Universidad Politécnica de Madrid (UPM)
ETS de Ingenieros Informáticos
Grado en Ciencia de Datos e IA

UNIVERSIDAD
POLITÉCNICA
DE MADRID

Proyecto Final

—

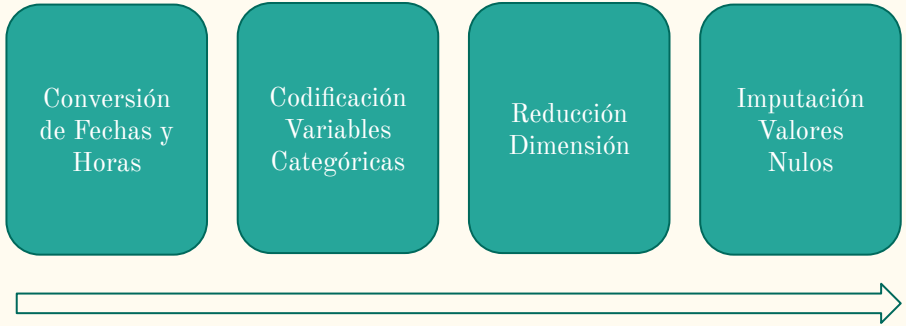
Por Laura García Perrín y Xiya Sun

Índice de Contenidos

Dominio y Procesamiento de Datos

- Dominio: clima
- Fuente de Datos: Kaggle [1]
- 96453 instancias o filas
- 12 variables o columnas

Pipeline del Procesamiento de Datos



Característica	Tipo de Datos	Descripción
Formatted Date	datetime64[ns, UTC]	Fecha y hora en formato UTC
Summary	object	Resumen del clima para el día
Precip Type	object	Tipo de precipitación
Temperature (C)	float64	Temperatura en C°
Apparent Temperature (C)	float64	Temperatura aparente en C°
Humidity	float64	Humedad relativa en porcentaje
Wind Speed (km/h)	float64	Velocidad del viento en km/h
Wind Bearing (degrees)	float64	Dirección del viento en grados
Visibility (km)	float64	Visibilidad en km
Loud Cover	float64	Cobertura de nubes
Pressure (millibars)	float64	Presión atmosférica en milibares
Daily Summary	object	Resumen del clima para el día

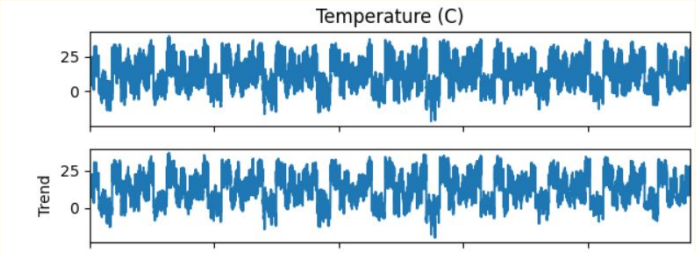
[1] https://www.kaggle.com/datasets/muthuj7/weather-dataset/data

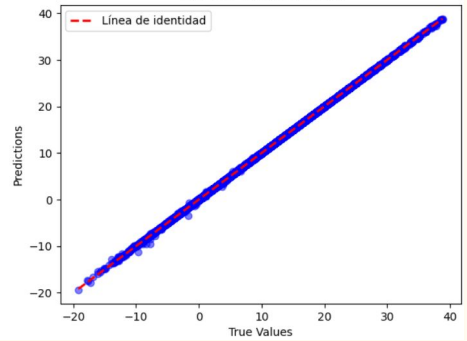
Objetivos del Proyecto

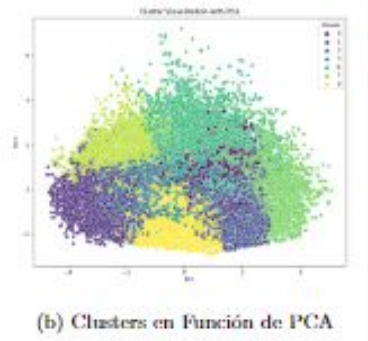
Objetivo principal → análisis temporal de tendencias climáticas.

