Présentation du projet

Table des matières

[1 Introduction 1](#_Toc159687078)

[2 Le jeu de données 1](#_Toc159687079)

[2.1 Source de données 1](#_Toc159687080)

[2.2 Exploration des données 2](#_Toc159687081)

[2.2.1 Variables catégorielles 3](#_Toc159687082)

[2.2.2 Variables numériques 4](#_Toc159687083)

[2.2.3 Corrélation entre les variables numériques 4](#_Toc159687084)

[2.2.4 Variables cibles 5](#_Toc159687085)

[2.2.5 Valeurs manquantes 6](#_Toc159687086)

[3 Pre-processing et feature engineering 7](#_Toc159687087)

[3.1 Nettoyage des données 7](#_Toc159687088)

[3.2 Transformation des données 7](#_Toc159687089)

[3.3 Ajout de variables 7](#_Toc159687090)

[4 Modélisation et prédiction 8](#_Toc159687091)

[4.1 RainTomorrow 9](#_Toc159687092)

[4.1.1 Résultats de la classification par les modèles classiques de machine learning : 9](#_Toc159687093)

[4.1.2 Résultats des modèles Deep Learning 10](#_Toc159687094)

[4.2 Prédiction de la pluie à un horizon de temps 10](#_Toc159687095)

[4.3 MaxTemp 12](#_Toc159687096)

[5 Conclusion 15](#_Toc159687097)

|  |  |  |
| --- | --- | --- |
| Partie | Temps | Qui ? |
| Intro | 2 |  |
| 2 Source de données | 2 |  |
| 3.1/3.2 Exploration des données | 2 |  |
| 3.3 Ajout de variables | 1 | Sophie |
| 4.1 Modelisation RT | 4 |  |
| 4.2 Prédiction horizon de temps | 3 | Sophie |
| 4.3 MaxTemp | 3 | Sophie |
| Conclusion | 3 |  |

# Introduction

# Le jeu de données

## Source de données

Le data set est celui disponible sur Kaggle pour le projet « Rain in Australia » ( <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package> ).

Ce dataset contient presque 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 01/11/2007 au 25/06/2017.

## Exploration des données

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nom de colonne** | **Unité** | **Explication** |
| 1 | Date | Timestamp | Date d’observation |
| 2 | Location | Chaîne de caractères | Nom du lieu 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 points cardinaux | 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 points cardinaux | Direction de vent à 9am |
| 11 | WindDir3pm | 16 points cardinaux | 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 ? |

### Variables catégorielles

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

’ Une image contenant diagramme, cercle

Description générée automatiquement

### Variables numériques

Une image contenant texte, conception, origami

Description générée automatiquement

### Corrélation entre les variables numériques

Une image contenant texte, capture d’écran, diagramme

Description générée automatiquement

### Variables cibles

**Rain Tomorrow**

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

Description générée automatiquement

Une image contenant carte, atlas, texte

Description générée automatiquement

### Valeurs manquantes

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

Description générée automatiquement

Une image contenant texte, capture d’écran, Caractère coloré, conception

Description générée automatiquement

# Pre-processing et feature engineering

## Nettoyage des données

* Suppression des doublons
* Suppression des lignes avec de données manquantes pour la variable cible.
* Suppression des lignes avec une forte proportion de données manquantes
* Imputation des données manquantes par KNN

