AIF
Prédiction de cumuls de pluie

Juan AYALA, Jeong Hwan KO, Alice LALOUE, Aldo MELLADO AGUILAR, Nicolas PREVOT
5A MA
2021 - 2022

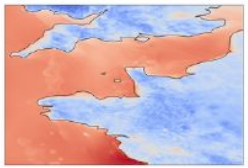
Sommaire

1. Objectifs
2. Données
3. Pre-processing
4. Feature Engineering
5. Modèles de machine learning
6. Apprentissage
7. Scores

Objectifs

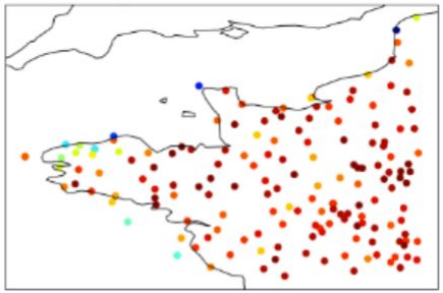
But : Prédire le cumul sur 24h de pluie pour le lendemain et pour les stations données.

Approche : Correction de la prédiction du modèle de Météo-France en fonction des conditions météorologiques.



+

->



On pose :

Y : Erreur commise par le modèle sur le cumul de pluie

Y = Ground truth - baseline forecast

Cumul : baseline forecast + Y

Objectifs

Score à minimiser :

Mean Average Percentage Error (MAPE) : $$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Améliorer baseline forecast :

