Rapport projet Data Visualisation &

Data Mining

Créé par :   
-Alaoui Mdaghri Ahmed  
-Khoulti Wissal   
-Ben Azzouz El Mehdi   
-Chafik Ayman



# Introduction

|  |
| --- |
| Dans le cadre de notre projet Data Visualisation et Data Mining on à choisit le thème environnement pour cadrer notre travail et choisir une dataset.  La dataset à été mise en disposition dans le cadre d’une compétition de machine learning dont l’objectif était de proposer une prédiction comme solution pour le problème, elle regroupe différente observation de météo selon 5 régions, avec plusieurs features comme la moyenne de vitesse de vent ou la direction de vent et aussi un label qui regroupe les différents résultats de ces observations selon 3 modalités  (N : pas de pluie, L : pluie modéré, H : forte pluie) |
| Pour remédier à cela on poursuivra à appliquer les différentes méthodes et stratégies mises à notre disposition dans les cours de datamining et datavisualisation pour explorer les données et par la suite créer un modèle de machine learning pour prédire la sévérité de pluie. |

# Data Exploration and Visualisation

Pour visualiser les données et explorer nos features on utiliseras l’outil PowerBI et puis Matplotlib sous Python

Dans un premier temps on visualise les données avec PowerBI :



Problématique :

La pluie a un impact significatif sur l'industrie agricole et à cause de cela, être capable de la prédire, aide les agriculteurs dans leurs décisions quotidiennes telles que la façon de planifier efficacement, de minimiser les coûts et de maximiser les rendements.

Une grande entreprise agricole a besoin de vous pour l'aider à maximiser l'efficacité de sa croissance, à économiser les ressources et à optimiser sa production.

Pour y parvenir, l'entreprise doit disposer d'un algorithme de prévision météorologique précis qui améliorera sa prise de décision sur les activités agricoles typiques telles que la plantation et l'irrigation.

À l'aide des informations météorologiques historiques de leur région, **pouvez-vous prévoir le temps qu'il fera dans les prochains jours** ?

Nous avons maintenant un objectif clair.

Le but 🥅

Prédisez la météo du lendemain sur la base de trois étiquettes.

N- Pas de pluie

L- Pluie légère

H— Fortes pluies

Sources de données :

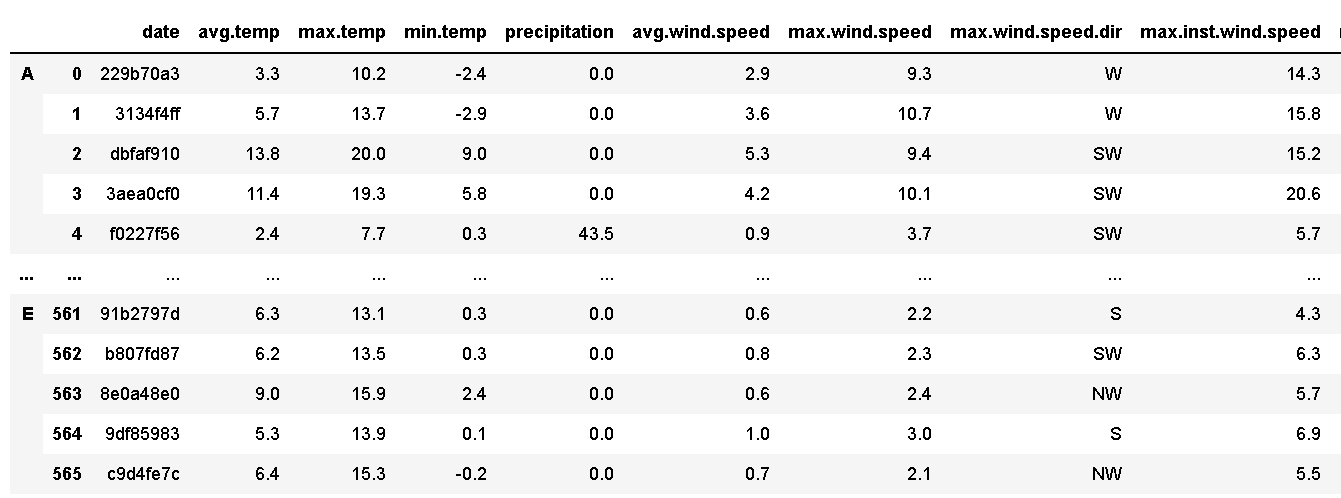
📂 **train**  
 ├── region\_A\_train.csv   
├── region\_B\_train.csv   
├── region\_C\_train.csv   
├── region\_D\_train.csv   
├── region\_E\_train.csv   
├── solution\_format.csv   
└─.─📂 **test**  
 ├── region\_A\_test.csv   
├── region\_B\_test.csv   
├── region\_C\_test.csv   
├── region\_D\_test.csv   
└── region\_E\_test.csv

