

ENTREGA # 1 – FUNDAMENTOS DE DEEP LEARNING

PREDICCIÓN DE VARIABLES CLIMATICAS MAXPLANK

CONTEXTO

La predicción de variables climáticas, día a día a toma mayor relevancia a nivel global, esto gracias a que en diferentes ciudades se han venido implementando sistemas de identificación de variables climáticas, los cuales permiten la descripción ambiental enfocado principalmente en la mitigación de riesgos. Dado que día a día se generan un gran numero de datos es necesario la implementación de modelos los cuales nos permitan su interpretación y aprovechamiento. Particularmente, un enfoque de esto es el aprovechamiento de las variables para la generación de modelos predictivos los cuales permitan la predicción de las variables climáticas en tiempos futuros.

Basados en esta necesidad, se aborda este enfoque desde el planteamiento de modelos basados en redes neuronales, y para evaluarlo se propone aprovechar el dataset publico Max Planck Weather, este conjunto de datos contiene 14 características diferentes, como la temperatura del aire, la presión atmosférica, la humedad, entre otros; los cuales fueron recopilados cada 10 minutos, a partir de 2003.

OBJETIVO

Predicir variables climáticas basado en el dataset publico Max Planck Weather utilizando modelos de redes neuronales

DATASET

Fuente: Kaggle, Max planck Weather Dataset <https://www.kaggle.com/datasets/arashnic/max-planck-weather-dataset>.

El dataset en cuestión corresponde a mediciones climáticas tomadas por el Max Planck Institute for Biogeochemistry (<https://www.bgc-jena.mpg.de/>) ubicado en Hamburgo Alemania. El dataset contiene mediciones de variables climáticas relevantes como los son las distintas mediciones de temperaturas, presiones, velocidades y direcciones de corrientes de aire entre otras. Los modelos propuestos a continuación buscan realizar la predicción de las variables seleccionadas del Dataset a través de análisis de redes neuronales.

En primer lugar, se realiza un análisis del dataset para conocer las variables que lo componen, la cantidad de datos presentes y proponer una ruta de análisis y datos a selecciona. Los datos que componen el dataset y su interpretación son las siguientes:

- DateTime: Fecha en el formato MM:HH:DD hh:mm:ss desde el año 2009 al año 2016.
- p (mbar): Presión atmosférica en milibares,
- T (degC): Temperatura en grados Celsius
- Tpot (K): Temperatura en grados Kelvin.

- Tdew (degC): Punto de rocío, es una variable correspondiente a la temperatura a la cual el aire no puede contener la humedad presente en el aire y se el agua se condensa (se genera el rocío).
- rh (%): porcentaje de humedad relativa, corresponde a la saturación de vapor de agua en el aire.
- VPmax (mbar): Presión de saturación del vapor
- VPact (mbar): Presión de vapor (medida)
- VPdef (mbar): Déficit de presión de vapor (diferencia entre las dos variables previas)
- sh (g/kg): Humedad específica.
- H2OC (mmol/mol): Concentración de vapor de agua.
- rho (g/m**3): Densidad del aire.
- wv (m/s): Velocidad del viento.
- max. wv (m/s): Máxima velocidad del viento.
- wd (deg): Dirección del viento medida en grados.

El dataset elegido cuenta con 420551 celdas sin faltantes y con un 0.1% de celdas repetidas, que pueden considerarse como momentos en el tiempo en los cuales las variables meteorológicas coinciden más no en datos repetidos que deban ser removidos de la base de datos; por esto podemos considerarla una base de datos bastante limpia y fácil de utilizar para el análisis planteado.

Después de reconocer las variables del Dataset se analizan las correlaciones presentes entre estas como puede apreciarse en la figura 1. Esta muestra la correlación de Spearman entre las variables previamente mencionadas.