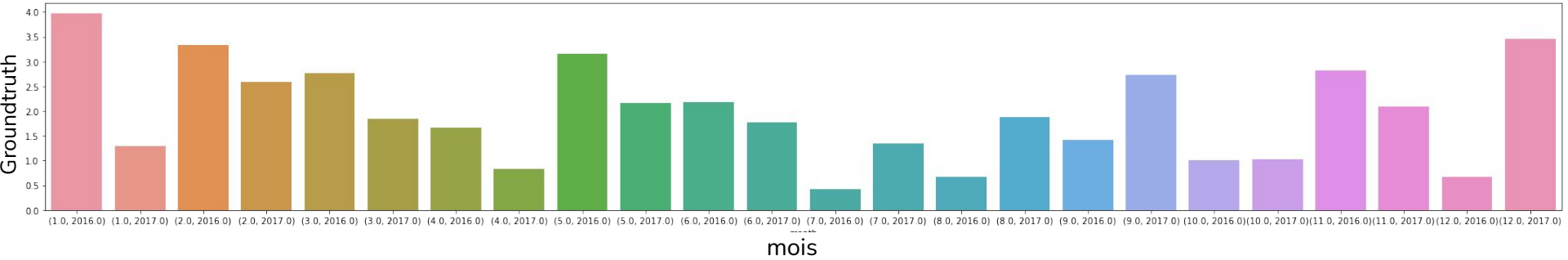
52	Baseline_forecast		47.51932	1	2mo

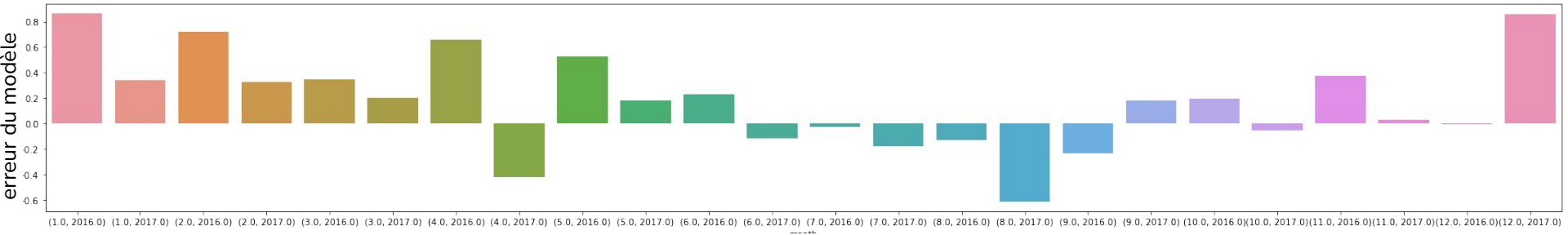
Données

- **Ground truth (Données réelles de cumul de pluie)**
- **Baseline observée et Baseline prédite**
- **Données horaires des stations de la veille (D-1)**
  - Température, Température de rosée, Vitesse et direction du vent, Précipitations, Humidité
- **Données 3D horaires du modèle arpège pour le lendemain (D)**
  - Température potentielle à 850 hPa, Température à 500 hPa, Géopotentiel à 500 hPa, Humidité relative à 700 hPa, Vitesse du vent à 1000 hPa, Vitesse verticale à 950 hPa.
- **Données 2D horaires du modèle arpège pour le lendemain (D)**
  - Vent zonal à 10m, Vent méridien à 10m, Température à 2m du sol, Humidité relative, Direction du vent, Température de rosée à 2m, Pression au niveau de la mer, Vitesse du vent, Précipitation totale