# Data Exploration and Cleaning

Les données ont été commodément divisées en ensembles de données d'entraînement et de test.

Dans chaque train et test, vous recevez des données météorologiques composées d'emplacements anonymisés nommés de la région A à la région E, qui sont toutes des régions voisines.

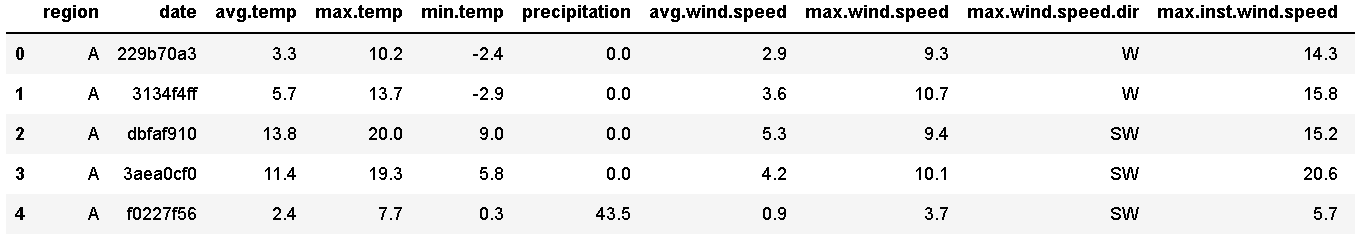
Voici un aperçu des cinq premières rangées deregion\_A\_train.csv



La première chose que vous devriez remarquer est que la date colonne n'est pas une date mais a été anonymisée pour être une valeur aléatoire.

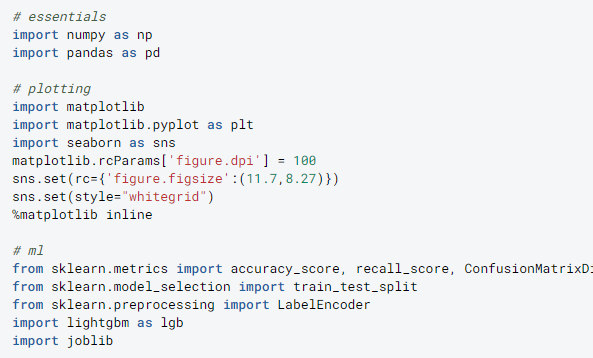
Il y a un total de 10 caractéristiques, qui sont composées de la température, des précipitations, de la vitesse du vent, de la direction de la vitesse du vent et de la pression atmosphérique

Puis, en regardant solution\_format.csv



## Charger les bibliothèques

Ensuite, nous chargeons quelques bibliothèques essentielles pour les visualisations et l'apprentissage automatique.