Otros propósitos:

- ★ Lanzar predicciones
- ★ Identificar fenómenos climáticos
- ★ Identificar patrones y grupos






(b) Clusters en Función de PCA

Técnicas de Data Mining

KMeans

✅Análisis

10 grupos climáticos

(S)ARIMA

✅Predicciones

Resultados mejorables de precisión y de latencia

Random Forest Regressor

✅Predicciones

Buenos resultados de precisión

CNNs

✅Análisis

Detecta eficazmente características de ST

Evaluación y Resultados

KMeans

- Silhouette Score: 0.2447
- Davies-Bouldin Index: 1.3159
- Calinski-Harabasz Index: 22616.9146

(S)ARIMA

- MAE: 197.4030
- MSE: 50431.0512
- RMSE: 224.5685
- MAPE: 62.6810

Random Forest Regressor

- MAE: 0.0122
- MSE: 0.0019
- RMSE: 0.0443
- MAPE: 0.0076
- R-squared: 0.9999

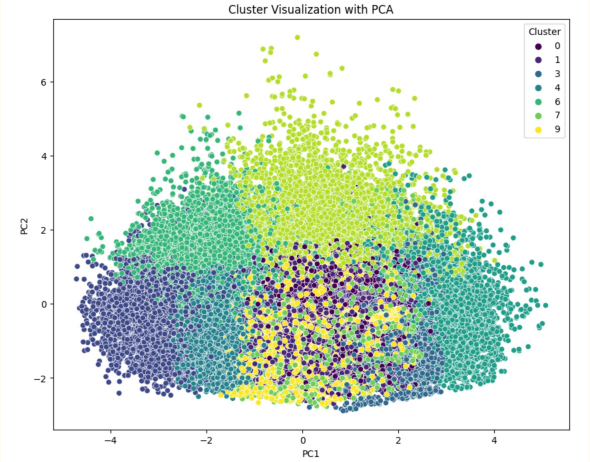
CNNs

- Test loss: 0.4592
- Test accuracy: 0.9743

Consideraciones - Clustering KMeans

Satisface los objetivos de:

- Identificar patrones y tendencias; predecir condiciones climáticas futuras.
- Identificar clústeres asociados con eventos climáticos extremos.
- Agrupar datos en patrones distintivos; identificar tendencias y cambios a lo largo del tiempo.



Guarda una serie de **contrapartidas**:

- Sensible a los outliers
- Capacidad limitada para captar patrones complejos

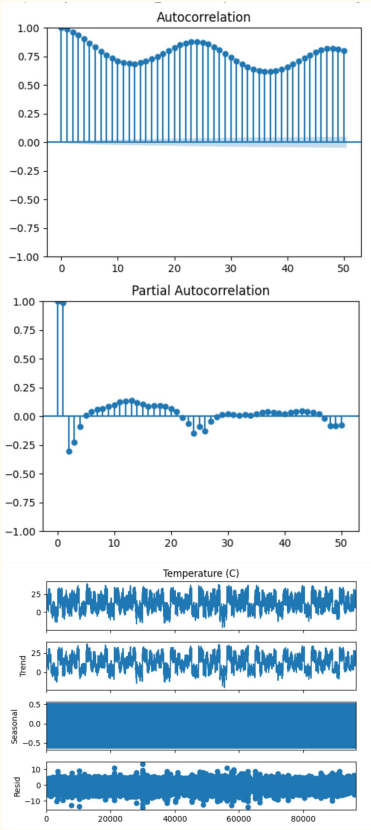
Consideraciones - (S)ARIMA

Satisface los objetivos de:

- Modelizar la evolución de una variable en el tiempo
- Identificar características o eventos climáticos.
- Predecir valores futuros basado en valores pasados (autoregresión).

Guarda una serie de **contrapartidas**:

- Malos resultados
- Muy lento; poco optimizado
- Requiere de un análisis previo algo exhaustivo



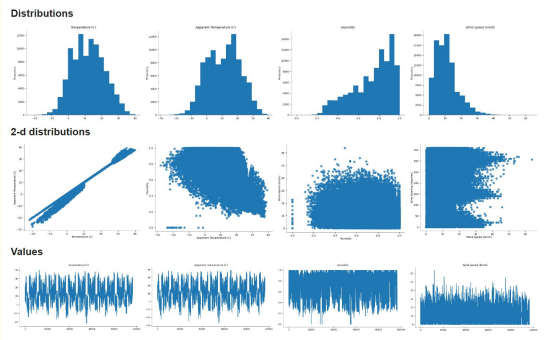
Consideraciones - Random Forest Regressor

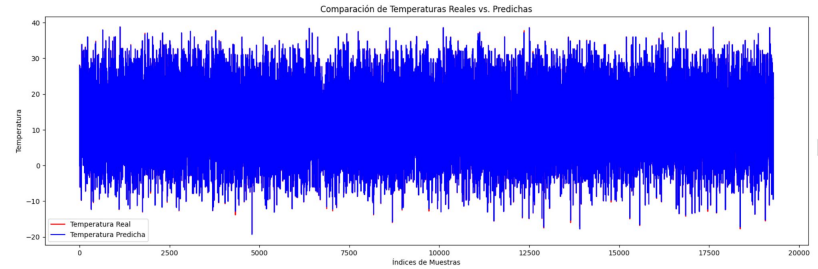
Satisface los objetivos de:

- Lanzar predicciones.
- Identificar patrones y tendencias en los datos climáticos.