Données

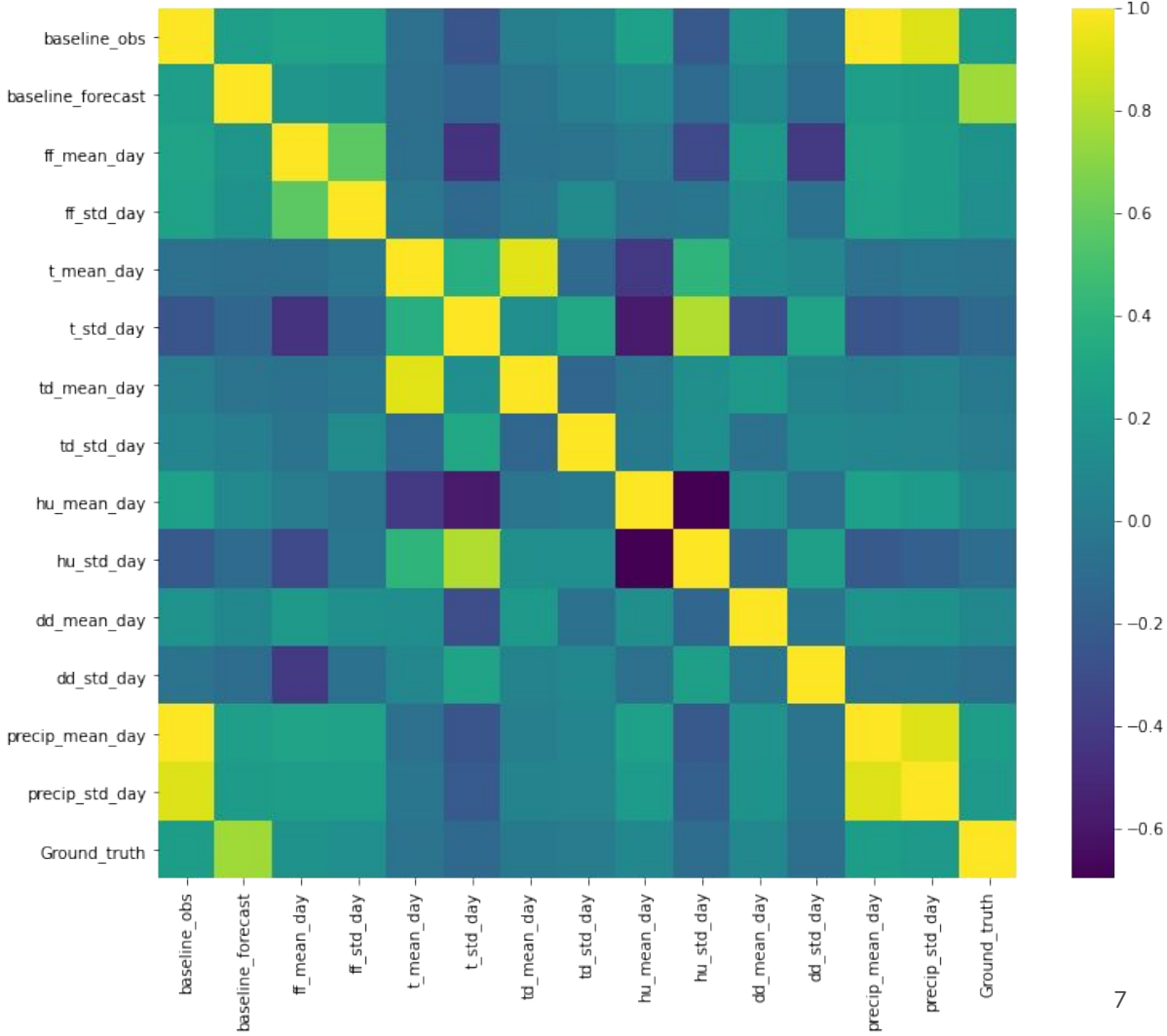
1. Statistiques descriptives





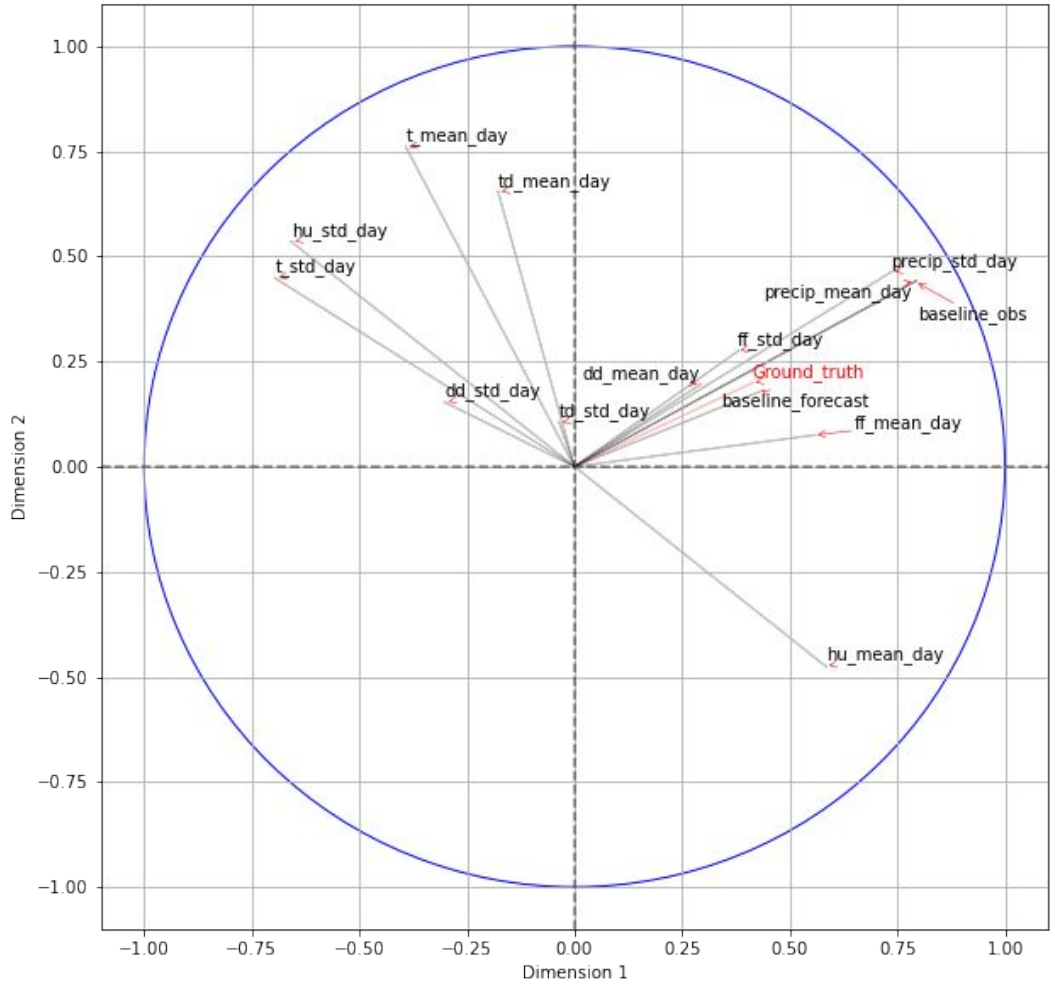
Données

2. Corrélogramme des variables des stations



Données

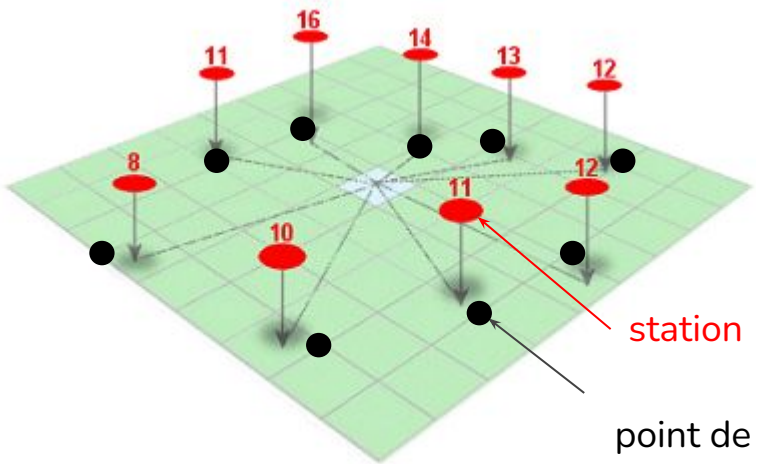
3. Cercle de corrélations (ACP) pour les variables des stations



Preprocessing

1. Récupération des données modèles pour chaque station