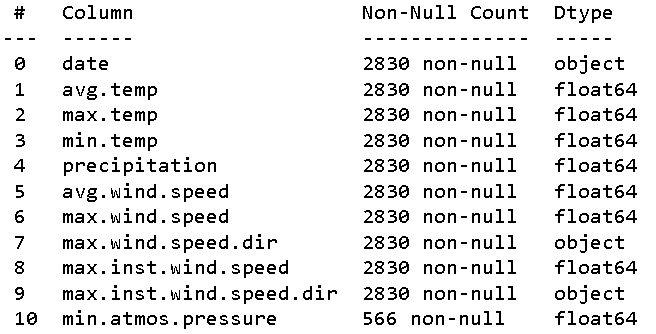
## Charger les données

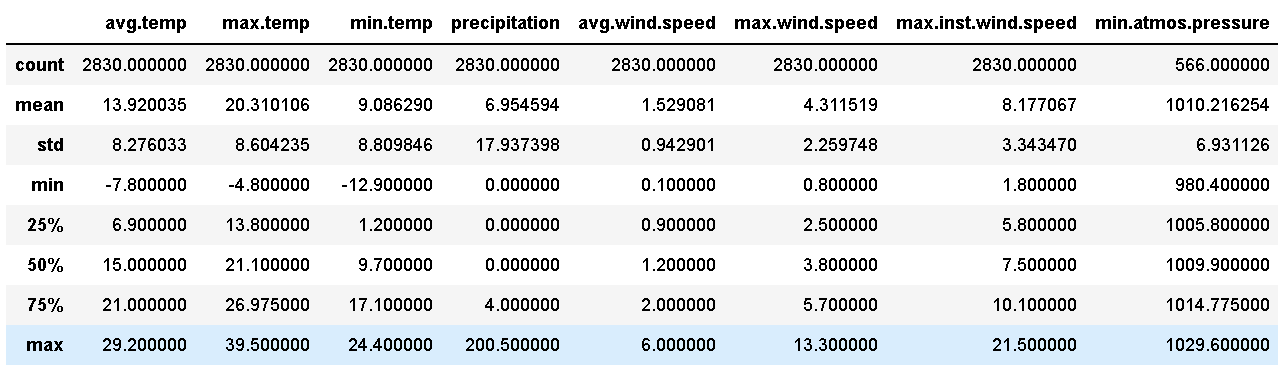
Lisons toutes les données dont nous disposons.

Une image contenant texte

Description générée automatiquement

Par la suite en pourra afficher le type de colonnes et des statistiques :





On veut pas que les régions soient l'index, nous pouvons donc réinitialiser l'index, puis renommer certaines colonnes pour obtenir les données dans la bonne forme.

Une image contenant table

Description générée automatiquement

## Répartition de la classe cible

Visualisons d'abord notre classe cible.



Il semble que nous ayons entre les mains une classe déséquilibrée, car l’étiquette domine le reste des classes.

Pourquoi c'est un problème ?

Le modèle sera biaisé vers les classes avec une plus grande quantité d'échantillons.

Cela se produit parce que le classificateur a plus d'informations sur les classes avec plus d'échantillons, il apprend donc à mieux prédire ces classes tout en restant faible sur les classes plus petites.

Dans notre cas, l'étiquette Nsera prédite plus que les autres classes.

Gardons cela à l'esprit et passons à d'autres visualisations.

## Tracer les fonctionnalités

Nous avons dix fonctionnalités dans chaque région.

Une image contenant texte

Description générée automatiquement

Il est temps de concocter quelques intrigues.

Étant donné que toutes les régions sont utilisées pour prédire la météo du lendemain, voyons si toutes les régions partagent des modèles similaires dans les entités et s'il existe des valeurs aberrantes ou des anomalies.

