7-4-2024

Proyecto Deep Learning

Entrega 1

Profesor: Raúl Ramos P

YESID ORLANDO RAMIREZ CASTANO
ESTUDIANTE DE POSGRADO , MAESTRIA EN INGENIERIA

Universidad de Antioquia

Medellin-Antioquia

2024

Contexto de la aplicación:

Una de las variables con mayor relevancia en el cambio climático es la temperatura. Por lo que se requiere una predicción precisa de esta. Varios estudios se han centrado en el uso de técnicas de regresión para predecir el cambio climático y su impacto en la temperatura.

Para este proyecto los datos están dados por el instituto de investigación y servicio operativo europeo ECMWF (European Centre for Medium-Range Weather Forecasts) es una reconocida organización de investigación meteorológica con sede en el Reino Unido. Es conocido por sus modelos avanzados de predicción meteorológica numérica y su función a la hora de proporcionar previsiones meteorológicas a medio plazo para Europa y otras partes del mundo (Andrés, 2011).

El Centro desempeña un papel clave en Copernicus (programa meteorológico), el componente de observación de la tierra del programa espacial de la Unión Europea, y ofrece información de calidad sobre el cambio climático, la composición atmosférica, las inundaciones y el peligro de incendio (ECMWF, 2024).

El objetivo planteado en la competencia de Kaggle, donde los datos fueron dispuestos , remite corregir errores en los pronósticos de temperatura de la superficie, del modelo ECMWF. En concreto, desean crear predicciones post-procesadas en unas 500 estaciones de superficie en Alemania (Kaggle, 2024).

Objetivo de machine learning.

Reducir los sesgos de pronósticos de temperatura.

Dataset

Tipo de datos

La competencia brinda los siguientes datos estructurados de la siguiente manera.

1. *pp_train.csv* contiene el dataset de entrenamiento incluida la variable objetivo t2m_obs
2. *pp_test.csv* contiente el dataset de tes sin la variables objetivo
3. *pp_sample_submission.csv* es un ejemplo de un archivo de envío con el formato adecuado

Tamaño (número de datos y tamaño en disco)

Con respecto a la forma y tamaño en disco con ayuda de Python se obtuvo:

```python
import numpy as np
import pandas as pd

train = pd.read_csv("pp_train.csv")
test = pd.read_csv("pp_test.csv")
subM = pd.read_csv("pp_sample_submission.csv")

print(f"Forma de train: {train.shape}")
print(f"Memoria en uso de train: {train.memory_usage().sum()} bytes")
print(f"Forma de test: {test.shape}")
print(f"Memoria en uso de test: {test.memory_usage().sum()} bytes")
print(f"Forma de sum¿bM: {subM.shape}")
print(f"Memoria en uso de subM: {subM.memory_usage().sum()} bytes")
```

✓ 9.6s

```
Forma de train: (980562, 26)
Memoria en uso de train: 203957028 bytes
Forma de test: (182218, 25)
Memoria en uso de test: 36443732 bytes
Forma de sum¿bM: (182218, 2)
Memoria en uso de subM: 2915620 bytes
```

Los tipos de datos al provenir de una entidad que se enfoca en hacer reanálisis de datos climáticas y meteorológicos, se encontraran que los data_sets ya estan preprocesados y para este caso no se cuenta con variables categóricas como se evidencia.

En la siguiente tabla se describe las abreviaciones de las columnas

Tabla 1. Abreviaciones y descripción de todas las características (Rasp & Lerch, 2018)

Feature	Description
Ensemble predictions (mean and std dev)	
t2m	2-m temperature
cape	Convective available potential energy
sp	Surface pressure
tcc	Total cloud cover
sshf	Sensible heat flux
slhf	Latent heat flux
u10	10-m *U* wind
v10	10-m *V* wind
d2m	2-m dewpoint temperature
ssr	Shortwave radiation flux
str	Longwave radiation flux
sm	Soil moisture
u_pl500	*U* wind at 500 hPa
v_pl500	*V* wind at 500 hPa
u_pl850	*U* wind at 850 hPa
v_pl850	*V* wind at 850 hPa
gh_pl500	Geopotential at 500 hPa
q_pl850	Specific humidity at 850 hPa
Station-specific information	
station_alt	Altitude of station
orog	Altitude of model grid point
station_lat	Lat of station
station_lon	Lon of station

