Reto | Predicción de Contingencias Ambientales en el Área Metropolitana de la Ciudad de Monterrey

Indicaciones:

- Para la entrega de tu actividad subirás el **enlace a tu libreta de Google Colab** y una versión en PDF, al que llamarás **DS_C7_SC1_NOMBRE (sin espacios).**
- Para obtenerlo, ubícate en la libreta desde tu *drive* y presiona *Get link* desde el menú contextual. No olvides darle permisos al archivo para poder evaluarlo (Opción: *Anyone with the link*).
- Tu reporte será evaluado con base al cumplimiento de los requerimientos y a su contenido, pero también por su presentación, por lo que **errores ortográficos o de redacción serán penalizados**.
- El archivo de datos "Monterrey Pollution Data 2.csv" contiene las lecturas de una estación de monitoreo de calidad ambiental durante el año 2015 situada en el centro de la ciudad de Monterrey. Hay una lectura por hora. Cada renglón reporta las siguientes variables (Tabla 1):

Abrev	Variable	Unidades
Date	Fecha en que se tomó la lectura	
Month	Mes del año	
Day	Dia del mes	
DayWeek	Día de la semana en texto	
Weekday	Día de la semana (Domingo=1)	
Hour	Hora del dia	
CO	Monóxido de Carbono	Ppm
NO	Monóxido de Nitrógeno	Ppb
NO2	Dióxido de Nitrógeno	Ppb
NOx	Suma de NO y NO2	Ppb
O3	Ozono	Ppb
PM10	Particulas menores a 10 microns	g/m^3
PM2.5	Particulas menores a 2.5 microns	g/ m^3
PRS	Presión	Mmhg
RAINF	Lluvia	Mm/hr
RH	Humedad Relativa	%
SR	Radiación Solar	KW/m2
TOUT	Temperatura	DegC
WSR	Velocidad del Viento	Km/hr
WDV	Dirección del viento	Deg

Tabla 1. Variables de contaminación, tiempo y clima reportados por la estación de monitoreo

El gobierno del estado de Nuevo León está preocupado particularmente por las variables O3 que no debe exceder 120 ppb y por partículas que diámetro menor a 2.5 micrones (PM2.5) que no debe exceder 40.5 g/m^3. El resto de los contaminantes son considerados precursores.

1. **Para concretar este proyecto realiza los siguientes pasos:**

a) En algún entorno Spark, carga los datos y elimina las variables innecesarias.
b) Realiza un análisis de correlación y establece qué variables ambientales o de tiempo afectan la concentración de contaminantes O3 y PM2.5. Algunas correlaciones son negativas. Toma en cuenta que algunas correlaciones son positivas y algunas negativas indicando que los contaminantes incrementan y otras bajan.
c) Confirma estas relaciones usando gráficos de dispersión.
d) Crea modelos de regresión para predecir estos contaminantes con la menor cantidad de variables atributos posible. Puedes utilizar cualquier técnica: regresión lineal (o polinomial) multivariable, random forrests, gradient boost, o cualquier otra técnica que hayas investigado que funcione en Spark.

2. **Evalúa los modelos y escribe la reflexión de tu evaluación.**

Se evaluaron modelos de *Liner Regression, Random Forest y Gradient Boosting* para el pronóstico de las partículas de interés. Para el caso de partículas *O3*, se encontró que el modelo en base a el método Gradient Boosting fue el que produjo la predicción con menos error, como se puede ver en la tabla siguiente:

Tabla 1.- Métricas de Desempeño Prediciendo Partículas O3

	LinearRegressor	RandomForest	GradientBoosting
rmse	9.410127	8.062031	6.906035
mse	88.550483	64.996347	47.693320
mae	7.399044	6.084978	5.021190
r2	0.600913	0.707069	0.785051

Por otro lado, para el caso de la predicción de partículas *PM2.5*, ningún modelo arrojo un desempeño adecuado para usarse como solución, desempeños que se presentan en la siguiente tabla:

Tabla 2.- Métricas de Desempeño Prediciendo Partículas PM2.5

	LinearRegressor	RandomForest	GradientBoosting
rmse	12.457806	11.827384	11.643198
mse	155.196931	139.887007	135.564069
mae	9.748390	9.412532	9.275617
r2	0.157060	0.240215	0.263694

No obstante, si necesario realizar predicciones en base a los modelos analizados, se utilizaría el modelo en base a Gradient Boosting, dado que es el que presenta menor error, como se puede observar en la tabla anterior. En las siguientes figuras se presenta el desempeño, de forma gráfica, al realizar predicciones con todos los modelos entrenados para la predicción de las partículas de interés.

