MAR)**:
* **Not Missing at Random (NMAR)**:

La Figure 6 représentant le pourcentage de valeurs manquantes pour chaque variable montre que toutes les variables, sauf *Location* qui est complete, contiennent des valeurs manquantes. Nous pouvons voir qu’un certaines varibles, comme Sunshine, *Evaporation*, *Cloud9am*, *Cloud3pm*, comportent un grand nombre de valeurs manquantes.

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

Figure : Le % de valeurs manquantes pour chaque variable

La Figure 7 représente une matrice des valeurs manquantes de chaque variable. La couleur de chaque cellule de la matrice est basée sur l’existence ou non des données. Si la couleur est noire, les données existent. Si la couleur est blanche, les données sont manquantes. A partir de ce graphique, nous avons une image de la proportion de données manquantes dans une ligne (observation) ou une colonne (variable).

Comme le montre le graphique résultant, les colonnes *Evaporation*, *Sunshine*, *Cloud9am* et *Cloud3pm* affichent de grandes parties de données manquantes. Cela a été identifié dans le graphique à barres ci-dessus, mais l’avantage supplémentaire est qu’on peut voir comment ces données manquantes sont distribuées dans le dataframe.

Sur le côté droit de la matrice se trouve une sparkline qui va de 0 à gauche au nombre total de colonnes dans le cadre de données à droite. Lorsqu’une ligne a une valeur dans chaque colonne, la ligne sera à la position maximale à droite. A mesure que les valeurs manquantes commencent à augmenter dans cette ligne, la ligne se déplacera vers la gauche. On peut observer qu’il y a des lignes (observations) contenant un grand nombre de valeurs manquantes.

Une image contenant croquis, art, bâtiment, noir et blanc

Description générée automatiquement avec une confiance moyenne

Figure : Matrice des valeurs manquantes

Le Figure 8 représente le pourcentage de données manquantes par observation.

Une image contenant texte, capture d’écran, affichage, diagramme

Description générée automatiquement

Figure : Le % de valeurs manquantes pour chaque observation

La Figure 9 représente les corrélations de nullité entre chacune des différentes variables.

* Les valeurs proches de 1 indiquent que la présence de valeurs manquantes dans une variable est corrélée à la présence de valeurs manquantes dans une autre variable.
* Les valeurs proches de -1 indiquent que la présence de valeurs manquantes dans une variable est anti-corrélée à la présence de valeurs manquantes dans une autre variable. Autrement dit, lorsque des valeurs manquantes sont présentes dans une variable, des valeurs de données sont présentes dans l’autre variable et l’inverse.
* Les valeurs proches de 0 indiquent qu’il y a peu ou pas de relation entre la présence de valeurs manquantes dans une variable et dans une autre.

Nous pouvons voir dans l’ensemble de données que la variable *WindGustSpeed* et la *WindGustDir* ont une corrélation de 1, ce qui souligne que si la valeur de *WindGustSpeed* est manquante, la valeur de *WindGustDir* sera également manquante. On observe le même effet entre la *RainToday* et la *Rainfall*.

Une image contenant capture d’écran, diagramme, conception

Description générée automatiquement

Figure : Corrélation de nullité entre les variables

Le Figure 10 représente un regroupement hiérarchique des variables qui ont de fortes corrélations de nullité. Si plusieurs variables sont regroupées au niveau zéro, la présence de valeurs manquantes dans l’une de ces variables est directement liée à la présence ou à l’absence de valeurs manquantes dans les autres colonnes. Plus les variables sont séparées dans l’arbre, moins les valeurs manquantes sont susceptibles d’être corrélées entre les variables. Dans le graphique de dendrogram, nous pouvons voir qu’il y a deux groupes distincts. Le premier se trouve sur le côté gauche (*Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm*) qui ont tous un degré élevé de la valeurs manquantes. La seconde est à droite, avec le reste des variables qui sont plus complètes.

Une image contenant croquis, diagramme, dessin, Dessin technique

Description générée automatiquement

Figure : Dendrogram de nullité entre les variables

## Traitement des données manquantes

Le traitement de données manquantes est une étape essentielle de la préparation des données pour l’analyse et la modélisation. Dans notre ensemble de données, plusieurs stratégies ont été adoptées pour traiter ces valeurs manquantes.

### Suppression des lignes contenant des données manquantes

Dans le cadre de cette analyse, nous avons adopté une approche progressive pour la suppression des lignes contenant des données manquantes.