Valeur du point de grille le plus proche de la station au temps t



$$d = r\,\mathrm{hav}^{-1}(h) = 2r\arcsin\left(\sqrt{h}\right)$$

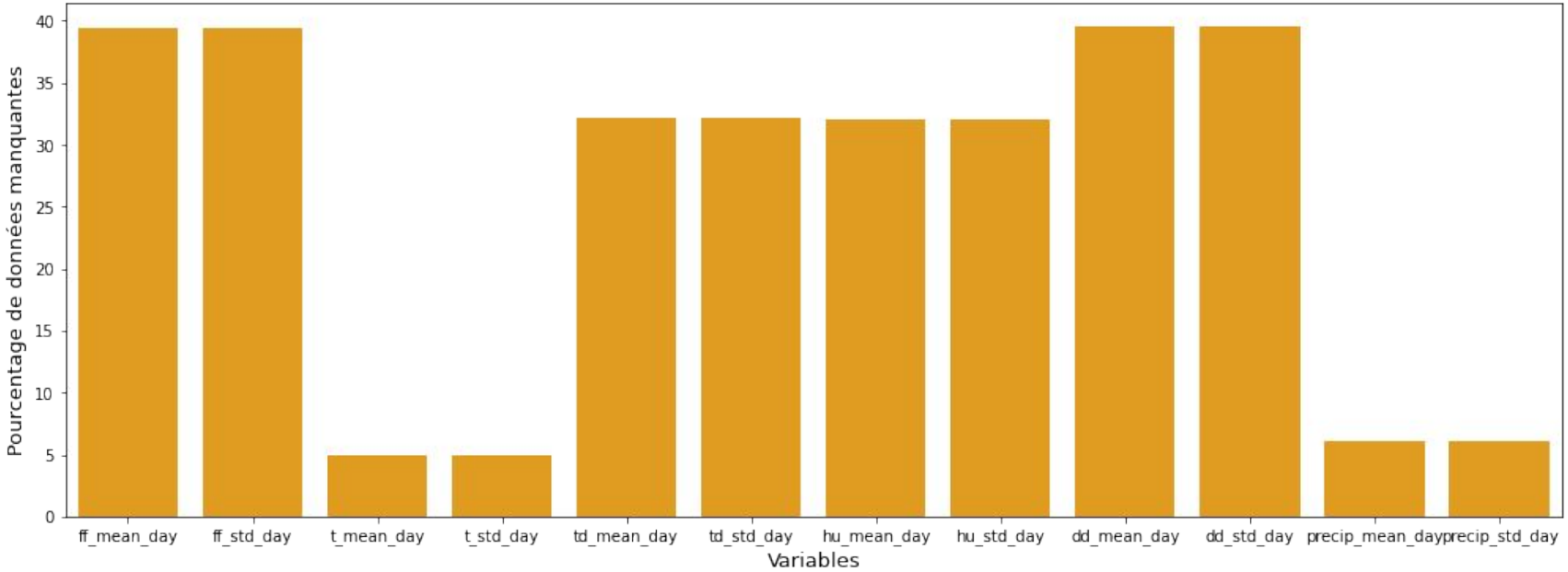
où h vaut $\mathrm{hav}(\frac{d}{r})$, ou plus explicitement:

$$d = 2r\arcsin\left(\sqrt{\mathrm{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\,\mathrm{hav}(\lambda_2 - \lambda_1)}\right)$$

