Prédiction de SO_2

Data Challenge Veolia

Fancello Marie Clara, Germini Eva

Avril 2022

Objectif et périmètre

OBJECTIF :

Prédire les émissions de dioxyde de soufre à la station MAS dans le Havre sur les 12 prochaines heures

DONNEES A DISPOSITION :

- Émissions de S02 à d'autres stations Normandes
- Coordonnées GPS
- Informations météorologiques
- Classes d'occupation des sols

Plan de la présentation

1. Présentation des données
2. Analyse descriptive: identification de deux problématiques majeures
3. Présentation des modèles et de leurs performances associées

Description et traitement des données

Dimensions:
- 6089 observations dans le dataset d'entraînement
- 168 observations dans le dataset de test *(environ 3% du train)*

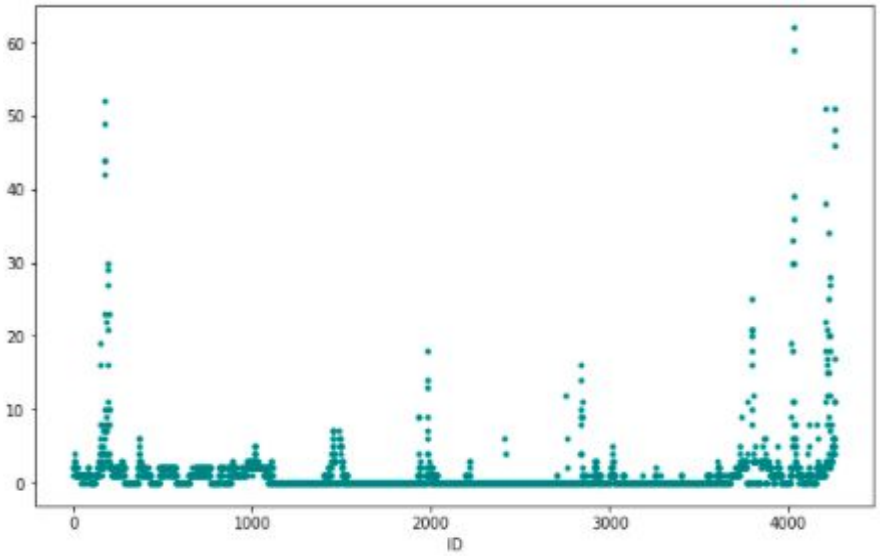
Un jeu de données ayant 2064 variables :
- 1776 variables numériques dont 672 variables ayant une seule valeur
- 288 variables catégorielles, ayant toutes une seule valeur

Peu de données manquantes :
- 0.03% dans le dataset d'entraînement imputées par la méthode *backward*
- 0.00% dans le dataset de test

Analyse des variables expliquées: SO2 MAS

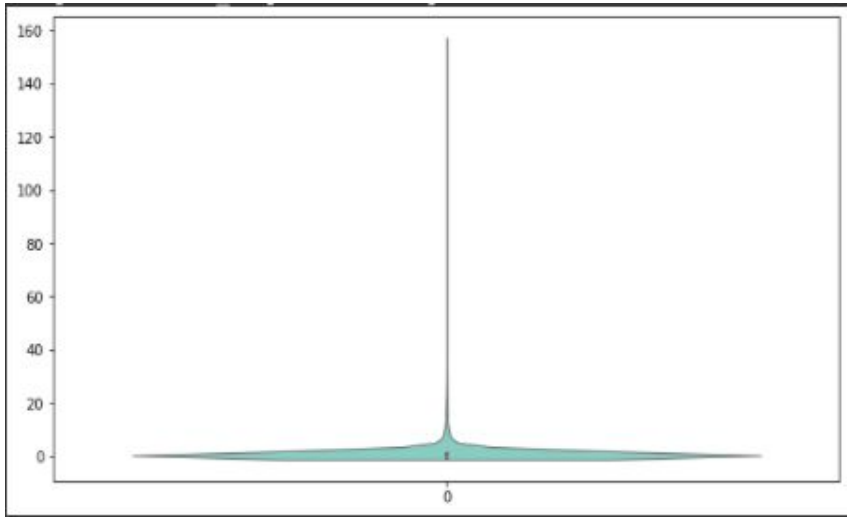
- Pas de tendance, ni de saisonnalité
- En présence d'outliers, l'émission de SO2 reste forte quelques heures.
- **Première problématique:** forte concentration de 0

