RAPPORT DU PROJET « DÉFI-IA 2022 » ÉLABORÉ PAR :

Thomas FRAMERY

Yoann MAAREK

Dorian VOYDIE

Dans le cadre du Mastère Spécialisé VALDOM

Table des matières

[Introduction 3](#_Toc91525008)

[Analyse exploratoire des données 3](#_Toc91525009)

[Les jeux de données 3](#_Toc91525010)

[Statistiques du jeu X\_train 4](#_Toc91525011)

[Statistiques du jeu Y\_train (target feature = « Ground\_truth ») 4](#_Toc91525012)

[Transformation des datasets 5](#_Toc91525013)

# Introduction

Dans le cadre de l’UE « IA Frameworks » du master VALDOM, nous avons pu travailler sur un projet en Data Science en collaboration avec METEO FRANCE. En effet, grâce aux données fournies par l’organisme météorologique, notre objectif était de prédire la quantité cumulée de pluie tombée sur une certaine station au sol. Par conséquent, nous avons élaboré une stratégie de conception pour obtenir la meilleure précision dans nos prédictions. Vous pouvez trouver les données sur le site MeteoNet : <https://meteonet.umr-cnrm.fr/>. L’évaluation de cette compétition était effectuée à travers la plateforme Kaggle : <https://www.kaggle.com/c/defi-ia-2022>

# Analyse exploratoire des données

## Les jeux de données

Dans cette première partie, nous allons analyser le contenu des données (la forme, les dimensions, la cardinalité, …) proposées par Météo France. Cette première phase est cruciale car elle permet de prendre en main les données et de savoir véritablement ce que l’on manipule.

Initialement, METEO FRANCE nous a fourni 2 datasets d’entrainement à travers la plateforme Kaggle :

* X\_station\_train (= X\_train)
* Y\_train

*Figure 1 : Les 5 premières lignes du dataset X\_station\_train*

Une image contenant table

Description générée automatiquement

Ici, ce qu’il a été important de remarquer, ce sont les informations contenues au niveau de l’identifiant du ligne. Par exemple, pour la première ligne, l’identifiant est :

Numéro de station 14066001\_0\_0 Heure

Jour

Ainsi, une ligne peut se lire comme suit : « Pour cette station, ce jour, cette heure, nous avons ces différents paramètres météorologiques ».

*Figure 2 : Les 5 premières lignes du dataset Y\_train*

*Une image contenant table

Description générée automatiquement*

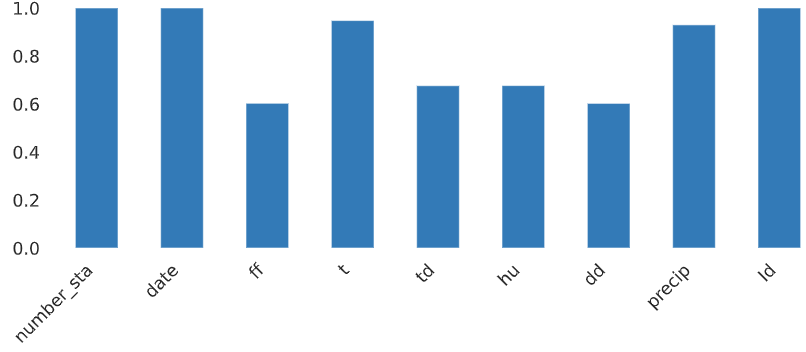
Ici, on voit que contrairement à précédemment, l’identifiant contient le numéro de station ainsi que le jour.

Par conséquent, il va falloir effectuer un travail de preprocessing sur X\_train et plus précisément au niveau de l’identifiant pour qu’il contienne seulement le numéro de station et le jour.

## Statistiques du jeu X\_train

|  |  |
| --- | --- |
| Number of Variables | 9 |
| Number of rows | 4409474 |
| Missing cells (%) | 17.4% |
| Variable types | Numerical : 7  Categorical : 2 (date & number\_sta) |

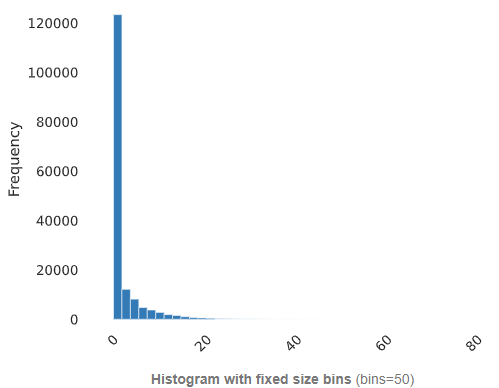
Missing Values (in **blue** percentage of not NaN values) :



## Statistiques du jeu Y\_train (target feature = « Ground\_truth »)

|  |  |
| --- | --- |
| Number of Variables | 4 |
| Number of rows | 183747 |
| Missing cells (%) | 2.9% |
| Variable types | Numerical : 2  Categorical : 2 (Id & number\_sta) |

Distribution :



# Transformation des datasets

Tout d’abord, cette transformation se subdivise en plusieurs étapes :

1. Écrire les fonctions de preprocessing
2. Appliquer ces fonctions aux datasets associés
3. Merge les 2 datasets résultants pour construire le trainset
4. Introduire les données AROME et ARPEGE (fournies par MeteoNet)
5. Introduire les coordonnées des stations pour plus de précision

Conformément à la stratégie mentionnée ci-avant, examinons les fonctions de preprocessing.