$$= 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Données

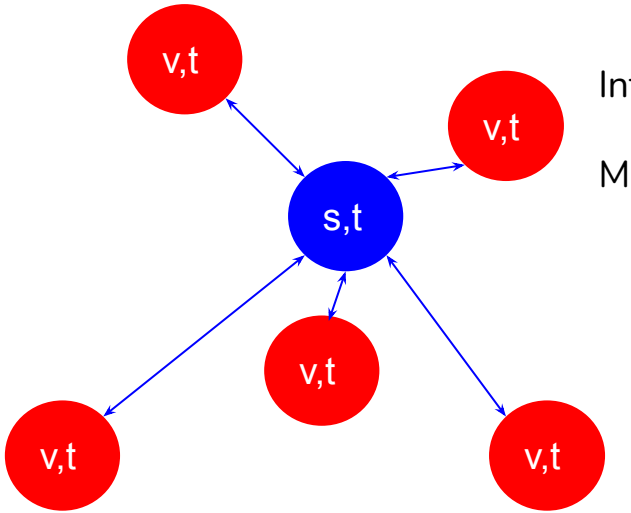
Pourcentages de données manquantes pour les données stations



Preprocessing

2. Traitement des données manquantes

Données stations : Valeurs manquantes globalement aléatoirement

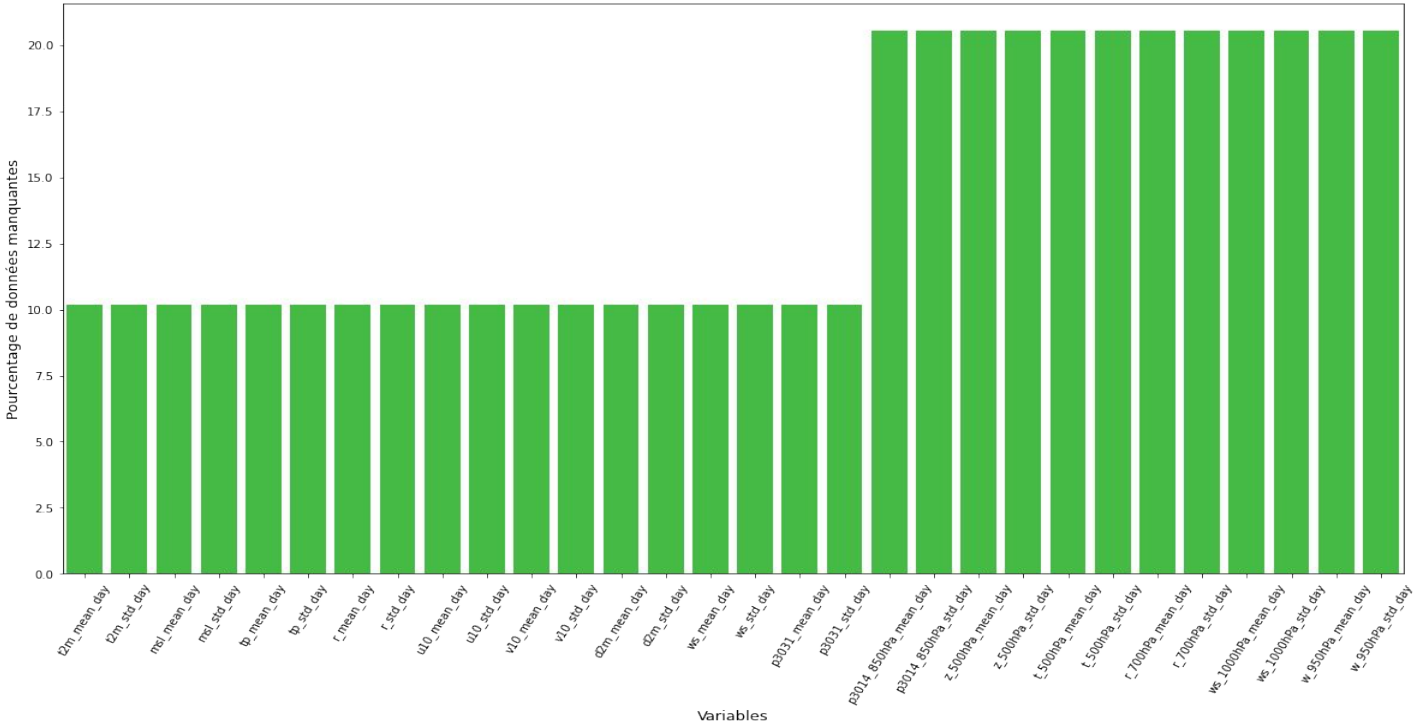


Interpolation spatiale des données des stations :

Moyenne des 5 stations les plus proches pondérées par la distance.