Datos de train y test:

```
RangeIndex: 980562 entries, 0 to 980561
Data columns (total 26 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        980562 non-null  int64
 1   t2m_obs           908857 non-null  float64
 2   time              980562 non-null  object
 3   station           980562 non-null  int64
 4   t2m_fc_mean       980562 non-null  float64
 5   orog              980562 non-null  float64
 6   station_alt       980562 non-null  float64
 7   station_lat       980562 non-null  float64
 8   station_lon       980562 non-null  float64
 9   u_pl500_fc_mean   980562 non-null  float64
 10  v_pl500_fc_mean   980562 non-null  float64
 11  gh_pl500_fc_mean  980562 non-null  float64
 12  u_pl850_fc_mean   980562 non-null  float64
 13  v_pl850_fc_mean   980562 non-null  float64
 14  q_pl850_fc_mean   980562 non-null  float64
 15  cape_fc_mean      980562 non-null  float64
 16  sp_fc_mean        980562 non-null  float64
 17  tcc_fc_mean       980562 non-null  float64
 18  sshf_fc_mean      980562 non-null  float64
 19  slhf_fc_mean      980562 non-null  float64
 20  u10_fc_mean       980562 non-null  float64
 21  v10_fc_mean       980562 non-null  float64
 22  ssr_fc_mean       980562 non-null  float64
 23  str_fc_mean       980562 non-null  float64
 24  d2m_fc_mean       980562 non-null  float64
 25  sm_fc_mean        886600 non-null  float64
dtypes: float64(23), int64(2), object(1)
memory usage: 194.5+ MB
```

```
RangeIndex: 182218 entries, 0 to 182217
Data columns (total 25 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        182218 non-null  int64
 1   time              182218 non-null  object
 2   station           182218 non-null  int64
 3   t2m_fc_mean       182218 non-null  float64
 4   orog              182218 non-null  float64
 5   station_alt       182218 non-null  float64
 6   station_lat       182218 non-null  float64
 7   station_lon       182218 non-null  float64
 8   u_pl500_fc_mean   182218 non-null  float64
 9   v_pl500_fc_mean   182218 non-null  float64
 10  gh_pl500_fc_mean  182218 non-null  float64
 11  u_pl850_fc_mean   182218 non-null  float64
 12  v_pl850_fc_mean   182218 non-null  float64
 13  q_pl850_fc_mean   182218 non-null  float64
 14  cape_fc_mean      182218 non-null  float64
 15  sp_fc_mean        182218 non-null  float64
 16  tcc_fc_mean       182218 non-null  float64
 17  sshf_fc_mean      182218 non-null  float64
 18  slhf_fc_mean      182218 non-null  float64
 19  u10_fc_mean       182218 non-null  float64
 20  v10_fc_mean       182218 non-null  float64
 21  ssr_fc_mean       182218 non-null  float64
 22  str_fc_mean       182218 non-null  float64
 23  d2m_fc_mean       182218 non-null  float64
 24  sm_fc_mean        175102 non-null  float64
dtypes: float64(22), int64(2), object(1)
memory usage: 34.8+ MB
```

Distribuciones de las clases

Con respecto a las distribuciones de las clases, el objetivo de proyecto se enfoca en realizar un pronostico de temperatura por lo que no se cuentan con clases en este caso.

Métricas de desempeño

Machine learning

Ya que el objetivo principal radica en reducir el sesgo de error la métrica que se va a usar el MSE.

De Negocio

Reducción de Errores y Desperdicios: Evalúa el impacto del modelo en la reducción de errores, desperdicios o pérdidas en la operación del negocio, al realizar predicciones con mayor precisión que puedan usarse para prever procesos donde el factor de la temperatura sea determinante , e.g. agricultura, energía solar,etc

Referencias y resultados previos

En la predicción de temperatura Mishra y otros realizaron un estudio mediante regresión polinomial para predecir la temperatura futura basándose en un conjunto de datos de la NASA, lo que permitió lograr una alta precisión en el entrenamiento y las pruebas (Subhra et al., 2023). Malakouti utilizó algoritmos de aprendizaje automático para crear un modelo de predicción del cambio de temperatura global, siendo el algoritmo Extra Trees el que mejor rendimiento . Otro estudio utilizó la regresión lineal y polinomial para pronosticar el aumento de la temperatura hasta 2059, teniendo en cuenta factores como las emisiones de CO2 y el índice de calidad del aire (Malakouti, 2023). Estos estudios destacan la eficacia de las técnicas de regresión y aprendizaje automático para predecir los cambios de temperatura debidos al cambio climático.

En otro estudio explican que debido a que las predicciones meteorológicas de conjunto requieren un postprocesamiento estadístico de errores sistemáticos para obtener pronósticos probabilísticos confiables y preciso. Proponen un modelo basado en redes neuronales que puede incorporar relaciones no lineales entre variables predictoras arbitrarias y parámetros de distribución de pronóstico que se aprenden automáticamente de una manera basada en datos en lugar de requerir una especificación previa funciones de enlace. (Rasp & Lerch, 2018)

Bibliografía

ECMWF. (03 de 04 de 2024). Obtenido de https://www.ecmwf.int/en/about

Kaggle. (05 de 04 de 2024). *Weather Postprocessing*. Obtenido de https://www.kaggle.com/competitions/weather-postprocessing/overview

Andrés, C. G. (2011). *Desarrollo y verificación de un modelo regional de clima para su aplicación en la Península Ibérica*. https://api.semanticscholar.org/CorpusID:162741730

Malakouti, S. M. (2023). Utilizing time series data from 1961 to 2019 recorded around the world and machine learning to create a Global Temperature Change Prediction Model. *Case Studies in Chemical and Environmental Engineering*, *7*(December 2022), 100312. https://doi.org/10.1016/j.cscee.2023.100312

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Subhra, S., Mishra, S., Alkhayyat, A., Sharma, V., & Kukreja, V. (2023). Climatic Temperature Forecasting with Regression Approach. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, 1–5. https://doi.org/10.1109/ICIEM59379.2023.10166883