Persiguiendo el clima: conocimientos y predicciones basados en datos

Daniela Mancilla
Gabriela Martínez
Francisca Quijada

Introducción:

El clima tiene una gran influencia en diversos aspectos vitales:

- Recursos hídricos
- Agricultura
- Desastres naturales

Introducción:

Objetivo general: *estudiar y comprender cómo las variables atmosféricas y oceánicas influyen en las condiciones climáticas*.

- **¿Se pueden realizar pronósticos o estimaciones?**
- **¿Existen patrones complejos entre las variables?**
- **¿Se pueden detectar anomalías?**

Los datos

Lugar: Estación Metereológica del Faro Extremo Molo de Abrigo, Valparaíso.

Cantidad: se consideran registros realizados cada una hora durante el año 2022, los que ascienden a una cantidad igual a 8.760.

Los datos

	fecha	TA	HR	PP	PA	VV	RV	DV	PRS	TW
0	2022-01-01 00:00:00	11.2	81.8	0.0	975.0	3.9	14.0	177.0	3.11	15.47
1	2022-01-01 01:00:00	11.0	81.5	0.0	974.0	2.3	8.6	208.0	3.04	14.90

TA: temperatura del aire (ºC)
PP: precipitación acumulada (mm)
VV: rapidez del viento (km/h)
DV: dirección del viento (º)
TW: temperatura del agua (ºC)

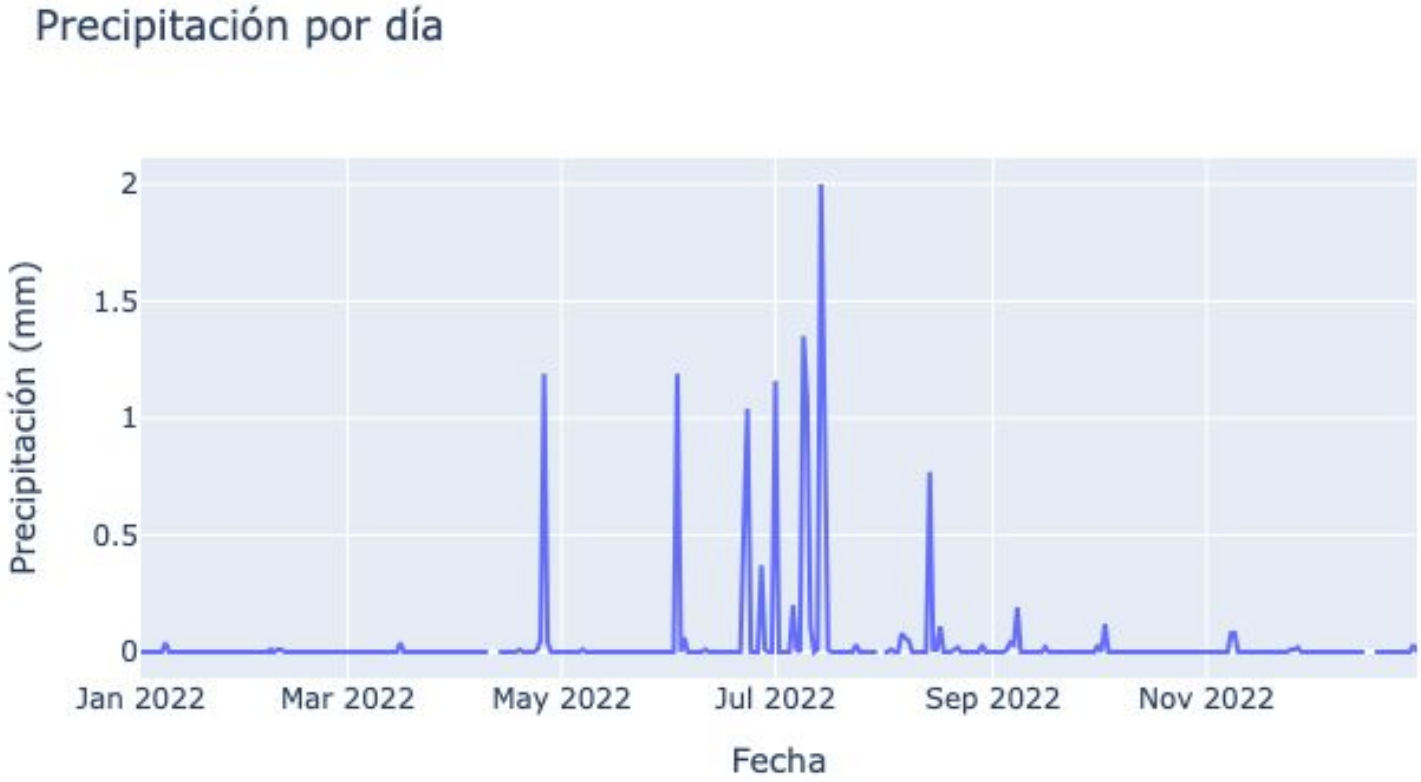
HR: humedad relativa (%)
PA: presión atmosférica (mbar)
RV: ráfaga de viento (km/h)
PRS: nivel del mar (m)

