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](#_Toc92359349)

[Analyse exploratoire des données 3](#_Toc92359350)

[Les jeux de données 3](#_Toc92359351)

[Statistiques du jeu X\_train 4](#_Toc92359352)

[Statistiques du jeu Y\_train (target feature = « Ground\_truth ») 4](#_Toc92359353)

[Transformation des datasets 4](#_Toc92359354)

# 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*

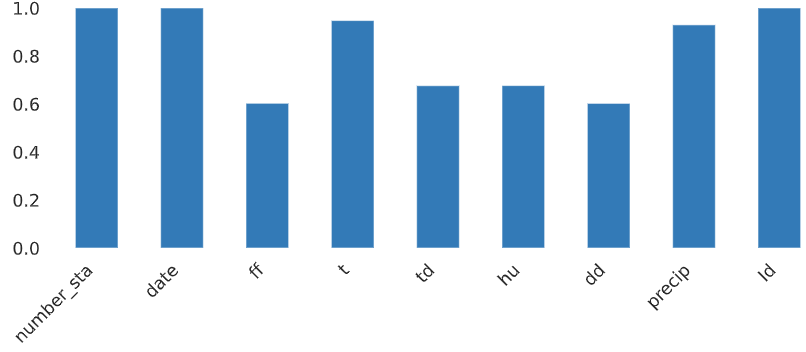
Ici, on voit que contrairement à précédemment, l’identifiant contient le numéro de station ainsi que le jour.

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) |

# 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

Création de colonnes « jour », « mois », « heure »

Mise en ordre des colonnes et au format int

Création d’une colonne pour merge avec Y\_train

Tri des lignes chronologique

*Figure 4 : Application du preprocessing 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 la façon dont le dataset Y\_train a été mis en forme

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

Une image contenant texte

Description générée automatiquement

On ne garde que le label Ground\_truth, on crée l’Id\_merge pour merge avec X\_train et on enlève les lignes NaNs 🡺 perte = 3%

*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

Une image contenant texte

Description générée automatiquement

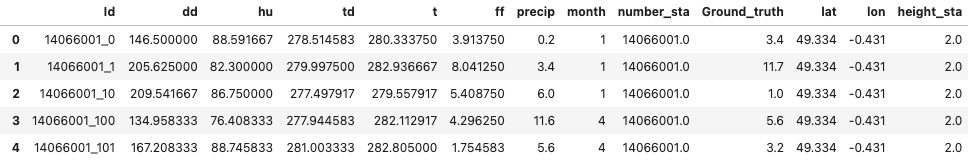
Jointure des coordonnées de stations

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

Moyenne des features sur la journée

Somme des précipitations sur un jour

*Figure 8 : Visualisation du trainset*



Features Target Features

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.

*Figure 9 : Imputation*

|  |  |
| --- | --- |
| Une image contenant texte  Description générée automatiquement | Nous avons essayé plusieurs imputations, d’abord dropna, puis un IterativeImputer afin d’aller chercher les valeurs manquantes dans des lignes similaires. Le KNNImputer étant très long nous ne l’avons pas testé |

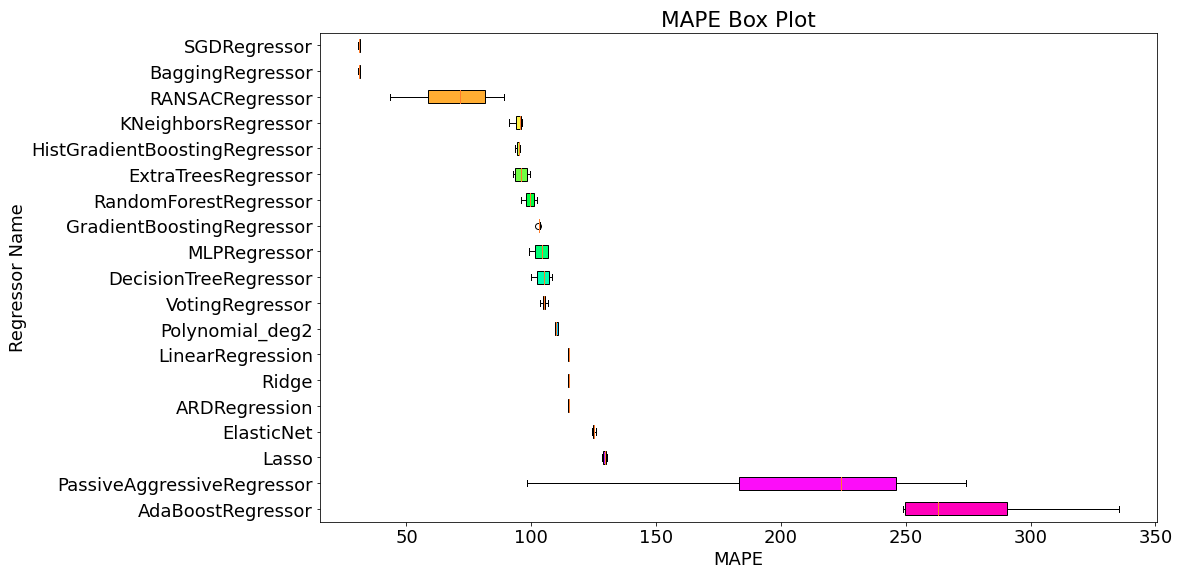
*Figure 10 : Encodage*

|  |  |
| --- | --- |
| Une image contenant texte  Description générée automatiquement | Encodage de la seule colonne qualitative : number\_sta |