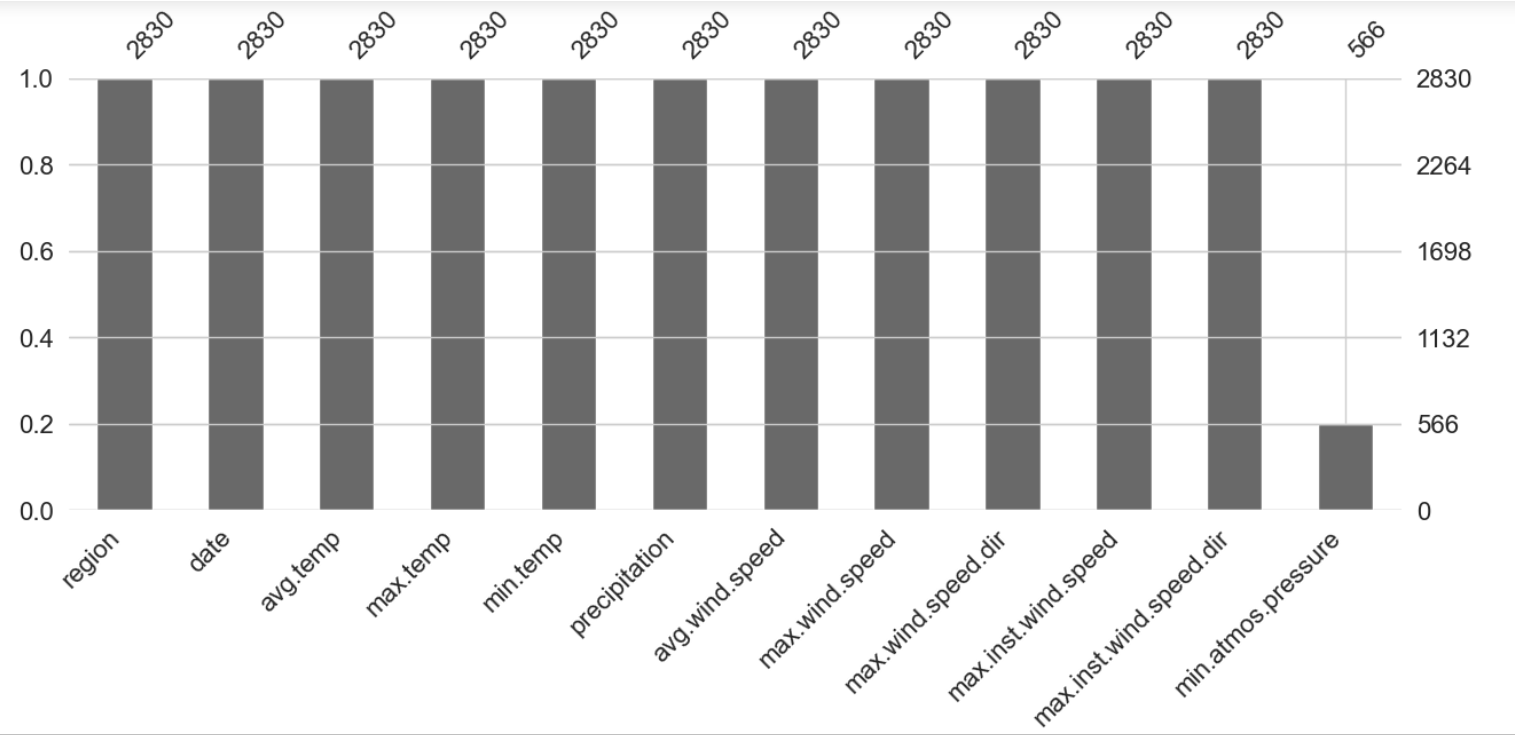


D'après les graphiques, nous voyons que les modèles dans les données sont très similaires à l'exception des régions C, D, E pour min.wind.speedet avg.wind.speedqui sont à l'échelle inférieure.

Maintenant que nous avons un peu exploré les données, nous vérifions les valeurs manquantes dans nos données.

# Valeurs manquantes

En utilisant ma fonction d'assistance personnalisée, il semble y avoir une grande quantité de données manquantes pour la min.atmos.pressure variable.



Regardons d'abord la distribution de la colonne avec des valeurs manquantes.