Données

Pourcentages de données manquantes pour les variables de 2D et 3D arpège

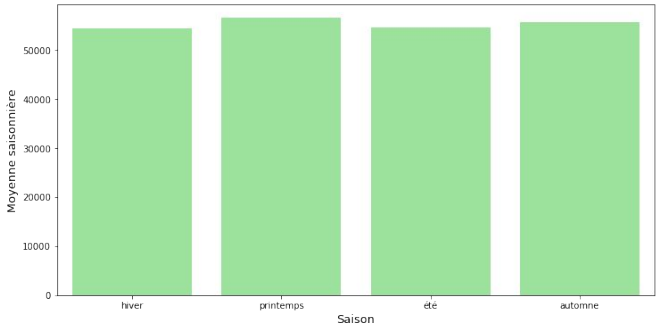


Preprocessing

2. Traitement des données manquantes

Données 3D arpège : Valeurs manquantes liées à l'absence de données certains jours entiers.

Remplissage par la moyenne saisonnière de chaque paramètre.



Données 2D arpège : Assez peu de données manquantes par rapport aux autres.

Abandon des lignes présentant des données 2D arpège manquantes.

Feature Engineering

Transformation des données

Données horaires :

- **Moyennes journalières**
- **Écart-types journaliers**

ff, td, t, hu, dd, precip sur 24h → moyenne et écart-type sur les 24h

t 500hPa, r 700 hPa, z 500hPa, p3014 850hPa, ws 1000hPa, w 950hPa sur 17 valeurs de temps → moyenne et écart-type sur les 17h

d2m, t2m, u10, v10, tp, p3031, msl, ws, r → moyenne et écart-type sur les 24h

Modèles de Machine Learning

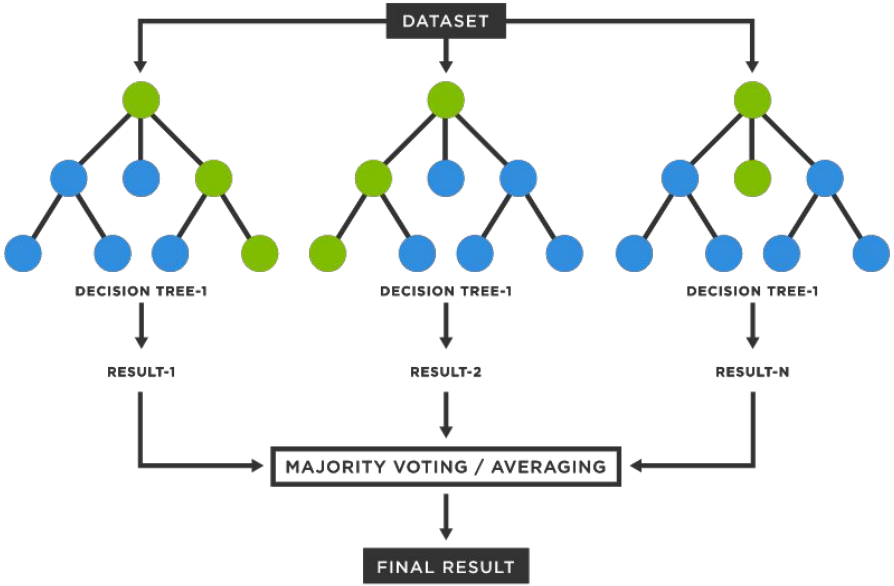
Voici les modèles de Machine Learning que nous avons implémentés :