# Modélisation

En l’état actuel du jeu de données, nous avons réussi à entrainer plusieurs modèles de machine learning via la bibliothèque SKLearn et voici les résultats (MAPE) :

*Figure 11 : Boxplot MAPE (sans AROME et ARPEGE) (N\_FOLDS = 10)*



Les meilleurs résultats sont ceux du KNRegressor, RandomForestRegressor, ExtraTreesRegressor, à savoir que le SGD ainsi que le BAGGING supposés être aux alentours de 35 ne prédisent en réalité que 0 (impertinent) :

|  |  |
| --- | --- |
| RandomForestReg |  |
| SGDReg |  |
| ExtraTreesReg |  |

# Introduction des datasets AROME et ARPEGE (site MeteoNet)

L’introduction de ces deux datasets a joué un rôle majeur dans les performances de nos prédictions :



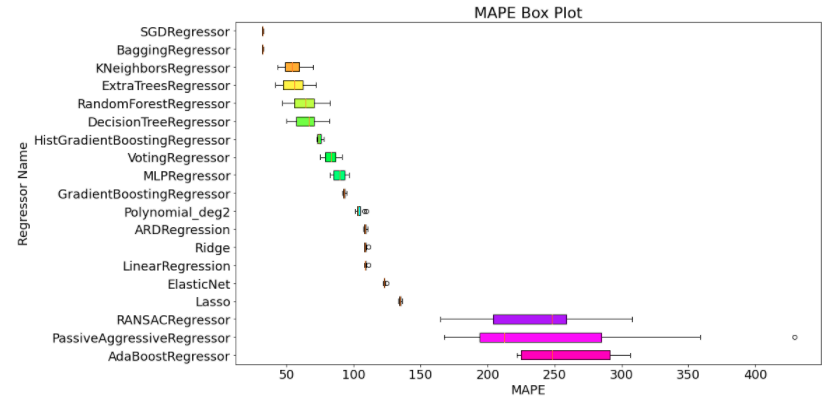
Une image contenant table

Description générée automatiquement

De 11 features nous sommes passés à 28. Ce qui a considérablement augmenté nos possibilités. Nous avons donc commencé par dresser une table des corrélations avec la Target :

|  |  |
| --- | --- |
|  | Il en ressort qu’une bonne partie des variables du jeu de données ARPEGE semblent importants pour prédire notre cible. |

Avec les mêmes modèles de machine learning que sur le jeu de données initial, voici le boxplot des MAPE que nous avons retenu :



Les valeurs sont bien meilleures, nos algorithmes tournent à présent autour de 50 (contre 100 pour le dataset initial)

Nous avons ensuite effectué de la features importance selon les modèles que nous testions, en voici quelques exemples qui nous montrent la variation de la MAPE en fonction des features retenues par ordre d’importance :

|  |  |  |
| --- | --- | --- |
|  | | Distribution Prediction |
| DecisionTreeReg |  |  |
| RandomForestReg |  |  |

Globalement la MAPE se stabilise pour les 21 plus importantes features, à savoir :

['dd', 'hu', 't', 'ff', 'precip', 'month', 'ws', 'p3031', 'u10', 'v10', 't2m', 'r', 'msl', 't\_3D\_arpege', 'r\_3D\_arpege', 'ws\_3D\_arpege', 'p3031\_3D\_arpege', 'u\_3D\_arpege', 'v\_3D\_arpege', 'w\_3D\_arpege', 'z\_2D\_arpege']

# Le Deep Learning

Le modèle nous ayant donné jusqu’alors les meilleurs résultats reste le réseau de neurones profond. En effet, alors que les modèles de Machine Learning classiques, même avec les hyper-paramètres optimisés peinaient à descendre en deçà de 35 de MAPE. Or grâce aux réseaux de neurone, **nous avons réussi à atteindre 26 de MAPE** pour le validation-set. Voici comment nous avons procédé :

from sklearn.model\_selection import train\_test\_split

trainset, testset = train\_test\_split(trainset, test\_size=0.3)

X\_train = trainset.drop(['Ground\_truth','Id'],axis=1)

y\_train = trainset['Ground\_truth']

X\_test = testset.drop(['Ground\_truth','Id'],axis=1)

y\_test = testset['Ground\_truth']

ann = Sequential()

ann.add(BatchNormalization())

ann.add(Dense(350, activation="relu", kernel\_initializer='normal'))

ann.add(Dropout(0.3))

ann.add(Dense(512, activation="relu", kernel\_initializer='normal'))

ann.add(Dropout(0.3))

ann.add(Dense(512, activation="relu", kernel\_initializer='normal'))

ann.add(Dropout(0.3))

ann.add(Dense(128, activation="relu", kernel\_initializer='normal'))

ann.add(Dense(1))

ann.compile(loss='mean\_absolute\_percentage\_error', optimizer='adam', metrics=['mean\_absolute\_percentage\_error'])

history = ann.fit(X\_train,y\_train,epochs=2048, batch\_size=8192, verbose = 2 ,validation\_data=(X\_test,y\_test))

history\_dict = history.history

Voici le résultat obtenu :

|  |  |
| --- | --- |
| **ann train MAPE : 20.94**  **ann val MAPE : 25.25** |  |

Voici également les prédictions effectuées sur le jeu de validation comparé à la labellisation :

Une image contenant texte, ciel, jour, plusieurs

Description générée automatiquement

Nous avons donc réussi à créer un modèle qui n’overfit pas nos données et qui a une performance pouvant nous propulser dans le top10 du leaderboard sur Kaggle (sur 68 équipes) :

Une image contenant texte

Description générée automatiquement