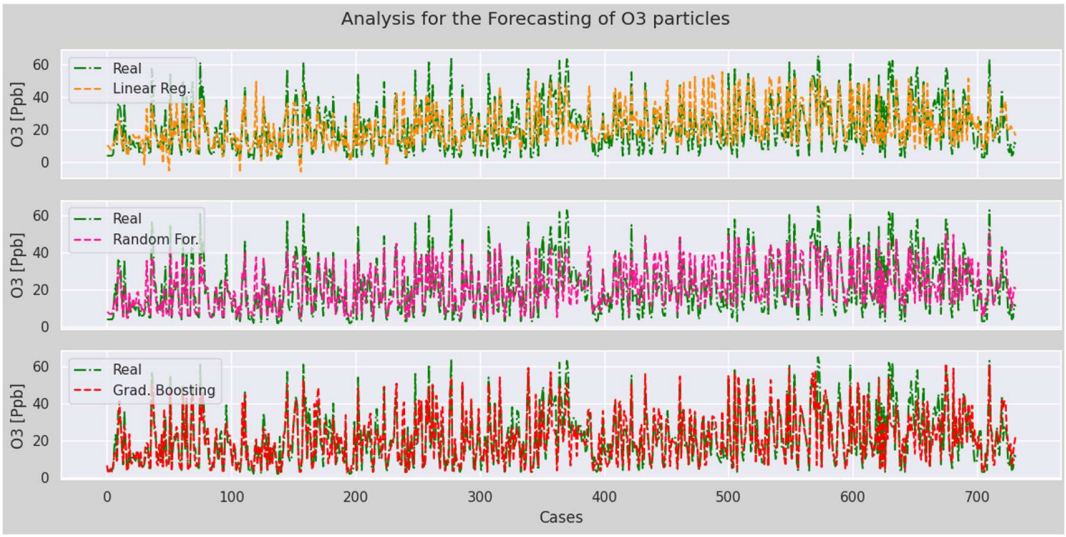

Ilustración 1.- Comparativo al utilizar los modelos entrenados para predecir partículas O3

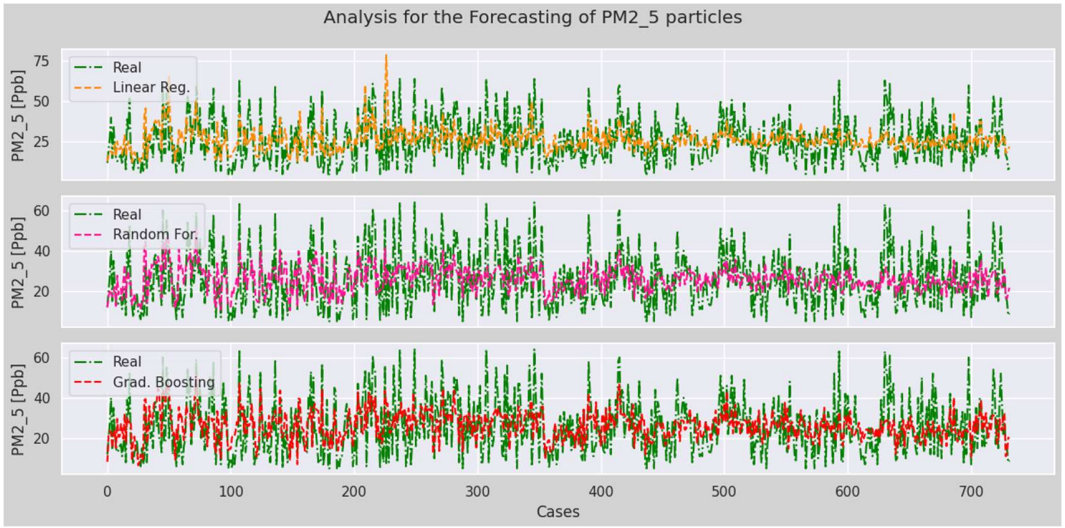

Ilustración 2.- Comparativo al utilizar los modelos entrenados para predecir partículas PM2.5

Conclusiones, contesta lo siguiente y justifica tus respuesta:

a) ¿Puedes decir que la contaminación por O3 o PM2.5 está ligada al tráfico vehicular?

