



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

Thomas FRAMERY

Yoann MAAREK

Dorian VOYDIE

dans le cadre du Mastère Spécialisé VALDOM

# Introduction

Dans le cadre de l’UE « IA Frameworks », nous avons dû travailler sur un projet de type 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.

# Analyse exploratoire des données

Dans cette première partie, nous allons présenter comment nous avons analysé les données de base fournies par METEO FRANCE.

Cette première phase demeure cruciale dans le sens où elle permet de prendre en main les données et de savoir véritablement ce qu’on manipule.

Initialement, METEO FRANCE nous a fourni 3 datasets :

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

*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 pré-processing sur X\_train et plus précisément au niveau de l’identifiant pour qu’il contienne seulement le numéro de station et le jour.

# Transformation des datasets

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

1. Écrire les fonctions de pré-processing
2. Appliquer ces fonctions aux datasets associés
3. Réunir les 2 datasets résultants pour construire le trainset
4. Introduire les données AROME (fournies par METEO FRANCE)
5. Introduire les coordonnées des stations pour plus de précision

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

*Figure 3 : Fonction de pré-processing concernant X\_train*

Une image contenant texte

Description générée automatiquement

*Figure 4 : Application du pré-processing sur X\_train*

Attribut de jointure

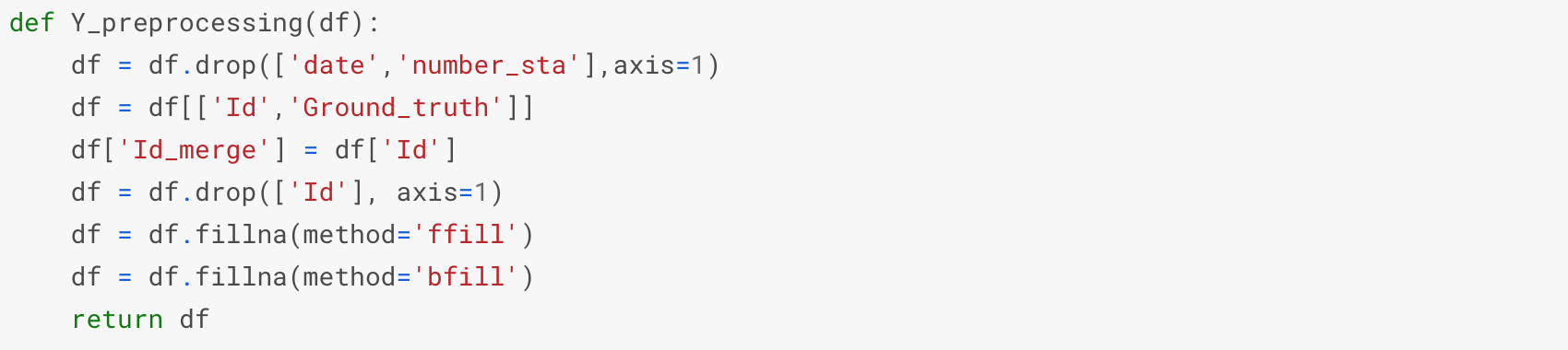
Une image contenant table

Description générée automatiquement

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 le dataset Y\_train « pré-processé »

*Figure 5 : Fonction de pré-processing concernant Y\_train*



*Figure 6 : Application du pré-processing sur Y\_train*

Une image contenant table

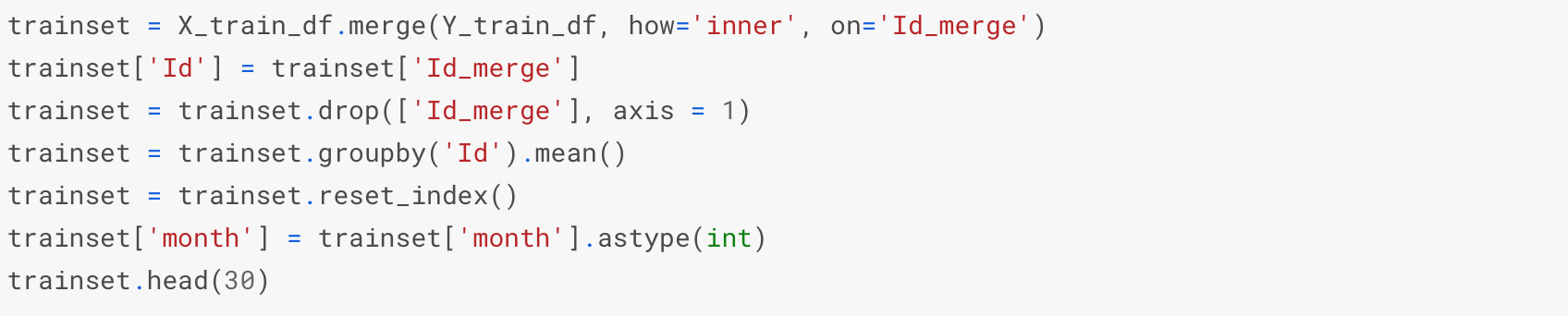
Description générée automatiquement

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 et en particulier des précipitations

Moyenne des précipitations sur un jour

*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.

Par contre, pour introduire les coordonnées des stations, nous ne disposons que du numéro de station pour effectuer la jointure avec notre trainset.

# Modélisation

­