Método ETL

- Eliminación de columnas y filas
- Formatos → de object a float, de object a datetime
- Unión de archivos → 2 .csv (atmosféricos) + 12 .csv (oceánicos)
- Análisis de nulos y duplicados

Temperatura a lo largo del año

Precipitaciones



Correlaciones

Correlación entre los distintos atributos

Regresión

Para responder a las preguntas: ¿cuánto lloverá? o ¿cuál será la temperatura?, se utilizan algoritmos de ***regresión.***

ATRIBUTOS X
VARIABLE DE RESPUESTA y

	fecha	TA	HR	PP	PA	VV	RV	DV	PRS	TW
0	2022-01-01 00:00:00	11.2	81.8	0.0	975.0	3.9	14.0	177.0	3.11	15.47
1	2022-01-01 01:00:00	11.0	81.5	0.0	974.0	2.3	8.6	208.0	3.04	14.90

	fecha	TA	HR	PP	PA	VV	RV	DV	PRS	TW
0	2022-01-01 00:00:00	11.2	81.8	0.0	975.0	3.9	14.0	177.0	3.11	15.47
1	2022-01-01 01:00:00	11.0	81.5	0.0	974.0	2.3	8.6	208.0	3.04	14.90

Temperatura ambiente

TA	HR	PP	PA	VV	RV	DV	PRS	TW
11.2	81.8	0.0	975.0	3.9	14.0	177.0	3.11	15.47
11.0	81.5	0.0	974.0	2.3	8.6	208.0	3.04	14.90

```
Model: LinearRegression, Mean score: 0.562
Model: RidgeCV, Mean score: 0.562
Model: LassoCV, Mean score: 0.55
Model: DecisionTreeRegressor, Mean score: 0.263
Model: RandomForestRegressor, Mean score: 0.584
Model: GradientBoostingRegressor, Mean score: 0.617
Model: SVR, Mean score: 0.197
Model: AdaBoostRegressor, Mean score: 0.552
Model: ExtraTreesRegressor, Mean score: 0.592
```

Evaluación de métricas y visualización

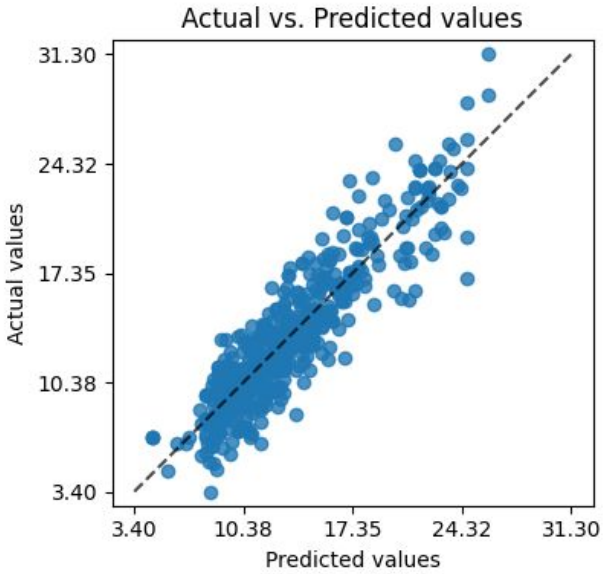
GradientBoostingRegressor + GridSearchCV

