# contaminAND_gr_congresos_RandomForests_datasetcompleto

Para acceder al codigo completo entrar en el siguiente link (Es un Databricks notebook.):
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3373319835044876/3115698141222039/5920673496494609/latest.html
Para usarlo necesitas cargar este archivo(el dataset):
https://github.com/oslugr/contaminAND/blob/master/datos/contaminAND-gr-congresos.csv

UTILIZAMOS RANDOM FORESTS PARA ANALIZAR LA RELACION ENTRE LAS DISTINTAS VARIABLES, ASI COMO LA IN
FLUENCIA QUE TIENEN LA HORA DEL DIA O EL DIA DE LA SEMANA. ESTE ALGORITMO NOS VA A DAR
PARA CADA VARIABLE QUE ANALICEMOS (NO2,CO2,PART,03 Y SO2) CUALES SON LAS VARIABLES QUE MEJOR "EXP
LICAN" A LA ANALIZADA. DE ESTA FORMA PODEMOS VER DE MANERA RAPIDA SI LOS DATOS TIENEN
SENTIDO Y POR LO TANTO SE PUEDEN CONSIDERAR LAS MEDICIONES COMO FIABLES, O SI PRESENTAN ALGUNAS I
NCONGRUENCIAS. AQUI ANALIZAMOS SOLO LOS DATOS DEL PALACIO DE CONGRESOS PARA TODOS LOS
DATOS QUE HAY DESDE 2009 HASTA MARZO DE 2017

Partimos del siguiente archivo de datos(07/2009-03/2017) provenientes del medidor de calidad del aire instalado en el Palacio de Congresos de Granada

```
+------------------+---+----+----+---+---+
|              date|NO2|  CO|PART| O3|SO2|
+------------------+---+----+----+---+---+
|2009-07-20T00:10:00| 28|1708|  50| 74|  6|
|2009-07-20T00:20:00| 28|1694|  51| 76|  6|
|2009-07-20T00:30:00| 27|1700|  45| 77|  6|
|2009-07-20T00:40:00| 25|1662|  33| 79|  6|
|2009-07-20T00:50:00| 26|1680|  23| 76|  6|
|2009-07-20T01:00:00| 25|1673|  23| 78|  6|
|2009-07-20T01:10:00| 28|1692|  29| 66|  6|
|2009-07-20T01:20:00| 34|1729|  31| 48|  6|
|2009-07-20T01:30:00| 35|1723|  35| 53|  6|
|2009-07-20T01:40:00| 33|1717|  42| 58|  6|
+------------------+---+----+----+---+---+
only showing top 10 rows
```

Out[77]:

| | NO2 | CO | PART | O3 | SO2 | WEEKDAY | TIME_minutesofday | WEEKEND_VD | WEEKEND_SD |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 28.0 | 1708.0 | 50.0 | 74.0 | 6.0 | 0 | 10 | 1 | 1 |
| 1 | 28.0 | 1694.0 | 51.0 | 76.0 | 6.0 | 0 | 20 | 1 | 1 |
| 2 | 27.0 | 1700.0 | 45.0 | 77.0 | 6.0 | 0 | 30 | 1 | 1 |
| 3 | 25.0 | 1662.0 | 33.0 | 79.0 | 6.0 | 0 | 40 | 1 | 1 |
| 4 | 26.0 | 1680.0 | 23.0 | 76.0 | 6.0 | 0 | 50 | 1 | 1 |

```
--------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

Nuestros datos presentan cerca de 390000 filas. Podemos ver a continuacion como la columna relati
va a la concentracion de particulas("PART") es la que presenta una mayor cantidad de
filas vacias/erroneas. Aun asi esto una perdida de datos no muy significativa: ~ 3-4%

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 388892 entries, 0 to 388891
Data columns (total 7 columns):
NO2                  381736 non-null float64
CO                   383612 non-null float64
PART                 374120 non-null float64
O3                   386068 non-null float64
SO2                  384789 non-null float64
WEEKDAY              388892 non-null int64
TIME_minutesofday    388892 non-null int64
dtypes: float64(5), int64(2)
memory usage: 20.8 MB
```

```
Out[11]:
NO2                   True
CO                    True
PART                  True
O3                    True
SO2                   True
WEEKDAY              False
TIME_minutesofday    False
dtype: bool
```

```
Out[12]:
NO2                   7156
CO                    5280
PART                 14772
O3                    2824
SO2                   4103
WEEKDAY                  0
TIME_minutesofday        0
dtype: int64
```

```
--------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

El dataset queda de la siguiente forma despues de:
    Limpiarlo y dejarlo preparado para que se pueda aplicar Random Forests.
    Añadir columnas relativas al dia de la semana, hora del dia y si es fin de semana o no (consi
derando V-D y solo S-D)