Evolution des émissions de SO2 au cours du temps



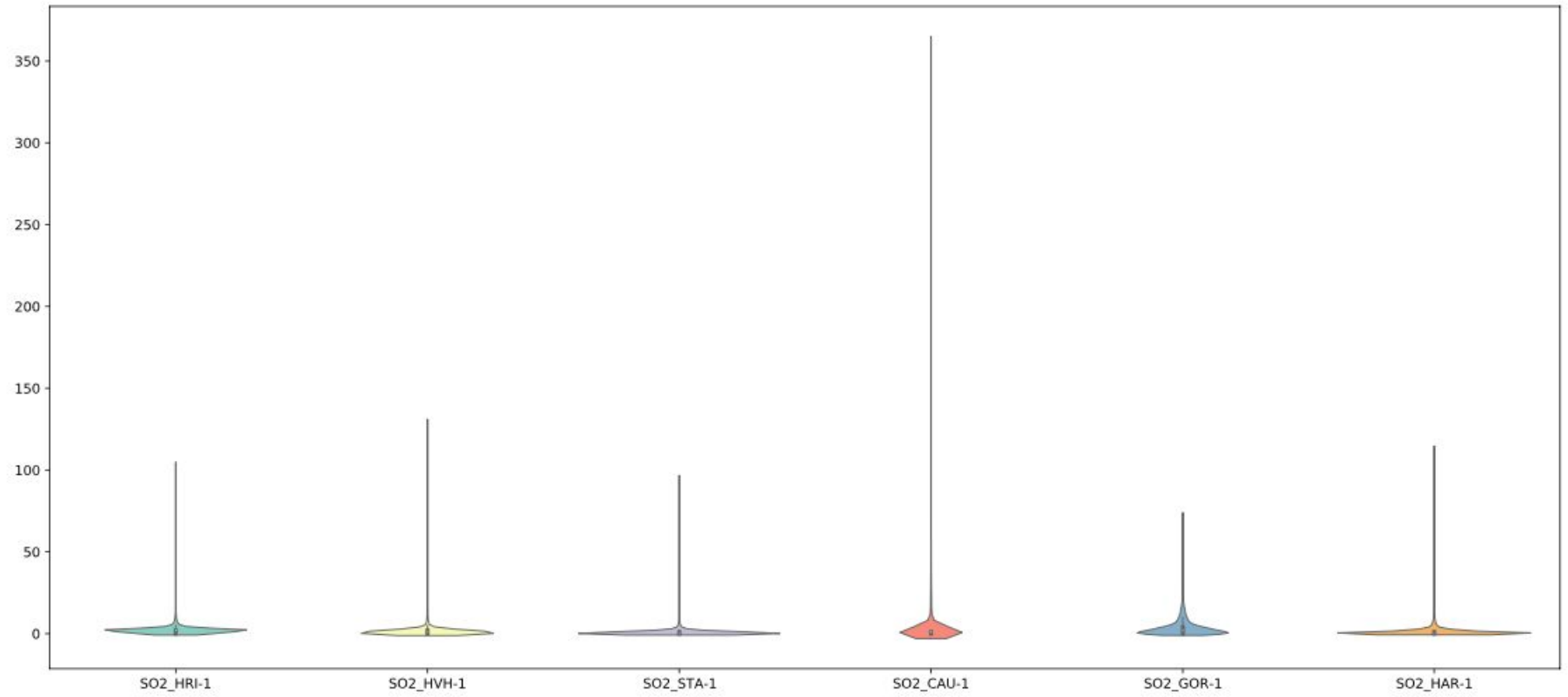
	SO2_MAS+0
count	6089.0
mean	1.33
std	5.86
min	0.0
25 %	0.0
50 %	0.0
75 %	1.0
max	156.0

Distribution des émissions de SO2



Émissions de SO2 aux autres stations

1 Analyse de la distribution des variables à l'instant h-1



Émissions de SO2 aux autres stations

2 Analyse de leurs relations avec les variables SO2 MAS

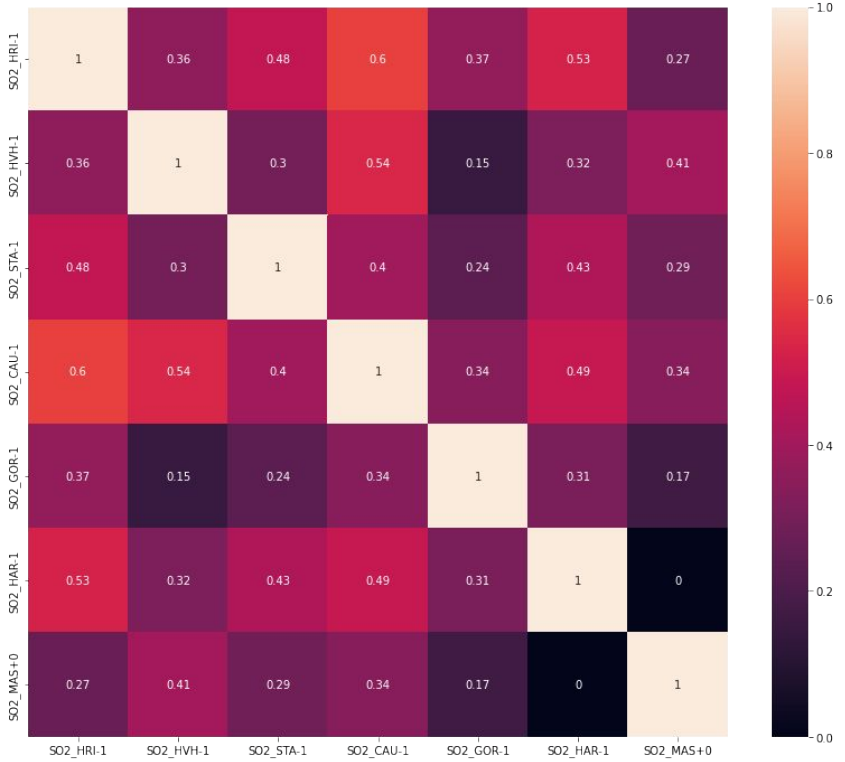
Distance avec la station MAS:

1. CAU
2. HRI
3. HVH
4. STA
5. HAR
6. GOR

Nombre d'outliers communs:

1. HVH et STA: 53
2. HRI: 46
3. CAU: 32
4. HAR: 14
5. GOR: 10

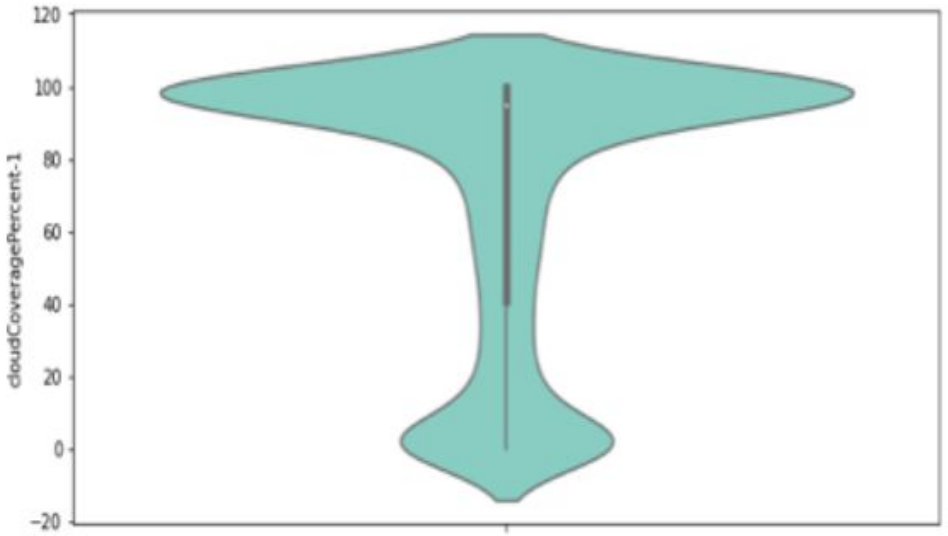
Heatmap des corrélations



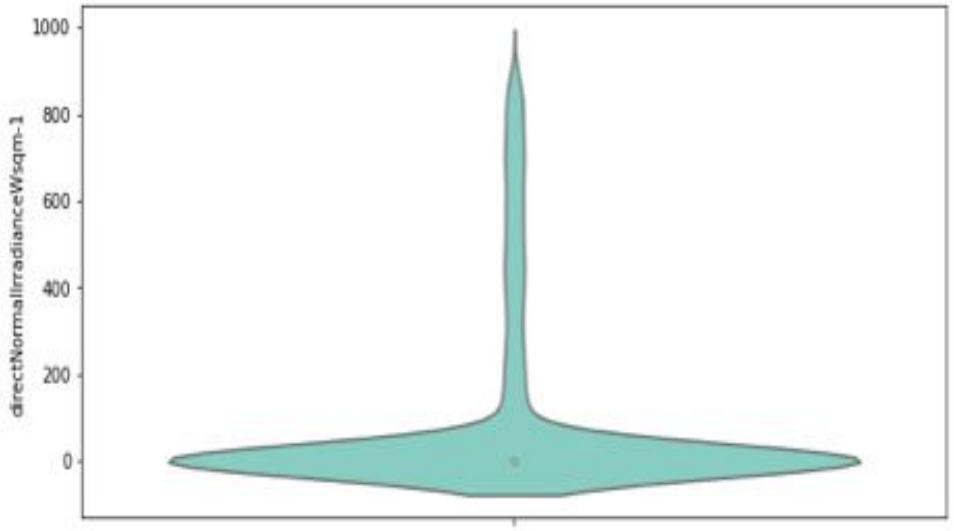
Etude des autres variables explicatives

- Corrélations très faibles entre les variables d'input et d'output
- **Seconde problématique:** aucune relation claire n'est identifiable entre les variables explicatives et les variables expliquées

Distribution de la couverture nuageuse



Distribution des radiations directes du soleil



Séparation des données

Séparation de la base d'entraînement :

- **De manière aléatoire**
- **90%** alloués à la base d'entraînement: *5480 observations*
- **10%** alloués à la base de validation: *609 observations*

Présentation des modèles

1. Régression linéaire avec pénalité “ElasticNet”
2. Modèle de Boosting: CatBoost
3. Modèles de Deep Learning: LSTM et GRU

Présentation des résultats

Modèle retenu: LSTM

MSE	ElasticNet	CatBoost	LSTM	GRU	Benchmark Veolia
Train	20.62	10.64	21.93	24.82	NR
Valid	14.51	10.57	14.10	14.99	NR
Test	45.78	47.25	43.02	49.03	46.34

Merci pour votre attention !