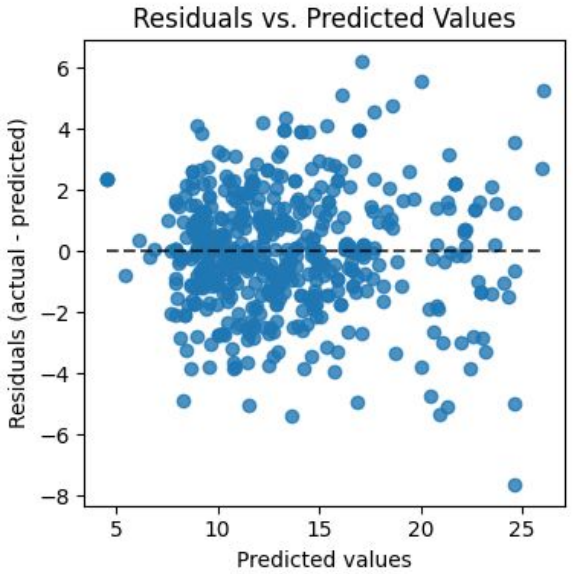
Coeficiente de determinación

R^2 = 0.819

Error absoluto medio

MAE= 1.50 ºC





Eliminación de características recursiva

```
R^2 para k= 1 : 0.635
R^2 para k= 2 : 0.719
R^2 para k= 3 : 0.803
R^2 para k= 4 : 0.835
R^2 para k= 5 : 0.845
R^2 para k= 6 : 0.845
R^2 para k= 7 : 0.847
R^2 para k= 8 : 0.848
```

