03/11/2021

Mathys Clerget

Universite du Québec à Chicoutimi

Rapport TP2

Dataset DailyClimate

# Code Permanent :

Mathys Clerget : CLEM14050107

Contents

[Code Permanent : 1](#_Toc88691121)

[Introduction : 1](#_Toc88691122)

[Description : 1](#_Toc88691123)

[Collecte des données : 1](#_Toc88691124)

[Variables : 1](#_Toc88691125)

[Classes : 1](#_Toc88691126)

[Valeurs manquantes ou dupliquées : 2](#_Toc88691127)

[Étapes de création de l‘ensemble de données : 2](#_Toc88691128)

[Fenêtre temporelle glissante : 2](#_Toc88691129)

[Réduction de dimension : 2](#_Toc88691130)

[Conclusion : 4](#_Toc88691131)

[Conclusion générale : 4](#_Toc88691132)

[Source : 4](#_Toc88691133)

# Introduction :

Ce TP a été réalisé par Mathys Clerget avec comme support : un dataset nommé « Daily Delhi Climate ». C‘est en réalité deux dataset « Daily Delhi Climate Train» et « Daily Delhi Climate Test» qui sont regroupé afin d‘effectué un travail sur l’ensemble des données comprise entre 2013 et 2017.  
(2013-2016-Train et 2017-Test)

Ce dataset possède la température moyenne, l’humidité, la vitesse du vent, et la pression atmosphérique quotidienne de la ville de Delhi entre le 1er janvier 2013 et le 24 avril 2017.

# Description :

## Collecte des données :

Ce dataset provient de Weather Undergroud API.

Il est à disposition de tous sur Kaggle afin de travailler sur les Times Series. Il est à l’origine utilisé par l’Université de Bangalore pour son cours sur l’analyse de données.

Le premier fichier « Daily Delhi Climate Train » est composé de 1463 instances.

Le second fichier « Daily Delhi Climate Test» est lui composé de 115 instances.

## Variables :

Ce dataset possède 5 variables :

|  |  |  |
| --- | --- | --- |
| Variable | Description | Valeurs |
| Date | Date au format YYYY-MM-DD | Date |
| MeanTemp | Température moyenne calculée en fonction des résultats obtenus toutes les trois heures. L’unité n’est pas donnée mais on suppose quelle est en degré Celsius | Float |
| Humidity | L’humidité absolue de la journée. L’unité est le grammes de vapeur d'eau par mètre cube de volume d'air | Float |
| Wind\_Speed | La vitesse du vent. L’unite est le kilomètre par heure | Float |
| Mean\_Pressure | Pression atmosphérique. L’unite est l’atm, atmosphère normale | Float |

## Classes :

J‘ai décidé de séparer les classes par les différentes années que possède ce dataset. Il y a donc 5 classes différentes, les classes de 2013 à 2015 sont composées de 365 valeurs, la classe 2016 de 366 et la classe 2017 est, plus petit, composée de 115 instances.

## Valeurs manquantes ou dupliquées :

Il n‘existe aucune valeur manquante dans ce dataset.

Mais il existe une ligne dupliquée dans ce dataset : Les valeurs du premier janvier 2017. C’est la ligne qui joint les deux fichiers csv. Il a été décidé de conserver la deuxième ligne dupliquée.

# Étapes de création de l‘ensemble de données :

## Fenêtre temporelle glissante :

Deux fenêtres temporelles glissantes ont été créées, une première avec une fréquence d’échantillonnage de 30 éléments et un chevauchement de 23 éléments soit de 77 %.

La seconde fenêtre temporelle glissante avait une fréquence d’échantillonnage de 20 éléments et un chevauchement de 15 éléments soit de 75 %.

## Réduction de dimension :

Le script de réduction de dimension a été appliqué aux deux fenêtres temporelles glissantes précédemment créées. 16 caractéristiques ont été conservé. Et les résultats possèdent quelques différences.