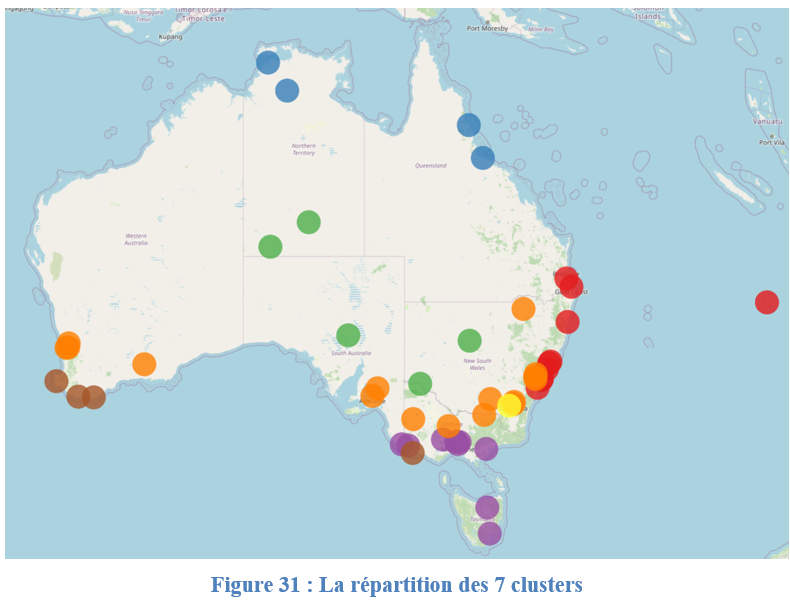
## Transformation des données

* Pour les trois variables *WindGustDir*, *WindDir9am*, *WindDir3pm* qui concernent la direction de vent, nous avons effectué 2 transformations différentes :
* Encodage OneHot
* Approche trigonométrique

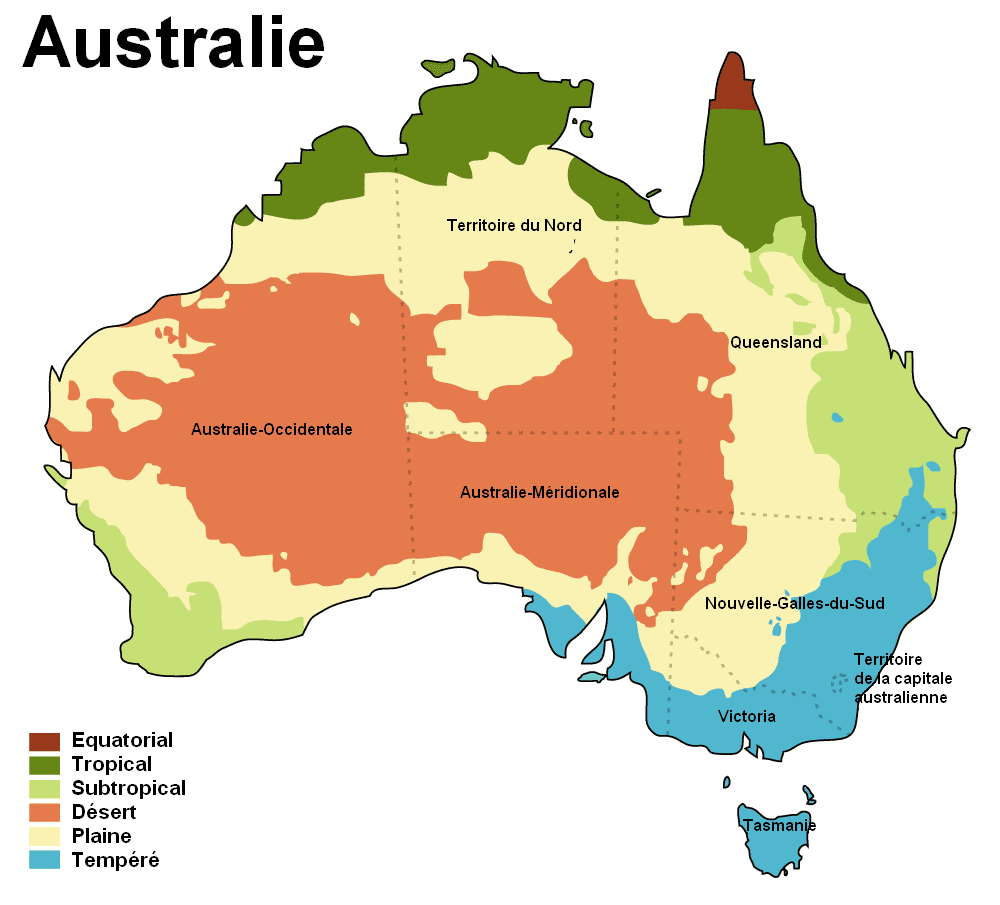
## Ajout de variables

Nous avons enrichi les features de nouvelles variables:

* Latitude et longitude de chaque Location
* Amplitude thermique (=MaxTemp-MinTemp)
* Cos (2 pi \* numéro du jour/365) (=approche trigonométrique du nb de j dans l’année)
* Cos (4 pi \* numéro du jour/365) (=idem, mais en mettant deux cycles dans une seule année, du fait d’un constat fait sur les prédictions sur une année)
* Zone climatique de chaque Location, issue de la clusterisation



Comparaison avec climats d’après wikipédia



# Modélisation et prédiction

Trois niveaux de modélisation/prédiction :