- Humedad Relativa
- Presión Atmosférica
- Velocidad Viento
- Temperatura Océano

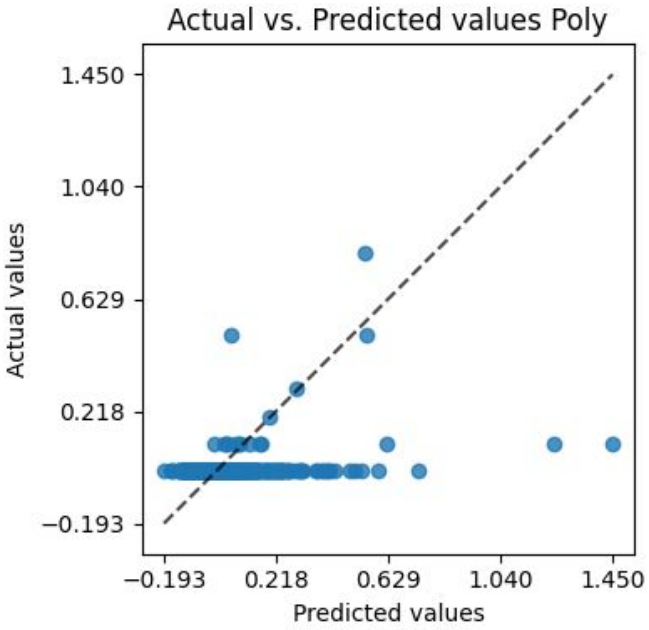
Precipitaciones

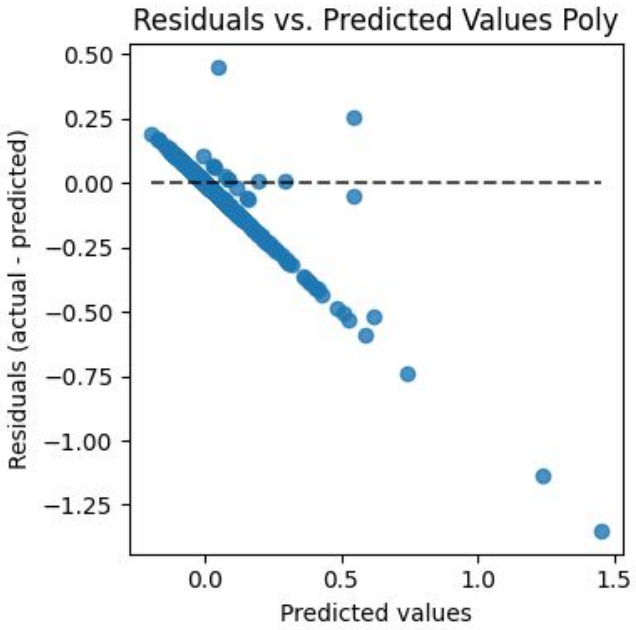
TA	HR	PP	PA	VV	RV	DV	PRS	TW
11.2	81.8	0.0	975.0	3.9	14.0	177.0	3.11	15.47
11.0	81.5	0.0	974.0	2.3	8.6	208.0	3.04	14.90

Modelo → Regresión polinomial de orden 2

Coeficiente de determinación

$R^2 = 0.193$





Precipitaciones

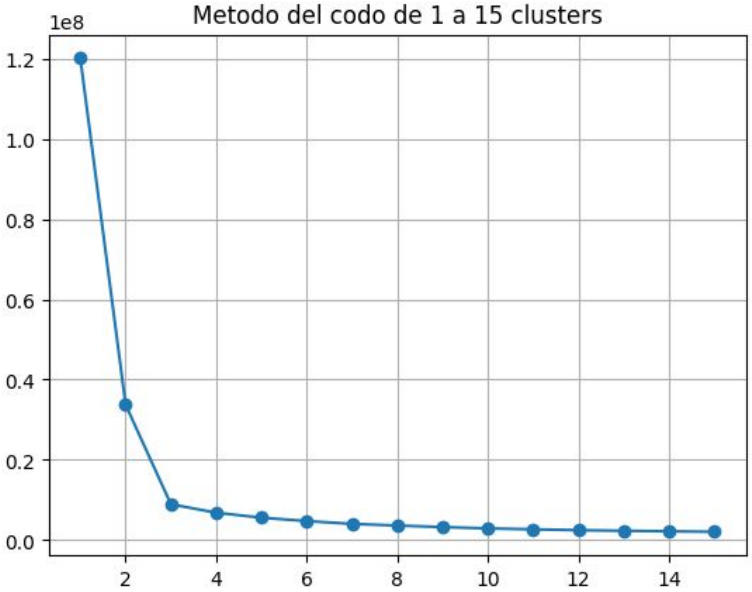
Validación cruzada

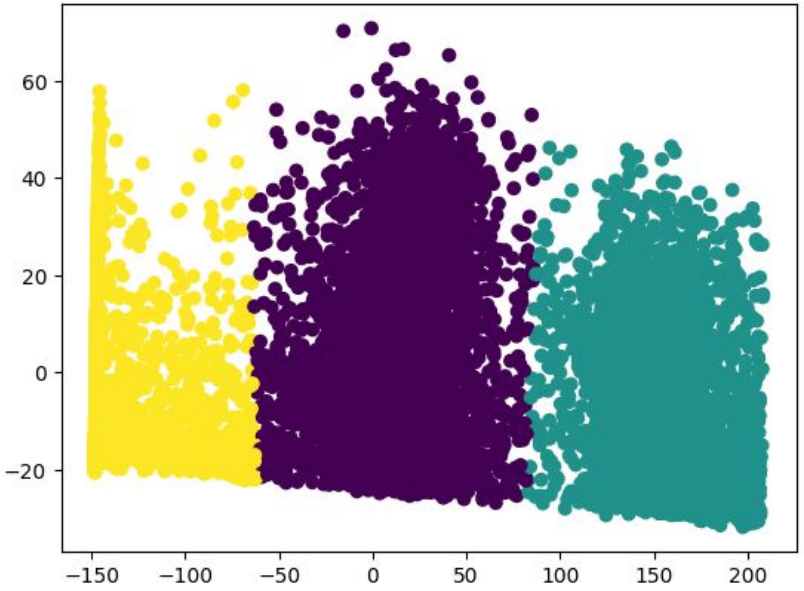
```
Model: LinearRegression, Mean score: -21.284
Model: RidgeCV, Mean score: -21.276
Model: LassoCV, Mean score: -20.266
Model: DecisionTreeRegressor, Mean score: -62.895
Model: RandomForestRegressor, Mean score: -135.382
Model: GradientBoostingRegressor, Mean score: -31.967
Model: SVR, Mean score: -14.364
Model: AdaBoostRegressor, Mean score: -382.929
Model: ExtraTreesRegressor, Mean score: -14.425
```