- Random Forest
- XGBoost
- SVR
- MLP

Modèles de Machine Learning

Random Forest avec scikit-learn

- 100 arbres de décision
- max features = ‘auto’
- profondeur maximale



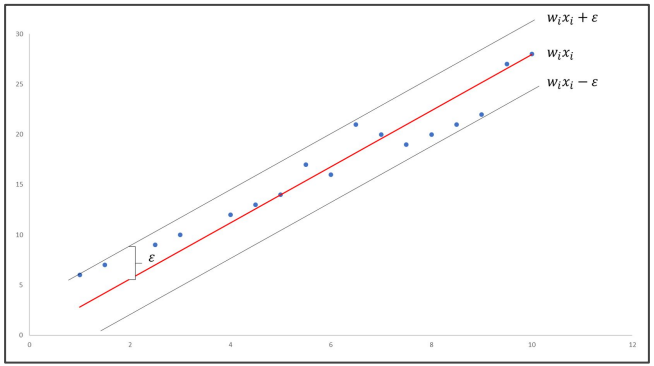
Modèles de Machine Learning

XGBoost

- 1000 arbres
- Profondeur maximale = 200
- Taux d’apprentissage de boosting = 0,1
- Ratio de sous-échantillon de l'instance d'entraînement = 0,7
- Ratio de sous-échantillon de colonnes pour chaque arbre = 0,8

Modèles de Machine Learning

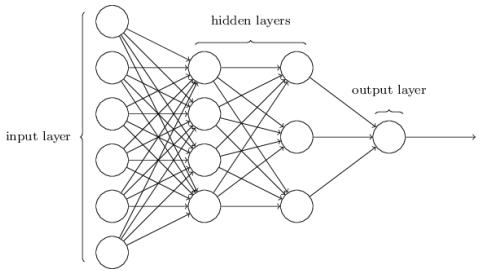
Linear SVR avec scikit-learn



Hyperparamètre ajusté : tolérance = 10^{-5}

Modèles de Machine Learning

Multi-Layer Perceptron (MLP) ou réseau de neurones avec tensorflow / keras



Notre modèle est constitué de :

- 20 couches cachées, avec fonction d'activation ReLU et initialiseur de noyau 'he_uniform'
- 32 neurones dans chacune des couches
- Couche de sortie avec un neurone et activation linéaire car problème de régression
- Fonction de perte : *mean absolute error*

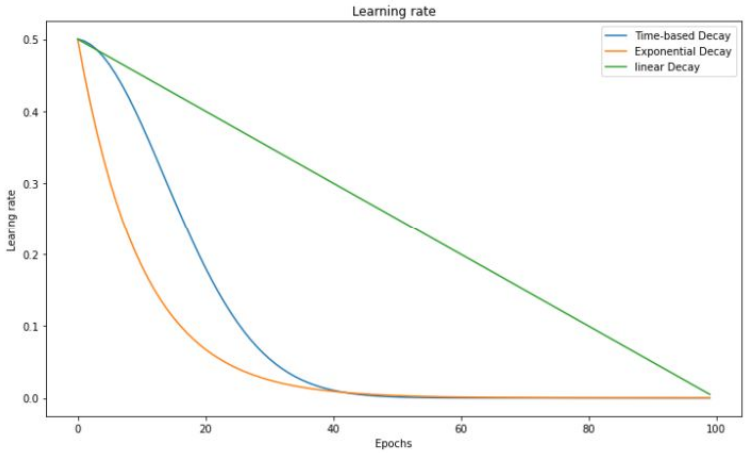
Modèles de Machine Learning

Multi-Layer Perceptron (MLP) ou réseau de neurones avec tensorflow / keras

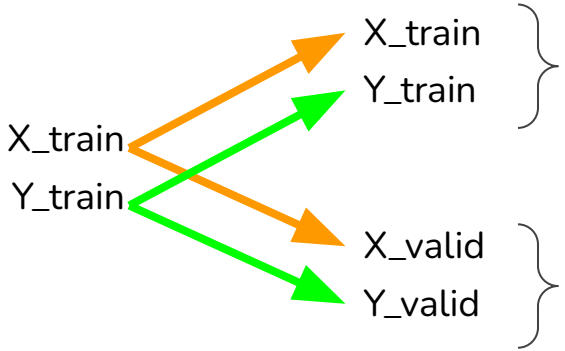
Taux d'apprentissage choisi : ~~diminution exponentielle~~

constante (0.001)

Optimiseur choisi : ~~SGD~~ RMSProp



Apprentissage



Entrainement du modèle

Validation du modèle

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

	Random Forest	XGBoost	MLP	SVR
MAPE	35	24	39	41

Optimisation par GridSearch (CV=3)

Loss function = MAPE

Scores

Modèle	MAPE
Baseline forecast	47.52
Random Forest	41.18
XGBoost	44.37
SVR	41.36
MLP	40.26