*Figure 3 : Fonction de preprocessing concernant X\_train*

Une image contenant texte

Description générée automatiquement

Une image contenant table

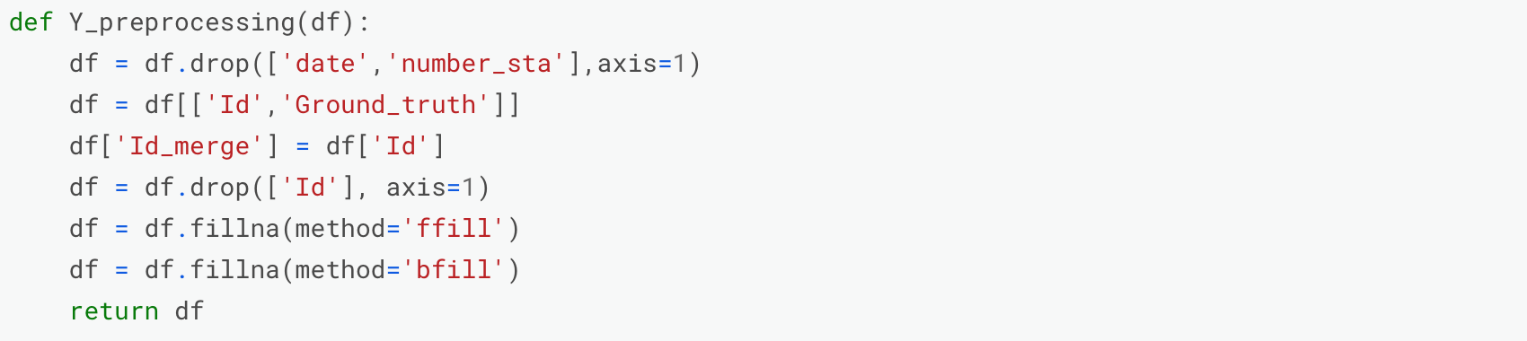
Description générée automatiquement*Figure 4 : Application du preprocessing sur X\_train*

Attribut de jointure

Ici, l’enjeu était de transformer le dataset X\_train pour que l’identifiant soit composé du numéro de station ainsi que du jour. Ce nouvel identifiant, nommé « Id\_merge » a servi d’attribut de jointure avec le dataset Y\_train.

L’autre enjeu concernait la colonne « precip ». Nous avons décidé de la multiplier par 24 mais cette opération prend tout son sens après la jointure des datasets et plus précisément lors du calcul de la moyenne.  
Mais avant cela, visualisons la façon dont le dataset Y\_train a été mis en forme

*Figure 5 : Fonction de preprocessing concernant Y\_train*



*Figure 6 : Application du preprocessing sur Y\_train*

Une image contenant table

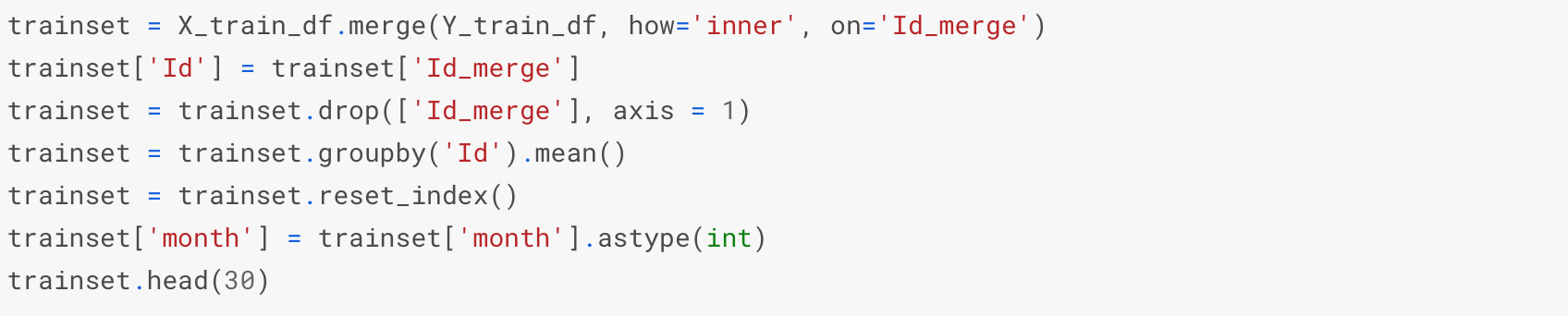
Description générée automatiquement

Attribut de jointure

Ensuite, nous devons donc effectuer la jointure des 2 datasets et créer ainsi le trainset.