* Un niveau macro, avec des modèles portant sur l’ensemble des données australiennes du jeu de données
* Un niveau micro, où nous génèrerons des modèles spécifiques pour chaque *Location*
* Un niveau intermédiaire, dans lequel nous aurons clusterisé l’Australie en plusieurs zones climatiques

## RainTomorrow

* Les modèles utilisés :
* des modèles de classification comme Logistic Regression, Decision Tree, Random Forest, XGBoost
* des modèles de Deep Learning comme DNN, RNN

### Résultats de la classification par les modèles classiques de machine learning :

* Présenter les scores des 4 modèles en fonction de métrique à optimiser 🡪 en déduire que le modèle XGBoost est le meilleur.
* Présenter les résultats (les scores) du modèle XGBoost pour les 3 niveaux de modélisation/prédiction.

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

Description générée automatiquement

* Interpréter le modèle XGBoost avec SHAP

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

Description générée automatiquementUne image contenant texte, capture d’écran

Description générée automatiquement

* Pour chaque niveau ajouter dans l’application streamlit les modèles déjà entraînés pour pourvoir faire une prédiction en direct et voir la contribution de chaque variable dans la prédiction

### Résultats des modèles Deep Learning

* Modèle DNN :

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

Description générée automatiquement

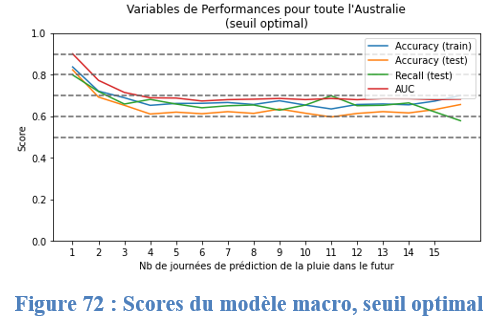
Ajouter dans l’application une partie avec des différentes options pour les paramètres comme

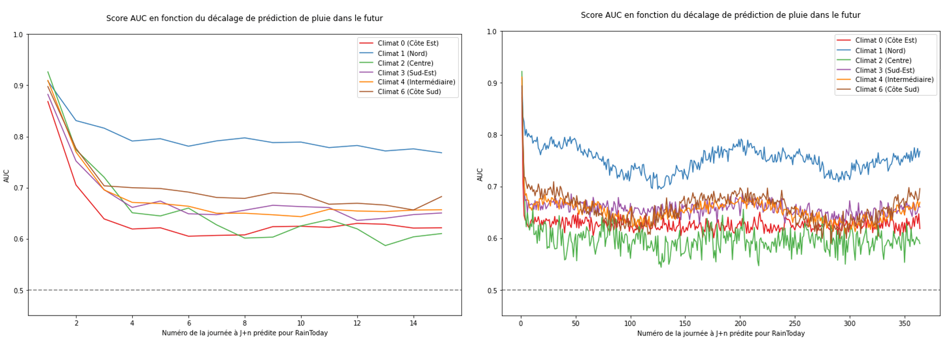
* le nombre de couches
* le nombre de neurones
* learning rate
* fonctions d’activation
* batch size