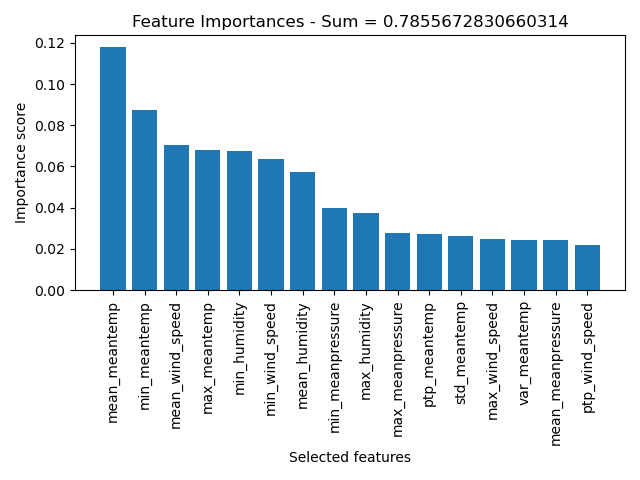
La première fenêtre atteint au score d’importance un maximum de presque 0.12 avec mean\_meantemp. La seconde, un maximum aux alentours de 0.10 avec max\_meantemp.

Les caractéristiques conservées sont les suivantes :

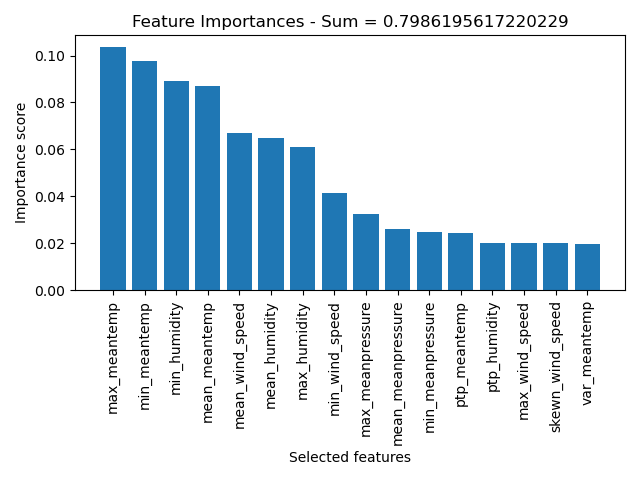
|  |  |
| --- | --- |
| Fenêtre de 30 échantillons, et chevauchement de 77% | Fenêtre de 20 échantillons, et chevauchement de 75% |
| * Mean\_meantemp * Min\_meantemp * Mean\_wind\_speed * Max\_meantemp * Min\_humidity * Min\_wind\_speed * Mean\_humidity * Mean\_meanpressure * Max\_humidity * Max\_meanpressure * Ptp\_meantemp * Std\_meantemp * Max\_wind\_speed * Var\_meantemp * Mean\_meanpressure * Ptp\_wind\_speed | * Max\_meantemp * Min\_meantemp * Min\_humidity * Mean\_meantemp * Mean\_wind\_speed * Mean\_humidity * Max\_humidity * Min\_wind\_speed * Max\_meanpressure * Mean\_meanpressure * Min\_meanpressure * Ptp\_meantemp * Ptp\_humidity * Max\_wind\_speed * Skewn\_wind\_speed * Var\_meantemp |

Il existe 4 caractéristiques différentes entre les 2 résultats, les autres caractéristiques ont seulement des valeurs différentes.

### Fenêtre de 30 échantillons, et chevauchement de 77%



### Fenêtre de 20 échantillons, et chevauchement de 75%



# Conclusion :

Malgré les choix différents d’échantillons et de chevauchement de fenêtre temporelle. Les résultats du script de réduction de dimension semblent similaires : les caractéristiques conservées sont presque les mêmes ou proviennent de la même colonne d’origine, par exemple : toutes les caractéristiques de meantemp sont présentes au moins une fois.

# Conclusion générale :

J’ai apprécié travailler avec les times series, avoir un repaire temporel dans ses données permettent d’avoir un certain regard sur son projet, se rendre compte ce a quoi correspondent les données enregistrées.

# Source :

* Dataset :
  + <https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data>
* Autres informations :
  + <https://www.edinformatics.com/math_science/what-is-humidity.html>
  + <https://fr.wikipedia.org/wiki/Atmosph%C3%A8re_(unit%C3%A9)>