*Figure 7 : Création du trainset*

Jointure des datasets



Moyenne des attributs de X\_train sur 24h car X\_train est décomposé en heures et Y\_train en jours.

Somme des précipitations sur un jour

Moyenne des features sur la journée

*Figure 8 : Visualisation du trainset*

Une image contenant table

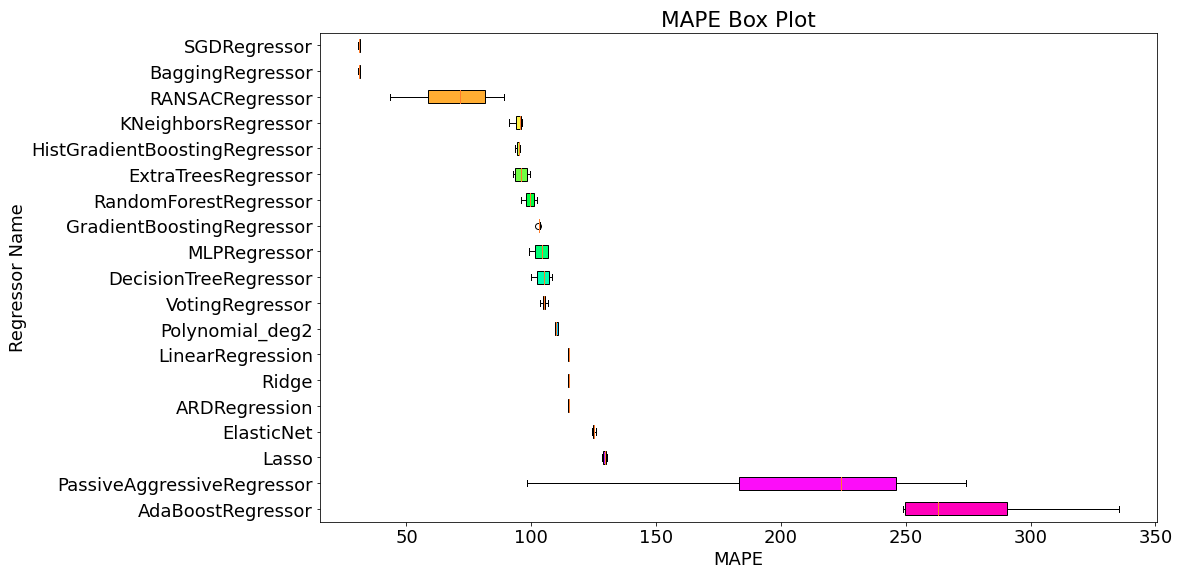
Description générée automatiquement

Features Target

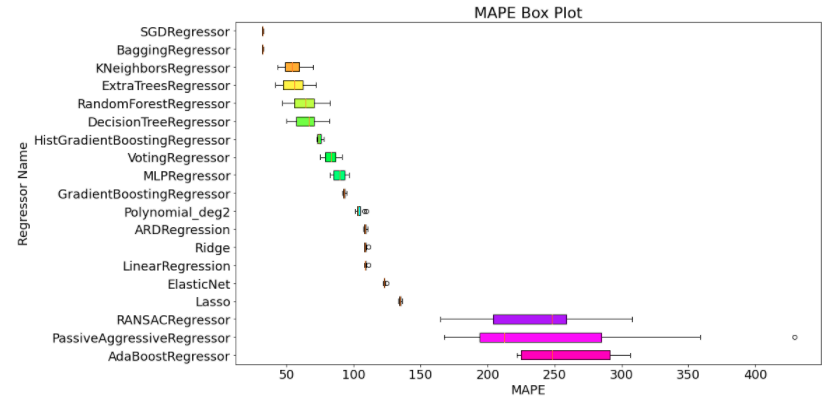
Si nous suivons notre stratégie mentionnée en début de partie, elle comporte l’introduction des données AROME (données supplémentaires) fournies par METEO FRANCE.

L’identifiant d’un dataset Arome étant composé du numéro de station ainsi que du jour, nous n’avons pas de modifications particulières à apporter pour l’intégrer dans notre trainset.

A exploiter : MAPEs pour le trainset Kaggle uniquement



A exploiter : MAPEs pour le trainset complet (avec arome et arpege)



­