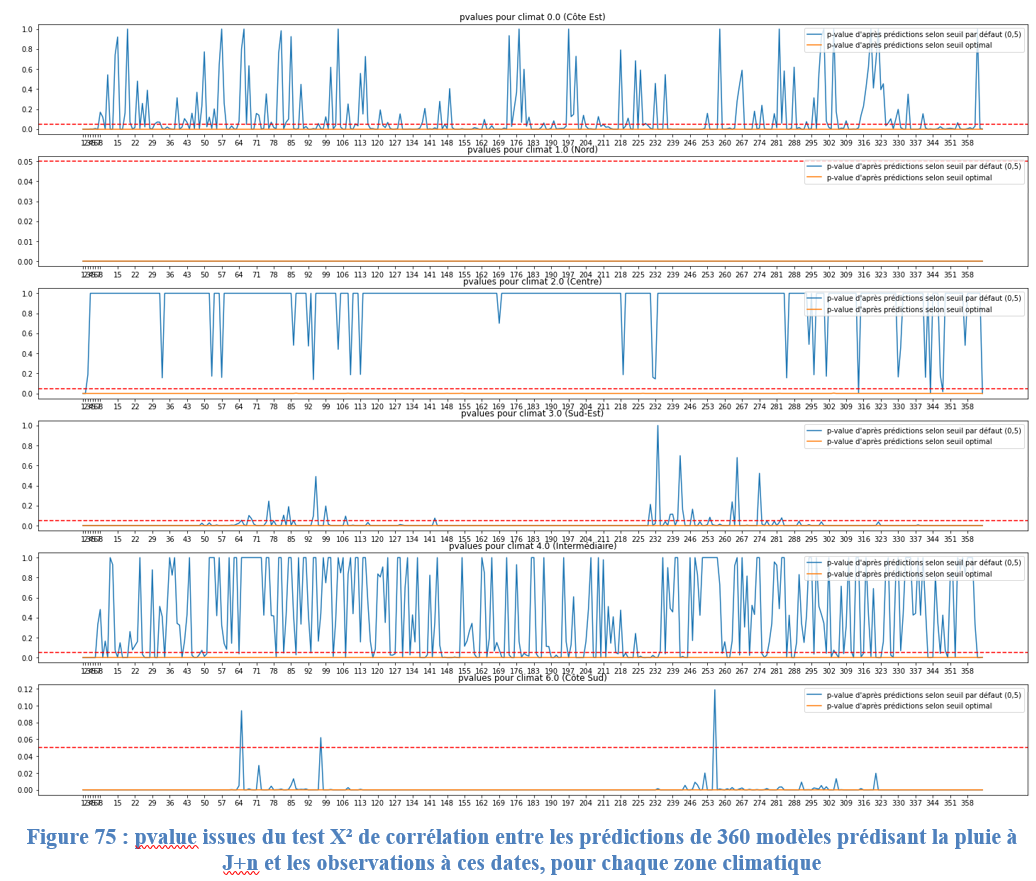
Cela permettra de choisir une combinaison des paramètres et afficher les graphiques d’évolution de la fonction de perte et de l’accuracy.

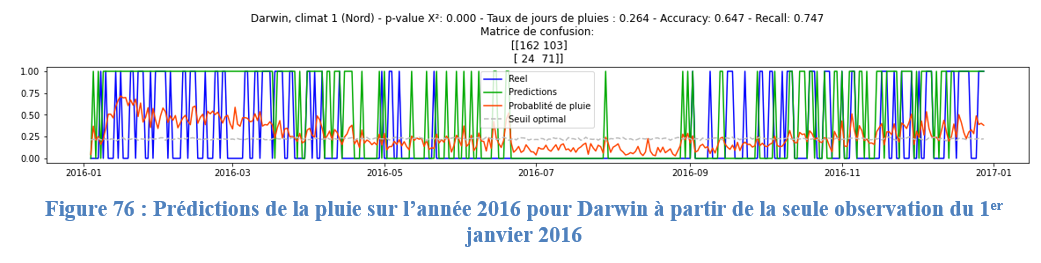
* Modèle RNN : résultats peu intéressants. Voire plutôt RNN sur MaxTemp

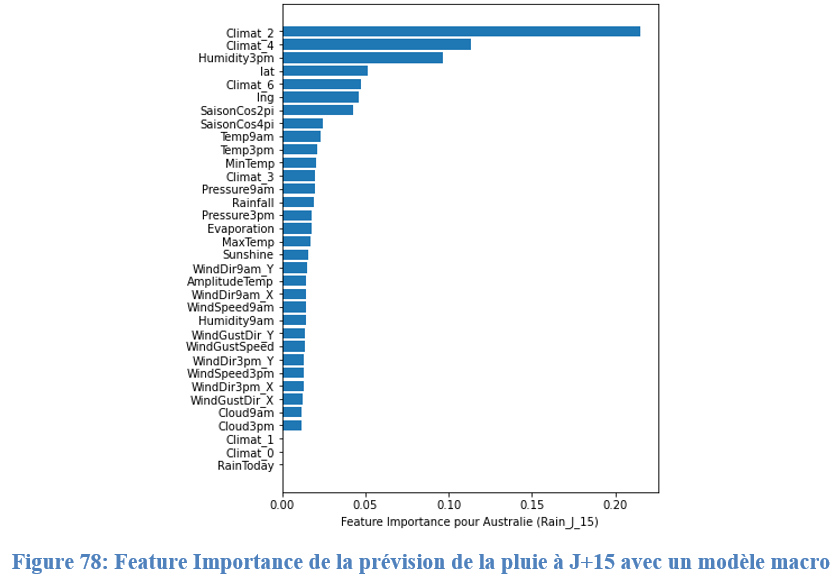
## Prédiction de la pluie à un horizon de temps

