

Notes d’avancement

Groupe : 4AM\_SID



Master 2 – Statistiques et informatique décisionnelle

Université Paul Sabatier

Version 1.0

1. **Introduction :**

Le présent document a pour objet de présenter l’état d’avancement du Challenge Big Data 2018 organisé pour les étudiants en Sciences des Données de niveau Master du site toulousain, défi élargi cette année à Bordeaux et Rennes.

Le nouveau défi, est le fruit d'une collaboration entre les enseignants / chercheurs du département GMM (INSAT) et des chercheurs de Météo-France. Il portera sur des prévisions de températures à court terme dans 7 villes de la métropole. Ces prévisions ont un impact direct sur celles des consommations électriques avec des conséquences économiques majeures pour ERDF.

La compétition se déroulera du 5 octobre au 16 janvier 2018.

Présentation du jeu de données :

Présentation des variables :

|  |  |
| --- | --- |
| Variables | Description |
| tH2\_obs | Observation de la température à 2 mètres in situ- au point station (prédicant) |
| ech | Echéance de validité |
| capeinsSOL0 | Energie potentielle convective |
| ciwcH20 | Fraction de glace nuageuse à 20 mètres |
| clwcH20 | Fraction d'eau nuageuse à 20 mètres |
| nH20 | Fraction nuageuse à 20 mètres |
| pMER0 | Pression au niveau de la mer |
| rr1SOL0 | Précipitation horaire au niveau du sol |
| rrH20 | Précipitation horaire à 20 mètres |
| tpwHPA850 | Température potentielle au niveau 850 hPa |
| ux1H10 | Rafale 1 minute du vent à 10 mètres composante zonale |
| vapcSOL0 | Colonne de vapeur d'eau |
| vx1H10 | Rafale 1 minute du vent à 10 mètres composante verticale |
| ddH10\_rose4 | Direction du vent à 10 mètres en rose4 |
| ffH10 | Force du vent à 10 mètres en m/s |
| flir1SOL0 | Flux Infra-rouge en J/m2 |
| fllat1SOL0 | Flux de chaleur latente en J/m2 |
| flsen1SOL0 | Flux de chaleur sensible en J/m2 |
| flvis1SOL0 | Flux visible en J/m2 |
| hcoulimSOL0 | Hauteur de la couche limite en mètres |
| huH2 | Humidité 2mètres en % |
| iwcSOL0 | Réservoir neige kg/m2 (équivalent en eau liquide des chutes de neige) |
| nbSOL0\_HMoy | Nébulosité basse (moyenne sur les 6 points de grille autour de la station) (fraction en octat du ciel occulté) |
| ntSOL0\_HMoy | Nébulosité totale (moyenne sur les 6 points de grille autour de la station) |
| tH2 | Température à 2 mètres du modèle AROME |
| tH2\_VGrad\_2.100 | Gradient vertical de température entre 2 mètres et 100 mètres |
| tH2\_XGrad | Gradient zonal de température à 2 mètres |
| tH2\_YGrad | Gradient méridien de température à 2 mètres |
| mois | Le mois |

A méditer :

Comment gérer les périodes manquantes ?

1. **Journal de bord**
   1. **Etape 1 – Pre-Processing : 13/10 - ?**

La première étape consiste à préparer nos données pour pouvoir leur appliquer nos algorithmes. On parle alors de nettoyage de données en plusieurs sous-étapes :

* Fixer les problèmes de formats :
* Lorsqu’on import avec la librairie Pandas le jeu de données, on observe que les valeurs ont un séparateur décimal différent. Par exemple la variable « tH2 » a une virgule comme séparateur décimal alors que la variable « ddH10\_rose4 » a un point comme séparateur décimal.

Le nettoyage ici consistera à corriger cela de façon à ce que toutes nos variables soient reconnus comme « Float »

* La colonne « date » représente la date au format JJ/MM/AAAA, or ce format n’est pas exploitable. Pour le gérer, deux solutions s’offre à nous :
* L’utilisation de variables ordinales,
* L’utilisation du Timestamp format pour la création de nouvelles colonnes comme : le jour, le mois, l’année, la semaine de l’année, le jour de l’année…

Les deux solutions vont être mis en place dans notre fichier, mais une seule sera utilisé dans le modèle. L’idée derrière est de pouvoir dire laquelle est la mieux.

* Gestion des données manquantes :

Il est fréquent que certaines valeurs soient manquantes dans les jeux de données. Cela signifie généralement qu'une information n'a tout simplement pas été recueillie.

* Quand on regarde le 1er jeu de données on se rend compte que plusieurs informations manquent et que plusieurs colonnes sont concernées (dans ce premier cas on ne parle pas des périodes manquantes mais uniquement des valeurs enregistrées manquantes). Il existe plusieurs options pour gérer les données manquantes mais celle qui aura le plus de sens pour notre modèle sera de les remplacer par la moyenne de la colonne. De plus, sachant qu’on parle ici de mesure sur des lieux différents et que la notion de date est très importante, on ne prendra pas toute la colonne pour calculer sa moyenne mais uniquement les données relatives à un code INSEE et a un mois donné. Prenons l’exemple de la variable « tH2 » qui corresponds à la température à 2 mètres du modèle AROME, pour remplacer la valeur manquante du 15/01/2015 de la station de Nice on commencera par filtrer les données en ne sélectionnant que le mois de Janvier ainsi que le code INSEE pour la station de Nice, il ne restera qu’à calculer la moyenne de « tH2 » de remplir les données manquantes avec cette valeur. Pour faire cela, c’est la fonction « groupby » et « transform » que nous avons utilisé.
* Gestion des valeurs manquantes sur les périodes
* Correction des valeurs incorrectes :

Pour certaines colonnes, certaines valeurs peuvent être identifiées comme manifestement incorrectes. Dans notre cas, de telles valeurs n’ont pas été identifié, on suppose alors que toutes les valeurs sont correctes.

* Normalisation :