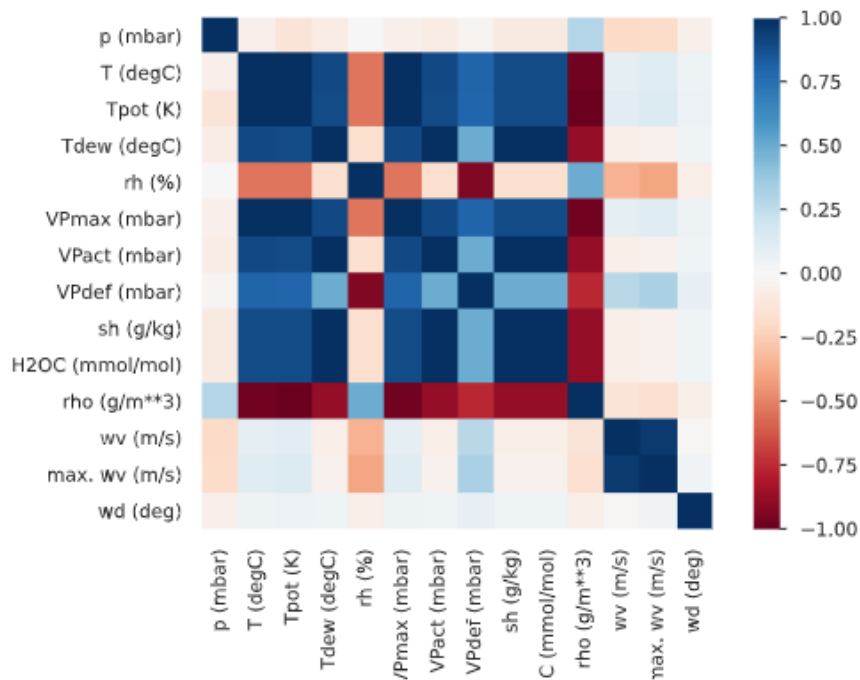


Figura 1 Correlación de Spearman entre las variables del meteorológicas del dataset

Se puede apreciar la elevada correlación entre algunas variables, esto debido a que fenomenológicamente las variables climáticas tienen relaciones como:

- La temperatura en grados Kelvin T (degC) es la temperatura en grados Celsius T (degC) con un Offset de 273,15°. Además, el punto de rocío o T_{dew} (degC) es una resultante de las condiciones de presión y temperaturas atmosféricas.
- Las diferentes presiones (atmosféricas, presión de vapor, presión de saturación de vapor) se encuentran correlacionadas físicamente entre sí.
- Variables como la densidad del aire o el porcentaje de humedad relativa son producto de las variables mencionadas previamente.

Las variables definidas para el análisis de series de tiempo son la **temperatura en grados Celsius**, la **densidad** del aire en g/m³ y la **presión** en milibares. Esta definición se hace para comparar variables con una clara tendencia de estacionalidad como los son las dos primeras y otra de baja estacionalidad como lo es la última (Figura 2).

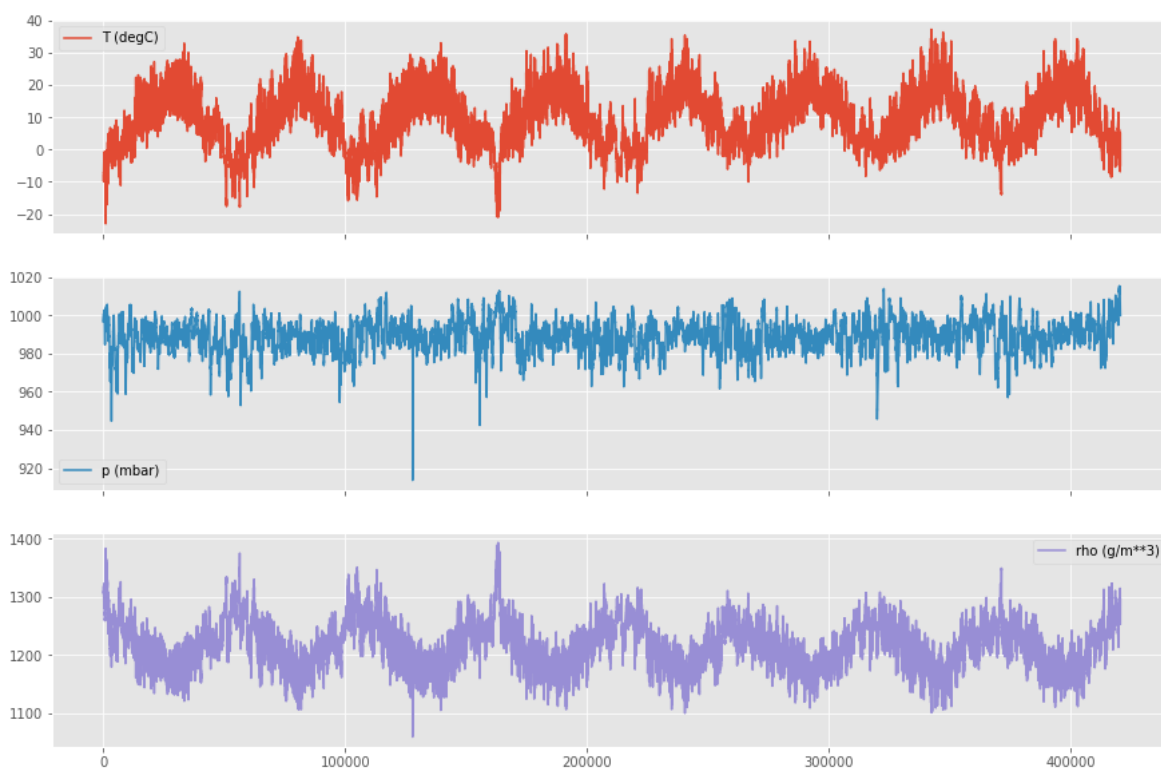


Figura 2 Comportamiento de las variables seleccionadas.

REFERENCIAS

[1] Prabhanshu Attri, Yashika Sharma, Kristi Takach, Falak Shah, Timeseries forecasting for weather prediction (https://keras.io/examples/timeseries/timeseries_weather_forecasting/)