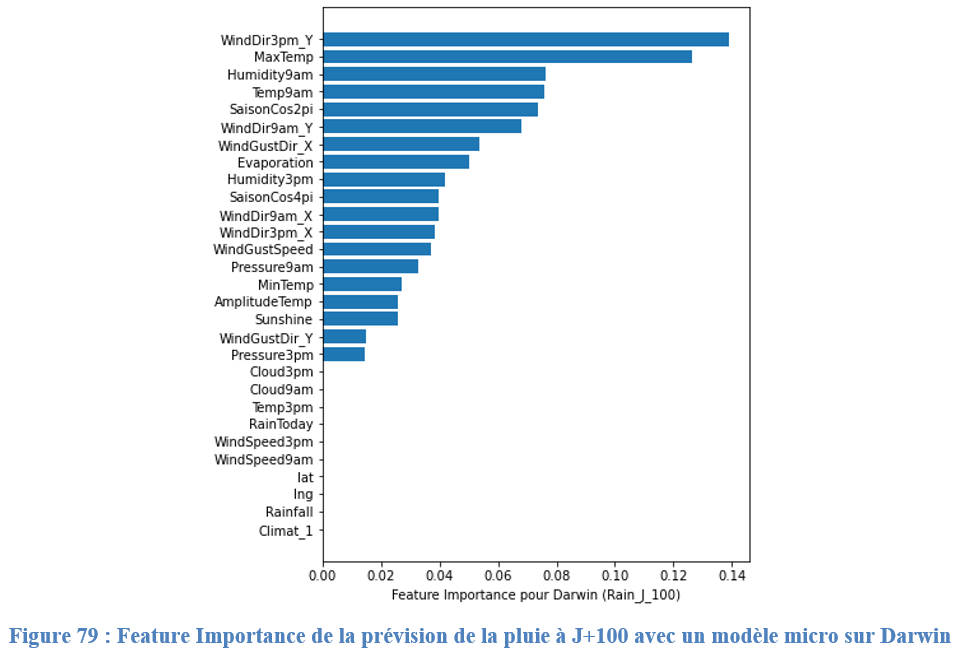




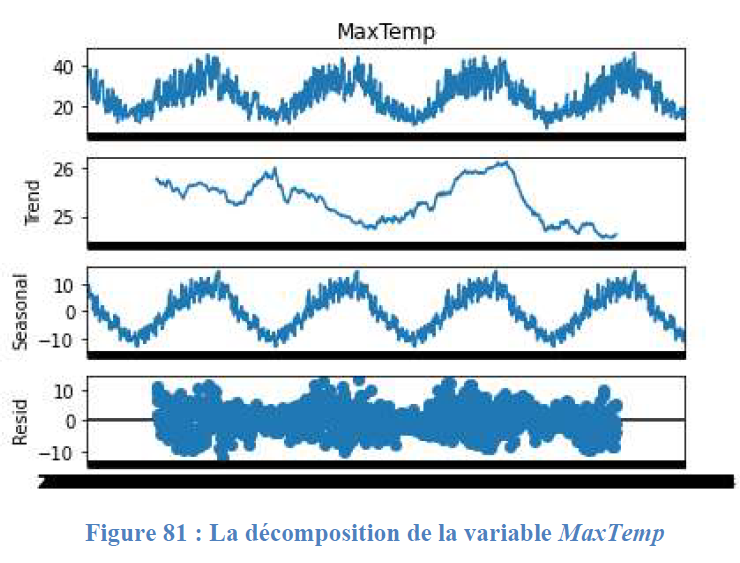




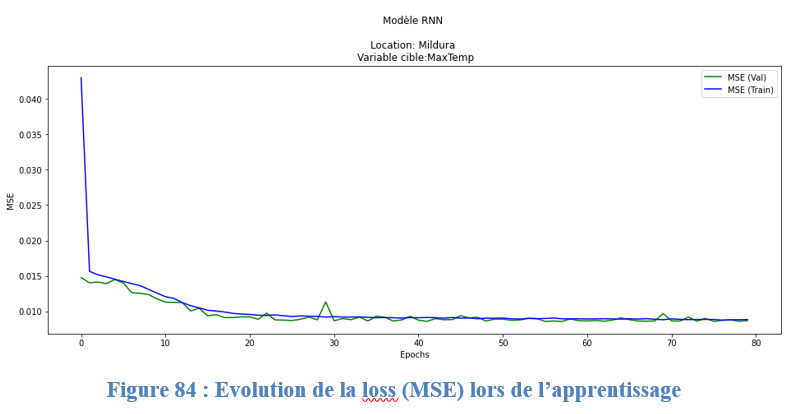


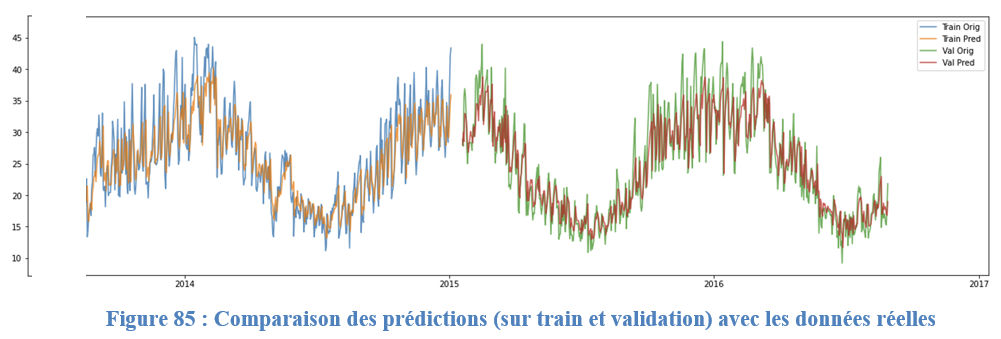


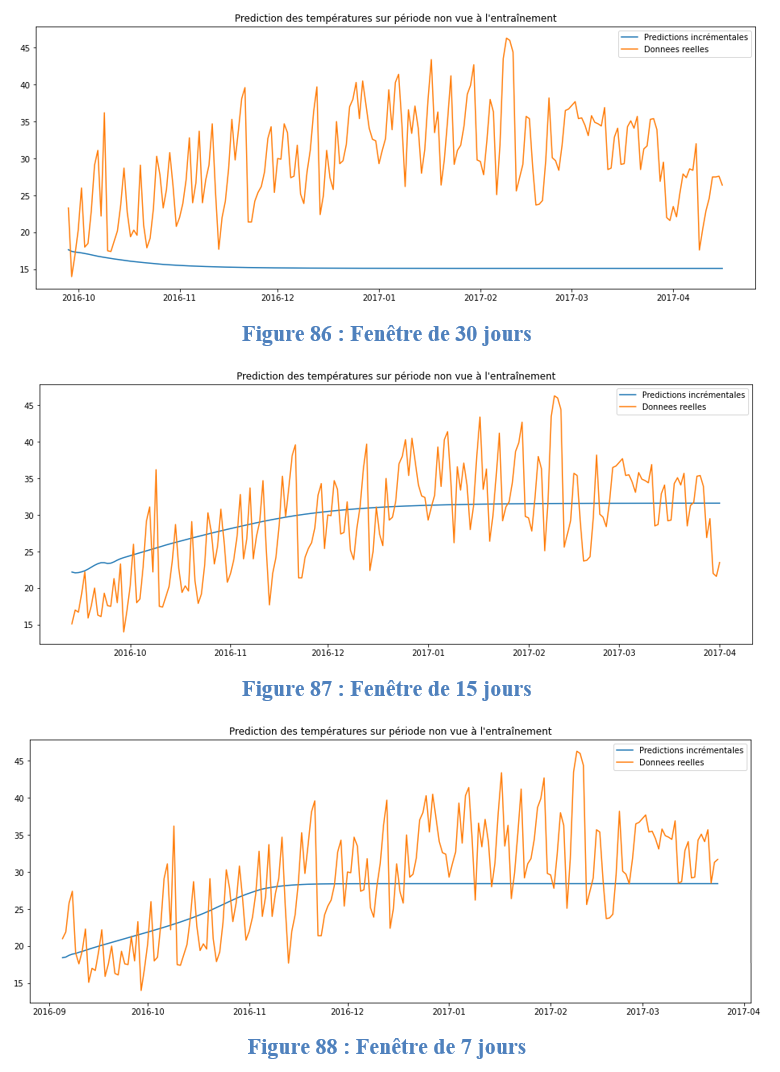
## MaxTemp



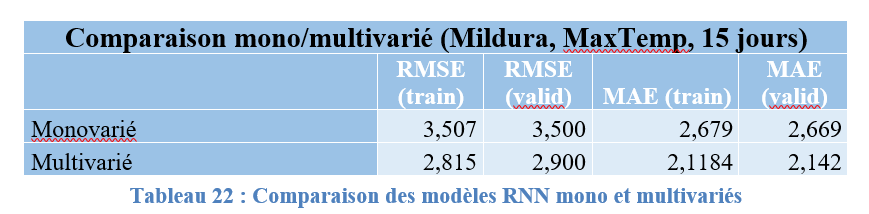
* Modèle RNN :
* Monovarié

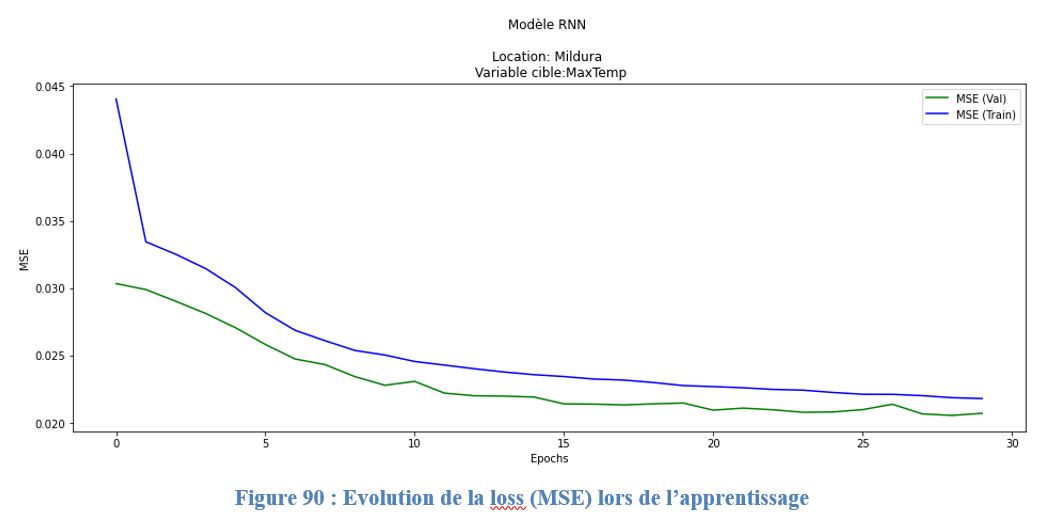