✅Parece arrojar buenos resultados de predicción

➡️Latencia mejorable





Consideraciones - CNN

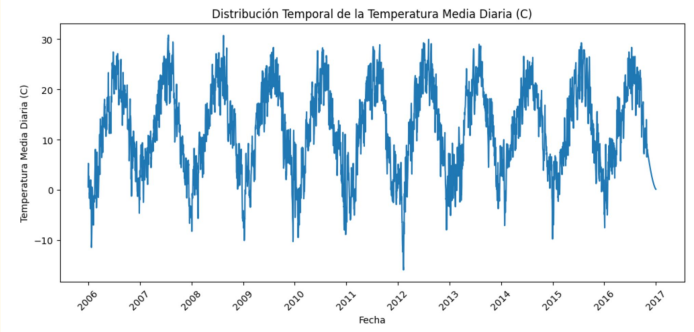
Satisface los objetivos de:

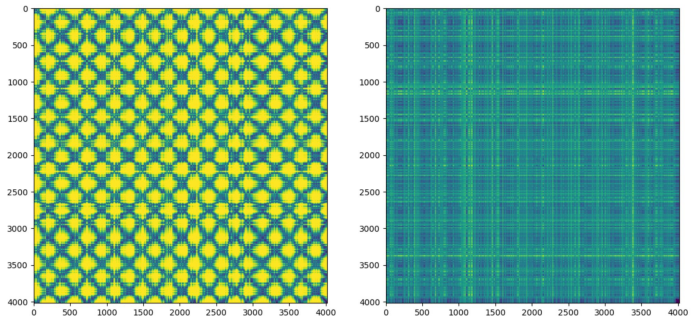
- Extraer características de las series temporales y clasificarlas
- Identificar patrones visuales que reflejan cambios climáticos significativos.

✅Diferencia bien entre series temporales

✅Latencia deseable

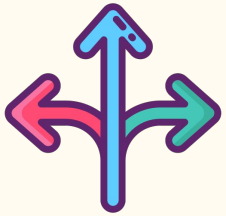
✅Método novedoso y de actualidad





Aspectos Mejorables

Aspecto Mejorable	Propuesta
Latencia en Métodos Estadísticos	Paralelizar las funciones o recurrir a otras herramientas y/o lenguajes de programación, como R o Weka.
Flexibilidad	Elaborar de métodos y funciones que sean trasladables a la hora de entrenar y evaluar modelos.
Innovar y Explorar	Investigar nuevas metodologías que sean de interés en la actualidad dada su relevancia, eficiencia y rendimiento.



23 funciones implementadas

Método	Tiempo de Ejecución (segundos)
`impute_precip_type_na_as_snow`	5.15
`preprocess_weather_data`	6.13
`apply_kmeans`	9.56
`fit_sarima`	131.50
`train_sarima`	97.52
`forecast_sarima`	1.57
`prepare_rf`	42.45
`crear_series_temporales`	0.13
`recurrence_plot`	0.22
`prepare_training_data`	22.50
`train_and_evaluar_model`	11.48

Conclusiones finales

- ★ El proceso KD (*Knowledge Discovery*) incluye varias etapas clave
- ★ No hay una única metodología para el KD
  - Depende del dominio y de los objetivos
  - Siempre hay espacio para mejoras
- ★ La finalidad subyacente es descubrir conocimientos útiles no evidentes
- ★ Las técnicas de Data Mining se alinean con los objetivos y tienen diferentes contribuciones
  - Ofrecen una visión integral del problema y/o conjunto de datos
  - Son perspectivas diferentes

Universidad Politécnica de Madrid (UPM)
ETS de Ingenieros Informáticos
Grado en Ciencia de Datos e IA

UNIVERSIDAD
POLITÉCNICA
DE MADRID

Gracias por su Atención

Laura García Perrín y Xiya Sun