```
Out[79]:
    NO2      CO  PART    O3  SO2  WEEKDAY  TIME_minutesofday  WEEKEND_VD  \
0  28.0  1708.0  50.0  74.0  6.0        0                 10           1
1  28.0  1694.0  51.0  76.0  6.0        0                 20           1
2  27.0  1700.0  45.0  77.0  6.0        0                 30           1
```

```
3  25.0  1662.0  33.0  79.0  6.0        0                 40           1
4  26.0  1680.0  23.0  76.0  6.0        0                 50           1

   WEEKEND_SD
0           1
1           1
2           1
3           1
4           1
```

---

ANALIZAMOS AHORA QUE VARIABLES INFLUYEN MAS EN LOS NIVELES DE NO2

Precision del modelo para el analisis de la variable "NO2"  0.810692459138

Variables que explican los valores de NO2

```
O3                  0.483569
TIME_minutesofday   0.206430
CO                  0.102506
PART                0.081853
SO2                 0.071660
WEEKDAY             0.048280
WEEKEND_SD          0.002895
WEEKEND_VD          0.002806
dtype: float64
```

LA CANTIDAD DE O3(48%) Y LA HORA DEL DIA(20%) ES LO QUE MAS INFLUYE EN LOS NIVELES DE NO2. TAMBIE
N TIENE INFLUENCIA EL CO.EL RESTO DE VARIABLES SON INSIGNIFICANTES

---

ANALIZAMOS AHORA QUE VARIABLES INFLUYEN MAS EN LOS NIVELES DE PARTICULAS

Precision del modelo para el analisis de la variable "PART":  0.464859024404

Variables que explican los valores de PART

```
  CO                  0.247480
NO2                 0.217162
TIME_minutesofday   0.164688
O3                  0.149365
SO2                 0.124526
WEEKDAY             0.070919
WEEKEND_VD          0.013407
WEEKEND_SD          0.012453
dtype: float64
```

HAY 5 VARIABLES QUE PARECEN TENER IMPORTANCIA EN LOS NIVELES DE PARTICULAS. EL C0,NO2,HORA DEL DI
A, 03 Y S02.

---

```
ANALIZAMOS AHORA QUE VARIABLES INFLUYEN MAS EN LOS NIVELES DE S02


('Precision del modelo para el analisis de la variable "S02": ', 0.64684927680372972)

Variables que explican los valores de S02

NO2                  0.257351
TIME_minutesofday    0.235044
CO                   0.199331
O3                   0.130597
PART                 0.117834
WEEKDAY              0.046667
WEEKEND_VD           0.007258
WEEKEND_SD           0.005917
dtype: float64

Hay 5 variables que parecen tener importancia en los niveles de S02. N02 y HORA DEL DIA sobre tod
o. Tambien C0, 03, PARTICULAS

--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------

ANALIZAMOS AHORA QUE VARIABLES INFLUYEN MAS EN LOS NIVELES DE O3


Precision del modelo para el analisis de la variable "O3":  0.817197651244

Variables que explican los valores de 03

NO2                  0.477318
TIME_minutesofday    0.251348
CO                   0.112602
PART                 0.069005
SO2                  0.054399
WEEKDAY              0.027242
WEEKEND_VD           0.004140
WEEKEND_SD           0.003947
dtype: float64

La cantidad de N02(47%) es lo que mas influye en los niveles de 03. Influyen tambien la hora del
dia (25%) y la cantidad de CO(11%). El resto de parametros son insignificantes

--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------

CONCLUSIONES INICIALES:

Los analisis de N02 Y O3, cuyos modelos han alcanzado una precision por encima del 80% son acepta
bles. Sin embargo los analisis de SO2 (con un ~65% de precision) y sobre todo de PARTIculas (45%)
 son demasiado pobres y habria que realizar distintas iteraciones variando los parametros del alg
oritmo para alcanzar mayores precisiones
```