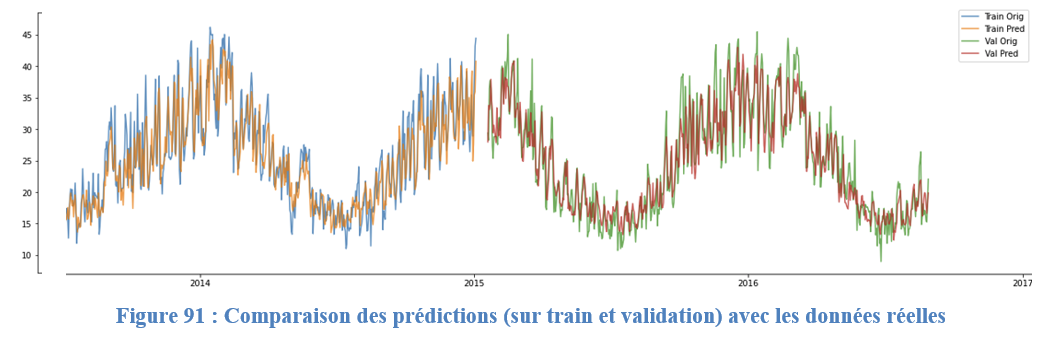




* Mulivarié







# Conclusion

* XGBoost est le meilleur modèle : très rapide et meilleurs scores (86,6% d’accuracy en global pour 77,6% de journées pluvieuses, AUC de 0,90)
* Temps très important en feature engineering pour un gain faible (85,7% d’accuracy, AUC de 0,88)

Pistes d’améliorations :

* Possibilité d’affiner les qualités en fonction d’objectifs précis (maximiser TP vs minimiser FP)
* Approfondir RNN
* Explorer d’autres bibliothèques de séries temporelles (GluonTS, TSFresh,…)
* Transformers
* Enrichir dataset (+ d’années, + de Locations, - de NA, + de features,…)