Patrones

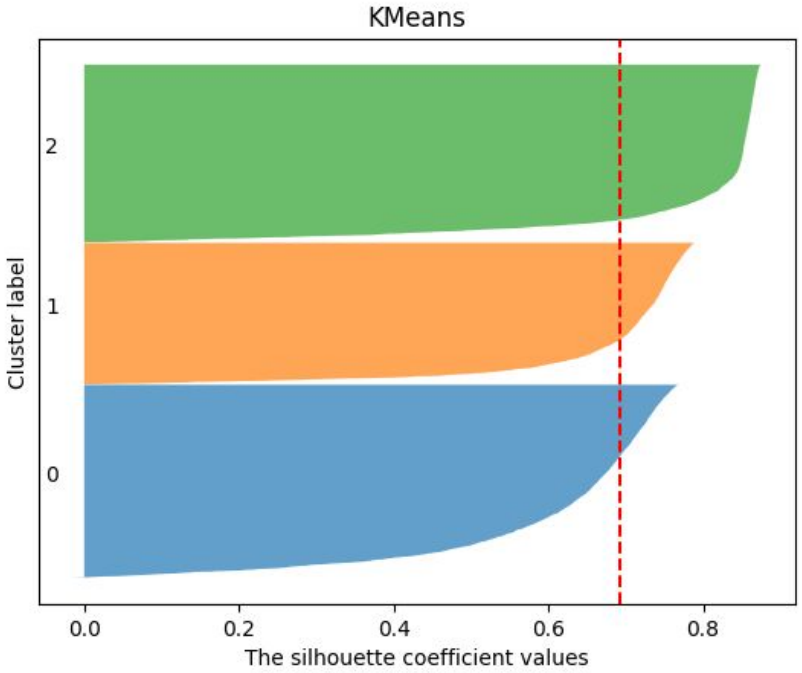
Para identificar patrones se utilizaron algoritmos de ***clustering***.

K-means





Coeficiente de silhouette



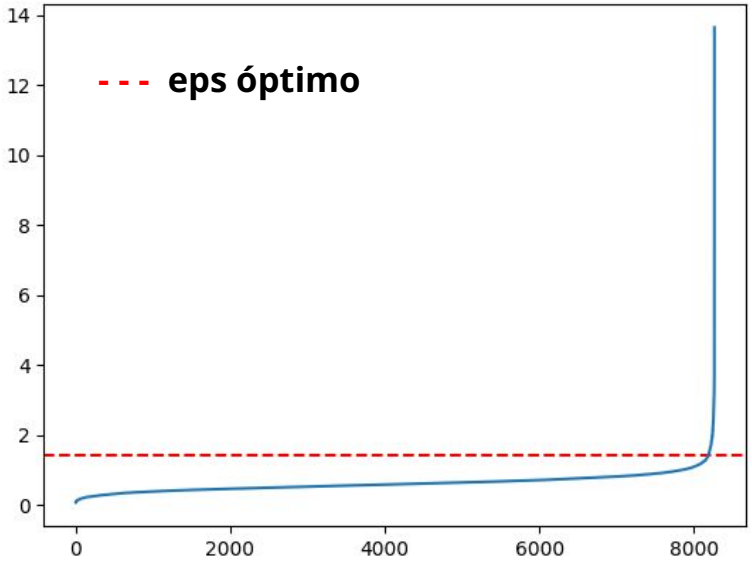
Características de cada grupo:

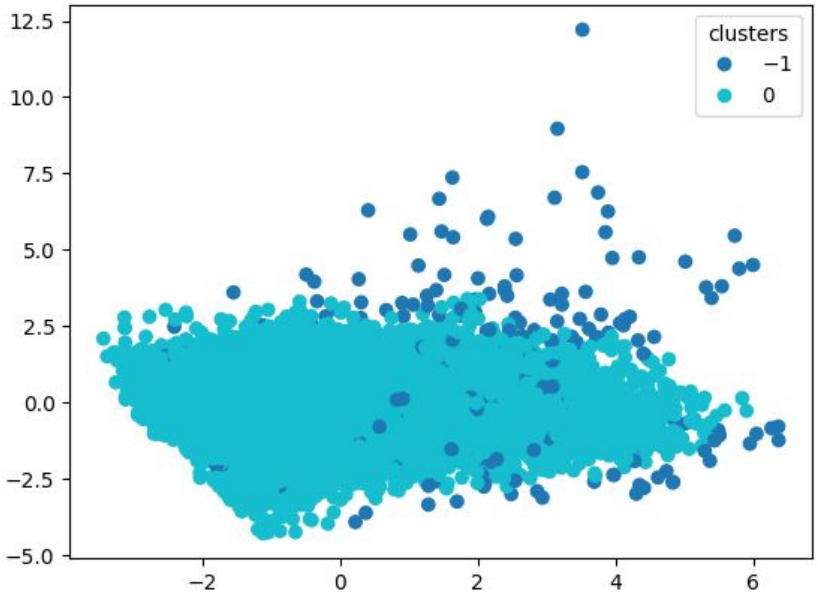
Cluster	count	mean TA	mean HR	mean VV	mean DV
0	3121.0	14.16	69.90	6.64	160.64
1	2283.0	14.46	76.94	8.53	302.73
2	2866.0	11.08	83.89	2.15	8.88

Anomalías

Se identifican “anomalías” como aquellas instancias que se desvían significativamente del comportamiento general de los grupos formados.

DBSCAN





Características de outliers

```
df_new.groupby(['Cluster'])["PP"].describe()
```

Cluster	count	mean	std	min	25%	50%	75%	max
-1	263.0	1.077567	2.144403	0.0	0.0	0.0	1.1	17.0
0	8007.0	0.004484	0.032409	0.0	0.0	0.0	0.0	0.6

```
df_new.groupby(['Cluster'])["VV"].describe()
```

Cluster	count	mean	std	min	25%	50%	75%	max
-1	263.0	12.600760	7.531871	0.0	6.8	12.1	17.4	34.8
0	8007.0	5.372711	4.605777	0.0	1.5	4.3	8.3	23.8

Los puntos que se alejan del comportamiento normal de los datos se caracterizan por presentar altas precipitaciones y alta rapidez de viento.

Discusión y Conclusiones

Regresión

Si bien los modelos lograron predecir de manera aceptable la temperatura ambiente, no lo fue así para las precipitaciones.

- Considerar atributos adicionales, ej: nubosidad, radiación solar.
- Considerar intervalo de tiempo mayor, ej: 10 años.
- Las condiciones climáticas no son un fenómeno local.

Discusión y Conclusiones

Clustering

- El método DSBCAN nos permitió identificar los datos que se escapan de la norma.
- Para mejorar los modelos, se podrían eliminar los datos que se escapan de la norma.
- El método K-means nos permitió identificar tres grupos de datos.
- Escalar los datos no ayudó a obtener un mejor rendimiento cuando usamos K-means, ya que el coeficiente de silhouette disminuía.

Discusión y Conclusiones

- Si bien la presión atmosférica y el nivel del mar varían poco, esto no indica que no jueguen un papel importante en la predicción de otros atributos.

- Los objetivos se cumplieron parcialmente, siendo relativamente exitosos para la temperatura ambiental pero no así para las precipitaciones.

- Predecir comportamiento futuro → utilizar el atributo fecha.

¿PREGUNTAS?