Si se asume que el tráfico vehicular se encuentra ligado a la variable Hora, se puede decir que las partículas O3 se encuentran ligadas, hasta cierto punto, al tráfico, lo cual no es el caso para las partículas PM2.5

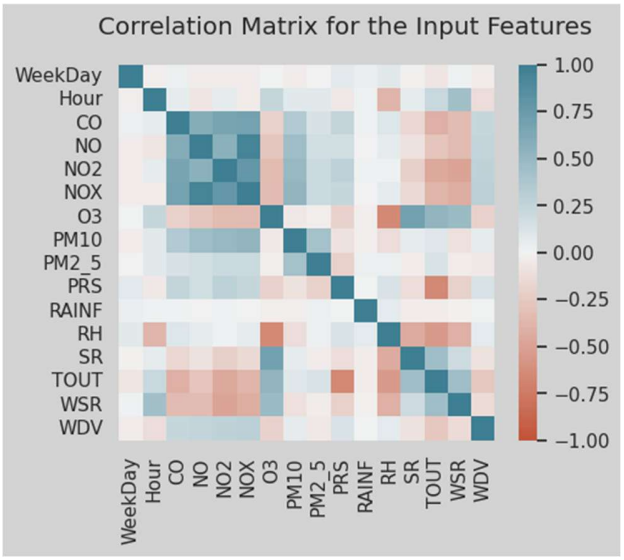

Ilustración 3.- Matriz de correlación de las características proporcionadas para esta solución

b) ¿Consideras que tendría efectos sobre la contaminación implantar un esquema de verificación vehicular?

La correlación entre el trafico vehicular y las partículas O3 existe, pero es pequeño. El beneficio depende de el tiempo en el cual se estime que el tráfico vehicular seguirá siendo un problema. Por ejemplo, si todo se mantiene igual a largo plazo se genera un beneficio a un esquema de verificación vehicular, sin embargo, si se promueven acciones como el uso de vehículos eléctricos o un mayor uso de transporte publico al aumentar su calidad, se podrían disminuir los beneficios de un esquema de verificación vehicular. No obstante, el tener más información acerca de las características de los vehiculas también será importante, para determinar si el generar incentivos para modernizarlos, si fuera el caso, también podría reducir la emisión de partículas O3.

c) Con un reporte del pronóstico del clima dado en la mañana, ¿puedes predecir que habrá una contingencia ambiental debido a que los contaminantes en el aire rebasaron los límites permitidos por la norma? ¿Por qué razón (es)?

La siguiente tabla presenta coeficientes de correlación de Pearson entre las partículas de interés y las características proporcionadas.

Tabla 3.- Correlación de Pearson entre las variables de interés y las características proporcionadas

	WeekDay	Hour	CO	NO	NO2	NOX	O3	PM10	PM2_5	PRS	RAINF	RH	SR	TOUT	WSR	WDV
O3	0.002597	0.241675	-0.205874	-0.278821	-0.33767	-0.332650	1.000000	-0.061473	-0.028865	-0.199972	-0.029271	-0.656093	0.689583	0.518518	0.475125	-0.200398
PM2_5	-0.002806	0.083095	0.140396	0.175017	0.20915	0.207492	-0.028865	0.426536	1.000000	-0.209827	0.027960	0.024929	-0.035409	0.119632	-0.043882	-0.048551

Se puede concluir que únicamente la presencia de partículas O3 se ven afectadas de manera significativa por las condiciones climáticas, no es el caso para la presencia de partículas PM2.5. En conclusión, se puede decir que, en caso de que exista presencia de partículas de contaminantes O3, se puede determinar, con cierto grado de certeza, la presencia de partículas O3 al conocer las condiciones climáticas en la mañana.

d) En tus propias palabras, ¿cuáles consideras qué son las condiciones climáticas se deben cumplir para tener altos niveles de contaminación de O3? ¿Y PM2.5?

Se puede observar, mediante los datos compartidos, que la intensidad del viento afecta en gran medida la presencia de partículas O3, al igual que la radiación solar y la lluvia, y en cierta manera, también se ven afectadas por la dirección del aire.

Por otro lado, se observa en los datos presentados que las partículas PM2.5 únicamente se ven afectadas por la presión ambiental. Sin embargo, se sabe que este tipo de partículas son afectadas por el viento y la lluvia en mayor medida a lo que se observa en los datos proporcionados. Por lo tanto, se puede concluir los datos no tienen la mejor calidad posible o que están sesgados a casos en que las partículas PM2.5 no se vieron afectadas por las condiciones climáticas. Se necesita mejorar la calidad de los datos.