* **Suppression des lignes avec des données manquantes pour la variable cible**: La variable cible *RainTomorrow* est essentielle pour notre modèle prédictif. Puisque c’est ce que nous cherchons à prédire, les lignes contenant des valeurs manquantes pour cette variable ont été supprimées, représentant 2.2% de l’ensemble de données. Cette suppression est justifiée car imputer la variable cible pourrait introduire un biais ou une inexactitude dans nos prédictions.
* **Suppression des lignes avec une forte proportion de données manquantes** : Avant de procéder à l’imputation des données manquantes, il est essentiel d’évaluer si certaines lignes ont une proportion exorbitante de valeurs manquantes, au point que leur utilité est mise en question. Dans notre jeu de données, toutes les lignes contenant plus de 50% de valeurs manquantes ont été jugées comme n’ayant pas suffisamment d’information pour être utiles et ont donc été supprimées.

### Méthodes d’imputation

Une fois les étapes initiales de suppression terminées, trois méthodes d’imputation distinctes ont été envisagées pour traiter les données manquantes restants dans les variables numériques.

**Imputation par la Moyenne** : cette méthode remplace les valeurs manquantes par la moyenne de la colonne correspondante.

**Imputation par la Médiane** : les valeurs manquantes sont remplacées par la médiane de la colonne.

Les graphiques ci-dessous illustrent les distributions des variables avant et après l’imputation. Sur chaque graphique :

* La courbe noire dépeint la distribution de la variable en présence des données manquantes.
* Les courbes rouge et bleue illustrent respectivement les distributions après l’imputation par moyenne et par médiane.

Les méthodes d’imputation par la moyenne ou par la médiane sont certes simples à mettre en œuvre. Toutefois, comme le montrent les graphiques elles peuvent altérer considérablement la distribution des variables, en particulier lorsque celles-ci présentent un taux élevé de données manquantes.

Une image contenant diagramme, texte, origami, motif

Description générée automatiquement

Figure : Distribution des variables avant et après l’imputation par moyenne/médiane

Une image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquement

**Imputation KNN** : L’imputation basée sur les k plus proches voisins est une méthode plus sophistiquée qui prend en compte les similarités entre les observations pour imputer les données manquantes. Au lieu de remplir avec une valeur unique (comme la moyenne ou la médiane), elle utilise les *k* observations les plus similaires pour estimer la valeur manquante. Bien que cette méthode puisse être plus précise, elle peut être assez gourmande en temps et en ressources, en particulier pour de grands ensembles de données. La raison est en effet, pour chaque point avec une valeur manquante, l’algorithme KNN essaie de trouver les « k » voisins les plus proches en calculant la distance entre le point cible et tous les autres points. Ensuite, il utilise ces « k » voisins pour imputer la valeur manquante. Pour un ensemble de données de taille 145 460 lignes dont il y a environ 60% de lignes contenant au moins une valeur manquante, cela signifie potentiellement des milliards de calculs de distance.

Pour déterminer la valeur optimale du le paramètre *k* dans l’imputation KNN, nous avons testé différentes valeurs pour k, allant de 2 à 4. Les graphiques ci-dessous illustrent les distributions des variables avant et après l’imputation. Sur chaque graphique :

* La courbe noire dépeint la distribution de la variable en présence des données manquantes.
* Les courbes verte, rouge et bleue illustrent respectivement les distributions après l’imputation KNN pour k = 2, 3, 4.

Il est à noter que pour les variables, *Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm*, qui présentent un pourcentage élevé de valeurs manquantes, l’imputation KNN impacte plus sensiblement leurs. En revanche, pour les autres variables avec moins de 10% de valeurs manquantes, l’imputation KNN ne perturbe pas de manière significative leur distribution initiale.

Une image contenant texte, diagramme, motif, conception

Description générée automatiquement

Figure  : Distribution des variables avant et après l’imputation KNN, k=2 , 3Une image contenant texte, capture d’écran, diagramme, Caractère coloré

Description générée automatiquement

Figure : Boxplot des variables avant et après l’imputation KNN, k = 2, 3, 4

Après avoir analysé les différentes méthodes d’imputation, il est clair que l’imputation KNN offre des résultats plus fidèles et cohérents comparativement aux méthodes d’imputation par la moyenne ou la médiane. Les distorsions introduites par ces deux dernières techniques, particulièrement visibles dans notre contexte, rendent l’imputation KNN nettement supérieure en matière de préservation de la structure initiale des données. Par conséquent, nous avons décidé d’adopter l’imputation KNN comme méthode privilégiée pour traiter les données manquantes dans cet ensemble de données.

# Outliers

Je n’ai pas encore rédigé cette partie