¡Has cumplido con el Reto!

Si deseas reforzar tus conocimientos, puedes seguir el procedimiento siguiente:

Encontrarás que crear los modelos de predicción no es fácil. Para mejorar los niveles de predicción, toma en cuenta lo siguiente:

- Si los niveles de los contaminantes son bajos, no es relevante. Lo único que importa es cuando los niveles son altos.
- Modifica la tabla agrupando los datos por día y obteniendo el valor máximo para O3 y PM2.5 en cada día.
- Ahora obtén el mínimo, el promedio y el máximo de cada variable del clima, creando una tabla como la que sigue:

	MAX									MIN							AVERAGE						
Date	MXO3	MXPM	MXPRS	MXRAI	MXRH	MXSR	MXTOU	MXWS	MXWD	MNPRS	MNRAI	MNRH	MNSR	MNTOU	MNWS	MNWD	AVPRS	AVRAIN	AVRH	AVSR	AVTOU	AVWSP	AVWD
15/01/2015	17	80	726	0.02	96	0.18	7.43	8.4	358	721	0	93	0	3.78	2.8	2	723	0	95.1	0.04	5.27	5.3	171
16/01/2015	10	52	722	0.01	96	0.17	7.53	5.9	354	718	0	87	0	4.19	1.7	1	720	0	93.8	0.04	5.8	3.65	60.6
17/01/2015	20	133	723	0	96	0.36	14.6	5.9	356	718	0	73	0	5.84	1.1	3	720	0	90.1	0.05	8.4	2.83	167
18/01/2015	31	118	733	0.01	94	0.44	17.4	12.7	352	724	0	46	0	6.84	1.4	3	728	0	70.9	0.12	11.9	5.96	178
19/01/2015	36	55	734	0	80	0.46	12.9	13.3	119	730	0	62	0	6.87	3.3	51	732	0	72.2	0.09	9.7	7.3	99.4
20/01/2015	49	105	730	0	85	0.53	17.3	8.2	327	727	0	58	0	7.5	2.4	18	728	0	73.6	0.11	12.2	4.37	96.2
21/01/2015	29	49	735	0.02	94	0.46	17.9	13.9	360	727	0	71	0	7.98	0.9	0	730	0	85.7	0.1	12.4	6.23	221
22/01/2015	31	52	735	0	88	0.16	7.49	11.1	358	726	0	69	0	3.66	1.5	0	731	0	79.1	0.04	4.69	7.16	142
23/01/2015	13	40	730	0.07	93	0.08	7.44	9.5	359	724	0	85	0	4.17	1.9	281	727	0	88.2	0.02	5.66	5.61	328
24/01/2015	14	18	730	0.05	94	0.13	4.91	7.4	359	725	0	91	0	2.86	1.8	1	727	0.01	93.3	0.03	3.84	5.26	222
25/01/2015	41	66	725	0.01	95	0.58	16.4	9.3	356	721	0	65	0	3.04	0.9	6	723	0	84.3	0.18	9.19	3.78	141

Ahora intenta crear nuevos modelos de regresión que, dadas los valores de variables del clima máximo, promedio y mínimas, predigan los valores máximos de los contaminantes y reflexiona:

¿Mejoraron los parámetros de desempeño de los modelos?

Al realizar los cambios recomendados se redujo la variabilidad de los valores a predecir, ya que se usa el máximo valor durante el día en lugar de mediciones obtenidas cada hora. Sin embargo, la mayor estabilidad tiene como costo la reducción de los datos a utilizar, por la tanto los modelos tendrán menos datos para poder generalizar el comportamiento de los datos en casos que no se le hayan presentado. Este efecto se puede apreciar en los desempeños de los modelos entrenados, que son los mismo que se usaron en la solución del reto, los cuales no mejoraron, como se muestra a continuación para la predicción de ambas partículas de interés:

Tabla 4.- Métricas de desempeño de los modelos usados para predecir partículas O3

	LinearRegressor	RandomForest	GradientBoosting
rmse	11.810044	12.781511	15.676879
mse	139.477138	163.367026	245.764542
mae	9.070269	9.826226	12.311408
r2	0.629869	0.566473	0.347815

Tabla 5.- Métricas de desempeño de los modelos usados para predecir partículas PM2.5

	LinearRegressor	RandomForest	GradientBoosting
rmse	26.156175	20.361725	20.999909
mse	684.145515	414.599841	440.996161
mae	17.548898	14.762894	15.874406
r2	-0.146480	0.305220	0.260986

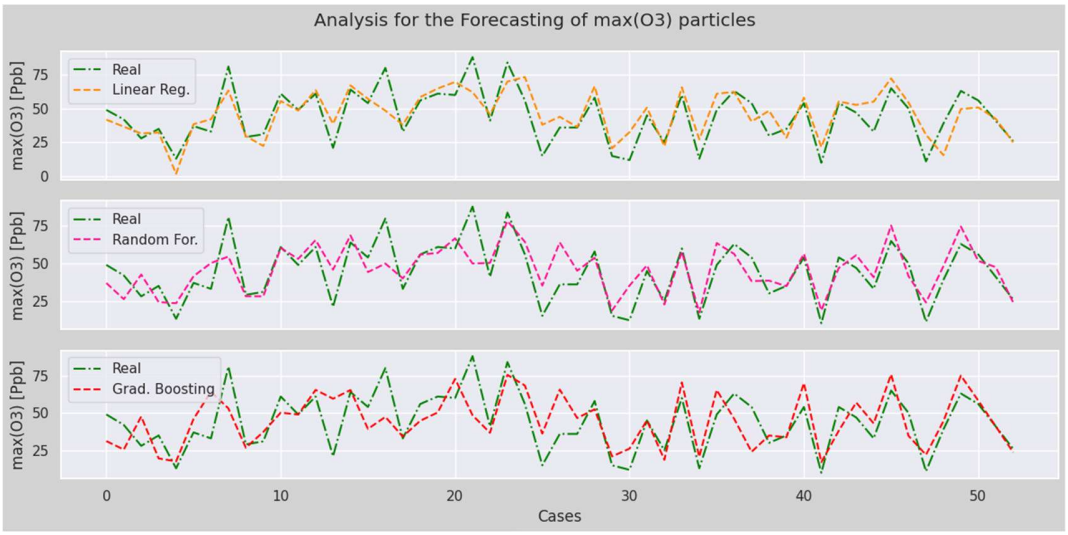

Ilustración 4,- Comparativo al utilizar los modelos entrenados para predecir partículas O3

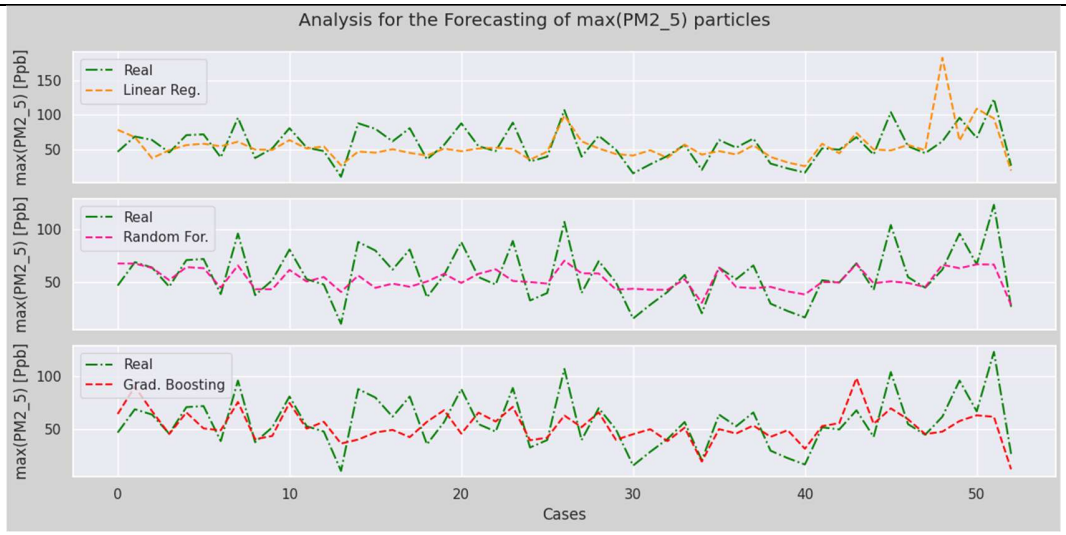

Ilustración 5.- Comparativo al utilizar los modelos entrenados para predecir partículas PM2.5